2008

An Integrated Knowledge Discovery and Data Mining Process Model

Sumana Sharma
Virginia Commonwealth University

School of Business
Virginia Commonwealth University

This is to certify that the dissertation prepared by SUMANA SHARMA entitled "AN INTEGRATED KNOWLEDGE DISCOVERY AND DATA MINING PROCESS MODEL" has been approved by her committee as satisfactory completion of the dissertation requirement for the degree of Doctor of Philosophy

_____
Dr. Kweku-Muata Osei-Bryson (Committee Chair), School of Business
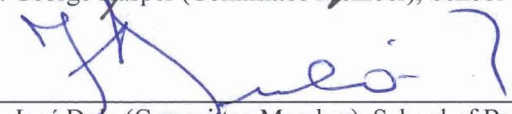
_____
Dr. Richard Redmond (Committee Co-Chair), School of Business

_____
Dr. Roland Weistroffer (Committee Member), School of Business

_____
Dr. George Kasper (Committee Member), School of Business

_____
Dr. José Dula (Committee Member), School of Business

_____
Dr. Richard Redmond (Chair – Department of Information Systems), School of Business

_____
Dr. Allen S Lee, Associate Dean for Research and Graduate Studies, School of Business

_____
Dr. F. Douglas Boudinot, Dean of the School of Graduate Studies

11/05/2008

**"AN INTEGRATED KNOWLEDGE DISCOVERY
AND DATA MINING PROCESS MODEL"**
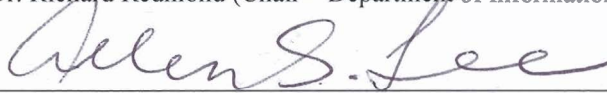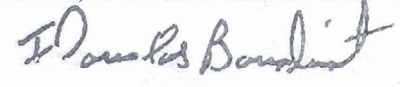
A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at Virginia Commonwealth University.

by

SUMANA SHARMA
Post Graduate Diploma (Honors) in Telecom Management
Concentration: Systems and Applied Technologies
Symbiosis Institute of Telecom Management, India, 2002

Bachelor of Engineering (Honors) in Electronics and Communication,
Oriental Institute of Science and Technology, India, 2000

Directors:

Dr. Kweku-Muata Osei-Bryson
Professor, Information Systems

Dr. Richard Redmond
Associate Professor and Department Chair, Information Systems

Virginia Commonwealth University
Richmond, Virginia
December 2008

Dedication

For my loving husband Gaurav
And the world's best parent's ~
Gyan Vardhan Pathak (my dad) and Sudha Pathak (my mom)

Acknowledgement

As my journey in the PhD program comes to an end, I fondly remember all the individuals who have made this piece of work possible.

I must begin with my dissertation chair Dr. Kweku-Muata Osei-Bryson under whose tutelage, I completed my dissertation. I feel extremely fortunate in that I had the opportunity to work with him, and regard the moment when I got introduced to him as the turning point in my academic life. He has taught me how to do research and become an independent thinker. He has inspired me through his love for teaching and the amount of energy and emotions that he invests in his student's welfare. Dr. Osei-Bryson, you are a true scholar with a gentle soul, and I am so grateful that I had you as my advisor.

I was also very fortunate to have Dr. Richard Redmond as my dissertation co-chair. Each one of our discussions helped me clarify my understanding on the subject matter as well as find ways to tackle roadblocks and issues. We students at VCU's Information Systems department are also very fortunate to have him as our department chair. As a PhD student, I often turned to him for guidance and help and he always responded with compassion and concern for my smallest of problems. Dr. Redmond, I hope that someday I can make the same positive difference in the lives of other students that you made in my life.

I would also like to express my sincere gratitude towards my committee members, Dr. Roland Weistroffer, Dr. George Kasper and Dr. Jose Dula. Dr. Weistroffer diligently read various drafts of this work and helped me in addressing several issues. He asked me thought provoking questions that helped improve the quality of this work. Dr. George Kasper helped me clarify and frame my research questions and research objective, prodded me about the suitability of various statistical tests and encouraged me to do detailed data analysis. By being outside of the department of Information Systems, Dr. Jose Dula brought an independent perspective to this research. I benefitted from each one of the discussions where he asked me numerous questions describing my work. The coherence and soundness of my responses alerted me towards areas needing further work and clarification.

I would like to also express my sincere gratitude towards some other faculty, who have had profound positive influence over me during the past four years. Special thanks to Dr. Allen Lee, Dr. Geoffrey Hubona, Dr. Steven Custer, Dr. Larry Williams, Dr. Peg Williams and Dr. Pam Kovacs. The presence of these individuals, taught me the skills and perspective required to succeed in the program.

I thank the entire staff, and past and present PhD students at VCU's Information Systems department who have influenced me and provided intellectual nourishment. I

thank you for truly shining like the north star, eternally constant, without a flicker. I love you!

I would be remiss if I did not thank the two little babies who have been growing inside me for the past eight months. You have been incredibly supportive and have allowed me to complete this work in a timely manner. Many anxious moments were relieved through your gentle kicking and squirming, and you helped me stay cognizant of Almighty's larger scheme of things. My precious little ones, your Dad and I cannot wait to hold you in our arms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

DM	Data Mining
KDD	Knowledge Discovery in Databases
KDDM	Knowledge Discovery and Data Mining
CRISP-DM	Cross Industry Standard Process for Data Mining
IKDDM process model	Integrated Knowledge Discovery and Data Mining Process Model
BSC	Business success criteria
DMSC	Data mining success criteria
VFT	Value focused thinking
SMART	Specific, measurable, achievable, relevant and timely
GQM	Goal Question Metric
BUSREQ	Business Requirements
SAS EM	SAS Enterprise Miner
PEOU	Perceived Ease of Use
PU	Perceived Usefulness
US	User satisfaction
PSQ	Perceived Semantic Quality

ABSTRACT

Enterprise decision making is continuously transforming in the wake of ever increasing amounts of data. Organizations are collecting massive amounts of data in their quest for knowledge nuggets in form of novel, interesting, understandable patterns that underlie these data. The search for knowledge is a multi-step process comprising of various phases including development of domain (business) understanding, data understanding, data preparation, modeling, evaluation and ultimately, the deployment of the discovered knowledge. These phases are represented in form of Knowledge Discovery and Data Mining (KDDM) Process Models that are meant to provide explicit support towards execution of the complex and iterative knowledge discovery process. Review of existing KDDM process models reveals that they have certain limitations (fragmented design, only a checklist-type description of tasks, lack of support towards execution of tasks, especially those of the business understanding phase etc) which are likely to affect the efficiency and effectiveness with which KDDM projects are currently carried out. This dissertation addresses the various identified limitations of existing KDDM process models through an improved model (named the Integrated Knowledge Discovery and Data Mining Process Model) which presents an integrated view of the KDDM process and provides explicit support towards execution of each one of the tasks outlined in the model. We also evaluate the effectiveness and efficiency offered by the IKDDM model against CRISP-DM, a leading KDDM process model, in aiding data mining users to execute various tasks of the KDDM process. Results of statistical tests

indicate that the IKDDM model outperforms the CRISP model in terms of efficiency and effectiveness; the IKDDM model also outperforms CRISP in terms of quality of the process model itself.

1   INTRODUCTION

We are drowning in information, but starving for knowledge.

- John Naisbett

1.1   Background

Data has emerged as a new found source of competitive advantage in an era where traditional bases of competition have largely evaporated (Davenport and Harris 2007). This competitive advantage is based on the knowledge gained from analysis of data and has catapulted to the forefront, fields like data mining and knowledge discovery, that offer techniques and processes for extracting this knowledge. Given the recognition that data needs to be first collected before it can be mined for knowledge has resulted in explosive growth in the size of databases (Fayyad, Piatetsky-Shapiro et al. 1996a) – some even argue that 'our ability of collecting and storing different types of data, has far outpaced our abilities to analyze and extract knowledge from this data' (Fayyad and Uthurusamy 2002).

Regardless the quest for discovering knowledge (interesting patterns) from large amounts of data remains the sole motive behind the vast mountains of data being

1

created by companies (Han and Kamber 2006). But how do we mine data to reach the often elusive end goal - knowledge? The guidance for conducting the 'knowledge discovery process' is encapsulated in form of knowledge discovery and data mining (KDDM) process models, sometimes also referred to as methodologies, which act as a road map for implementing the knowledge discovery process. Most process models recommend development of domain or business understanding, data understanding, data preparation, modeling and evaluation as building blocks to discovering knowledge.

The purpose of these KDDM process models is to guide the user through each step of mining data to discover knowledge. Given this role, the design of such models significantly affects the efficiency and effectiveness with which the knowledge discovery process can be executed. Existing process models suffer from certain limitations. Many process models only describe the process of knowledge discovery in form of a small number of tasks, which are not representative of the reality of this complex process. The only KDDM process model which is an exception (CRISP-DM) offers minimal support towards execution of the long list of tasks recommended by it. Despite the significant impact KDDM process models have by way of their design on the outcome of the KDDM process, existing models at best provide only broad guidance to the user in terms of "how" this process can be executed.

Accordingly the purpose of this dissertation is to systematically uncover deficiencies in existing KDDM process models and address them through an improved model design. The new KDDM model will be designed such that it can be relied upon

to provide explicit support to even the average user in implementing the seemingly complex and technical, knowledge discovery process.

The remainder of this chapter is organized as follows. We start out in Section 1.2 by discussing the difference between the terms knowledge discovery, knowledge discovery in databases (KDD) and knowledge discovery and data mining (KDDM), in an attempt to dispose off any terminological ambiguity surrounding the usage of these (related, but distinct) terms in this dissertation. Section 1.3 presents the conceptual framework used by this dissertation to study and evaluate existing KDDM process models, as well as to design a new KDDM model. Section 1.4 presents the research objective, and Section 1.5 presents the guiding research questions. Section 1.6 presents the organization of the remaining chapters of this dissertation.

1.2 Important Definitions

Discovery of knowledge nuggets requires both the use of 'tools' that can aid in analyzing these volumes of data as well as a 'process' that prescribes how the journey from data to discovering knowledge is to be completed. A cursory glance at the literature reveals three definitional issues created due to the tool/process distinction (Figure 1-1). Readers who are familiar with the definitional issues cited below can directly proceed to section 1.3.

1.2.1 Data Mining (DM) versus Knowledge Discovery in Databases (KDD)

There are three prevalent interpretations of data mining in the literature. (1) Data Mining is used to represent a set of tools used for analyzing data; (2) Data Mining is used to describe the process of discovering nuggets of knowledge in data; and (3) Knowledge Discovery in Databases (KDD) is used to describe the process of discovering nuggets of knowledge in data. The latter two are more problematic as they are described using the same definition "non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" leading to an impression that data mining and knowledge discovery in databases are synonymous concepts.

Fayyad et al. (1996a) who are credited with proposing the above definition used it to describe the knowledge discovery in databases (KDD) process ranging from incorporation of prior knowledge, to creation of target data set, its cleaning and preprocessing, to application of data mining algorithms, identification of interesting patterns, evaluation of the patterns and consolidation of the discovered knowledge [Figure 1-1 (c)]. They specifically positioned data mining as a step in the overall KDD process where the user applied selected data mining algorithms to identify interesting patterns. Clearly, they did not envision data mining as a process, but rather as a step in the KDD process. however Fayyad et al. (1996a) among others (Reinartz 2002; Han and Kamber 2006; Kurgan and Musilek 2006) acknowledge that today data mining and KDD have come to be used interchangeably in the literature.

4
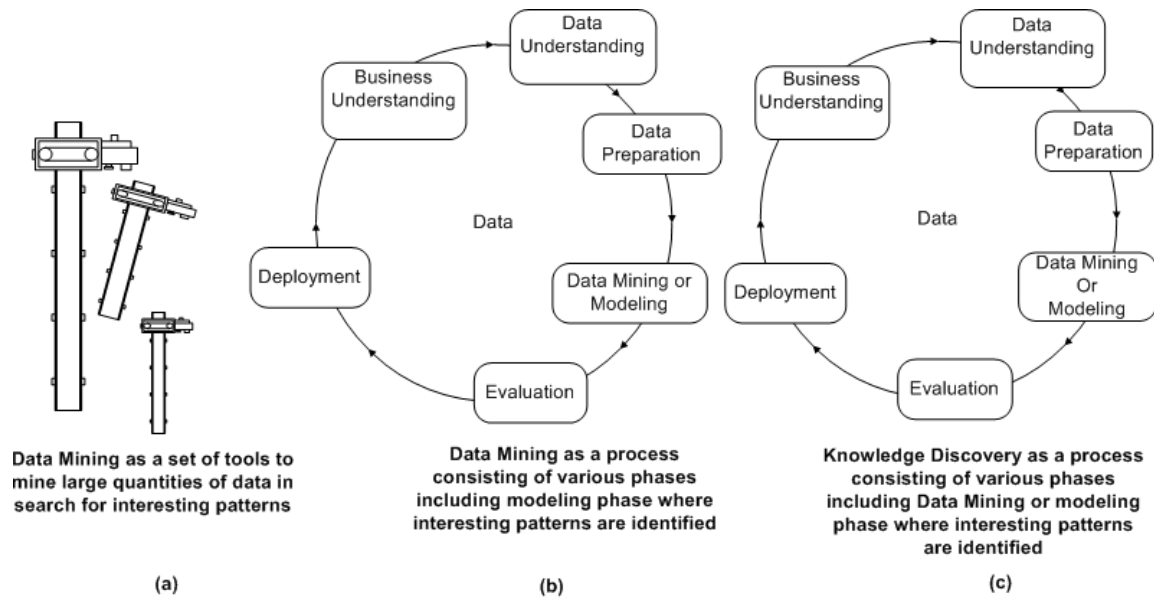
Figure 1-1: Various interpretations of Data Mining

1.2.2 Knowledge Discovery and Data Mining (KDDM)

Fayyad et al. (1996a) are of the opinion that while the MIS community adopted the term data mining (DM), the machine learning community continued with using KDD to describe the knowledge discovery process. They attempted to build bridges between the two communities by proposing the use of 'knowledge discovery and data mining' (KDDM) and argued that this term was more appropriate than data mining or KDD alone, as it signified the importance of two equally critical aspects: the (1) the overall knowledge discovery process which includes pre-processing and post-

5

processing of data as well as interpretation of the discovered patterns as knowledge, and (2) particular data mining methods and algorithms aimed at solely extracting patterns from data' (p. 4).

Review of the IS academic and practitioner data mining and knowledge discovery literature published during the last decade however reveals that the former term (i.e. data mining) has continued to become more popular and has even stimulated further adoption of this term even though many researchers acknowledges the history and difference between these terms (Han and Kamber 2006). This dissertation adopts the view that careful usage of various terms is essential to avoid ambiguous interpretations of these related but distinct concepts.

It appears that use of the term *data mining* (widely utilized in the Information Systems community) may blind us to the importance of the context and the overall knowledge discovery process resulting in 'data dredging' or 'fishing', the blind application of data mining methods (Fayyad, Piatetsky-Shapiro et al. 1996a). Fortunately, there has recently been a renewed call for use of the term *Knowledge Discovery and Data Mining (KDDM)* in favor of the terms *knowledge discovery in databases* or *data mining* (Reinartz 2002; Kurgan and Musilek 2006) to describe the knowledge discovery process.

1.2.3   Rationale for adopting the term KDDM over DM or KDD

We concur with the above authors and adopt the term *Knowledge Discovery and Data Mining (KDDM)* in this dissertation to describe the overall knowledge discovery process. The rationale for adoption of this term is summarized below:

1. Inclusion of the term 'knowledge discovery' reminds us of the importance of context in implementing the knowledge discovery process and can therefore help to avoid the blind application of data mining methods that may result if we use the term *Data Mining* alone. Inclusion of the term 'data mining' can help maintain or even enhance its appeal in the Information Systems Community where this term is well understood and popularly used by researchers and practitioners.

2. The combined term emphasizes the holistic nature of the knowledge discovery process and acknowledges *data mining* as one of its important constituents.

In the next section we discuss the importance of KDDM and the process models that can be used to implement the KDDM process and present the research opportunities identified and addressed in this dissertation.

1.3   Nature of KDDM process models

In the context of KDDM process models, the term 'process' is used in the activity-oriented sense and refers to a set of partially ordered steps intended to reach a goal

7

(Feiler and Humphrey 1993). Process models are processes of the same nature that are classified together in a model. The goals of process models can be descriptive, prescriptive or explanatory (Rolland 1998). These are briefly described below:

- Descriptive – takes the viewpoint of an external observer and tracks what actually happens in a process.
- Prescriptive – defines the desired processes and how they should/could/might be performed.
- Explanatory – explores and evaluate several possible courses of actions based on rational arguments.

KDDM process models belong to the category of 'prescriptive' process models. The KDDM process model being proposed in this dissertation is also prescriptive in nature. The motivation for designing a new KDDM process model stems from two main observations: (1) that existing process models are too broad and generic and (2) the enactment of the processes specified by them is not supported through suitable tools and/or detailed steps thereby leading to difficulties in implementing such process models.

1.4 Conceptual framework for analyzing KDDM process models

We use the process domain architecture proposed by Dowson (1993) as a conceptual framework to critically analyze existing KDDM process models and to design an improved KDDM process model (Figure 1-2).

8
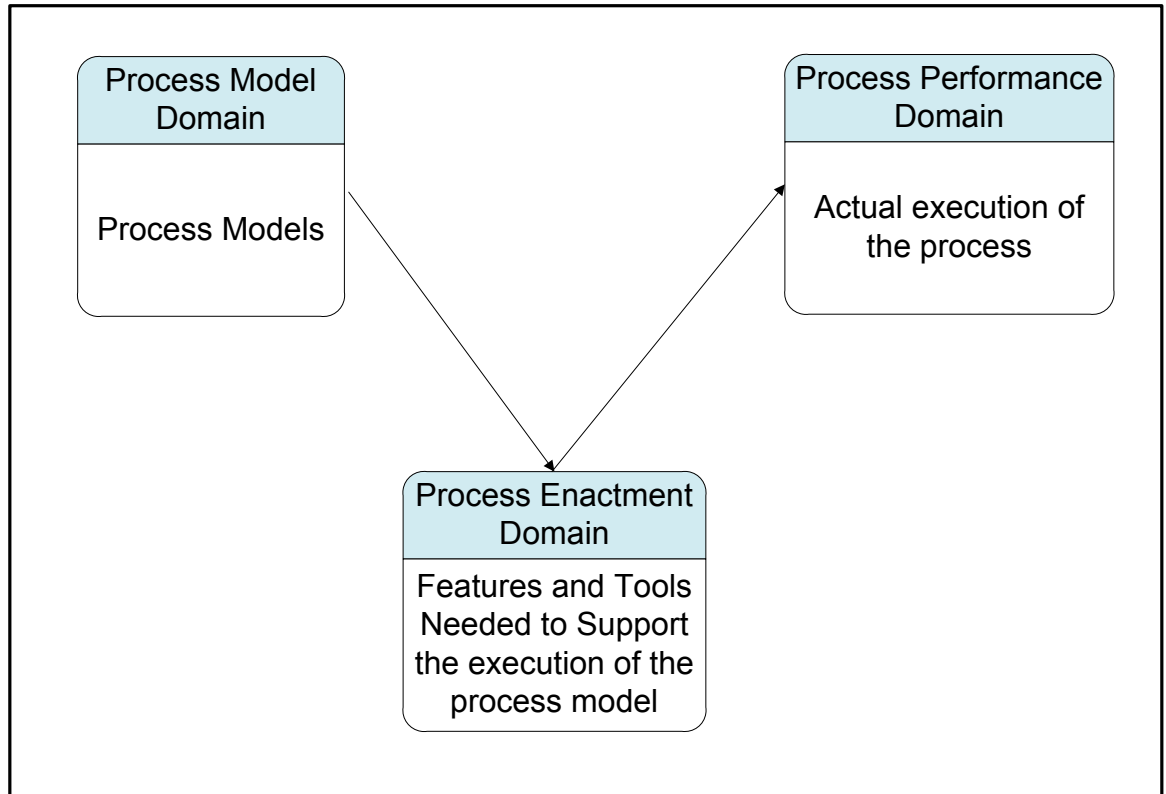
Figure 1-2: Interacting Process Domains (Dowson 1993)

According to Dowson, the usage world of a process (where goals of processes are established, and the range of facilities for process performance are determined) can be viewed as comprising of three interacting domains: (1) the process model domain; (2) the process enactment domain and (3) the process performance domain.

The *process model domain* contains the process models. A process model influences the manner in which the process is performed. Thus the design of the process

9

model is likely to significantly affect the outcome of the process. The *process enactment domain* contains the features and tools needed to support the execution of the processes recommended by the process model. The *process performance domain* deals with the 'actual' activities that are conducted by human users, software tools etc when the process is actually executed. It is important to note that the actual activities conducted during model execution may be different from what are recommended by the process model, but these do not become obvious till the model is actually executed in a relevant context.

This research is focused on the deficiencies in existing KDDM process models as found in the first two domains; the process model domain which contains the process model that directs how the process should be performed; and the process enactment domain which contains the features, tools to support the implementation of processes recommended by the process model. As stated earlier, the limitation in the process model domain lies in the low level of granularity in the design of existing process models. As a result, the models are highly generic and do not specify the complete set of processes required to implement the KDDM process. The limitation in the process enactment domain lies in the fact that the processes recommended by the KDDM process models are not adequately supported by features or tools that can be used to implement them. Lack of support for process enactment is also a serious issue as it is likely to lead to critical processes not being executed. Given the numerous dependencies in the KDDM process (where tasks are dependent on output of other tasks

10

for their execution), this has an effect on the quality of the outcome of the KDDM project. In Chapter 2 of the dissertation, we will be using the process domain architecture as a conceptual framework to systematically uncover deficiencies in the modeling domain and enactment domain, of existing KDDM process models. These limitations will be used to formulate a set of requirements for the new KDDM model being designed through this research. But first we define the research objective and research questions guiding this dissertation.

1.5   Research Objective and Scope

The research objective of this dissertation is to present a new Knowledge Discovery and Data Mining process model, and the set of features and tools that support its enactment.

The scope of the evaluation of this model will include the following phases of the KDDM process: business understanding, data understanding, data preparation, modeling and evaluation. The final phase of the process, 'deployment' is excluded from the scope of the evaluation presented in this research.

1.6   Research Questions

The following research questions anchor the research effort addressed by this dissertation:

11

- What are the limitations of existing KDDM process models and how do they affect the outcome of the KDDM process?

- How can an improved KDDM process model be developed to address these limitations?

1.7 Outline of dissertation

Chapter 2 presents a description of KDDM process models proposed in the academic and practitioner literature. Using Dowson's (1993) process domain architecture as the guiding conceptual framework, the process models proposed in the prior literature are analyzed and their deficiencies uncovered. The limitations identified are used to develop a set of design requirements to be fulfilled through the artifact (a new KDDM process model) being designed through this research. The significance of the new KDDM process model is also discussed.

In Chapter 3, we review existing literature and present a discussion of concepts and techniques relevant to the execution of the KDDM process. These concepts and techniques are used as the foundation for populating the process enactment domain of the KDDM process model being designed through this research.

Chapter 4 presents the design science research methodology that is being utilized by this research to design the artifact in form of the improved KDDM process model. The rationale for adopting this methodology its research guidelines and the application of these guidelines in the dissertation are being presented.

12

Chapter 5 presents the new KDDM process model, named the Integrated Knowledge Discovery and Data Mining (IKDDM) process model that has been designed through this dissertation. The detailed design of the process model, along with the features and tools to support its enactment are also presented. The chapter concludes with a summarization of how the IKDDM process model fulfills the design requirements established earlier in Chapter 2.

Chapter 6 presents the results of evaluation of the IKDDM process model to assess its utility and efficacy as compared to a leading KDDM process model.

Using the guidelines of design science research methodology as an anchor, Chapter 7 recapitulates the contribution and significance of this research. The chapter concludes with a discussion of limitations of this research and directions for future research endeavors.

2 PRIOR RESEARCH: IDENTIFYING GAPS & FORMULATING REQUIREMENTS

2.1 Survey of Existing KDDM Process Models

In this section we discuss five leading KDDM process models that have been proposed in the extant literature. These include a nine step model proposed by Fayyad, Piatetsky-Shapiro et al. (1996a); a five step model proposed by Cabena et al. (1998); a six step model proposed by Cios et al. (2000) and a multi-step model in form of CRISP-DM (2003). We also discuss the model proposed by Berry and Linoff (1997) authors of the book 'data mining techniques for marketing, sales and customer relationship management' who have done some early work in this area. Of these models, CRISP-DM has been proposed in the practitioner literature, while all others models have been proposed in the academic literature.

- **Fayyad, Piatetsky-Shapiro et al. (1996a)**

The Fayyad et al's. (1996a) KDDM model consists of nine steps, which are outlined below.

1. Developing and understanding the application domain: This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.

14

2. Creating a target data set: Here the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.

3. Data cleaning and preprocessing: This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.

4. Data reduction and projection: This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

5. Choosing the data mining task: Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.

6. Choosing the data mining algorithm: The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.

7. Data mining: This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.

8. Interpreting mined patterns: Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.

9. Consolidating discovered knowledge: The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge.

- **Berry and Linoff (1997)**

Berry and Linoff (1997) presented a four step methodology consisting of following steps: Identifying the Problem: Analyzing the Problem, Taking Action, and Measuring the outcome. They also specify the following 11 steps to further describe their proposed approach.

1. Translate the business problem into a data mining problem.
2. Select appropriate data.
3. Get to know the data.
4. Create a model set.
5. Fix problems with the data.
6. Transform data to bring information to the surface.
7. Build models.
8. Assess models.
9. Deploy models.
10. Assess results.

11. Begin again.

- **Cabena et al. (1998)**

Step 1: Business Objectives Determination: This step involves clearly defining the business problem or challenge. The minimum requirements are a perceived business problem or opportunity and some level of executive sponsorship. This step in the process is also the time at which to start setting expectations.

Step 2: Data Preparation: Cabena et al. (1998) note that Data preparation is the most resource-consuming step in the process, typically requiring up to 60% of the effort of the entire project. This step comprises 3 sub-tasks:

1. *Data Selection*: Identify all internal or external sources of information and select which subset of the data is needed for the data mining application.

2. *Data Preprocessing*: Study the quality of the data to pave the way for further analysis and to determine the kind of mining operation that will be possible and worth performing.

3. *Data Transformation*: During data transformation, the preprocessed data is transformed to produce the analytical data model. The analytical data model is an informational data model, and it represents a consolidated, integrated, and time-dependent restructuring of the data selected and preprocessed from the various operational and external sources. This is a crucial phase as the accuracy

17

and validity of the final result depend vitally on how the data analyst decides to structure and present the input.

Step 3: Data Mining: This is the step in which the actual data mining takes place. The objective is clearly to apply the selected data mining algorithm or algorithms to the preprocessed data. The actual details of the data mining step will vary with the kind of application that is under development. The author presents the example that while in the case of database segmentation, one or two runs of the algorithm may be sufficient, development of a predictive model will be a cyclical process where the models will be repeatedly trained and retrained on sample data before being tested against the real database.

Step 4: Analysis of Results: According this process model the analysis of results is inseparable from the data mining step in that the two are typically linked in an interactive process. The specific activities in this step depend very much on the kind of application that is being developed. However, the main objective remains the same, that is, to interpret and evaluate the output from the data mining step.

Step 5: Assimilation of Knowledge: This step closes the loop, which was opened when the business objectives were set at the beginning of the process. The objective now is to put into action the commitments made in the opening step, according to the new, valid and actionable information from the previous process steps. The two main challenges in this step are: to present the new findings in a convincing,

18

business-oriented way, and to formulate ways in which the new information can be best exploited.

- **CRISP-DM (2003)**

CRISP-DM (an acronym for CRoss Industry Standard Process was data mining) is an industry-neutral, tool-neutral data mining process model that was conceived in late 1996 by three leaders of the then immature data mining market: Daimler (then Daimler-Benz), SPSS (then ISL) and NCR. At the time, Daimler was ahead of other industrial and commercial organizations as it had already gained experience in data mining by applying it to its business operations. SPSS too had data mining experience owing to the data mining services it had been providing since the 1990's. It was also the first vendor to launch commercial data mining workbench called 'Clementine' in 1994. NCR too brought in data mining expertise owing to its experience of offering data mining services through its teams of consultants and technology specialists, in order to deliver added value to its Teradata data warehouse customers.

In 1997, a consortium was formed with the goal of formalizing the experience of the various real-world organizations that had been practicing data mining, in form of a process model. One of the prime characteristics of this project was the focus on creating a non-proprietary and freely available model that would assist in execution of data mining projects.

19

CRISP-DM describes the life cycle of a data mining project in form of six different phases, namely, business understanding, data understanding, data preparation, modeling, evaluation and deployment (Figure 2-1). It also describes the tasks and activities that need to be carried out in each of these phases (

Table 2-1). A description of the six phases of the CRISP-DM process model is presented next.

Different phases of the CRISP-DM process model

Phase 1 - Business understanding: This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Phase 2 - Data understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Phase 3 - Data preparation: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Phase 4 - Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. The CRISP-DM documentation points out that typically, there are several techniques for the same data mining problem

type. Some techniques have specific requirements on the form of data and therefore, stepping back to the data preparation phase is often necessary.

Phase 5 - Evaluation: This phase of the project consists of thoroughly evaluating the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Phase 6 - Deployment: Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. According to the CRISP-DM process model, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Feedback Loops Described in the CRISP-DM Process Model

It also describes various feedback loops to emphasize how certain phases should be revisited to leverage the new information or knowledge gained in the phase succeeding them. These have also been highlighted in Figure 2-1. For instance, while Data Preparation typically precedes Modeling, there may be a need to revisit Data Preparation as a chosen Modeling technique may require data to be prepared in a certain way.

21
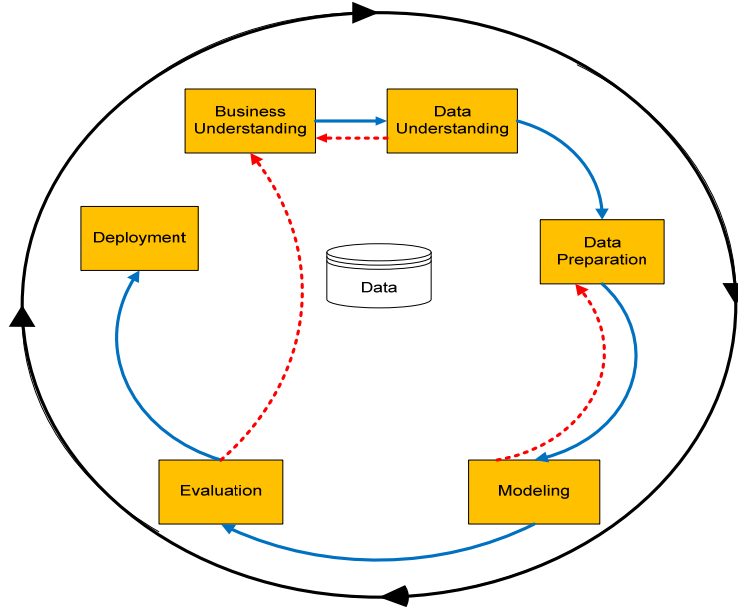
Figure 2-1 CRISP Process Model (CRISP-DM, 2003)

CRISP-DM is the most detailed of existing KDDM models. The documentation associated with CRISP-DM v 1.0 is divided in two parts. The first part provides a description of the reference model, its phases, general tasks and outputs. The second part called the user guide aims to provide detailed guidance about how to perform activities associated with each task.

The user guide portion of CRISP DM methodology (CRISP-DM 2003) aims to provide detailed advice about "how" to execute DM activities. That is, the user guide is expected to provide tools for implementing the vast number of activities suggested in the process model. However analysis of the user guide reveals that does not meet its

22

intended objective and only proposes a checklist of activities to be performed to accomplish the tasks associated with each phase. Tool support is only provided towards only two of the total twenty four tasks mentioned in the model and it appears that even these are not sufficient for efficiently executing the corresponding tasks. These are described below:

1. Tool support for task of selection of modeling techniques (modeling phase)

The CRISP-DM v1.0 documentation (CRISP-DM 2003) includes some support towards the modeling phase by providing a list of modeling techniques relevant to various types of data mining problems. However, it does not provide any support towards selection of appropriate techniques. Clearly, the list of techniques enumerated in the process model could be narrowed down further using output from previous tasks such as business objectives and data mining objectives, but that it is not considered by the process model.

2. Tool support for task of identification of divisions and manager's name and responsibilities (business understanding phase)

Analysis of the foundational business understanding phase reveals the use of just one tool - an organizational chart, to "identify divisions, manager's names and responsibilities etc". Clearly, organizations also need support for the diverse array of other activities associated with this important phase. Besides, the usefulness of

23

organizational charts, a primarily static entity, to identify organizational actors and their

interrelationships can be also be debated.

Table 2-1: Phases, Tasks and Outputs - CRISP-DM process model

Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
Determine Business Objectives - Background - Business Objectives - Business Success Criteria	**Collect Initial Data** - Initial Data Collection Report	**Select Data** - Rationale for Inclusion/ Exclusion	**Select Modeling Technique** - Modeling Technique - Modeling Assumptions	**Evaluate Results** - Assessment of Data Mining Results with respect to Business Success Criteria - Approved Models	**Plan Deployment** - Deployment Plan
Assess Situation - Inventory of resources - Requirements Assumptions and Constraints Risks and Contingencies - Terminology - Costs and Benefits	**Describe Data** - Data Description Report	**Clean Data** - Data Cleaning Report	**Generate Test Design** - Test Design	**Review Process** - Review of Process	**Plan Monitoring and Maintenance** - Monitoring and Maintenance Plan
Determine Data Mining Goals - Data Mining Goals - Data Mining Success Criteria	**Explore Data** - Data Exploration Report	**Construct Data** - Derived Attributes - Generated Records	**Build Model** - Parameter Settings Model - Model Description	**Determine Next Steps** - List of Possible Actions - Decision	**Produce Final Report** - Final report - Final Presentation
Produce Project Plan - Project Plan - Initial Assessment of Tools and Techniques	**Verify Data Quality** - Data Quality Report	**Integrate Data** ▪ Merged Data **Format Data** ▪ Reformatted data	**Assess Model** ▪ Model Assessment ▪ Revised parameter settings		**Review Project** - Experience - Documentation

- **Cios and Kurgan (2005)**

The process model proposed by Cios and Kurgan (2005) is shown in Figure 2-2.

1. Understanding the problem domain

In this step one works closely with domain experts to define the problem and determine the project goals, identify key people, and learns about current solutions to the problem. It involves learning domain-specific terminology. A description of the problem including its restrictions is done. The project goals then need to be translated into the DMKD goals, and may include initial selection of potential DM tools.

2. Understanding the data

This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the DMKD goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.
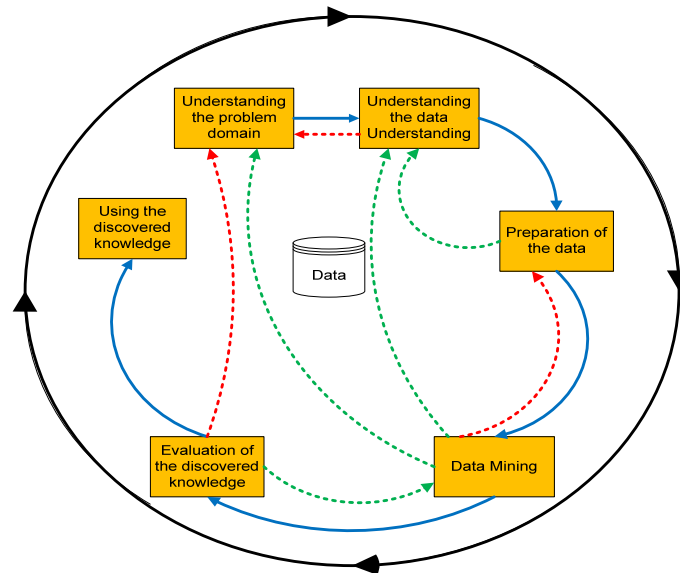
Figure 2-2: KDDM process model proposed by Cios and Kurgan (2005)

3. Preparation of the data

According to this process model, data preparation is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire project effort. In this step, decisions regarding which data will be used as input for data mining tools of step 4 are made. It may involve sampling of data, running correlation and significance tests, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say by discretization), and by summarization of data (data granularization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

4. Data mining

This is also regarded as a key step in the knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, such as rough and fuzzy sets, Bayesian methods, evolutionary computing, machine learning, neural networks, clustering, preprocessing techniques, etc. This step involves the use of several DM tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one

27

of the chosen DM tools; the generated data model is verified by using testing procedures.

5. Evaluation of the discovered knowledge

This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models (results of applying many data mining tools) are retained. The entire DMKD process may be revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.

6. Using the discovered knowledge

This step consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains. A plan to monitor the implementation of the discovered knowledge should be created, and the entire project documented. Cios and Kurgan (2005) also specify four additional feedback loops as compared to the CRISP-DM model and the situations under which the loops would get triggered.

Table 2-2: Specific Feedback Loops described in KDDM process model proposed by Cios and Kurgan (2005)

Feedback Loop	Condition Under Which Feedback Loop Should Be Triggered
2→1	From Step 2 to Step 1: execution of this loop is triggered by the need for additional domain knowledge to improve data understanding.
3→2	From Step 3 to Step 2: execution of this loop is triggered by the need for additional or more specific information about the data to guide choice of specific data preprocessing.
4→1	From Step 4 to Step 1: the loop is performed if results generated by selected DM methods are not satisfactory and modification of project's goals is required.
4→2	From Step 4 to Step 2: the most common reason is poor understanding of the data, which results in incorrect selection of DM method(s) and its subsequent failure (e.g., data was misclassified as continuous and discretized in Understanding the Data step).
4-→3	From Step 4 to Step 3: the loop is motivated by the need to improve data preparation; this is often caused by specific requirements of used DM method, which may have been unknown during Step 3.
5→1	From Step 5 to Step 1: the most common cause is invalidity of the discovered knowledge; there are several possible reasons including misunderstanding or misinterpretation of the domain, incorrect design or misunderstanding of problem restrictions, requirements, or goals. In these cases the entire KDDM process needs to be repeated.
5→4	From Step 5 to Step 4: this loop is executed when the discovered knowledge is not novel, interesting, or useful; the least expensive solution is to choose a different DM tool and repeat the DM step

2.2 Limitations of Existing KDDM Process Models

In this section, we use Dowson's process domain architecture to identify issues with respect to the process model domain (i.e. design of the model) and the process enactment domain (tools and features needed to support enactment of processes recommended by a model) of the KDDM process models described earlier. Our review of existing Knowledge Discovery and Data Mining Process Models (Fayyad, Piatetsky-

Shapiro et al. 1996a; Berry and Linoff 1997; Anand and Buchner 1998; Cabena, Hadjinian et al. 1998; Cios, Teresinska et al. 2000; Han and Kamber 2001; CRISP-DM 2003; Cios and Kurgan 2005) reveals some common serious deficiencies. We believe that the deficiencies apply to all existing KDDM models. Any exceptions are duly noted and discussed in the section below.

1. Description of the KDDM Process in a Checklist Manner

While nearly all KDDM process models acknowledge the complexity of the KDDM process, they still describe the complicated KDDM process in terms of a list of steps or tasks (Brachman and Anand 1996; Fayyad, Piatetsky-Shapiro et al. 1996b; Berry and Linoff 1997; Anand and Buchner 1998; Cabena, Hadjinian et al. 1998; Cios and Kurgan 2005). The number of steps suggested by each model may vary but the range is restricted to between four steps (Berry and Linoff 1997) and nine steps (Fayyad, Piatetsky-Shapiro et al. 1996b). Table 2-3 lists the steps specified by these two models for comparison purposes.

Analysis of the list of steps reveals that while these steps are valid, these models make broad assumptions about the users involved in carrying out the KDDM project. The steps are, at best, a broad guideline, a checklist that could be used by users to remind themselves of important stages of the KDDM process.

Table 2-3 Description of the complicated KDDM process in a small number of steps

KDDM process Model; Number of steps/tasks	List of steps/tasks specified by the KDDM model
Berry and Linoff (1997); 4 steps	1. Identifying the Problem 2. Analyzing the Problem 3. Taking Action 4. Measuring the Outcome.
Fayyad et al. (1996b) 9 steps	1. Developing and understanding the application domain 2. Creating a target data set 3. Cleaning and preprocessing data 4. Data reduction and projection 5. Choosing the data mining task 6. Choosing the data mining algorithm 7. Data mining 8. Interpreting mined patterns 9. Consolidating discovered knowledge.

CRISP-DM (2003) the leading KDDM model used in the industry (KDNuggets 2007) is different from the KDDM process models described above in that it divides the lifecycle of a KDDM project over six different phases and specifies the tasks (and their desired outputs) and activities needed to execute these tasks. However, while CRISP-DM (2003) discusses the KDDM process in detail, it also prescribes the various tasks in a similar 'checklist' manner. In fact, the checklist approach is even more problematic in case of CRISP-DM due to the large number of activities prescribed by the process model. Table 2-4 presents the list of activities (for each task) prescribed by the CRISP-DM model. It can be seen that the model recommends executing a total of 288 activities, which when presented in a checklist approach may seem very cost prohibitive to implement.

Table 2-4 Number of activities in each phase of CRISP-DM (2003)

Phase (Number of tasks)	Number of Activities
Business Understanding (4)	67
Data Understanding (4)	47
Data Preparation (3)	27
Modeling (4)	34
Evaluation (3)	25
Deployment (4)	28
Total number of activities	**228**

2. Fragmented View of the KDDM Process

The existing KDDM process models present a fragmented view of the KDDM process. In other words, the process models do not capture or highlight the important dependencies existent in a typical KDDM process. By dependencies we mean the interrelationships between the various steps, or between the various phases and tasks (of the same and different phases) of a KDDM project. For instance, process models that structure a KDDM process in form of a sequence of steps, do not clearly discuss how the steps are related to each other.

That a particular step is recommended to be executed at the beginning and another one towards the end signifies that the step performed at the end may be dependent on the execution of the one performed at the beginning of the project; specifically, it may utilize the output of the particular step directly or indirectly (using the output of a step which in turn uses the output of the step at the beginning). However, these dependencies are not discussed in the process models. We discuss the same issue with respect to CRISP-DM which presents a KDDM project in terms of a number of phases and tasks (instead of steps like the model discussed above) before proceeding to discuss the serious repercussions of not identifying the dependencies in a KDDM process.

CRISP-DM structures a KDDM process in form of phases and their corresponding tasks. The CRISP-DM process model is shown in Figure 1-2. The

33

various phases described by the model include business understanding, data understanding, data preparation, modeling, evaluation and deployment.
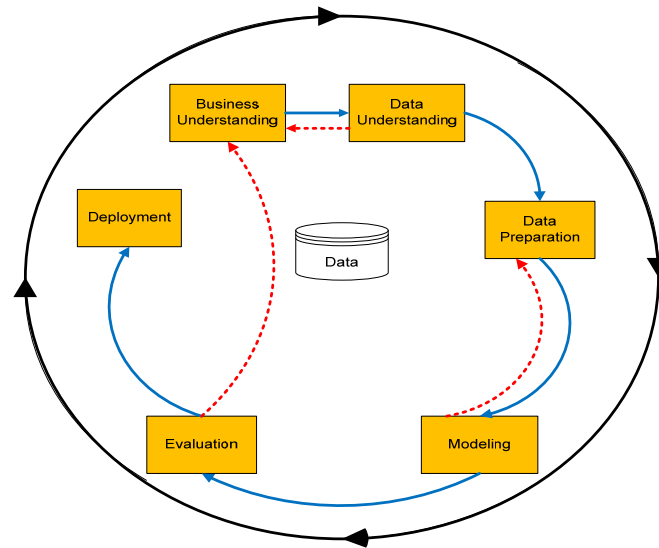
Figure 2-3 CRISP Process Model (CRISP-DM, 2003)

The dependency which is most obvious from this model is the phase-phase dependency resulting from the ordering of phases proposed by the model. That the model recommends executing the business understanding phase ahead of the data understanding phase suggests that data understanding phase must be utilizing the output of the business understanding phase. These dependencies are critical as they cannot be reversed without leading to detrimental effects or even inability of executing a particular phase. Further, it is important to consider that a phase really comprises of various tasks. Therefore, the output of a phase really comprises of the output of the diverse array of tasks that lie within it.

34

Clearly, a task-level view of a process model should explicate and highlight these dependencies. These dependencies are not obvious from the phase-level view of the process model as shown in Figure 2-3. It is relevant to note that task-task dependencies exist both due to interrelationships between the tasks of the same and phase as well as the tasks of the different phases of the model. Therefore the task level view of the process model should explicate both of these; in other words it should represent a complete view of the KDDM process.
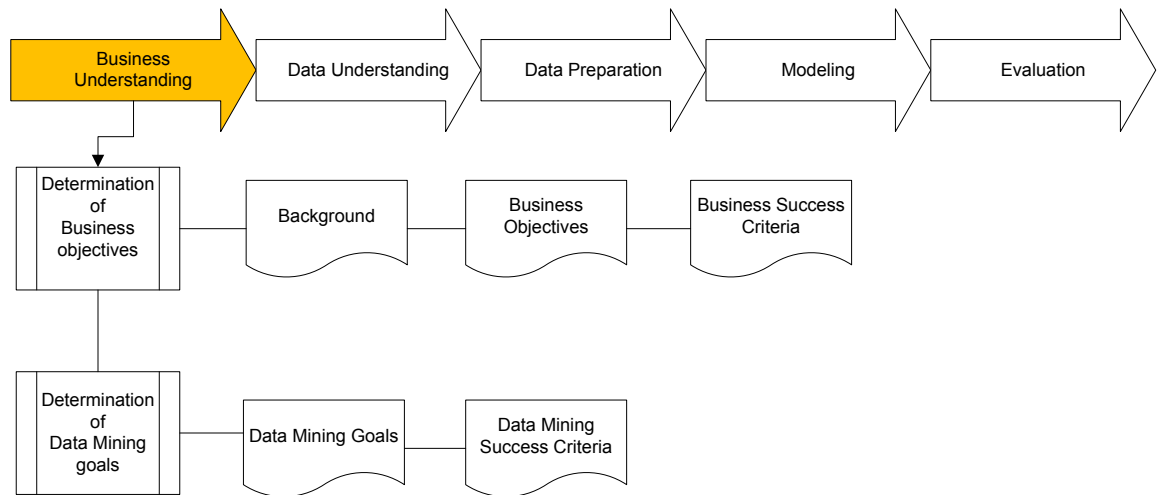
Figure 2-4 CRISP-DM - partial view of Business Understanding Phase

In Figure 2-4 and Figure 2-5 we present the task-level view of the CRISP-DM process model for a subset of tasks belonging to business understanding and data understanding phases. For the purpose of discussion, we only present a partial view of each phase in these figures. It can be seen that neither of the two types of dependencies highlighted above are obvious from these figures. The dependencies between the tasks

35

of different phases are not captured at all as each phase is described in standalone manner. The dependencies between the tasks of the same phase are also not obvious from these figures.
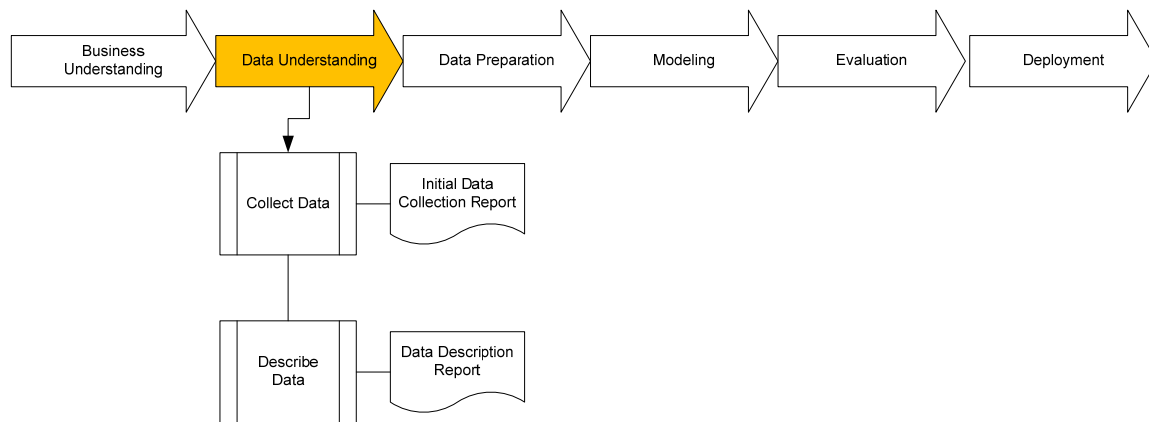
Figure 2-5 CRISP-DM - partial view of Data Understanding Phase

CRISP-DM presents the remaining four phases in a similar manner and does not present an integrated process model that shows all the dependencies. It can be argued that this is only a presentation issue as the documentation also describes the various tasks in detail. Careful analysis of the documentation reveals that while some dependencies can be implied from the (brief) description of tasks such that business objective can be translated into a data mining objective, the model does not make an effort at explicating the large number of dependencies that exist in the KDDM process or presenting them in form of an integrated model.

The repercussion of not explicating various dependencies existent in the context of a KDDM project could lead to inefficient implementation of projects. For instance, an organization may embark on a particular task and realize that it cannot be completed; this could translate into unnecessary costs and overhead worst still the task may be executed disregarding the output from a relevant task that should have been carried out prior to this task's execution. An example of this situation could be selection of a modeling algorithm without clearly first setting up the business objective. This is an important task-task dependency which if neglected can lead to the project take a completely different direction than intended.

3. Fragmentation: a Hindrance to Building an Integrated Process Model and "Semi-Automating" Well Understood tasks

Identification of task-task dependencies (between tasks of the same phase and different phases) is the first step towards building an integrated process model, the importance of which has been acknowledged in the literature (Brachman and Anand 1996; Kurgan and Musilek 2006). The integrated process model can then be used for enabling the semi-automation (Kurgan and Musilek 2006) or automation of some of the well understood tasks of the process. There is a general understanding that it is only the task of implementation of data mining methods (modeling phase) which can be automated (Berry and Linoff 2000). Recently however, researchers have also attempted to automate certain other tasks such as selection of appropriate modeling techniques or algorithms (Bernstein, Hill et al. 2005), which were once performed manually by the

human user. Clearly, the same opportunity lies in the other phases of the knowledge discovery process where certain tasks could be semi-automated if not completely automated to increase the overall efficiency and effectiveness of the knowledge discovery process.

Continuing with the example presented in the above section, we argue that the identification of dependency between two tasks such as a business and data mining objective should be leveraged to drive the execution of the latter task. For instance, effort should be made to examine whether output of business objects can be used to semi-automate tasks, such as determination of data mining objectives, that utilize it as its input (Figure 2-6).
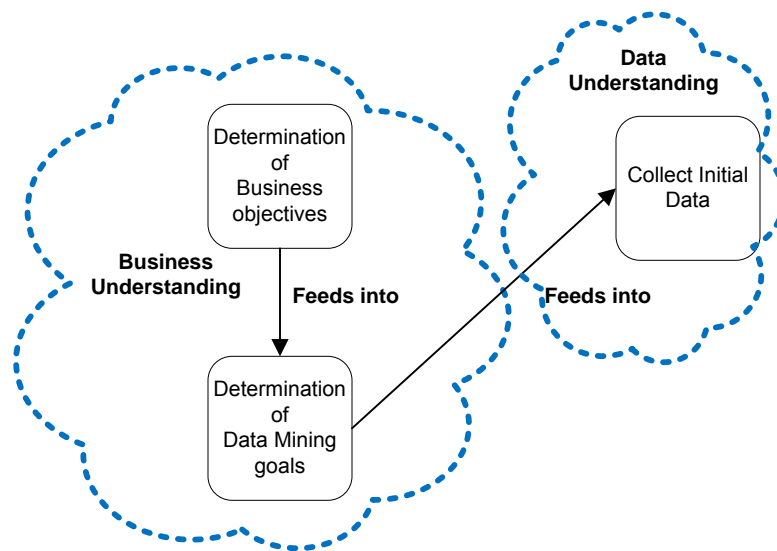
Figure 2-6: Explicating of dependencies as a first step towards enabling semi-automation

4. Lack of support for the end-to-end KDDM process

Existing KDDM models do not provide enough support towards "how" to implement the long list of tasks and activities suggested by them (Charest, Delisle et al. 2006). Given that a KDDM process requires a user to make numerous decisions (Fayyad, Piatetsky-Shapiro et al. 1996b), it is only necessary that the process models be complemented by support in form of appropriate tools and techniques for carrying out the various tasks. Charest et al. (2006) note that existing process models 'only provide general directives, however what a non-specialist really needs are explanations, heuristics and recommendations on how to effectively carry out the particular steps of the methodology'.

Lack of decision support towards tasks can result in non-execution of these tasks during the knowledge discovery process. Given the discussion of dependencies in the earlier section, we know that each task helps drive other tasks (who use its output as its input), and it is therefore non-execution of a task can translate into incapability to implement or ineffective implementation of succeeding tasks in the model.

Overtly, it may appear that this issue is less problematic in case of the data mining or modeling phase that has benefitted from the rapid advancement in development of plethora of data mining techniques. However, even this phase requires careful selection of the techniques is required if the objectives of the project are to be accomplished (Pyle 2003). Simoudis et al. (1996) note that a single data mining technique is often insufficient for extracting knowledge from a data set. They

recommend that in such a situation several techniques must be employed cooperatively to support a single data mining application. Clearly, support is needed to aid the user in selecting these techniques and the order in which they should be used if the KDDM project is to be effectively executed.

5. Conspicuous Lack of Support Towards Execution of Business Understanding Phase - the Foundational Phase of the KDDM Process

All process models recommend launching a KDDM process with gathering an understanding of the business domain. This phase includes making determinations about business and data mining objects, assessing resources and generating a project plan for the remainder of the project. Clearly, the importance of this phase cannot be overemphasized. However, different researchers estimate that very little time is actually devoted to the execution of this phase (see graph below).

Relative effort (%)

Legend:
- Estimates by Cabina et al. (1998)
- Estimates by Shearer (2000)
- Estimates by Cios and Kurgan (2005)

KDDM Steps: Understanding of Domain, Understanding of Data, Data Preparation, DM, Evaluation of Results, Deployment of Results
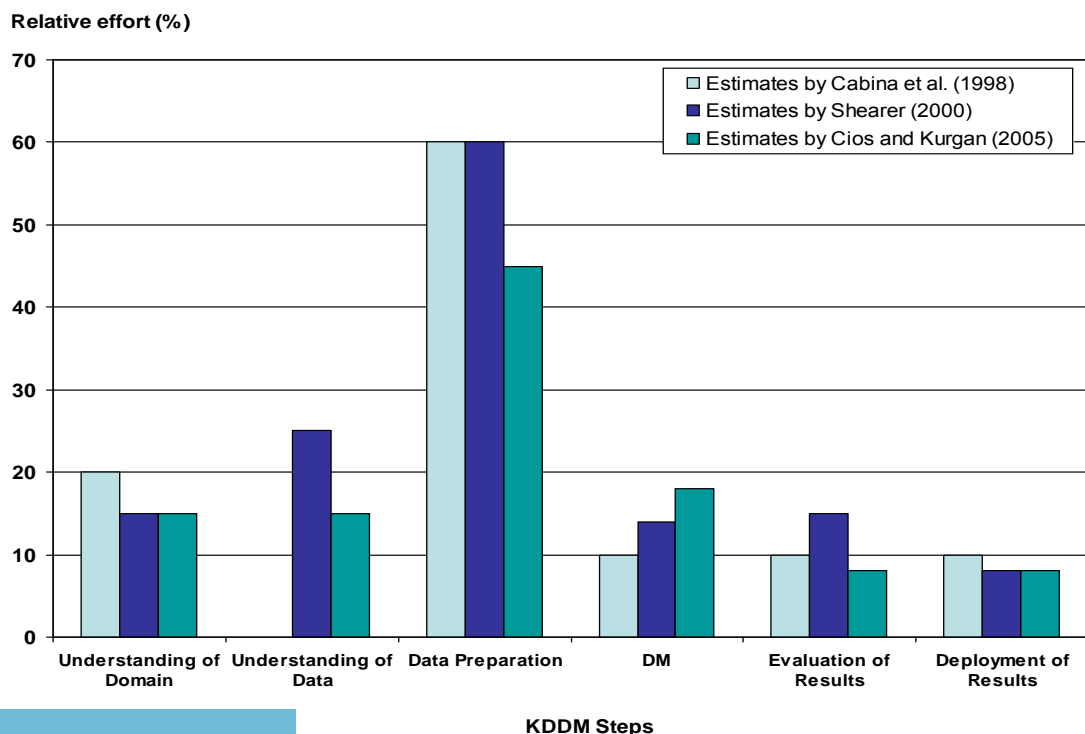
However, our review of published data mining case studies reveals that that the business understanding phase of KDDM projects is often implemented in an ad hoc manner. Hardly any published data mining case studies actually provide a detailed description of how this phase was formally implemented. We believe that the reason for such an unstructured approach is because of the general lack of support towards how the tasks of this phase can be implemented.

This issue has been highlighted and somewhat addressed by Pyle (2003) who describes how real world business problems (to be addressed through data mining) can be modeled. While the author has not based his approach on any particular DM methodology, he discusses various tools to carry out many (though not all) of the activities prescribed under the BU phase of the CRISP-DM methodology. However, these are only presented in a linear fashion, with the description of each activity followed by a brief description of a proposed tool. The overall framework which consists of nested sequences of action boxes, discovery boxes, technique boxes and example boxes is complicated to navigate, and may appear to be cumbersome or even cost prohibitive to actors involved in carrying out the critical business understanding phase.

 The description of the user guide portion of CRISP DM methodology (CRISP-DM 2003) also purports to provide detailed advice about "how" to execute KDDM activities outlined in the model. The only applicable tool mentioned in this phase is the use of an organizational chart, to "identify divisions, manager's names and

41

responsibilities etc". Clearly, organizations also need support for the diverse array of other activities associated with this important phase. Besides, the usefulness of organizational charts, a primarily static entity, to identify organizational actors and their interrelationships can be also be debated.

Formally implementing the Business Understanding phase is just as important as implementing the Modeling phase or any other phase of the data mining project (Sharma and Osei-Bryson 2008). Perhaps, the Business Understanding Phase is even somewhat more important than other phases given that a number of decisions about other phases, such as the Modeling as well as other phases (such as data preparation, data understanding, evaluation etc) are made, or ideally *should* be made, during the BU phase (Figure 2-7)
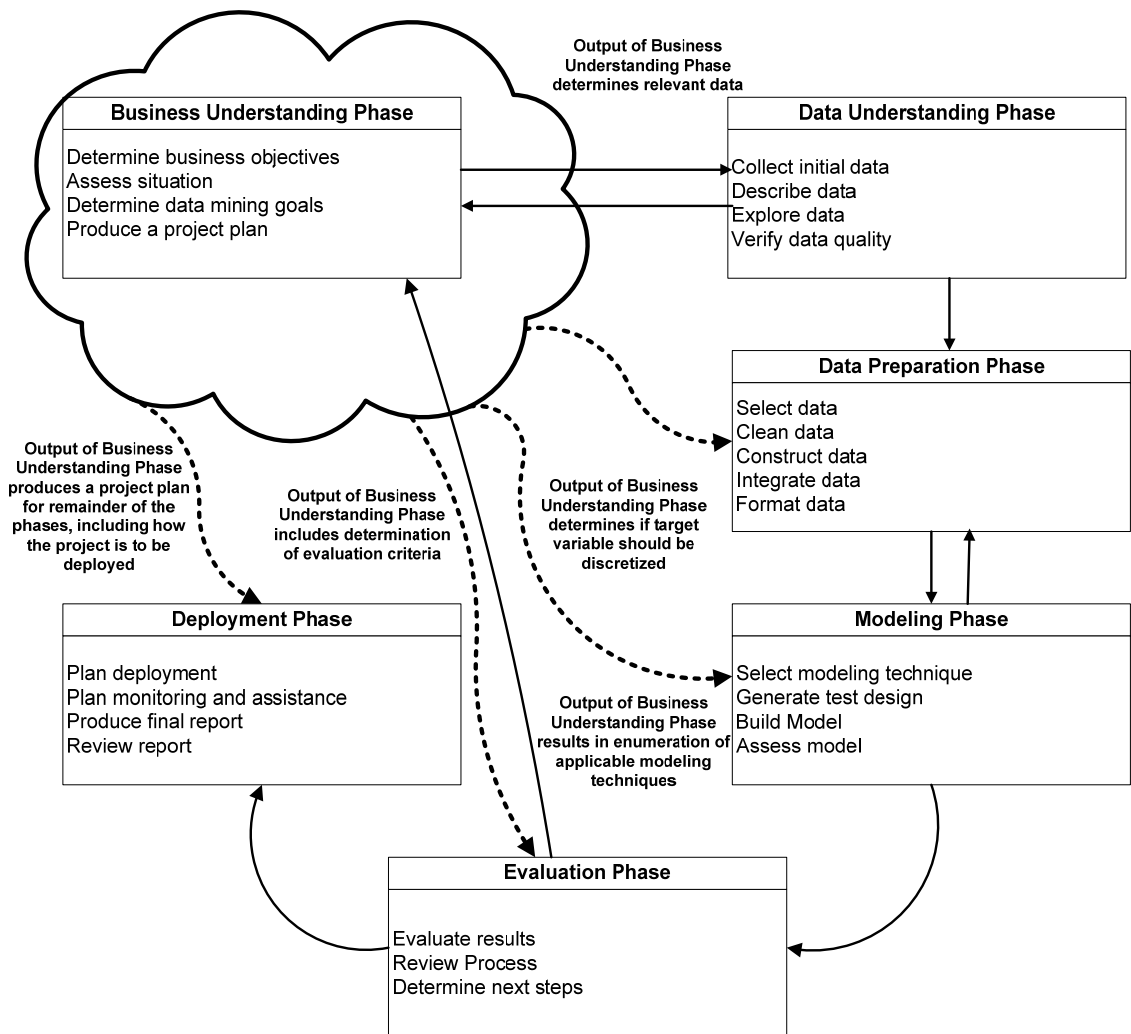
Figure 2-7: Business Understanding phase pervades all other phases of the KDDM process

(Adopted from Sharma and Osei-Bryson, 2008)

Not making appropriate decisions during the BU phase seems to lead to two problems. First, it creates inefficiencies as these decisions have to be dealt with in later phases taking away the time and resources that were allocated to accomplish the tasks

associated with that phase. The second problem is even more detrimental as not making certain decisions during the BU phase can lead to the DM project taking a completely different direction than what was intended. The second problem originates from issues of dependencies (existing between the various phases and tasks of a data mining project) and has been discussed earlier. These dependencies need to be clearly identified and effectively managed in order to formally implement this phase and in turn the entire KDDM project.

2.3 Design Requirements for the Integrated Knowledge Discovery and Data Mining (IKDDM) Model

Summarizing key issues identified in the previous section we see that existing KDDM models suffer from the following limitations:

- Description of the KDDM Process in a Checklist Manner

- Fragmented View of the KDDM Process

- Emphasis on feedback loops prior to completely understanding the primary sequencing of phases and tasks in a KDDM process

- Fragmented view acts as a hindrance to building an integrated process model and "semi-automating" tasks

- Lack of support for the end-to-end KDDM process

- Conspicuous lack of support towards execution of Business Understanding phase - the foundational phase of a KDDM process

44

As stated earlier, the research objective of this dissertation is to design an improved Knowledge Discovery and Data Mining Process Model. We name this model the Integrated Knowledge Discovery and Data Mining (IKDDM) process model. The fulfillment of the research objective requires the design of a solution that addresses the limitations in existing KDDM models identified above. Design is a goal oriented activity (Simon 1996). The requirements that the proposed model must address are described in Table 2-5 below. The table also shows whether the particular limitation is an aspect of the process model domain or the process enactment domain.

Table 2-5: Design Requirements for an improved KDDM model

Issues Identified with existing KDDM Process Models (As-is situation)	Aspect of	Design Requirements for the IKDDM model (To-be situation)
Description of the KDDM Process in a Checklist Manner	Process Model Domain	Present a user-oriented coherent description of the KDDM process
Fragmented View of the KDDM Process	Process Model Domain	Develop an integrated view of the KDDM process by explicating the various phase-phase and task-task dependencies
Emphasis on feedback loops prior to completely understanding the primary sequencing of phases and tasks in a KDDM process	Process Model Domain	Explicate sequencing of the various phases and their tasks before identifying feedback loops and establishing conditions under which the loops would get triggered
Fragmented view acts as a hindrance to building an integrated process model and "semi-automating" tasks	Process Model Domain	Leverage the dependencies explicated in the integrated process model to drive semi-automation of tasks, wherever possible
Lack of support for the end-to-end KDDM process	Process Enactment Domain	Prescribe approaches for offering decision support towards all tasks in all phases, described in the integrated KDDM model
Visible lack of support towards execution of tasks of the Business Understanding phase - the foundational phase of a KDDM process	Process Enactment Domain	Provide support for tasks of this foundational phase and use them as a basis for developing the integrated model

Data Mining Projects as an Instantiation of the KDDM Process

The KDDM process described in the above sections is generally referred to as a data mining project in Information Systems Research (Berry and Linoff 1997; Pyle 2003). Truly speaking, a data mining project is an instantiation of the knowledge

discovery process. Due to the wider usage of the term data mining project as opposed to knowledge discovery in the Information Systems Community, the dissertation will also make use of the former term in discussion of various concepts. However, it must be emphasized that the research objective and the solution proposed in this dissertation applies to the generic KDDM process and is not restricted to data mining projects, which are only an instantiation of this process.

2.4   Significance of the IKDDM process model

This dissertation addresses important research objectives that are relevant to both academicians and practitioners of Knowledge Discovery and Data Mining research. The KDDM process as implemented by these communities makes use of KDDM process models. These models play an important role in discovering of knowledge, a critical challenge facing today's business organizations that are awash in mountains of data (Han and Kamber 2006; Kurgan and Musilek 2006; Davenport and Harris 2007). A robust, understandable and comprehensive process model is required to adequately address this critical challenge.

The dissertation draws attention towards some common serious limitations (for example, checklist description, fragmented view of the KDDM process, lack of support for execution of the end-to-end KDDM process) of existing KDDM process models (Fayyad, Piatetsky-Shapiro et al. 1996a; Berry and Linoff 1997; Anand and Buchner 1998; Cabena, Hadjinian et al. 1998; Cios, Teresinska et al. 2000; Han and Kamber 2001; CRISP-DM 2003) based on a detailed survey of the relevant literature.

The integrated KDDM process model described in this dissertation extends the existing KDDM process models by addressing their limitations including checklist-type description of tasks and activities and neglect of critical dependencies existing between various tasks (of the same phase and different phases) of the knowledge discovery process. The integrated KDDM process model streamlines the list of tasks relevant in each phase and captures dependencies in its design. The importance of integration of KDDM process models has also been highlighted in the literature (Uthurusamy 1996; Han and Cercone 2000). Kurgan and Musilek (2006) who conducted a detailed review of existing KDDM models, acknowledge that the future of KDDM process models is in achieving integration of the whole process.

The dependencies highlighted in the integrated model can be used for semi-automating the execution of relevant tasks and can thereby result in more efficient and effective implementation of the knowledge discovery process. Further, the dissertation also proposes techniques that can be used for providing decision support in form of appropriate tools and techniques for the various tasks (excluding tasks belonging to deployment phase) belonging to the integrated KDDM process model.

The identification and description of relevant techniques can serve to ensure that all the tasks of the process model are executed and that no task is inadequately executed (or not executed) due to lack of support towards its implementation. This is also likely to result in improving the efficiency and effectiveness of execution of the KDDM process.

3 LITERATURE REVIEW: CONCEPTS RELEVANT TO THE KDDM PROCESS

In this chapter we review the literature with a goal of studying concepts and techniques that are relevant to the main components of the KDDM process. An important simultaneous consideration is to understand, "how" each one of these component of the process can be executed. This stems directly from out observation that there exist deficiencies in the process enactment domain of existing process models, which make it difficult to implement the processes recommended by the model.

Based on the discussion of KDDM process in the earlier chapters, we identify some main components of the KDDM process. This is followed by a discussion of concepts and/or techniques relevant to each component. The work described in this chapter is intended to build a foundation towards populating the enactment domain of the IKDDM model being designed by this research.

3.1 Main Components of the KDDM process

The main components of the KDDM process are described below:

1. Gathering background information about the problem to be addressed through data mining

2. Formulating (business and data mining) objectives

50

3. Formulating success criteria or evaluation criteria for the business and data mining objectives

4. Identifying relevant individuals (key stakeholders, project participants)

5. Understanding data and relationships between variables

6. Integrating data in preparation for modeling

7. Understanding data mining problem type(s) to be addressed through modeling

8. Analysis of characteristics of various modeling techniques

9. Evaluating output of modeling techniques to determine whether or not it meets requirements

3.2 Discussion of Relevant Concepts

In the section below we present concepts and techniques relevant to the main components of the KDDM process identified in the above section.

1. Gathering background information about the problem to be addressed through data mining

Before formally embarking on the KDDM project, background information about the problem to be addressed may need to be collected. This is a type of intelligence gathering and therefore intelligence gathering techniques may be relevant towards the execution of this component. Nutt (2007) present an approach for gathering intelligence during the decision making process (Table 3-1).

51

Table 3-1: Intelligence Gathering Approach for Decision Making - Proposed by Nutt (2007)

Variables	Data collection	Approaches used
Signal coding	1. Informants answered two open-ended questions: "What first captured your attention" and "Why was this important?" 2. Specifics about performance levels and expectations (e.g., norms or performance benchmarks) were inferred from what was said	Signal coding used "performance gaps" that were: 1. *Quantitative*—both norms and performance were determined factually. 2. *Qualitative*—both norms and performance were noted, but either the norm or the performance was not factually determined 3. *Impressionistic*—no norms or performance indicators were cited. Signals were described as an arena of action
Signal interpretation	1. Decision makers described the motivation to act. Determine whether this was performance or action driven 2. Questionnaire data rating the decision's importance and urgency, on a 1–5 scale (1 = low, 5 = very high) by the two secondary informants	Interpretations: 1. *Need*—performance driven, calling for better results 2. *Opportunity*—action driven, calling for a particular action 3. *Defined threat*—opportunity with urgency and importance both rated very high 4. *Undefined threat*—need with both urgency and importance rated very high
Search behavior evoked	1. Decision makers were asked to specify the steps undertaken to uncover alternatives that were considered before a course of action was selected	Search approaches uncovered 1. *Negotiated*—stakeholders meet to uncover options 2. *Rational*—outcome target set and formal protocol followed to find alternatives that can produce expected results 3. *Problem solving*—a variation of the rational approach in which the target is stated as a problem to be overcome 4. *Opportunity*—an idea noted in the signal prompting action was adopted 5. *Emergent opportunity*—the adopted idea emerged before a search could be completed 6. *Redevelopment*—the idea found in the signal was abandoned and a search undertaken to find a replacement

2. *Formulating (business and data mining) objectives*

Determination of business and data mining (technical) objectives is an important component of the KDDM process. This component represents the starting point of the KDDM process. Given this fact, it is easy to understand that improper formulation of objectives can lead to jeopardizing the entire KDDM project. Data Mining literature and process models recognize the significance of this component, but do not provide any approaches for implementing it. We identify some approaches proposed in the literature that can be used to do this. First, we discuss value focused thinking or VFT proposed by Keeney (1996) as means of formulating objectives and goals. Second, we discuss SMART approach for formulating objectives that is often recommended in the practitioner literature.

- Value Focused Thinking

Value focused thinking (VFT) considers the role of values in decision making and can be differentiated from conventional decision making which focuses on enumeration of alternatives. The concept of value focused thinking was first proposed by Keeney (1992) who argues that conventional decision making approaches are reactive in nature as they emphasize identification of alternatives ahead of articulation of values that are important to the particular decision situation.

According to Keeney it is important to make the values explicit and use them to guide the decision making process. Keeney (1996) offers a methodology for creating and structuring values in form of objectives and using the objectives to guide decision

53

making. Keeney's work has helped to address an important gap in research namely the lack of support towards formulation of objectives to characterize a decision situation. Keeney (1996) notes that while all experts on decision making agree that it is crucial to list your objectives, they are not specific about how to do it or how to use the objectives to guide your thinking. Keeney's work on value focused thinking provides explicit guidance towards formulation of objectives, an indispensable task in any decision making situation.

Value focused thinking includes three different types of objectives: fundamental objectives, means objectives and strategic objectives. Fundamental objectives concern the ends that decision makers value in a particular decision context whereas means objectives are the methods to achieve the ends. Strategic objectives provide common guidance for more detailed fundamental objectives.

Thinking about these different types of objectives can lead to enumeration of alternatives relevant to a decision situation. Keeney (1996) also contends that there is value in thinking about certain decision situations as opportunities rather than problems. He states that a decision opportunity can help alleviate problems or allow avoiding of future problems. Value focused thinking has found applications across a wide variety of decisions belonging to diverse domains including environmental engineering (Hassan 2004), military operations (Keeter and Parnell 2005), homeland security (Pruitt 2003), tourism management (Kajanus, Kangas et al. 2004), and systems engineering (Boylan, Tollefson et al. 2006) to name a few.

- SMART approach

The SMART acronym proposed by Doran (1981) is commonly recommended for setting objectives. The approach underlying SMART suggests that objectives should be specific, measurable, achievable, relevant and timely (Table 3-2).

Table 3-2: SMART approach for setting up Objectives

Criterion	Description
Specific	The objective must lead to an observable action, behavior or achievement
Measurable	The objective must be measurable through
Achievable	The business objective must be achievable within the constraints of the available resources, knowledge and time
Relevant	The objective must be relevant to the organizational goals
Time-Bound	There should be clear deadlines for the achievement of the objective

- Peter F Drucker's work on Management by Objectives

Drucker's (1954) work also offers guidance towards the process of formulating business objectives in data mining projects. Acknowledging the popularity of profit maximization as a business objective, Drucker cautions that emphasizing only profits as business objectives is likely to misdirect managers and result in poor decisions. He suggests setting objectives (in terms of performance and results) in eight different areas

(Table 3-3). These include (1) market standing, (2) innovation, (3) productivity, (4) physical and financial resources, (5) profitability, (6) manager performance and development, (7) worker performance and attitude and (8) public responsibility. Of these while the first five objectives are tangible, the remaining three are intangibles.

Table 3-3: Categories of Objectives proposed by Drucker (1954)

Categories of Objectives	Examples of Types of Objectives
Market Standing	What is the firm's market, who is the customer, where he is, what he buys, what he considers value, what his unsatisfied needs are
Innovation	New products or services needed to attain marketing objectives, new products needed because existing ones may become obsolete, new processes and improvement in old processes
Productivity	What is the best product mix, how much do products yield versus what is their utilization of raw materials
Physical and Financial Resources	Investment management, how much capital is needed and where will it come from
Profitability	The return on investment, break even point analysis, net profit

3. *Formulating success criteria or evaluation criteria for the business and data mining objectives*

Evaluation criteria or success criteria provide a means of measuring whether or not the business and data mining objectives were achieved. Our review of the literature reveals that the Goal Question Metric Approach (Basilli and Weiss, 1984) can be used to implement this important component of the KDDM process. It is described below.

▪ Goal Question Metric Approach

Goal Question metric (GQM) approach (Basili and Weiss 1984; Basili and Rombach 1988) was originally proposed as a mechanism for defining and evaluating a set of operational goals using measurement. While the technique was originally developed for identification of defects in software projects, its scope could be extended to serve other purposes such as corporate goal setting and evaluation. In this sense the technique can be used formally implementing the creation of organizational and project goals.

The GQM approach consists of a top-down hierarchical structure consisting of three components: Goals, questions and metrics (Figure 3-1). A goal specifies the purpose of measurement, object to be measured, issue to be measured and view point from which the measure is taken. The goal can be refined into a set of questions that characterize the goal in a quantifiable way. Finally each question can be refined into a

set of quantitative and/or qualitative metrics. These metrics can be regarded as the evaluation criteria or success criteria for the stated objectives.
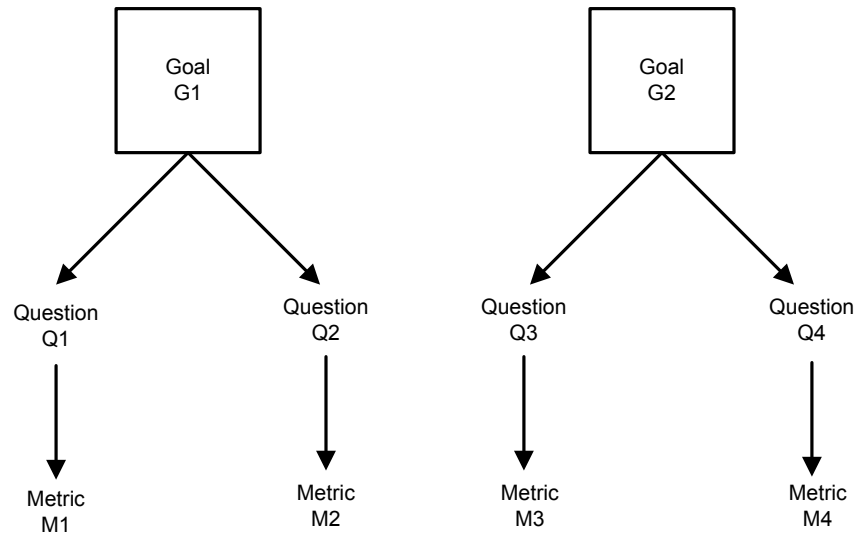
Figure 3-1: Goal Question Metric Approach proposed by Basili and Weiss (1984)

4. *Identifying relevant individuals (key stakeholders, project participants)*

Human users play an important role in the KDDM process and therefore efficient identification of relevant individuals is crucial to the implementation of KDDM projects. While an organization may make use of a conventional technique such as an organizational chart to identify relevant individuals, they are likely to be

limited by the functionalities of such a tool. For instance it may only be possible to search for individuals by their title and that too through browsing the hierarchy of the chart. Based on our review, it appears that an organization ontology can serve as a much more effective means of identifying relevant individuals. Below we explain this technique in more detail.

- Organization Ontology

Let us start be considering the definition of ontology itself. One of the most cited definition of ontology is the one provided by Gruber (1993). He defines ontology as "an explicit specification of a conceptualization". In essence, an ontology is the formal specification of concepts belonging to a certain domain, and their interrelationships. An 'organization ontology' models an organization in form of an information system (Fox, Barbuceanu et al. 1998). By "formalizing" the presence and relationships between various concepts and entities, it is able to facilitate their fast and easy retrieval.

The organization ontology proposed by Fox, Barbuceanu et al. (1998) consists of the following classes: *Organization, Organization Goal, Sub Goal, Division, Sub Division, Organization Agent, Team, Communication Link, Role, Skill Authority, Activity, Resource and Constraint*. Upon navigating their organization ontology, we find that an organization consists of divisions, and divisions consist of

59

sub-divisions. Organizational actors or agents are members of division(s) and also part of teams set up to pursue specific tasks. In contrast to divisions, teams are temporary in nature and set up when needed. Note that the concept of a team is especially important in the context of DM projects, where a variety of organizational actors come together to implement a DM project. Organizational agents play one or more roles within the organization. Each role is associated with one or more sub-goal(s) which are decomposition of the organizational goals. Each role requires certain skills and is allocated proper authority at the level that the role can achieve its goals.

5. *Understanding data and relationships between variables*

Prior to analyzing data through modeling algorithms, it must first be understood. This process can be accomplished by studying the metadata behind the data and by analyzing the data through visualization techniques. Both of these concepts are discussed below.

- Metadata

Metadata is data about data and describes how information is structured within the data warehouse. If new data is loaded old is archived or current data is moved within the data warehouse, metadata needs to be generated or updated to keep track of where that data resides. One important component of metadata is the variable type. The

information in form of the variable type is used to make several important decisions in a KDDM process. Table 3-4 summarizes the variable types on the basis of their main characteristics, amount of information supplied and whether or not the variable type has an order.

Table 3-4: Understanding data (studying data types)

Type of Variable	Characteristics	Amount of Information	Order
Nominal	Just names things	Least	No inherent order
Categorical	Names groups of things, not individual entities	Very little	No particular order
Ordinal	Gives order to the categories	Contains much more information than Nominal or Categorical	Meaningful order
Interval	- Includes order and differences in size - Measured using numbers	More information than Nominal, Categorical or Ordinal	Meaningful order

Ratio	- Two types:
	   a. Scale must be named
	   b. Scale need not be named
	- Knowledge of the unit of measurement is required
	- Quantitative
		Most information	Meaningful order

- Visualization using OLAP and MOLAP

In 1993, E. F. Codd, the acknowledged founder of relational databases, introduced the term *online analytical processing* (OLAP). Codd et al. (1993) developed a set of twelve rules for the development and use of multidimensional

61

databases intended to assist decision makers within an organization in freely manipulating their enterprise data models across many simultaneous dimensions. Codd's 12 Rules for OLAP are summarized below:

1. Multidimensional view

2. Transparent to the user

3. Accessible

4. Consistent reporting

5. Client-server architecture

6. Generic dimensionality

7. Dynamic sparse matrix handling

8. Multi-user support

9. Cross-dimensional operations

10. Intuitive data manipulation

11. Flexible reporting

12. Unlimited levels of dimension and aggregation

Multidimensional OLAP or MOLAP cube can be thought of as a common spreadsheet with two extensions: (1) support for multiple dimensions, and (2) support for multiple concurrent users. In contrast, relational OLAP or ROLAP contains both detailed and summarized data, thus allowing for "drill down" techniques to be applied to the data sets.

6. Integrating data (from multiple sources) in preparation for modeling

A typical KDDM process requires collection and integration of data from multiple data sources of different types. Two relevant approaches towards integrating data are presented below.

- Data Warehousing

As discussed in Chapter 1, today companies are trying to gain a competitive edge through the proactive use of information that they have been collecting and storing in their operational systems. These systems were never designed to support multidimensional analysis. Data warehouse have been evolved to support these new informational needs.

A data warehouse (DW) is a collection of integrated, subject-oriented databases designed to support the DSS (decision support) function, where each unit of data is non-volatile and relevant to some moment in time (Inmon 1992a). The Data Warehouse consists of operational data stores and data marts. The operational data store (ODS) is the most common component of the DW environment. Its primary day-to-day function is to store the data for a single, specific set of operational applications. The data mart is often viewed as a way to gain entry into the realm of data warehouses and to make all mistakes on a smaller scale.

Marakas (2003) describes the following four main characteristics of a Data Warehouse: Subject oriented, Data integrated, Time variant, and Nonvolatile. These are discussed below.

A. Subject Orientation: The first feature of the DW is its orientation toward the major subjects of the organization, which clearly contrasts with the more functional orientation of the various applications associated with the firms' legacy systems. The operational world of the organization is typically designed around processes and functions such as inventory or human resources, each of which exhibit specific data needs with most of the data elements local to that process or function. The DW, on the other hand, contains data primarily oriented to decision making and, as such, is organized more around the major subject areas relevant to the firm, such as customers or vendors.

B. Data Integrated: According to Inmon (1992b) the essence of the DW environment is that the data contained within the boundaries of the warehouse are integrated. This integration manifests itself through consistency in naming convention and measurement attributes, accuracy, and common aggregation.

C. Time Variant: The data are assumed to be accurate at the moment they were loaded into the DW. In this regard, data within a data warehouse are said to be time variant.

D. Non-volatility: Typical activities of inserts, deletes, and changes performed regularly in an operational application environment are completely non existent in a DW environment. Only two data operations are ever performed in the data warehouse: data loading and data access.

- ETL (Extract, Transform and Load)

Typical Project flow within a Data Warehouse consists of the following types of processes (Anahory and Murray 1997)

- Extract and load the data
- Clean and transform data in a form that can cope with large data volumes and provide good query performance
- Back up and archive data
- Manage queries and direct them to the appropriate data sources

The two main processes Extract and Load and Clean and transform are described below:

Extract and Load: Data extraction takes data from source systems and makes it available to the data warehouse. Data load takes the extracted data and loads it into the data warehouse. It is important to ensure that data is in a consistent state when it is extracted from the source system. Once the data is extracted it is typically loaded into a temporary data store in order for to be cleaned up and made consistent. Performing the load operations in the temporary data store allow the data warehouse to be kept up and running.

Clean and transform data: This system process takes the loaded data and structures it for query performance and for minimizing operational cost. Data is cleaned to ensure the following:

- That data is consistent within itself
- That data is consistent with other data within the same source
- That data is consistent with the data in other source systems
- That data is consistent with the information already in the warehouse

Transformation of data into effective structures: The transform process converts the source data in the temporary data store into a structure that is designed to balance query performance and operating cost.

7. *Understanding data mining problem type(s) to be addressed through modeling*

66

Once the objectives have been formulated, and relevant data has been understood and integrated from various sources, data mining models can be run to analyze this data for the purpose of uncovering knowledge. Running of models requires an understanding of data mining problem types. Below we present various data mining problems types described in the extant literature.

- Data Mining problem types

Data Mining problem types are generally classified into classification, estimation, prediction, association rules, clustering and visualization (Berry and Linoff 1997). Pyle (2003) a slightly different scheme and classify problems based on the modeling intent as (a) modeling to understand, (b) modeling to classify, and (c) modeling to predict. The classification of problem types discussed by other researchers (Cabena, Hadjinian et al. 1998) are generally a subset of the problem categories proposed by Berry and Linoff (1997) and are therefore not discussed separately.

```
                        Data Mining Problem Types
```

Classification	Estimation	Prediction	Association Rules	Clustering	Association Rules
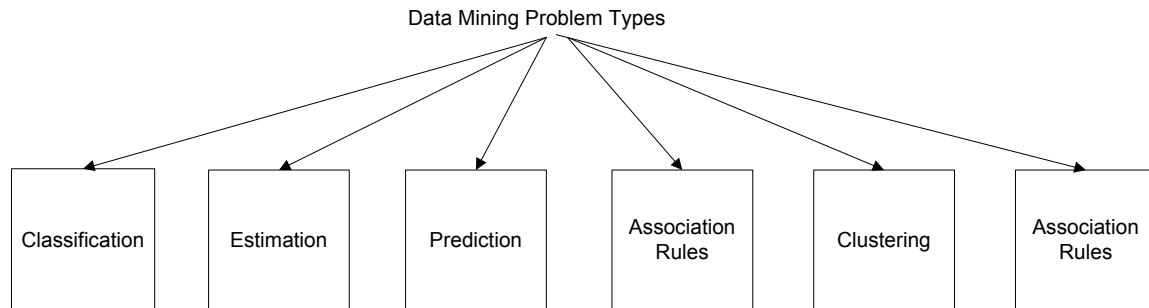
Figure 3-2: Data Mining Problem Types (proposed by Berry and Linoff, 1997)

A. Classification

Classification consists of examining the features of a newly presented object and assigning it to one of a predefined set of classes. The objects to be classified are generally represented by records in a database table or a file, and the act of classification consists of adding a new column with a class code of some kind. The classification task is characterized by a well-defined definition of the classes, and a training set consisting of preclassified examples. The task is to build a model of some kind that can be applied to unclassified data in order to classify it. Some examples of classification tasks include, Classifying credit applicants as low, medium, or high risk; Choosing content to be; Determine which phone numbers correspond to fax machines;

B. Estimation

While classification deals with discrete outcomes such as yes or no, estimation deals with continuously valued outcomes. Therefore, estimation can help derive a value for some unknown continuous variable such as income, height, or credit card balance. Berry and Linoff (1997) point out that in practice, estimation is often used to perform a classification task. They present an example of a credit card company wishing to sell advertising space in its billing envelopes to a ski boot manufacturer might build a classification model that put all of its card-holders into one of two classes, skier or non skier or assign each cardholder a "propensity to ski score" ranging from 0 to 1 that indicates the estimated probability that the cardholder is a skier.

C. Prediction

Prediction is the same as classification or estimation, except that the records are classified according to some future predicted behavior or estimated future value. In a prediction task, the only way to check the accuracy of the classification is to wait and see. Berry and Linoff (1997) note that any techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior. Some examples of predictive tasks include

- Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer

- Predicting which customers will leave within the next 6 months

D. Affinity Grouping or Association Rules

The task of affinity grouping is to determine which things go together. A common example includes determining what things go together in a shopping cart at the supermarket. Affinity grouping can also be used to identify cross selling opportunities and to design attractive packages or groupings of product and services. Affinity grouping is often regarded as a simple approach to generating rules from data. Patterns derived from these algorithms are generally expressed as "90% of all transactions that contain items, A, B and C, also contain item D"

E. Clustering

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters. Clustering does not rely on predefined classes. The records are grouped together on the basis of self similarity. Clustering is often done as a prelude to some other form of data mining or modeling to improve the performance of the predictive modeling technique. Analysis of members of the same cluster could help to derive particular rules.

F. Profiling

Sometimes the purpose of data mining is simply to describe what is going on in a complicated database using visualization techniques. It is assumed that such profiling can help to suggest an explanation for the behavior.

- Description of data mining problem types proposed by Pyle (2003)

Pyle (2003) describes data mining problem types on the basis of modeling intent as (a) modeling to understand, (b) modeling to classify, and (c) modeling to predict. These are described below (Figure 3-3).

A. Modeling to understand

This problem type represents the situation when the miner needs to answer an underlying question about a data set in terms of "why?" This requires developing an explanation and communicating it effectively to the business user. Pyle recommends crafting explanations in one of the following ways: (a) explain one variable at a time, (b) explain linear relationships, and (c) refer to labeled aggregates as wholes (creating new variables to generate meaningful concepts).
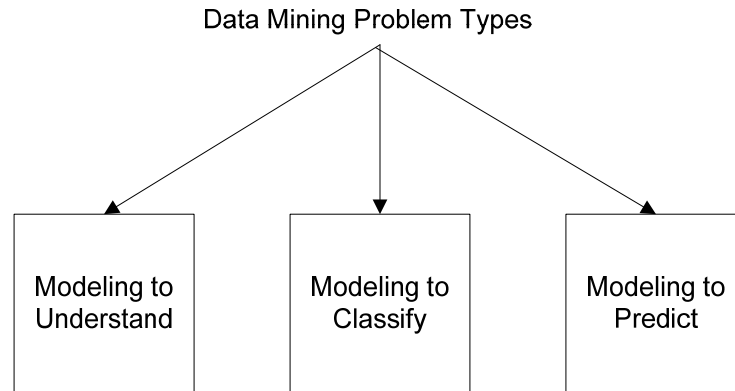
Figure 3-3: Data Mining Problem Types (proposed by Pyle, 2003)

B. Modeling to classify

Modeling to classify typically presents itself as a prediction problem which involves deciding how a dataset should be divided into a set of classes. The predictive model so built could also be used to assign a score to the classified records.

C. Modeling to predict

Prediction is about intelligently forecasting states that have not yet been encountered in existing data. In this sense, it is completely different from classification. Pyle (2003) presents the following example to explain the difference between classification and prediction: a classification model would be used if the goal is to find who will respond to a marketing solicitation. On the other hand, a prediction model would be built if the

goal is to predict who would respond to a marketing offer that has never been tried before.

8. *Analysis of characteristics of various modeling techniques*

Different modeling techniques can be used to address different data mining problem types. Below we summarize some of the popular modeling techniques and their unique characteristics.

A. Decision Trees

Decision Trees (DT's) are a popular technique used for *classification and prediction*. DT is not very popular for estimation; although it can estimate values of continuous variables (Regression and Neural Networks do a better job at estimation of values of continuous variables). They are popular for *exploration* (exploring data to gain insight into relationships between a large number of input variables to a target variable). DT is often used for initial data exploration even when final model is built using some other technique (can be used for selecting the best set of input variables).

- They have high explanatory power and are able to provide easy to understand rules. DT is the preferred choice when presence of rules in data set is suspected.

- They have low sensitivity to outliers and skewed distribution of numeric variables (not sensitive – only uses rank order and not absolute values)

- Missing values: DT can handle missing values in both numeric and categorical input fields by considering null to be a possible value with its own branch. This method of handling missing values is superior to throwing out records with missing values (leads to a biased training set) or replacing missing values with imputed values (important information provided by the fact that a value is missing will be ignored by the model. See p. 174 B&L)

DT starts by finding which of the input fields make the best split. Measure to evaluate a potential split is purity. High purity means that members of a single class dominate while low purity means that the set contains a representative distribution of classes. Best split (1) Increases purity of record sets by the greatest amount; (2) Does not create nodes containing very few records. If no split is possible, the node becomes a leaf node.

It is important to note that the splitting criterion depends on the type of the target variable and not the type of the input variable.

- If Categorical target variable, then splitting criterion = Gini, Information Gain, Chi Square is appropriate for evaluating the split (regardless of whether input variable providing the split is numeric or categorical)

- If Continuous numeric target variable, then splitting criterion = Variance reduction or F test is appropriate for evaluating the split (regardless of whether input variable providing the split is numeric or categorical).

- if continuous target variable, then we can also bin it and apply Gini, Information Gain, or Chi square.

Effectiveness of classification DT is determined by applying it to a test set: the user is typically interested in assessing the percentage of records classified correctly. If the misclassification rates on training and validation are very different, then it indicates an unstable tree. The performance of a tree is typically evaluated by its lift or error rate on the test data set. The performance of regression decision trees can be evaluated using an accuracy measure such as mean square error or average square error.

Decision Trees suffer from the following limitations:

- DT is not as popular for estimation; when used for estimating values of continuous variables, DT is called a regression tree. But it generates lumpy estimates; all records reaching the same leaf are assigned the same estimated value.

- Theoretically DT can assign records to an arbitrary number of classes, but in reality they become very error prone if the training examples per class get small. Small nodes can lead to big problems.

- In not making use of actual numeric values of variables (which makes them less sensitive to outliers and skewed distributions), DT's throw away valuable information which can be better utilized by other models like NN or regression.

75

▪ DT's can not point to interactions amongst variables; any derived variables should be manually created and modeled in the DT.

B. Neural Networks

Neural networks is regarded as one of a few algorithms that can inherently predict multiple outputs simultaneously; that is, predict values for more than one output variable (Pyle 1999). However, this is not good practice in general since the joint predictions tend to be of lower quality than two separate networks each predicting a single output variable.

Neural networks are a class of powerful, general purpose tools readily applied to prediction, classification, and clustering. Neural networks are used for prediction and estimation problems. A good problem has the following three characteristics:

- The inputs are well understood (the user has a good idea of which features of the data are important, but not necessarily how to combine them)

- The output is well understood (the user knows is to be modeled).

- Experience is available (dataset contains numerous examples where both the inputs and the output are known are available. These known cases are required in order to train the network.)

Neural networks can learn patterns that exist only in the training set, resulting in overfitting. The problem of keeping a neural network model up-to-date is made more

difficult by two factors. First, the model does not readily express itself in the form of rules, so it may not be obvious when it has grown stale. Second, when neural networks degrade, they tend to degrade gracefully making the reduction in performance less obvious. The following issues must be considered when modeling neural networks.

Coverage of values for all features: The most important of these considerations is that the training set needs to cover the full range of values for all features that the network might encounter, including the output.

Number of features: The number of input features affects neural networks in two ways. First, the more features used as inputs into the network, the larger the network needs to be, increasing the risk of over fitting and increasing the size of the training set. Second, the more the features, the longer it takes the network to converge to a set of weights. And, with too many features, the weights are less likely to be optimal.

Size of Training Set: The more features there are in the network, the more training examples that are needed to get a good coverage of patterns in the data.

Number of outputs: It is very important that there be many examples for all possible output values from the network.

C. Memory based Reasoning

The idea of similarity is the central idea. Along with finding the similar records from the past, there is the challenge of combining the information from the neighbors. These are the two key concepts for nearest neighbor approaches. Measures of distance and similarity are important to nearest neighbor techniques. MBR works by searching a database of known records is searched to find pre classified records similar to a new record. These neighbors are used for classification and estimation. MBR methods have two chief strengths:

- One of the strengths of MBR is its ability to use data "as is". Format of the records does not have any impact on usage. However, it is important to have the two operations: A distance function capable of calculation distance between any two records and a combination function capable of combining results from several neighbors to arrive at an answer.

- Another strength of MBR is its ability to adapt. Merely incorporating new data into the historical database makes it possible for MBR to learn about new categories and new definitions of old ones. MBR also produces good results without a long period devoted to training or to massaging incoming data into the right format.

However, MBR tends to be a resource hog since a large amount of historical data must be readily available for finding neighbors. There is also the challenge of

finding good distance and combination functions, which often requires a bit of trial and error and intuition. The following changes facing MBR must be considered each time this solution is considered.

Choosing a balanced set of historical records: The training set is a set of historical records. It needs to provide good coverage of the population so that the nearest neighbors of an unknown record are useful for predictive purposes. A random sample may not provide sufficient coverage for all values. Some categories are much more frequent than others and the more frequent categories dominate the random sample.

Representing the training data: The simplest method for finding nearest neighbors requires finding the distance from the unknown case to each of the records in the training set and choosing the training records with the smallest distances. As the number of records grows, the time needed to find the neighbors for a new neighbor grows quickly. This is especially true if the records are stored in a relational database.

Determining the distance function, combination function and number of neighbors: The distance function, combination function and number of neighbors are the key ingredients in using MBR. The same set of historical records can prove very useful or not very useful for predictive purposes, depending on these criteria.

MBR is a k-nearest neighbor approach. Determining which neighbors are near requires a distance function. Investigating different numbers of neighbors using the

validation set can help determine the optimal number of neighbors to choose. There is no right number of neighbors. The number depends on the distribution of data and the problem being solved. The basic combination function, weighted voting, does a good job for categorical data, using weights inversely proportional to distance. The analogous operation for estimating numeric values is a weighted average.

D. Automatic Cluster Detection

Automatic cluster detection is used for finding meaningful patterns in data. Clustering provides a way to learn about the structure of complex data. Once the proper clusters have been defined, it is often possible to find simple patterns within each cluster. Berry and Linoff (1997) discuss the following characteristics of automatic cluster detection

In clustering, there is no preclassified data and no distinction between independent and dependent variables. In a broader sense, however, clustering can be a directed activity because clusters are sought for some business purpose. In marketing, clusters formed for a business purpose are usually called "segments", and customer segmentation is a popular application of clustering. Automatic cluster detection is a data mining technique that is rarely used in isolation because finding clusters is not often an end in itself.

If all problems had so few dimensions, there would be no need for automatic cluster detection algorithms. As the number of dimensions (independent variables) increases, it becomes increasingly difficult to visualize clusters.

The K-means algorithm is one of the most commonly used clustering algorithms. The "K" in its name refers to the fact that the algorithm looks for a fixed number of clusters which are defined in terms of proximity of data points to each other. Three steps of the K–Means algorithm consist of the following: (1) In the first step, the algorithm randomly selects K data points to be the seeds; (2) The second step assigns each record to the closest seed; and (3) The third step is to calculate the centroids of the clusters; these now do a better job of characterizing the clusters than the initial seeds.

Clusters essentially describe the underlying structure in data. However, there is no one right description of that structure. These tests can be automated, but the clusters must also be evaluated on a more subjective basis to determine their usefulness for a given application.

Formal measures of assessing similarity: Geometric distance between two points is often used for assessing similarity. If two points are close in distance, the corresponding records are similar. Most common way to measure the distance is Euclidian distance. Other commonly used methods are angle between two vectors, Manhattan distance and number of features in common

81

Drawbacks of K-Means method are: (1) it does not do well with overlapping clusters; (2) the clusters are easily pulled off-center by outliers; (3) Each record is either inside or outside of a given cluster.

Clustering techniques are two main types: agglomerative clustering algorithm and divisive clustering. In agglomerative clustering, the first step is to create a similarity matrix. Initially, the similarity matrix contains the pair-wise distance between the individual pairs of records. Various approaches to measure distance between clusters are (1) single linkage, in this method, the distance between two clusters is given by the distance between the closest members; (2) complete linkage method, here the distance between two clusters is given by the distance between their most distant members; (3) centroid distance, where the distance between two clusters is measured between the centroids of each. In divisive clustering, decisions made earlier on in the process are never revisited, which means that some fairly simple cluster may not be detected if an early split or agglomeration destroys the structure.

E. Rule Extraction

Decision trees work by dividing the whole of state space into chunks, so that the data in each chunk characterizes the whole chunk in some particular way. Rule extractors typically are not concerned with state space as such but search for common features among the vectors. Rule extraction works by generating *covering rules*. These rules cover a certain number of instances. These have nominal sensitivity while some algorithms are ordinal. Numerical input usually has to be binned, and binning always

removes some information. All the values in a bin are considered to be equivalent. Sometimes binning is invisible to the miner, but the quality of the rules is dependent on the quality of the binning strategy. Rule extraction is generally limited to producing rules about binary splits of the output variable.

F. Linear Regression

Linear regression is an archetypal statistical technique and one of the most powerful and useful data mining algorithms. It enables the miner to make valuable and insightful discoveries in data. Essentially, linear regression is a way of fitting a single straight line through state space so that the line is as close as possible to all of the points in the space. Fitting a straight line to a data set this way works well for explanation and prediction, so long as the data does not bunch up in a state space in at least a rough approximation to a line. Since the line is fitted to be as close to all the points as possible, if a prediction were needed, and one variable's value is known, the value of the other variable has to be nearby in state space. So returning the value of the line's position for that variable is a good estimate of a reasonable value.

Linear Regression has numeric sensitivity. Advanced variable transformations allow the algorithm to regress all variable types. It finds only linear relationships, and is widely perceived not to be a data mining technique. This technique is very sensitive to anomalous fluctuations in the data, although robust versions of the algorithm are available which are less sensitive to such fluctuations. This algorithm struggles with the

co-linearity in the input variables and also cannot deal with missing data. It only produces an explanation and is sensitive to additive interactions.

G. Bayesian Methods

Bayesian methods are a way of starting with one set of evidence and arriving at an assessment of a justifiable estimate of the outcome probabilities given the evidence. A Bayesian method called naïve Bayes requires all of the variables to be "independent" of each other. There are other Bayesian methods that do not require independence of the input variable, but these become enormously computationally intensive for large data sets in their full form. Bayesian based probability models very often work remarkably well in practice, even when many of the theoretical constraints are obviously breached. Bayesian algorithms have nominal sensitivity and can deal with all variable types, but only through binning. It makes a lot of assumptions about the data that almost certainly don't hold up in the real world. However, this doesn't seem to matter in the results, which very often work well. Bayesian networks can present insights as well as make predictions. It can be very difficult to set up these networks with the exception of naïve Bayes network.

9. *Evaluating output of modeling techniques to determine whether or not it meets requirements*

Evaluation is an important component of the KDDM process. During this step, the output of modeling algorithms is assessed to determine whether or not it meets the

required evaluation criteria. The analytic hierarchy process appears to be a useful technique for the purpose of conducting evaluation. It supports simultaneous assessment of multiple criteria which often characterize KDDM projects.

- Analytical Hierarchy Process (AHP)

The Analytical Hierarchy Process (AHP) developed by Saaty (1980) is a decision making framework used for multi-criteria and multi-objective decision situations. The main idea underlying AHP is that human judgment can be used for performing evaluation amongst a set of alternatives.

AHP recommends decomposing a problem into a set of elements, assigning numerical weights or priorities to those elements, and comparing different alternatives according on the basis of their scores on the chosen set of elements. These various alternatives can then be rank ordered to make a selection. One of the chief strengths of AHP is that it can capture both subjective as well as objective evaluation criteria.

While AHP has been across a wide variety of decision situations, it is not without criticism. Critics of AHP have pointed to unreliability of results owing to use of arbitrary scales (Pöyhönen, Hämäläinen et al. 1997), rank reversals (Dyer 1990; French 1998), and Inducement of Nonexistent Order (Schenkerman 1997) etc. Debates between the critics and proponents have also been presented in the literature (Holder 1990; Holder 1991; Saaty 1991). However, AHP continues to be used as a popular decision

making tools by practitioners and academicians. It has also been incorporated in form of the commercial software Expert Choice.

Osei-Bryson (2004) prescribes a multi criteria decision making approach to guide selection of the best decision tree from a large set of decision trees. The prescribed approach describes the types of criteria that could be used for evaluating the performance of decision trees and using them in a multi-criteria decision making framework to aid selection of the best mode.

- Delphi Technique

Delphi (Linstone and Turoff 1975) may be characterized as a method for structuring a group communication process so that the processes effective in allowing a group of individuals, as a whole, to deal with a complex problem. It is proven to be a popular tool in IS research in identifying and prioritizing issues for managerial decision making. Delphi is also relevant to the evaluation step of the KDDM process, as selected evaluation criteria need to be prioritized before data mining models can be selected. The steps associated with the Delphi approach are as follows:

1. Assemble members based upon expertise in the problem context.

2. Send a survey instrument to all members to collect their views regarding the decision at hand.

3. Organize and analyze the survey results.

4. Send a second survey instrument to each member along with a summary of the results obtained from the first questionnaire. Ask the members to consider the summary results and to fill out the second survey instrument after this activity. Should a particular member's view still be significantly different than the majority, he or she should include an explanation of the rationale behind the different position. This position should be forwarded to all other MDM members.

5. Repeat steps 2 through 5 until a consensus is reached among the members. Should no consensus emerge within an established time limit, the most preferred choice becomes the final decision.

- Nominal Group Technique (NGT)

Developed by Delbecq and Van de Ven (1971) the nominal group technique works best in a consensus context such as group or committee structures. Like Delphi this technique can also be adapted to set up evaluation criteria, their weights and threshold values. The approach requires each participant to perform his or her activities using the following procedures:

1. Each participant writes down his or her opinions and ideas relating to what the decision or choice should be.

2. Using a round-robin approach, each participant presents the ideas on his or her list. Each idea is recorded in a summary list using a flip chart or whiteboard so

that all participants can view the list as it develops. At this point, no discussion regarding the desirability of the idea presented is conducted.

3. After all ideas are presented and listed, the participants ask questions of each other for classification of any of the alternatives on the list.

4. Each participant votes on each idea in the list using a predetermined scale or ranking system. The votes are tallied and the collective's choice is revealed.

The nominal group technique can be performed in a non automated fashion as described in the list of steps or it can be easily computerized so that the entire process is managed and conducted electronically.

88

4 RESEARCH METHODOLOGY

"The natural sciences are concerned with how things are. Design, on the other hand is concerned with how things ought to be, with devising artifacts to attain goals"

- Simon (1996)

4.1 Behavioral Science and Design Science Paradigms

The distinction between and also the complementary nature of natural science and design science is grounded in Herbert Simon's work on the 'Sciences of the Artificial' (Simon 1996). Simon argued that just like natural science is about knowledge of natural objects and phenomena, there should also be artificial science or knowledge about artificial objects and phenomena. Simon's groundbreaking work overthrew the popular paradigm that restricted the task of science to teaching about natural things and elevated the task of creation of artifacts (man made objects) to a scientific status.

Simon explained the relation between natural objects and artificial objects (artifacts) by noting that artifacts are not apart from nature. In fact, he argued "they have no dispensation to ignore or violate natural law, but are at the same time adapted

to human goals and purposes" (p. 3). The same view is expressed by March and Smith (1995) who position natural science and design science as complementary, rather than opposing species of research within Information Systems. They describe natural science as aimed at understanding reality; and design science as aimed at creating artifacts that serve to attain some goal and are technology-oriented.

The relationship between behavioral science (with its roots in natural science) and design science is eloquently summarized by Hevner et al.(2004) in the following way:

> Information Systems are implemented within an organization for the purpose of improving the effectiveness and efficiency of that organization… The behavioral science paradigm seeks to develop and justify theories that explain or predict organizational and human phenomenon surrounding the analysis, design, implementation, management, and use of information systems… The design science paradigm [on the other hand] seeks to create innovations that define the ideas, practices, technical capabilities and products through which the analysis, design, implementation, management, and use of information systems can be effectively and efficiently accomplished (p. 76).

The research objective of this dissertation is to provide an integrated KDDM model with a goal of facilitating more effective and efficient implementation of the KDDM process than what is offered by existing process models. Analysis of this

research objective reveals that it aims to create an artifact (in form of an integrated KDDM process model) that addresses a solved problem (implementation of the KDDM process) but in more effective and efficient ways.

Given the design-oriented research objective, it can be adequately addressed using the design science research paradigm which aids fulfillment of identified business needs through building and evaluating appropriate artifacts (March and Smith 1995; Järvinen 2000; Hevner, March et al. 2004). Further, design science research addresses important unsolved problems in unique or innovative ways or solved problems in more effective or efficient ways (Hevner, March et al. 2004). We can see that the research objective of this dissertation corresponds to the latter situation, solving solved problems (namely implementation of the knowledge discovery and data mining process) but in more effective or efficient ways. It must be pointed out that the dissertation covers all phases of the KDDM process except the deployment phase which is excluded from the scope of the dissertation.

It is relevant to note the analytical evaluation of the designed artifact (see Table 3-1) will include the use of qualitative techniques in form of semi-structured interviews and structured surveys to assess the static properties of the artifact. As recommended by Hevner et al. (2004) we will evaluate components of the artifact in terms of its completeness, consistency, performance, usability, efficacy and ease of use and as recommended by Norman (1988) in terms of its simplicity.

In the remainder of this chapter we discuss the state of design science research in Information Systems and how the dissertation makes use of this methodology to achieve the set research objectives.

4.2 State of Design Science Research in Information Systems

The result of design science research in IS (Gavish and Gerdes 1998; Markus, Majchrzak et al. 2002; Aalst and Kumar 2003) is, by definition, a purposeful IT artifact created to address an important organizational problem (Hevner, March et al. 2004). Utility is the hallmark of design science research and for the artifact to be considered useful it must help address a relevant organizational problem. This characteristic of design science research is reflected in the form of emphasis on relevance or practical significance of the outputs of design science work.

Rigor is achieved in design science research by appropriately applying foundations (such as theories, frameworks, instruments, constructs, models, methods and instantiations) during the building of the artifact and methodologies (such as analytical, observational, testing etc) during its evaluation (Hevner, March et al. 2004). For example, Gavish and Gerdes (1998) designed five anonymity mechanisms for a group decision support system and provided a set of formal proofs that the claims made by them were correct and drew their validity from the knowledge base of related past research. It is important to note that while achieving a combination of rigor and relevance is also stated as the desired objective of design science research, achievement of rigor at the cost of relevance is highly discouraged.

92

These characteristics of design science research make it less susceptible to the general criticism of IS research as being 'too rigorous but hardly relevant' (Applegate and King 1999; Benbasat and Zmud 1999). However, this leads to another issue in form of dominance of behavioral science research as compared to design science research, at least in the context of IS research in North America (Lee 2000).

Orlikowski and Iacono (2001) who conducted a survey of articles published in leading IS journals, concluded that 'IS researchers tend to give central theoretical significance to the context within which some usually unspecified technology is seen to operate'. They make a call for an increased focus on technology by stating that the IS field should 'begin to take technology as seriously as its effects, context, and capabilities. Benbasat and Zmud (1999) regard this dominance of behavioral research with its overemphasis on rigor as problematic and caution that the relevance of IS research is directly related to its applicability in design.

However, it would be incorrect to equate design science research as being highly relevant and behavioral science research with lack thereof. March and Smith (1995) note that it is important to study the research intent behind behavioral or design science research before drawing conclusions about their relevance or practical usefulness". They present an example that "a [natural science] account of failure of an information system may be more relevant to practice than the development of a new data modeling formalism [design science]".

Clearly, for IS research to be effective, there is need for emphasis on both the technological and behavioral aspects of information systems. As Lee (2000) notes, technology and behavior in information systems, are not dichotomous but inseparable. It follows that dominance of any one research paradigm (behavioral or design science) is likely to be problematic. The change of focus of the IS research field from a fixation on behavioral science to a balanced mix of behavioral and design science research, in effect requires application of principles of design. It is design that offers us the capability of changing existing situations into preferred ones' (Simon 1996).

According to Simon, what professionals do is to "transform an existing state of affairs, a problem, into a preferred state, a solution" (Schon 1990). IS researchers who have made the call for radically revising the existing state of affairs, by encouraging IS authors to rethink about research topics and readability of manuscripts and for IS journal editors to rethink acceptance/rejection criterion for research papers (Benbasat and Zmud 1999), encouraging theorizing of the IT artifact (Orlikowski and Iacono 2001), encouraging IS academicians to study methods used by IS consultants (Davenport and Markus 1999) are all engaging in the process of design.

4.3 Application of Design Science Research Guidelines

This dissertation utilizes the IS research framework and seven guidelines proposed by Hevner et al. (2004) for conducting design science research. The research framework can be seen in Figure 4-1. These guidelines are summarized below. The research framework proposed by Hevner et al. (2004) incorporates the processes related

to natural (behavioral) and design science research proposed by March and Smith (1995). These include two processes 'develop' and 'justify' related to natural (behavioral) science research and the processes 'build' and 'evaluate' related to design science research.
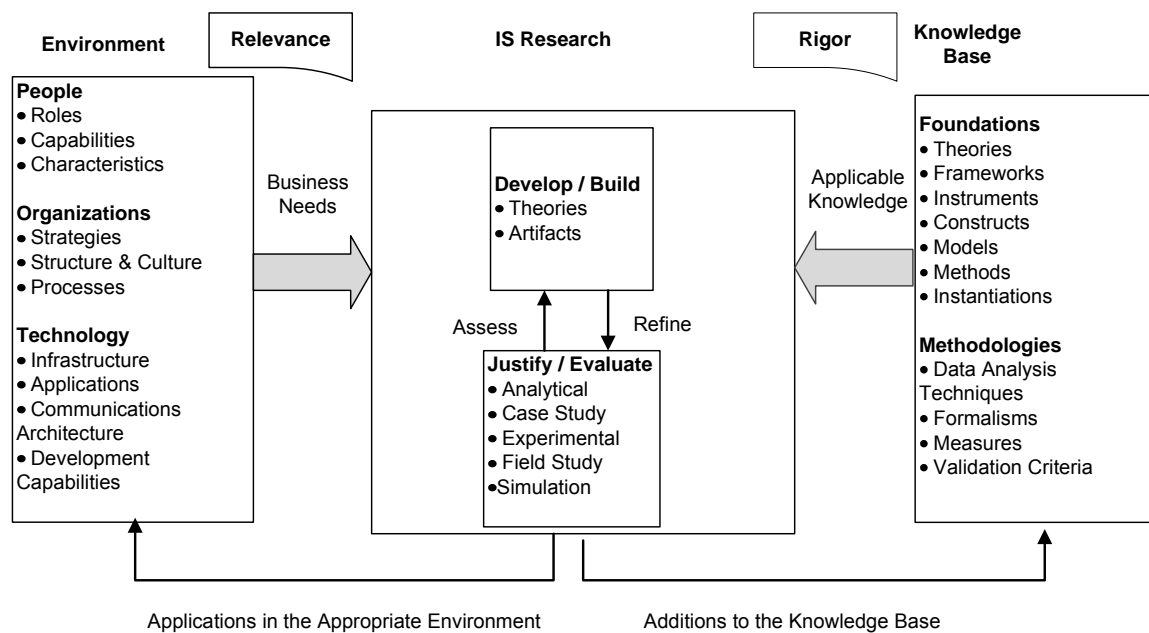
Figure 4-1: IS Research Framework proposed by Hevner et al. (2004)

- **Guideline 1: Design as an Artifact**

The output of design science research in Information Systems is to create an artifact that addresses an important organizational problem. The artifact can be a

construct, method, model or instantiation. Both the artifact and the process by which it is created form part of the design science research process.

Application of Guideline1 in the dissertation: The integrative KDDM process model is an artifact (method) that addresses an important organizational problem - the need for more effective and efficient guidance and decision support towards implementation of data mining projects, than what is afforded by existing process models. In chapter 1, we discussed the limitations of existing KDDM process models and highlighted that,

(1) The existing models do not capture the dependencies existent in a knowledge discovery and data mining process. The dependencies exist because of the interrelationships between the various phases and tasks of a KDDM process. In other words, various tasks/phases require the output of preceding tasks/phases as their input. A checklist approach reflected in the existing process models neglects these important dependencies and is therefore likely to lead to sub-optimal results from data mining projects.

(2) The models do not offer support in form of relevant tools and techniques for the diverse array of tasks described by them (Charest, Delisle et al. 2006). The lack of support can be regarded as responsible for the ad hoc implementation of the KDDM process wherein not all tasks are actually executed. This lack of tool support also exists in the Modeling Phase where although a variety of modeling techniques are available, no support is provided for the critical task of selection of the appropriate set of techniques that are relevant to the problem domain.

The IT artifact described in this research can be categorized as a method under the design science research framework offered by Hevner et al. (2004). It is a purposeful artifact as it aims to address the limitations of existing models thereby leading to more efficient and effective data mining projects.

- **Guideline 2: Problem Relevance**

The research problem addressed through design science research in IS should be of relevance to practitioners who deal with information systems and the technologies that enable the development and implementation of information systems. The research problem becomes relevant by addressing the problems faced by this community. The artifact proposed through the design science effort solves a relevant problem and can be used by organizations that are constantly in need of appropriate artifacts.

Application of Guideline 2 in the dissertation: The research objective of this dissertation relates to an important organizational problem and is of relevance to both academicians and practitioners who deal with implementation of various aspects of the knowledge discovery process. The KDDM process model can be used by practitioners to execute a real world data mining project. The decision support tools and techniques identified in the dissertation can be used for enabling the execution of the various tasks and thereby overcome the problem of lack of decision support towards tasks which may result in neglect of their execution during the KDDM process.

- **Guideline 3: Design Evaluation**

Evaluation is an important component of the research process. The utility, quality and efficacy of a design artifact must be demonstrated through well executed evaluation methods. Given that design is an iterative process, the evaluation phase provides essential feedback to the construction phase which can then be used to refine the artifact so constructed. A design artifact is considered complete and effective when it satisfies the requirements and constraints of the problem it is supposed to solve. The design artifact can be evaluated through the following design evaluation methods.

- Observational (through case studies and field studies)

- Analytical (through static analysis, architecture analysis, optimization and dynamic analysis)

- Experimental (through controlled experiments and simulation)

- Testing (through functional or black box and structural or white box testing)

- Descriptive (through informed arguments and scenario construction)

Application of Guideline 3 in the dissertation: This dissertation will utilize four of the five evaluation methods outlined in the above section. These include analytical, experimental, testing and descriptive approaches and are described in the section below.

The dissertation does not include evaluation of the artifact using observational methods such as case study or field study as the logistics of arranging and implementing such evaluation could be challenging if not insurmountable given the competitive business environments in which DM is used.

Given that the implementation of an artifact can cause enormous change, this study focuses on first studying the viability and usefulness of the artifact through other methods which do not require its implementation. The feedback obtained from the evaluation methods described in the dissertation can be used to improve the design of the artifact. In the future, the tested and improved artifact could be deployed in a real setting and its effectiveness may be tested using the observational methods. Below we provide more details about each of the evaluation techniques proposed to be applied in the dissertation.

Analytical

It is relevant to note that the analytical evaluation of the proposed artifact will employ both quantitative (structured survey to assess static qualities such as perceived usefulness, ease of use etc. of the model) and qualitative methods (Semi-structured interviews with expert users). The exploration of multi-method research has been emphasized in the IS literature (Mingers 2001a). A list of static criteria for evaluation of the proposed artifact has been shortlisted on the basis of relevant extant literature.

User Evaluations of Static Qualities of the Artifact

This dissertation proposes using the evaluation criteria for assessing quality of conceptual models proposed by Maes and Poels (2006) to assess the quality of the proposed model. A conceptual model defines user requirements and is used for designing information systems. The artifact in form of the integrated KDDM process model can also be regarded as a conceptual model which could ultimately be used to design an information system to implement the KDDM process. Hence it is reasonable to evaluate it according to guidelines for assessing quality of conceptual models.

Conceptual model quality is the totality of the features and characteristics of a conceptual model that bear on its ability to satisfy stated or implied needs (Sheer and Hars 1992). Maes and Poels's (2006) model is based on Seddon's (1997) variant of DeLone and McLean's (1992) Information Systems Success Model. The model incorporates the same dimensions as Seddon's model (perceived ease of use, perceived usefulness, and user satisfaction) but replaces the Information Quality dimension of the original model with a Perceived semantic quality construct. Maes and Poels (2006) contend that the Information Quality of a conceptual model users will perceive the semantic quality of the model as how valid and complete it is with respect to (their perception of) the problem domain. Validity means that all information conveyed by the model is correct and relevant to the problem whereas completeness entails that the model contains all information about the domain that is considered correct and relevant. In Table 4-1 below we present the measurement instrument for assessing conceptual model quality proposed by Maes and Poels (2006). The language has been modified to

include KDDM process model instead of conceptual model that is part of the original

instrument.

Table 4-1 Measurement instruments for Perceived Ease of Use, Perceived

Usefulness, User satisfaction and Perceived Semantic Quality constructs proposed

by Maes and Poels (2006)

PEOU1	It was easy for me to understand what the KDDM model was trying to model.	PU1	Overall, I think the KDDM model would be an improvement to a textual description of the KDDM process.
PEOU2	Using the KDDM model was often frustrating.	PU2	Overall, I found the KDDM model useful for understanding the process modeled.
PEOU3	Overall, the KDDM model was easy to use.	PU3	Overall, I think the KDDM model improves my performance when understanding the process modeled.
PEOU4	Learning how to read the KDDM model was easy.	PSQ1	The KDDM model represents the KDDM process correctly.
US1	The KDDM model adequately met the information needs that I was asked to support.	PSQ2	The KDDM model is a realistic representation of the KDDM process.
US2	The KDDM model was not efficient in providing the information I needed.	PSQ3	The KDDM model contains contradicting elements.
US3	The KDDM model was effective in providing the information I needed.	PSQ4	All the elements in the KDDM model are relevant for the representation of the KDDM process
US4	Overall, I am satisfied with the KDDM model for providing the information I needed.	PSQ5	The KDDM model gives a complete representation of the KDDM process

Descriptive: This approach consists of construction of detailed Scenarios around the

artifact to demonstrate its utility. Relevant literature will be used to build an argument

101

about the utility of the proposed KDDM process model and at least one detailed scenario will be built around the proposed KDDM process model to demonstrate its utility.

The analytical and white box testing based evaluation methods propose to employ both experts (such as decision support users, domain experts, data experts, analytical experts) and naïve users in comparing the performance of the proposed KDDM process model to a competing model. The experts are defined as users with more than 1 year of experience with one or more aspects of the knowledge discovery process. The naïve users on the other hand are defined as users with less than 1 year experience with one or more aspects of the knowledge discovery process.

The rigorous evaluation using both expert and naïve users can help to provide insight into the perceptions of the different types of users towards the proposed KDDM process model and if they deem it to be more efficient and effective than other competing KDDM process models. It also helps to assess whether or not an integrative KDDM process model is regarded as more efficient and effective than competing process models by the expert user. The interpretation of differences will provide valuable insight into the utility of the proposed artifact.

The results of evaluation so obtained can help to further both the knowledge base (theories and methodologies for implementing the knowledge discovery process) and to improve upon the design artifact in form of the KDDM process model.

102

- **Guideline 4: Research Contributions**

Every design science research effort must provide one or more of the following contributions:

- The design artifact itself which enables the solution of unsolved problems or solved problems in more effective and efficient ways;

- the extension and improvement of the knowledge base through the development of novel, appropriately evaluated constructs, methods, models or instantiations; and the methodologies in form of use of evaluation methods and proposal of new evaluation metrics.

Application of Guideline 4 in the dissertation: This dissertation contributes to research by presenting an integrated KDDM model which will be rigorously evaluated to demonstrate that it is likely to lead to more effective and efficient implementations of the KDDM process than what is possible through existing models. The designed artifact contributes to the knowledge about the knowledge discovery and data mining process and contributes to the knowledge about the tools and techniques relevant to various aspects of this process that could be used to provide decision support to relevant users and avoid the ad hoc implementation or worse still, neglect of these tasks during the execution of a data mining project.

Guideline 5: Research Rigor

Design science research requires application of rigorous methods in both the construction and evaluation of artifacts. Rigor and relevance should be balanced in the construction and evaluation of the artifact. Knowledge of theoretical foundations is necessary for construction of the artifact and the use of adequate evaluation techniques as outlined in guideline 3 are necessary for its evaluation.

Application of Guideline 5 in the dissertation: The dissertation will conduct the task of building the proposed artifact using relevant knowledge discovery and data mining theories and concepts. The artifact will be tested using adequate metrics specified in the theory to justify that it is a satisfactory solution towards the research objective of the dissertation.

Guideline 6: Design as a Search Process

Design is a search process to discover an effective solution to a problem. Design science research often simplifies a problem by decomposing it into simpler sub problems. The solutions to the sub problems can be regarded as a starting point and progress can be made by expanding the scope and solving the larger design problem. In cases when it is not possible to enumerate all the possible design solutions, heuristics strategies can be used for constructing an artifact that works well for the specified class of problems.

104

Application of Guideline 6 in the dissertation: The design process of the artifact proposed in this dissertation will be developed in an iterative manner. Only the first phase of knowledge discovery process namely domain understanding will be studied in its entirety and all interrelationships between the tasks of this phase will be utilized. Next, the process model will be expanded to cover succeeding phases till all the phases have been considered and the dependencies captured. The dependencies so identified will be logically analyzed to discover flaws in the model (redundant paths, more paths than needed) that may lead to inefficiencies in the implementation of the KDDM process. Any identified limitations will be addressed by refining the model until no further deficiencies can be identified or until the model represents at least a satisfactory solution (Simon 1996) towards the set research objectives.

- **Guideline 7: Communication of Research**

Design science research must be effectively communicated to both technology oriented as well as management oriented audiences. The former need enough details about the artifact to be constructed and used within an actual organization whereas the latter need enough details to determine whether or not organizational resources should be committed to constructing or purchasing the artifact.

Application of Guideline 7 in the dissertation: The results of the dissertation will be presented in a manner suitable to both technology and management oriented audiences. The former will be presented with detailed information about the techniques used in building the model and the results of its experimental evaluations while the latter will be

105

presented with results demonstrating its utility and effectiveness in solving the set research objective.

5 Towards an Integrated Knowledge Discovery and Data Mining (IKDDM) Process Model

This chapter presents the design of the proposed solution offered by this dissertation in form of a new Knowledge Discovery and Data Mining (KDDM) process model, hereafter referred to as IKDDM or the Integrated Knowledge Discovery and Data Mining Process Model. The proposed process model follows the generally accepted sequence of phases in a typical data mining project, ranging from business (or domain) understanding, to data understanding, data preparation, modeling and evaluation. The final phase, deployment (or implementation) wherein the outcome of the data mining project is deployed in a real world organization, has been excluded from the scope of the proposed KDDM model. The sequence of phases in a typical KDDM model is shown below (Figure 5-1). The design requirements for the IKDDM model established earlier are restated in Table 5-1.
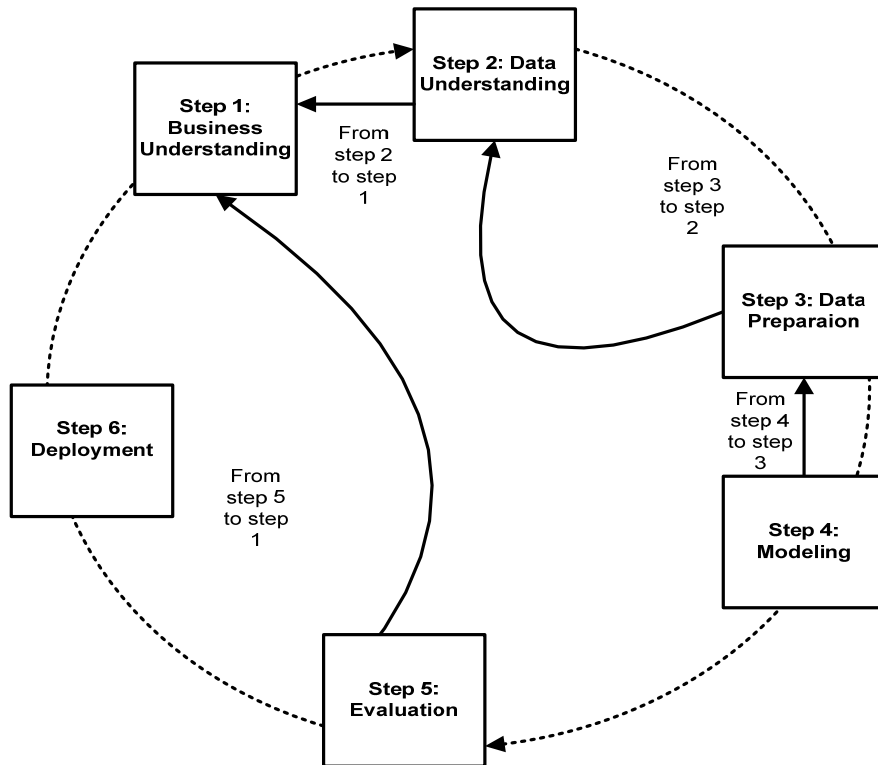
Figure 5-1: Sequence of phases in a typical KDDM process model

Table 5-1: Design Requirements for the Integrated KDM model

Issues Identified with existing KDDM Process Models (As-is situation)	Aspect of	Design Requirements for the IKDDM model (To-be situation)
Description of the KDDM Process in a Checklist Manner	Process Model Domain	Present a user-oriented coherent description of the KDDM process
Fragmented View of the KDDM Process	Process Model Domain	Develop an integrated view of the KDDM process by explicating the various phase-phase and task-task dependencies
Emphasis on feedback loops prior to completely understanding the primary sequencing of phases and tasks in a KDDM process	Process Model Domain	Explicate sequencing of the various phases and their tasks before identifying feedback loops and establishing conditions under which the loops would get triggered
Fragmented view acts as a hindrance to building an integrated process model and "semi-automating" tasks	Process Model Domain	Leverage the dependencies explicated in the integrated process model to drive semi-automation of tasks, wherever possible
Lack of support for the end-to-end KDDM process	Process Enactment Domain	Prescribe approaches for offering decision support towards all tasks in all phases, described in the integrated KDDM model
Visible lack of support towards execution of tasks of the Business Understanding phase - the foundational phase of a KDDM process	Process Enactment Domain	Provide support for tasks of this foundational phase and use them as a basis for developing the integrated model

5.1   Steps for Creating the IKDDM Process Model

The design of the proposed model incorporates treating each phase and its constituent tasks to understand the task-task dependencies existing amongst the various tasks of the same phase. The next step was to integrate the various phases together by

109

linking the task-task dependencies existing among tasks of different phases. The final step was to carefully analyze all the task-task dependencies (dependency relationships between tasks of the same phase and between tasks of different phases), to identify opportunities for leveraging the dependencies so identified through semi-automation, wherever possible. An equally important simultaneous consideration was to prescribe techniques and/or tools for implementing each task of the KDDM process. The steps for creating the IKDDM model are summarized in Table 5-2 below.

Table 5-2: Summary of Steps for Creating the IKDDM Model

1. Study each phase in detail and identify all existing task-task dependencies between tasks of each phase
2. Identify task-task dependencies between tasks of different phases
3. Suggest semi-automation of execution of tasks by leveraging task-task dependencies, wherever possible
4. Propose clearly defined techniques for implementing the remaining tasks
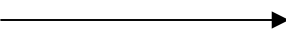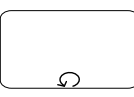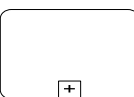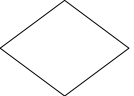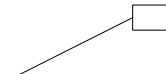
The description of each phase starts with a listing of the tasks of that phase, their output, steps or methods for implementing tasks and an asterisk mark indicating that the task is a candidate for semi-automation. A process model schematic for each phase is also included. Finally after the discussion of each of the independent phases, the process model schematic for the IKDDM model is presented. Each process model schematic included in this chapter is drawn using the Business Process Modeling notation, the

www.manaraa.com

standardized graphical notation for drawing business processes in a workflow [see White and Miers, 2008 for a reference guide for BPMN]. The process models depicted in this dissertation make use of a subset of the total set of graphical elements in the BPMN notation. The graphical elements used and their meaning are included in Table 5-3.

Semi-automation of tasks

The IKDDM model proposes a total of 16 candidate tasks for semi-automation. Of these, 13 tasks belong to phases other than the modeling phase. This is an important contribution of the IKDDM model, given that popular opinion suggests that it is only the tasks of the modeling phase (and to some extent tasks of the data preparation phase) that can be semi-automated or completely automated.

Table 5-3: Graphical Elements (and their meanings) of BPMN Notation Used in this Dissertation

BPMN Element	Meaning
◯	Start Event
◯	End Event
◉	Error resulting from sequence flow
◉	Indicates a message resulting from a sequence flow
→	Sequence flow
▭	Task
▭	Looping (Indicates a repeated task)
▭	Collapsed Sub Process (compound activity included in a process)
▯	Data Object
◇	Decision Box (Indicates flow can take two or more alternate paths)
	Annotation (text box to add callouts or notes)

5.2   Business Understanding Phase

Table 5-4: Business Understanding Phase: Tasks, Methods/Approaches for implementation, Output, and Dependencies

Business Understanding Phase	Approaches/Steps	Output
Creation of Business Objectives	Steps specified based on: - Value Focused Thinking - Goal Question Metric Approach - SMART criteria	Well Formulated Business Objective
Identification of Business Benefits	Steps specified	Clearly identified Business Benefits
Setting up of Business Success Criteria*	Steps specified Based on: Goal Question Metric Approach	List of Business Success Criteria
Formulation of Data Mining (DM) Objectives	Steps specified Based on: Goal Question Metric Approach	Well formulated Data Mining Objectives
Setting up of Business Requirements	Steps specified	List of Business Requirements
Identification of relevant Personnel*	Ontologies, Organization Charts, Skills/Competency Base	List of Available Personnel
Assessment of Policy, Legal and Budgetary Constraints	Steps specified	List of Policy, Legal and Budgetary Requirements
Setting up DM Success Criteria*	Goal Question Metric Approach Cross Reference Matrix	List of Data Mining Success Criteria or Evaluation Measures
Identifying Applicable Modeling Techniques*	Steps Specified Cross Reference Matrix	Array of Applicable Modeling Technique
Assessment of applicable modeling techniques against Data Mining success criteria*	Steps Specified Cross Reference Matrix	List of Data Mining Success Criteria supported by chosen techniques
Analysis of applicable DM tools*	Steps Specified Cross Reference Matrix	List of tools, applicable techniques and selected DMSC supported by tool
Determination of Preference Function*	Steps specified Preference Function Elicitation Tool (e.g. AHP)	Preference Function (e.g. weights for Evaluation Measures)
Determination of Value Functions for Relevant Evaluation Measures*	Steps specified	Value Function(s)
Identification of Applicable Data Resources	Steps specified Metadata Repository	List of Required Data sets
Estimation of Data Collection, Implementation and Operational Costs	Project Management Cost estimation Tools	Statement of Expected Costs
Cost-Benefit Analysis*	Automated Cost Benefit Analysis Tools	Statement of Costs & Benefits

Below we describe various tasks of the business understanding phase, how these can be implemented, and dependencies amongst the various tasks.

Formulation of Business Objectives

DEPENDENCY WITH TASK
Setting up of Organizational Objectives

Business objectives of a data mining project mark the start of a data mining project. Their importance cannot be overemphasized as it is the business objectives that determine the direction for the entire data mining project. The business objective of the DM project cannot be created in isolation from the overall organizational objective(s). It must be ensured that the business objective of the Data Mining Project satisfies one or more of the organizational objectives. Unless this is the case, a judicious management would not sanction use of resources towards a DM project. The SMART acronym proposed by Doran (1981) is commonly recommended for setting objectives. The approach underlying SMART suggests that objectives should be specific, measurable, achievable, relevant and timely. We argue that while the SMART approach provides a way of assessing the 'goodness' of a statement of objective(s), it does not provide detailed guidance towards formulation of objectives, particularly how business objectives of projects originate or how the business objectives can be translated into data mining objectives (a related and highly important task).

We therefore recommend the use of the SMART criteria to evaluate the preliminary statement of business objective(s) and to refine and reformulate it until it satisfies various SMART criteria. But for guidance towards formulating objectives themselves, we propose the use of (1) Value Focused Thinking (Keeney 1992) to stimulate discussion about the setting up of business objectives of the project , followed by the use of (2) Goal Question Metric Approach (Basili and Weiss 1984) for step-by-step guidance towards setting up a well formulated business objective. Below we present all three approaches (VFT → GQM → SMART) and how they could be used in the context of a data mining project. The sequence of steps for applying these approaches is also depicted in Figure 5-2

Step 1: Stimulating discussion about Business Objectives: Applying Value focused thinking

Keeney (1992) highlighted that decision making often focuses on alternatives and only afterwards addresses objectives or criteria to assess these alternatives. He labeled such reactive thinking as 'alternatives focused thinking' and argued that such thinking takes away the control of the decision situation from the decision maker. Keeney contended that since various alternatives are after all only a *means* to achieve *values*, it should be values that drive the decision making process of selecting amongst alternatives and not vice versa.

```
┌─────────┐
│ Step 1  │
└─────────┘
┌──────────────────────────────────────────────────────────────────────┐
│ Apply Value Focused Thinking to Stimulate Discussion about Business    │
│ Objective                                                              │
└──────────────────────────────────────────────────────────────────────┘
                                    │
                                    ▼
┌─────────┐
│ Step 2  │
└─────────┘
┌──────────────────────────────────────────────────────────────────────┐
│ Apply GQM to Generate Preliminary Statement of Business Objective      │
└──────────────────────────────────────────────────────────────────────┘
                                    │
                                    ▼
┌─────────┐
│ Step 3  │
└─────────┘
┌──────────────────────────────────────────────────────────────────────┐
│ Assess Preliminary Statement of Business Objective against SMART       │
│ criteria                                                               │
└──────────────────────────────────────────────────────────────────────┘
                                    │
                                    ▼
┌─────────┐
│ Step 4  │
└─────────┘
┌──────────────────────────────────────────────────────────────────────┐
│ Refine Preliminary Statement to Add Information from Assessment of      │
│ Measurability Criterion (if non quantitative focus) and Assessment of  │
│ Time-Boundedness Criterion                                             │
└──────────────────────────────────────────────────────────────────────┘
                                    │
                                    ▼
              FINAL STATEMENT OF BUSINESS OBJECTIVE
```

Figure 5-2: Steps for Formulating Business Objective: Application of VFT, GQM and SMART Approaches

It appears that data mining projects also often suffer from alternatives focused thinking. Study of case studies reveals that often a brief description of a problem situation is quickly followed up by discussion of alternatives in terms of what type of model (classification, estimation, prediction, association rules etc), or what type of data mining technique (decision tree, neural network, regression etc), would best address the

116

problem scenario at hand. There is typically no guidance provided towards how the business objective in the context of a DM project was or could be formulated.

Keeney (1996) acknowledges the same concern and asks that while 'clear objectives are useful, how should they be created?'. He defines an objective as a statement of something one wants to achieve in a particular decision context. He proposes that each statement of objectives must contain three features: a decision context, an object and a direction of preference. In his work on Value focused thinking (VFT), Keeney discusses two types of objectives: fundamental objectives and means objectives. Fundamental objectives are ends that decision makers value in a particular decision context; means objectives are methods to reach towards those ends.

We posit that in the context of DM projects, fundamental objectives are the business objectives of the data mining project. The data mining objectives (the technical objectives) are the methods for accomplishing the business objectives or the ends. Consider, the following commonly used data mining objective as an example: predict which customers are most likely to respond to a promotional offer? Is this objective, a fundamental or means objective? In the absence of any approach, different individuals may categorize it differently.

To overcome such issues, Keeney (1994) recommends applying the "why is that important?" (WITI) test to distinguish between fundamental and means objectives. If the decision maker answers that a particular objective is essential to a decision context,

then that objective is a fundamental objective. If however, he or she says that a particular objective is important due to its implication for other objectives then it is a means objective. With respect to the above example, a decision maker might answer that this objective is important because the company wishes to increase the response rate from customers, which in turn would mean that it is the desire to increase response rates that is the fundamental objective. By accurately predicting which customers are most likely to respond, the company can direct offers towards the customers most likely to apply for the offers, thereby accomplishing the fundamental objective of increasing response rates.

Keeney's approach provides a starting point for stimulating discussion towards setting up of business objectives and suggests that organizations ask what they value most in a particular decision context to formulate the objective. While the importance of the approach cannot be undermined, it may be difficult to implement by business users involved in setting up the objective. More specifically, the business users may find it difficult to formulate a statement of business objectives for their project using this approach. We posit that the GQM (Goal Question Metric) approach described below should be utilized for step-by-step guidance in formulating the business objectives.

Step 2: Creating a Well Formulated Business Objective: Applying the Goal Question Metric Approach

The GQM approach provides a process for setting goals, and measures for evaluating the goals, and is supported by specific methodological steps. In the context of DM projects, the approach can be applied to determine business objectives, data mining objectives, and business success criteria (measures) for evaluating the business objectives. The latter two are discussed in the following sections as they are independent tasks in the DM process.

The GQM approach was originally developed to establish a goal driven measurement system for software development (Basili and Weiss 1984). It is a top down approach in that a team starts with organizational goals, defines measurement goals (conceptual level), poses questions to address the goals (operational level), and identifies metrics that can provide answers to the questions (quantitative level). While originally developed as a measurement methodology for software development, its use has now been extended to many other contexts.

In applying the approach, we had to adapt it for an entirely new context, knowledge discovery and data mining. To the best of our knowledge, this is the first implementation of the GQM approach to formulate objectives (business and technical/data mining) and success criteria of a data mining project. While we follow the tenets of the approach in formulating the objectives and success criteria, we also suggest some enhancements to the steps, which have been duly noted in the description.

Five Components of Goals proposed by GQM approach

According to the GQM approach, each goal should include five facets of information, namely Object, Purpose, Focus, Viewpoint and Context. The five facets and their examples are included in the Table below.

Table 5-5: Five Information Facets of Goals (Per GQM Approach)

Five facets of Information to Formulate Goals	Example
1. **Object:** the product or process under study	Testing phase or the subsystem of an end product
2. **Purpose:** motivation behind the goal (why the goal is being pursued)	better understanding, guidance, control prediction, improvement
3. **Focus:** the quality attribute of the object under study;	reliability, effort, error slippage
4. **Viewpoint:** perspective of the goal (from whose viewpoint is goal being formulated)	Project manager, developer, customer, project team
5. **Context:** context or scope of the program	Project X, division B

Five Components of Business Objectives in the Context of Data Mining Projects

Let us consider the five components of goals (business objectives) in the context of DM projects.

- *Purpose:* the purpose signifies the motivation behind formulating the objective, or why the objective is being formulated. In the context of Data Mining projects, purpose can be of the following five types:

1. Increase/Improve
2. Decrease/Reduce
3. Identify
4. Understand
5. Determine (Hypothesis Testing)

- *Object Name and Defining Characteristic:* the object is the entity under the study. Examples of objects can include: (1) Customers, (2) Suppliers, (3) Products, (4) Employees, (5) Transactions, etc.

In selecting the object it is important to provide further qualifying information in form of the defining characteristic of the object. For instance, if the object is chosen as simply 'customers', it is may not be clear as to which customers of the firm are of interest and a resultant data mining endeavor may be based on the entire customer base of the firm. However, the results of data mining so obtained are likely to be diluted as it is well known that different types of customers behave differently. So when specifying

121

the object, we must augment it by adding more information, such as (see examples for various types of objects and their defining characteristics in Table 5-6.

Table 5-6: Objects and their Defining Characteristics

Objects	Defining Characteristics
Customers	Wireless internet Customers
	Customers with tenure > 1
	Customers acquired though marketing channel
	most loyal Customers
Suppliers	Suppliers for Eastern Region
	Suppliers of small moving parts
	Suppliers of parts X
Products	co-selling Products
	Products from a particular line (baby care or feminine products)
Employees	internal Hires
	part time Employees
	full time Employees
	Contract Employees
	Employees with tenure > 5
Transactions	Transactions that occurred in last week/month/year
	Transactions valued at >$250

- *Focus:* the focus is the variable or the quality attribute of the entity under study, i.e. what is being studied through the data mining project. The focus of a DM project can be on a tangible or quantitatively measurable behavior, or on an intangible attribute. Below we provide examples of both types.

- **Quantitative focus:** such a focus variable can be measured in terms of %, rate, amount etc. For e.g., churn rate or loss rate of a CUSTOMER [OBJECT]

- **Assuming constancy of other variables:** When focus is a quantitatively measured variable, other variables may have to be treated as constant. Constancy of other variables may or may not apply, but the user must be asked to provide this information, whenever applicable. For example, a credit card provider may be interested in increasing approval rates while maintaining the same loss rates. If the latter is not specified, data mining models that lead to an increase in approval rate, but at the cost of increasing bad rates may be created.

- **Qualitative focus:** such a focus variable cannot be measured in terms of %, rate, amount etc. For e.g., factors affecting motivation of EMPLOYEES [OBJECT]

Relation between Purpose and Focus of a Business Objective

Note that the focus of a business objective is closely related to the purpose of the business objective. When the purpose is to 'increase', 'decrease' or 'reduce', the focus is often on a quantitative variable. On the other hand, when the purpose is to 'identify' or 'understand' the focus is typically on a qualitative variable. When the purpose is hypothesis testing, the focus can be quantitative or qualitative depending on what is being hypothesized. Table 5-7 shows examples of some preliminary business objectives with three components, namely purpose, focus, and object (and their defining characteristic) identified.

Table 5-7: Preliminary Business Objectives (Purpose, Focus and Object identified)

Purpose	Focus/Issue	Object
Increase	Approval rates	- New Credit card applicants - Customers acquired through alternate channels
Decrease/ Reduce	Loss/Bad/Charge-off	- Customers with tenure >2 - Sub-prime credit card customers
	Churn Rate	- Handset customers
Identify	List of probable churners	- Customers with tenure > 5
	List of responders to a new offer	- prospective customers
	Factors affecting churn rate	- Handset customers
	Characteristics	- Most loyal customers
	List of yogurt lovers	- Overall customer population
	Co-selling products	- Complete line of products - Line of health and fitness products
	Occurrence of fraud	- Transactions > $250 - Online transactions > $150
Understand	Characteristics	- High risk customers
	Factors affecting retention	- Existing customers with tenure > 3 years
	Reasons behind charge off	- Sub-prime credit card customers
Determine if	Difference in price sensitivity	- Frequent roamers versus other customers
	Difference in likelihood of response to a home equity offer	- Families with children versus others

- *Viewpoint:* the viewpoint reflects the entity from whose perspective the objective is being designed. For e.g., (1) Project manager, (2) Project team, (3) Project sponsor, etc.

- *Context:* Context represents the scope or the environment where the data mining project is being carried out. For e.g., (1) a particular project (project "Manage Churn", project "Retain Customers"), (2) a particular division (Marketing division, Credit Risk Management division, Customer Relationship Management division).

Figure below shows how the five facets of information, can be put together to create

a preliminary statement of business objective of a DM project.

Select Purpose
Increase
Reduce
Identify
Understand
Determine If

Select Focus
Increase [Approval Rate…Other]
Reduce [Bad Rate, Churn Rate, … Other]
Identify [Factors, Characteristics….Other]
Understand [Factors, Characteristics, Reasons… Other]

Select Entity
Increase Approval Rate of [Sub-Prime Customers, Handset Customers… Other]
Reduce Churn Rate of [Hand Set Customers…Other]
Identify characteristics of [Good Customers… Other]
Understand Reasons for [Churn by Customers with Tenure>3… Other]

Select Viewpoint
Increase Approval Rate of Sub-Prime Customers from the viewpoint of [project manager...Other]
Reduce Churn Rate of Hand Set Customers from the viewpoint of the [Marketing Team… Other]
Identify characteristics of Good Customers from the viewpoint of the [project manager… Other]
Understand Reasons for Churn by Customers with Tenure>3 from the viewpoint of the [customers… Other]

Select Context
Increase Approval Rate of Sub-Prime Customers from the viewpoint of project manager within the context of [project 'bring more customers' … Other]
Reduce Churn Rate of Hand Set Customers from the viewpoint of the Marketing Team within the context of [project 'retain existing customers'… Other]
Identify characteristics of Good Customers from the viewpoint of the project manager within the scope of the [Credit Risk Management Division… Other]
Understand Reasons for Churn by Customers with Tenure>3 from the viewpoint of the customers within the context of [project 'Reduce Churn'… Other]

Preliminary Statement of Business Objective
Increase Approval Rate of Sub-Prime Customers from the viewpoint of project manager within the context of project 'bring more customers'
Reduce Churn Rate of Hand Set Customers from the viewpoint of the Marketing Team within the context of project 'retain existing customers'
Identify characteristics of Good Customers from the viewpoint of the project manager within the scope of the Credit Risk Management Division
Understand Reasons for Churn by Customers with Tenure>3 from the viewpoint of the customers within the context of project 'Reduce Churn'

Figure 5-3: Formulating preliminary statement of Business Objective (based on GQM approach)

Screen Shots for Assisting User in Formulating the Preliminary Business Objective

Task: Setting up of Business Objective - Using GQM Approach

The business objective is the foundation of the data mining project. The following screens will guide you step by step, to create a well formulated business objective. The steps are based on a modified implementation of the GQM approach.

You can press SAVE and EXIT at any time.

Save and Exit   < Back   Next >   Cancel

Creating a Well Formulated Business Objective

Step 1: Selection of Purpose (motivation behind the goal)

Select the Purpose that best represents this problem

○ **Increase/Improve**

◉ **Decrease/Reduce**

○ **Identify**

○ **Understand**

○ **Determine (Hypothesis Testing)**

Save and Exit   < Back   Next >   Cancel

Creating a Well Formulated Business Objective

Step 2: Select the Focus (variable or quality attribute under study)

Specify focus (target variable) variable

Loss Rate
Churn Rate
Approval Rate
Other [Specify Below]

Enter Text

You have selected a quantitative focus variable. Do you wish to Reduce 'Churn Rate' while keeping some other variable CONSTANT?

○ Yes

◉ No

Save and Exit   < Back   Next >   Cancel

Creating a Well Formulated Business Objective

Step 3: Select the Object and its Defining Characteristics (entity under study)

Specify Object

Customers
Suppliers
Employees
Other [Specify Below]

Select Defining Characteristic

Type
Age
Products Purchased
Tenure

Select Type

HandSet
Internet
Satellite Television

Save and Exit   < Back   Next >   Cancel

Creating a Well Formulated Business Objective

Step 4: Select the Viewpoint (entity from whose perspective the objective is being designed)

Select Viewpoint

○ Customers

○ Project Manager

○ Developer

○ Analyst

◉ Project Team

○ Suppliers

○ Regulators

Save and Exit   < Back   Next >   Cancel

Creating a Well Formulated Business Objective

Step 5: Select the Context (the name of the project and the division where the project will be executed)

Enter name of Project

PROJECT REDUCE CHURN

Select name of Division where project is being carried out

Marketing
Finance
Customer Relationship Management
Operations

Save and Exit   < Back   Next >   Cancel

Statement of Preliminary Business Objective

Formulation of statement based on information provided by you in steps 1-5

Preliminary Statement of Business Objective

Decrease Churn Rate of Handset Customers from the viewpoint of the project team within the context of Project 'Reduce Churn' of Marketing Division

In subsequent steps, this objective will be refined to create a final statement of business objective. Press Next to Continue.

Save and Exit   < Back   Next >   Cancel

132

Step 3: Applying SMART Criteria to Refine Statement of Preliminary Business Objective

The preliminary statement of business objectives formulated in the above two steps should be assessed against the various criteria underlying the SMART approach. The definitions of the five criteria underlying the SMART acronym are presented in Table 5-8.

Table 5-8: SMART Criteria for Evaluating Business Objectives

Criterion	Description
Specific	The business objective must lead to an observable action, behavior or achievement which is also often linked to a rate, frequency, number or percentage
Measurable	Concrete, clearly defined criteria should be laid down for measuring the attainment of the proposed business objective. These criteria are referred to as business success criteria and are described as a separate task
Achievable	The business objective must be achievable within the constraints of the available resources, knowledge and time
Relevant	The business objective must be relevant to the broader goals of the organization
Time-Bound	There should be clear deadlines for the achievement of the business objective

Figure 5-4 shows the partial process model (showing subset of tasks) of the Business Understanding Phase. The final view of the process model for this phase is presented at the end of the section and includes all the tasks of this phase.
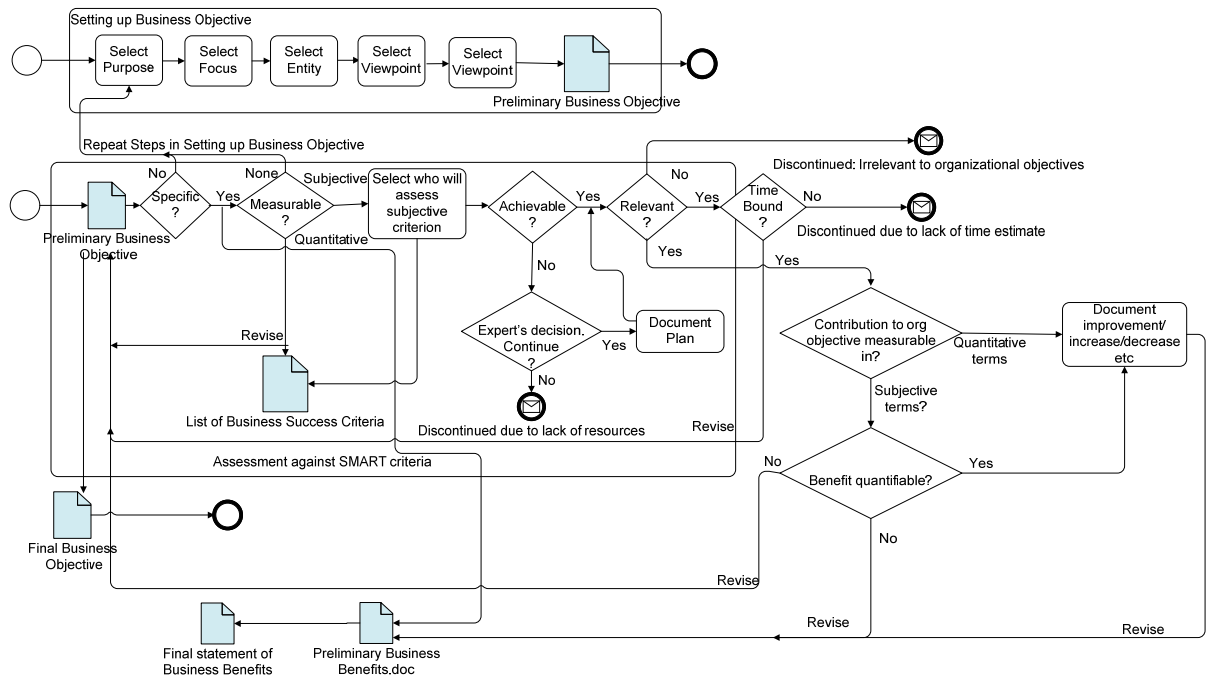
Figure 5-4: Partial View of Process Model for Business Understanding Phase

Step 3-1: Assessing Specificity

The 'specificity' criterion requires that the objective should lead to an observable action, behavior or achievement. In the context of DM projects, such observable action is often specified in quantitative terms, such a percentage improvement in profit, percentage reduction in losses or charge offs etc. It could also be specified in non-

134

quantitative terms such as improvement in customer's perception of the company's product(s), improvement in employee morale etc.

If the preliminary statement of business objectives does not satisfy the specificity criterion, then the steps related formulation of the objective should be repeated. This ensures that the objective will lead to a concrete identifiable outcome.

Assessment of Preliminary Business Objective against SMART criteria

Step 1: Assessment of 'Specificity' (A business objective must be specific, i.e. it must result in an observable action, behavior or outcome)

Will the Business Objective "Decrease Churn Rate of Handset Customers from the viewpoint of the project team within the context of Project 'Reduce Churn' of Marketing Division", result in,

◉ observable action, behavior or outcome measurable in quantitative terms
○ observable action, behavior or outcome measurable in non-quantitative terms
○ Neither

Save and Exit   < Back   Next >   Cancel

Step 3-2: Assessing Measurability

The assessment of measurability criterion helps to determine the business success criteria associated with the project. This criterion stipulates that the business objective must be measurable in quantitative or non-quantitative terms. This step ensures that the

objective formulated is indeed measurable. Based on the focus variable (Set up during the formulation of the objective), two situation arise: focus variable is quantitative in nature or focus variable is non-quantitative in nature.

- In case of *objectives with a non-quantitative focus*, the assessment of measurability should be performed by a domain expert who should set subjective criteria for assessing whether or not the objective was achieved.
- In case of *objectives with a quantitative focus*, the step 2 of the GQM approach (Basili and Weiss 1984) can be used for assessment of measurability, as explained below.

Using GQM for Assessment of Measurability for Objectives with Quantitative Focus

The GQM approach proposes refining the overall goal (business objective in the case of a DM project) into a set of questions, and then refining the questions into a set of metrics that could be objective or subjective. Figure 5-5 shows an example of how questions and metrics can be formulated from a statement of business objective.
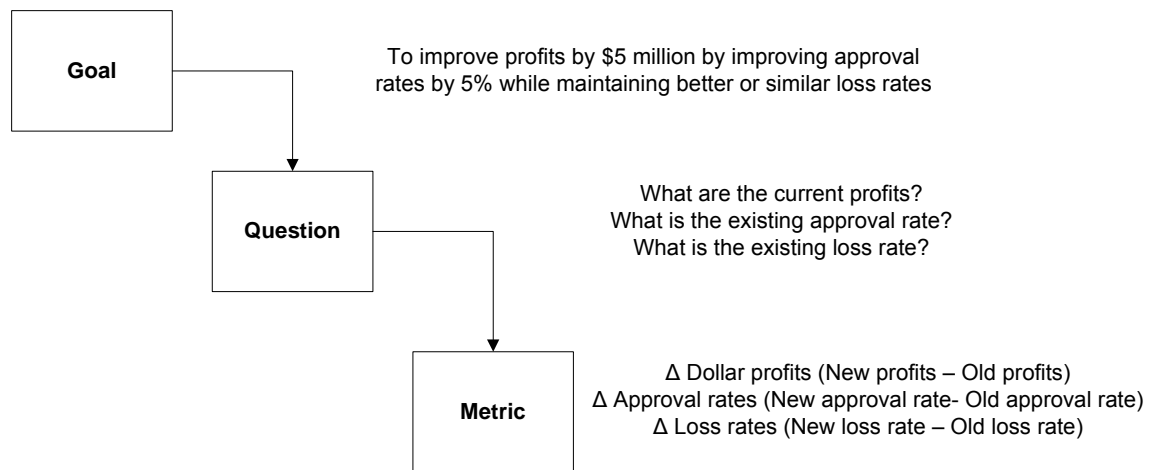
Figure 5-5: GQM approach for setting up of Business Success Criteria

Note that the 'questions' in the GQM approach are based directly on the focus variable or attribute. The metrics describe the business success criteria and must meet the threshold values specified in the statement of objectives. For instance, in case of the example shown in Figure 3: (a) Δ Dollar profits should be >= $5 million; (b) Δ Approval rate >= 5%; and (c) Δ Loss rate <= 0 since the objective is to maintain better or similar loss rates. The sequence of steps is also summarized below.

- Select existing value for focus variable and desired value of focus variable
- The Delta (or difference) between existing and desired values is a Business Success Criterion
- For example, if existing value for charge-off rate is 5% and desired value is 2%,
- Then business success criteria = (5 – 2)/5 = 60% reduction in charge off rate or Δ Charge off Rate = 60%

www.manaraa.com

- Project will only be deemed as successful if it leads to a 60% reduction in charge off rate. Anything less than that will be deemed unsatisfactory.

Assessment of Preliminary Business Objective against SMART criteria

Step 2: Assessment of 'Measurability' (A business objective must be measurable in quantitative or non-quantitative terms)

Based on information provided during selection of FOCUS, the Business Objective "Decrease Churn Rate of Handset Customers from the viewpoint of the project team within the context of Project 'Reduce Churn' of Marketing Division", can be measured in quantitative terms.

Provide existing churn rate in %   5%

Provide desired churn rate in %   2%

Preliminary Business Success Criteria is   60%

Δ Churn rate =

Save and Exit   < Back   Next >   Cancel

Step 3-3: Assessing Achievability

The assessment of achievability criterion helps to establish whether the business objective can be achieved within the constraints of the available resources, knowledge and time. This is an important step because unless this criterion is satisfied we cannot be sure that the business objective could get fulfilled. It is possible that the stakeholders involved in assessing achievability may only have a limited picture of available resources at this point in the project; however they must still use their expertise to consciously assess whether or not the firm possesses (or will be able to secure) the

138

necessary resources in the time frame required by the project. The assessment of achievability incorporates identifying relevant personnel and their availability, which is regarded as an independent task in the DM process, and therefore discussed separately.

Assessment of Preliminary Business Objective against SMART criteria

Step 3: Assessment of 'Achievability' (The business objective must be achievable within the constraints of available knowledge resources and time)

Is the Business Objective "Decrease Churn Rate of Handset Customers from the viewpoint of the project team within the context of Project 'Reduce Churn' of Marketing Division", achievable within the constraints of available knowledge, resources and time?

⦿ Yes

○ No

| Save and Exit | | < Back | Next > | Cancel |

Step 3-4: Assessing Relevance

This criterion ensures that the business objective is relevant to a higher order organizational objective. Unless this is the case, the project cannot be regarded as useful for the organization, making it difficult to approve any funding for its execution. The

stakeholders involved in assessing this objective must clearly specify the particular organizational objective that would be fulfilled if the business objective was carried out.

Step 3-5: Assessing Time-Boundedness

This criterion ensures that there is a clear timeline for the execution of the project and delivery of final results. Unless this criterion is satisfied, it will be difficult, if not impossible to track the progress of the project, allocate critical resources. The information provided by stakeholders in assessing this criterion must be used in refining

the preliminary business objective, to also reflect the time frame during which the project must be completed.

Step 4: Finalizing Statement of Business Objective: Updating Information Gained Through Assessment of Measurability and Time-Boundedness

The statement of business objective generated at the end of step 2 should be revised to add information gained through assessment of Measurability and Time-Boundedness.

Step 4-1: Updating information from Assessment of Measurability

When the focus is quantitative, the measurability criterion helps determine the quantitative improvement/reduction that must materialize in order for the project to be considered successful. This business objective must be updated to reflect this information. When the focus is not quantitative, this step should be skipped. Here is an example:

- Suppose that the statement of business objective formulated using GQM approach is: to Reduce Churn Rate of Hand-Set Customers, from the viewpoint of the Marketing Team, in the context of Project 'Reduce Churn'.

- Let us assume that the assessment of Measurability reveals that the churn rate must be reduced by 2% from 5% to 3%.

- The business objective should be updated as follows: "Reduce Churn Rate of Hand-Set Customers [INSERT DELTA OR DESIRED AND EXISTING VALUES FOR FOCUS VARIABLE] from the viewpoint of the Marketing Team, in the context of Project- Reduce Churn".

- The statement would read: Reduce Churn Rate of Hand-Set Customers by 2% from the viewpoint of the Marketing Team, in the context of Project 'Reduce Churn'.

- Alternatively, the business objective can be stated as: Reduce Churn Rate of Hand-Set Customers from 5% to 3%, from the viewpoint of the Marketing Team, in the context of Project 'Reduce Churn'.

- If any variables were assumed as constant while setting up the focus variable, that information should also be reflected in the business objective. For example, let us assume that after adding information gained during assessment of measurability the statement of business objective is: "Increase approval rates of sub prime customers while 5%, from the viewpoint of the Credit Risk Management Team, in the context of Project 'Bring More Customers'.

- The business objective should be updated as follows: "Increase approval rates of sub prime customers by 5% [INSERT ANY CONSTANT VARIABLES] from the viewpoint of the Credit Risk Management Team, in the context of Project 'Bring More Customers'. Assuming that the loss rate or bad rate was assumed to be constant, the revised statement would read:

- "Increase approval rates of sub prime customers while 5%, while maintaining better or similar loss rates, from the viewpoint of the Marketing Team, in the context of Project 'Reduce Churn'.

Formulation of final statement based on information provided by you in steps 1-5

Final Statement of Business Objective

> **Decrease Churn Rate of Handset Customers by 60% from the viewpoint of the project team within the context of Project 'Reduce Churn' of Marketing Division over the time frame 21-Aug-08 to 15-Aug-09**

If you agree with this statement, press FINALIZE; to go to previous steps press back; to revise statement, enter in text box below

Save and Exit   < Back   Finalize   Cancel

Step 4-2: Updating information from Assessment of Time-Boundedness

Review of the business objective so formulated reveals that it possesses all characteristics of a well formulated business objective, except the time frame during which this objective must be accomplished for the project to be considered successful. This information is collected during the assessment of Time Boundedness.

- Let us assume that the assessment of Time-Boundedness revealed that the project must be accomplished over September 08-August 09. The business objective can be updated as follows. "Reduce Churn Rate of Hand-Set

Customers by 2% from the viewpoint of the Marketing Team, in the context of Project 'Reduce Churn' over [INSERT TIME FRAME]".

- The final statement of business objective would read: Reduce Churn Rate of Hand-Set Customers by 2% from the viewpoint of the Marketing Team, in the context of Project 'Reduce Churn' over September 2008-August 2009.

Formulating Business Objectives: Use of DM Case Base as a Search Tool

A case base or repository of past data mining projects (if available), can also be used as an aid to formulate a business objective. A simple approach could be to search for key words describing problem scenario at hand with problem scenario of past projects in order to identify similar past projects. A study of statement of business objectives can be used to develop business objectives for an existing project and also study any information available about how that objective was formulated.

The importance of such reuse of knowledge from data mining projects has been discussed in literature (Engels 1999; Wirth and Hipp 2000; Becker and Ghedini 2005; Domingos 2007). The systematic documentation of previous knowledge, experiments, data and results, is valuable for management of existing data mining projects (Brachman and Anand. 1996; Zantout and Marir 1999). While organizations and individuals often evolve their own personal strategy of documenting data mining projects, (Wirth, Shearer et al. 1997) such an approach is likely to be suboptimal as it would be guided by what an individual perceives as important enough to be recorded.

145

Becker and Ghedini (2005) have proposed a documentation infrastructure for data mining projects that allows for capture of data mining process related information and all the associated artifacts. If a repository based on such an approach exists, it could provide valuable insights into formulation of business objectives, although even then techniques such as the ones described earlier would be needed to develop the business objective specific to a new problem scenario.

Identification of Business Benefits

DEPENDENCY WITH TASK (OF PHASE)
Determination of Business Objectives (Business Understanding)

Formulation of each business objective should be followed by elicitation of the business benefit(s) that will result from pursuing the objective. Unlike models such as CRISP-DM, the IKDDM model positions this task early in the project because identification of benefits of the project is an important task whose output affects the determination of business success criteria (task 4.1.3). These business benefits may be financial (increase in profits, ROI) or non financial (improvement in customer morale, customer loyalty) in nature. Together, information gained from assessment of measurability and business benefits helps formulate business success criteria. The following steps should be followed depending on whether the business benefits are financial or non-financial in nature.

If expected business benefits are quantifiable in monetary terms

- Present amount of benefit in monetary terms.

- For example, ROI of 10%, profit increase of $1 million, reduction in losses of $2 million etc.

If expected business benefits are NOT quantifiable in monetary terms

- Clarify who will be assessing whether or not the intangible benefits were achieved

- For example, the project manager Ms. Eesha Bansal will assess if a particular campaign led increase in morale among key stakeholders.

Note that sometimes even though these benefits are intangible, a monetary value may be assigned to them, in which case they would belong to the above category (i.e. quantifiable benefits). If however, they cannot be quantified, then the names of personnel who will be involved in making the subjective decision regarding whether the benefits were really achieved must be clarified.

Setting up of Business Success Criteria

DEPENDENCY WITH TASK(S)	PHASE
Assessment of Measurability of Business Objectives	Business Understanding
Business Benefits	Business Understanding

The output of Assessment of Measurability of Business Objectives and determination of business benefits helps in semi-automating the setting up of business success criteria.

- The Assessment of measurability results in delta value signifying difference between a desired and existing values for a quantitative focus variable or the subjective criteria for assessing achievement of a non-quantitative focus along with details about personnel responsible for assessing its achievement.

- The determination of business benefits results in a monetary value for expected business benefits, or the criteria for assessing intangible business benefits along with details about personnel responsible for assessing its achievement.

Given the dependency of the task determination of business success criteria with the assessment of measurability and determination of business benefits, it is clear, that not formally executing either or both of these, will result in failure to set up an accurate set of criteria. This dependency is an example of a dependency that can be leveraged to semi-automate the determination of business success criteria because the output of two preceding tasks directly determines the business success criteria for a project. It is recommended that the domain stakeholders assess the list of business success criteria so generated for any inconsistency or incompleteness and revise them using their domain knowledge

Determination of Data Mining Objective

DEPENDENCY WITH TASK(S)	PHASE
Business Objectives	Business Understanding
Business Benefits	Business Understanding

A data mining objective is often defined as the technical translation of the business objective but this definition by itself does not provide the user with enough guidance regarding creation of a well formulated data mining objective. The IKDDM model proposes that the Goal Question Metric (GQM) Approach can also be used for formulating data mining objectives.

As described earlier (under formulation of business objectives), the formulation of a goal requires information about five different components: (1) purpose (motivation behind the goal); (2) focus (variable or quality attribute under study); (3) object (entity under study); viewpoint (entity from whose perspective the goal is being designed); and (5) context (scope or environment). Below we provide a list of steps to be followed in setting up a data mining goal. Examples of these five components with respect to setting up of data mining objectives are also provided.

149

Step 1: Selection of Purpose

Purpose relates to the data mining problem type. Data mining literature generally distinguishes between six main problem types: classification, estimation, prediction, description (or visualization), clustering and association rules. The IKDDM model proposes removing Prediction and substituting by Prediction (Classification) or Prediction (Estimation). The rationale for doing so is that prediction problems can be modeled as either classification or estimation. The choice of the data mining problem type directly affects numerous other tasks and it is therefore necessary to extract this information when formulating a problem.

Classification, Prediction and Estimation are regarded as instances of supervised or directed data mining as the data mining endeavor is directed at a target variable. In case of unsupervised or undirected data mining (clustering, association rules and visualization, there is no target variable involved). The various supervised and unsupervised data mining problem types are presented in Table 5-9 and Table 5-9 respectively.

The steps related to selection of PURPOSE (data mining problem type) are described below.

Select Purpose from one of the following:

1. Classification

2. Estimation

3. Prediction (Classification)

4. Prediction (Estimation)

5. Clustering

6. Visualization

7. Affinity grouping or association rules (including sequential patterns)

✓ IF PURPOSE selected is Classification, clarify if classification is for the purpose of developing a scoring model (i.e. will the estimated probability values be rank ordered and cut offs applied to classify records into groups or classes.

- Propensity (probability) scoring is an example of this category. While the data mining objective will still be framed as a classification problem, this clarification will help clarify certain other steps in the process (such as determination of thresholds etc).

✓ IF PURPOSE SELECTED IS PREDICTION, then clarify if it is a classification or estimation problem (prediction problems can be modeled as classification or estimation depending on whether we are estimating the future value of a continuous variable or classifying records into classes based on some predicted future behavior.

- Example of 'prediction – classification' is studying characteristics of credit card applicants and dividing them into good or bad classes. Note that the actual

behavior of the applicants whether he was good or bad would only become evident in the future, but we use the existing records (classified as good or bad), and compare the new record (for which the target class is not yet known) to the existing ones to divide into the two classes.

- Example of 'prediction – estimation' is the amount of balance that will be transferred if a customer accepts a credit card offer. Again we can only estimate the value of the continuous variable "balance transferred", but the amount actually transferred only becomes evident in the future.

Table 5-9: Supervised Data Mining problems (with target variable specified)

Problem Type	Definition	Example
Classification	Dividing unseen records into predefined classes	Divide customers into • risky and non risky • loyal and not loyal • good and bad
Estimation	Estimating value of a continuous variable	Estimate annual income of households in zip code 23233
		Estimate amount of balance that a customer will transfer if she accepts a credit card offer
Prediction (Classification)	Classifying records into predefined classes based on "future behavior"	Classify customers into classes 'churn' and 'no churn'
Prediction (Estimation)	Estimating the "future" value of a continuous variable	Predicting the amount of balance that a customer will transfer if he accepts a credit card offer

Table 5-10: Unsupervised Data Mining problems (with no target variable)

Problem Type	Definition	Example
Clustering/Segmentation	Dividing records into clusters or segments	Identify different types of customers from overall customer base
Visualization	Study features, characteristics, factors, relationships	Identify characteristics of most loyal customers
Affinity grouping or association rules	Study co-occurrence of products or variables	Identify co-selling products from line of baby products

Step 2: Selection of Focus

The focus of a data mining goal cannot be divided into a finite set of categories like the purpose of a data mining project. However similar to the focus of a business objective, it can be quantitative or non-quantitative in nature. Examples of a quantitative focus include: a target variable such as bad rate, likelihood of churn, likelihood of charge-off, size of balance transferred, annual income of a household, etc. examples of a non-quantitative focus may include, co-selling products, different types of customers, general characteristics of a sample, etc. The examples of focus for different data mining problem types are presented below.

- For Classification, estimation or prediction (classification or estimation) problems, the focus is the 'target variable' under the study.

- For Classification and Prediction (classification) problems, focus may be a 'Categorical Target' with two classes such as "good" or "bad"

153

- For Estimation and Prediction (estimation) problems, focus may be a continuous target variable such as "household income", or "amount of balance transferred" etc.

- For Clustering problems, the focus is on the 'Types of Clusters or Segments (clusters of OBJECTS' with similar buying habits, of same age, having same spending pattern, buying similar products etc)

- For Affinity Grouping, the focus or the attribute under study is the 'co-occurrence of OBJECTS'

- For Visualization, the focus is on the 'factors, characteristics, relationships'

Step 3: Select OBJECT (entity under study), **OBJECT TYPE** (distinguishing characteristic of the entity) and **TIME FRAME** (period for which the object is to be studied).

- The OBJECT can be (1) customers, products, employees, suppliers, household, etc.

- The OBJECT TYPE can be sub prime applicants, bathing products, contract employees, small parts suppliers', households in zip code 19701.

- The TIME FRAME can be reflected as follows: sub prime credit card applicants 12 months from point of booking, bathing products sold in 2007-2008, contract employees with tenure > 2 years, small parts suppliers with tenure > 3 years, households in zip code 19701 for may 07-may 08.

154

The object and object type of the data mining objective is the same as object and object type of the business objective. Therefore this information is already available at the time of setting up of data mining objective.

Step 4: Select VIEWPOINT (entity from whose perspective the objective is being designed).

- The viewpoint could be that of the project manager, the customer, the project sponsor etc.

The viewpoint of the data mining objective may or may not be the same as the viewpoint of the business objective. for example, while the latter may be from the viewpoint of a project sponsor, the former may be from the viewpoint of the technical project manager.

Step 5: Select CONTEXT (PROJECT and the ENVIRONMENT or DIVISION where the project is being carried out).

- For example, the context could be project "increase visibility" under the Marketing division.

The context of the data mining objective may or may not be the same as the context of the business objective. For example, while the latter may be in the context of the Marketing Team, the former may be in the context of the Decision Science Team.

155

Sequencing steps for Classification and Clustering/Segmentation problems: elicit information about entities before eliciting information about focus

The above described sequence of steps applies to all problems, except classification and clustering/segmentation where a user may be able to identify the entity ahead of identification of the focus of the goal Figure 5-6. Consider for example, the following DM goal for a classification problem: to classify (purpose), customers (entity) into good versus bad (focus). In classification problems, the focus only appears after identification of entities and so it will be more useful to elicit information about the entity to be classified first.

Next, consider the following DM goal for a clustering problem: segment or cluster (purpose) customers (entity) into loyal and not loyal types (focus). Similar to classification the focus only appears after identification of entities and so it will be more useful to elicit information about the entity to be clustered first.

For all other types of purposes (estimation, prediction for estimation tasks, association rules, and description or visualization), the five steps can be performed in the order specified.

If Purpose = Classification, Prediction (Classification) or Clustering	If Purpose = Estimation, Prediction (Estimation), Association, or Description
⇓	⇓
Select **Entity**	Select **Focus**
⇓	⇓
Select **Focus**	Select **Entity**
⇓	⇓
Select Viewpoint	Select Viewpoint
⇓	⇓
Select Environment	Select Environment
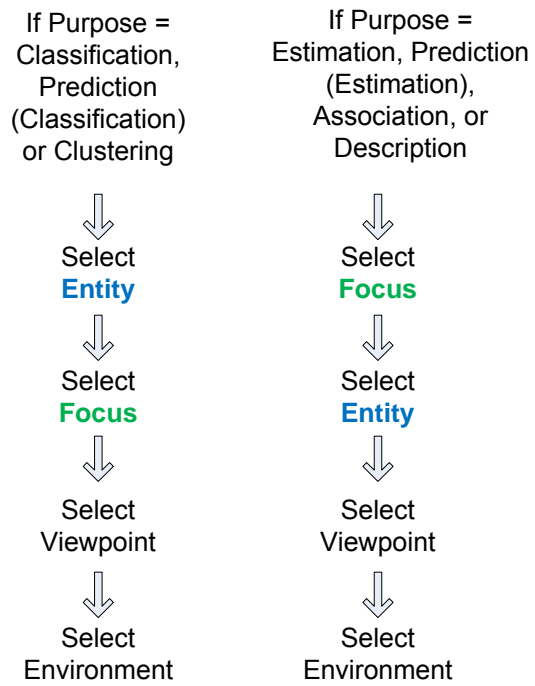
Figure 5-6: Sequence of Steps for Formulating DM Goal for Different Problem Types
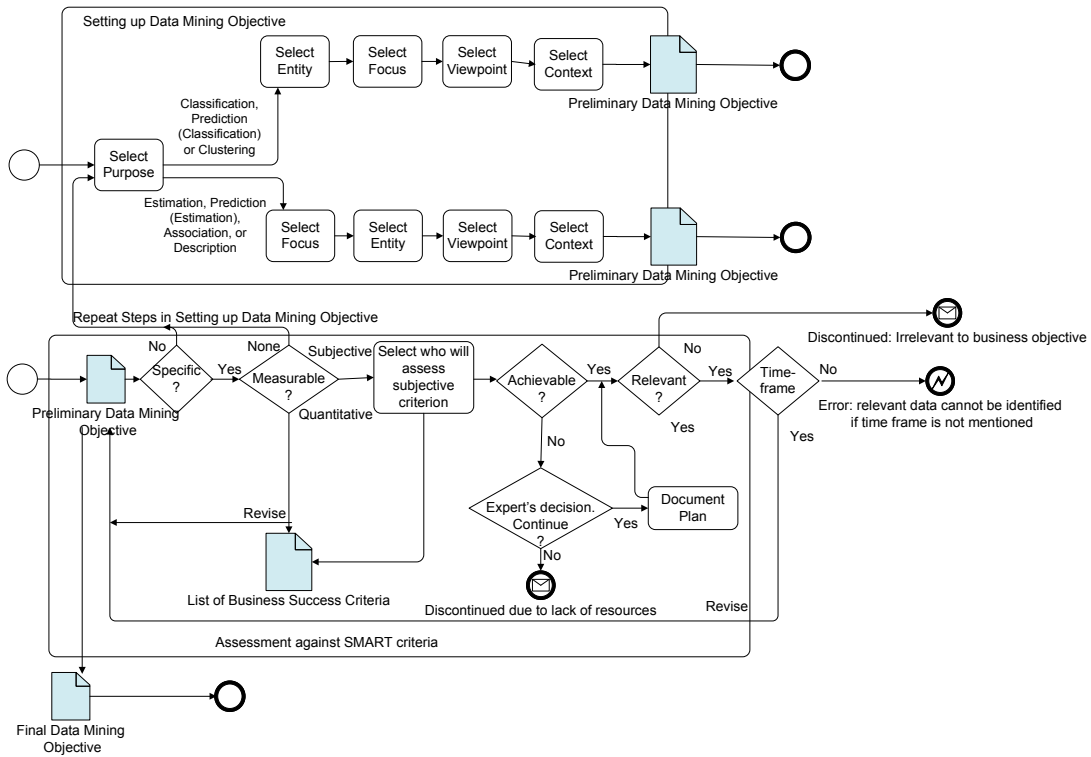
Figure 5-7: Creating Data Mining Objectives: Partial View of Business Understanding Phase

Assessment of need to discretize target variable (applicable if Purpose selected is Classification or Estimation)

If the user selects purpose of the data mining objective as Classification, Estimation or Prediction (Classification), they should be prompted to assess the need to discretize the target variable (i.e. convert it into a finite set of classes). Both categorical and continuous target variables may be discretized based on the business and data mining objectives.

The discretization of target variables is likely to lead to simplification of problem and improvement in results. However, the significance of this important task and methods for implementing it are not discussed by existing KDDM process models.

Note that the motivation behind discretization of target variables is completely different from motive behind discretization of input variables. While the former are discretized strictly based on business and data mining objectives, the latter are discretized based on requirements of underlying modeling technique (e.g. neural network), and are discussed as part of the Data Preparation stage.

Discretizing a Categorical Target Variable:

Discretization of a categorical target variable can help in reducing the number of classes in the target variable. Given its definition, it applies when the target consists of several levels. Discretization is therefore a moot point when the target variable is already discretized into classes, or is binary in nature. To understand how and why a multi-level categorical target variable can be discretized, consider the following example. An educational institute is involved in a data mining project that aims at classifying student applicants based on certain characteristics (input variables) into four categories (best, good, average, poor) as part of its admissions process. This is a Classification problem where target variable is the student rating (best, good, average, poor), and consists of four levels.

In an effort to increase the visibility of the institution as an excellent university, key stakeholders are only interested in intake of students whose student rating puts them in the 'best' category. The plan is to only send acceptance notices to students of this category and deny the rest. In developing the data mining model, the only target level class we are interested in is 'best'; this means that we can discretize the original target level with multiple classes (4 in this case) into a target level with two classes: best and 'other'. 'Other' here would contain students with ratings of good, average, or poor. This is an example of discretization of categorical target variable. Using thus approach we have converted a 4-level classification problem to a binary classification problem.

Discretizing a Continuous Target Variable:

Discretizing a continuous target variable can help to convert an estimation (or regression type problem) into a classification problem. Again, such a need is based on business and data mining objectives. In some cases, the goal may only be to estimate the value of the continuous target variable, whereas in some other cases we may wish to discretize the continuous target variable into discrete classes. For example, consider the case of the educational institute discussed earlier. Let us assume that the institute is involved in a data mining project aimed at estimating the expected cumulative GPA of student applicants. The project is also being conducted as part of the admissions process. This is an example of an Estimation problem where the target variable (GPA) is continuous in nature. It can take on any value between 0 and 4. The estimated values are calculated up to 2 decimal places such as 2.25, 3.65, 4.00 etc.

In the present year, the admissions officials are only interested in sending acceptance notices to students whose expected GPA's are higher than 3.00. The rest will be sent rejection letters. In developing the data mining model, the institute is interested in records where value of target is equal to or greater than 3. Given the objective of this project, the continuous target variable can be discretized into two classes: one with values $>$ or $= 3$, and one with values < 3. The former will be tagged as 'accept' and the latter will be tagged as 'reject'. This is an example of discretization of continuous target variable. Using thus approach we have converted an estimation problem to a binary classification problem.

Clarification of Business Requirements: Relation to Data Mining Objectives and Technical Success Criteria

Of the various KDDM process models, only the CRISP-DM (2003) uses the term 'business requirements' in its discussion of execution of the KDDM process. While CRISP does not define the term 'business requirements', it suggests capturing requirements on comprehensibility, accuracy, deployability, maintainability and repeatability of DM project and resulting models as part of business requirements. No other details regarding how this list of requirements was generated, or how these could be collected are provided.

IKDDM considers capturing of business requirements as closely related to the business goals and business success criteria. Often all desired requirements in a solution may not be discussed at the time of determination of objectives. However they must be clarified before proceeding to next steps in the process through consultations with relevant business personnel. Specifically, the business users may wish to assess:

- Ease of use of the solution
- Ease of Deployability of the solution

These business requirements and how they could be assessed are explained below.

Ease of use: Business personnel may point out during requirements elicitation that solution must be easy to use. Given that the data mining solution is to be ultimately

162

used by human users, the acceptance of the solution is likely to depend on how easy it is to use by the average user.

Ease of deployability: Business personnel may also point out, during a requirements gathering exercise that the solution must be easy to deploy using existing hardware and software. This is also an important requirement, as data mining solutions can vary in scope and technical sophistication. Some may not be easy to deploy in a given firm and business users may lay down their preferences in form of business requirements about the deployability of the solution.

To enable objective evaluation of whether or not such requirements was met, business users may be asked to provide a desired Likert-scale rating of desired levels of deployability or ease of use. Suppose that they point out that deployability and ease of use should each be 4 or above (with 5 being the highest and 1 being the lowest). Then during the evaluation phase an assessment of the solution could be done to ensure that these requirements are indeed being satisfied.

Business Requirements Applicable for Predictive Data Mining

The types of business requirements to be collected also have a strong relationship with the data mining problem type (PURPOSE). Table 5-11 summarizes the business requirements for various supervised data mining problems, often referred to as predictive data mining problems. If the user selects any of the supervised data

163

mining problem types as the purpose, then his input about the following requirements can be elicited.

Table 5-11: Relevant Business Requirements for Supervised Data Mining Problems

Problem Type	Relevant Business Requirements
Classification And Prediction (Classification)	Nature of desired output from Model – Explanatory, Non Explanatory, Either?
	Desired improvement in accuracy
	Amount of Quantitative Improvement over old Model (assessed through LIFT)
	Level of simplicity (or tolerable level of complexity) of the model
	Generalization of results over different population than the one used for building the model (assessed through STABILITY)
Estimation And Prediction (Estimation)	Accuracy
	Improvement over old Model (assessed through LIFT; applicable only if the estimated values are divided into two classes, i.e. if the target variable is discretized)
	Level of simplicity (or tolerable level of complexity) of the model
	Generalization of results over different population than the one used for building the model (assessed through STABILITY)

To enable objective evaluation of such business requirements, they must be associated with relevant DM success criteria. This is because a requirement such as accuracy will need to be assessed through a measure such as the correct classification rate or using the confusion matrix, information for which is technical in nature. Likewise, information pertaining to lift, accuracy, and stability, is only available through the modeling results.

164

Note that the determination of data mining or technical success criteria is a separate task that has likely not been completed at the stage of identification of business requirements. Nevertheless identification of certain business requirements (because of their nature) leads us to early identification of certain data mining success criteria. In all likelihood, the formal setting up of data mining goals may result in identification of more data mining success criteria.

Both of these set of criteria (those identified directly through determination of data mining goals and indirectly through identification of certain business requirements) must be considered during the evaluation phase. Commercially available requirements elicitation tools can be used to aid the execution of this task.

Analysis of Inventory of Business Personnel and Other Business Resources

An assessment of inventory of business resources available to the proposed project must be performed before delving deeper into the DM project. This task ensures that the business personnel, key high level stakeholders, domain experts and other organizational actors who will be part of the project team are available for the duration of the project. An organization ontology (Sharma and Osei-Bryson 2008), organization charts, or organizational memory can act as tools in identifying the relevant personnel for Data Mining projects. Below we discuss how each of these tools could be used.

An ontology is the formal specification of concepts and entities belonging to a particular domain, and their interrelationships. An organization ontology models an

organization in form of an information system (Fox, Barbuceanu et al. 1998). Since an ontology formally specifies all relations, users could simple browse the ontology or pose simple queries to get an answer. For instance, if the high level stakeholders are interested in finding out what roles a particular agent P plays, they can use a query such as plays (P, ?r). Having discovered that the particular agent's role belongs to executing models in the decision science department, they may proceed to finding out if the agent requires permission to perform the above activity, or is empowered to perform it without explicit permission [See Fox et al. (1998) for examples of different types of queries). Such information may come handy in determining whether a particular agent could be immediately bought on board or if permission for his involvement in the project would need to be channeled through his supervisor. The task of assessment of business personnel and resources belongs to the first phase of the KDDM process, namely the business understanding or domain understanding phase. Sharma and Osei-Bryson (2008) propose an organization ontology based framework to execute all tasks contained within this phase, including identification of business personnel and resources. If an organization has an organization ontology available, then tasks such as identification of personnel with specific roles, skills, and competencies, as well as different types of resources (data, information, knowledge etc) is readily available.

Alternatively organizational charts can also be used for the purpose of identifying different types of personnel. While organizational charts are more frequently available than organization ontologies, they provide far less information. Browsing an

www.manaraa.com

organization chart typically requires knowing the name of the agent before his title/role could be known; organization ontology on the other hand allows stakeholders to simply look up available personnel by their role, or personnel with certain set of skills (say SAS Enterprise Miner or Angoss Knowledge Seeker experts).

An organizational memory if available can also be used to identify relevant personnel. It offers much more information than plain organizational charts although it may not be as easy to search or navigate for information as an organization ontology. An organizational memory can be described as the way organizations store knowledge from the past to support present activities (Stein 1995). Nevo and Wand (2005) apply the transactive memory model towards creation of organizational memories. They describe three types of knowledge: 1) Role knowledge—this is knowledge that is required by the definition of the knowledge retainer; 2) Instance knowledge—this is knowledge that is not required by the formal definition of the knowledge retainer's role, but that the individual has acquired through his or her experiences over a period of time; and 3) Transactive knowledge—this is the directory knowledge a retainer has about group members. They note that the availability of transactive knowledge enables retainers to effectively extend the knowledge available to them by being able to access their group members' knowledge. If an organization memory based on such as model was available, it could be used in the KDDM process for identification of business personnel having particular skills and knowledge.

The available tools for identifying business personnel can be stored in a Tool Repository. The information about the capabilities (such as documentation of individuals by role, skills of a particular type, project worked on etc) that the tool can support must also be included. High level stakeholders including project sponsors who are looking for relevant business personnel to staff a Data Mining Project can look up the information on the basis on certain criteria, leading to simplification of the staffing process. Table 5-12 shows the snapshot of data recorded in the Tool Repository.

Table 5-12: Selecting Tools to Assist with Business Personnel Identification - Snap-Shot of Tool Repository

Look Up Criteria	Organization Ontology	Organization Chart	Organization Memory
Individuals by Name	Yes	Yes	Yes
Teams by Name	Yes	No	No
Individuals by Role	Yes	Yes	Yes
Individuals by DM projects participated in	Yes	No	Yes
Individuals by Business Skills in Data Mining	Yes	No	Yes

Screen Shots To Assist User In Identifying Business Personnel

Task: Identification of Relevant Business Personnel

Identifying Personnel Resources

You wish to search for relevant individuals by [Select ALL that apply]

☐ Name

☐ Name of Team

☑ Role

☑ DM Projects Participated in

☑ Business Skills in Data Mining

| Save and Exit | | < Back | Next > | Cancel |

www.manaraa.com

Task: Identification of Relevant Business Personnel

Identifying Personnel Resources

Looking up Tool Repository ...

Please select from following tools

☐ Organization Chart
☑ Organization Ontology
☐ Organizational Memory

Launching Organization Ontology...

[Save and Exit]   [< Back] [Next >] [Cancel]
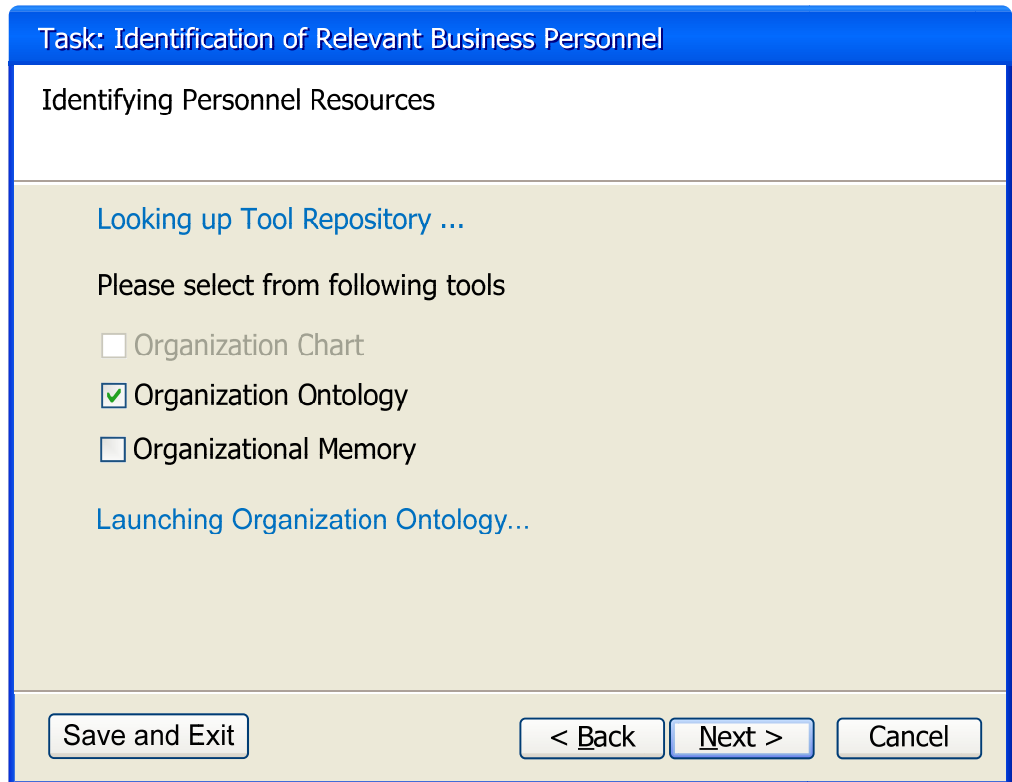
Clarification of Policy, Legal and Budgetary Constraints

Business constraints such as policy constraints, legal and budgetary constraints as well as availability of business personnel and business resources (described above) must be undertaken during this step (Figure 5-8). The policy constraints may reflect in the organization's business rules base. The potential solutions designed during the succeeding phases such as data preparation and modeling as well as tasks such as identification of necessary data that are performed during the BU phase must be in accordance with the business rules laid down by organization. For example, business rules may dictate that a predictive model may only use first three digits for zip code and

170

not all five. Rules can also be used to lay down conditions. For example, in determining if a credit card transaction is fraudulent, a rule may specify that point must be added to final score if a $1 transaction occurs at a gas station. Yet another example would be a rule that lays down that all marketing models directed at R rated movies must exclude people under 18 from the model. It is interesting to note that the rules stored in the business rules base may themselves have been generated through use of data mining modeling techniques such as decision trees.

Legal constraints may prohibit an organization from using certain variables in a certain manner and must be satisfied in the naming of solutions. For example, financial credit granting institutions are often prohibited from discriminating amongst applicants on the basis of their sex or nationality. Therefore, the company may be legally bound to exclude such variables from their decision making models. Severe penalty may be levied if it is found that a company has utilized such information in making its decisions.

Budgetary constraints are also an important type of business constraints and must present details about the funds available for the given project. Often the high level stakeholders including the project sponsor would decide on a budget for the project. The lower level stakeholders involved in the project then need to ensure that all expenses in form of resources (personnel, data, tools etc) can be satisfied within the confines of the allocated budget. The figure below shows partial view of the process model of the Business Understanding Phase.
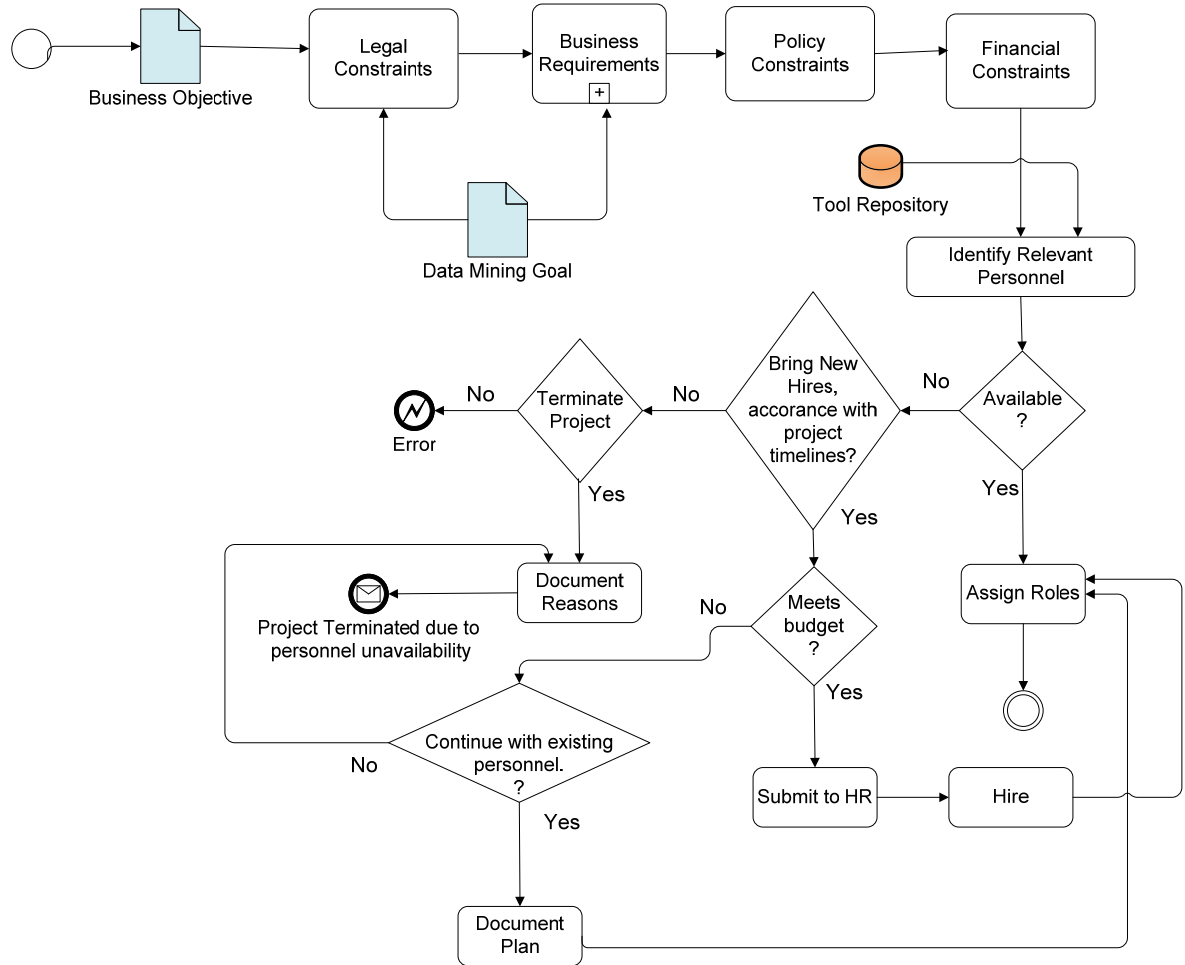
Figure 5-8: Clarification of Business Constraints and Setting up of Business Requirements: Partial View of Business Understanding Phase

Setting up of Data Mining Success Criteria

DEPENDENCY WITH TASK(S)	PHASE
Data Mining Goals	Business Understanding
Business Requirements	Business Understanding

Data Mining Success Criteria (DMSC) are used to evaluate the results of implementation of modeling techniques. These criteria must be set up before the implementation of modeling phase. We suggest using the GQM or goal question metric approach to move from data mining objectives to data mining success criteria. In this case the GQM approach can help translate the data mining objective into a set of questions which can then be refined into a set of objective or subjective metrics. These metrics are the evaluation criteria that can be used for assessing the results of the modeling phase to establish whether or not the selected model was helping accomplish the data mining objectives of the project. Data Mining success criteria influence the critical decision of whether or not a model should be deployed. Technical personnel in consultation with business users must be involved in setting up these criteria. We have used the GQM approach to shortlist data mining success criteria (measures) for various directed and undirected data mining problems.

Table 5-13 shows relevant evaluation criteria in the context of supervised or directed data mining. We present only classification and estimation as instances of directed data mining problems as prediction can be modeled as either of these problems

(Berry and Linoff 2001). The information about the data mining problem type has already been clarified during the task of formulation of objectives. Therefore, the user can be provided with guidance using the contents of the table below in setting up data mining success criteria. Note that of the various evaluation criteria, simplicity is not relevant if a non explanatory black box model such as neural network is used.

Table 5-13: Data Mining Success Criteria for Directed DM

Data Mining Problem Type	Data Mining Success Criteria
Classification	Accuracy, Profit and Loss, Lift, Simplicity*, Stability, Speed, Training Time and Memory Usage
Estimation	Mean Square Error, Variance (Standard Deviation), Simplicity*, Stability, Speed, Training Time and Memory Usage

* Simplicity is not relevant in case of Non Explanatory, Black Box Models

Table 5-14 shows relevant evaluation criteria in the context of unsupervised or undirected data mining problems. Note that in case of undirected data mining, particularly, description or visualization, the evaluation criteria are really a measure of the software tool used when addressing such tasks. The criteria presented here are discussed in (Redpath and Srinivasan 2003). The criteria associated with clustering and

association rules can however be used for evaluating the results from modeling techniques, similar to supervised data mining problems.

Table 5-14: Data Mining Success Criteria for Undirected DM

Data Mining Problem Type	Data Mining Success Criteria
Clustering	Normalized cluster means, Variable Importance Vectors, Ou Usefulness
Association Rules	Lift, Simplicity (Rule length), Support, Confidence, Recall, Precision, Interest Factor, Expected Monetary Factor, Incremental Monetary Factor
Description or Visualization	Number of instances in data set, Number of dimensions, Overlapping data instances, Ability to reveal patterns in dataset, Ability to reveal clusters of two or three dimensions, Number of clusters present, Amount of background noise, Variance of clusters, Ability to manipulate display automatically, Ease of Use

Setting up of Data Mining Success Criteria is also influenced by the Business Requirements. For instance, if the users expressed an interest in having a simple solution, then simplicity should be included as one of the data mining success criterion. The usefulness of the tables presented above is that it helps address any criteria that

may not have been uncovered during the setting up of business requirements. For instance, a business user at the time of setting up of requirements pertaining to accuracy may not be sure about the desired amount of accuracy of the new model, but the formal setting up of data mining success criteria using cross reference tables such as the ones above ensures that all important success criteria are set up before any analysis of results takes place.

In the section below we present definitions for the various criteria for supervised and unsupervised data mining problems.

Definitions of data mining success criteria for Supervised Data Mining Problems

1. **Accuracy**

Accuracy is an important criteria for both classification/prediction and estimation problems.

Accuracy for classification and estimation problems is measured in terms of the error rate, or the percentage of records classified incorrectly (Berry and Linoff 1997). In some domains, such as the world medical, false negatives and false positives may have entirely different implications. In some cases, a false negative may result in the patient not receiving treatment, whereas a false positive may cause him to unnecessarily undergo an invasive medical procedure. A confusion matrix or classification matrix sorts out false positives from false negatives. Different costs may be assigned to false

positives and false negatives, and models could be built to minimize the misclassification cost. Calculation of accuracy using confusion matrix is explained below.

Calculating accuracy using Confusion Matrix: the confusion matrix can be used for assessing the accuracy of classification models. It is calculated by applying the model to test data for which target values are already known. A confusion matrix is a square with n dimensions where n is the number of classes in the target data set. Therefore, a model where target variable has two classes will have a 2x2 confusion matrix, whereas a model where target variable has three values will have a 3x3 confusion matrix. An example of how accuracy can be calculated from a 2x2 confusion matrix is specified below. It shows results from a model used to classify applicants into good and bad customers.

		Actual	
		Good	Bad
Predicted	Good	200	15
	Bad	10	400

Table 5-15: Example Confusion Matrix

- The model made 600 correct predictions (200 + 400)

- The model made 25 incorrect predictions (10 + 15)

- The model scored total of 625 cases (600 + 25)

- The model error rate is 25/625 = 0.04

- The accuracy rate is 600/625 = 0.96 (it can also be calculated as 1-error rate or 1-0.04, also equal to 0.96)

Accuracy for estimation problems is expressed as the difference between the predicted score and the actual measured result (Berry and Linoff 1997). Accuracy of one estimate as well as accuracy of the entire model is of importance. A model that only provides good accuracy for a certain range of input values cannot be regarded as a good estimator. Generally, the 'average' is not used to assess the accuracy of an estimator because positive and negative values may cancel out each other. The variance (sum of squared differences), and the standard deviation (square root of the variance) are used in assessing the accuracy of estimators. Measures such as Mean Square Error and are also used. Sometimes R^2 is used to express the accuracy of an estimator. Really, R^2 represents the amount of variance in the model that is explained by the predictors and not the accuracy of the estimate. In other words, it is a measure of the goodness of the model, with a model with higher R^2 being regarded as better than one with lower R^2. When R^2 is used to assess to assess variability of estimation errors with variability of original values, then following formula should be employed:

$R^2 = 1 - SS_E / SS_T$

Where SS_E is the error sum of squares

And SS_T is the total sum of squares

Sometimes R^2 is expressed as the ratio of SS_R/SS_T (Field 2000), where SS_R is the residual sum of squares and SS_T is the total sum of squares. SS_R is the difference between SS_T and SSE.

Therefore,

$$R^2 = SS_T - SSE / SS_T = 1 - SSE / SS_T$$

2. **Simplicity**

Simplicity is an important evaluation criterion in data mining models. In simple terms, it highlights the preference for simple versus overly complex models, which are often known to be unstable and difficult to implement. It is estimated differently in different approaches. Below we discuss how simplicity could be estimated in regression, tree based models and techniques where simplicity is not applicable.

Simplicity in Regression Models: Simplicity in regression models can be estimated using the number of predictors and the number of interactions involved in the model. It is known that R^2 of a model may increase by simply adding new variables, creating an impression that a model with a higher R^2 is better than a model with lower R^2. The adjusted R^2 statistic on the other hand, increases only if the addition of a new term

179

improves the model, more than what would be expected by only chance (Draper and Smith 1998).

In other words, the Adjusted R^2 statistic adjusts for the degree of complexity of the model and penalizes an unnecessarily complex model. It can be calculated using following formula:

Adjusted R^2 = 1- SS_E/df_{error} /1- SS_T/df_{total}

Simplicity in Tree Based Models: Simplicity has been used as an evaluation criterion for tree based models. It can be calculated as the number of leaves or the length of the rule. The former is based on the assumption that the fewer the leaves the better the model. (Osei-Bryson 2004) points out that while a simple tree with fewer leaves is desirable, a tree with only one or two leaves may not be useful.

3. Lift

Lift is a measure of the performance of the model at segmenting the population. Lift measures the change in concentration of a particular class when the model is used to select a group from the general population. It can be calculated as follows:

Lift $_{subset\ of\ the\ population}$ = Predicted response rate for the subset/Predicted response rate for the population

For example, suppose that the population has a response rate of 2%, but a model has identified a subset with a predicted response rate of 20%, then the lift is 10. In developing response models in marketing it is common to divide the population into ten deciles and rank the deciles by lift. By comparing the profits (based on the predicted response), and the cost (of mailing out the offer), the firm can decide whether or not it will be profitable to mail an offer to a given decile.

4. **Sensitivity and Specificity**

Both of these measures are applicable to classification problems involving binary target variables.

Sensitivity (or Recall rate) measures the proportion of actual positives which are correctly identified. It is calculated as the ratio of true positives to true positives and false negatives

Specificity measures the proportion of negatives that are correctly identified. It is calculated as the ratio of true negatives to false positive and true negatives.

Let us use the confusion matrix described in Table 23 above to calculate values for these parameters. Let us assume that we are interested in predicting which customers are good customers from a pool of good and bad customers. Note that our target variable Y is a dichotomous response $(Y = 1, 0)$. The value for Y is based on a cut off c, where $0 \leq c \leq 1$

A decision rule may be created as follows. If $\Pi_i > c$, then $Y_i = 1$ (good)

$$\Pi_i \leq c, \text{ then } Y_i = 0 \text{ (bad)}$$

		$Y_i = 1$	$Y_i = 0$
		Good	Bad
$Y_i = 1$	Good	200 (A)	15 (B)
$Y_i = 0$	Bad	10 (C)	400 (D)

Table 5-16: Example Confusion Matrix

Sensitivity (True Positive Rate) = A/ (A+B)

$$= 200/ (200+15)$$

$$= 200/215$$

$$= 0.930$$

Specificity (True Negative Rate) = D/ (C+D)

$$= 400/ (10+400)$$

$$= 400/410$$

$$= 0.975$$

5. Precision and Recall

Precision and Recall are widely used measures in information retrieval and statistical classification. Precision is seen as a measure of exactness whereas Recall is seen as a measure of completeness. In a statistical classification task, precision for a class is the number of true positives divided by the sum of true positives and false negatives. Recall in this context is defined as the number of true positives divided by the sum of true of positives and false negatives.

Usually precision and Recall are not measured in isolation. Instead either values for one measure are compared for a fixed level at the other measure (for e.g., Precision at a Recall Level of 0.80), are combined into one measure such as the F measure defined below.

F_1 measure = 2 (Precision x Recall)/(Precision + Recall)

This represents the case when Precision is weighted as equal to Recall and is a specialized case of,

$F_{\beta} = (1+ \beta^2)$ (Precision x Recall)/(β^2.Precision + Recall)

Precision = A/(A+C) = 200/(200+10) = 200/210 = 0.952

Recall (Sensitivity, True Positive Rate) = A/ (A+B) = 200/ (200+15) = 200/215 = 0.93

6. KS statistic

KS statistic is a popular measure used by financial services firms and measures the maximum vertical difference between two population distributions. It is relevant to classification type problems where target variable is discrete in nature. In building credit scorecards, a firm often has interest in separate its good population (consisting of non-defaulters) from its bad population (consisting of defaulters).

- If a model can partition the population into two separate groups in which one group contains all the defaulted accounts and the other all the non-defaulted accounts, then the K-S is 100. In such a case, there would be no overlap between the curves for the populations and they would lay side by side.

- If the model can not differentiate between non-defaulted and defaulted accounts, then it is as if the model selects individuals randomly from the population. There would be no difference in the location of the distributions; they would lie on top of each other, leading to a K-S of 0.

- Generally, the K-S value will fall between 0 and 100, and the higher the value the better the model is at separating the non-defaulted from defaulted accounts.

The KS statistic is calculated as the maximum difference between the cumulative percent good distribution (non-defaulters) and cumulative percent bad distribution (defaulters).

Cumulative Percent Distribution for Good and Bad Loans
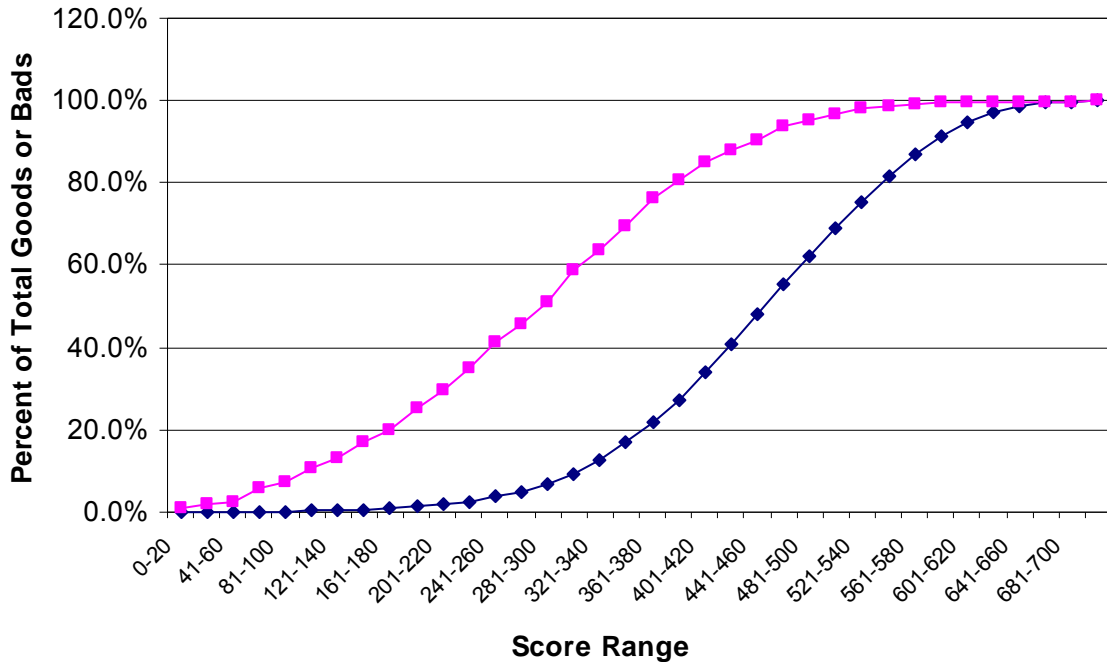
Figure 5-9: KS statistic

Mays (2001) points out that it is not judicious to rely on only KS to make any judgments as the statistic does not tell us about the ranking ability of a scorecard. Typically we expect the scorecard to show a higher bad rate for lower scores and a decreasing bad rate as scores increase. This is based on the simple logic that people with lower credit scores perform worse (have a high probability of defaulting or charging off) than people with higher credit scores. If however, our scorecard is not ranking the customers well, it may result in a lower bad rate at lower scores. Therefore it is recommended that the KS Statistic be only assessed after studying the distribution of

185

goods and bads and only when the distribution appears as expected should decisions regarding setting of cut offs be made.

7. ROC curve and AUC (Area Under Curve)

ROC curves were originally introduced in signal detection theory (Egan 1975) and are now also being utilized in data mining applications. In DM applications, dealing with classification models, the ranking quality of a classifier is considered very important. The criterion widely used to measure the ranking quality of a classification algorithm is the area under an ROC curve (AUC).

The receiver operating characteristic (ROC) chart displays the sensitivity and 1-specificity of a classifier for a range of cutoffs. Sensitivity is a measure of accuracy for predicting events. It is equal to the true positive / total actual positive or the proportion of event observations that the model predicts to be events for a given probability cutoff point.

Specificity is a measure of accuracy for predicting nonevents. It is equal to the (true negative / total actual negative) or the proportion of nonevent observations that the model predicts to be nonevents for a given probability cutpoint. One minus specificity is simply the number of false positives (the number of nonevent observations that the model incorrectly predicts as events for a given probability cutoff point) divided by the number of nonevents.

Each point on the curve represents a cutoff probability. The cutoff choice represents a trade-off between sensitivity and specificity. Ideally we would like to have high values for both sensitivity and specificity, so that the model can accurately predict both events and nonevents. A lower cutoff typically gives more false positives. A high cutoff gives more false negatives, a low sensitivity, and a high specificity.

For a fully random classification, the ROC curve is a straight line connecting the origin to (1, 1). Any improvement over random classification results in an ROC curve at least partially above this straight line. Cortes and Mohri (2004) highlight that for ROC curves to be useful, we need to measure and report the AUC properly. They suggest determining an interval of confidence for its value. The AUC is defined as the area under the ROC curve.

The performance quality of a model is demonstrated by the degree the ROC curve pushes upward and to the left. This can be quantified by the area under the curve. The area will range from 0.50, for a poor model, to 1.00, for a perfect classifier. For models with a high predictive accuracy, the ROC curve would rise quickly (sensitivity increases rapidly, specificity stays at 1). Therefore, the area under the curve is closer to 1 for a model with high predictive accuracy. Conversely, the ROC curve rises slowly and has a smaller area under the curve for models with low predictive accuracy. A ROC curve that rises at 45 degrees is a poor model. It represents a random allocation of cases to the classes and should be considered a baseline model.

187

Data Mining Success Criteria for Unsupervised Data Mining Problems

Data Mining Success Criteria for Association Rules

1. **Confidence**

Confidence is the ratio of the number of the transactions supporting the rule to the number of transactions where the conditional part of the rule holds (Berry and Linoff, 1997). In other words, confidence is the ratio of the number of transactions with all the items to the number of transactions with just the "if" items. Consider the following rule: "if B and C then A". Assume that confidence is 0.20. This means that when B and C appear in a transaction, there is a 20% percent chance that A also appears in it. That is, one times out of five, A occurs with B and C, and other four times, B and C appear without A. The most confident rule is the best rule.

2. **Lift**

Lift helps assess the improvement we can expect by using the rule rather than relying on chance. In other words, lift tells us how much better a rule is at predicting the result than just assuming the result in the first place.

Lift is the ratio of the density of the target after application of the left hand side to the density of the target in the population, or the ratio of the records that support the entire rule to the number that would be expected, assuming that there is no relationship between the products. Lift is a good measure of how much better the rule is doing.

188

Since It is the ratio of the density of the target (using the left hand side of the rule) to density of the target overall, the formula is:

Lift = (p (condition and result) / p (condition)) / p (result)

$\qquad$ = p (condition and result) / (p (condition) p (result))

When lift is greater than 1, then the resulting rule is better at predicting the result than guessing whether the resultant item is present based on item frequencies in the data. When lift is less than 1, the rule is doing worse than informed guessing.

3. Excess

Excess is a measure similar to lift and is defined as the difference between the number of records supported by the entire rule minus the expected value (Berry and Linoff, 1997).

4. Support (Pruning)

Pruning is a technique for reducing the number of items and combinations of items being considered at each step. AT each stage, the algorithm throws out a certain number of combinations that do not meet some threshold criterion. The most common pruning threshold is called minimum support pruning. Support refers to the number of transactions in the database where the rule holds. Minimum support pruning requires that a rule hold on a minimum number of transactions. For instance, if there are one

million transactions and the minimum support is 1 percent, then only rules supported by 10,000 transactions are of interest.

Choi et al. suggest three 1st level data mining success criteria for association rules: recency, frequency and monetary value. They propose measuring these through a set of 2nd level criteria. For example, they define recency as the time trend of a rule between time intervals in this study; a higher value implies a higher worth of attention to a rule. This factor can be measured with the attribute of the degree of change in support, the definition for which is included below.

5. Degree of change (DoC)

Even though most of data mining techniques usually give attention to the rules which have a large frequency of occurrence and ignore time trend, the rules with a large growth rate or decreasing rate in occurrence may give significant implications to managers in changing business environment in spite of their relatively small occurrence.

Another first level criteria discussed by Choi et al. is frequency and is defined as statistical significance of a rule in a time interval in this study; with higher frequency indicates greater statistical significance of a rule. They suggest that this factor can be measured through support, confidence, and interest factor. Definitions for support and confidence have been included earlier. The definition for interest factor is provided below.

6. **Interest factor (IF).** Interest factor is another widely used measure for association patterns (Brin et al., 1997). This metric is defined to be the ratio between the joint probability of two variables with respect to their expected probabilities under the independence assumption. The interest factor is a non-negative real number with a value of 1 corresponding to statistical independence.

The third 1st level criteria discussed by Choi et al. is monetary value and is defined as the profitability of a rule in this study; a higher value indicates that the company should focus more on that rule. This factor can be measured with two attributes, expected monetary value and incremental monetary value, the definitions for which are included below.

7. **Expected monetary value (EMV).** If we assume mutual independence between products, then the expected profit (expected monetary value) after buying a product X is equal to the probability of buying Y given X, multiplied by the profit of Y (Kitts et al., 2000).

8. **Incremental monetary value (IMV).** The idea behind incremental profit of Kitts et al. (2000) is the expected profit minus the profit you would expect to receive due to the natural course of a customer's purchasing. Incremental profit maximizes the profit of the item, minus the baseline profit associated with the item (Kitts et al., 2000).

Data Mining Success Criteria for Description/Visualization techniques

191

As stated earlier, data mining success criteria for description/visualization actually apply to the software used to execute this particular class of data mining problems. This is understandable since unlike other data mining problems types, description/visualization are not a type of modeling technique (or algorithm), but rather just a way of visualizing the relationships in the data. All the data mining success criteria presented below have been discussed by Redpath and Srinivasan

1. **Number of instances in data set**:

It is important to know whether the visualization techniques deal with data sets of different sizes. Data sets may range in size from a few hundred instances to many millions of instances. Not all the techniques will successfully deal with large numbers of instances. The concern here is not the capacity of the computer hard-ware used. Rather the visualization technique may be unable to effectively display and distinguish large numbers of instances. The capability of the visualization techniques to deal with large numbers of instances without overlap and the possible loss of information must therefore be considered.

2. **Support For Multi- Dimensional Data**:

Some of the visualization techniques are able to display many dimensions in a single display and others have an upper limit of two, three or four dimensions. Simple scatter plots can display only two or three dimensions. If the point plotted has other distinguishing features such as color or is an icon, which relates to further dimensions

192

through some aspect of its shape, a greater number of dimensions can be represented in a single display. Other techniques use multiple windows to display a large number of dimensions or a number of straight line axes as in the case of parallel co-ordinates.

3. Ability to reveal patterns in dataset:

The purpose of the visualization tools is primarily to gain knowledge through the recognition of patterns in the data. The technique must be able to reveal patterns inthe data set if they are present. If the visualization is unable to do this it has failed in its basic purpose. It would be desirable to be able to distinguish between different types of pattern. The criteria following consider more particular aspects of the ability to reveal patterns in the data set.

4. Ability to reveal clusters of two or three dimensions:

Clusters indicate the presence of relationships between attributes. They may be indicative of associations or classes also. For the visualization technique to be useful it is expected that as a minimum requirement two and three-dimensional clusters would be revealed.

5. Number of clusters present:

Most patterns manifest as clusters in the visualizations. The visualization techniques must be able to distinguish between clusters if a number of clusters are present. We are

concerned as to whether the clusters obscure each other or are clearly revealed as separate clusters.

6. Amount of background noise:

Another important consideration is how the visualization technique performs against a background of noise instances. Real data will usually have many instances that do not contribute to any pattern. If presence of background noise, as such instances are termed, obscures what patterns are present the visualization technique is less useful. It is necessary to test the visualization techniques against various levels of background noise to determine the usefulness in the presence of such noise.

7. Variance of clusters:

The instances that contribute to a cluster may be tightly packed or spread out in the region of space where the cluster or pattern appears. Given that there is usually some background noise clusters that are spread out will be more difficult to detect. It would be interesting to know if some visualization techniques are better than others at dealing with clusters that are more spread out.

8. Ability to manipulate display automatically

Ease of Use: The ease of use of the display or visualization technique relates to a combination of factors. These factors include the flexibility of data set format that can be imported. It also relates to how efficiently the data is displayed. If significant delays

exist in changing the display the user will have difficulty using the visualization techniques. If the design of the controls is awkward, not obvious, or fails to follow common conventions, the tool will not be easy to use.

Initial Assessment of Modeling Techniques

DEPENDENCY WITH TASK(S)	PHASE
Data Mining Goals	Business Understanding
Business Requirements	Business Understanding

The data mining problem type and target variable specified during the formulation of data mining objective, business requirement (related to whether or not an explanatory model is desired) can be used for generating an array of modeling techniques applicable in the context of the data mining project. Table 5-17 describes the applicable modeling techniques associated with various directed data mining problem types (such as classification, prediction and estimation) based on target variable type (binary, ordinal, nominal and interval). If there is no particular business requirement for an explanatory model, then any of the modeling techniques mentioned in the table could be applicable. However, if the requirement is for an explanatory model, then the black box techniques such as neural networks cannot be employed.

On further analysis, it was found that using the above method of identifying applicable techniques does not take into consideration the situation when a combination

195

of techniques (say explanatory and non explanatory) could be used for generating a model than outperforms the individual explanatory or non explanatory model.

For instance, it may be better to generate the lost of applicable techniques using the data mining problem type and the target variable and not impose the business requirement until all the models explanatory and non explanatory have been tried. Next if the non explanatory model's performance exceeds that of the explanatory model (more accurate, stable, etc), then using a combination of techniques could be considered. For instance, the output from a technique such as Neural Network could be explained using an explanatory technique such as a decision tree (Medina and Pratt 1995), or it could be explained using logistic regression (Wong, P.J.Fos et al. 2003). The decision tree could then be run over the unseen test data and its performance assessed. See cross reference matrix (Table 5-18) for an example of applicable modeling techniques generated for classification problems with a binary target variable.

Review of published case studies reveals that combination of models during modeling phase is not always pursued. Combining predictive models can lead to improvement in predictive accuracy (Berry and Linoff 2000). The simple principle underlying model combinations is that a predictive model can take a set of inputs and produce one or more outputs. These outputs in turn can be used as an input for another predictive model(s). The combination of models must however proceed with caution. Berry and Linoff (2000) describe four ways of combining models and the rationale for these approaches:

Multiple Model Voting: In multiple model voting, the individual predictions made by different models are compared. The model results are then combined to form what is called an "ensemble model". Multiple model voting allows us to have more confidence in the results. Such voting can be employed for combining several models of the same type (say decision trees) produced by varying parameters, or for combining results from models of different types such as decision trees, neural networks, and logistic regression. In a simple form of voting, a majority vote of the results (for categorical target variables) or average vote of results (for numeric targets) are considered. The various predictions can also be combined by using the statistics for predicted values and predicted confidence.

Segmented Input Combination Models: In this approach different models are built for different classes of input in the model. The difference between this approach and the previous approach is that in multiple model voting all models were applied to the complete set of input. Given that segmentation often results in smaller data sets, an effort must be made to avoid risk of over fitting by assessing appropriate parameters (minimum leaf size in decision trees or number of hidden nodes in neural networks). Segmentation can be mainly useful in two types of situations: (1) when data is available only for some records but not all of them; and (2) when the different segments are inherently different from each other (e.g., different types of customers) and warrant creation of different models for each segment.

Modeled segmentation: This approach is a variant of the segmented input combination approach described earlier. In the former approach, the segments (for which different models are built) are known in advance. However, in modeled segmentation, a model is first used to segment the input and then another model is used to build a model on the so identified segment.

Error fixing combination models: This approach also referred to as "boosting" cascades models based in their confidence. In the event that the prediction from a model has a low level of confidence, a different model is used to determine the outcome. Note that in this approach the second model (or the error fixing models) is trained using the rejects from the first models, where rejects are identified through the low level of confidence.

Data Enhancement combination models: In this approach a model is used to add new features to the input (say adding a cluster field or a propensity score such as the FICO score) or for replacing missing values.

The enumeration of techniques using the proposed approach indicates that this task of generating a list of applicable techniques can be semi-automated. The proposed approach utilizes the business requirement, data mining problem type and the target variable type (metadata) to generate the list of applicable techniques. This approach is different from that proposed by Bernstein et al. (2005) who start at the level of the data itself and propose that the data type can be used for making decisions about the

applicable techniques. Use of their approach can result in enumeration of those techniques that clash with the business requirement. So even if these techniques were tried, the results would not eventually be accepted, resulting in inefficient usage of resources. Also their approach results only in enumeration of single techniques and combination of techniques is not accommodated in their approach.

Table 5-17: Applicable Modeling Techniques for Various DM problem Types

Problem Type Target variable	Prediction	
	Classification	Estimation
binary	Logistic regression Classification Tree k-nearest neighbor Naïve Bayes* Neural network* Support Vector Machines* Genetic algorithm*	Not applicable
ordinal	Ordinal Logistic regression Classification Tree k-nearest neighbor Naïve Bayes* Neural network* Support Vector Machines* Genetic algorithm*	Not applicable
nominal	Multinomial Logistic regression Classification Tree k-nearest neighbor Naïve Bayes* Neural network* Support Vector Machines* Genetic algorithm*	Not applicable
Interval	Not Applicable	Regression Regression Tree k-nearest neighbor Memory Based Reasoning Neural Networks*

* This modeling technique is not applicable if Business Requirement is for an explanatory model

Table 5-18: Ensemble Modeling Techniques for Classifications Problems with

Binary Target Variable

Model Input	Model Output
Neural Network	Logistic Regression, Classification Tree, K-Nearest Neighbor, Memory Based Reasoning
Support Vector Machines	Logistic Regression, Classification Tree, K-Nearest Neighbor, Memory Based Reasoning
Genetic algorithm	Logistic Regression, Classification Tree, K-Nearest Neighbor, Memory Based Reasoning

201

Assessing Selected Modeling Techniques against Selected Data Mining Success Criteria

It is important to note that the data mining success criteria also have an important relationship with the data mining techniques or algorithms (such as decision tree, neural networks etc). This is so because output of different techniques can be assessed using different parameters. These different data mining techniques (decision trees, neural networks) may both belong to a common problem type such as classification, but may still need to be evaluated using a slightly different combination of data mining success criteria. For instance a criteria such as simplicity which is useful in evaluating the performance of a classification data mining technique such as decision tree, does not apply to another classification data mining technique such as neural networks. The user can be presented with the cross reference Table 5-19 and Table 5-20 to assess which criteria are applicable for which data mining techniques.

The purpose of this table is to remind the user that it may not be possible to calculate a particular data mining success criterion, if a certain technique is used. This has a direct effect on the calculation of value functions for data mining success criteria (a separate task) and will be explained under the relevant section.

202

Table 5-19: Summary Tables: Data Mining Success Criteria for Classification

Modeling Techniques

		Classification Modeling Techniques			
		Classification Tree	Logistic Regression	Naïve Bayes'	Neural Network
Data Mining Success Criteria	Accuracy (Misclassification Rate)	✓	✓	✓	✓
	Lift	✓	✓	✓	✓
	Precision	✓	✓	✓	✓
	Recall	✓	✓	✓	✓
	Simplicity	✓	✓	✓	×
	Stability	✓	✓	✓	✓
	Sensitivity	✓	✓	✓	✓
	Specificity	✓	✓	✓	✓
	ROC curve	✓	✓	✓	✓
	Area Under ROC curve	✓	✓	✓	✓
	KS Statistic	✓	✓	✓	✓
	Profit/Loss	✓	✓	✓	✓

Table 5-20: Summary Table: Data Mining Success Criteria for Regression Modeling Technique

		Estimation Techniques		
		Regression Tree	Linear Regression	Neural Network
Data Mining Success Criteria	Accuracy (Average Squared Error)	✓	✓	✓
	Simplicity	✓	✓	×
	Stability	✓	✓	✓
	Profit/Loss	✓	✓	✓

Analysis of Applicable Software Tools for Addressing the Data Mining Project

During this task the lead technical personnel must analyze the availability of technical resources in form of software tools for implementing the chosen data mining problem type (and the modeling techniques in case of supervised data mining problems). Analysis of tools can be simplified by storing the various modeling techniques supported by all the data mining tools (such as SAS Enterprise Miner, SPSS Clementine etc) available to the organization. If no available tools support the selected problem type then the relevant actors may propose sourcing of a relevant tool to the project sponsor or other key high level stakeholder who can then make the decision about whether or

not the budget would support the purchase of a new tool and ensuing training and implementation costs.

Analysis of Available Software Tools to Support Selected Data Mining Success Criteria

It is also pertinent to note that there is a relationship between the data mining success criteria that can be used for evaluating a particular data mining technique and the software tool used for implementing the particular technique. Some tools may provide output that yields the data mining success criteria (such as lift, accuracy etc) explicitly, others may only yield these criteria implicitly or indirectly with the user being responsible for calculating the exact values, still others may not provide the criteria at all (not even implicitly).

This means that there is a relationship between data mining techniques (e.g., decision trees, naïve bayes), data mining tools (e.g., SAS Enterprise Miner, SPSS Clementine) and the data mining success criteria (e.g., accuracy, Area Under ROC curve). Clearly we need to have detailed support towards all three aspects when dealing with a data mining project. With this goal in mind, IKDDM offers tabular summaries of different data mining techniques, success criteria that can be used to evaluate results from these techniques, software tools that can be used for implementing these techniques and whether or not or how the tools allow for the criteria to be calculated. Summary tables (Table 5-21, Table 5-22, Table 5-23, Table 5-24, Table 5-25, Table

205

5-26   and

) based on Data Mining Success Criteria for various modeling techniques along with details about example software tools are included below.

Table 5-21: Data Mining Success Criteria for Classification Trees

Measure	Source for Calculating Measure	SAS EM 4.3	SPSS Clementine 12.0
Accuracy	Test Misclassification Rate	Implicit Calculate using 1-Test Misclassification Rate	Explicit (Modeling results)
	Confusion Matrix	Implicit	Implicit
Lift or Gains Index	Visual Inspection of Lift Chart up to a particular Decile	Explicit-Visual	Explicit-Visual
	Lift Value can be estimated through analysis of lift chart	Implicit Calculate using Tree/Exact	Explicit (Modeling results)
Profit and Loss	Profit and Loss Matrix	Explicit (Modeling Results)	Explicit (also provides additional measures)
Simplicity	User Defined	Implicit (Calculate using Number of leaves, and/or Minimum Rule length)	Implicit (Calculate using Number of leaves)
Stability	User Defined	Implicit Calculate using a coarse measure such as Min $[ACCT_V/ACC_T, ACC_T/ACC_V]$ Where $ACCT_V$ is accuracy of validation data and ACCT is accuracy on training data	Implicit Models (by default) are built with generality. For assessing stability, validate against hold out sample
	Visual Inspection of Lift Chart at a particular decile	Explicit-Visual	Explicit-Visual
ROC curve	Plot of 1-specificity on x-axis and sensitivity on y axis.	Explicit-Visual Visual inspection of chart must be used to employ ROC as an evaluation measure	Explicit-Visual
Area under ROC Curve or AUC	Area calculated using trapezoidal rule	No	Explicit (Empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve and its 95% CI)
KS statistic (Komogorov-Smirnov)	Maximum KS value	No	No
Average Squared Error	Modeling Results	Explicit	No
Sensitivity	Confusion Matrix	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)
Specificity	Confusion Matrix	Implicit Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate	Implicit Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate

208

Table 5-22: Data Mining Success Criteria for Neural Networks

Measure	Source for Calculating Measure	SAS EM	SPSS Clementine
Accuracy	Test Misclassification Rate	Implicit Calculate using 1-Test Misclassification Rate	Explicit (Modeling Results)
	Confusion Matrix	Implicit	Implicit
Lift or Gains Index	Visual Inspection of Lift Chart up to a particular Decile	Explicit-Visual	Explicit-Visual
	Lift Value can be estimated through analysis of lift chart	Implicit Calculate using Tree/Exact	Explicit (Modeling Results)
	Confusion Matrix	Implicit Calculate using	Implicit Calculate using
Profit and Loss	Profit and Loss Matrix	Explicit (Modeling Results)	Explicit (Modeling Results)
Stability	User Defined	Implicit Calculate using a coarse measure such as Min $[ACCT_V/ACC_T, ACC_T/ACC_V]$ Where $ACCT_V$ is accuracy of validation data and ACCT is accuracy on training data	Implicit Models (by default) are built with generality. For assessing stability, validate against hold out sample
	Visual Inspection of Lift Chart at a particular decile	Explicit-Visual	Explicit-Visual
ROC curve	Plot of 1-specificity on x-axis and sensitivity on y axis.	Explicit-Visual	Explicit-Visual
Area under ROC Curve or AUC	Area calculated using trapezoidal rule or the statistic c in the "Association of Predicted Probabilities and Observed Responses" table. The value of the statistic is the area under the curve.	No	Explicit (Empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve and its 95% CI)
KS statistic (Komogorov-Smirnov)	Maximum KS value	No	No
Average Squared Error	Difference between predicted values and actual values	Explicit (Modeling results)	No
Sensitivity	Confusion Matrix	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)
Specificity	Confusion Matrix	Implicit (Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate)	Implicit (Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate)

Table 5-23: Data Mining Success Criteria for Naive Bayes

Measure	Source for calculating measure	Clementine
Accuracy	Conditional probabilities	Explicit (modeling results) Probabilities relate predicted classes (columns) and predictor-variable-value combinations (rows)
	Confusion Matrix	Implicit
Lift or Gains Index	Visual Inspection of Lift Chart up to a particular Decile	Explicit-Visual
	Lift Value can be estimated through analysis of lift chart	Explicit (Modeling Results)
	Confusion Matrix	Implicit
Profit and Loss	Modeling Results	Explicit (Modeling Results)
Simplicity	Implicit (Calculate using Minimum Description Length)	Explicit (Modeling Results of Adaptive Bayes Network)
Stability	User Defined	Implicit Models (by default) are built with generality. For assessing stability, validate against hold out sample
	Visual Inspection of Lift Chart at a particular decile	Explicit-Visual
ROC curve	Plot of 1-specificity on x-axis and sensitivity on y axis.	Explicit-Visual
Area under ROC Curve or AUC	Area calculated using trapezoidal rule	Explicit (Empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve and its 95% CI)
KS statistic (Komogorov-Smirnov)	Maximum KS value	No
Average Squared Error	Difference between predicted values and actual values	No
Sensitivity	Confusion Matrix	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)
Specificity	Confusion Matrix	Implicit (Calculate using TN/[FP+TN]

		Where TP is the true positive rate and FN is the false negative rate)

Table 5-24: Data Mining Success Criteria for Logistic Regression

Measure	Source for calculating measure	SAS EM 4.3	Clementine 12.0
Accuracy	Test Misclassification Rate	Implicit Calculate using 1-Test Misclassification Rate	Explicit (Modeling Results)
	Confusion Matrix	Implicit	Implicit
Lift or Gains Index	Visual Inspection of Lift Chart up to a particular Decile	Explicit-Visual	Explicit-Visual
	Lift Value can be estimated through analysis of lift chart	Implicit Calculate using Tree/Exact	Explicit (Modeling Results)
Profit and Loss	Modeling Results	Explicit (Modeling Results)	Explicit (Modeling Results)
Stability	User Defined	Implicit Calculate using a coarse measure such as Min $[ACCT_V/ACC_T, ACC_T/ACC_V]$ Where $ACCT_V$ is accuracy of validation data and ACCT is accuracy on training data	Implicit Models (by default) are built with generality. For assessing stability, validate against hold out sample
	Visual Inspection of Lift Chart at a particular decile	Explicit-Visual	Explicit-Visual
ROC curve	Plot of 1-specificity on x-axis and sensitivity on y axis.	Explicit-Visual	Explicit-Visual
Area under ROC Curve or AUC	Area calculated using trapezoidal rule	No	Explicit (Empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve and its 95% CI)
KS statistic (Komogorov-Smirnov)	Maximum KS value	No	No
Average Squared Error	Modeling Results	Explicit (Modeling results)	No
Sensitivity	Confusion Matrix	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)
Specificity	Confusion Matrix	Implicit (Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate)	Implicit (Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate)

Table 5-25: Data Mining Success Criteria for Linear Regression

Measure	Source for calculating measure	SAS EM 4.3	SPSS Clementine 12.0
Explainability of model	R^2	Explicit (Modeling Results)	Explicit (Modeling Results)
	Adjusted R^2	Implicit (Calculated using Adjusted R^2	Explicit (Modeling results)
Profit and Loss	Profit and Loss Matrix	Explicit (Modeling Results)	Explicit (Modeling Results)
Simplicity	Number of variables	Implicit (Calculate using number of variables, interaction effects, adjusted $R^{2\ or}$ Schwarz Bayesian Criterion)	Implicit (Calculate using number of variables, interaction effects, adjusted R^2)
Stability	User Defined	Implicit (Assess using predictor equations – beta values are different from sample to sample indicate instability)	Implicit (Assess using predictor equations – beta values are different from sample to sample indicate instability)
	DFBeta	No	Implicit (Assess using DFBeta to check if one or more cases are biasing regressions results in any way)
Area under ROC Curve or AUC	Area calculated using trapezoidal rule	No	Explicit (Empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve and its 95% CI)
Multicollenearity	Tolerance and VIF (Variable inflation factor)	No	Explicit (Modeling Results)
KS statistic (Komogorov-Smirnov)	Maximum KS value	No	No
Average Squared Error	Modeling Results	Explicit (Modeling Results)	No

213

Table 5-26: Data Mining Success Criteria for Association Rules

Measure	Source for Calculating Measure	SAS EM 4.3	SPSS Clementine 12.0
Lift	Ratio of confidence to the prior probability of having the consequent	Explicit (Modeling results)	Explicit (Modeling results)
Excess	Lift-1	Implicit Calculate using lift-1	Implicit Calculate using lift-1
Simplicity	Length of Rule	Implicit Calculate using length of rule	Implicit Calculate using length of rule
Support	Proportion of ID's for which entire rule, antecedents, consequents are true	Explicit (Modeling results)	Explicit (Modeling results)
Confidence	Ratio of rule support to antecedent support	Explicit (Modeling results)	Explicit (Modeling results)
Interest Factor	ratio between the joint probability of two variables with respect to their expected probabilities under the independence assumption	No	No
Monetary Value	Profitability of a rule	Explicit (Modeling Results)	Explicit (Modeling Results)
Deployability	% of training data that satisfies the conditions of the antecedent but does not satisfy the consequent	No	Explicit (Modeling Results)

Table 5-27: Data Mining Success Criteria for Regression Trees

Measure	Source for Calculating Measure	SAS EM 4.3	Clementine SPSS 12.0
Accuracy	Average Squared Error	Explicit (Modeling Results)	No
Profit and Loss	Profit and Loss Matrix	Explicit (Modeling Results)	Explicit (Modeling Results)
Lift	Visual Inspection of Lift Chart up to a particular Decile	Explicit-Visual	Explicit-Visual
	Lift Value can be estimated through analysis of lift chart	Implicit (Calculating using Tree/Exact)	Explicit (Modeling Results)
Stability	User Defined	Implicit Calculate using a coarse measure such as Min [$ACCT_V/ACC_T$, ACC_T/ACC_V] Where $ACCT_V$ is accuracy of validation data and ACCT is accuracy on training data	Implicit Models (by default) are built with generality. For assessing stability, validate against hold out sample
	Visual Inspection of Lift Chart at a particular Decile	Explicit-Visual	Explicit-Visual

Elicitation of preference functions and Creation of a value function

Once the criteria have been defined using a value function (for e.g. accuracy can be defined using a value function such as 1-test misclassification rate), a tool such as AHP or analytic hierarchy process could be used for weighting the various evaluation criteria based on the input of domain experts or criteria used in similar past projects. The relevant actors involved in this process must also finalize on the acceptable threshold values for the various criteria and a formula for creating the composite score. The formula represents the value function associated with the data mining objective.

One way of creating the composite score is to calculate the weighted sum of different criteria. In

Table 5-28, we present the data mining success criteria for classification problems (where business requirement is to produce an explanatory model) to illustrate the concepts of value functions, weights, thresholds and composite score. Note that not all criteria need to be weighted and included in the composite score. For instance, there may be no weight associated with a criterion such as training time and the only requirement may be that the chosen model(s) should not exceed the threshold value associated with the training time.

Table 5-28: DMSC for Classification problems (BusReq = Explanatory)

Applicable Data Mining Success Criteria (description)	Value Function	Thresholds	Weights
Accuracy (Proportion Correctly Classified)	1-Test Misclassification Rate)	>0.75	0.60
Profit and Loss (unequal misclassification costs)	(Average Worst Possible Loss – Average Loss of Model)/(Average Worst Possible Loss – Average Best Possible Loss)	>0.75	
Lift (Cumulative %Captured Response at the kth Decile)	(Model-Baseline)/(Exact-Baseline)	>0	0.20
Stability (Visual inspection of the non-cumulative %Response Lift Chart)	Stability is binary, with 1 indicating a stable model and 0 indicating an unstable model	>0	0.15
Simplicity (Number of leaves or number of rules)	If No Of Leaves <=2 or >=13, then score = 0; If No Of Leaves =3 or =4 Then score = (NoOfLeaves-2)/3 If 5<=No Of Leaves> <=8; score=1 If 9<=No Of Leaves<=12 Then score =(13-NoOfLeaves)/5		0.05
Speed (Run Time)	Number of minutes	< 25	
Training Time (Time taken to train the model)	Number of hours	<5	
Memory Usage (Memory occupied in executing the model)	Number of GB's	<1	
Formula For Creating Composite Score	(0.60*Accuracy Score) +(0.20 Lift Score) + (0.15*Stability Score) + (0.05*Simplicity Score		

An organization should follow a similar methodology for other problem types such as prediction, estimation, association rules, clustering and visualization. Due to space constraints we have only presented the example of classification problems. It is

important to point out that while the data mining success criteria described through Table 5-13 and Table 5-14 are meant to assess the various models that would be generated under the chosen problem type (such as classification, association rules etc), the success criteria for visualization techniques are to be used for selection of a particular visualization tool. Since visualization includes visually exploring the data, it does require generation of multiple models. However these criteria can be used to select from the various evaluation techniques available to relevant actors.

This methodology of setting up evaluation criteria (Osei-Bryson 2004) reflects the fact that data mining success criteria are (or should be) determined before selection and implementation of a modeling technique (such as neural networks). By encouraging the actors to think about relevant success criteria it helps to eliminate any biases resulting from setting up success criteria after the decision regarding the modeling technique has been made.

Analysis of Applicable Data Resources (Using existing new variables, ratio variables or collecting data)

The business objective and data mining objective provide a glimpse into the applicable data resources. The DM projects base can also be used to identify applicable data by searching for similar past projects. It is important to note that as business situations change, new variables may need to be brought in to execute the set data mining objectives. These new variables may be available to the organization, may need to be

created (by combining existing variables), or may have to be sourced (either by purchase from an external data vendor or through data collection for the purpose of the data mining project). In the former two cases, there will be a cost associated with extracting the data and ensuring that it will be available to relevant personnel for the duration of the project. In the latter case, there will be a cost associated with collecting the data. The costs in both instances should be analyzed in accordance with the budget and should be approved before proceeding to the next task.

Consider the example of an organization involved in a data mining project aimed at studying credit worthiness of its customers. After some discussions the stakeholders identify a new variable namely number of total trade lines which they presumed could play an important role in discerning good versus bad accounts. Previously the company had only included the variable "number of delinquent trade lines" in its decision making model. The discussion among the technical personnel reveals that this variable may not provide the full picture, as it does not provide an idea of how delinquent a person was. For instance, they wanted to discern between a customer who had 3 trade lines and was delinquent on 1, versus a person who had 7 trade lines and was delinquent on 1 of these lines. Given their objective of improving approval rates, they wanted to closely analyze the latter individuals with a higher total trade line to delinquent trade line ratio, to see if they could qualify for some amount of credit. Accordingly they decided to introduce a new variable, the total trade lines to delinquent trade lines ratio, to increase the predictive accuracy of their new model.

221

Dependency with BU

Data Mining Objective

Target Variable

Repeat for all techniques and tools

Tool Supports DMSC?   Yes   No

Pass to Modeling

Array of modeling techniques

Tool Support available?   Yes   No

Notify relevant business/ technical personnel

Determine Business Objective +

Determine Data Mining Objective +

Data Mining Success Criteria

Determine preference functions

Determine value functions

Pass to Evaluation

Identify Data Sources

Determine Business Requirements +

Estimate data collection, operating and implementation costs

Approximate Data Related Costs.doc

Personnel Related Costs

Approximate Tool Related Costs

Total Costs

Total Benefits

Dependency with BU

Perform Cost Benefit Analysis

Document Reasons

Project Terminated by Sponsor

Sponsor's decision   No/Go   Go

Develop Project Plan

Figure 5-10: Process Model of Business Understanding Phase

222

5.3 Data Understanding Phase

During this phase an initial understanding of the data collected during the Business Understanding phase is performed. The goal is to identify data quality issues and to analyze the gross properties of the data. The execution of this phase is dependent on various tasks of the preceding Business Understanding Phase whereas its outcome is directly relevant to tasks in data understanding and data preparation. The various tasks are described below.

Table 5-29: Tasks of Data Understanding Phase

Tasks	Approaches/Steps	Output
Studying data sources and assessing data sufficiency	Steps specified	Data availability and sufficiency report
Assessing need for derived attributes	Steps specified	List of derived attributes to be created and their respective formulae
Documentation of data sources	Steps specified	List of all relevant data sources
Survey of data quality	Steps specified	Data Quality Report

Task 1: Studying data sources and assessing data sufficiency

DEPENDENCY WITH TASK(S)/PHASE
Identification of data (Business Understanding)
Data Mining Goal (Business Understanding)

During this task the data identified during the previous stage should be assessed for "sufficiency", i.e. to determine whether or not it would help meet the data mining goals. If the analysis reveals that data selected cannot help meet the said data mining objectives from the Business Understanding phase then the domain expert may consider acquiring the necessary data. In some cases, the required data may be available for purchase from an external data vendor. Financial institutions such as credit card companies often purchase data from vendors such as Acxiom etc. However, data may not always be available for purchase, but may in fact have to be collected using appropriate data collection techniques.

- If the organizational decision makers decide to acquire the data, follow task 1-1
- If the organizational decision makers decide to continue project with existing data, follow task 1-2

Task 1-1: Assessment of data acquisition costs and timelines against project budget and deadlines, respectively

DEPENDENCY WITH TASK (OF PHASE)
Project Plan (Business Understanding)
Financial Constraints (Business Understanding)

Decision to acquire additional data must be assessed against the costs involved and the time it will take to acquire the data.

- the time needed to acquire additional data must be assessed against the time allocated for this phase in the project plan. If time needed to acquire additional data meets the time allocated for this phase in the project plan, then costs of data execution must be assessed, as described below. If the time needed exceeds timelines outlined in project plan, then the decision makers will have to decide if they want to continue with the existing data (go to 1-2).

- the cost of data acquisition (whether by purchasing it from a vendor, or collecting it from relevant sources), must be assessed against the project budget outlined in the preceding phase. If the cost meets budget, then decision may be made to acquire new data. If the cost exceeds the budget, then organization will have to decide if they want to continue with the existing data, as described below.

Task 1-2: If organization decides to continue with existing data

225

If the decision makers decide to continue with existing data (either because it could not be acquired in accordance with timeline or in accordance with project budget), then possible effects on project outcome, quality and results must be documented. This is important since it is now known that the data that will now be used for analysis is insufficient. However, based on their knowledge experts may still decide to continue with the existing data.

Task 2: Assessing need for derived attributes

DEPENDENCY WITH TASK (OF PHASE)
Data Mining Goal (Business Understanding)
Policy Constraints (Business Understanding)

During this task, the decision makers must assess the data to make decisions regarding creation of derived attributes that are needed to adequately address the data mining objective. A meta database containing business metadata can be helpful for analysis of possibility of derived attributes. The business metadata helps assess (1) whether or not aggregating certain variables makes business sense and (2) ensures that the policy constraints (often laid out as business rules) are not being violated. The formulae and reasoning behind creation of derived attributes must be clearly documented.

Siddiqi (2005) highlights that users involved in creating derived attributes should avoid the "carpet bombing" approach which involves taking all variables and dividing them by everything else, and then generating a list of ratios that may be predictive but

are unexplainable. He emphasizes that all ratios should be justified and should be backed by good business reasons.

Task 3: Documentation of Data Sources

Once data sufficiency and need for derived variable has been assessed, data sources must be properly documented. This step is important and is directly relevant to the succeeding data preparation phase where the list of data sources identified during this phase is merged to create the dataset for analysis.

Task 4: Survey of Data Quality

This task comprises of a survey to assess data quality. A data quality report should be generated at this time which includes the description of any missing values and outliers existing in the data. It is recommended that the data quality issues such as missing values and outliers do not be addressed during this stage as the results of different modeling techniques are affected by the handling of the data quality issues. The data quality survey must also identify the ranges for various variables, variances and standard deviation as well as the density of each type of variable in the dataset.

The process model for this phase is shown in Figure 5-11
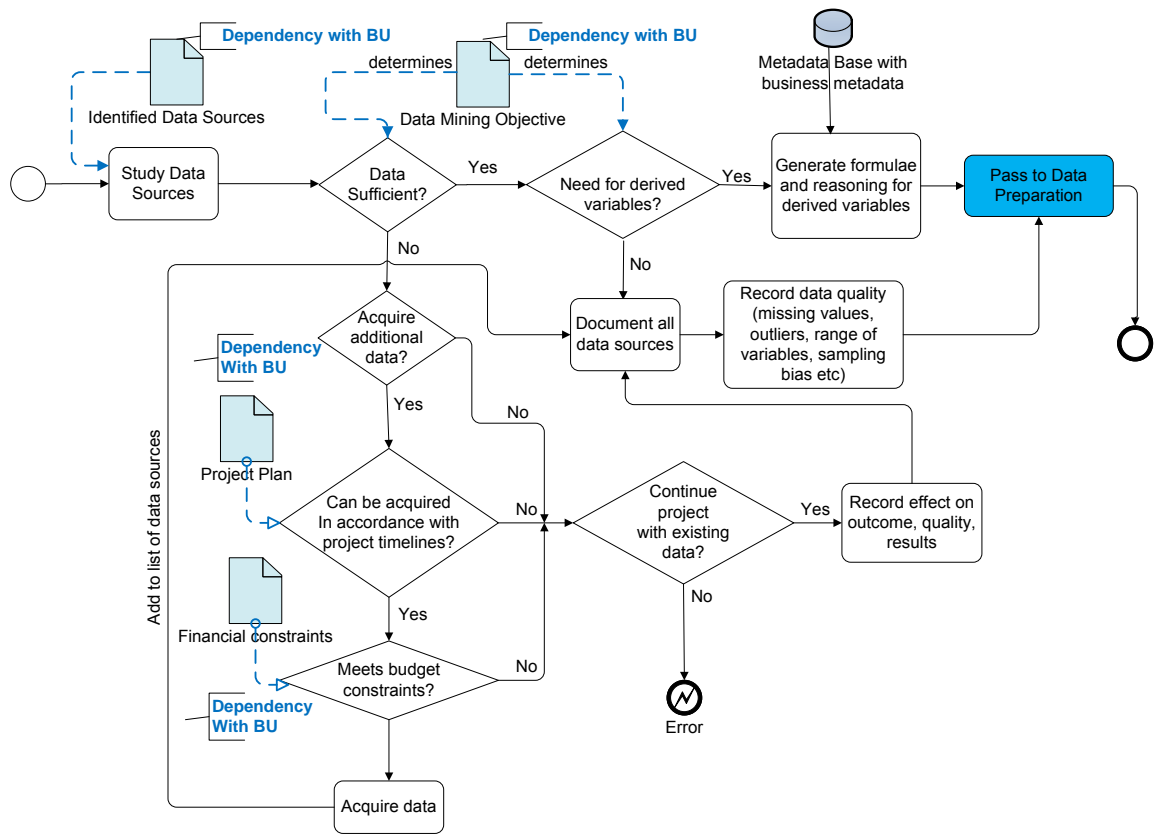
Figure 5-11: Data Understanding Phase

5.4   Data Preparation Phase

Table 5-30: Tasks of Data Preparation Phase

Tasks	Approaches/Steps	Output
Construction of Dataset	Steps Specified	Integrated data set containing the relevant data
Application of Policy and Legal Constraints	Steps Specified	Dataset after application of policy and legal constraints
Addition of Derived Variables*	Steps Specified	Dataset with derived variables added
Discretization of target variable*	Steps Specified	Dataset with target variable discretized (if applicable)
Fetch rank ordered array of modeling techniques (from Modeling Phase) and format the data	Steps Specified	Output data set compatible with requirements of modeling techniques
Loading data in software tool and applying tool specific formatting	Steps Specified	Output data set compatible with requirements of tool
Ensuring that tool can handle required number of rows and columns	Steps Specified	Output a dataset that can handle number of rows and columns

229

During this stage the final dataset is constructed from the raw initial data identified during business understanding and assessed during the data understanding phases. Several tasks in this phase share dependencies with tasks in business understanding, data understanding and modeling. The various tasks are described below.

Task 1: Construction of dataset

DEPENDENCY WITH TASK (OF PHASE)
Documented data sources (Data Understanding)

During this task, the final data set is constructed from the data sources documented in the preceding phase. This also includes additional data that may have been acquired. Note that such data was not available during the business understanding phase. The dataset so constructed is not the final dataset but must go through a series of refinements as described below.

Task 2: Application of Policy and Legal Constraints

DEPENDENCY WITH TASK (OF PHASE)
Determination of Policy Constraints (Data Understanding)
Determination of Legal Constraints (Business Understanding)

During this task the dataset created through various data sources is applied with policy and legal constraints to make sure that these constraints are not being violated.

As an example of policy constraints, an organization may have a policy that a product would only be offered to individuals 18 years or older in age. In such a case, any individuals whose age is less than 18 must be removed from the dataset to be used for analysis. As an example of legal constraints, law may require a firm to not make any decisions regarding offering products to customers on the basis of their sex or gender. In such a case, such variables must be removed before data is analyzed using modeling techniques. .

Task 3: Addition of derived variables

DEPENDENCY WITH TASK (OF PHASE)
Assessment of need for derived variables and their formulae (Data Understanding)

During this task, the derived variables identified during the preceding stage must be specifically added to the dataset. For example, experts may have determined Debt-to-Income Ratio as an important derived attribute for a predictive scoring model. In such a case, the variable debt-to-income must be created by dividing the values for debt by income.

Task 4: Discretization of target variable

DEPENDENCY WITH TASK (OF PHASE)
Data Mining Goal formulation (Business Understanding)

The data mining goal determines whether or not a target variable is applicable and if yes, if it needs to be discretized. The latter can specifically be answered while selecting the "purpose" when formulating the data mining goal. Once these four steps have been completed, move to Modeling Phase as described in Task 5.

Task 5: Fetch rank ordered array of modeling techniques and format the data

DEPENDENCY WITH TASK (OF PHASE)
Rank ordering modeling techniques (Modeling)
Survey of Data Quality (Data Understanding)

During this task, the data miner must jump ahead to the next phase, Modeling and fetch the rank ordered array of modeling techniques. Next, data must be formatted in accordance with the first modeling technique in the array. This step must be repeated for all techniques in the array or the top x number of techniques identified by the experts. Note that different techniques require data to be formatted in a particular way. For example if data is to be formatted for neural network processing then all variables may need to be mapped to a small range, such as 0 to 1 or -1 to +1, etc. note that the formatting according to techniques is also affected by the data quality survey conducted in the data understanding phase. For instance, data quality survey may have revealed the presence of certain missing values in the data. If the data is being formatted for decision trees, then the data miner and experts may decide to leave the missing values intact (and

not impute them), as (1) decision trees can handle them and (2) they can have predictive value and replacing them may affect the quality of decisions made by the tree.

Below we present a detailed description of how data can be formatted for various supervised and unsupervised modeling techniques. These guidelines can be stored in the modeling techniques base that can serve as a store-and look up source for formatting requirements for various techniques.

Formatting guidelines for Supervised and Unsupervised Modeling techniques

Data Preparation for Decision Trees

Data Preparation for decision trees is regarded as very simple. The following must be considered:

- There should be at least one input and at least one outcome variable.
- The input variables and target variables can be categorical or continuous. A tree with a categorical target variable is called a classification tree whereas a tree with a continuous target variable is called a regression tree.
- If business objectives and requirements, suggest that a continuous target variable should be binned, then such binning should be performed prior to running the tree. This is because the tree utilizes the nature of target variable in selecting its splitting criteria.

233

- Do not make hasty decisions regarding missing values as decision Trees can handle missing values and therefore missing values do not need to be imputed. This characteristic of decision trees allows for accommodating the fact that a value of null can often have predictive value and therefore records with missing values need not be thrown out or replaced with imputed values (Berry and Linoff 1997)

- Decision Trees are not sensitive to outliers or skewed distribution of numeric variables. This is because the decision trees only use rank order and not the absolute values.

- Use domain knowledge and expert's input to add derived variables to the list of input variables. Decision Trees cannot discover such relationships themselves and therefore the derived attributes must be created before decision tree modeling is undertaken.

- If decision trees are being used for prediction of sequential events, then data must be enriched with trend information by using fields such as differences and rates of change that explicitly represent change over time.

Data Preparation for Neural Networks

- Neural networks only accept numeric inputs. The inputs must be restricted to a small range such as -1 to 1. Such mapping of continuous and categorical variables should be done prior to training the network.

- The output from a neural network is also a number between 0 and 1 or -1 and 1 and should be remapped to get to the original scale of the target variable. This is done by applying an inverse of the transformation used for training the network.

- Continuous variables can be binned into ordered discrete values. Categorical variables must be treated more carefully as mapping to numbers may introduce a certain ordering (that although does not exist), will be taken into account by the neural network. Berry and Linoff (1997) point out that such ordering may or may not have an effect and should be cautiously considered. Another way is to break categories into flags, by assigning one flag to each value. Yet another approach (and perhaps most recommended) is to replace the categorical variable, if possible, with some numeric variable describing them.

- The training data set should cover the full range of features that the network might encounter including the output. This would include having several examples for each value of categorical variable and several examples for values of continuous and ordered discrete variables. While there is no simple rule to express relationships between the number of features and size of training data set, minimum of few hundred examples of each feature are needed to prevent over fitting

- Since the number of input variables affects the amount of time it takes to train the network, the choices regarding which input variables and derived variables must be included should be made judiciously. Decision Trees can be used to

identify important predictor variables and these can be subsequently modeled using a neural network.

- If the number of examples of a particular value for the output is less, then oversampling should be used to increase the proportion of rare cases.

- If any variables are showing a skewed distribution then this issue should be resolved prior to training the network. Neural networks are sensitive to skewed distributions since they make use of the actual values for the variables and not just the rank ordering like decision trees do. One way of addressing the issue of skewed distributions is to discretize or bin the relevant field. Taking logarithms is a good way of handling variables with wide ranges. Another approach is to standardize the variable (by subtracting the mean and dividing by the standard deviation). However, standard deviation must be used carefully if there are several large outliers as these can lead to many of the values falling within a small range making it difficult for the neural network to differentiate between them.

- If there are any missing values in the data, these should be replaced with imputed values. This is because neural network omits records with missing values in input or target variables.

Data Preparation for Association Rules

- Remember that there is no target variable in association rule mining.

- It should be ensured that the transaction data contains at least the following three different entities: customers, orders (also referred to as baskets or item sets) and items.

- It should be ensured that the data set is sparse. This means that only a small fraction of the attributes are non-zero or non-null in any given row. Examples of sparse data include market basket and text mining data. For example, in a market basket problem, there might be 1,000 products in the company's catalog, and the average size of a basket (the collection of items that a customer purchases in a typical transaction) might be 20 products. In this example, a transaction/case/record has on average 20 out of 1000 attributes that are not null. This implies that the fraction of non-zero attributes on the table (or the density) is 20/1000, or 2%. This density is typical for market basket and text processing problems. Data that has a significantly higher density can require extremely large amounts of temporary space to build associations.

- Missing values are not used in association rule modeling and therefore missing values should be imputed and replaced by non null values. Some authors (Ragel and Crémilleux 1998; Shintani 2006) have also proposed partitioning a database to deal the issue of missing values.

- Outliers should be treated with caution because when external equal width binning is used, all data will be concentrated in a few bins. In such a case, a single outlier may land in a bin. Outliers should be removed of this is the case.

- If data set is dense or has a large number of attributes then alternate techniques should be considered. Association rules do not deal with such data sets efficiently.

- If data set involves rare events, then association rules modeling is not recommended and alternative techniques such as classification modeling should instead be employed.

- If association rules are being used to perform sequential analysis, then the transaction data must have two additional features: a timestamp or sequential information to determine when the transactions occurred relative to each other and identifying information (such as account number, customer ID, household ID etc) that identifies different transactions as belonging to the same customer or household (Berry and Linoff 1997).

- If association rules are being used to compare different stores, then data must be augmented by adding virtual items. Such items describe the transaction though they are not themselves a product or service[1].

- Ensure that the items occur in roughly the same number of transactions. This presents data from being dominated by most common items. Consider the creation of a product hierarchy that can help roll up rare items (if any) to a higher level in the hierarchy so that they become more dominant.

Data Preparation for Linear Regression

---

[1] http://youngcow.net/doc/oracle10g/datamine.102/b14339/4descriptive.htm

- There should be at least one input and one output variable. This is regarded as simple regression. Multiple regression includes having several predictor variables (continuous or categorical) and one outcome variable (continuous).

- Ensure that al predictors have some variance (variables with zero variances should be excluded).

- If multi-collinearity (linear relationship between two or more predictor variables) is an issue, then this issue should be resolved prior to running the regression model. The correlation matrix can be checked to see if any variables correlate highly with each other; collinearity diagnostics such as VIF or variable inflation factor should also be considered [see Field (2000) for a review].

- It should be ensured that none of the predictors are found to correlate with external (or confounding) variables.

- Ensure that the relationship between the outcome and predictor variables is a linear one. If not then consider other alternatives such as curvilinear regression or other techniques such as decision trees or neural networks[2]. Linearity can be assessed by checking the box plot of observed versus predicted values (points should be symmetrically distributed around the diagonal) or residuals versus predicted values ((points should be symmetrically distributed around the horizontal). If non linearity is found and regression is still the choice of

---

[2] http://www.duke.edu/~rnau/411home.htm

technique, then non linear transformations should be applied to the input and/or output.

Data Preparation for Logistic Regression[3]

- Logistic regression involves discrete or continuous input variables and a dichotomous target variable. The target variable must be discrete

- There are no assumptions regarding predictors and therefore predictors do not have to be normally distributed, linearly related or having equal variance in each group.

- Assess the ratio of cases to variables, i.e. there should be enough responses for each category. If this is not ensured then it is likely that the standard errors will increase.

- Assess linearity in the logit, i.e., check that the regression equation has a linear relationship with the logit form of the discrete target variable (Ainsworth).

- Similar to linear regression, outliers can have a strong effect on the results of logistic regression. Outliers should be removed or modeled separately. The plot of residuals provides insights about the presence of outliers.

- If presence of interaction terms is suspected, these must be explicitly included in the model by adding them as independent variables.

---

[3] http://www.ats.ucla.edu/STAT/spss/output/logistic.htm

- In order to ensure meaningful results, all logit coefficients must be appropriately coded. The convention for binomial logistic regression is to code the dependent class of greatest interest as 1 and the other class as 0, and to code its expected correlates also as +1 to assure positive correlation. For multinomial logistic regression, the class of greatest interest should be the last class. Logistic regression is predicting the log odds of being in the class of greatest interest (Menard 2002).

Data Preparation for k Means Clustering

- Remember that there are no target variables in clustering and there is no distinction between independent and dependent variables.

- Different variables must be scaled such that there values fall in the same range. This can be done by normalizing, indexing or standardizing the values.

- If the business user believes one variable to be more or less important than the others, then different weights can be applied to encode such information. However such encoding should be preceded by first scaling the variables by standardizing them. In this sense, while scaling helps to remove bias due to the different measurement scales for the inputs, the weights added through encoding help to introduce bias based on domain knowledge and the business context.

- The data miners must consider creation of entity signatures such as town signatures, customer signatures etc. a signature is simply the collection of descriptive attributes about a particular entity. Creating such a signature requires

www.manaraa.com

aggregating data, normalizing it, calculating trends and adding derived variables (Berry and Linoff 1997).

Task 6: Loading data into software tool and applying any tool specific formatting

Once the above tasks have been completed, data should be loaded into the software tool for processing and any tool specific formatting should be applied. A tool repository can be used to store and look up information pertaining to formatting of techniques.

Task 7: Ensuring that tool can handle the required number of rows and columns

DEPENDENCY WITH TASK (OF PHASE)
Intermediate dataset (prior to formatting for modeling) (Data Preparation)

Next, the intermediate dataset created during this phase (after adding derived variables, imposing legal and policy constraints) should be assessed to study whether or not the tool can handle the desired number of rows (observations or records in the data) and columns (variables in the study). If the tool can handle the desired number of rows and columns, then data can be loaded into the tool and passed onto Modeling for running the algorithm on the prepared data set. If the tool cannot handle the required amount of rows, then proceed to task 6-1.

Task 6-1: when tool cannot handle the required amount of rows

When the tool cannot handle the required amount of data, it should be assessed if an alternate tool that can handle the number of rows is also available in the organization. If such a tool is available, check,

- if the tool can help assess (implicitly or explicitly), the data mining success criteria for the given project.
- If the personnel have knowledge and skills to use this tool
- if answer to any of these is no, then proceed to task 6-2

Task 6-2: when tool cannot handle the required amount of rows and no alternate tool exists (or existing alternate tool is not fit for use)

In such a situation, the organization must consider if they wish to buy a new tool. If they wish to not buy a new tool, then proceed to 6-3. If they wish to buy a new tool, then a list of alternate tools must be generated. Next, the experts must assess for each tool in the list, whether

- it can help provide for the project's data mining success criteria

- if the price of the tool meets the budget

- if the tool can be purchased in accordance with project's timelines

- if the personnel have knowledge and skills to use this tool.

Document list of all tools for which answers to all of the above is positive. Rank order these tools according to how well they fulfill the above four criteria (and any other criteria important to the organization). Request quotes for tool pricing from the vendor. Next, pass all the information to the project sponsor for decision.

- If the sponsor approves the purchase, then buy the tool, load data in the tool and pass to Modeling.

- If the sponsor does not approve purchase of the tool, then proceed to step 6-3

Task 6-3: when organization does not wish to buy a new tool (or sponsor does not approve purchase)

In this situation (wherein the buying of a new tool has been ruled out), the only option left is to consider reducing the number of rows and/or columns in accordance with the capabilities of the tool. The ultimate decision is made by the expert.

- if the expert decides to reduce the number of rows and/or columns, then he or she must also document the effect on project outcome, quality and results, before loading the data in the existing tool and passing to Modeling for analysis.

- The expert may also decide to exclude the modeling technique (for which tool support is an issue) itself from consideration and proceed to the next modeling techniques in the rank ordered array of techniques. All above steps will need to be repeated for the remaining techniques.

It can be seen that data preparation and modeling require several reiterations given that different modeling techniques require data to be prepared in particular ways. Figure 5-12 shows a schematic of the data preparation phase, its relations with two preceding phases, namely business and data understanding and its output to the modeling phase.
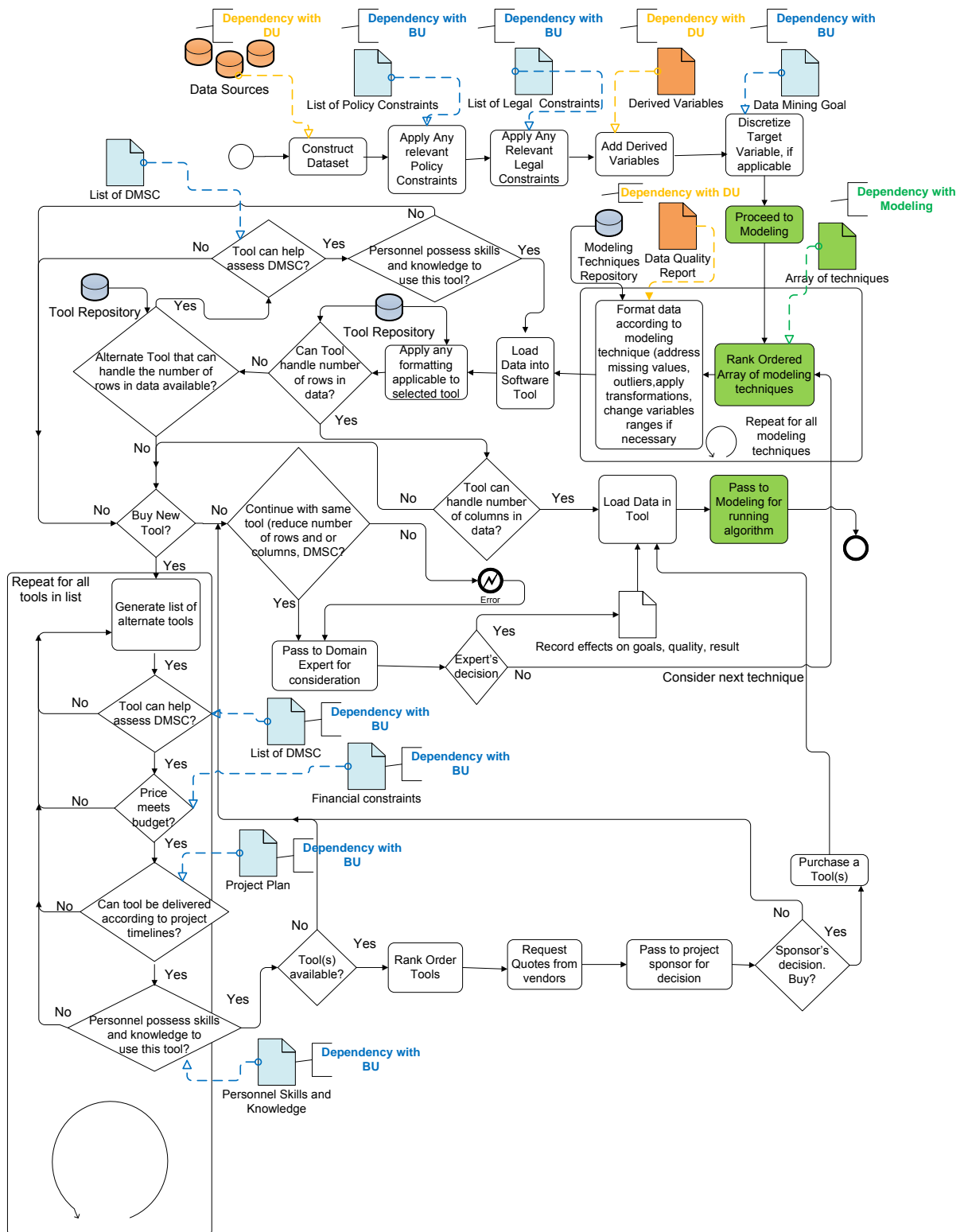
Figure 5-12: Process Model of Data Preparation Phase

246

5.5 Modeling Phase

Table 5-31: Tasks of Modeling Phase

Modeling Phase	Approaches/Steps	Output
Calculating values for accuracy and resource constraints for each modeling technique in the array of modeling techniques*	Steps specified	Rank ordered list of modeling techniques
Generate preference functions for resource constraints and setting up formula for creating composite score	Steps specified	Preference functions and formula for creating composite score
Rank ordering array of modeling techniques and making final selection of techniques*	Steps specified	Modeling techniques rank ordered by composite score
Select final set of modeling techniques from rank ordered list of modeling techniques*	Steps specified	Final set of modeling techniques
Fetch formatted data from Data Preparation phase (repeat for all techniques from finalized set of techniques)	Steps specified	Formatted Data loaded in software tool
Set up Model parameters (refine parameters on basis of objectives and success criteria, wherever applicable)	DM Software	Modeling techniques with parameters set up
Run modeling techniques and tabulate modeling results for all selected techniques in accordance with DMSC and DM Software used	Steps specified	Modeling results tabulated for DMSC

247

Task 1: Calculating values for accuracy and resource constraints for each modeling technique in the array of applicable modeling techniques

During this phase each modeling technique (or their combinations) would be applied to the formatted data. Note from the earlier discussion that often more than one technique may be applicable to a given data mining problem. For instance, if the data mining problem type has been identified as "Estimation" then both linear regression and neural networks may apply. In certain other problem types such as classification, the list of applicable techniques may be even larger (for e.g., any or all of decision trees, logistic regression, neural networks, naïve bayes, support vector machines etc may be used). While in an ideal scenario all applicable techniques (and their relevant ensembles) should be tried upon, real world constraints existing in business organizations, may prevent the execution of the complete set of applicable techniques.

In such as situation a decision needs to be made regarding which techniques should be tried out of the total set of techniques. This would require an approach for rank ordering the set of applicable techniques. Setting up objective criteria to speed up this decision is one possible approach. The solution offered by this dissertation suggests that the case base of historical performance of the various techniques on data sets of different sizes, the ease of analysis of the results, the speed of the technique (i.e. learning algorithm), among others could be used in rank ordering the list of applicable techniques.

Under conditions of time and resource constraints, the project owners could make use of only some of the techniques from this list. By having a Case base of historical performance of tools, the task of rank ordering of applicable techniques can be semi-automated. Below we provide further discussion, of how the various techniques could be compared using the above mentioned criteria. We also record the performance of different techniques on varying data sets to explain how a query optimizer-type logic could be used to select between these techniques when it is not possible to try all techniques.

One main type of constraint comes in form of computing resources such as training time and memory usage, with different techniques taking different amounts of time to train the model and using different amounts of memory in the process. It is expected that organizations may wish to optimize on these scarce resources and be able to rank order the techniques that must be tried upon. While computing resources are certainly an important constraint and making their best utilization is important, such optimization must also take into the account the relative accuracy offered by these techniques as accuracy is expected to be an important criteria in selection of techniques. A recommender system working on same logic as a query optimizer can assist in the process of ranking these techniques on the basis of their training time, memory usage, and accuracy.

As an organization makes a determination about which modeling techniques to execute on the basis of these variables, they can make use of a case base of past projects

that stores these values for similar datasets. The values for training time, memory usage and accuracy obtained on those data sets, can be used as a proxy for the existing data set and help in rank ordering the modeling techniques.

The following section presents a discussion of how the computing resources in form of training time and memory usage can be computed and how the relative accuracy for various techniques can be calculated.

Estimation of Training time:

The computer time and memory required for an analysis depend on the number of cases, the number of variables, the complexity of the model, and the training algorithm. For many modeling methods, there is a trade-off between time and memory. For all modeling nodes, memory is required for the operating system, the software supervisor, and the modeling diagram and programs, resulting in an overhead. This overhead amounts to 20 to 30 megabytes in case of SAS Enterprise Miner Software. The estimation of training time and memory usage is based on formulae provided by the SAS Enterprise Miner Help Manual.

Let:

- N be the number of cases.
- V be the number of input variables.

- I be the number of input terms or units, including dummy variables, intercepts, interactions, and polynomials.

- W be the number of weights in a neural network.

- O be the number of output units.

- D be the average depth of a tree.

- R be the number of times the training data are read in logistic regression or neural nets, which depends on the training technique, the termination criteria, the model, and the data. R is typically much larger for neural nets than for logistic regression. In regard to training techniques, R is usually smallest for Newton-Raphson or Levenberg-Marquardt, larger for quasi-Newton, and still larger for conjugate gradients.

- S be the number of steps in stepwise regression, or 1 if stepwise regression is not used.

For the Tree node, the minimum additional memory required for an analysis is about 8N bytes. Training will be considerably faster if there is enough RAM to hold the entire data set, which is about 8N(V+1) bytes. If the data will not fit in memory, they must be stored in a utility file. Memory is also required to hold summary statistics for a node, such as means or a contingency table, but this amount is usually much smaller than the amount required for the data.

For the Regression node, the memory required depends on the type of model and on the training technique. For linear regression, memory usage is dominated by the SSCP

matrix, which requires $8I^2$ bytes. For logistic regression, memory usage depends on the training technique as documented in the SAS/OR Technical Report: The NLP Procedure, ranging from about 40I bytes for the conjugate gradient technique to about $8I^2$ bytes for the Newton-Raphson technique.

For the Neural Network node, memory usage depends on the training technique as documented in the SAS/OR Technical Report: The NLP Procedure. About 40W bytes are needed for the conjugate gradient technique, while $4W^2$ bytes are needed for the quasi-Newton and Levenberg-Marquardt techniques. For a network with biases and H hidden units in one layer, $W=(I+1)H+(H+1)O$.

For both logistic regression and neural nets, the conjugate gradient technique, which requires the least memory, must usually read the training data many more times than the Newton-Raphson and Levenberg-Marquardt techniques. The formulae for memory usage for various techniques are summarized in Table 5-32.

Table 5-32: Estimating Memory Usage for Various Modeling Techniques

Name of Technique	Memory Usage
Tree Based Models	8N bytes preferred 8N (V+1) bytes
Linear Regression	$8I^2$ bytes
Logistic Regression (Conjugate Gradient technique)	40I bytes
Logistic Regression (Newton-Raphson technique)	$8I^2$ bytes
Neural Network (Conjugate Gradient technique)	40W bytes
Neural Network (quasi-Newton technique)*	$4W^2$ bytes
Neural Network (Levenberg-Marquardt technique)*	$4W^2$ bytes

* For a network with biases and H hidden units in one layer, W= (I+1)H+(H+1)O.

Estimation of Training Time

Assuming that the number of training cases is greater than the number of inputs or weights, the time required for training is roughly proportional to (see Table 5-33:

Table 5-33: Estimating Training Time for Various Modeling Techniques

NI^2 for linear regression.

SRNI   For logistic regression using conjugate gradients.

$SRNI^2$ For logistic regression using quasi-Newton or Newton-Raphson. R is usually considerably less for these techniques than for conjugate gradients.

DNI   for tree-based models.

RNW   for neural nets using conjugate gradients.

RNW^2 for neural nets using quasi-Newton or Levenberg-Marquardt. R is usually considerably less for these techniques than for conjugate gradients

Estimation of Accuracy

Of the various approaches (linear and logistic regression, trees and neural networks), neural networks and linear regression can be used for estimation type problems (where target variable is continuous). Logistic regression and tree based models can be used for classification type problems (where target variable is categorical). We propose to use the 1-test misclassification rate as a measure of

accuracy for classifier approaches and the Mean Square Error as a measure of accuracy for the estimation approaches. The various models are run on various SAS data sets. The name of the data sets and results obtained are tabulated below (Table 5-34 and Table 5-35).

Table 5-34: Performance of Classification Modeling Techniques (accuracy, training time and memory usage) –

Data Set	Classification Techniques	Accuracy	Training Time	Memory Usage
DMAGECR	Decision Trees	0.80	210000	176000
DMAHEQ	Decision Trees	0.70	696800	600320
DMAGECR	Logistic Regression	0.75	196800	146900
DMAHEQ	Logistic Regression	0.80	234000	259400

Table 5-35: Performance of Regression Modeling Techniques (accuracy, training time and memory usage)

Data Set	Estimation Techniques	Accuracy	Training Time	Memory Usage
FITNESS	Neural Network with conjugate gradient	3.65	4340	400[4]
FITNESS	Linear Regression	1.219	1519	392

Note that the values for training time and memory usage are non-normalized and must be normalized before the performance of techniques can be compared.

---

[4] Assuming three hidden nodes

Task 2: Generating preference functions (assigning weights) for resource constraints

The technical and/or business stakeholders should assign weight to performance criteria such as accuracy, training time and memory usage, based on the importance placed on these parameters. An AHP methodology similar to generating weights for data mining success criteria can be applied to generate the preference function. Next the formula for computing a composite score must be generated.

Task 3: Rank ordering array of modeling techniques and making final selection of techniques

The normalized results for accuracy, training and memory usage from output of task 1 and preference functions and formula for creating composite scores for these parameters from task 2, can be used to generate the final scores for different techniques. These scores can be rank ordered and a selection of modeling techniques can be made on the basis of the scores.

Task 4: Select final set of modeling techniques from rank ordered list of modeling techniques

The output of the previous tasks will help assess how the techniques fare on the criteria such as accuracy, memory usage and training time. In real world data mining numerous techniques and their combinations will be relevant. In such a case, in the interest of

managing resource constraints such as memory usage and training time and balancing them against accuracy, an organization may only decide to actually use only a subset of models from the array of applicable techniques.

Task 5: Fetch formatted data from Data Preparation phase (repeat for all techniques from finalized set of techniques)

Once the techniques have been finalized, the analysts will have to iterate between the modeling phase and the data preparation phase. The iteration back to data preparation is necessary as the data preparation phase helps generate data in a form suitable for modeling by the modeling technique. Since different techniques require data to be formatted in a particular way, the step will have to be repeated for all techniques in the array of modeling techniques. The formatting of data is a task of the data preparation phase and has been discussed in detail in the data preparation phase.

Task 6: Setting up Model parameters (refine parameters on basis of objectives and success criteria, wherever applicable)

Once the data has been formatted for a modeling technique, the parameters of the modeling technique must be set up before running the modeling technique. The setting up of parameters and their significance is acknowledged but no existing KDDM process model deals with this important task in detail. With respect to this task, the CRISP-DM (2003) user guide states that "With any modeling tool, there are often a large number of parameters that can be adjusted. List the parameters and their chosen values along with

reasons for their choice". The CRISP-DM process model provides no other guidance regarding how these parameter values can be chosen.

The IKDDM model discusses this task in detail. Analysis of parameters of modeling techniques such as decision trees, neural networks etc reveals that parameter settings are of two types:

(1) Parameters whose values are dependent on the specific objectives of the project and/or the data mining success criteria

(2) Parameters whose values are not directly dependent on the specific objectives of the project and/or the data mining success criteria.

As an example of the latter consider the number of hidden neurons in a neural network. SAS EM requires the user to specify the number of neurons. The screen shot is shown below.
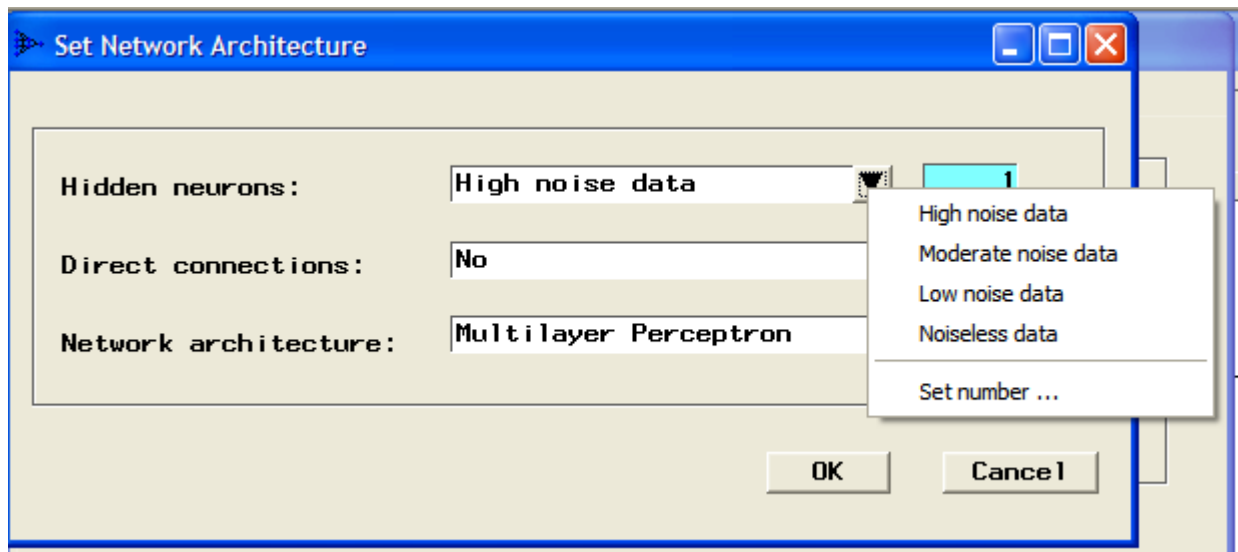
SAS EM Screen Shot 1: Setting up Number of hidden neurons in a Neural Network

If the user selects the number of hidden neurons based on the noise in the data (any of the first four items in the drop down menu), the number of neurons is determined at run time and based on the total number of input levels, total number of target levels, and the number of training data rows in addition to the noise level. Else the user can also set the number of neurons herself. The number of hidden neurons helps the neural network perform complex internal calculations, which are what make a neural network so powerful. However for the business or technical user interested in developing a model, the number of hidden neurons has no direct relationship with the objectives or success criteria of the project. While this parameter has its own importance, its values cannot be estimated on the basis of project objectives. In case of

such parameters, the user can vary parameter values and determine which one leads to the most desirable model.

Our interest lies in parameter values that are affected by the objectives and success criteria of the project. Simply varying such parameters from their default values is not likely to lead to a good model. Below we discuss such parameters and explain their relationship with objectives and success criteria.

Modeling parameters dependent on output of data mining objectives and success criteria

Selecting purity measures for evaluating splits

Splitting criterion depends on the target variable, which is determined by the business and data mining objective (Berry and Linoff, 1997). If target variable is categorical, then Gini, Information Gain or Chi Square may be used. If the target variable is continuous, then Variance reduction or F test may be used. However, if the business and data mining objectives required discretizing a target variable, then one of the three measures applicable to categorical targets may be used. In SAS EM, the variance reduction and F test are not available for selection if a categorical target is selected. However, if the user decides based on the objective that discretizing the target is needed, then she must be cognizant about the importance of the splitting criterion and its relationship to the target and make the appropriate selection.

259

Number of leaves

The number of leaves of a tree is regarded as a measure of simplicity of a tree. However, if the target variable is continuous then a tree can only generate as many discrete values as there are leaves in the tree. This means that if the number of leaves in the tree is set at 4, then all the values for the target variable will be grouped into four discrete categories. If the continuous target variable is the yearly income of a household and the range in the sample varies from [30,000 to 200,000], then each the value of target variable for each target variable will belong to one of four categories. This may or may not be desired based on the data mining objective. due to the above reasons, the value for the parameter number of leaves should be based on the data mining objective.
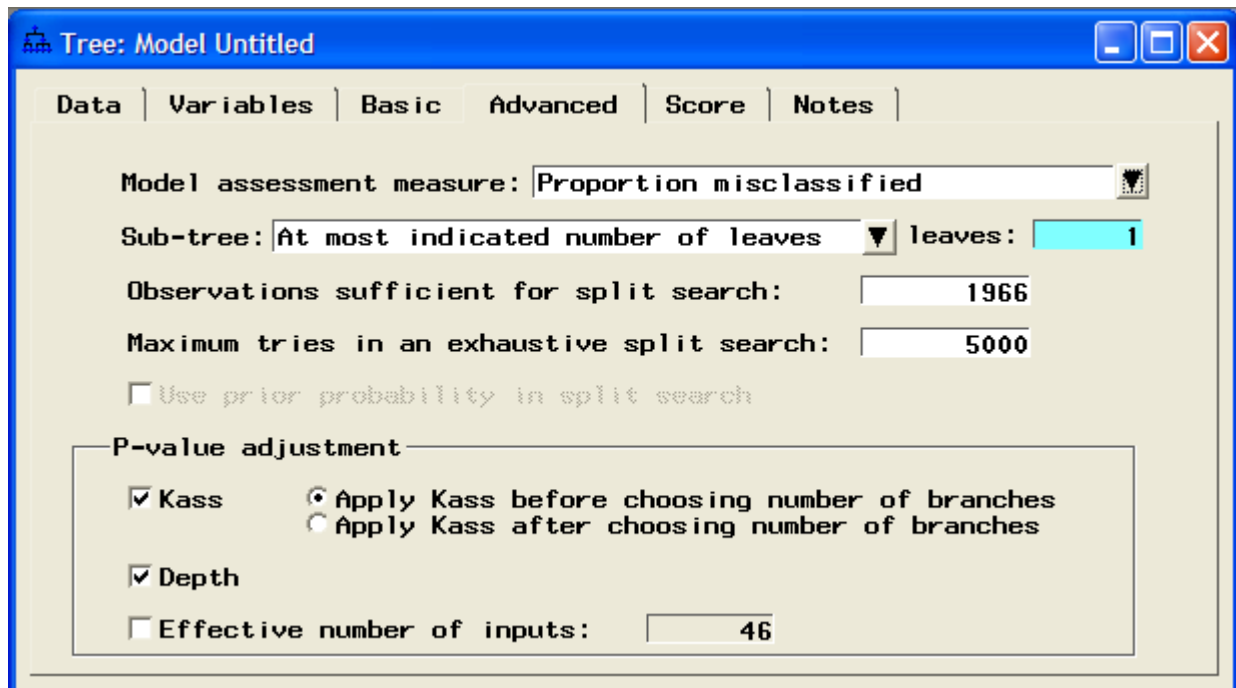
SAS EM Screen Shot 2: Setting Number of leaves in the sub-tree

Relationship between number of leaves and Stability

If number of leaves becomes very small, then it is likely that a large difference will appear between the performance of the training and validation data sets. If stability is an important data mining success criterion, then the value for the parameter number of leaves should be set up based on the acceptable levels of stability required by the user for the model to be considered successful.

Relationship between number of records at a node and Stability

261

Decision trees with nodes that have too few records are likely to be unstable. If stability is an important data mining success criterion, then the value for the parameter number of records should be set up based on the acceptable levels of stability required by the user for the model to be considered successful.

Relationship between number of leaves and objectives of the project

If goal is to generate scores, then having a large number of leaves is useful since each leaf generates a different score. If on the other hand, the goal is to generate rules, then it is better to have fewer rules.

Relationship between depth of a tree and efficiency of a tree

The average number of layers from the root to the terminal nodes is referred to as the *average depth* of the tree. In general, the average depth of the tree will reflect the weight given to efficiency.

Relationship between breadth of a tree and accuracy of a tree

The average number of internal nodes in each level of the tree is referred to as the *average breadth* of the tree. In general, the average breadth of the tree will reflect the relative weight given to classifier accuracy (Safavian and Landgrebe 1991)

Regression Models

Relationship of Model Selection criteria to Business Success Criteria and Data

Mining Success Criteria

If the user selects back, forward or stepwise regression methods, then he must specify model selection criteria. The choice of the model selection criteria stems directly from business and data mining success criteria. For example, if the business success criterion includes profit or loss, then it must be selected as the model selection criteria. The output will be a model that maximizes the profit or minimizes the loss.

If simplicity is one of the data mining success criterion, then AIC (Akaike's Information criterion) and SBC (Schwarz Bayesian criterion) must be selected. One way of assessing simplicity is based on the number of variables used in the model which directly affects the number of parameters of the model. Both of these criteria penalize for adding parameters to the model. The screenshot below shows model selection criteria for linear and logistic regression models in SAS EM.

263
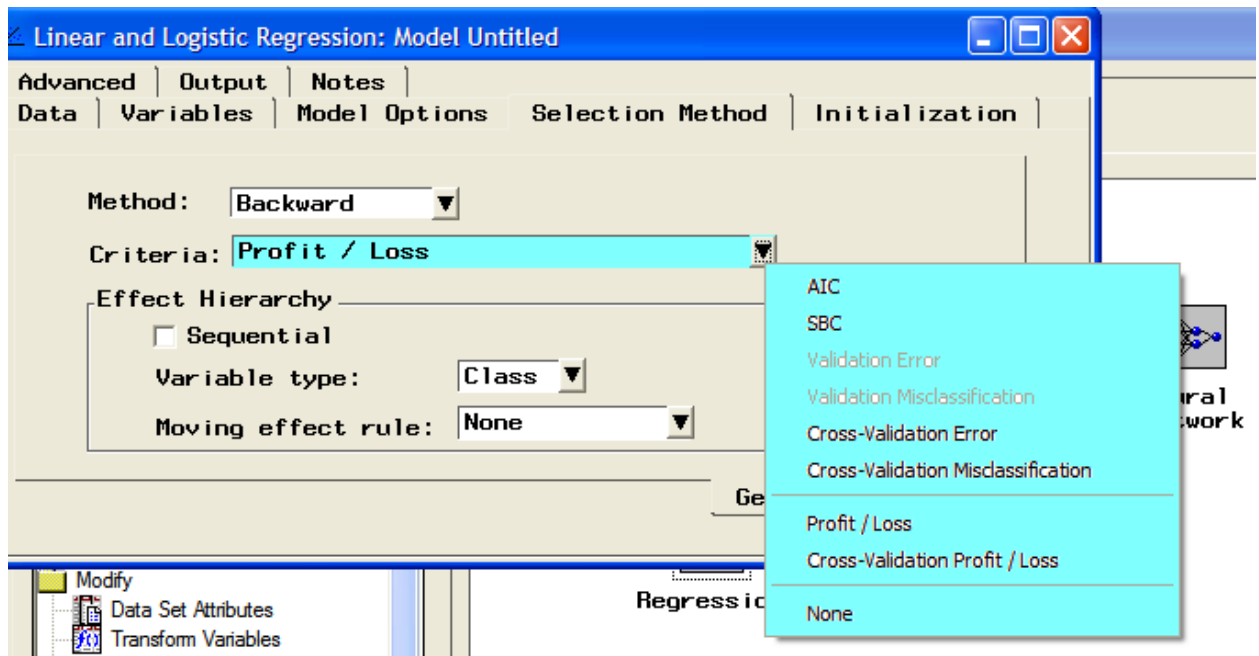
SAS EM Screen Shot 3: Model selection criteria for Linear and Logistic

Regression Models

Relationship of number of effects to data mining success criterion Simplicity

The number of effects in the model has a direct relationship to the data mining success criterion simplicity. Depending on whether simplicity is a data mining success criterion and the weight assigned to it, the user should select the number of effects in the model, and also specify (if possible) the effects that must be considered in the model. the screenshot below shows how the candidate effect can be specified by the user in SAS EM.

264
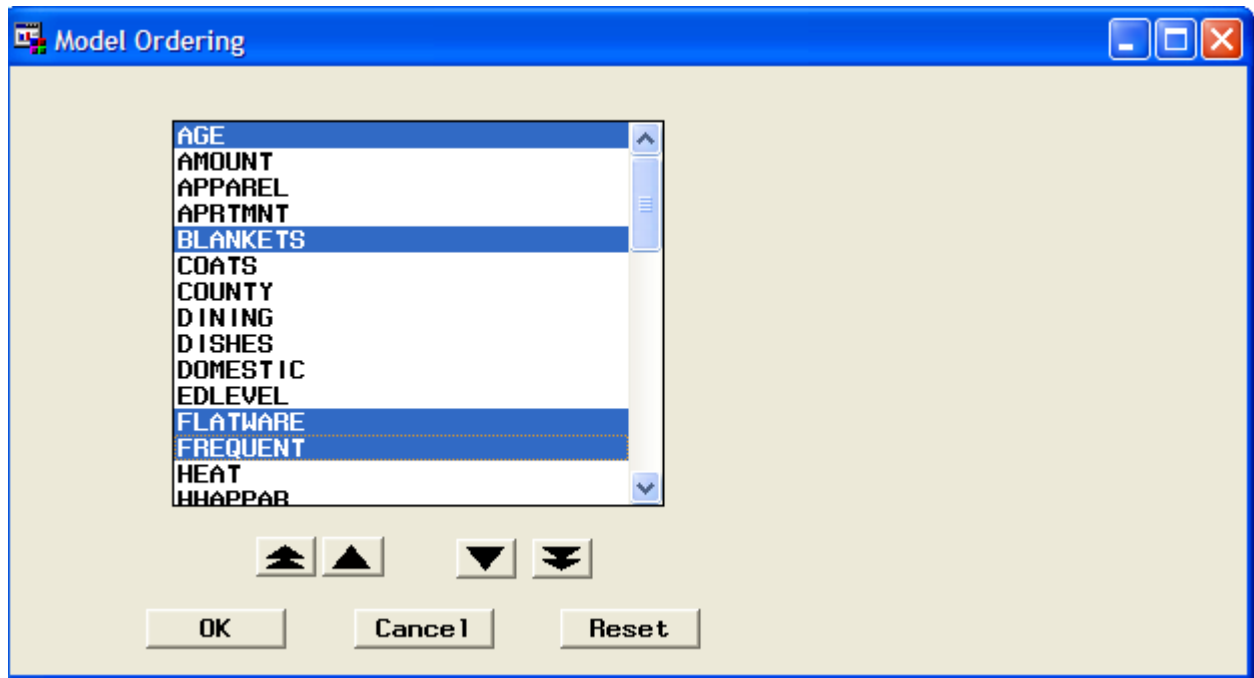
SAS EM Screen Shot 4: Selecting number of candidate effects to be used in the model

If the user has reason to believe on the basis of data understanding that certain effects are important and must be included in the model, she can move them up in the effect's hierarchy. Note that if this choice is made then the selected effects will be included even if they turn out to be non-significant. The screenshot below shows how this can be done in SAS EM.
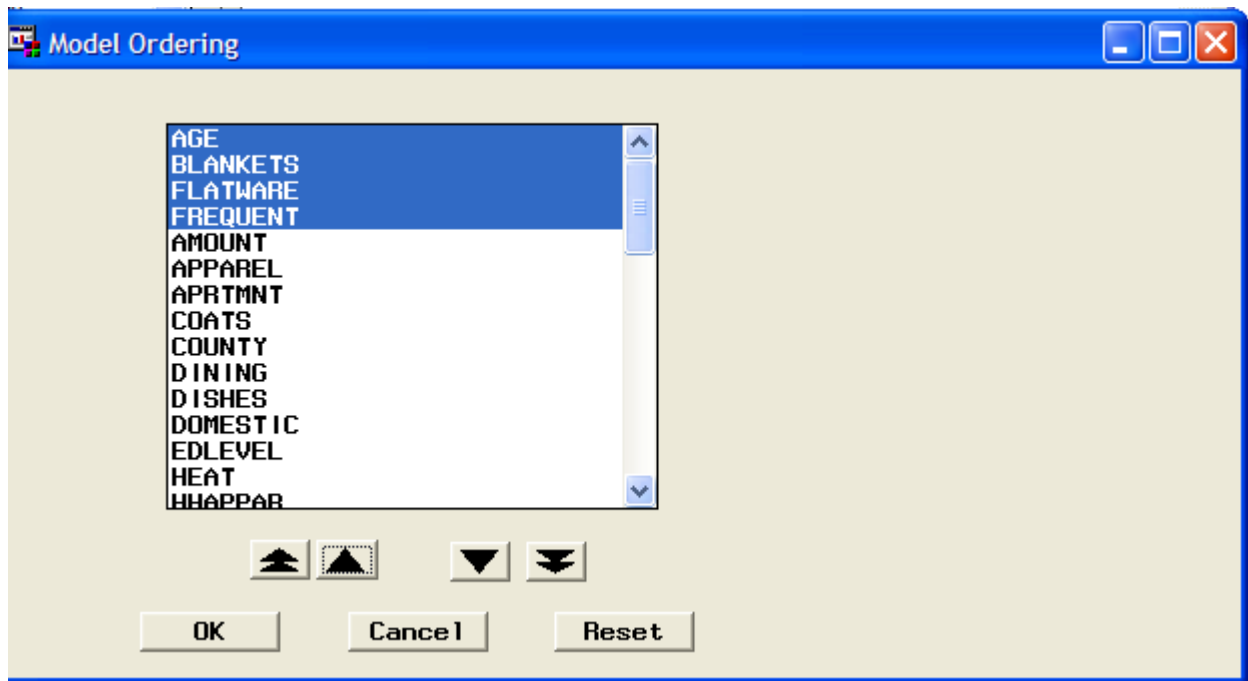
SAS EM Screen Shot 5: Forcing effects in the model

Relationship of optimization methods to size of problem

Regression problems require the user to select the optimization method to be used in building the model. The choice of the optimization method is related to the size of the data mining problem or the number of parameters which is known at this stage of the KDDM project. The screenshot below shows how the optimization methods must be selected in SAS EM.

www.manaraa.com
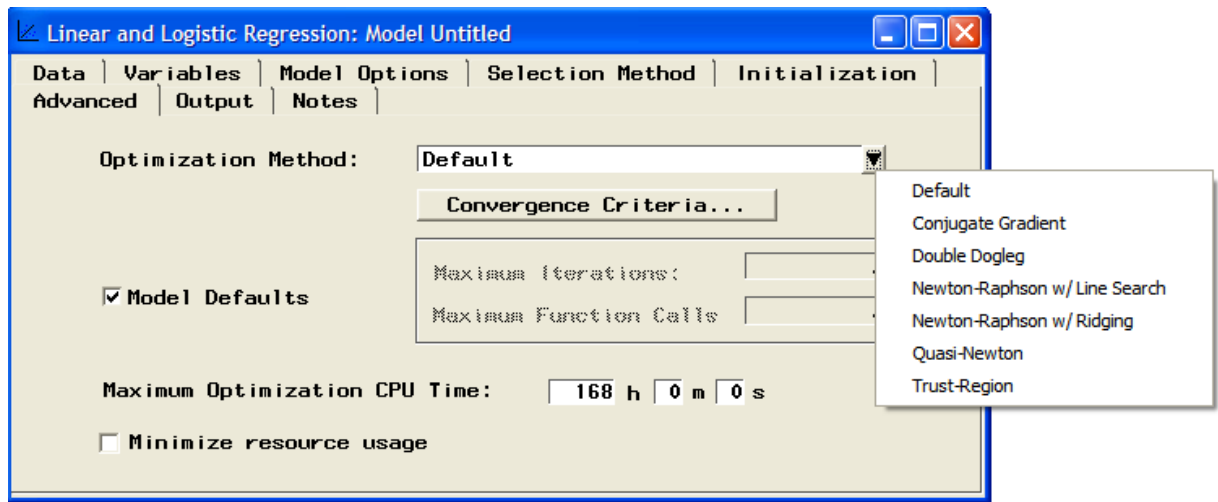
SAS EM Screen Shot 6: Selecting Optimization Method

The SAS user guide recommends the following choices for optimization methods based on the number of parameters:

- For small to medium problems (number of model parameters up to 40), Trust-Region, Newton-Raphson with Ridging, and Newton-Raphson with Line Search optimization methods should be used.

- For Medium Problems (number of model parameters up to 400), the Quasi-Newton and Double Dogleg methods are appropriate

- For Large Problems (number of model parameters greater than 400), the Conjugate Gradient method is most appropriate

Association/Sequencing Models

Setting up Minimum transaction frequency based on the data mining success criterion frequency

If frequency is a data mining success criterion, then the value for the parameter minimum transaction frequency should be set up based on this criterion. By setting up the value for this parameter the user can filter out any infrequent associations. The screenshot below shows how the parameter transaction frequency in SAS EM's association node can be set up.
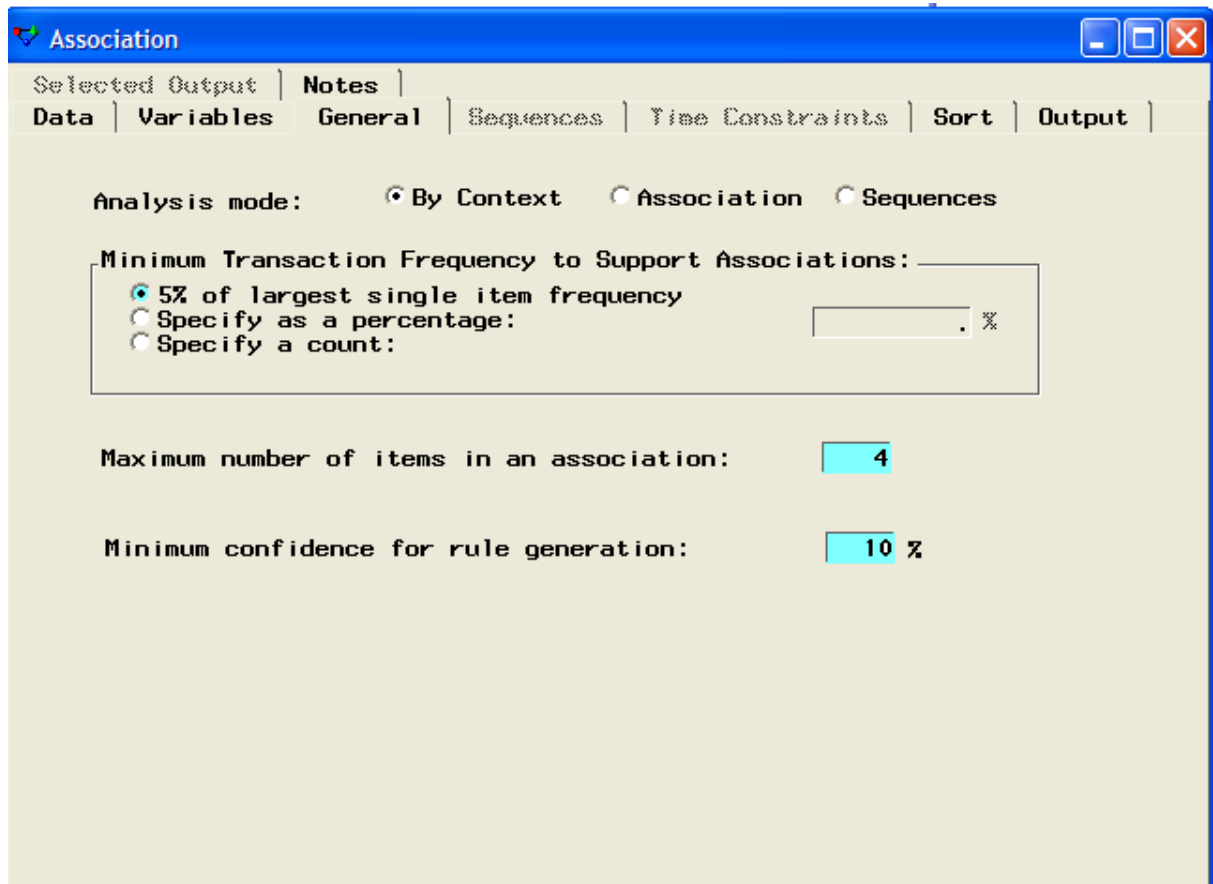
SAS EM Screen Shot 7: Setting up transaction frequency, minimum number of

items in an association

and minimum % confidence level

Relationship of Minimum confidence for rule generation to Data Mining Success

Criteria

This parameter specifies the minimum confidence level to generate a rule. In SAS EM,

the default value for this parameter is 10%. However this parameter is directly based on

the data mining success criterion confidence. For example if the criterion stipulates that only rules with a certain level of confidence, such as 70% are relevant, then the user must set up this parameter accordingly. Not setting up the value of this parameter in accordance with the data mining success criterion on confidence kevel, will result in generation of too many rules, even those that do not meet data mining success criteria. The screenshot above shows how the parameter minimum confidence level in SAS EM's association node can be set up.

Setting up umber of items in the longest chain of a sequence based on data mining objective

This parameter enables you to set the maximum number of items to include in a sequence. The user should select the value for this parameter based on the data mining objective. For example, owing to business reasons the user may only be looking for a maximum of 5 items in the longest chain. In such a case, the parameter maximum number of items must be set accordingly and not left at the default value.
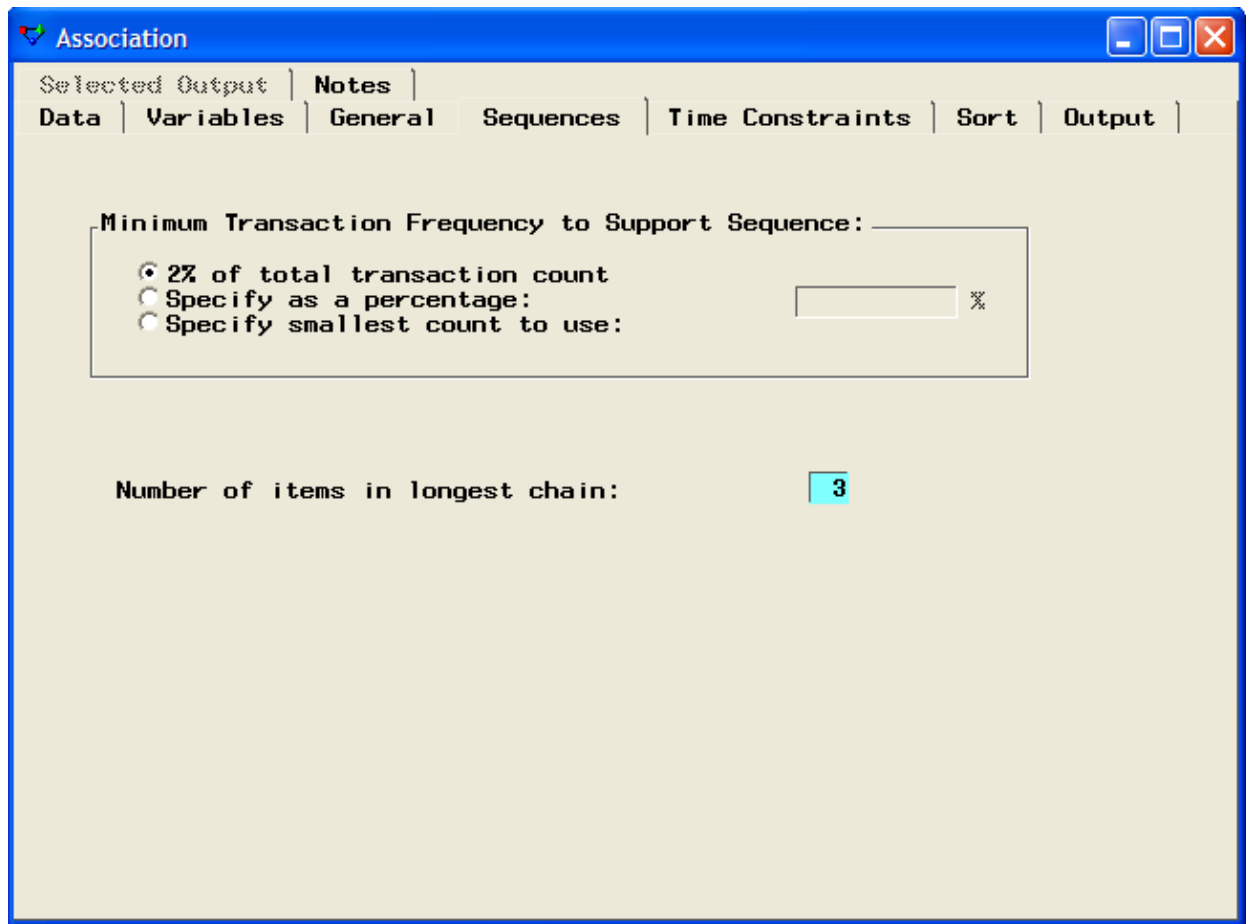
SAS EM Screen Shot 8: Setting up number of items in longest chain of a sequence

Task 7: Run modeling techniques and tabulate modeling results for all selected techniques in accordance with DMSC and DM Software used

After the Modeling parameters have been set up, the modeling technique can be run using the selected data mining software. The output of the modeling techniques must be presented in tabular form showing the results for all the data mining success criteria. As discussed earlier, while some data mining success criteria are output

271

explicitly by the tool (meaning they can be directly obtained from the modeling output), others may only be provided implicitly (meaning that the user will have to calculate values for these criteria using certain formulae). The various data mining success criteria (both explicit and implicit) supported by data mining software have been discussed in detail in Table earlier.

The figure below shows a schematic of the modeling phase, its relation to two preceding phases, namely business understanding and data preparation, and its output to evaluation phase.
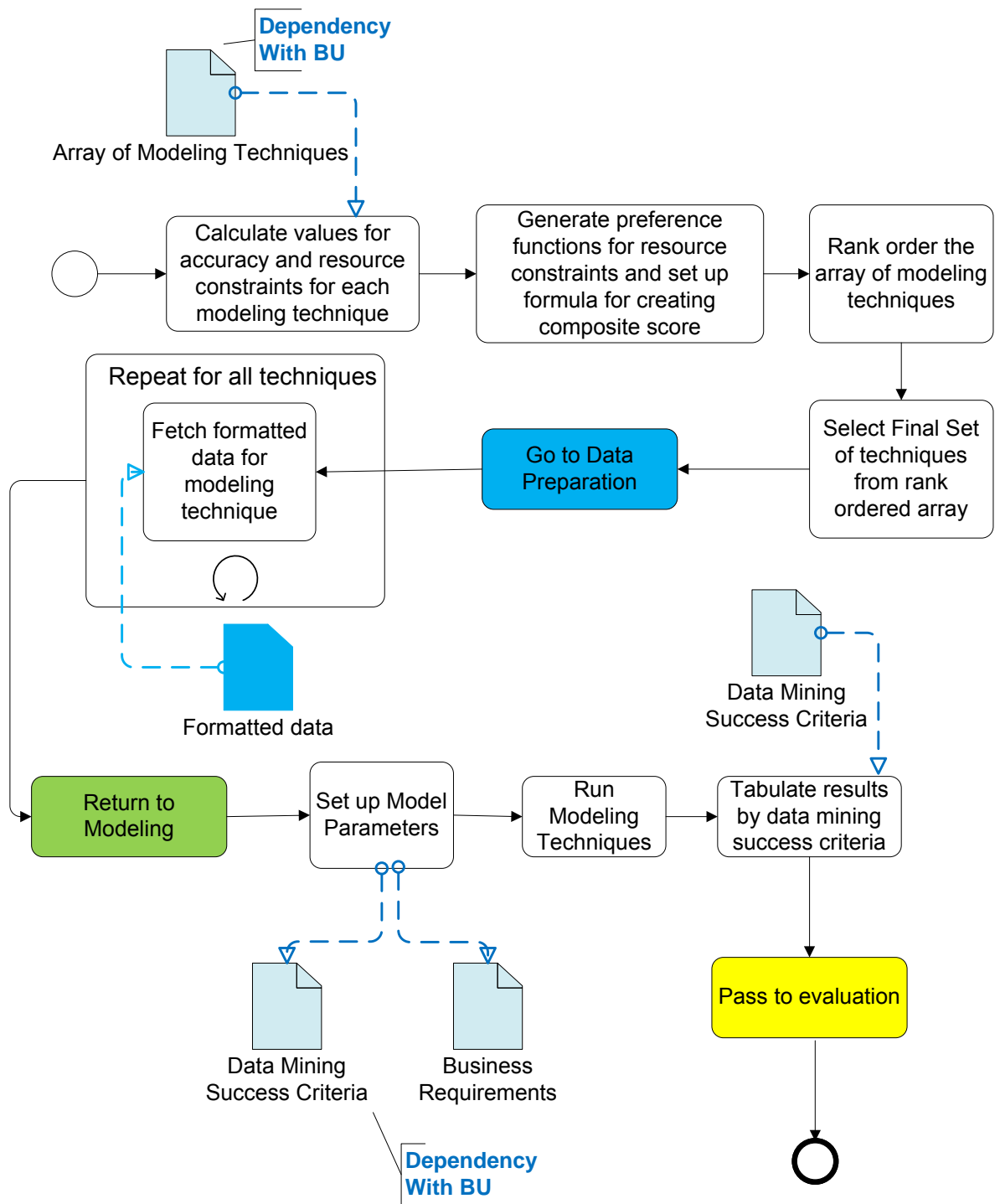
Figure 5-13: Process Model of Modeling Phase

5.6   Evaluation Phase

Table 5-36: Tasks of Evaluation Phase

Tasks	Approaches/Steps	Output
Assessment of Modeling results against data mining success criteria*	MS Excel, DM software	Model results assessed with respect to business and technical success criteria
Assessment of Modeling results against business success criteria*	DM Software	Summary of results of testing chosen model on real application
Using value functions to create composite scores for selected models *	Steps specified	Models rank ordered by composite scores
Compare models with the same composite score against different data mining success criteria (if applicable)*	Steps specified	Models rank ordered by performance on DMSC
Determine next steps for the project	Steps specified	List of next steps for the project

* Candidate tasks for semi-automation

During this phase, the results of the chosen modeling technique (output by the modeling phase) are evaluated against the business and technical success criteria. If the chosen solution only has technical merit and satisfies the DMSC but does not fulfill the business objectives (assessed via the accomplishment of business success criteria) then it cannot be regarded as a feasible solution. Also, vice versa if the solution satisfies business requirements but does not meet the technical success criteria, it cannot be regarded as an acceptable solution. A rigorous check is needed to provide evidence that the solution indeed meets both types of success criteria. The recommended tasks for this phase are documented below.

Task 1: Assessment of modeling results against data mining success criteria

DEPENDENCY WITH TASK (OF PHASE)
Setting up Data Mining Success Criteria (Business Understanding)

This task comprises of assessing each model tried during the modeling phase, M = [1, 2, ….m] against data mining success criteria. Following sub-steps are included

- assess modeling results against threshold values for different data mining success criteria.

- Store models that meet threshold values for all criteria in list of approved models, M = [1,2,…k], where k < or = m

- Store models that do no meet threshold values in list of 'models rejected for technical reasons'

- If no models, meet the technical success criteria, follow step 1-1

Task 1-1: Suggested solution when no models meets data mining success criteria (if applicable)

FEEDBACK TO TASK (OF PHASE)
Setting up Data Mining Success Criteria (Business Understanding)

If no models, meet technical success criteria, then business and technical stakeholders who set up the data mining success criteria during the business understanding phase must consult to finalize new threshold values for data mining success criteria

- select models that meet the new threshold values in list of approved models
- if it is not possible to change threshold values, then the decision makers may opt to continue with the existing model (challenger model) and closing the project. The reasons for closing the project must be documented

Task 2: Assessment Of Modeling Results Against Business Success Criteria

DEPENDENCY WITH TASK (OF PHASE)
Setting up Business Success Criteria (Business Understanding)

This task comprises of assessing each model from list of approved models generated above against business success criteria

- store models that meet threshold values for all business success criteria in list of approved models, M = [1,2,…f], where f < or = k

- store models rejected in list of 'models rejected for business reasons'

- if no models, meet the business success criteria, then follow step 2-2

Task 2-2: Suggested solution when no models meets business success criteria (if applicable)

FEEDBACK TO TASK (OF PHASE)
Setting up Business Success Criteria (Business Understanding)

If no models, meet business success criteria, then business stakeholders who set up the business success criteria during the business understanding phase must consult to finalize new threshold values for business success criteria

- select models that meet the new threshold values in list of approved models

- if it is not possible to change threshold values, then the decision makers may opt to continue with the existing model (challenger model) and closing the project.

Task 3: Using Value function(s) to create composite scores for selected models

277

DEPENDENCY WITH TASK (OF PHASE)
Setting up Data Mining Success Criteria (Business Understanding)

This task comprises of applying the value function set up during the business understanding phase to determine a composite score for all approved models

- rank order all models according to their composite scores

- select model with highest score as the best model.

- Assess best model against business requirements. If model selected as best model meets business requirement, then continue to step 5

- If model selected as best model does not meet business requirements, then proceed to step 3-1

- If models meet business requirements, but in case of a tie between two models, follow step 4.

Task 3-1: Suggested solution when no model meets business requirement (explanatory/ non-explanatory model), if applicable

FEEDBACK TO TASK (OF PHASE)
Creating Models (Modeling)

If model selected as best model on the basis of the composite score is one that does not meet business requirement

- after consultation with technical and business stakeholders, submit to modeling phase to construct a 2 stage model where output of non explanatory model is explained using an explanatory model
- business and technical stakeholders may wish to use this model as the final model or use the best explanatory model from modeling results available in previous step as the best model.

Task 4: Compare models with the same composite score against different data mining success criteria (if applicable)

DEPENDENCY WITH TASK (OF PHASE)
Threshold Values for data mining success criteria (Business Understanding)

- document which model performs best on each criteria
- document which model performs best on criteria with highest weight
- Recommend model that performs best on criteria with highest weight as the best model. Present results from the competing models with same composite results as well. The final decision of selecting between the two models rests with the domain experts
- Assess the best model to see if it meets business requirements. If model meets business requirements, then recommend it as final model; else, repeat task 3-1

Task 5: Publish list of next steps for the project

- if a model was chosen on the basis of the data mining project, then domain experts and high level stakeholders are expected to publish a list of next steps detailing how exactly the chosen model will be implemented via operations.

- If no model was chosen (either due to not meeting data mining or business success criteria), then relevant personnel must be informed about the decision to continue with the challenger model (if any).

- High level business and technical stakeholders must also discuss reasons for inability to find a suitable model: they may wish to specifically study (1) incorporation of new variables that may have lead to improving the model and resulting in an acceptable model (2) purchase of data mining software, if existing software did not allow for running the higher ranked modeling techniques or if a tool could not be chosen due to its inability to support data mining success criteria. (3) hiring of relevant personnel (internal or external) if knowledge or skills of personnel may have contributed to failure in discovering appropriate model.

The following schematic shows the process model for the Evaluation Phase
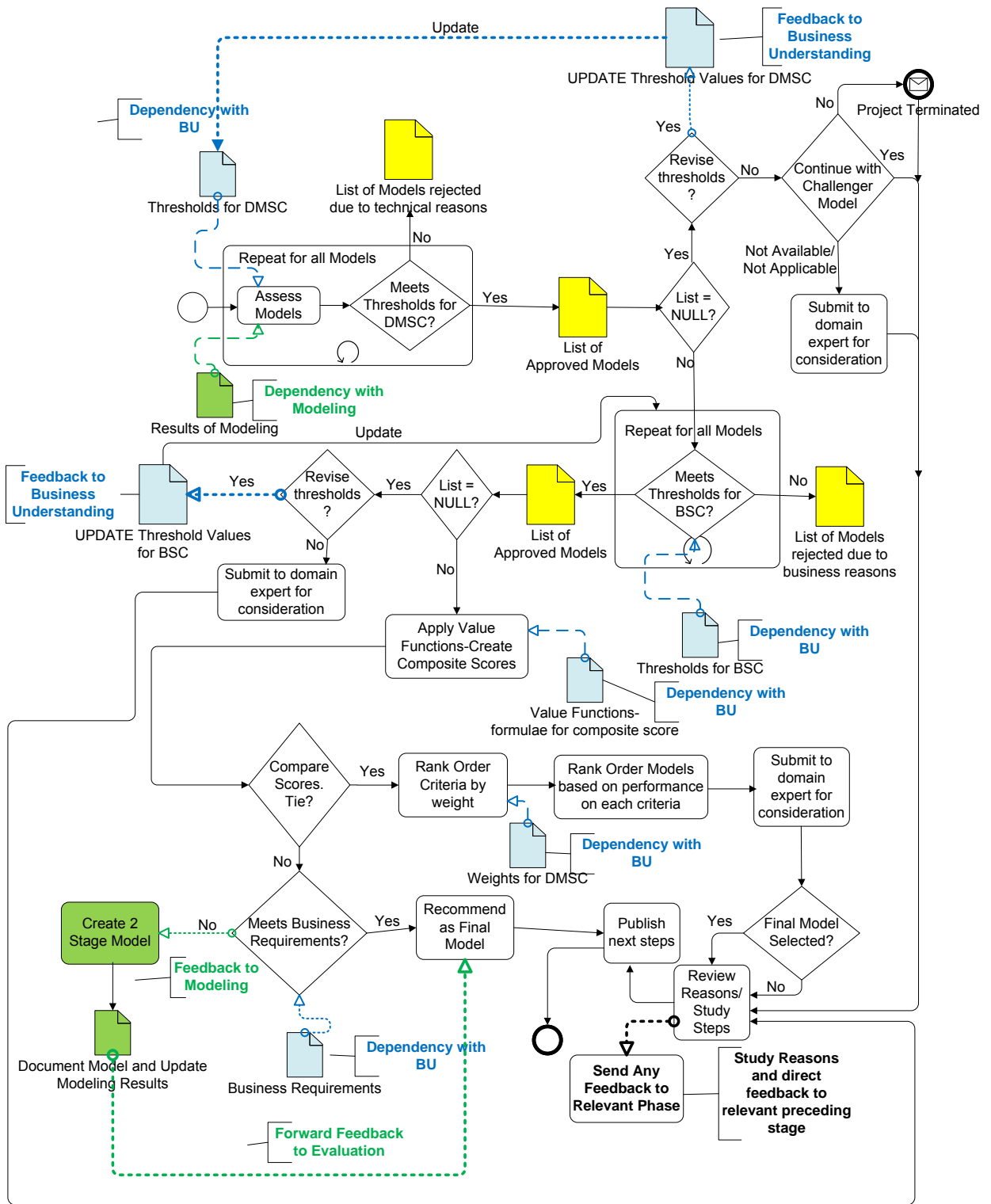
Figure 5-14: Process Model of Evaluation Phase

5.7 Deployment Phase

Deployment is the final phase of the KDDM process. Given this fact, the tasks of this phase exhibit numerous dependencies with tasks of the previous phases. The model(s) selected at the end of the evaluation phase are now deployed or implemented. The actual implementation results in valuable feedback for the preceding phases. The IKDDM model proposes the following list of tasks as part of this phase.

Task 1: Documentation of project activities

During this task, the personnel must work to ensure that a systematic account of the lifecycle of the project has been recorded. This is critical knowledge which if captured can be reused leading to improvement in efficiency and effectiveness of the KDDM project. As we have seen in the preceding phases, there are a wide variety of tasks that are accomplished as part of the preceding five phases. We recommend documenting the output of some of these tasks in the project report. These include the following:

1. Business objectives
2. Business Success Criteria
3. Data Mining Objectives
4. Data Mining Success Criteria
5. Initial cost benefit analysis

283

6. Assessment of data sufficiency

7. Assessment of data quality

8. Rationale for inclusion of derived variables (if applicable)

9. List of any tool specific formatting changes applied to data

10. Any reduction in size of dataset due to tool related constraints (if applicable)

11. Array of applicable modeling techniques

12. Array of modeling techniques used and justification for excluding any others

13. Modeling results tabulated by data mining success criteria

14. Modeling results tabulated by business success criteria

15. List of models rejected for technical reasons

16. List of models rejected for business reasons

17. List of approved models

18. Details of selected model, results, parameter settings and justification for selection

19. Final cost benefit analysis and deviations from initial analysis at the beginning of the KDDM process

Note that each of these tasks has a dependency with the corresponding task of the previous stage. Given the design of the IKDDM model wherein these tasks were systematically implemented in the earlier stages, the generation of such a documentation report can be easily automated, thereby removing any source of documentation burden on the users.

Task 2: Deployment of model on a test sample

The models generated through the KDDM process is deployed on a test sample (subset of the overall population) to see how well the model work in reality. The test sample is not the same as test data that was used for conducting the modeling. The test data used for modeling is a hold out sample. So for instance for supervised data mining problems, we know the outcome for each record, but hold it to see how the model does on this data. The test sample however is the new population, the population for which the model was built. But since there is some risk in deploying a model directly on the complete population, organizations often deploy it on a small percentage of this population, and assess the model's performance, how well it matches up to expectations and if it should indeed be deployed on the overall population. Results from this analysis should also be added to the project's documentation report.

Task 3: Creating a model maintenance plan

The models generated in the KDDM process use data captured in a particular time frame. Given this fact, the performance of the model is affected by time. It is likely that the performance decay with the progression of time. An organization must have a model maintenance plan that explicates how the organization plans on dealing with this issue. This can be done automatically or semi-automatically by refreshing the data used for

285

building the model and reassessing the model's performance on this new data. The new data may become relevant due to significant changes in attributes of the object (such as customer demographics or buying patterns) being modeled, either due to the passing of time or any event in the external (change in regulatory laws, change in competitor's lending policies) or internal environment (change in organization's lending policies).

Task 4: Summary of project for key stakeholders

The results of the execution of the KDDM process must be summarized for the key stakeholders. The management may not be interested in the detailed report generated through task 1, but only in the key findings of the KDDM process. The following items must be included in the report.

1. Business Objectives and Business Success Criteria
2. Data Mining Objectives and Data Mining Success Criteria
3. List of models that met the business objective, data mining objective, business success criteria and data mining success criteria, along with relevant details
4. Results of deployment on test sample
5. Results of deployment on overall population (if completed at this time)
6. Cost benefit analysis

Note again that the generation of such a summary can be easily semi-automated as following the steps in the IKDDM model, has already led to capturing of this relevant information.

Task 5: Lessons learned and feedback to preceding phases

As noted earlier, the KDDM process is iterative and reaching the deployment phase does not mean that the project can be considered as over. The cycle of knowledge discovery continues with feedback to different phases on the basis of events encountered during the execution of the KDDM process. This task of the deployment phase consists of reflecting on the tasks preceding it (in this phase and in other phases) and sending appropriate feedback that can help improve the execution of the preceding tasks in the future.

5.8 Schematic of the IKDDM Process Model

The IKDDM model has been developed to meet the design requirements outlined earlier. These are summarized in table below. The table also shows how the design requirements were addressed by the IKDDM model.

Table 5-37: Summary of Design Requirements Addressed by the IKDDM Model

Issues Identified with existing KDDM models	Design Requirements for the IKDDM model	How the Design Requirement was addressed?
Description of the KDDM Process in a Checklist Manner	Present a user-oriented coherent description of the KDDM process	Description of KDDM process is presented so as to provide guidance to the average business/technical user in executing the end-to-end process, not missing any step or stage of the process. Description of various tasks is followed by screenshots to show how the user can easily use the IKDDM model to understand the highly complex and iterative KDDM process
Fragmented View of the KDDM Process	Develop an integrated view of the KDDM process by explicating the various phase-phase and task-task dependencies	Each of the phases and their tasks have been studied to identify dependency relationships between tasks of the same and different phases
Emphasis on feedback loops prior to completely understanding the primary sequencing of phases and tasks in a KDDM process	Explicate sequencing of the various phases and their tasks before identifying feedback loops and establishing conditions under which the loops would get triggered	Each of the phases and their tasks has been carefully analyzed and the most optimal sequencing of tasks of different phases has been proposed. In some cases feedback loops have been identified, however these have been only been uncovered after explicating the primary sequencing (forward paths in the process model). Clearly explicating the primary sequencing has

288

		ensured that only necessary feedback loops have been retained, thereby ensuring optimal utilization of resources.
Fragmented view acts as a hindrance to building an integrated process model and "semi-automating" tasks	Leverage the dependencies explicated in the integrated process model to drive semi-automation of tasks, wherever possible	Various dependencies between tasks have been used to propose semi-automation of certain tasks. The tasks that have been semi-automated are not limited to the modeling phase.
Lack of support for the end-to-end KDDM process	Prescribe approaches for offering decision support towards all tasks in all phases, described in the integrated KDDM model	Every single one of the tasks outlined by the model has been supported through techniques/approaches for implementing it. Some of the approaches have been adapted from the literature to suit the context of the KDDM process. In other cases, the approach (in form of clearly defined sequence of steps) has been proposed by the IKDDM model itself.
Visible lack of support towards execution of tasks of the Business Understanding phase - the foundational phase of a KDDM process	Provide support for tasks of this foundational phase and use them as a basis for developing the integrated model	Given that the business understanding phase is the foundation of the KDDM process, it has been analyzed first, all tasks have been studied in detail (and their dependencies with tasks in the same phase and other phases identified), and support provided for executing each of the tasks outlined in this phase.

289

As stated in the beginning of the chapter, the IKDDM model was designed by detailed analysis of each of the phases of the KDDM process (business understanding, data understanding, data preparation, modeling and evaluation), their constituent tasks, dependencies between the tasks of the various phases and dependencies across phases (based on task-task dependencies between phases). A simultaneous focus is maintained on providing support for executing every single one of the tasks outlined by the model.

The dependencies between tasks of the same and different phases can sometimes be leveraged through semi-automation, speeding up the efficiency with which certain data mining tasks can be carried out. In other cases, support in form of tools or approaches is required for the execution of the tasks.

The discussion of individual phases contained description of dependency relationships of each task followed by a phase level process model. The schematic below shows the integrated view of the IKDDM model created by combining the various phases of the KDDM process. The schematic of the integrated schematic shows various phases and their tasks, as well as the dependencies between and across phases.
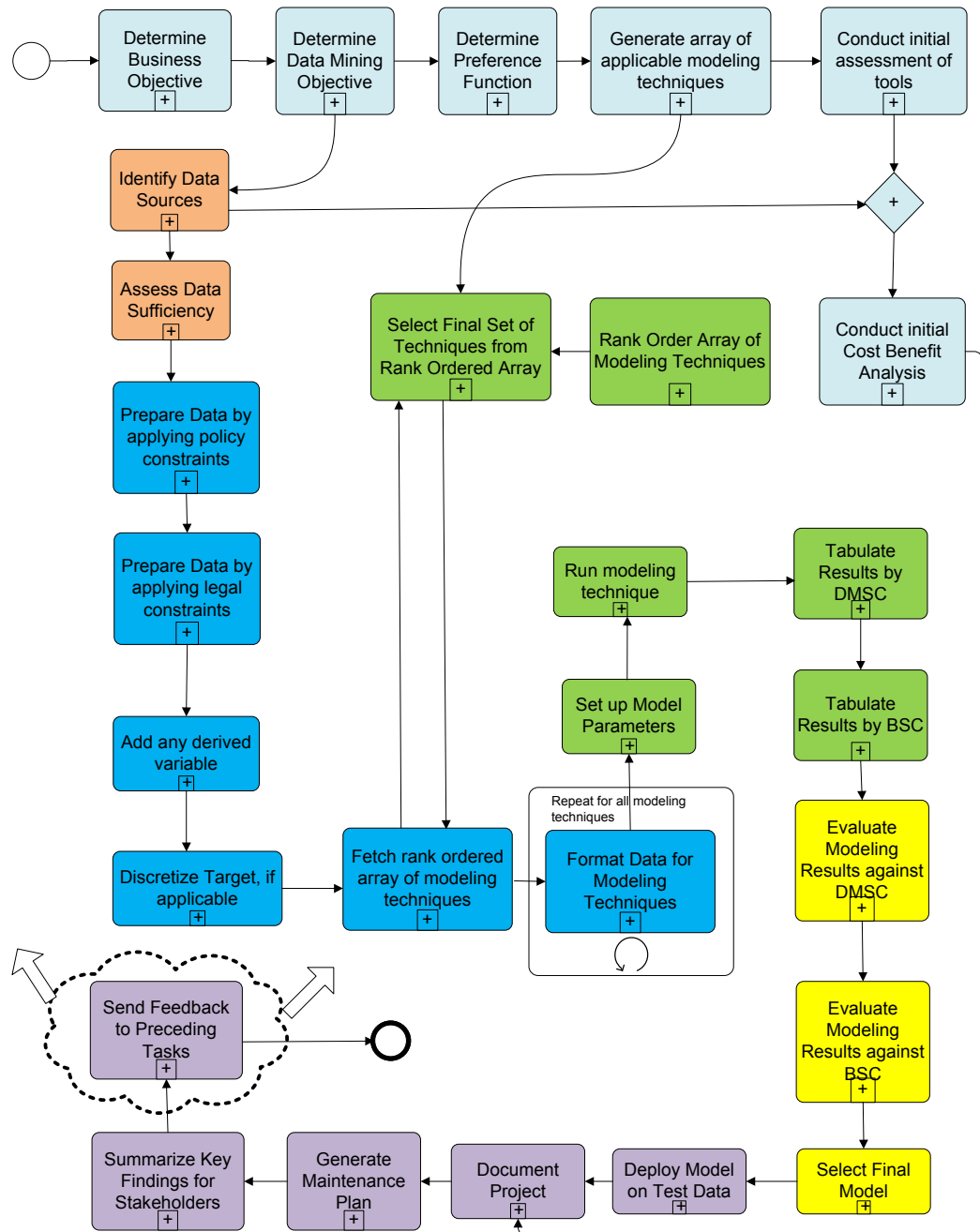
Figure 5-15: Overall schematic of IKDDM process model

6   EVALUATION OF THE IKDDM PROCESS MODEL

This chapter will describe the evaluation of the proposed KDDM model based on the guidelines prescribed by the Design Science Research Methodology (Hevner et al. 2004. The following evaluation methods are used:

- Analytical – Evaluation of the structure of the artifact for its static qualities
- Descriptive – Demonstration of utility of the artifact by constructing a detailed scenario

6.1   Analytical Testing

Analytical Testing comprises of the examination of the structure of artifact for static qualities such as ease of use, complexity, usability etc (Hevner et al. 2004). Clearly, prior to soliciting the input of users for analytical testing, the artifact (here the IKDDM process model) must first be made available to them for experimentation and use for executing data mining tasks. Additionally we wanted to compare the performance and static qualities of the artifact proposed in this dissertation (the IKDDM model) versus the performance and static qualities of a leading competing artifact, the CRISP-DM process model. The following methodology was selected for performing the analytical testing:

1. Identify and recruit 30 study participants and randomly divide them in two groups

2. Present one group of users with a test questionnaire, which includes data mining tasks posed as multiple choice questions. Provide them with the documentation of the CRISP-DM process model to assist in answering the questions (i.e. in executing tasks of a data mining project).

3. After the completion of the test questionnaire, record their perception of the static qualities of the artifact (i.e. the CRISP-DM process model) used by them through a set of survey questions.

4. Present the second group of users with the same test questionnaire, which includes data mining tasks posed as multiple choice questions. Provide them with the documentation of the IKDDM process model to assist in answering the questions (i.e. in executing tasks of a data mining project).

5. After the completion of the test questionnaire, record their perception of the static qualities of the artifact (i.e. the IKDDM process model) used by them through a set of survey questions.

6. Record each participant's gender, role/designation, number of years of experience in data mining, and time taken to complete the test. A numeric id will link the responder's test to the survey. No identifying detail, such as name of the participant, or name of the organization that the individual is affiliated with are to be recorded.

6.2   Statistical tests for evaluating the results of analytical testing

Independent Means t-test for comparing performance of IKDDM model versus CRISP-DM model

One of the goals of the evaluation was to compare the performance of the group that used the CRISP-DM model to answer the test questionnaire to that of the group that used the IKDDM model to answer the same test questionnaire. The performance of the two groups is a proxy for the effectiveness and reliability of the model used by them for answering the test. The results for each group will be computed by assigning a score of 2 points for every correct answer and 0 points for every incorrect answer.

The performance of the two groups (each with n = 21) will be compared using an independent mean t-test to determine if there was any statistical difference between the two groups. The statistical data analysis software SPSS 15 will be used for conducting the test. An overview of the independent means t-test test is included below.

Rationale for using Independent Means t-test

An independent mean t-test is used when there are two experimental conditions and different subjects were assigned to each situation. This test is also sometimes referred to as independent measures or independent samples t test. In contrast, a dependent means t-test is used when there are two experimental conditions but the same

subjects took part in both conditions of the experiment. This test is sometimes referred to as matched pairs or paired samples t-test.

In this study, there exist two experimental conditions, use of the CRISP-DM model or use of the IKDDM model to execute data mining tasks. Two different sets of individuals will be participating in each experimental condition. That is, each individual will either use the CRISP-DM model or the IKDDM model to execute the data mining tasks. Therefore, an independent means t-test was found to be appropriate for this scenario.

Steps for implementing Independent Means t-test

Field (2000) specifies following steps for conducting the independent means t-test:

1. Two samples of data are collected and the sample means are calculated. These means can the same, differ by either a little bit or a lot.

2. If the samples come from the same population, then we expect their means to be roughly equal. Under the null hypothesis, we assume that the "experimental manipulation has no effect on the subjects and therefore we expect the sample means to be identical or very similar".

3. The difference in sample means is compared to difference in sample means that we would expect to obtain by chance.

295

4. As the observed difference gets larger, the more confident we become that the null hypothesis should be rejected.

5. If the null hypothesis is incorrect then we can conclude that the two sample means differ because of the experimental manipulation imposed on each sample

The general equation for a t-test is,

t = (observed difference between sample means) − (expected difference when null hypothesis is true) / estimate of standard error

In mathematical notation, it can be expressed as:

$$t = (\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2) / estimate\ of\ standard\ error$$

Equation 1: t test – general equation

The null hypothesis is that $\mu_1 = \mu_2$ and therefore $\mu_1 - \mu_2 = 0$

$$t = (\overline{X}_1 - \overline{X}_2) / estimate\ of\ standard\ error$$

Equation 2: t –test – equation for independent means t-test

The standard error can be estimated as follows:

SE of sampling distribution of population 1 = $\dfrac{s_1}{\sqrt{N_1}}$

SE of sampling distribution of population 2 = $\dfrac{S_2}{\sqrt{N_2}}$

Since variance is simply the standard deviation squared, we can calculate the variance of each sampling distribution:

Variance of sampling distribution of population 1 = $\left(\dfrac{S_1}{\sqrt{N_1}}\right)^2 = \dfrac{S_1^2}{N_1}$

Variance of sampling distribution of population 2 = $\left(\dfrac{S_2}{\sqrt{N_2}}\right)^2 = \dfrac{S_2^2}{N_2}$

The variance sum law means that to find the variance of the sampling distribution of differences we merely add together the variances of the sampling distributions of the two populations:

Variance of the sampling distribution of differences = $\dfrac{S_1^2}{N_1} + \dfrac{S_2^2}{N_2}$

To find out the standard error of the sampling distribution of differences we merely take the square root of the variance (because variance is the standard deviation squared):

SE of the sampling distribution of differences = $\sqrt{\left(\dfrac{S_1^2}{N_1} + \dfrac{S_2^2}{N_2}\right)}$

Substituting value in equation 1, t becomes,

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)}}$$

Mann-Whitney Test for comparing difference in groups' perception about static qualities of KDDM process models

As stated earlier, the static qualities of the KDDM process model employed by the users to execute the data mining tests (in the test questionnaire) will be assessed through a set of survey questions with 7 point Likert-scale options ranging from Strongly Agree to Strongly Disagree.

The goal of the evaluation is to determine any difference in user's perception of the static qualities (such as perceived usefulness, ease of use etc.) of CRISP process model versus the IKDDM process model.

Rationale for Using Mann-Whitney Test

The data generated from the survey is in Likert scale form. Such data violates the assumptions of parametric tests that assume that the underlying data is interval or ratio in nature. A non-parametric test (sometimes referred to as an assumption-free test) makes no assumptions about the data on which they can be used. It is used for testing differences between means when there are two conditions and different subjects have been used in each condition.

Field (2000) points out that this ingenuity comes at a price: since non-parametric tests work by ranking the data, they lose information about the magnitude of difference between scores, making them less effective at detecting effects as compared to parametric tests. When using parametric tests there could be an increased chance of type-II error (i.e. more chance of accepting there is no difference when in a reality a difference exists).

However owing to the fact that Likert-scale data violates the assumptions of parametric tests, this dissertation employs the non-parametric test - Mann-Whitney for determining differences between groups when studying the survey data. We acknowledge that parametric tests such as MANOVA are frequently used by researchers to determine differences between groups, even when data is generated from Likert Scale and is therefore in ordinal form. Accordingly we refer the interested reader to the Appendix for the results of MANOVA performed using the Likert-Scale survey data from the participants. The study satisfies the test's assumption that there are two conditions (use of CRISP versus IKDDM) and different subjects have been assigned to each condition.

Pilot test of Test Questionnaire and Survey

Prior to conducting the actual evaluation, a pilot test of the test questionnaire and survey was conducted. 4 users with expertise in data mining participated in the pilot test. The average number of years of data mining experience of these users was 4 years. The following approach was adopted for conducting the pilot test:

- The four users were randomly divided into two groups of two users each.

- Each user was provided with a multiple choice test questionnaire consisting of 15 questions. The questions were based on typical tasks included in data mining projects, such as determination of business and data mining objectives, determination of data mining success criteria, selection of appropriate modeling techniques, verifying assumptions of data mining modeling techniques, evaluation of modeling results etc.

- The users in the group, labeled the CRISP-DM$_{pilot}$ were provided with the extract documentation of the CRISP-DM process model. The extract document was created from the user guide portion of the CRISP-DM process model and contained relevant portions from the model for answering each of the questions. A copy of the CRISP-DM extract documentation is included in the Appendix.

- The users in the group, labeled the IKDDM$_{pilot}$ were provided with the extract documentation of the IKDDM process model. The extract document was created from the design of the IKDDM model as described in Chapter 4 of this dissertation

300

and contained relevant portions from the model for answering each of the questions. A copy of the IKDDM extract documentation is included in the Appendix.

- The users were asked to use the extract documentation of the model provided to them in answering each of the questions. They were also asked to report on:

1. Adequacy of coverage of the tasks presented in the test questionnaire

2. The wording of the questions/options

3. Time taken by them to answer the test questionnaire.

- Once the users returned completed the test questionnaire, they were sent a survey with 16 questions to assess their perception of the static qualities of the model used by them in answering the questions. The survey instrument has been adapted from instrument for measuring quality of process model proposed by Maes and Poels (2006). Their instrument defines quality of a process model along four dimensions: perceived ease of use, perceived usefulness, user satisfaction and perceived semantic quality. These dimensions are examples of static qualities and can be used for assessing these qualities in the CRISP-DM and IKDDM models.

- Once the user's have experienced the artifact (CRISP-DM or proposed KDDM model), they will be asked to answer questions pertaining to the static qualities of the artifact. The measurement instrument for measuring conceptual model quality, proposed by Maes and Poels (2006) will be used for this purpose. Their instrument assesses qualities such as perceived ease of use, perceived usefulness, user satisfaction and perceived semantic quality. The wording of the items in the original

301

instrument has been modified to include the term 'KDDM process model' instead of the term 'conceptual model' that is part of the original instrument. No other changes have been made. The measurement instrument is shown in Table . More details can be found in the Research Methodology chapter (chapter 4) of this dissertation.

▪ The test questionnaire used in the evaluation is included in the Appendix

Table 6-1: Measurement Instrument for Assessing Quality of Process Models Proposed by Maes and Poels (2006)

PEOU1	It was easy for me to understand what the KDDM model was trying to model.	PU1	Overall, I think the KDDM model would be an improvement to a textual description of the KDDM process.
PEOU2	Using the KDDM model was often frustrating.	PU2	Overall, I found the KDDM model useful for understanding the process modeled.
PEOU3	Overall, the KDDM model was easy to use.	PU3	Overall, I think the KDDM model improves my performance when understanding the process modeled.
PEOU4	Learning how to read the KDDM model was easy.	PSQ1	The KDDM model represents the KDDM process correctly.
US1	The KDDM model adequately met the information needs that I was asked to support.	PSQ2	The KDDM model is a realistic representation of the KDDM process.
US2	The KDDM model was not efficient in providing the information I needed.	PSQ3	The KDDM model contains contradicting elements.
US3	The KDDM model was effective in providing the information I needed.	PSQ4	All the elements in the KDDM model are relevant for the representation of the KDDM process
US4	Overall, I am satisfied with the KDDM model for providing the information I needed.	PSQ5	The KDDM model gives a complete representation of the KDDM process

PEOU: Perceived Ease of Use   PU: Perceived Usefulness

PSQ: Perceived Semantic Quality   US: User Satisfaction

Results of the Pilot Test

On the basis of feedback from the users the test questionnaire was slightly revised, and a final version was created for use in the actual evaluation. It was also determined on the basis of feedback received from the pilot test that a time limit such as 1 or 2 hours should not be imposed, but rather that the users be provided with the test and survey at the beginning of the business day and be asked to return it by the end of the business day. They should still be asked to record the time when they started the test and the time when they had completed both the test and the survey.

The feedback received was also used to refine the extract documentations for both the models. At the time of the pilot the extract documentation for CRISP model was at 26 pages, and that of IKDDM was at 19 pages. Both of the extract documents were revised to remove information that was not directly relevant for answering the test questions. All pertinent information was retained for both models, but the exercise helped in bringing down the number of pages in both the models. The final version of the extract documentation provided had 11 pages for the CRISP model and 11 pages for the IKDDM model.

Analysis of Performance of CRISP-DM$_{pilot}$ versus IKDDM$_{pilot}$ on Test Questionnaire

Mean Accuracy Rate

- CRISP-DM$_{pilot}$ = 11

- IKDDM$_{pilot}$ = 18

Analysis of user's perception of static qualities of process model of CRISP-DM$_{pilot}$ versus IKDDM$_{pilot}$

The analysis of static qualities of the process model was accomplished using the survey instrument described above. The scoring technique for survey responses is presented in Table 6-2.

Table 6-2: Scoring technique Used for Likert-Scale Based Survey Items

Strongly disagree	
Disagree	
Moderately disagree	
Undecided	
Moderately Agree	
Agree	
Strongly agree	

The overall score of users on survey aimed at assessing their perception of the quality of the process model used by them in executing tasks of the KDDM process are presented below (Table 6-3)

Table 6-3: Pilot Test: Survey Scores of Expert Users

	CRISP-DM$_{pilot}$	IKDDM$_{pilot}$
User 1	46	85
User 2	73	75

Assessment of artifact by Users with Experience in Data Mining

Following the approach described earlier, the artifact, i.e. a KDDM process model and its extract documentation was made available to individuals with experience in data mining. They were asked to use the artifact by applying it to execute the various tasks of a hypothetical data mining project aimed at reducing churn at a telecommunications company. 42 individuals with varying levels of experience in data mining participated in the study. IRB approval was sought prior to conducting this study (*Ref Number HM 11636*). Based on the IRB guidelines, each participant was presented with a consent form, prior to soliciting their input through the test and the survey.

21 users were randomly assigned to use the leading KDDM process model, CRISP-DM to answer the various questions related to the data mining project whereas 21 users were randomly assigned to use the IKDDM model to answer the various

questions. Hereafter the two groups are referred to as CRISP-DM$_{eval}$ and IKDDM$_{eval}$ respectively.

The following information was recorded for each participant:

- Date on which data was collected from the individual
- Participant's Gender
- Participant's Role/Title
- Participant's number of years of data mining experience
- Start Time for the test
- End Time for the test

The start and end times for the test were used to estimate the total time taken by the participants to answer the test. The summary of participant's profile based on gender, years of data mining experience and the time taken by participants is tabulated below.

Table 6-4: Summary of participant's profile

	CRISP$_{eval}$ (N=21)	IKDDM$_{eval}$ (N=21)
Gender distribution	28.5 % females 71.4 % males	23.8 % females 76.1 % males
Average years of data mining	2.5 years	2.6 years

experience		
Average time taken to answer the test	36.52 minutes	31.38 minutes

Prior to running any tests and interpreting results, an assessment of validity of the measurement instrument was conducted. The methodology for conducting the assessment is described in the next section.

6.3   Assessment of validity of measurement instrument

The measurement instrument used in this research has been adopted from Maes and Poels (2007). They proposed an instrument to measure the quality of conceptual models and tested hypotheses pertaining to relationships between four constructs: perceived ease of use, perceived usefulness, user satisfaction and perceived semantic quality. This research made use of this instrument to assess the perception of users about the quality of the process model used by them (CRISP or IKDDM) to execute tasks in data mining. Unlike Maes and Poels (2007) our goal was not to test any structural model or hypotheses after validating the instrument. Nevertheless, it is important to assess the validity of the measurement instrument and if the results appear to be in line with recommendations.

We conducted the validity assessments in Smart-PLS (Ringle, Wende et al. 2005). Following Maes and Poels we conducted separate validity assessments for the reflectively (PEOU, PU, US) and formatively modeled (PSQ) construct. In Smart-PLS software, results of path analysis include factor loadings for reflective constructs and weights for formative constructs.

307
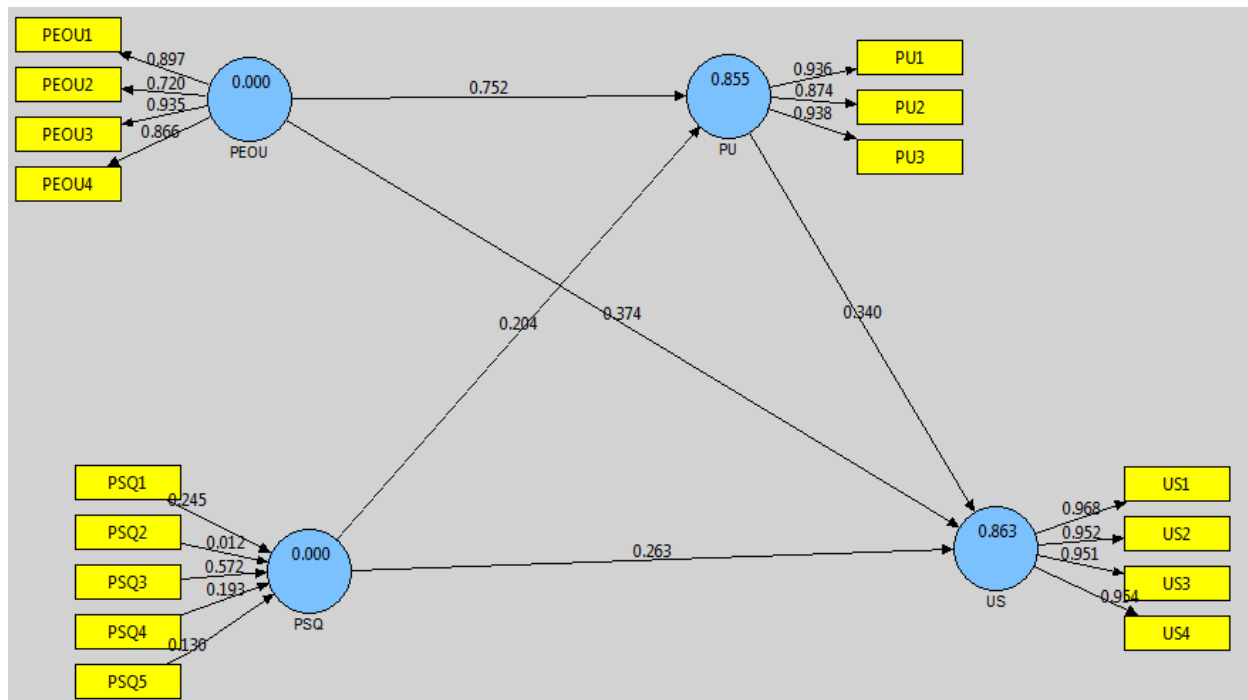
Figure 6-1: Path Model showing loadings for reflective constructs (PEOU, US, PU) and weights for formative construct (PSQ)

Validity assessments of Reflective Constructs: Perceived Ease of Use, Perceived Usefulness, User satisfaction

Table 6-5: AVE, Composite Reliability and Cronbach's Alpha

	AVE	Composite Reliability	R Square	Cronbachs Alpha
PEOU	0.737	0.9174	0	0.8787
PU	0.84	0.9402	0.8554	0.9045

308

US	0.9146	0.9772	0.8629	0.9689

The results obtained from testing the measurement model provide evidence of the robustness of the measures as indicated by their internal consistency reliabilities (indexed by the composite reliabilities). The composite reliabilities of the measures range from 0.917 to 0.972. All of these reliabilities exceed the recommended threshold of 0.70 suggested by Nunnally (Nunnally 1978). The reliability can also be confirmed through the values for Cronbach's alpha, ranging from 0.878 to 0.968, which exceed the minimum threshold of 0.7. These are shown in table above. Also, the average variances extracted (AVEs) for the measurement constructs range from 0.737 to 0.914 Consistent with the recommendation of Fornell and Larcker (Fornell and Larcker 1981), the AVE for each measure well exceeds the lower bound threshold value of 0.50.

Factor loadings

Table 6-6: Factor cross loadings

	PEOU	PU	US
PEOU1	**0.8975**	0.8514	0.7683
PEOU2	**0.7201**	0.6137	0.5426
PEOU3	**0.9351**	0.8813	0.9152
PEOU4	**0.8658**	0.7719	0.8087
PU1	0.889	**0.9363**	0.8816
PU2	0.7761	**0.8738**	0.729
PU3	0.8504	**0.938**	0.8456
US1	0.8606	0.8822	**0.9676**
US2	0.8153	0.8054	**0.9522**

US3	0.8481	0.8644	**0.9511**
US4	0.9101	0.8763	**0.9545**

Finally, to complete the psychometric assessment of our measurement model discriminant validity was examined. Discriminant validity refers to the extent to which the items proposed to measure a given construct differ from the items intended to measure other constructs in the same model. A cross-loading check indicated that all items loaded higher on the construct they were supposed to measure than on any other construct. A common rule of thumb to indicate convergent validity is that all items should load greater than 0.7 on their own construct, and should load more highly on their respective construct than on the other constructs (Yoo and Alavi 2001). Furthermore, each item's factor loading on its respective construct was highly significant ($p < 0.01$). This was true for items for all reflective constructs. Another means of assessing the discriminant validity is using the factor correlations and AVE. evidence of discriminant validity is found if the square root of AVE is greater than the factor correlations. The factor correlations matrix is a symmetric matrix with 1 along the diagonals (correlation of a factor with itself is 1). This is presented in table.

Table 6-7: Factor correlations matrix

	PEOU	PU	US
PEOU	1	0	0

310

PU	0.917	1	0
US	0.8987	0.8969	1

The method for conducting analysis of dicriminant validity consists of replacing the diagonal elements by the square root of the variance, and assessing if this value is greater than the factor's correlation with other factors.

Table 6-8: Assessment of discriminant validity (replacing diagonals of factor correlations matrix with square root of AVE)

	PEOU	PU	US
PEOU	0.858487	0	0
PU	0.917	0.916515	0
US	0.8987	0.8969	0.956347

In this case, it can be seen that discriminant validity holds true for all factors, except for PU or perceived usefulness because the square root of PU is the same as the correlation between PU and PEOU. From this analysis it appears that these two factors are not distinct, however the cross loadings confirm discriminant validity.

Validity assessments of Formative Construct: Perceived Semantic Quality

311

Because of the formative structure of the PSQ construct, traditional validity assessments can not be used (Diamantopoulos and Winklhofer 2001). Observed correlations among the items may not be meaningful (Diamantopoulos and Winklhofer 2001) and as a consequence, assessment of internal consistency and convergent validity become irrelevant (Chin 1998; Hulland 1999). The PSQ measure can be considered as valid if the PSQ indicator coefficients are significantly different from zero (Diamantopoulos and Winklhofer 2001). This can be determined by running a bootstrapping procedure in Smart-PLS. The output of the path model shows the values for t-statistic for all paths and coefficients.
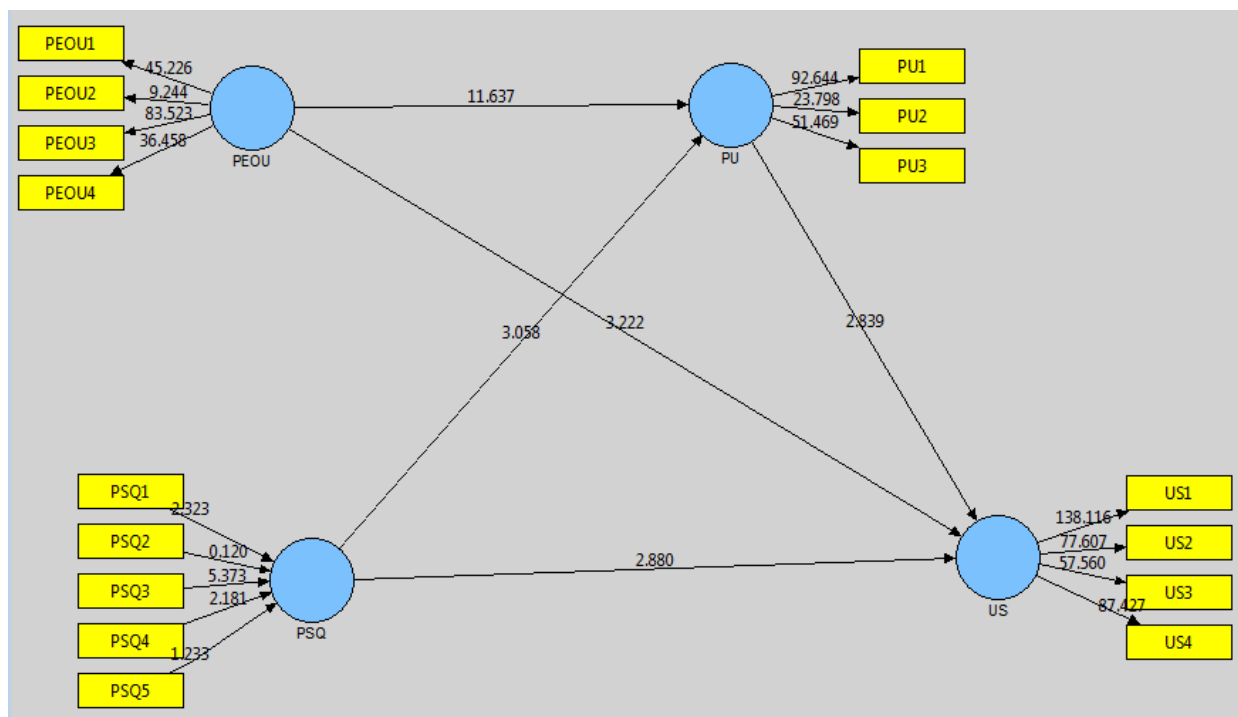
Figure 6-2: Output of Bootstrapping t-statistics for indicator coefficients and paths

PLS analysis indicates that not all PSQ indicators have a coefficient significantly different from zero (t>2). Such indicators should be deleted from the model if a structural model is to be tested. In this sample, PSQ1, PSQ3, and PSQ4 turned out to be significantly different from zero, but PSQ2 and PSQ5 are not significantly different from zero. On the basis of these results it appears that only PSQ1, PSQ3, and PSQ4, are relevant formative indicators of perceived semantic quality.

Table 6-9: Weights and t-values for formative indicators

	Weight	t-statistic	Significant?
PSQ1 -> PSQ	0.2451	2.3228	Significant
PSQ2 -> PSQ	0.0123	0.1205	Non significant
PSQ3 -> PSQ	0.5723	5.373	Significant
PSQ4 -> PSQ	0.1935	2.1812	Significant
PSQ5 -> PSQ	0.1304	1.2334	Non significant

6.4 Independent means t-test to assess differences based on gender distribution, years of data mining experience, and time taken

We also ran independent means t-test to assess if there were any differences between the two groups based on the gender distribution, years of data mining experience or

time taken to answer the test. The group statistics output by the t-test (Table 6-10) shows the summary statistics for the two experimental conditions.

Table 6-10: Group Statistics (comparing groups on the basis of gender distribution, years of data mining experience, and time taken to answer the test)

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
YRSOFEXP	CRISP	21	2.50	2.012	.439
	IKDDM	21	2.68	2.645	.577
TIMETAKEN	CRISP	21	36.52	17.180	3.749
	IKDDM	21	31.38	11.805	2.576
GENDER	CRISP	21	.29	.463	.101
	IKDDM	21	.24	.436	.095

314

Table 6-11: Independent means t-test (comparing groups on the basis of gender distribution, years of data mining experience, and time taken to answer the test)

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Upper	Lower
YRSOFEXP	Equal variances assumed	1.218	.276	-.249	40	.804	-.181	.725	-1.647	1.285
	Equal variances not assumed			-.249	37.342	.804	-.181	.725	-1.650	1.288
TIMETAKEN	Equal variances assumed	6.263	.017	1.131	40	.265	5.143	4.549	-4.050	14.336
	Equal variances not assumed			1.131	35.442	.266	5.143	4.549	-4.087	14.373
GENDER	Equal variances assumed	.471	.496	.343	40	.733	.048	.139	-.233	.328
	Equal variances not assumed			.343	39.862	.733	.048	.139	-.233	.328

The second table shows the actual test statistics. There are two rows containing values for test statistics: one row is labeled equal variances assumed, whereas other is labeled equal variances not assumed. Parametric tests assume that variances in experimental groups are roughly equal. The Levene's test tests the hypothesis that the

315

variances in the two groups are roughly equal (i.e. the difference between variances is zero). If Levene's test is significant, then the null hypothesis is incorrect and we have to conclude that the variances are significantly different. If, however, Levene's test is non-significant, then it can be concluded that the differences in variances is zero and the assumption of equal variances is tenable. For our data, the Levene's test is not significant for years of data mining experience (YRSOFEXP) or for Gender. The p values for these variables are 0.276 and 0.496 respectively which are greater than 0.05 and so we can read the test statistics in the row labeled equal variances assumed (Table 6-11). The 2-tailed significance for years of experience is 0.804 and for gender is 0.733, both of which are non-significant. We can therefore conclude that there were no significant differences between the groups on the basis on number of years of experience or the gender distribution of the sample.

Referring to the table again, Levene's test is significant for time taken to answer the test (TIMETAKEN). The p value for this variable is 0.017 which is smaller than 0.05, and therefore it can be concluded that the assumption of equal variables is not tenable. This means that we must read the statistics in the row labeled 'equal variances not assumed'. The 2-tailed significance for time taken is 0.266 which is non-significant. We can therefore conclude that there also no significant differences between the groups in terms of time taken to answer the test.

6.5 Results of Independent Means t-test – Analysis of performance

Analysis of Performance of CRISP-DM$_{eval}$ versus IKDDM$_{eval}$ on Test Questionnaire: Using Independent Mean t-test

The performance of the participants in the two groups (CRISP-DM versus IKDDM) was measured by calculating the accuracy of their response. An independent means t-test was used for determining the statistical difference in performance between the two groups. SPSS 15 was used for running the t-test. The sequence of steps followed are shown in screenshots below.

317

Setting up Independent Means t-test in SPSS (step 1 of 2)

Setting up Independent Means t-test in SPSS (step 2 of 2)

The group statistics output by the t-test (Table 6-12) shows the summary statistics for the two experimental conditions. From this table, we can see that both groups had 21 subjects. The group that was assigned to the CRISP model had a mean score of 12.67, whereas the group that was assigned to IKDDM had a mean score of 26.57.

Table 6-12: Group statistics: Independent means t-test

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
TESTSCORE	CRISP	21	12.67	3.967	.866
	IKDDM	21	26.57	2.908	.635

The second table shows the actual test statistics. There are two rows containing values for test statistics: one row is labeled equal variances assumed, whereas other is labeled equal variances not assumed. Parametric tests assume that variances in experimental groups are roughly equal. The Levene's test tests the hypothesis that the variances in the two groups are roughly equal (i.e. the difference between variances is zero). If Levene's test is significant, then the null hypothesis is incorrect and we have to conclude that the variances are significantly different. If, however, Levene's test is non-significant, then it can be concluded that the differences in variances is zero and the assumption of equal variances is tenable. For our data, the Levene's test is not significant (p = 0.107) which is greater than 0.05 and so we can read the test statistics in the row labeled equal variances assumed (Table 6-13).

Table 6-13: Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Upper	Lower
TESTSCORE	Equal variances assumed	2.726	.107	-12.955	40	.000	-13.905	1.073	-16.074	-11.736
	Equal variances not assumed			-12.955	36.681	.000	-13.905	1.073	-16.080	-11.729

Having established the assumption of homogeneity of variances we can look at the t-test itself. SPSS produces exact significance value for t and we are interested in whether or not this value is less than or greater than 0.05.

In this case the two tailed value of p is .000, which is much smaller than 0.05, and therefore we can conclude that there was a highly significant difference ($p = 0.000$) between the performance of the group that used the IKDDM model to execute data mining tasks versus the group that used the CRISP model to execute the same set of tasks.

The sample for both IKDDM and CRISP group included few naïve users; specifically the CRISP group had 5 naïve users whereas IKDDM group had 6 naïve users. Given the small number of naïve users, their performance cannot be separately assessed through a procedure like the independent means t-test, and so we instead compare their mean accuracy rate to gain insights into their relevant performance. These are presented in Table 6-14.

Table 6-14: Mean Accuracy Rate of Naïve Users

Naïve User	CRISP	IKDDM
1	6	30
2	12	26
3	14	26
4	6	28
5	10	30
6	N.A.	26
Mean Accuracy Rate of naïve users in each group	**9.6**	**28**
Maximum possible points	**30**	**30**

321

6.6 Discussion of Results of Independent Means t-test

The results of the Independent Means t-test confirms that the IKDDM group outperformed the CRISP group in terms of its performance on the test which asked users to utilize the process model assigned to them to execute data mining tasks. Since the tasks were formulated as multiple choice questions with only one correct answer, the performance of users in both the groups could be estimated using the accuracy of their responses. The performance provides insights into the effectiveness and efficiency offered by the IKDDM model over the CRISP model.

We also compared the mean accuracy rate of naïve users on the test to estimate how accurately they executed the tasks of the KDDM process. The mean accuracy rate of naïve users (users with 0 years of data mining experience) in the IKDDM group was 28 and was much higher than the mean accuracy rate of 9.6 obtained by naive users in the CRISP group. This is also an important finding and indicates that the IKDDM model was equally effective in supporting the information needs of the naïve users as well as experienced users and allowed for effective and efficient implementation of tasks by both types of users.

6.7 Results of Mann-Whitney Test

Analysis of perception about static qualities of process model of CRISP-DM$_{eval}$ versus IKDDM$_{eval:}$ Using Mann-Whitney Test

Figure 6-3 and Figure 6-4 show how the test was set up in SPSS 15. The Mann-Whitney test works by looking at differences in the ranked positions of scores in different groups. The first part of the output, shown in the Ranks table (Table 6-15), shows the average and total ranks for each condition. The group with the lowest means rank is also the group with the greatest number of lower scores in it. In the context of this study, the group with the lowest means rank is the group that was assigned to use the CRISP process model.

It can be seen that IKDDM (group 2) fared significantly better than the CRISP model in terms of user's perception of the quality of process model.
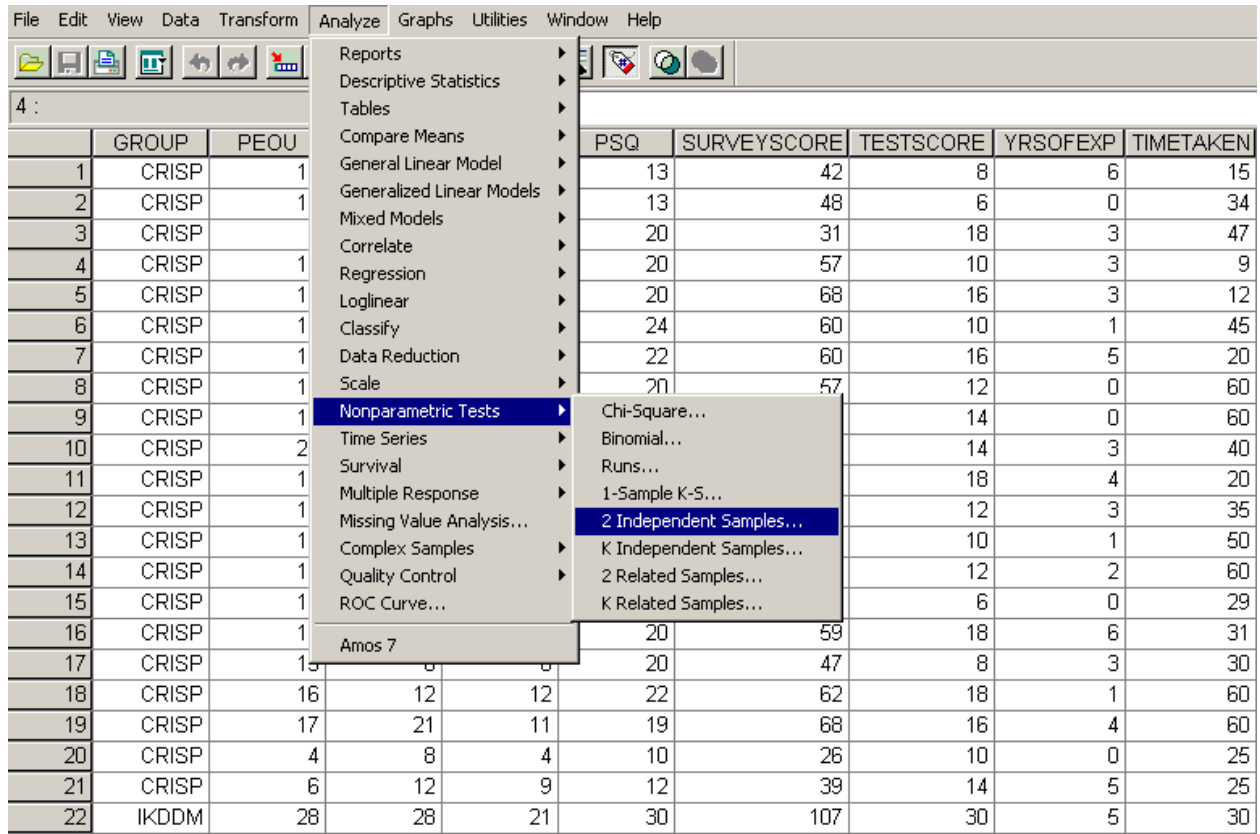
Figure 6-3: Setting up Mann-Whitney Test in SPSS (step 1 of 2)
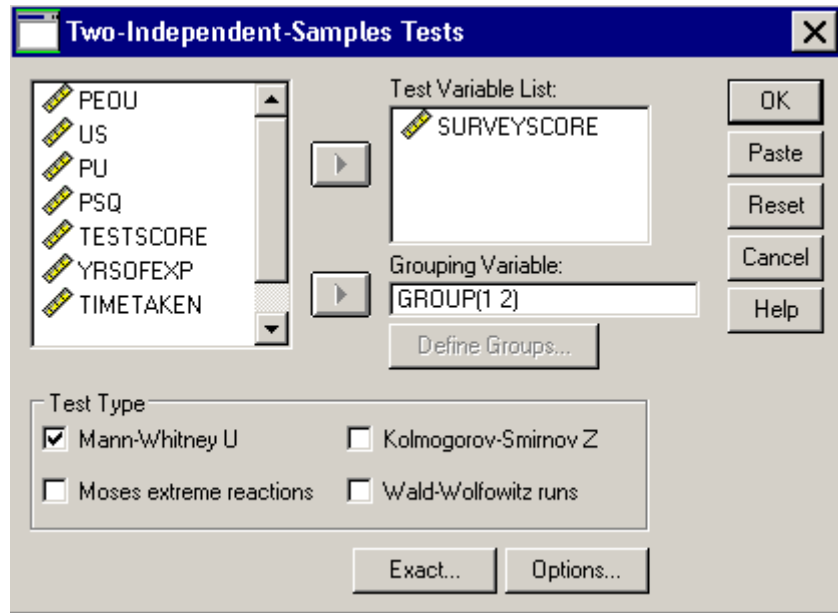
Figure 6-4: Setting up Mann-Whitney Test in SPSS (step 2 of 2)

Table 6-15: Ranks Table for Mann Whitney Test (N=42)

	GROUP	N	Mean Rank	Sum of Ranks
SURVEYSCORE	CRISP	21	11.76	247.00
	IKDDM	21	31.24	656.00
	Total	42		

The second table shows the actual test statistics for the Mann-Whitney test. The SPSS output has a column for the dependent variable (here, the survey score), and rows showing the value of Mann Whitney's U statistic, Wilcoxon's W statistic, and the associated z approximation. The table also contains the significance value of the test

www.manaraa.com

which gives the two-tailed probability that the magnitude of the test statistic is a chance result. For this test, the Mann-Whitney test is highly significant ($p<0.0001$) for the survey scores of the two groups (Table 6-16). The value of the means rankings indicates that the quality of the IKDDM process model was rated as significantly higher than the quality of the CRISP process model. This conclusion is reached by noting that for the survey scores representing model quality, the average rank is higher in the IKDDM group (31.24) than in the CRISP group (11.76).

Table 6-16: Test Statistics for Mann-Whitney (N=42)

	SURVEYSCORE
Mann-Whitney U	16.000
Wilcoxon W	247.000
Z	-5.146
Asymp. Sig. (2-tailed)	.000

a Grouping Variable: GROUP

6.8   Results of Mann Whitney Test to assess difference between groups on individual constructs

The Mann Whitney test was also used to assess if there were differences between the two groups (CRISP versus IKDDM) when the four constructs: perceived

326

ease of use, user satisfaction, perceived usefulness, and perceived semantic quality were analyzed separately. The earlier test, established that a significant difference existed between the groups on the combined score on the survey but did not tell us if this was true for each construct as well. The test was set up the same way, except the scores on items for the four different constructs were summed up for each of the two groups and differences examined. The results are shown below. These have been interpreted in the same manner as the results in the previous section.

Results for Perceived Ease of Use

The Mann-Whitney test is highly significant (p<0.0001) for the perceived ease of use scores of the two groups (

327

Table 6-17). The value of the means rankings indicates that the perceived ease of use of the IKDDM process model was rated as significantly higher than the perceived ease of use of the CRISP process model (Table 6-18). This conclusion is reached by noting that for the survey scores representing perceived ease of use, the mean rank is higher in the IKDDM group (30.98) than in the CRISP group (12.02)

Table 6-17: Ranks Table for Mann Whitney (comparing groups on perceived ease of use)

	GROUP	N	Mean Rank	Sum of Ranks
PEOU	CRISP	21	12.02	252.50
	IKDDM	21	30.98	650.50
	Total	42		

Table 6-18: Test Statistics for Mann Whitney (comparing groups on perceived ease of use)

	PEOU
Mann-Whitney U	21.500
Wilcoxon W	252.500
Z	-5.015
Asymp. Sig. (2-tailed)	.000

a Grouping Variable: GROUP

329

The Mann-Whitney test is highly significant (p<0.0001) for the user satisfaction scores of the two groups (Table **6-20**). The value of the means rankings indicates that the user satisfaction with the IKDDM process model was rated as significantly higher than the user satisfaction with the CRISP process model (Table 6-19). This conclusion is reached by noting that for the survey scores representing user satisfaction, the mean rank is higher in the IKDDM group (30.67) than in the CRISP group (12.33)

Table 6-19: Ranks Table for Mann Whitney (comparing groups on user satisfaction)

	GROUP	N	Mean Rank	Sum of Ranks
US	CRISP	21	12.33	259.00
	IKDDM	21	30.67	644.00
	Total	42		

Table 6-20: Test Statistics for Mann Whitney (comparing groups on user satisfaction)

	US
Mann-Whitney U	28.000
Wilcoxon W	259.000
Z	-4.860
Asymp. Sig. (2-tailed)	.000

a Grouping Variable: GROUP

Results for Perceived Usefulness

The Mann-Whitney test is highly significant (p<0.0001) for the perceived usefulness scores of the two groups (Table 6-22). The value of the means rankings indicates that the perceived usefulness of the IKDDM process model was rated as significantly higher than the perceived usefulness of the CRISP process model (Table 6-21). This conclusion is reached by noting that for the survey scores representing perceived usefulness, the average rank is higher in the IKDDM group (31.48) than in the CRISP group (11.52)

Table 6-21: Ranks Table for Mann Whitney (comparing groups on perceived usefulness)

	GROUP	N	Mean Rank	Sum of Ranks
PU	CRISP	21	11.52	242.00
	IKDDM	21	31.48	661.00
	Total	42		

Table 6-22: Test Statistics for Mann Whitney (comparing groups on perceived usefulness)

	PU
Mann-Whitney U	11.000
Wilcoxon W	242.000
Z	-5.294
Asymp. Sig. (2-tailed)	.000

a Grouping Variable: GROUP

Results for Perceived Semantic Quality

The Mann-Whitney test is highly significant (p<0.0001) for the perceived semantic quality scores of the two groups (Table 6-24). The value of the means rankings indicates that the perceived semantic quality of the IKDDM process model was rated as significantly higher than the perceived semantic quality of the CRISP process model (Table 6-23). This conclusion is reached by noting that for the survey scores representing semantic quality, the mean rank is higher in the IKDDM group (29.60) than in the CRISP group (13.40)

Table 6-23: Ranks Table for Mann Whitney (Comparing Groups on Perceived Semantic Quality)

	GROUP	N	Mean Rank	Sum of Ranks
PSQ	CRISP	21	13.40	281.50
	IKDDM	21	29.60	621.50
	Total	42		

Table 6-24: Test Statistics for Mann Whitney (Comparing Groups on Perceived Semantic Quality)

	PSQ
Mann-Whitney U	50.500
Wilcoxon W	281.500
Z	-4.319
Asymp. Sig. (2-tailed)	.000

a Grouping Variable: GROUP

6.9 Discussion of Results of Mann-Whitney Test

The results of Mann-Whitney test on overall survey scores representing quality of the process model indicate that a significant difference existed between the CRISP and IKDDM models. The test results clearly indicate that the IKDDM model outperformed the CRISP model by a highly significant margin ($p<0.001$). This is an important result and signifies that users rated the efficacy of IKDDM model as much higher than the CRISP model. The results of Mann-Whitney test across the four constructs also indicated that the IKDDM group and CRISP group significantly differed in their perceptions of ease of use, usefulness, semantic quality and levels of user satisfaction of the model employed by them to execute tasks in data mining. The IKDDM group reported significantly higher levels of perceived ease of use, perceived usefulness, semantic quality and user satisfaction as compared to the CRISP group.

The results confirm that IKDDM is more effective and efficient than the CRISP model in executing tasks of the KDDM process. The limitations of existing KDDM process models (such as use of only a checklist approach, or lack of explicit support towards execution of tasks) as identified in this research are certainly also perceived as problematic by the data mining users.

In keeping with the essence of design science research, the present design of the artifact can only be regarded as a "satisfactory solution" (Simon 1996). However the initial results of testing of IKDDM against CRISP (a leading model which is the most detailed of existing models) has generated promising results. These can be regarded as a

333

measure of the significance of the designed artifact, and its contribution to the existing knowledge base.

6.10   Descriptive Testing

Descriptive testing can be performed by the construction of detailed scenarios around the artifact to demonstrate the artifact's utility. This dissertation presents a detailed scenario around the IKDDM model to illustrate how the proposed model could be used for implementing an illustrative data mining project.

The construction of the detailed scenario includes various tasks ranging from business understanding phase to the evaluation phase. Bank loan data set from SPSS Clementine v 12.0 has been used for the construction of the scenario

Background

The scenario described in this dissertation is based on how a bank uses data mining to make decisions regarding granting of loans to applicants. Essentially, the bank wishes to identify customers whose loan request should be granted, and those whose loan request application should be denied.

We use the steps recommended by the IKDDM model to execute each of the tasks, starting from formulation of business objectives to evaluation of results. The steps are categorized under the phases of the IKDDM model: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

1. Formulation of Business Objective

Creation of preliminary business objective using an adaptation of GQM approach: the IKDDM model suggests a modified GQM based approach to assist in creating the preliminary statement of business objectives. The various steps recommended by the approach are implemented below to formulate the preliminary business objective based on the project.

Step 1: *Selection of Purpose:* The stakeholders discuss the purpose of the project and agree that of the five categories (1) Increase (2) Decrease (3) Identify (4) Understand and (5) Determine, "Decrease" best represents the purpose of the given project.

Step 2: *Selection of Focus variable:* The focus variable is the "loss rate". When asked to specify if any other variables were being assumed constant, the stakeholders proposed assuming constancy of the variable "approval rate".

Step 3: *Selection of Object and Defining Characteristic:* The object in this project was customers and their defining characteristic was their type. So in this case the object and defining characteristic is the bank's "personal loan customers".

Step 4: *Selection of viewpoint:* the stakeholders agree that the project is from the viewpoint of the bank's risk management division

Step 5: *Selection of context:* the initiative of lowering loss rates while keeping the same approval rates is being carried out under the banner of the project "curb losses".

On the basis of the information provided above the *preliminary business objective* can be formulated as follows:

> To reduce loss rates (while keeping approval rates constant) of personal
> loan customers, from the viewpoint of the risk management division,
> within the context of the project 'Curb Losses'

Assessment of business objective against SMART criteria: the IKDDM model recommends refining the preliminary business objective by assessing it against the criteria stipulated by the SMART approach. This approach recommends that we assess the business objective to ensure that it is specific, measurable, achievable, relevant and timely.

Step 1: *Assessing specificity:* the stakeholders attest to the specificity of the preliminary business objective by indicating that it will result in a specific outcome: lowering of loss rates

Step 2: *Assessing Measurability:* the stakeholders confirm the existing value for the focus variable (here the loss rate) as 5%. They express the desired value of focus variable at 3%. Therefore, the delta loss rate (business success criterion) is 2%. An understanding is reached that the business objective will be considered accomplished, when the business success criterion of delta loss rate of 2% is reached.

Step 3: *Assessing Achievability:* the stakeholders confirm that the stated project is achievable within the constraints of knowledge, resources and time.

Step 4: *Assessing Relevance:* the stakeholders agree that the business objective of the stated project is relevant to the organizational goals. The particular organizational goal that the business objective would help meet is that of increasing revenues.

Step 5: *Assessing Time-Boundedness:* the stakeholders confirm that the stated project should be completed over financial year 2008-2009 and provide specific dates as 21st August 2008 to 15th August 2009.

According to IKDDM, the information from measurability (when focus variable is quantitative such as the variable here, loss rate), and time Boundedness must be used to refine the statement of preliminary business objective formulated earlier, into a final statement of business objective.

The final statement of business objective is:

> To reduce loss rates (while keeping approval rates constant at 60%) of personal loan customers by 15%, from the viewpoint of the risk management division, within the context of the project 'Curb Losses' of Risk management division, over the time frame 21st August 2008 to 15th August 2009.

2. Identification of Business Benefits

The stakeholders confirm that the business benefits to be gained from this project are quantifiable in monetary terms. In accordance with the IKDDM steps, they specify the amount of benefit in monetary terms as an increase in profits through loss savings of $80 million.

3. Setting up of Business Success Criteria

The IKDDM model leverages the dependencies between the tasks, namely the assessment of measurability (conducted during formulation of business objectives) and the identification of business benefits to drive the semi-automation of the task identification of business success criteria. The inputs provided by stakeholders towards these tasks, is used to identify the following as business success criteria:

Delta loss rate = 15%

Loss savings = $80 million

4. Formulation of Data Mining Objective

Creation of preliminary business objective using an adaptation of GQM approach: the IKDDM model suggests a modified GQM based approach to assist in creating the preliminary statement of data mining objective. The various steps recommended by the approach are implemented below to formulate the preliminary business objective based on the project.

Step 1: *Selection of purpose:* The stakeholders discuss the purpose of the project and agree that of the seven categories (1) Classification (2) Estimation (3) Prediction (Classification) (4) Prediction (Estimation) (5) Clustering (6) Visualization or (7) Affinity grouping, "Prediction (Classification)" best represents the purpose of the given project. Based on the definitions of these terms (data mining problem types) provided

by the IKDDM model, Prediction (Classification) i.e. when goal is to classify but based on some future behavior, appears as the most adequate representation of the purpose of the project. In the case of this project the ultimate goal is to be able to classify customers into those who were likely to default and those who were not likely to default.

Step 2: *Selection of Focus variable:* The focus variable is the variable under study. In the context of this project, the bank is interested in the probability of default of its personal loan customers, as it is the values for the likelihood of default that is used to classify an applicant as good or bad, thereby paving the way for the decision of granting or rejecting the applicant's loan application respectively.

Based on the information about the purpose and focus variable and the information about the object and defining characteristic (specified by stakeholders earlier during formulation of business objective), the preliminary statement of data mining objective can be created as follows:

To predict the probability of charge-off of personal bank loan customers

Assessment of data mining objective against SMART criteria: the IKDDM model recommends refining the preliminary data mining objective by assessing it against foru of the five criteria stipulated by the SMART approach. This approach recommends that

340

we assess the objective to ensure that it is specific, measurable, achievable, relevant and timely. IKDDM recommends assessing the data mining objective to ensure that it is specific, achievable, relevant and time bound.

Step 1: *Assessing specificity:* the stakeholders attest to the specificity of the preliminary data mining by indicating that it will result in a specific result: better identification of customers who have a high probability of charging off and becoming delinquent accounts.

Step 2: *Assessing Achievability:* the stakeholders confirm that the stated project is achievable within the constraints of knowledge, resources and time.

Step 4: *Assessing Relevance:* the stakeholders agree that the data mining objective of the stated project is relevant to the business objective of the project, namely a reduction in loss rates. By more accurately predicting the likelihood of charge-off the bank can better differentiate between good and bad customers and bring down the loss rates.

Step 5: *Assessing Time-Boundedness:* the stakeholders confirm that the time frame for the data pertaining to the object of this project (personal loan customers) is 12 months from the point of booking. According to the IKDDM model, this is a valid piece of information, especially fir supervised data mining projects, and in the absence of such

information the data mining objective cannot be finalized. This in turn means that relevant data cannot be identified, and that the project cannot proceed.

According to IKDDM, the information from time-boundedness must be used to refine the statement of preliminary data mining objective formulated earlier, into a final statement of data mining objective.

The final statement of data mining objective is:

Predict the probability of charge-off of personal loan customers, 12 months from the point of booking.

5. Assessment of need to discretize target variable

IKDDM model recommends that the target variable (whether categorical or continuous) be discretized if the decision makers agree that this is in line with their objectives. The model suggests that discretization is a moot point for categorical targets if there are only two levels in the target variable. This is applicable to the present case where the target variable default is binary and can take on only two values, 1 or 0. therefore, no further action regarding discretization is necessary in this case.

6. Clarification of Business Requirements

342

First, in line with the recommendation of IKDDM, the stakeholders discuss if any requirements need to be laid down in terms of ease of use or ease of deployability of the solution. In the present case, the bank does not wish to set up any requirements related to these two.

In the next step, IKDDM recommends eliciting certain set of requirements especially if the project is related to supervised data mining. These requirements include the following [table 5-11]

- ✓ Nature of desired output from Model – Explanatory, Non Explanatory, Either?
- ✓ Desired improvement in accuracy
- ✓ Amount of Quantitative Improvement over old Model (assessed through LIFT)
- ✓ Level of simplicity (or tolerable level of complexity) of the model
- ✓ Generalization of results over different population than the one used for building the model   (assessed through STABILITY)

IKDDM model states that while the stakeholders may or not have input towards these business requirements, an effort must be made to capture these at this point of the project. The business requirements set up by the bank's stakeholders are included below.

Table 6-25: Setting up Business Requirements (Descriptive Testing)

Business Requirements	Response of bank's stakeholders
Nature of desired output	Explanatory
Desired improvement in accuracy	At least 5% over challenger model
Amount of Quantitative Improvement over old Model	Not specified at this point
Level of simplicity	Not specified at this point
Generalization of results over different population than the one used for building the model	Yes

7. Analysis of inventory of business personnel and other resources

Having established a business and data mining objective of the project, the stakeholders wished to formally create a team of individuals who had the necessary skills for seeing the project through to completion. This can be accomplished using tools such as organizational charts, organizational ontology, organizational memory bases etc. In case of the bank, all three tools were available. The stakeholders made use of the tool repository proposed by IKDDM to select the tool that most adequately met their needs. Since they wanted to look for individuals by their role and the data mining projects they had participated in, they had two choices, organization ontology and organizational

memory. The bank's stakeholders decided to make use of the organizational ontology to search for relevant individuals. This meant browsing through the ontology to identify individuals.

The organization ontology helped in identification of Ms. Julie Thomas as the key technical stakeholder. She was named as the central contact point for all technical issues, and for acting as a liaison between the business and technical teams for the project. Her counterpart on the business side was the business manager Mr. Gilbert Wright who was named as the central contact point for all business related issues pertaining to the projects. After their appointment, Ms. Thomas and Mr. Wright were asked to make use of the organization ontology to identify two individuals each for their teams. They were asked to use familiarity with supervised data mining projects and experience with data mining projects in the risk management division of the bank as criteria for selection of relevant individuals. The key stakeholders identified Mr. Robert Berry as the project sponsor, who agreed to be the project sponsor after reviewing the information from tasks already completed (such as the business and data mining objectives of the project, business benefits and business success criteria, and the members of the business and technical teams involved in the project).

8. Clarification of Policy, Legal and Budgetary constraints

At this stage, the business and technical managers interacted to clarify the policy and legal constraints applicable to the project. The technical manager Ms. Thomas made use of the company's business rules base to identify the applicable policy constraints. This project was centered on personal loan customers, and the bank has a policy of granting home loan only to individuals who were 21 years or older. This policy needed to be applied in the later stages as data would be collected. No other policy constraints were applicable.

The business manager Mr. Wright worked on identification of applicable legal constraints. He is aware that the legal rules have a major ramification in the banking industry. Together with his team he identified the following as legal constraints applicable to the project: variables such as individual's gender, nationality, and religion should be excluded from the analysis. At the time of the initial application, applicants are asked to voluntarily reveal any information about these fields, and are assured that it will not be used in the decision making process in any way. While the bank stores this information to build the customer's profile and to target him or her with only meaningful offers, the same variables cannot be used in making a decision such as granting of a loan and will therefore be excluded from the analysis.

The project sponsor Mr. Robert Berry is asked to specify the financial or budgetary constraints on the project. He allows for a total expenditure of $20,000 including amount spent on hiring process (if new individuals were needed), on new data (may

need to be collected or purchased from external data vendor), purchase and installation of data mining software.

9. Setting up of Data Mining Success Criteria

The IKDDM model identifies dependencies between this task and two preceding tasks, namely formulation of data mining goals and requirements to semi-automate the execution of this task. Both the business and technical managers, Mr. Wright and Ms. Thomas work together to finalize the set of applicable criteria.

From the discussion of business requirements held earlier, they are aware that the stakeholders are interested in developing a solution (response model) that offers at least 5% more accuracy than the previous (challenger model) and provided for results that generalize well over population different from the one used to construct the model. At the time, the stakeholders did not specify requirements on simplicity of the solution.

Therefore, it was clear that accuracy and stability were to be included as data mining success criteria. Mr. Wright and Ms. Thomas made use of the look up table proposed by IKDDM to identify other data mining success criteria that were applicable to the data mining problem type under consideration, namely, Prediction (Classification). By referring to cross reference table 5-19 they find that the applicable data mining success criteria include:

✓ Accuracy (Misclassification Rate)

- ✓ Lift

- ✓ Precision

- ✓ Recall

- ✓ Simplicity

- ✓ Stability

- ✓ Sensitivity

- ✓ Specificity

- ✓ ROC curve

- ✓ Area Under ROC curve

- ✓ KS Statistic

- ✓ Profit/Loss

After some discussion they identify, Simplicity, Lift, Precision, Area under ROC curve and KS static, besides accuracy and stability (identified earlier) as data mining success criteria. They refer to the IKDDM model to confirm the meanings of each of these terms.

10. Initial assessment of modeling techniques

Having completed the preceding tasks, the next step is to perform an initial assessment of modeling techniques applicable to the project. The IKDDM model identifies dependencies between this task and two preceding tasks, namely formulation of data

mining goals and business requirements pertaining to the nature of output from the model, to semi-automate the execution of this task. The business and technical managers, Mr. Wright and Ms. Thomas work together to finalize the set of applicable modeling techniques.

They refer to the cross reference matrices 5-17 and 5-18 proposed by the IKDDM model to execute this task. Using information about the target variable type (binary in this case) and the data mining problem type (here prediction –classification). The initial set of techniques (non-ensemble) are identified as

- ✓ Logistic regression
- ✓ Classification Tree
- ✓ k-nearest neighbor
- ✓ Naïve Bayes*
- ✓ Neural network*
- ✓ Support Vector Machines*
- ✓ Genetic algorithm*

The IKDDM model also recommends ensembles based on using non-explanatory techniques as input (marked in asterisks) and explanatory techniques as output, if the performance of non-explanatory techniques exceeds that of explanatory techniques.

349

Mr. Wright and Ms. Thomas identify following three techniques from the above set of techniques:

✓ Logistic Regression

✓ Classification Tree

✓ Neural Network

Since there business requirement is for an explanatory model, they are presented with the following as the applicable ensemble technique.

Neural network as input and logistic regression or classification tree as output

11. Assessment of selected modeling techniques against data mining success criteria

In the next step, Mr. Wright and Ms. Thomas need to assess the modeling techniques selected by them against the data mining success criteria that can be used for assessing the output of these techniques. The IKDDM model semi-automates the execution of this task using look up tables. The output of this task helps confirm that while logistic regression and classification trees can be assessed using all the data mining success

criteria established earlier; neural networks can be assessed using all but the simplicity criterion which does not apply to non-explanatory techniques such as neural networks.

12. Analysis of applicable software tools to implement the modeling techniques

In this task, Mr. Wright and Ms. Thomas review the software tools available to the bank to implement the selected modeling techniques. Using the look up table proposed by IKDDM model as a guide, they can see that the bank has two tools available to implement all three techniques. These include SAS Enterprise Miner 4.3 and SPSS Clementine 12.0. Ms. Thomas indicates that her team members who would be working on the modeling phase of the project are more experienced with SAS EM 4.3 and therefore a decision is made to use this tool for the modeling phase of the project.

13. Analysis of available software tools to support data mining success criteria

During this task, Mr. Wright and Ms. Thomas work to ensure that the SAS EM 4.3 tool selected by them also yields (implicitly or explicitly) the data mining success criteria established for this project. The IKDDM model semi-automates the execution of this task using cross reference tables. It is found that the chosen software tool SAS EM 4.3 will support identification of following data mining success criteria out of the total set of data mining success criteria. The cross reference table also provided information

351

about whether the tool outputs the criteria directly (explicitly) or implicitly (i.e. tools outputs information which can be used for estimating the values for the criteria). In some cases when the tool outputs criteria only implicitly, the user is expected to define the calculation of the criteria.

- ✓ Accuracy (Implicit)
- ✓ Simplicity (Implicit – User defined)
- ✓ Stability (Implicit – User defined)
- ✓ Precision (Implicit – Confusion Matrix)
- ✓ Lift (Explicit – Lift Chart)

However, the software tool SAS Enterprise Miner 4.3 does not yield the following two data mining success criteria, namely,

- ☒ Area Under ROC Curve
- ☒ KS statistic

Mr. Wright holds some discussion with his business team and the key stakeholders to discuss if these two criteria established earlier could be removed from the list of business success criteria. The discussion reveals that the area under ROC curve can be removed from the list of applicable criteria but that the KS statistic was an important criterion and needed to be used in at least the model selection phases. Mr. Wright relays

this information back to Ms. Thomas who indicates that while SAS EM 4.3 does not directly yield the KS statistic, her technical team members will be able to generate values for the KS curve needed by the key stakeholders outside of the tool in a Microsoft Excel spreadsheet. But since this requires additional effort, they would calculate these values only for the response model(s) and not for all models that are tried during the modeling phase.

14. Elicitation of a preference function and creation of a value function

During this stage Mr. Wright and Ms. Thomas consult to create a preference function and a value function for the data mining success criteria. They use the steps suggested in the IKDDM model to execute this task. Through several round of consultations, they establish value functions, thresholds and ways of creating a composite score based on the weighted data mining success criteria.

Table 6-26: Data Mining Success Criteria: Value function, threshold and weights (Descriptive Testing)

Data Mining Success Criteria	Value Function	Threshold	Weights
Accuracy	1-test misclassification rate	≥ 0.75	0.35
Simplicity	Based on number of leaves for tree model Score = 0 if # of leaves is <3 or ≥ 8 Score = 1 if $3 \leq$ # of leaves ≤ 5 Score = 0.5 if $6 \leq$ # of leaves ≤ 8	>0	0.15
	Based on number on the number of interactions for logistic regression model Score = 0 if # of interactions is <3 or ≥ 8 Score = 1 if $3 \leq$ # of interactions ≤ 5 Score = 0.5 if $6 \leq$ # of interactions ≤ 8		
Stability	Visual inspection of non cumulative % response lift chart up to the 50th percentile	>0	0.15
Lift	Visual inspection of cumulative % captured response chart at the prior probability	>0	0.20
Sensitivity	Using Confusion Matrix: True positives / (Sum of True Positives and False Negatives)	≥ 0.90	0.15

The information provided for the weights for each of the criteria yields the formula for composite score. However, prior to creating the formula, A check should be made to ensure that the sum of weights for different data mining success criteria adds to 1. This is true for the above weighting system as 0.35+0.15+0.15+0.20+0.15 =1

354

Formula for Composite score is the weighted scores for the various data mining success criteria. Once the weights have been assigned, the generation of the formula can be easily automated.

Composite Score = $[(Accuracy_{score} * 0.35) + (Simplicity_{score} * 0.15) + (Stability_{score} * 0.15) + (Lift_{score} * 0.20) + (Sensitivity_{score} * 0.15)]$

15. Analysis of Applicable Data Resources (Using existing new variables, ratio variables or collecting data)

During this task Mr. Wright analyzes the applicable data resources with his business team and the key stakeholders involved in the project. The stakeholders indicate that it is their belief that certain key variables were missing from the challenger model which may have led to its poor performance and the increasing bad rates for the company. Mr. Wright's business team presents the list of variables used by the old (challenger) mode. The challenger model was a logistic regression model that made use of the following fields.

The stakeholders notice that although debt to income was an important ratio variables, one of its constituent variables was not included in the model. Mr. Wright and his business team research other variables applicable to the project.

355

DATA UNDERSTANDING PHASE

Studying data sources and assessing data sufficiency

During this phase, the business and technical team members interact to determine whether or not the available data is sufficient to address the given data mining problem. The analysis done by the team members reveals that some key variables are missing from the analysis. For example, the data does not include the other debt owed by the personal loan customers. Ideally this debt should be considered along with the credit debt owed by the customer to build a complete profile.

This information is passed on by the team members to their team leads Mr. Wright and Ms. Thomas who decide to acquire this data from an external data vendor, named Acxiom. They contact Acxiom for the availability of the data and if the data could be made available by 21st October 2008. Next they enquire about the cost of buying this data. Acxiom quotes an amount of $5000 for the data, which is within the budget constraints of the company and is approved by the project sponsor.

Assessing the need for derived attributes

In this task the business team manager consults with his business team to assess any derived attributes that may be relevant to the problem being addressed by this project. They consider the various variables and identify debt to income as a meaningful

356

attribute that is likely to improve the predictive accuracy of the model. This information is then passed on by Mr. Wright to Ms. Thomas and the members of the technical team.

Documentation of data sources

During this step, the technical team members document all the data sources for the given project. The data in the past had also been drawn from various sources, some available to the bank directly, other obtained through credit bureaus and external data vendors. A record of trace is created to document the exact source of the data.

Survey of data quality

During this stage, the technical team members survey the data to assess the data quality. They find that the data contains no outliers or missing values. The distribution of the variables, their standard deviation etc is also noted by the members and results compiled in a data quality report.

DATA PREPARATION PHASE

Construction of dataset

During this task the technical team members construct the data set by using the data sources documented in the data understanding phase. This includes the data that the

bank already had and the other variables that were purchased from credit bureaus. This led to the creation of a preliminary data set. In succeeding tasks this data set was further refined and made ready for modeling.

Application of Policy and Legal Constraints

During this task, the technical team members applied the policy and legal constraints to the data set. As part of applying the policy constraints, they removed from the data set, all applicants whose age was less than 21. as part of applying the legal constraints, they removed variables such as gender, nationality and religion, information about which wax voluntarily submitted by some applicants during the application process.

Addition of derived variables

During this task, the technical team added the derived variable 'debt-to-income' ratio. This was created by dividing the sum of credit debt and other debt by the income. Since no other derived variables were identified, the technical team moved on to the next task.

Discretization of target variable

This was considered during the business understanding phase. However, this task is not applicable to the given problem scenario as the target variable has only two levels, and therefore discretization is not applicable.

Fetch rank ordered array of modeling techniques (from modeling phase) and format the data

This task requires formatting data in accordance with the various techniques in the rank ordered array of modeling techniques. However, the rank ordering of modeling techniques is a task that is performed during the modeling phase. so, at this time, the technical team members moved ahead to the modeling phase and implemented this task. Then they iterated between the data preparation and modeling phase to format data in accordance with each modeling technique (refer to task 1-4 of modeling stage)

Task 4 of the modeling phase redirects us back to data preparation. The technical team formats the data in accordance with the first technique in the array, namely neural networks and passes this formatted data onto task 5 of modeling phase where the parameters of the modeling technique are set up. Two more iterations are made between this task and task 5 of modeling phase for the remaining two modeling techniques, classification trees and logistic regression respectively.

The data preparation for all three techniques was done according to the recommendations of the IKDDM model (see chapter 5). These are also summarized below.

Loading data in software tool and applying tool specific formatting

SAS EM 4.3 does not require any additional formatting beyond the formatting for the modeling techniques which has already been completed at the end of the preceding task. Thereafter the technical team moves on to the next task.

Ensuring that tool can handle required number of rows and columns

During this task, the technical team works to ensure that the tool selected can handle the number of observations or rows (N=700) and the number of input variables or columns (I=9). This task is a check to ensure that the data set can be handled by SAS EM 4.3. The assessment reveals that SAS EM 4.3 can handle the required number of rows and columns.

MODELING PHASE

Calculating values for accuracy and resource constraints for each modeling technique in the array of modeling techniques

During this task the modeling team searched through a case base of past projects to assess the training time, memory usage etc of data sets similar to one being used in this analysis. They use the number of cases, type of target variable, and number of input variables to search for a similar data set. The closest match is found with a data set used by the credit risk division for distinguishing between good and bad customers.

Table 6-27: Search for similar data set from past projects (Descriptive Testing)

	Number of cases	Number of input variables (excluding ID)	Type of target variable
Data set for this project [BANKNEW]	700	9	Binary
Data set for past project [CREDITDATA]	800	10	Binary

Modeling techniques such as logistic regression, neural networks and classification trees were also tried for this data set and therefore results for training time, memory usage etc. were available. These were used to gain an estimate for accuracy and resource constraints for the present project.

Table 6-28: Accuracy and Resource Utilization for CREDITDATA (Descriptive Testing)

Data Set	Classification Techniques	Accuracy	Training Time$_{normalized}$	Memory Usage$_{normalized}$
CREDITDATA	Decision Trees	0.80	0.285714286	0.881818
	Logistic Regression	0.78	0.628571429	0.045455
	Neural Network	0.83	0.857142857	0.536364

2. Generate preference functions for resource constraints and setting up formula for creating composite score

During this task, Mr. Wright and Ms. Thomas work together to set up preference functions and a formula for creating a composite score that could be used to rank order the techniques. This involved setting up of weights and thresholds for accuracy, training time and memory usage. the discussion between the two managers reveals that all three criteria are important, but that accuracy is slightly more important than training time and memory usage. The values for weights and thresholds for various criteria finalized by them are summarized in table.

Table 6-29: Preference functions for Accuracy and Resource Utilization (Descriptive Testing)

Criteria	Accuracy	Training Time$_{normalized}$	Memory Usage$_{normalized}$
Weights	0.40	0.30	0.30
Composite score = (0.40*accuracy) + (0.30*training time) + (0.30*memory usage)			

3. Rank ordering array of modeling techniques and making final selection of techniques

The IKDDM model automates this task based on the output of the previous two tasks in this phase. Given that the users have already provided the preference functions for the various criteria, the generation of composite scores simply involves multiplication of values for accuracy and normalized values of training time and memory usage by their respective weights. The composite scores for techniques are used to rank order the techniques (highest to lowest) and will be made use of in selecting the final set of techniques in the next task.

Table 6-30: Rank Ordering modeling techniques by Accuracy and Resource Utilization (Descriptive Testing)

Data Set	Classification Techniques	Acc. score	Acc. Weight	TT$_{norm}$ score	TT weight	MU$_{norm}$ score	MU weight	Comp. score	Rank
CREDIT DATA	Decision Trees	0.80	0.40	0.2857	0.30	0.8818	0.30	0.6702	2
	Logistic Regression	0.78	0.40	0.6285	0.30	0.0454	0.30	0.5142	3
	Neural Network	0.83	0.40	0.8571	0.30	0.5363	0.30	0.7500	1

Acc – Accuracy; TT $_{norm}$ – normalized training time; **MU$_{norm}$** – normalized memory usage; Comp. score – composite score; TT - training time; MU – memory usage

3. Select final set of modeling techniques from rank ordered list of modeling techniques

The rank ordered array of modeling techniques reveals that neural networks is ranked first followed by decision trees and finally logistic regression. Mr. Wright and Ms. Thomas discuss the different criteria and the resource constraints and are assured that they could try all three techniques and do not have to leave out any technique from the array. However, they decide to run the modeling techniques in the order specified by the array and so in case that there were any disruptions,

4. Fetch formatted data from Data Preparation phase (repeat for all techniques from finalized set of techniques)

At this stage, the project team reverts back to the data preparation stage where data will be formatted in accordance with all the modeling techniques.

The formatted data is then used to run the modeling techniques in the next step. In the first iteration formatted data for neural networks is fetched for running a neural network model, followed by iterations 2 and 3 where formatted data for classification trees and logistic regression are fetched respectively.

5. Set up Model parameters (refine parameters on basis of objectives and success criteria, wherever applicable)

During this task, the technical team sets up the parameters for the various modeling techniques. The setting of parameters (in SAS EM 4.3) for the various techniques is described below.

The data set consisting of 700 observations is sampled using stratified sampling. 40% of the observations are used for training, 30% for validation and the remaining 30% for test.

Setting up Parameters for neural networks based on recommendations of IKDDM

During this task, the technical team sets up the neural network using following two network architectures: Multilayer Perceptron and General Linear Model. They select conjugate-gradient as the training technique as it requires the least amount of memory.

Next in accordance with the recommendation of IKDDM, they select the model selection criteria as the misclassification rate as this has been set up as a relevant data mining success criterion in this project. No other parameters of neural network appear to have such a direct dependency and therefore other values are left at default. The team managers concur that if time was not a constraint, it would also have been a good idea to try other training techniques as well. But for the purpose of this project they only use conjugate gradient.

Setting up Parameters for classification trees based on recommendations of IKDDM

During this task the technical team sets up the classification trees using following two purity measures: Chi-Square and Entropy reduction as these measures are applicable to a categorical target. Given that accuracy as judged by the misclassification rate is a data mining success criteria, 'proportion misclassified' is selected as the model assessment measure. The assessment for sub-tree is based on the best assessment value. Best assessment value was chosen over 'at most indicated number of leaves'. The number of leaves is directly related to simplicity which has been set up as a data mining success criterion in this project. However for the purposes of pruning the team decides to use the tree that gives the best assessment value and not impose a constraint on the number of leaves on the sub tree. Given the importance of simplicity, the depth of the tree is set at 6. Each tree is also tested with a depth of 5 to assess impact on performance. The

366

minimum number of observations is set at 7 or 2.5% of the observations in the training set and the observations required for a split search are set at twice this value, at 14. (Berry and Linoff, 1997). Other parameters do not appear to have a direct dependency with the objectives or success criteria. However these are necessary for the internal working of the modeling technique and are also set up by the team.

Setting up Parameters for logistic regression based on recommendations of IKDDM

During this task the technical team sets up the logistic regression model. given that accuracy is an important data mining success criterion, the team decides to use 'validation misclassification' selection criteria. Given the importance of memory usage (as evidenced during rank ordering of modeling techniques), conjugate gradient is selected as the optimization method. The team managers concur that if time was not a constraint, it would also have been a good idea to try other training techniques as well. But for the purpose of this project they only use conjugate gradient. Other parameters do not appear to have a direct dependency with the objectives or success criteria. However these are necessary for the internal working of the modeling technique and are also set up by the team.

6 Run modeling techniques and tabulate modeling results for all selected techniques in accordance with DMSC and DM Software used

367

During this task, the modeling team runs each of the three applicable techniques: neural network, classification trees and logistic regression. The modeling output is assessed and values for data mining success criteria are tabulated for the purpose of evaluation in the next phase.

Table 6-31: Tabulation of Modeling Results by Data Mining Success Criteria (Descriptive Testing)

DMSC Model	Accuracy		Simplicity		Stability		Lift		Sensitivity	
	test misclassification rate	Score	# of leaves or # of interactions	Score	Value	Justification	Value	Score	TP/(TP+FN)	Score
C_2_6 (tree)	**0.2511**	**0.7489**	**7**	**0.5**	**0**	**increases from 10th to 20th**	**1.241**	**0.438**	**134/155**	0.864516
C_2_5 (tree)	**0.2511**	**0.7489**	**7**	**0.5**	**0**	**increases from 10th to 20th**	**1.241**	**0.438**	**134/155**	0.864516
E_2_6 (tree)	**0.2606**	**0.7394**	**3**	**1**	**0**	**increases from 10th to 20th**	**1.152**	**0.645**	**144/155**	0.929032
E_2_5 (tree)	**0.2511**	**0.7489**	**7**	**0.5**	**0**	**increases from 10th to 20th**	**1.241**	**0.438**	**134/155**	0.864516
C_4_6 (tree)	**0.2511**	**0.7489**	**4**	**1**	**0**	**increases from 10th to 20th**	**1.347**	**0.545**	**138/155**	0.890323
C_4_5 (tree)	**0.2606**	**0.7394**	**4**	**1**	**0**	**increases from 10th to 20th**	**1.144**	**0.488**	**131/155**	0.845161
E_4_6 (tree)	**0.2417**	**0.7583**	**5**	**1**	**1**	**increases from 50th to 60th**	**1.247**	**0.429**	**142/155**	0.916129
E_4_5 (tree)	**0.2417**	**0.7583**	**5**	**1**	**1**	**increases from 50th to 60th**	**1.247**	**0.429**	**142/155**	0.916129
SW_VM (logistic regression)	**0.2417**	**0.7583**	**4**	**1**	**1**	**stable or decreasing until 80th**	**1.349**	**0.556**	**143/155**	0.922581
F_VM (logistic regression)	**0.2417**	**0.7583**	**4**	**1**	**1**	**stable or decreasing until 80th**	**1.349**	**0.556**	**143/155**	0.922581
B_VM (logistic regression)	**0.2274**	**0.7726**	**9**	**0**	**0**	**increases from 20th to 30th**	**1.331**	**0.426**	**140/155**	0.903226
MLP_MR (neural network)	**0.2274**	**0.7726**	**N.A.**	**0**	**0**	**increases from 20th to 30th**	**1.349**	**0.37**	**139/155**	0.896774
GLM_MR (neural network)	0.2085	0.7915	N.A.	0	0	increases from 20th to 30th	1.349	0.296	138/155	0.890323

Label for Tree – Splitting criterion_Number of branches_depth (Splitting criterion: C – chi-square; E – entropy reduction)

Label for Logistic Regression – Selection method (forward, backward, stepwise)_model selection (validation misclassification)

Label for Neural Network – Training technique (multilayer perceptron, general linear model)_model selection (misclassification rate)

EVALUATION PHASE

1. Assessment of Modeling results against data mining success criteria*

During this task the modeling results of various modeling techniques should be assessed against data mining success criteria. This means assessing the output of the modeling techniques to assess whether or not the thresholds for each data mining success criteria are being satisfied. The IKDDM model proposes this semi-automation of this task as it only requires comparison of scores for different criteria against the threshold values. At present, SAS EM 4.3 does not allow for automated comparison of modeling results against success criteria and therefore the technical team conducted the analysis in a Microsoft Excel spreadsheet. The results of the analysis are tabulated in

Table 6-32. As can be seen from this table, three models meet thresholds for all data mining success criteria: these include a tree model using entropy reduction as the splitting criterion, having four branches, and five levels; (2) the stepwise logistic regression model; and (3) the forward logistic regression model.

371

Table 6-32: Assessment of Modeling Results against Data Mining Success Criteria

(Descriptive Testing)

DMSC Model	Accuracy		Simplicity		Stability		Lift		Sensitivity		Mod meets Thresh ?
	Score	Meets Threshold?	Score	Meets Threshold?	Score	Meets Threshold?	Score	Meets Threshold?	Score	Meets Threshold?	
C_2_6 (tree)	**0.7489**	**No**	0.5	**Yes**	**0**	**No**	0.438	**Yes**	**0.8645**	**No**	No
C_2_5 (tree)	**0.7489**	**No**	0.5	**Yes**	**0**	**No**	0.438	**Yes**	**0.8645**	**No**	No
E_2_6 (tree)	**0.7394**	**No**	1	**Yes**	**0**	**No**	0.645	**Yes**	**0.9290**	**Yes**	No
E_2_5 (tree)	**0.7489**	**No**	0.5	**Yes**	**0**	**No**	0.438	**Yes**	**0.8645**	**No**	No
C_4_6 (tree)	**0.7489**	**No**	1	**Yes**	**0**	**No**	0.545	**Yes**	**0.8903**	**No**	No
C_4_5 (tree)	**0.7394**	**No**	1	**Yes**	**0**	**No**	0.488	**Yes**	**0.8451**	**No**	No
E_4_6 (tree)	**0.7583**	**Yes**	1	**Yes**	**1**	**Yes**	0.429	**Yes**	**0.9161**	**Yes**	No
E_4_5 (tree)	**0.7583**	**Yes**	1	**Yes**	**1**	**Yes**	0.429	**Yes**	**0.9161**	**Yes**	**Yes**
SW_VM (logistic regression)	**0.7583**	**Yes**	1	**Yes**	**1**	**Yes**	0.556	**Yes**	**0.9225**	**Yes**	**Yes**
F_VM (logistic regression)	**0.7583**	**Yes**	1	**Yes**	**1**	**Yes**	0.556	**Yes**	**0.9225**	**Yes**	**Yes**
B_VM (logistic regression)	**0.7726**	**Yes**	**0**	**No**	**0**	**No**	0.426	**Yes**	**0.9032**	**Yes**	No
MLP_MR (neural network)	**0.7726**	**Yes**	**0**	**No**	**0**	**No**	0.37	**Yes**	**0.8967**	**No**	No
GLM_MR (neural network)	0.7915	Yes	0	No	0	No	0.296	Yes	0.8903	No	No

Calculation of values for KS-statistic for E_4_6, Forward Logistic Regression and

Stepwise Logistic Regression Models

In the next step, the technical team members work toward calculating the KS statistic

for these three models that meet thresholds for all data mining success criteria. Ms.

Thomas had assured Mr. Wright that the team would be able to calculate this statistic outside of the SAS EM 4.3 tool, which does not output values for this statistic in its modeling results. The following steps were used by the technical team to calculate the values for the KS statistic. Note that the same steps were followed for each of the models. Here we present the tabulated results from the tree model (E_4_6) to illustrate the process. The results for the forward and stepwise logistic regression models are included in Appendix E.

- The first step was to rank order applicants from lowest to highest probability of default as predicted by each competing model (note that the probability of default is obtained by using each model to score the data. These are output by SAS EM 4.3 as P_DEFAULT1).

- After rank ordering, applicants were divided into deciles, decile 1 being the group of applicants with lowest probability of default while decile 10 being the group of applicants with highest probability of default.

- For each decile, starting from decile 1 to decile 10, cumulative percentage of good applicants and cumulative percentage of bad applicants was calculated. The percentage of bads for a decile can be obtained by dividing the total number of bads for that decile to the overall number of bads in the sample. The number of bads for each decile can be calculated using a pivot table in Excel (Table 6-33). For example, percentage of bads for decile 5 is 12/183. However we are interested in the cumulative % of bads in each decile. This is calculated as sum

of number of bads from decile 1 to decile 5 divided by total number of bads in the sample. Similar methodology was used to calculate cumulative good percentage for each decile.

-

Table 6-33: Pivot Table (calculating default accounts for each decile)

decile	Data	
	Sum of default	Sum of n
1	7	70
2	12	70
3	3	70
4	3	70
5	13	70
6	16	70
7	15	70
8	23	70
9	41	70
10	50	70
Grand Total	183	700

Table 6-34: Calculating cumulative % of good and bad accounts

Decile	cumulative bad	cumulative good	difference
1	3.8%	12.2%	8.4%
2	10.4%	23.4%	13.0%
3	12.0%	36.4%	24.3%
4	13.7%	49.3%	35.7%
5	20.8%	60.3%	39.6%
6	29.5%	70.8%	41.3%
7	37.7%	81.4%	**43.7%**
8	50.3%	90.5%	40.2%
9	72.7%	96.1%	23.5%
10	100.0%	100.0%	0.0%

- In the next step, difference between cumulative percentage of bad and cumulative percentage of good in each decile was calculated.

- The maximum difference between the cumulative bad percentage and cumulative good percentage, which is defined as the KS statistic of the model, was calculated for each of the competing model. For the tree model, E_4_6, the KS statistic was 43.7%. The chart below shows the KS statistic curve for this model.
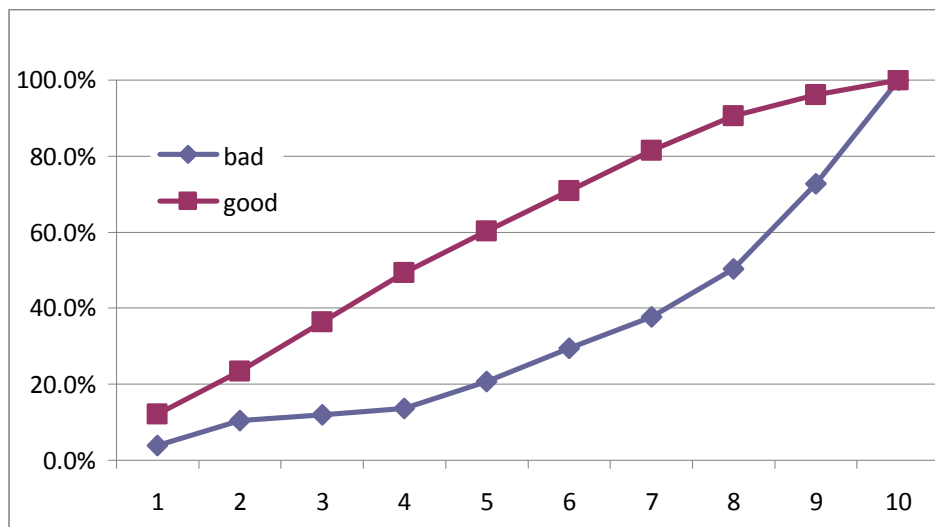
Figure 6-5: KS curve for E_4_6

Assessment of Modeling results against business success criteria

During this task the technical team assesses the three models that met all thresholds for data mining success criteria against business success criteria. This required a check to ensure that the models meet the desired values for loss rate and the desired decrease in losses.

Table 6-35: Assessment of Modeling Results against Business Success Criteria

(Descriptive Testing)

BSC Model	Loss Rate			Loss savings			Model meets all Thresholds ?
	Value	Threshold	Meets Threshold ?	Amount	Threshold	Meets Threshold ?	
E_4_5 (tree)	**20%**	**15%**	**No**	**($107,142,857)**	**$80,000,000**	**No**	No
SW_VM (logistic regression)	**15.6%**	**15%**	**Yes**	**$83,333,333**	**$80,000,000**	**Yes**	Yes
F_VM (logistic regression)	15.6%	15%	Yes	$83,333,333	$80,000,000	Yes	Yes

During this task, the technical team members and Ms. Wright assess the three models that passed all thresholds for data mining success criteria to determine whether or not they also meet the business success criteria. For this project, the business success criterion had been set up as follows:

- A 15% reduction in loss rate at the same approval rate (60%)
- Increase in profits (via loss savings) of at least $ 80 million

The technical team members again used the predicted probabilities of the three selected models to determine the loss rate for each decile. Since SAS EM 4.3 does not provide support for this analysis, it was performed outside of this tool in a Microsoft Excel Spreadsheet. The steps followed by the technical team are listed below. Here we only

present tabulated results for the tree model, E_4_6. The results from the stepwise and forward logistic regression models are included in Appendix E.

- Using the probability of default predicted by each model, applicants were rank ordered from lowest to highest probability of default.

- Then these applicants were divided into deciles, decile 1 being the group of applicants with lowest probability of default while decile 10 being the group of applicants with highest probability of default.

- The loss rate for each decile was calculated by dividing the total number of defaults for that decile by the number of applicants in that decile. For this data, each decile had 70 applicants). This data is also based on the pivot table (Table 6-33). Overall default rate of top 6 deciles of each competing model was calculated by dividing the total number of actual defaults in top 6 deciles by total number of applicants in those deciles.

Table 6-36: Loss rate by decile

Decile	bad rate
1	0.1
2	0.171429
3	0.042857
4	0.042857
5	0.185714
6	0.228571
7	0.214286
8	0.328571
9	0.585714
10	0.714286

- Default rate of top 6 deciles of each competing model was compared to see which model lead to the highest reduction in loss rate. Figure 6-6 presents a comparison of loss rates output by the various models and compare it to the loss rate of the challenger model.

378
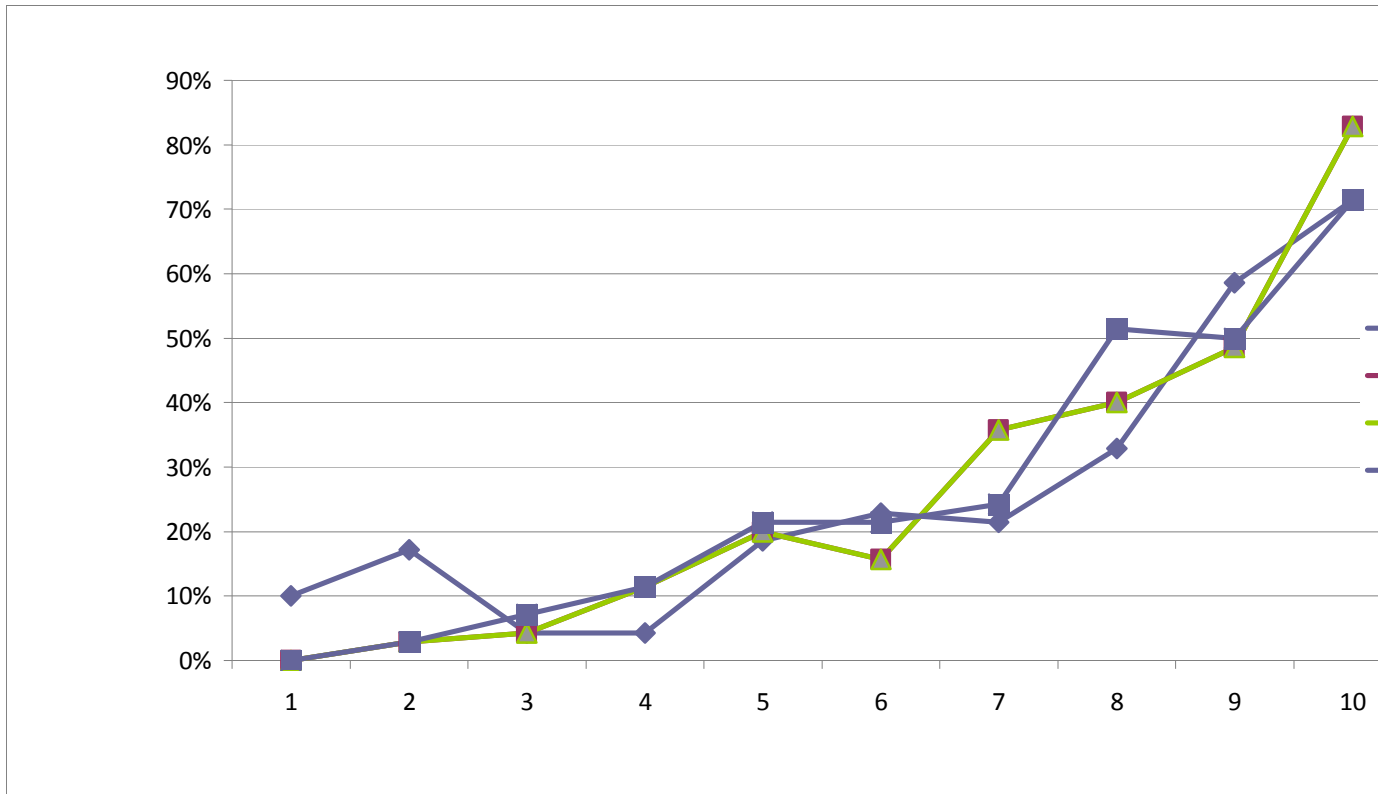
Figure 6-6: Loss rates of different models (Descriptive Testing)

The results (Table 6-37) reveal that the loss rate from the tree model (E_4_6) is 12.9% which is higher than the loss rate of the existing model (10.7%). Clearly this model does not meet the business success criteria of at least a 15% reduction in loss rate, but in fact leads to an increase of 20% in the loss rate. Therefore the technical team rejects this model and continues with the assessment of the two logistic regression (response) models.

Table 6-37: Loss rate and Loss savings from selected models (Descriptive Testing)

	Existing Model	Tree	Stepwise	Forward
Loss rate for 60% Approval rate:	10.7%	12.9%	9.0%	9.0%
Improvement	-	-20.0%	15.6%	15.6%
# Accounts booked per year:	1,000,000	1,000,000	1,000,000	1,000,000
# Charged off accounts:	107,143	128,571	90,476	90,476
Avg. Balance per Charged off account:	$5,000	$5,000	$5,000	$5,000
Year 1 Dollar Loss:	$535,714,286	$642,857,143	$452,380,952	$452,380,952
Year 1 Loss Saving:	-	($107,142,857)	$83,333,333	$83,333,333

The loss rate for the step wise logistic regression model is 9% and it leads to a 15.6% reduction in loss rate, and meets the business success criterion of at least a 15% reduction in loss rate. The team then uses the loss rate to determine the overall increase in profits that can be expected by deploying this model.

They know that the bank books a total of 1 million accounts every year. Using the stepwise logistic regression model, the number of charged off accounts would be 90,476. Given that the average balance per charged off account is $5000, the year 1 dollar loss would translate into $452,380,952. The year 1 dollar loss for the challenger model (existing logistic regression model) is $535,714,286. Subtracting the two, we can get the loss savings (or incremental profits) of $83,333,333, which exceeds the desired increase in profits of at least $80 million.

The technical team notices that the loss rate from the forward logistic regression models also yields the same figures, and that the two models (even though they were created using different techniques) are yielding the exact same results.

Using value functions to create composite scores for selected models

During this task, the technical team applies weights to scores for data mining success criteria for the models that meet thresholds for both business as well as data mining success criteria. This results in two models with the same composite scores.

Table 6-38: Assessment of Modeling Results against Data Mining Success Criteria (Descriptive Testing)

DMSC Selected Models	Accuracy		Simplicity		Stability		Lift		Sensitivity		Composite Score
	Score	Weight	Score	Weight	Score	Weight	Score	Weight	Score	Weight	
E_4_5 (tree)	**0.7583**	**0.35**	**1**	**0.15**	**1**	**0.15**	**0.429**	**0.2**	**0.9161**	**0.15**	0.638624
SW_VM (logistic regression)	**0.7583**	**0.35**	**1**	**0.15**	**1**	**0.15**	**0.556**	**0.2**	**0.9225**	**0.15**	0.664992
F_VM (logistic regression)	0.7583	0.35	1	0.15	1	0.15	0.556	0.2	0.9225	0.15	0.664992

Compare models with the same composite score against different data mining success criteria (if applicable)

Ms. Thomas and the technical team observe that both the forward and stepwise logistic regression models have the same composite score. Comparison of their scores across different data mining success criteria (

Table 6-32) reveals that these two models have the exact same score for all of these criteria. Recall from the earlier step, that these two models were also equivalent in terms of the improvement in loss rate and overall loss savings. Ms. Thomas discusses these results with Mr. Wright.

Determine next steps for the project

As next steps in the project, the team mangers decide to use the stepwise model on a small test sample comprising of 5% of the population. This is meant to see how well the model does on this population. The key stakeholders and Mr. Wright and Ms. Thomas decide to send the model into actual implementation if the results are aligned with their expectations and the model is helping them achieve their business and data mining objectives.

7 CONCLUSION

The goal of this chapter is to recapitulate the problem addressed by this dissertation, the motivation behind the research and the solution proposed. The chapter concludes with a discussion of open issues and work for the future. Since the dissertation has been following Hevner et al's (2004) Design Science Research methodology, the same is used to summarize the results.

7.1 Problem Identification and Motivation

A Knowledge discovery and Data Mining (KDDM) Process Model plays a significant role in the effective and efficient execution of KDDM projects. By its very definition, a KDDM process model is meant to assist the user through every single one of the multitude of tasks that underlie complex and iterative KDDM projects. A review of existing KDDM process models reveals that they provide only limited assistance to the user involved in executing such projects, and that too in a checklist manner. While the checklist presents the users with tasks to consider in the course of a KDDM projects, there is no detailed assistance provided as to "how" the long list of tasks in the checklist can be executed. The lack of support is likely to result in failure to execute tasks, a serious problem compounded by the fact that there exist numerous dependencies between tasks, i.e. many tasks in KDDM tasks are dependent on the output of previous tasks as their input. This means that not executing or not adequately implementing a

task can snowball into failure to execute or execute properly, a task dependent on the output of the former task as input. These dependencies are not sufficiently explored or highlighted in existing models, leading to a fragmented model design. The lack of support for execution of tasks is particularly evident in the Business Understanding phase which is the first phase of KDDM projects. This is particularly problematic since this phase is the foundational phase in the KDDM process and affects all other phases of the project.

This dissertation addresses the deficiencies in existing models by designing an artifact in form of a new KDDM model, called the Integrated Knowledge Discovery and Data Mining (IKDDM) Process Model. IKDDM was designed by a thorough exploration of the dependencies existing between the tasks of the same phase as well as the tasks of the different phases of KDDM projects. The execution of every single one of the tasks outlined by the model is supported by semi-automating the dependency relationships of this tasks with other tasks and/or through a set of approaches/clearly defined steps that can be followed by the user to adequately implement each task.

7.2   Design as an Artifact

The result of design science research is a purposeful artifact (construct, model, method or instantiation) that created to address an important organizational problem. The organizational problem addressed can be a heretofore unsolved problem that is being addressed by design science research in unique or innovative ways, or solved

problems in more effective or efficient ways. This dissertation presents an artifact, specifically a method, in form of a new KDDM process model called the Integrated Knowledge Discovery and Data Mining Process Model. The artifact addresses a problem addressed by previous research (namely supporting execution of KDDM process), but does so in more effective and efficient ways. The artifact designed is a prescriptive process model which provides both point and flow guidance towards execution of KDDM projects. Unlike existing KDDM models, the enactment domain of the IKDDM model contains the full set of features (task-task dependencies, steps or adaptations of relevant approaches from the literature) to support the implementation of the process recommended by the process model.

7.3 Design Evaluation

The utility, quality, and efficacy of a design artifact must be demonstrated via well executed evaluation methods. In this dissertation, descriptive testing and analytical testing methods were used to conduct the evaluation. Descriptive testing, via construction of a detailed scenario was used to illustrate how the IKDDM model would guide the execution of a KDDM project. The KDDM project is based on the context of a financial loan granting institution's attempt at discerning between customers in order to identify those who should be granted loans.

A two step analytical testing approach was used to demonstrate the efficacy and quality of the design artifact. In the first step, users with varying level of experience in

386

data mining were asked to execute key tasks in data mining (such as formulation of business and data mining objectives, selection of modeling techniques based on problem type, selection of data mining success criteria based on modeling techniques, setting of modeling parameters etc.) using either the IKDDM model or the CRISP-DM process model as a guide. The tasks were presented in form of multiple choice questions. Users were randomly assigned to the CRISP-DM and the IKDDM group and each group was presented with the exact same set of questions. Final scores were computed by assigning a score of +2 points for a correct answer and a score of 0 for an incorrect answer. The performance of the users in the two groups (IKDDM versus CRISP-DM) was compared by computing their scores. An independent means t-test revealed that there was a significant difference between the performance of the two groups and IKDDM clearly outperformed the CRISP-DM model. The performance of the group that used the IKDDM model is a clear indicator of the efficacy of the design artifact.

In the second step of the analytical testing approach, the users were presented with a survey to assess their perception of the quality of the process model used by them to execute the data mining tasks. The survey used a 7-point Likert scale with options ranging from strongly agree to strongly disagree. The instrument for process model quality was based on four dimensions: perceived usefulness, perceived ease of use, user satisfaction and perceived semantic quality. A Mann-Whitney test revealed a significant difference in the quality of CRISP-DM and IKDDM models and IKDDM model's

quality was rated as significantly higher and different from that of the CRISP-DM model. This second step of the analytical testing approach provided evidence of the superior quality of the design artifact.

7.4   Research contributions

Research contributions from design science research can be in one or more of the following three forms: the design artifact itself, extensions and improvements to the knowledge base, and creative development and use of evaluation methods and new evaluation metrics for evaluating the design artifact. This research provides research contributions in all three forms.

(1) Design as an artifact – the most important contribution of this research is the design artifact, the IKDDM process model itself. The artifact (a method) has been designed to address an important organizational problem, namely the execution of the complex and iterative KDDM process.

(2) Contributions to the knowledge base – the development of an appropriately evaluated, comprehensive KDDM process model, with a detailed documentation supporting its implementation has contributed to the knowledge base containing KDDM process models. The process of building the design artifact has also provided contributions to the knowledge base through uncovering and explicating many new tasks in the KDDM process not described by existing process models. Another contribution to the knowledge base of the KDDM process is through the methodical

388

exploration and propositioning of semi-automation of tasks (beyond those of the modeling and data preparation phases) through leveraging the dependencies explicated in the design of the IKDDM model.

(3) Evaluation methods - given the uniqueness of the KDDM process, the fact that they are prescriptive in nature, and the role of the human user in the KDDM process makes many of the conventional methods of evaluation of artifacts (controlled experimentation, simulation, black box testing) inapplicable. This dissertation implemented a two-step approach for analytical testing which can be used for cross-comparison of various KDDM process models. The first step of the approach whereby users are asked to implement key tasks in data mining using a KDDM model as their guide, can be used to systematically evaluate the breadth of tasks covered by a KDDM process model. The coverage testing of tasks supported by the model, provides an estimate of bounds on the behavior of the artifact if it were to be implemented in an actual KDDM project.

7.5 Research Rigor

Rigor is derived from the effectiveness use of the existing knowledge base and theoretical foundations. In this dissertation, rigor in construction was achieved through comprehensive analysis of the knowledge base containing KDDM process models and the foundations for each model. A set of design requirements were formulated prior to constructing the artifact and were later assessed to ensure that the IKDDM model met

389

each one of the design guidelines. KDDM process models proposed by both academicians and practitioners were considered during the construction of the artifact. The enactment domain of the IKDDM model was supported through series of well formulated steps or adaptations of relevant approaches proposed in the literature.

Rigor in evaluation was achieved through comparison of the design artifact against an existing artifact, the CRISP-DM process model which is considered as a leading methodology for implementing the KDDM process. Other process models are near subsets of this model. Evaluation was conducted to assess the utility, efficacy, and quality of the design artifact in comparison to similar existing artifacts. Similar to many other designed artifacts, the artifact designed in this research, is part of a human machine problem solving system. Various aspects of the KDDM process are intertwined, with some requiring human intelligence, others requiring machines running complex data mining algorithms and spanning large databases, and still others requiring the interaction of both humans and machines to execute particular processes. Design science research recommends getting appropriate subject groups to achieve rigor in evaluation of such artifacts. In this research, subjects groups consisting of users with varying levels of experience in data mining, participated in the evaluation. Given the tenets of design science research, the focus of the analytical evaluation was kept on determining how well the artifact works and not on theorizing or justifying why it works. The significantly better performance of the group that used the IKDDM model

to execute tasks in an illustrative KDDM project provides evidence that the artifact effectively and efficiently to support the information requirements of this group.

7.6 Design as a search process

Design science research is inherently iterative (Hevner et al. 2004). In this dissertation, design science research was used to design a KDDM process model which in itself is highly iterative in nature. The design of the model was initiated by using key tasks in the KDDM process that are common across various existing process models. The dependencies between tasks were explored by studying every task and (1) how its output affects other succeeding tasks in the same and different phases, and (2) which preceding task's output is used by this task as input. This exercise was done iteratively and gradually the granularity of tasks was refined to describe tasks at a greater level of detail than what is offered by existing process models. Each time a task at a finer level of granularity was included for the purpose of more effective implementation of the KDDM process, the dependency relationship of this task with other tasks of the same phase as well tasks of the different phases was examined. Dependency relationships between all other tasks previously considered were also reexamined. The final design of the model includes what appeared to be the most optimal sequencing of tasks, and one which also helped achieve the goal of semi-automating the execution of tasks, wherever possible. Based on the tenets of design science research the search for the optimal design, was based on heuristic search strategies, and was concluded when the artifact's design met all the design requirements.

7.7 Communication of research

Design science research must be effectively communicated to both technology oriented and management oriented audiences. The KDDM process is an example of a process where a particularly close interaction between managerial and technical users is required. In fact many participants in the KDDM process are responsible for executing both business oriented tasks such as setting up of business and data mining objectives as well as for technical tasks such as setting up and running modeling techniques and evaluating their results against both technical measures (for e.g., accuracy) as well as business measures (for e.g., profits or return on investments). Such users will directly benefit from the IKDDM model's description of tasks, and "how" they can be implemented. Yet another aspect of IKDDM model that will be found to be useful by both management and technology oriented audiences is the linkage of technical tasks (such as setting up of model parameters) to the business objectives of the KDDM project. This leads to greater understanding of the technical aspects of the process by the management audiences; the improvement in understanding of the technical audience can ensure that they implement the technical aspects in congruence with the foundational business objectives.

7.8 Limitations of this research

One limitation of this research is that the design of the artifact does not include the final phase of the KDDM process, namely deployment or implementation of

discovered knowledge. The final phase can only be executed when the outcome of the KDDM process is deployed in an actual organizational setting. Such implementation was outside the scope of this research.

The research utilized analytical testing to assess the static qualities of the designed artifact. The output of the evaluation confirmed the effectiveness and quality of the artifact, however the relatively small sample size (N=42), may have affected the results.

The IKDDM model discusses modeling techniques, their relevance in the context of different data mining problems, evaluation criteria for assessing the output of different techniques, setting up of parameters, but does not discuss the intricacies of different variants of modeling algorithms (say, the multitude of decision tree algorithms). Not covering the intricacies of modeling techniques (algorithms) or data preparation is not to undermine their importance. It is just that the goal of this dissertation was different. It was to design a KDDM process model, where the significance of each task is positioned in the context of the larger picture.

7.9 Directions for Future Research

Following avenues for future research are identified.

(1) *Implementation of artifact in an organizational setting:* the implementation of the IKDDM model in an actual organizational setting can reveal information about

393

the performance of the process model (process performance domain) and the effectiveness and degree to which the tasks of the model can be implemented. It can serve to provide information about any bounds on the behavior of the artifact.

(2) *Artifact improvements:* the information gained from the implementation can also be used to iteratively refine the designed artifact until a satisfactory solution is found. Artifact improvements can also come through examination of the artifact's design by researchers who may identify improvements in approaches for implementing tasks recommended by the IKDDM model as well as improvements in the sequencing of tasks by refining or extending existing dependency relationships.

(3) *Implementing the end-to-end KDDM process through data mining software:* at present data mining software such as SAS Enterprise Miner, SPSS Clementine, Angoss Knowledge Studio, IBM Intelligent Miner etc only support the execution of the modeling phase of the KDDM process, and to some extent the data preparation phase. This is evident right from the moment the user launches the data mining software, and is asked to select the data to be used in building the models. However, in the actual KDDM process the selection of data is preceded by numerous other tasks which are not presently supported by data mining software.

Likewise, while data mining software have helped in automating the running of modeling techniques, the evaluation of the results generated by the plethora of modeling techniques is still largely the responsibility of the user. In many cases the output of

394

modeling techniques has to be studied outside the data mining software if large numbers of models are to be compared. Future research is needed in the area of data mining software that "implements" the end-to-end KDDM process as highlighted by KDDM process models in their design. SPSS Clementine has refined its interface to include a projects tool which provides a central location for storing all the material related to the various phases of the CRISP-DM process model (Source: Clementine Help Manual). However, the user is still responsible for carrying out all the tasks and the projects tool only provides a repository for storing any files, tables, graphs, white papers etc associated with the project.
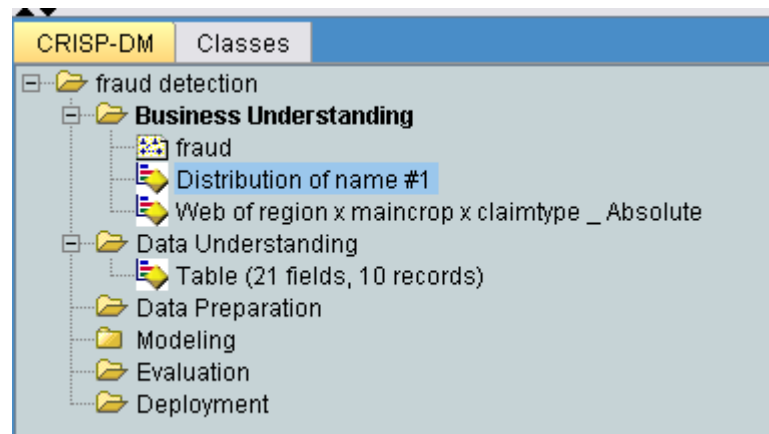
Figure 7-1: SPSS Clementine 12.0 interface – projects tool

(4) *Architecture to support the implementation of KDDM process:* As shown in this research, the implementation of the KDDM process requires heterogeneous components to work together. Examples include: requirements elicitation tools (to support elicitation of business, legal, technical requirements), AHP-type tools (to set up

www.manaraa.com

preference functions using data mining success criteria), organization-ontology (to support identification of relevant individuals) etc. This research explores tools and techniques applicable to various tasks. The next step would be to integrate the various components to design an architecture to implement the KDDM process, and to provide a proof-of-concept to demonstrate the workability of the artifact.

(5) *Extending PMML to include more than modeling results:* PMML or predictive markup modeling language is an XML based language to describe statistical and data mining models. It was developed by the Data Mining Group (www.dmg.org). PMML describes the inputs to data mining models, the transformations used prior to prepare data for data mining, and the parameters which define the models themselves. The general structure of a PMML document is presented in Figure 7-2 (Source: http://www.dmg.org/v3-0/GeneralStructure.html). Its main selling point is that it provides a means of sharing and deploying data models with other PMML aware tools (Swoyer 2005).

```
<?xml version="1.0"?>
<PMML version="3.0"
  xmlns="http://www.dmg.org/PMML-3_0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >

  <Header copyright="Example.com"/>
  <DataDictionary> ... </DataDictionary>

  ... a model ...

</PMML>
```

Figure 7-2: General structure of a PMML document

In this sense, it contributes towards the concept of knowledge reuse as it allows for the documentation of the models. While at present PMML only documents some aspects of modeling, it can be extended to include results from other phases such as the business objectives and data mining objectives behind the model, the business and data mining success criteria, the individuals involved in the project, the results of data understanding, etc. This would allow for documentation of the whole data mining project which is acknowledged as difficult to implement in practice (Becker and Ghedini 2005) due to the documentation burden involved.

397

References

Aalst, W. and A. Kumar (2003). "XML Based Schema Definition for Support of Interorganizational Workflow." Information Systems Research **14**(1): 23-46.

Ainsworth. "Logistic Regression: Class Presentation (PSY 524)." Retrieved 04/15/2008, from www.csun.edu/~ata20315/psy524/docs/Psy524%20lecture%2018%20logistic.ppt.

Anahory, S. and D. Murray (1997). Data Warehousing in the Real World: a Practical Guide for Building Decision Support Systems. England, Addison-Wesley.

Anand, S. and A. Buchner (1998). Decision Support Using Data Mining. London, Financial Times Pitman Publishers.

Applegate, L. M. and J. L. King (1999). "Rigor and relevance: Careers on the line." MIS Quarterly **23**(1): 17-18.

Basili, V. R. and H. D. Rombach (1988). "The TAME Project: Towards improvement–oriented software environments." IEEE Transactions on Software Engineering **14** (6): 758–773.

Basili, V. R. and D. M. Weiss (1984). "A methodology for collecting valid software engineering data." IEEE Transactions on Software Engineering **10**(6): 728–738.

Becker, K. and C. Ghedini (2005). "A documentation infrastructure for the management of data mining projects." Information and Software Technology **47**: 95–111.

Benbasat, I. and R. W. Zmud (1999). "Empirical Research in Information Systems: The Practice of Relevance." MIS Quarterly **23**(1).

Bernstein, A., S. Hill, et al. (2005). "Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification." IEEE Transactions on Knowledge and Data Engineering **17**(4): 503-518

Berry, M. and G. Linoff (1997). Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons.

Berry, M. and G. Linoff (2000). Mastering Data Mining: The Art and Relationship of Customer Relationship Management

John Wiley and Sons

Boylan, G. L., E. S. Tollefson, et al. (2006). "Using value-focused thinking to select a simulation tool for the acquisition of infantry soldier systems." Systems Engineering **9**(3): 199 - 212

Brachman, R. and T. Anand (1996). The Process of Knowledge Discovery in Databases. Advances in Knowledge Discovery and Data Mining. U. M. Fayyad, G. Paitetsky-Shapiro, P. Smyth and R. Uthuruswamy, AAAI Press/MIT Press.

Brachman, R. and T. Anand. (1996). The process of knowledge discovery in databases: a human-centered approach. Advances in Knowledge Discovery and Data

Mining. U. Fayyad, G. Paitestsky-Shapiro, P. Smith and R. Uthuruswamy. Menlo Park, AAAI Press**:** 36–57.

Cabena, P., P. Hadjinian, et al. (1998). Discovering Data Mining: From Concepts to Implementation., Prentice Hall.

Charest, M., S. Delisle, et al. (2006). Intelligent Data Mining Assistance via CBR and Ontologies. Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06).

Chin, W. (1998). The partial least squares approach for structural equation modeling. Modern Methods for Business Research,. G. A. Marcoulides and N. J. Mahwah, Lawrence Erlbaum Associates**:** 295–336.

Cios, K. and L. Kurgan (2005). Trends in Data Mining and Knowledge Discovery. Advanced Techniques in Knowledge Discovery and Data Mining. N. Pal and L. Jain, Springer**:** 1-26.

Cios, K., A. Teresinska, et al. (2000). "Diagnosing myocardial perfusion from PECTbull's-eye maps - a knowledge discovery approach." IEEE Engineering in Medicine and Biology Magazine, Special Issue on Medical Data Mining and Knowledge Discovery **19**(4): 17-25.

Codd, E. F., S.B.Codd, et al. (1993). "Beyond Decision Support." Computerworld **26**(July).

Cortes, C. and M. Mohri (2004). Confidence Intervals for the Area Under the ROC Curve Advances in Neural Information Processing Systems: Proceedings of the 2004 Conference

CRISP-DM. (2003). "Cross Industry Standard Process for Data Mining 1.0: Step by Step Data Mining Guide." Retrieved 01/10/07, from http://www.crisp-dm.org/.

Davenport, T. H. and J. G. Harris (2007). Competing on Analytics, Harvard Business School Press.

Davenport, T. H. and M. L. Markus (1999). "Rigor vs. relevance revisited: Response to Benbasat and Zmud." MIS Quarterly **23**(1): 19-23.

Delbecq, A. and A. H. V. d. Ven (1971). "A Group Process Model for Problem Identification and Program Planning." Journal of Applied Behavioral Science **7**: 466-492.

DeLone, W. H. and E. R. McLean (1992). "Information systems success: The quest for the dependent variable." Information Systems Research **3**(1): 60-95.

Diamantopoulos, A. and H. Winklhofer (2001). "Index Construction with Formative Indicators: An Alternative to Scale Development." Journal of Marketing Research **38**(2): 269-277.

Domingos, P. (2007). "Towards Knowledge-Rich Data Mining " Knowledge Discovery and Data Mining **15**(1): 21-28.

Doran, G. T. (1981). "There's a S.M.A.R.T. Way to Write Management Goals and Objectives." Management Review (AMA Forum): 35-36.

Dowson, M. (1993). Software Process Themes and Issues. Proceedings of the IEEE 2nd International Conference on the software process.

Draper, N. R. and H. Smith (1998). Applied Regression Analysis, Wiley-Interscience.

Drucker, P. F. (1954). <u>The Practice of Management</u>, Harper and Row.

Dyer, J. S. (1990). "Remarks on the Analytic Hierarchy Process." <u>Management Science</u> **36**(3): 249-258.

Egan, P. (1975). <u>Signal Detection Theory and ROC Analysis</u>, Academic Press.

Engels, R. (1999). Component-based user guidance for knowledge discovery and data mining processes. Karlsruhe, University of Karlsruhe. **PhD**.

Fayyad, U., G. Piatetsky-Shapiro, et al. (1996b). From Data Mining to Knowledge Discovery: An Overview. <u>Advances in Knowledge Discovery and Data Mining</u>, AAAI Press**:** 1-34.

Fayyad, U. and R. Uthurusamy (2002). "Evolving Data Mining into Solutions for Insights." <u>Communications of the ACM</u> **45**(8): 28-31.

Fayyad, U. M., G. Piatetsky-Shapiro, et al., Eds. (1996a). <u>Advances in Knowledge Discovery and Data Mining</u>, AAAI Press/MIT Press.

Feiler, P. H. and W. H. Humphrey (1993). <u>Software Process Development and Enactment: Concepts and Definitions</u> Proceedings of the 2nd International Conference on Software Process.

Field, A. (2000). <u>Discovering Statistics using SPSS for Windows</u> Sage Publications

Fornell, C. and D. Larcker (1981). "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error." <u>Journal of Marketing Research</u> **18**: 39-50.

Fox, M. S., M. Barbuceanu, et al. (1998). An Organization Ontology for Enterprise Modeling. <u>Simulating Organizations: Computational Models of Institutions and Groups</u>. Menlo Park CA, AAAI/MIT Press**:** 131-152.

French, S. (1998). <u>Decision Theory: An Introduction to the Mathematics of Rationality</u>. Chichester.

Gavish, B. and J. Gerdes (1998). "Anonymous Mechanisms in Group Decision Support Systems Communiciation." <u>Decision Support Systems</u> **23**(4): 297-328.

Gruber, T. R. (1993). "A Translation Approach to Portable Ontology Specifications." <u>Knowledge Acquisition</u> **5**(2): 199-220.

Han, J. and N. Cercone (2000). <u>RuleViz: a model for visualizing knowledge discovery process</u>. 6th ACM SIGKDD International Conference on Knowledge Discover and Data Mining.

Han, J. and M. Kamber (2001). <u>Data Mining: Concepts and Techniques</u>, Morgan Kaufmann.

Han, J. and M. Kamber (2006). <u>Data Mining: Concepts and Techniques</u>, Elsevier.

Hassan, O. A. B. (2004). "Application of value-focused thinking on the environmental selection of wall structures " <u>Journal of Environment Management</u> **70**(2): 181-187.

Hevner, A. R., S. T. March, et al. (2004). "Design Science in Information Systems Research." <u>MIS Quarterly</u> **28**(1): 75-105.

Holder, R. D. (1990). "Some Comment on the Analytic Hierarchy Process." <u>Journal of the Operational Research Society</u> **41**(11): 1073-1076.

Holder, R. D. (1991). "Response to Holder's Comments on the Analytic Hierarchy Process: Response to the Response"." The Journal of the Operational Research Society **42**(10): 914-918

Hulland, J. (1999). "Use of partial least squares (PLS) in strategic management research: a review of four recent studies." Strategic Management

Journal of Applied Behavioral Science **20**(2): 195–204.

Inmon, W. H. (1992a). Building the Data Warehouse. New York, Wiley.

Inmon, W. H. (1992b). "EIS and the Data Warehouse." Database Programming and Design **November**.

Järvinen, P. (2000). Research Questions Guiding Selection of an Appropriate Research Method European Conference on Information Systems, Vienna.

Kajanus, M., J. Kangas, et al. (2004). "The use of value focused thinking and the A'WOT hybrid method in tourism management." Tourism Management **25**(4): 499-506

KDNuggets. (2007, 08/2007). "Poll: Data Mining Methodology " Retrieved 11/6/2007, from http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.

Keeney, R. L. (1992). Value Focused Thinking: A path to creative decision making. Cambridge, MA, Harvard University Press.

Keeney, R. L. (1994). "Creativity in Decision Making with Value-Focused Thinking." Sloan Management Review **35**: 33-41.

Keeney, R. L. (1996). "Value-Focused Thinking: Identifying Decision Opportunities and Creating Alternatives." European Journal of Operations Research **92**: 537-549.

Keeter, R. R. and G. Parnell (2005). Applying Value-Focused Thinking to Effects Based Operations

Military Academy West Point NY, Department Of System Engineering.

Kurgan, L. A. and P. Musilek (2006). "A survey of Knowledge Discovery and Data Mining Process Models." The Knowledge Engineering Review **21**(1): 1-24.

Lee, A. S. (2000). Systems Thinking, Design Science, and Paradigms: Heeding Three Lessons from the Past to Resolve Three Dilemmas in the Present to Direct a Trajectory for Future Research in the Information Systems Field. Keynote Address, Eleventh International Conference on Information Management. Taiwan.

Linstone, H. A. and M. Turoff (1975). The Delphi Method: Techniques and Applications.

Maes, A. and G. Poels (2006). Evaluating Quality of Conceptual Models Based on User Perceptions 25th International Conference on Conceptual Modeling, Tucson, AZ.

Marakas, G. M. (2003). Decision Support Systems in the 21st Century. Upper Saddle River, NJ, Prentice Hall.

March, S. T. and G. F. Smith (1995). "Design and Natural Science Research on Information Technology." Decision Support Systems **15**: 251-266.

Markus, M. L., A. Majchrzak, et al. (2002). "A Design Theory for Systems that Support Emergent Knowledge Processes." <u>MIS Quarterly</u> **26**(3): 179-212.

Medina, C. and L. Pratt (1995). <u>NNES: A Neural Network Explanation System for Transforming Trained Neural Networks into Decision Trees</u>. Proceedings of the Twelfth National Conference on International Joint Conference on Artificial Intelligence

Menard, S. (2002). <u>Applied logistic regression analysis</u>. Thousand Oaks, CA:, Sage Publications.

Mingers, J. (2001a). "Combining IS research methods: towards a pluralist methodology." <u>Information Systems Research</u> **12**: 240–259.

Nevo, D. and Y. Wand (2005). "Organizational memory information systems: a transactive memory approach." <u>Decision Support Systems</u> **39**(4): 549-562.

Norman, D. A. (1988). <u>The Design of Everyday Things</u>. New York, Basic Books.

Nunnally, J. C. (1978). <u>Psychometric Theory</u>. New York, McGraw Hill.

Nutt, P. C. (2007). "Intelligence gathering for decision making." <u>Omega</u> **35**(5): 604-622.

Orlikowski, W. J. and C. S. Iacono (2001). "Research Commentary: Desperately Seeking the "IT" in IT Research — A Call to Theorizing the IT Artifact." <u>Information Systems Research</u> **12**(2): 121–134.

Osei-Bryson, K.-M. (2004). "Evaluation of Decision Trees." <u>Computers and Operations Research</u> **31**: 1933-1945.

Pöyhönen, M. A., R. P. Hämäläinen, et al. (1997). "An Experiment on the Numerical Modelling of Verbal Ratio Statements." <u>Journal of Multi-Criteria Decision Analysis</u> **6**(1): 1-10.

Pruitt, K. A. (2003). Modeling Homeland Security: A value focused thinking approach. Ohio, Air Force Institute of Technology. **Master of Science in Operations Research:** 237.

Pyle, D. (1999). <u>Data Preparation for Data Mining</u>, Morgan-Kaufmann.

Pyle, D. (2003). <u>Business Modeling and Data Mining</u>, Morgan Kaufmann Publishers.

Ragel, A. and B. Crémilleux (1998). <u>Treatment of Missing Values for Association Rules</u> Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Redpath, R. and B. Srinivasan (2003). <u>Criteria for a Comparative Study of Visualization Techniques in Data Mining</u>. IEEE 3rd International Conference On Intelligent Systems Design and Application, Tulsa, USA, Springer-Verlag, Berlin.

Reinartz, T. (2002). Stages of the Discovery Process. <u>Handbook of Data Mining and Knowledge Discovery</u>. W. Klosgen and J. Zytkow, Oxford University Press**:** 185-192.

Ringle, C. M., S. Wende, et al. (2005). SMART-PLS. Hamburg, Germany, University of Hamburg.

Rolland, C. A. (1998). <u>A Comprehensive View of Process Engineering</u>. Proceedings of the 10th International Conference, CAiSE

Saaty, T. L. (1980). <u>The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation</u>, McGraw-Hill

Saaty, T. L. (1991). "Response to Holder's Comments on the Analytic Hierarchy Process." The Journal of the Operational Research Society **42**(10): 909-914

Safavian, S. R. and D. Landgrebe (1991). "A Survey of Decision Tree Classifier Methodology." IEEE Transactions on Systems, Man, and Cybernetics **21**(3): 660-674.

Schenkerman, S. (1997). "Inducement of nonexistent order by the analytic hierarchy process." Decision Sciences **28**(2): 475-482.

Schon, D. A. (1990). The Design Process. Varieties of thinking: Essays from Harvard's Philosophy of Education Research Center. V. A. Howard. New York, Routledge**:** 110-141.

Seddon, P. B. (1997). " A respecification and extension of the DeLone and McLean model of IS success." Information Systems Research **8**(3): 240- 253.

Sharma, S. and K.-M. Osei-Bryson (2008). "Framework for Formal Implementation of the Business Understanding Phase of Data Mining Projects." Expert Systems with Applications **In press, corrected proof**.

Sharma, S. and K.-M. Osei-Bryson (2008). Organization-Ontology Based Framework for Executing the Business Understanding Phase of Data Mining Projects. Hawaii International Conference on Systems Sciences, Hawaii.

Sheer, A.-W. and A. Hars (1992). "Extending data modelling to cover the whole enterprise." Communications of the ACM **35**(9): 166–172.

Shintani, T. (2006). Mining Association Rules from Data with Missing Values by Database Partitioning and Merging. International Workshop on Component-Based Software Engineering,Software Architecture and Reuse.

Siddiqi, N. (2005). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring John Wiley and Sons, Inc.

Simon, H. A. (1996). The Sciences of the Artificial. Cambridge, MA, MIT Press.

Simoudis, E., B. Livezey, et al. (1996). Integrating Inductive and Deductive Reasoning for Data Mining. Advances in Knowledge Discovery and Data Mining U. Fayyad, G. Paitetsky-Shapiro, P. Smyth and R. Uthurusamy, AAAI Press/MIT Press.

Stein, E. W. (1995). "Organizational Memory: Review of concepts and recommendation of management." International Journal of Information Management **15**(1): 17-32.

Swoyer, S. (2005). PMML: Data Mining for the Masses? Enterprise Systems.

Uthurusamy, R. (1996). From Data Mining to Knowledge Discovery: Current Challenges and Future Directions. Advances in Knowledge Discovery and Data Mining. U. Fayyad, G. Paitetsky-Shapiro, P. Smyth and R. Uthurusamy, AAAI Press/The MIT Press.

Wirth, R. and J. Hipp (2000). CRISP-DM: towards a standard process model for data mining. Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining.

Wirth, R., C. Shearer, et al. (1997). Towards process-oriented tool support for knowledge discovery in databases. <u>Principles of Data Mining and Knowledge Discovery</u>, Springer Berlin / Heidelberg. **1263/1997:** 243–253.

Wong, W., P.J.Fos, et al. (2003). "Combining the performance strengths of the logistic regression and neural network models: a medical outcomes approach." <u>Scientific World Journal</u> **3**: 455-476.

Yoo, Y. and M. Alavi (2001). "Media and group cohesion: Relative influences on social presence, task participation, and group consensus,." <u>MIS Quarterly</u> **25**: 371-390.

Zantout, H. and F. Marir (1999). "Document management systems from current capabilities towards intelligent information retrieval: an overview." <u>International Journal of Information Management</u> **19**(6): 471–484.

APPENDIX A: TEST INSTRUMENT

Test Instrument
Evaluation of a Knowledge Discovery and Data Mining Process Model

Please note:

- There are a total of 15 multiple choice questions on this test.

- Please **SELECT THE BEST ANSWER** for these multiple choice questions. Each question has only one correct answer.

- **PLEASE USE THE EXTRACT DOCUMENT CREATED TO ASSIST YOU IN ANSWERING THESE QUESTIONS.** For each question, only the relevant portion from the documentation of a Data Mining Process Model has been included.

- **AFTER COMPLETING THE TEST, PLEASE ANSWER THE QUESTIONS ON THE BRIEF SURVEY** aimed at assessing your experience with using the process model to answer the questions on the test.

- **FOR EACH QUESTION ON THE TEST AND SURVEY, PLEASE ENCIRCLE OR PUT A CHECK MARK AGAINST THE OPTION OF YOUR CHOICE.**

All questions are based on data mining projects typically carried out by organizations engaged in furthering their Sales/Marketing/CRM type applications. Please use only the information provided in each question to answer the question. Please do not make any assumptions, or employ information beyond what is explicitly provided in the question.

ID (please leave blank; for investigator use only:

Before starting the test, please enter the following information about yourself

Gender:

Role/Title (if you are a student, please enter 'Student' and the degree program you are engaged in; if you are working)**:**

Number of years of Data Mining Experience (if you are a student, please enter number of years you have been studying Data Mining)**:**

Start Time:
End Time:

Questions Start on next page
****

Use the following hypothetical problem scenario to answer questions 1-6

Consider the case of a telecommunications services firm called ABC Global. The firm is facing the issue of losing its existing customers to its competitors. On further analysis the firm finds that it is the customers who have been with the firm for more 2 years (i.e. whose tenure is >2 years), who are most likely to leave (or churn). At present, 7% of the customers are churning away and this is resulting in a loss of $1 million for the company. The company wishes to bring this rate of churn down to 3% over the Financial year 2008-2009.

In order to deal with this situation the company wishes to identify the 10,000 customers who are most likely to leave in the next three months, in order to target them in time with new offers, enticing them to stay. The firm has applied data mining techniques to such projects in the past with varying degrees of success. Given the importance of this project, the firm wishes to use a process model to guide the formal execution of various tasks of the project.

Question 1: Which of the following statements of business objectives reflects the business objective of the data mining project being pursued by ABC Global?

    A) Reduce Churn rate of existing customers to 4% by 2009
    B) Reduce Churn rate of customers with tenure >2 to 3% over 2008-2009
    C) Predict the probability to churn of customers with tenure > 2 over 2008-2009
    D) Increase profits by reducing churn rate of customers with tenure >2 to 4% over 2008-2009

Question 2: What are the business success criteria for the above project?

    A) Δ (Delta) churn rate of 4%
    B) Δ (Delta) profits of $1 million
    C) Δ (Delta) losses of $2 million
    D) Δ (Delta) churn rate of 3%

Question 3: Which of the following data mining problem type best represents the given problem scenario?

    A) Classification
    B) Prediction
    C) Visualization/Description
    D) Association Rules Mining/Dependency analysis

Question 4: Which of the following statements of data mining objectives can be regarded as the most appropriate one for this business scenario?

    A) Predict the likelihood to churn of customers with tenure >2 years
    B) Estimate the churn rate of customers with tenure > 2 years
    C) Cluster customers with tenure > 2 to identify those that are likely to churn
    D) Predict the churn rate of customers with tenure >2 years

Question 5: Which of the following modeling techniques can be used to address the given problem?

    A) Logistic Regression
    B) Linear Regression
    C) Regression (decision) Tree
    D) K means clustering

Question 6: If a modeling team decided to use regression decision trees (target variable: continuous) to address a data mining problem, which of the following data mining success criterion could be used to assess the performance of the model?

    A) Accuracy
    B) Lift chart
    C) Number of clusters
    D) Confusion matrix

Question 7: The modeling team of a major retailer is employing the SAS Enterprise Miner 4.3 in developing their models. The technical team head asks his team to report to him the KS Statistic (a data mining success criterion) of the model along with other details. Which of the following statements is true?

    A) The modeling team cannot report this statistic as it does not apply to decision trees
    B) The modeling team does not have direct access to the statistic but can calculate it based on the confusion matrix
    C) The modeling team cannot report this statistic as it is not available in SAS EM 4.3
    D) The modeling team can calculate the statistic based on the misclassification rate on the test data.

Question 8: Which of the following data mining success criterion applies to both classification problems and association rules?

      A) Area under ROC curve
      B) KS (Kolmogorov-Smirnoff) statistic
      C) Support
      D) Lift

Question 9: Select which of the following statements is TRUE?

      A) Accuracy is an important data mining success criteria for both classification and estimation problems
      B) The business objective is derived from the data mining objective and is in fact the technical translation of the data mining objective.
      C) Legal requirements must be addressed after running of modeling algorithms
      D) None of the above.

Question 10: Which of the following data mining success criterion can be used to assess output of clustering algorithms?

      A) Interest factor
      B) Mean square error
      C) Variable importance vectors
      D) Number of dimensions

Question 11: The business team head of a major retailer consults with his team to find out any variables that do not exist in the present data set. He suggests that they look at incorporating some derived variables or ratio variables in the new model. Which of the following is true?

      A) This is a good idea and addition of any new variables not existing in the old model are likely to lead to improved results
      B) This is a bad idea as variables not present in the old model cannot be used in the new model
      C) This is a good idea, but to eliminate bias, all possible ratio variables must be created (by dividing each variable by other variables) and using these in developing the new model
      D) This can be a good or bad idea depending on whether or not it is backed by business reasons.

Question 12: How are the modeling parameters depth and breadth of a decision tree related to the accuracy and efficiency of the tree?

A) Modeling parameter breadth is related to the accuracy of the tree, whereas depth is related to the efficiency of the tree.
B) Modeling parameter depth is related to the accuracy of the tree, whereas breadth is related to the efficiency of the tree.
C) Both depth and breadth are related to accuracy of the tree, but neither is related to the efficiency of the tree.
D) There is no relation between modeling parameters depth and breadth and the accuracy and efficiency of the tree.

Question 13: In studying the data during the data understanding phase, the technical team of a leading bank, finds various outliers and missing values. Assuming that they intend to use logistic regression during the modeling phase, which of the following apply?

A) Outliers need not be removed as logistic regression is also unaffected by the skewed distribution of variables
B) Outliers need to be removed as they can have a strong effect on the results.

Question 14: Which of the following data mining success criteria apply to classification modeling techniques (for e.g., classification trees, logistic regression, Naïve Bayes etc)?

A) Accuracy and Lift
B) Area under ROC curve and Support
C) Lift and Frequency
D) Accuracy, Lift, and KS statistic

Question 15: While evaluating the modeling results, it must be ensured that,

A) At least the data mining success criteria are satisfied
B) All business success criteria and all data mining success criteria must be satisfied
C) At least one business success criterion and one data mining success criterion is satisfied
D) At least the business success criteria are satisfied

****
**

End of Test

Thank you for your participation!

APPENDIX B: Survey Instrument

Id:

Follow Up Survey

Thank you for taking the time out for completing the test questionnaire. Please answer the following survey questions based on your experience with using the (Knowledge Discovery and Data Mining) KDDM process model. For each statement, please select your response from the 7 options listed below each statement.

1. It was easy for me to understand what the KDDM model was trying to model.

> A. Strongly Disagree
> B. Disagree
> C. Moderately Disagree
> D. Undecided
> E. Moderately agree
> F. Agree
> G. Strongly Agree

2. Overall, I think the KDDM model would be an improvement to a textual description of the KDDM process.

> A. Strongly Disagree
> B. Disagree
> C. Moderately Disagree
> D. Undecided
> E. Moderately agree
> F. Agree
> G. Strongly Agree

3. Using the KDDM model was often frustrating.

> A. Strongly Disagree
> B. Disagree
> C. Moderately Disagree
> D. Undecided
> E. Moderately agree
> F. Agree
> G. Strongly Agree

410

4. Overall, I found the KDDM model useful for understanding the process modeled.

   A. Strongly Disagree
   B. Disagree
   C. Moderately Disagree
   D. Undecided
   E. Moderately agree
   F. Agree
   G. Strongly Agree

5. Overall, the KDDM model was easy to use.

   A. Strongly Disagree
   B. Disagree
   C. Moderately Disagree
   D. Undecided
   E. Moderately agree
   F. Agree
   G. Strongly Agree

6. Overall, I think the KDDM model improves my performance when understanding the process modeled.

   A. Strongly Disagree
   B. Disagree
   C. Moderately Disagree
   D. Undecided
   E. Moderately agree
   F. Agree
   G. Strongly Agree

7. Learning how to read the KDDM model was easy.

   A. Strongly Disagree
   B. Disagree
   C. Moderately Disagree
   D. Undecided
   E. Moderately agree
   F. Agree
   G. Strongly Agree

8. The KDDM model represents the KDDM process correctly.

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

9. The KDDM model adequately met the information needs that I was asked to support.

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

10. The KDDM model is a realistic representation of the KDDM process.

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

11. The KDDM model was not efficient in providing the information I needed.

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

12. The KDDM model contains contradicting elements.

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

13. The KDDM model was effective in providing the information I needed.

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

14. All the elements in the KDDM model are relevant for the representation of the KDDM process

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

15. Overall, I am satisfied with the KDDM model for providing the information I needed.

A. Strongly Disagree
B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

16. The KDDM model gives a complete representation of the KDDM process

A. Strongly Disagree

413

B. Disagree
C. Moderately Disagree
D. Undecided
E. Moderately agree
F. Agree
G. Strongly Agree

APPENDIX C: Extract Document for CRISP-DM Process Model

Extract Document with Relevant Portions for Each Question

Question 1.

Output	**Business objectives** Describe the customer's primary objective, from a business perspective, in the data mining project. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor, while secondary business objectives might be to determine whether lower fees affect only one particular segment of customers.
Activities	▪ Informally describe the problem which is supposed to be solved with data mining. ▪ Specify all business questions as precisely as possible. ▪ Specify any other business requirements (e.g., the business does not want to lose any customers). ▪ Specify expected benefits in business terms.

Question 2.

Output	Business success criteria

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and readily measurable, such as reduction of customer churn to a certain level or general and subjective such as "give useful insights into the relation-ships." In the latter case it should be indicated who would make the subjective judgment.

Activities

- Specify business success criteria (e.g., improve response rate in a mailing campaign by 10 percent and sign-up rate increased by 20 percent).
- Identify who assesses the success criteria.

Question 3.

(A) Classification

Classification assumes that there is a set of objects – characterized by some attributes or features – which belong to different classes. The class label is a discrete (symbolic) value and is known for each object. The objective is to build classification models (sometimes called classifiers), which assign the correct class label to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling.

(B) Prediction

Another important problem type that occurs in a wide range of applications is *prediction*. Prediction is very similar to classification. The only difference is that in prediction the target attribute (class) is not a qualitative discrete attribute but a continuous one. The aim of prediction is to find the numerical value of the target attribute for unseen objects. In the literature, this problem type is sometimes called regression. If prediction deals with time series data then it is often called forecasting.

(C) Visualization/Description

Data Description and Summarization aims at the concise description of characteristics of the data, typically in elementary and aggregated form. This gives the user an overview of the structure of the data. Sometimes, data description and summarization alone can be an objective of a data mining project. For instance, a retailer might be interested in the turnover of all outlets broken down by categories. Changes and differences to a previous period could be summarized and highlighted. This kind of problem would be at the lower end of the scale of data mining problems.

(D) Association Rules Mining/Dependency Analysis

Dependency analysis consists of finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of a data item given information on other data items. Although dependencies can be used for predictive modeling, they are mostly used for understanding. Dependencies can be strict or probabilistic.

Associations are a special case of dependencies, which have recently become very popular. Associations describe affinities of data items (i.e., data items or events which frequently occur together). A typical application scenario for associations is the analysis of shopping baskets. There, a rule like "in 30 percent of all purchases, beer and peanuts have been bought together" is a typical example for an association.

Question 4.

Task	**Determine data mining goals** A *business goal* states objectives in business terminology; a *data mining goal* states project objectives in technical terms. For example, the business goal might be "Increase catalogue sales to existing customers" while a data mining goal might be "Predict how many widgets a customer will buy, given their purchases over the past three years, relevant demographic information and the price of the item."
Output	**Data mining goals** Describe the intended outputs of the project that enable the achievement of the business objectives. Note that these are normally *technical* outputs.
Activities	■ Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified). ■ Specify data mining problem type (e.g., classification, description, prediction and clustering). For more details about data mining problem types, see Appendix V.2, where they are described in more detail.

Question 5.

Task	**Select modeling technique**
	As the first step in modeling, select the actual modeling technique that is to be used. Whereas you possibly already selected a tool in business understanding, this task refers to the specific modeling technique, e.g., decision tree building with C4.5 or neural network generation with back propagation. If multiple techniques are applied, perform this task for each technique separately.
Outputs	**Modeling technique**
	Document the actual modeling technique that is to be used.

Question 6.

Task	**Assess model**
	The model should now be assessed to ensure that it meets the data mining success criteria and the passes the desired test criteria. This is a purely technical assessment based on the outcome of the modeling tasks.
Output	**Model assessment**
	Summarize results of this task, list qualities of generated models (e.g. in terms of accuracy) and rank their quality in relation to each other.
Activities	Evaluate result with respect to evaluation criteria

Question 7.

Output	**Initial assessment of tools and techniques**

At the end of the first phase, the project also performs an initial assessment of tools and techniques. Here, you select a data mining tool that supports various methods for different stages of the process, for example. It is important to assess tools and techniques early in the process since the selection of tools and techniques possibly influences the entire project.

Activities
- Create a list of selection criteria for tools and techniques (or use an existing one if available).
- Choose potential tools and techniques.
- Evaluate appropriateness of techniques.
- Review and priorities applicable techniques according to the evaluation of alternative solutions.

Question 8.

Output	**Data mining success criteria**

Define the criteria for a successful outcome to the project in technical terms, for example a certain level of predictive accuracy or a propensity to purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

Activities
- Specify criteria for model assessment (e.g., model accuracy, performance and complexity).
- Define benchmarks for evaluation criteria.
- Specify criteria which address subjective assessment criteria (e.g. model explain ability and data and marketing insight provided by the model).

Question 9.

A.

Output	**Data mining success criteria** Define the criteria for a successful outcome to the project in technical terms, for example a certain level of predictive accuracy or a propensity to purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.
Activities	▪ Specify criteria for model assessment (e.g., model accuracy, performance and complexity). ▪ Define benchmarks for evaluation criteria. ▪ Specify criteria which address subjective assessment criteria (e.g. model explain ability and data and marketing insight provided by the model).

B.

Task	**Determine data mining goals** A *business goal* states objectives in business terminology; a *data mining goal* states project objectives in technical terms. For example, the business goal might be "Increase catalogue sales to existing customers" while a data mining goal might be "Predict how many widgets a customer will buy, given their purchases over the past three years, relevant demographic information and the price of the item."

Output	**Data mining goals** Describe the intended outputs of the project that enable the achievement of the business objectives. Note that these are normally *technical* outputs.
Activities	▪ Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified). ▪ Specify data mining problem type (e.g., classification, description, prediction and clustering). For more details about data mining problem types, see Appendix V.2, where they are described in more detail.

C.

Output	**Requirements, assumptions and constraints** List all requirements of the project including schedule of completion, comprehensibility and quality of results and security as well as legal issues. As part of this output, make sure that you are allowed to use the data. List the assumptions made by the project. These may be assumptions about the data, which can be checked during data mining, but may also include non-checkable assumptions about the business upon which the project rests. It is particularly important to list the latter if they form conditions on the validity of the results. List the constraints made on the project. These constraints might involve lack of resources to carry out some of the tasks in the project within the timescale required or there may be legal or ethical constraints on the use of the data or the solution needed to carry out the data mining task.
Activities	**Requirements** ▪ Specify target group profile. ▪ Capture all requirements on scheduling. ▪ Capture requirements on comprehensibility, accuracy, deploy ability, maintainability and repeatability of the data mining project and the resulting model(s). ▪ Capture requirements on security, legal restrictions, privacy, reporting and project schedule.

Question 10.

Task	**Assess model** The model should now be assessed to ensure that it meets the data mining success criteria and the passes the desired test criteria. This is a purely technical assessment based on the outcome of the modeling tasks.
Output	**Model assessment** Summarize results of this task, list qualities of generated models (e.g. in terms of accuracy) and rank their quality in relation to each other.
Activities	Evaluate result with respect to evaluation criteria

Question 11.

Output	**Derived attributes** Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. An example might be area = length * width.

Why should we need to construct derived attributes during the course of a data mining investigation? It should not be thought that only data from databases or other sources is the only type of data that should be used in constructing a model. Derived attributes might be constructed because:

- Background knowledge convinces us that some fact is important and ought to be represented although we have no attribute currently to represent it.

- The modeling algorithm in use handles only certain types of data, for example we are using linear regression and we suspect that there are certain non-linearities that will be not be included in the model.

- The outcome of the modeling phase may suggest that certain facts are not being covered.

Activities	**Derived attributes**
	- Decide if any attribute should be normalized (e.g. when using a clustering algorithm with age and income in lire, the income will dominate).
	- Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization).
	- How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction)].
	- Add new attributes to the accessed data.

Question 12.

Output	**Parameter settings**
	With any modeling tool, there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice.
Activities	- Set initial parameters.
	- Document reasons for choosing those values.

Question 13.

Output	**Modeling assumptions**
	Many modeling techniques make specific assumptions about the data, data quality or the data format.
Activities	- Define any built-in assumptions made by the technique about the data (e.g. quality, format, distribution).
	- Compare these assumptions with those in the Data Description Report.
	- Make sure that these assumptions hold and step back to the Data Preparation Phase if necessary.

Question 14.

Output	**Data mining success criteria** Define the criteria for a successful outcome to the project in technical terms, for example a certain level of predictive accuracy or a propensity to purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.
Activities	▪ Specify criteria for model assessment (e.g., model accuracy, performance and complexity). ▪ Define benchmarks for evaluation criteria. ▪ Specify criteria which address subjective assessment criteria (e.g. model explain ability and data and marketing insight provided by the model).

Question 15.

Model assessment
This section describes the results of testing the models according to the test design.

Topics to be covered:
- Overview of assessment process and results including any deviations from the test plan. For each model:
    - Detailed assessment of model including measurements such as accuracy and interpretation of behavior.
    - Any comments on models by domain or data experts.
    - Summary assessment of model.
    - Insights into why a certain modeling technique and certain parameter settings led to good/bad results.
    - Summary assessment of complete model set.

Evaluation

425

Assessment of data mining results with respect to business success criteria

This report compares the data mining results with the business objectives and the business success criteria.

Topics to be covered:

- Review of Business Objectives and Business Success Criteria (which may have changed during and/or as a result of data mining).
  For each Business Success Criterion:
    - Detailed comparison between success criterion and data mining results.
    - Conclusions about achievability of success criterion and suitability of data mining process.
- Review of Project Success; has the project achieved the original Business Objectives?
- Are there new business objectives to be addressed later in the project or in new projects?
- Conclusions for future data mining projects.

Review of process

This section assesses the effectiveness of the project and identifies any factors that may have been overlooked that should be taken into consideration if the project is repeated.

List of possible actions

This section makes recommendations regarding the next steps in the project.

APPENDIX D: Extract document for IKDDM Process Model

Question 1: Setting up Business Objectives

Consider the following steps to formulate a business objective:

Step 1: Select Purpose

Purpose: the purpose signifies the motivation behind formulating the objective, or why the objective is being formulated. In the context of Data Mining projects, purpose can be of the following five types:

1. Increase/Improve

2. Decrease/Reduce

3. Identify

4. Understand

5. Determine (Hypothesis Testing)

Step 2: Select Object of study and its defining characteristic

Object Name and Defining Characteristic: the object is the entity under the study. Examples of objects can include: (1) Customers, (2) Suppliers, (3) Products, (4) Employees, (5) Transactions, etc.

427

In selecting the object it is important to provide further qualifying information in form of the defining characteristic of the object. For instance, if the object is chosen as simply 'customers', it may not be clear as to which customers of the firm are of interest and a resultant data mining endeavor may be based on the entire customer base of the firm. However, the results of data mining so obtained are likely to be diluted as it is well known that different types of customers behave differently. So when specifying the object, we must augment it by adding more information. See examples of various types of objects and their defining characteristics in table 1 below).

Table 1: Objects and their Defining Characteristics

Objects	Defining Characteristics
Customers	Wireless internet Customers
	Customers with tenure > 1
	Customers acquired though marketing channel
	most loyal Customers
Suppliers	Suppliers for Eastern Region
	Suppliers of small moving parts
	Suppliers of parts X
Products	co-selling Products
	Products from a particular line (baby care or feminine products)
Employees	internal Hires
	part time Employees
	full time Employees
	Contract Employees
	Employees with tenure > 5
Transactions	Transactions that occurred in last week/month/year
	Transactions valued at >$250

Step 3: Select Focus Variable (the variable of interest)

Focus: the focus is the variable or the quality attribute of the entity under study, i.e. what is being studied through the data mining project. The focus of a data mining

428

project can be on a tangible or quantitatively measurable behavior, or on an intangible attribute. Below we provide examples of both types.

Quantitative focus: such a focus variable can be measured in terms of %, rate, amount etc. For e.g., churn rate or loss rate of a CUSTOMER [OBJECT]

Qualitative focus: such a focus variable cannot be measured in terms of %, rate, amount etc. For e.g., factors affecting motivation of EMPLOYEES [OBJECT]

Step 4: Formulate Preliminary Business Objective using PURPOSE, OBJECT, AND FOCUS variable selected earlier

For example the preliminary business objective can be: Increase (PURPOSE) the approval rate (FOCUS) of sub-prime customers (OBJECT AND DEFINING CHARACTERISTIC

Step 5: Finalize business objective by:

- Adding information about **Time Frame** over which objective must be achieved

- Adding information about the **delta change if focus variable is quantitative**

For example the business objective can be refined as: Increase (PURPOSE) the approval rate (FOCUS) of sub-prime customers (OBJECT AND DEFINING CHARACTERISTIC by 4% (DELTA CHANGE IN FOCUS VARIABLE) over 2009-2010 (TIME FRAME)

429

This statement can be regarded as **FINAL statement of business objective**

Question 2: Setting up of Business Success Criteria

The Business Success Criteria can be calculated as:

Delta (change) in quantitative focus variable: if goal is to reduce loss rate from 5% to

2% then the business success criterion becomes achieving a Δ **loss rate = 3%** (i.e. if

loss rate reduces by 3%, business success criterion will be satisfied.

Question 3: The different data mining problem types are summarized below

Table 2: Supervised Data Mining problems (with target variable specified)

Problem Type	Definition	Example
Classification	Dividing unseen records into predefined classes	Divide records into • High, medium, low • Republican and Democrat States
Estimation	Estimating value of a continuous variable	Estimate annual income of households in zip code 23233
Prediction (Classification)	Classifying records into predefined classes based on "future behavior"	Classify customers into classes 'churn' and 'no churn'
Prediction (Estimation)	Estimating the "future" value of a continuous variable	Predicting the amount of balance that a customer will transfer if he accepts a credit card offer

Table 3: Unsupervised Data Mining problems (with no target variable)

Problem Type	Definition	Example
Clustering/Segmentation	Dividing records into clusters or segments	Identify different types of customers from overall customer base
Visualization	Study features, characteristics, factors, relationships	Identify characteristics of most loyal customers
Affinity grouping or association rules	Study co-occurrence of products or variables	Identify co-selling products from line of baby products

Question 4: Formulating Data Mining Objective

Consider the following steps to formulate a business objective:

<u>**Step 1: Select Purpose**</u>

Select Purpose from one of the following (see tables above for definitions)

8. Classification

9. Estimation

10. Prediction (Classification) if goal is to classify but based on future behavior

11. Prediction (Estimation) if goal is to estimate but based on future behavior

431

12. Visualization

13. Clustering

14. Affinity grouping or association rules (including sequential patterns)

Step 2: Select Focus Variable

The focus of a data mining goal cannot be divided into a finite set of categories.

- For Classification, estimation or prediction (classification or estimation) problems, the focus is the 'target variable' under the study.

- For Classification and Prediction (classification) problems, focus may be a 'Categorical Target' with two classes such as "good" or "bad", "churn" or "no churn" etc.

- For Estimation and Prediction (estimation) problems, focus may be a continuous target variable such as "household income", or "amount of balance transferred" etc.

- For Clustering problems, the focus is on the 'Types of Clusters or Segments (clusters of OBJECTS' with similar buying habits, of same age, having same spending pattern, buying similar products etc)

- For Association Rules/Affinity Grouping, the focus or the attribute under study is the 'co-occurrence of objects'

- For Visualization, the focus is on the 'factors, characteristics, relationships'

Step 3: Select Object and its defining characteristic

Step 3: Select OBJECT (entity under study), **OBJECT TYPE** (distinguishing characteristic of the entity) and **TIME FRAME** (period for which the object is to be studied).

- The OBJECT can be (1) customers, products, employees, suppliers, household, etc.
- The OBJECT TYPE can be sub prime applicants, bathing products, contract employees, small parts suppliers', households in zip code 19701.
- The TIME FRAME can be reflected as follows: sub prime credit card applicants 12 months from point of booking, bathing products sold in 2007-2008, contract employees with tenure > 2 years, small parts suppliers with tenure > 3 years, households in zip code 19701 for may 07-may 08.

Step 4: Formulate Data Mining Objective using PURPOSE, OBJECT, AND FOCUS variable selected earlier

For example the data mining objective can be: Predict (PURPOSE) the probability of charge-off (FOCUS) of student loan customers 18 months from the point of booking (OBJECT, DEFINING CHARACTERISTIC AND TIME FRAME)

433

Question 5: Look Up Table for Modeling Techniques by Data Mining Problem Type

Table 4: Modeling techniques based on target variable type and data mining problem type

Problem Type Target variable	Prediction	
	Classification	**Estimation**
binary	Logistic regression Classification Tree k-nearest neighbor Naïve Bayes* Neural network* Support Vector Machines* Genetic algorithm*	Not applicable
ordinal	Ordinal Logistic regression Classification Tree k-nearest neighbor Naïve Bayes* Neural network* Support Vector Machines* Genetic algorithm*	Not applicable
nominal	Multinomial Logistic regression Classification Tree k-nearest neighbor Naïve Bayes* Neural network* Support Vector Machines* Genetic algorithm*	Not applicable
Interval	Not Applicable	Regression Regression Tree k-nearest neighbor Memory Based Reasoning Neural Networks*

* Non explanatory technique (cannot produce rules)

Question 6: Look Up Table for Data Mining Success Criteria for Regression Techniques

Table 5: Data Mining Success Criteria for Regression Modeling Techniques

		Estimation Techniques		
		Regression Tree	Linear Regression	Neural Network
Data Mining Success	Accuracy (Average Squared Error)	✓	✓	✓
	Simplicity	✓	✓	×
	Stability	✓	✓	✓
	Profit/Loss	✓	✓	✓

Question 7: Data Mining Success Criteria supported by different Data Mining Software

Table 6: Data Mining Success Criteria for Classification Trees provided by Data Mining Software (SAS EM, Clementine)

435

www.manaraa.com

Measure	Source for Calculating Measure	SAS EM 4.3	SPSS Clementine 12.0
Accuracy	Test Misclassification Rate	Implicit Calculate using 1-Test Misclassification Rate	Explicit (Modeling results)
	Confusion Matrix	Implicit	Implicit
Lift or Gains Index	Visual Inspection of Lift Chart up to a particular Decile	Explicit-Visual	Explicit-Visual
	Lift Value can be estimated through analysis of lift chart	Implicit Calculate using Tree/Exact	Explicit (Modeling results)
Profit and Loss	Profit and Loss Matrix	Explicit (Modeling Results)	Explicit (also provides additional measures)
Simplicity	User Defined	Implicit (Calculate using Number of leaves, and/or Minimum Rule length)	Implicit (Calculate using Number of leaves)
Stability	User Defined	Implicit Calculate using a coarse measure such as Min $[ACCT_V/ACC_T, ACC_T/ACC_V]$ Where $ACCT_V$ is accuracy of validation data and ACCT is accuracy on training data	Implicit Models (by default) are built with generality. For assessing stability, validate against hold out sample
	Visual Inspection of Lift Chart at a particular decile	Explicit-Visual	Explicit-Visual
ROC curve	Plot of 1-specificity on x-axis and sensitivity on y axis.	Explicit-Visual Visual inspection of chart must be used to employ ROC as an evaluation measure	Explicit-Visual
Area under ROC Curve or AUC	Area calculated using trapezoidal rule	No	Explicit (Empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve and its 95% CI)
KS statistic (Komogorov-Smirnov)	Maximum KS value	No	No
Average Squared Error	Modeling Results	Explicit	No
Sensitivity	Confusion Matrix	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)	Implicit (Calculate using TP/[TP+FN] Where TP is the true positive rate and FN is the false negative rate)
Specificity	Confusion Matrix	Implicit Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate	Implicit Calculate using TN/[FP+TN] Where TP is the true positive rate and FN is the false negative rate

Question 8:

See Table provided with Question 7 for DATA MINING SUCCESS CRITERIA FOR CLASSIFICATION TECHNIQUES.

DATA MINING SUCCESS CRITERIA FOR ASSOCIATION RULES are included below

Table 7: Data Mining Success Criteria for Association Rules

Measure	Source for Calculating Measure	SAS EM 4.3	SPSS Clementine 12.0
Lift	Ratio of confidence to the prior probability of having the consequent	Explicit (Modeling results)	Explicit (Modeling results)
Excess	Lift-1	Implicit Calculate using lift-1	Implicit Calculate using lift-1
Simplicity	Length of Rule	Implicit Calculate using length of rule	Implicit Calculate using length of rule
Support	Proportion of ID's for which entire rule, antecedents, consequents are true	Explicit (Modeling results)	Explicit (Modeling results)
Confidence	Ratio of rule support to antecedent support	Explicit (Modeling results)	Explicit (Modeling results)
Interest Factor	ratio between the joint probability of two variables with respect to their expected probabilities under the independence assumption	No	No
Monetary Value	Profitability of a rule	Explicit (Modeling Results)	Explicit (Modeling Results)
Deployability	% of training data that satisfies the conditions of the antecedent but does not satisfy the consequent	No	Explicit (Modeling Results)

Question 9:

A Accuracy is an important criteria for both classification/prediction and estimation problems.

Accuracy for classification problems is measured in terms of the error rate, or the percentage of records classified incorrectly (Berry and Linoff 1997).

Accuracy for estimation problems is expressed as the difference between the predicted score and the actual measured result (Berry and Linoff 1997). Accuracy of one estimate as well as accuracy of the entire model is of importance. A model that only provides good accuracy for a certain range of input values cannot be regarded as a good estimator.

B The data mining objective is the technical translation of the business objective. Given this relationship the business objective must be created before creating the data mining objective

C Application of Policy and Legal Constraints

This is an important task of the data preparation phase. During this task the dataset created through various data sources is applied with policy and legal constraints to make sure that these constraints are not being violated. As an example of policy constraints, an organization may have a policy that a product would only be offered to

individuals 18 years or older in age. In such a case, any individuals whose age is less than 18 must be removed from the dataset to be used for analysis. As an example of legal constraints, law may require a firm to not make any decisions regarding offering products to customers on the basis of their sex or gender. In such a case, such variables must be removed from analysis.

Question 10:

Table 8: Look Up Table for Unsupervised Data Mining techniques including Clustering, Association Rules and Description/Visualization.

Data Mining Problem Type	Data Mining Success Criteria
Clustering	Normalized cluster means, Variable Importance Vectors, Ou Usefulness
Association Rules	Lift, Simplicity (Rule length), Support, Confidence, Recall, Precision, Interest Factor, Expected Monetary Factor, Incremental Monetary Factor
Description or Visualization	Number of instances in data set, Number of dimensions, Overlapping data instances, Ability to reveal patterns in dataset, Ability to reveal clusters of two or three dimensions, Number of clusters present, Amount of background noise, Variance of clusters, Ability to manipulate display automatically, Ease of Use

Question 11: Rationale for creating Derived Attributes or Ratio Variables

During this task, the decision makers must assess the data to make decisions regarding creation of derived attributes that are needed to adequately address the data mining objective. A meta database containing business metadata can be helpful for analysis of

possibility of derived attributes. The business metadata helps assess (1) whether or not aggregating certain variables makes business sense and (2) ensures that the policy constraints (often laid out as business rules) are not being violated. The formulae and reasoning behind creation of derived attributes must be clearly documented.

Siddiqi (2005) highlights that users involved in creating derived attributes should avoid the "carpet bombing" approach which involves taking all variables and dividing them by everything else, and then generating a list of ratios that may be predictive but are unexplainable. He emphasizes that all ratios should be justified and should be backed by good business reasons.

Question 12:

Relationship between depth of a tree and efficiency of a classification tree

The average number of layers from the root to the terminal nodes is referred to as the *average depth* of the tree. In general, the average depth of the tree will reflect the weight given to efficiency.

Relationship between breadth of a tree and accuracy of a classification tree

The average number of internal nodes in each level of the tree is referred to as the *average breadth* of the tree. In general, the average breadth of the tree will reflect the relative weight given to classifier accuracy

Question 13: Data Preparation steps for logistic regression

- Logistic regression involves discrete or continuous input variables and a dichotomous target variable. *The target variable must be discrete*

- There are *no assumptions regarding predictors* and therefore predictors do not have to be normally distributed, linearly related or having equal variance in each group.

- *Assess the ratio of cases to variables*, i.e. there should be enough responses for each category. If this is not ensured then it is likely that the standard errors will increase.

- *Assess linearity* in the logit, i.e., check that the regression equation has a linear relationship with the logit form of the discrete target variable (Ainsworth).

- Similar to linear regression, outliers can have a strong effect on the results of logistic regression. *Outliers should be removed or modeled separately.* The plot of residuals provides insights about the presence of outliers.

- If presence of *interaction terms* is suspected, these must be explicitly included in the model by adding them as independent variables.

- In order to ensure meaningful results, all *logit coefficients must be appropriately coded.* The convention for binomial logistic regression is to code the dependent class of greatest interest as 1 and the other class as 0, and to code its expected correlates also as +1 to assure positive correlation. For multinomial logistic regression, the class of greatest interest should be the last class. Logistic

441

regression is predicting the log odds of being in the class of greatest interest (Menard 2002).

Question 14: Data Mining Success Criteria for Classification Modeling Techniques

Table 9: Data Mining Success Criteria for Classification Modeling Techniques

		Classification Modeling Techniques			
		Classification Tree	Logistic Regression	Naïve Bayes'	Neural Network
Data Mining Success Criteria	Accuracy (Misclassification Rate)	✓	✓	✓	✓
	Lift	✓	✓	✓	✓
	Precision	✓	✓	✓	✓
	Recall	✓	✓	✓	✓
	Simplicity	✓	✓	✓	×
	Stability	✓	✓	✓	✓
	Sensitivity	✓	✓	✓	✓
	Specificity	✓	✓	✓	✓
	ROC curve	✓	✓	✓	✓
	Area Under ROC curve	✓	✓	✓	✓
	KS Statistic	✓	✓	✓	✓
	Profit/Loss	✓	✓	✓	✓

Question 15: Assessment of Business and Data Mining Success Criteria during evaluation Phase

During this phase, the results of the chosen modeling technique (output by the modeling phase) are evaluated against the business and technical success criteria. If the chosen solution only has technical merit and satisfies the data mining success criteria but does not fulfill the business objectives (assessed via the accomplishment of business success criteria) then it cannot be regarded as a feasible solution. Also, vice versa if the solution satisfies business success criteria but does not meet the technical success criteria, it cannot be regarded as an acceptable solution. A rigorous check is needed to provide evidence that the solution indeed meets both types of success criteria.

444

APPENDIX E: Results of Analysis of Variance (ANOVA) and Multivariate Analysis of Variance (MANOVA) of Survey Results

MANOVA has two main assumptions:

- *Multivariate normality:* It is assumed that the dependent variable is normally distributed within each group.

- *Homogeneity of covariance matrices:* It is assumed that variances in each group are roughly equal and also that the correlation between any two dependent variables is the same in all groups. This assumption is examined by testing whether the population variance covariance matrices are equal.

- **Checking the assumption of multivariate normality**

Since the assumption of multivariate normality cannot be tested on SPSS, the alternative approach is to check the assumption of univariate normality for each dependent variable. The results of the analysis are shown below. A significance value of less than 0.05 indicates a deviation from normality.

As we can see, the dependent variable User Satisfaction is not normally distributed in the IKDDM group (p=0.012) whereas the dependent variable perceived semantic quality is not normally distributed in the CRISP group (p =0.005).

Tests of Normality - MANOVA on survey data

		Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	MODEL	Statistic	Df	Sig.	Statistic	df	Sig.
PEOU	CRISP	.191	16	.121	.926	16	.210
	IKDDM	.158	16	.200(*)	.920	16	.170
US	CRISP	.179	16	.182	.956	16	.597
	IKDDM	.243	16	.012	.799	16	.003
PU	CRISP	.179	16	.180	.905	16	.096
	IKDDM	.191	16	.121	.889	16	.055
PSQ	CRISP	.260	16	.005	.862	16	.021
	IKDDM	.190	16	.126	.943	16	.383

* This is a lower bound of the true significance.

a Lilliefors Significance Correction

▪ **Checking the assumption of equality of covariance matrix**

This assumption can be tested using Levene's test. As a preliminary check Levene's test should not be significant for any of the dependent variables. As can be

seen from table below, Levene's test is not significant for perceived ease of use, perceived usefulness, or user satisfaction, but is significant for perceived semantic quality.

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
PEOU	.622	1	30	.437
US	.235	1	30	.631
PU	1.274	1	30	.268
PSQ	5.469	1	30	.026

Levene's test does not take into account covariances which must be checked using Box's test. Results of Box's test are also shown below. It tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups. This test should be non-significant if the variance-covariance matrices are the same. The Box's test results in p value of 0.05, which is significant, but can be regarded as just barely significant. Even a slightly higher value would have made this result non-significant. Nevertheless, given these results we will have to conclude that the assumption of homogeneity of variances is not being met by this data

447

Box's Test of Equality of Covariance Matrices

Box's M	21.238
F	1.814
df1	10
df2	4302.789
Sig.	.053

a Design: Intercept+group

Armed with information about the tenability of the assumptions of MANOVA, we now proceed to running the actual analysis. The results are presented in table below.

Results of Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.973	245.720(a)	4.000	27.000	.000
	Wilks' Lambda	.027	245.720(a)	4.000	27.000	.000
	Hotelling's Trace	36.403	245.720(a)	4.000	27.000	.000
	Roy's Largest Root	36.403	245.720(a)	4.000	27.000	.000
group	Pillai's Trace	.660	13.086(a)	4.000	27.000	.000
	Wilks' Lambda	.340	13.086(a)	4.000	27.000	.000
	Hotelling's Trace	1.939	13.086(a)	4.000	27.000	.000
	Roy's Largest Root	1.939	13.086(a)	4.000	27.000	.000

a Exact statistic

b Design: Intercept+group

Test statistics are quoted for the intercept of the model and the group variable. For our purpose, the group effects are of interest as we are interested in knowing whether or not the KDDM process model had an effect on the assessment of model quality by data mining users. SPSS provides us with four different multivariate test statistics (Pillai's trace, Wilk's Lambda, Hotelling's Trace, and Roy's largest root), all of which are highly significant ($p<0.001$). The test statistic used determines whether or not the null hypothesis that there are no differences between groups can be rejected. However, here all four statistics are significant and it can therefore be safely concluded that the type of KDDM process model has a significant effect on the performance of the groups and their assessment of model quality.

It is recommended that if MANOVA is significant, then it should be followed by an Analysis of Variance or ANOVA (Field 2000). When we run MANOVA in SPSS 15, we are also presented with the ANOVA summary table (see below) for each of the dependent variables. The row of interest is the row labeled group. The values in this row are the same as the row labeled corrected model. This is because the model fitted to the data has only one independent variable 'group'. The row labeled GROUP contains an ANOVA summary table for each of the dependent variables. The columns labeled F and Significance contain the F-ratio for each univariate AANOVA. The p values indicate that there was a significant difference between the two groups (CRISP and IKDDM) in terms of perceived ease of use (PEOU), perceived usefulness (PU), user

449

satisfaction (US), and perceived semantic quality (PSQ). Thus the ANOVA also leads

to the conclusion that the type of model had a significant impact on user's perceptions

of ease of use, usefulness, semantic quality as well as the level of user satisfaction.

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	PEOU	981.167(a)	1	981.167	52.422	.000
	US	1131.524(b)	1	1131.524	50.126	.000
	PU	624.857(c)	1	624.857	70.949	.000
	PSQ	704.381(d)	1	704.381	31.489	.000
Intercept	PEOU	13357.167	1	13357.167	713.651	.000
	US	13321.524	1	13321.524	590.132	.000
	PU	8064.857	1	8064.857	915.718	.000
	PSQ	21942.857	1	21942.857	980.947	.000
GROUP	PEOU	981.167	1	981.167	52.422	.000
	US	1131.524	1	1131.524	50.126	.000
	PU	624.857	1	624.857	70.949	.000
	PSQ	704.381	1	704.381	31.489	.000
Error	PEOU	748.667	40	18.717		
	US	902.952	40	22.574		
	PU	352.286	40	8.807		
	PSQ	894.762	40	22.369		
Total	PEOU	15087.000	42			
	US	15356.000	42			
	PU	9042.000	42			
	PSQ	23542.000	42			
Corrected Total	PEOU	1729.833	41			
	US	2034.476	41			
	PU	977.143	41			
	PSQ	1599.143	41			

a R Squared = .567 (Adjusted R Squared = .556)

b R Squared = .556 (Adjusted R Squared = .545)

c R Squared = .639 (Adjusted R Squared = .630)

d R Squared = .440 (Adjusted R Squared = .426)

APPENDIX F: Tabulated results for stepwise and forward logistic regression models (Descriptive Testing)

Tabulated results for stepwise logistic regression model

	Data	
decile	Sum of default	Sum of n
1	0	70
2	2	70
3	3	70
4	8	70
5	14	70
6	11	70
7	25	70
8	28	70
9	34	70
10	58	70
Grand Total	183	700

	bad	good	diff
1	0.0%	13.5%	13.5%
2	1.1%	26.7%	25.6%
3	2.7%	39.7%	36.9%
4	7.1%	51.6%	44.5%
5	14.8%	62.5%	47.7%
6	20.8%	73.9%	**53.1%**
7	34.4%	82.6%	48.2%
8	49.7%	90.7%	41.0%
9	68.3%	97.7%	29.4%
10	100.0%	100.0%	0.0%

	bad rate
1	0
2	0.028571
3	0.042857
4	0.114286
5	0.2
6	0.157143
7	0.357143
8	0.4
9	0.485714
10	0.828571

451

Tabulated results for forward logistic regression model

	Data	
decile	Sum of default	Sum of n
1	0	70
2	2	70
3	3	70
4	8	70
5	14	70
6	11	70
7	25	70
8	28	70
9	34	70
10	58	70
Grand Total	183	700

	bad	good	diff
1	0.0%	13.5%	13.5%
2	1.1%	26.7%	25.6%
3	2.7%	39.7%	36.9%
4	7.1%	51.6%	44.5%
5	14.8%	62.5%	47.7%
6	20.8%	73.9%	**53.1%**
7	34.4%	82.6%	48.2%
8	49.7%	90.7%	41.0%
9	68.3%	97.7%	29.4%
10	100.0%	100.0%	0.0%

	bad rate
1	0
2	0.028571
3	0.042857
4	0.114286
5	0.2
6	0.157143
7	0.357143
8	0.4
9	0.485714
10	0.828571

452

VITA

SUMANA SHARMA

34 Loch Lomond Street, Bear, Delaware 19701
Cell: 804-519-8085 Email: sharma_sumana@yahoo.com

EDUCATION

PhD in Information Systems, Minor in Decision Science (Overall GPA: 4.0 / 4.0)
VIRGINIA COMMONWEALTH UNIVERSITY (SCHOOL OF BUSINESS), Richmond, VA
- <u>Thesis:</u> "An Integrated Knowledge Discovery and Data Mining Process Model"
- <u>Selected Coursework:</u> Research Methods in Business, Decision Support and Intelligent Systems, IS Strategy, IS Security, Knowledge Management, Decision Support Systems, Human Computer Interaction, Data Mining, Databases and Information Management, Applied Multivariate Statistics, Data Warehousing, Information Reengineering, Forecasting Methods, Operations Management.

Post Graduate Diploma (Honors) in Telecom Management, 2002 (71.2%)
SYMBIOSIS INSTITUTE OF TELECOM MANAGEMENT, Pune, India

Bachelor of Engineering (Honors) in Electronics & Telecommunications Engineering, 2000 (82%)
ORIENTAL INSTITUTE OF SCIENCE & TECHNOLOGY, Bhopal, India

TEACHING EXPERIENCE

VIRGINIA COMMONWEALTH UNIVERSITY, Richmond, VA   8/2004-5/2008
Instructor - Business Information Systems (Fall 2007)
Graduate Teaching Assistant (2004-2008)
- List of Courses: Data Mining, Data Warehousing, Database Management Systems, Business Process Reengineering, Information Systems for Managers, Information Reengineering, Enterprise Resource Management, Business Information Systems.

PROFESSIONAL EXPERIENCE

DATA BLUEPRINT, Richmond, VA   5/2006-8/2006 & 5/2007-8/2007
Summer Intern

INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT
AND JOURNAL OF INFORMATION SYSTEMS SECURITY, Richmond, Virginia   8/2004-5/2005
Editorial Assistant

DISHNET DSL LTD, Chennai, India   4/2002-10/2002
Executive – Networking (Telecommunications)

REFEREED PUBLICATIONS

- Sumana Sharma and Kweku-Muata Osei-Bryson (2009). "Implementation of Business Understanding Phase of Data Mining Projects," Expert Systems with Applications, 36 (2:2), pp. 4114-4124

452

- Sumana Sharma and Kweku-Muata Osei-Bryson (2009). "Role of Human Intelligence in Domain Driven Data Mining" in Data Mining for Business Applications. Longbing Cao, Philip S. Yu, Chengqi Zhang, Huaifeng Zhang (eds), Springer *(Forthcoming)*.
- Sumana Sharma, Long Li, Manoj Thomas. "Socio-Organizational Aspects of Information Systems Security: A Review" Proceedings of the 39th Annual Meeting of DSI, Baltimore, Maryland, Nov 2009.
- Sumana Sharma and Kweku-Muata Osei-Bryson. "An Organization-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects," Proceedings of the 41st Hawaii International Conference on Computers and Systems Sciences, Hawaii. CD ROM and IEEE Digital Library (http://www.IEEE.org), Jan. 2008.
- Sumana Sharma and Kweku-Muata Osei-Bryson. "Importance of Cognitive Aspects of Managerial Decision Making: Extending and Operationalizing the Cognitive DSS Model of Chen and Lee," Proceedings of 9th IBIMA Conference on Information Management in Modern Organizations, Morocco, p. 725-735, Jan. 2008.
- Sumana Sharma. "Designing an Integrated Knowledge Discovery and Data Mining Process Model", Doctoral Consortium, Proceedings of Southern Association of Information Systems Conference, 2008

PAPERS UNDER REVIEW

- Sumana Sharma and Kweku-Muata Osei-Bryson. "Designing an Integrated Knowledge Discovery and Data Mining Process Model". Under review (round 2) at *Knowledge Engineering Review* Journal.

PAPERS UNDER PREPARATION

- Sumana Sharma and Kweku-Muata Osei-Bryson. "An Ontology-Centric Documentation Infrastructure for Data Mining Projects", Target Journal: Decision Support Systems
- Sumana Sharma "Deficiencies in existing Knowledge Discovery and Data Mining Process Models: Lessons from Software Process Engineering", Target Journal: Expert Systems with Applications

TEACHING INTERESTS

Information Systems	Database Management	Telecommunications
MIS	Data Mining	Systems Analysis & Design
Knowledge Management	Decision Support Systems	IS Strategy

AWARDS & HONORS

- Phi Kappa Phi Honor Society, 2007
- Awarded Phi Kappa Phi Graduate Scholarship (one of 10 PhD students chosen university-wide,), 2007
- National Scholars Honors Society, 2007
- Golden Key International Honor Society, 2008
- Graduate Scholarship Recipient, Virginia Commonwealth University, 2004-2008
- Placed in University Merit List of outstanding students in seven out of eight semesters of undergraduate studies, 1997-2000
- National-level Merit Scholarship, for being in the top 0.01% students in English during the All India Secondary School Examination, 1996.

SERVICE- PEER REVIEWING

- *Information Systems Frontiers,* Special Issues on Decision Models for Information Systems Management, Forthcoming 2008
- Southern Association for Information Systems (SAIS) Conference, 2008

- *Journal of Strategic Information Systems,* Special Issue on Privacy and Security, 2007
- Hawaii International Conference on Systems Science, Mini-track on Human Computer Interaction, 2007
- *International Journal of Information Management,* 2005
- *Journal of Information Systems Security,* 2005

CERTIFICATIONS & AFFILIATIONS

- Decision Science Institute, 5/2007-Present
- Association of Information Systems, 3/2006-Present
- CCNA (Cisco Certified Network Associate), 2001 (Scored 96.7%)

EXTRACURRICULAR LEADERSHIP

- Executive Council, Graduate Students Association, Virginia Commonwealth University, 8/2005-12/2007
- Graduate Student Representative, VCU Library Advisory Committee, 8/2006-8/2007
- Academic Committee, Symbiosis Institute of Telecom Management, 2001-2002
- Class Mentor, Oriental Institute of Science and Technology, 7/1998-6/2000
- Cultural Music Committee, Oriental Institute of Science and Technology, 7/1997-6/1998