4-10-2008

Detecting Financial Statement Fraud: Three Essays on Fraud Predictors, Multi-Classifier Combination and Fraud Detection Using Data Mining

Johan L. Perols
University of South Florida

Detecting Financial Statement Fraud: Three Essays on Fraud Predictors, Multi-Classifier Combination and Fraud Detection Using Data Mining

by

Johan L. Perols

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Information Systems and Decision Sciences
Department of Accountancy
College of Business
University of South Florida

Co-Major Professor: Kaushal Chari, Ph.D.
Co-Major Professor: Jacqueline L. Reck, Ph.D.
Uday S. Murthy, Ph.D.
Manish Agrawal, Ph.D.

Date of Approval:
April 10, 2008

Keywords: Earnings Management, Discretionary Accruals, Unexpected Productivity, Information Markets, Combiner Methods, Machine Learning

Dedication

To Becca who provided support (in many ways), encouragement and motivation, helped me with my ideas, and believed in me more than I sometimes did; and to family and friends for providing the motivation for completing this dissertation.

Acknowledgments

To the faculty, administrators and fellow Ph.D. students in the information systems and decision sciences department and the accounting department, thank you for all the support and for creating an excellent learning environment. I especially want to acknowledge Dr. Chari, Dr. Reck, Dr. Murthy, Dr. Agrawal, Dr. Bhattacherjee and Ann Dzuranin. To my dissertation committee, thank you for reviewing, improving, supporting, criticizing and editing my dissertation. I am grateful to Dr. Chari for mentoring me, for emphasizing the importance of producing quality research, for guiding me in design science research, for working with me on the first essay and for prioritizing my development as a researcher over obtaining research assistance. I am thankful to Dr. Reck for introducing me to and guiding me in the archival research method, for helping me shape and organize my dissertation, for working with me on the second essay, and for providing advice on how to combine the accounting and information systems concentrations. I am grateful to Dr. Murthy for introducing me to AIS research, for encouraging and supporting the addition of the accounting concentration, and for working with me on various projects. I thank Dr. Agrawal for helping me with my first conference paper, for working with me on the first essay, for coaching and helping me prepare for conference presentations, and for taking time to discuss my research. Outside of my dissertation committee I also want to acknowledge Dr. Bhattacherjee for teaching me, both in and outside of class, about research and the importance of theory, and Ann Dzuranin for encouraging and pushing me to add the accounting concentration, for the many great discussions and for being a great friend.

Table of Contents

List of Tables

List of Figures

Detecting Financial Statement Fraud: Three Essays on Fraud Predictors, Multi-Classifier Combination and Fraud Detection Using Data mining

Johan L. Perols

ABSTRACT

The goal of this dissertation is to improve financial statement fraud detection using a cross-functional research approach. The efficacy of financial statement fraud detection depends on the classification algorithms and the fraud predictors used and how they are combined. Essay I introduces IMF, a novel combiner method classification algorithm. The results show that IMF performs well relative to existing combiner methods over a wide range of domains. This research contributes to combiner method research and, thereby, to the broader research stream of ensemble-based classification and to classification algorithm research in general. Essay II develops three novel fraud predictors: total discretionary accruals, meeting or beating analyst forecasts and unexpected employee productivity. The results show that the three variables are significant predictors of fraud. Hence Essay II provides insights into (1) conditions under which fraud is more likely to occur (total discretionary accruals is high), (2) incentives for fraud (firms desire to meet or beat analyst forecasts), and (3) how fraud is committed and can be detected (revenue fraud detection using unexpected employee productivity). This essay contributes to confirmatory fraud predictor research, which is a sub-stream of research that focuses on developing and testing financial statement fraud predictors. Essay III compares the utility of artifacts developed in the broader research streams to which the first two essays contribute, i.e., classification algorithm and fraud predictor research in detecting financial statement fraud. The results show that logistic regression and SVM perform well, and that out of 41 variables found to be good predictors in prior fraud research, only six variables are selected by three or more classifiers: auditor turnover, Big 4 auditor, accounts receivable and the three variables introduced in Essay II. Together, the results from Essay I and Essay III show that IMF performs better than existing combiner methods in a wide range of domains and better than stacking, an ensemble-based classification algorithm, in fraud detection. The results from Essay II and Essay III show

that the three predictors created in Essay II are significant predictors of fraud and, when evaluated together with 38 other predictors, provide utility to classification algorithms.

Chapter 1. Dissertation Overview

The Association of Certified Fraud Examiners (ACFE 2006) estimates that occupational fraud totals $652 billion per year in the U.S. or about 5% of total revenues. A national survey conducted by KPMG in 2003 reported that 75% of organizations had experienced fraud in the three months leading up to the study (KPMG 2003). The potential benefit of fraud reduction is staggering; a 33% reduction in fraud would result in a 26% increase in average profits of American organizations. Detecting fraud is, however, difficult and according to ACFE (2006) 25% of discovered fraud is detected by accident as compared to proactive measures such as internal audits (20%), internal controls (19%) or external audits (12%).

1.1. Research Framework

The overarching goal of this dissertation is to improve financial statement fraud detection. Figure 1.1 shows how the three essays are related and how they contribute to the goal of improving financial statement fraud prediction. The efficacy of the detection depends on the classification algorithms and the fraud predictors used and how they are combined. Essay III, Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms, evaluates the utility of different classification algorithms and fraud predictors for predicting financial statement fraud. This research pushes the research frontier in data mining fraud detection research in the functional area of accounting information systems. Essay I, Information Market Based Decision Fusion, introduces a new classification algorithm and contributes to classification algorithm research in the functional area of information systems. Three new fraud predictors are developed in Essay II, The Effect of Discretionary Accruals, Earnings Expectations and Unexpected Productivity on Financial Statement Fraud. This research adds to fraud predictor research in the functional area of accounting. Thus, I use a cross-functional research approach focusing on two functional areas, accounting and information systems, and their confluence, accounting information systems, to improve financial statement fraud detection.

In the classification algorithm essay (Essay I) and the fraud predictor essay (Essay II) I approach specific research sub-streams with the intention of moving each sub-stream forward.

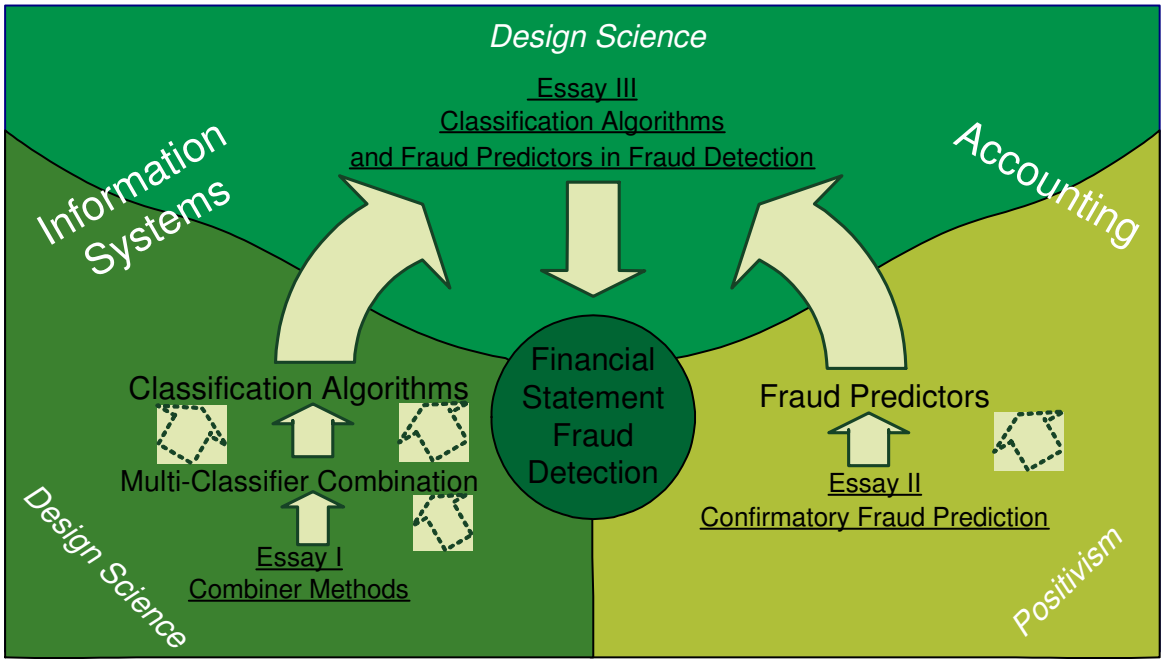


Figure 1.1: Research Framework

Essay I focuses on multi-classifier combination combiner method research, which is a research sub-stream within multi-classifier combination research that, in turn, is a sub-stream of the broader research stream of classification algorithms. Essay II extends confirmatory fraud predictor research within the broader research stream of financial statement fraud predictors. The third essay represents the nexus of the first two essays, and brings the two broader research streams together with the sole objective of improving fraud detection. This essay examines the utility of different combinations of classification algorithms and fraud predictors. Thus, the first two essays can be seen as contributing to the basic research needed for the third essay to accomplish its objective of improving financial statement fraud detection. Note, however, that the goal of the third essay is not to test the utility of the algorithm developed in the first essay or to test the efficacy of the predictors developed in the second essay in detecting financial statement fraud. Rather, the third essay takes findings from the broader research streams to which the first two essays contribute, i.e., classification algorithms and fraud predictors, and examines the efficacy of various artifacts developed within these research streams in detecting financial statement fraud. Of course, the algorithm developed in the first essay and the predictors developed in the second essay are part of these research streams, and are, thus, included in the examination in the third essay. This might seem like a subtle difference but it is important. By not focusing on the artifacts developed in the first two essays, the third essay is not tied to these artifacts. This allows me to choose among a larger number of artifacts and does not limit my

evaluation to the efficacy of the artifacts developed in the first two essays. Furthermore, this allows Essay I to develop an artifact that contributes to combiner method research without being tied to a specific domain. Similarly, by focusing on the confirmatory fraud research stream, Essay II can make a more theoretical contribution by developing artifacts that contribute to our understanding of conditions and incentives related to financial statement fraud, as opposed to focusing on developing artifacts that outperform existing predictors. It is also important to note that the contributions of the three essays are primarily dependent on the results reported in the respective essays. For example, the contribution of the multi-classifier combination combiner method developed in Essay I is determined based on an evaluation in Essay I of the performance of the proposed combiner method to combiner methods developed in prior research. This evaluation does not focus specifically on the fraud domain and instead evaluates the contribution of the combiner method across multiple domains.

In addition to my dissertation being cross-functional, I also use two paradigms to accomplish my research objective of improving financial statement fraud detection. Essay II is confirmatory hypotheses testing grounded in positivism, while Essays I and III follow the design science paradigm. I design and evaluate a novel IT artifact in Essay I, and in Essay III, I evaluate the utility of multiple classification algorithm and fraud predictor artifacts in the financial statement fraud domain.

1.2. Overview of the Three Essays

Essay I is titled: Information Market Based Decision Fusion. In this essay, I design a novel combiner method based on theoretical and empirical findings in information market research to improve the performance over existing combiner methods. Combiner methods are used in multi-classifier combination to improve the classification performance of individual classifiers by combining the decisions of many individual classifiers, like artificial neural networks (ANN), logistic regression and decision trees (Kittler and Roli 2000). I show through extensive experiments that when the true classes of objects are only revealed for objects classified as positive, IMF outperforms three benchmark combiner methods, Majority, Average and Weighted Average when the positive ratio is low, and outperforms Majority and performs on par with Average and Weighted Average, when the positive ratio is high. When the true classes of all objects are revealed, IMF outperforms Weighted Average and Majority, and at marginal level of significance, outperforms Average.

Essay II is titled: The Effect of Discretionary Accruals, Earnings Expectations and Unexpected Productivity on Financial Statement Fraud. The research objective in this essay is to

improve our understanding of conditions and incentives behind financial statement fraud. I hypothesize that (1) earnings management in prior years is positively related to financial statement fraud; (2) firms that meet or exceed analyst forecasts are more likely to have committed fraud than firms that fail to meet analyst forecasts; and (3) unexpected productivity is positively related to financial statement fraud. I use an archival research approach to compare a set of fraud firms, hand-collected from SEC enforcement actions, to a set of matched non-fraud firms. The empirical results show support for all three hypotheses.

Essay III is titled: Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. The research objective in this essay is to compare the utility of a fairly comprehensive set of classification algorithms and fraud predictors in financial statement fraud prediction. With this objective in mind I pose two specific research questions. (1) What classification algorithm provides the most utility given different assumptions about prior probabilities and costs of false positive and false negative classification errors? (2) What predictors are useful to these algorithms for detecting financial statement fraud? I find that logistic regression and support vector machines (SVM) perform well relative to C4.5 (a decision tree), MultilayerPerceptron (a backpropagation neural network), stacking (an ensemble method), bagging (also an ensemble method) and IMF (an ensemble method combiner method), while stacking and C4.5 consistently perform relatively poorly, where performance is measured using estimated relative cost. Furthermore, logistic regression and SVM provide the best performance under what is believed to be the most relevant prior probability and relative cost estimates. The results also show that out of 41 variables that have been found to be good predictors in prior fraud research, only six variables are selected by three or more classifiers: auditor turnover, total discretionary accruals, Big 4 auditor, accounts receivable, meeting or beating analyst forecasts, and unexpected employee productivity.

The reminder of the dissertation is organized as follows. Chapters 2, 3 and 4 contain the three essays, Essay I, Essay II and Essay III, respectively. Each essay is written as a stand alone paper, thus the three essays can be read in any order. Chapter 5 recaps the primary findings in the three essays and concludes with a discussion of how these results fit together.

Chapter 2. Information Market Based Decision Fusion

2.1. Introduction

In many decision-making scenarios, decisions of multiple human experts or classifiers are fused to determine the overall decision. Examples include: a group of accounting experts and classifiers making going-concern decisions and an ensemble of classifiers in a fraud detection application making decisions on whether a transaction is fraudulent. Multi-classifier combination (MCC) is a technique that can be used to improve the classification performance in various classification problems by combining the decisions of multiple individual classifiers (Suen and Lam 2000). In MCC, individual classifiers, commonly referred to as base-classifiers, classify objects based on inputs consisting of object feature vectors (see Figure 2.1). These classifications or decisions are then combined using a combiner method into a single decision about the object's class label.

The basic premise behind MCC is that different classifiers in an ensemble have different strengths and weaknesses, and therefore provide complementary information (referred to as diversity in MCC) about the classification problem. These differences can be leveraged to improve classification performance by combining base-classifiers' decisions (Kittler et al. 1998). Different combiner methods have been proposed and examined in the literature, and can be

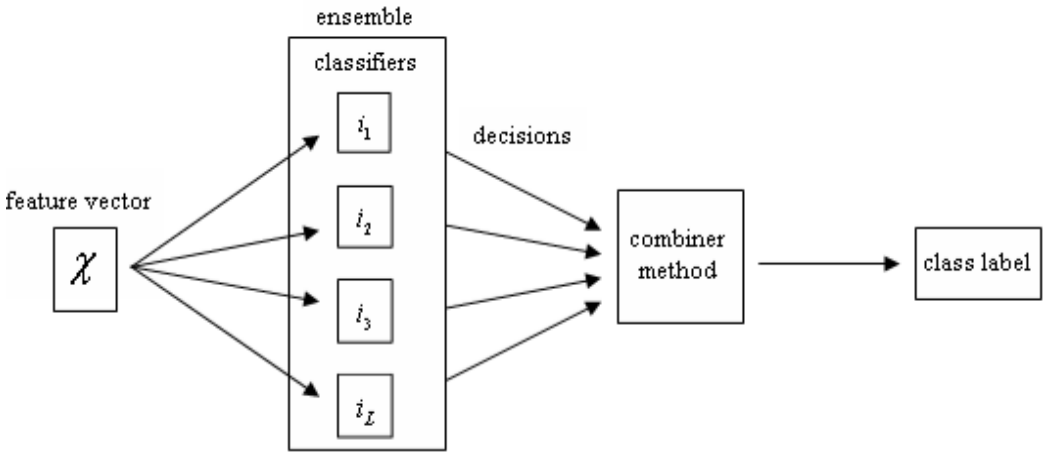


Figure 2.1: Generic Classifier Combiner Architecture

categorized based on whether they require training data. For example, Naive Bayes, Decision Templates and Weighted Average (WAVG) require training data, while Average (AVG), Majority (MAJ) and Product do not require training data. Existing combiner methods that require training data have limitations including the requirement for training data, and restrictive assumptions such as: 1) constant ensemble base-classifier composition; and 2) training data performance being a good proxy for subsequent actual performance. Experimental results generally indicate that MCC provides performance benefits, and that the performance of MAJ and AVG methods are comparable or superior to that of methods requiring training (Duin and Tax 2000).

To improve performance while overcoming these limitations, I propose an information market based fusion approach for multi-classifier combination that 1) has superior performance, 2) does not require training data, and 3) can adapt to changes in ensemble composition and base-classifier performance. In evaluating the effectiveness of the proposed approach, I compare IMF against three combiner methods, AVG, MAJ and WAVG. These methods have performed relatively well in prior research (Duin and Tax 2000) and have been used as benchmarks in recent MCC research[1]. For example, Zheng and Padmanabhan (2007) use AVG, which they refer to as Unweighted Average, and a version of WAVG where the weights are variance based, which they refer to as Variance Based Weighting. The experimental evaluation was performed using computational experiments with 17 datasets that were obtained from the UCI Machine Learning Repository (Newman et al. 1998) and 22 different base-classifiers from Weka (Witten and Frank 2005).

The rest of the chapter is organized as follows. In Section 2.2, I provide a review of related research. IMF is introduced in Section 2.3 along with an overview of information markets. I then present details on the computational experiments and results in Sections 2.4 and 2.5 respectively. In Section 2.6, I discuss these results and conclude in Section 2.7 with a review of my contributions and suggestions for future research.

---

[1] In Duin and Tax (2000), AVG is referred to as Mean and MAJ is referred to as Majority; eight additional combiner methods are evaluated: Bayes Rule (two different implementations), Nearest Mean, Nearest Neighbor, Maximum, Median, Minimum and Product. When combining the decisions of different base-classifiers trained using the same feature set, which is comparable to the MCC architecture that I use, their results (p. 23) show that Majority and Mean perform on par with or better than the other combiner methods.

2.2. Related Research

A classifier is a model that makes decisions about an object's class membership based on the object's feature set. Examples of classifiers include neural networks, logistic regression, decision trees and Bayesian classifiers (Witten and Frank 2005). Classifier performance is typically dependent on the problem domain as well as on the calibration of the classifier. Multiple classifiers are therefore typically tested in order to identify the best classifier for a given problem domain. However, it is generally difficult to determine which classifier(s) will perform well in subsequent classifications. Furthermore, classification for certain cases may even be improved by an "inferior" classifier (Kittler et al. 1998). Thus, by combining the decisions of diverse classifiers, it is possible to improve the overall performance.

Prior MCC research has primarily focused on one of two areas: (1) training and selection of ensemble base-classifiers; or (2) combination of base-classifier decisions. Methods such as bagging, boosting and stacking fall into the first category (Witten and Frank 2005), while combiner methods such as MAJ, AVG and WAVG fall into the second category. Recent research within the former stream has used ROC analysis to select dominant classifiers (Provost and Fawcett 2001), and Data Envelopment Analysis to select efficient classifiers (Zheng and Padmanabhan 2007) under various cost and class distributions, and then combine these classifiers' decisions. Zhao and Ram (2004) have investigated the appropriate cascading depth in cascade generalization, a variation of stacking, where classifiers are trained sequentially using the original input data and lower level classifiers' decisions. This essay does not focus on classifier selection and training; but on the combiner methods. Prior research within the combiner method research stream has found that methods that use measurement data are typically more accurate than methods that handle unique labels; methods that require training data typically outperform methods that do not require training (Jain et al. 2000), but that MAJ and AVG, which do not require training, perform either at the same level or significantly better than more complex methods Duin and Tax (2000).

Another important, but largely overlooked aspect of combiner methods is how well they fit with different system architectures. Software agents offer a new paradigm to support decision-making (Nissen and Sengupta 2006) where human-driven or autonomous software agents embodying classifiers and other intelligent algorithms can leverage their individual strengths to make collective decisions. The base-classifiers, combiner method and providers of object features in a MCC can be implemented as software agents in multi-agent systems. Research in data mining has implemented MCC agent systems for credit card fraud detection (Stolfo et al. 1997) and network intrusion detection (Lee et al. 2000).

In MCC multi-agent systems that are implemented in dynamic real world settings, the relative performance of base-classifiers and the ensemble composition can change over time as agents are retired, added or temporarily unavailable. Existing combiner methods that require training do not take this into consideration, and assume that the ensemble composition is static, and that individual classifier performance does not change subsequent to training and validation. I next introduce IMF, a combiner method that takes these issues into consideration.

2.3. Information Market Based Fusion

IMF is theoretically grounded in information markets. More specifically, the IMF aggregation mechanism used in this essay is based on pari-mutuel betting markets.

2.3.1 Information Markets

Information markets are markets specifically designed for the purpose of information aggregation. Equilibrium prices, derived using conventional market mechanisms, provide information based on private and public information maintained by the market participants about a specific situation, future event or object of interest (Hanson 2003). Although the concept of information markets is fairly recent, the underlying notion of markets being capable of aggregating information is not new (Hayek 1945), and the efficient market hypothesis states that all private and public information is reflected in equilibrium prices (Fama 1970). Empirical research has found support for the efficient market hypothesis, and for information aggregation in information markets in general (Berg and Rietz 2003), and pari-mutuel betting markets in particular (Plott et al. 2003).

The combiner method presented in this essay is based on pari-mutuel betting, which originated in horserace gambling in France in 1865, and since then has become a popular betting mechanism in the horseracing world. Pari-mutuel means “wager mutual” and comes from the fact that in pari-mutuel betting, a winning wager (i.e., bet) receives a share of the total wagers (winning and loosing bets less track commission) as a proportion of this winning wager to all winning wagers. The final track odd for a given horse is the total amount bet on all the horses in the race divided by the total amount bet on the given horse. The payout for a winning horse is the product of the amount bet on it and its odd (less track commission). From a MCC perspective, the odd associated with a horse is of great importance as it represents the aggregated market information about the probability estimate of that horse winning the race. I use pari-mutuel betting over mechanisms such as continuous double auctions since pari-mutuel betting does not suffer from liquidity problems that could potentially impact continuous double auction markets when there

are large bid-ask spreads or when bid-ask queues are empty (Pennock 2004). Hence pari-mutuel mechanisms would work effectively, even when the ensemble of base classifiers is small.

Plott et al. (2003) experimentally examined information aggregation and different betting behaviors in pari-mutuel betting markets using two private information models, Decision Theory Private Information (DTPI) and Competitive Equilibrium Private Information (CEPI), and one model with belief updating- Competitive Equilibrium Rational Expectations (CERI). Plott et al. (2003) found that DTPI and CEPI best described the behavior of human participants in their Probabilistic Information Condition experimental pari-mutuel betting market.

In DTPI, agents only consider their own private information and ignore market prices when deciding on their bets and in forming beliefs. In CEPI, agents base their bets on the current market price, although they do not update their beliefs based on market prices. In both models, agents maximize their conditional expected utility given their private probability estimates and constraints such as available funds. In both DTPI and CEPI, prices are assumed to be in equilibrium; however, as each betting round starts without prices defined, the equilibrium must be obtained before the agents can place their final bets. Assuming no track take, in equilibrium, all potential payouts are equal to the total amount bet across all events.

2.3.2 Information Market Based Fusion

IMF is a multi-classifier combiner method based on a pari-mutuel betting information market that can be used in any classification application domain. I present IMF in the context of a fraud detection application. In this application, object t (i.e., transaction t) can be classified as fraudulent (j=1) or non-fraudulent (j=2) by an ensemble E of agent classifiers. In this application, the set J={1,2} is the index set of the two classes (i.e., fraudulent and non-fraudulent). The ensemble E has m agents embodying different base-classifiers (referred to as agents) represented by indices i in the index set $D = \{1,..,m\}$. While determining the class membership of object t, agent $i \in D$ uses the feature vector associated with t to determine the posterior probability estimate $p_{itj} \in [0, 1]$ that t belongs to class $j \in J$. Agent i bets q_{itj} that object t belongs to class j and is paid according to the pari-mutuel mechanism based on four factors: (1) the agent's bets, q_{itj}; (2) the total bets on class j, $Q_{tj} = \sum_{i \in D} q_{itj}$; (3) the total bets on all classes, $Q_t = \sum_{j \in J} \sum_{i \in D} q_{itj}$; and (4) the true class of object t. Ensemble E's overall probability estimate that t belongs to $j \in J$ is given by $1/O_{tj} \in [0,1]$, where O_{tj} is the odd that t belongs to $j \in J$. The odd O_{tj}, which is equal to Q_t/Q_{tj}, is in equilibrium when the potential payouts $Q_{tj}O_{tj}$ for each $j \in J$ and the total bets Q_t are equal (assuming no house commission), i.e., O_{tj} is in equilibrium when $Q_{tj}O_{tj} = Q_t$.

Figure 2.2 provides an overview of IMF when the true class of objects is only determined for objects classified as positive. When all objects are investigated, Figure 2.2 is changed by eliminating the decision box, i.e., going straight from *Classify Object* to *Distribute Payout*. Investigations are however expensive, and in the real world only objects classified as positive are typically investigated. Unless otherwise noted, I will henceforth assume that only objects classified as positive are investigated.

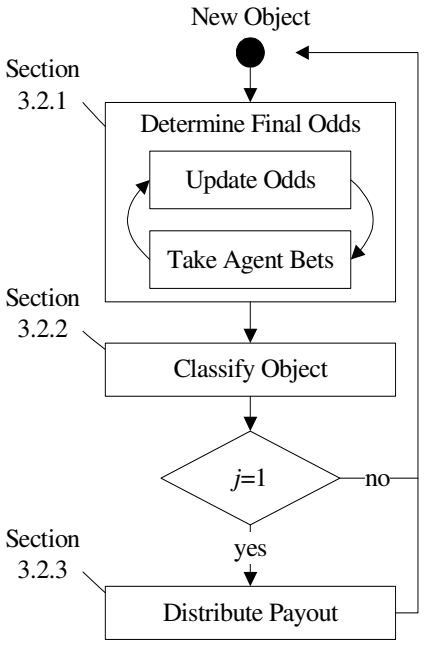


Figure 2.2: IMF Flowchart

In Figure 2.2, for each new object t, IMF first determines the final odds O^f_{tj} that are equilibrium or near-equilibrium odds. Establishing equilibrium odds is a nontrivial task because of the recursive relationship between Q_{tj} and O_{tj}, where odds are based on agent bets and agents base their bets on odds. Therefore, multiple rounds of betting are required to determine the final odds that can then be used by agents to make their actual bets. In each round, odds are first updated based on all the agents' prior bets and then agents place new bets based on the current updated odds. After the final odds have been established, ensemble *E's* overall probability estimate $1/O^f_{tj}$ is compared to a threshold value C_j to determine if t should be classified as belonging to class j. If object t is classified as fraudulent (j=1) then the true class of t is determined and winnings are distributed to the agents.

In addition to establishing ensemble *E's* probability estimate $1/O^f_{tj}$, IMF facilitates the redistribution of wealth among the agents based on the agents' bets and winnings. From an MCC perspective, IMF produces decisions that are wealth-weighted probability estimates of the occurrence of event j. I next describe the components of IMF in detail as per the major steps depicted in Figure 2.2.

2.3.2.1 Determining Final Odds

The problem to determine odds O_{tj} for object t is given by P1.

P1: $$Z_1 = \min_{O_{tj} M_j} \sum_{j \in J} M_j \qquad (1)$$

s.t. $$Q_{tj} O_{tj} - M_j = Q_t \qquad \forall j \in J \qquad (2)$$

$$M_j \geq 0 \text{ and } O_{tj} \geq 1 \qquad \forall j \in J \qquad (3)$$

The objective function Z_1 minimizes dummy variables M_j that represent the differences between the total bets by all agents and the total payout for each outcome (2). At equilibrium, M_1 and M_2 are equal to zero.

Due to the recursive relationship between Q_{tj} and O_{tj}, I solve P1 using binary search (see Figure 2.3) to determine equilibrium or near-equilibrium odds. Binary search starts with a lower bound P^l=0 and an upper bound P^u=1 for the probability that object t belongs to class j=1. O_{t1} is then computed using: O_{t1}=2/(P^l+P^u). It can be easily verified that $O_{t2} = O_{t1}/(O_{t1}-1)$ in the case of two class problems. The agents then place bets that maximize their individual utility, given their current wealth and probability estimates, and the current odds (Lemma 1 and Lemma 2 describe the optimal bets).

Set search space bounds
$P^l = 0$ and $P^u = 1$
set $O_{t1} = 2/(P^l + P^u)$
Take agent bets
Do
  If $Q_{t1}O_{t1} > Q_t$ then
    set $P^l = 1/O_{t1}$
  else if $Q_{t1}O_{t1} < Q_t$ then
    set $P^u = 1/O_{t1}$
  else if $Q_{t1}O_{t1} = Q_t$ then
    set P^l and P^u to $1/O_{t1}$
  set $O_{t1} = 2/(P^l + P^u)$
  Take agent bets
Until ($P^u - P^l \leq \varepsilon$)

Figure 2.3: Binary Search

The odds and bets are then used to evaluate whether the current odds are too high or too low. If the potential payout for j=1, i.e., $Q_{t1}O_{t1}$, is greater than the total bets Q_t, then odd O_{t1} is too high, and the lower search space boundary P^l is raised to the reciprocal of O_{t1}, i.e., P^l=$1/O_{t1}$. On the other hand, if the potential payout for j=1 is less than the total bets Q_t, then the odd O_{t1} is too low and the upper search space boundary P^u is lowered to the reciprocal of O_{t1}, i.e., P^u=$1/O_{t1}$. If the potential payouts for j=1 is the same as the total bets Q_t, then the potential payouts for j=1 and j=2 are equal, i.e., $Q_{t1}O_{t1}$=$Q_{t2}O_{t2}$, the odds are in equilibrium and the search space is set to this single value P^l=P^u=$1/O_{t1}$. O_{t1} is then set to the reciprocal of the mean of P^l and P^u and the agents place bets based on these odds. The updating of odds and agent bets continues iteratively until the search space is within tolerance ε, i.e., P^u-$P^l \leq \varepsilon$. When binary search terminates, it is known that the optimal odds are within bounds $1/P^u$ and $1/P^l$.

Determining Agent Bets - Given the current market odds O_{tj}, the agent's probability estimates p_{itj} of object t being in class j, the agent's current wealth w_{it} plus the periodic endowment m, and multiplier k that determines the house enforced maximum bet km, agent i solves the expected utility maximization problem P2 to determine the amount q_{itj} to bet on classes j=1,2. The periodic endowment m is given to all the agents in order to prevent them from running out of funds. Given the utility function U_i of agent i as a function of wealth, problem P2 can be stated as follows:

$$\text{P2:} \quad Z_2 = \max_{q_{itj}} \; p_{it1}U_i(w_{it}+m - q_{it1} - q_{it2}+q_{it1}O_{t1}) + p_{it2}U_i(w_{it}+m - q_{it1} - q_{it2}+q_{it2}O_{t2}) \qquad (4)$$

S.T. $$q_{it1} + q_{it2} = \begin{cases} w_{it} + m & \text{if } (w_{it} + m) \le km \\ km & \text{if } (w_{it} + m) > km \end{cases} \quad (5)$$

$$q_{itj} \ge 0 \quad (6)$$

The objective function in P2 represents the expected utility of agent i when it bets $q_{itj} \ge 0$ on event j. Constraint (5) dictates that the total amount of bets placed by agent i on events j=1 and j=2 equals the lower of the agents' available funds $m + w_{it}$ and km, the house enforced maximum bet. km limits the amount of influence the best performing agents in the ensemble could exert on ensemble decision, due to the need to have all agents, not just the best performing agents, contribute to improving the success of the ensemble (Kittler et al. 1998).

P2 is general enough to incorporate any utility function to model an agent's risk aversion. I utilize a natural logarithm (ln) utility function (hence forth simply referred to as log utility), which has been widely used in prior research (Rubinstein 1976), for the following reasons: (1) log utility enables agents to place bets that yield optimal long run growth rates (Kelly 1956); (2) it is twice-differentiable and non-decreasing concave, leading to a decreasing absolute risk aversion (Rubinstein 1976); and (3) depending on which betting constraint is binding (see Lemma 1 and 2 below), log utility bets are either increasing in p_{itj} and $w_{it} + m$ but not a function of O_{tj}, a betting behavior corresponding to DTPI (Plott et al. 2003), or increasing in p_{itj}, $w_{it} + m$ and O_{tj}, a betting behavior corresponding to CEPI (Plott et al. 2003).

Using log utility, problem P2 is transformed to either P3 or P4 depending on the binding constraint in (5) for a given agent. If $w_{it} + m \le km$ then $q_{it1} + q_{it2} = w_{it} + m$ and $w_{it} + m - q_{it1} - q_{it2} = 0$, leading to P3.

P3: $$Z_3 = \max_{qitj} \; p_{it1} ln(q_{it1} O_{t1}) + p_{it2} ln(q_{it2} O_{t2}) \quad (7)$$

S.T. $$q_{it1} + q_{it2} = w_{it} + m \quad (8)$$

$$q_{itj} \ge 0 \quad (9)$$

***Lemma 1**:* *The optimal bets of agent i in P3 while classifying t is:*

$$q^*_{itj} = p_{itj}(w_{it}+m) \; \forall j \in J.$$

Proof: See the Appendix 1.

If $w_{it} + m > km$ then $q_{it1} + q_{it2} = km$ and $w_{it} + m - q_{it1} - q_{it2} = w_{it} + m - km$, which I denote by constant a_{it}. Thus P2 can be transformed to P4.

P4: $$Z_4 = \max_{qitj} p_{it1} ln(a_{it} + q_{it1} O_{t1}) + p_{it2} ln(a_{it} + q_{it2} O_{t2}) \quad (10)$$

S.T. $$q_{it1} + q_{it2} = km \quad (11)$$

$$q_{itj} \ge 0 \quad (12)$$

Lemma 2: *The optimal bets of agent i in P4 while classifying t is:*

Solution a: $q^*_{it1} = p_{it1}\, km + a_{it}\dfrac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}}$ *and*

$q^*_{it2} = p_{it2}\, km + a_{it}\dfrac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}}$ *, when*

$0 < p_{it1}\, km + a_{it}\dfrac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}} < km$, *and*

$0 < p_{it2}\, km + a_{it}\dfrac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}} < km;$

Solution b: $q^*_{it1} = km$, *and* $q^*_{it2} = 0$, *when* $km \le p_{it1}\, km + a_{it}\dfrac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}}$ *; and*

Solution c: $q^*_{it2} = km$, *and* $q^*_{it1} = 0$, *when* $km \le p_{it2}\, km + a_{it}\dfrac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}}$

Proof: See Appendix 2.

Equilibrium Odds - The final odds O^f_{tj}, are equilibrium odds or near-equilibrium odds, where near-equilibrium is defined as being within bounds $1/P^u$ and $1/P^l$, and P^u-P^l is less than or equal to tolerance ε. When binary search terminates, it is known that the optimal odds are within these bounds. As described in binary search, the final odds O^f_{tj} are found for each object t by iteratively updating the odds, and requiring agents to place bets using these odds until the odds provided to the agents and their subsequent bets result in $Q_{tj}O^f_{tj} \approx Q_t$, at which time the market closes. The following observations are made. First, bets placed in betting rounds before the final odds have been established are only used for the purpose of updating the odds[2]. Second, if agent bets are discontinuous over O_{tj} then the existence of equilibrium odds cannot be guaranteed (Carlsson et al. 2001). Lemma 3 show that in IMF, when agents $i \in D_1$ bet as per Lemma 1, and agents

[2] I make the assumption that agents do not act strategically by attempting to bluff about their private information, i.e., placing bets that do not maximize their utility given the current odds. This assumption is made to make the utility maximization problem more tractable. In defense of this assumption, the agents do not know when the market closes, i.e., they never know if the current odds are the final odds, and strategic behavior is therefore less likely even if allowed. Furthermore, Plott et al. (2003) found that strategic behavior was negligible among their human subjects in the Probabilistic Information Condition experiment even though the subjects knew that the market would stay open at least until an announced time.

$i \in D_{2a}$, $i \in D_{2b}$ and $i \in D_{2c}$ bet as per Lemma 2, solutions a, b and c, respectively, then equilibrium exists. However, even when equilibrium odds do exist, IMF may not always find it due to the recursive nature of O_{tj} and Q_{tj}. Binary search used in IMF, nevertheless, guarantees a result that is at most ε (a tolerance parameter) from the optimal probability.

***Lemma 3**:* *Given any combination of betting behaviors as per Lemma 1 and Lemma 2, equilibrium exists, and the equilibrium odd for j=1 is:*

$$O_{t1} = \frac{\sum_{i \in D1 \cup D2a} p_{it2}(w_{it} + m) + \sum_{i \in D2c}(km)}{\sum_{i \in D1 \cup D2a} p_{it1}(w_{it} + m) + \sum_{i \in D2b}(km)} + 1$$

Proof: See Appendix 3. See Appendix 4 for an empirical evaluation of IMF when agent bets are discontinuous over O_{tj} and an empirical evaluation of the ability of IMF to find the equilibrium odds when the agents bet as per Lemma 1 and Lemma 2.

2.3.2.2 Classifying Objects

Once the final odds O^f_{t1} and O^f_{t2} are available, the decision rule in (13) can be used to classify[3] object *t*.

If $(1/O^f_{t1} \geq C_1)$ then classify class of *t* as *j*=1; else classify class of *t* as *j*=2 (13)

In (13), if the reciprocal of final odd for *j*=1 is higher than the threshold C_1 then object *t* is classified as a member of the positive class, i.e., *j*=1 and agent *i*'s wealth is decreased by the amount of *i*'s final bets:

$$w_{it} = w_{it} - \sum_{j \in J} q_{itj} \quad (14)$$

The true class of *t* is then investigated, and agent *i*'s wealth is updated with any potential winnings (see Section 2.3.2.3 below). If the object is classified as a member of the negative class i.e., *j*=2, then the verification of the object class is not pursued further as investigations are not typically carried out for negative classifications. In this case, agent wealth is not updated with bets or winnings.

2.3.2.3 Distributing Payout

Whenever object *t* is classified as belonging to the positive class, detailed investigations are necessary to establish the true class of *t*. While final bets are deducted from the agents' wealth

[3] MCC users might prefer rankings or raw probabilities (Saar-Tsechansky and Provost 2004). In these situations the generated ensemble probability estimates can be presented directly to the users.

immediately, due to the time taken for investigations, there is a time lag corresponding to v elapsed object classifications before winnings can be paid out. This mechanism is similar to sports betting (and other types of futures markets) where bets are collected when bets are placed and winnings are paid out after the game/race has been decided. In Figure 2.4, t is the current object being classified, t' is the object for which the investigation has just been completed, and v is the number of objects that have been classified since t' was classified. Based on the investigation, if t' is found to be a positive, then agent i's wealth is updated using (15), else t' is negative and agent i's wealth is updated using (16). This is also followed when the true classes of both positive and negative classifications are investigated.

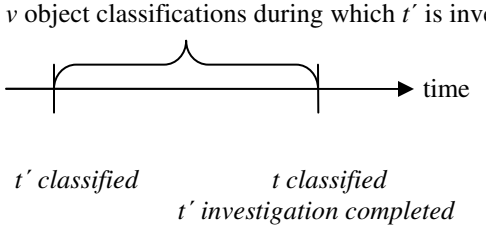


Figure 2.4: Payout Distribution Time Lag

$$w_{it} = w_{it} + (q_{it'1} / Q_{t'1})Q_{t'} \tag{15}$$

$$w_{it} = w_{it} + (q_{it'2} / Q_{t'2})Q_{t'} \tag{16}$$

2.4. Experimental Setup

2.4.1 Base-Classifiers and Data

Using Weka (version 3.3.6), 22 heterogeneous base-classifiers were created using their default settings (see Table 2.1). The base-classifiers were trained and evaluated using 10-fold cross validation on each of 17 datasets obtained from the UCI Machine Learning Repository (see Table 2.2). Datasets that included more than two classes were modified by either creating multiple subsets with only two classes in each subset

Table 2.1
Base-Classifiers

ADTree	MultilayerPerceptron
BayesNet	NaiveBayes
ConjunctiveRule	NBTree
DecisionStump	Nnge
DecisionTable	OneR
Ibk	PART
J48	RandomForest
JRip	RBFNetwork
KStar	Ridor
LMT	SimpleLogistic
LWL	SMO

Table 2.2
Datasets

Dataset	*Instances*	*Attributes*	*Positive Rate*	*Ensemble Diversity*	*Base-Classifier Accuracy*			
					Min	*Avg*	*Max*	*Std*
Adult	32,561	14	24.1%	0.847	75.9%	82.0%	86.0%	3.5%
Wisconsin Breast Cancer	699	110	34.5%	0.894	76.8%	94.0%	97.1%	4.2%
Contraceptive Choice	1,473	10	57.3%	0.698	60.0%	66.1%	71.0%	2.6%
Horse Colic	368	22	37.0%	0.868	78.0%	81.7%	86.1%	2.3%
Covertype (class 1 & 2)	10,000	11	72.9%	0.883	78.9%	86.6%	92.1%	3.8%
Covertype (class 3 & 4)	10,395	11	6.8%	0.954	93.0%	95.2%	97.5%	1.6%
Covertype (class 5 & 6)	10,009	11	66.9%	0.906	89.9%	96.2%	99.7%	3.5%
Australian Credit App.	690	15	55.5%	0.857	76.2%	83.7%	85.8%	2.6%
German Credit Approval	1,000	20	30.0%	0.778	63.7%	72.0%	75.7%	2.9%
Pima Indians Diabetes	768	8	34.9%	0.842	68.8%	73.5%	77.9%	2.5%
Thyroid Disease	3,772	5	7.7%	0.994	89.1%	92.7%	93.4%	1.2%
Labor	57	16	64.9%	0.487	68.4%	80.0%	93.0%	7.4%
Mushrooms	8,124	5	48.2%	0.775	77.0%	91.2%	96.6%	7.4%
Sick	3,772	12	6.9%	0.956	93.9%	96.8%	98.4%	1.2%
Spambase	4,601	58	39.4%	0.708	78.9%	87.6%	94.2%	5.6%
Splice-junction Gene Seq.	3,190	20	51.9%	0.498	53.4%	62.3%	67.1%	3.7%
Waveform	3,345	40	49.4%	0.790	77.6%	86.9%	92.7%	4.5%

or by combining classes. In order for computationally complex base-classifiers to complete the classification using a reasonable amount of resources, datasets with a large number of observation and/or attributes were filtered randomly based on records and/or attributes[4].

With an average dataset size of 5,350 records, a total of 2,000,900 base-classifiers validation decisions (5,350 records × 17 datasets × 22 base-classifiers) were generated from the 10-fold cross validation. These decisions were imported into Microsoft Access where combiner methods, implemented using Visual Basic, combined the data. Furthermore, since IMF, MAJ, AVG and dynamic WAVG did not require training, I did not use n-fold cross-validation in the combiner method experiments. Each dataset was combined 96 times as described in Section 2.4.2, for a

[4] The number of attributes/instances to delete was determined iteratively by first deleting only a few attributes/instances, running the most resource constraining base-classifier algorithm and then deleting more attributed/instances if needed. Attributes/instances selected for deletion were determined by assigning a random number to each attribute/instance using Microsoft Excel and then deleting the attributes/instances assigned the lowest number.

total of 8,745,888 (96 dataset combinations × 5,350 average dataset size × 17 datasets) ensemble decisions.

2.4.2 Experimental Design and Factors

The primary purpose of the computational experiments was to compare the effectiveness of IMF against MAJ, AVG and WAVG methods. As such, combiner method is the primary factor of interest. The experiment also included six other independent variables, two factors (cost-to-benefit ratio and number of agents) and four covariates (dataset positive ratio, dataset size, dataset average base-classifier accuracy and ensemble diversity), used to evaluate the sensitivity of the results (see Table 2.3).

As only main effects and second order interactions were investigated, and only interactions involving the combiner method factor, a full factorial design is not needed. Instead, two factorial block designs (4 combiner methods × 11 sets of agents, and 4 combiner methods × 13 cost-to-benefit ratios) were used, for a total of 96 treatment groups. The cost-to-benefit ratio factor was held constant at 1:10 in the 4×11 factorial design. The number of agents factor was held constant at 10 in the 4×13 factorial design. Net benefit and the covariates were measured for each of the 17 datasets within each of the 96 treatment cells for a total of 1,632 observations (4×11×17 + 4×13×17).

Table 2.3
Experimental Variables

Variable	*Function*	*Description*
Net Benefit	DV	FN cost avoidance * number of TP - investigation cost * (number of FP + number of TP)
Combiner Method	Main IV	IMF, AVG, WAVG, MAJ
Number of Agents	Manipulated Moderator	2, 4, 6,… , 22 agents in the ensemble
Cost-to-Benefit Ratio	Manipulated Moderator	1:100; 1:50, 1:25, 1:10, 1:7.5, 1:5, 1:4, 1:3, 1:2, 1:1.5, 1:1, 1.5:1, and 2:1
Dataset Size	Measured Moderator	Number of dataset records
Dataset Average Agent Accuracy	Measured Moderator	Average dataset accuracy of the base-classifiers
Dataset Positive Ratio	Measured Moderator	Positive records/total number of records in dataset
Ensemble Diversity	Measured Moderator	Dataset average pair-wise diversity measured using Yule's Q statistic

2.4.2.1 Dependent Measure

Performance measures used in MCC combiner method research include hit-rate (TP/(TP+FN)) and accuracy ((TP+TN)/(TP+TN+FP+FN)). However, accuracy and hit-rate only provide accurate measures of combiner method effectiveness under one specific scenario - when the number of positive and negative instances is the same, and the cost of FP and FN is the same. This is rarely true (Provost et al. 1998). More recently, Receiver Operating Characteristic (ROC) curves and the associated measure Area Under the ROC curve (AUC) have gained popularity, partially because they show how well algorithms handle the trade-off between true positive rate (TP/(TP+FN)) and false positive rate (FP/(FP+TN)), i.e., benefits and costs, without having to define a specific class distribution and cost assumption. ROC and AUC do not however allow for easy comparisons of combiner methods under specific distribution and cost assumptions that I am interested in. ROC also does not provide a single measure that allows us to assess the statistical significance and sensitivity of the relative combiner method performance results to various factors such as the number of base-classifiers in the ensemble, cost-to-benefit ratio, and dataset size, average agent accuracy, positive ratio and diversity (Drummond and Holte 2006). Furthermore, ROC curves are created using true positive rate and false positive rate, and therefore cannot be used in situations where TN and FN are not identified, i.e., in domains where negative classifications are not investigated.

Another common performance measure - misclassification cost, which is the total cost of FP and FN classifications - overcomes many of these shortcomings (Lin et al. 2003). However, this measure still requires knowing FN and ignores the costs associated with TP classifications, such as investigation costs. Chan et al. (1999) use Cost Savings (CS), which takes into account costs associated with TP, FP and FN, but still requires knowing FN. I use a measure very similar to CS[5] that I call net benefit (NB). NB, like CS allows us to overcome the problems described earlier, and in contrast to CS, does not require knowing FN. NB is calculated as the benefit derived from TP classifications (FN costs avoided) minus costs of investigating positive classifications, see (17). Like ROC curves, NB captures the trade-off between true positive rate

[5] It is shown in Appendix 5 that NB is equivalent to CS, given that Transaction Amount and Overhead (Chan et al. 1999) is defined as being equivalent to FN Cost Avoidance and Investigation Cost, respectively. Also note that my definition of cost-to-benefit ratio is based on the same idea used in Chan et al.'s (1999) rule: only transactions with transaction amounts > overhead should be investigated.

and false positive rate. To maximize NB, the classification threshold has to be selected so that it strikes an appropriate balance between net benefit of TP and cost of FP classifications.

NB = FN cost avoidance*number of TP - investigation cost*(number of FP and TP) (17)

The experiments compare the performance of various combiner methods using optimal thresholds for each treatment in order to isolate the treatment effect from noise introduced by using other mechanisms to determine the threshold. To determine the optimal thresholds, I run the MCC experiment 101 times for each treatment using a different threshold level (0, 0.01, 0.02,… 1) for each run. The threshold from the run that generates the highest total net benefit is then labeled as the optimal threshold for that specific treatment. By finding the best threshold for each combiner method, dataset, ensemble and cost-benefit ratio combination, the combiner methods are compared at optimal trade-off levels for that specific combination, which I believe is more relevant than comparing the sensitivity, specificity, hit-rate, etc, of the combiner methods at other sub-optimal levels. Furthermore, by comparing the combiner methods at a number of different cost-to-benefit ratios the generalizability of the results to different domains that have different cost-to-benefit ratios is improved.

2.4.2.2 Combiner Method Factor

Since the primary objective is to compare the performance of IMF to existing combiner methods, combiner method is included as a factor that is manipulated at four levels, IMF, AVG, MAJ and WAVG. IMF is compared to MAJ, AVG and WAVG since prior research indicates that AVG and MAJ perform well compared to other existing combiner methods (Duin and Tax 2000). WAVG is included primarily because of its similarity to IMF, since IMF generates a wealth weighted average. In MAJ, each base-classifier casts a vote on the class for which the base-classifier's probability estimate is higher than the classification threshold. The class with the most votes is then selected as the ensemble's decision. In AVG the mean of all the base-classifiers probability estimates is compared to the threshold and the class with a mean probability estimate that is higher than the threshold is selected as the ensemble's decision. In WAVG, different weights are assigned to the different base-classifiers' probability estimates when averaging these estimates. In order to maintain uniformity while comparing IMF to WAVG, I implement a dynamic version of WAVG where the weights are updated based on positive classifications only. The weights are determined as the ratio of an individual classifier's precision (TP/(TP+FP)) to the total precision of all the classifiers in the ensemble. I also test two alternative weighting schemes as detailed in Section 2.4.2.4.

2.4.2.3 Sensitivity Analysis

Number of Agents – The number of agents factor is manipulated at 11 levels: 2, 4, 6,…, 22 agents. This manipulation is done since there is evidence from prior research that the number of agents in an ensemble could impact ensemble classification performance (Lam 2000). The agents are randomly selected at each of the treatment levels, but the selection process is cumulative in nature. For ensembles consisting of two agents, the two agents are randomly selected from the 22 existing base-classifiers, for ensembles with four agents, two additional agents are randomly selected from the remaining 20 base-classifiers and added to the existing ensemble, and so on. To test the sensitivity of the combiner method performance to the number of agents, I examine if the relative combiner method performance is moderated by the number of agents, while holding the cost-to-benefit ratio constant at 1:10.

Cost-to-Benefit Ratio - The benefit derived from TP classifications (FN cost avoidance minus investigation cost) and the cost of FP classifications (investigation cost) impact the net benefit provided by any classification effort. As the cost-to-benefit ratio is domain specific, I use a wide range of cost-to-benefit ratios, 13 in total, to explore the generalizability of the results: 1:100; 1:50, 1:25, 1:10, 1:7.5, 1:5, 1:4, 1:3, 1:2, 1:1.5, 1:1, 1.5:1, and 2:1. To clarify, the 1:100 ratio indicates that the net benefit of a TP classification (cost of fraud minus investigation costs of detecting a fraud) is 100 times the cost of investigating a transaction (for example, cost of fraud = $10,100 vs. cost of investigation = $100). Note that the range of cost-to-benefit ratios used assumes that the net benefit of a TP is always positive, i.e., the FN cost avoided when making a TP classification is always more than the investigation cost. To examine the sensitivity of the combiner method performance to cost-to-benefit ratio, I investigate whether the relative combiner method performance is moderated by cost-to-benefit ratio, holding the number of agents constant at 10 agents.

Dataset Average Accuracy - Average base-classifier accuracy, measured as the percentage of all objects classified correctly for each dataset, is included as a possible interaction term given the possibility that relative combiner method performance could depend on the average accuracy of the base-classifiers in a given dataset. Thus, this interaction tests if the relative combiner method performance is moderated by the dataset average base-classifier accuracy.

Dataset Size - Dataset size refers to the number of records in the dataset, which varies from 57 to 32,561 records. Dataset size is included as a covariate primarily to examine the impact of size on the relative performance of IMF and WAVG to the other combiner methods. For example, if the data size is very small, the extent of adjustment of weights in WAVG and redistribution of

wealth in IMF is small. I therefore evaluate if the relative combiner method performance is moderated by the dataset size.

Dataset Positive Ratio - The positive ratio of the dataset refers to the number of positive class objects divided by all objects in the dataset. Positive ratio is included as a covariate to test if the relative combiner method performance depends on the dataset positive ratio. Theoretically, a performance difference, if any among combiner methods should be evident in datasets with positive ratios in the medium range, but not necessarily in datasets with very low (high) positive ratios where any trivial classifier that always predicts the object as negative (positive) does well. Thus, differences in performance among the combiner methods is only expected when the trivial rule is ineffective. The range of dataset positive ratios over which the trivial rule is effective is, furthermore impacted by the cost-to-benefit ratio level. The trivial rule that classifies everything as *positive* is effective over a wider range of dataset positive ratios (i.e., medium and high rather than just high dataset positive ratios) when the cost-to-benefit ratio is low (Witten and Frank 2005). Conversely, the trivial rule that classifies everything as *negative* is effective over a smaller range of dataset positive ratios (i.e., just extremely low rather than low positive ratios) when the cost-to-benefit ratio is low. Considering that the median of the experimental cost-to-benefit manipulations is close to 1:5, i.e., in the low range, I expect combiner method performance differences for low to medium dataset positive ratios, but not for medium to high dataset positive ratios. I therefore evaluate if the relative combiner method performance is moderated by the dataset positive ratio.

Ensemble Diversity - Base-classifier diversity describes the degree to which the ensemble base-classifiers differ in the errors they make. Diversity among the base-classifiers is incorporated in the experiment by using different learning algorithms for each base-classifier. Diversity is measured using Yule's Q statistic (Yule 1900) for each dataset. By measuring diversity I can evaluate if the relative performance of combiner methods is impacted by the level of complimentary information provided by the base-classifiers in the different datasets.

2.4.2.4 Investigating the True Class of All Objects

To evaluate the external validity of the result to domains where the true object class is revealed for all objects I perform an experiment where the performances of combiner methods are evaluated using both positive and negative classifications. In this experiment I examine a version of WAVG where wealth is updated for both positive and negative classifications, as well as aWAVG. In aWAVG, the weights are determined based on AdaBoost: ln((1-error rate)/error rate), where error rate is equal to (FP+FN)/(FP+FN+TP+TN).

2.4.3 Time Lag, IMF Parameters and Base-Classifier Cost-Benefit Retraining

IMF is a multi-classifier combiner method based on a pari-mutuel betting information market

2.4.3.1 Time Lag and Performance

In the main experiment the true class of t is given instantly after t is classified, but in reality, it usually takes some time to determine the true class of t. In order to determine the performance impacts of such time lags, I perform an experiment where wealth w_{it} is not updated until v additional objects have been classified. v is manipulated at six different levels: 0%, 1%, 5%, 10%, 25% and 50% of the size of the dataset, for each of the 17 datasets, while the main experiment factors are held constant as follows: combiner method = IMF; number of agents = 10; and cost-to-benefit ratio = 1:10. Using these treatments I investigate if the net benefits from 0%-IMF (no time lag) and the net benefits from 1%-IMF, 5%-IMF, 10%-IMF, 25%-IMF and 50%-IMF are significantly different.

2.4.3.2 Selection of IMF Parameters

Binary Search Stopping Parameter ε - The tolerance value ε is used in binary search to determine when to stop the search. To gain a better understanding of how to select an appropriate value for ε and to investigate if this selection is domain dependent, I run an experiment where different values of ε are tested. For a given value of ε (manipulated at 0.01, 0.001,… 0.00000000001), I run IMF on each of the datasets while holding other factors constant as follows: number of agents = 10; and cost-to-benefit ratio = 1:10. I am interested in investigating interactions between ε and the different dataset characteristics in order to assess whether ε is domain dependent. Also, if no interactions exist, I am still interested in investigating the direct impact of ε on net benefit.

Maximum Bet Multiplier *k* - To ensure that the ensemble is not completely dominated by a minority of better performing agents, while at the same time weighing the inputs of better performing agents more heavily, appropriate values of k are required to be used. For a given value of k (manipulated at 1, 2, 5, 10, 25, 50, 75, 100, 125, 150, 200, 250, 350, 500, and 1000), I run IMF on all datasets with the number of agents factor set at 10 and the cost-to-benefit ratio set at 1:10. I am also interested in investigating interactions between k and dataset characteristics to determine whether the choice of k is domain specific.

2.4.3.3 Base-Classifier Cost-Benefit Retraining

The ensemble base-classifiers in the experiments are not trained using cost sensitive learning and they are not retrained for each cost-to-benefit treatment level. The ensemble results are likely to change if the base-classifiers are retrained for different cost-to-benefit ratios. However, since

all four combiner methods are tested using the same base-classifiers, I do not believe that this will systematically bias the relative performance of the combiner methods. Nevertheless, I perform an experiment where the classification performances of various combiner methods are evaluated at five different cost-to-benefit ratios using an ensemble of five crisp base-classifiers used in two different modes: cost-to-benefit ratio retrained or not retrained. The retrained crisp base-classifiers are obtained by hardening measurement level base-classifiers at optimal thresholds for the different combinations of datasets and cost-benefit ratios. The base-classifiers that are not retrained are obtained by hardening the same base-classifiers using a threshold of 0.5. Holding the number of agents constant at five, I evaluate the effect of the interaction between base-classifier mode and combiner method on combiner method performance.

2.5. Results

2.5.1 Relative Combiner Method Performance

2.5.1.1 Overview

Table 2.4 provides an overview of the result data organized by the three result datasets used in the experiments. Two of the statistical analysis datasets are based on the 5×11 (combiner method by number of agents) and the 5×13 (combiner method by cost-to-benefit) factorial designs, while the third dataset is obtained by pooling the two statistical analysis datasets (possible as the interactions are not significant, as discussed in 5.1.3). I report two-tailed p-statistics throughout the chapter. For significance testing I use an alpha of 0.05, and 0.1 for marginal significance. To retain an experimentwise error rate of 0.05, while balancing the risk of type II errors, I use a modified Bonferroni procedure (Jaccard and Wan 1996).

2.5.1.2 Combiner Method Main Effect

The combiner method main effect is tested using the model shown in (18) and the pooled result set described earlier. Note that for each combination of UCI dataset*number of agents and UCI dataset*cost-to-benefit ratio, the same four combiner methods are tested. I therefore block for the dataset, number of agents, cost-to-benefit ratio, dataset*number of agents and dataset*cost-to-benefit ratio effects.

$$\ln(\text{net benefit}) = \beta_0 + \beta_1 * \text{combiner method} + \text{block} \qquad (18)$$

The combiner method main effect is significant ($p<0.0001$) and the post-hoc analysis show that IMF significantly outperforms AVG ($p=0.0042$), WAVG ($p=0.0229$) and MAJ($p<0.0001$). See Table 2.5 for parameter estimates and standard errors.

Table 2.4
Statistical Analysis Data

	Combiner Method x Number of Agents		*Combiner Method x Cost-to-Benefit Ratio*		*Combiner Method*	
	net benefit	*ln (benefit)*	*net benefit*	*ln (benefit)*	*net benefit*	*ln (benefit)*
Low	350	2.54	30	1.48	30	1.48
High	71,274	4.85	763,879	5.88	763,879	5.88
Mean	18,748	3.82	31,602	3.50	25,946	3.64
Standard Deviation	24,471	0.68	96,569	1.00	74,323	0.89
Number of Treatments	44		52		96	
N	748		884		1,632	

2.5.1.3 Sensitivity Analysis

The sensitivity of the relative performance of the combiner methods to the number of agents in the ensemble and the cost-to-benefit ratio are respectively tested using the model shown in (19) for the 5×11 factorial design, and model (20) for the 5×13 factorial design. Thus each combiner method is tested for all combinations of UCI dataset*number of agents and UCI dataset*cost-to-benefit ratio in the respective models. I therefore block for these interactions in the respective models.

$$\ln(\text{net benefit}) = \beta_0 + \beta_1 \text{combiner method} + \beta_2 \text{number of agents} + \beta_3 \text{combiner method*number of agents} + \text{block} \quad (19)$$

$$\ln(\text{net benefit}) = \beta_0 + \beta_1 \text{combiner method} + \beta_2 \text{cost-to-benefit ratio} + \beta_3 \text{combiner method*cost-to-benefit ratio} + \text{block} \quad (20)$$

The combiner method*number of agents (p=0.1407) and combiner method*cost-to-benefit ratio (p=0.7552) interactions are insignificant. This indicates that the performance advantages of IMF over AVG, WAVG and MAJ are not moderated by the number of agents in the ensemble or by the domain dependent cost-to-benefit ratio. Because prior research has not evaluated AVG, MAJ and WAVG under different cost assumptions, I perform further inspections of the interaction results using scatter plots and find that the performance differences among the combiner methods are stable over the different cost-to-benefit ratios tested. Thus, the earlier conclusion based on the statistical results is corroborated.

The sensitivity of the combiner method performance result to dataset average agent accuracy, size, positive ratio and ensemble diversity are tested using the same blocking factor and result set used for (18):

$$\ln(\text{net benefit}) = \beta_0 + \beta_1 \text{combiner method} + \beta_2 \text{dataset average agent accuracy} + \beta_3 \text{dataset size} + \beta_4 \text{dataset positive ratio} + \beta_5 \text{ensemble diversity} + \beta_6 \text{combiner method*dataset} \quad (21)$$

average agent accuracy + β_7 combiner method*dataset size + β_8 combiner method*dataset positive ratio + β_9 combiner method*ensemble diversity + block

The results do not show that relative combiner method performance is sensitive to the dataset size (p=0.7325) or dataset average agent accuracy (p=0.9803). I do however find that the combiner method*dataset ensemble diversity (p=0.0342) and combiner method*dataset positive ratio (p<0.0001) interactions are significant.

The interaction involving diversity appears to be driven by MAJ, as MAJ has a significant parameter estimate for the interaction (p=0.0039), while AVG (p=0.6363) and IMF (p=0.3227) are insignificant. This is verified by noting that the combiner method*diversity interaction is insignificant (p=0.7813) when MAJ is excluded from the analysis. When only including MAJ and IMF in the analysis the interaction is significant (p=0.0099). We, therefore, only perform a detailed analysis of IMF vs. MAJ.

Based on visual comparison (Figure 2.5) it appeared that IMF outperforms MAJ at all diversity levels, however, the performance difference is less at low diversity levels (high Yule Q). However, even at low diversity levels (Q>75) IMF outperforms MAJ (p<0.0138). Thus, at all diversity levels IMF outperformsMAJ as per this test, and AVG and WAVG as per the insignificant interaction and significant main effect.

I explore the significant combiner method*positive ratio interaction (p=0.0342) by dividing the datasets into two groups based on the dataset positive ratio, a high group with about half the datasets, positive ratio (>40%) and a low group with the remaining datasets (<=40%). In each group, a model with the combiner method factor and the blocking variables as in (18) are then tested. IMF significantly outperforms AVG (p=0.0005), MAJ (p<0.0001) and WAVG (p=0.0021) in the low group. In the high group, IMF significantly outperforms MAJ (p=0.0139), but the performance advantage is insignificant with respect to AVG (p=0.3439) and WAVG (p=0.6023).

2.5.1.4 Investigating the True Class of All Objects

The results that are obtained when the true classes of both positive and negative classifications are revealed are statistically equivalent to the results presented in Sections 2.5.1.2. and 2.5.1.3 with the following exceptions: 1) results are not sensitive to either diversity (p=0.4599) or positive ratio (p=0.0847), 2) IMF still significantly outperforms MAJ (p=0.0078) and WAVG (p=0.0767), and the performance advantage over AVG is now only marginally significant (p=0.1264). Note, that the p-values are two-tailed. The results also show that IMF outperforms aWAVG (p<0.0001). Table 2.5 summarizes these results.

2.5.2 *Time Lag, IMF Parameters and Base-Classifier Cost-Benefit Retraining Overview*

The impact of time lag on net benefit is tested using the model shown in (22) and a statistical analysis dataset derived from holding the number of agents and cost-to-benefit ratio constant at 10 and 1:10 respectively. Since all UCI datasets are used for all the treatments in the model I block for the dataset effect.

$$\ln(\text{net benefit}) = \beta_0+\beta_1 v+\text{block} \quad (22)$$

The lag-level main effect (p=0.9962) is insignificant, thereby indicating that time lag does not impact IMF performance.

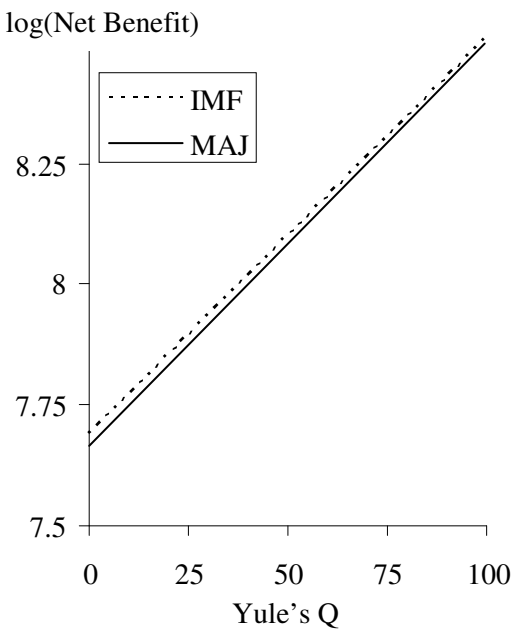


Figure 2.5: Combiner Method (MAJ and IMF) x Diversity Interaction

Using the model shown in (23), I do not find any evidence that the value of the binary search stopping parameter ε, within the tested range (0.01, 0.001,… 0.00000000001), impacts the performance of IMF (p=0.5071). The impact of ε on the performance of IMF is also not domain dependent. More specifically, while blocking for the dataset effect on net benefit, the ε*dataset positive ratio (p=0.1248), ε*dataset size (p=0.1856), ε*dataset average agent accuracy (p=0.5989) and ε*dataset diversity (p=0.8897) interactions are insignificant. Based on the results, in all experiments ε is set to the middle value tested ε = 0.000001.

$$\begin{aligned}\ln(\text{net benefit}) = {} & \beta_0 + \beta_1\varepsilon + \beta_2\text{dataset average agent accuracy} + \beta_3\text{dataset} \\ & \text{positive ratio} + \beta_4\text{dataset size} + \beta_5\text{dataset diversity} + \\ & \beta_6\varepsilon^*\text{dataset average agent accuracy} + \beta_7\varepsilon^*\text{dataset positive} \\ & \text{ratio}+\beta_8\varepsilon^*\text{dataset size}+\beta_9\varepsilon^*\text{dataset diversity} +\text{block}\end{aligned} \quad (23)$$

In order to choose appropriate values for the maximum bet multiplier k, and to investigate if the choice of k is domain dependent, I use the model shown in (24), where the block factor is dataset:

$$\begin{aligned}\ln(\text{net benefit}) = {} & \beta_0 + \beta_1 k + \beta_2\text{dataset average agent accuracy} + \beta_3\text{dataset} \\ & \text{positive ratio} + \beta_4\text{dataset size} + \beta_5\text{dataset diversity} + \\ & \beta_6 k^*\text{dataset average agent accuracy} + \beta_7 k^*\text{dataset positive} \\ & \text{ratio} + \beta_8 k^*\text{dataset size} + \beta_9 k^*\text{dataset diversity} + \text{block}\end{aligned} \quad (24)$$

Table 2.5
Summary of Primary Results

Effects	*Results*	*Sig[a].*
Net Benefit (only true class of objects classified as positive known)	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{AVG}$	p=0.0042
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{WAVG}$	p=0.0229
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{MAJ}$	p<0.0001
Parameter Estimates (standard errors)		
AVG		-0.00101 (0.00154)
IMF		0.00621 (0.00154)
MAJ		-0.00568 (0.00154)
Sensitivity Analysis		
Number of Agents	Not Sensitive	p=0.1407
Cost-to-Benefit Ratio	Not Sensitive	p=0.7552
Dataset Size	Not Sensitive	p=0.7325
Average Agent Accuracy	Not Sensitive	p=0.9803
Diversity*Method		
> without MAJ	Not Sensitive	p=0.7813
> without AVG/WAVG	Sensitive	p=0.0099
Net Benefit: Low Diversity	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{MAJ}$	p=0.0138
Net Benefit: High Diversity	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{MAJ}$	p<0.0001
Net Benefit: Low Positive Ratios	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{AVG}$	p=0.0005
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{WAVG}$	p=0.0021
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{MAJ}$	p<0.0001
Net Benefit: High Positive Ratios	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{AVG}$	p=0.3439
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{WAVG}$	p=0.6023
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{MAJ}$	p=0.0139
Parameter Estimates (standard errors)		
AVG		-0.00210 (0.00180)
IMF		0.00799 (0.00180)
MAJ		-0.00509 (0.00180)
Net Benefit (true classes of all objects known)	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{AVG}$	p=0.1264
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{WAVG}$	p=0.0767
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{MAJ}$	p=0.0078
	$\text{Net Benefit}_{IMF} > \text{Net Benefit}_{AWAVG}$	p<0.0001
Sensitivity Analysis		
Number of Agents	Not Sensitive	p=0.9848
Cost-to-Benefit Ratio	Not Sensitive	p=0.9944
Dataset Size	Not Sensitive	p=0.8293
Positive Ratio	Not Sensitive	p=0.0847
Average Agent Accuracy	Not Sensitive	p=0.4030
Ensemble Diversity	Not Sensitive	p=0.4599

[a] all p-values are two-tailed

The *k**dataset average agent accuracy (p=0.0036) and *k**dataset diversity (p=0.0012) interactions are significant, while the *k**dataset positive ratio (p=0.1581) and *k* *dataset size (p=0.1812) interactions are insignificant. Scatter plots with trend lines and the raw data tables for

the standardized log net benefit of the different datasets at the 15 different *k* values indicate that the significant interactions are driven by extreme values of *k*. For low *k* values net benefit decreases as the diversity decreases or the average agent accuracy increases, and vice versa for high *k* values. However, *k*=50 consistently provides relatively good results, even when compared to extreme *k* values at their best performance levels. Furthermore, k=25 and k=75 also perform well. Results show that when only using *k*=25, *k*=50 and *k*=75, the *k**dataset average agent accuracy (p=0.5697) and *k**dataset diversity (p=0.9212) interactions are no longer significant. Based on these results, I set *k*=50 in all experiments.

Using the model shown in (25), the base-classifier cost-benefit retraining experiment results show an insignificant (p=0.2037) base-classifier mode and combiner method interaction, blocking for dataset, cost-to-benefit ratio and dataset*cost-to-benefit ratio effects. Thus, as expected I do not find evidence of relative combiner method performance being moderated by base-classifier mode[6].

$$\ln(\text{net benefit}) = \beta_0+\beta_1\,\text{combiner method} +\beta_2\,\text{base-classifier mode} + \beta_3\,\text{combiner method*base-classifier mode} + \text{block} \quad (25)$$

2.6. Discussion Overview

2.6.1 Combiner Method Performance

Based on the β coefficients and standard errors from the main effects test (Table 2.5), IMF on average provides a 0.72, 1.19 and 0.55 percentage[7] greater impact on Net Benefit than AVG, MAJ and WAVG, respectively. These results are not sensitive to the number of agents in the ensemble, cost-to-benefit ratio, dataset size, dataset average agent accuracy or ensemble diversity. I however do find that the relationship between combiner method and net benefit is moderated by the dataset positive ratio when assuming that the true class of objects is only revealed for objects classified as positive. The results show that IMF outperforms the other combiner methods at low to medium positive ratios, and that there is no significant difference at medium to high positive

[6] Nevertheless, the reader should be aware that some of the crisp base-classifiers used in the experiments under extreme cost-benefit ratios could reduce the benefits of MCC. The effect of this potential problem is unknown, but based on the experiment just discussed it does not bias the relative performance results one way or the other.

[7] Based on estimation using β coefficients and standard errors of: -0.00101 and 0.00154 for AVG, 0.00621 and 0.00154 for IMF, and -0.00568 and 0.00154 for MAJ (Kennedy 1981).

ratio levels as theoretically expected. Thus, IMF performs well at all positive ratio levels and outperforms the other combiner methods when it matters the most, i.e., in skewed datasets with low to medium positive ratios for which, given the tested cost-benefit ratios, trivial rules classifying all objects as either positive or negative are likely to be ineffective. For low to medium positive ratios, IMF on average has a 1.01%, 1.31% and 0.71%[8] greater impact on Net Benefits than AVG, MAJ and WAVG, respectively.

To put this into perspective assume a fraud classification task where the average cost savings from a fraud detection is $20,000, the average cost of investigation is $500, the positive rate is 1% (a low positive rate), there are 40,000 transactions per year, and using IMF 50% of the positive instances and 98% of the negative instances are classified accurately. The benefit from this classification is $4,000,000 ($20,000*200), the cost is $496,000 (200*500+792*500), and the net-benefit is $3,504,000. In this example, IMF provides an additional benefit per year of $35,390, $45,902 and $24,878 over AVG, MAJ and WAVG, respectively. However, note that IMF also consumes more resources, an average of 6.21 msec of CPU time[9] to classify one object as compared to 5.07 msec for WAVG, 1.68 msec for AVG, and 1.69 msec for MAJ. The slightly larger CPU time consumed by IMF is minor given that in most settings IMF classifies on average, 579,415 objects per hour of CPU time on an off-the-shelf PC.

IMF also outperforms MAJ, WAVG, and aWAVG, and AVG (at a marginal level of significance), when true classes of objects classified as positive as well as negative are revealed. These results are robust within a wide range of cost-to-benefit ratios, number of agents in the ensemble, ensemble diversity, and dataset size, positive ratio and average base-classifier accuracy.

To understand why IMF has superior performance compared to the other combiner methods, we need to understand the workings of IMF. Because of the log utility function, IMF should perform on par with AVG if all the agents have the same amount of funds available for placing

---

[8] Based on estimation using β coefficients and standard errors of: -0.002053 and 0.001751 for AVG, 0.0079922 and 0.001751 for IMF, and -0.005092 and 0.001751 for MAJ (Kennedy 1981).

[9] I used GetProcessTimes from Kernel32.lib, which measures CPU time used, rather than actual time to run the algorithms. CPU time excludes time that the process is waiting for other processes to complete. The resource consumption experiment is performed on computers ranging from a desktop computer with a Pentium 4 2.0 GHz processor with 256 MB of RAM to a personal laptop with an AMD Turion 64 X2 Mobile Technology TL-56 processor and 2,048 MB of RAM. The computer is held constant within each treatment group.

bets in the market, i.e., if the aggregation is not wealth weighted (Wolfers and Zitzewitz 2006). However, as more accurate agents become wealthier, these agents end up influencing the market prices to a greater degree than the less accurate agents, and as the equilibrium prices represent the aggregated probabilities of the ensemble, the more accurate agents have a greater impact on the ensemble's decision than the less accurate agents. Thus the ensemble decision in IMF is a performance-weighted average, which explains why there is a difference between IMF and AVG, and also perhaps why IMF outperforms AVG, and to some extent MAJ, since MAJ is also a non-weighted combiner method with performance similar to AVG.

When comparing IMF to WAVG, we need to examine three major differences between IMF and WAVG: 1) WAVG assigns weights solely based on the precision of the base-classifiers relative to the precision of the other base-classifiers. IMF in contrast, places progressively greater weight on better performing agents' decisions, as agents with wealth above what they are allowed to bet hedge their bets to a lesser degree than other agents; 2) In IMF, weights are adjusted based on the degree of agent's performance as opposed to WAVG where the weights are adjusted solely based on the ratio of an individual classifier's precision to the total precision of all the classifiers in the ensemble. To clarify, in IMF, an agents' wealth increases (decreases) to a greater degree the more (less) accurate the agent is in each bet as agent bets are increasing in agent probability estimates. Thus, agents that are correct and more certain, receive a higher payout than agents that are correct but less certain, since the bets of more certain agents are higher, and vice versa; 3) The weights in IMF, but not in WAVG, are adjusted based on agents' relative contribution to the ensemble diversity. In IMF, agents with correct bets receive a greater payout if the odds are higher for that class, which occurs when the bets are higher for the other class.

2.6.2 Time Lag, IMF Parameters and Base-Classifier Cost-Benefit Retraining

The results do not show that time lag between object classification and object determination impact the performance of IMF within the range tested (0 to 50% of the records in the dataset). Thus, there is no evidence to suggest that the performance of IMF deteriorates with time lags between object classifications and object true class determination. The results also do not indicate that the binary search stopping parameter ε and maximum bet parameter k should be set to different values for different classification domains. In the experiments ε was held constant at 0.000001 and k was held constant at 50. I also do not find any evidence that there is a systematic bias in the relative performance of the combiner methods from not retraining the base-classifiers for the different cost-to-benefit ratios.

2.6.3 Combiner Method Design Considerations

For multi-agent system MCC implementations, IMF handles changes in ensemble composition and base-classifier performance. The market mechanism used in IMF functions independently of any specific agents that participate in the market. Furthermore, changes in an agent's relative performance impact the agent's wealth, and therefore also the weight given to the agent's decisions in the decision fusion process. IMF also provides market participants incentives to truthfully provide their private decisions. This is especially useful in multi-agent systems based on competitive agents (Ygge and Akkermans 1999).

2.7. Conclusions and Future Research Directions

In this essay, I present IMF, a new and novel combiner method based on information markets for multi-classifier combination. I show through extensive experimentation that IMF provides additional utility compared to three benchmark combiner methods AVG, WAVG and MAJ.

For future research, the effectiveness of IMF can be compared to other combiner methods in other multi-classifier combination architectures, such as bagging and boosting. Other research extensions include: investigating the performance impacts of other types of agent behavior using utility functions such as Constant Absolute Risk Aversion, Constant Relative Risk Aversion, etc; modeling agents to update their beliefs based on market signals or ensemble consensus; mixing agents with different utility functions; using a combination of human and software agent experts. IMF can also be extended for the more general k-class classification problem using the pari-mutuel betting mechanism. Finally, future research can explore the possibility of integrating the cost-benefit ratio into IMF itself.

Chapter 3. The Effect of Discretionary Accruals, Earnings Expectations and Unexpected Productivity on Financial Statement Fraud

3.1. Introduction

The Association of Certified Fraud Examiners (ACFE 2006) estimates that occupational fraud totals $652 billion per year in the U.S. Within occupational fraud, financial statement fraud (henceforth fraud) has the highest per case cost and total cost to the defrauded organization, with an estimated total cost of $340 billion per year in the U.S.[10] In addition to the direct impact on the defrauded organizations, fraud adversely impacts employees, investors and lenders. Fraud also has broader, indirect negative effects on market participants by undermining the reliability of corporate financial statements, which results in higher risk premiums. Despite recent legislation aimed at reducing fraud, fraud remains a prevalent problem and is considered to have remained at about the same level (ACFE 2006) or to have even risen lately (Oversight 2005).

Accounting professionals are increasingly assuming, through mandates and self-regulation, the responsibility for detecting fraud. Statement on Auditing Standard (SAS) No. 53, did not directly address the auditors' responsibility for providing a reasonable assurance that the financial statements are free of material misstatements due to fraud, but did so indirectly through reference to "irregularities" (AICPA 1988). However, starting with SAS No. 82, auditing standards refer to fraud directly; auditors should provide "reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud" (AICPA 1997, AU 110.02). SAS No. 99 reiterates this responsibility and further requires that analytical procedures be used specifically for the purpose of identifying risks related to fraud (AICPA 2002). Auditing Standard (AS) 2 (PCAOB 2004) specifies that managers should design and implement internal controls to address fraud risk (primarily for fraud prevention and detection),

[10] The ACFE (2006) report provides estimates of total fraud cost, mean cost per fraud category and number of cases. To derive the estimate for total cost of financial statement fraud, I assumed that the relative difference in mean is similar to the relative difference in median cost among the different occupational fraud categories.

and auditors should evaluate these internal controls. Finally, AS5 (PCAOB 2007) adopts a top-down audit approach and highlights that a fraud risk assessment should be taken into account when planning and performing the audit of internal control over financial reporting, which in turn impacts the audit. To summarize, the importance of fraud from an audit perspective has shifted from auditing standards only containing implicit reference to fraud, to fraud being one of the primary considerations of auditing standards.

Research that adds to our knowledge about fraud antecedents and detection is important to defrauded organizations, their employees, investors, lenders and financial markets, in general, as this knowledge can help curb costs associated with fraud and improve market efficiency. This knowledge is also important to auditors when providing a reasonable assurance about whether the financial statements are free of material misstatements caused by fraud, especially during client selection and continuation judgments, and audit planning. My research objective is to improve our understanding of antecedents of fraud, and thereby improve our ability to detect fraud. More specifically, I address three research questions not previously examined in the fraud literature: (1) what is the relation between the usage of discretionary accruals in prior years and fraud; (2) are managers that meet or exceed analyst forecasts more likely to have committed fraud; and (3) are firms with unexpected increases in revenue per employee more likely to have committed fraud?

The results of my research confirm that the likelihood of fraud is significantly higher for firms that meet or exceed analyst forecasts, are constrained by prior year earnings management, or have high labor productivity. These findings add to our theoretical understanding of fraud and at the same time make a practical contribution by improving our ability to detect fraud.

The chapter is organized as follows. A brief definition of earnings management, financial statement fraud and earnings manipulation as used in this study is provided in Section 3.2 along with a review of related fraud research. The research hypotheses are developed in Section 3.3. I present the research design, including descriptions of the sample selection, measures and descriptive statistics, in Section 3.4. The results are reported in Section 3.5 and additional analyses are provided in Section 3.6. Section 3.7 concludes the chapter with a discussion of research contributions, limitations and future research opportunities.

3.2. Related Research

Healy and Wahlen (1999) state that: “earnings management occurs when managers use judgment in financial reporting and in structuring transactions to alter financial reports to either mislead some stakeholders about the underlying economic performance of the company or to influence contractual outcomes that depend on reported accounting numbers” (p. 368). While

fraud has the same objective as earnings management, i.e., to alter financial reports with the intention of misleading its users, it differs from earnings management in that fraud is outside of generally accepted accounting principles (GAAP), whereas, earnings management is within GAAP. While this definition is clear cut, the distinction in reality is less clear. Rather than defining earnings management and fraud as two distinct classes, I view earnings management and fraud as being on opposite ends of a continuum, where the extremes are represented by earnings alterations that are either within or outside of GAAP. I furthermore use the term earnings manipulation to refer to the entire continuum, i.e., I consider both earnings management and fraud to be sub-categories of earnings manipulation. Fraud and earnings management also differ on two other important dimensions: 1) earnings management reverses over time but fraud does not; and 2) there are potential legal costs associated with fraud but not with earnings management. These dimensions will be further discussed in subsequent sections.

I now turn to the prior literature. Because of the importance of understanding fraud antecedents and improved fraud detection, a stream of research has focused on developing new predictors that explain and predict fraud. This research stream has taken either a confirmatory or exploratory approach. The confirmatory predictor research, the approach followed in this essay, has focused on testing specific fraud hypotheses primarily grounded in earnings management and corporate governance literature. The exploratory predictor research has taken a large number of variables, for example red flags proposed in SAS No. 53 and No. 82, and financial statement ratios, and either mapped these variables to fraud frameworks and/or tested their explanatory power. There has, however, been relatively little agreement in the results from the exploratory research as to what variables are significant predictors of fraud. To reduce the risk of obtaining statistically significant findings with low generalizability I follow the confirmatory predictor research approach, and propose and evaluate three novel fraud predictors.

The next section (Section 3.2.1) reviews research examining the impact of earnings management on fraud. This research proposes, but only partially tests, that the act of earnings management increases the likelihood of subsequent fraud (Beneish 1997, Lee et al. 1999). Section 3.2.2 reviews fraud predictor research that leverages earnings management hypotheses, more specifically the debt covenant and the bonus plan hypotheses. This research has examined whether earnings management motivations also provide incentives for managers to commit fraud (Dechow et al. 1996, Beneish 1999). Finally, Section 3.2.3 describes research that has examined predictors related to the revenue account, which is the most commonly manipulated financial statement account (Beneish 1997).

3.2.1 Fraud Motivated by Prior Years' Earnings Management

Prior fraud research has made the argument that as income-increasing accruals at some point reverse (Healy 1985), managers with income increasing accruals in prior years either have to deal with the consequences of the accrual reversals or commit fraud to offset the reversals (Dechow et al. 1996, Beneish 1997, Beneish 1999, Lee et al. 1999). Prior year income-increasing discretionary accruals might also cause the managers to run out of ways to manage earnings. When faced with these earnings reversals and decreased earnings management flexibility, managers can resort to fraudulent activities to achieve objectives that were earlier accomplished by managing earnings. A positive relation is, therefore, expected between prior discretionary accruals and fraud. I name this relation, and henceforth refer to it as the earnings reversals hypothesis.

The earnings reversals hypothesis was graphically depicted in Dechow et al. (1996) (see Figure 3.1). Fraud firms appeared to have greater total and discretionary accruals to assets in the three years, t_{-3}, t_{-2} and t_{-1}, leading up to the first fraud year, t_0, than did non fraud firms. The statistical analysis in Dechow et al. (1996), however, only examined the relation between total accruals in year t_0 and fraud in year t_0, rather than in the years prior to t_0, as predicted by the earnings reversals hypothesis and indicated graphically by Dechow et al. (1996). Dechow et al. (1996) found a significant positive relation between total accruals in year t_0 and fraud in year t_0. Contrary to this result, Beneish (1997) found a negative relation between total accruals in year t_0

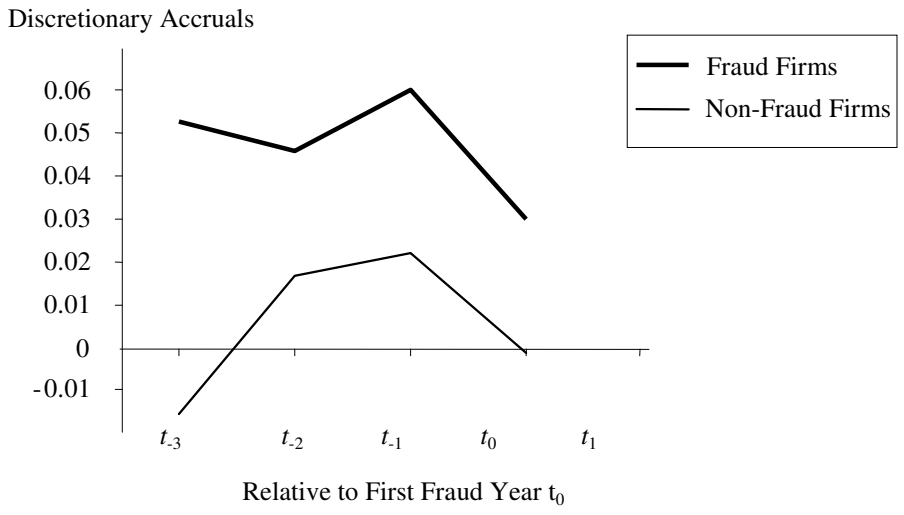


Figure 3.1: Income-Increasing Discretionary Accruals of Fraud and Non-Fraud Firms

and fraud in year t_0. The likelihood of fraud in year t_0 was, however, positively related to a dummy variable measuring whether the firm had positive accruals in both year t_{-1} and t_0 (Beneish 1997). A more recent paper by Beneish (1999), reported a positive relation between total discretionary accruals in year t_{-1} and fraud in year t_0. Lee et al. (1999) found a positive and significant relation between fraud and the difference between operating accruals summed over a three-year time span prior to the fraud being discovered by the SEC. However, the SEC fraud discovery on average lags the first fraud occurrence by 28 months (Beneish 1999). Thus, for the average firm, the discretionary accruals measure used in Lee et al. (1999) was for total accruals summed over years t_{-1}, t_0 and t_{+1}. More recently, Dechow et al. (2007) found indications, though not supported based on a statistical test, of accruals reversing subsequent to t_0, thus providing further support for the earning reversal hypothesis.

To summarize, prior fraud research examining the earnings reversals hypothesis has primarily used current accruals, but also accruals one year prior to the first year of the fraud. However, the earnings reversals hypothesis used in these studies does not specify a relation between current income-increasing discretionary accruals and fraud. Furthermore, the pressure to commit fraud due to accrual reversals should be higher when the firms have used income-increasing accruals to boost income over multiple years rather than just one year. The graphical analysis (See Figure 3.1 for a similar analysis based on this study's data) in Dechow et al. (1996) indicates that an appropriate time period to measure income-increasing accruals is three years prior to the first fraud year rather than only one year prior to the first fraud year. I extend the fraud literature by validating the previously discussed, but not yet hypothesized and fully examined, relation between positive discretionary accruals in prior years and fraud.

3.2.2 Fraud and Earnings Management Motivations

Given the shared objective of fraud and earnings management, fraud research has examined whether the same incentives that motivate earnings management also motivate fraud. This research has focused on examining incentives related to the debt covenant hypothesis and the bonus plan hypothesis in a fraud context.

Beneish (1999) and Dechow et al. (1996) examine fraud incentives related to the debt covenant hypothesis. In earnings management, the debt covenant hypothesis predicts that when firms are close to violating debt covenants managers will use income-increasing discretionary accruals to avoid violating the covenants (Dichev and Skinner 2002). Beneish (1999) and Dechow et al. (1996) hypothesize a positive relation between demand for external financing and fraud, and between incentives related to avoiding debt covenant violations and fraud. Demand for

external financing is measured in both studies as whether the difference between cash flow from operations and average capital expenditures to current assets is less than -0.5, and whether securities were issued in the fraud period. Incentives related to avoiding debt covenant violations are measured in both studies using leverage and actual instances of technical default. The results of the studies are mixed with one study (Dechow et al. 1996) finding support for the hypothesized relationships and the other (Beneish 1999) finding no support.

Beneish (1999), Summers and Sweeney (1998), and Dechow et al. (1996) also examine fraud incentives related to the bonus plan hypothesis. In earnings management, the bonus plan hypothesis predicts that earnings based bonuses provide managers with an incentive to manage earnings to increase their bonuses over multiple years. More specifically, if bonuses are (not) increasing in earnings then managers will use income-increasing (income-decreasing) discretionary accruals to increase their current (future) bonuses (Healy 1985).

In terms of the bonus plan hypothesis, Dechow et al. (1996) and Beneish (1999) posit that managers have greater incentives to commit fraud when they can benefit from the fraud either through their compensation agreements or through insider trading. Both Dechow et al. (1996) and Beneish (1999) measure compensation agreement using a dummy variable of whether a bonus plan exists. Beneish (1999) also uses stock appreciation rights. Neither study finds support for the hypothesis that the existence of a bonus plan increases the likelihood of fraud. Beneish (1999), however, does find support for a positive relation between fraud likelihood and whether managers redeem stock appreciation rights.

While both Dechow et al. (1996) and Beneish (1999) also examine insider trading incentives, they use different measures for this construct. In Dechow et al. (1996) insider trading, measured as insider sales divided by market value of equity, is not found to be a significant predictor of fraud. Beneish (1999) argues that in addition to insider sales, insider purchases should be included as fraudsters have incentives to both sell more and purchase less of their companies' stock when committing fraud. Beneish (1999), therefore, uses the difference between insider purchases and sales, and divides this difference by total trading activity. Beneish (1999) also uses the percentage of firm security offerings sold by insiders. Unlike Dechow et al. (1996), Beneish (1999) obtains significant results for insider trading. Beneish (1999) also finds support for whether managers redeem stock appreciation rights. In a similar study, Summers and Sweeney (1998) examine insider sales and purchases. In addition to dollar amounts sold and purchased, Summers and Sweeney (1998) measure the number of shares and number of transactions in insider sales and purchases. They find that only the number of shares sold is a moderately significant predictor of fraud.

As shown, prior fraud research focusing on earnings management motivations in a fraud context has examined compensation and debt incentives but not fraud incentives related to capital market expectations; specifically, the relation between analyst forecasts and fraud. In earnings management, capital market expectation hypotheses predict that managers have incentives to manipulate earnings to meet or exceed analyst forecasts when these forecasts would not otherwise have been met or exceeded. These incentives are related to manager performance and compensation, and firm performance in general, which are often evaluated based on meeting or exceeding analyst expectations (Burgstahler and Eames 2006). I extend fraud research by examining fraud incentives related to capital market expectations.

3.2.3 Fraud in the Revenue Account

Prior fraud literature has identified the revenue account as being the primary target for financial statement fraud (Beneish 1997). Given that the revenue account is typically manipulated, unusual revenue levels or changes in revenue might be indicative of revenue fraud. However, considering that revenue varies from year to year and among firms for reasons other than fraud, straight revenue is a relatively noisy measure of fraud. For example, it is very difficult to disentangle differences in revenue due to fraud from differences in revenue due to the size of the firm and the successfulness of the firm. To detect revenue fraud, SAS No. 99 highlights the need to analyze and identify unusual relationships involving revenue, for example between revenue and production capacity.

Prior research has included sales in various ratios that are not, typically, designed for the purpose of detecting revenue fraud. Nevertheless, the results from these studies are largely consistent with fraud firms manipulating the revenue account. For example, sales growth, used as a proxy for firm growth, has been used as a predictor of fraud based on the idea that high-growth firms have incentives to sustain their high growth levels and that slow-growth firms have incentives to increase growth (Erickson et al. 2006; Brazel et al. 2007). Erickson et al. (2006) found a positive relation between sales growth and fraud. Brazel et al. (2007) examined the relation between performance improvements and fraud in more detail and found a negative relation between sales growth and fraud, and a positive relation between sales growth minus growth measured using a non-financial measure and fraud. Together these results indicate that firms that increase revenue fraudulently are more likely to have abnormally high growth rates, and that poorly performing firms, i.e., firms with low actual growth rates, are more likely to commit fraud.

Both Chen and Sennetti (2005) and Fanning and Cogger (1998) examine the relation between gross profit margin and fraud. Chen and Sennetti (2005) examine this relation to detect inflated sales, while Fanning and Cogger (1998) argued that it is an indication of deflated cost of goods sold. Both these studies find a positive relation between gross profit margin and fraud, providing an indication of revenue fraud or manipulation of cost of goods sold. Chen and Sennetti (2005) also find that fraud firms have lower ratios of research and development expenditures to sales, and sales and marketing expenditures to sales than non-fraud firms. These results seemingly argue that fraud firms are more likely to be financially distressed and therefore, less likely to invest in research and development, and sales and marketing. Alternatively, these relations could indicate revenue manipulation as revenue fraud decreases both these ratios.

While these studies support the conjecture that managers fraudulently increase sales, both Summers and Sweeney (1998) and Fanning and Cogger (1998) examine ratios of sales that do not show evidence of revenue manipulation. Summers and Sweeney (1998) find a positive relation between change in inventory to sales and fraud, which they interpret to be evidence of fraudulent inventory manipulation. Note that a fraudulent increase in sales would reduce the ratio of inventory to sales in the fraud year. Fanning and Cogger (1998) examine the ratio of sales to assets with the idea that firms with relatively low sales to asset ratios are in financial distress and therefore, more likely to commit financial statement fraud. As they expect, they find a negative relation between the ratio of sales to assets and fraud. Note that a fraudulent increase in sales would increase sales to assets, if it is assumed that assets is not changed.

I extend this research by developing a productivity based measure that is designed specifically for the purpose of detecting financial statement fraud. I use a productivity measure because firms use resources, for example assets, to generate revenue. Thus, some of the noise associated with using revenue as a predictor can be removed by deflating revenue by the resources used to produce the revenue. Because resources are used to generate sales, the relation between sales and resources should be relatively stable over time compared to straight sales. Given the identified importance of the revenue account in fraud, the difficulty in using straight revenue for fraud detection, the inability of sales to assets (capital productivity) to detect fraudulent revenue manipulation (Fanning and Cogger 1998) and the appeal of using a productivity measures to detect revenue fraud, I examine the use of a different productivity measure, labor productivity. The rationale for using labor productivity rather than capital productivity is provided in the discussion leading up to the third hypothesis.

3.3. Hypotheses Development

3.3.1 Prior Years' Discretionary Accruals and Fraud

Managers can use discretionary accruals to transfer earnings between periods but over time discretionary accruals sum to zero (Healy 1985). Thus, income-increasing behavior in one period decreases the amount of discretionary accruals that can be used to increase earnings in subsequent periods as the prior income-increasing discretionary accruals eventually reverse (Dechow et al. 1996; Beneish 1997). For example, managers make judgments about the amount of outstanding accounts receivables that are uncollectible and adjust allowance for uncollectible accounts based on this judgment by debiting bad debt expense. The manager can manage earnings by deciding to establish the allowance level below the manager's actual estimate, thereby lowering bad debt expense and increasing earnings. However, assuming that the initial judgment about the correct allowance level was more accurate than the established allowance, the allowance account will not be sufficient and has to be increased at some point to cover actual receivables that could not be collected, thereby increasing future bad debt expense and decreasing future earnings.

When confronted with accrual reversals, managers can choose to either face the consequences of net income-decreasing accruals or fraudulently manipulate earnings to offset or more than offset the reversals (Beneish 1997). Given that managers facing accrual reversals can resort to fraudulent activities to achieve similar objectives that were earlier accomplished by managing earnings, I expect a positive relation between prior discretionary accruals and fraud. This relation was graphically depicted but not tested in Dechow et al. (1996), where fraud firms appeared to have higher total and discretionary accruals in the three years leading up to the first fraud year than did non-fraud firms (see Figure 3.1).

Based on this I posit[11] that the pressure of accruals reversal is greater and that earnings management flexibility is reduced the more earnings were managed in prior years. The pressure

[11] Note that firms with strong performance are less likely to resort to fraudulent activities to offset earnings reversals as their strong performance offsets the reversals and vice-versa for firms with poor performance. However, on average, firms facing accrual reversals are more likely to commit fraud than firms that are not facing accrual reversals. Although the posited relation could be further refined by taking into consideration firm performance, I do not hypothesize an interaction between performance and accrual reversals as firms that commit fraud also report higher performance. That is, while firms with low performance are more likely to commit fraud when faced with accrual reversal, firms that commit fraud are also more likely to report better performance.

from earnings reversals provides an incentive to manipulate earnings and the earnings management inflexibility increases the likelihood that fraud, rather than earnings management, is used to manipulate earnings. I define total prior discretionary accruals as discretionary accruals summed over three years prior to the first fraud year. In accordance with the earnings reversals hypothesis:

H1*: Total prior discretionary accruals is positively related to the likelihood of fraud.*

3.3.2 Capital Market Expectations and Fraud

Firm performance, and consequently market value, is partially determined by the firms' ability to meet or exceed analyst expectations. Managers, therefore, have incentives to manipulate earnings to meet or exceed analyst forecasts when these forecasts would not otherwise have been met or exceeded (Burgstahler and Eames 2006). Managers can manipulate earnings to meet or exceed analyst forecasts by managing earnings or by committing fraud.

When earnings are manipulated using earnings management, managers are likely to manage earnings to just meet analyst forecasts (Burgstahler and Eames 2006). While there are incremental benefits associated with exceeding forecasts, managers prefer to just meet analyst forecasts as the costs of earnings management also increase when forecasts are exceeded (Burgstahler and Eames 2006). One such cost relates to future earnings being negatively impacted by current earnings management. To be able to meet analyst forecasts in future periods managers are, therefore, likely to manage earnings to meet, rather than exceed analyst forecasts.

While prior research has not examined the relation between analyst forecasts and fraud, Dechow et al. (2007) show that fraud firms have unusually strong stock price performance prior to committing fraud, and indicate that this may put pressure on the firm to commit fraud to avoid disappointing investors and losing their high stock prices. Additionally, a large number of SEC Accounting and Auditing Enforcement Releases (AAER) provide anecdotal evidence of specific cases where fraud was committed to meet or exceed analyst forecasts. Thus, there are reasons to believe that managers may fraudulently manipulate earnings to meet or exceed analyst forecasts. As in earnings management, both the incremental benefits from meeting or exceeding analyst forecasts and expected costs associated with fraud are increasing in the magnitude of the fraud. However, earnings manipulated using fraud, as opposed to earnings manipulated using earnings management, do not reverse in future periods; therefore, it is difficult to predict whether managers prefer to fraudulently manipulate earnings to meet or to exceed forecasts. Since the exact nature of the utility managers derive from meeting or exceeding analyst forecasts when committing fraud is unknown, I define meeting or exceeding analyst forecasts as a dummy

variable that equals one if analyst forecasts are met or exceeded rather than attempting to define a cut-off as is done in earnings management research (Burgstahler and Eames 2006). Based on this discussion I hypothesize:

H2: *Firms that meet or exceed analyst forecasts are more likely to have committed fraud than firms that fail to meet analyst forecasts.*

3.3.3 Unexpected Labor Productivity and Fraud

The revenue account is the most commonly manipulated account in fraud (Beneish 1997). Thus, unusual increases in revenue could be an indication of fraud. To reduce some of the noise associated with this measure, revenue can be deflated by assets (capital productivity). Prior research has found capital productivity to be a significant predictor of fraud (Fanning and Cogger 1998; Kaminski et al. 2004).

However, capital productivity is still a somewhat noisy measure given constant changes in assets that do not directly impact revenue. Furthermore, and more importantly, given that accounting information systems are double-entry based, the utility of this measure in detecting fraud is reduced; for example, fictitious revenue will increase both the numerator (sales) and the denominator (assets) in capital productivity. The direction and magnitude of change in capital productivity resulting from revenue fraud depends on the level of a firm's actual capital productivity and profit margins. As an illustration, take firm A and firm B that both fraudulently increase sales by $10 million, which in turn increases assets by $5 million. Further assume that: (1) both firms have $100 million in assets before manipulating sales; (2) firm A has pre-manipulation sales of $50 million; and (3) firm B has pre-manipulation sales of $250 million. Under these assumptions, sales to asset *increases* from 0.5 to 0.57 for firm A and *decreases* from 2.5 to 2.48 for firm B. Thus, because revenue fraud increases both the numerator and the denominator of capital productivity, the ability of capital productivity to predict revenue manipulations is reduced.

In support of this discussion, Fanning and Cogger (1998) did not find a positive relation between capital productivity and fraud. They instead found a negative relation, which was described as showing that firms in financial distress are more likely to commit fraud. Thus, it is questionable whether it is possible to use sales to assets as evidence of revenue manipulation.

Labor productivity, another form of productivity, is measured as the amount of output per employee. Like capital productivity, labor productivity reduces the noise associated with sales by scaling sales by the input that is used to generate the output. However, unlike capital productivity, the denominator in labor productivity is not impacted by double-entry systems. Therefore, labor

productivity should be a less noisy predictor of revenue fraud. A recent working paper by Brazel et al. (2007) provides additional support for use of the number of employees as the denominator. This study examines the efficacy of nonfinancial measures, including the number of employees, in predicting fraud. They argue that nonfinancial measures that are strongly correlated to actual performance and at the same time relatively difficult to manipulate, like number of employees, can be used to assess the reasonableness of performance changes. The results in Brazel et al. (2007) show a positive relation between fraud and the difference between change in revenue and change in the nonfinancial measures.

Based on this discussion, I propose that firms that have high unexpected labor productivity are more likely to have committed fraud. I measure unexpected labor productivity as the percentage change in firm labor productivity from year t_{-1} to year t_0, minus the percentage change in industry labor productivity from year t_{-1} to year t_0, and hypothesize that:

H3: *Unexpected labor productivity is positively related to the likelihood of fraud.*

3.4. Research Design

3.4.1 Variable Construction

3.4.1.1 Total Discretionary Accruals

To test H1 a measure of total prior discretionary accruals that captures the pressure of earnings reversals and earnings management inflexibility is needed. I define *Total Discretionary Accruals*$_{j,t}$ as the total amount of discretionary accruals in the three years prior to the first fraud year deflated by assets at the beginning of each year:

$$\textit{Total Discretionary Accruals}_{j,t} = \sum_{t-3}^{t-1} DA_{j,t} \,/\, A_{j,t-1}\,, \tag{26}$$

where discretionary accruals $DA_{j,t}$ is calculated as the difference between total accruals $TA_{j,t}$ and estimated accruals, typically referred to as nondiscretionary accruals, $N\hat{D}A_{j,t}$:

$$DA_{j,t} \,/\, A_{j,t-1} = TA_{j,t} \,/\, A_{j,t-1} - N\hat{D}A_{j,t} \,/\, A_{j,t-1}\,, \tag{27}$$

where total accruals, $TA_{j,t}$, is defined as income before extraordinary items (#18)[12] minus cash flow from operations (#308). Nondiscretionary accruals, $NDA_{j,t}$, for firm j in year t_0 is estimated using the extended version of the modified Jones model (Jones 1991; Dechow et al. 1995) proposed in Kasznik (1999). To derive $NDA_{j,t}$ the regression parameters in model (28) are

[12] Numbers in parentheses refer to the Compustat number for the variable identified and is provided first time the variable is used in the essay and in footnotes in tables.

estimated for firm j using all firms in J, where J is the two-digit SIC code industry of j. These estimates are then used to calculate estimated $NDA_{j,t}$ for firm j using model (29):

$$TA_{j,t} / A_{j,t-1} = \alpha_0 / A_{j,t-1} + \alpha_1(\Delta REV_{j,t} - \Delta REC_{j,t}) / A_{j,t-1} + \alpha_2 PPE_{j,t} / A_{j,t-1} + \alpha_3 \Delta CFO_{j,t} / A_{j,t-1} \quad (28)$$

$$\hat{NDA}_{j,t} / A_{j,t-1} = \hat{\alpha}_{0,J} / A_{j,t-1} + \hat{\alpha}_{1,J}(\Delta REV_{j,t} - \Delta REC_{j,t}) / A_{j,t-1} + \hat{\alpha}_{2,J} PPE_{j,t} / A_{j,t-1} + \hat{\alpha}_{3,J} \Delta CFO_{j,t} / A_{j,t-1}, \quad (29)$$

where $\Delta REV_{j,t}$ is the change in revenue (#12), $\Delta REC_{j,t}$ is the change in receivables (#2) and $\Delta CFO_{j,t}$ is the change in cash flow from operations of firm j from year t_{-1} to year t_0; $PPE_{j,t}$ is firm j's gross property, plant and equipment (#7) at time t_0; and all values are deflated by $A_{j,t-1}$, firm j's assets (#6) at time t_{-1}.

3.4.1.2 Forecast Attainment

I develop a measure of whether firms meet or exceed analyst forecasts to test H2. I define *Forecast Attainment*$_{j,t}$ as a dummy variable that measures whether or not analyst forecasts were met or exceeded:

$$Forecast\_Attainment_{j,t} = \begin{cases} 1, & if\ (EPS_{j,t} - AF_{j,t}) \geq 0 \\ 0, & if\ (EPS_{j,t} - AF_{j,t}) < 0, \end{cases} \quad (30)$$

where for firm j, $EPS_{j,t}$ is actual earnings per share in year t_0, $AF_{j,t}$ is the first one year ahead analyst consensus forecast of earnings per share for firm j in year t_0 based on mean I/B/E/S earnings forecasts.

3.4.1.3 Unexpected Revenue per Employee

To test H3 I develop an unexpected labor productivity measure. I define *Unexpected Revenue per Employee*$_{j,t}$ as the difference in percentage change in revenue per employee between firm j and industry J:

$$Unexpected\ Revenue\ per\ Employee_{j,t} = \%\Delta RE_{j,t} - \%\Delta RE_{J,t}, \quad (31)$$

where revenue per employee, RE, defined as total revenue to total number of employees (#29), is measured for firm j and for firm j's industry J in year t_0 and year t_{-1}. The percentage difference between revenue per employee in year t_0 and t_{-1} is labeled percentage change in revenue per employee:

$$\%\Delta RE_t = \frac{RE_t - RE_{t-1}}{RE_{t-1}} \quad (32)$$

3.4.1.4 Control Variables

Confirmatory fraud research typically relies on matching non-fraud firms to fraud firms based on size and year of fraud, and includes measured variables, to control for potential omitted variable bias. However, the use of control variables is not standard. For example, Beneish (1999) and Summers and Sweeney (1998) included additional control variables, while Dechow et al. (1996) did not. Further, the control variables have not been used consistently. Without a theoretical basis or empirical support for the most appropriate set of control variables, I rely on variables that, given my hypotheses, conceptually are omitted variables.

Exploratory fraud research offers a rich set of variables from which to select controls that could conceptually be omitted variables in my study. Fanning and Cogger (1998) investigated the predictive value of 62 potential fraud indicators. Using stepwise logistic regression[13] they derived a model with eight significant fraud predictors: percent of outside directors; non-Big 4 auditor; whether CFO changed in the last three years; whether LIFO was used; debt to equity; sales to assets; whether accounts receivable was greater than 1.1 of last year's accounts receivable; and whether gross margin percentage was greater than 1.1 of last year's. Bell and Carcello (2000) matched 305 non-fraud cases to the 77 fraud cases in Loebbecke et al. (1989) and evaluated the discriminatory power of the indicators used in Loebbecke et al. (1989). All variables were measured using a survey of auditors with questions asking about the existence of risk factors.[14] Based on univariate results and testing a number of different logistic regression models, the final model contained five significant risk factors: weak internal control environment, rapid company growth, undue emphasis on meeting earnings projections, management lied or was overly evasive, and whether the company is public. Using univariate tests Kaminski et al. (2004) found that five out of 20 financial ratios tested were significant predictors of fraud during the year of fraud and one, two and three years prior to the fraud: fixed assets to assets, sales to accounts receivable, inventory to current assets, inventory to sales, and sales to assets. However, for these four years, a total of 84 tests were calculated, thereby, greatly increasing the chances of finding some significant relations by chance alone. Further, it is likely that many of these ratios were

---

[13] Fanning and Cogger (1998) also used an artificial neural network and two versions of discriminant analysis in their multivariate analyses. However, statistical significance was not reported for the selected variables for these models.

[14] Many of the examined factors cannot be obtained from public sources and require actual audits to be conducted, for example, the risk factor indicating whether the internal control environment was weak.

correlated, possibly giving rise to multicollinearity issues. Based on their findings, Kaminski et al. (2004) concluded that the red-flag approach only provides limited utility in detecting fraud. Chen and Sennetti (2005) examined 17 computer industry specific fraud predictors selected based on the most common fraud types in the computer industry and found significance for eight variables: research and development to sales, gross profit margin, net profit margin, sales and marketing to sales, tax benefits from exercising of employee stock options to operating cash flows, changes in free cash flow, accounts receivable turnover, and return on assets.

In general, these studies find significant predictors of fraud that can provide utility in the detection of financial statement fraud. However, 87% of the tested variables in this literature are insignificant predictors of fraud, and there is very little overlap between studies as to the variables that are identified as significant predictors of fraud. I, therefore, select variables from one study and supplement these variables with variables that could conceptually be considered omitted variables from this study. I select controls from Fanning and Cogger (1998), who compared a relatively comprehensive set of 62 potential predictors covering a wide number of potential fraud predictor types ranging from corporate governance to financial ratios. From the eight significant predictors in Fanning and Cogger (1998) I use three variables, *CFO Change*, *Auditor* and *Sales to Assets*. I also add two controls that were not among the 62 variables examined in Fanning and Cogger (1998), *Asset Growth* and *Current Discretionary Accruals*.

CFO Change is a dummy variable that measures whether the CFO has changed during the three years leading up to the first fraud year. While Fanning and Cogger (1998) were expecting a positive relation based on the idea that some CFOs committing fraud will leave their firms to avoid getting caught or are fired because of fraud suspicion, they found a negative relation but do not provide an explanation for this finding. A possible explanation for the negative relation is that CFOs committing fraud are less likely to leave as by leaving they relinquish control over evidence of the fraud and expose themselves to scrutiny by the incoming CFO. I include *CFO Change* to control for the possibility that both *Total Discretionary Accruals* and *Fraud* are related to ineffective corporate governance. Based on the empirical results in Fanning and Cogger (1998) I expect a negative relation between *CFO Change* and *Fraud*. Note that Fanning and Cogger (1998) examined 31 variables related to corporate governance and found in multivariate analysis that only *CFO Change*, *Auditor* and the percentage of insiders on the board were significant predictors of fraud. To reduce the number of control variables I chose to include *CFO Change* to control for corporate governance effectiveness in the main analysis but not the percentage of insiders on the board. I made this selection after empirically comparing the predictive ability of the two variables: *CFO Change* (p=0.113) and percentage of outsiders on the board (p=0.267).

Auditor is a dummy variable that measures whether the firm's auditor is a Big 4 auditor. Big 4 auditing firms are believed to provide a higher quality audit, which in turn is expected to increase the effectiveness of the monitoring function provided by the auditors and thereby decrease the likelihood of fraud. Thus, *Auditor* is expected to be negatively related to *Fraud* (Fanning and Cogger 1998). Like *CFO Change*, *Auditor* is included to provide a measure of a corporate governance mechanism that could conceptually explain the hypothesized relation between *Total Discretionary Accruals* and *Fraud*.

Sales to Assets is the ratio of sales to assets (capital productivity). *Sales to Assets* is expected to be negatively related to fraud given that low *Sales to Assets* is an indicator of financial distress (Fanning and Cogger 1998). I include *Sales to Assets* to examine my argument in H3 that *Unexpected Revenue per Employee* is a better predictor of revenue fraud than *Sales to Assets*. The inclusion of *Sales to Assets* also allows me to examine whether *Sales to Assets* and *Unexpected Revenue per Employee* capture different dimensions of productivity that can lead to fraud – *Sales to Assets* capturing low productivity and financial distress that drives fraud, and *Unexpected Revenue per Employee* capturing productivity that is artificially high as a result of revenue fraud (H3).

I add two additional variables, *Current Discretionary Accruals* and *Asset Growth*. *Current Discretionary Accruals* is the discretionary accruals in the first fraud year, t_0, calculated using the extended version of the modified Jones model (Jones 1991; Dechow et al. 1995) proposed in Kasznik (1999), see (27), (28) and (29) in section 3.4.1.1. As an indication of management attitude towards fraud, I expect *Current Discretionary Accruals* to be positively related to fraud. Attitude (henceforth management character) is difficult to measure and as in prior fraud research, I have to assume that management character is not an omitted variable. However, *Current Discretionary Accruals* might proxy for management character given that management character is positively related to management's use of discretionary accruals. This argument is based on the assumption that a manager's attitude towards earnings management is an indication of the manager's attitude towards fraud. I include this management character proxy to control for the possibility that management character, i.e., poor set of ethical values, explains both *Total Discretionary Accruals* and *Fraud*.

Asset Growth is a dummy variable that measures whether assets exceed 110% of the previous year's assets. *Asset Growth* is expected to be positively related to *Fraud* given that small growth firms are more likely to be investigated by the SEC (Beneish 1999) than larger slower growing firms. I include *Asset Growth* to control for the possibility that *Asset Growth* explains the positive

relations between *Unexpected Revenue per Employee* and *Fraud*, or *Forecast Attainment* and *Fraud*. Firm age and firm size are controlled in the matching procedure employed in my study.

3.4.2 Model for Hypotheses Testing

Model 33 is used to evaluate the hypotheses. More specifically, H1, H2 and H3 predict that α_1, α_2 and α_3, respectively, are positive and significant.

$$Fraud = \alpha_0 + \alpha_1 Total\ Discretionary\ Accruals + \alpha_2 Forecast\ Attainment + \alpha_3 Unexpected\ Revenue\ per\ Employee + \alpha_n control\ variables + \varepsilon, \quad (33)$$

where *Fraud* is a dependent dichotomous variable, equal to 1 if the firm was investigated by the SEC for fraud and otherwise 0, *Total Discretionary Accruals* is the total discretionary accruals in years t_{-1}, t_{-2} and t_{-3}, *Forecast Attainment* is a dummy variable, equal to 1 if analyst forecasts were met or exceeded and 0 otherwise, and *Unexpected Revenue per Employee* is the difference between a firm and its industry in the percentage change in revenue per employee from year t_{-1} to t_0. Five controls are used: *Current Discretionary Accruals*, *Sales to Assets*, *Auditor*, *CFO Change* and *Asset Growth*. Please refer to the previous section (Section 3.4.1.4) for descriptions of the control variables.

3.4.3 Data Sample

3.4.3.1 Experimental Sample

The fraudulent observations were located based on firms investigated by the SEC for fraud and reported in AAER from the 4th quarter of 1999 through 2005. I searched AAERs for explicit reference to section 10(b) and rule 10b-5, or descriptions of fraud. From this search a total of 745 potential observations were obtained (see Table 3.1). This initial selection was then reduced by eliminating: duplicates, financial companies, firms without the first fraud year specified in the SEC release, non-annual fraud, foreign corporations, AAERs focusing on auditors, not-for-profit organizations, and fraud related to registration statements, 10-KSB or IPO. Financial companies were, as is typically done, excluded from the sample as the rules and regulations governing financial firms are substantially different from other firms. An additional 75 fraud firms[15] from Beasley (1996) were added to the remaining 139 fraud firms, for a total of 214 fraud firms. Finally, 160 firms with missing data in Compustat for the fraud year or four prior years, Compact

[15] These fraud firms were kindly provided by Mark Beasley. Beasley (1996) collected the data from 348 AAERs released between 1982 and 1991 (67 of the 75 fraud firms) and from the Wall Street Journal Index caption of "Crime—White Collar Crime" between 1980 and 1991 (8 additional fraud firms).

Table 3.1
Sample Selection

Panel A: Fraud Firms	
Firms investigated by the SEC for fraudulent financial reporting from 4Q 1998 through 3Q 2005	745
Less: Financial companies	(35)
Less: Not annual (10-K) fraud	(116)
Less: Foreign companies	(9)
Less: Not-for-profit organizations	(10)
Less: Registration, 10-KSB and IPO related fraud	(78)
Less: Fraud year missing	(13)
Less: Duplicates	(287)
Remaining Fraud Observations	197
Add: Fraud firms from Beasley (1996)	75
Less: Not in Compustat or CompactD for first fraud year or four prior years or I/B/E/S for first fraud year	(218)
Usable Fraud Observations	54
Panel B: Non-Fraud Firms	
Firms in the same SIC industry as fraud firm in the year the fraud was committed (firms included in count once for each year matched to one or more fraud firms)	12,423
Less: Firms with missing data in fraud year or in four years prior to the fraud	(2,705)
Less: Firms not most similar in age and size to the fraud firms	(9,664)
Usable Non-Fraud Observations	54

D/SEC for the fraud year or three prior years, or I/B/E/S for the fraud year, were deleted for a total of 54 useable fraud firms. Seventy-four of the 75 companies provided by Beasley (1996) were part of the 160 deleted fraud observations. The fraud year for these 75 companies ranged from 1978 to 1990. Governance, analyst forecasts, or financial statement data were missing for these firms. The governance data, gathered from Compact D/SEC, were only available from 1988 and forward, and analyst forecast data, obtained from I/B/E/S, were only available from 1980 and forward, and were relatively sparse until 1995. Note that both governance and financial statement data were needed for the three years prior to the first fraud year.

Table 3.2 shows the industry distribution of the fraud firms by one-digit SIC groups. Manufacturing is the largest group, making up 35.19% of the sample, followed by Personal and Business Services (24.07%) and Wholesale and Retail (16.67%). This industry distribution is similar to distributions of prior fraud research (Beneish 1997).

The remaining 54 fraud firms were then matched with 54 non-fraud firms based on two-digit SIC code, firm age group and firm size, as measured by total assets in year t_0. Three age groups, over 10 years, five through 10 years, and four years were created so that a number of firms would be available for selection when matching on size. Note that the smallest firm age is four as

Table 3.2
Industry Distribution of Fraud Firm[a]

2-digit SIC	*Industry Description*	*Number of Firms*	*%*
10-19	Mining and Construction	0	0.00%
20-29	Commodity Production	6	11.11%
30-39	Manufacturing	19	35.19%
40-49	Transportation and Utilities	2	3.70%
50-59	Wholesale and Retail	9	16.67%
60-69	Financial Services (excl. 60-63)	0	0.00%
70-79	Personal and Business Services	13	24.07%
80-89	Health and Other Services	4	7.41%
99	Nonclassifiable Establishments	1	1.85%
		54	100.00%

[a] Table adapted from Beneish (1997), industry names are from the Standard Industrial Classification Manual (1987)

Compustat and Compact/D data were required for the fraud year and the three years leading up to the first fraud year. The matching was based on firm age before firm size based on Beneish's (1999) finding that matches based on age reduce the potential for omitted variable problems. The SEC typically targets young growth firms for investigation. Thus, an omitted variable problem can be introduced when such a firm is compared to other firms of similar size that are not young growth firms (Beneish 1999). For example, a young growth firm could have both high *Unexpected Revenue per Employee* and increased fraud likelihood. By matching based on age and size, Beneish (1999) found that differences in terms of age, growth and ownership structure between fraud and non-fraud firms were better controlled than when matched on only size, while both types of matches controlled for size, liquidity, leverage, profitability and cash flows. Because young firms are more likely to be growth firms the pair-wise matching should, at least partially, control for growth in addition to age (Beneish 1999). In addition to matching, I also include *Asset Growth* to more directly control for growth as not all high (low) growth firms are young (old).

For the 54 matched pairs, financial statement data for the first year of the fraud and each of the four years leading up to the first fraud year, were collected from Compustat. One-year-ahead analyst earnings per share forecasts and actual earnings per share in the fraud year were collected from I/B/E/S and matched to financial statement data collected from Compustat. Finally, *CFO Change* and percentage of outsiders on the board, collected for use in sensitive analysis, were collected from Compact D/SEC and manually from proxy statements.

3.4.3.2 Comparing Treatment and Control Samples

Table 3.3 contains descriptive statistics for the two samples. There was no statistical difference between fraud and non-fraud firms for median *Age* (p=0.347) or *Assets* (p=0.702).

Fraud firms were, however, more likely have high asset growth; 61% of the fraud firms versus 46% of non-fraud firms had high asset growth (p=0.062). Thus, the matching procedure effectively matched fraud firms with similar non-fraud firms in terms of firm age and size. However, the matching procedure was not as effective at eliminating differences in growth, and I, therefore, include the variable *Asset Growth* to control for any possible difference in growth between fraud and non-fraud firms.

There was no statistical difference[16] between fraud and non-fraud firms for median *Total Discretionary Accruals* (p=0.125), *Unexpected Revenue per Employee* (p=0.125), *Current Discretionary Accruals* (p=0.222), *Sales to Assets* (p=0.222), *Auditor* (p=1.000) and *CFO Change* (p=0.110). *Forecast Attainment* was significant (p=0.010), thus fraud firms were more likely than non-fraud firms to meet or exceed analyst forecasts. Fifty-two percent of the fraud firms had earnings equal to or greater than consensus forecasts as opposed to 30% of non-fraud firms. These univariate results provide initial support for H2, *Forecast Attainment* (p=0.010), but not for H1, *Total Discretionary Accruals* (p=0.125), and H3, *Unexpected Revenue per Employee* (p=0.125).

The correlation matrix in Table 3.4 shows positive significant (alpha<0.05) correlations between *Fraud* and three independent variables: *Total Discretionary Accruals* (r=0.17), *Forecast Attainment* (r=0.23) and *Unexpected Revenue per Employee* (r=0.16); and marginally significant relation between *Asset Growth* and *Fraud* (r=0.15). Firms are seemingly more likely to have committed fraud if they have high *Total Discretionary Accruals*, meet or exceed analyst forecasts, have high *Unexpected Revenue per Employee* or have high *Asset Growth*.

3.5. Results

3.5.1 Hypotheses Testing

The dependent variable, whether a firm has committed fraud, is dichotomous; therefore logistic regression is used to evaluate the model. The primary assumptions for logistic regression are as follows. (1) Binomial Distribution - the dependent variable must follow a binomial distribution. As there are only two potential outcomes, fraud and not fraud, this assumption is satisfied. (2) Bernoulli Distribution - the dependent variable classes must be mutually exclusive. This assumption is satisfied, as financial statements are either fraudulent or non-fraudulent, and

[16] One-tailed tests reported for estimates in the direction predicted, all other two-tailed, unless noted otherwise.

Table 3.3
Sample Descriptive Statistics for Study Variables

	Fraud Observations (n=54)					*Non-Fraud Observations (n=54)*					*Diff*
Variables[b]	*Mean*	*Std*	*Min*	*Median*	*Max*	*Mean*	*Std*	*Min*	*Median*	*Max*	*p-stat*[a]
Total Discretionary Accruals	0.15	0.51	-1.25	0.07	2.65	0.02	0.23	-0.58	0.03	0.50	0.125
Forecast Attainment	0.52	0.50	0.00	1.00	1.00	0.30	0.46	0.00	0.00	1.00	0.010
Unexpected Revenue per Employee	0.04	0.38	-1.12	0.00	1.29	-0.07	0.26	-0.69	-0.02	1.07	0.125
Current Discretionary Accruals	0.00	0.20	-1.03	0.01	0.67	0.00	0.13	-0.32	0.00	0.53	0.222
Sales to Assets	1.16	0.64	0.09	1.09	3.42	1.24	0.76	0.30	1.16	4.13	0.222
Auditor	0.96	0.19	0.00	1.00	1.00	0.96	0.19	0.00	1.00	1.00	1.000
Asset Growth	0.61	0.49	0.00	1.00	1.00	0.46	0.50	0.00	0.00	1.00	0.062
CFO Change	0.15	0.36	0.00	0.00	1.00	0.07	0.26	0.00	0.00	1.00	0.110
Firm Age	15.3	10.1	4.00	13.0	33.0	11.1	5.76	4.00	11.5	22.0	0.347
Assets	3254	6993	21.8	386	33381	2595	5802	25.83	361	31749	0.702

[a] Median χ^2 pair-wise comparison between fraud and non-fraud sample for continuous variables, Pearson χ^2 for dichotomous variables. One-tailed tests reported for estimates in the direction predicted, all other two-tailed.

[b] ***Total Discretionary Accruals***$_{j,t}$, is the total amount of discretionary accruals deflated by assets in the beginning of the year in the three years leading up to the fraud year. Discretionary accruals in year t_0 is estimated using the extended version of the modified Jones model (Jones 1991; Dechow et al. 1995; Kasznik 1999). Discretionary accruals $DA_{j,t}$ is calculated as estimated nondiscretionary accruals minus total accruals. Total accruals is income before extraordinary items (#18) minus cash flow from operations (#308). To obtain nondiscretionary accruals, $NDA_{j,t}$, for firm j in year t_0 regression parameters are first estimated in cross section for all firms in the same major industry group J (two-digit sic): $TA_{j,t} = \alpha_0 / A_{j,t-1} + \alpha_1(\Delta REV_{j,t} - \Delta REC_{j,t}) + \alpha_2 PPE_{j,t} + \alpha_3 \Delta CFO_{j,t}$. These parameter estimates are then used to derive estimated nondiscretionary accruals: $\hat{NDA}_{j,t} = \hat{\alpha}_{0,j} + \hat{\alpha}_{1,j}(\Delta REV_{j,t} - \Delta REC_{j,t}) + \hat{\alpha}_{2,j} PPE_{j,t} + \hat{\alpha}_{3,j} \Delta CFO_{j,t}$, where $\Delta REV_{j,t}$ is the change in revenue (#12), $\Delta REC_{j,t}$ is the change in receivables (#2) and $\Delta CFO_{j,t}$ is the change in cash flow from operations from time t_{-1} to t_0; and $PPE_{j,t}$ is gross property, plant and equipment (#8) at time t_0. All values are deflated by $A_{j,t-1}$, firm j's assets (#6) at time t_{-1}. ***Forecast Attainment*** is a dummy variable, equal to 1 if analyst forecast were met or exceeded and 0 otherwise (I/B/E/S). ***Unexpected Revenue per Employee*** for firm j in industry J is the difference between the % change in revenue per employee, RE=total sales (#12) divided by the number of employees (#29), of j and the % change in revenue per employee of J: *Unexpected Revenue per Employee* = $(RE_{jt} - RE_{jt-1})/RE_{jt-1} - (RE_{Jt} - RE_{Jt-1})/RE_{Jt-1}$. ***Current Discretionary Accruals*** is the discretionary accruals in year t_0, see definition in Total Discretionary Accruals. ***Sales to Assets*** = net sales / assets. ***Auditor*** is a dummy variable equal to 1 if auditor was a Big 4 audit firm (#149) and 0 otherwise. ***Asset Growth*** is a dummy variable equal to 1 if total assets exceeds 110% of the previous year's value and 0 otherwise. ***CFO Change*** is a dummy variable equal to 1 if CFO has changed in the three years leading up to the first fraud year and 0 otherwise. ***Firm Age*** is the number of years between t_0 and the first year data are reported for the company in Compustat. ***Assets*** is total assets of firm j.

Table 3.4
Pearson and Spearman Correlations[a] for Study Variables[b]

	Fraud	*Total Discretionary Accruals*	*Forecast Attainment*	*Unexpected Revenue Per Employee*	*Current Discretionary Accruals*	*Sales to Assets*	*Auditor*	*Asset Growth*	*CFO Change*
Fraud	1.00	0.14	0.23	0.15	0.04	-0.06	0.00	0.15	0.12
		(0.14)	(0.02)	(0.11)	(0.69)	(0.57)	(1.00)	(0.13)	(0.23)
Total Discretionary Accruals	0.17	1.00	-0.05	-0.02	0.28	-0.01	-0.12	-0.08	-0.07
	(0.08)		(0.59)	(0.82)	(0.00)	(0.90)	(0.23)	(0.43)	(0.46)
Forecast Attainment	0.23	0.02	1.00	-0.11	-0.06	0.09	0.06	0.09	-0.05
	(0.02)	(0.86)		(0.25)	(0.55)	(0.35)	(0.52)	(0.36)	(0.58)
Unexpected Revenue per Employee	0.16	-0.01	-0.08	1.00	-0.05	0.13	-0.11	-0.03	0.06
	(0.10)	(0.93)	(0.43)		(0.64)	(0.17)	(0.24)	(0.76)	(0.55)
Current Discretionary Accruals	0.01	0.34	-0.02	-0.01	1.00	0.08	-0.01	0.13	0.02
	(0.90)	(0.00)	(0.82)	(0.92)		(0.40)	(0.92)	(0.18)	(0.82)
Sales to Assets	-0.06	0.06	0.04	0.05	0.03	1.00	-0.03	-0.09	0.03
	(0.54)	(0.52)	(0.68)	(0.64)	(0.75)		(0.77)	(0.38)	(0.77)
Auditor	0.00	-0.04	0.06	-0.12	0.01	-0.03	1.00	0.11	-0.24
	(1.00)	(0.72)	(0.52)	(0.21)	(0.96)	(0.75)		(0.25)	(0.01)
Asset Growth	0.15	0.03	0.09	-0.09	0.10	-0.02	0.11	1.00	-0.09
	(0.13)	(0.79)	(0.36)	(0.38)	(0.32)	(0.84)	(0.25)		(0.38)
CFO Change	0.12	-0.13	-0.05	0.07	0.02	0.01	-0.24	-0.09	1.00
	(0.23)	(0.19)	(0.58)	(0.49)	(0.81)	(0.94)	(0.01)	(0.38)	

[a] Pearson correlations are below and Spearman correlations are above the diagonal. Two-tailed p-values reported within parentheses.
[b] Please refer to footnotes in Table 3.3 for variable definitions.

cannot be both fraudulent and non-fraudulent. (3) Independent Observations - the observations are independent as the order of the observations in the sample is irrelevant, i.e., I do not have time-series data, and there is only one observation per firm.

The model estimates in logistic regression can be sensitive to outliers and multicollinearity. Although the firms were matched based on major industry, age and size, the descriptive statistics indicate the possibility of outliers in the sample, for example the median *Assets* for fraud (non-fraud) firms is $386 ($361) as compared to a mean of $3,254 ($2,595), with a minimum of $22 ($26) and a maximum of $33,381 ($31,749).

To evaluate the impact of potential outliers I used Pearson residuals. One observation had Pearson residuals above 2. For this observation all continuous measures were truncated at a plus minus two standard deviations.[17] I did not find any evidence of multicollinearity; the highest Variance Inflation Factor (VIF) was 1.11, which is relatively low, and the highest condition index among the continuous variables was 6.00. Based on these findings I did not discard any variables due to multicollinearity.

The results in Table 3.5 show that: (1) *Total Discretionary Accruals* is positively related to *Fraud* (p=0.009); (2) *Forecast Attainment* is positively related to *Fraud* (p=0.004); and (3) *Unexpected Revenue per Employee* is positively related to *Fraud* (p=0.016). The positive relation between *total prior discretionary accruals* and *fraud,* supports H1, which states that total prior discretionary accruals is positively related to the likelihood of Fraud (p=0.009). H2, hypothesizing that firms that meet or exceed analyst forecasts are more likely to have committed fraud than firms that fail to meet analyst forecasts, is supported by the positive relation between *Forecast Attainment* and *Fraud* (p=0.004). Finally, the results, showing a positive relation between *Unexpected Revenue per Employee* and *Fraud* (p=0.016), provide support for H3, stating that unexpected labor productivity is positively related to the likelihood of fraud. Thus, H1, H2 and H3 are supported.

The logit estimates for *Total Discretionary Accruals*, *Forecast Attainment* and *Unexpected Revenue per Employee* are 1.641, 0.573 and 1.445, respectively. Thus, as *Total Discretionary*

---

[17] I also examined the hypotheses after: (1) deleting the outlier from the sample; (2) deleting the outlier and the outlier's matched non-fraud firm from the sample; and (3) including the outlier in the model without truncating it. The results obtained from these three sensitivity analyses were equivalent to the reported results.

Table 3.5
The Effect of Total Discretionary Accruals, Forecast Attainment and Unexpected Revenue per Employee on Financial Statement Fraud Likelihood
Logistic Regression Results[a]

Variable[b]	*Prediction*	*Estimate*	*Standard Error*	χ^2	*prob*>χ^2
Intercept	(?)	0.065	0.746	0.01	0.931
Tests for Hypotheses 1, 2 and 3					
Total Discretionary Accruals	(+)	1.641	0.789	5.66	0.009
Forecast Attainment	(+)	0.573	0.225	6.86	0.004
Unexpected Revenue per Employee	(+)	1.445	0.713	4.55	0.016
Control Variables					
Current Discretionary Accruals	(+)	-0.875	1.554	0.34	0.560
Sales to Assets	(-)	-0.327	0.319	1.09	0.148
Auditor	(-)	0.192	0.629	0.09	0.759
Asset Growth	(+)	0.359	0.217	2.80	0.047
CFO Change	(+)	1.304	0.726	3.49	0.031
Pseudo R^2		0.135			
χ^2-test of model fit		20.15 (p=0.010)			
n		108			

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; please refer to footnotes in Table 3.3 for variable definitions.

Accruals increases by one unit, the odds of fraud increase by a factor of 5.16[18], holding the other variables constant. The odds of fraud are higher by a factor of 1.77 for companies that meet or exceed analyst forecasts than for companies that do not meet or exceed analyst forecasts, holding the other variables constant. As *Unexpected Revenue per Employee* increases by one unit, the odds of fraud increase by a factor of 4.24, holding the other variables constant. In terms of the control variables, the positive relation between *CFO Change* and *Fraud* (p=0.031) indicate that firms that have had CFO turnover in the three years leading up to the first fraud year are more likely to have committed fraud than firms that have had the same CFO during this period. *Asset Growth* was also positively related to *Fraud* (p=0.047), which indicates that high growth firms are more likely to have committed fraud than firms that are not high growth firms. *Sales to Assets* (p=0.148), *Current Discretionary Accruals* (p=0.560) and *Auditor* (p=0.759) were insignificant. It

[18] Odds ratios of fraud are calculated by applying the exponential function to the logit estimates, i.e., the base of the natural logarithm is raised by the logit estimates. For example, 5.16=*e*^1.641, where *e*≈2.718.

is interesting to note that *Total Discretionary Accruals* is positive and significant and that *Current Discretionary Accruals* is insignificant. This indicates, tentatively, that the relation between *Total Discretionary Accruals* and fraud is not driven by management character, i.e., poor set of ethical values, explaining both earnings management and current year fraud.

While *Sales to Assets* is insignificant (p=0.148) the direction is negative as expected. When removing *Sales to Assets* from the model, *Unexpected Revenue per Employee* remains positive and significant (p=0.019). When removing *Unexpected Revenue per Employee* from the model, *Sales to Assets* remains negative an insignificant (p=0.186). Thus, it appears, as discussed earlier, that *Unexpected Revenue per Employee* and *Sales to Assets* capture different aspects of productivity. The positive and significant result for *Unexpected Revenue per Employee* indicates that *Unexpected Revenue per Employee* captures fraudulent revenue manipulation. While the negative, though insignificant, result for *Sales to Assets* provides some tentative indication of *Sales to Assets* capturing the relatively poor actual productivity of fraud firms, which puts pressure on these firms to commit fraud.

3.6. Additional Analyses

I next examine the sensitivity of the reported results (Section 3.6.1) and the appropriateness of variable design choices (Section 3.6.2). The sensitivity of the relation between *Total Discretionary Accruals* and *Fraud* to the use of a different discretionary accruals measure and to the inclusion of two real activities manipulation measures is examined in Section 3.6.1.1 and Section 3.6.1.2, respectively. Section 3.6.1.3 evaluates the sensitive of the results to the inclusion of additional control variables, while Section 3.6.1.4 evaluates the sensitivity of the results to the exclusion of industries.

The labor productivity measures used in *Unexpected Revenue per Employee* is compared to two similar measures that have been proposed in concurrent research in Section 3.6.2.1. Section 3.6.2.2 compares the aggregation period used to calculate *Total Discretionary Accruals* to two shorter periods. Section 3.6.2.3 provides an evaluation of the appropriateness of using the earliest one-year ahead analyst consensus forecast for calculating *Forecast Attainment*.

3.6.1 Sensitivity Analyses

3.6.1.1 Discretionary Accruals

As a sensitivity analysis of the relation between *Total Discretionary Accruals* and *Fraud* I used an alternative cash flow statement based measure of discretionary accruals from Hribar and Collins (2002). This measure, *Total Cash Based Discretionary Accruals* calculates total accruals

as net income (#172) minus cash flow from operations. Discretionary accruals and non-discretionary accruals are estimated following equations (27), (28) and (29), respectively. The results (see Table 3.6) were qualitatively the same, more specifically *Total Discretionary Accruals* derived using this alternative accruals measure remains positively (p=0.014) related to *Fraud*.

Table 3.6
Alternative Total Discretionary Accruals Measure
Logistic Regression Results[a]

Variable[b]	*Prediction*	*Estimate*	*Standard Error*	χ^2	*prob*>χ^2
Intercept	(?)	0.081	0.745	0.010	0.913
Cash Based Discretionary Accruals	**(+)**	**1.527**	**0.785**	**4.827**	**0.014**
Forecast Attainment	(+)	0.565	0.224	6.728	0.005
Unexpected Revenue per Employee	(+)	1.422	0.710	4.443	0.018
Current Discretionary Accruals	(+)	-0.761	1.485	0.278	0.598
Sales to Assets	(-)	-0.330	0.319	1.120	0.145
Auditor	(-)	0.187	0.628	0.090	0.765
Asset Growth	(+)	0.351	0.215	2.706	0.050
CFO Change	(+)	1.345	0.738	3.609	0.029
Pseudo R^2		0.129			
χ^2-test of model fit		19.32 (p=0.013)			
n		108			

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***Total Cash Based Discretionary Accruals*** is calculated using a cash based measure of total accruals, net income (#172) minus total cash from operations. This cash based measure of TA is then used to estimate discretionary accruals and non-discretionary accruals following equations (27), (28) and (29), respectively. Please refer to footnotes in table 3.3 for definitions of all other variables.

3.6.1.2 Real Activities Manipulation

Research has shown that in addition to using discretionary accruals to manage earnings, managers use real activities manipulation (Roychowdhury 2006). Real activities manipulation could, conceptually, be positively related to both Total Discretionary Accruals and Fraud if real activities manipulation is captured in discretionary accruals and if this manipulation is subsequently detected or leads to fraud. Thus, I evaluate if real activities manipulation is an omitted variable. This evaluation also provides some insight into whether the earnings reversal hypothesis can be applied to real activities manipulation.

I add two real activities manipulation measures to model 33, Abnormal Production Costs and Abnormal Discretionary Expenditures (Roychowdhury 2006), each summed over the three years

leading up to the first fraud year. Production costs are the sum of cost of goods sold and change in inventory. Abnormal Production Costs is the residual from a regression model estimating normal production costs using current sales, change in sales between t_0 and t_{-1} and change in sales between t_{-1} and t_{-2}, all variables are deflated by beginning of the period assets. Discretionary Expenditures are the sum of advertising expenses, R&D expenses, and selling, general and administrative expense. Abnormal Discretionary Expenditures is the residual from a regression model estimating normal discretionary expenditures using sales in t_{-1}, all variables are deflated by t_{-1} assets. Please refer to Roychowdhury (2006) for details on how to compute these measures.

The results in Table 3.7 show that Total Discretionary Accruals remain positive and significant (p=0.002), and that both Abnormal Discretionary Expenditures (p=0.027) and Abnormal Production Costs (p=0.033) are positive and significant. The positive relation between Abnormal Discretionary Expenditures and fraud was not expected[19]. Based on the idea that managers that manipulate earnings using real activities will reduce discretionary expenditures, I was expecting a negative relation between Abnormal Discretionary Expenditures in prior years and Fraud. Managers will over time run out of ways to manipulate earnings using real activities manipulation just like they do when they manipulate earnings using discretionary accruals. For example, if discretionary expenditures, such as research and development, are reduced to increase earnings then further reductions will eventually become difficult as there are limits to how much these real activities can be manipulated. Furthermore, by manipulating earnings using real activities manipulation the firm does not operate at an optimal level, at least not what management would consider optimal, and the firm becomes less likely to perform well in subsequent years. The deterioration in performance will pressure management to increase revenue and as the flexibility to manipulate earnings using real activities manipulation is reduced through earlier manipulation, it becomes more likely that the manager will commit fraud to increase revenue. A potential explanation for the unexpected positive relation could be that abnormally high discretionary expenditures in prior years indicate inefficient use of resources in prior years that lead to poor performance in subsequent years, and this poor performance puts pressure on management to commit fraud to manipulate earnings. The relation between Abnormal Production Costs is in the direction that the earnings reversal hypothesis would predict, indicating that as managers manipulate earnings using real activities in prior years they are more likely to commit

[19] I did not find any evidence of multicollinearity; the highest VIF was 1.48, which is relatively low.

Table 3.7
Total Discretionary Accruals, Real Activities Manipulation and Financial Statement Fraud Logistic Regression Results[a]

Variable[b]	*Prediction*	*Estimate*	*Standard Error*	χ^2	*prob*>χ^2
Intercept	(?)	0.331	0.832	0.160	0.691
Total Discretionary Accruals	**(+)**	**2.286**	**0.870**	**8.766**	**0.002**
Forecast Attainment	(+)	0.527	0.247	4.793	0.014
Unexpected Revenue per Employee	(+)	1.300	0.744	3.262	0.035
Abnormal Production Costs	**(+)**	**0.802**	**0.437**	**3.361**	**0.033**
Abnormal Discretionary Expenditures	**(-)**	**0.885**	**0.431**	**4.888**	**0.027**
Current Discretionary Accruals	(+)	-0.433	1.660	0.070	0.791
Sales to Assets	(-)	-0.435	0.357	1.574	0.105
Auditor	(-)	0.077	0.697	0.012	0.912
Asset Growth	(+)	0.465	0.237	3.993	0.023
CFO Change	(+)	1.300	0.745	3.274	0.035
Pseudo R^2		0.183			
χ^2-test of model fit		25.14 (p=0.005)			
n		100[c]			

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***Abnormal Production Costs*** is the residual from a regression model estimating normal production costs using production costs (defined as the sum of cost of goods sold and change in inventory), current sales, change in sales between t0 and t-1 and change in sales between t-1 and t-2, all variables are deflated by t-1 assets. ***Abnormal Discretionary Expenditures*** is the residual from a regression model estimating normal discretionary expenditures using discretionary expenditures (defined as the sum of advertising expenses, R&D expenses, and selling, general and administrative expense) and sales in t-1, all variables are deflated by t-1 assets. Please refer to footnotes in table 3.3 for definitions of all other variables.

[c] The sample was reduced by four matched pairs because discretionary expenditures data were not available for all firms.

fraud in subsequent years. Finally, note that the relation between Total Discretionary Accruals and Fraud is robust to the inclusion of the two real activities manipulation measures.

3.6.1.3 Additional Control Variables

As a sensitivity analysis I also added the other five variables found to be significant predictors of fraud in Fanning and Cogger (1998), and total assets, total sales and sales growth, as controls. The five additional variables in Fanning and Cogger 1998 are as follows. (1) Accounts Receivable Growth measured as a dummy variable equal to one if accounts receivable exceeds 110% of the previous year's value and zero otherwise. Given that accounts receivables often increases as a result of fraud, a positive relation is expected between Accounts Receivable Growth and Fraud.

Note that this information is captured in Current Discretionary Accruals. (2) Debt to Equity is the ratio of debt to equity (leverage) and is expected to be positively related to Fraud given that higher levels of leverage put more pressure on management to meet debt covenants. (3) Gross Margin Percentage is a dummy variable that is one if the gross margin percentage exceeds 110% of the previous year's value and zero otherwise; assuming that the gross margin percentage improves as a result of fraud, a positive relation is expected between Gross Margin Percentage and Fraud. (4) LIFO is a dummy variable that is one if the last-in-first-out inventory method is used and zero otherwise. Given that prices were generally rising during the examined period, LIFO is expected to be negatively related to Fraud as the usage of last-in-first-out, relative to first-in-first-out, decreases earnings when prices are rising. (5) The Percentage of Executive Directors on the board of directors is expected to be positively related to Fraud as the independence and monitoring effectiveness of the board of directors is reduced by including company executives on the board. The other additional control variables in the sensitivity analysis are defined as follows. (1) Assets is expected to be negatively related to Fraud given that small growth firms are more likely to be investigated by the SEC (Beneish 1999) than larger slower growing firms. (2) Sales is expected to be negatively related to Fraud given that small growth firms are more likely to be investigated by the SEC (Beneish 1999) than larger, slower growing firms. (3) Sales Growth is a dummy variable that measures whether sales exceeds 110% of the previous year's sales and is expected to be positively related to Fraud given that small growth firms are more likely to be investigated by the SEC (Beneish 1999) than larger, slower growing firms. The results (see Table 3.8) obtained from the sensitivity analysis were equivalent to the reported results with the exception that Asset Growth became insignificant (p=0.463). Note that both Sales Growth (p=0.157) and Asset Growth are measures of firm growth. Debt to Equity was additionally positive and marginally significant at (p=0.062).

3.6.1.4 Industry Clustering

I next investigate potential industry differences in the effect of Total Discretionary Accruals, Forecast Attainment and Unexpected Revenue per Employee on Fraud. I first add dummy variables for the seven one-digit SIC industries represented in the sample to model 33. The results (Table 3.9) show insignificant industry dummies (p>0.398) and that the interpretation of the other variables does not change when the industry dummy's are included in the main model, except for that Sales to Assets becomes marginally significant (p=0.090).

Due to sample size limitations I cannot, however, test the significance and direction of Total Discretionary Accruals, Forecast Attainment and Unexpected Revenue per Employee within each industry. I instead create seven subsamples by excluding, from each subsample, all firms

belonging to one of the industries. Using model 33, I then examine the significance and direction of Total Discretionary Accruals, Forecast Attainment and Unexpected Revenue per Employee within each subsample. The results (untabulated) show that Total Discretionary Accruals remains positive and significant ($p<0.028$) in six subsamples, and positive and marginally significant ($p=0.054$) when Personal and Business Services firms are excluded. Forecast Attainment remains positive and significant ($p<0.008$) in six subsamples, and positive but insignificant ($p=0.106$) when Wholesale and Retail firms are excluded. Unexpected Revenue per Employee remains positive and significant ($p<0.034$) in six subsamples, and positive but insignificant ($p=0.161$)

Table 3.8
Additional Control Variables
Logistic Regression Results[a]

Variable[b]	*Prediction*	*Estimate*	*Standard Error*	χ^2	*prob*>χ^2
Intercept	(?)	-0.273	1.148	0.060	0.812
Total Discretionary Accruals	(+)	1.619	0.817	5.023	0.013
Forecast Attainment	(+)	0.486	0.251	3.877	0.025
Unexpected Revenue per Employee	(+)	1.226	0.725	2.993	0.042
Current Discretionary Accruals	(+)	-0.851	1.599	0.304	0.581
Sales to Assets	(-)	-0.334	0.385	0.772	0.190
Auditor	(-)	0.170	0.646	0.071	0.790
Asset Growth	(+)	0.032	0.339	0.009	0.463
CFO Change	(+)	1.336	0.777	3.149	0.038
Account Receivable Growth	(+)	0.279	0.274	1.040	0.154
Debt to Equity	(+)	0.135	0.109	2.357	0.062
Gross Margin Percentage	(+)	0.259	0.434	0.356	0.275
LIFO	(-)	0.013	0.495	0.001	0.979
Percentage of Executive Directors	(+)	1.200	1.276	0.889	0.173
Assets	(-)	0.000	0.000	0.001	0.490
Sales	(-)	0.000	000	0.008	0.929
Sales Growth	(+)	0.320	0.319	1.013	0.157
Pseudo R^2		0.171			
χ^2-test of model fit		25.56 (p=0.061)			
n		108			

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***Account Receivable Growth*** is a dummy variable equal to one if accounts receivable exceeds 110% of the previous year's value and zero otherwise. ***Debt to Equity*** is the ratio of debt to equity. ***Gross Margin Percentage*** is a dummy variable that is one if the gross margin percentage exceeds 110% of the previous year's value and zero otherwise. ***LIFO*** is a dummy variable that is one if the last in first out inventory method is used and zero otherwise. The ***Percentage of Executive Directors*** is the percentage of all directors on the board of directors that are company executives. ***Sales*** is total sales. ***Sales Growth*** is a dummy variable that measures whether assets exceeds 110% of the previous year's assets. Please refer to footnotes in table 3.3 for definitions of all other variables.

when Manufacturing firms are excluded. Note that the results only changed to marginally significant or insignificant in three out of 21 tests, and that the significance levels only dropped when one of the three industries with the largest number of observations was removed from the sample, i.e., when the sample size was the smallest. Thus, the effects of Total Discretionary Accruals, Forecast Attainment and Unexpected Revenue per Employee on Fraud appear to be relatively robust for different industries.

3.6.2 Alternative Measure Design

3.6.2.1 Revenue Fraud

The measure difference between revenue growth and employee growth (DiffEmp) introduced in a concurrent working paper (Brazel et al. 2007) is similar to percentage change in revenue per employee (%ΔRE), which is used to derive Unexpected Revenue per Employee. However, the

Table 3.9
Major Industry, Total Discretionary Accruals,
Forecast Attainment and Unexpected Revenue per Employee
Logistic Regression Results[a]

Variable[b]	*Prediction*	*Estimate*	*Standard Error*	χ^2	*prob*>χ^2
Intercept	(?)	0.514	0.875	0.350	0.557
Total Discretionary Accruals	(+)	1.576	0.773	5.532	0.009
Forecast Attainment	(+)	0.612	0.232	7.453	0.003
Unexpected Revenue per Employee	(+)	1.781	0.775	5.918	0.008
Current Discretionary Accruals	(+)	-1.210	1.664	0.587	0.444
Sales to Assets	(-)	-0.530	0.409	1.793	0.090
Auditor	(-)	0.047	0.670	0.005	0.944
Asset Growth	(+)	0.405	0.227	3.297	0.035
CFO Change	(+)	1.318	0.739	3.437	0.032
Major Industry = 2	?	0.221	0.628	0.120	0.724
Major Industry = 3	?	-0.226	0.447	0.250	0.614
Major Industry = 4	?	-0.122	0.994	0.020	0.902
Major Industry = 5	?	0.398	0.677	0.350	0.556
Major Industry = 7	?	-0.449	0.531	0.710	0.398
Major Industry = 8	?	0.578	0.753	0.590	0.443
Pseudo R^2		0.1475			
χ^2-test of model fit		22.08 (p=0.077)			
n		108			

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***Major Industry*** are dummy variables for the seven one-digit SIC code industries represented in the sample. Please refer to footnotes in table 3.3 for definitions of all other variables.

conceptual basis for using the measures differs, leading to differences in definitions and what is actually measured. DiffEmp, defined as $\frac{rev_t - rev_{t-1}}{rev_{t-1}} - \frac{emp_t - emp_{t-1}}{emp_{t-1}}$, is based on the idea that nonfinancial measures that are highly correlated to performance and that at the same time are difficult to manipulate can be used to evaluate the reasonableness of changes in firm performance. Unexpected Revenue per Employee started with the idea that revenue manipulation is difficult to detect in the revenue account, as revenue varies for reasons other than fraud, and that some of this variation can be removed by deflating revenue by a production process input variable. The number of employees was selected as the deflator rather than assets as the number of employees is not impacted by revenue fraud when the fraudulent revenue is recorded following the double-entry system. The primary difference between the two measures is how they adjust revenue growth using employee growth. Note that both measures are based on the idea that there is a relatively constant relation between the number of employees and revenue. Thus, if the number of employees grows by 10%, both measures assume that revenue should also grow by 10%, and vice versa. The two measures, however, differ in how the difference between expected and actual revenue is measured. Diffemp is increasing in the absolute difference between expected revenue growth and actual revenue growth, while $\%\Delta RE$ is increasing in the ratio of expected revenue growth to actual revenue growth. To clarify, take company A that is growing at a rate of 10% as indicated by the number of employees growing by a rate of 10% and company B that is growing at a rate of 100%. Further assume that both companies fraudulently increase revenue by 30% over what could be expected given prior revenue, prior number of employees and current number of employees. Thus, in absolute terms, company B manipulated revenue more than company A, but as a percentage of expected revenue there was no difference between the two firms. In this situation Diffemp is 0.33 for company A and 0.6 for company B, while $\%\Delta RE$ is 0.3 for both company A and company B. Assuming a constant percentage manipulation over expected revenue Diffemp is, while $\%\Delta RE$ is not, increasing in the percentage change in the number of employees. Based on this discussion I expect that $\%\Delta RE$ models will provide better fit and predictive ability than Diffemp models, when the models do not control for firm growth, and that their performance will be similar when the models control for firm growth using employee growth.

Using model 33 without a control for firm growth (removing Asset Growth from the model) and replacing Unexpected Revenue per Employee with $\%\Delta RE$, $\%\Delta RE$ is in the expected direction and significant (p=0.016), see column 1 in Table 3.10. When replacing Unexpected Revenue per Employee with Diffemp, Diffemp is in the expected direction, but only marginally significant (p=0.067), see column 2 in Table 3.10. Using the same models but adding a control for employee growth, $\%\Delta RE$ is in the expected direction and significant (p=0.003) and Diffemp is in the expected direction and significant (p=0.003), see Table 3.10 columns 3 and 4, respectively. These results support the analytical analysis by indicating that when controlling for employee growth the two measures are equivalent, but that without a control for employee growth, the percentage change in revenue per employee $\%\Delta RE$ is a better predictor of fraud than Diffemp.

Table 3.10
Comparison of %ΔRE and Diffemp
Logistic Regression Results[a]

		(1)		(2)		(3)		(4)	
Variable[b]	*Prediction*	*Estim-ate*	*prob >χ^2*	*Estim-ate*	*prob> χ^2*	*Estim-ate*	*prob >χ^2*	*Estim-ate*	*prob >χ^2*
Intercept	(?)	0.572	0.448	0.635	0.384	0.385	0.614	0.419	0.574
Total Discretionary Accruals	(+)	1.551	0.011	1.524	0.011	1.519	0.016	1.556	0.014
Forecast Attainment	(+)	0.575	0.004	0.575	0.004	0.518	0.009	0.537	0.007
%ΔRE	**(+)**	**1.690**	**0.016**			**2.552**	**0.003**		
Diffemp	**(+)**			**0.882**	**0.067**			**2.357**	**0.003**
Current Discretionary Accruals	(+)	-0.743	0.614	-0.690	0.634	-0.818	0.624	-1.11	0.529
Sales to Assets	(-)	-0.333	0.146	-0.312	0.160	-0.322	0.151	-0.31	0.159
Auditor	(-)	0.218	0.722	0.200	0.737	0.178	0.773	0.149	0.804
CFO Change	(+)	0.597	0.041	0.589	0.041	0.644	0.032	0.669	0.028
Employee Growth	(+)					1.282	0.029	1.989	0.008
Pseudo R^2		0.119		0.103		0.1426		0.1417	
χ^2-test of model fit		17.74 (p=0.013)		15.41 (p=0.031)		21.35 (p=0.006)		21.22 (p=0.007)	
n		108		108		108		108	

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***%ΔRE*** is defined as $(\frac{rev_t}{emp_t} - \frac{rev_{t-1}}{emp_{t-1}}) / (\frac{rev_{t-1}}{emp_{t-1}})$, ***DiffEmp***, is defined as $\frac{rev_t - rev_{t-1}}{rev_{t-1}} - \frac{emp_t - emp_{t-1}}{emp_{t-1}}$, and ***Employee Growth*** is defined as $\frac{emp_t - emp_{t-1}}{emp_{t-1}}$. Please refer to footnotes in table 3.3 for definitions of all other variables.

To further substantiate this claim I perform a nested F-test where the fit of a reduced model is compared to full models. The reduced model includes all variables reported in Table 3.10 except for $\%\Delta RE$ or Diffemp, see column 3 in Table 3.11. The full models include $\%\Delta RE$ or Diffemp, see columns 1 and 2 in Table 3.11, respectively. The log likelihood values of the full models are then compared to the log likelihood value of the reduced model. The results show that only the $\%\Delta RE$ model (column 1, Table 3.11) significantly (p=0.032) improves the fit of the reduced model.

I also evaluate the predictive ability of the $\%\Delta RE$ model to the predictive ability of the Diffemp model. The prediction errors of the $\%\Delta RE$ model are significantly lower than the

Table 3.11
Comparison of Model Fit and Predictive Ability of *%ΔRE* and *Diffemp*
Logistic Regression Results[a]

		(1)		(2)		(3)	
Variable[b]	*Prediction*	*Esti-mate*	*prob>χ^2*	*Esti-mate*	*prob>χ^2*	*Esti-mate*	*prob>χ^2*
Intercept	(?)	0.572	0.448	0.635	0.384	0.632	0.367
Total Discretionary Accruals	(+)	1.551	0.011	1.524	0.011	1.443	0.013
Forecast Attainment	(+)	0.575	0.004	0.575	0.004	0.530	0.006
%ΔRE	**(+)**	**1.690**	**0.016**				
Diffemp	**(+)**			**0.882**	**0.067**		
Current Discretionary Accruals	(+)	-0.743	0.614	-0.690	0.634	-0.557	0.689
Sales to Assets	(-)	-0.333	0.146	-0.312	0.160	-0.271	0.182
Auditor	(-)	0.218	0.722	0.200	0.737	0.171	0.767
CFO Change	(+)	0.597	0.041	0.589	0.041	0.591	0.040
Pseudo R^2		0.119		0.103		0.0879	
χ^2-test of model fit		17.74 (p=0.013)		15.41 (p=0.031)		13.16 (p=0.041)	
Diffχ^2_{full} -$\chi^2_{reduced}$		**4.58 (p=0.032)**		**2.29 (p=0.134)**			
Predictive Ability (Mean Diff.)				**p=0.045**			
Predictive Ability (Wilcoxon)				**z=0.006**			
n		108		108		108	

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***%ΔRE*** is defined as $(\frac{rev_t}{emp_t} - \frac{rev_{t-1}}{emp_{t-1}}) / (\frac{rev_{t-1}}{emp_{t-1}})$, and ***DiffEmp*** is defined as $\frac{rev_t - rev_{t-1}}{rev_{t-1}} - \frac{emp_t - emp_{t-1}}{emp_{t-1}}$.
Please refer to footnotes in table 3.3 for definitions of all other variables.

prediction errors of the Diffemp model (Wilcoxon, z=0.006; mean difference, p=0.045). The nested F-test evaluation and the predictive ability comparison provide further support that $\%\Delta RE$ is a better predictor of Fraud than Diffemp. Recall that Unexpected Revenue per Employee is defined as the difference between a firm's $\%\Delta RE$ and the %ΔRE of the firm's industry. Diffemp is, however, not adjusted for industry differences and as such I used %ΔRE instead of Unexpected Revenue per Employee in the comparison just described. The specific form of the capital productivity measure and whether the industry adjustment adds value are empirical questions that I leave unanswered for future research to explore.

Dechow et al. (2007) also introduce an employee based fraud indicator, Abnormal Change in Employees defined as the percentage change in employees minus the percentage change in assets. This measure is very similar to the measure introduced in Brazel et al. (2007), though neither study discusses the other measure. I focus on comparing Unexpected Employee Revenue to Diffemp in this supplemental analysis as Diffemp is closer in nature than Abnormal Change in Employees to Unexpected Employee Revenue. I nevertheless provide some brief results from analyses of Abnormal Change in Employees. Since the correlation between Abnormal Change in Employees and Unexpected Revenue per Employee is low (r=-0.103), I include these two variables in the same model. The results in Table 3.12 show that Unexpected Revenue per Employee is in the direction expected and significant (p=0.018), while Abnormal Change in Employees is insignificant (p=0.955).

3.6.2.2 Total Discretionary Accruals Aggregation Periods

The results of my research confirm that the likelihood of fraud is significantly (p=0.009) higher for firms that are pressured and constrained by earnings management in prior years (Table 3.5). Total Discretionary Accruals was measured as a firm's total discretionary accruals during the three years leading up to the first fraud year. I chose to aggregate discretionary accruals over multiple years based on the idea that over time discretionary accruals reverse. Three years was used since it conceptually seems to be an appropriate time frame and based on the graphical depiction of the relation between discretionary accruals and fraud in (Dechow et al. 1996). To evaluate the appropriateness of this decision, I examine the relation between fraud likelihood and discretionary accruals aggregated over the two years leading up to the first fraud year, Total Discretionary Accruals2, and discretionary accruals in the year prior to the first fraud year, Total Discretionary Accruals1. Using model 33 and one of the prior year(s) total discretionary accruals measures at a time, Total Discretionary Accruals2 is in the expected direction and marginal significant (p=0.068) and Total Discretionary Accruals1 is also in expected direction but insignificant (p=0.108), see Table 3.13 columns 2 and 3, respectively. The Total Discretionary

Table 3.12
Unexpected Revenue per Employee and Abnormal Change in Employees
Logistic Regression Results[a]

Variable[b]	*Prediction*	*Estimate*	*Standard Error*	χ^2	*prob*>χ^2
Intercept	(?)	0.702	0.800	0.770	0.380
Total Discretionary Accruals	(+)	1.641	0.791	5.600	0.009
Forecast Attainment	(+)	0.571	0.227	6.737	0.005
Unexpected Revenue per Employee	**(+)**	**1.440**	**0.718**	**4.398**	**0.018**
Abnormal Change in Employees	**(+)**	**-0.042**	**0.747**	**0.003**	**0.955**
Current Discretionary Accruals	(+)	-0.852	1.612	0.291	0.590
Sales to Assets	(-)	-0.320	0.342	0.901	0.171
Auditor	(-)	0.192	0.629	0.095	0.758
Asset Growth	(+)	0.354	0.232	2.376	0.062
CFO Change	(+)	0.650	0.364	3.441	0.032
Pseudo R^2		0.183			
χ^2-test of model fit		25.14 (p=0.005)			
n		100[c]			

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***Abnormal Change in Employees*** is defined as the percentage change in employees minus the percentage change in assets. Please refer to footnotes in table 3.3 for definitions of all other variables.

Accruals2 and the Total Discretionary Accruals1 models have Pseudo R^2 values of 0.110 and 0.107, respectively. The higher p-value of Total Discretionary Accruals (p=0.009) and the higher Pseudo R^2 value of the Total Discretionary Accruals model (0.135) (Table 3.13 column 1) indicates that Total Discretionary Accruals is a stronger predictor of Fraud than either Total Discretionary Accruals2 or Total Discretionary Accruals1.

To further substantiate this claim I perform nested F-tests where the fit of a reduced model is compared to a full model. The reduced model excludes the Total Discretionary Accruals variables (column 4, Table 3.13). The full models are the three models described above. The log likelihood of these models (columns 1, 2, and 3, Table 3.13) are then compared to the log likelihood of the reduced model. The results show that only the Total Discretionary Accruals model (column 1, Table 3.13) significantly (p=0.017) improves the fit of the reduced model.

I also evaluate the predictive ability of the Total Discretionary Accruals model against the Total Discretionary Accruals2 model and Total Discretionary Accruals1 model. The prediction errors of the Total Discretionary Accruals model are significantly lower than the prediction errors of the Total Discretionary Accruals2 model (Wilcoxon, z=0.0007; mean difference, p=0.0102)

Table 3.13
Three Years, Two Years and One Year Total Discretionary Accruals
Logistic Regression Results[a]

		(1)		(2)		(3)		(4)	
Variable[b]	*Prediction*	*Estimate*	*prob>χ^2*	*Estimate*	*prob>χ^2*	*Estimate*	*prob>χ^2*	*Estimate*	*prob>χ^2*
Intercept	(?)	0.065	0.931	0.034	0.963	0.054	0.941	0.181	0.800
Total Discretionary Accruals	**(+)**	**1.641**	**0.009**						
Total Discretionary Accruals2	**(+)**			**1.390**	**0.068**				
Total Discretionary Accruals1	**(+)**					**1.701**	**0.108**		
Forecast Attainment	(+)	0.573	0.004	0.524	0.007	0.509	0.009	0.530	0.006
Unexpected Revenue per Employee	(+)	1.445	0.016	1.393	0.018	1.466	0.015	1.341	0.020
Current Discretionary Accruals	(+)	-0.875	0.560	-0.923	0.527	-0.656	0.655	0.054	0.482
Sales to Assets	(-)	-0.327	0.148	-0.280	0.179	-0.262	0.193	-0.254	0.200
Auditor	(-)	0.192	0.759	0.207	0.736	0.187	0.762	0.115	0.848
Asset Growth	(+)	0.359	0.047	0.344	0.051	0.368	0.042	0.332	0.056
CFO Change	(+)	1.304	0.031	1.165	0.046	1.074	0.057	0.990	0.073
Pseudo R^2		0.135		0.110		0.107		0.097	
χ^2-test of model fit		**20.15 (p=0.010)**		**16.72 (p=0.033)**		**16.03 (p=0.042)**		14.50 (p=0.043)	
Diffχ^2_{full} -$\chi^2_{reduced}$		**5.65 (p=0.017)**		**2.22 (p=0.136)**		**1.53 (p=0.216)**			
Predictive Ability (Mean Difference)				**p=0.0102**		**p=0.0098**			
Predictive Ability (Wilcoxon)				**z=0.0007**		**z=0.0087**			
n		108		108		108		108	

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***Total Discretionary Accruals*** is the total amount of discretionary accruals deflated by assets in the three years leading up to the fraud year, see table 3.3 for calculation of discretionary accruals, while ***Total Discretionary Accruals2*** (***Total Discretionary Accruals1***) is the total amount of discretionary accruals in the two (one) years leading up to the fraud year. ***Diffχ^2_{full} -$\chi^2_{reduced}$*** is twice the difference in negative log-likelihoods between the reduced model in column (4) and the full models in columns (1), (2) and (3). ***Predictive Ability*** compares the prediction errors of the full model in column (1) to the prediction errors of the full models in column (2) and (3). Please refer to footnotes in table 3.3 for definitions of all other variables.

Table 3.14
Alternative Analyst Forecast Measure
Logistic Regression Results[a]

Variable[b]	*Prediction*	*Estimate*	*Standard Error*	χ^2	*prob*>χ^2
Intercept	(?)	-0.112	0.720	0.020	0.877
Total Discretionary Accruals	(+)	1.576	0.742	6.001	0.007
Forecast Attainment (average)	**(+)**	**0.336**	**0.214**	**2.514**	**0.056**
Unexpected Revenue per Employee	(+)	1.206	0.688	3.318	0.034
Current Discretionary Accruals	(+)	-0.875	1.481	0.375	0.540
Sales to Assets	(-)	-0.255	0.305	0.717	0.199
Auditor	(-)	0.228	0.609	0.141	0.707
Asset Growth	(+)	0.382	0.213	3.314	0.034
CFO Change	(+)	1.269	0.719	3.368	0.033
Pseudo R^2		0.106			
χ^2-test of model fit		15.81 (p=0.045)			
n		108			

[a] Effect Likelihood Ratio Tests, one-tailed tests reported for estimates in the direction of the prediction, all other two-tailed.

[b] Dependent variable is financial statement fraud likelihood; ***Forecast Attainment (average)*** is a dummy variable equal to 1 if analyst forecast were met or exceeded and 0 otherwise (I/B/E/S), where the analyst forecast is the average of all consensus forecasts made throughout the year leading up to the first fraud year. Please refer to footnotes in table 3.3 for definitions of all other variables.

and the Total Discretionary Accruals1 model (Wilcoxon Sign-Rank, z=0.0087; mean difference, p=0.0098).

To summarize, in addition to being supported conceptually and by prior empirical research, the individual variable statistics, the nested F-test evaluation and the predictive ability comparison, support the use of Total Discretionary Accruals over Total Discretionary Accruals1 and Total Discretionary Accruals2.

3.6.2.3 Analyst Forecast Period

I use the first forecast rather than the most current forecast because financial statement fraud can be an on going activity occurring throughout the year. To evaluate the appropriateness of this decision I also examined the average of all consensus forecasts made throughout the year leading up to the first fraud year. The results were weaker but in the same direction, more specifically meeting or exceeding analyst forecasts was positively (p=0.056) related to the likelihood of fraud (Table 3.14). Thus, it appears that using the first forecast in the period is preferable in fraud detection.

3.7. Conclusions

My research objective was to improve our understanding of antecedents of fraud, and thereby improve our ability to detect fraud. To realize this objective I developed a new measure that aggregates discretionary accruals over the three years leading up to the first fraud year to capture the pressure of earnings reversals and earnings management inflexibility. My results show that firms that are running out of ways to manage earnings, and are facing accruals reversals as a result of earnings management in prior years are more likely to commit financial statement fraud. I also perform more in depth analyses of the earnings reversal hypothesis that provides some initial indications that (1) the earnings reversal hypothesis also applies to real activities manipulation, and (2) discretionary accruals should be summed over three years, rather than two years or one year.

This study also adds to fraud research by examining whether capital market expectations provide incentives to commit financial statement fraud. My results show that evidence of a firm meeting or beating analyst forecasts can be used to detect financial statement fraud. Additionally, this study adds to earnings management research investigating capital market expectation, which typically assumes that distributional inconsistencies in reported earnings around analyst forecasts indicate that some firms manipulated earnings to meet analyst forecasts. The results provide more direct evidence of earnings manipulation incentives related to capital market expectations and corroborate the findings of earnings management research.

I also develop a new productivity based measure designed to capture revenue fraud. The results show that this measure provides utility in fraud prediction and that the inclusion of this measure in fraud detection models improves model fit and model predictive ability more than similar measures being proposed in contemporary research. This measure might also provide insights to the development of new and improved discretionary accrual measures.

These results should, however, be interpreted knowing that the sample of fraud firms was taken from SEC Accounting and Auditing Enforcement Releases. Thus, I am really improving our understanding of fraud firms investigated by the SEC and this knowledge might not fully generalize to fraud in general.

Future research can extend this study in a number of ways. I proposed that total discretionary accruals increase the likelihood of fraud through two processes, prior earnings management puts pressure on management as the accruals reverse and prior earnings management constrains current earnings management flexibility. Future research can explore these two dimensions further and examine if only one, or if both, processes increase the likelihood of fraud. Earnings reversals can be viewed as providing an incentive to commit fraud, while earnings management

inflexibility is a condition that increases the likelihood of fraud given a set of incentives. Thus, earnings management inflexibility should interact with other incentives. Future research can also examine fraud incentives related to other capital market expectations than analyst forecasts. For example, do firms commit fraud so they do not have to report small losses or small earnings growth declines? Future research can also develop additional account specific fraud measures, or even focus on specific types of fraud. I developed a measure for detecting revenue fraud because revenue is the most commonly manipulated account. However, other accounts are also fraudulently manipulated. Future research could design new measures for detecting fraud in these accounts as well. These account specific measures could then be combined with non-specific measures that measure general incentives, conditions and management character, to improve the utility of fraud classification models. Different models could be built to identify different types of fraud and the classifications of these individual models could then be combined into an overall classification. Multi-classifier combination is a rich research stream in machine learning that could provide the foundation for this research.

To summarize, the results show that the likelihood of fraud is significantly higher for firms that meet or exceed analyst forecasts, are running out of ways to manage earnings and are facing accrual reversals, or have high labor productivity. These findings can help the SEC, auditors, financial institutions and others improve fraud prevention and detection, and thereby curb costs associated with fraud and improve market efficiency. These findings are also important to auditors that need to provide reasonable assurance about whether the financial statements are free of material misstatements caused by fraud, especially during client selection and continuation judgments, and audit planning.

Chapter 4. Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms

4.1. Introduction

> *Undetected financial fraud is one of the greatest risks to an organization's viability and corporate reputation, and it has the capacity to draw into its sphere all associated people, not only the guilty* (KPMG 2006)
>
> *- Jeffrey Lucy, Chairman, Australian Securities and Investments Commission*

The cost of financial statement fraud to defrauded organizations is estimated at $340 billion[20] per year in the U.S. (ACFE 2006). In addition to these direct costs, employees and investors are impacted negatively by financial statement fraud. Fraud also undermines the trustworthiness of corporate financial statements, which results in higher transaction costs and a less efficient market. Auditors have, both through self regulation and legislation, assumed the responsibility of providing reasonable assurance that the financial statements are free of material misstatement due to error or fraud. Earlier auditing standards, i.e., Statement on Auditing Standards (SAS) No. 53, only indirectly addressed this responsibility through references to "irregularities" (AICPA 1988). However, more recent auditing standards, SAS No. 82 and later, make this responsibility explicit. The auditor must provide "reasonable assurance about whether the financial statements are free of material misstatements, whether caused by error or fraud" (AICPA 1997, AU 110.02).

A stream of accounting research focuses on testing various statistical and data mining models with the goal of improving fraud detection. Data mining research that focuses specifically on financial statement fraud detection is important since the financial statement fraud domain is

---

[20] The ACFE (2006) report provides estimates of total fraud cost, mean cost per fraud category and number of cases. To derive the estimate for total cost of financial statement fraud, I assumed that the relative difference in mean is similar to the relative difference in median cost among the different occupational fraud categories.

unique. Distinguishing characteristics that make this domain unique include: (1) the ratio of fraud to non-fraud firms is small (high class imbalance); (2) there are different types of fraud (target class disjunct); (3) the ratio of false positive to false negative classification error cost is small (cost imbalance); (4) the attributes used to detect fraud are relatively noisy where similar attribute values can signal both fraudulent and non-fraudulent activities; and (5) fraudsters actively attempt to conceal the fraud by making the fraud firm's attribute values look like the attribute values of non-fraud firms. Because of the unique characteristics it is not clear, without empirical evaluation, whether classifiers that perform well in other domains will also perform well in the financial statement fraud domain.

The financial statement fraud research typically uses logistic regression as the baseline model against which data mining models are tested. The data mining algorithm used in this line of research has almost exclusively been artificial neural networks (ANN), for example Green and Choi (1997), Lin et al. (2003), Fanning and Cogger (1998) and Feroz et al. (2000). A number of data mining algorithms that have been found to be good predictors in other domains have not been investigated in financial statement fraud research. Two more recent financial statement fraud studies have included additional algorithms in their comparisons (Kotsiantis et al. 2006; Kirkos et al. 2007). However, these comparisons were based on unrealistic ratios of fraud to non-fraud firms (1:1: and 1:3); assumed equal costs of Type I and Type II errors; and used accuracy and error rate to measure performance, which generally are considered inappropriate for a domain like fraud detection. Furthermore, the Kotisantis et al. (2006) and Kirkos et al. (2007) studies included only financial ratios as predictor variables, leaving out many measures developed in confirmatory financial statement fraud research. Thus, we know very little about what algorithms are useful for detecting financial statement fraud, under what specific circumstances one algorithm might be better than another and what predictors are useful for the different algorithms.

My research objective is to compare the utility of a fairly comprehensive set of data mining algorithms in financial statement fraud prediction. More specifically, my research questions are:

1. What algorithm(s) provide the most utility given different assumptions about classification costs and prior probabilities of fraud?
2. What predictors are useful to these algorithms when detecting financial statement fraud?

The answers to these questions are of practical value to auditors and institutions, like the Securities and Exchange Commision (SEC). The results provide guidance as to what algorithms and predictors to use when creating new models for financial statement fraud detection. Auditors can use these algorithms and predictors to improve client selection, audit planning and analytical

procedures; while the SEC can leverage the findings to target audit engagements where the auditee is more likely to have committed financial statement fraud.

I provide an overview of related research in Section 4.2. Section 4.3 contains a description of the experimental variables and data used to evaluate the classification algorithms, while Section 4.4 describes the experimental procedure and reports pre-processing results. The experimental results are reported in Section 4.5. The results are summarized in Section 4.6, along with a discussion of research contributions and limitations, and suggestions for future research.

4.2. Related Research

Research focusing on evaluating the effectiveness of different fraud prediction algorithms has typically introduced different variations of ANN and compared these algorithms to logistic regression. Green and Choi (1997) evaluated the performance of three ANNs with input variables preprocessed in different ways: simple percentage change, plain sum-of-the-years'-digit weighted average, and incremental sum-of-the-years'-digit weighted average. In an experiment with 86 SEC fraud cases matched with 86 non-fraud cases the ANNs were compared to random guessing, defined as Type I and Type II error rates of 0.5, and summed error rate below 1. The results showed that the ANNs performed better than random guessing on the training sample. On the evaluation sample, however, the ANNs did not perform significantly better in terms of either Type I or Type II errors. The summed error rate comparison did show that the ANNs performed significantly better than random guessing, but this comparison used classification results from a combined sample of both the training and evaluation samples.

Starting with 62 potential fraud predictors, Fanning and Cogger (1998) compared an ANN to quadratic discriminant analysis, linear discriminant analysis and logistic regression. The ANN correctly classified 63% of the cases, as compared to 52% for the highest benchmark algorithm (linear discriminant analysis). However, the ANN had a lower true positive rate, defined as the number of true positive classifications divided by the number of positive instances in the dataset, than all three benchmark algorithms.

In both Green and Choi (1997) and Fanning and Cogger (1998) the experiments assumed that the costs of false positive and false negative were equal and that the dataset was balanced with a prior probability of 0.5. In reality, the probability of fraud is much smaller than 50%[21], and the cost of false negatives is often much larger than the cost of false positives. The optimal threshold

[21] In one estimate the probability of fraud is 0.6% (Bell and Carcello 2000).

for classifying a financial statement as fraudulent in discriminant analysis and logistic regression analyses is, therefore, likely to be much lower than the default of 0.5 used in these studies. As false positive and false negative rates do not remain constant over different threshold levels and the relative performance of algorithms is often different at different thresholds, the assumptions made in these studies limit our understanding of the performance of these algorithms to one specific scenario that is unrealistic.

Using seven SAS No. 53 red-flags, Feroz et al. (2000) focused on comparing the utility of an ANN model with logistic regression based on hit-rate, overall error rate[22] and estimated relative costs of misclassification.[23] The results did not show that one algorithm was dominant at all treatment levels where prior probabilities[24] and the relative cost of different classification errors[25] were manipulated. Lin et al. (2003) introduced an existing data mining algorithm, fuzzy ANN, into the accounting domain and compared the fuzzy ANN model to logistic regression using hit rate,[26] overall error rate and estimated relative costs of misclassification. The results, as in Feroz et al. (2000), did not show that either of the algorithms dominated. The analysis was performed using seven financial ratios that were related specifically to the revenue cycle. The sample was, however, not adjusted accordingly, i.e., non-revenue based fraud was included in the sample. In the overall error rate and estimated relative costs of misclassification analyses in Feroz et al. (2000) and Lin et al. (2003), the optimal fraud classification probability cutoff level was not determined for each treatment group, i.e., algorithm, relative error cost and prior probability treatment combination. The analyses instead used error rates obtained based on optimal cutoffs

---

[22] Overall error rate takes into account differences in prior-probabilities of the different outcomes and the type of classification error.

[23] Estimated relative costs of misclassification takes into account: prior-probabilities, classification costs of different outcomes, and the type of classification error (false positive or false negative). Note, however, that if the prior-probabilities and relative costs are not adjusted during model building, i.e., training, then the models might not perform optimally for the different prior-probability and relative classification costs combinations. Thus, this performance measure might be misleading if the different models are not rebuilt/retrained for each prior-probability and relative classification cost combination examined using the measure.

[24] The prior-probability refers to the percentage of fraud firms.

[25] The cost of making a Type I error (false positive) compared to the cost of making a Type II error (false negative).

[26] Hit rate is the percentage of objects accurately classified.

when assuming no differences in classification costs and without finding the best prior training fraud probabilities for different evaluation prior fraud probability levels.

More recently, Kotsiantis et al. (2006) used 41 fraud and 123 non-fraud firms in Greece to examine seven classification algorithms: C4.5 (Decision Tree), RBF (ANN), K2 (Bayesian networks), 3-NN (nearest-neighbor), RIPPER (rule-learner) and SMO (support vector machine) and logistic regression. They also examined four ensemble based algorithms[27]: MP stacking (stacking with seven base-classifier and a tree learner meta-classifier that only learns from true class probabilities), MLR stacking (stacking with seven base-classifier and a multi-response linear regression meta-classifier that learns from all class probabilities), Grading (IBk base-classifier with 10-NN meta-classifier) and Simple Voting (stacking with seven base-classifier and a tree learner meta-classifier). The algorithms were trained using 10-fold cross validation using financial statement ratios as input variables. The results in terms of overall accuracy showed that MP stacking provided the best performance (95.1%) followed by MLR stacking (93.9%), and that all ensemble based methods outperformed the best non-ensemble algorithm. The best non-ensemble algorithm was C4.5 (91.2%), followed by RIPPER (86.8%). The accuracy of logistic regression (75.3%) and the ANN (73.4%) was relatively low. Accuracy was also reported for each class, i.e., the equivalent of Type I and Type II error rates. While Kotsiantis et al. (2006) evaluated a relatively comprehensive set of classifiers, the percentage of fraud firms in their dataset (25%) was much higher than estimates indicating that around 0.6% of all organizations are fraud firms (Bell and Carcello 2000). Furthermore, they assumed that costs associated with Type I and Type II errors were equivalent, while in reality it is likely that Type II errors were much more expensive than Type I errors. Based on financial statement fraud effects reported in Beneish (1999), Bayley and Taylor (2007) estimated that the ratio of Type I error classification costs to Type II error classification costs was between 1:20 and 1:40. These assumptions were also reflected in their training data, in that they did not examine the performance of the classifiers at different classification thresholds.

Kirkos et al. (2007) compared an ANN, a Bayesian belief network and a decision tree learner using 35 fraud and 38 non-fraud firms. The reported class accuracies (fraud accuracy, non-fraud accuracy) indicated that the Bayesian belief network (91.7%, 88.9%) outperformed the ANN (82.5%, 77.5%) and the decision tree (75%, 72.5%). As in Kotsiantis et al. (2006) the

[27] Ensemble based classification algorithms combines the decision output from multiple classifiers, i.e., they use an ensemble of classifiers (see Section 4.3.1.2. for further detail).

classification costs were assumed to be the same for Type I and II errors, and the dataset contained almost the same number of fraud firms as non-fraud firms.

I extend this literature by evaluating the performance of a relatively extensive set of algorithms selected based on their classification performance in both prior financial statement fraud research and in domains similar to the financial statement fraud domain. I also examine under what specific, realistic circumstances these algorithms perform well and what predictors provide utility to these algorithms in terms of improving classification performance.

4.3. Experimental Variables and Data

To answer my research questions, I ran experiments to evaluate the performance of a relatively comprehensive and representative set of classification algorithms. The classification algorithms were used to classify firms as either fraud or non-fraud firms based on firm attributes derived from the financial statements, analyst forecasts and proxy statements of the firms. Classifier performance was measured using estimated relative cost under different assumptions of relative costs of Type I and Type II errors and prior probability of fraud and non-fraud (Dopuch et al. 1987). The performance of each classifier configuration and training scenario combination was measured 10 times using 10-fold cross validation for each combination of classification costs and prior fraud probability. Section 4.3.1, 4.3.2 and 4.3.3 describe the three factors that were manipulated in the experiment: classification algorithms, classification costs and prior probability of fraud, respectively. The dependent measure, estimated relative cost, is described in Section 4.3.4. The dataset, which includes objects (fraud and non-fraud firms) and object features (fraud predictors), is described in Section 4.3.5.

4.3.1 Classification Algorithms

4.3.1.1 Overview

One of the goals of this research is to examine classification algorithm performance in fraud detection. The primary experimental factor of interest was, therefore, classification algorithms. The classification algorithms were obtained from Weka, an open source data mining tool that contains components for (1) preprocessing data, including data loading, data filtering and transforming object attributes, (2) object classification, clustering and association rule mining, (3) object attribute evaluation, and (4) result analysis and visualization. Using an open source tool facilitates the replication and extension of this study by future financial statement fraud data mining research. Weka implements a relatively complete set of classification algorithms, including many of the most popular algorithms. Based on prior financial statement fraud research

and prior data mining research in domains with imbalanced datasets, I selected six algorithms from Weka: J48, SMO, MultilayerPerceptron, Logistics, stacking and bagging. J48 is a decision tree learner and Weka's implementation of C4.5. SMO is a support vector machine (SVM) and Logistics is Weka's logistic regression implementation, both of these classifiers are linear functions. MultilayerPerceptron is Weka's backpropagation ANN implementation, and stacking and bagging are two ensemble based methods. Section 4.3.1.3 provides more in-depth descriptions of these algorithms. In addition to the algorithms implemented in Weka, I also included Information Marked based Fusion (IMF), described in Essay I.

Logistic regression, ANN and stacking were included as they had performed well in prior fraud research (Feroz et al. 2000; Lin et al. 2003; Kotsiantis et al. 2006). Note that it was, however, not clear if these algorithms would perform well under realistic conditions and relative to not yet examined classifiers. Bagging, J48, SMO and IMF were included because prior data mining research (Fries et al. 1998; Phua et al. 2004; West et al. 2005) and the research in Chapter 2 found that these algorithms performed well in domains with imbalanced data, i.e., where the majority class was larger than the minority class, which is true in the financial statement fraud domain. It was, however, not known how these classifiers would perform in fraud detection.

In the next section, I describe in greater detail why these specific algorithms were selected (Section 4.3.1.2). I then provide more in-depth descriptions of each algorithm and experimental classifier parameter settings (Section 4.3.1.3).

4.3.1.2 Algorithm Selection

Of the classifiers I selected, four were individual classifiers: J48 (C4.5), SMO (SVM), MultilayerPerceptron (ANN) and Logistics (logistic regression). I included logistic regression and ANN algorithms to allow for comparisons with prior financial statement fraud research (Green and Choi 1997; Fanning and Cogger 1998; Feroz et al. 2000). I included SVM and C4.5 as these algorithms were found, in domains other than fraud, to provide good classification performances (Fries et al. 1998; Fan and Palaniswami 2000; Phua et al. 2004).

Fries et al. (1998) examined the performance of SVM on a dataset with a 34.5% prior probability of patients having breast cancer. The accuracy of SVM (99.5%) was higher than that reported previously for CART (94.2%), RBF ANN (95.9%), linear discriminant (96.0%) and multi-layered ANN (96.6%). Fan and Palaniswami (2000) compared a SVM to an ANN, MDA and learning vector quantization in bankruptcy prediction, with a 49.4% prior probability of bankruptcy, and found that the SVM outperformed the other classifiers. In another bankruptcy prediction evaluation, Shin et al. (2005) used a balanced dataset of 1160 bankrupt firms and 1160 non-bankrupt firms and found that the SVM outperformed the ANN.

Data mining research has also investigated the efficacy of various algorithms in the medical field. Weiss and Kapouleas (1990) found that two rule-based classifiers, CART (0.0064[28]) and PVM (0.0067), outperformed ANN (0.0146), Bayes Independence (0.0394), Nearest Neighbor (0.0473), DA linear (0.0615), Bayes 2nd order (0.0756), and DA Quadratic (0.1161) on a thyroid disease dataset with a 7.7% prior probability of thyroid disease. They also found similar results on a breast cancer dataset with a 70% prior probability of the existence cancer. CART (0.229) and PVM (0.229) again had the best overall performance and outperformed ASSISTANT Tree (0.280), Bayes Independence (0.282), ANN (0.285), DA linear (0.294), Bayes 2nd order (0.344), DA Quadratic (0.344), and Nearest Neighbor (0.347). Based on these results they concluded that rule based classifiers provided the best performance, especially on imbalanced data. In addition to performance benefits, rule-based classifiers also generate output that is interpretable by humans.

Of the rule-based classifiers, and perhaps of all machine learning algorithms, C4.5 and its commercial successor C5 have arguably become the most commonly used off-the-shelf classifiers (Witten and Frank 2005). Using a 1:3 unbalanced dataset of Greek financial statement fraud firms, Kotsiantis et al. (2006) found that C4.5 (91.2%), followed by RIPPER (86.8%), two rule-based classifiers, were the best non-ensemble based algorithms in their experiment. Their experiment also included five other individual classifiers: an ANN (RBF), a Bayesian network (K2), a nearest-neighbor (3-NN), a SVM (SMO) and logistic regression. Phua et al. (2004) provided further support for including the C4.5 algorithm in this study. They found that C4.5 performed relatively well in a dataset with a 6% prior probability of auto insurance claims fraud. Using a cost sensitive performance measure (cost savings within parentheses), C4.5 ($165) performed better than bagging ($127), stacking ($104), Naive Bayesian ($94) and ANN ($89). Furthermore, the performance of C4.5 was relatively robust with respect to the sampling method.

In addition to these four individual classifiers, I also examined three ensemble based methods[29]: bagging, stacking and IMF. By combining the results of a group or ensemble of individual classifiers, classification performance can be improved over the best individual classifier in the ensemble. The basic idea behind ensemble based methods is that different

---

[28] Error rates, defined as the number of false positive and false negative classifications divided by all instances in the dataset, are reported within the parentheses.

[29] As the name indicates an ensemble based method is a type of classification method that uses a group (ensemble) of individual classifiers, so called base-classifiers, to classify objects. Different ensemble based methods include different base-classifiers, train the base-classifiers differently and use different algorithms to combine base-classifier decision or probability outputs into an overall ensemble decision or probability.

classifiers have different strengths and weaknesses, and therefore provide complementary information about the classification problem. These differences can be leveraged to improve classification performance by combining the individual classifiers' decisions (Kittler et al. 1998). Ensemble research has primarily focused on two areas: (1) ensemble architecture, examining what classifiers to include in the ensemble and how to train these classifiers; and (2) combiner method, examining how to combine the base-classifiers' decisions.

Using an ensemble of ANN base-classifiers, West et al. (2005) compared the performance of crossvalidation (CV), bagging and boosting, three popular methods from ensemble architecture research. Three datasets were used in this comparison: Credit Rating Australian (307 no credit given and 383 credit given), Credit Rating German (300 no credit given and 700 credit given), and Bankruptcy (93 bankrupt and 236 non-bankrupt companies). The results (error rates) for the three datasets showed that bagging performed relatively well. More specifically, in the Australian Credit Rating dataset, bagging (0.128) outperformed CV (0.131), the single best base-classifier (0.132) and boosting (0.148). In the German Credit Rating dataset, CV (0.242) outperformed bagging (0.251), but these two ensemble methods both performed better than the single best base-classifier (0.253) and boosting (0.255). In the Bankruptcy dataset, bagging (0.126) outperformed boosting (0.128), CV (0.129) and the single best base-classifier (0.131). Note that in addition to performing well, bagging was the best performing classifier in the Bankruptcy dataset. This dataset is the most similar of the three to the fraud domain; it is highly imbalanced with object attributes derived largely from financial data and with a classification objective similar to that in the fraud domain. Based on these findings I included bagging in my experiment. I additionally included stacking, since prior fraud research using Greek fraud firms found stacking to perform well (see the literature review in Section 4.2, in particular Kotsiantis et al. 2006).

Ensemble combiner method research has found that relatively simple methods like Majority Vote and Average perform either at the same level or significantly better than more complex methods (Duin and Tax 2000), but that Information Market based Fusion (IMF) outperforms these two combiner methods and Weighted Average when the true classes of all objects are revealed (see Chapter 2). Further, Chapter 2 shows that when only the true classes of objects classified as positive are known, which is a more realistic assumption in certain domains such as fraud detection, IMF outperforms Majority Vote, Average and Weighted Average for datasets with low prior probabilities of the minority class (defined as datasets with prior probabilities below 40%). Based on these findings I include IMF in my experiment.

4.3.1.3 Algorithm Overview and Tuning

J4.8 is Weka's implementation of C4.5 version 8. C4.5 generates a decision tree, which is a divide and conquer classification method. The algorithm examines the information gain provided by each attribute and splits the data using the attribute that provides the highest information gain. The created branches are then split further by again examining the information gain, within each branch, provided by the different attributes. If an attribute creates a split with a branch with no or only a small number of instances, then this attribute is not used. The minimum number of instances permissible at a leaf is set by default to two, but can be changed. To avoid overfitting, the branches are pruned using subtree-raising where internal decision nodes (branch splits) are subsumed by lower nodes, thereby eliminating one node. The pruning is performed based on estimates of classification errors established using a confidence value that is set by default to 25% (Witten and Frank 2005). The reader is referred to Quinlan (1993) for further detail on C4.5. Witten and Frank (2005) suggest tuning C4.5 by testing lower confidence values and higher minimum number of instances. I examined three confidence values, 15%, 20% and 25%, and three minimum number of instances at a leaf, 2, 3 and 5, for a total of nine C4.5 configurations.

Logistic regression is a statistical algorithm that estimates the probability of a certain event occurring by applying maximum likelihood estimation after transforming the dependent variable into the natural log of the odds of the firm being fraudulent. ANNs are non-linear machine learning algorithms designed based on biological neural networks with interconnected input, hidden and output nodes.[30] Both of these classifiers have been used extensively in prior accounting and fraud research. Following prior research, logistic regression was not tuned, meaning that logistic regression was used with parameters set to their default values. I manipulated the learning time (epochs), learning rate, the number of hidden nodes and momentum for the ANN (Feroz et al. 2000; Green and Choi 1997). A good learning time was first determined without manipulating the other settings, which were set at their default Weka values (learning rate = 0.3; momentum = 0.2; the number of hidden nodes = the sum of the number of attributes and the number of classes divided by 2). The learning time was determined by comparing the performance of ANNs created using different learning times, starting with 500 epochs (iterations through the training data) and then increasing the number of epochs by 500 in each new evaluation round. The learning time evaluation was terminated when the performance did not improve over three consecutive learning time levels. At this point the learning time was

[30] See Green and Choi (1997) for a good description of ANNs.

set to the lowest learning time generating the highest performance, and the other settings were manipulated. The learning rate and momentum were both manipulated at three levels: 0.1, 0.3 and 0.5. The number of hidden nodes was manipulated at four levels: 4, 8, 12 and 16. Thus, after the learning time was determined, a total of 27 ANN configurations were included in the experiment.

SVM algorithms classify data points by learning hyperplanes (linear models) that provide the best separation of a set of positive instances from a set of negative instances. In a classification context with n object features, a hyperplane is a linear model with n-1 dimensions intersecting a space of n dimensions into two parts. For example, in a two-dimensional space the hyperplane is a line, and in a three-dimensional space, the hyperplane is a plane. The objective is to find the hyperplane that maximizes the separation of the data points of the different classes. While the hyperplane is a linear model, the input data can be transformed before the hyperplane is constructed. The effect of learning the hyperplane on transformed data is a non-linear decision boundary in the original space (Witten and Frank 2005). The hyperplane that provides the best separation is found by solving a large quadratic programming optimization problem. To improve training speed, sequential minimal optimization (SMO) solves the quadratic programming problem by breaking it up into a series of smaller problems that are then solved analytically (Platt 1999). Weka implements SVM using SMO. Following Shin et al. (2005), the complexity parameter C was manipulated at five values: 1, 10, 50, 75 and 100. Shin et al. (2005) also manipulated a radial basis kernel parameter, but since I used a polynomial kernel function I instead manipulated the exponent of the polynomial kernel at five values: 0.5, 1, 2, 5 and 10. Thus, 25 potential SVM configurations were included in the experiment. Furthermore, buildLogisticModels was set to true. This enabled proper probability estimates by fitting logistic regression models to the SVM outputs.

Stacking is an ensemble based method that combines the output of heterogeneous base-classifiers, i.e., different types of classifiers, trained on the same data. The base-classifier output is combined using a meta-classifier. The meta-classifier can be any classification algorithm, but is typically a relatively simple linear model or decision tree. To avoid overfitting the training data, the meta-classifier is trained on base-classifiers evaluation output generated using test data rather than training data. This is typically accomplished using k-fold cross validation (Wolpert 1992; Witten and Frank 2005). In the experiment, stacking was configured using the default Weka setting for the number of cross-validation folds (set at 10). In selecting base-classifiers, I followed prior research (Kotsiantis et al. 2006) and used all the other experimental classifiers, including bagging, but excluding IMF. Note that IMF was implemented using all the other classifiers in the experiment except for stacking. Thus, the base-classifiers selected were C4.5, SVM, ANN,

logistic regression and bagging. I included all classifier configurations that provided the best performance for a given classifier at one or more experimental treatment levels. Based on recommendations to use a relatively simple meta-classifier (Wolpert 1992), and experiments performed by Chan et al. (1999) and Prodromidis et al. (2000), I used a Bayesian classifier as the meta-classifier. In a dataset with a 0.2 prior probability of credit card fraud, Chan et al. (1999) evaluated the performance of four individual classifiers, C4.5, Ripper, CART and Bayesian, and a meta-classifier, Bayesian. They found that the Bayesian meta-classifier provided the best performance followed by CART. In a similar credit card fraud detection study, Prodromidis et al. (2000) evaluated C4.5, Ripper, CART, Bayes and ID3, and a Bayesian meta-classifier. In addition to the dataset with the 20% fraud cases from Chan et al. (1999), they also included a dataset with 15% fraud cases. As in Chan et al. (1999), Prodromidis et al. (2000) found that the Bayesian meta-classifier performed better than the other classifiers. I, therefore, used Bayesian as the meta-classifier in the stacking implementation. In Weka, NaiveBayes can be configured to use either kernel estimation or a single normal distribution for modeling numeric attributes. There is also an option to use supervised discretization to process numeric attributes. I manipulated these parameter settings in the experiment for a total of four stacking configurations.

Bagging is an ensemble based method that combines the output of homogenous base-classifiers, i.e., all classification algorithms are of the same type, trained using different data. The training data for the base-classifiers are generated by sampling with replacement from the original training data. Thus, the base-classifiers learn on different subsets of the original training data and, therefore, predict test cases differently. By combining multiple base-classifiers trained on different data subsets, bagging reduces the variance of the individual base-classifiers, which is especially beneficial for unstable base-classifier algorithms (Witten and Frank 2005). Bagging combines the base-classifiers using average. I based my bagging implementation on Breiman (1996) and used decision trees as the base-classifiers, more specifically C4.5, and set the number of sampling iterations to 50. I then manipulated the size of each bag at 75%, 100% and 125% (the default, 100% was used in Breiman 1996), and whether to calculate the out-of-bag error (yes or no), for a total of six bagging configurations.

IMF is an ensemble combiner method that combines the base-classifier output using an information market based approach. The base-classifiers are implemented as software agents that participate in an information market. The information market uses a pari-mutuel betting market mechanism. In this market agents place bets on the true class of objects. The bets are based on agents' private probability estimates of object class membership and market odds, which specify the potential payout for winning bets. These market odds are a function of the total bets placed in

the market, thus IMF has to solve a recursive problem where agents place their bets based on market odds and where market odds are updated based on agents' bets. This problem is solved in IMF using binary search. The reciprocals of the market odds that minimize the difference between the total betting amount and the potential payouts for different classes represent the ensemble probability class estimates (see Chapter 2). Following Chapter 2, IMF was implemented using all the other experimental classifiers, including bagging, but excluding stacking.

4.3.2 Classification Cost

Given a binary problem like fraud, there are four potential classification outcomes: (1) True Positive (TP), a fraud firm is correctly classified as a fraud firm; (2) False Negative (FN), a fraud firm is incorrectly classified as a non-fraud firm; (3) True Negative (TN), a non-fraud firm is correctly classified as a non-fraud firm; and (4) False Positive (FP), a non-fraud firm is incorrectly classified as a fraud firm. Different classification costs are associated with TP, FN, TN and FP. TP and FP classifications have investigation costs C^i that are incurred in order to find out whether the firm was actually fraudulent. FN classifications have fraud costs C^f, i.e., we missed some fraudulent activity and the fraud is costly. FP classifications might have C^w costs related to wrongfully accusing a firm of fraud. All classifications have overhead costs, for example computer equipment, data loading, running the classification algorithm, etc. The ratio of these costs impact *training* and *evaluation* of classifiers (Provost et al. 1998). Two classifiers based on the same algorithm can produce different classification results if they are trained on data with different costs. The classifier configurations described in Section 4.3.1 were, therefore, further tuned by manipulating the relative error cost used when *training* the classifiers. However, as the Weka implementation of the classifiers examined in this essay were not cost-sensitive I under-sampled the class with the lower relative error cost. That is, I included fewer non-fraud firms in the sample, to achieve this objective.

When evaluating classifiers using specific assumptions about relative costs, the results might not generalize to the population of interest. The relative error cost used in the *evaluation*, therefore, has to be estimated to reflect the relative error costs in the population. These costs are very difficult to estimate accurately. Researchers, therefore, typically examine the classification performance over a wide range of relative error costs (Feroz et al. 2000; Lin et al. 2003), which reduces the risk of cost misspecification, and provides richer information to other researchers and practitioners. I followed prior research (Lin et al. 2003) and *evaluated* the classification performance over a wide range of relative error costs, specifically from 1:1 through 1:100. I also performed a focused analysis of classification performance using relative error cost estimates

from Bayley and Taylor (2007). Bayley and Taylor (2007) estimated that relative error costs were on average from 1:20 through 1:40. They based this estimate on an analysis of market reactions to fraud announcements reported by Beneish (1999). In the focused analysis, I used these estimates and examined the performance of the classifiers on three relative error costs 1:20, 1:30 and 1:40. Thus, this analysis provides insights into the relative performance of the classifiers under what is estimated to be realistic circumstances.

4.3.3 Prior Probability of Fraud

Like classification costs, the prior probability of fraud impacts both classifier training and evaluation. Two homogeneous classifiers can produce different results if they are trained on data with different prior probabilities, and the performance of a trained classifier can change if the prior probabilities change. The classifiers, therefore, have to be tuned by using different prior probabilities in the *training sample*. The classifier configurations described in Section 4.3.1 were, therefore, further tuned by manipulating the prior probability of fraud used when *training* the classifiers. Furthermore, to generalize to the population of interest, the prior probability of fraud in the *evaluation sample* should reflect the prior probability of fraud in the population. Prior financial statement fraud research has typically assumed that P(fraud) is 0.5 for training, evaluation or both. In reality, most organizations do not commit financial statement fraud. Bell and Carcello (2000) estimate that only around 0.6% of all firm years are fraudulent, i.e., P(fraud)=0.006. This estimate is however likely to change over time, and be different for different samples examined. I, therefore, manipulated the prior probability of fraud in the *evaluation sample* at three levels: low, medium and high. I defined medium as P(fraud) = 0.006, i.e., the estimate from Bell and Carcello (2000), low as 50% of medium or P(fraud) = 0.003, and high as 200% of medium or P(fraud) = 0.012.

4.3.4 Dependent Variable

Consistent with prior financial statement fraud research (Feroz et al. 2000; Lin et al. 2003), performance was measured using estimated relative cost (ERC). This measure was selected instead of net benefit (used in Essay I) in order to stay consistent with prior financial statement fraud research. The sum of ERC and net benefit per classified firm is equal to the average fraud cost of the classified firms (see Appendix 6), which is constant in a given dataset. Thus, as net benefit per classified firm increases ERC decreases by the same amount, and vice versa.

Given specific estimates of prior fraud probability and relative error costs for evaluation purposes, and specific classification results, ERC is calculated as:

$$ERC = n^{FN}/\, n^{P} \text{ x } C^{FN} \text{ x } \text{P(Fraud)} + n^{FP}/\, n^{N} \text{ x } C^{FP} \text{ x P(Non-Fraud)}, \quad (34)$$

where P(Fraud) and P(Non-Fraud) are the evaluation prior fraud and non-fraud probabilities, respectively; C^{FP} is the cost of false positive classifications (wrongful accusation costs, C^{w}, plus investigation cost, C^{i}) and C^{FN} is the cost of false negative classifications (fraud costs, C^{f}, minus investigation cost, C^{i}), both deflated by the lower of C^{FP} or C^{FN}; n^{FP} is the number of false positive classifications, n^{FN} is the number of false negative classifications, n^{P} is the number of positive instances in the dataset and n^{N} is the number of negative instances in the dataset. ERC is derived for each classification algorithm at the threshold that minimizes the ERC at a specific evaluation prior fraud probability and relative error cost.

4.3.5 Data

4.3.5.1 Classification Objects: Fraud and Non-Fraud Firms Data

The fraudulent observations were located based on firms investigated by the SEC for financial statement fraud and reported in Accounting and Auditing Enforcement Releases (AAER) from the fourth quarter of 1998 through the fourth quarter of 2005. A total of 745 potential observations were obtained from this initial search (see Table 4.1). The data set was then reduced by eliminating: duplicates; financial companies; firms without the first fraud year specified in the SEC release; non-annual financial statement fraud; foreign corporations; releases related to auditors; not-for-profit organizations; and fraud related to registration statements, 10-KSB or IPO. Financial companies were excluded from the sample as the rules and regulations governing financial firms are substantially different from other firms. Firms committing non-annual financial statement fraud were excluded as quarterly financial statements report financial information covering shorter time periods. Fraud related to registration statements were excluded as the purpose of these statements are different from annual financial statements reporting, and thus are likely to provide different incentives to commit fraud and to commit different types of fraud. An additional 75 fraud firms from Beasley (1996) were added to the remaining 197 fraud firms, for a total of 272 fraud firms. From these 272 fraud firms, 221 firms with missing data in Compustat or Compact D/SEC for the fraud year or four prior years, or with missing data in I/B/E/S for the fraud year, were deleted from the sample because these data were needed to create the measures described in section 4.3.5.2. For example, total discretionary accruals require data for five years, the current year and the four prior years, to calculate discretionary accruals for the current year and three prior years

To these remaining 51 fraud firms, I added 15,934 non-fraud firm years[31] to obtain P(fraud) ≈ 0.003 (0.00319). I used data from three sources to construct the object features: (1) financial statement data for the current year *t* and each of the four years leading up to the current year, t_{-1}, t_{-2}, t_{-3} and t_{-4}, were collected from Compustat; (2) one-year-ahead analyst earnings per share forecasts and actual earnings per share in the fraud year were collected from I/B/E/S; and (3) executive and director names, titles and company holdings were collected from Compact D/SEC.

4.3.5.2 Object Features – Financial Statement Fraud Predictors

Financial statement fraud predictor research has either been confirmatory or exploratory in nature. Confirmatory predictor research has focused on testing specific financial statement fraud hypotheses by developing and evaluating fraud predictors. The exploratory predictor research has taken a large number of variables, for example red flags proposed in SAS No. 53 and No. 82, and/or financial statement ratios, and either mapped these variables to fraud frameworks and/or tested their explanatory power. These two research streams have evaluated a large number of potential financial statement fraud predictors and found a number of significant financial statement fraud predictors, as shown in Table 4.2. I leveraged these findings and included in my experiment those predictors that had been found to be significant and that were easily available from electronic sources. Other variables were excluded since they were less likely to be used in practice due to the difficulty in obtaining them (these variables are italicized in Table 4.2). See Table 4.3 for the final selection of the 41 predictors included in the experiment and how these predictors were calculated.

[31] Note that matching is typically used to increase internal validity by controlling for variables not manipulated or measured in the experiment. My goal in this research is not to improve our understanding of factors that explain financial statement fraud, but rather to establish what classification algorithms are useful in predicting financial statement fraud. I, therefore, attempt to create a dataset that allows me to examine the performance impact of changing the prior probability of fraud during classifier training. Assuming that a lower prior probability in the training dataset than what is used for evaluation purposes will not improve performance when the minority class is already sparse, the minimum number of non-fraud firms is equal to the number of fraud firms divided by the lowest prior probability tested minus the number of fraud firms, i.e., (51/0.003)-51 = 16,949. Higher prior probabilities can then be obtained for training purposes by under-sampling the majority class, i.e., eliminating non-fraud firms from the sample.

Table 4.1
Sample Selection

Panel A: Fraud Firms	
Firms investigated by the SEC for fraudulent financial reporting from 4Q 1998 through 3Q 2005	745
Less: Financial companies	(33)
Less: Not annual (10-K) fraud	(116)
Less: Foreign companies	(9)
Less: Not-for-profit organizations	(10)
Less: Registration, 10-KSB and IPO related fraud	(78)
Less: Fraud year missing	(13)
Less: Duplicates	(287)
Remaining Fraud Observations	197
Add: Fraud firms from Beasley (1996)	75
Less: Not in Compustat or CompactD for first fraud year or four prior years or I/B/E/S for first fraud year	(221)
Usable Fraud Observations	51
Panel B: Non-Fraud Firms	
Non-Fraud Observations	15,934

4.4. Experimental Procedures and Preprocessing

4.4.1 Preprocessing

Before comparing the classifiers I determined the following: 1) prior fraud probability to use when training the classifiers; 2) method to use to filter the input data; 3) fraud predictors or attributes to include when training and evaluating the classifiers; and 4) how to tune the classifiers. Note that all these preprocessing steps were performed independently for each classifier. Thus, different training prior fraud probabilities, filtering methods and attributes, could be selected for different classifiers. I did not need to perform these steps for IMF since this technique took the output from the tuned classifiers at the evaluation stage and simply combined these outputs.

4.4.1.1 Training Data Prior Fraud Probability

In order to determine prior fraud probabilities for training the classifiers, the performance of the classifiers were compared at ten different prior fraud probability levels in the training set: 0.32%, 0.6%, 1%, 1.5%, 2.5%, 5%, 10%, 20%, 40% and 60%. Note that 0.3% was the lowest prior probability used in the evaluation data adjusted for a relative classification error cost of 1:1, and that 60% was the highest prior probability used in the evaluation, i.e., 1.2%, adjusted for a relative classification error cost of 1:100. At this initial step the classifiers were not tuned and

Table 4.2
Prior Research
Financial Statement Fraud Predictors

Author	*Dataset*[a]	*Determinants in Final Model*[b]	*Algorithm*[c]
Beasley (1996)	75 SEC fraud cases matched with 62 non-fraud cases	% outside directors	logit
Dechow et al. (1996)	92 SEC GAAP violators matched with 92 non-violators	value of issued securities to market value; total debt to total assets; demand for financing (ex ante); whether securities were issued; % insiders on board; insider holdings to total board holdings; *whether the board has an audit committee;* whether board has over 50% inside directors; *whether the CEO is the founder;* total accruals in year of manipulation	paired logit
Beneish (1997)	64 SEC fraud cases; 2,658 (1,989) aggressive accruers (with increasing sales)	days in receivables index; total accruals to total assets; positive accruals dummy	probit
Gerety and Lehn (1997)	62 SEC fraud cases matched with 62 non-fraud cases	Results did not show any significant determinants	paired
Green and Choi (1997)	86 SEC fraud cases matched with 86 non-fraud cases	AFDA to net sales; AFDA to accounts receivable; net sales to accounts receivable; gross margin to net sales; accounts receivable to total assets; net sales; accounts receivable; AFDA	ANN
Fanning and Cogger (1998)	102 SEC fraud cases matched with 102 non-fraud cases	% of outside directors; non-big X auditor; whether CFO changed in last three years; whether LIFO; debt to equity; sales to total assets; whether accounts receivable > 1.1 of last year's; whether gross margin % > 1.1 of last year's	ANN DA logit
Summers and Sweeney (1998)	51 WSJ fraud cases matched with 51 non-fraud cases	current minus prior year inventory to sales; prior year ROA to total assets current year	logit
Beneish (1999)	49 SEC fraud cases matched with 49 non-fraud cases	insider trading; *whether managers redeem stock appreciation rights;* holding period return in the violation period; discretionary accruals in violation period	probit

Table 4.2 (Continued)

Lee et al. (1999)	56 SEC fraud cases matched with 60,453 non-fraud cases	total accruals to total assets; total debt to total assets; whether new securities were issued; whether firm was listed on AMEX; whether SIC code >2999 and <4000	logit
Beasley (2000)	66 SEC fraud cases matched with unknown number of industry benchmark companies from National Association of Corporate Directors	*whether technology company and board has an audit committee;whether health care and audit committee has 100% outside directors*	uni-variate
Bell and Carcello (2000)	77 PRP fraud cases matched with 305 PRP non-fraud cases	*weak internal control environment;* rapid company growth; *undue emphasis on meeting earnings projections; management lied or was overly evasive;* whether company is public	logit
Feroz et al. (2000)	42 SEC fraud cases matched with 90 non-fraud cases	industry ROE minus firm ROE; times interest earned; accounts receivable to sales; Altman Z Score; the number of CEO turnovers; the number of CFO turnovers; the number of auditor turnovers	ANN logit
Lin et al. (2003)	40 SEC fraud cases matched with 160 non-fraud cases	net sales; accounts receivable; AFDA; AFDA to net sales; AFDA to accounts receivable; accounts receivable to net sales; accounts receivable to total assets; gross margin to net sales	ANN logit
Dunn (2003)	113 SEC and WSJ fraud cases matched with 113 non-fraud cases	*control philosophy*structure; motivation*	Logit
Kaminski et al. (2004)	79 SEC fraud cases matched with 79 non-fraud cases	fixed assets to total assets; sales to accounts receivable; inventory to current assets; inventory to sales; sales to total assets	DA
Uzun et al. (2004)	133 WSJ fraud cases matched with 133 non-fraud cases	% of outside directors; *% of gray audit committee directors; % of gray compensation committee directors; % of gray nominating committee directors*	Logit

Table 4.2 (Continued)

Chen and Sennetti (2005)	52 SEC fraud cases matched with 52 non-fraud cases	*research and development to sales*; gross profit margin; net profit margin; sales and marketing to sales; *tax benefits from exercising of employee stock options to operating cash flows;* changes in free cash flow; accounts receivable turnover; return on assets	Logit

[a] SEC = Dataset obtained from SEC releases; WSJ = Dataset obtained from Wall Street Journal news releases; and PRP = Dataset derived from proprietary sources

[b] Listed are variables that were: (1) significant in the primary multivariate analysis ($p<0.05$); or (2) included in the primary model if p-values were not reported in the multivariate analysis and the focus was on evaluating models or significant in univariate analyses if p-values were not reported in the multivariate analysis and the focus was on evaluating predictors. Variables in italics were relatively difficult to obtain and are, therefore, less likely to be used in actual, real world analyses. AFDA = Allowance for Doubtful Accounts; ROA = Return on Assets; and ROE = Return on Equity

[c] ANN = Artificial Neural Network; DA = Discriminant Analysis; and Paired = Paired t-test.

Table 4.3
Experimental
Financial Statement Fraud Predictors[a]

Variable	*Definition[f]*	*Datasource*
accounts receivable	(data2)	CompuSTAT
accounts receivable to sales	(data2/data12)	CompuSTAT
accounts receivable to total assets	(data2/data6)	CompuSTAT
AFDA	(data67)	CompuSTAT
AFDA to accounts receivable	(data67/data2)	CompuSTAT
AFDA to net sales	(data67/data12)	CompuSTAT
Altman Z score	3.3*(data18+data15+data16)/data6+0.999* data12/data6+0.6*data25*data199/data181+ 1.2*data179/data6+1.4*data36/data6	CompuSTAT
Big 4 auditor	IF 0 < data149 < 9 THEN 1 ELSE 0	CompuSTAT

Table 4.3 (continued)

current minus prior year inventory to sales	(data3)/(data12)-(data3$_{t-1}$)/(data12$_{t-1}$)	CompuSTAT
days in receivables index	(data2/data12)/(data2$_{t-1}$/data12$_{t-1}$)	CompuSTAT
debt to equity	(data181/data60)	CompuSTAT
demand for financing (ex ante)	IF ((data308-(data128$_{t-3}$+data128$_{t-2}$+ data128$_{t-1}$)/ 3) /(data4) < -0.5 THEN 1 ELSE 0	CompuSTAT
evidence of CEO change[b]	IF CEO_Name<>CEO_Name$_{t-1}$ OR CEO_Name$_{t-1}$<>CEO_Name$_{t-2}$ OR CEO_Name$_{t-2}$ <> CEO_Name$_{t-3}$ THEN 1 ELSE 0	CompactD
evidence of CFO change[c]	IF CFO_Name<>CFO_Name$_{t-1}$ OR CFO_Name$_{t-1}$ <>CFO_Name$_{t-2}$ OR CFO_Name$_{t-2}$ <> CFO_Name$_{t-3}$ THEN 1 ELSE 0	CompactD
fixed assets to total assets	data7/data6	CompuSTAT
four year geometric sales growth rate	(data12/data12$_{t-3}$)^(1/4)-1	CompuSTAT
gross margin to net sales	(data12-data41)/data12	CompuSTAT
holding period return in the violation period	(data199 – data199$_{t-1}$) / data199	CompuSTAT
industry ROE minus firm ROE	data172 / data 60	CompuSTAT
insider holdings to total board holdings	SUM(IF relationship code = CB, D, DO, H, OD THEN Insider_Holdings ELSE 0) / SUM(Insider_Holdings)	CompactD
inventory to sales	data3/data12	
net sales	data12	CompuSTAT
positive accruals dummy	IF (data18-data308) > 0 and (data18$_{t-1}$-data308$_{t-1}$) > 0 THEN 1 ELSE 0	CompuSTAT
percentage officers on the board of directors[d]	SUM(IF Executive_Name = Director_Name THEN 1 ELSE 0) / Number_Of_Directors	CompactD
prior year ROA to total assets current year	(data172$_{t-1}$ / data 6$_{t-1}$) / data6	CompuSTAT
property plant and equipment to sales	data8/data12	CompuSTAT

Table 4.3 (continued)

sales to total assets	data12/data6	CompuSTAT
the number of auditor turnovers	IF data149<>$data149_{t-1}$ THEN 1 ELSE 0 + IF $data149_{t-1}$<>$data149_{t-2}$ THEN 1 ELSE 0 + IF $data149_{t-2}$<>$data149_{t-3}$ THEN 1 ELSE 0	CompuSTAT
times interest earned	(data18+data15+data16) / data15	CompuSTAT
total accruals to total assets[e]	(data18-data308) / data6	CompuSTAT
total debt to total assets	data181/data6	CompuSTAT
total discretionary accrual	DA_{t-1} + DA_{t-2} + DA_{t-3}, where DA = TA/A-estimated(NDA); TA/A = (data18-data308) /$data6_{t-1}$; NDA = 1/$data6_{t-1}$ + (data12 - $data12_{t-1}$ - data2 + $data2_{t-1}$)/$data6_{t-1}$+ (data308 - $data308_{t-1}$)/$data6_{t-1}$ + data7/$data6_{t-1}$	CompuSTAT
unexpected employee productivity	FIRM((data12/data29 - $data12_{t-1}$/$data29_{t-1}$)/($data12_{t-1}$/$data29_{t-1}$)) - INDUSTRY((data12/data29 - $data12_{t-1}$/$data29_{t-1}$)/($data12_{t-1}$/$data29_{t-1}$))	CompuSTAT
value of issued securities to market value	IF data396 > 0 THEN data396*data199/(data25*data199) ELSE IF (data25-$data25_{t-1}$)>0 THEN ((data25 - $data25_{t-1}$)*data199) / (data25*data199) ELSE 0	CompuSTAT
whether accounts receivable > 1.1 of last year's	IF (data2/$data2_{t-1}$) > 1.1 THEN 1 ELSE 0	CompuSTAT
whether firm was listed on AMEX	IF ZLIST=5, 15, 16, 17, 18 THEN 1 ELSE 0	CompuSTAT
whether gross margin % > 1.1 of last year's	IF ((data12-data41) / data12) / (($data12_{t-1}$ - $data41_{t-1}$)/$data12_{t-1}$) > 1.1 THEN 1 ELSE 0	CompuSTAT
whether LIFO	IF data59=2 THEN 1 ELSE 0	CompuSTAT
whether meeting or beating analyst forecast	IF EPS - Analyst_Forecast >= 0 THEN 1 ELSE 0	I/B/E/S
whether new securities were issued	IF (data25-$data25_{t-1}$)>0 OR data396>0 THEN 1 ELSE 0	CompuSTAT

Table 4.3 (continued)

whether SIC code larger (smaller) than 2999 (4000)	IF 2999<DNUM<4000 THEN 1 ELSE 0	CompuSTAT

[a] Showing all predictors found to be significant determinants of financial statement fraud in prior research and that were relatively essay to obtained.

[b] Because of the similarity between *evidence of CEO change* and the *number of CEO turnovers*, the *number of CEO turnovers* was excluded.

[c] Because of the similarity between *evidence of CFO change* and the *number of CFO turnovers*, the *number of CFO turnovers* was excluded.

[d] Because of the similarity between the *percentage officers on the board of directors* and *percentage insiders on board*, *percentage insiders on board* was excluded.

[e] Because of the similarity between *total accruals to total assets* and *discretionary accruals in violation period*, and between *total discretionary accruals* and *discretionary accruals in violation period*, and because the violation period was not known in the sample, and because there was no violation period for non-fraud firms as non-fraud firms were not matched with fraud firms, *discretionary accruals in violation period* was excluded.

[f] data# refers to specific items in CompuSTAT based on the numbering system in existence as of April 17, 2008.

were instead implemented using their default settings as described in the classifier tuning section. For each classifier, the prior fraud probability of the training dataset that produced the lowest classification error cost was selected in each evaluation prior fraud probability and relative error cost treatment group. Note that optimal decision thresholds were used when calculating the ERC for each classifier and treatment group. This threshold was determined empirically by calculating the ERC for each classifier at each treatment group 101 times as the threshold was changed from 0 to 1 in 0.01 increments. Thus, in this experiment a total of 13,332 ERC were derived (101 decision thresholds times 11 relative error cost treatment levels times three evaluation prior fraud probabilities times four classifiers).

As seen in Table 4.4, ANN minimized ERC at a training prior fraud probability of 0.6% for evaluation cost ratios from 1:1 through 1:50, 1:1 through 1:20 and 1:1 through 1:10, and at an evaluation prior fraud probability of 0.3%, 0.6% and 1.2%, respectively. For the remaining evaluation cost ratio and prior fraud probability levels, ANN minimized ERC using a training set with 60% prior fraud probability. All the other algorithms also minimized ERC using two or more prior fraud probabilities in the training set, as shown in Table 4.4. In general, the results show, as expected, that the optimal prior fraud probability level in the training set increased as the evaluation relative cost and prior probability of fraud treatments increased.

Table 4.4
Training Prior Fraud Probabilities:
Selected Training Prior Fraud Probabilities for each Classifier at Different Levels of Evaluation Prior Fraud Probability and Evaluation Relative Error Cost[a]

Evaluation Factors		*Classifiers*					
Relative Error Cost	*Prior Fraud Probability*	*ANN*	*SVM*	*C4.5*	*Logistic*	*Bagging*	*Stacking*
1:1	0.003	0.006	0.2	0.05	0.015	0.6	0.6
1:10	0.003	0.006	0.2	0.05	0.015	0.6	0.6
1:20	0.003	0.006	0.2	0.05	0.015	0.6	0.6
1:30	0.003	0.006	0.2	0.05	0.015	0.6	0.6
1:40	0.003	0.006	0.2	0.05	0.015	0.6	0.6
1:50	0.003	0.006	0.2	0.05	0.015	0.6	0.6
1:60	0.003	0.6	0.2	0.05	0.015	0.6	0.6
1:70	0.003	0.6	0.2	0.05	0.015	0.6	0.6
1:80	0.003	0.6	0.2	0.05	0.015	0.6	0.6
1:90	0.003	0.6	0.2	0.05	0.015	0.6	0.6
1:100	0.003	0.6	0.6	0.05	0.015	0.6	0.6
1:1	0.006	0.006	0.2	0.05	0.015	0.6	0.6
1:10	0.006	0.006	0.2	0.05	0.015	0.6	0.6
1:20	0.006	0.006	0.2	0.05	0.015	0.6	0.6
1:30	0.006	0.6	0.2	0.05	0.015	0.6	0.6
1:40	0.006	0.6	0.2	0.05	0.015	0.6	0.6
1:50	0.006	0.6	0.6	0.05	0.015	0.6	0.6
1:60	0.006	0.6	0.6	0.1	0.015	0.6	0.6
1:70	0.006	0.6	0.6	0.1	0.015	0.6	0.6
1:80	0.006	0.6	0.6	0.1	0.015	0.6	0.6
1:90	0.006	0.6	0.6	0.1	0.015	0.6	0.6
1:100	0.006	0.6	0.6	0.4	0.015	0.6	0.6
1:1	0.012	0.006	0.2	0.1	0.015	0.6	0.6
1:10	0.012	0.006	0.2	0.1	0.015	0.6	0.6
1:20	0.012	0.6	0.2	0.1	0.015	0.6	0.6
1:30	0.012	0.6	0.6	0.1	0.015	0.6	0.6
1:40	0.012	0.6	0.6	0.1	0.015	0.6	0.6
1:50	0.012	0.6	0.6	0.4	0.015	0.6	0.6
1:60	0.012	0.6	0.6	0.4	0.1	0.6	0.6
1:70	0.012	0.6	0.6	0.4	0.1	0.6	0.6
1:80	0.012	0.6	0.6	0.4	0.1	0.6	0.6
1:90	0.012	0.6	0.6	0.6	0.2	0.6	0.6
1:100	0.012	0.6	0.6	0.6	0.2	0.6	0.6

[a] For each evaluation treatment group (two columns to the left) the classifiers were evaluated using different prior fraud probabilities in the training dataset. The training dataset prior fraud probability that generated the lowest ERC for each classifier in each evaluation treatment group was then selected. This probability is shown for each classifier and treatment group (six columns to the right)

4.4.1.2 Data Filtering

I continued the preprocessing by evaluating whether filtering the data, using one of three filtering methods that transformed the continuous fraud predictors, improved classifier performance. These methods normalized, discretized and standardized the data. The utility of these methods and no filter were compared for each classifier at the training prior fraud probabilities that minimized ERC (see Table 4.4) at a cost ratio of 1:50 and a prior fraud probability of 0.3%, i.e., the median treatment level of the two evaluation factors. To discretize the attributes, I used the PKIDiscretize procedure in Weka, which implements equal frequency binning with the number of bins set to the square root of the number of non-missing values. This approach has been shown to produce improved classification results (Witten and Frank 2005). The standardized data were obtained by subtracting attribute means from instance values and then dividing this difference by the standard deviation of the attribute. To normalize the data, the difference between each instance value and the minimum instance value was divided by the range of the attribute values, i.e., maximum minus minimum value. The standardized attributes had mean of zero and a standard deviation of one, while the normalized attributes had values that were between zero and one.

The results reported in Table 4.5 show a relatively clear trend indicating that classifiers trained with data that were normalized and not filtered produced lower ERC than classifiers trained with data that were standardized, which in turn produced lower ERC than classifiers trained with data that were discretized. More specifically, the performance benefits, measured using ERC, of normalization, standardization and no filter were the same for both ANN and logistic regression, and were better than discretization in 23 and 27 out of 33 comparisons, respectively. For C4.5 no filter was better than or as good as the other methods in 31 out of 33 comparisons, while normalization was better than or as good as the other methods at all evaluation levels for SVM. Normalization was also the best approach for stacking, for which normalization was better than or as good as the other methods in 30 out of 33 comparisons. Finally, no filter was superior in 22 out of 33 comparisons, and inferior to normalization and standardization in the remaining 11 comparisons for bagging.

4.4.1.3 Fraud Predictor Utility

One of my research objectives was to improve our understanding of what predictors provide utility to the different classifiers. Answering this question can facilitate more efficient data collection as predictors that provide little or no utility to the classifiers do not have to be collected. Furthermore, this knowledge can provide the foundation for reducing the dataset dimensionality (reducing the number of attributes), which can improve the performance of the

Table 4.5
Data Filtering:
ERC for each Combination of Classifier and Data Filtering Method at Different Levels of Evaluation Prior Fraud Probability and Evaluation Relative Cost[a]

		ANN				*Logistic Regression*			
Relative Error Cost	*Prior Fraud Probability*	*Norm-alized*	*Stand-ardized*	*Discr-etized*	*No Filter*	*Norm-alized*	*Stand-ardized*	*Discr-etized*	*No Filter*
1:1	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
1:10	0.003	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
1:20	0.003	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
1:30	0.003	0.090	0.090	0.090	0.090	0.088	0.088	0.090	0.088
1:40	0.003	0.120	0.120	0.120	0.120	0.115	0.115	0.120	0.115
1:50	0.003	0.150	0.150	0.150	0.150	0.142	0.142	0.150	0.142
1:60	0.003	0.168	0.168	0.180	0.168	0.169	0.169	0.180	0.169
1:70	0.003	0.186	0.186	0.203	0.186	0.193	0.193	0.210	0.193
1:80	0.003	0.205	0.205	0.223	0.205	0.217	0.217	0.240	0.217
1:90	0.003	0.223	0.223	0.244	0.223	0.241	0.241	0.270	0.241
1:100	0.003	0.241	0.241	0.265	0.241	0.265	0.265	0.300	0.265
1:1	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
1:10	0.006	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
1:20	0.006	0.120	0.120	0.120	0.120	0.115	0.115	0.120	0.115
1:30	0.006	0.168	0.168	0.180	0.168	0.169	0.169	0.180	0.169
1:40	0.006	0.204	0.204	0.223	0.204	0.217	0.217	0.240	0.217
1:50	0.006	0.241	0.241	0.264	0.241	0.265	0.265	0.300	0.265
1:60	0.006	0.277	0.277	0.306	0.277	0.313	0.313	0.360	0.313
1:70	0.006	0.314	0.314	0.347	0.314	0.362	0.362	0.420	0.362
1:80	0.006	0.350	0.350	0.388	0.350	0.409	0.409	0.480	0.409
1:90	0.006	0.387	0.387	0.429	0.387	0.454	0.454	0.540	0.454
1:100	0.006	0.411	0.411	0.470	0.411	0.500	0.500	0.600	0.500
1:1	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012
1:10	0.012	0.120	0.120	0.120	0.120	0.115	0.115	0.120	0.115
1:20	0.012	0.204	0.204	0.223	0.204	0.217	0.217	0.240	0.217
1:30	0.012	0.277	0.277	0.305	0.277	0.313	0.313	0.360	0.313
1:40	0.012	0.350	0.350	0.388	0.350	0.408	0.408	0.480	0.408
1:50	0.012	0.410	0.410	0.470	0.410	0.500	0.500	0.600	0.500
1:60	0.012	0.457	0.457	0.552	0.457	0.590	0.590	0.720	0.590
1:70	0.012	0.504	0.504	0.635	0.504	0.680	0.680	0.831	0.680
1:80	0.012	0.551	0.551	0.717	0.551	0.769	0.769	0.936	0.769
1:90	0.012	0.598	0.598	0.775	0.598	0.858	0.858	1.040	0.858
1:100	0.012	0.645	0.645	0.812	0.645	0.947	0.947	1.143	0.947

Table 4.5 (Continued)

		C4.5				*Support Vector Machines*			
Relative Error Cost	*Prior Fraud Probability*	*Norm-alized*	*Stand-ardized*	*Discr-etized*	*No Filter*	*Norm-alized*	*Stand-ardized*	*Discr-etized*	*No Filter*
1:1	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
1:10	0.003	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
1:20	0.003	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
1:30	0.003	0.090	0.090	0.090	0.090	0.090	0.090	0.090	0.090
1:40	0.003	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120
1:50	0.003	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150
1:60	0.003	0.180	0.180	0.180	0.180	0.180	0.180	0.180	0.180
1:70	0.003	0.210	0.210	0.210	0.210	0.210	0.210	0.210	0.210
1:80	0.003	0.240	0.240	0.240	0.240	0.239	0.240	0.240	0.239
1:90	0.003	0.270	0.270	0.270	0.270	0.257	0.270	0.270	0.257
1:100	0.003	0.300	0.300	0.300	0.300	0.276	0.300	0.300	0.276
1:1	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
1:10	0.006	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
1:20	0.006	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120
1:30	0.006	0.180	0.180	0.180	0.180	0.180	0.180	0.180	0.180
1:40	0.006	0.240	0.240	0.240	0.240	0.238	0.240	0.240	0.238
1:50	0.006	0.300	0.300	0.300	0.300	0.276	0.300	0.300	0.276
1:60	0.006	0.360	0.360	0.360	0.357	0.314	0.360	0.360	0.314
1:70	0.006	0.420	0.418	0.420	0.415	0.348	0.420	0.420	0.348
1:80	0.006	0.480	0.476	0.480	0.472	0.380	0.480	0.480	0.380
1:90	0.006	0.540	0.533	0.540	0.530	0.413	0.540	0.540	0.413
1:100	0.006	0.600	0.591	0.600	0.588	0.446	0.600	0.600	0.446
1:1	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012
1:10	0.012	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120
1:20	0.012	0.240	0.240	0.240	0.240	0.238	0.240	0.240	0.238
1:30	0.012	0.360	0.360	0.360	0.357	0.313	0.360	0.360	0.313
1:40	0.012	0.480	0.475	0.480	0.472	0.380	0.480	0.480	0.380
1:50	0.012	0.599	0.591	0.600	0.588	0.446	0.600	0.600	0.446
1:60	0.012	0.717	0.706	0.720	0.703	0.512	0.720	0.686	0.512
1:70	0.012	0.830	0.817	0.840	0.814	0.577	0.787	0.752	0.577
1:80	0.012	0.943	0.928	0.960	0.924	0.628	0.824	0.818	0.643
1:90	0.012	1.004	0.966	0.988	0.981	0.677	0.862	0.884	0.709
1:100	0.012	1.006	0.966	0.988	0.981	0.727	0.900	0.949	0.761

Table 4.5 (Continued)

Relative Error Cost	*Prior Fraud Probability*	*Bagging*				*Stacking*			
		Norm-alized	*Stand-ardized*	*Discr-etized*	*No Filter*	*Norm-alized*	*Stand-ardized*	*Discr-etized*	*No Filter*
1:1	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
1:10	0.003	0.029	0.030	0.029	0.029	0.030	0.030	0.030	0.030
1:20	0.003	0.059	0.060	0.059	0.058	0.060	0.060	0.060	0.060
1:30	0.003	0.088	0.090	0.088	0.086	0.090	0.090	0.090	0.090
1:40	0.003	0.118	0.120	0.118	0.115	0.120	0.120	0.120	0.120
1:50	0.003	0.147	0.150	0.147	0.144	0.150	0.150	0.150	0.150
1:60	0.003	0.176	0.180	0.176	0.173	0.180	0.180	0.180	0.180
1:70	0.003	0.202	0.206	0.206	0.202	0.210	0.210	0.210	0.210
1:80	0.003	0.227	0.230	0.235	0.220	0.235	0.235	0.240	0.234
1:90	0.003	0.244	0.244	0.265	0.236	0.250	0.250	0.270	0.252
1:100	0.003	0.258	0.258	0.294	0.253	0.264	0.264	0.300	0.270
1:1	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
1:10	0.006	0.059	0.060	0.059	0.058	0.060	0.060	0.060	0.060
1:20	0.006	0.118	0.120	0.118	0.115	0.120	0.120	0.120	0.120
1:30	0.006	0.176	0.180	0.176	0.173	0.180	0.180	0.180	0.180
1:40	0.006	0.227	0.230	0.235	0.219	0.235	0.235	0.240	0.234
1:50	0.006	0.258	0.258	0.293	0.252	0.264	0.264	0.300	0.270
1:60	0.006	0.286	0.286	0.329	0.285	0.293	0.293	0.360	0.307
1:70	0.006	0.315	0.315	0.364	0.318	0.323	0.323	0.420	0.343
1:80	0.006	0.343	0.343	0.399	0.351	0.352	0.352	0.480	0.379
1:90	0.006	0.369	0.369	0.435	0.384	0.382	0.382	0.540	0.416
1:100	0.006	0.393	0.393	0.470	0.417	0.411	0.411	0.593	0.452
1:1	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012
1:10	0.012	0.118	0.120	0.118	0.115	0.120	0.120	0.120	0.120
1:20	0.012	0.227	0.229	0.235	0.219	0.234	0.234	0.240	0.233
1:30	0.012	0.286	0.286	0.328	0.285	0.293	0.293	0.360	0.306
1:40	0.012	0.342	0.342	0.399	0.351	0.352	0.352	0.480	0.379
1:50	0.012	0.392	0.392	0.469	0.417	0.410	0.410	0.591	0.452
1:60	0.012	0.442	0.442	0.540	0.482	0.469	0.469	0.657	0.525
1:70	0.012	0.491	0.491	0.610	0.548	0.525	0.528	0.723	0.574
1:80	0.012	0.541	0.541	0.681	0.614	0.563	0.582	0.789	0.619
1:90	0.012	0.590	0.590	0.700	0.672	0.600	0.622	0.854	0.664
1:100	0.012	0.639	0.639	0.717	0.714	0.638	0.662	0.920	0.709

[a] For each evaluation treatment group (two columns to the left) the classifiers were evaluated using different data filtering methods. The data filtering method that generated the lowest ERC for each classifier in each evaluation treatment group was then selected. The ERC are displayed for each classifier, data filtering method and evaluation treatment group (eight columns to the left). The lowest ERC for each classifier and treatment group is highlighted in grey.

classifiers. To reduce the dimensionality, I used a Wrapper attribute selection technique, which has been shown to be effective (Hall and Holmes 2003). The Wrapper approach examines the utility of the different attributes to a specific algorithm, as opposed to attribute selection techniques that examine the attributes without considering the specific classifier that will use the attributes. To evaluate the utility of the attributes, the Wrapper uses internal cross-validation iterations to compare the accuracy of a classifier using different sets of attributes. However, as discussed earlier, accuracy is not a good measure of performance in the fraud domain unless the training prior fraud probability is altered to take into account the actual prior fraud probability and relative costs in the domain. Assuming an average prior fraud probability of 0.006 (Bell and Carcello 2000) and an average relative cost of 1:30 (Bayley and Taylor 2007), I used a dataset with 51 fraud firms and 283 non-fraud firms for a prior fraud probability of 0.18 (calculated as 0.006*30). A genetic search algorithm was used within the Wrapper to search for the optimal attribute set. To evaluate the robustness of this search, I used 10-fold cross-validation to examine the selected attributes (note that the Wrapper also uses 5-fold cross-validation internally). In this examination I normalized the dataset attributes for SVM, logistic regression, stacking and ANN, and used non-filtered dataset attributes for C4.5 and bagging.

For logistic regression, nine variables were selected in at least 40% of the folds: the number of auditor turnovers, total discretionary accruals, Big 4 auditor, accounts receivables, allowance for doubtful accounts, whether meeting or beating analyst forecasts, inventory to sales, unexpected employee productivity and value of issued securities to market value (see Table 4.6). Note that an additional twelve variables would have been added to this selection if variables selected in at least 30% of the folds had been included. The SVM Wrapper selected six of the variables selected for logistic regression: the number of auditor turnovers, total discretionary accruals, Big 4 auditor, accounts receivables, allowance for doubtful accounts and whether meeting or beating analyst forecasts. The SVM Wrapper additionally selected Altman Z score, percentage of executives on the board of directors, property plant and equipment to sales, fixed assets to total assets, allowance for doubtful accounts to accounts receivable, and total debt to total assets in at least 40% of the folds (10 additional variables in at least 30% of the folds). For C4.5, auditor turnover, Big 4 auditor and whether meeting or beating analyst forecasts were again selected, as well as accounts receivable to total assets, accounts receivable to sales, gross margin to net sales, property plant and equipment to sales, industry ROE minus firm ROE, and positive accruals dummy in at least 40% of the folds (10 additional variables in at least 30% of the folds). The results for ANN, bagging and stacking also showed some overlap with the variables selected for logistic regression, SVM and C4.5.

Table 4.6
Attribute Selection:
The Percentage of Folds in which
Predictor was Selected for Each Classifier[a]

Predictor	*SVM*	*Log[b]*	*ANN*	*C4.5*	*Bag[c]*	*Stack[d]*	*Avg[e]*
the number of auditor turnovers	**70%**	**40%**	10%	**70%**	**70%**	0%	46%
total discretionary accruals	**60%**	**60%**	**40%**	30%	**30%**	0%	39%
Big 4 auditor	**50%**	**40%**	**30%**	**40%**	**50%**	0%	38%
accounts receivable	**70%**	**50%**	**30%**	0%	10%	**100%**	35%
allowance for doubtful accounts	**60%**	**80%**	10%	20%	20%	0%	34%
accounts receivable to total assets	30%	30%	**60%**	**40%**	20%	0%	33%
accounts receivable to sales	20%	20%	20%	**60%**	**50%**	0%	31%
whether meeting or beating forecast	**50%**	**40%**	20%	**40%**	10%	0%	29%
evidence of CEO change	20%	30%	**30%**	30%	10%	**100%**	28%
sales to total assets	30%	30%	20%	30%	10%	**100%**	28%
inventory to sales	30%	**50%**	10%	30%	0%	**100%**	28%
unexpected employee productivity	20%	**40%**	**30%**	20%	**30%**	0%	26%
Altman Z score	**60%**	30%	0%	20%	20%	0%	24%
percentage of executives on the board of directors	**40%**	30%	10%	30%	20%	0%	24%
demand for financing (ex ante)	30%	30%	20%	30%	10%	0%	23%
if account receivable grew by more than 10%	20%	30%	**50%**	0%	20%	0%	23%
allowance for doubtful accounts to net sales	20%	20%	10%	30%	0%	**100%**	21%
current minus prior year inventory to sales	0%	30%	10%	10%	**30%**	**100%**	21%
gross margin to net sales	20%	10%	0%	**40%**	10%	**100%**	21%
evidence of CFO change	20%	20%	**40%**	30%	0%	0%	21%
holding period return in the violation period	30%	30%	**40%**	10%	0%	0%	21%
property plant and equipment to sales	**40%**	10%	20%	**40%**	0%	0%	21%
value of issued securities to market value	30%	**50%**	20%	0%	10%	0%	21%
fixed assets to total assets	**60%**	30%	0%	0%	10%	0%	19%
days in receivables index	0%	20%	20%	20%	0%	**100%**	18%
four year geometric sales growth rate	30%	30%	20%	10%	0%	0%	18%
industry ROE minus firm ROE	0%	0%	20%	**40%**	**30%**	0%	18%
positive accruals dummy	20%	10%	10%	**50%**	0%	0%	18%
times interest earned	30%	10%	10%	30%	10%	0%	18%
if firm was listed on AMEX	20%	20%	20%	10%	10%	0%	16%
if gross margin grew by more than 10%	20%	10%	**30%**	10%	10%	0%	16%

Table 4.6 (continued)

if new securities were issued	30%	10%	10%	10%	20%	0%	16%
allowance for doubtful accounts to accounts receivable	**40%**	10%	0%	20%	10%	0%	16%
debt to equity	20%	10%	10%	30%	10%	0%	16%
total debt to total assets	**40%**	10%	20%	10%	0%	0%	16%

[a] The percentage of folds in which the Wrapper included the predictor in the final set of predictors. The percentages in bold show which predictors were selected to be included in the final dataset for each classifier.

[b] Log = logistic regression

[c] Bag = bagging

[d] Stack = stacking. Stacking is relatively computationally expensive as it uses all the other classifiers as base-classifiers. When using the Wrapper with a genetic search algorithm, the Wrapper runs stacking using an external, in addition to the internal, cross-validation with genetic search in each fold. This procedure becomes very computationally expensive and external cross-validation was, therefore, not performed for stacking.

[e] Average percentage of folds in which the Wrapper included the predictor in the final set of predictors. To compute this average, selected stacking predictors were assigned a percentage of 0.17 if they were not selected in the tested fold and 0.47 if they were selected in the tests fold. These percentages were the averages of all the selected and not selected predictors for the other classifiers.

It is interesting to note that no variables were selected by all algorithms. Whether a firm had a Big 4 auditor was selected by all classifiers but stacking, while auditor turnover, total discretionary accruals and accounts receivable were selected by four out of six classifiers. Meeting or beating analyst forecasts and unexpected employee productivity were selected by three out of six classifiers. All the other predictors were selected by less than 50% of the classifiers. In terms of the average number of folds, auditor turnover was selected by the Wrapper most consistently and was on average selected in about half the folds examined (46%). Total discretionary accruals and Big 4 auditor followed closely at 39% and 38%, respectively.

Henceforth, I used the training prior fraud probabilities that minimized ERC at the different evaluation prior fraud probability and relative error cost treatment levels (see Table 4.4). For the different classifiers I filtered the data as follows: 1) normalized the data for SVM, logistic regression and ANN; 2) standardized and used no filter for C4.5; and 3) normalized and used no filter for stacking and bagging. Finally, for each classifier the data dimensionality was reduced by only using the classifier specific Wrapper selected attributes. These preprocessing results are summarized in Table 4.7.

4.4.2 Classifier Evaluation

For classifier tuning purposes, I examined the C4.5, SVM, ANN, logistic regression, stacking, bagging and IMF classifier configurations described in Section 4.3.1.1. The different

Table 4.7
Preprocessing Result Overview: Selected Training Prior Fraud Probabilities, Data Filtering Methods and Predictors

Classifiers	*Training Prior Fraud Probability*	*Data Filtering*	*Predictors*[a]
ANN	0.2, 0.6	Normalize	1, 2, 3, 4, 5, 8, 13, 14, 22, 24, 33, 35
SMO	0.006, 0.6	Normalize	2, 3, 4, 6, 9, 12, 16, 20, 21, 31
C4.5	0.05, 0.1, 0.4, 0.6	No Filter, Standardize	1, 3, 6, 7, 8, 19, 22, 27, 28
Logistic	0.015, 0.1, 0.2	Normalize	1, 2, 3, 4, 5, 8, 11, 12, 23
Bagging	0.6	No Filter, Normalize	1, 2, 3, 7, 12, 18, 27
Stacking	0.6	No Filter, Normalize	4, 9, 10, 11, 17, 18, 19, 25

[a] Predictor numbers represent the following predictors: 1=the number of auditor turnovers, 2=total discretionary accruals, 3=Big 4 auditor, 4=accounts receivable, 5=allowance for doubtful accounts, 6=accounts receivable to total assets, 7=accounts receivable to sales, 8=whether meeting or beating forecast, 9=evidence of CEO chance, 10=sales to total assets, 11=inventory to sales, 12=unexpected employee productivity, 13=Altman Z score, 14=percentage of executives on the board of directors, 15=demand for financing (ex ante), 16=whether account receivable grew by more than 10%, 17=allowance for doubtful accounts to net sales, 18=current minus prior year inventory to sales, 19=gross margin to net sales, 20=evidence of CFO chance, 21=holding period return in the violation period, 22=property plant and equipment to sales, 23=value of issued securities to market value, 24=fixed assets to total assets, 25=days in receivables index, 26=four year geometric sales growth rate, 27=Industry ROE minus firm ROE, 28=positive accruals dummy, 29=times interest earned, 30=whether firm was listed on AMEX, 31=whether gross margin grew by more than 10%, 32=whether new securities were issued, 33=allowance for doubtful accounts to accounts receivable, 34=debt to equity, and 35=total debt to total assets.

configurations were evaluated using ten-fold stratified cross validation with the preprocessing dataset stratified and then randomly split into ten mutually exclusive folds of approximately equal size that each contained approximately the same prior class probabilities as the original dataset. The classifiers were then trained and evaluated ten times, each time using a different fold for evaluation and the nine remaining folds for training. The classifier tuning result set is a combination of the results from all ten evaluation folds. For each classifier type, I compared the different classifier configurations using these result sets and selected the configuration with the lowest ERC for each relative error costs and evaluation prior fraud probability combination.

Using the selected configurations, the ten-fold stratified cross validation was repeated ten times. The cross-validation results from the ten iterations were then used to calculate ten ERC scores for each classifier configuration, relative error cost and evaluation prior fraud probability combination. The final results set used for classifier evaluation were generated by taking the ten ERC measures of each classifier type, relative error costs and evaluation prior fraud probability combination generated by the configuration selected for this specific experimental manipulation. Thus, I did not necessarily take the configuration with the best final results for a specific combination of relative error cost and evaluation prior fraud probability, but instead used the

results from the preselected classifier configurations. The final result set contained ten observations per classifier type, relative error costs and prior fraud probability treatment groups.

4.5. Results

Table 4.8 reports descriptive classifier performance statistics. The reported estimated relative cost is the average for each classifier at all treatment levels. Thus, it is not surprising that the range of ERC is high and that the standard deviation is almost as high as the mean. For example, the standard deviation and mean ERC for logistic regression are 0.2367 and 0.2916, respectively. The descriptive statistics provide an initial indication that logistic regression, bagging and SVM perform well. Logistic regression performs particularly well, performing significantly[32] better ($p<0.05$) than ANN, IMF, C4.5 and stacking. It is, however, important to remember that these are descriptive statistics that report on the performance of the classifiers on average. Thus, we do not know under what specific evaluation prior fraud probabilities and relative cost conditions logistic regression, bagging and SVM outperform the other algorithms, and even if perhaps these other algorithms are better performers under certain conditions.

To determine whether the differences noticed in the descriptive statistics depend on the level of evaluation prior fraud probabilities or relative cost, I examined the interactions between prior

Table 4.8
Descriptive Statistics of Classifier Estimate Relative Cost[a]

Classifier	*Min*	*Median*	*Max*	*Mean*	*Standard Deviation*	*Connecting Letters Report[b]*
Logistic	0.0026	0.2167	0.9100	0.2916	0.2367	D
Bagging	0.0028	0.2400	0.8858	0.2978	0.2275	D C
SVM	0.0025	0.2306	0.8946	0.2989	0.2453	D C
ANN	0.0030	0.2400	0.8912	0.3046	0.2320	C
IMF	0.0026	0.2400	0.9880	0.3053	0.2463	C
C45	0.0028	0.2400	1.0614	0.3301	0.2734	B
Stacking	0.0030	0.2400	0.9880	0.3414	0.2905	A

[a] Classifier performance is measured using Estimated Relative Cost. Note that lower values are preferred over higher values.

[b] Levels not connected by the same letter are significantly different at a p-value of 0.05 using Tukey-Kramer HSD and blocking for the effect of evaluation prior fraud probability and relative cost on estimated relative cost.

[32] Significance determined using Tukey-Kramer HSD and blocking for the effect of evaluation prior fraud probability and relative cost on estimated relative cost.

fraud probability and classification algorithm, and between relative error cost and classification algorithm using the following regression model:

$$\begin{aligned} ERC = {} & \alpha_0 + \alpha_1 \text{Classification Algorithm} + \alpha_2 \text{Prior Fraud Probability} + \\ & \alpha_3 \text{Relative Error Cost} + \alpha_4 \text{Classification Algorithm*} \\ & \text{Prior Fraud Probability} + \alpha_5 \text{Classification Algorithm*} \\ & \text{Relative Error Cost} + \varepsilon \end{aligned} \quad (35)$$

The interaction between prior fraud probability and classification algorithm ($p<0.001$), and the interaction between relative error cost and classification algorithm ($p<0.001$) were both significant. Thus, the relative performance of the classifiers depends on the level of evaluation prior fraud probability and on the level of relative error cost. The parameter estimates reported in Table 4.9 show that as prior fraud probability increases, the performance of bagging ($p=0.005$), ANN ($p=0.039$) and logistic regression ($p=0.039$) improves relative to the other classifiers (the magnitude of the performance improvement is in the order listed), the relative performance of IMF ($p=0.347$) and SVM ($p=0.175$) does not change, while the relative performance of stacking ($p<0.001$) and C4.5 ($p<0.001$) deteriorates. The change in relative performance is similar when relative error cost increases; i.e., as the cost of FN errors becomes higher relative to the cost of FP. It is also interesting to note that the intercepts of bagging ($p=0.007$), logistic regression ($p<0.001$) and SVM ($p=0.007$) are lower than that of C4.5 and stacking. These results indicate that bagging, logistic regression and SVM outperform C4.5 and stacking at all levels and that the performance advantage of bagging and logistic regression is increasing in both evaluation prior fraud probability and relative error cost.

Panel A in Figure 4.1 shows the relative performance of the classification algorithms at different levels of relative cost when the evaluation prior fraud probability is 0.003, Panel B at 0.006, and Panel C at 0.012. Figure 4.1 corroborates the statistical findings showing that the relative performance of stacking and C4.5 deteriorates as the relative error cost and prior fraud probability increases. While the other results are also supported, SVM appears to perform slightly better than what was indicated by the linear regression results. At both the low (0.003) and middle (0.006) evaluation prior fraud probability levels (Figure 4.1, Panel A and Panel B) logistic regression appears to dominate the other classifiers except for SVM at all relative error costs except for at very low relative error costs (high FN cost), where all classifiers appear to perform similarly. Note that the lowest relative error cost level examined assumed that the cost of not detecting a financial statement fraud is the same as the cost of wrongfully suspecting that a financial statement is fraudulent, a relatively unlikely scenario. When the evaluation prior

Table 4.9
Regression Results for Testing Interactions between Classifier and Prior Fraud Probability, and Classifier and Relative Error Cost[a]

Variable[b]	*Estimate*	*Std Error*	*t-ratio*	*Prob>|t|*
Intercept	-0.256	0.005	-52.66	<0.001
Classifier [ANN]	-0.005	0.004	-1.19	0.236
Classifier [Bagging]	-0.012	0.004	-2.72	0.007
Classifier [C45]	0.020	0.004	4.48	<0.001
Classifier [IMF]	-0.005	0.004	-1.03	0.303
Classifier [Logistic]	-0.018	0.004	-4.09	<0.001
Classifier [SVM]	-0.012	0.004	-2.71	0.007
Classifier [Stacking]	0.031	0.004	7.02	<0.001
Prior Fraud Probability	39.565	0.489	80.86	<0.001
Relative Error Cost	0.006	0.000	99.09	<0.001
Classifier [ANN]*Prior Fraud Prob.	-0.000	0.000	-2.07	0.039
Classifier [Bagging]*Prior Fraud Prob.	-0.000	0.000	-2.78	0.005
Classifier [C45]*Prior Fraud Prob.	0.000	0.000	3.40	0.001
Classifier [IMF]*Prior Fraud Prob.	-0.000	0.000	-0.94	0.347
Classifier [Logistic]*Prior Fraud Prob.	-0.000	0.000	-2.06	0.039
Classifier [SVM]*Prior Fraud Prob.	-0.000	0.000	-1.36	0.175
Classifier [Stacking]*Prior Fraud Prob.	0.001	0.000	5.63	<0.001
Classifier [ANN]*Relative Error Cost	-3.340	1.199	-2.79	0.005
Classifier [Bagging]*Relative Error Cost	-4.143	1.199	-3.46	0.001
Classifier [C45]*Relative Error Cost	4.019	1.199	3.35	0.001
Classifier [IMF]*Relative Error Cost	-1.334	1.199	-1.11	0.266
Classifier [Logistic]*Relative Error Cost	-2.021	1.199	-1.69	0.092
Classifier [SVM]*Relative Error Cost	-0.640	1.199	-0.55	0.585
Classifier [Stacking]*Relative Error Cost	7.236	1.199	6.04	<0.001
Adjusted R^2	0.878			
RMSE	0.088			
n	2310			

[a] Two-tailed tests reported as directional predictions are not made.
[b] Dependent variable is Estimated Relative Cost.

probability level is high (Figure 4.1, Panel C), logistic regression still performs well, especially at the non-extreme relative error costs, i.e., in the error cost range that is the most realistic. Furthermore, at high prior fraud probability levels, bagging performs either on par with logistic regression or better, at all relative error costs. Thus, bagging appears to provide the best overall performance when the prior fraud probability is high. At the high evaluation prior fraud probability level, ANN also performs relatively well, but not better than bagging at any relative

error cost level. Finally, IMF is consistently a robust middle performer that outperforms the worse performing classifiers but is outperformed by the best performing classifiers, regardless of which specific classifiers perform well or poorly.

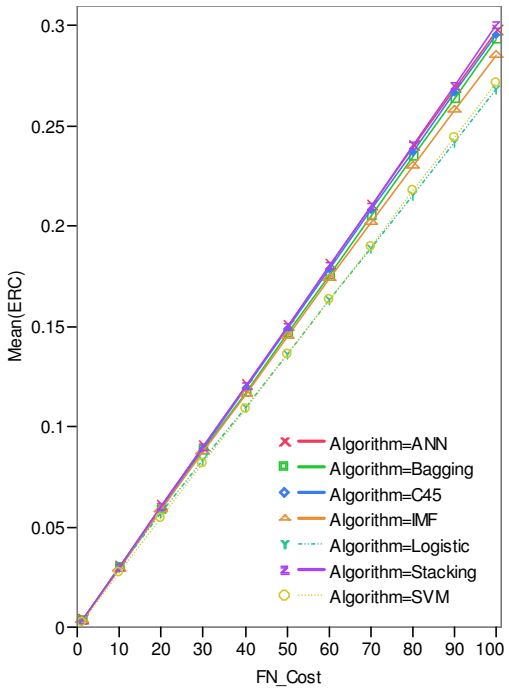
These results corroborate what was noted earlier in the regression analysis, and additionally indicates that SVM performs relatively well at non-extreme values of relative error cost when the prior fraud probability is either low or medium, and that IMF provides consistent, robust performance. Note that while SVM appears to perform relatively well at medium treatment levels, it appears to lose ground at high relative error costs, especially when the prior fraud probability is high. This finding explains why the interaction estimates are lower for SVM than for ANN, bagging and logistic regression. Based on these results it appears that logistic regression is a robust performer that often performs better than and rarely falls far behind the other classifiers. SVM appears to provide good performance over relevant ranges, but even so does not appear to provide any performance advantage when compared to logistic regression. Finally, bagging and ANN appear to perform relatively well at certain, though perhaps less relevant ranges, which explains why bagging and ANN overall performed relatively well.

To validate these observations I created three relative error cost groups, low (1:1, 1:10, and 1:20), middle (1:30, 1:40, 1:50, 1:60, and 1:70) and high (1:80, 1:90, and 1:100). Using the three relative error cost groups and the three original prior fraud probability levels, nine treatment groups were created. I examined an ANOVA model where the only main effect was the classifier algorithm within each of these nine treatment groups:

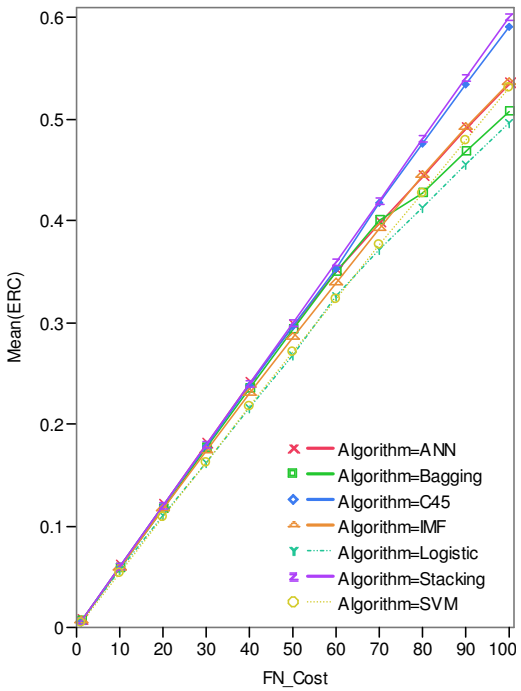
$$ERC = \alpha_0 + \alpha_1 \text{Classification Algorithm} + \varepsilon \quad (36)$$

The post-hoc analysis using Tukey-Kramer HSD reported in Table 4.10, shows that SVM significantly outperforms all other classifiers and that logistic regression significantly outperforms C4.5, stacking and ANN when the relative error costs and the prior fraud probability are low. Logistic regression and SVM significantly outperform all the other classifiers at: (1) middle and high relative error costs when the prior fraud probability is 0.003, (2) low and middle relative error costs when the prior fraud probability is 0.006 and (3) low relative error cost when the prior fraud probability is 0.012. When the prior fraud probability is 0.006, logstic regression significantly outperforms all the other classifiers except for bagging when the relative error cost is high. SVM significantly outperforms all the other classifiers except for stacking at middle relative error costs and a prior fraud probability of 0.012. At high relative error cost and high prior fraud probability stacking and C4.5 perform significantly worse than all the other classifiers. Overall, logistic regression and SVM perform well at all relative error cost and prior fraud probability levels.

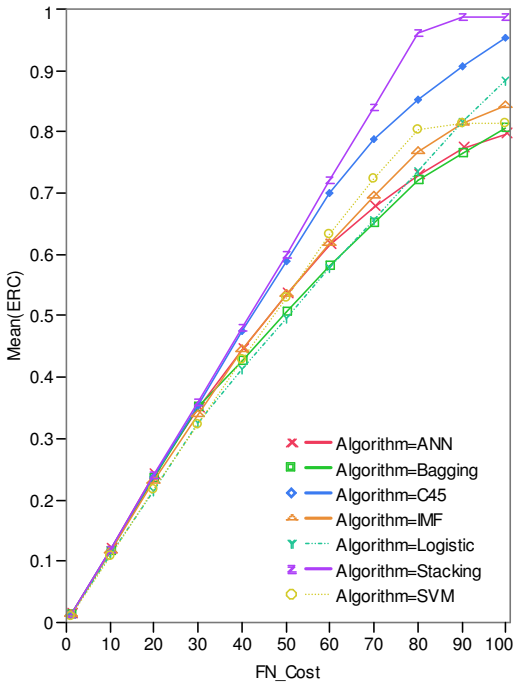
Panel A: Prior Fraud Probability = 0.003



Panel B: Prior Fraud Probability = 0.006



Panel C: Prior Fraud Probability = 0.012



Panel D: Prior Fraud Probability = 0.006

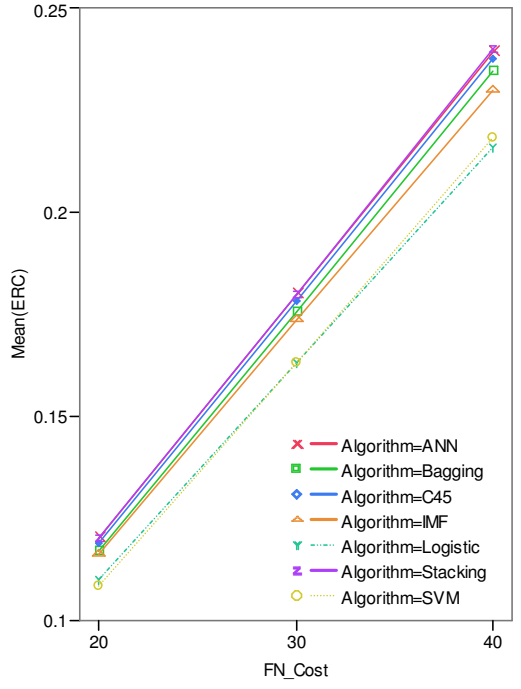


Figure 4.1: Classifier Comparison Estimated Relative Cost

Table 4.10
Comparison of Treatment Groups
Tukey-Kramer HSD Connected Letters Report

Panel A: Prior Fraud Probability = 0.003

	Relative Error Cost Range											
Classifier	**Low[a]**			**Classifier**	**Middle[a]**			**Classifier**	**High[a]**			
ANN	A			ANN	A			Stacking	A			
Stacking	A			Stacking	A			ANN	A			
C45	A			C45	A	B		C45	A			
Bagging	A	B		Bagging		B	C	Bagging	A	B		
IMF	A	B		IMF			C	IMF		B		
Logistic		B		Logistic				D	SVM			C
SVM			C	SVM				D	Logistic			C

Panel B: Prior Fraud Probability = 0.006

	Relative Error Cost Range											
Classifier	**Low[a]**				**Middle[a]**				**High[a]**			
ANN	A			Stacking	A				Stacking	A		
Stacking	A			C45	A	B			C45	A		
C45	A			ANN	A	B			IMF		B	
Bagging	A			Bagging		B	C		ANN		B	
IMF	A			IMF			C		SVM		B	
Logistic		B		SVM				D	Bagging		B	C
SVM		B		Logistic				D	Logistic			C

Panel C: Prior Fraud Probability = 0.012

	Relative Error Cost Range											
Classifier	**Low[a]**			**Classifier**	**Middle[a]**			**Classifier**	**High[a]**			
Stacking	A			ANN	A			Stacking	A			
ANN	A	B		Bagging	A			C45		B		
C45	A	B		C45		B		Logistic			C	
Bagging	A	B		IMF		B		SVM			C	
IMF		B		Logistic		B		IMF			C	
SVM			C	Stacking		B	C	ANN			C	
Logistic			C	SVM			C	Bagging			C	

[a] Levels not connected by the same letter are significantly different at a p-value of 0.05 using Tukey-Kramer HSD and blocking for the effect of evaluation prior fraud probability and relative cost on estimated relative cost.

As described earlier, relative error costs between 1:20 and 1:40, and prior fraud probability of 0.006 are believed to be good estimates of actual costs and prior probabilities associated with financial statement fraud (Bell and Carcello 2000; Bayley and Taylor 2007). Figure 4.1, Panel D shows the estimated relative cost of the classifiers at evaluation prior fraud probability of 0.006 and at relative error cost of 1:20, 1:30 and 1:40. It appears that logistic regression and SVM are superior when compared to the other classifiers over this relative error cost range. To validate this observation, I examined model (36) where the only main effect was classifier algorithm, within each of these three treatment groups.

The post-hoc analysis using Tukey-Kramer HSD and pair-wise t-test, reported in Table 4.11, confirm that logistic regression and SVM consistently outperform the other classifiers at what are believed to be good estimates of actual real world prior fraud probability and relative error cost.

4.6. Discussion

My experiments show that logistic regression, a relatively well-known and established classifier, and SVM outperform or perform as well as a relatively comprehensive set of data mining algorithms. This result is somewhat surprising considering that prior fraud research typically found ANN to either outperform or perform on par with logistic regression. However, this study differs from prior fraud studies in that it evaluates the classifiers using a highly imbalanced dataset, i.e., the minority class has a low prior probability, where the prior minority class probability is manipulated in both the training and the evaluation data. It also differs from most prior fraud research by examining the performance using optimal classification threshold levels for the different classifiers given a specific evaluation manipulation. Finally, this study differs from prior fraud research that compares classification algorithms by not only including a relatively complete set of attributes, but also using a Wrapper method to select attributes for each classifier. Thus, while the result that logistic regression and SVM outperform or perform as well as the other classifier is somewhat surprising it does not necessarily contradict these prior findings. Rather, the results show that when taking these additional factors into account logistic regression and SVM perform well in the fraud domain. A potential explanation as to why logistic regression performs well in this study is that logistic regression produces relatively accurate probability estimates (Perlich et al. 2003). Since the probability estimates generated by the different classifiers are compared in this study to various thresholds to find the threshold that minimizes ERC, the relative performance of logistic regression will be better than if performance is measured using classification results based on the default threshold of 0.5, which has been used in a majority of prior fraud research (Fanning and Cogger 1998; Feroz et al. 2000; Lin et al. 2003;

Table 4.11
Classifier Average Estimated Relative Cost at Best Estimates of Relative Error Cost and Prior Fraud Probability Levels

Panel A: Prior Fraud Probability = 0.006 and Relative Error Cost = 1:20

Classifier	*Tukey-Kramer HSD*[a]		*Pair-wise t-tests*					
			Logistic	*SVM*	*IMF*	*Bagging*	*C4.5*	*Stacking*
ANN	A		0.0100 (p<0.0001)	0.0113 (p<0.0001)	0.0037 (p=0.009)	0.0028 (p=0.0438)	0.0009 (p=0.4954)	0.0000 (p=1.000)
Stacking	A		0.0100 (p<0.0001)	0.0113 (p<0.0001)	0.0037 (p=0.009)	0.0028 (p=0.0438)	0.0009 (p=0.4954)	
C45	A		0.009 (p<0.0001)	0.0104 (p<0.0001)	0.0028 (p=0.0488)	0.0019 (p=0.1751)		
Bagging	A		0.0072 (p<0.0001)	0.0085 (p<0.0001)	0.0009 (p=0.5263)			
IMF	A		0.0063 (p<0.0001)	0.0076 (p<0.0001)				
SVM		B	0.0013 (p=0.3435)					
Logistic		B						

Panel B: Prior Fraud Probability = 0.006 and Relative Error Cost = 1:30

Classifier	*Tukey-Kramer HSD*[a]		*Pair-wise t-tests*					
			SVM	*Logistic*	*IMF*	*Bagging*	*C4.5*	*Stacking*
ANN	A		0.0169 (p<0.001)	0.0169 (p<0.001)	0.0065 (p=0.004)	0.0042 (p=0.059)	0.0015 (p=0.490)	0.0000 (p=1.000)
Stacking	A		0.0169 (p<0.001)	0.0169 (p<0.001)	0.0065 (p=0.004)	0.0042 (p=0.059)	0.0015 (p=0.490)	
C45	A		0.0154 (p<0.001)	0.0154 (p<0.001)	0.005 (p=0.027)	0.0027 (p=0.225)		
Bagging	A		0.0127 (p<0.001)	0.0127 (p<0.001)	0.0023 (p=0.304)			
IMF	A		0.0104 (p<0.001)	0.0104 (p<0.001)				
Logistic		B	0.0000 (p=0.994)					
SVM		B						

Table 4.11 (continued)

Panel C: Prior Fraud Probability = 0.006 and Relative Error Cost = 1:40

Classifier	*Tukey-Kramer HSD*[a]		*pair-wise t-tests*					
			Logistic	*SVM*	*IMF*	*Bagging*	*C4.5*	*Stacking*
ANN	A		0.0243 (p<0.001)	0.0219 (p<0.001)	0.010 (p=0.006)	0.0056 (p=0.112)	0.0026 (p=0.461)	0.0007 (p=0.833)
Stacking	A		0.0235 (p<0.001)	0.0212 (p<0.001)	0.0093 (p=0.010)	0.0049 (p=0.166)	0.0019 (p=0.598)	
C45	A		0.0217 (p<0.001)	0.0193 (p<0.001)	0.0074 (p=0.038)	0.0031 (p=0.387)		
Bagging	A		0.0186 (p<0.001)	0.0163 (p<0.001)	0.0044 (p=0.216)			
IMF	A		0.0143 (p=0.001)	0.0119 (p=0.001)				
SVM		B	0.0024 (p=0.501)					
Logistic		B						

[a] Levels not connected by the same letter are significantly different at a p-value of 0.05 using Tukey-Kramer HSD and blocking for the effect of evaluation prior fraud probability and relative cost on estimated relative cost.

Kotsiantis et al. 2006; Kirkos et al. 2007). Another potential explanation to why logistic regression performs well in this study is that logistic regression performs relatively well when it is difficult to separate signal from noise (Perlich et al. 2003). The area under the curve for logistic regression (AUC = 0.823), the measure of signal separability used in Perlich et al. (2003), is however, between the low- and high-separability groups found in their study.

Although the results are somewhat surprising, the experimental findings are encouraging since neither logistic regression nor SVM require extensive tuning and do not require a lot of resources for training and evaluation purposes. Furthermore, logistic regression is widely used and accepted, and produces results that are relatively easy to interpret and understand.

The experiment shows that out of 41 variables that have been found to be good predictors in prior fraud research, logistic regression uses a subset of only nine variables: auditor turnover, total discretionary accruals, Big 4 auditor, accounts receivable, allowance for doubtful accounts, meeting or beating analyst forecasts, inventory to sales, unexpected employee productivity and value of issued securities to market value. Across all classifiers only six variables are selected by three or more classifiers: auditor turnover, total discretionary accruals, Big 4 auditor, accounts

receivable, meeting or beating analyst forecasts, and unexpected employee productivity. These results, and the results reported for each classifier (see Table 4.7) can be used by practitioners as guidance for selecting variables to be included in fraud detection models. Another implication of this finding is that research developing new fraud predictors needs to examine the utility of the fraud predictor using more than one classification algorithm, i.e., in addition to using logistic regression other classifiers like SVM and C4.5 should be used when examining the utility of fraud predictors.

The findings in this essay need to be corroborated by future research using different datasets to evaluate the generalizability of the results. However, to increase the generalizability of the results I used 10-fold cross validation where the classification performance was measured on data not used for training. This 10-fold cross validation was repeated ten times to reduce the possibility of the results only pertaining to a specific 10-fold cross validation seed. Thus, the results should be generalizable to the population represented by the sample. However, datasets with different fraud firms can be used to validate that the sample is a good representative sample of fraud firms.

A natural extension of this research is to examine additional classification algorithms. While I select classification algorithms based on findings in prior research, it is possible that other classification algorithms will provide relatively good performance in financial statement fraud detection. Related to this extension is the possibility of not only tuning classification algorithms for the fraud domain, but designing novel classifiers for the specific purpose of detecting fraud. Finally, data mining research focusing on the class imbalance problem has proposed a number of sampling techniques such as SMOTE to improve classification performance (Chawla, et al 2002). The utility of these techniques in predicting fraud needs to be evaluated.

Chapter 5. Dissertation Conclusion

The first essay, Information Market Based Decision Fusion, introduces a novel combiner method based on theoretical and empirical findings in information market research. The results show that when the true classes of objects are only revealed for objects classified as positive and the positive ratio is low, IMF outperforms Majority, Average and Weighted Average. IMF outperforms Majority and performs on par with Average and Weighted Average, when the true classes of objects are only revealed for objects classified as positive and the positive ratio is high. Furthermore, IMF outperforms Weighted Average and Majority, and at a marginal level of significance, outperforms Average, when the true classes of all objects are revealed. This research contributes to multi-classifier combination combiner method research and, thereby, also to the broader research stream of ensemble-based classification and to classification algorithm research in general.

The second essay, The Effect of Discretionary Accruals, Earnings Expectations and Unexpected Productivity on Financial Statement Fraud: An Empirical Analysis, develops three novel fraud predictors: total discretionary accruals, meeting or beating analyst forecasts and unexpected revenue per employee. The results show that the three variables are significant predictors of fraud. This research contributes to the confirmatory fraud predictor research stream, which is part of a broader research area that focuses on developing and testing financial statement fraud predictors.

The third essay, Financial Statement Fraud Detection Using Data Mining: An Empirical Analysis, takes artifacts from the broader research streams to which the first two essays contribute, i.e., classification algorithm research and financial statement fraud predictor research, and compares the utility of artifacts developed in these research streams in detecting financial statement fraud. I find that logistic regression and SVM perform well relative to the other classification algorithms tested, i.e., C4.5, ANN, stacking, bagging and IMF. Logistic regression and SVM also provide the best performance under what is believed to be the most relevant prior probability and relative cost estimates. The results additionally show that out of 41 variables that have been found to be good predictors in prior fraud research, only six variables are selected by three or more classifiers: auditor turnover, total discretionary accruals, Big 4 auditor, accounts

receivable, meeting or beating analyst forecasts, and unexpected employee productivity. While other predictors are used by the classifiers their use is limited to only one or two classifiers. Thus, the utility of a given predictor, other than the six listed above, is dependent on the specific classifier used.

The results from Essay I in combination with the results from Essay III show that IMF performs better than existing combiner methods and better than stacking, an ensemble-based classification algorithm. Stacking is similar to IMF in that both methods use all the individual classifiers in the experiment as base-classifiers and then combine the results of these classifiers into an overall ensemble decision; stacking using a meta-learner and IMF using an information market based combiner method. Thus, the information market based combiner method developed in Essay I aggregates the base-classifiers decisions more effectively than the meta-learner used in stacking. The results also show that IMF performs on par with bagging, another ensemble-based classification algorithm. Bagging uses the combiner method AVG. Given that the positive ratio in the fraud domain is low and that the first essay shows that IMF outperforms AVG when the positive ratio is low, I expected IMF to outperform bagging. However, the homogeneous base-classifiers in bagging are trained using different data samples than the heterogeneous base-classifiers in IMF, which might explain why bagging performs on par with IMF even though bagging uses AVG. Assuming that IMF provides better performance than AVG and that the ensemble in bagging provides better performance than the ensemble used in IMF, it might be possible to improve the performance over bagging and IMF by combining the two algorithms. Future research can investigate the effectiveness of using IMF to combine the decisions of the base-classifiers in bagging and determine if bagging with IMF performs better than bagging with AVG.

The results from Essay II in combination with the results from Essay III show that the three predictors created in Essay II, unexpected revenue per employee, total discretionary accruals and meeting or beating analyst forecasts, are significant predictors of fraud and provide utility to classification algorithms. The three predictors provide insights into (1) conditions under which fraud is more likely to occur (total discretionary accruals is high), (2) incentives for fraud (firms desire to meet or beat analyst forecasts), and (3) how fraud is committed and can be detected (detection of revenue fraud using unexpected employee productivity). These three predictors are also among the group of six predictors selected by 50 percent or more of the classification algorithms. These results indicate that in a group of 41 fraud predictors that prior research has found to be significant predictors of fraud, the three predictors developed in Essay II are among the top six variables in terms of utility provided to the classification algorithms in fraud

prediction. Thus, the predictors developed in Essay II provide new knowledge about financial statement fraud and are useful in financial statement fraud classification.

To conclude, IMF performs well relative to existing combiner methods over a range of different domains, as shown in Essay I. In the fraud detection task, IMF is a robust performer and shows some promise when compared to other ensemble based methods. The three variables developed in Essay II were statistically significant predictors of fraud and these were shown to be robust. These variables made up half of the six variables selected from a group of 41 by 50 percent or more of the classification algorithms. I finally provide guidance for future fraud detection efforts by showing that logistic regression and SVM generally provide the best performance and specifically provide the best performance under what is believed to be the most realistic conditions. I also identified which predictors are overall most useful to the different classification algorithms. Six variables were selected by 50 percent or more of the classification algorithms: auditor turnover, total discretionary accruals, Big 4 auditor, accounts receivable, meeting or beating analyst forecasts, and unexpected employee productivity.

Chapter 6. References

ACFE, 2006, Report to the Nation on Occupational Fraud and Abuse, Association of Certified Fraud Examiners, Austin, TX.

AICPA, 1988, Statement on Auditing Standards (SAS) No. 53: The Auditor's Responsibility to Detect and Report Errors and Irregularities, American Institute of Certified Public Accountants, New York, NY.

AICPA, 1997, Statement on Auditing Standards (SAS) No. 82: Consideration of Fraud in a Financial Statement Audit, American Institute of Certified Public Accountants, New York, NY.

AICPA, 2002, Statement on Auditing Standards (SAS) No. 99: Consideration of Fraud in a Financial Statement Audit, American Institute of Certified Public Accountants, New York, NY.

Bayley, L., S., Taylor, 2007, "Identifying Earnings Management: A Financial Statement Analysis (Red Flag) Approach," Working Paper, January 20, 2007.

Beasley, M., 1996, "An Empirical Analysis of the Relation between the Board of Director Composition and Financial Statement Fraud," *The Accounting Review* 71(4), pp. 443-465.

Bell, T., J. Carcello, 2000, "A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting," *Auditing: A Journal of Practice & Theory* 19(1), pp. 169-184.

Beneish, M., 1997, "Detecting GAAP Violation: Implications for Assessing Earnings Management among Firms with Extreme Financial Performance," *Journal of Accounting and Public Policy* 16, pp. 271-309.

Beneish, M., 1999, "Incentives and Penalties Related to Earnings Overstatements That Violate GAAP," *The Accounting Review* 74(4), pp. 425-457.

Berg, J.E., T.A. Rietz, 2003, "Prediction Markets as Decision Support Systems", *Information Systems Frontiers* 5(1), pp. 79-93.

Breiman, L., 1996, "Bagging Predictors," *Machine Learning* 24(2), pp. 123-140.

Burgstahler, D., M. Eames, 2006, "Management of Earnings and Analysts' Forecasts to Achieve Zero and Small Positive Earnings Surprises," *Journal of Business Finance & Accounting* 33(5-6), pp. 633-652.

Carlsson, P., F. Ygge, A. Andersson, 2001, "Extending Equilibrium Markets," *IEEE Intelligent Systems* 16(4), pp. 18-26.

Chan, P.K., W. Fan, A.L. Prodromidis, S.J. Stolfo, 1999, "Distributed Data Mining in Credit Card Fraud Detection," *IEEE Intelligent Systems and Their Applications* 14(6), pp. 67-74.

Chawla, N.V., K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, 2002, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research* (16), pp. 321-357.

Chen, C., J. Sennetti, 2005, "Fraudulent financial reporting characteristics of the computer industry under a strategic-systems lens," *Journal of Forensic Accounting* 6(1), pp.23-54.

Dechow, P., R. Sloan, A. Sweeney, 1995, "Detecting Earnings Management," *The Accounting Review* 70(2), pp. 193-225.

Dechow, P., R. Sloan, A. Sweeney, 1996, "Causes and consequences of earnings manipulations: An analysis of firms subject to Enforcement Actions by the SEC," *Contemporary Accounting Research* 13(1), pp. 1-36.

Dichev, I., D. Skinner, 2002, "Large-sample evidence on the debt covenant hypothesis," *Journal of Accounting Research* 40, pp. 1091-1123.

Dopuch, N., R. Holthausen, R. Leftwich, 1987, "Predicting Audit Qualifications with Financial and Market Variables," *The Accounting Review* (62)3, pp. 431-454.

Drummond, C., R.C. Holte, 2006, "Cost Curves: An Improved Method for Visualizing Classifier Performance," *Machine Learning* 65(1), pp. 95-130.

Duin, P.W.R., M.J.D. Tax, 2000, "Experiments with Classifier Combining Rules," *International Workshop on Multiple Classifier Systems 2000*.

Fama, E., 1970, "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance* 25(2), pp. 383-417.

Fan, A., M. Palaniswami, 2000, "Selecting bankruptcy predictors using a support vector machine approach," *Neural Networks* (6), pp. 354-359.

Fanning, K., K. Cogger, 1998, "Neural network detection of management fraud using published financial data," *International Journal of Intelligent Systems in Accounting, Finance and Management* 7(1), pp. 21-41.

Feroz, E., T. Kwon, V. Pastena, K. Park, 2000, "The Efficacy of Red-Flags in Predicting the SEC's Targets: An Artificial Neural Networks Approach," *International Journal of Intelligent Systems in Accounting, Finance & Management* 9(3), pp. 145-157.

Fries, T., N. Cristianini, C. Campbell, 1998, "The kernel adatron algorithm: a fast and simple learning procedure for support vector machines," *In the Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

Green, B.P., J.H. Choi, 1997, "Assessing the Risk of Management Fraud Through Neural Network Technology," *Auditing: A Journal of Practice & Theory* 16(1), pp. 14-28.

Hall, M., G. Holmes, 2003, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," *IEEE Transactions on Knowledge and Data Engineering* (15)3, pp. 1-16.

Hanson, R., 2003, "Combinatorial Information Market Design," *Information Systems Frontiers* 5(1), pp. 107-119.

Hayek, F.A., 1945, "The Use of Knowledge in Society," *The American Economic Review* 35(4), pp. 519-530.

Healy, P., 1985, "The effect of bonus schemes on accounting decisions," *Journal of Accounting and Economics* 7, pp. 85-107.

Healy, P.M., J.M. Wahlen, 1999, "A review of the earnings management literature and its implications for standard setting," *Accounting Horizons* 13(4), pp. 365-383.

Hribar, P., D.W. Collins, 2002, "Errors in Estimating Accruals: Implications for Empirical Research," *Journal of Accounting Research* 40(1), pp. 105-134.

Jaccard, J., C.K. Wan, 1996, *LISREL approaches to interaction effects in multiple regression*, Sage Publications, Thousand Oaks, CA.

Jain, A.K., R.P.W. Duin, J. Mao, 2000, "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(1), pp. 4-37.

Jones, J., 1991, "Earnings management during import relief investigations," *Journal of Accounting Research* 29, pp. 193-228.

Kaminski, K., S. Wetzel, L. Guan, 2004, "Can financial ratios detect fraudulent financial reporting," *Managerial Auditing Journal* (19)1, pp. 15-28.

Kasznik, R., 1999, "On the Association between Voluntary Disclosure and Earnings Management," *Journal of Accounting Research* 37(1), pp. 57-81.

Kelly, J., 1956, "A New Interpretation of Information Rate.," *IEEE Transactions on Information Theory* 2(3), pp. 185-189.

Kennedy, P.E., 1981, "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations," *American Economic Review* 71(4), p. 801.

Kirkos, E., C. Spathisb, Y. Manolopoulos, 2007, "Data Mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications* (32)4, pp. 995-1003.

Kittler, J., M. Hatef, R.P.W. Duin, J. Matas, 1998, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(3), pp. 226-239.

Kotsiantis, S., E. Koumanakos, D. Tzelepis, V. Tampakas, 2006, "Forecasting Fraudulent Financial Statements using Data Mining," *International Journal of Computational Intelligence* (3)2, pp. 104-110.

KPMG, 2006, KPMG Forensic's 2006 survey of fraud in Australia and New Zealand, KPMG Forensic, Melbourne, Australia.

Lam, L., 2000, "Classifier Combinations: Implementations and Theoretical Issues," *Multiple Classifier Systems in Lecture Notes in Computer Science* 1857, pp. 77-86.

Lee, T.A., R.W. Ingram, T.P. Howard, 1999, "The difference between earnings and operating cash flow as an indicator of financial reporting fraud," *Contemporary Accounting Research* 16(4), pp. 749-786.

Lee, W., S.J. Stolfo, K.W. Mok, 2000, "Adaptive Intrusion Detection: A Data Mining Approach," *Artificial Intelligence Review* 14(6), pp. 533-567.

Lin, J., M. Hwang, J. Becker, 2003, "A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting," *Managerial Auditing Journal* 18(8), pp. 657-665.

Loebbecke, J.K., M.M. Eining, J.J. Willingham, 1989, "Auditors' experience with material irregularities: Frequency, nature, and detectability," *Auditing: A Journal of Practice and Theory* 9(1), pp. 1-28.

Newman, D.J., S. Hettich, C.L. Blake, C.J. Merz, 1998, UCI Repository of Machine Learning Databases http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA, University of California, Department of Information and Computer Science.

Nissen, M.E., K. Sengupta, 2006, "Incorporating Software Agents into Supply Chains: Experimental Investigation with a Procurement Task," *MIS Quarterly* 30(1), pp. 145-166.

Oversight, 2005, The 2005 Oversight System Report on Corporate Fraud, Oversight, Atlanta, GA.

PCAOB, 2004, Auditing Standard No. 2: An Audit of Internal Control Over Financial Reporting Performed in Conjunction With an Audit of Financial Statements, Public Company Accounting Oversight Board, Washington, DC.

PCAOB, 2007, Auditing Standard No. 5: An Audit of Internal Control Over Financial Reporting That Is Integrated with An Audit of Financial Statements, Public Company Accounting Oversight Board, Washington, DC.

Pennock, M.D., 2004, "A Dynamic Pari-Mutuel Market for Hedging, Wagering, and Information Aggregation," *In the Proceedings of the 5th ACM Conference on E-Commerce*, New York, NY.

Perlich, C., F. Provost, J. Simonoff, 2003, "Tree Induction vs. Logistic Regression: A Learning-Curve Analysis," *Journal of Machine Learning Research* (4), pp. 211-255.

Phua, C., D. Alahakoon, V. Lee, 2004, "Minority Report in Fraud Detection: Classification of Skewed Data," *SIGKDD Explorations* 6(1), pp. 50-59.

Platt, J., 1999, "Fast Training of support vector machines using Sequential Minimal Optimization," *Advances in Kernel Methods: Support Vector Learning*, pp. 185-208.

Plott, C.R., J. Wit, W.C. Yang, 2003, "Parimutuel Betting Markets as Information Aggregation Devices: Experimental Results," *Economic Theory* 22(2), pp. 311-351.

Prodromidis, A., P.Chan, S. Stolfo, 2000, "Meta-learning in distributed data mining systems: Issues and approaches," *Advances in Distributed and Parallel Knowledge Discovery*, (eds.) Kargupta, H. and Chan, P., Chapter 3, AAAI/MIT.

Provost, F., T. Fawcett, 2001, "Robust Classification for Imprecise Environments," *Machine Learning* 42(3), pp. 203-231.

Provost, F., T. Fawcett, R. Kohavi, 1998, "The Case Against Accuracy Estimation for Comparing Induction Algorithms," *In the Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, WI.

Quinlan, J.R., 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

Rubinstein, M., 1976, "The Strong Case for the Generalized Logarithmic Utility Model as the Premier Model of Financial Markets," *The Journal of Finance* 31(2), pp. 551-571.

Saar-Tsechansky, M., F. Provost, 2004, "Active Sampling for Class Probability Estimation and Ranking," *Machine Learning* 54(2), pp. 153-178.

Shin, K.S., T. Lee, H.J., Kim, 2005, "An application of support vector machines in bankruptcy prediction model," *Expert Systems with Application* (28), pp. 127-135.

Stolfo, S., A.L. Prodromidis, S. Tselepis, W. Lee, D.W. Fan, P.K. Chan, 1997, "JAM: Java agents for meta-learning over distributed databases," *In the Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA.

Suen, C.Y., L. Lam, 2000, "Multiple classifier combination methodologies for different output levels," *Lecture notes in computer science* 1857, pp. 52-66.

Summers, S.L., J.T. Sweeney, 1998, "Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis," *The Accounting Review* 73(1), pp. 131-146.

Weiss, S.M., I. Kapouleas, 1989, "An empirical comparison of pattern recognition, neural nets, and machine learning classification methods," *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI

West, D., S. Dellana, J. Qian, 2004, "Neural network ensemble strategies for decision applications," *Computer & Operations Research* (32), pp. 2543-2559.

Witten, I.H., E. Frank, 2005, *Data Mining: Practical machine learning tools and techniques*, San Francisco, CA.

Wolfers, J., E. Zitzewitz, 2006, "Interpreting Prediction Market Prices as Probabilities," *In the Proceedings of the Allied Social Science Association Annual Meeting*, Boston, MA.

Wolpert, D., 1992, "Stacked generalization," *Neural Networks* (5)2, pp. 241-259.

Ygge, F., J.M. Akkermans, 1999, "Decentralized Markets versus Central Control: A Comparative Study," *Journal of Artificial Intelligence Research* 11, pp. 301-333.

Yule, G.U, 1900, "On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, &c.," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 194, pp. 257-319.

Zhao, H., S. Ram, 2004, "Constrained Cascade Generalization of Decision Trees," *IEEE Transactions on Knowledge and Data Engineering* 16(6), pp. 727-739.

Zheng, Z., B. Padmanabhan, 2007, "Constructing Ensembles from Data Envelopment Analysis," *INFORMS Journal on Computing* 19(4), pp. 486-496.

Chapter 7. Appendices

Appendix 1: Proof Lemma 1

Lemma 1: The optimal bets of agent i in P3 while classifying t is:

$q^{*}_{itj} = p_{itj}(w_{it}+m)\ \forall j \in J.$

Proof Lemma 1

$$Z_3 = \max_{qitj}\ p_{it1} ln(q_{it1}O_{t1}) + p_{it2} ln(q_{it2}O_{t2}) \quad (37)$$

$$\text{S.T.} \quad q_{it1} + q_{it2} = w_{it}+m \quad (38)$$

$$q_{itj} \geq 0 \quad (39)$$

Using Lagrangian multipliers λ_1, λ_2, and λ_3, we get

$$L_3 = p_{it1} ln(q_{it1}O_{t1}) + p_{it2} ln(q_{it2}O_{t2}) + \lambda_1(w_{it} + m - q_{it1} - q_{it2}) + \lambda_2(q_{it1} - 0) + \lambda_3(q_{it2} - 0) \quad (40)$$

$$\partial L_3/\partial q_{it1} = 0 \Rightarrow p_{it1}O_{t1}/q_{it1}O_{t1} - \lambda_1 = 0 \quad (41)$$

$$\partial L_3/\partial q_{it2} = 0 \Rightarrow p_{it2}O_{t2}/q_{it2}O_{t2} - \lambda_1 = 0 \quad (42)$$

$$\partial L_3/\partial \lambda_1 = 0 \Rightarrow w_{it} + m - q_{it1} - q_{it2} = 0 \quad (43)$$

$$\lambda_2(q_{it1} - 0) = 0 \Rightarrow \lambda_2 q_{it1} = 0 \quad (44)$$

$$\lambda_3(q_{it2} - 0) = 0 \Rightarrow \lambda_3 q_{it2} = 0 \quad (45)$$

simplify (41) and (42)

$$p_{it1}/q_{it1} - \lambda_1 = 0 \quad (46)$$

$$p_{it2}/q_{it2} - \lambda_1 = 0 \quad (47)$$

combine (46) and (47)

$$p_{it1}/q_{it1} = p_{it2}/q_{it2} \quad (48)$$

combine (43) and (48)

$$p_{it1} / q_{it1} = p_{it2} / (w_{it} + m - q_{it1}) \quad (49)$$

$$p_{it2} / q_{it2} = p_{it1} / (w_{it} + m - q_{it2}) \quad (50)$$

simplify (49) and (50) (note that $p_{it1} = 1 - p_{it2}$)

$$p_{it1} (w_{it} + m) - p_{it1} q_{it1} = q_{it1} - p_{it1} q_{it1} \quad (51)$$

$$p_{it2} (w_{it} + m) - p_{it2} q_{it2} = q_{it2} - p_{it2} q_{it2} \quad (52)$$

simplify (51) and (52)

$$q_{it1} = p_{it1} (w_{it} + m) \quad (53)$$

Appendix 1: (Continued)

$q_{it2} = p_{it2}(w_{it} + m)$ (54)

Use the Hessian matrix for $p_{it1}ln(q_{it1}O_{t1}) + p_{it2}ln(q_{it2}O_{t2}) + \lambda_1(w_{it} + m - q_{it1} - q_{it2}) + \lambda_2(q_{it1} - 0) + \lambda_3(q_{it2} - 0)$ to verify that L_3 has a relative maximum at the critical point obtained in (53) and (54):

$$\begin{bmatrix} \partial^2 L_3 / q_{it1}^2 & \partial^2 L_3 /(q_{it1}, q_{it2}) \\ \partial^2 L_3 /(q_{it2}, q_{it1}) & \partial^2 L_3 / q_{it2}^2 \end{bmatrix}, \text{ where} \quad (55)$$

$$\partial^2 L_3 / q_{it1}^2 = -p_{it1}/q_{it1}^2 \quad (56)$$

$$\partial^2 L_3 / q_{it2}^2 = -p_{it2}/q_{it2}^2 \quad (57)$$

$$\partial^2 L_3 /(q_{it1}, q_{it2}) = 0 \quad (58)$$

$$\partial^2 L_3 /(q_{it2}, q_{it1}) = 0 \quad (59)$$

The determinant of (55) is:

$D_3 = (\partial^2 L_3 / q_{it1}^2)(\partial^2 L_3 / q_{it2}^2) - (\partial^2 L_3 /(q_{it1}, q_{it2}))(\partial^2 L_3 /(q_{it2}, q_{it1}))$

$D_3 = (-p_{it1}/q_{it1}^2)\,(-p_{it2}/q_{it2}^2)$ (60)

Simplify (60)

$D_3 = p_{it1}\, p_{it2}/q_{it1}^2 q_{it2}^2$ (61)

$\forall j \in J$, when $0<p_{itj}<1$, then $0< q_{itj} < (w_{it}+m)$ as per (53) and (54), therefore $D_3 > 0$. Further, since $\partial^2 L_3 / q_{it1}^2 < 0$ and $\partial^2 L_3 / q_{it2}^2 < 0$ (see (56), (57)), therefore the critical point is a relative maximum. When $p_{itj} = 0$ or 1, then $D_3=0$, i.e., the Hessian is indeterminate. It can be seen from (38), (45) and (53) that when $p_{it1} = 0$, then $q_{it1}=0$, $q_{it2} = (w_{it}+m)$ and $\lambda_3 = 0$. It can similarly be verified that when $p_{it2} = 0$, then $q_{it2}=0$, $q_{it1} = (w_{it}+m)$ and $\lambda_2 = 0$.

Appendix 2: Proof Lemma 2

Lemma 2: The optimal bets of agent i in P4 while classifying t is:

Solution a: $q^*_{it1} = p_{it1}\, km + a_{it}\dfrac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}}$ and

$$q^*_{it2} = p_{it2}\, km + a_{it}\frac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}}\text{, when}$$

$$0 < p_{it1}\, km + a_{it}\frac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}} < km \text{ and}$$

$$0 < p_{it2}\, km + a_{it}\frac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}} < km;$$

Solution b: $q^*_{it1} = km$, and $q^*_{it2} = 0$, when

$$km \le p_{it1}\, km + a_{it}\frac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}}\text{; and}$$

Solution c: $q^*_{it2} = km$, and $q^*_{it1} = 0$, when

$$km \le p_{it2}\, km + a_{it}\frac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}}$$

Proof Lemma 2

$Z_4 = \max_{qitj}\ p_{it1}ln(q_{it1}O_{t1} + a_{it}) + p_{it2}ln(q_{it2}O_{t2} + a_{it})$ (62)

S.T. $q_{it1} + q_{it2} = km$ (63)

$q_{itj} \ge 0$ (64)

Using Lagrangian multipliers λ_1, λ_2, and λ_3, we get

$L_4 = p_{it1}ln(q_{it1}O_{t1} + a_{it}) + p_{it2}ln(q_{it2}O_{t2} + a_{it}) + \lambda_1(km - q_{it1} - q_{it2}) + \lambda_2(q_{it1} - 0) + \lambda_3(q_{it2} - 0)$ (65)

$\partial L_4/\partial q_{it1} = 0 \Rightarrow p_{it1}O_{t1} / (q_{it1}O_{t1} + a_{it}) - \lambda_1 + \lambda_2 = 0$ (66)

$\partial L_4/\partial q_{it2} = 0 \Rightarrow p_{it2}O_{t2} / (q_{it2}O_{t2} + a_{it}) - \lambda_1 + \lambda_3 = 0$ (67)

$\partial L_4/\partial \lambda_1 = 0 \Rightarrow km - q_{it1} - q_{it2} = 0$ (68)

$\lambda_2(q_{it1} - 0) = 0 \Rightarrow \lambda_2 q_{it1} = 0$ (69)

$\lambda_3(q_{it2} - 0) = 0 \Rightarrow \lambda_3 q_{it2} = 0$ (70)

given (68), (69) and (70) Z_4 has three possible solutions:

Solution a: $0 < q_{it1} < km \rightarrow 0 < q_{it2} < km \wedge \lambda_2 = 0 \wedge \lambda_3 = 0$ (71)

Appendix 2: (Continued)

Solution b: $q_{it1} = 0 \rightarrow q_{it2} = km \wedge \lambda_3 = 0$ (72)

Solution c: $q_{it2} = 0 \rightarrow q_{it1} = km \wedge \lambda_2 = 0$ (73)

Agent i determines the optimal solution of (62) given only the constraint in (63) as given below.

To solve for Solution a combine (66), (67) and (71)

$p_{it1}O_{t1} / (q_{it1}O_{t1} + a_{it}) - \lambda_1 = p_{it2}O_{t2} / (q_{it2}O_{t2} + a_{it}) - \lambda_1$ (74)

simplify (74)

$p_{it1}q_{it2}O_{t1}O_{t2} + p_{it1}O_{t1}a_{it} = p_{it2}q_{it1}O_{t1}O_{t2} + p_{it2}O_{t2}a_{it}$ (75)

combine (68) and (75)

$kmp_{it1}O_{t1}O_{t2} - p_{it1}q_{it1}O_{t1}O_{t2} + p_{it1}O_{t1}a_{it} = p_{it2}q_{it1}O_{t1}O_{t2} + p_{it2}O_{t2}a_{it}$ (76)

$kmp_{it2}O_{t1}O_{t2} - p_{it2}q_{it2}O_{t1}O_{t2} + p_{it2}O_{t2}a_{it} = p_{it1}q_{it2}O_{t1}O_{t2} + p_{it1}O_{t1}a_{it}$ (77)

simplify (76) and (77) (note that $p_{it1} + p_{it2} = 1$)

$q_{it1}O_{t1}O_{t2} = kmp_{it1}O_{t1}O_{t2} + p_{it1}O_{t1}a_{it} - p_{it2}O_{t2}a_{it}$ (78)

$q_{it2}O_{t1}O_{t2} = kmp_{it2}O_{t1}O_{t2} + p_{it2}O_{t2}a_{it} - p_{it1}O_{t1}a_{it}$ (79)

simplify (78) and (79)

$q_{it1} = p_{it1}km + a_{it}(p_{it1}O_{t1} - p_{it2}O_{t2})/O_{t1}O_{t2}$ (80)

$q_{it2} = p_{it2}km + a_{it}(p_{it2}O_{t2} - p_{it1}O_{t1})/O_{t1}O_{t2}$ (81)

If $q_{it1} > 0$ and $q_{it2} > 0$ then Solution a is given by (80) and (81). When $q_{it1} \leq 0$, then agent i will bet as per Solution b, else when $q_{it2} \leq 0$, then agent i will bet as per Solution c.

Use the Hessian matrix for $p_{it1}ln(q_{it1}O_{t1} + a_{it}) + p_{it2}ln(q_{it2}O_{t2} + a_{it}) + \lambda_1(km - q_{it1} - q_{it2}) + \lambda_2(q_{it1} - 0) + \lambda_3(q_{it2} - 0)$ to verify that L_4 has a relative maximum at the critical point obtained in (80) and (81).

$$\begin{bmatrix} \partial^2 L_4 / q_{it1}^2 & \partial^2 L_4 /(q_{it1}, q_{it2}) \\ \partial^2 L_4 /(q_{it2}, q_{it1}) & \partial^2 L_4 / q_{it2}^2 \end{bmatrix}, \text{ where} \quad (82)$$

$$\partial^2 L_4 / q_{it1}^2 = - p_{it1}O_{t1}^2 /(q_{it1}O_{t1} + a_{it})^2 \quad (83)$$

$$\partial^2 L_4 / q_{it2}^2 = - p_{it2}O_{t2}^2 /(q_{it2}O_{t2} + a_{it})^2 \quad (84)$$

Appendix 2: (Continued)

$$\partial^2 L_4 /(q_{it1}, q_{it2}) = 0 \tag{85}$$

$$\partial^2 L_4 /(q_{it2}, q_{it1}) = 0 \tag{86}$$

The determinant of (82) is:

$$D_4 = (\partial^2 L_4 / q_{it1}^2)(\partial^2 L_4 / q_{it2}^2) - (\partial^2 L_4 /(q_{it1}, q_{it2}))(\partial^2 L_4 /(q_{it2}, q_{it1}))$$

$$D_4 = (- p_{it1} O_{t1}^2 /(q_{it1} O_{t1} + a_{it})^2)(- p_{it2} O_{t2}^2 /(q_{it2} O_{t2} + a_{it})^2) \tag{87}$$

Simplify (87)

$$D_4 = p_{it1} p_{it2} O_{t1}^2 O_{t2}^2 /((q_{it1} O_{t1} + a_{it})^2 (q_{it2} O_{t2} + a_{it})^2) \tag{88}$$

$\forall j \in J$, when $0<p_{itj}<1$, then $D_4 > 0$. Note that $O_{tj} \geq 1$ by definition, $a_{it}>0$ given $w_{it} > (k\text{-}1)m$. Further since $\partial^2 L_4 / q_{it1}^2 < 0$ and $\partial^2 L_4 / q_{it2}^2 < 0$ (see (83), (84)), therefore the critical point is a relative maximum. When $p_{itj} = 0$ or 1, then $D_4=0$, i.e., the Hessian is indeterminate. It can be seen from (63) CostSavings3 and (80) that when $p_{it1} = 0$, then $q_{it1} \leq 0$, $q_{it2} \geq km$, i.e., $p_{it2}\, km + a_{it} \frac{p_{it2} O_{t2} - p_{it1} O_{t1}}{O_{t1} O_{t2}} \geq km$, as per (81). It can similarly be verified that when $p_{it2} = 0$, then $q_{it2} \leq 0$, $q_{it1} \geq km$, i.e., $p_{it1}\, km + a_{it} \frac{p_{it1} O_{t1} - p_{it2} O_{t2}}{O_{t1} O_{t2}} \geq km$, as per (80). When $q_{it1} \leq 0$ then constraint (64) becomes binding and q_{it1} is set to 0 (Solution b). Similarly, when $q_{it2} \leq 0$ then constraint (64) becomes binding and q_{it2} is set to 0 (Solution c).

Appendix 3: Proof Lemma 3

Lemma 3*: Given any combination of betting behaviors as per Lemma 1 and Lemma 2, equilibrium exists, and the equilibrium odd for j=1 is:*

$$O_{t1} = \frac{\sum_{i \in D1 \cup D2a} p_{it2}(w_{it} + m) + \sum_{i \in D2c}(km)}{\sum_{i \in D1 \cup D2a} p_{it1}(w_{it} + m) + \sum_{i \in D2b}(km)} + 1$$

Proof Lemma 3

In IMF, for each object *t,* the house manipulates the market odds O_{t1} and O_{t2} to establish the equilibrium odds that occur when:

$$O_{t1}Q_{t1} = O_{t2}Q_{t2} \tag{89}$$

Using Lemma 1 and Lemma 2, the LHS and RHS of (89) are:

$$O_{t1}Q_{t1} = O_{t1}\sum_{i \in D1} p_{it1}(w_{it} + m) + O_{t1}\sum_{i \in D2a}(p_{it1}km + a_{it}\frac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}}) +$$
$$O_{t1}\sum_{i \in D2b}(km) + O_{t1}\sum_{i \in D2c}(0)\text{, and} \tag{90}$$

$$O_{t2}Q_{t2} = O_{t2}\sum_{i \in D1} p_{it2}(w_{it} + m) + O_{t2}\sum_{i \in D2a}(p_{it2}km + a_{it}\frac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}}) +$$
$$O_{t2}\sum_{i \in D2b}(0) + O_{t2}\sum_{i \in D2c}(km)\text{,} \tag{91}$$

where all agents $i \in D_1$ that bet per Lemma 1 are defined as $i \in D_1$ and all agents that bet per Lemma 2, solutions a, b and c are defined as $i \in D_{2a}$, $i \in D_{2b}$ and $i \in D_{2c}$, respectively.

On substituting the LHS of (89) with the RHS of (90) and the RHS of (89) with the RHS of (91):

$$O_{t1}\sum_{i \in D1} p_{it1}(w_{it} + m) + O_{t1}\sum_{i \in D2a}(p_{it1}km + a_{it}\frac{p_{it1}O_{t1} - p_{it2}O_{t2}}{O_{t1}O_{t2}}) +$$
$$O_{t1}\sum_{i \in D2b}(km) + O_{t1}\sum_{i \in D2c}(0) =$$
$$O_{t2}\sum_{i \in D1} p_{it2}(w_{it} + m) + O_{t2}\sum_{i \in D2a}(p_{it2}km + a_{it}\frac{p_{it2}O_{t2} - p_{it1}O_{t1}}{O_{t1}O_{t2}}) +$$
$$O_{t2}\sum_{i \in D2b}(0) + O_{t2}\sum_{i \in D2c}(km) \tag{92}$$

On Simplifying (92):

$$O_{t1}\sum_{i \in D1} p_{it1}(w_{it} + m) + O_{t1}\sum_{i \in D2a}(p_{it1}km) +$$

Appendix 3: (Continued)

$$\sum_{i\in D2a}(a_{it}\frac{p_{it1}O_{t1}}{O_{t2}})-\sum_{i\in D2a}(a_{it}p_{it2})+O_{t1}\sum_{i\in D2b}(km)=$$

$$O_{t2}\sum_{i\in D1}p_{it2}(w_{it}+m)+O_{t2}\sum_{i\in D2a}(p_{it2}km)+$$

$$\sum_{i\in D2a}(a_{it}\frac{p_{it2}O_{t2}}{O_{t1}})-\sum_{i\in D2a}(a_{it}p_{it1})+O_{t2}\sum_{i\in D2c}(km) \quad (93)$$

On Simplifying (93) by substituting O_{t2} for $O_{t1}/(O_{t1}\text{-}1)$:

$$O_{t1}\sum_{i\in D1}p_{it1}(w_{it}+m)+O_{t1}\sum_{i\in D2a}(p_{it1}km)+$$

$$\sum_{i\in D2a}(a_{it}p_{it1}(O_{t1}-1))-\sum_{i\in D2a}(a_{it}p_{it2})+O_{t1}\sum_{i\in D2b}(km)=$$

$$\frac{O_{t1}}{O_{t1}-1}\sum_{i\in D1}p_{it2}(w_{it}+m)+\frac{O_{t1}}{O_{t1}-1}\sum_{i\in D2a}(p_{it2}km)+$$

$$\sum_{i\in D2a}(a_{it}\frac{p_{it2}}{O_{t1}-1})-\sum_{i\in D2a}(a_{it}p_{it1})+\frac{O_{t1}}{O_{t1}-1}\sum_{i\in D2c}(km) \quad (94)$$

On simplifying (94):

$$O_{tl}(\sum_{i\in D1}p_{it1}(w_{it}+m)+\sum_{i\in D2a}(p_{it1}km)+\sum_{i\in D2a}(a_{it}p_{it1})+\sum_{i\in D2b}(km))-$$

$$\frac{O_{t1}}{O_{t1}-1}(\sum_{i\in D1}p_{it2}(w_{it}+m)+\sum_{i\in D2a}(p_{it2}km)+\sum_{i\in D2c}(km))-\sum_{i\in D2a}(a_{it}\frac{p_{it2}}{O_{t1}-1})=$$

$$\sum_{i\in D2a}(a_{it}p_{it2}) \quad (95)$$

On simplifying (95):

$$(O_{tl}\text{-}1)\,O_{tl}(\sum_{i\in D1}p_{it1}(w_{it}+m)+\sum_{i\in D2a}(p_{it1}km)+\sum_{i\in D2a}(a_{it}p_{it1})+\sum_{i\in D2b}(km))-$$

$$O_{t1}(\sum_{i\in D1}p_{it2}(w_{it}+m)+\sum_{i\in D2a}(p_{it2}km)+\sum_{i\in D2c}(km))-(O_{t1}-1)\sum_{i\in D2a}(a_{it}p_{it2})=$$

$$\sum_{i\in D2a}(a_{it}p_{it2}) \quad (96)$$

On simplifying (96):

$$(O_{tl}\text{-}1)O_{tl}(\sum_{i\in D1}p_{it1}(w_{it}+m)+\sum_{i\in D2a}(p_{it1}km)+\sum_{i\in D2a}(a_{it}p_{it1})+\sum_{i\in D2b}(km))-$$

$$O_{t1}(\sum_{i\in D1}p_{it2}(w_{it}+m)+\sum_{i\in D2a}(p_{it2}km)+\sum_{i\in D2a}(a_{it}p_{it2})+\sum_{i\in D2c}(km))=0 \quad (97)$$

Appendix 3: (Continued)

On simplifying (97):

$$(O_{t1}\text{-}1)\ (\sum_{i\in D1} p_{it1}(w_{it}+m)+\sum_{i\in D2a}(p_{it1}km)+\sum_{i\in D2a}(a_{it}p_{it1})+\sum_{i\in D2b}(km))=$$
$$\sum_{i\in D1} p_{it2}(w_{it}+m)+\sum_{i\in D2a}(p_{it2}km)+\sum_{i\in D2a}(a_{it}p_{it2})+\sum_{i\in D2c}(km) \quad (98)$$

On simplifying (98):

$$O_{t1}=\frac{\sum_{i\in D1} p_{it2}(w_{it}+m)+\sum_{i\in D2a} p_{it2}(km+a_{it})+\sum_{i\in D2c}(km)}{\sum_{i\in D1} p_{it1}(w_{it}+m)+\sum_{i\in D2a} p_{it1}(km+a_{it})+\sum_{i\in D2b}(km)}+1 \quad (99)$$

On simplifying (99), note that $a_{it} = w_{it} + m - km$:

$$O_{t1}=\frac{\sum_{i\in D1} p_{it2}(w_{it}+m)+\sum_{i\in D2a} p_{it2}(w_{it}+m)+\sum_{i\in D2c}(km)}{\sum_{i\in D1} p_{it1}(w_{it}+m)+\sum_{i\in D2a} p_{it1}(w_{it}+m)+\sum_{i\in D2b}(km)}+1 \quad (100)$$

On simplifying (100):

$$O_{t1}=\frac{\sum_{i\in D1\cup D2a} p_{it2}(w_{it}+m)+\sum_{i\in D2c}(km)}{\sum_{i\in D1\cup D2a} p_{it1}(w_{it}+m)+\sum_{i\in D2b}(km)}+1 \quad (101)$$

Since $O_{tj} = 1/P_{itj} \geq 1$ equilibrium odds exist when $\frac{\sum_{i\in D1\cup D2a} p_{it2}(w_{it}+m)+\sum_{i\in D2c}(km)}{\sum_{i\in D1\cup D2a} p_{it1}(w_{it}+m)+\sum_{i\in D2b}(km)} \geq 0$.

Thus, since $0 \leq p_{itj} \leq 1$, $(w_{it} + m) > 0$, and $km > 0$, equilibrium odds exist when agents bet as per Lemma 1 and Lemma 2.

Appendix 4: Empirical Experiments of Equilibrium Odds

Discontinuous Agent Bets in Odds **-** If agent bets are discontinuous over O_{tj} then the existence of equilibrium odds cannot be guaranteed (Carlsson et al. 2001). To provide some insights into the utility of IMF in situations such as above, when equilibrium odds might not exist (agents not betting as per Lemma 1 and Lemma 2), we run an experiment with risk neutral agents that bet their entire wealth on only one event j that satisfies $p_{itj}O_{tj}>1$, i.e., the bets are discontinuous and equilibrium odds do not always exist (verified empirically). The combiner method main effect is evaluated and the results are statistically equivalent to the results in the main experiment.

Existence of Equilibrium Odds **-** Equilibriums odds are defined in IMF as the odds that give $Q_{tj}O_{tj} = Q_t$, and where Q_{tj} and Q_t are functions of O_{tj}. The proof in Appendix 3 shows that the equilibrium odd O_{t1} is equal to $\frac{\sum_{i\in D1\bigcup D2a} p_{it2}(w_{it}+m)+\sum_{i\in D2c}(km)}{\sum_{i\in D1\bigcup D2a} p_{it1}(w_{it}+m)+\sum_{i\in D2b}(km)}+1$. However, this can not be used directly to determine the existence of equilibrium odds O_{tj} for a given object t because of the recursive nature of O_{t1} and Q_{t1} (note that q_{itj} is used to determine if $i\in D_1$, $i\in D_{2a}$, $i\in D_{2b}$ or $i\in D_{2c}$, and that the agents use O_{t1} to determine q_{itj}). To empirically validate the existence of equilibrium odds (defined as odds such as $|Q_{tj}O_{tj} - Q_t|/Q_t < 0.0000000001$), we use binary search with $\varepsilon = 0.000000000000001$. Based on this setting, we find equilibrium odds for 98.32% of the objects.

Appendix 5: Equivalence of Net Benefit and Cost Savings

In a given classification context, Cost Savings (CS) is defined as the difference between the costs that would result if no classification system is used and the costs that results when a classification system is used (Chan et al. 1999). In the fraud context CS is defined as:

CS = P*Fraud Cost - (FN*Fraud Cost + (TP+FP)*Investigation Cost), (102)

where P is the number of fraud instances and Fraud Cost is the cost of one fraud instance.

simplify (102):

CS = (P-FN)*Fraud Cost - (TP+FP)*Investigation Cost (103)

simplify (103):

CS = Fraud Cost*TP - Investigation Cost*(TP+FP) (104)

CS is the same as Net Benefit, since Fraud Cost is the same as FN cost avoidance.

Appendix 6: Relation between Estimated Relative Cost and Net Benefit

Estimated Relative Cost (ERC) is defined as the cost per classified firm of undetected instances of financial statement fraud plus the cost of investigating non-fraudulent firms:

$$ERC = n^{FN}/\, n^{P} \text{ x } C^{FN} \text{ x } \text{P(Fraud)} + n^{FP}/\, n^{N} \text{ x } C^{FP} \text{ x P(Non-Fraud)}, \quad (105)$$

where P(Fraud) and P(Non-Fraud) are the assumed population fraud and non-fraud probabilities, i.e., $n^P/(n^P+n^N)$ and $n^N/(n^P+n^N)$, respectively; C^{FP} is the cost of false positive classifications, and C^{FN} is the cost of false negative classifications; n^{FP} is the number of false positive classifications, n^{FN} is the number of false negative classifications, n^P is the number of positive instances in the dataset and n^N is the number of negative instances in the dataset.

simplify (105):

$$ERC = n^{FN}/\, n^{P} \text{ x } C^{FN} \text{ x } \; n^{P}/(n^{P}+n^{N}) \; + n^{FP}/\, n^{N} \text{ x } C^{FP} \text{ x } n^{N}/(n^{P}+n^{N}), \quad (106)$$

simplify (106):

$$ERC = (n^{FN} \text{ x } C^{FN} + n^{FP} \text{ x } C^{FP})/(n^{P}+n^{N}) \quad (107)$$

substitute n^{FN} in (107) with $n^P - n^{TP}$ (note that $n^{FN} + n^{TP} = n^P$):

$$ERC = ((n^{P} - n^{TP}) \text{ x } C^{FN} \; + n^{FP} \text{ x } C^{FP})/(n^{P}+n^{N}) \quad (108)$$

simplify (108):

$$(n^{P} \text{ x } C^{FN})/(n^{P}+n^{N}) = (n^{TP} \text{ x } C^{FN} \; - n^{FP} \text{ x } C^{FP})/(n^{P}+n^{N}) + ERC \quad (109)$$

substitute C^{FN} in the RH of (109) with $C^P - C^i$ and C^{FP} with C^i, where C^P is the cost of fraud and C^i is the cost of investigation:

$$(n^{P} \text{ x } C^{FN})/(n^{P}+n^{N}) = (n^{TP} \text{ x } (C^{P}-C^{i}) - n^{FP} \text{ x } C^{i})/(n^{P}+n^{N}) + ERC \quad (110)$$

simplify (110):

$$(n^{P} \text{ x } C^{FN})/(n^{P}+n^{N}) = (C^{P} \text{ x } n^{TP} - C^{i} \text{ x } (n^{TP} + n^{FP}))/(n^{P}+n^{N}) + ERC \quad (111)$$

substitute $C^P \text{ x } n^{TP} - C^i \text{ x } (n^{TP} + n^{FP})$ in (111) with NB, i.e., Net Benefit, given that FN cost avoidance and investigation cost in Net Benefit are equivalent to the cost of fraud, C^P, and the cost of investigation, C^i, in ERC

$$(n^{P} \text{ x } C^{FN})/(n^{P}+n^{N}) = NB/(n^{P}+n^{N}) + ERC \quad (112)$$

substitute $(n^P \text{ x } C^{FN})/(n^P+n^N)$ in (112) with constant a, the average fraud cost of all firms. Note that $(n^P \text{ x } C^{FN})/(n^P+n^N)$ is constant in any given dataset:

$$a = NB/(n^{P}+n^{N}) + ERC \quad (113)$$

ERC plus net benefit per classified firm is equal to constant a. In a given dataset when net benefit per classified firm increases as the result of the classification effort ERC decreases by the same amount, and vice versa.