International Journal of Information Systems and Project Management

Volume 8 | Number 2 Article 5

2020

Data mining approach to internal fraud in a project-based organization

Mirjana Pejic-Bach
University of Zagreb

Ksenija Dumičić
University of Zagreb

Berislav Žmuk
University of Zagreb

Tamara Ćurlin
University of Zagreb

International Journal of
Information Systems and Project Management
ISSN (print):2182-7796, ISSN (online):2182-7788, ISSN (cd-rom):2182-780X
Available online at www.sciencesphere.org/ijispm

Data mining approach to internal fraud in a project-based organization

Mirjana Pejić Bach
University of Zagreb, Faculty of Economics and Business
Croatia
mpejic@efzg.hr

Tamara Ćurlin
University of Zagreb, Faculty of Economics and Business
Croatia
tcurlin@efzg.hr

Ksenija Dumičić
University of Zagreb, Faculty of Economics and Business
Croatia
kdumicic@efzg.hr

Jovana Zoroja
University of Zagreb, Faculty of Economics and Business
Croatia
jzoroja@efzg.hr

Berislav Žmuk
University of Zagreb, Faculty of Economics and Business
Croatia
bzmuk@efzg.hr

Abstract:
Data mining is an efficient methodology for uncovering and extracting information from large databases, which is widely used in different areas, e.g., customer relation management, financial fraud detection, healthcare management, and manufacturing. Data mining has been successfully used in various fraud detection and prevention areas, such as credit card fraud, taxation fraud, and fund transfer fraud. However, there are insufficient researches about the usage of data mining for fraud related to internal control. In order to increase awareness of data mining usefulness in internal control, we developed a case study in a project-based organization. We analyze the dataset about working-hour claims for projects, using two data mining techniques: chi-square automatic interaction detection (CHAID) decision tree and link analysis, in order to describe characteristics of fraudulent working-hour claims and to develop a model for automatic detection of potentially fraudulent ones. Results indicate that the following characteristics of the suspected working-hours claim were the most significant: sector of the customer, origin and level of expertise of the consultant, and cost of the consulting services. Our research contributes to the area of internal control supported by data mining, with the goal to prevent fraudulent working-hour claims in project-based organizations.

Keywords:
project-based organization; internal control; fraud; data mining; CHAID; association and link analysis.

DOI: 10.12821/ijispm080204

Manuscript received: 8 April 2019
Manuscript accepted: 7 December 2019

1. Introduction

Internal fraud has become one of the crucial and increasingly serious problems in numerous organizations. Internal control encompasses various policies and procedures designed for detecting and preventing fraud conducted by the organization's employees or external hires, which have to be constantly updated and monitored [1], in order to efficiently support the organization in its risk management activities. Internal fraud control is widely used for the purpose of forecasting, detecting and preventing possible fraudulent behaviors conducted by organizations' employees [2]. However, numerous organizations still have inefficient internal control systems [3].

Data mining techniques are widely used for external fraud detection and prevention. Literature review regarding data mining methods for the detection of financial fraud revealed that data mining techniques have been mostly used for detecting insurance fraud, corporate fraud, and credit card fraud [4]. Studies about internal control fraud are mostly focused on financial organizations and accounting [5], [6]. One of the examples of utilization of data mining for combating internal fraud investigates the utilization of data mining methods for detecting fraud by employees in a financial organization [7]. Project-based organizations are especially prone to internal fraud since due to the lower level of control that is the result of the flatter organizational structure [8], and in some cases a poor management practices [9] or complex governance procedures [10]. However, research about fraud detection and prevention in project-based organizations are scarce [11], [12].

In order to shed some light on the usefulness of the data mining approach for the detection of internal fraud in project-based organizations, we develop a case study, based on the dataset from one project-based organization. The dataset contains the characteristics of the working-hour claims (client, expert, job characteristics) in one project-based organization, which is analyzed by chi-square automatic interaction detection (CHAID) decision tree and link analysis. Using these two methods, we develop data mining models that discover the client, expert and job characteristics that are significant predictors of fraudulent working-hour claims. The contributions of this paper are two-fold. First, we contribute to the area of internal fraud detection and prevention in project-based organizations, while most of the previous research has been oriented towards external fraud prevention. Second, we provide practical contributions, since our research results in the form of decision tree and association rules could enable organizations for developing their own solutions for automatic internal fraud-detection (e.g., using SQL code).

The paper is organized into five sections. After the introduction, we present the literature review, with the goal of internal control, data mining and fraud prevention. In the methodology section, we overview the characteristics of the dataset, as well as the used methods (link analysis and CHAID decision tree). In the fourth section, we present research results, with the extensive elaboration of the rules extracted from the decision trees and link analysis. The last section concludes the paper with an overview of research, practical contributions, paper limitations and future research directions.

2. Literature review

2.1 Internal controls and fraud

Fraud represents a severe problem in companies; whether committed outside or inside an organization. Many organizations from various industries such as credit transactions; telecom, insurance, and management are affected by fraudulent activities [13]. Fraudsters could even be financial or other institutions themselves, involved in money laundering or financial statement frauds. A pilot survey for measurin financial fraud in the USA found out that the fraudster most commonly executed frauds online (30%) with the credit card payment (32%) [14]. Consider that those numbers are not even accurate because fraud is often not reported because of the possible negative impact on the organizations' image. On the other side, fraud committed inside the organization is also common, generating a high loss, both in terms of money and loose of trust [15].

The purpose of internal control is to detect and prevent fraudulent behavior, and thus support the company's performance and achieve established goals. Opportunities for fraud occur in organizations, which have weak

compliance with internal controls [16]. Internal fraud is a growing problem in many companies and organizations, which indicates that it is necessary to investigate this problem further and deeper in order to get better internal control systems [17]. On the other hand, many organizations lack the strategy to develop and maintain an efficient internal control system. Insufficient and flouted internal controls give opportunities for personnel to commit unethical practices and fraud in an organization [3]. There are numerous recommendations related to increasing the efficiency of internal control systems, such as the usage of global positioning tracking units (GPS), monitoring of unutilized purchase orders and pre-approval of overtime work. However, progress is slow due to difficult access to data from previous cases, so it is hard for problem solvers to develop new methods and solutions [2].

2.2 Data mining

The main task for data mining is to extract the most significant patterns from databases in various organizations and institutions. Data mining is acting as a tool that delivers data for further investigation, interpretation, and understanding [18]. Kantrarzic et al. [19] define data mining as "iterative process within which progress is defined by discovery, either through automatic or manual methods", acknowledging that the exploratory analysis scenario, without predetermined notion on the possible results, is the domain where data mining is the most useful. There are three fundamental goals for data mining processes: description, prediction, and prescription. Data describing human-interpretable patterns are focused on the description, while the usage of variables in the database to predict unfamiliar or forthcoming values of other variables is primarily focused on prediction [20]. The main objective of prescription is providing the best solution to the actual problem. All three goals are possible to accomplish by data mining techniques, such as classification, prediction, outlier detection, optimization, and visualization.

A number of challenges occur when considering the development and implementation of data mining [21], who stress the following: performance time, management support, selection and execution of algorithms. Although the first concern is usually the performance time (the importance of real-time action, online vs. offline methods), another big challenge that emerges is the cost management related to employee costs, consultants, software and hardware. The second concern would be the choice of the data mining technique. Data mining techniques have their own challenges in the development process: not all the data needed to perform tests is available to the public, and there is also a big lack of well-researched methods, algorithms, and techniques. The chosen method will depend on the structure of the data and the type of results that are wanted from the analysis. Finally, the main concern is focused on the actual usage of data mining results in the decision-making, it is rarely technical and usually depends on management willingness to support the application of data mining.

2.3 Data mining for internal fraud detection

In the last decade, significant progress took place, and automated fraud detection systems based on data mining models have gained enormous popularity, especially within financial institutions [4]. In terms of data mining, fraud analysis is a process, which consists of a sequence of actions, or a group of characteristics that could be used for predicting or discovering potential or explicit threats of fraudulent activities. Data mining has remarkable results in diverse fields related to security and fraud, financial crime detection (money laundering, suspicious credit card transactions and financial reporting fraud), intrusion and spam detection [22]. However, data mining implementation in the area of internal fraud risk reduction is mostly focused on the analysis of financial statements [23], [24], [25]. Kranacher et al. [22] distinguish three categories of internal fraud on which most studies are focused: financial statement fraud, transaction fraud, and abuse of position. Data mining techniques can decrease the probability of internal fraud. Various methods have been used for developing data mining models for internal fraud prevention and detection, such as multivariate latent clustering, neural networks, logistic models and decision trees [26], [27].

Data mining has become one of the most important paradigms of advanced intelligent business analytics and decision support tools for internal fraud prevention [28], [29], [23]. Many organizations acknowledge data mining as one of the main technologies relevant to internal fraud prevention nowadays and in the future. The Institute of Internal Auditors – Australia [30] recommends the usage of data mining for auditing process, and The Chartered Global Management

Accountant has reported that data mining lies within the top ten focus priorities fundamental for the data-driven era of business and was ranked as relevant by more than half corporate leaders [31].

3. Methodology

3.1 Data

In order to inspect internal fraud, we have conducted a case study analysis on the data available from one large company. This company is organized using a project-based organizational structure, which means that projects present the key organizational activity [32], [33]. The company has more than 300 employees and implements and develops business-related software applications. Each month, employees working on a project-basis provide a report on their work including the number of hours, the characteristics of clients, the complexity of their work, and the amount claimed for an hour and in total. Based on this information, the working-hours claim is filled each month. The company has already developed its own methods for detecting suspicious working-hour claims, but those are focused on the detection of already committed fraudulent activities, while more research is needed in order to identify the characteristics of fraudulent claims in order to detect potential new ones. Therefore, the goal of this research is to determine the characteristics of the suspected working-hour claims, which are the candidates for in-depth fraud analysis, and to develop a model for preventing fraudulent behavior.

The company defines the suspect claims in the following manner. A working-hours claim is suspect if at least one of the following criteria has been met: (i) if a consultant is late in submitting the working-hours claim more than seven days from the day when the project is finished, and (ii) if a consultant cancels already claimed working-hours. In the case when at least one of the abovementioned criteria is fulfilled, the working-hours claim is considered as a suspect for fraud. The management of the company believed that it would be beneficial to identify the characteristics of the potential fraud (suspect) working-hour claims before the consultant is already late in submitting the claim.

Dataset consists of 1,194 working-hours claims, which comprise 5% of the total working-hours claims in the company in the observed year. According to Table 1, 294 working-hours claims, or 24.62%, were suspect for fraud whereas 900 working-hours claims, or 75.38%, were non-suspect for fraud. The variable Suspect defines these two categories of working-hours claims (if the claim is suspected it has value 1, otherwise it is equal to 0).

Table 1. Suspect and non-suspect working-hour claims in the sample

Variable Suspect	Count	Percent
Suspect (value 1)	294	24.62
Non-suspect (value 0)	900	75.38
Total	1,194	100.00

Source: Authors' work, based on the internal data source.

The independent variables in the working-hour claims are used for developing data mining models:

- Type of customer – variable Customer;
- Type of consultant – variable Consultant;
- The month when the working-hours were claimed – variable Month;
- The hourly-rate – variable UnitPriceCoded;
- The consultant's level of expertise – variable ExpertLevel;
- The number of hours claimed – variable NoHoursCoded;
- The total amount claimed – variable TotalAmountCoded.

The following analysis will present the distribution of the independent variables according to the fraudulent working-hour claims.

The distribution of the variable Customer is presented in Table 2. Customers ordering the work on the project (development and/or implementation of software applications) are divided into three categories: governmental institutions, internal projects, and private enterprises. Internal projects are suspected in a 50.68% case. The conducted chi-square test confirmed, at the significance level of 1%, that there is at least one category of customers whose structure according to the variable Suspect is different from the others (chi-square=77.435, df=2, p-value<0.001).

Table 2. Types of the customer – variable Customer

Customer origin	**Suspect**	**Not suspect**	Chi-square	P-value
Govern	4.76%	95.24%	77.435	<0.001
Internal	50.68%	49.32%		
Private	22.80%	77.20%		
Totals	24.62%	75.38%		

Source: Authors' work, based on the internal data source.

The variable Consultant describes the country of origin of experts, who have been claiming working-hours, since in some cases domestic consultants (from Croatia) and in some cases, foreign consultants are hired (Table 3). In cases when domestic consultants are observed, 23.46% of their working-hour claims were suspected, while foreign consultants were in 41.56% cases in the suspected working-hours claim category. The chi-square test has shown that, at the significance level of 1%, domestic and foreign employees have a statistically significantly different structure according to suspected and non-suspected working-hours claims (chi-square=12.719, df=1, p-value<0.001).

Table 3. Types of consultant – variable Consultant

Consultant origin	**Suspect**	**Not suspect**	**Chi-square**	**P-value**
Domestic	23.46%	76.54%	12.719	<0.001
Foreign	41.56%	58.44%		
Totals	24.62%	75.38%		

Source: Authors' work, based on the internal data source.

The variable Month represents the month in which a consultancy service was provided (Table 4). For the purpose of the analysis, months are coded as discrete values ranging from M1 to M12. The highest share of suspected working-hours claims can be found in months M1 (65.77%) and M12 (30.59%), which refer to January and December. It is highly probable that this large percentage of suspect claims are related to the beginning and the end of the fiscal year. On the other hand, the highest share of non-suspected working-hours claims is in months M10 (88.42%) and M4 (86.40%). According to the conducted chi-square test, those shares seem to be statistically significantly different, at the significance level of 1%, in different months (chi-square=134.670, df=11, p-value<0.001).

Table 4. The month when the working-hours were claimed – variable Month

The month of the claim	Suspect	Not suspect	Chi-square	P-value
M1	65.77%	34.23%	134.670	<0.001
M2	28.13%	71.88%		
M3	17.31%	82.69%		
M4	13.60%	86.40%		
M5	16.81%	83.19%		
M6	20.39%	79.61%		
M7	14.29%	85.71%		
M8	28.09%	71.91%		
M9	25.93%	74.07%		
M10	11.58%	88.42%		
M11	19.28%	80.72%		
M12	30.59%	69.41%		
Totals	24.62%	75.38%		

Source: Authors' work, based on the internal data source.

The variable UnitPriceCoded was used to take into account the cost of consultants (Table 5). In the analysis, this cost is expressed per hour. The minimum cost per hour is 19.9 EUR, and the highest is 173.9 EUR per hour. Because there are many different values, it has been decided that four groups of costs will be formed and that the unit price will be coded in four categories (1-50 EUR per hour, 51-100 EUR per hour, 101-150 EUR per hour, and 151-200 EUR per hour). The largest share of suspected working-hours claims was found in the category of the cost of 151-200 EUR (30.00%) whereas the largest share of non-suspected working-hours claims was found in the category of the cost of 1-51 EUR (89.19%). The chi-square test has shown that, at the significance level of 5%, the hypothesis of equal shares of suspected working-hours claims, or non-suspected working-hours claims, at all the four observed cost levels cannot be rejected (chi-square=6.278, df=3, p-value=0.099).

Table 5. The hourly-rate – variable UnitPriceCoded

The hourly rate	Suspect	Not suspect	Chi-square	P-value
1-50 EUR per hour	10.81%	89.19%	6.278	0.099
51-100 EUR per hour	25.59%	74.41%		
101-150 EUR per hour	18.75%	81.25%		
151-200 EUR per hour	30.00%	70.00%		
Totals	24.62%	75.38%		

Source: Authors' work, based on the internal data source.

The variable Expert Level (Table 6) reflects the five expert levels coded from L4 to L8, which refer to the experience and relevant knowledge of consultants claiming working-hours (L4 is the lowest level of expertise, while L8 is the top

level of expertise). The highest share of suspected working-hours claims is present at the expert level L5 (77.78%) whereas the highest share of non-suspected working-hours claims is present at the expert level L4 (89.19%). The chi-square test confirmed that, at the significance level of 1%, there is at least one expert level at which shares of suspected working-hours claims or non-suspected working-hours claims are statistically significantly different than at other expert levels (chi-square=33.147, df=4, p-value<0.001).

Table 6. The consultant's level of expertise – variable ExpertLevel

Consultant	Suspect	Not suspect	Chi-square	P-value
L6	24.69%	75.31%	33.147	<0.001
L5	77.78%	22.22%		
L8	30.00%	70.00%		
L4	10.81%	89.19%		
L7	18.75%	81.25%		
Totals	24.62%	75.38%		

Source: Authors' work, based on the internal data source.

Table 7 outlines the number of weekly working hours of employees or consultants (the variable NoHoursCoded). There is a quite large number of discrete values of weekly working hours. Consequently, they are classified into eight groups: 1-5; 6-10; 11-15; 16-20; 21-25; 25-30; 31-35; and 36-55. Due to some administrative problems, an additional category was introduced to incorporate negative weekly working-hours, which appeared due to some corrections conducted by consultants themselves. It is a company policy that, in the case of negative weekly working-hours, these working-hour claims are treated as suspected. The Chi-square test has shown that, at the significance level of 1%, there is at least one weekly working-hours category at which the share of suspected working-hours claims is statistically significantly different than at other weekly working-hours categories (chi-square=53.859, df=8, p-value<0.001).

Table 7. The number of hours claimed – variable NoHoursCoded

The number of hours	Suspect	Not suspect	Chi-square	P-value
Negative hours	100.00%	0.00%	53.859	<0.001
1-5 hours	22.63%	77.37%		
6-10 hours	26.06%	73.94%		
11-15 hours	26.90%	73.10%		
16-20 hours	23.39%	76.61%		
21-25 hours	17.65%	82.35%		
25-30 hours	9.68%	90.32%		
31-35 hours	23.08%	76.92%		
36-55 hours	21.05%	78.95%		
Totals	24.62%	75.38%		

Source: Authors' work, based on the internal data source.

The costs of consultants' working-hours are observed by the variable TotalAmountCoded (Table 8), and those costs have been categorized into 19 cost categories. A negative amount is claimed for the working-hour claims with negative hours, which was elaborated for the variable NoHoursCoded (Table 7). The conducted chi-square test has shown that, at the significance level of 1%, there is at least one total cost per consultant category at which the share of suspected working-hours claims is statistically significantly different (chi-square=80.068, df=18, p-value<0.001).

Table 8. The total amount claimed – variable TotalAmountCoded.

The total amount claimed	Suspect	Not suspect	Chi-square	P-value
1-100 EUR	25.30%	74.70%	80.068	<0.001
101-200 EUR	16.94%	83.06%		
201-300 EUR	24.56%	75.44%		
301-400 EUR	24.17%	75.83%		
401-500 EUR	24.21%	75.79%		
501-600 EUR	24.51%	75.49%		
601-700 EUR	31.52%	68.48%		
701-800 EUR	16.67%	83.33%		
801-900 EUR	13.33%	86.67%		
901-1000 EUR	42.42%	57.58%		
1001-1100 EUR	35.19%	64.81%		
1101-1300 EUR	27.87%	72.13%		
1301-1500 EUR	29.33%	70.67%		
1501-1600 EUR	5.56%	94.44%		
1601-1700 EUR	15.38%	84.62%		
1701-2000 EUR	22.22%	77.78%		
2001-3000 EUR	18.97%	81.03%		
3001-4000 EUR	13.79%	86.21%		
Negative	100.00%	0.00%		
Totals	24.62%	75.38%		

Source: Authors' work, based on the internal data source.

3.2 CHAID decision tree

In order to provide an understanding of the interrelation between working hours claim fraud and various characteristics, such as characteristics of customers, consultants, expert knowledge and others, a decision tree is developed using the CHAID algorithm. As the name reveals, the CHAID decision tree is based on the chi-square test, which is used to select the best split at each step. In order to construct a decision tree, the role of the dependent variable was given to the variable Suspect. All other observed variables have taken the role of independent variables (Customer; Consultant; Month; UnitPriceCoded; ExpertLevel; NoHoursCoded; TotalAmountCoded). In order to get a clear and easily understandable classification tree, it has been decided that the classification tree depth should go up to the third level, which is indicated by Bertsimas et al. (2017) [34], as the optimal depth of the tree. Furthermore, it has been defined that the main or parent node should have at least 100 cases whereas the following or child nodes should have at least 50

cases, which comprise approximately 8% and 4% respectively of the total sample (1,194 cases). The decision tree is developed using SPSS ver. 23.

3.3 Link analysis

Link analysis is a data analysis technique, which can be used for identification and evaluation of relationships between items that occur together, and which can be represented as "nodes." Different objects like enterprises, employees, customers, transactions, and similar can be referred to as nodes. Link analysis is used for the detection of potentially suspect working-hours claims based on characteristics of clients, consultants, and projects. By using link analysis, the association rules are extracted in order to detect significant relationships between suspect working-hours claims and various characteristics of customers, consultants, and projects. Association rules can be described as:

$$\text{If A=1 and B=1 then C=1 with probability p} \tag{1}$$

where A, B, and C are binary variables, p is a conditional probability defined as $p = p(C = 1|A = 1, B = 1)$. Furthermore, the association rule can be simply written as $A \Rightarrow B$, where A is the body of the rule and B is the head of the association rule [35].

In the analysis, all eight variables are included: Suspect; Customer; Consultant; Month; UnitPriceCoded; ExpertLevel; NoHoursCoded; TotalAmountCoded. Because there is no defined and strict order between variables and items, it has been decided that the non-sequential association analysis approach will be applied [36]. Link analysis has been conducted using Statistica Data Miner software ver. 13.5.

The minimum support value, which shows how frequently an itemset appears in the dataset, has been set to value 0.2 whereas the maximum value was set to 1.0. Support is calculated as:

$$\text{Support } (A \Rightarrow B) = p(A \cup B) \tag{2}$$

Items with support value lower than the minimum value will be excluded from the analysis. Similar, the minimum confidence value was set to 0.1 and the maximum value to 1.0. Confidence settings define how often the rule came out to be true. Again, items with confidence value lower than the minimum value will be excluded from the analysis. Confidence is calculated using the following equation:

$$\text{Confidence } (A \Rightarrow B) = p(B \mid A) = \text{Support } (A, B) / \text{Support } (A) \tag{3}$$

Additional, it has been defined that the maximum number of items in an item set is 10.

It has to be emphasized that there are no strict rules in the literature that minimum support value; minimum confidence value or the maximum number of items in an item set should be selected [37]. Other authors in their work use subjective criteria for selecting association rules [38], [39]. Therefore, the limits are here used as described before because the experiments with the different level of metrics indicated that they result in interesting rules.

4. Results

4.1 Decision tree

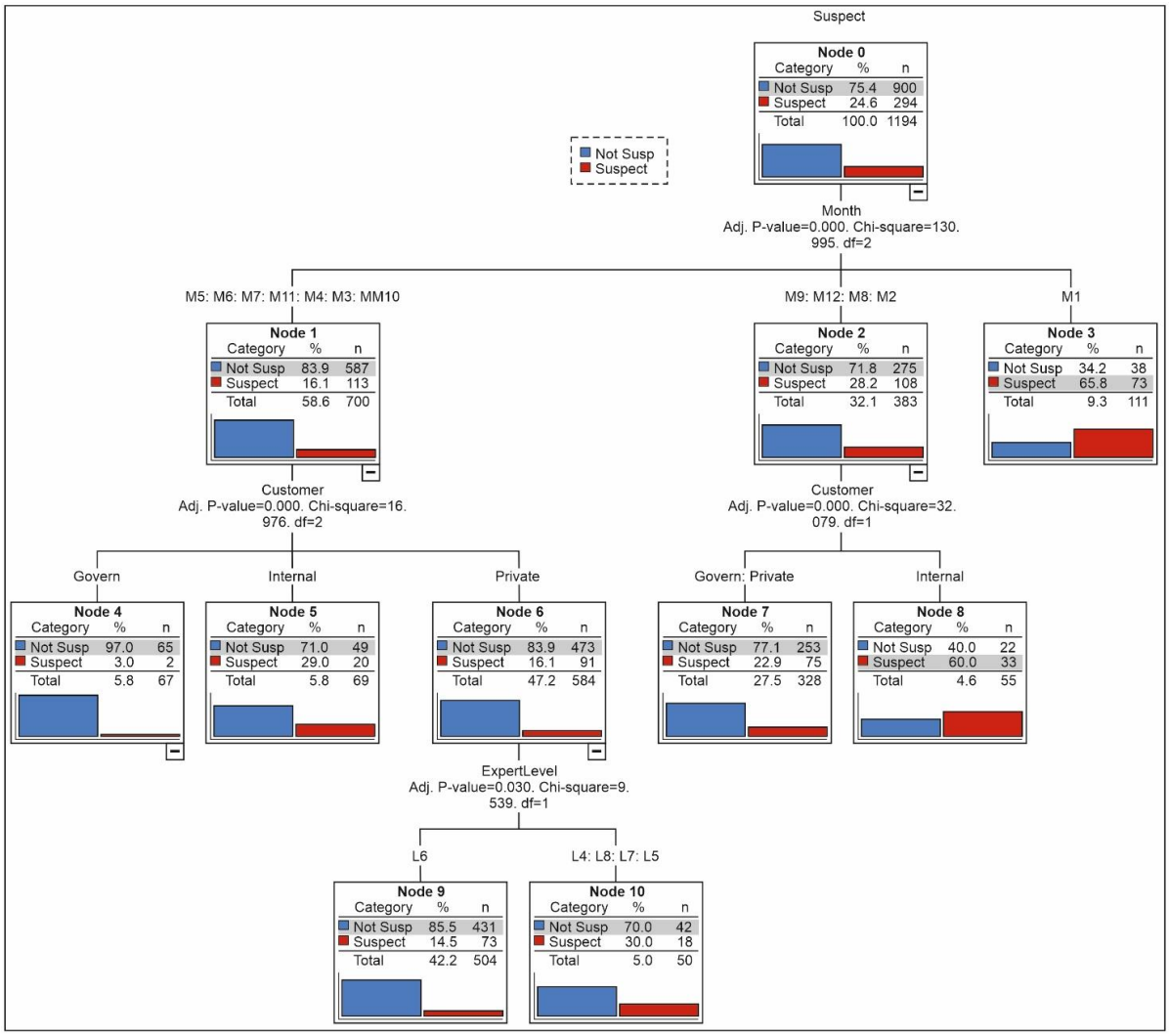
According to defined settings, the CHAID decision tree is developed (Figure 1). The resulting CHAID decision tree has 3 levels and overall 11 nodes out of which seven are considered as a terminal (they do not split further). Figure 1 also reveals that variables Month, Customer and ExpertLevel had the highest level of statistical significance and therefore they are used in building the classification tree.

The variable used for branching on the first level is the variable Month, which turned out to be statistically significant at the level of 1% (chi-square=130.995, p-value<0.001). This branching resulted in three new nodes (Node 1, Node 2, and Node 3). Node 1 includes categories M3, M4, M5, M6, M7, M10, and M11. That way Node 1 consists of 700 working-hours claims out of which 587 or 83.9% are treated as non-suspected whereas 113 or 16.1% are suspected. Node 2 includes the following categories of the variable Month: M2; M8; M9; and M12. Consequently, Node 2 has in total 383 working-hours claims out of which 275 or 71.8% are non-suspected whereas 108 or 28.2% are suspected. Node 3 includes only the category M1 and only at this node, the share of suspected working-hours claims (65.8%) is greater than the share of non-suspected working-hours claims (34.2%).

The variable Customer was used for branching on the second level. According to Figure 1, branching resulted in five new nodes with three of them (Node 4, Node 5, and Node 6) coming out from Node 1 and two of them (Node 7 and Node 8) from Node 2. Both branching processes are highly statistically significant at 1% (from Node 1 – chi-square=16.976, p-value<0.001; from Node 2 – chi-square=32.079, p-value<0.001). Node 4 includes only 67 customers of government institutions out of which 65 or 97.0% are connected with non-suspected working-hours claims, and two or 3.0% are connected with suspected working-hours claims. Node 5 consists of 69 customers of internal projects out of which 49 or 71.0% are connected with non-suspected working-hours claims and 20 or 29.0% are connected with suspected working-hours claims. Node 6 includes only customers of private enterprises, and it is the largest one among nodes of the second level. There are 473 or 83.9% customers of private enterprises that are connected with non-suspected working-hours claims and 91 or 16.1% that are connected with suspected working-hours claims. On the other hand, Node 7, which is related to Node 2, includes customers of government institutions and customers of private enterprises together. It has been shown that out of 328 customers 253 or 77.1% are non-suspected whereas 75 or 22.9% are suspected for working-hours claim fraud. Node 8 includes only customers of internal projects. When nodes of the second level are observed, it can be concluded that only at this node the share of suspected working-hours claims (60.0) is higher than the share of non-suspected working-hours claims (40.0%).

The third level branching variable is the variable ExpertLevel. This variable was used to branch Node 6 further into two new nodes (Node 9 and Node 10). This branching process is statistically significant at the 5% level (chi-square=9.539, p-value=0.030). Node 6 consists only of consultants with the expert level L6 whereas consultants with levels L4, L5, L7, and L8 can be found in Node 10. Node 9 is considerably larger than Node 10 and includes 431 or 85.5% non-suspected working-hours claims and 73 or 14.5% suspected working-hours claims. Furthermore, it has to be emphasized that Node 9 includes 42.2% of all observed working-hours claims whereas Node 10 includes only 5.0% of them. Therefore, Node 10 includes 60 working-hours claims out of which 42 or 70% are non-suspected whereas 18 or 30.0% are suspected.

The classification matrix, shown in Table 9, compares the observed and the predicted status of working-hours claims. The used algorithm was correct in 93.3% of cases for the non-suspect working-hour claims. In other words, out of 900 non-suspected working-hours claims, the algorithm has correctly classified 840 of them, whereas 60 working-hours claims were wrongly classified. The successfulness of the algorithm seems to be quite low in relation to suspected working-hours claims. Namely, out of 294 suspected working-hours claims, the algorithm correctly classified 106 working-hours claims or 36.1%.



Source: Authors' work, based on the internal data source.

Figure 1. CHAID decision tree

Table 9. The number of hours claimed – variable NoHoursCoded

Observed classification	**Predicted classification**		
	Non-suspect	**Suspect**	**Percent correct**
Non-suspect	840	60	93.3%
Suspect	188	106	36.1%
Overall percentage	86.1%	13.9%	79.2%

Source: Authors' work, based on the internal data source.

4.2 Link analysis

Using the selected metrics (minimum support value of 0.2; minimum confidence value of 0.1 and the maximum number of items in an item set of 0), the association rules have been developed. Table 10 presents the most frequent itemsets that contain Suspect item, indicating that the suspectable amount of working hours has been claimed. The item Suspect alone, with the frequency of 235, appears in the 27.71% of itemsets. Item Suspect in combinations with other items, such as Private, Domestic 51-100 and L6 can also be found in a significant number of projects. Consequently, it can be concluded that suspected working-hour claims are very closely related and linked with customers from private enterprises, with domestic consultants, with cost per hour between 51 and 100 EUR, and with expert level L6. Those relations are presented graphically in Figures 2 and 3 as well.

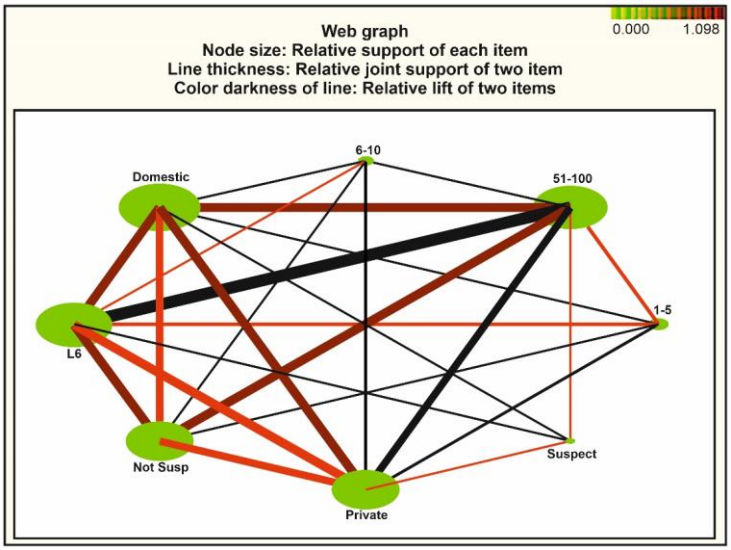
Table 10. Frequent itemsets that contain Suspect item

Frequent itemsets	Number of items	Frequency	Support (%)
Suspect	1	235	27.712
51-100, Suspect	2	225	26.533
51-100, L6, Suspect	3	221	26.061
L6, Suspect	2	221	26.061
Domestic, Suspect	2	220	25.943
51-100, Domestic, Suspect	3	210	24.764
51-100, Domestic, L6, Suspect	4	206	24.292
Domestic, L6, Suspect	3	206	24.292
Private, Suspect	2	193	22.759
Domestic, Private, Suspect	3	185	21.816
51-100, Private, Suspect	3	183	21.580
L6, Private, Suspect	3	180	21.226
51-100, L6, Private, Suspect	4	180	21.226
51-100, Domestic, Private, Suspect	4	175	20.636
51-100, Domestic, L6, Private, Suspect	5	172	20.283
Domestic, L6, Private, Suspect	4	172	20.283

Source: Authors' work, based on the internal data source.

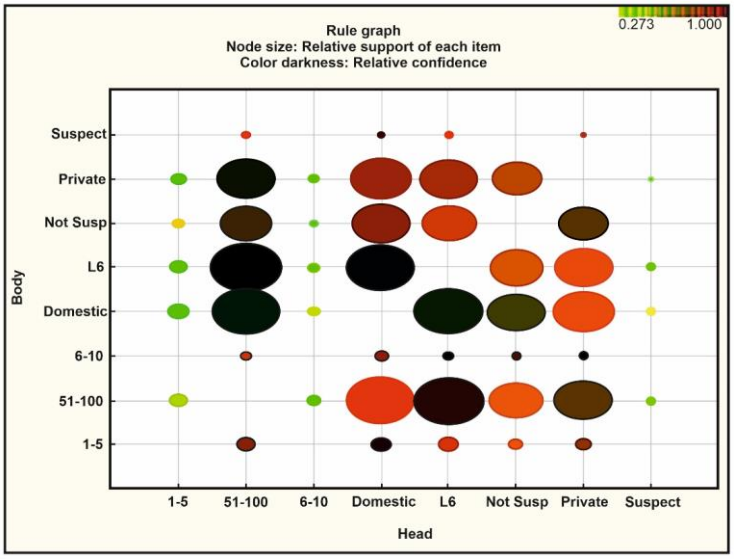
Figure 2 presents a Web graph of items generated by link analysis. Node size indicates the relative support for each item, line thickness relative joint support of two items, and color darkness of line a relative lift of two items. It can be observed that the most important nodes are related to the domestic experts, non-suspected claims, the lowest level of expertise (L6), private customers, and one of the low level of hourly paid rate (51-100 EUR). The strongest joint support is for the claims that are non-suspected and the domestic experts, the lowest level of expertise (L6), private customers, and one of the low level of hourly paid rate (51-100 EUR). As expected the darkest line presents the strength of the relationship between the lowest level of expertise (L6) and one of the low levels of hourly paid rate (51-100 EUR).

Figure 3 presents a rule graph of items generated by link analysis. Node size presents relative support of each item, and color darkness relative confidence. Again, the rule with the highest confidence and support is the relationship between the lowest level of expertise (L6) and one of the low level of hourly paid rate (51-100 EUR). It can be noted that the rules that contain the item Suspect are presented with small node sizes, and include the relationships between the item Suspect and the low level of hourly paid rate (51-100 EUR), domestic experts, the lowest level of expertise (L6), and private companies as customers.



Source: Authors' work, based on the internal data source.

Figure 2. Web graph of items generated by link analysis



Source: Authors' work, based on the internal data source.

Figure 3. Rule graph of items generated by link analysis

Table 11 presents association rules with the item Suspect in the body. The first rule shows that 26.53% of working-hours claims are suspected and with cost per hour between 51 and 100 EUR. Furthermore, it seems that 95.75% suspected working-hours claims are with cost per hour between 51 and 100 EUR. The second and third rules resulted in the same support and confidence levels.

Table 11. Frequent association rules with the item Suspect in the body

Body	==>	Head	Support (%)	Confidence (%)	Lift
Suspect	==>	51-100	26.533	95.745	1.052
Suspect	==>	51-100, L6	26.061	94.043	1.042
Suspect	==>	L6	26.061	94.043	1.042
Suspect	==>	Domestic	25.943	93.617	0.985
Suspect	==>	51-100, Domestic	24.764	89.362	1.031
Suspect	==>	51-100, Domestic, L6	24.292	87.660	1.021
Suspect	==>	Domestic, L6	24.292	87.660	1.021
Suspect	==>	Private	22.759	82.128	0.992
Suspect	==>	Domestic, Private	21.816	78.723	0.995
Suspect	==>	51-100, Private	21.580	77.872	1.025
Suspect	==>	51-100, L6, Private	21.226	76.596	1.016
Suspect	==>	L6, Private	21.226	76.596	1.016
Suspect	==>	51-100, Domestic, Private	20.637	74.468	1.027
Suspect	==>	51-100, Domestic, L6, Private	20.283	73.191	1.017
Suspect	==>	Domestic, L6, Private	20.283	73.191	1.017

Source: Authors' work, based on the internal data source.

Table 12 presents association rules with the item Suspect and one more item in the Body. If items Suspect and Private are in the Body, the strongest association is achieved with item Domestic. In that case, 21.82% of working-hours claims are suspected working-hours claims, with customers from private enterprises and with domestic consultants. It appears that 95.86% of suspected working-hours claims with customers from private enterprises include domestic consultants. If items Suspect and L6 are put together in the Body, the strongest association is achieved with item 51-100. It has been shown that all suspected working-hours claims with expert level L6 are related to cost per hour between 51 and 100 EUR. If items Suspect and Domestic are together in the Body, again the strongest association is achieved with item 51-100. However, 95.46% of suspected working-hours claims with domestic consultants have a cost per hour between 51 and 100 EUR.

Association rules with the item Suspect and two or more items in the Body are presented in Table 13. Suspected working-hours claims with customers from private enterprises and with expert level L6 have a cost per hour between 51 and 100 EUR. The same conclusion can be brought when items Domestic, L6 and Suspect are associated with item 51-100; and when items Domestic, L6, Private and Suspect are associated with item 51-100.

Table 12. Association rules with the item Suspect and one more item in the Body

Body	==>	Head	Support (%)	Confidence (%)	Lift
Private, Suspect	==>	Domestic	21.816	95.855	1.008
Private, Suspect	==>	51-100	21.580	94.819	1.042
Private, Suspect	==>	51-100, L6	21.226	93.264	1.034
Private, Suspect	==>	L6	21.226	93.264	1.034
Private, Suspect	==>	51-100, Domestic	20.637	90.674	1.046
Private, Suspect	==>	51-100, Domestic, L6	20.283	89.119	1.038
Private, Suspect	==>	Domestic, L6	20.283	89.119	1.038
L6, Suspect	==>	51-100	26.061	100.000	1.098
L6, Suspect	==>	51-100, Domestic	24.292	93.213	1.075
L6, Suspect	==>	Domestic	24.292	93.213	0.981
L6, Suspect	==>	51-100, Private	21.226	81.448	1.072
L6, Suspect	==>	Private	21.226	81.448	0.984
L6, Suspect	==>	51-100, Domestic, Private	20.283	77.828	1.073
L6, Suspect	==>	Domestic, Private	20.283	77.828	0.984
Domestic, Suspect	==>	51-100	24.764	95.455	1.049
Domestic, Suspect	==>	51-100, L6	24.292	93.636	1.038
Domestic, Suspect	==>	L6	24.292	93.636	1.038
Domestic, Suspect	==>	Private	21.816	84.091	1.016
Domestic, Suspect	==>	51-100, Private	20.637	79.545	1.047
Domestic, Suspect	==>	51-100, L6, Private	20.283	78.182	1.038
Domestic, Suspect	==>	L6, Private	20.283	78.182	1.038
51-100, Suspect	==>	L6	26.061	98.222	1.089
51-100, Suspect	==>	Domestic	24.764	93.333	0.982
51-100, Suspect	==>	Domestic, L6	24.292	91.556	1.066
51-100, Suspect	==>	Private	21.580	81.333	0.982
51-100, Suspect	==>	L6, Private	21.226	80.000	1.062
51-100, Suspect	==>	Domestic, Private	20.637	77.778	0.983
51-100, Suspect	==>	Domestic, L6, Private	20.283	76.444	1.063

Source: Authors' work, based on the internal data source.

Table 13. Association rules with the item Suspect and two or more items in the Body

Body	==>	Head	Support (%)	Confidence (%)	Lift
L6, Private, Suspect	==>	51-100	21.226	100.000	1.098
L6, Private, Suspect	==>	51-100, Domestic	20.283	95.556	1.102
L6, Private, Suspect	==>	Domestic	20.283	95.556	1.005
Domestic, Private, Suspect	==>	51-100	20.637	94.595	1.039
Domestic, Private, Suspect	==>	51-100, L6	20.283	92.973	1.031
Domestic, Private, Suspect	==>	L6	20.283	92.973	1.031
Domestic, L6, Suspect	==>	51-100	24.292	100.000	1.098
Domestic, L6, Suspect	==>	51-100, Private	20.283	83.495	1.099
Domestic, L6, Suspect	==>	Private	20.283	83.495	1.009
Domestic, L6, Private, Suspect	==>	51-100	20.283	100.000	1.098
51-100, Private, Suspect	==>	L6	21.226	98.361	1.090
51-100, Private, Suspect	==>	Domestic	20.637	95.628	1.006
51-100, Private, Suspect	==>	Domestic, L6	20.283	93.989	1.095
51-100, L6, Suspect	==>	Domestic	24.292	93.213	0.981
51-100, L6, Suspect	==>	Private	21.226	81.448	0.984
51-100, L6, Suspect	==>	Domestic, Private	20.283	77.828	0.984
51-100, L6, Private, Suspect	==>	Domestic	20.283	95.556	1.005
51-100, Domestic, Suspect	==>	L6	24.292	98.095	1.087
51-100, Domestic, Suspect	==>	Private	20.637	83.333	1.007
51-100, Domestic, Suspect	==>	L6, Private	20.283	81.905	1.087
51-100, Domestic, Private, Suspect	==>	L6	20.283	98.286	1.089
51-100, Domestic, L6, Suspect	==>	Private	20.283	83.495	1.009

Source: Authors' work, based on the internal data source.

5. Conclusions

A case study analysis was conducted using data related to suspected working-hour claims in one project-based company. We aim to identify the relationship of the suspect working-hour claims with selected variables, related to characteristics of customers, consultants, and work conducted (e.g., private and government customers; domestic or foreign consultants; the month of the work conducted and hourly rate). We develop two data mining models that identified the following characteristics of fraudulent working-hour claims: customers are private enterprises, consultants are of domestic origin and with the lowest level of expertise, and the cost of the consulting services are within the lowest range. First, the CHAID decision tree was developed in order to determine the relationships between numerous characteristics of the project (e.g., characteristics of the client and the expert), and suspect working-hour claims. The results of the decision tree showed a general rate of nearly 80% of correct classification. Second, the link analysis was used for the detection of potentially suspect working-hours claims. Both decision tree and link analysis indicate that suspected working-hours claims are related to customers from private enterprises, domestic consultants, cost per hour between 51 and 100 EUR, and the lowest level of expertise.

This paper contributes to the growing body of work that investigates internal fraud prevention and detection. However, most of the work conducted in this area is focused on the analysis of financial reports and accounting fraud [5], [6], [7], while in our work, we focus to project-based organizations. This research has demonstrated the use of a data mining methodology to detect internal fraud. Our proposition was that it is possible to develop a data mining application that could be useful for project-based organizations in predicting and detecting fraudulent working-hour claims. Although the decision tree algorithm is more efficient in predicting non-suspect working-hour claims than in suspect ones, and the confidence and support levels for suspect claims were rather low, the management from the company confirmed that the information derived is valid to them since it provided new insight into the characteristics of suspect working-hour claims. This information allows them to focus their efforts on the following categories identified by the decision tree as the most likely to be suspected: working-hour claims submitted in M1 by the internal experts. In addition, the general rate of correct classification of 79.2% can be observed as quite good [40]. Based on the presented results, it can be concluded that the decision tree and link analysis are recommended for use as a supportive instrument for the detection of suspect working-hour claims, in combination with other human-based and machine-based methods.

Our research has significant practical implications. Considering that auditors need non-accounting and non-financial data with no external standards to apply, it is likely that auditors will need to develop their own set of procedures to determine the quality of non-financial data [41]. Therefore, it is important that organizations expand usage and potentials of different data mining techniques, which could help them to be more effective and efficient in investigating and preventing internal fraud [17]. Project-based organizations often learn implicitly from experience [42], aiming to capture and share project-based knowledge, thus indicating that data mining could be widely accepted in their learning-oriented cultures [43]. One of the possible operationalizations of our work in this direction is the usage of SQL code that is generated by the software used for the development of the CHAID decision tree (Appendix 1), which can be used for the development of the solution for automatic internal fraud-detection.

Limitations of the paper derive mainly from sample characteristics since we presented one case study for one specific company and the usage of two data mining methods. Therefore, in order to test if our results are generally applicable, future research should be focused on datasets from organizations from different settings, using a broader set of data mining techniques, which would improve the knowledge regarding discovering patterns in internal fraud in project-based organizations using data mining techniques.

Acknowledgments

This paper extends the research on the internal fraud using the CHAID decision tree that was presented on CENTERIS - Conference on ENTERprise Information Systems [11]. This research has been fully supported by the Croatian Science Foundation under the PROSPER (Process and Business Intelligence for Business Performance) project (IP-2014-09-3729).

References

[1] ACFE (2018). *Report to the Nations: 2018 Global Study on Occupational Fraud and Abuse* [Online]. Available: https://www.acfe.com/uploadedFiles/ACFE_Website/Content/rttn/2018/RTTN-Government-Edition.pdf

[2] K. M. Zakaria, A. Nawawi and A. S. A. P. Salin, "Internal controls and fraud–Empirical evidence from oil and gas company," *Journal of Financial Crime*, vol. 23, no. 4, pp. 1154-1168, 2016.

[3] G. Baldock, "The perception of corruption across Europe, Middle East and Africa," *Journal of Financial Crime* vol. 23, no. 1, pp. 119-131, 2016.

[4] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559-569, 2011.

[5] M. Jans, N. Lybaert and K. Vanhoof, "Internal fraud risk reduction: Results of a data mining case study," *International Journal of Accounting Information Systems*, vol. 11, no. 1, pp. 17-41, 2010.

[6] M. Sánchez, J. Torres, P. Zambrano and P. Flores, "FraudFind: Financial fraud detection by analyzing human behavior," in *IEEE 8th Annual Computing and Communication Workshop and Conference*, Las Vegas, USA, 2018, pp. 281-286.
[7] J. B. Suh, R. Nicolaides and R. Trafford, "The effects of reducing opportunity and fraud risk factors on the occurrence of occupational fraud in financial institutions," *International Journal of Law, Crime and Justice*, vol. 56, pp. 79-88, 2019.
[8] C. Smith, *Making Sense of Project Realities: Theory, Practice and the Pursuit of Performance*. Abingdon, UK: Routledge, 2017.
[9] S. Bathallath, Å. Smedberg and H. Kjellin, "Managing project interdependencies in IT/IS project portfolios: A review of managerial issues," *International Journal of Information Systems and Project Management*, vol. 4, no. 1, pp. 67-82, 2016.
[10] D. R. Chandra and J. van Hillegersberg, "Governance of inter-organizational systems: A longitudinal case study of Rotterdam's Port Community System," *International Journal of Information Systems and Project Management*, vol. 6, no. 2, pp. 47-68, 2018.
[11] M. Pejić Bach, K. Dumičić, B. Žmuk, T. Ćurlin and J. Zoroja, "Internal fraud in a project-based organization: CHAID decision tree analysis," *Procedia Computer Science*, vol. 138, pp. 680-687, 2018.
[12] Z. Kahvandi, E. Saghatforoush, A. ZareRavasan and C. Preece, "Integrated project delivery implementation challenges in the construction industry," *Civil Engineering Journal*, vol. 5, no. 8, pp. 1672-1683, 2019.
[13] P. Gottschalk, "Categories of financial crime," *Journal of Financial Crime*, vol. 17, no. 4, pp. 441-458, 2010.
[14] M. DeLiema, G. Mottola and M. Deevy. (2017, February). *Findings from a pilot study to measure financial fraud in the United States* [Online]. Available: https://ssrn.com/abstract=2914560
[15] Kroll. (2019). *Global fraud and risk report 2019/20*. [Online]. Available: https://www.kroll.com/en/insights/publications/global-fraud-and-risk-report-2019
[16] A. Nawawi and A. S. A. P. Salin, "Employee fraud and misconduct: Empirical evidence from a telecommunication company," *Information & Computer Security*, vol. 26, no. 1, pp. 129-144, 2018.
[17] M. Jans, J. M. Van Der Werf, N. Lybaert and K. Vanhoof, "A business process mining application for internal transaction fraud mitigation," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13351-13359, 2011.
[18] J. Leskovec, A. Rajaraman and J. D. Ullman, *Mining of Massive Data Sets*. Cambridge, UK: Cambridge University Press, 2019.
[19] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, NJ: John Wiley & Sons, 2011.
[20] C. K. S. Leung, "Big data analysis and mining," in *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*, D.B.A. Mehdi Khosrow-Pour, Ed. Pennsylvania, USA: IGI Global, 2019, pp. 15-27.
[21] R. K. Saini, "Data mining tools and challenges for current market trends-A Review," *International Journal of Scientific Research in Network Security and Communication*, vol. 7, no. 2, pp. 11-14, 2019.
[22] M. J. Kranacher and R. Riley, *Forensic Accounting and Fraud Examination*. Hoboken, NJ: John Wiley & Sons, 2019.
[23] G. L. Gray and R. S. Debreceny, "A Taxonomy to guide Research on the Application of data mining to fraud detection in financial statement audits," *International Journal of Accounting Information Systems*, vol. 15, no. 4, pp. 357-380, 2014.
[24] M. Ahmed, A. N. Mahmood and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems*, vol. 55, no. C., pp. 278-288, 2016.
[25] F. A. Amani and A. M. Fadlalla, "Data mining applications in accounting: A review of the literature and organizing framework," *International Journal of Accounting Information Systems*, vol. 24, pp. 32-58, 2017.
[26] G. S. Temponeras, S. A. N. Alexandropoulos, S. B. Kotsiantis and M. N. Vrahatis, "Financial fraudulent statements detection through a deep dense artificial neural network," in *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Patras, Greece, 2019, pp. 1-5.
[27] A. Gepp, M. K. Linnenluecke, T. J. O'Neill and T. Smith, "Big data techniques in auditing research and practice: Current trends and future opportunities," *Journal of Accounting Literature*, vol. 40, pp. 102-115, 2018.

[28] C. Koh and C. Low, "Going concern prediction using data mining techniques," *Managerial Auditing Journal*, vol. 19, no. 3, pp. 462-476, 2004.
[29] S. Makki, R. Haque, Y. Taher, Z. Assaghir, G. Ditzler, M. S. Hacidm and H. Zeineddine, "Fraud analysis approaches in the age of big data-A review of state of the art," in *Foundations and Applications of Self* Systems (FAS* W), IEEE 2nd International Workshops*, Augsburg, Germany, 2017, pp. 243-250.
[30] The Institute of Internal Auditors – Australia. (2018). *Data mining using Excel* [Online]. Available: http://iia.org.au/sf_docs/default-source/technical-resources/2018-whitepapers/iia-whitepaper_data-mining-using-excel.pdf?sfvrsn=2
[31] Chartered Global Management Accountant. (2013). *Top ten focus priorities fundamental for the data-driven era of business* [Online]. Available: https://www.cgma.org/Resources/Reports/DownloadableDocuments/CGMA-briefing-big-data.pdf
[32] M. Hobday, "The project-based organization: An ideal form for managing complex products and systems," *Research Policy*, vol. 29, no. 7-8, pp. 871-893, 2000.
[33] M. Miterev, M. Mancini and R. Turner, "Towards a design for the project-based organization," *International Journal of Project Management*, vol. 35, no. 3, pp. 479-491, 2017.
[34] D. Bertsimas and J. Dunn, "Optimal classification trees," *Machine Learning*, vol. 106, no. 7, pp. 1039-1082, 2017.
[35] M. Pejić Bach, K. Dumičić and Z. Marušić, "Application of association rules method in tourism product development," in *10th International Symposium on Operational Research in Slovenia*, Bled, Slovenia, 2009, pp. 565-573.
[36] N. R. Mabroukeh and C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms," *ACM Computing Surveys (CSUR)*, vol. 43, no. 1, 3, 2010.
[37] P. Lenca, P. Meyer, B. Vaillant and S. Lallich, "On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid," *European Journal of Operational Research*, vol. 184, no. 2, pp. 610-626, 2008.
[38] T. Brijs, G. Swinnen, K. Vanhoof and G. Wets, "Building an association rules framework to improve product assortment decisions," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 7-23, 2004.
[39] Y. Boztuğ and T. Reutterer, "A combined approach for segment-specific market basket analysis," *European Journal of Operational Research*, vol. 187, no. 1, pp. 294-312, 2008.
[40] Y. Heryadi, L. A. Wulandhari and B. S. Abbas, "Recognizing debit card fraud transaction using CHAID and K-nearest neighbor: Indonesian Bank case," in *11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, Yogyakarta, Indonesia, 2016, pp. 1-5.
[41] W. F. Messier Jr., "Opportunities for task-level research within the audit process," *International Journal of Auditing*, vol. 14, no. 3, pp. 320-328, 2010.
[42] A. Trigo, J. Varajão, J. Barroso, P. Soto-Acosta, F. J. Molina-Castillo and N. Gonzalvez-Gallego, "Enterprise information systems adoption in Iberian large companies: Motivations and trends," in *Managing Adaptability, Intervention, and People in Enterprise Information Systems*, M. Tavana, Ed. Hershey, USA: Information Science Reference, 2011, pp. 204-228.
[43] M. Terzieva and V. Morabito, "Learning from experience: The project team is the key," *Business Systems Research Journal*, vol. 7, no. 1, pp. 1-15, 2016.

Appendix A. Selected SQL equations generated for the implementation of the CHAID decision tree

/* Node 4 */. DO IF (Month NE "M9" AND Month NE "M12" AND Month NE "M8" AND Month NE "M2" AND Month NE "M1") AND (Customer EQ "Govern"). COMPUTE nod_001 = 4. COMPUTE pre_001 = 'Not Susp'. COMPUTE prb_001 = 0.970149. END IF. EXECUTE.

/* Node 5 */. DO IF (Month NE "M9" AND Month NE "M12" AND Month NE "M8" AND Month NE "M2" AND Month NE "M1") AND (Customer EQ "Internal"). COMPUTE nod_001 = 5. COMPUTE pre_001 = 'Not Susp'. COMPUTE prb_001 = 0.710145. END IF. EXECUTE.

/* Node 9 */. DO IF (Month NE "M9" AND Month NE "M12" AND Month NE "M8" AND Month NE "M2" AND Month NE "M1") AND (Customer NE "Govern" AND Customer NE "Internal") AND (ExpertLevel NE "L4" AND ExpertLevel NE "L8" AND ExpertLevel NE "L7" AND ExpertLevel NE "L5"). COMPUTE nod_001 = 9. COMPUTE pre_001 = 'Not Susp'. COMPUTE prb_001 = 0.855159. END IF. EXECUTE.

/* Node 10 */. DO IF (Month NE "M9" AND Month NE "M12" AND Month NE "M8" AND Month NE "M2" AND Month NE "M1") AND (Customer NE "Govern" AND Customer NE "Internal") AND (ExpertLevel EQ "L4" OR ExpertLevel EQ "L8" OR ExpertLevel EQ "L7" OR ExpertLevel EQ "L5"). COMPUTE nod_001 = 10. COMPUTE pre_001 = 'Not Susp'. COMPUTE prb_001 = 0.700000. END IF. EXECUTE.

/* Node 7 */. DO IF (Month EQ "M9" OR Month EQ "M12" OR Month EQ "M8" OR Month EQ "M2") AND (Customer NE "Internal"). COMPUTE nod_001 = 7. COMPUTE pre_001 = 'Not Susp'. COMPUTE prb_001 = 0.771341. END IF. EXECUTE.

/* Node 8 */. DO IF (Month EQ "M9" OR Month EQ "M12" OR Month EQ "M8" OR Month EQ "M2") AND (Customer EQ "Internal"). COMPUTE nod_001 = 8. COMPUTE pre_001 = 'Suspect'. COMPUTE prb_001 = 0.600000. END IF. EXECUTE.

/* Node 3 */. DO IF (Month EQ "M1"). COMPUTE nod_001 = 3. COMPUTE pre_001 = 'Suspect'. COMPUTE prb_001 = 0.657658.

END IF. EXECUTE.

Biographical notes

Mirjana Pejić Bach
Mirjana Pejic-Bach Mirjana Pejic-Bach, Ph.D., is a Full Professor of System Dynamics Modelling, Managerial Simulation Games and Data Mining at the Faculty of Economics and Business, University of Zagreb, Department of Informatics. Her current research areas are system dynamics modeling, data mining and web content research. She is actively engaged in a number of scientific projects (FP7-ICT, bilateral cooperation, national projects).

Ksenija Dumičić
Ksenija Dumicic, Ph.D., is a Full Professor of Statistical Sampling, Business Statistics, Statistical Research Methods and SQC at the University of Zagreb Faculty of Economics and Business, Department of Statistics. Her current research areas are survey sampling and statistical education research. By now, she has been actively engaged in a number of national scientific projects, World Bank, UNHCR, UNICEF, WHO, IPA, and PHARE projects.

.h

Berislav Žmuk
Berislav Žmuk, Ph.D., is an Assistant Professor at the Faculty of Economics and Business, University of Zagreb, Department of Statistics. He was also educated at the University of Michigan in Ann Arbor, at GESIS in Cologne, the University of Essex in Colchester and at the University of Utrecht in Utrecht in the field of survey methodology and applied statistics. His main research interests are business statistics, survey methodology, and statistical quality control.

Tamara Ćurlin
Tamara Ćurlin is a Teaching Assistant and a Ph.D. student at the Faculty of Economics and Business, University of Zagreb, Department of Informatics. She received her BSc and MSc degrees from the Faculty of Economics and Business, University of Zagreb. She is teaching Informatics and Enterprise Information Systems courses exercises. Her current research interests include Information Technologies in Tourism, Mobile Technologies, Knowledge management, and Information management.

Jovana Zoroja
Jovana Zoroja, Ph.D., is an Assistant Professor at the Faculty of Economics and Business, University of Zagreb, Department of Informatics. She was also educated at the LSE in London in the field of Business Development and ICT Innovation. Her main research interests are information and communication technology, simulation games and simulation modeling. She participated in an Erasmus-preparatory-visit-program and is now engaged in an FP7-ICT project as well as bilateral cooperation.