INTERNAL FRAUD RISK REDUCTION
Results of a Data Mining Case Study

Mieke Jans, Nadine Lybaert
KIZOK Research Institute, Hasselt University Campus Diepenbeek, Agoralaan Building D, 3590 Diepenbeek, Belgium

Koen Vanhoof
Research Group Data Analysis and Modeling, Hasselt University Campus Diepenbeek, Diepenbeek, Belgium

Keywords: Internal Fraud, Data Mining, Risk Reduction.

Abstract: Corporate fraud these days represents a huge cost to our economy. Academic literature already concentrated on how data mining techniques can be of value in the fight against fraud. All this research focusses on fraud detection, mostly in a context of external fraud. In this paper we discuss the use of a data mining technique to reduce the risk of internal fraud. Reducing fraud risk comprehends both detection and prevention, and therefore we apply a descriptive data mining technique as opposed to the widely used prediction data mining techniques in the literature. The results of using a latent class clustering algorithm to a case company's procurement data suggest that applying this technique of descriptive data mining is useful in assessing the current risk of internal fraud.

1 INTRODUCTION

Saying that fraud is an important (however not loved) part of business, is nothing new. Fraud is a million dollar business, as several research studies reveal. Among them are an important survey of PriceWaterhouse&Coopers (PwC, 2007) and of the Association of Certified Fraud Examiners (ACFE, 2006). The study conducted in the United States by the ACFE in 2004-2005 and the worldwide study, held by PwC in 2006-2007 yield the following insights. 43% of companies worldwide have fallen victim to economic crime in the years 2006 and 2007. The average financial damage to companies subjected to the PwC survey was US$ 2.42 million per company over two years. Participants of the ACFE study estimate a loss of 5% of a company's annual revenues to fraud. These numbers all address corporate fraud.

Academic literature is currently investigating the use of data mining for the purpose of fraud detection. (Brockett et al., 2002), (Cortes et al., 2002), (Estévez et al., 2006), (Fanning and Cogger, 1998), (Kim and Kwon, 2006) and (Kirkos et al., 2007) are just a few examples of a more elaborated list of articles concerning the hot topic of fighting fraud. Although a lot of this research may be framed in different settings -going from different techniques to different fraud domains-, there are two characteristics that stand for all executed research up till now: the focus is on external fraud and a predictive data mining approach is applied for fraud detection. We however are interested in internal fraud, since this represents mainly these large costs in the PwC and ACFE surveys. Further, we are convinced that not fraud detection alone, but detection in combination with prevention, is of priceless value for organizations. We will use the term fraud risk reduction for encompassing both fraud detection and prevention.

We continue this study on the positive results of academic literature concerning the use of a data mining approach for the purpose of fraud detection. Since our aim is fraud risk reduction, we apply however another category of techniques than applied up till now. We believe in the value of descriptive data mining for the purpose of fraud risk reduction. In contrast to the explored predictive data mining techniques in current academic literature, descriptive data mining provides us with insights of the complete data set and can be of value for assessing the fraud risk in selected business processes.

In the following sections we explain the followed methodology of this study, the data set, the used latent

Jans M., Lybaert N. and Vanhoof K. (2008).
INTERNAL FRAUD RISK REDUCTION - Results of a Data Mining Case Study.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - AIDSS*, pages 161-166
DOI: 10.5220/0001679201610166

class clustering algorithm, and the results of investigating a business process of the case company. We end with a conclusion.

2 METHODOLOGY

The applied methodology can be summarized by Figure 1. As a first step, an organization should select a business process which it thinks is worthwhile investigating. Further is the implementation of advanced IT a breeding ground for employee fraud (Lynch and Gomaa, 2003). For this reason and because data needs to be electronically stored in order to mine it by means of a data mining approach, the selected business process needs to be one with an advanced IT integration. In a second step the stored data will be collected, manipulated and enriched. These are mainly technical transactions. During the third step, the technical data will be translated into behavioral data. This translation builds upon domain knowledge and is not just a technical transformation. The core of the methodology is then to apply descriptive data mining for getting more insights in this behavioral data. The descriptives should provide the researchers a recognizable pattern of procedures of the selected business process. In addition some other patterns of minor groups of observations in the data can arise, interesting to have a closer look at. By auditing observations part of such a subgroup, the domain expert can categorize the observations in four groups. The fraudulent observations (a first category) are part of fraud detection, while the observations that circumvent procedures or are created by mistake (two other categories) are part of fraud prevention. Fraud prevention is in this methodology primarily based on checking or taking away the fraud opportunity. The importance of opportunity is stressed by Cressey's fraud triangle with opportunity being the only element of fraud risk that an employer can influence. The other two elements, rationalization and incentive, are personal characteristics. The fourth category of audited observations, the ones with extreme values, can also occur, but are of no interest for internal fraud risk reduction.

3 DATA SET

The data set is established by performing the first three steps of our methodology. For this study, the corporation of a case company was acquired. This company, which chooses to stay anonymous in this study, is an international financial services provider, ranked in the top 20 of European financial institutions.

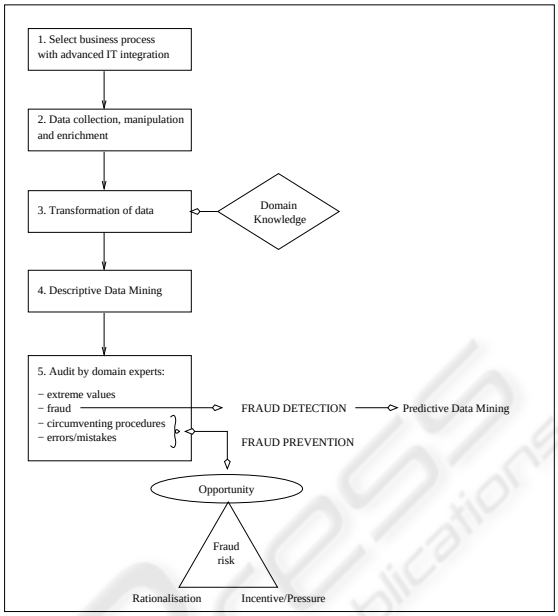

Figure 1: Methodology for internal fraud risk reduction.

The business process selected for internal fraud risk reduction is procurement, so data from the case company's procurement cycle is the input of our study. More specifically, the creation of purchasing orders (PO's) was adopted as process under investigation.

As a start, a txt-dump is made out of SAP. All PO's that in 2006 resulted in an invoice are the subject of our investigation. This raw data is then reorganized into appropriate tables to support meaningful analysis. After the creation of these new formats, additional attributes were created as enrichment and resulted in a data set of 36.595 observations. Based on domain knowledge, supported by descriptive statistics, a preclustering step is made. PO's are split in two groups: old PO's and new PO's. Old PO's are the ones created before July 2005. The fact that they are included in our data is because an invoice of the year 2006 can be linked to a PO created in 2005 or even before 2005. However, if a PO is from before July 2005 (there are PO's even from 2000), this PO shows a different life cycle than if it were younger (there are for example much more changes on such PO's). The subset of old PO's contains 2.781 observations while the subset of new PO's counts 33.814 observations. Both subsets of PO's were subjected to the proposed methodology. Since the latter group is the most prominent in assessing internal fraud risk (most recent) and given its magnitude, this paper gives detailed test results of the new PO's. The other side of the picture is that this large data set poses more problems in the fifth step of our methodology, namely the auditing of interesting observations. We restrict this study to provide recom-

mendations on this matter for the new PO's. For the subset of old PO's however, the audit step is effectively executed and these results will be reported after the discussion of the new PO's. In what follows, the term data set refers to the subset of new PO's (33.814 observations).

The most important attributes to describe a PO and its life cycle are the following: the name of the creator, the supplier, the purchasing group, the type of purchasing document, the number of changes, the number of changes after the last release and the number of price related changes after the last release. Concerning the numerical attributes, there are 91 creators recurring in the data set, 3.708 suppliers, 13 purchasing groups and 6 document types. (see Table 1)

Table 1: Categorical attributes.

Categorical	Recurrence in data set
Creator	91
Supplier	3.708
Purchasing Group	13
Document Type	6

Of the 91 creators, not all of them introduce equally as much PO's in the ERP system, because of the individual characteristics of each purchase. Some creators, responsible for a particular type of purchase, need to enter lots of PO's, while other creators, responsible for other types of purchase, only enter a few PO's. Also the turnover in terms of personnel has its reflection on the number of PO's per employee. Like creators, the frequency of suppliers in the data set is liable to the specific characteristics of the product or service supplied. There will be for example more PO's concerning monthly leasing contracts for cars than there will be for supplying desks. Hence the former supplier will be more frequently present in the data set than the latter. Concerning the 13 purchasing groups, there is no difference in expected fraud risk between the different groups. Some groups are more present than others in the data set, but this can all be explained by domain knowledge. The same goes for the six different purchasing document types. All types have their specific characteristics, but there is no expected difference concerning fraud risk.

The numerical attributes are described in Table 2. For each attribute, three intervals were created, based on their mean and standard deviation. For the first attribute, the intervals were [2-4], [5-8] and [9-...], for the second attribute [0-0], [1-2] and [3-...] and for the last attribute [0-0], [1-1] and [2-...]. In Table 2 we see that there is a highly skewed distribution for the three attributes, which is to be expected for variables that count these types of changes. The changes are supposed to be small in numbers.

After creating these attributes and providing descriptives, we turn to the third step of our methodology. For the specification of our model, we take into account the particular type of fraud risk we wish to reduce. The fraud risk linked with entering PO's into the ERP-system is connected with the number of changes one makes to this PO, and more specifically, the changes made after the last release. There is namely a built-in flexibility in the ERP system to modify released PO's without triggering a new release procedure. For assessing the related risk, we selected four attributes to mine the data. A first attribute is the number of changes a PO is subjected to in total. A second attribute presents the number of changes that is executed on a PO after it was released for the last time. The third attribute we created is the percentage of this last count that is price related. So what percentage of changes made after the last release is related to price issues? This is our third attribute. The last attribute concerns the magnitude of these price changes. Considering the price related changes, we calculate the mean of all price changes per PO and its standard deviation. On itself, no added value was believed to be in it. Every purchaser has its own field of purchases, so cross sectional analysis is not really an option. However, we combine the mean (μ) and standard deviation (σ) to create a theoretical upper limit per PO of $\mu + 2\sigma$. Next, we count for each PO how often this theoretical limit was exceeded. This new attribute is also taken into account in our model. In this core model, no categorical attributes were added. As a robustness check however, attributes like document type and purchasing group were included in the model. The results did not significantly change by these inclusions.

4 LATENT CLASS CLUSTERING ALGORITHM

For a descriptive data mining approach, we have chosen for a clustering algorithm, more specifically a latent class (LC) clustering algorithm. We prefer LC clustering to the more traditional K-means clustering for several reasons. The most important reason is that this algorithm allows for overlapping clusters. An observation is provided a probability to belong to each cluster, for example .80 for cluster 1, .20 for cluster 2 and .00 for cluster 3. This gives us the extra opportunity to look at outliers in the sense that an observation does not belong to any cluster at all. This is for example the case with probabilities like .35, .35 and .30. Other considerations to apply the LC clustering algorithm are the ability to handle attributes of mixed scale

Table 2: Descriptives of numerical attributes.

Attribute	Minimum	Maximum	Mean	Standard deviation	1st interval frequency (%)	2nd interval frequency (%)	3rd interval frequency (%)
Number of changes	1	152	4.37	3.846	71.3	21.5	7.2
Number of changes after last release	0	91	.37	1.343	80.9	11.9	7.2
Price related number of changes after last release	0	46	.15	.882	91.1	6.7	2.2

types and the presence of information criteria statistics to determine the number of clusters. For a more detailed comparison of LC clustering with K-means we refer to (Magidson and Vermunt, 2002). For more and detailed information about LC analysis, we refer to (Kaplan, 2004) and (Hagenaars and McCutcheon, 2002).

5 CASE STUDY RESULTS

5.1 Model Specifications

Using the four attributes described in Section 3, we executed the LC clustering algorithm with the number of clusters (K) set equal to 1 till 6. This yielded the information criteria (IC) values plotted in Figure 2. As you can see, the IC values drop heavily until the 3-cluster model. Beyond the 3-cluster model, the decreases are more modest. Based on these values, we decide to use this 3-cluster model.

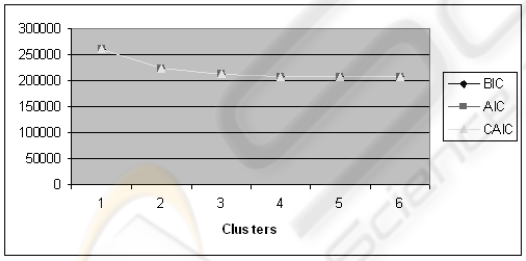

Figure 2: Information criteria values.

5.2 Results

The profile of the 3-cluster model is presented in Table 3. It gives the mean value of each attribute in each cluster. To compare with the data set as a whole, the mean values of the population are also provided.

Looking at the profile of the 3-cluster model, there is an interesting cluster to notice, the third cluster, if it was even only for its size. Cluster 1 comprehends 76.6% of the total data set, cluster 2 22.1% and cluster 3 only 1.3%. Why is there 1.3% of all PO's behaving

Table 3: Profile of data set and 3-cluster model.

	Population	Cluster 1	Cluster 2	Cluster 3
Cluster size	100	0.7663	0.2212	0.0125
Number of changes	4.37	3.3378	6.7608	25.459
Changes after release	0.37	0.0193	1.2376	6.1257
Percentage price related	0.0756	0	0.3185	0.4094
Count over limit	0.01	0.0072	0.0194	0.2725

differently than the remaining PO's? Regarding the mean attribute values of this small cluster, this cluster is, besides from its size, also interesting in terms of fraud risk. The mean number of changes per PO in this cluster, is 25, as opposed to a mean number of changes of 4 in the data set. Why are these PO's modified so often? Not only are these PO's changed so much in their entire life cycle, they are also modified significantly more after they were last released (6 times) in comparison with the mean PO in the data set (0.37 times). These are odd characteristics. The mean percentage in cluster 3 of changes after the last release that is price related is also the highest percentage of the three clusters (40.9%). All together this means that the average PO in cluster 3 is changed 25 times in total, of which 6 changes occur after the last release and 2.4 of those 6 changes are price related. Concerning the magnitude of the price related changes, we can conclude that these changes of PO's in cluster 3 are more often much larger than the average price change in that PO if we compare this with price related changes of PO's in the other clusters. In cluster 3, there are on average 0.2725 price related changes larger than $\mu + 2\sigma$ per PO, in comparison with 0.0072 and 0.00194 per PO in cluster 1 and 2 and 0.01 changes in the entire data set.

Taking these numerical characteristics into account, one can conclude that cluster 3 has a profile with a higher fraud risk than the other two clusters. Also categorical attributes behave in a different fashion than they behave in the data set as a whole. So there are the creators of the PO. One person for example created 39 out of the 408 PO's from cluster 3 (hereby representing 9.56% of cluster 3), while the same person only created 131 out of the 33.814 PO's, which counts only for 0.39% of the entire data set.

For calculating the probability of taking this per-

son (called xxx) by chance 39 times of 408, given the prior distribution, we use the hypergeometric distribution. This looks as follows.

$$h_m = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

The hypergeometric distribution is a discrete probability distribution that describes the number of successes m in a sequence of n draws without replacement, given a finite population N with M successes. In our situation concerning person xxx this leads to:

$$h_{39} = \frac{\binom{131}{39}\binom{33.814-131}{408-39}}{\binom{33.814}{408}} < 1^{-15}$$

So if we select 408 cases at random out of the population of 33.814 observations, there is a probability less than 1^{-15} that we pick 39 cases with user-id xxx, given the prior distribution of 131 successes in the population. This event is very unlikely to happen by coincidence.

Not only creators made such significant increases in representation, but also some suppliers are significantly more represented in cluster 3 than they are in the full data set. We screened all creators and suppliers on significant increases in representation between the data set and cluster 3 with a significance level of $h < 1^{-5}$. 14 suppliers and 12 creators met this criterium. Not all of them are however equally important since an increase of 0.03% representation to 0.98% is not as impressive as an increase of 1.47% to 7.6%. Table 4 gives us more insights into the importance of the 14 suppliers and 12 creators.

Table 4: Descriptives of creators and suppliers with a significant higher representation in cluster 3.

Representation (r) in cluster 3	Number of suppliers	Number of creators
$r < 1\%$	4	
$1\% < r < 2\%$	4	
$2.2\% < r < 4.5\%$	3	
$6\% < r < 7.5\%$	3	
Total	14	
$r < 2\%$		3
$2.9\% < r < 3.5\%$		3
$5\% < r < 10\%$		6
Total		12

Since it is more than likely that auditing all 408 PO's of cluster 3 is too time consuming, it would be interesting to take a sample of PO's that are made by one of the creators described above or involve one of those suppliers (or both). The smallest sample to extract from this cluster is to take only those PO's of the six creators and three suppliers that are most represented in the cluster. This yields a sample of 38 PO's. Why is it that they merely induce PO's in this small cluster than in the other two clusters. What makes these purchases this risky? Also the recurrence of a particular purchasing group and purchasing document type can shed an interesting light on deciding which PO's to audit. Auditing this kind of PO's can learn the company a lot about the opportunities that exist to commit fraud, in view of the fraud risky profile that the numerical attributes describe.

The possibility LC clustering provides to audit observations that do not belong to any cluster is also explored. 42 PO's were identified and audited in-depth. There was no uniform profile for these cases. The audit resulted in a few questions with regard to the use of the ERP-system. Nothing however showed misuse of procedures or any other fraud risk.

5.3 Audit by Domain Experts of Old PO's

As already mentioned, the audit step is not (yet) executed for the subset of new PO's. The entire methodology, provided in Figure 1, is however also applied on the subset of old PO's, including the fifth step. The results of the descriptive data mining step are similar to the discussed results. The small interesting cluster (in perspective of a fraud risky profile) of old PO's only contained 10 observations, with nine of them stemming from the same purchasing group and six of them created by the same employee. These 10 observations were audited by domain experts. The results of their investigation are summarized in Table 5.

Table 5: Summary of investigation by domain experts.

Category	Number of cases
Extreme values	0
Fraud	0
Circumventing procedures	9
Errors/Mistakes	1

These are very good results in the light of internal fraud risk reduction. Nine PO's, the ones in the particular purchasing group, are created and modified all over and over again. This is against procedures and makes investigating these PO's very difficult. By creating such complex histories of a PO, the opportunity of committing fraud increases. Only insiders can unravel what really happened with these PO's, since they are such a mess. This off course increases the opportunity and risk of internal fraud. Also, the investigation of this practice has put things in another perspective concerning the separation of functionalities. A follow-up investigation by the audit and investigations department of the case company for this matter is approved.

In the tenth PO a mistake is made. As explained before, a mistake that stays unnoticed creates a window of opportunity for internal fraud. The employee that first makes a mistake by accident, can afterwards consider how to turn this opportunity to one's advantage.

By investigating the 10 selected observations, additional odd practices came to light, which also induced extra investigations. On top of this, the case company gave priority on auditing the procurement cycle in depth.

5.4 Comparing Results with Reporting Results

The results of using a descriptive data mining technique, provides us with interesting results. In the smaller subset we encounter PO's that are changed over and over again. Also in the larger subset, changing the PO a lot of times is a primal characteristic of the selected observations. However, one could wonder if this outcome was not much easier to obtain, simply by applying some form of reporting. For example, maybe we get the same results when we just take the observations with lots of changes? We do not go into the discussion about one method being generally better than another. What we can and want to say however, is that in our case, we did not find the same results by using basis reporting as we found by applying the LC clustering algorithm. The multivariate analysis was indispensable to come to the presented results. Another remark to consider about reporting, is that you do not know in advance which attribute to go on. Also analyzing the influence of categorical variables is not easy with reporting tools.

6 CONCLUSIONS

In this paper, a methodology for reducing internal fraud risk is presented. This is a contribution to the literature in that it concerns internal fraud whereas the literature focusses on external fraud. Further we broaden our scope from fraud detection to fraud risk reduction, which encompasses both fraud detection as prevention. We were able to apply our suggested methodology in a top 20 ranked European financial institution. The results of the case study suggest that the use of a descriptive data mining approach and the latent class clustering technique, can be of additional value to reduce the risk of internal fraud in a company. Using simple reporting tools did not yield the same results, nor would be able to provide similar insights in current use of procedures. The application of the suggested methodology at the case company produced a tone of more concern about the topic of internal fraud along with concern about the opportunity of committing this crime.

REFERENCES

ACFE (2006). 2006 ACFE Report to the nation on occupational fraud and abuse. Technical report, Association of Certified Fraud Examiners.

Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A., and Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. *The Journal of Risk and Insurance*, 69(3):341–371.

Cortes, C., Pregibon, D., and Volinsky, C. (2002). Communities of interest. *Intelligent Data Analysis*, 6:211–219.

Estévez, P., Held, C., and Perez, C. (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications*, 31:337–344.

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

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press.

Kaplan, D. (2004). *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks: Sage Publications.

Kim, H. and Kwon, W. J. (2006). A multi-line insurance fraud recognition system: a government-led approach in Korea. *Risk Management and Insurance Review*, 9(2):131–147.

Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32:995–1003.

Lynch, A. and Gomaa, M. (2003). Understanding the potential impact of information technology on the susceptibility of organizations to fraudulent employee behaviour. *International Journal of Accounting Information Systems*, 4:295–308.

Magidson, J. and Vermunt, J. K. (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research*.

PwC (2007). Economic crime: people, culture and controls. the 4th biennial global economic crime survey. Technical report, PriceWaterhouse&Coopers.