Aalto University
School of Business

PREDICTING THE TYPE OF FINANCIAL STATEMENT FRAUD

Master's Thesis
Asko Jokinen
Aalto University School of Business
Department of Accounting

Fall 2019

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Asko Jokinen		
Title of thesis Predicting the type of financial statement fraud		
Degree Master of Business		
Degree programme Accounting		
Thesis advisor Emma-Riikka Myllymäki		
Year of approval 2019	**Number of pages** 100	**Language** English

Abstract

Fraud is a problem for the all kinds companies, both large and small. According to a study be Association of Certified Fraud Examiners could be even 5% of the whole world Gross Domestic Product leading to approximately $4 trillion losses. The financial statement fraud is the costliest form of fraud, when it occurs with a median loss of $800.000 per case. However, in 22% of the cases of financial statement fraud the loss is over $1.000.000. The problem is that the main way of finding fraud has been whistleblowing. There is a clear need of other effective methods to finding fraud. In case of financial statement fraud one can attempt to use artificial intelligence methods to predict whether a financial statement is fraudulent or not. Usually this has been studied using models, which only whether the financial statement is fraudulent or not. Here also the type of fraud is studied, so that one could start to use the information for predicting in which part of the financial statement the fraud is in.

We use dataset combined from Audit Analytics and Compustat datasets from Wharton Research Data Services. The data is for years 1995-2016 and consists of prediction variables, which are formed using financial statement data and other public data for the companies. Altogether there are 347 fraudulent financial statements and 58.892 non-fraudulent financial statements in the final dataset. 9 different predictive models are formed using regularized logistic regression and 35 predictive variables. 1 predictive model is for fraud as a whole, 8 are for different fraud types. Finally a predictive model of fraud is built using 3 different fraud types and compared whether it produces better results than modelling fraud directly. Of the 35 predictive variables 7 turn out to appear in at least 8 of the 9 different models: whether new securities were issued, value of issued securities to market value, accounts receivable, accounts receivable to total assets, is the auditor one of Big 4, net sales and whether standard industry classification code is between 3000-3999 or not.

The performance of the models to predict fraud or fraud type is measured using expected relative cost of misclassification, accuracy, precision, sensitivity, receiving operating curves and areas under the receiving operating curves. Receiving operating curves for fraud and fraud types are quite similar, so are their areas under the operating curves, which is 0,71 for fraud and 0,68 for the combination of 3 fraud types. The rest of the results depend on the prior fraud probability in the world, which is taken to be between 0,1% - 10%, and the ratio of cost of misclassifying fraud as non-fraud to cost of misclassifying non-fraud as fraud, which varies between 1:1 and 100:1. The accuracy, which measures the percentage of correct classifications among all cases, is between 80% - 99% for the combination of three types and 81% - 99% for fraud. The precision, which measures the percentage of correct fraud classifications among all predicted fraud cases, varies between 1,3% - 3,5% for fraud and 1,4% - 4,2% for the combination of three types, these numbers are low because of the huge imbalance between fraudulent and non-fraudulent cases. The sensitivity, which measures the percentage of correct fraud classifications among all the actual fraud cases, varies between 1,4% - 42% for fraud and between 1,7% - 48% for the combination of three types. The expected relative cost of misclassification for the combination of three types by -3,7% - +0,05% compared to fraud depending on prior fraud probability and relative costs of misclassification. The combination of three types perform better in predicting fraud than direct fraud prediction in most cases prior fraud probability and relative cost of misclassification.

Keywords fraud, financial statement, logistic regression, classification cost, accuracy, precision

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and background

Financial statements form the basis of valuation of business entities Palepu and Healy (2007). Especially the stock market uses them and even demands quarterly updates in addition to the usual annual financial statements. The stock price can change quite a lot when the financial figures are published for the business entity if it shows either better than expected result or worse than expected result. Therefore it is of utmost importance that the financial statements contain accurate information. Whenever mistakes in the financial statements are found and published, the stock price tends to do a correction. There are also other stakeholders who want accurate financial information like banks considering a loan for the entity want typically to see the financial statement of the entity and make the loan decision based on that information. If the financial statement contains wrong information, the decision is made on false premises and may lead into defaulting the loan.

One of the most costly mistakes of the financial statement is fraud. Fraud can exist in three forms according to the report by the Association of Certified Fraud Examiners (ACFE), see ACFE (2018). These are asset misappropriation, corruption and financial statement fraud which occurred in 89%, 38% and 10% of the cases ACFE observed in their study.[1] Median losses per case for the three categories of fraud are $114.000, $250.000 and $800.000 respectively. Therefore the most occurring type is least costly and the least occurring financial statement fraud is the most costly of the types of fraud. ACFE found in their study that the total losses due to fraud were at least $7.1 billion dollars in 2017.[2] However, they asked estimates of the amount fraud what the financial professionals think and came up with 5% of the total. Comparing this to the world GDP in 2017 $79.6 trillion, the 5% figure results into approximately $4 trillion, almost a thousand times the number ACFE identified in their study. So fraud is a serious problem and financial statement fraud, although the least common is the most costly to an entity.

[1] The percentages do not add up to 100% but are more than that because more than 1 type of fraud can co-exist in a case.

[2] The number is produced by changing the bottom 1% to the 1% value and the top 1% to the 99% value since the high numbers might have identified the entities in question.

The problem of the fraud is that it is hard to find. According to the ACFE report fraud is initially found through tips in 40% of the cases, 15% by internal audits, 13% by management review, 7% by accident, external audits in 4% of the cases and IT controls in 1% of the cases. There are also other means by which fraud is found but these are more rarely. Although IT controls find fraud only in 1% of the cases, they find it fastest, typically in 5 months, whereas internal audits, management review and tips take 12, 14 and 18 months respectively. Most typically the one reporting fraud is an employee in 53% of the cases, customer 21%, anonymous 14% and other categories each under 10%.[3] The costs of fraud are distributed such that in 55% of the cases the amount $200.000 or less, $200.000 - $1.000.000 in 23% of the cases and cost of fraud over $1.000.000 happens in 22% of the cases.

In this study we concentrate on financial statement fraud and how to find that. As mentioned earlier it has the least amount of occurrence, but it is the most costly type of fraud. Since fraud is hard to find there have been attempts to find financial statement fraud using statistical and artificial intelligence (AI) methods, most notably the study by Perols (2011). Perols and others have attempted to find financial statement fraud by using earlier data on financial statements and observed frauds to make a prediction whether the latest financial statement is fraudulent or not. Similar attempts have been made to predict misstatements in financial statements, see for example Dutta et al. (2017).

Practically all the articles observed have asked the question whether a financial statement is fraudulent or not, or in case of misstatements whether there is a misstatement or not in the financial statement. If you are someone who is actually trying find fraud, you would be glad if you get some kind of prediction whether a financial statement is fraudulent or not, but that does not get you very far. It does not tell anything about where the fraud is except that it is somewhere in the process of making the financial statement. If I were the one looking for it, I would like to know a bit more about where to look for it: revenue, purchases, receivables, payables etc. Something to narrow down the search. On the other hand an outside stakeholder might not be so interested in knowing what part of the financial statement is fraudulent, but rather what is its financial impact on the financial statement, how much is the profit affected

[3] Note that these numbers contain all types of fraud, not just financial statement fraud.

by the fraud. Also someone looking for the fraud might want to know the financial impact in order to determine how much effort and resources to use for finding fraud.

1.2 Research question and purpose of the study

Being able to know whether a financial statement is fraudulent or not is definitely good to know. However, it is not enough. Knowing that financial statement is fraudulent does not tell, in which part of the financial statement the fraud is. Therefore this study attempts to go further to predict also the type of financial statement fraud, which hopefully gives information on which part of the financial statement the fraud is. In the end one would like to know in what part of the accounting the fraud is, but research has to be restricted because of the data available. Usually one does not have available anything other than the public financial statement, so one has to be satisfied in finding out which part of the financial statement the fraud is. Anyone having the whole accounting information accessible might be able to go further.

Wharton Research Data Services (WRDS) contains financial statement information in COMPUSTAT database and restatement information in the Audit Analytics database. The restatements contain the reason for making the restatement. In most cases the reason is a regular misstatement, but in a few cases the reason is fraud. There is also information on what type of fraud was the reason for the restatement, for example one category is Revenue Recognition Issues. The type of fraud is used as a proxy for where to find the fraud. One might also think about using Audit Analytics data to predict the financial impact on the financial statement, but this is not possible. WRDS just does not have this data, but it is possible to get from commercial side of Audit Analytics. For this reason the financial impact prediction has to be left for possible future research. Personally I would find the financial impact prediction to be even more interesting and useful for wider audience than predicting the type of fraud.

Perols mentioned in his article that data on fraud is quite noisy and the same signal could indicate fraud and non-fraudulent activities, see Perols (2011). It is worth looking into whether prediction of fraud type produces better results or not. This type of thing has actually been

done by Perols et al. (2017) with the same dataset that was used by Perols (2011). If even one fraud type is observed, then the financial statement is fraudulent. However, since the fraud type prediction is not perfect, rather than making a fraud prediction based on just one fraud type prediction, one might get a better result by implementing a voting system. If majority of fraud type predictions predict the corresponding fraud type in the financial statement, then one predicts the financial statement to be fraudulent, otherwise it is predicted non-fraudulent. This might help in the fraud detection since different fraud types might be sensitive to fewer signals than aggregating all types of fraud into a single fraud category. This is also the basis for division into 4 different types in Perols et al. (2017). Their division was based on the fraud belonging to a particular side of the balance sheet or being revenue or cost. Here the fraud types are based on the categories given by the Audit Analytics dataset.

Most of the previous work, except Perols et al. (2017), has gone like

Variables -> Predict fraud or non-fraud

Instead we modify the procedure to two phases

For all fraud types make a fraud type prediction:

Variables -> Predict is of fraud type or not of fraud type

Results of fraud type predictions -> Predict fraud if majority of fraud type predictions predict that fraud type, otherwise non-fraud

After the introduction we suggest the following research questions

RQ1: Can the type of financial statement fraud be predicted?

RQ2: If yes, can the financial statement fraud type prediction be used to find fraud more effectively than using fraud as a single category?

Research question 1 deals with the first phase of trying to predict fraud through fraud types. Research question 2 handles the second phase of predicting fraud once the predictions of fraud types are made and comparing it to the direct method of predicting fraud from the variables directly.

1.3 Research method

The method used is quantitative. The data is gathered from WRDS using Compustat and Audit Analytics datasets. The two datasets are then combined on the company and financial year levels to find which financial statements are fraudulent and which types of fraud they contain. The datasets are then partitioned into 5 folds of training and test sets. The training data is fitted using logistic regression and then predictions are made on the corresponding test that was held out of fitting. All the presented results are based on test set. However, due to how cross validation works the union of test sets over the 5 folds results into the original dataset with predicted probabilities for each fraud type. More on this in the section 3 on data and methods.

1.4 Structure of the study

The thesis is structured as follows: in section 2 literature is reviewed with subsections on the fraud and misstatement literature, and based on the review the variables which are used to detect fraud are defined and the reasons for using them are presented, in section 3 the method of study and the performance measures of the model are reviewed, and how the dataset is formed, in section 4 the results presented, in section 5 the results are discussed and in section 6 conclusions and future research propositions are presented.

2 Literature review and theory

Financial statement fraud is part of the larger financial misstatements category. Both have their own literature, but there is overlap in the methods used. Therefore both categories are reviewed. The misstatements are reviewed only when relevant to the study here. Other methods, besides the ones used, are also reviewed in the last subsection.

2.1 Literature review

2.1.1 Financial statement fraud literature

The most relevant article for the study here is Perols (2011). Perols points out in his study many of particular features of financial statement fraud: 1) the ratio of fraud to non-fraud firms is small i.e. there are much more non-fraudulent financial statements than there are fraudulent financial statements, 2) The ratio of false positive to false negative misclassification costs is small meaning making a mistake of classifying non-fraud as fraud is much less expensive than making a mistake of classifying fraud as non-fraud, 3) the attributes used to detect fraud are noisy, same values may signal both fraudulent and non-fraudulent activities and 4) persons committing fraud try to conceal their actions making financial statements look non-fraudulent. His main point of study was to compare different machine learning algorithms while taking into account the distinctive features of the problem. He combined data from Compustat, Compact D/SEC and I/B/E/. He had in his final sample 51 fraud firms and 15934 non-fraud firms, so that the prior fraud probability was 0.3% (51 / 15934). He found that instead of the much more complicated machine learning models, the simpler models, logistic regression and support vector machine, performed best in low prior fraud probability environment. This was in contrast to other studies, where the ratio of fraudulent to non-fraudulent financial statements was much closer to 1 and for example they found that neural networks perform better than simple models. The performance measure used was expected relative costs of making false predictions.

Green and Choi (1997) studied the problem using neural networks with a balanced sample namely between 86 and 95 fraudulent financial statements and the same amount of non-fraudulent financial statements. They obtained an aggregate error rate of 25%. When one compares to making a coin toss, where the error rate is 50%, this is definitely lower. However, other types of measures taking into account the different costs were not used. Besides accuracy or error rate (error rate = 1 – accuracy) as a whole may not be a good performance measure in reality. In the balanced sample with the same amount of fraudulent and non-fraudulent financial statements this works, but with highly imbalanced samples, which the fraud in reality is, the accuracy can be a bad measure. For example if there are 1% of fraud in reality, one can get a 99% accurate classifier by classifying every case as non-fraud. This kind of classifier would find no fraud cases whatsoever, so as such it is clearly a bad classifier for the purpose.

Lin et al. (2003) used a fuzzy neural network with 40 fraudulent and 160 non-fraudulent financial statements and compared it with the logit model. The results for logit were overall accuracy 79%, actual fraud over total predicted fraud 5% and actual non-fraud over total predicted non-fraud 97,5%. On the other hand the fuzzy neural network had overall accuracy of 76%, actual fraud over total predicted fraud 35% and actual non-fraud over total predicted non-fraud 86,3%. So fuzzy neural network found fraud better than logit model even though the overall accuracy of fuzzy neural network was lower. Although the sample was not really of the realistic type, 20% fraud and 80% non-fraud, they analysed the overall error rate using realistic prior probabilities of fraud, namely they estimated prior probability to be 1%. They also calculated the expected costs with relative costs of misclassifying errors predicting fraud as non-fraud over predicting non-fraud as fraud from 1:1 to 100:1. They found that fuzzy neural network performed better than logit model, when the relative cost exceeded 40:1. Below that the logit model performed better.

Perols and Lougee (2011) study the relationship between earnings management and financial statement fraud.[4] They found that the firms, which commit fraud, are more likely to have committed earnings management in years prior to committing fraud. They also found an association between earnings management in the prior years and higher likelihood of firms

[4] Financial statement fraud is referred as here as fraud.

meeting or beating analyst forecasts or inflating their revenues are committing fraud, too. In addition fraud firms are more likely to meet or beat analyst forecasts and inflate revenue than non-fraud firms even when there is no evidence of prior earnings management. The reason why prior earnings management can lead to later fraud is that initially earnings are managed by manipulating accruals. Now that the accruals are in the balance sheet they have to be dealt with later on either by reversing them and dealing with the consequences of it or by committing fraud to hide them. As Perols and Lougee point out the purpose of earnings management and financial statement fraud is very similar. They cite Healy and Wahlen (1999) in the definition of earnings management: "earnings management occurs when managers use judgement 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 rely on reported accounting numbers". Perols and Lougee (2011) define financial statement fraud as "financial statement fraud occurs when managers use accounting practises that do not conform to generally accepted accounting principles (GAAP) to alter financial reports to either mislead some stakeholders about the underlying economic performance of the company or to influence contractual outcomes that rely on reported accounting numbers". As seen in the two definitions the purpose of earnings management and financial statement fraud is the same, but earnings management happens within GAAP and is legal, whereas fraud happens outside GAAP and is illegal. Based on this the variables that are used to detect earnings management can also be used to detect fraud.

Dechow et al. (1996) study the causes and consequences of earnings manipulation. This is not a study of fraud in essence. However, as mentioned above by Perols and Lougee, earnings management and fraud are associated with each other. Their main findings are that important motivation for earnings management is the desire to attract external financing at low cost. And firms engaging in earnings management are more likely to have boards of directors dominated by management, i.e. to have a Chief Executive Officer that is at the same time the Chairman of the Board and also the firm's founder. They are less likely to have an audit committee and an outside blockholder. When the earnings management becomes public knowledge, the firms engaging in it, are more likely to have their costs of capital increased significantly. The main value of this study for fraud detection is that the variables they develop for finding earnings

management can also be used in fraud detection and a few of them were included in Perols (2011).

Beneish (1997) also studies earnings management. However, the sample actually consists of firms that either SEC has charged with violating GAAP or which have publicly admitted to violating GAAP. So if one uses the definition in Perols and Lougee (2011), this study should rather be classified as study of financial statement fraud. There are two primary results: the model used provides means to assess the likelihood or earnings management among firms with large discretionary accruals, and adding lagged total accruals and a measure of past price performance as explanatory variables can help in isolating the discretion among firms with extreme performance. Besides suggesting variables that can also be used for fraud detection, Beneish also used the expected misclassification costs to assess the model performance, as was done originally in Dopuch et al. (1987) and later in Perols (2011).

Fanning and Cogger (1998) used an artificial neural network (ANN) with 20 variables to predict fraud. The corresponding logistic regression model was not successful in contrast to Perols (2011). However, Fanning and Cogger had a sample where there were 102 fraudulent financial statements and 102 non-fraudulent financial statements. So the sample is balanced instead of what Perols pointed out, fraud is rare and the sample should be highly imbalanced to reflect that. Using a balanced sample to estimate the performance of the models may lead to results that are not correct. The model can be fitted on the balanced sample but performance evaluation has to be done on the imbalanced sample. Nevertheless the 20 predictor variables contain good candidates for the fraud detection, many of which were used in Perols' study.

Feroz et al. (2000) made another study with artificial neural networks and logistic regression. They also found that the ANNs perform better than logistic regression. Their sample contained 42 fraudulent financial statements and 90 non-fraudulent financial statements. Their results are in accordance with Fanning and Cogger (1998), but would seem to be in contrast to Perols (2011). However, like Fanning and Cogger the sample is almost balanced and far from an actual situation. In order to remedy this they actually tested with imbalanced samples, too, by changing proportion of fraud and non-fraud samples from 10 % and 90% division with 10% steps to a balanced 50% and 50% sample. They used only 7 predictor variables for fraud. They

provided results both for classification accuracy and expected relative costs. Surprisingly classification accuracy is better with an ANN, but expected relative cost tends to be lower for logistic regression, when the cost of classifying fraud as non-fraud is at most 40 times as large as the cost of classifying non-fraud as fraud. However, above this the ANN is less costly. Therefore the expected relative cost is actually in accordance with the results of Perols.

Kaminski et al. (2004) used discriminant analysis with 79 fraud firms and 79 non-fraud firms, which were of similar size and industry type to the fraud firms. They had 21 predictor variables, of which they found 16 to be significant based on discriminant analysis. Again the usage of balanced sample for estimating performance as a problem.

Lee et al. (1999) use logistic regression model with earnings minus operating cash flow as the predictor variable with control variables to test its usefulness for predicting financial statement fraud. They find that results with this variable included are much better than without it. Their sample consisted of financial statements covering years 1978 – 1991 with 56 fraud cases and 60453 non-fraud cases. Originally they had 21 predictor variables, of which they chose 13 into the final model mainly based on not having missing data. Here is one of a few studies with realistic sample sizes and imbalance between fraud and non-fraud.

Kanapickiene and Grundiene (2015) is a Lithuanian study with 40 fraudulent and 125 non-fraudulent financial statements. They use financial ratios as predictor variables and logistic regression. They study 51 different variables and choose 32 in the end. They report 84,8% classification accuracy with their model. Question is of course whether this is at a realistic level since the sample is quite small and sample is close to balanced. Unless fraud in Lithuania is much more widespread than generally believed by ACFE (2018), the sample should contain more non-fraudulent financial statements. Second point is that the study does not take into account the different costs related to mistaking fraud as non-fraud compared to mistaking non-fraud as fraud.

Perols et al. (2017) continues the work done in Perols (2011). They study advanced subsampling methods, multi-subset observation and variable undersampling, in order to deal with the rareness of financial statement fraud. They also do a variation of variable undersampling, where variables are divided into smaller groups according to the fraud type in

question. This article seems to be the first to have studied fraud types. They use the same sample, which Perols (2011) originally used in his study, but the number of predictor variables is increased from 42 to 109. The result is that the expected relative costs with observation undersampling are reduced by 10,8 percent relative to the best benchmark in Perols (2011), and with variable undersampling including fraud types by 9,6 percent relative to the best performing variable undersampling benchmark. Combining observation and variable undersampling with fraud types improves performance further under certain conditions. One difference in procedure compared to what was used in Perols (2011) is that in the previous study Perols made the variable selection on the whole dataset, and after that divided the dataset into training and test sets. If one wants the test set results to be generalizable to finding new cases of fraud, the test set should be held out, so that it is not part of the variable selection and fitting process. The variable selection and fitting process should be done in the training set and use its results on the test set. In Perols et al. (2017) the procedure is to first divide into training and test sets, and then do the variables selections and model fitting on the training set, and finally measure the performance on the test set, so they do it in a way, where the results are generalizable.

2.1.2 Financial misstatement literature relevant for fraud detection

Dechow et al. (2011) use as a source the SEC's AAERs from the years 1982 – 2005. The study is on misstatements. The final sample consists of 676 firms with at least one annual or quarterly misstated financial statement. They develop a scaled probability, F-score, which can be used as a red flag for a misstatement. They found that all measures of accrual quality are unusually high in misstating years compared to the population of non-misstated. They also found that the percentage of soft assets is high giving more flexibility to change and adjust assumptions to influence short-term earnings. Accrual reversals are also an important signature of a misstatement. The variables used here were included in the study Perols et al. (2017).

Dutta et al. (2017) study the financial misstatements using the same kinds of methods as Perols (2011) for financial statement fraud. Among the variables they use 34 of the 42 predictor variables used by Perols. The total number of predictor variables is 115, which they reduce to 15 using stepwise forward selection to remove less significant or redundant variables. Among the final 15 variables are 5 variables used by Perols for fraud detection, 5 variables used in Dechow et al. (2011) and the remaining 5 are traditional financial or accounting variables. They used 5 different machine learning methods, logistic regression is not among the methods they used. They give the performance of the models using the usual performance measures for binary classification: accuracy, precision, recall or sensitivity, false positive rate, specificity, F-measure and area under the curve. The financial statements they use are obtained from Compustat and information about misstatements from Audit Analytics restatements. The same source has been used in this study and the procedure to clean the data in this study follows the procedure laid out in Dutta et al. (2017) and modifies it where relevant, since this study concentrates on fraud detection. The data sample by Dutta et al. consisted of financial statements between 2001 – 2014. They also made a study of performance by splitting the data to 2001-2007 and 2008-2014 and found no significant difference between the performances over the whole sample and the two subsamples.

2.2 Theory

As pointed out by Perols (2011) fraud is being actively concealed by the perpetrators of it. Fraud is also rare. These two factors make it difficult to find. In order to find it using analytical/statistical methods a lot of variables have been suggested in the literature for predicting fraudulent financial statements. In Perols et al. (2017) there were 109 predictor variables to start with. According to Green and Choi (1997) an auditor needs to use professional judgement to choose variables that predict fraud. In most of the literature the variables chosen have been previously found to be significant. The starting point in this study are the predictor variables used by Perols (2011), which contained 42 variables. Here only 35 of them are used, because the data for the other 7 variables was not available in Compustat. Why so many variables. Well as Perols pointed out a lot of other things, legitimate things, can

make the financial statement look similar to a fraudulent financial statement. With enough many variables one might be able to get fraud pop out somewhere as a combination of its effect on different variables. Perols gives the definitions of the variables and citations of publications where they were originally, but no theoretical reasoning for them. Next some theoretical reasoning is given for using the variables. In particular we try to give reasons for the signs of the coefficients.

In principle the signs of all of the coefficients can be justified to be undetermined, because according to Perols (2011) fraud is being concealed, and according to Perols and Lougee (2011) fraud does not necessarily follow GAAPs. If one does not follow GAAP, one can put any transaction almost anywhere in accounting, so that the fraudulent part can appear anywhere in the financial statement, at least in principle. Therefore any sign, which one could think of based on any theory, could be changed to opposite, because fraud does not have to follow GAAP. In practise the violation of GAAP cannot be too obvious, like moving all the fraudulent transactions to another part of the financial statement in one big chunk. It is likely that auditors would find this. But moving everything piece by piece as part of something legitimate might well go undetected and the effect on financial statement is the same as moving it in one big chunk. As conclusion predicting signs based on theories, when dealing with fraud, may be futile.

In the next subsection the possibility of changing signs due to fraud is not considered. The signs are determined based on any other theory if possible. If the sign turns out be different from the predicted one, the reason might be concealment of fraud or that the theory does not hold up. Fraud types might also have a different sign than fraud itself. This due to the competing

2.2.1 Predictor variables

Accounts receivable. This was used by Green and Choi (1997) and Lin et al. (2003). Green and Choi mentioned that this is relevant for risk assessment of revenue and collection cycle. Lin et al. reasoned that this is one of the most often used account trends in practise. Accounts

receivable has also been mentioned in auditing standards, ISA 805 attachment 1, IAASB (2018), concerning issues requiring special attention in financial statements. The risk is increased, when the accounts receivable increases, so the probability of fraud should increase at the same time. Therefore the predicted sign of the coefficient is (+).

Accounts receivable to sales. This is defined as the ratio of accounts receivable and sales. The variable was used by Green and Choi (1997), Feroz et al. (2000), Lin et al. (2003) and Kaminski et al. (2004). Green and Choi actually had net sales to accounts receivable as their variable, their reasoning being the risk assessment of revenue and collection cycle. Feroz et al. use this for auditing red flag, more specifically difficult to audit transactions class. Lin et al. reasoned that this is one of most commonly used financial ratios in audit. It was also reported to be useful for financial statement misstatements. Kaminski et al. cite empirical evidence of prior studies, but they do not have theoretical reasoning. In ISA 240 attachment 3, IAASB (2018), the unusual changes in the financial ratios can indicate misconduct, especially changes of receivables compared to net sales. Both large accounts receivable and large sales are a source of risk. Therefore the ratio could go up or down and increase the probability of fraud, so the predicted sign of the coefficient is undetermined.

Accounts receivable to total assets. This is the ratio of accounts receivable and total assets. This was used by Green and Choi (1997) and Lin et al. (2003). Green and Choi mentioned that this is relevant for risk assessment of revenue and collection cycle. Lin et al. reasoned that this is a comparative ratio, which is often used in the examination of accounts receivable. Accounts receivable is a riskier asset than some other assets, so if there are a high percentage of them of the total assets, the financial statement contains more risk. On the other hand total assets contain also other sources of risk like inventory, so again the ratio can go up or down and increase the probability of fraud.

Allowance of doubtful accounts (AFDA). This is a reduction of accounts receivable, which represents the amount of receivables that managers believe will not be paid. This was used by Green and Choi (1997) and Lin et al. (2003). Green and Choi mentioned that this is relevant for risk assessment of revenue and collection cycle. Lin et al. reasoned that this is one of the

common trends used in the audit of the sales and receivables cycle. Since the effect of this is opposite to the accounts receivable, the predicted sign of the coefficient is (-).

AFDA to accounts receivable. This is the ratio of AFDA and accounts receivable. This was used by Green and Choi (1997) and Lin et al. (2003). Green and Choi mentioned that this is relevant for risk assessment of revenue and collection cycle. Lin et al. reasoned that this measures the relationships between contra-account and applicable aggregate accounts. The bigger AFDA is compared to accounts receivable, the more riskiness of accounts receivable is reduced. Therefore one expects the probability of fraud to decrease with increasing AFDA to accounts receivable ratio. So the predicted sign of the coefficient is (-).

AFDA to net sales. This is the ratio of AFDA and sales. This was used by Green and Choi (1997) and Lin et al. (2003). Green and Choi mentioned that this is relevant for risk assessment of revenue and collection cycle. Lin et al. reasoned that this measures the relationships between contra-account and applicable aggregate accounts. Risk is increased if net sales are increased. On the other hand risk is reduced with increasing AFDA, because it decreases the effect of accounts receivable. Therefore the behavior of ratio is undetermined. The predicted sign of the coefficient is not determined

Altman Z-score. This is defined as follows by Perols (2011)

$$\begin{aligned}\text{Altman Z-score} &= [3{,}3 \\ &\cdot (\text{Income before extraordinary items} \\ &+ \text{Total interest and related expenses} + \text{Total income taxes}) \\ &+ 0{,}999 \cdot \text{Net sales} + 1{,}2 \cdot \text{Working capital} + 1{,}4 \\ &\cdot \text{Retained earnings}] \,/\, \text{Total assets} + 0{,}6 \\ &\cdot \text{Common shares outstanding} \\ &\cdot \text{Annual close price} / \text{Total liabilities}\end{aligned} \tag{1}$$

This was used by Feroz et al. (2000), and Fanning and Cogger (1998). They used this as financial red flag for auditing, more specifically as an indicator of going concern or financial distress. Small values typically indicate that the company is headed for bankruptcy. Since

increase of bankruptcy risk may be connected to fraud, the probability of fraud increases with decreasing Altman Z-score. So the predicted sign of the coefficient is (-).

Big 4 auditor. This variable is defined as "is the auditor one of the big 4 auditing firms". This actually consists of 8 different auditing firms: Arthur Andersen, Arthur Young, Coopers & Lybrand, Ernst & Young, Deloitte & Touche, KPMG Peat Marwick, Pricewaterhousecoopers and Touche Ross. Originally the firms were the Big 8 and then in time they have dropped to Big 4 for various reasons. All of original big auditing firms are included in this variable. This was used by Fanning and Cogger (1998). They called it as big 6 auditor, because at the time there were still 6 of the 8 big auditing firms around. This is relevant variable, because the larger auditing firms have invested more reputational capital than smaller ones, so they have greater incentives to reduce errors. Moreover the possibility of losing an audit is not so big of an issue for the revenue of a larger firm. They may also be able to provide higher quality, because they have more resources and experience with different industries. The models used here are based on data, where fraud has been found. Since Big 4 auditors are more likely to find fraud, the predicted sign of the coefficient is (+).

Current minus prior year inventory to sales. This is simply the difference of the ratios of inventory to sales in the current year and previous year

$$\text{Current minus prior year inventory to sales} = \frac{\text{Inventory (current)}}{\text{Net Sales (current)}} - \frac{\text{Inventory (previous)}}{\text{Net Sales (Previous)}} \quad (2)$$

This was used by Summers and Sweeney (1998). They base it on auditing standards. Any account whose value requires subjective judgement increases audit risk. Inventory is such an account. Because of subjectivity the management may use it for financial statement manipulation. Large changes in inventory compared to sales are typically suspicious, so the sign of the coefficient is undetermined.

Days in receivables index. Days in receivables is simply 365 times the accounts receivables over net sales. Days in receivables index is just the ratio of days in receivables in the current and previous year

$$\text{Days in receivables index} = \frac{\text{Account receivable (current)} \,/\, \text{Net sales (current)}}{\text{Accounts receivable (previous)} \,/\, \text{Net sales (previous)}} \quad (3)$$

This was used by Beneish (1997). He used it to measure whether accounts receivable is out of balance. A large increase in it may indicate that accounts receivable is inflated. Since the increase in accounts receivable likely increases the probability of fraud, the predicted sign of the coefficient is (+).

Debt to equity. This is the ratio of total liabilities and equity. This was used by Fanning and Cogger (1998). Since research suggests that the potential for wealth transfers from debt holders to managers increases as leverage increases. The managers may manipulate financial statements to meet debt covenants. Therefore the more debt there is compared to equity the higher risk for fraud. Therefore the predicted sign of the coefficient is (+).

Demand for financing (ex ante). This is a dummy variable, which is defined based on whether the following condition is true or not

$$\frac{\text{Cash flow from op. activities} - \text{Mean capital expenditure in 3 previous years}}{\text{Total current assets in previous year}} < -0{,}5 \quad (4)$$

If the condition in equation (4) is true, the value is 1, otherwise 0. This variable was created by Dechow et al. (1996) to show whether the firm requires external financing within the next two years (value = 1) or will the internal financing be sufficient for the next two years (value = 0). Dechow et al. used this variable in the study of earnings manipulation. As mentioned by Perols and Lougee (2011) the earnings management and fraud have the same purpose. Since the demand for financing increases the probability of fraud, the predicted sign of the coefficient is (+).

Declining cash sales dummy. This is a dummy variable, which is defined based on the following condition

$$\text{Cash sales (current year)} < \text{Cash sales (previous year)} \quad (5)$$

Where cash sales is net sales minus receivables in the current year plus receivables in the previous year. If the condition in equation (5) is true, the value is 1, otherwise the value is 0. This was used by Beneish (1997). This is used by analysts to measure earnings quality. If cash sales decline, extra financing may be required. This on the hand increases the likelihood of manipulating earnings. Therefore the probability of fraud increases and the predicted sign of the coefficient is (+).

Fixed assets to total assets. This is defined as the ratio of gross total of property, plant and equipment and the total assets. This was used by Kaminski et al. (2004). They cite earlier empirical studies as the basis. They do not have theoretical reasoning which they plainly say. Since there is no real reasoning for this, there is no reasoning for the sign of coefficient, so it is undetermined.

Four year geometric sales growth rate. This is defined as follows

$$\text{Four year geometric growth rate} = \left(\frac{\text{Net sales (current year)}}{\text{Net sales (current -3 years)}}\right)^{\frac{1}{4}} - 1 \qquad (6)$$

This was used by Fanning and Cogger (1998). They actually used geometric growth for the previous two years. Perols (2011) used the definition in equation (6). Reasoning of Fanning and Cogger is that continued growth is motivation for fraud. Rapid growth can also lead to a decrease in the effectiveness of internal controls making it easier to commit fraud. The higher this quantity is, the more likely is fraud, so the predicted sign of the coefficient is (+).

Gross margin. This is defined as the difference of net sales and cost of goods sold, divided by net sales. This was used by Green and Choi (1997) and Lin et al. (2003). Green and Choi mention that this is used for risk assessment of the revenue and collection cycle. Lin et al. actually use this, because it is one of the most commonly used financial ratios in audit. Typically large changes are looked for in this quantity, so the predicted sign of the coefficient is undetermined.

Holding period return in the violation period. This is defined as the difference of annual closing price in the current year and previous, divided by annual closing price in the current year. This was used by Beneish (1999). He uses this as one of the surrogates for an increased

likelihood that a firm is investigated and charged by the SEC. The larger the difference, the more likely that SEC investigates. Therefore the probability of fraud is increased with increasing difference, so the predicted sign of the coefficient is (+).

Industry ROE minus firm ROE. This is defined as the name says, and the return on equity (ROE) defined as the ratio of net income and equity. This was used by Feroz et al. (2000). They use as financial red flag for audit, more specifically to measure profitability. If the ROE of the firm deviates a lot from the industry ROE, the probability of fraud is increased. If firm ROE is larger than industry ROE, it is possible that firm is committing fraud by increasing its revenues or decreasing its costs. If firm ROE is less than industry ROE, it may indicate financial distress, which is a reason for committing fraud. Therefore the sign of the coefficient is undetermined.

Inventory to sales. This is the ratio of inventory and sales. This was used by Kaminski et al. (2004). They had no theoretical justification but based it on earlier empirical results. Both large inventories and sales contain risks. So it is unclear how this ratio affects the probability of fraud. Therefore the predicted sign of the coefficient is undetermined.

Net sales. This the total revenue from sales. This was used by Green and Choi (1997) and Lin et al. (2003). Green and Choi mentioned that this is relevant for risk assessment of revenue and collection cycle. Lin et al. stated that this is one of the most often used account trends in practise. This was also reported to be the most effective for detecting misstatements in revenue cycle. According to ISA 240 attachment 3, IAASB (2018), this is one of the accounts to look for changes in trends for misconduct. The risks are increased with higher net sales, so the predicted sign of the coefficient is (+).

Positive accruals dummy. This is a dummy variable, which is 1, if income before extraordinary items is larger than net cash flow from operating activities both in the current and previous year, otherwise it is 0. This was used by Beneish (1997). He points out that if managers have previously made income increasing accruals, they might attempt to avoid accrual reversals or run out of ways to increase earnings. Since having these accruals makes the financial statement more risky, the probability of fraud is increased with them. Therefore the predicted sign of the coefficient is (+).

Prior year ROA to total assets. The return on assets (ROA) is defined as net income divided by totals assets. The variable itself is the ratio ROA in the previous year and total assets in the current year. This was used by Summers and Sweeney (1998). They use ROA to measure financial performance. Since managers try to keep the financial performance the same or make it better, the larger the ratio, the larger the probability of fraud. Therefore the predicted sign of the coefficient is (+).

Property, plant and equipment to total assets. This is the ratio of net total property, plant and equipment and total assets. This was used by Fanning and Cogger (1998). They tested several variables for their ability to predict fraud and this was one of them. They did not provide theoretical reasoning for this variable although most other times they did provide theoretical reasoning. It is unclear how this should affect the probability of fraud. On the one hand the valuation of property, plant and equipment could contain fraud, so the same fraud would be in total assets. In this case the ratio increases. On the other hand fraud could be in other parts of total assets, in which case the ratio decreases. So the sign of the coefficient is left undetermined.

Sales to total assets. This the ratio of net sales and total assets. This was used by Fanning and Cogger (1998) and Kaminski et al. (2004). Like previous variable Fanning and Cogger tested this variable for its ability to predict fraud and mentioned that it had been previously observed to be significant. Sales is a variable that they mention is more likely to be manipulated by management. Due to two sided accounting the manipulation in sales has a corresponding item in the receivables and therefore the ratio is a useful variable to use. Kaminski et al. do not have theoretical reasoning, but they just cite empirical evidence from previous studies. High sales increase risk of fraud. On the other hand high total assets increase that risk, too. So the predicted sign of the coefficient is left undetermined.

The number of auditor turnovers. This is defined as a sum of auditor turnover in the current year, previous year and 2nd previous year. Auditor turnover in current year is 1, if auditor in the current year is different from the auditor in the previous year, otherwise 0. When the three turnover years are summed, the variable can have four different values: 0 (no auditor changes), 1 (one auditor change in the past three years), 2 (two auditor changes in the past three years)

and 3 (auditor changed every time in the past three years). This was used by Feroz et al. (2000) as an audit oriented red flag. It seems reasonable that if there is a large turnover of auditors, the likelihood of fraud increases, because presumably the auditor is changed in order to conceal fraud. However, turnover might also be due to other reasons like client not paying his bills and auditor gets changed for that reason. The more there is auditor turnover, the more likely it is that there is fraud, too. So the predicted sign of the coefficient is (+).

Times interest earned. This is the sum of total interest and related expenses, income before extraordinary items and total income taxes, divided by the total interest and related expenses. This was used by Feroz et al. (2000) as auditing red flag, more specifically indicating sensitivity to interest rates. There could be problems with any of the items that form this quantity. Therefore the predicted sign of the coefficient is undetermined.

Total accruals to total assets. This is the difference of income before extraordinary items and net cash flow from operating activities, divided by total assets. This was used by Beneish (1997), Dechow et al. (1996) and Beneish (1999). Beneish uses this variable to capture how much of accounting earnings is cash based. Firms violating GAAP tend to have larger accruals. Dechow et al. use only the accruals for detecting earnings management. Since larger accruals contain more risk, the probability of fraud is increased with large accruals. Therefore the predicted sign of the coefficient is (+).

Total debt to total assets. This is the ratio of total liabilities and total assets. This was used by Dechow et al. (1996). They use this as a proxy for the closeness of debt covenants in studying earnings management. Covenants can trigger the payment of the debt. Therefore the larger the debt, the larger the risk of earnings management and fraud. So the predicted sign of the coefficient is (+).

Total discretionary accruals. Total discretionary accruals in the current year are defined as

$$\text{Total discretionary accruals} = DA_{t-1} + DA_{t-2} + DA_{t-3} \tag{7}$$

$$DA_t = \frac{TA_t}{A_{t-1}} - NDA_t \tag{8}$$

$$\frac{TA_t}{A_{t-1}} = \frac{\text{Income before extraordinary items}_t - \text{net cash flow from operating activities}_t}{\text{Total assets}_{t-1}} \quad (9)$$

$$\begin{aligned} NDA_t = (1 &+ \text{Sales in the current year} - \text{Sales in the previous year} \\ &- \text{Receivables in the current year} \\ &+ \text{Receivables in the previous year} \\ &+ \text{Net cash flow from opearating activities in the current year} \\ &- \text{Net cash flow form operating activities in the previous year} \\ &+ \text{Gross total property, plant and equipment in the current year}) \\ &/ \text{Total assets in the previous year} \end{aligned} \quad (10)$$

This was used by Perols and Lougee (2011). They used this to capture the pressure of earnings reversals and earnings management limitations. This is the part of the accruals that the management can use for the earnings management. The larger the total discretionary accruals, the more likely is the earnings management and with it fraud. So the predicted sign of the coefficient is (+).

Whether accounts receivable > 1,1 * of last year's accounts receivable. This is a dummy variable, which is 1 if the condition in the name is true, otherwise it is 0. This was used by Fanning and Cogger (1998). Earlier work in trend analysis has established that auditors and analysts use 10% change as a threshold for material change in accounts or ratios according to them. This is also mentioned in ISA 240 attachment 3, IAASB (2018), as an example of potential misconduct. Increase in accounts receivable is connected to the increased probability of fraud. So the predicted sign of the coefficient is (+).

Whether gross margin percent > 1,1 * of last year's gross margin percent. This is a dummy variable, which is 1 if the condition in the name is true, otherwise 0. This was used by Fanning and Cogger (1998). Earlier work in trend analysis has established that auditors and analysts use 10% change as a threshold for material change in accounts or ratios, according to them. This is the reasoning as with the change in account receivable in the previous variable. Therefore in the same way the predicted sign of the coefficient is (+).

Whether new securities were issued. This is a dummy variable, which is 1, if common shares outstanding in the current year is greater than common shares outstanding in the previous year or common shares issued in the current year is greater than 0, or both, otherwise the binary variable is 0. This was used by Dechow et al. (1996). They use this variable to measure the demand for external financing, when earnings have already been manipulated. Demand for external financing is related to the increased probability of fraud. So the predicted sign of the coefficient is (+).

Whether Standard Industry Classification Code larger than 2999 and smaller than 4000. This is a binary variable, which is 1, if standard industry classification code is in the range mentioned in the name, otherwise it is 0. This was used by Lee et al. (1999). They provided descriptive statistics, which showed that with firms having SIC in the range 3000 – 3999 the fraud percentage was larger than with firms outside it. No theoretical reasoning was given for this. They obtained better results using this indicator variable than using separate dummy variables for the two digit SIC groups. Since the range is riskier, the predicted sign of the coefficient is (+).

Value of Issued Securities to Market Value. This is defined as the market value of common shares issued divided by the market value of common shares outstanding, if there are common shares issued. If common shares are not issued, then if there are more common shares outstanding in the current year than in the previous year, the variable is the difference of market values of common shares outstanding in the current and previous year, divided by the market value of common shares outstanding in the current year. If both previous conditions fail, meaning no common shares were issued and the number of common shares does not change between previous and current year, then the value of the variable is 0. This was used by Dechow et al. (1996). They used this for earnings management in order measure the need for external financing while the earnings management is ongoing. This is the corresponding real variable to the dummy whether new securities are issued. The more financing is needed, the more value should the issued securities have. Therefore the probability of fraud should increase with increasing value of issued securities, so the predicted sign of the coefficient is (+).

Unexpected Employee Productivity. This is defined as the employee productivity in the firm minus the average employee productivity in the industry, where the employee productivity is defined as follows

$$\text{Employee productivity} = \frac{\frac{\text{Net sales (current)}}{\text{Number of employees (current)}} - \frac{\text{Net sales (previous)}}{\text{Number of employees (previous)}}}{\frac{\text{Net sales (current)}}{\text{Number of employees (current)}}} \quad (11)$$

This was used by Perols and Lougee (2011). They used it to identify unusual relations between revenue and the number of employees. Large deviation from industry average is unusual. Since large deviation can occur in both directions, the predicted sign of the coefficient is undetermined.

3. Data and Methods

The third chapter of this study outlines the research methods and data used in creating answers to the research questions.

3.1 Logistic regression as the method of fraud classification

The task at hand is to classify whether an observation is fraud or not. This is a binary task where the dependent variable y is

$$y = \begin{cases} 1, & if\ fraud \\ 0, & if\ non-fraud \end{cases} \tag{12}$$

The logistic model is used to produce predictions $\hat{y}$ from variables $x_i, i = 1, \dots, m$ with parameters $\beta_0, \beta_1, \dots, \beta_m$, see p. 575 in Wooldridge (2009). The logistic function produces a probability

$$\begin{gathered} p(y = 1|\boldsymbol{x}) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m) \\ p(y = 0|\boldsymbol{x}) = 1 - p(y = 1|\boldsymbol{x}) \\ G(z) = \frac{1}{1 + e^{-z}} \end{gathered} \tag{13}$$

The logistic function G maps the usual linear regression result to the interval [0, 1] making it possible to interpret as probability. The prediction of the model is

$$\hat{y} = \begin{cases} 1, & if\ p(y = 1|\boldsymbol{x}) > 0.5 \\ 0, & if\ p(y = 1|\boldsymbol{x}) \leq 0.5 \end{cases} \tag{14}$$

The usual method for solving the parameters is the maximum likelihood estimation where the log-likelihood function is maximized

$$\begin{gathered} L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log\big(G(z_i)\big) + (1 - y_i) \log\big(1 - G(z_i)\big) \\ z_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i} \end{gathered} \tag{15}$$

However, the number of variables here is so much that the model easily overfits and there might be multi-collinearity issues with the model. In order to reduce these and to get a better generalization error a regularization term is added and the objective function to be used in the optimization is

$$J(\boldsymbol{\beta}) = -L(\boldsymbol{\beta}) + \frac{\lambda}{2}|\boldsymbol{\beta}|^2, \boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p) \tag{16}$$

Since the likelihood term is preceded by a negative sign the objective function is minimized to find the optimal solution. Chapter 7 in Goodfellow et al. (2016) deals with issues of regularization. The regularization chosen here is of the simplest type. In the book by Goodfellow et al. there are more choices presented, like absolute values of coefficients are taken instead of squared values. The squared values have better mathematical behaviour, so they are used here. Many times the coefficient of the intercept, β_0, is not included in the regularization term.

The addition of regularization has two effects: it keeps coefficients smaller because it penalizes the high values of the coefficients (note intercept term is not being penalized) and it reduces effects of multi-collinearity. The latter effect can be seen by forming the covariance matrix which is just the inverse of the information matrix, which is just the Hessian matrix of the negative likelihood. For logistic regression the covariance matrix is

$$\begin{aligned} I(\boldsymbol{\beta})_{ij} &= -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_i\,\partial\beta_j}, \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p) \\ Cov(\boldsymbol{\beta}) &= I(\boldsymbol{\beta})^{-1} \end{aligned} \tag{17}$$

For logistic regression the information matrix has been calculated in p. 35 of Hosmer and Lemeshow (2000)

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix}$$

$$V = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{pmatrix} \tag{18}$$

$$I(\boldsymbol{\beta}) = X^T V X$$

where $\hat{\pi}_i$ is the estimated probability of case *i*. Since for covariance the information matrix *I* has to be inverted, the matrix has to be invertible. If there are collinearities between variables, then the matrix may well fail to be invertible. The problem persists even when the lowest eigenvalue of the information matrix is close to 0. Without matrix *V* the information matrix would match the one in linear regression. With the addition of the regularization term the information matrix[5] becomes

$$I(\boldsymbol{\beta}) = X^T V X + \lambda \mathbb{1} \tag{19}$$

Since the first part of the information matrix is positive definite and symmetric and $\lambda > 0$, the eigenvalues of the information matrix have a lower bound

$$Eigenvalues(I) \geq \lambda \tag{20}$$

which guarantees the invertibility of the information matrix and the covariance matrix now exists and thus the multi-collinearity issue is reduced. However, the regularization term λ can be small and in this case the multi-collinearity might again become an issue, just like in linear regression it is not required that there is perfect collinearity but rather a strong correlation between variables. Therefore the regularization cannot be too small, so that one does not run into numerical instability.

[5] This is no longer an actual information matrix, because with regularization equation (16) is no longer an actual likelihood function.

Unfortunately maximizing the likelihood in equation (15) or minimizing the regularized likelihood in equation (16) do not have closed form solutions, so one has to resort to numerical methods. A typical method is Newton-Rhapson, which results into estimators for the coefficients, see for example section 5 in van Wieringen (2015) how to do this. From section 5.3 of van Wieringen we get also the covariance and bias of the estimator

$$\begin{aligned} E(\boldsymbol{\beta_{new}}) &= (X^T W X + \lambda \mathbb{1})^{-1}\left(X^T W X \boldsymbol{\beta_{old}} + X^T(E(\boldsymbol{y}) - \boldsymbol{\pi_{old}})\right) \\ Var(\boldsymbol{\beta_{new}}) &= (X^T W X + \lambda \mathbb{1})^{-1}\, X^T W X\, (X^T W X + \lambda \mathbb{1})^{-1} \end{aligned} \tag{21}$$

where the subscripts new and old refer to the Newton-Rhapson iterative algorithm, in which a new value is calculated based on the old one. The bias depends on the old coefficients directly, whereas the variance depends on the old coefficients only through the estimated probabilities in the matrix W. Although the regularization solves multicollinearity, the estimator is biased. Furthermore the bias actually depends on the actual coefficients, not the estimated ones, but we do not know the actual ones. The bias can be estimated using the estimated coefficients. As can be seen the variance goes down with increasing regularization parameter, this can also be seen in Figure 5.2 of van Wieringen (2015).

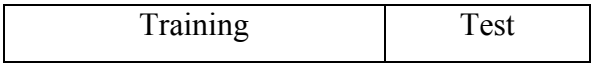
Another way to get rid of multicollinearity would be to remove the variables suffering from collinearity issues. Since the interest here is the prediction of fraud or fraud type, which means that fraud or fraud type is the dependent variable, not the independent one, the problem of collinearity is not so serious. If the interest were to explain the fraud or fraud type with particular variables, the approach to throw away the collinear variables would be better. But the main question is whether a financial statement is predicted to be fraudulent or not. Since the question is about the dependent variable, it can be answered even with perfectly collinear variables. Namely the coefficient can be estimated with suitable methods, which do not rely on inverting the singular covariance matrix, such as gradient descent, see chapter 4 in Goodfellow et al. (2016). With perfectly collinear variables without regularization the actual estimate would be an estimate of a sum of the coefficients of the collinear variables. A numerical method would find an estimate for the two coefficients such that their sum would equal the actual estimate. The numerical method could arrive in any possible combination, where the sum of the coefficients equals the actual estimate, at least within numerical accuracy of the computation. Whichever of these combinations is used is inconsequential for the dependent

variable, which only sees the effect of the sum of the coefficients of the perfectly collinear variables. So it is uniquely defined even with collinear variables. And its statistics is the one of binary variables in the end. The problem arises, if one wishes to ask how much each variable contributes to the dependent variable. And how significant are the coefficients. The regularization is just another way of doing the analysis with the collinear variables, where the regularization causes a unique choice of a combination of the coefficients, but introduces a bit of bias along the way. The conclusion is that even with collinear variables questions about the dependent variable can be handled.

3.2 Training and testing sets

Typically in machine learning one uses training and testing sets, chapter 5 in Goodfellow et al. (2016). Idea is that the model is developed and optimized on the training set and then the results, that presumably generalize to new data, are estimated based on test set. The way to define the training and test sets is random sampling to the two sets. There are no exact rules on how big the divisions should be but typically the training set takes 70-80% of the data and test set is left with the rest 20-30%. One may also have to use stratified sampling, if the classes are imbalanced. Essentially the data is divided in the following manner

Figure 1 Training and test set partition



However, there is not always enough data for having training and test sets. Then one can use cross-validation where one divides the data into k separate parts, called folds, and the kth fold is used as a test set and the folds 1, …, k-1 are used as a training set where a model is fitted, then the (k-1)th fold is used as test set and the other k-1 folds as training set where another model is fitted, etc. After this has been repeated k times the results are then averaged over the folds. This way all the data is used and nothing is wasted but nevertheless training and test sets

do not overlap per fold. This is called *k*-fold cross-validation. Below is an example of 3-fold cross-validation

Figure 2 Example: 3-fold cross validation

1. Fold	Training		Test
2. Fold	Training	Test	Training
3. Fold	Test	Training	

so at every fold 2/3 of data is used for training and 1/3 for testing. Since the data set is imbalanced, before fitting the training data is balanced by replicating the minority class. Perols (2011) did the balancing using undersampling on the non-fraud data and in Perols et al. (2017) the undersampling was done in much more sophisticated ways, that improved the results of Perols (2011) considerably. In Dutta et al. (2017) the balancing was done using SMOTE algorithm, which oversamples the fraud data and creates new samples, which are not exact replicas of the old ones.

3.3 Performance measures

The problem with unbalanced samples is that the accuracy is not necessarily a good measure of performance. With binary classification there are 4 possibilities: true positives model predicts fraud when fraud is present, false positives model predicts fraud when it is not fraud, true negatives model predicts non-fraud when it is not fraud and false negatives model predicts non-fraud when it is fraud. This is contained in the following table

Table 1 Confusion matrix

	Actual fraud	Actual non-fraud
Model predicts fraud	True positives (TP)	False positives (FP)
Model predicts non-fraud	False negatives (FN)	True negatives (TN)

Here the false negatives are the problem, because the costs of misclassifying them are much higher than the costs related false positives. Usual performance measures for this table are

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FP + FN + TN} \\ Precision &= \frac{TP}{TP + FP} \\ Sensitivity\ or\ recall &= \frac{TP}{TP + FN} \\ Specificity &= \frac{TN}{TN + FP} \end{aligned} \tag{22}$$

Accuracy tells the probability of predicting correctly over all observations. Precision tells the probability of predicting fraud correctly among all the cases where the classifier predicted fraud, so how often is the fraud classification correct. Sensitivity tells the probability of the model of finding fraud among all the fraud cases. Specificity is for non-fraud the analogue of what sensitivity is for fraud, namely it tells the probability of model finding non-fraud among all the cases of non-fraud. On top of these quantities one can define many more. Also there are other names for these quantities, too. See for example the web-page about Precision and Recall (2019), which contains a thorough list of the different performance measures and the different names for them.

The standard errors for these quantities can be calculated as

$$s.e.(PM) = \sqrt{\frac{1}{N} PM\ (1 - PM)} \tag{23}$$

where PM is any one of the above defined performance measures and N is the number of cases in the whole class for the performance measure

$$N = \begin{cases} TP + FP + FN + TN, & PM = Accuracy \\ TP + FP, & PM = Precision \\ TP + FN, & PM = Sensitivity \\ TN + FP, & PM = Specificity \end{cases} \tag{24}$$

Which performance measure should one use? There is no simple answer to this. One has to look at them in combination, since the performance measures describe different aspects of the classifier. One might think that accuracy is the most important performance measure. However, since we are more interested in getting the fraud cases predicted, the precision is the more interesting measure of performance. As an example let us suppose that there are 1% of fraud cases. Then one can easily have a classifier that is 99% accurate, just predict everything to be non-fraud. This results into TP = FP = 0, FN = 0,01 * Total and TN = 0,99 * Total. Putting into the equation for accuracy the result is 0,99. However, 0 fraud cases are observed with this classifier. If the purpose is to find fraud the 99% accuracy of the classifier meant nothing. This does not mean that accuracy is not important but rather that all the performance measures have to be looked at. On the other hand if everything is predicted to be fraud, then FN = TN = 0 and sensitivity = 1, specificity = 0, accuracy = precision would be small since the number of frauds is typically much smaller than number non-frauds TP << FP.

All the performance measures that have so far been defined are made for just single threshold probability of classification, which so far has been taken to be 0.5. However, it may be useful use a different threshold, like if one wants to find more fraud and is willing to sacrifice accuracy for it, the threshold may be dropped below 0.5. With a new threshold probability all the quantities above would have to be recalculated. There would have to be values for all different threshold probabilities. This is not practical. Therefore another tool is used for the changing threshold probability: a graphical device called a Receiver Operating Characteristic (ROC) curve where 1-Specificity vs. Sensitivity are drawn into the same figure while changing the threshold probability p_0

$$\hat{y} = \begin{cases} 1, & if\ p(y = 1|\boldsymbol{x}) > p_0 \\ 0, & if\ p(y = 1|\boldsymbol{x}) \leq p_0 \end{cases} \tag{25}$$

As an example let us produce ROC curve for coin toss or random choice. For every value of p_0 there are TP + FN fraud samples and FP + TN non-fraud samples. If the choice is truly random then predicted numbers will be split in fraud and non-fraud samples on threshold probability ratio

$$
\begin{aligned}
TP &= (1 - p_0)(TP + FN) \\
FN &= p_0\,(TP + FN) \\
FP &= (1 - p_0)(FP + TN) \\
TN &= p_0\,(FP + TN)
\end{aligned}
\tag{26}
$$

Now we can form the sensitivity and specificity

$$
\begin{aligned}
Sensitivity &= \frac{TP}{TP + FN} = 1 - p_0 \\
Specificity &= \frac{TN}{TN + FP} = p_0 \\
1 - Specificity &= 1 - p_0
\end{aligned}
\tag{27}
$$

The ROC curve traces out a curve (1-p_0, 1-p_0) when $0 \leq p_0 \leq 1$. This is just a straight line starting from (1, 1) and ending to (0, 0). A perfect classifier would have FP = FN = 0, so that the ROC curve shrinks to a point (0, 1). In general a classifier traces a curve between (1, 1) and (0, 0). The classifier performs better than another if it is above and to the left of the other one.

A performance measure related to the ROC curve is the Area Under the Curve (AUC). This is just the area between the ROC curve and the horizontal axis. For random choice this is 0.5 (area of the triangle with sides 1). Typically a classifier has to beat this number in order to be of any value. Standard error for AUC can be calculated as, see Hanley and McNeil (1982),

$$
\begin{aligned}
&s.e.(AUC) \\
&= \sqrt{\frac{AUC\,(1 - AUC) + \big(n_{fraud} - 1\big)(Q_1 - AUC^2) + \big(n_{non-fraud} - 1\big)(Q_2 - AUC^2)}{n_{fraud}\,n_{non-fraud}}} \\
&Q_1 = \frac{AUC}{2 - AUC} \\
&Q_2 = \frac{2\,AUC^2}{1 + AUC}
\end{aligned}
\tag{28}
$$

The assumption in Hanley-McNeil formula for standard error is that the two classes, here fraud and non-fraud, are both normally distributed. This may well be a reasonable assumption, if there are a lot cases in both classes, but here the number of fraud cases is typically small. Since there are not much better solutions for standard error of AUC, the above formula is used for it. Any kind of inferences based on this formula have to be taken with the caveat, that class of fraud may not be normally distributed due to small number of cases. This may make the inference unreliable.

All the performance measures so far do not take into account the cost of misclassifying fraud. Typically the cost of misclassifying fraud as non-fraud is much higher than the cost of misclassifying non-fraud as fraud. Therefore in Perols (2011) expected cost of misclassification is presented, which has originally been used by Dopuch et al. (1987) for predicting audit qualifications. Their formula can be derived by assuming a joint probability distribution of predictions and actual cases of fraud. That probability distribution has an expected value

$$\begin{aligned} ERC = E(C) = C_{f,f}\, P(f,f) + C_{nf,f}\, P(nf,f) + C_{f,nf}\, P(f,nf) \\ + C_{nf,nf}\, P(nf,nf) \end{aligned} \tag{29}$$

where $P(i, j)$ is the probability of predicting i when actual condition is j and $i, j = f$, nf with $f =$ fraud and $nf =$ non-fraud, $C_{i,j}$ is the extra cost of predicting i when the condition is j. If the prediction equals the actual condition there is no extra cost involved, so

$$\begin{aligned} C_{f,f} = C_{nf,nf} = 0 \\ C_{nf,f} = C_{FN}, \qquad C_{f,nf} = C_{FP} \end{aligned} \tag{30}$$

The expected relative cost becomes

$$\begin{aligned} ERC = C_{FN}\, P(nf,f) + C_{FP}\, P(f,nf) \\ = C_{FN}\, \frac{P(nf,f)}{P(f)}\, P(f) + C_{FP}\, \frac{P(f,nf)}{P(nf)}\, P(nf) \end{aligned} \tag{31}$$

where $P(i)$ refers to the probability of actual condition being i. The ratios of probabilities can be estimated with the frequencies of the model, which are evaluated by

$$\frac{P(nf,f)}{P(f)} = \frac{FN}{TP+FN}$$
$$\frac{P(f,nf)}{P(nf)} = \frac{FP}{FP+TN} \quad (32)$$

Putting the results of equation (31) into equation (30), we get the result, which has been presented by Perols (2011) and Dopuch et al. (1987)

$$ERC = \frac{FN}{TP+FN}\, C_{FN}\, P(fraud) + \frac{FP}{TN+FP}\, C_{FP}\, P(non-fraud) \quad (33)$$

where C_{FN} is the cost of classifying fraud as non-fraud and C_{FP} the cost of classifying non-fraud as fraud, P(fraud) is the prior fraud probability that exists at evaluation time (should be the real probability of fraud) and P(non-fraud) the prior probability of non-fraud. It is expected that $C_{FN} >> C_{FP}$ and P(fraud) << P(non-fraud). Trouble is that these quantities are not known and are hard to estimate. Therefore one typically estimates equation (32) with different values of prior fraud probability and different ratios of C_{FP} / C_{FN} leaving either of the costs undetermined.

Expected relative cost is calculated for the fraud prediction only, not for fraud type prediction. The reason is that with fraud type prediction there can be more general types of errors. If the model predicts fraud type, the actual situation can be that the fraud type was predicted correctly. But if it is not correct, there are two types of mistakes here now: it is actually non-fraud so it is not of any fraud type, it is actually fraud but of different type than what the model was predicting. Second if the model predicts not of this fraud type, there are now three situations: condition is actually non-fraud so it is not of any fraud type and prediction is correct, condition is actually not of this fraud type but is in reality fraud of other type, and last the condition is actually this fraud type so the prediction was not correct. If the cost of false positives and false negatives is equal, then one could just put the same cost for everything and one could easily use the above formula for the expected relative costs. However, if they are different and usually the cost of false negatives, predicting fraud as non-fraud, is much higher than the cost of false positives, predicting non-fraud as fraud, it seems that using fraud type prediction cost of false negatives for the case of predicting not of this fraud type when the case is of different fraud type than what is being looked for in the model. This issue is not raised in

Perols et al. (2017) but rather they use the same expected relative cost as with just fraud prediction.

3.4 Data

The data is obtained from WRDS in two parts: one part from Compustat taking financial announcements from the period 1.1.1991 – 31.12.2016 and the other part from Audit Analytics taking restatements from the period 1.1.1995 – 9.5.2019 (all the data obtainable).[6] The datasets are combined per Company Identity Key and the financial period. Restatements can be given for much longer periods than the financial period, for example one of the fraud examples has restatement for period 1.1.2001 – 31.12.2004 meaning that all the 4 years of financial statements contain fraud. The matching is made in such a way that Company Identity Keys have to match and the financial period of the financial statement has to be contained in restatement period meaning

$$\begin{aligned} &\text{Restatement company identity key} = \text{Financial statement company identity key} \\ &\text{Restatement end date} \geq \text{Financial statement begin date} \\ &\text{Restatement begin date} \leq \text{Financial statement end date} \end{aligned} \tag{34}$$

If the above conditions hold, then the financial statement is marked fraudulent using a dummy variable with value 1 for fraud, otherwise it is not fraudulent with value 0. If it is fraudulent, also its types are marked using their own dummy variables.[7]

The dummy variables describing fraud and its types are named according to the key number. The names can be found in table 3 for the fraud categories used here. Altogether Audit Analytics data contained 42 categories of fraud and 301 cases of fraud in the beginning.

[6] The reason in the difference of time periods is that Audit Analytics has data from 1.1.1995 onwards and some of the variables used require financial data from 4 previous years, so Compustat data is taken from 1.1.1991 to take this into account. Second the restatements are really given only afterwards, so the financial statement data has to be restricted to some latest date, which was chosen here to be 31.12.2016.

[7] One restatement can contain several fraud types. One dummy variable per fraud type is added.

From Compustat financial data 35 variables used to predict fraud are formed that were defined in section 3.2.1. These are the same as used Perols (2011) minus the 8 variables that could not be defined. Next table contains all the 35 variables used in this thesis. The first column contains the number with which to identify the fraud predictor, the second column the actual name and the third column the definition in terms of Compustat variables.

Table 2 Prediction variables, t refers to the end date of financial period, t-1 to the end date of previous financial period, and so on.

Number	Name	Definition using Compustat variables
1	Accounts receivable	$RECT_t$
2	Accounts receivable to sales	$RECT_t / SALE_t$
3	Accounts receivable to total assets	$RECT_t / AT_t$
4	Allowance of doubtful accounts (AFDA)	$RECD_t$
5	AFDA to accounts receivable	$RECD_t / RECT_t$
6	AFDA to net sales	$RECD_t / SALE_t$
7	Altman Z-score	$3,3 * (IB_t + XINT_t + TXT_t) / AT_t + (0,999 * SALE_t + 1,2 * WCAP_t + 1,4 * RE_t) / AT_t + 0,6 * CSHO_t * PRCC_t / LT_t$
8	Big 4 auditor	IF $0 < AU_t < 9$ THEN 1 ELSE 0
9	Current minus prior year inventory to sales	$INVT_t / SALE_t – INVT_{t-1} / SALE_{t-1}$
10	Days in receivables index	$(RECT_t / RECT_{t-1}) * (SALE_{t-1} / SALE_t)$
11	Debt to equity	LT_t / CEQ_t
12	Demand for financing (ex ante)	IF $(OANCF_t – (CAPX_{t-3} + CAPX_{t-2} + CAPX_{t-1}) / 3) / ACT_{t-1} < -0,5$ THEN 1 ELSE 0
13	Declining cash sales dummy	IF $(SALE_t – RECT_t + RECT_{t-1}) < (SALE_{t-1}$

		– $RECT_{t-1}$ + $RECT_{t-2}$) THEN 1 ELSE 0
14	Fixed assets to total assets	$PPEGT_t$ / AT_t
15	Four year geometric sales growth rate	$(SALE_t / SALE_{t-3})^{1/4} – 1$
16	Gross margin	$1 – COGS_t / SALE_t$
17	Holding period return in the violation period	$1 – PRCC_{t-1} / PRCC_t$
18	Industry ROE minus firm ROE	INDUSTRY(NI_t / CEQ_t) – FIRM(NI_t / CEQ_t)
19	Inventory to sales	$INVT_t / SALE_t$
20	Net sales	$SALE_t$
21	Positive accruals dummy	IF (($IB_t – OANCF_t$) > 0 AND ($IB_{t-1} – OANCF_{t-1}$) > 0)) THEN 1 ELSE 0
22	Prior year ROA to total assets	$(NI_{t-1} / AT_{t-1})/ AT_t$
23	Property, plant and equipment to total assets	$PPENT_t / AT_t$
24	Sales to total assets	$SALE_t / AT_t$
25	The number of auditor turnovers	IF $AU_t \neq AU_{t-I}$ THEN 1 ELSE 0 + IF $AU_{t-1} \neq AU_{t-2}$ THEN 1 ELSE 0 + IF $AU_{t-2} \neq AU_{t-3}$ THEN 1 ELSE 0
26	Times interest earned	$1 + (IB_t + TXT_t) / XINT_t$
27	Total accruals to total assets	$(IB_t – OANCF_t) / AT_t$
28	Total debt to total assets	LT_t / AT_t
29	Total discretionary accruals	$DA_{t-1} + DA_{t-2} + DA_{t-3}$ where $DA_t = TA_t / A_t$ – estimated(NDA_t) $TA_t / A_t = (IB_t – OANCF_t) / AT_{t-1}$ $NDA_t = (1 + SALE_t – SALE_{t-1} – RECT_t + RECT_{t-1} + OANCF_t – OANCF_{t-1} + PPEGT_t) / AT_{t-1}$
30	Whether accounts receivable >	IF ($RECT_t / RECT_{t-1}$) > 1,1 THEN 1 ELSE

	1,1 * of last year's accounts receivable	0
31	Whether gross margin percent > 1,1 * of last year's gross margin percent	IF $(1 - COGS_t / SALE_t) / (1 - COGS_{t-1} / SALE_{t-1}) > 1{,}1$ THEN 1 ELSE 0
32	Whether new securities were issued	IF $(CSHO_t - CSHO_{t-1}) > 0$ OR $CSHI_t > 0$ THEN 1 ELSE 0
33	Whether Standard Industry Classification Code larger than 2999 and smaller than 4000	IF $SIC_t > 2999$ AND $SIC_t < 4000$ THEN 1 ELSE 0
34	Value of Issued Securities to Market Value	IF $(CSHI_t > 0)$ THEN $(CSHI_t / CSHO_t)$ ELSE IF $(CSHO_t - CSHO_{t-1} > 0)$ THEN $(1 - CSHO_{t-1} / CSHO_t)$ ELSE 0
35	Unexpected Employee Productivity	$FIRM((SALE_t / SALE_{t-1}) * (EMP_{t-1} / EMP_t) - 1) - INDUSTRY(((SALE_t / SALE_{t-1}) * (EMP_{t-1} / EMP_t) - 1)$

Once the variables are formed, all the cases with missing values are deleted and the data is restricted to cases whose financial period starts at 1.1.1995 or later. The continuous variables contain some extreme values. These are dealt with by applying winsorization, where the bottom 1% of values are set to 1 percentile value and the top 1% of values are set to 99 percentile value. Dummy variables and variable 25 (auditor turnover) were not winsorized. Winsorization was also used by Dutta et al. (2017).

Variables 18 and 35 contain industry average. These are calculated per year and per standard industry classification code (SIC). Once they are calculated the final variable values are calculated. The variables are named according to the number in the above table. The original Compustat dataset contained 260282 cases. After the variables are formed there are 59239 cases left.

After the variables are formed the Audit Analytics and Compustat datasets are combined into one per Company Identity Code and financial period. The resulting sample has 347 cases of fraud and 58892 non-fraud cases. Thus 0,59% of cases contain fraud. In comparison Perols (2011) had fraud in 0,32% of cases. The amount of fraud cases with Perols was 51 in the final sample with 15934 non-fraud cases. Perols studied only years from the fourth quarter of 1998 to the fourth quarter of 2005 and his data source did not include Audit Analytics but used SEC's Accounting and Auditing Enforcement Releases as a source of identifying fraud cases. Furthermore he had 43 prediction variables compared to the 35 here, which have more possibilities of containing missing data. Another benchmark is the misstatement research done by Dutta et al. (2017). Fraud is one part there and their study contains 109 cases of fraud among 3513 cases of misstatements with 60720 non-restatement cases. This seems to be far off. However they start from 260 cases of fraud covering years 1995-2014 in the Audit Analytics dataset (I tried with the time period and got 269 cases, so some new ones have appeared). They further restrict to the years 2001-2014 and they have 112 variables in their study because they study misstatements and not just fraud study, the same choice is made by Perols (2011). Another difference is that Dutta et al. keep only one restatement year per restatement case, same was done by Hennes et al. (2014), in order to reduce firm-level effects. This is not possible to do here because the amount of fraud types would drop so low that any kind of analysis with them becomes useless and in the study of misstatements there are many more cases available than for fraud. Perols et al. (2017) likely get away with the problem of having fewer fraud types, because of the sophisticated undersampling that they are using. Taking into account the reduced time period, this study contains 1995 – 04/2019 restatements in Audit Analytics dataset, and the many more variables that can contain missing values whose cases are deleted, the numbers are not really that far off.

Next some of the fraud categories are combined into one because they contain too few cases to be of use. The ones with over 70 observations in the type category are kept, the rest are combined into a common category called Other. The following table describes the fraud type categories present in the final sample

Table 3 Fraud types

Category key	Frequency	Category title
6	137	Revenue recognition issues
7	74	Expense (payroll, SGA, other) recording issues
11	132	Foreign, related party, affiliated, or subsidiary issues
12	99	Liabilities, payables, reserves and accrual estimate failures
14	96	Accounts/loans receivable, investments & cash issues
20	94	Inventory, vendor and/or cost of sales issues
44	94	Foreign, subsidiary only issues (subcategory)
Other	138	All the other categories that are not listed above combined into here
Fraud	347	

The original Audit Analytics dataset on restatements contained 301 fraud cases. Here the number is 347. This happens because one restatement can contain several years i.e. more than one financial period. For example some restatements covered four years which could lead to possibly 4 or 5 financial statements containing fraud depending on how the financial statement periods compared to the restatement period, restatement might start in the middle of the first year and end in the middle of the 5th year thus affecting 5 financial statements.

It is also of interest to see how the fraud types are distributed through financial periods

Table 4 Fraud cases and its types distributed through 1996 – 2016

Year	**6**	**7**	**11**	**12**	**14**	**20**	**44**	**Other**	**Fraud**
1996	1	1	2	2	2	0	2	2	4
1997	3	1	4	2	4	0	3	4	7
1998	10	2	7	6	9	4	5	7	17
1999	11	2	8	7	11	7	5	8	22

2000	11	5	15	9	12	7	9	14	29
2001	15	6	14	12	8	9	9	16	35
2002	15	5	9	7	5	9	4	10	28
2003	12	9	10	9	4	5	6	11	26
2004	9	8	9	6	4	2	5	9	20
2005	7	4	7	4	3	2	4	5	14
2006	7	4	5	2	3	4	4	4	15
2007	8	2	5	2	4	6	3	4	15
2008	6	3	3	3	2	5	2	2	11
2009	4	3	3	2	3	4	3	4	11
2010	5	3	6	2	6	6	6	5	15
2011	3	3	5	4	3	6	5	5	12
2012	3	2	6	3	3	6	6	6	14
2013	2	2	5	6	3	5	5	8	16
2014	2	4	2	4	2	3	2	5	13
2015	3	4	5	4	4	3	4	5	15
2016	0	1	2	3	1	1	2	4	8
Total	137	74	132	99	96	94	94	138	347

It can be seen that during 1999 – 2004 there have been over 20 fraud cases per year. After that the number has dropped below 20 cases per year. During 2008 and 2009 when the latest financial crisis started, there have the lowest numbers of cases which is a bit surprising. On the other hand poor economic conditions lead to less money available and therefore less possibilities to commit fraud.

4. Findings

4.1 Descriptive statistics

4.1.1 Continuous predictor variables

Next table contains the descriptive statistic for all the continuous predictor variables of the final sample. For the final sample the data has been winsorized to 1% and 99% values of variables.

Table 5 Continuous variables descriptive statistics

count = 59239	Descriptive statistics						
Fraud predictor number	Mean	Std	Min	25%	Median	75%	Max
1	393,264	1219,445	0,058	7,034	38,613	191,376	8918,0
2	0,170	0,107	0,008	0,107	0,155	0,210	0,688
3	0,172	0,122	0,006	0,082	0,148	0,232	0,599
4	15,415	51,860	0,000	0,200	1,106	5,902	382,000
5	0,068	0,128	0,000	0,014	0,031	0,065	0,922
6	0,010	0,019	0,000	0,002	0,004	0,010	0,138
7	2,344	8,990	-53,690	1,269	2,782	4,715	29,690
9	-0,002	0,051	-0,244	-0,010	0,000	0,008	0,211
10	1,068	0,509	0,217	0,869	0,993	1,131	4,372
11	1,280	4,373	-18,532	0,359	0,888	1,775	26,273
14	0,534	0,420	0,023	0,214	0,424	0,750	2,152
15	0,091	0,232	-0,358	-0,011	0,053	0,138	1,487
16	0,353	0,316	-1,663	0,225	0,352	0,516	0,912

17	-0,362	1,393	-9,214	-0,424	0,008	0,259	0,875
18	-7,13E-17	0,985	-6,075	-0,213	-0,022	0,149	6,078
19	0,116	0,123	0,000	0,014	0,093	0,169	0,667
20	2712,64	8020,28	0,546	49,380	279,57	1395,95	57428,0
22	-0,028	0,170	-1,496	0,000	0,000	0,000	0,024
23	0,252	0,217	0,005	0,081	0,185	0,364	0,879
24	1,186	0,841	0,070	0,614	0,994	1,513	4,656
26	193,480	789,413	-236,98	-1,473	3,350	13,747	3535,3
27	-0,099	0,230	-1,616	-0,111	-0,055	-0,014	0,266
28	0,593	0,487	0,067	0,338	0,521	0,694	3,780
29	-2,701	2,609	-19,235	-3,366	-2,135	-1,267	0,491
34	0,943	0,354	0,000	1,000	1,000	1,039	1,743
35	5,632E-18	0,354	-1,654	-0,142	-0,023	0,074	2,642

There are variables with different scales in the above table. For this reason the training set is standardized before fitting the logistic regression model, the corresponding test set is transformed using the training set mean and standard deviation used in the standardization.

It is quite difficult to make determinations of the variables based on the table above. This is why the continuous predictor variables are divided into 4 classes based on the quartile ranges in the fraud type. The resulting fraud and non-fraud classes are tested with the χ^2-homogeneity test which tests the equality of the different ratios belonging to different classes. The χ^2-homogeneity test used here

$$
\begin{aligned}
H_0\colon\ & \pi_{fraud\ type}(x < x_1) = \pi_{not\ fraud\ type}(x < x_1) \\
& \pi_{fraud\ type}(x_1 \le x < x_2) = \pi_{not\ fraud\ type}(x_1 \le x < x_2) \\
& \pi_{fraud\ type}(x_2 \le x < x_3) = \pi_{not\ fraud\ type}(x_2 \le x < x_3) \\
& \pi_{fraud\ type}(x \ge x_3) = \pi_{not\ fraud\ type}(x \ge x_3) \\
H_a\colon\ & \textit{At least one of the above equalities is not true}
\end{aligned}
\tag{35}
$$

where x is a predictor variable under study, π are the proportions in the corresponding class (defined below) and typically $x_{1,2,3}$ are the 1st, 2nd and 3rd quartile.[8] The χ^2-homogeneity test for fraud is done as follows using the frequency table

Table 6 The frequency table of fraud with the quartile ranges defined using fraud class

	Fraud	Non-fraud	Total
$x < x_1$	$n_{f,1}$	$n_{nf,1}$	$N_1 = n_{f,1} + n_{nf,1}$
$x_1 \leq x < x_2$	$n_{f,2}$	$n_{nf,2}$	$N_2 = n_{f,2} + n_{nf,2}$
$x_2 \leq x < x_3$	$n_{f,3}$	$n_{nf,3}$	$N_3 = n_{f,3} + n_{nf,3}$
$x \geq x_3$	$n_{f,4}$	$n_{nf,4}$	$N_4 = n_{f,4} + n_{nf,4}$
Total	$N_f = \sum_{i=1}^{4} n_{f,i}$	$N_{nf} = \sum_{i=1}^{4} n_{nf,i}$	$N = N_f + N_{nf} = N_1 + N_2 + N_3 + N_4$

The χ^2-statistic is calculated as follows

$$\chi^2 = \sum_{i=1}^{4} \sum_{j=f,nf} \frac{\left(n_{j,i} - E_{j,i}\right)^2}{E_{j,i}}$$
$$E_{j,i} = \frac{N_i * N_j}{N} \tag{36}$$
$$dof = (4-1) * (2-1) = 3$$
$$\pi_j(i) = \frac{n_{j,i}}{N_j}, \qquad i = 1,2,3,4, \;\; j = f, nf$$

The test requires that all the observed, $n_{j,I}$, and expected, $E_{j,I}$, values are at least 5. The test for fraud types is run similarly, just replace fraud above with the corresponding fraud type. The results of the tests and the summary of p-values for each continuous predictor variable are presented in the next table. The complete table of p-values is in the Appendix.

[8] Sometimes one of the quartiles have to be replaced by some other value in order to have at least 5 observations and expected number of observations in all the classes.

Table 7 Frequency table of p-values of the continuous predictor variables under the $\chi 2$ homogeneity test, if Bonferroni correction is taken into account the limit of significance is 0,05 / 9 = 0,0056

		Frequency of fraud types		
Predictor	Name of the predictor variable	p-value > 0,05	0,0056 < p-value ≤ 0,05	p-value < 0,0056
1	Accounts receivable	1	---	8
2	Accounts receivable to sales	1	1	7
3	Accounts receivable to total assets	1	---	8
4	Allowance of doubtful accounts (AFDA)	1	1	7
5	AFDA to accounts receivable	8	---	1
6	AFDA to net sales	1	4	4
7	Altman Z-score	1	1	7
9	Current minus prior year inventory to sales	8	1	---
10	Days in receivables index	4	---	5
11	Debt to equity	2	3	4
14	Fixed assets to total assets	4	2	3
15	Four year geometric sales growth rate	7	2	---
16	Gross margin	1	5	3
17	Holding period return in the violation period	9	---	---
18	Industry ROE minus firm ROE	9	---	---
19	Inventory to sales	4	3	2
20	Net sales	1	---	8
22	Prior year ROA to total assets	3	3	3
23	Property, plant and equipment	3	4	2

	to total assets			
24	Sales to total assets	3	4	2
26	Times interest earned	1	3	5
27	Total accruals to total assets	7	2	---
28	Total debt to total assets	6	2	1
29	Total discretionary accruals	4	---	5
34	Value of Issued Securities to Market Value	2	1	6
35	Unexpected Employee Productivity	9	---	---

The results of the test show that predictor variables 17, 18 and 35 (holding period return in the violation period, industry return on equity minus firm return on equity and unexpected employee productivity) do not have any p-value below 0,05. The fact that unexpected employee productivity produces this result is surprising, because it is one of the variables that was found to be among selected variables with logistic regression by Perols (2011). This does not mean that this is necessarily a contradiction with Perols. It may well be that unexpected employee productivity needs other variables to show difference between fraud and non-fraud. On the other hand predictor variables 1, 2, 3, 4, 7, 20 and 34 (accounts receivable, accounts receivable to sales, accounts receivable to total assets, allowance of doubtful accounts, Altman Z-score, net sales and value of issued securities to market value) have at least 6 of the tests with p-values below 0,05 / 9, where the division is based on making 9 different tests, the so called Bonferroni correction. These are therefore expected to show variation in the corresponding predictor variable distribution for the fraud (type) and non-fraud (type) classes. Accounts receivable, allowance of doubtful accounts and value of issued securities to market value were among the variables that Perols found for logistic regression.

4.1.2 Fraud predictor 25: auditor turnover

Auditor turnover variable has only 4 different values since it counts the number of auditors changing in the past 3 years: 0, 1, 2 and 3.

Table 8 Auditor turnover descriptive statistics in the past 3 years

Fraud predictor 25 Auditor turnover	Frequency	Fraud	Non-fraud	% of fraud	% of non-fraud	%-difference
0	46185	261	45924	75,21	77,98	-2,77
1	11007	68	10939	19,60	18,57	1,03
2	1756	16	1740	4,61	2,95	1,66
3	291	2	289	0,58	0,49	0,09
Total	59239	347	58892	100,00	99,99	

As an impression the auditor turnover variable (fraud predictor 25) seems to have a slightly larger portion turnovers = 0 in the non-fraud category than in the fraud category, whereas with turnover > 0 the fraction is slightly higher in the fraud category. It is also to the direction one would expect, if one wants to have the positive coefficient for the logistic regression. Namely the percentage of fraud should be higher than the percentage of non-fraud when auditor turnover is greater than 0, and the other way around when it is 0. However, the differences are not large. This is one of the variables, which Perols (2011) found significant, so maybe the tilt in the distribution is enough. This can be tested with the χ^2 homogeneity test, equation (34). Because fraud class has only 2 cases with auditor turnover = 3, it has to be combined with auditor turnover = 2 class to get at least 5 cases for the each combination of auditor turnover and fraud type. Furthermore expected number has to be at least 5 and with some fraud types auditor turnover = 1, 2, 3 had to be combined into one class. The results are in the following table

Table 9 Auditor turnover p-values for $\chi 2$-statistic

Fraud type	6	7	11	12	14	20	44	other	Fraud
P-value	0,887	*0,021*	0,739	0,939	0,101	0,762	0,194	0,550	0,169

Only the fraud type 7, Expense (payroll, SGA, other) recording issues, has a p-value below 0,05, the rest have a non-significant p-value. Since multiple hypotheses are tested here, it is possible that is happening just by chance. Applying Bonferroni correction to the significance level 0,05 / 9 = 0,0056 and now even the fraud type 7 is not significant. This is another surprising result because auditor turnover is also one of the predictor variables that got chosen in the study of Perols (2011) for logistic regression. Similarly to unexpected employee productivity this does not mean that one should discard auditor turnover, but rather that it does not differentiate between fraud (type) and non-fraud (type) by itself but could do it in combination with other variables.

4.1.3 Binary predictor variables

The rest of the predictor variables are of binary type. The frequencies are in the next table

Table 10 Binary variables for fraud

	All		Fraud		Non-fraud		Fraud	Non-fraud	
Fraud predictor	0	1	0	1	0	1	% of 1s	% of 1s	%-diff.
8	15777	43462	49	298	15728	43164	85,9	73,3	12,6
12	46833	12406	301	46	46532	12360	13,3	21,0	-7,7
13	38583	20656	230	117	38353	20539	33,7	34,9	-1,2
21	55036	4203	323	24	54713	4179	6,9	7,1	-0,2

30	34032	25207	207	140	33825	25067	40,3	42,6	-2,3
31	48566	10673	285	62	48281	10611	17,9	18,0	-0,1
32	2192	57047	1	346	2191	56701	99,7	96,3	3,4
33	38378	20861	261	86	38117	20775	24,8	35,3	-10,5

The binary variables defined seem to differentiate between fraud and non-fraud with varying degrees. Fraud predictors 8 and 33 have over 10% difference in their distributions in the classes of fraud and non-fraud respectively, fraud predictor 12 is between 5-10% and the rest are below 5% difference. According to the predicted signs of the coefficients in section 2 all the predictor variables should have positive percentage difference. However, only fraud predictors 8 (big 4 auditor) and 32 (whether new securities were issued) have positive difference. The dummy variables are built in such a way that one would expect fraud cases correspond to having a larger proportion of 1 and non-fraud cases. This does not seem to be the case for most of the dummy variables. Especially fraud predictor 33, whether standard industry classification code is between 3000-3999 or not, has over 10% less of values 1 for fraud cases than for non-fraud cases. The usage of this variable was based on empirical findings by Lee et al. (1999). It seems that since then the situation has changed completely. But the 10% difference still means that this can be quite a good variable to use. The role it plays just has to be reversed. The same argument can be used for other binary variables.

The differences were tested with the above mentioned χ^2-statistic. With fraud predictor 32 the statistic is unreliable since the statistic requires at least 5 cases in the observed and expected classes and there is only one case with fraud and fraud predictor 32 = 0. The p-values associated with the statistic are in the below table for fraud and all the fraud types

Table 11 P-values of the χ2-statistic for the binary predictor variables, significant values with Bonferroni correction applied are bolded, significant values without Bonferroni correction are italized

	Fraud type								
Fraud predictor	6	7	11	12	14	20	44	other	fraud
8	**5,0E-4**	0,956	**4,3E-7**	**7,2E-4**	**1,1E-5**	**0,002**	**1,5E-5**	**1,0E-6**	**1,7E-7**
12	0,275	0,086	0,188	0,954	0,097	0,419	0,764	*0,049*	**5,3E-4**
13	0,961	0,941	0,519	0,830	0,054	0,478	0,354	0,409	0,693
21	0,942	0,213	0,469	*0,032*	0,901	0,946	0,052	0,923	0,980
30	0,316	0,998	0,638	0,629	0,893	0,465	0,465	0,579	0,436
31	0,249	0,062	0,231	0,486	0,394	0,700	0,144	0,234	0,998
32	0,106	0,445	*0,043*	0,249	0,267	0,103	0,103	0,103	**0,001**
33	**2,1E-5**	0,066	**1,3E-4**	0,121	*0,028*	0,931	0,063	*0,012*	**5,7E-5**

The bolded values are significant even with the Bonferroni correction. The p-values for fraud predictor 32 are unreliable because there is class with less than 5 observable cases. Also fraud type 7 and predictor variable 21 the result is unreliable because there is class with less than 5 observable cases. Big 4 auditor, fraud predictor 8, has 8 out of 9 tests with significant p-values. Fraud predictor 33, is the standard industry classification code between 3000 and 3999 or not, has 3 significant p-values with Bonferroni correction. This is also one of the variables that was found by Perols (2011) for logistic regression.

4.1.4 Summary of descriptive statistics

It was found that the variables accounts receivable, accounts receivable to sales, accounts receivable to total assets, allowance of doubtful accounts, Altman Z-score, net sales, value of issued securities to market value and big 4 auditor are significant including Bonferroni correction with at least 6 out 9 χ^2-homogeneity tests. Accounts receivable, allowance of

doubtful accounts, value of issued securities to market value and big 4 auditor were among the variables that were also found by Perols (2011). The other variables Perols found were auditor turnover, total discretionary accruals, whether meeting or beating a forecast, inventory to sales and unexpected employee productivity. Of these latter whether meeting or beating a forecast used data that is not available in Compustat, so it is not included here at all. Auditor turnover and unexpected employee productivity had p-values in the χ^2-homogeneity tests such that none of the 9 tests were significant when including the Bonferroni correction. Inventory to sales have 2 tests with significant p-values including Bonferroni correction, 3 significant when Bonferroni correction is not included and 4 not significant in any case. Total discretionary accruals have 5 tests significant with Bonferroni correction and 4 tests insignificant.

4.2 Model results for fraud and types separately

All the models are fitted by minimizing the objective function with the regularization term included. The dependent variable is fraud or one of the fraud value variables. Furthermore 5-fold cross-validation is being used, so there are actually 5 models fitted per fraud type case. Training set is first balanced by replication and then variables chosen by recursive feature elimination while regularization parameter is set to 1. Feature elimination uses 5-fold cross-validation inside the balanced training set. After this the model is fitted with the variables chosen in the previous step and fitting uses another 5-fold cross-validation in the balanced training set so that an optimal value of regularization parameter can be set, too. The continuous variables are standardized before fitting so that there are no different scales between variables. Test set variables are transformed with the corresponding training set means and standard deviations before prediction.

4.2.1 Predictor variables from model fitting

When fitting the models, there are actually 5 different models fitted, 1 per fold. On top of that the training set uses itself 5-fold cross validation in recursive feature elimination to find the predictor variables that give the best results. Then another 5-fold cross validation is used to find the best regularization parameter value λ with the predictor variables that were found in the previous step. Once these steps are done the model is fitted with chosen predictor variables and regularization parameter λ over the whole training set. Different folds tend to choose different variables. In table 12 the predictor variables that were chosen in at least 4 folds are reported

Table 12 Predictor variables chosen in at least 4 folds in each fraud category

Fraud type	Category title	Number of folds	Predictor variables chosen	Number of variables
Fraud	Fraud	5	1, 2, 3, 4, 7, 8, 10, 12, 14, 16, 20, 22, 23, 24, 25, 27, 32, 33, 34	19
		4	6, 21, 26, 28, 31, 35	6
6	Revenue recognition issues	5	3, 6, 7, 8, 10, 12, 14, 22, 27, 32, 33	11
		4	2, 4, 5, 18, 20, 24, 26, 29, 30, 31, 34, 35	12
7	Expense (payroll, SGA, other) recording issues	5	3, 9, 10, 15, 21, 22, 31, 32, 34	9
		4	1, 4, 5, 6, 11, 20, 23, 25, 30, 33	10
11	Foreign, related party, affiliated, or subsidiary issues	5	1, 2, 3, 5, 6, 8, 14, 15, 16, 20, 22, 24, 25, 32, 33	15
		4	12, 19, 27, 34	4
12	Liabilities, payables,	5	1, 2, 3, 4, 6, 8, 12, 14, 20, 23,	14

	reserves and accrual estimate failures		32, 33, 34, 35	
		4	7, 16, 26, 29	4
14	Accounts/loans receivable, investments & cash issues	5	3, 8, 21, 22, 24, 25, 26, 33	8
		4	1, 13, 16, 20, 32, 34	6
20	Inventory, vendor and/or cost of sales issues	5	1, 6, 8, 14, 16, 20, 22, 24, 26, 27, 29, 32, 34	13
		4	5, 30	2
44	Foreign, subsidiary only issues (subcategory)	5	1, 2, 3, 8, 9, 15, 20, 22, 24, 28, 30, 32, 33, 34	14
		4	5, 6, 7, 13, 14, 16, 21, 27, 31	9
Other	All the other categories that are not 6, 7, 11, 12, 14, 20 or 44	5	1, 2, 3, 4, 7, 8, 10, 12, 14, 23, 32, 34	12
		4	18, 21, 24, 25, 27, 31, 33, 35	8

Comparing the results for fraud with the results of Perols (2011), here there are 19 predictor variables chosen in all folds, Perols had 9. There are 6 more variables that were chosen in 4 out of 5 folds, but is missing from one fold. Perols made the variable selection over the whole dataset before the split to training and test sets, so he had only one set of variables to work with. The difference may well come from the fact that he had a smaller sample 51 fraud firms and 15934 non-fraud firms. Also the time span of Perols' study is shorter, the years spanned are from 4th quarter of 1998 to 4th quarter of 2005. Here the sample size is 347 fraud cases and 58944 non-fraud cases spanning years 1995 – 2016 in order to have enough data for the different fraud types. It is possible that there are also time effects here, which require more predictor variables to be used. The nature of fraud may have changed in time.

In the following (P) means that Perols observed that variable, too. The 19 predictor variables chosen in all 5 folds are accounts receivable (P), accounts receivable to total assets, (AFDA) allowance for doubtful accounts (P), Altman Z-score, big 4 auditor (P), days in receivables index, demand for financing (ex ante), fixed assets to total assets, gross margin, net sales, prior

year ROA to total assets, property, plant and equipment to total assets, sales to total assets, the number of auditor turnovers (P), total accruals to total assets, whether new securities were issued, whether Standard Industry Classification Code larger than 2999 and smaller than 4000 and value of issued securities to market value (P). The 6 variables that were chosen in 4 out of 5 folds are AFDA to net sales, positive accruals dummy, times interest earned, total debt to total assets, whether gross margin percent > 1,1 * of last year's gross margin percent and unexpected employee productivity (P). In addition Perols observed total discretionary accruals, whether meeting or beating a forecast (requires data not available in Compustat) and inventory to sales. Even if some of the variables were not selected with fraud model, they got selected in fraud type models. Total discretionary accruals got selected in all folds in fraud type 20 (inventory, vendor and/or cost of sales issues), and in 4 out of 5 folds in fraud type 6 (revenue recognition issues) and fraud type 12 (liabilities, payables, reserves and accrual estimate failures). Inventory to sales got selected in 4 out of 5 folds in fraud type 11 (foreign, related party, affiliated, or subsidiary issues). Perols et al. (2017) did not give a list of variables, which the models had chosen, so no comparison can be made with their results.

Overall all the models contain more variables than what Perols had in his study for logistic regression, even the fraud type models. The fraud type models have fewer variables in general in them than in the fraud model. This is as expected. Many forms of fraud may have competing effects that may cancel each other to some degree, so more variables are needed to cover this. Fraud type models look for the effects on the single type, so having fewer variables for them is natural. However, since the fraud type models still contain more variables than what Perols had, there is clearly room for improvement. Another reason for Perols having fewer variables is that he used forward selection to choose the variables. Here backward selection, recursively removing variables one by one until no improvement is observed, was used. It is natural to have more variables remaining with this method.

In the following table 13 the information of table 12 is re-written in terms of the predictor variables. In addition the signs of the coefficients have been produced, too.

Table 13 Fraud types that the predictor variables was chosen in, (m+, n-) means m folds with + and n folds with -, if no numbers, coefficient on all folds of the sign given, fraud types in bold have the sign of the predictor variable coefficient opposite of the predicted sign

Predictor	Name	Number of folds	Fraud types chosen in (sign)	Predicted sign
1	Accounts receivable	5	fraud (+), 11 (+), 12 (+), 20 (+), 44 (+), other (+)	+
		4	7 (+), 14 (+)	
2	Accounts receivable to sales	5	fraud (+), 11 (4-, 1+), 12 (+), 44 (-), other (+)	
		4	6 (-)	
3	Accounts receivable to total assets	5	fraud (+), 6 (+), 7 (+), 11 (+), 12 (+), 14 (+), 44 (+), other (+)	
4	Allowance of doubtful accounts (AFDA)	5	**fraud (+), 12 (+),** other (-)	-
		4	**6 (+),** 7 (-)	
5	AFDA to accounts receivable	5	11 (-)	-
		4	6 (-), 7 (-), 20 (-), 44 (-)	
6	AFDA to net sales	5	6 (+), 11 (+), 12 (-), 20 (+)	
		4	fraud (3+, 1-), 7 (+), 44 (+)	
7	Altman Z-score	5	fraud (-), 6 (-), other (-)	-
		4	12 (-), 44 (-)	
8	Big 4 auditor	5	fraud (+), 6 (+), 11 (+), 12 (+), 14 (+), 20 (+), 44 (+), other (+)	+
9	Current minus prior year inventory to sales	5	7 (+), 44 (+)	
10	Days in receivables index	5	**fraud (-), 6 (-), 7 (-), other (-)**	+
11	Debt to equity	4	7 (+)	+
12	Demand for financing	5	**fraud (-), 6 (-), 12 (-), other (-)**	+

	(ex ante)	4	**11 (-)**	
13	Declining cash sales dummy	4	**14 (-), 44 (-)**	+
14	Fixed assets to total assets	5	fraud (-), 6 (-), 11 (-), 12 (-), 20 (-), other (-)	
		4	44 (-)	
15	Four year geometric sales growth rate	5	**7 (-), 11 (-), 44 (-)**	+
16	Gross margin	5	fraud (-), 11 (-), 20 (-)	
		4	12 (-), 14 (-), 44 (-)	
17	Holding period return in the violation period	---	---	+
18	Industry ROE minus firm ROE	4	6 (3+, 1-), other (+)	
19	Inventory to sales	4	11 (-)	
20	Net sales	5	**fraud (-), 11 (-), 12 (-), 20 (-), 44 (-)**	+
		4	6 (3+, 1-), **7 (-), 14 (-)**	
21	Positive accruals dummy	5	**7 (-), 14 (-)**	+
		4	**fraud (-), 44 (3-, 1+), other (-)**	
22	Prior year ROA to total assets	5	fraud (+), 6 (+), 7 (+), 11 (+), 14 (+), 20 (+), 44 (+)	+
23	Property, plant and equipment to total assets	5	fraud (+), 12 (+), other (+)	
		4	7 (+)	
24	Sales to total assets	5	fraud (+), 11 (-), 14 (-), 20 (+), 44 (-)	
		4	other (+)	
25	The number of auditor turnovers	5	fraud (+), 11 (+), **14 (-)**	+
		4	7 (+), other (+)	

26	Times interest earned	5	14 (-), 20 (-)	
		4	fraud (-), 6 (-), 12 (-)	
27	Total accruals to total assets	5	fraud (+), 6 (+), 20 (+)	+
		4	11 (+), 44 (+), other (+)	
28	Total debt to total assets	5	**44 (-)**	+
		4	fraud (+)	
29	Total discretionary accruals	5	**20 (-)**	+
		4	6 (+), 12 (+)	
30	Whether accounts receivable > 1,1 * of last year's accounts receivable	5	**44 (-)**	+
		4	6 (+), 7 (+), **20 (3-, 1+)**	
31	Whether gross margin percent > 1,1 * of last year's gross margin percent	5	7 (+)	+
		4	fraud (+), **6 (-), 44 (-), other (-)**	
32	Whether new securities were issued	5	fraud (+), 6 (+), **7 (4-, 1+),** 11 (+), **12 (4-, 1+)**, 20 (+), 44 (+), other (+)	+
		4	14 (+)	
33	Whether Standard Industry Classification Code larger than 2999 and smaller than 4000	5	**fraud (-), 6 (-), 11 (-), 12 (-), 14 (-), 44 (-)**	+
		4	**7 (-), other (-)**	
34	Value of Issued Securities to Market Value	5	fraud (+), 7 (+), 12 (+), 20 (+), 44 (+), other (+)	+
		4	6 (+), 11 (+), 14 (+)	
35	Unexpected Employee Productivity	5	12 (-)	
		4	fraud (-), 6 (-), other (-)	

Based on the descriptive statistics accounts receivable, accounts receivable to sales, accounts receivable to total assets, allowance of doubtful accounts, Altman Z-score, net sales, value of issued securities to market value and big 4 auditor were expected to be found among the predictor variables. All these had at least 6 out 9 χ^2-homogeneity tests with p-values below the significance level with Bonferroni correction (0,05 / 9). Of these accounts receivable is found in 6 (5 folds) and 2 (4 folds) models, accounts receivable to sales in 5 (5 folds) and 1 (4 folds) models, accounts receivable to total assets in 8 (5 folds) models, allowance for doubtful accounts in 3 (5 folds) and 2 (4 folds) models, Altman Z-score in 3 (5 folds) and 2 (4 folds) models, big 4 auditor in 8 (5 folds) models, net sales in 5 (5 folds) and 3 (4 folds) models, and value of issued securities to market value in 6 (5 folds) and 3 (4 folds) models. In some cases the variables chosen are the same that had significant p-values in the χ^2-homogeneity tests, but in some cases it is not. For example accounts receivable to total assets and big 4 auditor appear in the same fraud types as tests of differences indicated. But for example accounts receivable does not appear in at least 4 folds in fraud type 6 (revenue recognition issues), but instead appears in fraud type 7 (expense (payroll, SGA, other) recording issues) which had non-significant p-value in tests of difference, see table 38 in Appendix. However, majority of the variables appearing correspond to the results of tests of differences.

In table 13 the signs of coefficients have also been produced. The signs differ from the predicted ones in Section 2 for 14 out 34 variables (1 was not chosen at all). However, with 8 of the 14 predictor variables the sign issue is with some of the fraud types, not all. With 6 predictor variables the sign issue happens with all the fraud types. With most of the variables with the wrong sign of the coefficient the descriptive statistics supports the sign. However, with net sales the situation is different. Net sales has descriptive statistics clearly supporting the positive sign for the coefficient, the mean value for fraud cases is larger than for non-fraud cases and all percentiles, too. So the whole distribution for fraud cases supports higher net sales than for non-fraud cases. Net sales appear in many variables, in some variables in the nominator and in other variables in the denominator. So the net effect of net sales is a bit ambiguous.

The tables of coefficients of the fitted models are not produced here, because there would be 5 of them for each fold times, 9 for each fraud type and fraud itself, each table containing as

many rows as there are variables in the model. Second as explained in the Data and methods section 3.1 the regularization parameter decreases the variance of the coefficients, in practise all of the coefficients tend to be highly significant because of the regularization. Altogether there are 1117 coefficients including intercepts and all the 5 folds. Of these 1103 have p-value below 0,01. The fitting is done on the training set and the variances and significance tests are based on the training set. The coefficients are not fitted on the test set, unlike the performance measures, which are estimated using the test set. Furthermore the performance measures relate directly to the question predicting fraud as well as possible. Therefore the performance measures should be considered as the test of the coefficients, too.

4.2.2 Results with probability threshold 0,5

Here the probability threshold of the model is 0,5 meaning that the predictions are formed as follows

$$\hat{y} = \begin{cases} 1, & if\ p(y = 1|\boldsymbol{x}) > 0{,}5 \\ 0, & if\ p(y = 1|\boldsymbol{x}) \leq 0{,}5 \end{cases} \tag{37}$$

Furthermore the test sets over the folds are pooled together and then the measures are calculated. This is micro averaging. Another way is to calculate the performance measures separately for each fold and then average the results, called macro averaging. The micro averaging results were chosen here because the macro average does not make sense in all cases to follow, like with ROC curves.

Table 14 Average performance measures when the threshold probability p_0=0,5, standard errors are in parentheses

Fraud type	Accuracy (s.e.)	Precision (s.e.)	Sensitivity (s.e.)	Specificity (s.e.)
Fraud	0,6554 (0,0020)	0,0107 (7,18E-4)	0,6311 (0,0259)	0,6556 (0,0020)
6	0,6946 (0,0019)	0,0050 (5,25E-4)	0,6642 (0,0403)	0,6947 (0,0019)
7	0,7346 (0,0018)	0,0027 (4,16E-4)	0,5811 (0,0574)	0,7347 (0,0018)
11	0,7025 (0,0019)	0,0051 (5,38E-4)	0,6894 (0,0403)	0,7025 (0,0019)
12	0,7404 (0,0018)	0,0043 (5,26E-4)	0,6667 (0,0474)	0,7405 (0,0018)
14	0,6723 (0,0019)	0,0032 (4,04E-4)	0,6458 (0,0488)	0,6724 (0,0019)
20	0,6385 (0,0020)	0,0027 (3,55E-4)	0,6170 (0,0501)	0,6385 (0,0020)
44	0,7163 (0,0019)	0,0039 (4,81E-4)	0,7021 (0,0472)	0,7163 (0,0019)
Other	0,6812 (0,0019)	0,0051 (5,16E-4)	0,6957 (0,0392)	0,6811 (0,0019)

Looking at the table one can see that the accuracy with the fraud types tends to be higher than the accuracy of the fraud class itself. The sole exception is the fraud type 20 which has approximately 1,7% lower accuracy than fraud. The best accuracy is with fraud type 12 which 74% accuracy. The reason may well be that there are more non-fraud cases with the types than with fraud class itself. Precision on the other hand is better with fraud than with the types. Reason may again be the lower number cases with the types. With sensitivity the picture is more complicated, fraud is roughly in the middle with 63% whereas the types vary from 58%

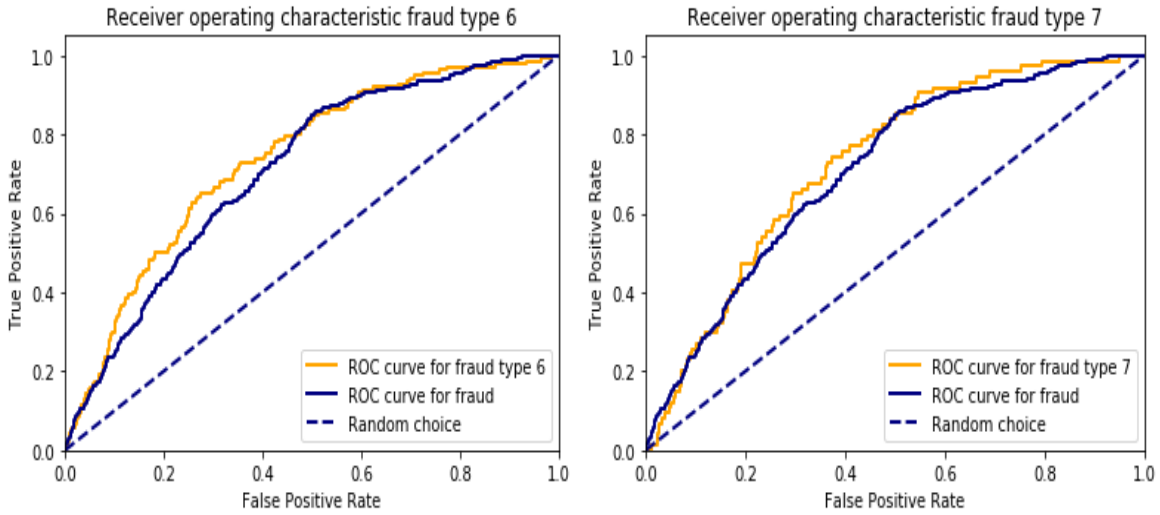
with type 7 to 70% with type 44. Fraud class tends to have lower specificity than the types. Fraud specificity is 66% while the lowest value is 64% with type 20 and the highest 74% with type 12.

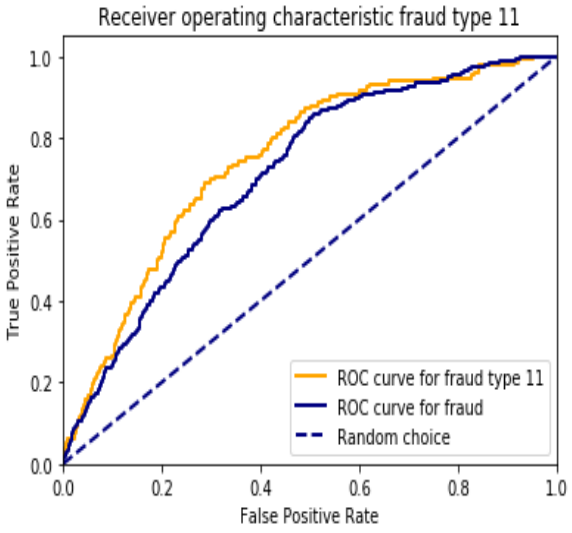
If one were to compare the results of micro or macro averaging, one sees that the performance measures itself do not change much if at all. When there is a change, it is smaller than the corresponding standard errors.

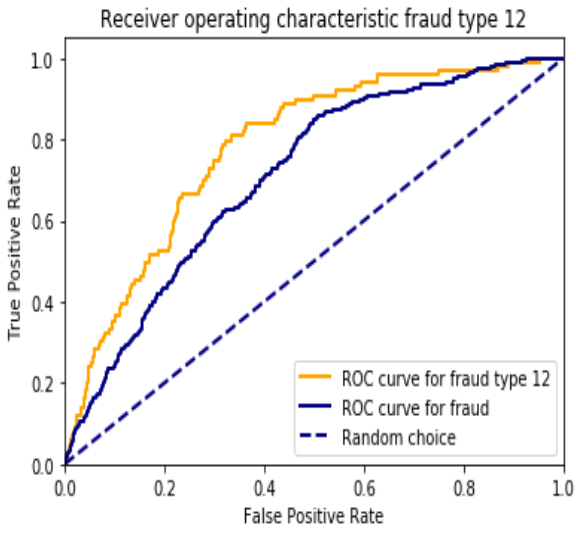
4.2.3 Receiving operating curves

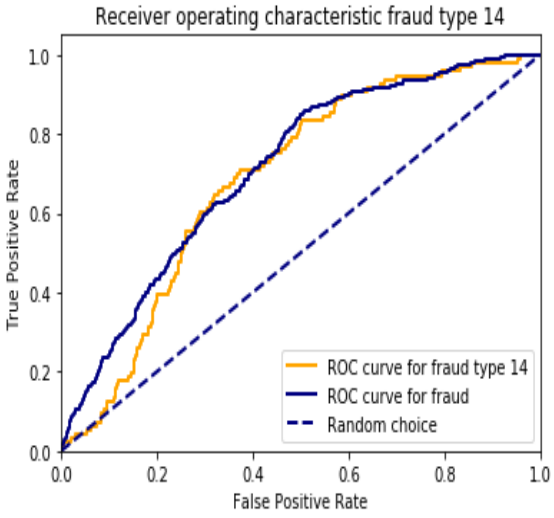
Receiving operating characteristic curves (ROC) give the performance of a classifier when the threshold probability varies from 0 to 1. The results here are based on pooling the test results over the folds back together to form the original test set with the probabilities for each case.

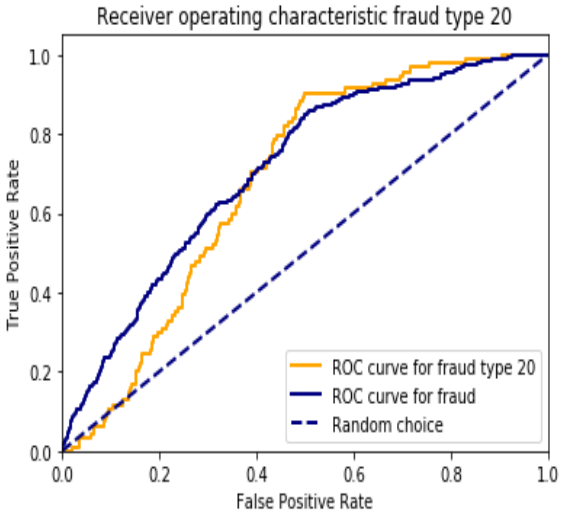
Figure 3 Receiving operating characteristic curves for fraud and fraud types

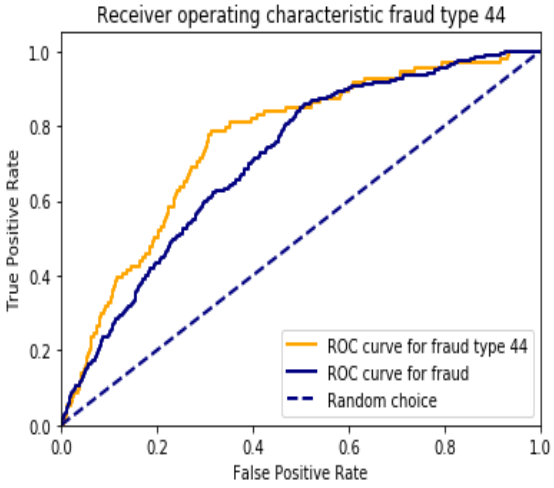


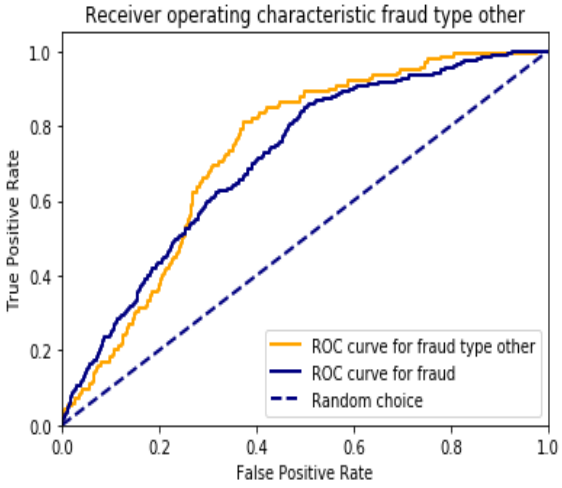

Receiver operating characteristic fraud type 11
True Positive Rate
False Positive Rate
ROC curve for fraud type 11
ROC curve for fraud
Random choice


Receiver operating characteristic fraud type 12
True Positive Rate
False Positive Rate
ROC curve for fraud type 12
ROC curve for fraud
Random choice


Receiver operating characteristic fraud type 14
True Positive Rate
False Positive Rate
ROC curve for fraud type 14
ROC curve for fraud
Random choice


Receiver operating characteristic fraud type 20
True Positive Rate
False Positive Rate
ROC curve for fraud type 20
ROC curve for fraud
Random choice


Receiver operating characteristic fraud type 44
True Positive Rate
False Positive Rate
ROC curve for fraud type 44
ROC curve for fraud
Random choice


Receiver operating characteristic fraud type other
True Positive Rate
False Positive Rate
ROC curve for fraud type other
ROC curve for fraud
Random choice

In the figures above the fraud type ROC curve is drawn along with fraud and random choice ROC curves for comparison. Typically a classifier is better than another if its ROC curve is to the left and above of the ROC curve of the other classifier. All the fraud types and fraud definitely perform better than random choice. With fraud types 14 and 20 there are ranges in the small false positive rate (FPR) region where they seem to perform only at level of random choice, but over the majority of the range they perform better than random choice. Comparison between fraud and fraud types is much more involved. There are ranges where the fraud types perform better than fraud class and vice versa. Fraud types 6, 7, 11, 12 and 44 have a range where they are better than fraud and outside perform similarly to fraud, so these can be considered to perform better than fraud based on the ROC curves. Fraud types other has a range, where it performs better than fraud, and a range, where it performs worse than fraud. Altogether one can say that fraud other performs equally well as fraud. Fraud types 14 and 20 have a large range, where they perform worse than fraud, and outside it at the same level as fraud. Based on this it can be said that fraud types 14 and 20 perform worse than fraud.

The differences in the performance in different regions can be taken into account by calculating the area under the ROC curve (AUC), area between horizontal axis and the curve. A perfectly performing classifier would have a ROC curve connecting the points (1, 1), (0, 1) and (0, 0) and have an area 1. Random choice has an area 0,5. The AUCs for the pooled test set is in the table below

Table 15 Areas under the ROC curves

Fraud type	6	7	11	12	14	20	44	other	fraud
AUC	0,7367	0,7277	0,7466	0,7802	0,6871	0,6809	0,7548	0,7287	0,7103
AUC s.e.	0,0245	0,0336	0,0248	0,0276	0,0302	0,0306	0,0291	0,0246	0,0157

The AUC results support the conclusion obtained through studying ROC curves. However, the standard errors on them are such that fraud AUC would be contained within 2 standard error confidence interval of fraud type AUCs, with the exception of fraud type 12.

Although some fraud type classifiers seem to perform better than fraud itself, one cannot infer from this that they are better. First of all the classifiers do not measure the same thing. Second the probability threshold to be used has to be determined somehow and then compare the results.

4.2.3 Expected relative costs

Here the probability threshold p_0 is chosen in such a way that the expected relative costs are minimized

$$ERC = C_{FN}\frac{FN}{TP+FP}\,P(fraud) + C_{FP}\frac{FP}{FN+TN}\,\big(1-P(fraud)\big) \qquad (38)$$

with several different values for relative costs *r* and prior fraud probability *P*(*fraud*). The amounts of *TP*, *FP*, *FN* and *TN* change when the probability threshold changes. The tables below contains the results for fraud and the corresponding performance measures

Table 16 Minimum expected relative costs and the threshold probabilities when CFP = 1 and CFN is set according to the table below

	ERC				**p_0**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0047	0,015	0,021	0,046	0,933	0,894	0,880	0,834
0,01	0,015	0,046	0,065	0,142	0,894	0,828	0,810	0,734
0,1	0,044	0,135	0,187	0,387	0,828	0,717	0,677	0,605

The expected relative costs increase when prior fraud probability and the difference between the costs of false positives and false negatives increase. One can look at that when prior fraud probability goes from 0,001 to 0,1 the expected relative costs become approximately 8-10 times the starting value. The same can be observed when the relative cost ratio changes from 1 to 0,01, the expected relative cost is approximately 8-10 times the starting value. At the same time the threshold probability decreases. Essentially the more there is fraud to begin with, the larger the prior fraud probability, and the more making the mistake of classifying fraud as non-fraud costs compared to the mistake of classifying non-fraud as fraud, the more is the expected relative cost. In Table 5 of Perols (2011) the expected relative costs were between 0,0026 and 0,91, with mean 0,2916 and standard deviation 0,2367. It is unclear over what range has Perols computed his values. It seems like he has aggregated over the prior fraud probabilities and cost ratios. However, the values here are compatible to values produced by Perols.

Standard errors are not produced here for the expected relative costs because only one minimum value is produced by the test set. Some statistics can be calculated if one takes the averages over the 5 folds, as in the table 17 below.

Table 17 Expected relative costs averaged over 5 folds, standard errors in parentheses

	C_{FP} / C_{FN}		
P(fraud)	**1**	**0,1**	**0,01**
0,001	0,0047 (7,8E-5)	0,015 (2,3E-4)	0,046 (0,0017)
0,01	0,015 (2,2E-4)	0,046 (0,0016)	0,141 (0,0048)
0,1	0,044 (0,0016)	0,134 (0,0041)	0,382 (0,0203)

There is no practical difference in the averages over the folds compared to the minimizing over the whole test set. If one wanted better statistics, one would need to repeat the process multiple times. For example Perols (2011) used the 10-fold cross validation, which was

repeated 10 times. Here 5 folds were used, because the time it took to run 5 folds was a couple of minute, but with 10 folds the time increase was not double but much more than that.

Next we look at the other performance measures with the threshold probabilities obtained through minimizing the expected relative costs. Here the comparison is done between fraud and fraud type 6 only. The other types show similar behaviour

Table 18 Accuracy of fraud and fraud type 6 with threshold probability minimizing expected relative cost, standard errors in parentheses

	Fraud				**Fraud type 6**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,9919 (3,7E-4)	0,9870 (4,7E-4)	0,9841 (5,1E-4)	0,9726 (6,7E-4)	0,9910 (3,9E-4)	0,9821 (5,5E-4)	0,9786 (6,0E-4)	0,9625 (7,8E-4)
0,01	0,9870 (4,7E-4)	0,9704 (7,0E-4)	0,9639 (7,7E-4)	0,9290 (1,1E-3)	0,9821 (5,5E-4)	0,9600 (8,1E-4)	0,9530 (8,7E-4)	0,9110 (1,2E-3)
0,1	0,9704 (7,0E-4)	0,9188 (1,1E-3)	0,8865 (1,3E-3)	0,8124 (1,6E-3)	0,9600 (8,1E-4)	0,8998 (1,2E-3)	0,8688 (1,4E-3)	0,8056 (1,6E-3)

The accuracy decreases when the prior probability of fraud increases and when the relative cost, C_{FP}/C_{FN}, decreases. This is connected to the behaviour of the probability threshold and is what one would expect. One should note that the accuracy for fraud is better than the accuracy for fraud type 6. This is in contrast to the results with threshold probability 0,5. Similar results are seen with other fraud types.

Table 19 Precision with threshold probabilities minimizing the expected relative cost for fraud and fraud type 6, standard errors in parentheses

	Fraud				**Fraud type 6**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0347 (0,0153)	0,0204 (0,0067)	0,0194 (0,0056)	0,0217 (0,0040)	0,0075 (0,0043)	0,0085 (0,0030)	0,0070 (0,0025)	0,0076 (0,0019)
0,01	0,0204 (0,0067)	0,0211 (0,0038)	0,0193 (0,0032)	0,0151 (0,0019)	0,0085 (0,0030)	0,0079 (0,0019)	0,0074 (0,0017)	0,0065 (0,0011)
0,1	0,0211 (0,0038)	0,0148 (0,0018)	0,0146 (0,0015)	0,0131 (0,0011)	0,0079 (0,0019)	0,0070 (0,0011)	0,0070 (9,5E-4)	0,0060 (7,2E-4)

The behavior of precision when prior fraud probability and the relative cost are changed. There is no clear cut picture. The behavior between different columns and rows is not the same. The only consistent thing is that the precision for fraud is higher than precision for fraud type 6. All the other fraud types have similar results to fraud type 6.

Table 20 Sensitivity with threshold probabilities minimizing the expected relative cost for fraud and fraud type 6, standard errors in parentheses

	Fraud				**Fraud type 6**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0144 (0,0064)	0,0259 (0,0085)	0,0346 (0,0098)	0,0836 (0,0149)	0,0219 (0,0125)	0,0584 (0,0200)	0,0584 (0,0200)	0,1168 (0,0274)
0,01	0,0259 (0,0085)	0,0893 (0,0153)	0,1037 (0,0164)	0,1729 (0,0203)	0,0584 (0,0200)	0,1314 (0,0289)	0,1460 (0,0302)	0,2482 (0,0369)
0,1	0,0893 (0,0153)	0,1960 (0,0213)	0,2767 (0,0240)	0,4179 (0,0265)	0,1314 (0,0289)	0,2993 (0,0391)	0,3942 (0,0418)	0,5037 (0,0427)

For fraud type 6 the sensitivity is in every case higher than sensitivity for fraud. This does not happen with all the fraud types, some of them have higher sensitivity than fraud just like type 6, but some of them have lower sensitivity than fraud. Sensitivity increases when prior fraud increases and the difference between costs increase. Also the difference between sensitivity of fraud and fraud type 6 tend to be larger when prior fraud is larger and difference between costs is larger.

Table 21 Specificity with threshold probabilities minimizing the expected relative cost for fraud and fraud type 6, standard errors in parentheses

	Fraud				**Fraud type 6**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,9976 (2,0E-4)	0,9926 (3,5E-4)	0,9897 (4,2E-4)	0,9778 (6,1E-4)	0,9933 (3,4E-4)	0,9842 (5,1E-4)	0,9807 (5,7E-4)	0,9645 (7,6E-4)
0,01	0,9926 (3,5E-4)	0,9756 (6,4E-4)	0,9689 (7,2E-4)	0,9335 (0,0010)	0,9842 (5,1E-4)	0,9619 (7,9E-4)	0,9548 (8,5E-4)	0,9126 (0,0012)
0,1	0,9756 (6,4E-4)	0,9231 (0,0011)	0,8901 (0,0013)	0,8147 (0,0016)	0,9619 (7,9E-4)	0,9012 (0,0012)	0,8699 (0,0014)	0,8063 (0,0016)

Specificity is consistently 0,01-0,02 higher for fraud than for fraud type 6. This holds also for other types. The tendency here is that specificity decreases when prior fraud increases and then difference between the two costs increase. Difference in specificity is smaller for smaller prior fraud and small difference between costs.

With the performance measures one should remember that they are obtained for minimum expected relative costs of fraud. In a true comparison one should calculate the minimum expected relative costs for fraud type 6 and then do the comparison. However, since there are troubles with defining the expected relative costs for fraud types themselves, this is not done and for the same reason the expected relative costs for fraud types are not produced. Other possibility is to define the relevant threshold probability for the fraud types. One way to do

this is using the threshold probability coming from minimizing the expected relative cost of combining the fraud types, which is done in the next section.

Overall the expected relative costs of fraud are on the same level with what was found by Perols (2011). The other performance measures vary a lot when the prior fraud probability and ratio of costs are changed. Accuracy and specificity for fraud are slightly better than for fraud type 6, similar behavior holds for other fraud types. Precision is clearly much better for fraud than for fraud type 6, and the same is true for other fraud types. Sensitivity is much more involved. For some types the sensitivity is better than for fraud, but for other types sensitivity of fraud is better. Second the difference between sensitivity of fraud and fraud type 6 changes a lot when prior fraud probability and cost ratio are changed. They are at the same level when cost ratio = 1 and prior fraud probability = 0,1% but the difference can be even 0,12 for other combinations of cost ratio and prior fraud probability.

4.3 Combined model results

In this section some of the fraud types are combined to produce a new predictor of fraud. The method used is a majority vote: if the majority of the fraud types models predict a case to be of that fraud type, the case is predicted to be fraud, otherwise it is non-fraud. Here we would like to have the combination that covers the fraud class as much as possible and to have the fraud type ROC curves to be better than fraud ROC curve. Looking at the results in section 4.2.3 we can see that types 6, 7, 11, 12 and 44 have ROC curves that are better than fraud ROC curve in the region where the expected relative costs of fraud are minimized. A combination of types 6, 11 and 12 is chosen because the union of these covers the fraud class most completely of the 3 type combinations, 253 cases of the 347 total.

4.3.1 Results where threshold probability is 0,5

The results for using the common threshold probability 0,5 for all the fraud types 6, 11 and 12 and then using a vote to determine whether prediction is fraud or not. The same can be accomplished by using the median of the probabilities for types 6, 11 and 12 and testing whether it is larger or smaller than 0,5.

Table 22 Performance measures of voting with the same threshold probability for all types

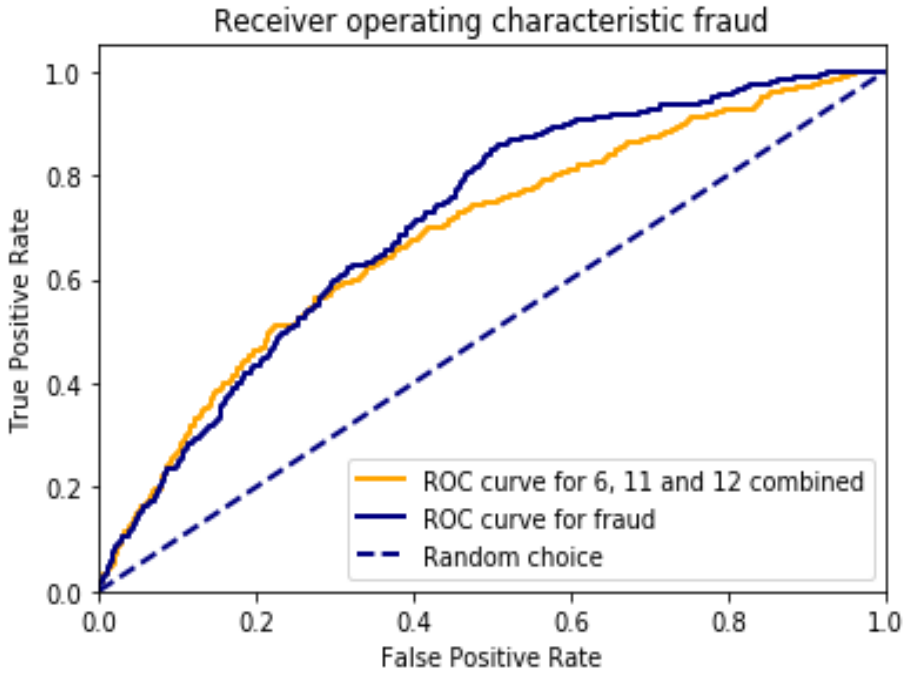
	Accuracy (s.e.)	Precision (s.e.)	Sensitivity (s.e.)	Specificity (s.e.)
Combined 6, 11 and 12	0,7351 (0,0018)	0,0118 (8,6E-4)	0,5331 (0,0268)	0,7363 (0,0018)
Fraud	0,6554 (0,0020)	0,0107 (7,2E-4)	0,6311 (0,0259)	0,6556 (0,0020)

As can be seen all the performance measures are better for the voting result than for fraud itself except sensitivity which gives the proportion of predicting fraud correctly among all the fraud cases. Essentially one gets more false negatives with voting than with predicting fraud directly. The likely reason is the fact that voting relies on just 253 cases of fraud when there are 347 fraud cases altogether.

4.3.2 Receiving operating curves

The ROC curve for the combined case is drawn by using the median probability prediction among the three types 6, 11 and 12

Figure 4 ROC curve for the combination of types 6, 11 and 12



Comparison between ROC curves of fraud and the combined fraud types shows that in the small false/true positive rate region the combined case seems to perform better, but getting higher false positive rates direct fraud prediction produces better results.

Table 23 Area under the ROC curve for the 6, 11 and 12 fraud types combined

	Combined 6, 11 and 12	Fraud
AUC	0,6830	0,7103
AUC s.e.	0,0159	0,0157

Area under the ROC curve is smaller for the combined case than for the fraud itself. Overall this would indicate that fraud is predicted better directly than through the types combined.

4.3.3 Expected relatives costs, common threshold between types

In this case the expected relative costs are calculated while the fraud prediction for the types 6, 11 and 12 are made with the common threshold probability and the minimization of expected relative cost is made using this common threshold probability. The resulting performance measures and expected relative costs are

Table 24 Expected relative costs for the combination of fraud types 6, 11 and 12 and fraud

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0047	0,0149	0,0212	0,0466	0,0047	0,0150	0,0212	0,0464
0,01	0,0149	0,0464	0,0646	0,1411	0,0149	0,0462	0,0649	0,1423
0,1	0,0441	0,1342	0,1842	0,3786	0,0439	0,1353	0,1870	0,3865

Table 25 Change in expected relative costs for the combination of fraud types 6, 11 and 12 compared to fraud

	Change in			
	C_{FP} / C_{FN}			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**
0,001	-0,53 %	-0,47 %	0,28 %	0,43 %
0,01	-0,52 %	0,46 %	-0,55 %	-0,84 %
0,1	0,42 %	-0,81 %	-1,52 %	-2,04 %

For most cases the relative costs of the combined type are lower than for fraud. However, there are 4 case in the table 25 where the costs of the combined type are actually higher than for fraud. The largest decrease of 2,04% is observed for when the prior fraud probability is 0,1

and the cost ratio is 0,01. On the other hand for prior fraud probability 0,01 and cost ratio 0,1 the expected relative cost of the combined type is 0,46% higher than for fraud.

Table 26 Accuracy for the combined 6, 11 and 12 fraud types and fraud with thresholds minimizing the corresponding expected relative costs, the values in bold have p-value below 0,05/12 and values italized have p-values between 0,05/12 and 0,05 under binomial difference test

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,9920 (3,7E-4)	0,9871 (4,6E-4)	0,9839 (5,2E-4)	*0,9701* (7,0E-4)	0,9919 (3,7E-4)	0,9870 (4,7E-4)	0,9841 (5,1E-4)	*0,9726* (6,7E-4)
0,01	0,9871 (4,6E-4)	0,9701 (7,0E-4)	0,9633 (7,7E-4)	**0,9239** (0,0011)	0,9870 (4,7E-4)	0,9704 (7,0E-4)	0,9639 (7,7E-4)	**0,9290** (0,0011)
0,1	0,9701 (7,0E-4)	**0,9239** (0,0011)	*0,8902* (0,0013)	0,8140 (0,0016)	0,9704 (7,0E-4)	**0,9188** (0,0011)	*0,8865* (0,0013)	0,8124 (0,0016)

Compared to fraud the accuracy with combining types 6, 11 and 12 is pretty much the same. Differences are in the 3rd or 4th decimal. Using binomial difference test only two of the accuracies are significantly different i.e. p-value below 0,05/12, in bold font, and two values with p-values between 0,05/12 and 0,05. In 5 out of 12 cases the accuracy is better for fraud than for the combination in the same choices of prior fraud probability and cost ratio as in Tables 24 and 25 for expected relative costs. However, the differences are quite small.

Table 27 Precision for the combined 6, 11 and 12 fraud types and fraud with thresholds minimizing the corresponding expected relative costs

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0426 (0,0170)	0,0251 (0,0075)	0,0175 (0,0052)	0,0196 (0,0036)	0,0347 (0,0153)	0,0204 (0,0067)	0,0194 (0,0056)	0,0217 (0,0040)
0,01	0,0251 (0,0075)	0,0196 (0,0036)	0,0205 (0,0032)	0,0158 (0,0019)	0,0204 (0,0067)	0,0211 (0,0038)	0,0193 (0,0032)	0,0151 (0,0019)
0,1	0,0196 (0,0036)	0,0158 (0,0019)	0,0157 (0,0016)	0,0139 (0,0011)	0,0211 (0,0038)	0,0148 (0,0018)	0,0146 (0,0015)	0,0131 (0,0011)

Precision is better for combined case than for fraud, except in 4 cases, the same choices of parameters as in Tables 24, 25 and 26 for expected relative costs and accuracy. The precision tends to be better for small prior fraud probability and when the costs are equal. There were no significantly different precision values. In 4 out of 12 cases the precision of fraud is better than with the combined type, in the rest the combined model has better precision. This is in contrast to the single fraud types, where the precision of fraud was clearly better. The combination has improved the precision. Here the differences are in 2nd decimal already. However, precision contains only the cases that are predicted to be fraud, both true and false positives, whose amounts vary between 100 - 11000. The accuracy uses the whole set of observations, which contains 59239 cases in total resulting into more significant p-values.

Table 28 Sensitivity for the combined 6, 11 and 12 fraud types and fraud with thresholds minimizing the corresponding expected relative costs

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0173 (0,0070)	0,0317 (0,0094)	0,0317 (0,0094)	0,0836 (0,0149)	0,0144 (0,0064)	0,0259 (0,0085)	0,0346 (0,0098)	0,0836 (0,0149)
0,01	0,0317 (0,0094)	0,0836 (0,0149)	0,1124 (0,0170)	0,1960 (0,0213)	0,0259 (0,0085)	0,0893 (0,0153)	0,1037 (0,0164)	0,1729 (0,0203)
0,1	0,0836 (0,0149)	0,1960 (0,0213)	0,2882 (0,0243)	0,4409 (0,0267)	0,0893 (0,0153)	0,1960 (0,0213)	0,2767 (0,0240)	0,4179 (0,0265)

Combining 6, 11 and 12 fraud types produces better sensitivity values than fraud itself in the same cases as in the previous tables 24-27. In 7 out of 12 cases the combination has better sensitivity than fraud. However, the standard errors are high enough that none of the values is significantly better than other. All p-values under the binomial difference test are above 0,05. Note sensitivity contains only the actual fraud cases, 347 altogether, which explains why the large differences observed are not significant.

Table 29 Specificity for the combined 6, 11 and 12 fraud types and fraud with thresholds minimizing the corresponding expected relative costs, the values in bold have p-value below 0,05/12 and values italized have p-values between 0,05/12 and 0,05 under binomial difference test

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,9977 (2,0E-4)	0,9927 (3,5E-4)	0,9895 (4,2E-4)	**0,9754** (6,4E-4)	0,9976 (2,0E-4)	0,9926 (3,5E-4)	0,9897 (4,2E-4)	**0,9778** (6,1E-4)
0,01	0,9927 (3,5E-4)	0,9754 (6,4E-4)	0,9683 (7,2E-4)	**0,9282** (0,0011)	0,9926 (3,5E-4)	0,9756 (6,4E-4)	0,9689 (7,2E-4)	**0,9335** (0,0010)
0,1	0,9754 (6,4E-4)	**0,9282** (0,0011)	*0,8938* (0,0013)	0,8162 (0,0016)	0,9756 (6,4E-4)	**0,9231** (0,0011)	*0,8901* (0,0013)	0,8147 (0,0016)

Specificity seems to be better in the combined case when cost ratio is 1 or prior fraud probability is 0,1. This time the behaviour does not follow the same pattern as in the previous tables. However, the differences are small, of the same order as the standard errors. Nevertheless there are 3 specificity values that are significantly different with p-values below 0,05/12 and one p-value between 0,05/12 and 0,05.

4.3.4 Expected relatives costs, different threshold for each type

In this section we repeat the results of section 4.3.3 but with the threshold probabilities for each type set separately.

Table 30 Expected relative costs for the combination of fraud types 6, 11 and 12 and fraud

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,00470	0,01492	0,02112	0,04630	0,00473	0,01500	0,02116	0,04640
0,01	0,01486	0,04610	0,06444	0,13930	0,01494	0,04615	0,06492	0,14230
0,1	0,04393	0,13219	0,18110	0,37230	0,04391	0,13529	0,18702	0,38650

Table 31 Change of expected relative costs for combination of fraud types 6, 11 and 12 over fraud

	C_{FP} / C_{FN}			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**
0,001	-0,53 %	-0,53 %	-0,19 %	-0,22 %
0,01	-0,54 %	-0,11 %	-0,74 %	-2,11 %
0,1	0,05 %	-2,29 %	-3,17 %	-3,67 %

The decrease in expected relative costs with the combination of fraud types 6, 11 and 12 and the majority voting compared to fraud decreases between 0,11% - 3,67% except in one case it increases by 0,05%. It should be noted that the minimum values for the combination were obtained numerically by starting from the common threshold value for all types and then minimizing one type at a time. This does not necessarily result into a global minimum, just a local minimum. So the results here for the combination present an upper limit for expected relative cost and a minimum for the decrease. The actual global minima may result into even larger decrease. The threshold probabilities related to the minimum expected relative costs

Table 32 Threshold probabilities for the combination of fraud types 6, 11 and 12, threshold probability for the combination with common threshold and threshold probability for fraud

		Combined 6, 11 and 12 Different threshold probabilities			Combined Common threshold	Fraud
C_{FP} / C_{FN}	P(fraud)	6	11	12		
1	**0,001**	0,957	0,958	0,957	0,958	0,933
	0,01	0,914	0,916	0,916	0,916	0,894
	0,1	0,842	0,834	0,85	0,841	0,828
0,1	**0,001**	0,915	0,916	0,916	0,916	0,894
	0,01	0,852	0,834	0,85	0,841	0,828
	0,1	0,695	0,78	0,692	0,730	0,717
0,05	**0,001**	0,904	0,888	0,911	0,898	0,880
	0,01	0,821	0,816	0,824	0,820	0,810
	0,1	0,656	0,754	0,639	0,673	0,677
0,01	**0,001**	0,841	0,834	0,868	0,841	0,834
	0,01	0,735	0,78	0,63	0,730	0,734
	0,1	0,546	0,544	0,636	0,579	0,605

The threshold probabilities do not show much difference between fraud and types. There is a general tendency of threshold probabilities to become lower when the cost ratio decreases and the prior fraud probability increases. When cost ratio is 1 the threshold probabilities of combined case with common threshold for types tend to be higher than for fraud. But when the cost ratio decreases, the threshold probabilities of the combined case tend to drop below fraud. With different threshold probabilities there is spread around the common threshold probability. The spread is larger for the lower cost ratios and larger prior fraud probabilities.

Table 33 Accuracy for the combined 6, 11 and 12 fraud types with different threshold probabilities and fraud, with thresholds minimizing the corresponding expected relative costs, the values in bold have p-value below 0,05/12 and values italized have p-values between 0,05/12 and 0,05 under binomial difference test

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,9919 (3,7E-4)	0,9870 (4,7E-4)	0,9840 (5,2E-4)	0,9714 (6,9E-4)	0,9919 (3,7E-4)	0,9870 (4,7E-4)	0,9841 (5,1E-4)	0,9726 (6,7E-4)
0,01	0,9870 (4,7E-4)	0,9715 (6,8E-4)	0,9632 (7,7E-4)	*0,9263* (0,0011)	0,9870 (4,7E-4)	0,9704 (7,0E-4)	0,9639 (7,7E-4)	*0,9290* (0,0011)
0,1	0,9702 (7,0E-4)	*0,9220* (0,0011)	**0,8985** (0,0012)	**0,8035** (0,0016)	0,9704 (7,0E-4)	*0,9188* (0,0011)	**0,8865** (0,0013)	**0,8124** (0,0016)

The behaviour of accuracy is different from the common threshold case in Table 26. Now there is no clear behaviour on which accuracy is better. Significant differences in accuracies under binomial test are observed for low cost ratios and high prior fraud probability.

Table 34 Precision for the combined 6, 11 and 12 fraud types with different threshold probabilities and fraud, with thresholds minimizing the corresponding expected relative costs

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0423 (0,0169)	0,0248 (0,0074)	0,0208 (0,0057)	0,0213 (0,0039)	0,0347 (0,0153)	0,0204 (0,0067)	0,0194 (0,0056)	0,0217 (0,0040)
0,01	0,0247 (0,0073)	0,0214 (0,0039)	0,0209 (0,0033)	0,0175 (0,0020)	0,0204 (0,0067)	0,0211 (0,0038)	0,0193 (0,0032)	0,0151 (0,0019)
0,1	0,0209 (0,0037)	0,0176 (0,0020)	0,0172 (0,0017)	0,0143 (0,0011)	0,0211 (0,0038)	0,0148 (0,0018)	0,0146 (0,0015)	0,0131 (0,0011)

Precision values for combined classifier are better than for fraud, except in 2 cases. However, none of the p-values of the binomial difference test are significant, which was also the case in Table 27.

Table 35 Sensitivity for the combined 6, 11 and 12 fraud types with different threshold probabilities and fraud, with thresholds minimizing the corresponding expected relative costs

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,0173 (0,0070)	0,0317 (0,0094)	0,0375 (0,0102)	0,0865 (0,0151)	0,0144 (0,0064)	0,0259 (0,0085)	0,0346 (0,0098)	0,0836 (0,0149)
0,01	0,0317 (0,0094)	0,0865 (0,0151)	0,1153 (0,0171)	0,2104 (0,0219)	0,0259 (0,0085)	0,0893 (0,0153)	0,1037 (0,0164)	0,1729 (0,0203)
0,1	0,0893 (0,0153)	0,2248 (0,0224)	0,2911 (0,0244)	0,4784 (0,0268)	0,0893 (0,0153)	0,1960 (0,0213)	0,2767 (0,0240)	0,4179 (0,0265)

Like precision the sensitivity values are better for combined classifier than for fraud classifier, except in one case. However, none of the sensitivities are significantly different under the binomial difference test, as with the common threshold case in Table 28.

Table 36 Specificity for the combined 6, 11 and 12 fraud types with different threshold probabilities and fraud, with thresholds minimizing the corresponding expected relative costs, the values in bold have p-value below 0,05/12 and values italized have p-values between 0,05/12 and 0,05 under binomial difference test

	Combined 6, 11 and 12				**Fraud**			
	C_{FP} / C_{FN}				**C_{FP} / C_{FN}**			
P(fraud)	**1**	**0,1**	**0,05**	**0,01**	**1**	**0,1**	**0,05**	**0,01**
0,001	0,9977 (2,0E-4)	0,9926 (3,5E-4)	0,9896 (4,2E-4)	0,9766 (6,2E-4)	0,9976 (2,0E-4)	0,9926 (3,5E-4)	0,9897 (4,2E-4)	0,9778 (6,1E-4)
0,01	0,9926 (3,5E-4)	0,9767 (6,2E-4)	0,9682 (7,2E-4)	*0,9305* (0,0010)	0,9926 (3,5E-4)	0,9756 (6,4E-4)	0,9689 (7,2E-4)	*0,9335* (0,0010)
0,1	0,9753 (6,4E-4)	*0,9261* (0,0011)	**0,9020** (0,0012)	**0,8054** (0,0016)	0,9756 (6,4E-4)	*0,9231* (0,0011)	**0,8901** (0,0013)	**0,8147** (0,0016)

The specificity for the combined classifier is sometimes better than with fraud and sometimes worse. This behaves in the similar manner to accuracy in Table 33. The same cases have significant differences under the binomial test as with accuracy. Altogether the specificity in combined case and fraud are similar.

Overall the expected relative costs decrease with combined classifier with different threshold probabilities compared to fraud classifier even more than with combined classifier with common threshold. This is no surprise because different thresholds contain the common threshold as a subcase where one puts all the threshold probabilities equal. Therefore the minimization over different thresholds should produce a result that is at least as good as with the common threshold. At the same time the precision and sensitivity become better for the combined classifier than for fraud classifier. The accuracy and specificity are similar between the combined and fraud classifiers. Since the expected relative costs are lower and the other performance measures better or at the same level, the combined classifier with different probability thresholds is better than fraud classifier. However, one cannot make that claim based on statistical tests for expected relative costs, because one gets only one value from this process. It would be possible to achieve the statistics with repeating the process multiple

times. However, that would require more computing resources and was not done for that reason. For the other performance measures one is able test the difference using binomial difference test. There were 4 cases with accuracy and specificity, where the differences were significant corresponding to low cost ratio and high prior fraud probability. This could be improved, too, by repeating the process multiple times.

5. Discussion

5.1 Comments on results

Section 4.2 deals with research question 1. The variables that were chosen in at least 4 of the 5 folds are presented first. As can be seen there are differences between fraud types in the variable selection. What is not seen in the results is that there is quite a lot of variation in the number of variables chosen in each fold. For fraud types the number of variables chosen by all folds is between 8 – 15, with fraud itself number of variables is 19. If we include the variables chosen by just 4 folds, for fraud types the number of variables varies between 4 – 12, and with fraud itself it is 6. Typically variables chosen for fraud appeared in at least 3 fraud types, too, only total debt to total assets appeared only in one fraud type model 44 and fraud model. The results vary when partitioning into folds is varied, different variables get chosen in all 5 folds or in 4 folds. One would likely get more stable results with 10 folds, since with 5 folds training sets in different folds have 60% of observations the same, with 10 folds training sets in different folds would have 80% of observations the same.

The signs of the coefficients in the logistic regressions (with regularization) are listed in the table 13. In most cases the sign is what is predicted in section 2. The variables, which have the wrong sign, typically have descriptive statistics showing that the sign is actually what one would expect. Predictor variable net sales is an exception. It has descriptive statistics supporting positive sign and it is also the predicted sign based on auditing standards. However, the actual sign is negative for fraud and fraud types, except the fraud type revenue recognition issues has positive sign for the coefficient. Net sales appears with many other variables which also contain net sales either in the nominator or denominator in the definitions of the variables. So the wrong sign may well be the result of the complicated non-linear dependence on the net sales. It may also be that this is an effect of trying to conceal fraud. Two sided accounting guarantees that anything wrong in one place of the financial statement causes there to be a corresponding mistake in at least one other place. So with the all the other fraud types except revenue recognition issues the primary mistake may well be covered with a mistake in net sales, which decreases the net sales.

Most of the predictor variables have the same sign of the coefficient in all the fraud and fraud type logistic regressions. However, 14 of the variables have different signs of the coefficients for different fraud types and fraud itself. This may be an indication that these variables are not good variables to use directly for predicting fraud, because there are effects that cancel each other. This is an issue that could be studied further.

The results for the naïve threshold probability 0,5 show that the accuracy for fraud is 65,5% and for fraud types varies between 63,9% - 74,0%. It is seen that fraud types tend to produce better accuracy than fraud itself, only fraud type 20 (inventory, vendor and/or cost of sales issues) has lower accuracy than fraud. Precision for fraud is 1,1% whereas fraud types vary from 0,27% - 0,51%, so fraud wins this comparison to fraud types. However, if one used just weighted coin toss to choose, one would expect the fraud precision to be 0,59% and for fraud types it is between 0,12% - 0,23%.[9] So compared to that benchmark both fraud types and fraud precision is roughly twice as much as with weighted random choice. Looking at it from this point of view the lower precision results for fraud types seem to be the result of having fewer number of cases than fraud totally has. Same reasoning can be applied to the accuracy results. It may be that the better accuracy is just a phenomenon of lower amount of fraud types than frauds. The sensitivity for fraud is 63,1% and for fraud types it is between 58,1% - 70,2%. Since sensitivity goes through only observations of fraud and fraud types, these are very much comparable and differences cannot be explained away with the number of observations. Specificity for fraud is 65,6% and for fraud types it is between 63,9% - 74,1%, nearly the same values as with accuracy. Specificity is over the non-fraud or non-fraud type observations, and since the relative difference between total number of observations and non-fraud observations is small, the results are close to each other. As a total it seems that fraud types might function a bit better than fraud, when the naïve probability threshold 0,5 is used for all them.

When the threshold probability is changed, the results are expected to change. The ROC curves are made for the changing threshold probability. From them one can see that there are

[9] In this case the coin would be weighted according to the frequencies of fraud and total number of observations. The probability of any case to be fraud would be the fraction total number of fraud (type) divided by the total number of observations. For fraud this would be 347 / 59239 = 0,59%. For fraud types the results are obtained similarly from table 3.

ranges, where fraud ROC curve is better than fraud type ROC curve, and then there are ranges, where the situation is opposite the fraud type ROC curve is better than fraud ROC curve. Based on ROC curves it seems that 5 fraud types (6, 7, 11, 12 and 44) have better ROC curve than fraud, 1 fraud type (other) has ROC curve that is about as good as fraud ROC curve, and 2 fraud types (14 and 20) have a ROC curve that is worse than fraud ROC curve. The areas under the ROC curve give just one number that measures quality of the ROC curve over the whole range of threshold probabilities. The AUC for fraud is 0,710 and for fraud types varies between 0,681 – 0,780. The fraud types that were seen to have better ROC curves have also larger AUC than fraud, the fraud type other, which had equally good ROC curve, has larger AUC than fraud, and fraud types 14 and 20 have a lower AUC than fraud, as can be seen from table 15. Based on this it seems that in general the prediction for 5 fraud types is better than for fraud.

The expected relative costs are calculated for fraud only. This is due to the fact that was already discussed in section 3. With fraud types one should include more variety in the mistakes. Since mistaking the fraud type for non-fraud type could still be another fraud type, its cost should reflect this and not be just the cost of false negatives. Likely the cost of mistaking a fraud type, when it is of another fraud type in reality, is much smaller than mistaking fraud as non-fraud. The ERC for fraud varies between 0,0047 and 0,39 when prior fraud probability varies between 0,1% - 10% and the ratios of costs between 0,01 – 1. The prior fraud probability 0,1% is likely too small and 10% is too large, the same with the relative cost bounds. The minimization of ERC chooses a threshold probability and one can look at how the performance measures behave at this threshold probability. The results produced compared fraud only to fraud type 6 but the other fraud types had similar results to fraud type 6. The accuracy of fraud varies between 81,2% and 99,2%, with fraud type 6 the accuracy is between 80,6% - 99,1%. The high accuracies are due to high threshold probabilities, which can be seen in table 32. The higher the threshold probability, the more predictions are made as non-fraud and since the sample is highly imbalanced, the accuracy grows. The precision for fraud is between 1,3% - 3,5%, for fraud type 6 it is between 0,6% - 0,85%. This is a result that is similar to threshold probability 0,5. In the same way the likely reason for the lower precision is the fact that there are less observations for fraud types than for fraud. The

sensitivity for fraud varies between 1,4% - 41,8% and for fraud type 6 it is consistently better 2,2% - 50,4%. Sensitivity increases when prior fraud probability increases and relative cost decreases. Specificity is again pretty similar to accuracy, for fraud it is between 81,5% - 99,8% and for fraud type 6 it is between 80,6% - 99,3%. One should note that the results in this paragraph from section 4.2.3 are calculated by optimizing the expected relative costs of fraud, not of the fraud types. Fraud type performance measures should really be calculated using ERC for them, but this case has not been handled in the literature previously and would really require extending the definition of ERC to include also the cost of mistaking fraud type to another fraud type. It was easier to bypass the problem with making the combined type, with which one can operate with the usual ERC definition.

Combining of fraud types 6, 11 and 12 to predict fraud is based on the ROC curves, and the fact that these three types covered the largest amount of fraud cases among the three type combinations in the five types, which had better ROC curves than fraud ROC curve. With more types included one could get better coverage of fraud cases, but since the minimum ERC had to be found numerically, the number of types needed to be as low as possible. With better method one might be able to use larger combinations.

When using the voting for the combination and common threshold probability 0,5 for all three types, the performance measures are better for combination with the exception of sensitivity, as can be seen in table 55. The ROC curve for the combination is worse than the fraud ROC curve and this is also shown by AUC for the combination is 0,683 and for fraud it is 0,710. Although in general the ROC curve for the combination is worse than for fraud, in the low false positive region which is the region that matters most here, the ROC curve for the combination is slightly better than for fraud.

In general the ERC of the combination for the common threshold probability for all the three types decreases from the ERC for fraud. However, there are 4 cases of prior fraud probability and relative cost ratio out of 12, where the ERC for the combination increases compared to fraud, as can be seen in table 25. The change in ERC varies between -2,0% and 0,46% depending on the prior fraud probability and the cost ratio. With the threshold probabilities minimizing ERC the accuracy of the combination is better than with fraud in 6 cases out of 12,

but the differences are very small. There are only 2 cases where the difference is significant as can be seen in table 26. The precision is better for the combination than for the fraud in exactly the same cases as with the ERC, but none of the differences are significant. The sensitivity has the same behaviour pattern as precision, but some of the cases have equal sensitivity. The specificity has otherwise the same behaviour pattern as accuracy, but 3 of the differences in specificities are significant. Altogether the differences in the results between the combination type and fraud are small. In order to get ERC decrease larger than 1 percent the prior fraud probability has to be unrealistically large 0,1. With realistic values the ERC differences are below 1 percent. However, since the threshold probabilities are kept the same, the minimum may not be the true one. With different threshold probabilities a true minimum is obtained.

Since there is no reason why one should use the same threshold probability for all the types, the condition is relaxed. However, the minimum with same threshold was calculated using a grid of threshold values 0, 0,001, … 1, which gives a good approximation of the minimum. With different thresholds this would result into a grid with 10^9 values. This goes beyond computer ability to handle the grid in memory, and would take a lot of time to go through. For this reason the minimization with different thresholds is done numerically, one type at a time until convergence starting from the common threshold probability for all the types. There might be a result that is even lower than what is obtained, since there is no guarantee that a global minimum is found for ERC in this way, only a local minimum. Now the ERC decreases for all cases but one, prior fraud probability 0,1 and relative cost ratio 1 results to increase of 0,05% for ERC. For the other cases the change in ERC is between -0,11% and -3,7%. ROC curve cannot be drawn for this case since it requires common threshold. The accuracy is better for fraud than for the combination in 5 cases, in 3 cases they are equal, and in 4 cases the combination has better accuracy than fraud. The differences are significant only in cases where the prior fraud probability is high 0,1 or relative cost ratio is 0,01, meaning likely unrealistic cases. The precision is higher for the combination in 10 cases out of 12. However, the differences are not significant and less than 1%. Sensitivity is also higher for the combination than for fraud in 10 cases out of 12 with not significant differences, but the size

of differences varies between 0,1% - 6,0%. The specificity has the same behaviour pattern as accuracy.

ERC significance could not be tested because one gets only one minimum from this process. By repeating the same process multiple times one could produce statistics for ERC, too. Repeating the process could also improve the statistics of performance measures so that differences might become significant. Also having 10-fold cross validation instead of 5-fold cross validation should help, too, because different folds have less variation between them in their training sets.

5.2 Comparison with published articles

The main article, which was followed here, was done by Perols (2011). Perols produced the results for expected relative costs. His results are on the range that were produced here. He does not give other performance measures. He does have area under the curve 0,823 which is much higher than obtained here for fraud 0,706 and higher than AUC for best fraud type 0,780. The reasons for this could be that Perols did much more thorough preprocessing than was done in this study. He undersampled the non-fraud cases before fitting, whereas in this study the fraud or fraud type cases were replicated, so the approach here is oversampling. He also used 10-fold cross validation that was repeated 10 times to get his results. One difference is that he made the predictor variable selection with the whole dataset before partitioning to training and test sets. Second difference is that when fitting the model in the training set, Perols used ERC as the criterion for setting fitting parameters. This is of course more preferable instead of using the accuracy as the criterion as was done in this study. The choice was based on the problems with using ERC for fitting fraud types as explained earlier. This is an issue that needs development. In any case if one were to use these methods in real situation, one should use the cost matrix of the stakeholder, whose point of view is taken.

A continuation of Perols' work is Perols et al. (2017) which uses advanced undersampling methods. In this study there is also hidden the usage of fraud types as one case. With prior fraud probability 0,6% and cost ratio 1:30 = 0,033 they had 9,6% and 10,8% decreases in

ERC. The prior fraud probability used is based on article by Bell and Carcello (2000) and the cost ratio on article by Bayley and Taylor (2007). The prior fraud probability is based on article from 2000, whereas ACFE (2018) mentioned that financial professionals think that fraud could be as large as 5%. Of course this number is fraud totally, not just the financial statement fraud, so it is unclear what numbers should be used. Nevertheless their numbers are much better than what is produced here. On the other hand the results have been obtained using support vector machines, not logistic regression and the number of predictor variables is 109 to start with in their study. However, they did produce results for AUC and logistic regression when using the methods for financial statement misclassification. Their results for logistic regression AUC varies between 0,741 – 0,770. This is at the same order as the best fraud type AUCs produced in this study. In this study 10-fold cross validation was used, but in contrast to (Perols, 2011) the variable selection is done after the partition to training and test sets, as is done in this study, too. As mentioned by Perols et al. (2017) it is expected, that the fraud types can have different variables in the fitting. This is seen in this study since different variables got selected for fraud types compared to fraud. They obtained the same thing by partitioning the variables into fraud types from the start. Maybe that should have been done here, too, instead of relying recursive feature elimination.

Fanning and Cogger (1998) had the test set accuracy, as they called their performance measures, for logit model 50% for total sample, 67% for fraud and 33% for non-fraud cases. The first is the accuracy, the second is the sensitivity and the third is the specificity used here. These are calculated with threshold probability 0,5 according to Perols (2011). Feroz et al. (2000) calculated ERC for logit model and the result varied between 0,095 – 0,688 when prior fraud probability varied between 0,1% – 0,5% and cost ratio varied between 0,02 – 1. It seems that in this study better results are obtained. The accuracy they had for 0,1% prior fraud probability was 88%, lower than what was obtained here.

6. Conclusions

6.1 Research summary

This thesis studied using the fraud types obtained from the Audit Analytics dataset to predict both fraud type and fraud itself by combining the fraud types using majority voting. It was found that the fraud types produce similar performance as fraud when measured using the accuracy, precision, sensitivity and specificity, both with threshold probability 0,5 and the threshold probability minimizing the expected relative costs. The performance is dependent on the prior fraud probability and the cost ratio of false positives to false negatives used. There are some differences in the variables that get chosen when using fraud types instead of fraud. Fraud had 19 variables in the 5 folds chosen and 6 more in just 4 out of 5 folds. The results are in Table 12. Fraud types had between 8-15 variables chosen in 5 folds and between 2-12 chosen in 4 out of 5 folds. So altogether fraud had 25 variables chosen in at least 4 out of 5 folds, fraud types had between 14-23 variables chosen in at least 4 out of 5 folds. It seems that fraud types do not need as many variables as fraud, which is an expected result. Holding period return in the violation period did not get chosen at all in fraud types or fraud. Whether new securities were issued and value of issued securities to market value were chosen in all fraud types and fraud models. Accounts receivable, accounts receivable to total assets, big 4 auditor, net sales, and whether standard industry classification is between 3000-3999 or not were chosen in 8 out of 9 fraud types and fraud models. Therefore these 2+5 variables could be considered the most important for detecting fraud. The significance of the coefficients were based on a biased estimator, which has its variance restricted by the regularization parameter, so one has to take the significance results with some prejudice.

The usage of fraud types allows one to answer questions, which one is not able answer when using just binary fraud classification. It only predicts whether financial statement is fraudulent or not. With the fraud types one can direct attention to where the problem is: revenues, costs, receivables, payables etc. This was discussed by Perols et al. (2017), too, who used just 4 types revenues, costs, assets and liabilities. Here the fraud types were based on the data itself. They are listed in Table 3 in section 3.4 and they are 1. revenue recognition issues, 2. expense recording issues, 3. foreign, related party, affiliated, or subsidiary issues, 4. liabilities,

payables, reserves and accrual estimate failures, 5. accounts/loans receivable, investments and cash issues, 6. inventory, vendor and/or cost of sales issues, 7. foreign, subsidiary only issues and 8. all the other issues not in previous categories. Categorization here is more detailed than with Perols et al. Although the using fraud types improved the results compared to fraud, the improvement was not as large as with Perols et al. So it seems that having more detailed categorization does not help. Furthermore the performance measures of the fraud types are comparable to the performance measures of fraud. The AUC for fraud types varies between 0,68-0,78, the AUC for fraud is 0,71.

Three fraud types were combined using majority voting to make a new fraud predictor. The performance of this was compared to fraud classification performance both with the same probability threshold for all types and then with different threshold probabilities for the types. With common thresholds the change in the expected relative costs is between -2,0% and 0,46% compared to fraud. With different threshold probabilities the change is between -3,7% and 0,05% compared to fraud. The other performance measures are of the same order as with fraud, sometimes better, sometimes worse. The accuracy varies between 0,80-0,99 for the combined type and between 0,81-0,99 for fraud. The precision varies between 0,014-0,042 for the combined type and between 0,013-0,035 for fraud. The sensitivity varies between 0,017-0,48 for the combined type and between 0,014-0,42 for fraud. The specificity varies between 0,805-0,998 for the combined type and between 0,815-0,998 for fraud. Differences are not generally significant. The ROC curve for fraud was slightly better as was the AUC, for fraud 0,71 for the combination 0,68. With more fraud types combined to use voting one might get even better results. At least one would have a better coverage of fraud cases than with just three fraud types.

6.2 Limitations of the study

The results cannot be generalized to different countries. The data is for US firms and fraud is defined by its laws. What might be fraudulent behaviour in US according to their laws might be allowed in other countries. The laws have also changed during the years, so there might be

problems in the data due to this. What was criminal behaviour 20 years ago might not be that today and vice versa. The cross validation method uses all the data to fit and test the model and the partition is made randomly. Because of this in some folds one uses the fitted model to predict future based past, and in other folds model fitted on future data to predict past. Therefore there may be bias in the results because of this. However, this approach is used in the published articles, too.

The predictions should be used only as an indication that one should look more carefully in the financial statement and into the accounting records, if one has access to them. Fraud is a criminal act that has to go through legal process. Predictions have errors and depending on how the probability thresholds are set, the amount of errors can be smaller or larger. Accusing of fraud based on a prediction of a model can lead into adverse legal consequences.

6.3 Suggestions for further research

Some ideas were identified already in the introduction. Namely if one gets hold of the financial impact data of Audit Analytics, one might try to predict the financial impact of fraud, which would be useful to people trying find fraud and other stakeholders. People trying to find fraud could use financial impact prediction to determine how much resources should be assigned for the task. Outside stakeholders would be more interested in how much is the financial statement wrong. If the result of the company is in millions and financial impact prediction is less than one hundred thousand, maybe it is not so important in the big picture. As far as I know this has not been attempted before, likely due to lack of data, which would contain the financial impact information.

One could enlarge the fraud type prediction to a misstatement type prediction in the same way as was done by Dutta et al. (2017). They essentially repeated the same with misstatements what Perols (2011) did with fraud . To some degree Perols et al. (2017) have already done this, but they did not use the same data source as Dutta et al. (2017).

One could add other variables. Although the 109 variables used by Perols et al. (2017) are probably more than enough. Rather one should think more towards specializing in the fraud type prediction. One should think about theoretical reasons for the variables in fraud type fitting. Perols et al. did not give much reasoning for putting variables into different fraud type classes or give a list of what belongs where. In connection to this one should consider the differences in the signs of the coefficients between the different fraud types and fraud, which was observed here. Some variables may well require to be considered with respect to a fraud type.

In this study recursive feature elimination was used. One might consider using stepwise forward selection where one adds variables one by one until no improvement is observed. At least this would likely lead to models with fewer variables.

Combination of fraud types can be done in several ways. One can use hard majority vote as was done in this study. Perols et al. (2017) seem to have used the method: fraud is predicted if one fraud type is predicted, although it is not clear. One could build several voting mechanisms: at least one type predicted, 2 types predicted then fraud, 3 types etc. The number fraud types used can also vary. In connection to this one should develop the expected relative costs for fraud types to include the different possibilities of mistakes, which exist in this case, or give clear reasons why the fraud based expected relative costs is sufficient.

One can also test whether more sophisticated models like neural networks would produce better results with more data than what was used by Perols (2011) and Perols et al. (2017). If the results of Dutta et al. (2017) for the study misstatements is any indication, this could be the case. However, the data imbalance with misstatements is much lower than with fraud.

One might make a training and test set partition based on time using the latest years as the test set. However, if one does this, one should leave the last few years out of the dataset, because fraud is typically found only afterwards. The latest years likely contain cases of fraud that have not been found yet, thus biasing the results. If the training is made on past data and tested on future, the training data likely contains all the fraud found and the test data does not unless one leaves the last years out of the dataset. Nevertheless it might be an interesting to test it.

Bibliography

ACFE. (2018). *Report to the Nations, 2018 Global Study on Occupational Fraud and Abuse.* Association of Cerfied Fraud Examiners.

Bayley, L.;& Taylor, S. (2007). Identifying earnings management: A financial statement analysis (red flag) approach. *Proceedings of the American Accounting Association Annual Meeting*, -.

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

Beneish, M. D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy, 16*, 271-309.

Beneish, M. D. (1999). Incentives and Penalties Related to Earnings Overstatements That Violate GAAP. *The Accounting Review, 24*(2), 425-457.

Dechow, P. M.;Ge, W.;Larson, C. R.;& Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research, 28*(1), 17-82.

Dechow, P. M.;Sloan, R. G.;& Sweeney, A. P. (1996). Causes and consequences of earnings manipulations: An analysis of firms subject to enforcement actions by the SEC. *Contemporary Accounting Research, 13*(1), 1-36.

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

Dutta, I.;Dutta, S.;& Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems With Applications, 90*, 374-393.

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

Feroz, E. H.;Kwon, T. M.;Pastena, V. S.;& Park, K. (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 and Management, 9*(3), 145-157.

Goodfellow, I.;Bengio, Y.;& Courville, A. (2016). *Deep Learning.* MIT Press. Available at web address http://www.deaplearningbook.org

Green, B. P.;& Choi, J. H. (1997). Assessing the risk of management through neural network technology. *Auditing: A Journal of Practise & Theory, 16*(1), 14-28.

Hanley, J. A.;& McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiving Operating Characteristic (ROC) Curve. *Radiology, 143*(1), 29-36.

Healy, P. M.;& Wahlen, J. M. (1999). A Review of the Earnings Management Literature and Its Implications for Standard Setting. *Accounting Horizons, 13*(4), 365-383.

Hennes, K. M.;Leone, A. J.;& Miller, B. P. (2014). Determinants and Market Consequences of Auditor Dismissals after Accounting Restatements. *The Accounting Review, 89*(3), 1051-1082.

Hosmer, D. W.;& Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons, Inc.

IAASB, I. B. (2018). *Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements, 2018 Edition.* New York, New York, USA: International Federation of Accountants.

Kaminski, K. A.;Wetzel, S. T.;& Guan, L. (2004). Can financial ratios detect fraudulent financial reporting? *Managerial Auditing Journal, 19*(1), 15-28.

Kanapickiene, R.;& Grundiene, Z. (2015). The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. *Procedia - Social and Behavioral Sciences, 213*, 321-327.

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

Lin, J. W.;Hwang, M. I.;& Becker, J. D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal, 18*(8), 657-665.

Palepu, K. G., & Healy, P. M. (2007). *Business analysis and valuation.* Cengage Learning EMEA.

Perols, J. L. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practise & Theory, 30*(2), 19-50.

Perols, J. L.;& Lougee, B. A. (2011). The relation between earnings management and financial statement fraud. *Advances in Accounting, incorporating Advances in International Accounting, 27*, 39-53.

Perols, J. L.;Bowen, R. M.;Zimmermann, C.;& Samba, B. (March 2017). Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction. *The Accounting Review, 92*(2), 221-245.

Precision and Recall. (3. June 2019). Looked 24. June 2019 at web address Wikipedia: https://en.wikipedia.org/wiki/Precision_and_recall

Summers, S. L.;& Sweeney, J. T. (1998). Fraudulently misstated financial statements and insider trading: An empirical analysis. *The Accounting Review, 73*(1), 131-146.

van Wieringen, W. N. (2015). *Lecture notes on ridge regression.* Downloaded from arXiv e-print repository: https://arxiv.org/abs/1509.09169v4

Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach.* Mason, OH: South-Western Cengage Learning.

Appendix

Table 37 Descriptive statistics of continuous predictor variables separately for fraud and non-fraud observations

Fraud predictor	Fraud value	count	mean	Std	min	25%	50%	75%	max
1	1	347	802,6	1850,7	0,453	23,18	122,9	357,6	8918,0
	0	58892	390,9	1214,4	0,058	6,983	38,34	190,4	8918,0
2	1	347	0,201	0,118	0,0082	0,133	0,183	0,237	0,688
	0	58892	0,167	0,107	0,0082	0,106	0,155	0,210	0,688
3	1	347	0,217	0,121	0,012	0,128	0,195	0,277	0,599
	0	58892	0,171	0,121	0,0058	0,082	0,148	0,232	0,599
4	1	347	35,80	90,41	0,000	0,588	2,745	16,69	382,0
	0	58892	15,30	51,53	0,000	0,200	1,100	5,870	382,0
5	1	347	0,067	0,105	0,000	0,014	0,032	0,071	0,922
	0	58892	0,068	0,128	0,000	0,014	0,031	0,065	0,922
6	1	347	0,013	0,019	0,000	0,0020	0,0060	0,016	0,138
	0	58892	0,010	0,019	0,000	0,0018	0,0044	0,0099	0,138
7	1	347	2,792	5,298	-53,69	1,877	2,571	4,065	19,11
	0	58892	2,341	9,007	-53,69	1,263	2,784	4,719	29,69
9	1	347	9,9E-5	0,032	-0,162	-0,0077	0,000	0,0073	0,210
	0	58892	-1,8E-4	0,051	-0,244	-0,0098	0,000	0,0076	0,211
10	1	347	1,035	0,407	0,217	0,904	0,987	1,096	4,372
	0	58892	1,068	0,560	0,217	0,868	0,993	1,131	4,372
11	1	347	1,695	4,227	-18,53	0,576	1,272	2,348	26,27
	0	58892	1,278	4,374	-18,53	0,358	0,886	1,771	26,27
14	1	347	0,429	0,295	0,038	0,185	0,341	0,636	1,527
	0	58892	0,534	0,421	0,023	0,214	0,424	0,750	2,152
15	1	347	0,085	0,176	-0,358	-0,012	0,047	0,143	1,066
	0	58892	0,092	0,232	-0,358	-0,011	0,053	0,138	1,487
16	1	347	0,322	0,262	-1,663	0,175	0,328	0,452	0,872
	0	58892	0,353	0,316	-1,663	0,225	0,352	0,516	0,912
17	1	347	-0,327	1,227	-9,214	-0,409	-0,0059	0,272	0,875
	0	58892	-0,363	1,394	-9,214	-0,424	0,0085	0,25	0,875
18	1	347	-0,019	0,861	-5,084	-0,191	-0,023	0,111	5,267
	0	58892	1,1E-4	0,986	-6,075	-0,213	-0,022	0,150	6,078
19	1	347	0,095	0,095	0,000	0,011	0,078	0,139	0,457
	0	58892	0,116	0,123	0,000	0,014	0,093	0,170	0,667
20	1	347	4126,9	9501,0	2,901	127,6	740,8	2437,5	57428,0
	0	58892	2704,3	8010,1	0,546	49,12	278,1	1390,6	57428,0
22	1	347	-0,0053	0,070	-1,021	-5,0E-6	2,0E-5	1,4E-4	0,025

	0	58892	-0,028	0,170	-1,496	-1,9E-4	1,5E-5	1,9E-4	0,025
23	1	347	0,215	0,178	0,0093	0,077	0,163	0,305	0,820
	0	58892	0,252	0,217	0,0053	0,081	0,185	0,364	0,879
24	1	347	1,320	0,875	0,070	0,668	1,115	1,614	4,656
	0	58892	1,185	0,841	0,070	0,614	0,993	1,513	4,656
26	1	347	98,86	520,7	-237,0	0,055	2,815	7,434	3535,3
	0	58892	194,0	790,7	-237,0	-1,484	3,352	13,78	3535,3
27	1	347	-0,067	0,126	-1,076	-0,090	-0,047	-0,014	0,266
	0	58892	-0,010	0,231	-1,616	-0,111	-0,055	-0,014	0,266
28	1	347	0,600	0,390	0,067	0,398	0,576	0,7106	3,780
	0	58892	0,593	0,488	0,067	0,338	0,521	0,694	3,780
29	1	347	-2,301	2,121	-19,23	-2,790	-1,734	-1,090	0,491
	0	58892	-2,703	2,611	-19,23	-3,370	-2,137	-1,268	0,491
34	1	347	1,032	0,204	0,000	1,000	1,004	1,067	1,743
	0	58892	0,942	0,355	0,000	1,000	1,000	1,038	1,743
35	1	347	-0,034	0,239	-0,880	-0,136	-0,028	0,062	1,369
	0	58892	2,0E-4	0,355	-1,654	-0,142	-0,023	0,074	2,642

Table 38 P-values of continuous predictor variables under the $\chi 2$ homogeneity test

	Fraud type								
Pred.	6	7	11	12	14	20	44	other	Fraud
1	**1,5E-7**	0,710	**5,6E-11**	**5,5E-6**	**5,6E-8**	**6,1E-7**	**5,8E-14**	**1,1E-7**	**1,4E-13**
2	**7,6E-8**	**1,2E-4**	**4,3E-9**	**5,4E-7**	**1,6E-5**	0,085	**2,6E-6**	*0,009*	**6,1E-8**
3	**7,5E-4**	**3,4E-8**	**1,1E-7**	**5,6E-7**	**5,2E-4**	0,112	**2,3E-8**	**3,9E-4**	**9,0E-11**
4	**1,3E-12**	*0,021*	**8,4E-11**	0,058	**1,5E-10**	**1,3E-7**	**7,2E-9**	**1,4E-4**	**2,3E-11**
5	**3,3E-4**	0,734	0,794	0,084	0,791	0,500	0,251	0,804	0,790
6	**2,0E-9**	*0,007*	*0,009*	**1,7E-6**	**3,2E-4**	*0,023*	*0,009*	0,104	**1,8E-6**
7	*0,002*	0,670	**3,9E-7**	**2,8E-4**	**1,7E-5**	**1,9E-5**	**2,9E-6**	**5,1E-6**	**5,1E-13**
9	0,660	0,051	0,215	0,727	0,823	0,186	0,151	*0,031*	0,212
10	**0,003**	0,320	**3,4E-4**	0,163	**0,004**	0,288	**1,8E-4**	0,277	**6,2E-4**
11	**0,002**	0,353	**0,002**	*0,011*	0,119	*0,021*	*0,019*	**0,003**	**1,1E-6**

14	*0,013*	*0,049*	0,319	**7,3E-4**	0,458	**0,002**	0,108	0,392	**0,004**
15	0,799	*0,032*	0,235	0,220	0,085	*0,036*	0,115	0,211	0,436
16	0,071	**2,8E-4**	*0,045*	*0,020*	*0,024*	*0,040*	*0,021*	**6,4E-5**	**3,7E-5**
17	0,201	0,798	0,913	0,139	0,790	0,918	0,852	0,634	0,591
18	0,080	0,961	0,612	0,263	0,073	0,212	0,859	0,150	0,299
19	0,162	*0,016*	**6,5E-5**	0,644	0,053	*0,039*	*0,025*	0,572	**0,004**
20	**4,2E-4**	0,324	**2,1E-5**	**7,4E-4**	**1,0E-7**	**4,0E-4**	**1,8E-8**	**1,7E-4**	**3,2E-9**
22	*0,040*	0,307	**3,1E-4**	0,180	**0,001**	*0,010*	*0,009*	0,053	**2,0E-4**
23	**3,0E-4**	0,428	0,070	*0,047*	*0,006*	**0,001**	*0,027*	0,120	0,080
24	**8,6E-5**	*0,033*	*0,045*	0,569	**4,4E-4**	*0,017*	*0,018*	0,299	0,072
26	**8,1E-4**	0,212	**1,4E-4**	*0,023*	**9,2E-5**	*0,010*	**0,003**	*0,039*	**5,3E-8**
27	*0,014*	0,950	0,220	0,080	0,183	0,051	0,334	0,126	*0,031*
28	*0,026*	0,678	0,095	0,182	0,336	0,478	0,145	*0,014*	**0,002**
29	**1,1E-5**	0,058	**0,002**	**8,2E-6**	0,860	0,156	**8,9E-6**	0,702	**1,2E-4**
34	**2,7E-5**	0,120	**1,5E-7**	**1,6E-4**	*0,016*	**5,9E-5**	0,064	**1,1E-10**	**1,1E-9**
35	0,141	0,369	0,646	0,754	0,801	0,324	0,266	0,881	0,672