ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

MASTER THESIS
MSc Accounting, Auditing and Control
Specialization: Accounting and Auditing

Audit Quality from a Different Perspective: The Use of Machine Learning to Predict Restatements

Mandy Dalebout | 475530
Supervisor: dr. J. Bae
Second assessor: prof. dr. J.P.M. Suijs

Date: 19-11-2021

Abstract

In the last decades, audit quality became a trending topic in accounting research especially since the occurrence of world-famous scandals (e.g. Enron). Audit quality is often analysed by linear- or logistic regressions, but this research uses machine learning to predict restatements, which is the best proxy for audit quality. The aim of this research is to examine the most accurate algorithm to predict audit quality in order to create consensus in the accounting literature. The sample contains 78,917 firm-year observations of 10,078 different American firms between 2000 and 2019. The random forest predicts audit quality the most accurate out of the six supervised machine learning algorithms. Further, the firm characteristics are better in explaining audit quality than the audit related variables. The users of the financial statements might be interested in the findings of this research. They have now more sophisticated tools to assess and predict audit quality which secures a more transparent working of the capital market. Furthermore, the audit profession might be interested in this research. Auditors can more easily identify clients with a higher likelihood to file a restatement and can therefore decide not to audit the client by which they can avoid potential litigation costs and reputation damage.

Keywords: supervised machine learning, audit quality, restatements, undersampling, stratified k-fold cross validation, artificial neural network, k-nearest neighbour, decision tree, random forest, naïve Bayes, logistic regression, confusion matrix, accuracy, G-mean, AUC, ROC curve

Table of contents

1. Introduction

In the last decades, audit quality became a trending topic in the field of accounting research. The interest is strongly increased since the Enron scandal in 2001 followed by the bankruptcy of Arthur Anderson and the World-Com scandal in 2002 (Francis, 2004). The challenge is to measure audit quality, because audit quality is unobservable and therefore hard to measure. Audit quality is best measured by material misstatements, which can be best proxied by restatements (Aobdia, 2019; DeFond & Zhang, 2014; Rajgopal, Srinivasan & Zheng, 2021). Restatements are namely direct and give strong evidence of poor audit quality, because they reveal the failure of the firm's financial reporting process, the failure of the internal controls over the financial reporting and the failure of the audit process (DeFond & Zhang, 2014; Draeger, Hermann & Lawson, 2016). In the existing literature, audit quality is often analysed using linear- or logistic regression models. However, this research introduces machine learning as methodology to analyse audit quality, because machine learning can detect complex patterns in large datasets, can determine the most predictive variables to explain audit quality and can make accurate out-of-sample predictions (Bertomeu, Cheynel, Floyd and Pan, 2020).

Thus, machine learning is a more sophisticated methodology compared to linear- or logistic regressions. Nevertheless, machine learning is currently barely used in the accounting literature. Therefore, this research aims to combine machine learning and audit quality to gain new insights in the accounting literature. This leads to the following research question:

Which machine learning algorithm can accurately predict audit quality and which variables can explain audit quality?

The existing literature of the use of machine learning to predict audit quality is limited. Dutta, Dutta and Raahemi (2017) use firm characteristics to predict restatements between 2001 and 2014. The artificial neural network is the best algorithm to predict restatements according to them. Zhang (2019) uses both firm characteristics and audit related variables to predict restatements between 2008 and 2016. She concludes that the random forest is the best algorithm to predict restatements and that the audit related variables are the best variables to explain restatements. Bertomeu et al. (2020) predict Accounting and Auditing Enforcement Releases (AAERs), another proxy of material misstatements, with the help of financial-, audit-, credit rating-, opinion divergence- and corporate governance variables between 2001 and 2014. They conclude that the k-nearest neighbour is the best algorithm to predict AAERs and that the audit variables are the most predictive variables of AAERs. The preceding illustrates that the researches have no consensus about which algorithm predicts audit quality the most accurate. Therefore, the aim of this research is to examine the most accurate algorithm to predict audit quality to create consensus in the accounting literature.

The research question is answered using 78,917 firm-year observations of 10,078 different American firms between 2000 and 2019. The dataset consists of 13 audit related variables and 22 firm characteristics to predict restatements, which are selected on the basis of the literature. The variables are the input of the six supervised machine learning algorithms, which are the artificial neural network, k-nearest neighbours, decision tree, random forest, naïve Bayes and logistic regression. The six supervised machine learning algorithms are fully optimized using the training set and used to make predictions for the observations in the testing set. The performance of the algorithms is then evaluated using three evaluation metrics, namely accuracy, geometric mean (G-mean) and Area Under the receiver operating characteristics Curve (AUC). The performance of the algorithm is better if the evaluation metrics are higher.

The findings of the research illustrate that the random forest is unanimously the most accurate supervised machine learning algorithm to predict restatements and therefore poor audit quality, which is consistent with the finding of Zhang (2019). The most predictive variables of restatements, and therefore of poor audit quality, are the firm characteristics, which is not in line with the conclusions of Zhang (2019) and Bertomeu et al. (2020). This could be due to the random forest in combination with the types of variables. Random forest can namely more easily split continuous variables than dummy variables, because continuous variables have larger ranges. It is possible that the firm characteristics ended as the most predictive variables of audit quality, because the firm characteristics mainly consist of continuous variables, while the audit related variables primarily contain dummy variables.

The initial findings of the research are quite satisfactory, but the performance of the supervised machine learning algorithms can always get better. Therefore, further research is conducted by adding gross domestic product (GDP), GDP growth and inflation to the dataset to control for their effect (Hu, Sun, Vaserhelyi and Zhang, 2021). The additional findings of the research do not differ from the initial findings, although the performance of the algorithms is increased. Thus, the random forest is still the most accurate supervised machine learning algorithm to predict restatements and therefore poor audit quality and the firm characteristics are still the most predictive variables of audit quality. However, GDP is a predictive variable of audit quality, which is in line with the findings of Hu et al. (2021). A potential negative relationship between GDP and restatements could exist, but should be examined further, because the sample only consists of American firms whereby the GDP is the same for every firm. In summary, random forest is the most accurate supervised machine learning algorithm to predict audit quality and the firm characteristics are the best variables to explain audit quality.

The findings of the research might be of interest to the different users of the financial statements (e.g. lenders, investors and regulators) and the audit profession. As a result of this research, the decision-making process of the users of the financial statements is improved, because they now have more sophisticated tools, compared to linear- or logistic regressions, to assess and predict audit quality. In this way, the information asymmetry between firms and users of the financial statements is reduced which secures a more transparent working of the capital market. For example, lenders can now more easily identify firms with low audit quality and can therefore more easily decide whether they should lend money to a firm or not. The audit profession can also benefit from this research. The auditor can use more sophisticated tools in the planning phase of (new) clients, where the engagement risk is assessed. Auditors can identify (new) clients with a higher likelihood to file a restatement. The auditor can assess that, partly due to the higher likelihood to file a restatement, the engagement risk is too high and decide not to audit the (new) client by which they can avoid potential litigation costs and reputation damage.

In future research, the sample could be adjusted to other countries or regions in the world. This research and almost all existing researches predict audit quality using American firms in their sample. For example, a sample consisting of European firms could be used in further research to predict audit quality in order to get more generalizable conclusions about analysing audit quality with the help of machine learning. Furthermore, this research finds a potential negative relationship between GDP and restatements, but this needs further investigation. This could be done by using a sample with, for example, European firms where the GDP is different for every country. In this way, it can be investigated whether the GDP is still in the top ten of the most predictive variables of audit quality after which it can be concluded whether the relationship between GDP and restatements exists.

Besides, quite simple supervised machine learning algorithms are used in this research which makes the performance of the algorithms not spectacularly high. A suggestion for further research is to use more advanced machine learning algorithms, such as gradient boosting algorithms, which can detect more complex patterns in large datasets and therefore make more accurate predictions about audit quality. Lastly, more (continuous) audit related variables could be included in the dataset, because the amount of firm characteristics and the amount of audit related variables is very unbalanced in this research. In this way, it could be examined whether the firm characteristics are still the most predictive variables of audit quality. Examples are the auditor market share, but also some variables on engagement level, such as the age, gender, education level and years of experience of the engagement partner.

The research continues with the literature review where audit quality, machine learning and the use of machine learning in accounting and auditing (research) are further explicated. Next, the data and methodology are explained, where the selection of the variables is described, the descriptive statistics are presented and where the six supervised machine learning algorithms and the three evaluation metrics are explained. Thereafter, the results of both the initial and the additional are presented and elucidated. Finally, the research question is answered in the conclusion.

2. Literature review

The literature review explains the relevant concepts to answer the research question. Firstly, the definition and measurement of audit quality is explicated, after which various factors influencing audit quality are discussed. Secondly, the definition and workflow of machine learning is given, followed by an explanation of supervised machine learning. Lastly, the use of machine learning in accounting- and auditing (research) is reviewed.

2.1. Audit quality

2.1.1. Definition of audit quality

The literature shows that audit quality is a complex concept which is hard to define (Francis, 2011). The most commonly used definition is from DeAngelo (1981) who states that audit quality is "the market-assessed joint probability that a given auditor will both discover a breach in the client's accounting system and report the breach" (p. 186). The discovery of a breach is linked to the auditor's competence and the level of effort, while reporting the breach is related to the objectivity, professional scepticism and independence of the auditor (Knechel, Kirshnan, Pevzner & Velury, 2013). Despite the fact that the definition points out two important components of audit quality, the definition of DeAngelo contains two problems according to Knechel et al. (2013). The first problem is that the definition is not linked to the audit risk model, which is the guidance during an audit and reflects the perception of the auditor. The second problem is that market perception could be incorrect. Another problem of DeAngelo's definition is the understatement of the benefits of high audit quality, because the definition characterizes the audit process as a binary process, where the auditor solely has to detect and report unequivocal U.S. GAAP violations (DeFond & Zhang, 2014). High audit quality can be achieved if auditors also examine whether the financial statements faithfully reflect the firm's underlying economics, rather than only examine whether the financial statements are in accordance with the applicable accounting principles.

The definition of DeAngelo is the most commonly used definition by accounting researchers. But, the overview by Watkins, Hillison and Morecroft (2004) indicates that accounting researchers also use other definitions for audit quality. Audit quality is also defined in the literature as (1) the probability that the auditor will not issue an unqualified opinion due to material errors in the financial statements, (2) the accuracy of the firm's information that is assessed by the auditor and (3) a way to measure if an audit reduces noise and bias and improves the fineness in the accounting data. In addition, governmental institutions also attempted to define audit quality. For example, the Government Accountability Office (2003) defines audit quality as giving a reasonable assurance that the financial statements and disclosures are (1) presented in accordance with U.S. GAAP and (2) are free from material misstatements due to fraud or error. Although the different definitions of audit quality, they have some commonalities. These definitions refer to competence and independence and how both are perceived (Watkins et al., 2004) and suggest that audit quality exists on a continuum where more is better than less (Francis, 2011; Knechel et al., 2013). In summary, there is no uniform definition of audit quality yet, but the definitions have definitely commonalities.

2.1.2. Measurement of audit quality

The measurement of audit quality is, just like defining it, problematic, because audit quality is unobservable (DeFond & Zhang, 2014; Wooten, 2003). DeFond and Zhang (2014) and Knechel et al. (2013) provide a framework of audit quality. They examine several proxies of audit quality in their frameworks and state that many of them represent a valid measurement of audit quality. In reality, every proxy has its strengths and weaknesses, measures audit quality (in)directly and the best proxy is dependent on the research design. Despite the fact that there is currently no consensus which proxy captures audit quality the best, Aobdia (2019) conclude that restatements, whether the issuer beet or meet the analyst's forecast and audit fees are the best proxies to accurate measure audit quality. Besides, Rajgopal et al. (2021) conclude that restatements and total accruals consistently and positively predict each of the six most cited audit deficiencies. Restatements still consistently and positively predict the six audit deficiencies when all audit quality proxies, 14 in total, are included in one single regression. Therefore, restatements are the best proxy for poor audit quality.

Restatements are classified as an output-based proxy (Aobdia, 2019; DeFond & Zhang, 2014; Knechel et al., 2013). The framework of DeFond and Zhang (2014) compares the most commonly used output-based audit quality proxies in the literature based on their directness and gives the strengths and weaknesses for every proxy. Material misstatements, which can be proxied by both restatements and Accounting and Auditing, Enforcement Releases, are relatively more direct and more egregious compared to other proxies. An advantage is that the proxies are discrete with a relatively high measurement consensus which lead to a relatively low measurement error. Another strength of restatements is that it gives strong evidence of poor audit quality. The last advantage is that AAERs and restatements can identify the existence of management fraud, which is important, because the perception is that auditors should prevent and detect fraud. A limitation of material misstatements is that they are rare and sometimes undetected.

In this research, restatements are used as a proxy for audit quality based on the previous literature. Restatements are corrections to misstatements in earlier financial statements of a firm (Palmrose, Richardson & Scholz, 2004). Restatements reveal both the failure of the firm's financial reporting process and the failure in the internal controls over the financial reporting (Draeger et al., 2016). Furthermore, restatements reveal a failure in the audit process, because the auditor issued an unqualified audit opinion for material misstated financial statements (DeFond and Zhang, 2014). The firm, the Securities and Exchange Commission (SEC), the independent auditor or a combination of these can request a restatement (Palmrose et al., 2004). The firm can detect misstatements during the internal audit, the SEC during an inspection of the firm's filings and the independent auditor during the external audit. A restatement can be disclosed in several ways. Some restatements are reported in a press release, some in 8-K (current events) filings and some are reported in 10-K (adjusted financial statements) filings.

2.1.3. Drivers of audit quality

In the literature many frameworks have been developed to determine drivers that could affect audit quality. The Financial Reporting Council (FRC) (2008) establishes the first formal framework in which the FRC identifies five drivers for audit quality. The drivers that influences the audit quality are (1) the culture within an audit firm, (2) the skills and personal qualities of audit partners and staff, (3) the effectiveness of the audit process, (4) the reliability and usefulness of audit reporting and (5) factors that are outside the control of auditors. The FRC mentions various indicators for each driver that could positively contribute to the audit quality. For example, an indicator for the skills and personal qualities of audit partners and staff is that audit personnel needs to be sufficient trained in auditing, accounting and specific industry issues. Another indicator is an active and professional audit committee that relates to factors that are outside the control of auditors.

Francis (2011) develops a framework for studying factors that are associated with the level of audit quality. The factors that are identifies by Francis are (1) audit inputs, (2) audit process, (3) accounting firms, (4) audit industry and audit markets, (5) institutions and (6) economic consequences of audit outcomes. The audit quality is higher when auditors are competent and independent, which is related to the audit input. The audit inputs have an effect on the audit process, where the audit quality is higher when the audit engagement team chooses the right tests and appropriately evaluate the evidence of these tests. The accounting firm contributes to higher audit quality by developing test procedures and creates incentives to make personnel behave ethically and professionally. The accounting firms constitute an industry, which can affect markets and economic behaviour. The audit takes place in an institutional context where institutions (e.g. SEC) fine auditors and accounting firms for low quality audits to encourage high quality audits. Lastly, the outcome of an audit affects the decisions of the users of financial statements, e.g. whether investors should invest in specific firms or not.

Knechel et al. (2013) create a framework based on the frameworks of the FRC and Francis. The framework consists of four dimensions, namely (1) inputs, (2) process, (3) outcomes and (4) context. Examples of inputs are the level of professional scepticism, knowledge and expertise. The audit quality is significantly influenced by the inputs, because improvements in inputs should lead to improvements in the other dimensions (i.e. process and outcomes). The quality of the audit is dependent on the level of professional judgement during the audit process. The outcome of the audit is uncertain and unobservable, as mentioned before. Many contextual characteristics can influence the audit quality. For example, a long relationship between auditor and client (i.e. auditor tenure) can negatively influence the audit quality, because the independence of the auditor may be harmed by a long relationship between auditor and client. Nevertheless, a long tenure can also increase audit quality, because short tenure means that the auditor barely knows something of the client. The context indicators directly influence inputs and/or process, so they indirectly influence the outcome of an audit. In summary, there are many factors that could positively or negatively influence the audit quality. The factors can mainly be classified as characteristics of the auditor (e.g. expertise) and characteristics of the audit engagement (e.g. auditor tenure).

2.2. Machine learning

2.2.1. Definition of machine learning

In the last two decades, machine learning has become increasingly popular. However, the development of machine learning already started in 1950 by the introduction of the Turing Test which determines whether a computer is capable to think like a human (Alzubi, Nayyar & Kumar, 2018). The first definition of machine learning is introduced by Samuel (1959) who states that machine learning is the capability of computers to learn without being explicitly programmed. Approximately forty years later Mitchell (1997) came up with the following definition of machine learning: "a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" (p. 2). Currently, machine learning is described as a category of artificial intelligence which makes computers able to learn patterns from large and complex data with minimal human intervention (Alzubi et al., 2018; Cho, Vasarhelyi, Sun & Zhang, 2020; Kavakiotis et al., 2017). Since this description is almost identical to the definition of Mitchell, machine learning is defined in this paper as the ability of computers to learn patterns from large and complex data with minimal human intervention.

2.2.2. Workflow of machine learning

The workflow of machine learning, shown in Figure 1, consists of six different stages and is similar for all different kinds of machine learning algorithms (Alzubi et al., 2018; Kotsiantis, Zaharakis & Pintelas, 2007). The first stage is the collection and preparation of the dataset, where the data is collected, structured and pre-processed. The second stage is known as the feature selection where unimportant and irrelevant variables are removed from the dataset. The third stage consists of choosing the best machine learning algorithm for the specific problem. Subsequently, the proper values for the parameters in the algorithm are chosen in the fourth stage. The algorithm is trained in the fifth stage and the performance of the algorithm is evaluated by different evaluation metrics in the sixth and final stage.

2.2.3. Supervised machine learning

Supervised machine learning algorithms use historical data to make predictions (Alzubi et al., 2018; Kotsiantis et al., 2007). The prediction of restatements belongs to the supervised machine learning, because historical data of companies is used to predict restatements. In supervised machine learning, the dataset is splitted in a training-, validation- and testing set to avoid memorizing and overfitting, which occurs when the algorithm exactly fits the training data (Elkan, 2012). The training set is used to train the algorithm, while the validation set is utilized for frequent evaluation of the algorithm and to optimize the parameters in the algorithm. Finally, the testing set is used to evaluate the performance of the final algorithm. Thus, the testing set is only used for the evaluation of the final algorithm and the data of the testing set remains unseen until the evaluation point to avoid overfitting. Examples of supervised machine learning algorithms are decisions trees, naïve Bayes and k-nearest neighbours (Kotsiantis et al., 2007). In practice, supervised machine learning algorithms are used by Netflix, where the algorithms recommend users which movies and series they should watch based on their historical data (Alzubi et al., 2018; Gomez-Uribe & Hunt, 2015).

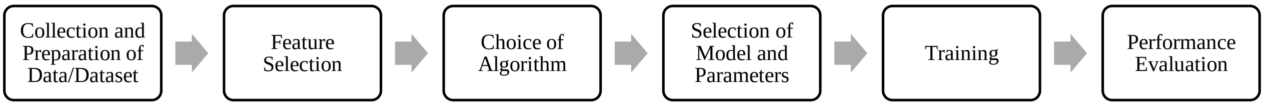

Figure 1 Workflow of machine learning
Source: Alzubi et al., 2018.

2.3. Machine learning in accounting and auditing (research)

2.3.1. Machine learning in accounting

Machine learning will make the accounting practice easier and more efficient by automating many steps in the accounting cycle (Cho et al., 2020). For example, machine learning can enhance the ordering process, where a chatbot talks with the customer, places the order, generates the invoice and processes the data into the firm's accounting information system. Furthermore, some accounting procedures become outdated, for instance the physical count of the inventory. The process can be automated by a drone which can identify objects and examine the quality and condition of the inventory. Machine learning can also be used to improve the accuracy of accounting estimates (Ding, Lev, Peng, Sun & Vasarhelyi, 2020). Examples of accounting estimates are depreciation, pension, employee stock options expenses, bad-debt allowance and fair value measurement. Currently, these accounting estimates are based on managerial judgement and are therefore sensitive to estimation errors and managerial manipulation. Ding et al. (2020) use machine learning to demonstrate that machine learning can improve the managerial estimates. They used data on loss reserves (future customer claims) from insurance companies. The loss estimates that were generated by machine learning were better than the managerial estimates. The foregoing examples show that the accounting practice will become easier, better and more efficient by the use of machine learning techniques.

2.3.2. Machine learning in auditing

The machine learning techniques can also be applied in the audit practice (Sun, 2019). The framework that Sun has developed, demonstrates how the capabilities of deep learning can be used in all audit phases. Speech recognition can be used in the internal control evaluation phase to understand oral responses from the interviewer which can create new insightful information. The use of "maybe" in answers or changes in the voice can imply deception. Machine learning techniques can also be used to automate substantive testing procedures such as asking suppliers for confirmation letters about the accounts payable and reviewing those confirmation letters. In this manner, machine learning assists the auditor to verify that the total amount of liabilities matches with the amount in the accounting records. The use of machine learning in auditing will grow in the upcoming years, but machine learning will always support the auditor, because professional judgement and professional scepticism are always needed which cannot be replaced by machine learning (Brown-Liburd, 2017; Sun, 2019).

2.3.3. Machine learning to predict accounting restatement

The existing literature of the use of machine learning to predict audit quality is limited. One of the first researches is conducted by Dutta et al. (2017) where they use firm characteristics to predict restatements. The artificial neural network predicts restatements the best concerning the accuracy and the AUC. The sample period covers the period 2001-2014, where the results are consistent after controlling for the financial crisis. Zhang (2019) uses firm characteristics, traits of the auditor and characteristics of the audit engagement to predict restatements between 2008 and 2016. Out of the ten algorithms, the random forest occurs to be the best algorithm to predict restatements. Lastly, the research shows that the audit related variables are the best variables to predict restatements.

A recent research, where machine learning is used to predict audit quality, was conducted by Bertomeu et al. (2020). They predict AAERs with financial-, audit-, credit rating-, opinion divergence- and corporate governance variables between 2001 and 2014. As found in the literature review, audit quality can be measured by material misstatements, which can be proxied by restatements, but also by AAERs. They compare several algorithms with each other where the k-nearest neighbour performs as the best algorithm. Finally, the research indicates that the audit variables are the most important variables to explain AAERs.

The three previous researches have no consensus about which algorithm predicts audit quality the best. The aim of this research is to examine the most accurate algorithm to predict audit quality in order to create more consensus in the accounting literature. The most accurate algorithms to predict audit quality of the three aforementioned researches are utilized. Thus, the artificial neural network of Dutta et al. (2017), the random forest of Zhang (2019) and the k-nearest neighbour of Bertomeu et al. (2020) are used to examine which algorithm performs best. Further, the decision tree and naïve Bayes are selected as supervised machine learning algorithms, because these are classic supervised machine learning algorithms and can be used to compare the performance of the different algorithms. In addition, the logistic regression is added as supervised machine learning algorithm in order to demonstrate that the other supervised machine learning algorithms probably perform better than the logistic regression.

3. Data and methodology

In this chapter, the data and methodology are described that are used to answer the research question empirically. First, the selection and collection of the variables are described. Thereafter, the pre-process of the dataset is explained, followed by the descriptive statistics. Afterwards, a short description of the training- and testing set follows. Finally, the six supervised machine learning algorithms and the three evaluation metrics are explained.

3.1. Data

Table 1 shows that 35 variables are selected, on basis of the literature, to predict restatements. The variables consist of 13 audit related variables and 22 firm characteristics. The choice and collection of these variables are explained in the subsequent sections.

3.1.1. Audit related variables

The relationship between (non-)audit fees and restatements is ambiguous. Lobo and Zhao (2013) find a negative association between audit effort, measured by audit fees, and restatements, because more audit effort means that the auditor is more likely to find errors. Nevertheless, Li and Lin (2005) find a positive association between audit fees and restatements, because higher audit fees increases the economic bond between client and auditor which impairs the auditor independence and lowers the audit quality. Further, the negative relationship between non-audit fees and restatements could exist due to knowledge spillover (Kinney, Palmrose & Scholz, 2004; Wahab, Gist & Majid, 2014). However, the economic dependence on non-audit fees may decrease the willingness of auditors to challenge misstatements which can increase the likelihood of a restatement (Kinney et al., 2004). Four variables are included to examine, absolutely and relatively, the relationship between (non-)audit fees and restatements: (1) the audit fees, (2) the non-audit fees, (3) the ratio between non-audit fees and audit fees and (4) the ratio between non-audit fees and total fees (Srinidhi & Gul, 2006; Wahab et al., 2014).

Eshleman and Guo (2014) conclude that clients of big 4 auditors are less likely to issue a restatement compared to clients of other auditors. Therefore, the big 4 auditor variable is equal to one whenever the auditor is Deloitte, EY, KPMG or PwC. Prior to 2002, the variable is also equal to one in case that the auditor is Arthur Anderson, which was a Big 5 auditor, but render their audit license in 2002 due to accounting scandals (Ainslie, 2006). Further, the change of the auditor is included, because Lazer, Livnat and Tan (2004) find more restatements for firms that switched auditors compared to companies that did not switch to another auditor.

Furthermore, Stanley and DeZoort (2007) find that short auditor tenure increases the likelihood of a restatement due to the lack of specific client knowledge, while Khang (2016) finds that short auditor tenure increases audit quality because of more independence in the early years which decreases the likelihood of a restatement (Knechel et al., 2013). The auditor tenure is included as continuous variable and as dummy variables, since some supervised machine learning algorithms prefer continuous variables over dummy variables, while other algorithms prefer dummy variables. Hence, the auditor tenure is trained in the supervised machine learning algorithms with both the continuous variable and the dummy variables.

In addition, auditor industry specialization decreases the likelihood of a restatement (Romanus, Maher & Fleming, 2008). The auditor is an industry specialist when their market share, measured by audit fees, is in the top three of an industry (DeFond, Francis & Wong, 2000). Finally, the receipt of a qualified audit opinion or going concern opinion is often followed by various negative events, such as restatements, bankruptcy or auditor changes (Cipriano, Hamilton & Vandervelde, 2017). Therefore, the qualified audit- and going concern opinions are included as dummy variables.

3.1.2. Firm characteristics

Dechow, Ge, Larson and Sloan (2011) develop an extensive framework of variables that could lead to material misstatements which are proxied by restatements (DeFond and Zhang, 2014). The accounting numbers are often misstated through accruals, thus it is expected that the working capital accruals, change in receivables and change in inventory are higher in restatement years (Dechow et al., 2011). Furthermore, firms with a large proportion of soft assets are more likely to have a misstatement (Barto & Simko, 2002).

The change in cash sales and the change in cash margin examine whether companies misstate the financial statements to hide bad performance, where a decline can be made up by managers to boost the accruals (Dechow et al., 2011). The deferred tax expense is the difference between the accounting income and the taxable income and could imply more manipulation of the accounting income. The abnormal change in employees and the abnormal change in order backlog also examine whether managers reduce the amount of employees or reduce the amount of order backlog to hide bad performance.

The leverage, ex-ante financing need, market-to-book ratio and price-to-earnings ratio describe the incentive of companies to maintain a high stock price. First, a high stock price will reduce the cost of capital. Therefore firms with high leverage and firms with an ex-ante financing need are more likely to misstate their financial statements to get a high stock price and raise capital on more favourable terms. Second, the management compensation is dependent on the stock price performance which means that higher market-to-book ratios and price-to-earnings ratios can be used to hide bad performance.

Furthermore, Palmrose et al. (2004) describe that companies with restatements are smaller, less profitable, less likely to have an audit committee and have higher debt. Therefore, the size of the firm is included in terms of total assets and market capitalization, the profitability is represented by the return on assets and return on equity (Alali & Wang, 2017). The presence of an audit committee is included as dummy variable and the total debt is added to the dataset. In addition, the board size is included, because Hasnan, Marzuki and Shuhidan (2017) find that board size and the likelihood of a restatement have a negative relationship due to the improvement of monitoring when the board size increases.

Various researches over the years have shown that computer-, retail- and service firms have more restatements than other industries relative to their share in the sample (Bertomeu et al., 2020; Dechow et al., 2011). The industry variable is an indicator variable, where the Standard Industrial Classification (SIC) codes are used to divide the firms into sixteen different industries. Lastly, the shift from rules-based standards (e.g. U.S. GAAP) to more principle-based standards (e.g. IFRS) may increase the amount of restatements, because auditors will become more tolerant of management judgement (Plumlee & Yohn, 2010). Therefore, the accounting standard variable is equal to one whenever the accounting standard is IFRS.

Table 1 Variable definitions

Variable	Abbreviation	Description	Source
Restatement	res	A dummy variable equal to one if the firm files a restatement in year t and zero otherwise.	Audit Analytics - Restatements
Audit related variables			
Audit fees	audit_fees	Audit fees	Audit Analytics – Audit Fees
Non-audit fees	non_audit_fees	Non-audit fees	Audit Analytics – Audit Fees
Fee ratio	fee_ratio	The ratio between non-audit fees and audit fees.	Audit Analytics – Audit Fees
Non-audit fee ratio	nfee_ratio	The ratio between non-audit fees and total fees.	Audit Analytics – Audit Fees
Big 4 auditor	big_4	A dummy variable equal to one if the auditor is Deloitte, EY, KPMG or PwC (or Arthur Andersen until 2002) and zero otherwise.	Compustat – North America – Fundamentals Annual
Auditor change	auditor_change	A dummy variable equal to one if the firm changes their auditor and zero otherwise.	Compustat – North America – Fundamentals Annual
Auditor tenure	auditor_tenure	The duration of the relationship between auditor and client.	Compustat – North America – Fundamentals Annual
Short auditor tenure	short_tenure	A dummy variable equal to one if the auditor tenure is three years or less and zero otherwise.	Compustat – North America – Fundamentals Annual
Medium auditor tenure	medium_tenure	A dummy variable equal to one if the auditor tenure is between four and eight years and zero otherwise.	Compustat – North America – Fundamentals Annual
Long auditor tenure	long_tenure	A dummy variable equal to one if the auditor tenure is nine years or longer and zero otherwise.	Compustat – North America – Fundamentals Annual
Auditor industry specialization	specialist	A dummy variable equal to one if the market share is in the top three of an industry.	Auditor Analytics – Audit Fees and Compustat – North America – Fundamentals Annual
Qualified audit opinion	qao	A dummy variable equal to one if the auditor issued a qualified audit opinion and zero otherwise.	Compustat – North America – Fundamentals Annual
Going concern opinion	gc	A dummy variable equal to one if the auditor issued a going concern opinion and zero otherwise.	Audit Analytics – Audit Opinions
Firm characteristics			
Working capital accruals	WC_acc	(Δcurrent assets - Δcash and short-term investments) - Δcurrent liabilities - Δdebt in current liabilities - Δtaxes payable) / average total assets	Compustat – North America – Fundamentals Annual
Change in receivables	ch_rec	Δaccounts receivable / average total assets	Compustat – North America – Fundamentals Annual

Change in inventory	ch_inv	Δinventory / average total assets	Compustat – North America – Fundamentals Annual
Percentage of soft assets	soft_assets	(total assets – PP&E – cash and cash equivalents) / total assets	Compustat – North America – Fundamentals Annual
Change in cash sales	ch_cs	Percentage change in cash sales (sales - Δaccounts receivables).	Compustat – North America – Fundamentals Annual
Change in cash margin	ch_cm	Percentage change in cash margin (1 – ((cost of goods sold - Δinventory + Δaccounts payable) / (sales - Δaccounts Receivable))).	Compustat – North America – Fundamentals Annual
Deferred tax expense	tax	Deferred tax expense_t / total assets_{t-1}	Compustat – North America – Fundamentals Annual
Abnormal change in employees	ch_emp	Percentage change in number of employees minus the percentage change in total assets.	Compustat – North America – Fundamentals Annual
Abnormal change in order backlog	ch_back	Percentage change in order backlog minus the percentage change in sales.	Compustat – North America – Fundamentals Annual
Leverage	lev	Long-term debt / total assets	Compustat – North America – Fundamentals Annual
Ex-ante financing need	exfin	A dummy variable equal to one if (CFO – past three year average capital expenditures) / current assets < - 0.5 and zero otherwise.	Compustat – North America – Fundamentals Annual
Market-to-book ratio	mb	Market capitalization / book value of equity	Compustat – North America – Fundamentals Annual
Price-to-earnings ratio	pe	Market value per share / earnings per share	Compustat – North America – Fundamentals Annual
Total assets	ta	Total assets	Compustat – North America – Fundamentals Annual
Market capitalization	mc	Stock price * shares outstanding	Compustat – North America – Fundamentals Annual
Return on assets	roa	Earnings / average total assets	Compustat – North America – Fundamentals Annual
Return on equity	roe	Earnings / average shareholders' equity	Compustat – North America – Fundamentals Annual
Audit committee	ac	A dummy variable equal to one if the firm has an audit committee and zero otherwise.	BoardEx – Board and Director Committees
Total debt	td	Total debt	Compustat – North America – Fundamentals Annual
Board size	board_size	Board size	BoardEx – Organization Summary - Analytics
Industry	industry	The distinction of the SIC code in different industries: 1. Agriculture = 100 – 999 2. Mining and Construction = 1000 – 1299 and 1400 – 1999 3. Food and Tobacco = 2000 – 2141 4. Textiles and Apparel = 2200 – 2399 5. Lumber, Furniture and Printing = 2140 – 2796 6. Chemicals = 2800 – 2824 and 2840 – 2899 7. Refining and Extractive = 1300 – 1399 and 2900 – 2999 8. Durable Manufacturers = 3000 – 3569 and 3580 – 3669 and 3680 – 3999	Compustat – North America – Fundamentals Annual

		9. Computers = 3570 – 3579 and 3670 – 3679 and 7370 – 7379 10. Transportation = 4000 – 4899 11. Utilities = 4900 – 4999 12. Retail = 5000 – 5999 13. Services = 7000 – 7369 and 7380 – 9999 14. Banks and Insurances = 6000 – 6999 15. Pharmaceuticals = 2830 – 2836 and 3829 – 3851 0. Other Industries = if the SIC code is not in 1 until 15	
Accounting standard	acctstd	The variable is equal to one if the accounting standard is IFRS and zero otherwise.	Compustat – North America – Fundamentals Annual

3.1.3. Data collection

The variables are extracted from six different datasets within Audit Analytics, Compustat and BoardEx. The last column of Table 1 indicates which dataset is used to calculate each variable. The dependent variable, restatements, needs some more explanation. The restatement variable is retrieved from the Audit Analytics Restatement Dataset which consists of the financial restatement disclosures of all SEC registrants since 1 January 2000. The restatement variable is equal to one in every year that a firm files a restatement.[1]

The construction of the final dataset starts in the six separate datasets by cleaning the datasets and creating the variables in each dataset. By cleaning the datasets is meant that the duplicated observations are removed and each firm has one observation per year. As mentioned before, the sample period runs from 2000 and 2019. The starting year is 2000, because the Restatement Dataset of Audit Analytics starts at 1 January 2000. The last year is 2019, because not all variables were available for 2020, but in this way the most recent data is used. Lastly, the six datasets are merged into one dataset which resulted in 90,000 observations of 10,762 firms.

3.2. Pre-process the dataset

The dataset needs to be pre-processed in such a way that the data can be used in the supervised machine learning algorithms. First, the observations with an infinite- or not a number value are removed, because these values cause problems in the algorithms. The final dataset consists of 78,917 observations of 10,078 firms after removing the observations with an infinite value and/or a not a number value.

The dataset contains a number of variables that have missing values (e.g. going concern opinion and board size). The missing values are computed with the help of imputation where the missing values are estimated based on the existing part of the dataset. Furthermore, several variables, such as the fee ratio and the change in cash margin, have outliers (see Appendix A, Figure 2). The variables with outliers are winsorized at the 1% and 99% percent to reduce the effect of the outliers in the results. By winsorization is meant that data below the 1^{st} percentile is set to the 1^{st} percentile and data above the 99^{th} percentile is set to the 99^{th} percentile. Lastly, the dataset contains variables with skewed distributions (see Appendix A, Figure 3). These variables are standardized with a log transformed variable which pulls the extreme values to the centre.

Before evaluating the descriptive statistics, the dataset is divided into the training- and the testing set, because the data of the testing set remains unseen until the final algorithms are evaluated (Elkan, 2012). The division is a stratified random split which ensures that the event rate, the restatement rate, is equal in both the training- and testing set. The data is divided into 80 percent and 20 percent for the training- and testing set respectively which is suggested in the literature (e.g. Gholamy, Kreinovich & Kosheleva, 2018; Rácz, Bajusz & Héberger, 2021).

[1] Audit Analytics and Bertomeu et al. (2020) confirmed by email that the period start and end dates encompass the duration of the specific restatement. For example, a restatement period starts at 1/1/2011 and ends at 31/12/2013, then the restatement variable is equal to one in 2011, 2012 and 2013.

3.3. Descriptive statistics

The descriptive statistics of the training set are provided in Table 2. The dataset contains 63,135 observations of 9,825 firms where 13.5 percent of the firm-year observations filed a restatement. The descriptive statistics related to the audit related variables reveal that 68.3 percent of the firm-year observations is audited by a big 4 auditor, 0.02 percent of the firm-year observations received a qualified audit opinion and 9.5 percent of the firm-year observations received a going-concern opinion. Besides, the average auditor tenure is approximately 8 years with a maximum auditor tenure of 46 years. The descriptive statistics concerning the firm characteristics illustrate that 94.9 percent of the firm-year observations has an audit committee and 4.7 percent of the firm-year observations uses IFRS as accounting standard. Lastly, the average board size is approximately 8 with a maximum board size of 23.

The trend of the restatements over the years is plotted in Figure 4 and listed in Table 3. The percentage of restatements varies from 4.05 percent to 21.79 percent. The restatement rate peaks with 21.79 percent in 2004, but decreases to 11.20 percent in 2008. Since 2008, the restatement rate increases steadily to 14.76 percent in 2012. After 2012, the restatement rate decreases to 4.05 percent in 2019. The peak of 2004 could be because of stricter SEC inspections after the implementation of SOX 404. The increase of the restatement rate since 2008 could be due to stricter SEC inspections after the financial crisis in 2007 (Zhang, 2019).

Figure 5 and Table 4 provide some additional information about the restatements across different industries. Bertomeu et al. (2020) and Dechow et al. (2011) find that the computer-, retail- and services industry have disproportionately more restatements compared to their share in the sample. Figure 5 and Table 4 confirm that the computer-, retail- and service industry have disproportionately more restatements relative to their share in the sample which is similar to the aforementioned researches. Moreover, in this sample, the durable manufacturers- and transportation industry have relatively more restatements compared to their share in the sample.

The left graph of Figure 6 illustrates that two third of the firm-year observations is audited by big 4 auditors and that two third of the restatements can be attributed to big 4 auditors. However, the amount of restatements for big 4 auditors is not significantly smaller than for non-big 4 auditors (0.566 > 0.1). This differs from the conclusion of Eshleman and Guo (2014) who state that clients of big 4 auditors are less likely to issue a restatement. The right graph of Figure 6 illustrates that the amount of restatements is relatively lower for firm-year observations with IFRS as accounting standard than for firm-year observations with other accounting standards. Also statistically, the amount of restatements for firm-year observations with IFRS as accounting standard is not significantly larger than the amount of restatements for firm-year observations with other accounting standards (1.000 > 0.1). This is not in line with Plumlee and Yohn (2010) who state that the shift from rules-based standards (e.g. U.S. GAAP) to more principle-based standards (e.g. IFRS) may increase the amount of restatements. However, domestic companies in the United States must use U.S. GAAP, while IFRS is required or permitted for listings by foreign companies (IFRS, 2021). Thus, the shift towards more principle-based standards has not yet taken place in the United States, which makes it necessary to be cautious when drawing conclusions.

Table 2 Descriptive statistics

Variable	**N**	**Mean**	**Median**	**St. Dev.**	**Min**	**Max**
Restatement	63,135	0.135	0	0.342	0	1
Audit related variables						
Audit fees	63,135	13.305	13.331	1.512	9.940	17.070
Non-audit fees	63,135	10.074	11.339	4.458	0.000	16.195
Fee ratio	63,135	0.349	0.160	0.566	0.000	3.489
Non-audit fee ratio	63,135	0.189	0.138	0.188	0.000	0.983
Big 4 auditor	63,135	0.683	1	0.465	0	1
Auditor change	63,135	0.108	0	0.310	0	1
Auditor tenure	63,135	8.300	6	7.783	1	46
Short auditor tenure	63,135	0.315	0	0.465	0	1
Medium auditor tenure	63,135	0.335	0	0.472	0	1
Long auditor tenure	63,135	0.349	0	0.477	0	1
Auditor industry specialization	63,135	0.532	1	0.499	0	1
Qualified audit opinion	63,135	0.0002	0	0.013	0	1
Going concern opinion	63,135	0.095	0	0.294	0	1
Firm characteristics						
Working capital accruals	63,135	-0.008	0.001	0.149	-0.910	0.490
Change in receivables	63,135	0.014	0.005	0.073	-0.252	0.333
Change in inventory	63,135	0.004	0.000	0.038	-0.152	0.182
Percentage of soft assets	63,135	0.554	0.581	0.293	0.015	0.987
Change in cash sales	63,135	0.156	0.053	1.848	-8.088	12.059
Change in cash margin	63,135	-0.049	-0.015	2.142	-11.896	12.063
Deferred tax expense	63,135	-0.001	0.000	0.019	-0.103	0.068
Abnormal change in employees	63,135	-0.092	-0.030	0.545	-3.542	1.349
Abnormal change in order backlog	63,135	0.097	-0.003	0.732	-1.488	4.260
Leverage	63,135	0.197	0.107	0.247	0.000	1.333
Ex-ante financing need	63,135	0.157	0	0.364	0	1
Market-to-book ratio	63,135	2.612	1.656	7.266	-30.376	43.994
Price-to-earnings ratio	63,135	10.029	9.844	48.999	-209.000	261.000
Total assets	63,135	19.937	20.103	2.683	12.680	26.024
Market capitalization	63,135	19.655	19.689	2.456	13.759	25.253
Return on assets	63,135	-0.168	0.011	0.666	-4.843	0.391
Return on equity	63,135	-0.053	0.065	1.302	-7.017	6.669
Audit committee	63,135	0.949	1	0.221	0	1
Total debt	63,135	14.618	17.772	8.037	0.000	24.216
Board size	63,135	8.211	8	2.664	1	23
Accounting standard	63,135	0.047	0	0.212	0	1

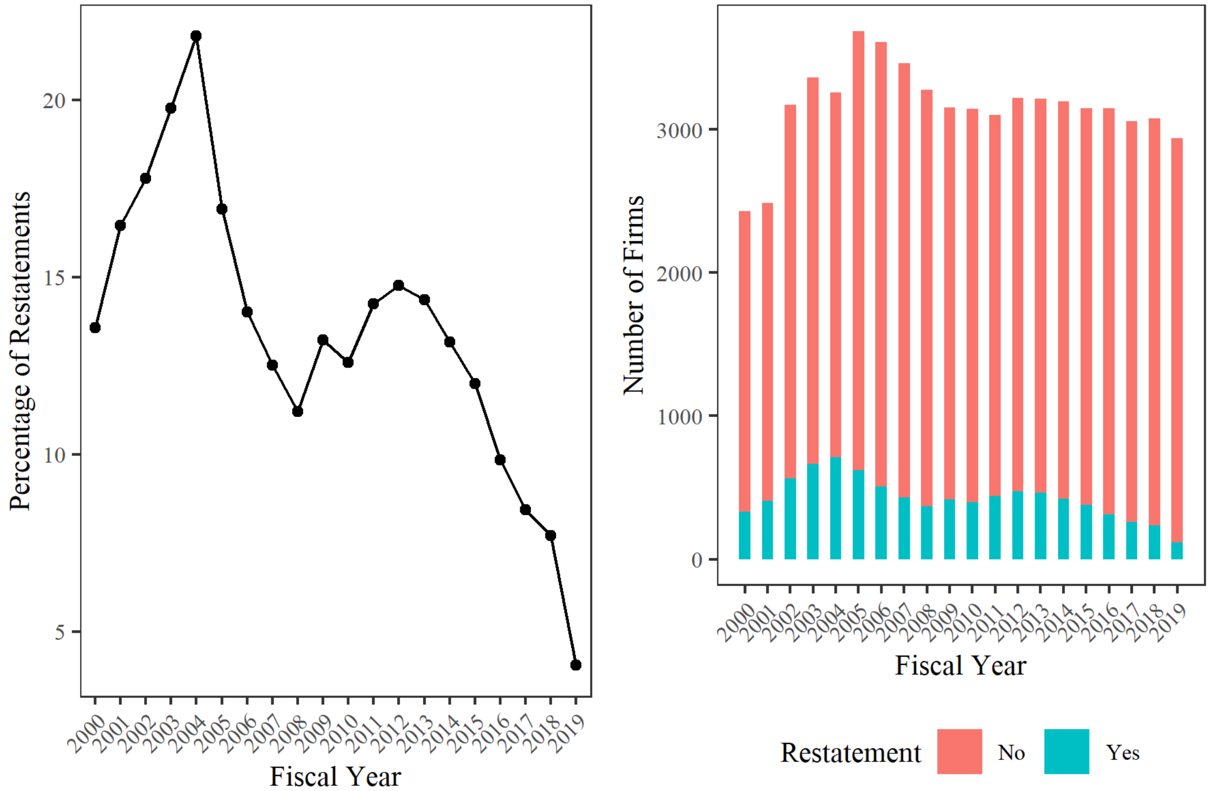

Figure 4 The percentage of restatements over the years

Table 3 The amount of restatements over the years

Fiscal year	**Number of restated firms**	**Percentage of the total restatements**	**Compustat firms**	**Percentage of Compustat firms**
2000	330	3.87	2,431	13.57
2001	409	4.80	2,487	16.45
2002	564	6.62	3,171	17.79
2003	664	7.79	3,360	19.76
2004	710	8.33	3,258	21.79
2005	623	7.31	3,685	16.91
2006	506	5.94	3,608	14.02
2007	433	5.08	3,462	12.51
2008	367	4.31	3,278	11.20
2009	417	4.89	3,154	13.22
2010	396	4.65	3,144	12.60
2011	442	5.19	3,101	14.25
2012	475	5.57	3,218	14.76
2013	462	5.42	3,214	14.37
2014	421	4.94	3,197	13.17
2015	378	4.44	3,149	12.00
2016	310	3.64	3,147	9.85
2017	258	3.03	3,058	8.44
2018	237	2.78	3,075	7.71
2019	119	1.40	2,937	4.05
Total	8,521	100.00	63,135	13.42

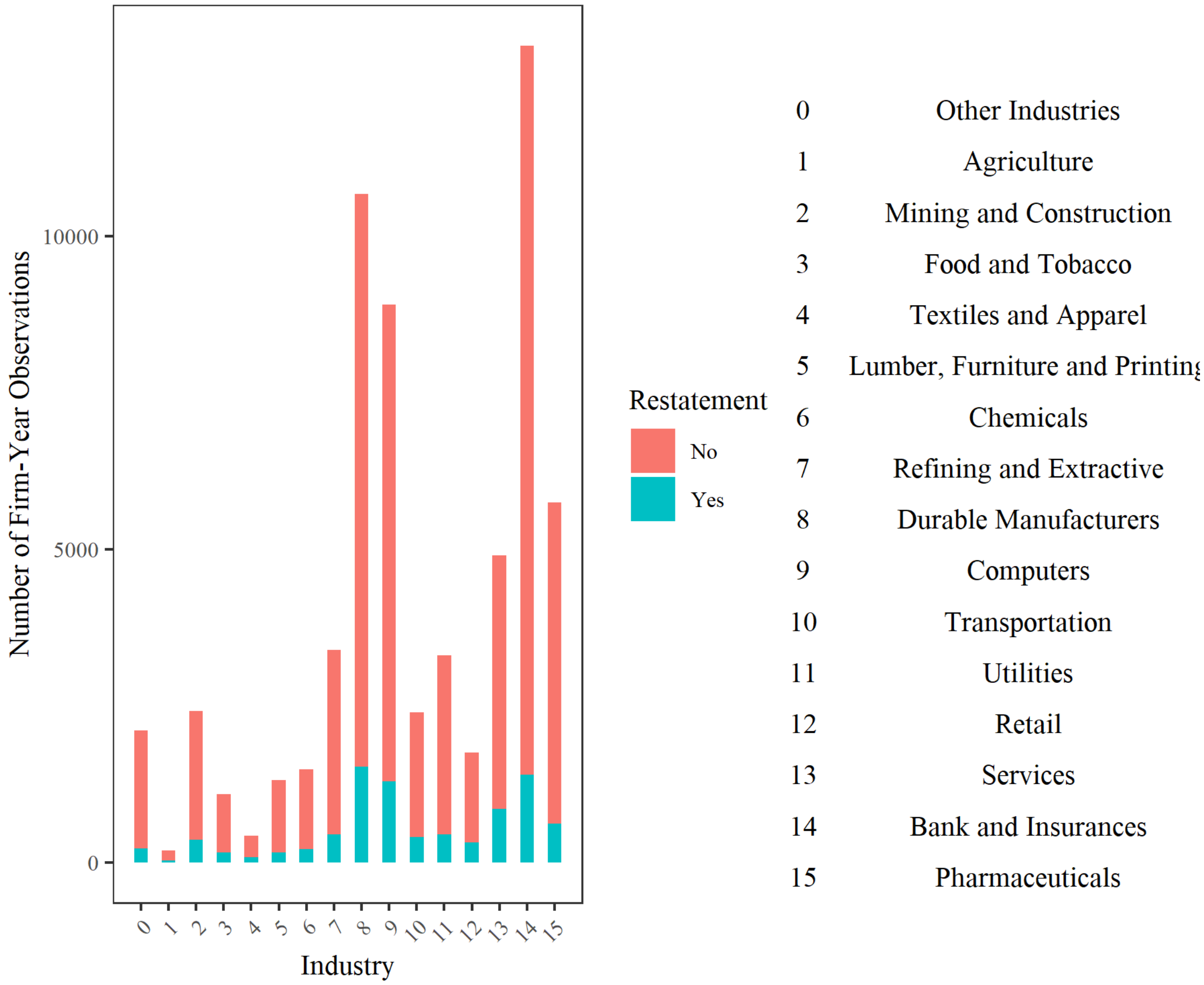

Figure 5 Firm-year restatements by industry

Table 4 Firm-year restatements by industry

Industry	Number of restatements	Percentage	Number of firm-years	Percentage in sample
Agriculture	26	0.31	192	0.30
Mining and Construction	360	4.22	2,420	3.83
Food and Tobacco	153	1.80	1,092	1.73
Textiles and Apparel	87	1.02	430	0.68
Lumber, Furniture and Printing	154	1.81	1,311	2.08
Chemicals	214	2.51	1,489	2.36
Refining and Extractive	449	5.27	3,390	5.37
Durable Manufacturers	*1,523*	*17.87*	*10,670*	*16.90*
Computers	**1,295**	**15.20**	**8,900**	**14.10**
Transportation	*407*	*4.78*	*2,397*	*3.80*
Utilities	444	5.21	3,309	5.24
Retail	**321**	**3.77**	**1,754**	**2.78**
Services	**857**	**10.06**	**4,902**	**7.76**
Bank and Insurances	1,396	16.38	13,034	20.64
Pharmaceuticals	618	7.25	5,743	9.10
Other Industries	217	2.55	2,101	3.33
Total	8,521	100.00	63,135	100.00

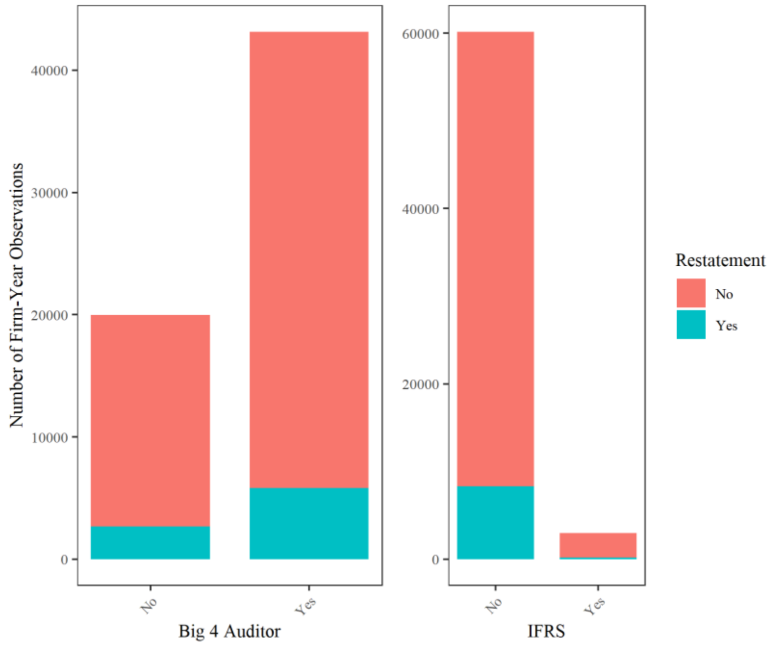

Figure 6 Firm-year restatements by auditor and accounting standard

Lastly, the means of the variables for firms without restatements and firms with restatements are compared with each other in Table 5. The audit fees (0.000 < 0.01) and non-audit fees (0.000 < 0.01) are significantly higher for firms with restatements which indicates a low independence between auditor and client (Kinney et al., 2004; Li & Lin, 2005). The amount of qualified audit opinions is not significantly higher for firms with restatements (0.164 > 0.10). Nevertheless, the amount of going concern opinions is significantly higher for firms with restatements (0.000 < 0.01) which confirms that going concern opinions are often followed by negative events such as restatements (Cipriano et al., 2017).

Dechow et al. (2011) conclude that accounting numbers are often misstated through accruals, which means that the accruals should increase more for firms with restatements. Table 5 affirms that the change in receivables (0.011 < 0.05) and the change in inventory (0.001 < 0.01) significantly increases more for firms with restatements compared to firms without restatements, but the working capital accruals are insignificant. The percentage of the soft assets (0.000 > 0.01) is significantly higher for firms with restatements, which confirms the conclusion of Barto and Simko (2002) who state that firms with a higher percentage of soft assets are more likely to file a restatement. Dechow et al. (2011) state that the change in cash sales and the change in cash margin should decrease more for firms with restatements, so firms can hide bad performance. Nevertheless, the variables are both increasing and insignificant. They also expect that the deferred tax expenses increase during the restatement years which implies more manipulation of the accounting income. However, the deferred tax expenses is insignificant. The abnormal change in employees is significantly higher for firms with restatements (0.000 < 0.01) which indicates that managers reduce the amount of employees to hide bad performance. Dechow et al. (2011) also conclude that managers reduce the amount of order backlog to hide bad performance, but Table 5 shows that the abnormal change in order backlog is insignificant.

Table 5 Differences between non-restatement and restatement years

	Non-restatement		Restatement			
Variable	**N**	**Mean**	**N**	**Mean**	**Difference in mean**	**One tailed p-value**
Audit fees	54,613	13.291	8,521	13.401	-0.110	0.000***
Non-audit fees	54,613	10.029	8,521	10.365	-0.336	0.000***
Qualified audit opinion	54,613	0.0001	8,521	0.0004	-0.0003	0.164
Going concern opinion	54,613	0.093	8,521	0.112	-0.019	0.000***
Working capital accruals	54,613	-0.007	8,521	-0.011	0.004	0.978
Change in receivables	54,613	0.014	8,521	0.016	-0.002	0.011**
Change in inventory	54,613	0.004	8,521	0.005	-0.001	0.001***
Percentage of soft assets	54,613	0.552	8,521	0.572	-0.020	0.000***
Change in cash sales	54,613	0.146	8,521	0.220	-0.074	0.999
Change in cash margin	54,613	-0.054	8,521	-0.016	-0.038	0.934
Deferred tax expense	54,613	-0.0005	8,521	-0.001	0.0005	0.999
Abnormal change in employees	54,613	-0.087	8,521	-0.126	0.039	0.000***
Abnormal change in order backlog	54,613	0.096	8,521	0.101	-0.005	0.695
Leverage	54,613	0.193	8,521	0.225	-0.032	0.000***
Ex-ante financing need	54,613	0.157	8,521	0.159	-0.002	0.347
Market-to-book ratio	54,613	2.633	8,521	2.483	0.150	0.951
Price-to-earnings ratio	54,613	10.058	8,521	9.849	0.209	0.634
Total assets	54,613	19.956	8,521	19.814	0.142	0.000***
Market capitalization	54,613	19.668	8,521	19.577	0.091	0.000***
Return on assets	54,613	-0.162	8,521	-0.209	0.047	0.000***
Return on equity	54,613	-0.055	8,521	-0.041	-0.014	0.784
Audit committee	54,613	0.948	8,521	0.950	-0.002	0.718
Total debt	54,613	14.562	8,521	14.979	-0.417	0.000***
Board size	54,613	8.246	8,521	7.983	0.263	0.000***

Note. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$

In addition, the leverage is significantly higher for restatement firms (0.000 < 0.01), but the ex-ante financing need is not significantly higher for firms with restatements. Dechow et al. (2011) confirm that firms with a high leverage are more likely to misstate the financial statements to maintain a high stock price and raise capital on favourable terms. They also conclude that management compensation is dependent on the stock price performance which means that managers could manipulate the numbers to get a higher market-to-book ratio and/or higher price-to-earnings ratio in order to hide bad performance. Table 5 does not support this, because the market-to-book ratio (0.951 > 0.10) and price-to-earnings ratio (0.634 > 0.10) are both lower in the years with restatements compared to the years without restatements.

Further, the findings of Palmrose et al. (2004) are partially confirmed, because firms with restatements are significantly smaller, considering the total assets (0.000 < 0.01) and market capitalization (0.000 < 0.01), and have significantly higher debts (0.000 < 0.01) compared to firms without restatements. Nevertheless, the firms with restatements are more likely to have an audit committee (0.718 > 0.10), which is inconsistent with the results of Palmrose et al. (2004). Further, Palmrose et al. (2004) state that firms with restatements are less profitable, which is confirmed by the return on assets (0.000 < 0.01), but not by the return on equity (0.784 > 0.10). Lastly, the conclusion of Hasnan et al. (2017) is confirmed, because the average board size is significantly lower for firms with restatements (0.000 < 0.01).

3.4. Imbalanced dataset

The descriptive statistics in Table 2 illustrate that the dataset is imbalanced, since the amount of restatements is in minority. An unbalanced dataset is considered as a problem in machine learning, because algorithms have difficulty with recognizing patterns in the training phase, since most algorithms expect a balanced dataset (Lemaître, Nogueira & Aridas, 2017). Many approaches have been developed to deal with imbalanced datasets, such as random undersampling or random oversampling. Random undersampling is applied to the dataset, because random oversampling can lead to overfitting (Lemaître et al., 2017). Random undersampling entails randomly selecting observations in the majority class that are removed from the dataset, so that the amount of observations in the majority class is reduced to the amount of observations in the minority class.

3.5. Training- and testing set

The original dataset consists of 78,917 firm-year observations which is divided into 80 percent for the training set (i.e. 63,135 firm-year observations) and 20 percent for the testing set (i.e. 15,782 firm-year observations) with a stratified random split. The amount of restatements is 8,463 in the training set. The unbalanced training set is transformed to a balanced training set by randomly reducing the amount of non-restatements to 8,463, which results in a training set with 16,926 firm-year observations. The testing set is not subjected to undersampling, because the trained algorithms should be evaluated on reality.

The machine learning algorithms should be evaluated in between to investigate whether the accurate parameters are chosen. But, the interim evaluation should not be conducted on the testing set, since the testing set remains unseen until the algorithms are fully trained. Therefore, there should be a validation set to examine the algorithms in between. Hence, stratified 10-fold cross validation is applied to the training set, which is illustrated in Figure 7. The training set is divided in ten random folds, where one fold is used as validation set and the other folds are used as training set where the restatement rate is equal in every fold. The cross validation process is repeated ten times, where every fold is used once as validation set. The average of the ten results is used for the evaluation metrics. In summary, the supervised machine learning algorithms are trained on the training set and the interim evaluation is performed with the help of stratified 10-fold cross validation. The predictions are made for the testing set when the algorithms are fully trained and the performance of the algorithms is evaluated by different evaluation metrics.

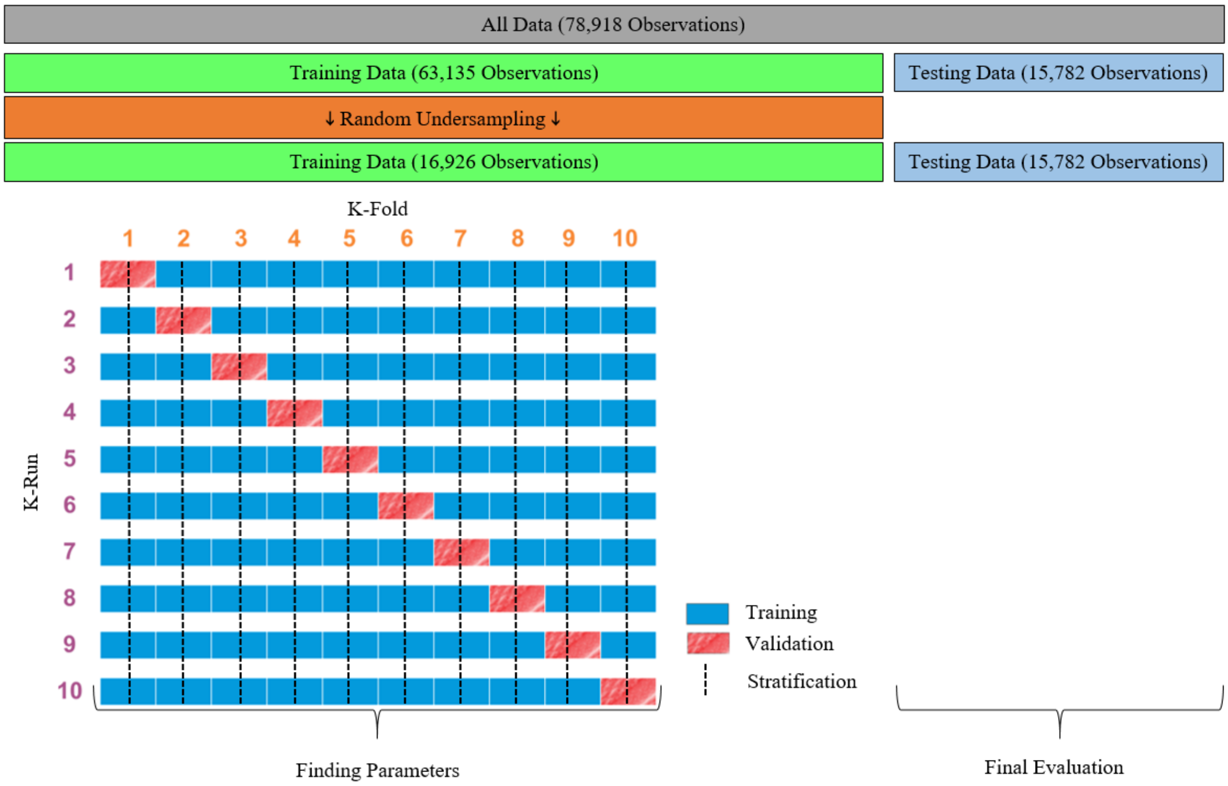

Figure 7 Example of stratified 10-fold cross validation

3.6. Supervised machine learning algorithms

The following sections briefly explain the six supervised machine learning algorithms that are utilized in this research.

3.6.1. Artificial neural network

The artificial neural network is inspired by the brain function and is the best algorithm to predict restatements according to Dutta et al. (2017). The artificial neural network consists of an input layer, hidden layer and output layer. The inputs are converted into outputs as a function of the weighted sum of the inputs (Dutta et al., 2017; Osisanwo et al., 2017; Zhang, 2019). The inputs are the variables from Table 1 and the output is either a restatement or a non-restatement. The weights of the inputs need to be trained in the hidden layer through backpropagation to minimize the error between the prediction and the actual value (Dutta et al., 2017; Zhang, 2019).

3.6.2. K-nearest neighbours

The k-nearest neighbours is the best in predicting restatements according to Bertomeu et al. (2020). The algorithm uses data points in the dataset which are grouped in several classes (Crisci, Ghattas & Perara, 2012). The classes in this research are restatements and non-restatements. The algorithm identifies the k-nearest neighbours of a new instance and decides in which class the new instance belongs (Alzubi et al., 2017). The number of nearest neighbours (k) is preferably odd since the majority rule applies. The rule of thumb for the choice of k is equal to the square root of the number of instances (Hassanat, Abbadi & Altarawneh, 2014).

3.6.3. Decision tree and random forest

The decision tree is a commonly used algorithm in classification problems and is easy to understand (Alzubi et al., 2017; Osisanwo et al., 2017). The decision tree splits the data based on features in the dataset. The decision tree consists of decision nodes and leaf nodes. A decision node represents a decision, e.g. the auditor is a big four auditor. The leaf nodes are the possible outcomes, e.g. yes or no. The objective is to create a decision tree with a minimal number of nodes. The prediction, whether the instance will file a restatement or not, is shown at the bottom of the decision tree. The random forest is largely similar to the decision tree and is according to Zhang (2019) the best algorithm to predict restatements. The difference with the decision tree is that the random forest creates many decision trees with a random subset of the data (Alzubi et al., 2017). The results of the different decision trees are merged together and then the prediction, whether the instance will file a restatement or not, is made.

3.6.4. Naïve Bayes

Naïve Bayes is a well-known statistical learning algorithm which is useful for both large datasets and high dimensional datasets (Alzubi et al., 2017; Dutta et al., 2017; Osisanwo et al., 2017). The algorithm is based on the Bayes Theorem with the naïve assumption, which indicates that there is rigid independence between the variables. (Osisanwo et al., 2017; Zhang, 2019). The model is as follows:

$$P(Restatement|X_1, \dots, X_{35}) = \frac{P(Restatement) * P(X_1, \dots, X_{35}|Restatement)}{P(X_1, \dots, X_{35})}$$

The probability of having a restatement conditional the variables X_1 to X_{35} (see Table 1) can be calculated by the probability of three events. The first event is the probability of having a restatement. The second event is the probability of being X_1 to X_{35} conditional having a restatement and the last event is the probability of being X_1 to X_{35}.

3.6.5. Logistic regression

The logistic regression is one of the most commonly used algorithms in discrete data analytics (Draeger et al., 2016; Osisanwo et al., 2017). The logistic regression models the relationship between the dependent variable and the independent variable(s) (Alzubi et al., 2017; Smith, 2017). The logistic regression is used to predict categorical dependent variables and can only predict values between zero and one (Osisanwo et al., 2017). In this research, the dependent variable restatement is categorical and therefore the logistic regression is used. The model is as follows:

$$Restatement_{it} = \beta_0 + \sum_{1}^{14} \beta \ (audit\ related\ variables)_{it} + \sum_{15}^{35} \beta \ (firm\ characteristics)_{it} + \varepsilon_{it}$$

The audit related variables and firm characteristics for firm *i* in year *t* from Table 1 are used to predict the likelihood of filing a restatement for firm *i* in year *t*. The model should not contain multicollinearity, which occurs when two or more variables are highly correlated, because the performance of the logistic regression decreases due to multicollinearity.

3.7. Evaluation metrics

The last stage is to evaluate the performance of the different algorithms. The three evaluation metrics of this research are briefly explained in the subsequent sections.

3.7.1. Accuracy

The confusion matrix, shown in Table 6, summarizes the possible outcomes of the testing set (Dutta et al., 2017; Zhang, 2019). The instances are true positives or true negatives when the algorithm correctly predicts restatements or non-restatements. Nevertheless, the algorithm can also predict incorrectly and then the instances are false negatives or false positives. The accuracy can be derived from the confusion matrix by adding the correctly predicted instances (i.e. true positives and true negatives) and divide them by the total number of instances. The higher the accuracy, the better the performance of the algorithm.

Table 6 Example of a confusion matrix

	Actual restatement	**Actual non-restatement**
Predicted restatement	True positive	False positive
Predicted non-restatement	False negative	True negative

3.7.2. G-mean

Accuracy is one of the most commonly used evaluation measures. However, the accuracy can be misleading for imbalanced datasets, because the accuracy presumes that the true positives and true negatives are equally important (Akosa, 2017). Therefore, G-mean is selected as evaluation metric, because the G-mean tries to correct for the imbalance in the data. The G-mean is calculated as:

$$\sqrt{sensitivity * specificity}$$

Sensitivity is the percentage of the actual restatements that is correctly identified as restatement, while specificity is the percentage of the actual non-restatements that is correctly identified as non-restatement. The G-mean varies from zero to one, where a lower score indicates poor performance of the algorithm in classifying restatements even if the non-restatements are correctly classified (Akosa, 2017). Thus, the G-mean is important to avoid overfitting of the non-restatements and to avoid underfitting of the restatements, where a higher G-mean is desirable.

3.7.3. Area under the curve

The performance can be visualized by the Receiver Operating Characteristics (ROC) curve (Bertomeu et al., 2020; Dutta et al., 2017; Zhang, 2019). The ROC curve plots the true positive rate, which is the percentage of the actual restatements that is correctly identified as restatement (i.e. sensitivity), on the y-axis. The false positive rate, which is the percentage of the actual non-restatements that is identified as restatement, is plotted on the x-axis of the ROC curve.

Figure 8 is an example of a ROC curve. The performance of the ROC curve is optimal in the upper left corner (0, 1), because there the algorithm predicts every instance correct without making a mistake. The diagonal line represents a random guess whether an instance files a restatement or not. The corresponding evaluation measure is the AUC, which is a robust metric to evaluate imbalanced datasets (Wardhani, Rochayani, Iriany, Sulistyono & Lestantyo, 2019). The performance of the algorithm is better if the AUC is higher.

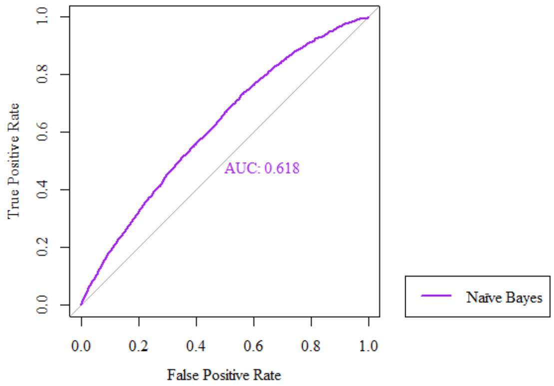

Figure 8 Example of a ROC curve

4. Results

In the present chapter, the results of the empirical analysis are presented. First, the results of the initial research are discussed, which are followed by the results of further research. The research question is answered by examining which supervised machine learning algorithm is the most accurate to predict audit quality when audit quality is proxied by restatements. In addition, the best variables to explain restatements, and thereby explain poor audit quality, are examined.

4.1. Initial results

The results of the initial research are presented in Table 7, Table 8, Table 9 and Figure 9, where Table 9 can be found in Appendix B. Table 7 shows the three evaluation metrics for every optimized supervised machine learning algorithm on the testing set. The ROC curve of each algorithm is displayed in Figure 9. Table 9 describes which variables are selected in every supervised machine learning algorithm to get the optimized model. Finally, Table 8 illustrates the ten best variables of the selected variables of each algorithm to explain restatements, where the statistical software identifies the relative importance of each variable to the model. The method to identify the relative importance of each variable differs per algorithm.

The most accurate supervised machine learning algorithm to predict restatements, according to accuracy, is the naïve Bayes which has an accuracy of 0.678. However, the accuracy is misleading for imbalanced datasets, which makes the G-mean and AUC more accurate evaluation metrics to evaluate imbalanced datasets, because they are robust to imbalanced datasets. The random forest scores the highest G-mean (0.627) and AUC (0.678). Furthermore, the ROC curve of the random forest is the closest to the optimal point of the ROC curve (0, 1). Altogether, the random forest is unanimously the most accurate supervised machine learning algorithm to predict restatements. This results is consistent with the finding of Zhang (2019).

Table 7 Initial results: testing results using balanced training set

Algorithm	Accuracy	G-mean	AUC
Artificial neural network	0.544	0.585	0.623
K-nearest neighbour	0.553	0.582	0.603
Decision tree	0.586	0.579	0.603
Random forest	0.585	0.627	0.678
Naïve Bayes	0.678	0.515	0.581
Logistic regression	0.576	0.577	0.606

Table 8 Initial results: top ten variables to explain restatements

Artificial neural network	K-nearest neighbour	Decision tree	Random forest	Naïve Bayes	Logistic regression
Market capitalization	Non-audit fees	Price-to-earnings ratio	Percentage of soft assets	Non-audit fees	Audit fees
Audit fees	Leverage	Percentage of soft assets	Audit fees	Leverage	Accounting standard
Total assets	Board size	Non-audit fees	Market capitalization	Board size	Board size
Industry	Audit fees	Leverage	Return on assets	Audit fees	Non-audit fee ratio
Big 4 auditor	Return on assets	Return on assets	Price-to-earnings ratio	Return on assets	Market capitalization
Accounting standard	Non-audit fee ratio	Return on equity	Abnormal change in employees	Non-audit fee ratio	Industry
Non-audit fees	Fee ratio	Market capitalization	Return on equity	Fee ratio	Leverage
Auditor industry specialization	Industry	Industry	Market-to-book ratio	Return on equity	Return on assets
Audit committee	Return on equity	Abnormal change in employees	Change in cash sales	Industry	Percentage of soft assets
Abnormal change in employees	Price-to-earnings ratio	Total assets	Total assets	Price-to-earnings ratio	Abnormal change in employees

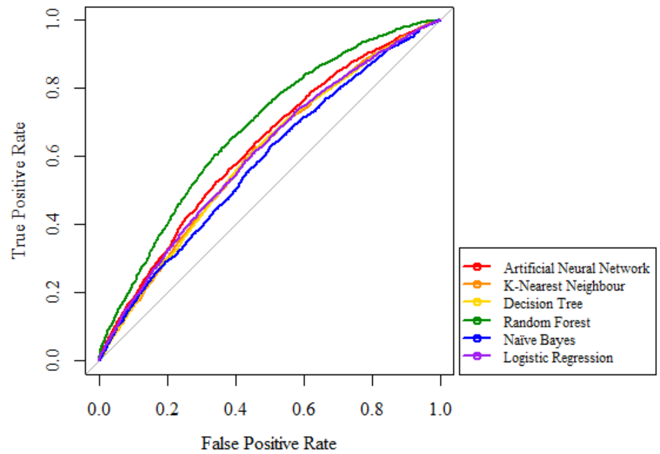

Figure 9 ROC curve of the initial results

Table 9 describes which variables are selected in the random forest to get the optimized model. Random forest consists of 22 firm characteristics and 10 audit related variables, where auditor tenure is included as continuous variable. The ten most predictive variables of restatements are listed in Table 8 and contain nine firm characteristics, which is 40.9 percent of the total firm characteristics in the model, and one audit related variable, which is 10 percent of the total audit related variables in the model. Unfortunately, random forest does not give the direction of the relationship between the independent variables and restatements. Random forest only mentions the variables that tend to enter the splits more often and thus have more relationship with restatements than the variables that have very little relationship with restatements.

Firstly, the accrual-based metrics, which are widely used in researches about audit quality, are surprisingly not represented in the predictive variables of audit quality. The percentage of soft assets is the most predictive variable of restatements in random forest, which confirms the finding of Barto and Simko (2002) and the findings of Table 5, namely that the percentage of soft assets between firm with restatements and firms without restatements differ. Furthermore, the size of firms, measured by total assets and market capitalization, and the profitability of firms, measured by return on assets and return on equity, are predictive variables of restatements. The findings of Palmrose et al. (2004) and Table 5 are confirmed by the random forest, namely that the size and profitability between firms with restatements and firms without restatements differ.

Besides, the price-to-earnings ratio, market-to-book ratio, abnormal change in employees and change in cash sales are predictive variables of restatements. Dechow et al. (2011) describe that these can be used by managers to hide bad performance by lower the amount of employees, lower the cash sales and higher the market-to-book ratio and higher the price-to-earnings ratio. The random forest confirms that these variables have a relationship with restatements, but this is partially confirmed by Table 5, which illustrates that only the abnormal change in employees differs between restated- and non-restated firms. Lastly, the audit fees are the only audit related variable in the predictive variables of restatements. Table 5 illustrates differences in audit fees between firms with restatements and firms without restatements as well.

In summary, the audit related variables are not the most predictive variables of audit quality in the initial research, which is in contrast to the findings of Zhang (2019) and Bertomeu et al. (2020). The audit related variables mainly contain dummy variables and the firm characteristics primarily consist of continuous variables. Random forest can split the firm characteristics more easily, because these have larger ranges and therefore it is possible that more firm characteristics ended in the top ten of predictive variables of audit quality.

4.2. Additional results

The initial results are quite satisfactory, but there is always room for improvement in the performance of the supervised machine learning algorithms. Therefore, further research is conducted by adding three macro-economic variables to the dataset. Hu et al. (2021) use machine learning to measure audit quality in China where they selected GDP, GDP growth and inflation as variables to control for their effect. The three variables appear in the top 30 of important variables to explain audit quality, where GDP is even in the top ten. Therefore, GDP, GDP growth and inflation are included in the dataset to examine whether these variables also explain restatements in the United States and whether the performance of the algorithms increases.

The additional results are presented in Table 10, Table 11, Table 12 and Figure 10, where Table 12 turns up in Appendix B. The structure of the tables and figure is the same as for the initial results. Thus, Table 10 shows the three evaluation metrics for each supervised machine learning algorithm on the testing set and Figure 10 displays the ROC curves. Table 12 describes which variables are used in each algorithm to get the optimized model. Lastly, Table 11 illustrates the ten best variables of the selected variables of each algorithm to explain restatements.

In general, the performance of the algorithms is increased, but the additional results do not differ from the initial results. Again, the naïve Bayes has the highest accuracy, which has increased from 0.678 in the initial research to 0.705 in the additional research. However, the G-mean and AUC are more appropriate evaluation measures, because these are robust to imbalanced datasets. Once more, the random forest has the highest G-mean (0.641), the highest AUC (0.696) and the ROC curve of the random forest still rises head and shoulders above the other ROC curves. Thus, the random forest is also the most accurate supervised machine learning algorithm to predict restatements in the additional research.

As shown in Table 12, the optimized random forest contains 22 firm characteristics, 3 macro-economic variables and 10 audit related variables, where auditor tenure is included as continuous variable. Table 11 lists the ten most predictive variables of restatements with one difference compared to the initial results. The total assets fall outside the top ten and is replaced by GDP. Therefore, the ten most predictive variables of restatements consist of eight firm characteristics, which is 36.4% of the total firm characteristics in the model, one macro-economic variable, which is 33.3% of the total macro-economic variables in the model, and one audit related variables, which is 10% of the total audit related variables in the model.

Table 10 Additional results: testing results using balanced training set

Algorithm	**Accuracy**	**G-mean**	**AUC**
Artificial neural network	0.604	0.603	0.648
K-nearest neighbour	0.546	0.584	0.617
Decision tree	0.589	0.600	0.625
Random forest	0.608	0.641	0.696
Naïve Bayes	0.705	0.537	0.621
Logistic regression	0.594	0.595	0.635

Table 11 Additional results: top ten variables to explain restatements

Artificial neural network	**K-nearest neighbour**	**Decision tree**	**Random forest**	**Naïve Bayes**	**Logistic regression**
Audit fees	GDP	GDP	Percentage of soft assets	GDP	GDP
Market capitalization	Non-audit fees	Audit fees	Audit fees	Non-audit fees	Audit fees
GDP	Leverage	Return on assets	Price-to-earnings ratio	Leverage	Board size
Big 4 auditor	Board size	Non-audit fees	Market capitalization	Board size	Accounting standard
Total assets	Audit fees	Price-to-earnings ratio	Return on assets	Audit fees	Market capitalization
Industry	Return on assets	Leverage	Abnormal change in employees	Return on assets	Big 4 auditor
Leverage	Non-audit fee ratio	Market capitalization	GDP	Non-audit fee ratio	Industry
Ex-ante financing need	Fee ratio	Percentage of soft assets	Return on equity	Fee ratio	Inflation
Auditor industry specialization	Industry	Total assets	Market-to-book ratio	GDP growth	GDP growth
Inflation	GDP growth	Accounting standard	Change in cash sales	Inflation	Leverage

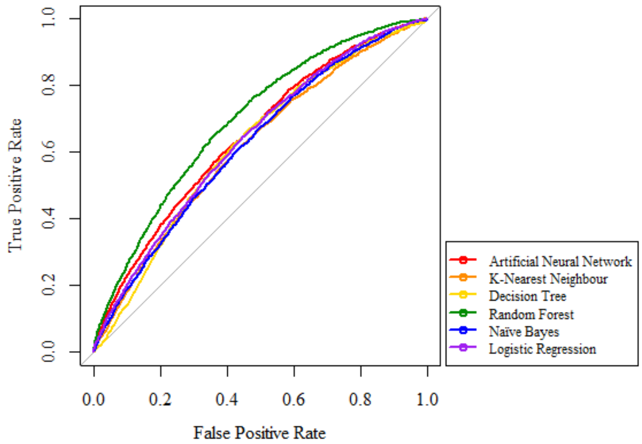

Figure 10 ROC curve of the additional results

The predictive variables of the additional research are almost the same as the predictive variables in the initial research. The GDP appears in the top ten of predictive variables of restatements, which is similar to the research of Hu et al. (2021). Figure 11 illustrates the relationship between GDP and the amount of restatements over the years. In general, the amount of restatements is decreasing, while the GDP is increasing. Around 2007, the GDP is decreasing and the amount of restatements increases which coincides with the financial crisis. Thus, when the welfare in the United States is lower, which is reflected by the GDP, firms are sometimes financially more distressed, which could be a trigger to misstate the financial statements, which eventually could lead to restatements. However, this conclusion should be drawn with extremely caution, because the GDP is the same for every firm, because the sample of this research only consists of American firms. Therefore, the negative relationship between GDP and restatements could exist, but this potential relationship should be examined further.

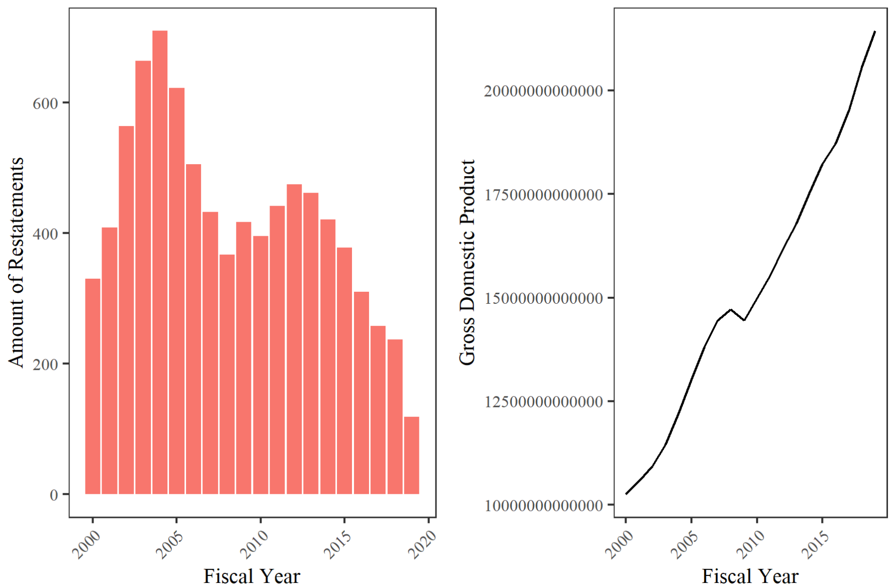

Figure 11 Relationship between GPD and restatements over the years

5. Conclusion and discussion

Currently, there is no consensus in the accounting literature about which supervised machine learning algorithm predicts audit quality the most accurate. Therefore, the aim of this research is to examine the most accurate algorithm to predict audit quality in order to achieve more consensus in the accounting literature. Hence, the research question is as follows:

Which machine learning algorithm can accurately predict audit quality and which variables can explain audit quality?

In this research, audit quality is proxied by restatements which is the best proxy for audit quality according to Aobdia (2019), DeFond and Zhang (2014) and Rajgopal et al. (2021). The sample consists of 78,917 firm-year observations of 10,078 different American firms in the period from 2000 and 2019. In total, six supervised machine learning algorithms are optimized and the performance of these algorithms is evaluated with the help of three evaluation metrics.

The performance of the random forest is superior, in both the initial research and the additional research, to the performance of the other algorithms. Therefore, the random forest is the most accurate algorithm to predict restatements and therefore poor audit quality. The aim of this research, to create more consensus in the accounting literature about which algorithm predicts audit quality the most accurate is achieved, because Zhang (2019) also concludes that random forest is the most accurate algorithm to predict audit quality. The firm characteristics are, in both the initial research and the additional research, the most predictive variables to explain restatements and thus poor audit quality. This is not in line with the findings of Zhang (2019) and Bertomeu et al. (2020) who state that the audit related variables are the most predictive variables of audit quality. The difference can be explained by the fact that random forest can more easily split continuous variables than dummy variables, because continuous variables have larger ranges. The firm characteristics mainly consists of continuous variables, while the audit related variables primarily contain dummy variables. Therefore, more firm characteristics could end in the most predictive variables of audit quality. The additional findings, where three macro-economic variables are added to the dataset, illustrate that GDP is a predictive variable of audit quality which is consistent with the findings of Hu et al. (2021). The potential negative relationship between GDP and restatements could exist, but should be examined further, because the sample only contains American firms where the GDP is the same for every firm.

The findings of the research might be of interest to the different users of the financial statements (e.g. lenders, investors and regulators) and the audit profession, because they have now more sophisticated tools, compared to linear- or logistic regressions, to assess and predict audit quality. In this way, the information asymmetry between firms and users of the financial statements is reduced which secures a more transparent working of the capital market. Furthermore, the auditor can use more sophisticated tools in the planning phase of (new) clients, where the engagement risk is assessed. Auditors can identify (new) clients with a higher likelihood to file a restatement. The auditor can assess that, partly due to the higher likelihood to file a restatement, the engagement risk is too high and decide not to audit the (new) client by which they can avoid potential litigation costs and reputation damage.

6. Literature list

Ainslie, E.K. (2006). Indicting corporations revisited: Lessons of the Arthur Andersen prosecution. *American Criminal Law Review, 43*(1), 107-142.

Akosa, J. (2017). *Predictive accuracy: A misleading performance measure for highly imbalanced data.* Retrieved from https://support.sas.com/resources/papers/proceedings 17/0942-2017.pdf

Alali, F., & Wang, S. (2017). Characteristics of Financial Restatements and Frauds: An Analysis of Corporate Reporting Quality from 2000-2014. *The CPA Journal, 87*(11), 32-41.

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series, 1142*, 1-15. doi: 10.1088/1742-6596/1142/1/012012

Aobdia, D. (2019). Do practitioner assessments agree with academic proxies for audit quality? Evidence from PCAOB and internal inspections. *Journal of Accounting and Economics, 67*(1), 144-174. doi: 10.1016/j.jacceco.2018.09.001

Barto, J., & Simko, P.J. (2002). The Balance Sheet as an Earnings Management Constraint. *The Accounting Review, 77*(s-1), 1-27. doi: 10.2308/accr.2002.77.s-1.1

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2020). Using machine learning to detect misstatements. *Review of Accounting Studies, 26*, 468-519. doi: 10.1007/s11142-020-09563-8

Brown-Liburd, H.L. (2017). Discussion of “Encouraging Professional Skepticism in the Industry Specialization Era”. *Journal of Business Ethics, 142*, 257-258. doi: 10.1007/s10551-016-3146-2

Cho, S., Vasarhelyi, M.A., Sun, T., & Zhang, C.A. (2020) Learning from Machine Learning in Accounting and Assurance. *Journal of Emerging Technologies in Accounting, 17*(1), 1-10. doi: 10.2308/jeta-10718

Cipriano, M.C., Hamilton, E.L., & Vandervelde, S.D. (2017). Has the lack of use of the qualified audit opinion turned it into the “Rotten Kid” threat? *Critical Perspectives on Accounting, 47*, 26-38. doi: 10.1016/j.cpa.2016.10.001

Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling, 240*, 113-122. doi: 10.1016/j.ecolmodel.2012.03.001

DeAngelo, L.E. (1981). Auditor size and audit quality. *Journal of Accounting and Economics, 3*(3), 183-199. doi: 10.1016/0165-4101(81)90002-1

Dechow, P.M., Ge, W., Larson, C.R., & Sloan, R.G. (2011). Predicting Material Accounting Misstatements. *Contemporary Accounting Research, 28*(1), 17-82. doi: 10.1111/j.1911-3846.2010.01041.x

DeFond, M.L., Francis, J.R., & Wong, T.J. (2000). Auditor Industry Specialization and Market Segmentation: Evidence from Hong Kong. *Auditing: a Journal of Practice & Theory, 19*(1), 49-66. doi: 10.2308/aud.2000.19.1.49

DeFond, M., & Zhang. J. (2014). A review of archival auditing research. *Journal of Accounting and Economics, 58*(2-3), 275-326. doi: 10.1016/j.jacceco.2014.09.002

Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M.A. (2020). Machine learning improves accounting estimates: evidence from insurance payments. *Review of Accounting Studies, 25*, 1098-1134. doi: 10.1007/s11142-020-09546-9

Draeger, M., Hermann, D., & Lawson, B.P. (2016). Changes in Audit Quality under Auditing Standard No. 5. *Accounting & the Public Interest, 16*(1), 57-83. doi: 10.2308/apin-51676

Dutta, I., Dutta, S., & Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications, 90*, 374-393. doi: 10.1016/j.eswa.2017.08.03051676

Elkan, C. (2012). *Evaluating Classifiers.* Retrieved from https://cseweb.ucsd.edu//~elkan/250Bwinter2012/classifiereval.pdf

Eshleman, J.D., & Guo, P. (2014). Do Big 4 Auditors Provide Higher Audit Quality after Controlling for the Endogenous Choice of Auditor? *Auditing: a Journal of Practice & Theory, 33*(4), 197-219. doi: 10.2308/ajpt-50792

Financial Reporting Council. (2008). *The Audit Quality Framework.* Retrieved from https://www.frc.org.uk/getattachment/a5650f5c-02d3-4467-a1c9-9ef433c2cd20/-;.aspx

Francis, J.R. (2004). What do we know about audit quality? *The British Accounting Review, 36*(4), 345-368. doi: 10.1016/j.bar.2004.09.003

Francis, J.R. (2011). A Framework for Understanding and Researching Audit Quality. *Auditing: A Journal of Practice & Theory, 30*(2), 125-152. doi: 10.2308/ajpt-50006

Gholamy, A., Kreinovich., & Kosheleva, O. (2018). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation.* Retrieved from https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techrep

Gomez-Uribe, C.A., & Hunt, N. (2015). The Netflix Recommender System: Algorithms, Business Value and Innovation. *ACM Transactions on Management Information Systems, 6*(4), 1-19. doi: 10.1145/2843948

Government Accountability Office. (2003). *Required Study on the Potential Effects of Mandatory Audit Firm Rotation.* Retrieved from https://www.gao.gov/assets/gao-04-216.pdf

Hasnan, S., Marzuki, H., & Shuhidan, S.M. (2017). Effects of Board Characteristics on Financial Restatement in Malaysia. *Pertanika Journal of Social Sciences and Humanities, 25*, 255-264.

Hassanat, A.B., Abbadi, M.A., & Altarawneh, G.A. (2014). Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. *International Journal of Computer Science and Information Secutirty, 12*(8). 33-39.

Hu, H., Sun, T., Vasarhelyi, M.A., & Zhang, M. (2021). A Machine Learning Approach of Measuring Audit Quality: Evidence from China. *Available at SSRN 3732563*.

IFRS. (2021). *United States – View Jurisdiction.* Retrieved from https://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/view-jurisdiction/united-states/

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal, 15*, 104-116. doi: 10.1016/j.csbj.2016.12.005

Khang, T.D. (2016). *Audit firm rotation and audit firm tenure, improvement or deterioration of the financial reporting quality for Dutch listed firms?* Retrieved from https://theses.ubn.ru.nl/bitstream/handle/123456789/2381/Tran_Duc%2c_Khang_1.pdf?sequence=1

Kinney, W.R., Palmrose, Z.V., & Scholz, S. (2004). Auditor Independence, Non-Audit Services and Restatements: Was the U.S. Government Right? *Journal of Accounting Research, 42*(3), 561-588. doi: 10.1111/j.1475-679X.2004.t01-1-00141.x

Knechel, W.R., Krishnan, G.V., Pevzner, M., Shefchik, L.B., & Velury, U.K. (2013). Audit Quality: Insights from the Academic Literature. *Auditing: A Journal of Practice & Theory, 32*(1), 385-421. doi: 10.2308/ajpt-5035

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

Lazer, R., Livnat, J., & Tan, C.E.L. (2004). Restatements and Accruals after Auditor Changes. *Working paper, New York University.*

Lemaître, G., Nogueira, F., & Aridas, C.K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research, 18*, 1-5.

Li, J., & Lin, J.W. (2005). The relation between earnings management and audit quality. *Journal of Accounting and Finance Research, 13*(1), 1–11.

Lobo, G.J., & Zhao, Y. (2013). Relation between Audit Effort and Financial Report Misstatements: Evidence from Quarterly and Annual Restatements. *The Accounting Review, 88*(44), 1385-1412. doi: 10.2308/accr-50440

Mitchell, T.M. (1997). *Machine learning.* McGraw-Hill.

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology, 48*(3), 128-138.

Palmrose, Z.V., Richardson, V.J., & Scholz, S. (2004). Determinants of market reactions to restatement announcements. *Journal of Accounting and Economics, 37*(1), 59-89. doi: 10.1016/j.jacceco.2003.06.003

Plumlee, M., & Yohn, T.L. (2010). An Analysis of the Underlying Causes Attributed to Restatement. *Accounting Horizons, 24*(1), 41-64. doi: 10.2308/acch.2010.24.1.41

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules, 26*(4), 1-16. doi: 10.3390/molecules26041111

Rajgopal, S., Srinivasan, S., & Zheng, X. (2021). Measuring audit quality. *Review of Accounting Studies, 26,* 559-619. doi: 10.1007/s11142-020-09570-9

Romanus, R.N., Maher, J.J., & Damon, M. (2008). Auditor Industry Specialization, Auditor Changes, and Accounting Restatements. *Accounting Horizons, 22*(4), 389-413.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210-229. doi: 10.1147/rd.33.0210

Srinidhi, B., & Gul, F.A. (2006). The Differential Effects of Auditors’ Non-Audit and Audit Fees on Accrual Quality. *Contemporary Accounting Research*, *24*(2), 595-629.

Stanley, J.D., & DeZoort, F.T. (2007). Audit firm tenure and financial restatements: An analysis of industry specialization and fee effects. *Journal of Accounting and Public Policy, 26*(2), 13-159. doi: 10.1016/j.jaccpubpol.2007.02.003

Sun, T. (2019). Applying Deep Learning to Audit Procedures: An Illustrative Framework. *Accounting Horizons, 33*(3), 89-109. doi: 10.2308/acch-52455

Wahab, E.A.A., Gist, W.E., & Majid, W.Z.N.A. (2014). Characteristics of non-audit services and financial restatements in Malaysia. *Journal of Contemporary Accounting & Economics, 10*(3), 225-247. doi: 10.1016/j.jcae.2014.10.001

Wardhani, N.W.S., Rochayani, M.Y., Iriany, A., Sulistyono, A.D., & Lestantyo, P. (2019). Cross-validation Metrics for Evaluation Classification Performance on Imbalanced Data. *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 14-18. doi: 10.1109/IC3INA48034.2019.8949568.

Watkins, A.L., Hillison, W., & Morecroft, S.E. (2004). Audit Quality: A Synthesis of Theory and Empirical Evidence. *Journal of Accounting Literature, 23*, 153-193.

Wooten, T.C. (2003). Research about audit quality: Certified Public Accountant. *The CPA Journal, 73*(1), 48-51.

Zhang, A. (2019). *Predict Audit Quality Using Machine Learning Algorithms.* Retrieved from https://zicklin.baruch.cuny.edu/wp-content/uploads/sites/10/2019/12/Abigail-Zhang-Predict-Audit-Quality-Using-Machine-Learning-Algorithms.pdf

Appendix A: Pre-process the dataset

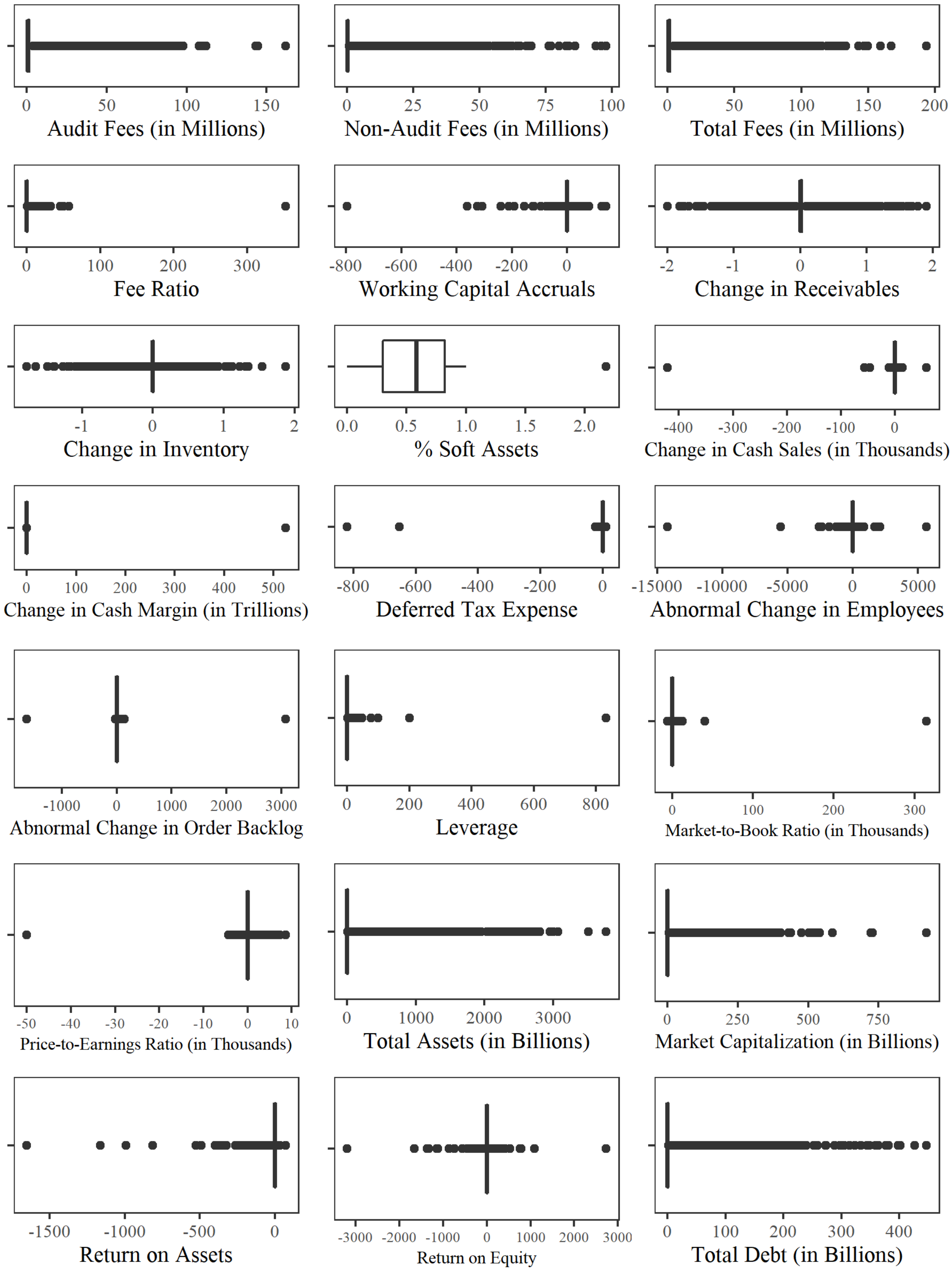

Figure 2 Boxplots of the variables with outliers

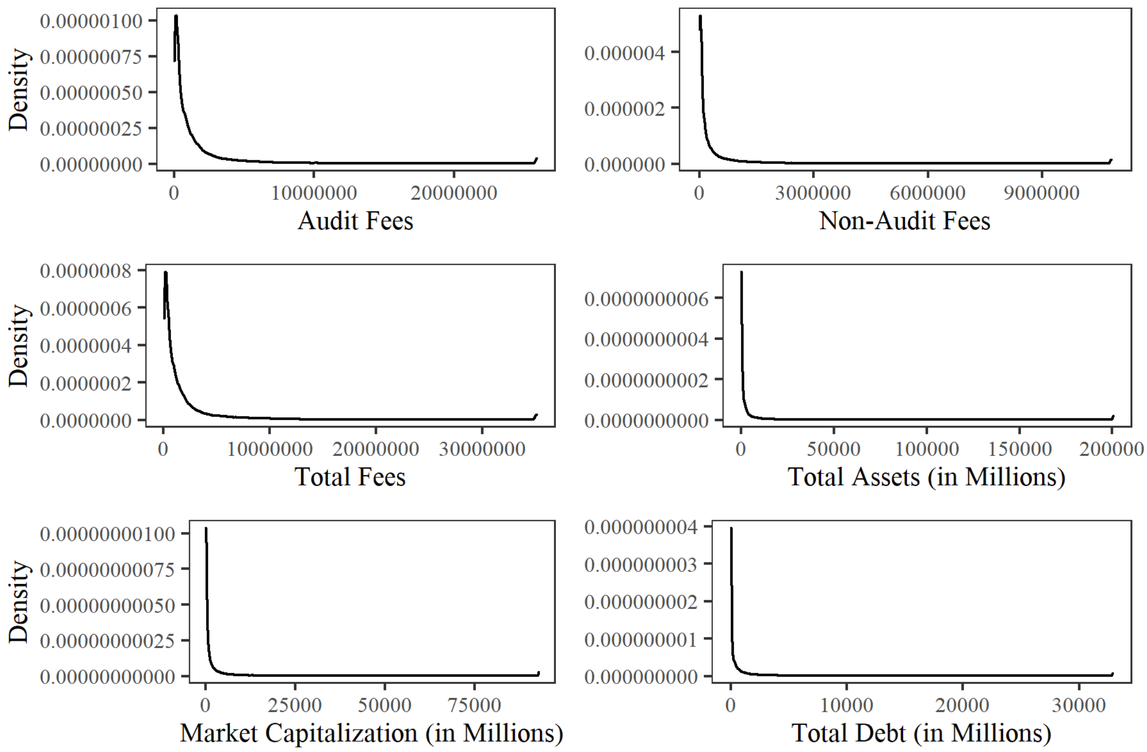

Figure 3 Density plot of the variables with a skewed distribution

Appendix B: Variables in the optimal model

Table 9 Variables in the optimal model for the initial results

	Artificial neural network	K-nearest neighbour	Decision tree	Random forest	Naïve Bayes	Logistic regression
Audit related variables						
Audit fees	✔	✔	✔	✔	✔	✔
Non-audit fees	✔	✔	✔	✔	✔	✔
Fee ratio	✔	✔	✔	✔	✔	✖
Non-audit fee ratio	✔	✔	✔	✔	✔	✔
Big 4 auditor	✔	✔	✔	✔	✔	✔
Auditor change	✔	✔	✔	✔	✔	✔
Auditor tenure	✔	✔	✖	✔	✖	✖
Short auditor tenure	✖	✖	✔	✖	✔	✔
Medium auditor tenure	✖	✖	✔	✖	✔	✔
Long auditor tenure	✖	✖	✔	✖	✔	✖
Auditor industry specialization	✔	✔	✔	✔	✔	✔
Qualified audit opinion	✔	✔	✔	✔	✔	✔
Going concern opinion	✔	✔	✔	✔	✔	✔
Total audit related variables	10	10	12	10	12	10
Firm characteristics						
Working capital accruals	✔	✔	✔	✔	✔	✔
Change in receivables	✔	✔	✔	✔	✔	✔
Change in inventory	✔	✔	✔	✔	✔	✔
Percentage of soft assets	✔	✔	✔	✔	✔	✔
Change in cash sales	✔	✔	✔	✔	✔	✔

Change in cash margin	✔	✔	✔	✔	✔	✔
Deferred tax expense	✔	✔	✔	✔	✔	✔
Abnormal change in employees	✔	✔	✔	✔	✔	✔
Abnormal change in order backlog	✔	✔	✔	✔	✔	✔
Leverage	✔	✔	✔	✔	✔	✔
Ex-ante financing need	✔	✔	✔	✔	✔	✔
Market-to-book ratio	✔	✔	✔	✔	✔	✔
Price-to-earnings ratio	✔	✔	✔	✔	✔	✔
Total assets	✔	✔	✔	✔	✔	✖
Market capitalization	✔	✔	✔	✔	✔	✔
Return on assets	✔	✔	✔	✔	✔	✔
Return on equity	✔	✔	✔	✔	✔	✔
Audit committee	✔	✔	✔	✔	✔	✔
Total debt	✔	✔	✔	✔	✔	✔
Board size	✔	✔	✔	✔	✔	✔
Industry	✔	✔	✔	✔	✔	✔
Accounting standard	✔	✔	✔	✔	✔	✔
Total firm characteristics	22	22	22	22	22	21
Total variables	**32**	**32**	**34**	**32**	**34**	**31**

Table 12 Variables in the optimal model for the additional results

	Artificial neural network	K-nearest neighbour	Decision tree	Random forest	Naïve Bayes	Logistic regression
Audit related variables						
Audit fees	✓	✓	✓	✓	✓	✓
Non-audit fees	✓	✓	✓	✓	✓	✓
Fee ratio	✓	✓	✓	✓	✓	✗
Non-audit fee ratio	✓	✓	✓	✓	✓	✓
Big 4 auditor	✓	✓	✓	✓	✓	✓
Auditor change	✓	✓	✓	✓	✓	✓
Auditor tenure	✓	✓	✓	✓	✓	✗
Short auditor tenure	✗	✗	✗	✗	✗	✓
Medium auditor tenure	✗	✗	✗	✗	✗	✓
Long auditor tenure	✗	✗	✗	✗	✗	✗
Auditor industry specialization	✓	✓	✓	✓	✓	✓
Qualified audit opinion	✓	✓	✓	✓	✓	✓
Going concern opinion	✓	✓	✓	✓	✓	✓
Total audit related variables	10	10	10	10	10	10
Firm characteristics						
Working capital accruals	✓	✓	✓	✓	✓	✓
Change in receivables	✓	✓	✓	✓	✓	✓
Change in inventory	✓	✓	✓	✓	✓	✓
Percentage of soft assets	✓	✓	✓	✓	✓	✓
Change in cash sales	✓	✓	✓	✓	✓	✓
Change in cash margin	✓	✓	✓	✓	✓	✓
Deferred tax expense	✓	✓	✓	✓	✓	✓

Abnormal change in employees	✓	✓	✓	✓	✓	✓
Abnormal change in order backlog	✓	✓	✓	✓	✓	✓
Leverage	✓	✓	✓	✓	✓	✓
Ex-ante financing need	✓	✓	✓	✓	✓	✓
Market-to-book ratio	✓	✓	✓	✓	✓	✓
Price-to-earnings ratio	✓	✓	✓	✓	✓	✓
Total assets	✓	✓	✓	✓	✓	✗
Market capitalization	✓	✓	✓	✓	✓	✓
Return on assets	✓	✓	✓	✓	✓	✓
Return on equity	✓	✓	✓	✓	✓	✓
Audit committee	✓	✓	✓	✓	✓	✓
Total debt	✓	✓	✓	✓	✓	✓
Board size	✓	✓	✓	✓	✓	✓
Industry	✓	✓	✓	✓	✓	✓
Accounting standard	✓	✓	✓	✓	✓	✓
Total firm characteristics	22	22	22	22	22	21
Macro-economic variables						
GDP	✓	✓	✓	✓	✓	✓
GDP growth	✓	✓	✓	✓	✓	✓
Inflation	✓	✓	✓	✓	✓	✓
Total variables	**35**	**35**	**35**	**35**	**35**	**34**