POLITECNICO DI TORINO

MASTER's Degree in DATA SCIENCE AND ENGINEERING

MASTER's Degree Thesis

Machine Learning for money laundering detection

Supervisors

Prof. ALESIO SACCO

Prof. MARCHETTO GUIDO

Prof. AKMAL RUSTAMOV

Candidate

OLLOSHUKUR ATADJANOV

MARCH 2025

Summary

This thesis explores the application of machine learning techniques for the detection of corporate fraud activities, a persistent challenge in the financial sector that undermines the integrity of economic systems worldwide.Corporate fraud continues to be one of the most prevalent financial crimes, with reputational harm resulting from its commission frequently outweighing the profit to the offenders by a factor of several. According to CFO estimates, it can amount to roughly 5% of a company's yearly sales. Cases of corporate fraud in some areas, particularly Russia, continue to rise notwithstanding efforts to tighten control of business activity and corporate information disclosure. Furthermore, corporate fraud schemes are getting more complicated everywhere since multiple kinds of fraud are being merged into one crime. In this sense, it is imperative to create strategies for spotting corporate fraud; artificial intelligence is helping more and more data to be accessible for this purpose. The paper looked at the key forms, causes, effects, and strategies of preventing corporate fraud. This study examined cases of any kind of corporate fraud since, generally, the motives for each type of fraud are the same and the choice of a particular type depends mostly on the possibility of doing it. Furthermore investigated were elements influencing the probability of corporate fraud: more conventional elements related to corporate governance, financial indicators, and factors becoming more and more relevant in connection with the evolution of machine learning, such text features.

Acknowledgements

I would like to express my deepest gratitude to my thesis advisor, Akmal Rustamov, for their unwavering support, guidance, and invaluable expertise throughout the research process. Their insightful feedback and constructive criticism significantly contributed to the refinement of this thesis. I am also thankful for the resources and facilities provided by Politecnico di Torino, which enabled me to conduct comprehensive and meaningful research. Additionally, I extend my appreciation to the participants who willingly shared their insights and experiences, contributing crucial data to the study. I am indebted to the pioneers in the field of machine learning and fraud detection whose groundbreaking work laid the foundation for this research.

Atadjanov Olloshukur

Table of Contents

Acronyms

ML
Machine Leaning

AML
Anti-Money Laundering

SAR
Suspicious Activity Reports

GDP
Gross Domestic Product

FATF
Financial Action Task Force

KYC
Know Your Customer

SVM
Support Vector Machines

RNN
Recurrent Neural Networks

ROC
Receiver Operating Characteristic

AUC
Area Under the Curve

LIME

Local Interpretable Model-Agnostic Explanations

API

Application Programming Interface

SMOTE

Synthetic Minority Over-sampling Technique

GPU

Graphics Processing Unit

Chapter 1

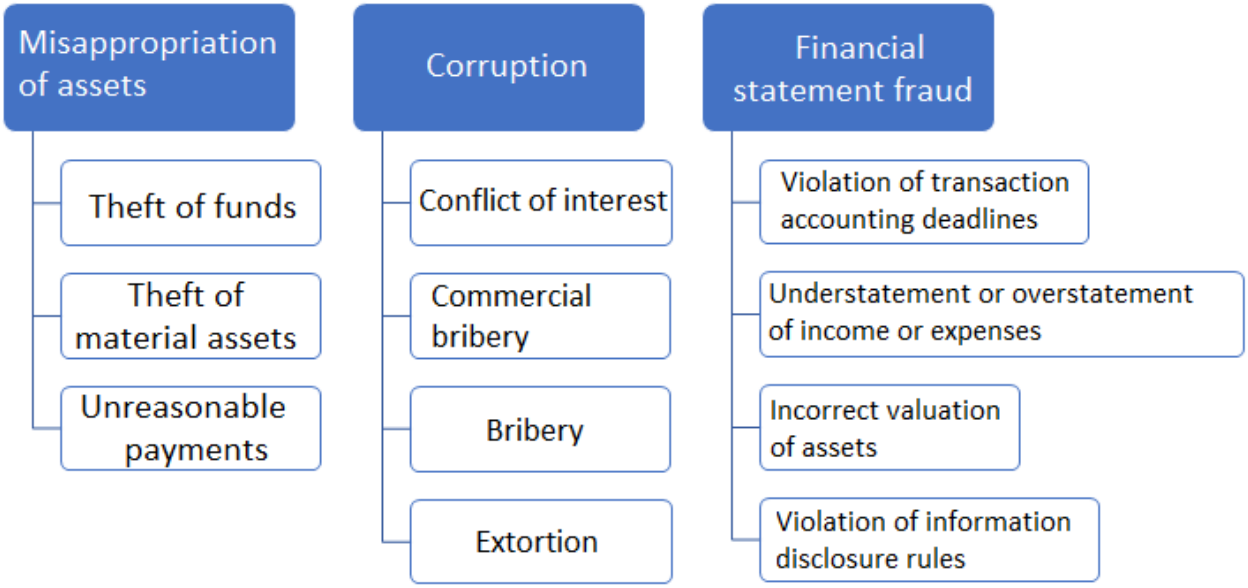
Introduction

1.1 The Phenomenon of Corporate Fraud

Corporate fraud is the intentional incorrect representation of a company's financial information or the actions of one or more individuals involved in the company, such as management, those charged with governance, employees, or third parties, to misrepresent information to the public in order to obtain an unfair or illegal advantage. Corporate fraud schemes are contrary to an employee's job responsibilities and are complex and have a negative economic impact on the business, other employees, and third parties (AICPA, 2019).

1.1.1 Types of Corporate Fraud

Corporate fraud can take many different forms, including bribery, tax evasion, bid rigging (when a company employee helps a supplier win a contract without actual competition), and asset theft. The ACFE (2023) categorizes corporate fraud into three main types: asset misappropriation, corruption, and financial reporting fraud (Figure 1).



Fugure 1. Major Types of Fraud. Source: ACFE, 2023

Asset misappropriation is a fraudulent scheme in which an employee steals or misappropriates company resources. In order to steal company funds, an employee may conspire to overreport or underreport sales, submit invoices for goods or services that were never produced or rendered, forge checks or electronic payments, or fabricate paperwork, such as for overtime, in order to transfer company funds to their personal bank accounts. Employees, managers, or outside parties may also steal inventory or other company assets by, for instance, buying other assets, arranging fictitious inventory shipments or sales, or abusing their official positions.

Corruption involves the abuse of official duties and the receipt of direct or indirect benefits directly from transactions that an employee conducts on behalf of the company, for example, by demanding bribes or inflated bonuses from counterparties for their work. Also, receiving bribes from third parties, even if the employee did not demand them, is also corporate fraud and belongs to this type.

Financial statement fraud refers to schemes in which employees intentionally misrepresent or omit material information from financial statements. This type of corporate fraud is often combined with other types of fraud because it is used to conceal the fraud.

The Criminal Code of the Republic of Uzbekistan, 1996, has a number of articles that address various forms of corporate fraud. These include: theft of another person's property, fraud, and corruption (Articles 158, 159, and 160 of the Criminal

Code, respectively); causing property damage through deception or abuse of trust in the absence of signs of theft (Article 165 of the Criminal Code); restricting competition (Article 178 of the Criminal Code); money laundering (Article 174 of the Criminal Code); abusing authority (Article 201 of the Criminal Code); commercial bribery (Article 204 of the Criminal Code); and accepting or providing a bribe (Articles 290 and 291). It is on the basis of these articles that the presence of corporate fraud in Uzbekistan companies will be determined in the future (Skipin et al., 2019; Stefan & Bykova, 2012).

Thus, corporate fraud has many types, but they are all crimes against property and the people who own it. Both employees and management of the company and third parties take advantage of their official duties and in one way or another withdraw assets from the company and have a negative impact on both the company itself and its counterparties.

1.1.2 Causes of Corporate Fraud

The reasons for corporate fraud depend on the specific type of fraud, but in most cases, employees, managers, and third parties seek personal gain and enrichment at the expense of the company. According to the fraud triangle model (Cressey, 1953), the main factors that cause corporate fraud are motivation, opportunity, and rationalization, with the decision to commit fraud depending on whether a person has each of these factors. Wolfe & Hermanson in their study (2004) expand this list and include a fourth factor - the ability to commit fraud. Let's consider each factor in more detail.

Motivation. According to research, the main motivation for employees, management, and third parties to commit corporate fraud is pressure, both external and internal. For example, this pressure can be associated with the fraudsters themselves: financial problems in the family, difficult life circumstances, low wages, and fear of losing a job (Sihombing & Shiddiq, 2014). Pressure may also be company-related: the need to meet analyst forecasts, performance standards, or industry average growth rates (Manurung & Hardika, 2015). In addition, if a company is in dire straits, even if the industry as a whole is in dire straits, managers may be motivated to overstate profits or understate losses and distort other metrics in order to reduce investor outflows (Davidson, 2016).
Opportunity. Fraud requires opportunity in addition to motivation, and this is caused by weak internal and external controls, such as the lack of corporate governance mechanisms or their ineffective operation, low compensation for internal and external auditors, inadequate process documentation, and inadequately developed

accounting policies. Opportunity also arises in the case of a complex, opaque production structure and the absence of strict sanctions for committing fraud (Abdullahi & Mansor, 2015). The fraudster must also have access tothe company's pertinent assets, the capacity to acquire the required information, fabricate documentation, and so forth.
Rationalization. The incidence of corporate fraud also depends on the beliefs of the person committing it: his internal beliefs must not contradict the commission of this crime, he must have a justification for his actions (Rae & Subramaniam, 2008). For example, the justification may be the belief that the employee should receive a higher salary than the company pays him; also, the fraudster may think that his actions will not have significant consequences for the company and thus justify his actions.
Ability. A fraudster must have certain knowledge and skills to commit fraud, especially if it is large-scale and long-term. Thus, to commit corporate fraud without being detected, it is necessary to have a good understanding of internal control and audit processes, and to have a good understanding of the weaknesses of existing control systems (Wolfe & Hermanson, 2004).

Thus, the likelihood of corporate fraud can be reduced by eliminating at least one of the factors: improving internal and external control systems or reducing employee motivation to commit fraud by maintaining a decent level of wages. However, to understand how much companies really need to combat fraud, it is necessary to assess the scale of losses caused by it.

1.1.3 Consequences and methods of combating corporate fraud

Corporate fraud has many consequences for a company and its employees: reputational damage and, as a result, customer churn, increased borrowing costs for the company, a drop in the market value of the stock, and damages due to litigation and government sanctions.

The most serious consequence of corporate fraud is reputational damage to the company, which in turn affects the behavior of stakeholders. For example, investors may begin to sell the company's shares on the stock market, which will cause their price to fall, the same may happen with bonds and other securities. This will lead to the company having a harder time raising investment and, in particular, obtaining loans, since investors will demand a higher risk premium when issuing new bonds, banks will issue loans at higher interest rates due to the same risks (Gong, 2021), and shareholders will be unwilling to invest their funds in a company

whose financial position is expected to deteriorate. Counterparties may stop doing business with the company, and consumers may stop using its goods or services, as they will lose trust in it, which may lead to losses or difficulties in conducting operations in general (Johnson, Xie, & Yi, 2012). As many researchers have noted, reputational damage is many times greater than government sanctions and, in particular, fines for companies (Karpoff, Lee, & Martin, 2008).

As mentioned above, corporate fraud may result in litigation, which also comes with significant costs for the company (Karpoff, Lee, & Martin, 2008). First, the company must pay for the work of lawyers and attorneys who will represent its interests in court; second, the court may oblige the company to pay compensation to those who suffered from its fraudulent actions; third, fines for violation of the law may be provided for the company, which must also be paid. In addition, litigation may increase the effects of reputational costs, for example, it may be necessary to change the company's management, which will slow down decision-making and development.

It is important to note that corporate fraud has a negative impact on society as a whole. Thus, one of the most common types of corporate fraud is tax evasion and other mandatory fees. Taxes and other mandatory fees are the main source of income for the state budget, and the largest portion of taxes is collected from large companies, and therefore the prevalence of tax evasion can lead to a deficit in the state budget and a reduction in spending on social programs. Also, companies create more and more jobs in the course of their development, but if the company's money is used to enrich third parties through fraudulent schemes, the company's development will slow down or stop altogether, which can lead to both a lack of new jobs and a reduction in existing ones.

Thus, corporate fraud causes serious damage to both an individual company and the economy as a whole, therefore, it is in the interests of the company itself, society and the state to combat cases of corporate fraud. Methods of prevention, timely detection and investigation are used to combat corporate fraud.

Measures to prevent and detect corporate fraud include the establishment and development of an internal and external control system, which includes the creation of regulations within the company describing prohibited actions and liability for their commission; the creation of special committees and security services in the company that monitor all transactions carried out with the company's assets, and regular audits; it is also necessary to involve external auditors to conduct independent audits (Skipin et al., 2019).

To investigate already identified or suspected cases of fraud, forensic specialists are involved, who independently analyze financial statements, check counterparties and transactions conducted with them, etc. Also, in some cases, law enforcement agencies may be involved in the investigation.

1.2 Methods for detecting corporate fraud

1.2.1 Review of existing methods for detecting corporate fraud

Let us proceed to the analysis of existing scientific approaches to identifying corporate fraud. The studies under consideration can be divided into several groups depending on the type of corporate fraud being studied and the characteristics that influence its occurrence.

Thus, with regard to the type of corporate fraud studied, most studies consider financial reporting fraud, while newer studies may analyze a specific type of fraud - tax evasion (Zainal et al., 2020; Jarboui et al., 2019), money laundering (Mousavi et al., 2022), etc. In the first study, the researchers concluded that board gender diversity is negatively associated with tax evasion, the second study concluded that larger companies are more likely to encounter corporate fraud, in particular tax evasion, and the last mentioned study proved that companies with a more developed corporate governance system have better compliance with anti-money laundering programs and are less likely to experience this type of fraud. Although conducting research taking into account different types of corporate fraud expands the sample and increases the reliability of the results obtained, different types of fraud may have different methods of detection, so researchers rarely analyze corporate fraud as a whole.

It is important to note that most studies use information from databases, such as the SEC for the United States, CSMAR for China, to determine the dependent variable of corporate fraud or its specific type. These databases collect information on criminal cases related to corporate fraud. However, not all countries have such databases, which is why some researchers collect data through the press (Uzun, Szewczyk, & Varma, 2004).

Researchers can also consider various characteristics that are related to the prevention and detection of fraud: the structure of the board of directors, the socio-demographic characteristics of its members and the company's management; financial indicators; characteristics of the text of the annual report or its individual

parts, for example, the company's management addresses, the section of the management analysis of the company's results of operations (MD&A); characteristics of the text of news about the company, etc.

One of the earliest studies of financial reporting fraud examined the impact of board composition and structure on the occurrence of corporate fraud (Beasley, 1996). The characteristics considered included the proportion of independent and foreign board members, and the presence of an audit committee on the board. It was shown that the proportion of independent and foreign board members was higher for companies that did not commit fraud, but the effect of the presence of an audit committee was not confirmed.

Further studies have used other board characteristics, such as the number of board members, the proportion of women on the board, the presence of oversight, compensation, and nomination committees, and the number of committee meetings per year. The results showed that fraudulent companies are less likely to have an audit committee and have a lower proportion of independent directors on their oversight committees (Uzun, Szewczyk, & Varma, 2004). Gender diversity has also been shown to reduce the likelihood of fraud, and the relationship is nonlinear and stronger in male-dominated industries (Cumming, Leung,& Rui, 2015).

In addition, studies have analyzed the impact of demographic characteristics of board members and company management. For example, Sun, Kent, Baolei, and Wang (2019) found that older and college-educated CFOs are less likely to commit fraud. In addition, Zhong, Ren, and Song (2022) showed that managers' foreign, academic, and military experience weaken the positive impact of lower operational efficiency on the occurrence of corporate fraud.

Thus, both the structure of the board of directors and the socio-demographic characteristics of its members and the company's management as a whole are important for identifying corporate fraud, as they are elements of corporate governance. In addition, various factors can strengthen or weaken each other's influence on the occurrence of corporate fraud.

Some researchers use changes or abnormal values of financial reporting indicators of companies to detect corporate fraud. Thus, the classic M-score model (Beneish, 1999) and its modified versions were developed. The classic Beneish model includes the dynamics of 8 indicators: the period of receivables collection (DSRI), gross profitability (GMI), asset quality (AQI), revenue growth (SGI), depreciation (DEPI), general and administrative expenses (SGAI), leverage (LVGI), total accruals to total assets (TATA). Based on them, the M-score was calculated using the following

formula:
"M-score"= 4.84 + 0.920 DSRI + 0.528 GMI + 0.404 AQI + 0.892 SGI + 0.115 DEPI – 0.172 SGAI – 0.327 LVGI + 4.697 TATA

Based on the value of this indicator, one can judge the presence of financial reporting fraud, so researchers often use it to determine the dependent variable of fraud in their studies. They consider fraud to be present if the M-score exceeds -2.22 (Khamainy, Mahrus, & Setiawan, 2022). Modified versions of this indicator are used by researchers for samples from other countries, such as China (Lu & Zhao, 2020) or Russia (Shtefan & Bykova, 2014; Fedorova & Gudova, 2019). However, this approach has its drawbacks, since it cannot always correctly identify a company as fraudulent or not, which is why using the values predicted by it as a dependent variable for the study can lead to irrelevant results.

1.2.2 Applying Text Data Mining and Machine Learning Methods to Detect Corporate Fraud

The studies presented earlier used traditional statistical models, such as logistic and probit regression, but with the development of artificial intelligence, machine and deep learning methods have increasingly been used in research. Their use allows us to identify both linear and nonlinear relationships between variables, which increases the predictive power of models. Machine learning methods also make it possible to expand the list of study variables by supplementing them with text characteristics and other information that is inaccessible to analysis using traditional methods. A description of the main machine learning methods and ways to assess the quality of their predictions is presented on pages 45-49.

Thus, Song et al. in their work (2014) used machine learning methods in combination with financial and non-financial indicators to assess the risk of financial reporting fraud: characteristics of financial stability and operational efficiency, quality of management and corporate control systems, organizational structure, industry conditions were used. Financial indicators included the amount of company assets, the financial dependence ratio, the ratio of fixed assets to assets, changes in inventories, operating income, accounts receivable, and others; in total, more than 20 financial indicators were used in the work. Among them, the size of the company (logarithm of the company's assets), financial dependence and current liquidity ratios, indicators of inventory growth, accounts receivable, and others turned out to be significant. It was also proven that machine learning methods have a higher predictive ability than the traditional logistic regression method.

Slightly less than 40 financial indicators, the full list of which is presented in Appendix 1, were used to detect financial reporting fraud by Chimonaki et al. (2019). In their study, they compared two machine learning methods - K nearest neighbors (KNN) and Naive Bayes (NB), with the former method showing higher quality (89.11 and 68.29 accuracy, respectively). For him, significant financial indicators were net working capital (NCWC), financial leverage ratio, sales growth, and others.

In addition, Gupta and Mehta (2021) conducted an analysis of studies that used machine learning to detect financial reporting fraud and found that machine learning methods can achieve higher fraud detection accuracy than traditional methods (the percentage of correct predictions for traditional methods, including logistic regression, is about 74% on average, and for machine learning methods - 85% on average), and, unlike traditional methods, can be used with small samples due to low data availability without a significant loss in quality (the percentage of correct predictions for machine learning on small samples is about 84% on average, on large ones - 87% on average). A comparative table on the basis of which conclusions are drawn is presented in Appendix 1.

Researchers also often use text characteristics both separately and together with financial indicators to identify corporate fraud. The text analyzed by researchers can be divided into four main groups: the text of annual reports, analytical reports, social networks of companies, and news about them. Researchers most often analyze the text of annual reports, the least work in the areas of analyzing analytical reports and news about companies (Gandía & Huguet, 2021).

Thus, one of the first such studies (Goel, Gangolly, & Faerman, 2010) used the texts of companies' annual reports. Using Natural Language Processing (NLP) methods, which include dividing the text into individual phrases or words (tokenization), reducing words to a universal form (stemming and lemmatization), removing stop words, semantic analysis, etc., and the Support Vector Machine (SVM) machine learning model, they proved that the text of annual reports contains information that is important for identifying corporate fraud. For example, fraudulent annual reports contain more sentences in the passive voice, more uncertainty markers, have greater lexical diversity, and are more difficult to read and understand than non-fraudulent ones.

Studies also often analyze the text of the Management Discussion and Analysis (MD&A) section or the company's management's letters to shareholders from annual reports, since this volume of text alone may be sufficient to detect fraud. For example, one study (Humpherys et al., 2011) analyzed the text of the MD&A

section of annual reports: such text characteristics as difficulty to read and understand, lexical diversity, as well as variables indicating the part of speech of the word, and sentiment variables (the number of positive and indefinite words) were used. It was shown that fraudulent texts contain more words, while having less lexical diversity, they also contain more positive words and complex constructions. It is important to note that the conclusions about lexical diversity of this and the previous study contradict each other, which may indicate the need to test hypotheses on different samples, since a company's affiliation with a certain country can affect the writing style of its annual reports.

Purda and Skillicorn in their work (2015) continued to analyze the text of the MD&A section of annual reports, their study consisted of several stages: first, a table of word frequencies was created using NLP methods, then these words were used to train a Random Forest model and determine the most significant ones for predicting corporate fraud, then the 200 most significant words were used as variables for the SVM model, which classified the observations as fraudulent or not. As a result, the resulting model correctly classified the observations in 82% of cases. Also in this study, on one sample, the authors compared the quality of this method with models based on text characteristics (the proportion of negative, uncertain and controversial words was calculated based on the Loughran and McDonald dictionary), and came to the conclusion that the calculated frequency of occurrence of negative, uncertain and other words is not enough to detect corporate fraud. It is important to note that despite the findings of this paper about the low predictive power of variables calculated based on sentiment dictionaries, more recent studies have increasingly used sentiment dictionaries because they simplify text preprocessing. However, different combinations of sentiment variables can eventually lead to a model that can classify observations as fraudulent with high accuracy.

Goel & Uzuner (2016) tested different combinations of sentiment variables using MD&A texts of annual reports for analysis. For this purpose, they applied the SVM machine learning method and several different sentiment dictionaries: LIWC (Linguistic Inquiry and Word Count Categories) is one of the basic dictionaries suitable for assessing the sentiment of texts on any topic, which contains a set of positive, negative words, as well as words indicating anxiety, anger, or sadness; MPQA (Multi-Perspective Question Answering Subjectivity Lexicon), which is also basic and contains both an indication of sentiment (subjective, as well as positive, negative, or neutral) and its degree (strong or weak); and the Loughran and McDonald dictionary, suitable for analyzing financial texts and containing, in addition to a list of positive and negative words, a list of constraining, uncertain, and litigious words. The authors also added variables indicating the part of speech of words to their

model to test whether they were associated with the presence of fraud in a company.

The study revealed that using only sentiment variables does not allow achieving high predictive ability of the model (the share of correctly classified observations - accuracy - was about 58%), however, with the addition of part-of-speech variables, the quality increases (accuracy of about 72% was achieved). After a series of experiments with variables, the authors managed to obtain a model with an accuracy of about 82% on a small sample of 360 companies, while among the sentiment dictionaries, Loughran and McDonald turned out to be the most relevant - the shares of positive and negative words calculated on its basis were in the top 5 significant variables. This is due to the fact that the same word, depending on the context, can be both positive and negative, therefore, for analyzing the text of annual reports, it is optimal to use a dictionary suitable for working with financial texts.

Goel & Gangolly (2012) also used dictionaries to determine the sentiment of annual report texts. In their work, they used the Diction 5.0 (Hart, 2000), STYLE (Cherry & Vesterman, 1991), and LIWC (Pennebaker, Booth, & Francis, 2007) dictionaries to analyze the text of annual reports for 405 fraudulent observations and 622 non-fraudulent observations in the first case, and 6741 in the second. Two groups of observations allowed us to compare the results of the models on samples of different sizes; also, in the second case, the sample was unbalanced, which could also negatively affect the results and required verification. The authors tested and ultimately confirmed 5 hypotheses: that fraudulent companies more often use complex sentences, negative words, passive voice words, uncertainty markers, adverbs, and that the text of fraudulent companies is more difficult to read and understand than for non-fraudulent ones. It was also proven that fraudulent texts contain fewer positive words. Despite the difference in the size and balance of the samples, the results for both samples were the same, which indicates the possibility of reducing the number of observations in the sample without a significant loss in the quality of the results.

Regarding management's addresses to shareholders, Bel, Bracons, and Anderberg (2021) analyzed them to check whether these texts exhibit the same features as annual report texts in general. Text characteristics such as length, lexical diversity, frequency of third-person pronouns, and words with strong emotional connotations, in particular adjectives, were calculated. Using a small sample, the SVM method showed that a large number of both positive and negative words can indicate the presence of fraud in the company, which contradicts the results of previous studies, which indicate a higher proportion of negative and a lower proportion of positive words for fraudulent texts.

Thus, text characteristics can indicate the presence of fraud in a company, and it is important to use a dictionary suitable for working with financial texts. In addition, using a part of the text of annual reports - the MD& A section or the company's management's addresses to shareholders - may be sufficient to obtain relevant results. It is also possible to use small and unbalanced samples without a significant loss in the quality of the model's predictions.

A number of researchers use text characteristics in combination with financial indicators to improve the quality of model predictions, and financial indicators can also be used as control variables to test models for relevance. For example, Hájek and Henriques (2017) used financial indicators such as company size, business reputation, profitability ratios, liquidity, activity, debt burden and market value, as well as asset structure and industry situation in their study. They also used variables of the proportion of positive and negative words, text sentiment, the proportion of uncertain, restrictive, controversial and other words. To test the hypotheses, the authors compared several machine learning models (random forest, naive Bayesian classifier, support vector machine, etc.) according to several criteria (accuracy, precision, recall, F1-score, AUC-ROC, etc.), which allowed them to select the optimal model and obtain reliable results. Thus, in terms of fraud detection, the best results were shown by the random forest and naive Bayes models, and in terms of the proportion of correct predictions, by the support vector method. At the same time, no significant difference in the quality of models using financial indicators with and without text features was found. As for the text characteristics, the proportion of negative words and the overall tonality of the text turned out to be significant; it was found that non-fraudulent texts contain fewer negative words than fraudulent ones, while the overall tone of the text for fraudulent annual reports is more informal due to the high proportion of emotionally charged words.

Finally, a number of studies have simultaneously used both financial metrics and text characteristics, as well as vectorized text itself. For example, one study (Craja, Kim,& Lessmann, 2020) used MD& A texts from annual reports and compared machine learning models in combination with different combinations of the mentioned variables. It was found that adding vectorized text to financial variables and text characteristics significantly improved the quality of model predictions, with both deep learning models (accuracy of about 97% on a sample of 208 fraudulent and 7,341 non-fraudulent observations) and the random forest machine learning model (accuracy of about 96%) proving optimal. In the study, the authors concluded that fraudulent texts, on average, contain 3 times more positive and 4 times more negative words, and they also contain more adjectives and adverbs. Similar results were obtained in the work of Xiuguo and Shengyong (2022), where the best quality

was demonstrated by deep learning models (correctly predicted about 92% of fraud cases on a sample of 5130 observations) and the SVM model (correctly predicted about 82% of cases) in combination with financial variables and vectorized text.

It is important to note that newer studies use machine learning, deep learning, and text mining techniques on samples from different countries, as they may have different cultural and institutional characteristics, including those reflected in the text of annual reports, and may yield different results. For example, Bel, Bracons, and Anderberg (2021) used a sample of Spanish firms, while Zhang et al. (2021) used a sample of Chinese firms. Both studies emphasized the need to continue using samples from different countries in future studies.

Studies using text analysis of company management's appeals to shareholders were conducted on a sample of Uzbekistan companies, but to study the impact of text characteristics on financial results (Fedorova et al., 2017) and the company's capital structure (Fedorova et al., 2019). The impact on corporate fraud, to the best of the authors' knowledge, has not been previously studied. It is also important to note that there are currently no dictionaries for determining the sentiment of words from financial texts, so studies with samples of Uzbek companies used texts in English.

1.3 Practical conclusions based on the studied literature and formulation of hypotheses

Thus, corporate fraud exists in various forms, each with its own characteristics and detection methods, making the process of identifying them long and complex. In this regard, many researchers focus on one type of corporate fraud: financial reporting fraud, tax evasion, money laundering, and so on. In this study, any type of corporate fraud will be used to determine the dependent variable, in order to examine whether it is possible to predict different types of fraud in a company using one set of factors.

Many researchers also study the impact of corporate governance characteristics and company financial indicators on the occurrence of corporate fraud. In connection with the development of artificial intelligence, new research areas have also emerged, one of which is the analysis of text information, such as annual reports of companies, MD& A sections, company management appeals to shareholders, company news, etc. This work will analyze the texts of company management appeals to shareholders from annual reports using machine learning methods.

To analyze financial texts, researchers use both sentiment dictionary-based approaches and other approaches using machine learning methods. Rarely, vectorized text variables are included in the models to improve the predictive power. In this study, text analysis will be performed using the Loughran & McDonald (LM) dictionary, and vectorized text using NLP methods will also be used. Control variables will be added to the model: financial indicators that have demonstrated a significant impact on the occurrence of corporate fraud in previous studies.

This paper will use a sample of Uzbek companies, as such studies have not been conducted on them before. Data will be collected manually using the press, as has been done in some studies in the absence of ready-made databases on corporate fraud (Uzun, Szewczyk, & Varma, 2004). The texts of annual reports will be used in English, as there are currently no tonality dictionaries for the Uzbek and Russian language suitable for working with financial texts.

Since there are inconsistencies in the existing literature regarding the relationship between the proportion of positive and negative words and the occurrence of corporate fraud, this study will also examine this relationship, assuming that fraudulent report texts will generally be more emotionally charged, consistent with the findings of Bel, Bracons, and Anderberg (2021).

Hypothesis 1. The higher the proportion of positive words in the company's management's addresses to shareholders, the higher the probability of corporate fraud.
Hypothesis 2. The higher the proportion of negative words in the company's management's addresses to shareholders, the higher the probability of corporate fraud.

Also, according to a number of studies (Humpherys et al., 2011; Goel & Gangolly, 2012), fraudulent texts are more difficult to read and understand, and longer, this work also puts forward this assumption. It is assumed that the company's management will seek to hide the fact of fraud and convince readers of the veracity of their statements, for this purpose, an increase in the text of appeals can be used by including additional information in it, and an increase in the number of complex grammatical constructions.

Hypothesis 3. The longer the management's address to shareholders, the higher the probability of corporate fraud.
Hypothesis 4. The more difficult it is to read and understand the management's address to shareholders, the higher the probability of corporate fraud.

Previous studies (Gupta & Mehta, 2021; Xiuguo & Shengyong, 2022) also found that machine learning methods can predict corporate fraud more accurately than traditional models. This study also plans to test whether this trend holds true for a sample of Uzbek companies.

Hypothesis 5. Identifying nonlinear relationships through the use of machine learning methods to predict the likelihood of corporate fraud allows us to increase the predictive power of models compared to the traditional logistic regression model.

Chapter 2

Main Part.Conducting an empirical study.

2.1 Research Methodology

2.1.1 Description of the sample and variables used

To conduct the study, data were collected from 2014 to 2021 on Uzbek non-financial companies that are included in the Rating of the largest Uzbekistan companies TOP-20 (UZA, 2024) and whose securities are traded on the UZ Exchange. In total, the initial sample included 260 companies, for each of which the texts of the CEO and chairman of the board of directors' addresses to shareholders were collected from the annual reports in English. Also, some financial indicators that have demonstrated a significant impact on the occurrence of corporate fraud in previous studies (Chimonaki et al., 2019; Xiuguo & Shengyong, 2022) were used in this work as control variables, their description is presented in Table 1. To calculate these indicators, information was collected on the following balance sheet items: assets, current assets, fixed assets, current liabilities; Information on revenue, net profit, profit before interest, taxes, depreciation and amortization, and depreciation expenses was also collected from the profit and loss statements. The sources of information used were the official websites of the companies and corporate information disclosure services, such as UZA (e-disclosure), as well as the Cbonds database.

Coefficient	Calculation formula
Size of company	Logarithm of assets
Profitability	EBIT / Assets
	Net Profit / Revenue
Asset structure	Fixed assets / Assets
Liquidity	Current assets / current liabilities

Table 1.Description of control variables

Source: compiled by the author

To determine the dependent variable of corporate fraud in a company, the work was carried out in several stages. First, keywords were selected by which one can search for information in the media about criminal cases related to corporate fraud. The following keywords were used: "fraud", "criminal case", as well as Articles 158, 159, 160, 165, 178, 179, 201, 204, 290 and 291 of the Criminal Code, which, according to previous studies (Skipin et al., 2019), are usually classified as articles related to corporate fraud. Secondly, a search for news about corporate fraud was conducted. The search was conducted for each company separately, while only those criminal cases were selected for which the investigation had already been completed, all court hearings had been held and a verdict had been rendered, acquittals were not taken into account. Thirdly, the judicial databases "Justice" and "Electronic Justice" were used to verify the collected information. As a result, a dependent variable was obtained that takes the value 1 if there were cases of fraud in the company in a particular year, and 0 otherwise.

Due to the Uzbek government's exemption from disclosing consolidated financial statements (Uzbek Government, 2022), and the fact that only a small number of companies publish annual reports in English, there were gaps in the data that were removed for further analysis. It is also important to note that the sample did not include companies for which annual and financial reports for 1-2 years can be found online, as this period is not representative enough compared to the main review period of 8 years. Thus, the final sample consists of 386 observations, including 55

companies for the 8-year period from 2014 to 2021.

The sample mainly includes companies from the electric power, metallurgy, oil and gas, and transport industries (Fig. 2). In general, this distribution corresponds to the sectoral structure of the Uzbek economy (Uzstat, 2022), with the exception of the high share of electric power, which is observed in the sample due to the inclusion of several companies from the National Electric Networks of Uzbekistan. Since almost every company from this group publishes financial statements and annual reports, and many of the distribution grid companies of this group have their own securities in circulation on the stock exchange and are included in the rating of the largest companies separately, they were included in the sample separately.

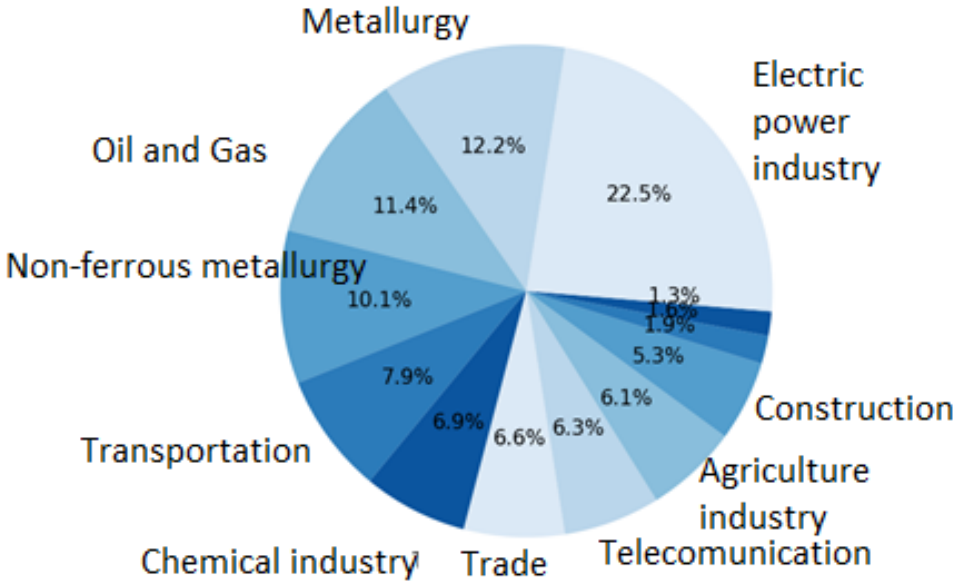


Figure 2.Distribution of observations in the sample by industry of companies. Source: author's calculations

There are 97 corporate fraud cases in the sample, which is about 25% of all observations. Unbalanced samples have been used in other similar studies, sometimes with a smaller proportion of fraud observations, and the studies have emphasized the possibility of using machine learning methods on small samples, so it is assumed that this proportion is sufficient to obtain relevant results. As for the distribution of corporate fraud cases in the sample, the largest number of cases occurs in the electric power, oil and gas, transport and metallurgy industries, which corresponds to the overall distribution of observations by industry in the sample (Fig. 3).

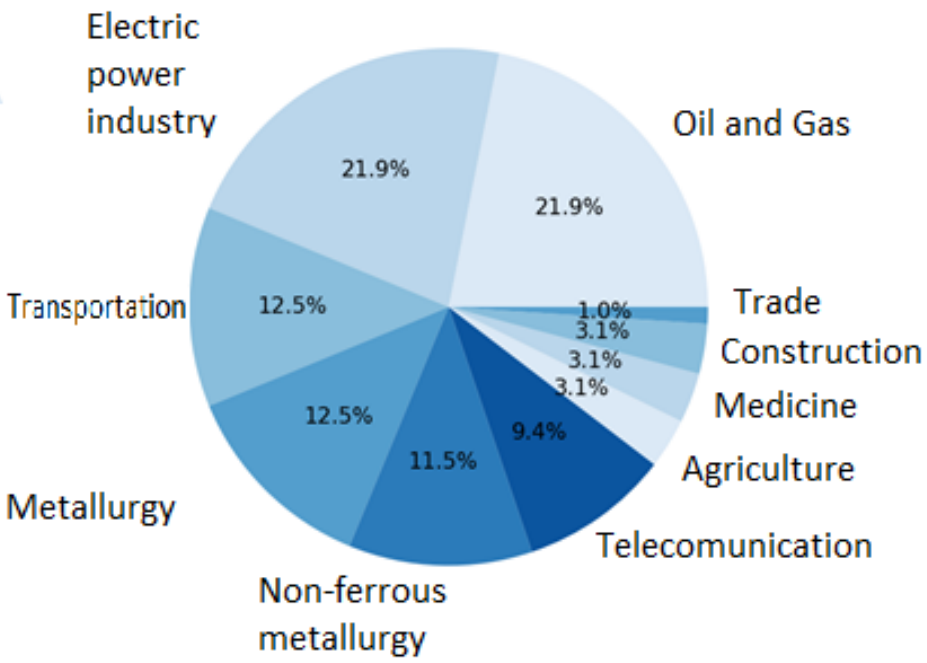


Figure 3. Distribution of corporate fraud cases in the sample by industry. Source: author's calculations

By year, both observations in general and cases of fraud are distributed evenly (Fig. 4). Each year, there are about 12 cases of fraud and about 50 observations in general. Small deviations are noticeable only in cases of fraud in 2017 and in observations in general in 2021. The latter is due to the fact that many companies have stopped publishing new annual reports (for 2021-2022), but left annual reports for past periods on information disclosure services.

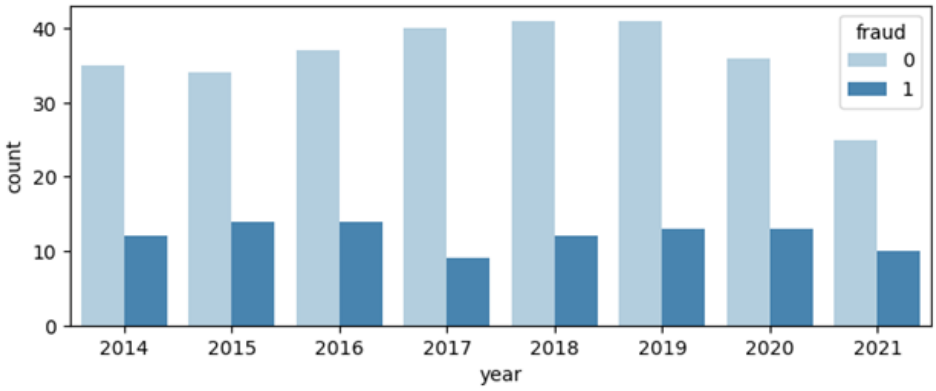


Figure 4.Distribution of corporate fraud cases in the sample by year. Source: author's calculations

Let us move on to the text characteristics used in this study. They can be divided into two groups: tonality variables and text difficulty for reading and understanding.

Variable tonality is necessary to determine the emotional coloring of the text: for this purpose, both positive and negative words are used, as well as uncertain, controversial and limiting ones (Loughran & McDonald, 2011); (Bodnaruk, Loughran, & McDonald, 2013). Positive words include the following: “achieve”, “efficient”, “improve”, “profitable”, etc. Negative words include “failure”, “lose”, “misunderstood”, etc.

Uncertainty words indicate uncertainty about the future, using words that indicate general uncertainty rather than solely the presence of risks. Examples of uncertain words include approximate, depend, fluctuate, and variability.

Constraining words include the following: “required,” “obligations,” “permitted,” and so on. It is assumed that a company’s management, expecting or experiencing financial problems, including due to fraud in the company, will use more constraining words to convey their concerns to shareholders and reduce reputational costs and the likelihood of subsequent litigation.

Litigious words are words that indicate the presence of legal disputes or legal proceedings. If a company has had high-profile lawsuits related to corporate fraud in a given year, then their mention in the text of the company’s management’s address to shareholders can be tracked using this category of words. Examples of limiting words are "claimant", "deposition", "tort", "legislation" and "regulation". A description of the sentiment variables used is presented in Table 2.

Variable	**Calculation formula**
Share of positive words	Number of positive words / text length
Share of negative words	Number of negative words / text length
Key	(Number of positive words – number of negative words) / (Number of positive words + number of negative words)
Share of unknown words	Number of undefined words / text length
Percentage of restrictive words	Number of limiting words / text length
Share of controversial words	Number of disputed words / text length

Table 2. Description of sentiment variables. Source: compiled by the author

As for the variables of text difficulty for reading and comprehension, it is assumed that when a company experiences financial difficulties, including due to corporate fraud, management is motivated to hide the fact of their existence, for example, to delay a fall in stock prices. To do this, the writing style and presentation of the text may change to make it more difficult for readers to understand. Longer sentences and longer words complicate the process of text comprehension, so studies use variables calculated on this basis: the proportion of complex (long) words in the text, the length of the text, and the average length of sentences in it (Humpherys et al., 2011). It is important to note that words with more than three syllables are considered complex, while suffixes are not taken into account when counting syllables, and multi-root words are divided into two words according to the roots, after which syllables are counted. The Fog-index is also used, which is calculated on the basis of the described indicators and determines how many years of education a person needs to understand the text on the first reading (Li, 2008). Thus, texts for a wide audience usually have a fog-index index of about 12, and texts that are understandable to almost everyone have a fog-index index of 8 or less. The description of the variables used to determine the complexity of the text for reading and understanding is presented in Table 3.

Variable	**Calculation formula**
Percentage of compound words	Number of compound words / Text length
Average number of words per sentence	Number of words in a sentence / Number of sentences in a text
Number of words in the text	Total number of words in the text
Fog index	0.4 * (Average number of words in a sentence + Percentage of complex words)

Table 3. Description of the variables of text difficulty for reading and comprehension.Source: compiled by the author.

In addition to text characteristics, this work will also use variables obtained by text vectorization: using TF-IDF Vectorizer, the frequency of each individual word in the text is calculated and then used as a variable for machine learning algorithms. This approach allows us to identify “red flag” words: words that are more common in fraudulent texts and are also identified by models as significant for detecting corporate fraud.

2.1.2 Preprocessing of text data

To calculate the variables of sentiment, reading difficulty, and comprehension, and to obtain vectorized text, text preprocessing is necessary. In this study, text processing was performed using Python 3.9.16 and its built-in libraries NumPy, Pandas, Scikit-Learn, NLTK, Re, and Readability.

First, the variables of text difficulty for reading and understanding are calculated, since this process requires minimal pre-processing: the text must be converted to lower case, then all punctuation marks except periods must be removed, and the text must be divided into separate vectors corresponding to sentences. Then, methods from the Readability library are used to calculate the corresponding variables.

To determine the sentiment variables, further processing is required: it is necessary to remove the remaining punctuation marks and stop words, tokenize the text and lemmatize the words. Stop words include frequently used words that do not determine the meaning of the text, the tone, etc. Examples of English stop words are "so", "can", "did", "which" and others. It is important to note that in this study, stop words did not include "no", "not" and other conjunctions that imply the negation of something, since they are negative and affect the tone. Tokenization involves dividing the text into separate parts (tokens), in this case the text will be divided into separate words. Lemmatization allows you to bring words to a standardized form, for example, converting plural words to singular. Also, to calculate the sentiment variables, it is necessary to pre-process the dictionary used: create lists of positive, negative, restrictive and other words on its basis, and lemmatize these words.

Finally, to create vectorized text variables, the TF-IDF method is required. This method is based on two metrics: word frequency (TF) and inverse frequency (IDF). TF shows how often a certain word occurs in a specific text, IDF shows how unique a word is for all analyzed texts. Thus, the method allows you to reduce the weight in the model of words that are frequently used for a specific sample.

$$(w\_ij = \text{tf}\_ij \times \log\left(\frac{N}{\text{df}_i}\right))$$

where w_{ij} is the TF-IDF weight of term i in document j. It reflects the importance of a word in a particular document within the entire corpus.

tf_{ij} is the term frequency of term i in document j. It represents how many times term i appears in document j.

df_i is the document frequency of term i, which counts how many documents in

the entire corpus contain term i.

N is the total number of documents in the corpus, essentially the size of the dataset.

Appendix 2 provides an illustrative example of the text pre-processing carried out in this study.

2.1.3 Descriptive and correlation analysis

The Pandas, NumPy, Scikit-Learn, Seaborn and Matplotlib libraries were used to conduct the data analysis. First of all, descriptive analysis was conducted, which consists of calculating and interpreting the following descriptive statistics: the average value of variables, their minimum and maximum. It also includes plotting the statistical distribution graphs of variables. This type of analysis is necessary to form a general idea of the values of variables, as well as to determine the need to change the functional forms of variables to bring their distribution to the standard normal.

Let us begin the analysis with the variables of text sentiment and its difficulty to read and understand. Descriptive statistics for these variables are presented in Table 4. Despite the fact that the dictionary used (LM) contains significantly fewer positive words than negative ones, the share of positive words in the texts in the sample is higher (5% versus 1%). It can be assumed that the managers of the companies included in the sample tend to present their results to shareholders in a more positive light. As for the overall sentiment of the text, on average it is 0.53, which also confirms this assumption. It is important to note that the sample contains both strictly positive texts (the maximum value of the sentiment variable is 1) and a number of texts where negatively colored words predominate (the minimum value was -0.404), which is also reflected in the high value of the standard deviation.

As for other tonality variables, their values are quite small (the shares of different words are less than 1% on average). This is due to the fact that the dictionary used has significantly fewer lists of restrictive, indefinite and controversial words than, for example, negative ones. In general, the number of words of these types in the language is small, since they are strongly limited by context (for example, the context of legal proceedings).

	count	mean	std	min	max
positive	386	0,053	0,016	0,015	0,109
negative	386	0,016	0,010	0,000	0,054
tonality	386	0,530	0,256	-0,404	1,000
uncertainity	386	0,006	0,004	0,000	0,020
constraining	386	0,005	0,004	0,000	0,022
litigious	386	0,003	0,003	0,000	0,017
wordcount	386	1612	838	286	5133
complex_pr	386	0,261	0,030	0,174	0,350
words_ps	386	24,134	3,255	15,316	37,855
fog_index	386	20,112	1,955	14,060	26,439

Table 4. Descriptive statistics of text sentiment variables and its difficulty for reading and comprehension.
Source: author's calculations

The variables of text difficulty for reading and understanding are less limited by the calculation method and therefore their values are more interpretable. Thus, on average, the number of words in the company's management's addresses to shareholders is 1612, while a large spread of this variable is noticeable: the minimum value is only about 300 words, while the maximum value reaches 5133. This is also noticeable by the fairly high value of the standard deviation. Based on this, we can assume the existence of differences in the length of the text for fraudulent and non-fraudulent companies, this assumption will be tested during the analysis of correlation and coefficients, as well as the significance of variables in the models. The share of complex words also has a fairly wide range of values: it can be 17% or reach 35% while on average it is 26% which approximately corresponds to the median and allows us to assume a normal distribution of this variable. As for the fog index, it averaged 20 years, with the minimum value being about 14 years and the maximum being 26. This suggests that for the available sample, the texts of the company management's addresses to shareholders are quite difficult to understand and require special training, in particular, a higher education in economics. Thus, this variable has the potential to be one of the most significant in the models.

	count	mean	std	min	max
Ln(Total Assets)	386	19,524	1,507	14,183	24,021
EBIT / Total Assets	386	0,091	0,303	-5,195	0,896
Net Income / Revenue	386	0,088	0,153	-0,822	1,003
Fixed Assets / Total Assets	386	0,573	0,232	0,016	1,147
Current Debt / Current Assets	386	1,366	0,969	0,136	11,183

Table 5. Descriptive statistics for control variables. Source: author's calculations

Let us also consider the distribution graphs of some variables to check how the presence of outliers affects them. Let us start with the variables of sentiment and text difficulty for reading and understanding (Fig. 5). The distribution of both variables is quite close to normal, which is generally typical for all variables reflecting text characteristics (see Appendix 3). In this regard, the functional forms of these variables will not change in any way, for example, by taking logarithms. From the distribution graph of the sentiment variable, it can also be seen that most of the texts have a positive tone. The Fog-index variable is distributed quite symmetrically, with most of the observations concentrated in the region of 18-22 years of education required to understand the text.

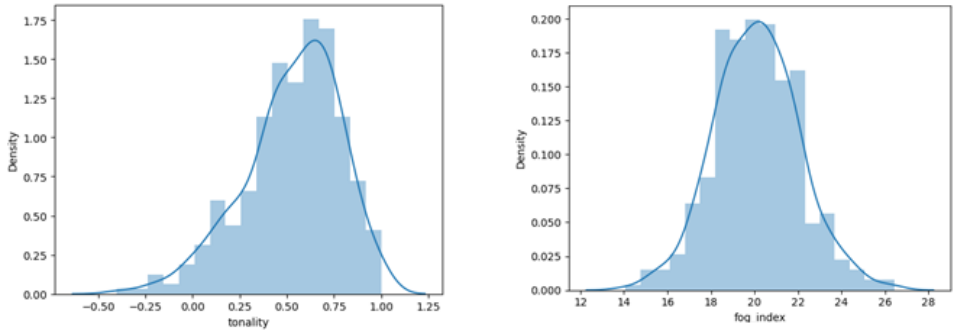


Figure 5. Distribution of the variables of text sentiment (tonality) and its difficulty for reading and understanding (fog-index). Source: author's calculations

As for the control variables, the presence of outliers is also visible in the distribution graphs (Fig. 6): for the EBIT / TA variable, a shift of some values to the left is visible, which confirms the assumption that the reason for the shift is the presence of small losses, due to which the value of the indicator becomes negative. As for the CD / CA variable, some of its values are shifted, on the contrary, to the right, and it is clear that there are several values in the range from 10 to 12, which may indicate the presence of one or more companies for which such a high value of this coefficient is typical. After logarithmizing the CD / CA variable, its distribution became closer to normal (see Appendix 3). This method was not used for the EBIT / TA variable, since logarithmization is not applicable in the presence of negative values of the variable, and therefore this variable will be used unchanged in the models below. Distribution graphs of the remaining variables are presented in Appendix 3; their distributions are closer to normal and do not require additional processing.

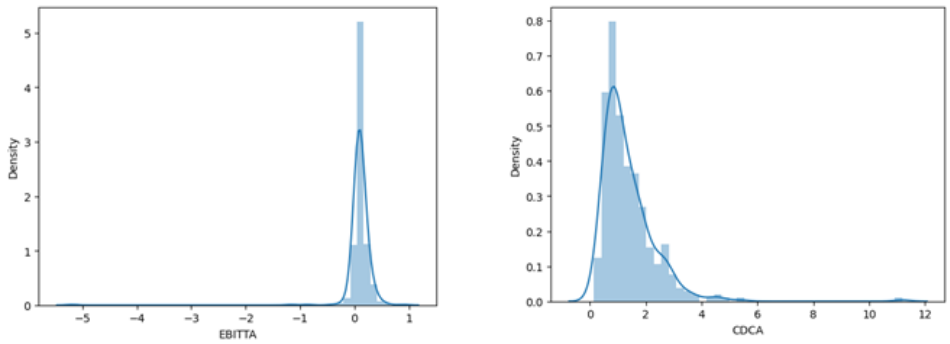


Figure 6. Distribution of profitability variables by EBIT (EBIT / TA) and current liquidity (CD / CA). Source: author's calculations

Thus, after processing, outliers in the sample are present only for one control variable. It is assumed that anomalous values of indicators may indicate the presence of corporate fraud, so these observations will not be removed from the sample. In addition, taking the logarithm in this case will not help to solve the problem, since it will remove from the sample those observations for which the indicator value was negative. It is also important to note that the values of all variables were scaled (brought to a single scale from approximately -2 to 2) for further use in machine learning models.

Now, let us consider the correlation values between the variables used. Let us start with the variables of text sentiment and its difficulty for reading and understanding (Fig. 7). The highest correlation values are observed between the sentiment variable and the variables of the proportion of positive and negative words, as well as the Fog-index variable and the variables of the proportion of complex words and the average number of words in a sentence. This is due to the calculation method, since in both cases the first variable includes the other two. Such a high correlation between the variables can cause multicollinearity in the model, due to which the variables of text sentiment (tonality), the proportion of complex words (complex) and the average number of words in a sentence (words_ps) will be removed when building the models. Also, the variable of controversial words will not be used when building the models, since it is relatively strongly interconnected with the Fog-index variable and, therefore, negatively affects the quality of the model results. The remaining variables are correlated with each other quite weakly, and they are also weakly connected with the dependent variable of the presence of fraud in the company (fraud). A possible reason for this may be the non-linear relationship between text characteristics and the occurrence of corporate fraud, which can be identified when building models using machine learning methods.

	fraud	positive	negative	tonality	uncertain	constr	litigious	fog_index	words	complex	words_ps
fraud	1,000	-0,069	-0,033	0,016	-0,097	-0,032	0,064	0,096	-0,072	0,109	0,044
positive	-0,069	1,000	-0,236	0,578	0,035	0,008	-0,163	0,058	-0,071	0,063	0,029
negative	-0,033	-0,236	1,000	-0,892	0,238	0,148	0,135	-0,019	0,007	-0,004	-0,024
tonality	0,016	0,578	-0,892	1,000	-0,174	-0,107	-0,182	0,017	-0,047	0,018	0,008
uncertain	-0,097	0,035	0,238	-0,174	1,000	0,078	-0,028	-0,108	0,188	-0,216	0,036
constr	-0,032	0,008	0,148	-0,107	0,078	1,000	-0,023	0,041	0,035	0,092	-0,023
litigious	0,064	-0,163	0,135	-0,182	-0,028	-0,023	1,000	0,241	-0,011	0,230	0,151
fog_index	0,096	0,058	-0,019	0,017	-0,108	0,041	0,241	1,000	0,124	0,761	0,803
words	-0,072	-0,071	0,007	-0,047	0,188	0,035	-0,011	0,124	1,000	-0,128	0,304
complex	0,109	0,063	-0,004	0,018	-0,216	0,092	0,230	0,761	-0,128	1,000	0,225
words_ps	0,044	0,029	-0,024	0,008	0,036	-0,023	0,151	0,803	0,304	0,225	1,000

Figure 7. Correlation matrix for the variables of text sentiment and its difficulty for reading and understanding. Source: author's calculations.

As for the control variables (Fig. 8), they are also quite weakly interconnected both among themselves and with the dependent variable. Relatively high correlation values are observed only between the profitability variables, as well as between them and the current liquidity variable. However, this does not negatively affect the quality of the models, while it is important to take into account different types of coefficients, so these variables will be used in constructing the models.

	fraud	lnTA	EBIT/TA	NI/R	FA/TA	lnCDCA
fraud	1,000	0,195	0,050	-0,004	0,138	-0,088
lnTA	0,195	1,000	0,244	0,217	0,104	0,106
EBIT/TA	0,050	0,244	1,000	0,338	-0,062	0,212
NI/R	-0,004	0,217	0,338	1,000	-0,030	0,436
FA/TA	0,138	0,104	-0,062	-0,030	1,000	-0,328
lnCDCA	-0,088	0,106	0,212	0,436	-0,328	1,000

Figure 8. Correlation matrix for control variables. Source: author's calculations.

Thus, the following variables will be used in further analysis: the proportion of positive, negative, restrictive and vague words, as well as the number of words in the text, Fog-index and all control variables.

$$\text{fraud}_{it} = \beta_0 + \beta_1 \cdot \text{tone}_{it} + \beta_2 \cdot \text{readability}_{it} + \beta_3 \cdot \text{vect\_words}_{it} + \beta_4 \cdot \text{control}_{it} + \epsilon_{it}$$

where tone are the tonality variables, readability are the text difficulty variables for reading and understanding, vect_ words are the vectorized words, control are the control variables

Almost all variables, except one control variable, are distributed close to normal or are reduced to such using logarithm. At the same time, the linear relationship

of the available variables with the occurrence of corporate fraud is quite weak, which confirms the need to use machine learning methods to identify nonlinear relationships and, therefore, improve the quality of the forecast models.

2.1.4 Description of the models used

This study will examine four machine learning methods in comparison with the traditional logistic regression model: Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), and K Nearest Neighbors (KNN).

Machine learning methods can be divided into classification and regression methods, classification methods are used if the dependent variable is several classes. In the case of corporate fraud, the dependent variable takes the values 0 (no fraud) or 1 (presence of fraud), so classification methods are used for this object of study. Further, all machine learning methods will be considered as binary classifiers.

Logistic regression is a traditional statistical model that can be used as a linear classifier. This model predicts the probabilities of objects belonging to certain classes. The formula for calculating the probabilities is presented in Appendix 4. The maximum likelihood method is used to estimate the weights of the various variables in the model.

The main advantages of this method are the simplicity of constructing the model and its further interpretation, also, if the relationship is linear, this method allows obtaining high-quality results. However, in the case of nonlinear relationships, the method can give irrelevant results, and to use it, it is necessary to remove outliers and any abnormal values, check the variables for the absence of multicollinearity. In the case of corporate fraud detection tasks, abnormal values of financial reporting indicators can be identifiers of fraud in the company, so it is better to leave them in the sample, which worsens the work of logistic regression.

K nearest neighbors as a method is based on the idea that objects belonging to the same class will be closer to each other in the plane than objects of different classes. To calculate the distance between nearest neighbors, the Euclidean distance is used, the formula for which is presented in Appendix 4.

Unlike logistic regression, this method is able to detect nonlinear relationships between features and the dependent variable, but it is still easy to interpret. However, its results depend on the distance between observations, which is why it may perform poorly in the presence of outliers and class imbalances.

Decision tree as a machine learning method is a sequence of questions and answers to them, with the help of which a decision on data classification is ultimately made. The statistical basis of the method are algorithms that, for example, determine the questions asked at each stage (the information criterion), limit the number of these questions (the stopping criterion), and so on. This model is trained by minimizing the heterogeneity index, which can be entropy, the Gini index, the misclassification index, and others. The formula for calculating the index used in this study is presented in Appendix 4.

The advantage of this method is that it can be used to identify nonlinear relationships and classify observations based on them. At the same time, it remains understandable and fairly interpretable. However, it is also unstable to outliers, just like logistic regression and the nearest neighbor method. It also has a tendency to overfit, which is why it is necessary to carefully select parameters and control the training of the model.

Random forest is an ensemble method that uses multiple decision trees to classify data, with each tree built on its own subsample of the main data set, which can improve prediction quality and reduce overfitting.

Compared to a single decision tree, a random forest allows for higher prediction accuracy, while being more robust to outliers since it averages the results of several trees. This method also allows for working with small data samples and a large number of features. The main drawback of this method is the difficulty of interpretation due to the use of several trees.

The support vector machine is based on the idea of finding a plane that would allow the observations to be divided into different groups of classes. The method is based on a function that maximizes the distance between the closest objects of different classes. If the data cannot be divided linearly, then a kernel suitable for the specific task and sample is selected: linear, polynomial, radial basis, or based on the sigmoid function. The formula for the kernel used in this study is presented in Appendix 4.

In general, the advantages of the support vector method coincide with the advantages of the random forest, and it is also more flexible when working with different types of data, since it allows you to select a kernel depending on this. However, the complexity of this method lies in the selection of the kernel, the selection can be quite complex and time-consuming. Also, if the data is unbalanced, which can be the case when working with corporate fraud data, this method can give poor results.

Thus, machine learning methods, unlike logistic regression, allow you to identify nonlinear relationships between independent and dependent variables, which can lead to better results. However, this can lead to overtraining of models, which is why you need to carefully select parameters for each of them. Also, each model has its own characteristics that need to be taken into account when using them.

To test models for overfitting and assess the quality of their predictions, the sample is divided into training and testing, as well as quality metrics, such as the proportion of correct predictions (accuracy), precision, recall, f1-score, and area under the error curve (auc roc). The training sample is used to teach models to predict corporate fraud, while the test sample is used to assess how well the models are able to predict it on new data, and the quality of the forecast is compared, and the presence of overfitting is determined based on this change. Thus, if the model always correctly predicts fraud on the training sample, this may indicate the presence of overfitting.

Let's consider how the values of the mentioned quality metrics are calculated, as well as the meaning behind them. Accuracy is calculated as the proportion of correctly classified observations. This metric is often used in research, but if the sample is unbalanced, then a high value of this indicator can be achieved by predicting all observations as a class that is more common. Also, this indicator may not be suitable if one class is more important to predict correctly than another, as in the case of corporate fraud detection, where it is more important to correctly predict the presence of fraud. The remaining metrics are calculated using the error matrix, which is shown in Figure 9.

Accuracy (recall) is calculated as TP/(TP+FP), recall (precision) as TP/(TP+FN)¦ f1-score as 2*precision*recall/(precision+recall). If we maximize accuracy when choosing the optimal model, this means that the model will focus on the correct prediction of the class that is equal to one. If we maximize recall, then, on the contrary, the model will focus on the correct prediction of 0. Maximizing f1-score allows us to focus on the correct prediction of both classes.

		True values	
		1 (Positive)	0 (Negative)
Predicted	1 (Positive)	TP (True Positive)	FN (False Negative)
	0 (Negative)	FP (False Positive)	TN (True Negative)

Figure 9. Confusion matrix for binary classification.Source:by the author.

In this study, the recall metric will be maximized, as it reflects how well the model can predict the presence of fraud in a company. Other metrics will also be considered to avoid overfitting and comparison of models.

2.1.5 Feature Engineering

Extraction of pertinent information from raw data to make it fit for machine learning models depends critically on feature engineering. From both linguistic and financial data, the following traits emerged:

Textual Qualities:

- **Tonality:** is computed as

$$\frac{\text{Number of Positive Words} - \text{Number of Negative Words}}{\text{Total Words}}$$

  This ability captures the general text's tone.

- **Words of uncertainty and constraint:** We estimated proportions of terms indicating uncertainty (e.g., "might," "possibly") or restrictive language (e.g., "must," "require").

Financial Traits:

- **Size of Company:** The logarithm of total assets helps to normalize the data, therefore facilitating comparison of different sized businesses.

- **Key Indicators:** Financial performance ratios like EBIT/Total Assets and net profit/revenue.

- **Current Assets and Liabilities:** These show a company's capacity to meet temporary needs.

2.2 Scalability and Normalization

StandardScaler from the sklearn library adjusted numerical features including text length and financial ratios to a standard range. This stage guarantees that all features are on a similar scale, therefore avoiding any features from unduly impacting the predictions of the model.

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import StandardScaler
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

# Tokenization and Stopping Word Removal
stop_words = set(stopwords.words("english"))

def preprocess_text(text):
    tokens = word_tokenize(text.lower())
    tokens = [word for word in tokens if word not in stop_words and
    word.isalpha()]
    return tokens

# Sentiment Analysis (An Example)
positive_words = ["profit", "success", "growth"]
negative_words = ["risk", "decline", "loss"]

def compute_sentiment(text):
    tokens = preprocess_text(text)
    pos_count = sum(1 for word in tokens if word in positive_words)
    neg_count = sum(1 for word in tokens if word in negative_words)
    return (pos_count - neg_count) / (pos_count + neg_count + 1e-9)
    # Avoid division by zero

# TF-IDF Vectorizing text-based data
vectorizer = TfidfVectorizer(max_features=5000, stop_words="english")
X_text = vectorizer.fit_transform(text_data)

# Standardization of numerical data
scaler = StandardScaler()
X_numeric = scaler.fit_transform(numeric_data)
```

2.3 Methodologies in Machine Learning

This work used several machine learning models to forecast corporate fraud. These models comprise logistic regression, random forest, support vector machine (SVM), decision tree, and k-nearest neighbors (k-NN).

2.3.1 Models of Selection

- **Logistic Regression:** We compared a baseline model using logistic regression. It fits binary classification projects and offers a probabilistic interpretation.

- **Random Forest:** A strong ensemble learning method based on several decision trees used to raise prediction accuracy. It can manage vast amounts of data and is less likely to overfit.

- **Support Vector Machine (SVM):** Effective in high-dimensional environments, SVM was used to identify a decision boundary optimizing the margin between fraud and non-fraud situations.

- **Decision Tree:** Simple yet understandable, a decision tree divides the data into subsets depending on feature values. Though it can readily overfit, it helps one to grasp feature relevance.

- **k-Nearest Neighbors (k-NN):** A non-parametric model predicted on the majority class of the closest neighbors. For small datasets, it is straightforward and efficient; for large datasets, it is computationally costly.

2.3.2 Hyperparameter Tuning

Grid Search methodically assesses a range of hyperparameters to identify the combination that produces the highest performance for each model.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV

# Random Forest with Grid Search
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5]
}
rf = RandomForestClassifier()
grid_search = GridSearchCV(rf, param_grid, cv=5, scoring='recall')
grid_search.fit(X_train, y_train)
```

2.4 Model Assessment

Examining the models' performance across several criteria came next once they had been trained. This guarantees that the model generalizes effectively to unprocessed data in addition to fit the training data.

2.4.1 Rating Systems of Measurement

- **Recall:** Reflecting the percentage of real fraud incidents the model accurately predicted, recall is the main evaluation indicator.
- **Precision:** This statistic evaluates the actual percentage of expected fraud instances turned out as fraudulent.
- **F1-Score:** Offering a fair assessment, the harmonic mean of recall and accuracy.
- **AUC-ROC:** The area under the Receiver Operating Characteristic curve gauging the model's capacity to discriminate between non-fraud and fraud situations.

2.4.2 ROC Curve and Confusion Matrix

The models' performance was visualized using a confusion matrix. It displays real positives, real negatives, false positives, and false negatives. Plotting ROC curves helped one to evaluate several model performance.

```
from sklearn.metrics import confusion_matrix, roc_curve, auc
import matplotlib.pyplot as plt

# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:\n", cm)

# ROC Curve
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
roc_auc = auc(fpr, tpr)

# Plotting ROC Curve
plt.plot(fpr, tpr, label=f"ROC Curve (AUC = {roc_auc:.2f})")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.legend()
plt.show()
```

Chapter 3

Results and Discussion

3.1 Description of research results

3.1.1 Logistic Regression Results

Let us proceed to the analysis of the model results and start with the traditional logistic regression model, the results of which were obtained using the Statsmodels library (Table 6). First of all, an LR test was conducted for the significance of the regression variables as a whole, the p-value of which was 0.000, therefore, the null hypothesis of the simultaneous insignificance of all model variables is rejected. As for pseudo-R2, its value was about 10% . However, since R2 is the proportion of the variance of the dependent variable that can be explained by the model, and the dependent variable is binary (0 or 1), the pseudo-R2 indicator should be interpreted with caution. In general, it can be said that the set of variables used can indeed predict corporate fraud.

	coef	std err	P>|z|	[0,025	0,975]
const	-9,9154	2,576	0,000***	-14,964	-4,86700
Ln(TA)	0,3410	0,090	0,000***	0,165	0,51700
EBIT/TA	5,4714	1,817	0,003**	1,909	9,03400
NI/R	-3,5627	1,536	0,020**	-6,572	-0,55300
FA/TA	1,0548	0,622	0,090*	-0,165	2,27400
Ln(CD/CA)	-0,4349	0,206	0,035**	-0,839	-0,03100
positive	-12,5626	8,699	0,149	-29,613	4,48800
negative	-6,0617	14,568	0,677	-34,614	22,49000
uncertainity	-23,5593	34,590	0,496	-91,354	44,23600
constraining	-50,1065	35,598	0,159	-119,878	19,66500
fog_index	0,1460	0,069	0,035**	0,010	0,28200
wordcount	-0,0004	0,000	0,030**	-0,001	-0,00004

Table 6. Logistic Regression Results. Source:Compiled by the author

p <0,1; **p <0,05; 6*p <0,01*

As for the variables individually, all control variables were significant, which indicates the adequacy of the model, and the fog index and text length were significant at a confidence level of 5% . At the same time, the fog index was positively associated with the occurrence of fraud, which allows us to confirm hypothesis 4, and the text length was negative, and therefore hypothesis 3 was rejected. Thus, with an increase in the required years of education by 1, the probability of a negative outcome increased by 1.1572 times, and with an increase in the text length by 1 word, it decreased by 0.9996 times. Conclusions on hypotheses 1 and 2 based on the results of logistic regression cannot be made, since the variables were insignificant.

3.1.2 Comparison of results obtained using logistic regression and machine learning models

Next, machine learning models were built, for each of which the corresponding hyperparameters were selected using the Grid Search method. The final hyperparameters of each model are presented in Appendix 5. To determine the most relevant model, quality metrics were compared (Table 7), the most important of which is recall (the proportion of correctly predicted fraud). The most relevant model turned out to be a random forest, while logistic regression shows the worst quality of fraud prediction than all the presented machine learning methods, which allows us to confirm hypothesis 5.

	Logit	KNN	Tree	Forest	SVM
accuracy	0,808	0,803	0,837	0,824	*0,883*
precision	0,829	0,684	0,685	0,624	*0,882*
recall	0,299	0,402	0,649	*0,753*	0,619
f1-score	0,439	0,506	0,667	0,682	*0,727*
roc-auc	0,639	0,670	0,775	*0,800*	0,795

Table 7. Comparison of quality metrics for the models used. Source: Author's calculations.

Let us compare the results of the optimal machine learning model (random forest) and logistic regression in more detail. Thus, according to the error matrix presented in Figure 10, it is clear that logistic regression correctly classifies only 29 fraud cases out of 97 presented in the sample, for the random forest this value is 73, which is 2.5 times more. However, the random forest model is more often wrong - 44 observations are classified as fraudulent, although they are not. The reason for this may be that fraud can be detected several years after it was committed, so the sample may include fraudulent observations for which the case of corporate fraud has not yet been detected, but is already reflected in the values of the variables. Another likely reason for frequent classification errors is the characteristics of the sample: it is quite small and unbalanced, which complicates the process of training models.

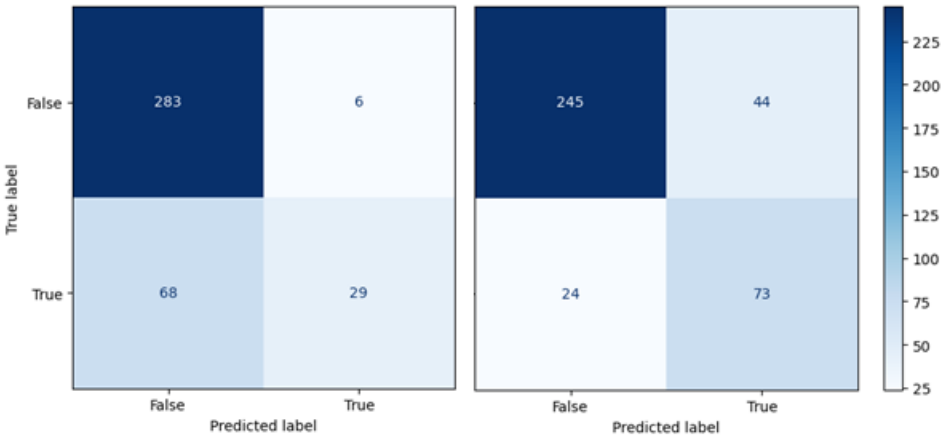


Figure 10. Confusion matrices for logistic regression and random forest, respectively. Source: author's calculations.

Also, the area under the ROC curve for logistic regression is smaller than for other models (Fig. 11), which indicates a lower ability of this model to correctly classify observations as fraudulent or not. The highest AUC ROC value is observed for the support vector machine (SVM), but in this case there is likely to be a small

overfitting (see Appendix 6), when the model is strongly adjusted to the existing sample, due to which the model can predict classes almost perfectly on it, but on new data the quality will be much lower.

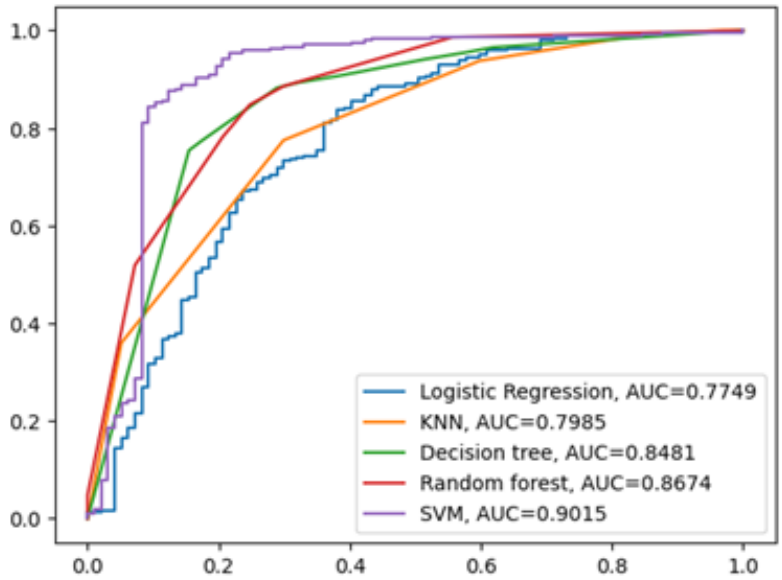


Figure 11. Comparison of AUC ROC for different models. Source: author's calculations.

Since the variable shares of positive and negative words in the text turned out to be insignificant for the logistic regression model, we will test hypotheses 1-2 for the random forest model. It is important to note that machine learning models are more difficult to interpret; they do not have the concept of feature significance in the statistical sense. However, for the random forest model, one of the typical trees can be visualized, based on which the observations were classified. At each step, the tree checks whether the observations meet a certain criterion and divides them into two groups; this continues until the sample is maximally divided into two classes.

It is important to note that due to the scaling of the features, the values of the variables in the criteria formed by the tree require careful interpretation (since they are always in the range from -2 to 2). Thus, it is assumed that the more the criterion formed on the basis of the variable divides the sample, the stronger the influence of this variable on the occurrence of fraud.

Thus, the criterion for dividing the sample based on the variable proportion of positive words was as follows: the proportion of positive words is less than 0.081,

taking into account scaling. Figure 12 shows a part of the decision tree illustrating the division of the sample based on this criterion. Thus, among the observations for which the proportion of positive words is higher than the value specified in the criterion, non-fraudulent ones prevail (106 observations versus 32 fraudulent ones), therefore, hypothesis 1 can be rejected.

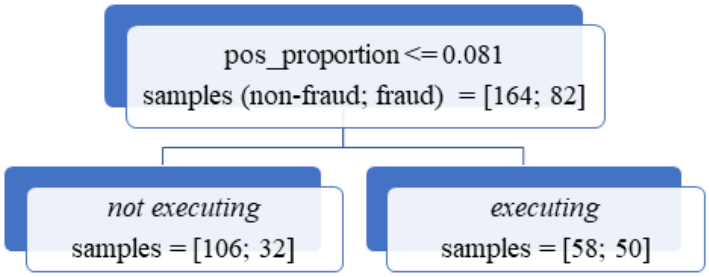


Figure 12. Illustration of the division of a sample into classes by a decision tree model based on a criterion formed using the variable proportion of positive words. Source: author's calculations.

As for the share of negative words, the following criterion was formed based on this variable: the share of negative words is less than -0.887, taking into account scaling. Thus, based on the results of the decision tree (Fig. 13), it can be assumed that for non-fraudulent companies, the share of negative words in the text of management's appeals to shareholders is less. However, conclusions regarding hypothesis 2 cannot be made based on this information, since the share of fraudulent observations that meet and do not meet the criterion does not differ so significantly. It is important to note that the number of observations before dividing the sample in the presented figures is different, since the decision tree divides the sample into classes sequentially.

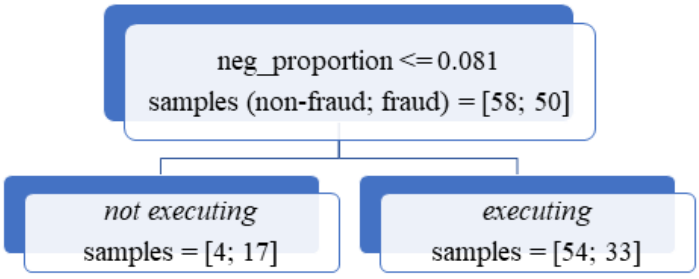


Figure 13. Illustration of the division of the sample into classes by the decision tree model based on a criterion formed using the variable proportion of negative words. Source: author's calculations.

Also, for the SVM model, which showed almost the same high results as the random forest, word clouds were built indicating the presence or absence of fraud

in the company. They are presented in Figures 14 and 15. The words indicating the presence of fraud were: “weaken”, “abandon”, “minimize”, “remain”, “become”. As for the words indicating the absence of fraud, they are “attract”, “differ”, “celebrate”, “motivate”, “recover”. It can be seen that the first group of words contains more negative ones, and the second group contains more positive ones, which is consistent with the results of the random forest model and the results of previous studies (Goel & Gangolly, 2012).

Figure 14. Word cloud indicating the presence of fraud in a company. Source: author’s calculations.

Figure 15. Word cloud indicating the absence of fraud in a company. Source: author’s calculations.

Thus, hypotheses 4 and 5 were confirmed, 1 and 3 were refuted, and no clear

conclusions can be drawn regarding hypothesis 2 based on the results obtained. The results of this study are consistent with the results obtained in the following studies: a higher proportion of positive words for companies in which there is no fraud (Goel & Gangolly, 2012); a lower proportion of negative words for companies in which there is no fraud (Hájek & Henriques, 2017); that the texts of company management's addresses to shareholders are more difficult to read for fraudulent companies (Humpherys et al., 2011); and that machine learning models, by identifying nonlinear relationships, allow obtaining higher quality results (Xiuguo & Shengyong, 2022). New results were also obtained: using a sample of large Russian companies, it was found that for companies in which there is no fraud, the texts of management's addresses to shareholders are longer.

Hypothesis 1. *The higher the proportion of positive words in the company's management's appeals to shareholders, the higher the likelihood of corporate fraud.* ***Refuted***

Hypothesis 2. *The higher the proportion of negative words in the company's management's messages to shareholders, the higher the likelihood of corporate fraud.* ***The results were mixed***

Hypothesis 3. *The longer the management's address to shareholders, the higher the probability of corporate fraud.* ***Refuted***

Hypothesis 4. *The more difficult it is to read and understand a company's management's address to shareholders, the higher the likelihood of corporate fraud.* ***Not refuted***

Hypothesis 5. *Identifying nonlinear relationships using machine learning methods to predict the likelihood of corporate fraud improves the predictive power of models compared to the traditional logistic regression model.* ***Not refuted***

Conclusion

Corporate fraud remains one of the most common types of financial crime, with reputational damage to a company often far exceeding the benefit to the perpetrators and, according to CFO estimates, can amount to approximately 5% of a company's revenue annually (ACFE, 2023).

Despite measures to strengthen regulation of companies' activities and disclosure of corporate information, cases of corporate fraud in some regions, including Russia, continue to grow. In addition, corporate fraud schemes are becoming more

complex around the world, and a combination of several types of fraud in one crime is increasingly common. In this regard, it is necessary to develop methods for detecting corporate fraud, and with the development of artificial intelligence, more and more information is becoming available for this.

The study examined the main types, causes, consequences and methods of combating corporate fraud. Since the causes for each type of corporate fraud are generally the same and the fraudster's choice of a specific type depends mainly on the possibility of committing it, this study analyzed cases of any corporate fraud. The factors influencing the likelihood of corporate fraud were also studied: these include both more traditional factors - corporate governance characteristics, financial indicators; and factors that are gaining popularity due to the development of machine learning, such as text characteristics.

To conduct the study, data were collected on 20 non-financial Uzbek companies for 2014-2021. Due to the lack of ready-made databases on corporate fraud for companies, data were collected manually to determine the dependent variable of corporate fraud, using information from the media and criminal case databases. The texts of company management's appeals to shareholders from annual reports in English were also manually collected, which is why there were many gaps in the sample. After preliminary data processing, the final sample included 386 observations for 18 companies over 8 years.

This study identified the relationship between the • Second, a number of hypotheses were tested using this model and a logistic regression model. Of these, the hypothesis that fraudulent texts are more difficult to understand and read was confirmed. It was also found that, compared to fraudulent companies, for companies that did not commit fraud, the texts of management communications to shareholders would contain more positive words, fewer negative words, and would be longer overall. This is consistent with a number of past studies (Goel & Gangolly, 2012; Hájek & Henriques, 2017; Humpherys et al., 2011; Xiuguo & Shengyong, 2022).

• Thirdly, using machine learning models, word clouds were obtained that signal the presence or absence of fraud in a company, which have practical significance for both audit and forensic specialists, government agencies, and for the company's stakeholders, who can make investment and other decisions based on them.characteristics of the text of CEO and board chairman appeals and the occurrence of corporate fraud using machine learning methods. Both the characteristics of the text tone and its difficulty to read and understand, which are quite common in previous studies, and vectorized words, which are quite rare as variables, were

used. In addition, similar studies have not previously been conducted on a sample of Uzbek companies, which together constitutes the scientific significance of this study. Despite the fact that this work uses a small sample of data, the study was able to obtain relevant results:

• First, an optimal model for detecting corporate fraud was obtained – the Random Forest machine learning model, which demonstrated the highest quality compared to other machine learning models (KNN, SVM, Decision Tree) and the traditional logistic regression model. This model was able to correctly predict corporate fraud in 75% of cases, which is 2.5 times higher than the same indicator for logistic regression. In the future, such a model can be used by audit and forensic specialists for the timely detection of corporate fraud.

• Second, a number of hypotheses were tested using this model and a logistic regression model. Of these, the hypothesis that fraudulent texts are more difficult to understand and read was confirmed. It was also found that, compared to fraudulent companies, for companies that did not commit fraud, the texts of management communications to shareholders would contain more positive words, fewer negative words, and would be longer overall. This is consistent with a number of past studies (Goel & Gangolly, 2012; Hájek & Henriques, 2017; Humpherys et al., 2011; Xiuguo & Shengyong, 2022).

• Thirdly, using machine learning models, word clouds were obtained that signal the presence or absence of fraud in a company, which have practical significance for both audit and forensic specialists, government agencies, and for the company's stakeholders, who can make investment and other decisions based on them.

The fact that the resulting model was able to achieve such high quality even on a small data set demonstrates the need for further research in this area. Future research could focus on the following:

1. First of all, by studying wider data samples. This can be achieved by expanding the time period towards earlier observations, including small and medium-sized enterprises in the sample, and expanding it geographically;

2. It is equally important to use more advanced machine learning methods – neural networks, which do not require the use of dictionaries and allow increasing the predictive power of models;

3. To improve the interpretability and ease of use of models, not all vectorized words can be used, but only the most significant of them;

4. It is possible to use other sources of text information – analytical notes, company news, etc.;

5. Along with the previous point, you can expand the list of sentiment dictionaries used, test the work of dictionaries available in Russian.

Chapter 4

Conclusion

There are other limitations of this study. First, the estimates obtained may be biased due to the fact that the sample only included the largest companies, and because annual reports in general and in English in particular were unavailable for some companies. Second, not all cases of corporate fraud for the presented sample have been disclosed to date, so fraud may be present in the company but not reflected in the dependent variable, which affects the model training process. Third, the texts of the company's management's addresses to shareholders are written in a fairly neutral manner, which is why a dictionary approach to determining their tone may be ineffective; however, the use of vectorized words helps compensate for this problem. Finally, each language has its own characteristics, and the text of the same annual reports may have a different writing style in Russian and English. Consequently, annual reports in English may not fully reflect both the emotional coloring and the difficulty of reading and understanding the original text in Russian.

4.1 Appendix 1.

Information on past corporate fraud studies

Indicator	Description
Total Debt	Solvency indicator
The logarithm of Total Debt	Solvency indicator
Equity	Structure indicator
Debt to Equity	Solvency indicator
Total Debt / Total Assets	Solvency indicator
Long Term Debt / Total Assets	Solvency indicator
Short-Term Debt / Total Assets	Solvency indicator
Account Receivable/Sales	Turnover indicator
Inventory/Sales	Turnover indicator
Inventory / Total Assets	Turnover indicator
Sales Growth	Turnover indicator
Sales	Turnover indicator
Gross margin	Return on sales
Sales minus Gross Margin	Turnover indicator
Total assets	Structure indicator
The logarithm of Total Assets	Structure indicator
Net fixed assets/ total assets	Structure indicator
Gross Profit/Total Assets	Return on investment
Net Profit/Total Assets	Return on investment
Net Profit/Sales	Return on sales
Working Capital	Liquidity indicator
Working Capital/Total Assets	Liquidity indicatior
Sales to total assets	Turnover indicator
Current Assets/ Current Liabilities	Liquidity indicatior
Net Income/Fixed Assets	Return on investment
Cash/Total Assets	Liquidity indicatior
Quick Assets/Current Liabilities	Liquidity indicatior
Earnings Before Interest and Taxes	Return on sales

Net Income/Fixed Assets	Return on investment
Cash/Total Assets	Liquidity indicatior
Quick Assets/Current Liabilities	Liquidity indicatior
Earnings Before Interest and Taxes	Return on sales
Ebit /Total Assets	Return on investment
Equity/Total Liabilities	Structure indicator
Z-score	Return on investment
Inventory	Turnover indicator
Net profit after tax	Return on sales
Sector	
P/E	
Price/book value	Investment Ratio (Coefficient)

Table 1. Description of financial indicators used in studies to detect financial reporting fraud.

Source: Chimonaki, Papadakis, Vergos, & Shahgholian, 2019

№	Author	Used methods	Selected volume	Accuracy
1	Persons (1995)	LR	206 companies; 103 fraud и 103 non-fraudulent	71.5%
2	Green and Choi (1997)	NN	95 companies; 46 fraud и 49 non-fraudulent	71.7%
3	Summers and Sweeney (1998)	Cascade Logistic Regression	102 companies; 51 fraud и 51 non-fraudulent	59.8%
4	Beneish (1999)	Probit regression	2,406 companies; 74 fraud и 2,332 non-fraudulent	89.5%
5	Feroz et al. (2000)	NN, LR	132 companies; 42 fraud и 90 non-fraudulent	NN: 81% LR: 70%
6	Spathis (2002)	LR, Univariate and multivariate statistical tools	76 companies; 38 fraud и 38 non-fraudulent	75.4%
7	Lin et al. (2003)	LR, NN	200 companies; 40 fraud и 160 non-fraudulent	NN: 79% LR: 76%
8	Kaminski et al. (2004)	Discriminant analysis	158 companies; 79 fraud и 79 non-fraudulent	53.8%
9	Kirkos et al. (2007)	Decision tree, NN and Bayesian networks	76 companies; 38 fraud и 38 non-fraudulent	DT: 73.6% NN: 80% BBN: 90.3%
10	Lenard and Alam. (2009)	LR	30 companies; 15 fraud и 15 non-fraudulent	77%
11	Gaganis (2009)	Discriminant analysis (UTADIS), logistic regression, method KNN, ANN	398 companies; 199 fraud и 199 non-fraudulent	PNN: 82.93% UTADIS: 87.20%

12	Cecchini et al. (2010)	SVM using a custom financial kernel	6,632 firms; 205 fraudulent and 6,427 non-fraudulent	90.4%
13	Ravisankar et al. (2011)	PNN в Neuroshell 2.0, genetic programming, logistic regression	202 firms; 101 fraudulent and 101 non-fraudulent	PNN: 98.09% GP: 94.14% LR: 71%
14	Dechow et al. (2011)	LR	79,651 firms; 293 fraudulent and 79,358 non-fraudulent	63.7%
15	Mehta et al. (2012)	LR	60 firms; 30 fraudulent and 30 non-fraudulent	71.5%
16	Dalnial et al. (2014)	Multiple linear regression	130 firms; 65 fraudulent and 65 non-fraudulent	74.7%
17	Kanapickienė and Grundienė (2015)	LR	165 firms; 40 fraudulent and 125 non-fraudulent	64%
18	Lin et al. (2015)	LR, Decision tree (CART), ANN	576 cases; 129 fraudulent and 447 non-fraudulent	LR: 88.5% CART: 90.3% ANNs: 92.8%
19	Dong et al.(2016)	SVM model on text features together with 84 financial coefficients	1610 firm-year observations; 805 fraudulent and 805 non-fraudulent	82.49%
20	Zainuddin and Hashim (2016)	LR	30 companies; 15 fraudulent and 15 non-fraudulent	73%
21	Hajek and Henriques (2017)	14 different methods applied on 30 different datasets, logistic regression, Bayesian networks	622 firms; 311 fraudulent and 311 non-fraudulent	Accuracy LR: 74.54% BBN: 90.32%

22	Jan (2018)	ANN + Decision tree (CART CHAID, C5.0, QUEST)	160 firms; 40 fraudulent and 120 non-fraudulent	CART: 91% CHAID: 90% C5.0: 88% QUEST: 84%
23	Jofre and Gerlach (2018)	Quadratic Discriminant Analysis (QDA), Logistic Regression, AdaBoost Decision Trees (AB DT), Gradient Boosting (BT) and Random Forests (RT)	3,188 firm-year observations; 1,594 fraudulent and 1,594 non-fraudulent	QDA and BT were the most accurate models, with both achieving 87.5% accuracy.
24	Hajek (2019)	Fuzzy rule-based system	622 firms, 311 fraudulent and 311 non-fraudulent	Accuracy: 86.8%

Source:Gupta & Mehta, 2021

4.2 Appendix 2.

Illustration of the steps of text data preprocessing

Original text

In the reporting period, Acron Group's financial results improved each quarter, supported by sales volume, a weaker rouble, and price recovery. Eventually, due to the Group's weak performance in H1 2020, total revenue decreased 6% to USD 1,661 million, EBITDA was down 11% to USD 489 million at 29% EBITDA margin. Net profit was USD 53 million, down from USD 383 million year-onyear. This decline in profit was mainly due to non-monetary factors, including a foreign exchange loss. The Group's debt burden grew in 2020, so we took measures to save and cut capex. Its actual amount was USD 249 million against the scheduled USD 300 million. By the end of 2020, dollar-denominated net debt/EBITDA was 2.8.

Text for calculating the variables of difficulty for reading and understanding

['reporting', 'period', 'financial', 'result', 'improved', 'quarter', 'supported', 'sale', 'volume', 'weaker', 'rouble', 'price', 'recovery', 'eventually', 'weak', 'performance', 'h', 'total', 'revenue', 'decreased', 'usd', 'million', 'ebitda', 'usd', 'million', 'ebitda', 'margin', 'net', 'profit', 'usd', 'million', 'usd', 'million', 'onyear', 'decline', 'profit', 'mainly', 'non', 'monetary', 'factor', 'including', 'foreign', 'exchange', 'loss', 'debt', 'burden', 'grew', 'took', 'measure', 'save', 'cut', 'capex', 'actual', 'amount', 'usd', 'million', 'scheduled', 'usd', 'million', 'end', 'dollar', 'denominated', 'net', 'debt', 'ebitda']

Vectorized text using tf-idf method

	sentence 1	sentence 2	sentence 3	sentence 4	sentence 5	sentence 6	sentence 7
actual	0	0	0	0	0	0,377225	0
amount	0	0	0	0	0	0,377225	0
burden	0	0	0	0	0,360632	0	0
capex	0	0	0	0	0,360632	0	0
cut	0	0	0	0	0,360632	0	0
debt	0	0	0	0	0,299355	0	0,368759
decline	0	0	0	0,321262	0	0	0
decreased	0	0,260081	0	0	0	0	0
denominated	0	0	0	0	0	0	0,444241
dollar	0	0	0	0	0	0	0,444241
ebitda	0	0,43178	0	0	0	0	0,368759
end	0	0	0	0	0	0	0,444241
eventually	0	0,260081	0	0	0	0	0
exchange	0	0	0	0,321262	0	0	0
factor	0	0	0	0,321262	0	0	0
financial	0,27735	0	0	0	0	0	0
foreign	0	0	0	0,321262	0	0	0
grew	0	0	0	0	0,360632	0	0
h	0	0,260081	0	0	0	0	0
improved	0,27735	0	0	0	0	0	0
including	0	0	0	0,321262	0	0	0
loss	0	0	0	0,321262	0	0	0
mainly	0	0	0	0,321262	0	0	0
margin	0	0,260081	0	0	0	0	0
measure	0	0	0	0	0,360632	0	0
million	0	0,369071	0,56069	0	0	0,535305	0
monetary	0	0	0	0,321262	0	0	0
net	0	0	0,327978	0	0	0	0,368759
non	0	0	0	0,321262	0	0	0
onyear	0	0	0,395114	0	0	0	0
performance	0	0,260081	0	0	0	0	0
period	0,27735	0	0	0	0	0	0
price	0,27735	0	0	0	0	0	0
profit	0	0	0,327978	0,266675	0	0	0
quarter	0,27735	0	0	0	0	0	0
recovery	0,27735	0	0	0	0	0	0
reporting	0,27735	0	0	0	0	0	0
result	0,27735	0	0	0	0	0	0
revenue	0	0,260081	0	0	0	0	0
rouble	0,27735	0	0	0	0	0	0
sale	0,27735	0	0	0	0	0	0
save	0	0	0	0	0,360632	0	0
scheduled	0	0	0	0	0	0,377225	0
supported	0,27735	0	0	0	0	0	0
took	0	0	0	0	0,360632	0	0
total	0	0,260081	0	0	0	0	0
usd	0	0,369071	0,56069	0	0	0,535305	0
volume	0,27735	0	0	0	0	0	0
weak	0	0,260081	0	0	0	0	0
weaker	0,27735	0	0	0	0	0	0

4.3 Appendix 3.

Distribution graphs of the variables used
Variable Tonalities

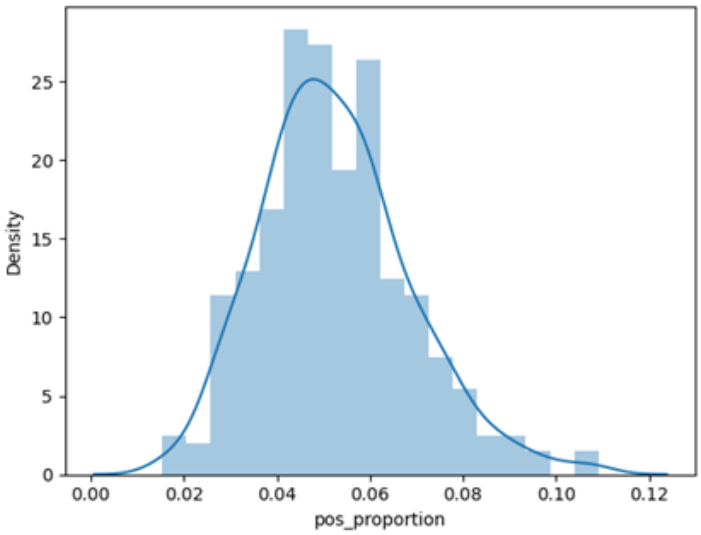


Figure 1. Distribution of the variable proportion of positive words in the text. Source: author's calculations

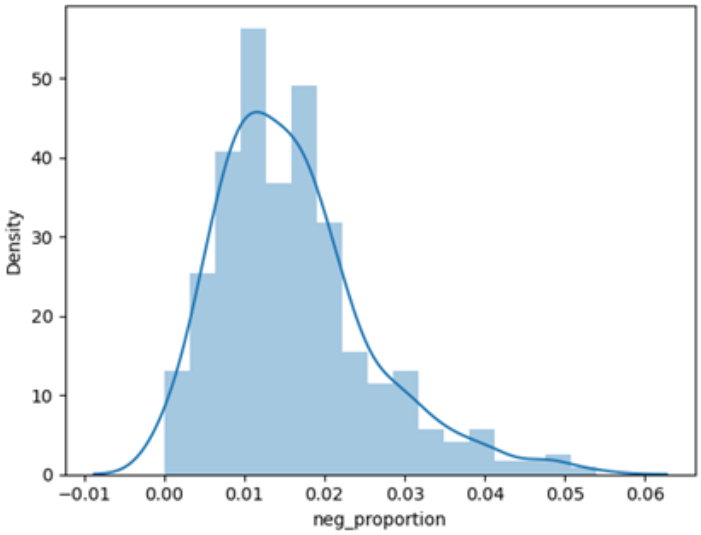


Figure 2. Distribution of the variable proportion of negative words in the text. Source: author's calculations.

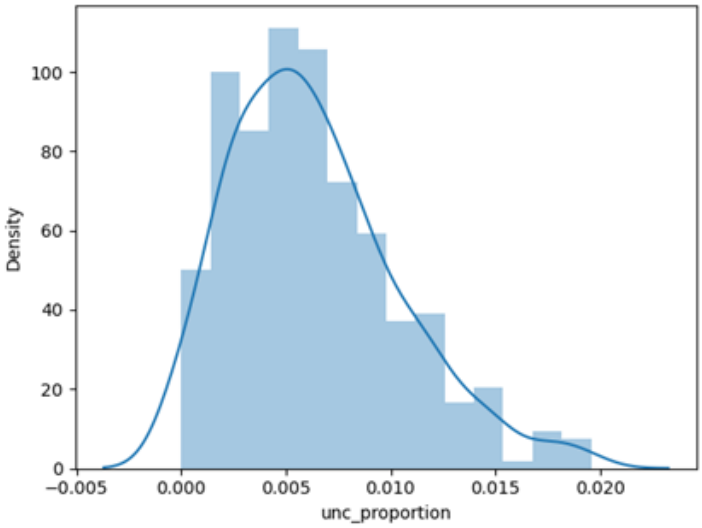


Figure 3. Distribution of the variable proportion of undefined words in the text. Source: author's calculations

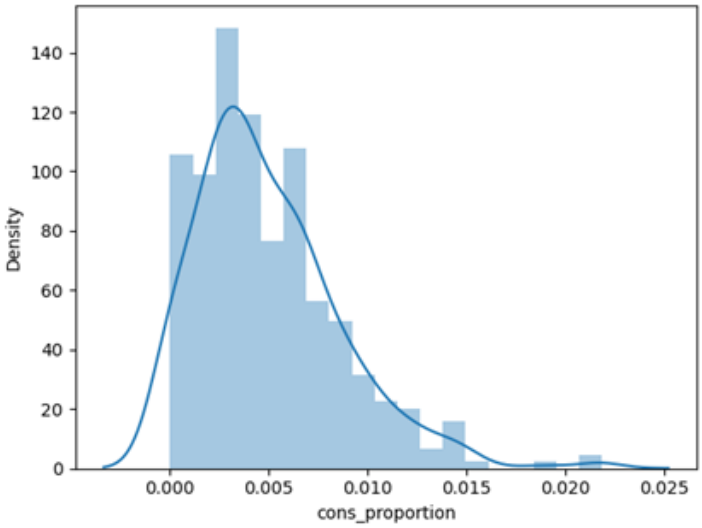


Figure 4. Distribution of the variable proportion of restrictive words in the text. Source: author's calculations

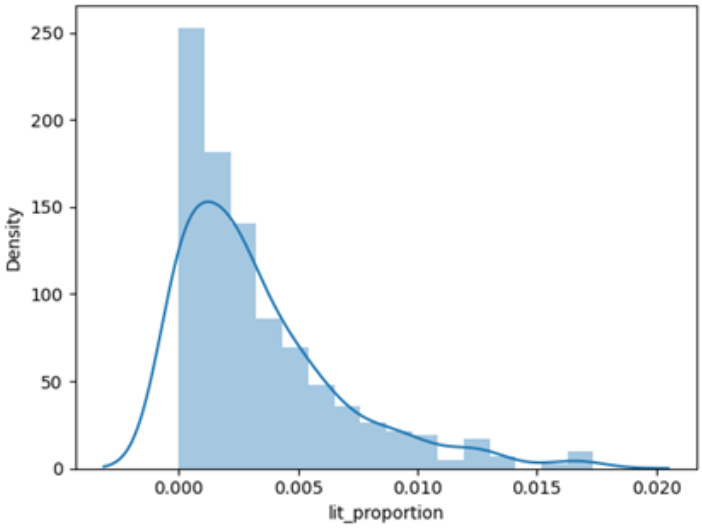


Figure 5. Distribution of the variable share of controversial words in the text. Source: author's calculations

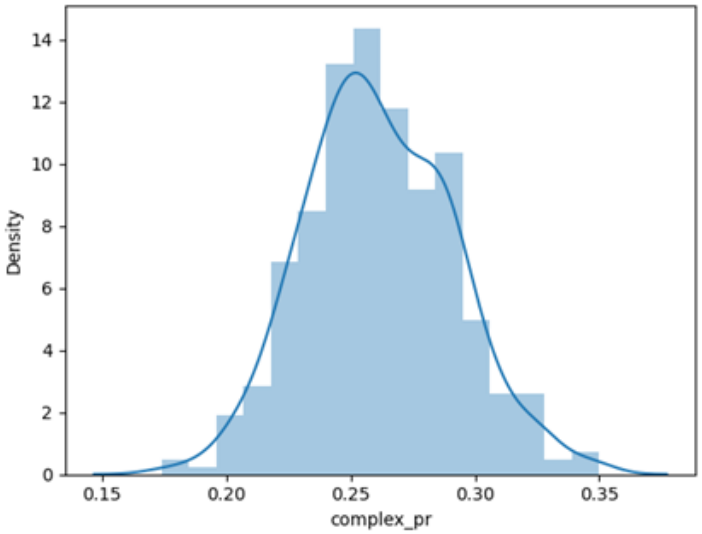


Figure 6. Distribution of the variable proportion of complex words in the text. Source: author's calculations

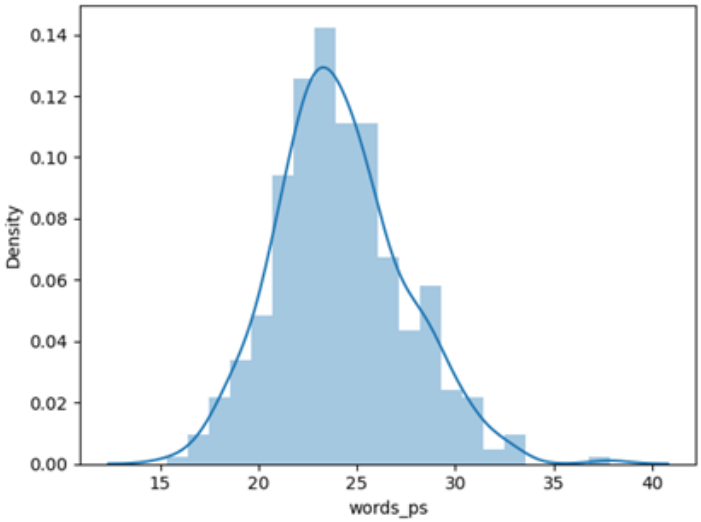


Figure 7. Distribution of the variable of the average number of words per sentence. Source: author's calculations

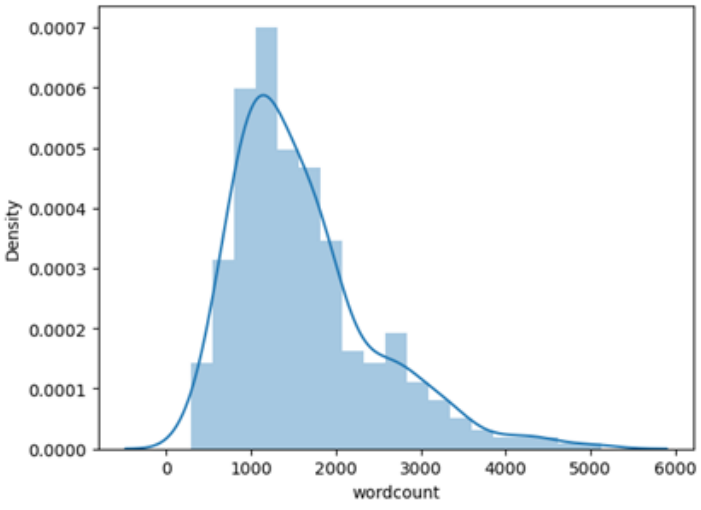


Figure 8. Distribution of the variable of the number of words in the text. Source: author's calculations

Control variables

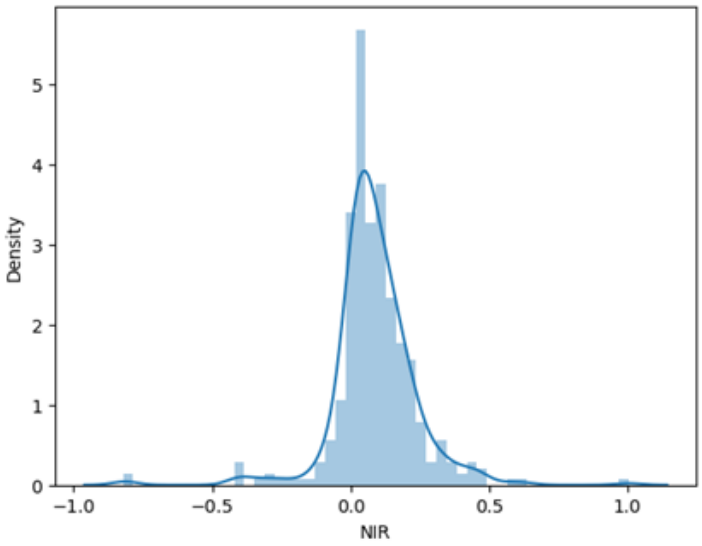


Figure 10. Distribution of variable profitability (Net Income / Revenue). Source: author's calculations

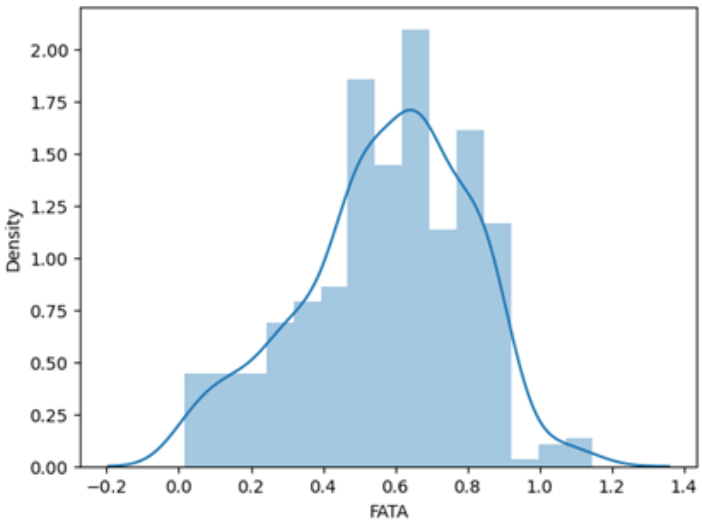


Figure 11. Distribution of the asset structure variable (Fixed Assets / Total assets). Source: author's calculations

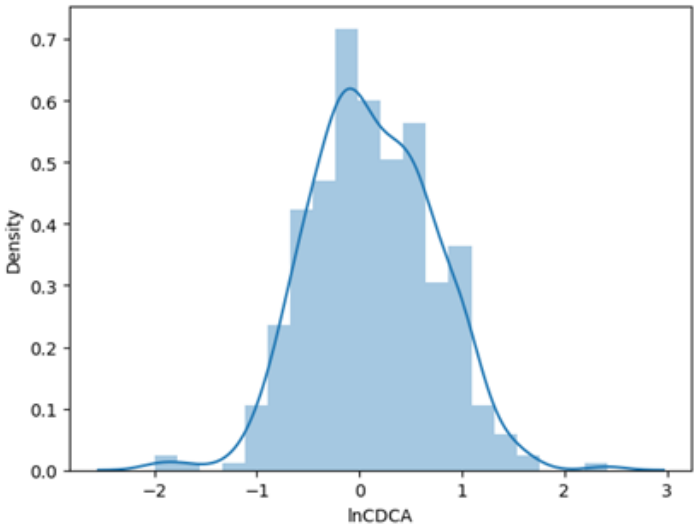


Figure 12. Distribution of the current liquidity variable (Current Debt / Current Assets). Source: author's calculations

4.4 Appendix 4.

Statistical basis of the models used

Logistic regression

The basis of logistic regression is the logistic function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

It is used in predicting probability as follows:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

where P is the probability of belonging to class 1, xn is the value of the dependent variable of the n-th feature, and beta n is the coefficient in the model for the n-th feature

KNN

Check for retraining	Metrics on the sample
accuracy train: 0.810 accuracy test: 0.784 AUC ROC train: 0.686 AUC ROC test: 0.619	accuracy: 0.803 precision: 0.684 recall: 0.402 f1-score: 0.506 AUC ROC: 0.670

Table 2. Quality metrics calculated for K prediction models including neighbors. Source: Author's calculations

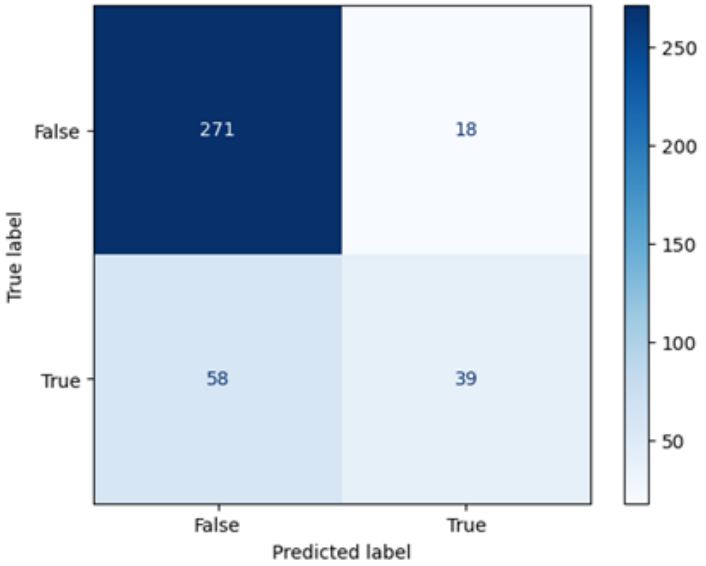


Figure 2. Confusion matrix calculated for K-nearest neighbor predictions. Source: author's calculations

Decision tree

Check for retraining	Metrics on the sample
accuracy train: 0.903 accuracy test: 0.639 AUC ROC train: 0.860 AUC ROC test: 0.509	accuracy: 0.837 precision: 0.685 recall: 0.649 f1-score: 0.667 AUC ROC: 0.775

Table 3. Quality metrics calculated for decision tree model predictions. Source: author's calculations

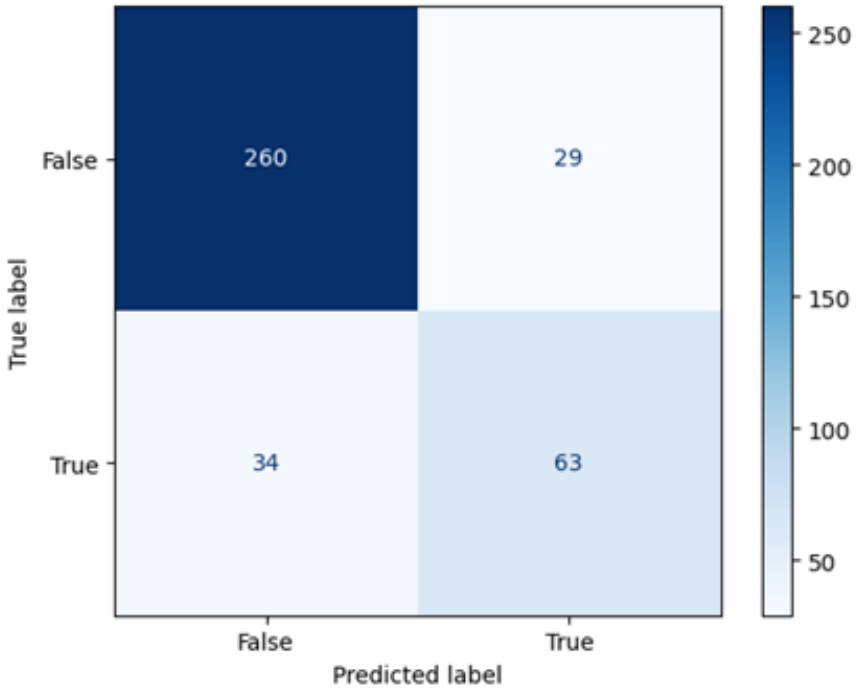


Figure 3. Confusion matrix calculated for decision tree model predictions. Source: author's calculations

Random forest

Check for training	Metrics on the sample
accuracy train: 0.893 accuracy test: 0.619 AUC ROC train: 0.866 AUC ROC test: 0.600	accuracy: 0.824 precision: 0.624 recall: 0.753 f1-score: 0.682 AUC ROC: 0.800

Table 4. Quality metrics calculated for random forest model predictions. Source: author's calculations

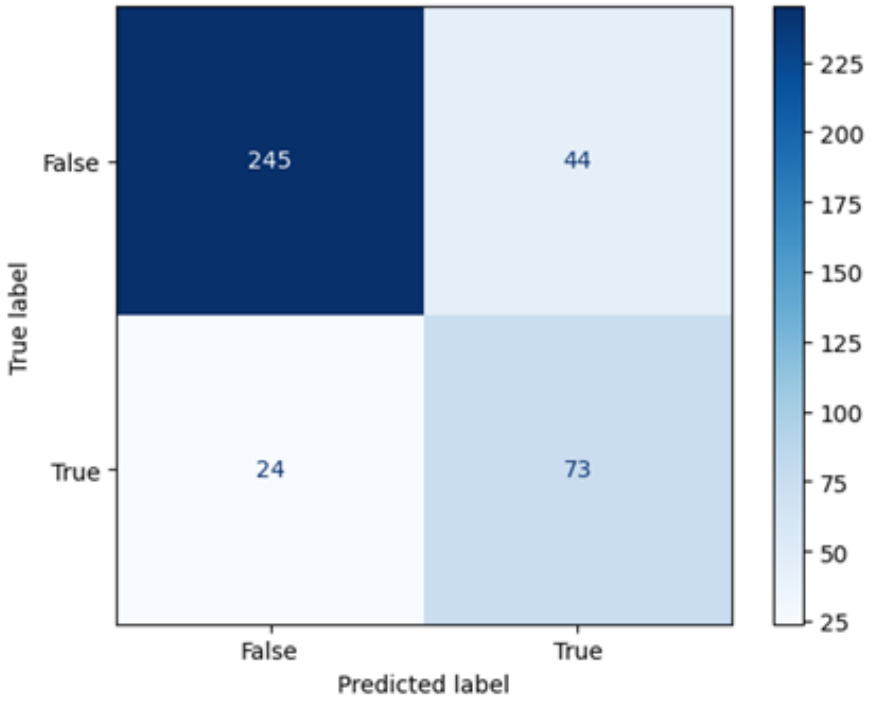


Figure 4. Confusion matrix calculated for random forest model predictions. Source: author's calculations

Support Vector Machine

Checking for training	Metrics on sample
accuracy train: 0.920 accuracy test: 0.773 AUC ROC train: 0.858 AUC ROC test: 0.597	accuracy: 0.883 precision: 0.882 recall: 0.619 f1-score: 0.727 AUC ROC: 0.795

Table 5. Quality metrics calculated for support vector machine predictions

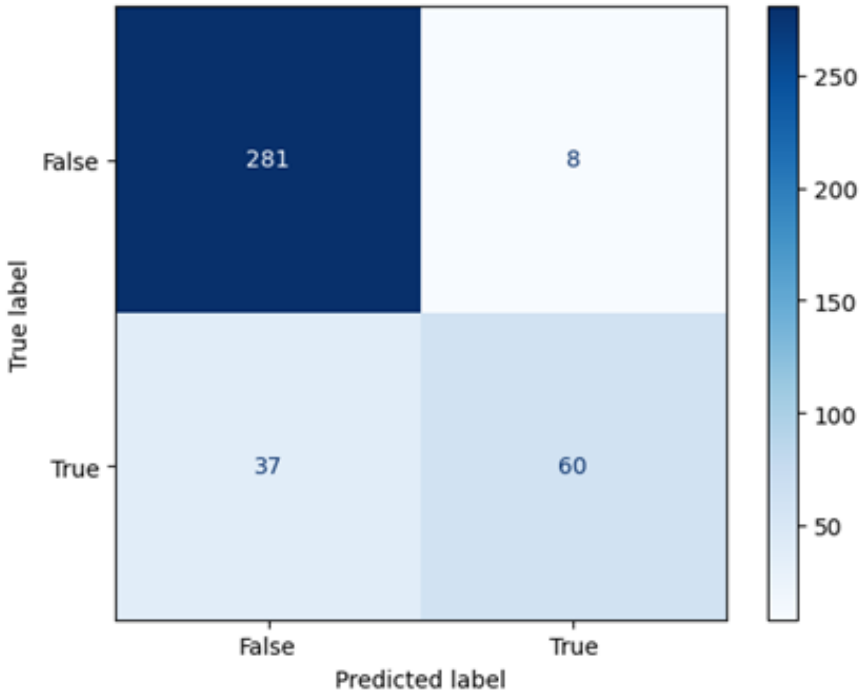


Figure 5. Confusion matrix calculated for SVM model predictions. Source: author's calculations

Comparison of the models

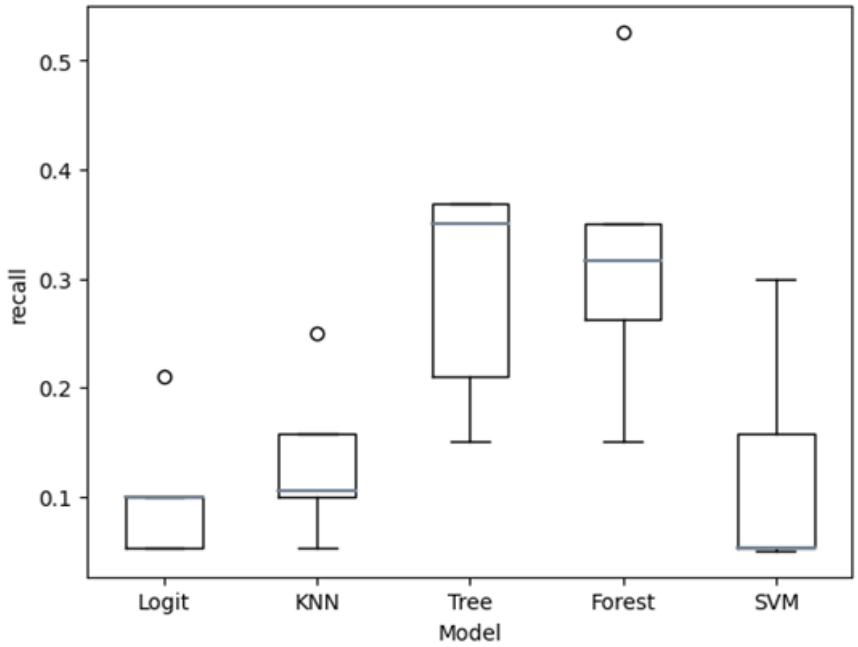


Figure 6. Comparison of the proportion of correctly classified fraud for different models. Source: author's calculations

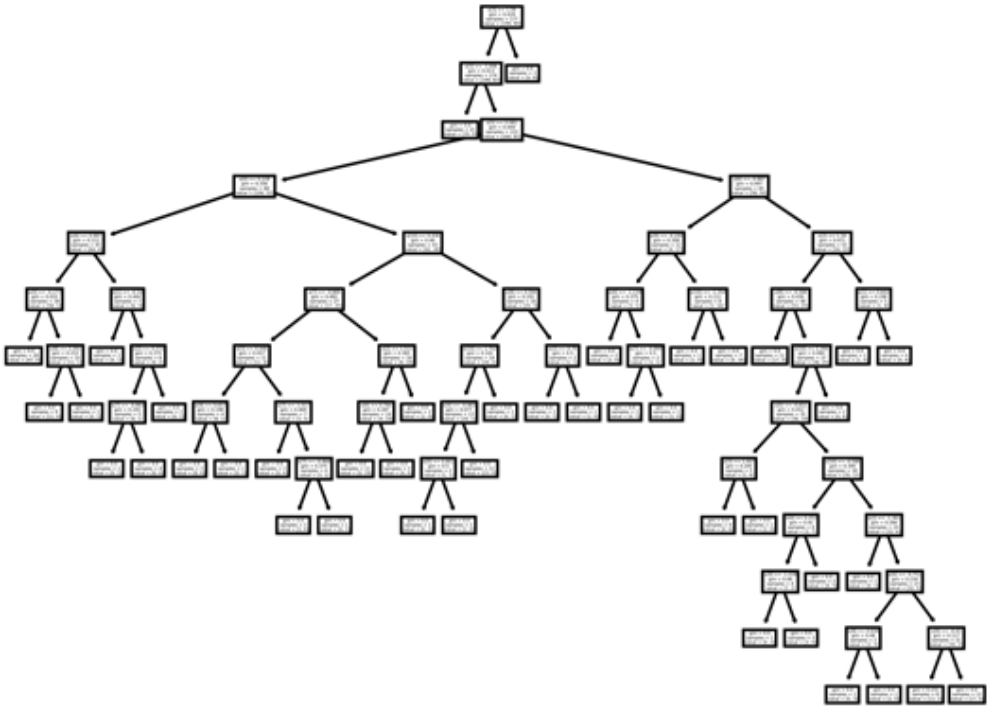


Figure 7. One of the decision trees used by the optimal random forest model to classify observations. Source: author's calculations