TALLINN UNIVERSITY OF TECHNOLOGY
School of Business and Governance

Artur Luik

DETERMINANTS OF ANNUAL REPORT SUBMISSION TIMELINESS IN ESTONIA

Masters's thesis
Programme Finance and Accounting, specialisation Finance

Supervisor: Laivi Laidroo, PhD
Co-supervisor: Tõnn Talpsepp, PhD

Tallinn 2024

I hereby declare that I have compiled the thesis independently
and all works, important standpoints and data by other authors
have been properly referenced and the same paper
has not been previously presented for grading.

The document length is 13048 words from the introduction to the end of the conclusion.

Artur Luik 07.05.2024
(date)

TABLE OF CONTENTS

ABSTRACT

The significant rate of non-submission and late submission of annual reports in Estonia compromises the accuracy of economic statistics and subsequently affects the quality of decision-making in both the public and private sectors. The objective of this thesis is to identify the determinants of annual report submission timeliness through the random decision forests ensemble learning method alongside the logistic regression model.

The initial dataset originates from an impact assessment study conducted for the Ministry of Finance in Estonia, consisting of 1 289 352 data rows with 285 company variables (e.g., age, size etc) and covering the years 2008 to 2018. Random forest classification algorithm alongside logistic regression model is implemented in Python, utilizing the Scikit-learn data analysis library. The dependent variable is timeliness which is equal to one if the annual report was submitted before or exactly on the deadline, otherwise, it is a non-timely submission equal to zero.

In line with previous studies, the results indicate that larger and older companies tend to prioritize timely reporting. Financial health indicators such as liquidity, absence of tax arrears, and lower leverage correlate positively with timely reporting, especially among small/medium enterprises. VAT obligation and longer reporting periods are associated with higher probabilities of timely annual report submissions across all company groups. Micro enterprises exhibit a nuanced relationship with fiscal year length, with both low and high values showing negative associations with timeliness. However, other factors such as the number of rulings related to late filing, market entry barriers, employment costs, and audit-related aspects may vary in their impact across different company groups. On average logistic regression and random forest behave similarly. For small/medium and micro groups non-timely predictions with random forest are 2% and 6% more precise across the years on average. In general random forest can predict 60% of non-timely submissions correctly.

Keywords: Timeliness, Annual report submissions, Mandatory disclosure, Financial reporting, Random forest ensemble learning, Logistic regression, Scikit-learn, SHAP.

INTRODUCTION

Timeliness of annual reports has recently emerged as an important topic in the context of European private companies, especially smaller ones, due to delays or non-submissions observed in private company reporting behaviours (Clatworthy, et al., 2016; Strouhal, et al., 2014). Similarly to their EU counterparts, according to the Accounting Act currently in force in Estonia, all legal entities must prepare and present their annual report to the registry six months after the end of a financial year. The failure or delay in the submission of the annual reports has a direct impact on the business activity statistics, which is an important input to the government (Bolívar & Galera, 2012). The coronavirus pandemic that started in December 2019 is a great example where the timeliness of the reports mattered for companies themselves. In a short timeframe, decisions to support the most impacted sectors were taken. In the tourism sector support package – "Support for partial compensation of losses resulting from the outbreak of the coronavirus causing the disease COVID-19 for tourism sector entrepreneurs" (Government of the Republic of Estonia, Regulation No 12, 2020) one of the requirements was the submission of last year's annual report. In such situations, the effectiveness of the support depends on the annual report submission rate.

Between 2010 and 2018 47-55% of Estonian legal entities failed to submit their annual report on time (Laidroo et al., 2020). The reasons for delayed annual report submissions in Estonia have been previously investigated in the 2020 report to the Ministry of Finance (Laidroo et al., 2020) using a survey alongside econometric analysis covering panel and logistic regression models. However, as the previous investigation was not heavily focused on quantitative analysis of the factors affecting the timeliness of annual report submissions, only part of the existing data was used in distinguishing companies that were late / did not submit the annual report from companies that submitted the annual report on time. The random forests method enables us to include more variables into consideration but offers lower levels of interpretability than logit models, in addition to that, random forest models are prone to overfit more easily than linear regression models. Random forest method usage in predicting fiscal stress events demonstrated 5 to 10 percentage points higher average accuracy in previous studies (Jarmulska, 2020).

The objective of this thesis is to identify the determinants of annual report submission timeliness through the random decision forests ensemble learning method alongside the logistic regression model. The thesis attempts to answer the following research questions

1. Which variables explain the timeliness of the annual report in three groups of companies (including micro enterprises, small and medium enterprises, audited enterprises)?
2. How does the annual report timeliness prediction process differ between the groups?
3. Which model can predict the timely submission of the annual report submission better?

The dataset used to answer these questions originates from the impact assessment of annual report submission timeliness for the Ministry of Finance (Laidroo et al., 2020). It is restricted for research purposes only and sourced from the RIK Information System, providing details on legal entities, including structure and financial data in XBRL format. Supplementary insights on VAT liability, tax arrears, and ownership structures were gathered from the Estonian Tax and Customs Board and RIK. Initially comprising 1 289 352 data rows and 285 variables, the dataset was accompanied by an Excel file detailing variable descriptions and data sources. Python programming language is used for running random forest and logistic regression algorithms with Scikit-learn (Pedregosa et al., 2011), Shap (Lundberg & Lee, 2017) for model interpretability, Pandas (McKinney, 2010) for data manipulation, and Matplotlib (Hunter, 2007) for data visualization. By utilizing these tools, the aim is to predict the timely submission of annual reports through binary classification. Annual reports submitted before or exactly on the deadline are considered timely submissions (1), everything else is considered not a timely submission (0). The author decides to use random forest because it is often preferred over logistic regression for its ability to handle complex, non-linear relationships between and target variables through ensemble learning, offering robustness against overfitting and capturing interactions among predictors more effectively (Breiman, 2001).

The thesis begins with a theoretical framework discussing mandatory disclosure and disclosure timing, followed by empirical evidence on the association between company attributes and financial reporting timeliness. The legal framework and existing empirical studies in Estonia are then overviewed. The second chapter focuses on data and methodology. The last chapter presents random forest and logistic regression model results together with interpretation to answer the research questions.

1. TIMELINESS OF ANNUAL REPORT SUBMISSIONS

1.1. Theoretical framework for mandatory disclosure and disclosure timing

Information disclosure refers to the act of companies sharing their financial and operational information with stakeholders. This can be mandatory, as required by law (e.g., annual report), or discretionary (e.g., sustainability report), where the company chooses what information to share (Verrecchia, 2001). This paper focuses on mandatory disclosures, specifically annual reports filed by private companies. These reports are meant to reduce information asymmetry, and the gap in knowledge between companies and stakeholders (investors, creditors, etc.). By disclosing information, companies might seek to improve capital allocation and social welfare (Minnis, 2017) or, alternatively, try to minimise the proprietary costs (Jacobson & Elliott, 1994).

It is possible to distinguish three different categories of disclosure research in accounting: "association-based disclosure", "discretionary-based disclosure", and "efficiency-based disclosure" (Verrecchia, 2001). Association-based disclosure investigates the relationship between exogenous disclosure, which is disclosure mandated by accounting standards or regulations, and investors' behaviour in financial markets. This type of research examines how the required disclosure of information is associated with stock prices and trading activity. For instance, studies in association-based disclosure might explore how the disclosure of a company's financial performance affects its stock price or how the disclosure of new accounting standards influences trading volume. Discretionary-based disclosure focuses on how managers exercise their judgment to disclose information that is not necessarily mandated by accounting standards. This type of research investigates the factors that motivate managers to disclose information, such as a desire to improve transparency, mitigate potential legal risks, or influence investor perceptions. For example, studies in discretionary-based disclosure might examine how a company's litigation history or the uncertainty of its financial reporting affects its decision to disclose additional information. Efficiency-based disclosure examines the question of which disclosure arrangements are most beneficial, assuming that we don't have any prior knowledge about the specific

information being disclosed. This type of research analyzes unconditional disclosure decisions, it considers the disclosure choices that would be optimal in general, without being influenced by the specifics of a particular company or situation. For instance, studies in efficiency-based disclosure might explore the relative advantages of mandatory disclosure versus voluntary disclosure, or they might investigate the costs and benefits of disclosing different levels of detail in financial statements. (Verrecchia, 2001). This thesis shares the greatest similarities with discretionary disclosure research. However, the context of mandatory disclosure creates some distinct differences from discretionary disclosure research.

Mandatory disclosure is usually approached from the perspective of benefits and costs for the information disclosure as explained in Eirele (2008) and Wittmann (2020). At the level of legal entities, the preparation and submission timing of annual reports are influenced by the following associated costs (Leuz & Wysocki, 2016):

- Direct costs – administrative costs related to compiling, auditing, and publishing reports.
- Indirect costs – this is about sharing information about a company with other people or organizations. It could mean secret business details getting out to competitors, facing lawsuits for giving out wrong information, changes in how much it costs to borrow money, managers losing their jobs if the company does poorly, and owners losing their privacy.

Direct costs associated with the preparation, auditing, and publishing of financial statements can be significant (Leuz & Wysocki, 2016). Usually, these costs could have a greater impact on small and micro-enterprises as well as non-profit organizations and foundations. The costs of preparing financial statements may not be considered necessary for operations because the failure of submission might not increase the total costs. To create monetary incentives, governments have introduced monetary sanctions for non-submission. The stronger the sanctions, the bigger the motivation for companies to meet the deadline (Clatworthy and Peel, 2016; Luypaert et al., 2016). At the same time, the indirect costs may also influence the disclosure decisions of companies, increasing the likelihood of delayed disclosure (Wittmann, 2020).

The time dimension of the costs/benefits tradeoff associated with disclosure timing for private companies is covered in Laidroo et al. (2024). It highlights that companies weigh the costs (administrative, compliance, and non-compliance) against the benefits of disclosure to determine the optimal disclosure time. Non-compliance costs are the financial penalties and operational disturbances incurred due to failing to meet regulatory obligations, such as fines and late

submission fees from missed deadlines for mandatory information provision (Laidroo, et al., 2024). Compliance costs may involve privacy and proprietary expenses. Privacy costs are common in small private companies, where revealing the owner's wealth through annual reports can jeopardize their privacy and raise concerns about tax evasion (Arruñada, 2011). Proprietary costs relate to disclosing sensitive company information like strategies and operational statistics that may compromise competitive advantage, highlighting the delicate balance between transparency and protecting proprietary interests (Jacobson & Elliott, 1994). Benefits of disclosure relate to annual report disclosure advantages by addressing agency costs, which encompass expenses incurred due to the mitigation of information imbalances among a company's stakeholders (Laidroo et al., 2024).

Factors such as company size, ownership structure, industry competitiveness, and regulatory environment influence the size of costs and benefits (Laidroo et al., 2024). As seen from Figure 1, late filing may occur if the benefits outweigh the costs, with the optimal disclosure time (t*) being after the official deadline. However, if administrative costs are disproportionately high, some companies may find it more beneficial to become non-filers because no matter how much time has passed, the total costs exceed the benefits of disclosure. Although the model is limited with its one-period nature, inability to account for multi-period impacts, and variations in costs and benefits across companies and countries it is the only framework fully capturing the mandatory disclosure timing.

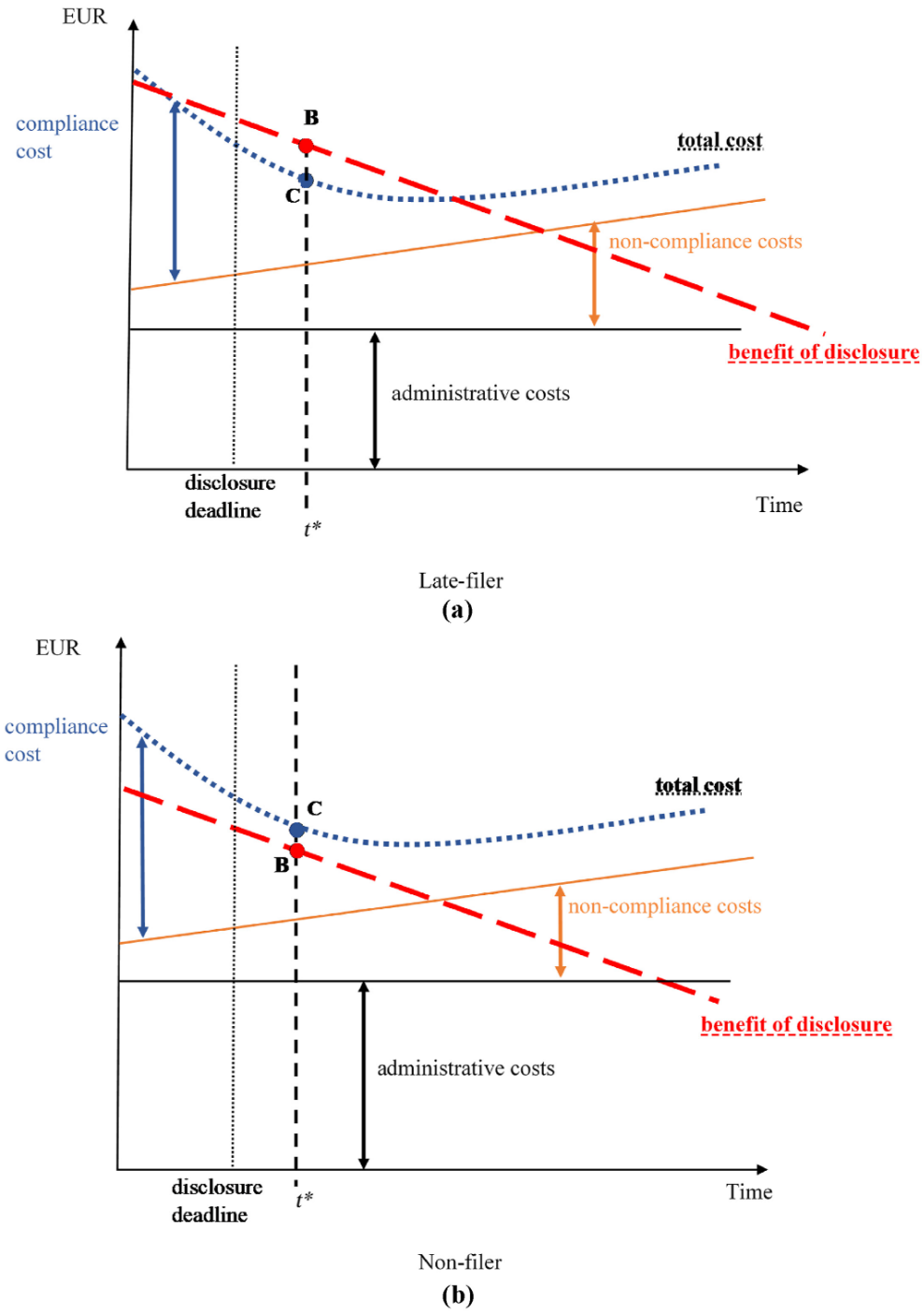


Figure 1 Cost-benefit framework for mandatory annual report disclosure timing (Laidroo et al., 2024)

Weetman (2004) links the timeliness of financial reporting with theories of proprietary costs, information cost saving and relative good/bad news. According to Weetman (2004), while regulatory requirements mandate that all companies release their financial statements by a certain deadline, there is a wide variation in reporting timeliness within this deadline. He found that some

companies release their statements very quickly, while others wait until the maximum allowable time. Companies with higher information costs, such as those with a lot of trading activity or that have recently issued shares, tended to release their financial statements more quickly. This is because timely disclosure reduces information asymmetry between the company and investors, which can lead to lower trading costs for the company. Companies in industries with lower barriers to entry or that are less concentrated were more likely to release their financial statements quickly. This is because they have less to fear from competitors learning about their financial performance. Companies with more favourable news tended to release their financial statements more quickly. This is because they want to capitalize on positive market reactions. Conversely, companies with unfavourable news tended to delay the release of their financial statements in order to avoid negative market reactions.

Overall, the theoretical models show that different costs and benefits explain some of the disclosure timing aspects. As noted by Laidroo et al (2024) disclosure decisions may also be influenced by the information environment (e.g., the presence of report submission reminders) and legal requirements (e.g., approvals needed for filing the reports). The main empirical findings are discussed in the following sub-section.

1.2 Empirical evidence on the association between company attributes and financial reporting timeliness

Numerous studies have examined company-specific factors' association with the timeliness of the reporting in the context of private companies outside of Estonia. One of the most recent papers is by Wittman (2020) which examines why private firms delay the public release of their financial statements despite being subject to the same disclosure rules as listed firms. Using a sample of 1000 large private companies in Germany, the study shows that firms facing higher proprietary costs, reporting losses or outperforming peers, tend to delay filings. Similarly, firms with perceived greater competition or family ownership delay disclosure. Some companies accept monetary sanctions for missing deadlines, suggesting that the benefits of delayed disclosure outweigh the costs. The study highlights the importance of understanding proprietary costs and the role of disclosure timing in shaping transparency among private firms.

The timing of reports may be influenced by the level of experience and expertise within the company, which tends to be lower in younger firms that may lack specialized personnel. This proposition is based on the learning curve theory (Owusu-Ansah, 2000). Several empirical studies have confirmed the observation that younger private companies tend to take longer to release their financial reports (Breuer et al., 2020; Wittman, 2020; Clatworthy & Peel, 2016).

The timeliness of annual reports has been shown to vary based on the size of the company - larger firms, which typically engage with a greater number of stakeholders, should have a higher benefit from fast disclosure (Breuer et al., 2020). Consequently, they often minimize delays in releasing their annual reports. A similar result was found in Eierle (2008) covering 22 108 firm-year observations on small and medium-size Belgian and Luxembourgian firms. Larger firms, potentially facing more severe agency problems, have a greater need for monitoring, leading to shorter reporting lags (Eierle, 2008). However, in some cases, the extensive complexity of reporting within very large companies may prolong the disclosure process, counteracting the presumed benefits of their size (Wittman, 2020).

The relationship between a company's capital structure and liquidity can also affect disclosure. When a company is heavily leveraged, the agency costs increase, necessitating faster disclosure of annual reports. Bigus and Hillebrand (2017) have demonstrated that firms with multiple bank relationships tend to disclose their reports more promptly. Conversely, highly leveraged firms facing financial distress may seek to delay disclosure to conceal poor performance from stakeholders. Other studies among private firms, such as those by Breuer et al. (2020), Clatworthy and Peel (2016), Luypaert et al. (2016), and Lukason and Camacho-Miñano (2019), consistently find longer delays in disclosure among more leveraged firms. In contrast, considerations regarding liquidity may lead to different outcomes compared to those for leverage. Firms with higher liquidity requirements may not feel pressured to submit their annual reports quickly, as they have less need for immediate capital. Conversely, less liquid firms might delay disclosure to avoid revealing their financial difficulties. Clatworthy and Peel (2016), Luypaert et al. (2016), and Bigus and Hillebrand (2017) found no significant relationship between liquidity and annual report disclosure delay, indicating ambiguity in this regard. However, Breuer et al. (2020), Clatworthy and Peel (2016), and Lukason (2019) support a negative association between liquidity and reporting delays. Additionally, Luypaert et al. (2016) have linked longer reporting delays to a higher likelihood of corporate default.

Based on the research conducted by Luypaert et al. (2006) and Clatworthy and Peel (2016), it can be concluded that companies which have previously submitted their annual financial reports on time are more likely to continue doing so in the future. Similarly, the probability of late submission of financial reports is higher for those companies that have been late in previous years as well. These findings suggest a pattern of consistency or lack thereof in the timing of financial reporting, indicating that past behaviour is indicative of future behaviour in this regard.

Annual reports are important input for credit issuers (e.g. banks). Maingot and Zeghal (2006) find that Canadian small firms consider borrowing as the main reason for preparing financial statements. Therefore, there should be a positive relationship between leverage and financial statement submission rate. This claim is supported by Bigus and Hillebrand (2017). However, companies with poor financial standing feel less motivated to expose their finances. This has been demonstrated by Luypaert et al. (2016) – a higher probability of corporate default has been associated with longer reporting delays.

The empirical evidence from public companies comes in various forms. Previous studies have investigated a number of firm-, economic-, and auditor-specific characteristics associated with financial reporting timeliness. Sufiyati (2017) investigated the empirical evidence regarding the influence of profitability, size, financial leverage, liquidity, and age on the timeliness of financial reporting in manufacturing companies listed on the Indonesia Stock Exchange from 2011 to 2013. The research employed a purposive sampling method and included 195 companies in its analysis. Data processing was conducted using SPSS software version 20.00 for logistic regression. The findings indicate that firm size has a positive and significant effect on the timeliness of financial reporting, suggesting that larger companies tend to report their financial information in a more timely manner. Firm age was found to have a negative and significant effect on the timeliness of financial reporting, implying that older companies may experience delays in their reporting processes. However, the study did not find any significant effects of profitability, financial leverage, or liquidity on the timeliness of financial reporting within the context of the manufacturing companies studied.

Türel (2010) analysis of 211 listed companies in Turkey sheds light on various factors affecting timely financial reporting. Empirical results indicate that company size, auditor firm, income status, audit opinion, and the industry significantly influence reporting lead times. Specifically, companies reporting net income with standard audit opinions tend to release their financial

statements earlier. Conversely, those audited by the Big Four audit firms and operating in the manufacturing industry are more likely to be late reporters. Furthermore, smaller audit firms exhibit efforts to avoid delays, while the extensive clientele of Big Four audit firms may contribute to delays in their auditing processes. Similar evidence can be found in Nigeria where Okougbo (2014) examined factors affecting the timeliness of financial reporting among Nigerian financial institutions, focusing on samples from the banking and insurance sectors between 2005 and 2008. Thirty-three institutions were analyzed using Generalized Least Square (GLS) regression. On average, financial institutions in Nigeria took about four months to release their financial reports, with banks performing better (three months) than insurance companies (five months). Results showed that company size had a negative and significant impact on reporting timeliness, while company age had a positive and significant effect. Leverage and performance had negative effects on timeliness at certain significance levels. Audit type did not significantly influence reporting timeliness, as auditors cannot alter reporting timelines without their clients' cooperation.

These results indicate that several company-specific factors are associated with the timeliness of financial reporting by private companies. Factors such as firm size, age, leverage, liquidity, profitability, and past filing history all play a role in how quickly companies release their annual reports. Previous empirical research relies mostly on regression models set to extract meaningful insights from data as concluded in Table 1.

Table 1 Methods used in empirical research

Paper	Method	Dependent variable
Wittman (2020)	Hazard model	Lag (days)
Luypaert et al. (2016)	Logistic regression	Late (0 or 1) and Ordlate (categories of lateness)
Clatworthy and Peel (2016)	Logistic regression	Lag (days)
Lukason and Camacho-Miñano (2019)	Logistic regression	Lag (days)
Laidroo et al. (2020)	Logistic regression	Ordlates; Lag; Late (0 or 1)
Laidroo et al. (2024)	Logistic regression	Filler (1) non-filler (0)
Sufiyati (2017)	Logistic regression	Late (0 or 1)
Türel (2010)	Multivariate regression	Lead time (Days)
Okougbo (2014)	Generalized Least Squares	Lag (days)

Source: Author based on papers listed

The main dependent variables in these papers have been lag in days or lateness (0 indicating late submissions or 1 indicating on-time submissions). The dependent variable in this paper is

timeliness which is equal to one if the annual report was submitted before or exactly on the deadline, otherwise, it is a non-timely submission equal to zero. Related papers do not use machine learning-based methods, however, Jarmeulska (2020) found that ensemble learning methods such as the random forest model demonstrate slightly higher effectiveness in predicting fiscal stress events. With an average prediction accuracy of nearly 80%, the random forest outperformed the logistic regression models, which achieved accuracies between 70 and 75%. Previous research also finds non-linear relationships between companies' capital structure and their profitability (Kohv, 2021). That indicates that there is room for new models for empirical research.

Previous studies concerning Estonia are discussed in the following section.

1.3 Overview of legal framework and empirical studies in Estonia

Financial reporting requirements in Estonia are based on The Accounting Act (RPS) § 14. At the end of every fiscal year, all business entities registered in Estonia must present an annual report comprising financial statements and a managerial overview (based on Commercial Code (ÄS) § 2 and 97, Non-profit Associations Act (MTÜS) § 1 and 36 and Foundations Act (SAS) § 1 and 34). According to RPS § 15, the number of components included in the report varies for companies depending on their size. According to RPS § 3 sections 14 to 17 companies are split into four groups

- A microenterprise is a private limited company (osaühing) meeting the following criteria on the balance sheet date of the financial year: total assets up to 175 000 €, liabilities not exceeding equity, one shareholder who also serves as a board member, and annual sales revenue up to 50 000 €.
- A small enterprise is a business entity registered in Estonia that is not classified as a microenterprise. For a small enterprise, only one of the following criteria may be exceeded on the balance sheet date of the financial year: total assets up to 4 000 000 €, annual sales revenue up to 8 000 000 €, and average number of employees during the financial year up to 50 individuals.
- A medium-sized enterprise is a business entity registered in Estonia that is not classified as a microenterprise or a small enterprise. For a medium-sized enterprise, only one of the following criteria may be exceeded on the balance sheet date of the financial year: total

assets up to 20 000 000 €, annual sales revenue up to 40 000 000 €, and average number of employees during the financial year up to 250 individuals.

- A large enterprise is a business entity registered in Estonia where at least two of the following criteria are exceeded on the balance sheet date of the financial year: total assets up to 20 000 000 €, annual sales revenue up to 40 000 000 €, and average number of employees during the financial year up to 250 individuals.

According to RPS § 14 microenterprises have simplified reporting obligations. Their annual financial statements consist of at least two primary statements (balance sheet, profit and loss account) and additional notes. Small enterprises have slightly more extensive reporting requirements compared to microenterprises. Their annual financial statements also consist of at least two primary statements (balance sheet, profit and loss account) and additional notes. For larger entities beyond micro and small enterprises, the reporting obligations are more comprehensive. These entities are required to prepare and disclose a full set of financial statements, including the balance sheet, profit and loss account, cash flow statement, and statement of changes in equity, along with additional notes. Regardless of the size, the purpose of financial reporting remains consistent, aiming to provide relevant and reliable information about the entity's financial position, performance, and cash flows to interested parties.

The Commercial Code complements the Accounting Act by creating the legal basis for reporting. According to Commercial Code § 97 subsection 1, § 179 subsection 4, § 334 subsection 2, Non-profit Associations Act (MTÜS) § 36 subsection 5, and Foundations Act (SAS) § 34 subsection 4, the annual report shall be submitted to the commercial register within 6 months from the end of the financial year. In this study, reports submitted within this deadline are considered timely. There is a separate rule for public interest entities, which include publicly listed companies, credit institutions, and insurance companies, who are required to publish their audited annual reports two months earlier than other companies, within four months after the end of the financial year (Securities Market Act § 110). If the annual report is not submitted on time or remains unsubmitted, the registrar has the right to impose fines without warning on both the legal entity and all persons obligated to submit the report under the Commercial Code § 71. If the annual report is not submitted within six months after the deadline, the registrar initiates a supervisory procedure against the legal entity under Commercial Code § 60, Non-profit Associations Act § 361, or Private Limited Companies Act § 341, which may result in the deletion from the register or compulsory dissolution of the legal entity.

In Estonia, there are four noteworthy publications that address the issue of financial reporting timeliness. Two papers - “Factors causing failure to submit annual reports (based on Estonian legal entities)” and “Reasons for Late Filings of Annual Reports in Estonia” – both used qualitative methods to understand reasons for late fillings. Both papers concluded that annual report fillings are late because the process is complex, and time-consuming and often companies don’t have available accountants (Kallakas, 2021; Kips, 2021). The 2019 article investigated the relationship between firms' reporting delays and bankruptcy risk in Estonia (Lukason and Camacho-Miñano, 2019). The findings suggest that firms with lower liquidity and profitability are more likely to delay reporting, and higher bankruptcy risk is associated with reporting delays, highlighting potential implications for stakeholders and the need for stricter measures by state institutions. The most comprehensive report to understand the situation in Estonia is “Impact assessment of annual report submission timeliness for the Ministry of Finance” (Laidroo et al., 2020) which uses a survey alongside econometric analysis covering panel and logistic regression models. The analysis has two dependent variables – filling delays (the number of calendar days from the end of the reporting year to the submission of the report) and binary variable late (1 – submission was late, otherwise 0). The variables affecting the filling delays are brought out in Table 2 and Table 3.

Table 2 Attributes having a positive relation with filing delays

Variable	Relationship
Previous Year's Late Submission	Late submission in the previous year leads to delays
Tax Arrears	Tax arrears lead to delayed submission
Reporting Period Length	Longer reporting periods lead to delayed submission
Report Quality	Lower quality reports lead to delayed submission
Consolidation	Consolidated reports lead to delayed submission
Year End Loss	Ending the year with a loss leads to delayed submission
Financial Distress	Net asset reduction below critical level leads to delayed submission

Source: Author, based on the impact assessment by Laidroo et al. (2020)

Table 3 Attributes having negative relation with filling delays

Variable	Relationship
Size	Larger companies submit faster
Age	Older companies submit faster
International Activity	Higher international sales result in faster submission
Diversification	More diversified revenue leads to faster submission
VAT Obligation	VAT obligation leads to faster submission
Board Size	Larger board size leads to faster submission
Fiscal Year Ending in December	Companies with fiscal years ending in December submit earlier
Submission Complexity	PDF or notary submission is faster compared to XBRL
Profitability	Higher profitability leads to slightly faster submission
Financial Leverage	Higher leverage leads to slightly faster submission
Cash Ratio in Assets	Higher cash ratio has negligible effect on submission timing

Source: Author, based on the impact assessment by Laidroo et al. (2020)

The purpose of the report was to identify the causes and consequences of non-submission and late submission of annual reports and determine the most suitable and effective solutions that would motivate entrepreneurs to submit annual reports by the deadline. Based on the objective the report also provided recommendations for improving the current situation. The analysis of data showed that from 2010 to 2018 on average 23-24% of legal persons failed to submit their annual report. 28-29% of legal submitted the report after the deadline. The econometric model for private limited companies across size indicated 27 statistically significant factors. The report found that as the age of the company increases, the probability of the company failing to submit a report decreases. The probability of a VAT-registered company failing to submit an annual report is 85% lower than that of a non-VAT-registered company. As the number of members of the company's board increases, the probability of the company failing to submit a report decreases. The probability of a company with a fiscal year ending in December failing to submit a report is 29% lower than that of a company with a fiscal year ending at other times. Significant differences in the likelihood of non-submission compared to the base value are observed across different sectors, compared to the agricultural sector.

2. DATA AND METHODOLOGY

2.1 Dataset

The dataset originates from the impact assessment of annual report submission timeliness for the Ministry of Finance (Laidroo et al., 2020). The data is not publicly available and is authorised to be used only for research purposes. The primary data source was the RIK Information System, which provided essential information about legal entities, including their legal structure and financial data from annual reports in XBRL format. Additionally, data from the Estonian Tax and Customs Board and the RIK was used to gather insights into VAT liability, tax arrears, and ownership structures of certain companies.

The initial dataset contained 1 289 352 data rows. The data file was accompanied by an Excel file that described the variables and the source of the data. There were 285 variables that include:

- Financial reporting and audit variables
    - Information regarding whether the financial statements are consolidated and audited.
    - Deadlines for submitting reports.
    - Information about delays in submitting reports.
    - Type of audit conducted (voluntary or mandatory).
    - Various financial figures such as revenue, expenses, assets, liabilities, and profits.
- Company characteristics
    - Legal form of the company.
    - Number of employees.
    - Age of the company.
    - Business segments and industry classifications.
    - Size of the company based on various metrics.
- Performance metrics
    - Profitability ratios (ROA, ROE).
    - Financial leverage ratios.

    - Liquidity ratios.
    - Growth rates of revenue or profits.
    - Altman's bankruptcy prediction score.
    - Zmijewski score.
    - Cash flow from operations.
    - Dividend information.
    - Market share information.
- Other variables
    - Dates and durations related to various reporting and auditing activities.
    - Submission status and timeliness of financial reports.
    - Specific indicators related to financial health and risk.
    - Information about external factors affecting financial performance.
    - Variables related to specific accounting practices and standards.

The initial dataset covers all registered business entities, non-profit organizations, and foundations in Estonia operating for at least one year from January 1, 2010, to December 31, 2018. Each entity was included in the dataset for each year it was operational during the specified period. Adjustments were made to exclude entities founded after June 30 of their establishment year if they failed to submit reports for that fiscal year and to remove observations associated with entities deregistered within the fiscal year without submitted reports. These changes resulted in a reduction of 127,686 observations in the final dataset. While there might be some margin of error due to multiple registrations and deregistrations of the same entity over the years, the dataset includes liquidated entities. The raw data was modified through feature engineering to accelerate the analysis and ensure that the data was well-prepared for the application of the statistical models. The original panel dataset was structured into distinct groups based on the research questions – micro, small and medium, large and audited groups that included Joint-stock company (AS) and Private limited company (OÜ) companies in Estonia. After the initial filtering, 999 385 company-year observations remained with 162 variables.

The data is unevenly distributed both across the years and groups. A visual representation in the form of Figure 2 showcases the count of observations in these groups, revealing an uneven distribution between the sizes of the groups. This unevenness can influence our analysis, introducing potential biases and challenges that must be carefully addressed when training the model.

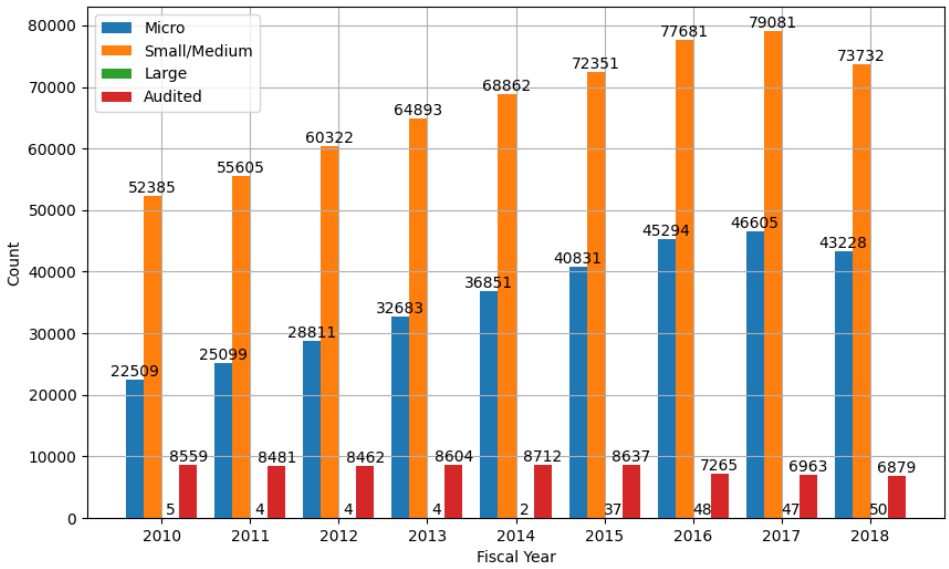


Figure 2 Count of companies in each group

In addition to the uneven group distribution, some variables exist exclusively within certain subsets of the data as demonstrated in Figure 3. During the variable selection, this characteristic of data will lead to different variable selection results for each group.

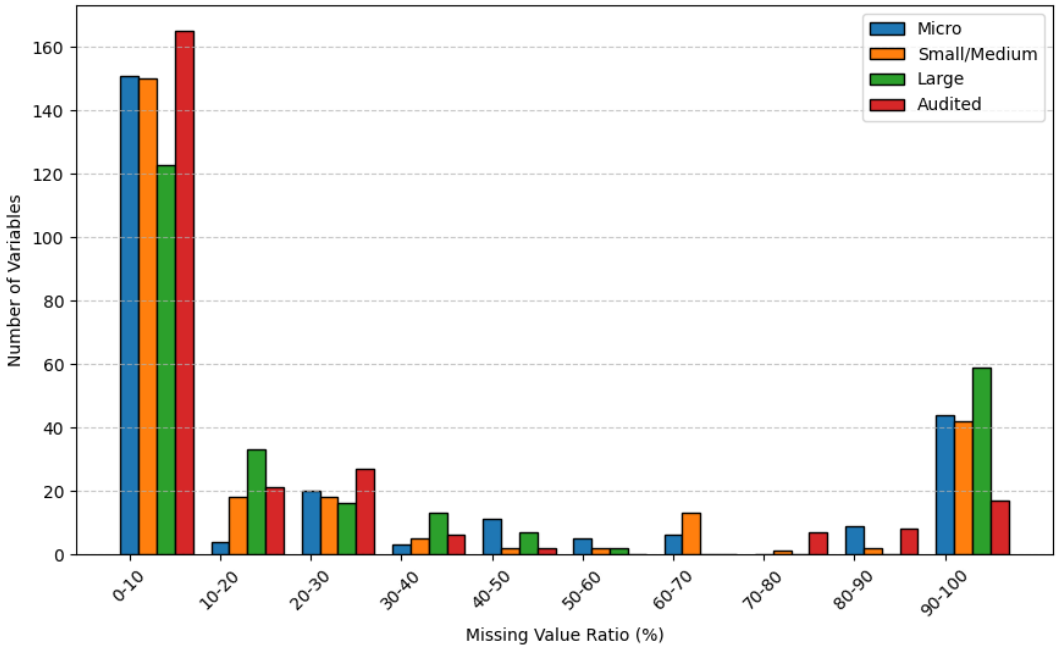


Figure 3 Distribution of missing value ratios for different groups

Appendix 1 presents all the variables and their descriptions taken into consideration across the groups. Appendix 2 provides descriptive statistics, including measures like mean, minimum, maximum and standard deviation for micro entities. Appendix 3 is the same for small and medium-sized entities and Appendix 4 is for audited entities. These statistics summarize data characteristics, aiding in data interpretation.

2.2 Data preparations

Data preparation involves cleaning, organizing, and transforming raw data to ensure its quality, consistency, and suitability for the intended research objectives, ultimately enhancing the reliability and validity of the final thesis findings (Pyle, 1999). Multiple filters were employed to reduce the variable count before employing the statistical models:

1. High Missing Value Ratio Filter: Fields with a high missing value ratio (>=70%) were removed. The missing value ratio is defined as the proportion of missing or incomplete data points in a particular variable within the dataset. High missing value ratios can distort the integrity of the dataset, potentially leading to biased results or reduced statistical power. Missing value distribution can be seen in Figure 3.
2. Temporal Variable Filter: Temporal variables related to time were eliminated. The model's training methodology, conducted on a yearly basis, removes the necessity of preserving specific dates for individual observations. Relevant dates with significance were already incorporated as distinct variables in the dataset. (For example, submit_time information already exists in the timely variable). The following variables were dropped
    a. Dummy variables (yr1, yr2 … to indicate the year of the observations)
    b. Audited date / submitted date
    c. Fiscal year start/end date (related to length of fiscal year)
    d. Registration date (related to company age already)
    e. Annual report-related deadlines/submission dates (not interesting, already incorporated into other variables)
3. Zero Variance Filter: A zero variance filter was applied to eliminate irrelevant variables that are unlikely to contribute to the model's performance. If all the values are constant in the dataset, there is no expected relevance for predicting the outcome.
4. Correlation Filter: Variables that have a high correlation (more than 0.9) were considered for removal to reduce the risks of overfitting. Variables that were known to be interesting

through previous research (e.g., size, age, financial leverage, fiscal year lengths) were "locked" and not allowed to be removed, so high correlation would not remove them accidentally. Due to the large amount of variables, detailed pairwise correlation after cleanup is presented as an electronic appendix (Luik, 2024).

The results of the variable reduction are displayed in Figure 4. After the ratio filter, the number of variables in the groups was reduced to approximately 200, down from the initial count of 284. The temporal variable and variance filter both had a marginal impact number of variables – together removing 10% of the variables. The impact of correlated variable cleanup, every group experienced a substantial reduction to approximately 100 variables.

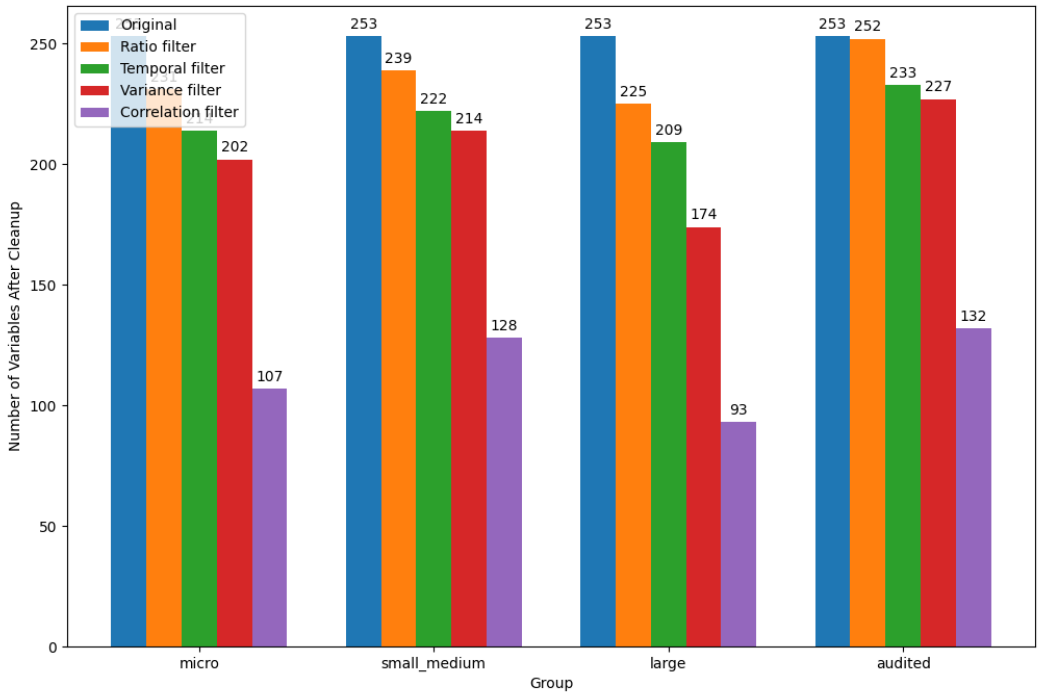


Figure 4 Number of variables remained after each cleanup iteration

Due to the lack of non-timely submissions in the large group companies as illustrated in Figure 5, the large group was left out of the analysis. It does not matter how many variables you take into account – it's always best for the model to predict that submission is timely.

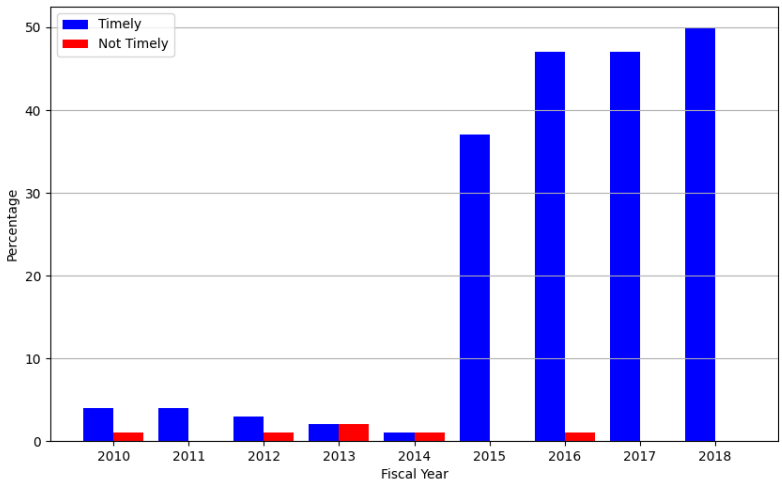


Figure 5 Number of timely and non-timely large group submissions of annual reports by year

2.3 Methodology

This thesis will use a random forest classification algorithm alongside the logistic regression model. Logistic regression has been the primary model in previous studies (Luypaert et al. (2016), Clatworthy and Peel (2016), Lukason and Camacho-Miñano (2019), Laidroo et al. (2020). However, the random forest has been shown to produce higher accuracy percentages in the finance domain for developing early warning for the fiscal stress (Jarmulska, 2020). Similar evidence suggesting that random forest outperforms linear models has been found in capital structure research (Amini, Elmore, Öztekin, & Strauss, 2021).

The classification models will be implemented using Python programming language and the Scikit-learn (Pedregosa et al., 2011) data analysis library. All descriptive statistics will be created with Matplotlib (Hunter, 2007) using the Pandas (McKinney, 2010) data manipulation library. Both methods will use various amounts of independent variables to predict a binary dependant variable – on-time submission. Annual reports submitted before or exactly on the deadline are considered on-time submissions (1), everything else is considered not a timely submission (0). This paper leverages Python 3.12 as the primary programming language, which is a natural choice for data science tasks due to its versatility, extensive libraries, and a large community of developers

and data scientists who continuously contribute to its ecosystem. Python has emerged as the de facto standard for data science in recent years. This leverages many community-maintained libraries. Scikit-learn (1.3.2) was used for machine learning algorithms. Shap (0.45.0) was used to understand the contributions of individual variables to model predictions. Also, it provided valuable insights for model interpretability and variable selection. Pandas (2.1.2) is a fundamental library for data manipulation and analysis in Python. It was used for cleaning, transformations, and exploration. Matplotlib (3.8.0) is a popular data visualization library in Python, it was chosen to create various plots and charts to visualize data, model performance, and results.

Random Forest is a supervised learning technique used for various tasks, including classification and regression (Breiman, 2001). It uses multiple decision trees during the training phase and then makes predictions based on the majority class for classification or the average prediction from these individual trees. A decision tree is a hierarchical structure. Each non-leaf node in the tree corresponds to a test of a variable, and each branch represents the range of values for that variable. The leaf nodes store specific categories or predictions. The decision tree starts at the root node, tests variable attributes related to the category to be classified, and selects branches based on the variable values until it reaches a leaf node. The category stored in the leaf node is considered the final decision. One example decision tree (with limited depth) is shown in Figure 6. This paper uses Gini index for the splitting condition. To calculate the Gini Index (Gastwirth, 1972), the formula is given by

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Where C is the total number of classes and $p(i)$ is the probability of picking the data point with the class i. During the decision tree training, the Gini Index guides the process of finding the optimal way to split the data at each node. The algorithm assesses various variables and thresholds to identify the split that minimizes impurity, as measured by the Gini Index. Once the optimal split is determined, it effectively separates the data into more homogenous subsets with respect to the target variable. This procedure is performed iteratively for each child node, ultimately leading to the creation of a tree structure that can make predictions based on the majority class or value at the leaf nodes. Random forest is expected to outperform logistic regression in this paper because it can capture complex, non-linear relationships in the data by aggregating multiple decision trees, making it more suitable for tasks where complex variable interactions and classification boundaries are involved (Breiman, 2001). At the same time, it's harder to interpret "black box"

models because the way they arrive at a decision is not straightforward. This is because they consist of many individual trees and it is difficult to determine how each tree influences the final outcome (Palczewska et al., 2013). Additionally, unlike linear models, random forests do not provide coefficients that can be easily interpreted.

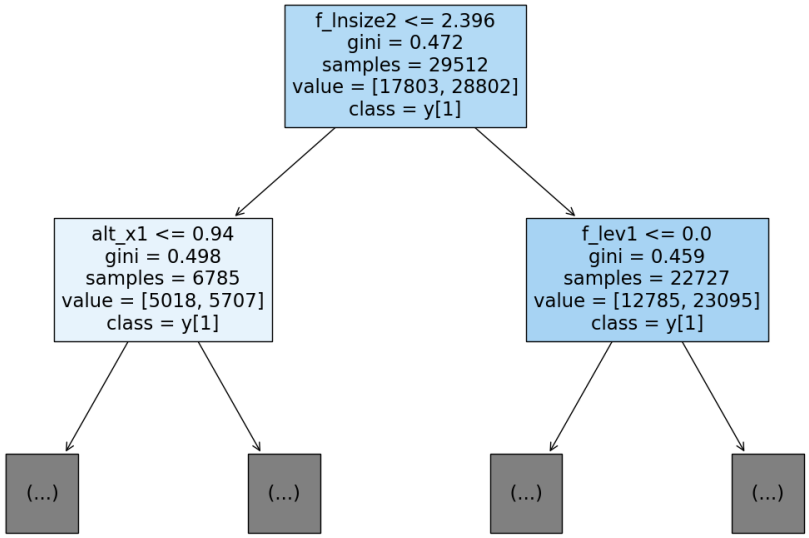


Figure 6 Example decision tree

We use a sliding window of two years to train the models. The sliding window splits the existing dataset into year ranges with the logic shown in Appendix 5. The use of a two-year sliding window for model training is a practical and performance-driven decision. The two-year sliding window decision was used based on how model accuracy behaved over the sliding window. For each window, a new model will be trained. This choice enables to achieve a balance between capturing sufficient historical data while keeping the dataset manageable. It also improves the model's performance by allowing it to incorporate recent trends and patterns as seen in Figure 7 – a window size of two is better than a window size of one 70% of the time. The model's performance will be assessed using data from the following year, mirroring real-world application scenarios.

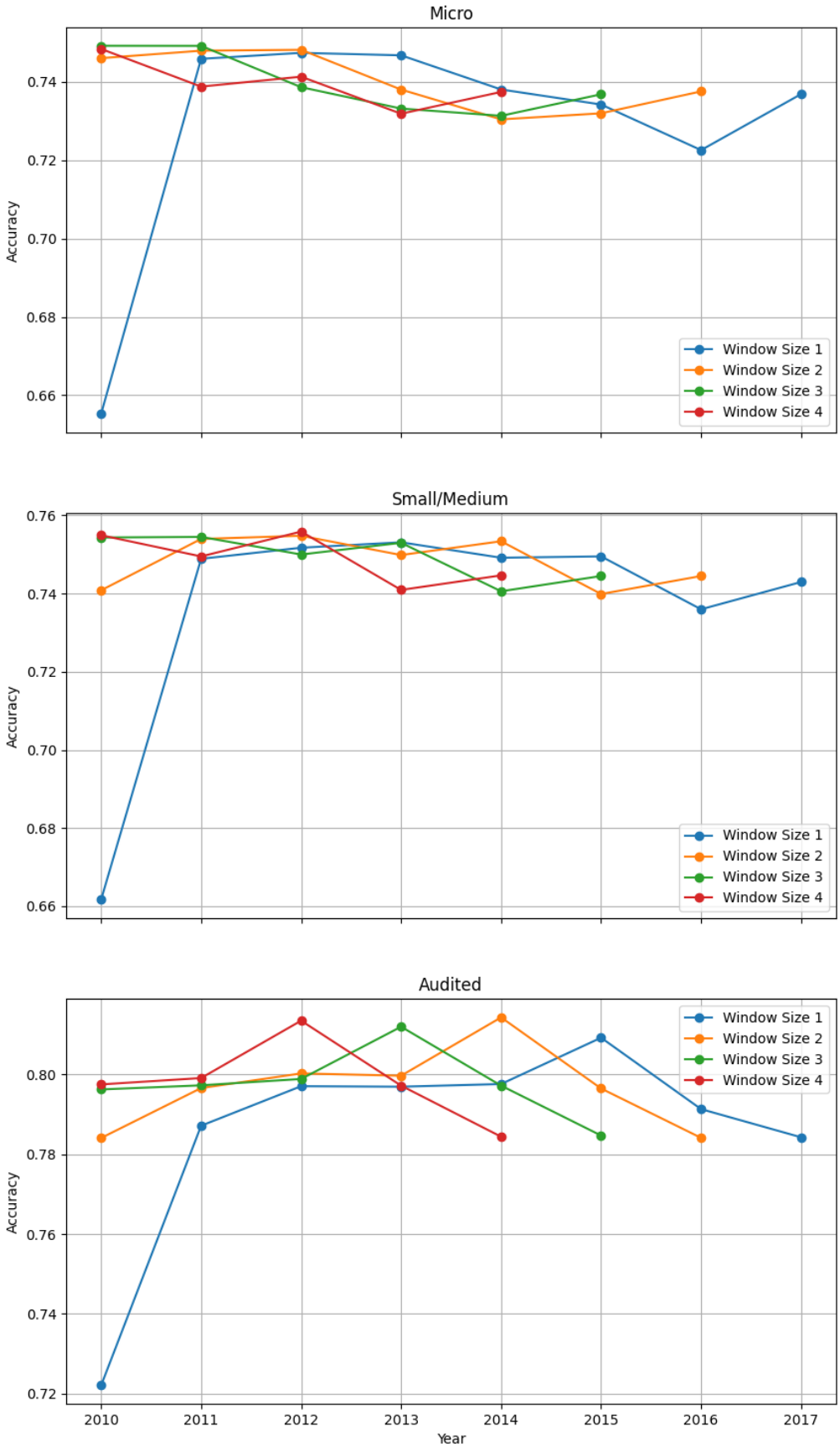


Figure 7 Sliding window size vs the accuracy

The random model was trained with the parameters shown in Table 4. To prevent the model from overfitting and adopting all variables given to the model, the optimal maximum tree depth and number of estimates were determined using cross-validation scores. The hyperparameter optimisation results are shown in Appendix 6 and 7.

Table 4 Random forest model parameters

Parameter name	Parameter value
window_size	2
n_estimators	50
max_depth	15
min_sample_leafs	5
min_samples_split	2

The second model used in this paper is logistic regression, which allows for the investigation of relationships between a binary or categorical dependent variable and one or more explanatory variables. The mathematical general form of a logistic regression model is the following (Hosmer & Lemeshow, 2000)

$$P(y = 1) = \frac{1}{1 + e^{-(a + x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k)}}$$

where

- a is the intercept,
- y is the binary dependent variable – timeliness that is equal to one if the annual report was submitted before or exactly on the deadline, otherwise it is non-timely submission equal to zero.
- $x_1, \ldots, x_k$ are the explanatory variables (differ by company groups)
- $\beta_1, \ldots, \beta_k$ are the estimated parameters of the explanatory variables

Logistical regression was brought in to create a comparison moment to the random forest model. The model has been the primary model in previous studies Luypaert et al. (2016), Clatworthy and Peel (2016), Lukason and Camacho-Miñano (2019), Laidroo et al. (2020).

To measure classification effectiveness, the author uses commonly known accuracy, precision, F1 and recall rate. Accuracy measures the ratio of correctly predicted instances to the total instances in the dataset (Müller & Guido, 2016).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- TP = True Positives (correctly predicted positive instances)
- TN = True Negatives (correctly predicted negative instances)
- FP = False Positives (incorrectly predicted as positive instances)
- FN = False Negatives (incorrectly predicted as negative instances)

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive (Ibid.).

$$Precision = \frac{TP}{TP + FP}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances out of all actual positive instances (Ibid.).

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is a measure of a model's accuracy, particularly when dealing with imbalanced classes. It considers both the precision and recall of the test to compute the score. It is calculated as the harmonic mean of precision and recall (Ibid.).

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

The author uses the backward elimination technique to remove unnecessary variables from the logistic regression model (Zellner, 2004). Backward elimination begins with all k variables in the model and progressively removes one variable at a time. At each step, the variable causing the least increase in the residual sum of squares is dropped. This process repeats until only one variable remains in the model or until a predetermined stopping criterion is met. In addition to backward elimination, the author removed parameters that have duplicated meaning or are correlated. Categorical variables are ignored due to high cardinality in the values.

The author trains a model for each 2-year window. The original variables are found for the 2010 window and carried over to other windows.

Similarly to logistic regression, the author uses backward elimination for random forests. The aim of the elimination is to enhance the model's ability to generalize by systematically discarding features that contribute minimally to predictive accuracy. This iterative process focuses on eliminating the least significant features, minimizing their impact on training errors while refining the model's overall performance.

Random forest model results are interpreted using SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) which is a popular method in the field of machine learning and interpretability. It is used to understand the contribution of individual variables to the prediction made by a machine learning model SHAP provides a way to explain model predictions by assigning each feature an importance score, indicating its impact on the prediction. This can be useful for gaining insights into the model's decision-making process and for making the model more transparent and trustworthy. The core idea of SHAP is based on game theory, specifically the concept of Shapley values (Lundberg & Lee, 2017).

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(X_{S \cup \{i\}}\right) - f_S(x_S) \right]$$

φ_i is the Shapely value for feature i. $\sum$ represents a sum over all possible combinations of features excluding feature i . S is the combination of features. F is the set of all features. $f(s)$ is the model's prediction when considering the combination of features s. $f_{S \cup \{i\}}\left(X_{S \cup \{i\}}\right)$ is the model's prediction when adding feature i to the combination s. Shapley values are used to fairly distribute the "payoff" of a game among its players, taking into account their individual contributions. In the context of machine learning, the game is the prediction task, and the "players" are the features. SHAP computes the Shapley values for each feature, determining how much each feature has contributed to a particular prediction. An example how SHAP scores influence particular observation timeliness prediction has been provided in Figure 8. Blue bar length indicates the magnitude of the negative effect on timely submission. For example, previous late submission (c_d_late1) reduces the likelihood of being timely for a given observation the most. Similar, yet smaller impact

is observed for count of tax arrears (count_tax). This force plot can be used to interpret how variable values impact each observation prediction.

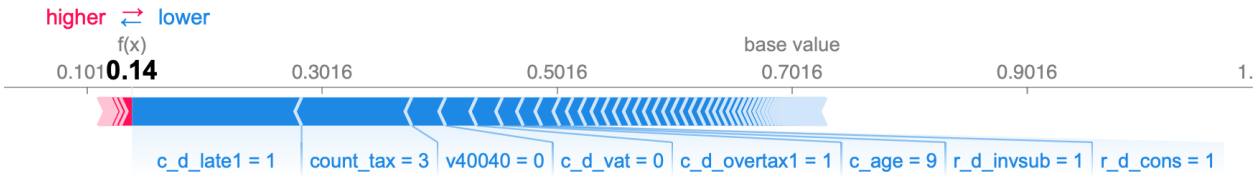


Figure 8 Example of SHAP scores influencing the timeliness

To see general trends, the SHAP violin plot will be used (Figure 9). The violin plot will display sampled population SHAP scores together on one plot. The colour indicates the magnitude of the variable value – red colours indicate corresponding variables having high values, while blue indicates variables having low values. The relative positioning to the vertical zero line (no impact on timeliness) will show how observations are positioned on the SHAP value plane – on the left the variable value impacts timeliness negatively, right side variable value impacts positively.

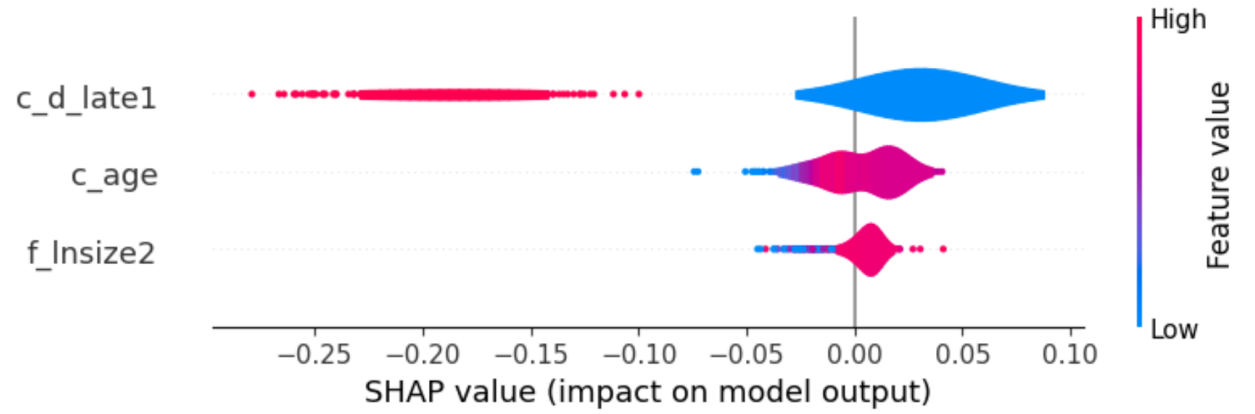


Figure 9 Example of SHAP violin plot

3. RESULTS AND DISCUSSION

3.1 Model results

Both logistic regression and decision tree models achieved similar accuracy (as detailed in Appendix 8 and 9). Examining the average precision, recall, F1-score, and overall accuracy metrics in Table 5, we see positive results for the true (timely) class. High precision (0.79 and 0.80) indicates a low false positive rate, meaning the models effectively identify true positives and avoid classifying negative instances as positive. Recall is also high (0.85 and 0.84), signifying a low false negative rate - the models capture most of the positive instances. The F1-score (0.82 for both) reflects a good balance between precision and recall, indicating strong performance on the true class.

Table 5 Average accuracy (over all years) scores for logit and random forest models

Metric	Logistic regression	Random forest
Precision (True)	0.79	0.80
Recall (True)	0.85	0.84
F1 Score (True)	0.82	0.82
Accuracy (True)	0.75	0.76
Precision (False)	0.67	0.67
Recall (False)	0.56	0.60
F1 Score (False)	0.61	0.63
Accuracy (False)	0.75	0.76

However, for the false (non-timely) class, precision is moderate (0.67 for both), suggesting both models might classify true positives as false positives. Recall for the false class is lower for logistic regression (0.56) compared to the decision tree model (0.60), suggesting the logistic regression model might misclassify more true negatives as false negatives. The F1-score for the false class is also slightly higher for the random forest model (0.61 vs 0.63).

While both models demonstrate good performance in classifying true and false classes, there is room for improvement in correctly identifying false negatives, particularly with the logistic regression model.

Table 6 outlines the results of the logistic regression models. The results show that all three models are statistically significant. The variables remaining in the model vary somewhat across the samples.

Table 6 Logistic regression coefficients (2010 window)

Variable	Micro		Small and Medium		Audited	
const	-0.1244	*** (0.147)	-0.4683	*** (0.330)	0.5462	*** (0.060)
c_age	0.0653	*** (0.003)	0.0592	*** (0.002)	0.0484	*** (0.004)
c_d_late1	-1.5355	*** (0.028)	-1.5404	*** (0.018)	-1.5609	*** (0.048)
maarusearv	-0.9302	*** (0.068)	-		-	
c_d_emplc	0.4868	*** (0.024)	-		-	
f_cash	0.0017	*** (0.000)	-		-	
c_d_vat	0.1131	*** (0.024)	0.0839	*** (0.019)	-	
c_d_overtax1	-0.7231	*** (0.042)	-0.7947	*** (0.022)	-0.9988	*** (0.065)
c_segcm	0.0026	*** (< 0.000)	-0.0010	*** (< 0.000)	-	
r_d_acc_diff	0.1298	*** (0.029)	-		-0.0762	** (0.039)
r_d_abper	-0.4774	*** (0.047)	-0.3358	*** (0.036)	-0.5664	*** (0.131)
end_quarter	-		0.1901	*** (0.025)	-	
f_lev4	-0.2931	*** (0.069)	-		-	
f_lev3	-		-0.0970	*** (0.029)	-	
f_lnsize2	0.0263	*** (0.005)	0.0441	*** (0.003)	0.0310	*** (0.005)
v30010	-		5.838e-07	*** (1.58e-07)	8.58e-08	*** (2.66e-08)
v40020	-		-5.57e-07	*** (6.7e-08)	-	
v50030	-		4.03	*** (1.04e-07)	-	
c_no_dir	-		-0.0702	*** (0.010)	-	
r_d_ifrs	-		-		0.6307	*** (0.146)
eitav_audit	-		-		-1.9993	*** (0.297)
loobumine_audit	-		-		-1.3023	*** (0.297)
c_d_mnc2	-		-		0.4807	*** (0.045)
f_d_loss	-		-		-1.457	*** (0.043)
No. of observations	47608		107790		17040	
Pseduo R-squ. (Cox & Snell)	0.111		0.108		0.116	
F-Statistic	627	***	1362	***	295	***

Source: author's calculations
Notes: *** 0.01 statistical significance, ** 0.05 statistical significance (Standard error in brackets). Marginal effects can be found in Appendix 14.

The feature importances of the decision tree are illustrated in Figure 10, showcasing their relative significance within the model. A higher feature importance denotes a greater impact on the model's predictions compared to features with lower importance, suggesting that features with higher scores yield more influence in predicting the targeted variable.

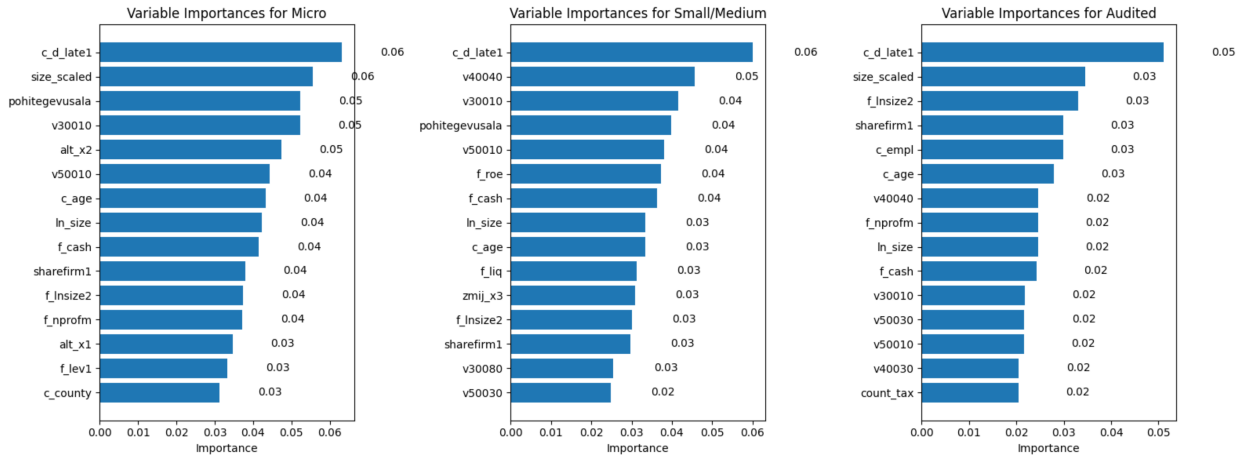


Figure 10 Top 15 variable importance for decision tree classifier (2010 window)

In the following chapter, we will focus more on the comprehensive explanation of the results of both the random forest and logistic regression models. The model results presented are based on the 2010 window, chosen for variable selection, and validated across the period of 2010-2016 to ascertain the persistence of significance (see Appendices 18, 19, 20). Any variables lacking significance across the years will be further commented on in the upcoming chapter.

3.2 Variables influencing the timeliness of the annual reports

3.2.1 Micro enterprises

Random forest variable impact on the timeliness of micro companies is illustrated in Figure 11, conclusions drawn from both models are presented in Table 7. In the case of micro companies, both models confirm previously established findings. Late submission for the previous year (c_d_late1) has consistently been identified as the most significant factor affecting timeliness negatively as also in Laidroo et al. (2020). It alone increases the likelihood of being non-timely by approximately 30 percentage points. Additionally, older companies (c_age) demonstrate a tendency to adhere to reporting schedules, confirming previous research (Laidroo et al., 2020; Breuer et al., 2020; Eierle, 2008).

Table 7 Variable impact on micro company timely submission

Variable	Random forest	Logistic regression	Related findings
c_d_late1	-	-	- (Laidroo et al., 2020)
c_age	+	+	+ (Laidroo et al., 2020; Breuer et al., 2020; Eierle, 2008)
maarusearv	-	-	
c_d_emplc	+	+	
f_cash	+ (Weak)	+ (Weak, seems like outlier impacted the regression)	+ (Laidroo et al., 2020; Breuer et al. 2020; Clatworthy and Peel, 2016; Lukason 2019) support a negative association between liquidity and reporting delays.
c_d_vat	+	+	+ (Laidroo et al., 2020)
c_d_overtax1	-	-	- (Laidroo et al., 2020)
c_segcm	+	+ (Weak)	
r_d_acc_diff	- (Weak)	+ (Weak)	
r_repper	-	+	- (Laidroo et al., 2020) Longer period length leads to filling delays
f_lev4	-	-	- (Laidroo et al., 2020; Bigus and Hillebrand, 2017; Clatworthy and Peel, 2016; Luypaert et al. 2016; Lukason and Camacho-Miñano, 2019)
f_lnsize2	+	+	+ (Laidroo et al., 2020)

Source: Author's calculations
Notes: Appendix 11 explains the partial dependency between the variable and timely submission

Companies with higher liquidity (f_cash), indicated by a larger percentage of cash, exhibit slightly better punctuality in reporting, similar to previous studies (Laidroo et al., 2020; Breuer et al., 2020; Clatworthy and Peel, 2016; Lukason, 2019). This observation remains statistically significant across five out of seven observed windows. However, its impact on timeliness is relatively modest, affecting less than one percentage point towards timeliness for companies with high liquidity. The obligation to pay VAT (c_d_vat) leads to a higher rate of on-time submissions, as in Laidroo et al. (2020), impacting the probability of timely reporting by around two percentage points across all observed windows. On the other hand, tax arrears (c_d_overtax1) have a negative effect on timely submissions, as in Laidroo et al. (2020), reducing timeliness by 15 percentage points. High leverage (f_lev4) is associated with delayed submissions, a trend observed in previous studies (Laidroo et al., 2020; Bigus and Hillebrand, 2017; Clatworthy and Peel, 2016; Luypaert et al., 2016; Lukason and Camacho-Miñano, 2019), was statistically significant only in four out of seven windows indicating no consistent pattern. Extreme cases of high leverage (100% of debt relative to liabilities) are associated with a 6 percentage point decrease in timeliness. Additionally, larger micro-companies by sales revenue (f_lnsize2) tend to submit reports on time as in Laidroo et al. (2020), the pattern did not only hold statistical significance for the year 2011 window.

In addition to the variables analyzed through logistic regression, the random forest algorithm has uncovered (Figure 11) significant relationships among multiple variables, which could be of interest (see Appendix 15) – the primary business area of a company (industry), count of quarters with tax arrears (count_tax) and relative profitability decrease compared to previous year (f_d_prof). The industry of a company (industry) also provides valuable insights into timeliness as revealed in Appendix 10, which highlights the uneven distribution of timeliness across different industry groups. Notably, the lowest scores are associated with managing sports facilities, loan operations, financial services, and bakeries. Further research is required to comprehensively understand these relationships.

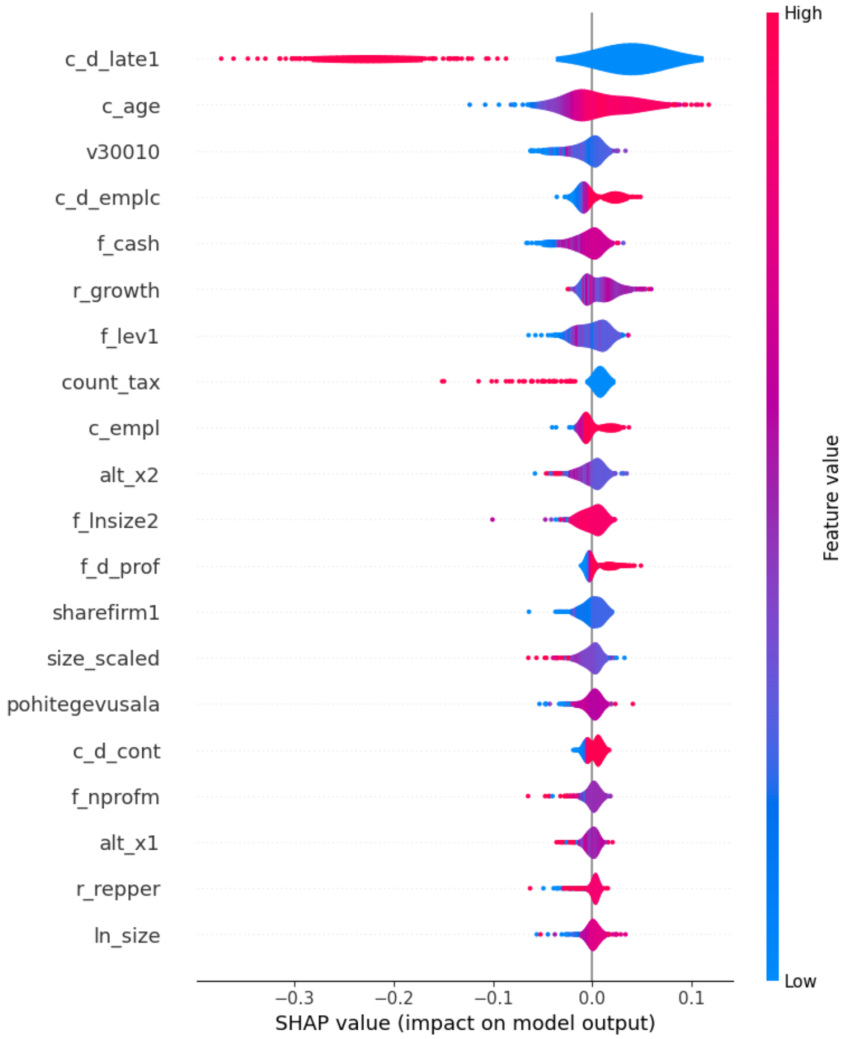


Figure 11 Micro group shap violin plot for random forest
Source: Author's calculations

Contrary to previous findings (Laidroo et al., 2020) and the random forest model, reporting period length was reported to have a positive relationship with timeliness in the logistic regression. However, according to the random forest model, deviations from the mean value (365 days) could lead to non-timely submissions. The number of rulings issued for a given company (maarusearv) tends to have a negative effect on timeliness, likely because many companies with rulings were also late in the previous year, particularly due to government-issued rulings about missing annual reports. Companies that report employment costs (c_d_emplc) tend to exhibit more timely behaviour. Random forest results also associate relative profitability loss (f_d_prof) with more timely reporting. It is plausible that the actual relationship does not exist, and the inclusion of this variable in the model may be coincidental. A similar issue arises with accounting difficulty – logistic regression associates accounting difficulty (r_d_acc_diff) with a 3 percentage point increase in favour of timeliness, however random forest finds that the relationship is actually negative.

3.2.2 Small and medium enterprises

Random forest variable impact on timeliness for small and medium entities is illustrated in Figure 12. Conclusions drawn from the results of both models are presented in Table 8. For small to medium-sized companies, both models align with previously identified findings. It has been established that late submissions for the previous year (c_d_late1) have a detrimental effect on timeliness as in Laidroo et al. (2020). This effect is analogous for micro companies, leading to a decrease in the likelihood of timely submissions by 30 percentage points. Moreover, older companies (c_age) tend to adhere to reporting schedules as in previous studies (Laidroo et al., 2020; Breuer et al., 2020; Eierle, 2008). With every extra year of activity, the company becomes one percentage point more likely to be on time with submissions. The VAT obligation (c_d_vat) correlates with increased punctuality in submissions as in Laidroo et al. (2020). The effect is similar to micro companies having approximately two percentage point advantage in terms of timeliness. Conversely, tax arrears (c_d_overtax1) have a negative impact on timely submissions similar to Laidroo et al. (2020). It has a 16 percentage point negative effect on timeliness, two percentage points stronger relative to micro companies. Larger companies by sales (f_lnsize2) tend to submit reports on time similarly to micro companies and there is a positive correlation between reporting period length (r_repper) and timeliness as in Laidroo et al. (2020). High leverage (f_lev3) is associated with delayed submissions, in line with previous studies (Laidroo et al., 2020; Bigus and Hillebrand, 2017; Clatworthy and Peel, 2016; Luypaert et al., 2016; Lukason and Camacho-

Miñano, 2019), only for one out of seven windows, therefore lacking evidence to prove the significance over the multi-period interval.

Table 8 Variable impact on small and medium company timely submission

Variable	Random forest	Logistic regression	Related findings
c_d_late1	-	-	- (Laidroo et al., 2020)
c_age	+	+	+ (Laidroo et al., 2020; Breuer et al., 2020; Eierle, 2008)
c_no_dir	-	-	+ (Laidroo et al., 2020)
r_repper	-	- (weak)	- (Laidroo et al., 2020) Longer period length leads to filling delays
c_segcm	+	- (weak)	
v30010	-	+	+ (Laidroo et al., 2020; Breuer et al., 2020; Clatworthy and Peel, 2016; Lukason and Camacho-Miñano, 2019) support a negative association between liquidity and reporting delays.
v40020	+	-	
v50030	-	+	
f_lnsize2	+	+	+ (Laidroo et al., 2020)
f_lev3	-	-	- (Laidroo et al., 2020; Bigus and Hillebrand, 2017; Clatworthy and Peel, 2016; Luypaert et al. 2016; Lukason and Camacho-Miñano, 2019)
c_d_overtax1	-	-	- (Laidroo et al., 2020)
c_d_vat	+	+	+ (Laidroo et al., 2020)

Source: Author's calculations
Notes: Appendix 12 explains the partial dependency between the variable and timely submission

In addition to logistic regression variables, random forest found important relationships (Figure 12) between multiple variables that could be interesting (see Appendix 16) - Similar to micro companies – count of quarters with tax arrears (count_tax) and the primary area of the business (industry) seem to be important variables. Unfortunately primary business area is carrying some information that we can't quantify with linear models. In addition, the market entry barrier (c_barr2) surfaced as a potentially interesting variable. The next paragraph will comment on those findings more.

There were several findings that presented questionable insights. For instance, cash in balance (v30010) emerged as a significant variable in both models; however, upon examination of partial dependence plots, it appears that the effect may not be significant for the random forest model.

This discrepancy could potentially stem from the sampling of observations that occurred prior to generating SHAP scores.

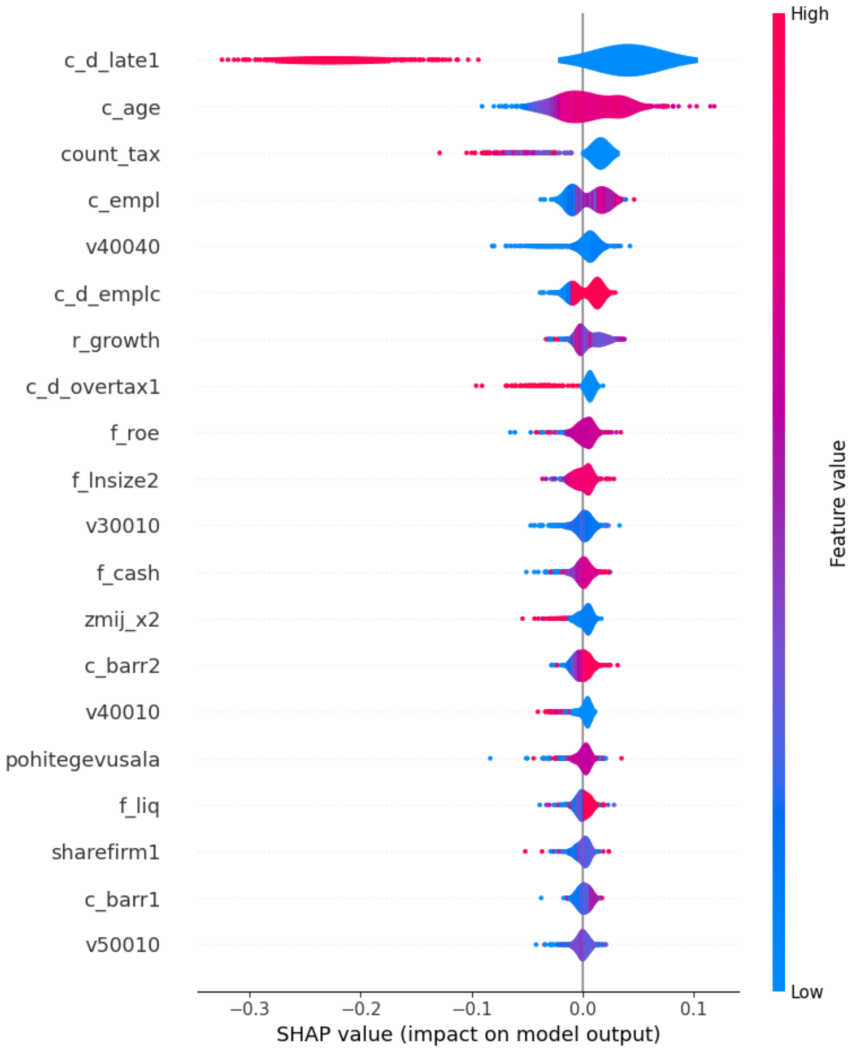


Figure 12 Small and medium group shap violin plot for random forest
Source: Author's calculations

Nevertheless, there seems to be a general trend indicating that high liquidity assets, such as cash, have a positive influence on the timeliness of annual reports because cash in balance seems to be a significant variable across all observed time windows for logistic regression. Similarly, both long-term debt (v40020) and cash flows from investment activities (v50030) yield conflicting results across the models and thus should be approached with caution in interpretation. The company's primary business area (industry) also provides valuable insights into timeliness, mirroring the need for further research to fully uncover its relationship, akin to the findings regarding micro companies. The random forest model reveals a negative relationship between timeliness and tax delays, specifically through the quarters of tax delays (count_tax). Furthermore,

a higher entry barrier to the market, as indicated by a larger percentage of assets tied up in machinery and equipment (c_barr2), correlates with a greater likelihood of timely submissions. There is a disagreement between models regarding the impact of companies with higher revenue from their primary business segment on timeliness. It is plausible that the actual relationship does not exist, and the inclusion of this variable in the model may have occurred by chance.

3.2.3 Audited enterprises

The impact of variables on the timeliness of audited companies is depicted in Figure 13, with conclusions drawn from logistic regressions presented in Table 9. For audited enterprises, findings remain also aligned closely with previous research.

Table 9 Variable impact on audited enterprises timely submission

Variable	Random forest	Logistic regression	Related findings
c_d_late1	-	-	- (Laidroo et al., 2020)
c_age	+	+	+ (Laidroo et al., 2020; Breuer et al., 2020; Eierle 2008)
eitav_audit	-	-	
loobumine_audit	-	-	
märkusega_rõhutamine_audit	-	-	
v30010	-	+	+ (Laidroo et al., 2020; Breuer et al., 2020; Clatworthy and Peel, 2016; Lukason and Camacho-Miñano, 2019) support a negative association between liquidity and reporting delays.
r_d_abper	-	-	- (Laidroo et al., 2020) Longer period length leads to filling delays
c_d_mnc2	+	+	
f_lnsize2	+	+	+ (Laidroo et al., 2020)
f_d_loss	-	-	
r_d_acc_diff	+	-	
c_d_overtax1	-	-	- (Laidroo et al., 2020)

Source: Author's calculations
Notes: Appendix 13 explains the partial dependency between the variable and timely submission

Notably, the previous year's lateness with annual report submission(c_d_late1) was consistently associated with decreased timeliness (Laidroo et al., 2020). The effect is similar to other company groups – reducing approximately 28 percentage points likelihood of being timely for companies that were late last year. Additionally, older companies (c_age) tend to adhere to reporting schedules, as supported by various studies (Laidroo et al., 2020; Breuer et al., 2020; Eierle, 2008), increasing one percentage point of timeliness likelihood per year. Furthermore, tax arrears

(c_d_overtax1) seem to have a detrimental effect, as in Laidroo et al. (2020), accounting for approximately 18 percentage point decrease in the likelihood of timeliness. Larger companies by sales revenue (f_lnsize2) also demonstrate a tendency to submit reports on time, and there appears to be a positive relationship between reporting period length (r_d_abper) and timeliness as in Laidroo et al. (2020). The relationship is even stronger than small and medium or micro enterprises – increasing by three percentage points per one log unit of size (measured in revenue). Auditor-related variables turned out to be significant variables across all the windows. For example, negative auditor decision (eitav_audit) emerged as a significant variable in both models, indicating a negative relationship with timeliness – decreasing timeliness by 35 percentage points. Similarly, enterprises that declined audit (loobumine_audit, 22 percentage points) or received audit results with remarks (märkusega_rõhutamine_audit, 11 percentage points) exhibited a similar negative relationship. International enterprises (c_d_mnc2) were found to be more likely to submit reports on time (8 percentage point positive impact for timeliness). That has also been noticed by previous research (Laidroo et al., 2020).

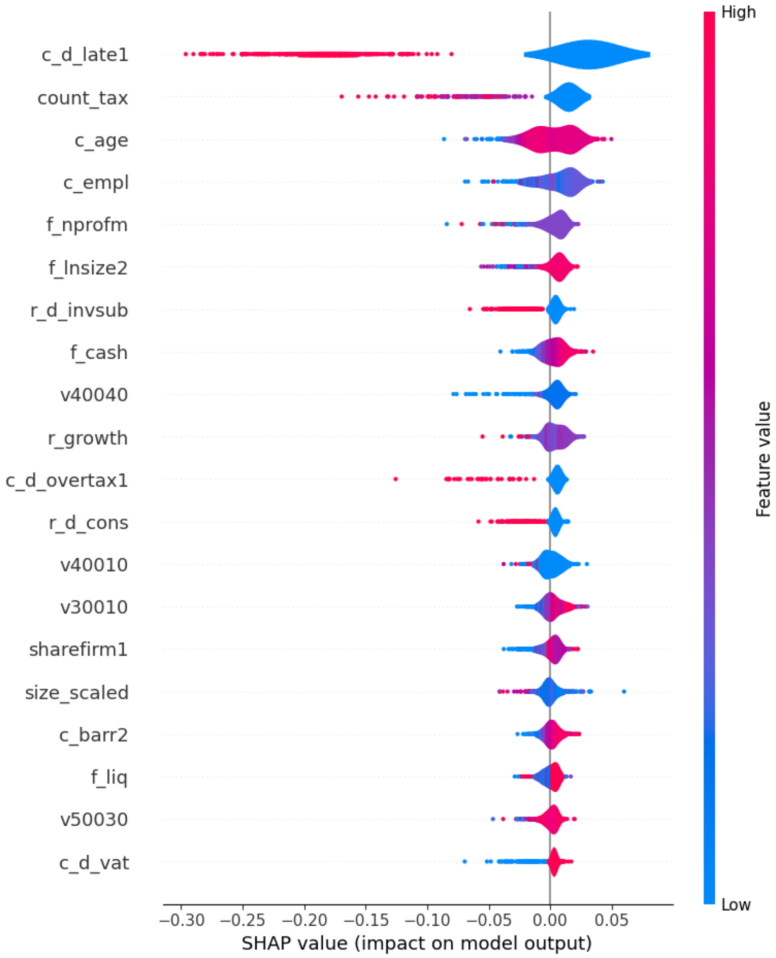


Figure 13 Audited group shap violin plot for random forest
Source: Author's calculations

In addition to the logistic regression variables, the random forest model uncovered (Figure 13) significant relationships between multiple variables, as detailed in Appendix 17. Similarly to small and medium companies – the count of quarters with tax arrears (count_tax) and market entry barrier (c_barr2) seem to be important variables. Similarly to other groups, the obligation to pay VAT (c_d_vat) leads to a higher rate of on-time submissions. Report consolidation (r_d_cons) and having subsidiaries (r_d_invsub) were negatively correlated with timeliness. Conversely, a higher entry barrier to the market (c_barr2) corresponded to a higher likelihood of timely submissions. Finally, enterprises with a higher number of employees (c_empl) were negatively associated with timeliness.

There were also conflicting indications - for instance, cash in balance (v30010) emerged as a significant variable in both models; however, when checking the variable significance across the multiple windows, it did not prove to be persistent – meaning it appeared in the model by chance. Conflicting results were also observed regarding losses reported during the previous year (f_d_loss) and accounting difficulty (r_d_acc_diff) between the random forest and logistic regression models. Partial dependence plots suggested weak or non-existent relationships, suggesting that these associations may not truly exist or are really weak.

3.3 Differences in variables affecting timeliness across company groups

This chapter explores how attributes impacting the timeliness of financial reporting vary across different company classifications: Micro, Small and Medium, and Audited Enterprises.
Table 10 below summarizes these relationships. Clearly conflicting / less significant variables across the company groups are ignored. A "+" symbol indicates a positive correlation with timeliness, meaning companies with that attribute tend to submit reports on time. A "-" symbol signifies a negative correlation, and "Non-binary" suggests the relationship is complex and requires more complicated analysis. A “+/-“ is significant for some groups, but conflicting relationship between models. Finally, a blank space indicates no significant link between the attribute and timeliness for that specific company group.

Table 10 Attributes affecting timeliness across company groups

Variable	Micro	Small / Medium	Audited
Previously late companies (c_d_late1)	-	-	-
Enterprise age (c_age)	+	+	+
Liquidity (f_cash, v30010)	+	+/-	+/-
VAT obligation (c_d_vat)	+	+	
Tax arrears (c_d_overtax1)	-	-	-
High leverage (f_lev3)	-	-	
Company size (f_lnsize2)	+	+	+
Main business area (industry)	Non-binary	Non-binary	
Reporting period length (r_repper, r_d_abper)	Non-binary	+	+
Number of rulings (maarusarv)	-		
Employment costs exists (c_d_emplc)	+		
Market entry barrier (c_barr2)		+	+
Audit complications (eitav_audit, loobumine_audit, märkusega_rõhutamine_audit)			-
Consolidation (r_d_cons)			-
Number of employees (c_empl)			-
International activity (c_d_mnc2)			+

Source: Author's calculations

Across the company classification, larger and older companies tend to prioritize timely reporting. This relationship has been reported by previous research by Laidroo et al. (2020), Breuer et al., (2020) and Eierle (2008). In terms of financial health, liquidity, absence of tax arrears, and lower leverage all contribute to a company's ability to dedicate resources towards timely reporting (positive correlation for Small and Medium enterprises). Similar results from Laidroo et al. (2020), Breuer et al. (2020), Clatworthy and Peel (2016), and Lukason (2019) support a negative association between liquidity and reporting delays. Regarding reporting requirements, VAT obligation and longer reporting periods are typically linked with a higher chance of timely annual report submissions all across the company groups. That has been already reported by Laidroo et al. (2020). This paper found one small nuance for micro-enterprises – low and high values for fiscal year length are associated with a negative association with timeliness, previous research has highlighted only a positive relationship with delays in general. Market entry barriers and specific business areas might influence timeliness. The business area has a non-binary relationship, meaning it will be hard to model with logistic regression due to the high cardinality of values. Through random forest, one could tell there is a significant relationship, but actual details need to be investigated more in detail. Market entry barrier seems to positively impact timeliness for small and medium and audited enterprises. Company structural complexity - the absence of audit complications, lack of consolidation requirements, and a low number of employees affect

timeliness positively for audited enterprises. In addition to that, international activity has a positive impact on timeliness.

The factors affecting timely financial reporting differ significantly across company groups. Size, age, financial health, and reporting requirements play a more prominent role for all entities. However, the number of rulings, market entry barriers, employment costs, audit related aspects might not have similar relationships across all the groups. It's important to note that these are general trends, and specific company circumstances can influence their reporting behaviour.

3.4 Prediction of timely submissions

This study employs two competing models: logistic regression and random forest. As illustrated in Table 5, the average F1 score and accuracy across models exhibit minimal deviation, with the random forest model demonstrating a slightly higher accuracy score on average by one percentage point. However, upon closer examination of the results within groups, as presented in Table 11, disparities emerge in how the models behave for "true" (timely) and false (non-timely) values. It is important to emphasize that, from an analytical standpoint, predictions of "false" values hold greater significance. This is because such findings can inform policymakers, practitioners, and researchers seeking to improve disclosure practices and mitigate delays in financial reporting. Conversely, predictions of "true" values hold less actionable value since these enterprises already comply with legal requirements. The disparity in F1 scores is depicted in Figure 14, highlighting that, on average, logistic regression and random forest models exhibit similar behaviour. However, in the small and medium group, "false" predictions with the random forest model are approximately 2% more precise on average across the years, while in the micro group, they are almost 6% more precise. Notably, there is no substantial difference between the models in predicting the timeliness of audited companies. These results suggest that, for non-timeliness predictions, the random forest model outperforms logistic regression.

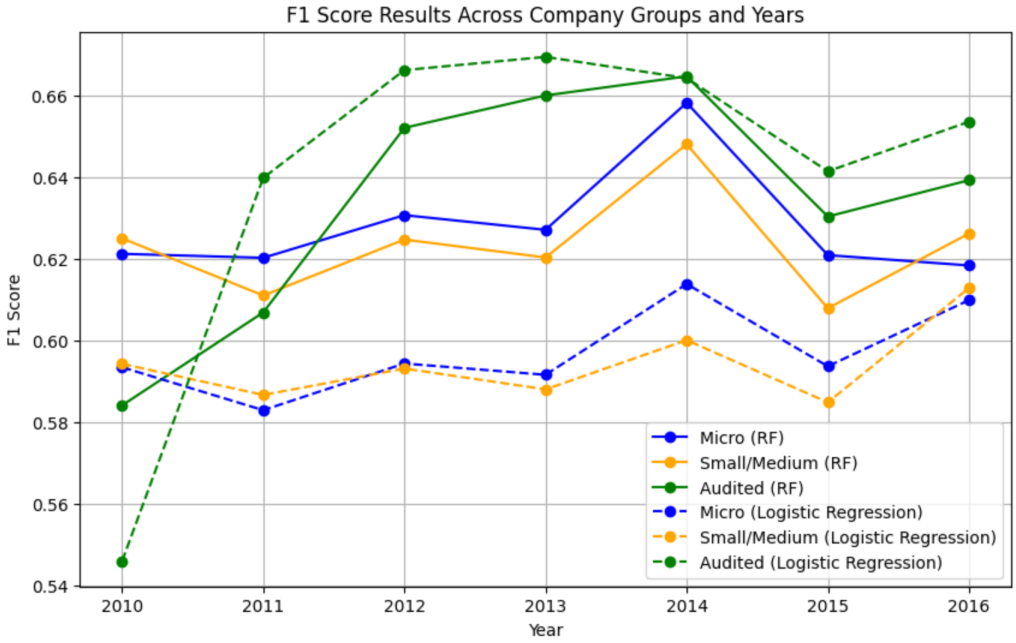


Figure 14 False F1 score across models, groups
Source: Author's calculations

Table 11 Accuracy performance between logit vs random forrest (% of overperformance of logit model over the random forest)

Year	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy
Micro - Predicting for Timely: "True"					Micro - Predicting for Timely: "False"			
'10	-2.53%	4.84%	1.10%	0.24%	4.31%	-13.12%	-4.68%	0.24%
'11	-3.01%	4.08%	0.47%	-0.60%	2.69%	-14.88%	-6.40%	-0.60%
'12	-2.84%	3.07%	-0.02%	-1.13%	1.86%	-12.89%	-6.11%	-1.13%
'13	-3.02%	3.54%	0.20%	-0.97%	1.55%	-13.08%	-6.00%	-0.97%
'14	-4.73%	7.36%	1.33%	-0.41%	5.17%	-19.76%	-7.23%	-0.41%
'15	-2.29%	3.35%	0.43%	-0.53%	2.22%	-10.54%	-4.58%	-0.53%
'16	-1.02%	2.65%	0.81%	0.49%	2.29%	-5.06%	-1.39%	0.49%
Small/Medium - Predicting for Timely: "True"					Small/Medium - Predicting for Timely: "False"			
'10	-2.00%	3.16%	0.55%	-0.09%	2.44%	-9.56%	-3.70%	-0.09%
'11	-0.51%	0.79%	0.12%	-0.05%	0.62%	-2.48%	-1.02%	-0.05%
'12	-1.03%	1.15%	0.02%	-0.35%	0.65%	-4.51%	-2.10%	-0.35%
'13	-0.94%	0.58%	-0.19%	-0.55%	-0.27%	-3.66%	-2.03%	-0.55%
'14	-2.42%	3.37%	0.46%	-0.32%	2.32%	-10.01%	-3.90%	-0.32%
'15	-0.11%	0.08%	-0.02%	-0.06%	0.00%	-0.43%	-0.24%	-0.06%
'16	-0.59%	1.15%	0.26%	0.07%	0.93%	-2.68%	-0.93%	0.07%
Audited - Predicting for Timely: "True"					Audited - Predicting for Timely: "False"			
'10	-1.97%	1.83%	-0.17%	-0.83%	1.66%	-13.84%	-7.19%	-0.83%
'11	2.11%	-2.38%	-0.04%	0.54%	-2.46%	10.68%	4.93%	0.54%
'12	1.09%	-0.99%	0.08%	0.39%	-0.75%	4.58%	2.10%	0.39%
'13	1.10%	-1.86%	-0.36%	-0.19%	-2.27%	4.88%	1.41%	-0.19%
'14	0.48%	-1.62%	-0.56%	-0.64%	-2.83%	2.69%	-0.04%	-0.64%
'15	0.68%	-2.33%	-0.75%	-0.79%	-4.72%	4.66%	0.59%	-0.79%
'16	0.76%	-2.34%	-0.71%	-0.64%	-3.6%	4.47%	0.78%	-0.64%

Source: Author's calculations

The difference in precision for small and medium and micro companies is likely due to non-linear relationship – company's primary business area (industry) that logistic regression did not incorporate in this research. Observation is based on the fact that for the audited companies group the overperformance of random forest is negligible. That suggests that employment of more complex model that can incorporate non-linear relationships for timeliness research might not yield to expected benefits.

CONCLUSION

The objective of this thesis was to identify the determinants of annual report submission timeliness through random decision forests ensemble learning method alongside the logistic regression model. The dependent variable was timeliness that was equal to one if the annual report was submitted before or exactly on the deadline, otherwise it is non-timely submission equal to zero. The initial dataset, sourced from an impact assessment study conducted for the Ministry of Finance in Estonia, consisted of 1 289 352 data rows with 285 variables. A random forest classification and logistic regression algorithms were implemented in Python, utilizing the Scikit-learn data analysis library. The data preparation involved a multi-step process, including the application of various filters such as high missing value ratio, temporal variable replacement, zero variance, and correlation filters. Recursive variable elimination was employed to prune less relevant variables based on the F-scores for logistic regression and feature importance for random forest. The final models were trained using a two-year sliding window approach. The analysis was done on three samples: micro, small and medium and audited enterprises. The large enterprises group was excluded from the comparison due to a lack of data on late filers in some years.

The results show that larger and older companies prioritize timely reporting. Financial health indicators such as liquidity, absence of tax arrears, and lower leverage positively correlate with timely reporting, particularly for small and medium enterprises. Additionally, VAT obligation and longer reporting periods are associated with higher chances of timely annual report submissions across all company groups. However, micro enterprises exhibit a nuanced relationship with fiscal year length, with both low and high values showing negative associations with timeliness. Factors such as market entry barriers and specific business areas also influence reporting timeliness, with complexities in modeling due to non-binary relationships and high cardinality values. Market entry barriers positively impact timeliness for small and medium and audited enterprises, while company structural simplicity, such as the absence of audit complications and low employee numbers, positively affects audited enterprises' timeliness. Furthermore, international activity is noted to have a positive impact on timeliness.

Across company groups, size, age, financial health, and reporting requirements consistently play significant roles in timely reporting. However, other factors such as number of rulings related to late-filing, market entry barriers, employment costs, and audit-related aspects may vary in their impact across different company groups.

For practical application, the timeliness of annual report submissions holds particular significance, with an emphasis on identifying instances of non-timely submissions. Logistic regression compared to random forest exhibits a suboptimal performance in this specific regard. On average logistic regression and random forest behave similarly. However, for small and medium group "False" predictions with random forest are 2 percentage points more precise across the years on average and for micro group almost 6 percentage point more precise. In general random forest can predict on 60% of non-timely submissions correctly. The implications of potential false negatives in this context could trigger unnecessary warnings or repercussions for enterprises. Further refinement and optimization to enhance the model's precision could improve non-timely prediction even more. XGBoost is good candidate model for futher improvements as other studies have found XGBoost overperforming random forests (Fauzan, 2018). The study suggests that the simpler logistic regression model may be sufficient for timeliness prediction due to the minimal improvement offered by the more complex random forest model.

KOKKUVÕTE

MAJANDUSAASTA ARUANNETE TÄHTAEGSE ESITAMISE MÕJURID EESTIS

Artur Luik

Aastaaruannete esitamise õigeaegsus on oluline teema Euroopa eraettevõtete, eriti väiksemate ettevõtete kontekstis, kuna on täheldatud aruannete esitamise viivitusi või esitamata jätmisi (Clatworthy, et al., 2016; Strouhal, et al., 2014). Sarnaselt nende ELi kolleegidega peavad kõik juriidilised isikud Eestis kehtiva raamatupidamisseaduse kohaselt koostama ja esitama oma aastaaruande registrile kuus kuud pärast majandusaasta lõppu. Aastaaruannete esitamise ebaõnnestumisel või viivitusel on otsene mõju äritegevuse statistikale, mis on oluline sisend valitsusele (Bolívar & Galera, 2012). Detsembris 2019 alanud koroonaviiruse pandeemia on suurepärane näide, kus aruannete õigeaegsus oli ettevõtete endi jaoks oluline. Lühikese aja jooksul võeti vastu otsused kõige rohkem mõjutatud sektorite toetamiseks. Turismisektori toetuspaketis - "Toetus koroonaviiruse põhjustatud haiguse COVID-19 puhangu tõttu turismisektori ettevõtjatele tekitatud kahjude osaliseks hüvitamiseks" (Government of the Republic of Estonia, Regulation No 12, 2020) üheks nõudeks oli eelmise aasta aastaaruande esitamine. Sellistes olukordades sõltub toetuse tõhusus aastaaruande esitamise määrast.

Perioodil 2010-2018 ei esitanud koguni 47-55% Eesti juriidilistest isikutest oma majandusaasta aruandeid tähtaegselt. Kuigi selle põhjuseid on eelnevalt käsitletud rahandusministeeriumile koostatud 2020. aasta raportis (Laidroo, et al., 2020), ei ole tehtud põhjalikumat kvantitatiivset analüüsi. Käesoleva lõputöö eesmärk on määratleda majandusaasta aruannete tähtaegset esitamist põhjustavad tegurid otsutusmetsa (ingl *random forest*) ja logistilise regressiooni meetoditega. Lõputöö raames otsistakse vastust järgmistele uurimisküsimustele:

1. Millised muutujad selgitavad aastaaruande õigeaegsust kolmes ettevõtete rühmas (sh mikroettevõtted, väikesed ja keskmise suurusega ettevõtted, auditeeritud ettevõtted)?
2. Kuidas erineb aastaaruande õigeaegsuse ennustamise protsess rühmade vahel?
3. Milline mudel suudab paremini ennustada aastaaruande esitamise õigeaegsust?

Töö on jagatud kolmeks peatükiks. Esimeses peatükis on esitatud teoreetiline raamistik, mis käsitleb kohustuslikku avalikustamist ja avalikustamise ajastust, millele järgneb empiiriline tõendus ettevõtte atribuutide ja finantsaruandluse õigeaegsuse seose kohta. Seejärel antakse ülevaade Eesti õigusraamistikust ja olemasolevatest empiirilistest uuringutest.

Teine peatükk keskendub andmetele ja metoodikale. Andmed pärinevad varasemast uuringust "Majandusaasta aruannete mitte esitamise mõjuanalüüs Rahandusministeeriumile" (Laidroo et al., 2020). Andmestik koosneb 1 289 352 andmereast ning 285 muutujast. Andmete ettevalmistamiseks kasutati mitmeastmelist protsessi. Eemaldati muutujad, kus oli puuduolevate väärtuste osakaal kõrge või variatsioon oli null. Ajalised muutujad asendati fiktiivsete muutujatega või eemaldati. Lõpuks eemaldati muutujad mis olid omavahel tugevasti korreleerunud. Autor kasutab Pythoni programmeerimiskeelt ostustusmetsa ja logistilise regressiooni rakendamiseks, kasutades Scikit-learn andmeanalüüsi teeki. Sõltuv muutuja on majandusaasta aruande õigeaegsus, mis on võrdne ühega, kui aastane aruanne esitati enne tähtaega või täpselt tähtajaks, vastasel juhul loetakse väärtus nulliks.
Mudelite treenimisel kasutati logistilise regressiooni jaoks rekursiivset muutujate elimineerimist F-skoori baasil ning muutujate olulisuse hinnangut otsustusmetsa jaoks. Lõplikud mudelid treenitakse kaheaastase libiseva akna lähenemisviisiga. Analüüs teostatakse grupiti eraldi: mikro-, väike- ja keskmise suurusega ning auditeeritud ettevõtted. Suurte ettevõtete grupp jäeti võrdlusest välja andmete puudumise tõttu.

Kolmandas peatükis esitatakse tulemused koos tõlgendustega. Tulemused näitavad, et suuremad ja vanemad ettevõtted esitavad majandusaasta aruanded tähtagsemalt. Finantsseisundi näitajad nagu likviidsus, maksuvõlgade puudumine ja madalam võlakoormus korreleeruvad positiivselt tähtaegse esitamisega, eriti väikeste ja keskmiste ettevõtete puhul. Lisaks on käibemaksukohustus ja pikemad majandusaastad seotud suurema tõenäosusega olla õigeaegne majandusaasta aruande esitamisel kõigis ettevõttegruppides. Siiski on majandusaasta pikkus mikroettevõtete puhul natuke keerukam - nii madalad kui ka kõrged väärtused näitavad negatiivset seost õigeaegsusega. Turu sisenemisbarjäärid ja konkreetse ettevõtte põhitegevus mõjutavad ettevõtte majandusaasta aruannete õigeaegsust. Põhitegevusala modelleerimine on raskendatud, sest eksisteerib mittelineaarseid seoseid ja väärtuste kardinaalsus on kõrge. Kõrged turu sisenemisbarjäärid on seotud kõrgema aruandluse õigeaegsusega väikeste ja keskmiste ning auditeeritud ettevõtete puhul. Probleemide puudumine auditeerimisel, rahvusvaheline tegevus ning madal töötajate arv, on positiivselt seotud auditeeritud ettevõtete majandusaasta aruannete õigeaegsusega.

Erinevate ettevõttegruppide puhul omavad tähtsat rolli ettevõtte suurus, vanus, finantsseisund ja aruandluse spetsiifika. Siiski võivad teised tegurid nagu ettevõttele tehtud määruste arv, turu sisenemisbarjäärid, tööjõukulude eksisteerimine ja auditit puudutavad aspektid omada erinevat mõju ettevõttegruppide lõikes.

Praktikas on majandusaasta mittesitamise ennustamisel oluline tähtsus. Logistiline regressioon näitab võrreldes otsustusmetsaga kehvemat ennustusvõimet. Samas keskmiselt käituvad logistiline regressioon ja otsustusmets sarnaselt. Siiski on väikeste ja keskmiste ettevõtete puhul otsustusmetsa meetod 2 protsendipunkti parem ning mikroettevõtete juures peaaegu 6 protsendipunkti parem. Üldiselt suudab otsustusmets ennustada umbes 60% mittetähtaegseid esitusi õigesti. Potentsiaalsed valenegatiivsed tulemused selles kontekstis võivad tuua ettevõtetele tarbetuid hoiatusi. Mudeli täpsuse parandamiseks on XGBoost hea kandidaat edasisteks täiustusteks, kuna teised uuringud on leidnud, et XGBoost ületab otsustusmetsa täpsust (Fauzan, 2018). Kuna otsustusmetsa mudel pakub minimaalset paremat tulemust, leiab autor, et lihtsam on kasutada logistilist regressiooni majandusaasta aruannete esitamise tähtaegsuse ennustamiseks.

LIST OF REFERENCES

Abernathy, J., Beyer, B., Masli, A., & Stefaniak, C. (2014). The association between characteristics of audit committee accounting experts, audit committee chairs, and financial reporting timeliness. *Advances in accounting, Elsevier, vol. 30(2)*, 283-297.

Accounting Act. (2021, November 23). Accounting Act. *Riigi Teataja*.

Amini, S., Elmore, R., Öztekin, Ö., & Strauss, J. (2021). Can Machines Learn Capital Structure Dynamics? *Journal of Corporate Finance (JCF), Vol. 70, No. 1*, pp. 1–22,.

Arruñada, B. (2011). Mandatory accounting disclosure by small private companies. *European Journal of Law and Economics Vol 32 No 3*, 377-413.

Bigus, J., & Hillebrand, C. (2017). Bank Relationships and Private Firms' Financial Reporting Quality. *European Accounting Review, 26(2), 379-409*.

Bolívar, M. P., & Galera, A. N. (2012). The Role of Fair Value Accounting in Promoting Government Accountability. *Abacus, 48.*, 348-386.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5-32.

Breuer, M., Hombach, K., & Müller, M. A. (2020). The Economics of Firms. Public Disclosure: Theory and Evidence SSRN 2020 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3037002.

Clatworthy, M., Peel, & M.J. (2016). The Timeliness of UK Private Company Financial Reporting: Regulatory and Economic Influences. *British Accounting Review, 48 (3), 297– 315*.

Commercial Code. (2017, May 9). RT I, 06.07.2023, 131.

Eierle, B. (2008). Filing Practice of Small and Medium-sized Companies: Empirical Findings from Austria. *International Small Business Journal*, 491-528.

Fauzan, M. A. (2018). The accuracy of XGBoost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl*, pp. 159-171.

Foundations Act. (2022, December 23). RT I, 23.12.2022, 31.

Gastwirth, J. (1972). The Estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics, Vol. 54, No. 3*, 306-316.

Government of the Republic of Estonia, Regulation No. 12, 2020. (2020, April 29). Support for partial compensation of losses resulting from the outbreak of the coronavirus causing the disease COVID-19 for tourism sector entrepreneurs.

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

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.

Jacobson, P., & Elliott, R. (1994). Costs and benefits of business information disclosure". *Accounting Horizons Vol. 8 No 4*, 80-96.

Jarmulska, B. (2020). Random forest versus logit models: which offers better early warning of fiscal stress? *Working Paper Series 2408, European Central Bank.*

Kallakas, K. (2021). *Factors causing failure to submit annual reports (on the basis of Estonian legal entities).*

Kips, K. (2021). *Reasons for Late Filings of Annual Reports in Estonia.*

Kohv, A. (2021). The Relationship Between Capital Structure and Profitability Based on Estonian Manufacturing Companies. Tallinn: Published: https://digikogu.taltech.ee/et/Item/8f1250af-88ca-4a6b-9ad4-a0f13554d16b.

Laidroo et al., L. (2020). *Impact assessment of annual report submission timeliness for the Ministry of Finance.* Talllinn: Estonian Research information System.

Laidroo, L., Küttim, M., Rumma, K., Paavo, S., & Avarmaa, M. (2024). Mandatory annual report filings of private companies – why late or missing? *Baltic Journal of Management Vol. 19 No. 1*, 123-144.

Leuz, C., & Wysocki, P. D. (2016, Februrary). The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research. *Journal of Accounting Research, 54*, 525-622.

Luik, A. (2024). *Electronic appendix for master thesis - Determinants Of Annual Report Submission Timeliness In Estonia*. Retrieved from GitHub: https://github.com/arturluik/thesis-timeliness

Lukason, O., & Camacho-Miñano, , M.-d.-M. (2019). Bankruptcy Risk, Its Financial Determinants and Reporting Delays: Do Managers Have Anything to Hide? *Risks, vol. 7(3)*, 1-15.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. Pages 4768–4777).

Luypaert, M., Caneghem, T. V., & Uytbergen, S. V. (2016). Financial statement filing lags: An empirical analysis among small firms. *International Small Business Journal*, 506-531.

Luypaert, M., Van, C., & Van, U. (n.d.). Financial Statements Filing lags: An Empirical Analysis Among Small Firms. *International Small Business Journal, 34(4), 506-531*.

Maingot, M., & Zeghal, D. (2006). Financial Reporting of Small Business Entities in Canada. *Journal of Small Business Management. 44. 513 - 530. 10.1111/j.1540-627X.2006.00191.x*.

McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

Minnis, M. a. (2017). Why regulate private firm disclosure and auditing? *Accounting and Business Research*, 473-502.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python.* O'Reilly Media, Inc.

Non-profit Associations Act . (n.d.). RT I, 23.12.2022, 15.

Okougbo, P. a. (2014). Timeliness of Financial Reporting in Nigeria. *South African Journal of Accounting Research*.

Owusu-Ansah, S. (2000). Timeliness Of Corporate Financial Reporting In Emerging Capital Markets: Empirical Evidence From The Zimbabwe Stock Exchange. *Accounting and Business Research. 30. 10.2139/ssrn.215929*.

Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2013). Interpreting random forest models using a feature contribution method. *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, (pp. 112-119). San Francisco, CA, USA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Pyle, D. (1999). *Data Preparation for Data Mining.* London: Morgan Kaufmann Publishers.

Strouhal, J., Nikitina-Kalamäe, M., & Gurvitš, N. (2014). Are Czech and Estonian Companies Willing To Publicly Present Their Financial Statements? Evidence from Czech and Estonian TOP100. *International Journal of Trade, Economics and Finance, 5 (4), 332–336*.

Sufiyati, S. (2017). The Impact of Corporate Attributes on the Timeliness of Financial Reporting in Indonesia Stock Exchange. *International Journal of Economic Perspectives Vol. 11*, 1720-1730.

Türel, A. G. (2010). Timeliness of Financial Reporting in Emerging Capital Markets: Evidence from Turkey. *European Financial and Accounting Journal*, 113-133.

Verrecchia, R. E. (2001, December). Essays on Disclosure. Journal of Accounting and Economics, 32 (1-3), 97-180.

Weetman, S. L. (2004). Timeliness of financial reporting: applicability of disclosure theories in an emerging capital market. *Accounting and Business Research*, 43-56.

Wittman, C. (2020). *Reporting Opaqueness of Private Firms.* PhD Disseratation, https://epub.uni-bayreuth.de/4630/.

Zellner, D. (2004). Variable Selection in Logistic Regression Models. *Communications in Statistics - Simulation and Computation, 33(3)*, 787-805.

APPENDICES

Appendix 1. Variable descriptions

Variable	M	SM	A	Description
maararv_V2E	1	1	0	Total number Cancellation of entry regulation: financial year report not submitted, none 0
r_d_ifrs	1	1	1	Reporting standard: 1 IFRS, 0 Estonian best practice
f_liq	0	1	1	Short. place coverage ratio - Current assets/ short place
loobumine_audit	0	0	1	1 waiver of audit opinion, other 0
end_quarter	1	1	1	Quarter of the end of the financial year
r_d_cons	1	1	1	Consolidated - 1 if consolidated, 0 otherwise
f_cash	1	1	1	Share of money in assets - money / assets
c_segc2	0	0	1	Proportion of the main non-concentrated business segment as emtak 2 (=perc_emtak2)
i_top4	1	1	1	top4 company's market share - Data of the Statistical Office EMTAK by letter 1 - proprietary costs (trade secret leakage cost)
c_barr2	1	1	1	Market entry barrier - tangible fixed assets/assets
c_d_cont	1	1	1	1 if the company has no sales revenue (c_d_sales), no labor cost (c_d_emplc), net asset problems (c_d_nassets1), deleted from the register in the same year or the following year, or an NGO 0 report (r_d_mtynull), otherwise 0
v40010	1	1	1	Short-term loan liabilities unconsolidated com. Debt consolidated ratios assume that must be >=0
v30010	1	1	1	Cash unconsolidated custom. Money consolidated ratios assume must be >=0
r_d_accf	0	0	1	Reporting format - 1/0 Does the reporting publication format used exceed the minimum required by law
rep_PDF	1	1	0	1 PDF, 0 muu
r_d_accanomaly	0	1	1	There is an anomaly in the records of the financial year report, which can be explained in very rare cases (not necessarily the case of a specific company) 1 otherwise 0
v30030	0	1	1	Inventories unconsolidated custom. Stocks consolidated, ratios assume must be >=0
c_d_late1	1	1	1	late
r_d_rmed	1	0	0	Selected report type: 1 medium company, 0 other (empty not selected)
r_d_rmicro	0	1	0	Selected report type: 1 microenterprise, 0 other (empty not selected)
industry	1	1	0	Company EMTAK code

Appendix 1. Variable descriptions (continuation)

Variable	M	SM	A	Description
c_d_sales	1	1	1	Sales revenue or existence of income - 1 if sales revenue or non-profit organization income is not i.e. = 0, otherwise 0
maa_algus	1	1	1	The beginning of the financial year according to the statute
c_d_o1comb	1	1	1	1 if the largest direct owner is combined (several owners with the same stake, some from Estonia and some from abroad), otherwise 0
vabatahtlikaudit	0	0	1	1 voluntarily audited, 0 unaudited
v40020	1	1	1	Long-term loan liabilities unconsolidated com. Debt consolidated ratios assume that must be >=0
v50010	1	1	1	Cash flows from operating activities unconsolidated com. Cash flows from operating activities consolidated
size_scaled	1	0	1	Industry-adjusted natural log of total assets (EMTAK 2 alusel)
c_segc1	0	0	1	Proportion of the main non-concentrated business segment as emtak 1 (=perc_emtak1)
r_d_abper	1	1	1	1 Fiscal year longer than calendar year, other 0
c_d_vat	1	1	1	1 if the legal person has been liable for VAT for more than 6 months in the previous financial year, otherwise 0
f_lev1	1	0	0	Leverage - Liabilities/assets
maararv_X1	1	1	1	Total number Warning order for deletion from the register: financial year report not submitted, none 0
c_d_o1ocomb	1	1	1	1 if the largest direct owner is a combined type, i.e. several largest owners with equal shares, and some of them are physical and some are legal entities, other 0
r_orddiff	0	1	1	The complexity of preparing the report based on the actually selected report format and, if it is missing, based on the legal requirement - 1 micro-enterprise, 2 small enterprise, 3 medium-sized enterprise, 4 large enterprise
f_lev3	1	1	1	Leverage - Long-term high/ (long-term high + ok)
f_cfoper	1	1	1	Cash flow from business activities - 1 if >0, otherwise 0
count_tax	1	1	1	In several quarters at the end of the calendar year, the tax debt is at least EUR 1
rep_tabelid	1	1	0	1 tables, 0 other
r_d_pdfnot	1	1	0	Reporting format pdf or notary - 1 if notarized or pdf, 0 other
c_d_o1est	0	1	1	1 if the largest direct owner from Estonia, otherwise 0
v30040	0	0	1	Current assets unconsolidated custom. Current assets consolidated
f_lev2	1	0	1	Gearing - Long-term loan/equity
v40040	0	1	1	Equity unconsolidated custom. Equity consolidated, ratios assume must be >=0
f_nprofm	1	0	1	Profitability - net profit/sales revenue, NGO result/total revenue
d_ülevaatus	0	0	1	1 review, 0 other
r_sizeg	0	1	0	selected report format 1 Micro, 2 small, 3 medium, 4 large
v20200	0	0	1	Interest expenses combined
v30080	0	1	1	Fixed assets unconsolidated custom. Fixed assets consolidated
f_d_altman	1	1	1	1 According to Altman's bankruptcy ratio, bankruptcy is expected, 0 other

Appendix 1. Variable descriptions (continuation)

Variable	M	SM	A	Description
r_d_invsub	1	1	1	1- daughter of inv and related e/v or has submitted a console report (r_d_cons), 0 if neither
c_finend	1	0	0	the month of the end of the company's financial year
f_lev5	1	0	1	(Loan amount incl. rent amount + either rent amount)/equity
oigvorm_initial	0	1	0	Legal form from the original Business Register inquiries by year with spaces: 1 joint-stock company (AS), 2 general partnership (UT), 3 FIE, 4 limited partnership (UÜ), 5 limited partnership (OÜ), 6 & 16 profit cooperative, 7, 8, foundation, 10 branch, 11 European economic interest association, 12 European business association (Societas Europaea), 17 European cooperative, 18 European territorial cooperation group, 20 executive state authority or other state institution, 21 public legal entity, 22 - locally owned. institution, 23 housing cooperatives, 24 land improvement cooperatives
puhas_rõhutamine_audit	0	0	1	1 pure audit verdict with emphasis, other 0
alt_x2	1	0	0	Altman bankruptcy multiplier - 2nd element
zmij_x3	0	1	1	Zmijewski skoor - 3rd element
f_dsr	0	0	1	The turnover margin of buyers' invoices
maararv_V2T	0	1	1	Total number Deletion from the register: annual report not submitted, no 0
c_county	1	1	1	County 1-16, 1- Tallinn, 2 Harju County, 3 Ida-Viru, 4 Lääne-Viru, 5 - Jõgeva, 6- Tartu, 7-Pölva, 8-Võru, 9-Valga, 10 Vilandi, 11 Järva, 12 Rapla , 13 Pärnu, 14 Lääne, 15 Hiiu, 16 Saare
rep_notar	1	1	0	1 through a notary, 0 other
r_d_rlarge	1	1	0	Selected report type: 1 large company, 0 other (unselected blank)
f_lnsize2	1	1	1	Size - ln sales revenue
oigvorm_arc	1	1	0	The legal form taken from the last request of the Business Register, i.e. it is unchanged over time
c_empl	1	1	1	the number of employees in the company
f_d_loss	1	1	1	Loss - 1 if the net profit or the result of the NGO is negative, otherwise 0
v50030	1	1	1	Cash flows from investment activities unconsolidated com. Cash flows from investing activities consolidated
maararv_X2	1	1	1	Total number Warning order for deletion from the register: financial year report not submitted, none 0
f_d_roe	0	1	1	1 if ROE y-o-y change was negative, 0 if positive
r_repper	1	1	1	Length of financial year in days (=year_length)
c_d_o1for	0	1	1	1 if the largest direct owner from abroad, otherwise 0
sharefirm1	1	1	1	The company's market share of the sales revenue, based on the EMTAK one-letter code
märkusega_rõhutamine_audit	0	0	1	1 audit opinion with emphasis, 0 other
r_d_finend	1	0	0	End of honor year - 1 if Dec, other 0

Appendix 1. Variable descriptions (continuation)

Variable	M	SM	A	Description
r_d_lowqual	0	1	0	1 if there are errors or anomalous entries in the report indicating low accounting quality (r_d_accerror, r_d_accanomaly), error corrections have been made (r_d_error) or the auditor's decision is not clean, otherwise 0
c_d_overtax1	1	1	1	1- is Debt at the end of the financial year (at least EUR 1), other 0
c_d_nassets2	1	1	0	1 netovara Oü-s alla 1250 Asil alla 12500, muu 0
eitav_audit	0	0	1	1 negative audit opinion, other 0
f_dsi	0	1	1	Inventory cost margin
c_d_nassets1	1	1	0	1 Net assets are 0 or negative, otherwise 0
r_growth	1	1	1	sales revenue growth y-o-y - (Sales revenue t - sales revenue t-1)/sales revenue t-1; In the case of an NGO, income growth
zmij_x2	1	1	0	Zmijewski skoor - 2nd element
v20100	1	0	0	Cost of goods sold equivalent for all companies
alt_x1	1	0	0	Altman's bankruptcy multiplier - 1st element
c_sizeg	0	1	1	1 Micro, 2 small, 3 medium, 4 large
maa_lopp	1	1	1	End of the financial year according to the statute
v40030	0	0	1	Short-term liabilities unconsolidated com. Current liabilities consolidated ratios assume must be >=0
c_d_mnc1	0	0	1	International ev - 1 foreign sales revenue 20% (assuming that the missing indication of foreign revenue indicates its absence), other 0
v30060	1	1	1	Tangible fixed assets unconsolidated custom. Tangible fixed assets consolidated, ratios assume must be >=0
f_d_prof	1	1	1	profit decreased compared to last year by 1, otherwise 0
r_d_rmissing	1	1	1	Selected report type: 1 not selected, 0 selected
c_d_emplc	1	1	1	Labor cost - not present in the report 1, otherwise 0 (is there any activity at all); labor costs are taken from three different records V20070, v62020 and v62030.
c_o1ctry	1	1	1	Country of residence of the largest direct owner (only AS and OÜ indicator)
c_d_end	1	1	1	1 company has been deleted from the register in the current year, 0 others
r_d_intang	1	1	1	1 if >0, 0 if absent
maararv_HM	1	1	1	Total number Fine warning order: financial year report not submitted, none 0
ln_size	1	1	1	logarithm of total sector sales revenue according to EMTAK 2 number
c_o1ctryt	1	1	1	Same as c_o1ctry only for Stata formatted as text
c_no_share	1	1	1	Number of owners
maararv_HP	1	1	1	Total number of Fine warning order: on deficiencies in the annual report, none 0
maararv_V2	1	1	0	Total number Entry order for deletion from the register: annual report not submitted, no 0
c_no_dir	1	1	1	Number of board members

Appendix 1. Variable descriptions (continuation)

Variable	M	SM	A	Description
r_d_accerror	1	1	1	There is an error in the entries of the financial year report, which cannot be 1 according to the accounting rules, otherwise 0
r_big4	0	0	1	Auditor big4 - 1 as auditor big4, 0 other
c_age	1	1	1	Age of the company - from establishment to the end of the reporting year
r_d_rsmall	0	1	0	Selected report type: 1 small business, 0 other (empty not selected)
c_barr1	1	1	1	Market entry barrier, trade secret disclosure rate - tangible fixed assets/sales revenue =v30060/(v10010/x*365)*100
r_accr	0	1	1	Share of assets not received from buyers - claims against buyers / assets
f_roe	0	1	0	ROE - EBT/Equity
r_aud2	0	0	1	1 unaudited report, 2 mandatorily audited report, 3 voluntarily audited report
c_d_mnc2	0	0	1	International ev - 1 foreign sales income greater than 1% (assuming that the missing indication of foreign income indicates its absence), other 0
f_d_liq	1	1	1	Illiquid - 1 if f_liq < 1, 0 otherwise
r_d_acc_diff	1	1	1	Accounting complexity 1 if intangible assets (r_d_intang), group (r_d_invsub), multiple business segments (c_d_segind2), goodwill (r_d_goodwill) =1, other 0
oigvorm	0	1	0	Legal form from the requests of the Business Register by year, where the blanks are filled: 1 joint-stock company (AS), 2 general partnership (UT), 3 FIE, 4 limited partnership (UÜ), 5 limited partnership (OÜ), 6 & 16 profit cooperative, 7, 8, foundation, 10 branch, 11 European economic interest association, 12 European company (Societas Europaea), 17 European cooperative, 18 European territorial cooperation group, 20 executive state authority or other state institution, 21 public legal entity, 22 - locally owned. institution, 23 housing cooperatives, 24 land improvement cooperatives
c_segcm	1	1	1	Proportion of the main non-concentrated business segment (=perc_mainact)
c_top1	1	1	1	The largest direct owner's participation in the share or share capital (only AS and OÜ indicator)
f_d_roa	0	1	1	1 if ROA y-o-y change was negative, 0 if positive
r_d_goodwill	1	1	1	1 if goodwill is different from 0, otherwise 0
rep_RMP	1	1	1	1 RMP, 0 other
r_risk	0	0	1	Financial reporting risk - (trade receivables + inventories)/assets
r_d_xbrl	1	1	1	Reporting format xbrl - 1 if xbrl, 0 otherwise
korr_aud_ev	0	0	1	The registry code of the auditing company has been corrected

Appendix 1. Variable descriptions (continuation)

Variable	M	SM	A	Description
v64040	1	1	1	Operating lease liability (all expenses, whether consolidated or unconsolidated, added together)
r_inv	0	1	1	Share of inventories to assets - Inventories/assets
f_lev4	1	1	1	Debt - Borrower/obligations
maarusearv	1	1	1	the total number of regulations for a given company in a given year
v30100	0	1	1	Trade receivables in the combined balance sheet and for anomalies adjusted from the notes

Source: Author's calculation

1 – indicating variable is considered in the given sample, 0 – excluded from the respective sample, M – sample of micro entities; SM – sample of small and medium entities; A – sample of audited entities

Appendix 2. Descriptive statistics – Micro companies

Variable	Mean	Std	Min	Max
maararv_V2E	0.00	0.03	0.00	1.00
r_d_ifrs	0.00	0.02	0.00	1.00
end_quarter	3.95	0.33	1.00	4.00
r_d_cons	0.00	0.03	0.00	1.00
f_cash	40.32	92.33	0.00	46600.00
i_top4	10.90	7.70	0.00	65.56
c_barr2	10.31	24.93	0.00	3333.33
c_d_cont	0.54	0.50	0.00	1.00
v40010	419.40	2753.60	-70302.81	84791.00
v30010	5292.77	11450.81	-52177.00	173975.00
rep_PDF	0.00	0.04	0.00	1.00
c_d_late1	0.29	0.45	0.00	1.00
r_d_rmed	0.01	0.09	0.00	1.00
industry	57583.10	25144.30	0.00	96099.00
c_d_sales	0.23	0.42	0.00	1.00
maa_algus	1.18	1.80	1.01	31.12
c_d_o1comb	0.00	0.01	0.00	1.00
v40020	300.07	3260.73	-46832.00	573208.00
v50010	1202.03	19493.49	-10000000.00	1284668.00
size_scaled	-1.13	1.41	-12.25	3.22
r_d_abper	0.04	0.20	0.00	1.00
c_d_vat	0.34	0.47	0.00	1.00
f_lev1	0.09	0.14	0.00	0.69
maararv_X1	0.00	0.04	0.00	1.00
c_d_o1ocomb	0.00	0.01	0.00	1.00
f_lev3	0.01	0.07	-4.25	17.77
f_cfoper	0.12	0.33	0.00	1.00
count_tax	0.16	0.62	0.00	4.00
rep_tabelid	0.98	0.13	0.00	1.00
r_d_pdfnot	0.00	0.05	0.00	1.00
f_lev2	0.01	0.07	0.00	1.76
f_nprofm	-48.43	12275.55	-5951700.00	1897900.00
f_d_altman	0.05	0.22	0.00	1.00
r_d_invsub	0.02	0.12	0.00	1.00
c_finend	11.77	1.32	1.00	12.00
f_lev5	0.04	0.19	-1.00	46.60
alt_x2	-1.39	236.82	-122499.01	1056.00
c_county	3.97	4.23	0.00	16.00

Appendix 2. Descriptive statistics – Micro companies (continuation)

Variable	Mean	Std	Min	Max
rep_notar	0.00	0.04	0.00	1.00
r_d_rlarge	0.00	0.01	0.00	1.00
f_lnsize2	6.84	3.96	-2.75	15.09
oigvorm_arc	5.00	0.12	1.00	5.00
c_empl	0.34	0.67	0.00	65.00
f_d_loss	0.26	0.44	0.00	1.00
v50030	-352.69	39258.85	-16508157.00	9995802.00
maararv_X2	0.03	0.17	0.00	1.00
r_repper	359.61	43.46	1.00	549.00
sharefirm1	0.00	0.24	-0.01	96.05
r_d_finend	0.96	0.19	0.00	1.00
c_d_overtax1	0.04	0.19	0.00	1.00
c_d_nassets2	0.05	0.22	0.00	1.00
c_d_nassets1	0.01	0.12	0.00	1.00
r_growth	132.97	25923.40	-55653.59	14616954.00
zmij_x2	0.02	0.08	0.00	6.04
v20100	-3734.30	22172.20	-10487432.00	99303.00
alt_x1	0.49	0.42	-1.17	11.11
maa_lopp	30.90	2.07	1.01	31.12
v30060	2322.22	9579.21	-3384.00	174127.80
f_d_prof	0.34	0.48	0.00	1.00
r_d_rmissing	0.59	0.49	0.00	1.00
c_d_emplc	0.31	0.46	0.00	1.00
c_o1ctry	235.23	12.12	0.00	567.00
c_d_end	0.00	0.00	0.00	1.00
maararv_HM	0.00	0.03	0.00	1.00
ln_size	20.47	1.47	0.00	23.55
c_o1ctryt	235.23	12.12	0.00	567.00
c_no_share	1.00	0.07	0.00	5.00
maararv_HP	0.00	0.04	0.00	1.00
maararv_V2	0.00	0.03	0.00	1.00
c_no_dir	1.08	0.32	0.00	10.00
c_age	5.96	5.14	0.00	23.00
c_barr1	117.08	19989.65	0.00	10197884.00
f_d_liq	0.03	0.18	0.00	1.00
r_d_acc_diff	0.15	0.35	0.00	1.00
c_segcm	74.64	40.80	0.00	100.00

Appendix 2. Descriptive statistics – Micro companies (continuation)

Variable	Mean	Std	Min	Max
c_top1	99.67	5.47	0.00	100.00
r_d_goodwill	0.00	0.01	0.00	1.00
rep_RMP	0.01	0.09	0.00	1.00
r_d_xbrl	0.01	0.08	0.00	1.00
v64040	32.82	712.17	-9092.00	95508.00
f_lev4	0.02	0.15	0.00	25.75
maarusearv	0.04	0.20	0.00	3.00

Source: Author's calculation

Notes: For variable descriptions see Appendix 1.

Appendix 3. Descriptive statistics – Small and medium companies

Variable	Mean	Std	Min	Max
maararv_V2E	0.00	0.03	0.00	1.00
r_d_ifrs	0.00	0.03	0.00	1.00
f_liq	78.35	5480.31	0.00	3628676.30
end_quarter	3.96	0.31	1.00	4.00
r_d_cons	0.00	0.06	0.00	1.00
f_cash	30.83	60.82	0.00	28615.39
i_top4	10.44	7.50	0.00	73.66
c_barr2	20.68	55.13	0.00	21369.23
c_d_cont	0.67	0.47	0.00	1.00
v40010	21112.86	614383.59	-169615.00	212793456.00
v30010	30109.43	859546.19	-278274.00	292635008.00
rep_PDF	0.00	0.04	0.00	1.00
r_d_accanomaly	0.00	0.07	0.00	1.00
v30030	15123.14	131096.04	-31548.00	61362000.00
c_d_late1	0.28	0.45	0.00	1.00
r_d_rmicro	0.03	0.18	0.00	1.00
industry	52711.26	23478.77	0.00	96099.00
c_d_sales	0.10	0.30	0.00	1.00
maa_algus	1.10	1.27	1.01	31.12
c_d_o1comb	0.02	0.12	0.00	1.00
v40020	31284.40	1228716.19	-350000.00	573670016.00
v50010	7425.04	394192.59	-169606000.00	101306000.00
r_d_abper	0.04	0.19	0.00	1.00
c_d_vat	0.65	0.48	0.00	1.00
maararv_X1	0.00	0.04	0.00	1.00
c_d_o1ocomb	0.01	0.11	0.00	1.00
r_orddiff	1.98	0.23	1.00	4.00
f_lev3	0.11	0.26	-7.98	56.07
f_cfoper	0.23	0.42	0.00	1.00
count_tax	0.32	0.88	0.00	4.00
rep_tabelid	0.98	0.13	0.00	1.00
r_d_pdfnot	0.00	0.06	0.00	1.00
c_d_o1est	0.91	0.28	0.00	1.00
v40040	93945.27	1155915.18	-55764000.00	333716000.00
r_sizeg	0.73	0.97	0.00	4.00
v30080	94285.20	1882581.18	-39976212.00	510912992.00
f_d_altman	0.31	0.46	0.00	1.00
r_d_invsub	0.04	0.19	0.00	1.00

Appendix 3. Descriptive statistics – Small and medium companies (continuation)

Variable	Mean	Std	Min	Max
oigvorm_initial	4.75	1.09	0.00	5.00
zmij_x3	78.35	5480.31	0.00	3628676.30
maararv_V2T	0.00	0.00	0.00	1.00
c_county	3.95	4.28	0.00	16.00
rep_notar	0.00	0.05	0.00	1.00
r_d_rlarge	0.00	0.02	0.00	1.00
f_lnsize2	8.51	4.70	-2.96	17.82
oigvorm_arc	4.97	0.33	1.00	5.00
c_empl	1.84	16.48	0.00	6613.45
f_d_loss	0.32	0.47	0.00	1.00
v50030	-6915.35	439880.06	-211879696.00	68947000.00
maararv_X2	0.02	0.16	0.00	1.00
f_d_roe	0.39	0.49	0.00	1.00
r_repper	362.23	35.00	1.00	634.00
c_d_o1for	0.09	0.28	0.00	1.00
sharefirm1	0.01	0.23	-0.02	100.00
r_d_lowqual	0.01	0.08	0.00	1.00
c_d_overtax1	0.07	0.26	0.00	1.00
c_d_nassets2	0.12	0.32	0.00	1.00
f_dsi	1492.07	199293.01	0.00	82001080.00
c_d_nassets1	0.08	0.28	0.00	1.00
r_growth	451.18	90005.65	-54401.67	58519048.00
zmij_x2	8.28	1158.02	0.00	619200.06
c_sizeg	2.00	0.02	2.00	3.00
maa_lopp	31.01	1.43	1.01	31.12
v30060	37330.88	487455.11	-20479.00	208560496.00
f_d_prof	0.36	0.48	0.00	1.00
r_d_rmissing	0.63	0.48	0.00	1.00
c_d_emplc	0.54	0.50	0.00	1.00
c_o1ctry	242.10	42.73	0.00	567.00
c_d_end	0.00	0.00	0.00	1.00
r_d_intang	0.05	0.21	0.00	1.00
maararv_HM	0.00	0.03	0.00	1.00
ln_size	20.86	1.45	0.00	23.55
c_o1ctryt	242.10	42.73	0.00	567.00
c_no_share	1.68	1.81	0.00	367.00
maararv_HP	0.00	0.05	0.00	1.00

Appendix 3. Descriptive statistics – Small and medium companies (continuation)

Variable	Mean	Std	Min	Max
maararv_V2	0.00	0.03	0.00	1.00
c_no_dir	1.45	0.68	0.00	11.00
r_d_accerror	0.00	0.03	0.00	1.00
c_age	7.51	5.68	0.00	23.00
r_d_rsmall	0.32	0.47	0.00	1.00
c_barr1	254.53	13831.07	0.00	6168783.00
r_accr	9.86	1274.48	0.00	616180.00
f_roe	70.73	140051.86	-32921896.00	84032096.00
f_d_liq	0.22	0.42	0.00	1.00
r_d_acc_diff	0.26	0.44	0.00	1.00
oigvorm	4.99	0.17	1.00	5.00
c_segcm	73.35	40.04	0.00	100.00
c_top1	76.94	30.80	0.00	100.00
f_d_roa	0.40	0.49	0.00	1.00
r_d_goodwill	0.00	0.03	0.00	1.00
rep_RMP	0.00	0.06	0.00	1.00
r_d_xbrl	0.01	0.09	0.00	1.00
v64040	513.20	10836.68	-125339.00	2321030.00
r_inv	10.82	50.12	0.00	28656.52
f_lev4	0.13	0.30	0.00	47.88
maarusearv	0.03	0.18	0.00	4.00
v30100	9793.94	49493.40	0.00	7687636.00

Source: Author's calculation

Notes: For variable descriptions see Appendix 1.

Appendix 4. Descriptive statistics – Audited companies

Variable	Mean	Std	Min	Max
r_d_ifrs	0.02	0.13	0.00	1.00
f_liq	346.42	19927.77	0.00	3828899.30
loobumine_audit	0.01	0.10	0.00	1.00
end_quarter	3.93	0.39	1.00	4.00
r_d_cons	0.15	0.35	0.00	1.00
f_cash	12.83	19.32	0.00	100.00
c_segc2	80.62	33.30	0.00	100.00
i_top4	10.52	9.69	0.00	68.87
c_barr2	22.48	28.58	0.00	100.00
c_d_cont	0.83	0.38	0.00	1.00
v40010	618853.50	5681160.25	0.00	677766976.00
v30010	503396.65	2621314.07	-6403.00	223547008.00
r_d_accf	0.06	0.24	0.00	1.00
r_d_accanomaly	0.00	0.02	0.00	1.00
v30030	681865.81	4948883.37	-2329.00	909749312.00
c_d_late1	0.24	0.43	0.00	1.00
c_d_sales	0.02	0.14	0.00	1.00
maa_algus	1.02	0.30	1.01	31.12
c_d_o1comb	0.01	0.12	0.00	1.00
vabatahtlikaudit	0.06	0.24	0.00	1.00
v40020	1298097.19	5874860.50	-26337.00	236528992.00
v50010	352957.10	2915999.06	-146819824.00	188299008.00
size_scaled	3.79	1.55	-9.04	11.64
c_segc1	82.47	33.13	0.00	100.00
r_d_abper	0.02	0.12	0.00	1.00
c_d_vat	0.89	0.31	0.00	1.00
maararv_X1	0.00	0.04	0.00	1.00
c_d_o1ocomb	0.02	0.12	0.00	1.00
r_orddiff	2.23	0.48	0.00	4.00
f_lev3	0.18	0.28	-0.25	1.00
f_cfoper	0.52	0.50	0.00	1.00
count_tax	0.27	0.74	0.00	4.00
c_d_o1est	0.77	0.42	0.00	1.00
v30040	2555517.12	13407220.90	0.00	1762546688.00
f_lev2	5.21	119.32	0.00	15122.71
v40040	3447599.32	13445807.15	-74630576.00	1077577728.00
f_nprofm	-255.39	102449.49	-23459634.00	9344270.00
d_ülevaatus	0.22	0.41	0.00	1.00

Appendix 4. Descriptive statistics – Audited companies (continuation)

Variable	Mean	Std	Min	Max
v20200	-58879.30	303836.77	-17417000.00	21482000.00
v30080	4034491.62	18572567.63	0.00	1057345024.00
f_d_altman	0.39	0.49	0.00	1.00
r_d_invsub	0.22	0.42	0.00	1.00
f_lev5	-5.38	1889.52	-485227.28	19039.26
puhas_rõhutamine_audit	0.04	0.19	0.00	1.00
zmij_x3	346.42	19927.77	0.00	3828899.30
f_dsr	283.07	23371.11	0.00	4934918.50
maararv_V2T	0.00	0.00	0.00	1.00
c_county	3.38	3.93	0.00	16.00
f_lnsize2	13.12	4.15	0.00	21.04
c_empl	38.43	167.24	0.00	18140.30
f_d_loss	0.24	0.43	0.00	1.00
v50030	-293109.56	2830367.04	-220226000.00	97016664.00
maararv_X2	0.01	0.11	0.00	1.00
f_d_roe	0.43	0.49	0.00	1.00
r_repper	364.39	20.06	3.00	549.00
c_d_o1for	0.22	0.42	0.00	1.00
sharefirm1	0.17	1.03	-0.09	99.92
märkusega_rõhutamine_audit	0.01	0.11	0.00	1.00
c_d_overtax1	0.05	0.21	0.00	1.00
eitav_audit	0.00	0.06	0.00	1.00
f_dsi	4954.09	580544.33	0.00	137765632.00
r_growth	4373.40	445232.47	-2128.67	81099904.00
c_sizeg	2.17	0.44	0.00	4.00
maa_lopp	31.09	0.31	1.01	31.12
v40030	1679611.50	10893945.84	-11614.00	796801024.00
c_d_mnc1	0.69	0.46	0.00	1.00
v30060	1631336.67	14852789.74	-499.00	1052772992.00
f_d_prof	0.38	0.48	0.00	1.00
r_d_rmissing	0.71	0.45	0.00	1.00
c_d_emplc	0.80	0.40	0.00	1.00
c_o1ctry	239.90	39.57	0.00	567.00
c_d_end	0.00	0.01	0.00	1.00
r_d_intang	0.24	0.43	0.00	1.00
maararv_HM	0.00	0.02	0.00	1.00
ln_size	21.18	1.40	9.70	23.55
c_o1ctryt	239.90	39.57	0.00	567.00

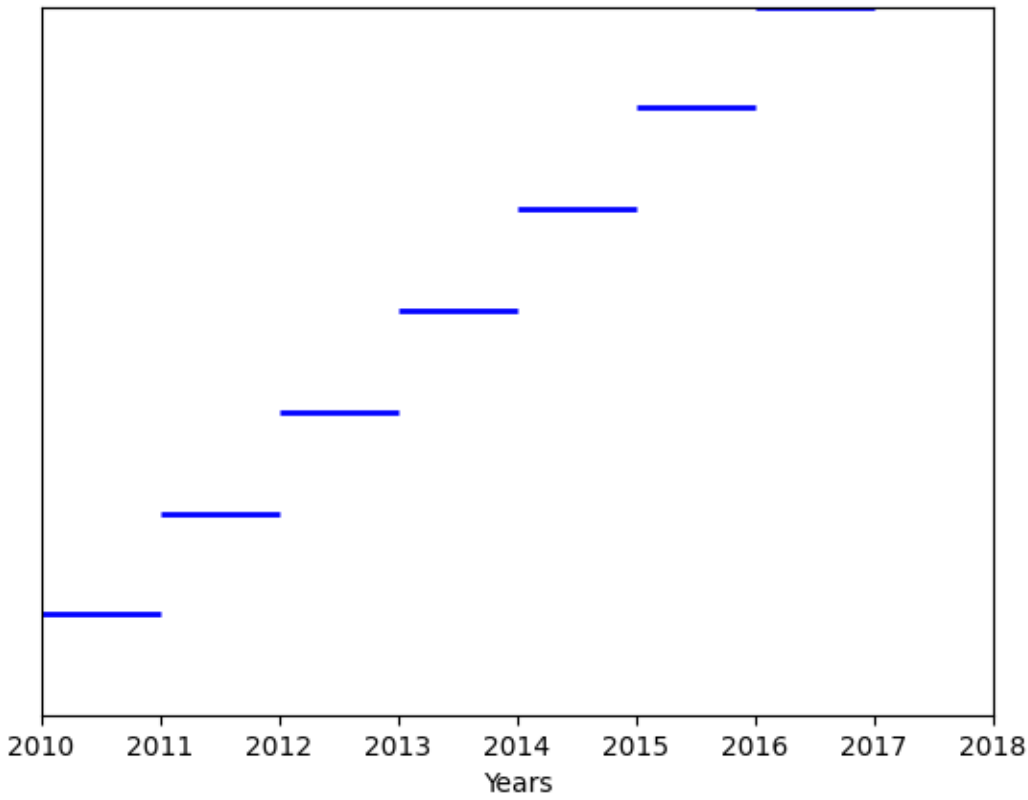
Appendix 4. Descriptive statistics – Audited companies (continuation)

Variable	Mean	Std	Min	Max
c_no_share	2.26	8.83	0.00	502.00
maararv_HP	0.00	0.05	0.00	1.00
c_no_dir	1.66	0.89	0.00	12.00
r_d_accerror	0.00	0.02	0.00	1.00
r_big4	0.07	0.26	0.00	1.00
c_age	11.71	5.74	0.00	23.00
c_barr1	788.75	70577.48	0.00	13512607.00
r_accr	17.43	965.32	0.00	245183.33
r_aud2	2.06	0.24	2.00	3.00
c_d_mnc2	0.77	0.42	0.00	1.00
f_d_liq	0.25	0.43	0.00	1.00
r_d_acc_diff	0.61	0.49	0.00	1.00
c_segcm	75.08	35.04	0.00	100.00
c_top1	73.77	35.89	0.00	100.00
f_d_roa	0.42	0.49	0.00	1.00
r_d_goodwill	0.03	0.18	0.00	1.00
rep_RMP	0.00	0.01	0.00	1.00
r_risk	20.04	31.38	0.00	1714.07
r_d_xbrl	0.00	0.07	0.00	1.00
korr_aud_ev	7483459.70	5059916.73	0.00	14849070.00
v64040	146493.05	5414219.18	-120154000.00	875420032.00
r_inv	12.91	20.64	0.00	100.00
f_lev4	0.31	0.37	0.00	14.09
maarusearv	0.02	0.14	0.00	2.00
v30100	503366.36	2482149.14	0.00	397090912.00

Source: Author's calculation

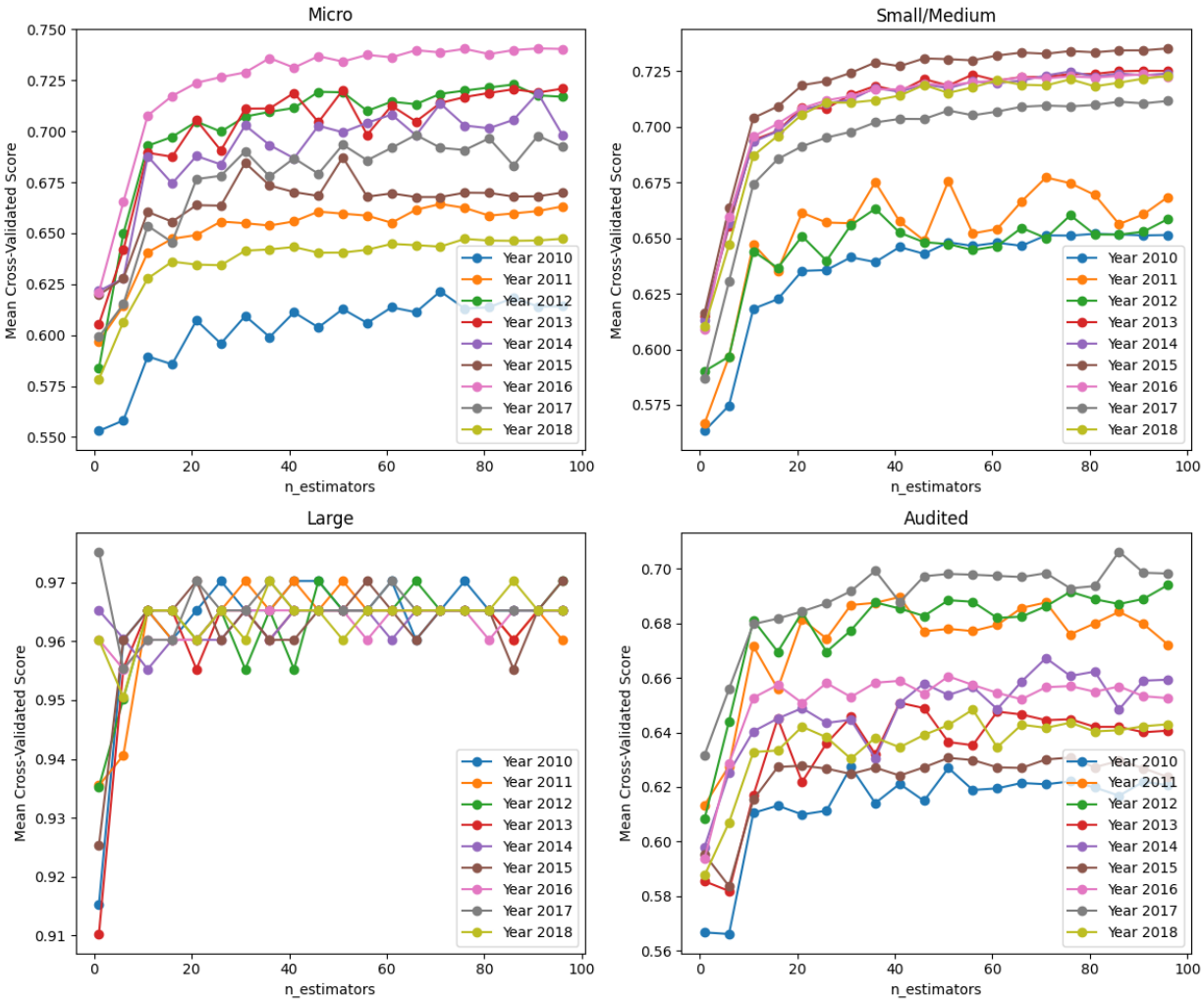
Notes: For variable descriptions see Appendix 1.

Appendix 5. Sliding window



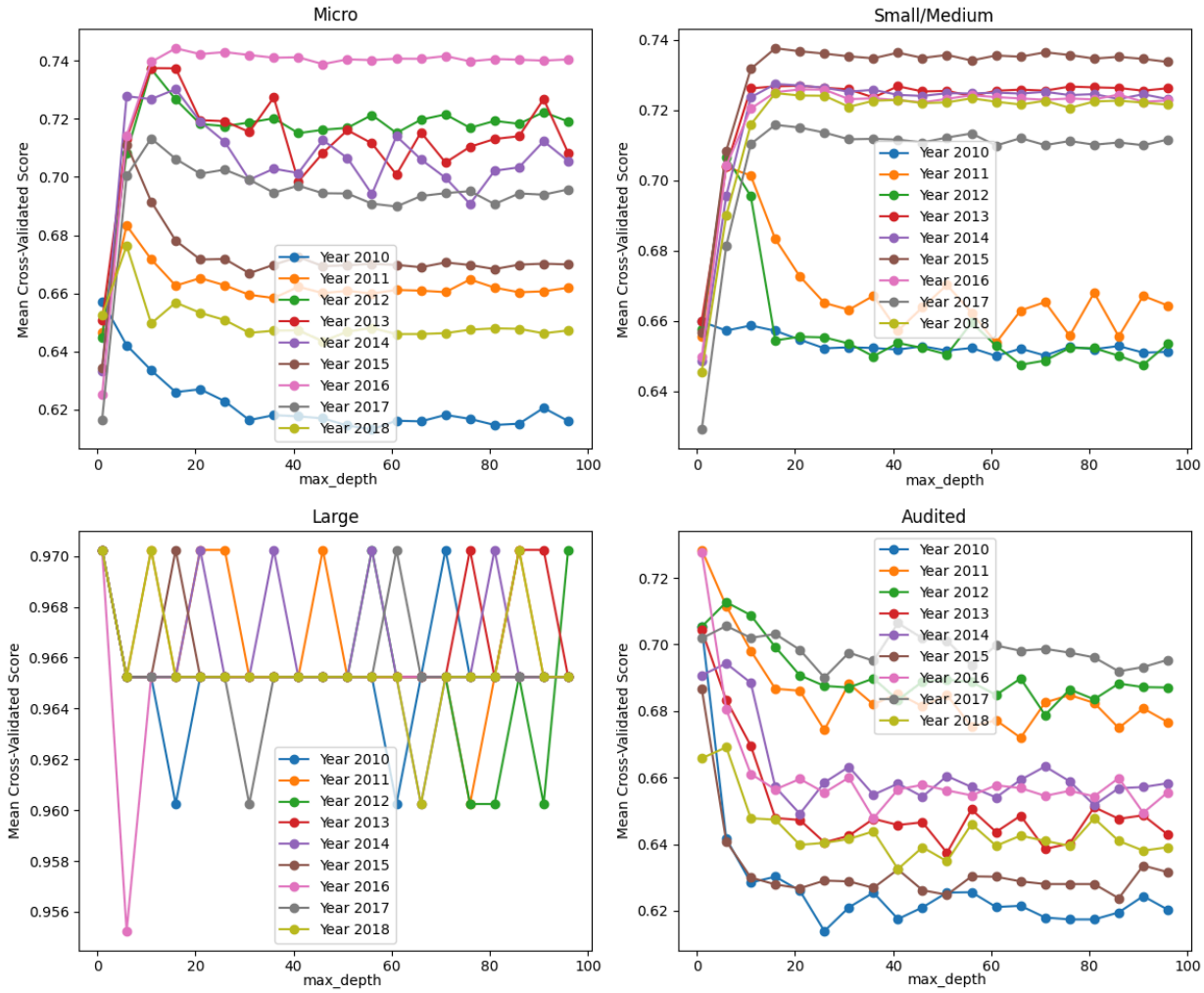
Source: Author's calculation

Appendix 6. Random forest – n_estimators hyperparameter sensitivity



Source: Author's calculation

Appendix 7. Random forest – max_depth hyperparameter sensitivity

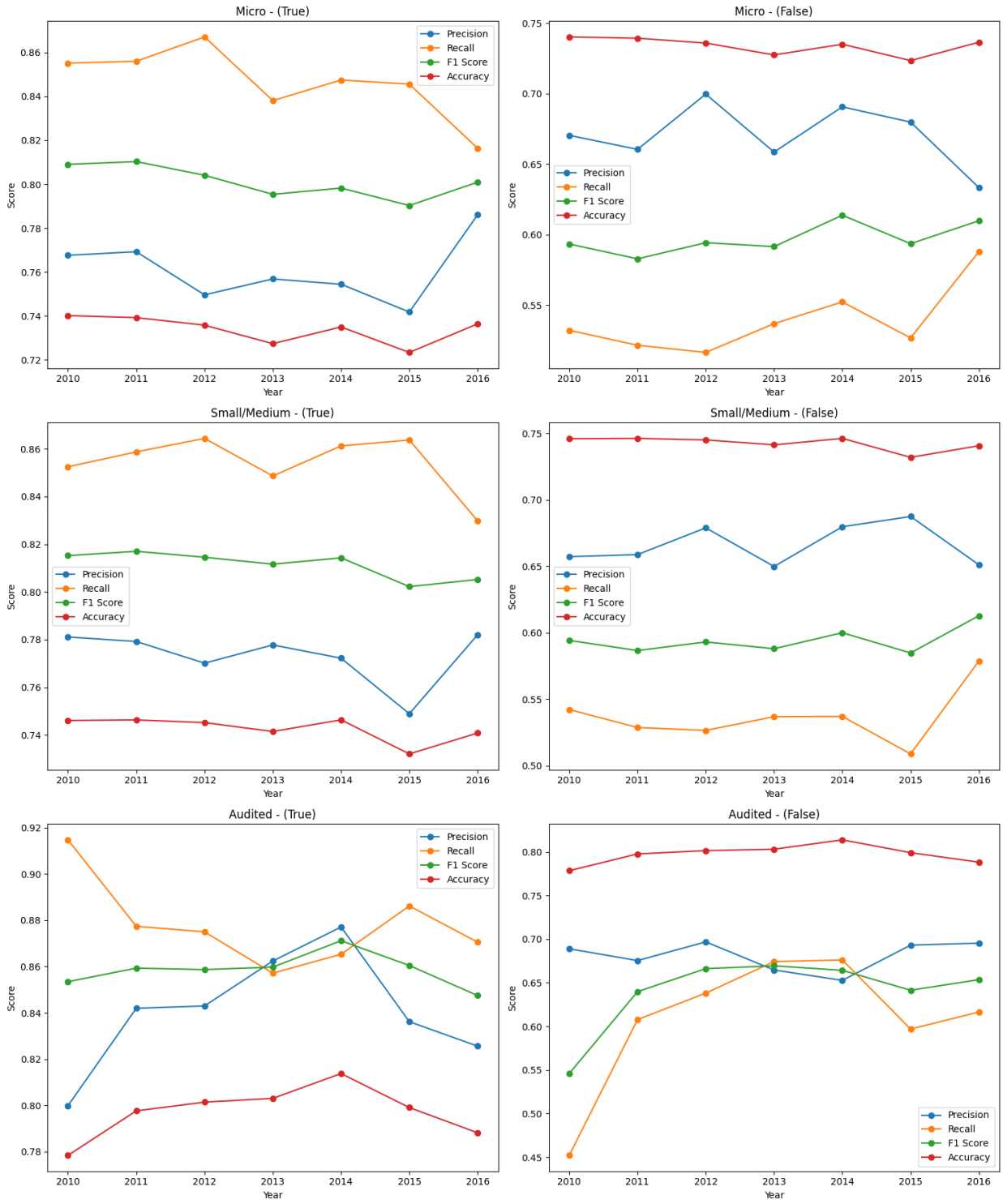


Source: Author's calculations

Appendix 8. Random forest – model accuracy

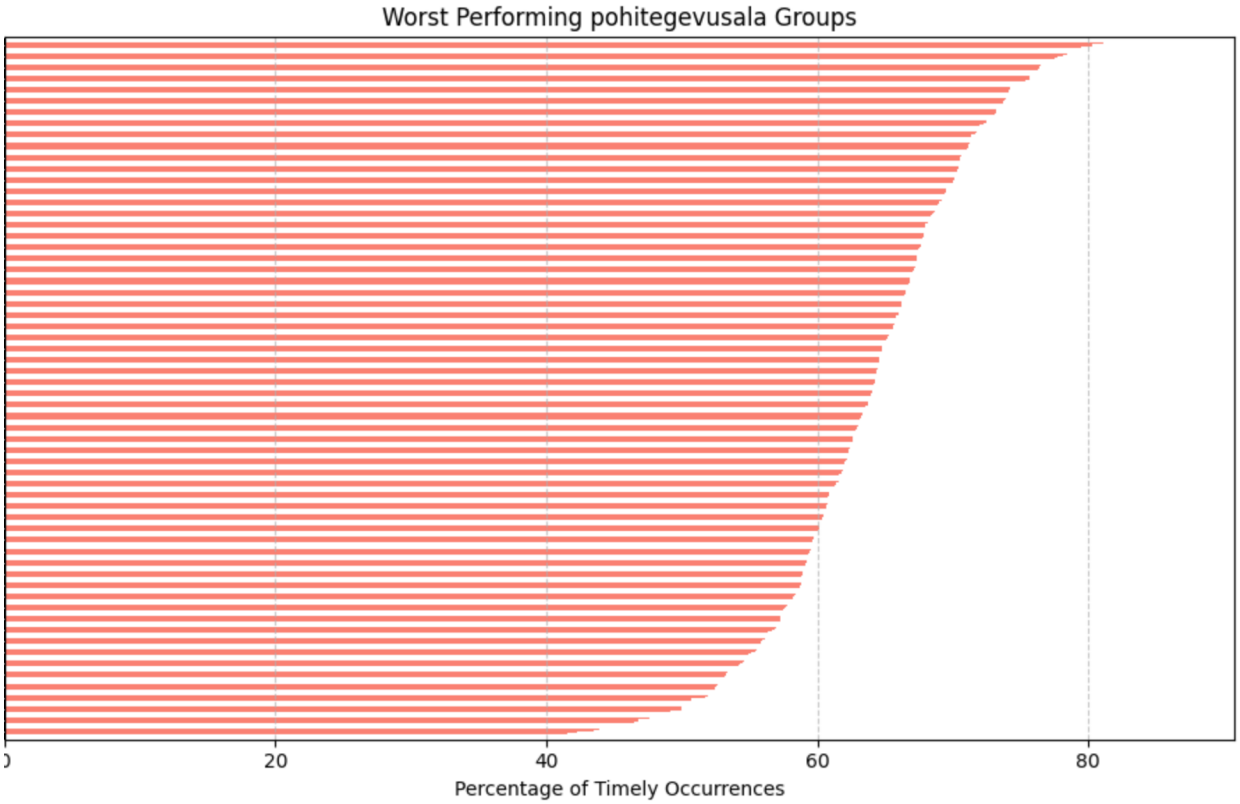
Source: Author's calculation

Appendix 9. Logistic regression - model accuracy



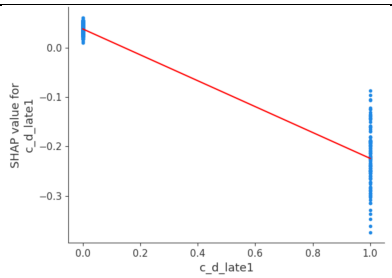
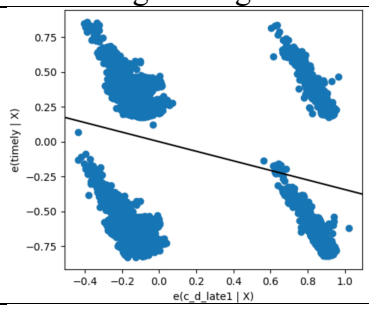
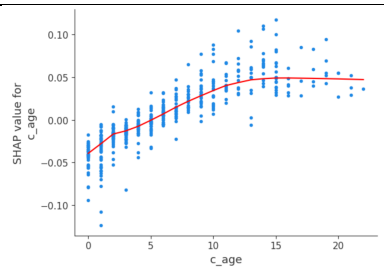
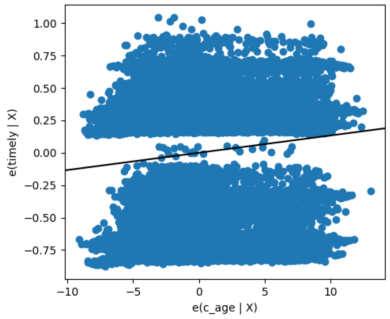
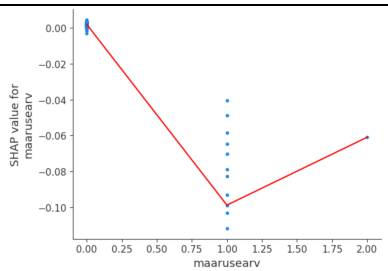
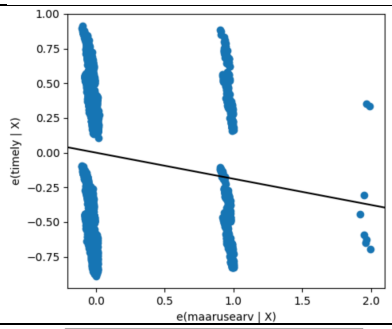
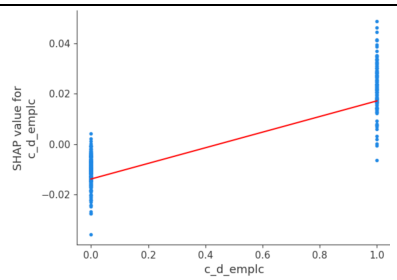
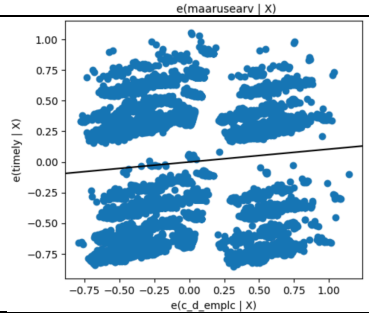
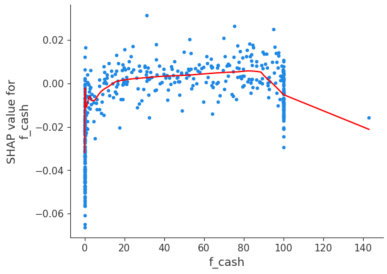
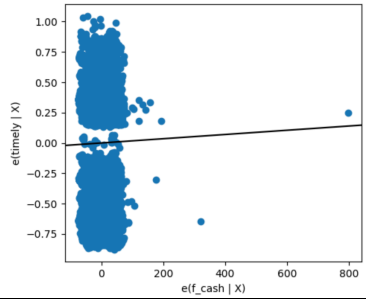
Source: Author's calculation

Appendix 10. Timely submission indicator distribution by industry among micro companies

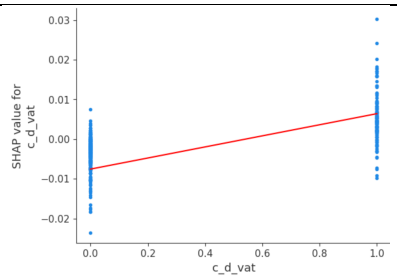
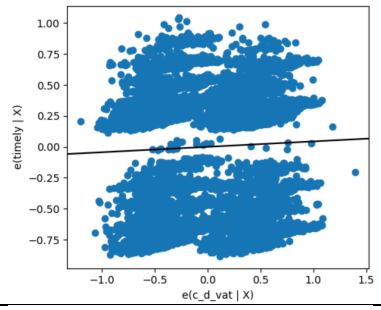
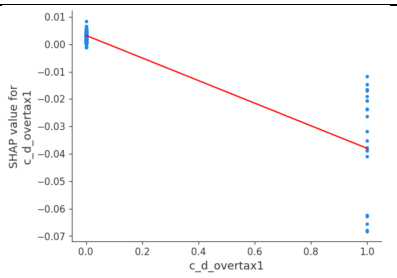
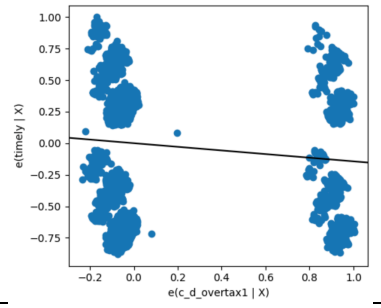
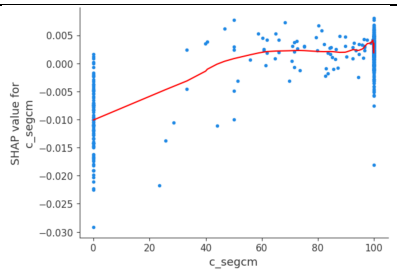
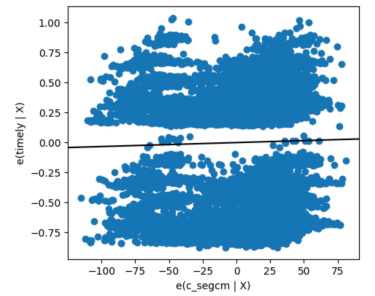
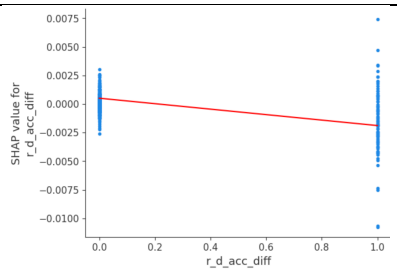
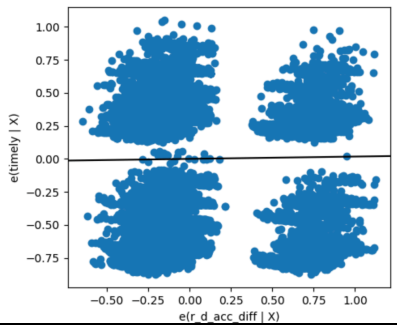


Source: Author's calculation

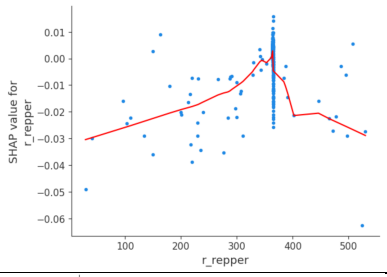
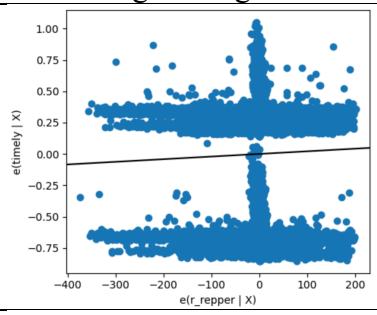
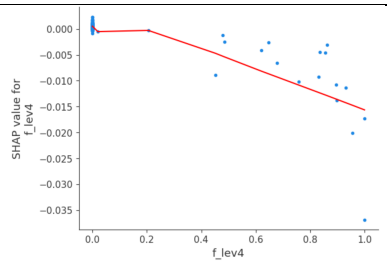
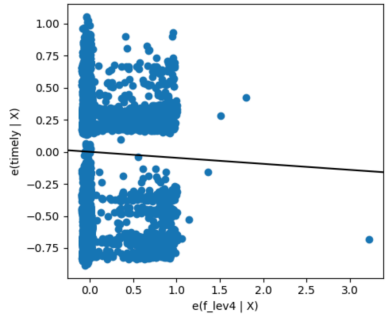
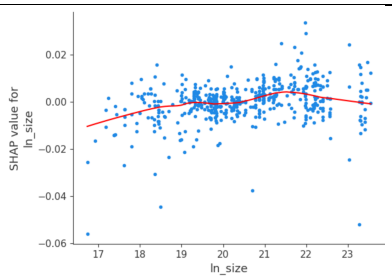
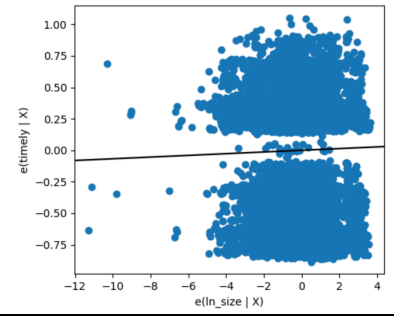
Appendix 11. Random forest and logistic regression partial dependence plots per variable (Micro, 2010 window)

Variable	Random forest	Logistic regression
c_d_late1		
c_age		
maarusearv		
c_d_emplc		
f_cash		

Appendix 11. Random forest and logistic regression partial dependence plots per variable (Micro, 2010 window) (continuation)

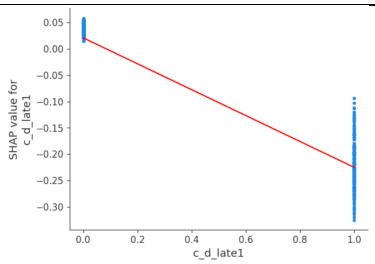
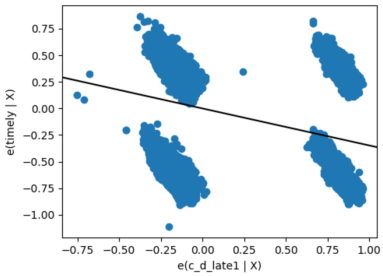
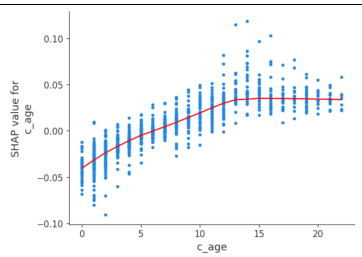
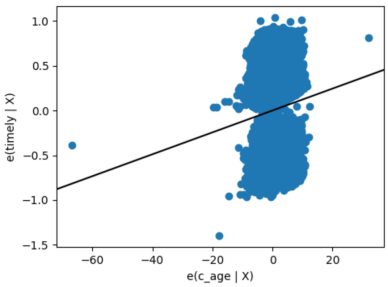
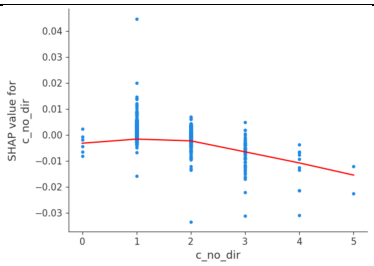
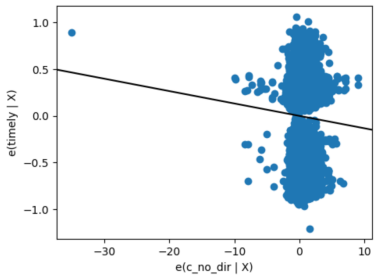
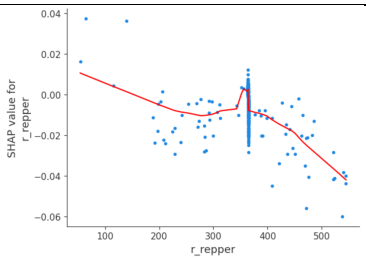
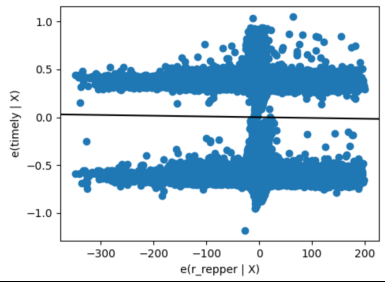
Variable	Random forest	Logistic regression
c_d_vat		
c_d_overtax1		
c_segcm		
r_d_acc_diff		

Appendix 11. Random forest and logistic regression partial dependence plots per variable (Micro, 2010 window) (continuation)

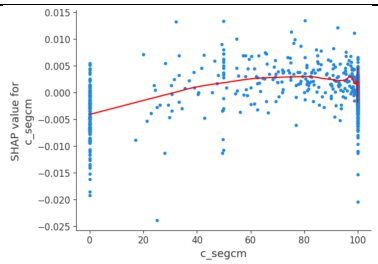
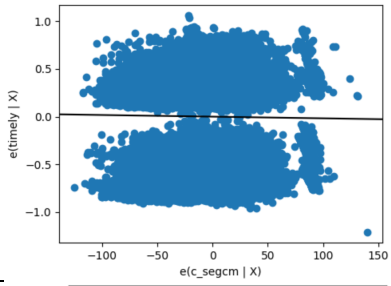
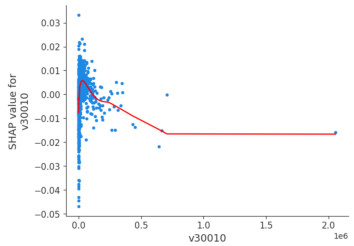
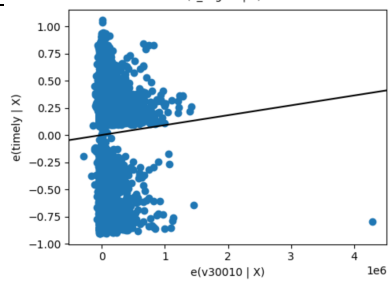
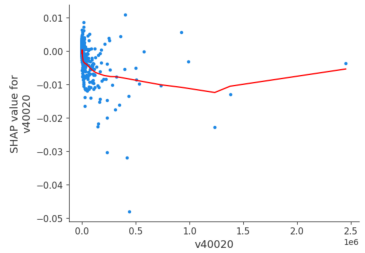
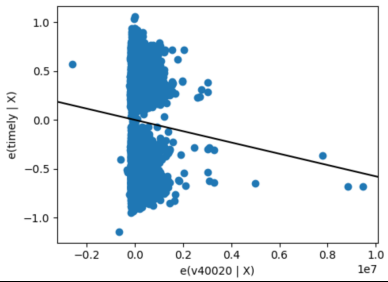
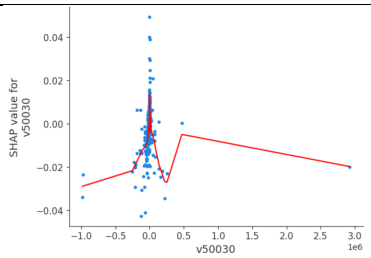
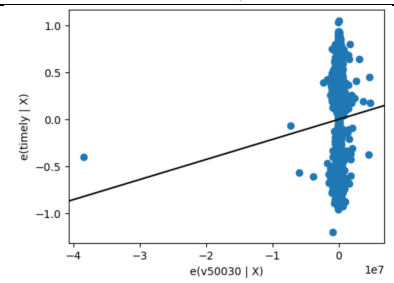
Variable	Random forest	Logistic regression
r_repper		
f_lev4		
f_lnsize2		

Source: Author's calculation

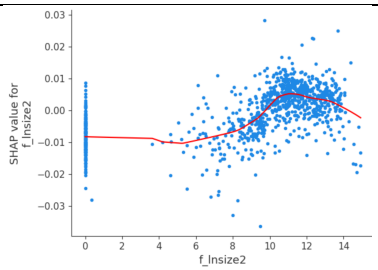
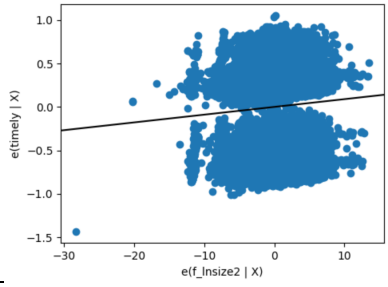
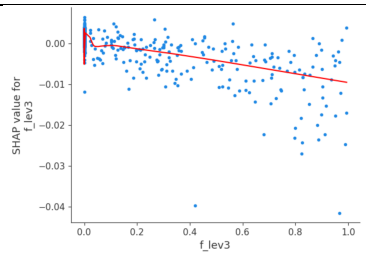
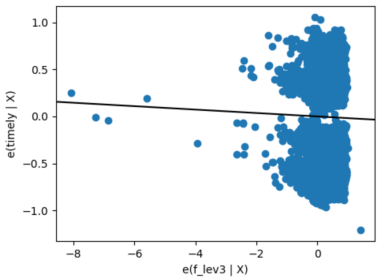
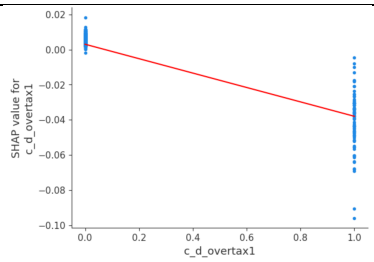
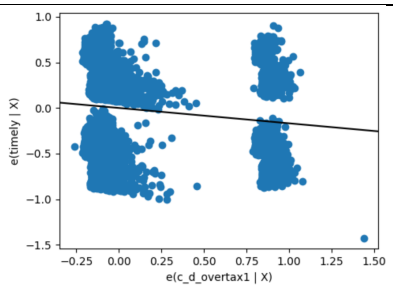
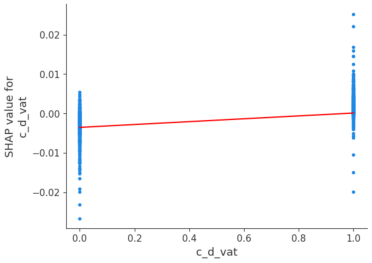
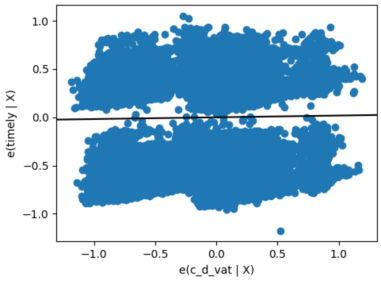
Appendix 12. Random forest and logistic regression partial dependence plots per variable (Small and medium, 2010 window)

Variable	Random forest	Logistic regression
c_d_late1		
c_age		
c_no_dir		
r_repper		

Appendix 12. Random forest and logistic regression partial dependence plots per variable (Small and medium, 2010 window) (continuation)

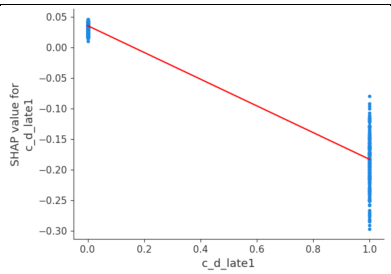
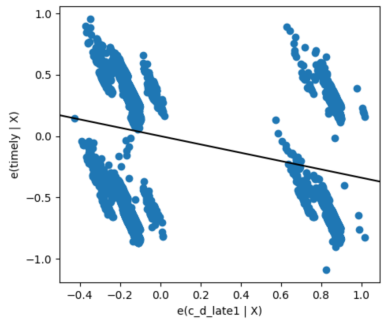
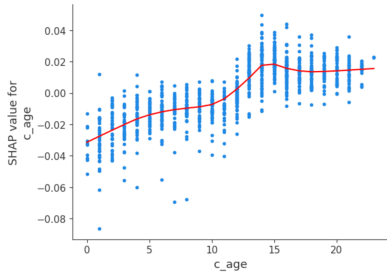
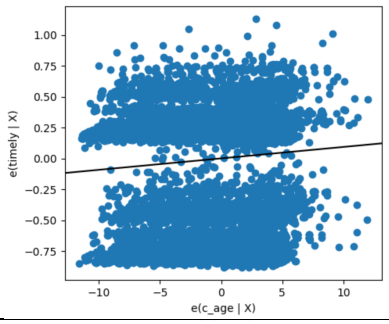
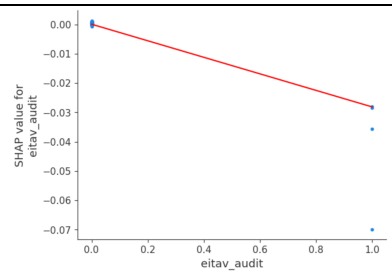
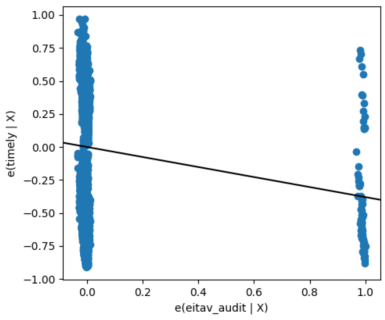
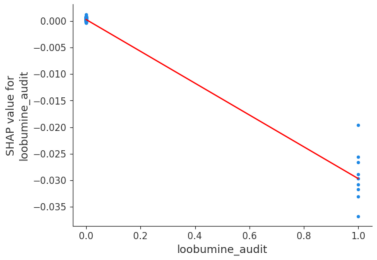
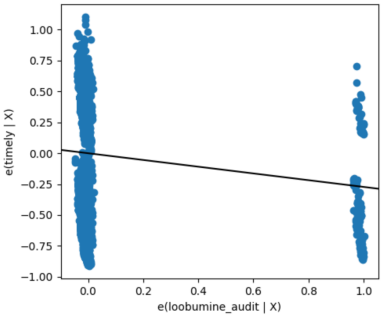
Variable	Random forest	Logistic regression
c_segcm		
v30010		
v40020		
v50030		

Appendix 12. Random forest and logistic regression partial dependence plots per variable (Small and medium, 2010 window) (continuation)

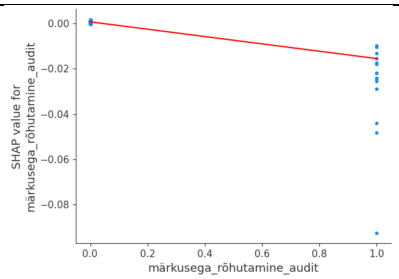
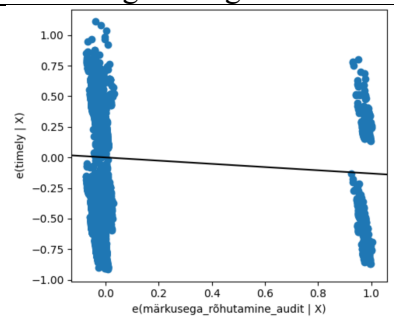
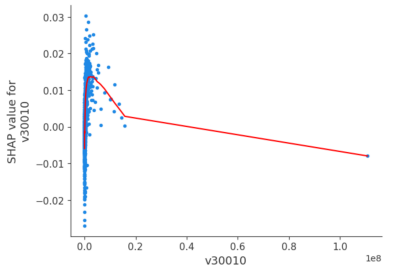
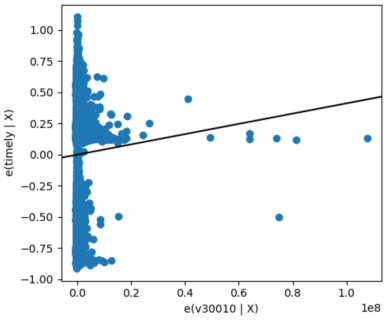
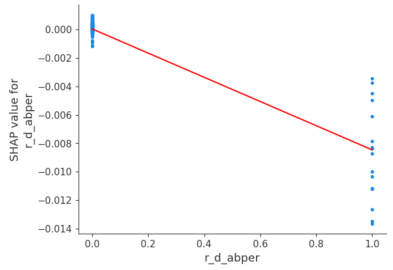
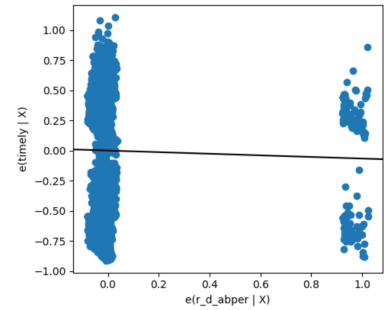
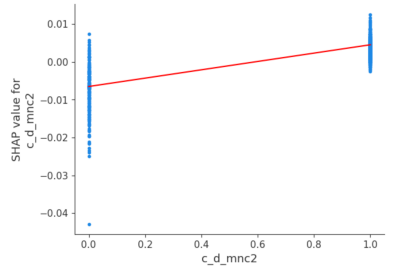
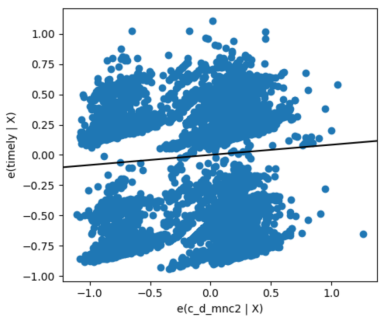
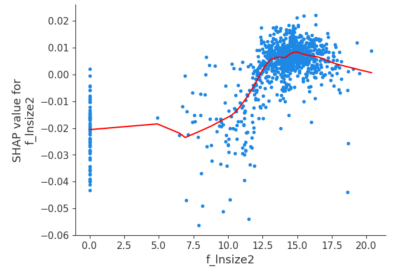
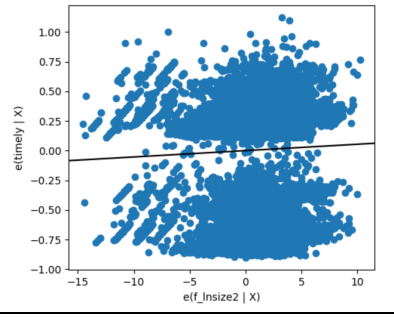
Variable	Random forest	Logistic regression
f_lnsize2		
f_lev3		
c_d_overtax1		
c_d_vat		

Source: Author's calculation

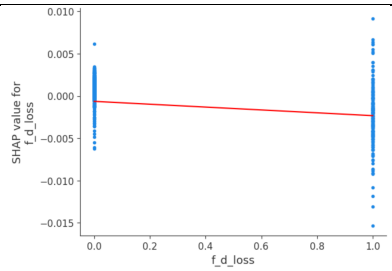
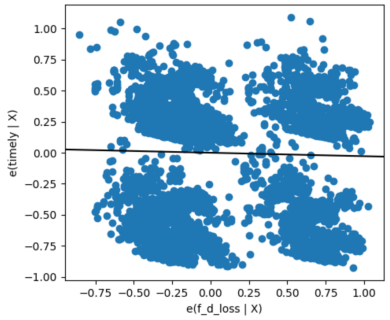
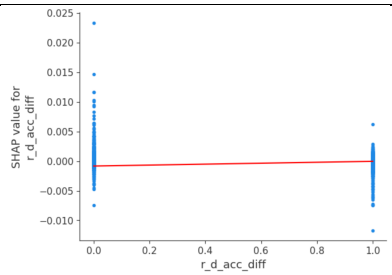
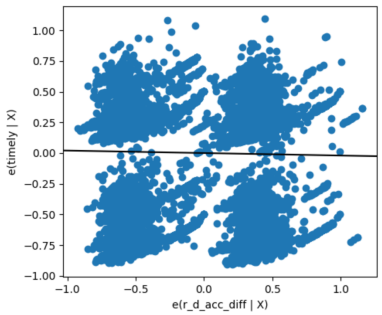
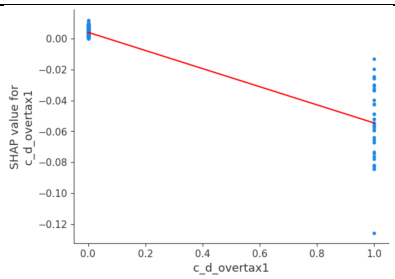
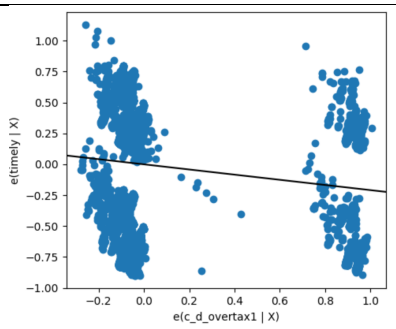
Appendix 13. Random forest and logistic regression partial dependence plots per variable (Audited, 2010 window)

Variable	Random forest	Logistic regression
c_d_late1		
c_age		
eitav_audit		
loobumine_audit		

Appendix 13. Random forest and logistic regression partial dependence plots per variable (Audited, 2010 window) (continuation)

Variable	Random forest	Logistic regression
märkusega_rõhutamine_audit		
v30010		
r_d_abper		
c_d_mnc2		
f_lnsize2		

Appendix 13. Random forest and logistic regression partial dependence plots per variable (Audited, 2010 window) (continuation)

Variable	Random forest	Logistic regression
f_d_loss		
r_d_acc_diff		
c_d_overtax1		

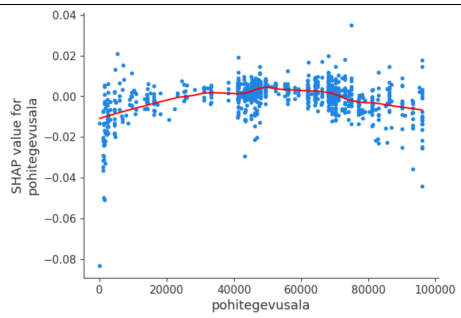
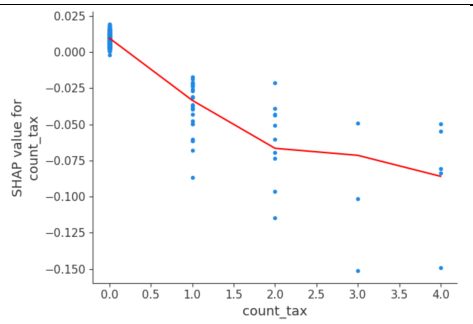
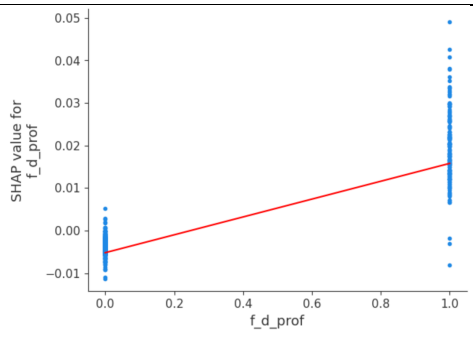
Source: Author's calculation

Appendix 14. Marginal affects for logistic regression model (2010 window)

Variable	Micro	Small and medium	Audited
c_age	0.0131*** (0.001)	0.0118*** (<0.000)	0.0089*** (0.001)
c_d_late1	-0.3074*** (0.005)	-0.3073*** (0.003)	-0.2765*** (0.008)
maarusearv	-0.1862*** (0.014)	-	-
c_d_emplc	0.0975 *** (0.005)	-	-
f_cash	0.0003*** (<0.000)	-	-
c_d_vat	0.0226*** (0.005)	0.0167*** (0.004)	-
c_d_overtax1	-0.1448*** (0.008)	-0.1585*** (0.004)	-0.1787*** (0.011)
c_segcm	0.0005*** (< 0.000)	-0.0002*** (< 0.000)	-
r_d_acc_diff	0.0260*** (0.006)	-	-0.0201*** (0.007)
r_d_abper	-0.0956 (0.009)	-0.067 (0.007)	-0.0544*** (0.023)
end_quarter	-	0.0442 *** (0.005)	-
f_lev4	-0.0587*** (0.014)	-	-
f_lev3	-	-0.0194*** (0.006)	-
f_lnsize2	0.0053***(0.001)	0.0088*** (0.001)	0.0310*** (0.005)
v30010	-	1.16e-07*** (3.11e-08)	1.27e-08*** (4.56e-09)
v40020	-	-5.58e-07*** (1.37e-08)	-
v50030	-	4.03e-08*** (1.92e-08)	-
c_no_dir	-	-0.0140 *** (0.002)	-
r_d_ifrs	-	-	0.1070*** (0.026)
eitav_audit	-	-	-0.3496*** (0.052)
loobumine_audit	-	-	-0.2296*** (0.034)
märkusega_rõhutamine_audit	-	-	-0.112***(0.024)
c_d_mnc2	-	-	0.0786*** (0.008)
f_d_loss	-	-	-0.0273*** (0.001)

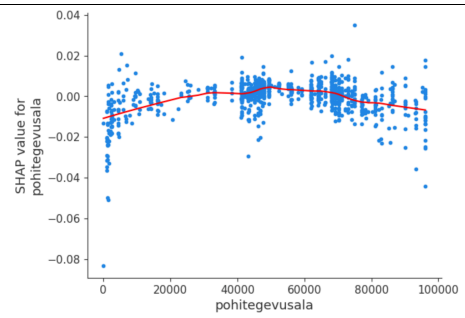
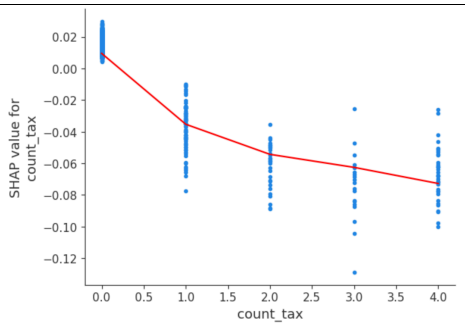
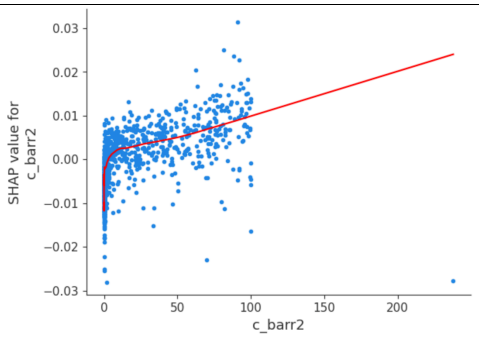
Source: Author's calculation

Appendix 15. Additional relationships found with random forest for micro entities in 2010 window

Variable	SHAP partial dependency	Relationship	Related findings
industry		?	
count_tax		-	- (Laidroo et al., 2020)
f_d_prof		+	

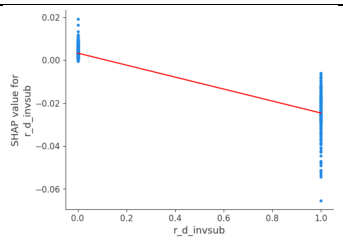
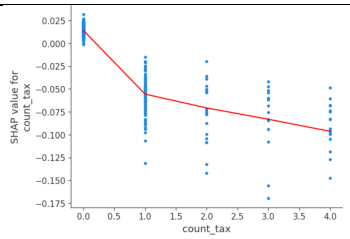
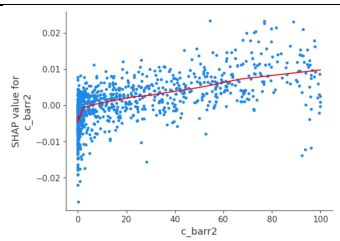
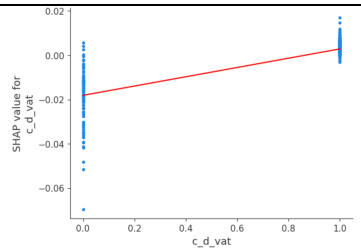
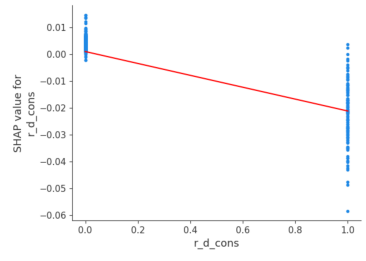
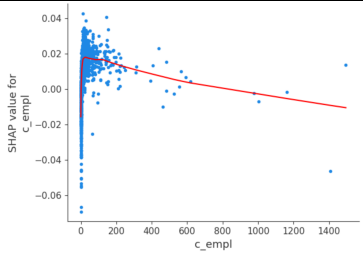
Source: Author's calculation

Appendix 16. Additional relationships found with random forest for small and medium entities in 2010 window

Variable	SHAP partial dependency	Relationship	Related findings
industry		?	
count_tax		-	- (Laidroo et al., 2020)
c_barr2		+	

Source: Author's calculation

Appendix 17. Additional relationships found with random forest for audited entities in 2010 window

Variable	SHAP partial dependency	Relationship	Related findings
r_d_invsub		-	
count_tax		-	- (Laidroo et al., 2020)
c_barr2		+	
c_d_vat		+	+ (Laidroo et al., 2020)
r_d_cons		-	
c_empl		-	

Appendix 18. Logistic regression variable significance persistence over the years for micro entities' sample

Variable Name	2010	2011	2012	2013	2014	2015	2016
const	***	***	***	***	***	***	***
c_age	***	***	***	***	***	***	***
c_d_late1	***	***	***	***	***	***	***
maarusearv	***	***	***	***	***	***	***
c_d_emplc	***	***	***	***	***	***	***
f_cash	***	***	-	-	***	***	***
c_d_vat	***	***	***	***	***	***	***
c_d_overtax1	***	***	***	***	***	***	***
c_segcm	***	***	***	***	***	***	***
r_d_acc_diff	***	***	-	*	***	***	***
r_d_abper	***	***	***	***	***	***	***
f_lev3	***	*	-	-	-	**	***
f_lnsize2	***	-	***	***	***	***	***

Source: author's calculations

*** 0.01 statistical significance, ** 0.05 statistical significance, * 0.1 statistical signficance

Appendix 19. Logistic regression variable significance persistence over the years for small and medium entities' sample

Variable Name	2010	2011	2012	2013	2014	2015	2016
const	***	***	***	***	***	***	**
c_age	***	***	***	***	***	***	***
c_d_late1	***	***	***	***	***	***	***
c_no_dir	***	***	***	***	***	***	***
r_d_abper	***	***	***	***	***	***	***
c_segcm	***	***	***	***	***	***	***
v30010	***	**	***	***	***	***	***
v40020	***	***	***	***	-	-	**
v50030	**	***	***	**	**	-	*
f_lnsize2	***	***	***	***	***	***	***
f_lev3	***	-	-	-	-	-	-
end_quarter	***	***	***	***	***	***	***
c_d_overtax1	***	***	***	***	***	***	***
c_d_vat	***	***	***	***	***	***	***

Source: author's calculations

*** 0.01 statistical significance, ** 0.05 statistical significance, * 0.1 statistical signficance

Appendix 20. Logistic regression variable significance persistence over the years for audited entities' sample

Variable Name	2010	2011	2012	2013	2014	2015	2016
const	***	***	***	***	***	***	***
c_age	***	***	***	***	***	***	***
c_d_late1	***	***	***	***	***	***	***
r_d_ifrs	***	***	-	-	-	-	***
eitav_audit	***	***	**	**	***	***	**
loobumine_audit	***	***	***	***	***	***	***
märkusega_rõhutamine_audit	***	***	***	***	***	***	-
v30010	***	-	-	-	*	-	-
r_d_abper	**	***	***	***	***	***	***
c_d_mnc2	***	***	***	***	***	***	***
f_lnsize2	***	***	***	***	-	***	***
f_d_loss	***	***	***	***	*	*	***
r_d_acc_diff	***	***	*	**	***	**	***
c_d_overtax1	***	***	***	***	***	***	***

Source: author's calculations

*** 0.01 statistical significance, ** 0.05 statistical significance, * 0.1 statistical signficance

Appendix 21. Non-exclusive licence

A non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Artur Luik (*author's name*)

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis Determinants of Annual Report Submission Timeliness in Estonia (*title of the graduation thesis*)

supervised by Laivi Laidroo, (*supervisor's name*)

1.1 to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

07.05.2024 (date)

[1] *The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period*