Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Evaluating the Credit Risk of SMEs Using Artificial Intelligence, Financial and Alternative Data

Master's Final Degree Project

Laura Miliūnaitė

Project author

Prof. Dr. Gerda Žigienė

Supervisor

Dr. Paulius Danėnas

Supervisor

Kaunas, 2023

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Evaluating the Credit Risk of SMEs Using Artificial Intelligence, Financial and Alternative Data

Master's Final Degree Project

Business Big Data Analytics (6213AX001)

Laura Miliūnaitė

Project author

Prof. Dr. Gerda Žigienė
Supervisor

Dr. Paulius Danėnas
Supervisor

Prof. Dr. Vytautas Snieška
Reviewer

Prof. Dr. Evaldas Vaičiukynas
Reviewer

Kaunas, 2023

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Laura Miliūnaitė

Evaluating the Credit Risk of SMEs Using Artificial Intelligence, Financial and Alternative Data

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Laura Miliūnaitė

Confirmed electronically

Miliūnaitė, Laura. Evaluating the Credit Risk of SMEs Using Artificial Intelligence, Financial and Alternative Data. Master's Final Degree Project / supervisors Prof. Dr. Gerda Žigienė and Dr. Paulius Danėnas; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics (Mathematical Sciences)

Keywords: *credit risk, small and medium-sized enterprises, artificial intelligence, alternative data.*

Kaunas, 2023. 69 pages.

Summary

Small and medium-sized enterprises (SMEs) are of major importance in world economies and job creation. Financing is one of the key issues for SME development since SMEs are often considered riskier than large companies. It is argued in the literature that artificial intelligence (AI) and alternative data could increase the financial inclusion of disadvantaged groups, such as SMEs. Thus, this study aimed to compare SMEs' credit risk prediction models incorporating alternative data with models using only traditional financial data. The dataset used in the study involved Lithuanian SMEs' observations from the 2015-2020 period and included traditional financial data as well as alternative data such as general characteristics of the company, macroeconomic indicators and payment behaviour data. Five different AI methods were employed in the model development process. The results showed that including alternative data in credit risk prediction models can increase the prediction performance of the models compared to models that use only financial data. Variable importance analysis revealed that payment behaviour data had the most significant impact of all alternative data-based variables.

Miliūnaitė, Laura. Mažų ir vidutinių įmonių kredito rizikos vertinimas naudojant dirbtinį intelektą, finansinius ir alternatyvius duomenis. Magistro studijų baigiamasis projektas / vadovai Prof. Dr. Gerda Žigienė ir Dr. Paulius Danėnas; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (Matematikos mokslai)

Reikšminiai žodžiai: *kredito rizika, mažos ir vidutinės įmonės, dirbtinis intelektas, alternatyvūs duomenys.*

Kaunas, 2023. 69 p.

Santrauka

Mažos ir vidutinės įmonės atlieka svarbų vaidmenį pasaulio ekonomikoje ir kuriant darbo vietas. Finansavimas yra viena iš pagrindinių mažų ir vidutinių įmonių plėtros problemų, kadangi mažos ir vidutinės įmonės dažnai laikomos rizikingesnėmis nei didelės įmonės. Literatūroje teigiama, kad dirbtinis intelektas ir alternatyvūs duomenys galėtų padidinti nepalankioje padėtyje esančių grupių, tokių kaip mažos ir vidutinės įmonės, galimybes gauti finansavimą. Todėl šiame tyrime siekiama palyginti mažų ir vidutinių įmonių kredito rizikos prognozavimo modelius, kuriuose naudojami alternatyvūs duomenys, su modeliais, naudojančiais tik tradicinius finansinius duomenis. Tyrime naudojami duomenys apėmė Lietuvos mažų ir vidutinių įmonių stebėjimus per 2015–2020 m. laikotarpį ir tradicinius finansinius bei alternatyvius duomenis, tokius kaip bendros įmonės charakteristikos, makroekonominiai rodikliai ir mokėjimo elgsenos duomenys. Kuriant modelį buvo išbandyti penki skirtingi dirbtinio intelekto modeliai. Rezultatai parodė, kad alternatyvių duomenų įtraukimas į kredito rizikos prognozavimo modelius gali padidinti modelių prognozavimo tikslumą, lyginant su modeliais, kuriuose naudojami tik finansiniai duomenys. Kintamųjų svarbos analizė atskleidė, kad mokėjimo elgsenos duomenys turėjo didžiausią įtaką iš visų alternatyviais duomenimis pagrįstų kintamųjų.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviations:

€ – euro;

AI – artificial intelligence;

AUC – area under the curve;

EBIT – earnings before interest and taxes;

EBITDA – earnings before interest, taxes, depreciation and amortisation;

FPR – false positive rate;

GDP – gross domestic product;

KNN – k-nearest neighbours;

LightGBM – light gradient boosting machine;

MDA – multivariate discriminant analysis;

MSE – mean squared error;

NACE – statistical classification of economic activities in the European Community;

PR – precision-recall curve;

Prof. – professor;

RBF – radial basis function SVM kernel;

ROA – return on assets;

ROC – receiver operating characteristic;

ROE – return on equity;

SMEs – small and medium-sized enterprises;

SMOTE – synthetic minority over-sampling technique;

SVM – support vector machines;

TPR – true positive rate;

XGBoost – extreme gradient boosting.

Introduction

Small and medium-sized enterprises (SMEs) are an essential part of global economies. More than 95% of all businesses worldwide are SMEs; they provide more than half of all job places and generate more than half of GDP in developed economies [1]. In addition, SMEs create the majority of new jobs and contribute to innovation. Despite the significance of SMEs for the global economy, financing is one of the biggest problems in the development of SMEs. As a result, SMEs lack funds to establish, sustain and grow the business and it is estimated that the total credit gap of formal SMEs might amount to approximately one trillion US dollars [1].

SMEs' financing issues are closely related to the perceived risk associated with smaller companies. SMEs are often considered riskier than large companies because they have fewer assets and their financial data are not deemed as reliable as the financial data of large companies. Since traditional credit risk assessment methods usually are based on financial statement data, this leads to information asymmetry between creditors and companies. Consequently, creditors make higher demands or refuse to provide financing at all.

Studies show that including alternative data in credit risk prediction models can improve the model's performance compared to traditional models that use only financial statement data. Meanwhile, better prediction performance could increase access to credit for SMEs. Traditionally, statistical models such as discriminant analysis or logistic regression have been used to predict credit risk. However, in recent years Artificial Intelligence (AI) methods started to gain popularity as an alternative to classical statistical methods. AI algorithms are superior to statistical approaches because they do not require satisfying assumptions and can adapt to different types of data, which is especially relevant when the model includes alternative data.

Aim of the project. To compare SMEs credit risk prediction models incorporating alternative data with models using only financial data.

Objectives of the project:

1. To overview the traditional statistical and AI based methods and data types used for credit risk prediction presented in the scientific literature.
2. To choose the most suitable methods for the research based on the literature review.
3. To prepare the dataset of Lithuanian companies for model development.
4. To implement the selected algorithms and compare their performance when the dataset includes alternative data with the dataset including only financial data, provide conclusions and recommendations.

The results of this thesis were presented at the 17th Prof. Vladas Gronskas International Scientific Conference and an article was submitted to the journal of the conference. In addition, part of the research of this thesis was conducted during the Student Summer Research project *("Studentų vasaros moksliniai tyrimai")* organised by the Research Council of Lithuania and has received funding based on the agreement No. P-SV-22-153.

1. Literature Overview

1.1. Small and Medium-Sized Enterprises Financing Issues

The credit risk assessment topic has been widely discussed in scientific literature since the 1960s. Credit risk prediction is an essential part of the credit risk process for financial institutions, as they must assess whether the potential borrower can fulfil its obligations. Historically, credit risk was evaluated using the expert judgement of employees of a financial institution. However, with the increasing number of applicants and competitors emerged a need for more efficient methods [2]. Additionally, since 1998 credit risk management of the financial institutions started to be coordinated internationally by Basel Accords [3]. According to Basel II, financial institutions must have a credit risk assessment mechanism for 12 months and the models must be transparent and interpretable [4]. The financial institution can employ a standardised approach, relying on external credit ratings, or develop its own risk assessment models [3]. However, external credit ratings provided by independent credit ratings agencies such as Fitch Ratings, Moody‘s and S&P usually are available only for large companies [5, 6]. In addition, credit risk prediction is vital not only for financial institutions but also for companies in other industries, which must decide whether to give trade credit to their clients and on which conditions. Although there are significantly more SMEs than large businesses worldwide, most research in the credit risk prediction area has focused on large companies [7].

According to the European Commission, SMEs are companies with less than 250 employees and annual revenue under €50 million or total assets below €43 million [8]. The literature describes several distinctions between small and large businesses. For example, R. R. Pettit and R. F. Singer [9] analysed differences between large and small companies and found that small companies have fewer assets, more frequently choose short-term debt, internal funds or shareholder loans as financing sources and less often use external equity. SMEs are often considered riskier than large companies and one of the main reasons lies in the quality and availability of financial data [7]. SMEs' financial data could be regarded as less reliable than large companies because it is usually not audited by external auditors. It is relatively expensive for small companies to hire external auditors and often financial statements of small companies are not legally required to be audited [6, 9]. In addition, frequently, small companies are even not obliged to disclose detailed financial statements [7]. The lack of qualitative financial data creates an information asymmetry between creditors and SMEs which seek financing, resulting in higher demands from creditors to hedge against potential risks [6, 9, 10]. For instance, the creditors might charge higher interest rates to compensate for higher risk, include additional covenants in the loan agreement, require guarantees from other parties or more collateral [9]. In practice, such requirements might be challenging for many SMEs due to limited available resources.

As smaller companies have a different risk profile and specifics than larger companies, credit risk prediction models created with the data of large companies might not be suitable to evaluate the credit risk of SMEs [7, 11]. In addition, usually, SMEs' financial figures are significantly smaller compared to large companies. Thus, even small changes in SMEs' financial position could substantially impact their financial ratios compared to large companies, which are much more stable financially [7]. Therefore, the model developed with large companies' data might lack predictive ability when applied on smaller company. Consequently, companies sometimes fall into a vicious circle when they cannot access financing because creditors do not have sufficient financial data about SMEs, and the company cannot afford to cover higher borrowing costs [12, 13]. The issue of SMEs’ access to financing falls under the broader topic of financial inclusion. Financial inclusion generally refers to the ease of access and availability of traditional financing to all people and businesses, especially those in disadvantaged groups [10]. The topic has become increasingly important in the global society in recent years. For example, financial inclusion is a part of United Nations Sustainable Development Goals, namely, the 8th goal concerning economic growth and employment has targets such as increasing access to banking and financial services for everybody and promoting SMEs development through access to finance[1]. It is argued in the literature that AI and alternative data could increase the financial inclusion of disadvantaged groups [10, 13]. The possibility to include alternative data is particularly relevant in the case of SMEs, which usually lack the financial information needed for credit risk evaluation. Alternative data sources in credit risk assessment models could help solve problems such as information asymmetry, adverse selection, and moral hazard and allow increased access to finance for previously excluded entities [10]. From an economic perspective, this could improve the financial inclusion of frequently credit-invisible SMEs and stimulate economic growth in related areas [14].

1.2. Definition of Credit Risk and Its Elements

Basel Committee on Banking Supervision defined credit risk as “the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms” [1 p. 15]. The elements of credit risk include default probability, exposure at default and loss given default [16]. Most credit risk prediction research focuses on estimating the probability of default, which is essential in preventing credit losses. Even though credit risk prediction has been extensively explored in the scientific literature for several decades, there is no clear consensus on the default definition. Many studies focus on bankruptcy prediction [17, 18, 19, 20, 21, 22, 23, 24], which is considered as an extreme case of default when the entity is declared bankrupt by a court and is dissolved [25]. Others use terms such as failure or failing company [4, 26, 27, 28, 29], referring to the inability to fulfil financial obligations, which do not necessarily end up as bankruptcy. The lack of agreement on default definition might be attributed to data availability since researchers tend to define default depending on the available data [30]. Basel Committee on Banking Supervision in Basel II accord states that default happens when either the debtor is unlikely to fully fulfil all its obligations (for example, in case of a bankruptcy), the debtor has missed a payment for more than 90 days or both conditions have occurred [31]. Exposure at default refers to the outstanding amount that could be lost at the time of default [16]. While loss given default is the loss incurred at the event of default after collection efforts (for example, after collecting funds from the guarantor or sale of collateral) [16].

Various internal and external reasons could trigger SMEs' default. For example, a default may arise due to the acts of other firms, suppliers and clients [30]. In addition, the lack of knowledge, experience and personal skills of managers of the companies may also lead to the default of the company [30]. J. Ropega [30] argues, that SMEs are more vulnerable to the risk of default than large companies due to smaller internal financial resources and limited access to bank financing.

Several situations are described in the literature which might increase the risk of SMEs default. Examples include a lack of control mechanisms and operational inefficiencies. Failure to establish internal controls leads to such problems as runaway costs, uncompetitive prices, slow turnover and poor liquidity [30, 32]. Another risk factor is related to expansion when management has an overly optimistic view of the company’s growth without clear goals and sufficient analysis. This situation

[1] https://sdgs.un.org/goals/goal8

might lead to increasing financial leverage due to growing capital expenditure, liquidity and solvency problems due to too high costs, idle production capacity and excess inventory [30, 32]. Furthermore, management lacking motivation, a forward-looking approach and failure to react to the changing environment could also increase the default risk for the SMEs [30]. Finally, the owner of the entity, which use the funds of the company for his or her personal needs, might also create risks [30]. According to P. du Jardin [33], most default causes are foreseeable and can be observed in financial accounts. Nevertheless, the default could not be predicted by relying only on financial indicators and other sources of information must also be considered.

Various financial and non-financial symptoms are described in the literature, which might indicate the declining financial position of the company. For example, J. Ropega [30] overviewed multiple studies and listed such financial symptoms as a decrease in revenue, profit, liquidity, market share, rising operating expenses, high level of debt and significant overcapacity as the factors which could signal a deteriorating financial situation. While non-financial symptoms could include issues with internal controls, distribution, productivity and production quality, absence of business plan, insufficient trust in partners and insufficient professional growth of employees [30].

1.3. Overview of Methods Used in Credit Risk Prediction

1.3.1. Classical Statistical Methods

1.3.1.1. Discriminant Analysis

One of the first studies in credit risk prediction was conducted by W. H. Beaver in 1966 [26], where he performed a univariate discriminant analysis of 158 companies. However, the analysis was very limited as it assessed only one financial ratio at a time. A couple of years later, in 1968, E. I. Altman used multivariate discriminant analysis (MDA) to predict the credit risk of the companies [17]. During later years MDA gained popularity in credit risk prediction studies [18], for example, studies that applied MDA include research works performed by E. B. Deakin [34], P. P. M. Pompe and J. Bilderbeek [35] and M. Blum [36]. However, MDA had several drawbacks; namely, it required satisfying assumptions of linearity, homogeneous variances of both defaulted and non-defaulted firms, predictors had to be normally distributed, the score provided by the MDA had no explicit meaning and the method required to "match" defaulted and non-defaulted companies based on arbitrary criteria [18, 22, 34, 37]. With actual financial data, all these assumptions were difficult to satisfy.

1.3.1.2. Logistic Regression

Due to the limitations of MDA, other methods were suggested in later studies. In 1970 Y. E. Orgler [38] used multivariate regression analysis, where dependent variable values were bounded to be between zero and one, as an alternative to MDA. The multivariate regression analysis was stated to be more accessible, prominent, and provide more statistical data. The model predicted 80% of the bad loans correctly.

In the 1980s, linear regression was replaced by logistic regression. J. A. Ohlson [18] argued that logistic regression avoids all the shortcomings of MDA. As a result, logistic regression became one of the most widely used methods for credit risk prediction in the coming years. For example, studies which applied logistic regression include but are not limited to research done by G. Dong et al. [39], N. H. Wellalage and S. Locke [23] and C. Yin et al. [40]. Logistic regression is superior to other traditional statistical models, such as linear regression or MDA, because it does not require satisfying as many assumptions. Namely, there are no assumptions related to the normality, linearity and homoscedasticity of independent variables [23]. In addition, the logistic regression is appreciated due to its simplicity and interpretability [39, 41, 42]. The latter properties have led to the widespread use of logistic regression in the banking industry [39]. Nevertheless, logistic regression requires fulfilling certain assumptions, such as a sample should be random and large, no collinearity between explanatory variables and observations should be independent [43].

1.3.2. Artificial Intelligence Methods

Limitations of classical statistical methods stimulated research of other credit risk prediction approaches, including but not limited to artificial intelligence (AI). One of the benefits of AI is that it does not require as much data processing and satisfying assumptions as traditional statistical methods and allows faster processing of potential debtors' applications without compromising the quality of processing [14]. For example, M. Fuster et al. [44] analysed mortgage application data in the United States and found that companies using financial technologies, including AI, process applications approximately 20% faster than other financial services providers without a significant difference in the risk assumed. In addition, AI can include alternative data, which could potentially yield more accurate scores. H. Sadok et al. [14] overviewed studies on the topic of credit risk analysis using AI and concluded that AI offers only minor performance improvements when applied to the same data sets as classical statistical methods, but the performance improves when the model includes other types or large volumes of data, which could not be used in classical methods.

Most AI algorithms are criticised for lack of opacity and interpretability [14, 41]. However, this issue could be at least partially overcome by applying specific tools. For example, the most recent studies address the issue of AI opacity by applying post-prediction explanation models such as the Shapley additive explanation (SHAP) [4].

Various AI methods are found in the credit risk prediction literature, such as neural networks, support vector machines (SVM), decision trees and ensemble methods, for instance, random forest or gradient boosting. It should be noted that this list is not finite and includes more frequently mentioned methods in the reviewed literature. The following chapters further discuss the most often cited algorithms.

1.3.2.1. Neural Networks

In the 1990s, AI gained popularity in credit risk prediction studies as an alternative to traditional statistical techniques. One of the first studies was published in 1990 by M. D. Odom and R. Sharda [19], where authors aimed to apply the neural networks with the same financial ratios as in E. I. Altman's study dated 1968 and compare it with the results of MDA. M. D. Odom and R. Sharda stated that the neural networks are better than MDA as it does not require the normality of variables [19]. The authors used a dataset from the period between 1975 and 1982 with 65 defaulted and 64 non-defaulted companies. The results showed that the neural networks were better at predicting defaulted companies than MDA. Results of studies comparing neural networks with other methods also showed that the neural networks could achieve equally well or higher accuracy. For example, K. Y. Tam and M. Y. Kiang [27] compared the performance of neural networks with a linear discriminant model, logistic regression, k-nearest neighbours (KNN) and decision tree (ID3 algorithm) with the dataset containing half defaulted and half non-defaulted US banks. The authors found that neural networks outperformed other models in prediction accuracy, adaptability and robustness. Another study conducted by T. B. Bell [28] applied neural networks and logistic regression to a sample of US banks. The findings demonstrated that both neural networks and logistic regression performed equally well. C. Charalambous et al. [21] compared four neural network algorithms (backpropagation, learning vector quantization, radial basis function and feedforward network) with logistic regression. The methods were applied to a dataset of US defaulted and non-defaulted firms matched on industry, size and the year of default. It was found that learning vector quantization, radial basis function and feedforward network algorithms performed better than logistic regression and backpropagation algorithms.

The neural networks are superior to other methods due to its non-linear, non-parametric and adaptive learning features [37]. In addition, neural networks do not require making any assumptions about parameter distributions beforehand since it determines the relationship between dependent and independent variables itself based on the given data [22]. However, the performance of neural networks depends on the size of the dataset. For example, Zhang et al. [20] conducted a study where they used financial ratios to compare neural networks performance with logistic regression performance in the credit risk prediction of 220 companies. They found that the neural networks outperformed logistic regression with smaller and larger datasets, but the difference in accuracy increased significantly when the authors increased the dataset size. In addition, the neural networks require expertise in data preprocessing to select control parameters, the model might be overfitted to the particular case making it difficult to make generalized conclusions and due to the opacity of the method, the possibilities of explaining the obtained results are very limited [37].

1.3.2.2. Support Vector Machines

Another widely applied method in credit risk prediction studies is support vector machines (SVM). Studies show that SVM is a promising algorithm in credit risk prediction, capable of outperforming more classical models. For example, Shin et al. [22] employed a dataset of 2,320 medium-size manufacturing firms in Korea (1,160 defaulted and 1,160 non-defaulted) to compare the performance of SVM and neural networks in credit risk prediction. The study's results revealed that the SVM outperformed the neural networks in accuracy and generalization performance, especially when the dataset is small. S. Lahmiri [45] compared SVM, neural networks (backpropagation algorithm and radial basis function), linear discriminant analysis and naïve Bayes to assess credit risk. The models were applied to the Australian credit approval dataset containing 690 records and 14 attributes, the German credit dataset with 1,000 records and 20 variables and the Japanese credit approval dataset containing 658 records and 15 attributes. The results showed that the SVM provided the best accuracy with all three datasets. The author concluded that the SVM appears very suitable for credit risk prediction tasks. However, when the performance results were evaluated in terms of other metrics, such as specificity and sensitivity, no single method was the best based on all metrics, so it is important to use multiple methods when assessing credit risk [45].

SVM has several benefits compared to other algorithms, such as neural networks. For example, the SVM does not require as much expertise in choosing control parameters since it has only two main free parameters (upper bound and kernel parameter) [22]. SVM gives a unique, optimal and global solution and does not require a large training set [22]. Conversely, due to its simple structure, SVM is unsuitable for dealing with high cardinality datasets. There are ways to circumvent this issue, such as reducing dimensionality or applying feature selection. However, applying these techniques results in losing part of the information [4]. In addition, it might be challenging to select appropriate parameters, it tends to be slow at a test stage and requires substantial memory reserves due to a highly complex algorithm [46].

1.3.2.3. Decision Trees

Another method used in credit risk prediction studies is decision trees. The decision trees are superior to other AI algorithms because it provides interpretable rules based on which classification was performed [46]. Also, decision trees can solve both regression and classification problems. However, the decision trees require large data samples and sufficient dependent variable cases [43, 46]. In addition, the model is prone to overfitting, might be susceptible to extreme outliers and in some cases, a minor modification in the data can result in a substantially different series of splits, making interpretation problematic [43, 47].

Examples of studies which have applied decision trees in credit risk prediction include studies by A. Ptak-Chmielewska [43], K. Y. Tam and M. Y. Kiang [27] and N. Gulsoy and S. Kulluk [5]. A. Ptak-Chmielewska [43] compared decision trees performance with SVM, gradient boosting, logistic regression and neural networks to predict credit risk of 806 small enterprises. The author concluded that other methods outperformed the decision trees, which appeared less stable and overfitted. K. Y. Tam and M. Y. Kiang [27] used neural networks, MDA, logistic regression, KNN and decision trees ID3 algorithm to predict the credit risk of financial institutions. The results showed that in certain situations, decision trees were better than KNN, however, compared to all other methods, decision trees have underperformed. While N. Gulsoy and S. Kulluk compared six different decision tree/rule-based algorithms (Multi Objective Evolutionary Fuzzy Classifier, Naïve Bayes Tree, PART, J48, Random Tree and Simple Cart) to evaluate SMEs' credit risk. Multi Objective Evolutionary Fuzzy Classifier and PART algorithms achieved the best performance results, while the worst performing was Random Tree.

1.3.2.4. Ensemble Methods

Another common method in credit risk prediction studies, random forest, is an ensemble learning model which constructs multiple decision trees [47]. With additional trees, random forest compensates for decision trees' propensity to overfit their training data and is less likely to change its output when small changes are made to the dataset [47]. Studies show that the random forest is a very effective AI technique [42]. However, random forest is a black box method, hence, the interpretation of obtained results is not straightforward.

Research performed by A. Malakauskas and A. Lakštutienė [48] and G. Yao, X. Hu, L. Xu and Z. Wu [49] support the statement that the random forest is an effective method in credit risk prediction. A. Malakauskas and A. Lakštutienė [48] applied logistic regression, neural networks and random forest to predict credit risk of 12,000 SMEs from Baltic countries. Results showed that random forest achieved superior results compared to logistic regression and neural networks based on the area under the curve (AUC) for Receiver Operating Characteristic (ROC) graph, AUC for the Detection Error Trade off graph and equal error rate. G. Yao, X. Hu, L. Xu and Z. Wu [49] employed SVM, random forest and extreme random tree for credit risk prediction in the supply chain performance of listed companies in China. The authors found that although the SVM achieved the highest accuracy (92.21%), the random forest had a higher precision score of 90.32%.

Another ensemble technique based on the decision trees is gradient boosting. Gradient boosting can handle large amounts of data with high dimensionality, has high predictive performance and is less vulnerable to outliers [4]. Nevertheless, like the random forest, gradient boosting is a black box method, with outputs that are impossible to define as a direct function of inputs, and without extra processing, gradient boosting can only be used for predictive but not descriptive purposes [4]. Gradient boosting has different implementations, such as Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (CatBoost) [4].

Empirical research examples include research performed by B. Lextrait [4], who applied three gradient boosting algorithms (LightGBM, XGBoost and CatBoost), SVM and logistic regression on the data set of French companies containing a wide range of several hundred financial variables from balance sheets and several simple alternative data-based variables. The research showed that gradient boosting outperforms SVM and logistic regression based on training time, evaluation metrics and economic accuracy. However, the author noticed that gradient boosting performs worse when the complexity of the data decreases due to overtraining. In another study A. Ptak-Chmielewska [43] compared decision trees, SVM, gradient boosting, logistic regression and neural networks performance and found that the gradient boosting outperformed all models except for SVM, but the model appeared to be overfitted. Another example is a study by Yin et al. [40], who compared the performance of logistic regression, random forest and XGBoost while predicting the credit risk of 1,091 Chinese SMEs, where 1,011 were non-defaulted and 80 defaulted. The results showed that based on the ROC AUC metric and the Kolmogorov-Smirnov statistic, the performance of all three models was similar, however, logistic regression slightly outperformed random forest and XGBoost.

The table below summarises the advantages and disadvantages of the overviewed credit risk prediction methods.

Table 1. Summary of advantages and disadvantages of overviewed methods

Method	**Advantages**	**Disadvantages**
MDA	Transparent and interpretable	Strict assumptions, the score provided by the MDA has no explicit meaning and requires to “match” defaulted and non-defaulted companies based on arbitrary criteria
Logistic regression	Transparent and interpretable, requires satisfying fewer assumptions compared to other statistical methods	Requires satisfying certain assumptions
Neural networks	Non-linear, non-parametric and adaptive learning features. Do not require making any assumptions about parameter distributions	Requires expertise in the parameter selection process, large data sets, prone to overfitting, opaque, uninterpretable
SVM	Does not require as much expertise in choosing control parameters as other methods, such as neural networks, gives a unique, optimal and global solution and does not require a large training set	Not suitable to deal with high cardinality datasets, it might be challenging to select appropriate parameters, it tends to be slow at a test stage and requires substantial memory reserves due to a highly complex algorithm
Decision trees	Provides interpretable rules based on which classification was performed	Requires large data samples, prone to overfitting, might be susceptible to extreme outliers, can be unstable

Ensemble methods (random forest and gradient boosting)	Less prone to overfitting the data than decision trees, more stable, can handle large amounts of data with high dimensionality, have high predictive performance and are less vulnerable to outliers	Opaque, prone to overfitting with smaller data sets

1.4. Common Issues in Credit Risk Prediction

1.4.1. Class Imbalance Issue

One of the common issues in credit risk prediction is a class imbalance in the dataset, as the target event (default of the company) usually is rare. An imbalanced dataset has significantly fewer observations in one class (minority class) than in another (majority class) [50]. Class imbalance can significantly impair the predictive performance of AI algorithms as they usually expect balanced datasets or equal misclassification costs [51]. Imbalanced dataset issues are further aggravated when the sample size is small and observations are defined by the relatively high number of features. Frequently this combination leads to overfitting [51].

Several approaches are described in the literature for dealing with imbalanced datasets, such as resampling or applying cost-sensitive methods [51]. Cost-sensitive methods usually attribute larger weight for misclassifying minority class than majority class with cost-sensitivity embedded inside the algorithm [52]. It is typical to assign no cost when both classes are correctly classified and a higher cost when a minority class is misclassified [52].

Contrary to cost-sensitive methods, resampling techniques work as wrapper-based methods, which can be applied to any AI model [52]. Resampling of the dataset can be done by either oversampling the minority class, undersampling the dominant class or applying a combination of both approaches [53]. Undersampling refers to reducing the number of observations in the majority class, while oversampling refers to increasing the number of observations in the minority class.

The most used resampling techniques include random oversampling, random undersampling and Synthetic Minority Over-sampling Technique (SMOTE) [50]. Random oversampling multiplies randomly selected minority observations [51]. The main drawback of random oversampling is that the method simply replicates the data, which could lead to overfitting [51]. SMOTE is an improved version of random oversampling because it makes slightly adjusted copies of minority class, making minority class decision regions more general [53]. However, SMOTE has some drawbacks, such as over-generalisation and variance [53]. Random undersampling removes randomly selected observations from the majority class [51]. The main issue with undersampling is the loss of data [51] and this could be particularly important when the dataset is small.

Generally, there is no consensus in the literature on which approach for dealing with imbalanced data is the best and, in some situations, original data can provide better performance [51]. For example, E. Burnaev, P. Erofeev and A. Papanov [50] used Bootstrap, random undersampling and SMOTE resampling techniques with decision trees, KNN and logistic regression with a pool of artificial and real datasets. The authors concluded that resampling could improve prediction performance in most cases and the best resampling technique varies with different datasets, AI algorithms and selected parameters. While G. M. Weiss, K. McCarthy and B. Zabar [52] compared cost-sensitive method, random oversampling and random undersampling using decision trees and 14 datasets. The results showed that no technique was distinctively better than the others. However, it was noticed that with smaller datasets, oversampling showed better results, while with larger datasets cost-sensitive algorithm outperformed resampling techniques.

1.4.2. Missing Data Issue

Another common issue in credit risk prediction is missing data. Most AI algorithms cannot work with missing observations [54]. Typical methods for dealing with missing data include deleting rows or columns with missing values or imputing missing values with another value [54]. Although the first option is the easiest to implement, it leads to data loss [55]. The choice of how to fill in the missing values usually depends on the type of missingness. D. B. Rubin [56] suggested classifying missing data into three types – missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). When missing data are MCAR, the reason why data are missing is entirely random and is not related to any observed or unobserved characteristics of an object. When the probability that the data are missing depends on other observed characteristics, the missing data are MAR. Finally, when the probability that the data are missing depends on both observed and unobserved characteristics, the missingness is called MNAR. When missing data are MAR, simple approaches, such as overall mean imputation, are unsuitable, but more advanced methods, such as single and multiple imputation, provide unbiased results [57].

Some frequently used single and multiple imputation techniques include k-nearest neighbours (KNN) imputation and multivariate imputation by chained equations (MICE) with various algorithms. For example, A. Jadhav et al. [58] compared the performance of several single imputation methods (mean imputation, median imputation and KNN imputation) and several multiple imputation algorithms from MICE (predictive mean matching, Bayesian Linear Regression, Linear Regression, non-Bayesian and random sample). The methods were compared using five different datasets. Results showed that KNN imputation outperformed all other imputation techniques with all datasets. O. Troyanskaya et al. [59] compared KNN, mean and singular value decomposition (SVD) techniques for missing data imputation. The authors concluded that both KNN and SVD provide fast and accurate results, but the KNN is more robust when the proportion of missing data increases and is less sensitive to the selected parameters. L. Yu et al. [60] compared four single imputation methods (mean, hot deck, KNN and linear regression imputation), three multiple imputation methods (random forest regression, random sample and multiple imputation integrated by four single model imputation methods) and one-hot encoding method using three credit risk datasets. The results showed that multiple imputation generally outperforms single imputation, but single imputation techniques are superior considering computation time.

1.5. Overview of Data Types Used in Credit Risk Prediction

This section will overview various data types found in credit risk prediction literature, where either classical statistical, AI or both types of methods were applied.

1.5.1. Financial Data

According to the reviewed literature, traditional financial data in credit risk prediction is typically financial ratios computed from the companies' financial statements. Financial ratios usually cover such areas as liquidity, profitability, turnover and solvency. Financial data dominated in overviewed studies, which employed simple statistical methods, namely univariate and multivariate statistical analysis [17, 26, 34, 35, 36]. While research works that employed algorithms capable of dealing with other forms of information (e.g., categorical data) frequently included alternative data-based variables in addition to or instead of financial variables (for example [23, 29, 40, 49]). Alternative data are discussed in more detail in the following subsection. Examples of financial ratios used in overviewed studies are presented in Table 2 below.

1.5.2. Alternative Data

The data not classified as financial is called alternative data in this study. With the growing popularity of AI-based methods for credit risk prediction, the inclusion of alternative data in assessing corporate credit risk has become increasingly popular. Alternative information could include noncomplex information such as company size, location and industry, which do not necessarily require sophisticated AI models and could be analysed using both AI and classical statistical methods. Alternatively, some studies used more complex alternative data, which requires more sophisticated analysis techniques. The data could be sizeable and include unstructured information, language and perceptions [14]. The specific examples from conducted studies are overviewed below.

1.5.2.1. General Characteristics of the Company

A typical example of simple alternative data is general company characteristics. For instance, I. M. Horta and A. S. Camanho [29] used information about the business activities, size and headquarters location beside six financial ratios related to liquidity, profitability, leverage and activity to predict contractor failure of Portuguese companies in the construction industry. The results showed that alternative data-based variables could improve ROC AUC and accuracy. In addition to company size, location and industry-related variables, some studies included ownership-related information. For example, N. H. Wellalage and S. Locke [23] used company size, geographical location, industry, owner involvement in the management process and whether the company is a franchise measuring variables besides traditional financial features to predict credit risk of New Zealand SMEs. Authors found that smaller, located in rural areas, non-franchised and with fewer inside ownership SMEs are more likely to default. Additionally, results showed that industry information also had an impact on credit risk prediction results.

1.5.2.2. Payment Behaviour Data

Another alternative data used in credit risk prediction is payment behaviour information. For instance, A. Malakauskas and A. Lakštutienė [48] applied several alternative data-based ratios such as historic overdue payments, company age and country beside traditional financial ratios measuring profitability, liquidity, activity, leverage, coverage, change in sales and short-term assets of more than 12,000 SMEs from Baltic countries to predict credit risk. The authors achieved the best results when the dataset included alternative data. Another example is a study by F. Ciampi et al. [7], where they found that including payment behaviour information in a model with financial ratios could improve prediction accuracy, especially when the company's size is small. Similarly, N. Wilson et al. [61] found that historic payment behaviour data substantially improves credit risk prediction performance compared to models constructed using only financial data.

1.5.2.3. Industry-Specific Information and Other Subjective Criteria

Other studies investigating credit risk prediction with companies from specific industries have included industry-specific alternative data-based variables. For example, H. S. Kim and S. Y. Sohn

[37] used simple alternative data-based features, such as SMEs characteristics, technology evaluation factors, and economic indicators, besides financial features measuring activity, profitability, liquidity and growth of 4,590 Korean SMEs from the technology field to predict credit risk. While N. Gulsoy and S. Kulluk [5] performed a data mining study where they used six subjective and six objective criteria to find a credit rating for a data sample of 103 SMEs from banking institutions in Turkey. Subjective criteria included morality (evaluated based on payment ethics and similar criteria), partnership structure (e.g., sector experience, the experience of the company partners), wealth (evaluates firm and its shareholders), financial asset (assessed financial liabilities), production (e.g., capacity utilization, performance) and sales (e.g., sales volume, sales conditions) sections. While objective criteria included liquidity, active structure, borrowing, profitability, performance and consolidated sections, which were measured based on quantitative data.

1.5.2.4. Textual Information

Some studies used variables constructed from textual information to evaluate credit risk. For instance, M. F. Tsai and C. J. Wang [62] used textual information from financial reports and sentiment analysis to evaluate the riskiness of publicly traded companies. The analysis confirmed that textual information from financial reports, especially financial sentiment words, is important in credit risk prediction. G. Yao, X. Hu, L. Xu and Z. Wu [49] tested whether social media data can improve credit risk prediction in the supply chain of listed companies in China. The authors applied sentiment analysis to analyse social media data retrieved from JRJ financial website and constructed four text sentiment features. Results confirmed the hypothesis that social media data improves credit risk prediction performance and even in most cases social media-based features, especially the time-weighted text sentiment feature, have better prediction power than financial data-based features.

Although the abovementioned methods showed promising results, they can be difficult to apply, especially to SMEs, due to specific data sources. Thus, other studies focused on SMEs. For instance, C. Yin et al. [40] examined the impact of publicly available legal judgment data on China's SMEs' credit risk prediction. The dataset used in the study included ten financial variables, six simple alternative data-based variables (age of the company, registered capital, number of insured, number of patents, city and number of shareholders) and variables constructed from legal judgements information. Authors found that adding features extracted from legal judgments to a model with conventional variables improves model performance. Another author, C. P. Sanchez et al. [63], extracted data from audit reports of 98 Spanish non-financial SMEs and created a model for predicting the credit risk. The model included only alternative data, such as the frequency of changing auditors, the average length of auditors' contracts and the proportion of qualified audit reports. The results showed that late submission of annual financial statements, the average duration of the contract with the auditors and the proportion of audit reports with a qualified opinion are statistically significant features in predicting credit risk.

1.5.2.5. Relational Data

The final data type found in the overviewed literature includes relational data. For example, E. Tobback et al. [24] conducted a study using relational data of Belgian SMEs to predict credit risk. In this study, relational data included information about current and past management and directors of the company and their links with other companies. Authors hypothesised that a company, related to many other bankrupt companies, have a greater failure rate, particularly when this company is in a bad financial position. In addition to relational data, the authors employed financial ratios (debt to total assets, current ratio, cash flow to equity, return on equity, profit (loss), return on total assets) and alternative data-based ratios such as date of incorporation and industry code. The authors found that relational data alone does not have enough predictive power, but supplementing a financial data-based model with relational data improves accuracy, especially when predicting the credit risk of the riskiest SMEs.

1.6. Summary of Overviewed Studies

The table below summarises credit risk prediction methods and data used in the overviewed literature.

Table 2. Summary of overviewed credit risk prediction studies

Author(s)	Year	Ref.	Method(s)	Data
Beaver	1966	[26]	Univariate analysis	**Financial:** Cash flow/total debt, net income/total assets, total debt/total assets, working capital/total assets, current ratio, no-credit interval
Altman	1968	[17]	MDA	**Financial:** Working capital/total assets, retained earnings/total assets, EBIT/total assets, market value equity/book value of total debt, sales/total assets
Orgler	1970	[38]	Multivariate regression analysis	**Financial:** Current assets/current liabilities, working capital, cash/current liabilities, inventory/current assets, quick ratio, working capital/current assets, net profit/sales, net profit/net worth, net profit/total assets, net profit, net worth/total liabilities, net worth/fixed assets, net worth/long-term debt, net worth, sales/fixed assets, sales/net worth, sales/total assets, sales/inventory, sales/receivables, various variable derived from a change in current ratio, net profit, sales, total assets, net worth. **Alternative:** Binary variables such as if the company is incorporated, if it has the latest balance sheet and profit and loss statement, if financial statements were audited, has the loan balance changed, has the loan been criticised by a bank examiner, is the loan secured, is it demand or time loan
Deakin	1972	[34]	MDA	**Financial:** Cash flow/total debt, net income/total assets, total debt/total assets, current assets/total assets, quick assets/total assets, working capital/total assets, cash/total assets, current assets/current liabilities, quick assets/current liabilities, cash/current liabilities, current assets/sales, quick assets/sales, working capital/sales, cash/sales
Blum	1974	[36]	MDA	**Financial:** Quick flow ratio, net quick assets/inventory, cash flow/total liabilities, net worth at fair market value/total liabilities, net worth at book value/total liabilities, rate of return to common stockholders who invest for a minimum of three years, a standard deviation of net income over a period, trend break for net income, a slope for net income, standard deviation, trend breaks and slope of net quick assets to inventory
Ohlson	1980	[18]	Logistic regression	**Financial:** Log(total assets/GNP price-level index), total liabilities/total assets, working capital/total assets, current liabilities/current assets, binary variable if total liabilities exceed total assets, net income/total assets, funds provided by operation/total liabilities
Odom & Sharda	1990	[19]	MDA, neural networks	**Financial:** Working capital/total assets, retained earnings/total assets, EBIT/total assets, the market value of equity/book value of total debt, sales/total assets

Tam & Kiang	1992	[27]	Neural networks, MDA, logistic regression, KNN, decision trees	**Financial:** 19 financial ratios measuring capital adequacy, asset quality, earnings and liquidity
Bell	1997	[28]	Neural networks, logistic regression	**Financial:** 28 variables measuring such financial features as bank size, loan exposure, capital adequacy, asset quality, operating performance, non-operating performance and liquidity
Zhang et al.	1999	[20]	Neural networks, logistic regression	**Financial:** Working capital/total assets, retained earnings/total assets, earnings before interest and tax/total assets, the market value of equity/ total debt, and sales/total assets, current assets/current liabilities
Charalambous et al.	2000	[21]	4 different neural networks algorithms, logistic regression	**Financial:** Cash and equivalents/total assets, current liabilities/total assets, change in accounts receivable, total debt/total assets, dummy variable showing if operating income was positive or negative for the last two years, change in cash flow from operations/the market value of equity at fiscal year-end, working capital from operations/market value of equity at fiscal year-end
Wilson et al.	2000	[61]	Logistic regression	**Financial:** Log of net sales, log of total debt, log of total assets, log of trade creditors, log of stock/work in progress, log of cash and liquid assets, dummy variables indicating whether each of financial variables had missing data **Alternative:** Industry, age, payment behaviour-based variables constructed using D&B Payment Score (Paydex) data - minimum Paydex last 12 months, number of increases in the last 12 months and 3 months, number of decreases in the last 6 months, variance in the last 12 months
Pompe & Bilderbeek	2005	[35]	MDA, neural networks	**Financial:** 11 profitability ratios, 28 activity ratios, 24 liquidity ratios, 10 solvency ratios
Shin et al.	2005	[22]	SVM, neural networks	**Financial:** Total asset growth, contribution margin, operating income/total asset, fixed assets/sales, owner's equity/total assets, net assets/total assets, net loan dependence rate, operating assets constitute ratio, working capital turnover period, net operating assets turnover period
Kim & Sohn	2010	[37]	SVM, neural networks, logistic regression	**Financial:** Net income/total assets, net income/shareholder's equity, net income/sales, total asset turnover, stockholder's equity turnover, total assets growth rate, stockholder's equity growth rate, sales growth rate, debt ratio **Alternative:** 8 variables related to SMEs characteristics (such as age, if financials are audited by an external auditor and if listed on the stock market), 16 variables related to technology evaluation (such as technology knowledge, the environment of technology development and market potential), 10 economic indicators (such as consumer price index and oil price)
Hewa Wellalage & Locke	2012	[23]	Logistic regression	**Financial:** Total debt/total assets, current assets/current liabilities, net income/total assets **Alternative:** Proportion of owners, which are also employees in the company, if the company is a franchise, geographical location, industry, company size
Horta & Camanho	2013	[29]	SVM, logistic regression	**Financial:** Total profit/sales, net profit/total assets, net profit/shareholders equity, current ratio, working capital/total assets, the net value of sales/average current assets **Alternative:** Company's main activity, company size and headquarter geographic location

Sánchez et al.	2013	[63]	Logistic regression	**Alternative:** Proportion of audited years, number of different auditors, change in auditor and opinion at the same time, the average length of auditors' contracts, the proportion of qualified audit reports, number of critical qualified audit reports, auditor type, explicit obstructionism, compliance with audit obligation, delay in submitting financial statements
Lahmiri	2016	[45]	SVM, 2 different neural networks algorithms, MDA, naïve Bayes classifier	Not available.
Tobback et al.	2017	[24]	SVM, relational data learner – weighted vote relational neighbour classifier	**Financial:** Debt/total assets, current ratio, cash flow/equity, ROE, profit (loss), ROA **Alternative:** Age of the company, industry, relational data (information about current and past directors and managers of the companies)
Tsai & Wang	2017	[62]	Support vector regression and ranking SVM	**Alternative:** Textual information retrieved from public financial statements
Ptak-Chmielewska	2019	[43]	Decision trees, SVM, random forest, logistic regression and neural networks	**Financial:** Current liquidity, quick ratio, liquidity cash, capital share in assets, gross margin, operating profitability of sales, operating profitability of assets, net profitability of equity, assets turnover, current assets turnover, receivables turnover, inventory turnover, capital ratio, coverage of short-term liabilities by equity, coverage of fixed assets by equity, the share of net financial surplus in total liabilities **Alternative:** Sector of the company's activity, company's legal form, region, age of the company, number of employees
Gulsoy & Kulluk	2019	[5]	6 different decision trees algorithms	**Financial:** Variables related to liquidity, activity, borrowing, profitability, performance, change in total bank loans **Alternative:** Subjective opinion-based variables related to morality, partnership structure, wealth, financial assets, production and sales of the company
Ciampi et al.	2020	[7]	Logistic regression, neural networks (Kohonen maps), Shumway's discrete time hazard model	**Financial:** Bank loans/turnover, net financial position/turnover, EBIT/turnover, interest expense/EBITDA **Alternative:** Past due and/or overdrawn exposures for more than 60 days, past due and/or overdrawn exposures for more than 60 days/EBITDA, number of cumulative non-remedied payment delays exceeding 60 days
Yin et al.	2020	[40]	Logistic regression, random forest, XGBoost	**Financial:** Current ratio, debt/assets, quick ratio, receivables turnover ratio, inventory turnover ratio, total assets turnover, operating profit ratio, rate of return on common stockholders' equity, ROA **Alternative:** Proportion of missing financial variables, age of the company, registered capital, number of insured, number of patents, city, number of shareholders and variables constructed from legal judgements information (if the company have judgements in each of four distinguished legal judgments categories and the number of judgements the company has in different categories)

Malakauskas & Lakštutienė	2021	[48]	Logistic regression, neural networks, random forest	**Financial:** Gross margin, profit margin, current ratio, quick ratio, cash/current liabilities, accounts receivable turnover, debt/total equity, debt-service coverage ratio, asset coverage ratio, change in sales, change in short-term assets **Alternative:** Historic overdue, company age, country
Lextrait	2022	[4]	LightGBM, XGBoost, CatBoost	**Financial:** Several hundred financial variables **Alternative:** Company age, sector, location, employees related variables, company size
Yao et al.	2022	[49]	SVM, random forest, extreme random tree	**Financial:** Operating cycle, total assets turnover ratio, ROE, profit margin, ROA, current ratio, quick ratio, cash ratio, debt/assets, net profit growth rate, operating revenue growth rate, total assets growth rate, accounts receivable turnover, accounts receivable collection period **Alternative:** 4 text sentiment features based on social media data, industry trends, industry concentration ratio, supply chain concentration ratio

The summary table reveals how the popularity of different methods changed over time. Classical statistical models such as MDA and logistic regression were dominant in early studies. Later, neural networks and SVM gained popularity, while ensemble techniques were frequently tested in the most recent studies. In addition, a variety of different statistical and AI algorithms in overviewed studies show that there is still no consensus on which method is the best. It might be related to the fact that all studies use different datasets, including different companies, different time periods, the different balance between defaulted and non-defaulted companies and different features. Thus, it is important to try multiple methods, as the best-performing models may vary depending on the data.

Random forest, gradient boosting (three different algorithms) and SVM were selected for this thesis's research part. All these models were previously successfully employed in credit risk prediction literature. More specifically, ensemble methods were chosen due to their efficiency in credit risk prediction and superiority compared to decision trees. The SVM algorithm was also shown as an effective method that performs well with smaller datasets. Although considered, classical statistical methods were not chosen due to a need to satisfy various assumptions, which are usually difficult to satisfy with actual data. Furthermore, the neural networks were also considered but not chosen since they are known to perform well in complex tasks but often require a large amount of data for training and extensive knowledge in parameter tuning. Therefore, other AI algorithms appear more suitable due to the limited dataset size available for this study and the scope of the research.

Various types of alternative data have already been tested and demonstrated promising results in credit risk prediction modelling. These results seem encouraging, as including alternative data in credit risk prediction could expand opportunities to receive financing for SMEs, which usually lack quality financial data for credit risk assessment. Although some more advanced data forms, such as relational or textual data, seem promising data sources, in practice, it might require a significant amount of time and expertise to collect and process such data, creating additional expenses for creditors. Also, some forms of alternative data were already questioned in the literature. For example, Wei et al. [64] concluded that initially models using social media information might enhance the accuracy of credit risk prediction and improve the financial inclusion of disadvantaged groups, however, later debtors might start to deliberately alter their social networks to improve their credit risk prediction results and the possibility to receive financing. Thus, general characteristics of the company and payment behaviour data appeared as the most promising alternative data types due to easier collection, processing and resistance to manipulation and were selected for the research part of this thesis. In addition, X. Dastile et al. [41] reviewed multiple studies on credit risk prediction and found that most of the research did not include macroeconomic variables, even though such indicators as interest rates or inflation might impact the borrower's risk. Thus, it was decided to include macroeconomic variables in this research to consider the impact of the macroeconomic environment.

Although payment data usually are abundant inside the companies, it is not available in public data sources. Thus, the research conducted in this thesis is subject to several limitations due to the data constraints. Despite concerted efforts to obtain a comprehensive dataset, the available data presented certain limitations. These limitations may have affected the accuracy and generalisability of the findings and are presented in more detail at the end of section 3. Hence, the conclusions drawn from the research should be interpreted within the context of these limitations.

2. Research Methods

This section provides specific information related to the empirical research. The research process included several steps, as listed below:

1. Initial data aggregation and processing – merging of different data tables, separating SMEs, creating class variables and independent variables based on literature review and available data;
2. Descriptive data analysis;
3. Splitting the data into train and test sets;
4. Preparation of the data for the models – missing data imputation, encoding of dummy variables, standardisation of data values;
5. Hyperparameter tuning using cross-validation;
6. Applying selected models on test datasets.

All the steps were carried out using Python programming language (version 3.10.9) and several Python libraries, such as scikit-learn[2] [65], imbalanced-learn[3], pandas[4] and NumPy[5].

2.1. Data

2.1.1. Data Source

Data for the empirical research was collected from a Lithuanian FinTech company engaged in credit risk assessment. The dataset was provided in Excel format, included 1,307 companies, and covered the period from 1 January 2014 to 4 October 2022. All provided data was encoded, i.e., all company names and other information based on which the company could be identified were replaced with arbitrary codes. The data was divided into four areas:

- General information about companies – company ID, legal form, legal status, industry information, number of employees, registration date and city;
- Financials – information from financial statements;
- Legal information – information about court cases in which the companies in the dataset are involved. It was decided not to include this information in the further analysis due to the limited number of concerned companies;
- Transaction data, including such data as document date, due date, payment date, transaction amount and supplier code.

In addition to the provided data, some macroeconomic data-based variables were added to the dataset:

- Inflation data from State Data Agency Statistics Lithuania[6];
- GDP growth data from State Data Agency Statistics Lithuania;
- Interest rates from the Bank of Lithuania[7].

[2] https://scikit-learn.org/

[3] https://imbalanced-learn.org

[4] https://pandas.pydata.org

[5] https://numpy.org

[6] https://osp.stat.gov.lt/statistiniu-rodikliu-analize#/

[7] https://www.lb.lt/lt/statistika

2.1.2. Definition of Default in the Research

The dataset did not include any variable indicating whether the company has defaulted, thus, it was necessary to create such variable. As described in the literature review section, the definition of default is still subject to discussion in academia. Many studies in the credit risk field have considered the company as defaulted in case of bankruptcy. However, the dataset did not include sufficient information about bankruptcy, namely when the bankruptcy occurred. In addition, companies may miss payments even when they are not in bankruptcy. Such unpredicted overdue receivables might pose a risk to the recipient's solvency and liquidity, while timely identification of the risk that the debtor will not pay on time allows the company to take measures to reduce risks. Therefore, in this study, it was decided to adopt the second part of the definition of default from the Basel II accord, which states that the company is considered defaulted on payment if it has more than 90 days overdue payment.

Transaction data was used to decide if the company has overdue payments of more than 90 days. Although transaction data covered various points during a year, financial data was provided only for a full year, therefore, it was decided to perform analysis on a yearly basis and predict if the company would default in the following 12-month period. The figure below illustrates the process for creating default status for each company in a particular year.

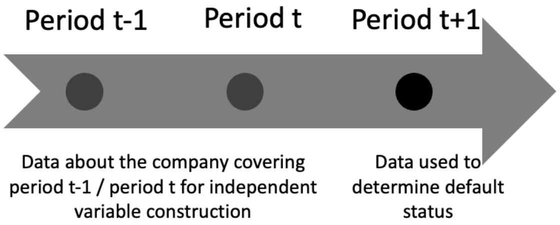
Period t-1
Period t
Period t+1
Data about the company covering period t-1 / period t for independent variable construction
Data used to determine default status

Fig. 1. Process for creating default status

2.1.3. Initial Data Processing

Initial data processing included the following steps:

- **Processing of transaction data**. Calculating overdue days of each transaction as a difference between the due date and actual payment date and grouping into five groups based on the number of overdue days – not overdue, overdue 1-30 days, overdue 31-60 days, overdue 61-90 days and overdue more than 90 days. If some transactions were still not paid, they were classified as overdue for more than 90 days. All transactions were assigned to a particular year based on their due date. Transactions from year 2022 were excluded from the dataset because the data for the entire year of 2022 was not provided.
- **Creating default status variable**.
- **Preparing the format of other separate tables for merging into one table**. Observations of a company in a particular year were included in the merged table only if it had required financial information and transaction information covering a three-year period, as presented in Figure 1.
- **Excluding large companies from a dataset**. SMEs were identified based on the SME definition by European Commission.

- **Creating independent variables**. Financial and alternative data-based variables were created based on variables identified in overviewed literature and available data in this dataset. After this step, the dataset included 44 variables – year, default status variable, company age, four categorical variables (legal form, industry, location, number of employees), three macroeconomic variables, 19 financial ratios and 15 variables based on transaction data. The year variable is not used in model development and was excluded from the dataset after splitting to train and test sets.

After initial data processing, the combined dataset had 2,266 observations (rows) from 825 different Lithuanian companies and covered the 2015-2020 period. Number of observations in the dataset is larger than the number of companies because some companies are repeated through several rows if there is sufficient data to cover different periods. Also, the same company could be defaulted multiple times. It was decided not to exclude any observations from the dataset due to the limited number of observations and the nature of the default event used in this analysis. Default is defined as overdue payments, which might occur as a result of serious financial problems (such as bankruptcy), however, overdue payments also can be caused by some temporary circumstances, which might be no longer relevant after some time. The dataset is imbalanced, as there were only 4.3% of cases when a company defaulted.

2.1.4. Variable Selection

After initial data processing, the dataset contained 42 independent variables and one dependent variable. As the number of independent variables is rather substantial, some variables are expected to be redundant. Removing redundant features decreases storage and computing costs while avoiding significant loss of information and a decrease in prediction performance [66]. Therefore, it was decided to calculate the correlations between numerical variables and remove the highly correlated variables. A correlation matrix was computed using the non-parametric Spearman correlation method, which measures the monotonic relationship between two variables [67]. Spearman correlation is estimated as:

$$\rho_{r_x,r_y} = \frac{cov(r_x,r_y)}{\sigma_{r_x}\sigma_{r_y}} \qquad (1)$$

If $cov(r_x, r_y)$ is the covariance of rank variables r_x, r_y and σ_{r_x}, σ_{r_y} are standard deviations of r_x, r_y. The non-parametric method was selected because, based on the literature overview, financial data usually does not satisfy linearity and normality assumptions. The correlation coefficient may vary from -1 to 1, where -1 and 1 show a perfect monotonic relationship and 0 shows that there is no monotonic relationship between variables.

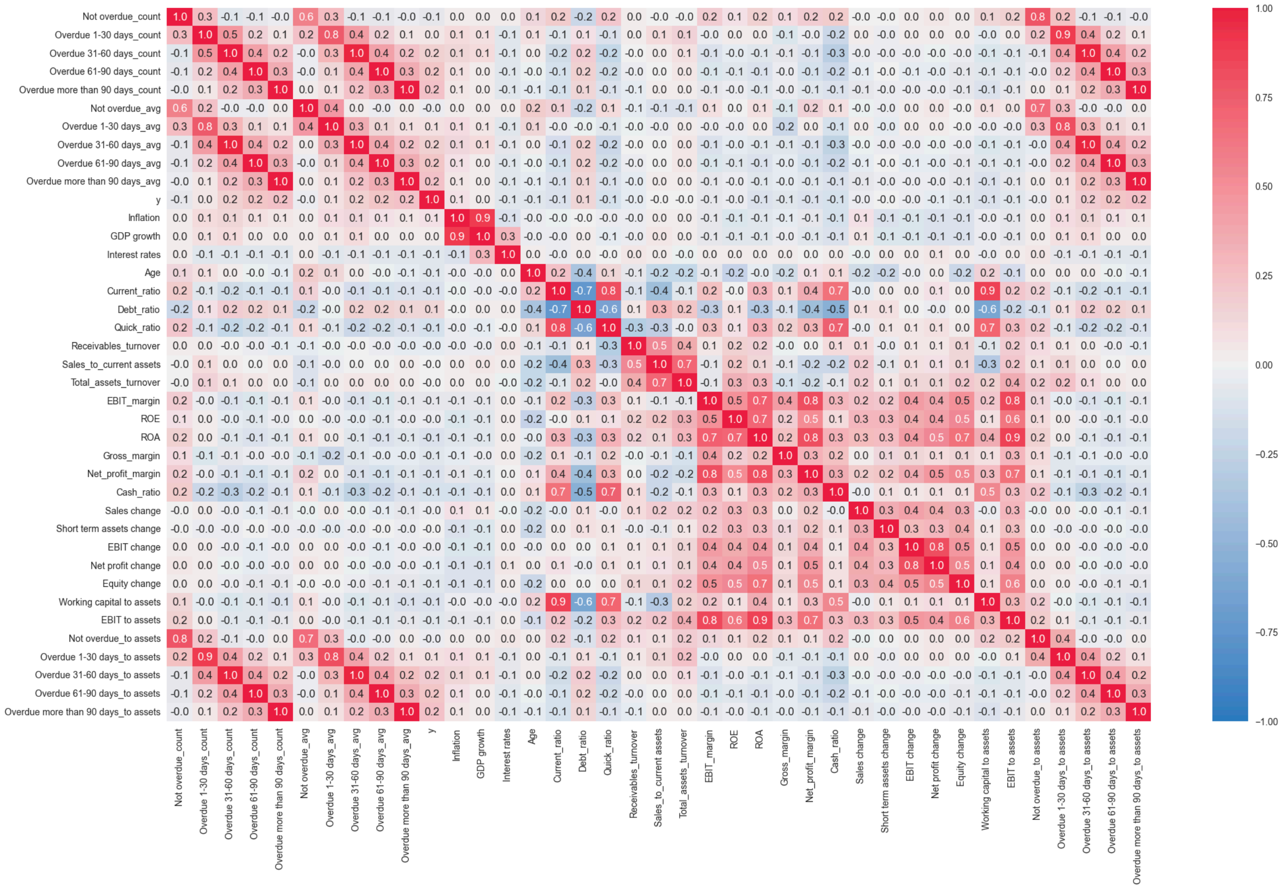

Fig. 2. Correlation matrix of numerical variables

The correlation matrix presented above shows that there are several correlated variables in the dataset. After analysing the matrix, it was decided to exclude variables with a correlation coefficient of 0.8 or higher. 16 variables were dropped from the dataset - Not overdue_count, Overdue 1-30 days_count, Overdue 31-60 days_count, Overdue 61-90 days_count, Overdue more than 90 days_count, Not overdue_avg, Overdue 1-30 days_avg, Overdue 31-60 days_avg, Overdue 61-90 days_avg, Overdue more than 90 days_avg, Inflation, Quick_ratio, EBIT_margin, Working capital to assets, ROA and EBIT change. The correlation coefficients of Not overdue_count and Not overdue_avg were below 0.8, but it was decided to exclude these variables for consistency reasons.

2.2. Description of Variables and Initial Data Analysis

After initial data cleaning, the dataset included 2,266 observations, one dependent variable and 26 independent variables. Independent variables included 12 alternative data-based variables, as presented in Table 3 and 14 financial data-based variables, as presented in Table 4.

Table 3. List of financial data-based variables

	Group	**Variable**	**Description**
Financial ratios	Liquidity & leverage	Current_ratio	$\frac{Total\ current\ assets}{Total\ current\ liabilities}$
		Debt_ratio	$\frac{Total\ assets - Total\ equity}{Total\ assets}$
		Cash_ratio	$\frac{Cash}{Total\ current\ liabilities}$

	Efficiency	Sales_to_current assets	$\frac{Sales}{Total\ current\ assets}$
		Receivables_turnover	$\frac{Sales}{Accounts\ receivable}$
		Total_assets_turnover	$\frac{Sales}{Total\ assets}$
	Profitability	ROE	$\frac{Net\ profit}{Total\ equity}$
		Gross_margin	$\frac{Sales - Cost\ of\ sales}{Sales}$
		Net_profit_margin	$\frac{Net\ profit}{Sales}$
		EBIT to assets	$\frac{EBIT}{Total\ assets}$
	Other	Sales change	$\frac{Sales_t - Sales_{t-1}}{\lvert Sales_{t-1}\rvert}$
		Short term assets change	$\frac{Total\ current\ assets_t - Total\ current\ assets_{t-1}}{\lvert Total\ current\ assets_{t-1}\rvert}$
		Net profit change	$\frac{Net\ profit_t - Net\ profit_{t-1}}{\lvert Net\ profit_{t-1}\rvert}$
		Equity change	$\frac{Equity_t - Equity_{t-1}}{\lvert Equity_{t-1}\rvert}$

Table 4. List of alternative data-based variables and descriptions

Category	**Variable**	**Description**
Class variable	y	y=0 the company will not default on any payments in the next 12 months y=1 the company will default on at least one payment in the next 12 months
General characteristics of the company	Employees group	Number of employees in the company, grouped into three categories - 10 - 49 employees, 50 - 249 employees and <10 employees
	Legal Form_aggregated	The legal form of the company
	Location	Location of the company
	Age	Age of the company
	NACE Code_letter	Industry based on NACE code
Macroeconomic indicators	GDP growth	Yearly change of GDP at constant prices (chain-linking method)
	Interest rates	Interest rates in new loan agreements with non-financial companies, an average of year t
Payment behaviour	Not overdue_to assets	$\frac{Sum\ of\ payments\ in\ EUR, which\ were\ paid\ on\ time\ in\ year\ t}{Total\ assets_t}$
	Overdue 1-30 days_to assets	$\frac{Sum\ of\ payments\ in\ EUR, which\ were\ paid\ up\ to\ 30\ days\ late\ in\ year\ t}{Total\ assets_t}$

	Overdue 31-60 days_to assets	$\frac{Sum\ of\ payments\ in\ EUR, which\ were\ paid\ 31-60\ days\ late\ in\ year\ t}{Total\ assets_t}$
	Overdue 61-90 days_to assets	$\frac{Sum\ of payments\ in\ EUR, which\ were\ paid\ 61-90\ days\ late\ in\ year\ t}{Total\ assets_t}$
	Overdue more than 90 days_to assets	$\frac{Sum\ of\ payments\ in\ EUR, which\ were\ paid\ more\ than\ 90\ days\ late\ in\ year\ t}{Total\ assets_t}$

Observations in the dataset are not evenly distributed during the 2015-2020 period, as most of the observations are from the years 2018-2020. The largest number of defaults occurred in 2017 (31 observations), while during 2018-2020, the number of default cases averaged around 21.

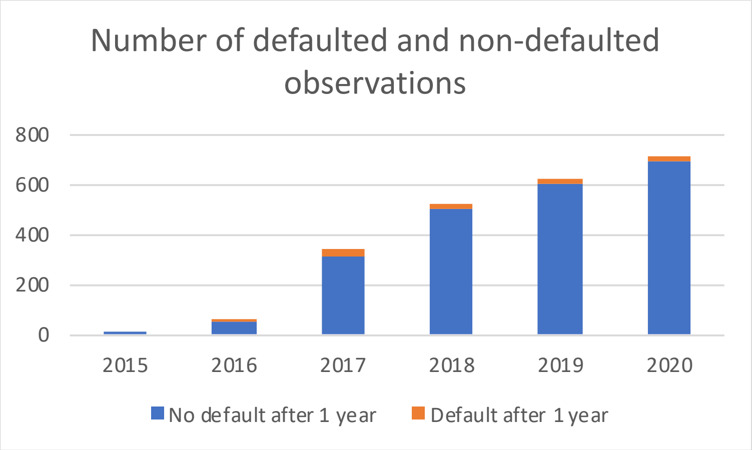

Fig. 3. Number of defaulted and non-defaulted observations

Descriptive statistics of the dataset were analysed to test whether there are differences between observations which will default on payment and will not default on any payments after one year. Mean, standard deviation, minimum, maximum, 1st and 3rd quartiles and median values were calculated for each numerical value. The table with descriptive statistics is provided in Appendix 1. Data analysis showed that with most variables, there are apparent differences between observations which will default on payment and will not default on any payments after one year. For illustrative purposes, a comparison of the median values of financial variables is provided below.

Figure 4 shows that median current ratio and cash ratio of the companies, which will default after one year, are lower, while the median debt ratio is higher compared to companies which will not default on any payments after one year. Such variations are anticipated, as it is expected that the company that defaults would have higher levels of liabilities.

Median receivables turnover is lower in observations, which will default after one year. This difference is also expected because the collection of accounts receivable in the company that will default is likely less effective than in the company that will not default. However, other median efficiency ratios, such as sales to current assets and total assets turnover, showed no visible differences between the two groups.

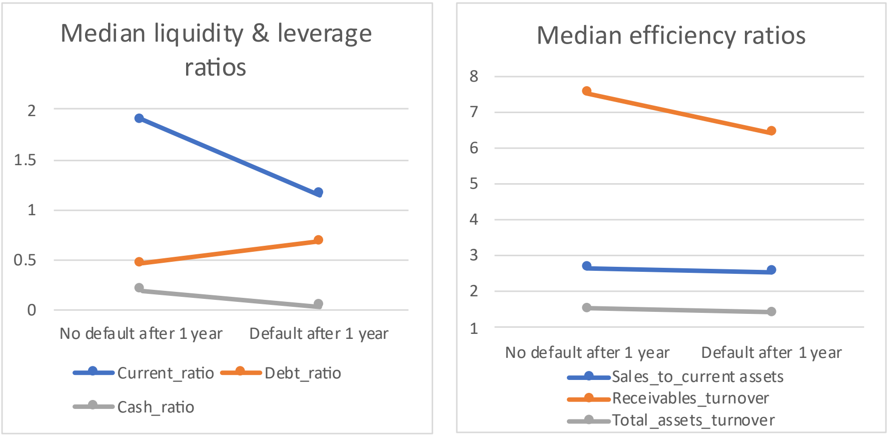

Fig. 4. Median liquidity, leverage and efficiency ratios

All median profitability ratios are lower in observations that will default after one year, with the gross profit margin showing the most noticeable decrease. These differences are also consistent with the expectation that companies that default will be less profitable than companies that pay on time.

Increases in sales, short-term assets, net profit, and equity during the year in observations that will default are either lower or negative when compared to observations that will not default. These differences are also expected because it is presumed that the company that defaults will have lower sales and profits.

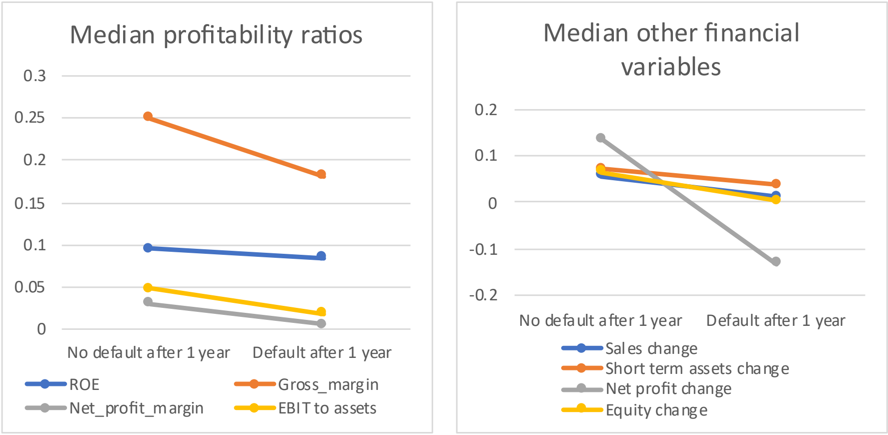

Fig. 5. Median profitability ratios and other financial variables

A comparison of alternative data-based variables indicates that observations that will default differ from those that will not. For example, a median company that will default after one year, made more late payments in the current year. Additionally, analysis shows that younger companies are more likely to default because the median age of the defaulted company is 14 years, compared to the median age of 19 years of non-defaulted companies.

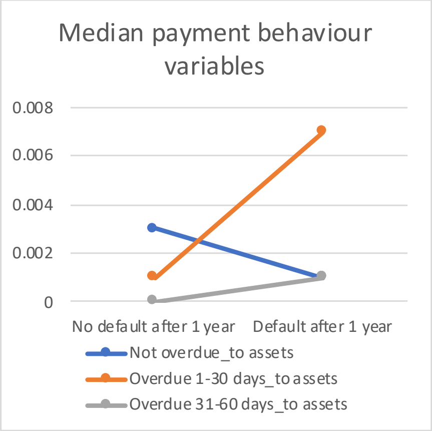

Fig. 6. Median payment behaviour variables

Analysis of categorical variables shows that the companies that will default more frequently have a minimal number of employees (less than 10) and more often are from the agriculture, forestry and fishing industries. While the distribution of locations in which companies are registered is similar for both types of observations.

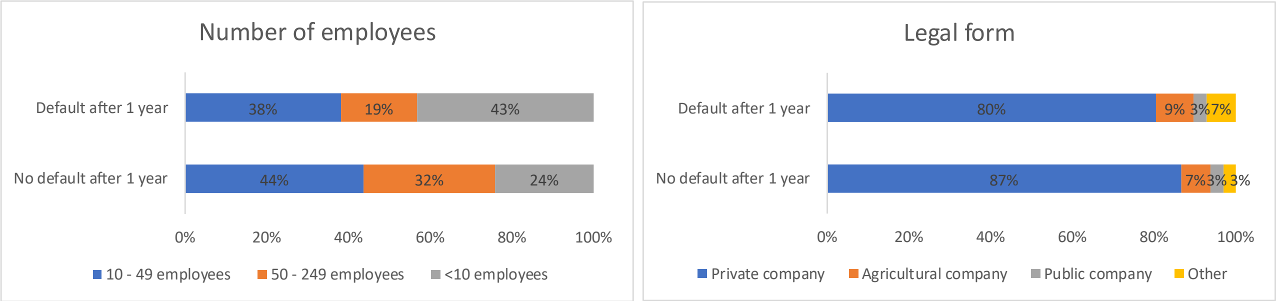

Fig. 7. Distribution of number of employees and legal forms

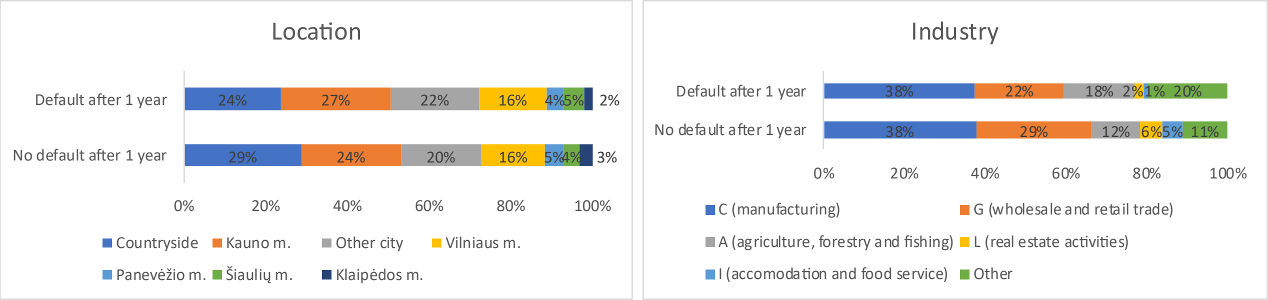

Fig. 8. Distribution of locations and industries

The dataset contains a few extreme minimum and maximum variables. In most cases, extreme values resulted after financial ratios calculation when the value in the denominator was significantly smaller than the ratio numerator. Since small financial values are typical in SMEs, it was decided not to discard any observations.

2.3. Missing Data and Imputation

1.3% of the data in the dataset was missing. The reasons why this data are missing are unknown. The figure below illustrates how missing data are distributed in the dataset (white dashes represent missing data). It can be noticed that Receivables_turnover and Cash_ratio variables have the most significant proportion of missing variables. These two variables differ from the others in that they require a full balance sheet to be calculated (accounts receivable and cash and cash equivalents items are required to calculate ratios). Thus, part of the missing data might be related to the fact that small companies are not obliged to file full financial statements in Lithuania [68].

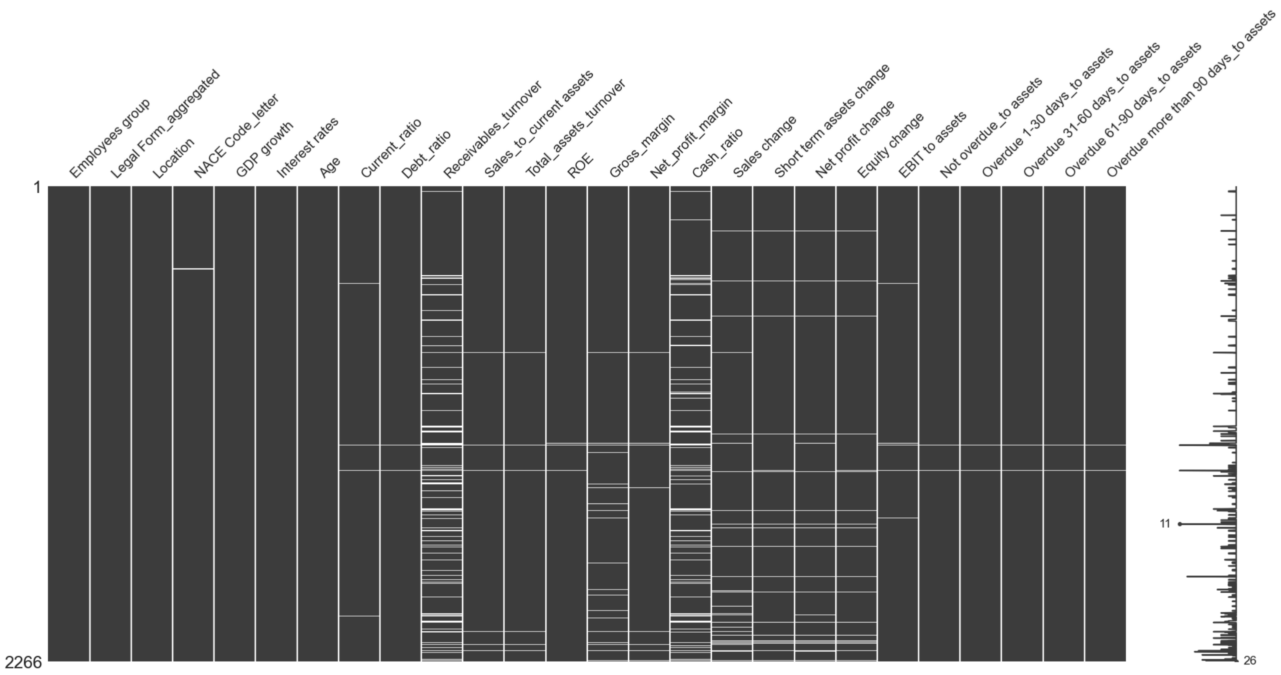

Fig. 9. Missing data in the dataset

The dataset used in this study is small, therefore, data imputation seems the most suitable option for handling missing values. Since the choice of the data imputation approach depends on the type of missingness, the type of missing data should be identified first. If part of the data is missing because small companies publish abbreviated financial statements, respective missing data should not be MCAR, as the probability that the data are missing will depend on the company's size. Furthermore, nullity correlations (which show the impact of variable’s presence or absence on another’s presence or absence) between missing values were estimated to test whether missing data are not MCAR. The figure provided below shows some significant correlations between missing values, thus, it can be concluded that missing values in the dataset are not MCAR. Whether the missing data are MNAR cannot be determined from the observed data as there is no test which shows if the missing data are MAR or MNAR [55] and would require further investigation of underlying reasons, which is out of the scope of this thesis. However, multiple significant correlations suggest that missing data could be classified as MAR [55]. Thus, single or multiple imputation should be the most suitable option.

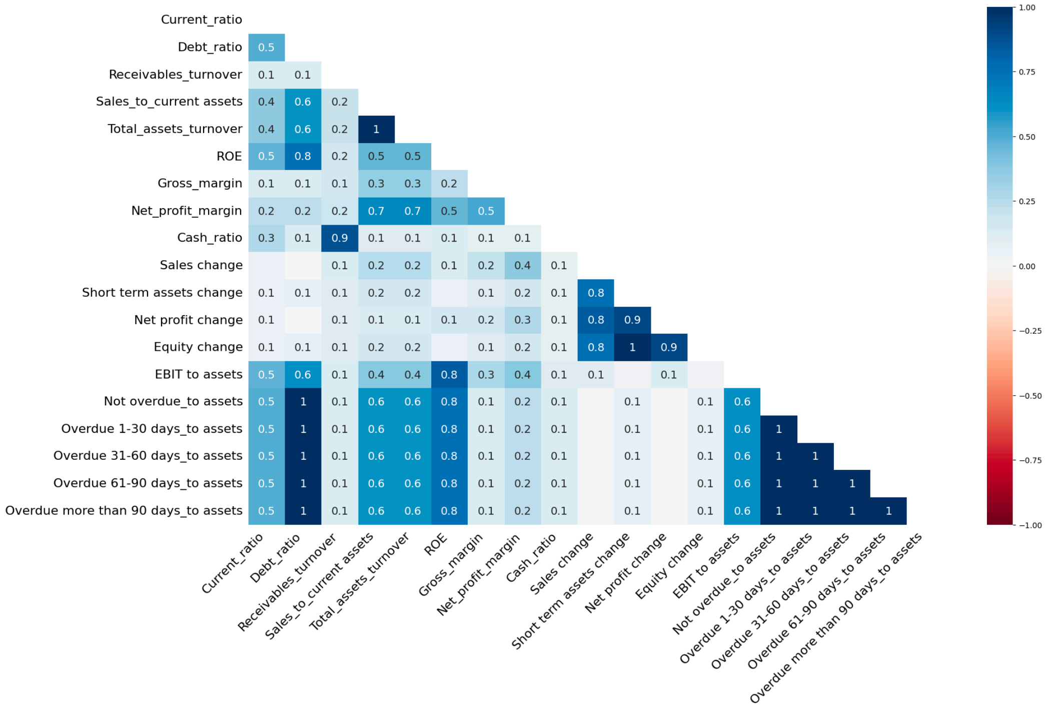

Fig. 10. Correlations between missing data

Specific imputation methods comparison is outside the scope of this thesis; hence, the imputation technique was chosen based on the literature overview. Based on performance results in previous studies, computation time and available methods in Python, it was decided to apply KNN for missing value imputation in this study. Imputation was done using the scikit-learn library. KNN imputation algorithm estimates the value for missing observation as the mean value from the k nearest neighbours found in the training set [59]. In this study, the five nearest neighbours were considered. Two samples are regarded as close based on Euclidean distance. Missing data imputation was performed after splitting the data to train and test sets to avoid data leakage into the test set. KNN imputation algorithm was trained on a train set and then applied to both data samples. Descriptive statistics of the imputed train dataset are provided in Appendix 2.

The selected KNN imputation algorithm cannot fill in missing categorical features, however, as the dataset contained only four missing values in the NACE Code_letter variable, it was decided to manage these missing values while performing one-hot encoding of categorical features, as described in the section 2.10.2.

2.4. Dataset Splitting

The dataset must be split for AI models to be trained and results to be evaluated. Train and test sets should be mutually exclusive and have similar variance as the entire dataset [69]. It is common to use 2/3 of the dataset as a train set and 1/3 as a test set [69]. However, there are no rules for the size of the train and test sets.

The dataset in this study was divided to train and test sets based on the year – all observations from the latest year of 2020 were put in the test set as in practice only historical data about the debtors would be available and the creditor would aim to assess the credit risk of future transactions.

After splitting the dataset, 68.4% of observations were assigned to the train set and 31.6% to the test set. Both train and test sets maintained a similar proportion of defaulted and non-defaulted observations as the full dataset.

Table 5. Distribution of defaulted and non-defaulted observations in data samples

y	**Full dataset**		**Train dataset**		**Test dataset**	
	Observations	Percentage	Observations	Percentage	Observations	Percentage
0	2169	95.7%	1475	95.1%	694	97.1%
1	97	4.3%	76	4.9%	21	2.9%

2.5. Used Algorithms

Based on the literature review presented in section 1, random forest, gradient boosting (gradient boosting decision trees, LightGBM and XGBoost) and SVM were selected for the research part of this thesis.

2.5.1. Random Forest and Gradient Boosting

The decision trees algorithm is the fundamental component of random forest and gradient boosting; thus, this section begins with the description of decision trees.

2.5.1.1. Decision Trees

The decision trees use recursive partitioning and metrics such as Gini impurity or entropy to divide the dataset into separate groups (form a decision tree) [43, 46]. The root of the decision tree is referred to as the root node, which is further split into intermediate nodes or leaf nodes (final nodes), as presented in the figure below.

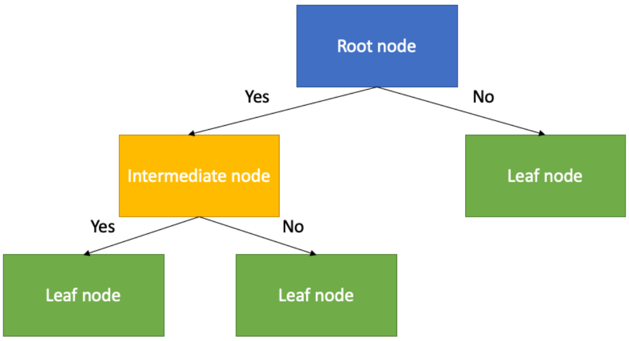

Fig. 11. Decision tree illustration

Decision trees are a non-parametric and require minimal data preparation, it is not necessary to scale or centre the data [54]. There are various algorithms for training decision trees, such as CART, ID3 and C5.0 Decision trees [54]. In this research, decision trees are trained with the CART algorithm and Gini impurity metric. CART algorithm divides the train set into two parts using a feature *k* and a threshold t_k. To find values for *k* and t_k an algorithm tries to minimise the cost function:

$$J(k, t_k) = \frac{m_i}{m} G_i + \frac{m_j}{m} G_j \tag{2}$$

If m is the number of observations in the train set, m_i and m_j are the number of instances in each part of the divided train set and G_i and G_j are Gini impurity metrics of each part of the divided train set. Gini impurity is estimated as:

$$G_x = 1 - \sum_{z=1}^{n} p_{x,z}^2 \tag{3}$$

If G_x is the Gini impurity of the x^{th} node, $p_{x,z}$ is the ratio of class z instances among training instances in the x^{th} node.

2.5.1.2. Random Forest

A random forest is an ensemble method that trains many decision tree classifiers on various dataset sub-samples and lets them vote for the most popular class [70]. The random forest algorithm used in this study estimates the average probabilistic prediction instead of voting. Usually, subsamples are generated with the bootstrapping method. The key idea of bootstrapping is to train the algorithm multiple times with different random subsets of the training dataset when the sampling is performed with replacement [54]. In addition, each decision tree used for training is assembled with a subset of randomly selected features to introduce extra randomness [41]. Introducing such randomness in the decision tree allows for reducing variance and overfitting.

2.5.1.3. Gradient Boosting

Gradient boosting decision trees is one of the boosting methods, which combines multiple weak learners into strong learner [54]. In gradient boosting, decision trees are trained sequentially, and each new predictor is trained on residual errors (also called negative gradients) made by the previous predictor. Gradient boosting was first introduced by L. Breiman [71] and later developed by J. H. Friedman [72]. The final prediction could be aggregated using summation, averaging or majority voting [4]. Gradient boosting prediction $\hat{y_i}$ estimated in this study could be written as follows:

$$\hat{y_i} = G_N(x_i) = \sum_{n=1}^{N} h_n(x_i) \tag{4}$$

If x_i is an input value and h_n is a weak learner. Each newly added tree h_n is fitted to minimise the total losses L_n:

$$h_n = \arg\min_h L_n = \arg\min_h \sum_{i=1}^{m} l(y_i, G_{n-1}(x_i) + h(x_i)) \tag{5}$$

If $G_{n-1}(x_i)$ is previous ensemble and $l(y_i, G(x_i))$ is a loss function. In this study, a negative log-likelihood loss function is used.

Gradient boosting has several optimized implementations, such as LightGBM and XGBoost. XGBoost is a gradient boosting algorithm with various optimization techniques, such as an approximate algorithm for splitting decision trees, parallelising tree building and early stopping of the boosting process to prevent overfitting [73]. These techniques help to increase the speed and accuracy of the model. While LightGBM is a gradient boosting approach that employs Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques to handle a large number of observations and variables [74]. The key idea of the GOSS technique is to help the model to select a subset of data for each iteration of boosting process, while EFB can group features with high correlation into a single feature.

2.5.2. Support Vector Machines

SVM was developed by V. N. Vapnik [75]. The SVM is based on statistical learning theory and can be applied to solve classification and regression tasks [22, 46]. The algorithm creates a line called binary classifier or hyperplane, separating the dataset into classes, even when the data are not linearly separable [22, 37]. The hyperplane of linear SVM is defined by the following equation [76]:

$$w^T \cdot x + b = 0 \tag{6}$$

If x is the training input vector, w is a weight vector and b is the bias term. An optimal hyperplane is found by selecting w and b to minimise the following equation:

$$Q(w, b) = \frac{1}{2} \|w\|^2 \tag{7}$$

Subject to the constraints

$$y_i(w^T x_i + b) \geq 1, \forall_i = 1, \ldots, n \tag{8}$$

The best possible hyperplane for separating classes is found by maximising the margin between the closest data points of different classes, as illustrated in the figure below.

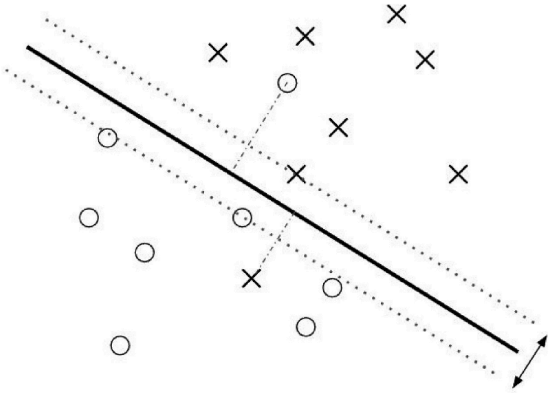

Fig. 12. Illustration of SVM in two-dimensional space [4]

The larger the margin width, the better the separator. However, in practice, usually, it is not possible to perfectly separate observations into classes, thus, observations classified to the wrong side are given a penalty, which depends on how far away wrongly classified points are positioned. Hence, SVM tries to find an optimal solution to achieve to widest margin and the smallest penalty [4]. To handle the data, which is non-linearly separable, SVM uses kernel functions, such as polynomial kernel or radial basis function (RBF) [22].

2.6. Performance Metrics

Multiple metrics for evaluating the classification performance are proposed in the literature. Generally, the metrics could be classified into three groups – metrics based on a threshold, probability and ranking [77]. Examples of threshold-based metrics include accuracy, F-score, precision and recall [77]. These metrics require a threshold level, where probabilities above the threshold are classified as positive and below the threshold as negative [78]. The results of these metrics can differ significantly based on the selected threshold. Metrics based on probability measure deviation from true probability and include such scores as mean absolute error, mean squared error (Brier score) and LogLoss (cross-entropy) [77]. While metrics based on ranking assess how well the model separates classes and include such measures as the area under the receiver operating characteristics curve (ROC AUC) and the area under the precision-recall curve (PR AUC) [77, 78]. These measures do not require setting a threshold and evaluate model performance under all possible thresholds [78].

The metrics chosen for model performance evaluation must consider the dataset's imbalance. However, some often-used traditional metrics, such as accuracy, are not suitable to imbalanced datasets because a model that classifies all observations to the majority class achieves high accuracy [53]. Thus, to evaluate the performance of models in this study, performance metrics were selected from each of the three performance metrics groups, which can consider the performance with an imbalanced dataset:

- Threshold metrics: F1 score, precision and recall;
- Probability metrics: Brier score;
- Ranking metrics: ROC AUC, PR AUC and average precision.

In addition to threshold metrics, confusion matrices are computed to determine the factors that influenced threshold metric scores.

2.6.1. Threshold Metrics

The confusion matrix shows the number of cases correctly and incorrectly categorised by the model [54]. The rows of the confusion matrix represent actual class, while columns represent predicted classes. Like threshold metrics, the confusion matrix requires setting a threshold to classify cases as positive or negative. The best possible model would have values only in true positive and true negative cells, while both false positive and false negative would be zero.

Table 6. Confusion matrix

	Predicted negative	**Predicted positive**
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

The precision score indicates the proportion of predicted positive class, which is actually positive [77]. The best value of the metric is 1, which means that all predicted positive cases were identified correctly, while the worst value is 0.

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

The recall score represents the ability of the model to identify all the positive cases [77]. The best possible value is 1, which means that the model managed to identify all positive classes, while the worst value is 0.

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

The F1 score could be interpreted as the weighted mean of the precision and recall [78]. The best possible value of the F1 score is 1, while the worst is 0. Precision and recall equally contribute to the F1 score.

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

The F1 score will be high only if both precision and recall are high. Usually, when the precision is higher, the recall will be lower and vice versa [54]. Thus, the F1 score was selected because it indicates how the model can predict both classes (defaulted and non-defaulted observations). The best value of the F1 score would be 1, while the worst – 0.

2.6.2. Probability Metrics

Brier score (also called mean squared error (MSE)) was introduced by G. W. Brier [79]. The following formula describes the Brier score:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 \quad (12)$$

Where m is the number of observations, y_i – actual probability and $\hat{y_i}$ – predicted probability. Brier score does not require classifying predictions into classes. The score can range from 0 to 1, where 0 is the best possible value. The score penalises for strong deviations of predicted probabilities from true probabilities. The Brier score was selected as it indicates how accurate are predicted probabilities.

2.6.3. Ranking Metrics

ROC AUC metric is commonly used in credit risk prediction studies (for example, used in [4], [48] and [40]). The metric is derived by estimating the area under the ROC curve. ROC curve plots the true positive rate (TPR) (also known as recall, which was presented above) against the false positive rate (FPR) at different thresholds [54]. The false positive rate is estimated as follows:

$$FPR = \frac{FP}{FP + TN} \quad (13)$$

Usually, there is a trade-off between TPR and FPR – the higher TPR, the lower FPR and vice versa [54]. The best ROC AUC score equals 1, while the worst is 0.5.

Similarly to ROC AUC, PR AUC is derived by estimating the area under the precision-recall (PR) curve. PR curve plots precision score against recall score at different thresholds. High precision is preferred when the cost of false positives is high, while high recall is preferred when there is a need to identify all positive cases and the cost of false positives is low [54]. The best possible value of PR AUC is 1, while the worst is 0 [4].

The average precision metric is similar to PR AUC, but it summarises a PR curve as the weighted mean of precisions obtained at each threshold, with the increase in recall from the preceding threshold used as the weight[8].

2.7. Managing Imbalanced Dataset

2.7.1. Dataset Balancing

The dataset used in this research has only 4.3% of defaulted observations, thus, it is imbalanced. As described in the literature review, class imbalance in the dataset might lead to worse performance of

[8] https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics

AI algorithms, therefore, it was decided to apply resampling to balance a dataset. A comparison of specific resampling techniques is out of the scope of this thesis; thus, the technique selection was done based on a literature overview.

It was decided to select resampling techniques over cost-sensitive methods, as resampling can be applied to any AI algorithm. Furthermore, since the dataset used in this study is relatively small, any strategies involving under-sampling would not be suitable. Thus, it was decided to choose SMOTE technique to balance a dataset, which based on the literature review is superior to random oversampling. The SMOTE algorithm works by (1) drawing a random sample from the minority class; (2) identifying k nearest neighbours in the sample; (3) finding the difference between the feature vector and its neighbour; (4) multiplying the difference by a random number ranging from 0 to 1; and (5) adding the multiplied difference to the feature vector [53]. The optimal number of k nearest neighbours for each model were selected during the hyperparameter tuning process by considering values between 2 and 7. The selected values in each model are presented in section 3.

Sometimes oversampling might lead to overfitting and consequently worse model performance [53]. Thus, in this study, different proportions of oversampling were tested with each model to find an optimal balance between the majority and minority classes. Five ratios between minority and majority class - 0.1, 0.3, 0.5, 0.7 and 1 were assessed in the hyperparameter tuning process. The selected optimal ratio in each model is presented in section 3.

2.7.2. Optimal Threshold Selection

Frequently, a classification threshold of 0.5 is unsuitable when a dataset is imbalanced [80]. By default, the threshold for classification into classes in case of binary outcome is 0.5 in all algorithms used in this study. This means that prediction will be classified as 0 (non-defaulted), if the probability predicted by a model is lower than 0.5 and when the probability is 0.5 or higher, the prediction will be classified as 1 (defaulted). Probability and ranking metrics do not require setting a threshold to evaluate the model's performance and consider the outcome of the model in probabilities format, but threshold metrics, such as the F1 score, evaluate the performance considering predicted classes. To find an optimal threshold for each model, F1 scores were calculated for each point in the PR curve and the point where the F1 score was highest was selected as the optimal threshold. Optimal thresholds for each model are provided in section 3.

2.8. Optimal Hyperparameters Selection

Optimal hyperparameters were selected using grid search. It was decided to use grid search with stratified 15-fold cross-validation on the train set due to a limited number of observations in the dataset. Using stratified k-fold cross-validation the dataset is divided into equal size k mutually exclusive subsets, also called folds [69]. The model is trained and evaluated k times. Every time the model selects a different fold for evaluation and uses the remaining folds for training. The folds are made by preserving the proportion of both classes. Grid search was performed using the scikit-learn library. During the grid search process, it was decided to consider only ROC AUC and average precision metrics, which do not require setting a threshold.

The best model was selected based on the mean average precision score of validation dataset. Due to limited computer resources, optimal hyperparameter values were found by considering two to three hyperparameters in every generation rather than exploring all hyperparameters at once. The selected hyperparameters for tuning each model and the optimal values of hyperparameters selected during the grid search process are presented in section 3.

2.9. Measuring Variable Importance

All models chosen for the research section are impossible to interpret without additional processing (they are known as black-box models). However, user trust in the model, which is typically gained by understanding model decisions, is a crucial aspect for the model to be used in practice [4]. Several methods are proposed in the literature for explaining black box models, such as LIME, DeepLIFT, Layer-Wise Relevance Propagation, Classic Shapley Value Estimation and Shapley Additive Explanation (SHAP) values [81]. S. M. Lundberg and S. Lee [81] compared LIME and DeepLIFT with SHAP values and concluded that results provided by SHAP values are more consistent with human intuition and have better computational performance. Furthermore, SHAP values are a unified framework which can be applied to different AI methods. Because of these advantages, it was decided to use SHAP values to determine which values influenced the results of the best-performing models.

SHAP values explain the contribution of each independent variable to a particular prediction by measuring how much each independent variable changes the prediction compared to its expected value [81]. SHAP values are based on Shapley values from cooperative game theory, which assigns a value to each player in a coalition game based on their contribution to the overall outcome [81].

2.10. Other Transformations

2.10.1. Data Scaling

Some of the features in the dataset have different scales. Most AI algorithms expect numerical variables to have similar scales and not perform well when the scales are significantly different [54]. The most common approaches for equalising the scales of numerical features include normalisation and standardisation. Using normalisation, values are rescaled to range from 0 to 1. In comparison, standardisation subtracts the variable mean from values and divides them by the variable's standard deviation [54]. Standardisation is less sensitive to outliers compared to normalisation; however, conventional standardisation uses mean and standard deviation statistics, thus, the outcome would still be affected by outliers. As described in section 2.2, the dataset used in this study includes some values which could be classified as outliers. Therefore, before building the models, numerical features were centred and scaled using the RobustScaler function from the scikit-learn library, which uses statistics robust to outliers. The function scales feature according to indicated quantile range and centres features by removing the median. After data analysis, a quantiles 0.1 to 0.9 were selected for data scaling.

2.10.2. Encoding of Categorical Features

The dataset used in this study has four categorical features - Employees group, Location, Legal Form_aggregated and NACE Code_letter. AI methods, such as SVM, cannot process categorical features [54], therefore one-hot encoding technique (OneHotEncoder function from the scikit-learn library) was applied to convert categorical features to binary columns. Some categorical features, such as NACE Code_letter, have a relatively small number of observations in some categories. A dummy column with a small number of 1 would add little value to the model's performance but would increase the dataset's dimension and the model's complexity. Thus, if the category contained less than

10% of train set observations, it was merged into one column with other infrequent categories. Four missing values in NACE Code_letter, described in section 2.3, were also classified as infrequent categories.

3. Research Results

This section discusses model development and performance results. First, the model development based on each AI approach is described, and then the results of all selected models are compared. Each AI model was developed on two datasets – the dataset including alternative and financial data and the dataset including only financial data, to assess if the inclusion of alternative data influences prediction performance. Also, the most important features of best-performed models are examined to support the conclusions. Finally, the section is finished with a discussion of research limitations.

3.1. Random Forest

The random forest models were built using the scikit-learn library. Five hyperparameters of random forest - maximum depth, maximum features, minimum samples needed to split a node, minimum samples in a leaf node and the number of estimators were considered while searching for an optimal random forest model as summarised in Table 7. In addition, different values of SMOTE oversampling proportion and a number of k nearest neighbours were tested as described in section 2.

Table 7. Selected hyperparameters of random forest

Hyperparameter	Default value	Description
Maximum depth	None	Maximum size of each tree in the model. Smaller values prevent overfitting
Maximum features	sqrt (square root of total number of features)	The maximum number of features considered for each split. A lower value reduces the variance
Minimum samples needed to split a node	2	The minimum number of samples needed to split an internal node. Larger values prevent overfitting
Minimum samples in a leaf node	1	The smallest number of samples required in a leaf node. Larger values prevent overfitting
Number of estimators	100	The number of trees in the random forest

Various values between 2 and 10 were considered during the maximum depth tuning process. Both models (model trained with both alternative and financial data and model trained with only financial data) achieved the best validation score using a maximum depth value of 3. During the maximum features tuning procedure, various proportions between 0.05 and 1 were tested. The best validation score was achieved using a maximum features proportion of 0.07 for a model with alternative data and a proportion of 0.15 for a model with only financial data. Multiple values between 2 and 50 were considered as minimum samples needed to split a node during the hyperparameter tuning process. The best validation score was obtained with a value of 15 for a model containing both alternative and financial data and a value of 10 for a model containing solely financial data. Various values between 1 and 15 were tested as minimum samples in a leaf node. The best validation score was reached using 10 minimum samples in a leaf node for both models. Values between 100 and 1000 of the number of estimators were tested during hyperparameter tuning procedure. The best validation score was achieved when using 700 trees in both models.

A model with alternative and financial data achieved the best results when the proportion of minority to majority class was 0.5 and the number of k nearest neighbours was 3, while the model with only financial data performed better when the proportion was 0.7 and the number of k nearest neighbours was 5. The table below summarises tested and selected hyperparameters values.

Table 8. Selected hyperparameter values in random forest models

Hyperparameter	**Alternative and financial data**	**Financial data**	**Values tested during hyperparameter tuning**
Maximum depth	3	3	2, 3, 4, 6, 8, 10, none
Maximum features	0.07	0.15	1.0, 0.7, 0.3, 0.1, 0.07, 0.05, sqrt
Minimum samples needed to split a node	15	10	2, 5, 10, 15, 20, 30, 50
Minimum samples in a leaf node	10	10	1, 2, 4, 6, 8, 10, 15
Number of estimators	700	700	100, 300, 500, 700, 1000
Oversampling proportion (SMOTE)	0.5	0.7	0.1, 0.3, 0.5, 0.7, 1
Number of k nearest neighbours (SMOTE)	3	5	2, 3, 4, 5, 6, 7

The average precision and ROC AUC scores of the train and validation sets presented in the table below indicate that both models might be overfitted, as the train scores are higher than the validation scores. However, when trying to reduce the gap between the train and validation sets, the validation set score decreased significantly, hence it was decided to choose the following models.

Table 9. The cross-validation results of selected random forest models

	Alternative and financial data		**Financial data**	
	Train	**Validation**	**Train**	**Validation**
ROC AUC	0.86	0.77	0.83	0.69
Average precision	0.41	0.35	0.30	0.17

The test set scores show that the model using alternative and financial data performed better based on all metrics. The results indicate that the model with alternative and financial data is superior in distinguishing defaulted and non-defaulted observations and minimising incorrectly identified cases. Higher recall scores than precision scores in both models show that they tend to identify more defaulted observations while misclassifying non-defaulted observations. All scores are provided in the table below. The best scores are highlighted in green.

Table 10. The test set results of selected random forest models

	Alternative and financial data	**Financial data**
F1	0.38	0.29
ROC AUC	0.85	0.78
Precision	0.35	0.24
Recall	0.43	0.38
PR AUC	0.37	0.17
Average precision	0.38	0.18
Brier score	0.08	0.13
Selected threshold	*0.41*	*0.54*

Table 11. The confusion matrix of selected random forest models

Alternative and financial data			**Financial data**		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	677	17	Actual 0	668	26
Actual 1	12	9	Actual 1	13	8

The ROC curve of the model incorporating alternative data indicates slightly better performance compared to the model based on solely financial data. However, both models outperform random guessing in detecting defaulted and non-defaulted instances, and they perform well in terms of capacity to discriminate between both classes. The shape of PR curves shows that the model based on alternative and financial data is superior to the model with only financial data. However, the shape of the PR curves of both models reveals that it is more difficult for the models to detect defaulting observations than non-defaulting observations.

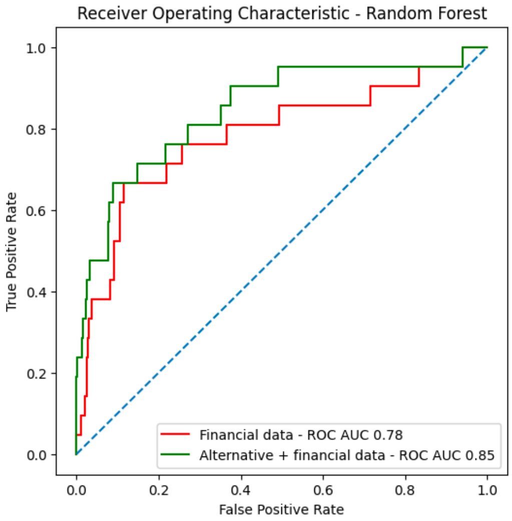

Fig. 13. ROC curves of selected random forest models

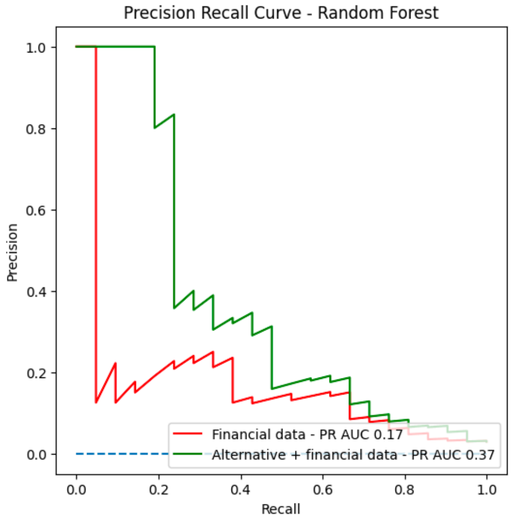

Fig. 14. PR curves of selected random forest models

3.2. Gradient Boosting

Three different gradient boosting algorithms were tested in this study: gradient boosting, LightGBM and XGBoost. The gradient boosting models were built using the scikit-learn library, the LightGBM models were built using a lightgbm library and the XGBoost models were built using an xgboost library. Five different hyperparameters of each model were considered (gradient boosting: maximum depth, number of estimators, learning rate, minimum samples in a leaf node and maximum features; LightGBM: learning rate, number of estimators, maximum depth, minimum samples in a leaf node and maximum number of leaves; XGBoost: learning rate, maximum depth, minimum node weight, subsample proportion and minimum loss reduction) to find the best-performing models. In addition to method-specific hyperparameters, SMOTE oversampling proportion and a number of k nearest neighbours were considered in the hyperparameter tuning process of each model.

Table 12. Selected hyperparameters of gradient boosting

Hyperparameter	**Default value**	**Description**
Maximum depth	3	Maximum size of each tree in the model. Smaller values prevent overfitting
Number of estimators	100	Number of weak learners in a model
Learning rate	0.1	Controls the contribution of each tree to the final prediction. Smaller values prevent overfitting
Minimum samples in a leaf node	1	The smallest number of samples required in a leaf node. Larger values prevent overfitting
Maximum features	1.0	The maximum number of features considered for each split. A lower value reduces the variance

Table 13. Selected hyperparameters of LightGBM

Hyperparameter	**Default value**	**Description**
Learning rate	0.1	Controls the contribution of each tree to the final prediction. Smaller values prevent overfitting
Number of estimators	100	Number of weak learners in a model
Maximum depth	None	Maximum size of each tree in the model. Smaller values prevent overfitting
Minimum samples in a leaf node	20	The smallest number of samples required in a leaf node. Larger values prevent overfitting
Maximum number of leaves	31	Maximum number of leaves in a base learner. This hyperparameter controls the complexity of each tree. Large values may lead to better accuracy but also overfitting

Table 14. Selected hyperparameters of XGBoost

Hyperparameter	**Default value**	**Description**
Learning rate	0.3	Controls the contribution of each tree to the final prediction. Smaller values prevent overfitting
Maximum depth	6	Maximum size of each tree in the model. Smaller values prevent overfitting
Minimum node weight	1	The minimum sum of instance weight in each node. Larger values prevent overfitting
Subsample proportion	1	The proportion of the train set used in every boosting iteration. Smaller values prevent overfitting
Minimum loss reduction	0	The minimum loss reduction needed for further node splits. Larger values prevent overfitting

The maximum depth was tuned in all three algorithms. Various values between 2 and 10 were considered during the tuning procedure. The best validation score was achieved using a value of 3 for both gradient boosting models and LightGBM and XGBoost models trained with alternative data. LightGBM and XGBoost models with solely financial data achieved the best results with a maximum depth value of 4.

The learning rate was also tuned in all three algorithms, considering multiple values between 0.0001 and 0.3. The highest validation score was achieved when using a learning rate value of 0.0001 for both gradient boosting models, a value of 0.01 for both LightGBM models, a value of 0.01 for the XGBoost model with alternative data and a value of 0.001 for the XGBoost model containing solely financial data.

Number of estimators and minimum samples in a leaf node were tuned in gradient boosting and LightGBM models. Multiple values between 100 and 400 were explored during the number of estimators tuning procedure. The best validation score was reached using 100 trees in both gradient boosting models and 300 trees in both LightGBM models. Various values between 1 and 40 were considered during the minimum samples in a leaf node tuning process. The best gradient boosting validation scores were achieved using a value of 8 for a model with alternative data and 20 for a model containing solely financial data. While the best LightGBM validation scores were reached when using 10 minimum samples in a leaf node for a model with alternative data and 30 samples in the model with financial data.

The maximum features were tuned only in the gradient boosting model. Various proportions between 0.05 and 1 were explored during the tuning process. The best validation score was obtained with a proportion of 0.07 for a model containing both alternative and financial data and 0.1 for a model containing solely financial data.

The maximum number of leaves was tuned only in LightGBM. Various values between 5 and 31 were tested during the hyperparameter tuning procedure. The best validation score was obtained with a value of 7 for a model trained with alternative data and 5 for a model with solely financial data.

The minimum node weight, subsample proportion and minimum loss reduction were tuned only in XGBoost. Various values between 1 and 10 were considered during the minimum node weight tuning process. Both models achieved the best validation score using a minimum node weight of 5. Various values between 0.1 and 1 were tested during the subsample proportion tuning procedure. The best validation score was achieved when using a subsample proportion of 1 for a model with alternative data and 0.7 for a model with only financial data. Multiple minimum loss reduction values between 0 and 0.5 were considered during the hyperparameter tuning process. The best validation score was obtained with a value of 0.3 for a model containing both alternative and financial data and 0.1 for a model containing solely financial data.

The gradient boosting model trained with a dataset containing both alternative and financial data achieved the best results when the proportion of minority to majority class was 1 and the number of k nearest neighbours was 2. The gradient boosting model trained with only financial data achieved the best results when the ratio between minority and majority classes was also 1 and the number of k nearest neighbours was 3. LightGBM model trained with alternative data achieved the best results when the proportion of minority to majority class was 0.5 and the number of k nearest neighbours was 5, while the model trained with only financial data performed better when the class proportion was 0.7 and the number of k nearest neighbours was 3. XGBoost model trained with both alternative and financial data performed the best when the proportion of minority to majority class was 0.5 and the number of k nearest neighbours was 6, while the model trained with solely financial data performed better when the proportion between classes was 0.3 and the number of k nearest neighbours was 5. The tables below summarise tested and selected hyperparameters values.

Table 15. Selected hyperparameter values in gradient boosting models

Hyperparameter	**Alternative and financial data**	**Financial data**	**Values tested during hyperparameter tuning**
Maximum depth	3	3	2, 3, 4, 5, 6, 8, 10
Number of estimators	100	100	100, 200, 300
Learning rate	0.0001	0.0001	0.0001, 0.001, 0.01, 0.1
Minimum samples in a leaf node	8	20	1, 2, 4, 6, 8, 10, 20, 30
Maximum features	0.07	0.1	1.0, 0.7, 0.3, 0.1, 0.07, 0.05
Oversampling proportion (SMOTE)	1	1	0.1, 0.3, 0.5, 0.7, 1
Number of k nearest neighbours (SMOTE)	2	3	2, 3, 4, 5, 6, 7

Table 16. Selected hyperparameter values in LightGBM models

Hyperparameter	**Alternative and financial data**	**Financial data**	**Values tested during hyperparameter tuning**
Learning rate	0.01	0.01	0.0001, 0.001, 0.01, 0.1
Number of estimators	300	300	100, 200, 300, 400
Maximum depth	3	4	2, 3, 4, 5, 6, 7, 8, 10, none
Minimum samples in a leaf node	10	30	5, 10, 15, 20, 30, 40
Maximum number of leaves	7	5	5, 7, 10, 15, 20, 31
Oversampling proportion (SMOTE)	0.5	0.7	0.1, 0.3, 0.5, 0.7, 1
Number of k nearest neighbours (SMOTE)	5	3	2, 3, 4, 5, 6, 7

Table 17. Selected hyperparameter values in XGBoost models

Hyperparameter	**Alternative and financial data**	**Financial data**	**Values tested during hyperparameter tuning**
Learning rate	0.01	0.001	0.0001, 0.001, 0.01, 0.1, 0.3
Maximum depth	3	4	2, 3, 4, 5, 6, 7, 8, 10
Minimum node weight	5	5	1, 5, 7, 10
Subsample proportion	1	0.7	0.1, 0.3, 0.5, 0.7, 1
Minimum loss reduction	0.3	0.1	0, 0.1, 0.3, 0.5
Oversampling proportion (SMOTE)	0.5	0.3	0.1, 0.3, 0.5, 0.7, 1
Number of k nearest neighbours (SMOTE)	6	5	2, 3, 4, 5, 6, 7

The average precision and ROC AUC scores of the train and validation sets provided in the table below show that all models might be overfitted, as the train scores are higher than the validation scores. However, when trying to reduce the gaps between the train and validation sets, the validation set scores decreased significantly, hence it was decided to keep the following models.

Results with the validation set of gradient boosting and LightGBM models are similar, while the results of XGBoost are slightly inferior, thus, it was decided not to include XGBoost in further analysis. Furthermore, gradient boosting appeared less overfitted than LightGBM; thus, it was decided to address only the gradient boosting results further.

Table 18. The cross-validation results of selected gradient boosting models

	Alternative and financial data		**Financial data**	
	Train	**Validation**	**Train**	**Validation**
Gradient boosting				
ROC AUC	0.86	0.76	0.83	0.69
Average precision	0.39	0.34	0.28	0.16
LightGBM				
ROC AUC	0.89	0.76	0.90	0.69
Average precision	0.48	0.33	0.36	0.15
XGBoost				
ROC AUC	0.82	0.71	0.90	0.68
Average precision	0.32	0.29	0.44	0.16

The test set scores show that the model using both alternative and financial data achieved better performance based on all metrics except the Brier score. The Brier score was equal in both models. The findings show that the model using alternative data outperforms the model with only financial data in distinguishing between both classes and minimising mistakenly detected classes. All scores are provided in the table below. The best scores are highlighted in green.

Table 19. The test set results of selected gradient boosting models

	Alternative and financial data	**Financial data**
F1	0.43	0.32
ROC AUC	0.86	0.80
Precision	0.38	0.30
Recall	0.48	0.33
PR AUC	0.37	0.18
Average precision	0.38	0.19
Brier score	0.25	0.25
Selected threshold	*0.50*	*0.50*

Table 20. The confusion matrix of selected gradient boosting models

Alternative and financial data			**Financial data**		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	678	16	Actual 0	678	16
Actual 1	11	10	Actual 1	14	7

Both models' ROC curves are similar, showing that they outperform random guessing in finding defaulted and non-defaulted cases and perform well in discriminating between both groups. The shape of the PR curves demonstrates that the model based on alternative and financial data outperforms the model based solely on financial data. However, the shapes of the PR curve of both models reveals that it is more difficult for the models to detect defaulting observations than non-defaulting observations.

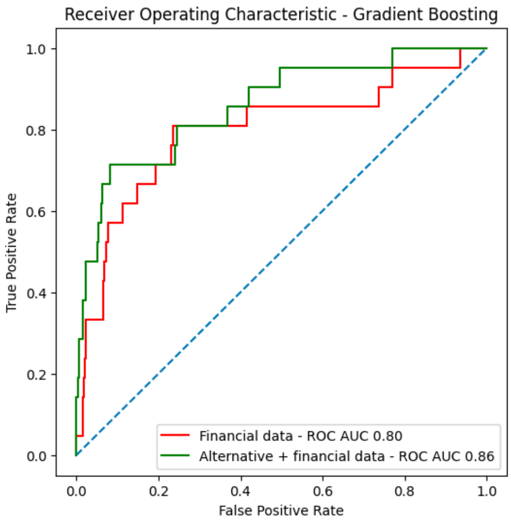

Fig. 15. ROC curves of selected gradient boosting models

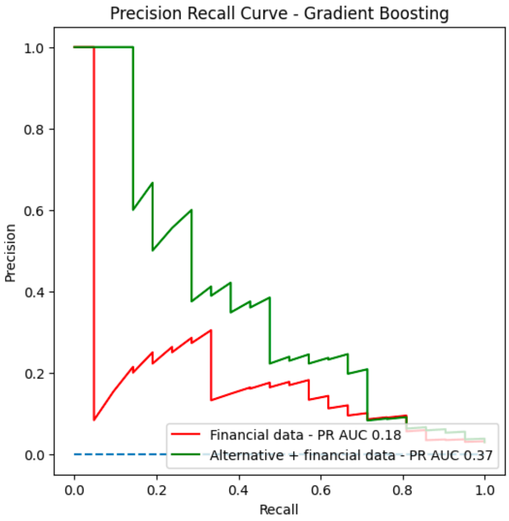

Fig. 16. PR curves of selected gradient boosting models

3.3. Support Vector Machines

The SVM models were built using the scikit-learn library. Regularisation parameter, kernel, kernel coefficient as well as SMOTE oversampling proportion and the number of k nearest neighbours were considered to find the best performing SVM models.

Table 21. Selected hyperparameters of SVM

Hyperparameter	**Default value**	**Description**
Regularisation parameter	1.0	Regularisation parameter. With lower values larger margin would be selected, which might lead to more misclassified observations
Kernel	rbf	Kernel type
Kernel coefficient	scale	Kernel coefficient for RBF, polynomial and sigmoid kernels, which controls the influence of an observation. The higher value of the kernel coefficient means that only those observations which are close to the hyperplane will be considered

Various values of regularisation parameter between 0.0001 and 100 were explored during the tuning procedure. The best validation score was attained using a regularisation parameter of 0.05 for a model with both alternative and financial data and 0.0001 for a model with only financial data. During the kernel selection procedure, linear, polynomial (with degrees 2 and 3), RBF and sigmoid kernel types were tested. The best validation score was achieved using an RBF kernel for models with both datasets. Various values of kernel coefficient between 0.01 and 0.15 were tested during the tuning procedure. The best validation score was achieved using a value of 0.01 with RBF kernel for models with both datasets. When the ratio of minority to majority class was 0.5 and the number of k nearest neighbours was 6, a model with alternative data performed best, whereas a model with only financial data performed better when the proportion was 0.7 and the number of k nearest neighbours was 5. The table below summarises the values of the tested and chosen hyperparameters.

Table 22. Selected hyperparameter values in SVM models

Hyperparameter	**Alternative and financial data**	**Financial data**	**Values tested during hyperparameter tuning**
Regularisation parameter	0.05	0.0001	0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 1, 2, 3, 5, 10, 100
Kernel	RBF	RBF	linear, polynomial, RBF, sigmoid
Kernel coefficient	0.01	0.01	scale, 0.01, 0.05, 0.1, 0.15
Oversampling proportion (SMOTE)	0.5	0.7	0.1, 0.3, 0.5, 0.7, 1
Number of k nearest neighbours (SMOTE)	6	5	2, 3, 4, 5, 6, 7

The average precision scores of the train and validation sets presented in the table below indicate that both models should be neither underfitted nor overfitted, as train and validation scores are equal. However, based on ROC AUC values, both models might be overfitted, as the train scores are higher than the validation scores. However, when trying to reduce the gap between the train and validation set, the validation set score decreased significantly, hence it was decided to choose the following models.

Table 23. The cross-validation results of selected SVM models

	Alternative and financial data		**Financial data**	
	Train	**Validation**	**Train**	**Validation**
ROC AUC	0.83	0.73	0.68	0.62
Average precision	0.20	0.20	0.15	0.15

The test set scores of models show that the model incorporating alternative data achieved better performance based on all metrics except for the recall. The recall score in both models was equal. The findings suggest that the model using alternative data outperforms the model with only financial data in distinguishing between defaulted and non-defaulted observations and minimising mistakenly detected cases. However, both models correctly identified the same number of defaulted observations. All scores are provided in the table below. The best scores are highlighted in green.

Table 24. The test set results of selected SVM models

	Alternative and financial data	**Financial data**
F1	0.35	0.19
ROC AUC	0.80	0.72
Precision	0.46	0.15
Recall	0.29	0.29
PR AUC	0.31	0.15
Average precision	0.32	0.16
Brier score	0.06	0.23
Selected threshold	*0.64*	*0.70*

Table 25. The confusion matrix of selected SVM models

Alternative and financial data			**Financial data**		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	687	7	Actual 0	659	35
Actual 1	15	6	Actual 1	15	6

The ROC curve shapes of both models show that the model including alternative data is slightly better at discriminating between different classes. However, both models outperform random guessing in detecting defaulted and non-defaulted instances. The shape of PR curves shows that the model based on both alternative and financial data is superior to the model with only financial data. Nevertheless, the shape of the PR curves of the models demonstrates that it is more difficult for the models to recognise defaulting observations than non-defaulting observations.

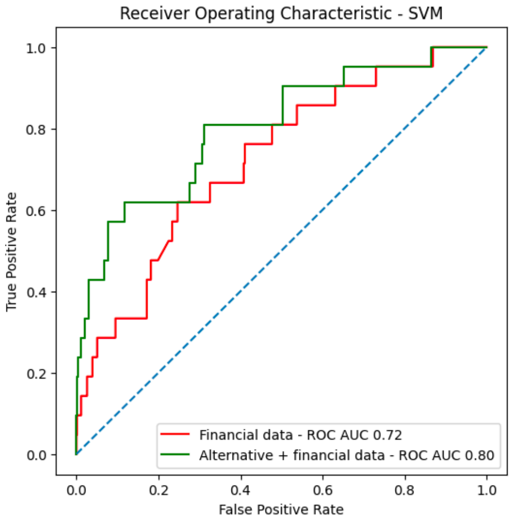

Fig. 17. ROC curves of selected SVM models

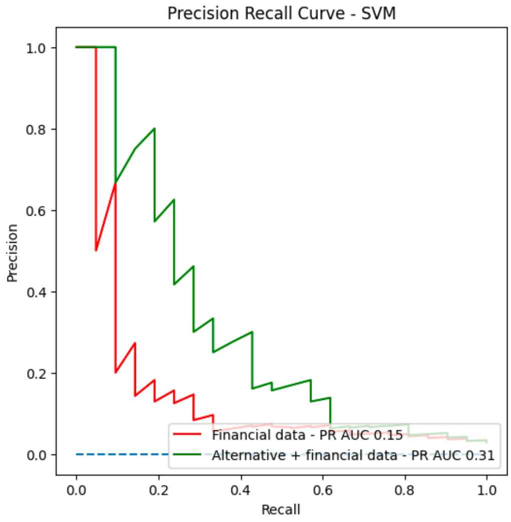

Fig. 18. PR curves of selected SVM models

3.4. Comparison of the Results and Discussion

The results of the models incorporating alternative data show that gradient boosting and random forest were the best-performing models, with gradient boosting having the highest ROC AUC, F1 and recall scores and both having the highest PR AUC and average precision scores. SVM was superior to the random forest and gradient boosting based on precision score. However, its recall score was relatively low, indicating that SVM was more conservative while predicting default cases. Also, SVM had the lowest Brier score, thus, its probabilistic predictions were the most accurate.

Table 26. Results comparison of the models with alternative and financial data

	Random forest	**SVM**	**Gradient boosting**
F1	0.38	0.35	0.43
ROC AUC	0.85	0.80	0.86
Precision	0.35	0.46	0.38
Recall	0.43	0.29	0.48
PR AUC	0.37	0.31	0.37
Average precision	0.38	0.32	0.38
Brier score	0.08	0.06	0.25

The performance results of the models with only financial data indicate that the gradient boosting and random forest models also were the best-performing models. Gradient boosting had the best overall performance, with the highest F1, ROC AUC, precision, PR AUC, average precision scores and second-highest recall score. The random forest model had the highest recall score of 0.38 and the lowest Brier score of 0.13, indicating that the model correctly identified more defaulted observations and its probabilistic predictions were more accurate.

Table 27. Results comparison of the models with financial data

	Random forest	**SVM**	**Gradient boosting**
F1	0.29	0.19	0.32
ROC AUC	0.78	0.72	0.80
Precision	0.24	0.15	0.30
Recall	0.38	0.29	0.33
PR AUC	0.17	0.15	0.18
Average precision	0.18	0.16	0.19
Brier score	0.13	0.23	0.25

Overall, the results show that models developed with both alternative and financial data generally performed better than those developed on the dataset with only financial data. The F1 scores of models trained with alternative and financial data were generally higher, showing that the models performed better in correctly identifying default cases and minimizing false default cases. The ROC AUC scores (which measure the ability of the model to distinguish between risky and non-risky companies) of models incorporating alternative data were also higher compared to models trained with only financial data. The PR AUC and average precision scores of the models including alternative data were also higher than those trained only with financial data, indicating that the models were more effectively ranking default cases against the non-default cases when the dataset included alternative variables. The Brier score of the models trained with the full dataset, which measures the overall accuracy of the model's predicted probabilities, was generally lower, hence, the models' predicted probabilities were more accurate than models trained only with financial data. These findings are in line with such research works as [7], [48] and [61], which also found that variables based on general characteristics of the company and payment behaviour data improve prediction performance.

The random forest and gradient boosting models showed the most significant increases in performance results when the alternative data was included in the dataset, indicating that these models performed better when the complexity of the dataset was increased. The increase in gradient boosting performance when the complexity of the dataset increase was also found by B. Lextrait [4].

However, most of the models employed in this study were overfitted, therefore, they might not perform well when applied to datasets with different companies. Also, although ROC AUC scores were generally relatively high, F1, precision, recall, PR AUC and average precision scores were relatively low even of best-performing models, which indicates that the models lack the ability to identify defaulted observations correctly. Both shortcomings might be related to the relatively small and imbalanced sample used in this study and the results might have been better if the sample size had been larger.

SHAP values were estimated for the best-performed models to understand which features were the most influential. The figures below present the 20 most important variables ranked based on their importance from the most important (at the top) to less important (at the bottom). Blue dots with positive SHAP values indicate that the lower value of the feature increases credit risk, while red dots with positive SHAP values show that the higher value of the feature increases credit risk.

Four of the five most influential variables in gradient boosting and random forest models trained with alternative data were based on payment behaviour data. Figures show that lower values of overdue payments were related to lower default risk, while higher values of overdue payments were related to higher default risk. Other alternative data-based variables, such as interest rates, GDP growth and number of employees, were among the top 10 most influential variables in both models. Lower GDP growth and a small number of employees (less than 10) were related to higher risk, while lower interest rates were related to lower risk. Considering financial variables, the current ratio and debt ratio were among the top 10 most influential variables in both models. The SHAP values show that low values of the current ratio and high values of the debt ratio increase credit risk.

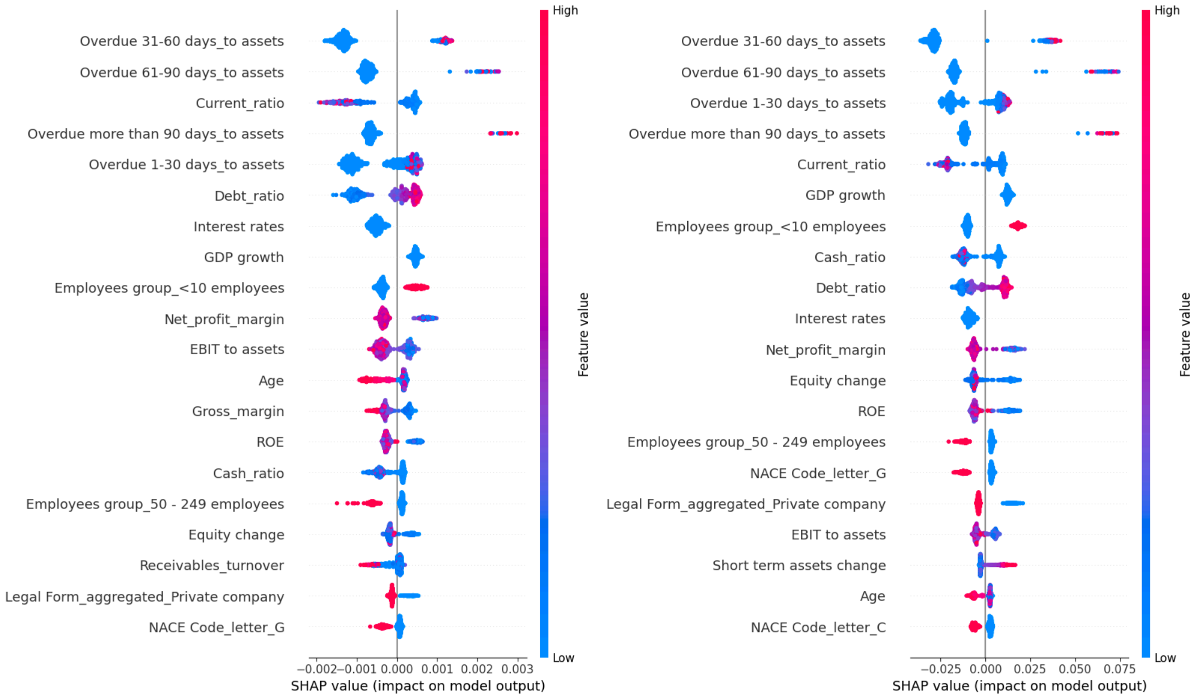

Fig. 19. Variable importance in gradient boosting (on the left) and random forest (on the right) models, including alternative data

Feature importance was also estimated for the gradient boosting and random forest models trained with only financial data. Consistent with the above results, current ratio and debt ratio were among the most important financial features.

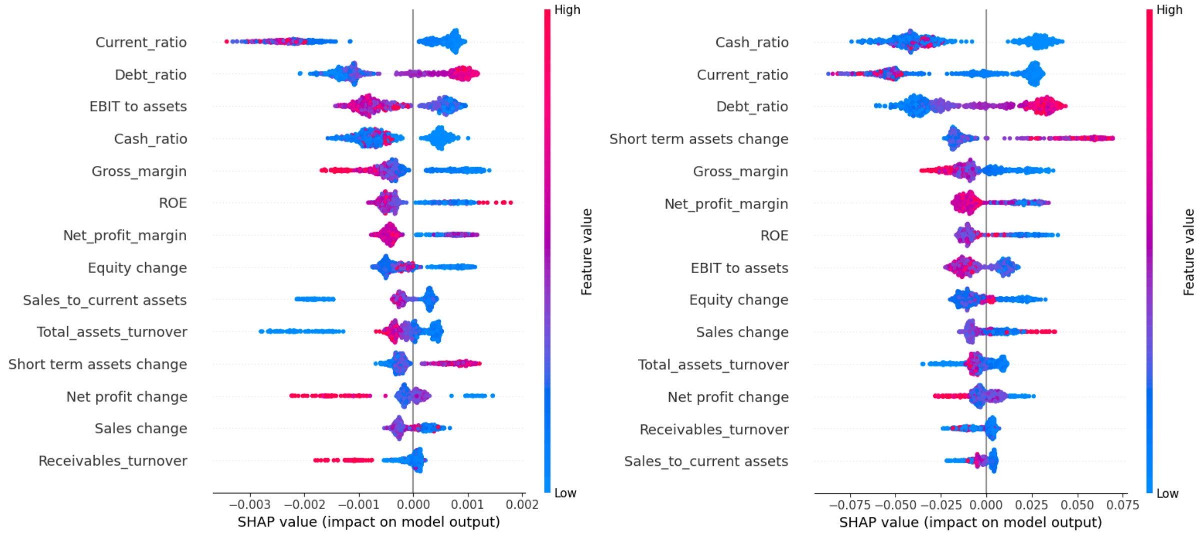

Fig. 20. Variable importance in gradient boosting (on the left) and random forest (on the right) models, including only financial data

Overall, the research findings suggest that alternative data could improve SMEs' credit risk prediction results. As a result, it is recommended for practitioners to include alternative data in their credit risk prediction models to increase the accuracy of credit risk projections, especially when the available financial data are limited. However, the inclusion of additional features increases the dimension of the dataset, thus, it is recommended to ensure that the sample used for model development is sufficiently large and to select methods capable of dealing with complex data.

3.5. Additionally Tried Methods

All models achieved relatively sound performance based on ROC AUC scores. However, PR AUC, average precision and F1 scores indicate that models have difficulty identifying defaulted observations. The literature overview presented in section 1 showed that worse performance might be related to high class imbalance, thus, it was decided to try several additional approaches for dealing with class imbalance.

First, it was stated in the literature that in some situations, no oversampling might provide better results than oversampling. Thus, it was tried to train all selected random forest, gradient boosting and SVM models with no oversampling. A comparison of validation set results showed that the average precision score deteriorated for all models trained with alternative data. Random forest trained only with financial data showed no change in average precision score. Gradient boosting trained only with financial data achieved a 0.01 higher average precision score (an improvement from 0.16 to 0.17), while the average precision score of SVM deteriorated.

In addition, it was decided to try a different resampling technique from the imblearn library – SMOTETomek, which combines oversampling and undersampling in all selected random forest, gradient boosting and SVM models. Five ratios between minority and majority class - 0.1, 0.3, 0.5, 0.7 and 1 were tested during the hyperparameter tuning process. The comparison of validation set results showed that for all models trained with alternative data, results of average precision either deteriorated or remained the same. With solely financial data, SVM showed no change in average precision score, while random forest and gradient boosting achieved 0.01 higher average precision scores (an improvement from 0.17 to 0.18 in the random forest and an improvement from 0.16 to 0.17 in the gradient boosting).

Finally, it was tried to increase minority class weight in selected random forest and SVM models instead of applying SMOTE (gradient boosting implementation in the scikit-learn library cannot adjust class weights; therefore, application of class weights in this model was not considered). Five proportions were tested during the hyperparameter tuning process – balanced (inversely proportional to class frequencies ratio), 1:5, 1:10 and 1:15 ratios. Similarly, results with the validation set showed either worse performance or no change in average precision scores of random forest and SVM models with both datasets.

Considering that all these experiments provided either no change, worse results or only minor improvement (up to 0.01 points), it was decided not to change the selected oversampling technique.

3.6. Limitations

The research described in this section is subject to several limitations. These limitations may have affected the accuracy and generalisability of the findings. Thus, the conclusions described above should be interpreted within the context of these limitations.

The main limitations of the research are related to the used dataset. The sample used in the study was relatively small. Because of the small number of SMEs in the sample, the study's findings may not generalise well to other datasets. This could be due to the over-representation of certain industries, such as agriculture and forestry and the absence of others, such as education. Also, the dataset included SMEs located only in Lithuania. Thus, the study's results should be interpreted with caution when applying them to other contexts. In addition, the small sample size may have led to overfitting and reduced prediction performance of the models used in the study. Future studies could address this issue by employing a bigger sample size to increase prediction accuracy and reduce overfitting. In addition, due to the small sample size, several observations of the same company were allowed in the dataset in a few instances. In general, all methods employed in this study require that the data would be independent and identically distributed. Financial results of the same SME can vary significantly from year to year, but repeated inspections of the same company may still violate this assumption. Increasing the sample size and eliminating repeated observations may produce more robust results.

Feature selection in the study was done based only on correlations and the number of selected variables was still relatively high. Further reducing the number of features or selecting the most important variables with more sophisticated methods or using dimension reduction techniques, such as principal component analysis, might improve prediction accuracy.

Conclusions

1. Although access to finance is critical for the success of SMEs, it is usually difficult for them to obtain funding since they are considered riskier than large corporations. The larger risk is often related to a lack of available data for creditors to assess the risk. An overview of studies showed that artificial intelligence and alternative data could be used to solve this issue and increase the financial inclusion of SMEs. The literature review covered multiple credit risk prediction methods, such as MDA, logistic regression, neural networks, SVM, decision trees and ensemble methods, as well as various data types used for assessing credit risk, such as general characteristics of the company, payment behaviour data, industry-specific information, textual information and relational data.
2. The literature overview showed that the most popular statistical and AI methods differed over time, and there is no consensus on which models are the best for credit risk prediction. Thus, random forest, three different gradient boosting algorithms and SVM were chosen for the research part due to their efficiency and potential applicability for the dataset. While general firm characteristics and payment behaviour data appeared to be the most viable alternative data types due to ease of collection, processing, and resistance to manipulation and were chosen for the research part of this thesis.
3. The dataset used in this study included observations of Lithuanian SMEs from the 2015-2020 period. Independent variables included 12 alternative data-based variables and 14 financial data-based variables. The company was considered defaulted on payment if it had more than 90 days overdue payment. The dataset was imbalanced, with only 4.3% of defaulted observations.
4. The results of empirical research showed that including alternative data in credit risk prediction models can increase prediction performance compared to models that only use financial data. According to variable importance analysis, alternative data-based features dominated among the most important variables. Payment behaviour data had the most significant impact of all alternative data-based variables. Other important alternative data-based features included macroeconomic indicators (interest rates and GDP growth) and number of employees in the company. Since alternative data could be used as additional data in credit risk prediction models, the findings of this thesis support the statement that alternative data could help increase SMEs' access to finance. Thus, it is recommended to supplement credit risk prediction models with alternative data to improve the accuracy of credit risk estimates, especially when available financial data are limited.

List of References

1. ALIBHAI, S., BELL, S. and CONNER, G. *What's Happening in the Missing Middle? : Lessons from Financing SMEs.* World Bank, Washington, DC. Mar 29, 2017.

2. HARRIS, T. Quantitative Credit Risk Assessment using Support Vector Machines: Broad Versus Narrow Default Definitions. *Expert Systems with Applications*, 2013, vol. 40, no. 11. pp. 4404-4413 DOI 10.1016/j.eswa.2013.01.044.

3. CHATTERJEE, S. *Modelling Credit Risk.* Bank of England, 2015 ISBN 1756-7270.

4. LEXTRAIT, B. Scaling Up SMEs' Credit Scoring Scope with LightGBM. *Applied Economics*, 2023, vol. 55, no. 9. pp. 925-943 DOI https://doi.org/10.1080/00036846.2022.2095340.

5. GULSOY, N. and KULLUK, S. A Data Mining Application in Credit Scoring Processes of Small and Medium Enterprises Commercial Corporate Customers. *WIREs Data Mining and Knowledge Discovery*, 2019, vol. 9, no. 3. pp. e1299 DOI https://doi.org/10.1002/widm.1299.

6. YOSHINO, N. and TAGHIZADEH-HESARY, F. A Comprehensive Method for the Credit Risk Assessment of Small and Medium-Sized Enterprises Based on Asian Data. *ADBI Working Paper Series*, 2018, vol. 907. Available from: http://hdl.handle.net/10419/222674.

7. CIAMPI, F., CILLO, V. and FIANO, F. Combining Kohonen Maps and Prior Payment Behavior for Small Enterprise Default Prediction. *Small Business Economics*, 2020, vol. 54. pp. 1007-1039 Springer Link. DOI https://doi.org/10.1007/s11187-018-0117-2.

8. European Commission. *Commission Recommendation of 6 may 2003 Concerning the Definition of Micro, Small and Medium-Sized Enterprises (Text with EEA Relevance) (Notified Under Document Number C(2003) 1422).* , May 6, 2003.

9. PETTIT, R.R. and SINGER, R.F. Small Business Finance: A Research Agenda. *Financial Management*, 1985, vol. 14, no. 3. pp. 47-60. Available from: https://www.jstor.org/stable/3665059 DOI https://doi.org/10.2307/3665059.

10. MHLANGA, D. Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment. *International Journal of Financial Studies*, 2021, vol. 9, no. 3. pp. 39 DOI https://doi.org/10.3390/ijfs9030039.

11. DIETSCH, M. and PETEY, J. Should SME Exposures be Treated as Retail Or Corporate Exposures? A Comparative Analysis of Default Probabilities and Asset Correlations in French and German SMEs. *Journal of Banking & Finance*, 2004, vol. 28, no. 4. pp. 773-788 DOI https://doi.org/10.1016/j.jbankfin.2003.10.006.

12. RYBAKOVAS, E. and ŽIGIENĖ, G. Is Artificial Intelligence a Magic Pill Enhancing SMEs Access to Finance?. *2021 IEEE International Conference on Technology and Entrepreneurship (ICTE)*, 2021. pp. 1-6 DOI 10.1109/ICTE51655.2021.9584833.

13. AITKEN, R. 'All Data is Credit Data': Constituting the Unbanked. *Competition & Change*, 2017, vol. 21, no. 4. pp. 274-300 DOI https://doi.org/10.1177/1024529417712830.

14. SADOK, H., SAKKA, F. and EL MAKNOUZI, M.,El Hadi. Artificial Intelligence and Bank Credit Analysis: A Review. *Cogent Economics & Finance*, 2022, vol. 10, no. 1 DOI 10.1080/23322039.2021.2023262.

15. *Principles for the Management of Credit Risk.* Basel Committee on Banking Supervision. 27 September, 2000.

16. BESSIS, J. *Risk Management in Banking.* 4th ed. Wiley, 2015 ISBN 978-1-118-66021-8.

17. ALTMAN, E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 1968, vol. 23, no. 4. pp. 589-609. Available from: https://www.jstor.org/stable/2978933 DOI https://doi.org/10.2307/2978933.

18. OHLSON, J.A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 1980, vol. 18, no. 1. pp. 109-131. Available from: https://www.jstor.org/stable/2490395 DOI https://doi.org/10.2307/2490395.

19. ODOM, M.D. and SHARDA, R. A Neural Network Model for Bankruptcy Prediction. *1990 IJCNN International Joint Conference on Neural Networks*, 1990, vol. 2. pp. 163-168 DOI 10.1109/IJCNN.1990.137710.

20. ZHANG, G., HU, M.,Y., EDDY PATUWO, B. and INDRO, D.,C. Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis. *European Journal of Operational Research*, 1999, vol. 116, no. 1. pp. 16-32 DOI https://doi.org/10.1016/S0377-2217(98)00051-4.

21. CHARALAMBOUS, C., CHARITOU, A. and KAOUROU, F. Comparative Analysis of Artificial Neural Network Models: Application in Bankruptcy Prediction. *Annals of Operations Research*, 2000, vol. 99. pp. 403-425 DOI https://doi.org/10.1023/A:1019292321322.

22. SHIN, K., LEE, T.S. and KIM, H. An Application of Support Vector Machines in Bankruptcy Prediction Model. *Expert Systems with Applications*, 2005, vol. 28, no. 1. pp. 127-135 DOI https://doi.org/10.1016/j.eswa.2004.08.009.

23. HEWA WELLALAGE, N. and LOCKE, S. Factors Affecting the Probability of SME Bankruptcy: A Case Study on New Zealand Unlisted Firms. *Business Journal for Entrepreneurs (Quarterly) June 2012*, 2012. Available from: https://ssrn.com/abstract=2073794.

24. TOBBACK, E., et al. Bankruptcy Prediction for SMEs using Relational Data. *Decision Support Systems*, 2017, vol. 102. pp. 69-81 ISSN 0167-9236. DOI https://doi.org/10.1016/j.dss.2017.07.004.

25. BROWN, K. and MOLES, P. *Credit Risk Management.* , 2014.

26. BEAVER, W.H. Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, 1966, vol. 4. pp. 71-111. Available from: https://www.jstor.org/stable/2490171 DOI https://doi.org/10.2307/2490171.

27. TAM, K.Y. and KIANG, M.Y. Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *Management Science*, 1992, vol. 38, no. 7. pp. 926-947. Available from: https://www.jstor.org/stable/2632376.

28. BELL, T.B. Neural Nets Or the Logit Model? A Comparison of each Model's Ability to Predict Commercial Bank Failures. *Intelligent Systems in Accounting, Finance and Management*, 1997, vol. 6. pp. 249-264 DOI https://doi.org/10.1002/(SICI)1099-1174(199709)6:3%3C249::AID-ISAF125%3E3.0.CO;2-H.

29. HORTA, I.M. and CAMANHO, A.S. Company Failure Prediction in the Construction Industry. *Expert Systems with Applications*, 2013, vol. 40, no. 16. pp. 6253-6257 DOI https://doi.org/10.1016/j.eswa.2013.05.045.

30. ROPEGA, J. The Reasons and Symptoms of Failure in SME. *International Advances in Economic Research*, 2011, vol. 17. pp. 476-483 DOI https://doi.org/10.1007/s11294-011-9316-1.

31. *International Convergence of Capital Measurement and Capital Standards.* Bank for International Settlements, 30 June, 2006.

32. SUN, J. and LI, H. Financial Distress Early Warning Based on Group Decision Making. *Computers & Operations Research*, 2009, vol. 36, no. 3. pp. 885-906 DOI https://doi.org/10.1016/j.cor.2007.11.005.

33. DU JARDIN, P. Bankruptcy Prediction Models: How to Choose the most Relevant Variables?. *Bankers, Markets & Investors*, 2009, no. 98. pp. 39-46.

34. DEAKIN, E.B. A Discriminant Analysis of Predictors of Business Failure. *Journal of Accounting Research*, 1972, vol. 10, no. 1. pp. 167-179. Available from: https://www.jstor.org/stable/2490225 DOI https://doi.org/10.2307/2490225.

35. POMPE, P.P.M. and BILDERBEEK, J. The Prediction of Bankruptcy of Small- and Medium-Sized Industrial Firms. *Journal of Business Venturing*, 2005, vol. 20, no. 6. pp. 847-868 DOI https://doi.org/10.1016/j.jbusvent.2004.07.003.

36. BLUM, M. Failing Company Discriminant Analysis. *Journal of Accounting Research*, 1974, vol. 12, no. 1. pp. 1-25. Available from: https://www.jstor.org/stable/2490525 DOI https://doi.org/10.2307/2490525.

37. KIM, H.S. and SOHN, S.Y. Support Vector Machines for Default Prediction of SMEs Based on Technology Credit. *European Journal of Operational Research*, 2010, vol. 201, no. 3. pp. 838-846 DOI https://doi.org/10.1016/j.ejor.2009.03.036.

38. ORGLER, Y.E. A Credit Scoring Model for Commercial Loans. *Journal of Money, Credit and Banking*, 1970, vol. 2, no. 4. pp. 435-445. Available from: https://www.jstor.org/stable/1991095 DOI https://doi.org/10.2307/1991095.

39. DONG, G., LAI, K.K. and YEN, J. Credit Scorecard Based on Logistic Regression with Random Coefficients. *Procedia Computer Science*, 2010, vol. 1, no. 1. pp. 2463-2468 DOI https://doi.org/10.1016/j.procs.2010.04.278.

40. YIN, C., JIANG, C., JAIN, H.K. and WANG, Z. Evaluating the Credit Risk of SMEs using Legal Judgments. *Decision Support Systems*, 2020, vol. 136. pp. 113364 DOI https://doi.org/10.1016/j.dss.2020.113364.

41. DASTILE, X., CELIK, T. and POTSANE, M. Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey. *Applied Soft Computing*, 2020, vol. 91. pp. 106263 DOI https://doi.org/10.1016/j.asoc.2020.106263.

42. ÓSKARSDÓTTIR, M., et al. The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion using Mobile Phone Data and Social Network Analytics. *Applied Soft Computing*, 2019, vol. 74. pp. 26-39 DOI https://doi.org/10.1016/j.asoc.2018.10.004.

43. PTAK-CHMIELEWSKA, A. Predicting Micro-Enterprise Failures using Data Mining Techniques. *Journal of Risk and Financial Management*, 2019, vol. 12, no. 1. pp. 30 DOI https://doi.org/10.3390/jrfm12010030.

44. FUSTER, A., PLOSSER, M., SCHNABL, P. and VICKERY, J. The Role of Technology in Mortgage Lending, 2018. Available from: http://www.nber.org/papers/w24500 DOI 10.3386/w24500.

45. LAHMIRI, S. Features Selection, Data Mining and Finacial Risk Classification: A Comparative Study. *Intelligent Systems in Accounting, Finance and Management*, 2016, vol. 23, no. 4. pp. 265-275 DOI https://doi.org/10.1002/isaf.1395.

46. RAVI KUMAR, P. and RAVI, V. Bankruptcy Prediction in Banks and Firms Via Statistical and Intelligent Techniques – A Review. *European Journal of Operational Research*, 2007, vol. 180, no. 1. pp. 1-28 DOI https://doi.org/10.1016/j.ejor.2006.08.043.

47. THIEL, D.v. and RAAIJ, W.F.v. Artificial Intelligence Credit Risk Prediction: An Empirical Study of Analytical Artificial Intelligence Tools for Credit Risk Prediction in a Digital Era. *Journal of Risk Management in Financial Institutions*, 2019, vol. 12, no. 3. pp. 268-286.

48. MALAKAUSKAS, A. and LAKŠTUTIENĖ, A. Financial Distress Prediction for Small and Medium Enterprises using Machine Learning Techniques. *Engineering Economics*, 2021, vol. 32, no. 1. pp. 4-14 DOI https://doi.org/10.5755/j01.ee.32.1.27382.

49. YAO, G., HU, X., XU, L. and WU, Z. Using Social Media Information to Predict the Credit Risk of Listed Enterprises in the Supply Chain. *Kybernetes*, 2022 DOI https://doi.org/10.1108/K-12-2021-1376.

50. BURNAEV, E., EROFEEV, P. and PAPANOV, A. Influence of Resampling on Accuracy of Imbalanced Classification. *Proc. SPIE 9875, Eighth International Conference on Machine Vision (ICMV 2015)*, 2015. pp. 987521 DOI https://doi.org/10.1117/12.2228523.

51. HE, H. and GARCIA, E.A. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 2009, vol. 21, no. 9. pp. 1263-1284 DOI 10.1109/TKDE.2008.239.

52. WEISS, G.M., MCCARTHY, K. and ZABAR, B. Cost-Sensitive Learning Vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?, 2007.

53. CHAWLA, N.V., BOWYER, K.W., HALL, L.O. and KEGELMEYER, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *The Journal of Artificial Intelligence Research*, 2002, vol. 16. pp. 321-357 DOI https://doi.org/10.1613/jair.953.

54. GÉRON, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* Second edition ed. O'Reilly, 2019 ISBN 9781492032649.

55. FLOREZ-LOPEZ, R. Effects of Missing Data in Credit Risk Scoring. A Comparative Analysis of Methods to Achieve Robustness in the Absence of Sufficient Data. *The Journal of the Operational Research Society*, 2010, vol. 61, no. 3. pp. 486-501 DOI 10.1057/jors.2009.66.

56. RUBIN, D.B. Inference and Missing Data. *Biometrika*, 1976, vol. 63, no. 3. pp. 581-592 DOI https://doi.org/10.2307/2335739.

57. DONDERS, A.R.T., van der Heijden, Geert J. M. G., STIJNEN, T. and MOONS, K.G.M. Review: A Gentle Introduction to Imputation of Missing Values. *Journal of Clinical Epidemiology*, 2006, vol. 59, no. 10. pp. 1087-1091 DOI 10.1016/j.jclinepi.2006.01.014.

58. JADHAV, A., PRAMOD, D. and RAMANATHAN, K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 2019, vol. 33, no. 10. pp. 913-933 DOI 10.1080/08839514.2019.1637138.

59. TROYANSKAYA, O., et al. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, 2001, vol. 17, no. 6. pp. 520-525 DOI 10.1093/bioinformatics/17.6.520.

60. YU, L., ZHOU, R., CHEN, R. and LAI, K.K. Missing Data Preprocessing in Credit Classification: One-Hot Encoding Or Imputation?. *Emerging Markets Finance and Trade*, 2022, vol. 58, no. 2. pp. 472-482 DOI 10.1080/1540496X.2020.1825935.

61. WILSON, N., SUMMERS, B. and HOPE, R. Using Payment Behaviour Data for Credit Risk Modelling. *International Journal of the Economics of Business*, 2000, vol. 7, no. 3. pp. 333-346 DOI 10.1080/13571510050197230.

62. TSAI, M. and WANG, C. On the Risk Prediction and Analysis of Soft Information in Finance Reports. *European Journal of Operational Research*, 2017, vol. 257, no. 1. pp. 243-250 DOI 10.1016/j.ejor.2016.06.069.

63. SÁNCHEZ, C.P., MONELOS, P.d.L. and LÓPEZ, M.R. A Parsimonious Model to Forecast Financial Distress, Based on Audit Evidence. *Contaduría Y Administración*, 2013, vol. 58, no. 4. pp. 151-173 DOI 10.1016/S0186-1042(13)71237-3.

64. WEI, Y., YILDIRIM, P., VAN DEN BULTE, C. and DELLAROCAS, C. Credit Scoring with Social Network Data. *Marketing Science*, 2015, vol. 35, no. 2. pp. 234-258 DOI 10.1287/mksc.2015.0949.

65. PEDREGOSA, F., et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011, vol. 12. pp. 2825-2830.

66. LI, J., et al. Feature Selection: A Data Perspective. *ACM Computing Surveys*, 2017, vol. 50, no. 6 DOI 10.1145/3136625.

67. HAUKE, J. and KOSSOWSKI, T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 2011, vol. 30, no. 2. pp. 87-93 DOI 10.2478/v10117-011-0021-1.

68. 1-Asis Verslo Apskaitos Standartas „Finansinė Atskaitomybė", May 28, 2015. Available from: https://www.e-tar.lt/portal/lt/legalAct/9f854fa005bd11e588da8908dfa91cac.

69. KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

70. BREIMAN, L. Random Forests. *Machine Learning*, 2001, vol. 45. pp. 5-32 DOI 10.1023/A:1010933404324.

71. BREIMAN, L. Arcing the Edge, 1997.

72. FRIEDMAN, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 2001, vol. 29, no. 5. pp. 1189-1232 DOI 10.1214/aos/1013203451.

73. CHEN, T. and GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. pp. 785-794 DOI 10.1145/2939672.2939785.

74. KE, G., et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems*, 2017. pp. 3149-3157.

75. VAPNIK, V.N. *The Nature of Statistical Learning Theory.* Springer New York, 1995 ISBN 978-1-4757-2440-0.

76. CORTES, C. and VAPNIK, V. Support-Vector Networks. *Machine Learning*, 1995, vol. 20. pp. 273–297 DOI https://doi.org/10.1023/A:1022627411411.

77. FERRI, C., HERNÁNDEZ-ORALLO, J. and MODROIU, R. An Experimental Comparison of Performance Measures for Classification. *Pattern Recognition Letters*, 2009, vol. 30, no. 1. pp. 27-38 DOI 10.1016/j.patrec.2008.08.010.

78. JENI, L.A., COHN, J.F. and DE LA TORRE, F. Facing Imbalanced Data Recommendations for the use of Performance Metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013. pp. 245-251 DOI 10.1109/ACII.2013.47.

79. BRIER, G.W. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 1950, vol. 78, no. 1. pp. 1-3 DOI 10.1175/1520-0493(1950)0782.0.CO;2.

80. ZOU, Q., et al. Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*, 2016, vol. 5. pp. 2-8 DOI 10.1016/j.bdr.2015.12.001.

81. LUNDBERG, S. and LEE, S. A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems*, 2017 DOI 10.48550/arxiv.1705.07874.

Appendices

Appendix 1. Descriptive statistics of numerical variables

Table 28. Descriptive statistics of numerical variables

	No default in 1 year								**Default in 1 year**							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
GDP growth	2169	0.029	0.02	0	0	0.04	0.046	0.046	97	0.033	0.018	0	0.025	0.043	0.043	0.046
Interest rates	2169	0.028	0.002	0.023	0.028	0.029	0.03	0.03	97	0.027	0.003	0.023	0.023	0.028	0.029	0.03
Age	2169	17.544	8.042	0	11	19	25	30	97	14.258	8.103	0	7	14	22	27
Current_ratio	2153	15.715	325.766	-437.5	1.162	1.894	3.764	10576.31	97	3.183	12.544	0.14	0.869	1.143	1.8	119.577
Debt_ratio	2165	0.529	0.633	0	0.237	0.46	0.689	15.931	97	0.721	0.346	0.008	0.488	0.678	0.948	1.8
Receivables_turnover	1960	141.18	3140.339	-87.721	4.675	7.546	13.785	116756	78	8.762	25.704	-104.619	4.728	6.422	11.661	151.131
Sales_to_current assets	2160	3.675	5.71	0	1.632	2.674	3.982	131.294	95	3.468	3.38	0.547	1.683	2.549	4.183	22.208
Total_assets_turnover	2160	2.022	2.436	0	0.663	1.524	2.569	42.831	95	2.203	3.051	0.128	0.61	1.421	2.43	20.236
ROE	2162	0.048	5.962	-194.877	0.019	0.095	0.235	120.377	97	0.63	2.589	-5.043	-0.051	0.085	0.306	17.212
Gross_margin	2112	0.283	0.406	-15.039	0.142	0.249	0.383	1.203	94	0.208	0.243	-0.735	0.063	0.181	0.349	0.786
Net_profit_margin	2154	-0.319	16.669	-772.715	0.004	0.031	0.084	25.337	95	-0.023	0.21	-1.636	-0.044	0.006	0.038	0.475
Cash_ratio	1970	1.876	23.993	-2.068	0.031	0.189	0.8	1018.375	82	1.791	13.19	0	0.004	0.028	0.115	119.115
Sales change	2108	0.97	24.155	-1	-0.05	0.056	0.183	1077.1	88	2.286	18.721	-0.538	-0.163	0.012	0.244	175.282
Short term assets change	2127	0.54	7.336	-1	-0.066	0.072	0.255	304.554	92	2.009	10.174	-0.661	-0.085	0.037	0.437	71.815
Net profit change	2124	12.829	283.084	-143.406	-0.482	0.135	1.18	10407.77	92	-4.246	19.51	-142.319	-1.104	-0.132	0.98	6.287
Equity change	2127	0.797	13.873	-30.184	-0.01	0.066	0.208	518.436	92	-0.287	2.768	-21.562	-0.195	0.004	0.217	5.25
EBIT to assets	2159	0.074	0.494	-10.699	0.003	0.048	0.128	14.578	97	0.063	0.377	-1.057	-0.043	0.019	0.063	1.724
Not overdue_to assets	2165	0.058	0.811	0	0	0.003	0.015	34.859	97	0.022	0.073	0	0	0.001	0.015	0.524
Overdue 1-30 days_to assets	2165	0.027	0.133	0	0	0.001	0.012	3.656	97	0.055	0.224	0	0	0.007	0.031	1.912
Overdue 31-60 days_to assets	2165	0.004	0.022	0	0	0	0	0.624	97	0.019	0.054	0	0	0.001	0.018	0.464
Overdue 61-90 days_to assets	2165	0.001	0.007	0	0	0	0	0.281	97	0.009	0.024	0	0	0	0.003	0.146
Overdue more than 90 days_to assets	2165	0.001	0.01	0	0	0	0	0.342	97	0.027	0.163	0	0	0	0	1.572

Appendix 2. Descriptive statistics of train dataset before and after missing data imputation

Table 29. Descriptive statistics of train dataset before and after missing data imputation

	Before imputation								**After imputation**							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
GDP growth	1551	0.04	0	0.02	0.04	0.04	0.05	0.05	1551	0.04	0	0.02	0.04	0.04	0.05	0.05
Interest rates	1551	0.03	0	0.02	0.02	0.03	0.03	0.03	1551	0.03	0	0.02	0.02	0.03	0.03	0.03
Age	1551	17.33	7.74	0	11	19	24	29	1551	17.33	7.74	0	11	19	24	29
Current_ratio	1539	19.12	383.9	-437.5	1.14	1.82	3.57	10576.31	1551	19.02	382.41	-437.5	1.14	1.82	3.57	10576.31
Debt_ratio	1547	0.53	0.57	0	0.24	0.48	0.71	15.93	1551	0.53	0.57	0	0.24	0.48	0.71	15.93
Receivables_turnover	1403	122.9	3142.44	-94.99	4.62	7.27	13.25	116756	1551	117.49	2989.55	-94.99	4.74	7.48	14.13	116756
Sales_to_current assets	1544	3.74	6.14	0	1.69	2.72	4.06	131.29	1551	3.76	6.14	0	1.69	2.73	4.06	131.29
Total_assets_turnover	1544	2.04	2.49	0	0.67	1.55	2.6	42.83	1551	2.05	2.5	0	0.68	1.55	2.6	42.83
ROE	1546	0.05	6.78	-194.88	0.01	0.08	0.2	120.38	1551	0.05	6.77	-194.88	0.01	0.08	0.2	120.38
Gross_margin	1514	0.27	0.45	-15.04	0.14	0.24	0.37	1.2	1551	0.27	0.44	-15.04	0.14	0.24	0.37	1.2
Net_profit_margin	1543	0.02	0.57	-15.05	0	0.03	0.07	9.59	1551	0.02	0.57	-15.05	0	0.03	0.07	9.59
Cash_ratio	1408	2.12	28.44	-2.07	0.02	0.13	0.6	1018.38	1551	2.06	27.14	-2.07	0.03	0.16	0.63	1018.38
Sales change	1502	1.04	28.19	-1	-0.04	0.06	0.18	1077.1	1551	1.18	28.28	-1	-0.04	0.06	0.18	1077.1
Short term assets change	1516	0.61	8.74	-1	-0.07	0.05	0.22	304.55	1551	0.67	8.69	-1	-0.07	0.06	0.23	304.55
Net profit change	1516	13.74	313.45	-143.41	-0.57	0.02	0.86	10407.77	1551	14.81	314.32	-143.41	-0.57	0.02	0.86	10407.77
Equity change	1516	0.54	9.23	-30.18	-0.03	0.05	0.16	311	1551	0.54	9.13	-30.18	-0.03	0.05	0.16	311
EBIT to assets	1544	0.07	0.53	-10.7	0	0.04	0.12	14.58	1551	0.07	0.53	-10.7	0	0.04	0.12	14.58
Not overdue_to assets	1547	0.06	0.95	0	0	0	0.02	34.86	1551	0.07	0.96	0	0	0	0.02	34.86
Overdue 1-30 days_to assets	1547	0.03	0.14	0	0	0	0.01	3.66	1551	0.03	0.14	0	0	0	0.01	3.66
Overdue 31-60 days_to assets	1547	0.01	0.03	0	0	0	0	0.62	1551	0.01	0.03	0	0	0	0	0.62
Overdue 61-90 days_to assets	1547	0	0.01	0	0	0	0	0.28	1551	0	0.01	0	0	0	0	0.28
Overdue more than 90 days_to assets	1547	0	0.01	0	0	0	0	0.34	1551	0	0.01	0	0	0	0	0.34