Suzi Mikulić, PhD
University of Mostar, Bosna and Herzegovina
E-mail: suzi.mikulic@sum.ba
Orcid: https://orcid.org/0000-0001-8730-7432

MACHINE LEARNING APPROACHES IN CREDIT RISK MODELLING IN B2B TRANSACTIONS IN THE NON-FINANCIAL SECTOR

UDC / UDK: 336.77:004.8:005.334
JEL classification / JEL klasifikacija: C45, G32, M41
https://doi.org/10.17818/EMIP/2026/34
Preliminary communication / Prethodno priopćenje
Received / Primljeno: February 4, 2026 / 4. veljače 2026.
Accepted / Prihvaćeno: July10, 2026 / 10. srpnja 2026.

Abstract

Credit risk in business-to-business (B2B) transactions can threaten financial stability if it is not managed effectively. This study evaluates the use of machine learning (ML) methods to predict credit risk in B2B transactions, using 4,828 observations from large companies in Bosnia and Herzegovina over a five-year period. Three ensemble-based ML models (Bagging Decision Tree, Random Forest, and Gradient Boosting) were compared with logistic regression. All ML models showed strong predictive performance, with Gradient Boosting performing slightly better overall. Liquidity, activity, and leverage indicators were the most important predictors across all models, while non-financial variables made only a limited contribution. The findings highlight the importance of accounting information in credit risk assessment and are relevant to IFRS 9, where probability of default is a key input to expected credit loss estimation under the general approach. These results provide a basis for further development of credit risk models for non-financial companies.

Keywords: machine learning, credit risk models, B2B transactions, IFRS 9, non-financial sector

1. INTRODUCTION

The goal of every company is to ensure long-term stable operations, with maintaining financial health playing a key role in achieving this (Rabega, 2024). A fundamental prerequisite for financial health is maintaining liquidity, which depends on the timely collection of receivables. Delays or failures in collecting

This work is licensed under a Creative Commons Attribution 4.0 International License.

these payments directly expose companies to credit risk, which can negatively impact cash flow and ultimately overall financial stability.

Credit risk is defined as the probability that the counterparty will fail to meet its financial liabilities under the agreed terms; in other words, it is the risk of loss arising from the debtor's failure to repay debt commitments within the stipulated deadlines (Nehrebecka, 2015). Although traditionally most prevalent in the banking sector, where it is strictly regulated by the Basel III framework, credit risk is also present in business-to-business (B2B) transactions, especially in the non-financial sector, where payment delays can significantly affect a company's liquidity and business stability.

In this context, credit risk prediction models are important tools for identifying and quantifying risk (Seitshiro & Govender, 2024). In the banking sector, they assess the probability of default (PD), the fundamental measure of credit risk, which reflects the likelihood that a borrower will fail to settle their current liabilities within a specific timeframe (Bandyopadhyay, 2016). PD is also a key component of the expected credit loss (ECL) model under the general approach of IFRS 9.

IFRS 9 is an international accounting standard that requires business entities to recognise impairment of financial instruments for ECL on each instrument exposed to credit risk, and has been mandatory since 2018 for financial and non-financial companies. Its application is based on a forward-looking approach, ensuring timely recognition of credit losses and increasing the transparency of financial reporting. In the non-financial sector, the most significant exposure to credit risk is accounts receivable, for which a simplified approach based on historical loss rates is applied to receivables without a significant financing component. However, within the general ECL approach, PD is a component without which it is not possible to determine the ECL, making credit risk assessment models a key tool for its operational application. According to this approach, financial instruments are classified into credit risk stages that reflect changes in their credit quality. Transfers between stages are triggered by increases or decreases in credit risk, with the number of days past due being one of the primary criteria used to assess such changes. Therefore, the development of such models could contribute not only to more accurate measurement of ECL, but also to strengthening the financial stability of non-financial sector companies and improving risk management (Grebenar, 2018). Previously, credit risk estimation models have been developed almost exclusively in a banking context, and their applicability to B2B transactions in the non-financial sector has not been empirically investigated, which represents a regulatory and research gap that this paper seeks to address. In the context of the non-financial sector in Bosnia and Herzegovina (B&H), the relevant legal framework defines credit risk as "the risk of loss of invested funds due to the debtor's delay in making payments to the company" (Article 6 of the Law on Financial Operations of the Federation of B&H, 2016). In this study, credit risk is considered not as a classic banking default, but as a legally defined delay in payments. Accordingly, this paper develops credit risk

assessment models for B2B transactions that use payment delay as an indicator of changes in customer credit quality. To identify operational default, the study adopts the 60-day past-due threshold prescribed by the B&H Financial Operations Act, which legally defines a state of illiquidity. In this context, operational default is not treated as a conventional banking default or a formal insolvency event. Instead, it reflects a significant deterioration in a customer's ability to meet payment obligations, which is particularly relevant in B2B trade relationships. By focusing on this type of credit event, the proposed framework provides an empirically grounded basis for estimating the PD parameters in the non-financial sector.

Recent advances in artificial intelligence and machine learning (ML) have created new opportunities to improve credit risk prediction (Alonso & Carbo, 2020). ML algorithms enable the modelling of nonlinear relationships, the processing of large datasets, and the identification of complex patterns that traditional statistical methods often cannot detect (Provenzano et al., 2020; de Oliveira & Basso, 2025). Given the specifics of the B2B environment, the heterogeneity of data, the nonlinear relationships between financial variables, and the absence of formal credit agreements, ML methods can provide a suitable analytical framework for credit risk modelling in this context.

Based on this, the paper analyses the applicability of ML models in predicting credit risk in B2B transactions within the non-financial sector. The study develops three ML models based on ensemble trees – Bagging Decision Tree, Random Forest, and Gradient Boosting to predict credit risk in B2B transactions at non-financial companies, alongside logistic regression as the traditional benchmark model commonly used in credit scoring. The research addresses four main questions:

1) Can ML techniques be used to develop credit risk models for B2B transactions in the non-financial sector, and how accurate are these models?
2) Do ML models demonstrate better performance compared to traditional statistical approaches, and which ML model achieves the highest performance?
3) Which indicators have the highest impact on predicting credit risk, and how does their importance vary across models?
4) What are the practical limitations of implementing these models in assessing credit risk for the non-financial sector?

Despite growing interest in applying ML to credit risk, existing literature largely focuses on the banking sector and credit products, while the B2B non-financial context remains under researched. In particular, studies modelling credit risk based on actual payment behaviour between companies are lacking. This gap is especially pronounced in transition economies, where payment delays are common and represent standard business practice. Transition markets are often characterised by liquidity constraints, challenges in debt collection, and weak financial discipline, which are linked to the underdeveloped institutional and financial frameworks of these economies (Machokoto et al., 2022).

In this regard, this paper contributes to the literature in several ways. First, it extends the existing literature on credit risk assessment to B2B transactions in the non-financial sector of a transition economy, where the aim is not to develop new ML algorithms, but to empirically validate existing ML methods within a specific institutional context. Unlike the dominant literature, which primarily focuses on the banking sector, this research analyses credit risk in an environment characterised by the absence of formal loan agreements, significant payment delays, and weaker institutional enforcement. Second, it introduces an institutionally based definition of a credit risk event, defined as payment delay exceeding 60 days, where one or more liabilities exceed 20% of total short-term liabilities, in line with the legal definition of illiquidity in Bosnia and Herzegovina (Article 17 of the Law on Financial Operations of the Federation of Bosnia and Herzegovina, 2016). This adapts the model to the actual characteristics of B2B relationships in the non-financial sector, which differ from standard banking definitions of default based on the 90-day threshold. Third, the value of this work derives from a unique dataset covering approximately 80% of large non-financial companies in B&H over a five-year period. Access to comprehensive firm level data in B&H remains relatively limited. Consequently, studies based on large scale datasets typically require substantial data collection efforts and remain relatively rare within the local research environment. Furthermore, the paper examines the practical applicability of ML models in real business conditions, including their ability to achieve high predictive accuracy using original, untransformed financial indicators, without the need for extensive data preprocessing or the strict linearity assumptions characteristic of logistic regression. The paper is also relevant in the context of IFRS 9, since PD is a key component of the ECL model under the general approach. The developed approaches may provide an empirical starting point for further research into PD estimation in non-financial companies, with potential relevance for credit risk measurement in accordance with accounting requirements.

Below is an overview of previous research and relevant literature on the development of credit risk assessment models. This is followed by a description of the methodology, including data preparation and the development of predictive ML models for credit risk assessment. The results section presents the outcomes of the model comparison and evaluation, while the discussion compares the findings and explores their practical implications.

2. THEORETICAL BACKGROUND

2.1. Historical development of credit risk assessment models

Throughout history, the concept of credit risk has evolved alongside the development of trade. What began as relatively simple deferred payment arrangements has gradually transformed into a complex framework for managing financial risk, increasingly supported by technological advancements. Before the twentieth century, deferred exchanges relied primarily on interpersonal trust and

the reputational standing of merchants. At the same time, banks expanded lending activities under conditions of limited transparency and imperfect information about future asset values. The industrial expansion of the late nineteenth and early twentieth centuries, followed by periods of economic instability, market stagnation, global integration, and increasing demands for standardisation, led to the proliferation of companies and financial instruments. These changes created a growing need for more flexible credit arrangements, extended payment terms, and increasingly complex financial transactions. Consequently, credit risk intensified, prompting greater focus on its measurement and management. Altman and Saunders (1998) argue that several long-term structural developments have made systematic credit risk assessment indispensable, including rising bankruptcy rates, increasing disintermediation among high-quality borrowers, heightened competitive pressures on lending margins, declining collateral values across multiple markets, and the rapid expansion of off-balance sheet instruments with embedded default risk.

Beaver (1966) and Altman (1968) developed the first predictive models of financial failure. These models relied on historical (fundamental) information from financial statements and applied univariate and multivariate discriminant analysis. Later, structural models, also known as theoretical asset value models, were developed, including the Black-Scholes (BS) model (1973) and the Black-Scholes-Merton (BSM) model (1974). These models estimate the likelihood of default by treating corporate debt as a put option on the company's assets, with bankruptcy occurring when the value of assets falls below total liabilities (Wang, Zhang & Liu, 2023). Structural models rely on stock price data, which are suitable for market-based valuation, but this also limits their applicability to companies that are not publicly traded. Throughout the 1970s and 1980s, logistic regression became increasingly popular as a statistical technique for assessing credit risk in banks, particularly for estimating the PD, a key component of credit risk assessment (Ohlson, 1980; Bandyopadhyay, 2016). Its use remains widespread today.

With advancements in technology and data analysis, and the need to incorporate an increasing number of potentially explanatory variables, modern models have been developed. These models use advanced algorithms such as ML and big data analysis (Alonso & Carbo, 2020). Recent research published in economic journals from Southeast European countries confirms the growing use of ML methods in various areas of economics (Ljubičić & Hell, 2023) and finance, particularly in financial forecasting and predictive modelling (Vlah Jerić, 2023). However, the use of these methods in modelling the credit risk in B2B transactions, based on a sample of companies from B&H, remains insufficiently explored. This paper applies ML algorithms to modelling credit risk in B2B transactions of non-financial companies from B&H, thus contributing to the existing literature by empirically comparing selected ML algorithms in the assessment of the credit risk of business entities.

2.2. Overview of recent research and applications of ML in credit risk modelling

As credit risk has traditionally been associated with lending activities, earlier research has predominantly focused on the banking sector. These studies have examined bank receivables from granted loans, observing credit risk either from the client's default status (Grebenar, 2018) or the level of non-performing loans (Chaibi & Ftiti, 2015; Kharabsheh, 2019). A similar approach is also present in research from Southeast European countries, where credit risk analyses often focus on the banking sector and factors that affect the level of non-performing loans (Džidić et al., 2022).

In recent years, artificial intelligence, particularly ML approaches, has significantly impacted credit risk assessment, as the use of ML techniques in credit risk management within the financial industry has increased (Alonso & Carbo, 2020). Numerous studies indicate that ML models, especially those employing ensemble techniques such as Random Forest and Gradient Boosting, outperform traditional statistical approaches in predicting defaults and evaluating corporate credit risk (Noriega, Rivera & Herrera, 2023; Suhadolnik, Ueyama & Da Silva, 2023; Ayari, Guetari & Kraïem, 2025). Tian et al. (2020) showed that the Gradient Boosting Decision Tree (GBDT) is more effective than traditional ML models commonly used for credit risk assessment, such as Support Vector Machine (SVM), classical decision tree, and Multi-Layer Perceptron (MLP). Their results indicate that GBDT has the greatest ability to distinguish between good and bad clients, as well as a strong ability to generalise to new data. These models are particularly effective at capturing the nonlinear relationships often present in financial data. Historically, banks have relied on logistic regression and other statistical techniques; however, the increasing complexity of corporate networks and credit portfolios has driven the growing adoption of ML based models (Chowdhury, 2025).

Recent literature also highlights the limitations of ML applications. Interpretability remains a challenge, particularly for complex models such as Gradient Boosting and Neural Networks, which limits their transparency for bank risk officers and regulators (Noriega et al., 2023). Chen, Ye, & Ye (2023) note that ML models are often treated as “black boxes”, making their results difficult to interpret. High model performance is frequently achieved at the expense of explainability, and users may not understand the logic behind the model's predictions. This lack of interpretability is one of the main challenges for both researchers and the credit industry, as credit application decisions cannot be based on discretionary models. In that context, methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are becoming more widely used, as they explain how individual variables influence model predictions and improve the transparency of complex ML models (Molnar, 2022; Lundberg & Lee, 2017). Practical implementation of such approaches may depend on the functionalities and integration capabilities of the software environment used for model development and evaluation. Additionally, explainability techniques such as feature importance measures can provide valuable insights by helping users understand which indicators most strongly influence model

predictions. Bücker et al. (2020) note that in credit scoring, logistic regression is still often used for its simplicity and interpretability, while increasing attention is being given to developing methods for explaining complex ML models, which can achieve comparable or better performance but with greater interpretative complexity.

Several studies highlight the practical application of ML in banking. Suhadolnik et al. (2023) empirically demonstrate that ensemble ML models significantly increase predictive accuracy for corporate clients' creditworthiness compared to traditional rating models. Similarly, Zhang et al. (2024) show that optimised Random Forest models can effectively enhance credit risk assessment by identifying the most influential features in large loan datasets.

Furthermore, the literature indicates a clear trend towards hybrid and stacking approaches. Chen et al. (2023) report that combining multiple models stabilises predictions and mitigates challenges associated with unbalanced credit portfolios, using ensemble techniques that integrate feature selection, data balancing, and algorithmic enhancements to achieve improved predictive performance. Although interpretability remains a challenge, ML algorithms and techniques offer practical solutions for assessing corporate credit risk efficiently and accurately (Chen et al., 2023; Suhadolnik et al., 2023; Liu, 2024).

Despite significant progress in developing credit risk models within the banking sector, these approaches have mainly been designed for financial institutions and loan-based lending, raising questions about their direct applicability to B2B transactions in the non-financial sector, where the nature of credit risk differs.

2.3. Credit risk in B2B transactions

In the B2B environment, credit risk arises from a counterparty's inability to settle liabilities on time. Unlike bank lending, where repayment terms are predefined in a contract, B2B relationships are characterised by delayed payments, which complicate standardised risk measurement. In this context, trade receivables management represents an important aspect of corporate financial management, as companies should establish appropriate credit policies regarding credit terms, credit standards, and collection procedures when selling products or services on credit (Kontuš, 2013). Payment delays may result from limited liquidity (Wu et al., 2020), insufficient operational efficiency (Deloof, 2003), an unfavourable capital structure (Hill et al., 2010), or strategic liquidity management through intentional deferral of payments (Wilner, 2000). Companies with lower liquidity are more likely to delay settling liabilities due to a lack of available cash, while high leverage, increases financial pressure and reduces flexibility in cash flow management. Operational efficiency, measured by inventory and accounts receivable turnover ratios, also affects the ability to generate sufficient liquidity to settle short-term liabilities (Tintor, 2020).

The literature on trade credit highlights that credit terms in B2B relationships serve not only as a means of deferred payment, but also as a key

mechanism for financing and coordination within supply chains. Trade credit can stimulate demand, increase the profitability of supply chain participants, and serve as an alternative to bank financing, particularly when access to external capital is limited (Seifert et al., 2013). In this context, decisions regarding the granting of trade credit and the tolerance of payment delays depend not only on the financial stability of the buyer, but also on the cost of capital, bargaining power, supply chain coordination, and strategic working capital management. Consequently, the financial indicators in this research acquire a broader theoretical foundation, serving not only as traditional predictors of financial distress, but also as indicators of the economic mechanisms underlying payment behaviour in the B2B environment and the company's ability to maintain trade credit, manage the operating cycle, and participate in financial coordination within the B2B transactions. For this reason, they constitute the majority of independent variables and potential predictors in credit risk assessment models for B2B transactions.

The characteristics of the B2B environment, data heterogeneity, non-linear relationships between financial variables, mixed types of predictors, and high variability in payment behaviour between subjects, can make classical statistical methods such as logistic regression methodologically limited in this context. Logistic regression assumes a linear relationship between the predictors and the log-odds of the dependent variable, exhibits sensitivity to extreme values, and requires extensive data preprocessing to address multicollinearity and outliers. In contrast, decision tree-based ML methods naturally model non-linear relationships and interactions between variables, exhibit robustness to outliers and heterogeneous distributions, and can handle numerical and categorical variables simultaneously without transformation. In the context of B2B transactions, where payment delays are influenced by liquidity pressures, operational efficiency, capital structure, and strategic behaviour, and these relationships are not necessarily linear or additive, ML methods therefore provide a theoretically grounded and methodologically appropriate analytical framework.

3. METHODOLOGY

In this study, credit risk refers to the risk that a customer will not settle its trade obligations within the legally prescribed payment period. As existing credit risk models have been developed primarily in the banking context, this paper uses banking literature as a methodological basis, while adapting the operational definition to the specifics of B2B transactions through the 60-day threshold prescribed by the Law on Financial Operations of the Federation of B&H. The developed models rely on payment delay dynamics as a key indicator of changes in customer credit quality.

The models have been developed using three decision tree methods based on an ensemble approach: (a) Bagging Decision Tree, (b) Random Forest, and (c) Gradient Boosting. Ensemble methods enhance prediction accuracy by combining multiple decision trees. Bagging constructs several decision trees from random bootstrap samples of the training data. Each tree is built independently, and

predictions are aggregated by majority voting for classification or probability averaging, which reduces variance compared to single-tree models (Breiman, 1996). Random Forest extends Bagging by introducing randomness in both row sampling and feature selection at each split, reducing correlation between trees and improving generalisation (Breiman, 2001). Boosting is a sequential method that iteratively builds trees, with each new tree focusing on correcting the errors of its predecessors. It adapts to data complexity and often achieves the highest accuracy among ensemble techniques (Friedman, 2002).

In the empirical analysis, logistic regression was also used as a reference (benchmark) model for the classical statistical approach, to compare the predictive power of traditional and modern ML methods.

The empirical analysis was conducted using SAS Enterprise Miner, a data mining and analysis software developed by the American SAS Institute. This software is designed to detect patterns and build accurate predictive and descriptive models based on the analysis of large datasets. Model development in SAS Enterprise Miner follows the SEMMA principle (Sample, Explore, Modify, Model, Assess) and is primarily driven by a graphical user interface. In line with the system architecture, the evaluation of results relies on built-in visual representations, ROC curves, lift graphs, confusion matrices, and intra-model measures of the importance. Advanced methods of model explainability, such as SHAP values or LIME, were not available in the software used, which represents a methodological limitation acknowledged in the paper and identified as a direction for future research using more flexible analytical platforms.

For data processing, tree-based ML models were developed using original, untransformed variables, as these algorithms are less sensitive to measurement scales and data distribution formats, and can internally handle different types of variables without additional transformation. However, this does not diminish the importance of data preprocessing; rather, it reflects the robustness of these models to certain forms of variability in the data.

In contrast, logistic regression was developed using previously transformed variables. Logarithmic transformations were applied to variables with pronounced asymmetry, while winsorisation was used to limit the influence of extreme values. Additionally, missing values in continuous variables were replaced by mean values, which is the standard and straightforward approach for handling missing data in classical statistical models.

ML models were trained using standard (default) hyperparameter values, without additional optimisation. Although ensemble methods such as Gradient Boosting and Random Forest possess mechanisms like multi-model aggregation or sequential learning, which often contribute to more stable and accurate predictions, their performance can be affected by the choice of hyperparameters. In this study, the aim was to compare models under standardised conditions, without additional optimisation, to ensure comparability of results, while optimisation of hyperparameters represents a direction for future research.

3.1. Data collection and sample structure

The research was conducted on a sample of large companies in B&H during a stable, pre-crisis five-year period from 2015 to 2019. This timeframe was chosen to avoid the effects of extreme macroeconomic shocks that occurred after 2020, primarily due to the COVID-19 pandemic and the ongoing wars in Ukraine and the Middle East. Given the relative rarity of studies focused on credit risk models for transactions in the non-financial sector, selecting this economically stable period provides a strong basis for future research and comparisons.

The initial dataset comprised 4.856 observation units, covering large companies from three key sectors of the non-financial sector in B&H: manufacturing, construction, and wholesale and retail. The manufacturing sector included 1.871 units, with annual distributions as follows: 364 (2015), 369 (2016), 372 (2017), 378 (2018), and 388 (2019), representing 38.53% of the sample. The construction sector comprised 465 units: 90 (2015), 92 (2016), 93 (2017), 94 (2018), and 96 (2019), accounting for 7.50% of the total. The wholesale and retail sector had the largest share, with 2.520 units: 485 (2015), 494 (2016), 501 (2017), 507 (2018), and 533 (2019), accounting for 40.66% of the companies observed. The inclusion of large companies in the sample reflects the regulatory framework (Law on Accounting and Auditing of the Federation of B&H, 2009; 2021), as the mandatory application of IFRS 9 applies to this segment, whereas for small and medium-sized enterprises (SMEs), application is voluntary. Therefore, extending the analysis to the SME segment, while taking into account structural differences in data availability and payment behaviour patterns, represents a direction for future research. The selected companies represented approximately 80% of all large non-financial companies in B&H, ensuring the representativeness and reliability of the sample for the credit risk model. Collecting data of this scope in B&H is particularly demanding, given that the country's fragmented administrative structure with the Federation of B&H, Republika Srpska, and Brcko District each having separate institutions responsible for collecting and processing financial statements which makes empirical research of this kind relatively rare and the data difficult to obtain. The classification of companies as large was based on the criteria from the B&H Law on Accounting and Auditing, which was in force during the research period. A company was classified as large if it met at least two of the following conditions: it employed more than 250 people, had a balance sheet exceeding €2 million, or had an annual turnover exceeding €4 million. The data required for the research were obtained from the database of LRC Ltd, Sarajevo.

3.2. Dependent variable

The dependent variable in this study is a credit risk event, defined as a delay in settling obligations exceeding 60 days. The model estimates PD, i.e. the probability of occurrence of a credit risk event as defined above. The 60-day threshold is adopted from the regulatory framework of the Federation of B&H (Article 17 of the Law on Financial Operations of the Federation of B&H, 2016),

where it serves as one of the criteria for determining illiquidity. In the context of IFRS 9, the estimated PD is an input parameter in the ECL general approach model, used to estimate the level of credit risk and potential credit losses over time. In this framework, the credit quality of exposures is assessed through three levels (Stage 1–3), which reflect changes in credit risk over the life of the exposure. Entities without a significant increase in credit risk correspond to Stage 1, entities with a significant increase in credit risk correspond to Stage 2, and entities with a delay of more than 60 days represent a group with highly elevated credit risk, conceptually close to Stage 3 under IFRS 9. This approach enables the empirical model to reflect real sector practice and serves as an institutionally based alternative to the banking definition of default, which is based on a delay of more than 90 days (Regulation (EU) No 575/2013). The dependent variable is binary (0/1), where 1 indicates that the legally prescribed delay threshold (>60 days) has been exceeded, and 0 indicates that the company settled its obligations within the prescribed period.

It is important to note that payment delays in B2B relationships can reflect various economic situations, including liquidity difficulties, an unfavourable capital structure, strategic cash flow management, or asymmetries in bargaining power between business partners (Wilner, 2000; Hill et al., 2010). Regardless of the cause, any delay exceeding the legal deadline of 60 days constitutes a legally defined illiquidity and, as such, represents a relevant credit risk for the creditor. The dependent variable in this study is not limited to "pure financial default" in the banking sense and should not be interpreted as a conventional banking default or insolvency event, but rather encompasses a broader, institutionally defined credit risk concept specific to the B2B environment. This distinction has important implications for the interpretation of the model: the identified predictors reflect the determinants of payment delay in a broader sense, and not exclusively the predictors of financial distress.

The dataset was randomly divided using stratified random sampling, with a binary indicator of credit risk events (0/1) used as the stratification variable. The data were split into two parts: 60% for model development (training sample) and 40% for validation (testing sample). This stratified approach ensured that the proportion of the value "1" in the dependent variable remained consistent in both subsets, supporting the robustness and comparability of the developed models. The credit risk event during the observed period ranged from 42.25% to 47.75%, with an average of 45.74%, indicating a relatively balanced distribution of the dependent variable and justifying the use of stratified sampling. Therefore, no data balancing techniques such as oversampling or undersampling were applied.

3.3. Independent variables

In this research, the independent variables are financial ratios and available non-financial information for the sampled companies, as shown in Table 1. The inclusion of both financial and non-financial indicators aligns with previous research in credit risk modelling, which demonstrates that combining quantitative

financial ratios with qualitative non-financial information improves predictive outcomes and model robustness (Giampaolo, Matthias & Giammarino, 2020; Wu, Li & Zhang, 2025). Financial variables have traditionally formed the basis of credit risk assessment, covering key aspects of liquidity, profitability, and leverage. Recent studies continue to confirm their importance. For example, Gafsi (2025) shows that key financial performance indicators, particularly ratios related to indebtedness, profitability, and asset quality, remain highly influential predictors in ML models such as Random Forest, XGBoost, CatBoost, and LightGBM.

Table 1 Table of Indicators (Independent Variables)

Indicator Name	Symbol	Indicator Name	Symbol
Accrual-based Indicators*			
Current Ratio	QLig	Net Profit Margin	NmProf
Quick Ratio	FLig	Gross Profit Margin	BmProf
Liquidity Ratio	CLig	Net Return on Assets	NmTA
Financial Stability Ratio	FSt	Gross Return on Assets (ROA)	BmTA
Net Working Capital	NWC	Return on Equity (ROE)	NmCap
Cash Gap	CGap	Total Asset Turnover	QtTA
Debt Ratio	QSol	Current Asset Turnover	QtCA
Equity Financing Ratio	QCap	Accounts Receivable Turnover	QtRec
Financing Ratio	LibCap	Average Inventory Days	DInv
Interest Coverage	CovInt	Average Collection Period	DRec
Indebtedness Factor	FqSol	Total Business Efficiency	Tec
Coverage Ratio I	CovD I	Operating Efficiency	Bec
Coverage Ratio II	CovD II	Other Operating Efficiency	Oec
Credit Debt/Equity	CBCap	Revenue Growth	GrRRev
Cash-based Indicators*			
Cash Interest Coverage	CCint	Financing Indicator	CFin
Cash Coverage of Current Liabilities	CCLib	Cash Return on Assets	CRoa
Cash Coverage of Total Liabilities	CCTLib	Cash Return on Capital and Liabilities	CRoce
Revenue Quality	QRev	Cash Return on Equity	CRoe
Earnings Quality	QPro	Free Cash Flow I	FCF I
Investment Indicator	CInv	Free Cash Flow II	FCF II
Non-financial Information			
Industry Sector	Sec	Number of Employees	Emp
Type of Organization (Ltd or JSC)	Org	Age	Age

*Source: * Gulin et al. (2004, 183, 195, 215-216); Gulin (2009, 65); Žager et al. (2017, 45-56); Gulin & Hladika (2018, 110)*

This variable design aligns with the research objective of assessing whether ML techniques can be effectively applied to develop accurate and reliable credit risk models for B2B transactions in the non-financial sector, thereby addressing the first research question (RQ1).

ML models based on decision trees, including Bagging Decision Tree, Random Forest, and Gradient Boosting, are suitable for this type of data as they enable the modelling of non-linear relationships and interactions between financial and non-financial variables. Previous research indicates that Gradient Boosting often outperforms classical and other ML approaches in credit risk prediction (Tian

et al., 2020; Suhadolnik et al., 2023), which justifies its use in addressing the second research question (RQ2) regarding the identification of models with the highest performance. In addition, logistic regression was developed as a benchmark model to compare the performance of the classical statistical approach and ML methods.

Furthermore, the heterogeneous structure of the dataset enables the third (RQ3) and fourth (RQ4) research questions to be addressed. It allows identification of the most important financial and non-financial predictors of credit risk (RQ3), and provides a basis for discussing the practical limitations of using these models in assessing credit risk in the non-financial sector (RQ4).

4. RESEARCH RESULTS

The results of empirical data processing for the developed credit risk models are presented below, structured according to the research questions.

4.1. Model comparison and performance evaluation

Table 2 compares ML models based on the ensemble approach with logistic regression, which serves as the benchmark model for the classical statistical approach in credit risk modelling. Model quality values are provided for both the training and validation samples. The evaluation and comparison of the models are primarily based on the results from the validation sample, as these offer a more objective assessment of the models' predictive performance on previously unseen data.

For the validation sample, the following indicators are presented: ROC index, average squared error (ASE), misclassification rate, and Kolmogorov–Smirnov (KS) statistic. For the training sample, the ROC index and ASE are shown. Using multiple evaluation criteria enables more comprehensive assessment of the model's predictive power, beyond its discriminatory capacity. ROC curves (Figure 1) are shown for both samples and do not display significant deviations, indicating model stability and the absence of overfitting. Therefore, the ROC index estimated on the validation sample was used as the main criterion for comparison, while other model quality indicators were also considered.

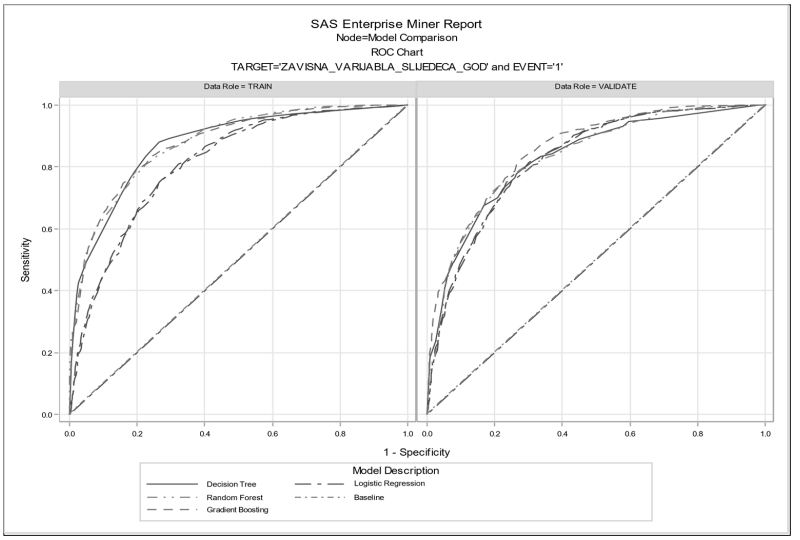
Table 2 Statistical Results for Model Comparison

Model Description	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error	Valid: Misclassification Rate	Valid: Kolmogorov-Smirnov Statistic	Train: Average Squared Error	Train: Roc Index
Gradient Boosting	0.856	0.15790	0.23641	0.548	0.14931	0.876
Random Forest	0.833	0.17968	0.23510	0.527	0.16456	0.878
Bagging Decision Tree	0.829	0.17045	0.24492	0.514	0.13809	0.875
Logistic Regression	0.828	0.16960	0.24885	0.520	0.17785	0.810

Source: model results obtained using SAS Enterprise Miner

All models achieve high ROC index values (0.828–0.856), which, according to the standard interpretation (Hosmer & Lemeshow, 2000), indicates excellent discriminatory ability. Gradient Boosting achieves slightly better results across all evaluation criteria (ROC = 0.856; ASE = 0.15790; KS = 0.548), while Random Forest (ROC = 0.833) and Bagging Decision Tree (ROC = 0.829) show comparable performance. Logistic regression achieves an ROC index of 0.828, which is close to the performance of the ML models. The differences in discriminatory power between all models are relatively small and should be interpreted with caution.

Figure 1 ROC curve



Source: model results obtained using SAS Enterprise Miner

The ROC curves in Figure 1 compare the discriminatory ability of the ML models and logistic regression. Consistent with the statistical results presented above, the ROC curves visually demonstrate the comparable performance of the evaluated models. Specifically, Gradient Boosting shows a slightly higher discriminatory power, though the differences between the models remain relatively small.

4.2. Variable importance

The third research question concerns identifying the most important predictor variables in credit risk prediction for B2B transactions, as shown in Tables 3–7.

Table 3 presents the predictor variables for the Bagging Decision Tree model. The importance of each predictor variable was estimated based on its frequency of occurrence and position within the tree, using both training and

validation samples. The predictors are ranked according to their validation sample importance, identifying QtCA, CGap, and LibCap as the most contributing variables.

Table 3 Important Predictor Variables in Bagging Decision Tree Model

Variable Name	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
QtCA	5	1.0000	0.9262	0.9262
CGap	6	0.9736	0.7850	0.8063
LibCap	1	0.9432	1.0000	1.0602
QtTA	1	0.6682	0.3981	0.5958
CLig	2	0.4020	0.2650	0.6593
QtRec	3	0.3532	0.2958	0.8375
QCap	1	0.3205	0.2643	0.8247
FqSol	2	0.2903	0.0000	0.0000
CovD I	1	0.2698	0.0000	0.0000
Age	1	0.2027	0.1808	0.8920
ESec	1	0.1860	0.1322	0.7109
NWC	1	0.1769	0.0666	0.3764
CovInt	1	0.1606	0.2630	1.6374
CovD II	1	0.1470	0.2104	1.4310

Source: research results from SAS Enterprise Miner

Table 4 presents important variables for the Random Forest model, showing each variable's contribution to reducing the Gini index of impurities in the training and validation samples. Ensemble trees, such as Random Forest, are known for their complexity and limited interpretability. To quantify the importance of individual predictors, Breiman (2001) proposed metrics that measure each variable's contribution to improving model performance. In this paper, the importance of each predictor is assessed using the Mean Decrease Impurity (MDI) metric, which is based on the Gini index. Mathematically, each time a node is split within a single tree, the selected variable reduces the Gini impurity at that node. The total importance of a variable is calculated as the sum of all reductions in Gini impurity achieved by that predictor, weighted by the number of samples passing through the node, and divided by the total number of trees in the ensemble. Variables that appear more frequently at higher hierarchical levels of the trees (closer to the root) provide greater reductions in impurity and thus receive higher importance scores. The importance of each predictor variable was estimated based on its frequency of occurrence and position within the tree, using both training and validation samples. Predictors are ranked according to their importance in the validation sample, reflecting their overall contribution to the model. The variables with the highest impact on index reduction in the validation sample are QCap, QSol, and QtTA. This result indicates that these specific indicators have the highest discriminatory power in distinguishing regular payers from clients in credit risk event, emphasising the importance of leverage and activity indicators in B2B credit risk assessment.

Table 4 Important Predictor Variables in Random Forest Model

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction
NmCap	105	0.004011	0.008022	-0.00039	0.00428
CovInt	84	0.001576	0.003152	-0.00225	-0.00115
QCap	65	0.022801	0.045602	0.02212	0.04746
QtTA	64	0.011581	0.023161	0.00801	0.02107
DRec	61	0.007271	0.014542	0.00411	0.01166
CCint	60	0.001015	0.002030	-0.00126	-0.00079
CGap	59	0.005643	0.011287	0.00159	0.00774
CBCap	49	0.002143	0.004286	-0.00063	0.00171
QSol	49	0.012456	0.024912	0.01242	0.02605
ESec	47	0.004876	0.009753	0.00261	0.00737
LibCap	44	0.004644	0.009288	0.00355	0.00878
BmProf	43	0.002863	0.005727	0.00127	0.00447
QtCA	40	0.008093	0.016186	0.00739	0.01610
CInv	39	0.000750	0.001499	-0.00087	-0.00009
DInv	38	0.002186	0.004372	0.00064	0.00369
CLig	35	0.008162	0.016324	0.00703	0.01641
BmTA	34	0.005241	0.010481	0.00496	0.01069
NmProf	31	0.001513	0.003026	-0.00025	0.00169
Oec	26	0.000710	0.001419	-0.00103	-0.00022
CFin	22	0.000592	0.001185	-0.00076	0.00004
FLig	19	0.001627	0.003254	0.00054	0.00241
FqSol	19	0.000989	0.001978	0.00028	0.00146
CCTLib	18	0.000880	0.001760	-0.00024	0.00088
GrRRev	18	0.000452	0.000903	-0.00036	0.00040
CovD I	18	0.000621	0.001243	0.00008	0.00058
QtRec	17	0.000800	0.001601	-0.00005	0.00080
NWC	17	0.005102	0.010203	0.00428	0.00991
CovD II	17	0.001100	0.002201	0.00035	0.00143
NmTA	16	0.002713	0.005426	0.00262	0.00569
CCLib	15	0.000947	0.001895	0.00037	0.00143
QPro	15	0.000340	0.000679	-0.00011	0.00037
Bec	14	0.000595	0.001190	-0.00020	0.00039
FSt	13	0.001573	0.003145	0.00087	0.00263
Ent	12	0.000548	0.001096	0.00001	0.00057
QLig	12	0.001648	0.003297	0.00189	0.00412
Age	11	0.000751	0.001502	-0.00016	0.00053
FCF II	10	0.000407	0.000814	-0.00036	0.00007
CRoce	10	0.000498	0.000995	-0.00027	0.00046
CRoe	10	0.000144	0.000288	-0.00000	0.00026
Emp	8	0.000612	0.001225	-0.00008	0.00056
Tec	7	0.000142	0.000285	-0.00014	-0.00006
CRoa	6	0.000170	0.000339	-0.00022	0.00002
FCF I	6	0.000227	0.000455	-0.00020	0.00002
Org	5	0.000160	0.000320	0.00006	0.00028

Source: research results from SAS Enterprise Miner

Table 5 presents the importance of predictor variables in the Gradient Boosting model, showing that a small number of variables account for most of the model's overall contribution. This aligns with the theoretical characteristics of ensemble methods, which automatically perform variable selection and reduce the

influence of less important predictors. The most significant predictors were identified using the measure of relative importance (Importance) and confirmed on the validation sample (Validation Importance). The results indicate that the variables CGap, QtCA, and CLig provide the largest contribution to the model. CGap achieved the highest importance value on the training sample (1.000), while QtCA had the highest importance on the validation sample (1.000), demonstrating its strong discriminatory ability. The CLig variable also showed a stable contribution to the model in both the training and validation samples. Additionally, the variables QtRec, QtTA, LibCap, and QCap make a significant contribution to the model, although their validation values are slightly lower than those on the training sample.

Table 5 Important Predictor Variables in Gradient Boosting Model

Variable Name	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
CGap	26	1.00000	0.82584	0.82584
QtCA	16	0.88802	1.00000	1.12610
CLig	7	0.77985	0.81435	1.04423
QtRec	14	0.72565	0.50487	0.69574
QtTA	13	0.71621	0.54724	0.76408
LibCap	5	0.59751	0.64476	1.07908
QCap	3	0.43982	0.46414	1.05530
DInv	8	0.41905	0.37839	0.90297
QSol	1	0.41432	0.48025	1.15915
DRec	3	0.37327	0.34537	0.92525
NWC	1	0.31257	0.31299	1.00134
FLig	2	0.24914	0.33456	1.34289
QLig	2	0.19990	0.00000	0.00000
FqSol	2	0.17805	0.15533	0.87244
CBCap	2	0.17562	0.00000	0.00000
Emp	2	0.16000	0.12486	0.78041
ESec	1	0.15007	0.04114	0.27412
CRoe	1	0.13734	0.00000	0.00000
CovD II	1	0.13446	0.00000	0.00000
GrRRev	1	0.11250	0.00000	0.00000
CCint	1	0.10852	0.00000	0.00000
CCTLib	1	0.10420	0.06876	0.65989

Source: research results from SAS Enterprise Miner

In addition, for comparison with the classical approach, a logistic regression model was developed, with the results shown in Table 6. Logistic regression is more sensitive to asymmetric distributions and extreme values of independent predictor variables. Therefore, standard pretreatment of variables was performed before modelling. A logarithmic transformation (19 variables) was applied to asymmetric variables with exclusively positive values, while winsorisation at the 1st and 99th percentiles (20 variables) was used for variables with both positive and negative values. Missing values were replaced by mean

values (24 variables), provided that the proportion of missing data did not exceed 5% for any variable. The extent of the required preprocessing, which included transforming or correcting almost all variables in the dataset, empirically confirms that financial data in the B2B non-financial sector is highly asymmetric and contains extreme values. This reflects the real heterogeneity between businesses in this environment and further justifies the use of tree-based ML methods, which are robust to such data characteristics without prior transformation. The results of the logistic regression show that the model includes seven statistically significant predictors ($p < 0.05$), which fall into three main groups of financial indicators: activities (QtTA, QtCA, DRec), leverage (QCap, CBCap), and liquidity (CLig, CGap).

Table 6 Regression Coefficients in a Logistic Regression Model

Variable Name	Level	Coefficient	T-value	P Value
Intercept	1	4.47864	10.3069	6.5554E-25
QCap		-2.53584	-7.7695	7,882E-15
LOG Clig		-1.38013	-8.0211	1.0477E-15
IMP LOG QtCA		-1.35877	-6.4500	1.1188E-10
LOG QtTA		-1.07879	-4.88822	.000001018
IMP LOG CBCap		-0.65290	-5.08918	.000000360
IMP LOG DRec		0.24995	4.39775	.000010938
IMP REP CGap		-0.00065	-4.67966	.000002874

Source: research results from SAS Enterprise Miner

Table 7 presents the most influential financial and non-financial variables identified by three ML models (Gradient Boosting, Random Forest, and Bagging Decision Tree) and logistic regression in assessing credit risk for large non-financial companies.

The results demonstrate the continued dominance of accrual-based financial indicators, particularly liquidity variables (CGap, CLig), activity measures (QtCA, QtTA), and leverage (QCap). Their consistent presence across all models indicates their robust capacity to explain credit risk in the non-financial sector.

Although there is a high degree of agreement between models regarding the identification of the most important variables, some differences have been observed, particularly in the Gradient Boosting model, which identifies more relevant predictors (22) than logistic regression (7). This suggests that ML models can identify more complex nonlinear relationships and interactions between variables that linear models cannot fully capture.

Table 7 Comparison of Variable Indicators (Predictors) Identified as Important in Credit Risk Assessment across ML Models and Logistic Regression

Variable Name	Gradient Boosting	Random Forest	Bagging Decision Tree	Logistic Regression
CGap	✓	✓	✓	✓
QtCA	✓	✓	✓	✓
CLig	✓	✓	✓	✓
QtRec	✓		✓	✓
QtTA	✓	✓	✓	✓
LibCap	✓		✓	
QCap	✓	✓	✓	✓
DInv	✓			
QSol	✓	✓		
DRec	✓	✓		
NWC	✓	✓	✓	
FLig	✓			
QLig	✓			
FqSol	✓		✓	
CBCap	✓			✓
Emp	✓			
ESec	✓		✓	
CRoe	✓			
CovD II	✓		✓	
GrRRev	✓			
CCint	✓			
CCTLib	✓			
CovD I			✓	
Age			✓	
CovInt			✓	
BmTA		✓		

Source: author

The dominance of liquidity, activity, and leverage indicators is not surprising and has a clear economic interpretation in the context of B2B transactions. More precisely, the cash gap (CGap) indicates the period between settling liabilities and collecting receivables. The longer this gap, the higher the pressure on the company's liquidity and its ability to pay suppliers on time. The current ratio (CLig) measures a company's ability to cover its short-term liabilities with its current assets, evaluating its capacity to maintain sufficient liquidity under unexpected financial pressures. Activity indicators (QtCA, QtTA) show the rate at which a company converts assets into liquid assets. High leverage, as measured by the capital financing ratio (QCap), further reduces flexibility in cash flow management and increases the company's vulnerability to disruptions in debt collection. Together, these indicators describe the financial condition of the company and reflect the underlying economic mechanisms that directly determine its ability to meet its liabilities to suppliers within the legal deadline.

The fact that the same groups of financial indicators dominate all models, regardless of methodological approach, further confirms the robustness of these findings and suggests that their explanatory role is a structural characteristic of B2B

credit risk in this institutional environment, rather than a result of the specific method used.

At the same time, the limited and model dependent importance of non-financial variables, such as sector, size measured by number of employees, and age of the company, indicates that in the context of large non-financial companies in B&H, financial information is the primary signal of credit risk, while contextual factors play a complementary role that depends on the specifics of each model.

4.3. Evaluation of the model's stability and predictive power

The stability and predictive power of the model were assessed across three complementary dimensions: discriminative ability, temporal stability, and calibration of predicted probabilities. The model developed using the Gradient Boosting algorithm was examined in greater detail, as it had shown the highest discriminatory performance in previous phases of the analysis.

The initial model comparison used a stratified hold-out approach (60/40), while an additional stability check was conducted using time validation over two periods (2015–2018 and 2015–2019). This approach provides insight into the robustness of the model as the dataset expands and the temporal structure changes. The classification results, including performance metrics for PD and Non-PD status prediction, are presented in the table below.

Table 8 Classification performance metrics for PD status prediction

FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)	FALSE POSITIVE (FP)	TRUE POSITIVE (TP)	TOTAL 2015-2018
424	1721	285	1385	3815
			Accuracy	81.42%
			Precision	82.93%
			Recall	76.56%
			F1 Score	78.91%
			Classification accuracy: PD status	82.93%
			Classification accuracy: Non-PD status	80.23%
FALSE NEGATIVE	TRUE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE	TOTAL 2015-2019
524	2216	375	1713	4828
			Accuracy	81.38%
			Precision	82.04%
			Recall	76.58%
			F1 Score	78.90%
			Classification accuracy: PD status	82.04%
			Classification accuracy: Non-PD status	80.88%

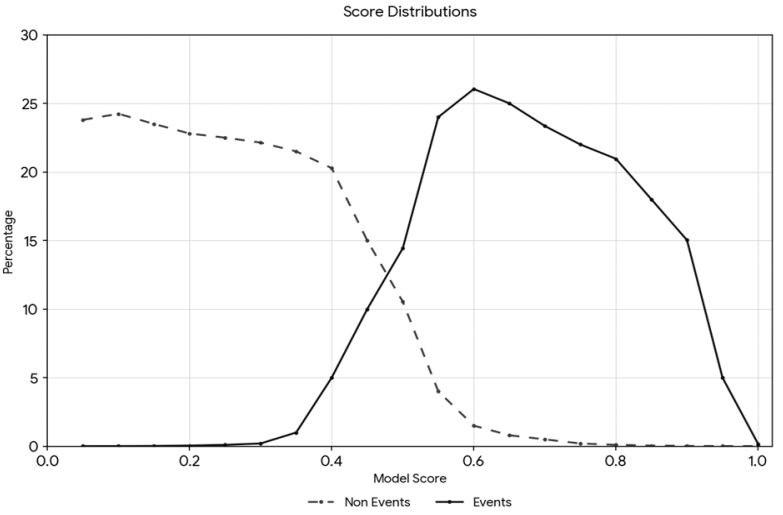
Source: research results from SAS Enterprise Miner

The results indicate high model stability, with overall accuracy consistently around 81.4% in both time cross-sections and the F1-score remaining at approximately 78.9%, confirming the model's good generalisation ability. However, detailed analysis shows that the model achieves higher accuracy for the

response (Recall), indicating that it relatively successfully minimises false positive classifications, but at the same time fails to identify some actual cases of non-fulfilment of liabilities (False Negatives). This result is particularly important in the context of credit risk management in the B2B environment, as misidentification of risky entities can have direct financial consequences for creditors. The results therefore suggest that future research should consider optimising the classification threshold or including additional variables to improve the model's sensitivity.

Additionally, model calibration was analysed using the Model Score approach (Figure 2). Unlike the ROC curve, which visualises cumulative event rates, this method compares the rates of events (PD = 1) and non-events (PD = 0) with estimated probabilities. The procedure involves ranking business entities by estimated probabilities and dividing them into twenty equal groups. Within each group, the proportions of events and non-events are analysed. For each 5% segment of the ranked data, the observed event rate (PD = 1) is calculated. The same procedure is applied to non-events (PD = 0).

Figure 2 Model calibration plot based on predicted probabilities



Source: model results obtained using SAS Enterprise Miner

The results display the expected pattern: the proportion of actual credit risk events increases as the estimated probability rises, while the proportion of regular payers decreases, indicating good discriminatory and calibration capabilities of the models. The intersection point of the curves falls within the range 0.45–0.55, which can be interpreted as an empirically optimal classification

threshold. At this level, the model balances sensitivity and specificity, providing a relevant starting point for business decisions regarding credit risk acceptance.

From a business application perspective, the choice of classification threshold directly affects the balance between false positive and false negative classifications. Entities with an estimated PD above this threshold carry significantly higher credit risk, while those below the threshold represent a more stable part of the portfolio. This enables the model to differentiate clients according to their risk level, which is particularly important in the B2B context, where individual incorrect decisions can significantly impact the liquidity and financial stability of creditors. It is important to emphasise that the optimal classification threshold in practice is not necessarily fixed, but depends on risk management preferences, particularly the relative costs of false positive and false negative decisions. In situations of higher credit risk aversion, the threshold could be lowered to reduce the number of unrecognised risky entities, while in other cases it could be set higher. Although no formal optimisation of the threshold based on cost functions was conducted in this study, the results indicate potential for its adjustment, representing an important direction for future research.

To further quantitatively assess the model's calibration using a formal statistical measure, the Brier score was calculated. This score measures the average squared deviation between the estimated probabilities and the actual outcomes, and is a standard metric for evaluating the calibration quality of probabilistic models (Brier, 1950; Hastie et al., 2009). Lower values indicate better calibration, with a value close to zero representing perfect prediction and a value of 0.25 corresponding to a non-informative model that always predicts the base rate.

Table 9 Calibration Statistics – Gradient Boosting Model

Measure	Value
Brier score	0.156628
ASE (validation sample, Table 2)	0.15790

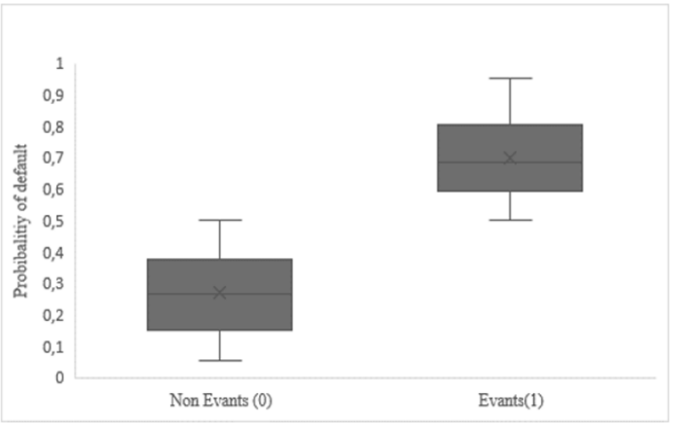
Source: research results from SAS Enterprise Miner

The obtained value of 0.156628 indicates a relatively good alignment between the estimated probabilities of the credit risk event and the observed event outcomes. This result is consistent with the average square error (ASE) of 0.15790 shown in Table 2 for the validation sample, with minimal difference, further confirming the stability and reliability of the estimated probabilities. A value below 0.2 in complex classification problems is generally considered satisfactory, especially in domains characterised by high data heterogeneity and pronounced dynamics in subject behaviour. The results are not directly comparable to banking models, where values lower than 0.1 are often achieved, as banking systems have highly structured, standardised, and regulatory controlled data with more stable patterns of client behaviour. In contrast, B2B transactional data are characterised

by higher volatility, heterogeneity of business relationships, and varying payment patterns, making the prediction task inherently more challenging.

Analysis of the distribution of estimated probabilities (box plot) enabled further segmentation of entities by risk level (Figure 3).

Figure 3 Boxplot of estimated probabilities for classification of subjects – Gradient Boosting Model



Source: model results obtained using SAS Enterprise Miner

The results indicate that the average estimated PD is approximately 0.27 for entities that regularly settle their payment liabilities, and about 0.70 for entities that do not settle their obligations within the legally defined 60-day threshold. Based on these values, subjects are ranked into three credit risk levels: those with values up to 0.27 are classified as low risk, those in the range from 0.27 to 0.70 as medium risk, and those with values above 0.70 as high credit risk. This segmentation does not replace the previously defined classification threshold (approximately 0.5), but complements it. While the classification threshold enables a binary decision on credit risk event status, segmentation based on the distribution of estimated probabilities allows finer differentiation of entities within the portfolio and can contribute to more detailed credit risk management. From a calibration perspective, the low-risk segment shows good separation from riskier entities, the high-risk segment demonstrates a higher concentration of realised credit risk events, confirming good alignment with observed outcomes, and the medium-risk segment represents the area of greatest uncertainty, where discriminatory power is naturally lower. This cross-segment pattern is consistent with the Brier score of 0.156628, suggesting reasonably stable calibration across risk levels.

These results are relevant in the context of IFRS 9, which prescribes ECL estimation based on PD and segmentation of exposure according to credit risk levels. The developed model may serve as a starting point for PD assessment and

portfolio segmentation within the IFRS 9 conceptual framework, although further external validation, calibration refinement, and governance procedures are required before any practical regulatory or accounting application.

Further improvements, such as model calibration or optimisation of the classification threshold based on cost functions, could enhance its practical applicability. Additional calibration methods, such as Platt scaling or isotonic regression, could further improve the alignment between estimated probabilities and actual outcomes, which is particularly relevant for IFRS 9 reporting. The validation framework combines a stratified hold-out approach (60/40) with temporal stability analysis across two periods, demonstrating consistent model performance. More advanced validation procedures, such as k-fold cross-validation, are a direction for future research.

5. DISCUSSION

The results of this study confirm that ML techniques can be successfully applied to credit risk assessment models in B2B transactions within the non-financial sector, thus addressing the first research question (RQ1). The findings show that all three ML models observed exhibit high discriminatory ability in distinguishing between subjects with compliant and non-compliant settlement of liabilities, which is consistent with previous research (Tian et al., 2020; Suhadolnik et al., 2023). This highlights the advantages of ensemble methods in capturing nonlinear relationships between financial and non-financial variables (Friedman, 2002; Noriega et al., 2023; Suhadolnik et al., 2023; Ayari et al., 2025). However, as the models are trained with default hyperparameters, it is not possible to rule out the possibility that further optimisation could affect the relative differences in performance.

Regarding the second research question (RQ2), the comparison of the models shows that all four methods achieve high and mutually comparable discriminatory power, with Gradient Boosting performing slightly better according to all observed evaluation criteria. However, the differences in performance between the models are relatively small and should be interpreted with caution. This finding is analytically interesting because it suggests that key predictors of credit risk in this B2B environment are sufficiently structured for logistic regression to identify them effectively, while ML methods retain an advantage in recognising a wider set of predictors and more complex nonlinear interactions between variables, as evidenced by the larger number of relevant variables identified in the Gradient Boosting model. It is important to emphasise that logistic regression, despite comparable discriminatory power, required extensive data preprocessing, including logarithmic transformation of 19 variables, winsorisation of 20 variables, and imputation of missing values for 24 variables, while ML models achieve high discriminatory accuracy on original, untransformed variables. This makes them operationally simpler to implement in practical applications, especially in business environments where data is often not ideally distributed. Transparency of interpretation remains an advantage of logistic regression, as each

predictor has a clear direction and magnitude of influence on the dependent variable, but this comes with methodologically more demanding data preparation. Furthermore, since ML models were trained using standard default hyperparameters, the reported advantage of Gradient Boosting should be understood as a conservative assessment of the potential of the ML approach, and the additional optimisation of hyperparameters represents direction for future research. Beyond discriminatory ability, the stability and predictive power of the model with the highest discriminatory power were further assessed through temporal validation and probability calibration. The Gradient Boosting model shows consistent performance across both validation periods, confirming good generalisation ability, while the Brier score confirms relatively good calibration of estimated probabilities, further supporting the reliability of the model's probability estimates.

The analysis of variable importance shows that financial indicators of liquidity, activity, and leverage are consistently the most important determinants of credit risk across all models, addressing the third research question (RQ3). These results confirm existing findings in the literature that emphasise the dominance of financial indicators in predicting the creditworthiness of companies (Giampaolo et al., 2020; Gafsi, 2025; Wu et al., 2025) and are consistent with theoretical expectations from the trade credit literature (Seifert et al., 2013). The interpretation of the identified predictors in the developed models should not be equated with traditional banking models of credit risk, even though the results align with existing literature on credit risk in banking and the prediction of financial difficulties. As the dependent variable reflects the exceeding of the legally defined threshold of 60 days, the statistical significance of the predictor should be interpreted within the broader context of commercial relations in the non-financial sector. In B2B relationships, entering the zone of legally defined illiquidity may indicate an asymmetry of bargaining power between business partners, administrative inefficiencies, or deliberate strategic management of cash flows (Wilner, 2000; Hill et al., 2010), and does not necessarily signal the debtor's final financial collapse. However, from the perspective of IFRS 9, such behaviour represents a deterioration in credit quality because it distorts the time value of money. Therefore, the identified predictors provide relevant information for credit risk assessment and the calculation of ECL in the non-financial sector. Companies facing liquidity pressures, weaker operational efficiency, and high leverage have a structurally reduced ability to meet vendor liabilities in a timely manner (Deloof, 2003; Hill et al., 2010; Wilner, 2000). The fact that the same financial indicators consistently dominate across all models, regardless of whether a simple or complex method is used suggests that these are genuine drivers of payment behaviour in this business environment, not a coincidence driven by the choice of modelling technique. This is particularly relevant in the B&H context, where late payments are a common feature of business relationships rather than an isolated occurrence. The contribution of this study therefore lies not in revealing an unexpected mechanism, but in providing empirical confirmation of these relationships within a specific and underexplored context, B2B transactions in the non-financial sector of a transitional economy where such evidence has been largely absent.

In contrast, non-financial variables appear as additional but less stable predictors, which may reflect the specifics of the B2B context, where financial information provides the main content regarding credit risk, and the limited availability of qualitative non-financial information in the database used. Moreover, the extensive data preprocessing required for developing the logistic regression model empirically confirms that financial data in the B2B non-financial sector are highly asymmetric and influenced by extreme values. This highlights the real heterogeneity among business entities in this environment and further supports the use of ML methods, which are inherently robust to such data characteristics. Future research incorporating a broader range of non-financial data could further clarify their role in explaining credit risk in the B2B environment.

Regarding the fourth research question (RQ4), key limitations of implementation have been identified. Although ensemble models achieve high performance, their interpretability remains structurally limited. Unlike logistic regression, where the effect of each predictor on the dependent variable is clearly defined in terms of direction and strength, ML models make decisions through complex combinations of trees, making their internal decision-making process difficult for end users, credit risk managers, auditors, or regulators to understand. In this paper, measures of variable importance were used as the primary tool for interpretation, but these do not reveal the direction of influence of a particular variable or quantify interactions between predictors. In this respect, logistic regression retains a clear interpretability advantage, as each predictor has a defined coefficient that directly indicates the direction and magnitude of its effect on the dependent variable. In contrast, tree-based ML models distribute predictive information across hundreds of trees and interactions, making it difficult to attribute model decisions to individual variables in a transparent and auditable manner. This trade-off between interpretability and predictive performance is particularly relevant in accounting and regulatory contexts, where model transparency is not merely desirable but may be required. Advanced methods of explainable artificial intelligence (XAI), such as SHAP values (Lundberg & Lee, 2017), would allow explanation of individual predictions, which is particularly important in the B2B context and within the IFRS 9 regulatory framework, where transparency of ECL estimation is a regulatory requirement. The application of these methods was limited by the features available in the SAS Enterprise Miner used for the empirical analysis; therefore, extending this research by incorporating advanced XAI techniques on more flexible software platforms represents a direction for future research.

In addition to the issue of interpretability, the practical application of the model faces challenges related to data quality. The extensive pretreatment required for logistic regression, the transformation or correction of almost all variables, empirically confirms the high asymmetry of financial data in a B2B environment. Although ML models demonstrate higher robustness to such characteristics, the quality of input data remains a key factor in the reliability of forecasts, which can present significant organisational and technical challenges in the business environment of transition economies. The analysis was conducted using data from

a single national market, so the results should be interpreted within the specific institutional and economic context of B&H. While the identified patterns may also be relevant for similar transition economies, their application to other markets should be considered with caution due to possible differences in the institutional framework, regulatory environment, and business practices. Furthermore, the analysis is based on historical data from stable economic conditions, and examining model behaviour under different economic conditions remains a direction for future research.

6. CONCLUSION

This paper analyses the application of ML techniques in credit risk modelling for B2B transactions within the non-financial sector, using logistic regression as a reference model for the classical statistical approach. Empirical results show that all models achieve a high level of discriminatory ability, with ensemble-based models, particularly Gradient Boosting, delivering slightly better results than logistic regression according to the observed evaluation criteria. However, the differences in performance between models are not large, suggesting that in this context, traditional statistical approaches remain highly competitive.

Comparison of the findings with the existing literature shows consistency with previous empirical research highlighting the advantages of ensemble methods in credit risk modelling. This research also contributes to the literature on several levels: it extends the empirical application of credit risk modelling to B2B transactions in the non-financial sector of a transition economy; introduces an institutionally based definition of legally defined payment delay (illiquidity), grounded in B&H legislation; and empirically demonstrates that ML models achieve high predictive accuracy using original, untransformed variables, without the extensive data preprocessing required for classical statistical approaches. The consistency of the most important predictor variables across different models provides further evidence of the robustness of the findings, although validation on independent datasets from different institutional contexts is necessary to draw firm conclusions about generalisability.

From a practical perspective, the results enable the differentiation of business entities by credit risk levels and support decision-making in credit risk management through an early warning framework for managers in sectors with a high proportion of B2B transactions. The identified financial indicators, particularly those related to liquidity, leverage, and operational efficiency, can serve as early signals of credit risk and inform decisions to grant payment deferrals or credit terms to business partners. The model's particular value lies in its focus on predicting legally defined payment delays and illiquidity, rather than conventional banking default, which increases its relevance for the non-financial sector. Exceeding the legal threshold of 60 days represents a relevant signal of elevated credit risk, regardless of its underlying causes - whether liquidity difficulties, bargaining power asymmetries, or operational factors - which underscores the need to interpret the results within a broader economic context. However, practical application requires adaptation to the specifics of the particular business environment and continuous revalidation of the model. The research also contributes to extending the application of

IFRS 9 to credit risk analysis for non-financial companies, providing an empirical basis for PD assessment in the context of B2B receivables and further operationalising the IFRS 9 framework outside the banking sector, with emphasis on trade receivables and the non-financial sector. The model is not a regulatory solution, but an empirical tool that requires further external validation, calibration, and robustness testing before possible application in reporting or monitoring frameworks.

Despite these contributions, this research has some limitations. The analysis was conducted on data from a single national context and during a relatively stable five-year economic period, which may limit the generalisation of results to other markets, particularly under conditions of economic shocks or periods of crisis. Although the selected sectors cover approximately 80% of all large non-financial companies in B&H, the applicability of the findings to other sectors, SMEs, and different institutional contexts may be limited. The research uses standard hyperparameters and a stratified hold-out approach, ensuring comparability and transparency, while more advanced optimisation techniques and cross-validation methods represent directions for future research. The interpretability of complex ML models remains limited despite the use of variable importance measures, and the application of advanced XAI methods such as SHAP values - not available within the software used - also represents a direction for future research.

Overall, the research results indicate the potential of ML methods for credit risk assessment in the B2B segment of the non-financial sector, but their implications should be interpreted with caution given these limitations. The paper provides a relevant empirical basis for further development, methodological improvements, and broader application in credit risk management.

Author Contributions: The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Funding: The research presented in this manuscript did not receive any funding from external sources.

Conflict of Interest: None.

Acknowledgement of AI or AI-assisted tools use: During the preparation of this paper, the author used InstaText tool in order to improve spelling, grammar correction, word choice, readability, and clarity. After using the tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgement: This paper is based on research conducted as part of a doctoral dissertation entitled “Risks effects on the expected credit losses in non-financial sector.”

REFERENCES

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Alonso, A., & Carbó, J. M. (2020). Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost. *Working Paper No. 2032*. Banco de España. https://doi.org/10.2139/ssrn.3724374

Altman, E. I., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance, 21*(11–12), 1721-1742. https://doi.org/10.1016/S0378-4266(97)00036-8

Ayari, H., Guetari, P. R., & Kraïem, P. N. (2025). Machine learning powered financial credit scoring: A systematic literature review. *Artificial Intelligence Review, 59,* 13. https://doi.org/10.1007/s10462-025-11416-2

Bandyopadhyay, A. (2016). Managing portfolio credit risk in banks. Cambridge University Press. https://doi.org/10.1017/CBO9781316550915

Basel Committee on Banking Supervision (BCBS). (2011, revised 2019). Basel III: A global regulatory framework for more resilient banks and banking systems. Bank for International Settlements.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71-111. https://doi.org/10.2307/2490171

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140. https://doi.org/10.1007/BF00058655

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

Chaibi, H., & Ftiti, Z. (2015). Credit risk determinants: Evidence from a cross-country study. *Research in International Business and Finance*, Elsevier, *33*(C), 1-16. https://doi.org/10.1016/j.ribaf.2014.06.001

Chen, D., Ye, J., & Ye, W. (2023). Interpretable selective learning in credit risk. *Research in International Business and Finance, 65*, 101940. https://doi.org/10.1016/j.ribaf.2023.101940

Chowdhury, R. H. (2025). A machine learning framework for credit risk mitigation: Assessing the impact of AI and blockchain integration. *International Journal of Research and Innovation in Applied Science,* 233-263. https://doi.org/10.51584/IJRIAS.2025.10060014

Deloof, M. (2003). Does working capital management affect profitability of Belgian firms? Journal of Business Finance & Accounting, *30*(3–4), 573–588. https://doi.org/10.1111/1468-5957.00008

Džidić, A., Živko, I., & Čolak, A. (2022). Macroeconomic Factors of Non-Performing Loans: The Case of Bosnia and Herzegovina. *Ekonomska misao i praksa*, *31*(2), 421–438. https://doi.org/10.17818/EMIP/2022/2.4

De Oliveira, N. A., & Basso, L. F. C. (2025). Advancing credit rating prediction: The role of machine learning in corporate credit rating assessment. *Risks, 13*(6), 116. https://doi.org/10.3390/risks13060116

European Parliament, & Council of the European Union. (2013). Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012. Official Journal of the European Union, L 176, 1–337.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*, 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gafsi, N. (2025). Machine learning approaches to credit risk: Comparative evidence from participation and conventional banks in the UK. *Journal of Risk and Financial Management, 18*(7), 345. https://doi.org/10.3390/jrfm18070345

Giampaolo, G., Matthias, M., & Giammarino, M. (2020). Modeling hard and soft facts for SMEs: Some international evidence. *Journal of International Financial Management and Accounting, 30*, 203-222. https://doi.org/10.1111/jifm.12108

Grebenar, T. (2018). A behavioral model for assessing the probability of default events and the rating of non-financial enterprises. (Bihevioralni model procjene vjerojatnosti nastanka događaja neispunjavanja obveza i rejting nefinancijskih poduzeća). *Istraživanja I-53*, Hrvatska Narodna Banka. Dostupno na: https://www.hnb.hr/documents/20182/2259329/i-053.pdf

Gulin, D. (2009). Cash flow management using net working capital (Upravljanje novčanim tokom pomoću neto obrtnog kapitala). *Računovodstvo i financije, 11*, 63-65.

Gulin, D., & Hladika, M. (2018). Cash flow report for 2017 (Izvještaj o novčanim tokovima za 2017. godinu). *Računovodstvo i financije, LXIV*(1), 104-113.

Gulin, D., Tušek, B., & Žager, L. (2004). *Business planning, control, and analysis (Poslovno planiranje, kontrola i analiza).* Zagreb: Hrvatska zajednica računovođa i financijskih djelatnika.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hill, M. D., Kelly, G. W., & Highfield, M. J. (2010). Net operating working capital behavior: A first look. Financial Management, 39(2), 783–805. https://doi.org/10.1111/j.1755-053X.2010.01092.x

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/0471722146

International Accounting Standards Board (IASB). (2014). IFRS 9 Financial Instruments. London: IFRS Foundation.

Kharabsheh, B. (2019). Determinants of Bank Credit Risk: Empirical Evidence from Jordanian Commercial Banks. Yarmouk University, *Academy of Accounting and Financial Studies Journal*, *23*(3).

Kontuš, E. (2013). Management of accounts receivable in a company. *Economic Thought and Practice*, *22*(1), 21–38.

Law on Accounting and Auditing of the Federation of Bosnia and Herzegovina (2009) "Official Gazette of the Federation of Bosnia and Herzegovina", No. 83/09.

Law on Accounting and Auditing of the Federation of Bosnia and Herzegovina (2021) "Official Gazette of the Federation of Bosnia and Herzegovina", No. 15/21.

Law on Deadlines for the Execution of Monetary Obligations in Business Transactions in the Republic of Srpska (2018) "Official Gazette of RS", No. 31/18.

Law on Financial Operations in the Federation of Bosnia and Herzegovina (2016) "Official Gazette of the Federation of B&H", No. 48/16.

Liu, S. (2024). Application analysis of machine learning models in credit risk. *Highlights in Science Engineering and Technology, 107,* 76–81. https://doi.org/10.54097/6qm0dk77

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS).

Ljubičić, T., & Hell, M. (2023). Student success prediction using artificial neural networks. *Economic Thought and Practice*, *32*(2), 361–374. https://doi.org/10.17818/EMIP/2023/2.3

Machokoto, M., Gyimah, D., & Ibrahim, B. M. (2022). The evolution of trade credit: New evidence from developed versus developing countries. Review of Quantitative Finance and Accounting, *59*(3), 857–912. https://doi.org/10.1007/s11156-022-01061-z

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Leanpub

Nehrebecka, N. (2015). Approach to the assessment of credit risk for nonfinancial corporations. Evidence from Poland, IFC Bulletins chapters, in: Bank for International Settlements (ed.), Combining micro and macro data for financial stability analysis, *41*, Bank for International Settlements.

Noriega, M., Rivera, A., & Herrera, L. (2023). Machine learning techniques for credit risk prediction: *A systematic literature review. Data, 8*(11), 169. https://doi.org/10.3390/data8110169

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109-131. https://doi.org/10.2307/2490395

Provenzano, A. R., Trifirò, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., Le Pera, G., Spadaccino, M., Massaron, L., & Nordio, C. (2020). Machine learning approach for credit scoring. *arXiv Papers,2008.01687,* 1-28.

Rabega, C. (2024). Financial management: The cornerstone of organizational success. *Business Studies Journal, 16*(2), 1-2.

Seifert, D., Seifert, R. W., & Protopappa-Sieke, M. (2013). A review of trade credit literature: Opportunities for research in operations. *European Journal of Operational Research*, *231*(2), 245–256. https://doi.org/10.1016/j.ejor.2013.03.016

Seitshiro, M. B., & Govender, S. (2024). Credit risk prediction with and without weights of evidence using quantitative learning models. *Cogent Economics & Finance, 12*(1). https://doi.org/10.1080/23322039.2024.2338971

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management, 16*(12), 496. https://doi.org/10.3390/jrfm16120496

Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science, 174,* 150–160. https://doi.org/10.1016/j.procs.2020.06.070

Tintor, Ž. (2020). Analysis of financial statements in the function of making better business decisions. *Education for Entrepreneurship-E4E Scientific and Professional Journal on Education for Entrepreneurship*, *10*(1), 84-103. https://doi.org/10.38190/ope.10.1.4

Vlah Jerić, S. (2023). Analysis of the financial performance of machine learning models for predicting the direction of changes in CEE and SEE stock market indices with different classification evaluation metrics. *Economic Thought and Practice*, *32*(2), 533–545. https://doi.org/10.17818/EMIP/2023/2.12

Wang, L., Zhang, M., & Liu, Z. (2023). The progress of Black-Scholes model and Black-Scholes-Merton model. *BCP Business & Management, 38*, 3405-3410. https://doi.org/10.54691/bcpbm.v38i.4314

Wilner, B. S. (2000). The exploitation of relationships in financial distress: The case of trade credit. The Journal of Finance, *55*(1), 153–178. https://doi.org/10.1111/0022-1082.00203

Wu, J., Rui, O. M., & Wu, W. (2020). Trade credit, cash holdings, and financial constraints: Evidence from China. *Journal of Banking & Finance*, 113, 105760. https://doi.org/10.1016/j.jbankfin.2020.105760

Wu, Y., Li, H., & Zhang, Q. (2025). Predicting credit spreads and ratings using machine learning: The role of financial and non-financial indicators. *arXiv Papers, 2509.19042.* https://doi.org/10.48550/arXiv.2509.19042

Zhang, X., Xu, L., Li, N., & Zou, J. (2024). Research on credit risk assessment optimization based on machine learning. *In Proceedings of the 6th International Conference on Computing and Data Science*. https://doi.org/10.54254/2755-2721/69/20241497

Žager, K., Mamić Sačer, I., Sever Mališ, S., Ježovita, A., & Žager, L. (2017). *Financial statement analysis: Principles, procedures, and cases (Analiza financijskih izvještaja: načela – postupci – slučajevi)* (3rd ed.). Zagreb: HZRIF.

Dr. sc. Suzi Mikulić
Sveučilište u Mostaru, Bosna i Hercegovina
E-mail: suzi.mikulic@sum.ba
Orcid: https://orcid.org/0000-0001-8730-7432

PRIMJENA TEHNIKA RAČUNALNOG UČENJA U RAZVOJU MODELA PROCJENE KREDITNOG RIZIKA U TRANSAKCIJAMA IZMEĐU PODUZEĆA NEFINANCIJSKOG SEKTORA

Sažetak

Kreditni rizik u transakcijama između poduzeća može ugroziti njihovu financijsku stabilnost ako se njime ne upravlja učinkovito. Ovaj rad istražuje primjenjivost metoda strojnog učenja u predviđanju kreditnog rizika u B2B transakcijama na uzorku od 4.828 opažanja velikih poduzeća iz Bosne i Hercegovine tijekom petogodišnjeg razdoblja. Razvijena su tri modela temeljena na stablima odlučivanja (Bagging Decision Tree, Random Forest i Gradient Boosting), dok je logistička regresija korištena kao referentni model. Rezultati pokazuju da svi modeli postižu visoku prediktivnu uspješnost, pri čemu je model Gradient Boosting ostvario najbolje rezultate. Najvažnijim prediktorima pokazali su se financijski pokazatelji likvidnosti, aktivnosti i zaduženosti, dok je doprinos nefinancijskih varijabli bio ograničen. Nalazi potvrđuju važnost računovodstvenih informacija u procjeni kreditnog rizika te mogu pridonijeti daljnjem razvoju modela za primjenu u okviru zahtjeva standarda MSFI 9, gdje je vjerojatnost neispunjenja obveza ključna komponenta procjene očekivanih kreditnih gubitaka prema općem pristupu.

Ključne riječi: računalno učenje, modeli procjene kreditnog rizika, B2B transakcije, MSFI 9, nefinancijski sector.

JEL klasifikacija: C45, G32, M41.