Aalto-yliopisto
Aalto-universitetet
Aalto University

Bachelor's Programme in Accounting

Hybrid Ensemble Model Based Approach to Bankruptcy Prediction in European SMEs

An Empirical Study Using Soft Voting Ensembles and SME Financial Data

---

Lucas Tepponen

Bachelor's thesis
2025

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Lucas Tepponen

Title of thesis Hybrid Ensemble Model Based Approach to Bankruptcy Prediction

Programme Aalto University School of Business

Major Accounting

Thesis supervisor Roope Keloharju

Date 11.5.2025 Number of pages 27+17 Language English

Abstract

The purpose of this thesis is to examine whether a simple hybrid model using logistic regression and XGBoost can effectively predict corporate bankruptcy while remaining interpretable and practical use in business environments.

The study uses a large, multi-country dataset of European SMEs from the Orbis database, and compares the performance, calibration, and interpretability of the individual models and their soft voting ensemble.

The results show that the hybrid model retains the high predictive accuracy of XGBoost while adding transparency through the logistic regression component. This approach offers a practical compromise between accuracy and explainability, especially for stakeholders without technical expertise.

Keywords Bankruptcy prediction, Logistic regression, XGBoost, Hybrid models, Interpretability, Financial distress, Machine learning

Tiivistelmä
Tämä kandidaatintutkielma tarkastelee, voiko yksinkertainen hybridimalli, joka yhdistää logistisen regressiomallin ja XGBoost-algoritmin, ennustaa yritysten maksukyvyttömyyttä tarkasti ja samalla tarjota riittävää tulkittavuutta käytännön liiketoimintaympäristöissä.

Tutkimuksessa hyödynnetään laajaa, monen Euroopan maan pk-yrityksistä koostuvaa aineistoa Orbis-tietokannasta. Mallien ennustetarkkuutta, kalibrointia ja tulkittavuutta verrataan yksittäin ja yhdistettynä.

Tulokset osoittavat, että hybridimalli säilyttää XGBoostin korkean suorituskyvyn ja samalla parantaa läpinäkyvyyttä logistisen regression ansiosta. Malli tarjoaa käytännönläheisen tasapainon tarkkuuden ja selitettävyyden välillä, erityisesti teknistä taustaa omaamattomille sidosryhmille.

Avainsanat Konkurssiennustaminen, Logistinen regressio, XGBoost, Hybridimallit, Pk-yritykset, Tulkittavuus, Koneoppiminen, Soft voting -menetelmä, Yritysriskien mallinnus.

Table of contents

1 Introduction

1.1 Background and motivation

As former astronaut and businessman Frank Borman put it: “Capitalism without bankruptcy is like Christianity without hell”. This quote underlines the fact, that the possibility of failure is not a systemic flaw, but rather a defining feature. Bankruptcy plays an important role in shaping markets and allocating resources. For lenders, auditors, and regulators, predicting it accurately and early is essential (Park et al., 2021). These predictions also support internal decision-making, since managers may use early warning signals to start restructuring, reduce operational risk, or seek external financing. From a management accounting perspective, predicting financial distress early is crucial for decision-making, budgeting, and risk management.

Due to its importance, bankruptcy prediction has been a popular topic in accounting and finance in the last century (Altman, 1968; Ohlson, 1980; Bellovary et al., 2007). Traditional models such as the Altman Z-score and logistic regression-based approaches, remain widely used even today, especially due to their simplicity and interpretability. However, with the rise of machine learning (ML) methods, newer models have demonstrated better predictive performance and propose a strong alternative to these traditional methods (Iparraguirre-Villanueva and Cabanillas-Carbonell, 2024; Sun et al., 2014). These techniques often outperform traditional models in terms of raw accuracy, but their complexity can make them difficult to implement or trust in business environments (Yusuf et al., 2024; Doshi-Velez and Kim, 2017).

This research is motivated by the need for predictive models that balance accuracy with accessibility and transparency. It explores whether

combining a powerful machine learning model with a familiar, interpretable one in a simple ensemble model can serve as a practical middle ground, one that is easier to develop, understand, and apply in a corporate setting compared to complex ML solutions.

1.2 Research problem and objectives

As mentioned earlier, the adoption of ML-methods remains challenging, particularly among small and medium-sized enterprises (Zavodna et al., 2024). Many of the most accurate models, such as ensemble methods or neural networks, are difficult to interpret, which can limit their transparency and trustworthiness in decision-making. Furthermore, implementing and understanding these methods often requires extensive knowledge in complex technical fields, which might make them less practical, especially for SMEs. It should also be considered that the results of these models often need to be explained to stakeholders with non-technical backgrounds.

As Doshi-Velez and Kim (2017) argue, the lack of interpretability in high-performing models creates barriers to trust, especially when decisions carry significant consequences. Simpler models such as logistic regression are interpretable but often struggle in modeling complex financial data. This trade-off between predictive performance and interpretability is a well-documented challenge in applied machine learning and financial risk modelling. More complex models, such as XGBoost, typically achieve higher predictive accuracy but are often viewed as “black-box” models, which can limit trust and adoption in high-stakes domains like finance (Carmona et al., 2022).

The objective of this study is to investigate whether a simple soft-voting ensemble model combining logistic regression and XGBoost can offer a practical compromise between accuracy and interpretability. The model is not

designed to outperform XGBoost on predictive metrics alone, rather, it aims to maintain XGBoost's high performance while enhancing interpretability with the inclusion of logistic regression. This approach provides users with a built-in, transparent component, reducing the need to rely solely on post-hoc explainability tools such as SHAP or LIME, which may be less accessible to non-technical stakeholders.

Unlike more complex models that require post-hoc explainability tools, this study deliberately focuses on a lightweight, interpretable hybrid that can be realistically implemented by practitioners without deep ML expertise. This practical emphasis distinguishes the approach from many prior studies that prioritize technical performance over usability.

1.3 Structure of the thesis

The structure of this thesis is as follows. Chapter 2 covers prior research in bankruptcy prediction, covering both traditional statistical methods and machine learning based techniques, with a focus on interpretability and hybrid model design. Chapter 3 covers the data sources, variable selection, and methodological choices, including the development of the models used in this study. Chapter 4 presents the results of the study and compares the predictive performance and calibration of the models. Chapter 5 discusses the findings in a more detailed manner, focusing on interpretability, practical applications, limitations and future research.

2 Literature review

2.1 Traditional approaches to bankruptcy prediction

Bankruptcy prediction models have been a popular topic in accounting and finance for decades. Before the 1960's, the methods used to predict bankruptcy often relied on comparing single financial factors between struggling and surviving firms (Beaver, 1966; Bellovary et al., 2007). This approach, while at the time novel, inherently struggles to capture the relationships between multiple financial variables. As a result, the next step was moving towards multivariate models (Bellovary et al., 2007).

2.1.1 Altman Z-Score

One of the most influential models in early bankruptcy prediction literature was the Altman Z-Score model introduced by Altman (1968). This model applied multiple discriminant analysis (MDA) to predict bankruptcy. It used five financial ratios related to liquidity, profitability, leverage, solvency, and efficiency. These ratios were combined into a weighted score to classify a company as bankrupt or non-bankrupt (see Appendix A for formula). His Z-score model performed well in predicting bankruptcy in manufacturing firms and quickly became a benchmark in academic research (Altman, 1968; Bellovary et al., 2007). Even so, the method relies on assumptions such as multivariate normality and equal covariance matrices between groups, which may not hold up in financial datasets (Bellovary et al., 2007).

2.1.2 Logistic regression

In the decades following Altman's Z-score model, researchers began to develop other statistical methods for bankruptcy prediction. One of the most notable ones was the O-score introduced by Ohlson (1980). In his model, Ohlson approached bankruptcy prediction by using logistic regression (LR). This approach made it possible to estimate bankruptcy probabilities without having to rely on the same assumptions as Altman's Z-score model did (Ohlson, 1980). In logistic regression models, the likelihood of bankruptcy is presented as a logistic function of a linear combination of predictor variables (see Appendix A for formula).

Ohlson's model was built using a relatively large dataset of 105 bankrupt and 2,058 non-bankrupt companies (Ohlson, 1980). The O-score used nine financial ratios based on financial statement information, focusing on firm size, leverage, performance, and liquidity (Ohlson, 1980). By using logistic regression rather than discriminant analysis, Ohlson addressed the statistical limitations of earlier models.

2.2 Limitations of Traditional Models

Together the Z-score and O-score models have laid the foundation for quantitative bankruptcy prediction research, but they are not without limitations. Both models primarily rely on linear relationships between financial ratios and firm failure, which may not fully capture the non-linear relationships between variables often present in financial distress (Sun et al., 2014). Furthermore, these models focus solely on accounting-based financial statement data, overlooking the potential value of non-financial information such as corporate governance, macroeconomic conditions, or market sentiment (Sun et al., 2014). In addition, as Altman et al. (2017) emphasize, the predictive performance of these models tends to deteriorate across different industries, geographies, and time periods, requiring recalibration.

Along with technological developments, these limitations have motivated researchers to explore machine learning (ML) techniques for bankruptcy prediction. ML models, such as support vector machines (SVM), decision trees (DT), and gradient boosting algorithms can model complex, non-linear interactions between variables without requiring strict statistical assumptions. As a result, ML methods have gained increasing attention in recent years for their potential to achieve higher predictive accuracy compared to traditional statistical approaches (Sun et al., 2014; Qu et al., 2019).

2.3 Machine learning in financial distress prediction

According to Bellovary et al. (2007), the earliest ML applications in financial distress prediction utilized artificial neural networks (ANNs) and decision tree algorithms. While early machine learning models initially suffered from challenges such as limited computational power and concerns about interpretability, their ability to model more complex data patterns made them popular in financial distress prediction tasks (Bellovary et al., 2007).

The range of techniques used in financial distress prediction is quite large (Qu et al., 2019). Artificial neural networks (ANN) have been widely used for their ability to model complex, non-linear relationships among financial variables (Alaka et al., 2018). Support Vector Machines (SVMs) have also gained attention, particularly due to their performance in smaller high-dimensional datasets (Qu et al., 2019). Decision Trees (DTs) offer an interpretable approach to classification tasks, although they are often prone to overfitting (Alaka et al., 2018). Random Forests (RFs), which form the predictions from multiple decision trees, have been introduced to improve model robustness and reduce overfitting (Qu et al., 2019). Furthermore, boosting methods such as AdaBoost and Gradient Boosting build ensembles in stages, each one attempting to correct the errors of its predecessor, achieving improved predictive performance (Alaka et al., 2018). Overall,

machine learning models have been shown to outperform traditional statistical methods, such as discriminant analysis and logistic regression, in bankruptcy prediction tasks (Qu et al., 2019; Sun et al., 2014).

2.3.1 Extreme gradient boosting

Among the models mentioned previously, the use of extreme gradient boosting (XGBoost) has become popular in bankruptcy prediction and other financial applications due to its strong performance in classification tasks, scalability and flexibility (Chen and Guestrin, 2016). XGBoost is an implementation of gradient boosting decision trees, designed to handle large datasets (Chen and Guestrin, 2016).

In bankruptcy prediction tasks, XGBoost has often performed better compared to traditional statistical models such as logistic regression. Son et al. (2019) showed that XGBoost significantly outperformed logistic regression in predictive performance across different datasets. These performance improvements are particularly important in bankruptcy datasets, which are often skewed, contain missing values, complex patterns that might prove challenging for linear models (Son et al., 2019).

However, despite its strong predictive performance, XGBoost is often criticized for being a "black-box" model due to its complexity. To address this limitation, researchers have developed post-hoc interpretability tools and techniques. For example, Carmona et al. (2022) applied XGBoost to predict business failure and combined it with post-hoc explainability tools. The study showed that it is possible to explain how the XGBoost model made its predictions, and which financial variables were the most significant ones affecting bankruptcy risk within their dataset.

2.3.2 Hybrid models

While individual machine learning models have demonstrated strong performance in bankruptcy prediction, recent research suggests that combining different models can enhance predictive accuracy and robustness (Ainan et al., 2024; Chou, 2019). Hybrid models aim to use the unique strengths of different algorithms, compensating for their individual weaknesses by combining their outputs or by blending different modeling approaches.

Bankruptcy datasets often have high dimensionality, class imbalances, and complex non-linear relationships between variables, making it difficult for any single model to consistently perform well across different settings (Wu et al., 2019). Furthermore, different models offer distinct advantages. For example, logistic regression is transparent and widely understood, but it lacks flexibility, while models like XGBoost are often highly accurate but operate as black boxes (Chen and Guestrin, 2016; Carmona et al., 2022). Thus, hybrid models can offer a promising strategy to strike a balance between the strengths and weaknesses of individual models.

Several studies support the use of hybrid machine learning techniques for financial distress prediction. Ainan et al. (2024) proposed a hybrid model combining XGBoost and artificial neural networks (ANN) which, when combined outperformed each individual component, especially when dealing with imbalanced datasets. Meanwhile, Wu et al. (2019) developed an ensemble model that combined feature engineering with multiple classifiers to improve model robustness and performance.

A particularly interesting example is the work by Chou (2019), who developed a hybrid model integrating a decision tree (DT) and deep neural network (DNN), enhanced with the LIME algorithm for post-hoc interpretability. Their model used decision trees for global explainability and DNNs to improve accuracy. The final model improved predictive accuracy to over

90% across different configurations. However, the approach required multiple complex components and interpretability tools that may be difficult to implement and maintain in business contexts, particularly in SMEs.

Ensembles are models that average the predicted probabilities of multiple models. These models offer a simple yet powerful way to improve accuracy in classification tasks without requiring additional model training. Hybrid approaches such as these have shown strong potential in improving financial distress prediction while balancing predictive accuracy and practical usability (Wu et al., 2019; Ainan et al., 2024).

Overall, the design of hybrid models can vary widely. Some studies combine models sequentially, using one model for feature extraction and another for prediction, while others integrate outputs through voting or stacking frameworks. That being said, combining traditional statistical models with machine learning ones is less common in academic literature especially in bankruptcy prediction.

2.4 Challenges and barriers to machine learning adoption

The adoption of modern ML techniques in financial distress prediction, especially for SMEs, faces several practical challenges. These challenges include the technical limitations of the machine learning models themselves along with organizational, technical and regulatory issues (Yusuf et al., 2024; Zavodna et al., 2024).

As discussed earlier, many ML models, such as gradient boosting techniques, SVMs and NNs are often criticized for being difficult to interpret (Carmona et al., 2022). In bankruptcy prediction, this can prove to be a problem, since the stakeholders who use these models for decision-making can't only rely on accurate predictions, they also need to understand the reasoning behind the models. Without interpretability and transparency, even highly accurate

models can be viewed as untrustworthy, affecting their usability in certain decision-making contexts. As Carmona et al. (2022) argue, the lack of transparency in black-box models such can lower trust and hinder managerial adoption, especially when the predictions made by these models are used in decisions that carry consequences.

Recent developments in ML have introduced more interpretable "glass-box" models, such as explainable boosting machines (EBMs), generalized additive models (GAMs), and interpretable neural networks (Caruana et al., 2015; Lou et al., 2013). These techniques aim to preserve predictive accuracy while making model structure and outputs easier to understand. Even so, these models do not come without their own challenges. They often underperform compared to more complex algorithms in capturing subtle nonlinear relationships, especially in imbalanced or high-dimensional datasets (Carmona et al., 2022). Furthermore, they still require specialized knowledge to implement and operate. These are capabilities that many SMEs lack (Zavodna et al., 2024). As a result, theoretically transparent models may not be feasible for smaller firms with limited technical staff.

According to Yusuf et al. (2024), resource constraints, lack of skilled personnel, and underdeveloped digital infrastructure are among the leading barriers to AI and ML adoption in developing-country SMEs. Even in Europe, Zavodna et al. (2024) note that trust, skill gaps, and unclear return on investment remain major concerns when firms consider deploying complex analytics systems.

Despite these challenges, there are several strategies to make ML models more accessible and implementable in business settings. One promising approach is the use of post-hoc interpretability tools such as SHAP (Shapley Additive Explanations), which can explain the contribution of individual variables to a specific prediction, even in black-box models (Lundberg and Lee, 2017). These techniques can help bridge the gap between model complexity

and decision-maker trust. In addition, initiatives to improve digital literacy and analytical capacity within SMEs through training programs, partnerships with AI vendors, or access to government-supported platforms are gaining traction (Yusuf et al., 2024; Zavodna et al., 2024).

2.5 Summary and research gap

The research in financial distress prediction has evolved throughout the decades from simpler statistical models like Altman's Z-score and Ohlson's O-score to more advanced machine learning-based methods (Bellovary et al., 2007; Altman, 1968; Ohlson, 1980). These modern methods often outperform traditional ones in terms of raw accuracy, however they tend to sacrifice interpretability, which might make them problematic in situations where decision-making needs to be explainable and auditable.

Hybrid models and explainable machine learning have been proposed to address this issue (Carmona et al., 2022; Wu et al., 2019). However, most focus on maximizing performance using methods like stacking or deep neural integrations. These approaches often require significant technical infrastructure and expertise to implement. They often overlook the importance of built-in transparency and the need for models that can be interpreted by both technical and non-technical users alike. To date, there appears to be no published research that combines logistic regression and XGBoost in a soft-voting ensemble, specifically to enhance interpretability and practical usability in bankruptcy prediction.

This gap is particularly relevant especially for small and medium-sized enterprises (SMEs), which face additional barriers in adopting advanced analytics, including limited technical resources, regulatory pressures, and the need for stakeholder trust. Most existing studies that examine bankruptcy prediction in SMEs often focus on single industries, countries, or models with limited transparency.

This thesis addresses that gap by developing and empirically evaluating a lightweight hybrid model that combines logistic regression's interpretability with the predictive power of XGBoost. Unlike post-hoc explanation tools, interpretability is included directly in the model's structure. The model is tested on a large, multi-country dataset of SMEs and aims to provide a practical, scalable, and transparent solution for financial distress prediction.

3 Data and methodology

3.1 Data source and sampling strategy

The data used in this study was collected from the Orbis database, which provides firm-level financial, ownership and legal information (Bureau van Dijk, 2024). The dataset includes private limited companies classified as small or medium-sized (SME) that operate in 11 European countries: Finland, Sweden, Denmark, Germany, France, Italy, the Netherlands, Austria, Spain, Belgium and Poland. The dataset covered financial years from 2017 to 2023.

To ensure diversity, firms were filtered based on their NAICS (North American Industry Classification System) codes (U.S. Census Bureau, 2022). The selected industries include Construction (23), Manufacturing (31–33), Wholesale and Retail Trade (42, 44–45), Transportation and Warehousing (48–49), Information (51), and Accommodation and Food Services (72).

The raw dataset initially contained financial statements for multiple years per company, however only the most recent year (e.g., the year preceding the bankruptcy if applicable) was retained for each key variable. This choice was made to simplify model design, as time series analysis was beyond the scope of this study.

Bankruptcy labels were based on Orbis's "Status" filter. Specifically, firms were included if they were marked as "Bankrupt" or "Dissolved (bankruptcy)". Non-bankrupt firms were selected by filtering for companies marked as "Active". Companies with other ambiguous statuses were excluded to ensure clean classification.

To address the class imbalance problem, where bankrupt firms are significantly underrepresented in the dataset, a stratified sampling approach was used. All available bankrupt companies from each country were included in

the dataset. Then, 5,000 non-bankrupt companies were randomly selected from each country by using Orbis's random selection tool. This was done to keep the dataset size manageable and ensuring a roughly balanced representation across different countries and preventing overrepresentation from large economies.

After applying basic data cleaning steps and excluding firms with too many missing variables (e.g., total assets, operating revenue, equity), the final dataset consisted of 15,768 bankrupt and 24,112 non-bankrupt companies, totalling 39,880 observations.

3.2 Variable selection and feature engineering

The variable selection of this study was guided by both prior literature on bankruptcy prediction (e.g., Altman, 1968; Ohlson, 1980; Son et al., 2019) and data availability within the Orbis database. The objective was to include variables covering profitability, leverage, liquidity, firm size, and operational performance.

An initial set of financial variables was extracted, including total assets, operating revenue (turnover), net income, EBIT, equity, financial expenses (used as a proxy for interest expense), current assets, current liabilities, and number of employees.

Based on these, several financial ratios and transformations were made to improve comparability across firms and capture variables commonly related to financial health as shown in Table 1.

Variable	Formula
Log-transformed total assets	log(total_assets)
Equity ratio	equity / total_assets
Debt ratio	total_liabilities / total_assets
Return on assets (ROA)	net_income / total_assets
Return on equity (ROE)	net_income / equity
Profit margin	net_income / turnover
Current ratio	current_assets / current_liabilities
Interest coverage ratio proxy	EBIT / financial_expenses

Table 1. Financial ratios and transformations derived from raw variables

After correlation and multicollinearity checks, several variables were excluded from the final model. Debt ratio was removed due to a near-perfect negative correlation with equity ratio. Additionally, EBIT and financial_expenses were excluded due to high collinearity with net_income and redundancy with int_coverage.

The final dataset included the following 10 predictor variables alongside the target variable (bankrupt):

- turnover
- employees
- total_assets
- net_income
- equity_ratio
- int_coverage
- current_ratio
- log_assets
- roa
- ebit_margin

To validate the final feature set, both a correlation heatmap and Variance Inflation Factor (VIF) scores were calculated. All retained variables had VIF values well below the commonly accepted threshold of 5 (most below 2.5), indicating low amounts of multicollinearity between variables. These figures are shown in Appendix B.

3.3 Missing value treatment

The final dataset contained missing values, which is to be expected for bankruptcy-related data due to incomplete or inconsistent reporting. Since the logistic regression model does not natively handle missing data, imputation was needed to ensure that the model was compatible with the dataset.

Multiple imputation techniques were tested to examine whether the choice of imputation method would meaningfully affect model results. However, this did not appear to be the case. Initially, all the models were tested with median imputation. Then, a more advanced method, Multiple Imputation by Chained Equations (MICE) was tested (Azur et al., 2011). The results showed no meaningful differences in model performance, logistic regression coefficients, or XGBoost's SHAP values.

XGBoost was also tested without any imputation, since it can handle missing values on its own (Chen and Guestrin, 2016). The model's performance was nearly identical to the imputed version. To ensure a consistent comparison across all models, the final analysis used median imputation for all the models.

3.4 Model design and rationale

Three models were chosen to predict bankruptcy. Logistic regression and XGBoost were used as the baseline models and a hybrid model was constructed by combining these two using a soft-voting ensemble approach.

Logistic regression was chosen for its familiarity among both technical and non-technical stakeholders, its widespread usage in bankruptcy prediction and its interpretability (Ohlson, 1980; Son et al., 2019;). XGBoost was selected due to its strong predictive performance in classification tasks (Chen and Guestrin, 2016; Carmona et al., 2022; Son et al., 2019).

From a theoretical perspective, these models complement each other. Logistic regression is a high-bias, low-variance model that performs well in situations where the relationship between the predicting factors and the outcome is mostly linear. In contrast, XGBoost is a low-bias, high-variance model (Hastie et al., 2009). By combining the two in a soft-voting ensemble, the model attempts to balance variance and bias while preserving interpretability and keeping the overall structure simple. Running the models separately would have resulted in conflicting predictions with no clear way to integrate them. The soft-voting ensemble offers a unified output for the individual predictions while retaining the strengths of both methods.

An equal-weighted (50/50) configuration was tested initially for the hybrid model. The model's performance (as measured by ROC-AUC) slightly decreased as the weight assigned to logistic regression increased as shown in Figure 1. However, the performance decrease was smaller at lower weights. Based on this, a 20% weight for logistic regression and 80% for XGBoost was selected for the final model.

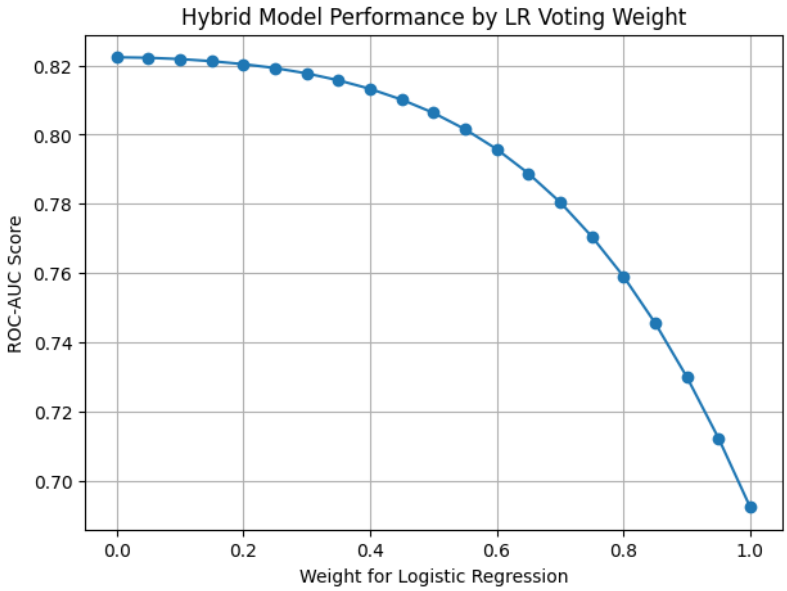


Figure 1 ROC-AUC performance of the hybrid model across different logistic regression weights

It was also hypothesized that including logistic regression could improve the calibration of predicted probabilities by lowering the overconfidence often associated with XGBoost. While this hypothesis was exploratory and not the primary design goal, its validity was examined in Section 4 using calibration curves.

3.5 Model evaluation

Model performance was evaluated using a set of standard classification metrics. The chosen metrics include precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). ROC-AUC was used as the primary metric for model comparison as it summarizes performance across all classification thresholds (Fawcett, 2006). All models were evaluated on the same 80/20 train-test split.

To assess how well each model's predicted probabilities matched actual risk, calibration curves were also generated. These plots show whether the probabilities predicted by the models match the true likelihood of bankruptcy. The interpretability and practical implementation of these models were assessed qualitatively. This analysis is discussed in detail in Chapter 5.

4 Results

4.1 Logistic regression results

The logistic regression model performed moderately in distinguishing between bankrupt companies and non-bankrupt ones. Its ROC-AUC score was 0.69. The ROC curve for the logistic regression model is shown below in Figure 2.

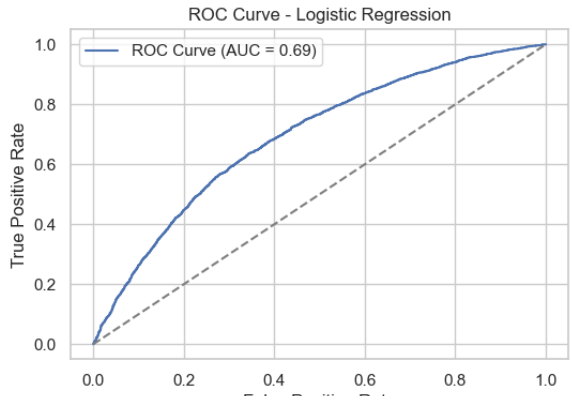


Figure 2 ROC Curve - Logistic Regression

The model's classification report (see Appendix C) shows an overall accuracy of 65%, a precision of 0.55 and recall of 0.61 for the bankrupt class. The confusion matrix shows that out of 3,154 actual bankrupt firms, the model correctly identified 1,938 but falsely predicted 1,573 non-bankrupt firms as bankrupt. The relatively higher recall indicates that the model was more successful at identifying bankrupt firms, while precision was slightly lower.

Based on the coefficients, the top positive predictors were equity ratio, net income, and return on assets (ROA), suggesting that stronger capital structure and profitability metrics reduce bankruptcy risk within the dataset. Alternatively, low current ratios and high log-transformed asset values showed negative associations. The full list of coefficients is presented in Appendix D.

While logistic regression offers the benefit of interpretability, its predicted probabilities showed signs of miscalibration, particularly underconfidence in the mid-range and slight overconfidence at the extremes. This is further

discussed in the model comparison section (see calibration curves in Section 4.4). Nonetheless, the model provides a transparent and replicable baseline for comparison against more advanced methods such as XGBoost.

4.2 XGBoost results

The XGBoost model achieved a ROC-AUC score of 0.822, outperforming logistic regression by a notable amount. The ROC curve is shown in Figure 3.

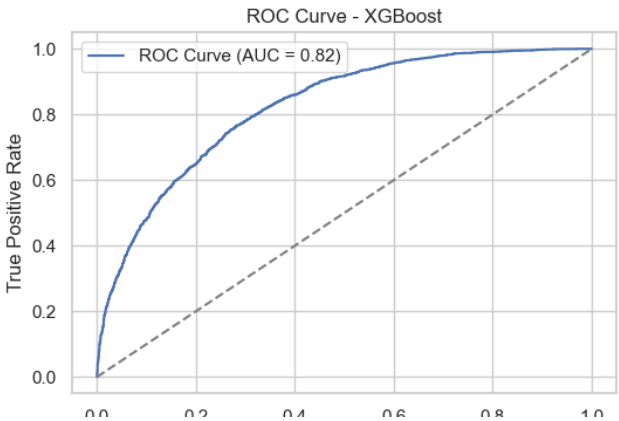


Figure 3 XGBoost ROC Curve

The model had an overall accuracy of 74%. For the positive class (bankrupt companies), it had a precision of 0.70 and a recall of 0.62. Compared to logistic regression, XGBoost improved precision substantially and slightly increased recall, resulting in an F1-score of 0.65 for the bankrupt class. The full classification report is shown in Appendix E.

The confusion matrix (see Appendix E) shows that XGBoost correctly identified 1,944 out of 3,154 bankrupt firms, while misclassifying 844 non-bankrupt firms as bankrupt.

4.3 Hybrid model results

The hybrid model achieved a ROC-AUC score of 0.820, which is extremely close to the standalone fine-tuned XGBoost model. Figure 4 below shows the ROC curve.

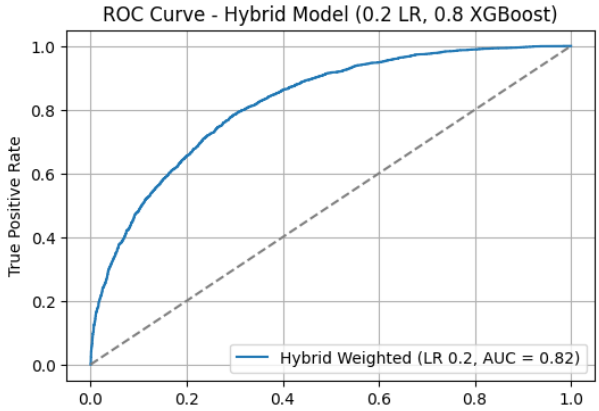


Figure 4. Hybrid model ROC curve

In terms of overall performance, the hybrid model had an accuracy of 74%, with precision of 0.69 and a recall of 0.62 for the positive class (bankrupt). As with the ROC-AUC score, these metrics are nearly identical to the XGBoost model. Full confusion matrix and classification report is shown in Appendix F.

4.4 Model comparison

4.4.1 Predictive performance

Metric	Logistic Regression	XGBoost (Tuned)	Hybrid (20% LR ; 80% XGB)
Accuracy	0,65	0,74	0,74
Precision (Bankrupt)	0,55	0,7	0,69
Recall (Bankrupt)	0,61	0,62	0,62
F1-Score (Bankrupt)	0,58	0,65	0,65
ROC-AUC	0,69	0,822	0,82

Table 2. Summary of model performance metrics

Table 2 presents a side-by-side comparison of the three models' performance scores. XGBoost demonstrated the highest scores across all metrics, validating prior research on the effectiveness of boosting methods in handling skewed and high-dimensional datasets (Chen & Guestrin, 2016; Carmona et al., 2022). The model's strength stems from how it learns from past misclassifications and detects patterns involving multiple financial variables at once (Chen & Guestrin, 2016).

The hybrid model achieved nearly identical performance compared to XGBoost, with a ROC-AUC of 0.820 and similar classification metrics. This suggests that introducing a minor logistic regression component primarily for interpretability did not affect the predictive accuracy of the model in a meaningful way. While earlier testing confirmed that increasing the logistic regression weight reduced the ensemble's ROC-AUC score, the chosen 80/20 split preserved XGBoost's strength while offering a built-in layer of interpretability within the model.

In contrast, the logistic regression model showed weaker performance across all metrics, with a ROC-AUC of 0.69, and lower precision and recall for the bankrupt class. Although the logistic regression model has its strengths as it is transparent and easy to implement, its reliance on linear assumptions can limit its ability to capture complex relationships between variables (Hastie et al., 2009). These limitations are consistent with previous studies emphasizing the difficulty of modeling the non-linear nature often present in bankruptcy datasets (Son et al., 2019).

4.4.2 Calibration

To assess the reliability of predicted probabilities, calibration curves were generated for all three models. These plots compare predicted probabilities to actual default frequencies, making it possible to evaluate how well each model quantifies risk.

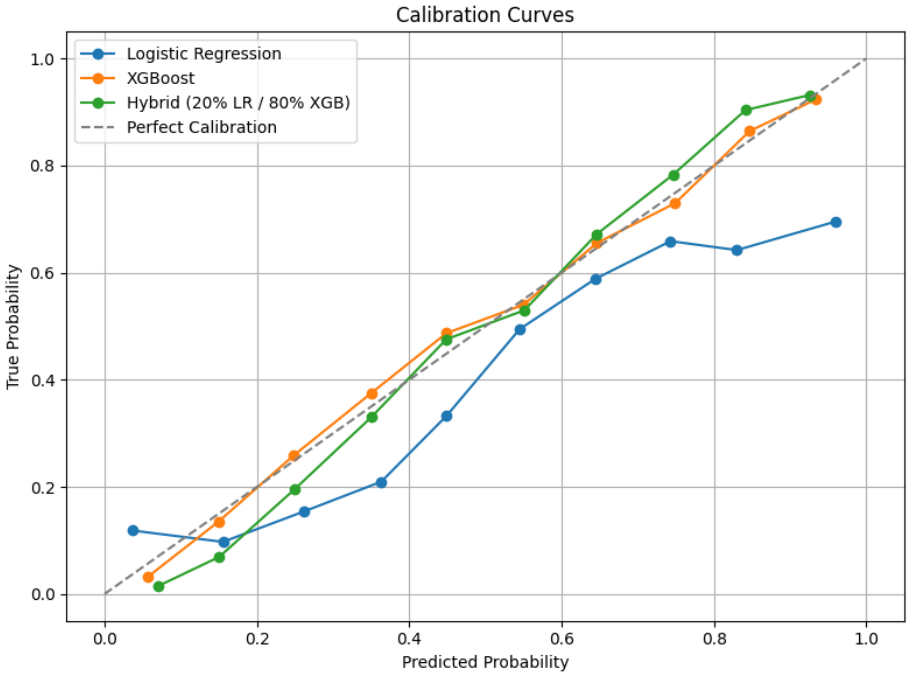


Figure 5 Calibration curves

As shown in Figure 3, the XGBoost model demonstrated the best overall calibration. It closely followed the ideal calibration line, particularly in the mid-to-high probability range. Meanwhile, logistic regression significantly underestimated bankruptcy risk in higher probability ranges and consistently deviated from the ideal line.

The hybrid model's calibration performance is between the two base models. In the mid-range (approximately 0.4–0.7), it outperformed logistic regression and closely followed XGBoost. However, in the lower and upper probability ranges, the hybrid curve deviated more than XGBoost's, suggesting that the addition of logistic regression slightly worsened calibration in those ranges.

While one goal of the hybrid model was to lower XGBoost's potential overconfidence, the results here suggest that this effect was minimal or inconsistent. Nonetheless, the hybrid retained good overall calibration and remained competitive with XGBoost.

5 Discussion

5.1 Interpretability

In machine learning, interpretability means a human's ability to understand how a model works and how the input variables affect its predictions. This is especially important in high-stakes domains such as credit risk, where decisions must often be explained to regulators, auditors, or executives without technical backgrounds (Doshi-Velez & Kim, 2017; Rudin, 2019).

There are generally two types of interpretability. Global and local. Global interpretability refers to understanding how a model works across all inputs and cases. For example, logistic regression is globally interpretable, as each coefficient directly tells us the effect of a predicting variable on the log-odds of the outcome. Local interpretability, on the other hand, explains individual predictions. In this case why a firm was classified as having high risk of bankruptcy. Local methods are commonly used with complex models like XGBoost, often via post hoc tools such as SHAP.

XGBoost's accuracy can be partly attributed to how it works by combining decision trees. However, this makes the model less transparent and hard to explain. Explaining the logic behind a prediction or understanding how a variable affects risk more generally requires additional tools and technical knowledge. Post hoc methods such as SHAP aim to address this by assigning importance scores to each variable per prediction. However, these tools add complexity and can be computationally intensive or difficult for non-technical stakeholders to interpret accurately (Rudin, 2019; Zeng et al., 2024).

The hybrid ensemble model used in this study is designed in a way which retains a level of built-in interpretability through its logistic regression component. Although the ensemble's final predictions are weighted combinations of the individual probability outputs from both models, the logistic

regression model can still be analysed independently to understand the model on a global level. Specifically, stakeholders can inspect the logistic regression coefficients to understand which financial variables are associated with increased or decreased bankruptcy risk, even if the final prediction is affected by the XGBoost component.

This hybrid approach offers a compromise. Instead of relying solely on black-box methods post-hoc estimations, the model embeds interpretability directly into its architecture. Furthermore, SHAP values can still be used alongside the hybrid model to provide local, company specific explanations when needed, complementing the global insights offered by the logistic regression coefficients. Figure 6 and Figure 7 illustrate examples of interpretability outputs from the hybrid model using SHAP for local explanation and logistic regression coefficients for global interpretability.

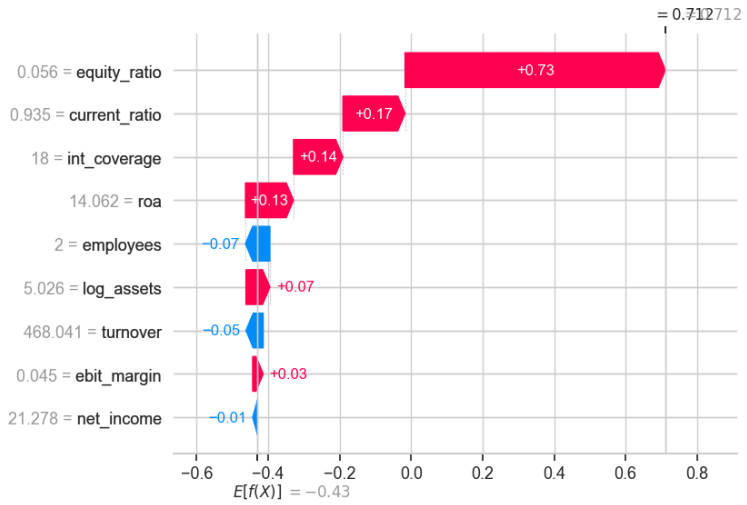


Figure 6 SHAP-based local explanation for single firm

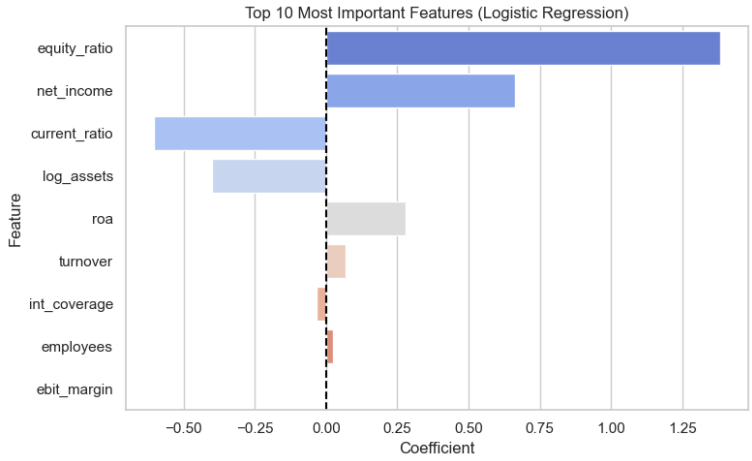


Figure 7. Logistic regression coefficients

While SHAP explanations can be valuable, they have limitations. They are not part of the actual model's decision-making, instead they are estimated after the fact. Additionally, SHAP values can be misinterpreted, especially when the features interact in non-obvious ways. Using SHAP also adds computational burden and may require specialized infrastructure depending on the model used (Zeng et al., 2024).

The main practical benefit of using logistic regression coefficients stems from the fact that they are direct, stable and easy to communicate. For example, a negative equity ratio coefficient immediately tells the decisionmaker that higher equity levels reduce bankruptcy risk. That being said, different situations might warrant different types of interpretability. While LR coefficients are suited for explaining general patterns, SHAP values can offer valuable case-specific information.

The hybrid model does not eliminate the need for post hoc explanation tools but reduces reliance on them. It integrates a transparent and well understood component into the ensemble architecture, offering a layer of accountability that is often missing in pure machine learning models.

While post hoc explanation tools such as SHAP are widely used for interpreting black-box models, several studies have highlighted important limitations. SHAP values rely on assumptions leveraging game theory that may not align with how humans understand causality or feature importance (Kumar et al., 2020). These explanations can assign influence to features that are only

proxies for other variables, especially in the presence of feature correlation, potentially leading to misleading interpretations (Rudin, 2019; Kumar et al., 2020). Furthermore, SHAP assumes feature independence and local linearity. These assumptions often do not hold up in datasets (Doumard et al., 2022).

Although SHAP is praised for its intuitive visualizations, these explanations are approximating the actual behaviour of the model (Wang et al., 2023). As such, reliance on SHAP can introduce an illusion of transparency while still requiring significant expertise to interpret correctly.

5.2 Practical applications

The hybrid model developed in this study is designed with practical implementation in mind, particularly for use in SME bankruptcy risk assessment. By combining a strong machine learning model with a transparent logistic regression component, the model strikes a balance between accuracy and interpretability. This structure is especially valuable in business contexts, where predictions must often be explained to non-technical decision-makers, auditors, or regulators (Doshi-Velez & Kim, 2017; Rudin, 2019).

For example, the model can be used in early warning systems by highlighting key financial risk indicators through regression coefficients, while also allowing deeper case by case analysis using SHAP values when needed. The model could be applied in tasks such as loan underwriting, risk scoring, and performance monitoring. Additionally, since both XGBoost and logistic regression are relatively lightweight, the model can be realistically deployed without heavy infrastructure or advanced machine learning expertise.

5.3 Limitations of the study

While the hybrid model presented in this study offers a balance between predictive accuracy and interpretability, there are several limitations.

The dataset used in the study was limited to the most recent financial year available for each firm, meaning that the model does not account for financial trends. Including time-series based variables could potentially improve prediction accuracy and overall performance. The dataset was restricted to SMEs in eleven European countries. The model's generalizability to other regions or larger firms also remains untested. Furthermore, the study could have tested the model's performance across individual industries or countries.

Additionally, only financial variables were included in the final model. The exclusion of qualitative or macroeconomic indicators, such as management quality, news sentiment, or regional business conditions, might limit the model's ability to capture broader drivers of financial distress.

While the hybrid model improves interpretability, it still introduces moderate complexity compared to a single-model approach, which could be a barrier for firms with limited technical expertise or infrastructure.

5.4 Suggestions for future research

There are several opportunities to extend the findings of this study. Future research could incorporate time-series financial data to capture changes in firm performance over time, potentially improving prediction accuracy. Also, adding macroeconomic indicators or qualitative variables such as management quality, board structure, or sentiment analysis could further enhance the prediction accuracy and provide meaningful insights on how these variables affect bankruptcy risk.

The current model uses a fixed soft-voting strategy. Future work could explore alternative ensemble techniques such as stacking or dynamic weighting to further improve calibration or performance. Additionally, researchers could compare this hybrid approach with other interpretable machine learning models, such as Explainable Boosting Machines or Generalized Additive Models.

Testing the model across different firm sizes, industries, or geographical regions would help examine its generalizability and practical value beyond the European SME context.

References

Ainan, U.H., Por, L.Y., Chen, Y.-L., Yang, J. and Ku, C.S., 2024. Advancing bankruptcy forecasting with hybrid machine learning techniques: Insights from an unbalanced Polish dataset. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3354173

Alaka, H.A., Oyedele, L.O., Owolabi, H.A., et al., 2018. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications, 94, pp.164–184. https://doi.org/10.1016/j.eswa.2017.10.040

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), pp.589–609. Available at: https://www.jstor.org/stable/2978933

Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E.K. and Suvas, A., 2017. Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. Journal of International Financial Management & Accounting, 28(2), pp.131–171. https://doi.org/10.1111/jifm.12053

Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J. (2011), Multiple imputation by chained equations: what is it and how does it work?. Int. J. Methods Psychiatr. Res., 20: 40-49. https://doi.org/10.1002/mpr.329

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

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N., 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1721–1730. https://doi.org/10.1145/2783258.2788613

Carmona, P., Dwekat, A. and Mardawi, Z., 2022. No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. Research in International Business and Finance, 61, 101649. https://doi.org/10.1016/j.ribaf.2022.101649

Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794. https://doi.org/10.1145/2939672.2939785

Chou, T.-N., 2019. *An explainable hybrid model for bankruptcy prediction based on the decision tree and deep neural network*. In: 2nd IEEE International Conference on Knowledge Innovation and Invention (ICKII 2019), 11–13 July 2019, Seoul, South Korea. IEEE, pp. 122–125. doi:10.1109/ICKII46350.2019.8941195

Doumard, E., Aligon, J., Escriva, E., Excoffier, J.-B., Monsarrat, P. and Soulé-Dupuy, C., 2022. A comparative study of additive local explanation methods based

on feature influences. arXiv preprint, arXiv:2202.06502. https://arxiv.org/abs/2202.06502

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27(8), pp.861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.

Iparraguirre-Villanueva, V. and Cabanillas-Carbonell, M., 2024. Predicting business bankruptcy: A comparative analysis with machine learning models. Procedia Computer Science, 232, pp.2231–2240. https://doi.org/10.1016/j.procs.2023.12.382

Kumar, I.E., Venkatasubramanian, S., Scheidegger, C. and Friedler, S.A., 2020. Problems with Shapley-value-based explanations as feature importance measures. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020), PMLR 119. Available at: https://arxiv.org/abs/2002.11097

Lou, Y., Caruana, R. and Gehrke, J., 2013. Intelligible models for classification and regression. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.150–158. https://doi.org/10.1145/2339530.2339556

Lundberg, S.M. and Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA. Available at: https://arxiv.org/abs/1705.07874

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

Park, M.S., Son, H., Hyun, C. and Hwang, H.J., 2021. Explainability of machine learning models for bankruptcy prediction. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3110270

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), pp.206–215. Available at: https://arxiv.org/abs/1811.10154

Son, H., Hyun, C., Phan, D. and Hwang, H.J., 2019. Data analytic approach for bankruptcy prediction. Expert Systems with Applications, 138, 112816. https://doi.org/10.1016/j.eswa.2019.07.033

Sun, J., Li, H. and Huang, Q.H., 2014. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Knowledge-Based Systems, 57, pp.41–56. https://doi.org/10.1016/j.knosys.2013.12.006

Wang, T., Liu, M., Wang, Y., Peng, B., Chen, S., Song, X. and Zhang, C., et al., 2023. Explainable AI across domains: Techniques, domain-specific applications, and future directions. IEEE Transactions on Neural Networks and Learning Systems, Early Access. https://doi.org/10.1109/TNNLS.2023.3273904

Wu, X., Yang, D., Zhang, W. and Zhang, S., 2019. A hybrid ensemble model for corporate bankruptcy prediction based on feature engineering method. International Journal of Information and Communication Sciences, 4(3), pp.63–69. https://doi.org/10.11648/j.ijics.20190403.12

Yusuf, M.T., Jibir, A. and Musa, H., 2024. AI and digital transformation for SMEs: Empirical evidence from developing economies. Technology in Society, 76, 102481. https://doi.org/10.1016/j.techsoc.2023.102481

Zavodna, Z., Hodinkova, M. and Urbanek, S., 2024. Barriers to AI implementation in SMEs: Evidence from Europe. Technological Forecasting and Social Change, 198, 122877. https://doi.org/10.1016/j.techfore.2023.122877

Zeng, X., 2024. Enhancing SHAP values interpretability using large language models. arXiv preprint, arXiv:2404.04281. https://arxiv.org/abs/2404.04281

Appendix

Appendix A. Z-score and logistic regression formulas

Altman Z-score (1968)

$$Z = 1.2\,X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1X_5$$

Where X_1 is Working capital / Total assets, X_2 is Retained earnings / total assets, X_3 is EBIT / Total assets, X_4 is Market value of equity / Book value of total liabilities and X_5 is Sales / Total assets.

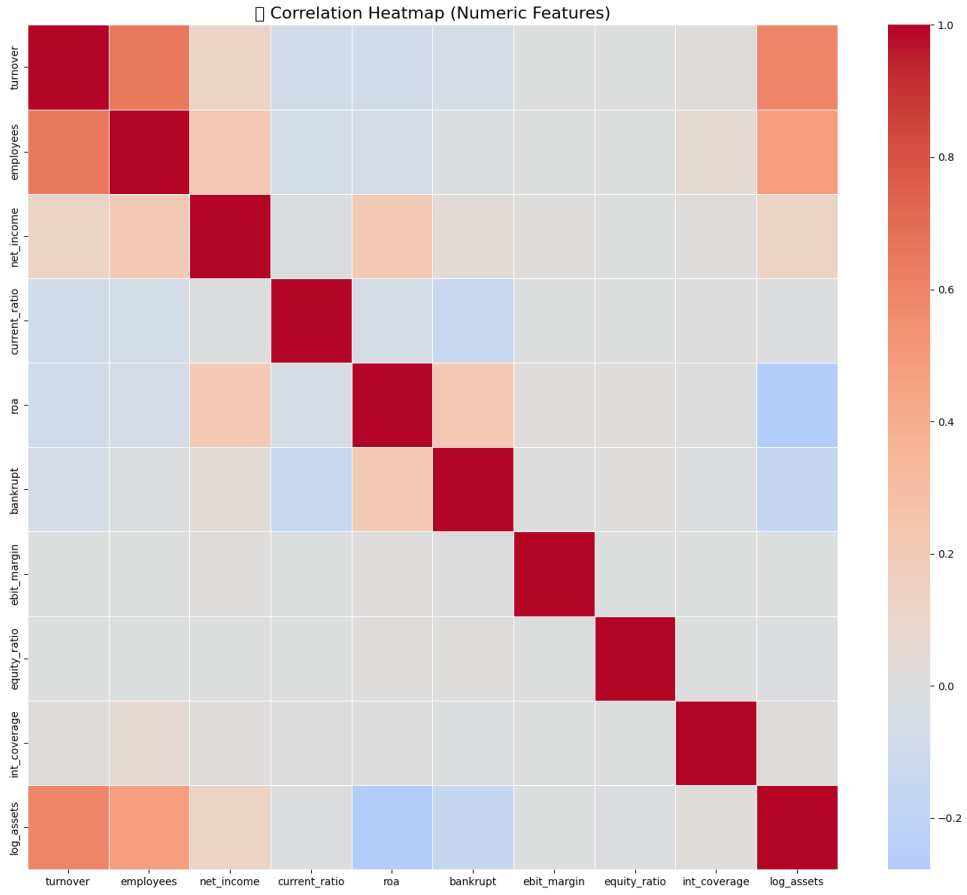
Ohlson's logistic regression model (O-score)

$$\text{logit(P)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Where P is the probability of bankruptcy, $X_1, X_2, \ldots, X_n$ are the financial ratios and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients estimated from training data.

Appendix B. Correlation heatmap and VIF scores

Feature	VIF
const	38.509936
turnover	2.442422
log_assets	2.304641
employees	1.758686
net income	1.428276
roa	1.391931
bankrupt	1.152182
current ratio	1.838377
equity ratio	1.020344
ebit_margin	1.006793
int_coverage	1.001903



Appendix C. Logistic regression confusion matrix and classification report

Confusion Matrix	Predicted: 0	Predicted: 1
Actual: 0 (non-bankrupt)	3,249	1,573
Actual: 1 (bankrupt)	1,216	1,938

Classification report	Precision	Recall	F1-score	Support
0 (non-bankrupt)	0.73	0.67	0.70	4,822
1 (bankrupt)	0.55	0.61	0.58	3,154
Accuracy			0.65	7,976
Macro avg	0.64	0.64	0.64	7,976
Weighted avg	0.66	0.65	0.65	7,976

Appendix D. Full list of logistic regression coefficients

Feature	Coeffi-cient
equity_ratio	1.382464
net income	0.661553
current ratio	-0.603321
log_assets	-0.401975
roa	0.280174
tumover	0.067263
int_coverage	-0.034637
employees	0.025433
ebit_margin	0.001042

Appendix E. XGBoost confusion matrix and classification report

Confusion Matrix	Predicted: 0	Predicted: 1
Actual: 0 (non-bankrupt)	3,978	844
Actual: 1 (bankrupt)	1,210	1,944

Classification report	Precision	Recall	F1-score	Support
0 (non-bankrupt)	0.77	0.82	0.79	4,822
1 (bankrupt)	0.70	0.62	0.65	3,154
Accuracy			0.74	7,976
Macro avg	0.73	0.72	0.72	7,976
Weighted avg	0.74	0.74	0.74	7,976

Appendix F. Hybrid model confusion matrix and classification report

Confusion Matrix	Predicted: 0	Predicted: 1
Actual: 0 (non-bankrupt)	3,954	868
Actual: 1 (bankrupt)	1,196	1,958

Classification report	Precision	Recall	F1-score	Support
0 (non-bankrupt)	0.77	0.82	0.79	4,822
1 (bankrupt)	0.69	0.62	0.65	3,154
Accuracy			0.74	7,976
Macro avg	0.73	0.72	0.72	7,976
Weighted avg	0.74	0.74	0.74	7,976