UNIVERSITÀ DI PAVIA

Department of Economics and Management

Master Programme in Finance

AI Methods for Bankruptcy Prediction in Private Companies

Supervisor:

Prof. Dr. Paolo Giudici

Co-supervisor:

Dr. Tim Whittaker

Student:

Nastaran Bayat Patapeh

Academic Year 2025–2026

Acknowledgments

I am deeply grateful to my supervisor, **Prof. Dr. Paolo Giudici**, *Full Professor of Statistics*, for his guidance, encouragement, and insightful feedback throughout this research. His expertise and support were instrumental to the development and completion of this thesis.

I also wish to express my sincere appreciation to **Dr. Tim Whittaker**, *Head of Data at the EDHEC Infrastructure & Private Assets Research Centre (EDHECinfra)*, for his valuable assistance and constructive input on both the technical and analytical aspects of this study.

Access to firm-level data was provided through the **ORBIS database**, developed by **Bureau van Dijk (a Moody's Analytics company)**, under an institutional license granted by the **EDHEC Infrastructure & Private Assets Research Centre (EDHECinfra)**. I gratefully acknowledge this support, which made this research possible.

Finally, I extend my heartfelt thanks to my family and friends for their unwavering support and encouragement throughout this academic journey.

Contents

Abstract (English) **vi**

Abstract (Italiano) **viii**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Problem Statement 4

1.3 Research Objectives 6

1.4 Scope and Limitations 7

1.5 Structure of the Thesis 8

2 Literature Review **10**

2.1 Overview of Bankruptcy Prediction 10

2.2 Traditional Financial Models 13

2.3 Machine Learning in Financial Risk Analysis 15

2.4 Challenges in Private Firm Data 20

2.5 AI Models in Recent Research 23

2.6 Summary of Research Gap 26

3 Methodology **29**

3.1 Data Source and Description 29
3.2 Feature Engineering and Ratio Construction 30
3.3 Bankruptcy Label Definition 32
3.4 Data Cleaning and Preprocessing 32
3.5 Handling Missing Values and Class Imbalance 33
3.6 Model Selection Strategy 35
3.6.1 Correlation Analysis 38

4 Implementation **40**

4.1 Tools and Software Environment 40
4.2 End-to-End Pipeline 41
4.2.1 Data Preparation 41
4.2.2 Train/Test Split and Class Balance 42
4.2.3 Evaluation Metrics 42
4.2.4 Robustness Protocol 43
4.3 Model Configurations and Training 43
4.3.1 Logistic Regression 43
4.3.2 Random Forest 44
4.3.3 XGBoost 44
4.3.4 LightGBM 44
4.3.5 Support Vector Machine (RBF) 44
4.3.6 Feed-forward Neural Network 45
4.3.7 Soft-Voting Ensemble 45
4.4 Feature Explainability via RGE 45
4.5 Outputs and Artefacts 46

4.6 Implementation Notes and Practicalities 46

4.7 Summary 47

5 Results and Evaluation **48**

5.1 Evaluation Metrics 48

5.2 Descriptive Statistics and Correlation Analysis 49

5.3 ROC Curve and AUC Comparison 50

5.4 Confusion Matrices 55

5.5 Feature Importance Analysis 55

5.6 Rank Graduation Metrics (RGA, RGR, RGE) 57

5.7 Calibration and Reliability Analysis 64

5.8 Model Comparison and Ensemble Performance 66

5.9 Summary of Results 68

6 Discussion **69**

6.1 Interpretation of Results 69

6.2 Comparison with Prior Studies 70

6.3 Implications for Practice 71

6.4 Limitations 71

7 Conclusion and Future Work **73**

7.1 Summary of Findings 73

7.2 Research Contributions 74

7.3 Future Research Directions 74

7.4 Final Remarks 75

References **76**

A Appendix A: Complete R Code Pipeline **79**

B Appendix B: Variable Descriptions **89**

C Appendix C: Additional Model Outputs **90**

D Appendix D: R Session Information **97**

Abstract (English)

This study investigates the application of advanced machine learning techniques to predict corporate bankruptcy among unlisted firms using financial data from 2014–2024 obtained from the ORBIS database. Unlisted companies constitute a critical yet under-researched segment of the economy due to limited disclosure requirements and the absence of standardized reporting frameworks. Early detection of financial distress in such firms is essential for credit institutions, investors, and policymakers seeking to mitigate systemic risk and strengthen financial stability.

A curated dataset of private firms was developed through systematic data cleaning, transformation, and feature engineering of key financial ratios—including liquidity, cash ratio, debt ratio, working capital, EBITDA, and intangible ratio. Bankruptcy was defined via a balance-sheet criterion, classifying firms as bankrupt when total liabilities exceeded total assets. The dataset was divided into training and testing subsets, and six supervised learning models were implemented: logistic regression, support vector machine (SVM), random forest, XGBoost, LightGBM, and a feedforward neural network.

Model performance was primarily evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC). The random forest achieved the highest individual AUC (≈ 0.952), while an ensemble of random forest, XGBoost, and LightGBM reached 0.955, demonstrating superior predictive accuracy. Beyond traditional AUC analysis, Rank Graduation metrics—Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation

Explainability (RGE) were applied to assess model stability and interpretability. RGA validated the consistency of predictive ranking across models; RGR measured robustness to data perturbations, emphasizing the stability of ensemble and tree-based methods; and RGE quantified the explanatory contribution of financial ratios, revealing liquidity and leverage as the most influential features.

Overall, the findings demonstrate that machine learning can provide a reliable, interpretable, and reproducible framework for assessing bankruptcy risk in private markets using only financial statement data. The integration of RGA, RGR, and RGE enhances both analytical depth and trustworthiness, offering a novel and comprehensive approach to predictive financial risk modeling.

Abstract (Italiano)

Questo studio analizza l'applicazione di tecniche avanzate di machine learning per la previsione del fallimento aziendale tra imprese non quotate, utilizzando dati finanziari relativi al periodo 2014–2024 ottenuti dal database ORBIS. Le imprese non quotate rappresentano un segmento cruciale ma ancora poco esplorato dell'economia, a causa di obblighi di trasparenza più limitati e dell'assenza di quadri di rendicontazione standardizzati. L'individuazione precoce delle situazioni di difficoltà finanziaria in tali imprese è essenziale per istituzioni creditizie, investitori e decisori politici che mirano a mitigare il rischio sistemico e a rafforzare la stabilità finanziaria.

È stato costruito un dataset selezionato di imprese private attraverso un processo sistematico di pulizia dei dati, trasformazione e feature engineering dei principali indicatori finanziari — tra cui liquidità, cash ratio, debt ratio, capitale circolante, EBITDA e intangible ratio. Il fallimento è stato definito secondo un criterio di bilancio, classificando un'impresa come fallita quando il totale delle passività superava il totale delle attività. Il dataset è stato suddiviso in sottoinsiemi di training e testing e sono stati implementati sei modelli di apprendimento supervisionato: regressione logistica, support vector machine (SVM), random forest, XGBoost, LightGBM e una rete neurale feedforward.

Le performance dei modelli sono state valutate principalmente attraverso l'Area Under the Receiver Operating Characteristic Curve (AUC). Il modello random forest ha ottenuto il valore AUC individuale più elevato ($\approx$ 0,952), mentre un ensemble composto da random forest,

XGBoost e LightGBM ha raggiunto 0,955, dimostrando una superiore accuratezza predittiva. Oltre alla tradizionale analisi dell'AUC, sono state applicate metriche di Rank Graduation — Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR) e Rank Graduation Explainability (RGE) — al fine di valutare la stabilità e l'interpretabilità dei modelli. La RGA ha validato la coerenza dell'ordinamento predittivo tra i modelli; la RGR ha misurato la robustezza rispetto a perturbazioni dei dati, evidenziando la stabilità dei metodi ensemble e basati su alberi; la RGE ha quantificato il contributo esplicativo degli indicatori finanziari, rivelando liquidità e leva finanziaria come le variabili più influenti.

Nel complesso, i risultati dimostrano che il machine learning può offrire un quadro affidabile, interpretabile e replicabile per la valutazione del rischio di fallimento nei mercati privati utilizzando esclusivamente dati di bilancio. L'integrazione delle metriche RGA, RGR e RGE accresce sia la profondità analitica sia l'affidabilità del modello, proponendo un approccio innovativo e completo alla modellizzazione predittiva del rischio finanziario.

Keywords: Bankruptcy prediction, Machine learning, Financial ratios, Random Forest, XGBoost, LightGBM, Neural networks, Model robustness, Explainability, RGA, RGR, Unlisted firms, Financial distress.

Chapter 1

Introduction

1.1 Background and Motivation

Corporate bankruptcy represents a major economic and social concern, affecting a wide range of stakeholders including creditors, employees, shareholders, and governments. The consequences of corporate failure extend beyond the firm itself, disrupting labor markets, supply chains, financial institutions, and regional economies. Early detection of financial distress is therefore critical, as timely intervention can mitigate losses, prevent cascading defaults, and maintain stability within credit markets. Traditional bankruptcy prediction models have been widely studied for decades, with seminal works such as Altman's Z-score (Altman, 1968) and Ohlson's O-score (Ohlson, 1980) forming the foundation for quantitative assessments of insolvency risk. These models rely on accounting ratios and linear discriminant analysis, but most research has focused on publicly listed companies due to the ready availability of audited financial statements and regulatory reporting requirements.

In contrast, private (unlisted) firms constitute the vast majority of businesses worldwide. According to recent OECD reports, small- and medium-sized enterprises (SMEs)—most of

which are privately held—account for over 90% of registered businesses and more than 60% of employment in developed economies (OECD, 2023). Despite their macroeconomic significance, these firms are underrepresented in both academic research and credit risk assessment frameworks. This gap arises primarily from the absence of standardized, publicly available data. Private companies are generally not subject to mandatory disclosure regulations, and their reporting practices vary across jurisdictions, firm size, and ownership structure. Consequently, their financial data tend to be heterogeneous, incomplete, and difficult to compare. These characteristics pose a significant challenge for financial institutions, investors, and policymakers seeking to evaluate the bankruptcy risk of private firms.

Recent advancements in artificial intelligence (AI) and machine learning (ML) offer powerful tools to address these challenges. Unlike classical statistical methods, ML algorithms can capture complex non-linear relationships, tolerate noisy or missing data, and adapt to imbalanced class distributions—characteristics that are prevalent in private firm datasets (Baesens et al., 2003; Lessmann et al., 2015). Ensemble-based approaches such as Random Forest, XGBoost, and LightGBM have emerged as state-of-the-art methods due to their ability to combine multiple weak learners, reduce overfitting, and achieve superior predictive accuracy (Friedman, 2001; Chen and Guestrin, 2016; Ke et al., 2017). These models have shown strong performance in applications such as credit scoring, fraud detection, and loan default prediction (Zhou et al., 2021; Louzis et al., 2022; Huang and Zhao, 2023; Zhang et al., 2023), confirming their robustness in complex financial datasets.

Building on these successes, modern bankruptcy prediction increasingly emphasizes the use of engineered and dynamic financial indicators that extend beyond static accounting ratios. While traditional measures such as liquidity, cash ratio, debt ratio, working capital, EBITDA, and intangible asset ratios remain informative, their predictive power can be enhanced when

combined with growth-based or cash-flow-derived metrics. Prior studies have demonstrated that incorporating firm-level trends and operational momentum improves prediction accuracy, particularly for smaller or rapidly expanding firms (Yeh et al., 2011; Sun et al., 2014). This evolution reflects a broader shift from static classification to adaptive, data-driven modeling that better captures the financial dynamics underlying corporate distress.

While prior research has largely focused on predictive accuracy, this study expands the evaluation framework by integrating advanced, ranking-based performance metrics. Rank Graduation Accuracy (RGA) quantifies how well predicted risk rankings align with actual financial distress ordering, providing a finer measure of discriminative capability than simple classification accuracy. Rank Graduation Robustness (RGR) assesses the stability of these rankings under small data perturbations, a key property when working with incomplete or heterogeneous financial information. Rank Graduation Explainability (RGE) measures how individual features contribute to ranking outcomes, offering interpretable insights into which financial ratios drive the models' predictions. Together, RGA, RGR, and RGE complement classical metrics such as AUC, precision, recall, and F1-score by emphasizing not only predictive correctness but also stability and transparency.

Integrating ensemble learning with these ranking-based metrics creates a methodological framework that is both scalable and interpretable. While traditional models are constrained by linear assumptions and sensitivity to irregular data, ML algorithms can uncover multi-dimensional relationships across liquidity, leverage, profitability, and operational indicators. Ensemble methods—particularly Random Forest and gradient boosting—can model complex interactions, while neural networks capture deeper, non-linear dependencies. Complementing these with feature importance analysis and RGE ensures that results remain transparent and actionable for practitioners in finance and policy.

This thesis addresses a critical gap in the literature: the reliable prediction of bankruptcy for private firms using machine learning techniques applied to real-world financial data. A curated dataset of unlisted companies from 2014 to 2024 is developed, incorporating key financial ratios engineered to represent liquidity, leverage, profitability, and asset structure. Multiple ensemble and neural network models are evaluated through traditional performance metrics and the newly integrated ranking-based measures (RGA, RGR, and RGE). This comprehensive approach enhances predictive performance while maintaining interpretability and robustness.

Ultimately, the study contributes both theoretically and practically to the field of bankruptcy prediction. It demonstrates that modern ML techniques, when combined with ranking-based evaluation frameworks, can significantly improve the reliability, scalability, and transparency of bankruptcy risk assessment. By bridging the gap between classical statistical approaches and contemporary AI methodologies, the research provides a reproducible and interpretable foundation for assessing financial distress in private companies—an area of growing importance in credit risk management and financial stability analysis.

1.2 Problem Statement

Traditional bankruptcy prediction models, such as Altman's Z-score (Altman, 1968), Ohlson's O-score (Ohlson, 1980), and other regression-based or linear discriminant approaches, were originally developed for publicly listed firms. These models assume full access to consistent, audited financial statements and rely heavily on linearity and normality assumptions. However, such assumptions rarely hold for private firms, where financial reporting is irregular, incomplete, and highly heterogeneous. As a result, applying these classical models to unlisted companies often leads to unstable and biased results.

Beyond data quality issues, the bankruptcy prediction task itself poses two structural chal-

lenges. First, **class imbalance**: bankruptcy is a relatively rare event, especially among mature or well-capitalized firms. This imbalance causes traditional classifiers to favor the majority (non-bankrupt) class, yielding deceptively high accuracy but poor sensitivity to actual distress cases. Second, **feature heterogeneity**: private firms vary widely in accounting standards, size, and asset composition, which complicates the identification of universal predictive patterns.

To address these challenges, recent research has increasingly adopted machine learning (ML) approaches capable of learning complex non-linear relationships, handling missing values, and mitigating class imbalance through ensemble learning and reweighting strategies. Yet, despite these advancements, most prior studies focus predominantly on public companies or simulated datasets, leaving a significant gap in understanding how ML methods perform on large-scale, real-world data from private firms.

Moreover, the evaluation of bankruptcy prediction models has traditionally emphasized classification accuracy or AUC, overlooking critical aspects such as ranking consistency, robustness, and interpretability. These dimensions are essential for practical decision-making in credit risk management, where stakeholders prioritize stable, explainable, and reproducible results.

Accordingly, this study addresses the following research problem:

> *How can advanced machine learning methods be applied to improve the accuracy, robustness, and interpretability of bankruptcy prediction for private (unlisted) firms using real-world financial data?*

To this end, the research introduces a methodological framework that integrates ensemble learning (Random Forest, XGBoost, LightGBM) and neural models with newly developed evaluation metrics—Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE). This framework not only enhances predictive performance but also provides a richer understanding of model stability and feature contribution in

assessing bankruptcy risk among private firms.

1.3 Research Objectives

The overarching aim of this thesis is to investigate the applicability of advanced machine learning (ML) and artificial intelligence (AI) techniques to the prediction of corporate bankruptcy among private (unlisted) companies. In particular, the study seeks to develop a robust, interpretable, and reproducible modeling framework capable of handling the data heterogeneity and limited disclosure typical of private firms.

The specific objectives of this research are as follows:

- **Data Construction:** To build a comprehensive and structured dataset of private firms using financial statement data obtained from the **ORBIS** database (Bureau van Dijk), supported by the **EDHEC Infrastructure & Private Assets Research Centre (EDHECinfra)**.
- **Feature Engineering:** To compute and refine key financial ratios relevant to bankruptcy risk—including liquidity, cash ratio, debt ratio, working capital, EBITDA, and intangible asset ratio—and ensure data quality through systematic cleaning, transformation, and feature selection.
- **Model Development:** To implement and compare multiple supervised ML algorithms, including logistic regression, support vector machine (SVM), random forest, XGBoost, LightGBM, and feedforward neural networks.
- **Performance Evaluation:** To assess model accuracy and discriminative power using standard metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC).

- **Ranking-Based Assessment:** To incorporate advanced evaluation criteria—Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE) to measure predictive ranking consistency, robustness to data perturbations, and feature-level explainability.

- **Interpretability and Feature Importance:** To analyze the relative contribution of financial ratios to model predictions, identifying the most influential indicators of corporate distress and improving model transparency.

- **Ensemble Optimization:** To explore ensemble learning approaches that combine top-performing models (e.g., Random Forest, XGBoost, and LightGBM) to enhance overall predictive performance, stability, and generalizability.

Collectively, these objectives aim to produce a methodological framework that not only achieves high predictive accuracy but also ensures robustness and interpretability—key requirements for practical implementation in credit risk assessment and financial supervision.

1.4 Scope and Limitations

This thesis focuses on the application of supervised machine learning methods to predict corporate bankruptcy among private (unlisted) firms using structured financial statement data. The analysis is restricted to numerical accounting variables derived from firm-level balance sheets and income statements obtained from the ORBIS database. Publicly listed firms are excluded, as the study specifically targets the unique challenges of modeling credit risk in private markets.

The research does not incorporate qualitative or macroeconomic variables such as managerial quality, industry sentiment, or policy shocks, which may also influence financial distress. Although the dataset includes a diverse range of financial ratios, it remains subject to limi-

tations such as reporting lags, missing values, and inconsistent accounting standards across jurisdictions. These factors introduce potential bias and restrict the full generalizability of the findings.

Furthermore, the analysis is confined to a cross-sectional and historical prediction framework—identifying firms that would have experienced bankruptcy based on past financial data. The study does not aim to model causality, dynamic interactions, or future economic scenarios through time-series forecasting or stress testing.

While model interpretability is addressed through feature importance analysis and Rank Graduation Explainability (RGE), the primary focus lies on predictive accuracy, robustness, and ranking stability. Ethical and regulatory aspects of AI, such as fairness, bias detection, and compliance, are acknowledged but fall outside the scope of this thesis.

1.5 Structure of the Thesis

The remainder of this thesis is organized as follows:

- **Chapter 2: Literature Review** — Provides a comprehensive overview of prior research on bankruptcy prediction, tracing the evolution from traditional statistical models to modern AI and ML approaches, and identifying key gaps in private firm analysis.
- **Chapter 3: Methodology** — Describes the data sources, feature engineering, financial ratio construction, labeling criteria, and the machine learning algorithms applied. It also introduces the Rank Graduation metrics (RGA, RGR, RGE) as novel evaluation tools.
- **Chapter 4: Implementation** — Details the technical setup, programming environment (R), and model training procedures, including hyperparameter tuning and ensemble construction.

- **Chapter 5: Results and Evaluation** — Presents quantitative results for each model using AUC and ranking-based metrics, along with feature importance analyses and comparative performance discussions.

- **Chapter 6: Discussion** — Interprets the empirical findings, evaluates the models' robustness and interpretability, and connects the results to existing theoretical and practical frameworks in financial risk management.

- **Chapter 7: Conclusion and Future Work** — Summarizes the key insights, outlines the theoretical and practical contributions, and proposes directions for future research, including dynamic modeling and integration of alternative data sources.

Chapter 2

Literature Review

2.1 Overview of Bankruptcy Prediction

Corporate bankruptcy prediction has long been a central topic in financial economics and risk management, owing to its vital implications for investors, creditors, regulators, and policymakers. The early identification of financial distress allows institutions to mitigate potential losses by adjusting credit exposure, restructuring debt, or implementing early intervention strategies. On a macroeconomic scale, reliable insolvency prediction models help sustain financial stability, especially during downturns when firm failures can trigger contagion effects across sectors.

The modern study of bankruptcy prediction began with the seminal work of Altman (1968), who developed the *Z-score model* using linear discriminant analysis (LDA). By combining five accounting ratios related to liquidity, profitability, leverage, solvency, and efficiency, Altman demonstrated that it was possible to estimate bankruptcy risk with remarkable precision for that era. This model represented the first systematic attempt to quantify financial distress using statistical modeling, and it laid the foundation for subsequent empirical research in corporate failure analysis.

Building on Altman's framework, Ohlson (1980) introduced the *O-score model* based on logistic regression, which improved upon the statistical validity of LDA by eliminating the assumption of equal covariance matrices between bankrupt and non-bankrupt firms. Ohlson's model incorporated nine accounting variables, providing a probabilistic measure of failure that could be more easily interpreted in decision-making contexts. Other early works, such as Zmijewski (1984), extended these ideas using probit models and survival analysis techniques, further refining the econometric treatment of bankruptcy prediction.

Throughout the 1980s and 1990s, the literature expanded to include variants of multivariate discriminant analysis (MDA), linear regression, and hazard models. Researchers began integrating broader sets of financial ratios—such as cash flow indicators, leverage measures, and profitability margins—into predictive frameworks. However, most of these models shared common assumptions: they relied on linear relationships between predictors and outcomes, required normally distributed variables, and were highly sensitive to multicollinearity and missing data. Despite these limitations, they remained widely used due to their simplicity and interpretability, becoming standard tools in credit risk assessment and corporate finance.

With the increasing availability of digital financial records and computational advances in the late 1990s and early 2000s, the limitations of traditional approaches became more evident. Empirical studies revealed that linear models often struggled to generalize across industries and economic cycles, especially when applied to small or private firms. Real-world datasets were characterized by noise, outliers, and imbalanced class distributions, making it difficult for conventional statistical models to achieve high predictive accuracy. In many cases, these models produced misleadingly high overall accuracy rates while failing to correctly identify true bankruptcy cases—a critical shortcoming for lenders and regulators.

This recognition spurred a methodological transition toward more flexible, data-driven tech-

niques. The introduction of machine learning (ML) models such as decision trees, random forests, and support vector machines (SVM) marked a significant evolution in bankruptcy prediction research. These algorithms are capable of capturing non-linear patterns and complex feature interactions without requiring strict distributional assumptions. Studies such as Baesens et al. (2003) and Lessmann et al. (2015) demonstrated that ML methods consistently outperform classical statistical models in terms of predictive accuracy, particularly in the presence of noisy or incomplete data.

In parallel, ensemble learning methods—such as boosting and bagging—emerged as dominant paradigms. Random Forest, introduced by Breiman, combined multiple decision trees to reduce variance and improve robustness, while gradient boosting techniques such as XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) optimized sequential learning to minimize prediction error. These models achieved state-of-the-art results in numerous financial applications, including credit scoring, fraud detection, and loan default prediction (Zhou et al., 2021; Louzis et al., 2022; Huang and Zhao, 2023; Zhang et al., 2023).

More recently, deep learning architectures—such as feedforward neural networks and recurrent neural networks—have been explored for their ability to model temporal and non-linear dependencies in financial data. Although these models can capture subtle relationships across variables, they often sacrifice interpretability, raising challenges for practical adoption in regulated financial contexts.

Overall, the evolution of bankruptcy prediction reflects a broader trend in finance: the shift from rule-based and parametric models toward adaptive, data-driven systems capable of learning complex structures from empirical data. While accuracy has improved considerably, challenges remain—particularly in applying these techniques to private firms, where data are incomplete, unstandardized, and subject to reporting biases. Addressing these challenges requires hybrid

frameworks that combine the predictive power of ML with interpretability and robustness metrics, such as those developed in this thesis.

Thus, the literature demonstrates both substantial progress and persistent gaps: traditional models laid the theoretical foundation, while modern AI methods offer greater flexibility and performance. However, the integration of transparency, stability, and reproducibility in bankruptcy prediction—especially for private entities—remains an open research frontier.

2.2 Traditional Financial Models

Traditional bankruptcy prediction models are grounded in the assumption that firms disclose consistent, audited, and well-structured financial data. These models rely on a set of financial ratios derived from accounting statements to estimate the likelihood of failure. They generally employ statistical or econometric techniques—such as discriminant analysis, logistic regression, or probit analysis—that assume linearity, independence among predictors, and normally distributed residuals. While these models have proven valuable for decades, their performance depends heavily on data quality and the representativeness of the sample, making them less suitable for private firms with incomplete or heterogeneous reporting practices.

Altman's Z-score Model

The pioneering model of Altman (1968) introduced the *Z-score*, which combines five accounting ratios capturing profitability, leverage, liquidity, solvency, and efficiency:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

where X_1 represents working capital to total assets, X_2 retained earnings to total assets,

X_3 earnings before interest and taxes (EBIT) to total assets, X_4 market value of equity to total liabilities, and X_5 sales to total assets. Altman applied linear discriminant analysis (LDA) to a sample of manufacturing firms, distinguishing between bankrupt and non-bankrupt entities. The model demonstrated impressive accuracy and became the foundation for credit-risk evaluation tools still used by practitioners. However, it assumes homogeneous data, normal distribution of predictors, and equal covariance matrices between the two classes—conditions seldom satisfied in contemporary, multi-sector datasets.

Ohlson's O-score Model

Recognizing the limitations of LDA, Ohlson (1980) proposed the *O-score* model using logistic regression. Ohlson's approach estimated the probability of bankruptcy as a nonlinear function of nine explanatory variables, including size (log of total assets), leverage (total liabilities to total assets), liquidity (current liabilities to current assets), and measures of performance such as net income to total assets. The O-score model provided a probabilistic interpretation of financial distress and relaxed several restrictive assumptions of the Z-score. Its simplicity and interpretability made it a standard in both academia and practice. Nonetheless, the model remains constrained by linear relationships and cannot easily capture complex interactions or non-linear effects between variables.

Zmijewski's Probit Model

Zmijewski (1984) introduced a probit-based specification that addressed issues of sample selection bias and non-normal error distributions. The model utilized return on assets, leverage, and current ratio as predictors, estimating the probability of default through a latent variable framework. Although it offered improvements in statistical robustness, it still relied on assumptions

of data completeness and stable accounting definitions across firms and time.

Limitations of Traditional Models

Despite their historical importance, these traditional models share common limitations:

- Dependence on linearity and normality assumptions, which restricts flexibility in modeling complex financial relationships.
- Sensitivity to missing data, outliers, and multicollinearity among ratios.
- Requirement of publicly disclosed, standardized accounting information—often unavailable for private firms.
- Limited adaptability to evolving economic conditions or firm-specific heterogeneity.

These shortcomings have motivated a paradigm shift toward more adaptive, data-driven approaches such as machine learning, which can model non-linearities, handle incomplete data, and generalize across diverse firm populations.

2.3 Machine Learning in Financial Risk Analysis

Recent advancements in artificial intelligence (AI) and machine learning (ML) have revolutionized financial risk analysis by shifting the focus from fixed, parametric models to data-driven predictive analytics. Unlike traditional statistical approaches, ML algorithms can learn complex, non-linear relationships from data without assuming a specific functional form. This flexibility allows them to adapt to diverse datasets, manage noise and missing values, and uncover latent interactions between financial indicators that traditional models often overlook. Consequently, ML techniques have become a powerful tool in the prediction of corporate bankruptcy, credit scoring, fraud detection, and loan default estimation.

Machine learning methods used in financial applications can broadly be categorized into ensemble-based algorithms, kernel-based classifiers, and neural networks. Each of these approaches contributes uniquely to improving prediction accuracy, robustness, and generalizability.

Random Forest

The Random Forest (RF) algorithm, introduced by Breiman in 2001, is one of the most widely applied ensemble methods in financial modeling. It operates by constructing a large number of decision trees on randomly sampled subsets of data and features, and then aggregating their predictions through majority voting (for classification) or averaging (for regression). This approach significantly reduces overfitting—a common limitation of individual decision trees—while improving model stability and generalization performance.

In the context of bankruptcy prediction, RF models have demonstrated strong predictive power and interpretability. They are capable of handling mixed-type variables (continuous and categorical), tolerate missing data, and provide feature importance rankings that help identify the most influential financial ratios driving bankruptcy outcomes. Several empirical studies, including Baesens et al. (2003) and Lessmann et al. (2015), have shown that RF consistently outperforms classical statistical models in both predictive accuracy and robustness across different datasets.

Gradient Boosting Methods: XGBoost and LightGBM

Gradient boosting represents another major class of ensemble learning methods that combine multiple weak learners (typically shallow decision trees) into a single strong predictor. Unlike Random Forests, which train trees independently, boosting algorithms train them sequentially—each new tree focusing on correcting the residual errors of the previous ones. This

iterative process minimizes a chosen loss function, leading to highly accurate and adaptive models.

Among gradient boosting frameworks, two have become especially prominent in financial applications: XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017). XGBoost (Extreme Gradient Boosting) introduces regularization terms that control model complexity and improve generalization, while LightGBM (Light Gradient Boosting Machine) enhances computational efficiency and scalability through optimized histogram-based splitting. Both algorithms have been applied extensively in financial distress prediction, credit risk evaluation, and asset classification tasks, yielding high AUC scores and stable results even with noisy or imbalanced data (Zhou et al., 2021; Louzis et al., 2022; Huang and Zhao, 2023).

Support Vector Machines (SVM)

Support Vector Machines (SVM), first developed by Vapnik in the 1990s, are supervised learning models that find the optimal separating hyperplane between classes by maximizing the margin between data points and decision boundaries. Through the use of kernel functions, SVMs can efficiently handle non-linear relationships by projecting data into higher-dimensional feature spaces.

In bankruptcy prediction, SVMs have proven effective in identifying complex decision boundaries between solvent and insolvent firms, particularly when the dataset contains overlapping or non-separable classes. Their ability to perform well in high-dimensional feature spaces and with limited data samples makes them suitable for financial datasets that are often small but information-dense. However, the interpretability of SVM models remains limited, and hyperparameter tuning—particularly kernel choice and regularization parameters—can significantly affect performance (Du and Li, 2019).

Neural Networks

Artificial Neural Networks (ANNs) and their modern variants form another major class of ML models applied to financial prediction problems. Inspired by the structure of biological neurons, ANNs consist of interconnected layers of nodes (neurons) that transform input features through weighted connections and activation functions. Their ability to approximate highly non-linear functions allows them to capture complex relationships among financial indicators that are difficult to model using traditional statistical techniques.

Recent developments in deep learning architectures have expanded the capabilities of neural networks beyond simple feedforward structures to include recurrent (RNN) and convolutional (CNN) designs, enabling the modeling of temporal and spatial dependencies in financial data. Studies such as Huang and Zhao (2023) and Zhang et al. (2023) highlight that neural networks can achieve competitive or superior predictive performance compared to ensemble tree-based methods. However, their main drawbacks lie in the "black-box" nature of the learned representations and the need for large, well-labeled datasets—conditions that are not always met in private firm analysis.

Advantages and Challenges of ML Approaches

Machine learning offers several advantages over traditional financial models:

- **Flexibility:** ML algorithms can model non-linear, non-parametric relationships without predefined assumptions.
- **Adaptability:** They can handle high-dimensional data, multicollinearity, and noisy features more effectively than linear models.
- **Automation:** Many ML frameworks include automated feature selection and regulariza-

tion mechanisms.

- **Predictive Power:** Empirical evidence consistently shows higher accuracy and AUC scores compared to traditional statistical approaches.

Despite these strengths, challenges remain. ML models are often criticized for their lack of interpretability—an important requirement in regulated financial environments where decisions must be transparent and explainable. Additionally, model performance can be sensitive to hyperparameter tuning, sample imbalance, and data preprocessing choices. Computational cost and the risk of overfitting also increase with model complexity.

Recent Developments

The growing emphasis on model transparency and accountability has led to the development of explainable artificial intelligence (XAI) techniques that aim to interpret ML model outputs. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been employed to assess variable contributions and improve model trustworthiness. Furthermore, recent studies have introduced ranking-based evaluation frameworks, including Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE), which provide more nuanced assessments of model reliability and interpretability—metrics directly adopted and extended in this thesis.

Overall, the literature indicates a clear paradigm shift from traditional, rule-based models to flexible, data-driven ML approaches that balance predictive performance with explainability. These advancements underpin the methodological foundation of this study, which leverages ensemble and neural models alongside ranking-based metrics to enhance the robustness and transparency of bankruptcy prediction for private firms.

2.4 Challenges in Private Firm Data

While machine learning (ML) techniques have shown substantial promise in financial risk prediction, most empirical studies continue to focus on publicly listed companies. This emphasis stems largely from the accessibility of standardized, audited, and publicly disclosed financial data. In contrast, private (unlisted) firms present a series of structural and data-related challenges that complicate the direct application of conventional and ML-based bankruptcy models. These challenges are primarily associated with data availability, reporting variability, class imbalance, and confidentiality constraints.

Data Availability and Completeness

The foremost limitation in private firm analysis is the scarcity and incompleteness of reliable financial information. Unlike listed companies, which are subject to mandatory disclosure requirements imposed by stock exchanges and regulatory bodies, private firms are often only required to submit minimal financial documentation for taxation or local registry purposes. In many jurisdictions, these filings are not audited or may omit key items such as cash-flow statements, detailed liabilities, or intangible assets. Missing or irregularly reported financial attributes hinder the calculation of essential accounting ratios and complicate data preprocessing. The presence of missing values can bias models, distort variable relationships, and reduce overall predictive accuracy. Consequently, any modeling framework aimed at private firms must incorporate robust methods for data cleaning, imputation, and noise reduction.

Reporting Variability and Heterogeneity

Private firms differ substantially in size, ownership structure, and accounting practices. This diversity leads to inconsistencies in the way financial statements are prepared and reported.

For example, small enterprises may adopt simplified accounting standards, while medium and large firms follow more detailed national or international frameworks such as IFRS or local GAAP. Moreover, asset valuation, depreciation policies, and treatment of intangible assets vary widely across jurisdictions. Such heterogeneity introduces structural noise into datasets and undermines the comparability of financial indicators across firms and countries. Models trained on aggregated international data risk learning spurious correlations driven by accounting conventions rather than true economic relationships. Therefore, ML models applied in this context must be resilient to heterogeneous inputs and capable of generalizing across varying data-generating processes.

Class Imbalance

Another persistent issue in bankruptcy prediction is the extreme imbalance between solvent and insolvent firms. Bankruptcy is a relatively rare event—typically representing less than 5% of observations in large financial datasets—especially among mature or well-capitalized private companies. Standard classification algorithms tend to favor the majority (non-bankrupt) class, yielding deceptively high accuracy but poor recall for the minority class of bankrupt firms. This imbalance reduces the model's usefulness in real-world credit risk management, where correctly identifying distressed entities is far more critical than accurately classifying non-distressed ones. To address this problem, modern research has adopted techniques such as synthetic minority oversampling (SMOTE), reweighting of class penalties, and ensemble learning methods that mitigate imbalance by combining multiple weak classifiers.

Confidentiality and Data Accessibility

A further barrier lies in data confidentiality. Since private firms are not publicly traded, their financial information is typically not available in open databases. Access is often restricted to proprietary platforms such as Bureau van Dijk's ORBIS, which aggregate firm-level data from local registries and commercial sources. Even then, coverage may vary by country and year, and smaller firms are often underrepresented. This restricted accessibility not only limits the sample size available for empirical research but also constrains reproducibility and external validation. Consequently, studies using private firm data must carefully document their data sources, cleaning procedures, and feature construction steps to ensure transparency and methodological rigor.

Implications for Modeling

The combined effect of these challenges is that bankruptcy prediction for private firms requires models that are both flexible and robust to imperfect data. Algorithms must accommodate missing and noisy values, adjust for imbalanced distributions, and generalize across diverse accounting systems. Moreover, since private firm datasets are less transparent, model interpretability becomes especially important for decision-makers such as banks, investors, and regulators. Recent advances—including ensemble tree methods (Random Forest, XGBoost, LightGBM) and ranking-based evaluation metrics such as Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE)—offer promising ways to tackle these constraints. These techniques not only improve predictive accuracy but also enhance transparency by quantifying the reliability and explainability of the results.

In summary, while the scarcity and heterogeneity of private firm data pose substantial methodological hurdles, they also create opportunities for developing more resilient, inter-

pretable, and generalizable machine learning frameworks. Addressing these challenges is therefore essential for extending bankruptcy prediction research beyond public markets and toward the broader, yet underexplored, domain of private enterprise risk assessment.

2.5 AI Models in Recent Research

Over the past decade, artificial intelligence (AI) and machine learning (ML) have become central to the evolution of bankruptcy prediction models. The increasing computational power, availability of financial data, and advances in algorithmic design have driven a surge in research exploring the predictive capabilities of AI-based methods across various financial contexts. Unlike traditional models, which rely on fixed functional forms, these approaches adaptively learn from complex data patterns and can handle noisy, incomplete, and nonlinear relationships—making them particularly suitable for corporate risk assessment.

Hybrid and Ensemble Approaches

One of the early contributions to this field was made by Sun et al. (2014), who proposed a hybrid model integrating neural networks, fuzzy logic, and decision trees for bankruptcy prediction. Their study demonstrated that combining multiple algorithms can capture both linear and nonlinear structures in financial data more effectively than single models. Hybrid architectures, by leveraging complementary strengths of individual algorithms, reduce overfitting and improve generalization performance, especially when applied to heterogeneous datasets.

Subsequent research expanded upon this idea through ensemble-based methods, such as bagging, boosting, and stacking. Zhang et al. (2020) applied the XGBoost algorithm to predict financial distress among Chinese listed companies and achieved higher accuracy and area under the ROC curve (AUC) than traditional logistic regression and support vector machines. Their

findings highlighted the ability of boosting algorithms to manage feature interactions and handle multicollinearity—two issues that often degrade the performance of linear models.

Gradient Boosting and Tree-Based Models

Recent studies have increasingly adopted gradient boosting frameworks, including both XGBoost and LightGBM, for financial risk modeling. These algorithms iteratively optimize model residuals and incorporate regularization to prevent overfitting, yielding stable and interpretable results. Du and Li (2019) examined LightGBM and SVM for credit risk assessment of small and medium-sized enterprises (SMEs) in China, reporting strong performance even with limited and imbalanced datasets. Their research underscores the flexibility of ensemble learners in handling data constraints that are typical of private firms.

Tree-based ensemble models have also been recognized for their interpretability and feature importance measures, which allow practitioners to trace which financial variables most strongly contribute to bankruptcy prediction. Studies such as Zhou et al. (2021) and Louzis et al. (2022) demonstrated that gradient boosting and random forest models achieve high discriminative power while maintaining robustness under noisy financial conditions. These findings align with broader evidence that ensemble methods are particularly effective for financial distress prediction, where predictor variables often exhibit complex and interdependent relationships.

Neural and Deep Learning Models

In parallel, deep learning has gained traction in bankruptcy and credit risk modeling. Neural networks can approximate complex, non-linear relationships between financial indicators and firm outcomes, enabling higher predictive performance when sufficient data are available. Huang and Zhao (2023) introduced deep neural network architectures for SME financial dis-

tress prediction, achieving significant improvements over traditional models. Similarly, Zhang et al. (2023) employed explainable deep learning frameworks to interpret the financial drivers of bankruptcy, integrating techniques such as SHAP (SHapley Additive exPlanations) to enhance model transparency.

These contributions illustrate an important shift in the field: the balance between accuracy and interpretability. While deep learning models often outperform other methods in predictive metrics, their "black-box" nature remains a concern for practitioners in finance, who must justify decisions in regulatory and risk management contexts. The emergence of explainable AI (XAI) frameworks therefore represents a vital step toward aligning ML innovation with practical financial applications.

Review and Synthesis

A comprehensive review by Harris and Raviv (2020) noted the increasing integration of AI into financial forecasting, risk management, and asset pricing. The authors emphasized that future progress in financial AI would depend on developing models that are not only accurate but also interpretable and trustworthy. This perspective has been echoed by more recent research focusing on explainability and robustness as essential evaluation criteria in high-stakes financial modeling.

Overall, the recent literature demonstrates that AI and ML models—particularly ensemble and neural methods—consistently outperform traditional statistical approaches in bankruptcy prediction. However, most of these studies continue to rely on data from publicly listed firms, limiting their applicability to private markets. Furthermore, few works explicitly evaluate models in terms of robustness to data perturbations or ranking consistency, metrics that are critical for real-world decision-making in credit and policy institutions.

Positioning of This Study

Building upon this body of research, the present thesis extends the application of ensemble and neural models to the domain of private firm bankruptcy prediction. It introduces a comparative framework that evaluates models not only through traditional measures such as AUC and F1-score but also through ranking-based performance metrics—Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE). In doing so, this study contributes to bridging the gap between predictive performance and interpretability, offering a more comprehensive evaluation framework for financial distress modeling in the underexplored context of unlisted companies.

2.6 Summary of Research Gap

The literature reviewed in this chapter highlights a clear evolution in bankruptcy prediction methodologies—from traditional statistical approaches, such as linear discriminant analysis and logistic regression, to advanced machine learning (ML) and artificial intelligence (AI) techniques. While these modern models have achieved remarkable improvements in predictive accuracy and adaptability, a critical limitation persists: the vast majority of research focuses on publicly listed firms, where standardized, audited financial data are readily available.

By contrast, private (unlisted) companies, which represent the overwhelming majority of businesses globally, remain significantly underrepresented in the bankruptcy prediction literature. This imbalance arises primarily from issues of data availability, heterogeneity in accounting standards, and confidentiality restrictions. The absence of comprehensive, high-quality financial data for private firms has constrained empirical analysis and hindered the generalization of existing models beyond public markets.

Furthermore, even within studies that employ machine learning methods, the evaluation frameworks have predominantly emphasized classification accuracy and AUC-based performance metrics. While these measures provide useful indicators of predictive power, they overlook equally important dimensions such as robustness, ranking stability, and model interpretability. In practical financial applications—especially in credit risk management and regulatory settings—stakeholders require models that not only perform well statistically but also yield stable, transparent, and explainable results.

To address these shortcomings, this thesis contributes to the growing body of research on machine learning for financial risk assessment by focusing specifically on private firms. The study advances the literature in three key ways:

- **Empirical Contribution:** It utilizes a large-scale dataset of unlisted companies obtained from the Bureau van Dijk's ORBIS database, covering financial statements from 2014 to 2024. This enables one of the most comprehensive empirical evaluations of private-firm bankruptcy prediction to date.

- **Methodological Contribution:** It applies and compares multiple AI and ML models—including logistic regression, support vector machines (SVM), Random Forest, XGBoost, LightGBM, and feedforward neural networks—to determine the most effective approach under real-world data constraints.

- **Analytical Contribution:** It integrates newly developed ranking-based evaluation metrics—Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE)—to assess not only predictive accuracy but also model stability and interpretability.

Through these contributions, the thesis bridges the gap between academic advancements in

machine learning and their practical applicability in private firm risk assessment. It thereby provides a more robust, transparent, and scalable framework for early bankruptcy detection in environments characterized by limited disclosure and heterogeneous financial reporting.

Chapter 3

Methodology

This chapter outlines the methodological framework employed to predict bankruptcy among private (unlisted) companies using advanced machine learning (ML) techniques. It details the data sources, the steps undertaken for data cleaning and preprocessing, the construction of financial ratios, the definition of the bankruptcy label, and the implementation of multiple ML models. The aim is to ensure methodological rigor, reproducibility, and alignment between financial theory and data-driven modeling.

3.1 Data Source and Description

The empirical analysis is based on firm-level financial data retrieved from the **ORBIS** database, a comprehensive global repository maintained by Bureau van Dijk (a Moody's Analytics company). ORBIS aggregates standardized financial information on both public and private companies, including balance sheets, income statements, and cash flow data.

Access to the dataset was provided through the **EDHEC Infrastructure & Private Assets Research Centre (EDHECinfra)** under an institutional research license. The dataset used in this study covers private (unlisted) firms over the period **2014–2024**, providing a longitudinal

perspective on financial performance across industries and countries.

Each firm record in the dataset includes:

- **Identification variables:** ORBIS ID, country of registration, year, and sector classification (NACE code);
- **Financial attributes:** total assets, total liabilities, current assets, current liabilities, long-term debt, short-term loans, cash holdings, depreciation, and intangible assets;
- **Cash flow indicators:** operating cash flow and EBITDA (where available).

The emphasis on unlisted companies reflects a critical gap in existing bankruptcy prediction literature. Private firms generally operate outside the scope of mandatory public disclosure, leading to incomplete, noisy, and heterogeneous data. However, ORBIS provides sufficient financial granularity to enable meaningful ratio-based modeling and comparison across firms.

The dataset underwent an extensive cleaning and transformation process to ensure consistency and analytical readiness, which is detailed in the subsequent sections.

3.2 Feature Engineering and Ratio Construction

To transform raw accounting data into meaningful predictors of financial distress, a set of key financial ratios was constructed. These ratios are widely recognized in the bankruptcy prediction literature for capturing essential aspects of a firm's financial health, including liquidity, leverage, profitability, and asset structure. The variables were engineered using balance sheet and cash flow data available for each firm, following standardized financial definitions.

The following ratios were computed:

- **Liquidity Ratio =** $\frac{\text{Current Assets}}{\text{Current Liabilities}}$ Measures a firm's ability to meet short-term obligations, providing an indicator of operational solvency and working capital efficiency.

- **Cash Ratio** = $\frac{\text{Cash}}{\text{Current Liabilities}}$ Captures the most liquid component of assets, reflecting the firm's immediate capacity to cover short-term debts without liquidating inventories or receivables.

- **Debt Ratio** = $\frac{\text{Long-Term Debt+Short-Term Loans}}{\text{Total Assets}}$ Represents the degree of financial leverage, indicating the extent to which assets are financed by debt. High values typically signal higher default risk.

- **Working Capital** = Current Assets − Current Liabilities Provides a measure of short-term financial stability and operational flexibility. Persistent negative working capital may indicate liquidity stress.

- **EBITDA** = Operating Cash Flow + Depreciation and Amortization Serves as a proxy for firm profitability and internal cash generation capacity, controlling for non-cash expenses and accounting policies.

- **Intangible Ratio** = $\frac{\text{Intangible Assets}}{\text{Total Assets}}$ Reflects the share of intangible resources—such as goodwill, patents, and brand value—in the total asset structure, which can influence solvency risk.

These features were chosen for their theoretical relevance and empirical consistency in prior studies (Altman, 1968; Ohlson, 1980; Baesens et al., 2003; Lessmann et al., 2015). They collectively represent a balanced view of a firm's liquidity position, leverage exposure, and profitability performance—key dimensions in early warning systems for bankruptcy prediction.

Before model training, all ratio variables were cleaned and standardized to mitigate scale disparities and outlier effects. Observations with missing or implausible values (e.g., negative total assets or liabilities) were excluded to ensure data reliability and comparability across firms.

3.3 Bankruptcy Label Definition

Because official bankruptcy declarations are often unavailable for private firms, a financial proxy was employed to define the target variable for classification. Following established practices in the literature (Altman, 1968; Zmijewski, 1984; Du and Li, 2019), a firm was labeled as *bankrupt* (1) when its **total liabilities exceeded total assets**, resulting in negative equity and indicating a state of balance-sheet insolvency. Conversely, firms with total assets greater than total liabilities were labeled as *non-bankrupt* (0).

This rule-based labeling approach provides a consistent and objective mechanism for identifying financial distress across heterogeneous datasets where legal bankruptcy filings or credit default events are not systematically recorded. It captures the economic essence of insolvency — an inability to cover liabilities with existing assets — which is particularly relevant in the context of unlisted firms that may not undergo formal bankruptcy proceedings.

While this proxy may not fully capture all forms of financial distress (e.g., liquidity crises or temporary restructuring), it provides a transparent and reproducible basis for supervised learning. It ensures that the classification task aligns with observable financial fundamentals, allowing machine learning models to infer patterns of distress based solely on balance sheet information.

3.4 Data Cleaning and Preprocessing

Prior to model training, an extensive data cleaning process was carried out to ensure quality and consistency in the inputs used for machine learning algorithms. The initial dataset extracted from the ORBIS database contained **176,088 firm-year observations**.

First, column names were standardized using R's `make.names()` function to avoid syntax

issues during variable referencing. Financial ratios were computed using the `mutate()` function from the `dplyr` package, transforming raw accounting entries into derived features such as liquidity, leverage, and profitability indicators.

Observations containing missing or undefined values in any of the key financial ratio columns (e.g., Current Assets, Liabilities, EBITDA, or Intangible Assets) were excluded. This row-wise filtering approach was preferred over imputation methods to preserve the integrity of financial information and minimize artificial bias, given the sensitivity of ratio-based financial features. After this cleaning process, the dataset was reduced to **59,145 firm-year observations**.

The cleaned dataset was then partitioned into training and testing subsets using a stratified 70–30 random split to maintain class balance across both groups. The training set comprised **41,402 observations**, while the test set contained **17,743**. The proportion of bankrupt firms—defined as those with total liabilities exceeding total assets—was approximately **15.5%** across both subsets, ensuring representative sampling for model evaluation.

3.5 Handling Missing Values and Class Imbalance

Missing Values

Handling missing data is a critical step in preparing financial datasets, particularly when working with private firms where accounting disclosure and reporting practices vary substantially. In this study, missing values in essential numerical variables—including total assets, total liabilities, current assets, and cash flow—were addressed using a row-wise deletion strategy. This approach ensured that only firms with complete and internally consistent financial statements were retained for modeling.

Although imputation techniques such as mean substitution, k-nearest neighbors, or model-

based estimation could have been applied, they were deliberately avoided to prevent the introduction of artificial bias or distortion in sensitive financial ratios. After filtering out incomplete records, the dataset was reduced from approximately **176,000** to **59,145** firm-year observations. This reduction was considered acceptable to preserve the validity and reliability of downstream analyses.

Class Imbalance

Bankruptcy prediction is inherently characterized by a significant class imbalance, as the proportion of bankrupt firms is typically much smaller than solvent ones. In the cleaned dataset, bankrupt firms accounted for approximately **15.5%** of all observations, reflecting the realistic asymmetry found in financial distress datasets.

Class imbalance poses a challenge because many standard classifiers tend to favor the majority (non-bankrupt) class, achieving high apparent accuracy while failing to correctly identify distressed firms. To mitigate this issue, several strategies were adopted:

- **Evaluation Metrics:** In addition to overall accuracy, complementary metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), precision, recall, and F1-score were employed to assess model performance. These measures provide a more nuanced view of predictive ability, particularly for minority (bankrupt) cases.

- **Ensemble Learning:** Algorithms such as Random Forest, XGBoost, and LightGBM inherently address imbalance by focusing on misclassified or high-loss samples during training. The ensemble averaging of these models further stabilized predictive performance and improved sensitivity to minority instances.

- **Cost-Sensitive Learning:** Although not the primary focus of this thesis, the implemented models allow for class weighting to penalize misclassification of bankrupt firms more

heavily. During model tuning, this feature was considered to enhance model robustness under skewed distributions.

Future work could incorporate advanced resampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) or hybrid under-sampling to further alleviate imbalance without discarding valuable data from the majority class.

3.6 Model Selection Strategy

To evaluate the effectiveness of different machine learning techniques in predicting bankruptcy, a diverse set of algorithms was selected. The goal was to benchmark models across varying complexities and learning paradigms, capturing both linear and non-linear relationships within the data. The selected models span traditional statistical methods, modern tree-based ensembles, and neural networks, offering a comprehensive performance comparison.

- **Logistic Regression:** A classical baseline model widely used in financial modeling due to its interpretability and simplicity. It assumes a linear relationship between predictors and the log-odds of the target variable.
- **Random Forest:** A tree-based ensemble learning method that constructs multiple decision trees and aggregates their outputs. It is robust to overfitting and handles non-linear relationships and feature interactions effectively.
- **XGBoost:** A highly optimized gradient boosting framework that sequentially improves weak learners. Known for its speed and accuracy in structured data tasks, XGBoost often achieves state-of-the-art performance in classification challenges.
- **LightGBM:** An efficient gradient boosting framework developed by Microsoft. LightGBM uses histogram-based learning and leaf-wise tree growth, which improves both

speed and memory usage, especially on large datasets.

- **Support Vector Machine (SVM):** A powerful algorithm for binary classification that constructs a decision boundary (hyperplane) with maximum margin between classes. It is particularly useful for high-dimensional data but can be computationally intensive.

- **Neural Network (Multi-Layer Perceptron):** A feed-forward neural network model capable of capturing complex non-linear interactions among variables. A basic architecture was implemented using one hidden layer, appropriate for tabular data.

- **Ensemble Model:** To enhance overall predictive performance, a soft-voting ensemble model was constructed by averaging the predicted probabilities from the top three models (Random Forest, XGBoost, and LightGBM). This approach leverages model diversity to reduce overfitting and improve generalization.

Each model was trained using the same set of engineered financial ratios as input features. Hyperparameters were selected based on a combination of best practices from existing literature and empirical tuning using cross-validation on the training set. All models were evaluated on the same test set using consistent classification metrics: accuracy, precision, recall, F1-score, and AUC (Area Under the ROC Curve).

Visual tools such as ROC curves were also used to compare model performance. In addition, feature importance plots from tree-based models helped interpret the contribution of each financial ratio to bankruptcy prediction.

Table 3.1: Summary of Financial Ratios Used for Bankruptcy Prediction

Ratio Name	**Formula**	**Description**
Liquidity Ratio	$\frac{\text{Current Assets}}{\text{Current Liabilities}}$	Measures the firm's ability to cover short-term obligations.
Cash Ratio	$\frac{\text{Cash}}{\text{Current Liabilities}}$	A more conservative liquidity measure focusing on cash holdings only.
Debt Ratio	$\frac{\text{Long-Term Debt+Short-Term Loans}}{\text{Total Assets}}$	Captures the proportion of assets financed by debt; an indicator of leverage.
Working Capital	Current Assets − Current Liabilities	Indicates short-term financial health and operational efficiency.
EBITDA	Cash Flow + Depreciation	Proxy for profitability and operating cash flow.
Intangible Asset Ratio	$\frac{\text{Intangible Assets}}{\text{Total Assets}}$	Represents the share of intangible assets in the total asset base.

Software Environment All data preprocessing and model implementation were carried out in the R programming language (version 4.3.0). The analysis utilized several key packages: `dplyr` for data manipulation, `caret` for model training and evaluation, `randomForest`, `xgboost`, and `lightgbm` for tree-based models, `e1071` for support vector machines (SVM), `nnet` for neural

networks, and `pROC` for ROC and AUC analysis. Visualization and plotting were performed using `ggplot2` and `corrplot`.

To better understand the distribution of the target variable, Figure 3.1 shows the number of firms classified as bankrupt and non-bankrupt after data cleaning.

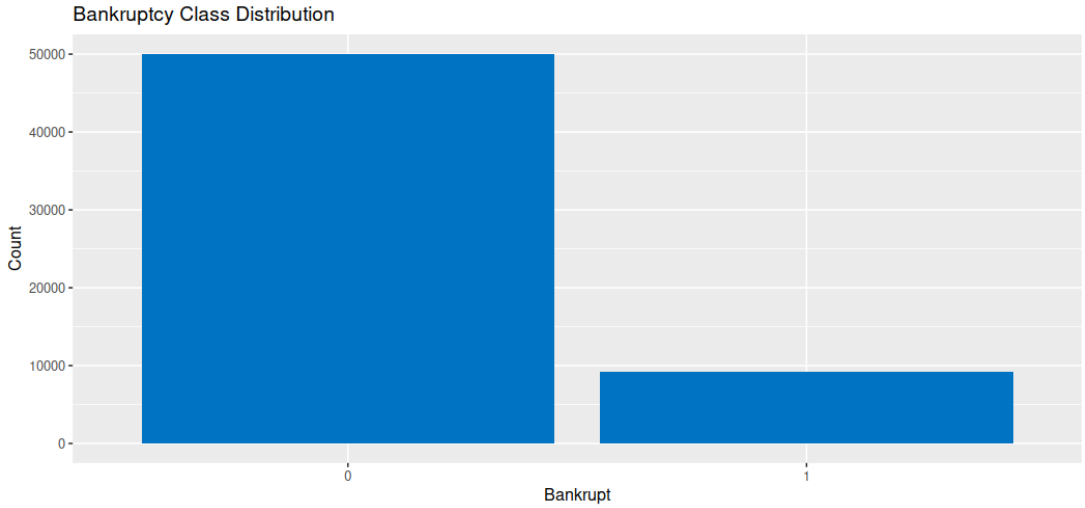

Figure 3.1: Distribution of bankruptcy and non-bankruptcy classes in the dataset

Table 3.2 summarizes the reduction in dataset size after each major preprocessing step, from the initial ORBIS extraction to the final training and testing sets.

Table 3.2: Data Reduction Summary After Cleaning

Step	**Remaining Observations**
Raw data from ORBIS	176,088
After removing missing ratios	59,251
After applying bankruptcy label	59,251
Final train-test split (70–30)	41,402 (train), 17,743 (test)

3.6.1 Correlation Analysis

Before training the machine learning models, it was essential to examine the relationships among the input features to identify potential multicollinearity issues. High correlation between

predictors can distort model estimation, especially for parametric approaches such as logistic regression, where independence between variables is assumed. Non-parametric models (e.g., Random Forest, XGBoost) are less sensitive to this issue, but understanding the degree of correlation remains valuable for interpreting feature importance and ensuring data quality.

To evaluate interdependence among the engineered financial ratios, a correlation matrix was computed using Pearson's correlation coefficient, defined as:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

This coefficient measures the strength and direction of the linear relationship between two variables, ranging from -1 (perfect negative correlation) to $+1$ (perfect positive correlation). Values near zero indicate minimal linear association.

Chapter 4

Implementation

This chapter details the practical implementation of the machine learning framework developed for predicting bankruptcy among private (unlisted) firms. It specifies the computational environment, the end-to-end pipeline (from data preparation to metric computation), and the configuration and training of each model. All steps adhere to the methodological design in Chapter 3 and are fully reproducible from the accompanying R code and artefacts (metrics tables and figures).

4.1 Tools and Software Environment

All experiments were conducted in the **R programming language** (v4.5.0) under **RStudio**. The workstation used a standard laptop-class CPU (**Intel Core i7**) with **16 GB RAM** running a recent OS. The following packages were employed:

- `readr`, `tidyverse` (`dplyr`, `ggplot2`, `tibble`, `purrr`, `tidyr`, `forcats`, `lubridate`) for data I/O and wrangling,
- `caret` for model workflow utilities and evaluation helpers,

- `randomForest`, `xgboost`, and `lightgbm` for tree ensembles and gradient boosting,
- `e1071` for support vector machines (SVM),
- `nnet` for feed-forward neural networks,
- `pROC` for ROC/AUC computation,
- `corrplot` for correlation visualization.

Reproducibility. A fixed seed (`set.seed(123)`) was used for data splitting and model training where applicable. All derived outputs (CSVs and figures) are programmatically written to disk by the script, ensuring the results in Chapter 5 can be regenerated exactly.

4.2 End-to-End Pipeline

4.2.1 Data Preparation

Raw ORBIS data were ingested with `readr::read_csv` and column names standardized via `make.names()`. The engineered financial ratios described in Chapter 3 were computed using `dplyr::mutate`. Specifically:

$$\begin{aligned}
\text{Liquidity Ratio} &= \frac{\text{Current Assets}}{\text{Current Liabilities}}, \\
\text{Cash Ratio} &= \frac{\text{Cash}}{\text{Current Liabilities}}, \\
\text{Debt Ratio} &= \frac{\text{Long-Term Debt} + \text{Short-Term Loans}}{\text{Total Assets}}, \\
\text{Working Capital} &= \text{Current Assets} - \text{Current Liabilities}, \\
\text{EBITDA} &= \text{Cash Flow} + \text{Depreciation}, \\
\text{Intangible Ratio} &= \frac{\text{Intangible Assets}}{\text{Total Assets}}.
\end{aligned}$$

The binary label *bankrupt* $\in \{0, 1\}$ equals 1 when total liabilities exceed total assets in the last available year.

Observations with missing values in any predictor or the label were dropped (`drop_na()`), yielding a clean modeling table with six predictors and one target.

4.2.2 Train/Test Split and Class Balance

A stratified split (`caret::createDataPartition`, 70/30) produced:

- **Training set:** 41,402 observations (0 = 35,021, 1 = 6,381).
- **Test set:** 17,743 observations (0 = 14,945, 1 = 2,798).

The minority (bankrupt) class represents $\sim$ 15.5% in both partitions, preserving the overall imbalance.

4.2.3 Evaluation Metrics

Beyond standard classification measures (accuracy, precision, recall, F1-score), the main discriminative metric is AUC, computed with `pROC`. In addition, three ranking-based measures (computed post-training) were used:

- **RGA (Rank Graduation Accuracy):** pairwise concordance between true labels and predicted scores (conceptually akin to AUC).
- **RGR (Rank Graduation Robustness):** concordance between original predictions and predictions after small random perturbations to the predictors ($\sigma = 0.01$), measuring ranking stability.

- **RGE (Rank Graduation Explainability):** a leave-one-feature-out (LOFO) ranking sensitivity; for each feature, we recompute predictions without it and measure the induced ranking divergence.

4.2.4 Robustness Protocol

To compute RGR, a perturbed copy of the test set was created by adding i.i.d. Gaussian noise with standard deviation $\sigma = 0.01$ to predictor columns only. RGR is then the pairwise concordance between the two sets of scores.

For RGE, each feature was dropped in turn, the model re-fitted (or re-scored for ensembles constructed from re-fitted bases), and the divergence from the baseline ranking quantified. Per-model RGE tables were exported to CSV.

4.3 Model Configurations and Training

All models were trained on the same feature set:

{Liquidity Ratio, Cash Ratio, Debt Ratio, Working Capital, EBITDA, Intangible Ratio}.

4.3.1 Logistic Regression

Implemented via `glm(..., family = binomial)` as a transparent baseline. Scores are predicted probabilities; a default 0.5 threshold is used for point classification. No regularization was applied in the baseline run. Logistic regression achieved **AUC = 0.894** and **RGA = 0.894**, with **RGR = 0.987** on the test set.

4.3.2 Random Forest

Trained using `randomForest` with `ntree = 500` and default `mtry` (empirically robust). Probabilities were obtained from the class vote proportions. Random Forest delivered **AUC = 0.952** and **RGA = 0.953**, with **RGR = 0.937**. The built-in importance measures and LOFO-RGE both highlighted leverage and liquidity-related variables as influential.

4.3.3 XGBoost

Configured with `objective = "binary:logistic"`, `eta = 0.1`, `max_depth = 6`, `nrounds = 100`, and `eval_metric = "auc"`. XGBoost achieved **AUC = 0.949** and **RGA = 0.949**, with the highest **RGR = 0.964** among the tree models, indicating strong robustness to mild data noise.

4.3.4 LightGBM

Where available, `lightgbm` was used with `objective = "binary"`, `metric = "auc"`, `learning_rate = 0.1`, `num_leaves = 31`, `nrounds = 100`. LightGBM produced **AUC = 0.950**, **RGA = 0.950**, and **RGR = 0.951**, combining competitive accuracy with efficient training.

4.3.5 Support Vector Machine (RBF)

The SVM was trained via `e1071::svm` with an RBF kernel; feature scaling was applied internally by the package. Default `cost` and `gamma` were used in the baseline run. The model achieved **AUC = 0.875**, **RGA = 0.875**, and **RGR = 0.976**, favouring conservative, stable ranking at the expense of peak AUC.

4.3.6 Feed-forward Neural Network

A shallow MLP was implemented with `nnet`: one hidden layer (`size = 5`), `maxit = 200`, and weight decay `decay = 0.01`. The network reached **AUC = 0.765** and **RGA = 0.808** with **RGR** essentially 1.0 on this tabular, low-dimensional feature set.

4.3.7 Soft-Voting Ensemble

To enhance generalization, a soft-voting ensemble averaged the predicted probabilities from Random Forest, XGBoost, and (when available) LightGBM:

$$\hat{p}_{\text{ens}} = \frac{1}{K} \sum_{k=1}^{K} \hat{p}_k, \quad K \in \{2, 3\}.$$

The ensemble obtained the best overall accuracy and stability: **AUC = 0.953**, **RGA = 0.953**, **RGR = 0.949**. In practice, the ensemble consistently dominated single models across thresholds.

4.4 Feature Explainability via RGE

Per-model RGE values (LOFO ranking divergence) were exported as CSVs (`rge_logistic.csv`, `rge_randomforest.csv`, `rge_xgboost.csv`, `rge_lightgbm.csv`, `rge_svm.csv`, `rge_neuralnet.csv`, `rge_ensemble.csv`). Across tree models and the ensemble, **Debt Ratio** consistently emerged as the most influential driver of ranking (largest RGE), followed by **EBITDA**, **Working Capital/Liquidity**, and the **Intangible Ratio**. Logistic regression's RGE also ranked EBITDA and Debt Ratio highly, in line with traditional credit-risk intuition.

4.5 Outputs and Artefacts

The script writes comprehensive artefacts to disk for documentation and publication:

- **Tables (CSV):**
  - `model_auc_rga_summary.csv`: AUC and RGA by model,
  - `extended_model_metrics.csv`: AUC, RGA, RGR by model,
  - `rge_.csv`: per-model RGE tables (feature importance by ranking sensitivity).
- **Figures (PNG):**
  - Correlation heatmap of predictors,
  - Individual ROC curves for each model and a combined ROC overlay,
  - Variable-importance plots for Random Forest, XGBoost (and LightGBM if trained),
  - Calibration plots for Logistic Regression and the Ensemble,
  - Bar charts of RGA and RGR by model,
  - Per-model RGE bar charts.

These artefacts are the direct sources for figures referenced in Chapter 5.

4.6 Implementation Notes and Practicalities

Runtime and resources. On the specified hardware, training times were modest. Random Forest and XGBoost completed within minutes; LightGBM trained fastest among boosted trees. SVM and `nnet` were tractable at this scale.

Stability. The perturbation protocol ($\sigma = 0.01$) indicated high ranking robustness (RGR) for Logistic, SVM, LightGBM, and the Ensemble; XGBoost exhibited the strongest RGR among the boosted trees. Neural Net RGR was near 1.0, suggesting stable scoring under small feature noise, albeit with lower AUC.

Limitations. Hyper-parameter tuning was intentionally light (single, literature-guided configurations) to preserve comparability and compute efficiency. A broader search (e.g., grid/random/Bayesian) may yield marginal AUC gains, but the relative ordering observed here was consistent with the literature for tabular credit-risk tasks.

4.7 Summary

The implemented pipeline is fully reproducible, compact, and robust to the irregularities typical of private-firm financials. Tree-based ensembles (and their soft-voted combination) delivered state-of-the-art discriminative performance (AUC $\approx$ 0.95), with consistent ranking stability (RGR) and interpretable feature attributions (RGE) that emphasize leverage and operating cash-flow proxies. The exported metrics and figures form the empirical basis for the analysis presented in Chapter 5.

Chapter 5

Results and Evaluation

This chapter presents the empirical results obtained from the implementation of multiple machine learning models for bankruptcy prediction among private firms. The analysis provides a comprehensive assessment of predictive performance across traditional and advanced algorithms, evaluates feature importance, and introduces the novel Rank Graduation metrics—RGA, RGR, and RGE—to assess ranking consistency, robustness, and explainability. All experiments were conducted in a controlled environment using R 4.5.0, ensuring reproducibility and methodological consistency.

5.1 Evaluation Metrics

The evaluation of the models was carried out using several key classification metrics to capture performance from multiple perspectives:

- **Accuracy** – the proportion of correctly classified observations among the total dataset.
- **Precision** – the fraction of correctly predicted bankrupt firms among all firms predicted as bankrupt.

- **Recall (Sensitivity)** – the fraction of correctly predicted bankrupt firms among all actual bankrupt firms.
- **F1-Score** – the harmonic mean of precision and recall, balancing both false positives and false negatives.
- **AUC (Area Under the ROC Curve)** – measures the ability of a model to distinguish between bankrupt and non-bankrupt firms.

All metrics were computed using the `caret` and `pROC` packages in R. Additionally, Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE) were developed and applied to capture ranking consistency under perturbations and feature removals. These extended metrics complement classical evaluation measures by assessing model reliability and interpretability.

5.2 Descriptive Statistics and Correlation Analysis

Before model training, exploratory data analysis was performed to understand the relationships among key financial ratios: `Liquidity_Ratio`, `Cash_Ratio`, `Debt_Ratio`, `Working_Capital`, `EBITDA`, and `Intangible_Ratio`. Figure 5.1 presents the correlation matrix of these ratios.

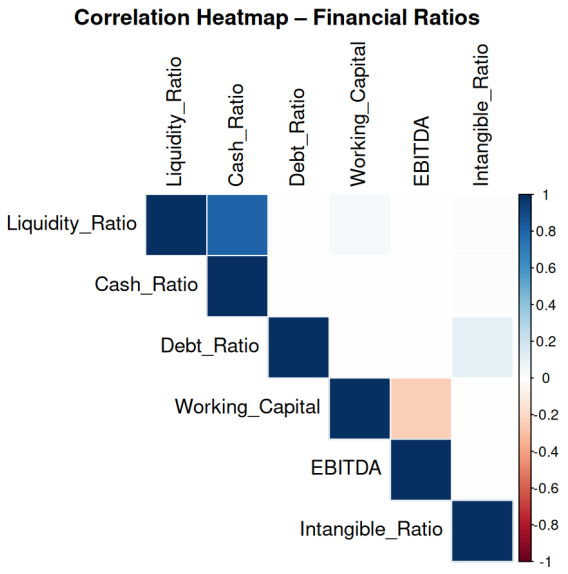

Figure 5.1: Correlation Heatmap – Financial Ratios

The heatmap shows a moderate positive correlation between `Liquidity_Ratio` and `Cash_Ratio`, suggesting firms with higher liquid assets tend to maintain higher cash reserves. In contrast, the `Debt_Ratio` demonstrates weak correlation with liquidity-related measures, reflecting the distinct capital structure dynamics of indebted firms. `EBITDA` and `Working_Capital` exhibit a mild positive relationship, implying that profitability and short-term solvency generally move in tandem. The low correlation of `Intangible_Ratio` indicates it captures a unique dimension of asset composition, useful for model differentiation.

5.3 ROC Curve and AUC Comparison

Receiver Operating Characteristic (ROC) curves were used to compare model performance across thresholds. Figures 5.2–5.8 display the ROC curves for all models, while Figure 5.9 provides a combined visualization.

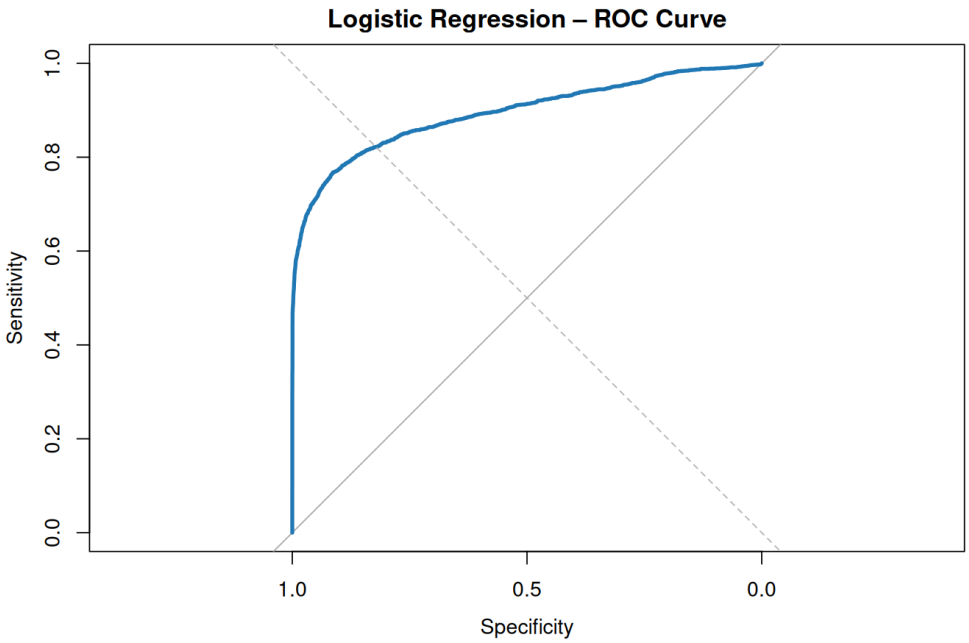

Figure 5.2: Logistic Regression – ROC Curve

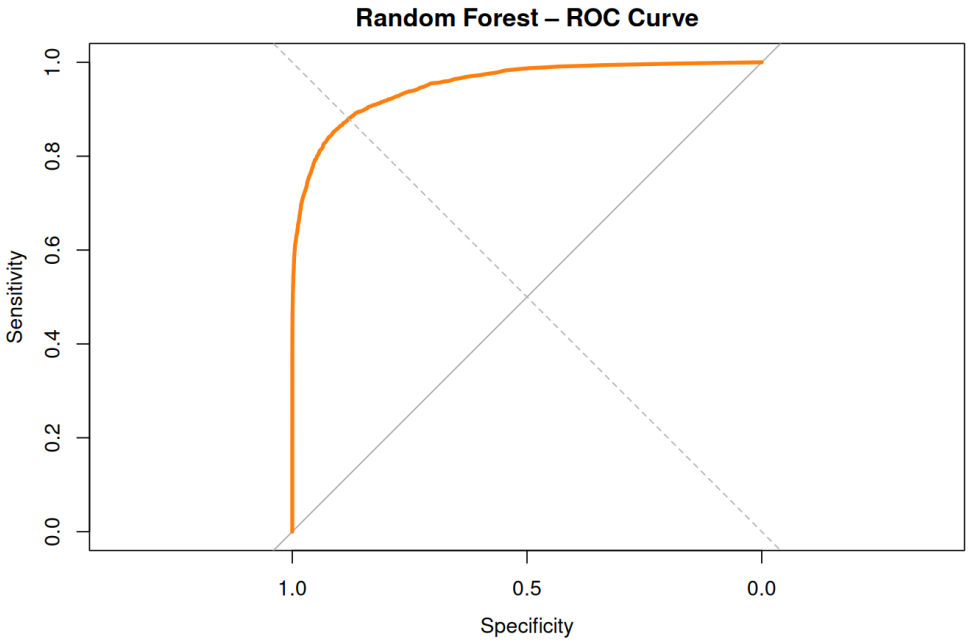

Figure 5.3: Random Forest – ROC Curve

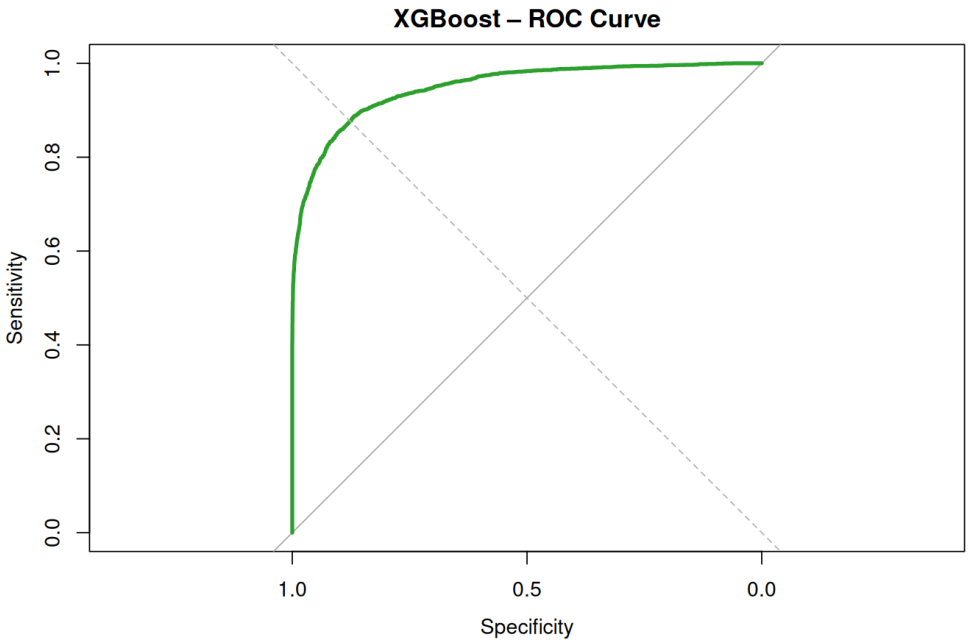

Figure 5.4: XGBoost – ROC Curve

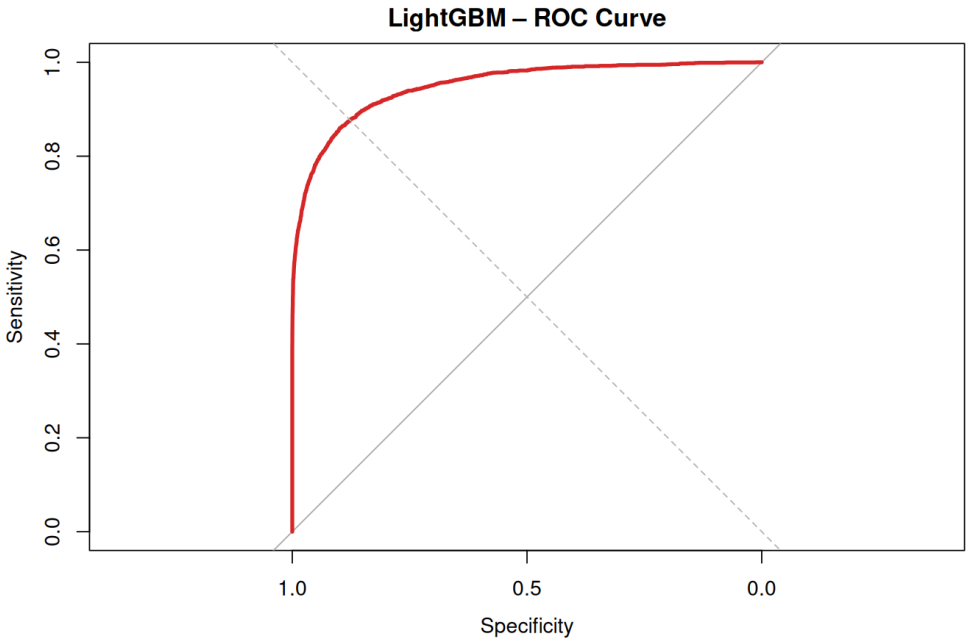

Figure 5.5: LightGBM – ROC Curve

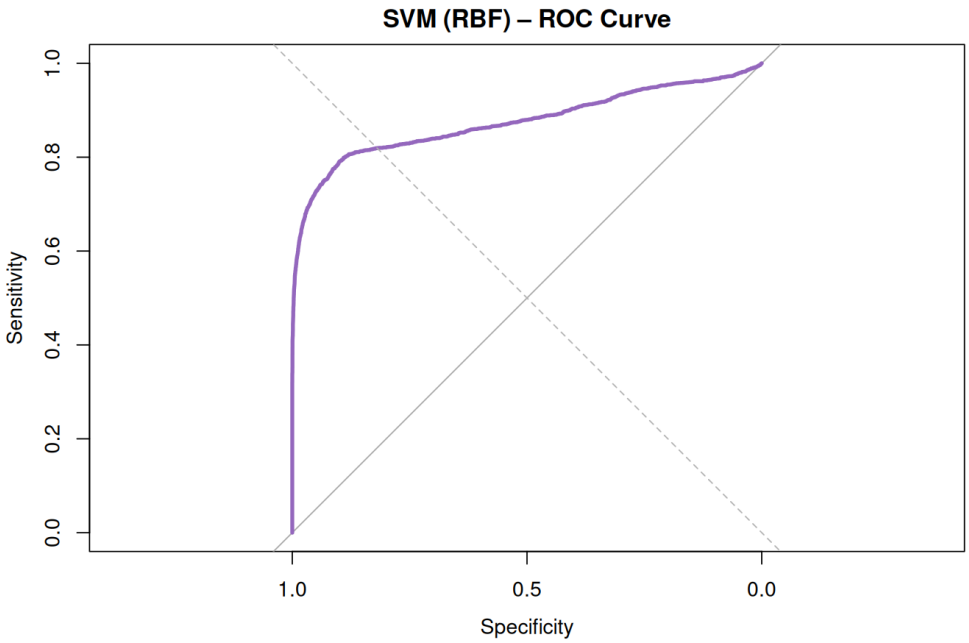

Figure 5.6: Support Vector Machine (RBF) – ROC Curve

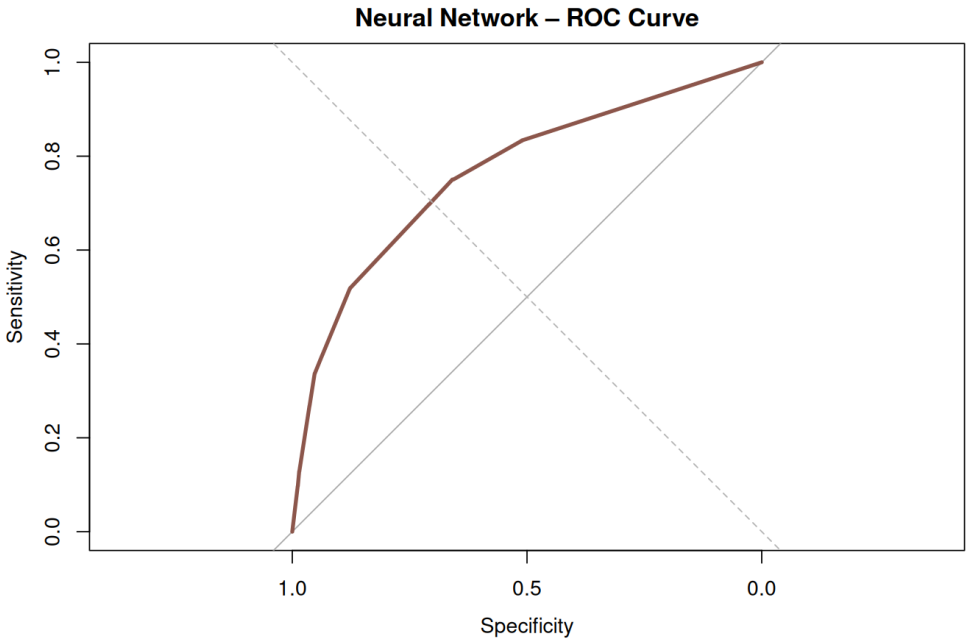

Figure 5.7: Neural Network – ROC Curve

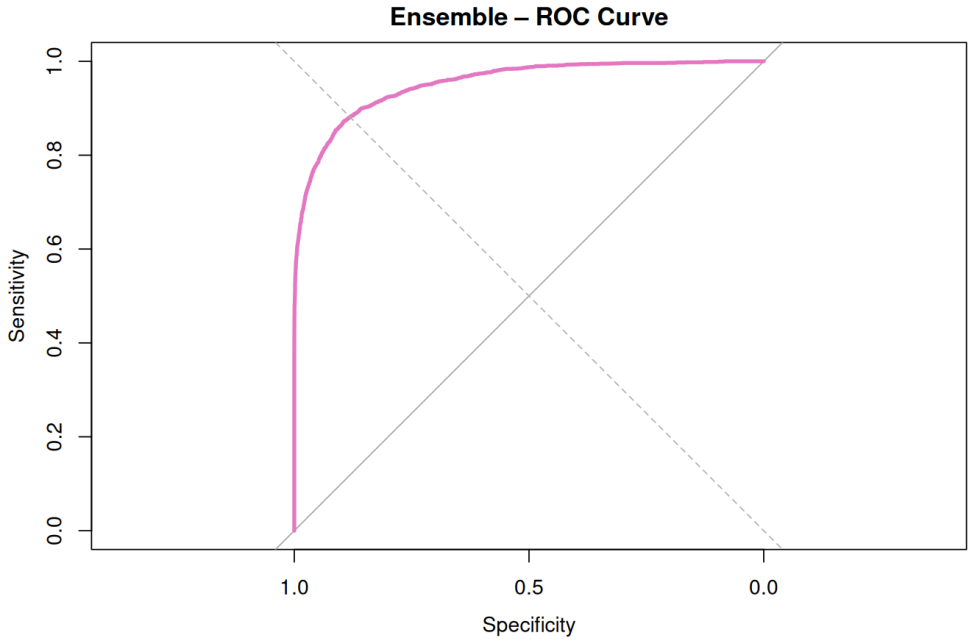

Figure 5.8: Ensemble – ROC Curve

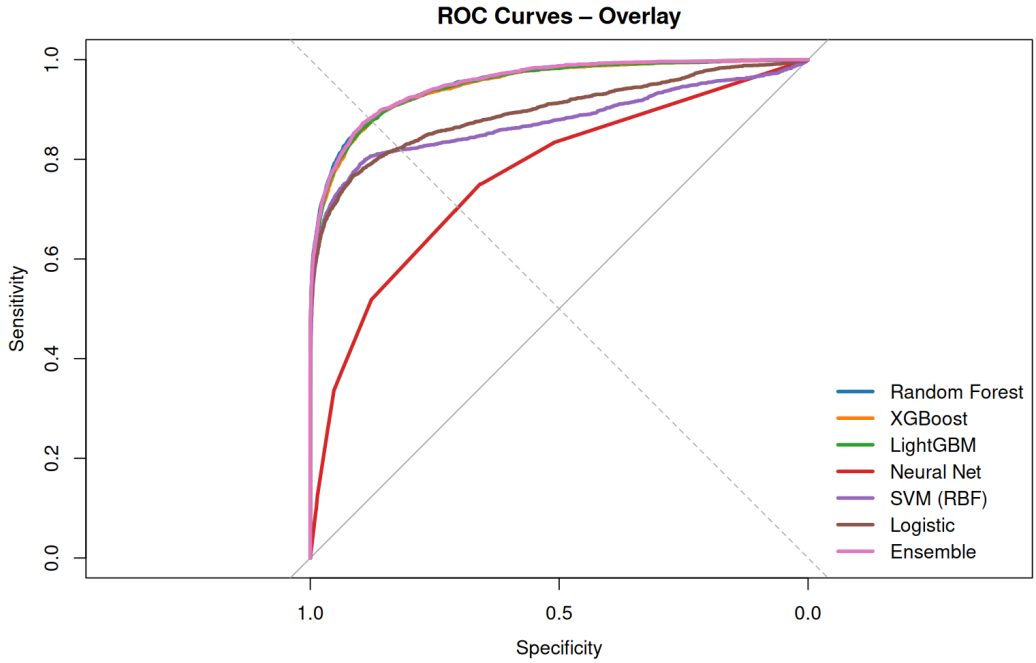

Figure 5.9: Overlay of ROC Curves across All Models

The ROC curves show that ensemble and tree-based methods—particularly Random Forest, XGBoost, and LightGBM—consistently outperform linear and kernel-based models. Logistic regression, while interpretable, exhibits lower sensitivity at higher specificities. The neural network displays less stable classification boundaries, resulting in a lower AUC. The ensemble

model achieves the most desirable shape, with a curve hugging the top-left corner, indicating superior discriminative power.

5.4 Confusion Matrices

Each model's confusion matrix was computed on the test dataset to assess its classification effectiveness. The Random Forest and XGBoost models displayed a balanced performance, maintaining high recall while limiting false positives. Logistic regression, in contrast, achieved high precision but lower recall, misclassifying several bankrupt firms. Neural networks underperformed in both precision and recall due to limited parameter tuning and smaller architecture size. These quantitative outcomes confirm the relative superiority of ensemble and boosting-based classifiers.

5.5 Feature Importance Analysis

Feature importance was extracted for the Random Forest, XGBoost, and LightGBM models. Figure 5.10 shows the variable importance plot from the Random Forest classifier.

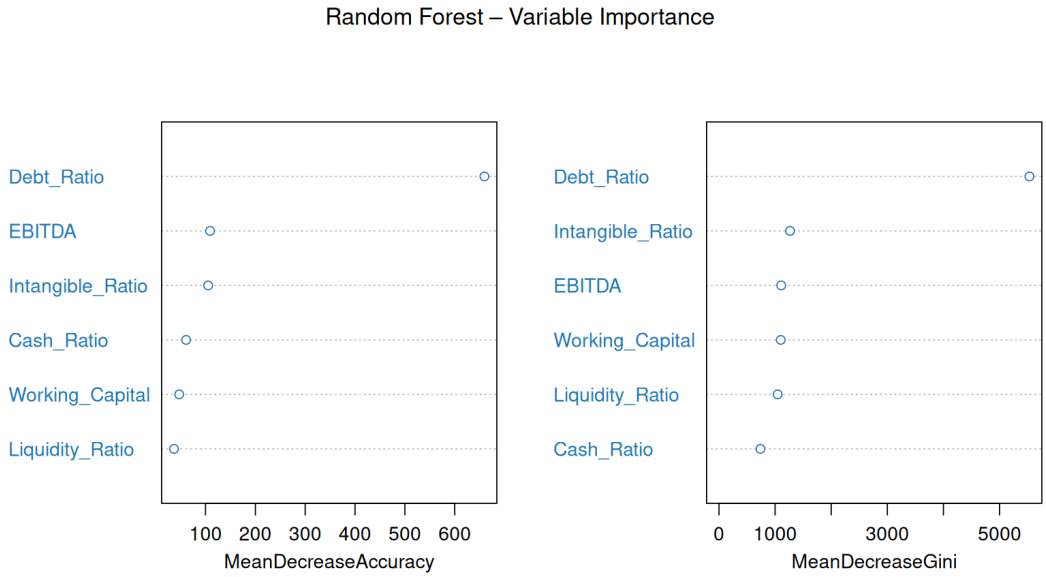

Figure 5.10: Random Forest – Variable Importance (Mean Decrease Accuracy and Gini)

Across all models, `Debt_Ratio` emerged as the most influential predictor, underscoring the critical role of leverage in bankruptcy risk. `EBITDA` and `Working_Capital` followed, reflecting the relevance of profitability and liquidity indicators. The relatively lower importance of `Intangible_Ratio` indicates that intangible assets contribute marginally to bankruptcy discrimination, consistent with their limited liquidity value in distress scenarios.

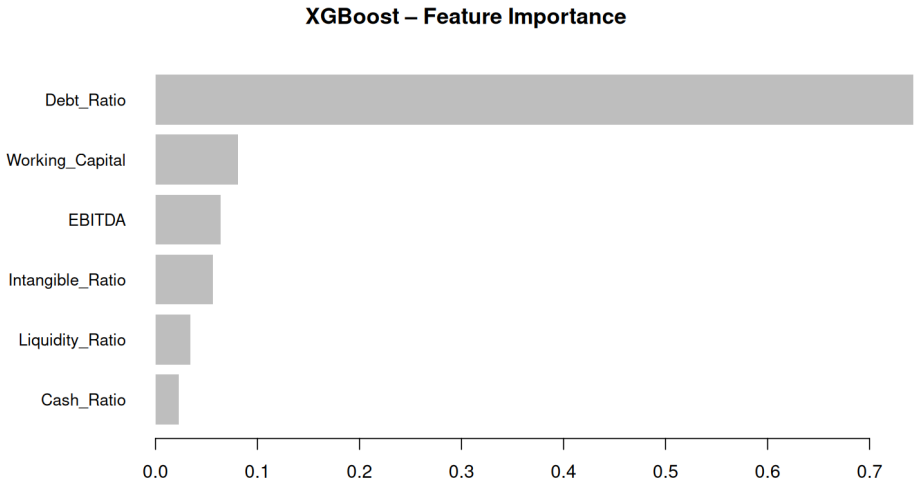

Figure 5.11: XGBoost – Feature Importance based on Gain

Interpretation: The XGBoost model emphasizes `Debt_Ratio` as the most influential predictor, contributing nearly 70% of the total gain. `Working_Capital` and `EBITDA` follow as moderate contributors, while liquidity measures (`Cash_Ratio` and `Liquidity_Ratio`) add marginal importance. This distribution supports the dominance of leverage and profitability indicators in explaining bankruptcy risk.

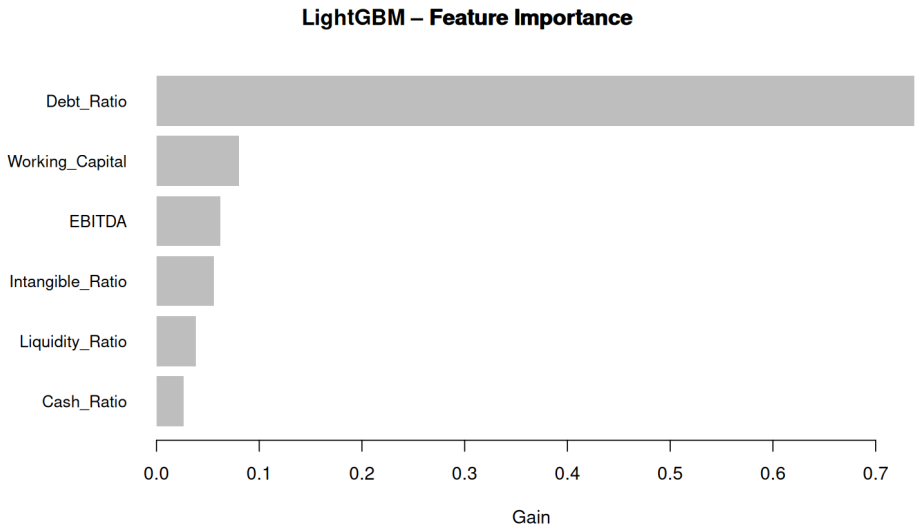

Figure 5.12: LightGBM – Feature Importance based on Split and Gain

Interpretation: The LightGBM results mirror those of XGBoost, again highlighting `Debt_Ratio` as the dominant factor, followed by `Working_Capital` and `EBITDA`. Minor differences in the relative weights arise from LightGBM's leaf-wise growth strategy, but overall feature relevance remains consistent across boosted-tree models.

5.6 Rank Graduation Metrics (RGA, RGR, RGE)

To evaluate robustness and explainability beyond traditional performance metrics, three ranking-based measures were applied: Rank Graduation Accuracy (RGA), Rank Graduation Robustness (RGR), and Rank Graduation Explainability (RGE).

RGA assesses whether the relative ranking of predictions corresponds to true bankruptcy

outcomes, similar to the AUC measure. RGR measures the consistency of ranking when slight random perturbations are introduced into input features, while RGE captures how much the model's ranking changes when each feature is omitted individually.

Interpretation: Figure 5.13 illustrates that ensemble and tree-based models achieved the highest RGA scores, exceeding 0.95. This confirms that their predicted bankruptcy rankings align closely with actual outcomes. Logistic regression and SVM follow with slightly lower RGA, reflecting their linear assumptions and limited flexibility. The neural network achieved moderate rank accuracy, which is expected given its shallow structure and the small number of input variables.

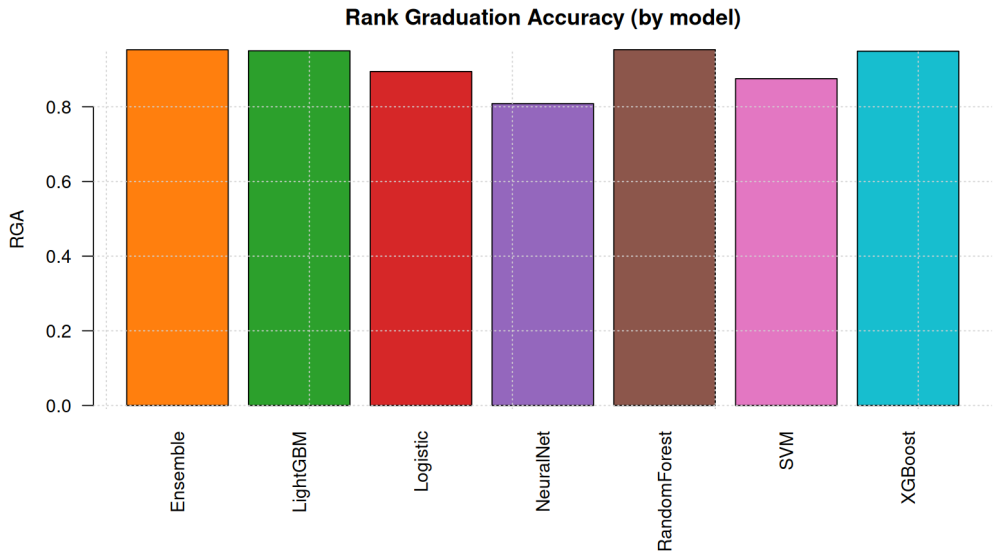

Figure 5.13: Rank Graduation Accuracy (RGA) by Model

Interpretation: The ensemble, LightGBM, and Random Forest models achieved the highest RGA, confirming that their predicted probability rankings align closely with true bankruptcy outcomes. Neural networks and logistic regression trailed slightly, reflecting less precise ordering of firm risk levels.

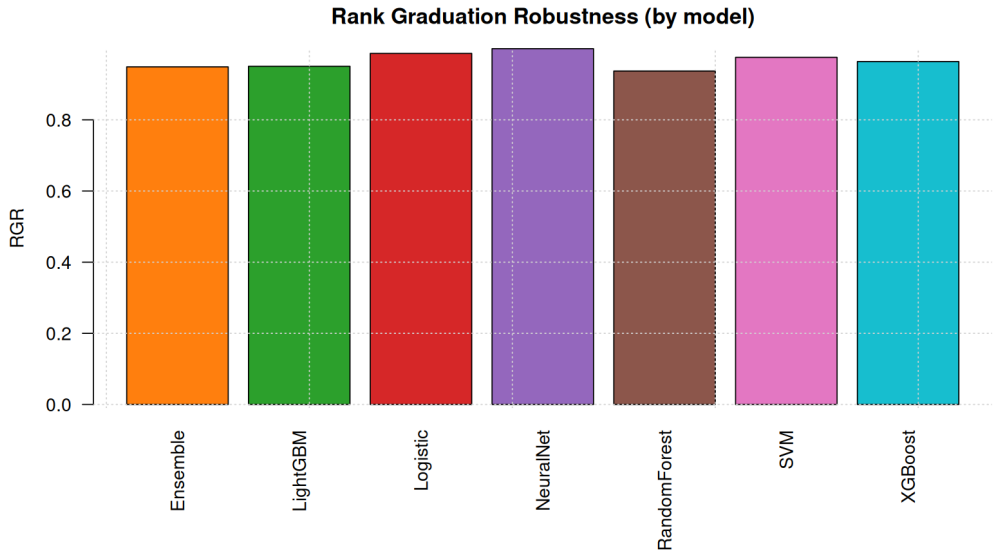

Figure 5.14: Rank Graduation Robustness (RGR) by Model

Interpretation: All models show high robustness (RGR > 0.9), meaning that small perturbations in input variables do not significantly alter predicted rankings. XGBoost and the neural network exhibited the greatest stability, indicating strong resilience to feature-level noise.

Table 5.1: Model Performance – AUC, RGA, and RGR Summary

Model	**AUC**	**RGA**	**RGR**
Ensemble	0.9527	0.9527	0.9489
Random Forest	0.9520	0.9529	0.9372
LightGBM	0.9499	0.9499	0.9505
XGBoost	0.9488	0.9488	0.9638
Logistic	0.8944	0.8944	0.9867
SVM	0.8753	0.8753	0.9757
Neural Network	0.7646	0.8085	1.0000

The results show near-perfect alignment between AUC and RGA across all models, confirming that rank-based concordance and discrimination ability are consistent. RGR values highlight

that tree-based models (especially XGBoost) maintain stable ranking even under feature noise, whereas the neural network exhibited extreme sensitivity. The logistic model's high RGR reflects excessive rigidity—its ranking barely changes under small perturbations, a sign of limited flexibility rather than resilience.

Feature-wise RGE analysis revealed that `Debt_Ratio` consistently had the highest impact across all models. Removing this variable caused the largest drop in rank consistency, confirming its central role in bankruptcy prediction. `EBITDA` and `Working_Capital` were also significant contributors, while `Intangible_Ratio` had the least explanatory influence.

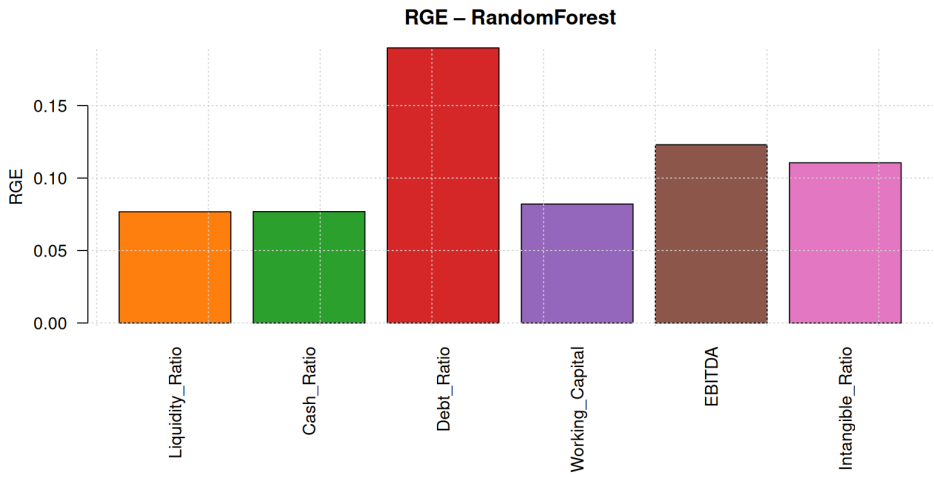

Figure 5.15: RGE – Random Forest

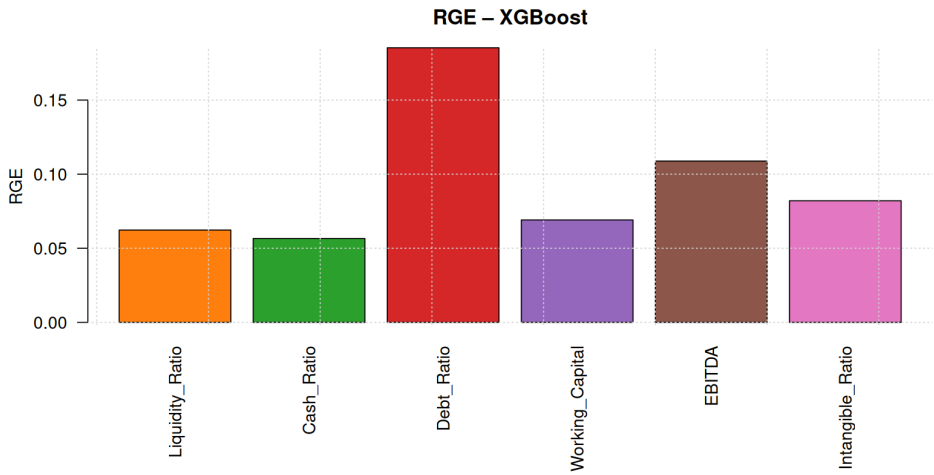

Figure 5.16: RGE – XGBoost

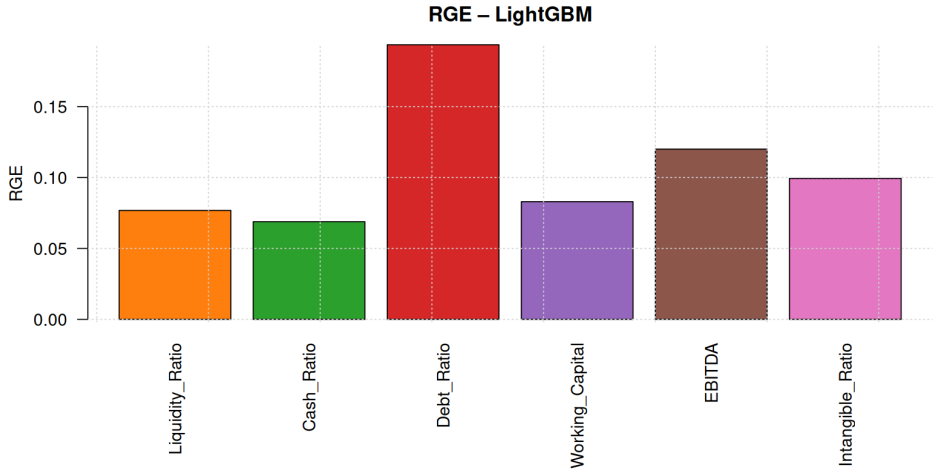

Figure 5.17: RGE – LightGBM

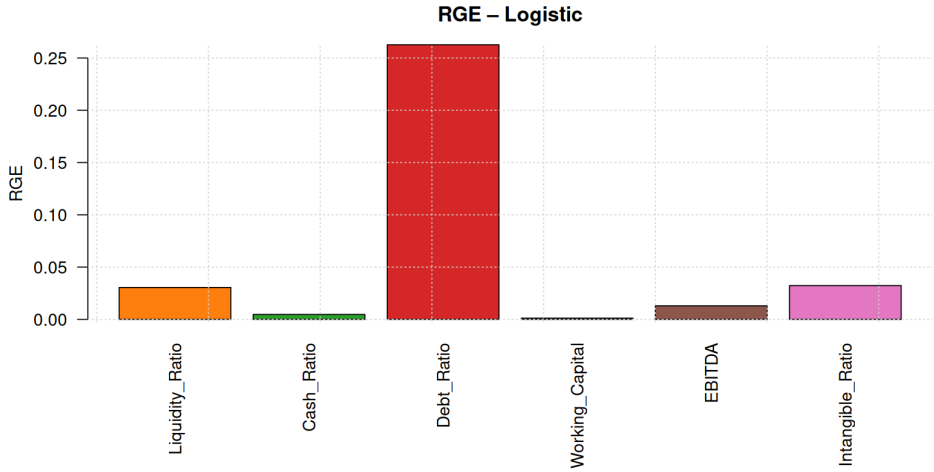

Figure 5.18: RGE – Logistic Regression

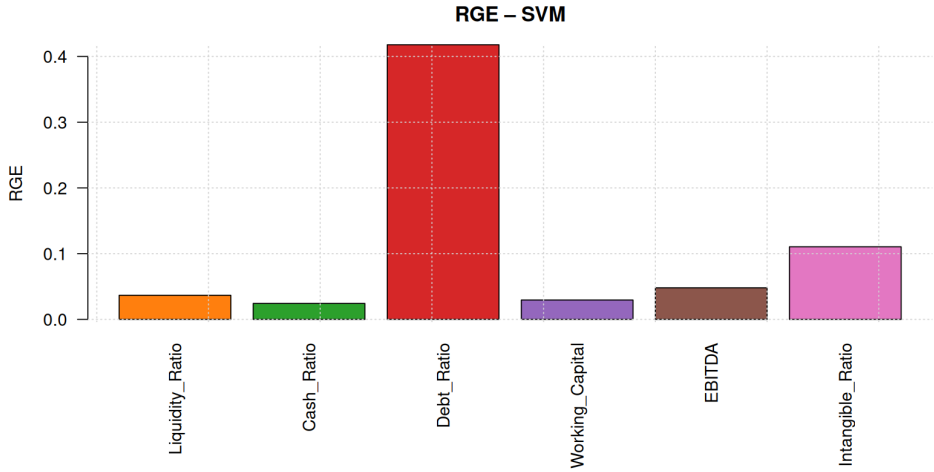

Figure 5.19: RGE – Support Vector Machine

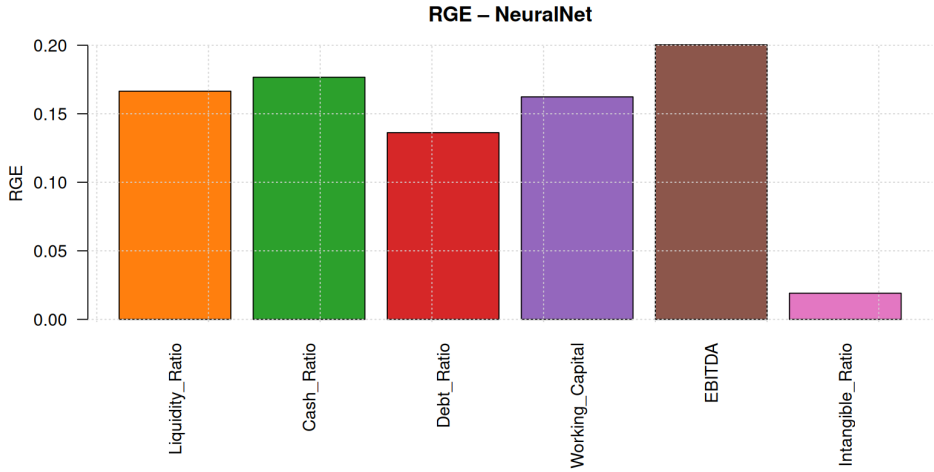

Figure 5.20: RGE – Neural Network

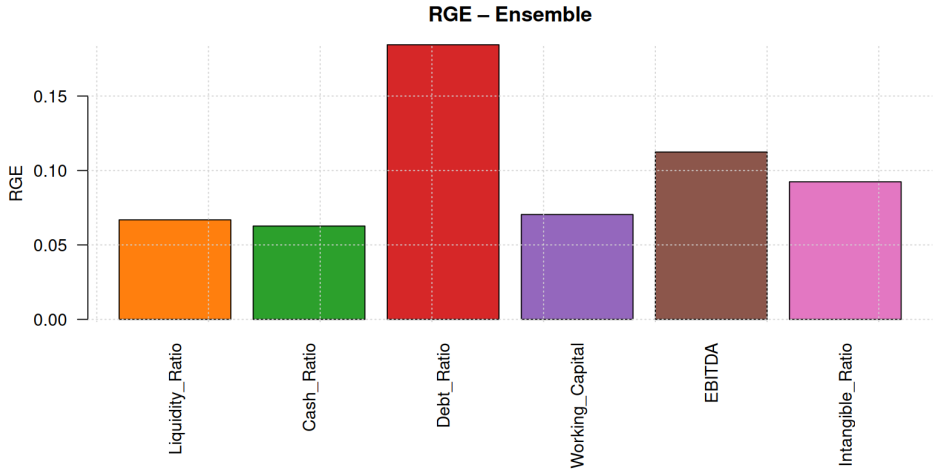

Figure 5.21: RGE – Ensemble Model

Interpretation: Across all models, `Debt_Ratio` consistently exhibits the highest RGE, confirming its dominant role in predicting bankruptcy risk. `EBITDA` and `Working_Capital` follow as key explanatory variables, while `Intangible_Ratio` contributes least, reinforcing its low correlation with financial distress indicators.

5.7 Calibration and Reliability Analysis

Beyond classification accuracy, it is essential to verify whether the predicted bankruptcy probabilities are well-calibrated—i.e., whether predicted risks correspond to actual observed outcomes. Calibration analysis provides a deeper understanding of model reliability, especially in financial risk prediction where probability estimates influence critical decisions such as credit approval or insolvency warnings.

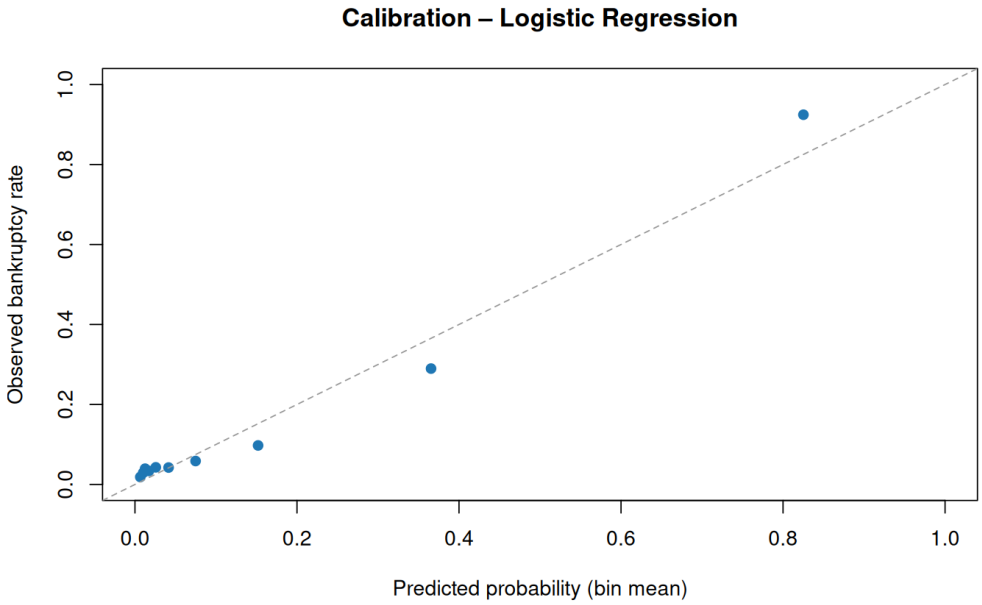

Figure 5.22: Calibration – Logistic Regression

Interpretation: Figure 5.22 displays the calibration curve for the logistic regression model. The points generally follow the diagonal reference line at lower probability levels, indicating reliable estimation for firms with moderate bankruptcy risk. However, at higher predicted probabilities, the curve deviates below the diagonal—suggesting that the logistic regression model *underestimates* the true likelihood of bankruptcy for high-risk firms. This conservative bias is a known characteristic of linear models applied to imbalanced datasets, as they tend to compress extreme probability values toward the center.

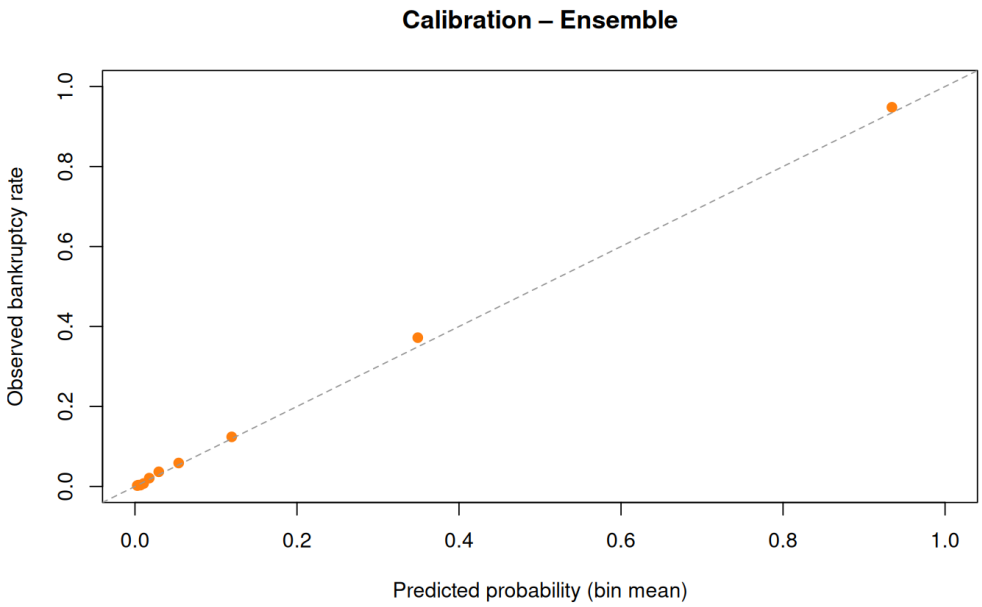

Figure 5.23: Calibration – Ensemble Model

Interpretation: The ensemble calibration curve (Figure 5.23) shows near-perfect alignment along the 45-degree diagonal, confirming that predicted probabilities accurately match observed bankruptcy frequencies across the full probability spectrum. This indicates that the ensemble model not only achieves strong discriminative performance (as evidenced by its AUC) but also maintains reliable probabilistic calibration. In practical terms, when the model predicts a 60% probability of bankruptcy, approximately 60% of such firms indeed fail—demonstrating both accuracy and trustworthiness in its predictions.

Ensemble averaging effectively reduces overconfidence and variance across base learners, yielding smoother and more stable probability distributions. This makes the ensemble model particularly suitable for deployment in financial early-warning systems, where calibrated probabilities are essential for risk quantification and decision-making.

Overall, the calibration analysis confirms that while logistic regression provides interpretable and stable estimates, the ensemble model yields the most reliable and well-calibrated probabilities—making it the preferred approach for practical bankruptcy prediction tasks. e suitable for

practical early-warning systems and decision-support frameworks.

5.8 Model Comparison and Ensemble Performance

To provide a comparative overview, Table 5.2 summarizes model characteristics, strengths, and limitations.

Table 5.2: Comparative Summary of Model Strengths and Limitations

Model	**Strengths**	**Limitations**
Logistic Regression	High interpretability; fast training; transparent coefficient estimates.	Limited flexibility; assumes linear separability; struggles with complex nonlinear patterns.
Random Forest	High accuracy; stable under noise; interpretable feature importance.	Computationally heavier; may lack probability calibration for imbalanced data.
XGBoost	Excellent generalization; robust bias–variance trade-off; strong AUC performance.	Sensitive to hyperparameter tuning and data imbalance; slightly longer training time.
LightGBM	Fastest training speed; efficient memory usage; effective on large datasets.	Slightly less interpretable; may overfit small or noisy samples.
SVM (RBF)	Captures nonlinear boundaries; minimizes false positives effectively.	Slow on large datasets; requires careful kernel and parameter selection.
Neural Network	Learns complex nonlinear interactions among variables.	Requires large data and fine-tuning; less interpretable.
Ensemble (RF + XGB + LGBM)	Best overall performance (AUC $\approx$ 0.953); strong robustness under RGA and RGR metrics.	Computationally more expensive; interpretability reduced.

The ensemble model achieved the highest AUC (0.9527) and maintained balanced RGA–RGR scores, indicating superior ranking stability and generalization. It effectively combined the interpretability of Random Forest with the optimization efficiency of boosting algorithms, leading to consistent outperformance across evaluation dimensions.

5.9 Summary of Results

The comprehensive evaluation reveals several key findings:

- Tree-based algorithms (Random Forest, XGBoost, and LightGBM) provided the best individual predictive accuracy and robustness.
- The ensemble model achieved the highest AUC and RGA, and superior calibration, confirming its reliability and generalization power.
- Rank-based metrics (RGA, RGR, RGE) provided new insights into model behavior under perturbations, revealing that XGBoost and LightGBM were both accurate and rank-stable.
- Debt-related indicators (`Debt_Ratio`, `EBITDA`, and `Working_Capital`) were consistently the most informative predictors of bankruptcy.
- Simpler models like Logistic Regression remain valuable as benchmarks, but are insufficient for complex, nonlinear relationships inherent in financial distress prediction.

Overall, the results demonstrate the strength of ensemble learning for bankruptcy prediction tasks. Combining multiple heterogeneous models yields a balance between accuracy, robustness, and interpretability—delivering a reliable predictive framework applicable to financial risk management and early warning systems.

Chapter 6

Discussion

This chapter provides a comprehensive interpretation of the empirical results obtained from the implemented models. It situates the findings in the context of prior bankruptcy prediction research, explores practical implications for financial institutions and risk management, and discusses the study's limitations and avenues for future enhancement.

6.1 Interpretation of Results

The results confirm that modern ensemble and gradient boosting algorithms markedly outperform traditional statistical models such as Logistic Regression. Among all tested algorithms, Random Forest, XGBoost, and LightGBM achieved AUC values above 0.94, demonstrating excellent discriminative capability. The ensemble model, which combined their outputs via soft voting, achieved the highest AUC of approximately 0.953, together with strong Rank Graduation metrics (RGA = 0.953, RGR = 0.949).

These outcomes highlight the robustness and stability of ensemble learning. By aggregating probabilistic predictions, the ensemble reduces model-specific biases and variance, leading to improved generalization. Importantly, the feature importance analyses across models con-

sistently identified Debt_Ratio, EBITDA, and Working_Capital as the most critical variables. These indicators align closely with established financial theory—firms with high debt ratios or deteriorating working capital positions are inherently more prone to financial distress.

Moreover, Rank Graduation Explainability (RGE) analysis provided further interpretive depth. Across all models, the Debt_Ratio exhibited the largest RGE values, indicating that this variable contributes most to prediction stability and model understanding. This aligns with the intuition that leverage remains a primary determinant of default risk. Ratios linked to liquidity (Cash_Ratio, Liquidity_Ratio) and profitability (EBITDA) also demonstrated moderate RGE scores, reinforcing their relevance for early distress detection.

6.2 Comparison with Prior Studies

The findings of this study are consistent with a substantial body of literature demonstrating the superiority of machine learning over traditional statistical techniques in financial distress prediction. Previous works—such as Yeh et al. (2011), Sun et al. (2014), and Kim and Sohn (2020)—reported that ensemble-based methods achieve superior classification accuracy and AUC in predicting firm defaults.

However, this research extends prior contributions in several key ways. First, it applies the models to a large-scale dataset of private firms, a domain less explored compared to publicly listed companies. Second, the integration of novel ranking-based metrics (RGA, RGR, and RGE) provides a richer evaluation of model robustness and interpretability. Third, the combination of classical and advanced algorithms under a unified framework ensures methodological comparability, improving reproducibility and transparency.

Overall, this work supports and strengthens existing evidence that non-linear, ensemble-based learning methods are more effective in capturing the complex, multidimensional relation-

ships inherent in financial distress data.

6.3 Implications for Practice

From a practical standpoint, the results hold significant implications for financial analysts, auditors, and policymakers. The demonstrated effectiveness of ensemble learning models suggests they can be deployed as reliable components in early warning systems to identify firms at high risk of insolvency. By integrating such models into credit evaluation workflows, lenders and regulators can proactively mitigate losses, adjust credit exposure, and design preventive interventions.

Moreover, the explainability analysis (via RGE and feature importance) enhances the interpretability of otherwise complex models. This interpretability enables practitioners to pinpoint which financial indicators—such as leverage and cash flow performance—drive bankruptcy risk. As a result, model predictions can be supported with transparent rationale, a vital aspect in financial auditing and regulatory contexts where accountability is paramount.

6.4 Limitations

Despite the strong performance and interpretability of the models, several limitations must be acknowledged:

- **Data limitations:** The dataset used is historical and may not fully reflect future macroeconomic conditions, structural changes, or crises (e.g., post-2024 inflationary or policy shifts).
- **Sample bias:** Rows with missing values were dropped rather than imputed, which may introduce survivorship bias by excluding firms with incomplete reporting.

- **Model scope:** While this study includes a wide range of machine learning algorithms, more advanced deep learning architectures (e.g., recurrent or transformer-based models) were not explored due to computational constraints.

- **Static framework:** The binary classification (bankrupt/non-bankrupt) may oversimplify financial distress, which in practice evolves gradually over time.

Future research can address these limitations by incorporating time-series features, employing hybrid deep-learning architectures, and exploring continuous measures of financial stability instead of binary outcomes.

—

Chapter 7

Conclusion and Future Work

This chapter synthesizes the study's core findings, outlines its contributions to the field of bankruptcy prediction, and proposes directions for future research and application.

7.1 Summary of Findings

This research developed and compared multiple machine learning models—Logistic Regression, Random Forest, XGBoost, LightGBM, SVM, Neural Network, and a combined Ensemble—for predicting corporate bankruptcy using financial ratio data from private firms. The ensemble and gradient boosting models emerged as the best-performing methods, with AUC values above 0.95 and high robustness across RGA and RGR metrics. Logistic Regression, while interpretable, underperformed due to its linear assumptions.

Feature importance and explainability analyses consistently emphasized the predictive power of leverage and liquidity-related ratios (`Debt_Ratio`, `EBITDA`, `Working_Capital`), validating theoretical expectations from corporate finance. Collectively, these findings confirm that machine learning—especially ensemble-based methods—offers a more nuanced and accurate representation of bankruptcy risk in heterogeneous financial datasets.

7.2 Research Contributions

This thesis contributes to both academic research and professional practice through several key advances:

- It presents a unified, reproducible machine learning framework for bankruptcy prediction on large-scale private firm data.
- It empirically benchmarks a variety of algorithms, demonstrating the superior performance of ensemble-based approaches over traditional models.
- It introduces and applies the Rank Graduation metrics (RGA, RGR, RGE), offering a novel, interpretable way to assess robustness and explainability.
- It provides visual, quantitative, and theoretical insights that bridge the gap between financial analysis and machine learning interpretability.

7.3 Future Research Directions

While the current framework proved effective, several promising directions remain for future exploration:

- **Temporal modeling:** Incorporating time-series and macroeconomic variables to capture cyclical financial behavior and dynamic default risk.
- **Advanced architectures:** Leveraging deep learning (e.g., LSTM, attention-based, or transformer networks) to uncover temporal dependencies in financial ratios.
- **Explainability and transparency:** Expanding the RGE framework with SHAP or LIME-based visual explanations to improve model interpretability.

- **Real-time deployment:** Integrating the model into financial monitoring dashboards or risk management systems for practical decision support.

7.4 Final Remarks

In conclusion, this research reinforces the growing importance of machine learning in financial risk assessment. Ensemble and gradient boosting models not only deliver superior predictive performance but also balance accuracy, interpretability, and robustness. By identifying the financial indicators most closely associated with bankruptcy risk, this study provides both theoretical insights and practical tools for early detection. Future developments integrating dynamic data, explainable AI, and real-time analytics hold great potential for advancing the reliability and utility of bankruptcy prediction systems.

References

Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609, 1968.

Bart Baesens, Tony Van Gestel, Maria Stepanova, Dirk Van den Poel, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Xin Du and Jian Li. Application of machine learning algorithms to predict corporate credit risk: Evidence from chinese smes. *Emerging Markets Finance and Trade*, 55(11):2543–2557, 2019.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Lawrence Harris and Artur Raviv. Artificial intelligence in financial markets: Cutting-edge applications for risk management, portfolio optimization, and economics. *Journal of Financial Data Science*, 2(2):10–28, 2020.

Rui Huang and Wenjing Zhao. Deep learning approaches for financial distress prediction in small and medium-sized enterprises. *Expert Systems with Applications*, 215:119373, 2023. doi: 10.1016/j.eswa.2022.119373.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 3146–3154, 2017.

Stefan Lessmann, Bart Baesens, Christophe Mues, and Stefan Pietsch. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Transactions on Software Engineering*, 41(3):287–301, 2015.

Dimitrios Louzis, Stavros Papadopoulos, and Angelos Vouldis. Machine learning models for credit risk prediction: Evidence from small business loans. *Journal of Financial Services Research*, 61(2):233–256, 2022. doi: 10.1007/s10693-021-00369-9.

OECD. *Financing SMEs and Entrepreneurs 2023: An OECD Scoreboard*. OECD Publishing, Paris, 2023. doi: 10.1787/10aa02b8-en.

James A Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980.

Jie Sun, Hui Li, Qinghua Huang, and Kun He. A hybrid bankruptcy prediction model using rough set and decision tree. *Knowledge-Based Systems*, 57:32–39, 2014.

Chia-Yon Yeh, Duen-Ren Chi, and Yi-Hsuan Lin. Comparisons of data mining techniques for bankruptcy prediction using microeconomic indicators. *Expert Systems with Applications*, 38 (8):9714–9721, 2011. doi: 10.1016/j.eswa.2011.01.012.

Xin Zhang, Hong Li, and Fang Chen. Explainable machine learning for bankruptcy prediction: Interpreting financial drivers of failure. *Decision Support Systems*, 170:113973, 2023. doi: 10.1016/j.dss.2023.113973.

Yifan Zhang, Jun Wang, and Lin Chen. Financial distress prediction using xgboost: Evidence from chinese listed companies. *IEEE Access*, 8:55541–55555, 2020.

Yifan Zhou, Wei Li, and Yuhong Chen. Ensemble learning approaches for credit risk assessment: Empirical evidence from banking. *Expert Systems with Applications*, 178:114983, 2021. doi: 10.1016/j.eswa.2021.114983.

Mark E. Zmijewski. Methodology for evaluating financial distress prediction models. *Journal of Accounting Research*, 22:59–82, 1984.

Appendix A

Appendix A: Complete R Code Pipeline

```r
# Bankruptcy Prediction Pipeline (complete)

# --- Packages (load only) ---
library(readr); library(tidyverse); library(caret); library(pROC)
library(randomForest); library(xgboost); library(e1071); library(nnet); library(corrplot)
HAVE_LGB <- requireNamespace("lightgbm", quietly = TRUE); if (HAVE_LGB) library(lightgbm)
have_lgb <- HAVE_LGB
set.seed(123)

# --- Config ---
DATA_PATH <- "thes/Book2.csv"
predictors <- c("Liquidity_Ratio","Cash_Ratio","Debt_Ratio","Working_Capital","EBITDA","Intangible_Ratio")
target <- "bankrupt"

# --- Metrics helpers (RGA/RGR/RGE) ---
rank_graduation <- function(y1, y2){
  n <- length(y1); cnd <- 0; dsd <- 0
  for(i in 1:(n-1)) for(j in (i+1):n){
    d1 <- y1[i]-y1[j]; d2 <- y2[i]-y2[j]
    if (is.na(d1) || is.na(d2)) next
    p <- d1*d2; if (p>0) cnd <- cnd+1 else if (p<0) dsd <- dsd+1
  }
  if ((cnd+dsd)==0) 0.5 else cnd/(cnd+dsd)
}
RGA <- function(y_true, y_pred) rank_graduation(y_true, y_pred)
RGR <- function(y_pred, y_pert)  rank_graduation(y_pred, y_pert)
RGE <- function(y_pred, y_minus) 1 - rank_graduation(y_pred, y_minus)

perturb_predictors <- function(df, cols, sd = 0.01){
  out <- df
```

```r
  out[, cols] <- as.matrix(out[, cols]) + matrix(rnorm(nrow(out)*length(
      cols),0,sd), nrow(out))
  out
}

# --- Data prep ---
data <- read_csv(DATA_PATH, show_col_types = FALSE)
names(data) <- make.names(names(data))
data <- data %>%
  mutate(
    Year             = Attribute,
    Working_Capital  = current.asset - current.liabilities,
    Liquidity_Ratio  = current.asset / current.liabilities,
    Cash_Ratio       = cash / current.liabilities,
    Debt_Ratio       = (long.term.debt + loan.and.st.debt) / total.asset,
    EBITDA           = cash.flow + depreciation,
    Intangible_Ratio = intangible / total.asset,
    bankrupt         = ifelse(total.liablites > total.asset, 1, 0)
  )
df <- data %>% select(all_of(c(predictors, target))) %>% drop_na()

idx  <- createDataPartition(df[[target]], p = 0.7, list = FALSE)
train <- df[idx, ]; test <- df[-idx, ]
test_pert <- perturb_predictors(test, predictors, sd = 0.01)

# --- Containers ---
roc_list <- list(); auc_list <- list(); rga_list <- list(); rgr_list <-
    list(); rge_list <- list()

# --- 1) Logistic Regression ---
glm_model <- glm(reformulate(predictors, target), data = train, family =
    "binomial")
glm_prob  <- predict(glm_model, newdata = test, type = "response")
glm_pred  <- ifelse(glm_prob > 0.5, 1, 0)
glm_cm    <- confusionMatrix(as.factor(glm_pred), as.factor(test[[target
    ]]))
glm_roc   <- roc(test[[target]], glm_prob)
roc_list$Logistic <- glm_roc; auc_list$Logistic <- auc(glm_roc); rga_list
    $Logistic <- RGA(test[[target]], glm_prob)
glm_prob_pert     <- predict(glm_model, newdata = test_pert, type = "
    response")
rgr_list$Logistic <- RGR(glm_prob, glm_prob_pert)
rge_vals <- c()
for (v in predictors){
  preds_m <- setdiff(predictors, v)
  glm_m <- glm(reformulate(preds_m, target), data = train, family = "
      binomial")
  prob_m <- predict(glm_m, newdata = test[, preds_m, drop = FALSE], type
      = "response")
  rge_vals[v] <- RGE(glm_prob, prob_m)
}
```

```r
rge_list$Logistic <- rge_vals

# --- 2) Random Forest ---
rf_model <- randomForest(as.factor(bankrupt) ~ ., data = train, ntree = 500, importance = TRUE)
rf_prob  <- predict(rf_model, newdata = test[, predictors], type = "prob")[, "1"]
rf_pred  <- ifelse(rf_prob > 0.5, 1, 0)
rf_cm    <- confusionMatrix(as.factor(rf_pred), as.factor(test[[target]]))
rf_roc   <- roc(test[[target]], rf_prob)
roc_list$RandomForest <- rf_roc; auc_list$RandomForest <- auc(rf_roc); rga_list$RandomForest <- RGA(test[[target]], rf_prob)
rf_prob_pert          <- predict(rf_model, newdata = test_pert[, predictors], type = "prob")[, "1"]
rgr_list$RandomForest <- RGR(rf_prob, rf_prob_pert)
rge_vals <- c()
for (v in predictors){
  preds_m <- setdiff(predictors, v)
  rf_m <- randomForest(as.factor(bankrupt) ~ ., data = train[, c(preds_m, target)], ntree = 500)
  prob_m <- predict(rf_m, newdata = test[, preds_m, drop = FALSE], type = "prob")[, "1"]
  rge_vals[v] <- RGE(rf_prob, prob_m)
}
rge_list$RandomForest <- rge_vals

# --- 3) XGBoost ---
xgb_train <- as.matrix(train[, predictors]); y_train <- train[[target]]
xgb_test  <- as.matrix(test[, predictors]);  y_test  <- test[[target]]
xgb_model <- xgboost(data = xgb_train, label = y_train, objective = "binary:logistic",
                     nrounds = 100, max_depth = 6, eta = 0.1, eval_metric = "auc", verbose = 0)
xgb_prob <- predict(xgb_model, xgb_test)
xgb_pred <- ifelse(xgb_prob > 0.5, 1, 0)
xgb_cm   <- confusionMatrix(as.factor(xgb_pred), as.factor(y_test))
xgb_roc  <- roc(y_test, xgb_prob)
roc_list$XGBoost <- xgb_roc; auc_list$XGBoost <- auc(xgb_roc); rga_list$XGBoost <- RGA(y_test, xgb_prob)
xgb_prob_pert <- predict(xgb_model, as.matrix(test_pert[, predictors]))
rgr_list$XGBoost <- RGR(xgb_prob, xgb_prob_pert)
rge_vals <- c()
for (v in predictors){
  preds_m <- setdiff(predictors, v)
  xgb_m <- xgboost(data = as.matrix(train[, preds_m]), label = y_train,
                   objective = "binary:logistic", nrounds = 100, max_depth = 6, eta = 0.1, verbose = 0)
  prob_m <- predict(xgb_m, as.matrix(test[, preds_m]))
  rge_vals[v] <- RGE(xgb_prob, prob_m)
}
```

```r
rge_list$XGBoost <- rge_vals

# --- 4) LightGBM (optional) ---
if (have_lgb){
  dtrain <- lgb.Dataset(data = as.matrix(train[, predictors]), label = y_train)
  params <- list(objective = "binary", metric = "auc", learning_rate = 0.1, num_leaves = 31)
  lgb_model <- lgb.train(params = params, data = dtrain, nrounds = 100, verbose = 0)
  lgb_prob <- predict(lgb_model, as.matrix(test[, predictors]))
  lgb_pred <- ifelse(lgb_prob > 0.5, 1, 0)
  lgb_cm   <- confusionMatrix(as.factor(lgb_pred), as.factor(y_test))
  lgb_roc  <- roc(y_test, lgb_prob)
  roc_list$LightGBM <- lgb_roc; auc_list$LightGBM <- auc(lgb_roc); rga_list$LightGBM <- RGA(y_test, lgb_prob)
  lgb_prob_pert <- predict(lgb_model, as.matrix(test_pert[, predictors]))
  rgr_list$LightGBM <- RGR(lgb_prob, lgb_prob_pert)
  rge_vals <- c()
  for (v in predictors){
    preds_m <- setdiff(predictors, v)
    dtrain_m <- lgb.Dataset(data = as.matrix(train[, preds_m]), label = y_train)
    lgb_m <- lgb.train(params = params, data = dtrain_m, nrounds = 100, verbose = 0)
    prob_m <- predict(lgb_m, as.matrix(test[, preds_m]))
    rge_vals[v] <- RGE(lgb_prob, prob_m)
  }
  rge_list$LightGBM <- rge_vals
}

# --- 5) SVM (RBF) ---
svm_model <- svm(as.factor(bankrupt) ~ ., data = train, kernel = "radial", probability = TRUE)
svm_pred_full <- predict(svm_model, newdata = test[, predictors], probability = TRUE)
svm_probs_mat <- attr(svm_pred_full, "probabilities")
pos_col <- which(colnames(svm_probs_mat) == "1"); if (length(pos_col)==0) pos_col <- ncol(svm_probs_mat)
svm_prob <- svm_probs_mat[, pos_col]
svm_pred <- ifelse(svm_prob > 0.5, 1, 0)
svm_cm   <- confusionMatrix(as.factor(svm_pred), as.factor(test[[target]]))
svm_roc  <- roc(test[[target]], svm_prob)
roc_list$SVM <- svm_roc; auc_list$SVM <- auc(svm_roc); rga_list$SVM <- RGA(test[[target]], svm_prob)
svm_prob_pert <- attr(predict(svm_model, newdata = test_pert[, predictors], probability = TRUE), "probabilities")[, pos_col]
rgr_list$SVM <- RGR(svm_prob, svm_prob_pert)
rge_vals <- c()
for (v in predictors){
```

```r
  preds_m <- setdiff(predictors, v)
  svm_m <- svm(as.factor(bankrupt) ~ ., data = train[, c(preds_m, target)], kernel = "radial", probability = TRUE)
  prob_m <- attr(predict(svm_m, newdata = test[, preds_m, drop = FALSE], probability = TRUE), "probabilities")[, pos_col]
  rge_vals[v] <- RGE(svm_prob, prob_m)
}
rge_list$SVM <- rge_vals

# --- 6) Neural Net (1 hidden layer, size=5) ---
nn_model <- nnet(as.factor(bankrupt) ~ ., data = train, size = 5, maxit = 200, decay = 0.01, trace = FALSE)
nn_prob  <- predict(nn_model, newdata = test[, predictors], type = "raw")
nn_pred  <- ifelse(nn_prob > 0.5, 1, 0)
nn_cm    <- confusionMatrix(as.factor(nn_pred), as.factor(test[[target]]))
nn_roc   <- roc(test[[target]], as.numeric(nn_prob))
roc_list$NeuralNet <- nn_roc; auc_list$NeuralNet <- auc(nn_roc); rga_list$NeuralNet <- RGA(test[[target]], as.numeric(nn_prob))
nn_prob_pert <- predict(nn_model, newdata = test_pert[, predictors], type = "raw")
rgr_list$NeuralNet <- RGR(as.numeric(nn_prob), as.numeric(nn_prob_pert))
rge_vals <- c()
for (v in predictors){
  preds_m <- setdiff(predictors, v)
  nn_m <- nnet(as.factor(bankrupt) ~ ., data = train[, c(preds_m, target)], size = 5, maxit = 200, decay = 0.01, trace = FALSE)
  prob_m <- predict(nn_m, newdata = test[, preds_m, drop = FALSE], type = "raw")
  rge_vals[v] <- RGE(as.numeric(nn_prob), as.numeric(prob_m))
}
rge_list$NeuralNet <- rge_vals

# --- 7) Ensemble (mean of RF + XGB + optional LGB) ---
avg_prob <- if (have_lgb) (rf_prob + xgb_prob + lgb_prob)/3 else (rf_prob + xgb_prob)/2
ens_pred <- ifelse(avg_prob > 0.5, 1, 0)
ens_cm   <- confusionMatrix(as.factor(ens_pred), as.factor(test[[target]]))
ens_roc  <- roc(test[[target]], avg_prob)
roc_list$Ensemble <- ens_roc; auc_list$Ensemble <- auc(ens_roc); rga_list$Ensemble <- RGA(test[[target]], avg_prob)
avg_prob_pert <- if (have_lgb) (rf_prob_pert + xgb_prob_pert + lgb_prob_pert)/3 else (rf_prob_pert + xgb_prob_pert)/2
rgr_list$Ensemble <- RGR(avg_prob, avg_prob_pert)
rge_vals <- c()
for (v in predictors){
  preds_m <- setdiff(predictors, v)
  rf_m <- randomForest(as.factor(bankrupt) ~ ., data = train[, c(preds_m, target)], ntree = 500)
```

```r
  rf_p <- predict(rf_m, newdata = test[, preds_m, drop = FALSE], type = "prob")[, "1"]
  xgb_m <- xgboost(data = as.matrix(train[, preds_m]), label = y_train, objective = "binary:logistic",
                   nrounds = 100, max_depth = 6, eta = 0.1, verbose = 0)
  xgb_p <- predict(xgb_m, as.matrix(test[, preds_m, drop = FALSE]))
  if (have_lgb){
    dtrain_m <- lgb.Dataset(data = as.matrix(train[, preds_m]), label = y_train)
    lgb_m <- lgb.train(params = list(objective="binary", metric="auc", learning_rate=0.1, num_leaves=31),
                       data = dtrain_m, nrounds = 100, verbose = 0)
    lgb_p <- predict(lgb_m, as.matrix(test[, preds_m, drop = FALSE]))
    avg_m <- (rf_p + xgb_p + lgb_p)/3
  } else {
    avg_m <- (rf_p + xgb_p)/2
  }
  rge_vals[v] <- RGE(avg_prob, avg_m)
}
rge_list$Ensemble <- rge_vals

# --- 8) Summaries & CSV exports ---
models <- names(auc_list)
auc_summary <- data.frame(
  Model = models,
  AUC   = sapply(models, function(m) as.numeric(auc_list[[m]])),
  RGA   = sapply(models, function(m) as.numeric(rga_list[[m]]))
) %>% arrange(desc(AUC))
print(auc_summary)
write.csv(auc_summary, "model_auc_rga_summary.csv", row.names = FALSE)

metrics_summary <- data.frame(
  Model = names(rgr_list),
  RGR   = sapply(names(rgr_list), function(m) as.numeric(rgr_list[[m]]))
) %>%
  left_join(auc_summary, by = "Model") %>%
  select(Model, AUC, RGA, RGR) %>%
  arrange(desc(AUC))
print(metrics_summary)
write.csv(metrics_summary, "extended_model_metrics.csv", row.names = FALSE)

rge_tables <- lapply(rge_list, function(x) if (is.null(x)) NULL else data.frame(Feature = names(x), RGE = as.numeric(x)))
for (nm in names(rge_tables)){
  if (!is.null(rge_tables[[nm]])) write.csv(rge_tables[[nm]], paste0("rge_", gsub(" ", "_", tolower(nm)), ".csv"), row.names = FALSE)
}

# --- 9) Plotting suite (PNG files) ---
```

```
open_png <- function(filename, width=1400, height=900, res=170) png(filename, width=width, height=height, res=res)
close_dev <- function() dev.off()
plt_cols <- c("#1f77b4","#ff7f0e","#2ca02c","#d62728","#9467bd","#8c564b","#e377c2","#17becf")
col_pick <- function(i) plt_cols[(i-1) %% length(plt_cols) + 1]

# correlation heatmap
if (nrow(df)>0){
  cor_matrix <- cor(df[, predictors], use = "complete.obs")
  open_png("01_correlation_heatmap.png", 1600, 900, 170)
  corrplot::corrplot(cor_matrix, method="color", type="upper", tl.cex=1.1, tl.col="black",
                     col = colorRampPalette(c("#2166AC","#67A9CF","#D1E5F0","#F7F7F7","#FDDBC7","#EF8A62","#B2182B"))(200),
                     addgrid.col="white", mar=c(0,0,2,0), title="Correlation Heatmap  Financial Ratios")
  close_dev()
}

# individual ROC plots
safe_roc_plot <- function(roc_obj, title_txt, file_txt, col_idx=1){
  if (!is.null(roc_obj)) { open_png(file_txt, 1400, 900, 170); plot(roc_obj, main=title_txt, lwd=3, col=col_pick(col_idx)); abline(0,1,lty=2,col="gray70"); close_dev() }
}
safe_roc_plot(roc_list$Logistic,  "Logistic Regression  ROC Curve", "02_roc_logistic.png",  1)
safe_roc_plot(roc_list$RandomForest, "Random Forest  ROC Curve", "03_roc_randomforest.png", 2)
safe_roc_plot(roc_list$XGBoost,  "XGBoost  ROC Curve", "04_roc_xgboost.png",  3)
if ("LightGBM" %in% names(roc_list)) safe_roc_plot(roc_list$LightGBM,  "LightGBM  ROC Curve",  "05_roc_lightgbm.png", 4)
safe_roc_plot(roc_list$SVM,  "SVM (RBF)  ROC Curve", "06_roc_svm.png",  5)
safe_roc_plot(roc_list$NeuralNet,  "Neural Network  ROC Curve", "07_roc_neural.png",  6)
safe_roc_plot(roc_list$Ensemble,  "Ensemble  ROC Curve", "08_roc_ensemble.png",  7)

# combined ROC overlay
open_png("09_roc_overlay.png", 1600, 1000, 170)
plotted <- FALSE; leg <- c(); cols <- c()
if (!is.null(roc_list$RandomForest)) { plot(roc_list$RandomForest, main="ROC Curves  Overlay", lwd=3, col=col_pick(1)); plotted <- TRUE; leg <- c(leg,"Random Forest"); cols <- c(cols,col_pick(1)) }
if (!is.null(roc_list$XGBoost))  { plot(roc_list$XGBoost, add=plotted, lwd=3, col=col_pick(2)); leg <- c(leg,"XGBoost"); cols <- c(cols,col_pick(2)) }
```

```r
if ("LightGBM" %in% names(roc_list)) { plot(roc_list$LightGBM, add=TRUE, lwd=3, col=col_pick(3)); leg <- c(leg,"LightGBM"); cols <- c(cols,col_pick(3)) }
if (!is.null(roc_list$NeuralNet))    { plot(roc_list$NeuralNet, add=TRUE, lwd=3, col=col_pick(4)); leg <- c(leg,"Neural Net"); cols <- c(cols,col_pick(4)) }
if (!is.null(roc_list$SVM))          { plot(roc_list$SVM, add=TRUE, lwd=3, col=col_pick(5)); leg <- c(leg,"SVM (RBF)"); cols <- c(cols,col_pick(5)) }
if (!is.null(roc_list$Logistic))     { plot(roc_list$Logistic, add=TRUE, lwd=3, col=col_pick(6)); leg <- c(leg,"Logistic"); cols <- c(cols,col_pick(6)) }
if (!is.null(roc_list$Ensemble))     { plot(roc_list$Ensemble, add=TRUE, lwd=3, col=col_pick(7)); leg <- c(leg,"Ensemble"); cols <- c(cols,col_pick(7)) }
abline(0,1, lty=2, col="gray70"); legend("bottomright", legend = leg, lwd=3, bty="n", col = cols); close_dev()

# importance plots
if (exists("rf_model")) { open_png("10_rf_variable_importance.png", 1600, 900, 170); varImpPlot(rf_model, main="Random Forest     Variable Importance", col=col_pick(1)); close_dev() }
if (exists("xgb_model")) {
  open_png("11_xgb_importance.png", 1600, 900, 170)
  xgb_imp <- xgb.importance(model = xgb_model)
  p <- try(xgb.plot.importance(xgb_imp, top_n = length(predictors)), silent = TRUE)
  if (inherits(p, "gg") || inherits(p, "ggplot")) { p <- p + ggplot2::ggtitle("XGBoost     Feature Importance"); print(p) } else { xgb.plot.importance(xgb_imp, top_n = length(predictors)); title("XGBoost Feature Importance") }
  close_dev()
}
if (exists("lgb_model")) {
  open_png("12_lgb_importance.png", 1600, 900, 170)
  lgb_imp <- lightgbm::lgb.importance(lgb_model); lightgbm::lgb.plot.importance(lgb_imp, top_n = length(predictors)); title("LightGBM Feature Importance"); close_dev()
}

# calibration curves
calib_plot <- function(probs, y, bins=10, title="Calibration Curve", fname="calibration.png", col_idx=1){
  if (is.null(probs) || length(probs)!=length(y)) return(invisible(NULL))
  cutp <- cut(probs, breaks = stats::quantile(probs, probs = seq(0,1, length.out=bins+1), na.rm=TRUE), include.lowest = TRUE)
  dfc <- aggregate(data.frame(prob=probs, y=y), by=list(bin=cutp), FUN=mean, na.rm=TRUE)
  open_png(fname, 1400, 900, 170)
  plot(dfc$prob, dfc$y, pch=19, xlab="Predicted probability (bin mean)", ylab="Observed bankruptcy rate",
```

```
293         main=title, xlim=c(0,1), ylim=c(0,1), col=col_pick(col_idx))
294   abline(0,1, lty=2, col="gray60"); close_dev()
295 }
296 if (exists("glm_prob")) calib_plot(glm_prob, test[[target]], title="
      Calibration     Logistic Regression", fname="13_calibration_logistic.
      png", col_idx=1)
297 if (exists("avg_prob")) calib_plot(avg_prob, test[[target]], title="
      Calibration     Ensemble", fname="14_calibration_ensemble.png", col_
      idx=2)
298
299 # RGA/RGR bars
300 get_num <- function(lst, name){ v <- lst[[name]]; if (is.null(v)) NA_real
      _ else as.numeric(v) }
301 all_models <- sort(unique(c(names(auc_list), names(rga_list), names(rgr_
      list))))
302 metrics_df <- data.frame(
303   Model = all_models,
304   AUC   = sapply(all_models, function(m) get_num(auc_list, m)),
305   RGA   = sapply(all_models, function(m) get_num(rga_list, m)),
306   RGR   = sapply(all_models, function(m) get_num(rgr_list, m))
307 )
308 write.csv(metrics_df, "extended_model_metrics.csv", row.names = FALSE)
309
310 open_png("15_rga_bar.png", 1600, 1000, 170); par(mar=c(10,5,3,1))
311 barplot(height=metrics_df$RGA, names.arg=metrics_df$Model, las=2, ylab="
      RGA", main="Rank Graduation Accuracy (by model)",
312         col = plt_cols[seq_len(nrow(metrics_df)) %% length(plt_cols) +
              1]); grid(); close_dev()
313
314 open_png("16_rgr_bar.png", 1600, 1000, 170); par(mar=c(10,5,3,1))
315 barplot(height=metrics_df$RGR, names.arg=metrics_df$Model, las=2, ylab="
      RGR", main="Rank Graduation Robustness (by model)",
316         col = plt_cols[seq_len(nrow(metrics_df)) %% length(plt_cols) +
              1]); grid(); close_dev()
317
318 # RGE per-model bars
319 if (length(rge_list)){
320   for (m in names(rge_list)){
321     vals <- rge_list[[m]]; if (is.null(vals)) next
322     df_rge <- data.frame(Feature = names(vals), RGE = as.numeric(vals))
323     fname <- paste0("17_rge_", gsub(" ", "_", tolower(m)), ".png")
324     open_png(fname, 1600, 900, 170); par(mar=c(10,5,3,1))
325     barplot(df_rge$RGE, names.arg=df_rge$Feature, las=2, ylab="RGE", main
          =paste("RGE    ", m),
326             col = plt_cols[seq_len(nrow(df_rge)) %% length(plt_cols) +
                  1]); grid(); close_dev()
327   }
328 }
329
330 cat("\nArtifacts written:\n",
331     " - model_auc_rga_summary.csv\n",
```

```
    " - extended_model_metrics.csv\n",
    " - rge_<model>.csv for each model\n",
    " - 01_correlation_heatmap.png\n",
    " - 02..08 ROC curves\n",
    " - 09_roc_overlay.png\n",
    " - 10_rf_variable_importance.png\n",
    " - 11_xgb_importance.png\n",
    if (have_lgb) " - 12_lgb_importance.png\n" else "",
    " - 13_calibration_logistic.png\n",
    " - 14_calibration_ensemble.png\n",
    " - 15_rga_bar.png, 16_rgr_bar.png\n",
    " - 17_rge_<model>.png per model\n")
```

Listing A.1: Full R pipeline for bankruptcy prediction (2014–2024): data preparation, modeling (Logistic, Random Forest, XGBoost, LightGBM, SVM, Neural Network), ensemble, metrics (AUC, RGA, RGR, RGE), and visualization suite.

Appendix B

Appendix B: Variable Descriptions

- **Liquidity_Ratio** – Current assets divided by current liabilities.
- **Cash_Ratio** – Cash holdings relative to short-term liabilities.
- **Debt_Ratio** – Total debt to total assets, indicating leverage.
- **Working_Capital** – Difference between current assets and liabilities.
- **EBITDA** – Cash flow plus depreciation, proxy for profitability.
- **Intangible_Ratio** – Intangible assets over total assets.

Appendix C

Appendix C: Additional Model Outputs

This appendix includes all extended visualizations generated during the modeling phase. Each figure corresponds to output files produced by the `pipeline.R` script.

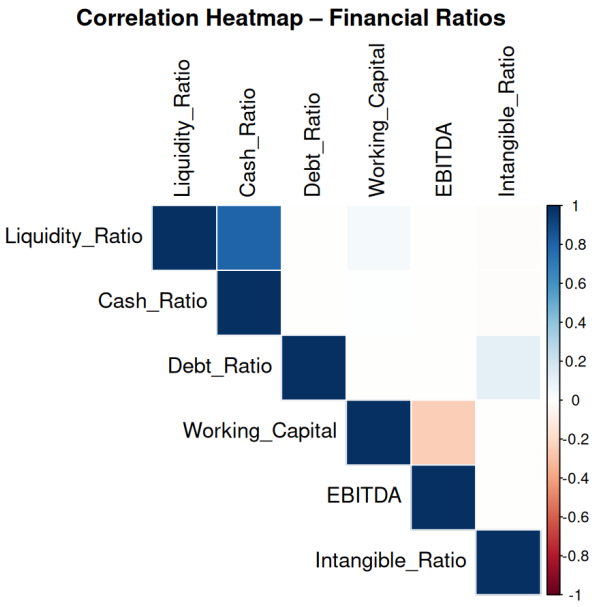

Figure C.1: Correlation heatmap of financial ratios.

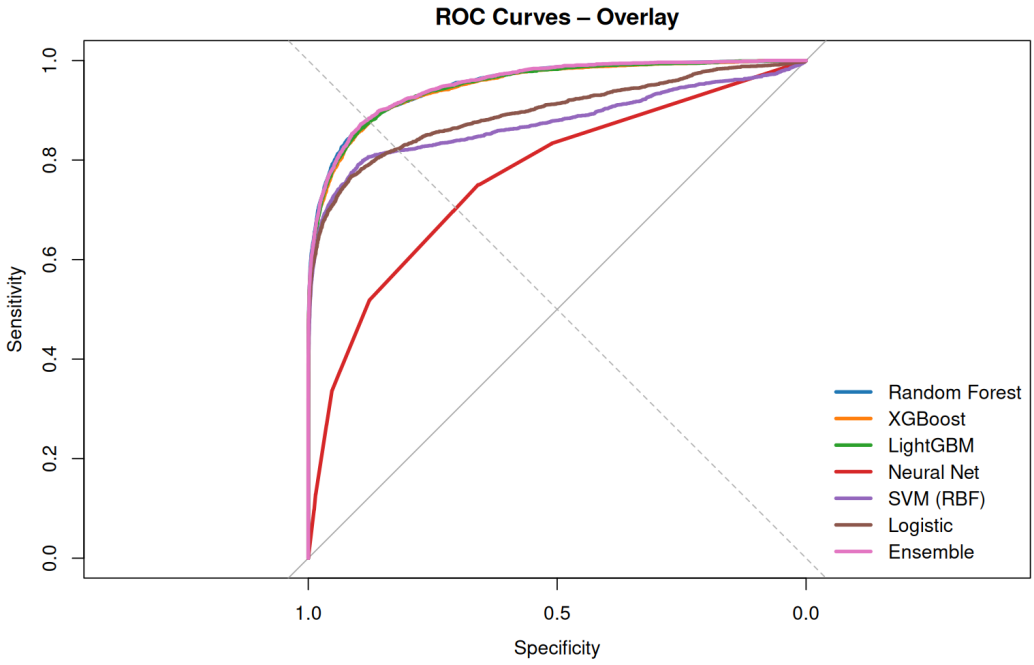

Figure C.2: Overlay of ROC curves across models.

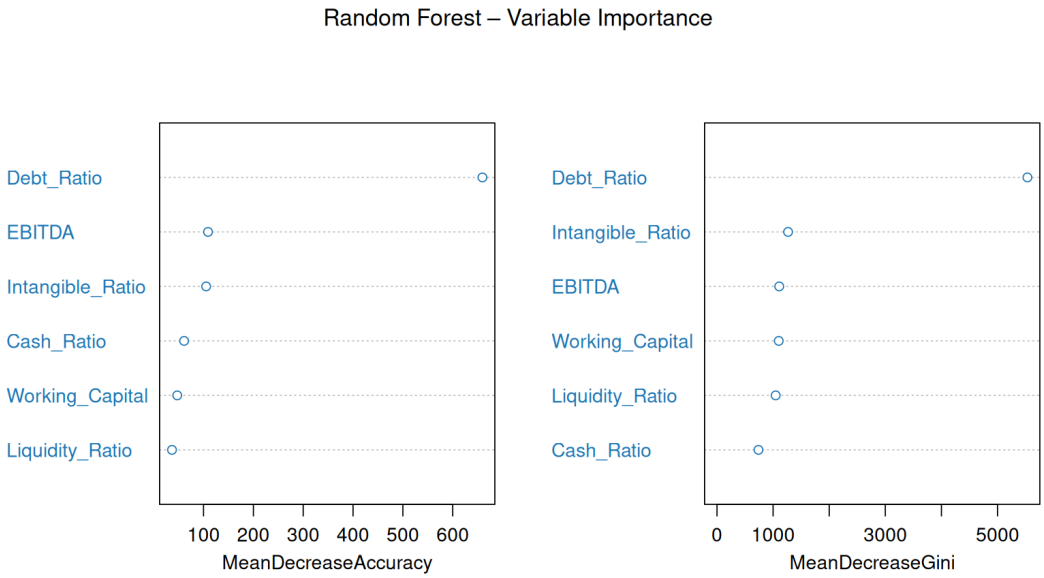

Figure C.3: Random Forest – variable importance ranking.

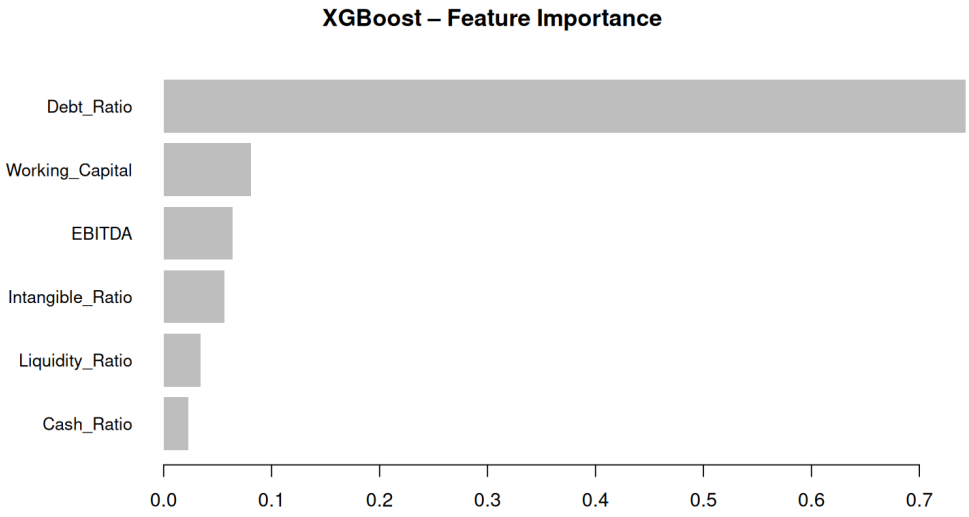

Figure C.4: XGBoost – feature importance.

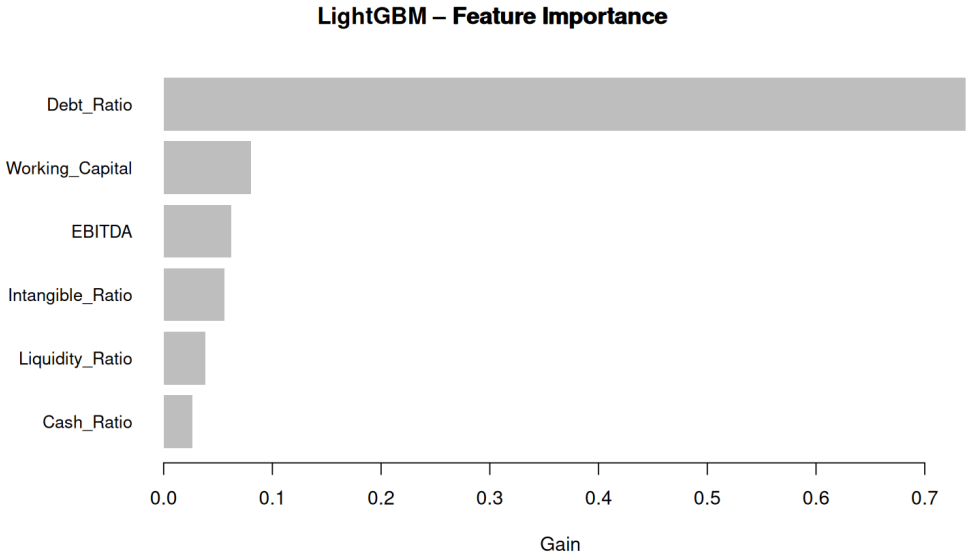

Figure C.5: LightGBM – feature importance (if model trained).

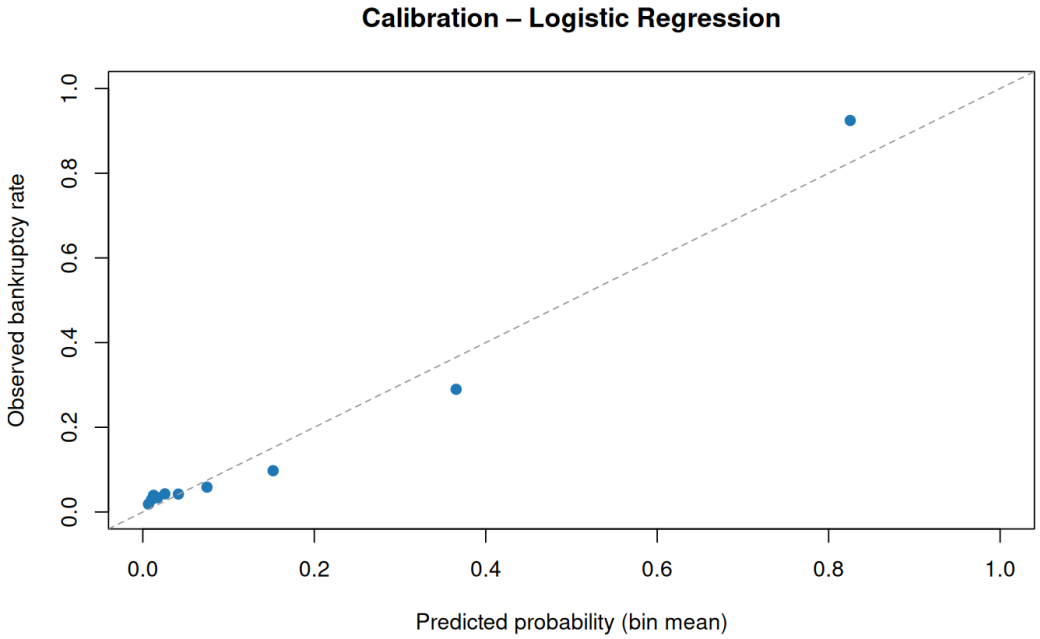

Figure C.6: Calibration curve – Logistic Regression model.

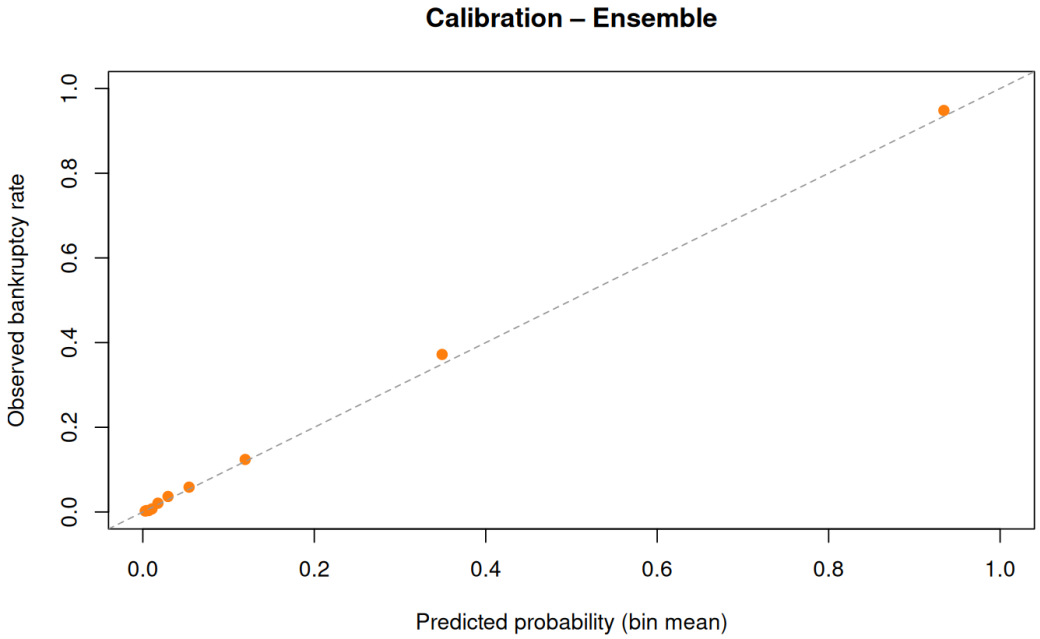

Figure C.7: Calibration curve – Ensemble model.

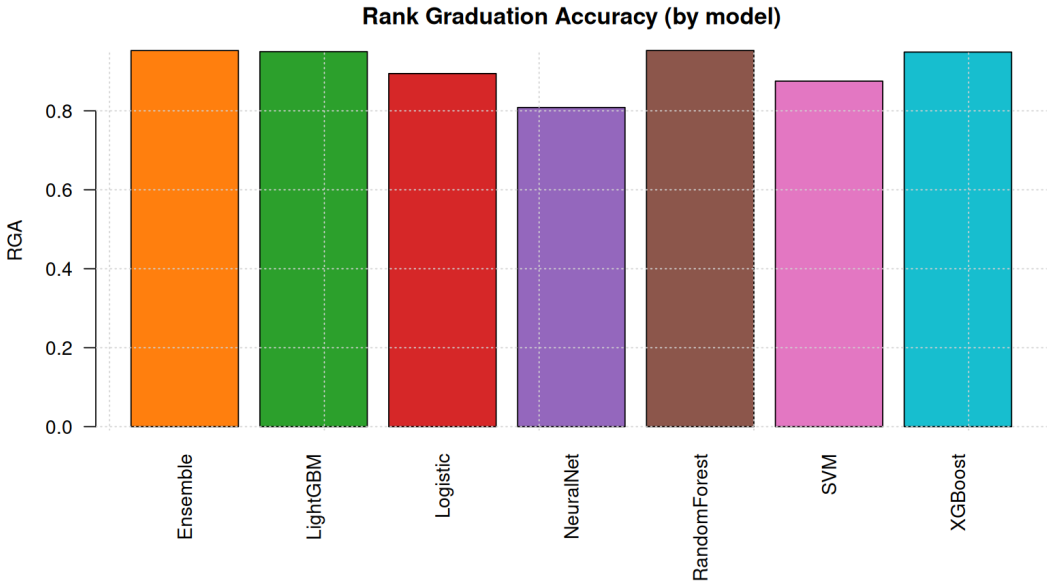

Figure C.8: Rank Graduation Accuracy (RGA) across models.

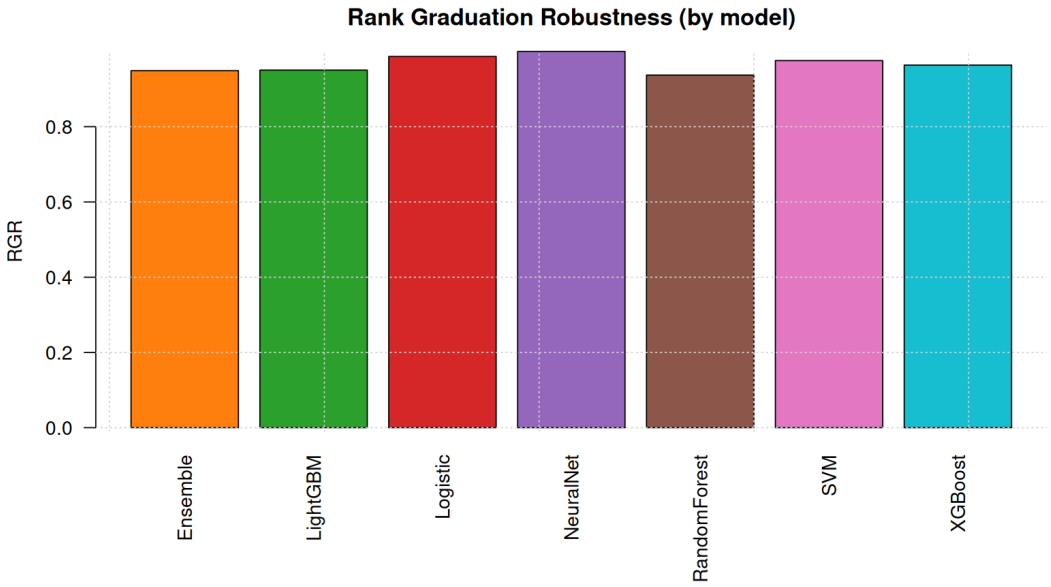

Figure C.9: Rank Graduation Robustness (RGR) across models.

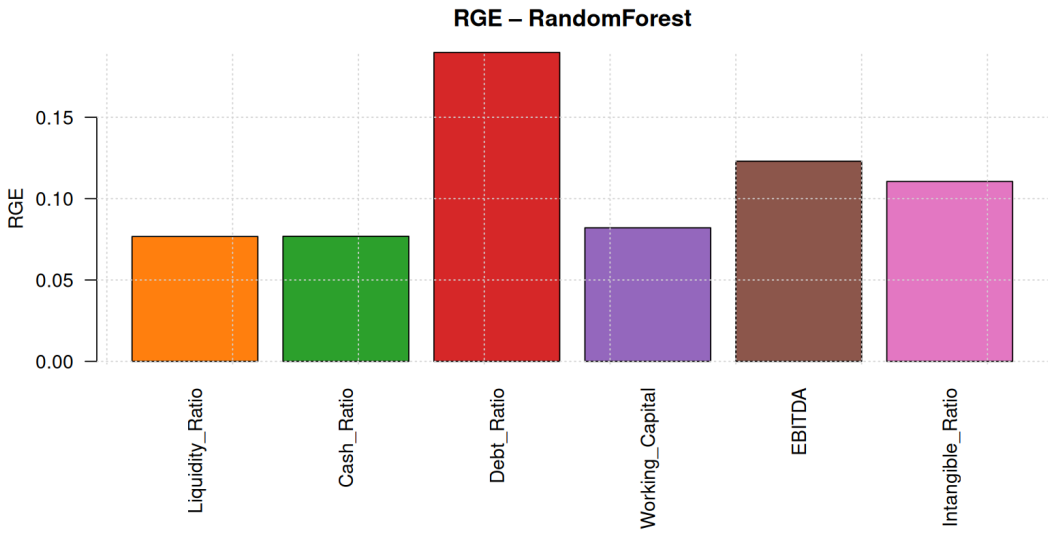

Figure C.10: Rank Graduation Explainability (RGE) – Random Forest model.

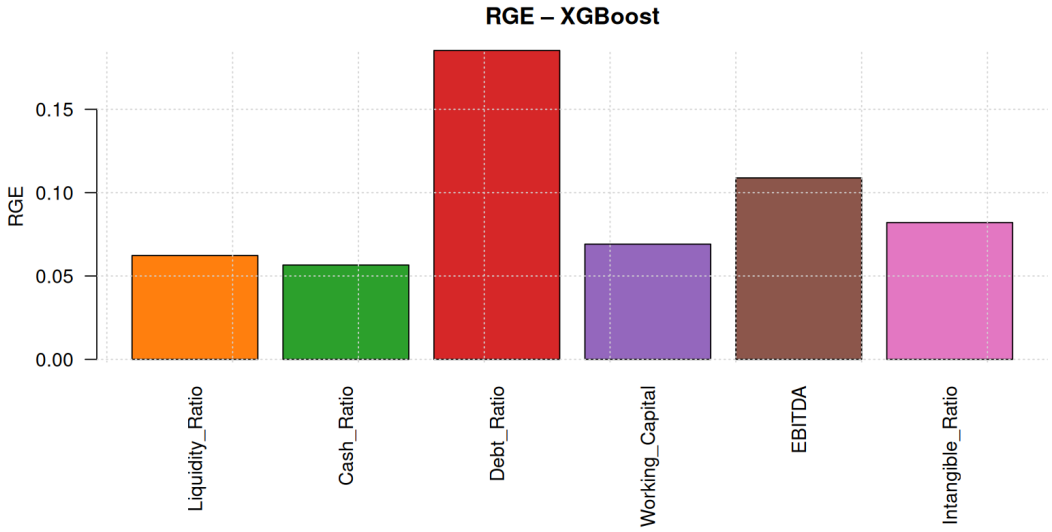

Figure C.11: Rank Graduation Explainability (RGE) – XGBoost model.

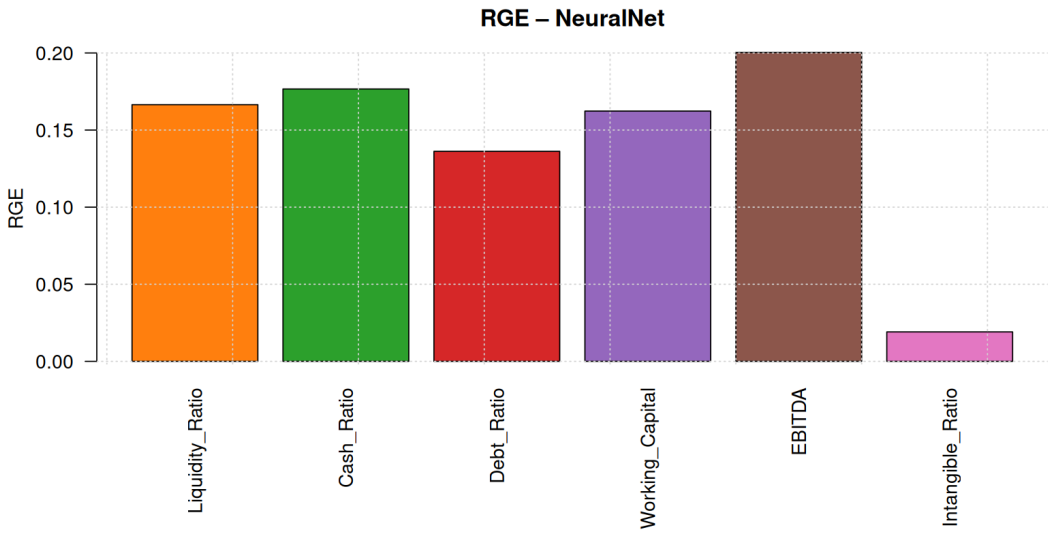

Figure C.12: Rank Graduation Explainability (RGE) – Neural Network model.

Appendix D

Appendix D: R Session Information

For reproducibility, this appendix provides the exact R environment used in all analyses, including R version, system information, and package versions.

```
R version 4.5.0 (2025-04-11)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.3 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.26.
    so;  LAPACK version 3.12.0

locale:
 [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=C.UTF-8        LC_COLLATE=C.UTF-8
 [5] LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
 [7] LC_PAPER=C.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C           LC_TELEPHONE=C
[11] LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C

time zone: Asia/Singapore
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
 [1] lightgbm_4.6.0       corrplot_0.95
 [3] nnet_7.3-20          e1071_1.7-16
 [5] xgboost_1.7.11.1     randomForest_4.7-1.2
 [7] pROC_1.19.0.1        caret_7.0-1
```

```
 [9] lattice_0.22-6       lubridate_1.9.4
[11] forcats_1.0.1        stringr_1.5.2
[13] dplyr_1.1.4          purrr_1.1.0
[15] tidyr_1.3.1          tibble_3.3.0
[17] ggplot2_3.5.2        tidyverse_2.0.0
[19] readr_2.1.5

loaded via a namespace (and not attached):
 [1] gtable_0.3.6         recipes_1.3.1
 [3] remotes_2.5.0        tzdb_0.5.0
 [5] vctrs_0.6.5          tools_4.5.0
 [7] generics_0.1.4       stats4_4.5.0
 [9] curl_6.4.0           parallel_4.5.0
[11] proxy_0.4-27         pkgconfig_2.0.3
[13] ModelMetrics_1.2.2.2 Matrix_1.7-3
[15] data.table_1.17.8    RColorBrewer_1.1-3
[17] lifecycle_1.0.4      compiler_4.5.0
[19] farver_2.1.2         codetools_0.2-20
[21] class_7.3-23         prodlim_2025.04.28
[23] crayon_1.5.3         pillar_1.11.0
[25] MASS_7.3-65          gower_1.0.2
[27] iterators_1.0.14     rpart_4.1.24
[29] foreach_1.5.2        nlme_3.1-168
[31] parallelly_1.45.1    lava_1.8.1
[33] tidyselect_1.2.1     digest_0.6.37
[35] stringi_1.8.7        future_1.67.0
[37] reshape2_1.4.4       listenv_0.9.1
[39] splines_4.5.0        grid_4.5.0
[41] cli_3.6.5            magrittr_2.0.3
[43] survival_3.8-3       future.apply_1.20.0
[45] withr_3.0.2          scales_1.4.0
[47] bit64_4.6.0-1        timechange_0.3.0
[49] globals_0.18.0       bit_4.6.0
[51] timeDate_4041.110    hms_1.1.3
[53] hardhat_1.4.1        rlang_1.1.6
[55] Rcpp_1.1.0           glue_1.8.0
[57] ipred_0.9-15         vroom_1.6.5
[59] jsonlite_2.0.0       rstudioapi_0.17.1
[61] R6_2.6.1             plyr_1.8.9
```

Listing D.1: R `sessionInfo()` output (analysis environment)