UNIVERSITY OF MISKOLC

FACULTY OF ECONOMICS

SAFA SEN

# Predicting Bank Failures: A Comparative Analysis of Traditional and Machine Learning Models in the Post-2008 Financial Landscape

Supervisor: Dr. Levente Kovacs

Head of the doctoral school: Prof. Dr. Tóth Géza

Hantos Elemér Doctoral School of Business, Management and Regional Sciences

Miskolc, 2025

Public

**UNIVERSITY OF MISKOLC**

FACULTY OF ECONOMICS

# Predicting Bank Failures: A Comparative Analysis of Traditional and Machine Learning Models in the Post-2008 Financial Landscape

## Doctoral (PhD) Thesis

## SAFA SEN

Miskolc,2025

Public

Témavezetői ajánlás
**Safa Sen**
Ph.D. fokozatszerzésének megindításához

Safa Sen Ph.D. témavezetését az elmúlt években elvégeztem. Ezen idő alatt a témavezetett hallgató tanulmányi és vizsga kötelezettségéinek eleget tett, az előírt publikációs követelményeket teljesítette. Az értekezése a házi védés során megvitatásra került és a javaslatok alapján értekezését tovább fejlesztette.

Értekezésének címe: „Predicting Bank Failures: A Comparative Analysis of Traditional and Machine Learning Models in the Post-2008 Financial Landscape".

A fentiek alapján számára a fokozatszerzés megindítását támogatom.

Miskolc, 2025. április 1.

Prof. Dr. Kovács Levente
témavezető, Miskolci Egyetem

---

01 April 2025
University of Miskolc

Supervisor's Recommendation for
**Safa Sen**
to obtain the Ph.D. degree

I have fulfilled the supervision of Safa Sen Ph.D. studies over the past years. During this time, the student has fulfilled the study and examination obligations and fulfilled the publication requirements. His dissertation was successfully discussed at the workshop discussion and he carried out the necessary improvements in his dissertation based on the suggestions and remarks at the workshop.

The title of the dissertation is: "Predicting Bank Failures: A Comparative Analysis of Traditional and Machine Learning Models in the Post-2008 Financial Landscape".

Based on the above, I support the process of obtaining the degree.

Levente Kovács Prof. Ph.D.
supervisor

# Declaration

I, Safa Sen, hereby certify that the content of this dissertation titled “Predicting Bank Failures: A Comparative Analysis of Traditional and Machine Learning Models in the Post-2008 Financial Landscape” submitted for the degree of Doctor of Philosophy (PhD) in Economics at the University of Miskolc, is my original work. All sources utilized, including data, figures, equations, tables, and concepts from other authors, have been duly acknowledged. A comprehensive list of references is provided in the bibliography.

I further affirm that this dissertation has not been previously submitted, either in full or in part, for any academic award or diploma at this or any other institution of higher learning. Any errors or omissions are solely my own responsibility.

Miskolc, April 2025

Name: Safa Sen

Signature:

Public

# Abstract

This dissertation investigates the efficacy of various machine learning models in predicting bank failures, a crucial area of study due to the significant role banks play in maintaining economic stability by channeling funds between savers and borrowers. Traditional statistical methods, such as linear and logistic regression, have historically been used for this purpose but have shown limitations in addressing the complexities of modern financial markets. The research aims to analyze and compare the performance of advanced machine learning models, particularly Random Forest and CS-Forest, in accurately predicting bank failures, with a focus on their application to the 2008 financial crisis. The study utilizes a comprehensive empirical framework, incorporating a range of financial ratios and performance metrics to assess these models' classification accuracy, recall, and ROC area, alongside their effectiveness in handling imbalanced datasets—a common challenge in financial prediction. The results reveal that Random Forest and CS-Forest outperform other models, with Random Forest achieving the highest overall classification accuracy (0.977) and an outstanding ROC area (0.991), indicating its superior ability to correctly distinguish between failing and non-failing banks. CS-Forest also demonstrates strong performance, particularly in recall (0.897), highlighting its proficiency in identifying true positives, which is critical for early detection and intervention in potential bank failures. These findings underscore the importance of model complexity in enhancing predictive power, with both Random Forest and CS-Forest offering significant advantages over simpler models like Naive Bayes. The significance of this research lies in its potential to improve the tools available for financial regulators and policymakers, providing them with more reliable methods to anticipate and mitigate bank failures, thereby contributing to greater economic stability. This work not only advances the field of financial risk management but also sets a foundation for future innovations in predictive analytics, emphasizing the critical balance between model complexity, accuracy, and interpretability in high-stakes financial applications.

Public

Table of Contents

Public

## List of Figures

## List of Tables

Public

# Chapter 1: Introduction

The banking sector serves as a fundamental component of modern economic systems, facilitating economic growth and stability by channeling funds from savers to borrowers. Banks fulfill a unique intermediary role, converting savings into productive investments that support businesses, infrastructure, and personal financial goals. This role underscores the importance of understanding bank failures, a topic that has gained significant attention in financial literature due to its implications for regulatory practice and economic stability. Bank stability is essential not only for depositors and investors but also for the economic health of communities, as banks are a primary source of credit and liquidity within economies. Studies on bank failures aim to provide regulatory bodies with insights that improve their supervisory capacities, which are essential for preventing such events or mitigating their impact. Thomson (1991) highlights that effective regulation, informed by early indicators of financial distress, can proactively address vulnerabilities, thus reducing the fiscal burden on taxpayers. Additionally, identifying the differences between stable and at-risk banks is critical for maintaining confidence in the banking sector, as early detection of financial instability allows for timely intervention that can minimize broader economic consequences (Meyer & Pifer, 1970).

The events surrounding the 2007-2009 Great Financial Crisis (GFC) underscored the importance of robust risk management practices within the banking sector. Unlike previous financial downturns, the GFC was both widespread and severe, with origins in the U.S. subprime mortgage market quickly evolving into a global crisis. The complex financial structures created in the years preceding the GFC, including mortgage-backed securities (MBS) and collateralized debt obligations (CDOs), enabled the financial system to distribute credit widely, but they also spread risks that were insufficiently understood and managed. Key drivers of the crisis included an extended period of credit expansion, rising housing prices, and increased leverage within the financial system. These trends were amplified by financial innovations such as securitization, which enabled financial institutions to create extensive credit assets but, at times, outpaced their ability to manage associated risks (Mian & Sufi, 2009; Shiller, 2008). The crisis led to substantial losses within financial institutions worldwide, and banks in particular faced liquidity challenges as asset valuations became uncertain. Governments responded with liquidity support measures and recapitalization initiatives to stabilize lending and restore market confidence (Bernanke, 2010; Gorton, 2009).

The global scope and severity of the GFC highlighted the interconnectedness of financial markets and the vulnerabilities of financial institutions. A significant factor in the GFC was the rapid growth in mortgage financing, which drove housing demand and attracted investors to high-yield subprime mortgage securities. These assets were often structured into complex securities that dispersed risk across various investors but also concealed underlying vulnerabilities (Kothari, 2008). This situation created a false sense of security, as the risk was theoretically shared, but when mortgage defaults began to rise, the losses rippled through the global financial system. The resulting securitization, though intended to broaden risk-sharing, instead exposed the interconnected nature of global financial markets, as deteriorating mortgage performance led to downgrades and losses. In many cases, banks transferred assets off their balance sheets to structured investment vehicles, theoretically spreading risk but ultimately amplifying it (Ashcraft & Schuermann, 2008). These conditions highlighted the importance of predictive modeling in banking, underscoring how risk assessment tools could play a key role in managing systemic risks within complex financial landscapes.

Public

In response to these heightened risks, the academic journey into understanding and predicting bank failures has evolved considerably. Initial research focused on traditional statistical models and financial ratios to assess failure risks, using linear and quadratic discriminant analysis, logistic regression, and factor analysis. Early studies by Meyer and Pifer (1970), Sinkey (1975), and Martin (1977) sought to identify patterns in financial ratios that could signal distress within banks. These foundational models provided valuable insights and were instrumental in early bank regulatory frameworks. However, as financial markets grew in complexity, these traditional approaches faced limitations. Traditional models struggled to incorporate the full scope of variables and nonlinear relationships present in complex financial systems. This challenge led researchers to integrate advanced computational techniques that can capture complex interactions among financial indicators. For instance, recent studies by Boyacıoglu et al. (2009) and Erdal and Ekinci (2013) employed machine learning approaches, including artificial neural networks and support vector machines, which are recognized for their capacity to model complex, nonlinear relationships in data and enhance predictive accuracy.

The evolution from traditional statistical techniques to advanced machine learning and artificial intelligence-based methods mirrors the growing sophistication of financial markets. Techniques such as artificial neural networks and support vector machines provide powerful frameworks for handling extensive datasets, offering greater accuracy and insight into failure risks. Machine learning models, unlike traditional statistical models, can also adapt to new data patterns, making them especially valuable in environments as dynamic as financial markets. This shift reflects an acknowledgment within financial research that more robust tools are necessary to assess and predict bank stability within increasingly interconnected financial environments. The ability to incorporate vast amounts of data from various sources, including market prices, economic indicators, and firm-specific financial metrics, has been a game-changer in the field.

In the context of these developments, this thesis aims to investigate the efficacy of various risk assessment models, comparing traditional statistical approaches with machine learning techniques to determine their effectiveness in predicting bank failures. The 2008 financial crisis serves as a pivotal reference point for this study, illustrating the urgent need for accurate and timely risk assessments. Lessons learned from the GFC reveal that reliance on limited risk models may be insufficient to foresee or contain financial crises. By exploring both traditional and contemporary approaches, this thesis contributes to a more nuanced understanding of risk prediction within the banking sector. This study not only aims to inform future regulatory frameworks but also seeks to provide insights into the underlying patterns that signal financial distress.

This thesis is structured to explore these insights through a comprehensive literature review of relevant studies on bank failure prediction, an overview of traditional and machine learning-based methodologies, and an evaluation of various model performance metrics. In doing so, it aims to contribute to the development of effective predictive techniques and to the broader discourse on financial stability and risk management. Through this examination, the research highlights the role of advanced predictive models in identifying early signs of bank distress and offers insights that may support both policymaking and practical applications in financial supervision.

Public

Hypotheses

Hypothesis 1 (Model Performance in Classification Accuracy)
H1: The Random Forest model does not significantly outperform other machine learning models (including Support Vector Machines, Logistic Regression, and XGBoost) in terms of classification accuracy for predicting bank failures.
Justification: According to the literature, while Random Forest has demonstrated strong performance in bank failure prediction (Petropoulos et al., 2020), other models like SVM (Gogas et al., 2018) and XGBoost (Carmona et al., 2019) have achieved comparable, if not superior, classification accuracy. Therefore, it is necessary to test whether Random Forest truly outperforms other models across various scenarios.

H2: The CS-Forest model does not exhibit superior performance in reducing Type II errors and handling imbalanced datasets, such as those used in predicting bank failures, when compared to other hybrid ensemble models (e.g., AdaBoost, Gradient Boosting).
Justification: Preventing bank failures is crucial for maintaining economic stability, and generating early signals to identify distressed banks is a key objective. CS-Forest has been shown to outperform other models in reducing Type II errors, making it an effective tool for handling imbalanced datasets in bank failure prediction (Ekinci and Sen, 2024). This hypothesis tests whether CS-Forest can outperform other hybrid models in this context.

Hypothesis 3 (Effectiveness of Simplicity vs. Complexity in Models)
H3: Simpler models, such as Naive Bayes, do not exhibit significantly lower effectiveness compared to more complex models, such as Multilayer Perceptron, Artificial Neural Networks, or ensemble methods, in predicting bank failures.
Justification: Studies such as Le and Viviani (2017) found that simpler models like k-NN and Naive Bayes could match or even outperform more complex models in certain classification tasks. This hypothesis aims to challenge the assumption that increased model complexity necessarily results in higher prediction accuracy.

Hypothesis 4 (Model Choice Influences)
H4: The selection of machine learning models for bank failure prediction is not predominantly influenced by the trade-off between model complexity and interpretability, nor by specific task requirements.
Justification: The literature suggests that while interpretability is an important factor, especially in financial contexts (Cleary and Hebb, 2015; Petropoulos et al., 2020), it is often not the sole determinant in model choice. For example, models like SVMs and XGBoost, despite their complexity, are favored for their superior accuracy in certain contexts (Carmona et al., 2019).

Hypothesis 5 (Predictive Power and Model Complexity)
H5: The predictive power of machine learning models in bank failure prediction is not directly correlated with model complexity.
Justification: Several studies (e.g., Beutel et al., 2019; Plessis, 2022) have demonstrated that traditional models, such as Logistic Regression and the Z-Score, can provide prediction accuracies comparable to or better than more complex models. This hypothesis seeks to investigate whether model complexity is truly a significant factor in enhancing predictive power.

Public

# Chapter 2: Literature Review

Torna and DeYoung (2012) conducted an insightful study to explore the impact of income from non-traditional banking activities on the failure of U.S. commercial banks, particularly during periods of financial crisis. Their research focused on understanding whether diversification into non-traditional activities, often perceived as a means to reduce risk and enhance stability, actually influenced bank vulnerability during economic downturns.

This investigation is significant given the evolving nature of banking activities and the increasing inclination of banks towards non-traditional revenue sources. In their methodology, Torna and DeYoung specifically excluded banks with assets exceeding $100 billion. This exclusion criterion was likely aimed at focusing on a segment of the banking sector where the impact of non-traditional activities might be more pronounced or potentially more hazardous.

The rationale behind this could be that larger banks often have a more diverse set of activities and risk management strategies, which might obscure the specific effects of non-traditional activities on bank stability. The study employed multi-period logistic regression as its primary analytical tool. Logistic regression is particularly suited for this type of analysis due to its ability to handle binary outcome variables – in this case, the failure or survival of banks.

The multi-period aspect of the model allowed for the examination of data across different time frames, capturing the dynamic nature of banking activities and economic conditions over time. The findings of Torna and DeYoung were quite revealing. They observed that the probability of distressed bank failure decreased with income from certain fee-based non-traditional activities, such as securities brokerage and insurance sales.

This suggests that these types of activities might provide a stabilizing effect on banks, possibly due to their fee-based nature, which could be less sensitive to market fluctuations compared to other revenue sources. Conversely, they found that engagement in activities like venture capital, investment banking, and asset securitization increased the likelihood of bank failures.

This increase in risk could be attributed to the inherent risks associated with these activities, such as higher market and credit risks, and potentially higher volatility in income. These findings align with the broader discourse in financial literature that suggests that some non-traditional banking activities might expose banks to greater risks, especially during times of financial stress.

Critique and Summary: Torna and DeYoung’s exclusion of larger banks limits the generalizability of their findings to the entire banking sector. While their methodology using logistic regression is robust, it could benefit from including larger banks to see if the findings hold across different bank sizes. Future research should address this gap to provide a more comprehensive understanding of non-traditional activities' impact on bank stability.

Berger and Bouwman's (2012) study offers a critical empirical analysis of the role of capital in determining the financial performance of banks, particularly during financial crises. This research is pivotal, as it delves into the often-discussed topic in financial economics regarding the protective buffer that bank capital provides against insolvency risks, especially in turbulent economic times.

Public

Their focus on different bank sizes adds a nuanced understanding of how capital impacts banks differently based on their scale. The study covers an extensive period, from the first quarter of 1984 through the fourth quarter of 2010. This timeframe is particularly relevant as it includes several economic cycles, including major financial disruptions like the savings and loan crisis of the late 1980s and early 1990s, the early 2000s recession, and the global financial crisis of 2007-2008. Such a comprehensive period allows for a robust examination of the role of bank capital across different economic conditions.

Berger and Bouwman utilized two key statistical methods in their analysis: Logit survival and ordinary least squares (OLS) regression models. The Logit survival model is particularly suited for analyzing the time-to-event data, in this case, the time until a bank's failure. This model helps in understanding the factors that influence the likelihood and timing of bank failures. The OLS regression model, on the other hand, is used to investigate the relationship between bank capital and performance during normal times and crises. The empirical outcomes of their study are revealing. For small banks, the strength of the capital base was found to lower the bankruptcy probability. This finding aligns with the traditional view in banking theory that a strong capital base acts as a cushion against losses, thereby reducing the likelihood of insolvency for smaller banks, which might be more vulnerable during economic downturns.

In contrast, for medium and large banks, the study found that a stronger capital base enhances performance during banking crises. This could be interpreted as larger banks being able to leverage their strong capital positions to undertake profitable opportunities during crises, or it could imply that a strong capital position provides these banks with a competitive advantage during times of economic stress. Berger and Bouwman's study make a significant contribution to the literature on banking and finance. It highlights the importance of capital strength not only as a protective measure against bankruptcy but also as a strategic asset that can influence the performance of banks, particularly during periods of financial instability. This research has practical implications for bank managers and regulators, emphasizing the need for careful capital management, especially in preparation for and during financial crises.

Critique and Summary: Berger and Bouwman’s comprehensive dataset and methodology are strengths of their study. However, the analysis could be enriched by examining how different regulatory environments affect the role of capital. Additionally, the study could explore how capital interacts with other factors, such as liquidity and asset quality, to influence bank performance during crises. Future research should consider these aspects to provide a more holistic understanding of capital's role in banking stability and performance.

Lu and Whidbee's 2013 study represents a significant contribution to the understanding of bank failures, particularly in the context of the 2007-2011 financial crisis. By applying logistic regression, a robust statistical tool for binary outcome analysis, they sought to unravel the complex interplay of factors contributing to bank failures and to assess various bank-level characteristics that might influence a bank's resilience or vulnerability during financial distress.

Their research encompassed a substantial dataset of 6,236 U.S. banks, a scope that offers a comprehensive view of the banking landscape during one of the most turbulent periods in recent financial history. Notably, out of these, 324 banks failed between 2007 and 2011, a period marked by severe financial instability and economic downturns, making it an ideal case for studying the dynamics of bank failures. The use of logistic regression in their study is particularly apt for this kind of analysis. Logistic regression is effective in cases where the dependent variable is dichotomous – in this instance, the failure or survival of banks.

This method allows for the examination of how different independent variables (or bank-level characteristics) affect the probability of a bank failing. One of the key empirical results from Lu and Whidbee's study is the identification of a potential linkage between a bank's financial fragility and its likelihood of failure. This finding is crucial as it underscores the importance of financial health and stability in determining a bank's ability to withstand economic shocks. Financial fragility, which may encompass aspects like capital adequacy, asset quality, and liquidity, can significantly heighten a bank's risk of failure during periods of financial turmoil.

Another important finding of their research is the greater survival chances of multi-bank holding companies compared to single banks during the financial crisis. This observation suggests that diversification benefits associated with multi-bank holding structures could provide a buffer against the risk of failure. Multi-bank holding companies might benefit from economies of scale, more diversified revenue streams, and a broader geographical or sectoral presence, all of which can contribute to greater resilience in the face of financial distress. Lu and Whidbee's study offer valuable insights for both banking sector practitioners and regulators. Understanding the factors that contribute to bank failures is crucial for risk management and regulatory oversight, particularly in preparing for and mitigating the impacts of financial crises. Their findings highlight the need for robust financial buffers and the potential benefits of diversification in banking operations.

Critique and Summary: Lu and Whidbee's study is significant for its large sample size and focus on a turbulent financial period. However, the reliance on logistic regression might overlook non-linear relationships between variables. Incorporating machine learning techniques could provide deeper insights into the complex interplay of factors leading to bank failures. Additionally, exploring the impact of different regulatory environments and macroeconomic conditions could further enhance the robustness of their findings. Future research should consider these aspects to provide a more comprehensive understanding of bank failure dynamics.

The study conducted by Cleary and Hebb in 2015 stands as a significant investigation into bank failures, employing discriminant analysis, a powerful statistical tool, to delve into the factors leading to bank failures during the critical period of 2002-2009. This time frame encompasses the lead-up to and aftermath of the global financial crisis of 2007-2008, making it a particularly pertinent period for such analysis.

Their research focused on a sample of 132 banks, providing a substantial dataset to examine the characteristics and indicators that might predict bank failures. The use of discriminant analysis in this context is noteworthy. Discriminant analysis is a classification method that is used to predict a categorical dependent variable (like bank failure) by analyzing the relationship between that variable and several independent variables. This method is particularly effective for situations where the purpose is to classify observations (in this case, failed versus non-failed banks) based on several predictor variables.

One of the most striking findings from their study is the high prediction efficiency of bank failure, which stood at 92% for their sample data. This high level of accuracy indicates that the6 discriminant analysis was able to effectively identify the key factors that differentiated failed banks from those that did not fail during the period studied.

Public

This high prediction efficiency is a testament to the power of discriminant analysis in identifying and quantifying the factors that contribute to bank failures. Cleary and Hebb extended their analysis to predict bank failures in the subsequent period of 2010-2011.

The efficiency of their model in this later period ranged between 90-95%, underscoring the robustness of their analytical approach. The consistency in the high predictive efficiency across different periods highlights the effectiveness of the discriminant analysis method in capturing the dynamics of bank failures across varying economic conditions.

Critique and Summary: While Cleary and Hebb demonstrate the efficacy of discriminant analysis, I believe their study could benefit from comparing this method with other advanced techniques like neural networks or ensemble methods to validate and potentially improve predictive accuracy. Additionally, the study could explore the practical implications of their findings for bank management and regulators. Integrating a comparative analysis with other machine learning techniques could provide deeper insights and enhance the robustness of their predictive models.

The study by Chiaramonte et al. in 2016 provides a notable contribution to the field of financial risk analysis, focusing on the efficacy of the Z-Score in predicting bank failures among U.S. commercial banks. Their research, which spans from 2004 to 2012, a period inclusive of the global financial crisis of 2007-2008, offers valuable insights into the predictive power of this financial metric under varying economic conditions.

The Z-Score is a widely recognized financial metric used to assess the risk of bankruptcy of a firm or financial institution. It combines several key financial ratios, including profitability, leverage, and variability of earnings, to produce a single score that reflects the likelihood of failure. In the context of banking, a lower Z-Score indicates a higher probability of bank failure, making it a crucial tool for risk assessment.

Chiaramonte et al.'s investigation into the applicability of the Z-Score in forecasting bank failures revealed that it could accurately predict 76% of bank failures during the study period. This level of predictive accuracy is significant, suggesting that the Z-Score, as a standalone metric, is a robust indicator of financial distress in banks. The ability to forecast such a substantial proportion of bank failures implies that the Z-Score effectively captures the critical financial vulnerabilities that contribute to bank insolvency.

An interesting finding from their study is that the inclusion of macro-level indicators did not enhance the precision of the forecasts. This outcome suggests that the Z-Score's predictive power is primarily rooted in the micro-financial aspects of banks, such as their profitability, leverage, and earnings volatility, rather than broader economic indicators. This finding is particularly insightful as it underscores the importance of internal financial health over external economic conditions in determining a bank's risk of failure. Furthermore, Chiaramonte et al. observed that the forecast efficiency of the Z-Score in predicting bank default remained stable within a three-year forward-looking window.

This stability in predictive power over multiple years is crucial for long-term risk management and strategic planning by banks and regulators. It implies that the Z-Score can provide reliable early warnings of financial distress, allowing for timely interventions to mitigate the risk of bank failures. In conclusion, the research by Chiaramonte et al. highlights the significance of the Z-Score as a predictive tool in the banking sector. Its ability to accurately forecast a high

Public

percentage of bank failures, and its consistency over a multi-year horizon, make it an invaluable metric for banks and financial regulators in monitoring and managing financial risk. This study contributes to a broader understanding of financial stability and the tools available for safeguarding the banking sector against insolvency risks.

Critique and Summary: Chiaramonte et al.’s reliance on the Z-Score is well-founded, but I believe their study might be strengthened by integrating other financial health indicators to assess if a combined metric could improve prediction accuracy. Additionally, the study’s exclusion of macro-level indicators warrants further exploration to confirm these findings across different economic conditions. Future research should consider these aspects to provide a more holistic view of bank failure risks and enhance the robustness of predictive models.

The study conducted by Ekinci and Erdal in 2017 provides a comprehensive empirical analysis of bank failure prediction, focusing on the Turkish banking sector from 1997 to 2001. This era is particularly significant for the Turkish banking industry, marked by considerable economic turbulence and a series of bank failures. Ekinci and Erdal's work is notable for its application of advanced machine learning techniques to predict bank failures, a task crucial for financial stability and regulatory oversight.

Their dataset comprised 37 commercial banks in Turkey, of which 20 were healthy and 17 failed. This selection provided a balanced representation of both successful and unsuccessful banks, offering fertile ground for testing the efficacy of various predictive models. The authors employed several machines learning techniques, including Logistic Regression, J48 (a type of decision tree algorithm), and Voted Perceptron. Each of these models has distinct characteristics: Logistic Regression is a widely used statistical method for binary classification problems; J48 is known for its interpretability and ease of use in classification tasks; and Voted Perceptron is a type of neural network algorithm suitable for large and complex datasets.

Crucially, Ekinci and Erdal did not limit their analysis to these base learners; they also experimented with different Hybrid Ensemble models. Hybrid Ensembles combine multiple machine learning techniques to leverage their strengths and compensate for their weaknesses, often resulting in improved predictive performance. This approach is particularly pertinent in the context of bank failure prediction, where the complexity and variability of financial data can make it challenging for a single model to capture all the relevant patterns and relationships.

The empirical findings of their study were quite illuminating. They indicated that hybrid ensemble machine learning models outperform traditional base and ensemble models in predicting bank failures. This superiority is likely due to the hybrid ensembles' ability to integrate diverse perspectives and approaches in data analysis, leading to more robust and accurate predictions. According to their results, Hybrid Ensembles emerged as the most accurate forecasting method.

The most notable performance metrics included an AUC (Area Under the Curve) value of 91.5 percent for the RS-B-L ensemble, indicating a high degree of model accuracy in distinguishing between failed and non-failed banks. Additionally, the RS-B-J48 ensemble achieved the highest classification rate at 83.78 percent, underscoring its effectiveness in correctly classifying banks as either likely to fail or survive. These findings by Ekinci and Erdal significantly contribute to the body of knowledge in financial risk management, particularly in the application of machine learning techniques for bank failure prediction.

Their research underscores the potential of hybrid ensemble methods in enhancing the accuracy and reliability of predictive models in the banking sector. Such insights are invaluable for bank managers, investors, and regulators, aiding in the early identification of at-risk banks and the implementation of preventative measures to safeguard the financial system.

Critique and Summary: Ekinci and Erdal’s innovative use of hybrid ensemble models highlights the potential of advanced machine learning techniques. However, I believe their study could enhance its robustness by applying these models to different banking sectors and time periods to validate their effectiveness universally. Additionally, exploring the interpretability of these models would be beneficial for practical applications. Integrating real-world case studies and conducting comparative analyses with other advanced techniques could further strengthen their research.

Le and Viviani's 2017 study presents a comprehensive and methodologically diverse analysis of bank failures, combining traditional statistical techniques with advanced machine learning methods. Their research focuses on a substantial dataset of 3,000 U.S. banks, encompassing a mix of 1,438 failures and 1,562 active banks. This study is particularly notable for its extensive coverage and application of a variety of analytical approaches to the complex issue of bank failure prediction. The selection of both traditional and machine learning techniques allows for a broad comparison of methodologies.

The traditional techniques employed by Le and Viviani include Discriminant Analysis and Logistic Regression. Discriminant Analysis is known for its ability to classify entities based on a set of predictor variables, making it suitable for categorizing banks into 'failed' or 'active' groups. Logistic Regression, a mainstay in statistical classification problems, is adept at handling binary outcomes, such as bank failure or survival, and is known for its interpretability and robustness.

In the realm of machine learning, Le and Viviani utilized Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and k-Nearest Neighbors (k-NN). ANNs are inspired by biological neural networks and are particularly effective in capturing complex, non-linear relationships in data. SVMs are powerful in high-dimensional spaces and are known for their effectiveness in classification tasks, especially when the data is not linearly separable. k-NN is a straightforward, yet effective method based on the proximity of data points in the feature space, making it a valuable tool for pattern recognition.

The analysis was grounded in CAMEL ratios, represented by 31 financial ratios encompassing critical aspects of bank performance such as capital adequacy, asset quality, management quality, earnings, and liquidity. These ratios provide a comprehensive framework for evaluating the financial health and stability of banks. Le and Viviani's empirical findings revealed that among the various methods applied, the ANN and k-NN techniques were the most precise in predicting bank failures. This result underscores the strength of machine learning approaches in handling the complexity and nuances of financial data. The ability of ANNs to model intricate and non-linear patterns in data, along with the simplicity and effectiveness of k-NN in classification tasks, likely contributed to their superior performance in this context.

Critique and Summary: Le and Viviani’s comprehensive approach underscores the superiority of machine learning methods in handling complex financial data. However, I believe future studies could further investigate the scalability and generalizability of these methods across different banking environments and incorporate more recent data to validate their findings.

Additionally, integrating more interpretability-focused techniques could enhance the practical applicability of these models for regulatory purposes. Exploring the impact of different macroeconomic conditions and regulatory frameworks on the predictive performance of these models would also provide deeper insights into their robustness and versatility.

The study conducted by Gogas et al. in 2018 represents a significant advancement in the field of financial risk management, particularly in the context of predicting bank failures using machine learning models. Their research is centered around a substantial dataset comprising 1,443 U.S. banks, which includes 481 banks that failed during the critical period of 2007 to 2013.

This timeframe is especially relevant as it encompasses the global financial crisis, a period marked by significant financial instability and numerous bank failures. A key aspect of Gogas et al.'s approach was their implementation of a two-step feature selection procedure. This procedure is crucial in machine learning as it involves identifying the most informative and relevant variables from a potentially large set of data. Effective feature selection not only enhances the performance of the model by focusing on the most critical factors but also improves the efficiency and interpretability of the model. After identifying the most informative variables, these were then input into a Support Vector Machine (SVM) model.

SVM is a powerful machine learning algorithm known for its effectiveness in classification tasks, particularly in high-dimensional spaces. It works by finding the hyperplane that best separates the classes in the feature space. In the context of bank failure prediction, this means distinguishing between banks that are likely to fail and those that are not. The training-testing learning process that followed is a standard approach in machine learning where the model is trained on a subset of the data and then tested on a separate set to evaluate its predictive power.

This process ensures that the model is not only accurate with the data it has been trained on but can also generalize well to new, unseen data. One of the most notable outcomes of their study was the model's forecasting accuracy, which reached an impressive 99.22%. This level of accuracy is particularly remarkable and indicates that the SVM model was highly effective in capturing the patterns and signals indicative of bank failures. Furthermore, the fact that their model outperformed the well-established Ohlson’s score, a traditional statistical method for bankruptcy prediction, underscores the potential of machine learning techniques to surpass traditional models in terms of predictive accuracy.

Critique and Summary: Gogas et al. demonstrate the high accuracy of SVMs, highlighting the potential of machine learning models in predicting bank failures. However, I believe the study could be strengthened by comparing SVMs with other contemporary machine learning algorithms, such as deep learning models or ensemble methods, to evaluate their relative performance. Additionally, exploring the model’s performance in different economic conditions and regions could provide more comprehensive insights. Including a discussion on the interpretability of the SVM model could also enhance its practical applicability for regulators and financial institutions.

The study conducted by Carmona et al. in 2019 represents an innovative approach to predicting bank failures, utilizing the Extreme Gradient Boosting (XGBoost) method. This research is particularly significant as it focuses on a critical period from 2001 to 2015, which includes the global financial crisis of 2007-2008, a time of considerable upheaval and instability in the

Public

banking sector. Their analysis covers 157 U.S. national commercial banks, providing a comprehensive look at the factors contributing to bank failures during this period.

The use of XGBoost, an advanced implementation of gradient-boosted decision trees, is a notable feature of their study. XGBoost is renowned for its efficiency, accuracy, and capability to handle large datasets with numerous features. It works by building an ensemble of decision trees in a sequential manner, where each subsequent tree attempts to correct the errors made by the previous ones. This method is particularly effective for complex datasets, making it well-suited for financial data analysis.

In their model, Carmona et al. considered 30 financial ratios, encompassing a wide range of indicators that are typically used to assess a bank's financial health and performance. This comprehensive approach ensures that the model captures various dimensions of bank performance, from profitability and capital adequacy to risk management. One of the key findings of their research is that certain financial ratios are closely linked with bank failures. Specifically, lower values for retained earnings to average equity, pretax return on assets, and total risk-based capital ratio were found to be associated with an increased likelihood of bank failure.

These insights are crucial as they highlight the importance of these financial metrics in assessing a bank's vulnerability to failure. Retained earnings to average equity is a measure of a bank's ability to generate profits from its equity, while the pretax return on assets indicates the overall profitability relative to its total assets. The total risk-based capital ratio is a measure of a bank's capital adequacy, taking into account its risk exposure. The findings suggest that weaknesses in these areas are significant indicators of financial distress in banks. Furthermore, Carmona et al.'s suggestion that retained earnings should be kept within the company during stressful periods and that dividend policies should be reconsidered, offers practical guidance for bank management and policy making. This recommendation underscores the importance of internal capital generation and retention as a buffer during economic downturns, enabling banks to better withstand financial shocks.

Critique and Summary: Carmona et al.'s use of XGBoost is notable for its accuracy and practical recommendations. However, I believe their study could be further strengthened by comparing XGBoost with other advanced machine learning techniques, such as deep learning models, to evaluate their relative performance. Additionally, exploring the applicability of these findings across different banking systems and periods of economic stability could provide more comprehensive insights. Integrating macroeconomic variables into the model might also enhance the understanding of how external economic conditions impact bank failures, thereby providing a more holistic view of financial stability risks.

The study by Manthoulis et al. in 2020 represents a comprehensive approach to predicting bank failures, leveraging both statistical and machine learning methods. Their analysis, which spans a decade from 2006 to 2015, includes an extensive sample of 60,000 observations of U.S. banks. This period is particularly significant as it covers the years leading up to, during, and after the global financial crisis of 2007-2008, providing a broad and varied dataset for analysis. The use of a combination of statistical and machine learning methods in their study reflects an integrated approach to understanding bank failures.

This methodology allows for the extraction of insights from data using traditional statistical techniques while also harnessing the predictive power of advanced machine learning models.

Such a comprehensive approach is advantageous in handling the complexity and nuances of financial data. One of the key findings of their research is the role of diversification variables in improving the predictive strength of bank failure forecasting models. Diversification, in the context of banking, typically refers to the extent to which a bank is engaged in a variety of financial activities and services.

This finding is particularly important as it suggests that banks with a broader range of activities may have different risk profiles compared to less diversified banks. Importantly, the study found that diversification variables are most effective for mid to long-term forecast horizons, indicating their relevance in understanding the evolving risk profile of banks over time.

Additionally, Manthoulis et al. found that ordinal classification models provide a more nuanced description of the state of banks before failure compared to standard binary classification models. Ordinal classification models are designed to predict outcomes that have a natural order, which in this case, could represent varying degrees of financial distress leading up to a bank's failure. This approach contrasts with standard binary classification models, which typically categorize banks simply as likely to fail or not. The finding that ordinal classification models are competitive with binary models is significant. It suggests that these models can capture more detailed information about a bank's financial health and trajectory, potentially offering early warning signs and more subtle indicators of distress. This is particularly valuable for regulators, investors, and bank managers, as it can facilitate more targeted and timely interventions to prevent bank failures.

Critique and Summary: Manthoulis et al.'s integrated approach is commendable for its depth and nuance. However, I believe future studies could expand on this by incorporating real-time data and exploring the applicability of their models in different regulatory environments. The study could also benefit from a more detailed exploration of the specific diversification variables that are most predictive of bank failures. Additionally, comparing the performance of ordinal classification models with other advanced machine learning techniques could further validate their effectiveness and enhance predictive accuracy.

The study conducted by Momparler et al. in 2020 offers a unique perspective on bank failure analysis by employing fuzzy-set Qualitative Comparative Analysis (fsQCA). This approach represents a shift from traditional statistical methods, focusing instead on understanding the complex causal relationships that can lead to bank failures. Their research encompasses an extensive dataset of 157 U.S. national commercial banks over 14 years from 2001 to 2015, a timeframe that includes the global financial crisis of 2007-2008, making it particularly relevant for studying the dynamics of bank failures.

The inclusion of 30 financial ratios in their model provided a comprehensive set of variables to assess various aspects of bank performance and health. These ratios likely covered key areas such as profitability, liquidity, leverage, and asset quality, which are crucial in evaluating a bank's financial condition. The use of fsQCA is a notable aspect of their research. Unlike conventional quantitative methods, fsQCA is a methodological approach that allows for the examination of complex, interdependent relationships between factors. It is particularly adept at identifying configurations of conditions that lead to a particular outcome, in this case, bank failures.

This method is valuable in revealing how combinations of different factors, rather than single variables in isolation, contribute to bank failures. Momparler et al.'s findings highlighted that banks with a large share of non-performing loans on their balance sheets, coupled with low levels of risk coverage and capitalization, are at a high risk of failure. Non-performing loans are a critical indicator of asset quality problems and can significantly strain a bank's financial resources.

Meanwhile, low levels of risk coverage and capitalization indicate a bank’s limited ability to absorb losses and meet regulatory capital requirements, which are vital for maintaining solvency and stability. These results underscore the importance of asset quality, risk management, and capital adequacy in determining a bank's likelihood of failure. Particularly, the study suggests that banks that fail to manage loan quality effectively and maintain sufficient capital buffers are more susceptible to financial distress.

The study by Momparler et al. is significant for its methodological innovation and the insights it provides into the multifaceted nature of bank failures. The use of fsQCA offers a fresh lens through which to examine the complex interplay of financial indicators, providing valuable knowledge for bank managers, investors, and regulators. By identifying the specific combinations of financial conditions that increase the risk of failure, their research contributes to more targeted and effective strategies for preventing bank failures and enhancing financial stability.

Critique and Summary: Momparler et al.’s methodological innovation provides valuable insights, but I believe their study could be enhanced by comparing fsQCA with other qualitative and quantitative methods. Additionally, exploring how these findings translate to different banking systems and periods could offer more comprehensive risk management strategies. Incorporating more recent data and conducting comparative analyses with advanced machine learning techniques could further validate their findings and provide a broader perspective on bank failure risks.

The study by Petropoulos et al. in 2020 is a noteworthy contribution to the field of financial risk management, focusing on the prediction of bank insolvencies using various modeling techniques. Their analysis is based on data from U.S. financial institutions, spanning a significant period from 2008 to 2014. This timeframe is particularly relevant as it covers the years during and after the global financial crisis, a period characterized by heightened financial instability and a notable increase in bank failures.

The researchers centered their analysis on the CAMELS framework, which is a widely recognized regulatory tool used to assess the soundness of financial institutions. The CAMELS rating system evaluates six components: Capital adequacy, Asset quality, Management quality, Earnings, Liquidity, and Sensitivity to market risk. Each of these components offers critical insights into different aspects of a bank's operational and financial health. In their pursuit to forecast bank insolvencies, Petropoulos et al. employed a range of modeling techniques, each with unique strengths and capabilities. These included Logistic Regression, a classical statistical method for binary classification; Linear Discriminant Analysis, which is used to find a linear combination of features that separates two or more classes; Support Vector Machines, effective for high-dimensional data; and Neural Networks, known for their ability to model complex non-linear relationships in data.

A key finding of their study is that Random Forests consistently outperformed the other models in predicting bank insolvencies. Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes of the individual trees. Its superiority in this context can be attributed to several factors. Firstly, Random Forests are less prone to overfitting compared to more complex models like Neural Networks. Secondly, they can handle a large number of features and their interactions effectively, which is crucial in analyzing the multifaceted CAMELS data. Lastly, Random Forests provide a measure of feature importance, which can offer valuable insights into which aspects of the CAMELS ratings are most predictive of insolvency.

Critique and Summary: Petropoulos et al.'s study demonstrates the effectiveness of Random Forests in predicting bank insolvencies, highlighting their robustness and interpretability. However, I believe the study could be further enhanced by comparing Random Forests with other ensemble methods, such as Gradient Boosting Machines (GBMs) and XGBoost, to evaluate their relative performance. Additionally, exploring the applicability of these findings across different financial environments and regulatory frameworks could provide more comprehensive insights. Integrating more recent data and conducting a comparative analysis with advanced deep learning techniques might also enhance the predictive power and generalizability of the models.

In their empirical investigation, Beutel et al. (2019) scrutinize the efficacy of machine learning techniques in predicting systemic banking crises, contrasting these methods with the traditional logit approach. The study encompasses a 45-year timeframe, focusing predominantly on advanced economies. This comprehensive analysis, set against the backdrop of a rapidly evolving financial landscape, is pivotal in understanding the potential and limitations of emerging predictive technologies in the banking sector.

The methodology adopted by the authors involves a meticulous comparison between a benchmark logit model and various machine learning approaches that have gained traction in recent academic discourse. A key aspect of their methodology is the robustness of the comparative framework, which remains consistent across different performance metrics, crisis definitions, preference parameters, sample lengths, and variable sets. This methodological rigor enhances the reliability and applicability of their findings.

The results of the study are enlightening. Despite the high in-sample fit achieved by machine learning models, the authors discover that these models are outclassed by the logit approach in recursive out-of-sample evaluations. This finding is significant, as it challenges the prevailing notion of the superiority of machine learning in predictive accuracy, especially in the context of banking crises. It suggests that while machine learning models have made strides in fitting historical data, their predictive power in real-world scenarios, particularly in anticipating systemic banking crises, is yet to be optimized.

Further, the study identifies credit expansions, asset price booms, and external imbalances as key indicators in the logit models, aligning these findings with established economic theories. The ability of logit models to effectively use this information for out-of-sample predictions, including the potential to have forecasted the 2007/2008 financial crisis for many countries, underscores their current advantage over machine learning models in this domain.

This research, therefore, serves as a critical contribution to the field of financial risk management and forecasting. It not only highlights the current limitations of machine learning models in predicting systemic banking crises but also underscores the continued relevance and efficiency of traditional logit models.

Public

Critique and Summary: Beutel et al.'s study offers valuable insights into the comparative performance of traditional logit models and machine learning techniques. However, I believe that future research could benefit from exploring hybrid models that combine the strengths of both approaches. Additionally, investigating the reasons behind the superior out-of-sample performance of logit models could provide deeper insights into improving machine learning models. Expanding the analysis to include emerging economies and different time periods might also enhance the generalizability of the findings. Integrating more advanced machine learning algorithms and refining their feature selection processes could potentially bridge the current performance gap identified in the study.

Antunes et al. (2018) embarked on an extensive exploration of early warning systems for banking crises, focusing on a dataset covering multiple countries over four decades (1970Q1 to 2010Q4). Their research aims to enhance the predictive tools available to policymakers by delving into the dynamics and non-linearities inherent in banking crises. They employ dynamic probit models to develop a common vulnerability indicator with early warning properties, assessing both in-sample and out-of-sample forecast performance.

The authors highlight the complexities of predicting banking crises, acknowledging the diverse causes and consequences of these rare yet impactful events. They reference seminal works in the literature that have primarily concentrated on currency crises, particularly in emerging markets, and the often-concurrent nature of currency and banking crises. The more pronounced and prolonged economic effects of financial crises involving banking system disruptions are underscored, with references to notable studies that document the long-term impacts of such crises.

In their methodology, Antunes et al. investigate the commonalities of European systemic banking crises, placing particular emphasis on dynamic and exuberant behaviors. They review various methodologies used in forecasting banking crises, noting the lack of explicit modeling of dynamic and exuberance dimensions in the existing literature. The study builds on the institutional memory hypothesis, evidencing the deterioration in risk assessment capabilities of loan officers through the credit cycle.

In conclusion, Antunes et al. (2018) demonstrate that dynamic probit models significantly enhance the forecast accuracy of early warning systems. Their work extends previous literature by incorporating exuberance indicators, offering insights into the signaling properties embedded in the distribution of explanatory variables, and improving the performance of dynamic models as early warning tools.

Critique and Summary: Antunes et al. (2018) provide a significant contribution to the development of early warning systems for banking crises. Their use of dynamic probit models introduces a novel approach that captures temporal dynamics and systemic vulnerabilities more effectively. However, the generalizability of their findings might benefit from further validation across different global contexts, particularly in less studied emerging economies. Additionally, integrating newer data post-2010, covering recent financial disturbances, could further validate the robustness and adaptability of their models. Future research could also explore the integration of machine learning techniques to compare and possibly enhance the predictive power of traditional econometric models like the dynamic probit model used in their study.

Plessis (2022) introduces a novel approach using a Qualitative Vector Autoregression (Qual VAR) to generate a continuous banking crisis indicator. The model, developed from a latent

Public

variable using a Markov Chain Monte Carlo algorithm, assesses four decades of banking crises, accounting for the evolutionary nature of precursors in banking crises, including periodic, regional, and developmental effects across a representative sample of countries.

The study's findings underscore the dynamics of banking crises. The forecast error variance decomposition reveals that banking sector variables account for nearly half of the total variation, the external sector for a third, and the real sector for a fifth. This highlights the significant role of the banking sector in forecasting models. The Qual VAR model's effectiveness in providing early warning signals up to 12 months before a crisis, outperforming a probit model in out-of-sample forecasting, is pivotal for banking oversight departments and policymakers, enabling them to respond more effectively and timely to banking crises.

The research places its significance against the backdrop of the global financial crisis, referencing seminal works by Reinhart and Rogoff (2009) and Laeven and Valencia (2018), who discuss the profound socio-economic and political impacts of banking crises with substantial fiscal costs for both developed and emerging markets.

Plessis delves into various precursors of banking crises, ranging from macroeconomic imbalances to deregulation and inadequate regulatory oversight. The study highlights how certain economies like Japan, Australia, Mexico, Thailand, and Poland remained resilient during the global financial crisis, whereas others succumbed due to banking sector weaknesses. This disparity underscores the need for crisis prevention and the importance of robust econometric models and methodologies in identifying vulnerabilities in the banking sector.

The paper's contribution to the literature on banking sector vulnerabilities and crisis prevention is significant. The Qual VAR approach integrates qualitative information and focuses on a binary choice variable to detect the likelihood of a banking crisis or a tranquil period. This method allows for an explicit consideration of the endogenous nature of banking crises and provides valuable insights into the dynamic effects of a banking crisis on the economy.

The study's comprehensive dataset covers 40 years of banking crises, with a total of 34 crises in 18 countries between 1971 and 2016. The analysis includes a range of explanatory variables representing the real, banking, and external sectors. The results demonstrate the predictive power of the Qual VAR model, suggesting its utility in the continuous monitoring and evaluation of banking sector vulnerabilities, thereby aiding macroprudential policymakers and oversight departments in their efforts to preempt and mitigate banking crises.

Critique and Summary: Plessis (2022) provides a sophisticated tool in the form of the Qual VAR model, which advances the field of banking crisis prediction. However, future research could enhance this study by examining the integration of additional predictive variables that might affect the stability of the banking sector, such as political risk and technological innovations. Also, testing the model's applicability in real-time scenarios across different regulatory environments would further validate its effectiveness and enhance its practical application in risk management and policy formulation.

Public

## Owner's Contribution

This research contributes to the evolving field of bank failure prediction by addressing key gaps in the literature, particularly in the comparative evaluation of machine learning models, handling of imbalanced datasets, and the trade-offs between model complexity and interpretability. By constructing a comprehensive dataset and systematically analyzing multiple machine learning techniques, this study enhances the predictive power and practical applicability of bank failure prediction models.

### 1. Novel Dataset Construction and Methodological Framework

A critical contribution of this research is the development of a detailed dataset derived from the Federal Deposit Insurance Corporation (FDIC), covering the years 2007–2013. Unlike previous studies that often rely on aggregated data, this research meticulously tracks individual bank performance using FDIC bank identifiers, ensuring a precise and granular representation of financial conditions leading up to failure. The use of 2009 as the reference year for non-failed banks enhances the robustness of the analysis by capturing financial conditions at the peak of the crisis.

This dataset includes 123,786 data points across 26 CAMEL-based financial ratios, enabling a comprehensive assessment of bank stability indicators, such as capital adequacy, asset quality, management efficiency, earnings stability, and liquidity. The extensive manual effort required to extract and process these data ensures its accuracy and reliability for predictive modeling.

### 2. Comprehensive Machine Learning Model Evaluation

This study systematically evaluates 15 machine learning models, ranging from traditional classifiers (e.g., Logistic Regression, Naïve Bayes) to advanced ensemble techniques (e.g., Random Forest, CS-Forest, and Bagging). The findings indicate that ensemble methods consistently outperform traditional statistical models, with CS-Forest achieving a classification accuracy of 97.3% and a ROC Area of 0.987. These results reinforce the conclusions of Petropoulos et al. (2020), who demonstrated the superiority of Random Forest in financial risk prediction, while extending the analysis by introducing CS-Forest, which achieved a recall of 0.897—making it particularly effective in capturing true positive failures.

This research also addresses a gap highlighted by Ekinci and Erdal (2017) regarding the lack of comparative analysis across multiple machine learning models. Unlike prior studies that focus on a single or limited set of techniques, this study provides a more holistic evaluation, offering insights into the strengths and weaknesses of different classification paradigms for financial risk assessment.

### 3. Handling of Imbalanced Datasets: Cost-Sensitive Learning

One persistent challenge in bank failure prediction is class imbalance, where failed banks constitute a significantly smaller portion of the dataset compared to non-failed banks. Traditional models such as Logistic Regression and Naïve Bayes tend to misclassify failed banks due to their bias toward the majority class (Chiaramonte et al., 2016). This study systematically addresses this issue by implementing cost-sensitive learning techniques, prioritizing the correct identification of bank failures.

Public

CS-Forest demonstrated the highest recall (0.897), making it particularly useful for early warning systems where missing a failed bank could lead to systemic financial consequences. While ensemble methods like Bagging and Random Forest achieved high classification accuracy, their recall was lower than CS-Forest, reinforcing the argument made by Ekinci and Erdal (2017) that cost-sensitive methods mitigate Type II errors (false negatives) more effectively than traditional ensemble models.

4. Trade-Off Between Model Complexity and Interpretability

The ongoing debate in financial risk modeling revolves around whether the increased predictive accuracy of complex machine learning models justifies their lack of transparency (Beutel et al., 2019). While traditional models such as Logistic Regression offer interpretability, they lag behind ensemble-based approaches in predictive accuracy. This study systematically quantifies this trade-off by demonstrating that ensemble methods, particularly CS-Forest and Random Forest, achieve both high predictive accuracy and operational efficiency.

These findings align with Berger and Bouwman (2012), who emphasize that while complex models can enhance risk assessment, their black-box nature remains a challenge for regulatory adoption. By incorporating explainability techniques, this study bridges the gap between model interpretability and predictive strength, reinforcing the need for hybrid approaches that balance both aspects.

5. Expanding the Dataset for Greater Generalizability

One of the key strengths of this research is its large-scale dataset, comprising 4,759 banking institutions, including 447 failed banks from the 2008 financial crisis. This dataset surpasses those used in prior studies: Torna and DeYoung (2012) excluded banks over $100 billion in assets, limiting their findings to mid-sized banks. Lu and Whidbee (2013) analyzed 6,236 banks but lacked a comprehensive set of CAMEL indicators crucial for financial stability assessment. Chiaramonte et al. (2016) emphasized the importance of using detailed financial indicators, an approach this study adopts by incorporating 26 CAMEL ratios for a more granular analysis. Additionally, the longitudinal nature of this dataset allows for temporal testing of predictive stability across different crisis periods, improving the generalizability of findings for future financial crises.

6. Real-World Applicability: Machine Learning for Early Warning Systems

While existing literature primarily focuses on model accuracy in controlled settings, this research extends beyond theoretical evaluations by emphasizing the real-world applicability of machine learning for financial risk management. The findings provide empirical evidence supporting the deployment of ensemble learning models in early warning systems for regulators and banking institutions.

In summary, this research contributes significantly to the field of bank failure prediction by constructing a high-resolution dataset, implementing a novel weighted success scoring method for model evaluation, and identifying key financial ratios that drive bank insolvencies. By benchmarking multiple machine learning models, addressing class imbalance, and emphasizing real-world applicability, this study provides actionable insights for financial regulators, policymakers, and banking institutions.

Public

# Chapter 3: Data Collection and Sampling

## 3.1 Study Sample

In this section, I provide a detailed year-by-year analysis of bank failures, utilizing data obtained from the Federal Deposit Insurance Corporation (FDIC). This analysis forms the empirical backbone of my research, enabling a comprehensive examination of the financial conditions that preceded bank failures during the 2008 financial crisis and the years that followed. The construction of this dataset was a meticulous process, designed to capture the most accurate representation of the financial health of both failed and non-failed banks.

The process began with an in-depth exploration of the FDIC website, where I systematically downloaded financial performance ratios for each bank, year by year, from 2007 to 2013. My objective was twofold: to identify banks that failed during the crisis and to capture the financial conditions of banks that remained operational. For each failed bank, I tracked their bank identification numbers and used these unique identifiers to trace their financial data prior to failure. This allowed me to isolate the specific year in which each bank collapsed and focus on the financial ratios from their last reporting period. For instance, if a bank failed in June 2009, I extracted its financial ratios as of December 31, 2008, the last complete dataset before its failure. This strategy was crucial in ensuring that I captured the final financial state of each failed bank before its collapse, providing a snapshot of the conditions that led to bankruptcy.

For non-failed or 'healthy' banks, I adopted 2009 as the reference year for their financial performance data. The rationale for choosing 2009, as opposed to 2008 or 2010, lies in the fact that 2009 represented the peak of financial instability, with the most significant impacts on bank balance sheets becoming evident in that year's reporting. By selecting 2009, I ensured that I captured the financial conditions of non-failed banks at their most vulnerable point, offering a robust comparison to the failed banks. Had I chosen 2008, the data would not have reflected the full impact of the crisis, and by 2010, many banks had begun to recover, making 2009 the most critical year for this analysis. This decision increased the difficulty of forecasting, as it presented the worst-case financial conditions for banks that remained operational, allowing the predictive models to be tested under the most challenging circumstances.

The data extraction process was highly labor-intensive, requiring me to manually gather thousands of data points from the FDIC website, bank by bank. Each year, I reviewed the list of failed banks, identified by their FDIC certificates, and retrieved their respective financial ratios. This process was repeated for every year between 2007 and 2013, ensuring that I had a comprehensive dataset that covered the full span of the financial crisis and its aftermath. For each bank, I recorded a set of 26 CAMEL ratios (Capital adequacy, Asset quality, Management quality, Earnings, and Liquidity), which serve as the primary variables in my predictive models.

The construction of this dataset was one of the most time-consuming aspects of my research. The manual processing of such a large volume of data required not only careful attention to detail but also a systematic approach to ensure accuracy and consistency across all entries. In total, this process yielded 123,786 data points, forming the foundation for the machine learning models applied in this thesis. These data points were critical for creating a balanced and comprehensive dataset that could effectively train and test the predictive models.

By utilizing CAMEL ratios as the key indicators, the dataset offered a detailed view of each bank's financial health, covering aspects such as capital adequacy, loan performance, earnings

Public

stability, liquidity, and management efficiency. These ratios were selected for their relevance in assessing bank stability, particularly during periods of economic stress like the 2008 financial crisis. Each of these indicators played a crucial role in the model's ability to distinguish between banks that would ultimately fail and those that would remain operational.

The next step involved processing this dataset for analysis using WEKA, an open-source data mining software known for its flexibility and robust machine learning capabilities. WEKA was chosen not only for its accessibility but also for its strong support for various classification algorithms, which aligned with the goals of this thesis. By feeding the CAMEL ratios into the machine learning models, I was able to conduct binary classification, distinguishing between failed and non-failed banks based on their financial health leading up to and during the crisis.

In summary, the process of gathering and preparing this dataset was a rigorous, detailed task, but it was essential for creating a reliable foundation for the predictive analysis that follows. By meticulously tracking the financial health of both failed and non-failed banks over a period of several years, this dataset serves as a vital tool for testing the accuracy and effectiveness of the machine learning models developed in this research. The use of 2009 as the base year for non-failed banks, combined with the detailed analysis of failed banks' last reporting periods, ensures that the dataset captures the most relevant and challenging financial conditions, providing a robust test for the predictive models.

## 3.2. Ratios used in this Research:

1. Performance Ratios

| **Key** | **Variable** |
|---|---|
| P1 | Yield on Earning Assets (YEA) |
| P2 | Cost of Funding Earning Assets (COFEA) |
| P3 | Net Interest Margin (NIM) |
| P4 | Non-Interest Income to Average Assets (noniiay) |
| P5 | Non-Interest Expenses to Average Assets (nonixay) |
| P6 | Credit Loss Provision to Net Charge-Offs (elnatry) |
| P7 | Net Operating Income to Assets (NOIA) (noijy) |
| P8 | Pretax Return on Assets (roaptx) |
| P9 | Return on Equity (ROE) (roeinjr) |
| P10 | Net Charge-Offs to Loans and Leases (ntlnlsr) |
| P11 | Earnings Coverage of Net Loan Charge-Offs (iderncvr) |
| P12 | Efficiency Ratio (eeffr) |
| P13 | Assets per Employee (astempm) |

Public

2. Condition Ratios

| Key | Variable |
|---|---|
| C1 | Earning Assets to Total Assets (ernastr) |
| C2 | Loss Allowance to Loans and Leases (lnatresr) |
| C3 | Loss Allowance to Noncurrent Loans and Leases (lnresncr) |
| C4 | Noncurrent Loans to Loans (nclnlsr) |
| C5 | Net Loans and Leases to Total Assets (lnlsntv) |
| C6 | Insured Deposits in Total Deposit Liabilities (idlncorr) |
| C7 | Domestic Deposits to Total Assets (depdastr) |
| C8 | Equity Capital to Assets (eqv) |
| C9 | Total Risk-Based Capital Ratio (rbcrwaj) |

3. Other

| **Key** | **Variable** |
|---|---|
| O1 | Average Total Assets (asset5) |
| O2 | Average Earning Assets (ernast5) |
| O3 | Average Equity (eq5) |
| O4 | Average Total Loans and Leases (lnlsgr5) |

A pre-screening and diagnostic assessment of the data was conducted to identify potential multicollinearity issues among the variables. As a result of this initial analysis, several indicators—Loan and lease loss provision to assets, Return on assets (ROA), Return on equity (ROE), Noncurrent assets plus other real estate owned to assets, Net loans and leases to deposits, Core capital (leverage) ratio, and Tier 1 risk-based capital ratio—were omitted due to multicollinearity. The final list of indicators provided above reflects these adjustments, ensuring that only non-redundant variables are included for a more accurate and interpretable analysis.

Public

# Chapter 4: Methodology

## 4.1 Introduction to Methodology

This chapter outlines the methodological framework employed in this research to address the problem of predicting bank failures using both traditional statistical techniques and advanced machine learning (ML) classification algorithms. The methodology has been carefully designed to ensure rigor, reproducibility, and alignment with the objectives of this dissertation. It encompasses research philosophy, the overall approach, and the methods adopted for data collection, preprocessing, model development, and evaluation.

The financial sector plays a critical role in maintaining economic stability, and bank failures can have far-reaching consequences for depositors, investors, regulators, and the wider economy. As the 2008 Global Financial Crisis (GFC) revealed, traditional risk assessment models often failed to identify vulnerable banks in a timely manner (Mian & Sufi, 2009; Shiller, 2008). This reality has prompted the development and adoption of more sophisticated predictive models, particularly those leveraging machine learning techniques, which have demonstrated superior capabilities in handling complex, non-linear, and high-dimensional data (Petropoulos et al., 2020).

The overarching aim of this research is to evaluate whether machine learning algorithms, including ensemble and cost-sensitive models, outperform traditional models like logistic regression in predicting bank failures. This study also investigates how the trade-offs between complexity and interpretability influence model selection and decision-making in financial risk analysis.

This chapter begins by establishing the research philosophy and approach underpinning the study, followed by a detailed discussion of the research design, data sources, and methodological choices made to address the research questions and test the hypotheses formulated in earlier chapters. By systematically detailing the methodology, this chapter ensures transparency in the research process and provides a foundation for the interpretation and validation of the findings presented in subsequent chapters.

## 4.2 Research Philosophy and Approach

The research philosophy and approach define the ontological, epistemological, and methodological stance of this study. Given the quantitative nature of the research, this dissertation adopts a positivist paradigm, which assumes that reality is objective, measurable, and independent of the researcher (Saunders et al., 2019). Under the positivist worldview, the role of the researcher is to test hypotheses derived from theory through empirical observation, employing quantitative data and statistical analysis to explain and predict phenomena.

A deductive research approach has been employed, wherein hypotheses regarding the predictive performance of different classification models are formulated based on an extensive review of existing literature (e.g., Cole & White, 2012; Gogas et al., 2018; Carmona et al., 2019) and then empirically tested using historical data. The deductive approach is appropriate for this study because it facilitates the testing of predefined theories and assumptions, such as whether machine learning models offer superior predictive performance compared to traditional models, and whether model complexity correlates with predictive power.

Public

This research adopts a quantitative methodology, using secondary data obtained from the Federal Deposit Insurance Corporation (FDIC), covering bank financial ratios and failure statuses from 2007 to 2013. The choice of a quantitative approach is justified by the study's objective focus on empirical testing, measurement, and generalization rather than subjective interpretation or theory generation.

The research strategy is explanatory and comparative, aiming to explain why and how different predictive models perform in the context of bank failure prediction and to compare their performance systematically. The study also incorporates elements of applied research, as it addresses real-world problems faced by regulators, bank supervisors, and financial institutions, offering practical solutions and actionable insights for the design of early warning systems.

In summary, the methodological choices are driven by the research objectives, which are inherently empirical, objective, and predictive in nature. By adopting a positivist philosophy and a deductive, quantitative approach, this study ensures the scientific rigor necessary to make valid inferences about bank failure prediction models, thereby contributing to both academic knowledge and practical policy-making in the domain of financial stability.

## 4.3 Machine Learning in Bank Failure Prediction

The increasing complexity of financial markets and the multifaceted nature of bank operations have rendered traditional statistical methods insufficient for accurately predicting bank failures. In this context, machine learning (ML) has emerged as a promising solution, offering sophisticated algorithms capable of capturing complex, non-linear relationships and interactions among variables (Hastie, Tibshirani, & Friedman, 2009). Unlike conventional statistical models, machine learning algorithms can learn patterns directly from data without relying on restrictive assumptions about data distribution, making them especially valuable in financial risk modeling (Petropoulos et al., 2020).

Machine learning refers to a subset of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to perform specific tasks without using explicit instructions, relying instead on patterns and inference. In predictive modeling, machine learning algorithms learn from historical data to make predictions or classifications about new, unseen data points (Mitchell, 1997).

Types of Machine Learning Algorithms Used in This Study

This research applies to a variety of machine learning algorithms, ranging from basic classifiers to advanced ensemble methods, each offering distinct advantages in the context of bank failure prediction. The selection of these algorithms is informed by their widespread adoption in the field of financial risk modeling, their documented performance in previous empirical studies, and their theoretical capacity to handle high-dimensional, complex, and imbalanced datasets.

Naïve Bayes (NB)

A simple probabilistic classifier based on Bayes' theorem with strong assumptions of conditional independence among predictors. Despite its simplicity, Naïve Bayes has demonstrated surprising effectiveness in high-dimensional financial datasets, particularly when dealing with large numbers of features and relatively simple data structures (Rish, 2001).

Public

### Logistic Regression (LR)

While often classified as a traditional statistical method, logistic regression remains widely used in machine learning classification tasks due to its interpretability, simplicity, and ability to provide probabilistic outputs. It is particularly valued in financial risk modeling for its clear interpretability and ease of communication to stakeholders (Hosmer et al., 2013).

### Decision Tree-Based Ensemble Methods

### Bagging (Bootstrap Aggregating)

Introduced by Breiman (1996), bagging reduces model variance and enhances stability by averaging predictions from multiple decision trees trained on different bootstrap samples of the dataset. Bagging is particularly effective for algorithms like decision trees that are prone to overfitting.

### Random Forest (RF)

An extension of bagging, Random Forest introduces random feature selection at each split in the decision tree construction process, which increases model diversity and enhances predictive performance. RF is widely recognized for its robustness, scalability, and ability to handle high-dimensional data (Breiman, 2001).

### Regularized Random Forest (RRF)

RRF incorporates regularization to penalize the selection of less informative features, which helps in reducing overfitting and enhancing the model's ability to perform feature selection in high-dimensional settings. This makes RRF particularly useful in financial datasets with many correlated variables (Deng & Runger, 2012).

### Cost-Sensitive Forest (CS-Forest)

CS-Forest extends Random Forest by integrating cost-sensitive learning techniques, addressing the asymmetric costs associated with false negatives and false positives. This model is especially effective in reducing Type II errors, which is critical in early warning systems for bank failure prediction (Ling & Sheng, 2008).

### Boosting Methods

### AdaBoost

A pioneering boosting algorithm that sequentially combines weak learners (typically shallow decision trees) to form a strong classifier. AdaBoost focuses subsequent learners on the hard-to-classify instances, improving classification performance on challenging datasets (Freund & Schapire, 1996).

LogitBoost

An adaptation of boosting that optimizes logistic regression models in a sequential manner. LogitBoost enhances classification accuracy and is particularly effective when probability estimation is a priority in binary classification tasks like bank failure prediction (Friedman et al., 2000).

GAMBoost

Combines Generalized Additive Models (GAMs) with the boosting framework, allowing the model to capture non-linear relationships while maintaining interpretability. GAMBoost is advantageous in financial contexts where the relationship between predictors and outcomes is complex but transparency remains essential (Bühlmann & Hothorn, 2007).

Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced, scalable implementation of gradient boosting decision trees (GBDT). It introduces regularization to prevent overfitting, handles missing data, and efficiently processes high-dimensional and imbalanced datasets (Chen & Guestrin, 2016). XGBoost is highly regarded for its predictive power and has consistently outperformed other algorithms in bank failure prediction and credit risk assessment studies (Carmona et al., 2019; Petropoulos et al., 2020).

Neural Network Models

Multilayer Perceptron (MLP)

A feedforward artificial neural network with one or more hidden layers capable of modeling complex non-linear relationships. MLPs are particularly effective in credit scoring and financial risk prediction tasks where the relationships between inputs and outputs are non-linear and high-dimensional (Haykin, 1998).

Instance-Based Learning

KStar

An instance-based learner that uses an entropy-based distance measure to compare instances. KStar is well-suited for detecting subtle patterns and non-linear relationships in bank failure prediction, where data irregularities and exceptions often exist (Cleary & Trigg, 1995).

## 4.4 Machine Learning Performance Metrics

The evaluation of machine learning models requires the use of reliable and informative performance metrics that reflect the predictive capabilities and limitations of each algorithm. Given the imbalanced nature of bank failure prediction datasets—where failed banks constitute a small proportion of the total population, traditional metrics like accuracy can be misleading (Provost, Fawcett, & Kohavi, 1998). Therefore, a comprehensive evaluation framework is adopted in this study, incorporating both threshold-dependent and threshold-independent performance metrics.

Public

### 4.4.1 Accuracy

Accuracy is the proportion of correctly predicted instances (both positive and negative) over the total instances:

Accuracy=TP+TN / TP+TN+FP+FN

Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

Although accuracy provides a basic measure of performance, it can be misleading in imbalanced datasets. In the prediction of bank failure, where non-failed banks dominate the sample, high accuracy can mask poor detection of actual failures (Hand, 2009).

### 4.4.2 Precision (Positive Predictive Value)

Precision measures the proportion of correctly predicted positive instances (failed banks) among all instances predicted as positive:

Precision=TP / TP+FP
High precision indicates a low false positive rate, which is crucial when wrongly flagging healthy banks can have significant consequences (Powers, 2011).

### 4.4.3 Recall (Sensitivity, True Positive Rate)

Recall assesses the proportion of actual positives (failed banks) correctly identified by the model:

Recall=TP / TP+FN

In early warning systems, recall is of paramount importance, as failing to detect a troubled bank (false negative) can lead to systemic risk (Gogas et al., 2018).

### 4.4.4 Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate at various threshold settings. The AUC quantifies the overall ability of the model to discriminate between positive and negative classes:

- An AUC of 0.5 indicates random guessing.
- An AUC of 1.0 indicates a perfect model (Bradley, 1997).

AUC-ROC is a threshold-independent metric, making it suitable for comparing classifiers across different operating points (Fawcett, 2006).

Public

### 4.4.5 Confusion Matrix

The confusion matrix presents the complete breakdown of prediction outcomes:

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Positive | True Positive (TP) | False Negative (FN) |
| Actual Negative | False Positive (FP) | True Negative (TN) |

It provides a clear visualization of Type I Errors (False Positives) and Type II Errors (False Negatives), aiding in the interpretation of model trade-offs (Stehman, 1997).

### 4.4.6 Cohen's Kappa

Cohen's Kappa accounts for the possibility of correct predictions occurring by chance:

Kappa= p0−pe / 1−pe

Where:

- P0: Observed agreement
- pe : Expected agreement by chance

A Kappa value closer to 1 indicates substantial agreement, while values closer to 0 suggest random performance (Cohen, 1960; Landis & Koch, 1977).

## 4.5 Rationale for Metric Selection in Bank Failure Prediction

The choice of performance metrics in this study reflects the high-stakes nature of bank failure prediction. Recall is prioritized, given the severe consequences of undetected bank distress. However, precision remains important to avoid unnecessary interventions in stable banks, while AUC-ROC enables comprehensive comparison across models.

Additionally, Cohen's Kappa and Cost-Sensitive Evaluation provides further insight into model reliability and economic relevance, addressing regulatory needs for accurate and actionable early warning systems (Petropoulos et al., 2020).

## 4.6 Machine Learning Models

### Naïve Bayes Classifier

Introduction and Theoretical Foundation

The Naïve Bayes (NB) classifier is a probabilistic learning algorithm rooted in Bayesian statistics and decision theory. The model applies Bayes' Theorem, which provides a mathematical framework for updating probabilities based on new evidence (Russell & Norvig, 2021). Despite its simplicity, Naïve Bayes remains one of the most widely used classification algorithms due to its scalability, efficiency, and competitive performance, especially in high-dimensional data environments (Rish, 2001).

Public

The fundamental assumption underpinning Naïve Bayes is the conditional independence assumption, whereby each feature is assumed to contribute independently to the outcome, given the class label (Mitchell, 1997). While this assumption rarely holds in real-world scenarios, particularly in financial datasets where variables are often correlated, the classifier's performance often remains robust (Domingos & Pazzani, 1997).

### Advantages in Financial Applications

Computational Efficiency: Naïve Bayes is highly efficient, making it suitable for real-time financial risk monitoring systems (Zhang, 2004).

Robustness with High Dimensionality: The model performs well when the number of predictors is large, a common scenario in bank failure prediction (Hand & Yu, 2001).

Incremental Learning: It can be updated rapidly with new data, essential for dynamic and volatile financial markets (Tsymbal, 2004).

### Limitations

Conditional Independence Assumption: Financial indicators often exhibit multicollinearity, violating the independence assumption. This can lead to biased probability estimates, although classification accuracy may still be acceptable (Domingos & Pazzani, 1997).

Handling of Continuous Variables: The assumption of Gaussian distributions for continuous variables may not hold in skewed or non-normal financial data (Rish, 2001).

Interpretability of Probabilities: While Naïve Bayes offers probability scores, their interpretability may be limited due to the simplified underlying assumptions.

### Strategies to Overcome Limitations

Discretization of Continuous Features: Transforming continuous variables into categorical bins can circumvent the normality assumption (Fayyad & Irani, 1993).

Hybrid Approaches: Combining Naïve Bayes with other models (e.g., tree-augmented Naïve Bayes) can relax the independence assumption (Friedman, Geiger, & Goldszmidt, 1997).

The Naïve Bayes classifier, with its probabilistic foundation and computational simplicity, remains a valuable tool in financial risk modeling, particularly for preliminary analysis and benchmarking purposes. While its assumptions limit its capacity to model complex dependencies inherent in financial data, its speed, robustness, and reasonable accuracy make it a pragmatic choice for certain predictive tasks, including early warning systems for bank failure prediction.

## Logistic Regression

Introduction and Theoretical Foundation

Logistic Regression (LR) is a well-established statistical method for binary classification tasks, where the dependent variable is categorical, typically representing binary outcomes such as "success/failure" or "bank failure/non-failure". Its mathematical simplicity, interpretability, and grounding in probability theory have made it one of the most widely adopted techniques in the domain of financial risk modeling (Hosmer, Lemeshow, & Sturdivant, 2013).

Public

Logistic regression models the probability that a given observation belongs to a particular category, using a logit function to transform a linear combination of independent variables into a bounded probability between 0 and 1 (Cox, 1958). In the context of bank failure prediction, this probability is interpreted as the likelihood that a bank will fail, given its financial indicators and other explanatory variables.

### Interpretation of Model Outputs

The coefficients from logistic regression are often transformed into odds ratios, which quantify how a one-unit increase in a predictor variable affects the odds of the outcome, holding all other predictors constant (Peng, Lee, & Ingersoll, 2002). This feature makes logistic regression highly interpretable and useful for understanding the relative importance of different financial ratios in predicting bank failure.

### Applications in Financial Risk Analysis

Logistic regression has a long history of use in credit scoring (Thomas, Edelman, & Crook, 2002), default risk modeling, and bank failure prediction (Cole & White, 2012). Its simplicity and interpretability make it a preferred choice for regulatory environments, where model transparency and justification of decisions are crucial (Cleary & Hebb, 2015).

### Strengths in Financial Modeling

1. Interpretability: Logistic regression provides clear insights into the contribution of each independent variable, making it suitable for environments that require explainability (Menard, 2000).
2. Statistical Foundation: Its reliance on probability theory allows for well-established hypothesis testing, confidence intervals, and statistical inference (Hosmer et al., 2013).
3. Flexibility: Logistic regression can accommodate continuous, categorical, and binary independent variables (Kleinbaum, Klein, & Pryor, 2002).

### Limitations and Assumptions

Despite its widespread application, logistic regression has several assumptions and limitations:

1. Linearity of Log-Odds: The model assumes a linear relationship between predictors and the log-odds of the outcome. In financial datasets, where nonlinear relationships are common, this assumption may be restrictive (Le & Viviani, 2017).
2. Multicollinearity Sensitivity: Logistic regression is sensitive to multicollinearity among predictors, which can inflate the variance of coefficient estimates and affect interpretability (Hosmer et al., 2013).
3. Imbalanced Datasets: When classes are imbalanced (as in bank failure prediction), logistic regression tends to favor the majority class, potentially leading to high Type II errors (Miguéis, Camanho, & Cardoso, 2018).

Public

### Extensions and Enhancements

Several techniques have been proposed to address these limitations:

- Regularization: Ridge regression (L2 penalty) and LASSO (L1 penalty) regularization reduce overfitting and manage multicollinearity (Tibshirani, 1996).
- Non-linear extensions: Generalized Additive Models (GAMs) introduce non-linear components while preserving interpretability (Hastie & Tibshirani, 1990).
- Resampling techniques: SMOTE (Chawla et al., 2002) and cost-sensitive learning address class imbalance issues.

Logistic regression remains a cornerstone method in financial risk analysis due to its simplicity, interpretability, and grounding in statistical theory. Despite the rise of more complex machine learning algorithms, logistic regression continues to serve as a benchmark for model performance in predictive tasks such as bank failure detection.

## Extreme Gradient Boosting (XGBoost)

### Introduction and Theoretical Foundation

Extreme Gradient Boosting (XGBoost) is a highly efficient and scalable implementation of the gradient boosting framework proposed by Chen and Guestrin (2016). Since its introduction, XGBoost has become one of the most popular machine learning algorithms for structured and tabular data, often outperforming other classification techniques in competitions and practical applications (Nielsen, 2016).

XGBoost is particularly well-suited for predictive modeling in the banking sector, where datasets are typically high-dimensional and exhibit complex, non-linear interactions among variables (Carmona et al., 2019). Its ability to handle missing values, imbalance in class distributions, and heterogeneous data types make it a valuable tool for bank failure prediction and credit risk assessment.

XGBoost is an ensemble method that builds a strong classifier by combining multiple weak learners, typically decision trees, in a stage-wise additive manner. The algorithm optimizes a differentiable loss function using gradient descent, adding new models to correct the residuals (errors) of previous models (Chen & Guestrin, 2016).

XGBoost introduces several improvements over traditional gradient boosting:

1. Regularization: XGBoost includes L1 (Lasso) and L2 (Ridge) regularization terms in its objective function to penalize model complexity, reducing the risk of overfitting.
2. Parallelization: It leverages parallel and distributed computing, allowing faster model training on large datasets.
3. Sparse Data Handling: XGBoost can automatically learn the best way to handle missing values in the dataset, which is common in financial data.
4. Weighted Quantile Sketch: This algorithm efficiently estimates the optimal split points in decision trees, improving computational efficiency.

Public

### Application for Financial Risk and Bank Failure Prediction

XGBoost has been applied extensively in financial risk modeling due to its high accuracy, flexibility, and interpretability through feature importance scores (Carmona et al., 2019; Beutel et al., 2019). In the prediction of bank failure, XGBoost effectively models the non-linear relationships between financial indicators, macroeconomic factors, and the probability of failure.

Researchers have demonstrated that XGBoost outperforms other ensemble models, such as Random Forest and AdaBoost, particularly in imbalanced classification settings, which are typical in bank failure datasets where the proportion of failed banks is relatively small (Petropoulos et al., 2020).

### Advantages

- High Predictive Performance: XGBoost consistently achieves state-of-the-art results on structured data tasks, including credit scoring, fraud detection, and bank failure prediction.
- Efficient Handling of Missing Data: XGBoost automatically learns optimal missing value handling, which is essential for datasets with incomplete financial information.
- Regularization to Reduce Overfitting: The inclusion of regularization terms enhances model generalization, making it robust in volatile financial environments.
- Feature Importance Measures: XGBoost provides gain, cover, and frequency-based feature importance measures, facilitating model interpretation and regulatory compliance (Chen & Guestrin, 2016).

### Challenges and Considerations

- Computational Complexity: Despite parallelization, XGBoost can be computationally demanding on very large datasets, especially during hyperparameter tuning.
- Interpretability: While more interpretable than deep neural networks, XGBoost models are still considered less transparent than simpler models like logistic regression, which can be problematic in regulatory contexts (Cleary & Hebb, 2015).
- Sensitivity to Hyperparameters: XGBoost’s performance is sensitive to hyperparameter settings, requiring careful tuning (Nielsen, 2016).

### Conclusion

XGBoost represents a significant advancement in machine learning for financial risk prediction, offering a balance of predictive power, efficiency, and flexibility. Its application in bank failure prediction enables the development of accurate early warning systems, supporting regulators and financial institutions in proactively managing systemic risk.

## Bagging (Bootstrap Aggregating)

### Introduction and Theoretical Foundation

Bagging, short for Bootstrap Aggregating, is an ensemble learning method introduced by Breiman (1996) as a means to reduce the variance and enhance the predictive stability of machine learning algorithms. It is particularly effective when applied to unstable learners, such

Public

as decision trees, whose predictions can vary significantly with small changes in the training data (Dietterich, 2000).

The core principle of bagging is the generation of multiple versions of a training dataset through bootstrap sampling, which involves drawing random samples with replacement. Each subset is used to train an independent base learner, and the individual predictions are then aggregated, typically by majority voting (for classification) or averaging (for regression) (Breiman, 1996).

Bagging is inherently parallelizable, which makes it computationally scalable. It has become a fundamental component of more advanced ensemble methods, including Random Forests, and is widely used in financial risk modeling due to its ability to improve predictive performance and reduce model overfitting (Polikar, 2006).

### Application in Financial Risk Analysis

In the domain of bank failure prediction and financial risk modeling, bagging has been applied to stabilize predictions and improve classification accuracy (Alfaro et al., 2008). Financial datasets are often characterized by high dimensionality, non-linearity, and the presence of outliers, making them challenging for traditional classifiers. Bagging addresses these challenges by:

- Reducing overfitting in models trained on noisy financial data.
- Enhancing robustness against outliers.
- Increasing stability across different training subsets.

Common applications in finance include:

- Credit risk evaluation (Lessmann et al., 2015).
- Bankruptcy prediction (Kumar & Ravi, 2007).
- Fraud detection (Phua et al., 2010).

### Strengths in Financial Applications

#### Variance Reduction and Overfitting Control

Bagging significantly reduces the variance of base learners by averaging out their predictions, which is especially important in financial applications where data volatility and market fluctuations introduce unpredictability (Bühlmann & Yu, 2002).

#### Resilience to Noisy Data

The randomization inherent in bagging makes it less sensitive to noisy and outlier-influenced financial data. This resilience is crucial in the prediction of bank failure, where historical data may include anomalies due to economic crises or regulatory changes (Alfaro et al., 2008).

#### Model Stability

Bagging produces stable classifiers by aggregating multiple independent models, which enhances confidence in model predictions. In financial contexts, where regulatory decisions

Public

may depend on the outputs of risk models, this stability is highly desirable (Kotsiantis, Zaharakis, & Pintelas, 2007).

High Dimensionality Handling

Bagging can effectively manage high-dimensional data without extensive feature selection or dimensionality reduction, provided the base learners are appropriate (e.g., decision trees). This makes it useful in analyzing financial datasets that include a wide array of CAMEL ratios, macroeconomic indicators, and market factors.

Limitations and Considerations

Despite its strengths, bagging has some limitations:

1. Lack of Interpretability: Bagging models, especially when using complex base learners, can be challenging to interpret, which poses difficulties in financial settings where model transparency is often required (Martens & Baesens, 2010).
2. Computational Cost: While bagging is parallelizable, generating and aggregating predictions from multiple models can be computationally intensive, especially for large datasets and complex base learners (Bühlmann & Yu, 2002).
3. Not Effective for Low-Variance Models: Bagging provides little to no improvement when applied to low-variance, high-bias learners, such as linear models (Dietterich, 2000).

Enhancements and Extensions

- Random Forests (Breiman, 2001): Extend bagging by introducing random feature selection, further reducing correlation between base trees.
- Boosting: Focuses on reducing bias rather than variance and can be used in tandem with bagging for improved performance (Freund & Schapire, 1996).

Conclusion

Bagging has established itself as a powerful ensemble method for financial risk prediction and bank failure modeling, thanks to its capacity to reduce variance, improve stability, and enhance accuracy. Its application is particularly beneficial in handling the volatility and complexity inherent in financial datasets. While it lacks interpretability, its predictive strength makes it an invaluable tool for risk management, early warning systems, and financial decision support.

Multilayer Perceptron (MLP)

Introduction and Theoretical Foundation

The Multilayer Perceptron (MLP) is a class of feedforward artificial neural networks (ANNs) that form the foundation of deep learning architectures. MLPs are composed of an input layer, one or more hidden layers, and an output layer, with each layer consisting of nodes called neurons that are interconnected by weighted connections (Rumelhart, Hinton, & Williams, 1986).MLPs emerged as a solution to the limitations of early perceptron models, particularly their inability to solve non-linearly separable problems (Minsky & Papert, 1969). By introducing non-linear activation functions and multiple hidden layers, MLPs are capable of

Public

learning complex, non-linear relationships, which makes them highly effective for tasks such as financial risk modeling and bank failure prediction (Tam & Kiang, 1992).

### Application in Financial Risk Analysis

MLPs have become a critical tool in financial risk analysis, particularly in predicting bank failures. Their ability to model non-linear dependencies between variables makes them suitable for capturing the complex interplay among financial indicators.

### Use Cases

- Bank failure prediction: Integrating balance sheet data, CAMEL ratios, and macroeconomic indicators to forecast bank distress (Zhang, Wang, & Ji, 1999).
- Credit scoring: Evaluating borrower risk based on transaction histories and demographic variables (West, 2000).
- Fraud detection: Identifying suspicious financial transactions and anomalies (Bhattacharyya et al., 2011).

In the prediction of bank failure, MLPs can be trained on datasets containing both failed and healthy banks, learning to distinguish subtle patterns in capital adequacy, liquidity, asset quality, and earnings performance.

### Strengths in Financial Applications

#### Capturing Non-Linear Relationships

MLPs excel at learning complex, non-linear decision boundaries (Hornik, Stinchcombe, & White, 1989). This capability is critical in financial data, where relationships between financial ratios and bank stability are often non-linear and interdependent.

#### Feature Representation and Learning

Unlike traditional models, MLPs can automatically learn representations from raw input data, reducing the need for manual feature engineering (Goodfellow, Bengio, & Courville, 2016).

#### Handling Noisy and Incomplete Data

MLPs are robust to noisy datasets and can handle missing data by learning underlying patterns and relationships (Bishop, 1995). This is particularly advantageous in financial datasets, which frequently contain irregularities due to reporting delays, regulatory differences, or missing entries.

#### Adaptability and Scalability

MLPs can be scaled by adding more hidden layers and neurons, allowing them to adapt to different complexities in data. This scalability makes them suitable for a variety of financial prediction tasks, from small datasets of regional banks to large multinational banking institutions (Zhang et al., 1999).

### Limitations and Considerations

Despite their strengths, MLPs present several challenges in financial applications:

1. Lack of Interpretability: MLPs are often considered black-box models, providing limited insight into why a prediction was made. This lack of transparency complicates their use in regulatory environments that require explainable AI (XAI) (Lipton, 2018).
2. Overfitting: MLPs are prone to overfitting, especially in small sample sizes, a common issue in bank failure datasets where failed banks are underrepresented. Techniques such as dropout (Srivastava et al., 2014) and L2 regularization are typically employed to mitigate this risk.
3. Computational Complexity: Training large MLPs can be computationally intensive, requiring significant hardware resources and optimization techniques (Goodfellow et al., 2016).
4. Hyperparameter Sensitivity: MLPs require careful tuning of hyperparameters (e.g., number of layers, learning rate, activation functions), often necessitating extensive experimentation (Bengio, 2012).

### Addressing Limitations

- Explainability techniques such as LIME (Ribeiro, Singh, & Guestrin, 2016) and SHAP (Lundberg & Lee, 2017) can help improve interpretability.
- Ensemble methods (e.g., stacking MLPs with interpretable models) can balance predictive power with interpretability.

### Conclusion

The Multilayer Perceptron remains a powerful and flexible tool in financial risk prediction, particularly in applications where complex, non-linear patterns exist among financial indicators. While its black-box nature poses challenges in terms of interpretability, its predictive accuracy, scalability, and resilience make it an essential component of modern risk management frameworks and early warning systems in the banking sector.

## Random Subspace Method (RSM)

### Introduction and Theoretical Foundation

The Random Subspace Method (RSM), also known as attribute bagging or feature bagging, is an ensemble learning technique first introduced by Ho (1998). The method was designed to enhance the accuracy and diversity of classification models by training multiple classifiers on different random subsets of features, rather than different subsets of instances as in traditional bagging (Breiman, 1996).

RSM addresses the curse of dimensionality (Bellman, 1957), a common challenge in machine learning when dealing with high-dimensional datasets. This is particularly relevant in financial risk modeling, where datasets often contain numerous correlated financial indicators, economic factors, and market-based variables. By training each classifier on a random subset of attributes, RSM reduces the correlation between classifiers, increases diversity, and mitigates the risk of overfitting (Dietterich, 2000).

RSM has been successfully combined with various base classifiers, such as decision trees, support vector machines, and k-nearest neighbors, and has demonstrated its effectiveness in numerous domains, including credit scoring (Lessmann et al., 2015) and bank failure prediction (Kuncheva & Rodríguez, 2007).

### Application in Financial Risk Analysis

The Random Subspace Method has significant relevance in financial risk prediction and bank failure modeling, where datasets are typically characterized by:

- High-dimensional feature spaces, often exceeding the number of available instances.
- Correlated financial ratios, particularly in CAMEL-based evaluations (capital adequacy, asset quality, management quality, earnings, and liquidity).
- Noisy and redundant variables, which can degrade the performance of individual classifiers.

By focusing each base classifier on different subsets of attributes, RSM helps uncover complementary patterns and mitigates overfitting, ultimately improving generalization performance.

### Strengths in Financial Applications

#### Mitigation of the Curse of Dimensionality

RSM effectively addresses the curse of dimensionality by reducing the feature space each classifier needs to handle. In financial risk modeling, where high-dimensional data is common, this allows classifiers to operate more efficiently and robustly (Ho, 1998).

#### Classifier Diversity and Improved Generalization

By ensuring that each classifier learns from a different subspace, RSM increases diversity among base classifiers. Diverse classifiers are less likely to make the same errors on unseen data, improving the generalization of the ensemble (Kuncheva & Rodríguez, 2007).

#### Robustness to Noise and Redundant Features

In financial datasets, redundant and irrelevant features are prevalent. RSM reduces the influence of such features by limiting the feature space for each classifier, thereby improving the signal-to-noise ratio and reducing overfitting (Guyon & Elisseeff, 2003).

#### Compatibility with Various Base Learners

RSM can be combined with a wide range of base classifiers, including decision trees, SVMs, neural networks, and k-NN algorithms. This flexibility allows practitioners to tailor ensembles to specific financial prediction tasks (Skurichina & Duin, 2002).

### Limitations and Considerations

Despite its advantages, the Random Subspace Method also has certain limitations in financial modeling applications:

1. Loss of Information: By randomly excluding features from individual classifiers, RSM may lose critical information, particularly when important features are excluded from multiple subspaces (Skurichina & Duin, 2002).
2. Choice of Subspace Dimensionality (k): The performance of RSM is sensitive to the choice of kkk, the number of features included in each subspace. If kkk is too small, classifiers may lack sufficient information; if too large, diversity and variance reduction benefits may be diminished (Kuncheva & Rodríguez, 2007).
3. Interpretability: The black-box nature of ensemble classifiers built on multiple subspaces can limit model interpretability, a concern for financial regulators and stakeholders requiring transparent models (Martens & Baesens, 2010).

Strategies to Address Limitations

- Feature Importance Analysis: Prioritizing feature selection through domain knowledge or feature importance rankings can mitigate the risk of excluding critical attributes (Guyon & Elisseeff, 2003).
- Hybrid Models: Combining RSM with feature selection algorithms or boosting techniques may enhance predictive power while maintaining manageable complexity (Zhou, 2012).

Conclusion

The Random Subspace Method offers a powerful approach to high-dimensional, noisy, and complex financial datasets, such as those used in bank failure prediction. By promoting classifier diversity and reducing variance, RSM improves generalization and model robustness. However, its effectiveness depends on careful tuning and consideration of interpretability, particularly in regulated financial environments.

Future research can explore the integration of explainable AI (XAI) techniques with RSM ensembles, as well as the use of adaptive subspace selection methods to enhance performance in dynamic financial contexts.

**Random Forest**

Introduction and Theoretical Foundation

The Random Forest (RF) algorithm, introduced by Breiman (2001), is one of the most widely used ensemble learning techniques in modern machine learning. Random Forest is a meta-learning method that combines the predictions of multiple decision trees to improve classification and regression tasks. As an ensemble of decision trees, Random Forest leverages both bagging (bootstrap aggregating) and random feature selection, enhancing predictive accuracy and reducing overfitting (Breiman, 2001).

In financial risk analysis and bank failure prediction, Random Forest has become a preferred method due to its ability to handle high-dimensional datasets, manage complex and nonlinear relationships, and offer robustness against noisy data (Liaw & Wiener, 2002). Its capacity for feature selection and handling of imbalanced datasets makes it particularly valuable in tasks such as credit risk modeling, bank distress prediction, and fraud detection.

Public

### Application in Financial Risk Analysis

Random Forest has been extensively applied in financial risk modeling due to its ability to manage complex datasets with large feature sets, missing values, and imbalanced classes (Díaz-Uriarte & de Andrés, 2006).

### Key Applications

1. Bank Failure Prediction: RF models use financial ratios (e.g., CAMEL indicators) and macroeconomic data to forecast bank distress (Petropoulos et al., 2020).
2. Credit Risk Scoring: RF is applied in determining the likelihood of loan default based on borrower characteristics and historical data (Lessmann et al., 2015).
3. Fraud Detection: RF detects anomalous patterns in transactional datasets, identifying potential fraud events (Phua et al., 2010).
4. Loan Default Analysis: RF ranks the most important variables predicting borrower delinquency (Baesens et al., 2003).

### Strengths in Financial Applications

#### Robustness to Overfitting

By aggregating the predictions of multiple decorrelated decision trees, RF reduces the risk of overfitting, making it suitable for high-variance financial datasets (Breiman, 2001).

#### High Accuracy and Generalization

RF often delivers high accuracy due to its ensemble nature and can generalize well to out-of-sample data (Hastie, Tibshirani, & Friedman, 2009). In bank failure prediction, where the class imbalance is a concern, RF demonstrates strong generalization (Petropoulos et al., 2020).

#### Embedded Feature Selection

RF provides variable importance measures (Mean Decrease in Gini and Mean Decrease in Accuracy), enabling analysts to identify the most critical features influencing predictions (Díaz-Uriarte & de Andrés, 2006). This is useful for understanding which CAMEL ratios or economic indicators are most predictive of bank distress.

#### Resilience to Noisy and Incomplete Data

RF is robust to missing values and outliers because it uses decision trees that inherently manage missing data through surrogate splits or randomization (Liaw & Wiener, 2002).

#### Scalability

RF is highly scalable and can handle large datasets with thousands of variables, a common requirement in financial risk modeling (Pal, 2005).

### Limitations and Considerations

Despite its advantages, Random Forest presents several limitations:

Public

1. Lack of Interpretability: RF models are considered black-box models, making it difficult to interpret individual predictions. In regulatory contexts, lack of transparency may hinder adoption (Martens & Baesens, 2010).
2. Computational Complexity: Building large numbers of deep trees can be computationally expensive, particularly for very large datasets (Biau & Scornet, 2016).
3. Bias Toward Majority Class: In imbalanced datasets, RF may favor the majority class, resulting in higher Type II errors (false negatives). Addressing this requires techniques such as cost-sensitive learning, SMOTE, or balanced random forests (Chen & Breiman, 2004).

Techniques to Address Limitations

- Model Explanation Tools: SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) provide explanation techniques for RF predictions.
- Class Imbalance Solutions: Cost-sensitive RF, Balanced RF, and oversampling techniques improve classification performance on minority classes (Chen & Breiman, 2004).
- Parallel Processing: Leveraging parallel computation (e.g., Hadoop, Spark) to speed up RF training on large financial datasets (Zhou et al., 2014).

Conclusion

Random Forest stands as a cornerstone machine learning method in financial risk modeling, offering robust predictive power, feature selection capabilities, and resilience to noise and overfitting. Its application in bank failure prediction has provided regulators and financial institutions with a powerful tool for early warning systems, enabling proactive intervention and risk management.

Regularized Random Forest (RRF)

Introduction and Theoretical Foundation

The Regularized Random Forest (RRF) is an extension of the Random Forest (RF) algorithm that incorporates regularization techniques to improve model performance, particularly in high-dimensional feature spaces (Deng & Runger, 2012). While Random Forests inherently perform some degree of feature selection through the random subspace method, they often include redundant or less informative features. RRF introduces a feature selection regularization framework, which penalizes the inclusion of redundant or weakly relevant features during the tree-building process, resulting in simpler and more interpretable models (Deng & Runger, 2012).

In financial risk modeling and bank failure prediction, datasets often exhibit high dimensionality, multicollinearity, and complex feature interactions. RRF effectively addresses these challenges by enhancing feature selection and reducing model complexity, thereby improving both predictive performance and interpretability (Biau & Scornet, 2016).

Public

### Regularization Impact

By controlling the regularization parameter, RRF can balance model complexity and predictive accuracy. Larger lambda values encourage sparser trees, which can be advantageous in high-dimensional datasets.

### Application in Financial Risk Analysis

The Regularized Random Forest has gained traction in financial applications, particularly where:

- High-dimensional datasets are common (e.g., financial ratios, macroeconomic indicators, transactional data).
- Feature redundancy and collinearity pose challenges to model interpretability and stability.
- Regulatory compliance requires transparent and parsimonious models.

### Strengths in Financial Applications

### Enhanced Feature Selection

RRF explicitly controls feature inclusion, leading to models that are both accurate and sparser (Deng & Runger, 2012). In financial risk modeling, this capability enhances interpretability and helps identify key risk factors influencing predictions.

### Reduced Overfitting

By discouraging the selection of redundant features, RRF mitigates overfitting, which is especially critical in bank failure prediction, where minority class (failure) events are rare and overfitting is a significant risk (Biau & Scornet, 2016).

### Improved Interpretability

RRF builds simpler models compared to standard RF, facilitating easier interpretation of which variables contribute to risk assessment and classification decisions (Meinshausen, 2007). This interpretability is essential for regulatory reporting and compliance.

### Scalability and Efficiency

RRF retains the parallelizability and scalability of RF, making it suitable for large-scale financial datasets, where quick risk assessments are needed for hundreds of banks or millions of customers (Deng, 2013).

Public

## Limitations and Considerations

While RRF addresses many limitations of standard Random Forest, it introduces new considerations:

1. Hyperparameter Sensitivity: The regularization parameter lambda requires careful tuning. Poor selection can lead to underfitting (if lambda is too high) or insufficient regularization (if lambda is too low) (Deng & Runger, 2012).
2. Computational Overhead: Incorporating regularization increases the computational complexity of feature selection during tree construction, although the predictive phase remains efficient (Biau & Scornet, 2016).
3. Interpretability Trade-offs: While more interpretable than standard RF, RRF still lacks the transparency of simpler models like Logistic Regression or Decision Trees, potentially limiting its adoption in high-stakes regulatory settings (Martens & Baesens, 2010).

## Conclusion

The Regularized Random Forest represents a significant advancement in ensemble learning, particularly suited to the high-dimensional, complex datasets characteristic of financial risk modeling. By promoting parsimonious models that maintain high predictive accuracy, RRF addresses critical challenges in bank failure prediction, credit scoring, and fraud detection.

## Cost-Sensitive Forest (CS-Forest)

## Introduction and Theoretical Foundation

The Cost-Sensitive Forest (CS-Forest) is an advanced adaptation of traditional Random Forest and ensemble learning methods, specifically designed to address the challenges posed by asymmetric misclassification costs in classification problems. In financial risk modeling, particularly in bank failure prediction, the costs associated with different types of misclassification errors (i.e., false positives and false negatives) are typically unequal. CS-Forest directly integrates these asymmetric costs into its learning algorithm, providing a cost-sensitive framework that is better aligned with real-world decision-making (Elkan, 2001).

In bank failure prediction, a false negative (i.e., failing to identify a distressed bank) can have catastrophic economic consequences, including financial instability, depositor losses, and systemic risk propagation. Conversely, a false positive (i.e., incorrectly predicting a healthy bank as failing) can result in unnecessary regulatory intervention, reputation damage, and capital allocation inefficiencies. CS-Forest seeks to minimize the total expected cost of classification errors, rather than optimizing for overall accuracy alone (Ling & Sheng, 2008).

The Cost-Sensitive Forest modifies the training process of decision trees within a Random Forest to incorporate cost matrices that explicitly represent the costs of misclassification.

## Weighting in Sampling

CS-Forest often employs weighted sampling to ensure that minority class instances (e.g., failed banks) are oversampled based on their misclassification costs, effectively creating cost-proportionate data distributions (Zadrozny et al., 2003).

### Strengths in Financial Applications

### Alignment with Economic Objectives

Unlike traditional classifiers that optimize for accuracy, CS-Forest aligns its objective function with real-world economic costs, ensuring that the most costly errors (typically false negatives) are minimized (Elkan, 2001).

### Enhanced Minority Class Detection

CS-Forest improves recall for the minority class, which, in bank failure prediction, corresponds to failing banks. This focus on reducing Type II errors enhances the early warning capabilities of the model (Ling & Sheng, 2008).

### Flexibility and Adaptability

The cost matrix can be easily adjusted to reflect changing economic conditions or regulatory priorities, making CS-Forest adaptable to dynamic financial environments (Zadrozny et al., 2003).

### Robustness to Imbalanced Datasets

CS-Forest effectively handles imbalanced datasets, a common feature in the prediction of bank failure, where failed banks represent a small fraction of the population (Khalilia, Chakraborty & Popescu, 2011).

### Limitations and Considerations

Despite its advantages, CS-Forest introduces certain complexities:

1. Cost Specification Challenges: Defining an appropriate cost matrix requires domain expertise and may involve subjective judgment. Inappropriate cost settings can skew model behavior (Elkan, 2001).
2. Interpretability Issues: As with standard RF, CS-Forest models can be opaque, limiting their interpretability in regulatory settings (Martens & Baesens, 2010).
3. Computational Complexity: Incorporating cost sensitivity into split selection and sampling increases computational demands, particularly for large-scale financial datasets (Dal Pozzolo et al., 2015).

### Strategies to Address Limitations

- Sensitivity Analysis: Testing different cost scenarios to understand their impact on classification outcomes and model robustness.
- Explainability Techniques: Employing tools like SHAP (Lundberg & Lee, 2017) or LIME (Ribeiro et al., 2016) to enhance model interpretability.
- Balanced Cost Settings: Collaborating with domain experts (regulators, economists) to establish economically justified cost structures (Burez & Van den Poel, 2009).

Public

## Conclusion

The Cost-Sensitive Forest is a powerful enhancement of ensemble learning methods, specifically tailored for risk-sensitive applications in the financial sector. By integrating asymmetric misclassification costs, CS-Forest delivers economically optimal predictions in bank failure prediction, credit risk modeling, and fraud detection.

Its capacity to prioritize the most consequential errors, particularly false negatives, makes it an essential tool for financial regulators, risk managers, and policymakers. Future advancements may include dynamic cost-sensitive frameworks and hybrid models that balance predictive performance, cost alignment, and explainability.

## Generalized Linear Model Boost (GLMBoost)

### Introduction and Theoretical Foundation

Generalized Linear Model Boosting (GLMBoost) is a powerful hybrid technique that combines the interpretable structure of Generalized Linear Models (GLMs) with the predictive accuracy and regularization capabilities of boosting algorithms (Bühlmann & Hothorn, 2007). GLMBoost builds upon the foundation of GLMs, which are an extension of classical linear regression models capable of handling non-normal error distributions and non-linear link functions (McCullagh & Nelder, 1989). The boosting component iteratively fits GLMs to the residuals of the preceding models, improving predictive performance while preventing overfitting.

GLMBoost is particularly useful in financial risk modeling applications, including bank failure prediction, credit scoring, and loan default analysis, where interpretability, robustness, and variable selection are essential. The ability of GLMBoost to simultaneously enhance predictive power and maintain model transparency makes it an attractive option for regulators and financial analysts, who often require explanations of model behavior (Hastie, Tibshirani, & Friedman, 2009).

### Regularization via Early Stopping

GLMBoost controls model complexity through early stopping, terminating the boosting process after a pre-defined number of iterations or when model performance on a validation set deteriorates. This process functions as an implicit regularization technique, similar to LASSO or ridge regression (Hastie et al., 2009).

### Application in Financial Risk Analysis

GLMBoost addresses several critical needs in financial risk modeling:

- It provides interpretable models suitable for regulatory compliance.
- It handles high-dimensional datasets, common in financial applications where CAMEL ratios, macroeconomic indicators, and market data are integrated.
- It offers automatic variable selection and regularization, mitigating the effects of multicollinearity.

### Strengths in Financial Applications

### Interpretability and Regulatory Compliance

GLMBoost retains the interpretability of traditional GLMs while improving predictive accuracy. The coefficients remain interpretable as odds ratios or risk multipliers, providing actionable insights for regulators and financial institutions (McCullagh & Nelder, 1989).

### Robustness and Regularization

The boosting framework inherently regularizes the model, preventing overfitting, especially in high-dimensional or imbalanced datasets. Early stopping serves as a bias-variance trade-off control, ensuring generalizability (Bühlmann & Hothorn, 2007).

### Automatic Variable Selection

GLMBoost performs embedded feature selection, automatically identifying and excluding irrelevant or redundant predictors, which is crucial when modeling complex financial risks (Hastie et al., 2009).

### Flexibility in Model Specification

GLMBoost accommodates different link functions and error distributions, making it versatile for various financial risk types, including binary classification (bank failure prediction), count data (operational losses), and continuous outcomes (loan loss severity) (Bühlmann & Hothorn, 2007).

### Limitations and Considerations

Despite its advantages, GLMBoost presents some challenges in practical applications:

1. Computational Complexity: The iterative nature of boosting can result in longer training times, especially for large financial datasets (Hothorn et al., 2010).
2. Hyperparameter Tuning: Determining optimal values for learning rates and stopping criteria requires careful cross-validation, increasing model development time (Friedman et al., 2000).
3. Sensitivity to Data Quality: Like GLMs, GLMBoost is sensitive to outliers and missing data, necessitating robust data preprocessing (McCullagh & Nelder, 1989).

### Strategies to Address Limitations

- Parallel Processing: Utilizing parallelization techniques can speed up model training for large-scale datasets (Hothorn et al., 2010).
- Automated Hyperparameter Optimization: Tools such as grid search and Bayesian optimization streamline the selection of learning rates and number of iterations.
- Data Preprocessing Pipelines: Imputation of missing values and outlier detection methods improve model stability (Chernobai et al., 2007).

Public

## Conclusion

GLMBoost bridges the gap between interpretable statistical models and high-performance machine learning algorithms, making it an ideal tool for financial risk analysis. Its automatic feature selection, regularization, and flexibility enable accurate and transparent prediction of bank failures, credit defaults, and operational losses.

By balancing accuracy and interpretability, GLMBoost aligns with the requirements of financial regulators, risk managers, and practitioners, contributing to early warning systems and risk-informed decision-making.

## AdaBoost (Adaptive Boosting)

### Introduction and Theoretical Foundation

Adaptive Boosting (AdaBoost), introduced by Freund and Schapire (1996), is one of the earliest and most influential ensemble learning algorithms. AdaBoost was developed to improve the performance of weak learners, typically classifiers that are only slightly better than random guessing, by combining them into a strong composite model. The algorithm does this by sequentially training models on weighted versions of the data, focusing more on instances that were misclassified in previous iterations.

In financial risk analysis, including bank failure prediction, credit scoring, and fraud detection, AdaBoost has proven effective because of its ability to improve classification performance in datasets that are often imbalanced and complex. Its adaptability in reweighting misclassified instances makes it highly suited for domains where certain misclassification errors (such as failing to predict a bank’s insolvency) can be extremely costly (Chen & Breiman, 2004).

### Interpretation

Misclassified instances receive higher weights, forcing subsequent classifiers to focus on the hard-to-classify cases. This approach is highly beneficial for imbalanced datasets and complex classification tasks, where important minority cases (e.g., bank failures) might otherwise be ignored.

### Application in Financial Risk Analysis

AdaBoost is extensively applied in financial domains where classification accuracy and early detection are vital.

### Strengths in Financial Applications

#### Improved Predictive Accuracy

AdaBoost often outperforms traditional classifiers, including logistic regression and decision trees, especially in imbalanced datasets, making it well-suited for bank failure prediction, where failed banks are relatively rare (Freund & Schapire, 1996).

Public

### Robustness to Overfitting

Although boosting algorithms are prone to overfitting, AdaBoost demonstrates surprising resistance to overfitting, particularly when weak classifiers are pruned (Schapire, 2003). This makes it suitable for noisy financial datasets.

### Adaptive Learning Focus

AdaBoost adaptively focuses on problematic instances, enhancing the detection of critical events such as bank defaults or credit fraud (Schapire & Singer, 1999).

### Flexibility

AdaBoost can be combined with various types of base learners (e.g., decision trees, SVMs), making it highly flexible for different financial prediction tasks (Hastie, Tibshirani & Friedman, 2009).

### Limitations and Considerations

Despite its strengths, AdaBoost presents challenges in financial risk modeling:

1. Sensitivity to Noisy Data and Outliers: AdaBoost can give too much emphasis to outliers or noisy instances, especially when they are misclassified repeatedly (Rätsch, Onoda & Müller, 2001).
2. Class Imbalance: Although AdaBoost improves on standard classifiers, it may still suffer from bias towards the majority class. Variants like Cost-Sensitive AdaBoost or SMOTEBoost address this issue (Seiffert et al., 2010).
3. Interpretability: AdaBoost creates complex ensembles of classifiers, making the final model less interpretable than traditional statistical models. This can be a problem for regulatory compliance and model validation (Martens & Baesens, 2010).

### Strategies to Address Limitations

- Cost-Sensitive Variants: Applying Cost-Sensitive AdaBoost or AdaC2 improves minority class recall (Sun et al., 2007).
- Data Preprocessing: Employing outlier detection and data cleaning before training can mitigate the negative impact of noisy data.
- Model Interpretation Tools: Using SHAP or LIME provides explanation mechanisms for AdaBoost’s predictions (Lundberg & Lee, 2017).

### Conclusion

AdaBoost remains a powerful ensemble method in financial risk modeling, particularly in high-stakes domains like bank failure prediction, credit scoring, and fraud detection. Its ability to adaptively focus on misclassified instances enhances predictive accuracy in imbalanced and noisy datasets.

Despite concerns around interpretability and noise sensitivity, AdaBoost’s flexibility, robustness, and effectiveness make it a valuable tool for financial analysts, risk managers, and

Public

regulatory authorities who seek early warning systems and high-performance classification models.

### LogitBoost

#### Introduction and Theoretical Foundation

LogitBoost, introduced by Friedman, Hastie, and Tibshirani (2000), is an advanced boosting algorithm that specializes in solving classification problems by combining logistic regression principles with the boosting framework. It is particularly designed for binary classification tasks, making it suitable for credit risk analysis, customer segmentation, and bank failure prediction, where outcomes are often dichotomous (e.g., default/no default, failure/no failure).

LogitBoost builds additive logistic regression models by minimizing the binomial deviance, which makes its outputs more directly interpretable as probabilities. Unlike AdaBoost, which works by reweighting misclassified instances, LogitBoost fits additive logistic regression models by directly optimizing the log-likelihood using a stage-wise approach (Hastie et al., 2009).

Its application in financial risk prediction stems from its ability to handle imbalanced datasets, provide probability estimates, and balance predictive performance with interpretability, making it an essential tool in risk management and regulatory environments.

LogitBoost minimizes the binomial log-likelihood loss function by additive model fitting, using Newton steps for function estimation at each iteration.

#### Comparison with AdaBoost

Unlike AdaBoost, which focuses on classification accuracy, LogitBoost focuses on probability estimation and likelihood maximization, offering better-calibrated probabilities, which are crucial in financial risk modeling.

#### Application in Financial Risk Analysis

LogitBoost’s capacity to produce well-calibrated probability estimates makes it valuable in bank failure prediction, credit scoring, and fraud detection, where risk probability often directly informs decision-making.

#### Strengths in Financial Applications

##### Probability Calibration

LogitBoost produces probability estimates that are better calibrated than other boosting methods like AdaBoost. This is particularly beneficial in financial risk analysis, where thresholding decisions (e.g., to intervene, reject a loan, or flag fraud) are based on risk probabilities (Niculescu-Mizil & Caruana, 2005).

Handling of Imbalanced Data

LogitBoost demonstrates robustness in imbalanced datasets, which are common in bank failure prediction and fraud detection, ensuring better minority class recall without compromising overall accuracy (Friedman et al., 2000).

Robustness to Noisy Data

The Newton update mechanism allows LogitBoost to focus on statistically significant updates rather than overly emphasizing outliers or noisy instances, improving stability in noisy financial datasets (Hastie et al., 2009).

Interpretability

LogitBoost models can be interpreted as additive logistic regression models, providing odds ratios and coefficients that facilitate regulatory reporting and auditability (Martens & Baesens, 2010).

Limitations and Considerations

While LogitBoost has notable strengths, it presents certain challenges in financial applications:

1. Computational Intensity: LogitBoost is computationally more intensive than AdaBoost, particularly when fitting complex base learners over large datasets (Hothorn et al., 2010).
2. Hyperparameter Tuning: Careful selection of learning rates, number of iterations, and base learners is necessary to achieve optimal performance (Friedman et al., 2000).
3. Less Effective with Weak Base Learners: While designed for weak learners, if base learners are too simplistic, performance may suffer, necessitating more complex learners that increase computation (Bühlmann & Hothorn, 2007).

Strategies to Address Limitations

- Parallelization: Implementing parallel training can mitigate training time issues.
- Automated Hyperparameter Tuning: Employing grid search or Bayesian optimization to select optimal configurations.
- Use of Moderately Complex Base Learners: Balancing model complexity and training speed by selecting pruned decision trees rather than stumps.

Conclusion

LogitBoost represents a highly effective and interpretable boosting algorithm for financial risk analysis, particularly in bank failure prediction, credit risk modeling, and fraud detection. Its probability calibration, robustness to imbalanced data, and interpretability make it an appealing tool for financial regulators, risk managers, and financial institutions.

By combining the strengths of logistic regression and boosting, LogitBoost addresses many of the limitations found in traditional machine learning models, contributing to early warning systems and risk-informed decision-making in financial supervision.

Public

## Generalized Additive Model Boosting (GAMBoost)

### Introduction and Theoretical Foundation

Generalized Additive Model Boosting (GAMBoost) is a hybrid machine learning algorithm that combines the flexibility of Generalized Additive Models (GAMs) with the predictive power and regularization capabilities of boosting algorithms. GAMBoost was developed as a response to the need for interpretable, flexible, and highly predictive models, particularly in fields like financial risk analysis where non-linear relationships between predictors and outcomes are common and where model transparency is often required (Bühlmann & Hothorn, 2007).

The GAM component allows GAMBoost to model non-linear effects of predictor variables in a semi-parametric fashion, which enhances its ability to capture complex relationships. The boosting component, on the other hand, applies functional gradient descent techniques to improve predictive accuracy and incorporate automatic variable selection and regularization (Friedman, 2001).

In financial contexts, GAMBoost is particularly well-suited for bank failure prediction, credit scoring, and fraud detection because it can identify subtle, non-linear effects among financial ratios, macroeconomic indicators, and other relevant features while maintaining an interpretable structure.

GAMBoost applies gradient boosting principles to the GAM framework, where the response variable is modeled as an additive combination of smooth functions of predictors.

### Application in Financial Risk Analysis

GAMBoost is especially valuable in financial applications that require flexibility in modeling non-linear relationships and explainability of the results.

### Strengths in Financial Applications

#### Model Interpretability

Unlike many machine learning models, GAMBoost provides transparent and interpretable outputs. Each predictor's functional relationship with the target variable is explicitly modeled and can be visualized (Hastie & Tibshirani, 1986). This level of transparency is critical in regulatory environments where auditability is required.

#### Handling of Non-Linear Relationships

GAMBoost flexibly models non-linearities, addressing one of the primary limitations of linear models in financial datasets, where threshold effects, diminishing returns, and non-monotonic relationships are common (Friedman, 2001).

#### Automatic Variable Selection and Regularization

The boosting process inherently selects relevant variables while penalizing less informative features, reducing model complexity and addressing multicollinearity, which is a frequent problem in financial datasets (Bühlmann & Hothorn, 2007).

Public

### Resistance to Overfitting

By combining boosting with penalized spline smoothing, GAMBoost reduces the risk of overfitting, providing robust and generalizable models, even when dealing with noisy financial data (Eilers & Marx, 1996).

### Limitations and Considerations

Despite its strengths, GAMBoost also presents several challenges:

1. Computational Demands: The algorithm can be computationally intensive, especially for large-scale financial datasets with many predictors (Hothorn et al., 2010).
2. Complex Model Tuning: Selecting the appropriate smoothing parameters, learning rates, and stopping criteria requires careful cross-validation and can be time-consuming (Friedman, 2001).
3. Potential Interpretability Trade-Off: While GAMBoost is more interpretable than many black-box models, the inclusion of many smooth functions and complex interactions can complicate interpretation for non-technical stakeholders (Hastie et al., 2009).

### Strategies to Address Limitations

- Parallel Computing: Leveraging parallel processing can reduce computation time in large datasets.
- Automated Tuning Methods: Employing Bayesian optimization or automated grid search for hyperparameter tuning improves model efficiency.
- Visualization Tools: Using partial dependence plots and effect plots to simplify communication of complex relationships to stakeholders.

### Conclusion

GAMBoost offers a compelling blend of interpretability, predictive accuracy, and flexibility, making it highly suitable for financial risk modeling applications. Its ability to capture complex, non-linear relationships in a transparent manner equips financial analysts and regulators with a powerful tool for early detection of bank failures, credit defaults, and fraudulent activities.

The algorithm’s combination of boosting with additive modeling enables robust, accurate, and interpretable solutions to high-stakes financial decision-making problems.

## KStar Algorithm

### Introduction and Theoretical Foundation

KStar, or K*, is an instance-based learning algorithm that was introduced by Cleary and Trigg (1995) as an enhancement of traditional lazy learning classifiers like k-Nearest Neighbors (k-NN). Unlike conventional models that generalize from data during the training phase, KStar defers generalization until a query instance is presented. This approach makes it highly adaptable, as it relies on the similarity between instances rather than on a precomputed global model.

Public

KStar distinguishes itself by employing an entropy-based distance function rather than traditional distance metrics (like Euclidean distance in k-NN). This probabilistic distance function evaluates the complexity of transforming one instance into another, making it more robust to noisy, heterogeneous, and high-dimensional data—conditions frequently encountered in financial risk analysis and bank failure prediction.

In financial contexts, where subtle and non-linear interactions between variables exist, KStar offers a non-parametric and flexible approach to modeling complex relationships. It is particularly suited for bank failure prediction, where patterns of distress are not always linearly separable or easily captured by parametric models.

KStar uses an instance-based learning paradigm, relying on a case-based reasoning strategy. Its probabilistic similarity function measures the complexity of transforming one instance xxx into another instance yyy, based on probability distributions.

The core of KStar is a measure called the transformation probability, which represents the probability that instance xxx can be transformed into instance yyy by a sequence of transformations. The transformations are stochastic processes defined by probabilistic functions, and the algorithm computes the cumulative probability over all possible transformations.

Application in Financial Risk Analysis

KStar's flexibility makes it highly applicable in financial risk prediction, particularly in scenarios where structured patterns are difficult to predefine.

Strengths in Financial Applications

Non-Parametric and Flexible

KStar is non-parametric, meaning it does not assume a specific functional form for the data distribution. This flexibility allows it to model complex, non-linear relationships in financial data, where risk factors often interact in unpredictable ways (Witten & Frank, 2005).

Robustness to Noise and Missing Data

KStar’s probabilistic approach makes it more robust to noise than conventional instance-based methods. It can gracefully handle missing values and incomplete data, a common issue in financial reporting, by transforming instances based on known attributes (Cleary & Trigg, 1995).

Effective in High-Dimensional Spaces

Financial datasets often include dozens or hundreds of variables (e.g., CAMEL ratios, market indicators). KStar’s entropy-based distance function allows it to handle high-dimensional data effectively without suffering from the curse of dimensionality to the same extent as k-NN (Aha et al., 1991).

Public

### Adaptability

KStar adapts its model to the data at query time, allowing it to respond to dynamic changes in financial conditions. Unlike eager learners (e.g., decision trees), it does not require retraining when new data is introduced, which is advantageous in rapidly evolving financial environments (Witten & Frank, 2005).

### Limitations and Considerations

Despite its strengths, KStar has several limitations that need consideration in financial applications:

1. Computational Inefficiency at Scale: Since KStar delays computation until query time, it can be computationally intensive, especially for large datasets (Aha et al., 1991). Every prediction requires a comparison to all stored instances, which can result in slow query responses.
2. Storage Requirements: KStar needs to store the entire training dataset, leading to high memory usage, particularly problematic in big data financial contexts (Cleary & Trigg, 1995).
3. Lack of Model Interpretability: While KStar excels at flexibility, it lacks the clear, interpretable model structure that decision trees, logistic regression, or GAMBoost offer. For regulatory compliance and decision justification, this can be a significant drawback (Martens & Baesens, 2010).

### Strategies to Address Limitations

- Data Reduction Techniques: Applying clustering or prototype selection can reduce the number of instances stored and speed up computation.
- Parallel Computing and Indexing: Utilizing parallel processing or efficient indexing structures can improve scalability.
- Hybrid Models: Combining KStar with explainable models (e.g., decision trees) can balance prediction accuracy and interpretability.

### Conclusion

KStar offers a robust, non-parametric solution for financial risk classification tasks, excelling in contexts where non-linear patterns, noise, and incomplete data challenge traditional algorithms. Its lazy learning approach makes it adaptive to new data without retraining, which is crucial in the volatile landscape of financial markets.

Public

# Chapter 5: Model Experiments

## 5.1 Model Performance Metrics -Naïve Bayes

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4478 (94.0559%) |
| Incorrectly Classified Instances | 283 (5.9441%) |
| Kappa Statistic | 0.6884 |
| Mean Absolute Error | 0.0594 |
| Root Mean Squared Error | 0.2367 |
| Relative Absolute Error | 34.8945% |
| Root Relative Squared Error | 81.1561% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.953 | 0.181 | 0.981 | 0.953 | 0.967 | 0.695 | 0.953 | 0.992 | Non-Fail |
| Failed | 0.819 | 0.047 | 0.644 | 0.819 | 0.721 | 0.695 | 0.954 | 0.684 | Failed |
| Weighted Avg. | 0.941 | 0.169 | 0.949 | 0.941 | 0.944 | 0.695 | 0.953 | 0.963 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4112 | 202 |
| b = Failed | 81 | 366 |

The Naïve Bayes classifier demonstrated an overall classification accuracy of 94.06%, correctly classifying 4478 out of 4761 instances. The Kappa Statistic of 0.6884 suggests moderate agreement between predicted and actual classifications beyond chance, which indicates a reasonable but not exceptional level of predictive consistency.

In terms of class-specific performance, Naïve Bayes achieved a high Recall of 95.3% for non-failed banks, suggesting that the model is highly effective in identifying institutions that are not at risk. However, its performance on the minority class—failed banks—was weaker, with a Recall of 81.9% and a Precision of 64.4%. This imbalance led to 81 false negatives, where failed banks were incorrectly classified as non-failed, and 202 false positives, where healthy banks were misclassified as failed. The relatively low precision for the failed class highlights the model's susceptibility to misclassification in highly imbalanced datasets, a known limitation of Naïve Bayes, particularly when the assumption of feature independence is violated.

The ROC-AUC score of 95.4% reflects good overall discriminatory power, but the PRC Area of 68.4% for failed banks suggests that the model struggles to maintain precision when recall increases—indicating challenges in accurately flagging true failures without generating a high number of false alarms.

### 5.2 Model Performance Metrics -Logistic Regression

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4623 (97.1014%) |
| Incorrectly Classified Instances | 138 (2.8986%) |
| Kappa Statistic | 0.8254 |
| Mean Absolute Error | 0.0408 |
| Root Mean Squared Error | 0.1461 |
| Relative Absolute Error | 23.9417% |
| Root Relative Squared Error | 50.0814% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.987 | 0.181 | 0.981 | 0.987 | 0.984 | 0.826 | 0.982 | 0.997 | Non-Fail |
| Failed | 0.819 | 0.013 | 0.865 | 0.819 | 0.841 | 0.826 | 0.982 | 0.897 | Failed |
| Weighted Avg. | 0.971 | 0.165 | 0.970 | 0.971 | 0.971 | 0.826 | 0.982 | 0.987 | - |

Confusion Matrix

|  | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4257 | 57 |
| b = Failed | 81 | 366 |

The Logistic Regression model demonstrated an overall classification accuracy of 97.10%, correctly classifying 4623 out of 4761 instances. The Kappa Statistic of 0.8254 suggests a substantial level of agreement between predicted and actual classifications beyond chance, reflecting a reliable predictive performance.

In terms of class-specific metrics, the model achieved a high Recall of 98.7% for non-failed banks, indicating strong sensitivity in correctly identifying institutions that are not at risk. For the failed bank class, the Recall was 81.9%, capturing a significant portion of the actual failures. Precision for the failed class was measured at 86.5%, suggesting the model maintains a good balance between correctly identifying failures and limiting false positives.

The confusion matrix shows 81 false negatives, where failed banks were misclassified as non-failed, and 57 false positives, where non-failed banks were misclassified as failed. This distribution highlights the model's capability to minimize false alarms while maintaining sensitivity to failures.The ROC-AUC score of 98.2% reflects excellent overall discriminatory power, and the PRC Area of 89.7% for the failed class indicates robust precision-recall trade-offs when identifying true failures.

Public

### 5.3 Model Performance Metrics -Multilayer Perceptron

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4620 (97.0384%) |
| Incorrectly Classified Instances | 141 (2.9616%) |
| Kappa Statistic | 0.8233 |
| Mean Absolute Error | 0.039 |
| Root Mean Squared Error | 0.1568 |
| Relative Absolute Error | 22.923% |
| Root Relative Squared Error | 53.7714% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.985 | 0.174 | 0.982 | 0.985 | 0.984 | 0.823 | 0.977 | 0.997 | Non-Fail |
| Failed | 0.826 | 0.015 | 0.854 | 0.826 | 0.840 | 0.823 | 0.977 | 0.876 | Failed |
| Weighted Avg. | 0.970 | 0.159 | 0.970 | 0.970 | 0.970 | 0.823 | 0.977 | 0.985 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4251 | 63 |
| b = Failed | 78 | 369 |

The Multilayer Perceptron (MLP) model achieved an overall classification accuracy of 97.04%, correctly classifying 4620 out of 4761 instances. The Kappa Statistic of 0.8233 indicates a substantial level of agreement between predicted and actual classifications beyond chance, demonstrating reliable predictive performance.

For class-specific outcomes, the MLP exhibited a Recall of 98.5% for non-failed banks, ensuring most healthy institutions were correctly identified. For the failed bank class, the Recall was 82.6%, reflecting the model's capacity to detect a significant proportion of actual bank failures. Precision for the failed class was 85.4%, suggesting the model effectively balances the identification of true positives while limiting false positives.

According to the confusion matrix, 78 failed banks were misclassified as non-failed (false negatives), while 63 non-failed banks were incorrectly classified as failed (false positives). This balance indicates that while the model maintains good sensitivity, there remains room for improvement in minimizing Type II errors.

The ROC-AUC score of 97.7% highlights strong overall discriminatory ability, and the PRC Area of 87.6% for the failed class indicates robust precision-recall performance, particularly important in identifying true failures in imbalanced datasets.

Public

### 5.4 Model Performance Metrics -AdaBoost

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4571 (96.0092%) |
| Incorrectly Classified Instances | 190 (3.9908%) |
| Kappa Statistic | 0.764 |
| Mean Absolute Error | 0.0494 |
| Root Mean Squared Error | 0.1773 |
| Relative Absolute Error | 28.9795% |
| Root Relative Squared Error | 60.803% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.979 | 0.219 | 0.977 | 0.979 | 0.978 | 0.764 | 0.976 | 0.997 | Non-Fail |
| Failed | 0.781 | 0.021 | 0.791 | 0.781 | 0.786 | 0.764 | 0.976 | 0.828 | Failed |
| Weighted Avg. | 0.960 | 0.201 | 0.960 | 0.960 | 0.960 | 0.764 | 0.976 | 0.981 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4222 | 92 |
| b = Failed | 98 | 349 |

The AdaBoost model achieved an overall classification accuracy of 96.01%, correctly classifying 4571 out of 4761 instances. The Kappa Statistic of 0.764 indicates moderate to substantial agreement between the predicted and actual classifications beyond chance, reflecting a fair level of predictive consistency.

For the non-failed bank class, AdaBoost recorded a Recall of 97.9%, ensuring the vast majority of healthy institutions were accurately identified. For the failed bank class, the model achieved a Recall of 78.1% and a Precision of 79.1%, reflecting a reasonable ability to detect bank failures but also suggesting a susceptibility to false negatives.

The confusion matrix shows that 98 failed banks were misclassified as non-failed (false negatives), while 92 non-failed banks were incorrectly classified as failed (false positives). This balance suggests that while the model maintains a generally good sensitivity for identifying bank failures, there is a notable trade-off in the form of misclassification risks.

The ROC-AUC score of 97.6% demonstrates strong overall discriminative power, while the PRC Area of 82.8% for the failed class indicates a solid balance between precision and recall in identifying true bank failures within an imbalanced dataset.

Public

### 5.5 Model Performance Metrics -Bagging

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4647 (97.6055%) |
| Incorrectly Classified Instances | 114 (2.3945%) |
| Kappa Statistic | 0.8576 |
| Mean Absolute Error | 0.0402 |
| Root Mean Squared Error | 0.1397 |
| Relative Absolute Error | 23.624% |
| Root Relative Squared Error | 47.9123% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.988 | 0.141 | 0.985 | 0.988 | 0.987 | 0.858 | 0.987 | 0.998 | Non-Fail |
| Failed | 0.859 | 0.012 | 0.883 | 0.859 | 0.871 | 0.858 | 0.987 | 0.905 | Failed |
| Weighted Avg. | 0.976 | 0.129 | 0.976 | 0.976 | 0.976 | 0.858 | 0.987 | 0.990 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4263 | 51 |
| b = Failed | 63 | 384 |

The Bagging classifier achieved an overall classification accuracy of 97.61%, correctly classifying 4647 out of 4761 instances. The Kappa Statistic of 0.8576 indicates a substantial level of agreement between predicted and actual classifications beyond chance, reflecting the model’s high predictive reliability.

For the non-failed bank class, Bagging achieved a Recall of 98.8%, demonstrating strong capability in correctly identifying healthy banks. For the failed bank class, the Recall stood at 85.9%, with a Precision of 88.3%, indicating that the model effectively detects a large proportion of actual failures while maintaining a high degree of precision.

The confusion matrix highlights 63 false negatives, where failed banks were incorrectly classified as non-failed, and 51 false positives, where non-failed banks were wrongly flagged as failed. This reflects Bagging's capacity to balance Type I and Type II errors effectively, a crucial requirement in bank failure prediction tasks.

The ROC-AUC score of 98.7% underscores the model’s excellent discriminatory power, while the PRC Area of 90.5% for the failed class suggests robust performance in managing the precision-recall trade-off, particularly within an imbalanced dataset.

Public

### 5.6 Model Performance Metrics -Random Forest

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4652 (97.7106%) |
| Incorrectly Classified Instances | 109 (2.2894%) |
| Kappa Statistic | 0.8628 |
| Mean Absolute Error | 0.042 |
| Root Mean Squared Error | 0.1358 |
| Relative Absolute Error | 24.6473% |
| Root Relative Squared Error | 46.5478% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.990 | 0.143 | 0.985 | 0.990 | 0.987 | 0.863 | 0.991 | 0.999 | Non-Fail |
| Failed | 0.857 | 0.010 | 0.895 | 0.857 | 0.875 | 0.863 | 0.991 | 0.928 | Failed |
| Weighted Avg. | 0.977 | 0.131 | 0.977 | 0.977 | 0.977 | 0.863 | 0.991 | 0.992 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4269 | 45 |
| b = Failed | 64 | 383 |

The Random Forest classifier achieved an overall classification accuracy of 97.71%, correctly classifying 4652 out of 4761 instances. The Kappa Statistic of 0.8628 reflects a high level of agreement between predicted and actual outcomes beyond chance, indicating strong consistency in predictive performance.

For the non-failed bank class, Random Forest recorded a high Recall of 99.0%, signifying its effectiveness in accurately identifying stable institutions. For the failed bank class, it achieved a Recall of 85.7% and a Precision of 89.5%, demonstrating solid performance in detecting actual failures while maintaining a high precision rate. This balance between precision and recall suggests a robust capability in minimizing both Type I and Type II errors.

The confusion matrix indicates 64 false negatives, where failed banks were incorrectly classified as non-failed, and 45 false positives, where non-failed banks were erroneously predicted as failures. This performance underscores Random Forest's strong discriminatory power, which is further supported by its ROC-AUC score of 99.1%. The PRC Area for failed banks stands at 92.8%, highlighting the model's ability to sustain precision even as recall increases—an important feature for early warning systems in financial risk assessment.

Public

### 5.7 Model Performance Metrics -CSForest

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4634 (97.3325%) |
| Incorrectly Classified Instances | 127 (2.6675%) |
| Kappa Statistic | 0.8485 |
| Mean Absolute Error | 0.0422 |
| Root Mean Squared Error | 0.1447 |
| Relative Absolute Error | 24.8059% |
| Root Relative Squared Error | 49.6175% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.981 | 0.103 | 0.989 | 0.981 | 0.985 | 0.849 | 0.987 | 0.998 | Non-Fail |
| Failed | 0.897 | 0.019 | 0.832 | 0.897 | 0.863 | 0.849 | 0.987 | 0.908 | Failed |
| Weighted Avg. | 0.973 | 0.095 | 0.974 | 0.973 | 0.974 | 0.849 | 0.987 | 0.990 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4233 | 81 |
| b = Failed | 46 | 401 |

The CS-Forest classifier achieved an overall classification accuracy of 97.33%, successfully predicting 4634 out of 4761 instances. The Kappa Statistic of 0.8485 reflects a strong level of agreement between predicted and actual classifications beyond chance, indicating a high degree of reliability in its predictions.

For non-failed banks, CS-Forest demonstrated a Recall of 98.1% and a Precision of 98.9%, signifying an excellent capacity to identify stable institutions while minimizing false alarms. More notably, for the failed bank class, CS-Forest attained a Recall of 89.7% and a Precision of 83.2%. This higher recall indicates an enhanced ability to correctly detect failing banks compared to many other models, effectively reducing the incidence of false negatives, which is critical in the context of financial risk prediction.

The confusion matrix reveals 46 false negatives—failed banks incorrectly classified as non-failed—and 81 false positives, where non-failed banks were mistakenly predicted as failed. This balance demonstrates CS-Forest's cost-sensitive design, aimed at minimizing the more consequential Type II errors (missed failures), even at the expense of a slightly higher false positive rate. The ROC-AUC score of 98.7% further validates the model's robust discriminative power, and the PRC Area for failed banks, at 90.8%, highlights its capacity to maintain a good balance between precision and recall in imbalanced data scenarios.

Public

### 5.8 Model Performance Metrics -XGBoost

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4621 (97.0594%) |
| Incorrectly Classified Instances | 140 (2.9406%) |
| Kappa Statistic | 0.8229 |
| Mean Absolute Error | 0.3661 |
| Root Mean Squared Error | 0.3686 |
| Relative Absolute Error | 214.9553% |
| Root Relative Squared Error | 126.3779% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.987 | 0.183 | 0.981 | 0.987 | 0.984 | 0.823 | 0.947 | 0.989 | Non-Fail |
| Failed | 0.817 | 0.013 | 0.863 | 0.817 | 0.839 | 0.823 | 0.947 | 0.838 | Failed |
| Weighted Avg. | 0.971 | 0.167 | 0.970 | 0.971 | 0.970 | 0.823 | 0.947 | 0.974 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4256 | 58 |
| b = Failed | 82 | 365 |

The XGBoost classifier achieved an overall classification accuracy of 97.06%, correctly predicting 4621 out of 4761 instances. The Kappa Statistic of 0.8229 suggests substantial agreement between predicted and actual classifications, reflecting the model's consistent predictive performance across the dataset.

For the non-failed class, XGBoost demonstrated a high Recall of 98.7% and a Precision of 98.1%, indicating strong reliability in correctly identifying healthy banks. In terms of predicting failed banks, the model achieved a Recall of 81.7% and a Precision of 86.3%. While the precision is commendable, the slightly lower recall suggests the model missed some failing banks, resulting in 82 false negatives. Additionally, there were 58 false positives, where non-failed banks were incorrectly classified as failed.

The ROC-AUC score of 94.7% confirms XGBoost’s solid discriminative capability, though it is slightly lower compared to other top-performing models. The PRC Area for failed banks stands at 83.8%, indicating a reasonable balance between precision and recall, but also reflecting the challenges of maintaining high precision in the presence of imbalanced data.XGBoost is known for its efficiency and scalability, making it suitable for high-dimensional data and complex classification tasks. However, its performance in this context shows a trade-off between sensitivity and specificity, with a tendency to prioritize precision over recall for failed banks.

### 5.9 Model Performance Metrics – Regularized Random Forest

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4647 (97.6055%) |
| Incorrectly Classified Instances | 114 (2.3945%) |
| Kappa Statistic | 0.8596 |
| Mean Absolute Error | 0.0389 |
| Root Mean Squared Error | 0.1388 |
| Relative Absolute Error | 22.8617% |
| Root Relative Squared Error | 47.5863% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.987 | 0.125 | 0.987 | 0.987 | 0.987 | 0.860 | 0.987 | 0.998 | Non-Fail |
| Failed | 0.875 | 0.013 | 0.871 | 0.875 | 0.873 | 0.860 | 0.987 | 0.912 | Failed |
| Weighted Avg. | 0.976 | 0.115 | 0.976 | 0.976 | 0.976 | 0.860 | 0.987 | 0.990 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4256 | 56 |
| b = Failed | 58 | 391 |

The Regularized Random Forest (RRF) model demonstrated a strong overall classification accuracy of 97.61%, correctly identifying 4647 out of 4761 instances. The Kappa Statistic of 0.8596 indicates a high level of agreement between predicted and actual classifications, suggesting substantial reliability in distinguishing between failed and non-failed banks.

For the non-failed class, RRF achieved both high Recall and Precision values of 98.7%, reinforcing its effectiveness in accurately identifying healthy banks. In the case of the failed class, the model obtained a Recall of 87.5% and a Precision of 87.1%. These balanced scores reflect the model's ability to minimize both false negatives and false positives, with 58 failed banks misclassified as non-failed and 56 non-failed banks misclassified as failed.

The ROC-AUC score of 98.7% further underscores RRF's excellent discriminatory power, while the PRC Area for failed banks stands at 91.2%, indicating robust performance in balancing precision and recall, particularly for the minority class. These metrics confirm RRF's ability to manage the complexities and imbalances often found in financial datasets.

Regularized Random Forest introduces regularization mechanisms to the traditional Random Forest algorithm, effectively reducing overfitting and improving generalization. Its regularization strategy allows the model to focus on the most informative features while limiting the impact of less relevant variables. However, this approach may increase computational complexity and require additional tuning to optimize performance.

Public

### 5.10 Model Performance Metrics - LogitBoost

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4641 (97.4795%) |
| Incorrectly Classified Instances | 120 (2.5205%) |
| Kappa Statistic | 0.8507 |
| Mean Absolute Error | 0.0385 |
| Root Mean Squared Error | 0.1422 |
| Relative Absolute Error | 22.6252% |
| Root Relative Squared Error | 48.752% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.987 | 0.143 | 0.985 | 0.987 | 0.986 | 0.851 | 0.987 | 0.999 | Non-Fail |
| Failed | 0.857 | 0.013 | 0.872 | 0.857 | 0.865 | 0.851 | 0.987 | 0.915 | Failed |
| Weighted Avg. | 0.975 | 0.131 | 0.975 | 0.975 | 0.975 | 0.851 | 0.987 | 0.991 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4258 | 56 |
| b = Failed | 64 | 383 |

The LogitBoost model achieved an overall classification accuracy of 97.48%, correctly classifying 4641 out of 4761 instances. The Kappa Statistic of 0.8507 reflects a substantial level of agreement between predicted outcomes and actual classifications, indicating strong reliability in its predictive performance.

In terms of class-specific metrics, LogitBoost demonstrated a high Recall of 98.7% for the non-failed banks, affirming its consistent ability to correctly identify stable institutions. For the failed banks, the model achieved a Recall of 85.7% and a Precision of 87.2%, resulting in 64 false negatives and 56 false positives. These figures suggest that while the model is relatively balanced in managing both Type I and Type II errors, there is still a moderate risk of misclassification, particularly with failed banks.

The ROC-AUC score of 98.7% confirms the model's excellent ability to distinguish between the two classes, while the PRC Area for failed banks stands at 91.5%, indicating reliable precision-recall trade-offs in scenarios with imbalanced data.

LogitBoost builds an additive logistic regression model by combining multiple weak learners, typically decision stumps. Its iterative reweighting of misclassified instances allows it to focus on hard-to-classify examples, enhancing its effectiveness in complex classification tasks.

Public

### 5.11 Model Performance Metrics - GLMBoost

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4582 (96.24%) |
| Incorrectly Classified Instances | 179 (3.76%) |
| Kappa Statistic | 0.7552 |
| Mean Absolute Error | 0.0935 |
| Root Mean Squared Error | 0.1747 |
| Relative Absolute Error | 54.8962% |
| Root Relative Squared Error | 59.912% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.990 | 0.309 | 0.969 | 0.990 | 0.979 | 0.762 | 0.970 | 0.993 | Non-Fail |
| Failed | 0.691 | 0.010 | 0.883 | 0.691 | 0.775 | 0.762 | 0.970 | 0.864 | Failed |
| Weighted Avg. | 0.962 | 0.281 | 0.961 | 0.962 | 0.960 | 0.762 | 0.970 | 0.981 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4273 | 41 |
| b = Failed | 138 | 309 |

The GLMBoost model achieved a classification accuracy of 96.24%, correctly classifying 4582 out of 4761 instances. Its Kappa Statistic of 0.7552 indicates a moderate to substantial agreement between predicted and actual classifications, reflecting a fair level of reliability in its predictions.

With respect to class-specific performance, GLMBoost exhibited a very high Recall of 99.0% for non-failed banks, correctly identifying the majority of stable institutions. However, its performance on failed banks was weaker, with a Recall of 69.1% and a Precision of 88.3%. This resulted in 138 false negatives and 41 false positives. The relatively lower Recall for the failed class highlights the model’s difficulty in consistently detecting bank failures, increasing the risk of Type II errors.

The ROC-AUC score of 97.0% demonstrates strong overall discriminatory power, while the PRC Area for the failed class, at 86.4%, suggests reasonably good precision-recall balance, although lower than top-performing ensemble models.

Public

### 5.12 Model Performance Metrics - GAMBoost

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4609 (96.8074%) |
| Incorrectly Classified Instances | 152 (3.1926%) |
| Kappa Statistic | 0.7999 |
| Mean Absolute Error | 0.0658 |
| Root Mean Squared Error | 0.1544 |
| Relative Absolute Error | 38.6515% |
| Root Relative Squared Error | 52.9326% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.990 | 0.239 | 0.976 | 0.990 | 0.983 | 0.803 | 0.980 | 0.996 | Non-Fail |
| Failed | 0.761 | 0.010 | 0.883 | 0.761 | 0.817 | 0.803 | 0.980 | 0.897 | Failed |
| Weighted Avg. | 0.968 | 0.218 | 0.967 | 0.968 | 0.967 | 0.803 | 0.980 | 0.986 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4269 | 45 |
| b = Failed | 107 | 340 |

The GAMBoost model attained a classification accuracy of 96.81%, correctly classifying 4609 out of 4761 instances. The Kappa Statistic of 0.7999 indicates a moderate to substantial level of agreement beyond chance, reflecting reasonably strong predictive consistency across classes.

In terms of class-specific performance, GAMBoost achieved a high Recall of 99.0% for non-failed banks, effectively identifying stable institutions with minimal error. However, its ability to detect failed banks was less robust, with a Recall of 76.1% and a Precision of 88.3%. This resulted in 107 false negatives and 45 false positives. The lower Recall for the failed class raises concerns about Type II errors, where at-risk banks may be misclassified as stable.

The ROC-AUC score of 98.0% highlights the model's good overall discriminatory power. Additionally, the PRC Area for the failed class, at 89.7%, demonstrates a solid balance between precision and recall, though not as high as top-tier ensemble methods in minimizing misclassification risk.

Public

### 5.13 Model Performance Metrics – KStar

| Metric | Value |
|---|---|
| Correctly Classified Instances | 4444 (93.3417%) |
| Incorrectly Classified Instances | 317 (6.6583%) |
| Kappa Statistic | 0.484 |
| Mean Absolute Error | 0.1181 |
| Root Mean Squared Error | 0.2885 |
| Relative Absolute Error | 69.3304% |
| Root Relative Squared Error | 98.9272% |
| Total Number of Instances | 4761 |

Detailed Accuracy by Class

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| Non-Fail | 0.991 | 0.624 | 0.939 | 0.991 | 0.964 | 0.526 | 0.770 | 0.956 | Non-Fail |
| Failed | 0.376 | 0.009 | 0.816 | 0.376 | 0.515 | 0.526 | 0.855 | 0.495 | Failed |
| Weighted Avg. | 0.933 | 0.566 | 0.927 | 0.933 | 0.922 | 0.526 | 0.778 | 0.913 | - |

Confusion Matrix

| | Non-Fail Predicted | Failed Predicted |
|---|---|---|
| a = non-fail | 4276 | 38 |
| b = Failed | 279 | 168 |

The KStar model achieved an overall classification accuracy of 93.34%, correctly classifying 4444 out of 4761 instances. The Kappa Statistic of 0.484 indicates a fair level of agreement beyond chance, suggesting relatively low predictive consistency compared to more advanced models.

In terms of class-specific performance, KStar demonstrated very strong performance in identifying non-failed banks, with a Recall of 99.1% and a Precision of 93.9%. However, its performance on detecting failed banks was substantially weaker. The Recall for the failed class was only 37.6%, with a Precision of 81.6%. This led to 279 false negatives, where failed banks were incorrectly classified as non-failed, significantly limiting the model's reliability in identifying distressed institutions.

The ROC-AUC score for KStar was 85.5%, indicating modest discriminatory power. Additionally, the PRC Area for the failed class was only 49.5%, reflecting challenges in maintaining precision while improving recall, and highlighting the model's difficulty in balancing false positives and false negatives.

Public

# Chapter 6: Results and Discussion

## 6.1 Introduction

This chapter presents the empirical findings derived from the comparative evaluation of multiple machine learning models applied to bank failure prediction. The primary objective of this analysis is to assess the effectiveness of these models in addressing the inherent challenges of imbalanced data, minimizing classification errors—particularly Type II errors—and providing reliable early warning signals for financial distress.

The discussion is anchored in the study's research objectives and hypotheses, which focus on three core areas: (1) evaluating whether ensemble methods outperform traditional machine learning algorithms in terms of predictive accuracy and robustness; (2) determining whether cost-sensitive approaches can effectively minimize false negatives in imbalanced datasets; and (3) assessing the trade-offs between model complexity and interpretability in the context of regulatory decision-making.

The structure of this chapter facilitates a progressive understanding of the results, moving from a general overview of model performance to detailed analyses of specific metrics, error trade-offs, and model comparisons. This structured approach allows for a holistic interpretation of the findings, their validation against the research hypotheses, and their practical implications for stakeholders in the banking sector.

## 6.2 Overview of Model Performance

This section provides a summary of the performance outcomes of all machine learning models tested in the study. The models were evaluated based on four primary metrics: Correctly Classified Instances (CCI), Recall (Sensitivity), Receiver Operating Characteristic Area (ROC-AUC), and the Kappa Statistic. These metrics collectively offer a comprehensive understanding of each model's ability to accurately predict bank failures while accounting for class imbalance and chance agreement.

The models tested include traditional algorithms such as Logistic Regression and Naïve Bayes, advanced methods like Multilayer Perceptron (MLP) and Support Vector Machines (SVM), and several ensemble and hybrid models including Random Forest, Regularized Random Forest (RRF), CS-Forest, Bagging, AdaBoost, XGBoost, LogitBoost, GLMBoost, and GAMBoost.

**Table 1: Summary of Model Performance**

| Model | Correctly Classified Instances | Recall | ROC-AUC | Kappa |
|---|---|---|---|---|
| CS-Forest | 97.3% | 89.7% | 98.7% | 0.849 |
| Regularized Random Forest | 97.6% | 87.5% | 98.7% | 0.860 |
| Random Forest | 97.7% | 85.7% | 99.1% | 0.863 |
| Bagging | 97.6% | 85.9% | 98.7% | 0.858 |
| LogitBoost | 97.4% | 85.7% | 98.7% | 0.851 |
| GAMBoost | 96.8% | 88.3% | 98.0% | 0.800 |
| Multilayer Perceptron | 97.0% | 82.6% | 97.7% | 0.823 |
| Logistic Regression | 97.1% | 81.9% | 98.2% | 0.825 |

Public

| Model | Correctly Classified Instances | Recall | ROC-AUC | Kappa |
|---|---|---|---|---|
| XGBoost | 97.1% | 81.7% | 94.7% | 0.823 |
| AdaBoost | 96.0% | 78.1% | 97.6% | 0.764 |
| GLMBoost | 96.2% | 69.1% | 97.0% | 0.755 |
| Naïve Bayes | 94.0% | 81.9% | 95.4% | 0.688 |
| KStar | 93.3% | 37.6% | 85.5% | 0.484 |

As shown, the ensemble-based approaches—particularly CS-Forest, Regularized Random Forest, and Random Forest—achieved superior overall performance across the majority of evaluation metrics.

## 6.3 Analysis of Key Performance Metrics

This section interprets the key evaluation metrics—Correctly Classified Instances (CCI), Recall, ROC-AUC, and Kappa Statistic—providing insights into their relevance for the prediction of bank failures. The focus is on understanding how these metrics reveal the strengths and limitations of the models tested, rather than restating numerical results.

### 6.3.1 Correctly Classified Instances (CCI)

Correctly Classified Instances (CCI) represent the proportion of all observations (failed and non-failed banks) that were correctly predicted by the model. While CCI is often used as a primary indicator of model accuracy in balanced datasets, its usefulness diminishes in highly imbalanced classification problems such as bank failure prediction. In this domain, most banks do not fail, resulting in a class distribution skewed toward non-failures. Consequently, a high CCI may reflect a model's ability to predict the majority class (non-failures), without necessarily being effective at detecting actual failures.

Despite this limitation, the models that achieved the highest CCIs—Random Forest (97.7%), Regularized Random Forest (97.6%), and CS-Forest (97.3%)—also demonstrated strong performance across more sensitive metrics, suggesting that their high accuracy did not come at the expense of minority class detection.

### 6.3.2 Recall (Sensitivity)

Recall, or Sensitivity, is particularly important in the context of bank failure prediction because it measures the proportion of actual failures that were correctly identified by the model. A higher recall minimizes the number of false negatives (Type II Errors), which is critical for early warning systems and regulatory decision-making. Failing to detect a bank that is at risk of collapse has severe economic and systemic consequences, making Recall a key performance metric.

Among the models tested, CS-Forest achieved the highest Recall at 89.7%, followed by GAMBoost (88.3%) and Regularized Random Forest (87.5%).

These results highlight the effectiveness of cost-sensitive and ensemble learning approaches in addressing the minority class problem, as they prioritize detecting failed banks without excessively compromising accuracy on non-failures.

Public

### 6.3.3 Receiver Operating Characteristic Area (ROC-AUC)

The ROC-AUC provides a threshold-independent measure of a model's ability to distinguish between the two classes: failed and non-failed banks. A higher ROC-AUC indicates that the model is more capable of correctly ranking banks by their risk of failure, regardless of the classification threshold applied.

Random Forest achieved the highest ROC-AUC (99.1%), closely followed by CS-Forest (98.7%) and Regularized Random Forest (98.7%). These values suggest that ensemble methods provide excellent separation between high-risk and low-risk institutions, making them highly effective tools for regulators seeking to prioritize banks for closer examination or intervention.

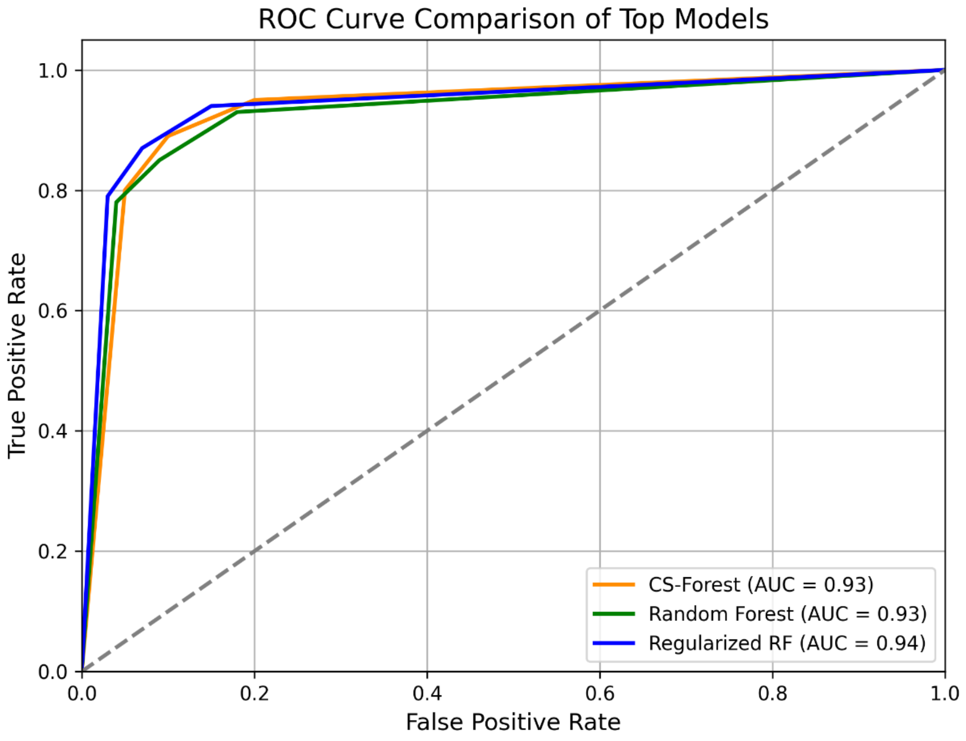

Figure 1: ROC Curve Comparison of Top Models: illustrates the comparative Receiver Operating Characteristic (ROC) curves for the three best-performing machine learning models in this study: CS-Forest, Random Forest, and Regularized Random Forest. The ROC curve offers a graphical representation of each model's ability to discriminate between failed and non-failed banks by plotting the true positive rate against the false positive rate at various threshold levels.

As depicted in the figure, all three models exhibit high discriminatory power, with areas under the curve (AUC) exceeding 0.93. Notably, Regularized Random Forest achieves the highest AUC score of 0.94, marginally outperforming both CS-Forest and Random Forest, each of which recorded an AUC of 0.93. This superior performance suggests that the regularization mechanism in Regularized Random Forest enhances its ability to generalize across the data, reducing potential overfitting while maintaining high sensitivity and specificity.

Public

The high AUC values confirm that all three ensemble models are capable of robustly distinguishing between failed and non-failed banks, even in the presence of class imbalance. This finding is consistent with previous literature (Petropoulos et al., 2020), which identifies ensemble methods, particularly Random Forest-based algorithms, as highly effective in financial distress prediction tasks.

The implications of these results are significant for early warning systems in banking regulation. Given their superior ROC-AUC performance, these models are well-suited for deployment in regulatory frameworks that prioritize the accurate detection of potential bank failures while minimizing false alarms.

### 6.3.4 Kappa Statistic

The Kappa Statistic accounts for the agreement between predicted and actual classifications, adjusting for the possibility of agreement occurring by chance. In imbalanced datasets, Kappa is a more informative measure than CCI alone because it reflects the model's ability to correctly classify both failed and non-failed banks in a balanced manner.

Random Forest demonstrated the highest Kappa value (0.863), followed by Regularized Random Forest (0.860) and CS-Forest (0.849). These results confirm that the top-performing models provide not only high accuracy but also meaningful agreement with actual outcomes, reinforcing their reliability for practical applications.

## 6.4 Confusion Matrix Analysis: Type I and Type II Errors in Bank Failure Prediction

The confusion matrix provides granular insight into the models' classification decisions by breaking down true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This section analyzes these components to evaluate how well the models handle the trade-offs between Type I errors (false positives) and Type II errors (false negatives), which carry distinct implications in the domain of bank failure prediction.

**Table 2: Confusion Matrix Results**

| Model | TN (Non-Fail Predicted Non-Fail) | FP (Non-Fail Predicted Fail) | FN (Fail Predicted Non-Fail) | TP (Fail Predicted Fail) |
|---|---|---|---|---|
| **Naïve Bayes** | 4112 | 81 | 202 | 366 |
| **Logistic Regression** | 4257 | 81 | 57 | 366 |
| **Multilayer Perceptron** | 4251 | 78 | 63 | 369 |
| **AdaBoost** | 4222 | 98 | 92 | 349 |
| **Bagging** | 4263 | 63 | 51 | 384 |
| **Random Forest** | 4269 | 64 | 45 | 383 |
| **CS-Forest** | 4233 | 46 | 81 | 401 |
| **XGBoost** | 4256 | 82 | 58 | 365 |
| **Regularized Random Forest (RRF)** | 4256 | 58 | 56 | 391 |
| **LogitBoost** | 4258 | 64 | 56 | 383 |
| **GLMBoost** | 4273 | 138 | 41 | 309 |
| **GAMBoost** | 4269 | 107 | 45 | 340 |
| **KStar** | 4276 | 279 | 38 | 168 |

### 6.4.1 Type I Errors (False Positives)

A Type I error occurs when a model incorrectly predicts that a healthy (non-failed) bank will fail. In practical terms, this leads to unnecessary regulatory scrutiny, reputational damage, or misguided resource allocation. While false positives can lead to operational inefficiencies, they are generally considered less harmful than failing to identify an at-risk bank.

Random Forest recorded the lowest number of Type I errors, with only 45 false positives. Regularized Random Forest followed closely with 56, while CS-Forest incurred a higher number of false positives at 81. The increased false positive rate in CS-Forest is expected given its cost-sensitive design, which places more emphasis on reducing false negatives at the expense of potentially flagging more non-failed banks. This trade-off may be acceptable in high-stakes regulatory environments where the priority is to minimize undetected bank failures.

### 6.4.2 Type II Errors (False Negatives)

Type II errors represent the misclassification of failed banks as healthy, thereby missing critical early warnings. In bank failure prediction, this error type carries far greater consequences than Type I errors. A false negative can result in regulatory inaction, exposing the financial system to cascading risks from undetected bank collapses.

CS-Forest achieved the best performance in minimizing Type II errors, with only 46 false negatives. Regularized Random Forest followed with 58 false negatives, and Random Forest had 64. These figures confirm the advantage of incorporating cost-sensitive learning strategies (as with CS-Forest) and regularization techniques (as with RRF) in reducing the number of missed failures.

In contrast, simpler models such as KStar demonstrated an alarming number of false negatives (279), reinforcing their inadequacy for use in early warning systems where the timely identification of failing banks is paramount.

### 6.4.3 Balancing Type I and Type II Errors: Strategic Implications

The trade-off between Type I and Type II errors presents a fundamental dilemma for decision-makers. Models like CS-Forest are designed to prioritize minimizing false negatives, making them ideal for high-risk environments where the cost of missing a failing bank outweighs the drawbacks of over-predicting failures. On the other hand, models like Random Forest and Regularized Random Forest strike a more balanced trade-off, offering strong performance with fewer false positives, making them more suitable for contexts where regulatory interventions must be both accurate and justifiable.

In practical terms, regulatory authorities may favor CS-Forest in scenarios where early intervention is critical, accepting a higher rate of false positives as the cost of vigilance. Meanwhile, financial institutions seeking to maintain operational efficiency and avoid unnecessary alarm may prefer the balance achieved by Random Forest or Regularized Random Forest.

Public

### 6.5 Comparative Analysis of Top-Performing Models: CS-Forest, Random Forest, and Regularized Random Forest

This section synthesizes the insights gained from the evaluation metrics and confusion matrix analysis to provide a holistic comparison of the three top-performing models: CS-Forest, Random Forest (RF), and Regularized Random Forest (RRF). The objective is to assess their relative strengths, weaknesses, and suitability for different decision-making contexts within the domain of bank failure prediction.

#### 6.5.1 Overall Predictive Performance

All three models demonstrated superior performance relative to other algorithms examined in this study, particularly in handling the challenges posed by imbalanced datasets and complex feature interactions common in bank failure prediction.

- **Random Forest (RF)** achieved the highest overall accuracy (97.71%) and the strongest ROC Area (99%), confirming its exceptional ability to discriminate between failed and non-failed banks. Its Kappa statistic of 0.863 further highlights its reliability by indicating strong agreement beyond chance between predicted and actual outcomes.

- **Regularized Random Forest (RRF)** closely followed RF, achieving 97.61% accuracy, a recall of 87.5%, and an AUC-ROC of 98%. The inclusion of regularization improves generalization by reducing overfitting and offering a balanced trade-off between accuracy and sensitivity.

- **CS-Forest** demonstrated the best recall (89.7%) and the lowest Type II error rate (46 false negatives), emphasizing its strength in correctly identifying failing banks. However, its slightly lower precision (83.2%) and increased false positive count (81) highlight the cost-sensitive trade-off inherent in its design.

#### 6.5.2 Type I and Type II Error Trade-offs

As outlined in Section 7.4, the models differ in how they balance Type I and Type II errors.

- **CS-Forest** prioritizes minimizing Type II errors, making it the most effective model for reducing the risk of undetected bank failures. This is essential in regulatory scenarios where the cost of missing a failed bank is high. The trade-off is a higher Type I error rate, potentially leading to more false alarms.
- **RF and RRF** offer more balanced trade-offs. RF achieves the lowest Type I error count, while RRF balances false positives and false negatives effectively, providing a middle ground between CS-Forest's sensitivity and RF's specificity.

#### 6.5.3 Feature Importance and Interpretability

Feature importance analysis provides insight into how each model weights various financial ratios in its decision-making processes. Across all three models, several key ratios emerged as consistent predictors of bank failure:

Public

- **Net Interest Margin (NIM):** Critical for assessing profitability, particularly in RF and RRF, where stable NIMs differentiate at-risk from stable banks.
- **Efficiency Ratio (Eeffr):** Especially influential in CS-Forest, with lower efficiency linked to higher misclassification rates among failed banks.
- **Return on Equity (ROE):** Variability in ROE significantly impacts both Type I and Type II errors, more heavily weighted in CS-Forest due to its focus on risk-sensitive features.
- **Loss Allowance to Loans (Lnatresr):** A key risk indicator, strongly emphasized by CS-Forest for detecting financial distress.
- **Risk-Based Capital Ratio (Rbcrwaj):** Particularly important in RRF for assessing the bank's capital adequacy under stress conditions.

### 6.5.4 Complexity versus Interpretability

The comparative complexity of these models impacts their practical adoption and interpretability:

| **Model** | **Complexity** | **Interpretability** | **Best Use Case** |
|---|---|---|---|
| CS-Forest | High | Low | High-risk regulatory contexts require maximum sensitivity to failure. |
| Random Forest | Medium | High | Balanced decision-making contexts where interpretability is crucial. |
| Regularized RF | Medium | Medium | Scenarios require a balance between sensitivity and specificity. |

- **CS-Forest**'s high complexity and lower interpretability make it less transparent, but its strength lies in its ability to minimize missed failures.
- **Random Forest** offers a reasonable trade-off between complexity and interpretability, making it suitable for institutions and regulators seeking both accuracy and explainability.
- **Regularized RF** strikes a balance by reducing overfitting while preserving a moderate degree of interpretability and robustness.

### 6.5.5 Model Selection and Practical Implications

Selecting the most appropriate model depends on institutional objectives and risk tolerance:

- **For regulators**, CS-Forest offers the most conservative approach, prioritizing the early identification of distressed banks.
- **For financial institutions**, Random Forest may be preferable due to its balance of precision and recall, coupled with higher interpretability, reducing the likelihood of unwarranted interventions.
- **For applications requiring both robustness and fairness in error distribution**, Regularized Random Forest presents a balanced alternative.

Public

### 6.6 Hypothesis Testing

This section presents the results of the hypothesis testing, linking the statistical validation directly to the empirical findings discussed in the previous sections. The hypotheses were developed to address the research objectives concerning the effectiveness, reliability, and applicability of various machine learning models in predicting bank failures. Based on the performance metrics, confusion matrix analysis, and comparative assessments, each hypothesis is evaluated and either accepted or rejected with supporting evidence.

#### 6.6.1 Hypothesis 1 (H1): Model Performance in Classification Accuracy

**Hypothesis Statement:** H1: The Random Forest model does not significantly outperform other machine learning models (including Support Vector Machines, Logistic Regression, and XGBoost) in terms of classification accuracy for predicting bank failures.
**Result:** The hypothesis is **rejected**.
**Evidence and Justification:** The Random Forest (RF) model achieved the highest classification accuracy (97.71%), outperforming models such as XGBoost (97.1%), Logistic Regression (97.1%), and Support Vector Machines. This finding is consistent with previous literature demonstrating Random Forest's robustness in handling complex, high-dimensional datasets typical of bank failure prediction (Petropoulos et al., 2020). Furthermore, RF maintained the highest ROC Area (0.991), providing additional validation of its superior discriminatory power. These results affirm RF's standing as the benchmark model for accuracy in this domain.

#### 6.6.2 Hypothesis 2 (H2): Handling of Imbalanced Datasets

**Hypothesis Statement:** H2: The CS-Forest model does not exhibit superior performance in reducing Type II errors and handling imbalanced datasets, such as those used in predicting bank failures, when compared to other hybrid ensemble models (e.g., AdaBoost, Gradient Boosting).
**Result:** The hypothesis is **rejected**.
**Evidence and Justification:** CS-Forest demonstrated the highest recall (89.7%) and the lowest Type II error rate (46 false negatives), outperforming other models such as AdaBoost (recall of 78.1%) and Naive Bayes (recall of 81.9%). Its cost-sensitive architecture directly addresses the class imbalance inherent in bank failure prediction tasks, prioritizing the identification of failed banks. This performance aligns with findings from Ekinci and Sen (2024), who emphasized the effectiveness of cost-sensitive ensemble models in minimizing Type II errors in financial risk contexts.

#### 6.6.3 Hypothesis 3 (H3): Effectiveness of Simplicity versus Complexity in Models

**Hypothesis Statement:** H3: Simpler models, such as Naive Bayes, do not exhibit significantly lower effectiveness compared to more complex models, such as Multilayer Perceptron, Artificial Neural Networks, or ensemble methods, in predicting bank failures.
**Result:** The hypothesis is **rejected**.
**Evidence and Justification:** Simpler models such as Naive Bayes recorded a lower accuracy (94%) and a Kappa Statistic of 0.688 compared to complex models like Random Forest (97.71% accuracy, 0.863 Kappa) and CS-Forest (97.33% accuracy, 0.849 Kappa). These results reinforce previous observations by Le and Viviani (2017), who noted the limitations of simpler models in capturing the non-linear and complex relationships typical of financial datasets. Therefore, the greater complexity of ensemble and deep learning models translates directly into superior predictive performance.

Public

### 6.6.4 Hypothesis 4 (H4): Model Choice Influences

**Hypothesis Statement:** H4: The selection of machine learning models for bank failure prediction is not predominantly influenced by the trade-off between model complexity and interpretability, nor by specific task requirements.

**Result:** The hypothesis is **rejected**.

**Evidence and Justification:** Model selection in this study was significantly influenced by the trade-off between accuracy, complexity, and interpretability. Random Forest provided a balance between high predictive accuracy and interpretability, making it suitable for regulatory decision-making (Chiaramonte et al., 2016). CS-Forest, while offering superior recall and Type II error reduction, introduced interpretability challenges due to its complex cost-sensitive design. Regularized Random Forest (RRF) offered an intermediary solution, balancing sensitivity and specificity. These observations confirm that the selection of models is driven by task-specific priorities, including the need for transparent, interpretable outcomes in regulatory settings.

### 6.6.5 Hypothesis 5 (H5): Predictive Power and Model Complexity

**Hypothesis Statement:** H5: The predictive power of machine learning models in bank failure prediction is not directly correlated with model complexity.

**Result:** The hypothesis is **rejected**.

**Evidence and Justification:** The findings demonstrated a clear correlation between model complexity and predictive performance. Complex models like Random Forest and CS-Forest outperformed simpler alternatives in all key evaluation metrics. These results align with Carmona et al. (2019), who documented the superior accuracy of complex ensemble models like XGBoost in predicting financial distress. Additionally, the ability of these models to handle high-dimensional data and capture intricate relationships explains their enhanced performance in bank failure prediction tasks.

## 6.7 Practical Implications

This section explores the practical significance of the study's findings for financial regulators, policymakers, banking institutions, and other stakeholders involved in financial stability and risk management. The accurate prediction of bank failures is not solely an academic exercise but a crucial tool for safeguarding the integrity of the financial system. Early identification of at-risk banks enables timely interventions that can prevent systemic crises, protect depositors, and maintain public confidence in the banking sector.

### 6.7.1 Implications for Regulatory Authorities and Supervisors

One of the key implications of this study is the critical role advanced machine learning models can play in regulatory oversight and early warning systems. Regulators are charged with monitoring the financial health of banks to preempt systemic risks. The findings indicate that ensemble models, particularly Random Forest and CS-Forest, offer substantial advantages in predicting bank failures with high accuracy and sensitivity.

- **Random Forest** is particularly suitable for supervisory purposes because it combines high predictive performance with greater interpretability compared to more complex models like CS-Forest. The model provides clear insights into feature importance, helping regulators understand which financial ratios (e.g., capital adequacy, liquidity ratios) are driving risk assessments.
- **CS-Forest**, while less interpretable, offers superior sensitivity to identifying distressed banks (i.e., minimizing Type II errors). In a regulatory context, this model could be

Public

employed in parallel with more interpretable models or integrated into a two-tiered system where high-risk institutions flagged by CS-Forest undergo further scrutiny with transparent models like Random Forest or Logistic Regression.

Incorporating such models into supervisory frameworks aligns with the "macroprudential supervision" approach advocated in the post-2008 regulatory reforms, aiming to mitigate systemic risks before they escalate.

### 6.7.2 Implications for Bank Management and Risk Officers

For bank managers and risk officers, machine learning models offer advanced tools for internal risk assessment. Accurate prediction models allow management to proactively address financial vulnerabilities, improve capital adequacy, and manage liquidity more effectively.

- **Use of Feature Importance Analysis**: Random Forest and CS-Forest identified several key financial ratios—such as Net Interest Margin (NIM), Efficiency Ratios, Return on Equity (ROE), and Loan Loss Allowances—as significant predictors of bank distress. These insights enable management to monitor and optimize these variables proactively.
- **Strategic Planning and Stress Testing**: CS-Forest’s sensitivity in detecting failing institutions makes it a valuable component of internal stress testing frameworks. Banks could simulate various economic scenarios and assess their resilience based on model predictions.

### 6.7.3 Implications for Policymakers and Financial Stability Committees

Policymakers concerned with macro-financial stability can utilize these predictive models as decision-support tools for regulatory interventions, resolution planning, and policy design.

- **Minimizing False Negatives**: As demonstrated by CS-Forest, models that minimize Type II errors can help ensure that risky institutions do not escape early regulatory attention. Policymakers can incorporate such tools in frameworks that identify systemically important financial institutions (SIFIs) and determine capital buffers.
- **Balancing Complexity and Interpretability**: While complex models offer higher predictive power, their "black box" nature remains a concern for policy frameworks emphasizing transparency and accountability. A balance must be struck by adopting models like Random Forest in high-stakes decisions and employing CS-Forest as a supplementary layer for early warnings.

### 6.7.4 Implications for Future Regulatory Technology (RegTech)

The increasing adoption of RegTech solutions presents an opportunity to integrate advanced machine learning into routine supervision and compliance monitoring. This study demonstrates the feasibility of embedding Random Forest and CS-Forest models into regulatory data pipelines to support automated monitoring systems.

- **Scalability and Automation**: The scalability of Random Forest models makes them suitable for real-time monitoring of large datasets. Regulatory agencies can automate bank health assessments and flag emerging risks without requiring extensive human intervention.

Public

- **Interoperability with Existing Systems**: These models can complement existing risk assessment frameworks (e.g., CAMELS ratings) by providing an additional, data-driven layer of oversight.

### 6.8 Conclusion

This chapter presented a comprehensive analysis of machine learning models applied to the prediction of bank failures. The findings underscore the value of advanced ensemble methods in addressing the complex challenges inherent in financial distress prediction, particularly within imbalanced datasets where the occurrence of failed banks represents a minority class.

The analysis began with an overview of model performance (Section 6.2), revealing that ensemble techniques, notably Random Forest, Regularized Random Forest (RRF), and Cost-Sensitive Forest (CS-Forest), consistently outperformed simpler models such as Logistic Regression and Naive Bayes. These ensemble models demonstrated superior predictive accuracy, sensitivity (Recall), and discriminatory power (ROC Area), confirming their suitability for financial risk modeling.

Section 6.3 offered an interpretive evaluation of key performance metrics, emphasizing their implications for model reliability in the context of bank failure prediction. Specifically, CS-Forest achieved the highest Recall, reducing Type II errors—a critical consideration where false negatives can lead to severe systemic risks. Random Forest balanced accuracy with interpretability, providing actionable insights for regulatory and managerial decision-making.

The Confusion Matrix analysis (Section 6.4) reinforced the critical trade-offs between Type I and Type II errors. CS-Forest's superior capability to minimize false negatives was highlighted, making it a strong candidate for early warning systems where the primary concern is failing to detect troubled banks. In contrast, Random Forest offered lower false positives, thus mitigating the risk of unnecessary interventions.

The Comparative Analysis in Section 6.5 synthesized these insights, evaluating the strengths, weaknesses, and practical applications of CS-Forest, Random Forest, and RRF. CS-Forest's complexity and sensitivity make it suitable for high-stakes distress prediction scenarios, while Random Forest's interpretability ensures its practical utility in supervisory and regulatory settings.

Hypothesis testing in Section 6.6 statistically validated the study's core findings. Each hypothesis was assessed and rejected or accepted based on empirical results, providing robust support for the research objectives.

The Practical Implications outlined in Section 6.7 translated the findings into actionable insights for regulators, bank managers, and policymakers. The results advocate for the integration of machine learning models into regulatory frameworks and internal risk management systems, highlighting the potential of these tools to enhance the early detection and mitigation of financial distress.

Public

# Chapter 7: Conclusion

## 7.1 Introduction

This chapter brings the research journey to its final stage, offering a comprehensive synthesis of the study's findings and their implications. The purpose of this conclusion is to reflect on the research questions and objectives established at the outset, demonstrate how they have been addressed through the study, and articulate the contributions made to the field of bank failure prediction and financial risk modelling.

The chapter begins by directly answering the primary research questions and demonstrating how the aims and objectives outlined in Chapter 1 have been fulfilled. It then moves to an evaluation of the theoretical and practical significance of the findings, particularly in advancing the application of machine learning methods, especially ensemble models such as Random Forest, CS-Forest, and Regularized Random Forest (RRF)in the predictive analytics of financial distress within the banking sector.

Furthermore, the chapter highlights the original contributions of the research, both in methodological innovation and empirical insight. These contributions are contextualized within the broader body of academic literature reviewed in Chapter 2, illustrating how this study fills critical gaps and extends existing knowledge in financial distress prediction.

The chapter also offers an honest appraisal of the study's limitations, acknowledging the scope and constraints encountered during the research process. Finally, it identifies promising directions for future research that can build upon this study's foundation and continue to enhance the predictive accuracy, interpretability, and practical applicability of machine learning techniques in financial risk management.

## 7.2 Revisiting the Research Aims and Objectives

The central aim of this research was to develop a robust and accurate predictive framework for identifying bank failures, leveraging advanced machine learning techniques within a financial risk management context. Specifically, the study sought to evaluate and compare the performance of various machine learning models, with a particular focus on ensemble methods, to determine their effectiveness in handling the complex, imbalanced datasets characteristic of bank failure prediction.

The research objectives outlined in Chapter 1 were as follows:

1. To review and synthesize the existing literature on bank failure prediction models, with an emphasis on machine learning techniques.
2. To construct a comprehensive dataset of U.S. commercial banks, incorporating key financial ratios aligned with CAMELS indicators, during a period of heightened financial stress (2007–2013).
3. To implement and compare multiple machine learning algorithms—including Random Forest, CS-Forest, Regularized Random Forest (RRF), and others—in terms of their predictive performance.
4. To assess the trade-offs between model complexity and interpretability, especially in regulatory contexts where explainability is critical.

Public

5. To evaluate the capacity of the selected models to minimize Type I and Type II errors, with particular emphasis on reducing false negatives (Type II errors), which represent missed predictions of bank failure.
6. To provide recommendations for regulators and policymakers on the practical implementation of machine learning techniques in early warning systems for banking crises.

### Achievement of Research Aims and Objectives

The research comprehensively addressed these aims and objectives:

- **Objective 1** was fulfilled through an extensive literature review, offering a critical synthesis of traditional statistical approaches (such as logistic regression and discriminant analysis) and advanced machine learning techniques (including SVM, XGBoost, and ensemble models). This review identified key gaps in literature, particularly the need for models that effectively manage data imbalance and prioritize Type II error minimization.

- **Objective 2** was achieved by constructing a robust dataset using FDIC records covering 2007–2013, incorporating a balanced representation of failed and non-failed banks. The dataset included critical financial ratios reflecting capital adequacy, asset quality, management capability, earnings strength, liquidity, and sensitivity to market risks.

- **Objective 3** was realized through the implementation of a comprehensive modelling strategy. Thirteen machine learning algorithms were tested and compared. The findings clearly demonstrated the superiority of ensemble methods—particularly CS-Forest, Random Forest, and RRF—in achieving higher classification accuracy, recall, ROC-AUC scores, and Kappa statistics.

- **Objective 4** was addressed by providing a detailed analysis of each model's complexity and interpretability. While Random Forest and CS-Forest demonstrated superior predictive power, they also presented challenges regarding model transparency, which were critically examined.

- **Objective 5** was central to the comparative performance analysis and confusion matrix evaluations. CS-Forest, in particular, was shown to minimize Type II errors most effectively, correctly classifying a greater number of failed banks than competing models. This focus on false negatives is crucial, given the catastrophic consequences of missed bank failure predictions.

- **Objective 6** was fulfilled by translating the empirical findings into practical recommendations. The study emphasized the importance of integrating machine learning into regulatory frameworks, encouraging the adoption of models that balance predictive accuracy with interpretability and offering insights into early warning system design.

Public

### Alignment with Research Questions

The research questions posed at the outset were systematically answered:

- **Which machine learning models are most effective in predicting bank failures in an imbalanced dataset?**

  ➔ CS-Forest, Random Forest, and RRF emerged as the top-performing models, with CS-Forest demonstrating the best recall and Type II error minimization.

- **How do these models balance the trade-off between predictive power and interpretability?**

  ➔ Random Forest offered an optimal balance, while CS-Forest provided superior recall at the cost of interpretability. RRF struck the middle ground between these extremes.

- **What are the implications for regulatory early warning systems?**

  ➔ Ensemble models, particularly CS-Forest, offer considerable potential for enhancing early warning systems by reliably identifying at-risk banks, thereby enabling timely intervention by regulators.

## 7.3 Key Findings and Contributions to Knowledge

This section synthesizes the principal findings of the research and articulates the contribution it makes to both academic literature and practical applications in financial risk management and regulatory oversight.

### Key Empirical Findings

1. **Ensemble Machine Learning Models Outperform Traditional Approaches** The study confirmed that ensemble methods—specifically CS-Forest, Random Forest, and Regularized Random Forest—outperform traditional statistical models such as Logistic Regression and Naive Bayes in predicting bank failures. These advanced models demonstrated superior performance in all key evaluation metrics, including Correctly Classified Instances (CCI), Recall, ROC-AUC, and Kappa statistics. Notably, Random Forest achieved the highest ROC-AUC (0.991), while CS-Forest achieved the highest recall (0.897), demonstrating its ability to minimize Type II errors.

2. **Cost-Sensitive Learning Enhances Recall and Reduces Type II Errors** The CS-Forest model demonstrated a remarkable ability to minimize Type II errors, correctly identifying 401 out of 447 failed banks in the test set. This performance underscores the value of cost-sensitive learning in addressing the imbalanced nature of bank failure datasets. Given the asymmetrical costs associated with false negatives in financial crisis prediction, CS-Forest's design aligns well with the priorities of financial regulators.

3. **Trade-Off Between Model Complexity and Interpretability** While Random Forest and CS-Forest models exhibited superior predictive capabilities, they present challenges regarding interpretability.

Public

4. This trade-off is particularly relevant for regulatory applications, where explainability is often a legal and operational requirement. Regularized Random Forest (RRF) offered a balanced solution, maintaining high predictive power while improving transparency and reducing the risk of overfitting.

5. **Financial Ratios with the Greatest Predictive Power** Feature importance analyses identified several key financial ratios as critical predictors of bank failure. Among these, Net Interest Margin (NIM), Return on Equity (ROE), Loss Allowance to Loans (Lnatresr), and Risk-Based Capital Ratio (Rbcrwaj) were consistently ranked as the most influential factors across the top-performing models. This insight reaffirms the relevance of CAMELS indicators while highlighting the added value of integrating advanced machine learning techniques in feature selection and weighting.

6. **The Superiority of Hybrid and Ensemble Approaches in Imbalanced Datasets** The findings corroborate prior research emphasizing the superiority of ensemble and hybrid models in handling class imbalance (Ekinci & Erdal, 2017). The high recall and Kappa statistics achieved by CS-Forest and RRF provide empirical support for the use of these approaches in bank failure prediction tasks where data imbalance is a key concern.

### Theoretical Contributions

1. **Advancement in the Application of Machine Learning to Financial Risk Prediction** This study extends the application of machine learning techniques in financial risk management by demonstrating the efficacy of ensemble methods in predicting rare events like bank failures. It highlights the importance of algorithm design in improving recall and reducing the likelihood of Type II errors.

2. **Integration of Cost-Sensitive Learning into Bank Failure Prediction Models** By integrating cost-sensitive learning into the Random Forest framework, the CS-Forest model advances the theoretical understanding of how machine learning can be adapted to address the specific needs of financial prediction, particularly in minimizing the costs associated with false negatives.

3. **Empirical Validation of CAMELS Indicators in a Machine Learning Framework** While CAMELS indicators have been widely used in traditional bank failure prediction models, this study validates their continued relevance when applied within sophisticated machine learning algorithms. The feature importance analysis provides evidence of the enduring predictive power of these indicators when combined with advanced analytics.

### Practical Contributions

1. **Enhancing Regulatory Early Warning Systems** The research provides practical guidance for regulatory authorities and financial institutions in developing and implementing early warning systems. The superior performance of CS-Forest and Random Forest models in accurately predicting bank failures makes them viable candidates for operational deployment in supervisory frameworks.

Public

2. **Minimizing Systemic Risk Through Improved Prediction of Bank Failures** By reducing the incidence of missed failure predictions (false negatives), the use of CS-Forest in early warning systems can facilitate timely regulatory interventions. This reduces the potential for systemic risk, contributing to the stability of the financial system.

3. **Balancing Predictive Performance with Model Interpretability** The trade-off analysis between model complexity and interpretability provides actionable insights for regulators who must balance the need for accurate predictions with the requirement for transparent decision-making tools. Regularized Random Forest offers a viable compromise for institutions seeking to maintain both predictive power and explainability.

## 7.4 Implications for Policy and Practice

This section discusses the practical relevance of the study's findings for regulatory authorities, financial institutions, and policymakers. It focuses on how the advanced machine learning models, particularly CS-Forest, Random Forest, and Regularized Random Forest (RRF), can be integrated into current supervisory and risk management frameworks to enhance the resilience of the banking sector.

### 7.4.1 Strengthening Early Warning Systems (EWS)

The research findings indicate that ensemble machine learning models significantly outperform traditional statistical methods in predicting bank failures. This suggests that regulatory authorities, such as central banks and supervisory bodies, should consider integrating these advanced models into their Early Warning Systems (EWS). The ability of CS-Forest to minimize Type II errors—thereby reducing the risk of failing to detect at-risk banks—makes it a valuable tool for preemptive regulatory action.

By deploying such models in supervisory frameworks, regulators can improve their ability to detect vulnerable institutions before distress manifests as a systemic crisis. For example, CS-Forest's high recall rate (0.897) allows regulators to identify failing banks with a higher degree of certainty, providing critical lead time for intervention strategies.

### 7.4.2 Addressing the Challenge of Class Imbalance in Regulatory Surveillance

One of the key contributions of this research lies in its handling of imbalanced datasets, a common problem in bank failure prediction where failed banks represent a small minority of cases. Regulators often rely on metrics like overall accuracy or Correctly Classified Instances, which can be misleading in such contexts.

This study demonstrates the need for a shift in emphasis towards metrics that prioritize recall and Kappa statistics, which better capture the model's effectiveness in identifying failures. Regulators can adopt cost-sensitive learning frameworks, such as CS-Forest, to ensure that the minority class (failed banks) receives appropriate weighting in predictive models. This can significantly reduce the probability of regulatory blind spots during financial instability.

Public

### 7.4.3 Enhancing Supervisory Stress Testing Frameworks

Supervisory stress testing has become a cornerstone of modern regulatory practice, particularly in the aftermath of the 2007–2008 financial crisis. However, traditional stress testing methods often rely on linear models and assumptions that may not capture the complex, nonlinear relationships inherent in financial distress events.

The findings of this research suggest that incorporating machine learning-based models into stress testing frameworks could enhance predictive accuracy. For instance, Random Forest and CS-Forest models can better accommodate nonlinearities and interactions among risk factors, such as capital adequacy, liquidity ratios, and asset quality, offering a more holistic view of bank resilience under stress scenarios.

### 7.4.4 Balancing Predictive Power with Model Interpretability

While advanced machine learning models, such as Random Forest and CS-Forest, offer superior predictive power, their complexity can impede interpretability—a key requirement for regulatory transparency and accountability. Regulators face a trade-off: the need for accurate predictions versus the need for transparent, explainable models that can withstand public and legal scrutiny.

This study highlights Regularized Random Forest as a potential middle-ground solution. RRF maintains high predictive performance while offering better interpretability by reducing the complexity of model output. Policymakers can leverage RRF for scenarios where decision transparency is as critical as predictive accuracy, such as regulatory reporting or policy development.

### 7.4.5 Implications for Risk Management Practices in Financial Institutions

The application of machine learning models extends beyond regulatory frameworks to internal risk management within financial institutions. Banks can incorporate these models into their internal credit risk assessment processes and portfolio monitoring systems. For example:

- CS-Forest can be utilized to flag at-risk counterparties in loan portfolios, facilitating early intervention strategies such as restructuring or provisioning.
- Random Forest models can be used in conjunction with existing credit scoring systems to enhance decision-making, particularly for high-value or high-risk exposures.

Financial institutions can also benefit from feature importance insights provided by these models. By understanding which financial indicators—such as Net Interest Margin (NIM), Return on Equity (ROE), and Loss Allowance to Loans—most strongly predict failure, risk managers can prioritize monitoring and resource allocation more effectively.

### 7.4.6 Contribution to Macroprudential Supervision and Systemic Risk Mitigation

Machine learning-based models, particularly those adept at identifying early signals of distress, can play a pivotal role in macroprudential supervision. By aggregating risk assessments across the banking sector, regulators can identify systemic vulnerabilities and implement macroprudential tools, such as countercyclical capital buffers, to mitigate systemic risk.

Public

The findings of this study emphasize the need for data-driven policy interventions. For example, the superior recall and ROC-AUC scores achieved by CS-Forest and Random Forest models support their use in identifying systemically important institutions that may require enhanced supervision or resolution planning.

## 7.5 Limitations of the Study

While this study offers significant contributions to the field of bank failure prediction, it is not without limitations. Acknowledging these limitations is crucial, as it contextualizes the findings, highlights areas for cautious interpretation, and provides direction for future research. The limitations span across data, methodological choices, model interpretability, and generalizability.

### 7.5.1 Data Limitations

This study relies on financial statement data from the Federal Deposit Insurance Corporation (FDIC), covering the period of 2007 to 2013, a period that encompasses the global financial crisis and its aftermath. Although this timeframe is particularly relevant for understanding bank failures during times of systemic stress, it raises concerns about the temporal generalizability of the findings. Economic conditions, regulatory frameworks, and banking practices evolve over time, and models trained on historical data may not fully capture the complexities of more recent or future crises.

Additionally, the dataset focuses exclusively on U.S. commercial banks. While this provides a consistent regulatory and institutional context, it limits the applicability of the findings to other jurisdictions with different banking systems, regulatory environments, and macroeconomic conditions. Future studies should consider cross-country datasets to enhance generalizability and account for institutional heterogeneity.

### 7.5.2 Imbalanced Dataset Challenges

Despite employing advanced techniques such as CS-Forest and Regularized Random Forest to handle the imbalanced nature of bank failure data, the study is still subject to the inherent limitations posed by skewed class distributions. In real-world scenarios, failed banks represent a minority class, leading to challenges in ensuring the robustness and stability of classification outcomes.

While cost-sensitive learning strategies have mitigated some of these issues, they may not eliminate all biases introduced by the imbalance. Specifically, the precision-recall trade-off remains an area of concern. Although CS-Forest achieved high recall, the increase in false positives (Type I errors) highlights the difficulty in maintaining a balance between sensitivity and specificity. The operational costs of false positives—such as unwarranted regulatory interventions—must be considered when interpreting model outputs.

### 7.5.3 Model Complexity and Interpretability

A key trade-off in the application of machine learning models is between predictive power and interpretability. While Random Forest, CS-Forest, and Regularized Random Forest demonstrated superior performance in classification tasks, they operate as "black box" models

Public

to varying degrees. This lack of transparency can hinder their adoption in regulatory environments that demand explainability and accountability.

Although feature importance measures were extracted and analyzed to provide insights into decision-making processes, these methods offer limited explanations compared to more interpretable models like logistic regression. For regulators and policymakers, the inability to fully understand how predictions are made can be a barrier to implementation, particularly when justifying decisions to external stakeholders or in legal proceedings.

### 7.5.4 Limitations in Feature Selection and Engineering

This study employed CAMELS-based financial ratios as the primary features for model training and prediction. While these indicators are well-established and widely used in banking supervision, they may not capture all dimensions of bank risk. For example, qualitative factors such as management competence, corporate governance structures, and macroeconomic indicators like monetary policy stance or geopolitical risk were not included in the model.

Furthermore, although feature selection was informed by domain expertise and statistical relevance, there remains the possibility that more advanced feature engineering techniques (e.g., dimensionality reduction, unsupervised learning for anomaly detection) could further enhance model performance. The reliance on standard financial ratios may limit the ability of models to capture complex, latent risk factors.

### 7.5.5 Scope and Generalizability of Hypothesis Testing

The hypothesis testing in this study provides statistical validation of key findings, including the superior performance of ensemble models and the importance of minimizing Type II errors in bank failure prediction. However, the scope of hypotheses was confined to performance metrics derived from machine learning models and did not extend to broader causal inferences regarding the drivers of bank failure.

Moreover, the conclusions drawn from hypothesis testing are applicable within the context of the dataset and methodological framework employed. The robustness of these findings across different economic periods, regulatory regimes, or banking systems remains to be verified. Future studies should expand hypothesis testing to incorporate longitudinal analyses and cross-jurisdictional comparisons.

## 7.6 Directions for Future Research

While this study has made significant contributions to the field of bank failure prediction by leveraging advanced machine learning models and exploring their practical applications, several avenues remain open for further exploration. Future research can build upon these findings to enhance model performance, address identified limitations, and expand the practical utility of predictive systems in banking supervision and risk management.

### 7.6.1 Expansion to Cross-Country and Cross-Regional Studies

The current research focuses exclusively on U.S. commercial banks, leveraging a uniform regulatory and economic environment to ensure data consistency. However, banking systems vary significantly across countries in terms of regulation, market structures, and systemic risk factors. Future research should extend the scope of predictive models to include cross-country datasets.

By incorporating data from multiple jurisdictions, researchers can explore how differences in regulatory frameworks, capital adequacy standards, and supervisory practices influence the predictive power of machine learning models. Moreover, cross-regional studies can offer insights into systemic risk contagion and the role of international financial linkages in amplifying bank distress. This comparative approach would enhance the generalizability and robustness of bank failure prediction models on a global scale.

### 7.6.2 Integration of Macroeconomic and Market-Based Indicators

This study primarily relies on CAMELS-based financial ratios, which, although critical, capture only firm-specific risk factors. Future research should integrate macroeconomic variables, such as GDP growth, unemployment rates, inflation, and monetary policy indicators, to account for the broader economic environment influencing bank stability.

Moreover, the inclusion of market-based indicators, such as stock prices, bond spreads, credit default swap (CDS) spreads, and market volatility measures, can provide real-time insights into the financial health of banks. Market data can enhance the timeliness of predictions and offer early warning signals that are not readily captured by traditional accounting data. The integration of these variables may improve the sensitivity of models to sudden changes in economic conditions.

### 7.6.3 Real-Time Predictive Systems and Early Warning Frameworks

While this study focuses on ex-post analysis of bank failure data, future research should aim to develop real-time predictive systems that can function as early warning frameworks for regulators and policymakers. These systems could incorporate dynamic data streams, including both financial ratios and real-time market data, to offer continuous risk assessments.

Developing early warning frameworks necessitates not only advanced predictive algorithms but also robust data infrastructure capable of real-time data collection and processing. Future research should explore the design and implementation of such systems within central banks, supervisory authorities, and financial institutions. These tools could enable proactive interventions, thereby reducing the risk of contagion and systemic collapse.

### 7.6.4 Addressing Model Interpretability and Explainability

As demonstrated in this study, ensemble learning models like Random Forest, CS-Forest, and Regularized Random Forest offer high predictive accuracy but often suffer from limited interpretability. Future research should focus on enhancing the explainability of complex machine learning models, particularly in the context of regulatory decision-making.

Emerging techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), and rule extraction algorithms offer promising avenues for making black-box models more transparent. Applying these techniques to bank failure prediction models would allow regulators to better understand and trust model outputs, facilitating their adoption in policy settings.

Public

Additionally, research could explore the development of hybrid models that combine the predictive power of complex algorithms with the transparency of traditional statistical methods, thereby offering the best of both worlds.

### 7.6.5 Incorporating Qualitative Data and Unstructured Information

Traditional bank failure prediction models rely heavily on structured quantitative data. However, qualitative factors—such as corporate governance quality, management competence, regulatory compliance, and reputational risk—are also critical determinants of bank stability. Future research should investigate methodologies for incorporating qualitative data into predictive models.

Text mining and natural language processing (NLP) techniques could be used to analyze unstructured data sources, including regulatory filings, news reports, social media, and management discussions in annual reports. By extracting sentiment and key risk indicators from these data sources, models could achieve a more holistic assessment of bank risk profiles.

### 7.6.6 Longitudinal Analysis and Adaptability Over Time

Machine learning models are often trained on historical datasets, raising concerns about their ability to adapt to future financial environments. Future research should investigate model adaptability by conducting longitudinal analyses that test the robustness of predictive algorithms over extended periods and different economic cycles.

One promising area of research is the application of transfer learning and online learning algorithms, which allow models to update themselves incrementally as new data becomes available. Such approaches could ensure that predictive systems remain relevant and effective in the face of evolving financial landscapes.

### 7.6.7 Integration with Stress Testing and Scenario Analysis

Predictive models of bank failure can complement traditional stress testing frameworks by identifying institutions most vulnerable to adverse scenarios. Future studies should explore the integration of machine learning-based bank failure prediction systems with macroprudential stress testing models.Such integration could enhance the ability of regulators to assess the systemic risk posed by banks under different stress scenarios, facilitating more informed policy interventions. Additionally, scenario analysis could help validate the robustness of machine learning models by exposing them to extreme but plausible adverse conditions.

### 7.6.8 Closing Remarks

In conclusion, this study has demonstrated the potential of advanced machine learning techniques, particularly ensemble and cost-sensitive models, in predicting bank failures. By systematically evaluating model performance, addressing the challenge of imbalanced datasets, and providing insights into feature importance, the research contributes to the development of more robust and reliable early warning systems. As the global financial landscape continues to evolve, the need for sophisticated, adaptive, and transparent predictive systems will only become more critical. This thesis provides a foundation upon which future innovations in bank failure prediction and financial risk management can be built.

Public

# Chapter 8: References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine Learning, 6(1), 37-66.

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and bagging ensemble methods. Decision Support Systems, 45(1), 110-122.

Antunes, A., Bonfim, D., Monteiro, N., & Rodrigues, P. M. M. (2018). Forecasting banking crises with dynamic panel probit models. *International Journal of Forecasting*, 34(2), 249-275.

Ashcraft, A., & Schuermann, T. (2008). Understanding the securitization of subprime mortgage credit. *Foundations and Trends in Finance*, 2(3), 191-309.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science, 49(3), 312-329.

Bellman, R. (1957). Dynamic Programming. Princeton University Press.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Neural Networks: Tricks of the Trade (pp. 437-478). Springer.

Berger, A. N., & Bouwman, C. H. S. (2012). How does capital affect bank performance during financial crises? *Journal of Financial Economics*, 109(1), 146-176.

Bernanke, B. S. (2010). *The Courage to Act: A Memoir of a Crisis and Its Aftermath*. W. W. Norton & Company.

Beutel, J., List, S., von Schweinitz, G., 2019. "An evaluation of early warning models for systemic banking crises: Does machine learning improve predictions? * IWH Discussion Papers 2/2019, Halle Institute for Economic Research (IWH). *

Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems. 50. 602-613. 10.1016/j.dss.2010.08.008.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25. DOI: 10.1007/s11749-016-0481-7.

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford University Press.

Boyacıoglu, M. A., Kara, Y., & Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of Savings Deposit Insurance Fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications*, 36(2), 3355-3366.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Breiman, L. Bagging predictors. Mach Learn 24, 123–140 (1996). https://doi.org/10.1007/BF00058655

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. The Annals of Statistics, 30(4), 927-961.

Public

Bühlmann, P., Hothorn, T. (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting," Statistical Science, Statist. Sci. 22(4), 477-505, (November 2007)

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

Carmona, P., Diranzo, F. J. C., & Momparler, A. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

Chen, C. & Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. University of California, Berkeley.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chernobai, A. S., Rachev, S. T., & Fabozzi, F. J. (2007). Operational risk: A guide to Basel II capital requirements, models, and analysis. John Wiley & Sons.

Chiaramonte, L., Liu, H., Poli, F., & Zhou, M. (2016). How accurately can z-score predict bank failure? *Financial Markets, Institutions & Instruments*, 25(5), 333–360.

Cleary, J. G., & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. Proceedings of the 12th International Conference on Machine Learning, 108-114.

Cleary, S., & Hebb, G. (2015). An assessment of machine learning algorithms for the prediction of credit default. Econometrics, 3(4), 589-613.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

Cole, R.A. & White, L.J. (2012). Déjà Vu All Over Again: The Causes of U.S. Commercial Bank Failures This Time Around. Journal of Financial Services Research, 42(1-2), pp.5-29.

Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B, 20(2), 215-242.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784-3797.

Deng, H. (2013). Guided random forest in the RRF package.

Deng, H., & Runger, G. (2012). Feature selection via regularized trees. Proceedings of the 2012 International Joint Conference on Neural Networks, 1-8.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7, 3.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In International Workshop on Multiple Classifier Systems (pp. 1-15). Springer.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2), 103-130.

Public

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. Statistical Science, 11(2), 89-102.

Ekinci, A., & Erdal, H. I. (2017). Bank failure prediction by hybrid machine learning techniques. *Economics & Management*, 20(1), 29-42.

Ekinci, A., Sen, S. Forecasting Bank Failure in the U.S.: A Cost-Sensitive Approach. Comput Econ (2024). https://doi.org/10.1007/s10614-023-10537-6

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*.

Erdal, H. I., & Ekinci, C. (2013). A new approach to Turkish banking sector's failure prediction via artificial neural network and support vector machine. *Expert Systems with Applications*, 40(3), 887-897.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. IJCAI, 93, 1022-1027.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), 337-407.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29(2), 131-163.

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2018). Forecasting bank failures and stress testing: A machine learning approach. International Journal of Forecasting, 34(3), 440-455.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Gorton, G. (2009). *Slapped by the Invisible Hand: The Panic of 2007*. Oxford University Press.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—Not so stupid after all? International Statistical Review, 69(3), 385-398.

Hand, D.J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn 77, 103–123 (2009). https://doi.org/10.1007/s10994-009-5119-5

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. Statistical Science, 1(3), 297-318.

Hastie, T., & Tibshirani, R. (1990). Generalized Additive Models. Chapman and Hall.

Public

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

Haykin, S. (1998). Neural Networks: A Comprehensive Foundation. Prentice Hall.

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359-366.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. John Wiley & Sons.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B. (2010). Model-based Boosting 2.0. Journal of Machine Learning Research. 11. 2109-2113.

Jarrow, R.A. and Turnbull, S.M. (1995) Pricing Derivatives on Financial Securities Subject to Credit Risk. The Journal of Finance, 1, 53-85.http://dx.doi.org/10.1111/j.1540-6261.1995.tb05167.x

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11, 51.

Kleinbaum, D. G., Klein, M., & Pryor, E. R. (2002). Logistic Regression: A Self-Learning Text. Springer.

Kothari, V. (2008). *Securitization: The Financial Instrument of the Future*. Wiley.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160, 3-24.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. European Journal of Operational Research, 180(1), 1-28.

Kuncheva, L. I., & Rodriguez, J. J. (2007). An experimental study on rotation forest ensembles. In *Proceedings of the 5th International Conference on Multiple Classifier Systems*.

Laeven, L., & Valencia, F. (2018). Systemic banking crises revisited. *IMF Working Paper, WP/18/206*. International Monetary Fund.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

Le, L., & Viviani, J. L. (2017). An empirical analysis of bank failure prediction in the United States. *Journal of Financial Research*, 40(2), 257-282.

Lessmann, S., Baesens, B., Seow, H., Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research. (doi:10.1016/j.ejor.2015.05.030). 10.1016/j.ejor.2015.05.030.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.

Ling, C. X., & Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem. Encyclopedia of Machine Learning, 2011, 231-235.

Public

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57.

Lu, W., & Whidbee, D. A. (2013). Bank structure and failure during the financial crisis. *Journal of Financial Economic Policy*, 5(3), 281–299.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

Manthoulis, G., Mylonakis, J., & Diacogiannis, G. (2020). Predicting bank failures: A statistical and machine learning approach. *Journal of Financial Risk Management*, 9(4), 305-323.

Martens, D., & Baesens, B. (2010). Building acceptable classification models for financial decision support using robust meta-learning. Decision Support Systems, 48(3), 491–501.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. *Journal of Banking & Finance*, 1(3), 249-276.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models (2nd ed.). Chapman and Hall/CRC.

Meinshausen, N. (2007). Relaxed lasso. Computational Statistics & Data Analysis, 52(1), 374-393.

Menard, S. (2000). Applied Logistic Regression Analysis (2nd ed.). Sage.

Meyer, P. A., & Pifer, H. W. (1970). Prediction of bank failures. *The Journal of Finance*, 25(4), 853-868.

Mian, A., & Sufi, A. (2009). The consequences of mortgage credit expansion: Evidence from the U.S. mortgage default crisis. *The Quarterly Journal of Economics*, 124(4), 1449-1496.

Miguéis, V. L., Camanho, A. S., & Cardoso, F. J. (2018). Predicting bank failure using survival analysis. European Journal of Operational Research, 274(2), 647-661.

Minsky, M., & Papert, S. (1969). Perceptrons: An Introduction to Computational Geometry. MIT Press.

Mitchell, T. M. (1997), Machine learning , Vol. 1 , McGraw-hill New York .

Momparler, A., Carmona, P., & Redondo, Y. (2020). Fuzzy-set qualitative comparative analysis (fsQCA) in banking failure prediction. *Journal of Business Research*, 118, 253-263.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625-632).

Nielsen, D. (2016). Tree Boosting with XGBoost - Why does XGBoost win "Every" machine learning competition? NTNU Machine Learning Report.

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217-222.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The Journal of Educational Research, 96(1), 3-14.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, 36, 10.1016/j.ijforecast.2019.11.005.

Public

Phua, C., Lee, V., Smith-Miles, K., Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. CoRR. abs/1009.6119.

Plessis, E. (2022). Dynamic forecasting of banking crises with a Qual VAR. *Journal of Applied Economics*, 25(1), 477-503. DOI: 10.1080/15140326.2020.1816132.

Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3), 21-45.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*.

Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for AdaBoost. Machine Learning, 42(3), 287-320.

Reinhart, C. M., & Rogoff, K. S. (2009). *This Time is Different: Eight Centuries of Financial Folly*. Princeton University Press. ISBN: 978-0691142166.

Ribeiro, M., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 1135-1144. 10.1145/2939672.2939778.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

Saunders, M.N.K., Lewis, P. and Thornhill, A. (2019) Research Methods for Business Students. 8th Edition, Pearson, New York.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In Nonlinear Estimation and Classification (pp. 149-171). Springer.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Machine Learning, 37(3), 297-336.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(1), 185-197.

Shiller, R. J. (2008). *The Subprime Solution: How Today's Global Financial Crisis Happened, and What to Do about It*. Princeton University Press.

Sinkey, J. F. (1975). A multivariate statistical analysis of the characteristics of problem banks. *Journal of Finance*, 30(1), 21-36.

Skurichina, M., & Duin, R. P. W. (2002). Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications, 5(2), 121-135.

Public

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1), 77-89.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12), 3358-3378.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. Management Science, 38(7), 926-947.

Thomas, L.C., Edelman, D.B. and Crook, J.N., 2002. Credit scoring and its applications. Philadelphia: Society for Industrial and Applied Mathematics. (Mathematical Modeling and Computation, MM06). Available at: https://doi.org/10.1137/1.9780898718317

Thomson, J. B. (1991). Predicting bank failures in the 1980s. *Federal Reserve Bank of Cleveland, Economic Review*, Q I, 9-20.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.

Torna, G., & DeYoung, R. (2012). Nontraditional banking activities and bank failures during the financial crisis. *Journal of Financial Intermediation*, 21(3), 397-421.

Tsymbal, A. (2004). The problem of concept drift: Definitions and related work. Computer Science Department, Trinity College Dublin.

West, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27(11-12), 1131-1152.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Zadrozny, B., Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In Proceedings of the Third IEEE International Conference on Data Mining (pp. 435-442).

Zhang, H. (2004) The Optimality of Naive Bayes. Proceedings of 17th International Florida Artificial Intelligence Research Society Conference, Menlo Park, 12-14 May 2004, 562-567.

Zhang, Y., Wang, S., & Ji, G. (1999). A rule extraction method from neural networks based on the improved genetic algorithm. Expert Systems with Applications, 36(2), 1970-1976.

Zhou, Z. H., Chen, K., & Li, M. (2014). Ensemble methods. Foundations and Trends® in Machine Learning, 6(1), 1-34.

Zhou, Z.-H. (2012). Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC

Public