Tilburg University

School of Economics and Management

Machine Learning for Corporate Bankruptcy Prediction

Integrating Firm-Level Data and Macroeconomic Stress Testing

Master Thesis Finance

Author	Anna Csányi
SNR	2144100
ANR	927718
Supervisor:	Denis Kojevnikov
Second reader:	Gianpaolo Parise
Date:	23.06.2025

Table of Contents

List of Tables

1. Introduction

Corporate bankruptcy prediction plays a crucial role in enhancing financial stability, managing credit risk, and supporting early intervention strategies. The failure of large publicly traded firms can have wide-reaching effects, increasing unemployment, disrupting supply chains, and triggering broader market instability. This risk increases in particular during periods of economic downturn, when even fundamentally sound companies may experience distress due to tightened credit conditions or collapsing demand. Considering the potential systemic consequences of large numbers of corporate insolvencies, bankruptcy prediction models are developed by both academics and industry professionals. Effective bankruptcy prediction models are now used by credit rating agencies, lenders, and regulators as part of early-warning systems, aiming to identify at-risk firms well before default occurs. Many of these models are designed to predict distress two to three years in advance, enabling proactive risk mitigation and informed policy responses.

Historically, the academic literature on bankruptcy prediction has concentrated primarily on firm-specific characteristics, particularly accounting-based indicators such as profitability, leverage, and liquidity ratios. These variables have served as the foundation for traditional statistical models and, more recently, machine learning approaches. While methodological advancements have significantly improved prediction accuracy, most models continue to treat corporate failure as an isolated phenomenon that is driven exclusively by internal company metrics. In doing so, they often overlook the broader economic environment in which firms operate. Macroeconomic variables such as interest rates, GDP growth, inflation, and financial market volatility can meaningfully influence corporate survival, especially under stress. During financial crises, these external shocks can simultaneously affect thousands of firms, increasing the risk of widespread default and destabilising entire sectors. However, the influence of macroeconomic conditions remains underexplored in most mainstream bankruptcy prediction frameworks.

This thesis aims to bridge this gap by analysing the interaction between firm-level vulnerabilities and macroeconomic stress in predicting corporate bankruptcies. The question being considered in this study is whether macroeconomic factors and financial crises influence corporate bankruptcies in the United States. The study uses an imbalanced dataset of over 14,000 US-incorporated firms from 1990 to 2023 retrieved from Compustat and CRSP using

WRDS, and Federal Reserve Economic Data. By integrating firm-level financial data with industry-level and macroeconomic indicators and applying a range of machine learning techniques alongside a stress-testing framework, the thesis investigates how external shocks interact with firm vulnerabilities to affect bankruptcy risk. Four modelling techniques—logistic regression, random forest, XGBoost, and a neural network—are applied across three stages: first using only firm-level features, then adding industry-level data, and finally incorporating macroeconomic variables. The analysis further includes a macroeconomic stress-testing framework that models firm-level risk under simulated crisis conditions based on three past crises.

The results show that machine learning models outperform traditional logistic regression in most settings, with XGBoost consistently achieving the highest predictive accuracy, particularly when only firm-level variables are used. Random forest performs especially well in limiting false positives, which is crucial in real-world applications. Remarkably, logistic regression demonstrates high recall and benefits significantly from the inclusion of macroeconomic data, while the neural network, although improving with macroeconomic inputs, still underperforms compared to the other machine learning models. The initial stress testing simulations reveal that bankruptcy risk increases most sharply under dotcom-like conditions, while remaining stable or even declining during 2008 and COVID-like crises, suggesting possible model limitations in capturing credit market or policy effects. However, the extended stress testing framework, which includes additional crisis-sensitive indicators, shows a substantial increase in predicted risk under the 2008 scenario as well, highlighting the importance of incorporating more specific macro-financial variables to improve scenario sensitivity.

These findings align with and extend previous work in the literature. Barboza, Kimura, and Altman (2017) extended the traditional modelling approach by comparing logistic regression with several machine learning algorithms, including random forests and support vector machines. Their results showed that machine learning models outperform traditional methods in terms of accuracy, particularly when additional short-term performance indicators are included. This thesis builds directly on their findings by incorporating both traditional financial ratios and dynamic growth metrics, and by evaluating four classification models: logistic regression, random forest, XGBoost, and neural network. Consistent with Barboza, Kimura, and Altman (2017), the results of the thesis show that tree-based models, especially XGBoost, offer superior predictive performance across most configurations, while logistic regression performs well in terms of recall but suffers from low precision. In contrast to these primarily firm-level approaches, Duffie, Saita, and Wang (2007) introduced a model that incorporates macroeconomic indicators such as interest rates and equity market returns into the estimation of corporate default probabilities. Their findings support the inclusion of macro-level variables for improving out-of-sample forecasts. In line with this, the present thesis integrates macroeconomic indicators into the predictive framework and extends it by using machine learning techniques and stress-testing scenarios along with traditional predictive models. While the inclusion of macro variables did not universally improve performance across all models, the stress testing analysis revealed how external shocks can significantly elevate predicted bankruptcy risk, highlighting the importance of economic context in firm-level forecasting.

These findings suggest that while firm-level indicators remain the strongest predictors of bankruptcy, incorporating macroeconomic variables can enhance model performance and enable forward-looking risk assessment under adverse scenarios. This thesis contributes to the literature by demonstrating the value of combining firm-specific and macro-level data within modern machine learning frameworks, and by illustrating how macroeconomic stress testing can provide practical insight into systemic financial risk. Although the analysis is limited to US-incorporated firms due to data availability, the modelling framework may be adapted to other countries or regions in future research.

The remainder of the thesis is structured as follows. Section 2 presents a review of the existing literature, outlining the development of bankruptcy prediction models and identifying gaps addressed in this study. Section 3 introduces the research hypotheses, followed by Section 4, which details the data sources, variables, and descriptive statistics used in the analysis. Section 5 explains the methodological framework, including the theoretical background for the selected statistical and machine learning models. Section 6 reports the results across all modelling configurations and stress-testing scenarios. Section 7 concludes.

2. Literature review

2.1. Traditional bankruptcy prediction

The groundwork for modern bankruptcy prediction was laid by Beaver (1966), who conducted one of the first empirical studies linking financial ratios to corporate failure. Using univariate analysis, Beaver demonstrated that certain accounting-based indicators—particularly cash flow to total debt and net income to total assets—could predict failure up to five years in advance. His work established financial ratios as early-warning signals, which Altman (1968) later expanded into a multivariate model through the development of the well-known Z-score. Altman examined thirty-three bankrupt and thirty-three non-bankrupt firms from the period between 1946 to 1965. In this study, five key financial ratios measuring liquidity, profitability, leverage, solvency, and activity were identified through multiple discriminant analysis as the most significant indicators for distinguishing between solvent and insolvent firms. Using these ratios, Altman constructed the Z-score formula, which demonstrates a high level of predictive accuracy in classifying firms as bankrupt or non-bankrupt. Altman's model represented a major change in the assessment of credit risk and laid the foundation for later research on bankruptcy prediction. Even though the Z-score was developed more than five decades ago, it remains a widely used benchmark in both academic and professional financial contexts.

Following these developments, Ohlson (1980) introduced the O-score model, a logistic regression-based framework for bankruptcy prediction designed to address some of the limitations of earlier models such as Altman's Z-score. Unlike discriminant analysis, which assumes multivariate normality and equal covariance matrices across groups, logistic regression offers greater flexibility without imposing such strict requirements. Ohlson built his model using a large, unpaired sample of US industrial firms, including both bankrupt and non-bankrupt companies, thus better reflecting the actual distribution of firms in the economy. He selected nine explanatory variables, incorporating financial ratios (such as leverage, liquidity, and profitability), firm size, and binary indicators for events like negative equity. A key innovation of the O-score model is its probabilistic nature: instead of assigning firms to binary categories, it outputs a probability of bankruptcy, allowing for more nuanced, risk-sensitive decision-making. Additionally, the model does not require assumptions about prior bankruptcy probabilities and can dynamically adapt to new information through re-estimation with updated data.

While early studies like those of Beaver (1966), Altman (1968), and Ohlson (1980) demonstrated the predictive power of firm-level accounting ratios, these models were inherently static and relied solely on historical financial statements. As markets and financial systems evolved, researchers increasingly recognised the limitations of ignoring market signals, leading to the development of models that incorporated forward-looking, market-based variables. Shumway (2001) proposed a hazard model for bankruptcy prediction that combines market-based variables (e.g., firm return and volatility) with accounting data. His work represents a shift from static models to dynamic ones, where information is updated continuously, mirroring real-world market conditions. Shumway's empirical findings reveal that models using both accounting and market data outperform those relying solely on financial statements.

By comparing traditional accounting-based models, such as the Z-score and O-score, with a market-based probability of default derived from Merton's structural model, Hillegeist et al. (2004) further advanced the field of bankruptcy prediction. Their findings suggest that market data offers a forward-looking view of financial distress that can outperform purely historical financial ratios. While this thesis focuses on firm-level accounting data and macroeconomic indicators, the paper provides valuable insight into alternative modelling approaches and underscores the potential of integrating market-based signals in future research.

Building on these studies, Chava and Jarrow (2004) further investigated how incorporating stock market variables can enhance bankruptcy prediction models. Their research demonstrates that incorporating market-based variables such as stock returns and volatility significantly improves the accuracy of bankruptcy prediction models and reduces the marginal contribution of traditional accounting-based indicators. However, in many practical applications, especially where market data is limited, delayed, or noisy, accounting information remains a critical input. Credit rating agencies, banks, and regulators continue to rely on financial ratios derived from firm reports to assess credit risk, particularly for private firms or those with low trading volumes. Moreover, while Chava and Jarrow's study focused on traditional econometric approaches, this thesis explores the performance of accounting-based features within non-linear machine learning models, where complex patterns and interactions can be captured more effectively. This allows for a re-evaluation of the predictive value of accounting data in modern, data-driven frameworks.

As the field evolved, machine learning and data mining techniques gained increasing attention in bankruptcy prediction research. Sun et al. (2014) provide a structured review of these developments, categorising models based on the definition of financial distress, approaches to modelling financial distress prediction, sampling methods, and common techniques used for forecasting distress. Their review highlights the growing importance of non-linear models and ensemble methods such as boosting, as well as the practical challenges associated with applying machine learning techniques to imbalanced datasets.

Expanding on this shift toward non-linear modelling techniques, Yeh, Chi, and Hsu (2014) examined the application of neural networks and support vector machines in bankruptcy prediction. They observed that these non-linear models significantly improved classification accuracy when trained on diverse datasets. The authors stress the importance of feature selection and data preprocessing in building effective predictive models, a consideration echoed in many modern machine learning workflows.

While much early machine learning research focused on firm-specific bankruptcy prediction, Malekipirbazari and Aksakalli (2015) applied machine learning models to predict borrower status in social lending. Their study incorporates a cost-sensitive analysis to account for the higher cost of misclassifying bad borrowers, and highlights that while random forests perform particularly well at identifying top-quality borrowers under low acceptance rates, their relative advantage decreases as acceptance rates rise. The findings underline the importance of both algorithmic robustness and practical considerations such as cost asymmetry and model transparency in financial prediction tasks. Although Malekipirbazari and Aksakalli (2015) focus on borrower classification in a social lending setting rather than on corporate bankruptcy prediction, their study remains relevant for this thesis by illustrating how random forest models can effectively handle financial risk prediction tasks. Their findings support the use of ensemble learning methods for improving predictive accuracy, which aligns with this thesis's approach of applying non-linear machine learning models to capture complex patterns in firm-level financial data.

More recently, Barboza, Kimura, and Altman (2017) compared traditional statistical methods (e.g., logistic regression) with machine learning algorithms, including decision trees, random forests, and support vector machines in predicting bankruptcies. By incorporating both classic financial ratios and additional features influencing short-term firm performance, they created more nuanced models. Their findings revealed that machine learning models consistently outperform traditional approaches in terms of accuracy and robustness. However, the authors acknowledged the trade-off between performance and interpretability, a core concern in the adoption of machine learning in regulated industries. This study illustrates the increasing preference for data-driven methods capable of identifying complex, non-linear patterns that traditional models may overlook.

2.2. Macroeconomic factors and stress testing

While much of the early research on bankruptcy prediction has focused on firm-specific financial indicators, there is growing recognition that broader economic conditions also play a critical role in determining corporate financial distress. Macroeconomic shocks, such as recessions or sudden shifts in interest rates, can significantly influence the likelihood of firm failure, even among otherwise healthy companies. In response, a growing body of literature has explored how macro-level variables—GDP growth, inflation, interest rate, unemployment rate, and the VIX index—can be integrated into default prediction frameworks. At the same time, methods such as scenario analysis and macroeconomic stress testing have become increasingly relevant for assessing risk under adverse economic conditions. This section reviews key contributions in this area, highlighting how macroeconomic indicators and stress-testing approaches have been used to enhance bankruptcy prediction models and where gaps remain in combining them with machine learning.

Mensah (1984) offers criticism against the stationarity of multivariate bankruptcy prediction models, arguing that financial ratios and their predictive power vary across different economic environments. A key limitation of earlier studies, including Altman's, is that they do not account for macroeconomic conditions such as inflation, interest rates, and business cycles, which influence firms' financial health. Mensah finds that different financial indicators become relevant depending on the overall state of the economy, suggesting that static models may lose accuracy over time. His work emphasises the necessity of adaptive models that account for dynamic economic conditions, proposing that failure to incorporate macroeconomic variability could lead to misleading predictions and poor financial decisions.

Building on the recognition of the impact of macroeconomic conditions on financial stability, subsequent research has started to incorporate macro variables directly into default prediction models. Duffie, Saita, and Wang (2007) developed a model to estimate the term structure of conditional corporate default probabilities, incorporating both firm-specific and macroeconomic variables. Using maximum likelihood estimation, they analyse an extensive dataset of US industrial firms spanning from 1980 to 2004, covering more than 390,000 firm-month observations. Their findings show that future default probabilities are significantly influenced by a firm's distance to default (a volatility-adjusted leverage measure), its trailing one-year stock return, the trailing one-year return on the S&P 500 index, and US interest rates. Among these factors, changes in the distance to default have the greatest impact on the evolution of default risk over time. The model also outperforms other benchmarks, including models based on credit ratings, in out-of-sample forecasting. These results are highly relevant to the first part of this thesis, particularly in guiding the selection of important predictor variables (including both financial and macroeconomic indicators), highlighting the value of multi-period forecasting, and supporting the motivation to apply machine learning techniques to advance traditional bankruptcy prediction approaches.

While Duffie, Saita, and Wang (2007) focused on firm-level default prediction, Giglio, Kelly, and Pruitt (2016) reviewed broader systemic risks accumulating in the financial sector. This paper empirically evaluates the relationship between systemic risk and the macroeconomic environment, providing insights relevant to the integration of macroeconomic factors and stress testing in this thesis. The study examines how systemic risk accumulated within the financial sector increases risks to the real economy, particularly by raising the probability of macroeconomic downturns. By constructing systemic risk indices from aggregated measures, they show that financial sector stress significantly predicts the lower tail of the distribution of future macroeconomic shocks. Their findings suggest that volatility in financial sector equity prices is particularly informative for identifying downside macroeconomic risks. The study emphasises the asymmetric and nonlinear effects of systemic risk on economic outcomes, offering a methodological perspective relevant for analysing the macro-financial channels influencing corporate bankruptcy risk.

Country-level default studies provide further analogues to corporate bankruptcy, particularly in their use of macro-financial indicators to predict fiscal collapse. Reinhart and Rogoff (2009) trace financial crises across centuries and nations, identifying recurring patterns in sovereign defaults. They argue that excessive debt accumulation, especially denominated in foreign currency, combined with declining economic output and weak institutional governance, are the strongest predictors of default. Although the book focuses on nations, the framework is applicable to corporate entities operating in globalised, multi-currency markets as well. The behavioural dynamics (over-leveraging, mispricing of risk, and delayed responses to macro signals) mirror those observed in corporate finance.

Although Reinhart and Rogoff (2009) emphasised sovereign risk patterns, more recent studies have turned to micro-level modelling using machine learning techniques to capture complex, non-linear interactions. Sirignano, Sadhwani, and Giesecke (2018) trained deep neural networks on a dataset of over 120 million US mortgages to detect payment difficulties. While their study focuses on mortgage risk, the methodology offers important guidance for this thesis, emphasising that properly trained deep learning models can uncover non-linear relationships in large financial datasets. The results also highlight the importance of macroeconomic interactions and show that borrower difficulties predictably increase during periods of economic downturn. This reinforces the idea that business cycles should play a central role in all default prediction models, including those for corporate bankruptcies. The paper demonstrates that deep learning can go beyond the limitations of traditional linear models by identifying complex, non-linear interactions and dependencies.

A closely related study by Bussiere and Fratzscher (2006) investigated early warning systems for financial crises using a multinomial logit approach. Although the paper's primary focus is on currency crises measured by exchange market pressure, its findings are highly relevant for corporate default modelling. The methodology relies heavily on leading indicators (foreign reserves, GDP growth, and inflation) that could similarly be adapted to assess firm-level stress in international markets. The study underlines the value of forward-looking, high-frequency macro data in early identification of financial distress.

Beyond traditional financial indicators, emerging risks such as climate change have been increasingly recognised as systemic threats. Bolton et al. (2020) argue that climate change creates systemic risks for the financial system and suggest that environmental stress tests should be built into regulatory frameworks. The study points out that traditional backward-looking risk assessments are not sufficient given the uncertainty around climate change. Instead, the paper recommends forward-looking, scenario-based approaches, such as the "climate stress tests" being developed by several central banks and supervisors. These tests aim to measure how resilient the financial system is to climate-related shocks, with examples including Dutch banks' method of translating climate shocks into macroeconomic and portfolio effects, and the Bank of England's PRA stress tests based on scenario-driven valuation shocks. Although this study does not focus specifically on bankruptcy prediction, the idea of incorporating non-traditional, systemic shocks into financial risk models is highly relevant. Companies exposed to climate risks—particularly in industries like energy, agriculture, and insurance—may face long-term solvency challenges. Predictive models that neglect these emerging macro-financial variables might underestimate long-term default probabilities.

As concerns about systemic risk and macroeconomic instability grew, regulatory frameworks began to formalise stress testing as a core tool for risk assessment—most notably through the Basel III (Basel Committee on Banking Supervision (2011)) guidelines. The concept of stress testing has demonstrated the value of assessing risk exposure under adverse macroeconomic scenarios. While originally developed for banks, this approach has inspired similar methods for evaluating corporate bankruptcy risk, where firm solvency can also be highly sensitive to macroeconomic downturns.

Despite growing attention to macroeconomic influences on corporate bankruptcy, most existing studies either rely on traditional econometric models or apply machine learning techniques without incorporating economic scenario analysis. While earlier research has established that macroeconomic factors can affect bankruptcy risk, few studies have combined them with machine learning over extended time periods or embedded them within a stress testing framework. This thesis contributes to the literature by applying modern machine learning methods to predict corporate bankruptcies using both firm-level financial data and macroeconomic indicators, including forward-looking measures such as the VIX index. It also incorporates macroeconomic stress testing to simulate how bankruptcy probabilities evolve under adverse economic conditions, which approach is still not widely used in machine learning-based bankruptcy prediction. By bringing together firm-level prediction and scenario-based macro-financial risk assessment, this study aims to offer a more comprehensive and forward-looking framework for evaluating corporate default risk.

3. Hypotheses: Development

Corporate bankruptcies play an essential role in credit risk assessment, financial stability, and investment strategies. While firm-level financial ratios such as leverage and liquidity are widely used in bankruptcy prediction models, the role of macroeconomic conditions is often overlooked in empirical research. Since economic crises tend to follow cycles, understanding how external shocks influence corporate bankruptcies can improve risk assessment models for investors, lenders, and policymakers.

Traditional bankruptcy prediction models such as Altman's Z-score (Altman (1968)) primarily rely on firm-specific financial indicators. However, the 2008 financial crisis—and subsequent studies such as Giglio, Kelly, and Pruitt (2016)—drew attention to the importance of macroeconomic factors in corporate failures as economic downturns significantly increased bankruptcy rates across industries.

Existing literature has exclusively focused on firm-level indicators for bankruptcy prediction, overlooking macroeconomic factors. This thesis aims to add to the literature by bridging the gap between firm-level bankruptcy prediction and macroeconomic risk assessment by incorporating macroeconomic stress-testing techniques into bankruptcy models. The research question being considered in this study is whether macroeconomic factors and financial crises influence corporate bankruptcies in the United States. To achieve this, the study combines firm-level financial data with macroeconomic indicators, employ machine-learning techniques, and apply stress-testing to examine the effect of economic stress on the probability of bankruptcies. This section develops the testable hypotheses that guide the empirical analysis.

According to previously published studies, firm bankruptcies are not purely the result of internal financial weaknesses but are also influenced by broader economic conditions (Duffie, Saita, and Wang (2007)). Periods of economic downturn, rising interest rates, or high inflation can increase the financial pressure on firms, even on otherwise financially sound companies. However, most traditional bankruptcy models either exclude macroeconomic factors or include them only indirectly. This thesis explicitly integrates key macro variables—GDP growth, interest rates, inflation, unemployment, and the VIX index—into predictive models to assess their incremental contribution to model performance.

Hypothesis 1: The inclusion of macroeconomic variables improves the predictive accuracy of bankruptcy models compared to models using only firm-level financial data.

This hypothesis tests whether macro-level data provides additional explanatory power in identifying firms at risk of failure. The expectation is that incorporating wider economic conditions will allow models to capture systemic pressures that influence firm stability.

Beyond their potential contribution to predictive performance, macroeconomic variables allow for the construction of hypothetical stress scenarios to evaluate the model's sensitivity to adverse economic conditions. Stress testing simulates external shocks such as sharp increases in interest rates or market volatility and examines their impact on predicted bankruptcy probabilities.

Hypothesis 2: Applying macroeconomic stress scenarios will result in significantly higher predicted bankruptcy rates, indicating that the models effectively capture the influence of systemic economic shocks.

This hypothesis evaluates whether the models are responsive to declining macroeconomic conditions, thus offering a means to quantify risk under different economic environments.

Prior literature shows that machine learning models can outperform traditional statistical methods in predictive tasks, particularly when dealing with non-linear relationships and high-dimensional data (Barboza, Kimura, and Altman (2017)). Given the complexity of bankruptcy risk, which involves interactions between various financial and macroeconomic factors, machine learning models may be better suited to capture these patterns.

Hypothesis 3: Machine learning models (random forest, XGBoost, and neural network) will outperform the traditional logistic regression model in predicting corporate bankruptcies.

This hypothesis is supported by the growing evidence that ensemble and tree-based algorithms can capture subtle interactions and non-linear relationships that linear models may miss. However, this benefit may come at the cost of interpretability.

While firm-level and macroeconomic indicators are crucial in understanding bankruptcy risk, industry conditions also play a significant role. Firms within the same industry often face similar external pressures, such as input cost shocks, regulatory changes, or shifts in consumer demand. Aggregated industry-level financial ratios can capture these sector-wide trends and provide additional context for firm-level data.

Existing theoretical and empirical literature suggests that including industry-level information can enhance the predictive accuracy of bankruptcy models (Chava and Jarrow (2004)). For example, a firm operating in a declining industry may be more likely to fail, even if its individual financial ratios appear healthy. By incorporating average financial ratios for each industry, the model can better identify firms that are underperforming relative to their peers.

Hypothesis 4: The inclusion of industry-level financial variables improves the predictive performance of bankruptcy models beyond firm-level data alone.

This hypothesis evaluates whether industry-wide financial context increases the predictive power of the bankruptcy prediction models.

While both industry-level and macroeconomic variables are expected to enhance the performance of bankruptcy prediction models, the comparison of the two types of contextual information remains an important area to consider. Industry-level variables may offer more immediate financial relevance, while macroeconomic indicators reflect broader economic pressures that apply across all industries.

Hypothesis 5: Industry-level financial variables improve bankruptcy prediction accuracy more than macroeconomic variables when added to firm-level models.

This hypothesis compares the relative contribution of two layers of contextual information. The expectation is that industry-level ratios may offer greater predictive value than broader macroeconomic indicators as they are more closely tied to firm operations.

Together, these hypotheses guide the empirical research by testing the added value of macroeconomic, industry-level, and machine learning enhancements to traditional bankruptcy prediction approaches.

4. Data

The goal of this thesis is to predict corporate bankruptcy using a combination of firm-level financial indicators, and industry-level and macroeconomic variables. To achieve this, a dataset is required that contains both historical financial information and explicit bankruptcy outcomes over a sufficiently long time period. The selected data sources meet these criteria by providing comprehensive firm-level accounting data, market information, and bankruptcy labels, along with industry classifications and relevant macroeconomic indicators.

Firm-level financial data is essential for bankruptcy prediction, as it directly reflects factors which have consistently been shown to correlate with default risk in the literature such as a company's capital structure, liquidity, profitability, and operational stability. Including bankruptcy outcomes is equally critical, as these labels define the classification task and allow supervised learning models to be trained and evaluated.

Beyond the individual firm, industry-level variables offer a way to account for sector-specific conditions that may affect firms' likelihood of failure. These were constructed using average data from the same firm sample, grouped by industry and year, to reflect broader economic pressures within each sector.

Finally, macroeconomic variables were included to capture the influence of national and global economic conditions on firm solvency. Economic downturns, interest rate changes, and shifts in market volatility can all affect bankruptcy rates across firms, and including these variables allows for both improved model accuracy and the possibility of conducting stress tests under hypothetical economic scenarios.

4.1. Variable definition

The main equation of the thesis uses Altman's Z-score indicators (Altman (1968)) that were identified by multiple discriminant analysis (Working Capital/Total Assets, Retained Earnings/Total Assets, EBIT/Total Assets, Market Value of Equity/Book Value of Total Liabilities, and Sales/Total Assets) and, in addition, ratios that have a stronger short-term impact on financial performance: Growth of Assets, Growth in Sales, Growth in the number of employees, Operational Margin, Change in Return on Equity, and Change in Price-to-Book ratio as financial ratios (Barboza, Kimura, and Altman (2017)). These variables serve as independent variables in supervised learning frameworks, where a binary bankruptcy dependent variable (bankrupt = 1, non-bankrupt = 0) is regressed on these financial metrics.

As suggested by previous studies (Chava and Jarrow (2004)), industry-level variables play an important role in capturing the broader financial context within which firms operate. In this thesis, industry-level indicators were constructed by aggregating firm-level financial ratios as the mean value by two-digit SIC code and year. This approach allows each firm to be evaluated relative to the financial performance of its industry peers at a given point in time. The inclusion of these variables captures sector-wide trends such as industry-specific leverage norms, profitability cycles, or liquidity pressures, which may not be fully reflected in a firm's individual financials. For instance, a firm with moderate leverage may still be at higher risk if it operates in an industry undergoing a systemic downturn. Calculating industry ratios in this way also aligns with the methodology used in prior empirical research, enabling comparability with established findings.

While firm-specific indicators provide valuable insights into bankruptcy risk, existing literature (Giglio, Kelly, and Pruitt (2016) and Mensah (1984)) suggests that macroeconomic conditions play an important role in determining firms' financial health as well. Therefore, in order to examine the effects of macroeconomic factors on corporate bankruptcies, GDP growth, interest rate, inflation, unemployment rates, and the VIX index are introduced to the model in the third part of the thesis. These variables were selected as they represent fundamental economic forces that impact corporate solvency. GDP growth reflects overall economic performance; interest rates affect borrowing costs; inflation impacts purchasing power; unemployment signals labour market strength, and the VIX index is an indicator of market volatility.

Additionally, the macroeconomic stress testing framework is extended to include three crisis-sensitive variables: the National Financial Conditions Index (NFCI), the NASDAQ Composite Index (NASDAQCOM), and industrial production (INDPRO). These indicators provide a more specific angle of view of financial system stress, equity market trends, and real economic output, respectively, thus enhancing the robustness of the stress scenario simulations.

4.2. Data sources

The dataset for this thesis was constructed using multiple financial and macroeconomic sources to provide comprehensive analysis of the relationship between corporate bankruptcies and economic conditions in the United States. To identify bankrupt firms, data was extracted from CRSP's delisting database via WRDS. Firms that filed for bankruptcy between 1990 and 2023 were identified based on specific delisting codes, firstly DLSTCD = 574, which directly indicates bankruptcy. To ensure that no relevant cases were missed, additional delisting codes reflecting financial distress (DLSTCD = 560, 561, 572, 585) were also included. These codes helped capture various types of financially distressed exits (namely insufficient capital or assets, liquidation, and delisting due to protection of investors and the public interest), broadening the scope of the analysis.

To examine financial indicators, Compustat (WRDS) was used to extract selected financial variables from the financial statements for both bankrupt and non-bankrupt firms. Key variables were extracted from firms' annual financial statements, including net income, total assets, total liabilities, and earnings before interest and taxes. These base values were later used to construct financial ratios commonly associated with bankruptcy prediction models, such as profitability, leverage, and liquidity metrics. Non-bankrupt firms were also included in the sample to enable supervised learning. To match the CRSP delisting data with Compustat financials, the CRSP–Compustat Merged (CCM) link table was used, ensuring accurate alignment using firm identifiers.

In addition to firm-level data, macroeconomic variables were collected from the Federal Reserve Economic Data (FRED) database. These include GDP growth, interest rates, inflation, unemployment rate, and the VIX index, which reflect the broader economic environment and were included to test the hypothesis that corporate bankruptcy risk is influenced not only by firm-specific fundamentals but also by changes in the macroeconomic context. To enhance the stress testing framework and better capture the dynamics of financial crises, three additional variables were incorporated: the National Financial Conditions Index (NFCI), the NASDAQ Composite Index (NASDAQCOM), and the Industrial Production Index (INDPRO). These variables provide further insight into financial system stability, equity market movements, and real economic activity, respectively. All macroeconomic indicators were matched with firm-year observations based on fiscal years.

The dataset was filtered to focus on US-incorporated firms only (fic = “USA”). Following standard practice in bankruptcy literature (Ohlson (1980)), financial firms (SIC 6000–6999) and utilities (SIC 4900–4999) companies were excluded from the analysis. These industries operate under different financial structures and regulatory conditions, and their inclusion might bias the modelling outcomes or reduce comparability. As Ohlson (1980) noted, these sectors are structurally different and present challenges in data consistency and interpretation.

4.3. Dataset construction

Firstly, the three datasets were imported into R: Compustat annual fundamentals, CRSP delisting information, and the CRSP–Compustat Merged link table. After conducting the above-mentioned filtering, these datasets were merged using the GVKEY and PERMNO identifiers. When linking Compustat firm data to CRSP stock identifiers, only the highest-priority link per firm was included. Primary links (LINKPRIM = “C”) were prioritised, followed by secondary links (“P”), and all others were used solely if no better match was available.

The merged dataset included firm-level accounting data, bankruptcy indicators, and industry classifications. After merging, non-bankrupt firms were labelled as non-bankrupt (bankrupt_firm = 0). To ensure that bankruptcy predictions were based on observable firm behaviour prior to failure, only firms with a sufficient history before bankruptcy were retained. This was achieved by filtering for only those firms that had financial data for three years before the delisting year. To avoid data leakage, all observations from post-bankruptcy years were removed. The final dependent variable, a binary bankruptcy indicator (bankrupt = 1) was assigned to observations in the year immediately prior to the firm’s delisting. All other years were labelled as non-bankrupt (bankrupt = 0). With this setup, the thesis dataset provides information for one-year-ahead bankruptcy predictions.

One of the major challenges in constructing financial datasets across long time periods is dealing with missing values. In this thesis, a three-stage imputation process was implemented to balance robustness with completeness. Non-financial variables—gvkey, sic2, date variables, and the two binary bankruptcy variables—were excluded before the process. Firstly, Kalman smoothing was applied to firm-level financial time series where sufficient historical data was available. This method estimates the unobserved value of a time series model based on the full sample of the data allowing for capturing trends and the underlying structure of the time series

(Harvey (2009)). It is particularly suited for economic and financial data, as it can account for noise and adapt to structural shifts in firm performance. Secondly, firm-level median imputation was used to fill in missing values where Kalman smoothing was not feasible due to shorter time-series. This step ensures that missing values are filled with values that reflect the firm's own historical norms, maintaining internal consistency. Thirdly, any remaining gaps were filled using industry-year medians, grouped by two-digit SIC code and fiscal year. This provides a reasonable benchmark based on peer firms operating under similar economic and regulatory conditions. By using industry-level information, this step helps sustain cross-sectional comparability and prevents firms from being excluded due to isolated missing values.

After the missing values were handled, the selected financial ratios were calculated. Following the methodology of Barboza, Kimura, and Altman (2017), these ratios are Altman's Z-score (Altman (1968)) (Working Capital/Total Assets, Retained Earnings/Total Assets, EBIT/Total Assets, Market Value of Equity/Book Value of Total Liabilities, and Sales/Total Assets) and, in addition, ratios that have a stronger short-term impact on financial performance: Growth of Assets, Growth in Sales, Growth in the number of employees, Operational Margin, Change in Return on Equity, and Change in Price-to-Book ratio as financial ratios. To handle outlier values, all independent variables were winsorized at the 0.5th and the 99.5th percentiles. These bounds were deliberately set narrowly to keep extreme but informative values that are particularly relevant in bankruptcy prediction, where financial distress may be captured through unusually poor performance. Prior to winsorization, several variables showed implausible values in the billions, likely due to data entry or reporting errors. This transformation was applied consistently across all variables. Extreme values such as infinite or undefined numbers resulting from problematic divisions (e.g., division by zero) were converted to missing. In cases where a zero value had real economic meaning (zero growth or margin) the zero value was kept. Additional transformations included the logarithm of total assets (log_at) to reduce skewness, and a categorical version of the two-digit SIC code (sic2_f) for use in modelling. Firms that lacked sufficient history to compute these features were excluded.

After these preprocessing steps, firm-year observations with invalid financial entries such as zero total assets were filtered out. This stage also helped remove firms where financial statements were not interpretable, ensuring that model inputs would be meaningful.

The final cleaned dataset includes firm identifiers, bankruptcy labels, financial ratios, growth metrics, and macroeconomic variables. Financial ratios include working capital to total assets, retained earnings to total assets, EBIT to total assets, market value of equity to book value of liabilities, and sales to total assets. Short-term performance measures such as return on equity, price-to-book ratio, and operating margin were also included, along with year-over-year changes in firm size, sales, and employee counts.

To avoid look-ahead bias and ensure that no future information influences model predictions, all macroeconomic indicators were lagged by one year prior to inclusion in the dataset. These lagged variables were then matched to firm-level observations based on the corresponding fiscal year. It is important to note that the VIX index, one of the selected macroeconomic indicators, only became available in 1990. As a result, the analyses involving macroeconomic variables (presented in the third and fourth part of the thesis) covers solely the period from 1991 to 2023.

The final dataset includes 1102 bankrupt and 13,221 non-bankrupt firms with complete and reliable financial data. Given the objective of this thesis to construct a realistic bankruptcy prediction model and to incorporate macroeconomic indicators and stress testing, the dataset was intentionally kept imbalanced and split using a time-based approach. This decision reflects the real-world distribution of bankruptcy events, where bankruptcies account for only a small fraction of all firms. It is also consistent with the findings of (Sun et al. (2014)), who highlight the importance of bankruptcy prediction models on imbalanced datasets as existing models mostly rely on balanced datasets. Maintaining the original class imbalance allows for the model to be trained under conditions that mirror actual use cases in financial risk prediction. While balanced datasets are common in the bankruptcy prediction literature for interpretability, this study prioritises real-world applicability.

The dataset was split into training and test sets based on each firm's last available fiscal year. For bankrupt firms, this was defined as the year of delisting (DLSTDT), while for non-bankrupt firms, it was the last year in which they appeared in the data based on the fyear variable. Firms with a reference year prior to 2018 were assigned to the training set, while those with reference years in 2018 or later were allocated to the test set. This split ensured that model evaluation was performed on future, unseen data, mimicking real-world forecasting conditions. Each firm's full time series was kept together in the respective split. The training dataset included 1042 bankrupt and 8964 non-bankrupt firms, while the test dataset included 60 bankrupt and 4256 non-bankrupt firms. This setup is particularly important for later stress testing scenarios, where predictions must be made on out-of-sample data under changing economic conditions. To support this, a narrower test set from 2018 to 2019 was used during the stress testing phase that ensured that the stress simulations were applied to a stable pre-crisis economic environment, allowing the effects of hypothetical macroeconomic shocks to be evaluated in isolation from real-world disruptions such as the COVID-19 pandemic.

4.4. Properties of the final dataset

To better understand the characteristics of the dataset used in the modelling process, this section presents descriptive statistics for the selected financial ratios. Table 1 shows summary statistics for the full sample, as well as separately for bankrupt and non-bankrupt firms. This comparison provides insight into potential differences in financial performance and growth patterns between the two groups and supports the development of appropriate predictive models.

As expected, bankrupt firms generally exhibit weaker financial positions across several dimensions. The median values of many ratios, for instance EBIT/Total Assets, Retained Earnings/Total Assets, and Operational Margin are all lower for bankrupt firms than for non-bankrupt firms, reflecting lower profitability and accumulated deficits. Additionally, bankrupt firms tend to show more negative or volatile changes in key performance indicators such as return on equity (ROE) and the price-to-book ratio. These patterns are consistent with prior literature, where declining profitability, earnings volatility, and declining asset efficiency often signal bankruptcy.

However, some counterintuitive results can be observed from the summary statistics as well. For instance, bankrupt firms report slightly higher average values in variables such as Sales/Total Assets, Growth in Assets, and Growth in Sales, and even in Working Capital/Total Assets compared to their non-bankrupt counterparts. While these may seem positive at first glance, they may also reflect riskier or more aggressive expansion strategies that are not supported by underlying profitability or capital structure and thus potentially contributing to financial instability. These findings highlight the importance of evaluating firm performance through a combination of static ratios and change-based indicators.

Table 1

Descriptive statistics (mean, standard deviation (SD) minimum, 1st quarter, median, 3rd quarter, and maximum) of the full sample. First, data including all firms. Second, only data from bankrupt firms. Third, only data from non-bankrupt firms.

Variables	Mean	SD	Min	Q1	Median	Q3	Max
Working Capital/Total Assets	-0.31	4.36	-51.25	0.02	0.21	0.44	0.97
Retained Earnings/Total Assets	-6.90	46.93	-566.05	-1.10	-0.05	0.22	1.08
EBIT/Total Assets	-0.46	2.82	-32.23	-0.15	0.04	0.11	0.51
MV Equity/BV Total Liabilities	12.04	43.65	0.00	0.81	2.30	6.93	466.29
Sales/Total Assets	1.11	1.02	0.00	0.42	0.90	1.49	7.00
Growth in Assets	0.40	2.07	-0.88	-0.05	0.02	0.20	22.55
Growth in Sales	0.29	1.50	-1.00	-0.02	0.04	0.21	16.33
Growth in num. of employees	0.10	0.53	-0.94	-0.04	0.00	0.12	5.00
Change in ROE	-0.01	4.54	-33.64	-0.13	0.00	0.07	33.43
Change in Price-to-Book ratio	0.00	42.39	-324.17	-0.68	0.00	0.54	352.17
Operational Margin	-3.21	22.29	-256.58	-0.09	0.03	0.10	0.58
Bankrupt firms	**Mean**	**SD**	**Min**	**Q1**	**Median**	**Q3**	**Max**
Working Capital/Total Assets	0.15	1.16	-51.25	0.01	0.19	0.43	0.97
Retained Earnings/Total Assets	-1.72	9.67	-566.05	-1.22	-0.18	0.09	1.08
EBIT/Total Assets	-0.25	1.27	-32.23	-0.25	-0.01	0.08	0.51
MV Equity/BV Total Liabilities	10.95	39.31	0.00	0.58	1.77	6.14	466.29
Sales/Total Assets	1.17	1.06	0.00	0.36	0.98	1.65	7.00
Growth in Assets	0.48	2.25	-0.88	-0.09	0.00	0.25	22.55
Growth in Sales	0.39	1.81	-1.00	-0.05	0.02	0.26	16.33
Growth in num. of employees	0.13	0.65	-0.94	-0.07	0.00	0.13	5.00
Change in ROE	-0.03	5.25	-33.64	-0.27	-0.01	0.09	33.43
Change in Price-to-Book ratio	-0.51	40.30	-324.17	-0.88	0.00	0.53	352.17
Operational Margin	-3.18	20.55	-256.58	-0.30	0.00	0.06	0.58
Non-bankrupt firms	**Mean**	**SD**	**Min**	**Q1**	**Median**	**Q3**	**Max**
Working Capital/Total Assets	-0.34	4.49	-51.25	0.02	0.21	0.44	0.97
Retained Earnings/Total Assets	-7.22	48.31	-566.05	-1.09	-0.04	0.23	1.08
EBIT/Total Assets	-0.47	2.89	-32.23	-0.14	0.04	0.11	0.51
MV Equity/BV Total Liabilities	12.10	43.91	0.00	0.83	2.33	6.98	466.29
Sales/Total Assets	1.10	1.02	0.00	0.43	0.90	1.48	7.00
Growth in Assets	0.40	2.06	-0.88	-0.05	0.03	0.20	22.55
Growth in Sales	0.28	1.48	-1.00	-0.02	0.04	0.20	16.33
Growth in num. of employees	0.10	0.52	-0.94	-0.04	0.00	0.12	5.00
Change in ROE	-0.01	4.50	-33.64	-0.13	0.00	0.07	33.43
Change in Price-to-Book ratio	0.04	42.52	-324.17	-0.67	0.00	0.54	352.17
Operational Margin	-3.21	22.39	-256.58	-0.08	0.04	0.11	0.58

Moreover, the standard deviations for most variables are higher among bankrupt firms, especially for growth in sales, ROE change, and price-to-book ratio change. This suggests greater heterogeneity and volatility in performance among distressed firms. In contrast, non-bankrupt firms tend to have somewhat more stable distributions, with tighter interquartile ranges and fewer extreme outliers.

It is also important to note that due to the previous winsorization at the 0.5th and the 99.5th percentiles using the full sample, all three tables have the same minimum and maximum values for all variables.

Overall, the descriptive statistics confirm that meaningful financial differences exist between bankrupt and non-bankrupt firms. At the same time, the presence of overlapping distributions and some unexpected averages indicates that a multivariate modelling approach—particularly using flexible machine learning techniques—may be better suited to capturing the complex, non-linear interactions that contribute to bankruptcy risk.

In addition to firm-level factors, the thesis used macroeconomic variables as independent variables in the bankruptcy prediction models as well. Table 2 provides descriptive statistics about these factors.

Table 2

Descriptive statistics (mean, standard deviation (SD) minimum, 1st quarter, median, 3rd quarter, and maximum) of the macroeconomic variables

Variables	Mean	SD	Min	Q1	Median	Q3	Max
Interest rate	0.031	0.023	0.001	0.010	0.032	0.053	0.081
GDP growth	0.026	0.017	-0.026	0.019	0.028	0.039	0.062
Inflation	0.028	0.009	0.007	0.024	0.027	0.030	0.069
Unemployment	0.058	0.015	0.036	0.046	0.055	0.069	0.096
VIX index	0.195	0.057	0.111	0.142	0.178	0.242	0.327

The macroeconomic variables indicate various economic conditions throughout the sample period. Interest rates and GDP growth showed moderate variability, consistent with a mix of stable economic expansion and periods of slowdown. Inflation remained within a relatively narrow band, suggesting overall price stability during the period. In contrast, the unemployment rate showed greater fluctuation, reflecting shifts in labour market dynamics. Among the macroeconomic variables, the VIX index demonstrated the highest variability, likely driven by sharp shifts in investors behaviour during uncertain times.

5. Methodology

This section outlines the modelling approach used in this thesis to predict corporate bankruptcies based on firm-level financial indicators, industry-level variables, and macroeconomic information. Given the binary nature of the outcome (bankrupt or non-bankrupt) the study focused on classification models. The analysis involved training and evaluating several machine-learning models, beginning with a simple logistic regression baseline and proceeding to more advanced techniques: random forest, extreme gradient boosting (XGBoost), and neural network. Furthermore, scenario analysis was performed on the final models including macroeconomic variables.

Logistic regression served as a baseline model, providing interpretability and allowing comparison with more advanced techniques. Random forest was used to reduce overfitting and improve model stability, while extreme gradient boosting model was included due to its high predictive accuracy in financial classification tasks. Additionally, neural network was tested to capture complex interactions between firm-level financials and macroeconomic conditions. The thesis evaluates each model on classification metrics commonly used in credit risk literature (AUC-ROC curve, precision, recall, and F1 metrics). After all models were trained, stress testing was completed to evaluate the influence of macroeconomic shocks on corporate bankruptcies.

5.1. Theoretical background

5.1.1. Logistic regression

Logistic regression is a widely used statistical method for modelling a binary dependent variable, where the outcome takes one of two possible categories (e.g., bankrupt or non-bankrupt). Unlike linear regression, which is suitable for continuous outcomes, logistic regression is appropriate when the response variable is qualitative and categorical. The primary objective of logistic regression is to estimate the probability that a given observation belongs to a particular category based on one or more explanatory variables (De Menezes et al. (2017)). This estimated probability is typically denoted as $\pi(x)$, and by definition, it must fall within the [0, 1] interval.

However, a linear combination of the independent variables can take any real value in the range of $(-\infty, \infty)$, which is not compatible with probability modelling (Hosmer, Lemeshow, and Sturdivant (2013)). To address this issue, logistic regression applies the logit transformation, which maps the probability π(x) to the log-odds (or logit) scale:

$$logit(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right).$$

This transformation allows the model to express the log-odds as a linear function of the independent variables:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Here, the term $\frac{\pi(x)}{1-\pi(x)}$ represents the chance (odds) of the event occurring, and the output of its logarithm can fall on the entire real number line, making it suitable for linear modelling.

The inverse of the logit transformation yields the logistic function, which maps the linear combination of predictors back to the [0, 1] probability scale:

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}.$$

This function ensures that all predicted values of $\pi(x)$ are valid probabilities. In the case of a single predictor, if $\beta_1 > 0$, the probability of the event increases with x_1; if $\beta_1 < 0$, the probability decreases accordingly.

5.1.2. Random forest

Random forest is an ensemble learning method that builds on the idea of bagging (bootstrap aggregating) to improve the performance of decision trees. While bagging trains multiple decision trees on different bootstrap samples of the training data and aggregates their predictions (Breiman (1996)), random forest introduces additional randomness during tree construction to reduce correlation between trees and enhance generalisation (Breiman (2001)).

In bagging, each model is trained on a bootstrap sample, which dataset is created by sampling with replacement from the original training set (Breiman (1996)). Once all models are trained, their outputs are aggregated: in classification tasks, the final prediction is made by majority vote, while in regression, predictions are averaged. This approach is particularly effective for high-variance base learners such as decision trees, where aggregating multiple different models reduces overfitting.

Random forest extends this idea by introducing a modification during the tree-building process. At each node of every decision tree, instead of evaluating all available predictors to determine the optimal split, only a randomly selected subset of features is considered (Breiman (2001)). This further increases the diversity among the trees in the ensemble, which lowers their correlation and improves predictive performance.

Formally, if $\phi_b(x)$ denotes the prediction of the b-th tree trained on a bootstrap sample, the random forest prediction for a given input x in classification is:

$$\hat{y} = mode\ \{\phi_1(x), \phi_2(x), \dots, \phi_B(x)\},$$

while the random forest prediction for a given input x in regression is:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} \phi_b(x),$$

where B is the total number of trees in the forest.

Random forests are relatively robust to noise and outliers, and they can handle datasets with a large number of weak or correlated predictors (Breiman (2001)). Besides their predictive performance, they also provide internal estimates of variable importance and classification error using out-of-bag (OOB) samples—observations that were not included in the bootstrap sample used to train each tree.

5.1.3. Extreme gradient boosting (XGBoost)

XGBoost is a powerful and efficient machine learning method based on gradient boosting. Like other boosting techniques, it builds a strong predictive model by combining the outputs of several weaker models—in this case, decision trees (Chen and Guestrin (2016)). In contrast to bagging-based methods such as random forest, boosting focuses on reducing bias by training models one after another, with each tree attempting to correct the errors made by the previous ones.

The final prediction for a given input x_i is the sum of the outputs of K regression trees:

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i).$$

Each tree is trained to reduce the remaining error from the previous trees. XGBoost improves the basic gradient boosting approach by adding a regularised objective function. This function has two parts: a loss function and a regularisation term. The loss function ($l(y_i, \hat{y})$) measures how far the prediction is from the true value, while the regularisation term ($\Omega(f_t)$) penalises complex trees to help avoid overfitting. A commonly used form of the regularisation term is:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum w_j^2 ,$$

where T is the number of leaves in the tree and w_j are the values (weights) predicted by the leaves.

To improve training efficiency and generalisation, XGBoost uses gradient-based optimisation, relying on both the first and second derivatives of the loss function. It also includes several regularisation techniques to prevent overfitting such as shrinkage (learning rate), column subsampling, and automatic handling of missing values. XGBoost has also been shown to perform well on imbalanced datasets (Velarde et al. (2023)). This is particularly useful in classification tasks where one class is much rarer than the other.

5.1.4. Neural network

Neural networks are computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons). Each neuron computes a weighted sum of its inputs, applies an activation function (e.g., ReLU, sigmoid), and passes the result forward through the network (Wilamowski (2009)). Training involves minimising a loss function via gradient descent and backpropagation. Key hyperparameters include the number of layers, neurons per layer, activation functions, learning rate, and regularisation strategies such as dropout. Even though neural networks can model complex, non-linear relationships, they require more data and tuning compared to simpler models like logistic regression or decision trees. They are particularly powerful in applications involving high-dimensional and unstructured data.

5.2. Firm-level predictions

5.2.1. Logistic regression

For the baseline model of the thesis, firm-level financial ratios were used as independent variables, while the binary bankrupt variable served as the dependent variable. The financial indicators included Working Capital/Total Assets, Retained Earnings/Total Assets, EBIT/Total Assets, Market Value of Equity/Book Value of Total Liabilities, Sales/Total Assets, Growth in Assets, Growth in Sales, Growth in the Number of Employees, Operational Margin, Change in Return on Equity, Change in the Price-to-Book Ratio, Total Assets (log-transformed), and a categorical industry variable based on the two-digit SIC code. Logistic regression was chosen due to its interpretability and historical use in early bankruptcy prediction models (Altman (1968); Ohlson (1980)). The model was estimated using the binomial family from the glm function in R. In some cases, the model produced fitted probabilities that approached 0 or 1 due to strong separation between bankrupt and non-bankrupt firms in certain industries. This is a common issue in rare event modelling and, while it triggers standard warnings, it does not affect the model's predictive performance (Heinze and Schemper (2002)).

5.2.2. Random forest

Following the baseline model, a random forest classifier was trained. Random forest is a popular ensemble learning method known for its robustness and ability to handle non-linear relationships (Breiman (2001)). For the hyperparameter tuning Bayesian Optimisation was applied using the ParBayesianOptimization package in R. This method is more efficient than traditional approaches like grid search, especially when model training is time-consuming. Although grid search was initially considered, its computational inefficiency made it unsuitable for the modelling approach used in this thesis. Bayesian Optimisation offered a more effective alternative, as it selects new parameter combinations based on the results of previous ones, which helps reduce the number of evaluations needed. Although some practical limitations of this method are discussed in the literature such as difficulties in choosing the right internal settings, these did not cause any issues in this case, and the method performed well for the purposes of the analysis (Snoek, Larochelle, and Adams (2012)).

The hyperparameter tuning for the random forest model focused on two key parameters that directly influence the model's complexity and variance: mtry (the number of features randomly selected at each split) and max_depth (the maximum depth of each decision tree). The number of features (mtry) was tuned within a range derived from the square root of the total number of numeric predictors, a common method recommended by Breiman (2001) and Liaw and Wiener (2002). Specifically, the minimum was set to half of the square root of the number of predictors, and the maximum to twice the square root, ensuring the values remained within realistic bounds. This approach helps reduce the tuning space while focusing on values that are likely to perform well. The maximum tree depth was tuned between 3 and 16, consistent with the ranges used in prior machine learning literature on financial classification problems, where moderate tree depths are typically sufficient to capture complex interactions without overfitting (Probst, Wright, and Boulesteix (2019)). The Bayesian optimisation method targeted the AUC-ROC score on the validation data and used an initial random exploration of 5 points (initPoints = 5), followed by 15 guided iterations (iters.n = 15) to search for optimal hyperparameters. The acquisition function was set to "ucb" (upper confidence bound), with a kappa value of 2.576 to control the balance between exploration and exploitation. The final selected parameters were: mtry = 5.593 and max_depth = 14.152, with 300 trees (ntree = 300) determined based on the OOB error plot. A sample fraction of 5% per tree was used to increase tree diversity and reduce computational costs to improve generalisation performance (Cutler et al. (2007)). A second random forest model was included as well to assess impurity-based variable importance calculation (importance = "impurity" in the ranger package), which led to a slight reduction in AUC.

5.2.3. Extreme gradient boosting (XGBoost)

The third model trained was XGBoost, a gradient boosting framework that is especially effective on imbalanced datasets (Chen and Guestrin (2016); Velarde et al. (2023)). XGBoost is known for its high predictive accuracy and efficiency, making it a common choice in financial prediction tasks. Like the Random Forest, the XGBoost model was also tuned with Bayesian Optimisation using the ParBayesianOptimization package in R. The hyperparameter ranges used in the tuning process were selected based on established practices in the literature and are consistent with applications in financial default prediction. In particular, Barboza, Kimura, and Altman (2017) and Verma (2024) provide empirical and theoretical support for the tuning of parameters such as eta, max_depth, subsample, and colsample_bytree within the ranges used in this thesis. The inclusion and tuning of scale_pos_weight is further justified by its demonstrated effectiveness in handling class imbalance in classification tasks (He and Garcia (2009)).

The tuning process maximised the AUC score on the validation set and focused on parameters that influence the model's complexity and generalisation capacity. The learning rate (eta) was tuned within a range of 0.01 to 0.3. Lower values of eta help to slow down learning, reducing the risk of overfitting by allowing the model to build more trees with incremental improvements. The maximum depth of individual trees (max_depth) was tuned between 3 and 10, a range supported by previous studies on financial applications of XGBoost, which have shown that moderate tree depths help to maintain a balance between model interpretability and accuracy (Verma (2024)). Additionally, the subsample and colsample_bytree parameters—which control the fraction of observations and features sampled for each tree—were tuned between 0.6 and 1. These parameters introduce randomness into the tree-building process, increasing model robustness by reducing variance. The min_child_weight parameter specifies the minimum sum of instance weights needed to form a new leaf node, was optimised within a range of 1 to 10. Higher values of this parameter encourage more conservative tree growth and reduce the model's sensitivity to noisy patterns. Finally, scale_pos_weight—which adjusts the penalty for misclassifying the minority class—was tuned between 1 and the actual imbalance ratio observed in the training data. This parameter is particularly important in bankruptcy prediction, where the dataset is heavily skewed toward non-bankrupt firms (He and Garcia (2009)).

The final model was trained with 100 boosting rounds and included early stopping, which terminated training if performance on the validation set did not improve for 10 consecutive rounds. Like in the random forest tuning, the Bayesian optimisation used an initial random exploration of 10 points (initPoints = 10), followed by 15 guided iterations (iters.n = 15) to search for optimal hyperparameters. The acquisition function was set to "ucb" (upper confidence bound), with a kappa value of 2.576 to control the balance between exploration and exploitation. The best hyperparameters identified during the optimisation process were as follows: eta = 0.256, max_depth = 8, subsample = 0.6, colsample_bytree = 0.776, min_child_weight = 10, and scale_pos_weight = 1.

Feature importance was measured using the Gain metric provided by XGBoost, which reflects the average improvement in model performance attributable to each feature (Chen and Guestrin (2016)). As this measure directly quantifies a feature's contribution to reducing prediction error, it was used as the primary basis for evaluating importance. Other metrics such as Cover and Frequency were available but not included, as Gain is the most interpretable and widely used metric in both academic and applied machine learning contexts. In the XGBoost model, the sic2_f industry variable was one-hot encoded into separate binary features for each industry group. As a result, feature importance was calculated for each individual level (e.g., sic2_f73, sic2_f48). For comparison with the random forest model, which handles categorical variables natively, these individual scores were aggregated to reflect the overall contribution of the industry factor.

5.2.4. Neural network

Finally, a neural network was trained to evaluate whether a flexible, non-linear modelling approach could further enhance predictive performance. Given the high-dimensional and potentially non-linear relationships among financial ratios, neural networks have become an increasingly popular choice for classification problems in finance (Barboza, Kimura, and Altman (2017)).

Before training the neural network, all continuous input variables were standardised based on the mean and standard deviation calculated from the training set. This transformation was applied to ensure consistency in feature scaling, which enables stable model convergence and supports the efficient use of activation functions such as ReLU (LeCun, Bengio, and Hinton (2015)). The output layer employed a sigmoid activation function suitable for binary classification, and model training was conducted using binary cross-entropy as the loss function. The neural network was implemented using the keras package in R, which provides a high-level interface for building and training deep learning models on top of TensorFlow. TensorFlow was also used directly to ensure reproducibility by setting random seeds.

Due to the limited availability of Bayesian Optimisation for deep learning models in R, hyperparameter tuning was conducted using random search, which is a commonly recommended and computationally efficient strategy for neural networks (Bergstra and Bengio (2012)). The search space included the number of units in two hidden layers (ranging from 16 to 128 and 8 to 32, respectively), dropout rates (0.2 to 0.5), learning rates (0.0005 to 0.01), and batch sizes (32 to 128). This configuration was chosen to balance model complexity and regularisation in the context of a relatively small and imbalanced dataset.

The final model had two dense hidden layers with ReLU activations, followed by dropout regularisation and a single sigmoid output layer. ReLU was selected due to its favourable convergence properties and ability to handle sparsity. Dropout was used to mitigate overfitting by randomly deactivating a fraction of units during training, promoting robustness. The model was trained over 50 epochs using the Adam optimiser with early stopping based on validation loss. The best-performing configuration achieved an AUC of 0.7818, and its parameters were: units1 = 32, units2 = 8, dropout = 0.2867, learning_rate = 0.0035, batch_size = 128. While neural networks can capture complex patterns in data, they present challenges in terms of interpretability and can be sensitive to data imbalances.

5.3. Model evaluation

To assess the performance of the bankruptcy prediction models under normal circumstances, the thesis used classification evaluation. The primary metrics include AUC-ROC (Area Under the Receiver Operating Characteristic Curve), Precision, Recall, and F1-score, which are widely used in credit risk and bankruptcy prediction literature (Barboza, Kimura, and Altman (2017)). Since bankruptcies are relatively rare events, accuracy alone may be misleading as a model could classify most firms as non-bankrupt and still appear highly accurate.

A confusion matrix was constructed for each model to provide a detailed breakdown of classification outcomes. The matrix compares predicted versus actual values, allowing for the evaluation of true positives (correctly predicted bankruptcies), false positives (non-bankrupt firms predicted as bankrupt), false negatives (missed bankruptcies), and true negatives. Due to the high class imbalance in the dataset, the default threshold of 0.5 would not be appropriate for classifying bankruptcies. Instead, the optimal threshold was selected based on the ROC curve, using the point that maximises the true positive rate while minimising the false positive rate. This corresponds to the best trade-off between sensitivity and specificity, calculated using Youden's J statistic (Youden (1950)).

The ROC curve plots the true positive rate (TP) against the false positive rate (FP) at various threshold levels. The AUC (Area Under the Curve) summarises the performance across all possible thresholds into a single value ranging from 0 to 1. An AUC of 0.5 indicates no discriminatory power (equivalent to random guessing), while a value of 1.0 corresponds to perfect classification (Fawcett (2006)). Since AUC-ROC is threshold-independent, it provides a robust overall measure of model performance, especially under class imbalance. In addition to AUC, threshold-dependent metrics were calculated using the optimal classification threshold selected from the ROC curve. These include precision, recall, and F1-score.

Precision measures the proportion of predicted bankruptcies that were correct:

$$Precision = \frac{TP}{TP + FP}.$$

Recall (or sensitivity) captures the proportion of actual bankruptcies correctly identified:

$$Recall = \frac{TP}{TP + FN}.$$

F1-score is the harmonic mean of precision and recall, which provides a balanced measure when both false positives and false negatives are important:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}.$$

These metrics provide a more detailed view of model performance than accuracy alone, which can be misleading in imbalanced classification problems. Together, they offer complementary insights into both the general and practical performance of the models under evaluation.

5.4. Industry ratios

Beyond firm-level indicators, existing literature suggests that industry conditions play major role in bankruptcy risk as well (Chava and Jarrow (2004)). Firms do not operate in isolation, and they are influenced by common sectoral dynamics such as demand fluctuations, competitive intensity, regulatory changes, and supply chain constraints. In line with this, the present analysis incorporates industry-level financial ratios to assess whether sector-wide financial health adds explanatory power beyond firm-level data.

To test this effect within the thesis dataset, each of the models described above was re-estimated with the inclusion of industry-level financial ratios alongside with the firm-level metrics. These were calculated as the yearly average of each financial ratio across firms within the same two-digit Standard Industrial Classification (SIC) code. This approach enables the model to account for variations in the broader financial environment within which each firm operates, such as cyclical risk, industry-level leverage, or profitability trends.

The modelling process itself remained unchanged to ensure comparability. Hyperparameter tuning, training procedures, and evaluation metrics were held constant with the firm-level models. This methodological consistency ensures that any observed changes in model performance can be described specifically by the inclusion of industry-level data, rather than differences in training procedures or model configurations.

Incorporating industry-level metrics not only reflects the broader financial landscape in which firms operate but also serves as a bridge between the purely micro-level (firm-specific) analysis and the macroeconomic modelling introduced in the next section.

5.5. Macroeconomic variables

In the third and final modelling phase, the analysis focused on incorporating macroeconomic variables into the baseline firm-level models. The goal was to examine whether broader economic conditions provide additional predictive power in forecasting corporate bankruptcies. Specifically, five macroeconomic indicators were selected based on their theoretical and empirical relevance: GDP growth, inflation rate, interest rate (central bank policy rate), unemployment rate, and the VIX index (a measure of market volatility). These variables represent key aspects of the macroeconomic environment that may influence firm-level default risk (Mensah (1984)). Including them allows for a direct comparison between the explanatory power of industry-level and macro-level factors when added to the same firm-level baseline.

To isolate the effect of macroeconomic information and allow for clear model comparisons, this phase of the analysis excluded industry-level variables. Models were constructed using only firm-level financial indicators and selected macroeconomic variables. This structure enabled an evaluation of the additional predictive power contributed by macro-level factors relative to industry-specific characteristics. In doing so, the analysis assessed whether macroeconomic conditions offer meaningful improvement in forecasting accuracy when added to the baseline model.

Before incorporating the macroeconomic variables into the models, preliminary correlation analysis was conducted to assess multicollinearity. The results showed a relatively high correlation between the inflation rate and the central bank interest rate (Pearson's $r = 0.74$), which posed a risk of redundancy and instability in the predictive model. To address this, the inflation variable was excluded. The interest rate was retained, as it has a more direct influence on corporate borrowing costs and is likely to have a stronger causal link to default risk. Although dimensionality reduction techniques such as principal component analysis (PCA) were considered to mitigate multicollinearity, these were ultimately rejected in favour of model interpretability and compatibility with scenario-based stress testing conducted later in the study. Using original, interpretable macroeconomic variables allowed for more transparent economic narratives to be applied in the stress analysis phase.

To ensure consistency across the different modelling stages, the training procedures, hyperparameter tuning methods, and performance evaluation metrics remained identical to those used in the previous models. This methodological consistency is important not only for fair comparison, but also for interpreting performance differences as arising specifically from the introduction of macroeconomic predictors.

Overall, this step in the modelling process supports a deeper understanding of the role macroeconomic conditions play in firm-level bankruptcy prediction. It also lays the foundation for macroeconomic stress testing, which allows the trained models to be exposed to simulated adverse economic scenarios to assess potential default risks under hypothetical future conditions.

5.6. Macroeconomic stress testing

The final stage of the thesis involved macroeconomic stress testing to assess the sensitivity of the bankruptcy prediction model to adverse macroeconomic conditions. Stress testing is widely used in financial risk management and policy (Basel Committee on Banking Supervision (2011); International Monetary Fund (2020)) and is increasingly applied in firm-level research to evaluate how systemic shocks impact individual firm outcomes (Acharya, Engle, and Pierret (2013); Almeida et al. (2011)). Building on the approaches of Hamerle, Liebig, and Scheule (2004) and Duffie et al. (2009), this thesis adopted a macroeconomic stress testing framework grounded in historical data, specifically employing a scenario-based static stress testing methodology. Three historical crises were used as benchmark scenarios: the dotcom crisis (2000–2001), the global financial crisis (2007–2009), and the COVID-19 pandemic crisis (2020–2021).

For the main evaluation of model performance, the test set covered the period from 2018 to 2023. However, for the stress testing analysis, a separate test set from 2018 to 2019 was used to ensure a stable economic baseline and to avoid overlapping with real-world crises such as the COVID-19 pandemic. This approach allows the effect of simulated macroeconomic shocks to be isolated more clearly.

To prepare the datasets for stress simulations, macroeconomic variables in both the training and test datasets were standardised based on the training set's mean and standard deviation. This was necessary to ensure that macroeconomic variable inputs (interest rates, GDP growth, unemployment, and the VIX index) were on the same scale across all observations and to preserve consistency between training and testing distributions.

Macroeconomic stress values were calculated as the average of each macro variable during the full duration of each crisis. For the dotcom scenario averages from 2000 and 2001 were used, for the global financial crisis from 2007 to 2009, and COVID from 2020 to 2021. This approach produces more robust crisis representations by smoothing out year-specific volatility and avoiding the use of potentially unrepresentative single-year values. Each scenario was simulated by replacing the test dataset's macroeconomic variables with the standardised values calculated from the crisis periods. A logistic regression model was then used to generate bankruptcy probabilities under each scenario. The model was trained on the full pre-2018 data (using firm-level and macro predictors), and predictions were generated on the baseline test set and on each stressed version of the test set.

To compare the impact of the scenarios, average predicted risk and number of firms at risk were used as evaluation metrics. Average predicted risk measures the mean of predicted bankruptcy probabilities across firms, while the firms-at-risk shows the number of firms whose predicted bankruptcy probability exceeds a classification threshold. The threshold was selected using Youden's J statistic on the ROC curve of the baseline model (Youden (1950)), following the same evaluation procedure described earlier in the thesis. This ensures that risk classification is not arbitrarily fixed at 0.5, which would be inappropriate given the class imbalance.

While the baseline stress testing framework incorporated key macroeconomic indicators, its performance during the 2008 and COVID-19 crisis simulations suggested the potential omission of crisis-specific risk dimensions. To address this, indicators more directly tied to each crisis were incorporated: NASDAQ Composite index performance as a proxy for market volatility and the equity-driven nature of the dotcom bubble (Ofek and Richardson (2003)), the National Financial Conditions Index (NFCI) to reflect credit and liquidity stress during the 2008 financial crisis (Brave and Butters (2012)), and industrial production to capture the economic disruption caused by the COVID-19 pandemic (McKibbin and Fernando (2020)). These variables were selected based on theoretical relevance and data availability and were added jointly to a revised model to assess whether they improve sensitivity to crisis conditions.

6. Results

This section presents and interprets the performance metrics of all bankruptcy prediction models developed in the thesis. As outlined in the methodology chapter, model evaluation was based on four widely used classification metrics: the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), precision, recall, and F1-score. These metrics were selected due to their ability to capture different aspects of predictive accuracy, particularly in imbalanced classification problems such as bankruptcy prediction. The AUC-ROC provides an aggregate measure of performance across all possible classification thresholds, while precision, recall, and F1-score provide more detailed insight into the model's ability to correctly identify bankrupt companies below a given threshold.

The results are presented in three main comparative tables corresponding to each modelling stage: models using only firm-level financial indicators, models combining firm-level and industry-level variables, and models incorporating both firm-level and macroeconomic factors. Each table reports the performance of four classification algorithms—logistic regression, random forest, XGBoost, and neural network—allowing for both horizontal (between-model) and vertical (within-model) comparisons. While the tables summarise model performance based on the full sample of 14,323 firms, the confusion matrices report higher observation counts as each firm-year was treated as a separate instance in the one-year-ahead bankruptcy prediction framework.

A different table is dedicated to the macroeconomic stress testing component. Unlike the classification metrics, this part of the analysis evaluates model sensitivity under adverse macroeconomic scenarios. For each simulated crisis, the average predicted bankruptcy probability and the number of firms classified as at-risk are reported. This allows for a scenario-based assessment of the models' ability to respond to systemic shocks and provides practical insights into their behaviour in periods of economic distress.

6.1. Firm-level predictions

Table 3 summarises the evaluation metrics—AUC-ROC, precision, recall, and F1-score—for the firm-level prediction models. Due to the high class imbalance in the dataset (i.e., the proportion of bankrupt firms is very low relative to non-bankrupt firms), relying on accuracy alone would be misleading. Therefore, metrics better suited to imbalanced classification problems were used. All ROC curves for the firm-level models are shown on Figure A1 in Appendix A.

Table 3

The 14,323 firms tested in the firm-level prediction models. AUC means the area under the ROC curve. Precision is the proportion of predicted bankruptcies that were correct. Recall is the proportion of actual bankruptcies that were classified correctly. F1-score is the harmonic mean of precision and recall.

Firm-level models	AUC (%)	Precision	Recall	F1-score
Logistic regression	66.48	0.0012	0.9000	0.0025
Random forest	86.86	0.0037	0.7833	0.0073
XGBoost	92.57	0.0050	0.8667	0.0100
Neural network	75.54	0.0018	0.7000	0.0036

As expected, all machine learning models outperformed the baseline logistic regression model in terms of AUC-ROC. Among them, extreme gradient boosting (XGBoost) achieved the highest overall performance. Although neural networks are typically well-suited for capturing complex relationships in high-dimensional data, the two tree-based models outperformed it in this case, most likely due to the extreme class imbalance in the dataset. XGBoost also produced the best precision and F1-score, whereas logistic regression had the highest recall. However, the latter's extremely low precision resulted in the lowest F1-score among the four models. Interestingly, the neural network showed only a slight improvement compared to logistic regression in terms of precision and F1.

Logistic regression achieved an AUC of 66.48%, which is only moderately better than random guessing. Its recall rate, however, was remarkably high at 90%, meaning it was able to correctly identify nearly all actual bankruptcies in the test dataset (54 out of 60 as shown in Table B1 in Appendix B). Despite the high true positive rate, this model also generated a very high number of false positives classifying more than 43,000 non-bankrupt firms as bankrupt, which substantially reduced its precision and F1-score. These results suggest that the model over-predicts bankruptcies. Despite its lower predictive accuracy, logistic regression offers interpretability: all financial ratios were statistically significant at the 1% level, except for Change in ROE, Change in Price-to-Book Ratio, and Operating Margin. The log of total assets was also significant, indicating that firm size contributes to bankruptcy prediction.

The random forest model produced a considerably higher AUC of 86.86%, which was consistent with the literature on tree-based models. While its recall was slightly lower at 78.33% (correctly classifying 47 bankrupt firms), its false positive rate dropped dramatically to 12,817 compared to 43,293 in logistic regression. This yielded a higher precision of 0.37% and a much-improved F1-score of 0.0073. These improvements reflect the strength of random forests in balancing sensitivity and specificity, especially when tuned appropriately. The use of a smaller sample fraction and controlled tree depth also helped reduce overfitting. An additional advantage of this model is its ability to provide variable importance rankings. Interestingly, the Change in ROE variable appeared as the most important predictor—despite being statistically insignificant in the logistic regression model—demonstrating the ability of ensemble methods to detect complex interactions that traditional models might miss. Other key variables included Market Value of Equity/Book Value of Liabilities, Growth in Assets, and Growth in number of employees.

Out of all models XGBoost achieved the highest predictive power based on the AUC-ROC curve with the AUC of 92.57%. It further decreased the number of false positive predicted firms, while increased the true positives compared to the random forest model. This strong performance was expected as XGBoost has consistently been shown by existing literature to have high accuracy in handling imbalanced datasets, especially when hyperparameters such as scale_pos_weight are appropriately tuned. With a precision of 0.50% and an F1-score of 0.01, it not only maintained high sensitivity (recall: 86.67%) but also delivered the highest overall balance between identifying bankruptcies and avoiding false alarms. The most influential features in the firm-level XGBoost model were Market Value of Equity/Book Value of Total Liabilities, Growth in Assets, and Total Assets (log-transformed), indicating that firm size and capital structure are central to predicting bankruptcy (Table C4, Appendix C). Dynamic indicators like Change in ROE and Price-to-Book ratio also ranked highly, emphasising the importance of financial performance volatility. Compared to the random forest model (Table C1), the order of top features was mostly consistent, although XGBoost gave more relative weight to profitability and asset-based metrics, while random forest ranked Change in ROE and employee growth the highest.

The neural network outperformed logistic regression but fell behind both tree-based models. It achieved an AUC of 75.54%, with a recall of 70% (42 true positives) and a precision of 0.18%. Although its recall was fairly strong, its precision remained low. One possible explanation the extreme class imbalance, which neural networks often find difficult to handle. Furthermore, unlike the tree-based models, which were tuned using Bayesian Optimisation, the neural network was tuned using random search due to technical limitations in R. This may have prevented the model from reaching its full potential. Neural networks typically require large datasets and deeper architectures to perform optimally, which were not fully met in this study's setting. However, the model still demonstrated potential and outperformed the logistic regression baseline, suggesting that with more advanced tuning or a larger dataset, its performance could improve further.

6.2. Industry-level factors

Table 4 presents the evaluation metrics for the prediction models that incorporate both firm-level and industry-level financial variables. As in the firm-only models, the machine learning algorithms continued to outperform the traditional logistic regression. XGBoost maintained the highest performance across all four metrics, while logistic regression remained the weakest. Overall, the inclusion of industry-level indicators produced mixed effects: some models showed improvements, while others experienced a slight decline in performance compared to the baseline models. The corresponding ROC curves are shown in Figure A2 in Appendix A.

Table 4

The 14,323 firms tested in the firm and industry-level prediction models. AUC means the area under the ROC curve. Precision is the proportion of predicted bankruptcies that were correct. Recall is the proportion of actual bankruptcies that were classified correctly. F1-score is the harmonic mean of precision and recall.

Industry-level models	AUC (%)	Precision	Recall	F1-score
Logistic regression	71.34	0.0017	0.6500	0.0034
Random forest	85.91	0.0033	0.7833	0.0065
XGBoost	92.29	0.0035	0.9333	0.0070
Neural network	77.97	0.0020	0.7167	0.0040

The logistic regression model showed an increase in AUC to 71.34%, indicating a modest improvement over the firm-only model. It became less biased towards predicting bankruptcies, which reduced the recall rate from 90% to 65% (correctly identifying 39 bankrupt firms compared to 54 previously; see in Table B2, Appendix B). At the same time, false positives dropped significantly from 43,293 to 23,043. This shift resulted in a noticeable improvement in precision (from 0.12% to 0.17%) and F1-score, despite the lower recall. Among the independent variables, those previously not significant—Change in ROE, Change in Price-to-Book ratio, and Operating Margin—remained insignificant. Moreover, Sales/Total Assets also became statistically insignificant. Several industry-level ratios, however, showed significance: the industry means of Retained Earnings/Total Assets, Market Value Equity/Book Value of Liabilities, and Sales/Total Assets were significant at the 1% level, while Change in ROE and EBIT/Total Assets reached significance at the 5% level.

The random forest model outperformed logistic regression again with an AUC of 85.91%. However, its performance slightly declined compared to the firm-only version across all evaluation metrics. It correctly identified the same number of bankrupt firms (47), but a minor rise in false positives led to a small decrease in both precision and F1-score. As shown in Table C2 (Appendix C), the Change in ROE remained the most important feature, confirming its central role. The three other leading predictors also remained the same but in a different order. The industry-level ratios contributed less to the model's predictive power, suggesting that firm-level variability continued to dominate.

XGBoost presented the strongest performance among the models again, with an AUC of 92.29%, recall of 93.33%, and the highest F1-score. However, compared to the firm-only XGBoost, its performance declined slightly in most areas except for recall. The model correctly identified 56 bankrupt firms (compared to 52 previously), but this improvement came with a notable increase in false positives from 10,314 to 15,887. This trade-off led to a reduction in both precision and the F1-score. While XGBoost is known for its capacity to balance precision and recall in imbalanced datasets, the inclusion of industry-level metrics may have introduced redundant or noisy information that impacted its overall precision. In this setting, EBIT/Total Assets was the most dominant variable in the XGBoost model, while Operating Margin also gained higher importance. Market Value of Equity/Book Value of Liabilities and Growth in Assets remained top performers (Table C5, Appendix C). XGBoost's top features aligned broadly with random forest's again, but the extreme gradient boosting model placed relatively less weight on industry ratios overall, suggesting it relies more heavily on firm-specific data.

The neural network performed in between the tree-based models and logistic regression again. Its AUC rose to 77.97%, with modest improvements across all other metrics. The model correctly identified 43 bankrupt firms (up from 42 previously), and its number of false positives declined from 23,213 to 21,250. These changes suggest that the model slightly benefited from the inclusion of industry-level information, although it continued to fall behind the more structured tree-based methods. Given the relatively simple architecture and limited dataset size, further improvements might require deeper networks or alternative regularisation techniques.

Overall, as logistic regression and the neural network benefited from the inclusion of industry indicators, whereas random forest and XGBoost showed a decline in performance, Hypothesis 4 is only partially supported based on the analysis.

6.3. Macroeconomic variables

Table 5 presents the evaluation metrics for models that incorporate both firm-level and macroeconomic variables. As in previous sections, XGBoost remained the best-performing model in terms of AUC-ROC. However, the neural network underperformed all other models in this stage including logistic regression. This marks the only case where not all machine learning models outperform the traditional baseline. The introduction of macroeconomic indicators had a mixed impact on model performance relative to both the baseline (firm-level only) and industry-level models. The corresponding ROC curves are shown in Figure A3 in Appendix A.

Table 5

The 14,323 firms tested in the firm and macro-level prediction models. AUC means the area under the ROC curve. Precision is the proportion of predicted bankruptcies that were correct. Recall is the proportion of actual bankruptcies that were classified correctly. F1-score is the harmonic mean of precision and recall.

Macro-level models	AUC (%)	Precision	Recall	F1-score
Logistic regression	79.27	0.0023	0.8333	0.0047
Random forest	86.98	0.0058	0.7500	0.0115
XGBoost	91.44	0.0042	0.8833	0.0083
Neural network	78.65	0.0020	0.9000	0.0040

Logistic regression achieved its best performance in this setting with an AUC of 79.27%. It also reached its highest precision across all model configurations, at 0.23%, which reflects a significant reduction in false positives (21,328 compared to 43,293 in the baseline; see in Table B3 in Appendix B). Although its recall decreased from 90% in the firm-only model to 83.33% in the macro-setting, this trade-off improved the F1-score overall. The number of true positives increased to 50, suggesting a more balanced classification strategy compared to the industry-level prediction. In terms of predictor significance, Sales/Total Assets regained statistical significance at the 1% level, while Change in ROE, Change in Price-to-Book Ratio, and Operating Margin remained insignificant. Among the macro variables, only the VIX index was statistically not significant.

The random forest model also reached its best performance in this configuration, with an AUC of 86.98% and the highest precision of 0.58%, resulting in the highest F1-score among all four models. However, the model recorded its lowest recall (75%), indicating that it correctly identified fewer bankruptcies than in earlier setups. Notably, it also achieved the lowest number of false positives (7736) among all models across all configurations, highlighting its particularly conservative classification behaviour. This conservative tendency suggests a strong bias toward non-bankruptcy predictions, which may be preferable in contexts where false alarms carry high costs. This is further reflected in the variable importance rankings (Table C3, Appendix C), where firm-level variables continued to dominate. Change in ROE remained the top predictor. The macroeconomic variables contributed modestly, with the VIX index ranking highest among them in importance, despite not being significant in logistic regression.

XGBoost maintained the highest AUC (91.44%), stressing its strong overall classification ability; however, its performance was the poorest in this configuration. Consequently, Hypothesis 1 is not supported by the data of this thesis. While most models showed increased AUC scores in the macroeconomic setting, XGBoost experienced a slight decline. Despite the high AUC score, XGBoost's performance on the other three metrics was exceeded by at least one other model in each case. While the macro-level XGBoost model improved upon the industry-level version in terms of precision, it had a lower recall, and its F1-score remained below that of the random forest. XGBoost achieved a true positive count of 53 and a false positive count of 12,682, which was more balanced than logistic regression or the neural network, but less conservative than random forest.

Firm-level features dominated this model in this setting as well: Market Value of Equity/Book Value of Liabilities, Total Assets, and EBIT/Total Assets being the top performers (Table C6, Appendix 6). Among macroeconomic variables, interest rate and GDP growth contributed most, although their impact remained smaller than any top firm-level metric. One possible explanation for the relative underperformance of XGBoost in this setting lies in its sensitivity to feature noise and redundancy. The addition of macroeconomic variables may have introduced less informative or weakly correlated features that diluted the impact of the most predictive firm-level variables. Moreover, XGBoost's reliance on sequential tree boosting can sometimes lead to overfitting when additional weak predictors are added, especially in imbalanced datasets. Future work could investigate whether feature selection or dimensionality reduction techniques might help XGBoost maintain performance consistency across varying input spaces.

The neural network model performed best within its own configurations in this setting, achieving its highest values across all metrics compared to the firm-only and industry-level models. It reached an AUC of 78.65% and an exceptionally high recall of 90%, correctly identifying 54 of the 60 bankrupt firms in the test dataset. However, despite these improvements, the model performed the weakest among the four tested models in this macroeconomic setup. Its precision remained low at 0.20%, which led to an F1-score of 0.0040, the lowest among the models. The extremely high recall was most likely driven by overfitting and a tendency to overpredict bankruptcies, as indicated by the considerable increase in false positive classifications (26,715, compared to 23,213 in the firm-only model and 21,250 in the industry-level configuration).

These results suggest that while the neural network benefited from the inclusion of macroeconomic variables, it struggled to balance sensitivity and precision effectively. One reason may be that neural networks are particularly sensitive to irrelevant or weakly correlated input features, and the macroeconomic variables—although relevant—had lower importance scores than the core firm-level metrics. Without strong signals from the new variables, the model may have overfit, especially given the small number of positive bankruptcy cases. Additionally, the model architecture and tuning method (random search) may not have been flexible enough to adapt to the more complex input structure, further limiting its performance relative to tree-based models.

Based on all model performances in all configurations, Hypothesis 3 is partially supported: machine learning models generally outperformed logistic regression, however, the neural network underperformed in the macroeconomic setting. Furthermore, Hypothesis 5 cannot be supported as the effect of added variable types (industry-level and macroeconomic indicators) was highly model-specific, with no clear dominance of one over the other.

6.4. Macroeconomic stress testing

Table 6 summarises the results of the macroeconomic stress testing, which provides a forward-looking extension to traditional bankruptcy prediction models. This approach enables the evaluation of systemic weaknesses by assessing how current firm profiles might respond under adverse macroeconomic conditions that mimic historical crises. Such analysis offers valuable insights for investors, regulators, and lenders aiming to predict financial instability before actual downturns occur.

The stress testing framework used in this thesis is scenario-based and static in nature. Simulations were conducted using a fixed predictive model, and macroeconomic variables were altered to reflect the average conditions during three historical crises: the dotcom crash, the global financial crisis, and the COVID-19 pandemic. Firm-level indicators were held constant, ensuring that observed changes in predicted risk can be derived solely from shifts in macroeconomic conditions.

Table 6

The four tested scenarios. Average risk is the mean predicted bankruptcy probability. Firms-at-risk is the number of firms exceeding the classification threshold.

Scenarios	Average risk	Firms-at-risk
Baseline	0.0057	3866
Dotcom	0.0089	5748
2008	0.0050	3363
COVID	0.0034	1790

Due to the standardisation of the macroeconomic variables prior to stress testing, the absolute risk values in Table 6 cannot be interpreted as real-world bankruptcy probabilities. Instead, these values represent relative changes in predicted risk under different crisis-like conditions, assuming all firm-level characteristics remain constant. In this context, the focus is on how risk shifts across scenarios, not the specific values themselves. Figure D1 in Appendix D visualises the scenario outcomes. Although standardised inputs limit interpretability of exact probability values, the patterns of change across scenarios remain analytically useful.

The dotcom crisis simulation resulted in the highest predicted risk and the largest number of firms classified as at-risk with an increase of over 50% relative to the baseline. This suggests that the model is particularly sensitive to the macroeconomic environment of the early 2000s. Although unemployment was below the historical average, the scenario featured a notable combination of elevated interest rates (+0.93), high market volatility (+0.77), and above-average GDP growth (+1.14). Despite the growth signal, the model may have reacted more strongly to the tightened monetary conditions and heightened investor uncertainty, interpreting them as signs of financial stress. This implies that in the model's structure, market-based indicators such as interest rate and the VIX index may have a stronger influence on predicted bankruptcy risk than labour market measures alone.

In contrast, the global financial crisis (2008) scenario showed a slight decrease in predicted risk compared to the baseline. This may initially appear counterintuitive, given the depth of the crisis. However, the standardised values for the macro variables in this simulation suggest that the scenario did not reflect an extreme macro shock in the model's input space. The model showed only modest increases in interest rate (+0.25) and VIX (+0.22), along with lower-than-average GDP growth (–0.63) and unemployment (–0.67). It is possible that the firms in the test set are relatively less sensitive to the specific macro configuration associated with 2008, or that some protective firm characteristics (e.g., size, cash flow, sector) offset the macro risk. Another explanation is that within the model's static structure the credit constraints and liquidity shortages that defined the 2008 crisis are not directly captured through the included macro variables.

The COVID-19 simulation produced the lowest risk levels of all scenarios. This outcome aligns with the strongly negative standardised values for interest rate (–0.92) and GDP growth (–1.56), suggesting strong policy support during the pandemic, which may have helped reduce financial pressure on firms even though the overall economy was severely affected. Although the VIX index was moderately elevated (+0.47), unemployment remained close to average (–0.06), and the macro shock may have been interpreted by the model as relatively small from a financial pressure perspective. This likely reflects the stabilising effect of ultra-low interest rates and aggressive policy interventions, which helped to preserve firm solvency despite broader economic disruptions. At the same time, the unexpectedly low risk also reveals a limitation of the model: it does not incorporate time dynamics or delayed responses. For example, long-term impacts of reduced industrial production, disrupted supply chains, or firm-level financial fragility unfolding over time are not captured in this static, one-step prediction.

To further examine model responsiveness under crisis conditions, an extended stress testing framework was implemented incorporating three additional variables: NFCI, industrial production, and NASDAQ Composite performance. As before, the average risk values are standardised and should be interpreted as relative rather than absolute risk levels. The revised model produced more differentiated results across scenarios (Table 7). Notably, the predicted average risk under the 2008 scenario nearly doubled compared to the baseline (from 0.0028 to 0.0056), suggesting that the inclusion of a financial stress indicator (NFCI) significantly enhanced the model's sensitivity to credit market conditions during the global financial crisis. Similarly, risk under the dotcom scenario increased further, consistent with the equity-driven nature of that crisis being more fully captured by the NASDAQ composite variable. Interestingly, the COVID-19 simulation yielded an even lower predicted risk than in the baseline model, possibly because the added variables introduced noise rather than signal for this particular episode. Figure D2 in Appendix D visualises the scenario outcomes. It is also worth noting that this extended model was estimated using only logistic regression, which may not be well-suited for capturing complex non-linear relationships in multivariate stress environments.

Table 7

The four tested scenarios. Average risk is the mean predicted bankruptcy probability. Firms-at-risk is the number of firms exceeding the classification threshold.

Scenarios	Average risk	Firms-at-risk
Baseline	0.0028	3899
Dotcom	0.0079	6742
2008	0.0056	6277
COVID	0.0016	1589

These findings highlight that the model does not uniformly predict increased bankruptcy risk under all historical crises. While the revised stress testing framework improved sensitivity in the dotcom and 2008 scenarios, the predicted risk still declined under the COVID-19 scenario. As a result, Hypothesis 2 is rejected, as the models do not consistently indicate heightened bankruptcy risk across all simulated macroeconomic shocks.

This modelling approach has clear limitations. As a static framework, it does not incorporate the unfolding timeline of crises or feedback loops between macroeconomic and firm-level indicators. It assumes that the relationship between macro conditions and firm outcomes remains constant over time and across different contexts. Furthermore, it does not capture sectoral exposure differences or heterogeneity in firm resilience. Finally, the use of past crises as scenarios provides realistic but backward-looking shocks and may not account for macroeconomic risks caused by geopolitical instability or climate-related disruptions.

Despite these limitations, the model is valuable in identifying firms that may be particularly vulnerable under reasonable stress scenarios. It offers a structured, reproducible tool for forward-looking risk management that complements traditional backward-looking bankruptcy models. By applying worst-case macroeconomic inputs even during stable periods, the method can be used to flag at-risk firms in advance of actual economic downturns.

Future research could expand this framework in several ways. Dynamic stress testing models, potentially based on time-series or recurrent neural networks, could incorporate the evolving nature of economic shocks and firm responses. Sector-specific macro sensitivities could also be integrated, allowing for separated stress testing across industries. Moreover, stress scenarios could be constructed synthetically using Monte Carlo simulations or hypothetical policy shifts, broadening the framework's applicability beyond historical crises.

7. Conclusion

This thesis examined the predictive performance of four modelling techniques—logistic regression, random forest, XGBoost, and neural networks—in forecasting corporate bankruptcies using firm-level financial data, enhanced by industry-level and macroeconomic indicators. Through a structured, multi-phase modelling approach and a stress testing framework, the study aimed to assess not only model accuracy under typical conditions but also their responsiveness under simulated economic crises.

Overall, the findings underscore the strength of machine learning models in handling imbalanced bankruptcy prediction tasks, with XGBoost consistently outperforming the others in terms of AUC and F1-score across most settings. Tree-based models, particularly random forest, demonstrated stable performance and a conservative classification style, favouring fewer false positives, which feature is potentially valuable in regulatory or credit assessment contexts. Logistic regression, while more limited in accuracy, showed strength in recall and interpretability, especially when enhanced with macroeconomic variables. Neural network yielded mixed results, with performance constrained by class imbalance and less flexible tuning, particularly in the macroeconomic setting.

The macroeconomic stress testing framework provided further insights into model sensitivity during adverse conditions. The initial version of the framework produced counterintuitive results under the 2008 and COVID-19 crisis scenarios, which improved with the addition of targeted, crisis-relevant variables. Nevertheless, the stress testing approach remained limited by its static nature and reliance on historical crises. These constraints highlight the complexity of capturing systemic risk through a purely cross-sectional, one-step model, and point toward future possible research for dynamic or time-aware stress testing designs.

The findings offer partial support for several of the proposed hypotheses. While machine learning models generally outperformed logistic regression, the neural network's underperformance in the macro-setting limits this conclusion. Similarly, the effects of adding industry- and macro-level variables were highly model-specific, with no clear evidence of consistent benefit across all configurations. Most notably, the assumption that macroeconomic indicators would enhance model accuracy was not supported, particularly for the best-performing XGBoost model.

Despite these caveats, the thesis makes several meaningful contributions. It reinforces the importance of firm-level financial indicators in bankruptcy prediction, demonstrates the practical usefulness of ensemble learning techniques in imbalanced classification settings, and introduces a replicable macroeconomic stress testing framework that reveals model behaviour under adverse scenarios. Importantly, the modelling approach retains real-world relevance by using an imbalanced dataset that mirrors actual bankruptcy proportions enhancing applicability for lenders, regulators, and risk managers.

While this thesis contributes meaningful insights to the field of corporate bankruptcy prediction, several limitations should be noted. The modelling framework relied on a one-year-ahead prediction horizon, which is a common approach in the literature, however it inherently constrains the ability to capture dynamic financial distress processes such as feedback loops or temporal dependencies. As financial health evolves over time, future research would benefit from adopting time-aware models such as recurrent neural networks (RNNs), which can model sequential data and better reflect the progression of risk. Similarly, the stress testing framework, while innovative in its scenario-based design, remains static. This limits the framework's realism in representing how firms might adjust under pressure or how macro shocks unfold dynamically. Future work could explore more flexible stress testing techniques, including dynamic models or simulation-based methods that allow for co-evolving firm and macroeconomic variables over multiple time periods.

In terms of model inputs, the exclusive use of accounting-based indicators—although interpretable and widely used—may restrict the models' responsiveness to early warning signals available in financial markets. Incorporating market-based data such as equity volatility, credit spreads, or structural default measures could enhance predictive accuracy and timeliness. However, such integration introduces challenges related to data availability and comparability across firms as well. Additionally, the neural network architecture used in this thesis was relatively shallow due to technical constraints in R, limiting its ability to capture complex patterns. More advanced frameworks such as TensorFlow in Python would allow for deeper, better-regularised architectures and more rigorous hyperparameter tuning. Moreover, the models in this thesis did not account for inter-firm dependencies, an increasingly relevant consideration in systemic risk research. Future studies could incorporate network-based predictors derived from stock return correlations, supply chain data, or financial linkages, enabling a richer understanding of how firm distress may propagate across the economic system. Lastly, although this paper focused exclusively on firms incorporated in the United States due to data availability constraints, extending bankruptcy prediction analysis to other countries and regions represents a valuable area for future research.

In conclusion, this thesis highlights the strengths and trade-offs of different modelling strategies in bankruptcy prediction and underscores the continuing relevance of firm-level fundamentals in shaping risk outcomes. By integrating modelling accuracy with forward-looking stress analysis, it offers both practical insight and a foundation for further academic development in financial risk forecasting.

8. Reference list

Acharya, Viral V., Robert Engle, and Diane Pierret, 2013, Testing Macroprudential Stress Tests: The Risk of Regulatory Risk Weights, . NBER working paper series.

Almeida, Heitor, Murillo Campello, Bruno Laranjeira, and Scott Weisbenner, 2011, Corporate Debt Maturity and the Real Effects of the 2007 Credit Crisis, *Critical Finance Review* 1.

Altman, Edward I., 1968, Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance* 23, 589–609.

Barboza, Flavio, Herbert Kimura, and Edward Altman, 2017, Machine learning models and bankruptcy prediction, *Expert Systems with Applications* 83, 405–417.

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

Beaver, William H., 1966, Financial Ratios As Predictors of Failure, *Journal of Accounting Research* 4, 71–111.

Bergstra, James, and Yoshua Bengio, 2012, Random Search for Hyper-Parameter Optimization, *Journal of Machine Learning Research* 13, 281–305.

Bolton, Patrick, Morgan Despres, Luiz Awazu Pereira da Silva, Frédéric Samama, and Romain Svartzman, 2020, *The Green Swan - Central Banking and Financial Stability in the Age of Climate Change* (Bank for International Settlements, Basel, Switzerland).

Brave, Scott, and R Andrew Butters, 2012, Diagnosing the Financial System: Financial Conditions and Financial Stress, *International Journal of Central Banking* 8, 191–239.

Breiman, Leo, 1996, Bagging predictors, *Machine Learning* 24, 123–140.

Breiman, Leo, 2001, Random Forests, *Machine Learning* 45, 5–32.

Bussiere, Matthieu, and Marcel Fratzscher, 2006, Towards a new early warning system of financial crises, *Journal of International Money and Finance* 25, 953–973.

Chava, Sudheer, and Robert A Jarrow, 2004, Bankruptcy Prediction with Industry Effects, *Review of Finance* 8, 537–569.

Chen, Tianqi, and Carlos Guestrin, 2016, XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, San Francisco California USA).

Cutler, D. Richard, Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler, 2007, Random Forest For Classification in Ecology, *Ecology* 88, 2783–2792.

De Menezes, Fortunato S., Gilberto R. Liska, Marcelo A. Cirillo, and Mário J.F. Vivanco, 2017, Data classification with binary response through the Boosting algorithm and logistic regression, *Expert Systems with Applications* 69, 62–73.

Duffie, Darrell, ANDREAS ECKNER, Guillaume Horel, and LEANDRO SAITA, 2009, Frailty Correlated Default, *Journal of Finance* 64, 2089–2123.

Duffie, Darrell, Leandro Saita, and Ke Wang, 2007, Multi-period corporate default prediction with stochastic covariates$, *Journal of Financial Economics* 83, 635–665.

Fawcett, Tom, 2006, An introduction to ROC analysis, *Pattern Recognition Letters* 27, 861–874.

Giglio, Stefano, Bryan Kelly, and Seth Pruitt, 2016, Systemic risk and the macroeconomy: An empirical evaluation, *Journal of Financial Economics* 119, 457–471.

Hamerle, Alfred, Thilo Liebig, and Harald Scheule, 2004, Forecasting Credit Portfolio Risk, *SSRN Electronic Journal*.

Harvey, Andrew C., 2009, *Forecasting, Structural Time Series Models and the Kalman Filter*. Transf. to dig. print. (Cambridge Univ. Press, Cambridge).

He, Haibo, and E.A. Garcia, 2009, Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering* 21, 1263–1284.

Heinze, Georg, and Michael Schemper, 2002, A solution to the problem of separation in logistic regression, *Statistics in Medicine* 21, 2409–2419.

Hillegeist, Stephen A, Elizabeth K Keating, Donald P Cram, and Kyle G Lundstedt, 2004, Assessing the Probability of Bankruptcy, *Review of Accounting Studies* 9, 5–34.

Hosmer, David W., Jr., Stanley Lemeshow, and Rodney X. Sturdivant, 2013, *Applied Logistic Regression*. 3rd ed. (John Wiley & Sons Inc., Hoboken, New Jersey, USA).

International Monetary Fund, 2020, *Stress Testing: Principles, Concepts, and Frameworks*. Ed. Li Lian Ong and Andreas A. Jobst (International Monetary Fund, Washington, DC).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, 2015, Deep learning, *Nature* 521, 436–444.

Liaw, Andy, and Matthew Wiener, 2002, Classification and Regression by randomForest, *R News* 2, 18–22.

Malekipirbazari, Milad, and Vural Aksakalli, 2015, Risk assessment in social lending via random forests, *Expert Systems with Applications* 42, 4621–4631.

McKibbin, Warwick, and Roshen Fernando, 2020, The Global Macroeconomic Impacts of COVID-19: Seven Scenarios, *CAMA Working Paper*.

Mensah, Yaw M., 1984, An Examination of the Stationarity of Multivariate Bankruptcy Prediction Models: A Methodological Study, *Journal of Accounting Research* 22, 380–395.

Ofek, Eli, and Matthew Richardson, 2003, DotCom Mania: The Rise and Fall of Internet Stock Prices, *The Journal of Finance* 58, 1113–1137.

Ohlson, James A., 1980, Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research* 18, 109–131.

Probst, Philipp, Marvin N. Wright, and Anne-Laure Boulesteix, 2019, Hyperparameters and tuning strategies for random forest, *WIREs Data Mining and Knowledge Discovery* 9, e1301.

Reinhart, Carmen M., and Kenneth S. Rogoff, 2009, *This Time Is Different: Eight Centuries of Financial Folly* (Princeton University Press).

Shumway, Tyler, 2001, Forecasting Bankruptcy More Accurately: A Simple Hazard Model, *The Journal of Business* 74, 101–124.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke, 2018, Deep Learning for Mortgage Risk, (arXiv).

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams, 2012, Practical Bayesian Optimization of Machine Learning Algorithms, (arXiv).

Sun, Jie, Hui Li, Qing-Hua Huang, and Kai-Yu He, 2014, Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches, *Knowledge-Based Systems* 57, 41–56.

Velarde, Gissel, Anindya Sudhir, Sanjay Deshmane, Anuj Deshmunkh, Khushboo Sharma, and Vaibhav Joshi, 2023, Evaluating XGBoost for Balanced and Imbalanced Data: Application to Fraud Detection, (arXiv).

Verma, Vibhu, 2024, Exploring Key XGBoost Hyperparameters: A Study on Optimal Search Spaces and Practical Recommendations for Regression and Classification, *International Journal of All Research Education and Scientific Methods* 12, 3259–3266.

Wilamowski, Bogdan, 2009, Neural network architectures and learning algorithms, *IEEE Industrial Electronics Magazine* 3, 56–63.

Youden, W. J., 1950, Index for rating diagnostic tests, *Cancer* 3, 32–35.

9. AI use disclosure

In the development of this thesis, AI tools (specifically ChatGPT) were used in a limited and transparent way to support the writing and research process. This use was consistent with the university's AI Policy and strictly supplemental in nature. At no point were AI tools used to generate substantive content, formulate research questions, conduct analysis, or write core sections of the thesis independently. All ideas, arguments, data handling, and empirical work presented in this study are entirely my own.

AI assistance was primarily employed for language refinement such as improving sentence structure and checking grammar. Furthermore, AI was used to organise certain sections more effectively and rephrase ideas for improved readability. Additionally, the tool was used to obtain general literature search guidance such as identifying relevant papers and authors to explore. However, all sources cited in the thesis were selected, read, and evaluated independently through academic databases and university resources. AI tools also supported the coding work by helping in identifying errors and understanding R functions.

10. Appendices

Appendix A: ROC curves

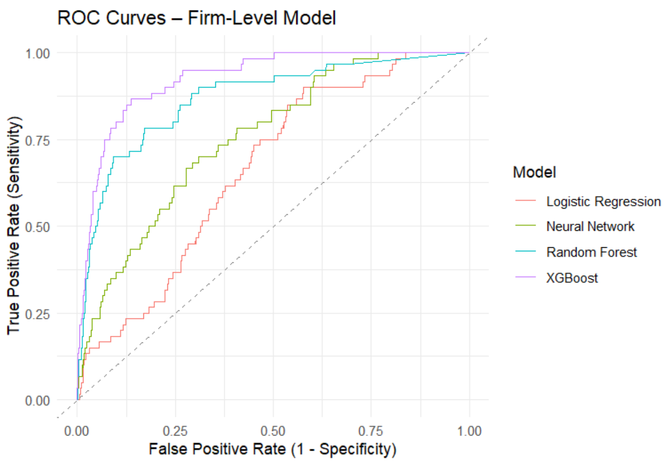

Figure A1: ROC curve for firm-level predictions

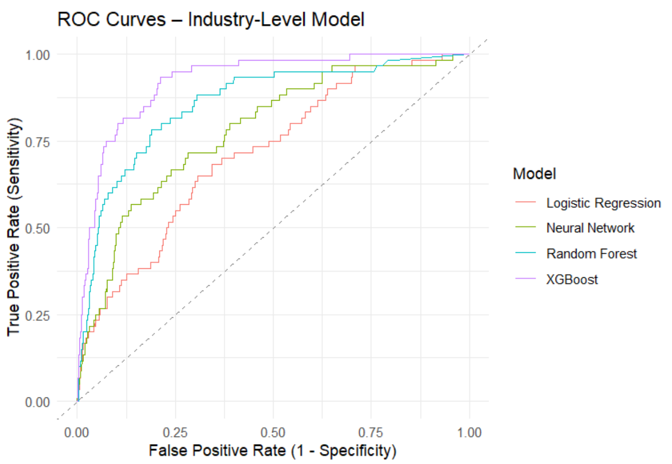

Figure A2: ROC curve for industry-level predictions

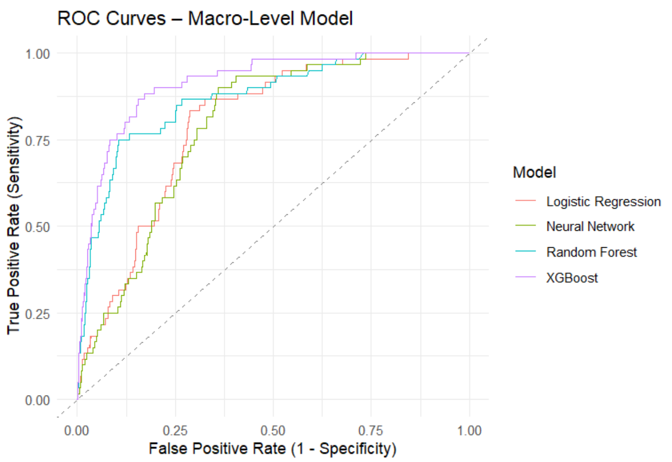

Figure A3: ROC curve for macro-level predictions

Appendix B: Confusion matrices

Logistic regression	Actual: 0	Actual: 1	Random forest	Actual: 0	Actual: 1
Predicted: 0	31,911	6	**Predicted: 0**	62,387	13
Predicted: 1	43,293	54	**Predicted: 1**	12,817	47

XGBoost	Actual: 0	Actual: 1	Neural network	Actual: 0	Actual: 1
Predicted: 0	64,890	8	**Predicted: 0**	51,991	18
Predicted: 1	10,314	52	**Predicted: 1**	23,213	42

Table B1: Confusion matrices of firm-level prediction models

Logistic regression	Actual: 0	Actual: 1	Random forest	Actual: 0	Actual: 1
Predicted: 0	52,161	21	**Predicted: 0**	60,957	13
Predicted: 1	23,043	39	**Predicted: 1**	14,247	47

XGBoost	Actual: 0	Actual: 1	Neural network	Actual: 0	Actual: 1
Predicted: 0	59,317	4	**Predicted: 0**	53,954	17
Predicted: 1	15,887	56	**Predicted: 1**	21,250	43

Table B2: Confusion matrices of industry-level prediction models

Logistic regression	Actual: 0	Actual: 1	Random forest	Actual: 0	Actual: 1
Predicted: 0	52,852	10	**Predicted: 0**	66,444	15
Predicted: 1	21,328	50	**Predicted: 1**	7,736	45

XGBoost	Actual: 0	Actual: 1	Neural network	Actual: 0	Actual: 1
Predicted: 0	61,498	7	**Predicted: 0**	47,465	6
Predicted: 1	12,682	53	**Predicted: 1**	26,715	54

Table B3: Confusion matrices of macro-level prediction models

Appendix C: Variable importance in random forest and XGBoost

Variable	Importance
Change in ROE	6.2410
MV of Equity/BV of Total Liabilities	5.8221
Growth in Assets	5.3054
Growth in number of employees	5.2573
Total Assets (logarithmic)	4.7176
Working Capital/Total Assets	4.6916
EBIT/Total Assets	4.6444
Change in Price-to-Book ratio	4.6038
Operating Margin	4.5570
Growth in Sales	4.4714
Retained Earnings/Total Assets	4.2510
Sales/Total Assets	3.7408
SIC2	3.2031

Table C1: Importance of variables in firm-level random forest model

Variable	Importance
Change in ROE	4.7935
Growth in number of employees	3.6215
Growth in Assets	3.4723
MV of Equity/BV of Total Liabilities	3.4129
Growth in Sales	2.8719
Change in Price-to-Book ratio	2.7795
Total Assets (logarithmic)	2.7186
Working Capital/Total Assets	2.6777
Operating Margin	2.6440
EBIT/Total Assets	2.5970
Industry Working Capital/Total Assets	2.3021
Retained Earnings/Total Assets	2.2787
Industry Change in ROE	2.1170
Industry Change in Price-to-Book ratio	2.1111
Sales/Total Assets	2.0804
Industry Growth in number of employees	1.9982
Industry Growth in Sales	1.9739
Industry Sales/Total Assets	1.8766
Industry Growth in Assets	1.8303
Industry Retained Earnings/Total Assets	1.7878
Industry MV of Equity/BV of Total Liabilities	1.6749
Industry Operating Margin	1.6696
Industry EBIT/Total Assets	1.6253
SIC2	1.4392

Table C2: Importance of variables in industry-level random forest model

Variable	Importance
Change in ROE	5.8303
MV of Equity/BV of Total Liabilities	5.4180
Growth in number of employees	4.8375
Growth in Assets	4.7671
Change in Price-to-Book ratio	4.3175
Total Assets (logarithmic)	4.1729
Working Capital/Total Assets	3.8668
Retained Earnings/Total Assets	3.8563
EBIT/Total Assets	3.8496
Operating Margin	3.8170
Growth in Sales	3.7377
Sales/Total Assets	3.1429
SIC2	3.0539
VIX index	2.0189
Interest rate	1.5132
GDP growth	1.4312
Unemployment	1.3984

Table C3: Importance of variables in macro-level random forest model

Feature	Gain
MV of Equity/BV of Total Liabilities	0,1484
Growth in Assets	0,1255
Total Assets (logarithmic)	0,1199
EBIT/Total Assets	0,1139
Change in ROE	0,0785
Working Capital/Total Assets	0,0710
Change in Price-to-Book ratio	0,0692
Growth in number of employees	0,0629
Operating Margin	0,0613
Retained Earnings/Total Assets	0,0519
Growth in Sales	0,0506
Sales/Total Assets	0,0385
SIC2	0,0084

Table C4: Importance of variables in firm-level XGBoost model

Feature	Gain
EBIT/Total Assets	0,1797
MV of Equity/BV of Total Liabilities	0,1731
Growth in Assets	0,1157
Operating Margin	0,0924
Total Assets (logarithmic)	0,0898
Change in ROE	0,0653
Working Capital/Total Assets	0,0437
Change in Price-to-Book ratio	0,0427
Retained Earnings/Total Assets	0,0354
Growth in number of employees	0,0322
Growth in Sales	0,0292
Industry Operating Margin	0,0137
SIC2	0,0116
Industry Working Capital/Total Assets	0,0098
Industry Sales/Total Assets	0,0095
Industry Growth in Assets	0,0093
Industry Change in ROE	0,0093
Sales/Total Assets	0,0078
Industry EBIT/Total Assets	0,0069
Industry Growth in number of employees	0,0063
Industry Retained Earnings/Total Assets	0,0062
Industry MV of Equity/BV of Total Liabilities	0,0040
Industry Growth in Sales	0,0033
Industry Change in Price-to-Book ratio	0,0032

Table C5: Importance of variables in industry-level XGBoost model

Feature	Gain
MV of Equity/BV of Total Liabilities	0,1386
Total Assets (logarithmic)	0,1308
EBIT/Total Assets	0,1191
Growth in Assets	0,1178
Operating Margin	0,0790
Change in ROE	0,0703
Working Capital/Total Assets	0,0441
Change in Price-to-Book ratio	0,0435
Interest rate	0,0416
Growth in number of employees	0,0401
Retained Earnings/Total Assets	0,0349
Sales/Total Assets	0,0339
GDP growth	0,0293
Unemployment rate	0,0256
VIX index	0,0204
Growth in Sales	0,0201
SIC2	0,0110

Table C6: Importance of variables in macro-level XGBoost model

Appendix D: Macroeconomic stress testing results

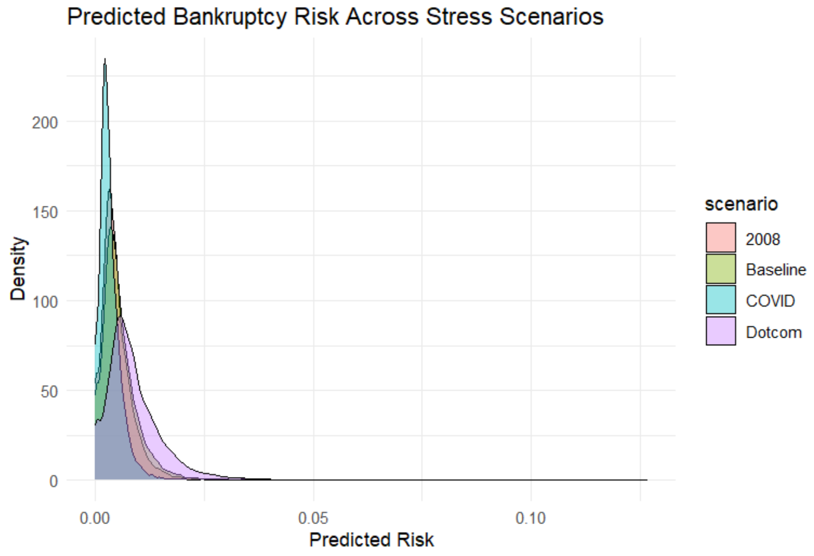

Figure D1: Predicted bankruptcy risk across the simulated stress scenarios

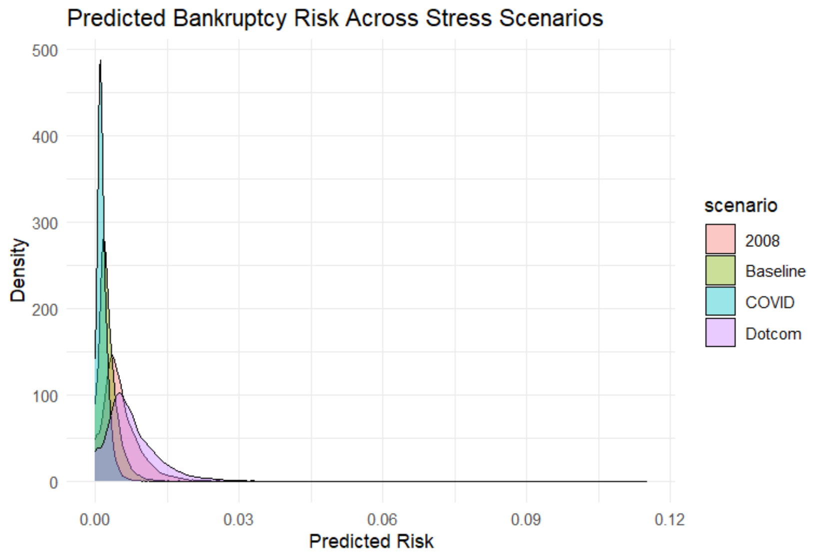

Figure D2: Predicted bankruptcy risk across the simulated stress scenarios with the additional crises-specific variables

Appendix E: R Code

Due to the overall length of the R code, only the baseline firm-level model and the initial stress testing framework are included in this appendix. These sections reflect the key structure and logic of the modelling process.

```
# 2. FIRM-LEVEL MODELS ---------------------------------------------------------
# Logistic regression ----
lr_model <- glm(bankrupt ~ wc_ta + re_ta + ebit_ta + mve_bvlt + sale_ta +
                  asset_growth + sales_growth + emp_growth +roe_change +
                  pb_change + op_margin + log_at + sic2_f,
                data = train_data, family = "binomial")
lr_pred <- predict(lr_model, newdata = test_data, type = "response")

summary(lr_model)

# AUC - ROC
auc_lr <- roc(test_data$bankrupt, lr_pred)
print(auc_lr)
plot(auc_lr, main = "ROC Curve for Logistic Regression")

# Confusion matrix
best_thresh_lr <- as.numeric(coords(auc_lr, "best", ret = "threshold", transpose =
FALSE))
pred_class_lr <- ifelse(lr_pred >= best_thresh_lr, 1, 0)
conf_matrix_lr <- table(Predicted = pred_class_lr, Actual = test_data$bankrupt)
print(conf_matrix_lr)

# Precision, recall, and F1
TP_lr <- conf_matrix_lr["1", "1"]
FP_lr <- conf_matrix_lr["1", "0"]
FN_lr <- conf_matrix_lr["0", "1"]

precision_lr <- TP_lr / (TP_lr + FP_lr)
precision_lr
recall_lr <- TP_lr / (TP_lr + FN_lr)
recall_lr
f1_lr <- 2 * (precision_lr * recall_lr) / (precision_lr + recall_lr)
f1_lr


# Random forest ----
rf_formula <- bankrupt ~ wc_ta + re_ta + ebit_ta + mve_bvlt + sale_ta +
  asset_growth + sales_growth + emp_growth + roe_change + pb_change +
  op_margin + log_at + sic2_f

set.seed(123)
# OOB for tree count selection
rf_oob_plot <- randomForest(
  bankrupt ~ wc_ta + re_ta + ebit_ta + mve_bvlt + sale_ta +
    asset_growth + sales_growth + emp_growth + roe_change +
    pb_change + op_margin + log_at,
  data = train_data,
  ntree = 1000,
  importance = TRUE
)

plot(rf_oob_plot, main = "OOB Error vs. Number of Trees")


# Count number of predictors used in your model
# Exclude: target + IDs + factor variables (if already dummy-encoded)
numeric_predictors <- train_data %>%
  select(where(is.numeric)) %>%
```

```
  select(-bankrupt_firm)

p <- ncol(numeric_predictors)
cat("Number of numeric predictors:", p, "\n")

# Suggest realistic bounds
suggested_mtry_min <- max(1, floor(sqrt(p) / 2))
suggested_mtry_max <- min(p, ceiling(sqrt(p) * 2))
suggested_depth_min <- 3
suggested_depth_max <- 16

cat("Suggested mtry range: ", suggested_mtry_min, "to", suggested_mtry_max, "\n")
cat("Suggested max.depth range: ", suggested_depth_min, "to", suggested_depth_max,
"\n")

# Define scoring function for RF
rf_cv_score <- function(mtry, max_depth) {
  model <- ranger(
    formula = bankrupt ~ wc_ta + re_ta + ebit_ta + mve_bvlt + sale_ta +
      asset_growth + sales_growth + emp_growth + roe_change + pb_change +
      op_margin + log_at + sic2_f,
    data = train_data,
    num.trees = 300,
    mtry = floor(mtry),
    max.depth = floor(max_depth),
    probability = TRUE
  )

  preds <- predict(model, data = test_data)$predictions[, "1"]
  auc <- pROC::auc(test_data$bankrupt, preds)
  return(list(Score = auc))
}

# Run Bayesian Optimization
set.seed(123)
opt_rf <- bayesOpt(
  FUN = rf_cv_score,
  bounds = list(
    mtry = c(suggested_mtry_min, suggested_mtry_max),
    max_depth = c(suggested_depth_min, suggested_depth_max)
  ),
  initPoints = 5,
  iters.n = 15,
  acq = "ucb",  # acquisition function
  kappa = 2.576 # controls exploration vs. exploitation
)

# Best parameters
getBestPars(opt_rf)
ntree <- 300

set.seed(123)
rf_model <- ranger(
  formula = rf_formula,
  data = train_data,
  sample.fraction = 0.05,
  num.trees = ntree,
  max.depth = opt_rf$max_depth,
  mtry = opt_rf$mtry,
  probability = TRUE)

# Evaluate on test set
rf_probs <- predict(rf_model, data = test_data)$predictions[, "1"]
roc_rf <- pROC::roc(test_data$bankrupt, rf_probs)
auc(roc_rf)
pROC::plot.roc(test_data$bankrupt, rf_probs, main = "ROC Curve for Random Forest")

# Importance
```

```
set.seed(123)
rf_model_imp <- ranger(
  formula = rf_formula,
  data = train_data,
  sample.fraction = 0.05,
  num.trees = ntree,
  max.depth = opt_rf$max_depth,
  mtry = opt_rf$mtry,
  probability = TRUE,
  importance = "impurity")

# Feature importance
importance_df <- data.frame(
  Variable = names(rf_model_imp$variable.importance),
  Importance = rf_model_imp$variable.importance
) %>%
  arrange(desc(Importance))

# Print the sorted importance table
print(importance_df)


# Confusion matrix
best_thresh_rf <- as.numeric(coords(roc_rf, "best", ret = "threshold", transpose =
FALSE))
pred_class_rf <- ifelse(rf_probs >= best_thresh_rf, 1, 0)
conf_matrix_rf <- table(Predicted = pred_class_rf, Actual = test_data$bankrupt)
print(conf_matrix_rf)

# 4. Precision, recall, F1-score
TP_rf <- conf_matrix_rf["1", "1"]
FP_rf <- conf_matrix_rf["1", "0"]
FN_rf <- conf_matrix_rf["0", "1"]

precision_rf <- TP_rf / (TP_rf + FP_rf)
recall_rf <- TP_rf / (TP_rf + FN_rf)
f1_rf <- 2 * (precision_rf * recall_rf) / (precision_rf + recall_rf)

precision_rf; recall_rf; f1_rf


# XGBoost ----
# Remove IDs and non-predictor columns
feature_cols <- c("wc_ta", "re_ta", "ebit_ta", "mve_bvlt", "sale_ta", "asset_growth",
                  "sales_growth",    "emp_growth",    "roe_change",    "pb_change",
"op_margin",
                  "log_at", "sic2_f")

# Ensure all predictors are numeric (convert factors like sic2_f to dummies)
train_matrix <- model.matrix(~ . - 1, data = train_data[, feature_cols])
test_matrix  <- model.matrix(~ . - 1, data = test_data[, feature_cols])

# Convert target to numeric vector (0/1)
train_label <- as.numeric(as.character(train_data$bankrupt))
test_label  <- as.numeric(as.character(test_data$bankrupt))

# Create DMatrix objects (XGBoost format)
dtrain <- xgb.DMatrix(data = train_matrix, label = train_label)
dtest  <- xgb.DMatrix(data = test_matrix, label = test_label)


# Hyperparameter tuning (XGBoost)
# 1. Data summary for tuning ranges
# Number of predictors (after model.matrix)
p <- ncol(train_matrix)
imbalance_ratio <- sum(train_label == 0) / sum(train_label == 1)

cat("Number of features:", p, "\n")
```

```
cat("Imbalance ratio (non-bankrupt / bankrupt):", round(imbalance_ratio, 2), "\n")

# 2. Define the scoring function
xgb_cv_score    <-    function(eta,    max_depth,    subsample,    colsample_bytree,
min_child_weight, scale_pos_weight) {
  params <- list(
    objective = "binary:logistic",
    eval_metric = "auc",
    eta = eta,
    max_depth = as.integer(max_depth),
    subsample = subsample,
    colsample_bytree = colsample_bytree,
    min_child_weight = min_child_weight,
    scale_pos_weight = scale_pos_weight
  )

  model <- xgb.train(
    params = params,
    data = dtrain,
    nrounds = 100,
    watchlist = list(eval = dtest),
    verbose = 0,
    early_stopping_rounds = 10
  )

  pred <- predict(model, newdata = dtest)
  auc_score <- pROC::auc(test_label, pred)
  return(list(Score = auc_score))
}

# 3. Run Bayesian Optimization
set.seed(123)

opt_xgb <- bayesOpt(
  FUN = xgb_cv_score,
  bounds = list(
    eta = c(0.01, 0.3),
    max_depth = c(3L, 10L),
    subsample = c(0.6, 1),
    colsample_bytree = c(0.6, 1),
    min_child_weight = c(1, 10),
    scale_pos_weight = c(1, imbalance_ratio)
  ),
  initPoints = 10,
  iters.n = 15,
  acq = "ucb",
  kappa = 2.576,
  parallel = FALSE,
  verbose = 1
)

# Final XGBoost model with best parameters
best_params_xgb <- getBestPars(opt_xgb)
print(best_params_xgb)

final_params_xgb <- list(
  objective = "binary:logistic",
  eval_metric = "auc",
  eta = best_params_xgb$eta,
  max_depth = as.integer(best_params_xgb$max_depth),
  subsample = best_params_xgb$subsample,
  colsample_bytree = best_params_xgb$colsample_bytree,
  min_child_weight = best_params_xgb$min_child_weight,
  scale_pos_weight = best_params_xgb$scale_pos_weight
)

xgb_model <- xgb.train(
  params = final_params_xgb,
```

```
  data = dtrain,
  nrounds = 100,
  watchlist = list(eval = dtest),
  print_every_n = 10,
  early_stopping_rounds = 10
)

xgb_pred <- predict(xgb_model, newdata = dtest)

# Get feature importance from the trained XGBoost model
importance_df_xgb <- xgb.importance(feature_names = colnames(train_matrix), model =
xgb_model)
importance_df_xgb <- importance_df_xgb %>%
  arrange(desc(Gain))

sic_importance <- importance_df_xgb %>%
  filter(grepl("sic2_f", Feature)) %>%
  summarise(total_sic_gain = sum(Gain))

print(importance_df_xgb)

# AUC
roc_xgb <- roc(test_label, xgb_pred)
auc_xgb <- auc(roc_xgb)
print(auc_xgb)
plot(roc_xgb, main = "ROC Curve for XGBoost")


best_thresh_xgb <- as.numeric(coords(roc_xgb, "best", ret = "threshold", transpose =
FALSE))[1]
pred_class_xgb <- ifelse(xgb_pred >= best_thresh_xgb, 1, 0)
conf_matrix_xgb <- table(Predicted = pred_class_xgb, Actual = test_label)
print(conf_matrix_xgb)

# Precision, Recall, F1
TP_xgb <- conf_matrix_xgb["1", "1"]
FP_xgb <- conf_matrix_xgb["1", "0"]
FN_xgb <- conf_matrix_xgb["0", "1"]

precision_xgb <- TP_xgb / (TP_xgb + FP_xgb)
recall_xgb <- TP_xgb / (TP_xgb + FN_xgb)
f1_xgb <- 2 * precision_xgb * recall_xgb / (precision_xgb + recall_xgb)

precision_xgb; recall_xgb; f1_xgb


# Neural Network ----
# Data preparation
feature_cols <- c("wc_ta", "re_ta", "ebit_ta", "mve_bvlt", "sale_ta", "asset_growth",
                  "sales_growth",    "emp_growth",    "roe_change",    "pb_change",
"op_margin",
                  "log_at", "sic2_f")

# One-hot encode factor variables
train_x <- model.matrix(~ . - 1, data = train_data[, feature_cols])
test_x  <- model.matrix(~ . - 1, data = test_data[, feature_cols])


# Scale the features (standardization)
# Use training data to compute scaling parameters
train_x <- scale(train_x)
test_x  <- scale(test_x,
                 center = attr(train_x, "scaled:center"),
                 scale = attr(train_x, "scaled:scale"))

# Convert target to numeric binary vector
train_y <- as.numeric(as.character(train_data$bankrupt))
test_y  <- as.numeric(as.character(test_data$bankrupt))
```

```
# Function to create and train a neural network with given hyperparameters
train_nn_model <- function(units1, units2, dropout, learning_rate, batch_size) {
  model <- keras_model_sequential() %>%
    layer_dense(units = units1, activation = "relu", input_shape = ncol(train_x)) %>%
    layer_dropout(rate = dropout) %>%
    layer_dense(units = units2, activation = "relu") %>%
    layer_dense(units = 1, activation = "sigmoid")

  model %>% compile(
    loss = "binary_crossentropy",
    optimizer = optimizer_adam(learning_rate = learning_rate),
    metrics = c("accuracy")
  )

  history <- model %>% fit(
    x = train_x,
    y = train_y,
    validation_split = 0.2,
    epochs = 50,
    batch_size = batch_size,
    verbose = 0
  )

  preds <- as.vector(predict(model, x = test_x))
  auc <- pROC::auc(test_y, preds)
  return(list(auc = auc, model = model))
}

set.seed(123)
tensorflow::tf$random$set_seed(123)

# Define search space
search_space <- data.frame(
  units1 = sample(c(16, 32, 64, 128), 10, replace = TRUE),
  units2 = sample(c(8, 16, 32), 10, replace = TRUE),
  dropout = runif(10, 0.2, 0.5),
  learning_rate = runif(10, 0.0005, 0.01),
  batch_size = sample(c(32, 64, 128), 10, replace = TRUE)
)

# Store results
results <- list()
best_auc <- 0
best_model <- NULL
best_config <- NULL

for (i in 1:nrow(search_space)) {
  cat("Training model", i, "...\n")
  config <- search_space[i, ]
  result <- train_nn_model(config$units1, config$units2, config$dropout,
                           config$learning_rate, config$batch_size)
  cat("AUC:", round(result$auc, 4), "\n\n")

  results[[i]] <- list(config = config, auc = result$auc)

  if (result$auc > best_auc) {
    best_auc <- result$auc
    best_model <- result$model
    best_config <- config
  }
}

cat("Best AUC:", round(best_auc, 4), "\n")
print(best_config)

# Final model
```

```r
set.seed(123)
tensorflow::tf$random$set_seed(123)

nn_model <- keras_model_sequential() %>%
  layer_dense(units = best_config$units1, activation = "relu", input_shape =
ncol(train_x)) %>%
  layer_dropout(rate = best_config$dropout) %>%
  layer_dense(units = best_config$units2, activation = "relu") %>%
  layer_dense(units = 1, activation = "sigmoid")

nn_model %>% compile(
  loss = "binary_crossentropy",
  optimizer = optimizer_adam(learning_rate = best_config$learning_rate),
  metrics = c("accuracy")
)

history <- nn_model %>% fit(
  x = train_x,
  y = train_y,
  validation_split = 0.2,
  epochs = 50,
  batch_size = best_config$batch_size,
  verbose = 0)

# Evaluate Model
results <- nn_model %>% evaluate(test_x, test_y)

# Do Predictions
nn_probs <- as.vector(predict(nn_model, test_x))

# AUC - ROC
roc_nn <- roc(test_y, nn_probs)
auc_nn <- auc(roc_nn)
cat("AUC:", round(auc_nn, 4), "\n")
plot(roc_nn, main = "ROC Curve for Neural Network")

# Confusion matrix
best_thresh_nn <- as.numeric(coords(roc_nn, "best", ret = "threshold", transpose =
FALSE))
pred_class_nn <- ifelse(nn_probs >= best_thresh_nn, 1, 0)
conf_matrix_nn <- table(Predicted = pred_class_nn, Actual = test_data$bankrupt)
print(conf_matrix_nn)

# Precision, recall, F1-score
TP_nn <- conf_matrix_nn["1", "1"]
FP_nn <- conf_matrix_nn["1", "0"]
FN_nn <- conf_matrix_nn["0", "1"]

precision_nn <- TP_nn / (TP_nn + FP_nn)
recall_nn <- TP_nn / (TP_nn + FN_nn)
f1_nn <- 2 * (precision_nn * recall_nn) / (precision_nn + recall_nn)

precision_nn; recall_nn; f1_nn


# Aggregated ROC curves ----
roc_lr  <- roc(test_y, lr_pred)
roc_rf  <- roc(test_y, rf_probs)
roc_xgb <- roc(test_y, xgb_pred)
roc_nn  <- roc(test_y, nn_probs)

roc_df <- rbind(
  data.frame(model = "Logistic Regression",
             specificity = 1 - roc_lr$specificities,
             sensitivity = roc_lr$sensitivities),
  data.frame(model = "Random Forest",
             specificity = 1 - roc_rf$specificities,
             sensitivity = roc_rf$sensitivities),
```

```
  data.frame(model = "XGBoost",
             specificity = 1 - roc_xgb$specificities,
             sensitivity = roc_xgb$sensitivities),
  data.frame(model = "Neural Network",
             specificity = 1 - roc_nn$specificities,
             sensitivity = roc_nn$sensitivities)
)

# Plot all ROC curves on the same graph
ggplot(roc_df, aes(x = specificity, y = sensitivity, color = model)) +
  geom_line() +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "gray50") +
  labs(title = "ROC Curves – Firm-Level Model",
       x = "False Positive Rate (1 - Specificity)",
       y = "True Positive Rate (Sensitivity)",
       color = "Model") +
  theme_minimal()

# 5. MACROECONOMIC STRESS TESTING ----------------------------------------------

model_data_stress <- model_data_macro

# Train - test split
train_data_stress <- model_data_stress %>%
  filter(fyear < 2018)

test_data_stress <- model_data_stress %>%
  filter(fyear >= 2018 & fyear <= 2019)

train_data_stress %>%
  group_by(bankrupt_firm) %>%
  summarise(firm_count = n_distinct(gvkey)) %>%
  ungroup() %>%
  print()
test_data_stress %>%
  group_by(bankrupt_firm) %>%
  summarise(firm_count = n_distinct(gvkey)) %>%
  ungroup() %>%
  print()

# Scale the macro variables
macro_vars <- c("interest_rate", "gdp_growth", "unemployment", "vix")
macro_means <- sapply(train_data_stress[, macro_vars], mean, na.rm = TRUE)
macro_sds   <- sapply(train_data_stress[, macro_vars], sd, na.rm = TRUE)

# Scale macro variables in train set
train_data_stress <- train_data_stress %>%
  mutate(
    interest_rate    =    (interest_rate    -    macro_means["interest_rate"])    /
macro_sds["interest_rate"],
    gdp_growth        =  (gdp_growth         -  macro_means["gdp_growth"])         /
macro_sds["gdp_growth"],
    unemployment      =   (unemployment      -   macro_means["unemployment"])      /
macro_sds["unemployment"],
    vix           = (vix           - macro_means["vix"])           / macro_sds["vix"]
  )

# Scale macro variables in test set using the same values
test_data_stress <- test_data_stress %>%
  mutate(
    interest_rate    =    (interest_rate    -    macro_means["interest_rate"])    /
macro_sds["interest_rate"],
    gdp_growth        =  (gdp_growth         -  macro_means["gdp_growth"])         /
macro_sds["gdp_growth"],
    unemployment      =   (unemployment      -   macro_means["unemployment"])      /
macro_sds["unemployment"],
    vix           = (vix           - macro_means["vix"])           / macro_sds["vix"]
  )
```

```
# Recreate a clean version of sic2_f with only observed levels in train
train_data_stress$sic2_f  <-  factor(as.character(train_data_stress$sic2),  levels  =
unique(as.character(train_data_stress$sic2)))
test_data_stress$sic2_f  <-  factor(as.character(test_data_stress$sic2),  levels  =
levels(train_data_stress$sic2_f))
test_data_stress <- test_data_stress %>% filter(!is.na(sic2_f))

# Baseline model - stress testing ----
stress_model <- glm(bankrupt ~ wc_ta + re_ta + ebit_ta + mve_bvlt + sale_ta +
                     asset_growth + sales_growth + emp_growth +roe_change +
                     pb_change + op_margin + log_at + sic2_f +
                     interest_rate + gdp_growth + unemployment + vix,
                   data = train_data_stress, family = "binomial")

pred_base <- predict(stress_model, newdata = test_data_stress, type = "response")
summary(stress_model)


# Define crisis periods
crisis_periods <- list(
  dotcom = c(2000, 2001),
  gfc    = c(2007, 2008, 2009),
  covid  = c(2020, 2021)
)

# Get average macro values for each crisis
macro_crisis_values <- lapply(crisis_periods, function(years) {
  model_data_macro %>%
    filter(fyear %in% years) %>%
    summarise(across(all_of(macro_vars), ~ mean(.x, na.rm = TRUE))) %>%
    as.list()
})

# Scale macro crisis values using training means and SDs
macro_crisis_scaled <- lapply(macro_crisis_values, function(values) {
  mapply(function(value, var) {
    (value - macro_means[[var]]) / macro_sds[[var]]
  }, values, names(values))
})

# Dotcom crisis ----
stress_dotcom <- test_data_stress %>%
  mutate(interest_rate = macro_crisis_scaled$dotcom["interest_rate"],
         gdp_growth    = macro_crisis_scaled$dotcom["gdp_growth"],
         unemployment  = macro_crisis_scaled$dotcom["unemployment"],
         vix           = macro_crisis_scaled$dotcom["vix"])

pred_dotcom <- predict(stress_model, newdata = stress_dotcom, type = "response")

# 2008 crisis ----
stress_2008 <- test_data_stress %>%
  mutate(interest_rate = macro_crisis_scaled$gfc["interest_rate"],
         gdp_growth    = macro_crisis_scaled$gfc["gdp_growth"],
         unemployment  = macro_crisis_scaled$gfc["unemployment"],
         vix           = macro_crisis_scaled$gfc["vix"])

pred_2008 <- predict(stress_model, newdata = stress_2008, type = "response")

# Covid crisis ----
stress_covid <- test_data_stress %>%
  mutate(interest_rate = macro_crisis_scaled$covid["interest_rate"],
         gdp_growth    = macro_crisis_scaled$covid["gdp_growth"],
         unemployment  = macro_crisis_scaled$covid["unemployment"],
         vix           = macro_crisis_scaled$covid["vix"])

pred_covid <- predict(stress_model, newdata = stress_covid, type = "response")
```

```
# Compare the models ----
# Get the threshold
roc_stress <- roc(test_data_stress$bankrupt, pred_base)
opt_threshold <- as.numeric(coords(roc_stress, "best", ret = "threshold", transpose
= FALSE))[[1]]
opt_threshold

compare_rates <- tibble(
  scenario = c("Baseline", "Dotcom", "2008", "COVID"),
  avg_risk = c(
    mean(pred_base),
    mean(pred_dotcom),
    mean(pred_2008),
    mean(pred_covid)),
  firms_at_risk = c(
    sum(pred_base > opt_threshold),
    sum(pred_dotcom > opt_threshold),
    sum(pred_2008 > opt_threshold),
    sum(pred_covid > opt_threshold))
)

print(compare_rates)

# Plot the results
# Combine all predictions into one dataframe
risk_df <- bind_rows(
  tibble(risk = pred_base,  scenario = "Baseline"),
  tibble(risk = pred_dotcom, scenario = "Dotcom"),
  tibble(risk = pred_2008,   scenario = "2008"),
  tibble(risk = pred_covid,  scenario = "COVID")
)

# Plot
ggplot(risk_df, aes(x = risk, fill = scenario)) +
  geom_density(alpha = 0.4) +
  labs(
    title = "Predicted Bankruptcy Risk Across Stress Scenarios",
    x = "Predicted Risk",
    y = "Density"
  ) +
  theme_minimal()
```