A Comparative Analysis of Distress Prediction in Public Firms:

The Impact of Macroeconomic Conditions, Industry Dynamics, and Nonlinear Modeling

By

Márton Molnár

Submitted to Central European University
Department of Undergraduate Studies

In partial fulfillment of the requirements for the degree of
Bachelor of Arts / Bachelor of Science in Data Science and Society

Supervisor: Professor Gábor Békés

Vienna, Austria
2025

Abstract

Reliable prediction of corporate financial distress is critical for effective risk management and maintaining economic stability, especially given the systemic vulnerabilities revealed by the 2008 financial crisis. Traditional models, that rely on accounting ratios, are limited by linear assumptions and poor adaptability across varying economic conditions. This thesis compares classical econometric models (logistic regression) with advanced machine learning algorithms (extreme gradient boosting and shallow neural networks) under a standardized framework to address three research questions: Do macroeconomic indicators meaningfully improve predictive accuracy beyond accounting metrics alone? Do sector-specific models outperform general cross-industry models? Do advanced machine learning models offer sufficiently better performance to justify their complexity? To answer these questions, a novel binary distress indicator is defined that reflects significant financial underperformance rather than legal insolvency and is scaled to align with historical large-cap bankruptcy rates. This label is then applied to a panel dataset of 490 listed U.S. firms (2015-2024) in sectors with historically higher bankruptcy rates. Results show that XGBoost significantly outperforms other models, achieving accuracy over 97% and precision-recall (PR-AUC) between 65-69% during expansionary periods, with slightly reduced but robust performance in volatile conditions (94% accuracy; 55% PR-AUC). While macroeconomic and sector-specific variables occasionally improved predictive accuracy, they generally introduced unnecessary complexity and reduced interpretability without significant practical advantages. Practitioners are recommended to prioritize flexible machine learning models, like XGBoost, relying on core financial ratios, and periodically retrain these models to ensure consistent accuracy in changing economic conditions. Key limitations, including moderate dataset size, class imbalance, and indirect validation of distress cases, suggest ways for further research, such as incorporating external distress validation or exploring recurrent neural architectures.

Author's Declaration

I, the undersigned, **Márton Molnár** candidate for the **Bachelor of Arts / Bachelor of Science degree in Data Science and Society** declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 25 May 2025

Márton Molnár
Signature

Acknowledgments

I would like to thank my supervisor, professor Gábor Békés, for his guidance and feedback throughout the development of this thesis. I also acknowledge the use of AI tools, such as OpenAI's ChatGPT (version 4.0), which were used for grammar correction, troubleshooting code-related issues, and formatting tables and figures. All content and analysis presented in this thesis were written and conducted by me.

Table of Contents

List of Figures

List of Tables

List of Abbreviation

API	Application Programming Interface
CAPM	Capital Asset Pricing Model
DL	Deep Learning
EBIT	Earnings Before Interest and Taxes
EBITDA	Earnings Before Interest, Taxes, Depreciation, and Amortization
EV	Enterprise Value
FD	Financial Distress
FDP	Financial Distress Prediction
FRED	Federal Reserve Economic Data
IPO	Initial Public Offering
LASSO	Least Absolute Shrinkage and Selection Operator
LSEG	London Stock Exchange Group
Logit	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NN	(Shallow Feedforward) Neural Network
OTC	Over-the-Counter (Markets)
PR-AUC	Area Under the Precision-Recall Curve
RIDGE	Ridge Regression
ROA	Return on Assets
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
SEC	Securities and Exchange Commission
SHAP	SHapley Additive exPlanations
WS	LSEG Workspace Software
XGBoost	Extreme Gradient Boosting

1. Introduction

Corporate financial distress (FD) has been an integral part of credit risk management and economic policy for decades. The failure of a single firm can cascade through the market, disrupting supply chains and regulatory frameworks. The 2008 financial crisis exposed how undetected corporate weakness can contribute to systemic risk, which incited renewed efforts to create reliable early-warning systems to identify the signs of FD. Continuous advancements in computing techniques and the rise of artificial intelligence resulted in more studies applying machine learning (ML)-based models to forecast and analyze distress events, suggesting an increased demand for tools that can process, high-dimensional and nonlinear data. (Zhao et al. 2024)

Early bankruptcy prediction models were centered on accounting ratios and used simple statistical methods to flag financially vulnerable firms. These frameworks demonstrated that measures of liquidity, profitability, and solvency could often distinguish between healthy and bankrupt firms, years before any legal filings. However, they relied on linear assumptions, oversimplified firm dynamics, and often used datasets with unrealistic proportions of bankrupt firms. As a result, they struggled to generalize across sectors and time periods, especially in volatile and unusual economic environments. This prompted the search for more flexible and robust models capable of capturing intricate firm dynamics and generalizing well on unseen data.

Later research improved these ratio-based models by integrating equity-market and macroeconomic variables, which yielded consistent improvements in predictive accuracy (Hernandez Tinoco and Wilson 2013). FD typically evolves over months or years and is influenced by business-cycle fluctuations, so models that incorporate temporal dynamics instead of a single cross-section produce more reliable forecasts across both expansionary and recessionary periods (Almamy et al. 2016; Prastyo et al. 2023). Adopting time-series methods can better capture downturn risk that static models overlook, but it also increases data requirements and modeling complexity (Liu et al. 2023).

Over the last decade, ML and deep learning (DL) techniques have begun to outperform traditional methods in financial distress prediction (FDP) by modeling complex, nonlinear relationships with ease. However, their practical adoption is hindered by concerns over their interpretability and computational needs. Moreover, there is no consensus on which features consistently deliver the greatest marginal benefit across modeling setups. For example, Hernandez Tinoco and Wilson (2013) showed that sectoral controls based on SIC[1] codes improved predictive accuracy when applied to listed firms in the U.K., allowing the model to capture industry-specific dynamics not reflected in accounting ratios alone. Similarly, Alaka et al. (2018) found that tailoring input variables to sector-specific contexts (e.g., services vs. manufacturing) led to lower misclassification rates, suggesting that feature relevance varies significantly across industries.

This study therefore aims to address three core questions: Do macroeconomic indicators significantly improve predictive performance over purely accounting-based models? Are sector-adjusted models superior to generalized, cross-industry frameworks? Do advanced machine learning algorithms that capture nonlinear and high-dimensional patterns offer sufficiently better performance to justify their added complexity and computational demands?

To answer these questions, a standardized empirical analysis is performed on a comprehensive multi-sector firm-year panel that incorporates both financial ratios and macroeconomic

[1] SIC (Standard Industrial Classification) codes are four-digit numerical codes used by U.S. and U.K. government agencies to classify industries by their primary business activities.

features. The target variable is a binary distress indicator, defined based on a detailed set of accounting criteria and benchmarked against established methods like Altman's Z-score. Three modeling approaches are applied: logistic regression (Logit), extreme gradient boosting (XGBoost), and a shallow feedforward neural network (NN), each evaluated using identical performance metrics. This approach quantifies the added value of macroeconomic variables while weighing the trade-offs between predictive accuracy, computational complexity, generalizability, and interpretability.

This research plans to deliver three main contributions. First, it measures how much macroeconomic and sector-specific factors improve distress prediction across different time horizons, to help guide feature selection. Second, it establishes a benchmarking framework that compares classical econometric and modern ML models under homogenous conditions. Third, it offers practical recommendations for risk managers and regulators on when to adapt generalized versus specialized models, considering resource constraints and performance needs.

The remainder of the paper is organized as follows: Section 2 presents a detailed literature review, tracing the evolution of traditional bankruptcy models (Section 2.1) and surveying ML applications (Section 2.2). Section 3 describes the data sources, preprocessing steps, and feature engineering, including the construction of the target variable. Section 4 outlines the modeling framework and evaluation methods. Section 5 reports the empirical results, covering classification performance, calibration accuracy, and feature importance analysis. Section 6 contextualizes the results, highlights limitations, and discusses implications for financial practitioners. Section 7 summarizes key findings and recommendations. A detailed description and technical details of the data collection and analysis can be found in the Appendix.

2. Background & Related Work

2.1 Evolution of Bankruptcy Modeling

Traditional bankruptcy models focused on detecting ratio-based signals through simple statistical techniques. Beaver (1966) introduced a univariate ratio analysis, which showed that certain liquidity and leverage metrics could meaningfully distinguish between healthy and bankrupt firms several years before any legal filings. However, this method assessed each ratio independently and used matched samples of bankrupt and non-bankrupt firms, ignoring potential interdependencies among features. Altman (1968) addressed these limitations in his Z-score model by applying five financial ratios in a multivariate discriminant analysis and aggregating them into a continuous index with fixed thresholds. While the Z-score achieved high in-sample accuracy, it was developed using a relatively small sample of manufacturing firms. Its reliance on normality and linearity assumptions also limits its generalizability. Ohlson (1980) proposed a similar model (Ohlson's O) but used logistic regression as a more robust alternative that relaxes these assumptions and returns directly interpretable probability estimates. Zmijewski (1984) later introduced an alternative specification (Zmijewski's X) based on a probit[2] model and emphasized how choice-based sampling and nonrandom selection bias[3] can distort parameter estimates and reduce out-of-sample reliability.

Recognizing that internal accounting measures alone fail to capture broader risk exposure, subsequent research expanded the predictor space beyond static financial ratios by incorporating temporal dynamics, as well as market and macroeconomic indicators. Hernandez

---

[2] Probit models are a type of regression used for binary outcomes, where the probability of an event is modeled using the cumulative distribution function of the standard normal distribution.

[3] Choice-based sampling means selecting equal number of bankrupt and healthy firms irrespective of their true population proportions, while nonrandom selection refers to systematically excluding firms with incomplete data; both practices make the sample unrepresentative of the population.

Tinoco and Wilson (2013) showed that combining accounting data with market-based and macroeconomic proxies consistently improves predictive accuracy. Since bankruptcy is typically preceded by long-term financial deterioration, several studies propose modeling distress as a time series, rather than treating it as cross-sectional (Liu et al. 2023; Prastyo et al. 2023). This is also reflected in the fact that distress patterns often follow business-cycle fluctuations, highlighting the importance of capturing macroeconomic effects over time (Alaka et al. 2018). Sector-specific calibration has also been shown to improve model performance, as different industries have distinct risk factors and regulatory constraints that affect distress dynamics and feature relevance (Kim et al. 2022; Alaka et al. 2018). These developments suggest that FD is shaped by a mix of firm-level, macroeconomic, and industry-specific dynamics that interact across time.

2.2 Machine Learning in Distress Prediction

The availability of high-dimensional financial data and the constraints of classical models have encouraged the adoption of ML algorithms. Tree-based classifiers, notably Random Forest and XGBoost, consistently outperform Logit models in both sensitivity and ROC-AUC (Barboza et al. 2017; Alaka et al. 2018; Huang and Yen 2019; Alam et al. 2021; Kim et al. 2022; Liu et al. 2023). While their ability to capture nonlinear relationships makes them well-suited for distress prediction, ensemble models are sensitive to overfitting[4], particularly with noisy and collinear[5] data. To address this, recent studies underline the importance of methods like regularization and cross-validation to improve generalizability (du Jardin 2009; Alaka et al. 2018; Zhao et al. 2024). Neural networks have also been utilized more frequently, including shallow

[4] Overfitting occurs when a model learns patterns specific to the training data, including noise, resulting in poor performance on new, unseen data.
[5] Collinearity refers to when two or more predictor variables are highly correlated, which can distort coefficient estimates and reduce model interpretability.

feedforward[6] models and more advanced architectures like LSTMs[7]. Although they excel in modeling complex temporal dynamics, their susceptivity to hyperparameters and "black-box" nature raises concerns in terms of reproducibility and interpretability, delaying their large-scale practical application (du Jardin 2009; Alaka et al. 2018; Huang and Yen 2019; Kim et al. 2022; Liu et al. 2023). Still, under the right circumstances, DL models can outperform simple econometric models, especially when temporal effects and latent interactions are prominent.

Feature selection is an integral part of distress modeling. Combining statistical filtering with domain knowledge generally improves out-of-sample performance (du Jardin 2009). As data availability expanded, automated feature selection methods, such as filters and wrappers[8], gained popularity (Zhao et al. 2024). Regularization-based models, like LASSO or RIDGE, shrink irrelevant coefficients, addressing both collinearity and overfitting. Tree-based models, like Random Forest, offer embedded mechanisms for ranking feature importance based on their contribution to reducing prediction error, allowing for efficient variable selection in large datasets. Several studies also suggest using resampling[9] techniques to mitigate class imbalance inherent in bankruptcy data, reduce Type II errors (failing to detect distress), and improve overall calibration (Alaka et al. 2018; Huang and Yen 2019; Alam et al. 2021; Zhao et al. 2024).

In FDP, the model evaluation must account for the rarity of distress cases and the asymmetric cost of errors. As a result, metrics beyond standard accuracy are preferred. ROC-AUC is used to measure overall discriminatory performance (Kim et al. 2022), while precision-recall curves (PR-AUC) are more sensitive to class imbalance (Alaka et al. 2018; Rahman and Zhu 2024).

[6] A shallow feedforward neural network (FNN) is a neural network with one or two hidden layers where data flows in one direction, from input to output, without feedback loops.
[7] Long Short-Term Memory (LSTM) networks are recurrent neural architectures with gated memory units that capture long-range temporal dependencies in sequential data.
[8] Filters and wrappers are feature selection methods: filter methods rank features based on statistical properties, while wrapper methods evaluate subsets using model performance as a criterion.
[9] Resampling techniques balance classes by either oversampling the minority (e.g., duplicating observations or generating synthetic samples via SMOTE) or undersampling the majority class, which reduces bias toward the dominant group and enhances the model's ability to detect rare events.

In this context, recall is often prioritized over precision, as failing to identify distress poses a greater financial and regulatory risk (Barboza et al. 2017). Brier scores[10] and calibration curves are used to assess how well predicted probabilities reflect actual outcomes (Békés and Kézdi 2021). For the interpretability of ML models, SHAP[11] values are useful for explaining feature importance (Liu et al. 2023). Additionally, evaluating model performance across decile rankings helps determine whether high-risk firms are identified early, which is especially useful when audit resources are limited (Qian et al. 2022).

Recent literature shows a clear trend toward integrating macroeconomic, market-based, and sector-specific features into distress modeling. However, few studies (Hernandez Tinoco and Wilson 2013; Kim et al. 2022) evaluate these components together under a standardized framework. Most papers assess macroeconomic and sectoral effects separately and use inconsistent evaluation metrics. Moreover, the trade-offs between generalizability and domain-specific tuning remain underexplored, despite their importance for practical applications. This study addresses these gaps by systematically comparing traditional econometric and ML models using a unified framework, testing the added value of macro- and sector-level effects, and prioritizing calibration, interpretability, and sensitivity to Type II errors.

[10] Brier score is the mean squared error between predicted probabilities and actual binary outcomes.
[11] SHAP (SHapley Additive exPlanations) values derive from cooperative game theory to attribute each feature's contribution to an individual prediction, offering consistent, local insights into model behavior.

3. Data Sources & Preprocessing

3.1 Data Sources

For this analysis, a panel dataset was constructed where each row corresponds to a specific firm-year observation spanning 2015 to 2024. The data combines firm-level financial information retrieved from the LSEG Workspace application programming interface (WS API) with macroeconomic indicators from the Federal Reserve Bank of St. Louis.

The companies were screened using CIQ Pro ("S&P Capital IQ Pro" 2024) based on the following criteria: operating, U.S.-incorporated firms that had completed their IPO before December 2014 and are listed on a major exchange[12] (excluding OTC markets). To focus on industries with historically higher bankruptcy rates (Cornerstone Research 2023; 2024), only firms from the financials, real estate, industrials, and materials[13] sectors were included. The final selection consists of 490 firms: *250 financials, 107 industrials, 106 real estate, and 27 materials*.

The primary source for financial data was annual 10-K filings retrieved through the WS API ("LSEG Workspace" 2023). Various items were collected from the income statement, balance sheet, cash flow, and valuation tables, and used to derive financial ratios addressing profitability, liquidity, solvency, and other dimensions of firm performance. In addition to accounting data, daily close prices were used to estimate each firm's annual market β and idiosyncratic risk under standard CAPM assumptions. Together, these features measure the

[12] Major exchanges in this context include: NASDAQ, NYSE, TMX, ASX, LSE, TXSE

[13] Based on S&P's industry classification.

internal financial health and external market exposure of a firm. A detailed list of variables is provided in Appendix A.

Macroeconomic indicators from FRED (Federal Reserve Bank of St. Louis 1997) were incorporated to reflect broader business-cycle dynamics and contextualize firm-level risk. Alongside general features, such as GDP, inflation, and interest rates, additional series were chosen to reflect sector-specific conditions. For instance, real estate firms are affected by construction spending and mortgage rates, while materials are linked to commodity prices. All macro series were aggregated to the calendar-year level and merged with the firm-level panel, effectively acting as year-fixed effects in the model.

3.2 Data Cleaning & Feature Engineering

After assembling the dataset, it underwent a structured cleaning and transformation process to address missing values, avoid data leakage, and ensure feature comparability. The dataset includes two main types of variables: firm-level financial and market indicators that vary by company and year, and macroeconomic variables that are consistent across firms within a given year.

The first step was addressing missing financial statement items. When feasible, missing values were reconstructed using accounting identities from the WS API documentation. Where complete reconstruction was not possible, proxy estimates were calculated using domain knowledge. Manual corrections were made in cases where data inconsistencies arose from API import errors, based on values from the SEC filings. For a handful of firms, regression imputation was applied to fill in the remaining gaps. Values missing due to structural irrelevance, such as interest expense for firms with no recorded debt, were treated as zeros. These cases were flagged with a binary indicator to distinguish between true zero values and imputed zeros, ensuring that models could account for their informational significance during training.

After cleaning the financial statement data, more than 40 financial ratios were derived, including quick ratio for short-term liquidity and ROA for profitability, among others. Ratios with denominators equal to zero were set to zero and flagged to indicate non-applicability, consistent with the treatment of financial statement items. This included, for example, the quick ratio for firms with no current liabilities and the cash flow coverage ratio for firms with no debt obligations. In addition to accounting ratios, three common distress scores were calculated for benchmarking the target variables (see Section 3.3).

There were no missing values among the macroeconomic features; however, where appropriate they were transformed to year-over-year percentage changes, to make them comparable to the financial ratios in scale. Macro indicators were also used to construct interaction terms with accounting ratios, like "*leverage ratio adjusted by the effective federal funds rate"* to capture financing risk under monetary contraction. A detailed list of the derived ratios and macroeconomic features is available in Appendices A.2 and A.3.

All approximations, transformations, and calculations were made using data from the same year or earlier. In the rare cases where interpolation was applied, the imputed values were not directly used during the analysis. As a result, each observation is represented by a consistent vector of firm- and macro-level features available at the time of prediction. The final dataset contains 5390 rows and 81 candidate variables. A comprehensive description of the cleaning process can be found in Appendix B.

3.3 Defining Financial Distress

In this study, financial distress is defined not as formally filing for bankruptcy, but as a state of significant operational and financial underperformance. Since the dataset includes only publicly listed, operating firms, real bankruptcy is unlikely to occur. However, these firms are not immune to distress. Instead, the aim is to identify sub-optimal performance, that if left unaddressed, may escalate into legal insolvency. This is captured using a binary target, constructed based on a set of 15 accounting and sector-adjusted criteria reflecting weaknesses in profitability, liquidity, and leverage (criteria listed in Appendix C.1).

Each criterion is weighed equally, and observations are labeled as distressed if they score at least 9 out of 15. This threshold was chosen to align the distribution of persistent distress, defined as three or more consecutive years, with the historical bankruptcy rates for U.S. public firms. Specifically, 1.02% of firms in the sample exhibit at least four consecutive years of distress, closely matching the estimated 1-1.5% annual bankruptcy rate for large-cap public firms (Cornerstone Research 2024; U.S. Courts 2025). Extending the window to a minimum of three years raises this share to 2.86%, a reasonable amount given that it captures firms whose distress likely came exclusively from the abnormal conditions of the COVID-19 period. Since the selected sectors historically experience higher default rates, a higher proportion is expected.

Overall, the dataset includes 210 distressed observations, accounting for approximately 3.9%, confirming the presence of class imbalance. While a single year of distress may reflect transitory shocks, extended periods of distress are more indicative of systemic weakness. This labeling approach offers a meaningful, forward-looking proxy for financial deterioration, even in the absence of formal bankruptcy filings.

To validate the labeling, the binary target was benchmarked against three established scoring models: Altman's Z-score, Ohlson's O-score, and Zmijewski's X-score. Each of these models outputs a distress classification based on its own threshold logic, and within our dataset, they classify more than half of the observations as distressed. This confirms a tendency toward overclassification, likely due to the differences in the sample period and sectoral characteristics. The binary distress indicator, by contrast, is more conservative and designed to reflect more realistic financial strain in a post-2008 and COVID economy. Agreement analysis (Appendix C.2) showed that in cases of disagreement, the scores unanimously flagged distress while the binary distress indicator did not, highlighting their lower specificity in this setting. This custom target variable, therefore, balances theoretical foundations with empirical calibration and practical relevance for forecasting distress in currently operating, listed firms.

4. Modeling Framework

4.1 Experimental Design

This study adopts three distinct models, representing different levels of complexity and practicality. Logit is employed first because its coefficients are easily interpretable, and the method is well documented in the literature. Next, XGBoost was selected as tree-based ensemble models are capable of handling nonlinear relationships better than regressions. Lastly, a shallow feedforward neural network was included to represent deep learning techniques. Neural networks can capture more abstract and high-dimensional relationships, like temporal dependencies, which econometric and tree-based models might struggle with.

Each model follows a standardized design to ensure comparability across all configurations. The dataset is divided into two non-overlapping five-year windows (2015-2019 and 2020-2024), the first reflecting a relatively stable macroeconomic environment and the second capturing structural shocks, such as the COVID-19 pandemic and rising interest rates. Within each window, models are trained on the first two years and then tested on the subsequent three. Three different feature sets are considered: financial ratios only; financial and macroeconomic indicators; financial, macroeconomic, and sector-specific variables. The combination of three algorithms and three feature sets produces nine distinct model setups. Each setup is then evaluated on both windows, resulting in eighteen models total, which allows direct comparison across algorithms, feature groups, and economic conditions.

4.2 Feature Selection

The feature selection uses a two-stage process. First, a Random Forest classifier provides an initial ranking of features. Second, a Variance Inflation Factor analysis removes collinear variables. This process is applied separately for financial ratios only, and financial ratios considering firm-macro interaction terms. In each case, the ten most informative and mutually independent features are retained. Further details are in Appendix D.1.

All three algorithms incorporated L1 (LASSO) regularization during training to remove redundant predictors and avoid overfitting, particularly useful when macroeconomic series introduce high collinearity. Since each training window spans only two years, macro variables exhibit limited variation, making conventional feature selection methods ineffective. L1 regularization allows the model to discard macro features that offer little or misleading signals. Furthermore, a single missingness indicator was added to each feature set, counting the number of non-applicable financial ratios per observation. This helps mitigate nonrandom selection bias by allowing the models to differentiate between valid zero values and undefined ratios.

4.3 Model Training

Models are trained with 5-fold stratified cross-validation to preserve the proportion of distressed observations. Within each fold, the minority class is randomly oversampled to mitigate imbalance, and features are standardized to prevent look-ahead bias[14]. After cross-validation, each model is refit on the full training window and evaluated on the corresponding test window. Decision thresholds are then tuned to minimize the Brier score through greedy search[15]. The optimal threshold allows for a more accurate separation between distressed and non-distressed firms, especially in disproportionate datasets.

[14] Look-ahead bias, a form of data leakage, occurs when information unavailable at the time of prediction is inadvertently used during model training, leading to overly optimistic estimates.
[15] Greedy search refers to an iterative method that evaluates a fixed set of candidate thresholds and selects the one that yields the best performance according to the chosen loss function.

Model performance is evaluated using several metrics: *accuracy* measures the ratio of correct classifications; *precision* indicates the proportion of flagged firms that are truly distressed; *recall* captures the proportion of distressed firms the model successfully detects; *specificity* reflects the rate at which healthy firms are correctly identified; the *F_1-score* balances precision and recall into a single indicator; *ROC-AUC* assesses the model's ability to rank distressed against healthy firms across all thresholds; and *PR-AUC* summarizes the precision-recall trade-off, which is especially informative in imbalanced datasets. Recall is prioritized to minimize Type II errors, because missing a distressed firm carries a higher cost than flagging healthy ones, while ROC-AUC and PR-AUC capture the overall discriminatory power and robustness to class imbalance. Probability calibration is assessed using the Brier score, which measures the mean squared difference between predicted and actual probabilities, and calibration curves, which visualize how well predicted probabilities align with observed frequencies.

Model interpretability is evaluated using standardized coefficients and p-values for Logit, built-in feature importance scores for XGBoost, and SHAP values for both XGBoost and NN. Further details about the methodology are provided in Appendix D.2.

5. Results

5.1 Classification Performance & Calibration

Tables I and II compare the performance of the different algorithms and feature sets, for the M1 (2017-2019) and M2 (2022-2024) evaluation windows respectively.

In M1, XGBoost models have the best performance in all three setups with accuracy above 0.97 and PR-AUC in the 0.65-0.69 range, indicating near-perfect classifications and an average precision of 67% across all recall levels, meaning that two-thirds of the firms flagged as distressed are true positives, which is well above the 2.3% baseline[16] for this window. The F_1-scores above 0.60 suggest a healthy balance between recall (0.65-0.71) and precision (0.55-0.56), while the specificity of 0.98-0.99 shows that they rarely misclassify healthy firms. NN models match XGBoost on accuracy and specificity but collapse on PR-AUC (0.12-0.35) revealing that they struggle with correctly identifying distressed firms. They seem to prioritize

Model	Setup	Threshold	Accuracy	Precision	Recall	Specificity	F1	ROC_AUC	PR_AUC	Brier
Logit	Financial	0.491	0.8952	0.1304	0.6000	0.9024	0.2143	0.7635	0.2356	0.0928
	Macro	0.279	0.8340	0.0996	0.7429	0.8362	0.1757	0.8038	0.2595	0.0848
	Sector	0.461	0.4680	0.0349	0.8000	0.4599	0.0668	0.7548	0.2153	0.4510
XGBoost	Financial	0.088	0.9796	0.5610	0.6571	0.9875	0.6053	0.9732	0.6523	0.0126
	Macro	0.079	0.9796	0.5556	0.7143	0.9861	0.6250	0.9723	0.6597	0.0130
	Sector	0.069	0.9796	0.5556	0.7143	0.9861	0.6250	0.9731	0.6888	0.0123
NeuralNet	Financial	0.895	0.9741	0.4516	0.4000	0.9882	0.4242	0.7793	0.3542	0.0284
	Macro	0.898	0.9741	0.2000	0.0286	0.9972	0.0500	0.6531	0.1242	0.0258
	Sector	0.893	0.9776	0.6667	0.1143	0.9986	0.1951	0.7101	0.2247	0.0225

Table I *Out-of-sample evaluation metrics for the M1 window (2017–2019 test period).*
Each model is trained on data from 2015–2016 using one of three feature sets: financial ratios, macroeconomic indicators, or sector-specific variables. Metrics reported include the optimal classification threshold (minimizing the Brier score), accuracy, precision, recall, specificity, F1-score, ROC-AUC, PR-AUC, and Brier score.
Source*: df_NN dataset.*

precision across feature sets with the financial setup having moderate detection (recall: 0.40; precision: 0.45), the macro setup misses distress entirely (recall: 0.03; precision 0.20), whereas

[16] The PR-AUC baseline equals to the positive-class presence: in M1 only 2.3% of the observations are distressed, so a random model would achieve 0.023 precision on average; in M2 the distress rate is 7%, yielding a 0.07 baseline.

the sector model catches only 11.4% of distress but at high precision (0.67). Logit, on the other hand, sacrifices accuracy (0.89, 0.83, 0.46 for financial, macro, sector) for higher recall (0.60, 0.74, 0.80), yet has lower PR-AUC (0.22-0.25), underlining that it finds more distressed firms only by generating many false alarms (precision at 0.03-0.13).

In M2, under COVID-era volatility, XGBoost remains the most reliable model with 93% accuracy, 96% specificity, and PR-AUC, precision, recall, and F_1-scores all clustered between 0.50-0.60. Logit models perform more consistently in this window. Accuracy and PR-AUC are around 0.80-0.84 and 0.24-0.35, while specificity and F_1-scores are near 0.79-0.81 and 0.32-0.38 respectively. However, they still lean toward recall (0.60-0.67) at the expense of precision (0.21-0.27). This shows that Logit behaves similarly to XGBoost but with slightly worse classification accuracy. NN again performs well only in the financial setup (recall: 0.61 and PR-AUC: 0.45) but collapses when macro and sector effects are added (PR-AUC: 0.16), catching essentially no distress (recall: 0.01) even if precision strikes to 1.0 (F_1-score: 0.02). Nevertheless, the financial NN is an exception, outperforming Logit on all metrics except recall.

Model	Setup	Threshold	Accuracy	Precision	Recall	Specificity	F1	ROC_AUC	PR_AUC	Brier
Logit	Financial	0.514	0.8483	0.2692	0.6796	0.8610	0.3857	0.8446	0.2808	0.1362
	Macro	0.513	0.8007	0.2376	0.8350	0.7981	0.3699	0.8483	0.3152	0.1517
	Sector	0.496	0.8068	0.2127	0.6505	0.8186	0.3206	0.7999	0.2465	0.1371
XGBoost	Financial	0.181	0.9388	0.5586	0.6019	0.9642	0.5794	0.9432	0.5574	0.0512
	Macro	0.188	0.9340	0.5268	0.5728	0.9612	0.5488	0.9426	0.5646	0.0513
	Sector	0.273	0.9361	0.5398	0.5922	0.9620	0.5648	0.9439	0.5565	0.0504
NeuralNet	Financial	0.637	0.9027	0.3795	0.6117	0.9247	0.4684	0.8733	0.4580	0.0792
	Macro	0.712	0.9306	1.0000	0.0097	0.0192	0.5680	0.1572	0.0692	0.0692
	Sector	0.532	0.9299	0.5000	0.0097	0.9993	0.0190	0.6586	0.1621	0.0700

Table II *Out-of-sample evaluation metrics for the M2 window (2022–2024 test period).*
Each model is trained on data from 2020–2021 using one of three feature sets: financial ratios, macroeconomic indicators, or sector-specific variables. Metrics reported include the optimal classification threshold (minimizing the Brier score), accuracy, precision, recall, specificity, F1-score, ROC-AUC, PR-AUC, and Brier score.
Source*: df_NN dataset.*

Probability calibration evaluates whether a model's estimated probabilities can be taken at face value. Here, Brier scores serve as a representation of prediction errors. XGBoost achieves the lowest errors (0.012 in M1, 0.05 in M2), indicating its probabilities closely match observed frequencies. NNs follow with slightly higher Brier scores (0.02 in M1, 0.07 in M2), suggesting that while their classification performance is weaker, their predicted probabilities are reasonably calibrated. In other words, they estimate the overall likelihood of distress accurately but fail to meaningfully differentiate between healthy and distressed firms. Logit's much higher Brier values (up to 0.15 and reaching 0.45 in the M1 sector model) reflect systematic overconfidence, its high-risk predictions tend to overestimate the true distress rate. Moreover, whereas including macro and sector indicators marginally improves calibration for XGBoost and NN, it generally worsens Logit's calibration, highlighting that added controls can sometimes introduce noise rather than enhance performance. Calibration and PR-curves are in Appendix E.

5.2 Feature Importance

Logit coefficients offer a direct measure of each predictor's impact on odds ratios, with negative coefficients reducing estimated probabilities and positive ones increasing them (see Appendix F for more details). Across both M1 and M2, core accounting ratios dominate the model, while all macroeconomic features are shrunk to 0 zero by L1 regularization. In M1, EV-to-EBITDA is the strongest predictor (β_ι = -10.68, -9.69, -10.66 for financial, macro, and sector), closely followed by the missingness flag (β_ι = -0.56, -13.91, -18.25), the latter significant at the 5% level in the macro and sector models, suggesting that firms with more unreported ratios face substantially lower estimated probability of distress. In M2, operating ROA (EBIT/Assets) becomes the leading variable (β_ι = -0.58, -11.29, -13.88), while again none of the macro or sector effects are significant. This pattern indicates that in normal economic conditions valuation metrics best separate distress risk, whereas in a crisis period, raw profitability becomes more reliable

XGBoost's feature importance ranking and SHAP values reinforce this hierarchy in a nonlinear setting. Operating cash flow-to-debt and operating ROA together account for nearly half of the total importance in M1, and operating ROA rises to 34-42% in M2. In the macro setup, the general macroeconomic features have marginal or no effect at all, while in the sector model, some sector-specific features gain importance; however, the SHAP analysis shows that their impact on the output remains minimal. Generally, profitability and liquidity ratios seem to have the most effect on the estimated probabilities.

For NNs, the SHAP values are much lower, between -1 and 1, and there are no features completely dominating the predictions. Financial ratios have stronger signals, but both macro and sector variables show considerable impact, which goes to show the ability of neural networks to handle complex data structures. However, since the macro- and sector-augmented NN models exhibit low PR-AUC and recall, their feature impact should be treated with caution.

5.3 Discrimination & Ranking

The receiver operating characteristic curve shows the true positive rate against the false positive rate across every possible classification threshold, and the area under the curve (ROC-AUC) quantifies how well a model ranks distressed against healthy firms irrespective of any single cutoff. Figure 1 shows that in M1 there is a clear distinction in discriminatory power between models. All three XGBoost models closely hug the top with ROC-AUC above 0.97, suggesting almost perfect ranking ability. Logit sits around 0.75-0.80, having slightly better separation with added macro features, hinting at modest benefits from capturing broader economic signals, but it worsens with the sector setup as the extra variables likely introduce noise. The financial NN model shows slightly better discrimination than most Logit models (ROC-AUC: 0.77), but the macro and sector models significantly underperform. In M2, XGBoost still has the highest ROC-AUC of 0.94 (Fig 2), but all three Logit models and the financial NN improve significantly with ROC-AUCs between 0.80-0.87, suggesting that they work better with more volatile data. Conversely, the macro and sector NNs drop in performance (ROC-AUC at 0.56 and 0.65 respectively), which is not much better than random guessing, and indicates overfitting.

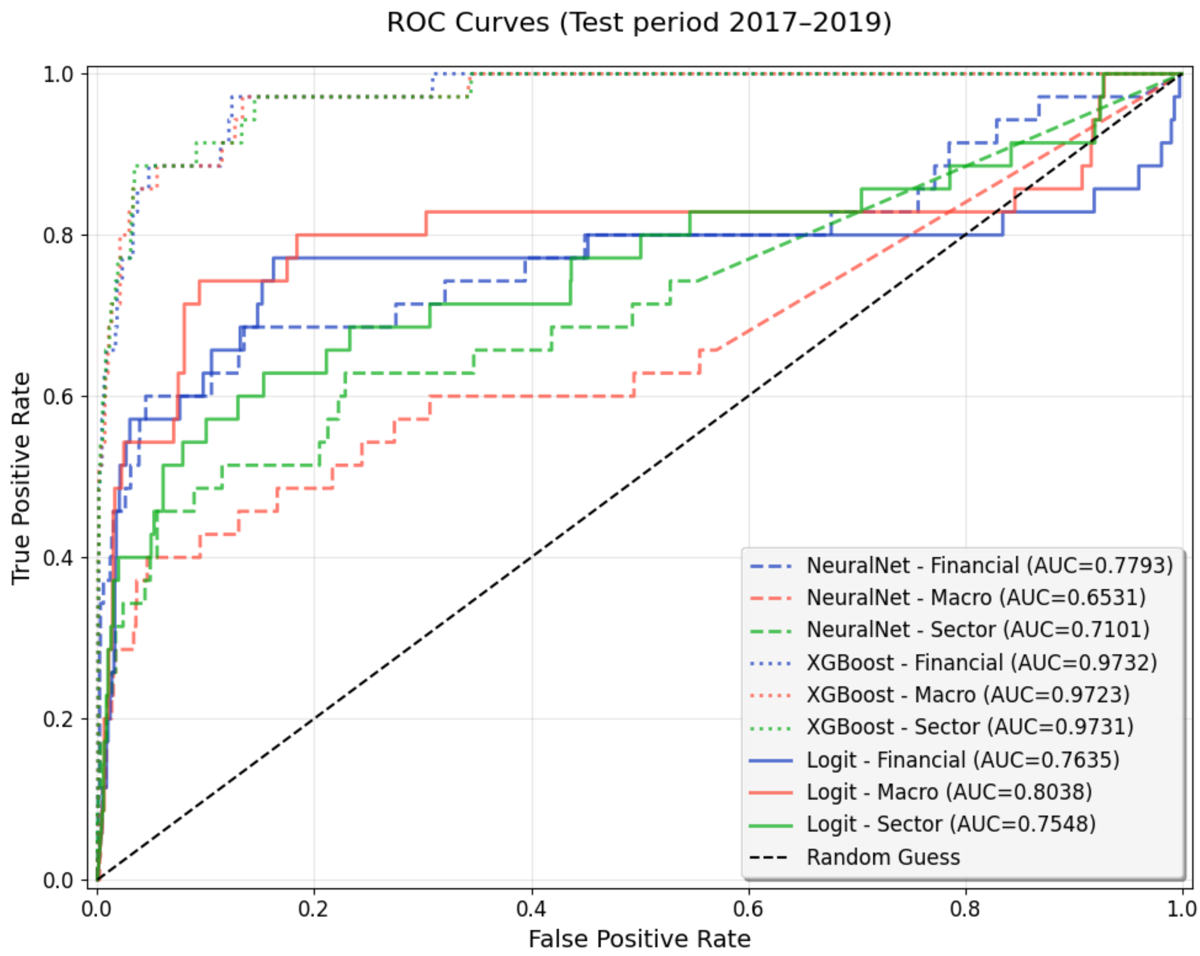

Figure 1 *ROC curves for models evaluated on the M1 test window (2017–2019).*
Each curve represents a classifier trained using either financial ratios, or considering additional macroeconomic, and sector-specific variables. The area under the curve (AUC) is reported in the legend for each model and feature set. The diagonal line denotes the performance of a random guess. Higher curves indicate stronger discrimination between distressed and non-distressed firms.
Source*: df_NN dataset.*

Lift and gain charts illustrate the proportion of distressed observations concentrated in each ranked decile[17], revealing how effectively models identify distress. A lift chart shows how much more concentrated distress is in each score decile compared to random, while a gain chart shows the cumulative share of distress cases captured as more firms are considered. In M1, XGBoost's

[17] Each decile represents a 10-percentile segment of observations ordered by predicted probability, with decile 1 being the top 10%.

top decile delivers a 9× lift and 90% gain (Fig 3 upper left and upper right), meaning that the

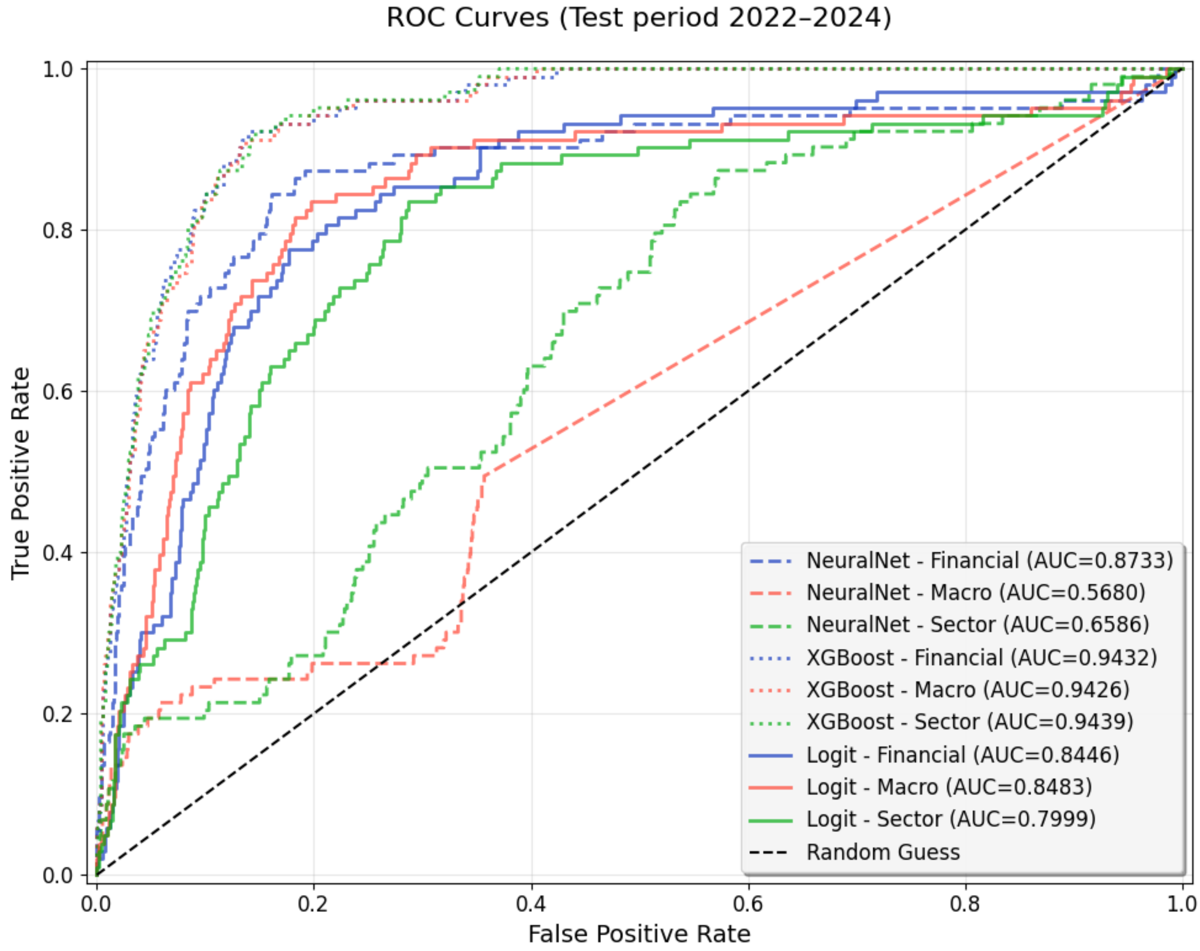

Figure 2 *ROC curves for models evaluated on the M2 test window (2022–2024).*
See the caption of Figure 1 for methodological details.
Source*: df_NN dataset.*

top 10% of firms ranked by the predicted score are nine times more likely to be distressed and they contain almost all distressed firms. Logit and NN models have much worse lifts and gains, with lower deciles still containing a significant number of true positives, and occasionally underperforming random guessing. In M2, all models perform slightly worse. XGBoost's first decile lift eases to 7× and its gain to 60% (Fig 3 lower left and lower right). Logit and the financial NN model show similar gains, reaching 80% by the second or third decile, whereas their lift ranges from 3-6 times. However, the macro and sector NN model show irregular lifts and gains, meaning their ability to rank distress before non-distress is not much better than random guessing. These charts confirm that ensemble trees are able to efficiently rank distressed firms before healthy ones, which is useful when resources for auditing are constrained.

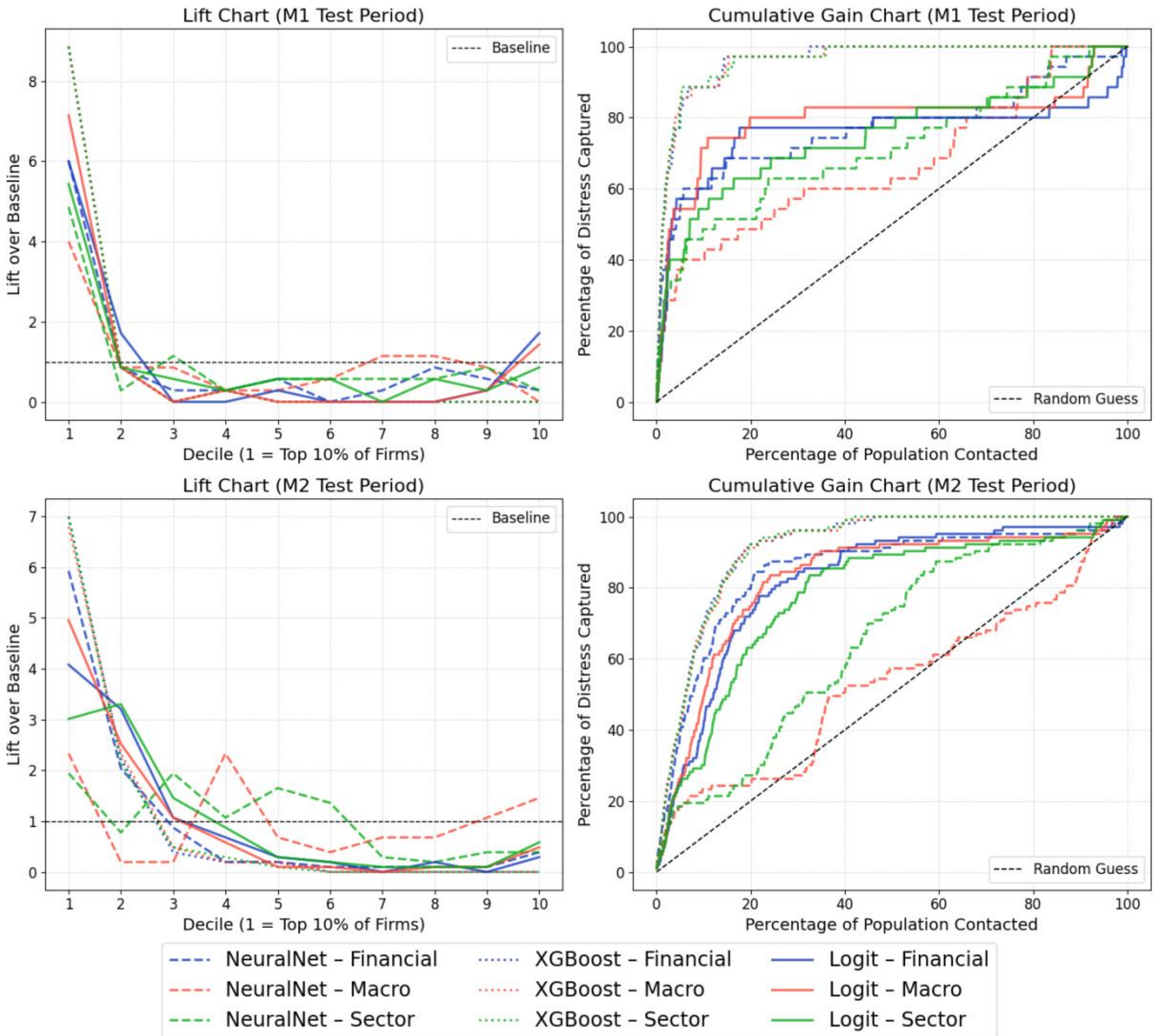

Figure 3 *Lift and cumulative gain charts for the M1 and M2 test windows.*
The top row shows results for the M1 test window (2017–2019), and the bottom row shows results for M2 (2022–2024). The left column displays lift charts, which plot how many times more likely distressed firms appear in each decile relative to the baseline distress rate. The right column shows cumulative gain charts, indicating the proportion of total distress captured as more of the population is targeted. Each line represents a model trained on one of three feature sets: financial ratios, macroeconomic indicators, or sector-specific variables. A higher curve indicates better ranking performance.
Source*: df_NN dataset.*

Across both stable and volatile economic regimes, XGBoost consistently outperforms Logit and NN in every dimension that matters for practical distress models. XGBoost delivers the highest precision-recall trade-off (PR-AUC of 0.65-0.69 in M1 and 0.55-0.56 in M2), the strongest calibration, and the steepest lift and gain curves. Logit remains interpretable and delivers a strong recall (up to 0.80), but pays heavily with Type I[18] errors and miscalibrated probabilities. Neural Nets only shine in the financial-only setting, and are prone to overfitting with the added features. In sum, fundamental accounting ratios drive nearly all predictive power, and tree-based ML learners translate them into the most robust FDP models.

[18] A Type I error, or false positive, occurs when a model incorrectly classifies a healthy (non-distressed) firm as distressed.

6. Discussion

6.1 Key Takeaways

This study first asked whether broad economic indicators add predictive value beyond firm-level accounting ratios. The results suggest that their impact is inconsistent. When macro time-fixed effects were included, they occasionally improved recall marginally, especially in Logit, yet they more often introduced noise without improving overall calibration or discrimination. Across both windows, financial ratios, like valuation multiples during expansions or profitability metrics in crises, drove nearly all decision-making. L1 regularization routinely shrunk macro coefficients to negligible in both Logit and XGBoost, while in neural nets even though they gained some importance, those models did not perform well enough to explicitly state the benefit of added macroeconomic effects. This suggests that for firm-level FDP, including macroeconomic features rarely justifies the extra data collection and modeling complexity.

The second question concerned the value of sector-specific models against a general cross-industry framework. In most cases, the sector-augmented models outperformed their general macro counterparts across all evaluation criteria, but this comes at the cost of interpretability. In the macro models, every feature applies to each firm the same way, so a rise in GDP would have the same effect on all companies, while in the sector setup, sectoral effects only influence certain firms, so their impact is localized to a handful of observations. Moreover, there was no single sector variable that remained important across models or windows, making it difficult to distinguish between the presence of true economic dynamics or just incidental correlation. In practice, these marginal improvements barely justify the added complexity and conditional interpretation required by sector-adjusted models.

Third, nonlinear models were compared with simpler methods. Extreme gradient boosting justified its complexity by outperforming Logit and NN on all metrics, including discriminatory power (ROC-AUC), precision-recall trade-off (PR-AUC), probability calibration (Brier score), and operational efficiency (decile ranking). The neural net limited to financial ratios slightly outperformed Logit, notably in PR-AUC, ranking ability, and calibration accuracy, showing its potential in FDP, albeit at the cost of a higher Type I error rate. However, once macro and sector effects were added, their performance collapsed, reflecting overfitting on small samples, while XGBoost remained robust in all settings. These patterns indicate, that for moderate datasets, ML algorithms strike an optimal balance between flexibility and stability, whereas DL architectures require substantially more data to improve upon simple financial frameworks.

Finally, comparing the two economic environments highlights the importance of adaptive modeling. In the stable 2015-2019 period, valuation multiples and cash flow metrics contributed most to distress signals, while in the more volatile COVID era, profitability and liquidity ratios proved more significant. Simultaneously, the proportion of distressed observations between the training and test data was at par in the first window (2.15% for training, and 2.38% for the test set), so thresholds and probability calibrations transferred smoothly without major performance loss. However, in the second window, the training set had 4.28% distress, while the test period jumped to 7%, causing the threshold tuned on the training data to underpredict distress and yield lower recall and precision out of sample. This widening gap in class distribution helps explain the drop in performance and underlines the need for cyclical adjustment in uncertain market conditions.

6.2 Limitations

The main data-related limitations are sample size, class imbalance, and the conceptual definition of distress. The modest sample of 980 observations per training window combined with low distress proportions (2.3%-4.28%) limits the robustness of the results. Expanding the dataset by including additional sectors or a longer timeframe would likely enhance model stability. Additionally, the distress indicator used here, while calibrated against historical bankruptcy rates, captures significant financial underperformance rather than explicit bankruptcy events. The dataset contains no real evidence of legal filings for Chapter 7 liquidation or Chapter 11 reorganization[19]. Integrating actual bankruptcy filings or credit-rating data could improve external validation and increase confidence in predicted probabilities.

Another limitation concerns hedge accounting[20], where fully hedged positions and derivative-related risks may not be explicitly reflected in financial statements. While this accounting approach accurately captures effective risk management, it also obscures underlying economic exposures that could otherwise provide valuable distress signals. Consequently, models relying heavily on financial statement volatility or profitability ratios may underestimate the true distress risk of firms applying hedge accounting, potentially introducing bias into empirical analyses.

Regarding modeling limitations, firm-year observations were treated independently, ignoring any potential autocorrelation. Time-series methods or recurrent architectures could better capture these temporal dynamics and improve forecasting performance. Moreover, the consistent use of L1 regularization and greedy threshold tuning, while effective for comparability, might not represent the optimal decision-making strategy for every algorithm. Additionally, neural network results exhibited slight variability across runs due to their

[19] Under the U.S. Bankruptcy Code, Chapter 7 refers to liquidation of assets for individuals or companies, while Chapter 11 allows businesses to reorganize their debts while continuing operations.
[20] Hedge accounting is an accounting practice that adjusts how hedging instruments are reflected in financial statements. It links the accounting treatment of a hedge (such as a derivative) with the item it is intended to offset, allowing firms to reduce volatility in reported earnings by deferring or reallocating gains and losses that would otherwise appear immediately in the income statement.

stochastic[21] training process, complicating reproducibility. Finally, while SHAP values enhance model interpretability, they fall short of the full transparency provided by linear coefficients. In regulatory or compliance-oriented settings, the preference for fully interpretable models may outweigh the performance of ML and DL algorithms.

6.3 Implications for Financial Practitioners

For risk managers and credit analysts operating with limited data and resources, gradient-boosted trees on core accounting ratios should form the basis of any FDP task. XGBoost delivers reliable probability estimates, strong discrimination, and steep lift curves, meaning audit and review efforts can concentrate on a handful of firms in the top deciles and guarantee to detect a significant portion of distress. Institutions requiring simpler models for explainability or audit compliance can still use logistic regression on those same ratios and have similar results, but they must build in rigorous calibration monitoring and expect higher false-positive rates.

The negligible effect of macroeconomic and sector-specific features in firm-level classifiers suggests that practitioners can simplify data pipelines by focusing on accounting metrics. If macro signals are desired, they may be better applied indirectly, such as in portfolio-level stress tests, rather than embedded within individual firm models. Moreover, as business cycles fluctuate, monitoring changes in feature importance and distress prevalence can guide dynamic threshold adjustments. Automating model retraining on rolling windows and integrating decile-based monitoring into risk dashboards can help keep predictions aligned with changing risk conditions and analytical capacity.

[21] Stochastic training refers to the use of random elements such as weight initialization, data shuffling, and dropout layers in neural networks, which can cause slight variations in results across different training runs, even when using the same data.

Ultimately, by matching model configuration to data availability and interpretability requirements, financial institutions can deploy distress-prediction systems that balance predictive power, operational efficiency, and regulatory transparency. Such systems will allow for more timely interventions, improve the allocation of audit resources, and strengthen the resilience of credit portfolios in all business cycles.

7. Conclusion

This thesis has demonstrated that accounting ratios are the foundation of effective corporate distress prediction. While macroeconomic and sector-specific indicators occasionally offered marginal improvements, they more often introduced additional complexity without delivering consistent or meaningful performance gains. Extreme gradient boosting consistently delivered the strongest discrimination, calibration, and operational efficiency across both stable and volatile market conditions. In contrast, logistic regression provided transparent decision boundaries and strong recall but struggled with miscalibration and high false positive rates. Shallow feedforward neural networks achieved comparable performance only in the simplest financial setting and quickly deteriorated when macroeconomic features were introduced. This highlights the limitation of neural networks in low-sample, high-noise environments where their performance is hindered by insufficient data variation. Moreover, the feature importance profiles shifted with macroeconomic conditions. Valuation metrics were driving predictions during expansionary periods, while core profitability and liquidity ratios became more prominent during downturns. This indicates the need for periodic retraining, dynamic threshold adjustment, and continuous monitoring of model behavior to maintain forecasting validity, as class proportions and business cycles evolve. Further studies could expand the dataset through higher frequency accounting data, longer time horizons, or the inclusion of international firms to improve generalizability and enable the use of temporal models like recurrent neural networks. Redefining distress through alternative labels such as credit rating downgrades, covenant breaches or formal bankruptcy filings could help validate the robustness of this analysis. Additionally, hybrid models that combine ensemble learners with scenario-based macro simulations or causal inference techniques may uncover more subtle interactions between firm-level health and external shocks. Ultimately, this study offers a scalable and interpretable framework for financial distress prediction. It provides a practical foundation for more adaptive risk management, helping credit analysts and risk officers detect early signs of financial distress.

Bibliography

Alaka, Hafiz A., Lukumon O. Oyedele, Hakeem A. Owolabi, Vikas Kumar, Saheed O. Ajayi, Olugbenga O. Akinade, and Muhammad Bilal. 2018. "Systematic Review of Bankruptcy Prediction Models: Towards a Framework for Tool Selection." *Expert Systems with Applications* 94 (March):164–84. https://doi.org/10.1016/j.eswa.2017.10.040.

Alam, Talha Mahboob, Kamran Shaukat, Mubbashar Mushtaq, Yasir Ali, Matloob Khushi, Suhuai Luo, and Abdul Wahab. 2021. "Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World." *Computer Journal* 64 (11): 1731–46. https://doi.org/10.1093/comjnl/bxaa056.

Almamy, Jeehan, John Aston, and Leonard N. Ngwa. 2016. "An Evaluation of Altman's Z-Score Using Cash Flow Ratio to Predict Corporate Failure amid the Recent Financial Crisis: Evidence from the UK." *Journal of Corporate Finance* 36 (February):278–85. https://doi.org/10.1016/j.jcorpfin.2015.12.009.

Altman, Edward I. 1968. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *The Journal of Finance* 23 (4): 589–609. https://doi.org/10.2307/2978933.

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. "Machine Learning Models and Bankruptcy Prediction." *Expert Systems with Applications* 83 (October):405–17. https://doi.org/10.1016/j.eswa.2017.04.006.

Beaver, William H. 1966. "Financial Ratios As Predictors of Failure." *Journal of Accounting Research* 4:71–111. https://doi.org/10.2307/2490171.

Békés, Gábor, and Gábor Kézdi. 2021. "Data Analysis for Business, Economics, and Policy." Higher Education from Cambridge University Press. Cambridge University Press. May 6, 2021. https://doi.org/10.1017/9781108591102.

Cornerstone Research. 2023. "Trends in Large Corporate Bankruptcy and Financial Distress: Midyear 2023 Update." New York, NY: Cornerstone Research. https://www.cornerstone.com/insights/reports/trends-in-large-corporate-bankruptcy-and-financial-distress-midyear-2023-update/.

———. 2024. "Trends in Large Corporate Bankruptcy and Financial Distress: Midyear 2024 Update." New York, NY: Cornerstone Research. 2025-01-07. https://www.cornerstone.com/insights/reports/trends-in-large-corporate-bankruptcy-and-financial-distress-midyear-2024-update/.

Federal Reserve Bank of St. Louis. 1997. "Federal Reserve Economic Data." https://fred.stlouisfed.org.

Hernandez Tinoco, Mario, and Nick Wilson. 2013. "Financial Distress and Bankruptcy Prediction among Listed Companies Using Accounting, Market and Macroeconomic Variables." *International Review of Financial Analysis* 30 (December):394–419. https://doi.org/10.1016/j.irfa.2013.02.013.

Huang, Yu-Pei, and Meng-Feng Yen. 2019. "A New Perspective of Performance Comparison among Machine Learning Algorithms for Financial Distress Prediction." *Applied Soft Computing* 83:105663-. https://doi.org/10.1016/j.asoc.2019.105663.

Jardin, Philippe du. 2009. "Bankruptcy Prediction Models: How to Choose the Most Relevant Variables?" MPRA Paper. January 2009. https://mpra.ub.uni-muenchen.de/44380/.

Kim, Hyeongjun, Hoon Cho, and Doojin Ryu. 2022. "Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data." *Computational Economics* 59 (3): 1231–49. https://doi.org/10.1007/s10614-021-10126-5.

Liu, Jiaming, Chengzhang Li, Peng Ouyang, Jiajia Liu, and Chong Wu. 2023. "Interpreting the Prediction Results of the Tree-based Gradient Boosting Models for Financial Distress Prediction with an Explainable Machine Learning Approach." *Journal of Forecasting* 42 (5): 1112–37. https://doi.org/10.1002/for.2931.

"LSEG Workspace." 2023. London Stock Exchange Group. https://www.lseg.com/en/data-analytics/products/workspace.

Ohlson, James A. 1980. "Financial Ratios and the Probabilistic Prediction of Bankruptcy." *Journal of Accounting Research* 18 (1): 109–31. https://doi.org/10.2307/2490395.

Prastyo, Dedy Dwi, Rizki Nanda Savera, and Danny Hermawan Adiwibowo. 2023. "CORPORATE FINANCIAL DISTRESS PREDICTION USING STATISTICAL EXTREME VALUE-BASED MODELING AND MACHINE LEARNING." *Media Statistika (Online)* 16 (1): 1–12. https://doi.org/10.14710/medstat.16.1.1-12.

Qian, Hongyi, Baohui Wang, Minghe Yuan, Songfeng Gao, and You Song. 2022. "Financial Distress Prediction Using a Corrected Feature Selection Measure and Gradient Boosted Decision Tree." *Expert Systems with Applications* 190:116202-. https://doi.org/10.1016/j.eswa.2021.116202.

Rahman, Md Jahidur, and Hongtao Zhu. 2024. "Predicting Financial Distress Using Machine Learning Approaches: Evidence China." *Journal of Contemporary Accounting & Economics* 20 (1): 1–25. https://doi.org/10.1016/j.jcae.2024.100403.

"S&P Capital IQ Pro." 2024. S&P Global Market Intelligence. https://www.spglobal.com/market-intelligence/en/solutions/products/sp-capital-iq-pro.

U.S. Courts. 2025. "Bankruptcy Filings Rise 14.2 Percent." United States Courts. https://www.uscourts.gov/data-news/judiciary-news/2025/02/04/bankruptcy-filings-rise-14-2-percent.

Zhao, Jinxian, Jamal Ouenniche, and Johannes De Smedt. 2024. "Survey, Classification and Critical Analysis of the Literature on Corporate Bankruptcy and Financial Distress Prediction." *Machine Learning with Applications* 15:100527-. https://doi.org/10.1016/j.mlwa.2024.100527.

Zmijewski, Mark E. 1984. "Methodological Issues Related to the Estimation of Financial Distress Prediction Models." *Journal of Accounting Research* 22:59–82. https://doi.org/10.2307/2490859.

Appendix

Appendix A: Financial & Macroeconomic Variables

Category	Name	Field Code
Income Statement	Revenue from Business Activities – Total	`TR.F.TotRevenue`
	Earnings before Interest & Taxes (EBIT)	`TR.F.EBIT`
	Earnings before Interest, Taxes, Depreciation & Amortization (EBITDA)	`TR.F.EBITDA`
	Depreciation, Depletion & Amortization – Total	`TR.F.DeprDeplAmortTot`
	Net Income after Tax	`TR.F.NetIncAfterTax`
	Income Taxes	`TR.F.IncTax`
	Interest Expense	`TR.F.IntrExpn`
	Operating Expenses – Total	`TR.F.OpExpnTot`
	Income before Taxes	`TR.F.IncBefTax`
Balance Sheet	Total Assets	`TR.F.TotAssets`
	Total Liabilities	`TR.F.TotLiab`
	Total Current Assets	`TR.F.TotCurrAssets`
	Total Current Liabilities	`TR.F.TotCurrLiab`
	Working Capital	`TR.F.WkgCap`
	Cash & Cash Equivalents – Total	`TR.F.CashCashEquivTot`
	Cash & Short-Term Investments – Total	`TR.F.CashSTInvstTot`
	Loans & Receivables – Total	`TR.F.LoansRcvblTot`
	Payables & Accrued Expenses	`TR.F.PbleAccrExpn`
	Debt – Total	`TR.F.DebtTot`
	Net Debt	`TR.F.NetDebt`
	Debt, Long-Term – Total	`TR.F.DebtLTTot`
	Short-Term Debt & Current Portion of Long-Term Debt	`TR.F.STDebtCurrPortOfLTDebt`
	Shareholders' Equity – Common	`TR.F.ShHoldEqCom`
	Retained Earnings – Total	`TR.F.RetainedEarnTot`
	Equity, Non-Contributed, Reserves & Retained Earnings	`TR.F.EqNonContribRsrvRetainedEarn`
	Total Comprehensive Income, Accumulated	`TR.F.ComprIncAccumTot`
	Other Equity Reserves – Total	`TR.F.OthRsrvEqTot`
	Deposits – Total	`TR.F.DeposTot`
	Collateralized Agreements (REPOs & Securities Borrowed) – Assets	`TR.F.CollatAgrmtAssets`
	Unearned Premiums – Gross	`TR.F.UnearnPremGross`
	Policy & Contract Claims – Gross	`TR.F.PolicyContrClaimsGross`
	Operating Lease Liabilities – Current Portion / Short-Term	`TR.F.OpLeaseLiabCurrPortST`
	Dividends Payable	`TR.F.DivPble`
Cash Flow	Net Cash Flow from Operating Activities	`TR.F.NetCashFlowOp`
	Net Cash Flow from Investing Activities	`TR.F.NetCashFlowInvst`
	Net Cash Flow from Financing Activities	`TR.F.NetCashFlowFin`
	Free Cash Flow	`TR.F.LeveredFOCF`
	Capital Expenditures – Total	`TR.F.CAPEXTot`
	Capital Expenditures, Net (Cash Flow)	`TR.F.CAPEXNetCF`
	Working Capital Increase/(Decrease) (Cash Flow)	`TR.F.WkgCapCF`
	Foreign Exchange Effects (Cash Flow)	`TR.F.FXEffectsCF`
	Net Cash from Discontinued Operations	`TR.F.NetCashDiscOps`
	Net Change in Cash – Total	`TR.F.NetChgInCashTot`
	Net Cash from Continuing Operations	`TR.F.NetCashContOps`
	Non-Classified Cash Flows	`TR.F.NonClassifCashFlows`
Valuation	Market Capitalization	`TR.F.MktCap`
	Price Close	`TR.PriceClose(SDate=-1)`
	Common Shares Outstanding – Total	`TR.F.ComShrOutsTot`
	Enterprise Value	`TR.F.EV`
	Debt inc. Preferred Equity & Minority Interest – Total	`TR.F.DebtInclPrefEqMinIntrTot`
	Weighted Average Cost of Capital (WACC, %)	`TR.WACC(SDate=-1)`
	WACC Cost of Debt (%)	`TR.WACCCostofDebt(SDate=-1)`
Other	Beta (derived from daily stock returns)	$\frac{\mathrm{Cov(stock, market)}}{\mathrm{Var(market)}}$
	Idiosyncratic risk (derived from Beta)	$\max\big(\mathrm{Var(stock)} - \beta^2_{\mathrm{stock}}\mathrm{Var(market)}, 0\big)$

Table III *Financial statement variables and corresponding LSEG Workspace field codes.*
This table lists the financial statement items used in the analysis, organized by reporting category. The "Name" column shows the official item names as defined by LSEG Workspace. The "Field Code" column provides the corresponding API identifier used to extract each variable. Beta and Idiosyncratic Risk are computed from daily stock returns on a calendar-year basis. ***Source:*** *LSEG Workspace – Data Item Browser.*

Group	Variable	Definition	Series ID / Formula
General	GDP	Gross Domestic Product	GDP
	UNRATE	Unemployment Rate (%)	UNRATE
	CPI	Inflation Rate (Consumer Price Index, 1984 = 100)	CPIAUCSL
	CCI	Consumer Confidence Index (1966 = 100)	UMCSENT
	FEDFUNDS	Federal Funds Rate (%)	FEDFUNDS
	STRFED	St. Louis Fed Financial Stress Index	STLFSI4
	YCSPRED	Yield Curve Spread (10 Y–2 Y Treasury) (%)	T10Y2Y
	VIXLOGR	VIX annualized log-returns	$\ln(\mathrm{VIX}_t/\mathrm{VIX}_{t-1})$
	VIXSTDR	VIX annualized relative standard deviation	σ_t/μ_t
	BetaMKT	Market beta (S&P 500 / SPY)	1.0 (default)
Financials	BNKASST	Commercial bank assets	TLAACBW027SBOG
	DLQRATE	Delinquency rate on commercial & industrial loans	DRALACBN
	TCLOANOUT	Total consumer loans and leases outstanding	TOTCI
	BetaFIN	Sector beta: Financials (XLF)	XLF
Real Estate	CRELOAN	Commercial real estate loans	REALLN
	MORTGAGE	30-year fixed-rate mortgage average (%)	MORTGAGE30US
	HOUST	Housing starts	HOUST
	CONSPEND	Total construction spending	TTLCONS
	IRCREPI	Commercial real estate price index	BOGZ1FL075035503Q
	BetaREI	Sector beta: Real Estate (VNQ)	VNQ
Industrials	MANEMP	All employees: Manufacturing	MANEMP
	DGORDER	Manufacturers' new orders for durable goods	DGORDER
	BetaIND	Sector beta: Industrials (XLI)	XLI
Industrials & Materials	BUSINV	Total business inventories	BUSINV
	CAPUT	Capacity utilization (%)	TCU
	INDPRO	Industrial production index (2017 = 100)	INDPRO
Materials	IPMINE	Industrial production: mining, quarrying, oil & gas extraction (2017 = 100)	IPMINE
	PPIACO	Producer Price Index: all commodities (1982 = 100)	PPIACO
	GPIACO	Global price index: all commodities (2016 = 100)	PALLFNFINDEXQ
	BetaMAT	Sector beta: Materials (XLB)	XLB

Table IV *Description of macroeconomic variables used in model training and evaluation.*
The table lists the macroeconomic and sector-level variables grouped by thematic category. The "Variable" column provides the internal name used throughout the analysis, while the "Definition" explains each indicator, and the "Series ID / Formula" shows its corresponding FRED code, ETF ticker, or computational formula. Variables marked with "(%)" are expressed as raw percentages. All other time series were transformed into year-over-year percentage changes, except for UNRATE, where the first difference was used. Volatility measures (VIXLOGR, VIXSTDR) are derived from the VIX closing price using log-returns and rolling standard deviation. Sector betas are estimated annually under CAPM assumptions using daily ETF returns.
Source: *df11_macro_data dataset; data from Federal Reserve Economic Data (FRED) and LSEG Workspace.*

Dimension	Ratio	Formula
Profitability	OpROA	EBIT / Total Assets
	ROA	Net Income / Total Assets
	Net Profit Margin	Net Income / Revenue
	ROE	Net Income / Equity
	EBITDA Margin	EBITDA / Revenue
	Operating Margin	EBIT / Revenue

Continued on next page

Dimension	Ratio	Formula
Liquidity	Current Ratio	Current Assets / Current Liabilities
	Quick Ratio	(Cash & STI + Loans & Receivables) / Current Liabilities
	Cash Ratio	Cash & STI / Current Liabilities
	Working Capital-to-Assets	Working Capital / Total Assets
Leverage	Debt-to-Assets	Total Liabilities / Total Assets
	Debt-to-Equity	Total Liabilities / Equity
	Net Debt-to-EBITDA	Net Debt / EBITDA
	Net Debt-to-Assets	Net Debt / Total Assets
	Interest Coverage Ratio	EBIT / Interest Expense
	Retained Earnings-to-Assets	Retained Earnings / Total Assets
	Market Cap-to-Liabilities	Market Cap / Total Liabilities
Efficiency	Asset Turnover	Revenue / Total Assets
	Revenue / Receivables	Revenue / Loans & Receivables
	Capital Intensity	CAPEX / Total Assets
	Working Capital Turnover	Revenue / Working Capital
Cash Flow	CF Coverage Ratio	Operating CF / Total Debt
	OCF Margin	Operating CF / Revenue
	OCF-to-Debt	Operating CF / Total Liabilities
	FCF-to-Debt	Free CF / Total Debt
Valuation	Market-to-Book Ratio	Market Cap / Equity
	EV-to-Revenue	Enterprise Value / Revenue
	EV-to-EBITDA	Enterprise Value / EBITDA
Growth	Revenue Growth	$(R_t - R_{t-1})/R_{t-1}$
	Total Assets Growth	$(A_t - A_{t-1})/A_{t-1}$
	Equity Growth	$(E_t - E_{t-1})/E_{t-1}$
	Liabilities Growth	$(L_t - L_{t-1})/L_{t-1}$
	OCF Growth	$(\mathrm{CF}_t - \mathrm{CF}_{t-1})/\mathrm{CF}_{t-1}$
	Cash Growth	$(C_t - C_{t-1})/C_{t-1}$
	Working Capital Growth	$(\mathrm{WC}_t - \mathrm{WC}_{t-1})/\mathrm{WC}_{t-1}$
	Net Debt Growth	$(\mathrm{ND}_t - \mathrm{ND}_{t-1})/\mathrm{ND}_{t-1}$
	Net Income Growth	$(\mathrm{NI}_t - \mathrm{NI}_{t-1})/\mathrm{NI}_{t-1}$
Earnings Quality	Cash Flow to Earnings Ratio	Operating CF / Net Income
	Earnings Quality	Operating CF / EBIT
Earnings Sensitivity	Operating Leverage	EBIT Growth / Revenue Growth
Firm–macro interaction	Inflation adjusted Liquidity	Quick Ratio × (1 + CPI)
	Short-term Interest Rate adjusted Leverage	Debt/Equity × FEDFUNDS
	GDP adjusted Returns	ROA × (1 + GDP)

Table V *Financial ratios, formulas, and feature engineering logic.*
This table lists the financial ratios used for modeling, grouped by dimension (e.g., profitability, leverage, efficiency). The "Ratio" column names the variable, while the "Formula" column defines its calculation using standard accounting elements. Missing values typically result from division by zero, and such cases are flagged using indicator variables (e.g., Debt Flag, Receivables Flag). The Any_NA_M1 and Any_NA_M2 variables count the number of flags per observation in each feature set. If the FS24 Flag is true (indicating a missing 2024 10-K), all features for that firm-year are set to zero; this applies to only three observations.
Source: *df_ratio_table and df_NN datasets.*

Appendix B: Data Cleaning Process

B.1 Accounting Approximation

Missing Variable	Imputation Formula	Penalty
Revenue from Business Activities	EBIT + Operating Expenses	0
Depreciation, Depletion & Amortization	EBITDA – EBIT	0
Income Taxes	Income before Taxes – Net Income after Tax	0
Interest Expense	EBIT – Income Taxes – Net Income after Tax	1
Total Liabilities	Total Assets – Shareholders' Equity	0
Comprehensive Income	assign 0	0
Other Reserves/Equity	assign 0	0
Retained Earnings	Equity, Non-Contributed Reserves & Retained Earnings – (Comprehensive Income) – (Other Reserves/Equity) assign 0 (history present) assign 0 (no history)	0 1 0
Debt Total	Net Debt + Cash & Short-Term Investments	0
Debt Long-Term	Debt Total – Debt Short-Term 0, if Debt Total = 0	0 0
Debt Short-Term	Debt Total – Debt Long-Term 0, if Debt Total = 0	0 0
Payables & Accrued Expenses	assign 0 (history present) assign 0 (no history)	1 0
Loans & Receivables	assign 0 (history present) assign 0 (no history)	1 0
Current Liabilities	assign 0, if Total Liabilities = 0	0
Current Assets	assign 0, if Total Assets = 0	0
Current Assets (Financials only)	Cash & Short-Term Investments + Loans & Receivables + REPOs & Collateralized Agreements (Assets)	0
Current Liabilities (Financials only)	Payables & Accrued Expenses + Short-Term Debt + Deposits + Unearned Premiums + Policy Claims + Operating Lease Liabilities (Short-Term) + Dividends Payable	0
Working Capital	Total Current Assets – Total Current Liabilities	0
Net Cash from Discontinued Operations (NCDO)	assign 0 if missing	0
Non Classified Cash Flows (NCCF)	Net Change in Cash – Net Cash from Continuing Operations – Net Cash from Discontinued Operations	1
Operating Cash Flow (OCF)	assign 0 if missing	0
Investing Cash Flow (CFI)	assign 0 if missing	0
Financing Cash Flow (CFF)	assign 0 if missing	0
Foreign Exchange Effects (FX)	assign 0 if missing	0
Any missing among OCF, CFI, CFF, FX	$\dfrac{\text{Net Cash from Continuing Ops} - \sum \text{non-missing items}}{\text{number of missing items}}$	$N-1$
Total Capital Expenditure (CAPEXTot)	OCF – FCF CAPEX Net – CFI	0 2 3
Free Cash Flow (FCF)	OCF – CAPEXTot OCF – Working Capital (CF)	0 2
Market Capitalization	Price Close × Shares Outstanding	0
Price Close	Market Capitalization / Shares Outstanding	0
Shares Outstanding	Market Capitalization / Price Close	0
Enterprise Value	Market Capitalization + Total Debt + Cash & Short-Term Investments	0

Table VI *Variable-specific reconstruction formulas and associated penalties.*
(Caption continues on the next page)

This table documents the formulas and approximations used to impute missing financial statement variables during the data cleaning process. For each variable, the most appropriate reconstruction rule was applied based on accounting identities, data structure, or economic logic. The final column indicates a penalty score that reflects the degree of approximation involved in the imputation. These penalties were used for internal documentation only and ***were not included*** *as features in model training or evaluation.*

- ***Penalty 0*** *– Exact accounting identity or value directly documented in LSEG Workspace API*
- ***Penalty 1*** *– Indirect identity or approximation relying on related reported items*
- ***Penalty 2*** *– Approximation using multiple proxy components or reconstructed inputs*
- ***Penalty 3*** *– Heavily estimated or interpolated values with no direct reporting support*
- ***Penalty N–1*** *– Applied to averages over N components, assigning a fractional penalty to each non-missing input (e.g., dividing by 2 gives 1 per component; dividing by 3 gives 2 per component)*

Some of the variable names in this table are shortened for readability; the exact LSEG Workspace field names and API codes are provided in Table III.
Source: *LSEG Workspace API documentation; reconstruction logic supported by ChatGPT.*

B.2 Imputation

Time series imputation was used selectively to address residual missing values for 6 firms, where reconstruction using accounting identities was not possible.

Interpolation of missing reports:
One firm had no cash flow reporting for 2019. In this case, cash flow metrics were linearly interpolated by taking the average of the surrounding years (2018 and 2020). These values were given a penalty of 3 (see Table VI).

Regression imputation:
Five firms with irregular missing values were imputed using ordinary least squares (OLS) regression. For each financial variable, a linear model was fitted using the firm's historical data, provided that at least 6 out of 11 years were available. This approach was used to estimate gaps in core financial statement items such as revenue or liabilities, resulting in trends consistent with related non-missing indicators. These imputed values were assigned a penalty score of 3. Since these variables (see Table III) were not directly used in any of the predictive models, the imputations had minimal impact on model outcomes.

The same method was used for estimating WACC and cost of debt for a larger set of firms. Due to their relative stability, the minimum required number of non-missing years was reduced to four. Although this step resulted in roughly 500 imputed values, these features were ultimately excluded from the model training, making them analytically irrelevant. Nonetheless, values were retained and assigned a penalty of 3 for completeness.

Appendix C: Target Variable Construction

C.1 Distress Criteria

Criteria	Definition
Net Income ≤ 0	Indicates a loss-making period, compromising basic profitability.
Revenue ≤ 0	Suggests operations are halted or the firm is in severely declining.
Revenue Growth < –10%	Significant contraction in performance.
Interest Coverage Ratio < 1	Indicates inability to cover interest expenses from EBIT.
Current Ratio < 1	The Firm may struggle to meet short-term obligations.
Asset Turnover < 0.5	Signals inefficient use of assets to generate revenue.
Operating Cash Flow < 0	Core operations are burning cash, not generating it.
Retained Earnings-to-Assets < –0.5	Long-term cumulative losses outweigh retained value.
Working Capital < 0	Current liabilities exceed current assets.
EBITDA Margin < 0	The Firm has a negative operating margin.
Cash Ratio < 25th percentile	Minimal immediate liquidity buffer.
Earnings Quality < 25th percentile	Weak alignment between earnings and cash flows.
Market Cap-to-Liabilities < 25th percentile (sector)	Weak market confidence relative to financial obligations.
Debt-to-Assets > 75th percentile (sector)	Highly leveraged firm compared to sector norms.
Capital Intensity < 25th percentile (sector)	Low investment in fixed assets, possibly underdeveloped.

Table VII *Financial distress criteria and descriptions.*
This table lists the 15 criteria used to construct the binary target variable. Each condition reflects a signal of financial weakness across profitability, liquidity, efficiency, solvency, or growth. Firm-year observations receive one point for each condition met, and are classified as distressed if they accumulate at least 9 out of 15 points. All criteria are weighted equally. Thresholds expressed in percentiles are computed annually across either the full sample or sector-specific distributions, as noted. Sector-adjusted thresholds account for industry variation in capital structure and investment intensity. The chosen 9-point threshold yields a distress rate of approximately 3.9%.
Source*: df_NN dataset*

C.2 Agreement Analysis

To evaluate how the classification method in this study compares to established approaches, we assess agreement with three traditional scoring models: Altman Z, Ohlson O, and Zmijewski X. These models estimate financial bankruptcy based on accounting ratios related to profitability, leverage, and liquidity. The formulas and classification thresholds are shown in Table 8, with each model assigning observations to one of three categories: distressed, grey zone (ambiguous), or healthy.

To enable comparison, each score was converted into a numeric vote: 1 for distress, 0 for non-distress, and 0.5 for ambiguous values. Averaging the three votes provided a consensus classification, which was then compared to our distress labeling. Due to missing inputs, only 5,385 firm-year observations could be evaluated. Among these, 726 were unanimously flagged as distressed by all three scores but not under the study's definition. Full agreement occurred in 713 cases, while the remaining observations showed mixed or partial alignment.

These results confirm that the traditional models, especially Altman Z and Zmijewski X, tend to overclassify distress in this sample. In contrast, the classification used here is more selective and focused on sustained underperformance, making occasional disagreements expected.

Altman Z-Score

$$Z = 1.2\,\frac{\text{Working Capital}}{\text{Total Assets}} + 1.4\,\frac{\text{Retained Earnings}}{\text{Total Assets}} + 3.3\,\frac{\text{EBIT}}{\text{Total Assets}} + 0.6\,\frac{\text{Market Cap}}{\text{Total Liabilities}} + 1.0\,\frac{\text{Revenue}}{\text{Total Assets}}$$

Ohlson O-Score

$$O = -1.32 - 0.407\,\ln(\text{Total Assets}) + 6.03\,\frac{\text{Total Liabilities}}{\text{Total Assets}} - 1.43\,\frac{\text{Working Capital}}{\text{Total Assets}} + 0.0757\left(1 - \frac{\text{Current Assets}}{\text{Current Liabilities}}\right)$$

$$- 2.37\,\frac{\text{Net Income}}{\text{Total Assets}} - 1.83\,\frac{\text{Operating Cash Flow}}{\text{Total Liabilities}} + 0.285\,\text{Ohl7} - 0.521\,\text{Ohl8} - 1.72\,\text{Ohl9}$$

Where

Ohl7 = indicator of negative net income in the past 2 years,
Ohl8 = net income momentum,
Ohl9 = indicator if liabilities exceed assets.

Zmijewski X-Score

$$X = -4.3 - 4.5\,\frac{\text{Net Income}}{\text{Total Assets}} + 5.7\,\frac{\text{Total Liabilities}}{\text{Total Assets}} + 0.004\,\frac{\text{Current Assets}}{\text{Current Liabilities}}$$

Thresholds

Score	Distress	Grey Zone	Healthy
Altman Z-Score	$Z < 1.81$	$1.81 \leq Z \leq 2.99$	$Z > 2.99$
Ohlson O-Score	$O > -5.16$	$-7.50 \leq O \leq -5.16$	$O < -7.50$
Zmijewski X-Score	$X > 0.65$	$-1.50 \leq X \leq 0.65$	$X < -1.50$

***Table VIII** Formulas and Thresholds for Traditional Financial Distress Scores.*
This figure presents the equations and classification thresholds for the Altman Z-Score, Ohlson O-Score, and Zmijewski X-Score. Each model outputs a continuous score based on accounting ratios that reflect profitability, leverage, liquidity, and operational risk. Observations are classified as distressed, healthy, or within a grey zone based on their position relative to the established thresholds. These thresholds are applied consistently across all firm-year observations in the agreement analysis (Appendix C.2).

Appendix D: Feature Selection & Model Structures

D.1 Feature Selection

To ensure both model efficiency and interpretability, a two-stage feature selection process was applied to the financial and interaction-based feature sets. In the first stage, a Random Forest classifier was trained on the full training window using all available financial variables. Random Forest is a tree-based ensemble method that estimates the importance of each feature based on its contribution to reducing classification error across many decision trees. Features that consistently split the data in ways that improve classification accuracy receive higher importance scores. Since the algorithm is non-parametric and robust to scale, it provides a reliable initial ranking of predictive power.

In the second stage, the top-ranked variables were filtered using Variance Inflation Factor (VIF) analysis to reduce multicollinearity. VIF quantifies how much the variance of a regression coefficient is inflated due to linear correlation with other predictors. A greedy selection procedure was implemented: features were added one by one based on importance, but only retained if the resulting VIF for all variables remained below a strict threshold of 3. This ensured that retained features were both informative and statistically independent.

Feature Set 1		Feature Set 2	
Variable	**Category**	**Variable**	**Category**
Operating ROA	Profitability	ROA	Profitability
ROE	Profitability	Net Profit Margin	Profitability
OCF-to-Debt	Cash Flow	ROE	Profitability
OCF Margin	Cash Flow	OCF-to-Debt	Cash Flow
CF Earnings Ratio	Earnings quality	CF Coverage Ratio	Cash Flow
EV-to-EBITDA	Valuation	CF Earnings Ratio	Earnings quality
Market-to-Book Ratio	Valuation	EV-to-EBITDA	Valuation
Working Capital-to-Assets	Liquidity	Working Capital-to-Assets	Liquidity
Quick Ratio	Liquidity	Quick Ratio	Liquidity
Equity Growth	Growth	Net Debt-to-Assets	Leverage
Any_NA_M1	Missingness	Any_NA_M2	Missingness

Table IX *Selected financial ratios per feature set.*
This table displays the final selected features for Feature Set 1 (financial ratios only) and Feature Set 2 (financial ratios with macro interaction terms), as determined by Random Forest importance ranking followed by VIF-based filtering (VIF < 3). Each set includes 10 variables spanning key financial dimensions, along with a missingness indicator. This process ensures that retained variables are both predictive and sufficiently independent for model training.

The process was applied separately to financial ratios alone and to ratios that included firm–macro interaction terms (Table V). In each case, the 10 most informative and non-collinear variables were selected. Both sets include indicators across categories such as profitability, liquidity, leverage, valuation, and cash flow (Table IX). A missingness indicator for non-applicable ratios (Any_NA) was also included to help models account for structural missingness without distortion.

D.2 Model Structure and Optimal Parameters

D.2.1 Logistic Regression

This study employs L1-penalized (LASSO) logistic regression to estimate the probability that a firm-year observation falls into financial distress. The model is based on the logistic function:

$$P(Y = 1 \mid \mathbf{X}) = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$$

where X is the vector of input features and β are the model coefficients. All predictor variables were standardized to have mean zero and unit variance, such that a unit increase in any variable corresponds to a one standard deviation change. L1 regularization was applied using the saga solver from scikit-learn, which introduces sparsity by shrinking less important coefficients to exactly zero. The regularization strength was controlled by the inverse parameter C, where $\alpha = 1/C$, with higher C values allowing more complexity and lower C values inducing stronger shrinkage. To select the optimal C, a grid search was performed over the following values: $C \in \{0.01, 0.1, 0.5, 1, 2, 5, 10, 100\}$. The value minimizing the F1-score loss on the upsampled training set was retained.

The class imbalance was addressed using the “RandomOverSampler” from the “imblearn” library. Oversampling was applied within each training fold before scaling to ensure a balanced representation of the minority class (distressed = 1). Each model was trained using 5-fold stratified cross-validation (“StratifiedKFold”) to preserve the class distribution. A StandardScaler was fit on the resampled training data and applied to both training and validation sets to avoid information leakage. After cross-validation, each model was retrained on the full upsampled training set, and the optimal decision threshold was selected by minimizing the **Brier score** using a greedy search over 1001 equally spaced values between 0 and 1.

The final trained model outputs a vector of probabilities, which were compared to the optimal threshold to generate binary distress classifications. Performance was evaluated on the out-of-sample test set (2017–2019 for M1 and 2022–2024 for M2) using several metrics: accuracy, recall, precision, F1-score, specificity, ROC-AUC, PR-AUC, and the Brier score. Confusion matrices were also generated and stored. Coefficients were later converted into odds ratios for interpretability and reported alongside p-values, derived from Wald tests, to assess statistical significance. Calibration curves were produced for further validation of probability estimates. Additional performance and calibration results for logistic models are reported in Tables I–II and Appendices E.1–E.3.

D.2.2 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a high-performance implementation of gradient-boosting decision trees, which is an ensemble technique that builds additive predictive models by sequentially fitting new trees to the residuals of prior iterations. At each stage, a new tree is trained to minimize the gradient of a specified loss function, effectively correcting the errors of the previous model. In this study, XGBoost is used with the logloss objective, which optimizes the log-likelihood for binary classification, making it well-suited for probabilistic outputs.

Each tree in the ensemble is built using a greedy splitting process that evaluates candidate splits based on their ability to reduce the loss function. The final prediction is the weighted sum of all trees in the model. Regularization is applied via parameters such as reg_alpha (L1 penalty), gamma (minimum loss reduction to split), and constraints on tree complexity like max_depth, subsample, and colsample_bytree. The model outputs a continuous probability between 0 and 1, interpreted as the predicted likelihood of financial distress for a given firm-year observation. To convert these scores into binary classifications, the optimal threshold is selected by minimizing the Brier score, using the same procedure outlined in Appendix D.2.1.

The training follows the same cross-validation structure used for the Logit model. Each training window is split into five stratified folds, and within each fold, the minority class is randomly oversampled. Features are standardized after oversampling to ensure comparability across models. For each model, a fixed set of hyperparameters is applied (Table X), selected via grid search over combinations of core tree-building parameters. Final models are retrained on the entire upsampled training set and evaluated on the corresponding test window.

Model	n_estimators	max_depth	learning_rate	subsample	colsample_bytree	gamma	reg_alpha
Grid search range							
–	{50,100,200,500}	{3,4,5}	{0.01,0.05}	{0.75,0.90,1.0}	{0.75,0.90,1.0}	{0,0.1,1}	{0,0.5,1}
Financial M1	500	5	0.05	1.0	0.90	0.1	0
Financial M2	500	5	0.05	1.0	0.75	0.0	0
Macro M1	500	5	0.05	1.0	0.75	0.1	0
Macro M2	500	5	0.05	1.0	0.75	0.0	0
Sector M1	500	4	0.05	1.0	0.75	0.0	0
Sector M2	500	5	0.05	1.0	0.90	0.1	0

Table X *Hyperparameter Grid and Selected Values for XGBoost Models.*
The top row shows the grid search space used to tune model parameters for each feature setup (financial, macroeconomic, and sector-specific) across both evaluation windows (M1: 2015–2019, M2: 2020–2024). The parameters include n_estimators (number of boosting rounds), max_depth (maximum tree depth), learning_rate (step size shrinkage), subsample (fraction of samples per tree), colsample_bytree (fraction of features per tree), gamma (minimum loss reduction to split), and reg_alpha (L1 regularization penalty). Selected values for each configuration reflect the best-performing combination based on cross-validated F1-score. All models were trained with the logloss objective.

D.2.3 Two-layer Feedforward Neural Network

Feedforward neural networks are a class of supervised learning algorithms that model complex, nonlinear relationships by passing input features through layers of interconnected nodes. Each node applies a weighted sum of its inputs followed by a nonlinear activation function, enabling the network to approximate flexible mappings between inputs and outputs. In this thesis, a shallow feedforward neural network is used, consisting of two hidden layers and one output layer. These models are particularly useful for capturing intricate interactions between financial and macroeconomic features that may not be well represented by linear or tree-based methods.

Each network is implemented using the Keras API with TensorFlow backend. The hidden layers use the ReLU activation function, which introduces nonlinearity while avoiding issues like vanishing gradients. The final output layer contains a single neuron with a sigmoid activation function, producing a probability estimate between 0 and 1 for each observation. These outputs reflect the model's predicted likelihood of financial distress. To prevent overfitting and stabilize training in relatively small and noisy datasets, two forms of regularization are applied. First, L2 regularization (also called RIDGE penalty) is used on the network weights. Unlike L1 regularization (used in the logistic model), which shrinks some weights to zero and thus performs variable selection, L2 regularization penalizes large weights without eliminating variables. This is more suitable for neural networks, where predictive power is distributed across many connections rather than individual features. Second, dropout layers are added after each hidden layer to randomly deactivate a fraction of neurons during training. This acts as a structural regularization mechanism by preventing the model from becoming overly reliant on any single pathway and encourages more robust learning. Models are compiled using the Adam optimizer with a fixed learning rate and trained using the binary cross-entropy loss function, which is appropriate for probabilistic binary classification. Early stopping is employed to halt training once validation loss stagnates for 10 epochs, with the model reverting to the best-performing weights. Training is performed for up to 100–200 epochs with a batch size of 32, using a 10% validation split from the upsampled training set.

Hyperparameters include the number of neurons in the first and second hidden layers (neurons_layer1, neurons_layer2), the dropout rate, the learning rate, and the L2 penalty. While the architecture was held constant across most models, macroeconomic setups used smaller hidden layers to reduce overfitting. Predictions are converted to binary classifications using a threshold selected via greedy search on the training set to minimize the Brier score. Final performance metrics are then computed on the test window using this fixed threshold.

Model	`neurons_layer1`	`neurons_layer2`	`dropout_rate`	`learning_rate`	`l2_penalty`
Grid search range	{32, 64}	{16, 32}	{0.3, 0.5}	{0.001, 0.0005}	{0.0001, 0.001, 0.01}
Financial M1	64	32	0.3	0.001	0.0001
Financial M2	64	32	0.3	0.001	0.0001
Macro M1	32	32	0.3	0.001	0.0001
Macro M2	32	32	0.3	0.001	0.0001
Sector M1	64	32	0.3	0.001	0.0001
Sector M2	64	32	0.3	0.001	0.0001

Table XI *Hyperparameter Grid and Selected Neural Network Architectures by Setup.*
This table summarizes the grid search range and optimal parameter values for the shallow feedforward neural network (NN) models across feature configurations and evaluation windows. Each model uses two hidden layers, with neurons_layer1 and neurons_layer2 indicating the number of units in the first and second layers, respectively. The dropout_rate controls the fraction of neurons randomly deactivated during training for regularization. learning_rate is the step size used by the Adam optimizer, while l2_penalty sets the strength of L2 weight regularization. Macroeconomic setups use a smaller architecture to reduce overfitting, while financial and sectoral setups rely on a larger configuration for improved representation capacity.

Appendix E: Modeling Results

E.1 Confusion Matrices

Model	Setup	True Negative	False Positive	False Negative	True Positive
Logit	Financial	1295	140	14	21
	Macro	1200	235	9	26
	Sector	660	775	7	28
NeuralNet	Financial	1418	17	21	14
	Macro	1431	4	34	1
	Sector	1433	2	31	4
XGBoost	Financial	1417	18	12	23
	Macro	1415	20	10	25
	Sector	1415	20	10	25

Model	Setup	True Negative	False Positive	False Negative	True Positive
Logit	Financial	1177	190	33	70
	Macro	1091	276	17	86
	Sector	1119	248	36	67
NeuralNet	Financial	1264	103	40	63
	Macro	1367	0	102	1
	Sector	1366	1	102	1
XGBoost	Financial	1318	49	41	62
	Macro	1314	53	44	59
	Sector	1315	52	42	61

Table XII *Confusion matrix counts for M1 (top) and M2 (bottom) test windows.*
This figure presents the confusion matrix counts for all model–setup combinations across two evaluation periods. The top panel (M1) corresponds to models trained on 2015–2016 and tested on 2017–2019, while the bottom panel (M2) reflects training on 2020–2021 and testing on 2022–2024. Each matrix reports the number of True Positives (correctly identified distressed firms), True Negatives (correctly identified healthy firms), False Positives (Type I errors), and False Negatives (Type II errors). These values form the basis for computing most evaluation metrics (accuracy, precision, recall, F1-score, specificity, ROC-AUC, and PR-AUC) reported in Tables I and II, respectively.
Source*: df_NN dataset.*

E.2 Precision-Recall Curves

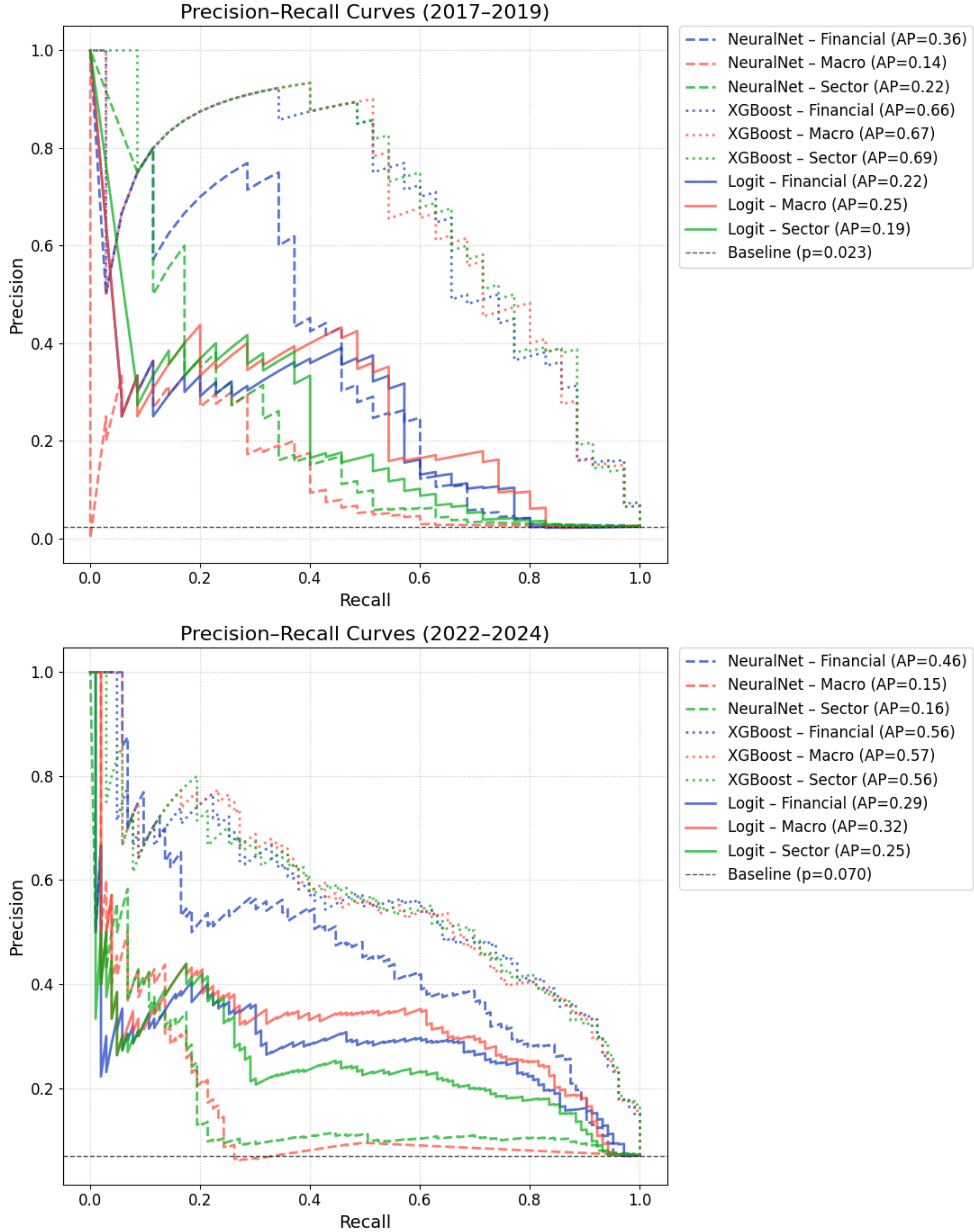

***Figure 4** Precision–Recall Curves for M1 and M2 Test Windows.*
This figure displays the precision-recall (PR) curves for each model–setup combination across two evaluation periods. The top panel (M1) shows results for the 2017–2019 test window (trained on 2015–2016), and the bottom panel (M2) for the 2022–2024 test window (trained on 2020–2021). The area under each curve (average precision) is reported in the legend. Dashed lines represent the baseline level of precision, which equals the proportion of distressed firms in each test set. PR curves are especially relevant in imbalanced datasets, as they emphasize the trade-off between correctly identifying distressed firms (recall) and avoiding false alarms (precision). These curves and average precision values correspond directly to the PR-AUC scores listed in Tables I and II.
***Source**: df_NN dataset.*

E.3 Calibration Curves

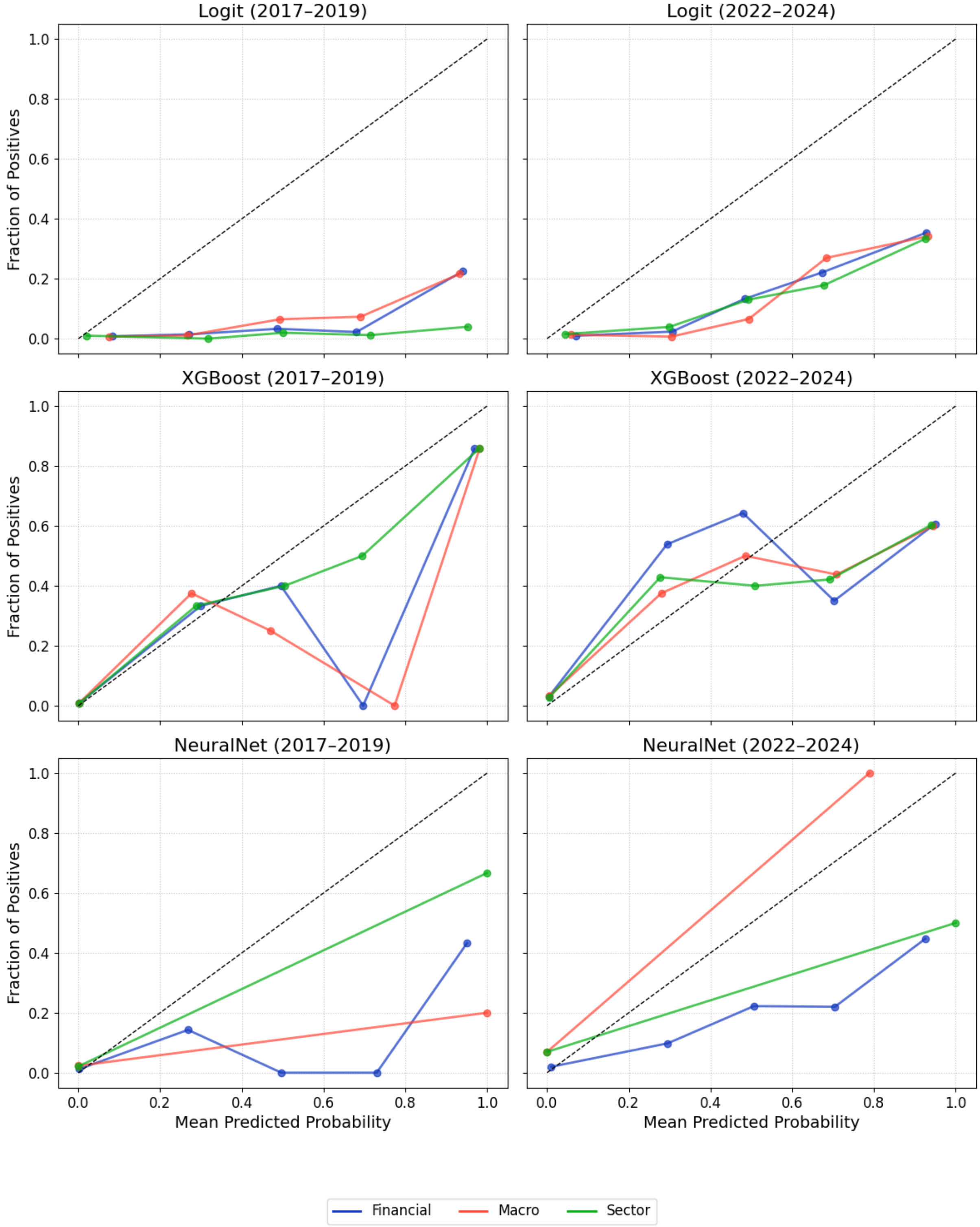

Figure 5 *Calibration curves by model and test window.*
This figure presents probability calibration curves for each model setup across the two evaluation periods. The left column shows results for the M1 window (2017–2019), and the right column for M2 (2022–2024). Each row corresponds to a different model family: Logit (top), XGBoost (middle), and NeuralNet (bottom). The x-axis shows the mean predicted probability of distress, while the y-axis indicates the actual fraction of positive outcomes in each bin. Perfect calibration corresponds to the diagonal dashed line. Curves above the diagonal suggest underconfidence; curves below indicate overconfidence. These plots align with Brier score results and provide insight into the reliability of predicted probabilities for each model and setup.
Source*: df_NN dataset.*

Appendix F: Feature Importance Results

F.1 Logistic Regression

Logistic regression models the probability that a firm-year observation is classified as distressed based on a linear combination of predictor variables. In this framework, the log-odds of distress (i.e., the natural logarithm of the odds ratio) is modeled as:

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

where p is the predicted probability of distress, x_i are the predictor variables, and β_i are the estimated coefficients. Each coefficient β_i represents the change in the log-odds of distress associated with a one-unit increase in the corresponding variable x_i, holding all other variables constant. Since all features are standardized in this analysis, a one-unit increase corresponds to a one standard deviation increase in the original scale of that variable. To facilitate interpretation, the coefficients are often transformed into odds ratios by exponentiation:

$$\text{Odds Ratio} = e^{\beta_i}$$

Each coefficient is accompanied by a p-value, calculated via the Wald test, which evaluates whether the estimated coefficient differs significantly from zero. The Wald statistic is computed as the square of the ratio of the coefficient to its standard error and follows a chi-squared distribution under the null hypothesis. A low p-value (typically < 0.05) suggests that the feature contributes meaningfully to the model, while high p-values indicate statistical insignificance. Although logistic regression does not capture nonlinearities or interactions, its transparency and coefficient-based structure make it useful for evaluating the directional influence and relative importance of predictors in a standardized setup. Coefficient values, odds ratios, and p-values for each model and feature set are reported in the tables below.

Variable	M1 Window Coefficient	M1 Window p-value	M1 Window Odds Ratio	M2 Window Coefficient	M2 Window p-value	M2 Window Odds Ratio
OpROA	-2.0589	0.243	0.1276	-9.5820	0.000	0.0001
OCF/Debt	0.9193	0.049	2.5075	2.1918	0.000	8.9513
OCFMrgn	-0.2501	0.047	0.7787	-0.4866	0.000	0.6147
ROE	-3.2847	0.000	0.0375	0.0131	0.869	1.0132
CFEarnRatio	-0.8024	0.000	0.4483	-0.1461	0.025	0.8641
EV/EBITDA	-10.6828	0.000	0.0000	-0.8790	0.063	0.4152
WkgCap/Assets	-3.6084	0.000	0.0271	-0.9885	0.000	0.3721
EquityGr	0.3233	0.130	1.3817	-0.1629	0.411	0.8497
Mkt/Book	-0.5845	0.049	0.5574	-0.0052	0.966	0.9948
QuickRatio	-0.0997	0.842	0.9051	-2.4684	0.004	0.0847
Any_NA_M1	-0.5622	0.503	0.5700	-1.3719	0.307	0.2536

Table XIII *Logistic regression coefficients, p-values, and odds ratios (Financial Setup).*
All predictors were standardized before model estimation. Coefficients represent the change in the log-odds of distress associated with a one standard deviation increase in each variable. Odds ratios are computed as e^{β} . The p-values are derived from the Wald test, assessing the statistical significance of each coefficient. Results are based on models trained using only financial ratios from Feature Set 1.
Source: *df_NN dataset.*

Variable	M1 Window Coefficient	M1 Window p-value	M1 Window Odds Ratio	M2 Window Coefficient	M2 Window p-value	M2 Window Odds Ratio
OpROA	-6.3172	0.000	0.0018	-11.2915	0.000	0.0000
OCF/Debt	0.1076	0.274	1.1136	1.1064	0.0003	3.0235
OCFMrgn	-2.1034	0.000	0.1220	-0.4861	0.000	0.6150
ROE	-2.9406	0.000	0.0528	0.0143	0.949	1.0144
CFEarnRatio	-0.5816	0.000	0.5590	-0.1350	0.028	0.8737
EV/EBITDA	-9.6855	0.000	0.0001	-0.6767	0.092	0.5083
WkgCap/Assets	-3.3381	0.000	0.0355	-0.9799	0.000	0.3753
EquityGr	0.5655	0.000	1.7603	-0.0638	0.743	0.9382
Mkt/Book	0.0945	0.345	1.0991	-0.0071	0.943	0.9929
QuickRatio	0.2990	0.049	1.3485	-2.1823	0.000	0.1128
Any_NA_M2	-13.9085	0.000	0.0000	-3.2773	0.000	0.0377
GDP	-0.0028	1.000	0.9972	-0.0139	1.000	0.9862
UNRATE	0.0028	1.000	1.0028	0.0139	1.000	1.0140
CPI	0.0028	1.000	1.0028	-0.0139	1.000	0.9862
FEDFUNDS	0.0028	1.000	1.0028	0.0139	1.000	1.0140
YCSPRED	-0.0028	1.000	0.9972	-0.0139	1.000	0.9862
CCI	-0.0013	1.000	0.9987	-0.0139	1.000	0.9862
STRFED	0.0011	1.000	1.0011	0.0128	1.000	1.0129
VIXSTDR	-0.0013	1.000	0.9987	0.0129	1.000	1.0130
VIXLOGR	-0.0016	1.000	0.9984	0.0129	1.000	1.0130

Table XIV *Logistic regression coefficients, p-values, and odds ratios (Macro Setup).*
The model includes both financial ratios from Feature Set 2 and general macroeconomic indicators from Table IV. Predictors were standardized before training; coefficients represent the effect of a one standard deviation increase on the log-odds of distress. Variables with a p-value of 1.0 were fully shrunk by L1 regularization and effectively excluded from the model. Odds ratios are computed as e^{β}.
Source*: df_NN dataset.*

	M1 Window			M2 Window		
Variable	**Coefficient**	**p-value**	**Odds Ratio**	**Coefficient**	**p-value**	**Odds Ratio**
OpROA	-6.8326	0.003	0.0011	-13.8814	0.000	0.0000
OCF/Debt	1.0054	0.012	2.7330	1.4524	0.000	4.2734
OCFMrgn	-2.8584	0.000	0.0574	-0.5177	0.000	0.5959
ROE	-3.7533	0.000	0.0234	-0.0166	0.790	0.9835
CFEarnRatio	-0.7326	0.000	0.4807	-0.1194	0.065	0.8875
EV/EBITDA	-10.6569	0.000	0.0000	-0.3632	0.150	0.6954
WkgCap/Assets	-4.3511	0.000	0.0129	-1.0899	0.000	0.3363
EquityGr	0.6829	0.000	1.9796	-0.0366	0.962	0.9641
Mkt/Book	0.0134	0.200	1.0135	-0.0020	0.985	0.9980
QuickRatio	0.4891	0.001	1.6308	-2.5126	0.000	0.0811
Any_NA_M2	-18.2474	0.000	0.0000	-3.9070	0.000	0.0201
GDP	-0.1437	1.000	0.8661	0.0000	-	1.0000
UNRATE	0.1437	-	1.1545	0.0000	1.000	1.0000
CPI	0.1437	1.000	1.1545	0.0000	1.000	1.0000
FEDFUNDS	0.1437	1.000	1.1545	0.0000	1.000	1.0000
YCSPRED	-0.1437	1.000	0.8661	0.0000	1.000	1.0000
CCI	-0.1437	1.000	0.8661	0.0000	-	1.0000
STRFED	0.1437	1.000	1.1545	0.0000	1.000	1.1545
VIXSTDR	-0.1437	1.000	0.8661	0.0000	1.000	1.0000
VIXLOGR	-0.1437	-	0.8661	0.0000	-	1.0000
BNKASST	-0.0047	1.000	0.9953	-0.0000	1.000	1.0000
DLQRATE	0.0855	1.000	1.0893	0.0000	1.000	1.0000
TCLOANOUT	-0.6545	1.000	0.5197	-0.1015	-	0.9035
BetaFIN	0.6933	1.000	2.0003	0.0000	1.000	1.0000
CONSPEND	0.3836	1.000	1.4676	0.0000	-	1.0000
HOUST	0.5296	-	1.6983	0.0000	-	1.0000
MORTGAGE	-0.3098	1.000	0.7336	0.1242	1.000	1.1322
IRCREPI	1.6402	-	5.1562	-0.2023	1.000	0.8168
CRELOAN	-0.3950	-	0.6737	-0.0338	1.000	0.9668
BetaREI	-0.5449	1.000	0.5799	0.3226	-	1.3807
INDPRO	0.1025	-	1.1079	0.0000	1.000	1.0000
CAPUT	0.0000	-	1.0000	0.0000	1.000	1.0000
BUSINV	0.0000	-	1.0000	0.0000	1.000	1.0000
DGORDER	-0.5118	-	0.5994	0.0642	1.000	1.0663
MANEMP	0.5733	-	1.7741	0.0000	-	1.0000
BetaIND	0.0805	-	1.0838	0.1182	-	1.1255
PPIACO	0.3033	1.000	1.3543	-0.9264	1.000	0.3960
IPMINE	0.2376	1.000	1.2682	0.7000	-	2.0138
GPIACO	0.2942	1.000	1.3421	-0.8928	-	0.4095
BetaMAT	-0.2860	1.000	0.7513	-1.7634	-	0.1715

Table XV *Logistic regression coefficients, p-values, and odds ratios (Sector Setup).*
This specification includes financial ratios from Feature Set 2 and all macroeconomic and sector-specific variables from Table IV. Predictors were standardized; coefficients indicate the effect of a one standard deviation increase on the log-odds of distress. Variables with no p-value or a p-value of 1.0 were excluded from the model due to L1 regularization shrinkage. Odds ratios are computed as e^{β}.
Source: *df_NN dataset.*

F.2 Extreme Gradient Boosting

To interpret the behavior of the XGBoost models, two complementary approaches were employed: built-in feature importance scores and SHAP values derived from the TreeExplainer function.

Feature importance in XGBoost reflects how frequently and how effectively a feature is used to split decision nodes across all trees in the model. The most commonly used metric is gain, which measures the average improvement in the model's objective function (e.g., log-loss) from splits involving that feature. Higher gain indicates greater overall influence on the model's predictions. These values are aggregated across trees and then normalized to create bar plots that summarize global feature relevance. However, these scores provide no insight into the direction (positive or negative) or the distribution of a feature's influence across observations.

To complement this, SHAP (SHapley Additive exPlanations) values are computed using TreeExplainer, a method specifically optimized for tree-based models. SHAP values apply principles from cooperative game theory to decompose each prediction into additive contributions from individual features. The SHAP value of a feature represents its marginal impact on the prediction relative to the average model output. Positive SHAP values indicate that the feature pushed the model toward classifying a firm-year as distressed, while negative values suggest the opposite.

SHAP summary plots visualize these values across all test observations. Each dot represents a single firm-year, plotted according to its SHAP value (horizontal axis). The color encodes the standardized value of the feature for that observation: red for high feature values, blue for low, and purple for intermediate values. This visual structure allows simultaneous assessment of a feature's magnitude, directional impact, and the distribution of effects across the sample.

Since there are a total of six XGBoost models (three setups and two evaluation windows), each producing both feature importance and SHAP plots, only two representative models are included in the appendix. These are the sector-specific setup from the M1 window (2017–2019) and the financial-only setup from the M2 window (2022–2024), selected to illustrate feature effects in both a high-performing and a broadly generalizable model.

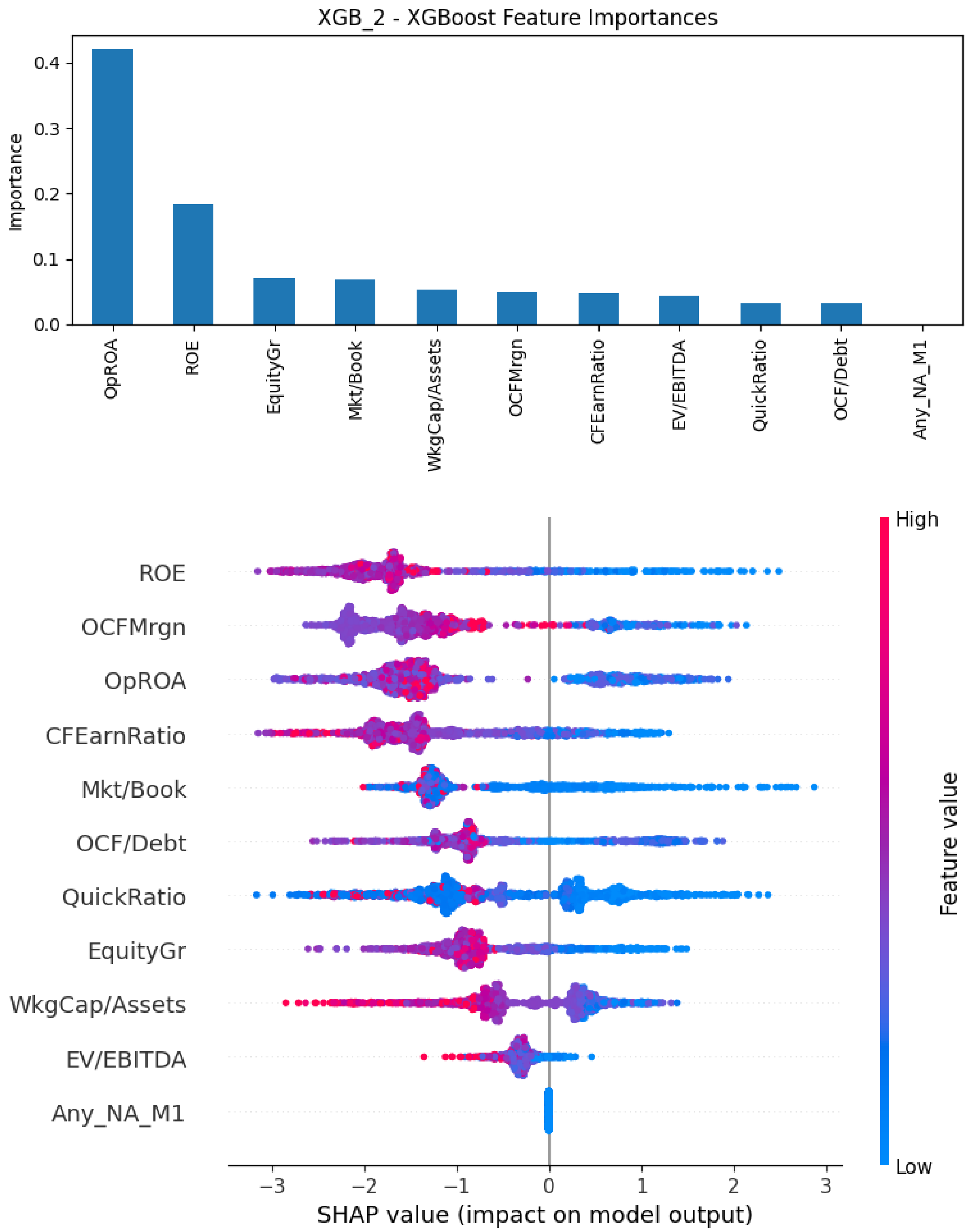

Figure 6a-b *XGBoost SHAP summary and feature importance plots (Financial Setup M2)*
6a (top): *SHAP summary plot displaying the marginal effect of each feature on the model's predicted output. Each dot represents a firm-year observation, colored by feature value (red = high, blue = low). The horizontal axis shows the SHAP value, indicating the feature's impact on increasing (right) or decreasing (left) the probability of distress.* ***6b (bottom):*** *Feature importance scores based on the frequency and gain of splits across all decision trees. Higher values reflect a greater contribution to the model's overall decision-making.*
These plots correspond to the M2 evaluation window using the financial-only feature set and are included as one of two illustrative XGBoost model interpretations.

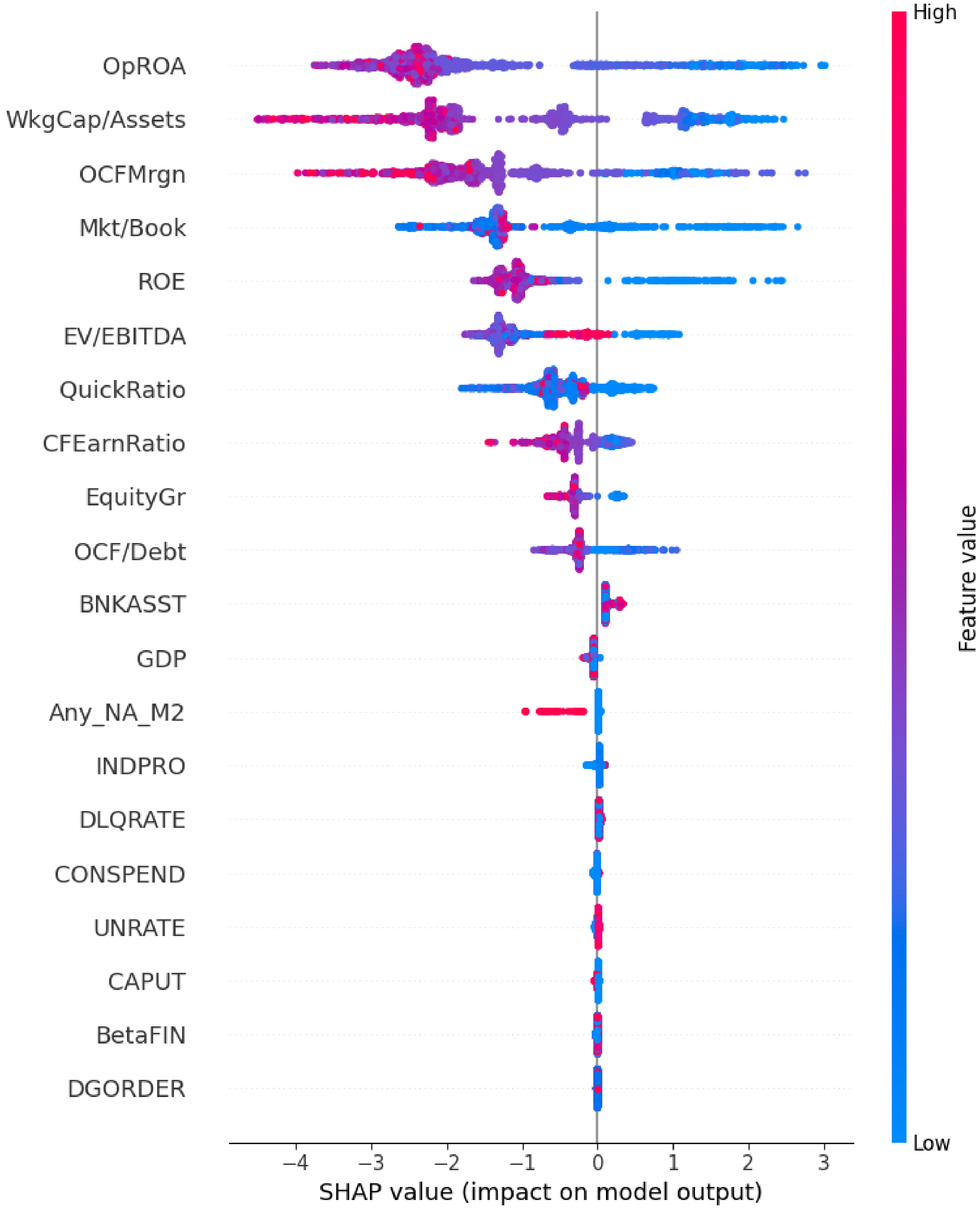

Figure 7a-b *XGBoost SHAP summary and feature importance plots (Sector Setup M1).*
7a (top): *SHAP summary plot displaying the marginal effect of each feature on the model's predicted output. Each dot represents a firm-year observation, colored by feature value (red = high, blue = low). The horizontal axis shows the SHAP value, indicating the feature's impact on increasing (right) or decreasing (left) the probability of distress.* ***7b (bottom, next page):*** *Feature importance scores based on the frequency and gain of splits across all decision trees. Higher values reflect a greater contribution to the model's overall decision-making.*
These plots correspond to the M1 evaluation window using the sector-augmented feature set and are included as the second of two illustrative XGBoost model interpretations.

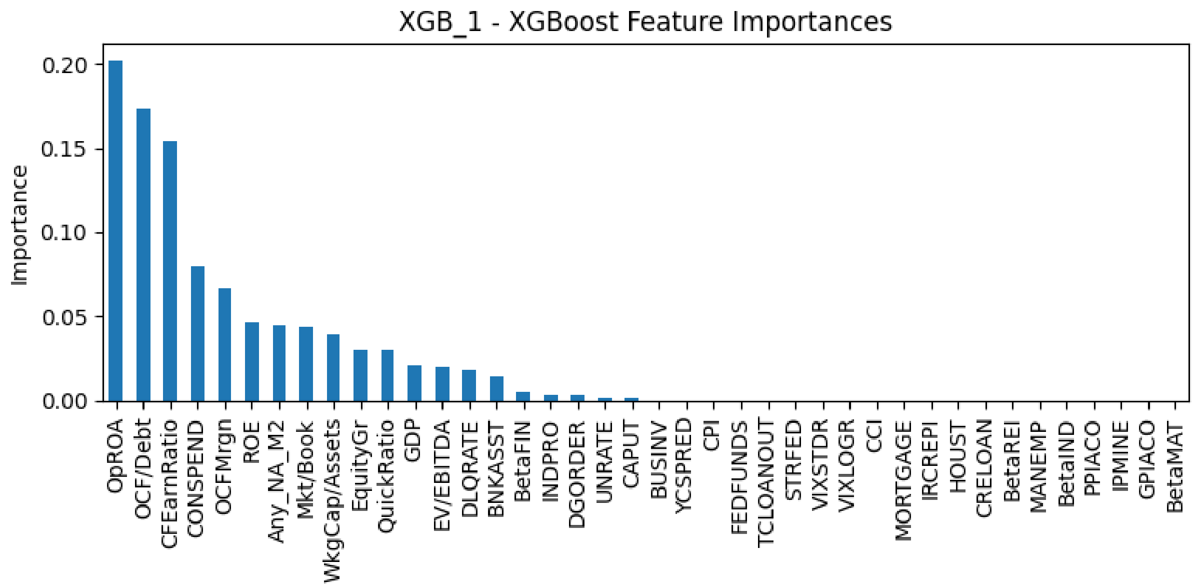

Figure 7b *(caption is on the previous page)*

F.3 Feedforward Neural Network

For neural networks, SHAP values were calculated using DeepExplainer, a method adapted to the internal structure of deep learning models. DeepExplainer combines background sampling with layer-wise relevance propagation to efficiently approximate Shapley values without requiring retraining or perturbing inputs. Unlike tree models, where importance can be derived from structure, neural networks lack inherent interpretability, making SHAP one of the few reliable tools to quantify how input features influence predictions. Traditional alternatives like weight magnitudes or gradient-based saliency maps fail to capture nonlinear interactions or offer directionality, whereas SHAP provides consistent, locally accurate explanations that reflect the model's actual decision logic.

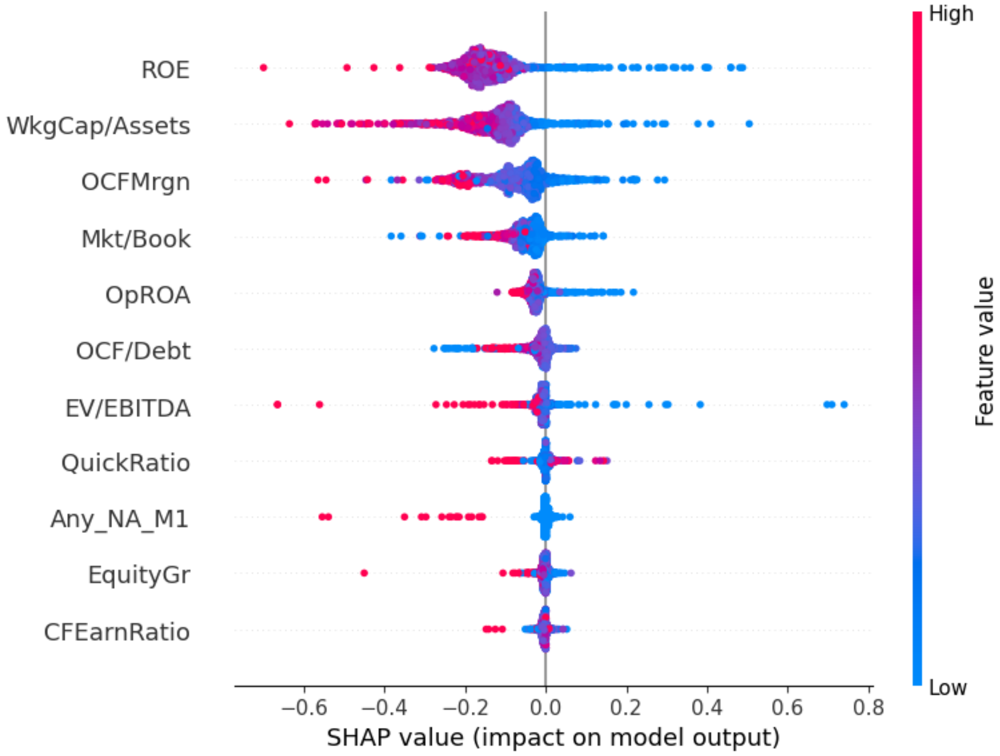

Figure 8 *Neural network SHAP summary plot (Financial Setup, M1).*
This SHAP summary plot illustrates the feature-level impact on predicted financial distress probabilities for the neural network model using the financial-only feature set during the M1 (2017–2019) evaluation window. Each dot represents a firm-year observation, with position on the x-axis indicating the SHAP value (positive = higher predicted distress). Color reflects the feature's standardized value (red = high, blue = low), revealing directional influence. This plot provides insight into how the neural network internally weighs features, despite its non-transparent structure.