Machine Learning Approaches to Corporate Failure Prediction

Angelina Schmitt

Dissertation written under the supervision of
Professor Dan Tran

Dissertation submitted in partial fulfillment of requirements for the Master in Finance, at the Universidade Católica Portuguesa, September 11th, 2025.

Abstract

Corporate failure prediction is a central topic in financial literature due to the substantial costs associated with firm failures. At the same time, advances in machine learning have introduced powerful new tools for addressing this challenge. This thesis examines the predictive performance of traditional and machine learning models in forecasting corporate failures, with an evaluation across different prediction horizons.

Two iterations of datasets were employed: one based on the established Campbell (Campbell et al., 2008) variables and another extended dataset incorporating accounting, market, and macroeconomic information. Firms were observed over consecutive months until disappearance, following a hazard model framework (Shumway, 2001). Logistic regression served as the baseline and was compared against random forest and extreme gradient boosting (XGBoost). Model performance was assessed using AUC and other evaluation metrics and feature importance was analyzed to shed light on the main drivers of corporate failure risk.

The results indicate that Logistic Regression with Campbell variables performs well over short horizons (one month). However, expanding the feature set and applying machine learning models, particularly ensemble models, substantially improves predictive accuracy and ensures greater stability across longer horizons. These findings underscore the importance of combining richer datasets with advanced algorithms, offering both theoretical contributions to the literature and practical implications for early warning systems, risk management, and corporate monitoring.

Author: Angelina Schmitt

Title: Machine Learning Approaches to Corporate Failure Prediction

Keywords: Machine Learning, Corporate Failures

Resumo

A previsão de falências corporativas é um tema central na literatura financeira devido aos custos substanciais associados às falências empresariais. Ao mesmo tempo, os avanços em machine learning introduziram novas poderosas ferramentas para enfrentar este desafio. Esta dissertação examina o desempenho preditivo de modelos tradicionais e de machine learning na previsão de falências corporativas, avaliados em diferentes horizontes de previsão.

Foram utilizadas duas iterações de bases de dados: uma baseada nas variáveis estabelecidas por Campbell (Campbell et al., 2008) e outra alargada, que incorpora informações contabilísticas, de mercado e macroeconómicas. As empresas foram observadas ao longo de meses consecutivos até a sua saída do mercado, seguindo uma estrutura de modelo de risco (hazard model) (Shumway, 2001). A regressão logística foi utilizada como linha de base e comparada com random forest e extreme gradient boosting (XGBoost). O desempenho dos modelos foi avaliado através do AUC e de outras métricas, e a importância das variáveis foi analisada para identificar os principais determinantes do risco de falência corporativa.

Os resultados indicam que a regressão logística com as variáveis de Campbell apresenta bom desempenho em horizontes curtos (um mês). No entanto, a ampliação do conjunto de variáveis e a aplicação de modelos de machine learning, em particular os modelos de ensemble, melhoram substancialmente o poder preditivo e asseguram maior estabilidade em horizontes mais longos. Estes resultados reforçam a relevância de combinar dados mais ricos com algoritmos avançados, oferecendo contribuições teóricas e implicações práticas para sistemas de alerta precoce, gestão de risco e monitoramento corporativo.

Autora: Angelina Schmitt

Título: Abordagens de Machine Learning para a Previsão de Falências Corporativas

Palavras-chave: Machine Learning, Falências Corporativas

Preface

Completing this thesis has been a long journey that stretched beyond initial plans. It would not have been possible without the patience, encouragement, and support I received along the way.

I would like to express my gratitude to my supervisor, Professor Dan Tran, for his guidance and valuable feedback throughout the research process. I am also thankful to the program director for the Master in Finance, José Faias, and to the Master Affairs office, in particular Carla Rocha, as well as the Universidade Católica Portuguesa as a whole, for their understanding and continuous support during this period.

My deepest thanks go to my dear family and friends. To my partner Philipp, to my parents and my brother Francesco and to my best friends Marie and Anna: thank you for your unwavering encouragement, care, and patience. Your support was not limited to the thesis itself, but extended far beyond, providing me with both emotional and practical help whenever I needed it most.

This thesis is as much a reflection of your support as it is of my own work, and I am profoundly grateful.

Contents

List of Tables

List of Figures

1 Introduction

Default prediction has been a longstanding research topic in finance, with interest intensifying during economic crisis. Accurate default prediction helps mitigate financial losses, improves capital allocation, and strengthens financial stability (Agarwal & Taffler, 2008). Many economic and societal stakeholders, such as lenders, investors or governments have a strong interest in enhancing its accuracy (Barboza et al., 2017).

The field gained prominence in the 1960s. Over the past decades, researchers made significant progress in identifying which variables most effectively predict default and which modeling techniques offer the best performance. This progress has been driven by improved data availability, the construction of larger datasets, and advances in computational power (Alonso & Carbó, 2021).

A particularly important recent development has been the emergence of Artificial Intelligence (AI) and Machine Learning (ML) techniques. These models capture nonlinear relationships and often outperform traditional statistical methods (Alonso & Carbó, 2021; Barboza et al., 2017; Jones et al., 2017).

Seminal models introduced by Shumway (2001) and Campbell et al. (2008) demonstrated that combining accounting and market variables significantly enhances predictive power. Further studies have emphasized the importance of macroeconomic conditions in predicting defaults (Bonfim, 2009; Pesaran et al., 2006). Building on these findings, Nam et al. (2008) extended the model of Shumway (2001) by incorporating macroeconomic variables, which notably improved performance (Nam et al., 2008).

This thesis builds on established approaches to corporate default prediction by combining accounting, market, and macroeconomic variables within modern machine learning frameworks. The primary focus is on ensemble methods — Random Forest and Extreme Gradient Boosting (XGBoost) — which have demonstrated strong predictive performance in prior studies (Alonso & Carbó, 2021; Barboza et al., 2017; Jones et al., 2017). Logistic Regression is employed as a benchmark, providing a linear baseline for comparison. In addition, the well-known Campbell specification (Campbell et al., 2008), which relies on a restricted set of accounting and market variables, serves as a second reference point. Extending beyond this baseline, the present

study incorporates a broader range of firm-level market and accounting characteristics together with macroeconomic indicators, thereby assessing whether richer information sets improve the accuracy and robustness of default prediction models.

Overall, the results show that Logistic Regression with the Campbell variable set already delivers strong predictive accuracy at short horizons ($t + 1$). However, expanding the information set to include additional accounting, market, and macroeconomic variables, together with the use of ensemble methods, substantially improves performance, particularly at medium- and long-term horizons ($t + 6$ and $t + 12$).

The remainder of this thesis is structured as follows. Chapter 2 presents a literature review, providing the theoretical and empirical background on corporate failure prediction. Chapter 3 introduces the dataset, detailing the construction of firm-level and macroeconomic variables as well as the procedures for variable selection. Chapter 4 presents the modeling framework, beginning with Logistic Regression and the Campbell specification as baselines, before turning to ensemble methods. Chapter 5 reports the empirical results, including predictive performance across horizons and an analysis of feature importance. Chapter 6 provides a discussion of these findings in light of the broader literature. Finally, Chapter 7 concludes with a summary of the main insights, limitations, and avenues for future research.

2 Literature Review

2.1 Bankruptcy Prediction Models

Bankruptcy prediction has been a central topic in finance for decades, attracting extensive research attention. Early studies by Beaver (1966) and Altman (1968) relied on financial ratios. Altman's Z-score, which applies multiple discriminant analysis (MDA) to five ratios, remains widely used today. Later, Ohlson (1980) introduced the O-score, a logit-based model that also became a standard benchmark.

A major shift came with the work of Black and Scholes (1973) and Merton (1974), who introduced market-based variables for firm valuation. In Merton's framework, which extends the model by Black and Scholes (1973), firm value measures can be inferred directly from market data. This market data reflects information embedded in financial statements (Agarwal & Taffler, 2008). Unlike accounting data, which is reported infrequently and with delays, market data is available in real time. Accounting data also depends on reporting standards, is susceptible to managerial manipulation, is created on a going-concern basis and reflects past performance. In contrast, market data captures investor expectations of future cash flows (Agarwal & Taffler, 2008).

Nevertheless, Agarwal and Taffler (2008) show that accounting- and market-based models perform similarly in predictive accuracy. However, traditional accounting-based models, such as Taffler's Z-score (Taffler, 1984) (based on the Z-Score by Altman (1968)), often prove more cost-efficient than market-based ones, which rely on restrictive assumptions. In practice, each type of model provides useful insights, but neither is sufficient alone for reliable default prediction (Agarwal & Taffler, 2008).

Shumway (2001) advanced the field with a discrete-time hazard model. This approach introduced two innovations. First, it combined accounting and market variables in a single framework. Second, it exploited time-varying information by treating each firm-month as an observation and estimating the probability of default at time (t), conditional on survived until $(t-1)$. Firms contribute observations until they default or disappear, allowing the model to capture changing conditions over time. Shumway's model outperformed earlier approaches

out-of-sample (Shumway, 2001).

Campbell et al. (2008) confirmed the strength of the hazard framework using a slightly different set of variables. Later, Bauer and Agarwal (2014) compared hazard, accounting-based, and contingent claims (market-based) models, finding that the hazard model not only delivered stronger economic performance but also subsumed the information contained in the alternatives.

In addition to accounting and market data, macroeconomic variables also shape default risk. Studies such as Bonfim (2009) and Pesaran et al. (2006) show that indicators like GDP and interest rates influence default probabilities. Tinoco and Wilson (2013) integrated accounting, market, and macroeconomic factors in a combined model. While the time fixed effects in hazard models already capture some macroeconomic variation, they did not explicitly include macroeconomic variables. Nam et al. (2008) extended Shumway's framework by explicitly modeling temporal and macroeconomic dependencies. This addition improved predictive performance, underscoring the importance of integrating firm-level and macroeconomic perspectives.

2.2 Variables

Since reproducibility and transparency are priorities in implementing the firm-level perspective, we follow the methodology of Jensen et al. (2023), who examined whether academic finance literature faces a 'replication crisis'. They reached the opposite conclusion, showing that a majority of research characteristics can, in fact, be replicated. To support this claim, they meticulously documented their procedures in detail and published replication instructions with direct links to WRDS (Wharton Research Data Services) (Jensen et al., 2022). This resource was used to collect data on 153 firm-specific characteristics. The variables are constructed with market and accounting data and cover a wide set of categories, including Investment, Value, Trading Frictions, Profitability, and Intangibles.

The characteristics span a broad range of firm-level information. Investment measures capture capital allocation decisions through variables such as capital expenditures (CAPEX), debt and equity issuance, asset growth, and accruals. Value characteristics focus on standard valuation ratios, including book-to-market equity, earnings-to-price, and dividend yield. Trading frictions capture market structure and risk, with metrics like volatility, beta, skewness, turnover, and bid-ask spreads. Profitability characteristics reflect performance, drawing on profits-to-assets, return on equity, operating cash flow, and profit margins. Intangible factors include R&D spending, liquidity, seasonality, and operating leverage. Additional characteristics extend beyond these groups. They include measures of mispricing, momentum, firm size, and age. Together, the full set of variables provides a rich picture of firm fundamentals, market behavior, and structural characteristics (Jensen et al., 2022).

2.3 Machine Learning

In their bibliometric review, Ahmed et al. (2022) observe that Artificial Intelligence (AI) has become an increasingly dominant techonology in finance, attracting significant research interest and investment. AI methods capture both linear and nonlinear patterns in financial data and have played a key role in improving bankruptcy prediction models. These improvements matter because bankruptcies impose heavy costs on both the economy and society (Ahmed et al., 2022).

Early applications of machine learning in bankruptcy prediction included Logistic Regression and Support Vector Machines (SVM) (Jabeur et al., 2021). More recent studies, such as Jones et al. (2017), demonstrate that ensemble methods often deliver stronger predictive performance. Ensemble models combine multiple individual models, called learners, to create more robust predictions and reduce both bias and variance (Jabeur et al., 2021). They also tend to be less sensitive to missing data and, are easier to interpret compared to neural networks (Jones et al., 2017). There are two major types of ensemble techniques: bagging and boosting.

Bagging, short for Bootstrap Aggregating, trains multiple models in parallel on random subsets of the data and then aggregates their predictions, typically by majority voting or averaging (Barboza et al., 2017). The most common bagging method in bankruptcy prediction is Random Forest, an ensemble of decision trees. Each tree is trained on a subset of data and features, making the model resistant to overfitting and noise while maintaining a level of interpretability (Barboza et al., 2017; Jabeur et al., 2021). Barboza et al. (2017) find that Random Forests outperform many other models in bankruptcy prediction.

Boosting builds models sequentially, with each new model correcting the errors of its predecessors. Final predictions result from a weighted combination of all models, with higher weights given to stronger learners (Jones et al., 2017). Like bagging, boosting is robust to noisy data and can remain interpretable. Early research used Adaptive Boosting (AdaBoost) (Barboza et al., 2017), while more recent work highlights Gradient Boosting algorithms such as Extreme Gradient Boosting (XGBoost) (Jabeur et al., 2021). XGBoost is well known for its computational efficiency and regularization features and has proven highly effective in default prediction (Alonso & Carbó, 2021).

This thesis evaluates two ensemble methods: Random Forest as a representative of bagging and XGBoost as a boosting approach. Their performance is compared against Logistic Regression with the Campbell specification (Campbell et al., 2008), which serves as a transparent and widely used baseline. This comparison allows for an assessment of whether ensemble methods, combined with broader sets of accounting, market, and macroeconomic variables, extend predictive power beyond the traditional, transparent and linear Campbell framework.

3 Data

3.1 Dataset Overview

This study utilizes a comprehensive dataset of 25,558 publicly traded U.S. firms from 1965 to 2022, totaling 3 million firm-year observations. The dataset integrates three groups of variables:

- **Firm-specific characteristics:** 153 accounting and market characteristics (independent variables)
- **Macroeconomic variables:** 6 indicators (independent variables)
- **Corporate failure indicator:** A binary variable capturing both bankruptcies and liquidations (dependent variable)

The dataset is highly unbalanced, with corporate failures representing less than 0.01% of observations. To address this, we built a data pipeline to improve quality and reproducibility, summarized in Figures 3.1 and 3.4. Variable selection follows established determinants of corporate failure, enabling analysis of accounting, market, and macroeconomic drivers of bankruptcy risk over several decades (Nam et al., 2008; Tinoco & Wilson, 2013).

3.2 Data Sources

3.2.1 Independent Variables: Firm-level Characteristics

The firm-specific variables used in this study consist of 153 characteristics that are widely employed in the asset-pricing and corporate finance literature. Professor Dan Tran, who also serves as the supervisor for this thesis, and his research team at the Católica Lisbon School of Business and Economics provided the data following Jensen et al. (2022). The characteristics draw on both market and accounting data and cover categories such as valuation ratios, profitability measures, and investment indicators.

The data were retrieved via the Wharton Research Data Services (WRDS) platform from two primary sources: Center for Research in Security Prices (CRSP) and Compustat. Observations from these sources were linked using the CRSP/Compustat Merged (CCM) linking table to ensure accurate alignment of market and accounting information. The linkage relies on three key identifiers:

- `gvkey` → Compustat firm identifier
- `permno` → CRSP security identifier
- `permco` → CRSP company identifier

We construct a custom firm-level identifier (`firm_id`) from unique `gvkey_permco` pairs and use it as primary firm identifier throughout the empirical analysis. The dataset covers all U.S. listed firms with a CRSP share code (`SHRCD`) of 10 or 11, which denote ordinary common shares traded on U.S. exchanges (Center for Research in Security Prices (CRSP), n.d.). The panel spans January 1965 to December 2022 at a monthly frequency, using end-of-month (`eom`) calendar dates as the time identifier.

The dataset includes industry classification based on Standard Industrial Classification (SIC) codes. Prior to exploratory data analysis, we grouped these SIC codes into both 1–4 and 1–10 categories following the approach of Chava and Jarrow (2004). These groupings allow parsimonious control of industry effects in the modeling stage. The full industry mapping and category definitions are provided in Table A.1 in the appendix.

3.2.2 Independent Variables: Macroeconomic Variables

The second set of independent variables consists of macroeconomic indicators, which serve as proxies for broad economic conditions. We obtained the value-weighted return of the S&P 500 from CRSP and a selected set of macroeconomic indicators, including the Consumer Price Index (CPI), Gross Domestic Product (GDP), long- and short-term interest rates, and the unemployment rate from the FRED (Federal Reserve Bank of St. Louis) API. Table A.2 in the appendix provides additional information about the variables. To ensure consistency with the firm-level panel, all time series were aligned to a monthly frequency and indexed using end-of-month (`eom`) timestamps.

Prior research motivates the inclusion of these variables. Bonfim (2009) highlight the predictive relevance of equity returns and GDP growth. Pesaran et al. (2006) use inflation and interest rates in forecasting, Mare (2015) include unemployment, and Tinoco and Wilson (2013) account for inflation-adjusted short-term interest rates. Other indicators such as coincident indicators, loan growth (Bonfim, 2009), oil prices (Pesaran et al., 2006), and VIX (Ang et al., 2006) were

considered but lacked consistent coverage for the full sample period.

We applied several transformations to enhance the relevance of the macroeconomic series and calculated the GDP growth rate (Bonfim, 2009), the slope of the yield curve (long- minus short-term interest rates) (Bluwstein et al., 2023), the inflation rate, and the inflation-adjusted short-term interest rate (Tinoco & Wilson, 2013). The transformed macroeconomic dataset includes the following variables:

- `gdp_growth`,
- `inflation`,
- `sp500_ret`,
- `st_interest`,
- `st_interest_inflation_adj`,
- `unemployment`,
- `yield_slope`

3.2.3 Dependent Variable: Bankruptcy Indicator

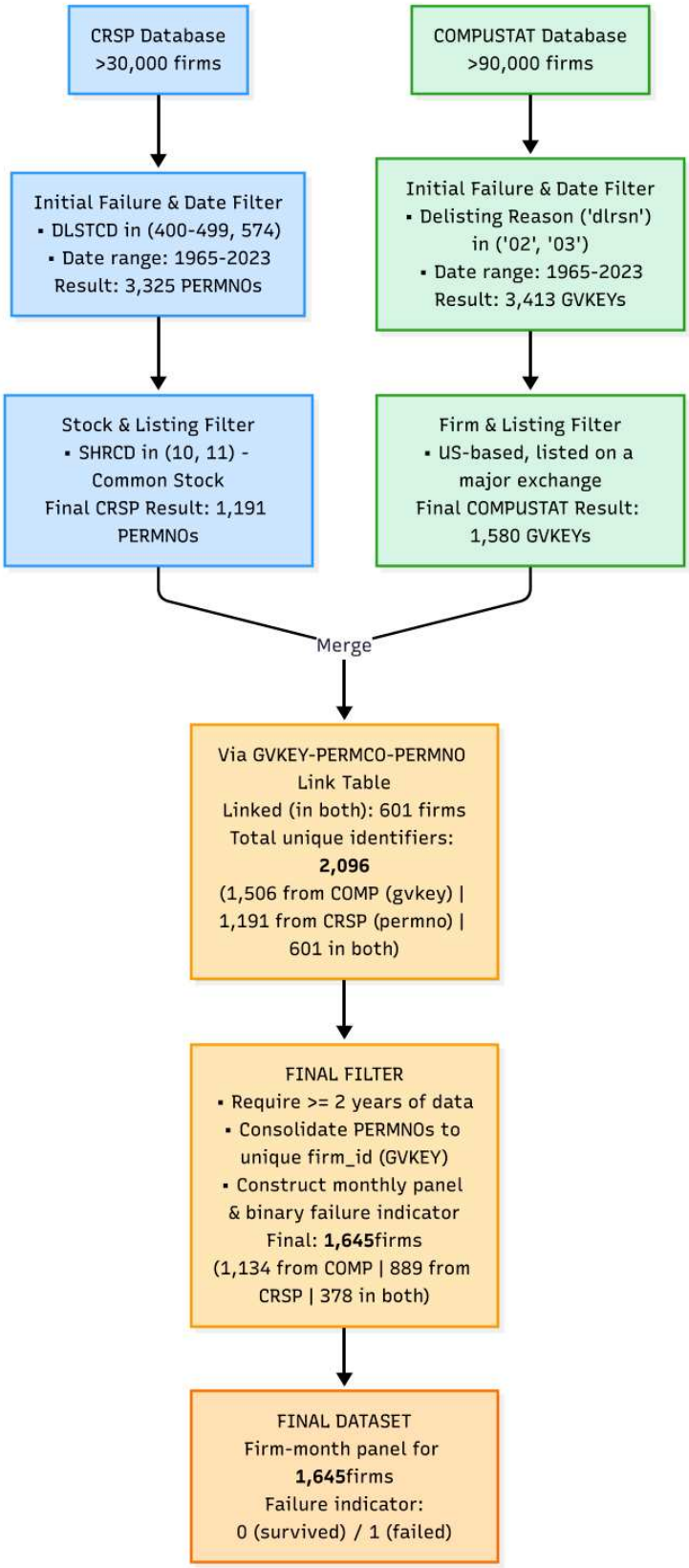
We constructed the dependent variable to indicate whether a firm experienced bankruptcy or liquidation during the sample period. We combined event data from Compustat and CRSP as follows:

- Compustat: `gvkey` identifier, deletion reason (`dlrsn`), and deletion date (`dldte`) → Firms with `dlrsn` equal to `02` (bankruptcy) or `03` (liquidation) were retained.
- CRSP: `permno` identifier, delisting code (`DLSTCD`), and delisting date `DLSTDT`) → Firms with `DLSTCD` in the 400 – 499 range (liquidation) or equal to `574` (bankruptcy) were selected.

We aligned firms to the same panel used for independent firm-level characteristics. We matched events to `firm_id` using the three-way link of `gvkey`, `permco`, and `permno` and resolved any conflicts with event dates and manual review. We then consolidated multiple listing attributes and retained firms with at least two years of consistent data, producing a final sample of 1,645 firms (Figure 3.1).

Figure 3.1

Pipeline of the Construction of the Corporate Failure Dataset



3.3 Data Cleaning

We identified several data issues during preparation, mainly temporal gaps and duplicate rows for the same `firm_id`, `eom` pair. These problems came from mismatches among firm and security identifiers (`gvkey`, `permco`, `permno`). To build a complete firm-month panel, we inserted missing months (up to a 12-month gap) using forward-fill; for longer gaps, we kept only the longest or most relevant continuous segment (typically close to a failure event).

For identifier inconsistencies, we found that firms very rarely had multiple rows for the same period mostly due to stock structures. Instead of dropping these cases, that occur proportionally more often in the minority class, we computed a market equity-weighted average of duplicates and added an `issues` column to flag consolidation events for transparency.

We then filtered out entities with short histories (less than 24 months of data), ensuring every included firm had enough observations for reliable modeling. The final cleaned dataset is a well-structured monthly panel: it has no temporal gaps, merged and harmonized identifiers, and sufficient data coverage across firms and time.

3.4 Merging Firm-Level Characteristics and Failure Data

The final sample consists of **1,645** distinct `firm_id` observations with recorded corporate failures, encompassing both bankruptcies and liquidation-related delistings. We define the event date as the earlier of the CRSP delisting date (`DLSTDT`) and the Compustat deletion date (`dldte`), and merge it with the firm-level dataset at the last available information of each `firm_id`, `eom`), creating a binary indicator that equals '1' if the firm has a corporate failure at time $t + 1$.

3.5 Selection and Exploration of Firm-Level Characteristics

After assembling the full dataset, we split the observations into training (50%), validation (20%), and testing (30%) subsets before conducting any exploratory data analysis or feature selection. We performed all variable exploration, preprocessing, feature engineering, and model development exclusively on the training set, ensuring that the test set remained untouched until the final evaluation.

The extensive exploratory data analysis (EDA) examined distributions, correlations, and missing data patterns across numeric, ordinal, and categorical variables. Results guided subsequent feature selection and model design and are detailed in the following subsections.

3.5.1 Baseline Feature Construction (Campbell variables)

We constructed a baseline feature set following the variable definitions in Campbell et al. (2008) for bankruptcy prediction. Each Campbell variable was matched to the closest equivalent in our dataset:

- **CASHMTA** (cash ratio from most recent quarter) → `cash_at` (cash-to-assets ratio)
- **EXRETAVG** (geometrically weighted average of returns) → constructed from `ret_1_0` (short-term reversal) and `sp500_ret` (value-weighted market return)
- **MB** (market-to-book at end of month $t-1$) → `be_me` (book-to-market ratio)
- **NIMTAAVG** (geometrically weighted average of profitability) → constructed from `niq_at` (quarterly return on assets)
- **PRICE** (stock price at end of month $t-1$) → constructed from `prc` (price per share)
- **RSIZE** (relative size: market capitalization at end of month $t-1$) → constructed from `me_company` (total company market equity)
- **SIGMA** (volatility over past three months) → constructed from `rvol_21d` (return volatility)
- **TLMTA** (leverage from most recent quarter) → `debt_me` (debt-to-market equity ratio)

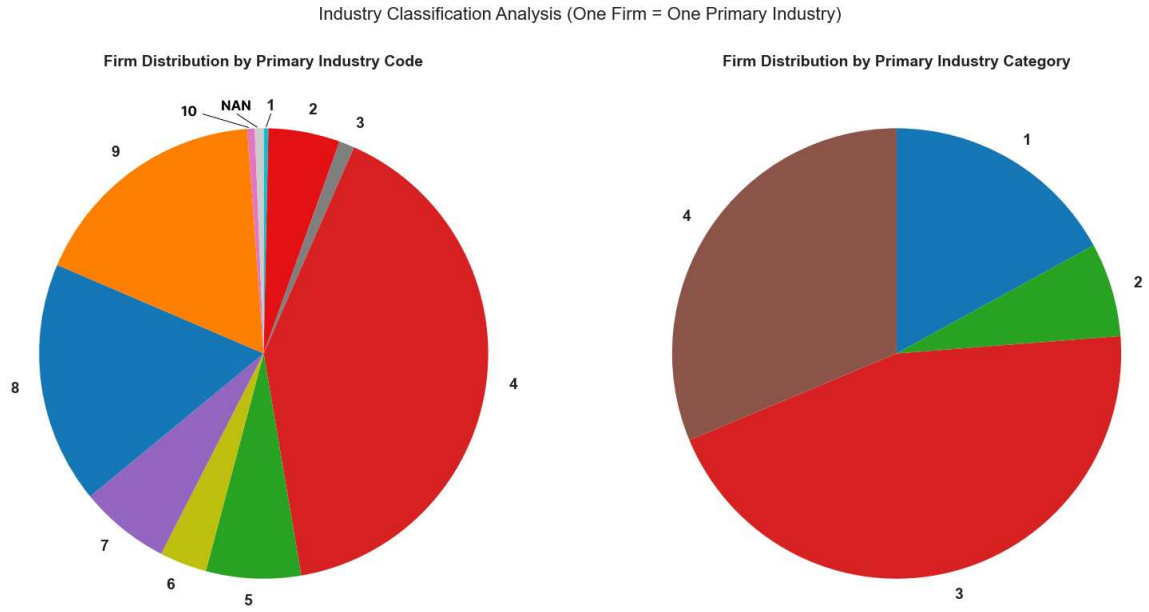
Following Campbell et al. (2008), we capped price at \$15 (`prc_capped`), transformed size into 5 categorical bins (`me_binned`, now an ordinal variable), and computed geometric weighted averages for profitability (`niq_at_avg`) and excess returns (`exc_ret_avg`). To approximate the three-month volatility metric used by Campbell, we averaged monthly volatility over three months (`rvol_21d_3m`). These variables serve as a robust baseline against which to compare additional features.

3.5.2 Categorical Variable: Industry

Two versions of industry classification were derived from SIC codes: a finer-grained 1–10 industry code and a broader 1–4 category grouping, following the scheme proposed by Chava and Jarrow (2004). Figure 3.2 shows the distribution of the industry classes.

Figure 3.2

Distribution of Firms by Industry Codes and Categories



To determine which version to retain, we performed chi-squared tests of independence using the dependent variable. Table A.3 in the appendix reports these results, showing that the 1–10 industry code has stronger associations with the outcome variable in both datasets, with higher chi-squared statistics and smaller p-values. Therefore, the 1–10 classification was retained for all subsequent analysis and modeling.

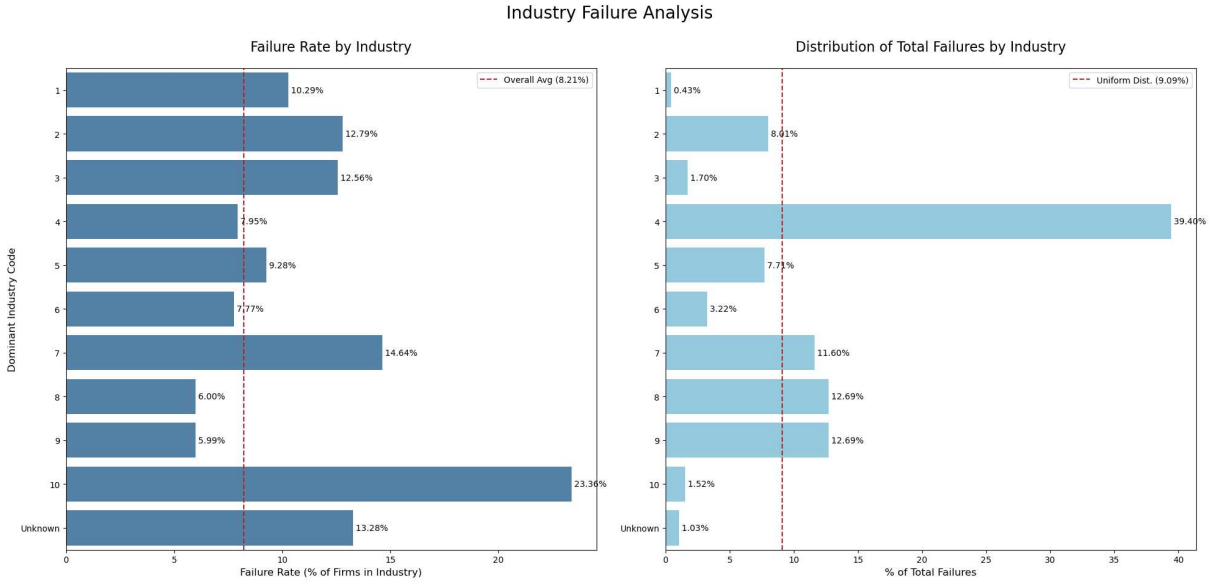
Figure 3.3 offers two perspectives on corporate failure rates across industry sectors. Public Administration (Industry 10) shows the highest risk (23.36%), though based on a small sample (107 firms, 25 failures post–2008), making it susceptible to variability. Retail Trade (14.64%) and unknown classifications (no code available) (13.28%) also exhibit elevated risk, while Finance/Real Estate (6.00%) and Services (5.99%) appear relatively stable.

Manufacturing (Industry 4) accounts for the largest share of failures (39.4%), despite a below-average failure rate (7.95%), reflecting its dominant presence in the dataset. Conversely, Public Administration's high individual risk translates into only 1.52% of total failures, consistent with its small size (Figure 3.2).

These findings show that high-risk sectors are not necessarily the largest contributors to aggregate failures. The distribution aligns broadly with Chava and Jarrow (2004), with fewer failures in Agriculture (Sector 1) and more in Finance/Real Estate (Sector 8). The appearance of Public Administration failures, absent from Chava and Jarrow (2004), likely reflects post–2008 economic and COVID–19–related distress, though its instability limits its overall impact.

Figure 3.3

Corporate Failure Rate by Industry



3.5.3 Descriptive Statistics and Missingness

We calculated descriptive statistics for all numeric firm-level characteristics to assess central tendency, spread, and potential outliers. Variables with more than 25% missing data were excluded from the analysis. For the remaining variables, we examined missing data patterns in detail and found that earlier years contained more missing observations, reflecting lower data quality in the past. Variables with widespread missingness were flagged for further preprocessing.

3.5.4 Multicollinearity Assessment and Variable Reduction

We computed a pairwise Pearson correlation matrix to assess multicollinearity and flagged pairs with absolute correlations above 0.70 for further review. Strong correlation clusters appeared among related accounting variables in the investment category. For example, `ncol_gr1a` (change in noncurrent operating liabilities) and `oaccruals_at` (operating accruals) showed very high positive correlation (0.9954), while `fnl_gr1a` (change in financial liabilities) and `nfna_gr1a` (change in net financial assets) showed very high negative correlation (-0.9999), reflecting underlying consistency in accounting measures. Similarly, for volatility and skewness, alternatives such as `ivol_capm_21d` and `ivol_ff3_21d` were strongly correlated (0.9946), indicating redundancy. Only representative measures were retained.

To decide which variables to keep, we considered three criteria: (i) missing data patterns, (ii)

theoretical relevance, and (iii) alignment with prior literature. We aimed to preserve key accounting concepts such as leverage, liquidity, profitability, operating efficiency, working capital, and accruals, alongside market features like volatility, returns, book-to-market ratios, and issuance, which have strong support in bankruptcy literature (Altman, 1968; Campbell et al., 2008; Ohlson, 1980; Shumway, 2001).

It is worth noting that modern machine learning algorithms are generally robust to correlated predictors and can accommodate multicollinearity without severe performance degradation. Accordingly, the primary goal of this screening was not to eliminate all correlated variables, but rather to avoid retaining features that carry essentially the same information content.

3.5.5 Variable Ranking and Selection

After removing 58 variables due to missingness and multicollinearity, we evaluated the remaining 96 numeric characteristics for their predictive relevance using three complementary methods:

- **Mann–Whitney U test**, to determine if variable distributions differ significantly between failing and non-failing firms.
- **Lasso regression**, applied within a preprocessing pipeline (imputation, 1% winsorization, normalization), to identify variables with strong linear relationships to failure.
- **Random forest**, used after imputation and winsorization, to capture non-linear patterns and rank variable importance based on predictive power.

While these statistical methods guided initial selection, we prioritized interpretability and theoretical relevance from the finance and bankruptcy literature. We favored simple, economically meaningful ratios over complex composite measures, since composite variables can obscure economic meaning ('black-box') and their constituent parts frequently overlap with other, more transparent variables. By emphasizing parsimony and clarity, our final set balances predictive strength and interpretability.

This process yielded a final list of 31 variables. Together, they offer a robust, transparent foundation for model development.

3.5.6 Final Firm-Level Characteristic Variables

Our final model uses 41 predictors: 8 Campbell baseline variables (including one ordinal), 31 numeric variables selected through a rigorous, multi-stage process, the categorical `industry_code` variable, and an ordinal `issues` variable that flags data construction concerns. These variables represent a balanced mix of financial and market indicators grounded in economic theory and corporate finance literature.

The predictors fall into several broad categories, including core distress measures, profitability, value, momentum, trading frictions, investment activity, firm age, and intangibles. For a detailed description and full list of included variables, see Appendix A.5. Figure 3.4 illustrates the variable selection process.

3.5.7 Distribution of Continuous Firm-Level Variables

The continuous firm-level variables differ markedly between failed and non-failed firms in terms of median, spread, and overall distributional shape. Variables such as age, leverage (`debt_me`), capped price (`prc_capped`), and price relative to the 52-week high (`prc_highprc_252d`) display clear shifts in central tendency and dispersion, suggesting their value as predictors of distress. Many features are highly skewed, contain extreme observations, and vary substantially across groups, underscoring the need for robust imputation and scaling during modeling.

These extreme observations are not random but systematically align with failure events. Earnings-to-price (`ni_me`) and price momentum (`ret_12_1`), for example, often show large outliers concentrated in failing firms. Such distributional shifts and recurrent extremes provide strong signals of financial weakness and enhance predictive performance.

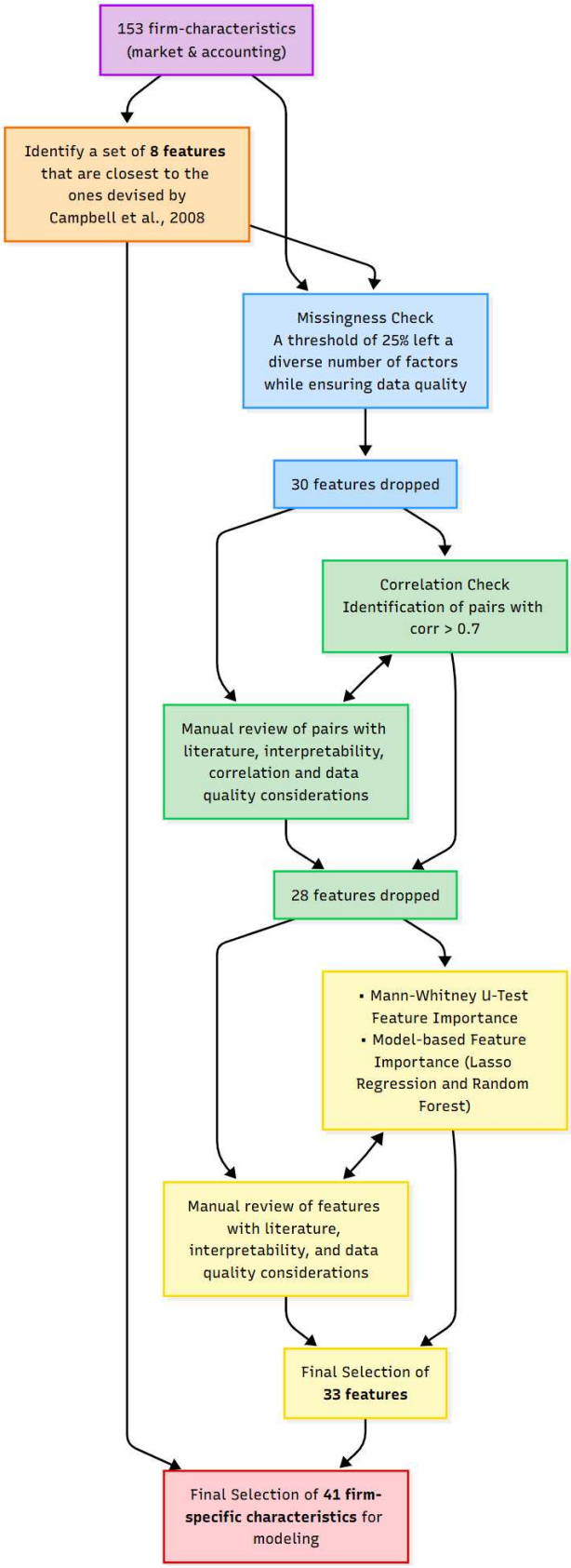
3.6 Selection and Exploration of Macroeconomic Variables

We retained all macroeconomic variables to assess their predictive power in explaining firm-level corporate failures. We created a monthly series of total firm failures from the training dataset as the outcome variable. To capture the impact of economic conditions over time, we engineered features from selected macroeconomic indicators: `gdp_growth`, `inflation`, `lt_interest`, `sp500_ret`, `st_interest`, `st_interest_inflation_adj`, `unemployment`, and `yield_slope` (the yield curve slope).

For each variable, we generated six lagged features (3, 6, 9, 12, 18, and 24 months) and four rolling statistics (mean and standard deviation over 6 and 12 months). We used only past data

Figure 3.4

Firm-Level Feature Selection Pipeline



for rolling statistics to avoid look-ahead bias. Because `gdp_growth` is quarterly, we computed rolling statistics over 4-quarter and 8-quarter windows with a 1-quarter shift.

We tested each generated feature with univariate correlation against monthly failure counts. We kept the top 30 most predictive features (see Appendix A.6) and checked them for multicollinearity. From each correlated cluster, we selected one representative feature, yielding six final macroeconomic predictors:

- `gdp_growth_avg4q`,
- `inflation_vol12m`,
- `sp500_ret_vol12m`,
- `st_interest_inflation_adj_lag24`,
- `unemployment_lag3`,
- `yield_slope_avg6m`

We merged these features with the firm-level panel dataset using the end-of-month (`eom`) timestamp.

A simple multivariate regression with these six predictors achieved an R^2 of 0.14, showing that these economic conditions explain a substantial share of the monthly failure count variance.

We examined the distribution and temporal patterns of these macroeconomic indicators through time-series plots. These visualizations clarified macro-level dynamics and supported using selected variables as exogenous covariates.

Figure 3.5

Yearly Failures with Overlaid Normalized Macroeconomic Indicators

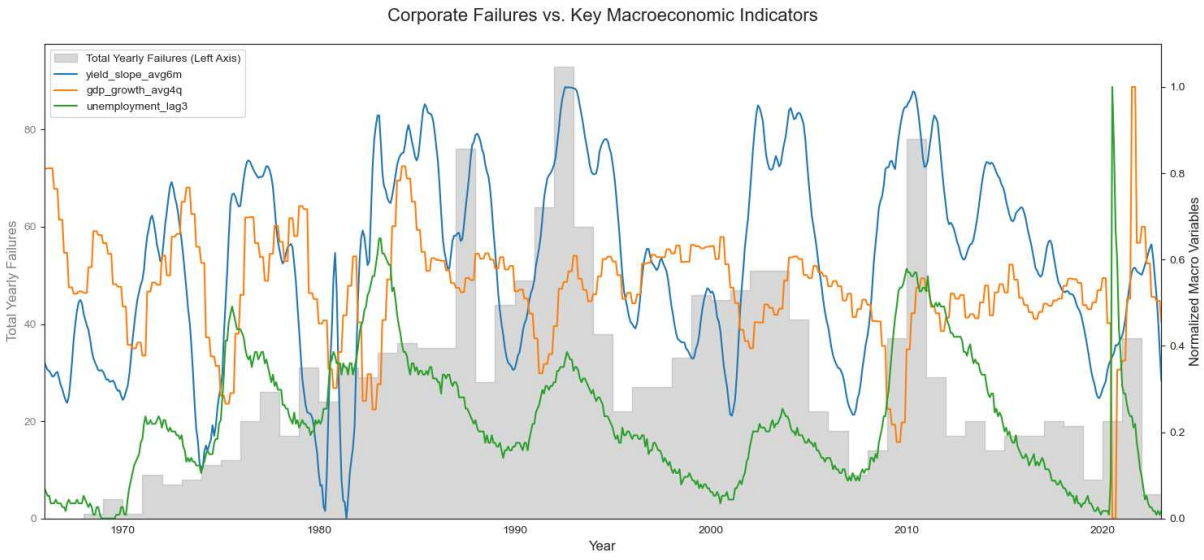


Figure 3.5 shows yearly corporate failure totals (bars) alongside normalized macroeconomic variables (`yield_slope`, `gdp_growth_avg4q` and `unemployment_lag3`) strongly linked to defaults. Macroeconomic changes correspond to firm distress, though not always perfectly synchronized. Unemployment (lagged 3 months) often rises before or during failure peaks, suggesting it may serve as an early signal. GDP growth generally moves inversely to failures, with downturns coinciding with spikes in defaults. The yield curve slope, a standard recession indicator, often flattens or inverts ahead of increased defaults, highlighting its predictive role.

Overall, the figure shows high corporate failure periods align closely with macroeconomic downturns. Some indicators, like the yield curve and unemployment, provide early warnings, while others, including GDP growth and market volatility (see Figure A.1 in the appendix), move alongside observed defaults.

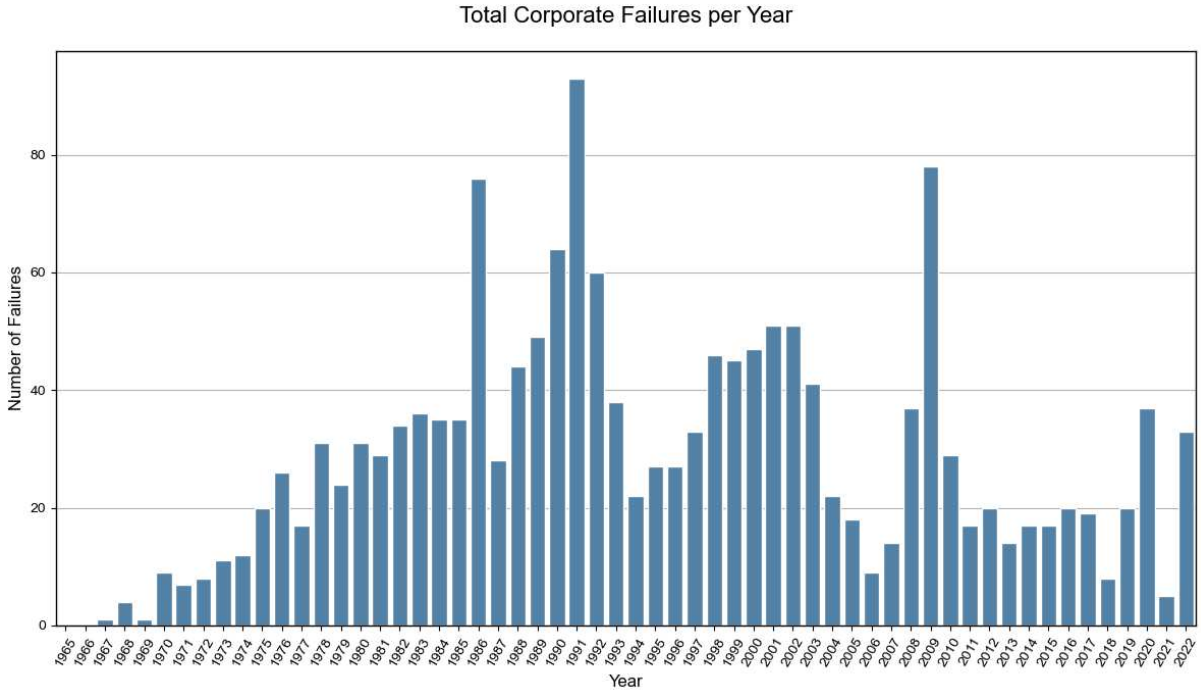
3.7 Exploration of Corporate Failure Indicator

We found a high class imbalance in both bankruptcy and failure datasets: over 99.9% of observations represent non-failure outcomes (failure = 0). We will account for this imbalance during model development.

Figure 3.6 shows the distribution of corporate failure event dates across the sample. Corporate failure counts rise noticeably during economic downturns and financial stress periods, such as the sharp increase during the financial crisis (2008–2009). These peaks demonstrate how adverse macroeconomic conditions raise failure risk, while stable economic times coincide with lower failure rates. The figure highlights the cyclical nature of firm distress linked to broader economic fluctuations.

Figure 3.6

Number of Corporate Failures per Year (1965–2022)



3.8 ID-specific Imputation Strategy

Lastly, a two-step imputation approach was implemented at the firm level (`firm_id`): initially, a forward-fill (`ffill`) method propagated the last available monthly observation, followed by within-firm median imputation to handle any remaining missing entries. This strategy preserved firm-level dynamics while improving data completeness. Overall, 86.5% of missing values were imputed in the Campbell variable set, compared to 71.9% in the full variable set.

3.9 Summary and Final Variables

This study uses three main variable types: (1) firm-level accounting and market characteristics, (2) macroeconomic variables, and (3) a binary outcome indicating corporate failure (bankruptcy or liquidation). The panel covers U.S. publicly traded firms from 1965 to 2022. Data sources include CRSP and Compustat for firm data, and FRED and CRSP for macroeconomic indicators. Each row in the dataset is identified by the columns `firm_id` and `eom` (end-of-month date).

We excluded firm-level characteristics with excessive missing data during preprocessing. Remaining variables underwent exploratory data analysis and correlation filtering to reduce multicollinearity, followed by model-based feature selection. We emphasized interpretability and

theoretical support when choosing variables. For macroeconomic indicators, we created lagged and rolling macroeconomic features and selected the top six predictors.

The final dataset forms the basis for bankruptcy prediction modeling, combining firm fundamentals and macroeconomic context. It includes 41 firm-level variables (38 continuous, 2 ordinal, 1 categorical) and 6 macroeconomic variables. Table A.4 in the appendix lists all final variables.

4 Predictive Modeling

4.1 Overview of Modeling

The modeling procedure predicts corporate failure using two datasets (Campbell and Full variable sets) and three machine learning algorithms: Logistic Regression, Random Forest, and XGBoost. The data splits chronologically into training (1965–1994, approximately 50%), validation (1995–2007, approximately 20%), and test sets (2008–2022, approximately 30%). This split prevents look-ahead bias since model fitting and hyperparameter tuning occur only on the training and validation sets, reserving the test set for final evaluation. The test period includes years after the publication of Campbell et al. (2008), allowing assessment of the original Campbell variables' predictive power over time.

We constructed two datasets. Both contain a firm identifier (`firm_id`), the corresponding month-end date (`eom`), and a binary corporate failure indicator.

- **Campbell dataset:** Based on the eight firm-specific variables originally proposed by Campbell et al. (2008).
- **Full dataset:** Contains 47 variables total: the eight Campbell variables, 33 additional firm-specific characteristics, and 6 macroeconomic variables. The additional features selection process is described in Chapter 3.

The binary failure variable indicated whether a firm experiences corporate failure at time $t + 1$. Modeling initially focused on a one-month prediction horizon, extending later to six-month ($t + 6$) and twelve-month ($t + 12$) horizons without retuning hyperparameters. This approach assessed model performance across prediction horizons while keeping the data chronological and maintaining interpretability. It revealed if predictive power remains stable as the horizon lengthens, providing insights on temporal robustness.

We applied three machine learning models to each dataset:

- **Logistic Regression**,

- **Random Forest**, and
- **XGBoost**,

yielding six model variations in total.

We evaluated model performance primarily using the Area Under the Receiver Operating Characteristic Curve (AUC) (Barboza et al., 2017). Additionally, we monitored Type I errors (false positives) to assess the model's capability of correctly identifying failures. Feature importance analysis identifies key predictors for each model.

4.2 Preprocessing Pipeline

We implemented a unified preprocessing pipeline to treat variables consistently, especially for logistic regression. The pipeline applies different procedures based on variable type:

- **Numeric variables:** We imputed missing values using the median, applied winsorization at the 1% tails to limit extreme outliers, and then standardized the variables.
- **Ordinal variables:** These variables had no missing values and required no preprocessing.
- **Categorical variables:** The only categorical feature was *industry_code*, which ranges from 1 to 10 or can be missing. We treated missing values as a separate category ('11') rather than imputing, to preserve the original industry distribution. We then applied one-hot encoding.
- **Macroeconomic variables:** These contained no missing values and were standardized directly.

We performed imputation using `SimpleImputer` with the `strategy='median'` to replace missing values. Standardization was applied using `StandardScaler`, which centers each variable to a mean of zero and scales it to unit variance. This process ensures that features contribute proportionally to the model and improves numerical stability during optimization.

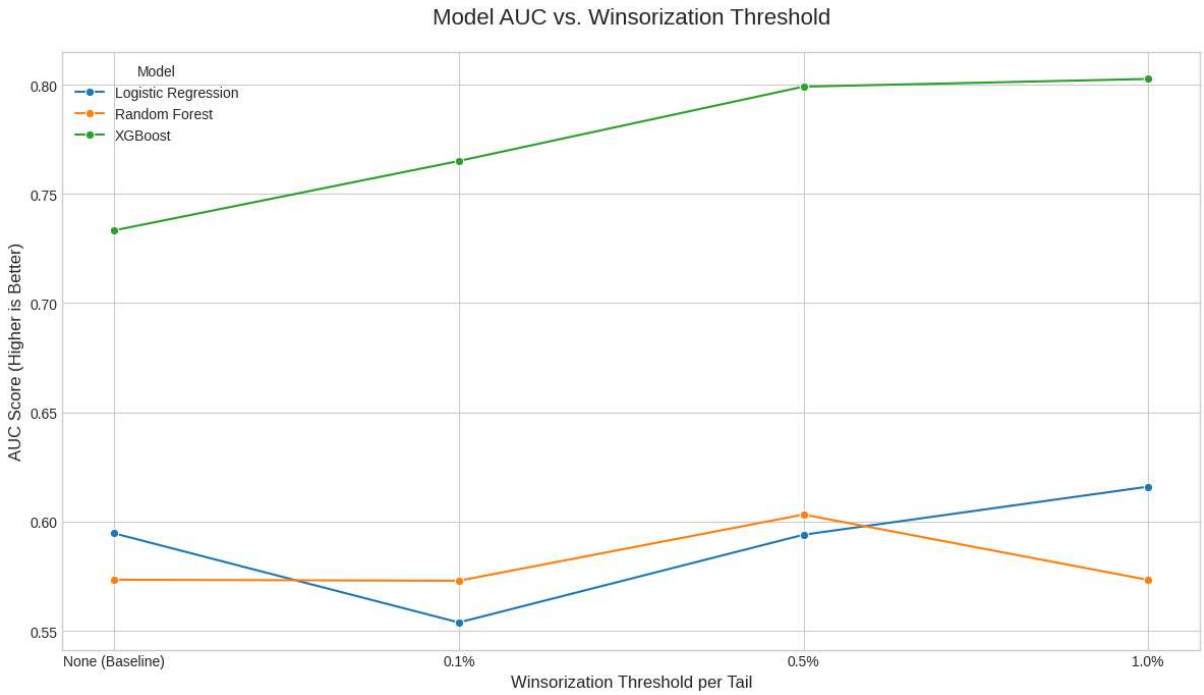
We tested how winsorization thresholds affect model performance using a sensitivity analysis on a training sample for all numeric variables. We evaluated four scenarios:

- No winsorization,
- Winsorization at the 0.1% tails,
- Winsorization at the 0.5% tails,
- Winsorization at the 1% tails.

For each scenario, we trained models and measured predictive performance via AUC on a random training sample. We identified the 1% tail winsorization as the best balance, providing stable and generally higher predictive performance. Figure 4.1 illustrates these results, so we applied this threshold in the final numeric preprocessing pipeline.

Figure 4.1

Model Performance across Winsorization Thresholds



We applied this preprocessing consistently for all logistic regression models to ensure comparability and stability. For ensemble models (Random Forest and XGBoost), we tested two approaches: applying this preprocessing and using raw data without preprocessing. This comparison verified whether preprocessing influenced performance and ensured consistent evaluation across modeling approaches.

4.3 Naive Baseline Model

We constructed a firm-level terminal-event naive baseline to benchmark model performance. This baseline respects the data structure, where failures occur only in a firm's final observed month. During validation, we applied the overall failure rate from the training set by randomly assigning the same proportion of firms as failed in their last observed month. For the test stage, we recalculated the failure rate based on the combined training and validation data and applied it

similarly. This approach prevented data leakage, mirrors the actual failure process, and provides a valid lower bound on performance.

As expected, the naive baseline yielded AUCs of 0.5046 on the validation set and 0.5099 on the test set, which is near random chance and showed no discriminative ability. Including this naive benchmark ensures that improvements from more complex models are meaningful.

4.4 Hyperparameter Tuning

Hyperparameters were tuned using the one-month prediction horizon ($t + 1$). Due to the imbalance in the binary outcome variable, the class_weight parameter (or the equivalent scale_pos_weight parameter for XGBoost) was set to 'balanced'. This ensures that misclassifications of the minority (failure) class are penalized more heavily, addressing the skewed class distribution.

4.4.1 Logistic Regression

Model performance was strong without hyperparameter tuning, as Campbell originally designed and optimized his framework using standard logistic regression. While we conducted hyperparameter optimization using RandomizedSearchCV with three-fold cross-validation on a random training subsample of 50,000 observations, this tuning yielded only minor improvements.

The optimal hyperparameters differed between datasets: the Campbell dataset favored the `liblinear` solver with L1 regularization (lasso) and a strong regularization parameter, while the larger full dataset preferred the `saga` solver with L2 regularization (ridge) and a weaker penalty. These choices reflect differences in dataset size and complexity: L1 regularization shrinks some coefficients to zero, helping variable selection in smaller feature spaces, while L2 regularization preserves more variables in larger, high-dimensional datasets.

This tuning preserved interpretability and mitigated overfitting while adapting to the evolving complexity of the extended dataset. See Appendix A.9.1 for detailed hyperparameter search spaces and results.

4.4.2 Random Forest

We tuned Random Forest models separately for each dataset (Campbell and extended) and for both preprocessing strategies (with and without preprocessing), resulting in four sets of optimal hyperparameters. Initial models showed clear overfitting signs, with training AUCs near 1.0

but substantially lower validation performance, highlighting the need for careful tuning and generalization control.

Using a representative subsample of 100,000 firm-month observations grouped by firm, we performed randomized hyperparameter searches with GroupKFold cross-validation (3 folds), respecting the panel data structure.

Key hyperparameters included the number of trees, maximum tree depth, minimum samples required to split nodes and form leaves, and feature subsampling methods. Preprocessing consistently allowed more complex tree structures (deeper trees and larger ensembles) without overfitting in the smaller Campbell dataset. Without preprocessing, models relied on shallower trees and more constrained leaf sizes to control variance. By contrast, the extended dataset favored relatively shallow trees regardless of preprocessing, reflecting the higher risk of overfitting in a high-dimensional setting.

Preprocessing influenced splitting rules differently across datasets: it enabled larger leaf sizes and feature subsampling by a fixed proportion in the extended case, promoting model conservatism, while without preprocessing, it favored smaller leaves and a logarithmic feature sampling approach.

Overall, Random Forest models trained on the extended dataset consistently outperformed Campbell models, with cross-validation AUCs ranging from 0.9125 to 0.9142 versus 0.8756 to 0.8798, demonstrating the substantial predictive benefit from a richer feature set. These results support the benefits of thorough hyperparameter tuning tailored to dataset size and preprocessing strategy.

Refer to Appendix A.9.2 for detailed search spaces and optimal hyperparameter values.

4.4.3 XGBoost

We performed hyperparameter tuning for XGBoost similar to Random Forest, utilizing a representative subsample of 100,000 firm-month observations grouped by firm and three-fold GroupKFold cross-validation. We explored key hyperparameters including boosting rounds, tree depth, learning rate, subsampling ratios, feature sampling, and regularization penalties, testing models both with and without preprocessing.

Before tuning, XGBoost exhibited signs of overfitting, though less severe than Random Forest, with training AUCs close to 1.0 but lower validation performance. The Campbell dataset favored very deep trees with moderate learning rates and minimal regularization, consistent across preprocessing configurations. In contrast, the extended dataset required more substan-

tial regularization, shallower tree structures, and conservative boosting parameters to control overfitting.

Preprocessing only slightly improved performance in the Campbell specification, while substantially enhancing the extended dataset results. The extended preprocessed model achieved the highest cross-validation AUC (0.9257), combining deeper trees and faster learning with stronger regularization and more selective feature sampling.

These results reflect the importance of balancing model complexity with generalization, tailoring hyperparameter tuning to dataset characteristics, and leveraging preprocessing to enable more effective learning dynamics. This combination produced the strongest predictive performance among all modeling approaches in our study.

Refer to Appendix A.9.3 for detailed hyperparameter tuning specifications and results.

4.5 Modeling Overview

In the final modeling stage, we estimated a total of 30 models to evaluate predictive performance across multiple prediction horizons, datasets, and modeling approaches. Specifically, for each horizon, $t + 1$, $t + 6$, and $t + 12$, we ran the following models: Logistic Regression, Random Forest, and XGBoost.

We considered two variable specifications: the original Campbell variables and an extended variable set. Additionally, for ensemble models (Random Forest and XGBoost), we trained versions both with and without preprocessing for each variable specification. This approach resulted in a comprehensive set of models capturing a wide spectrum of methodological choices.

We conducted hyperparameter optimization solely at the shortest horizon $t + 1$ for each model. For the longer horizons $t+6$ and $t+12$, we reused the optimized hyperparameter values obtained at $t + 1$ to maintain consistency across horizons and avoid overfitting to extended prediction outcomes.

This modeling framework enabled systematic comparison of predictive performance across model types, datasets, preprocessing strategies, and prediction horizons, while holding hyperparameter settings and evaluation procedures consistent for comparable assessment.

4.6 Evaluation Metrics

We computed standard evaluation metrics for each model run, including AUC, accuracy, precision, recall, F1 score, Type I error, Type II error, and the confusion matrix. We identified the most important variables using feature importance measures. For Logistic Regression, importance derives from the absolute magnitude of coefficients after preprocessing. Random Forest and XGBoost rely on their built-in feature importance scores. We assess top features from each model to compare predictive contributions across algorithms. For each model, we computed metrics and feature importance at two evaluation stages: first, training set (1965–1994) predictions on the validation set (1995–2007); then combined training and validation set (1965–2007) predictions on the test set (2008–2022).

4.6.1 Area Under the Curve (AUC)

The **AUC** measures a model's ability to distinguish positive from negative classes across all classification thresholds. It is the area under the Receiver Operating Characteristic (ROC) curve, which plots True Positive Rate against False Positive Rate. A higher AUC means the model ranks failing firms higher than non-failing firms, independent of any specific threshold (Barboza et al., 2017).

AUC suits bankruptcy prediction because it evaluates the overall ranking performance rather than accuracy at one threshold, making it robust for rare-event problems. It is also relatively insensitive to class imbalance. Values range from 0.5 (random guessing) to 1.0 (perfect discrimination). In bankruptcy literature, an AUC above 0.8 generally indicates excellent performance (Barboza et al., 2017).

4.6.2 Accuracy, Recall, Precision, and F1 score

Accuracy measures the fraction of correct predictions among all cases. Though intuitive, accuracy may mislead in imbalanced datasets like corporate failure, where predicting all firms as non-failing yields high accuracy (> 99%) but no practical value (Shumway, 2001).

Recall (also called sensitivity or True Positive Rate) measures the fraction of actual failures correctly identified, while **Precision** measures the proportion of predicted failures that are true failures. The **F1 score** combines precision and recall into a single metric, Together, they show how well the model detects failures without excessive false alarms (Bauer & Agarwal, 2014).

4.6.3 Type I and Type II Errors

Type I error (False Negative Rate) happens when a failing firm is wrongly classified as non-failing. This error is costly because it risks continued investment in troubled firms and delays interventions (Agarwal & Taffler, 2008). **Type II error** (False Positive Rate) occurs when a non-failing firm is wrongly classified as failing. Though inconvenient, these errors tend to be less costly since they prompt cautionary decisions reducing potential losses. Minimizing Type I error remains a priority to effectively identify at-risk firms (Agarwal & Taffler, 2008; Altman, 1968).

4.6.4 Confusion Matrix

The **confusion matrix** summarizes prediction outcomes comparing predicted and actual classes: true positives (correctly identified failures), false positives (non-failures labeled as failures), false negatives (missed failures), and true negatives (correct non-failures). It offers a straightforward summary of classification performance (Ohlson, 1980).

4.6.5 Evaluation Focus

This thesis emphasized AUC as the main metric for overall discriminative power and closely monitored Type I error to assess practical effectiveness in identifying failing firms, while also keeping a close watch on the confusion matrix, using the default classification threshold of 0.5 without any additional tuning.

We evaluate model types (Logistic Regression, Random Forest, XGBoost), variable sets (Campbell vs. Extended), and preprocessing strategies (with vs. without pipeline). We assess predictive generalization over time using a training period from 1965 to 1994, validation from 1995 to 2007, and test from 2008 to 2022, across prediction horizons $t+1$, $t+6$, $t+12$.

5 Results

5.1 Model Performance

Table 5.1

Model performance at $t+1$, $t+6$, $t+12$ horizons by model and variable set, with and without preprocessing. ***Bold*** *indicates best AUC and lowest Type I error per horizon.*

		$t+1$		$t+6$		$t+12$	
Model	Variables	AUC	Type I	AUC	Type I	AUC	Type I
With Preprocessing							
Logistic Regression	Campbell	0.9354	0.14	0.8953	0.16	0.8493	0.22
Logistic Regression	Full	0.8990	**0.12**	0.8831	**0.16**	0.8639	0.22
Random Forest	Campbell	0.9516	0.95	0.9149	1.00	0.8550	1.00
Random Forest	Full	**0.9560**	0.16	**0.9239**	0.18	**0.8899**	**0.21**
XGBoost	Campbell	0.9552	0.98	0.8988	1.00	0.8500	1.00
XGBoost	Full	0.8900	0.72	0.9029	0.64	0.8718	0.70
Without Preprocessing							
Random Forest	Campbell	0.9492	0.13	0.9110	**0.14**	0.8715	0.18
Random Forest	Full	**0.9616**	0.16	**0.9303**	0.17	**0.8997**	0.21
XGBoost	Campbell	0.9549	0.98	0.9035	1.00	0.8331	1.00
XGBoost	Full	0.9607	0.14	0.9212	0.15	0.8942	**0.18**

Note*: Top panel applies preprocessing; bottom panel omits it. Bold values mark highest AUC and lowest Type I error per horizon. Results reflect models trained on the combined training and validation period (1965–2007) and tested on the out-of-sample period (2008–2022).*

Table 5.1 highlights the comparative predictive performance of the models across multiple horizons. Logistic Regression using the Campbell variable set delivers robust performance at the one-month horizon ($t+1$) with a test AUC of 0.9354 and low Type I error, establishing a practical and interpretable baseline. The persistent strong test AUC confirms the enduring validity of the Campbell variable set, originally proposed by Campbell et al. (2008), even when applied

to out-of-sample data covering the post-publication period (2008–2022). This highlights that the classic financial distress predictors retain their power across economic cycles and evolving market conditions, supporting their continued use in early warning systems.

While ensemble models such as Random Forest and XGBoost generally achieve higher AUC values, signaling improved discrimination, these benefits must be balanced with greater computational complexity and extended data requirements stemming from the full variable set. Notably, ensemble methods exhibit a slower decline in performance over medium- to long-term horizons ($t + 6$, $t + 12$) compared to Logistic Regression, suggesting their superior capacity to model complex temporal dynamics and nonlinear interactions.

Preprocessing influences models unevenly: it remains essential for Logistic Regression, which depends on imputations and feature scaling, but it may sometimes reduce the effectiveness of ensemble models by potentially removing nonlinear signals. Patterns of overfitting, such as perfect scores in training accompanied by a decline in validation or testing AUC, appear more pronounced in the models where preprocessing enabled deeper tree structures during hyperparameter tuning. Although differences in data treatment imply that results with and without preprocessing are not fully comparable, this variation underscores important practical considerations when deploying these models.

Finally, the strong AUCs across all models indicate identifiable failure signals, yet relatively higher Type I errors, especially in ensemble methods, suggest that further optimization via threshold tuning could improve practical failure identification without compromising overall model discrimination. Such calibration could enhance operational utility in credit risk and bankruptcy early warning systems.

Table 5.2 reports the confusion matrices for Random Forest with preprocessing at the $t + 1$ horizon for both Campbell and Full variable sets. The Full variable set detects substantially more failing firms correctly, with 311 true positives compared to 20 for the Campbell set, at the expense of a larger false positive count. Conversely, the Campbell set yields fewer false positives but substantially underdetects failures, highlighting the trade-off between detection sensitivity and specificity. There needs to be a balance between false positives and false negatives, but considering false negatives come at a higher cost in risk management contexts (Agarwal & Taffler, 2008), the Full variable set outperforms the Campbell set in practical terms.

5.2 Variable Importance

The Table 5.3 presents the top variables selected by each model across different variable sets, horizons, and training periods, providing insight into the evolution of predictor importance in

Table 5.2

Confusion matrices for Random Forest with preprocessing at the $t + 1$ horizon using Campbell and Full variable sets.

Model	Variable Set	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
Random Forest	Campbell	650,675	176	350	20
Random Forest	Full	599,682	51,169	59	311

***Note**: Additional confusion matrices for other models and horizons can be found in the appendix. This thesis follows prior literature (Agarwal & Taffler, 2008) in prioritizing Type I error reduction due to the greater costs associated with false negatives in financial distress prediction.*

corporate failures forecasting.

Across all models and prediction horizons, a set of core variables consistently emerges as critical predictors of financial distress. Notably, the Campbell variables `prc_capped` (capped price), `rvol_21d_3m` (volatility), `niq_at_avg` (profitability, geometric average return on assets), `debt_me` (leverage, debt-to-market equity ratio), and `me_binned` (size, 1–5 market equity bins) maintain strong importance even when evaluated within the larger full variable set. This persistence underscores the enduring relevance of these classical financial measures as foundational predictors across differing model architectures, whether linear or nonlinear. Their robust presence validates their continued inclusion alongside newer variables from enriched datasets, suggesting that these Campbell variables capture core structural aspects of distress risk that remain predictive across time and modeling frameworks.

Additionally, the variable `issues`, a data creation feature capturing parallel issues for one firm, frequently appears in the linear models, but its role is linked to data construction rather than an economic driver of distress. Alongside it, the significance of `industry_code` as a predictor, consistent with findings by Chava and Jarrow (2004), manifests mainly in logistic regression, suggesting sectoral risk factors play a stronger role within linear frameworks. Other mixed variables such as `ni_me` (earnings-to-price ratio), accounting variables like `ni_be` (return on equity), and market variables like `prc_highprc_252d` (price relative to the highest price over the past 252 trading days) further contribute important predictive information,

Temporal shifts between the Train (1965–1994) and Train + Val (1965–2007) periods indicate evolving variable importance. Some variables, marked with asterisks, emerge or shift in ranking, particularly among market-related measures such as `me_binned` (market capitalization bins) and volatility indicators. These changes likely reflect structural shifts in market behavior and financial reporting over time, emphasizing the necessity of model retraining with updated data

Table 5.3

Top selected variables for each model, variable set, and horizon with & without preprocessing.

Model	Variables	$t+1$	$t+6$	$t+12$
With Preprocessing				
LR	Campbell	me_binned, debt_me, prc_capped	prc_capped, debt_me, me_binned*, rvol_21d_3m**	prc_capped, niq_at_avg, debt_me
LR	Full	industry_code, issues, ebit_bev, ni_be*, age*, aliq_at**, cop_atl1**	industry_code, issues, cop_atl1*, ebit_bev*, age*, prc_capped**, aliq_at**, ni_be**	industry_code, issues, ni_be*, div12m_me*, niq_at_avg*, prc_capped**, cop_atl1**, prc_highprc_252d**
RF	Campbell	niq_at_avg, prc_capped, me_binned*, debt_me**	niq_at_avg, prc_capped, me_binned	niq_at_avg, prc_capped, me_binned
RF	Full	ni_me, prc_capped, ni_be, me_binned, rvol_21d_3m	ni_me, prc_capped, ni_be, prc_highprc_252d, ebit_sale*, rvol_21d_3m**	ni_me, prc_capped, ni_be, at_be*, ebit_sale*, rvol_21d_3m**, prc_highprc_252d**
XGB	Campbell	rvol_21d_3m, niq_at_avg, prc_capped	rvol_21d_3m, niq_at_avg, me_binned	niq_at_avg, me_binned, prc_capped
XGB	Full	ni_me, prc_capped, at_be*, bidaskhl_21d*, at_me*, me_binned**, div12m_me**, dolvol_var_12d**	me_binned, prc_capped, ni_me, prc_highprc_252d, ni_be	prc_capped, ni_me, ni_be, me_binned*, age*, prc_highprc_252d**, at_me**
Without Preprocessing				
RF	Campbell	niq_at_avg, prc_capped, rvol_21d_3m	niq_at_avg, prc_capped, rvol_21d_3m	niq_at_avg, prc_capped, rvol_21d_3m
RF	Full	ni_me, niq_at_avg, prc_capped, rvol_21d_3m, bidaskhl_21d*, prc_highprc_252d**	ni_me, niq_at_avg, prc_capped, prc_highprc_252d, bidaskhl_21d*, rvol_21d_3m**	ni_me, niq_at_avg, prc_capped, prc_highprc_252d, bidaskhl_21d*, rvol_21d_3m**
XGB	Campbell	prc_capped, me_binned, niq_at_avg*, rvol_21d_3m**	niq_at_avg, prc_capped, rvol_21d_3m*, debt_me**	niq_at_avg, prc_capped, rvol_21d_3m*, debt_me**
XGB	Full	ni_me, prc_capped, niq_at_avg, prc_highprc_252d, rvol_21d_3m	ni_me, prc_capped, niq_at_avg, prc_highprc_252d, ni_be	prc_capped, niq_at_avg, ni_me, prc_highprc_252d, rvol_21d_3m*, ni_be**

* indicates the variable appears only in the Train model.

** indicates the variable appears only in the Train+Val model.

This does not imply unimportance; it simply did not rank in top 3 (Campbell) or top 5 (Full) for that model and dataset.

to capture new distress drivers. Similarly, variables exhibit modest variation across horizons, with shorter-term horizons favoring immediate market signals, while longer horizons show enhanced importance for fundamental accounting fundamentals such as leverage (`debt_me`) or profitability (`niq_at_avg`) measures.

Preprocessing exerts a subtle yet meaningful influence on variable importance rankings. For instance, models employing preprocessing often elevate accounting fundamentals such as profitability (`niq_at_avg`) and leverage (`debt_me`), likely because normalization and imputation reduce scale differences and missing data biases, thereby enhancing the stability and predictive contribution of these financial fundamentals. Conversely, market-based variables like price (`prc_capped`) and volatility (`rvol_21d_3m`), which capture raw market dynamics and complex nonlinearities, may be somewhat diminished in importance following preprocessing steps that smooth or standardize data distributions. For example, in ensemble models such as Random Forest and XGBoost, the importance of volatility measures sometimes declines after scaling. This dual effect highlights the need to carefully tailor preprocessing strategies to the modeling approach, balancing improved variable interpretability against potential loss of informative nonlinear patterns.

Categorizing variables into market, accounting, hybrid market-accounting, and macro groups reveals that market and accounting predictors dominate the top rankings, with macroeconomic variables less frequently selected. This pattern aligns with prior literature emphasizing firm-specific signals over broad economic indicators for distress prediction. The interplay between these categories, particularly where variables blend market and accounting information, reflects the multifaceted nature of default risk and the value of integrating heterogeneous data sources for robust early warning systems.

6 Discussion

6.1 Modeling Approaches and Predictive Accuracy

Ensemble methods, especially XGBoost and Random Forest, demonstrated superior predictive accuracy for longer-term predictive horizons on the given dataset. Logistic regression, paired with Campbell's variable set (Campbell et al., 2008), remained competitive at short-term forecasts. These findings align with Barboza et al. (2017) and with Alonso and Carbó (2021), who found that machine learning methods can deliver higher discriminatory power than traditional statistical models. However, such models may require richer variable sets and careful consideration of error trade-offs. We find that while ensemble models offer improved overall AUC, they also introduce more data complexity and higher computational needs.

6.2 Economic Interpretation of Key Predictors

Market-based variables, such as price and volatility measures, consistently ranked among the most important predictors in all models. These signals reflect current investor sentiment and liquidity risks, supporting the argument by Agarwal and Taffler (2008), Shumway (2001) and Campbell et al. (2008) that those market measures contains forward-looking distress information and that they work best in combination (or hybrid). Accounting variables, especially leverage and profitability measures, are prominent across all models, though leverage shows a slight decline in importance in tree-based approaches. This suggests that non-linear methods may capture distress signals from market interactions, slightly reducing the relative weight of traditional ratios (Altman, 1968; Beaver, 1966). Such a shift could be due to interactions and thresholds absorbing part of the risk information otherwise represented by linear terms.

Macroeconomic variables exhibited weak and inconsistent contributions. This is consistent with Tinoco and Wilson (2013) and Bonfim (2009), who found that firm-level indicators generally dominate such measures in bankruptcy prediction. One explanation is that macroeconomic effects are partly captured by market-based predictors or absorbed into baseline model components, such as the intercept in logistic regression. In addition, the hazard model specification

already incorporates temporal dynamics through time-varying baseline hazards and covariates, as shown by Shumway (2001) and Bauer and Agarwal (2014), which may diminish the incremental value of explicit macroeconomic variables. While these factors help explain why such indicators rarely emerged among the top-ranked features, ensemble models that included them consistently outperformed the Campbell variable specification, suggesting they might still provide complementary, though indirect, information.

6.3 Generalization and Stability

Overall, model performance is quite good and stable, with consistent results from training to testing datasets covering the period from 1965 to 2022, and most models exceeding the benchmark AUC of 0.8 reported by Barboza et al. (2017). Performance declined as the horizon extended, particularly for logistic regression. Ensemble models displayed greater stability, retaining discriminatory power in medium- and long-term forecasts. These results contrast with some earlier hazard-model findings (Bauer & Agarwal, 2014; Shumway, 2001) and highlight the ability of modern machine learning models to learn persistent predictive patterns, though at the cost of interpretability and heightened computational needs.

Importantly, even modest gains in predictive accuracy can translate into substantial financial savings for lenders and investors when these models are deployed at scale. As Agarwal and Taffler (2008) demonstrate, improved default prediction not only helps mitigate losses, but also enhances capital allocation and delivers significant strategic advantages by enabling more reliable long-term forecasting.

6.4 Practical Implications and Future Directions

The practical choice of modeling technique depends on the relative importance of predictive accuracy, error rates, and model transparency. For real-world applications, low Type I error remains essential for effective early warning and risk management (Agarwal & Taffler, 2008). Models combining extended feature sets and ensemble methods excel as forecasting tools but require careful validation to avoid excessive complexity.

Future research could focus on refining variable selection within the extended 47-variable dataset to balance parsimony and predictive performance, as reducing overly large datasets may enhance model generalizability and interpretability. Exploring alternative regularization techniques, threshold selection methods, or additional macroeconomic data sources may further improve both economic insight and robustness. Additionally, integrating machine

learning findings with established financial theory, as suggested by Ahmed et al. (2022), offers promising avenues to deepen understanding of firm distress and credit risk assessment. The role of preprocessing deserves further investigation since ensemble models showed a slight performance improvement without preprocessing, which, although not perfectly comparable in this case, might have practical implications for model deployment.

7 Conclusion

This thesis investigated the predictive modeling of corporate failure using machine learning techniques. The goal was to evaluate the performance of logistic regression, random forest, and XGBoost across different variable sets, preprocessing strategies, and prediction horizons, and to identify the key drivers of corporate failure.

A panel dataset of listed US firms was constructed at the firm–month level (identified by `firm_id` and `eom`) covering the period from 1965 to 2022, containing 41 firm-specific characteristics, including market, accounting, and mixed measures, along with 6 macroeconomic variables and a binary failure indicator. The dataset was carefully compiled through rigorous data gathering, cleaning, and variable selection procedures. From this full dataset, a smaller subset of eight Campbell-like variables was extracted to serve as a baseline specification for comparative modeling.

A total of 30 models were estimated, varying across modeling approaches (Logistic Regression, Random Forest, XGBoost), variable sets (Campbell vs. Full), preprocessing strategies (with vs. without pipeline), and prediction horizons ($t+1$, $t+6$, $t+12$). Hyperparameters were tuned separately for each combination of model, variable set, and preprocessing approach. The primary evaluation metrics were AUC, to assess overall discriminatory power, Type I error, to measure the models' ability to correctly identify failing firms, and the Confusion Matrix, which showed a detailed breakdown of classification outcomes. In addition, feature importance was recorded for every model run to identify the key predictors driving performance.

The baseline, Logistic Regression with Campbell variables, is strong. It provides robust predictive power across time. Nevertheless, ensemble models with an extended variable set deliver clear improvements. They achieve higher predictive accuracy and greater stability, particularly in medium- and long-term forecasts. While the Campbell variables remain predictive, additional features such as earnings-to-price (ni_me) and return-on-equity (ni_be) ratios, industry affiliation (industry_code), and momentum measures (prc_highprc_252d) emerge as key drivers of corporate failure, highlighting the value of a richer feature set. Even modest gains in predictive accuracy, particularly in predictions over longer term horizons, can translate into substantial financial savings and even broader benefits when these models are applied at scale.

Several avenues for further research emerge from this study. First, the impact of preprocessing on ensemble model performance warrants deeper investigation, as it may affect model performance, particularly the ability to identify corporate failures. Second, identifying a smaller, high-impact subset of features from the extended variable set could reduce computational complexity while maintaining predictive power. Finally, the role of macroeconomic variables should be explored more closely, as their contribution to model performance and failure prediction remains unclear.

Overall, these results confirm much of the established literature on the value of financial ratios, market signals, and careful model design (Bauer & Agarwal, 2014; Campbell et al., 2008; Shumway, 2001). They also underscore the practical potential of recent advances in machine learning, particularly ensemble methods, to enhance bankruptcy prediction (Alonso & Carbó, 2021; Barboza et al., 2017; Jones et al., 2017). This thesis emphasizes replicability, as advocated by Jensen et al. (2023), and provides a foundation for reliable model building and ongoing method development. This strong foundation can be readily replicated, allowing future research to build upon and optimize these models for improved predictive accuracy and practical relevance.

Bibliography

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of banking & finance*, *32*(8), 1541–1551.

Ahmed, S., Alshater, M. M., El Ammari, A., & Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, *61*, 101646.

Alonso, A., & Carbó, J. M. (2021). Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The journal of finance*, *61*(1), 259–299.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417.

Bauer, J., & Agarwal, V. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? a comprehensive test. *Journal of Banking & Finance*, *40*, 432–442.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71–111.

Black, F., & Scholes, M. S. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, *81*(3), 637–654. https://doi.org/10.1086/260062

Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, *145*, 103773.

Bonfim, D. (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of banking & finance*, *33*(2), 281–299.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of finance*, *63*(6), 2899–2939.

Center for Research in Security Prices (CRSP), A. R. C. a. C. B. (n.d.). Data description guide | crsp us stock and us indices databases [Accessed: 2025-07-08]. https://www.crsp.org/wp-content/uploads/guides/CRSP_US_Stock_%26_Indexes_Database_Data_Descriptions_Guide.pdf

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of finance*, *8*(4), 537–569.

Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). Catboost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*, 120658.

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, *78*(5), 2465–2518.

Jensen, T. I., Kelly, B. T., & Pedersen, L. H. (2022). Is there a replication crisis in finance? *Journal of Finance, Forthcoming*.

Jones, S., Johnstone, D., & Wilson, R. (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting*, *44*(1-2), 3–34.

Mare, D. S. (2015). Contribution of macroeconomic factors to the prediction of small bank failures. *Journal of International Financial Markets, Institutions and Money*, *39*, 25–39.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, *29*(2), 449–470.

Nam, C. W., Kim, T. S., Park, N. J., & Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting*, *27*(6), 493–506.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*.

Pesaran, M. H., Schuermann, T., Treutler, B.-J., & Weiner, S. M. (2006). Macroeconomic dynamics and credit risk: A global perspective. *Journal of Money, Credit and Banking*, 1211–1261.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, *74*(1), 101–124.

Taffler, R. J. (1984). Empirical models for the monitoring of uk corporations. *Journal of banking & finance*, *8*(2), 199–227.

Tinoco, M. H., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International review of financial analysis*, *30*, 394–419.

A Appendix

A.1 Industry Mapping

We mapped firm SIC codes into industry categories following Chava and Jarrow (2004). Two schemes were applied: a coarse classification with 4 broad categories and a detailed classification with 10 industry codes. This mapping enables analysis of corporate failure patterns across industries and comparison with prior research, and has been shown to predict default risk (Chava & Jarrow, 2004).

Table A.1
Industry Mapping

SIC Code Range	Industry Name	Code	Category
< 1000	Agriculture, Forestry & Fisheries	1	4
1000 – 1499	Mineral Industries	2	3
1500 – 1799	Construction Industries	3	4
2000 – 3999	Manufacturing	4	3
4000 – 4999	Transportation, Communications & Utilities	5	2
5000 – 5199	Wholesale Trade	6	4
5200 – 5999	Retail Trade	7	4
6000 – 6799	Finance, Insurance & Real Estate	8	1
7000 – 8899	Service Industries	9	4
9100 – 9999	Public Administration	10	4

Notes: Both the 1–10 industry codes and the broader 1–4 categories are adapted from Chava and Jarrow (2004). The broader categories are defined as follows: **1** = Finance, Insurance & Real Estate; **2** = Transportation, Communications & Utilities; **3** = Manufacturing & Mineral; **4** = Miscellaneous (all remaining industries).

A.2 Macroeconomic Variables

Following prior research on the relationship between macroeconomic conditions and corporate failure (Bonfim, 2009; Mare, 2015; Pesaran et al., 2006), we retrieved key variables from

FRED and CRSP. Other indicators, including coincident indices, loan growth (Bonfim, 2009), oil prices (Pesaran et al., 2006), and VIX (Ang et al., 2006), were considered but lacked consistent coverage for the full sample period. Table A.2 summarizes the variables, source codes, descriptions, reporting frequencies, and available start dates.

Table A.2

Macroeconomic Variables Used in the Analysis

Variable	Code	Description	Frequency	Start
Consumer Price Index	CPIAUCSL	Consumer Price Index for All Urban Consumers	Monthly	1947
GDP	GDPC1	Real Gross Domestic Product	Quarterly	1947
Long-Term Interest Rate	IRLTLT01USM156N	Long-Term Government Bond Yields: 10-Year	Monthly	1953
S&P 500 Return	vwretd	CRSP Value-Weighted Return (incl. dividends)	Daily	1926
Short-Term Interest Rate	TB3MS	3-Month Treasury Bill Rate	Monthly	1934
Unemployment	UNRATE	Unemployment Rate	Monthly	1948

Note: All variables except `vwretd`, which was retrieved from CRSP, were retrieved using the FRED API.

A.3 Chi-squared Test Results for Industry Categorizations

We conducted a chi-squared test to compare the predictive power of the two industry classifications (Chava & Jarrow, 2004) (4-category vs. 10-code). The 1–10 code produced a higher chi-squared statistic and a lower p-value, indicating stronger association with corporate failure and was therefore retained.

Table A.3

Chi-squared Test Results for Industry Classification Variables

Industry Variable	Chi-squared Statistic	Degrees of Freedom	p-value
Category (1–4)	47.02	3	3.44×10^{-10}
Code (1–10)	267.27	9	2.23×10^{-52}

A.4 Issues Flag

During data cleaning, we created an ordinal `issues` flag to capture the number of concurrent firm-level stock issues. This flag is a data processing indicator used to consolidate multiple issues per `firm_id` into a single measure. It is not a characteristic or inherent attribute of the firms. The `issues` flag shows weak correlations with other numeric predictors, mainly those related to size. Statistical tests confirm it differs across failure categories, providing unique and useful information. We kept this flag for later modeling.

A.5 Final Selection of Firm-Specific Characteristics

- **Campbell Core Distress Variables (8):** The foundational predictors based on the seminal work of Campbell et al. (2008), which form the baseline of the model.
  - `be_me`: Book-to-Market Equity (value)
  - `cash_at`: Cash-to-Assets (cash)
  - `debt_me`: Debt-to-Market Equity (leverage)
  - `exc_ret_avg`: Geometrically Weighted Average of Excess Returns (excess returns)
  - `me_binned`: Market Equity, transformed into a categorical variable with 5 bins (size)
  - `niq_at_avg`: Geometrically Weighted Average of Return on Assets (profitability)
  - `prc_capped`: Share Price, capped at $15 (price)
  - `rvol_21d_3m`: Return Volatility over 21 days, averaged over 3 months (volatility)
- **Profitability (7):** Additional measures of firm profitability, earnings quality, and overall financial health.
  - `at_be`: Book Leverage
  - `cop_atl1`: Cash-Based Operating Profits-to-Lagged Book Assets
  - `ebit_bev`: EBIT-to-Enterprise Value (Core Profitability)
  - `ebit_sale`: Profit Margin
  - `ni_be`: Net Income-to-Book Equity (Return on Equity)
  - `ocf_at`: Operating Cash Flow-to-Assets
  - `op_at`: Operating Profits-to-Book Assets
- **Value (6):** Measures comparing a firm's market valuation to its fundamental value or payouts.
  - `at_me`: Assets-to-Market Equity
  - `div12m_me`: Dividend Yield over 12 months
  - `eqnpo_me`: Net Payout Yield

  - `ni_me`: Earnings-to-Price (Earnings Yield)
  - `ocf_me`: Operating Cash Flow-to-Market
  - `sale_me`: Sales-to-Market

- **Momentum (2):** Indicators related to recent price trends.
  - `ret_12_1`: 12-Month Momentum (t-12 to t-1)
  - `prc_highprc_252d`: Price to 1-Year High Price

- **Trading Frictions (3):** Measures of market risk and trading activity (volatility, return and price measure included in Campbell variables).
  - `betadown_252d`: Downside Beta
  - `bidaskhl_21d`: High-Low Bid-Ask Spread
  - `dolvol_var_126d`: Dollar Volume Variability

- **Investment (8):** Indicators of firm growth and financing activities.
  - `at_gr1`: Asset Growth (1-Year)
  - `cowc_gr1a`: Change in Current Operating Working Capital (1-Year)
  - `dbnetis_at`: Net Debt Issuance-to-Assets
  - `eqnetis_at`: Net Equity Issuance-to-Assets
  - `eqnpo_12m`: Equity Net Payout (12 Months)
  - `fnl_gr1a`: Change in Financial Liabilities (1-Year)
  - `lti_gr1a`: Change in Long-Term Investments (1-Year)
  - `oaccruals_at`: Operating Accruals-to-Assets

- **Age (1):** A fundamental, non-financial characteristic of the firm.
  - `age`: Firm Age

- **Intangibles (4):** Measures related to the composition and liquidity of firm assets, as well as proxies for intangible value and past performance.
  - `aliq_at`: Liquidity of Book Assets
  - `aliq_mat`: Liquidity of Market Assets
  - `seas_1_1na`: Seasonality (Year 1 - Lagged Return, nonannual)
  - `tangibility`: Asset Tangibility

A.6 Macroeconomic Variables Selection

We used a univariate correlation test to assess the predictive power of macroeconomic features on monthly corporate failures and selected the 30 most predictive variables. From these, six

(highlighted in bold) entered the final dataset, representing the strongest feature set to avoid redundancy and multicollinearity and improve predictability.

- **sp500_ret_vol12m**: 0.2660
- sp500_ret_vol6m: 0.2525
- lt_interest_lag24: 0.2097
- **gdp_growth_avg4q**: 0.2084
- **st_interest_inflation_adj_lag24**: 0.2059
- st_interest_lag24: 0.1998
- lt_interest_lag18: 0.1917
- lt_interest_vol12m: 0.1804
- gdp_growth_avg8q: 0.1794
- lt_interest_lag12: 0.1762
- **yield_slope_avg6m**: 0.1675
- **unemployment_lag3**: 0.1669
- yield_slope_lag3: 0.1669
- yield_slope_avg12m: 0.1641
- st_interest_inflation_adj_lag18: 0.1633
- st_interest_lag18: 0.1628
- lt_interest_lag9: 0.1622
- unemployment_avg6m: 0.1570
- yield_slope_lag9: 0.1532
- **inflation_vol12m**: 0.1498
- lt_interest_avg12m: 0.1441
- yield_slope_lag6: 0.1396
- lt_interest_lag6: 0.1385
- gdp_growth_lag6: 0.1356
- yield_slope_vol12m: 0.1298
- lt_interest_vol6m: 0.1272
- st_interest_inflation_adj_lag12: 0.1269
- st_interest_lag12: 0.1236
- lt_interest_avg6m: 0.1225
- gdp_growth_lag9: 0.1210

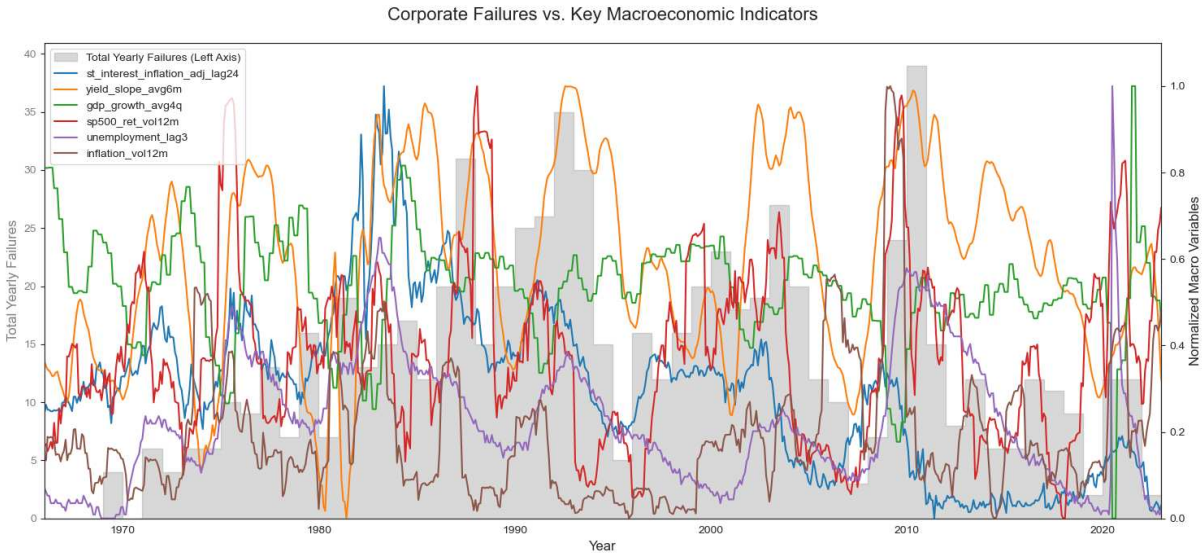
A.7 Yearly Failures with Overlaid Normalized Macroeconomic Indicators

Overall, Figure A.1 shows high corporate failure periods align closely with macroeconomic downturns. Some indicators, like the yield curve and unemployment, provide early warnings,

while others, including GDP growth and market volatility, move alongside observed defaults.

Figure A.1

Yearly Failures with All Normalized Macroeconomic Indicators Overlaid



A.8 Overview of Dataset Variables by Type

Table A.4 provides an overview of all final variables in the dataset, organized by type.

Table A.4

Overview of Dataset Variables by Type

Column Name	Description
Identifier / Time Index Variables	
`eom`	End of month - calendar date
`firm_id`	Unique firm identifier
Corporate Failure Indicator	
`failure`	Binary Failure (Bankruptcy or Liquidation) Indicator for `t+1`
Market / Accounting Variables	
`age`	Firm Age
`aliq_at`	Liquidity of Book Assets
`aliq_mat`	Liquidity of Market Assets
`at_be`	Book Leverage
`at_gr1`	Asset Growth (1-Year)
`at_me`	Assets-to-Market Equity
`be_me`	Book-to-Market Equity

Column Name	Description
betadown_252d	Downside Beta
bidaskhl_21d	High-Low Bid-Ask Spread
cash_at	Cash-to-Assets
cop_atl1	Cash-Based Operating Profits-to-Lagged Book Assets
cowc_gr1a	Change in Current Operating Working Capital (1-Year)
dbnetis_at	Net Debt Issuance-to-Assets
debt_me	Debt-to-Market Equity
div12m_me	Dividend Yield over 12 months
dolvol_var_126d	Dollar Volume Variability
ebit_bev	EBIT-to-Enterprise Value
ebit_sale	Profit Margin
eqnetis_at	Net Equity Issuance-to-Assets
eqnpo_12m	Equity Net Payout (12 Months)
eqnpo_me	Net Payout Yield
exc_ret_avg	Geometrically Weighted Average of Excess Returns
fnl_gr1a	Change in Financial Liabilities (1-Year)
industry_code	Industry Code (1-10)
issues	Number of simultaneous issues (Ordinal)
lti_gr1a	Change in Long-Term Investments (1-Year)
me_binned	Market Equity, transformed into a categorical variable with 5 bins
ni_be	Net Income-to-Book Equity (Return on Equity)
ni_me	Net Income-to-Price (Earnings Yield)
niq_at_avg	Geometrically Weighted Average of Profitability
oaccruals_at	Operating Accruals-to-Assets
ocf_at	Operating Cash Flow-to-Assets
ocf_me	Operating Cash Flow-to-Market
op_at	Operating Profits-to-Book Assets
prc_capped	Share Price, capped at $15
prc_highprc_252d	Price to 1-Year High Price
ret_12_1	12-Month Momentum (t-12 to t-1)
rvol_21d_3m	Return Volatility over 21 days, averaged over 3 months
sale_me	Sales-to-Market
seas_1_1na	Seasonality (Year 1 - Lagged Return, nonannual)
tangibility	Asset Tangibility
Macroeconomic Features	
gdp_growth_avg4q	GDP growth, averaged over 4 quarters
inflation_vol12m	Inflation volatility, over 12 months
sp500_ret_vol12m	S&P 500 return volatility, over 12 months

Column Name	Description
st_interest_inflation_-adj_lag24	Inflation-adjusted short-term interest rate, lagged 24 months
unemployment_lag3	Unemployment rate, lagged 3 months
yield_slope_avg6m	Yield curve slope, averaged over 6 months

A.9 Hyperparameter Tuning

This section provides a detailed overview of the hyperparameter tuning process. It includes the search spaces explored for each model and the results of the tuning experiments.

A.9.1 Logistic Regression

We performed hyperparameter tuning using `RandomizedSearchCV` with three-fold cross-validation on a random subsample of 50,000 observations from the training set. This approach efficiently explores the hyperparameter space by sampling parameter combinations rather than exhaustively searching all possibilities.

For logistic regression, we searched the following hyperparameter space:

- **solver:** {liblinear, saga}
- **penalty:** {l1, l2}
- **C (regularization strength):** {0.01, 0.1, 1, 10, 100}
- **max_iter:** {1000, 2000}

The tuning process identified optimal configurations for different variable sets:

- **Campbell dataset:**
  - solver = liblinear,
  - penalty = l1 (lasso),
  - C = 0.01,
  - max_iter = 1000,
  - best cross-val AUC: 0.8629
- **Full dataset:**

  - solver = saga,
  - penalty = l2 (ridge),
  - C = 0.1,
  - max_iter = 2000,
  - best cross-val AUC: 0.8497

The `liblinear` solver suits smaller feature spaces and supports L1 regularization, which performs variable selection by shrinking some coefficients exactly to zero. Conversely, the `saga` solver scales efficiently to larger datasets and uses L2 regularization, which shrinks coefficients toward zero without eliminating variables entirely, preserving information in high-dimensional data.

This tuning strategy balanced model complexity and interpretability while optimizing predictive performance across datasets of varying sizes.

A.9.2 Random Forest

Random Forest models were tuned separately for the Campbell and extended datasets, and for preprocessing applied or not, yielding four hyperparameter sets.

Before tuning, models showed overfitting, with training AUCs close to 1.0 but substantially lower validation AUCs (e.g., Campbell with preprocessing: train AUC = 1.00, val AUC = 0.7274).

We performed randomized search over 20 configurations using GroupKFold (3 folds) on a subsample of 100,000 firm-month observations, grouping by firm for temporal integrity. Key hyperparameters searched:

- **n_estimators:** {100, 200, 300, 500}
- **max_depth:** {3, 5, 7, 10, 15, 20, 30, 50, None}
- **min_samples_split:** {2, 5, 10, 20}
- **min_samples_leaf:** {1, 2, 4, 8, 16}
- **max_features:** {'sqrt', 'log2', 0.3}

Campbell dataset optimal configurations:

- Without preprocessing:
  - n_estimators: 300
  - min_samples_split: 2

  - min_samples_leaf: 8
  - max_features: log2
  - max_depth: 3
  - Best cross-val AUC: 0.8756
- With preprocessing:
  - n_estimators: 500
  - min_samples_split: 10
  - min_samples_leaf: 16
  - max_features: sqrt
  - max_depth: 30
  - Best cross-val AUC: 0.8798

Extended dataset optimal configurations:

- Without preprocessing:
  - n_estimators: 300
  - min_samples_split: 10
  - min_samples_leaf: 1
  - max_features: log2
  - max_depth: 5
  - Best cross-val AUC: 0.9125
- With preprocessing:
  - n_estimators: 300
  - min_samples_split: 2
  - min_samples_leaf: 16
  - max_features: 0.3
  - max_depth: 5
  - Best cross-val AUC: 0.9142

Preprocessing enabled deeper, larger forests without overfitting in the smaller dataset, but the extended dataset favored shallow trees regardless of preprocessing to mitigate overfitting. Preprocessing effects reversed between datasets: it increased complexity in the Campbell case but promoted stronger regularization in the extended case. Overall, extended dataset models attained superior predictive performance, reinforcing the value of the richer feature set.

A.9.3 XGBoost

We performed hyperparameter tuning for XGBoost using a random subsample of 100,000 firm-month observations, preserving temporal integrity by grouping at the firm level. We employed

GroupKFold cross-validation (3 folds) and randomized search over the main hyperparameter space:

- **n_estimators:** {100, 200, 300, 500}
- **max_depth:** {3, 5, 7, 10, 15}
- **learning_rate:** {0.01, 0.05, 0.1, 0.2, 0.3}
- **subsample:** {0.7, 0.8, 0.9, 1.0}
- **colsample_bytree:** {0.3, 0.5, 0.7, 0.8, 1.0}
- **gamma:** {0, 1, 5, 10}
- **reg_alpha (L1):** {0, 0.1, 1, 5}
- **reg_lambda (L2):** {1, 1.5, 2, 5, 10}

Campbell dataset optimal configurations:

- Without preprocessing:
  - n_estimators: 300
  - max_depth: 15
  - learning_rate: 0.05
  - subsample: 1.0
  - colsample_bytree: 0.7
  - reg_alpha: 1
  - reg_lambda: 1
  - gamma: 0
  - Best cross-val AUC: 0.8567
- With preprocessing:
  - n_estimators: 300
  - max_depth: 15
  - learning_rate: 0.05
  - subsample: 1.0
  - colsample_bytree: 0.7
  - reg_alpha: 1
  - reg_lambda: 1
  - gamma: 0
  - Best cross-val AUC: 0.8629

Extended dataset optimal configurations:

- Without preprocessing:
  - n_estimators: 200
  - max_depth: 3

  - learning_rate: 0.01
  - subsample: 0.9
  - colsample_bytree: 0.8
  - reg_alpha: 1
  - reg_lambda: 2
  - gamma: 5
  - Best cross-val AUC: 0.8689
- With preprocessing:
  - n_estimators: 300
  - max_depth: 5
  - learning_rate: 0.1
  - subsample: 1.0
  - colsample_bytree: 0.5
  - reg_alpha: 1
  - reg_lambda: 5
  - gamma: 10
  - Best cross-val AUC: 0.9257

XGBoost models on the Campbell dataset favored very deep trees with modest regularization; preprocessing yielded slight performance gains without changing the structure significantly. The extended dataset required shallower trees, lower learning rates, and stronger regularization to control overfitting. Preprocessing enabled increased tree depth and learning rates combined with stricter split penalties and feature subsampling. The extended preprocessed model achieved the highest cross-validation AUC (0.9257), illustrating the benefit of richer features and tailored training strategies.

These results highlight the importance of combining rich feature sets with appropriate preprocessing and model tuning to optimize predictive performance.

A.10 Results

A.10.1 Model Performance

This section in the appendix provides additional diagnostic metrics beyond those presented in the Chapter 5, offering a more granular view of model behavior.Tables A.5, A.6, and A.7 report the main results from 30 modeling runs (with Logistic Regression (LR), Random Forest (RF) and XGBoost (XGB)). The evaluation focuses on two key metrics: AUC, which reflects overall model performance, and the *Type I error rate*, which captures the ability to correctly identify

corporate failures (the minority class). Higher AUC values indicate stronger predictive accuracy, whereas lower Type I error rates signal better failure detection.

Each model was first trained on the training dataset (train: 1965–1994) and validated on the validation dataset (val: 1995–2007). A second iteration combined the training and validation sets (train + val: 1965–2007) to re-train the models, which were then evaluated on the test dataset (test: 2008–2022). The results highlight differences across modeling approaches (with ensemble methods generally outperforming logistic regression), variable sets (with ensemble models performing best on the Full set, and logistic regression performing well on the Campbell set), and prediction horizons (with performance degrading at longer horizons). They also reveal some differences between models trained with and without preprocessing.

Table A.5

Model performance at $t + 1$ horizon by model and variable set, with and without preprocessing. Metrics reported are the AUC and (Type I error).

Model	Variables	train	val	train + val	test
With Preprocessing					
LR	Campbell	0.8639 *(0.21)*	0.9392 *(0.09)*	0.8910 *(0.17)*	0.9354 *(0.14)*
LR	Full	0.9061 *(0.21)*	0.9360 *(0.13)*	0.9234 *(0.17)*	0.8990 *(0.12)*
RF	Campbell	1.0000 *(0.00)*	0.9305 *(0.97)*	1.0000 *(0.00)*	0.9516 *(0.95)*
RF	Full	0.9528 *(0.20)*	0.9554 *(0.15)*	0.9517 *(0.17)*	0.9560 *(0.16)*
XGB	Campbell	1.0000 *(0.00)*	0.9271 *(0.99)*	1.0000 *(0.00)*	0.9552 *(0.98)*
XGB	Full	0.8709 *(0.62)*	0.5879 *(0.96)*	0.9996 *(0.00)*	0.8900 *(0.72)*
Without Preprocessing					
RF	Campbell	0.9077 *(0.19)*	0.9409 *(0.14)*	0.9187 *(0.18)*	0.9492 *(0.13)*
RF	Full	0.9530 *(0.15)*	0.9546 *(0.17)*	0.9543 *(0.14)*	0.9616 *(0.16)*
XGB	Campbell	1.0000 *(0.00)*	0.9240 *(0.99)*	1.0000 *(0.00)*	0.9549 *(0.98)*
XGB	Full	0.9529 *(0.12)*	0.9503 *(0.15)*	0.9537 *(0.15)*	0.9607 *(0.14)*

Table A.6

Model performance at $t + 6$ horizon by model and variable set, with and without preprocessing. Metrics reported are the AUC and (Type I error).

Model	Variables	train	val	train + val	test
With Preprocessing					
LR	Campbell	0.8373 *(0.21)*	0.8737 *(0.18)*	0.8525 *(0.20)*	0.8953 *(0.16)*
LR	Full	0.8715 *(0.22)*	0.9260 *(0.16)*	0.8786 *(0.22)*	0.8831 *(0.16)*
RF	Campbell	1.0000 *(0.00)*	0.8510 *(1.00)*	1.0000 *(0.00)*	0.9149 *(1.00)*
RF	Full	0.9252 *(0.21)*	0.8995 *(0.27)*	0.9162 *(0.21)*	0.9239 *(0.18)*
XGB	Campbell	1.0000 *(0.00)*	0.8506 *(1.00)*	1.0000 *(0.00)*	0.8988 *(1.00)*
XGB	Full	0.9989 *(0.00)*	0.9499 *(0.44)*	0.9967 *(0.00)*	0.9029 *(0.64)*
Without Preprocessing					
RF	Campbell	0.8789 *(0.19)*	0.8785 *(0.20)*	0.8807 *(0.18)*	0.9110 *(0.14)*
RF	Full	0.9299 *(0.16)*	0.8990 *(0.26)*	0.9211 *(0.16)*	0.9303 *(0.17)*
XGB	Campbell	1.0000 *(0.00)*	0.8566 *(1.00)*	1.0000 *(0.00)*	0.9035 *(1.00)*
XGB	Full	0.9253 *(0.13)*	0.8900 *(0.15)*	0.9170 *(0.14)*	0.9212 *(0.15)*

Table A.7

Model performance at $t + 12$ horizon by model and variable set, with and without preprocessing. Metrics reported are the AUC and (Type I error).

Model	Variables	train	val	train + val	test
With Preprocessing					
LR	Campbell	0.8037 *(0.22)*	0.8052 *(0.29)*	0.8065 *(0.23)*	0.8493 *(0.22)*
LR	Full	0.8119 *(0.27)*	0.7960 *(0.38)*	0.8362 *(0.25)*	0.8639 *(0.22)*
RF	Campbell	1.0000 *(0.00)*	0.7856 *(1.00)*	1.0000 *(0.00)*	0.8550 *(1.00)*
RF	Full	0.8986 *(0.22)*	0.8416 *(0.39)*	0.8844 *(0.23)*	0.8899 *(0.18)*
XGB	Campbell	1.0000 *(0.00)*	0.7753 *(1.00)*	1.0000 *(0.00)*	0.8500 *(1.00)*
XGB	Full	0.9988 *(0.00)*	0.7976 *(0.89)*	0.9962 *(0.00)*	0.8718 *(0.70)*
Without Preprocessing					
RF	Campbell	0.8439 *(0.21)*	0.8202 *(0.28)*	0.8394 *(0.20)*	0.8715 *(0.18)*
RF	Full	0.9010 *(0.20)*	0.8780 *(0.36)*	0.8887 *(0.19)*	0.8997 *(0.21)*
XGB	Campbell	1.0000 *(0.00)*	0.7672 *(1.00)*	1.0000 *(0.00)*	0.8331 *(1.00)*
XGB	Full	0.8916 *(0.17)*	0.8429 *(0.32)*	0.8800 *(0.19)*	0.8942 *(0.18)*

A.10.2 Confusion Matrices

The following confusion matrices in Tables A.8, A.9 and A.10 extend the results presented in Chapter 5. They summarize classification outcomes across all models (Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB)), both variable sets (Campbell and Full), and all prediction horizons ($t+1$, $t+6$, and $t+12$). For ensemble models, results are reported with and without preprocessing. Each matrix corresponds to the test evaluation, where models were trained on the combined training and validation datasets (1965–2007) and evaluated on the test set (2008–2022).

The confusion matrices report the four possible classification outcomes: true negatives (non-failing firms correctly predicted), false positives (non-failing firms incorrectly predicted as failures), false negatives (failing firms incorrectly predicted as non-failures), and true positives (failing firms correctly predicted).

Table A.8

Test Confusion matrices at $t+1$ horizon by model and variable set, with and without preprocessing.

Model	Variable Set	Pred. Horizon	True **Negatives** (TN)	False **Positives** (FP)	False **Negatives** (FN)	True **Positives** (TP)
With Preprocessing						
LR	Campbell	$t+1$	552,216	98,635	51	319
LR	Full	$t+1$	471,366	179,485	43	327
RF	Campbell	$t+1$	650,675	176	350	20
RF	Full	$t+1$	599,682	51,169	59	311
XGB	Campbell	$t+1$	650,768	83	361	9
XGB	Full	$t+1$	646,803	4,048	265	105
Without Preprocessing						
RF	Campbell	$t+1$	576,986	73,865	49	321
RF	Full	$t+1$	602,563	48,288	58	312
XGB	Campbell	$t+1$	650,759	92	363	7
XGB	Full	$t+1$	586,720	64,131	53	317

Table A.9

Test Confusion matrices at $t + 6$ horizon by model and variable set, with and without preprocessing.

Model	Variable Set	Pred. Horizon	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
With Preprocessing						
LR	Campbell	$t + 6$	523,499	125,525	59	303
LR	Full	$t + 6$	522,272	126,752	57	305
RF	Campbell	$t + 6$	649,011	13	362	0
RF	Full	$t + 6$	562,416	86,608	65	297
XGB	Campbell	$t + 6$	648,987	37	362	0
XGB	Full	$t + 6$	635,105	13,919	231	131
Without Preprocessing						
RF	Campbell	$t + 6$	528,656	120,368	52	310
RF	Full	$t + 6$	561,783	87,241	63	299
XGB	Campbell	$t + 6$	649,008	16	361	1
XGB	Full	$t + 6$	541,729	107,295	54	308

Table A.10

Test Confusion matrices at $t + 12$ horizon by model and variable set, with and without preprocessing.

Model	Variable Set	Pred. Horizon	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
With Preprocessing						
LR	Campbell	$t + 12$	501,233	145,707	74	263
LR	Full	$t + 12$	528,939	118,001	73	264
RF	Campbell	$t + 12$	649,940	0	337	0
RF	Full	$t + 12$	532,204	114,736	72	265
XGB	Campbell	$t + 12$	646,934	6	337	0
XGB	Full	$t + 12$	630,967	15,973	235	102
Without Preprocessing						
RF	Campbell	$t + 12$	492,460	154,480	59	278
RF	Full	$t + 12$	531,150	115,790	72	265
XGB	Campbell	$t + 12$	646,938	2	337	0
XGB	Full	$t + 12$	508,973	137,967	59	278