The Report Committee for Soorya Sriram
certifies that this is the approved version of the following report:

Early Warning Credit Risk Default Prediction using Machine Learning for Small and Medium American Businesses

SUPERVISING COMMITTEE:

Dr. Guoming Lai, Supervisor

Dr. John Hasenbein, Co-supervisor

Early Warning Credit Risk Default Prediction using Machine Learning for Small and Medium American Businesses

by

Soorya Sriram

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Operations Research and Industrial Engineering

The University of Texas at Austin
December 2024

Epigraph

What starts here changes the world.

—University of Texas at Austin

Acknowledgments

I would like to express my sincerest gratitude to my academic supervisor Dr. Guoming Lai from the Information, Risk and Operations Management Department from the McCombs School of Business. I thank him for giving me the opportunity to work under his guidance and providing me with the flexibility in allowing to pursue the topic and project of my choice with utmost freedom. His constant encouragement was the main driving force for me to complete the project within the time provided. I would like to thank him for his patience and efforts put in to take up doubts.

I would like to thank Dr. John Hasenbein, graduate advisor of Operations Research and Industrial Engineering department for his valuable advice throughout my graduate studies and for being the co-supervisor for this report and graduate studies.

I would like to thank my undergraduate thesis professor Dr. Kalpana P, from Indian Institute of Information Technology Design and Manufacturing for cultivating an interest in Operations Research.

I would like to thank all the professors at UT Austin ORIE specifically Dr. Kutanoglu and Dr. Eric Bickel for their guidance during the ORIE Applied Projects courses enabling me to learn new skills and work with industry stalwarts to pick their brains working on cutting edge problems.

Last but not the least, I would like to thank my family, fellow friends and ORIE classmates for their continuous support during my graduate studies.

Abstract

Early Warning Credit Risk Default Prediction using Machine Learning for Small and Medium American Businesses

Soorya Sriram, MS ORIE
The University of Texas at Austin, 2024

SUPERVISORS: Dr. Guoming Lai, Dr. John Hasenbein

There has been a lot of research and variants in credit risk prediction in the past using traditional operations research and quantitative analytical models. The purpose of this applied research study was to understand industry standard modern machine learning models used for small and medium American businesses. The goal was to analyze 50 organizations belonging to five different sectors namely retail, fmcg, transportation, commercial services and utilities over 11 years from 2013-2023. Scraping 27 financial indicators from publicly available yet reliable sources, a study was made to identify the financial indicators which has the most impact on the accuracy of the model. Grey correlation analysis was used to find and rank weighted financial indicators which were then used to identify the number of principal components. Evaluating a number of different models, ensemble methods and correlation subsets to find the best combination of machine learning models, explainable AI techniques were used to perform a sector wise analysis on the misclassified data to derive valuable insights.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Credit risk management has long been a critical focus for banks and financial institutions worldwide, where they try to evaluate the likelihood of default by individuals and organizations. Credit risk assessment spans the entire life cycle of a loan, from initial application to final repayment. Financial institutions consider several key factors during their evaluation process, including repayment capacity, credit history, collateral, and asset holdings. In this study, the focus is on five of the most volatile industries, namely retail, fmcg, commercial services, transportation, and utilities, which have experienced significant changes post-COVID and have been heavily impacted by advances in AI. Traditionally, rule-based systems have been used for credit risk assessment, and while they have been effective, they often fail to identify complex patterns in customer spending, cash flow, new loans, deposits, and repayment behaviors. This research investigates the use of industry-standard Machine Learning models for credit risk assessment, comparing statistical approaches like logistic regression and linear discriminant analysis against a diverse selection of base classifiers. The goal is to improve predictive accuracy and capture the nuanced patterns that traditional methods tend to overlook.

1.1 Background

The credit risk project aims to develop an early warning system to predict corporate defaults by using advanced machine learning techniques. It is a critical concern for financial institutions, investors, and regulatory bodies. Accurate prediction models help mitigate potential losses by identifying high-risk factors before default occurs.

In this project, historical financial data for 50 companies over 11 years (2013-2023) is analyzed, focusing on key financial ratios and market indicators such as profitability, leverage, and liquidity. The models incorporate feature selection, dimensionality reduction through principal component analysis (PCA), and hyperparameter-tuned machine learning algorithms to predict defaults for the latest year. The project also explores the influence of industry-specific factors on default risk.

1.2 Motivation

The motivation for this study stems from the increasing need for more accurate and dynamic credit risk assessment models in the financial industry, particularly in the wake of the COVID-19 pandemic, which disrupted markets and heightened financial uncertainty. Traditional methods of credit risk evaluation, while useful, often fail to capture the complexities of modern financial behaviors, especially in industries that are highly volatile and influenced by emerging technologies like AI. With the rapid advancements in machine learning and artificial intelligence, there is a growing opportunity to leverage these tools to better predict defaults by identifying patterns in financial data that would otherwise remain undetected. This study aims to explore the potential of machine learning models to improve credit risk predictions and provide more robust risk management strategies for financial institutions, ultimately contributing to greater financial stability and more informed lending decisions.

1.3 Objectives

- Develop an early warning system for predicting corporate defaults using machine learning techniques, focusing on companies within five of the most volatile industries affected post-COVID namely retail, fmcg, transportation, utilities and Commercial Services.

- Use feature selection and dimensionality reduction techniques, such as Mutual Information and principal component analysis (PCA), to improve model interpretability and performance.

- Compare the performance of traditional statistical models, such as Logistic Regression and Linear Discriminant Analysis, with advanced machine learning models, including CatBoost, to determine the most effective approach for credit risk prediction.
- Analyze key financial ratios, market indicators, and industry-specific factors to assess their impact on credit risk.
- Utilise Explainable AI techniques to better understand the impact of certain factors on the prediction.

Chapter 2: Literature Review

2.1 Machine Learning for Credit Risk

Noriega et al. (2023) have mentioned in their research the latest developments in using Machine Learning for credit risk applications. Some factors that often influence credit risk prediction are risk level, credit terms and rate, fraud potential, sales data consistency, cultural and environmental influences, macroeconomic indicators (e.g., GDP growth, exchange rates), innovation capacity, economic activity, and organizational experience. They mentioned that relevant predictors to include would be borrower demographics (gender, education level), mortgage credit status, microfinance, debt balance, days past due, and days of default, particularly defaults exceeding 90 days.

Financial institutions often prioritize high-risk, income-generating loans due to the potential for greater profits but the tradeoff is that there are higher chances of payment defaults. Incorporating interactions within borrowers' supply chains improves predictive models, adding context to credit risk assessments. Utilizing diverse data sources enables a more holistic evaluation, incorporating dimensions like profitability, liquidity, asset quality, management indices, capital adequacy, and industry risks.

Unbalanced Data is a major problem faced and most common techniques used for addressing imbalanced data include Random Under Sampling, Near Miss, Cluster Centroid, Random Oversampling, Adaptive Synthetic, K-Nearest Neighbors Imputation, K-Nearest Neighbors SMOTE, SMOTE, Borderline SMOTE, and K-Fold. Optimizing hyperparameters through Genetic Algorithms, K-Fold CV, Random Search, and Grid Search is emphasized to enhance model accuracy. Metrics used include ROC - AUC, Accuracy, F1 score, precision, recall, true positive/negative rates, geometric mean, Kolmogorov-Smirnov, Brier Score, Gini Index, RMSE, Kappa, and MAE, with

ROC - AUC prioritized for unbalanced datasets.

Explainable AI is often used where the objective is to provide predictive models with transparency. Shapley Additive Explanations (SHAP) and Generalized Shapley Choquet Integral (GSCI) are used to make models transparent by revealing parameter dependencies. Methods such as penalized logistic trees and Slack-Based Measures aim to improve Logistic Regression interpretability and performance. Models like MLIA and logistic regression with variable coefficients enhance interpretability. Techniques such as genetic algorithms, boosting methods (AdaBoost, XGBoost, LightGBM), and bagging are explored for model optimization and performance enhancement.

2.2 Ensemble Methods for Credit Risk Performance

The research in Wang and Zhang (2023) indicates that sample index system and early warning models are two major research directions in credit risk prediction. Most credit risk early warning models do not interpret results well and qualitative analysis was conducted on the basis of prediction results a research focus. Feature selection and extraction are common in constructing the sample index system. They created a synthetic dataset where special treatments are chosen as samples with credit risk while normal listed companies are regarded as samples without credit risk. High-risk samples which included companies listed for over five years, which first received special treatment between 2012 and 2021. Normal samples include companies listed for over five years with no special treatment during that time.

Corporate credit risk early warning with financial indicator data in the form of time series, ranks financial institutions data from every company where entropy is used to weigh these indicators more effectively, improving the grey correlation analysis's accuracy in assessing each indicator's impact on credit risk. By using an entropy method, the weight distribution for each evaluation indicator is recalculated, leading to a more balanced indicator weighting and enhancing predictive accuracy. The process includes normalizing the input matrix, calculating entropy values for each indicator, and determining the improved entropy weight. Key steps include determining the sequence type, normalizing data, calculating correlation coefficients, and computing the grey relational degree for each factor.

2.3 Importance of Interpretability

Chen et al. (2024) mentions techniques like Local Interpretable Model-Agnostic Explanations (LIME) and SHAP are used to make advanced credit scoring models more interpretable, providing insights into model predictions. There is often a trade-off between improved accuracy and reduced interpretability when using complex credit scoring models.

Stability of model interpretation is evaluated through Sequential Risk Agreement and Coefficient of Variation, ensuring the consistency of interpretability across model versions or datasets. The Coefficient Stability Index and Variables Stability Index are introduced to measure the robustness of model coefficients and selected features over time, supporting the reliability of the model's predictive features.

2.4 Optimization Models for Credit Risk

Liang and Wang (2021) model priced vulnerable options by incorporating stochastic volatility using the Heston-Nandi GARCH process. The process estimates conditional variances for the underlying asset and issuer based on observable market prices. The model includes both idiosyncratic and systematic risks within the asset dynamics of both the underlying asset and the option issuer, providing a more comprehensive risk assessment. Two primary model types are used to analyze default risk in options: structural models (where default occurs when the issuer's asset value is lower than the option's value at maturity) and reduced-form models (where default is defined by a rate or intensity of default).

Vulnerable options are derivatives that carry the risk of issuer default. They differ from standard options due to the possibility that the issuer may fail to meet obligations if their asset value falls below a threshold. Some key research mentioned here is the following

- Hull and White (1995) proposed a pricing formula for vulnerable options using an independence assumption.
- Fard (2015) developed a closed-form solution assuming a mean-reverting default intensity.
- Antonelli et al. (2021) used correlation expansion to approximate option prices.
- Wang (2017) derived a closed-form solution in a discrete-time GARCH setting.
- Bakshi et al. (2006) introduced a reduced-form model using Vasicek variables (e.g., leverage, book-to-market ratio).
- Gu et al. (2014) enhanced reduced-form models to include impacts from trigger events and economic conditions.
- Boudreault et al. (2014) analyzed contagion effects, with default time and recovery rate tied to the firm's leverage.

2.5 Correlation Based Classifier Selection

The study by Xiong and Huang (2022) introduces the Maximum Information Coefficient-based Correlation Classifier Selection (MIC-CCS), a method that uses nonlinear optimization programming to select classifiers with low inter-classifier correlation and high correlation with the target variable. MIC-CCS based ensemble models outperform those built with unselected classifiers and other methods (Pearson correlation, forward selection, backward selection), with MIC-CCS delivering better accuracy due to its nonlinear focus.

Chapter 3: Keywords and Organizations

3.1 Financial Indicators in Dataset

- Market Capitalization (Market Cap): Market capitalization is the total value of a company's outstanding shares of stock, representing the market's perception of its overall value. It is calculated by multiplying the current stock price by the total number of outstanding shares. Market Cap is often used as an indicator of a company's size, with companies classified as large-cap, mid-cap, or small-cap based on their total valuation.

- Enterprise Value: EV is a measure of a company's total value, includes Market Cap as well as short, long term debt and preferred equity without the cash equivalents. It represents the actual cost to acquire a company accounting for both its debt obligations and available liquid assets.

- Price to Earnings Ratio: P/E Ratio is a valuation metric that measures a company's current share price relative to its earnings per share and is used to gauge whether a stock is overvalued or undervalued.

- Price to Sales Ratio: P/S Ratio is a valuation metric that compares a company's stock price to its revenue per share. It reflects how much investors are willing to pay for each dollar of sales generated by the company.

- Price to Book Ratio: P/B Ratio is a valuation metric that compares a company's current market price to its book value per share. It is commonly used to assess whether a stock is overvalued or undervalued relative to the net assets of the company.

- Price to Free Cash Flow: P/FCF Ratio is a valuation metric that compares a company's market price to its free cash flow per share assessing how much investors are willing to pay for each dollar of free cash flow generated by the company.

- Price to Operating Cash Flow: P/OCF Ratio is a valuation metric that compares a company's market price to its operating cash flow per share.
- Enterprise Value to Sales Ratio: EV/Sales Ratio is a valuation metric that compares a company's enterprise value to its revenue.
- Enterprise Value to EBITDA: EV/EBITDA Ratio is a valuation metric that compares a company's enterprise value to its earnings before interest, taxes, depreciation and amortization. It helps one to understand how many times a company's EBITDA the market is willing to pay for the business.
- Enterprise Value to EBIT: EV/EBIT Ratio is a valuation metric that compares a company's enterprise value to its earnings before interest and taxes. It helps one to understand the value of a company in relation to its core operated earnings providing an insight to its valuation without the influence of capital structure and taxes.
- Enterprise Value to Free Cash Flow: EV/FCF Ratio is a valuation metric that compares a company's enterprise value to its free cash flow. It helps one understand how much investors are willing to pay for each dollar of free cash flow the company generates providing an insight into the valuation relative to the actual cash available after operations expenses and capital expenditures.
- Debt to Equity: Debt/Equity Ratio is a valuation metric that compares a company's total liabilities to its shareholders' equity helping one understand the degree to which a company is financing its operations through debt versus wholly owned funds.
- Debt to EBITDA: Debt/EBITDA Ratio is a financial metric used to assess a company's ability to pay off its debt using its earnings before accounting for

interest, taxes, depreciation and amortization providing insight into a company's leverage and financial health.

- Debt to Free Cash Flow: Debt/FCF Ratio is a financial metric use to evaluate a company's ability to pay off its debt using the cash it generates from operations after accounting for capital expenditures providing insight into the company's liquidity and financial health.

- Quick Ratio: Quick Ratio is a financial metric that assesses a company's ability to meet its short term liabilities using its liquid assets. It is measured as the difference between its current assets and inventory relative to its current liabilities.

- Current Ratio: Current Ratio is a liquidity ratio that measures a company's ability to pay off its short-term liabilities with its short term assets providing insight into the financial health of a company and its capacity to cover obligations that are due within a year.

- Asset Turnover: Asset Turnover is a financial ratio that measures the efficiency of a company's use of its assets in generating sales revenue indicating the effectiveness of the a company in utilizing its assets to produce revenue and gauge operational performance.

- Return on Equity: Return on Equity is a financial metric that measures a company's ability to generate profits from its shareholders' equity indicating how effectively the company is using the equity invested by shareholders to generate earnings.

- Return on Assets: Return on Assets is a financial metric that measures how efficiently a company is using its assets to generate profits indicating how the management is converting the company's total assets into net income, providing insights into operational efficiency and profitability.

- Return on Invested Capital: ROIC is a profitability ratio that measures the return generated on the total capital invested in a company. It is calculated as the Net Operating Profit After Taxes or the operating income excluding the effects of interest relative to the invested capital which includes equity, debt or other capital.

- Earnings Yield: Earnings Yield is a financial metric that represents the earnings generated by each dollar invested in a stock. It is inverse of the PE Ratio represented as a percentage.

- Free Cash Flow Yield: FCF Yield is a financial ratio that measures a company's free cash flow per share relative to its market price per share providing an insight into how much free cash flow a company is generating in relation to its market valuation.

- Dividend Yield: Dividend Yield is a financial ratio that indicates how much a company pays out in dividends each year relative to its stock price and is a measure of the return on investment for shareholders from dividend payments.

- Payout Ratio: Payout Ratio is a financial metric that indicates the proportion of a company's earnings that is distributed to shareholders as dividends providing an insight into how much profit a company returns to its investors versus how much it retains for reinvestment, debt reduction or for other reason.

- Buyback Yield/Dilution: Buyback Yield measures the effect of share repurchases on the returns to shareholders, indicating the rate at which a company is buying back its own shares relative to its market capitalization. Dilution measures the impact of issuing additional shares, which can reduce existing shareholders' ownership percentage and potentially decrease earnings per share.

- Total Shareholder Return: Total Shareholder Return measures the overall financial benefit gained by shareholders, combining both capital gains (stock price increase) and income (dividends) over a specific period.

- Debt Issuance: Debt Issuance is the process by which a company raises capital by borrowing funds through the sale of bonds, notes or other debt instruments to investors allowing the organization to finance its operations, invest in project or restructure existing debt without diluting ownership through equity financing.

3.2 Training and Test Set

50 Small and Medium Size Businesses selected belonging to 5 different sectors mainly Energy, Retail, Commercial Services, Utilities and Transportation Services:

1. Amplify Energy Corp
2. DMC Global Inc
3. NACCO Industries, Inc
4. U.S. Silica Holdings, Inc
5. Teekay Corporation
6. Safe Bulkers, Inc
7. Tsakos Energy Navigation Limited
8. SFL Corporation Ltd
9. World Kinect Corporation
10. Navigator Holdings Ltd
11. Pitney Bowes Inc
12. Barrett Business Services, Inc
13. Viad Corp
14. VSE Corporation

15. The GEO Group, Inc

16. ICF International, Inc

17. MillerKnoll, Inc

18. CoreCivic, Inc

19. TTEC Holdings, Inc

20. Matthews International Corporation

21. Weyco Group, Inc

22. Winmark Corporation

23. Tile Shop Holdings, Inc

24. Monro, Inc

25. Guess', Inc

26. Sonic Automotive, Inc

27. The ODP Corporation

28. Hibbett, Inc

29. The Children's Place, Inc

30. Winmark Corporation

31. Universal Corporation

32. B&G Foods, Inc

33. Adecoagro S.A.

34. Fresh Del Monte Produce Inc

35. National Beverage Corp

36. Flowers Foods, Inc

37. Sanfilippo (John B.) & Son, Inc

38. Calavo Growers, Inc

39. Alico, Inc

40. Lifeway Foods, Inc

41. Universal Logistics Holdings, Inc

42. Jabil Inc

43. SkyWest, Inc

44. Allegiant Travel Company

45. Euroseas Ltd

46. Costamare Inc

47. Hawaiian Holdings, Inc

48. Golden Ocean Group Limited

49. Danaos Corporation

50. Limoneira Company

Chapter 4: Experiment and Results

4.1 Data Preprocessing

1. **Data Scraping:** Scraped data for 50 companies from 2013 - 2023 with 10 companies belonging to each sector namely Retail, FMCG, Commercial Services, Utilities and Transportation from publicly available source https://stockanalysis.com/

2. **Data Cleaning**:

   (a) Performed K-Nearest Neighbors Imputation which is used to fill in missing values based on the values of similar (neighboring) data points with a k=2 (closest observations).

   (b) Once the missing values in the numerical columns were filled, performed Outlier Analysis using Z scores > 3 and replaced outliers with the median of that features belonging to that specific organization.

   (c) Performed Standard Scaling and converted percentage columns back to decimals.

3. **Default Assignment**: Considering the following features and their values, whether that organization in that year would default or not was assigned (binary classes) as shown in Figure 4.1

   (a) Debt/Equity Ratio: Benchmark: Less than 1 is considered conservative. Over 2 may be considered high.

   (b) Debt/EBITDA Ratio: Benchmark: Generally, below 3 is considered good, but this can vary by industry.

   (c) Debt/FCF Ratio: Benchmark: Below 1 is considered favorable, but this can vary by industry.

(d) Quick Ratio: Benchmark: A ratio of 1 or higher is generally considered good. Below 1 may indicate potential liquidity issues.

(e) Current Ratio: Benchmark: A ratio of 2 or higher is generally considered good. Below 1 may indicate potential liquidity issues.

(f) Interest Coverage: Benchmark: A ratio of 3 or higher is generally considered good.

(g) Total Shareholder Return: Benchmark: Positive values are generally desired. A negative value may signal challenges.

(h) Earnings Yield: Benchmark: Positive values are expected. Negative values are uncommon.

(i) FCF Yield: Benchmark: Positive values are expected. A higher yield may be favorable.

(j) Payout Ratio: Benchmark: Below 60% is often considered reasonable, but this can vary by industry.

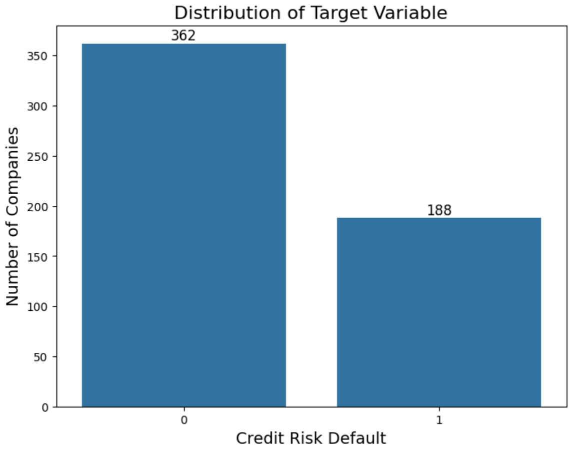


Figure 4.1: Distribution of Target Variable after Default Assignment

4. **Initial Correlation between Features:** A Correlation test was performed using a Heatmap as shown in Figure 4.2 to make sure the features were independent and were not correlated. Having a threshold of 0.8, Current Ratio and

Quick Ratio had a correlation of 0.806 and hence was allowed to remain the dataset for further experiments.

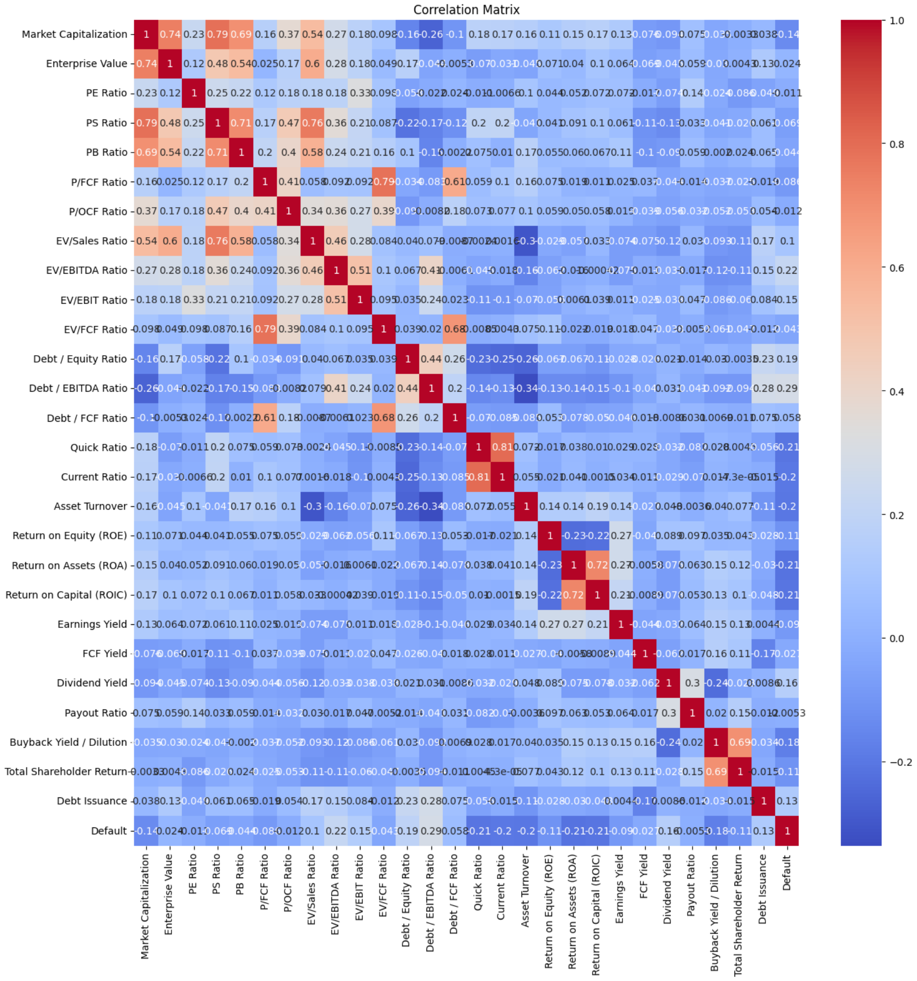


Figure 4.2: Correlation between Financial Indicators

4.1.1 Train Set - Target Distribution

5 sectors - 50 Companies - (2012 - 2022) 500 Data Points with 30 Features inclusive of Company Name, Ticker and Year columns as shown in Figure 4.3

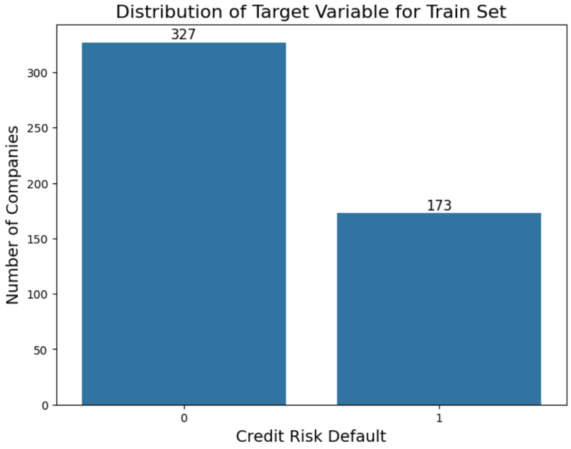


Figure 4.3: Distribution of Target Variable after Default Assignment

4.1.2 Test Set - Target Distribution

5 sectors - 50 companies - (2023) 50 Data Points with 30 Features inclusive of Company Name, Ticker and Year columns as shown in Figure 4.4

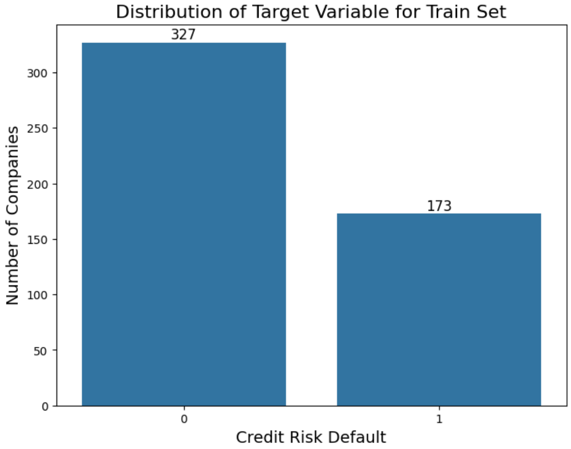


Figure 4.4: Distribution of Target Variable after Default Assignment

4.2 Grey Correlation Analysis

Grey Correlation Analysis (GCA) is a method used to quantify the strength of relationships between variables when data may be uncertain, incomplete, or complex.

Grey Correlation Analysis is applied to assess the relationships between financial indicators in the scaled training dataset.

1. **Entropy Based Weight Calculation:** Entropy reflects the amount of information a feature contains, higher entropy generally means more variation within that feature, and thus it has more impact on the analysis. Weights are normalized values of entropy, helping you assign a relative importance to each feature.
2. **Weighted Grey Correlation Matrix:** Weighted sum gives a Grey Correlation matrix that reflects the overall correlation level between each pair of features.
3. **Ranking Financial Indicators:** High-ranking indicators are strongly associated with other variables indicating that it is influential or central to the dataset's structure.
4. **Results:** Earnings Yield, FCF Yield, Return on Capital, Return on Assets and Return on Equity were ranked the highest while PB Ratio, P/OCF Ratio, EV/EBITDA Ratio, EV/Sales Ratio and PS Ratio were the least.

4.3 Principal Correlation Analysis

Using the weighted training set obtained after Grey Correlation Analysis, selecting a cumulative variance of 0.95 as the threshold, 15 principal components were selected and used for Model Selection as shown in Figure 4.5

4.3.1 Final Correlation between Principal Components:

The advantage of using Principal Components in Model Selection is that every Principal Component is independent and has low correlation between components however interpretability of the principal components is difficult.

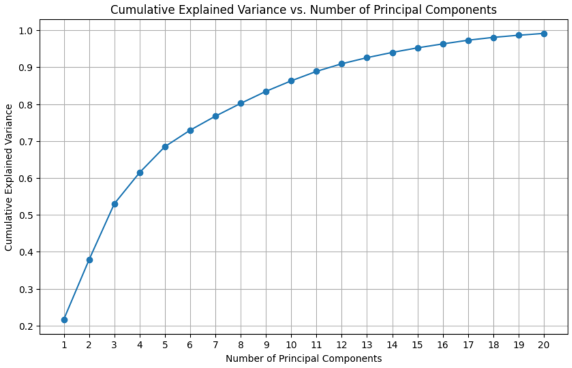
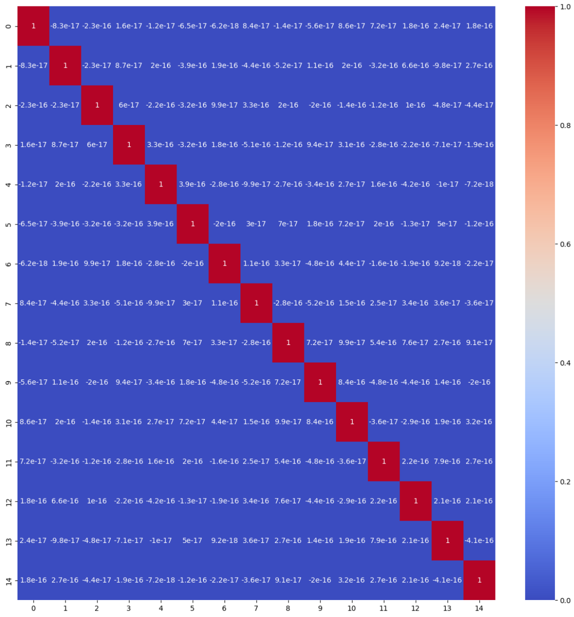


Figure 4.5: Distribution of Target Variable after Default Assignment



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	-8.3e-17	-2.3e-16	1.6e-17	-1.2e-17	-6.5e-17	-6.2e-18	8.4e-17	-1.4e-17	-5.6e-17	8.6e-17	7.2e-17	1.8e-16	2.4e-17	1.8e-16
1	-8.3e-17	1	-2.3e-17	8.7e-17	2e-16	-3.9e-16	1.9e-16	-4.4e-16	-5.2e-17	1.1e-16	2e-16	-3.2e-16	6.6e-16	-9.8e-17	2.7e-16
2	-2.3e-16	-2.3e-17	1	6e-17	-2.2e-16	-3.2e-16	9.9e-17	3.3e-16	2e-16	-2e-16	-1.4e-16	-1.2e-16	1e-16	-4.8e-17	-4.4e-17
3	1.6e-17	8.7e-17	6e-17	1	3.3e-16	-3.2e-16	1.8e-16	-5.1e-16	-1.2e-16	9.4e-17	3.1e-16	-2.8e-16	-2.2e-16	-7.1e-17	-1.9e-16
4	-1.2e-17	2e-16	-2.2e-16	3.3e-16	1	3.9e-16	-2.8e-16	-9.9e-17	-2.7e-16	-3.4e-16	2.7e-17	1.6e-16	-4.2e-16	-1e-17	-7.2e-18
5	-6.5e-17	-3.9e-16	-3.2e-16	-3.2e-16	3.9e-16	1	-2e-16	3e-17	7e-17	1.8e-16	7.2e-17	2e-16	-1.3e-17	5e-17	-1.2e-16
6	-6.2e-18	1.9e-16	9.9e-17	1.8e-16	-2.8e-16	-2e-16	1	1.1e-16	3.3e-17	-4.8e-16	4.4e-17	-1.6e-16	-1.9e-16	9.2e-18	-2.2e-17
7	8.4e-17	-4.4e-16	3.3e-16	-5.1e-16	-9.9e-17	3e-17	1.1e-16	1	-2.8e-16	-5.2e-16	1.5e-16	2.5e-17	3.4e-16	3.6e-17	-3.6e-17
8	-1.4e-17	-5.2e-17	2e-16	-1.2e-16	-2.7e-16	7e-17	3.3e-17	-2.8e-16	1	7.2e-17	9.9e-17	5.4e-16	7.6e-17	2.7e-16	9.1e-17
9	-5.6e-17	1.1e-16	-2e-16	9.4e-17	-3.4e-16	1.8e-16	-4.8e-16	-5.2e-16	7.2e-17	1	8.4e-16	-4.8e-16	-4.4e-16	1.4e-16	-2e-16
10	8.6e-17	2e-16	-1.4e-16	3.1e-16	2.7e-17	7.2e-17	4.4e-17	1.5e-16	9.9e-17	8.4e-16	1	-3.6e-17	-2.9e-16	1.9e-16	3.2e-16
11	7.2e-17	-3.2e-16	-1.2e-16	-2.8e-16	1.6e-16	2e-16	-1.6e-16	2.5e-17	5.4e-16	-4.8e-16	-3.6e-17	1	2.2e-16	7.9e-16	2.7e-16
12	1.8e-16	6.6e-16	1e-16	-2.2e-16	-4.2e-16	-1.3e-17	-1.9e-16	3.4e-16	7.6e-17	-4.4e-16	-2.9e-16	2.2e-16	1	2.1e-16	2.1e-16
13	2.4e-17	-9.8e-17	-4.8e-17	-7.1e-17	-1e-17	5e-17	9.2e-18	3.6e-17	2.7e-16	1.4e-16	1.9e-16	7.9e-16	2.1e-16	1	-4.1e-16
14	1.8e-16	2.7e-16	-4.4e-17	-1.9e-16	-7.2e-18	-1.2e-16	-2.2e-17	-3.6e-17	9.1e-17	-2e-16	3.2e-16	2.7e-16	2.1e-16	-4.1e-16	1

Figure 4.6: Distribution of Target Variable after Default Assignment

4.4 Model Selection

Target Class is a Binary Classification problem where Target Class 0 is not default while Target Class 1 is default.

Model	**HyperParameters**
K-Nearest Neighbors	k = 2,3,4,5,6,7,8,**9**
Naive Bayes	variation smoothing = $\mathbf{1 \times 10^{-9}}$, 1×10^{-8}, 1×10^{-7}, 1×10^{-6}, 1×10^{-5}
SVM	c = **0.1**, 1, 10, 100 Kernel = 'linear', $'rbf'$, 'poly', 'sigmoid' degree = **2**,3 gamma = $'scale'$, 'auto'
Decision Tree	criterion = gini, ***entropy***, max depth = None, **5** 10,15,20 min samples split = 2,5,**10** min samples leaf = 1,**2**,4
Random Forest	n estimators = **50**,100,150 max depth = ***None***,10,20,30 min samples split = 2,5,**10** min samples leaf = 1,2,**4**
XGBoost	n estimators = (100,300) max depth = (3,7) learning rate = (0.01,0.3)
LightGBM	num leaves = (20,40) depth = (3,7) learning rate = (0.01,0.3)
CatBoost	iterations = (100,300) max depth = (3,7) learning rate = (0.01,0.3)
Multi Layer Pecerptron NN	Hidden Layer with ReLU activation, Output Layer - Sigmoid Activation Function, learning rate = (0.001,0.1), L2 Regularization, Early Stopping with epochs = (100,500)
CNN	Conv1D (Filters:32, Kernel Size:3, ReLU), Max Pooling Layer (Pool Size:2), Flatten Layer, Dense Layer (UNits 64, ReLU), Output (Sigmoid), Optimizer (Adam, Batch Size - 5)
LSTM	1st LSTM layer(64 units), 2nd LSTM layer(32 units), Batch Normalization layer, Output Layer, Optimizer - Adam

Table 4.1: HyperParameters

Model	**Accuracy**	**ROC AUC Score**	**Precision-Class 0**	**Recall-Class 0**	**F1 Score-Class 0**	**Precision-Class 1**	**Recall-Class 1**	**F1 Score-Class 1**
Linear Discriminant Analysis	0.7	-	-	-	-	-	-	-
Logistic Regression	0.68	0.625	0.72	0.89	0.79	0.43	0.2	0.27
KNN Classifier	0.72	0.648	0.74	0.91	0.82	0.57	0.27	0.36
Naive Bayes	0.68	0.775	0.74	0.83	0.78	0.45	0.33	0.38
SVM	0.7	0.709	0.7	1.00	0.82	0	0	0
Decision Tree	0.7	0.65	0.76	0.83	0.79	0.5	0.4	0.44
RandomForest	**0.74**	**0.81**	**0.74**	0.97	**0.84**	0.75	**0.2**	**0.32**
XGBoost	0.74	0.7457	0.73	1	0.84	1	0.13	0.24
LightGBM	0.68	0.7467	0.72	0.89	0.79	0.43	0.2	0.27
CatBoost	**0.74**	0.781	0.73	**1**	**0.84**	**1**	0.13	0.24
Multi Layer Perceptron NN	0.7	0.5	0.7	1	0.82	0	0	0
CNN	0.72	0.5848	0.71	1	0.83	1	0.07	0.12
LSTM	0.7	0.3962	0.7	1	0.82	0	0	0

Table 4.2: Model Performance

4.4.1 Forward Selection Ensemble Model

The candidate models used for the ensemble model were the best models from the above models which are Logistic, K Nearest Neighbors, Naive Bayes, Simple Vector Machine, Decision Tress, Random Forest, XGBoost, LightGB and CatBoost.

Using the PCA Training Set, the best ROC-AUC score was found to be **0.7613**. The Ensemble Model selected was Logistic Regression, K-Nearest Neighbors, Naive Bayes, Simple Vector Machine, Decision Tree, Random Forest Classifier and XGBoost.

Model	**Accuracy**	**ROC AUC Score**	**Precision-Class 0**	**Recall-Class 0**	**F1 Score-Class 0**	**Precision-Class 1**	**Recall-Class 1**	**F1 Score-Class 1**
Forward Selection Ensemble Model	0.74	0.739	0.74	0.97	0.84	0.75	0.2	0.32

Table 4.3: Ensemble Model - Forward Selection Performance

4.4.2 Backward Selection Ensemble Model

The candidate models used for the ensemble model were the best models from the above models which are Logistic, K - Nearest Neighbors, Naive Bayes, Simple Vector Machine, Decision Tress, Random Forest, XGBoost, LightGB and CatBoost.

Using the PCA Training Set, the best ROC-AUC score was found to be **0.7536**. The Ensemble Model selected was Logistic Regression, K Nearest Neighbors, Naive Bayes, Simple Vector Machine, Random Forest Classifier, XGBoost and LightGBM.

Model	**Accuracy**	**ROC AUC Score**	**Precision-Class 0**	**Recall-Class 0**	**F1 Score-Class 0**	**Precision-Class 1**	**Recall-Class 1**	**F1 Score-Class 1**
Backward Selection Ensemble Model	0.7	0.7676	0.72	0.97	0.83	0.67	0.13	0.22

Table 4.4: Ensemble Model - Backward Selection Performance

4.4.3 Pearson Correlation Ensemble Model Selection

Attempting to choose models based on correlation based methods similar to one of the research previously cited, Pearson correlation was used for model selection with each model's predictions and its correlation with the actual target values. This provided insights into which models are likely to perform better based on their predictive accuracy. If the correlation is the highest, the best model is chosen and if it matches the highest, the model is added to the list of best models. If the final ensemble contains multiple models, Majority Voting is used to combine the predictions.

The Final Ensemble Model led only to the selection of the XGBoost Classifier.

Model	**Accuracy**	**ROC AUC Score**	**Precision-Class 0**	**Recall-Class 0**	**F1 Score-Class 0**	**Precision-Class 1**	**Recall-Class 1**	**F1 Score-Class 1**
Pearson Correlation Ensemble Model	0.74	0.7257	0.75	0.94	0.84	0.67	0.27	0.238

Table 4.5: Pearson Correlation Ensemble Model

4.4.4 Mutual Information Correlation Ensemble Model Selection

Attempting to choose models based on correlation based methods similar to one of the research previously cited, the alternate correlation method using Mutual

Information was used for model selection with each model's predictions and its correlation with the actual target values. MI is particularly useful when dealing with non-linear relationships. If models exhibit non-linear interactions with the target variable, MI can help identify these relationships better than Pearson correlation. A higher MI indicates that the model's predictions provide substantial information about the actual outcomes, which could suggest better predictive performance.

The Final Ensemble Model led to the selection of K Nearest Neighbors, Naive Bayes, Simple Vector Machine, Decision Tree and Light GBM Classifier.

Model	**Accuracy**	**ROC AUC Score**	**Precision-Class 0**	**Recall-Class 0**	**F1 Score-Class 0**	**Precision-Class 1**	**Recall-Class 1**	**F1 Score-Class 1**
Mutual Information Correlation Ensemble Model	0.68	0.52	0.71	0.91	0.8	0.4	0.13	0.2

Table 4.6: Mutual Information Correlation Ensemble Model

Having tried multiple approaches as well standalone models, the final two models that really stood out were Random Forest and CatBoost as highlighted above. Considering all the values selected, **CatBoost** was selected for it being a gradient boosted model as compared to Random Forest being a bootstrapped model where CatBoost trains trees sequentially learning from the mistakes of the previous ones.

4.5 Model Performance

The final analysis was to identify the sectors which CatBoost incorrectly predicts. This was performed by using the parameters of the best performing CatBoost Model from the previous set of experiments to the initial training set.

Model	**Accuracy**	**Overall Precision**	**Overall Recall**	**Overall F1 Score**	**ROC AUC**
Tuned CatBoost	0.82	0.7143	0.6667	0.6474	0.8267

Table 4.7: Overall Performance

Model	**Precision - Class 0**	**Recall - Class 0**	**F1 Score - Class 0**	**Precision - Class 1**	**Recall - Class 1**	**F1 Score - Class 1**
Tuned CatBoost	0.86	0.89	0.87	0.71	0.67	0.69

Table 4.8: Individual Class Performance

It shows excellent performance in predicting both classes where the train test split in the test set was 35 - 15.

Chapter 5: Inferences

5.1 Sector Wise Analysis

The objective of this project was to identify if there is an one fit all Machine Learning Model that would work for all the sectors chosen and identify the sectors in which this model under performs.

Company	Default	Predicted	Sector
NACCO Industries, Inc	0	1	Energy
Viad Corp	1	0	Commercial Services
Tile Shop Holdings, Inc	0	1	Retail
Monro, Inc	1	0	Retail
Limoneira Company	1	0	Retail
Sonic Automotive, Inc	1	0	Retail
Calavo Growers, Inc	0	1	FMCG
Alico, Inc	0	1	FMCG
SkyWest, Inc	1	0	Transportation

Table 5.1: Misclassified Organizations

CatBoost trained on these parameters performs well in predicting Energy, Commercial Services and Transportation Sectors. An overall accuracy of 0.82, ROC-AUC of 0.8268 and F1 scores of 0.87 and 0.69 for class 0 and 1 was achieved.

5.2 Explainability

5.2.1 SHAP Analysis

Higher absolute SHAP values (further from 0) indicate stronger influence on the prediction. A positive SHAP value means the features pushes the prediction higher while a negative value means it pushes it lower. A wide spread for features such as FCF Yield and Quick Ratio have highest impacts where higher the feature value the prediction is negatively impacted while PB Ratio and Market Capitalization have less impact due to their concentrated distribution around 0 as shown in Figure 5.1

Return on Invested Capital and Current Ratio shows clusters of points on both sides of the SHAP value axis suggesting interaction effects or varying impacts depending on the feature's value in combination with other factors.

5.2.2 Permutation Importance

The permutation importance analysis provides insights into the relative importance of various financial metrics in a predictive model. Liquidity and profitability metrics generally have higher importance than valuation ratios. Cash flow-related metrics (FCF Yield, EV/FCF Ratio) are more important than traditional earnings-based metrics (PE Ratio). The model appears to place more emphasis on short-term financial health (Quick Ratio) and shareholder returns (Payout Ratio, Earnings Yield) than on long-term debt metrics or company size.

5.2.3 Partial Dependency Plots

- Market Capitalization shows a positive trend, indicating an increase in the target variable.

- Enterprise Value displays a negative trend, suggesting a decrease in the target variable.

- A lower PE Ratio correlates with a higher probability of default, as shown by the steep increase in default probability when the PE Ratio decreases.

- The default probability increases significantly as the EV/EBITDA ratio decreases below zero, indicating higher risk with lower ratios.

- As the EV/Sales ratio moves from negative to positive, the default probability rises sharply, demonstrating increased risk with higher ratios.

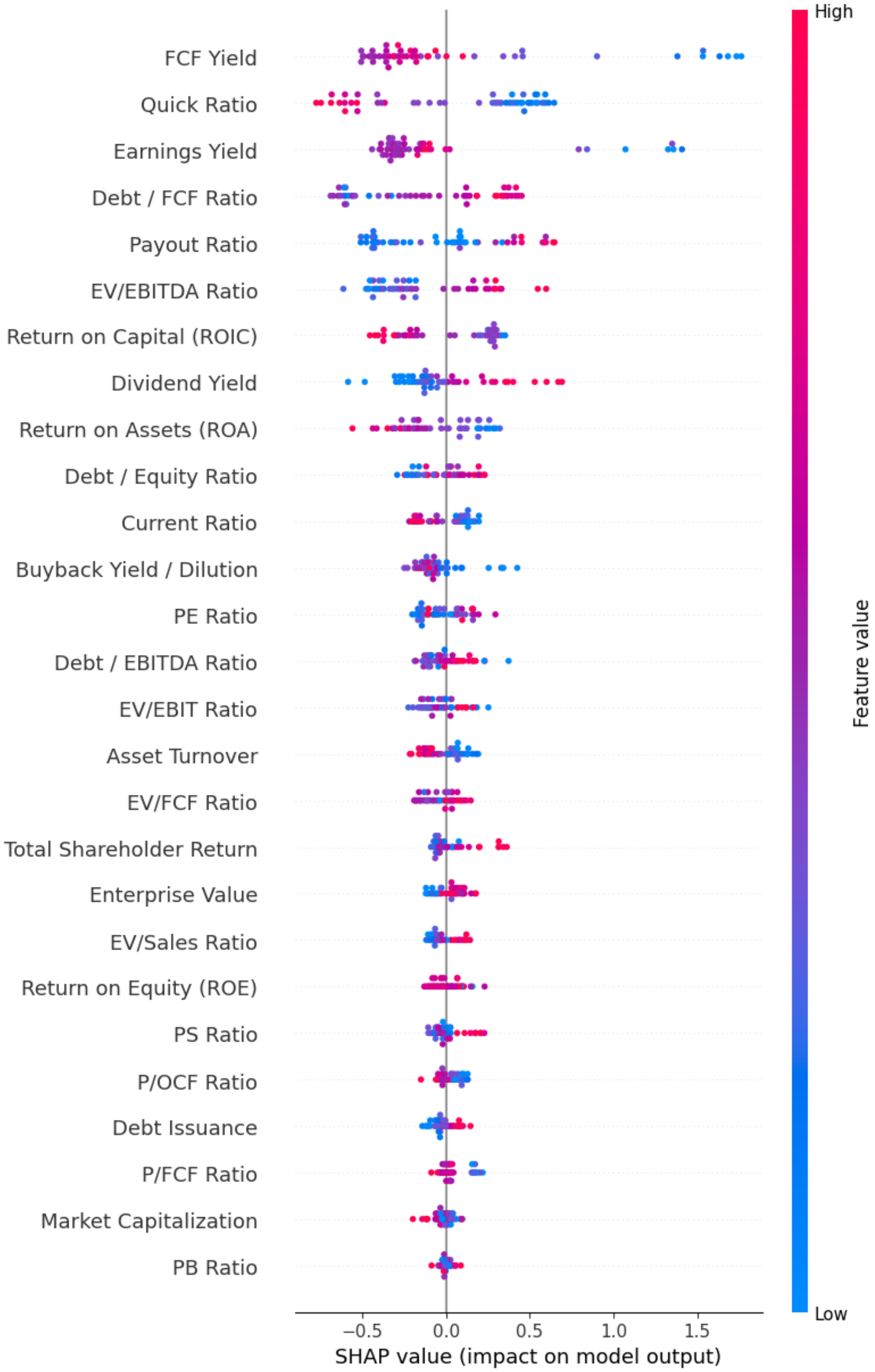


Figure 5.1: SHAP Analysis

- There is a sharp decrease in default probability around a threshold (0.05). Beyond this point, further increases in buyback yield do not significantly affect the probability.

- Higher debt ratios lead to an increased default probability after a certain threshold, reflecting greater financial risk with higher leverage.

- Metrics like Current Ratio, Asset Turnover, Return on Equity, Return on Assets, Return on Capital, Earnings Yield, Dividend Yield, Payout Ratio, and others show flat lines, indicating little to no effect on the target variable as shown in Figure 5.2

5.2.4 Feature Importance

- **Cash Flow and Earnings:** FCF Yield and Earnings Yield are the two most important features indicating the model heavily considers these cash generating factors.

- **Financial Health Indicators:** Quick Ratio and Debt/FCF Ratio highlights the importance of short term liquidity and debt levels relative to cash flow.

- **Valuation Metrics:** EV/EBITDA Ratio is ranked the highest compared to traditional metrics such PE Ratio, PS Ration and PB Ratio.

- **Profitability and Efficiency:** ROA and ROIC are the highest ranked profitability measures.

- **Shareholder Returns:** Dividend Yield and Buyback Yield/ Dilution is being considered by the model.

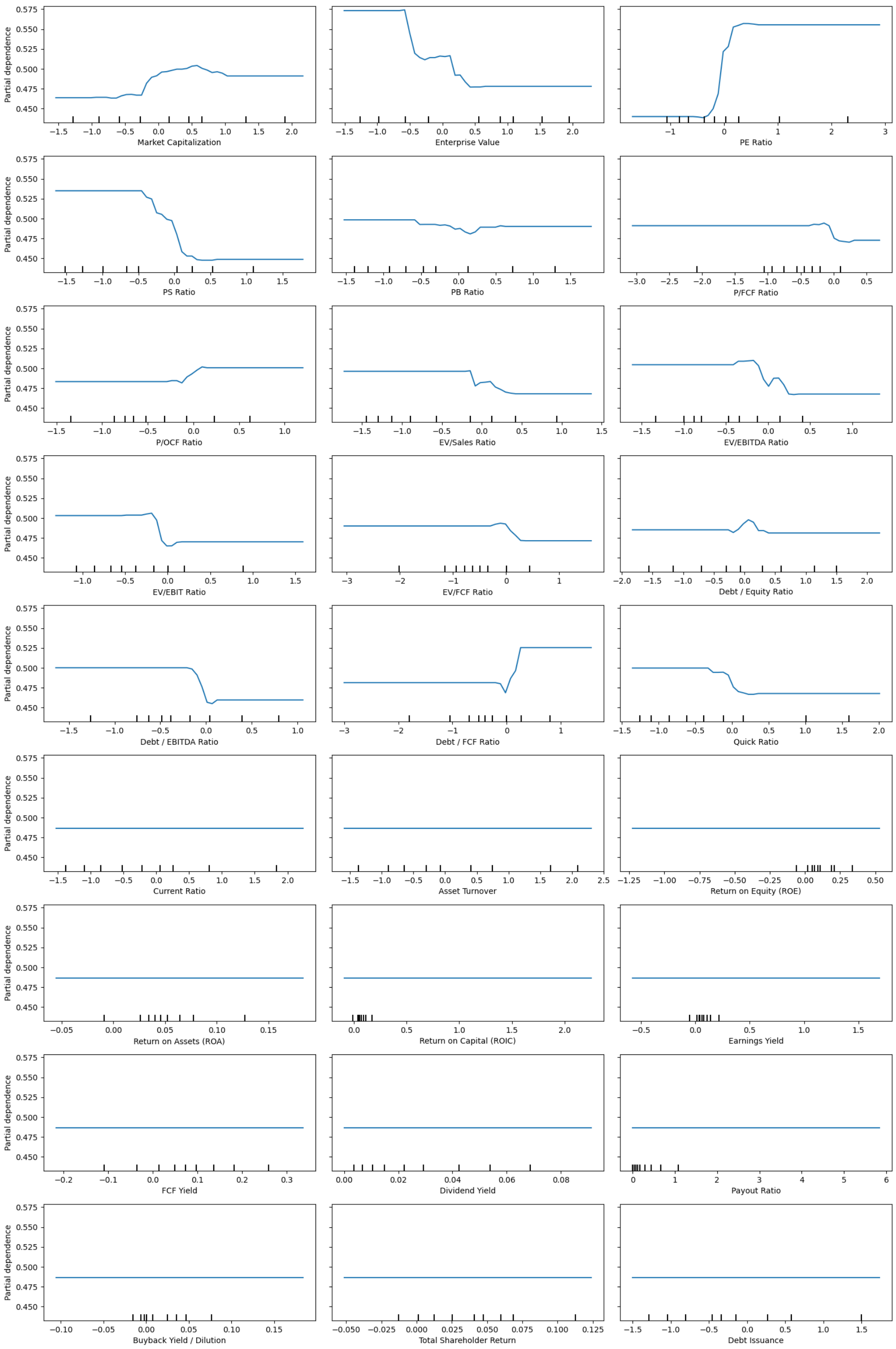


Figure 5.2: Partial Dependency Plots

Chapter 6: Conclusion

An undertaking with the objective of achieving an Early Warning Credit Risk Default prediction for 50 Small and Medium Size Businesses listed in the USA was achieved. 26 Financial Metrics belonging to Cashflow, Financial Health, Valuation, Profitability and Shareholder Returns were scraped from a reliable source, benchmarks were selected to assign the Default class to the 2013 - 2023 time series data. Ranking the indicators using entropy and Grey correlation analysis, the weighted financial indicators were then transformed to Principal Components. Testing a variety of industry standard Machine Learning Models, implementing Correlation based model subset ensemble models, CatBoost was found to perform the best in various metrics such ROC-AUC and F1 Score. The model selected was then trained on the entire training set of Financial Metrics leading to an accuracy of 0.82 and ROC-AUC of 0.83 on the smaller test set. Interpreting the model results for each individual class, sector and financial metric led to informative conclusions.

6.1 Future Scope

Selection of 10 organizations for each sector was an arduous task however if the limitation of publicly available data and sectors can be overcome, more data collection and analysis would lead to high impact results. Some regulations such Basel And IFRS 9 might have to be taken into account to make the results more industry relevant. Macroeconomic Factors such as GDP, Unemployment Rate, Interest Rate, Inflation Rate and Currency Exchange Rates need to be included and it would be interesting to combine a mathematical optimization model with an ML model output. There might not be a one fit all Machine Learning model for all sectors as identified with this small dataset.

Works Cited

Fabio Antonelli, Alessandro Ramponi, and Sergio Scarlatti. Cva and vulnerable options pricing by correlation expansions. *Annals of Operations Research*, 299 (1):401–427, 2021.

Gurdip Bakshi, Dilip Madan, and Frank Xiaoling Zhang. Investigating the role of systematic and firm-specific factors in default risk: Lessons from empirically evaluating credit risk models. *The Journal of Business*, 79(4):1955–1987, 2006.

Mathieu Boudreault, Geneviève Gauthier, and Tommy Thomassin. Contagion effect on bond portfolio risk measures in a hybrid credit risk model. *Finance Research Letters*, 11(2):131–139, 2014.

Yujia Chen, Raffaella Calabrese, and Belen Martin-Barragan. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 312(1):357–372, 2024. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2023.06.036. URL `https://www.sciencedirect.com/science/article/pii/S0377221723005088`.

Farzad Alavi Fard. Analytical pricing of vulnerable options under a generalized jump–diffusion model. *Insurance: Mathematics and Economics*, 60:19–28, 2015.

Jia-Wen Gu, Wai-Ki Ching, Tak-Kuen Siu, and Harry Zheng. On reduced-form intensity-based model with 'trigger'events. *Journal of the Operational Research Society*, 65(3):331–339, 2014.

John Hull and Alan White. The impact of default risk on the prices of options and other derivative securities. *Journal of Banking & Finance*, 19(2):299–322, 1995.

Gechun Liang and Xingchun Wang. Pricing vulnerable options in a hybrid credit risk model driven by heston–nandi garch processes. *Review of Derivatives Research*, 24(1):1–30, Apr 2021. ISSN 1573-7144. doi: 10.1007/s11147-020-09167-z.

Jomark Pablo Noriega, Luis Antonio Rivera, and José Alfredo Herrera. Machine learning for credit risk prediction: A systematic literature review. *Data*, 8(11), 2023. ISSN 2306-5729. doi: 10.3390/data8110169. URL `https://www.mdpi.com/2306-5729/8/11/169`.

Lu Wang and Wenyao Zhang. A qualitatively analyzable two-stage ensemble model based on machine learning for credit risk early warning: Evidence from chinese manufacturing companies. *Information Processing & Management*, 60(3):103267, 2023. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2023.103267. URL `https://www.sciencedirect.com/science/article/pii/S0306457323000043`.

Xingchun Wang. Analytical valuation of vulnerable options in a discrete-time framework. *Probability in the Engineering and Informational Sciences*, 31(1): 100–120, 2017.

Zhibin Xiong and Jun Huang. Prediction of credit risk with an ensemble model: a correlation-based classifier selection approach. *Journal of Modelling in Management*, 17(4):1078–1097, Jan 2022. doi: 10.1108/JM2-09-2020-0235.