A Project Synopsis on

Bankruptcy Detection

Done by

Anushka Churi

ABSTRACT

This project develops a machine learning framework to predict corporate bankruptcy using financial ratios. The dataset comprises 6,819 compnies with 96 features capturing profitability, liquidity, leverage, and operational metrics. After standardizing the data and addressing severe class imbalance with SMOTE, three models were evaluated: Logistic Regression, Random Forest, and XGBoost. Performance was assessed using Accuracy, Precision, Recall, F1-Score, and AUC-ROC, with Random Forest and XG- Boost demonstrating superior predictive power, particularly in identifying distressed firms. Feature importance analysis revealed liquidity and leverage ratios as the most critical indicators of financial instability. This aproach provides a data-driven method for banks, investors, and regulators to proactively manage credit risk and make informed financial decisions.

Contents

List of Figures iii

List of Tables iv

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Problem Statement . . . 1
- 1.3 Objectives . . . 1

2 Model Description 3
- 2.1 Logistic Regression (baseline model) . . . 3
  - 2.1.1 What Logistic Regression Does . . . 3
  - 2.1.2 Why Logistic Regression is Used . . . 3
  - 2.1.3 Limitations . . . 4
- 2.2 Random Forest (ensemble model) . . . 4
  - 2.2.1 What Random Forest Does . . . 4
  - 2.2.2 Advantages in Bankruptcy Prediction . . . 4
  - 2.2.3 Why It's Better than Logistic Regression . . . 4
- 2.3 XGBoost (gradient boosting model) . . . 5
  - 2.3.1 What XGBoost Does . . . 5
  - 2.3.2 Advantages for Bankruptcy Prediction . . . 5
  - 2.3.3 Why XGBoost is Often the Best Choice . . . 5

3 Business Impact 6
- 3.1 For Banks and Financial Institutions . . . 6
- 3.2 For Investors and Asset Managers . . . 6
- 3.3 For Corporates and Regulators . . . 6
- 3.4 Overall Strategic Value . . . 6

4 Observation 8
- 4.1 Logistic Regression . . . 8
- 4.2 Random Forest (Ensemble Model) . . . 8
- 4.3 XGBoost (Gradient Boosting Model) . . . 9
- 4.4 General Observations Across Models . . . 10

5 Conclusion 11

List of Figures

2.1 Logistic Regression Model Training..... 3
2.2 Random Forest Model Training. 4
2.3 Xgboost Model Training. 5

4.1 Logistic Regression Model Training Performance. 8
4.2 Random Forest Model Training Performance. 9
4.3 Random Forest Model Feature Training Performance. 9
4.4 Random Forest Model Bar Graph. 9
4.5 XgBoostForest Model Training Performance. 10
4.6 XgBoost Model Feature Training Performance. 10
4.7 Xgboost Model bar graph. 10

Chapter 1

Introduction

This project predicts corporate bankruptcy using machine learning techniques. The data set consists of 6,819 companies with 96 financial ratios, and the target variable indicates whether a company is bankrupt (1) or healthy (0). The preprocessing steps included standardization of the features and balancing of the classes using SMOTE.

1.1 Motivation

Corporate bankruptcies can cause significant losses for investors, banks, and other stake holders. Traditional financial analysis methods often fail to detect early warning signs of distress due to their reliance on static ratios and manual assessments. With the rise of machine learning, it is now possible to analyze complex patterns in financial data more effectively. This project is motivated by the need to develop a predictive model that not only identifies financially distressed firms with greater precision, but also provides insights into the key financial indicators driving the risk of bankruptcy. Using data science techniques, the project aims to support better credit risk management and strategic decision making.

1.2 Problem Statement

Financial institutions and stakeholders face challenges in identifying companies at risk of bankruptcy in a timely and reliable manner. Existing methods are either too simplistic or fail to generalize across industries. There is a need for a robust predictive model that:

1. Handles imbalanced data (since bankrupt firms are rare).
2. Accurately identify distressed companies while minimizing false negatives.
3. Provides insights into key financial indicators that drive bankruptcy.

1.3 Objectives

1. Preprocess financial data by standardizing ratios and addressing class imbalance using SMOTE.
2. Develop and compare machine learning models (Logistic Regression, Random Forest, XGBoost) for bankruptcy prediction.

3. Evaluate the performance of the model using metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

4. Identify the most influential financial ratios through analysis of significance of characteristics.

5. Provide actionable insights to support data-driven credit risk management and early intervention strategies.

Chapter 2

Model Description

2.1 Logistic Regression (baseline model)

Logistic regression is a fundamental supervised learning algorithm, widely used as a base line model for classification tasks in machine learning. It is a simple, interpretable, and efficient model, making it a good starting point before exploring more complex algorithms. Despite its name, it predicts the probability of a categorical outcome (e.g. yes/no, true/false) rather than a continuous one.

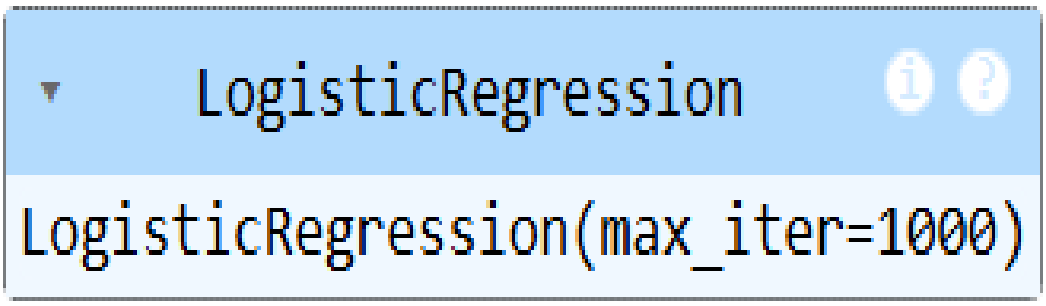

Figure 2.1: Logistic Regression Model Training.

2.1.1 What Logistic Regression Does

1. It predicts the probability of bankruptcy (1) vs. financially healthy (0).
2. Uses a sigmoid function to map financial ratios to probabilities between 0 and 1.
3. The coefficients represent how each feature affects the odds of bankruptcy.

2.1.2 Why Logistic Regression is Used

1. Baseline model to compare with more complex algorithms.
2. Interpretable results, important for financial regulators and managers.
3. Helps identify which financial ratios matter the most.

2.1.3 Limitations

1. Assumes a linear relationship between ratios and bankruptcy risk.
2. Might perform less than expected compared to tree-based models like Random Forest or XGBoost, which capture non-linear interactions.

2.2 Random Forest (ensemble model)

Random Forest is a set of machine learning models that utilize multiple decision trees to improve prediction accuracy and mitigate overfitting. It can be used for classification and regression tasks.

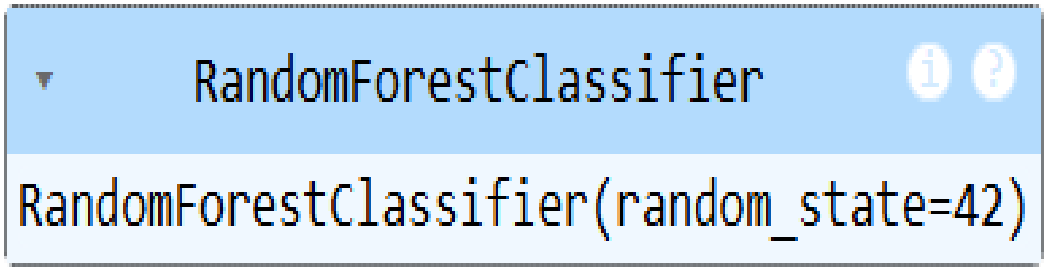

Figure 2.2: Random Forest Model Training.

2.2.1 What Random Forest Does

1. It is an ensemble model that builds multiple decision trees and averages their predictions.
2. Each tree learns from a random subset of features and data, reducing overfitting.
3. Great for tabular financial data because it captures non-linear relationships and interactions between ratios (something Logistic Regression cannot do well).

2.2.2 Advantages in Bankruptcy Prediction

1. Better manages class imbalance (still works best with SMOTE).
2. Captures complex patterns such as how high debt is risky only when profitability is also low.
3. Provides importance of features, helping to identify which financial ratios drive predictions.
4. Less sensitive to scaling, so minor preprocessing errors matter.

2.2.3 Why It's Better than Logistic Regression

1. Captures non-linear effects (e.g., bankruptcy risk increases sharply only after certain thresholds).
2. More robust to noise in financial data.

3. Usually higher Recall and AUC, making it more effective in detecting bankrupt companies.

2.3 XGBoost (gradient boosting model)

XGBoost, or eXtreme Gradient Boosting, is an open source distributed machine learning library that implements gradient-boosted decision trees. It is known for its speed, performance, and scalability, making it a popular choice for various machine learning tasks, especially with structured or tabular data.

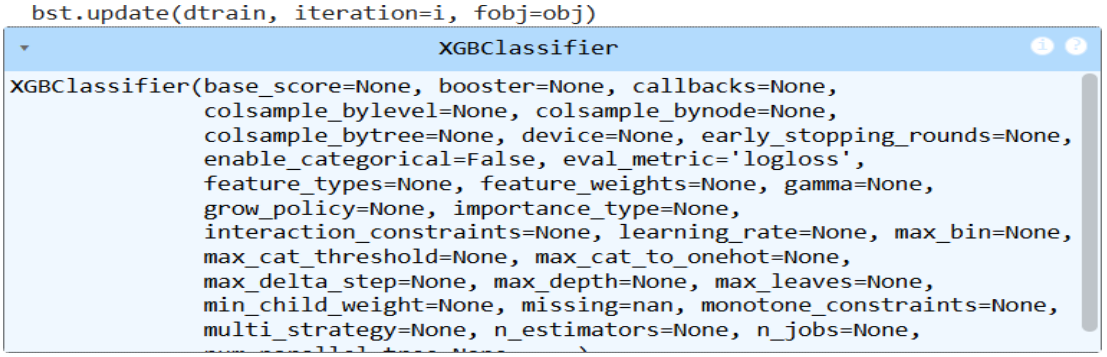

Figure 2.3: Xgboost Model Training.

2.3.1 What XGBoost Does

1. Gradient Boosting Algorithm: Build trees sequentially, where each new tree corrects the errors of the previous ones.
2. Optimized for speed and performance: Faster and more accurate than basic Gradient Boosting.
3. Handles non-linear relationships and complex interactions between financial ratios better than Logistic Regression and often better than Random Forest.

2.3.2 Advantages for Bankruptcy Prediction

1. High predictive power: Often achieves the best Recall and AUC, crucial for detecting bankrupt firms.
2. Handles class imbalance with parameters like scale pos weight (but SMOTE already helps).
3. The importance of the feature shows which financial ratios are most influential.
4. Built-in regularization (reduces overfitting compared to Random Forest).

2.3.3 Why XGBoost is Often the Best Choice

1. Captures complex patterns that simpler models miss.
2. The best AUC and Recall → are usually to catch bankrupt companies even at the cost of a few false positives.
3. Faster training and tuning compared to Random Forest on large datasets.

Chapter 3

Business Impact

3.1 For Banks and Financial Institutions

The model improves credit decision making by enabling improved credit risk management, allowing banks to identify high-risk firms before granting loans, and thus reducing non performing assets (NPAs). It also supports proactive monitoring by detecting early signs of financial distress, giving institutions the opportunity to intervene through measures such as debt restructuring. In addition, it enables better capital allocation by directing lending toward financially stable firms, which strengthens overall portfolio quality and reduces exposure to potential defaults.

3.2 For Investors and Asset Managers

The model empowers investors to make informed investment decisions by identifying companies with a high risk of bankruptcy, thereby protecting portfolios from sudden losses. Enhances valuation analysis by complementing traditional fundamental assessments with predictive modeling, offering a clearer view of a company's long-term financial health. In addition, it supports risk-adjusted returns by enabling better portfolio diversification through the exclusion or hedging of high-risk firms, ultimately leading to more stable and resilient investment strategies.

3.3 For Corporates and Regulators

The model serves as an early warning system for companies, enabling them to self-assess their financial health and implement corrective measures such as improving liquidity before reaching critical distress. For regulators, it strengthens oversight by providing tools to monitor systemic risks and prevent large-scale corporate failures. In addition, it aids policy planning by identifying trends in corporate distress, supporting the development of financial stability frameworks that improve the resilience of the broader economy.

3.4 Overall Strategic Value

The model promotes data-driven decision making by replacing subjective judgment with evidence based predictions, allowing stakeholders to act with greater confidence. It supports cost reduction by minimizing financial losses arising from defaults, bankruptcies, and poor investment choices. In addition, it contributes to overall market stability by strengthening investor confidence and fostering healthier and more resilient capital markets.

Chapter 4

Observation

4.1 Logistic Regression

Logistic Regression performed reasonably well as a starting point, but was limited in its ability to capture the complex, nonlinear relationships present in financial data. Its main strength lies in providing interpretable coefficients that highlight how different financial ratios influence risk bankruptcy. However, it achieved a lower recall compared to ensemble models, which means it failed to identify a larger proportion of bankrupt firms, a critical drawback for practical applications where missing distressed companies can lead to significant financial losses.

```
Logistic Regression Performance:
Accuracy: 0.8922287390029325
Precision: 0.2023121387283237
Recall: 0.7954545454545454
F1 Score: 0.3225806451612903
AUC-ROC: 0.9147727272727273
Confusion Matrix:
 [[1182  138]
 [   9   35]]
```

Figure 4.1: Logistic Regression Model Training Performance.

4.2 Random Forest (Ensemble Model)

Random Forest improved predictive performance by capturing nonlinear relationships between financial ratios, offering a more sophisticated understanding of patterns leading to bankruptcy. It achieved higher recall and AUC-ROC compared to Logistic Regression, making it more effective and reliable for detecting financially distressed firms. Feature importance analysis revealed that liquidity and leverage ratios were the dominant predictors driving model decisions. Although the model showed slight tendencies toward overfitting, this issue can be effectively managed through parameter tuning.

```
Random Forest Performance:
Accuracy: 0.9618768328445748
Precision: 0.43548387096774194
Recall: 0.6136363636363636
F1 Score: 0.5094339622641509
AUC-ROC: 0.9373880853994491
Confusion Matrix:
 [[1285   35]
 [  17   27]]
```

Figure 4.2: Random Forest Model Training Performance.

	Feature	Importance
39	Borrowing dependency	0.083886
90	Liability to Equity	0.051615
85	Net Income to Total Assets	0.050565
36	Debt ratio %	0.042066
18	Persistent EPS in the Last Four Seasons	0.041178
9	Continuous interest rate (after tax)	0.040066
7	After-tax net Interest Rate	0.038034
89	Net Income to Stockholder's Equity	0.033305
37	Net worth/Assets	0.032316
67	Retained Earnings to Total Assets	0.028188

Figure 4.3: Random Forest Model Feature Training Performance.

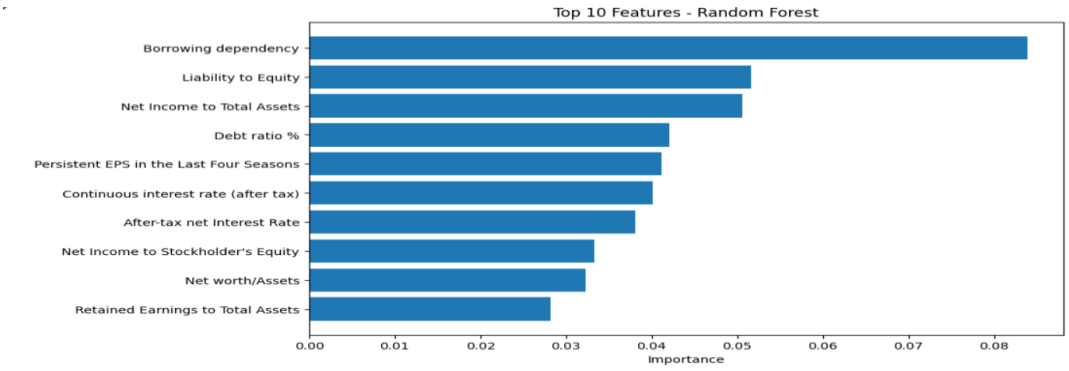

Figure 4.4: Random Forest Model Bar Graph.

4.3 XGBoost (Gradient Boosting Model)

XGBoost delivered the best overall performance, particularly in recall and AUC-ROC, which are crucial to minimizing false negatives in bankruptcy prediction. It effectively captured complex interactions among financial ratios and provided clear feature importance rankings, similar to Random Forest but with higher precision. By balancing predictive power and generalization, XGBoost emerged as the most suitable model for deployment in a real-world corporate risk assessment system.

```
XGBoost Performance:
Accuracy: 0.966275659824047
Precision: 0.4791666666666667
Recall: 0.5227272727272727
F1 Score: 0.5
AUC-ROC: 0.9491219008264463
Confusion Matrix:
 [[1295   25]
 [  21   23]]
```

Figure 4.5: XgBoostForest Model Training Performance.

	Feature	Importance
39	Borrowing dependency	0.233128
18	Persistent EPS in the Last Four Seasons	0.080226
9	Continuous interest rate (after tax)	0.068055
85	Net Income to Total Assets	0.042599
8	Non-industry income and expenditure/revenue	0.028331
0	ROA(C) before interest and depreciation befor...	0.024962
20	Revenue Per Share (Yuan ¥)	0.022371
91	Degree of Financial Leverage (DFL)	0.020703
35	Total debt/Total net worth	0.019371
45	Accounts Receivable Turnover	0.015881

Figure 4.6: XgBoost Model Feature Training Performance.

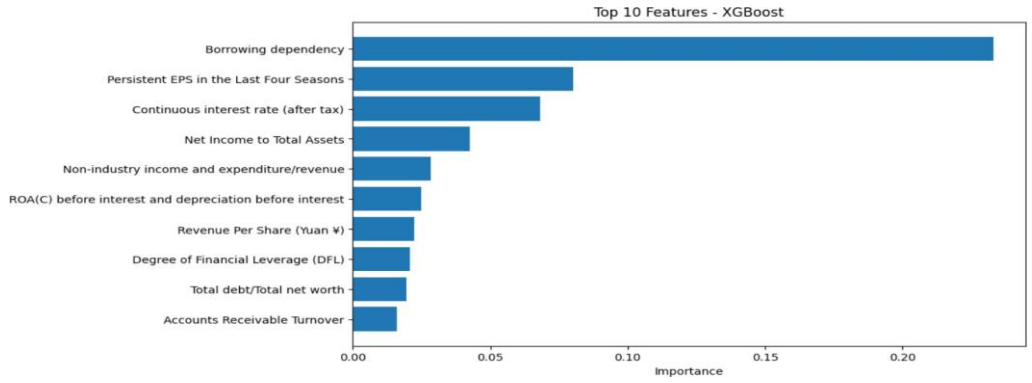

Figure 4.7: Xgboost Model bar graph.

4.4 General Observations Across Models

The application of SMOTE significantly improved model performance by addressing class imbalance, leading to notable gains in recall and enhancing the detection of bankrupt firms. In all models, liquidity, leverage, and profitability ratios consistently emerged as the most influential drivers of bankruptcy prediction, underscoring their critical role in the assessment of financial health. Furthermore, ensemble methods outperformed the linear baseline, confirming that financial distress is best explained by complex, nonlinear patterns and interactions rather than simple linear relationships.

Chapter 5

Conclusion

This project demonstrates how machine learning can significantly enhance the prediction of corporate bankruptcy using financial ratios. By standardizing data, addressing class imbalance with SMOTE, and comparing Logistic Regression, Random Forest, and XGBoost models, the analysis shows that ensemble methods, particularly XGBoost, provide superior accuracy and recall for identifying financially distressed firms. Feature importance analysis highlights liquidity, leverage, and profitability ratios as key indicators of corporate failure. The findings underscore the value of data-driven approaches in improving credit risk management, guiding investment strategies, and supporting regulatory oversight. Future work can integrate macroeconomic variables, real-time financial data, and model deployment tools to further improve predictive performance and usability in practical settings.