Suk Won Choi

PREDICTING CORPORATE BANKRUPTCY USING FINANCIAL RATIOS

A PREDICTIVE APPROACH FOCUSED ON MINIMIZING FALSE NEGATIVES IN BANKRUPTCY FORECASTING

Minimizing Type II errors is crucial for effective risk management, as it significantly reduces the chances of overlooking potential bankruptcies that could lead to substantial financial losses.

Conclusion and Insights

i. Bankruptcy is an extreme event, so feature outliers play a critical role in prediction.

ii. Minimizing false negatives in bankruptcy prediction is vital to avoiding significant financial losses.

iii. Random Forest-based feature selection consistently improves model performance, especially for Logistic Regression.

iv. The Balanced Random Forest model shows strong recall and F-beta scores, making it highly effective for identifying bankruptcies.

MODEL UTILIZATION STRATEGIES

How to Use the Modeling Results

i. The Balanced Random Forest and logistic regression model's high recall can help banks avoid loans to high-risk companies, protecting stability and reducing potential losses

ii. Investors can use the model to identify and exclude high-risk companies from their portfolios, improving investment safety and performance.

iii. This predictive model aids financial institutions in accurately reporting bankruptcy risk, ensuring regulatory compliance.

iv. These systems can update data quarterly based on financial statements, allowing for more timely and informed decision-making.

I. Introduction

1. Problem Statement

Assessing the bankruptcy risk of potential loan or investment targets is extremely important for lenders such as banks and investors in credit. The ability to accurately predict bankruptcy is not only essential for risk management but also crucial for making informed financial decisions. However, predicting bankruptcy is a complex task due to the multifaceted nature of financial data and the intricate interactions between various financial indicators.

The challenge lies in developing a robust and reliable model that can accurately predict the likelihood of corporate bankruptcy using imbalanced datasets, where the number of bankrupt companies is significantly lower than non-bankrupt ones, leading to biased predictions.

Moreover, there is significant asymmetry in the costs associated with prediction errors, specifically in false negative and false positive cases. It is crucial to find a way to incorporate this cost asymmetry into the model.

2. Goal

The primary goal of this project is to develop a robust and accurate model for predicting corporate bankruptcy using financial data. To achieve this, the following specific objectives have been identified:

- Feature Selection and Engineering
- Selection of evaluation metrics considering the costs associated with errors
- Model Development and Evaluation
- Performance Assessment

By accomplishing these goals, the project aims to provide a reliable tool for predicting corporate bankruptcy that not only performs well in terms of accuracy but also considers the financial and operational implications of prediction errors.

II. Dataset

This dataset comes from the UCI Machine Learning Repository[1]. The original data source is the Taiwan Economic Journal, covering data from 1999 to 2009. Corporate bankruptcy is defined based on the business regulations of the Taiwan Stock Exchange.

This dataset's target variable is bankruptcy which indicates whether a company is bankrupt (1) or not (0). The dataset consists of 95 features, which include one categorical variable and the rest are numerical financial ratios, which are normalized.

The financial ratios can be categorized into six groups, each with its own significance:

- **Profitability Ratios:** These ratios assess a company's ability to generate profits relative to its revenue, assets, or equity. High profitability ratios indicate efficient operations and strong financial health. In the context of bankruptcy, low profitability ratios could signal that a company is struggling to maintain its profitability, which may lead to financial distress and increase the risk of bankruptcy.

[1] Source: https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction

- **Liquidity Ratios:** Liquidity ratios measure a company's ability to meet its short-term obligations using its current assets. A high liquidity ratio suggests that a company has sufficient assets to cover its short-term debts. Companies with low liquidity ratios might face difficulties in paying their immediate liabilities, which could be a precursor to bankruptcy if the situation persists.
- **Solvency Ratios:** Solvency ratios evaluate a company's ability to meet its long-term debt obligations. These ratios typically compare a company's debt levels to its assets, equity, or earnings. A company with high solvency ratios is generally considered financially stable. Conversely, low solvency ratios may indicate that a company is over-leveraged and could struggle to meet its long-term obligations, increasing the risk of bankruptcy.
- **Activity Ratios:** Activity ratios analyze how effectively a company utilizes its assets and liabilities to generate revenue and profit. High activity ratios suggest that the company is efficiently managing its resources, leading to improved financial performance. Conversely, poor activity ratios indicate operational inefficiencies, which can adversely affect profitability and liquidity, ultimately increasing the risk of bankruptcy.
- **Growth Ratios:** Growth ratios measure how well a company is expanding its assets, revenue, or profitability over time. Positive growth ratios indicate that a company is growing, which is typically a sign of financial health. On the other hand, negative or stagnant growth ratios might suggest that a company is not expanding or is even contracting, which could lead to financial instability and raise the risk of bankruptcy.
- **Other Financial Metrics:** This category includes various other financial ratios that don't neatly fit into the above categories but are still crucial for analyzing a company's financial performance. Depending on their values, they can provide additional insights into a company's financial condition and potential risk of bankruptcy. For instance, a low interest coverage ratio could indicate that a company might struggle to cover its interest expenses, which could be a warning sign of financial distress.

III. Exploratory Data Analysis

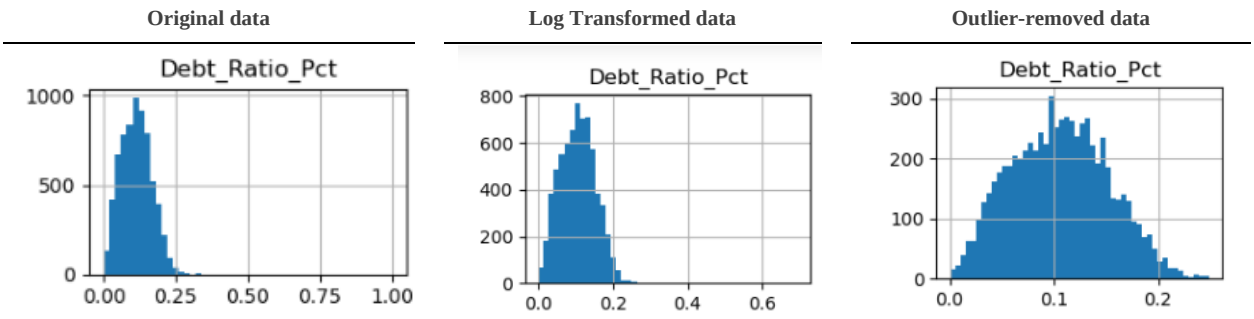
1. Univariate Analysis

1.1 Target Variable

The target variable is highly imbalanced with 96.9% of 0 and 3.1% of 1, which can lead to various issues in model performance and evaluation. Such imbalance might cause the model to be biased towards the majority class, resulting in poor prediction accuracy for the minority class. In a later section, we will address this issue by implementing data preprocessing techniques specifically designed to handle imbalanced datasets. These techniques will help ensure that our model performs well across all classes.

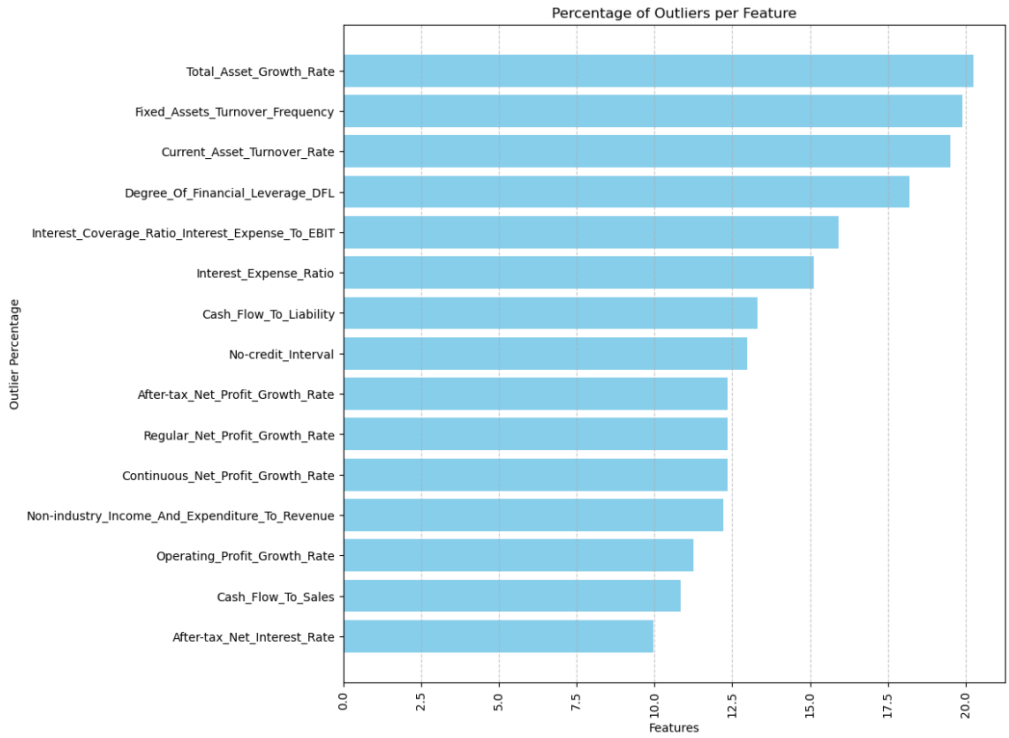
1.2 Numerical Features

Two noteworthy observations were made regarding the distribution of the numerical features. First, most variables deviated from a normal distribution, exhibiting asymmetrical distributions. Although log transformation was attempted to address this, it had minimal effect. However, removing outliers significantly reduced the asymmetry in the distributions. <Figure 1> illustrates an example of such a case.



<Figure 1.Distribution Example: Debt Ratio >

The second noteworthy observation is the presence of a large number of outliers. Using the interquartile range to identify them, it was found that 15 features had outliers comprising more than 10% of the data. (Figure 2)



<Figure2. Proportion of Outliers Per Feature>

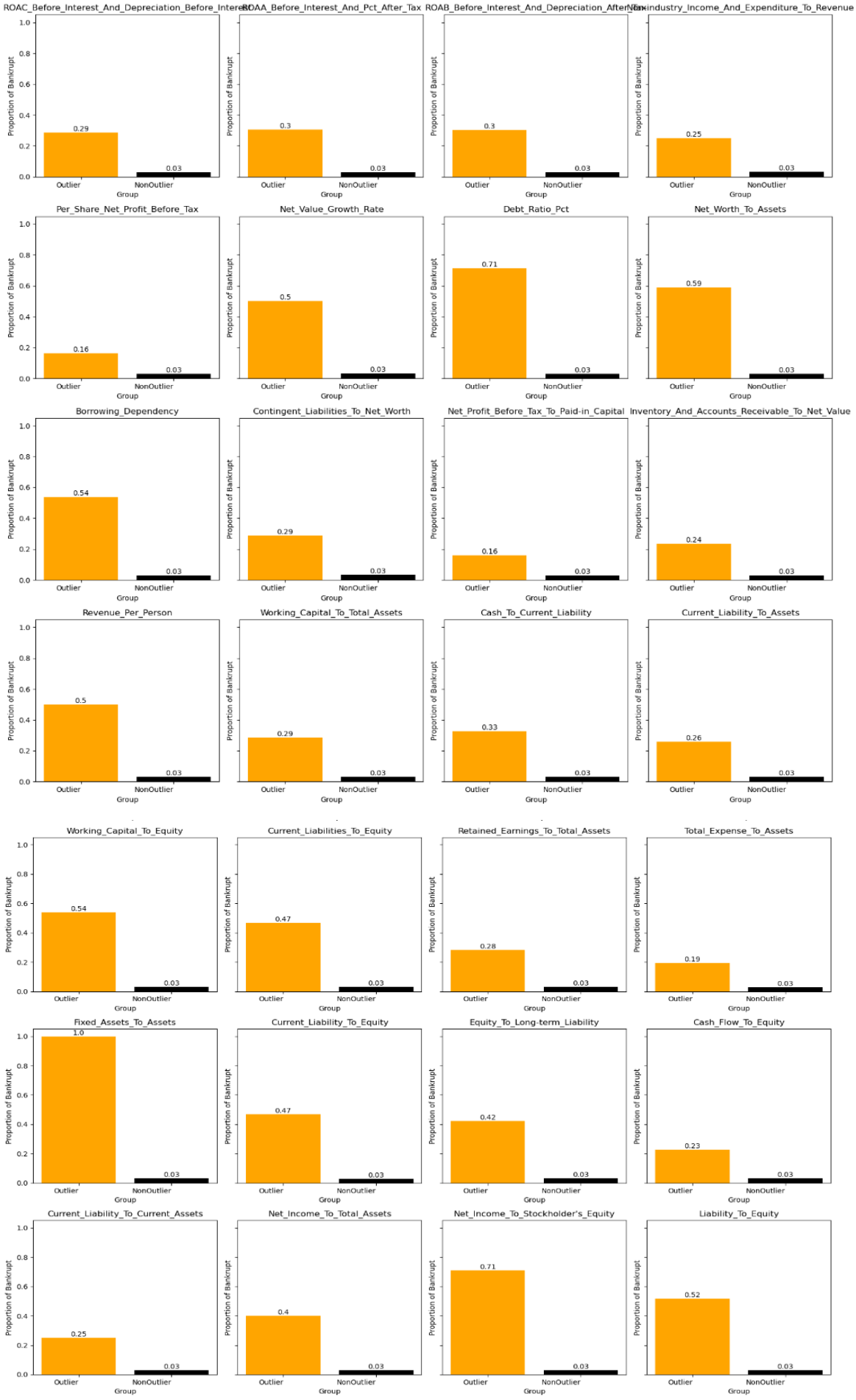
Summarizing these two findings, while removing outliers improves the distribution of the features, the large number of outliers suggests that doing so could result in significant information loss. To address this, we examined how effectively the presence of outliers could predict the target variable in the following section.

2. Bivariate Analysis

2.1 Feature Outliers and Bankruptcy Rate

By dividing each numerical feature into outlier and non-outlier groups and measuring the bankruptcy rate, we found that, for many features, the bankruptcy rate among non-outliers was similar to the overall rate but significantly higher among outliers. <Figure 3> displays graphs for features where the bankruptcy rate in the outlier group is at least 5 times higher than that in the non-outlier group. A total of 28 features fall into this category.

This indicates that outliers contain important information for predicting the target variable. Therefore, the outliers will be retained in the dataset without any further processing.

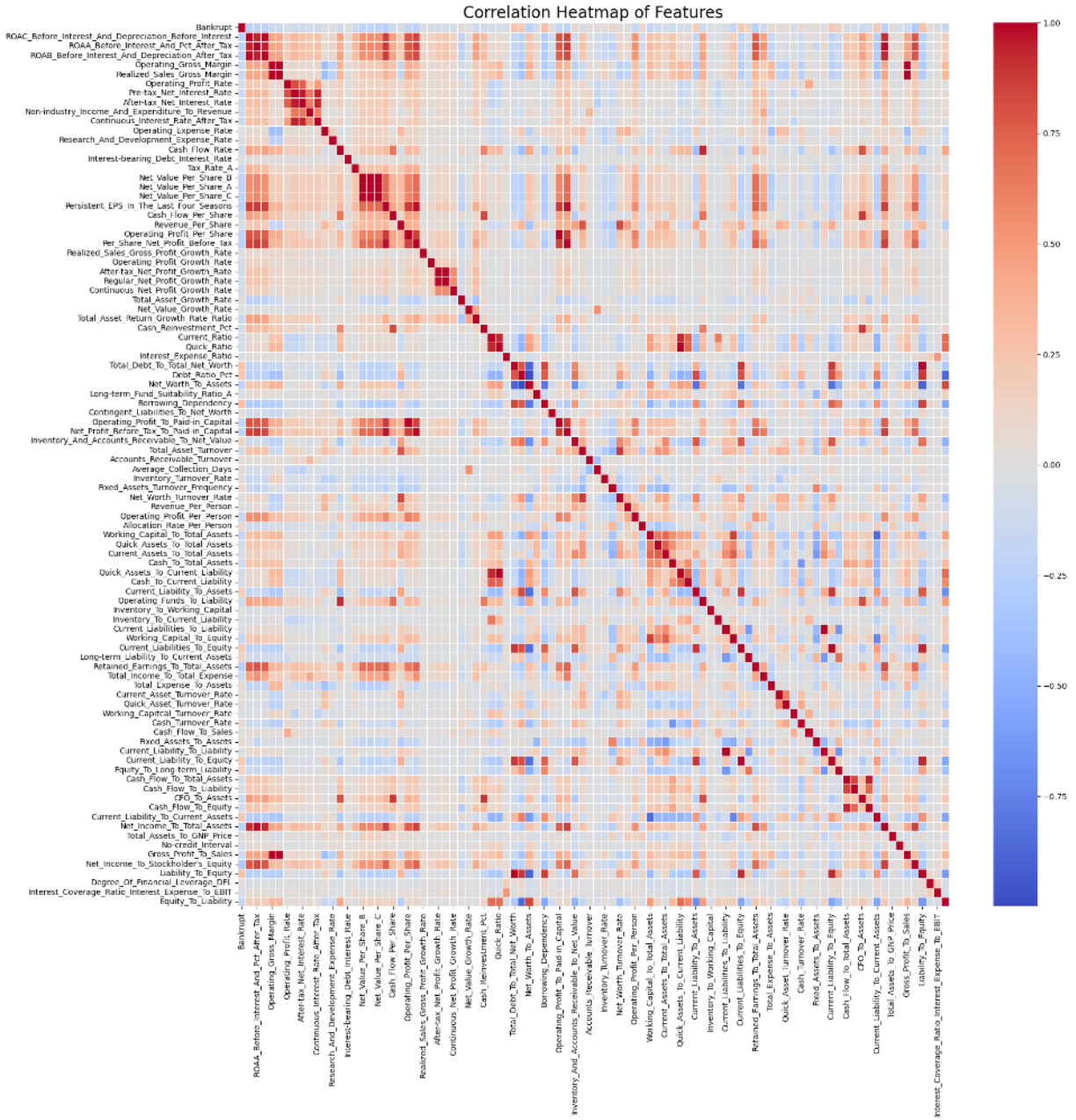


<Figure3. Feature Outliers and Bankruptcy Rate>

2.2 Correlation Analysis: Relationship between Features

Next, we checked the correlations between numerical features. As a result, we observed high correlations in many feature pairs (Figure 4). We found that 131 pairs of features have an absolute correlation above 0.7. The high absolute correlation values are due to these features being highly linearly related to each other. This often occurs when features are derived from similar underlying metrics or represent the same concept in different ways. This suggests that there may be multicollinearity issues in subsequent modeling steps.

Since there are many variables and multicollinearity issues, we could solve this using Feature Selection or PCA in the modeling section.

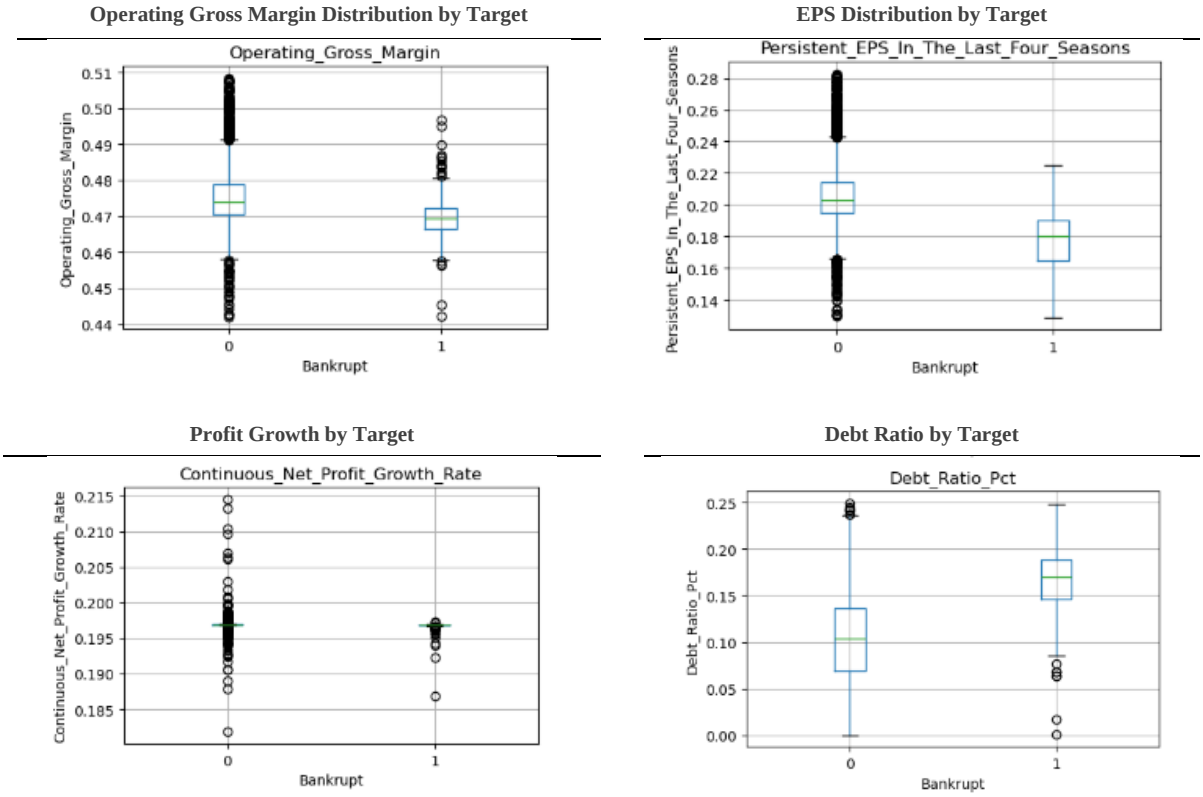


<Figure 4. Correlation Matrix of Numerical Features>

2.3 Relationship between Numerical Features and Target Variable

If the distribution characteristics of a feature differ according to the target variable value, that feature can be considered useful for prediction. Looking at the box plots, variables included in the categories of profitability, profit, growth, and financial stability will likely be useful for bankruptcy prediction (Figure 5).

- Profitability: Operating Gross Margin, ROA related features, Sales Gross Margin, etc.
- Profit: EPS, Operating Profit, Net Profit, Cash Flow, etc.
- Growth: Net Profit Growth Rate, Operating Profit Growth, etc.
- Stability (Solvency or Liquidity Ratios): Current Ratio, Quick Ratio, Debt Ratio, Net Worth to Asset, etc.



<Figure 5. Boxplots of Selected Features>

IV. Preprocessing and Training Data Development

1. Train-Test Splitting and Scaling

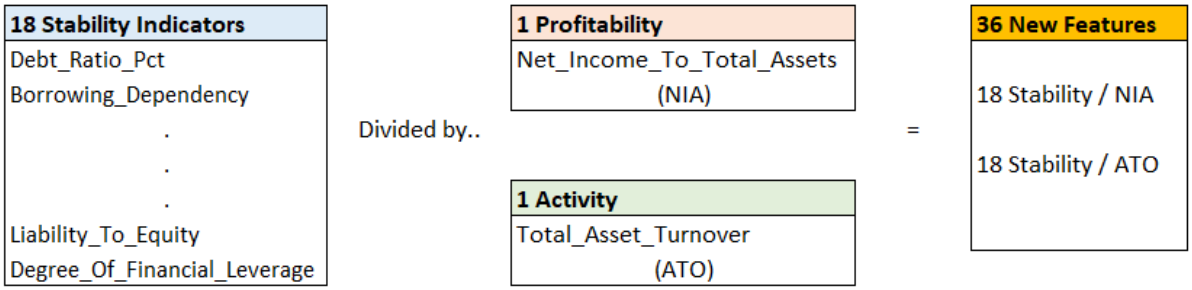
Before preprocessing data, we first perform a train-test split to prevent data leakage and then apply Min-Max scaling to the numerical features. Scaling before modeling is crucial because it ensures that all features contribute equally to the model, preventing bias due to differing scales.

2. New Features Generation

2.1 Ratio between Features

Although there are already 93 features, new ones can be created and then selected through feature selection to identify the most useful ones. If financial stability deteriorates while profitability and activity remain low, the probability of bankruptcy is expected to increase. Considering this, features created by dividing stability indicators by profitability and activity indicators are likely to improve model performance.

However, given the large number of possible combinations of these indicators, it is impractical to consider every combination. Additionally, a high correlation was observed among features within the same category. Therefore, a few key stability indicators will be used to create new features by dividing them by representative profitability and activity features, such as Net_Income_To_Total_Assets and Total_Asset_Turnover.



<Figure 6. New Feature Generation 1>

2.2 Outlier Dummy Features

During the exploratory data analysis, we observed that outliers in some features were significant for predicting the target variable. Therefore, dummy features were created to indicate whether these variables were outliers, particularly those with a high bankruptcy ratio.

The process involved creating dummy variables based on outlier detection for numerical features in the dataset. First, the z-scores for each numerical feature were calculated to standardize the data, which allows for the identification of outliers. An outlier was defined as a data point with a z-score above or below a certain threshold, typically 3 or -3. For each feature, a dummy variable was created where a value of 1 was assigned if the data point was an outlier, and 0 otherwise.

By creating feature ratios and dummy variables, 66 new features were added, resulting in a total of 158 features used for model building.

V. Modeling and Results

1. Balancing Data Using SMOTE

As previously discussed, our target variable is highly imbalanced, with only 3.2% of the observations being 1 and the rest being 0. If this imbalance is not addressed, bias towards the majority class and overfitting may arise during feature selection or modeling.

Therefore, we will use SMOTE to address the imbalance. SMOTE is an oversampling technique that resolves the problem of data imbalance by generating additional data for the minority class in imbalanced datasets.

2. Feature Selection: Lasso Regression and Random Forest

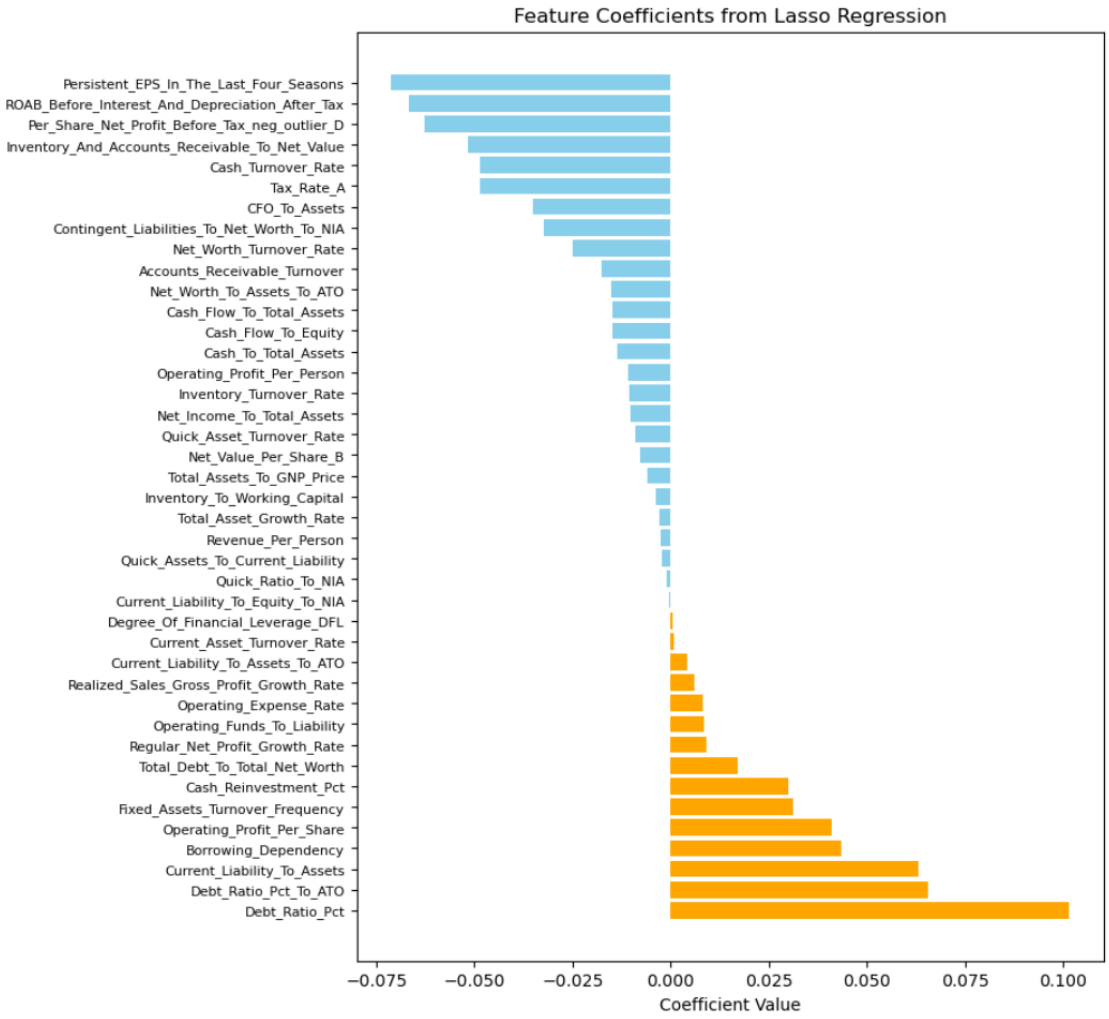
Currently, the dataset consists of 158 features, which presents a challenge in terms of model complexity and performance. We select features using Lasso regression and Random Forest to address this.

The motivation behind choosing these methods is to reduce the dimensionality of the dataset while retaining the most informative features. Lasso regression is known for its ability to penalize less important features, effectively shrinking their coefficients to zero, while Random Forest provides insight into feature importance based on the contribution of each feature to the model's prediction accuracy. By applying these methods, we can streamline the dataset, improve model interpretability, and enhance predictive performance.

2.1 Feature Selection Result (1): Lasso Regression

Through feature selection using Lasso regression, 41 features were selected for modeling, which is less than a third of the original 158 features.

Figure 7 shows the selected features and coefficients of them. Profitability and activity indicators were selected as having an inverse relationship with bankruptcy, while features related to solvency showed a positive relationship with bankruptcy, with several indicators being chosen.



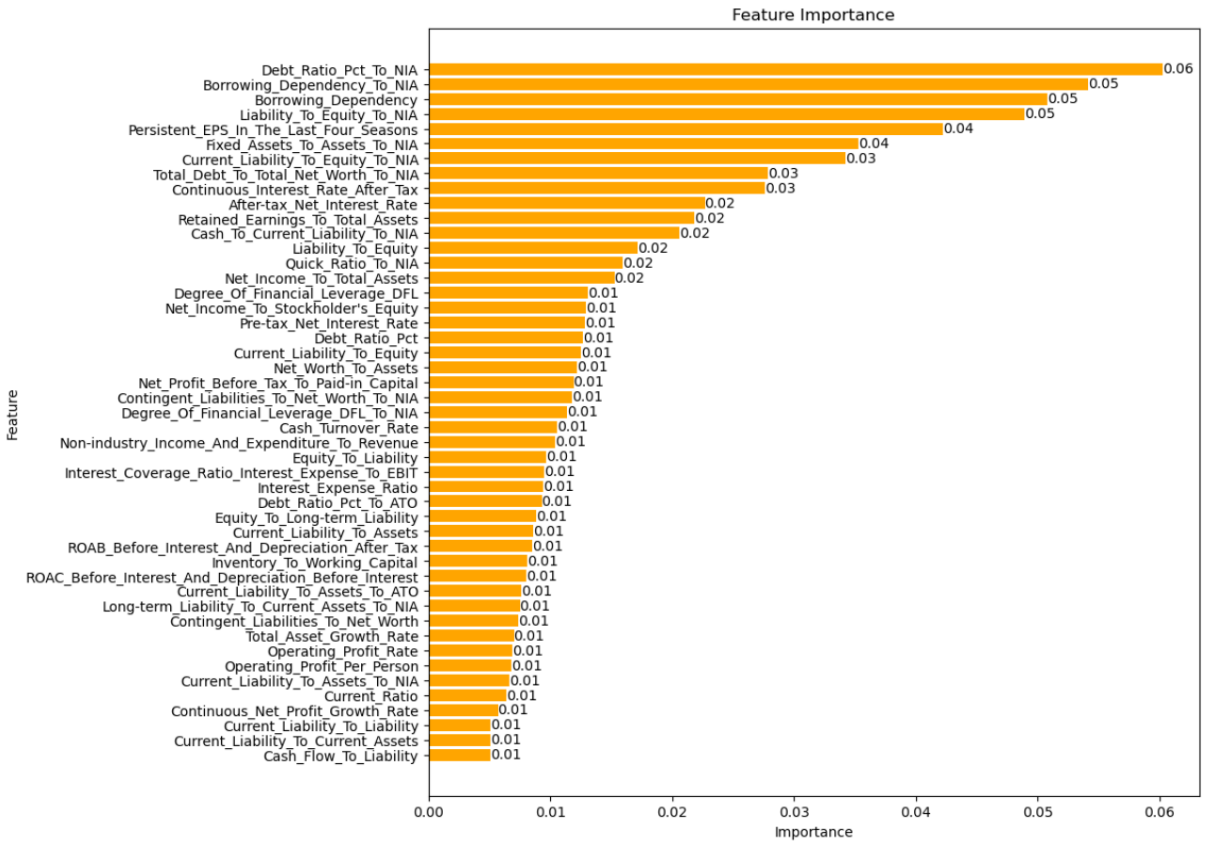
<Figure 7. Coefficients of Lasso Regression>

2.2 Feature Selection Result (2): Random Forest

When using the feature importance from the Random Forest, removing features with low importance results in a total of 47 features remaining. This is a bit more than the number selected using Lasso regression. Notably, many of

the high-importance features include those created by dividing solvency indicators by net profit margin, whereas none of the dummy variables were selected(Figure 8).

Now, we will proceed with modeling using the two selected feature sets in the following section.



<Figure 8. Feature Importance from Random Forest>

3. Model Building and Evaluation

3.1 Setting-Up Evaluation Metrics: Cost Implications of Error Types

Before proceeding with modeling, it is crucial to determine the performance metrics that will be used to evaluate the model. When modeling the likelihood of corporate bankruptcy from the perspective of a lending or investment institution, the cost of incorrect predictions can vary significantly depending on whether the error is a false negative or a false positive. Let’s examine the costs associated with each error type.

- **False Negative (Type II Error) Costs:**
  A FN occurs when a loan is issued to a company that will actually default, but the model incorrectly predicts that it will not. This results in significant losses, including the inability to recover the principal and interest. Furthermore, there would be additional costs such as recovery expenses, legal fees, and damage to reputation.

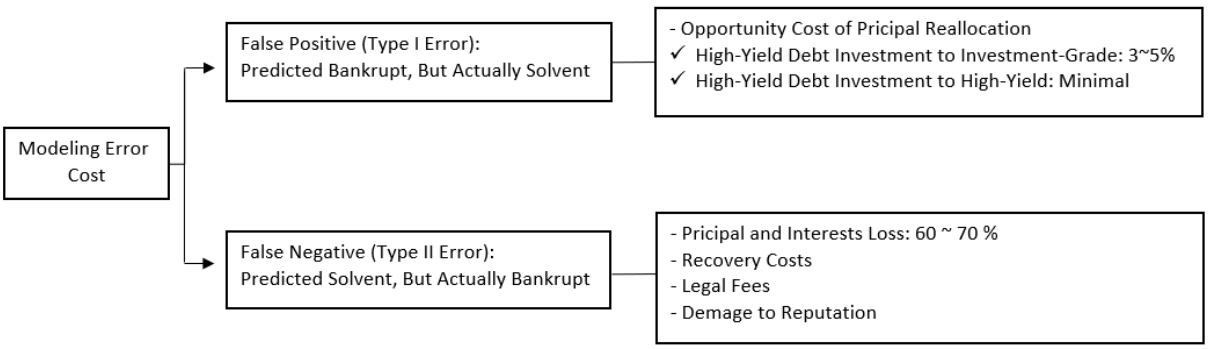
- **False Positive (Type I Error) Costs:**

  An FP occurs when a company that will not default is incorrectly predicted to default, leading to the rejection of a loan. This can result in opportunity costs, as the bank or investor might miss out on potential returns from this company.

- **Comparing FN and FP Costs**

  While FP can lead to missed opportunities, these costs are generally much smaller compared to FN costs. Even if a default was incorrectly predicted (FP), the opportunity cost would be small because the funds could have been reallocated to loans or investments in other companies.

  If the yield spread between speculative-grade corporate bonds and investment-grade corporate bonds is 3–5%, then if a speculative-grade company is incorrectly predicted to be bankrupt and investment-grade bonds are instead invested in, the opportunity cost can be considered to be 3–5%. If invested in another lower-rated company, this opportunity cost might be smaller. On the other hand, if a company is incorrectly predicted not to go bankrupt, the risk of principal loss is significant. For speculative-grade bonds, in general, the average recovery rate is between 30% and 40%, meaning losses could range from 60% to 70%. Additional costs such as legal fees and opportunity costs due to time can lead to even greater losses. Thus, the cost difference between these two types of errors can be ten to several dozen times greater.



<Figure 9. Cost Difference between Type I Error and Type II Error>

To reflect the cost differences associated with these types of errors in machine learning modeling, it is essential to adjust the evaluation metrics used. In this case, where the costs of false negatives (FN) and false positives (FP) differ significantly, recall is far more important than precision. Additionally, in cases of severe imbalance, if the predictions for the minor class are all incorrect but the predictions for the major class are correct, the accuracy will still be high. Therefore, accuracy should not be used as a single major performance metric.

Instead, the F-beta score can be adjusted by modifying the beta parameter to reflect the greater importance of recall (minimizing FN) over precision (minimizing FP). In the F-beta score, the beta parameter is set to a value greater than 1 when the recall is more important and less than 1 when the precision is more important, reflecting the relative importance of the two metrics. If the cost of FN errors is much greater than FP errors, a higher beta value should be chosen to emphasize recall.

The F-beta score function is as follows:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$

<Equation 1. F-beta Score Function>

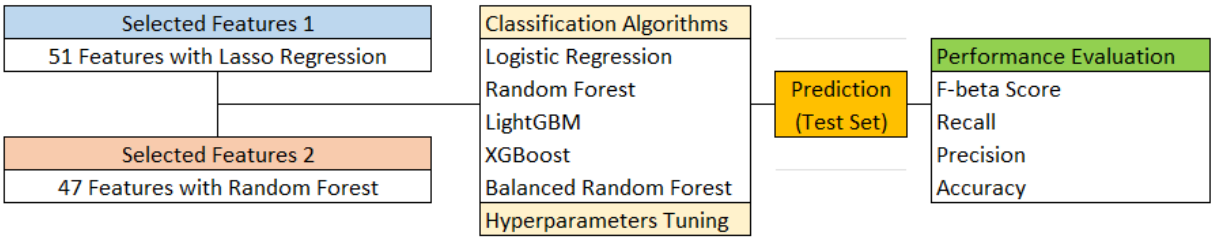
In the above equation., we can set the beta parameter to represent the ratio of recall importance to precision importance. Specifically, the square of beta equals the ratio of the weights assigned to recall and precision[2]. As discussed earlier, the cost difference between FN and FP can be tenfold or more. Therefore, in this analysis, we set beta to 4, meaning recall is weighted 16 times more heavily than precision.

3.2 Model Building and Performance Comparison

The machine learning algorithms for modeling the data include Logistic Regression, Random Forest, LightGBM, XGBoost, and Balanced Random Forest.

The first four models utilize a dataset balanced using SMOTE, while the Balanced Random Forest model is applied to the original imbalanced data. That is, in a Balanced Random Forest, the minority class remains unchanged while the majority class is undersampled to achieve balance. This balanced data is then modeled using Random Forest, and the process is repeated multiple times, with the final result being derived by voting on the outcomes.

With selected features using Lasso Regression and Random Forest, hyperparameters for each algorithm were tuned using the F-beta score function. The models were then evaluated on the test set, and their performance was compared.



<Figure 10. Modeling Process>

The results obtained from the above process are shown in <Figure 11>, and the following conclusions can be drawn.

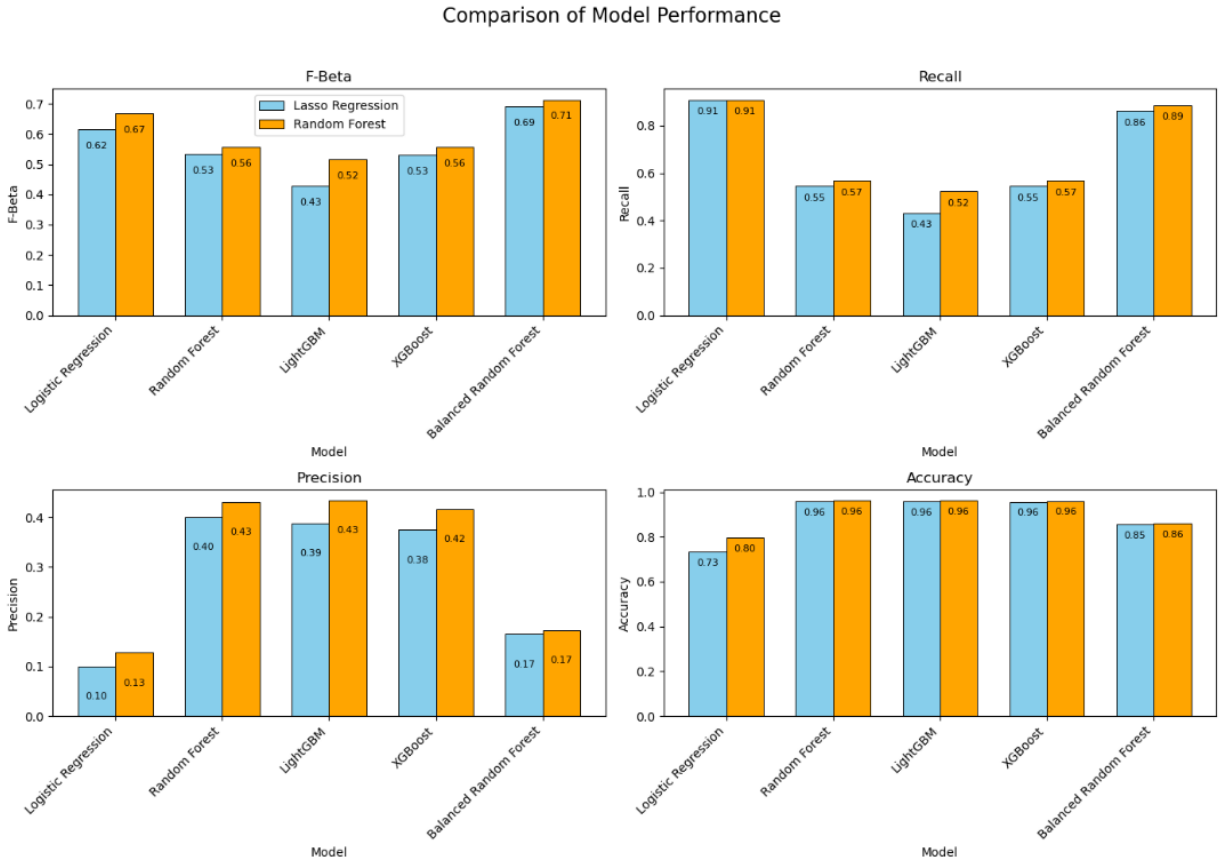
- **Lasso vs. Random Forest as Feature Selection Method**

  Logistic Regression benefits the most from Random Forest feature selection, improving all of the performance metrics. This suggests Random Forest feature selection better aligns with Logistic Regression's strengths. For models like LightGBM and XGBoost, the differences are subtle but generally favor Random Forest feature selection, implying that it better captures the nuances of the data.

[2] F-beta Score in Keras Part I. Creating custom F1 score for binary classification problems in Keras (https://towardsdatascience.com/f-beta-score-in-keras-part-i-86ad190a252f)

- **Best Model: Balanced Random Forest vs Logistic Regression**

While its accuracy is not the highest with 86.0% with Random Forest Feature Selection, its high F-beta score(0.713) and recall(88.6%) highlight its effectiveness in identifying bankrupt companies. However, from a more conservative perspective aiming to reduce false negative errors, Logistic Regression with a higher recall of 91% could be an alternative.



<Figure 11. Comparison Model Performance>

VI. Practical Applications of Findings

The insights gained from the model evaluation and comparison have practical implications for financial institutions and investment firms involved in assessing corporate bankruptcy risk. Here, we explore how these findings can be applied in real-world scenarios to enhance decision-making and risk management.

1. Enhanced Risk Management Strategies

The model results indicate that the Balanced Random Forest model provides the best performance in terms of the F-beta score, despite not having the highest accuracy. This suggests that the Balanced Random Forest model is highly effective at identifying bankrupt companies, which is crucial for financial institutions looking to minimize their exposure to high-risk clients. Implementing this model can lead to more informed lending decisions, helping institutions avoid loans to companies with a high likelihood of default.

2. Strategic Investment Decisions

Investors and asset managers can use the model results to refine their investment strategies. The high recall rate of the Balanced Random Forest model ensures that most bankrupt companies are identified, which can guide

investment decisions and portfolio management. For instance, an investment firm can use this model to screen out companies with a high risk of bankruptcy from their investment portfolios, thereby safeguarding their investments and enhancing overall portfolio performance.

Additionally, if an investor's strategy prioritizes minimizing false negatives to avoid missing potential bankruptcies, the Logistic Regression model with a high recall rate can serve as an alternative. This model would help in situations where the cost of failing to identify a default is more critical than incorrectly predicting a default.

3. Regulatory Compliance and Reporting

Financial institutions are often required to comply with regulatory standards and report on the risk profiles of their loan portfolios. The findings from this analysis can assist institutions in meeting these requirements by providing a robust model for assessing bankruptcy risk. The Balanced Random Forest model, with its emphasis on minimizing false negatives, can help ensure that institutions accurately report their exposure to high-risk entities, which is crucial for regulatory compliance.

4. Decision Support Systems

Incorporating the best-performing models into decision support systems can provide valuable insights for executives and decision-makers. These systems can update data quarterly based on financial statements, allowing for more timely and informed decision-making.

VII. Further Research

While the current study has provided valuable insights into the effectiveness of different machine learning models for bankruptcy prediction, there are several areas for further research that could enhance model performance and broaden the applicability of the findings. Below are key areas where additional exploration could lead to improved results and deeper understanding:

1. Incorporating Analysts' Financial Estimates

Current models primarily rely on historical financial data, which can be limited in predicting future bankruptcy risks. A promising avenue for further research is the integration of analysts' financial estimates into the modeling process. Financial analysts provide forward-looking estimates based on industry trends, economic forecasts, and company-specific information. These estimates often include projected earnings, revenue growth, and other key financial metrics that can offer additional insights into a company's future performance.

Incorporating these estimates could enhance the models' ability to predict bankruptcy by providing a more comprehensive view of a company's financial health. For instance, projected earnings growth or analyst downgrades can serve as leading indicators of potential financial distress that historical data alone might not capture. By integrating analysts' estimates, models could become more sensitive to emerging financial trends and potential risks.

2. Enhancing Model Complexity

The current models, while effective, can benefit from further exploration of more complex or hybrid modeling approaches. Future research could focus on:

- Ensemble Methods: Combining multiple models or using advanced ensemble techniques could improve predictive performance. For instance, blending the Balanced Random Forest with models like XGBoost or LightGBM might leverage the strengths of each algorithm.
- Deep Learning Approaches: Investigating deep learning methods such as neural networks could capture complex patterns in the data that traditional models might miss. Recurrent neural networks (RNNs) or transformers could be explored for their potential to handle sequential and temporal data.
- Feature Engineering and Selection: Further research into feature engineering and selection techniques could uncover additional predictors of bankruptcy. Techniques such as feature interaction, polynomial features, or domain-specific knowledge could enhance model inputs.