Springboard Data Science Capstone Project

Predicting Corporate Bankruptcy Using Financial Ratios

A Focus on Minimizing Type II Errors

Suk Won Choi

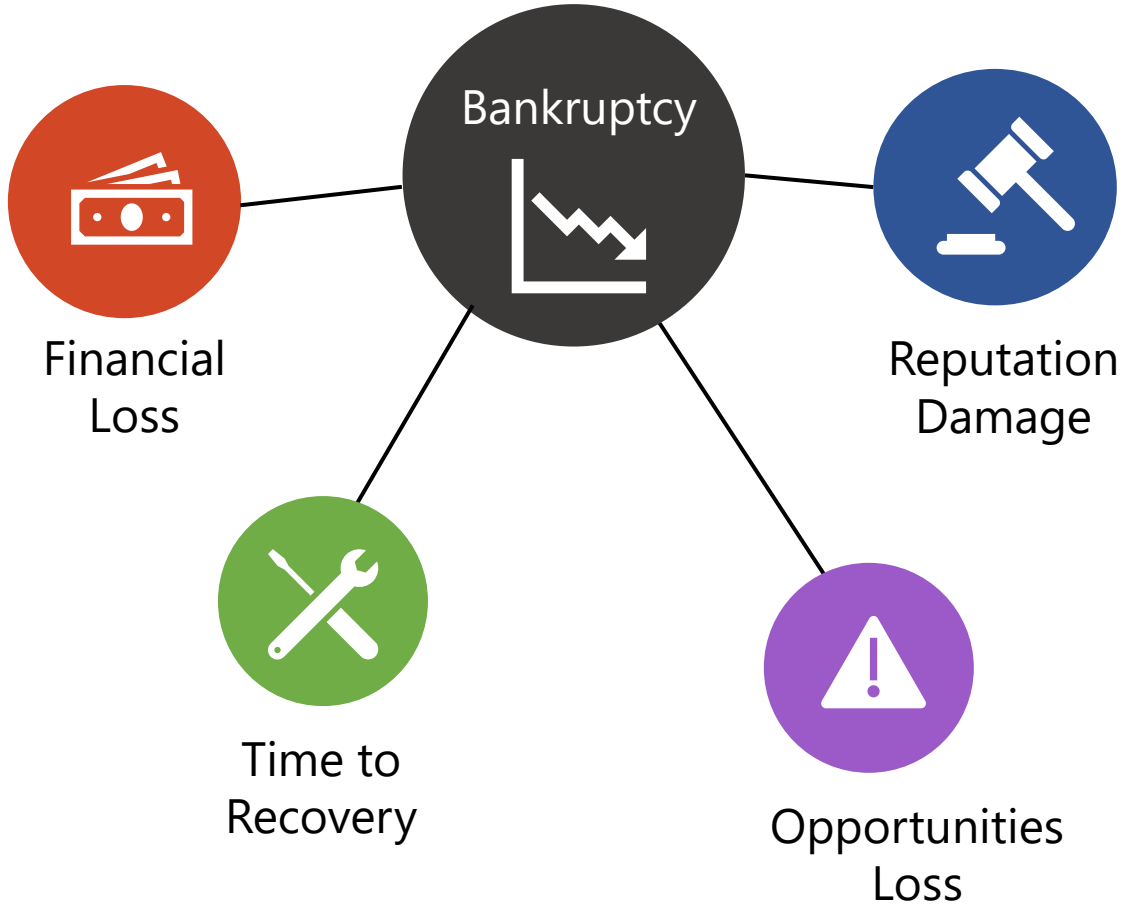
Problem & Goal

Corporate Bankruptcy Can Lead To:

- Financial Loss
- Damage In Reputation
- Time & Efforts To Recovery
- Loss In Opportunities

Develop a Reliable Model That:

- Effectively Distinguish Between Companies At High Risk of Bankruptcy and Those That Are Not
- Minimize Costly Errors



Overview

At-a-glance

DATA

Provide a summary of the dataset, including key features, class distribution, and any initial insights.

EDA

Analyze the data to uncover patterns, relationships, and anomalies, and visualize key trends

FEATURE ENGINEERING

Create new features, handle outliers, and address multicollinearity to enhance model performance.

MODELING

Build and compare multiple machine learning models to predict bankruptcy, optimizing for the best performance.

PERFORMANCE EVALUATION

Assess relevant model metrics to determine the most reliable prediction model.

Dataset Overview (1) Data Source

Data Source: UCI Machine Learning Repository

Taiwanese Bankruptcy Prediction

Donated on 6/27/2020

The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification
Feature Type	**# Instances**	**# Features**
Integer	6819	95

Data Source: https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction

Dataset Overview (2) Binary Target Variable

High Imbalance: Bankrupcy vs Non-Bankruptcy

96.9%

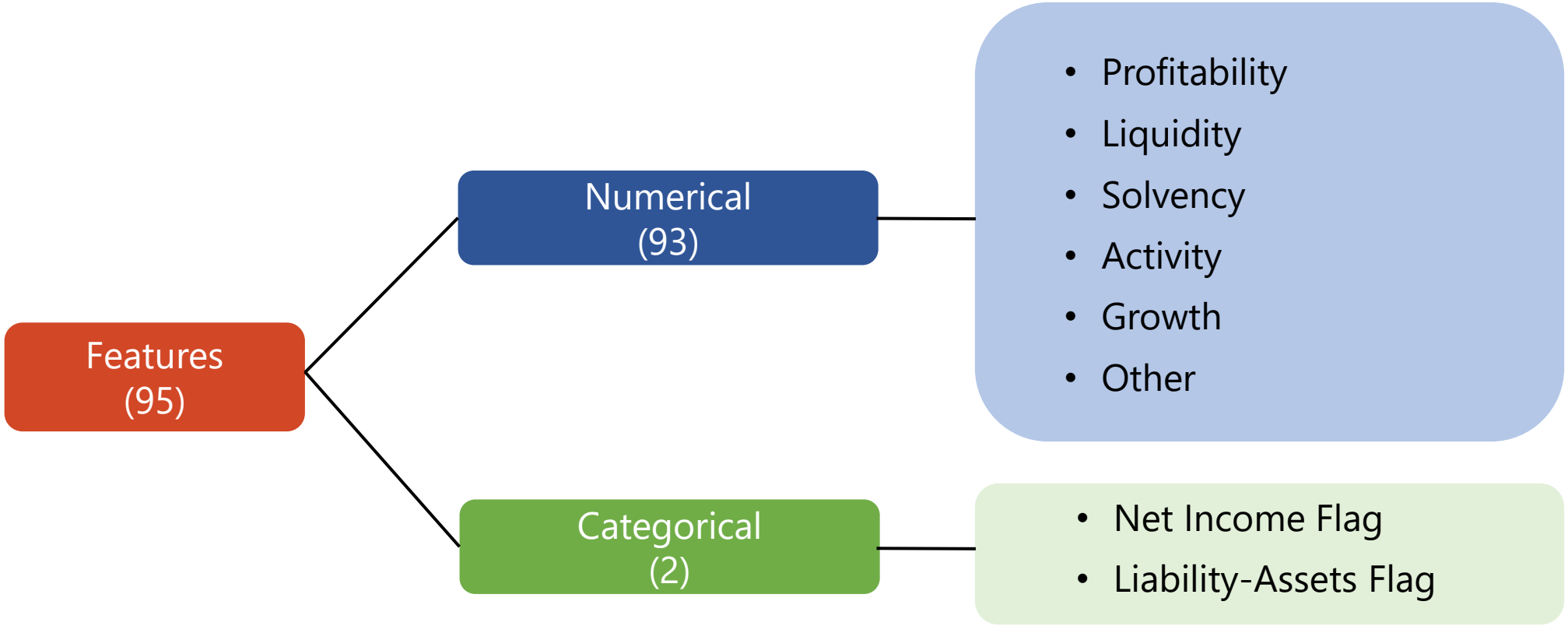
Bankruptcy

3.1%

Non-Bankruptcy

Dataset Overview (3) Features

93 Numerical Features and 2 Categorical Features

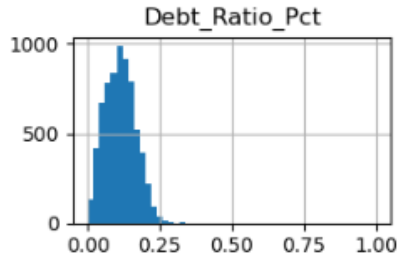


Exploratory Data Analysis (1) Numerical Features

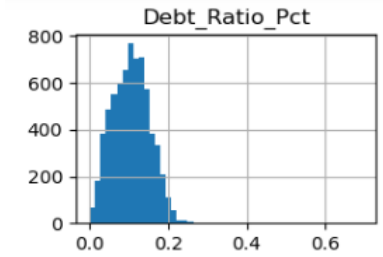
Trade-off: Clean Distribution vs. Retaining Critical Information

- Skewed Distribution by Outliers

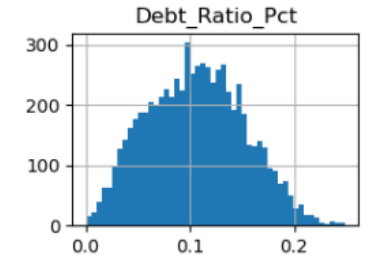
Original



Log Transformed

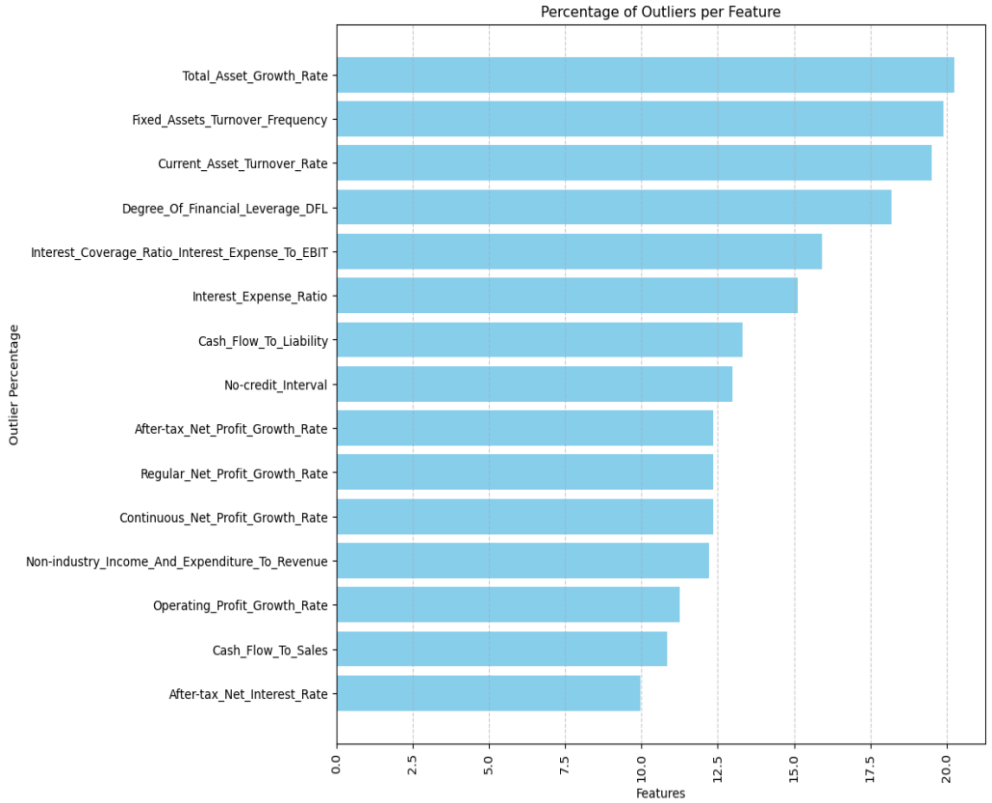


Outlier Removed



VS

- High Proportion of Outliers



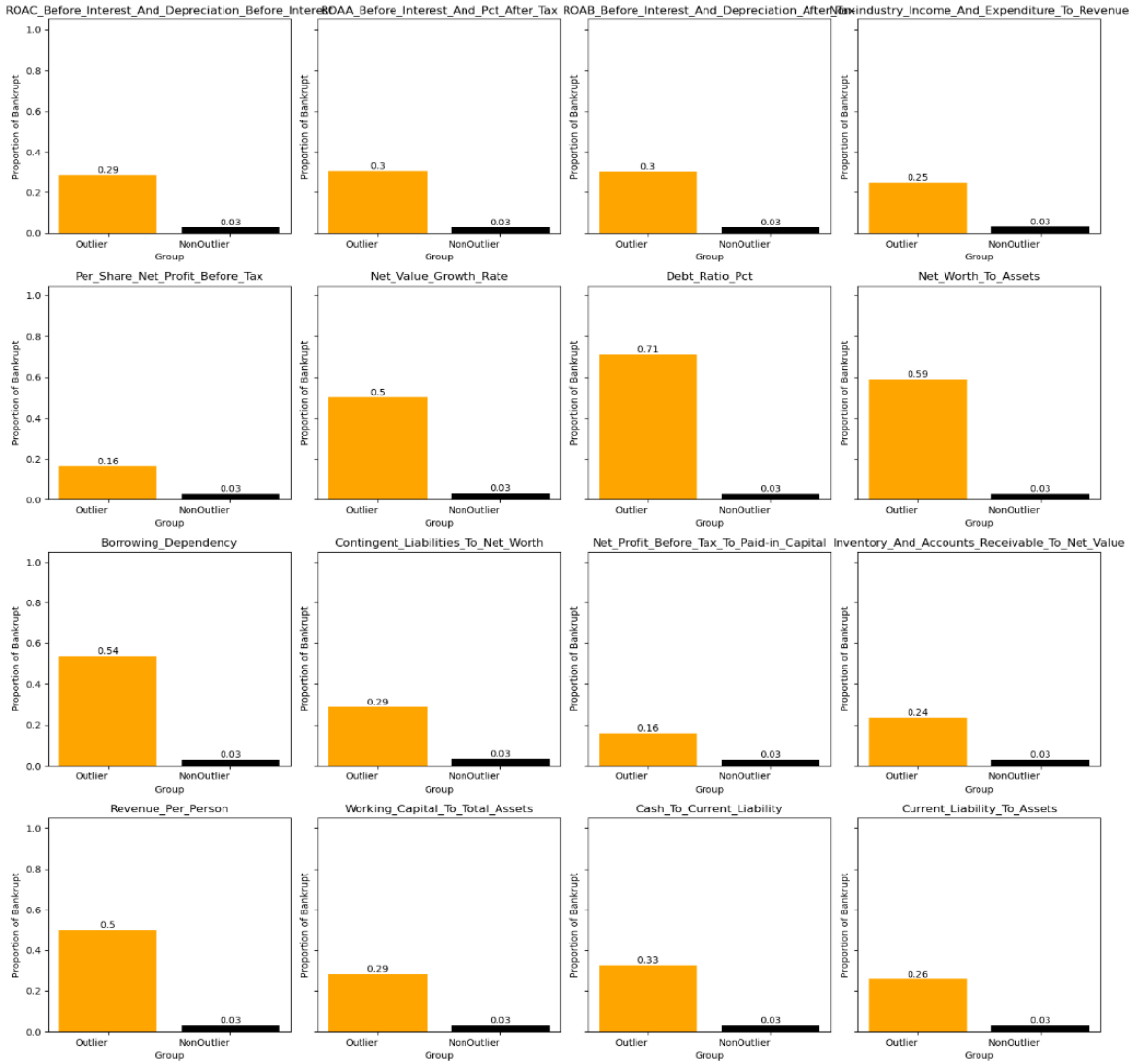
Exploratory Data Analysis (2) Outliers

Outliers Matter:

How Rare Events Drive Predictions

- Critical Insights Hidden in Outliers:

  Essential for Predicting Extreme Events



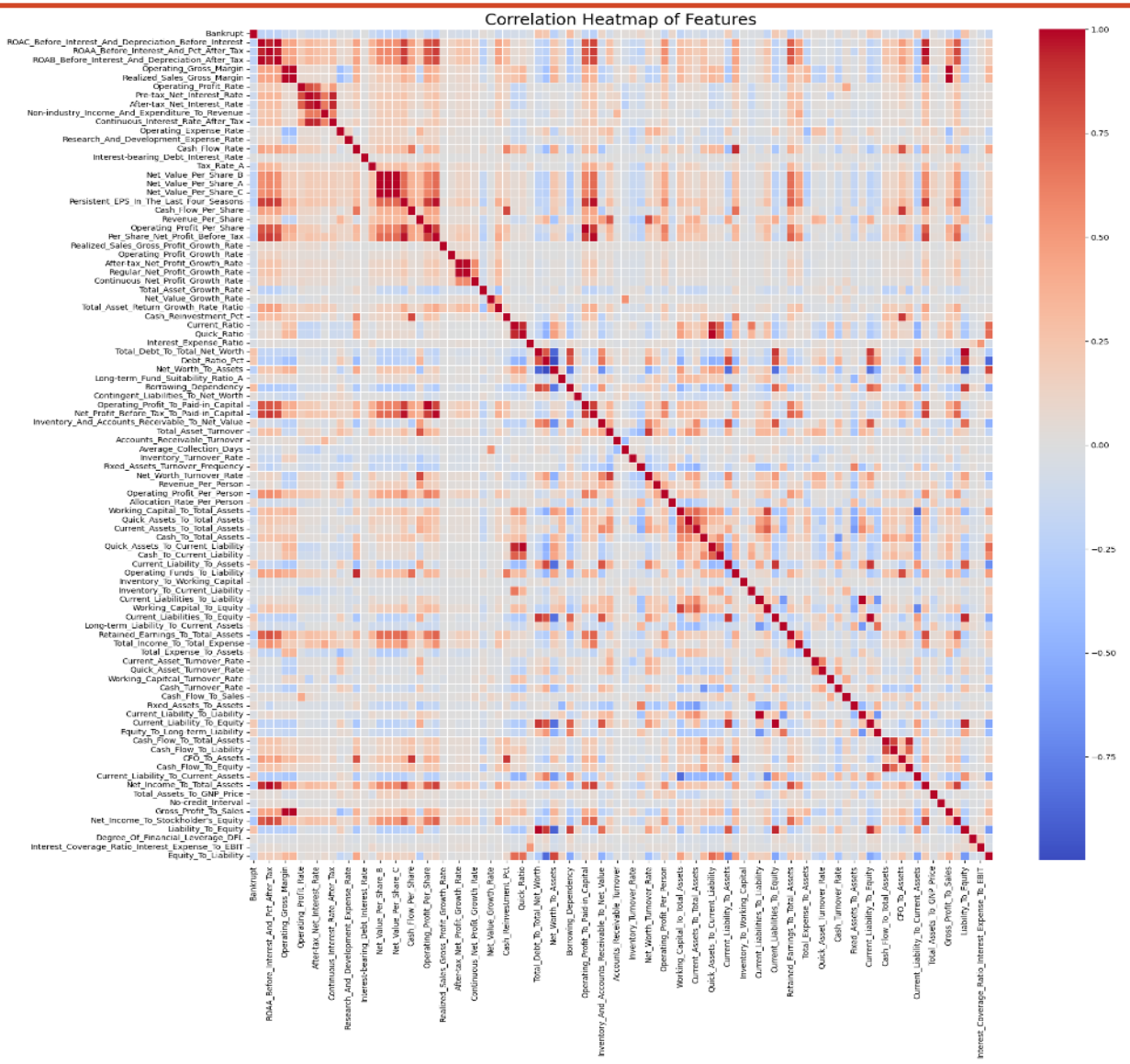
Exploratory Data Analysis (3) Relationship Between Features

Multicollinearity Alert

- 131 Feature Pairs with High Correlation (>0.7)

Tackling Multicollinearity

- Applying Feature Selection or Dimensionality Reduction



Exploratory Data Analysis (4) Numerical Features and Target Variable

Useful Features For Prediction

- **Profitability**:

✓ Operating Gross Margin, ROA related features, Sales Gross Margin, etc.

- **Profit**:

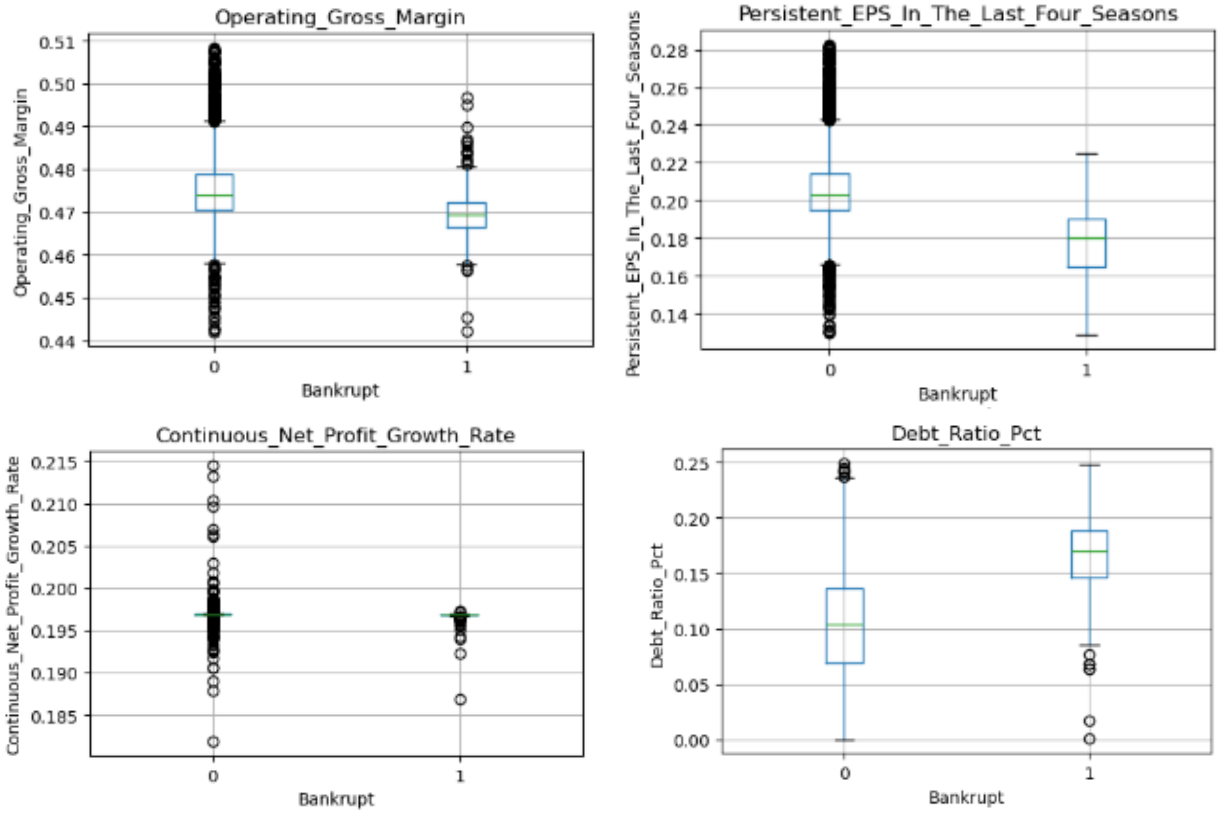
✓ EPS, Operating Profit, Net Profit, Cash Flow, etc.

- **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.



Feature Engineering: New Feature Generation

1 Ratio Between Features: 36 New Features

18 Stability Indicators
Debt_Ratio_Pct
Borrowing_Dependency
.
.
.
Liability_To_Equity
Degree_Of_Financial_Leverage

Divided by..

1 Profitability
Net_Income_To_Total_Assets (NIA)

1 Activity
Total_Asset_Turnover (ATO)

=

36 New Features
18 Stability / NIA
18 Stability / ATO

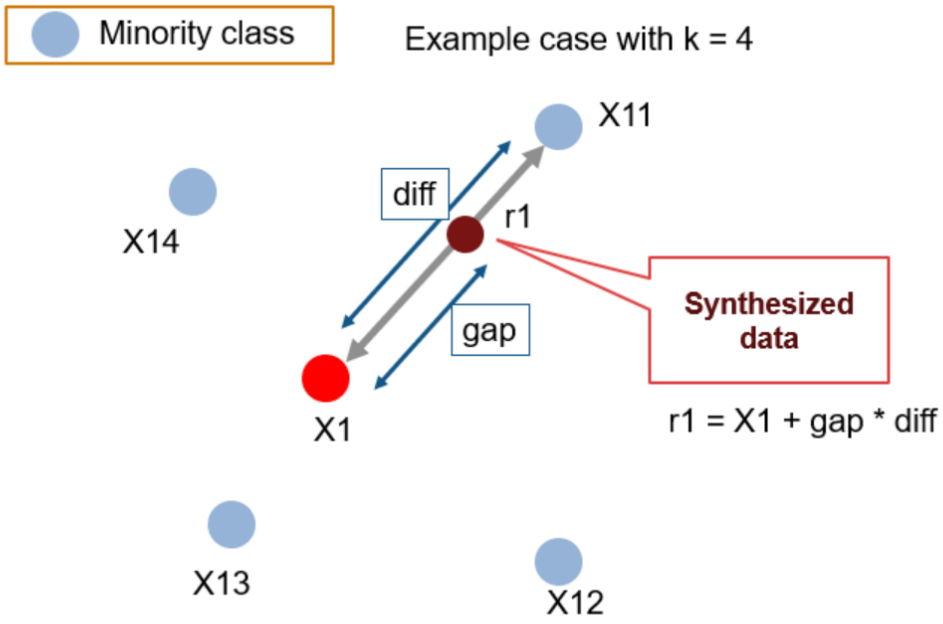
2 Outlier Dummy Features: 30 New Features

Modeling (1): Balancing the Dataset

SMOTE:

Synthetic Minority Oversampling Technique

- Generating synthetic examples for the minority class to improve model performance
- Analyzing existing minority data points and then generating new ones similar to them

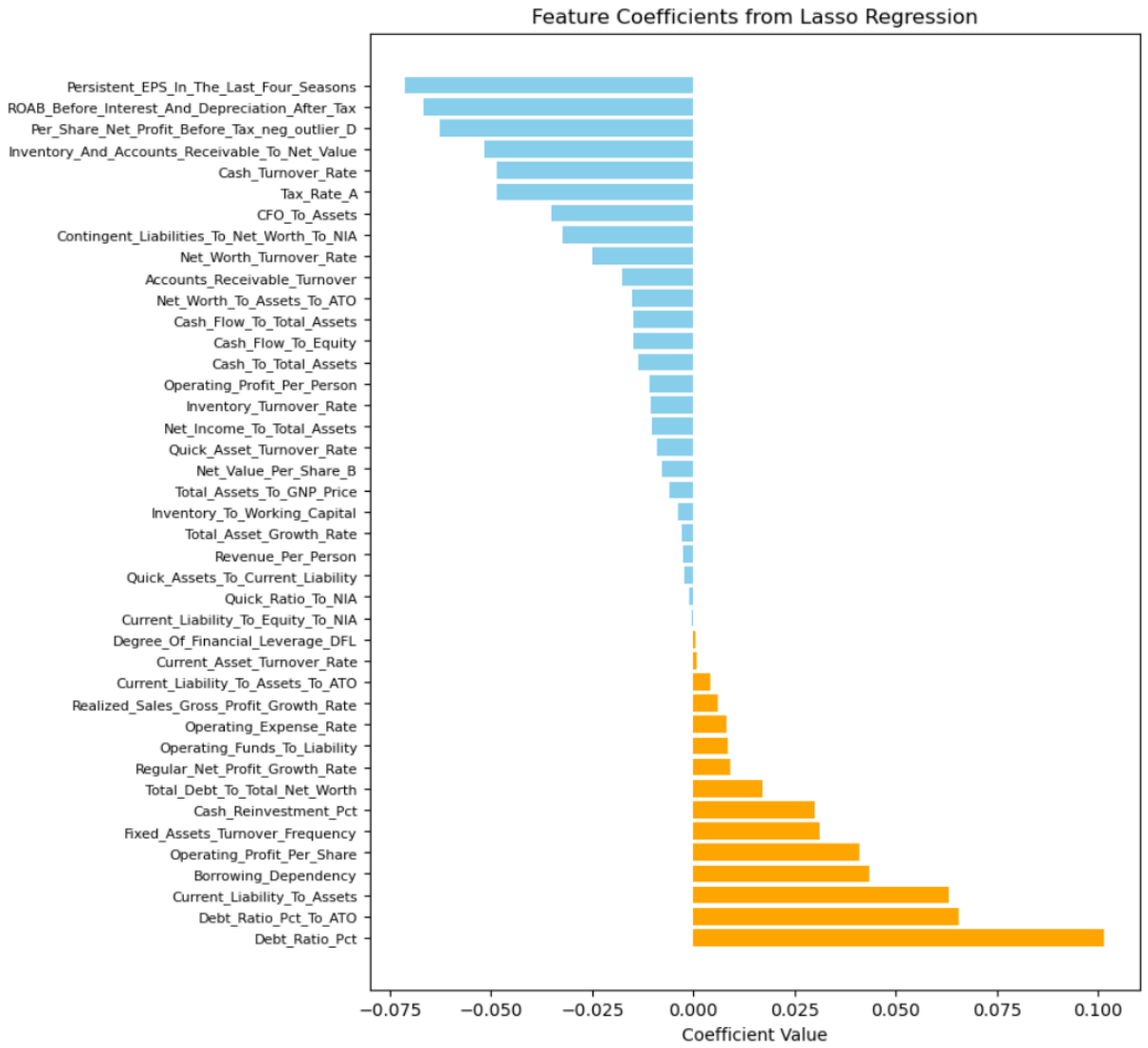


Source: GitHub

Modeling (2): Feature Selection

1 Lasso Regression

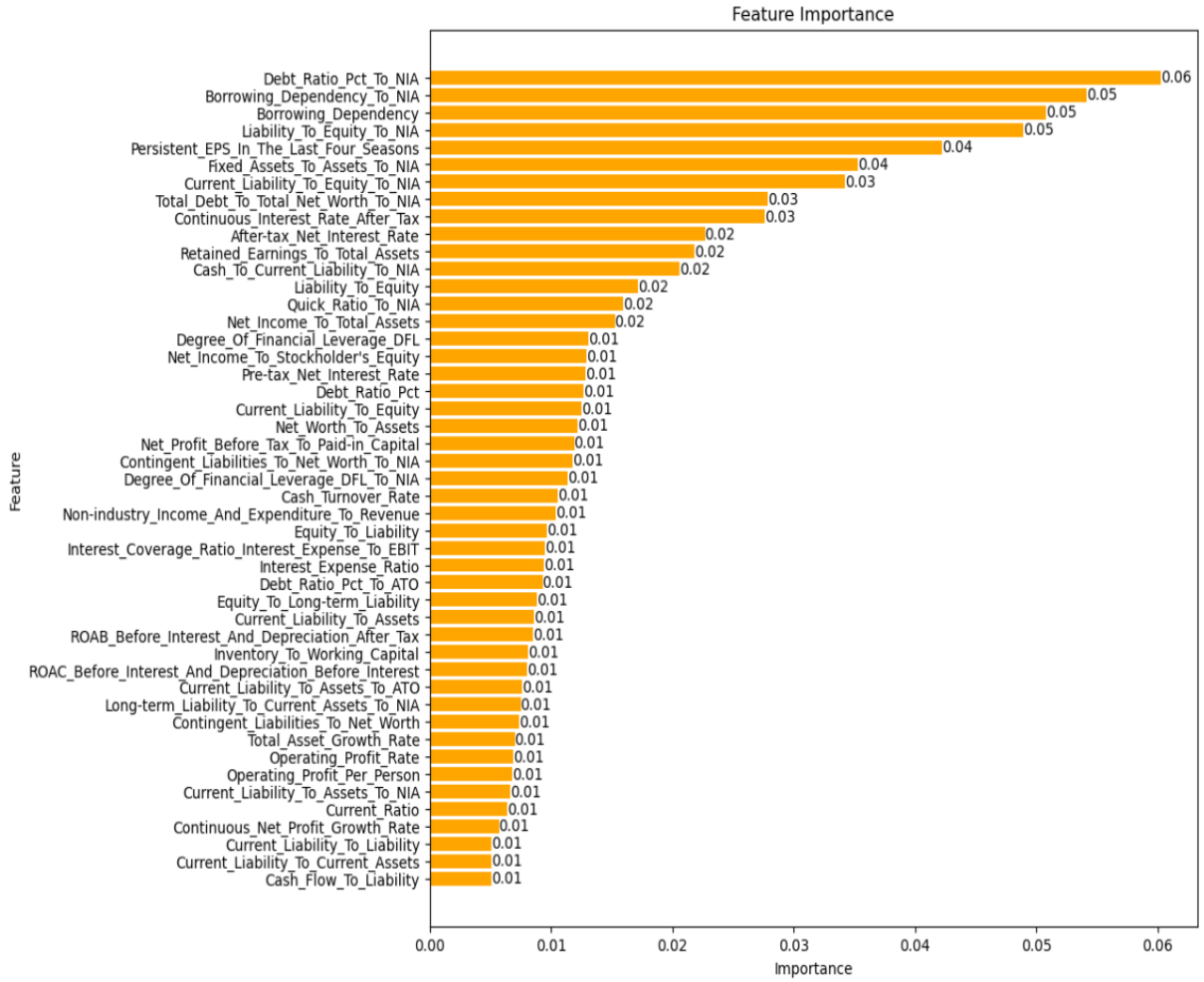
- Penalizing less important features, effectively shrinking their coefficients to zero
- Selected 41 features for modeling, which is almost a quarter of the original 158 features
- Inverse relationship with bankruptcy: Profitability and Activity Ratios
- Positive relationship with bankruptcy: Solvency and Liquidity Ratios
- 7 newly generated indicators included



Modeling (2): Feature Selection

2 Random Forest

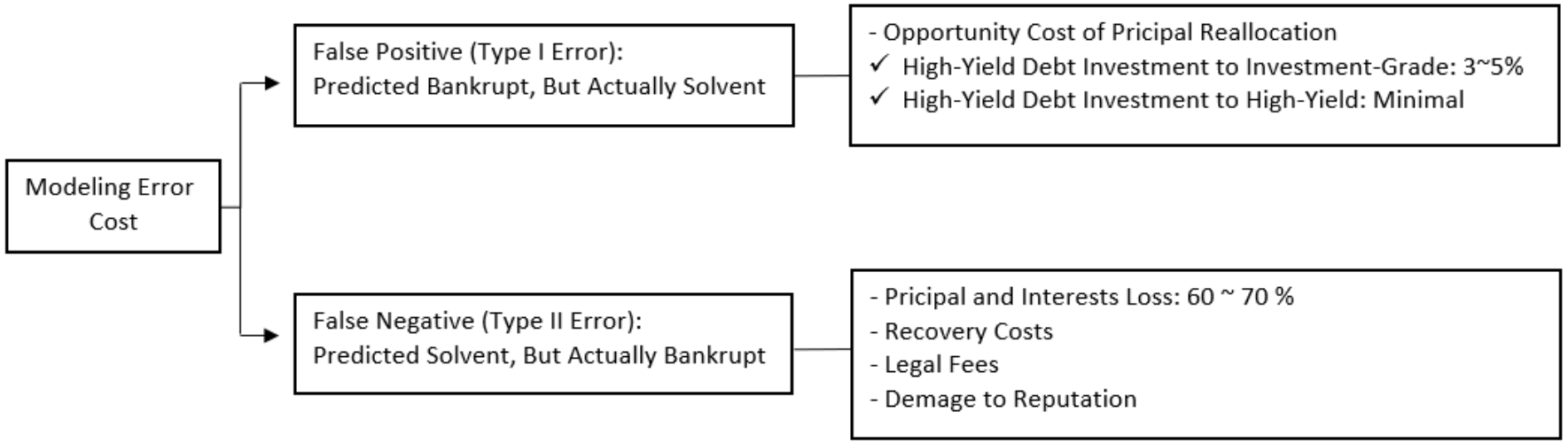
- Evaluated feature importance by measuring how much each feature reduces impurity
- Selected 47 features for modeling
  - ✓ Included 14 newly generated indicators based on interaction terms
  - ✓ 6 newly generated indicators among top 10 most important features



Modeling (3): Setting-Up Evaluation Metrics

Cost Implications of Error Types

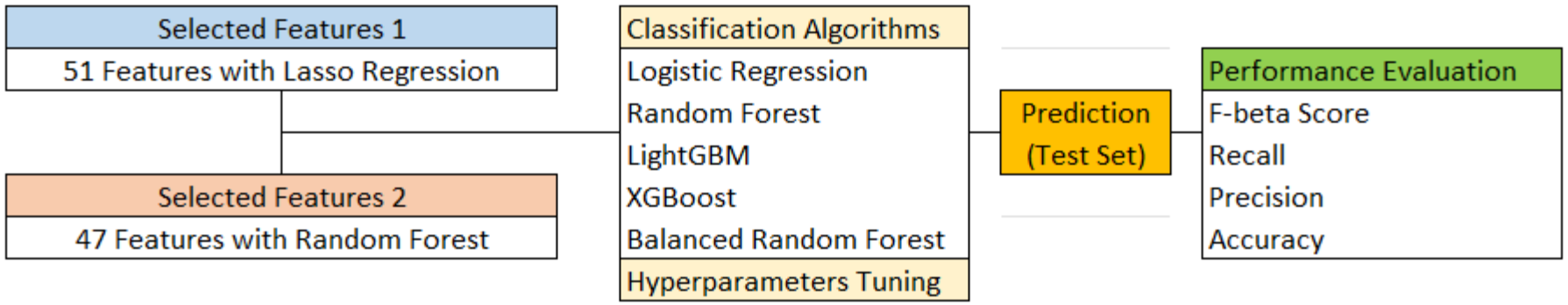
- False Negative Error Cost > False Positive Error Cost
- The most reliable performance metric: Recall > Precision > Accuracy
  - ✓ F-Beta Score: Enabling weight more on Recall than Precision



Modeling (4): Model Building

Fitting Models With 5 Different Algorithms

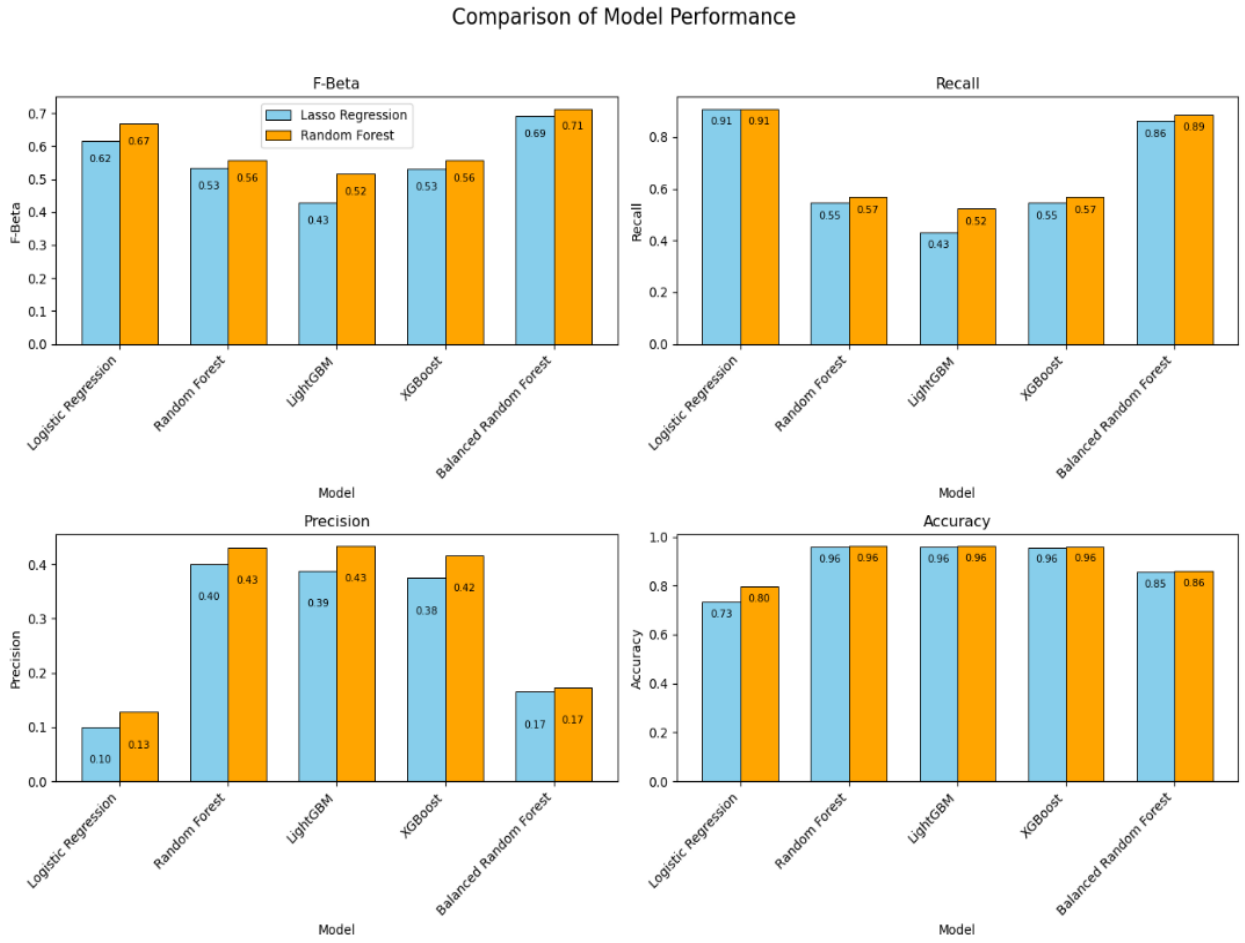
- Two-Sets of Selected Features
- Logistic Regression, Random Forest, LightGBM, XGBoost, and Balanced Random Forest
- Comparing 10 Results Using F-beta, Recall, Precision, and Accuracy



Modeling (5): Performance Evaluation

What is the Best Model?

- Random Forest Is Better for Feature Selection
- Best F-beta score achieved with Balanced Random Forest
  - ✓ Logistic Regression considered as a conservative alternative



Application of the Model

1. **Enhanced Risk Management Strategies**
2. **Strategic Investment Decisions**
3. **Regulatory Compliance and Reporting**
4. **Decision Support Systems**

Further Research

1 Incorporating Additional Data Than Historical Financial Data

- ✓ Analysts' Financial Estimates
- ✓ Stock Market Data
- ✓ Macro Economic Data

2 Enhancing Model Complexity

- ✓ Combining multiple models or using advanced ensemble techniques
- ✓ Investigating deep learning methods such as neural networks