Tanvi Mehta

Project Report
Bankruptcy Prediction

Problem Statement

Bankruptcy is a problem that has concerned entrepreneurs, researchers and even government for years. It is important to know the important to detect early signs for a company to evade bankruptcy. The task is to identify the most important features involved in the company bankruptcy.

The data has been sourced from Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. The dataset included about 7000 rows and 95 features.

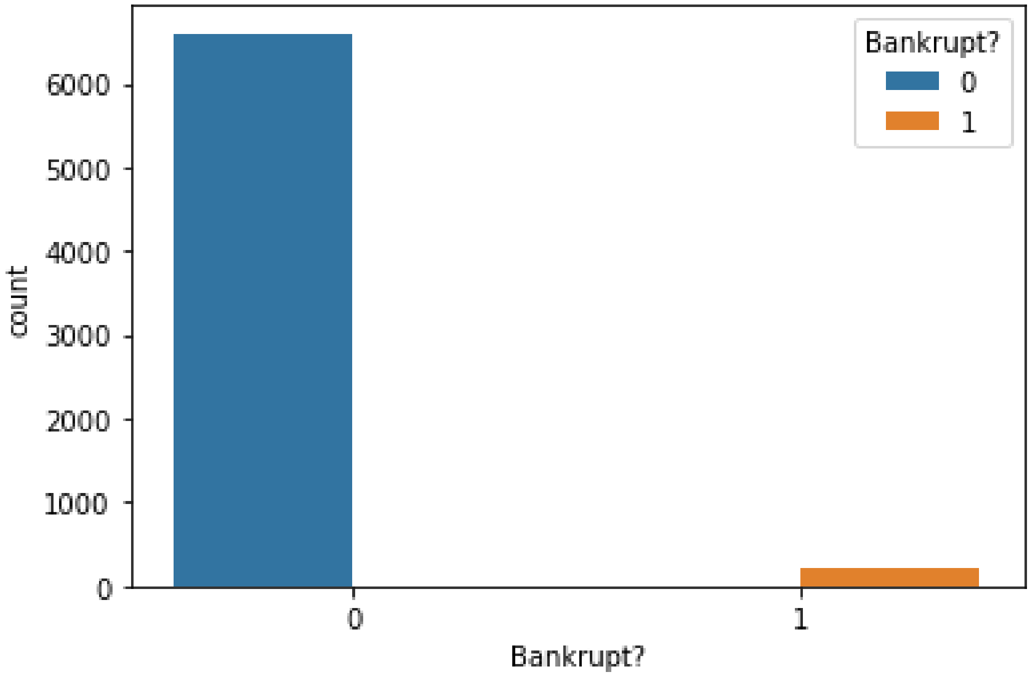
Data Wrangling

The first step of the project was to observe the data, looking at both the length and width of the dataset and accounting for any missing values. The dataset did not contain any missing values to be accounted for.

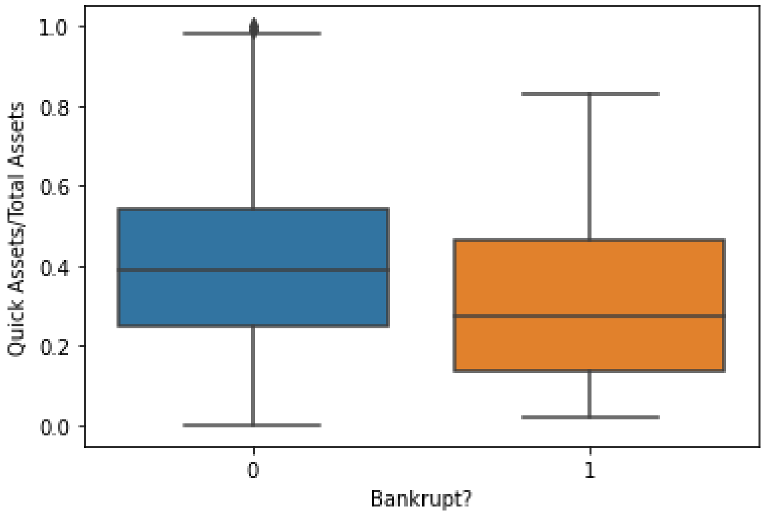
The next step was to look for outliers using boxplots. The outliers could not be excluded because it was shrinking the dataset by more than half, and this could have resulted in a huge loss in important data points.

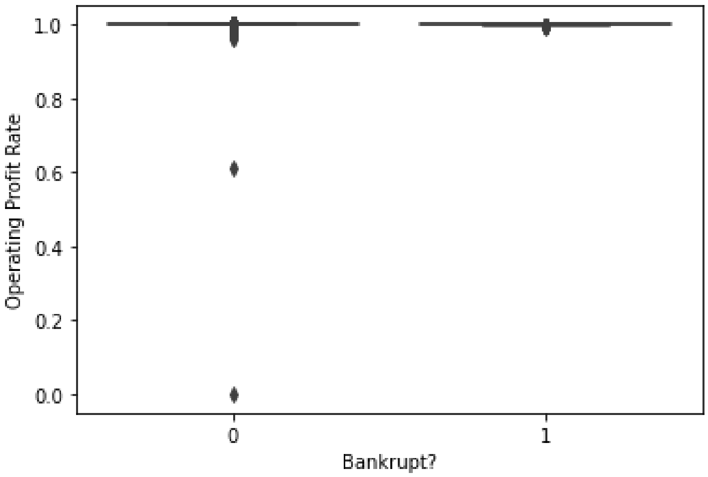
Exploratory Data Analysis

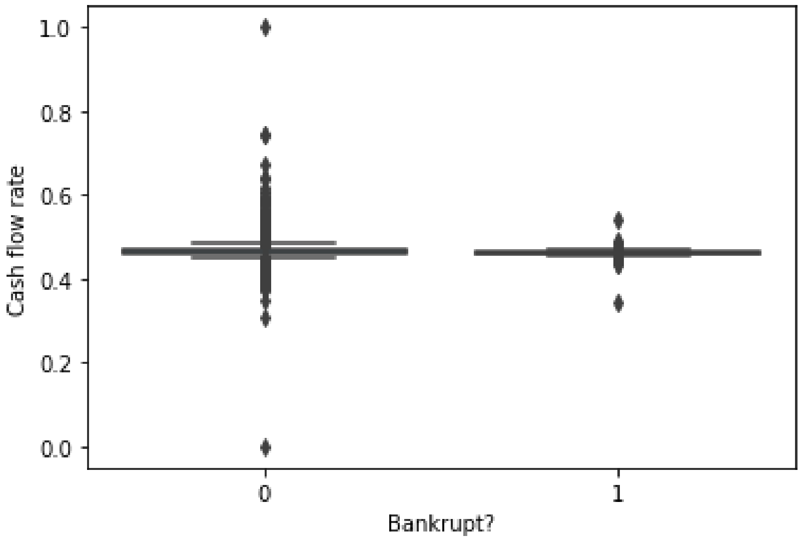
In the Exploratory Data Analysis phase, the first step was to look at the two class labels, the companies that went bankrupt and the companies that did not. The was a huge class imbalance because of the difference between the Yes and No class label. Out of the 6818 companies, 6599 were a No class label, i.e., did not go bankrupt and 220 were a Yes class label, i.e., did go bankrupt. The Yes class label constitutes just for ~ 3% of the overall data.

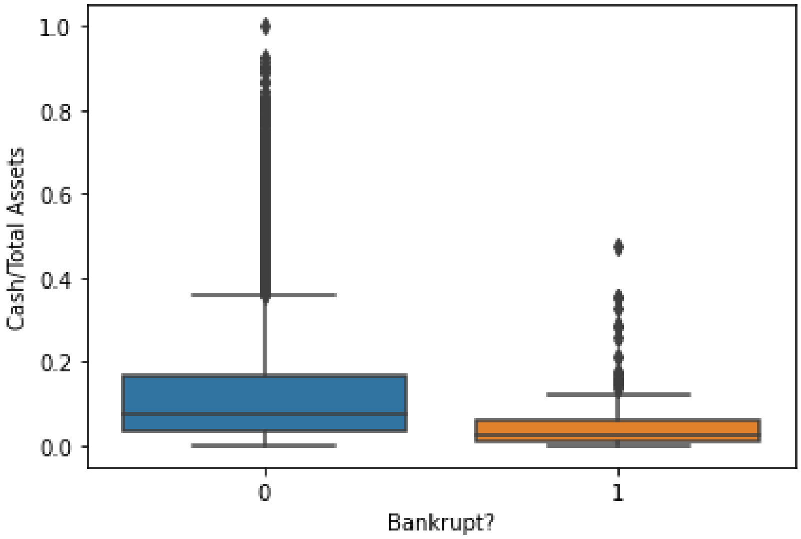


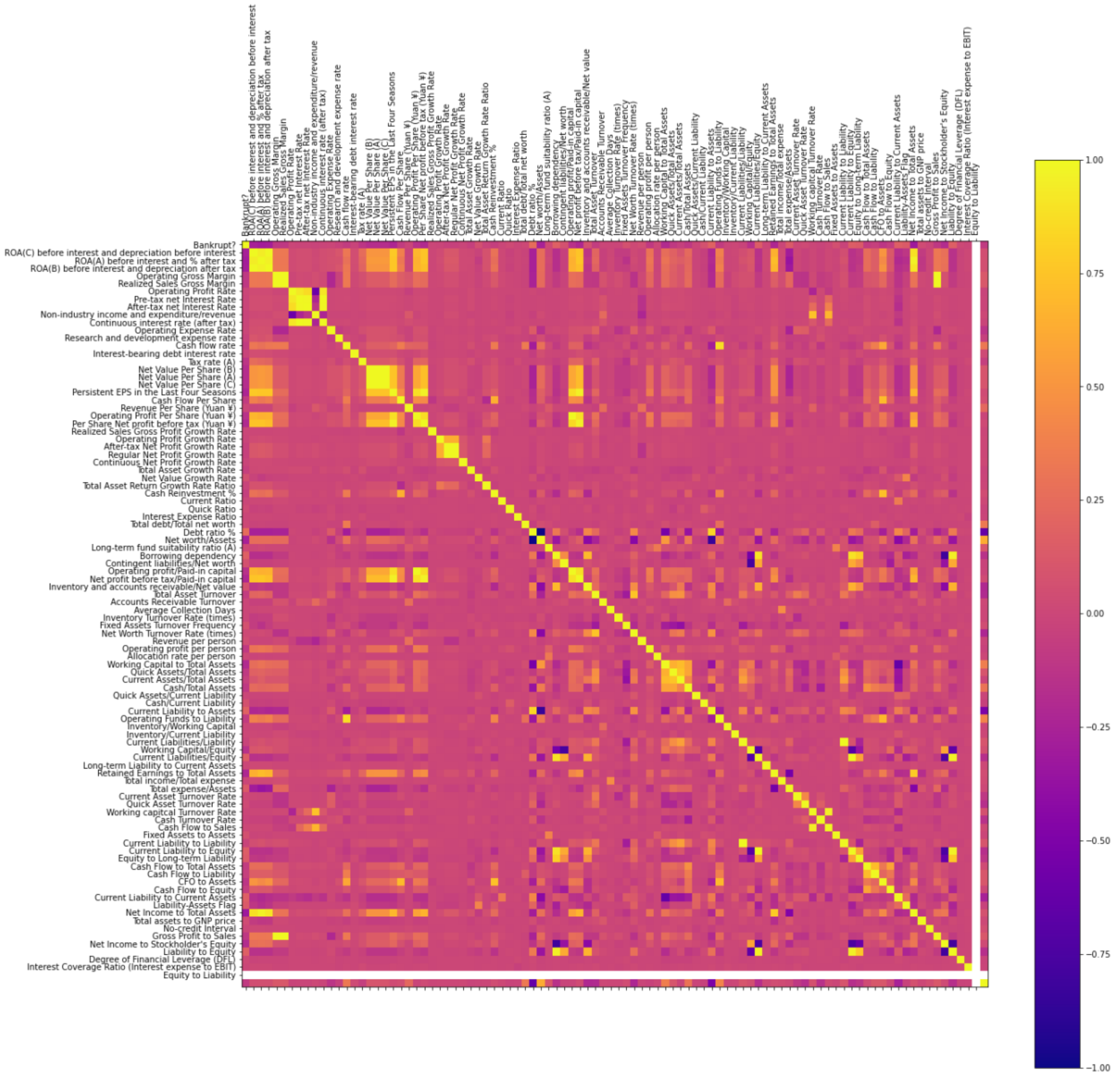
The next step was to explore the correlation between the features for which a heatmap was used. Then there was exploring the correlation between the different features and the class labels. For this box plots were used with a feature on the y axis and Yes/No class label on the x axis. The features did not seem to have a strong correlation among themselves but there were some differences observed between the features and the class label comparison.

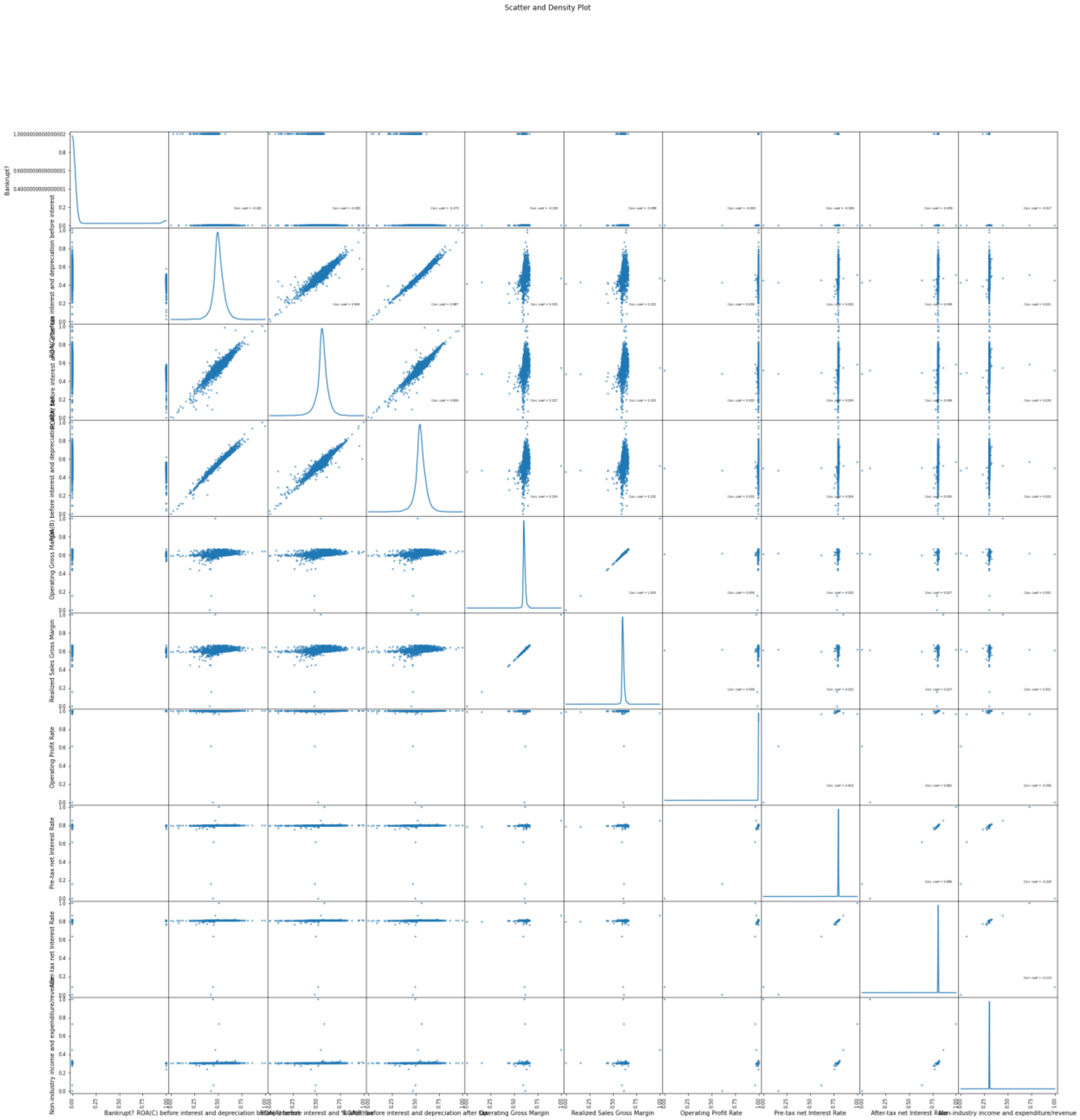



Operating Profit Rate
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Bankrupt?


Cash flow rate
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Bankrupt?


Cash/Total Assets
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Bankrupt?


Bankrupt?
ROA(C) before interest and depreciation before interest
ROA(A) before interest and % after tax
ROA(B) before interest and depreciation after tax
Operating Gross Margin
Realized Sales Gross Margin
Operating Profit Rate
Pre-tax net Interest Rate
After-tax net Interest Rate
Non-industry income and expenditure/revenue
Continuous interest rate (after tax)
Operating Expense Rate
Research and development expense rate
Cash flow rate
Interest-bearing debt interest rate
Tax rate (A)
Net Value Per Share (B)
Net Value Per Share (A)
Net Value Per Share (C)
Persistent EPS in the Last Four Seasons
Cash Flow Per Share
Revenue Per Share (Yuan ¥)
Operating Profit Per Share (Yuan ¥)
Per Share Net profit before tax (Yuan ¥)
Realized Sales Gross Profit Growth Rate
Operating Profit Growth Rate
After-tax Net Profit Growth Rate
Regular Net Profit Growth Rate
Continuous Net Profit Growth Rate
Total Asset Growth Rate
Net Value Growth Rate
Total Asset Return Growth Rate Ratio
Cash Reinvestment %
Current Ratio
Quick Ratio
Interest Expense Ratio
Total debt/Total net worth
Debt ratio %
Net worth/Assets
Long-term fund suitability ratio (A)
Borrowing dependency
Contingent liabilities/Net worth
Operating profit/Paid-in capital
Net profit before tax/Paid-in capital
Inventory and accounts receivable/Net value
Total Asset Turnover
Accounts Receivable Turnover
Average Collection Days
Inventory Turnover Rate (times)
Fixed Assets Turnover Frequency
Net Worth Turnover Rate (times)
Revenue per person
Operating profit per person
Allocation rate per person
Working Capital to Total Assets
Quick Assets/Total Assets
Current Assets/Total Assets
Cash/Total Assets
Quick Assets/Current Liability
Cash/Current Liability
Current Liability to Assets
Operating Funds to Liability
Inventory/Working Capital
Inventory/Current Liability
Current Liabilities/Liability
Working Capital/Equity
Current Liabilities/Equity
Long-term Liability to Current Assets
Retained Earnings to Total Assets
Total income/Total expense
Total expense/Assets
Current Asset Turnover Rate
Quick Asset Turnover Rate
Working capitcal Turnover Rate
Cash Turnover Rate
Cash Flow to Sales
Fixed Assets to Assets
Current Liability to Liability
Current Liability to Equity
Equity to Long-term Liability
Cash Flow to Total Assets
Cash Flow to Liability
CFO to Assets
Cash Flow to Equity
Current Liability to Current Assets
Liability-Assets Flag
Net Income to Total Assets
Total assets to GNP price
No-credit Interval
Gross Profit to Sales
Net Income to Stockholder's Equity
Liability to Equity
Degree of Financial Leverage (DFL)
Interest Coverage Ratio (Interest expense to EBIT)
Equity to Liability
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

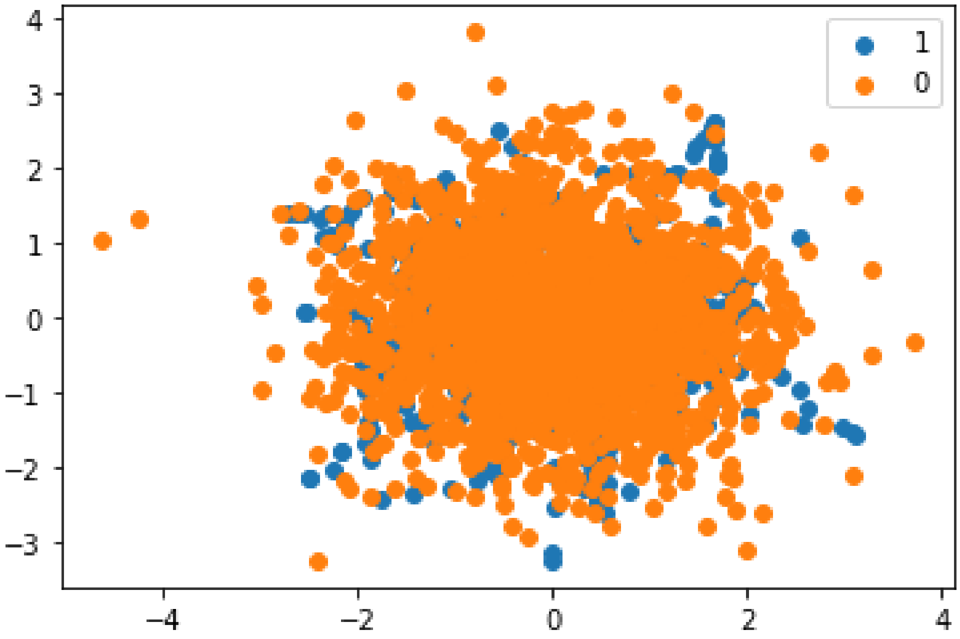


Preprocessing and Training Data

During the preprocessing and training data phase the data was split using an 80-20 split and models were applied to look at the initial prediction. Logistic regression, Gradient boosting, Decision trees, and XGboost as the initial models for prediction and used F1-score to predict the accuracy. It is important to use F1-score because when dealing with a class imbalance the overall accuracy could be rally high. In this case we can get an accuracy of 97% if we predict all class label as No. Therefore, it is important to look at both precision and recall.

As a result, F1 score for the No class label was 0.98 and F1-score for Yes class label was 0.32 for logistic regression. For the rest of the models the F1-score for the Yes class label ranged between 0.3 – 0.4.

The goal now was to increase the F1-score for the Yes class label. The reason for the low F1-score for the Yes class label was probably because of the class imbalance. To solve this is issue, I used SMOTE (Synthetic Minority Oversampling Technique) which increased the No class label in the data through oversampling.



Modelling

After applying SMOTE, different models were applied to the data to check the accuracy. The F1-score for the No class label increased significantly. All models showed a significant increase in the F1-score of No class label (between 0.95-0.97) The highest increase in the F1-score for No class label was seen in the XGboost model with a F1-score for No class label to be 0.97 and an overall accuracy of 0.98.

```
      precision    recall  f1-score   support

           0       0.99      0.98      0.99       387
           1       0.96      0.99      0.97       207

    accuracy                           0.98       594
   macro avg       0.98      0.98      0.98       594
weighted avg       0.98      0.98      0.98       594

[[379   8]
 [  3 204]]
```

In the Decision tree the main features used for the final prediction were Liability to Equity, Working capital turnover rate, Revenue per share, ROA(A) before interest and % after tax, Operating profit rate, Quick Assets/ Total Assets, Net worth, Average collection days, Interest-bearing debt interest rate, Cash flow to sales, and Debt ratio %.