CS 6375 Machine Learning - Final Project

Darsh Vaghasia DXV210035
Nand Patel NXP21002

8th May 2023

Contents

1 Project Topic

Compare different Machine Learning models for the Bankruptcy dataset from the Taiwan Economic Journal for the years 1999-2009, find out the best one. Also, find out which companies have a higher probability of declaring bankruptcy.

2 Motivation

Bankruptcy prediction is an important problem for modern economies because early warnings of bankrupt help not only the investor but also public policy makers to take proactive steps to minimize the impact of bankruptcies.

3 A Quick Overview

Firstly, performed preprocessing steps, checked the class distributions of the dataset, and performed the exploratory data analysis and tried to remove the outliers, then we compared different Machine Learning models (Logistic Regression, Naive Bayes, Decision Tree, KNN, SVM, AdaBoost, XGBoost, Neural Network and Random Forest) across various metrics (Accuracies, Cross Fold Validation, ROC Curves, Confusion Matrices, Precision, Recall, and F1 score).

4 About the Data Set(s)

4.a Description

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.

- **data.csv**

The original data set comprises of about 6820 rows of data spanned over 96 columns. Each data point includes the details of each company. We have filtered out unwanted noisy data (handled missing values, removed matches from other tournaments, etc.). The preprocessed data set contains 6819 data points and uses 8 features for building the models. The features and classes are discussed in the following section.

4.b Data Variety

- Across different industries (electronic manufacturing, retail, shipping, tourism...).
- Each industry has sufficient number of companies in the similar size in order to do the comparison.
- 95 features (X1-X95, business regulations of Taiwan Stock Exchange).
- 1 labels (bankrupt or not).

4.c Features

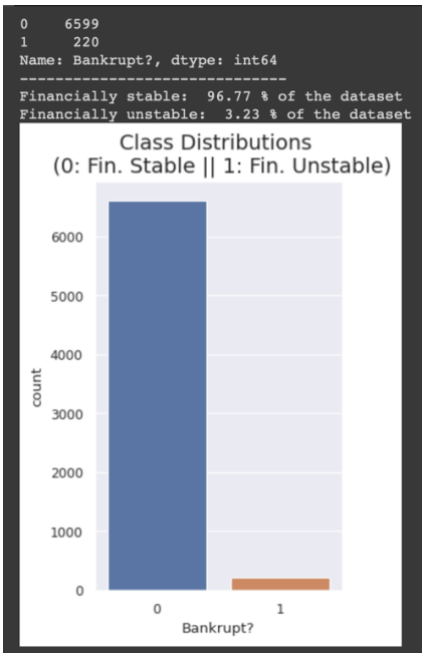
Y = Output Features, X= Input Features

Y - Bankrupt?: Class label
X1 - ROA(C) before interest and depreciation before interest: Return On Total Assets(C)
X2 - ROA(A) before interest and % after tax: Return On Total Assets(A)
X3 - ROA(B) before interest and depreciation after tax: Return On Total Assets(B)
X4 - Operating Gross Margin: Gross Profit/Net Sales
X5 - Realized Sales Gross Margin: Realized Gross Profit/Net Sales
X6 - Operating Profit Rate: Operating Income/Net Sales
X7 - Pre-tax net Interest Rate: Pre-Tax Income/Net Sales
X8 - After-tax net Interest Rate: Net Income/Net Sales
X9 - Non-industry income and expenditure/revenue: Net Non-operating Income Ratio
X10 - Continuous interest rate (after tax): Net Income-Exclude Disposal Gain or Loss/Net Sales
X11 - Operating Expense Rate: Operating Expenses/Net Sales
X12 - Research and development expense rate: (Research and Development Expenses)/Net Sales
X13 - Cash flow rate: Cash Flow from Operating/Current Liabilities
X14 - Interest-bearing debt interest rate: Interest-bearing Debt/Equity
X15 - Tax rate (A): Effective Tax Rate
X16 - Net Value Per Share (B): Book Value Per Share(B)
X17 - Net Value Per Share (A): Book Value Per Share(A)
X18 - Net Value Per Share (C): Book Value Per Share(C)
X19 - Persistent EPS in the Last Four Seasons: EPS-Net Income
X20 - Cash Flow Per Share
X21 - Revenue Per Share (Yuan ¥): Sales Per Share
X22 - Operating Profit Per Share (Yuan ¥): Operating Income Per Share
X23 - Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share
X24 - Realized Sales Gross Profit Growth Rate
X25 - Operating Profit Growth Rate: Operating Income Growth
X26 - After-tax Net Profit Growth Rate: Net Income Growth
X27 - Regular Net Profit Growth Rate: Continuing Operating Income after Tax Growth
X28 - Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
X29 - Total Asset Growth Rate: Total Asset Growth
X30 - Net Value Growth Rate: Total Equity Growth
X31 - Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
X32 - Cash Reinvestment %: Cash Reinvestment Ratio
X33 - Current Ratio
X34 - Quick Ratio: Acid Test
X35 - Interest Expense Ratio: Interest Expenses/Total Revenue
X36 - Total debt/Total net worth: Total Liability/Equity Ratio
X37 - Debt ratio %: Liability/Total Assets
X38 - Net worth/Assets: Equity/Total Assets
X39 - Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
X40 - Borrowing dependency: Cost of Interest-bearing Debt
X41 - Contingent liabilities/Net worth: Contingent

Liability/Equity
X42 - Operating profit/Paid-in capital: OperatingIncome/Capital
X43 - Net profit before tax/Paid-in capital: Pretax Income/Capital
X44 - Inventory and accounts receivable/Net value: (Inventory+Accounts Receivables)/Equity
X45 - Total Asset Turnover
X46 - Accounts Receivable Turnover
X47 - Average Collection Days: Days Receivable Outstanding
X48 - Inventory Turnover Rate (times)
X49 - Fixed Assets Turnover Frequency
X50 - Net Worth Turnover Rate (times): Equity Turnover
X51 - Revenue per person: Sales Per Employee
X52 - Operating profit per person: Operation Income Per Employee
X53 - Allocation rate per person: Fixed Assets Per Employee
X54 - Working Capital to Total Assets
X55 - Quick Assets/Total Assets
X56 - Current Assets/Total Assets
X57 - Cash/Total Assets
X58 - Quick Assets/Current Liability
X59 - Cash/Current Liability
X60 - Current Liability to Assets
X61 - Operating Funds to Liability
X62 - Inventory/Working Capital
X63 - Inventory/Current Liability
X64 - Current Liabilities/Liability
X65 - Working Capital/Equity
X66 - Current Liabilities/Equity
X67 - Long-term Liability to Current Assets
X68 - Retained Earnings to Total Assets
X69 - Total income/Total expense
X70 - Total expense/Assets
X71 - Current Asset Turnover Rate: Current Assets to Sales
X72 - Quick Asset Turnover Rate: Quick Assets to Sales
X73 - Working capitcal Turnover Rate: Working Capital to Sales
X74 - Cash Turnover Rate: Cash to Sales
X75 - Cash Flow to Sales
X76 - Fixed Assets to Assets
X77 - Current Liability to Liability
X78 - Current Liability to Equity
X79 - Equity to Long-term Liability
X80 - Cash Flow to Total Assets
X81 - Cash Flow to Liability
X82 - CFO to Assets
X83 - Cash Flow to Equity
X84 - Current Liability to Current Assets
X85 - Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
X86 - Net Income to Total Assets
X87 - Total assets to GNP price
X88 - No-credit Interval
X89 - Gross Profit to Sales
X90 - Net Income to Stockholder's Equity
X91 - Liability to Equity
X92 - Degree of Financial Leverage (DFL)
X93 - Interest Coverage Ratio (Interest expense to EBIT)
X94 - Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
X95 - Equity to Liability

4.d Exploratory Data Analysis

Checking Stability of the data



As we see the class label distributions in the data set, So the dataset is heavily skewed, and this issue can be solved by resampling the data.

Spearman Heatmap

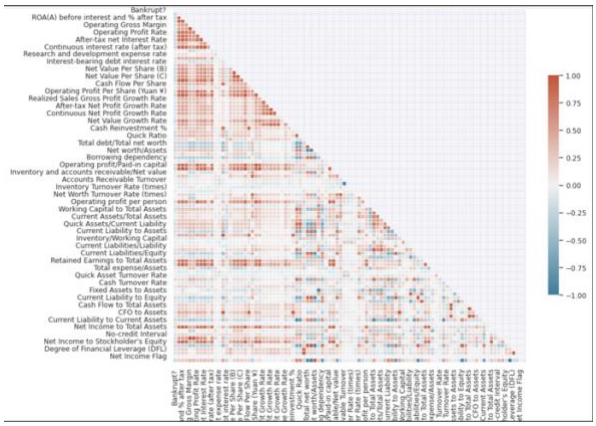
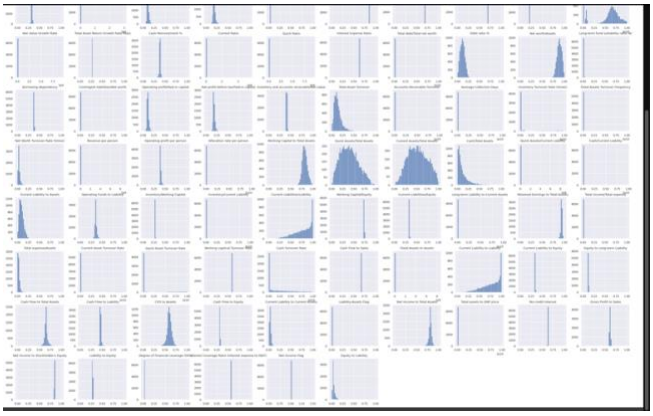


Figure 1: Spearman Heatmap

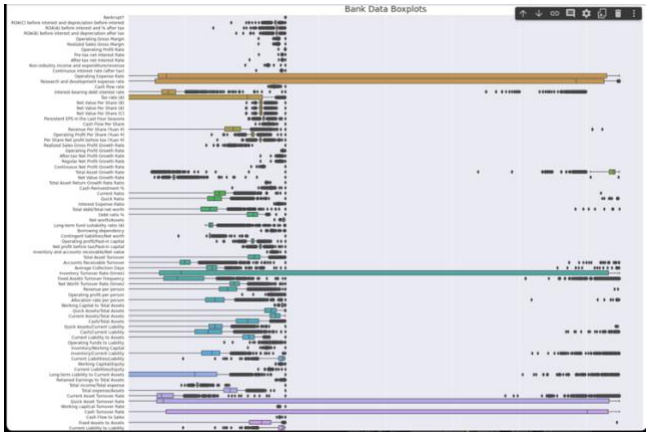
Using the spearman heatmap, calculated the spearman rank correlation which is a statistical measure to assess the strength and direction of the monotonic relationship between two features. The coefficient value between two columns range in the value between -1 and +1. We see some features are highly correlated to each other.

Histogram Data Analysis



Performed Histogram Frequency distribution observation to explore the distribution of values in each column, including the range, shape, and central tendency of the data.

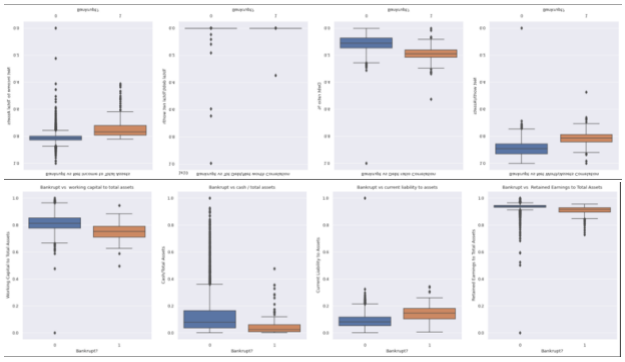
Box Plots



Boxplot Visualization for visualizing the distribution of data in a dataset, especially when comparing multiple numerical features. Besides observing the outliers, the data points that are beyond the range of 1.5 times the IQR of the box which is the middle line representing the median value. Thus identified the outliers in the data and compared the distribution of data across different features in the dataset.

4.e Plotting Primal Features

Plotted Some of the Interesting features which according to us might be the potential features which are correlated to each other and might be the key for predicting bankruptcy of the company. Some of the features are **"Net Income to Total Assets, Total debt/ Total net worth, Debt Ratio %, Net worth/ Assets, Working Capital to Total Assets, Cash/ Total Assets, Current Liability to Assets, Retained Earnings to Total Assets."**



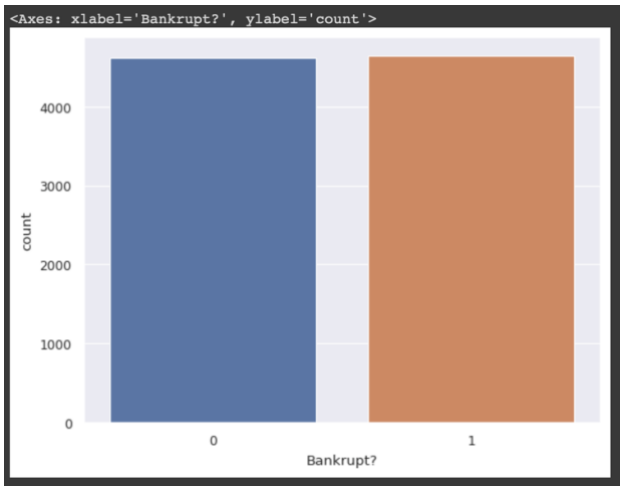
Later performed the task of removing the outliers from all the features.

4.f Target Class

We want to predict whether the company is on the verge of declaring bankruptcy or not in advance which will help the companies to take precautionary measures well in advance.

4.g Oversampling

We performed Oversampling over the minority class since the classes in the dataset were highly skewed hence making the dataset unbalanced. For oversampling we used Adaptive Oversampling technique which is a variant of SMOTE technique. It basically generates synthetic samples based on the density distribution of the minority class.



Distribution of classes after performing oversampling using ADASYN

5 Comparison between Algorithms (scikit-learn)

5.a Algorithms used.

- Logistic Regression
- Decision Tree
- K Nearest Neighbor
- Support Vector Machines (Linear Kernel, RBF Kernel)
- Neural Network
- Random Forest
- AdaBoost
- XGBoost

5.b Train / Test Accuracies

Computing Train/Test accuracies based on the train and test errors for the above models, we can see the following graph,

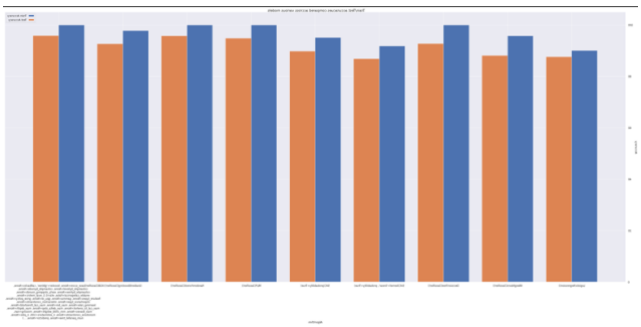


Figure 2: Train / Test Accuracies compared across various models.

From the above figure, it is evident that XGBoost, Random Forest, Decision Tree perform equally well on the train set but XGBoost performs the best among all models in test set with an accuracy of approximately 96.6%. Multiple Layer Perceptron (Neural Network) have high accuracy on train set but relatively less accuracy on Test set to XGBoost & Decision Trees. Decision Trees appears to have high accuracy on the train set but fails to accurately capture the predictions on the test set in comparison to XGBoost. Logistic Regression have relatively less accuracies.

5.c 10-Fold Cross Validation

We are running k-fold cross validation (k=10) and checking the average performance of the models by dividing the data into 10 different blocks and iteratively running the models on 10 combinations of train/test (9:1) sets.

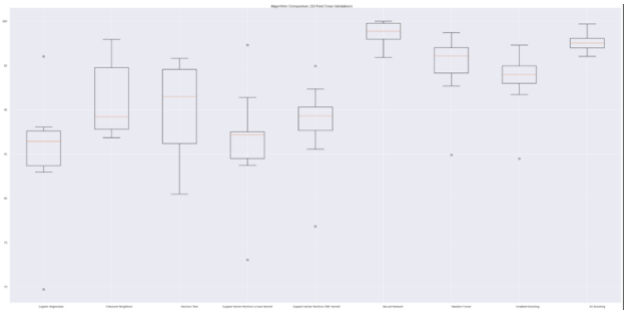


Figure 3: Algorithm comparison over 10-fold cross validation

We can see that Neural Net and XGBoost perform well with cross validation with the median accuracy being roughly around 98% and 97% respectively. Logistic Regression and SVM produce average results over cross validation. Random Forest and AdaBoost perform second best performance over all other models.

5.d ROC/AUC Curves

Generating ROC curves for the following algorithms, we get the below plot.

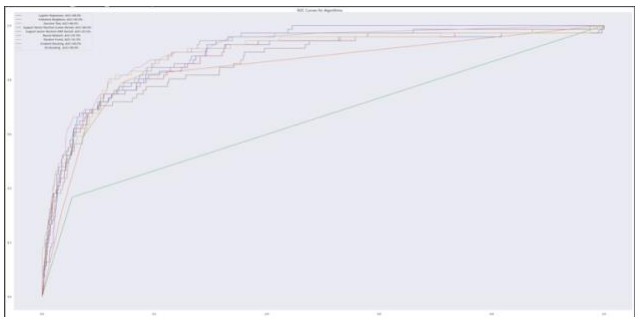
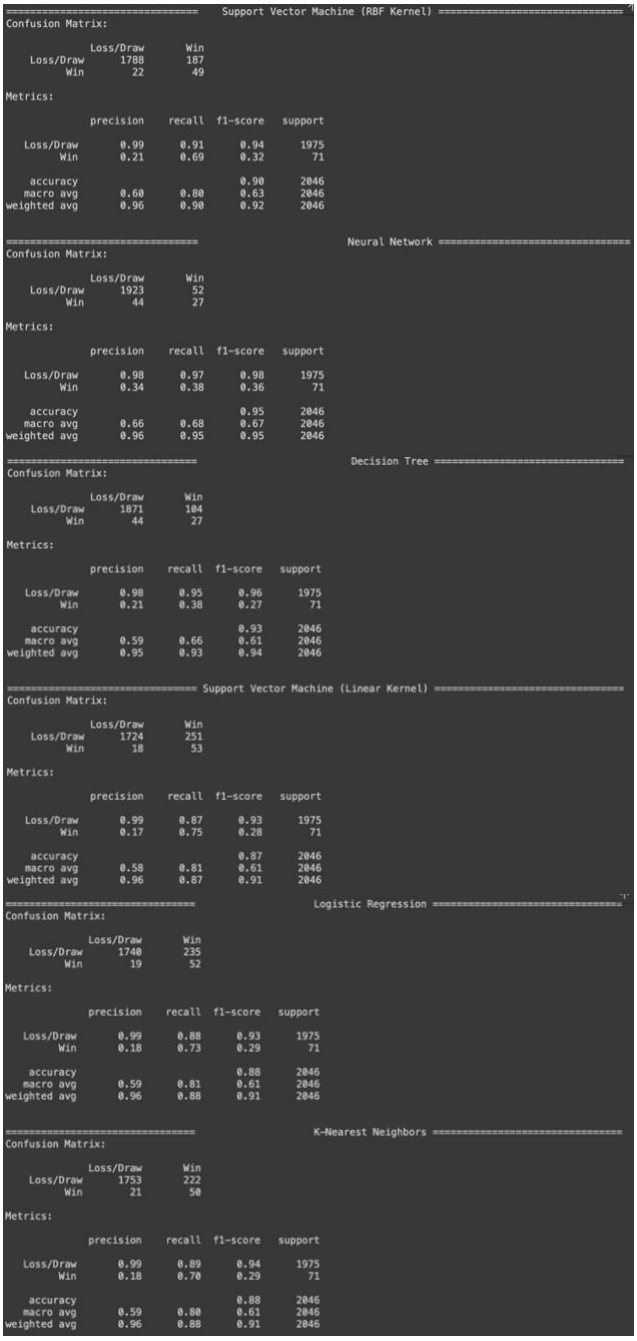


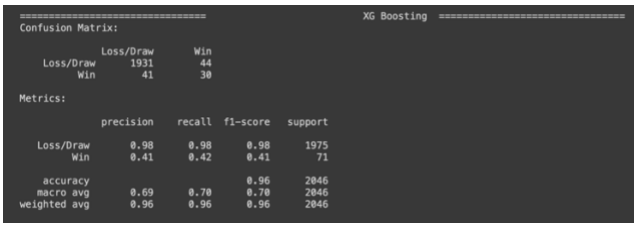
Figure 4: Comparison of ROC curves for different algorithms

We can see that Support Vector Machine (RBF Kernel) and Random Forest has the highest Area Under the Curve (AUC). Decision Trees has the least Area Under the Curve.

5.e Confusion Matrix, F1, Precision and Recall

```
================================ Random Forest ================================
Confusion Matrix:

            Loss/Draw      Win
Loss/Draw        1919       56
Win                39       32

Metrics:

              precision    recall  f1-score   support

   Loss/Draw       0.98      0.97      0.98      1975
         Win       0.36      0.45      0.40        71

    accuracy                           0.95      2046
   macro avg       0.67      0.71      0.69      2046
weighted avg       0.96      0.95      0.96      2046

============================== Gradient Boosting ==============================
Confusion Matrix:

            Loss/Draw      Win
Loss/Draw        1855      120
Win                30       41

Metrics:

              precision    recall  f1-score   support

   Loss/Draw       0.98      0.94      0.96      1975
         Win       0.25      0.58      0.35        71

    accuracy                           0.93      2046
   macro avg       0.62      0.76      0.66      2046
weighted avg       0.96      0.93      0.94      2046
```





6 Results

After comparing the above algorithms, we can clearly see that XGBoost, Neural Network, Random Forest perform better than almost all the other algorithms. We expected them to have more accuracy over the others. Therefore, XGBoost performs the best among all, and we consider it as a base learning algorithm.

7 Conclusion

Thus, we performed Exploratory Data Analysis and found that the dataset is unbalanced, Performed ADASYNC Oversampling techniques to generate synthetic data of the minority class with the goal to make the dataset balanced first. After that, we have compared different Machine Learning Models over various metrics. After assessing the performance of all the models using various metrics such as F-1, Precision, Recall, Confusion Matrix and 10-fold Cross Validation we found XGBoost, Random Forest and Neural Network as the best performing algorithms with the accuracy range of 95-96%.

8 What did we learn?

Making predictions solely based on historical data has its limitations. There are a lot of other factors (including human intervention) that determine the outcome of real-life events.

9 Scope of Improvement

- Real-time Prediction: We can consider implementing a real-time prediction system, which can provide timely and accurate predictions of bankruptcy risk for individual companies based on their financial data. This can be useful for financial institutions to make more informed lending decisions.

- Ensemble Methods: Ensemble methods combine multiple machine learning models to improve the overall predictive power. We can consider implementing ensemble methods such as stacking or blending, which can combine the predictions of different models to generate a more accurate and reliable final prediction.

10 Discussions

- Consideration of Class Imbalance: Bankruptcy prediction is often characterized by class imbalance, with a relatively small number of bankruptcy cases compared to non-bankruptcy cases. It is important to consider techniques such as oversampling, under sampling, or cost-sensitive learning to address this imbalance and improve the accuracy of the models.

- Need for Domain Expertise: Bankruptcy prediction requires a deep understanding of financial data and industry-specific factors. It is important to work with domain experts to identify the most relevant financial indicators and ensure that the models are effectively capturing the nuances of the problem.

- Overall, Implementation of multiple machine learning models and identifying best performing models is a positive step in developing an effective bankruptcy prediction. Further improvements can be made by considering the complexity of the models.