Company Bankruptcy Prediction

Sanjeev Hegde
Data science Trainee, AlmaBetter

Abstract: Predicting the possibility of bankruptcy of a company is of paramount importance to businesses such as investment banks, banks assessing credit risk while understanding the overall picture of the market segment and economy. However, predicting whether a company could go bankrupt depends on several factors such revenue, debt owned by the company, offerings by the competitors, company management, external factors such as recession or natural calamities, fraud etc. In this regard, various machine learning models are developed based on the company bankruptcy data gathered from Taiwan Economic Journal during 1999 to 2009 consisting of various financial ratios. Supervised machine learning classifications algorithms logistic regression, Bernoulli-Naive Bayes classifier, Support Vector Machine Classifier, XGBoost Classifier have been utilized to make predictions of bankruptcy on companies and it has been found that Support Vector Machine Classifier(SVM) performed the best with 0.9994 recall and 0.9847 ROC AUC score.

Key words: Bankruptcy, Investment Banking, Credit Risk, economy, supervised machine learning, Classification algorithms, Taiwan Economic Journal, financial ratios, SMOTE, logistic regression, Bernoulli-Naive Bayes classifier, Support Vector Machine Classifier(SVM), Extreme Gradient Boosting Classifier(XGB aka XGBoost), recall, ROC AUC score

Problem statement: Company bankruptcy prediction using data collected from Taiwan Economic Journal from 1999 to 2009.

Introduction: Taiwan Economic Journal is the largest financial information company in Taiwan. It was founded in the year 1990 and provides various fundamental financial data of companies in Taiwan. The data set used to build the model is based on the data collected from the Taiwan Economic Journal during 1999 to 2009 consisting of various financial ratios and whether the company is bankrupt or not. A company in the data set is classified as bankrupt as per the definitions of bankruptcy based business regulations of Taiwan Stock Exchange. It is to be noted that the data collected is during the period of dot com bubble crash (end of 20th century) in US and also during the housing market collapse in US which led to recession worldwide during 2008. Thus, it is essential to understand what factors led companies to bankruptcy and how the rest survived. Based on the collected data various machine learning models can be built.and can be used to explain the reasons for companies going bankrupt during the time period which could be company specific or due to external factors thereby explaining the reasons which led to fall of companies. In this regard, let us understand the features in the data set, process used to build the machine learning models and their outcomes, evaluation metrics used and chosen to explain the superiority of a machine learning model over other models in the upcoming sections.

1. *Feature description:* The data set contains 96 features including the dependent feature. Their details are as below:

- Bankrupt?: Class label 1 : Yes , O: No
- ROA(C) before interest and depreciation before interest: Return On Total Assets(C)
- ROA(A) before interest and % after tax: Return On Total Assets(A)
- ROA(B) before interest and depreciation after tax: Return On Total Assets(B)
- Operating Gross Margin: Gross Profit/Net Sales
- Realized Sales Gross Margin: Realized Gross Profit/Net Sales
- Operating Profit Rate: Operating Income/Net Sales
- Pre-tax net Interest Rate: Pre-Tax Income/Net Sales
- After-tax net Interest Rate: Net Income/Net Sales
- Non-industry income and expenditure/revenue: Net Non-operating Income Ratio
- Continuous interest rate (after tax): Net Income-Exclude Disposal Gain or Loss/Net Sales
- Operating Expense Rate: Operating Expenses/Net Sales
- Research and development expense rate: (Research and Development Expenses)/Net Sales
- Cash flow rate: Cash Flow from Operating/Current Liabilities
- Interest-bearing debt interest rate: Interest-bearing Debt/Equity
- Tax rate (A): Effective Tax Rate
- Net Value Per Share (B): Book Value Per Share(B)
- Net Value Per Share (A): Book Value Per Share(A)
- Net Value Per Share (C): Book Value Per Share(C)
- Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share
- Realized Sales Gross Profit Growth Rate
- Operating Profit Growth Rate: Operating Income Growth
- After-tax Net Profit Growth Rate: Net Income Growth
- Regular Net Profit Growth Rate: Continuing Operating Income after Tax Growth
- Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
- Total Asset Growth Rate: Total Asset Growth
- Net Value Growth Rate: Total Equity Growth
- Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
- Cash Reinvestment %: Cash Reinvestment Ratio
- Current Ratio
- Quick Ratio: Acid Test
- Interest Expense Ratio: Interest Expenses/Total Revenue
- Total debt/Total net worth: Total Liability/Equity Ratio
- Debt ratio %: Liability/Total Assets
- Net worth/Assets: Equity/Total Assets
- Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
- Borrowing dependency: Cost of Interest-bearing Debt
- Contingent liabilities/Net worth: Contingent Liability/Equity
- Operating profit/Paid-in capital: Operating Income/Capital
- Net profit before tax/Paid-in capital: Pretax Income/Capital
- Inventory and accounts receivable/Net value: (Inventory+Accounts Receivables)/Equity
- Total Asset Turnover
- Accounts Receivable Turnover
- Average Collection Days: Days Receivable Outstanding

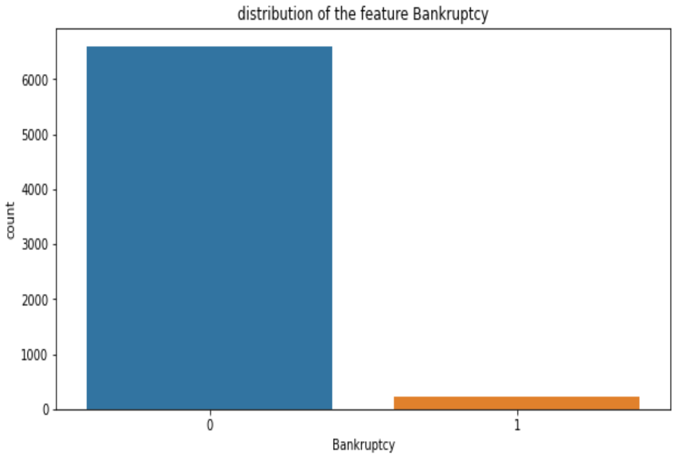
- Persistent EPS in the Last Four Seasons: EPS-Net Income
- Cash Flow Per Share
- Revenue Per Share (Yuan ¥): Sales Per Share
- Operating Profit Per Share (Yuan ¥): Operating Income Per Share
- Inventory Turnover Rate (times)
- Fixed Assets Turnover Frequency
- Net Worth Turnover Rate (times): Equity Turnover
- Revenue per person: Sales Per Employee
- Operating profit per person: Operation Income Per Employee
- Allocation rate per person: Fixed Assets Per Employee
- 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: Current Assets to Sales
- Quick Asset Turnover Rate: Quick Assets to Sales
- Working capital Turnover Rate: Working Capital to Sales
- Cash Turnover Rate: Cash to Sales
- 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: 1 if Total Liability exceeds Total Assets, 0 otherwise
- Net Income to Total Assets
- Total assets to GNP price
- No-credit Interval
- Gross Profit to Sales
- Net Income to Stockholders' Equity
- Liability to Equity
- Degree of Financial Leverage (DFL)
- Interest Coverage Ratio (Interest expense to EBIT)
- Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
- Equity to Liability

2. ***Data Wrangling:*** Data wrangling is the process of cleaning, transforming and organizing raw data in order to make it appropriate for visualization, analytics and machine learning. In this regard, following points can be made about the considered data set:

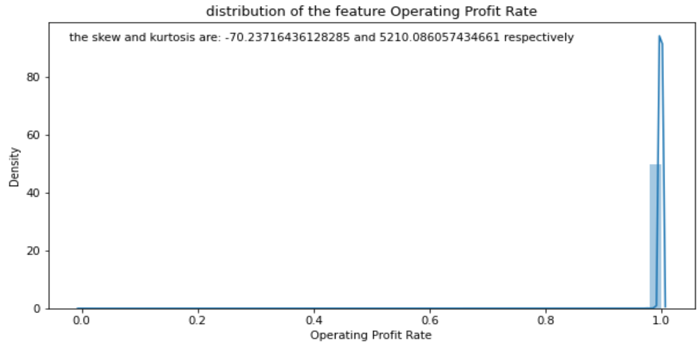
- Consists of 6819 entries against 96 features inclusive of the dependent variable.
- There are no duplicate entries in the data set.
- There are no null values in the data set.
- There are 3 categorical variables. They are 'Bankruptcy?', 'Liability-Asset Flag' and 'Net Income Flag'.
- There are extra whitespaces and non essential characters such as "?" in the name of the features. These have been stripped.

3. ***Data Visualization:*** Data visualization is essentially the best way to represent complex data in the easiest possible ways. In this regard, following visualizations have been made and results are as summarized below:

- **Univariate Analysis:** Histograms and bar plots have been plotted for all the features. Categorical features were plotted using bar plots and other features using histograms. Plots essential to develop the following conclusions have been shown below.

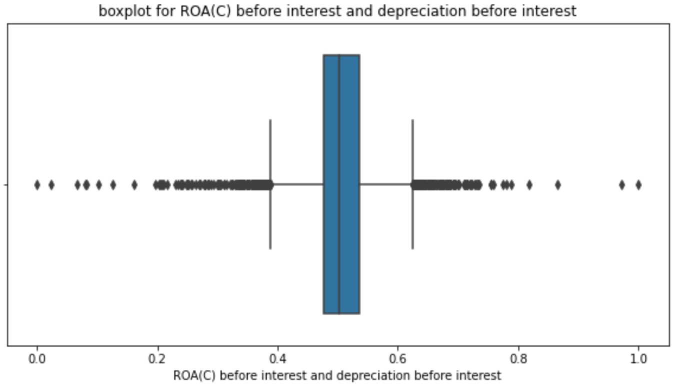


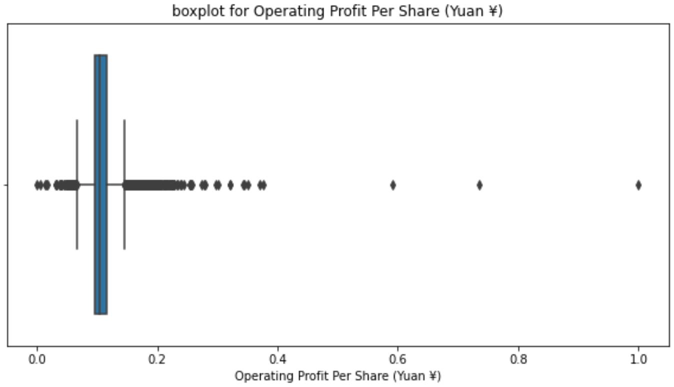
It is clear that the data set is highly imbalanced from the above bar plot. There are a few hundreds of companies that have filed bankruptcy among the 6819 companies.

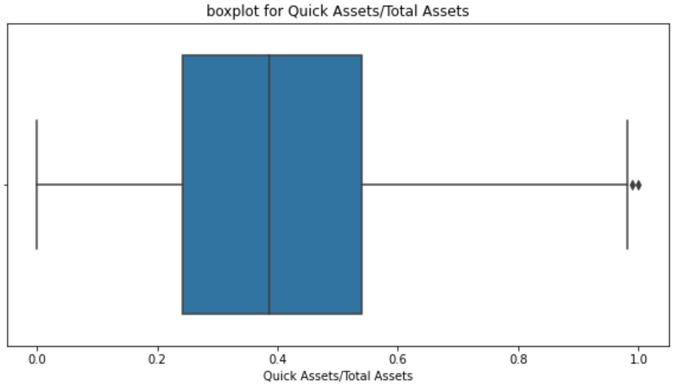


Also, it is to be noted that most features are either positively or negatively skewed, exhibiting leptokurtic behavior as represented above for Operating Profit Rate feature.

Further, Boxplots have been plotted to understand the distribution of the features. Examples sufficient to summarize the results have been shown below.

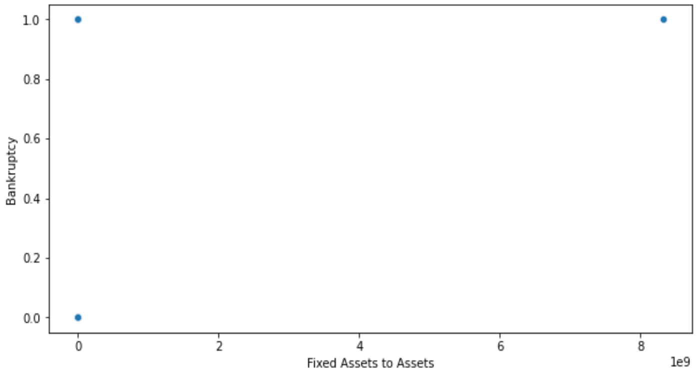


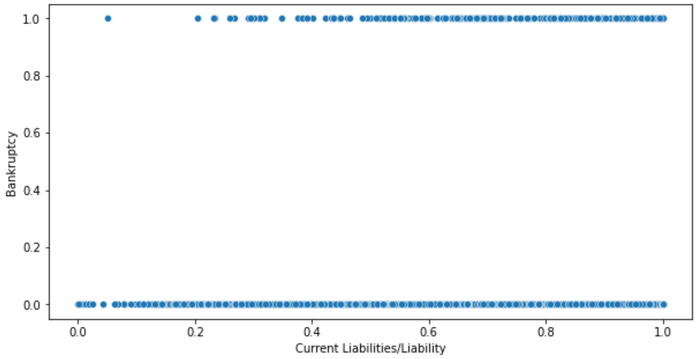


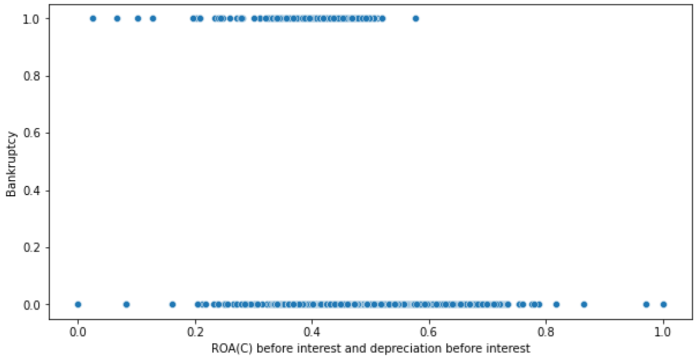


It is to be noted that most features have considerable presence of outliers in them. It is expected to be this way since different companies can have a range of corresponding values for these features depending on the valuation of the company and relative position of the company in the market it operates.

- **Bivariate Analysis:** Scatter plots of all the features against the dependent feature Bankruptcy have been plotted. Following samples of the plots of features can clearly make it evident that most features are continuous while only 2 features are categorical.

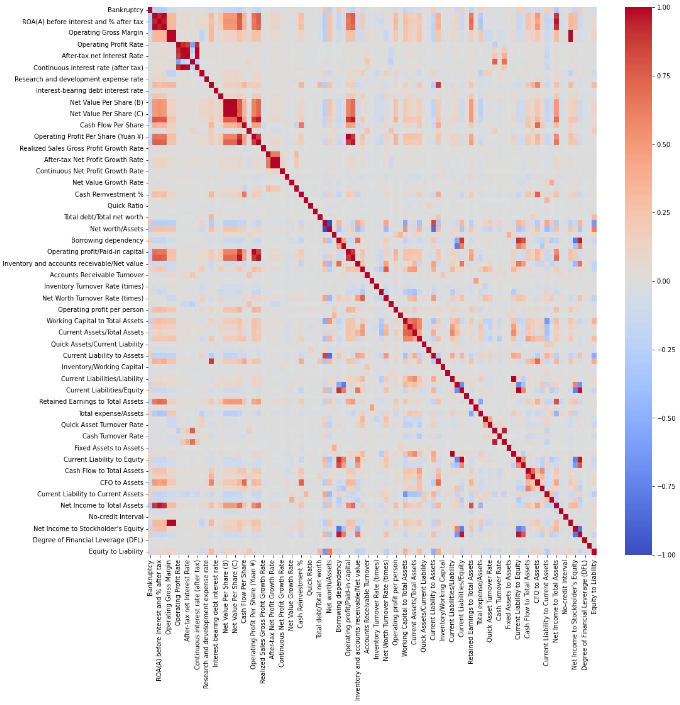






Moreover, as mentioned earlier it is clearly evident that only two features are categorical in nature apart from the dependent variable i.e., ‘Bankruptcy’.

- **Multivariate Analysis:** Correlation heatmaps are one of the most convenient methods to analyze the relationship between multi variables. In this regard, correlation heatmap has been plotted and following points are summarized.



➢ ROA(C) before interest and depreciation before interest, ROA(A) before interest and % after tax and ROA(B) before interest and depreciation after tax features can be summed up and be considered as one feature: Return on Assets since they exhibit high correlation.

➢ Net Value Per Share (B), Net Value Per Share (A), Net Value Per Share (C) features exhibit high correlation with each other. Hence, we can calculate the average of these values and assign it as the new variable.

➢ Current Liability to Liability and Current Liabilities/Liability are the same. Similarly, Current Liability to Equity and Current Liabilities/Equity are the same. Hence, it is appropriate to drop one of these.

➢ It is to be noted there are other features that exhibit higher correlation up to certain extent. However, since they are mostly related to taxation and no further information has been given in the data set regarding taxation, it is good to keep these features unaltered.

4. ***Feature Engineering:*** Following changes have been brought into the data set to make it ready for training on machine learning algorithms.

- **Encoding:** One Hot Encoding is used to encode the feature 'Liabilty-Asset FLag'. Its value is found to be 1 if total Liability exceeds Total Assets, 0 otherwise.

- Feature 'Net Income Flag' has only one value of 1 in the data set and it is thus dropped..

- **Feature Manipulation:** Following steps have been taken in order to ensure reduced multicollinearity:

➢Returns on Assets A, B and C have been summed to reduce the multicollinearity thereby creating a new feature Return on Assets.

➢Net Value Per Share of A,B and C have been averaged to drop the multicollinearity further thereby forming new feature 'avg_Net Value Per Share' feature

➢It is to be noted that after creation of new features, parent features have been dropped from the data set.

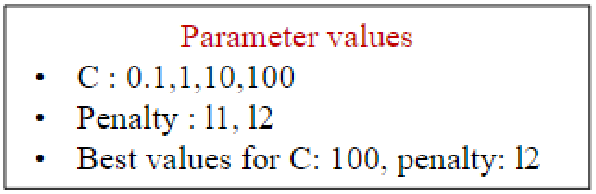
➢'Current Liability to Liability', 'Current Liability to Equity' features have been dropped as they are repetitive

- **Declaration of variables:** 'Bankruptcy' feature is declared as the dependent variable while the rest of the variables are declared as independent variables.

5. ***Data set balancing:*** The data set is balanced using Synthetic Minority Oversampling Technique(SMOTE) which essentially uses KNN to create synthetic data points for undersampled category i.e., Bankruptcy values equal to 1 instance in case of the given data set. It is to be noted that there were 6599 instances for non bankrupt companies in the data set while bankrupt instances were only 220. As a result of fitting smote on to the data set, we are now presented with 6599 values of both non bankrupt and bankrupt instances in the data set.

6. ***Data Splitting and Scaling:*** Scaling the data before fitting any kind of machine learning models is essential to mitigate the effect of different range of values in the features. With regard to the same, data is initially split into training and testing data with training data size being 0.7 times over all data using train_test_split from sklearn. MinMaxScaler is used in this project to achieve the objective of scaling.

7. ***Logistic Regression:*** A logistic regression model with 5 fold cross validation was developed based on the training data set. The list of parameters used during cross validation are as below:

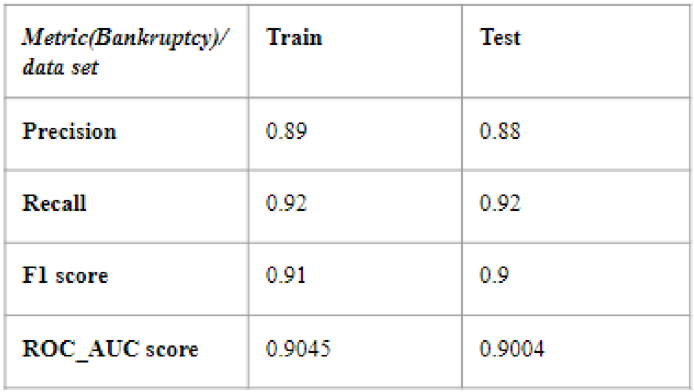


Parameter values

- C : 0.1,1,10,100
- Penalty : l1, l2
- Best values for C: 100, penalty: l2

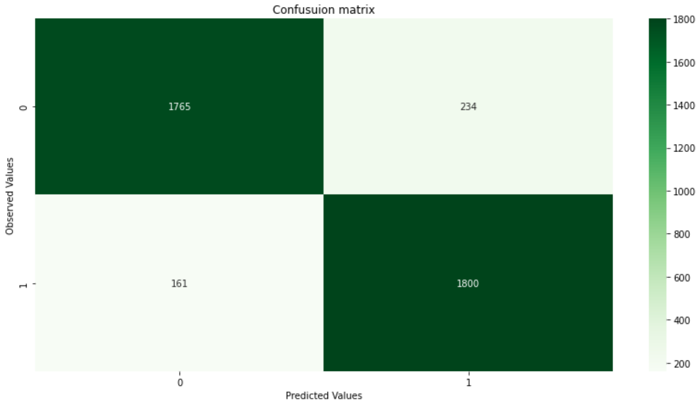
Further, precision, recall, F1 score and ROC AUC score have been used as the evaluation metric as shown in the next page.

The evaluation metric scores for logistic regression model are as below:

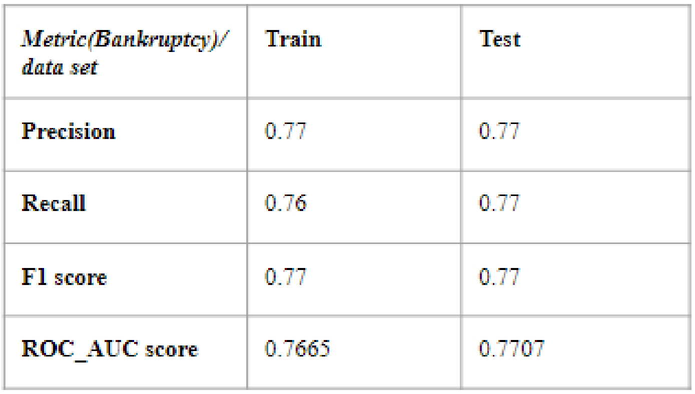


Metric(Bankruptcy)/ data set	Train	Test
Precision	0.89	0.88
Recall	0.92	0.92
F1 score	0.91	0.9
ROC_AUC score	0.9045	0.9004

The Confusion matrix for logistic regression model test data is shown below:

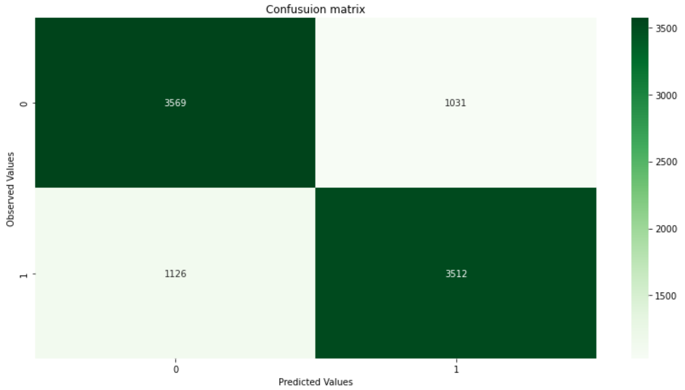


8. ***Bernoulli-Naive Bayes Classifier:*** Since the data set involves binomial classification Bernoulli-Naive Bayes classifier is used. The evaluation metric scores are shown below:



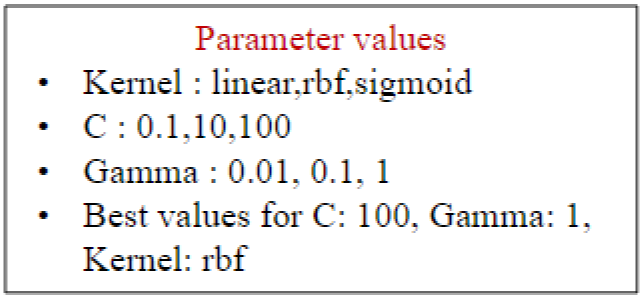
Metric(Bankruptcy)/ data set	Train	Test
Precision	0.77	0.77
Recall	0.76	0.77
F1 score	0.77	0.77
ROC_AUC score	0.7665	0.7707

Further, the confusion matrix for test data of Bernoulli-Naive Bayes is plotted below:



Thus, it is evident that results of Bernoulli-Naive Bayes model are not satisfactory with respect to the data set.

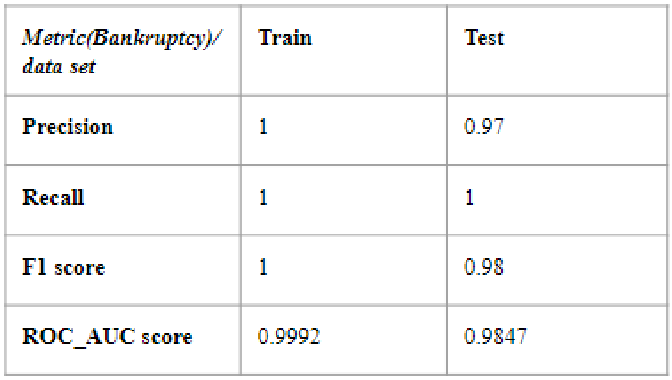
9. ***Support Vector Machine Classifier:*** Support Vector Machine Classifier was trained with 5 fold cross validation. The parameters used and the best parameters during cross validation are summarized below:



Parameter values

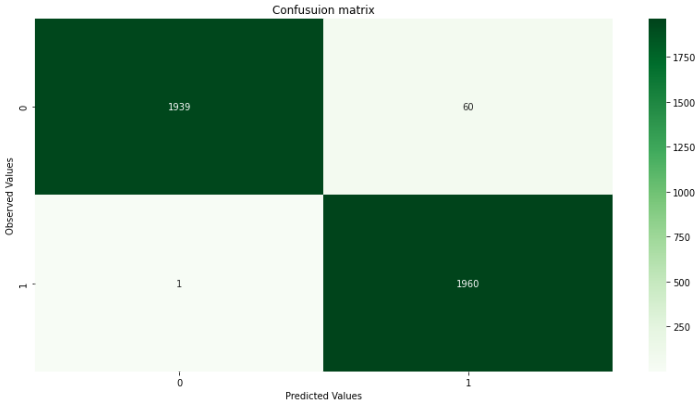
- Kernel : linear,rbf,sigmoid
- C : 0.1,10,100
- Gamma : 0.01, 0.1, 1
- Best values for C: 100, Gamma: 1, Kernel: rbf

Further, the result of evaluation metric scores are summarized below:



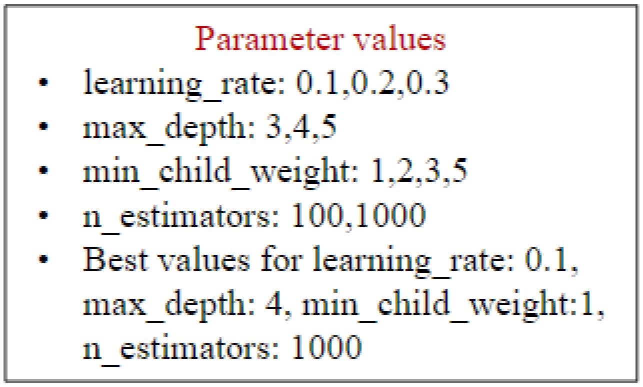
Metric(Bankruptcy)/ data set	Train	Test
Precision	1	0.97
Recall	1	1
F1 score	1	0.98
ROC_AUC score	0.9992	0.9847

The Confusion matrix of SVM Classifier for test data is shown below:



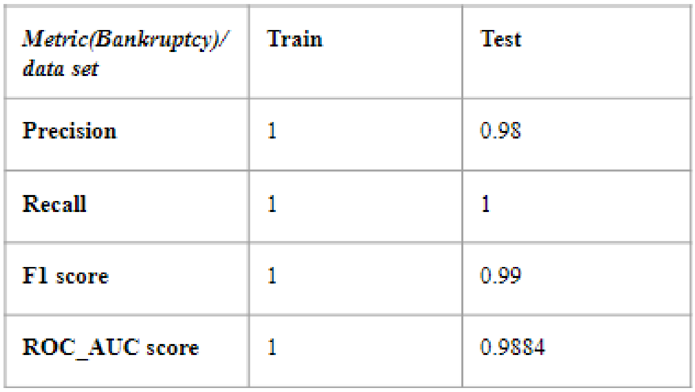
Hence, it is evident that there is just one false negative.

10. *XGBoost Classifier:* XGBoost classifier was trained with 5 fold cross validation. Following are the list of parameters used and best parameters:


Parameter values

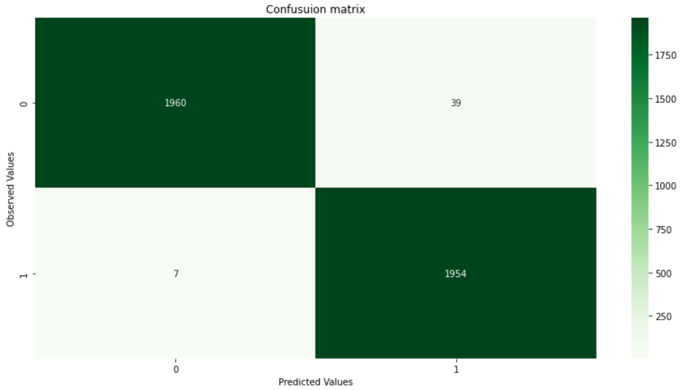
- learning_rate: 0.1,0.2,0.3
- max_depth: 3,4,5
- min_child_weight: 1,2,3,5
- n_estimators: 100,1000
- Best values for learning_rate: 0.1, max_depth: 4, min_child_weight:1, n_estimators: 1000

Further, evaluation metric scores for XGBoost classifier are as below:



Metric(Bankruptcy)/ data set	**Train**	**Test**
Precision	1	0.98
Recall	1	1
F1 score	1	0.99
ROC_AUC score	1	0.9884

The Confusion matrix of XGBoost classifier for test data is as shown below:

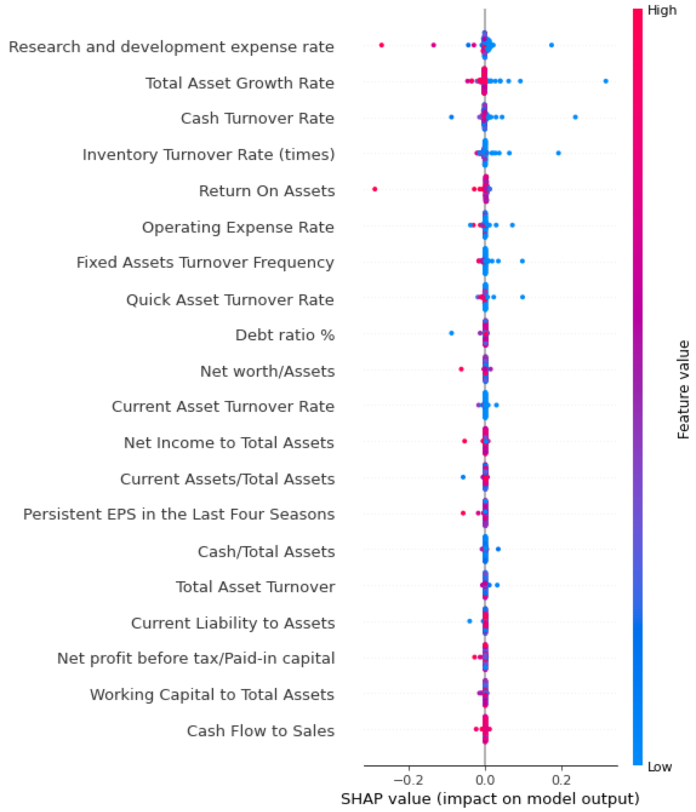


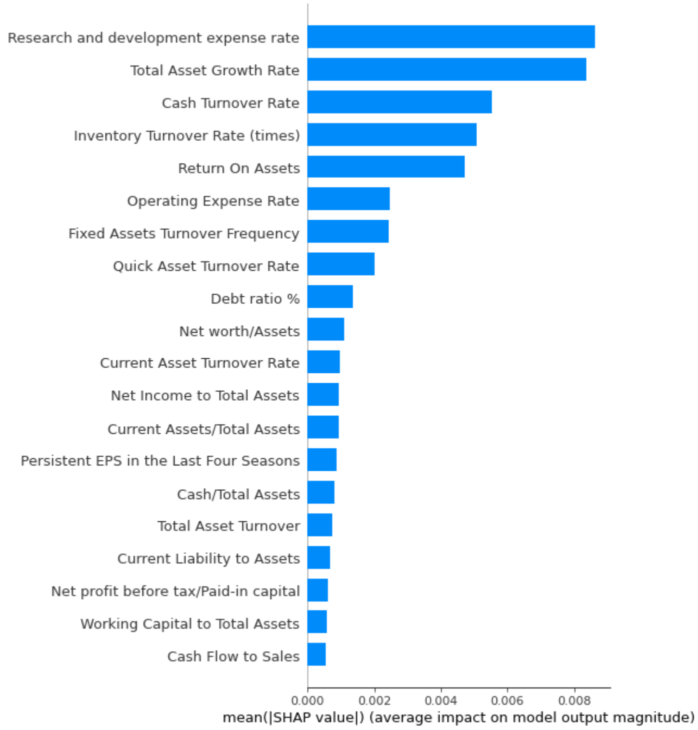
Hence, we can see that there are 7 false negatives.

11. *Choice of Model and Evaluation metric:* Predicting whether a company could go bankrupt is a critical question to ask for any investment bank or credit lending business. In this regard, there is very little room for error as bad choice of company implies blowing up the capital. Thus, we can conclude that recall score is the appropriate metric as it is the ratio of true positives to the sum of true positives and false negatives. Thus, on calculating the recall score, it was found that SVM classifier had the best recall score of 0.9994 while XGBoost Classifier had slightly less recall score of 0.9964. Hence, it can be concluded that the Support Vector Machine(SVM) model is the best performing model.

12. *Model Explainability using SHAP:* SHapley Additive exPLanation is a method used in game theory concept of Shapley values to explain the importance of features. Using the shap library, important features were calculated for the SVM classifier model considering 150 samples.

Important features as per shap values are shown below:





Thus, top 5 important features are:

- Research and development rate
- Total Asset Growth Rate
- Cash Turnover Rate
- Inventory Turnover Rate (times)
- Return On Assets

12. ***Conclusion:*** The following points can be concluded regarding company bankruptcy prediction based on Taiwan Economic Journal data set:

- Most features in the data set are leptokurtic, some being positively skewed and others being negatively skewed
- Most features have considerable amount of outliers
- Only two features apart from dependent variable are categorical
- Multicollinearity in the data set is reduced by appropriate aggregate combination of features, repetitive features were dropped keeping only single instances.
- One hot encoding is used to encode categorical features
- SMOTE is used to balance the data set
- MinMaxScaler is used to scale the data post splitting the data with training size of 0.7
- Logistic regression, Support Vector Machines classifier and XGBoost Classifier were fitted with 5 fold cross validation and used to make predictions.
- Bernoulli-Naive Bayes classifier was also fitted on the training data set. However it performed poorly compared to other models.
- Recall score is considered to be the most important evaluation metric and SVM Classifier is the best model with best recall score of 0.9994.
- Shap values were considered to explain the feature importance and following were top 5 factors causing bankruptcy: Lower value of Research and Development Expense Rate,

Lower Total Assets growth rate, Lower Cash Turnover rate, Lower inventory turnover rate.and Low returns on assets.

13. ***Scope for future work:*** There is still a decent amount of improvement that can be done with better data and models. They include:

- Better models can be built if there is an extra categorical feature pointing to the industry in which company operates
- It is to be noted that the world went through the dot com bubble crash during the end of 20th century and US housing market collapse causing recession during 2009. Hence, another feature which details the year when a company filed bankruptcy could be helpful for making better insights.
- Better model may be built if one has computational capacity to compute probabilities while fitting the SVM model.

14. ***References:***

- *https://github.com/*
- *https://stackoverflow.com/*
- *https://pandas.pydata.org/docs/*
- *https://www.almabetter.com/*
- *https://www.kaggle.com/*
- https://www.w3schools.com/
- https://www.geeksforgeeks.org/
- https://www.wikipedia.org/