**Machine Learning Classification Capstone project on**

# Company Bankruptcy Prediction

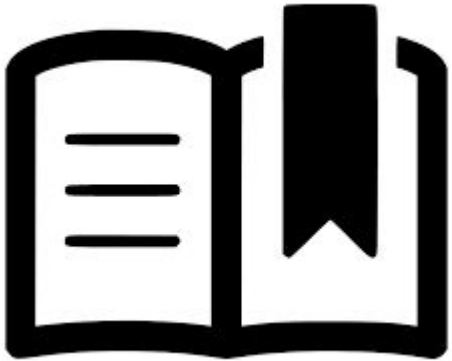
Capstone Project by :

Sanjeev Hegde

Data Science Trainee, AlmaBetter

# Table of Contents:





# Need for analysis:

Investment Decisions

Credit Risk (Banks)

Why Analyse companies for Bankruptcy??


Market Segment health

Devising Strategies: For any company entering newly into the market

# Problem Statement:

Predicting whether company could go bankrupt is essential for several kinds of businesses. However, the process of predicting the same depends on several factors. Examples of few factors are as listed below:

- Cash flow / revenue
- Debt owed by the company
- Offerings by competitors
- Weak Management
- External factors in economy such as recession
- Fraud / scams
- Natural calamities

However, there are countless other factors which are summarised in various financial ratios and effect the bankruptcy as well as overall financial health of the company.

Let us build a classification model which will successfully predict the bankruptcy of the company given several financial factors in the upcoming sections.

# Introduction to data set:

The data set consists of details of the companies along with various financial ratios collected from Taiwan Economic Journal during the period of 1999 to 2009. These companies are classified to be bankrupt or non bankrupt based on the definition of bankruptcy as per the business regulations of Taiwan Stock Exchange.

TAIWAN
STOCK EXCHANGE

# Attributes in the data set:

The data set has total of 96 features including the dependent variable. They include:

- 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)
- 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
- 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

# Attributes in the data set:

Continued…

- 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
- 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

# Attributes in the data set:

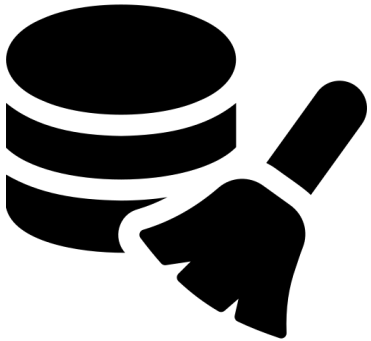
Continued…

- 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

# EDA : Data Wrangling

A clean data set free from abnormalities, missing values and errors is is the foundation of a great machine learning algorithm. In this regard, following points need to be noted for the data set we look forward to build the machine learning classification model:

- Data set consists of 6819 entries against 96 features including the dependent variable
- There are no duplicates in the data set
- There are no null values in the data set
- There are three categorical variables: ‘Bankruptcy?’, ‘Liability-Assets Flag’ and ‘Net Income Flag’
- Extra space and characters in feature names have been removed.

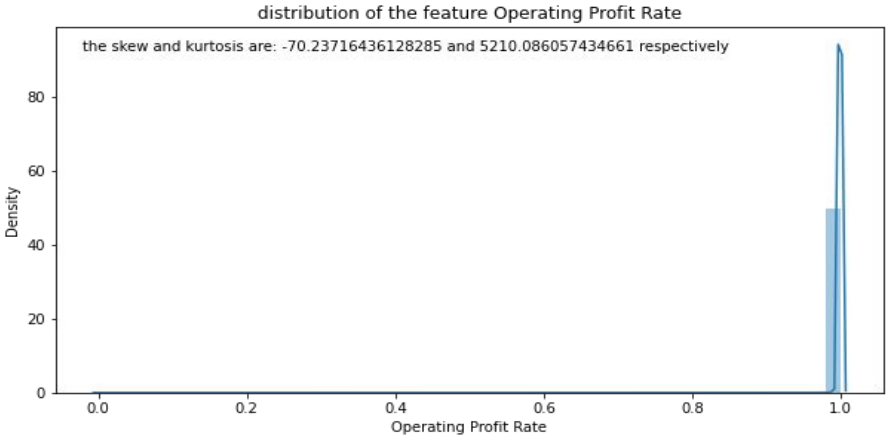


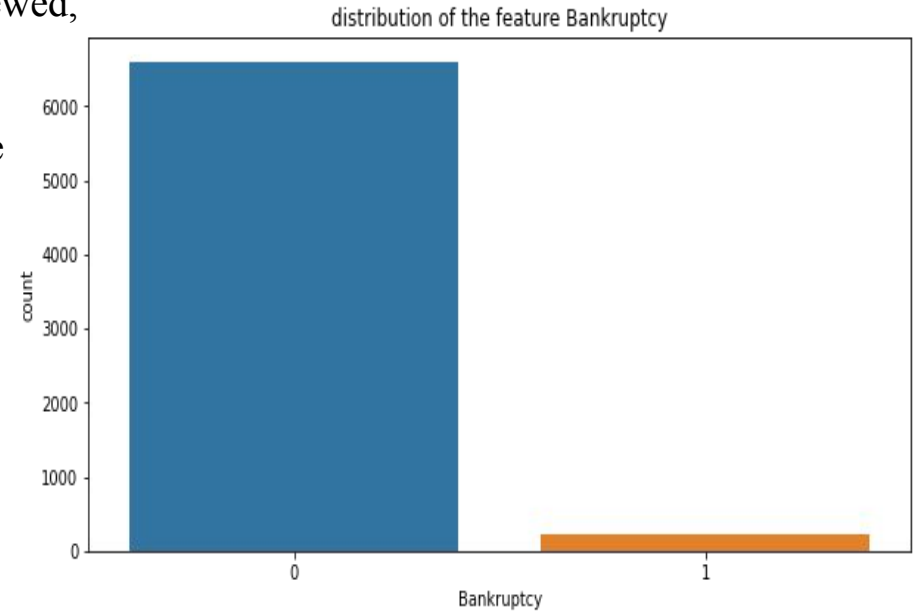
# EDA : Data Visualization

1. Univariate Analysis: Analysis using histograms and barplots have been performed on the features. Categorical features have been plotted using bar plot and continuous variables using histograms. An important conclusion visible from the chart was the imbalance in the data set as shown below.

2. Most features are negatively or positively skewed, Mostly leptokurtic (referred: figure below).

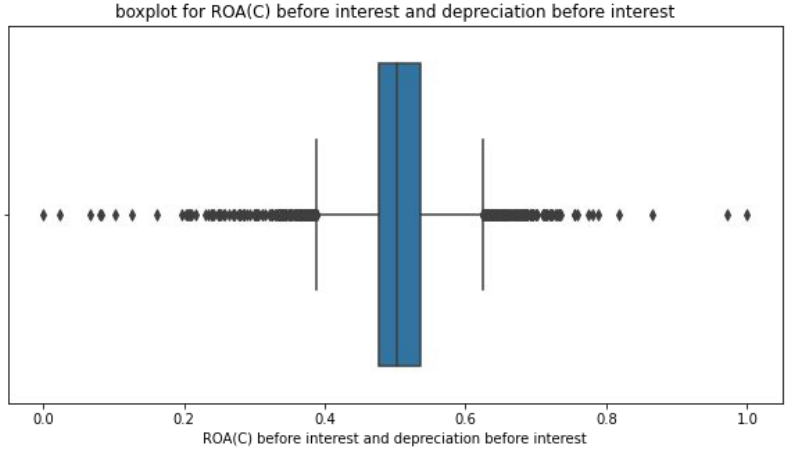
3. Feature ‘Net Income Flag’ has only one value of 1 in the data set and it is thus dropped.

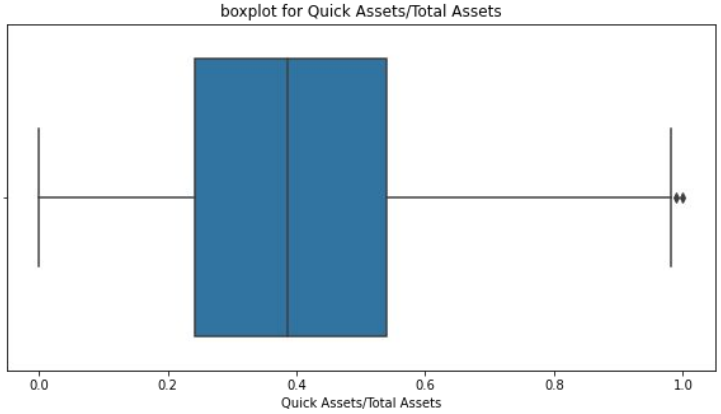


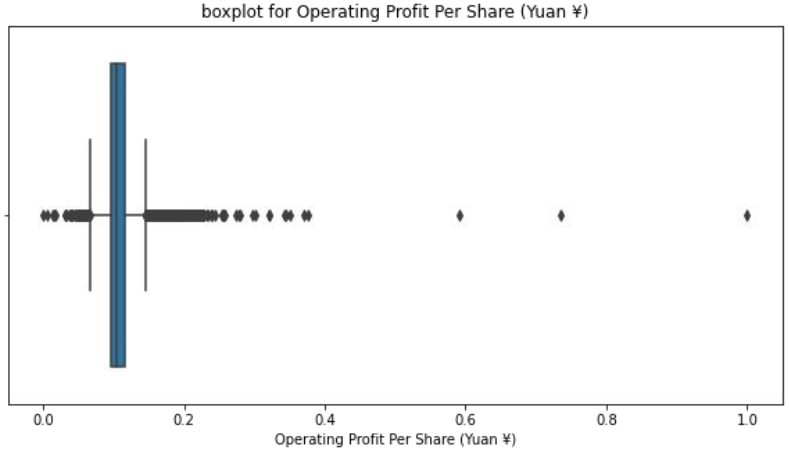



# EDA : Data Visualization

Box Plot: It can be seen that most features in the given data set have considerable amount of outliers in them. It is expected to be this way as various features can have different range of values depending on the firm's valuation, what it does in the market.



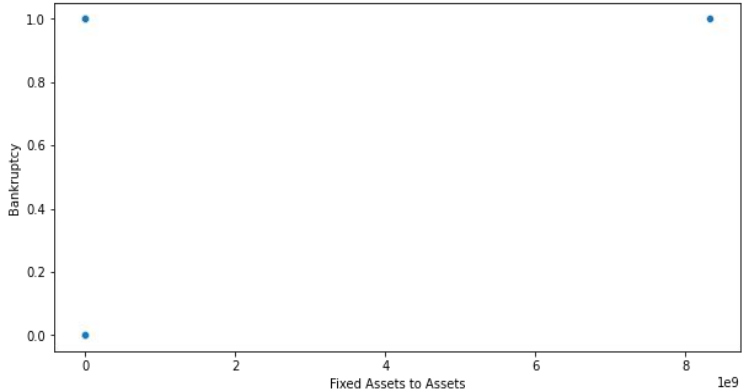



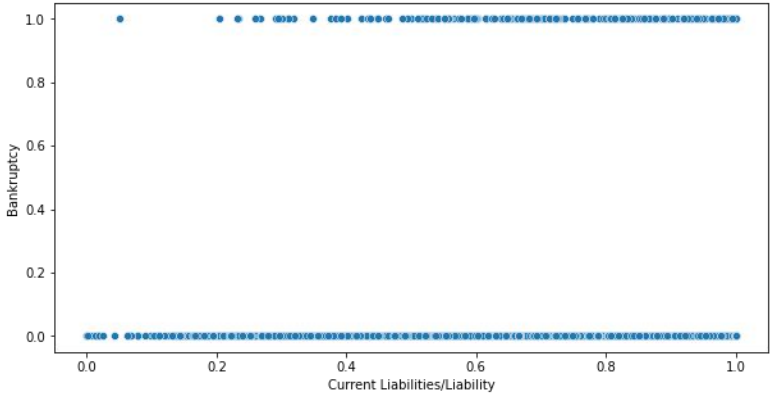



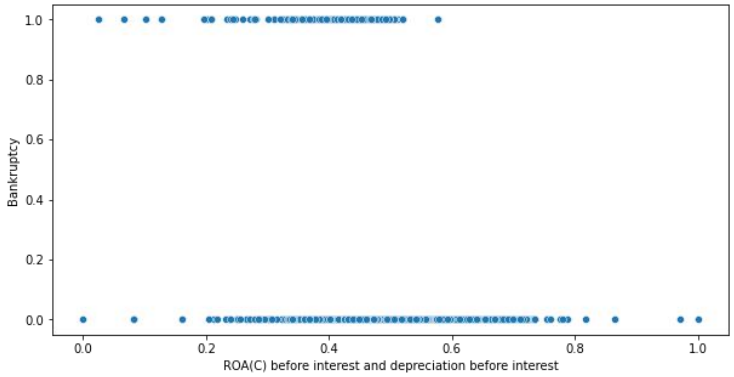
# EDA : Data Visualization

Bivariate Analysis: Scatter Plots

From the scatter plots of all features against dependent variable Bankruptcy, it is clear that most features are continuous variables and 2 are categorical variables.



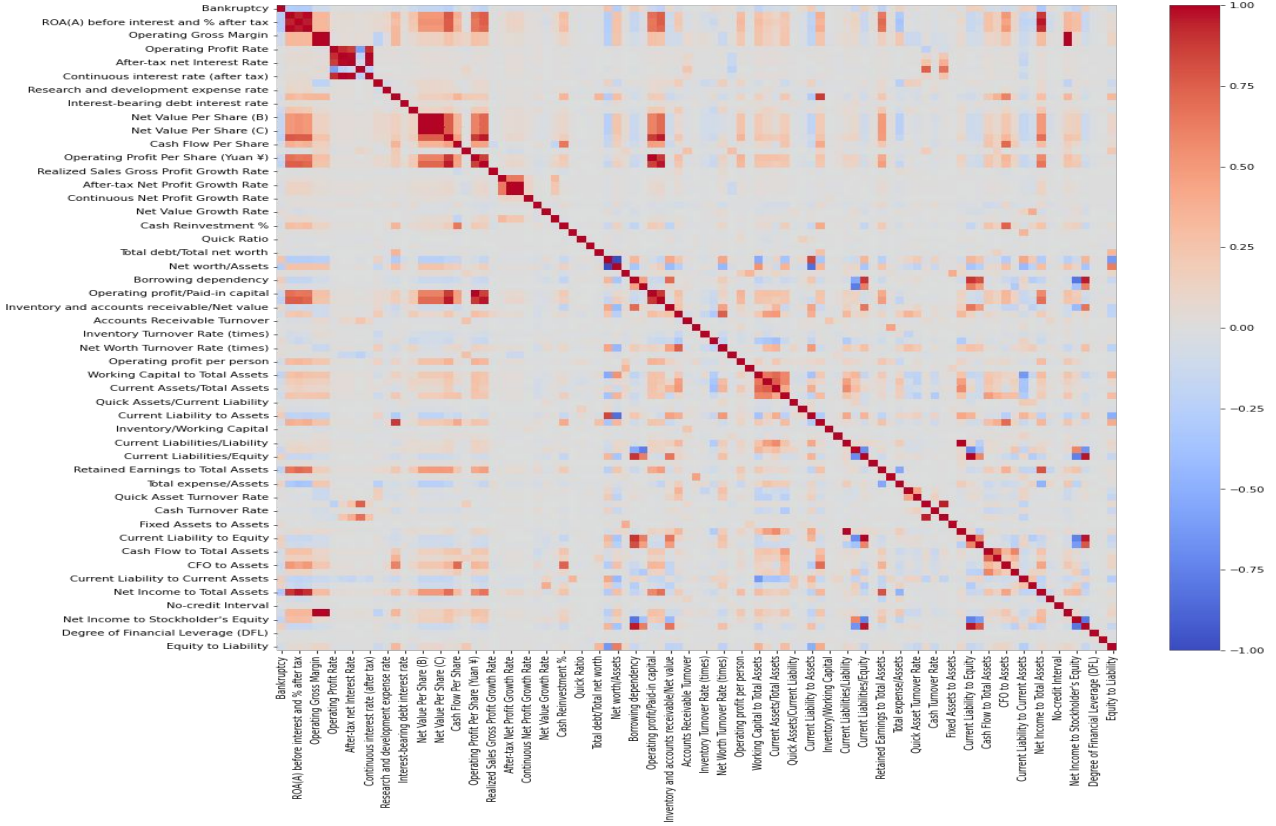
# EDA : Data Visualization

## Multivariate Analysis: Correlation Heatmap

From correlation heatmap it was evident that:

1. 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
2. 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.
3. Current Liability to Liability and Current Liabilities/Liability are same. Similarly, Current Liability to Equity and Current Liabilities/Equity are same. Hence, it is appropriate to drop one of these.

There are other features which can be potentially optimized. However, since most of these features are related to taxation, it is a good idea to leave them untouched.



# Feature Engineering

1. Encoding: One hot encoding has been used to encode the categorical features in the data set i.e., ‘Liability-assets flag’. Its value is found to be 1 if total Liability exceeds Total Assets, 0 otherwise.

2. Feature Manipulation and Selection:

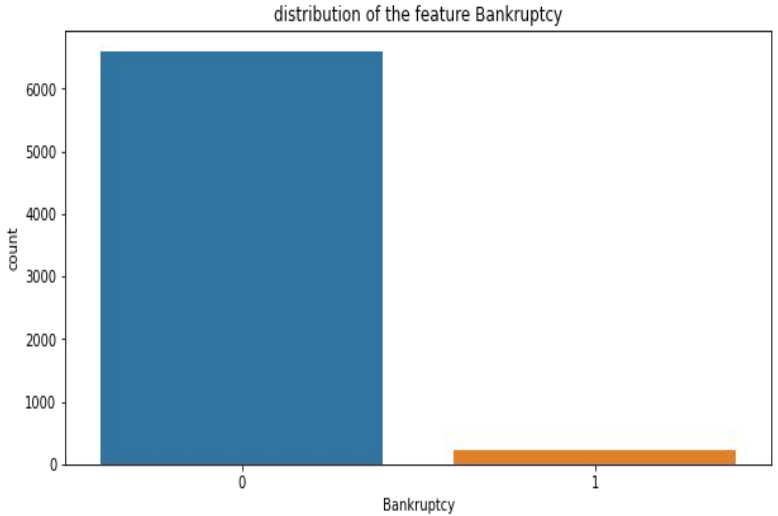
- returns on Assets A, B and C have been summed to reduce the multicollinearity
- Net Value Per Share of A,B and C can be averaged to drop the multicollinearity further
- It is to be noted that after creation of new features, parent features have been dropped from the data set.
- 'Current Liability to Liability', 'Current Liability to Equity' features have been dropped as they are repetitive

3. Bankruptcy feature is declared as the dependent feature and all other features are declared as independent variables

# SMOTE

SMOTE stands for Synthetic Minority Oversampling Technique which is essentially a method to synthetically reproduce minority data points using KNN.

In case of the given data set, it is clearly observed that data is imbalanced with 6599 entries for non bankrupt companies and 220 entries for bankrupt companies. In order to overcome the same, SMOTE is applied on the data set as result of which, we have a balanced data set for making predictions.

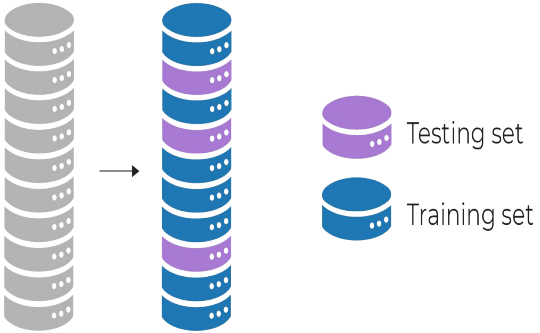



Imbalanced original data set

# Data Splitting and Scaling:

Scaling the data before fitting machine learning model is extremely essential in order to mitigate the effects different range of numeric features. In this regard, Data is initially split using train_test_split from sklearn.

- Training data size is set to be 0.7
- MinMax Scaler is used to scale the data



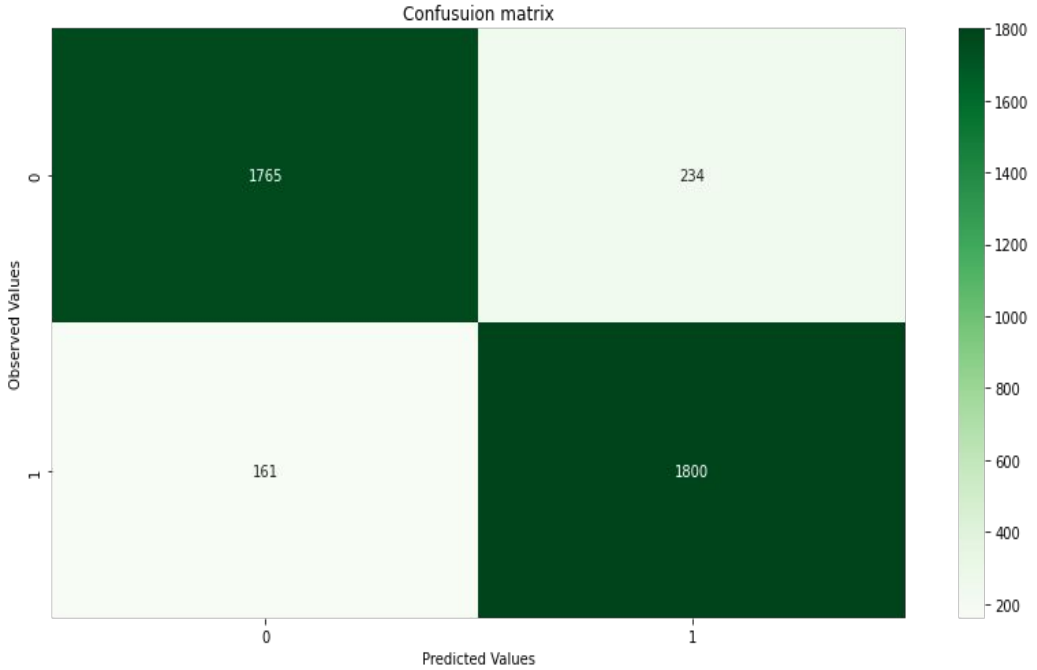
# Logistic Regression

Logistic regression with 5 fold Cross validation has been performed and results are summarised below

**Parameter values**

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

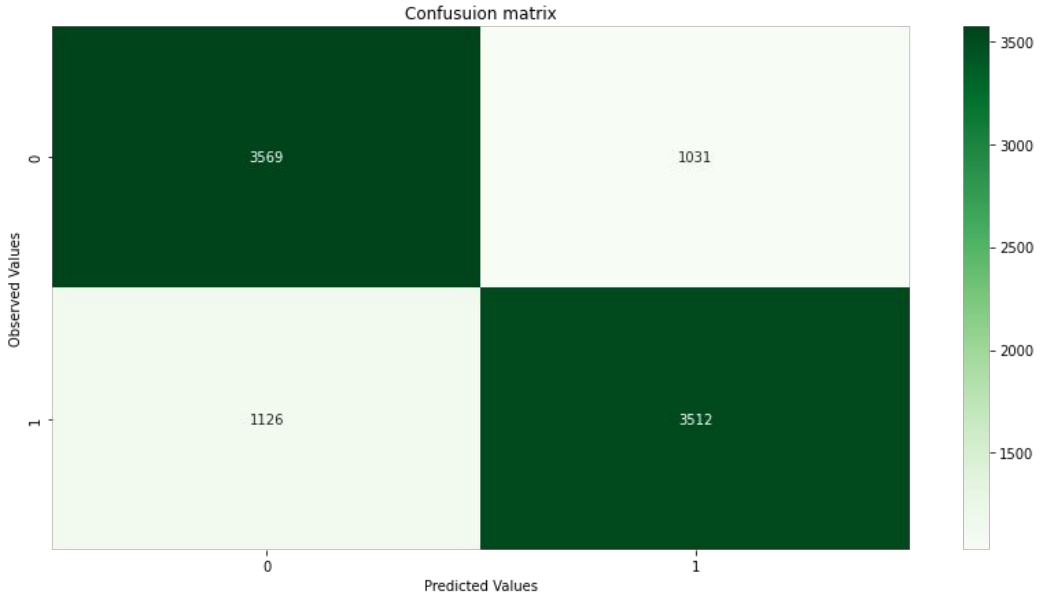
| *Metric(Bankruptcy)/ data set* | **Train** | **Test** |
|---|---|---|
| **Precision** | 0.89 | 0.88 |
| **Recall** | 0.92 | 0.92 |
| **F1 score** | 0.91 | 0.9 |
| **ROC_AUC score** | 0.9045 | 0.9004 |



# Bernoulli-Naive Bayes Classifier

Bernoulli-Naive Bayes Classification has been performed and results are summarised 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 |




It is to be noted that Bernoulli-Naive Bayes Classifier performed poorly compared to Logistic Regression

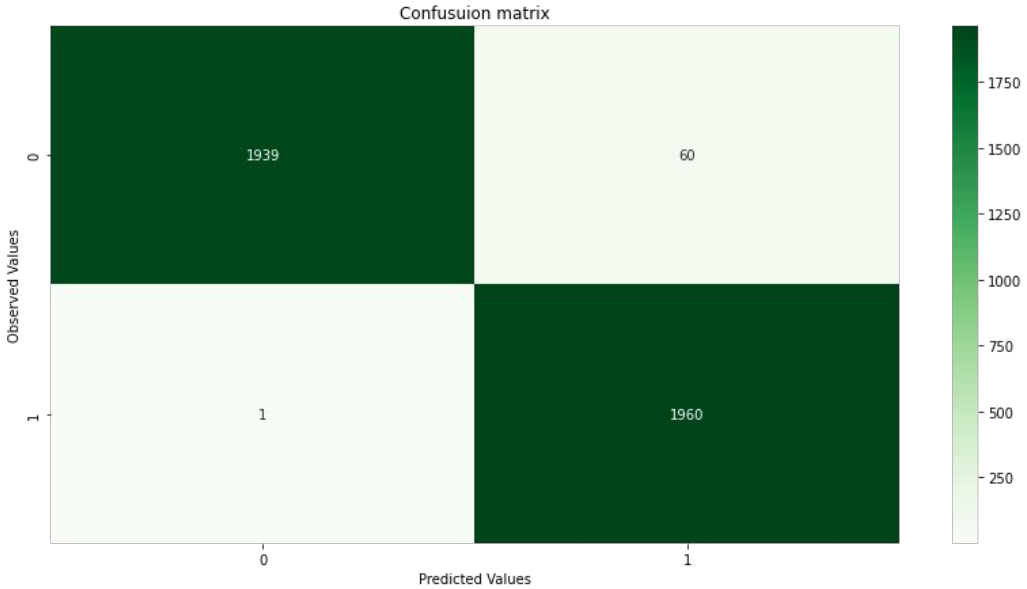
# Support Vector Machines Classifier

Support Vector Machine Classifier with 5 fold Cross validation has been performed and results are summarised 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

| *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 |



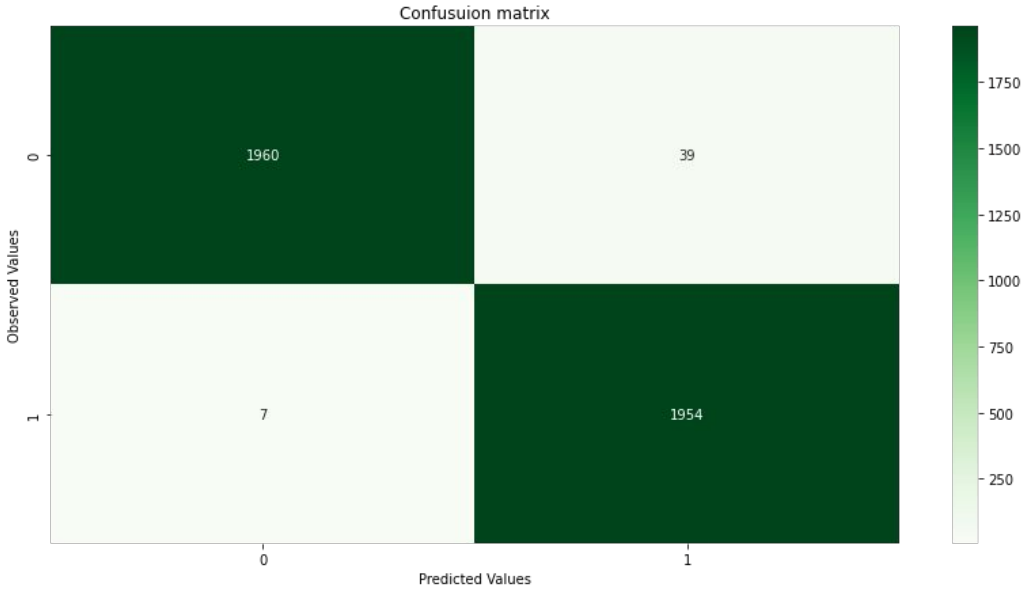
# XGBoost Classifier

XGBoost Classifier with 5 fold Cross validation has been performed and results are summarised below

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

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



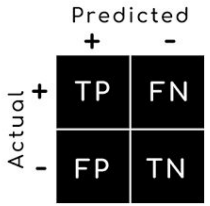
# Choice of Model and Evaluation Metric

Predicting whether a company could go bankrupt is the critical question businesses ask while making investments. In this regard, there is minimal room for error while making such investment decisions.

Hence, using recall score i.e., ratio of true positives to sum of true positive and false negatives is appropriate evaluation metric. Higher recall score indicates lower number of false negatives (predicting that company is not bankrupt by algorithm even though it was observed to be bankrupt) which is extremely essential from the point of view of an Investment Banking firm or any investor in general.

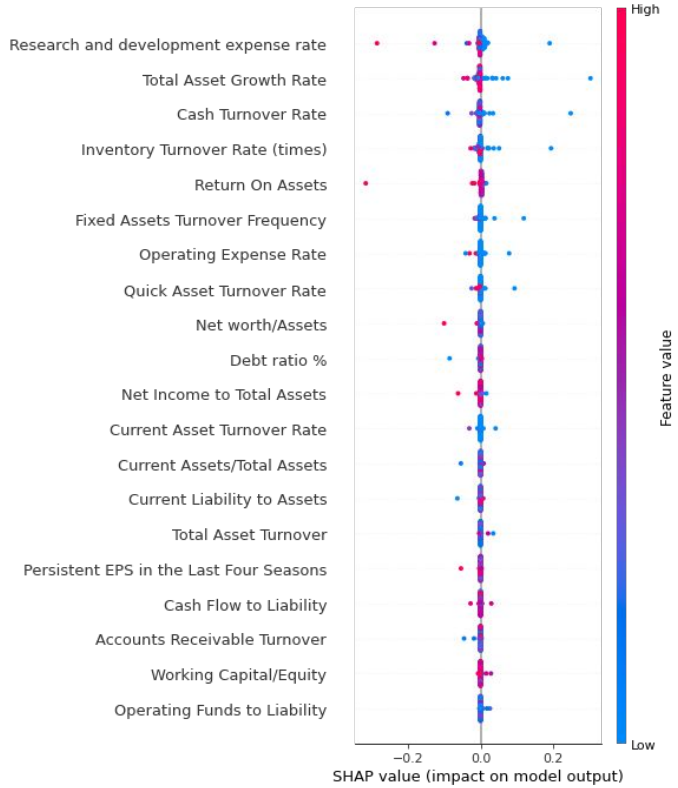
The recall scores for best performing models are as below:

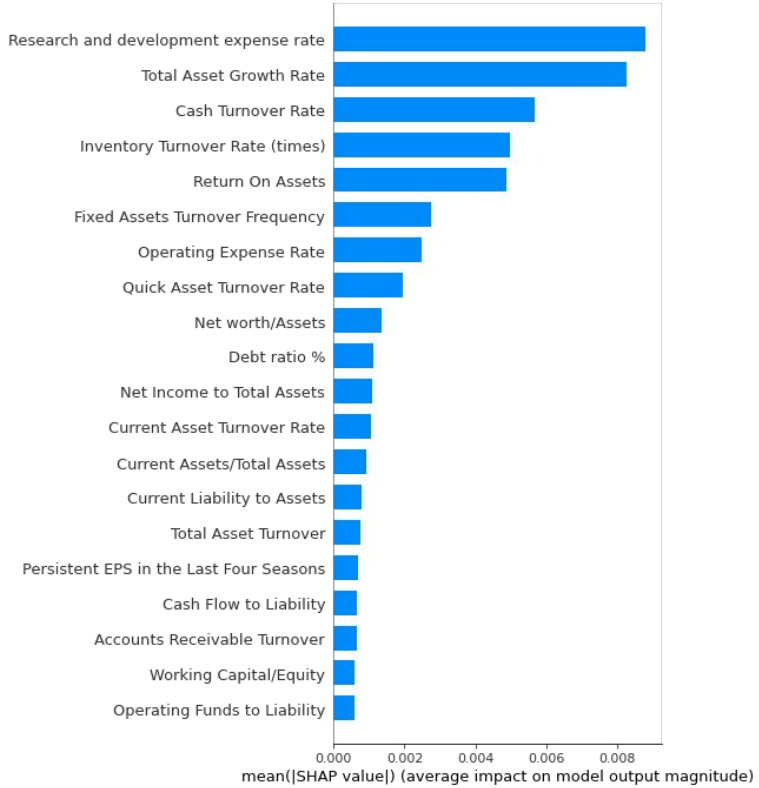
- recall score for SVM model on testing data set is 0.9994
- recall score for XGB model on testing data set is 0.9964
- The above fact is evident from the confusion matrix of these models as well wherein SVM model has just 1 false negative while XGB model has 7.



# Model Explainability using SHAP

SHapley Additive exPlanations is a method based on game theory concept of shapley values used to explain the importance of features.



# Conclusion:

- Given data set is initially studied using EDA techniques and it was found that data did not contain null values or duplicate entries. Data had categorical independent features along with numerical features. Conclusions regarding multicollinearity were made. Also, it was clear that the data set was highly imbalanced.
- One hot encoding was used to encode categorical independent features.
- Features such as returns on different assets, net value per share for various assets were combined respectively to single feature.
- Features with meaning the same with different names were dropped, keeping only one such feature in the data set.
- Data set was balanced using SMOTE.
- Data was split initially and then scaled using MinMaxScaler.
- Logistic regression with 5 fold cross validation was performed with 0.9004 ROC AUC score on test data
- Bernoulli-Naive Bayes classifier was built and it performed considerably poorer compared to other models with ROC AUC score of 0.7707 on test data
- Support Vector Machine Classifier was fit on training data with 5 fold cross validation and 0.9847 ROC AUC score on test data (RBF kernel)
- XGBoost classifier was fit on training data with 5 fold cross validation and ROC AUC score of 0.9884 on test data.
- However, from business point of view, Recall was considered as the evaluation metric and SVM classifier had marginally better recall at 0.9994 while XGB classifier had recall of 0.9964.
- Shap was used to evaluate the feature importance (150 samples) and it was found that 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 were top 5 features driving the company to bankruptcy.

# Future Work and Challenges faced:

- Data collected is insufficient if one decides to have an industry wise outlook by looking at the number of bankrupt companies in specific industry.
- Hence, essentially an additional feature containing the industry in which company operates can provide more detailed insight.
- It is to be noted that world went through dot com bubble crash during the end of 20th century and US housing market collapse causing recession during 2009. Hence, another feature which details regarding the year when a company filed bankruptcy could be helpful for making better insights.
- Computational limits were frequently a barrier to calculate probability using ‘predict_proba’ in SVM model.
- Data set was highly imbalanced and had to be countered using SMOTE