Predicting Company bankruptcy based on Financial and Economic ratios

By

David Song

In partial fulfillment of the requirements for the Springboard's Curriculum

Table of Contents

1. Introduction

Bankruptcy prediction is an area of financial research which explores the ability to predict financial distress and bankruptcy in many public firms. Constructing an effective bankruptcy prediction has been a significant focus in both academic and financial sectors for many years. The literature on bankruptcy prediction dates back to the 1930's beginning with the initial studies concerning the use of ratio analysis to predict future bankruptcy [1]. The subject has been a major concern for entrepreneurs, researchers and even governments for years, since detecting early signs that a company is going to enter bankruptcy involuntarily and being able to save it from that process, can help reduce the economic losses that bankruptcy entails, both in quantitative and qualitative terms [2].

There are many different approaches to developing bankruptcy predictions. Many academic researchers previously used traditional statistics techniques to evaluate the predictive modeling however, as computational science is rapidly advancing, the development methods are also widening to machine learning models and early artificial intelligence models for better and more accurate predictions. For present project, statistical and machine learning techniques (Logistic Regression, Random Forest and Support Vector Machine) have been tested to predict the bankruptcy and were carried out utilizing financial data based on business regulation of the Taiwan Stock Exchange obtained from Taiwan Economic Journal from the years 1999 to 2009.

2. Background and Objective

There are 2 primary objectives for this project. First is to observe which financial ratios are most critical in determining the outcome of the bankruptcy and second, is to determine which supervised learning models produce the best performance predicting the likelihood of bankruptcy.

2.1. Summary of the Data

The data collected from Taiwan Economic Journal consists of features (or input variables) defined as financial ratios which provides information about certain aspect of a business' condition and prospects and are in general used to evaluate and analyze company's performance and financial health. The ratios presented in the data can be categorized into 7 sub-groups namely: solvency, profitability, capital structure ratios, turn-over ratios, cash-flow ratios, growth,

and others. In addition, there were 2 criteria used in collecting the data samples. First, the sample companies had to have at least 3 years of complete public information before the occurrence of the financial crisis. Second, there should be a sufficient number of comparable companies of similar size in the same industry for comparison of the bankrupt and non-bankrupt cases [3]. The data samples are composed of sectors including as manufacturing companies in industrial and electronics, the service industry composed of shipping, tourism, and retail companies and others with exception of financial companies.

2.2. Data structure

There are total of 95 features composed of financial ratios and 1 target feature labelled as 'Bankrupt' which is designated column for classifying the listed companies between bankrupted and not bankrupted. This column consists of labels '1' and '0' where label '1' representing the bankrupt condition of the company and label '0' representing the condition of surviving company. The resultant division of the data set with pre-defined input variables and target variables categorizes this problem as binary classification problem. For the row of the samples, there are total of 6819 entries of the company data thus, the shape of the dataset is structured as 6819 rows with 96 columns.

Inside of the target variable 'Bankrupt' there are only 220 entries of companies that are classified as bankrupted out of 6819 entries and the remainder of 6599 entries are valid and operating companies. By observation, it is apparent that the data-set poses a challenge of imbalanced ratio between the classes where the distribution of majority class (labelled as 0) is proportionally dominant over minority class (labelled as 1). This issue will be further analyzed in the latter section of the project.

3. Data inspection and cleaning

3.1. Overview of the data types

There are only 2 types of data present in the dataset; continuous variables and categorical variables. There are 93 features corresponding to 'float' data type and 3 features composed of 'int' data type including the target variable 'Bankrupt'.

3.2. Inspecting for missing values

Two methods were applied for inspecting for missing values. First, missingno matrix was applied for quick data visualization of any patterns of NaN values in the dataset. And secondly, seaborn heatmap was applied for validation and for better visualization by appointing null check into the function. Both the matrix and the heatmap did not detect any patterns or occurrences of missing/null values in the dataset as observed in Figure 1. Alternatively, by making quick observation from the info function on the data frame, it is recognizable that all columns are identified with 6819 non-null counts which confirm that the dataset has no missing values.

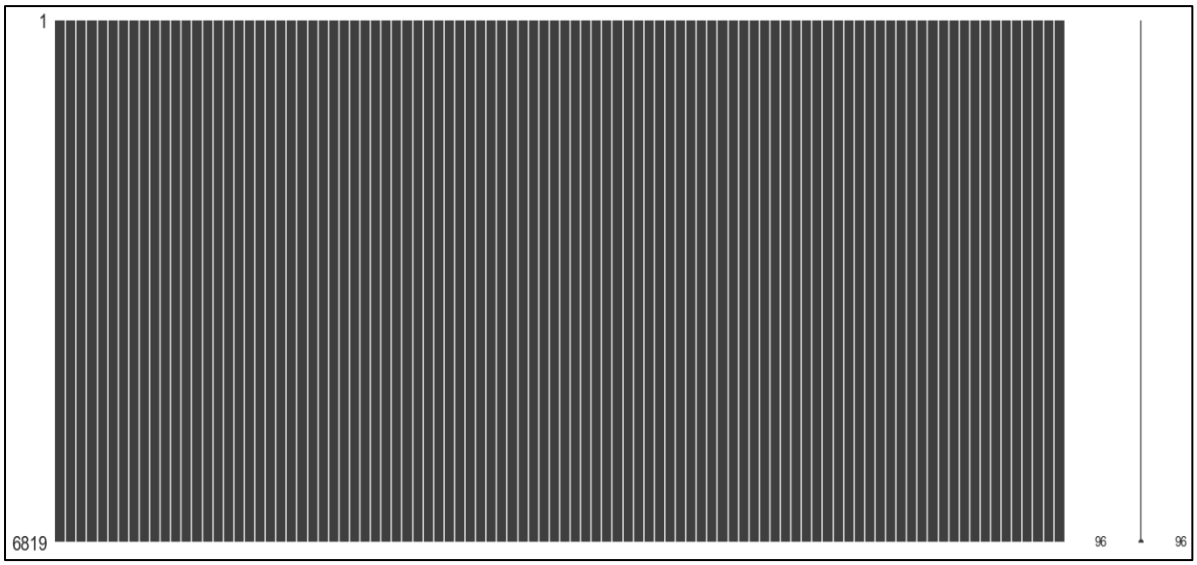

Figure 1. Missingno matrix of the dataset presenting no occurrence of null value.

3.3. Inspecting categorical variables

The 3 categorical features in the dataset are columns labelled as ‘Bankrupt’, ‘Liability-Assets Flag’, and ‘Net Income flag’. All three columns are data type of ‘int’ and it is only classified of 0 and 1s. Since ‘Bankrupt’ column is the target feature for present project, it has been discarded from any evaluation or manipulation of the data. Simple value counts function was applied for the evaluation of the ‘Liability-Assets Flag’ and ‘Net Income flag’ to identify the categories of each feature. Conclusively, ‘Net Income Flag’ column only contained 1 constant value (Refer to Figure 2) and was removed from the dataset as the variable did not provide any significant importance to the analysis.

```
Liability-Assets_Flag  Net_Income_Flag
0                      1                  6811
1                      1                     8
dtype: int64
```

Figure 2. Evaluation table of 2 categorical variables in the dataset

4. Exploratory Data Analysis

There are numerous features to select in the dataset and each feature depicts its own significant influence to the target variable. Although it would be most accurate to examine each individual column for significant relation to the Bankruptcy variable, inspecting all 93 features would be tedious and laborious task. For the present project, selective methods have been implemented to determine most critical features which impacts the target variables the most while reducing the overall number of features for simpler exploratory analysis.

4.1. Scatter plots

Before exploring further into the data, scatter plots have been plotted on few interesting features for quick examination to observe if there are any patterns or correlation to be identified. Total of 9 features have been randomly selected from the sub-groups discussed in the Section 2.1 and shown in below table:

Constant variable:	**Comparison Variables:**
ROA(A) before interest and % after tax return	ROA(B) before interest and depreciation after tax
	ROA(C) before interest and depreciation before interest
	Cash/Total Assets
	Net profit before tax/Paid in capital
	Net worth/Assets
	Total Assets Growth Rate
	Realized Sales Gross Margin
	Debt ratio%

The method of displaying the features onto the scatter plot is by selecting 1 constant variable as independent variable (x-axis) and arranging the remainder of variables (Comparison Variable) to dependent variable (y-axis) while distinguishing the class variables between 'Bankrupt' and 'Not bankrupt'. Hence, 'ROA(A) before interest and % after tax return' is repeated in each scatter plots as a constant parameter in conjunction to all comparison parameters. Figure 3 and 4 represents the scatter plot for the selected features. The class labels 0 and 1 have been reassigned as 'No' and 'Yes' for easier interpretation of bankruptcy.

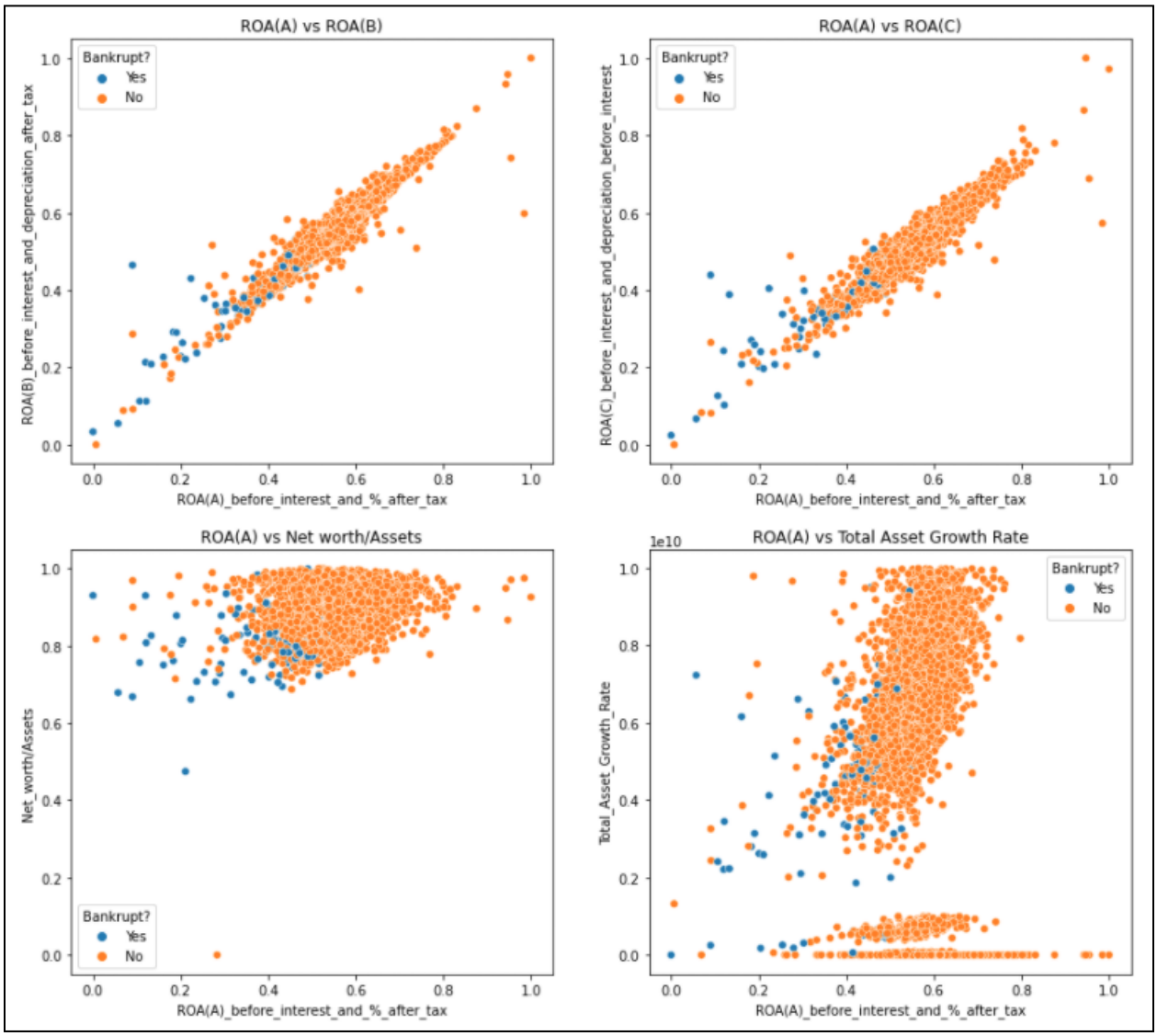

Figure 3. Scatter plots representation of 'ROA(A) vs. ROA(B)', 'ROA(A) vs. ROA(C)', 'ROA(A) vs. Net worth/Assets', ROA(A) vs. Total Asset growth Rate'

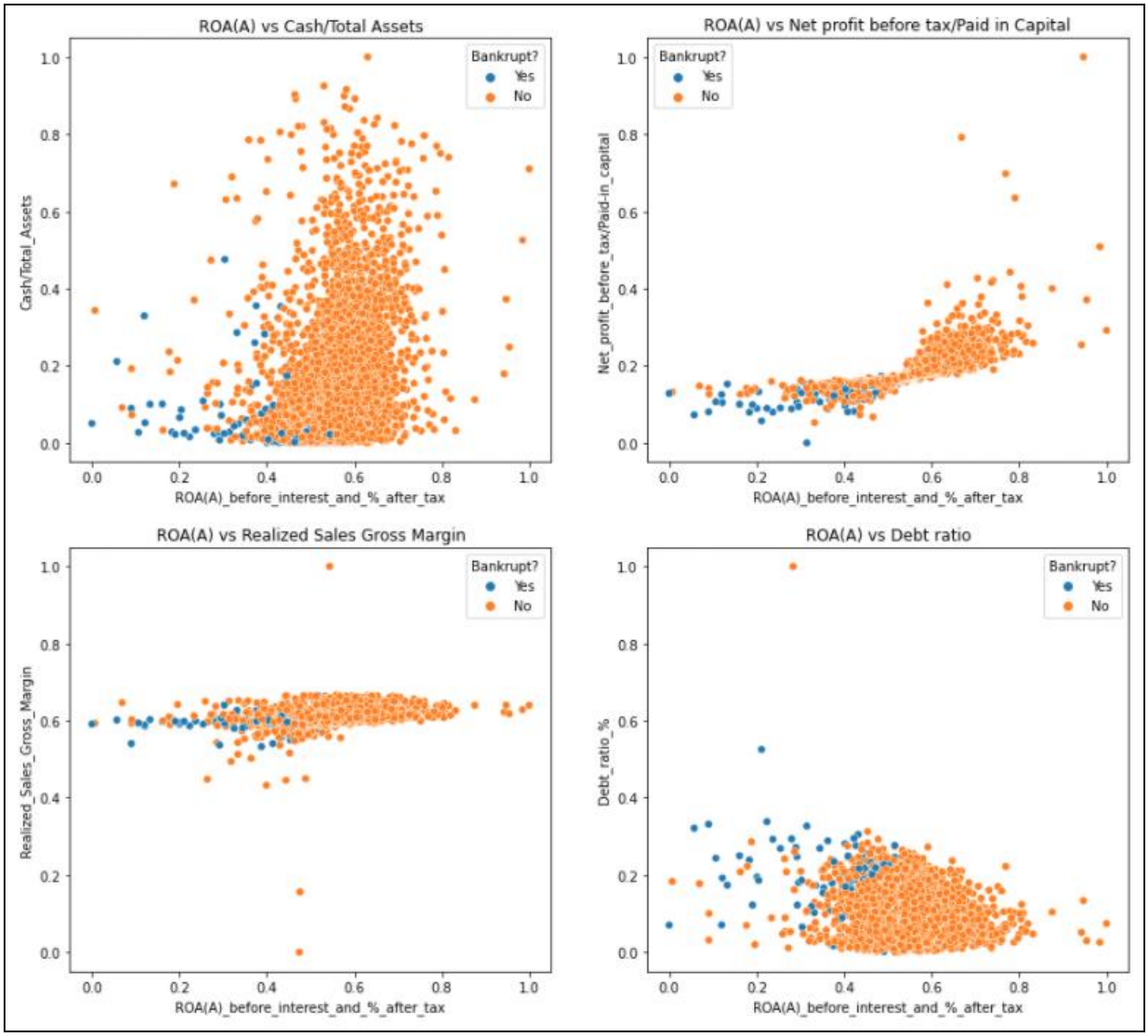

Figure 4. Scatter plots representation of 'ROA(A) vs. Cash/Total Assets', 'ROA(A) vs. Net profit', 'ROA(A) vs. Realized Sales Gross Margin', 'ROA(A) vs. Debt ratio'

On both of the figures there is noticeable pattern on the overall scatter plots. At instance, the plots appear to display perceptible boundary between bankrupt and not bankrupt labels as majority of the companies labelled as bankrupt are spotted in the lower range (between 0.0 and 0.6) of the Return on Assets (ROA) values. The inference constructed at this instant could signify that bankrupt companies are less efficient in managing its assets to generate earnings as ROA is a indicator of the profitability in relation to company's total assets. However, it is difficult to justify the outcome of bankruptcy sorely on a single feature. Exploiting through the individual scatter plots, following assessment can be considered:

1. ROA(A) vs. ROA(B) & ROA(C) reveals a linear pattern as bankrupt class is concentrated in the lower range of the plot as opposed to the not bankrupt class. These variables depict high correlation to each other and could contain overlapping information between the features.
2. In ROA(A) vs. Cash/Total Assets plot, the surviving companies evidently show higher Cash/Total assets ratio. Cash/Total Assets ratio measures the portion of company's assets held in cash or marketable securities. The indication in the pattern could signify that the likelihood of survival from the bankruptcy increases with higher degree of cash or assets in the company as safety net.
3. In ROA(A) vs. Debt Ratio plot, the bankrupt companies contain moderately higher values of Debt Ratio compared to surviving companies. As debt ratio defines the ratio of total debt to total assets, this could indicate that probability of bankruptcy increases in the companies with higher proportion of liabilities in relation to assets.
4. The bankrupt class is comprised of lower values in the plots of Net worth/Assets, Total Asset growth rate and Realized sales gross margin however, the difference is inconsiderable and the pattern appears to be insignificant for evaluating the target variable.

The scatter plots exhibits interesting patterns in association to the target variable. It is noticeable that profitability of the financial ratio, ROA(A), ROA(B) and ROA(C) reveals the most significant difference between the class variables. Conversely, these features are also highly correlated to each other as it can seen by the near perfect linear relationship in the plots, hence it is difficult to determine accurately if corresponding features are truly the most critical features affecting the target variable. The Solvency groups, Cash/Total Assets and Debt Ratio plots also present apparent distinction between bankrupt and not bankrupt. As the plots indicate, the company with higher ROA have tendency to maintain higher Cash/Total Assets ratio thus chances of bankruptcy decreases. Inversely, the companies with lower ROA have higher Debt ratio on average indicating bankrupt companies have higher proportion of liabilities in relation to assets.

4.2. Correlation heat-map

Correlation heat-map is a great tool for demonstrating graphical representation of correlation between each feature. As previously seen in the scatter plots, some of the features are highly correlated with each other. Figure 5 represents the correlation heat-map of all the features presented in the data. At first glance of the graph, there are 2 aspects to be noted. First, there is excessively abundant number of features presented in the data set and it is hard to distinguish and analyze the graph. Not all of the features will have significant influence on the analysis and the features which deliver the least impact to the target variable should be identified and removed from the data. This task will be executed through Feature selection.

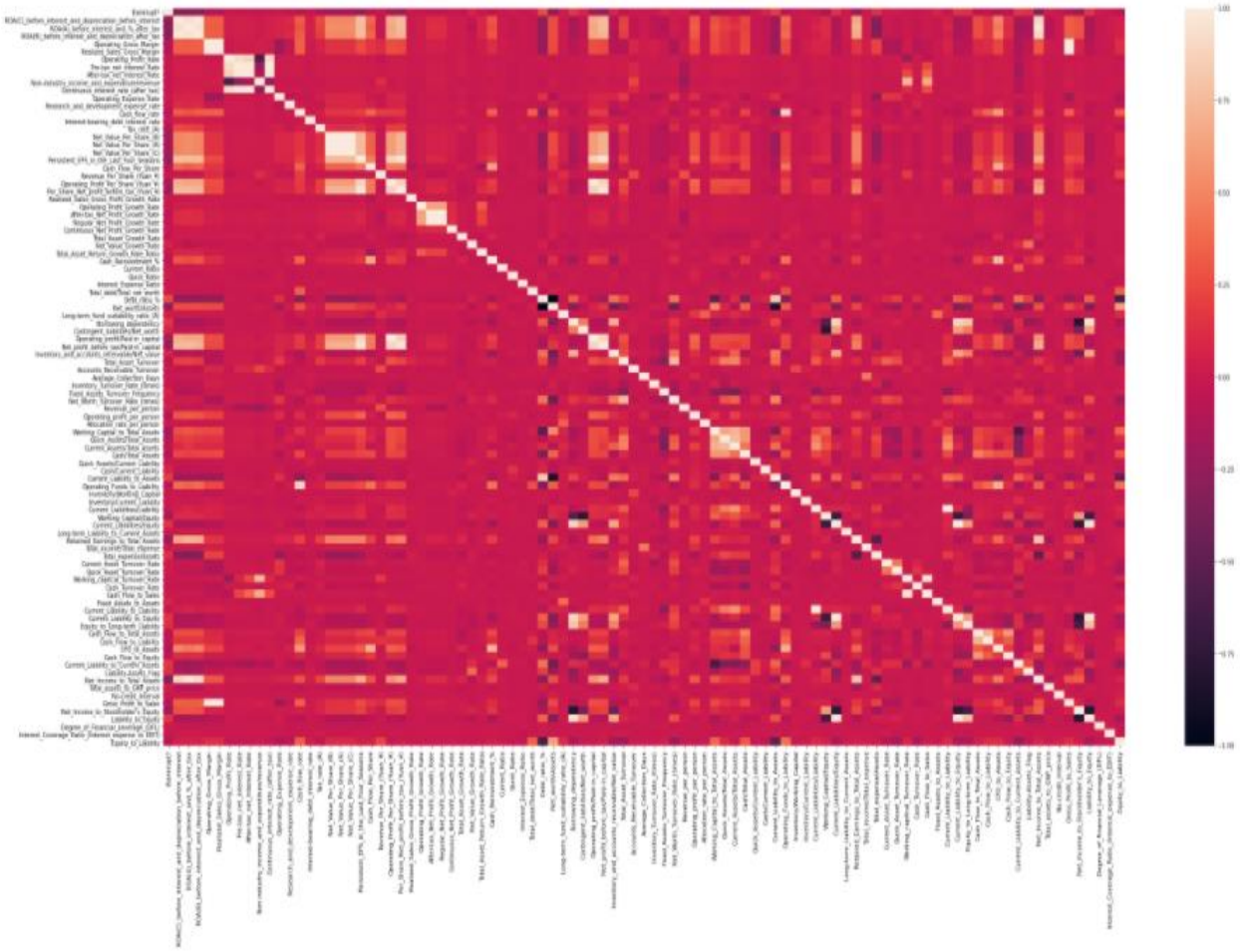

Figure 5. Correlation heat-map of all the features selected in the data frame

Secondly, there are variables which measures extremely close to correlation value of 1 or -1. This extremity signifies that some features are highly correlated with each other and possibly have high colinearity and redundant. High correlation between the independent variables in the regression analysis could impact overall interpretation of the result and can lead to multicollinearity.

4.3. Multicollinearity check

Multicollinearity is the occurrence of high intercorrelation among 2 or more independent variables in regression model and can lead to skewed or misleading results. It describes a perfect or exact relationship between regression exploratory variables. An example of this relation was seen in the scatter plot of ROA(A) vs. ROA(B) and (C). The heat-map does not specify or reveal if the information in one variable is explained through multiple other variables however, the high correlated value of 1 and -1 between the features represent that some of the features may provide redundant or duplicate information of each other. In order to address the issue, features which have correlation coefficient above 0.95 and below -0.95 have been identified and separated into a new list and dropped from the main data frame. By doing so, it reduces the number of features to avoid multicollinearity in the variables. The final number of the columns after filtering the multicollinearity in the dataset is 78.

4.4. Feature Selection

The aim of feature selection or dimensionality reduction is to reduce irrelevant or redundant features by selecting more representative features having more discriminatory power over a given data set [4]. The number of the features has been reduced to 78 through multicollinearity check method however, the data-set still contains large number of features. For the feature selection technique, Backward Elimination method was considered. Backward Elimination method is a common technique of wrapper method which begins with all the features and removes the least significant feature at each iteration with the intention of improving the performance of the model. The process repeats until there are no further improvements observed on removal of the features. The metric used to evaluate the performance of the feature is by using P-value. The steps taken for the elimination method is as follows:

1. Select the significance level (P-value) as 5%(0.05).
2. Fit the model with all the features selected.

3. Identify the predictor with the highest P-value
4. If the P-value identified is greater than the set significance level, the feature is removed from the column
5. Once the feature is removed, the data-set is inserted back into the model
6. The process is iterated until the highest P-value in the data-set is less than 0.05 and all features are filtered out.

With the above technique, the features are eliminated down to 29 from 78 and new column are designated into new data frame presented in Figure 6. The below identified columns are the most critical features for distinguishing the target class defined the elimination.

#	Column	Non-Null Count	Dtype
0	Bankrupt?	6819 non-null	int64
1	Operating_Profit_Rate	6819 non-null	float64
2	Non-industry_income_and_expenditure/revenue	6819 non-null	float64
3	Continuous_interest_rate_(after_tax)	6819 non-null	float64
4	Tax_rate_(A)	6819 non-null	float64
5	Net_worth/Assets	6819 non-null	float64
6	Contingent_liabilities/Net_worth	6819 non-null	float64
7	Operating_profit/Paid-in_capital	6819 non-null	float64
8	Inventory_and_accounts_receivable/Net_value	6819 non-null	float64
9	Fixed_Assets_Turnover_Frequency	6819 non-null	float64
10	Net_Worth_Turnover_Rate_(times)	6819 non-null	float64
11	Revenue_per_person	6819 non-null	float64
12	Operating_profit_per_person	6819 non-null	float64
13	Working_Capital_to_Total_Assets	6819 non-null	float64
14	Quick_Assets/Total_Assets	6819 non-null	float64
15	Current_Assets/Total_Assets	6819 non-null	float64
16	Cash/Current_Liability	6819 non-null	float64
17	Current_Liability_to_Assets	6819 non-null	float64
18	Operating_Funds_to_Liability	6819 non-null	float64
19	Retained_Earnings_to_Total_Assets	6819 non-null	float64
20	Cash_Turnover_Rate	6819 non-null	float64
21	Fixed_Assets_to_Assets	6819 non-null	float64
22	Current_Liability_to_Liability	6819 non-null	float64
23	Cash_Flow_to_Equity	6819 non-null	float64
24	Current_Liability_to_Current_Assets	6819 non-null	float64
25	Liability-Assets_Flag	6819 non-null	int64
26	Net_Income_to_Total_Assets	6819 non-null	float64
27	Net_Income_to_Stockholder's_Equity	6819 non-null	float64
28	Liability_to_Equity	6819 non-null	float64
29	Equity_to_Liability	6819 non-null	float64

Figure 6. New columns filtered through the methods of Multicollinearity check and Feature selection

The result of the selection can be seen in new correlation heat-map presented in Figure 7. It is apparent that there is significant decrease in the correlation values between the features compared to the original heat-map discussed in section 4.2.

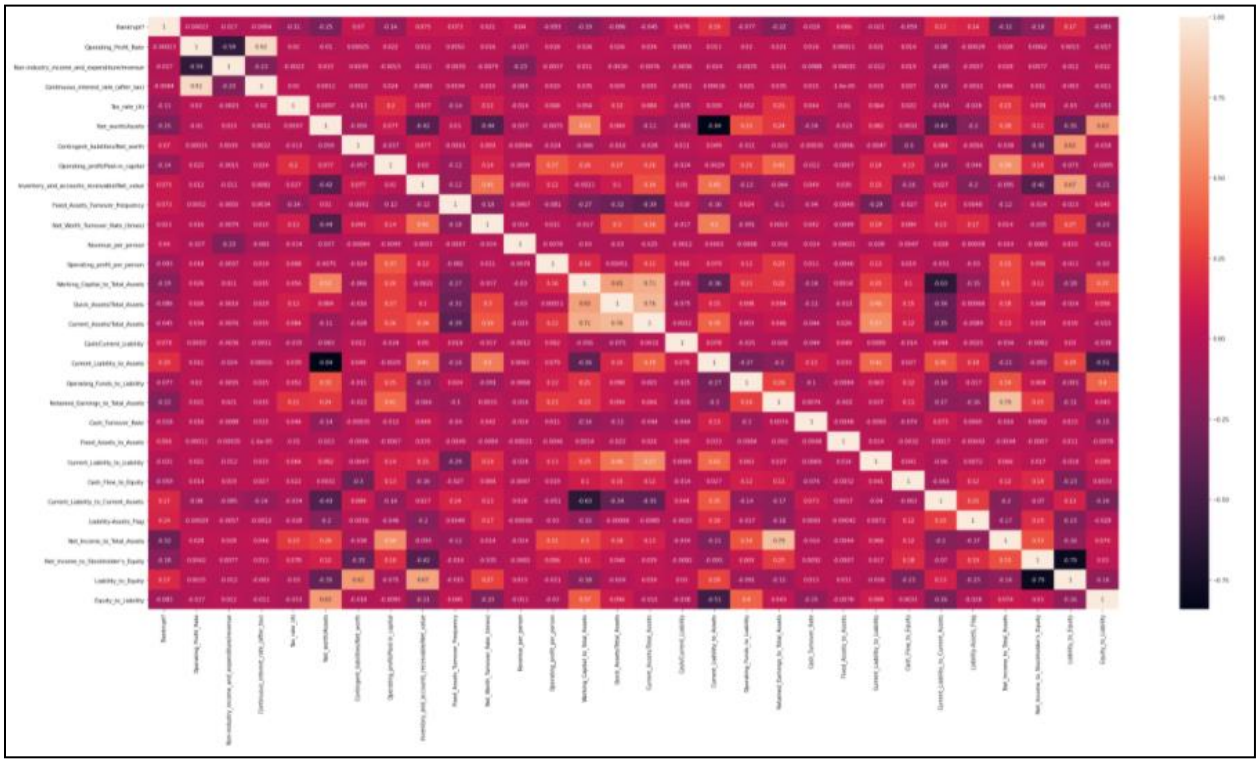

Figure 7. Correlation heat-map of new reduced features refined through Multicollinearity check and Feature selection.

4.5. Box plots

For further analysis in the individual features, total of 16 features were selected from the new columns for deeper examination through the box plots. The new columns are divided into 4 major categories namely: Solvency (Asset ratios), Solvency (Liability ratios), Turnover and Cash flow ratios and Operation measures.

4.5.1. Solvency: Asset ratios

Asset ratios are the features which presents percentage of the company's assets financed by creditors. Following features are presented in this sub-category:

1. Current Assets/Total Assets - A ratio to determine economic value for or within one year to all current assets. Ratio describes the measurement of the short-term liquidity of the

company, or its ability to generate enough cash to pay off all debts should they become due at once.

2. Quick Assets/Total Assets - Similar to Current assets ratio, Quick assets ratio helps to measure the liquidity of the company but excluding inventory, and other less liquid assets and focuses on the company's most liquid assets.
3. Net Worth/Assets – Measures the amount of equity the business has when compared to the total assets owned by the business. Higher the Net worth to asset ratio, the less leveraged the company is and indicates that a larger percentage of its assets are owned by the company and its investors.
4. Working Capital/Total Assets – A ratio that compares the net liquid assets to the total assets of the firm. The ratio is an indicator of the short term liquidity and financial strength of the business and its ability to finance short term obligations

It can be observed that overall box plots between the bankrupt and not bankrupt companies reveal significant difference in the ratios. Surviving companies have higher ratio values on both Current Assets/Total Asset ratio & Quick Assets/Total Assets on average which presents that these companies are more capable of generating cash to pay off debts should they become due at once. Liquidity is critical to any company and if a company cannot meet its financial obligations, it increases the change of bankruptcy. The Net worth/Assets and Working capital to Total Assets reveals even larger gap between bankrupt and not bankrupt companies as the surviving companies reveal much higher ratio on interquartile range.

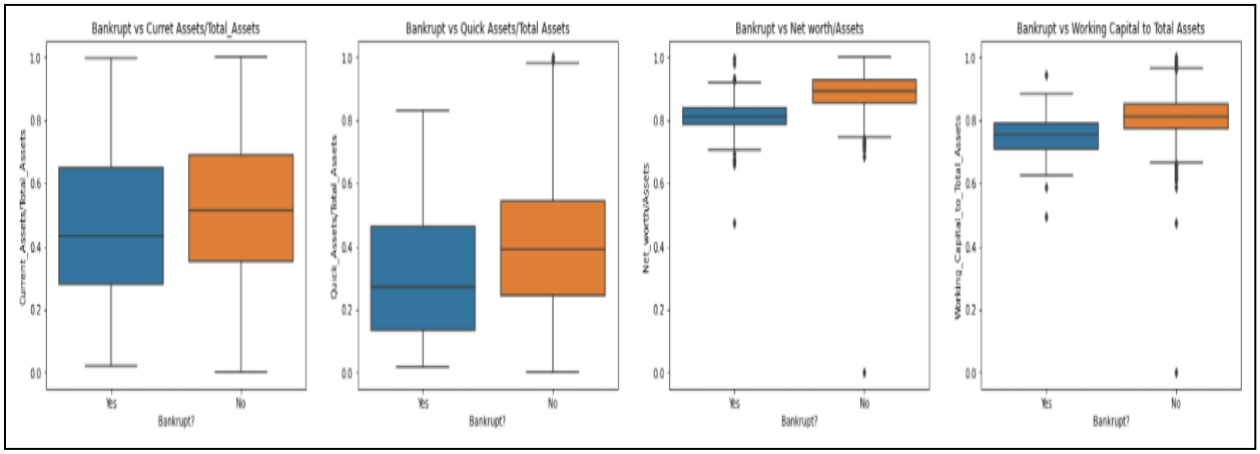

Figure 8. Box plots of Solvency: Asset ratios between Bankrupt vs Not bankrupt classes

4.5.2. Solvency: Liability ratio

Liability ratios are the features that define the financial obligation of the company. In this sub-category, following features are presented:

1. Current Liability to Assets – Debt to Asset ratio, current liabilities divided by the total amount of the company has in assets, whether short-term investments or long-term and capital assets.
2. Current Liability to Current Assets – Current liabilities are obligations expected to be paid within one year to current assets are those which can be converted into cash within one year.
3. Equity to Liability – Also known as Debt-to-equity, and is used to evaluate a company's financial leverage. It is measure of the degree to which company is financing its operations through debt vs. wholly owned funds.
4. Current Liability to Liability – Company's short-term financial obligations that are due within one year or within a normal operating cycle to total liability.

The liability ratios displays similar patterns compared to the Assets ratio analysis but in reverse order. Asset ratios overviews financial strength while Liability ratios overview debts and obligations of the business. As expected, bankrupt companies are comprised of higher ratios of liabilities compared to surviving companies as observed in Figure 9. Current liability to assets ratio presents clear gap between the target variables and the interquartile range on bankrupt companies have much higher values. This indicates that the bankrupt companies have tendency to have higher proportion of current liability in relation to their total assets. The similar pattern is also examined in Current Liability to Current Assets however, the difference between the target variable is minor. As both above ratios are defined as company's ability to convert assets into cash to pay off obligation in long term and short term, this signifies that surviving company exhibits greater strength in financial stability and are able to avoid bankruptcy as liability is proportionally smaller in respect to assets. Another interesting factor to be noted is that the difference between the target classes in liability ratio box plots are not noticeably perceivable as compared to box plots for Asset ratios. This factor implies that features defining assets ratios are comparatively more influential in determining the bankruptcy than the liability ratios.

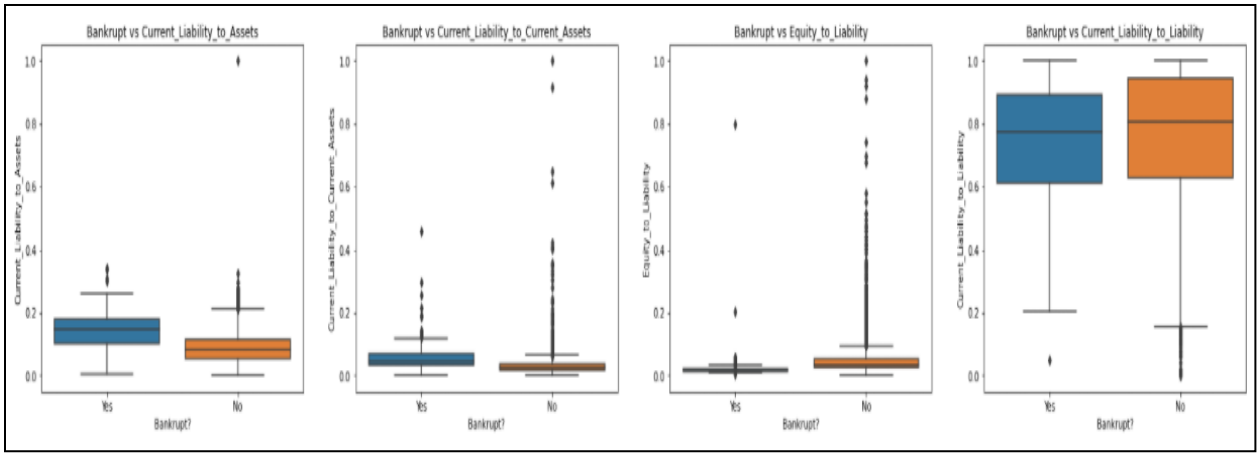

Figure 9. Box plots of Solvency: Liability ratios between Bankrupt vs. Not bankrupt classes

4.5.3. Turnover and Cash flow ratios

In this sub-category, following variables are presented:

1. Net Income to Total Assets – A ratio which measures percentage of profit a company earns in relation to its overall resources. It is used to define the profitability of a company in relative to its total assets.
2. Cash turnover rate – Efficiency ratio that reveals the number of times that cash is turned over in an accounting period. Used to determine the proportion of cash required to generate sales.
3. Retained Earnings to Total Assets = Depicts the financial leverage of the entities. This ratio indicates how assets were financed from retention of profit instead of paying profit out as dividends and acquiring loans.
4. Cash Flow to equity – A ratio calculating how much cash is available to the equity shareholders of a company after all expense, reinvestment and debt are paid.

It is apparent that there is no significant pattern to be made in the box plots apart from the few aspects. Both the net income to total assets and Retained earnings to total assets display moderately higher values for the surviving companies. This pattern indicates that surviving companies are generally more profitable and proficient in retaining its profits to finance assets instead of paying out dividends or converting debt and new capital to fund its operations.

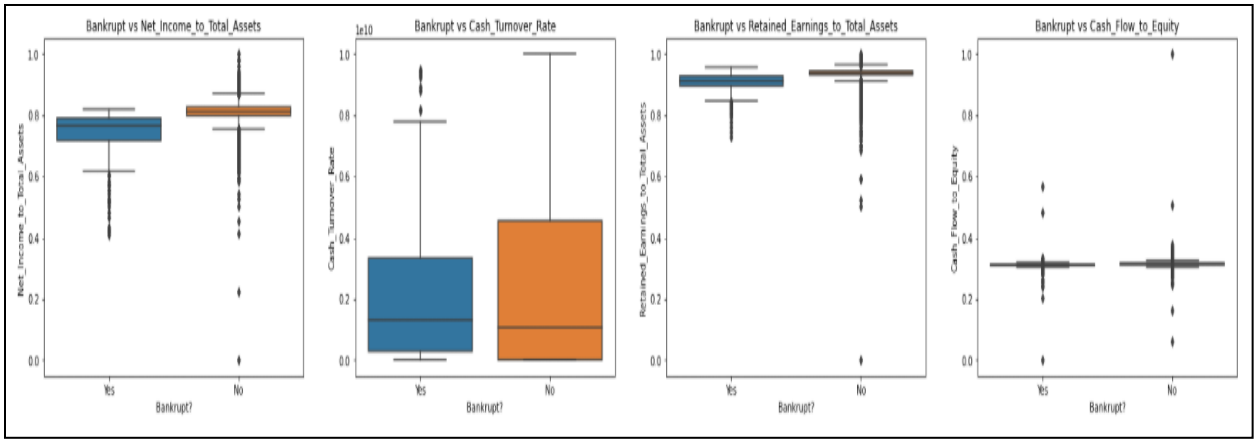

Figure 10. Box plots of Turnover and Cash flow ratios between Bankrupt vs. Not bankrupt classes

4.5.4. Operation Measures

The contents of the last sub-categories inspect the ratios that define the operational competence of the company:

1. Operating Profit rate – Profitability or performance ratio that reflects the percentage of profit a company produces from its operations
2. Operating Funds to Liability – A ratio which measures how well a company can pay off its current liabilities with the cash flow generated from its core business operations
3. Operating profit/Paid in Capital – Measurement of profit in operation to the funds raised by the business through selling its equity and not from ongoing business operations.
4. Fixed Assets Turnover Frequency – A ratio that measures operating performance of the company. Efficiency ratio compares net sales (Income Statement) to fixed assets (Balance sheet) and measures a company's ability to generate net sales from its fixed-asset investment (property, plant, equipment).

The box plots for operation measure category displays almost no difference between the bankrupt and not bankrupt classes as shown in Figure 11. In contrast to both solvency and turnover/cash flow plots, the operation measure plots are measured evenly between 2 classes. The insignificance in the plots interprets that operation measuring ratios are not a critical contributing factor in determining the bankruptcy.

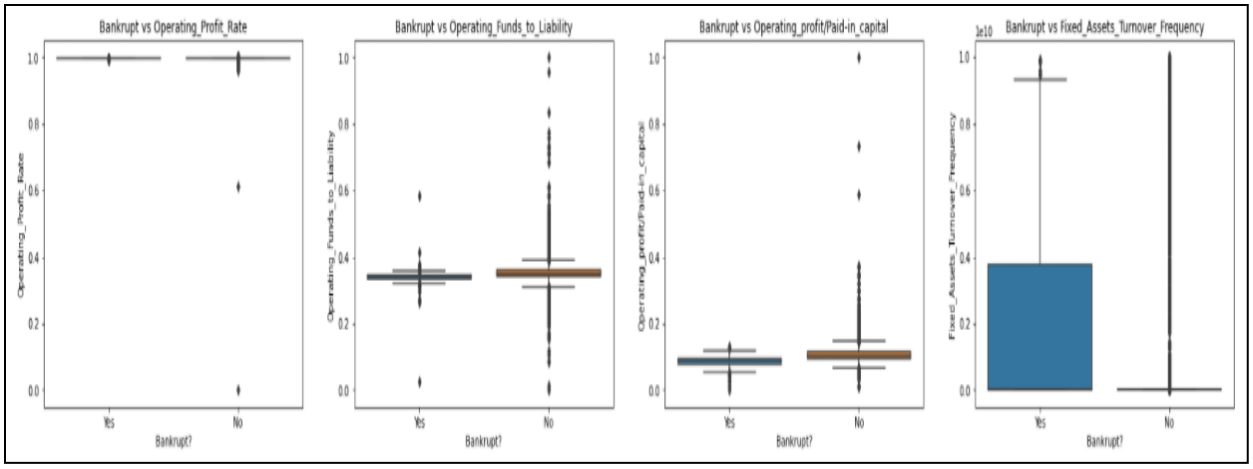

Figure 11. Box plots of Operation measure ratios between Bankrupt vs. Not bankrupt classes

5. Pre-processing and Modelling

There are many common and well known techniques for which can be employed to develop prediction models. In present project, 3 related methods are compared namely, Logistic Regression, Random Forest Classification and Support Vector Machine. To evaluate the performance of the prediction models, 2 metrics have been considered; average prediction accuracy rate which is calculated by how many data samples are correctly classified by the prediction model over a given testing set and precision/recall score which determines the fraction of relevant instances among the retrieved instances and relevant instances that were retrieved. The focus is centered towards recall score as the project objective is to accurately predict the class label 1

5.1. Dealing with imbalance data

In Data wrangling section, it was observed that the target variable 'Bankrupt' column is highly imbalanced where minority class (Class 1) only contain sample of 220 compared to sample of 6599 (Class 0) out of total 6819 companies. This class imbalance problem poses a challenge for predictive modelling as many machine learning algorithms used for classifications were designed around the assumption of an equal number of examples from each class. This may cause poor predictive performance in models, specifically for the minority class. Given that the objective of the project is to predict the bankruptcy of the company (minority class of variable 1), the data-set should be addressed of the imbalance issue.

5.1.1 Synthetic Minority Oversampling Technique

The method applied to resolve the imbalance issue is to apply the technique known as SMOTE (Synthetic Minority Oversampling Technique). Oversampling is achieved by simply duplicating examples from the minority class in the dataset which can balance the class distribution and does not provide any additional information to the model. This technique is utilized to generate synthetic samples from the minority class to imbalanced target problem. The final count of the target variable after applying the technique is 6599 for both 0 and 1 resulting equal balance of samples between Bankrupt and Not bankrupt classes.

```
1    6599
0    6599
Name: Bankrupt?, dtype: int64
```

Figure 12. Final count of the target variable 'Bankrupt' after applying SMOTE.

5.2. Logistic regression

For logistic regression, 2 models have been compared. First model is tested with default parameters and only with oversampling (SMOTE) data set while the second model is optimized with hyperparameter tuning.

5.2.1. Logistic regression base model

With the default setting and oversampling of the target variable, the accuracy of the base model is evaluated to be 61.4% as shown in Figure 13. The precision and recall score for bankrupt class (class 1) is 64% and 50% respectively. The recall performance in the confusion matrix presents 968 samples were correctly evaluated as bankrupt companies out of 1955 predictions. The performance of the base model is reasonably adequate for predicting not bankrupt classes as recall percentage is 73% however, it is not so precise for predicting bankrupt classes and could be improved for higher accuracy.

Accuracy: 0.6138888888888889

	precision	recall	f1-score	support
0	0.60	0.73	0.66	2005
1	0.64	0.50	0.56	1955
accuracy			0.61	3960
macro avg	0.62	0.61	0.61	3960
weighted avg	0.62	0.61	0.61	3960

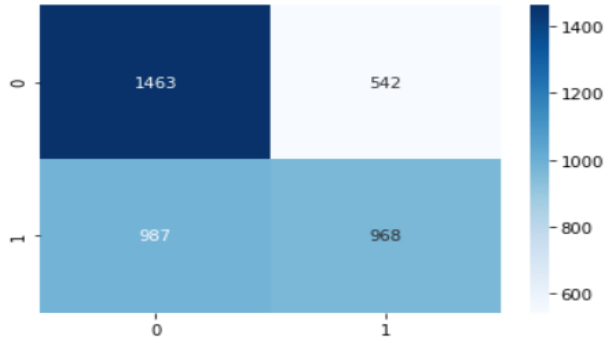

Figure 13. Average accuracy, classficiation report and confusion matrix of precision and recall of Base model

5.2.2. Logistic regression Hyperparameter tuning

In order to find the most optimal parameter for the Logistic Regression, grid search with cross-validation have been applied with pipeline initiating scaling the data-set before fitting through the model. The parameter grid is defined to search through inverse of regularization strength (C) from 0.01 to 100, Regularization specified as Ridge (L2) and solver option between Limited-memore-Broyden-Fletcher-Goldfarb-Shanno(lbfgs) and Library for larger Linear classification.

```
#Defining parameter grid to be evaluated
params = {'classifier__C':[1.0,10,100,0.1,0.01],
          'classifier__penalty':['l2'],
          'classifier__solver':['lbfgs', 'liblinear']}
```

Figure 13. Parameter grid defined for Logistic Regression GridsearchCV

As a result, the most optimal parameter defined by the model was applying C value of 1.0, Regularization as Ridge (L2) and solver option of Limited-memore-Broyden-Fletcher-Goldfarb-Shanno(lbfgs).

5.2.3. Final comparison and ROC curve

The hyperparameter tuned model resulted significant increase in both accuracy and precision and recall score as illustrated in Figure 14. The accuracy improved to 91.2% from 61.4% in the base model and recall value for bankrupt variable increased to 93% as opposed to 50% in the base model and was able to perform 1820 samples out of 1955 prediction.

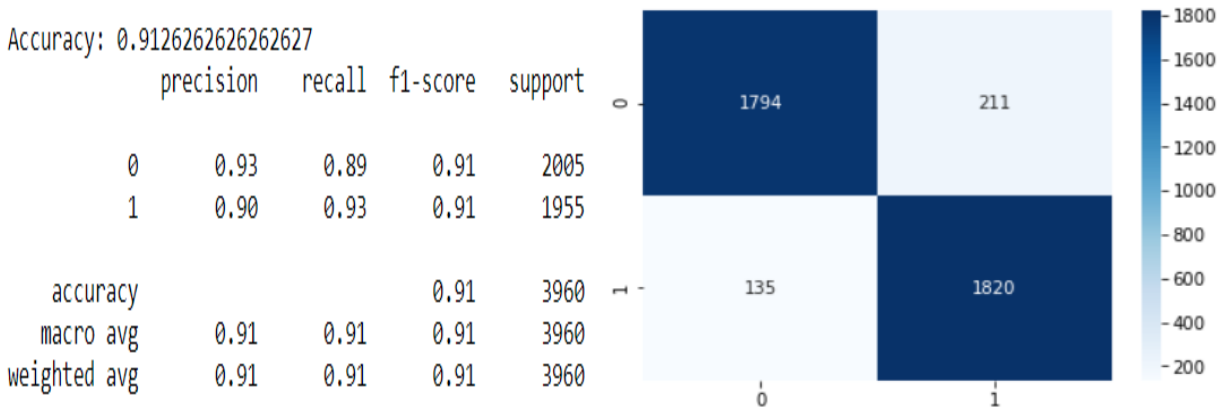

Figure 13. Average accuracy, classfication report and confusion matrix of precision and recall of Hyperparameter tuned model

The receiver operating characteristic (ROC) curve also illustrates a significant difference between the 2 Logistic regression models as true positive rate increased drastically in the hyperparameter tuned model.

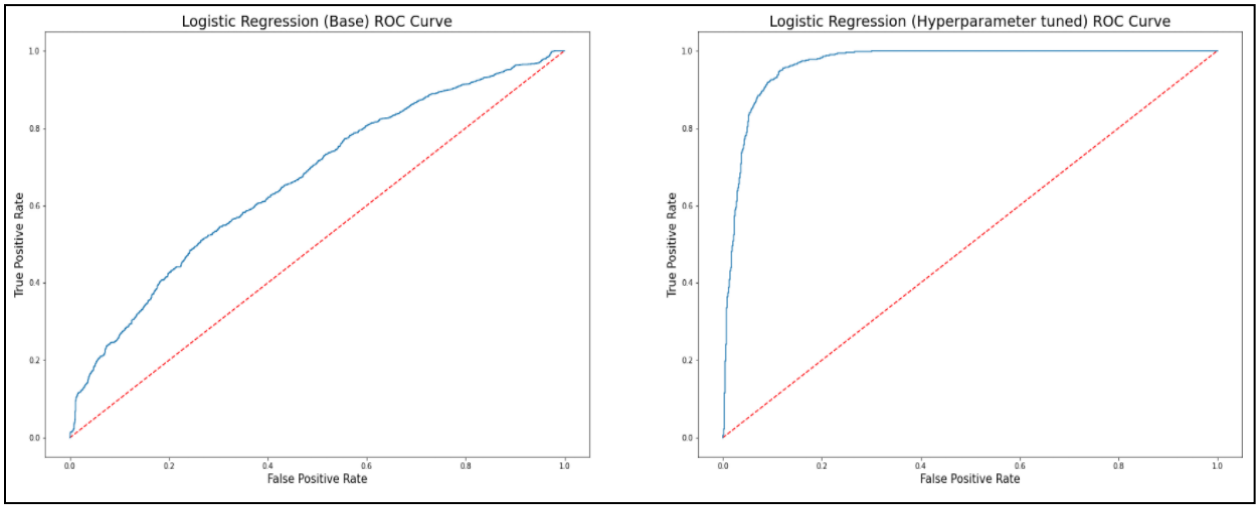

Figure 14. ROC curve comparison between Logistic Regression (Base) vs. Logistic Regression (Hyperparameter Tuned) models

5.3. Random Forest Classification

For Random Forest Classifier a different approach is engaged as opposed to the Logistic Regression model. The first model is tested with default parameters in Random forest with original data-set without oversampling and scaling of the data. The second model is tested with hyperparameter tuning and data-set applied with SMOTE and scaling technique. To find the

most optimal parameters for the model, both Randomized Search and Grid Search were implemented. Random Forest Classifier has extensive range of parameter to consider. To search through the set of hyperparameter values computing for each combination as how Grid Search performs could be computationally complex. By contrast, Randomized search explicitly control the number of parameter combinations that are attempted by selecting random combination with highest score to train the model thus, creating more efficient search iteration.

5.3.1. Random Forest Classifier base model

From the default setting with original data-set, the accuracy of the Random Forest model is evaluated to be 96% as shown in Figure 15. To all appearances the accuracy presents extraordinary results with Random Forest base model, however when examined closely, Precision and Recall score on the bankruptcy class reveals 57% and 10% respectively. The percentage on recall for the bankrupt class indicates that the performance of the model is precise when predicting for surviving companies but is not reliable on prediction of the bankrupt companies. Due to the reason the base model cannot be considered as an appropriate model to perform bankrupt prediction.

Accuracy: 0.9628543499511242

	precision	recall	f1-score	support
0	0.97	1.00	0.98	1968
1	0.57	0.10	0.17	78
accuracy			0.96	2046
macro avg	0.77	0.55	0.58	2046
weighted avg	0.95	0.96	0.95	2046

Figure 15. Average accuracy and classification report of Random Forest Classifier base model

5.3.2. Randomized Search CV

The parameter grid for the randomized search is defined in Figure 16. The Estimator which describes the number of trees to build is set between 200 to 1000 with combination of max depth of the tree from 10 to 100. The criterion which is function to measure the quality of a split is compared between Gini impurity vs. Entropy and maximum number of feature comparison is

evaluated between Auto (simply takes all the features which is logical in every tree) and Sqrt (square root of total number of features in individual run).

```
param = {
    'n_estimators': estimators,
    'max_depth': depth,
    'criterion': ['gini','entropy'],
    'max_features':['auto','sqrt']
```

Figure 16. Parameter grid defined for Random Forest Classifier RandomizedsearchCV

The best parameter estimated through the search is defined with estimator of 555, tree max depth of 50 and max feature as Auto and criterion as Gini impurity. The accuracy assessed with the estimated parameters is resulted to be 97.9% with Precision and Recall score of 96% and 100% for bankrupt prediction as shown in Figure 17. The results present astonishing improvement on the recall score for the bankrupt class from 10% to 100%.

	precision	recall	f1-score	support
0	1.00	0.96	0.98	2005
1	0.96	1.00	0.98	1955
accuracy			0.98	3960
macro avg	0.98	0.98	0.98	3960
weighted avg	0.98	0.98	0.98	3960

Figure 17. Classification report of Random Forest Classifier tuned with RandomizedsearchCV parameters

The rank test scores of the cross validation results can be displayed in data frame to examine on the additional information regarding on the settings as presented in Figure 18. This represents the rank of each combination of the random hyperparameter values defined by the randomized search sorted by its test scores. It can be observed that the parameters which produce 5 best scores are comprised of estimators between 400 to 700, max tree depth between 20, 50 and 70, criterion as Gini impurity and Auto as max feature.

rank_test_score	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_n_estimators	param_max_features	param_max_depth	param_criterion
1	30.731771	0.980944	0.435334	0.005313	555	auto	50	gini
2	33.543887	1.532408	0.488334	0.033865	555	sqrt	50	gini
3	26.455957	0.246322	0.363333	0.006945	466	sqrt	50	gini
4	40.729141	0.314030	0.489665	0.021642	733	auto	70	gini
5	30.343781	0.969560	0.392333	0.036059	466	sqrt	20	gini
6	72.001199	1.630633	0.773375	0.107397	1000	auto	70	entropy
7	52.865824	0.797367	0.554000	0.037480	733	auto	40	entropy
8	13.988406	0.209135	0.173334	0.021172	200	auto	50	entropy

Figure 18. Cross validation results of the Random Forest Classifier randomized search

5.3.3. Grid Search CV

With the narrowed down list of parameter identified through the test scores, these parameters can be searched through Grid Search CV once more with the aim to find more enhanced parameters for the model. The parameter grid settings are described in Figure 19.

```
param = {
    'n_estimators': [400,500,600,700],
    'max_depth': [20,50,70],
    'criterion': ['gini'],
    'max_features':['auto']
```

Figure 19. Parameter grid defined for Random Forest Classifier GridSearchCV

The outcome of the search produced estimator of 500 and max depth of 50. The scores generated through the grid search parameter is almost identical compared to the previous scores as the accuracy score is calculated to be 97.9% and Precision and Recall score as 96% and 100%. This results indicates that there is no significant difference between Random search and Grid Search defined parameters and hence narrowing down the parameter grid is unnecessary for

testing Random Forest model with given data-set. The tuned model of Random Forest Classifier outputs astounding result predicting 1949 samples out of 1955 for the bankrupt class.

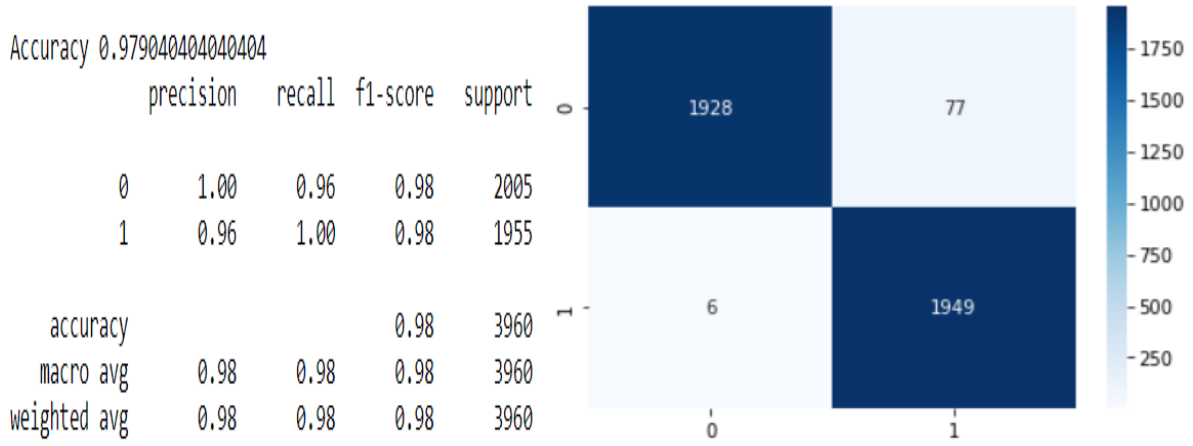

Figure 20. Average accuracy, classfication report and confusion matrix of precision and recall of Random Forest Classifier Hyperparameter tuned model

5.4. Support Vector Machine

The Support Vector Machine (SVM) is tested in similar pattern to Random Forest classifier. It is first computed without oversampling (SMOTE) and scaling of the data, and then optimized through Grid search for most appropriate parameters.

5.4.1. Support Vector Machine base model

The default setting and parameter of the Support Vector Machine outputs very similar results to assessment concluded in Random Forest. In Figure 21, the accuracy of the model yields 96.1% with Precision and Recall score of 0% for bankrupt class. Again the accuracy of the base model is exceptionally high yet, it is not successful in predicting the outcome of the bankrupt variables. As a matter of fact the performance of SVM model is worse than the Random Forest model as the model failed to correctly interpret the bankrupt class and consequently produced recall score of 0%.

```
0.9618768328445748
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      1968
           1       0.00      0.00      0.00        78

    accuracy                           0.96      2046
   macro avg       0.48      0.50      0.49      2046
weighted avg       0.93      0.96      0.94      2046
```

Figure 21. Classification report of Support Vector Machine base model

5.4.2. Grid Search CV

The parameter grid setting for the SVM model is given to evaluate the inverse of regularization strength (C) from 0.1 to 100, Gamma (spread of the kernel) from 0.001 to 1 with Radian Basis Function as gamma parameter. The result of the most optimal parameter search defined by the model was utilizing C value of 10, Gamma value of 0.1.

```
param_grid = {'C': [0.1, 1, 10, 100],
              'gamma': [1, 0.1, 0.01, 0.001],
              'kernel':['rbf']}
```

Figure 22. Parameter grid defined for Support Vector Machine GridSearchCV

The improvement in hyperparameter tuned model is considerable in both precision and recall score as illustrated in Figure 23. The SVM models performed very similar to Random Forest Classifier models as both the precision and recall score increased from 0% to 98% and 99% respectively. The model successfully estimated prediction of 1933 samples out of 1955 from the testing set.

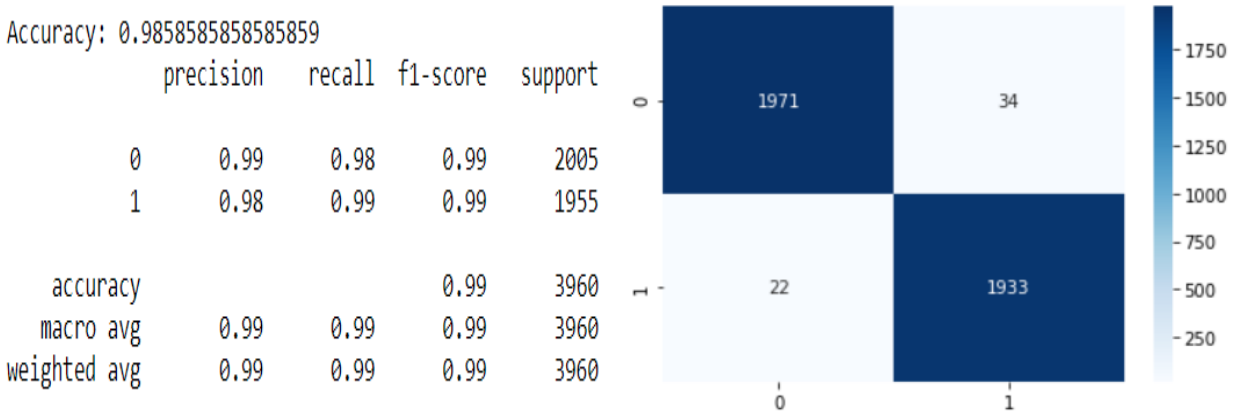

Figure 23. Average accuracy, classfication report and confusion matrix of precision and recall of Support Vector Machine Hyperparameter tuned model

5.5. Final comparison and ROC analysis

The final comparison for the performance of the hyperparameter tuned models is described in Table 1. These results are the scores of the hyperparameter tuned models categorizing only the resulting cases of the bankrupt class (Target variable of 1).

Models	Accuracy	Precision	Recall
Logistic Regression	91.1%	90%	93%
Random Forest Classifier	97.9%	96%	100%
Support Vector Machine	98.6%	98%	99%

Table 1. Final comparison of Accuracy, Precision and Recall scores of the models on target variable of bankrupt class

All 3 models demonstrated remarkable improvement on the scores with synthetic oversampling and scaling of the data-set. The best accuracy is performed by Support Vector Machine resulting 98.6% followed by the Random Forest model with 97.9% and Logistic Regression model in last place with 91.1%. Although the performance obtained with Support Vector Machine has the highest accuracy and precision scores, Random Forest Classifier has the highest Recall score predicting 1949 samples out of 1955 bankruptcy counts from the test data-set. Since the main objective of the project is to accurately predict and classify the bankruptcy, Random Forest Classifier exhibits the best performance out of all the models.

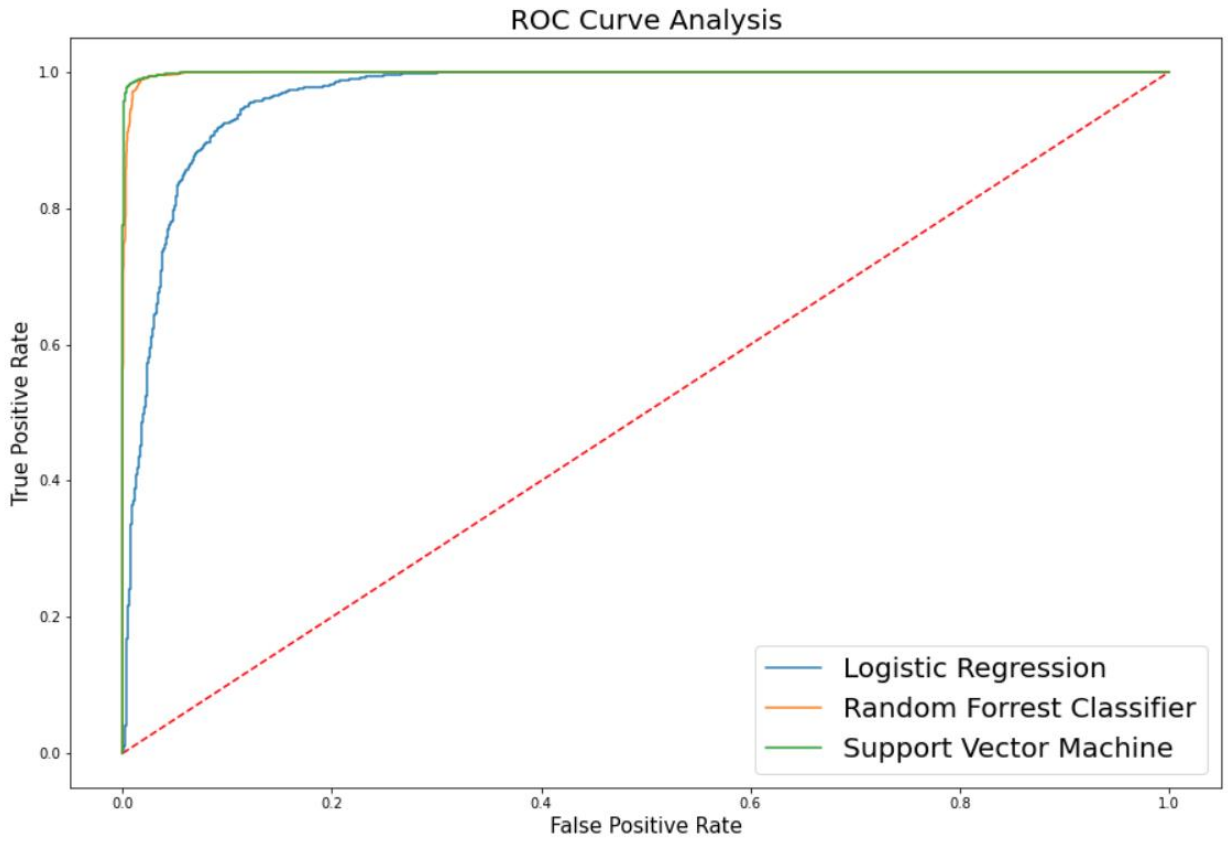

Figure 24. ROC curve comparison between hyperparameter tuned Logistic Regression, Random Forest Classifier and Support Vector Machine

6. Conclusion

The present project pursues on evaluating the combination of different categories of financial and economic ratio data of corporate business regulations to determine the outcome of bankruptcy. In particular, 7 sub-groups of financial ratio are considered namely: solvency, profitability, capital structure ratios, turn-over ratios, cash-flow ratios, growth, and others. Through exploratory data analysis, we discovered that the profitability and solvency are the most critical features affecting the company's financial stability. Furthermore, the largest divergences of the bankrupt and not bankrupt companies were observed in the ratio which classifies assets and liabilities. The surviving companies have higher chance of avoiding bankruptcy due to higher proportion of assets in relation to liabilities and in consequence are able to meet its financial obligations. Through pre-processing and modelling, 3 prediction models were developed with implementation of Synthetic Minority oversampling technique to address

imbalance issue, scaling and hyperparameter tuning for optimization. The Support Vector Machine produced the highest scores for both accuracy and precision however, Random Forest Classifier model provided the highest score for recall score. In conclusion, as the main goal of the project is to identify the variables of bankruptcy class, Random Forest Classifier demonstrates the best performance of bankruptcy prediction.

References

1. Jodi L. Bellovary, Don E. Giacomino and Michael D.Akers, "A Review of Bankruptcy Prediciton Studies: 1930 to Present", *University of Wisconsin-Madison, Marquette University*
2. Javier Parra, Maria E. Perez-Pons, Jorge Gonzalez, "The Importance of Bankruptcy Prediciton in the Advancement of Today's Businesses and Economies", 2020, *Distributed Computing and Artificial Intelligence, 17th International conference.*
3. Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, Guan-An Shih. "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study", *European Journal of Operational Research*
4. M. Dash, H.Liu. "Feature selection for classification", 1997 *Intelligent Data Analysis*