Bankruptcy Prediction based on Financial and Economic ratios

David Song

ABOUT DATA

Financial data based on business regulation of the Taiwan Stock Exchange obtained from Taiwan Economic Journal from the years 1999 to 2009. Financial ratios refers to information related to business condition and prospects and are in general utilized to evaluate and analyze company's performance and financial health.

Data Summary

- Total of 95 input variables(features) representing financial ratios
- 1 Target variable named ‘Bankrupt’: 1 for bankrupt and 0 for not bankrupt
- Features can be categorized into 7 Sub-groups namely: Solvency, Profitability, Capital Structure ratios, Turn over ratios, Cash flow ratios, Growth and others
- Total of 6819 entries of company data
- 93 columns are data type of ‘float’ : Continuous variables
- 3 columns are data type of ‘int’ : Categorical variables

OBJECTIVE

The primary goals of this project is to analyze:

- Which financial ratio are most critical in determining the outcome of the bankruptcy?

- Which Supervised Learning models produce the best performance predicting the likelihood of bankruptcy?

DATA IMBALANCE PROBLEM

TOTAL NUMBER OF SAMPLES : 6819

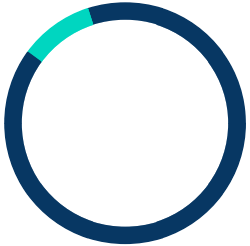
220 BANKRUPT

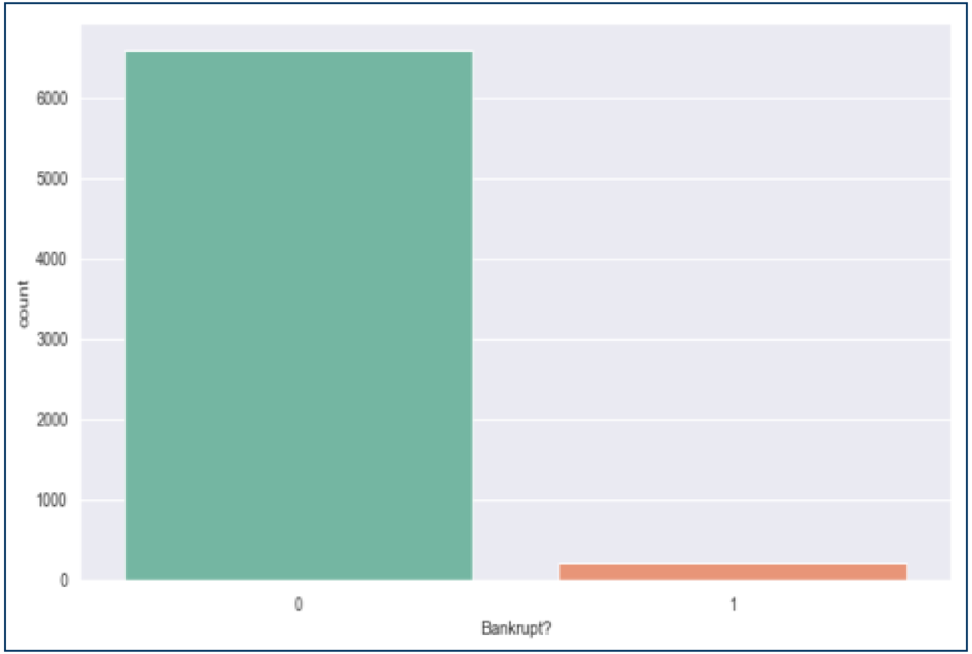
6599 NOT BANKRUPT

PERCENTAGE OF TARGET VARIABLE

3.2%

PERCENTAGE





SCATTER PLOTS

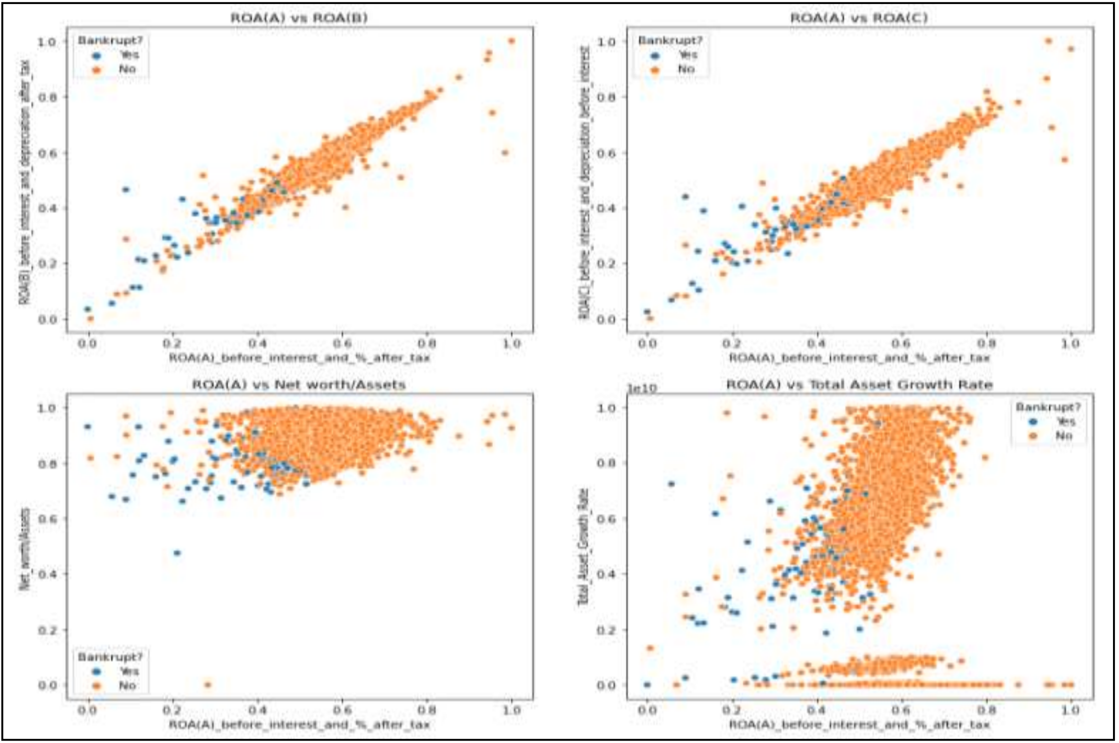
Constant variable:	**Comparison Variables:**
ROA(A) before interest and % after tax return	ROA(B) before interest and depreciation after tax
	ROA(C) before interest and depreciation before interest
	Cash/Total Assets Net profit before tax/Paid in capital
	Net worth/Assets Total Assets Growth Rate Realized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

- 9 features have been randomly selected for analysis for any patterns to be observed.

- ROA(A) is selected as independent variable (x-axis)

- All other variables are arranged to dependent variable for comparison

SCATTER PLOTS (CONTINUE)

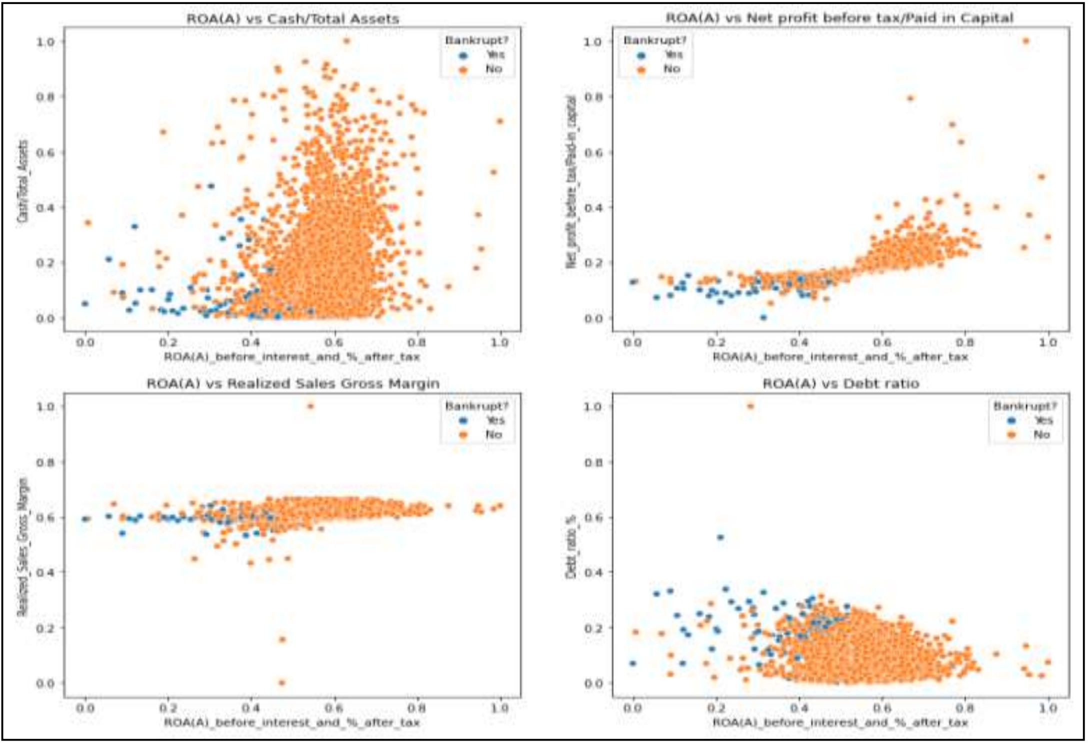


▪Overall scatter plots presents clear difference on Return on Assets (ROA) between target variables. Bankrupt companies show weakness in profitability

▪ROA(B) and ROA(C) reveals most significant difference between target variables

▪Near perfect linear relationship in ROA(B) and ROA(C) presents they are highly correlated to each other.

SCATTER PLOTS (CONTINUE)

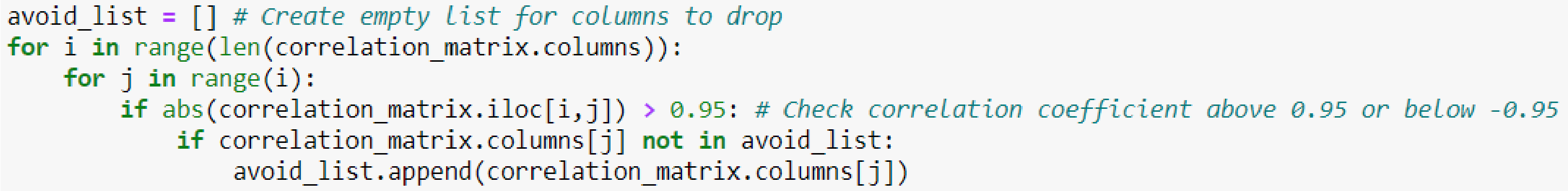


▪Cash/Total Assets ratio values are evidently lower for bankrupt companies

▪Debt ratio values are moderately higher for bankrupt companies. This signifies that chance of bankruptcy increases as debt and liability increases

▪The difference of the target variable on the remaining scatter plots are inconsiderable.

MULTICOLLINEARITY



```
avoid_list = [] # Create empty list for columns to drop
for i in range(len(correlation_matrix.columns)):
    for j in range(i):
        if abs(correlation_matrix.iloc[i,j]) > 0.95: # Check correlation coefficient above 0.95 or below -0.95
            if correlation_matrix.columns[j] not in avoid_list:
                avoid_list.append(correlation_matrix.columns[j])
```

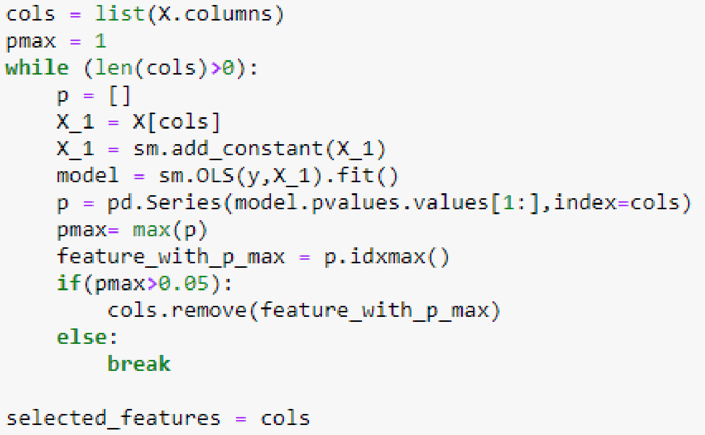
MULTICOLLINEARITY CHECK

- High correlation between the independent variables in the regression analysis could impact overall interpretation of the result.
- The features that have correlation coefficient above 0.95 or below -0.95 are dropped to avoid multicollinearity in the variables.

FEATURE SELECTION

▪Backward Elimination method was considered as feature selection technique to reduce down the least significant features.

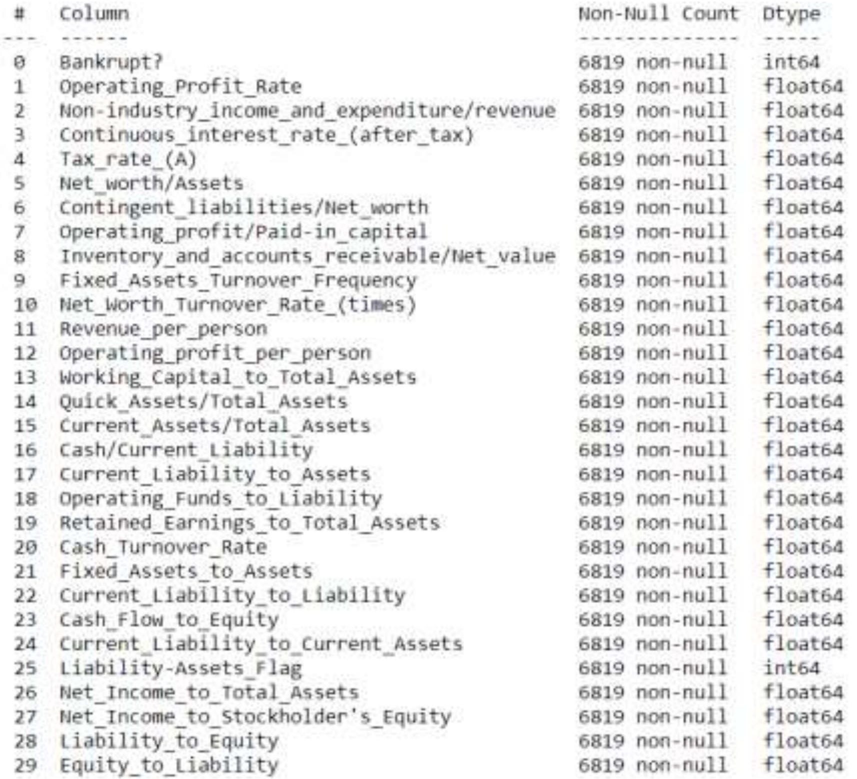
▪The final count of columns filtered through both Multicollinearity check and Feature section is 29.



```
cols = list(X.columns)
pmax = 1
while (len(cols)>0):
    p = []
    X_1 = X[cols]
    X_1 = sm.add_constant(X_1)
    model = sm.OLS(y,X_1).fit()
    p = pd.Series(model.pvalues.values[1:],index=cols)
    pmax= max(p)
    feature_with_p_max = p.idxmax()
    if(pmax>0.05):
        cols.remove(feature_with_p_max)
    else:
        break

selected_features = cols
```

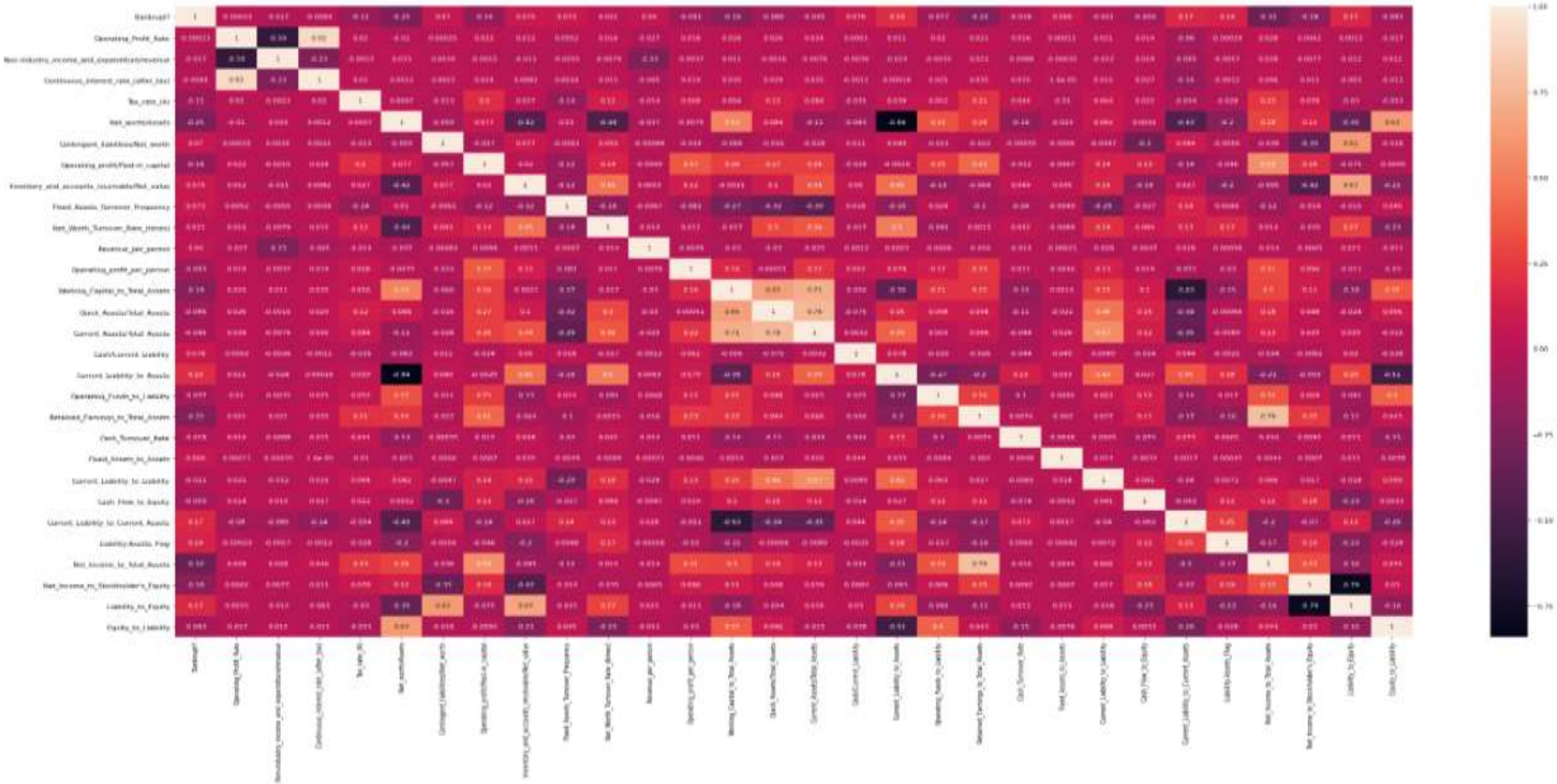
BACKWARD ELIMINATION METHOD



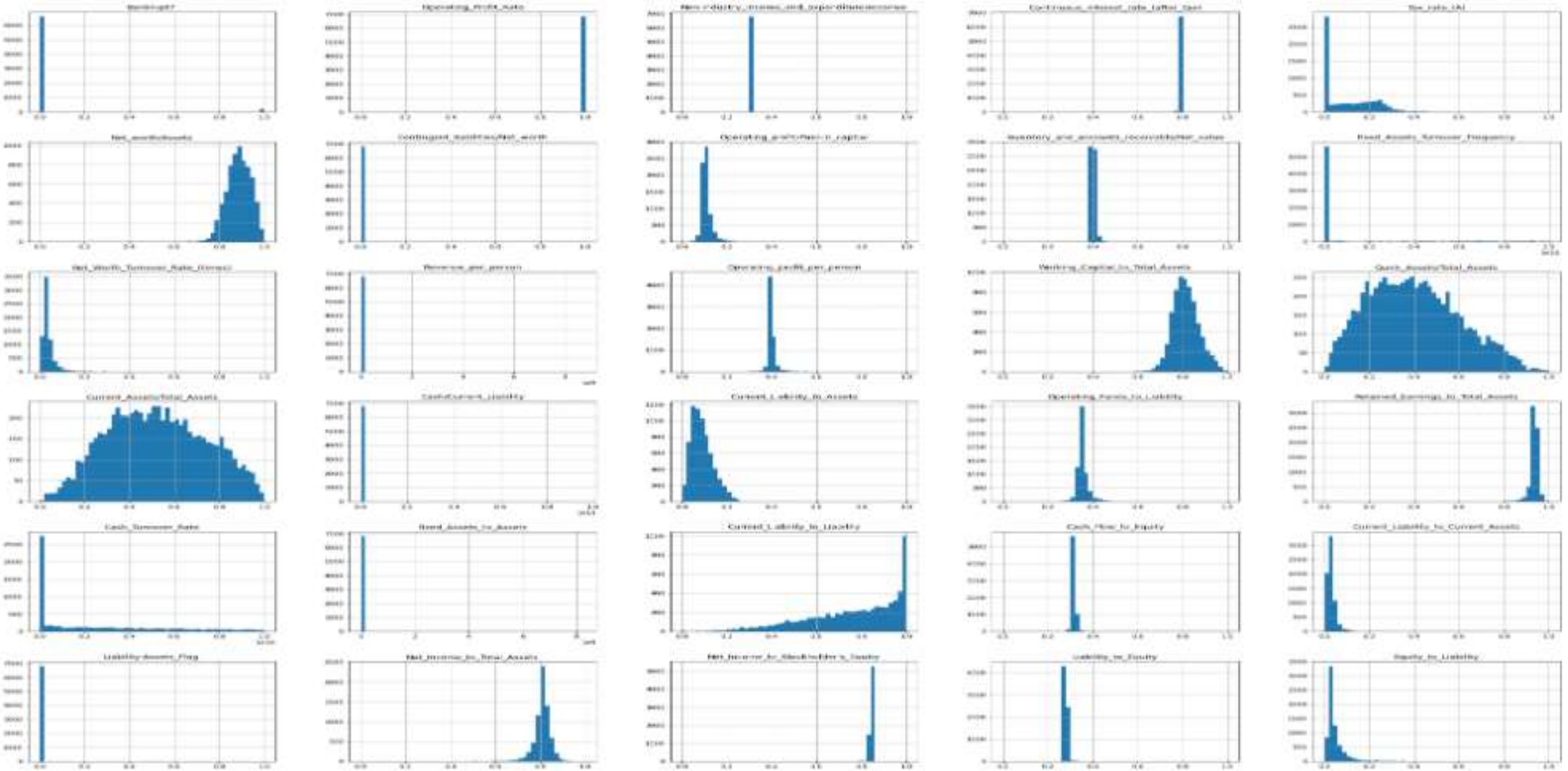
```
 #   Column                                    Non-Null Count  Dtype  
---  ------                                    --------------  -----  
 0   Bankrupt?                                 6819 non-null   int64  
 1   Operating_Profit_Rate                     6819 non-null   float64
 2   Non-industry_income_and_expenditure/revenue 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
```

SELECTED FEATURES

HEATMAP


Bankrupt?
Operating_Profit_Rate
Non-industry_income_and_expenditure/revenue
Continuous_interest_rate_(after_tax)
Tax_rate_(A)
Net_worth/Assets
Contingent_liabilities/Net_worth
Operating_profit/Paid-in_capital
Inventory_and_accounts_receivable/Net_value
Fixed_Assets_Turnover_Frequency
Net_Worth_Turnover_Rate_(times)
Revenue_per_person
Operating_profit_per_person
Working_Capital_to_Total_Assets
Quick_Assets/Total_Assets
Current_Assets/Total_Assets
Cash/Current_Liability
Current_Liability_to_Assets
Operating_Funds_to_Liability
Retained_Earnings_to_Total_Assets
Cash_Turnover_Rate
Fixed_Assets_to_Assets
Current_Liability_to_Liability
Cash_Flow_to_Equity
Current_Liability_to_Current_Assets
Liability-Assets_Flag
Net_Income_to_Total_Assets
Net_Income_to_Stockholder's_Equity
Liability_to_Equity
Equity_to_Liability
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75

HISTOGRAM

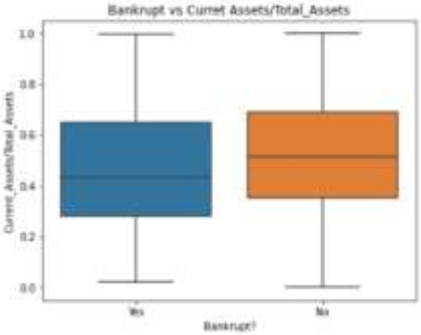

Bankrupt?
Operating_Profit_Rate
Non-industry_income_and_expenditure/revenue
Continuous_interest_rate_(after_tax)
Tax_rate_(A)
Net_worth/Assets
Contingent_liabilities/Net_worth
Operating_profit/Paid-in_capital
Inventory_and_accounts_receivable/Net_value
Fixed_Assets_Turnover_Frequency
Net_Worth_Turnover_Rate_(times)
Revenue_per_person
Operating_profit_per_person
Working_Capital_to_Total_Assets
Quick_Assets/Total_Assets
Current_Assets/Total_Assets
Cash/Current_Liability
Current_Liability_to_Assets
Operating_Funds_to_Liability
Retained_Earnings_to_Total_Assets
Cash_Turnover_Rate
Fixed_Assets_to_Assets
Current_Liability_to_Liability
Cash_Flow_to_Equity
Current_Liability_to_Current_Assets
Liability-Assets_Flag
Net_Income_to_Total_Assets
Net_Income_to_Stockholder's_Equity
Liability_to_Equity
Equity_to_Liability

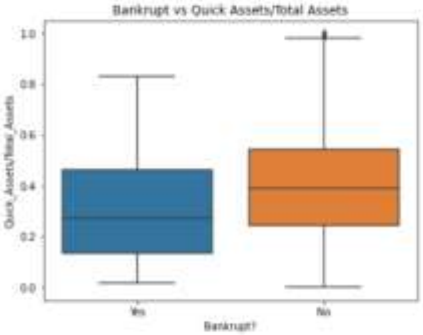
DIVIDING CATEGORIES OF NEW FEATURES

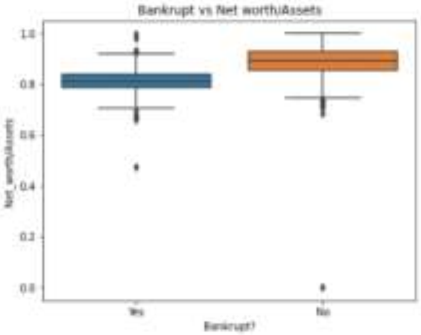
From the new columns, 16 ratios are selected and separated into 4 Major categories:

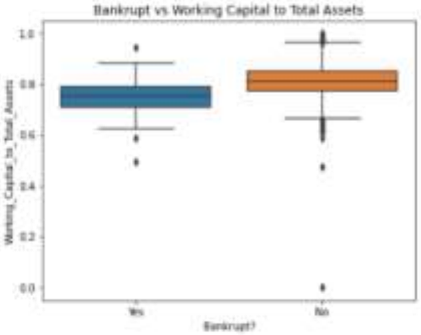
- Solvency : Asset ratios
  - Current Assets/Total Assets
  - Quick Assets/Total Assets
  - Net Worth/Assets
  - Working Capital/Total Assets
- Solvency : Liability ratios
  - Current Liability to Assets
  - Current Liability to Current Assets
  - Equity to Liability
  - Current Liability to Liability
- Turnover and Cash flow ratios
  - Net Income to Total Assets
  - Cash turnover rate
  - Retained Earnings to Total Assets
  - Cash Flow to Equity
- Operation Measures
  - Operating Profit rate
  - Operating Funds to Liability
  - Operating Profit/Paid in Capital
  - Fixed Assets Turnover Frequency

BOX PLOTS (SOLVENCY : ASSETS)





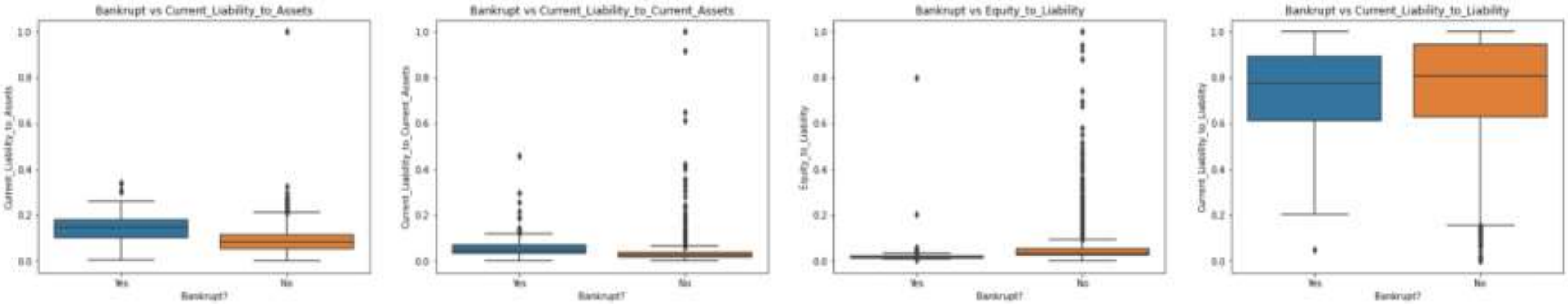




Asset ratios are the features which presents percentage of the company's assets financed by creditors

- Overall, there is noticeable difference between Bankrupt vs. Surviving companies as surviving companies extends higher measurement for all of the Asset ratios.
- The plots indicates that surviving companies are more capable of generating cash to pay off debts and meet its financial obligations thus maintaining the financial strength to avoid bankruptcy

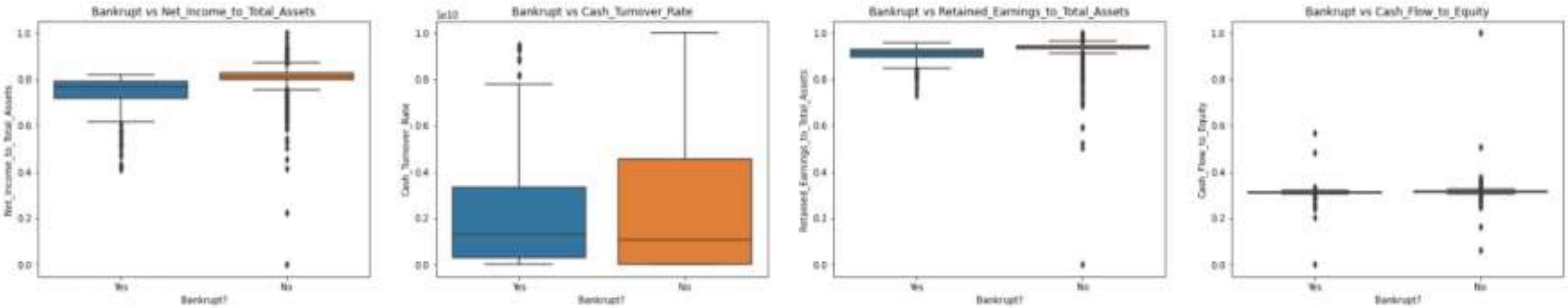
BOX PLOTS (SOLVENCY : LIABILITY)



Liability ratios are the features that define the financial obligation of the company

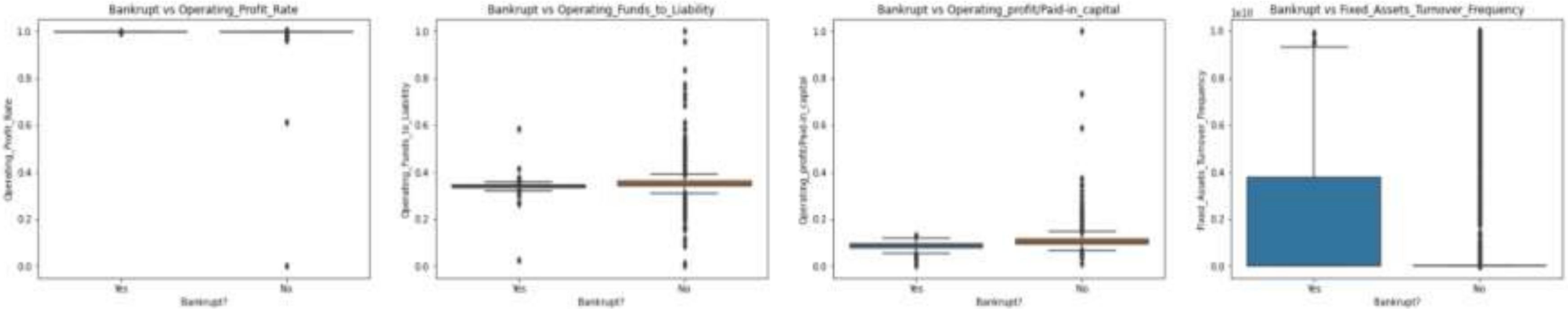
- As expected, bankrupt companies are comprised of higher ratios of liabilities compared to surviving companies
- Bankrupt companies have higher measurement in Current Liability to Assets & Current Liability to Current Assets ratios
- Observation presents that surviving companies exhibits greater strength in financial stability and are able to avoid bankruptcy due to proportionally smaller liabilities in respect to assets

BOX PLOTS (TURNOVER AND CASH FLOW RATIOS)



- Net Income to Total Assets and Retained Earnings to Total Assets display moderately higher values for the surviving companies.
- The pattern signifies 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.

BOX PLOTS (OPERATION MEASURES)



- ▪Operation measure category portrays no difference between the 2 classes. This insignificance in the plots interprets that operation measuring ratios are not a critical contributing factor in determining bankruptcy.

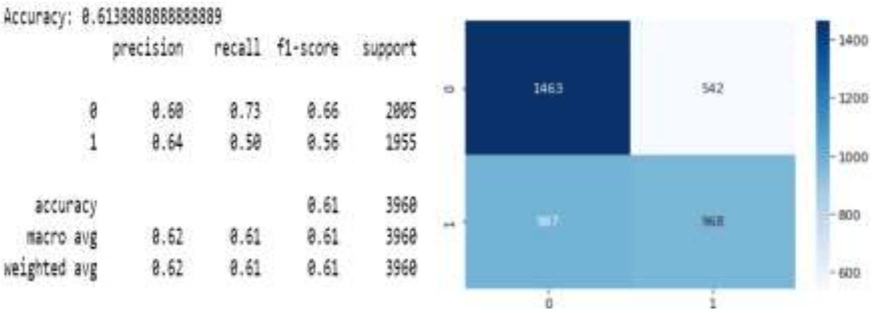
MODELLING

- 3 Supervised Learning Models are considered:
  1) Logistic Regression
  2) Random Forest Classification
  3) Support Vector Machine

- All models are be compared between the Base Model vs. Hyperparameter Tuned Model

- Binary Classification problem | “Bankrupt” vs. “Not Bankrupt”
  - Bankrupt companies are labelled as 1
  - Not Bankrupt companies are labelled as 0

LOGISTIC REGRESSION

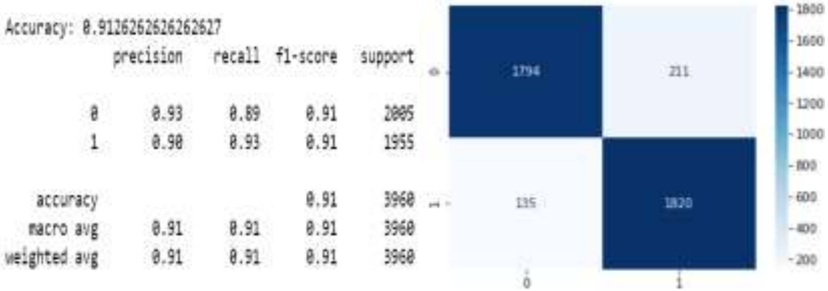


```
Accuracy: 0.6138888888888889
              precision    recall  f1-score   support

           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
```

BASE MODEL



```
Accuracy: 0.9126262626262627
              precision    recall  f1-score   support

           0       0.93      0.89      0.91      2005
           1       0.90      0.93      0.91      1955

    accuracy                           0.91      3960
   macro avg       0.91      0.91      0.91      3960
weighted avg       0.91      0.91      0.91      3960
```

HYPERPARAMETER TUNED

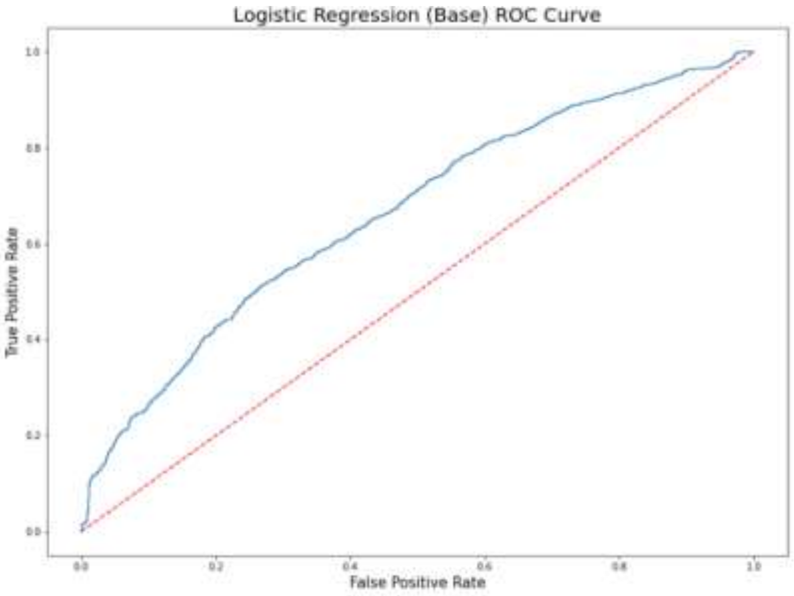
OPTIONS:

- ❑ Default parameters
- ❑ Synthetic Minority Oversampling Technique applied to solve imbalance problem

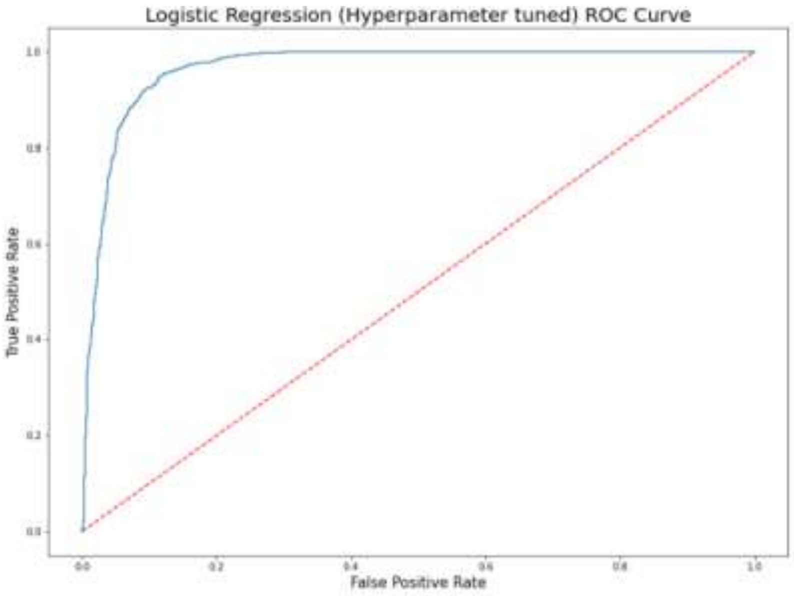
OPTIONS:

- ❑ Inverse of regularization strength(C) = 1.0
- ❑ Regularization = Ridge (L2)
- ❑ Solver option = lbfgs
- ❑ Synthetic Minority Oversampling Technique applied
- ❑ Scaling of Data-set

LOGISTIC REGRESSION (CONTINUED)

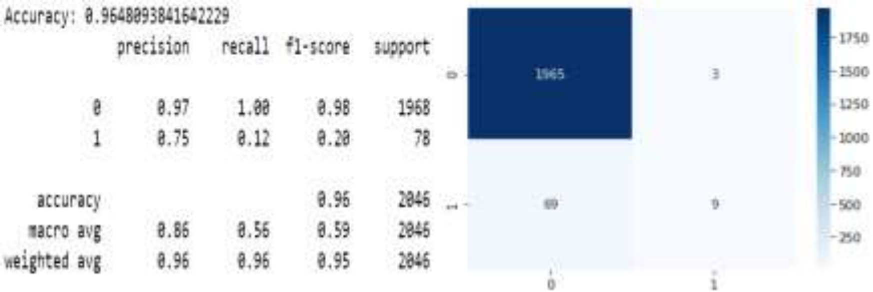


Logistic Regression (Base) ROC Curve



Logistic Regression (Hyperparameter tuned) ROC Curve

RANDOM FOREST CLASSIFIER

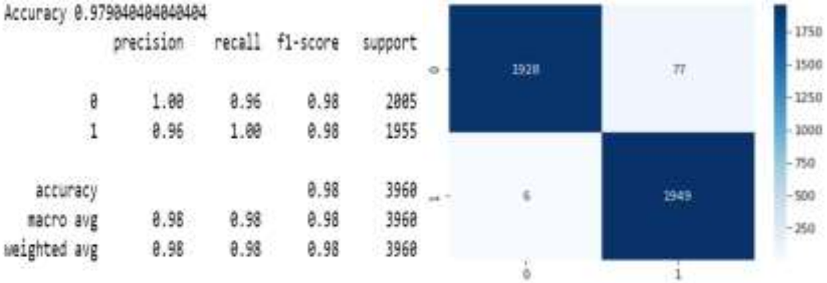


```
Accuracy: 0.9648093841642229
              precision    recall  f1-score   support

           0       0.97      1.00      0.98      1968
           1       0.75      0.12      0.20        78

    accuracy                           0.96      2046
   macro avg       0.86      0.56      0.59      2046
weighted avg       0.96      0.96      0.95      2046
```

BASE MODEL



```
Accuracy 0.9790404040404040
              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
```

HYPERPARAMETER TUNED

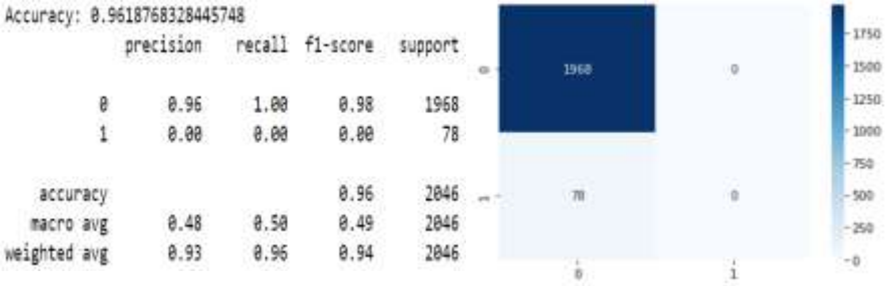
OPTIONS:

- ❑ Default parameters
- ❑ Original Data-set

OPTIONS:

- ❑ Estimators = 500
- ❑ Max tree depth = 50
- ❑ Max feature = Auto
- ❑ Criterion = Gini Impurity
- ❑ Synthetic Minority Oversampling Technique applied
- ❑ Scaling of Data-set

SUPPORT VECTOR MACHINE

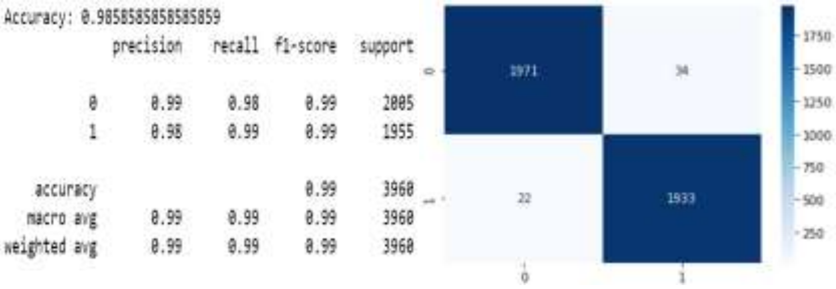


```
Accuracy: 0.9618768328445748
              precision    recall  f1-score   support

           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
```

BASE MODEL



```
Accuracy: 0.9858585858585859
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      2005
           1       0.98      0.99      0.99      1955

    accuracy                           0.99      3960
   macro avg       0.99      0.99      0.99      3960
weighted avg       0.99      0.99      0.99      3960
```

HYPERPARAMETER TUNED

OPTIONS:

- ❑ Default parameters
- ❑ Original Data-set

OPTIONS:

- ❑ Inverse of regularization strength (C) = 10
- ❑ Gamma = 0.1
- ❑ Kernel= Radian Basis Function (rbf)
- ❑ Synthetic Minority Oversampling Technique applied
- ❑ Scaling of Data-set

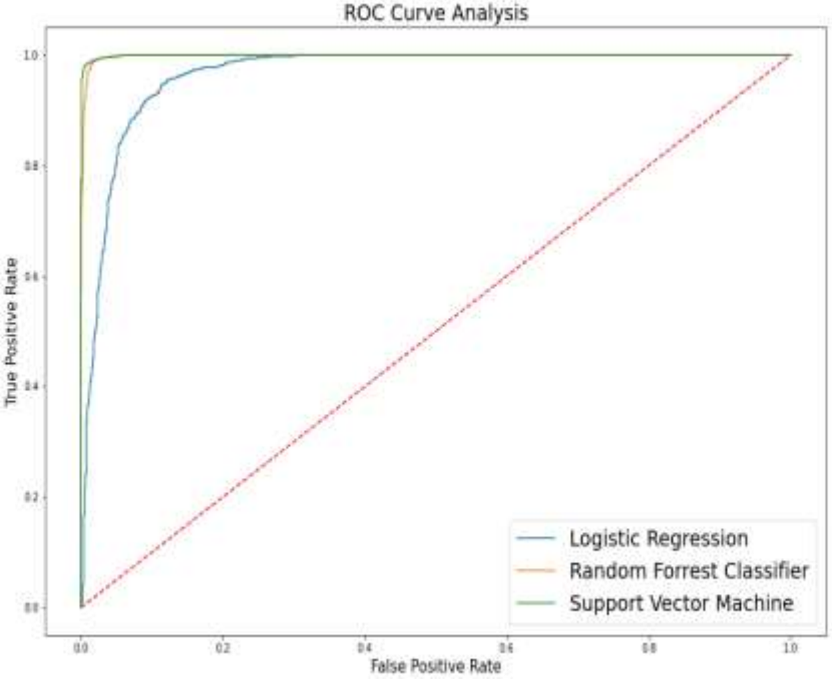
MODEL COMPARISON

Models	Accuracy	Precision	Recall	Recall Prediction
Logistic Regression	91.1%	90%	93%	1820/1955
Random Forest Classifier	97.9%	96%	100%	1949/1955
Support Vector Machine	98.6%	98%	99%	1933/1955

HYPERPARAMETER TUNDED MODEL RESULTS (TARGET VARIABLE OF 1 ONLY)

❑ Support Vector Machine produces the highest accuracy score however,

❑ Random Forest Classifier has the highest Recall score predicting 1949 samples out of 1955 bankruptcy counts from the test data-set



CONCLUSION

Which financial ratio are most critical in determining the outcome of the bankruptcy?

- ❖Profitability and Solvency are most critical features affecting company’s financial stability
- ❖Asset and Liability ratios has the largest divergences between Bankrupt vs. Not Bankrupt companies

Which Supervised Learning models produce the best performance predicting the likelihood of bankruptcy?

- ❖ Random Forest Classifier demonstrates the best performance of Bankruptcy prediction
- ❖ Recall score of 100% | Prediction: 1949 out of 1955