December 2021

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Financial Risk Analytics

Business Report

Thakur Arun Singh

Contents

[Problem 1: 2](#_Toc90239216)

[Problem 1.8 3](#_Toc90239217)

[Problem 1.9 6](#_Toc90239218)

[Problem 1.10 9](#_Toc90239219)

[Problem 1.11 11](#_Toc90239220)

[Problem 1.12 14](#_Toc90239221)

[Problem 1.13 19](#_Toc90239222)

[Problem 2.1 19](#_Toc90239223)

[Problem 2.2 20](#_Toc90239224)

[Problem 2.3 21](#_Toc90239225)

[Problem 2.4 22](#_Toc90239226)

[Problem 2.5 23](#_Toc90239227)

## Problem 1:

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

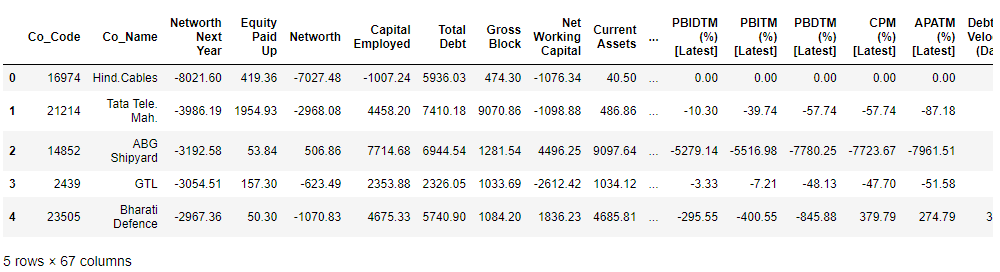
Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

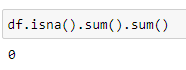
**Exploratory data analysis**

Dataset has 67 variables of which 63 are of float data type, 3 are integer type and 1 is object type.

The head of the dataset is as below:



The data has 3586 Rows and 67 Columns. No duplicate data is present in the data set.

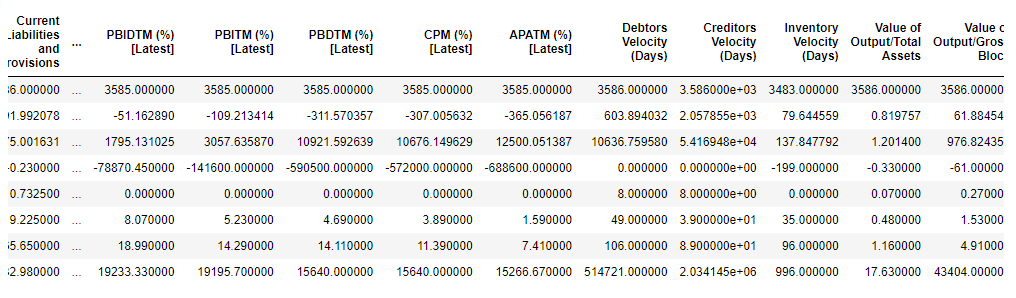


The data has 3586 Rows and 67 Columns. No duplicate data is present in the data set.

We dropped unrequired columns like Co\_Code and Co\_Name since they do not add value to the analysis.

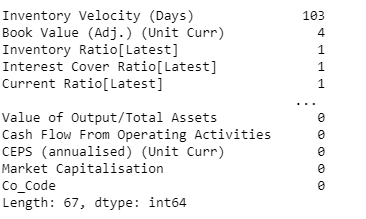
Descriptive statistics / 5 point summary is shown below.





The values of mean, standard deviation, minimum and maximum, 25th, 50th and 75th percentile are mentioned in the above tables.

Next we checked for null values.



Further details on missing values is covered under 1.2

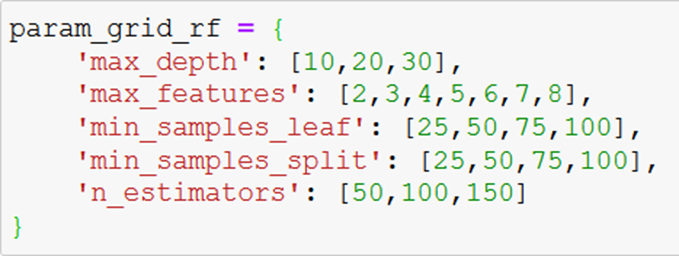
### Problem 1.8

**Build a Random Forest Model on Train Dataset. Also showcase your model building approach**

**Resolution:**

We have created two models using Random Forest one without SMOTE data and one with SMOTE data. We did not scale the data for Random Forest as tree based models are not distance based models and can handle varying ranges of features.

We also used GridSearchCV for hyper parameter tuning. PFB grid which was used, same grid was used for both the models i.e. with and without smote to maintain uniformity.



A custom function named apply\_evl which takes below arguments was created. This function is able to handle modeling for different type of models and returns the performance metrics as the output. The same function has been used to build and evaluate the different models here.



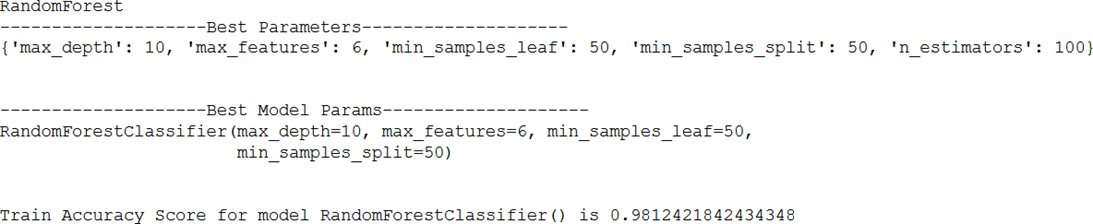
**name** --> Name of the model

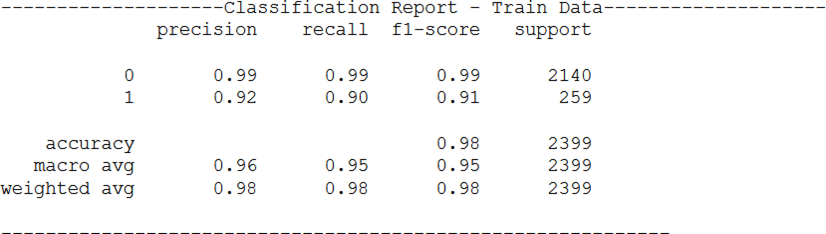
**model** --> Model object which is created and passed to the function

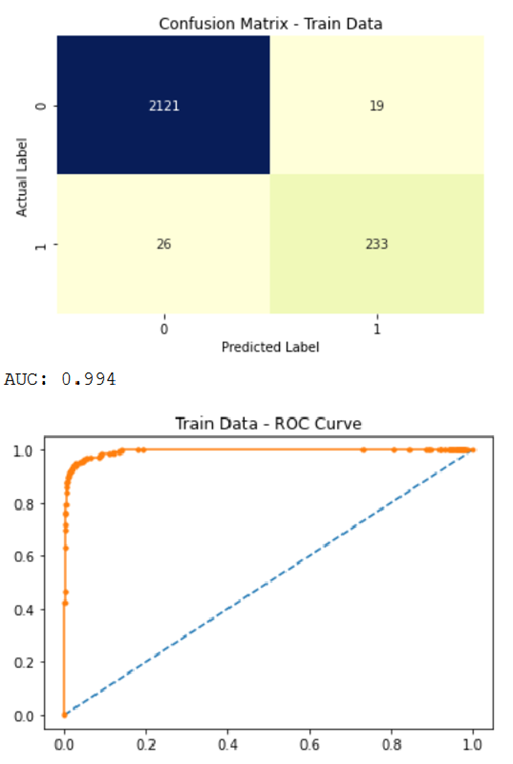
**pram\_grid** --> this takes the grid as a dictionary object. This can also be passed as *None* in case a grid search is not required.

**X\_Train, X\_test,y\_train,y\_test** --> split up data set, in 67:33 ratio with seed as 42 as per project notes.

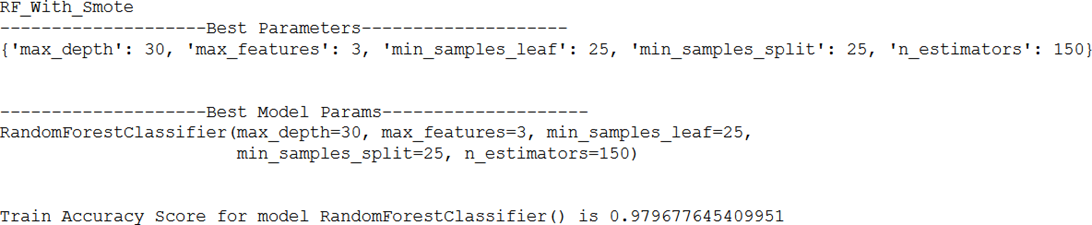
Below are the performance metrics on **train data** for random forest model **without smote data.**

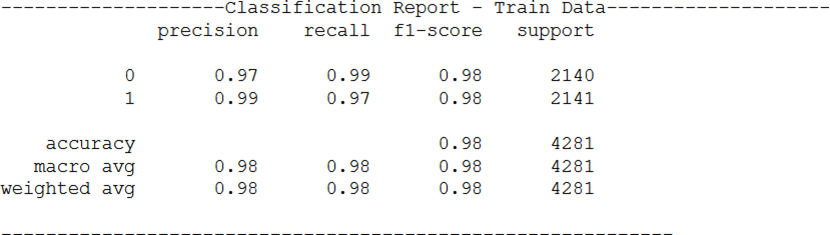


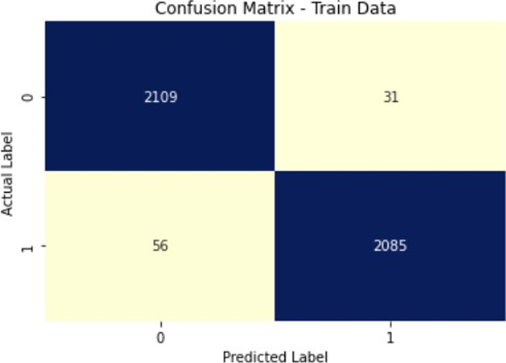


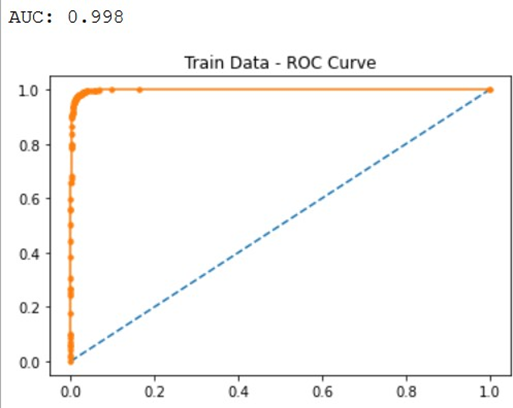


Below are the performance metrics on train data for random forest model with smote data.







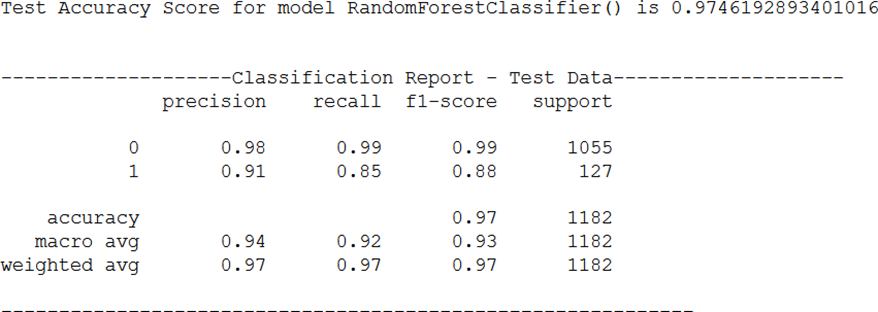


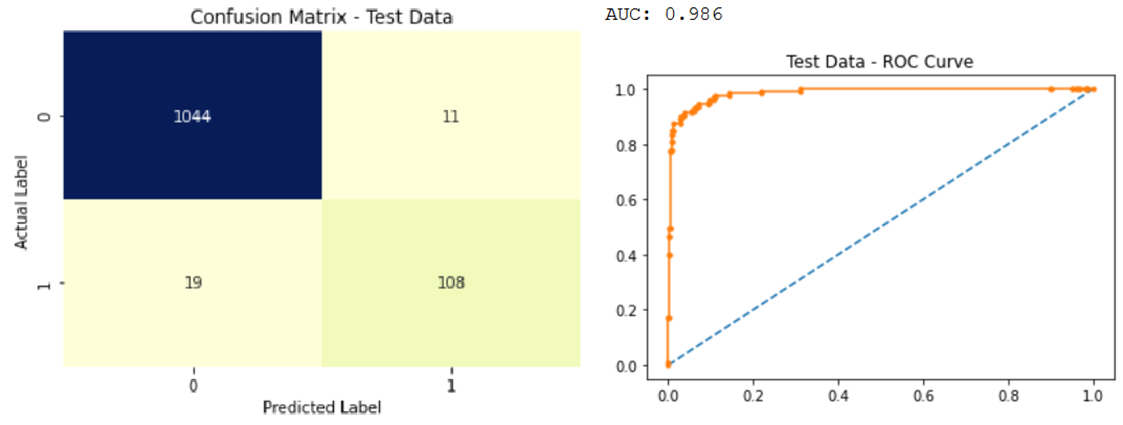
### Problem 1.9

Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model

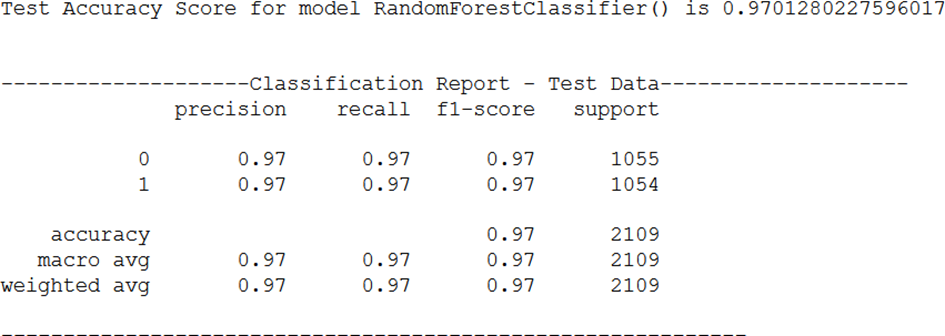
**Resolution:**

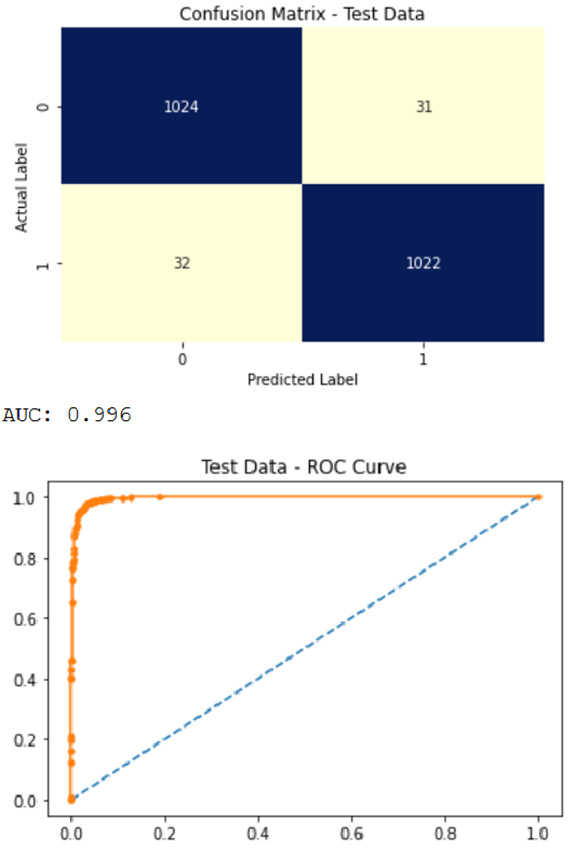
Both the random forest models were then evaluated on the test data. Below are the results for **test data** for random forest model **without smote data.**



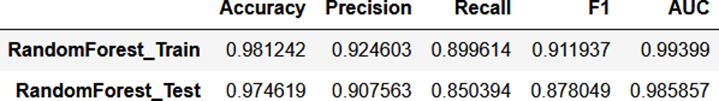


Below are the performance metrics on **test data** for random forest model **with smote data.**

****

****

Below are the comparison of various evaluation metrics in tabular format for both the models i.e. with and without smote.





We can see in the above comparison that even though Random Forest without SMOTE has higher accuracy, but Random Forest with SMOTE surpasses the other model significantly in terms of Precision, Recall, F1 score as well as AUC score.

Also there is not a significant difference in the accuracies of both the models.

Hence we can say Random Forest with SMOTE data is the better of the two models with 97.01% accuracy, 97.05% precision, 96.96% recall and a recall and F1 score of 0.9701 and 0.996038 respectively.

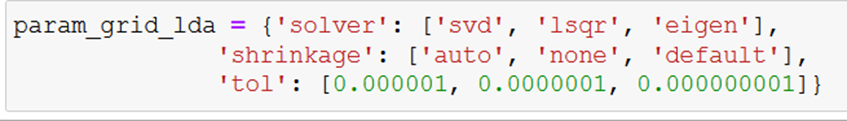
### Problem 1.10

Build a LDA Model on Train Dataset. Also showcase your model building approach

**Resolution:**

we have created two models using LDA, one without SMOTE data and one with SMOTE data. We did not scale the data for LDA as it finds it's coefficients using the variation between the classes, hence scaling doesn't matter.

We also used GridSearchCV for hyper parameter tuning. PFB the grid which was used. Same grid was used for both the models i.e. with and without smote to maintain uniformity.



A custom function named **apply\_evl** which takes below arguments was created. This function is able to handle modeling for different type of models and returns the performance metrics as the output. The same function has been used to build and evaluate the different models here.



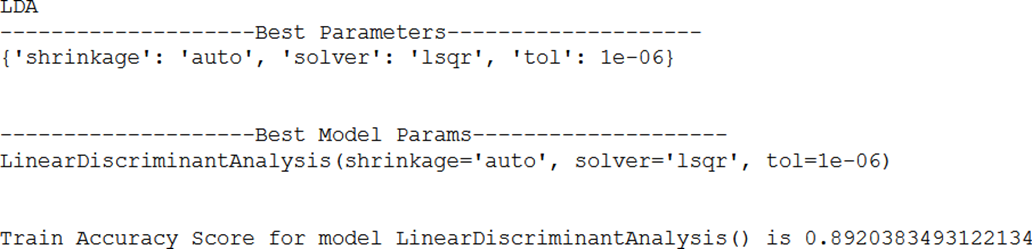
**name** --> Name of the model

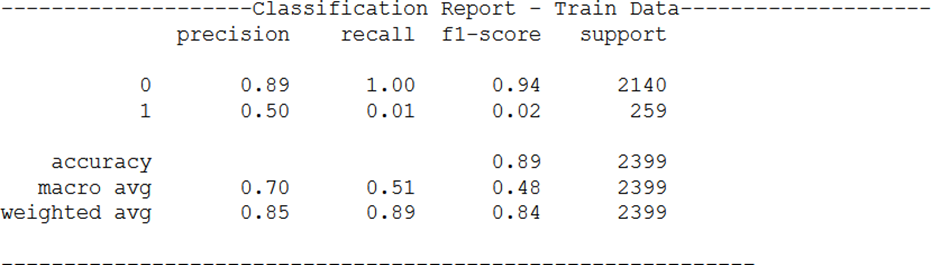
**model** --> Model object which is created and passed to the function

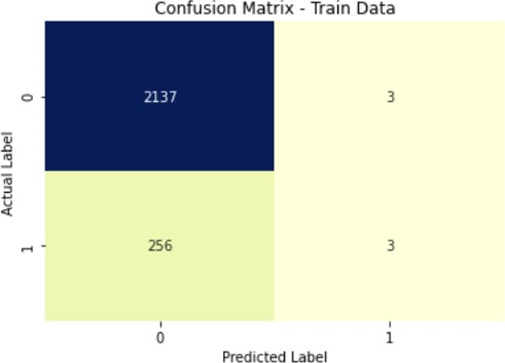
**pram\_grid** --> This takes the grid as a dictionary object. This can also be passed as *None* in case a grid search is not required.

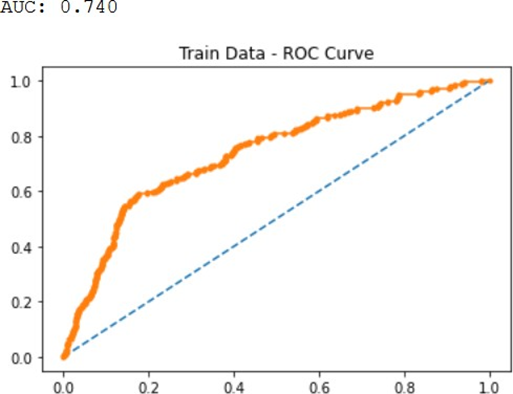
**X\_Train, X\_test,y\_train,y\_test** --> split up data set, in 67:33 ratio with seed as 42 as per project notes.

Below are the performance metrics on **train data** for LDA model **without smote data.**



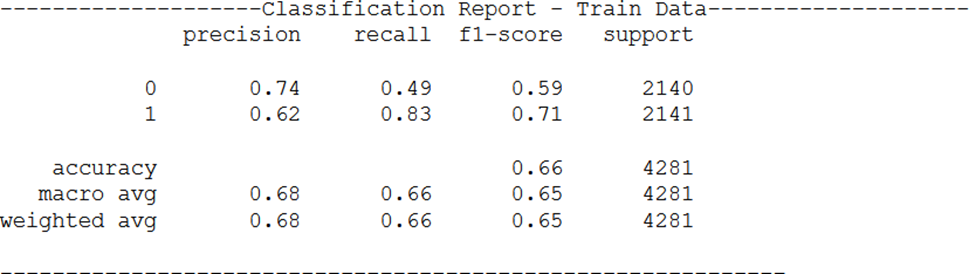


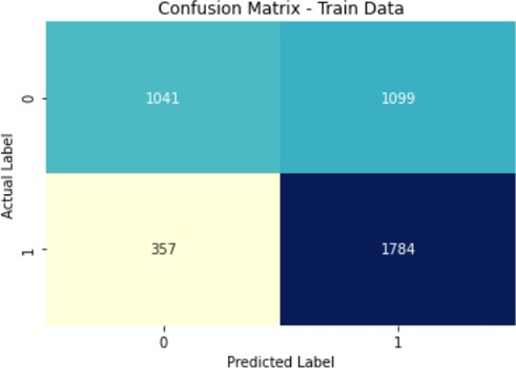


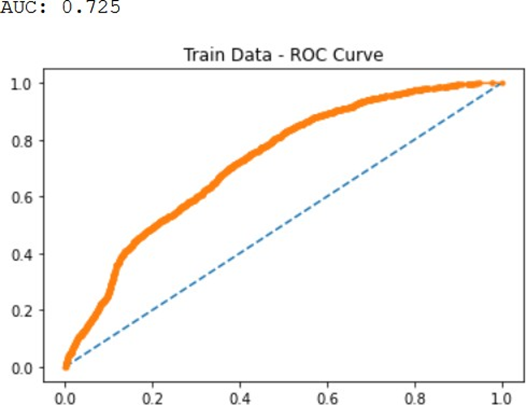


Below are the performance metrics on **train data** for LDA model **with smote data.**







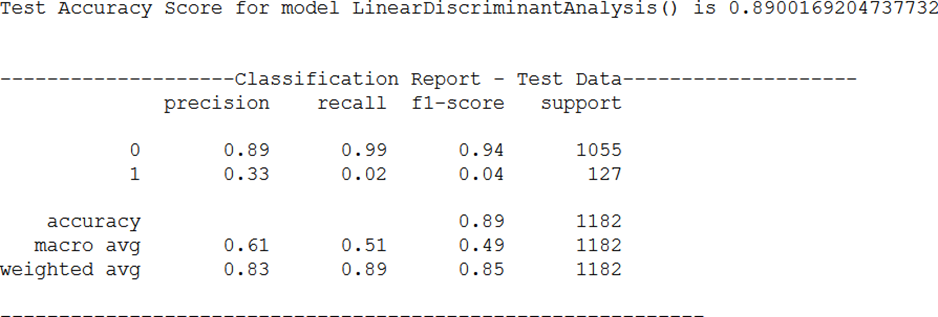


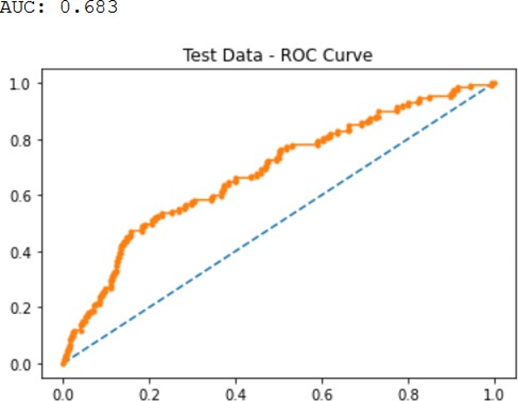
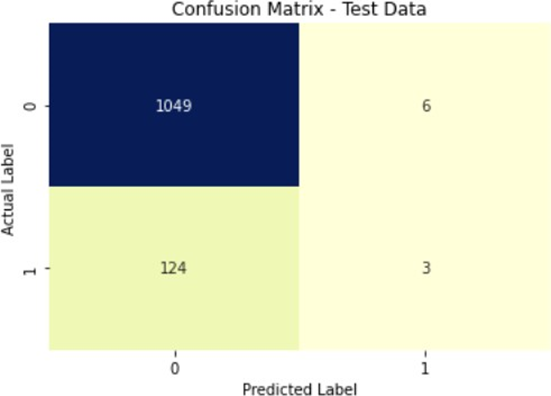
### Problem 1.11

Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model

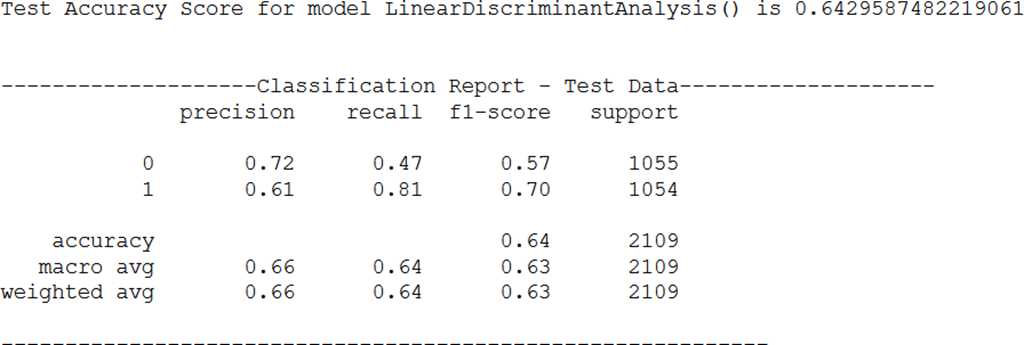
**Resolution:**

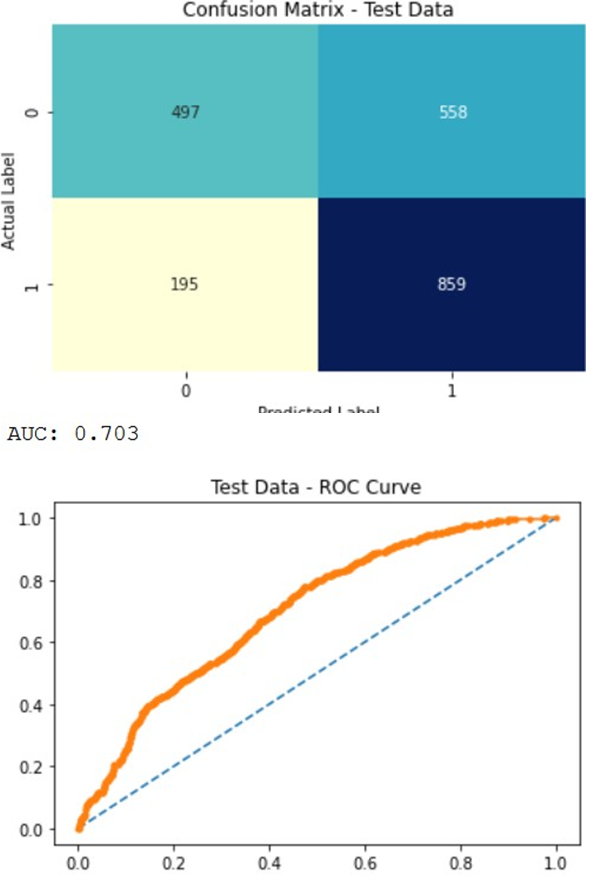
Both the LDA models were then evaluated on the test data. Below are the results for **test data** for LDA model **without smote data.**

****

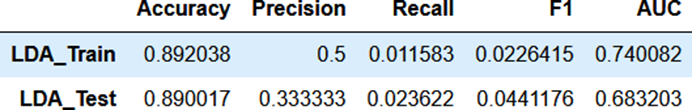
****

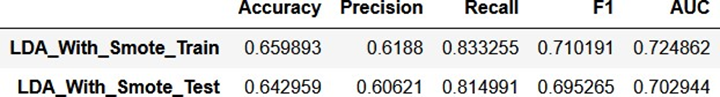
Below are the performance metrics on **test data** for LDA model with **smote data.**





Below are the comparison of various evaluation matrics in tabular format for both the models i.e. with and without smote.

****



Comparing the two models, we can see LDA model without SMOTE data has higher accuracy on test data. However it has really bad precision, recall and F1 scores compared to the other model.

On purely accuracy perspective model without smote data performs better. But in real world scenario model with smote will perform better given better recall and precision values.

### Problem 1.12

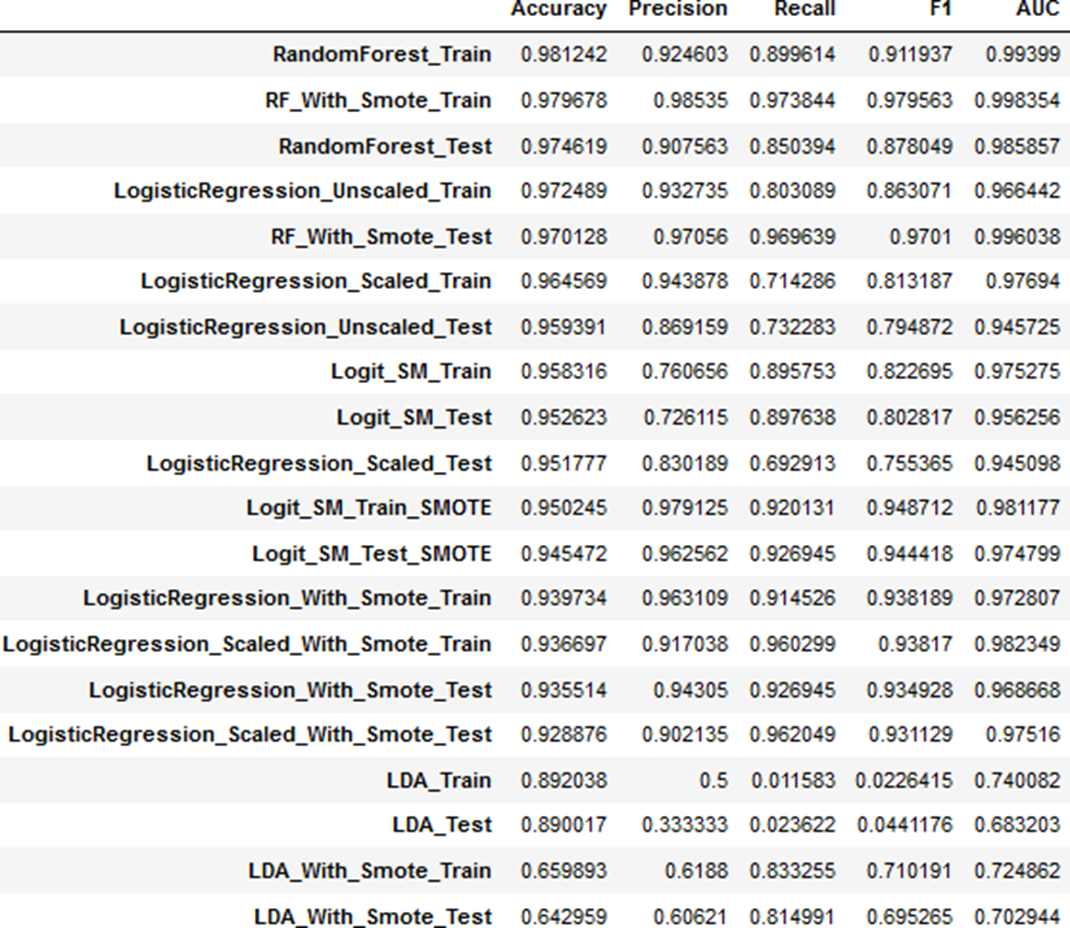
Compare the performances of Logistics; Radom Forest and LDA models (include ROC Curve)

**Resolution:**

We made following models as part of this exercise. For Random Forest and LDA scaling was not taken into account as these models are not impacted by varying magnitudes of the variables.

* Logistic regression using statsModel without SMOTE
* Logistic regression using statsModel with SMOTE
* Logistic Regression using sklearn With Unscaled Variables Without SMOTE
* Logistic Regression using sklearn With Scaled Variables Without SMOTE
* Logistic Regression using sklearn With Unscaled Variables With SMOTE
* Logistic Regression using sklearn With Scaled Variables With SMOTE
* LDA using sklearn without SMOTE
* LDA using sklearn with SMOTE
* Random Forest using sklearn without SMOTE
* Random Forest using sklearn with SMO

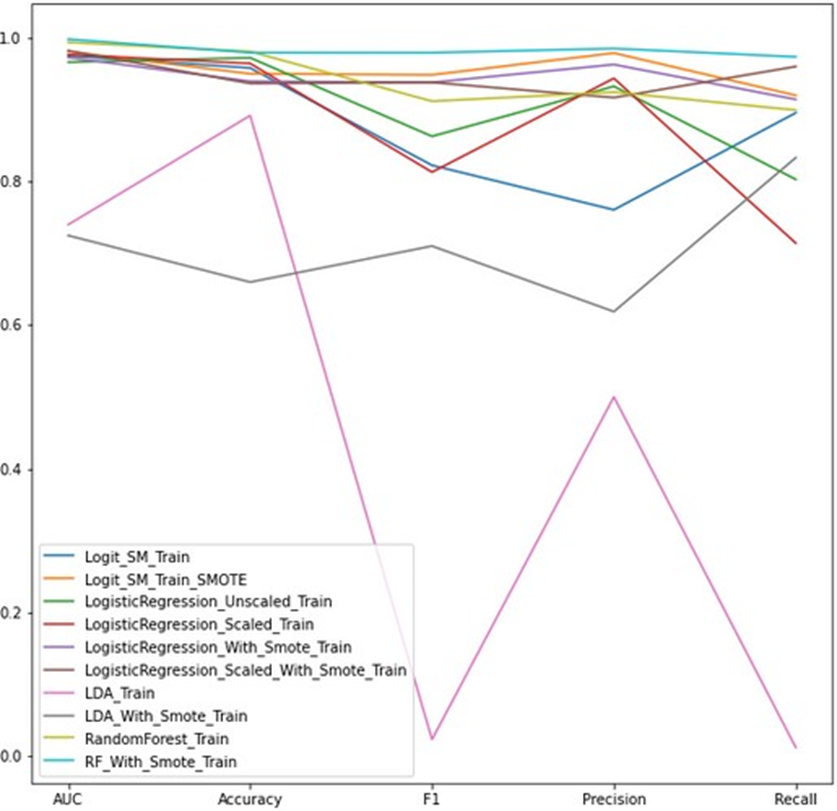
Below is the comparison chart for all the models.



Looking at the above comparison chart we can see Random Forest model has surpassed all other models in terms of accuracy. Both Random Forest with and without smote feature in the top 2 models.

While Random Forest was the best performing model, LDA models were the worst performers of all. LDA with or without were both really bad in terms of accuracies, LDA with smote being the worst of the lot.

Below is a comparison of different evaluation metrics for all the models on the training data set.



Looking at the above chart we clearly see the blue line indicating "Random Forest With SMIOTE" being on top of all other models, consistently for all the evaluation parameters.

While the pink line for LDA without smote is slightly higher than LDA with smote in terms of accuracy and auc score, it slips down significantly on all other parameters, making it the worst performing model of the lot.

Below is a comparison of different evaluation metrics for all the models on the test data set.

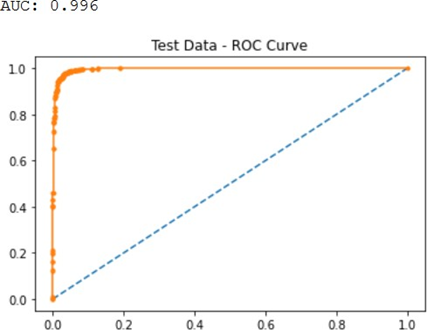
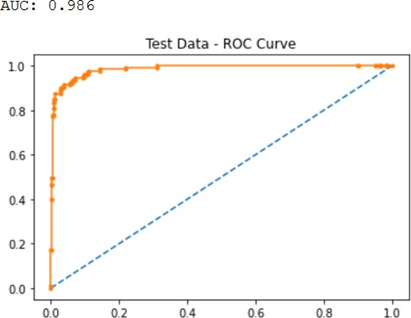


On the test data set, which is the deciding factor, here too we see Random Forest with SMOTE indicated by blue line performing the best on almost all the fronts. We can see the blue line dip only very slightly on accuracy compared to Random Forest without smote, on all other fronts the blue line stays on the top and is the clear winner amongst all the models.

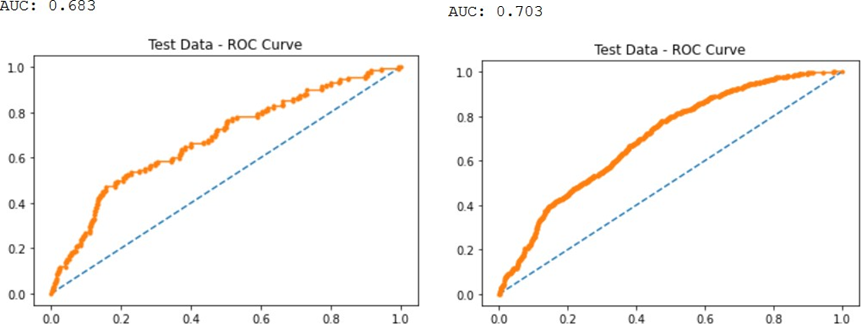
**Random Forest model with smote dataset is the best model amongst all the models with 97.01% accuracy, 97.05% precision, 96.96% recall and a recall and F1 score of 0.9701 and 0.996038 respectively.**

Also the ROC curves for Logit, LDA and random forest are plotted below. ROC curves for the best performing models on test data set have been plotted for comparison.

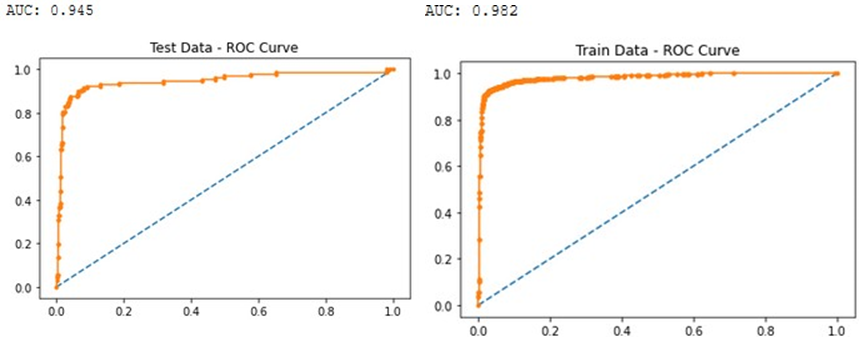
Comparing AUC scores too we see Random forest with smote data clearly wins compared to the others.



RF without SMOTE RF with SMOTE



LDA Without SMOTE LDA with SMOTE



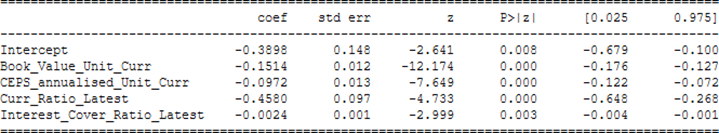
Logit without SMOTE (Scaled) Logit with SMOTE (Scaled)

### Problem 1.13

State Recommendations from the above models

**Resolution:**

Using the information gained from above exercise, we can say Random Forest with smote data is the best model. We also looked at the coefficients derived from the best Logit model built using Stats model to derive some more insights.



From the above analysis, we can infer below business insights. Following things should be kept in mind while investing in these companies.

* Lower the Book\_value\_unit\_curr i.e. Net assets, higher is the chance of a default, which would mean the net worth next year for this company is expected to be negative.
* Lower the CEPS\_annualised\_Unit\_Curr i.e. Cash earning per share, higher is the change of a default.
* Higher the Curr\_Ratio\_Latest, i.e. the companies ability to pay short term dues, lower are its chances of defaulting or having a negative net worth in next year.
* Higher the Interest\_Cover\_Ratio\_Latest lower the chances of default. Which means easier the company is able to pay the interest on its outstanding debt, lower are its chances to default.
* Curr\_Ratio\_Latest is most important criteria amongst the above parameters, while Interest\_Cover\_Ratio\_Latest is the least important when considering only these 4 parameters. However all these 4 parameters remain important compared to the other variables in the data set.

### Problem 2.1

Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference

**Resolution:**

The dataset contains 6 years of information (weekly stock information) on the stock prices of 10 different Indian Stocks. Calculate the mean and standard deviation on the stock returns and share insights. Please find attached the files to be referred.

Stock price graphs of Infosys and Mahindra & Mahindra are plotted below vs Time.



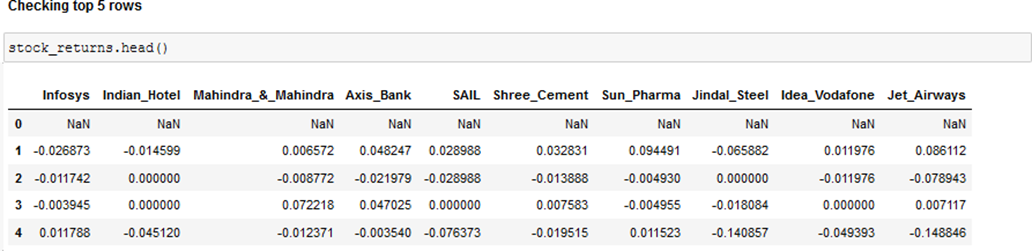
### Problem 2.2

Calculate Returns for all stocks with inference

**Resolution:**

Returns for all the stocks i.e. difference of log of price at t and the log of price at t-1 are shown below.





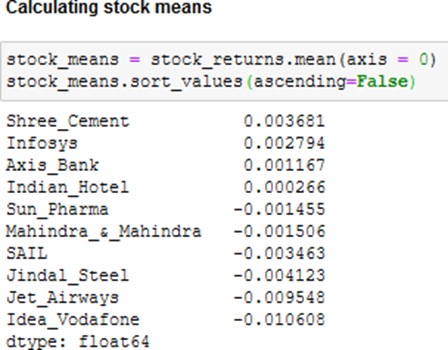
Week over week returns for all the stocks has been given in the above figure. We have given only the top 5 rows due to space constraints here. For complete data refer Jupyter notebook.

### Problem 2.3

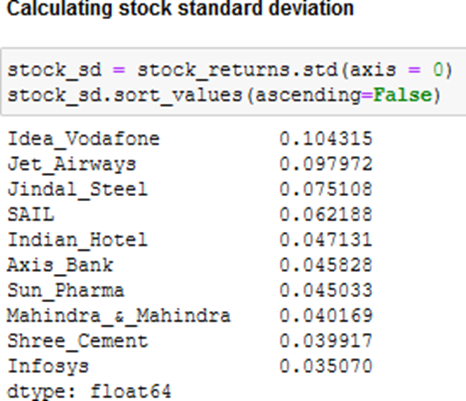
Calculate Stock Means and Standard Deviation for all stocks with inference.

**Resolution:**

Stock means and standard deviations have been calculated below. The values have been sorted in descending order.



Idea\_Vodafone has the lowest returns, while shree cements have the highest returns.



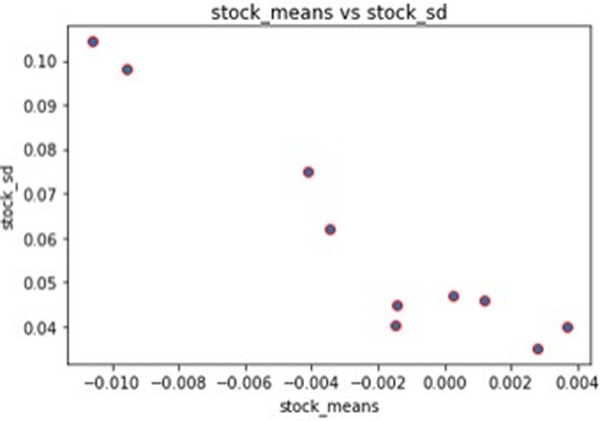
Idea\_Vodafone has the highest risk factor while Infosys is the least risky investment option.

### Problem 2.4

Draw a plot of Stock Means vs Standard Deviation and state your inference

**Resolution:**

Below is the plot of stock means vs standard deviation.



Stoks higher up but on the far left indicate high volatility and low returns, while the stocks on the bottom right indicate low volatility and high returns.

This is a useful graph to find a balance between risk and reward when it comes to investing in different companies.

### Problem 2.5

Conclusion and Recommendations

**Resolution:**

We can conclude by saying the below.

Stock with a lower mean & higher standard deviation do not play a role in a portfolio that has competing stock with more returns & less risk.

Thus for the data we have here, we are only left few stocks:

* One with highest return and lowest risk &
* One with lowest risk and highest return

Therefore from pure Returns perspective, Shree\_Cement followed by Infosys & Axis\_Bank looks good in this dataset.

From pure Risk perspective (as measured by standard deviation), Infosys followed by Shree\_Cement & Mahindra\_&\_Mahindra looks good in this dataset.

We would recommend using the stock means vs standard deviation plot to asses the risk to reward ratio. More volatile stock might give short term gains but might not be a good investment in long term. Whereas a low volatile stock might not be a good investment in short term, but might give a good return in long term.

Hence based on the type of investment that one is looking for, a inference should be made from the above mentioned plot.

The End

Thakur Arun Singh

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*