fra MILESTONE 2

S. VISHAL

july pgpdsba 2021

BUSINESS REPORT

[Company name] | [Company address]

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SNO** | **DESCRIPTION** | **PAGE NO** |
| 1 | Build a Random Forest Model on Train Dataset.  Also showcase your model building approach | 2 |
| 2 | Validate the Random Forest Model on test Dataset and  state the performance matrices.  Also state interpretation from the model | 2 |
| 3 | Build a LDA Model on Train Dataset.  Also showcase your model building approach | 3 |
| 4 | Validate the LDA Model on test Dataset and  state the performance matrices.  Also state interpretation from the model | 5 |
| 5 | Compare the performances of  Logistics, Radom Forest and LDA models (include ROC Curve) | 6 |
| 6 | State Recommendations from the above models | 7 |
| 7 | Draw Stock Price Graph (Stock Price vs Time)  for any 2 given stocks with inference | 8 |
| 8 | Calculate Returns for all stocks with inference | 9 |
| 9 | Calculate Stock Means and Standard Deviation  for all stocks with inference | 9 |
| 10 | Draw a plot of Stock Means vs Standard Deviation and  state your inference | 10 |
| 11 | Conclusion and Recommendations | 11 |

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

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

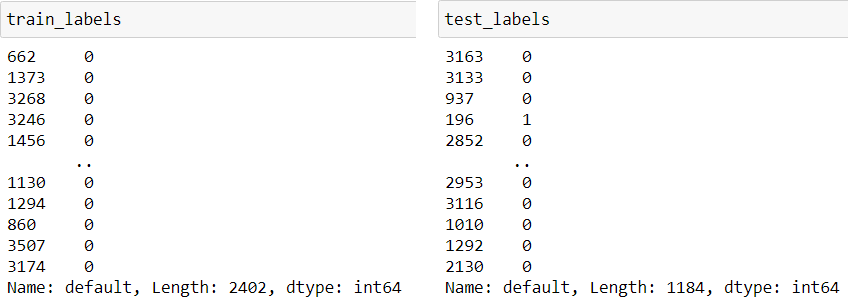
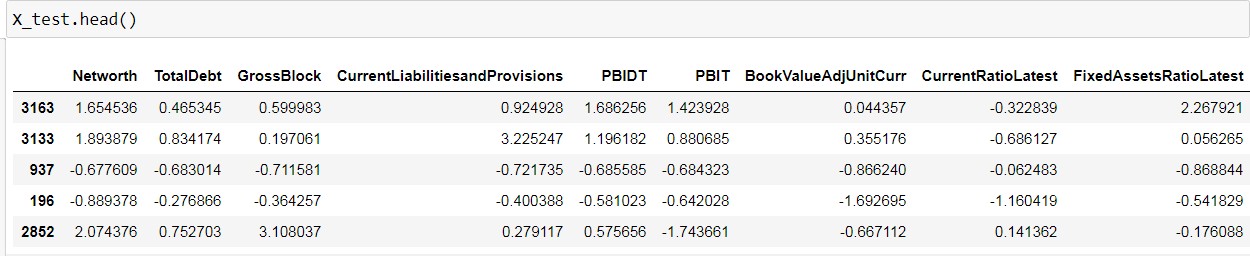
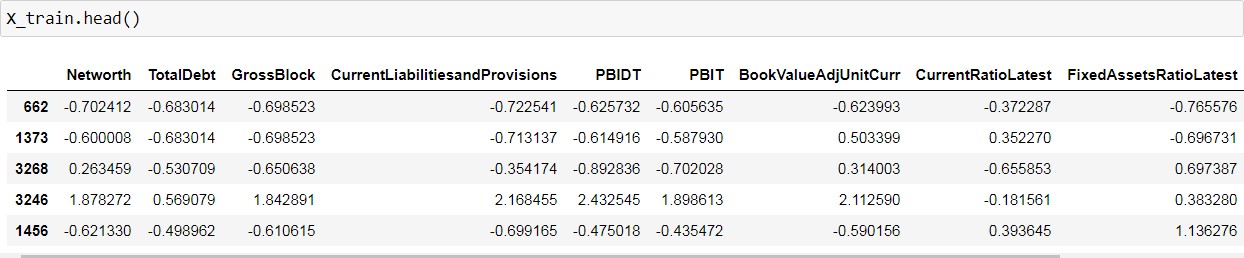
We got the following best parameters through GridSearchCV for the dataset,

RandomForestClassifier (max\_depth=**8**, max\_features=**4**, min\_samples\_leaf=**30**, min\_samples\_split=**90**, n\_estimators=**200**)



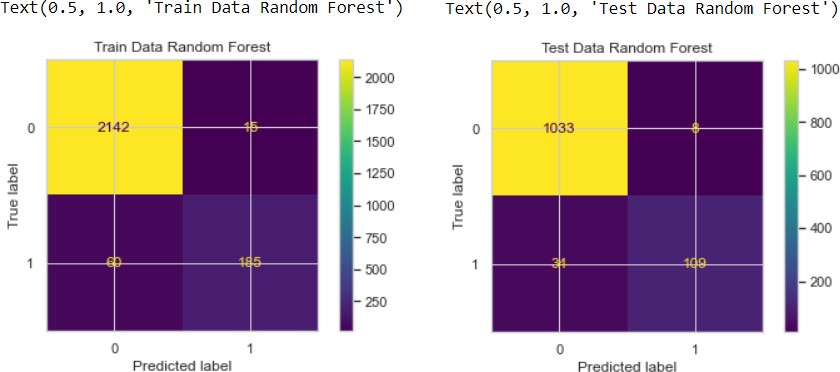
The highest importance feature denoted by this method is **Net worth (49.21% importance).**

# Train\_Test Split:

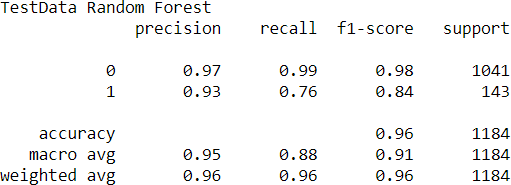
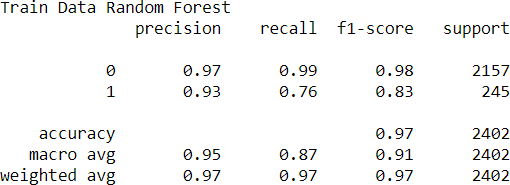


* 1. **Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model**

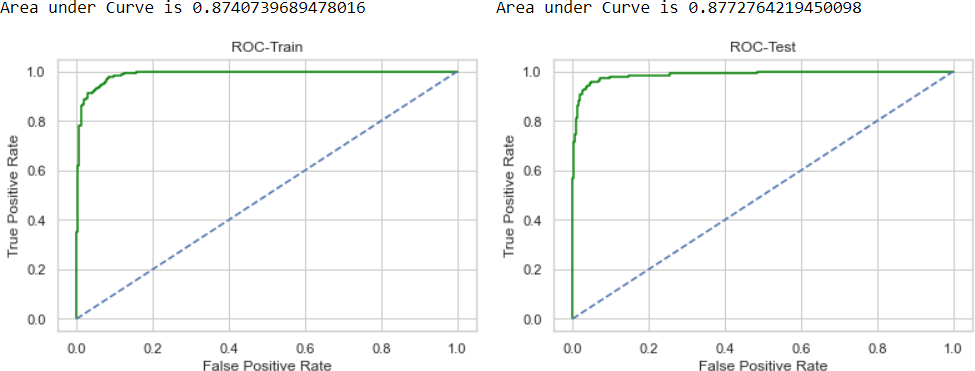
Confusion Matrix:



Classification Report:



ROC Curve:



Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

**Net worth** is again the most important variable for predicting default status.

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

Linear discriminate analysis and logistic regression are the most widely used statistical methods for analyzing categorical outcome variable. While both are appropriate for the development of linear classification models, linear discriminate analysis makes more assumptions about the underlying data and LDA is preferred when it is nominal (more than two groups).



The Coefficients of different variables as per LDA model are as below:

A screenshot of a computer

Description automatically generated with medium confidence

The highest importance feature denoted by this method is **BookValueAdjUnitCurr (-2.0044213).**

**Train\_Test Split:**

Table

Description automatically generated

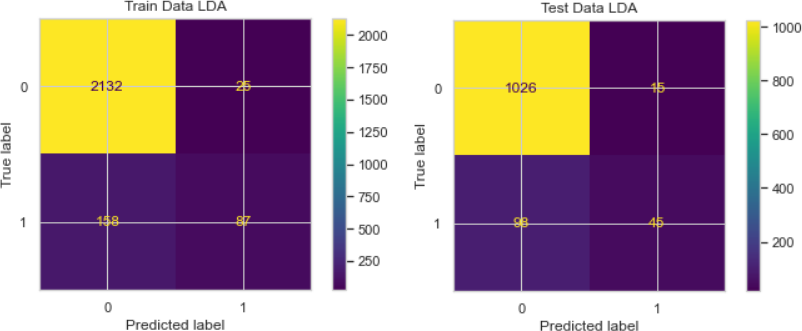
Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

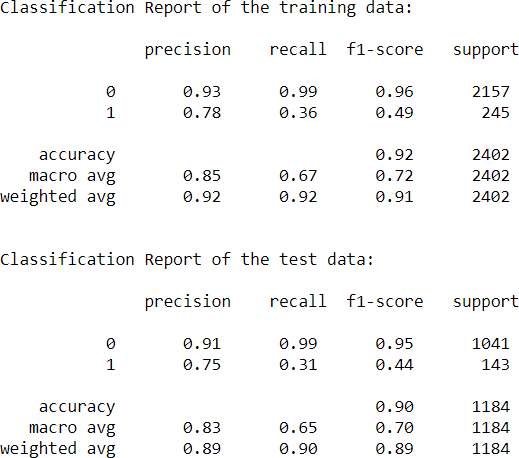
Description automatically generated

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

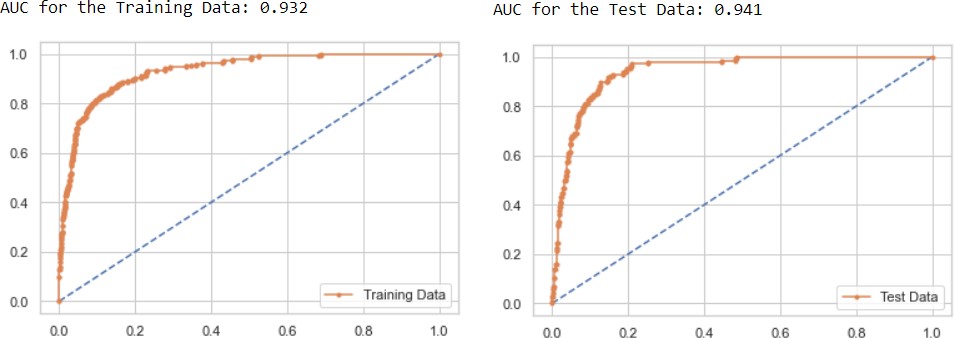
Confusion Matrix:



Classification Report:



ROC Curve:



Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

**BookValueAdjUnitCurr** is again the most important variable for predicting default status.

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

Comparing the performance metrics from the three models, we can summarize as below,

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Train** | **Logistic Test** | **LDA Train** | **LDA Test** | **Random Forest Train** | **Random Forest Test** |
| **Accuracy** | 0.94 | 0.93 | 0.92 | 0.9 | 0.97 | 0.96 |
| **AUC** | 0.934 | 0.946 | 0.932 | 0.941 | 0.87 | 0.87 |
| **Recall** | 0.49 | 0.58 | 0.36 | 0.31 | 0.76 | 0.76 |
| **Precision** | 0.83 | 0.83 | 0.78 | 0.75 | 0.93 | 0.93 |
| **F1 Score** | 0.62 | 0.68 | 0.49 | 0.44 | 0.83 | 0.84 |

Looking at the details got from **test data** from the three models,

Accuracy: Random Forest models has highest value of 0.96

AUC: Logistic Regression model has highest value of 0.946 and Random Forest model has least value of 0.87

Recall: Random Forest model has highest value of 0.76 and LDA model has least value 0.31

Precision: Random Forest has highest value of 0.93 and LDA model has least value of 0.75

F1 Score: Random Forest has highest value 0.84 and LDA model has least value of 0.44

Training and Test set results are almost similar in all the three models and overall measures are high in Random Forest.

# Therefore, Random Forest has slightly better performance than the Logistic Regression and LDA model

Overall, all the 3 models are reasonably stable enough to be used for making any future predictions. From Logistic and LDA Model, the variable **BookValueAdjUnitCurr** is found to be the most useful feature amongst all other features for predicting default status.

Chart, line chart

Description automatically generated

# State Recommendations from the above models.

Due to the importance of understanding and managing the risks in volatile business

domains, it is required to find an effective aid in making decisions. The result from model shows that the above algorithm is a promising opportunity in predicting whether a company will go for the default or not through the cause-and-effect relationship between the independent and dependent variables of the given dataset.

Also, Random Forest has proven to be a great algorithm if the dataset is in tabular format. Random Forests requires less pre-processing, and the training process is also much simpler. Moreover, hyper-parameter tuning is easier with random forest when compared to other models. This gives random forest the edge above remaining models

The above model will be helpful in predicting the dependent variables through the independent variables by assigning the probability of company going for the default to every predictor variable to give the best predictive/dependent variable.

As per predictions of the model, the variable **BookValueAdjUnitCurr** is found to be the most useful feature amongst all other features for predicting default status.

We must look about company based on the feature importance to get the better results in predicting whether a company will go for the default or not.

So, The Overall analysis of given dataset helped to get insights that would help in predicting about the company.

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

Stock Price Graph for Infosys:

Chart, scatter chart

Description automatically generated

The Stock price for the Infosys is on increasing trend from 2014 to 2021. There is an almost increase of 500 points within the span of 7 years.

Chart, scatter chart

Description automatically generatedStock Price Graph for Sun\_Pharma:

The Stock price for the Sun\_Pharma is on decreasing trend from 2014 to 2021. There is an almost decrease of 700 points within the span of 7 years.

# Calculate Returns for all stocks with inference.

**Returns** is the **difference** between two consecutive week prices for the stock.

Graphical user interface, application

Description automatically generated

The **negative** value of Return means there is **decrease** in price compared to previous week and the **positive**

value of Return means there is **increase** in price compared to previous week.

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

* + - **Stock Means:** Average returns that the stock is making on a week-to-week basis.

Text

Description automatically generated

**Shree\_Cement** has **highest** Stock Means and **Jet\_Airways** has **lowest** Stock Means.

* + - **Stock Standard Deviation:** It is a measure of volatility meaning the more a stock's returns vary from the stock's average return, the more volatile the stock.

Text

Description automatically generated

**Idea\_Vodafone** has **highest** Volatility and **Infosys** has **lowest** Volatility.

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

Plot between Stock Means & Stock standard Deviation:

Chart, scatter chart

Description automatically generated

From above plot, we can understand that stock with higher average value has lower volatility. There is a decrease in the average value with the increase in the volatility.

Table

Description automatically generated

# Conclusion and Recommendations.

Of all the above stocks, only the following stocks are having positive average means.

Infosys – 0.002794

Indian\_Hotel – 0.000266

Axis\_Bank – 0.001167

Shree\_Cement – 0.003681

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:

|  |  |  |
| --- | --- | --- |
|  | **Average** | **Volatility** |
| **Infosys** | 0.0028 | 0.0351 |
| **Shree\_Cement** | 0.0037 | 0.0399 |
| **Axis\_Bank** | 0.0012 | 0.0458 |
| **Indian\_Hotel** | 0.0003 | 0.0471 |

Among the above stocks, **Infosys & Shree\_Cement** stocks are having **best average** with **low volatility.**

Therefore, the stocks with higher return for a comparative or lower risk are considered better among all the available stocks.