BUSINESS REPORT ON

FRA Project (Milestone-2)

TABLE OF CONTENTS

INTRODUCTION:

PROBLEM 1

The purpose of this is to explore the dataset. Do the exploratory data analysis. Treating outliers and missing values. Scaling of the dataset. Split the data into Train and Test dataset in a ratio of 67:33 and use random_state =42. Model Building is to be done on Train Dataset and Model Validation is to be done on Test Dataset using different models viz., logistic regression, random forest and LDA and finally drawing a best model for identifying defaults.

PROBLEM 2

MARKET RISK ANALYSIS

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. To perform market risk analysis using python.

PROBLEM 1

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

BUSINESS INSIGHT:

Investors want to invest in such companies where there are minimum defaults and business or companies can fall prey to defaults if the debt obligations are not met. Thus defaults can lead to lower credit rating and have to pay higher interest rates on existing debts. 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.

The data of financial statement of the companies for the previous year (2015)and also information about the Networth of the company in the following year (2016) is provided . This is provided as Company_Data2015-1.xlsx. Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

The companies with minimum defaults are beneficial from investors point and are the best. To build a Random Forest Model on Train Dataset and showcase model building approach.

APPROACH USED:

- Import necessary libraries .'numpy, python, matplotlib, seaborn, sklearn metrics.
- Read the dataset (Company_Data2015.xlsx)
- Have a glimpse on the dataset viz., head, shape
- Fixing messy column names (like %, spaces etc.) for ease of use and fixing visualise the data head for the changes made.
- Data information on datatypes, and describing structure of the dataset.
- Eliminating or dropping the redundant variables.
- Checking for duplicates in the data set and removing them.
- Check the number of missing values in the dataset
- Checking for outliers in the dataset.
- Outlier treatment.

• Missing value treatment.
• Creating a binary target variable'default' using 'Networth_Next_Year' and taking the value of 1 when net worth next year is negative & 0 when net worth next year is positive.
• Scaling the predictors using StandardScaler from sklearn.
• Splitting of the data into Train and Test dataset in a ratio of 67:33 and using random_state =42.
• Build random forest model as mentioned below:

RANDOM FOREST MODEL BUILDING

1. Import randomforestclassifier and gridsearchcv from sklearn.
2. Consider param_grid for the following parameters of max_depth, min_samples_leaf, min_samples_split and n_estimators.
3. Perform gridsearch for randomforestclassifer as estimator and param_grid.
4. Fit model for X_train and y_train i.e., train dataset.
5. Predict the confusion matrix on train dataset.
6. Print the classification report for train dataset.
7. Plot AUC and ROC curve for the train data.

Parameters Used In Building model on train dataset:
{ max_depth, min_samples_leaf, min_samples_split and n_estimators} and estimator is RandomForestClassifier().

INFERENCE:

The best parameters to build the random forest model are:
{'max_depth': 5, 'min_samples_leaf': 5, 'min_samples_split': 15, 'n_estimators': 50}
As observed from the classification metrics, the accuracy of the model i.e. %overall correct predictions is 98%.
Sensitivity of the model is 87% i.e. 87% of those defaulted (default=1) were correctly identified as defaulters by the model.
The AUC for the train data is 0.931.
The model is built for the train dataset and the classification report is mentioned in Table 1.8a

OUTPUT:

Confusion Matrix for the train data using random forest:

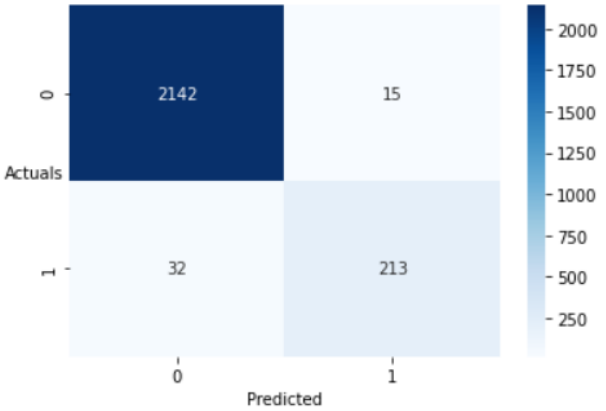
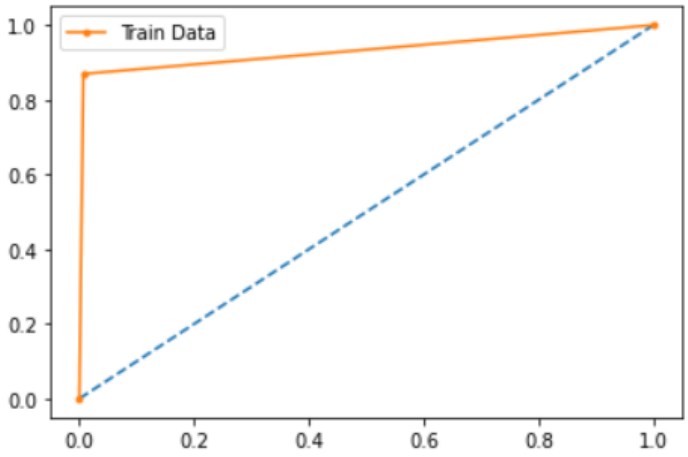


Table 1.8a: Classification Report for the train data using random forest:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	2157
1	0.93	0.87	0.90	245
accuracy			0.98	2402
macro avg	0.96	0.93	0.94	2402
weighted avg	0.98	0.98	0.98	2402

AUC-ROC Curve for the train data using random forest:

AUC for the Train Data: 0.931



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

BUSINESS INSIGHT:

Investors want to invest in such companies where there are minimum defaults and business or companies can fall prey to defaults if the debt obligations are not met. Thes defaults can lead to lower credit rating and have to pay higher interest rates on existing debts. 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.

The data of financial statement of the companies for the previous year (2015)and also information about the Networth of the company in the following year (2016) is provided . This is provided as Company_Data2015-1.xlsx. Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

The companies with minimum defaults are beneficial from investors point and are the best. To validate the Random Forest Model on test Dataset and stating the performance matrices and also interpreting from the model.

APPROACH USED:

• Import necessary libraries .'numpy, python, matplotlib, seaborn, sklearn metrics.
• Read the dataset (Company_Data2015.xlsx)
• Have a glimpse on the dataset viz., head, shape
• Fixing messy column names (like %, spaces etc.) for ease of use and fixing visualise the data head for the changes made.
• Data information on datatypes, and describing structure of the dataset.
• Eliminating or dropping the redundant variables.
• Checking for duplicates in the data set and removing them.
• Check the number of missing values in the dataset
• Checking for outliers in the dataset.
- Outlier treatment.

• Missing value treatment.
• Creating a binary target variable'default' using 'Networth_Next_Year' and taking the value of 1 when net worth next year is negative & 0 when net worth next year is positive.
- Scaling the predictors using StandardScaler from sklearn.

• Splitting of the data into Train and Test dataset in a ratio of 67:33 and using random_state =42.
- Build random forest model on train dataset.
- Validate the random forest model on the test dataset using the model built for train dataset.
- The optimal threshold value is calculated and the test data set is validated above this threshold.
- Evaluating the performance matrices along with interpretation.

Performance Metrics:

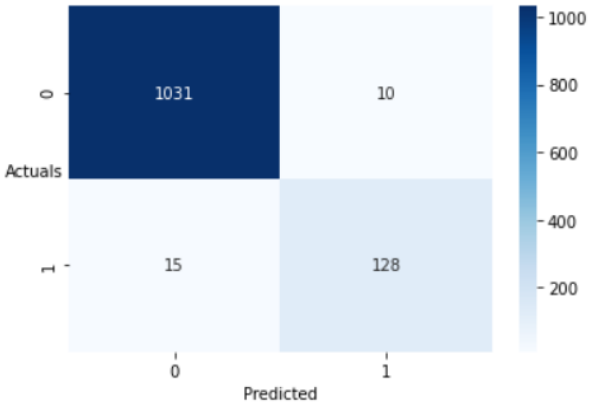
- Confusion Matrix
- Classification metrics
- AUC-ROC

INFERENCE:

- Model is validated by considering the optimal threshold value of 0.0512 for the test dataset.
- The accuracy of the model on test data i.e. %overall correct predictions is 98% Sensitivity of the model is 90% i.e. 90% of those defaulted (default=1) were correctly identified as defaulters by the model.
- The AUC for the test data is 0.946.

OUTPUT:

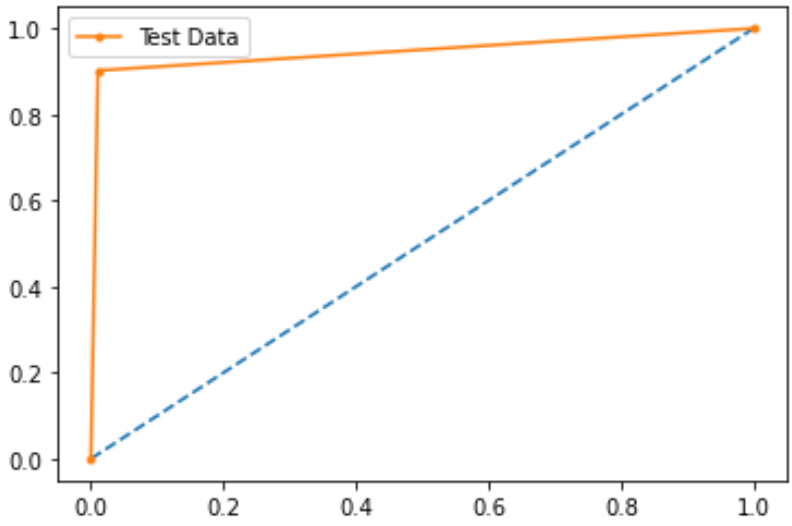
Confusion Matrix for the test data using random forest:



Classification Report for the test data using random forest:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1041
1	0.93	0.90	0.91	143
accuracy			0.98	1184
macro avg	0.96	0.94	0.95	1184
weighted avg	0.98	0.98	0.98	1184

AUC-ROC Curve for the test data using random forest:



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

BUSINESS INSIGHT:

Investors want to invest in such companies where there are minimum defaults and business or companies can fall prey to defaults if the debt obligations are not met. Thes defaults can lead to lower credit rating and have to pay higher interest rates on existing debts. 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.

The data of financial statement of the companies for the previous year (2015)and also information about the Networth of the company in the following year (2016) is provided . This is provided as Company_Data2015-1.xlsx. Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

The companies with minimum defaults are beneficial from investors point and are the best.

To build a LDA Model on Train Dataset and showcase model building approach.

APPROACH USED:

• Import necessary libraries .'numpy, python, matplotlib, seaborn, sklearn metrics.

• Read the dataset (Company_Data2015-1.xlsx)

• Have a glimpse on the dataset viz., head, shape

• Fixing messy column names (like %, spaces etc.) for ease of use and fixing visualise the data head for the changes made.

• Data information on datatypes, and describing structure of the dataset.

• Eliminating or dropping the redundant variables.

• Checking for duplicates in the data set and removing them.

• Check the number of missing values in the dataset

• Checking for outliers in the dataset.

- Outlier treatment.

• Missing value treatment.

• Creating a binary target variable'default' using 'Networth_Next_Year' and taking the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

- Scaling the predictors using StandardScaler from sklearn.

• Splitting of the data into Train and Test dataset in a ratio of 67:33 and using random_state =42.

- Build LDA model as mentioned below:
    1. Import LinearDiscriminantAnalysis from sklearn.
    2. Fit model for X_train and y_train i.e., train dataset.
    3. Calculating the optimal threshold value for building LDA model
    4. Predict the confusion matrix on train dataset.
    5. Print the classification report for train dataset.
    6. Plot AUC and ROC curve for the train data.

INFERENCES:

Model Building by considering the optimal threshold value of 0.0512

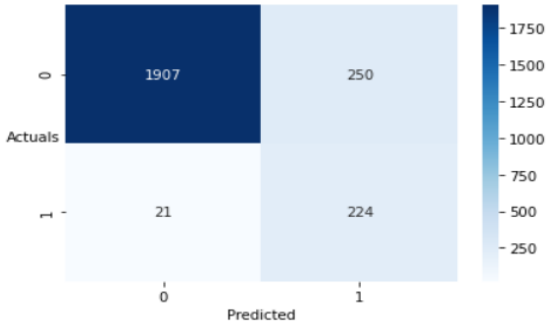
As observed from the classification metrics, the accuracy of the model i.e. %overall correct predictions is 91%.

Sensitivity of the model is 88% i.e. 88% of those defaulted (default=1) were correctly identified as defaulters by the model.

The AUC for the train data is 0.899.

OUTPUT:

Confusion Matrix for the train data using LDA:

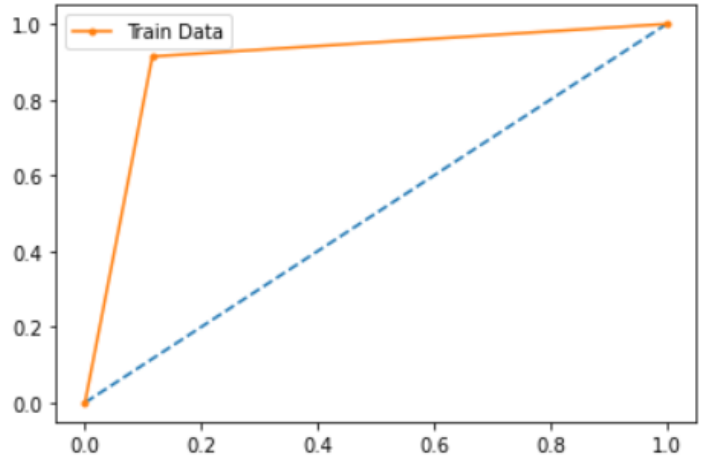


Classification Report for the train data using LDA:

	precision	recall	f1-score	support
0	0.989	0.884	0.934	2157
1	0.473	0.914	0.623	245
accuracy			0.887	2402
macro avg	0.731	0.899	0.778	2402
weighted avg	0.936	0.887	0.902	2402

AUC-ROC Curve for the train data using LDA:

```
AUC for the Train data is: 0.899
```



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

BUSINESS INSIGHT:

Investors want to invest in such companies where there are minimum defaults and business or companies can fall prey to defaults if the debt obligations are not met. Thes defaults can lead to lower credit rating and have to pay higher interest rates on existing debts. 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.

The data of financial statement of the companies for the previous year (2015)and also information about the Networth of the company in the following year (2016) is provided . This is provided as Company_Data2015-1.xlsx. Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

The companies with minimum defaults are beneficial from investors point and are the best.

- **Validate the LDA model on the test dataset using the model built for train dataset.**
- **The optimal threshold value is calculated and the test data set is validated above this threshold.**
- Evaluating the performance matrices along with interpretation.

  Performance Metrics:
    - Confusion Matrix
    - Classification metrics
    - AUC-ROC

APPROACH USED:

• Import necessary libraries .'numpy, python, matplotlib, seaborn, sklearn metrics.
• Read the dataset (Company_Data2015.xlsx)
• Have a glimpse on the dataset viz., head, shape
• Fixing messy column names (like %, spaces etc.) for ease of use and fixing visualise the data head for the changes made.
• Data information on datatypes, and describing structure of the dataset.
• Eliminating or dropping the redundant variables.
• Checking for duplicates in the data set and removing them.
• Check the number of missing values in the dataset
• Checking for outliers in the dataset.
- Outlier treatment.

• Missing value treatment.
• Creating a binary target variable'default' using 'Networth_Next_Year' and taking the value of 1 when net worth next year is negative & 0 when net worth next year is positive.
- Scaling the predictors using StandardScaler from sklearn.

• Splitting of the data into Train and Test dataset in a ratio of 67:33 and using random_state =42.
- Build random forest model on train dataset.

- Validate the LDA model on the test dataset and evaluating the performance matrices along with interpretation.

Performance Metrics:

- Confusion Matrix
- Classification metrics
- AUC-ROC

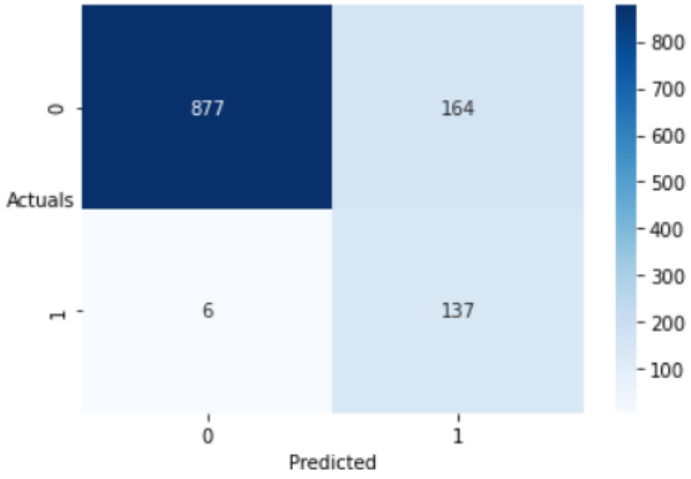
INFERENCE:

- It is inferred that the accuracy of the model i.e. %overall correct predictions is 85.6%.
- Sensitivity of the model is 95.8% i.e. 96% of those defaulted (default=1) were correctly identified as defaulters by the model.
- The AUC for the train data is 0.900.

OUTPUT:

True Negative: 877, False Positives: 164,
False Negatives: 6, True Positives: 137

Confusion Matrix for the test data using LDA:

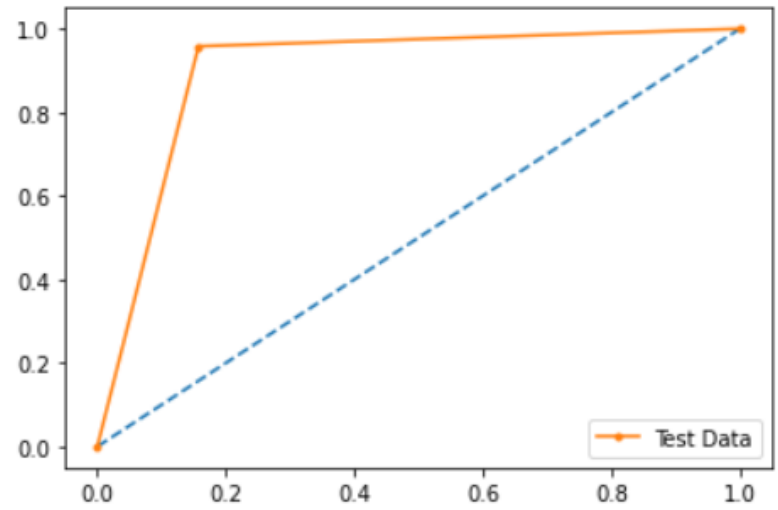


Classification Report for the test data using LDA:

	precision	recall	f1-score	support
0	0.993	0.842	0.912	1041
1	0.455	0.958	0.617	143
accuracy			0.856	1184
macro avg	0.724	0.900	0.764	1184
weighted avg	0.928	0.856	0.876	1184

AUC-ROC Curve for the test data using LDA:

```
AUC for the Test Data: 0.900
```



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

BUSINESS INSIGHT:

Investors want to invest in such companies where there are minimum defaults and business or companies can fall prey to defaults if the debt obligations are not met. Thes defaults can lead to lower credit rating and have to pay higher interest rates on existing debts. 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.

The data of financial statement of the companies for the previous year (2015)and also information about the Networth of the company in the following year (2016) is provided . This is provided as Company_Data2015-1.xlsx. Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

The companies with minimum defaults are beneficial from investors point and are the best. To compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve)

APPROACH USED:

• Import necessary libraries .'numpy, python, matplotlib, seaborn, sklearn metrics.
• Read the dataset (Company_Data2015.xlsx)
• Have a glimpse on the dataset viz., head, shape
• Fixing messy column names (like %, spaces etc.) for ease of use and fixing visualise the data head for the changes made.
• Data information on datatypes, and describing structure of the dataset.
• Eliminating or dropping the redundant variables.
• Checking for duplicates in the data set and removing them.
• Check the number of missing values in the dataset

• Checking for outliers in the dataset.
- Outlier treatment.
• Missing value treatment.
• Creating a binary target variable'default' using 'Networth_Next_Year' and taking the value of 1 when net worth next year is negative & 0 when net worth next year is positive.
- Scaling the predictors using StandardScaler from sklearn.
• Splitting of the data into Train and Test dataset in a ratio of 67:33 and using random_state =42.
- Build logistic regression, random forest, LDA model on train dataset.
- Validate the logistic regression, random forest, LDA model on the test dataset and evaluating the performance matrices along with interpretation.

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

- ROC curves of both train and test data are plotted and the best model is selected among the 3 models.

INFERENCE:

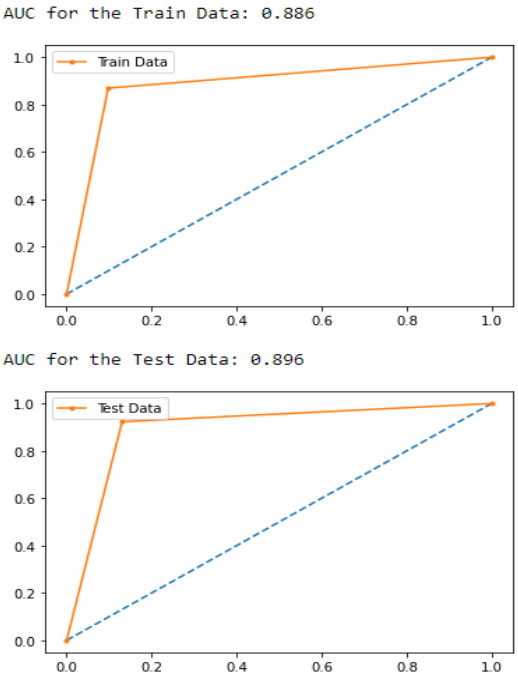
AUC- ROC curve is a performance measurement for the classification problems at various threshold settings. The higher the AUC value, the higher is the capability of model to differentiate the classes.

	AUC-Train Data	AUC-Test Data
Logistic Regression	0.886	0.896
Random Forest	0.931	0.946
LDA	0.899	0.900

Thus it is observed that **Random Forest model** has better performance in distinguishing between defaulters and non- defaulters.

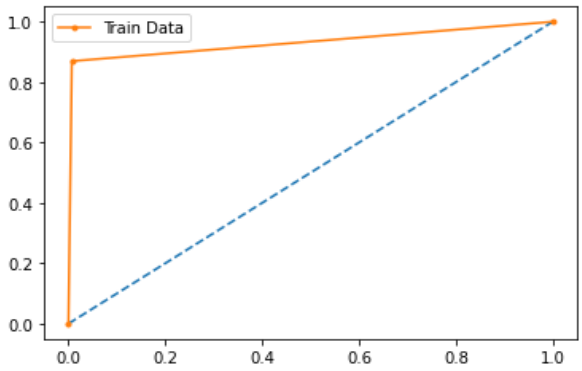
OUTPUT:

AUC- ROC curves for Logistic Regression

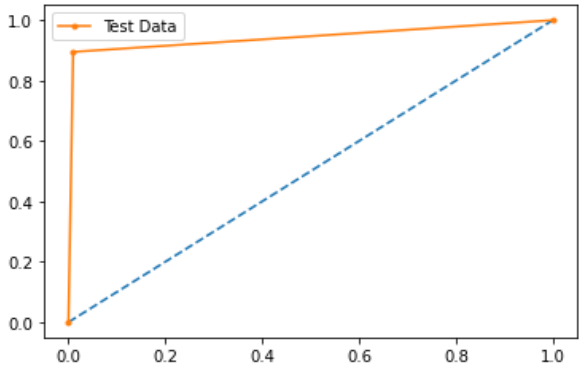


AUC- ROC curves for Random Forest

AUC for the Train Data: 0.931

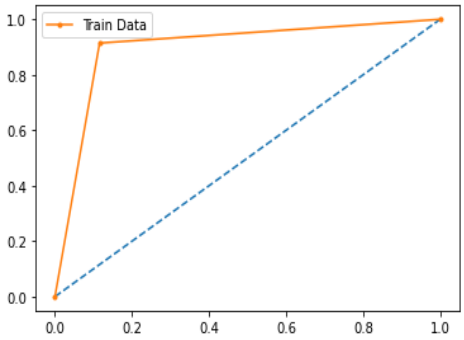


AUC for the Test Data: 0.943

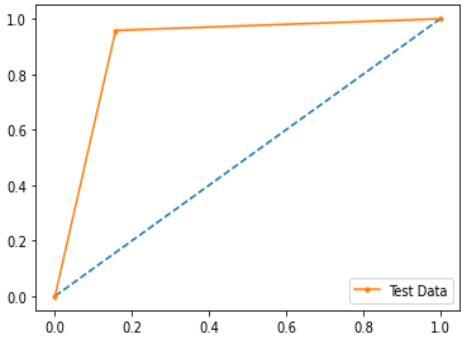


AUC- ROC curves for LDA

AUC for the Train data is: 0.899



AUC for the Test Data: 0.900



1.13 State Recommendations from the above models

BUSINESS INSIGHT:

Investors want to invest in such companies where there are minimum defaults and business or companies can fall prey to defaults if the debt obligations are not met. Thes defaults can lead to lower credit rating and have to pay higher interest rates on existing debts. 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.

The data of financial statement of the companies for the previous year (2015)and also information about the Networth of the company in the following year (2016) is provided . This is provided as Company_Data2015-1.xlsx. Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

The companies with minimum defaults are beneficial from investors point and are the best. To state recommendation from logistic regression, randomforest and LDA models.

APPROACH USED:

• Import necessary libraries .'numpy, python, matplotlib, seaborn, sklearn metrics.
• Read the dataset (Company_Data2015.xlsx)
• Have a glimpse on the dataset viz., head, shape
• Fixing messy column names (like %, spaces etc.) for ease of use and fixing visualise the data head for the changes made.
• Data information on datatypes, and describing structure of the dataset.
• Eliminating or dropping the redundant variables.
• Checking for duplicates in the data set and removing them.
• Check the number of missing values in the dataset
• Checking for outliers in the dataset.
- Outlier treatment.

• Missing value treatment.
• Creating a binary target variable'default' using 'Networth_Next_Year' and taking the value of 1 when net worth next year is negative & 0 when net worth next year is positive.
- Scaling the predictors using StandardScaler from sklearn.

• Splitting of the data into Train and Test dataset in a ratio of 67:33 and using random_state =42.
- Build logistic regression, random forest, LDA model on train dataset.
- Validate the logistic regression, random forest, LDA model on the test dataset and evaluating the performance matrices along with interpretation.

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

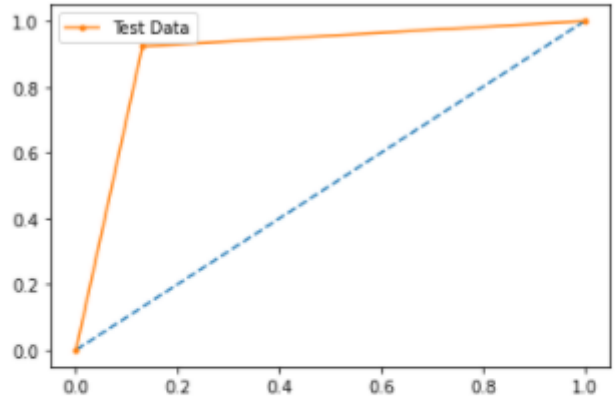
INFERENCE:

Observing test data AUC values
AUC for the Test Data: 0.896 - Logistic Regression
AUC for the Test Data: 0.946 - Random Forest
AUC for the Test Data: 0.900 – LDA

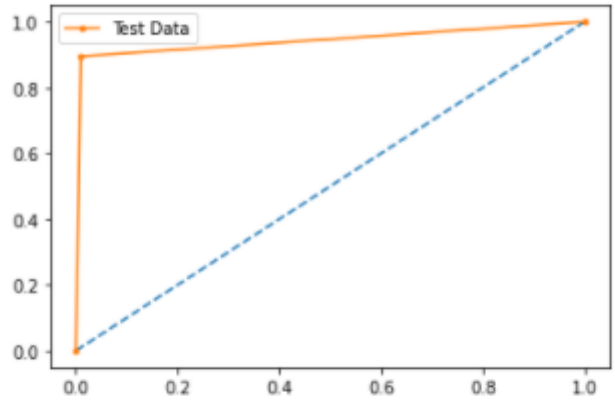
Thus it is observed that **Random Forest model** has better performance in distinguishing between defaulters and non- defaulters.

OUTPUT:

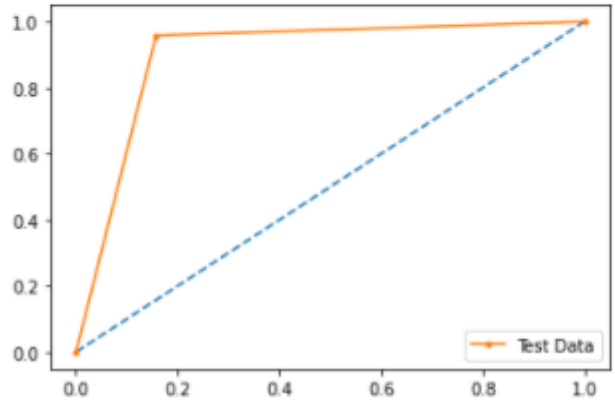
```
AUC for the Test Data using Logistics: 0.896
```




```
AUC for the Test Data using RandomForest: 0.943
```




```
AUC for the Test Data using LDA: 0.900
```



RECOMMENDATIONS (LOGISTIC REGRESSION):

For the test dataset:

- The accuracy of the model i.e. %overall correct predictions is 87%.
- Sensitivity of the model is 92.3% i.e. 92% of those defaulted were correctly identified as defaulters by the model.

RECOMMENDATIONS (RANDOM FOREST):

For the test dataset:

- The accuracy of the model on test data i.e. %overall correct predictions is 98%.
- Sensitivity of the model is 90% i.e. 90% of those defaulted (default=1) were correctly identified as defaulters by the model.

RECOMMENDATIONS (LDA):

For the test dataset:

- It is inferred that the accuracy of the model i.e. %overall correct predictions is 85.6%.
- Sensitivity of the model is 95.8% i.e. 96% of those defaulted (default=1) were correctly identified as defaulters by the model.

OVERALL RECOMMENDATION:

- **Random Forest model** has better performance in distinguishing between defaulters and non- defaulters.
- The company must ensure that to attract investors, it needs to minimise the defaulter section in the company
- The company must devise schemes and plans to reduce the defaulters as it results in high risk and loss for the company.

PROBLEM-2

MARKET RISK ANALYSIS-PROBLEM STATEMENT

Market Risk

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.

Dataset provided is Market Risk Dataset for doing the Market Risk Analysis using Python.

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

INSIGHTS:

Weekly stock information pertaining to stock prices of 10 different Indian Stocks are provided in the dataset of "Market Risk Dataset" as an excel file.

To draw stock price graph for stock price vs time for 2 stocks and give inferences accordingly.

APPROACH:

1. Importing the libraries of numpy, pandas, matplotlib, seaborn sklearn metrics.
2. Read the dataset 'Market+Risk+Dataset.csv'.
3. Fixing messy column names for easy using.
4. Glimpse on the head of the dataset.
5. Shape of the dataset.
6. Information on datatypes.
7. Checking basic measures of descriptive statistics.
8. To plot & see price trend over time for i.e., Stock Price garph for any two companies viz., Infosys and Sun-Pharma.
9. A scatter plot is plotted with date on x-axis and stockprices on y-axis

INFERENCE:

Weekly datasets are provided . The shape of the dataset has 314 rows and 11 columns. There are no null entries in the dataset. The basic measures of descriptive statistics for the continuous variables shows the five point summary with mean and standard deviation.
The date-time object is dealt and Dates are taken on x-axis and stock prices on y-axis.

The graph showing Stock Price vs Time for Infosys is shown in Fig 2.1a which shows an upward trend over the years.
The graph showing Stock Price vs Time for Sun_Pharma is shown in Fig 2.1b which shows declining trend over the years.

OUTPUT:

Figure 2.1a: Graph showing Stock Price vs Time for Infosys

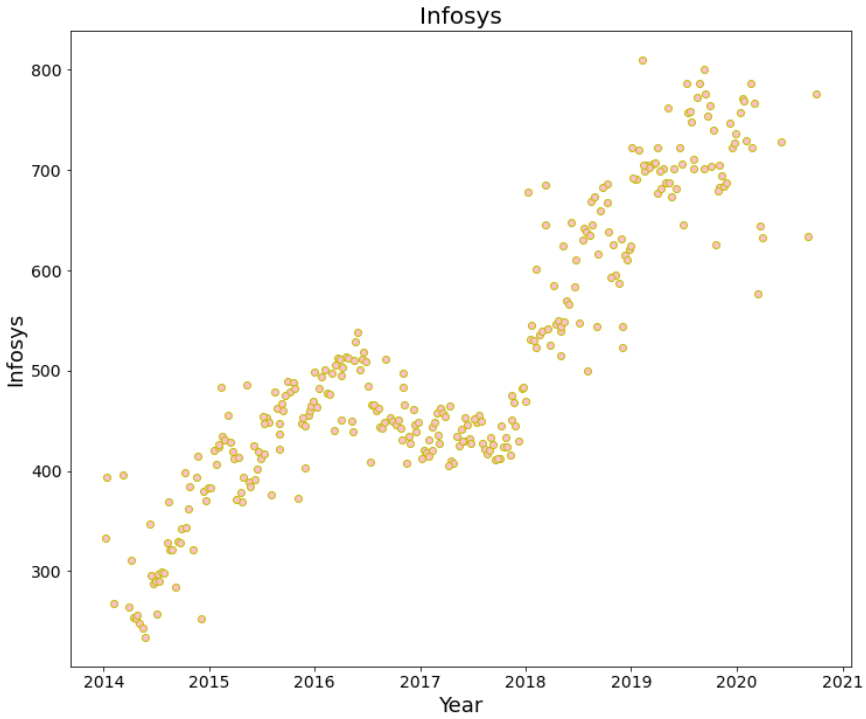
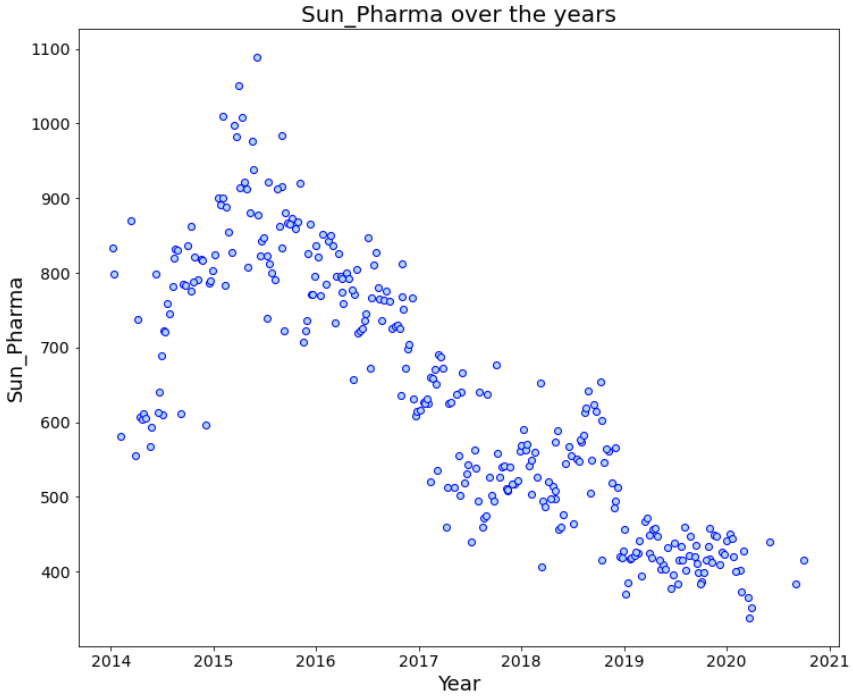


Figure 2.1b: Graph showing Stock Price vs Time for SUN-PHARMA



2.2 Calculate Returns for all stocks with inference

INSIGHTS:

6 years of weekly stock information pertaining to stock prices of 10 different Indian Stocks are provided in the dataset of "Market Risk Dataset" as an excel file.

To calculate returns for all stocks with inference.

APPROACH:

1. Importing the libraries of numpy, pandas, matplotlib, seaborn sklearn metrics.
2. Read the dataset 'Market+Risk+Dataset.csv'.
3. Fixing messy column names for easy using.
4. Glimpse on the head of the dataset.
5. Shape of the dataset.
6. Information on datatypes.
7. Checking basic measures of descriptive statistics.
8. Steps for calculating returns from prices:
   - Take logarithms
   - Take differences

The logarithmic returns are a difference between two consecutive week prices.

9. Check the shape and head of the dataset for all stocks after calculating returns.

INFERENCES:

The shape of the dataset is 314 rows and 10 columns.

The head of the dataset is shown in Table 2.2a. The first row is showing NaN as the earlier week value is not there to obtain difference. The remaining rows have all the logarithmic return values for the 10 different stocks.

OUTPUT:

	Infosys	Indian_ Hotel	Mahindra_ &_ Mahindra	Axis_ Bank	SAIL	Shree_ Cement	Sun_ Pharma	Jindal_ Steel	Idea_ Vodafone	Jet_ Airways
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	-0.0269	-0.0146	0.0066	0.0482	0.0290	0.0328	0.0945	-0.0659	0.0120	0.0861
2	-0.0117	0.0000	-0.0088	-0.0220	-0.0290	-0.0139	-0.0049	0.0000	-0.0120	-0.0789
3	-0.0039	0.0000	0.0722	0.0470	0.0000	0.0076	-0.0050	-0.0181	0.0000	0.0071
4	0.0118	-0.0451	-0.0124	-0.0035	-0.0764	-0.0195	0.0115	-0.1409	-0.0494	-0.1488

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

INSIGHTS:

6 years of weekly stock information pertaining to stock prices of 10 different Indian Stocks are provided in the dataset of "Market Risk Dataset" as an excel file.

To calculate Stock Means and Standard Deviation for all stocks with inference.

APPROACH:

1. Importing the libraries of numpy, pandas, matplotlib, seaborn sklearn metrics.
2. Read the dataset 'Market+Risk+Dataset.csv'.
3. Fixing messy column names for easy using.

4. Glimpse on the head of the dataset.
5. Shape of the dataset.
6. Information on datatypes.
7. Checking basic measures of descriptive statistics.
8. To calculate Stock Means which is an Average returns that the stock is making on a week to week basis
9. To calculate Stock Standard Deviation which is a measure of volatility meaning the more a stock's returns vary from the stock's average return, the more volatile the stock.
10. Creating a dataframe for the average and volatility i.e, mean and standard deviation.

INFERENCE:

Average returns of different stock and volatility i.e., stock standard deviation is shown in Tables 2.3 a.

Idea_Vodafone is having more volatility whereas average returns were higher for Shree_Cement.

OUTPUT:

Table 2.3a: Average and Volatility for different stocks

	Average	Volatility
Infosys	**0.002794**	**0.035070**
Indian_Hotel	**0.000266**	**0.047131**
Mahindra_&_Mahindra	**-0.001506**	**0.040169**
Axis_Bank	**0.001167**	**0.045828**
SAIL	**-0.003463**	**0.062188**
Shree_Cement	**0.003681**	**0.039917**
Sun_Pharma	**-0.001455**	**0.045033**
Jindal_Steel	**-0.004123**	**0.075108**
Idea_Vodafone	**-0.010608**	**0.104315**
Jet_Airways	**-0.009548**	**0.097972**

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

INSIGHTS:

6 years of weekly stock information pertaining to stock prices of 10 different Indian Stocks are provided in the dataset of "Market Risk Dataset" as an excel file.

To calculate Stock Means and Standard Deviation for all stocks with inference.

APPROACH:

1. Importing the libraries of numpy, pandas, matplotlib, seaborn sklearn metrics.
2. Read the dataset 'Market+Risk+Dataset.csv'.
3. Fixing messy column names for easy using.
4. Glimpse on the head of the dataset.
5. Shape of the dataset.
6. Information on datatypes.
7. Checking basic measures of descriptive statistics.

8. Create a scatter plot by adding specific lines to the scatter plot corresponding to the mean (y-axis) and standard deviation (x-axis)

INFERENCES:

The plot of stock means vs standard deviation are shown in Table 2.4a

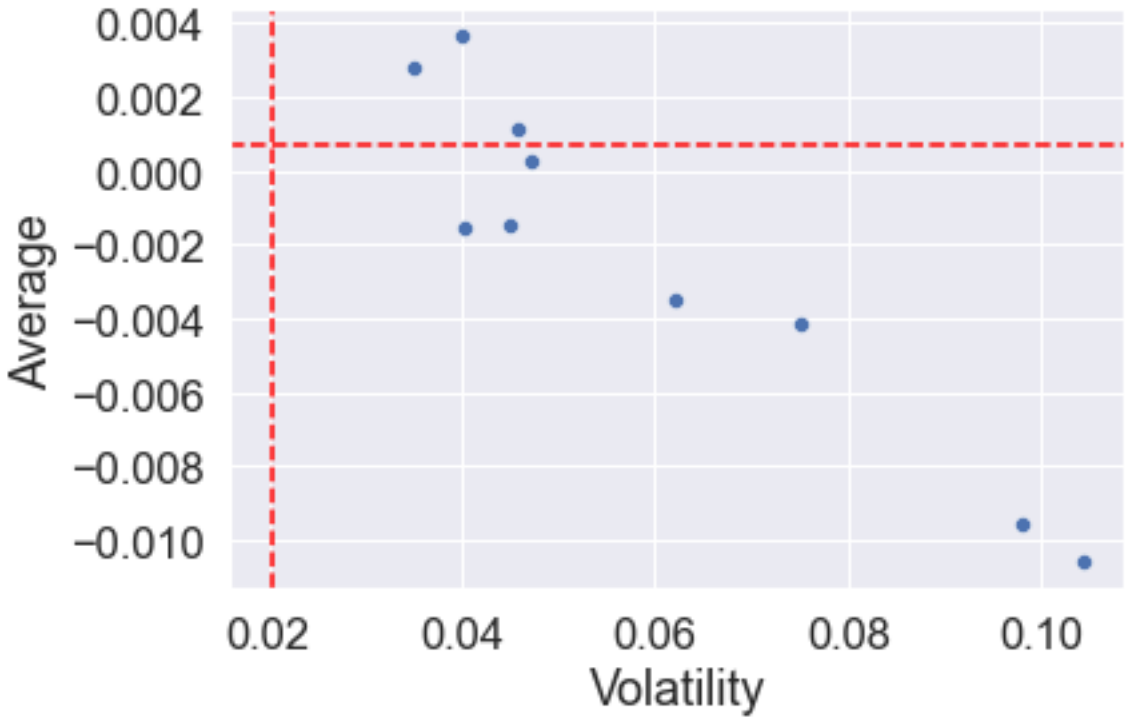
The intersection point of two lines is sensex.

Low risk and high returns are preferred.

Stocks having lower volatility of less than 0.04 have higher returns.

OUTPUT:

Table 2.4a. Scatter plot for stock means and standard deviation



2.5 Conclusion and Recommendations

INSIGHTS:

6 years of weekly stock information pertaining to stock prices of 10 different Indian Stocks are provided in the dataset of "Market Risk Dataset" as an excel file.

To calculate Stock Means and Standard Deviation for all stocks with inference.

APPROACH:

1. Importing the libraries of numpy, pandas, matplotlib, seaborn sklearn metrics.
2. Read the dataset 'Market+Risk+Dataset.csv'.
3. Fixing messy column names for easy using.
4. Glimpse on the head of the dataset.
5. Shape of the dataset.
6. Information on datatypes.
7. Checking basic measures of descriptive statistics.
8. Calculate average (mean) and volatility (standard deviation) for the dataset.

9. Create a scatter plot by adding specific lines to the scatter plot corresponding to the mean (y-axis) and standard deviation (x-axis)

INFERENCES:

RECOMMENDATION:

Based on the average and volatility as well as evident from plot Shree_Cement, Infosys and Axis_bank have high returns on positive side. Among these Shree_Cement is having low risk followed by Infosys.

CONCLUSION:

Low risk and high returns are considered as best stocks and for the data provided, Shree_Cement is best followed by Infosys, Axix_Bank.