**Business Report on**

***FRA Project (Milestone-2)***

***Submitted to***



**Great Learning Olympus**

***Prepared by***

**Mrinmoy Majumdar**

**June-D Batch**

**Post Graduate Program in Data Science & Business Analytics**

**From**

****

UT Austin

**May, 2022**

TABLE OF CONTENT

**1.**8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach………………………………………………………………4

**1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model**………………6

**1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach……………………………………………………………………….**7

###### **1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model**…………………………….9

**1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve) …………………………………………………………**10

1.13 State Recommendations from the above models………………………...12

2.1 Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference……………………………………………………………………….13

2.2 Calculate Returns for all stocks with inference……………………………13

2.3 Calculate Stock Means and Standard Deviation for all stocks with inference………………………………………………………………………15

2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference………………………………………………………………………17

2.5 Conclusion and Recommendations……………………………………….17

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **Images** | **Page No** |
| **Confusion matrix** | **5** |
| **Classification report** | **5** |
| **ROC curve** | **5** |
| **Shape of dataset** | **12** |
| **Information on dataset** | **13** |
| **Stock price graph** | **14** |
| **Stock means** | **16** |
| **Stock means vs stock standard deviation** | **17** |

**LIST OF TABLES**

|  |  |
| --- | --- |
| Table 1 | 10 |
| Table 2 | 12 |
| Table 3 | 12 |
| Table 4 | 13 |
| Table 5 | 15 |
| Table 6 | 16 |

**Problem Statement:**

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 Net worth of the company in the following year (2016) is provided which can be used to drive the labelled field.

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

**Q1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach**

Grid Search CV technique is used to fit Random Forest model into the dataset. Below is the list of hyper parameters used for training the model –

* N estimators = number of trees in the forest
* Max depth = maximum number of levels in each decision tree
* Min samples split = minimum number of data points placed in a node before the node is split
* Min samples leaf = minimum number of data points allowed in a leaf node

Grid parameters selected for the model -

GridSearchCV (estimator=RandomForestClassifier (),

param\_grid= {'max\_depth': [3, 5, 7],

'min\_samples\_leaf': [5, 10, 15],

'min\_samples\_split': [15, 30, 45],

'n\_estimators': [25, 50]})

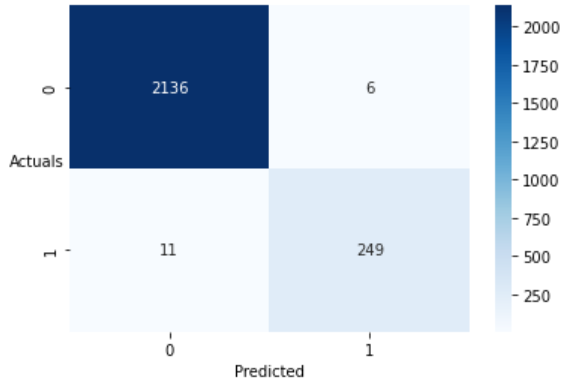
Best grid parameters –

{'max\_depth': 7,

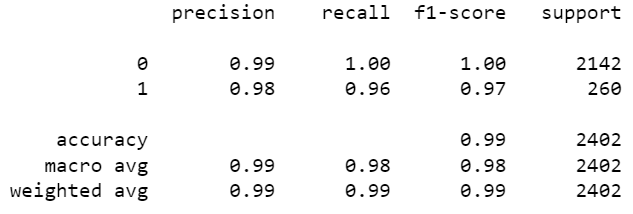
'min\_samples\_leaf': 5,

'min\_samples\_split': 30,

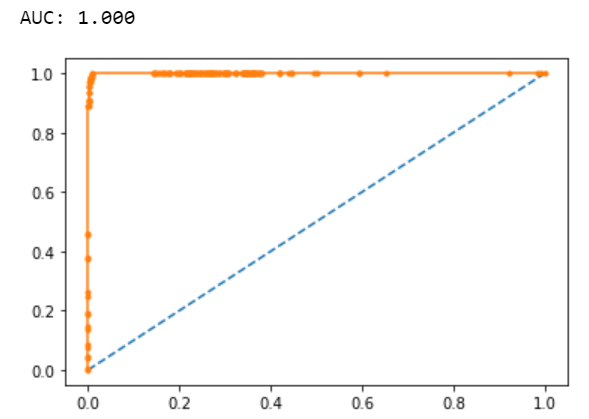
'n\_estimators': 25}



**Fig 1: Confusion matrix for RF training data**

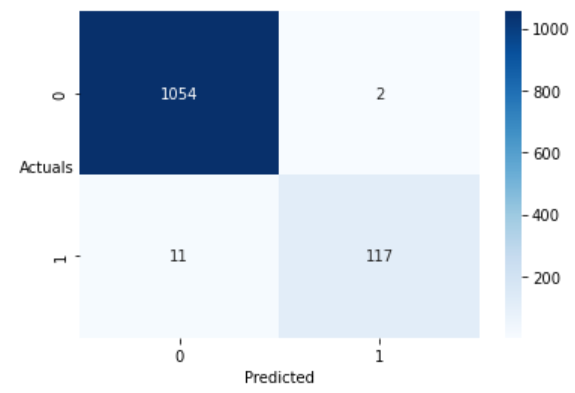


**Fig 2: Classification report for RF training data**

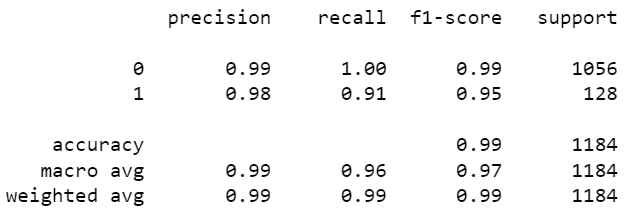


**Fig 3: ROC curve for RF training data**

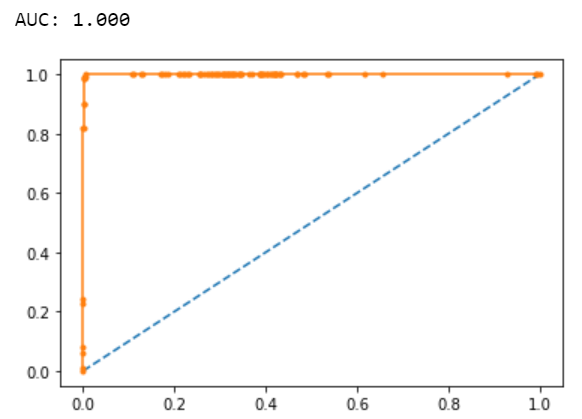
**Q1.9. Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model**



**Fig 4: Confusion Matrix for RF testing data**



**Fig 5: Classification report for RF testing data**



**Fig 6: ROC curve for RF testing data**

Performance metrics used for Random Forest Model –

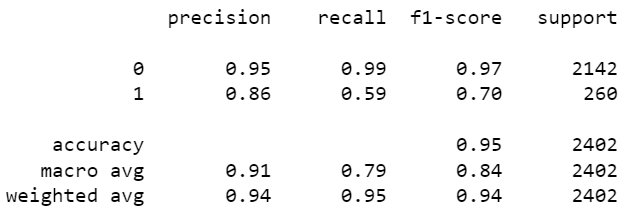
* Confusion Matrix
* Classification Report
* ROC-AUC curve

Interpretations of the model –

* The model performs well based on train and test data based on the accuracy, recall, precision and F1-scores
* ROC curves are identical for train and test samples indicating a perfectly built model.

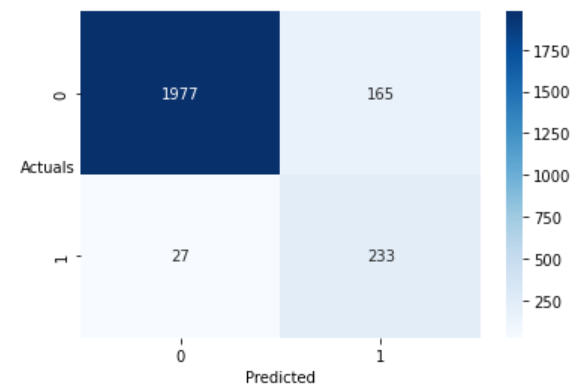
**Q1.10. Build a LDA Model on Train Dataset. Also showcase your model building approach**

Linear Discriminant Analysis (LDA) is a dimensionality reduction technique which is commonly used for the supervised classification problems. It is used to project the features in higher dimension space into a lower dimension space. Basic hyperparameters were used to fit the LDA model. The threshold value is assumed to be 0.05.

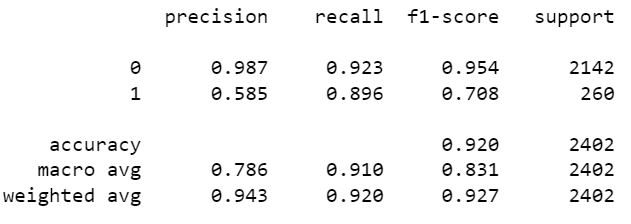


**Fig 7: Classification report for LDA training data**

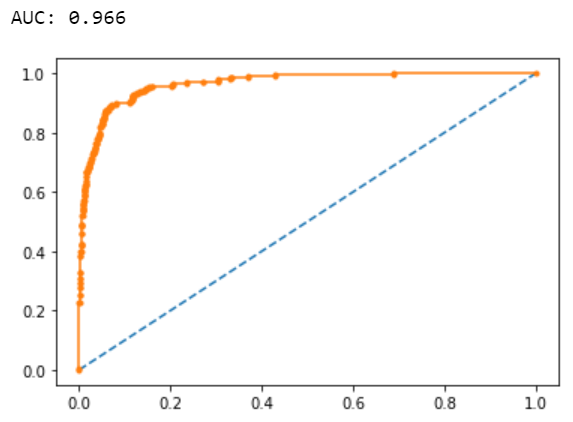
Optimal threshold is calculated as 0.08 and the model is built on the threshold value.



**Fig 8: Confusion matrix for LDA training data**

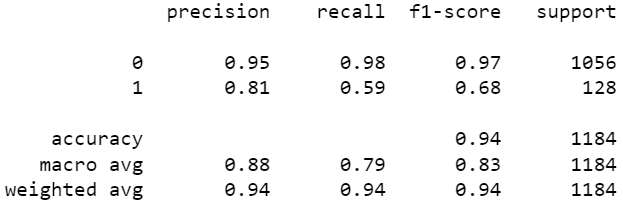


**Fig 9: Classification report for LDA training data with revised threshold**

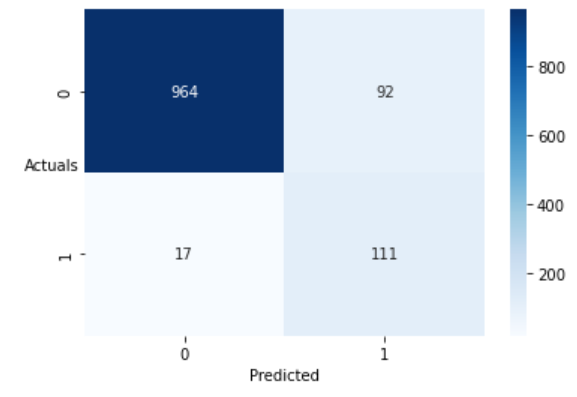


**Fig 10: ROC curve for LDA train**

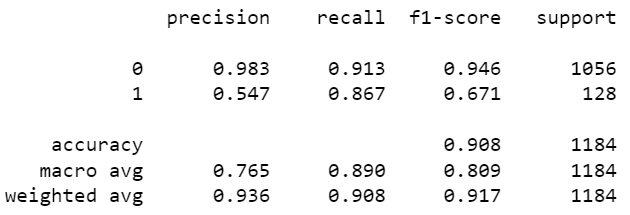
**Q1.11. Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model**



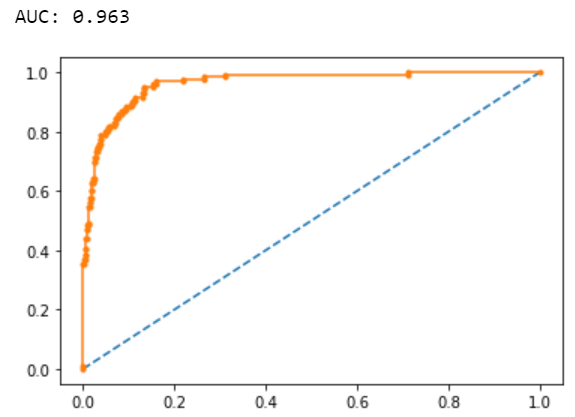
**Fig 11: Classification report for LDA test**



**Fig 12: Confusion matrix for LDA test**



**Fig 13: Classification report for LDA test with revised threshold**

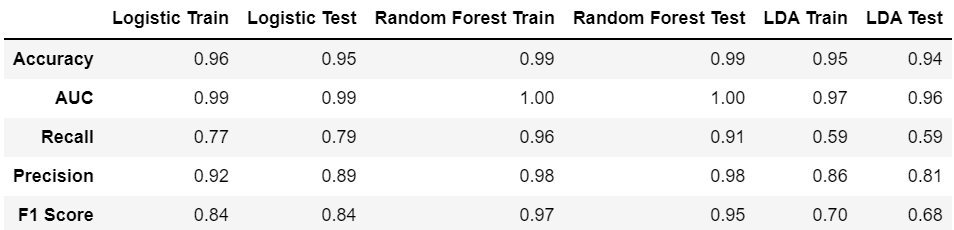


**Fig 14: ROC curve for LDA test**

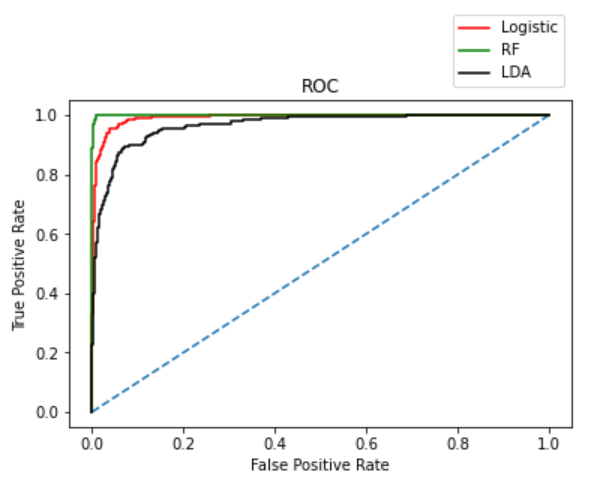
Interpretations of the model -

* Initially the threshold was considered as 0.05 which yielded a poor recall score for both train and test samples.
* The threshold was revised to 0.08 which improved the recall score considerably.
* The accuracy and ROC scores are identical for train and test data which signifies a well build model.

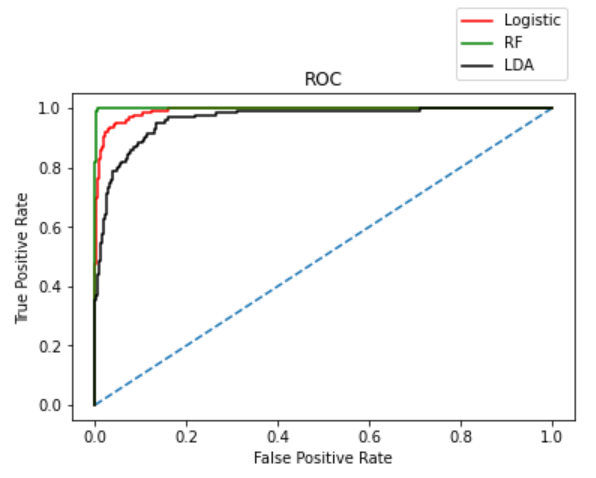
**Q1.12. Compare the performances of Logistics, Random Forest and LDA models (include ROC Curve)**



**Table 1: Performance comparison of models**



**Fig 15: ROC curve comparison with train data**



**Fig 16: ROC curve comparison with test data**

Inferences –

* Random Forest is the best fit model as it has cleanly predicted the Defaulters and Non-Defaulters without any False Predictions.
* Logistic Regression and Linear Discriminant Analysis Models do not have overfitting or underfitting issues.

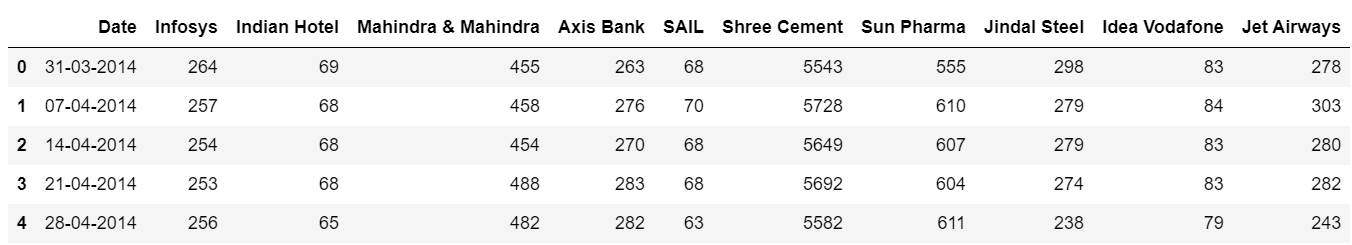
**Q 1.13. State Recommendations from the above models**

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

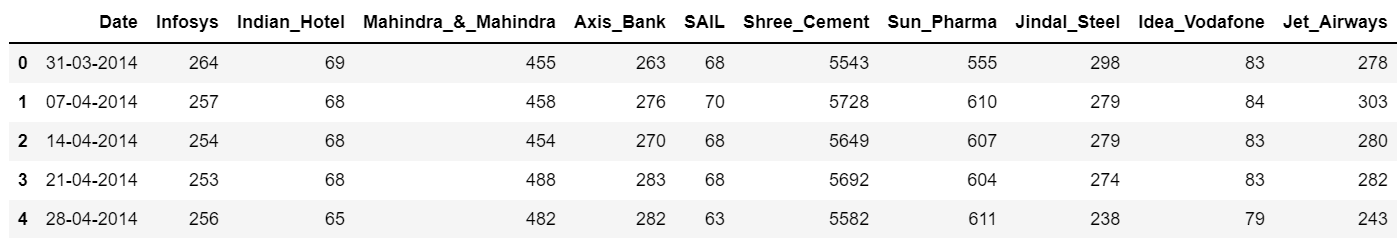
**Problem Statement:**

**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.**



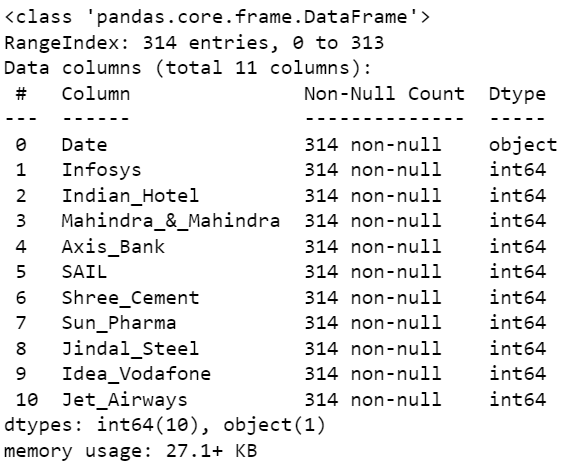
**Table 2: Head of stock price table**



**Table 3: Head of dataset with messy columns fixed**



**Fig 17: Shape of stock dataset**



**Fig 18: Information on stock dataset**

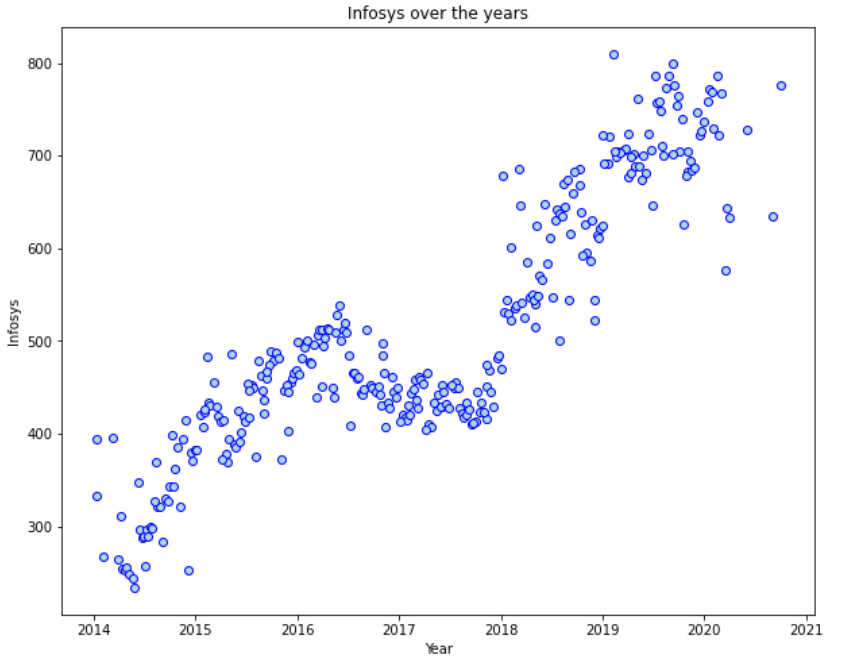


**Table 4: Description of dataset**

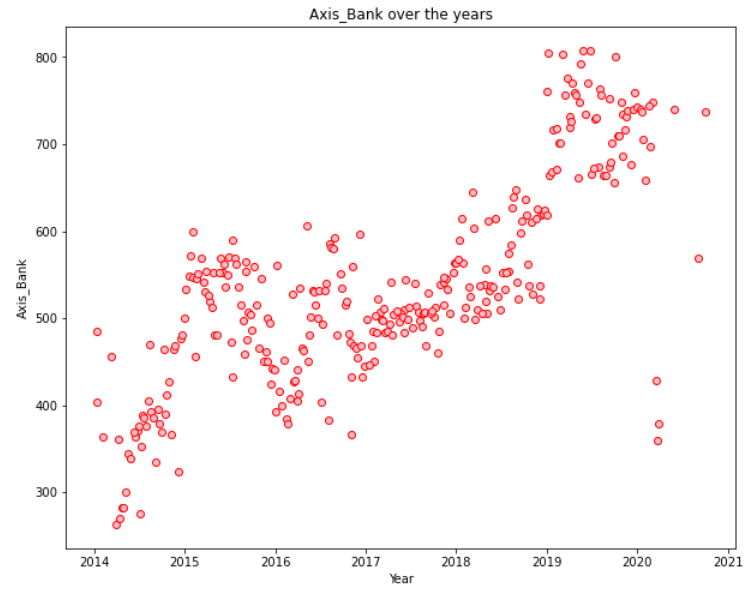
Insights from the data–

* There are no null values present in the dataset.
* Duplicate values are absent from the data provided.
* Dataset contains 314 rows and 11 columns.
* 10 variables are of integer type and 1 is of object type.
* Summary of data indicates that most of the features have normal distribution.

**Q2.1. Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference**



**Fig 19: Stock price graph for Infosys**



**Fig 20: Stock price graph for Axis Bank**

Inference -

* Axis Bank has sharp increasing trend that can be observed from Year 2014 till mid 2015 then some dip in stock price.
* Again, an increasing trend can be observed from 2017 which has continued till 2020.
* Infosys has a sharp increasing trend which can be observed from year 2014 to 2015 year-end followed by a dip in stock price between 2016 - 2018.
* An increasing trend can be observed from 2018 which has continued till 2020.

**Q2.2. Calculate Returns for all stocks with inference**

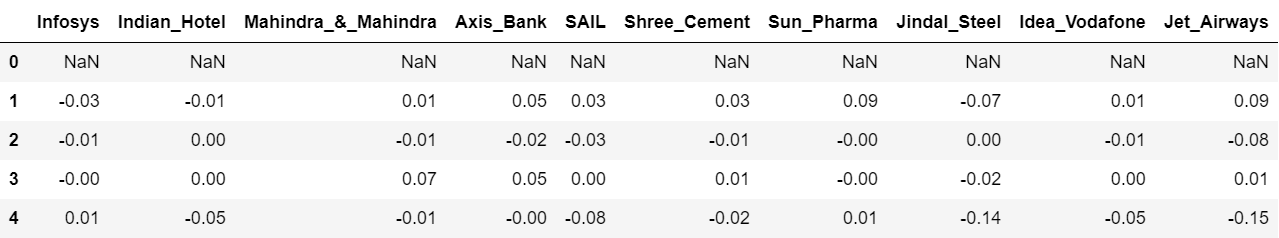
Steps for calculating returns from prices:

* Take logarithms
* Take differences

Here, we will use logarithm method to analyse the returns of these stock prices available which is difference between two consecutive days prices. In this method, we are calculating Returns for all the stocks i.e., difference of log of price at t and the log of price at t-1.

We will be using np.log of stock prices and dropping the “Dates” column, thereafter checking shape and head of data to observe the changes made.

Shape of dataset – (314,10)



**Table 5: Returns for all stocks**

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.

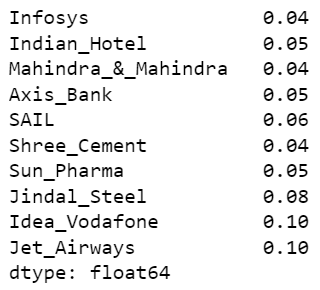
**Q2.3. Calculate Stock Means and Standard Deviation for all stocks with inference**

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

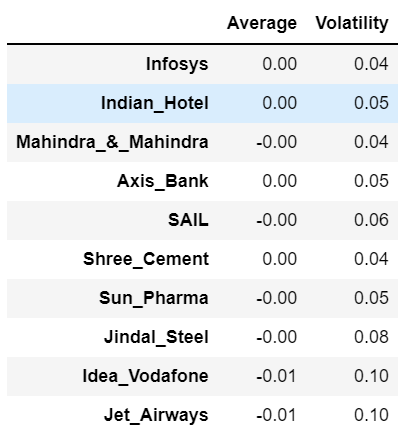
Stock Standard Deviation: It is a measure of volatility meaning the more a stock's returns vary from the stock's average return, the more volatile the stock.



**Fig 21: Stock means for features**



**Fig 22: Stock standard deviation for features**

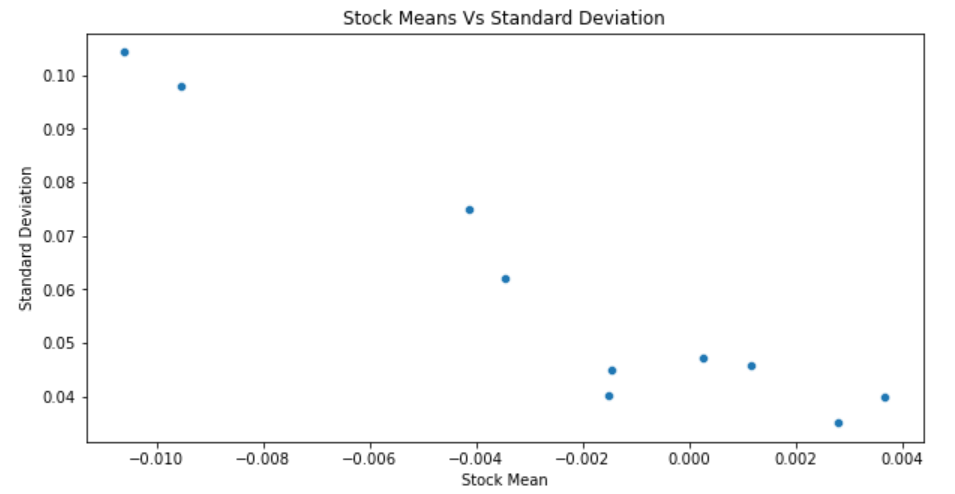


**Table 6: Average and Volatility table**

Inferences –

* Idea Vodafone and Jet Airways have highest volatility and lowest return which signifies higher risk.
* Shree Cement has the highest returns and one of the least risky stocks.
* Infosys is the least risky investment option.

**Q2.4. Draw a plot of Stock Means vs Standard Deviation and state your inference**



**Fig 23: Stock means vs standard deviation**

Stocks 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. During the investment, this graph is very useful in analysing the risk from different companies.

**Q2.5. Conclusion and Recommendations**

Conclusion -

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: Ones with higher return for a comparative or lower risk are considered better.

Shree Cement followed by Infosys & Axis Bank looks good in this dataset based on returns.

Infosys followed by Shree Cement & Mahindra & Mahindra looks good in this dataset based on risk.

Recommendation –

We would recommend using the stock means vs standard deviation plot to assess the risk to reward ratio. The smaller the standard deviation, an investment will be the less risky. On the other hand, the larger the variance and standard deviation, the more volatile a security. 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. As with anything else, the greater the number of possible outcomes, the greater the risk of choosing the wrong one. Recommendations can be made based on the significance and type of investment that an investor is interested in.