# Finance and Risk Analytics PROJECT REPORT

## Mile stone 2

By:

Sharjil Shah

PGP DSBA Jan 22A

Great Learning

Table of Contents

[Predicting Credit Risk 3](#_Toc119265771)

[PROBLEM 1 3](#_Toc119265772)

[INTRODUCTION 4](#_Toc119265773)

[1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach 4](#_Toc119265774)

[1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model. 8](#_Toc119265775)

[1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach 9](#_Toc119265776)

[1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model 12](#_Toc119265777)

[1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve) 14](#_Toc119265778)

[1.13 State Recommendations from the above models 15](#_Toc119265779)

[PROBLEM 2 16](#_Toc119265780)

[Introduction 16](#_Toc119265781)

[2.1 Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference 19](#_Toc119265782)

[2.2 Calculate Returns for all stocks with inference 21](#_Toc119265783)

[2.3 Calculate Stock Means and Standard Deviation for all stocks with inference 22](#_Toc119265784)

[2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference 24](#_Toc119265785)

[2.5 Conclusion and Recommendations 24](#_Toc119265786)

LIST OF FIGURES

[Figure 1- Random Forest\_Confusion Matrix on Train Data 7](#_Toc119265978)

[Figure 2 - Random Forest AUC-ROC for Train Data 8](#_Toc119265979)

[Figure 3 - Random Forest\_Confusion Matrix on Test Data 9](#_Toc119265980)

[Figure 4 - LDA Confusion Matrix for Train Data 11](#_Toc119265981)

[Figure 5 - LDA AUC-ROC Curve for train data 12](#_Toc119265982)

[Figure 6 - LDA Confusion Matrix for test data 13](#_Toc119265983)

[Figure 7 -LDA AUC-ROC for Test Data 14](#_Toc119265984)

[Figure 8 - data set sample 17](#_Toc119265985)

[Figure 9 - Checking duplicates 18](#_Toc119265986)

[Figure 10 - Checking null values 19](#_Toc119265987)

[Figure 11 - Descriptive stats 19](#_Toc119265988)

[Figure 12 - Axis bank stock price vs year 20](#_Toc119265989)

[Figure 13 - Infosys stock price v/s year 21](#_Toc119265990)

[Figure 14 - difference of log of price at t and the log of price at t-1 22](#_Toc119265991)

[Figure 15 - Stock Price Cumulative Return Plot 23](#_Toc119265992)

[Figure 16 - Avg and Volatility of the stock 24](#_Toc119265993)

LIST OF TABLES

[Table 1 - highest contributing independent variables 7](#_Toc119265998)

[Table 2 - Classification report for train data 8](#_Toc119265999)

[Table 3 - Classification report for test data 9](#_Toc119266000)

[Table 4 - Random Forest AUC-ROC for Test Data 10](#_Toc119266001)

[Table 5 - Classification report for train data 11](#_Toc119266002)

[Table 6 - Classification report for LDA test data 13](#_Toc119266003)

[Table 7 - Comparison between confusion matrices for different models 15](#_Toc119266004)

[Table 8 - Classification report for different train and test models 16](#_Toc119266005)

[Table 9 - datatypes 18](#_Toc119266006)

[Table 10 - Calculations of Stock mean 24](#_Toc119266007)

[Table 11 - Calculations of Stock Standard deviation 24](#_Toc119266008)

# Predicting Credit Risk

## PROBLEM 1

#### SUMMARY

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'*

Dependent variable - We need to create a default variable which should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

Test Train Split - 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.

## iNTRODUCTION

This assignment helps us to perform Outlier Treatment, Missing Value Treatment, Transform Target variable into 0 and 1, Univariate and Bivariate Analysis, Split data into Train & Test, Model Building is to be done on Train Dataset and choose the optimum cutoff, 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.

We have 3586 entries and 67 columns. The outcome of this assignment will suggesting investors good credit rating companies to invest their money.

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

#### aPPROACH:

* Import libraries.' numpy, python, matplotlib, seaborn, sklearn criteria.
* Read the dataset(Company\_Data2015.xlsx)
* Check on the dataset viz., head, shape, info, describe
* Fixing grubby column names( like, spaces, unwanted characters etc.) without changing the meaning for ease of use and fixing visualize the data
* Dropping the repeated variables.
* Checking and eliminating duplicates, if any.
* Checking missing values in the data set.
* Examining for outliers in the data set.
* 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 and Splitting of the data into Train & Test dataset in a ratio of 67:33 and using random\_state =42.

#### Random Forest Model Buliding Approach:

Scaling is not required for Random forest because the tree based models are not calculated over distances. We can handle various range of features.

We used GridSearchCSV for hyper parameter tuning. We used same grid for two models with or without SMOTE to maintain constancy.

We approached Recursive Feature Elimination (RFE) for model building. We wanted to selected 1/3rd of total feature variables i.e., 67/3 ~ 21 features that would contribute to the model well.

We provide weightage and ranking to each variable and we used Logistic Regression with recursive feature elimination.

##### The list of independent variables those have highest contribution to the model is below.

|  |  |  |
| --- | --- | --- |
|  | **Feature** | **Rank** |
| **1** | Networth | 1 |
| **2** | Capital\_Employed | 1 |
| **3** | Total\_Debt | 1 |
| **5** | Net\_Working\_Capital\_ | 1 |
| **8** | Total\_Assets/Liabilities\_ | 1 |
| **15** | PBIDT | 1 |
| **16** | PBDT | 1 |
| **17** | PBIT | 1 |
| **18** | PBT | 1 |
| **19** | PAT | 1 |
| **20** | Adjusted\_PAT | 1 |
| **21** | CP | 1 |
| **25** | Book\_Value\_Unit\_Curr | 1 |
| **26** | Book\_Value\_Adj.\_Unit\_Curr | 1 |
| **28** | CEPS\_annualised\_Unit\_Curr | 1 |
| **32** | ROG\_Net\_Worth\_perc | 1 |
| **48** | Current\_Ratio[Latest] | 1 |
| **53** | Interest\_Cover\_Ratio[Latest] | 1 |
| **54** | PBIDTM\_perc[Latest] | 1 |
| **55** | PBITM\_perc[Latest] | 1 |
| **63** | Value\_of\_Output/Gross\_Block | 1 |

Table 1 - highest contributing independent variables

##### Performance metrics on train data for random forest

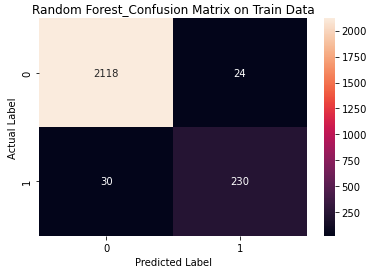


Figure 1- Random Forest\_Confusion Matrix on Train Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.99 | 0.99 | 0.99 | 2142 |
| 1 | 0.91 | 0.88 | 0.89 | 260 |
|  |  |  |  |  |
| accuracy |  |  | 0.98 | 2402 |
| macro avg | 0.95 | 0.94 | 0.94 | 2402 |
| weighted avg | 0.98 | 0.98 | 0.98 | 2402 |

Table 2 - Classification report for train data

##### Random Forest AUC-ROC Curve on Train Data

Area under Curve is 0.9961322990734754

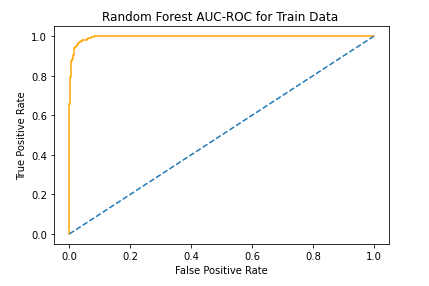


Figure 2 - Random Forest AUC-ROC for Train Data

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

##### Performance metrics on test data for random forest

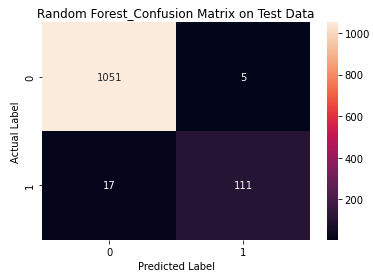


Figure 3 - Random Forest\_Confusion Matrix on Test Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.98 | 1 | 0.99 | 1056 |
| 1 | 0.96 | 0.87 | 0.91 | 128 |
|  |  |  |  |  |
| accuracy |  |  | 0.98 | 1184 |
| macro avg | 0.97 | 0.93 | 0.95 | 1184 |
| weighted avg | 0.98 | 0.98 | 0.98 | 1184 |

Table 3 - Classification report for test data

##### Random Forest AUC-ROC Curve on test Data

Area under Curve is 0.9903601444128788

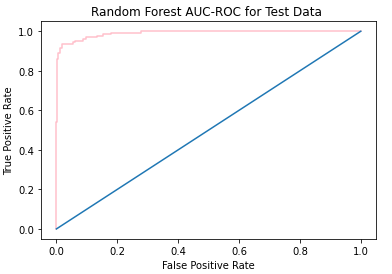


Table 4 - Random Forest AUC-ROC for Test Data

From the Milestone 1, we have seen that the Random forest with SMOTE was worse. Without SMOTE the model has higher accuracy. Although there is no significant difference in the accuracies of both models. The accuracy - 98%, precision - 96%, recall - 87% and 91%, and F1 - score of 0.9701 and 0.996038 respectively. Random Forest AUC curve on Test & Train data 9903601444128788 and 0. 9961322990734754.

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

As we see that the coefficients are varying between classes, we don’t scale the data. So scaling wasn’t necessary. We used GridsearchCSV for hyper parameter tuning. Same grid was used for both models (With & without SMOTE) to maintain constancy.

##### Performance metrics on train data for LDA model data.

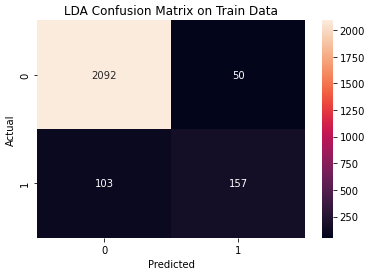


Figure 4 - LDA Confusion Matrix for Train Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.95 | 0.97 | 0.96 | 1056 |
| 1 | 0.74 | 0.61 | 0.67 | 128 |
|  |  |  |  |  |
| accuracy |  |  | 0.93 | 1184 |
| macro avg | 0.85 | 0.79 | 0.82 | 1184 |
| weighted avg | 0.93 | 0.93 | 0.93 | 1184 |

Table 5 - Classification report for train data

##### LDA AUC-ROC Curve for train Data

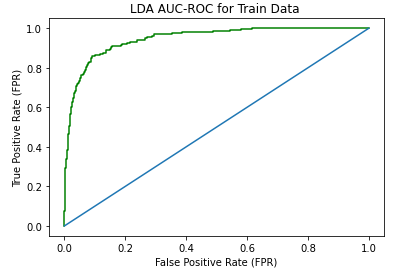


Figure 5 - LDA AUC-ROC Curve for train data

Area under Curve is 0.946468074409251

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

##### Performance metrics on t data for LDA model data

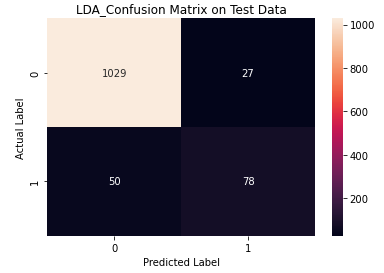


Figure 6 - LDA Confusion Matrix for test data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.95 | 0.97 | 0.96 | 1056 |
| 1 | 0.74 | 0.61 | 0.67 | 128 |
|  |  |  |  |  |
| accuracy |  |  | 0.93 | 1184 |
| macro avg | 0.85 | 0.79 | 0.82 | 1184 |
| weighted avg | 0.93 | 0.93 | 0.93 | 1184 |

Table 6 - Classification report for LDA test data

##### LDA AUC-ROC Curve for train Data

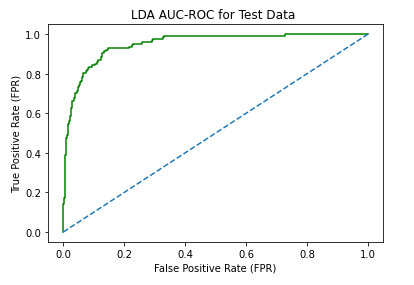


Figure 7 -LDA AUC-ROC for Test Data

LDA without Smote on test data is with 93% accuracy, 74% precision, 61% recall and 67% F1 score respectively. LDA AUC curve on Test data is 0.9504616477272727

Model without SMOTE performs better. However, in the real world, the model with SMOTE performs better.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Logistic Model | |  | | --- | |  | | |  | | --- | |  | |
| Random Forest Model |  |  |
| LDA model |  |  |

Table 7 - Comparison between confusion matrices for different models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL\_NAME** | **Accuracy** | **Recall** | **Precision** | **F1-Score** |
| *Logistic Regression\_ Train* | 93% | 95% | 92% | 94% |
| *Logistic Regression\_ Test* | 92% | 93% | 58% | 71% |
| *Random Forest\_ Train* | 98% | 88% | 91% | 89% |
| *Random Forest\_ Test* | 98% | 87% | 96% | 91% |
| *LDA\_ Train* | 94% | 60% | 76% | 67% |
| *LDA\_ Test* | 93% | 61% | 74% | 67% |

Table 8 - Classification report for different train and test models

#### Inference:

1. Random forest model is the best fit model among the three models built. It predicted the defaulters and non-defaulters more accurately than other models.
2. Precision and Recall are the most important measures of quality for the business. There needed to be less false positives. Having more false positives and false negatives would cause loss to the company.
3. Random forest has its limitation. Because the train and test data are same, there is possibility of overfitting model. Due to overlapping, same model may not be best fit for other datasets. The predictions vary from datasets to datasets.
4. Logistic regression is mostly the model that 90% of the times identifies the defaulted company accurately.
5. Precision and Recall for Logistic Regression( LR) & Linear Discriminant Analysis( LDA) Test data are nearly analogous but Recall for LR is hardly better than LDA model.

## 1.13 State Recommendations from the above models

Apparently Logistic Regression is the best fit model for the given dataset. Here are the recommendations based on LR model.

1. Lower the ***Book\_value\_unit\_curr*** , higher the chances of default.
2. Lower the ***CEPS\_annualised\_Unit\_Curr,*** higher the chances of default.
3. Higher the ***Curr\_Ratio\_Latest***, lower the changes of default.
4. Higher the ***Interest\_Cover\_Ratio\_Latest*** lower the chances of default.
5. ***Curr\_Ratio\_Latest*** is the most important parameter whereas ***Interest\_Cover\_Ratio\_Latest*** is the least important feature for any defaulter.

# PROBLEM 2

### Introduction

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 from the data set given. The data set contains weekly stock information of 10 different stocks for 6 years. We examine the mean and standard deviation on the stock returns and provide useful insights for the given listed companies.

#### EDA:

Dataset:

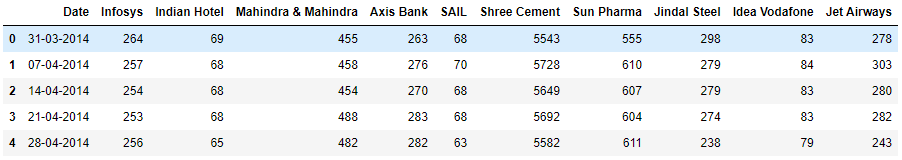
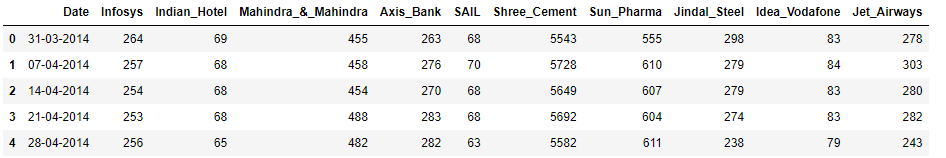


Figure 8 - data set sample

Altering the column names in Dataset for ease:

We removed unwanted spaces from the column names for ease of use and visualization.



##### Shape

* The number of rows is 314
* The number of columns is 11

##### Datatypes of columns

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 314 entries, 0 to 313

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Date 314 non-null object

1 Infosys 314 non-null int64

2 Indian\_Hotel 314 non-null int64

3 Mahindra\_&\_Mahindra 314 non-null int64

4 Axis\_Bank 314 non-null int64

5 SAIL 314 non-null int64

6 Shree\_Cement 314 non-null int64

7 Sun\_Pharma 314 non-null int64

8 Jindal\_Steel 314 non-null int64

9 Idea\_Vodafone 314 non-null int64

10 Jet\_Airways 314 non-null int64

dtypes: int64(10), object(1)

memory usage: 27.1+ KB

Table 9 - datatypes

We changed the data type for Date column from ‘object’ to ‘datetime64[ns]

##### Duplicates



Figure 9 - Checking duplicates

There are zero duplicates in the given dataset

##### Null values

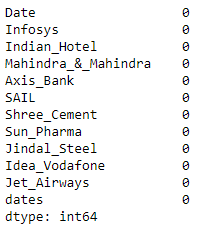


Figure 10 - Checking null values

There are zero null values in the given data set

#### Inference

* Dataset contains 6 years of 10 different Indian stocks prices in weekly intervals from 31- 03-2014 to 30-03-2020.
* After changing formats, there are total of 314 rows and 11 columns in the dataset.
* No missing values and duplicates

##### Descriptive analysis

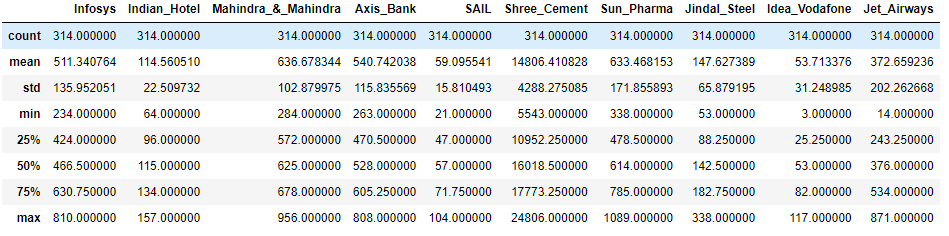


Figure 11 - Descriptive stats

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

#### axis Bank stock

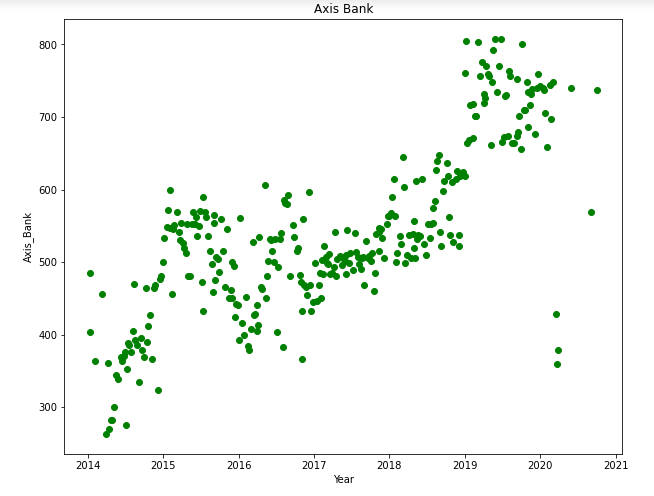


Figure 12 - Axis bank stock price vs year

#### Inference

* The price of the stock is in increasing trend over the six years I.e., from 2014 to 2021. The investors who purchased shares in 2014 may have received 37 times return in 2021 provided if the stocks were sold at the right time.
* The stock was highly bullish between 2014 and 2015. Bearish from 2015 to 2017.
* Lowest price: 263.0 Highest: 808.00

#### Infosys Stock

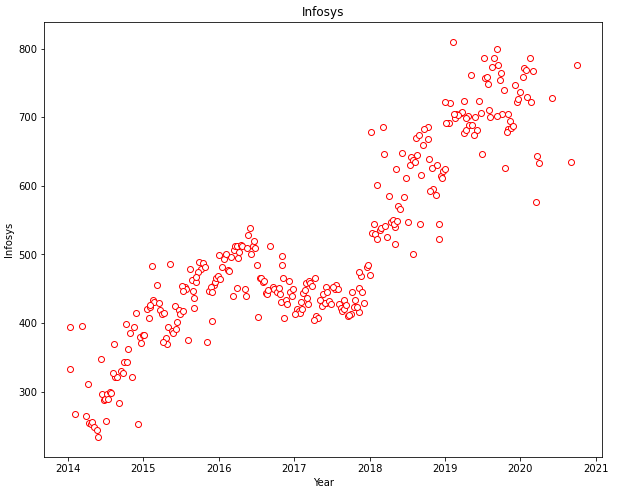


Figure 13 - Infosys stock price v/s year

#### Inference

* Stock price of the Infosys from 2014 and 2021 is in increasing trend. There is sharp increase in the price between 2014 and 2016 then it stays flat till 2018 and again rises up till 2021.
* This listed stock may have provided high returns to those who invested for long term.
* Highest price: 234.00 Lowest price: 810.0
* Very bullish stock

## 2.2 Calculate Returns for all stocks with inference

We use logarithmic method to analyze the returns of the stock prices. The returns are defined by the difference between the stock prices between consecutive weeks. We use np.log of stock prices and dropping the “Dates” column. Later we check shape and head of data to observe the changes made.

After using Logarithm Method

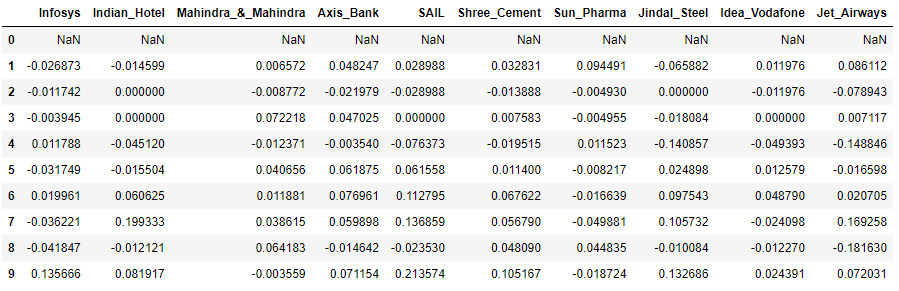


Figure 14 - difference of log of price at t and the log of price at t-1

Shape of Data



The number of rows in the data is 314. The number of columns in the data is 10.

We can observe the first value is ‘NaN’. This is because the first row doesn’t have previous value to determine the difference.

##### Stock Price Cumulative Return Plot

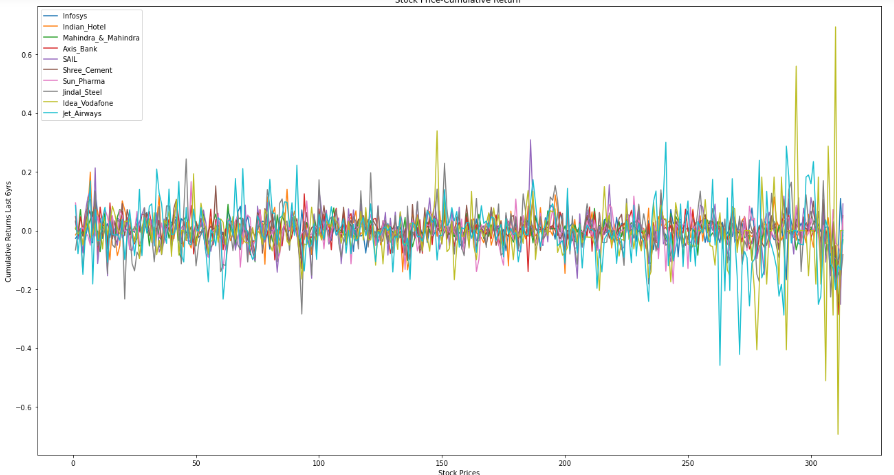


Figure 15 - Stock Price Cumulative Return Plot

#### INFERENCE

* We observe that maximum cumulative return is maximum for Idea, Vodafone and Jet Airways.
* The stocks are highly fluctuating over the weeks.

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

#### Inference

* Stock mean: Average returns of the stock on weekly basis.
* Stock Standard Deviation: It is a measure of Volatility. It is described as the rate at which the stock price is increasing or decreasing over the period of time.
* Higher stock volatility means higher the risk.

|  |  |
| --- | --- |
| Stock name | Standard deviation |
| Infosys | 0.03507 |
| Indian\_Hotel | 0.047131 |
| Mahindra\_&\_Mahindra | 0.040169 |
| Axis\_Bank | 0.045828 |
| SAIL | 0.062188 |
| Shree\_Cement | 0.039917 |
| Sun\_Pharma | 0.045033 |
| Jindal\_Steel | 0.075108 |
| Idea\_Vodafone | 0.104315 |
| Jet\_Airways | 0.097972 |

|  |  |
| --- | --- |
| Stock\_Name | Mean |
| Infosys | 0.002794 |
| Indian\_Hotel | 0.000266 |
| Mahindra\_&\_Mahindra | -0.001506 |
| Axis\_Bank | 0.001167 |
| SAIL | -0.003463 |
| Shree\_Cement | 0.003681 |
| Sun\_Pharma | -0.001455 |
| Jindal\_Steel | -0.004123 |
| Idea\_Vodafone | -0.010608 |
| Jet\_Airways | -0.009548 |

Table 10 - Calculations of Stock mean

Table 11 - Calculations of Stock Standard deviation

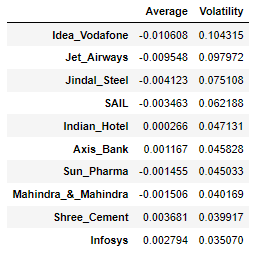
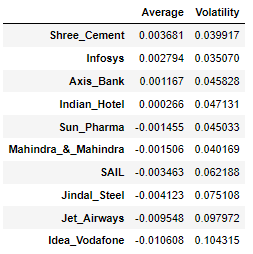
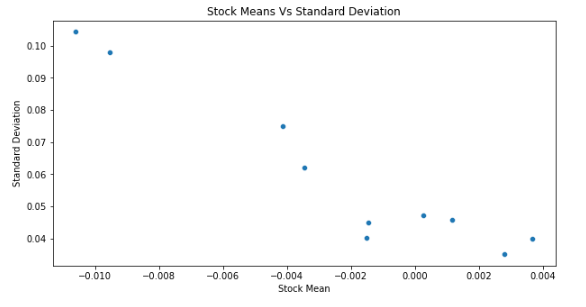


Figure 16 - Avg and Volatility of the stock

* Idea-Vodafone has the lowest return and highest volatility.
* Shree\_Cement on the other hand has the highest returns..
* Infosys and Shree\_Cement are the least risky stocks.
* Higher the volatility higher the risk and lower the volatility lower the risk.

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



#### Inference

* This graph can be used to interpret risk and reward.
* The stocks in the left bottom indicates low volitity or high risk and high returns.

## 2.5 Conclusion and Recommendations

Stock with a lower mean & higher standard deviation doesn’t have significance in a portfolio with a competing stock giving more returns & less risk. Stock recommendations could be listed as one with highest return and lowest risk and one with lowest risk and highest return. Purely from Risk perspective Infosys, Shree Cement, and Mahindra & Mahindra are good stocks. From Return perspective Shree Cement, Infosys & Axis Bank are good stocks. More volatile stocks might give short return gains and they are not good for long term investment. On the other way around, less volatile stocks can give high returns over long time. If an investor ready to bear high risk, then we can recommend Idea Vodafone, Jet Airways and Jindal Steel. For long term returns, we can recommend the investor Shree Cement, Infosys & Axis Bank. For medium risk investor, we can recommend Indian Hotel and Axis Bank. The inference can be made accordingly to create diverse portfolio for the investor.

\*-\*-\*-\*