FRA Milestone 2 Project

PGP – DSBA

May ‘22

Date: 02/04/2023

Swaathi Ramakrishnan

**Table of Contents**

|  |  |
| --- | --- |
| **Contents** | **Page Number** |
|  |  |
| Problem Statement | 4 |
|  |  |
| 1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach | 5 |
|  |  |
| 1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model | 7 |
|  |  |
| 1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach | 8 |
|  |  |
| 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 | 11 |
|  |  |
| 2. Market Risk | 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 | 15 |
|  |  |
| 2.3 Calculate Stock Means and Standard Deviation for all stocks with inference | 16 |
|  |  |
| 2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference | 17 |
|  |  |
| 2.5 Conclusion and Recommendations | 18 |

**List of Tables/Figures**

|  |  |
| --- | --- |
| Fig 1. Random Forest best estimators | 5 |
| Fig 2. RF Confusion Matrix and Classification Report for Model 1 Train data | 5 |
| Fig 3. Random Forest SMOTE best estimators | 6 |
| Fig 4. RF SMOTE Confusion Matrix and Classification Report for Model 2 Train data | 6 |
| Fig 5. RF Confusion Matrix and Classification Report for Model 1 Test data | 7 |
| Fig 6. ROC\_AUC curve and score for RF Model 1 test data | 7 |
| Fig 7. ROC\_AUC curve and score for RF Model 2 SMOTE test data | 7 |
| Fig 8. LDA Confusion Matrix and Classification Report for Model 1 Train data | 8 |
| Fig 9. LDA SMOTE Confusion Matrix and Classification Report for Model 2 Train data | 8 |
| Fig 10. LDA Confusion Matrix and Classification Report for Model 1 Test data | 9 |
| Fig 11. LDA SMOTE Confusion Matrix and Classification Report for Model 2 Test data | 9 |
| Fig 12. ROC\_AUC Curve for LDA Model 1 | 9 |
| Fig 13. ROC\_AUC Curve for LDA Model 2 | 9 |
| Fig 14. Comparing models | 10 |
| Fig 15. ROC Curve Comparison | 10 |
| Fig 16. Statsmodel Summary | 11 |
| Fig 17. Data Head | 12 |
| Fig 18. Descriptive stats of the dataset | 12 |
| Fig 19. Shree\_Cement stock price graph | 13 |
| Fig 20. Idea vodafone stock price graph | 13 |
| Fig 21. Data Tail | 15 |
| Fig 22. Stock return data description | 15 |
| Fig 23. Stock return mean and SD | 16 |
| Fig 24. Plot of Mean vs SD | 17 |
| Fig 25. Stock Price Movement | 18 |

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

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

Random forest model on the train dataset is built first and then on the oversampled training dataset created using SMOTE.

**Model 1:**

Random forest classifier is run for various parameters and hyper -parameter tuning is performed using gridsearch on train dataset.

* Criterion - gini and entropy
* Train Dataset - 2402 observations
* min\_samples\_leaf - 1% of observations, so around 24
* Hypertuning - 20,30 and 40 - minimum sample leaf
* min\_samples\_split - 3 times split, hence we can use values 60,90 and 120
* minimum depth - 10,15 and 20
* On performing the gridsearch, with cross-validation 5 times on the train dataset, the optimal random forest classifier model has been identified as below:

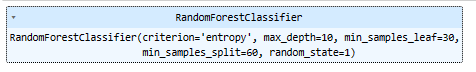


Fig 1. Random Forest best estimators

On predicting the response with above model on training data gives the below confusion matrix and classification report:

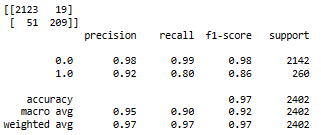


Fig 2. RF Confusion Matrix and Classification Report for Model 1 Train data

The overall accuracy is good at 0.97 and the recall for defaulters are at 0.8 which is good.

The results need to be verified against the test dataset.

**Model 2:**

Random forest classifier is run for various parameters and hyper -parameter tuning is performed using gridsearch on train dataset using SMOTE.

* Criterion - gini and entropy
* Train Dataset - 4284 observations
* min\_samples\_leaf - 1% of observations, so around 24
* Hypertuning – 30, 40 and 50 - minimum sample leaf
* min\_samples\_split - 3 times split, hence we can use values 90, 120 and 150
* minimum depth - 10,15 and 20
* On performing the gridsearch, with cross-validation 5 times on the train dataset, the optimal random forest classifier model has been identified as below:

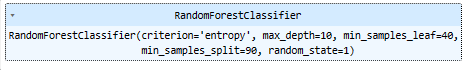


Fig 3. Random Forest SMOTE best estimators

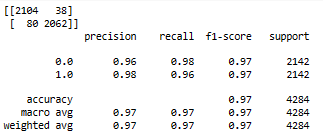


Fig 4. RF SMOTE Confusion Matrix and Classification Report for Model 2 Train data

We can see that the overall accuracy is good at 0.97 and the recall for defaulters are at 0.96 which is better than model1.

We still need to validate these results against the test dataset.

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

Using the 2 random forest models (Model 1 on train dataset and Model2 on the oversample train dataset using SMOTE) on the test dataset and verifying if the models gave good results

**Model 1:**

Test dataset confusion matrix and classification report values

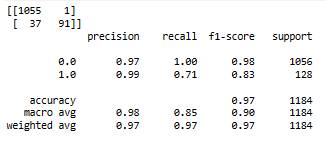


Fig 5. RF Confusion Matrix and Classification Report for Model 1 Test data

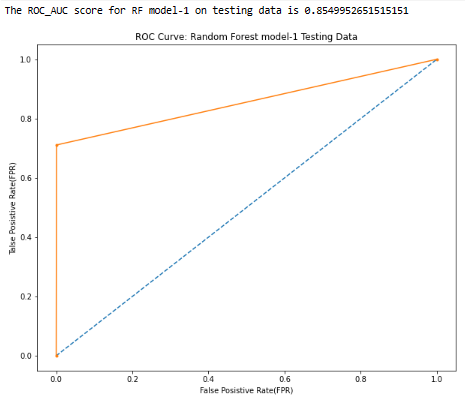
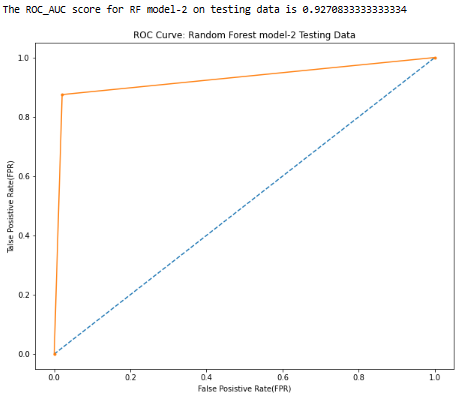


Fig 6. ROC\_AUC curve and score Fig 7. ROC\_AUC curve and score

for RF Model 1 test data for RF Model 2 SMOTE test data

Comparing the 2 models we can see accuracy is same at 0.97. But the ROC score for model 2 (0.927) is better than model 1 (0.855). Model 2 has a much better recall score of 0.88 than model1 at 0.71. Both the models are not either over or underfit. We will consider Model2 as the better one in this scenario.

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

We will build LDA model for train dataset and then on SMOTE.

**Model 1:**

On running LDA on different parameters and after doing hyper-parameter tuning using grid search, the optimal LDA model is been identified as : {'solver': 'svd', 'tol': 0.01}

On predicting the response with above model on training data gives the below confusion matrix and classification report:

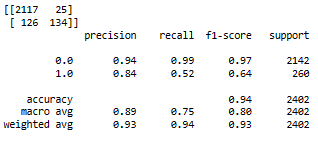


Fig 8. LDA Confusion Matrix and Classification Report for Model 1 Train data

Even though the accuracy is good at 0.94, recall is very low. We need to verify this model against the test dataset.

On running LDA on different parameters and after doing hyper-parameter tuning using grid search, the optimal LDA model is been identified as: {'solver': 'svd', 'tol': 0.01}

On predicting the response with above model on training data gives the below confusion matrix and classification report:

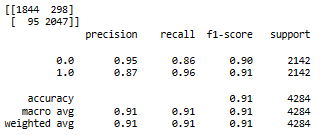


Fig 9. LDA SMOTE Confusion Matrix and Classification Report for Model 2 Train data

We can see that the overall accuracy is good at 0.91 and the recall for defaulters are at 0.96 which is better than model1. We can comment about the effectiveness of the model once its results are verified against the test dataset.

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

Using the 2 LDA models (Model 1 on train dataset and Model2 on the oversample train dataset using SMOTE) on the test dataset and verifying if the models gave good results

**Model 1:**

Test dataset confusion matrix and classification report values

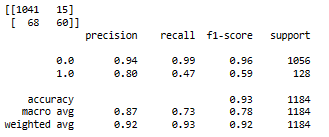


Fig 10. LDA Confusion Matrix and Classification Report for Model 1 Test data

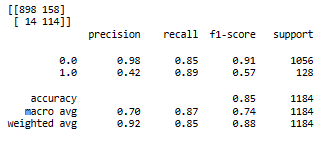
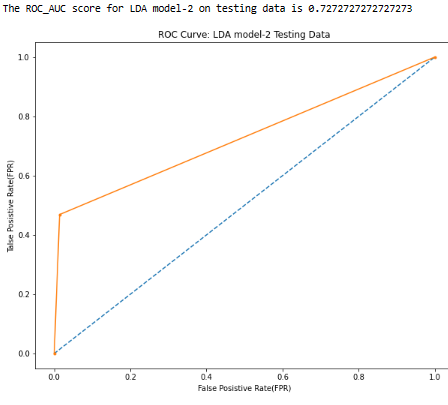


Fig 11. LDA SMOTE Confusion Matrix and Classification Report for Model 2 Test data



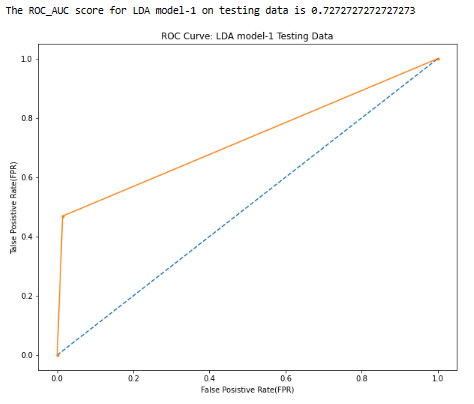


Fig 12. ROC\_AUC Curve for LDA Model 1 Fig 13. ROC\_AUC Curve for LDA Model 2

Comparing the 2 models we can see accuracy of model 1 is higher than model 2. But the ROC score for model 2 (0.87) is better than model 1 (0.72). Model 2 has a much better recall score of 0.89 than model1 at 0.47. Both the models are not either over or underfit. We will consider Model2 as the better one in this scenario.

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

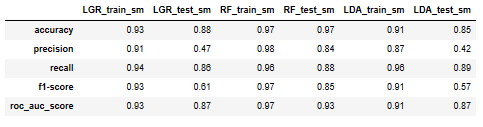


Fig 14. Comparing models

The above table shows the performance of the best model amongst logistic regression(LR), random forest(RF) and linear discriminant analysis(LDA) against the train and test data. We can see that in each of the models, SMOTE has corrected the imbalance of data and delivered better results. Between each models, we can see that accuracy between train/test is lesser than 0.1, which tells us that the models are neither overfit or underfit.

If we compare the performance of the model metrics against the test dataset, we can see that Random Forest has higher values than other models in accuracy, precision, recall (which is almost same for all models), f1-score and roc\_auc\_score. The ROC curve below shows that the second curve which is random forest is able to predict most of the data correctly.

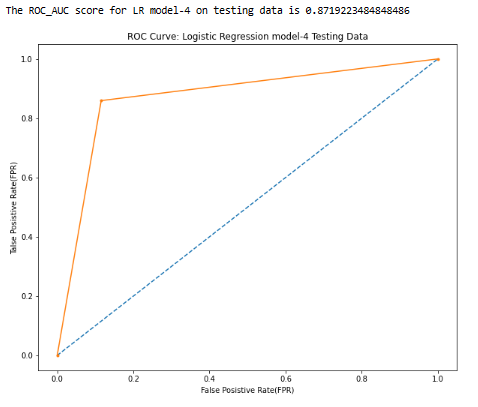
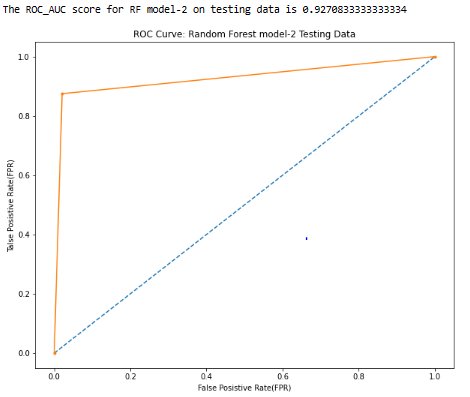
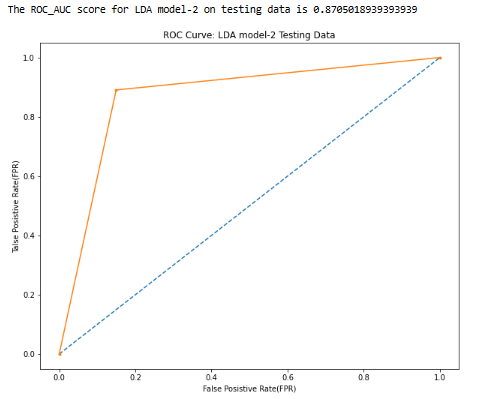


Fig 15. ROC Curve Comparison

**1.13 State Recommendations from the above models**

From the above models we can see that Random Forest with SMOTE oversampling on training dataset has yielded the best predictive results on test dataset.

From the statsmodel summary done as part of logistic regression, we can make the below inferences:

* Book\_value\_Adj\_Unit\_Curr has the highest negative coeffiecient which indicates that as the value of this predictor increases the chance for default decreases by 10% (i.e. 1-e^(-3.6150)).
* Total\_debt and Other\_Income has positive coefficient which tells us that higher these values the chances of payment default also increases.
* As the Current\_RatioLatest increases, which translates to company’s agility clearing of short term dues, it reduces the chance of defaulting.
* Groos\_Block has the highest positive coefficient and increase in this factor will lead to less chances of payment default.

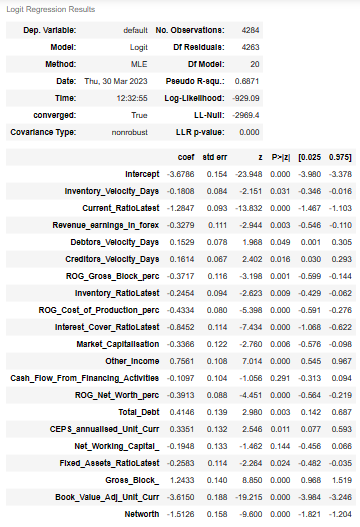


Fig 16. Statsmodel Summary

**2. Market Risk**

Initial validation of data

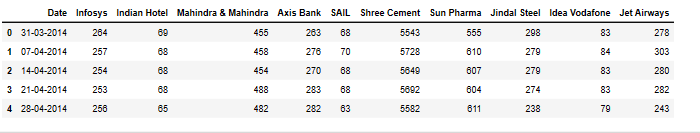
****

Fig 17. Data Head

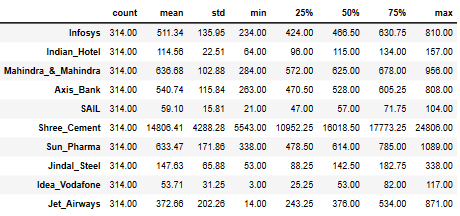


Fig 18. Descriptive stats of the dataset

The number of rows (observations) is 314

The number of columns (variables) is 11

We have one date variable(as object, will need to typecast as date variable), and stock price of 10 stocks. There are no missing values in the dataset.

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

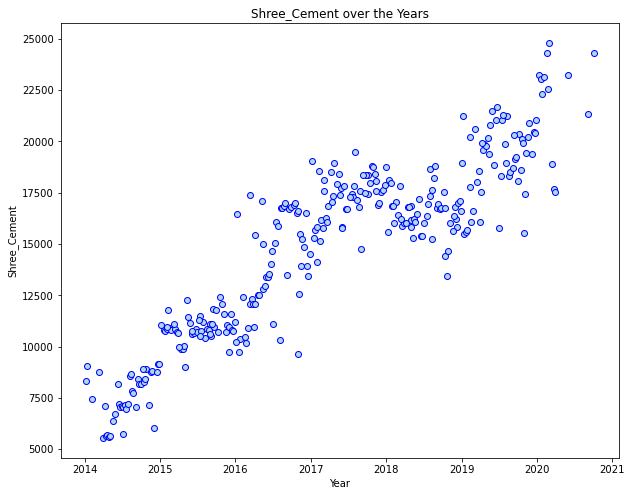


Fig 19. Shree\_Cement stock price graph

Figure 19 shows the price graph of Shree\_Cement from the year 2014 to 2021. Overall we can see a increasing trend in price from 5543 to 24806. In the periods, 2015-2016 and 2018-2019 the stock has maintained its rates or dipped slightly, but in all other time periods it has a very positive trend.

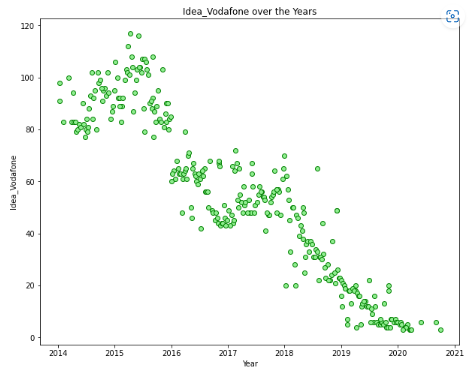


Fig 20. Idea\_Vodafone stock price graph

Figure 20 shows the price graph of Idea\_Vodafone from the year 2014 to 2021. Overall we can see a decreasing trend in price from 80 to 3. In the periods,2014 to mid 2015 the price has increased and reached max of 120. From then on till 2021 the price has decreased sharply, except for 2016 to 2017 when the price has not moved much.

**2.2 Calculate Returns for all stocks with inference**

We can calculate return by taking the log of values and calculating the difference of price value with price value of previous week. Lets look at the head and tail of this data.

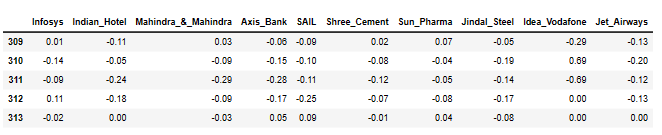


Fig 21. Data Tail

The returns are all calculated on the base of natural log. If we look at table 2.6, and consider rows 312 and 313, we can see Infosys price has reduced less than say Mahindra&Mahindra in the same week, on their respective rates.

Let us look at the central measures of tendency of the stock returns:

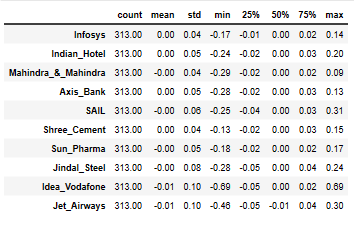


Fig 22. Stock return data description

If we look at the mean, it shows over the time period, on a weekly basis the price of Infosys is a positive value and hence the price is on an increasing trend, whereas the weekly mean is negative for Mahindra&Mahindra, which says that the price should be on a negative trend.

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

Calculating Mean and SD for all stocks

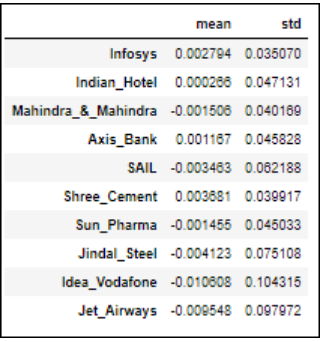
****

Fig 23. Stock return mean and SD

The mean shows the trend of overall price movement whereas SD indicates the volatility of the stock. If mean is positive the stock shows price appreciation and if mean is negative then price depreciation. Similarly when standard deviation is low, it shows that the price does not fluctuate drastically, whereas higher standard deviation indicates a stock which is very volatile and with high price fluctuations.

On comparing Infosys and Indian\_Hotel stock, we can see that mean of Indian\_Hotel is lesser than Infosys by an order of 10. This means that Infosys price has appreciated much more from its starting price than Indian\_Hotel has. If we look at the standard deviation Infosys has lesser than Indian\_Hotel which tells us that Infosys stock movement is much more stable than Indian\_Hotel. If one were to choose investment between both the stocks for a longer duration, Infosys is a good choice as it has higher price appreciation and is more stable.

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

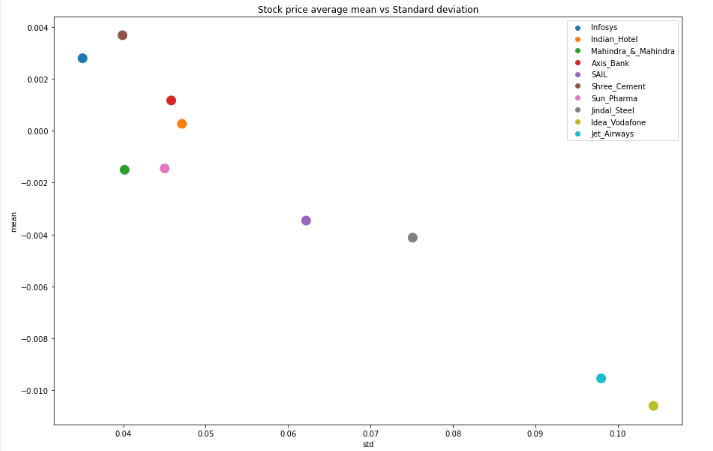


Fig 24. Plot of Mean vs SD

All stocks that are below y-axis value of 0, have their price depreciated, whereas all stocks above 0 have price appreciated.

All stocks with least value of x axis are better in terms of stability and price volatility is lesser.

If we look at price appreciation alone the top 3 stocks are Shree\_Cement, Infosys and Axis Bank whereas if we see the top 3 stocks in terms of price volatility, it is Infosys,Shree\_Cement and Mahindra&Mahindra.

Idea\_Vodafone has the lowest mean and highest standard deviation, which tells us the price for this stock has fallen the maximum and the stock is the most volatile, hence this stock should be least preferred of the lot.

**2.5 Conclusion and Recommendations**

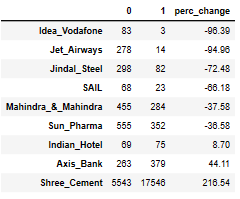


Fig 25. Stock Price Movement

* As we have surmised in section 2.4, Idea\_Vodafone stock has crashed, the price reducing by 96% in the period 2014 to 2021, whereas Shree\_Cement price has increased the most among the stock by 216% in the period 2014 to 2021.
* For long term investment in stocks, it is safer to invest in stock with lower volatility and higher average mean stock return. Shree\_Cement, Infosys and Axis Bank seems to be the pick of the lot from this perspective.
* Volatile stocks can also be invested with short term duration in mind, as the price varies a lot and if the trade is performed diligently, profits can be earned. Intraday trading and options are mostly traded with this fact in mind.
* The stock returns plot can be used by portfolio managers to get a better sense of the stocks in their portfolio and devise strategies accordingly.