FINANCE AND RISK ANALYSIS

Milestone-2

KIRAN.N - GREAT LEARNING

Table of Contents

Table of Figures 2

List Of Tables 3

Credit Risk Problem 4

Data Pre-Processing 4

Q 1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach 4

Q 1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model 5

Q 1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach 6

Q 1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model 6

Q 1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve) 7

Q 1.13 State Recommendations from the above models 11

Market Risk Problem 12

Q 2.1 Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference 12

Q 2.2 Calculate Returns for all stocks with inference 13

Q 2.3 Calculate Stock Means and Standard Deviation for all stocks with inference 14

Q 2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference 15

Q 2.5 Conclusion and Recommendations 15

Table of Figures

Figure 1: ROC Curve for LDA Training Set with 0.5 Threshold 8
Figure 2: ROC Curve for LDA Training Set with 0.07 Threshold 8
Figure 3: ROC Curve for LDA Testing Set with 0.5 Threshold 8
Figure 4: ROC Curve for LDA Testing Set with 0.07 Threshold 9
Figure 5:ROC Curve for RF Training Data 9
Figure 6: ROC Curve for RF Testing Data 9
Figure 7: LR Model ROC for Train data with 0.5 Threshold 10
Figure 8: LR Model ROC for Train data with 0.2 Threshold 10
Figure 9: LR Model ROC for Test data with 0.5 Threshold 10
Figure 10: LR Model ROC for Test data with 0.2 Threshold 11
Figure 11: Infosys Stock Price vs Time 12
Figure 12: Axis Bank Stock Price vs Time 13
Figure 13: Stock Means vs Stock Std Deviation 15

List Of Tables

Table 1 : Classification Report of Training Set 5
Table 2: Classification Report of Testing Set 5
Table 3: Classification Report of Training Set for 0.5 threshold 6
Table 4: Classification Report of Training Set for 0.07 threshold 6
Table 5: Classification Report of Testing Set with 0.5 threshold 7
Table 6: Classification Report of Testing Set with 0.07 threshold 7
Table 7: Comparison Of different Model Performances 11
Table 8: Sample Data Set 12
Table 9: Sample of Stock Returns 13
Table 10: Mean and Std Deviation of Stocks 14
Table 11: Stock Recommendations 15

Credit Risk Problem

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.

Data Pre-Processing

Before proceeding with Model Building activity, we do the following:

- Remove unwanted columns which do not have significant scope in model building. In this data set we have removed Co_Name and Co_Code which are mainly present for identification purpose.
- Check for null values and treat them if any. In this dataset we had 118 null values, we treated them using 'ffill' functionality.
- Check for outliers and treat them if any. On this case we had outliers in almost all numeric columns and we have treated them by replacing with boundary values.
- Split the data into 2 sets, Train set and Test set. We use train set to build the model and test set to verify the built model. Here we have split the given data in 67:33 ratio randomly. Post splitting, we have 2402 rows in train set and 1184 rows in test set.

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

Random Forest technique is an ensemble technique wherein we construct multiple models and take the average output of all the models to take final decision/make prediction. For constructing multiple models/decision trees using same dataset we go for boot strapped dataset

The Prediction strength of every individual tree must be high. The decision trees must not be correlated to each other.

To avoid this overfitting of each tree and to optimize the model building activity, we pass following parameters during model building:

- max_depth - indicates max levels up to which trees can extend
- min_samples_leaf - indicate minimum number of samples to be present on each leaf.
- min_samples_split -indicate minimum number of samples to be required for splitting current node.
- n_estimators-Number of Decision Trees to be constructed
- max_features-Number of columns randomly selected for decision making at each stage.

When we are not sure with the optimal value for each of the parameters, we can pass a list of values to each parameter. We create a dictionary with each parameter as a key and list of values to each key. Using the GridSearchCV functinality available in sklearn.model_selection module we pass the above dictionary and build the Random Forest model. The resultant model will select the optimal values for each parameter and build the model. The optimal values set for each parameter will be stored in grid_search.best_params_ variable.

```
{'max_depth': 3,
 'max_features': 25,
 'min_samples_leaf': 5,
 'min_samples_split': 15,
 'n_estimators': 10}
```

Following is the confusion matrix of training data.

```
array([[2157,    0],
       [   2,  243]], dtype=int64)
```

Following is the Classification report of training data.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2157
1	1.00	0.99	1.00	245
accuracy			1.00	2402
macro avg	1.00	1.00	1.00	2402
weighted avg	1.00	1.00	1.00	2402

Table 1 : Classification Report of Training Set

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

Upon validating the above built model against testing data, confusion matrix is as follows:

```
array([[1042,    0],
       [   0,  142]], dtype=int64)
```

Following is the Classification report of testing data.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1042
1	1.00	1.00	1.00	142
accuracy			1.00	1184
macro avg	1.00	1.00	1.00	1184
weighted avg	1.00	1.00	1.00	1184

Table 2: Classification Report of Testing Set

Interpretation from the model:

- By referring to confusion matrix and classification report of train data it seems the model is a over fitted model.
- But upon validating the model against the testing data the results are similar. We observe similar classification report as that of training data.
- We observe a clear imbalance in data where nearly 11% of given dataset has default values as '1' and remaining 89% as '0'. Due to this imbalance we might have got this over fitted model.

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

We build a Linear Discriminant Analysis model with a default threshold of 0.5. Following is the classification report for training data.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	2157
1	0.84	0.55	0.66	245
accuracy			0.94	2402
macro avg	0.90	0.77	0.82	2402
weighted avg	0.94	0.94	0.94	2402

Table 3: Classification Report of Training Set for 0.5 threshold

Using ROC, we found optimal threshold as 0. 0.071

	precision	recall	f1-score	support
0	0.988	0.929	0.957	2157
1	0.588	0.898	0.711	245
accuracy			0.925	2402
macro avg	0.788	0.913	0.834	2402
weighted avg	0.947	0.925	0.932	2402

Table 4: Classification Report of Training Set for 0.07 threshold

With the optimal Threshold value, we observe there is a better Recall value compared to previous one.

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

Upon validating the above built LDA model against Testing data we get following Classification report for default 0.5 threshold value.

	precision	recall	f1-score	support
0	0.94	0.99	0.96	1042
1	0.87	0.56	0.68	142
accuracy			0.94	1184
macro avg	0.91	0.77	0.82	1184
weighted avg	0.93	0.94	0.93	1184

Table 5: Classification Report of Testing Set with 0.5 threshold

Using ROC, we found optimal threshold as 0. 0.071

	precision	recall	f1-score	support
0	0.979	0.906	0.941	1042
1	0.555	0.859	0.674	142
accuracy			0.900	1184
macro avg	0.767	0.883	0.808	1184
weighted avg	0.928	0.900	0.909	1184

Table 6: Classification Report of Testing Set with 0.07 threshold

With the optimal Threshold value, we observe there is a better Recall value compared to previous one.

Interpretation from the model:

- The Accuracy of the LDA model (threshold = 0.07) with training data is 92.5% which is a good value also we find similar value of 90% for test data.
- We find Recall values of both test and train data to be more than 0.85.
- So overall, the built LDA model is a good model with high accuracy and there is no much deviation observed on test and train set.
- Model is neither underfit nor overfit model.

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

Linear Discriminant Analysis ROC Curves:

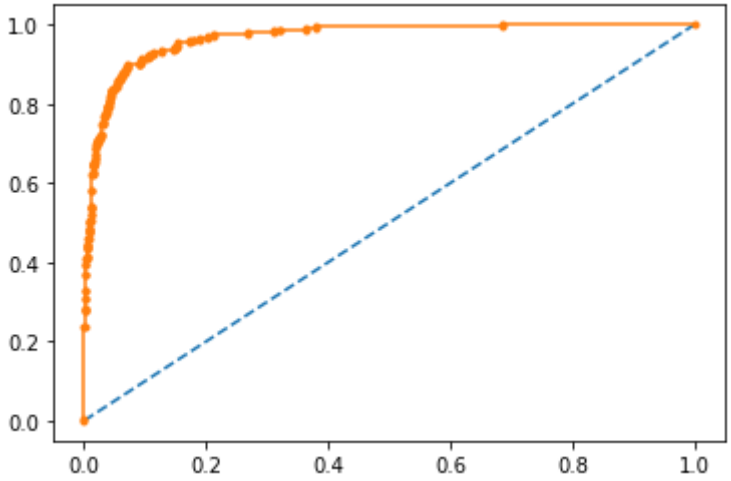

Figure 1: ROC Curve for LDA Training Set with 0.5 Threshold

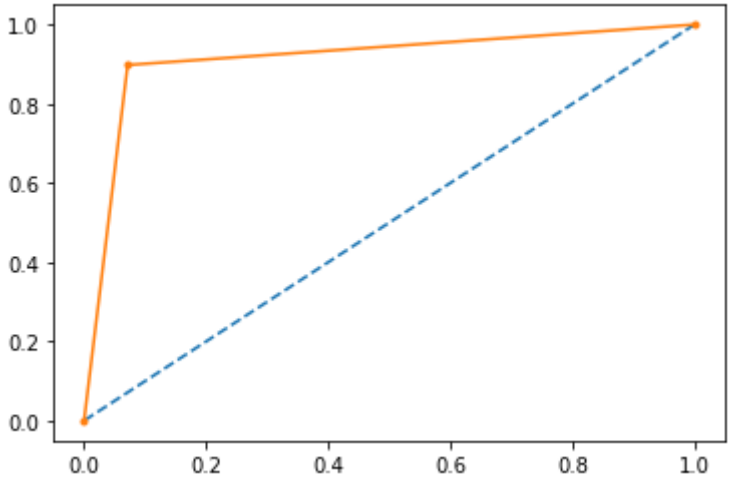

Figure 2: ROC Curve for LDA Training Set with 0.07 Threshold

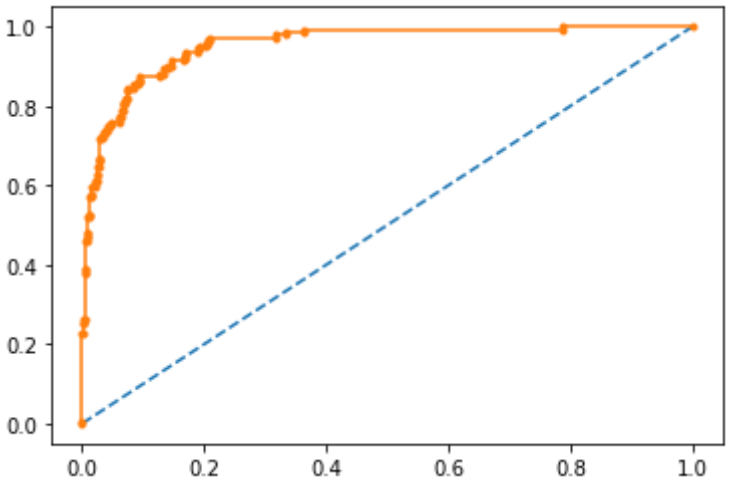

Figure 3: ROC Curve for LDA Testing Set with 0.5 Threshold

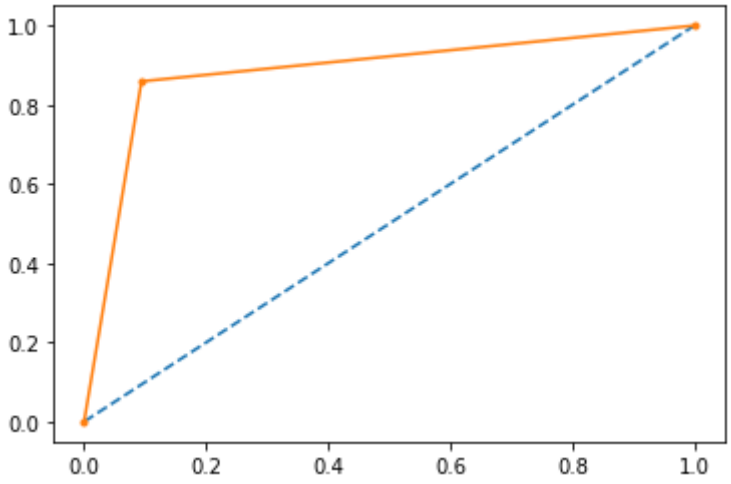

Figure 4: ROC Curve for LDA Testing Set with 0.07 Threshold

Random Forest Model ROC Curves:

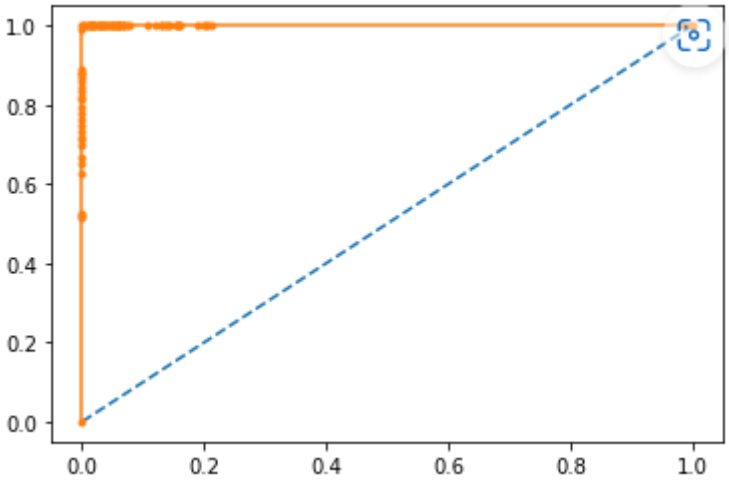

Figure 5:ROC Curve for RF Training Data

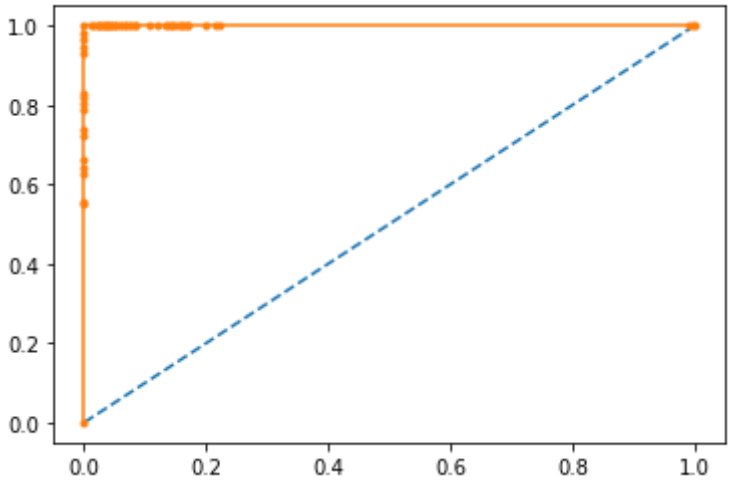

Figure 6: ROC Curve for RF Testing Data

Logistic Regression Analysis ROC Curves:

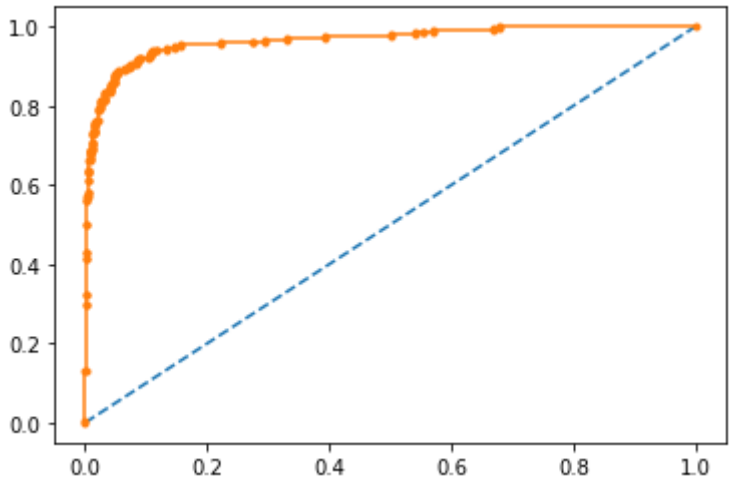

Figure 7: LR Model ROC for Train data with 0.5 Threshold

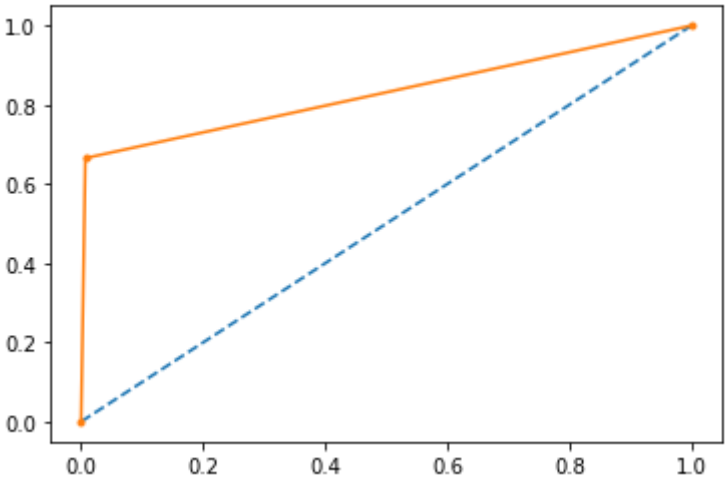

Figure 8: LR Model ROC for Train data with 0.2 Threshold

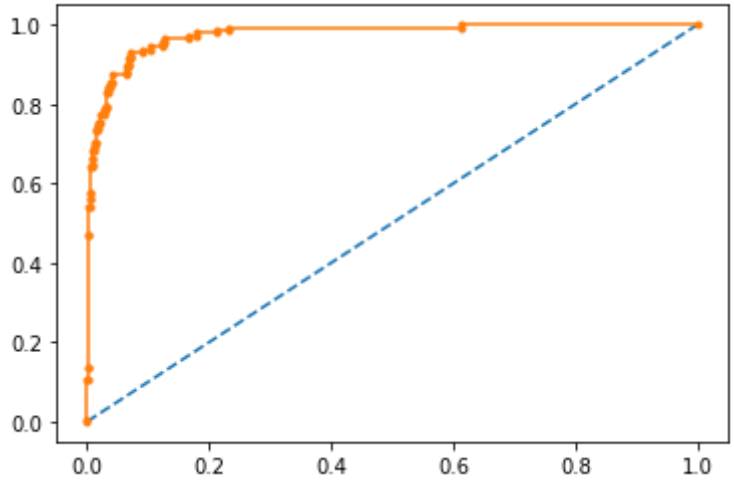

Figure 9: LR Model ROC for Test data with 0.5 Threshold

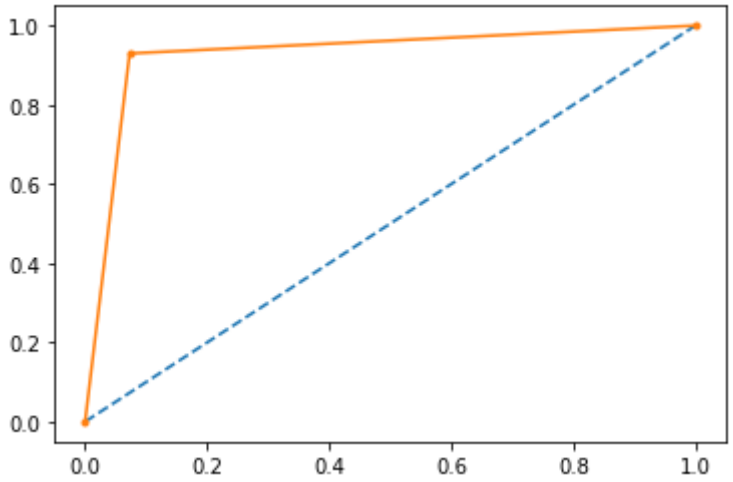

Figure 10: LR Model ROC for Test data with 0.2 Threshold

Models		AUC	Accuracy	Recall	F1-score
Logistic Regression - 0.5 Threshold	Train	0.966	0.96	0.66	0.77
	Test	0.976	0.95	0.73	0.79
Logistic Regression - 0.2 Threshold	Train	0.829	0.94	0.88	0.75
	Test	0.928	0.93	0.93	0.75
Linear Discriminant analysis - 0.5 threshold	Train	0.968	0.94	0.55	0.66
	Test	0.955	0.94	0.56	0.68
Linear Discriminant analysis - 0.07 threshold	Train	0.913	0.93	0.89	0.71
	Test	0.883	0.9	0.86	0.67
Random Forest	Train	1	1	0.99	1
	Test	1	1	1	1

Table 7: Comparison Of different Model Performances

Comparing to all models Logistic Regression model with 0.2 looks better compared to other model. Ignoring Random Forest model with highest Accuracy, F1-score and Recall value equal to 1 since it doesn't seem to be same in all practical cases.

Q 1.13 State Recommendations from the above models

Referring to Table-7

- Random Forest model has the highest Accuracy, F1-score and Recall equal to 1. But this is a ideal scenario which can't be true in all scenarios.
- Logistic Regression model with 0.2 threshold has better Accuracy, F1-score and Recall values for both test and train data.
- So Logistic Regression model with 0.2 threshold value is the preferred model out of all models built.

Market Risk Problem

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.

There are no null values present in given dataset.

	Date	Infosys	Indian_Hotel	Mahindra_&_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways
0	31-03-2014	264	69	455	263	68	5543	555	298	83	278
1	07-04-2014	257	68	458	276	70	5728	610	279	84	303
2	14-04-2014	254	68	454	270	68	5649	607	279	83	280
3	21-04-2014	253	68	488	283	68	5692	604	274	83	282
4	28-04-2014	256	65	482	282	63	5582	611	238	79	243

Table 8: Sample Data Set

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

1) Infosys

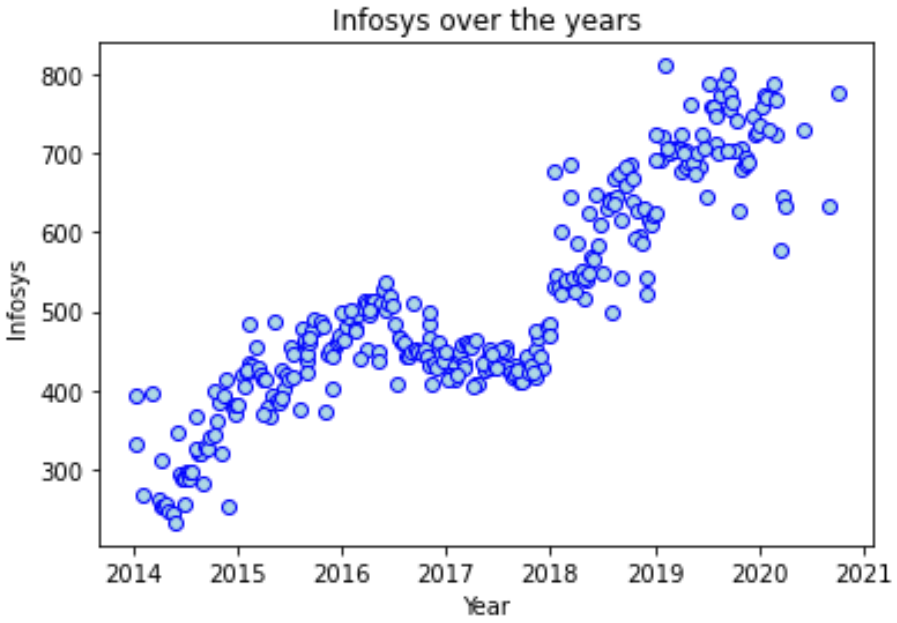

Figure 11: Infosys Stock Price vs Time

2) Axis Bank

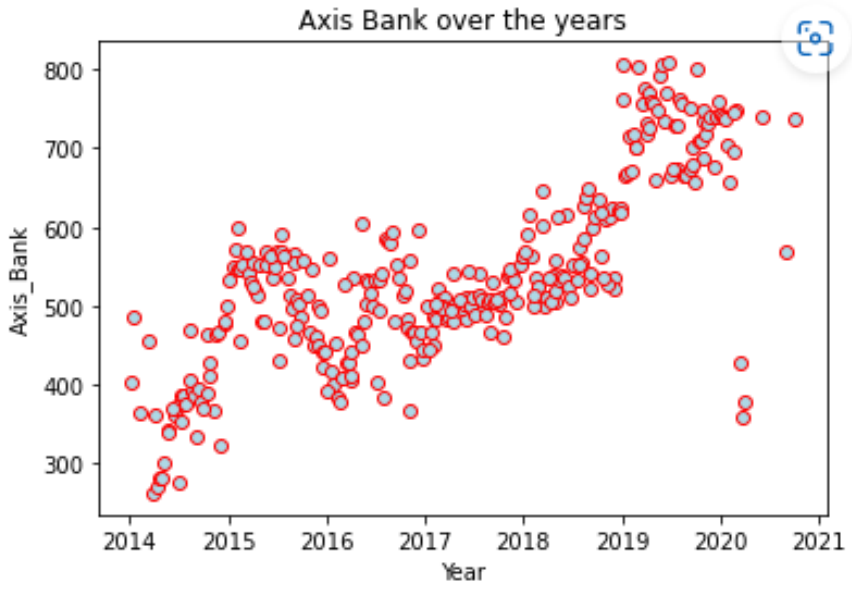

Figure 12: Axis Bank Stock Price vs Time

Inferences from graphs:

- Infosys stock price has increased over a course of time. It has moved in uptrend with a minor dip in between. Considering overall timeline 2014 -2021 stock price which was below 300 has moved to 800+
- Axis Bank stock price has increased over a course of time. Though there were ups and downs in the price, in overall timeline 2014 -2021 stock price which was below 300 made a high of about 800+ and settled around 700+.

Q 2.2 Calculate Returns for all stocks with inference

Stock return for an interval is defined as the difference between logarithmic value of their prices at the start and end of the interval.

	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	-0.026873	-0.014599	0.006572	0.048247	0.028988	0.032831	0.094491	-0.065882	0.011976	0.086112
2	-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011976	-0.078943
3	-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000000	0.007117
4	0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049393	-0.148846

Table 9: Sample of Stock Returns

Summing up all the returns we get:

```
Infosys                 0.874521
Indian_Hotel            0.083382
Mahindra_&_Mahindra    -0.471323
Axis_Bank               0.365382
SAIL                   -1.084013
Shree_Cement            1.152290
Sun_Pharma             -0.455337
Jindal_Steel           -1.290374
Idea_Vodafone          -3.320228
Jet_Airways            -2.988564
```

- From the above data we can infer Infosys, Indian Hotel, Axis Bank, Shree Cement has given Positive returns while other stock have given negative returns which means the investment on other stocks have reduced their investment value.
- Shree Cement has given the highest returns.
- Idea Vodafone has given the worst returns.

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

- Stock Means: Average returns that the stock is making on a month-on-month 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 is.

	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

Table 10: Mean and Std Deviation of Stocks

- Infosys, Indian Hotel, Axis Ban Shree Cement stocks have positive mean values which means they have given positive returns.
- While other stocks have negative mean values which means they have given negative returns.
- Infosys has least Std deviation of 0.035 which means it risk involved in investing on this stock is less compared to other stocks.

- Idea Vodafone has highest Std deviation of 0.104 which means it has highest risk in investment.
- Shree Cement has highest mean 0.003 which means it has given good returns.
- Idea Vodafone has given the worst returns with least mean of -0.01

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

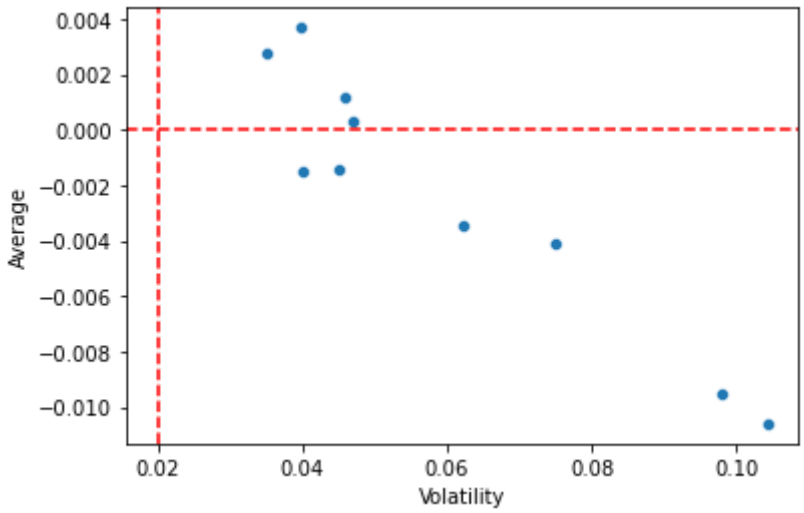

Figure 13: Stock Means vs Stock Std Deviation

- Stocks with a lower mean & higher standard deviation is not preferred.
- Stocks with higher mean & lower standard deviation is preferred.

Q 2.5 Conclusion and Recommendations

Following is the stock which are recommended. These stocks have positive means. Also these stocks are arranged in the increasing order of their standard deviation which means in the increasing order of the risks.

	Average	Volatility
Infosys	0.002794	0.035070
Shree_Cement	0.003681	0.039917
Axis_Bank	0.001167	0.045828
Indian_Hotel	0.000266	0.047131

Table 11: Stock Recommendations

Infosys is the best stock with lower risk and better returns.