FRA Project Milestone 2

Ashwani Balyan
Batch- G4
balyanashwani@gmail.com

Table of Contents:

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 labeled field.

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

- We have build the model without using any parameters.
- This model will be validated on test dataset and the parameters will be updated using gridsearch, if required.

Classification Report for train data

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2157
1	1.00	1.00	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 1.8.1

Classification Report for train data

	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

Table 1.8.2

- This model seems to be overfitting so we will use grid search and use the parameters to build a new model.

Using grid search

- Best parameters are: max_depth=7, min_samples_leaf=3, min_samples_split=10,n_estimators=15

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

Performance Matrices - Random Forest Model after using grid search

Classification report for Train Data

	precision	recall	f1-score	support
0	0.99	0.99	0.99	2157
1	0.95	0.89	0.92	245
accuracy			0.98	2402
macro avg	0.97	0.94	0.96	2402
weighted avg	0.98	0.98	0.98	2402

Table 1.9.1

Confusion matrix for Train Data

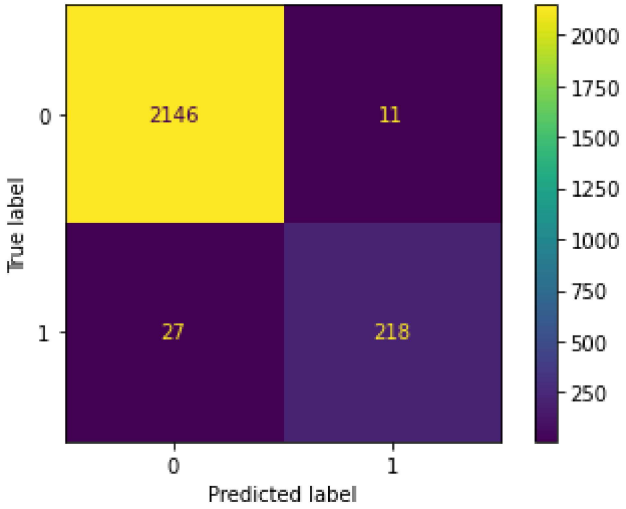

Fig 1.9.1

Classification report for Test Data

	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	0.95	0.95	1184
weighted avg	0.98	0.98	0.98	1184

Table 1.9.2

Confusion matrix for Test Data

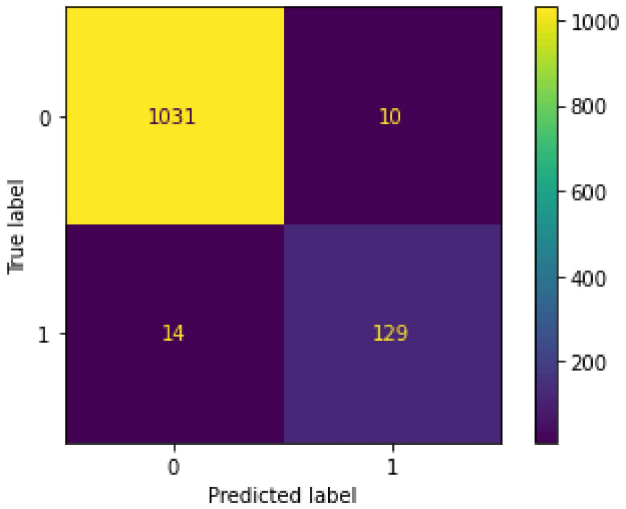

Fig 1.9.2

Interpretation from the model;

- The model built without using grid search is overfitting.
- Model build after putting parameters using grid search is giving decent results.
- Recall and precision values are good along with accuracy.
- The model is able to predict 129 true defaulters correctly and 14 defaulters are predicted wrongly as non-defaulters.

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

- We have built the model with default threshold value.
- The model will be validated on test data and the threshold value will be updated if required.

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

Performance Matrices - LDA Model

Classification report for Train Data

	precision	recall	f1-score	support
0	0.94	0.99	0.96	2157
1	0.81	0.42	0.55	245
accuracy			0.93	2402
macro avg	0.87	0.70	0.76	2402
weighted avg	0.92	0.93	0.92	2402

Table 1.11.1

Confusion matrix for Train Data

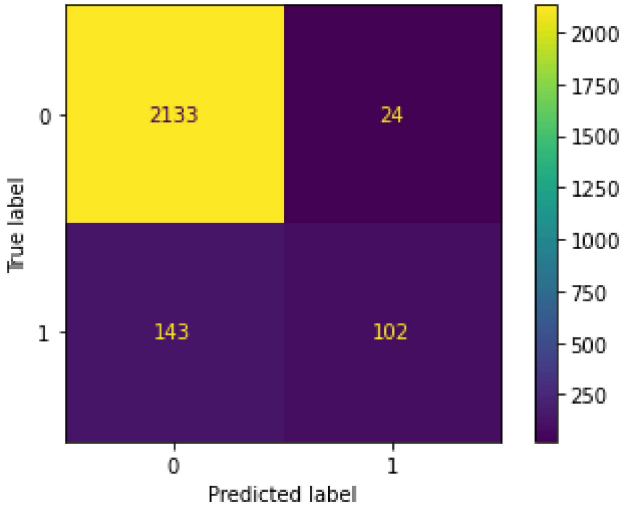

Fig 1.11.1

Classification report for Test Data

	precision	recall	f1-score	support
0	0.92	0.99	0.95	1041
1	0.78	0.38	0.51	143
accuracy			0.91	1184
macro avg	0.85	0.68	0.73	1184
weighted avg	0.90	0.91	0.90	1184

1.11.2

Confusion matrix for Test Data

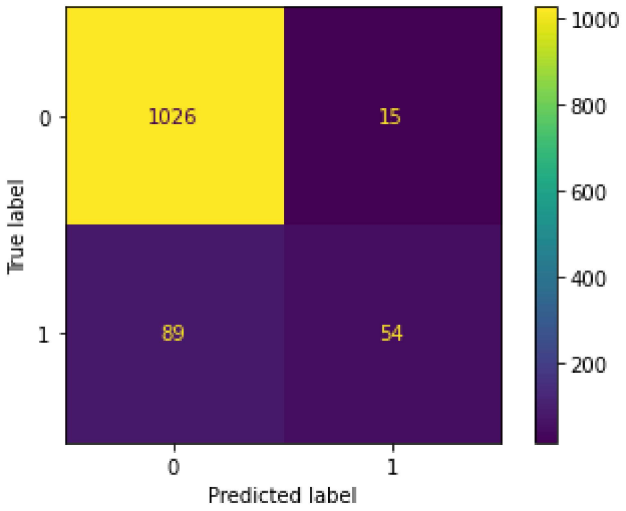

Fig 1.11.2

- The recall values are too low so we will build another model by updating threshold value to optimum.

Choosing the optimal threshold

Performance Matrices - LDA Model with optimum threshold of 0.046

Classification report for Train Data

	precision	recall	f1-score	support
0	0.99	0.88	0.93	2157
1	0.46	0.89	0.60	245
accuracy			0.88	2402
macro avg	0.72	0.89	0.77	2402
weighted avg	0.93	0.88	0.90	2402

Table 1.11.3

Confusion matrix for Train Data

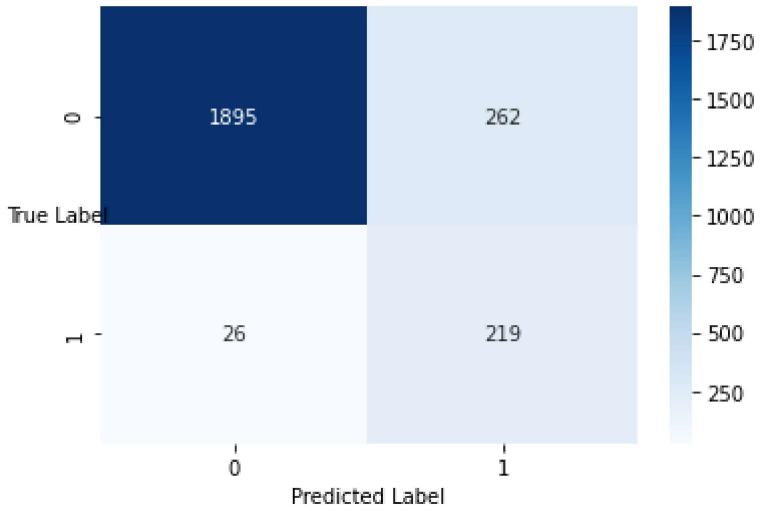

Fig 1.11.3

Classification report for Test Data

	precision	recall	f1-score	support
0	0.99	0.82	0.90	1041
1	0.41	0.91	0.57	143
accuracy			0.83	1184
macro avg	0.70	0.87	0.73	1184
weighted avg	0.92	0.83	0.86	1184

Table 1.11.4

Confusion matrix for Test Data

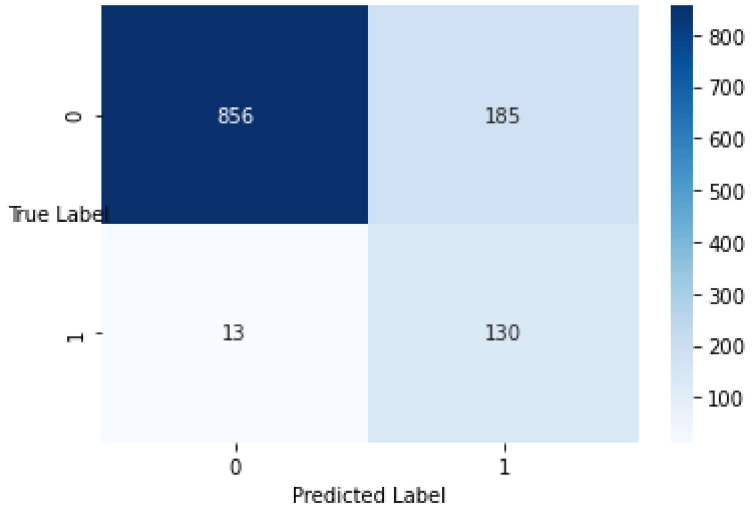

Fig 1.11.4

Interpretation from the model;

- The initial model gave poor results on Recall value.
- The model built using optimized threshold value gave better results.
- Though the Recall value increased but the precision decreased a lot.

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

AUC and ROC for Logistics Model

Train Data

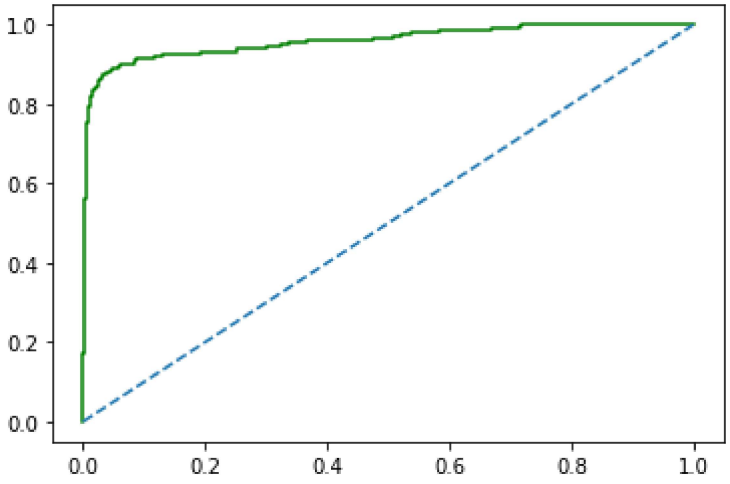

Fig 1.12.1

Area under Curve is 0.9596226807830225

Test Data

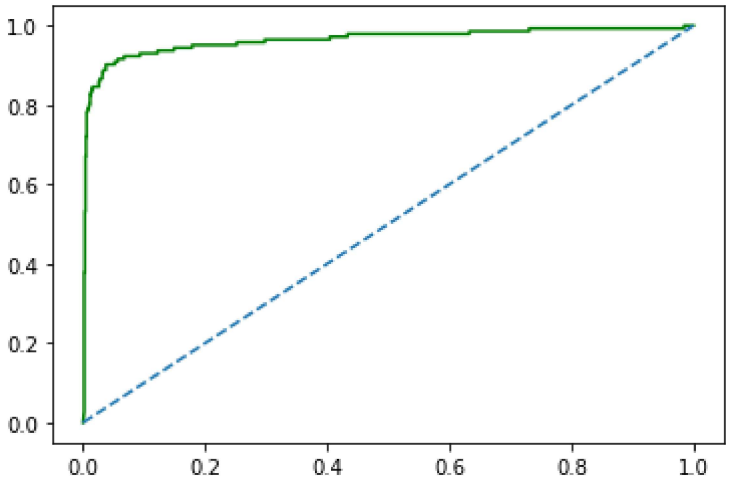

Fig 1.12.2

Area under Curve is 0.9640877854134338

AUC and ROC for Random Forest Model

Train Data

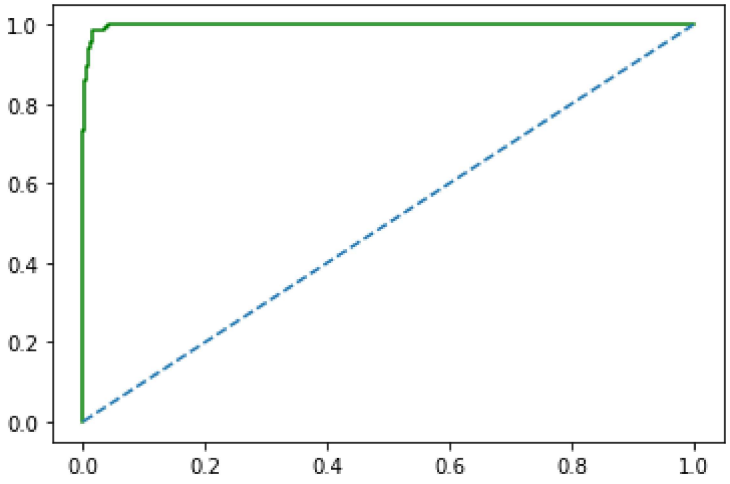

Fig 1.12.3

Area under Curve is 0.9981569261918953

Test Data

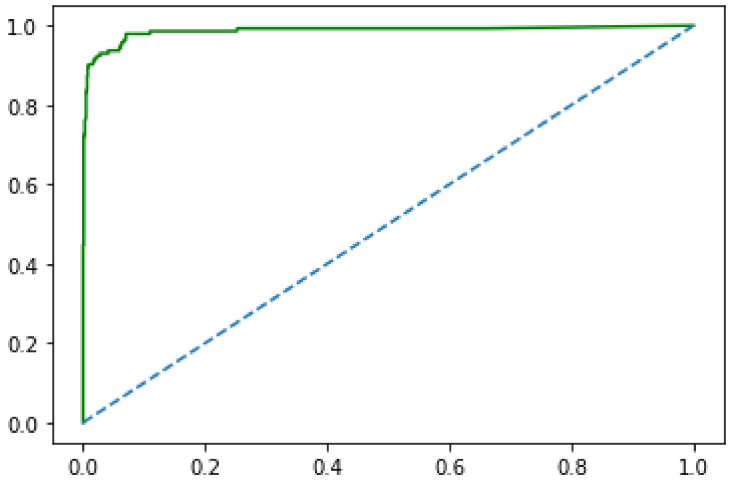

Fig 1.12.4

Area under Curve is 0.9868402490880875

AUC and ROC for LDA Model

Train Data

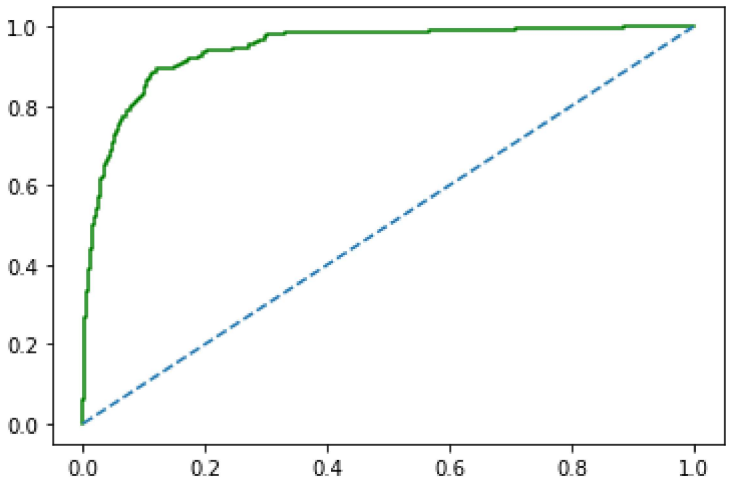

Fig 1.12.5

Area under Curve is 0.944062520696735

Test Data

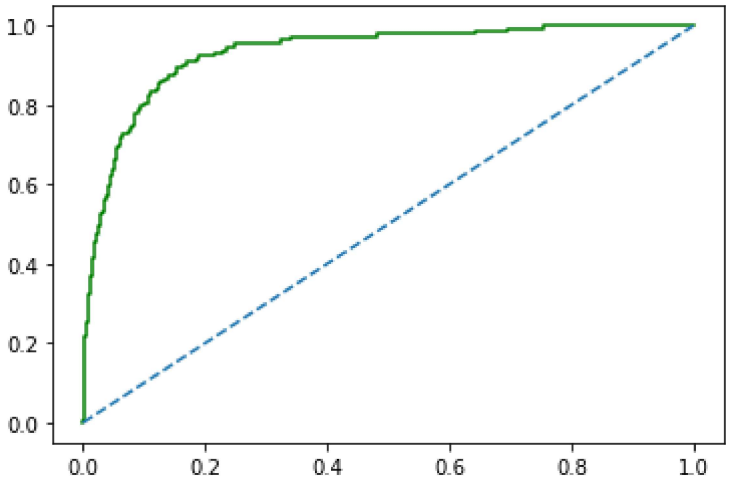

Fig 1.12.6

Area under Curve is 0.9313261186460035

Head to Head Model Comparision

	Logit_model_train	Logit_model_test	RF_model_train	RF_model_test	LDA_model_train	LDA_model_test
Precision	0.67	0.70	0.95	0.92	0.46	0.41
Recall	0.89	0.91	0.89	0.90	0.89	0.91
Accuracy	0.95	0.94	0.98	0.98	0.88	0.83
AUC	0.96	0.96	1.00	0.99	0.94	0.93

Table 1.12.1

From above table it is clear that ;

- Best Precision for class '1' (i.e for default) is given by Random Forest model
- Best Recall for class '1' (i.e for default) is given by Logistic Regression model and LDA model
- Best accuracy is given by Random Forest model
- Best AUC is given by Random Forest model

Since, the Recall is more important measure for our model (predicting the defaulters), so Logistic Regression model is suggested for this problem, as it has higher precision and accuracy than LDA model

1.13 State Recommendations from the above models

Top 10 important features of Random Forest Model

	Imp
Book_Value_Unit_Curr	0.353573
Networth	0.320431
Book_Value_Adj_Unit_Curr	0.073209
ROG_Net_Worth_perc	0.026742
PBDT	0.023714
PBDTM_perc_Latest	0.022299
CP	0.021406
PBITM_perc_Latest	0.018135
Capital_Employed	0.017135
Total_Asset_Turnover_Ratio_Latest	0.014666

Table 1.13.1

Top important features of Logistic Regression Model

- Book_Value_Adj_Unit_Curr
- Current_Ratio_Latest
- Debtors_Ratio_Latest
- Interest_Cover_Ratio_Latest
- PBDTM_perc_Latest

Recommendations:

- Both the models i.e Random forest and Logistic regression has the feature "Book_Value_Adj_Unit_Curr" as topmost important feature. Thus, the investors may consider this feature as a very important one to analyse the credit risk.
- The second feature which appears as important in both models is "PBDTM_perc_Latest".
- The other important features may be considered are; "Networth", "Capital_Employed","Current_Ratio_Latest".
- Investors are recommended to use these features for predicting the might be "Defaulters" before investing money in companies.

Market Risk

Problem Statement

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.

- The datatype of column 'date' is object type, which is required to be changed to 'Datetime' datatype

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

Stock Price Graph for Infosys

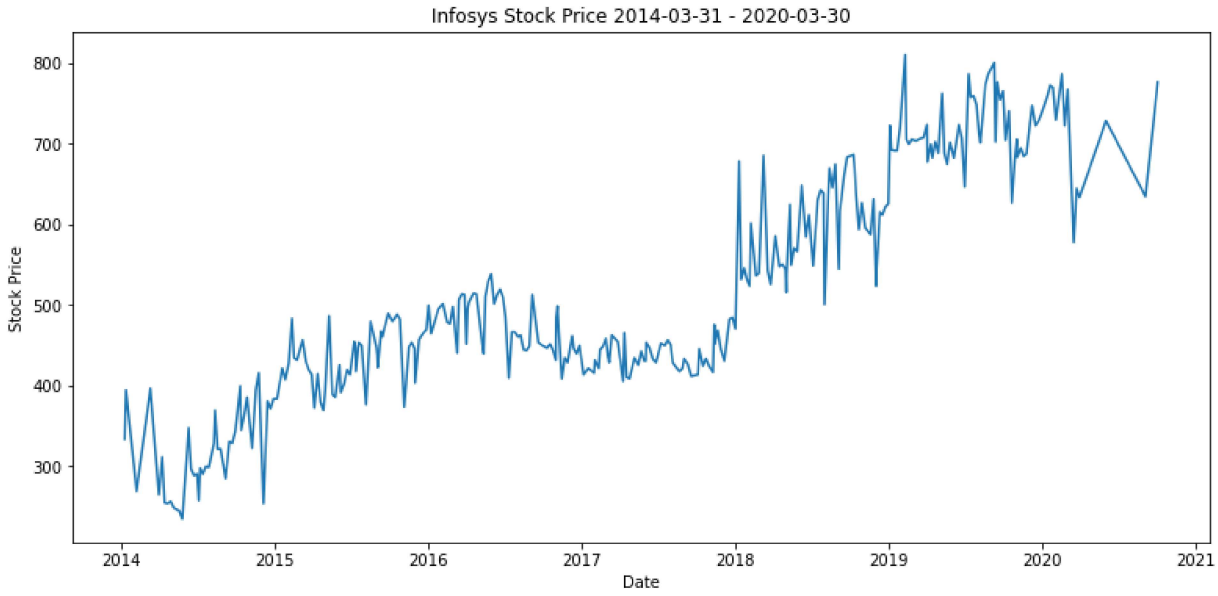

Fig 2.1.1

Stock Price Graph for SAIL

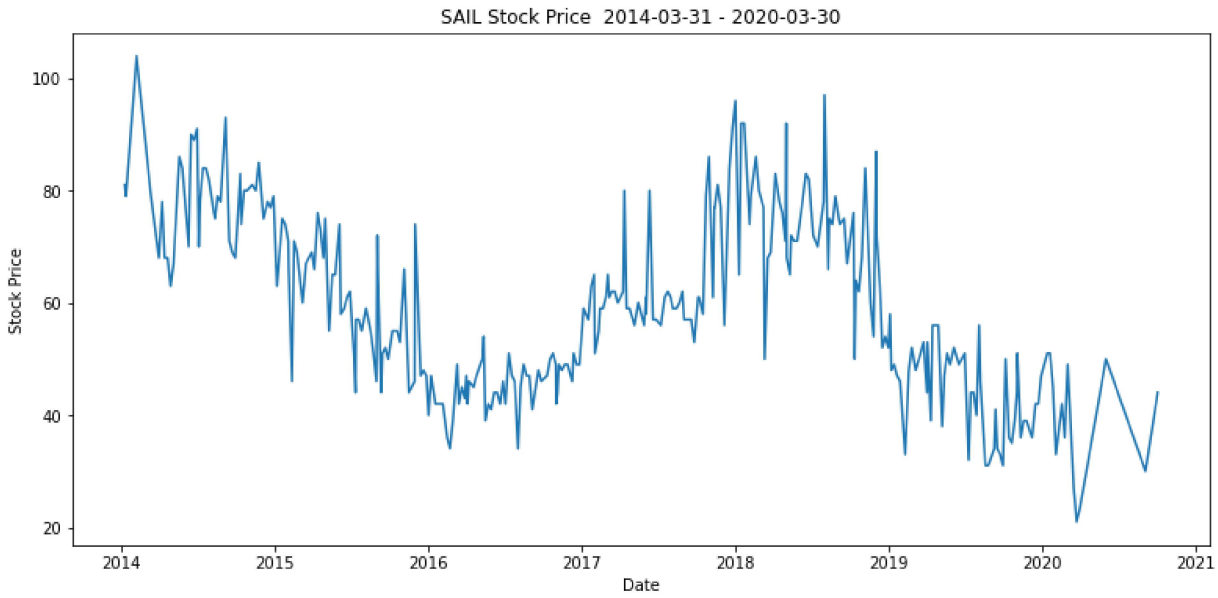

Fig 2.1.2

Infrences;

- Stock price graph of Infosys;
  - Stock price graph of Infosys shows an overall increase in stock prices with time from 2014-03-31 to 2020-03-30.
  - There is an increase in stock prices till mid 2016 and the stock prices decreased till 2017 year end.
  - The stock prices are generally increasing after 2018 year start.
- Stock price graph of SAIL;
  - Stock price graph of SAIL shows an overall decrease in stock prices with time from 2014-03-31 to 2020-03-30.
  - There is steep decrease in stock prices till first quarter of 2016 and then prices increased till last quarter of 2018.
  - The stock prices are generally decreasing after 2018 last quarter of year start.

2.2 Calculate Returns for all stocks with inference

Top 5 rows of calculated returns for all stocks

	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 2.2.1

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

Stock means

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 2.3.1

Stock standard deviation

Infosys	0.035070
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

Table 2.3.2

Inferences;

- The lowest Stock Means is of Idea_Vodafone with value -0.010608, i.e the lowest mean returns

- The highest Stock Means is of Shree_Cement with value 0.003681, i.e the highest mean returns
- The lowest Stock standard deviation is of Idea_Vodafone with value 0.104315, i.e the lowest volatility
- The highest Stock standard deviation is of Infosys with value 0.035070, i.e the highest mean returns

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

Plot of Stock Means(Average) vs Standard Deviation(Volatility)

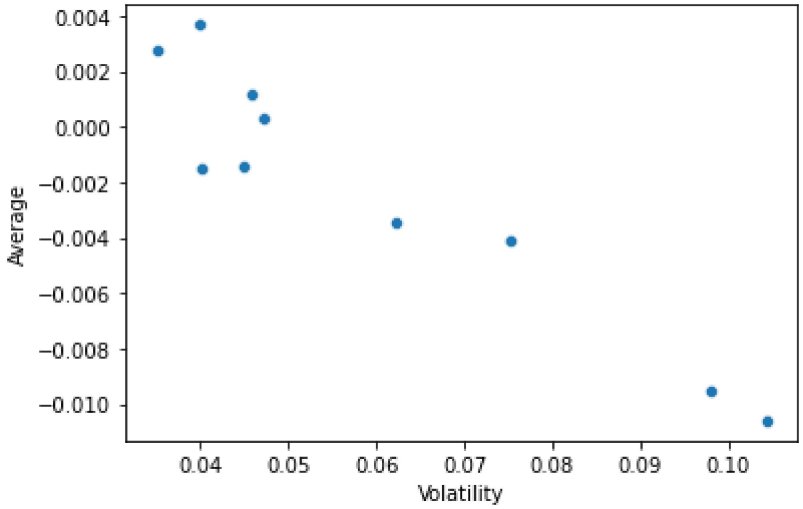

Fig 2.4.1

Inferences;

- There seems to be just four companies with positive mean returns (Stock Means).
- The Standard Deviation for these companies is also comaparatibily low.
- It seems that the volatility is inversly proportional to Average return values.

2.5 Conclusion and Recommendations

Stocks sorted by Volatility

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

Table 2.5.1

Stocks sorted by Average return

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

Table 2.5.2

Getting the stocks with relatively higher mean and low standard deviation(Volatility)

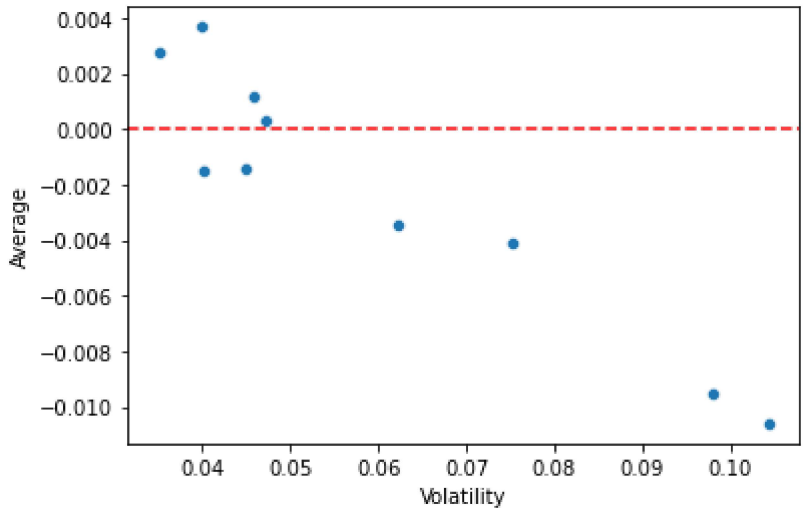

Fig 2.5.1

Stocks with positive Average value, sorted by volatility value

	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 2.5.3

Conclusion and Recommendations;

- From above graph and tables;
  - The lowest volatility is for Infosys and highest for Idea_Vodafone.
  - The lowest average return is for Idea_Vodafone and highest for Shree_Cement.
  - There are only four stocks with non negative average value of stocks and less volatility, namely,Infosys, Shree_Cement, Axis_Bank and Indian_Hotel.
  - There seems to be a relation between the returns and risk. The low returns have high risk values and vice-versa.
- Hence, from above analysis; the worst stock to be invested seems like that of Idea_Vodafone and the best seems to be that of Infosys. If somewhat higher risk is acceptable, than the stocks of Indian_Hotel will provide better returns.

Thanks.....