Jan 2021 | PG-DSBA-Online

FINANCE AND RISK ANALYTICS PROJECT REPORT

SUBMITTED BY
DEV TRIPATHI

Contents

Problem 1 1

Credit Default Dataset Analysis *(Cont.)* 1

Information about dataset 1

Data Dictionary for Credit Default Dataset 1

Random Forest Model 1

Linear Discriminant Analysis 4

Performance Comparison 6

Conclusion and Recommendations 7

Problem 2 8

Stock Price Dataset Analysis 8

Information about the dataset 8

Sample of the dataset 8

Stock Prices over time 8

Returns Calculation 9

Stock Means vs Standard Deviation 11

Conclusions and Recommendations 13

List of Figures

Figure 1: Performance metrics for Random Forest Classifier train set (top) and test set (bottom) 2
Figure 2: Performance metrics for tuned Random Forest model on train set (Left) and test set (Right)......3
Figure 3: Performance metrics for tuned RFCL model on balanced data 4
Figure 4: Feature importance and their corresponding coefficients for LDA model 5
Figure 5: Performance Metrics for LDA model after using RFE and balanced set (train set is on the left and test set performance is on the right) 5
Figure 6: ROC Curves and AUC score for train (left) and test (right) set 6
Figure 7: Sample of Stock Price dataset...8
Figure 8: Stock Prices of Infosys 8
Figure 9: Stock Prices of Idea Vodafone 9
Figure 10: Returns (%) calculated for Stock Prices 10
Figure 11: Statistical description of the Returns (%)..11
Figure 12: Stock Price Means and standard Deviation...11
Figure 13: Stock Means vs Standard Deviation on Normal scale...12
Figure 14: Stock Means vs Standard Deviation on Log scale 12

Problem 1

Credit Default Dataset Analysis *(Cont.)*

Information about dataset

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 labeled field.

Data Dictionary for Credit Default Dataset

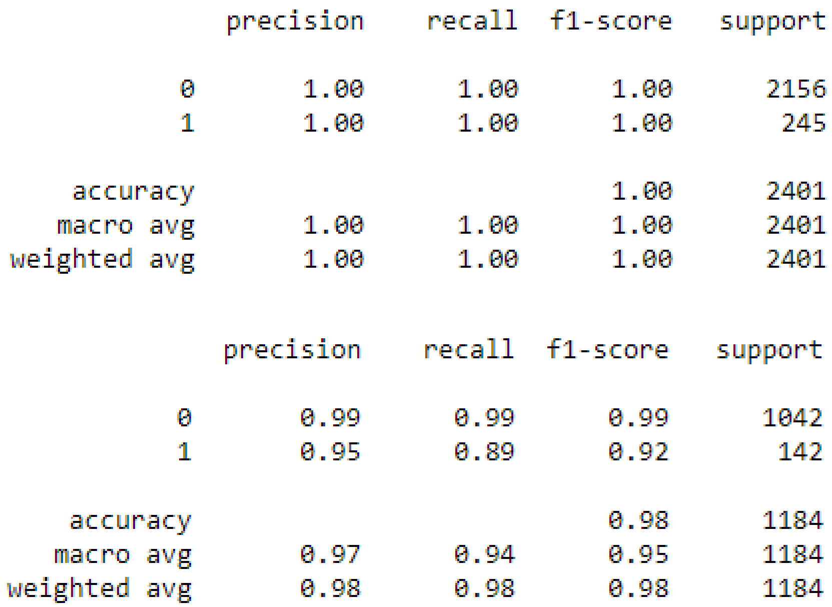
The dataset contains total 67 features.

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

Random Forest Model

Now, continuing the modelling part for the credit risk dataset, we can go for ensemble methods such as random forest. Random forest classifier works well in general compared to one single decision tree as single decision trees overfit on train set if not pruned properly.

For training random forest model, scikit learn library was used. Firstly, we tried a random forest classifier without tuning of hyperparameters. The results obtained are given below:



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2156
           1       1.00      1.00      1.00       245

    accuracy                           1.00      2401
   macro avg       1.00      1.00      1.00      2401
weighted avg       1.00      1.00      1.00      2401

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1042
           1       0.95      0.89      0.92       142

    accuracy                           0.98      1184
   macro avg       0.97      0.94      0.95      1184
weighted avg       0.98      0.98      0.98      1184
```

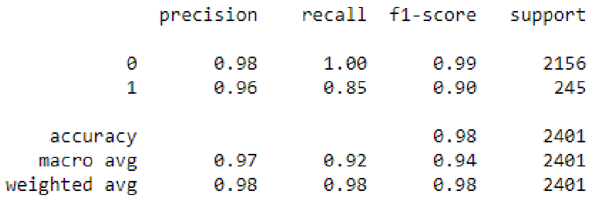
Figure 1: Performance metrics for Random Forest Classifier train set (top) and test set (bottom)

As we can observe from the forementioned metrics, the Random Forest classifier is overfitting on train data set and performing poor on test set (the recall value for train set is 100% while on test set it is 89% only for class 1 which is our minority class i.e., Defaulters).

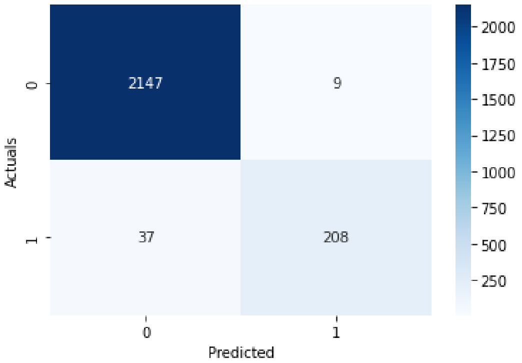
Further, we have tried to tune the model using Grid Search method provided in scikit learn package. Grid Search is basically a Brute force method. It simply trains the model for the different combinations in a given set of parameters. The best parameters obtained after using Grid search are mentioned below:

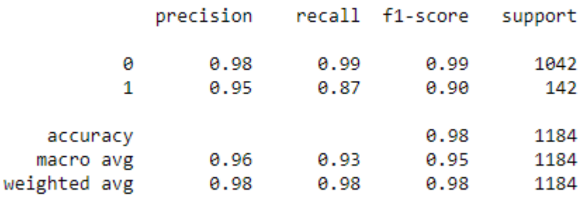
```
'max_depth': 7,
'max_features': 5,
'min_samples_leaf': 15,
'min_samples_split': 2,
'n_estimators': 300
```

The performance metrics obtained after using this set of hyperparameters is mentioned below:



```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      2156
           1       0.96      0.85      0.90       245

    accuracy                           0.98      2401
   macro avg       0.97      0.92      0.94      2401
weighted avg       0.98      0.98      0.98      2401
```





```
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      1042
           1       0.95      0.87      0.90       142

    accuracy                           0.98      1184
   macro avg       0.96      0.93      0.95      1184
weighted avg       0.98      0.98      0.98      1184
```

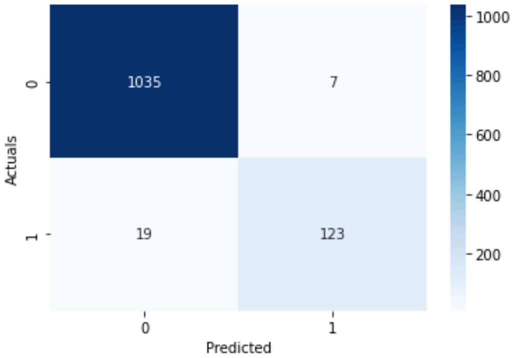
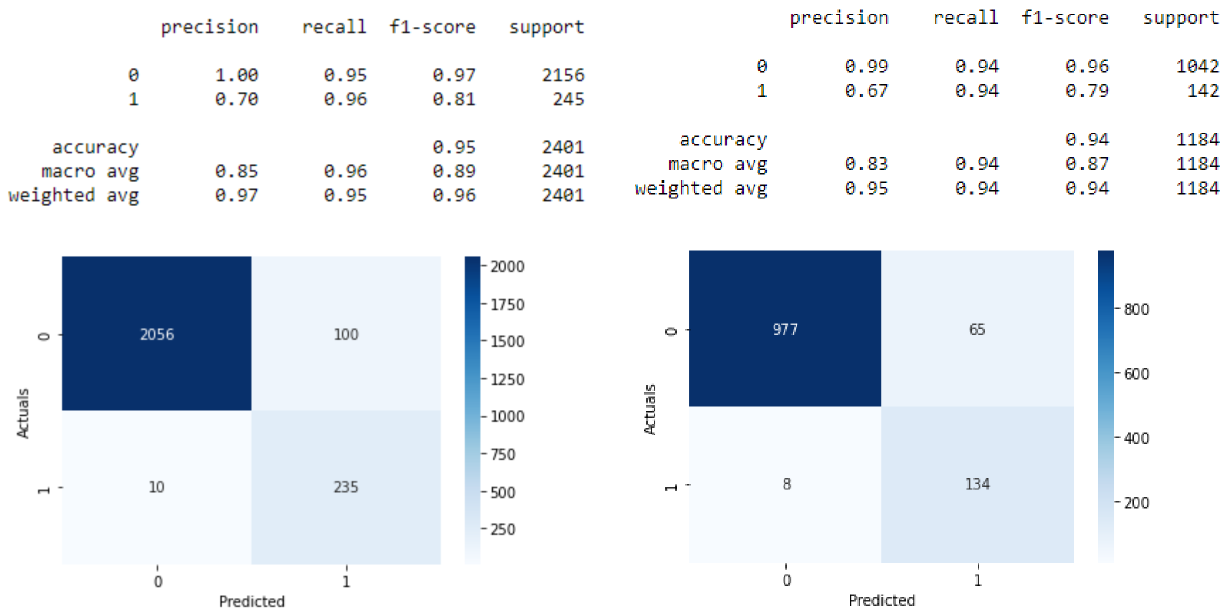


Figure 2: Performance metrics for tuned Random Forest model on train set (Left) and test set (Right)

As we can observe, the performance is comparable for both train and test set. Although, the model performance has slightly decreased if we compare it with the earlier trained Random Forest Classifier results.

Now, we can also try to enhance the performance further by using the balanced dataset which was generated earlier by using SMOTE. The Grid Search method was also used for training this Random Forest model also. The performance metrics obtained for this model are given below:



```
              precision    recall  f1-score   support

           0       1.00      0.95      0.97      2156
           1       0.70      0.96      0.81       245

    accuracy                           0.95      2401
   macro avg       0.85      0.96      0.89      2401
weighted avg       0.97      0.95      0.96      2401
```

```
              precision    recall  f1-score   support

           0       0.99      0.94      0.96      1042
           1       0.67      0.94      0.79       142

    accuracy                           0.94      1184
   macro avg       0.83      0.94      0.87      1184
weighted avg       0.95      0.94      0.94      1184
```

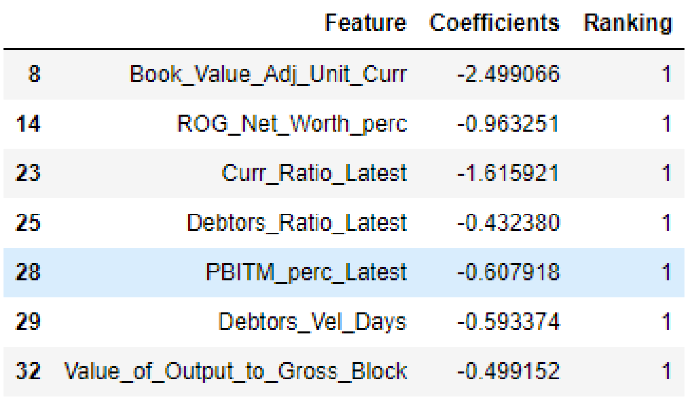
Figure 3: Performance metrics for tuned RFCL model on balanced data

As we can see that the model is performing significantly better for the recall metric which is our main goal for this particular problem. Although, the accuracy of the model has decreased slightly but the increase in recall values is more significant those. The parameters obtained for this model are given below:

```
{'max_depth': 5,
'max_features': 4,
'min_samples_leaf': 5,
'min_samples_split': 5,
'n_estimators': 100}
```

Linear Discriminant Analysis

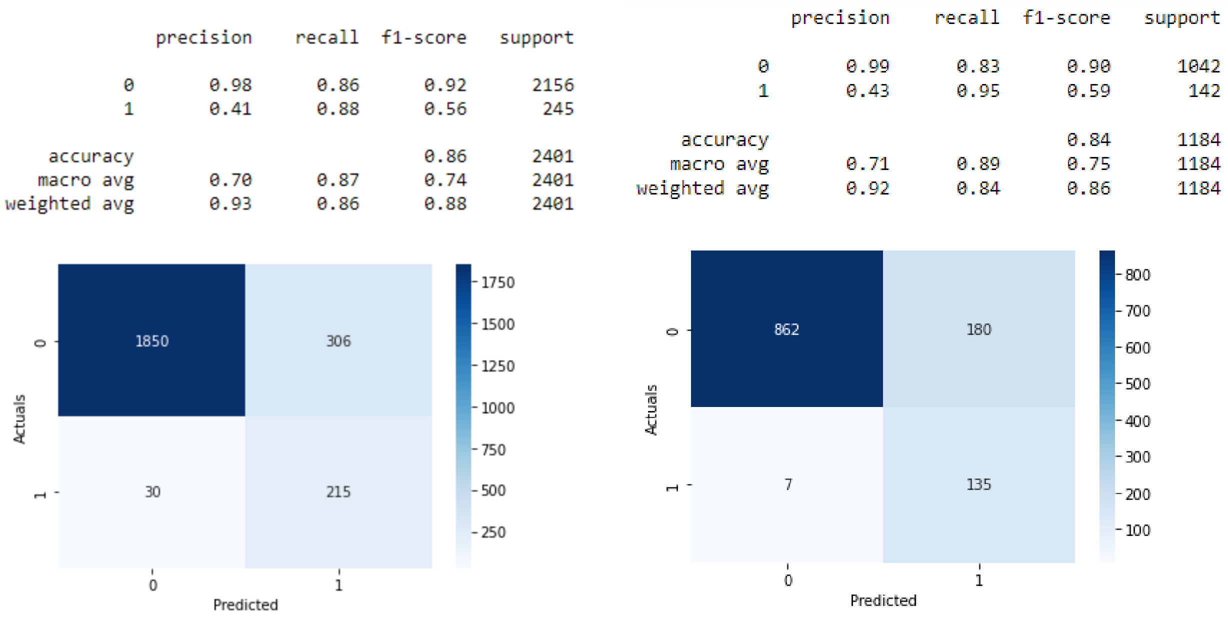
For training LDA model, similar process was followed as it was in Logit model. The best model, which performed well among the trained LDA models, was the one which included all the variables (Features) present in the dataset and was trained on Balanced dataset. But, even after using RFE technique and reduction of variables, the model performed quite similar to the “best LDA model” mentioned earlier. The feature importance for predicting class 1 (minority class i.e., defaulters) obtained after using RFE and their corresponding coefficients are given below:



	Feature	Coefficients	Ranking
8	Book_Value_Adj_Unit_Curr	-2.499066	1
14	ROG_Net_Worth_perc	-0.963251	1
23	Curr_Ratio_Latest	-1.615921	1
25	Debtors_Ratio_Latest	-0.432380	1
28	PBITM_perc_Latest	-0.607918	1
29	Debtors_Vel_Days	-0.593374	1
32	Value_of_Output_to_Gross_Block	-0.499152	1

Figure 4: Feature importance and their corresponding coefficients for LDA model

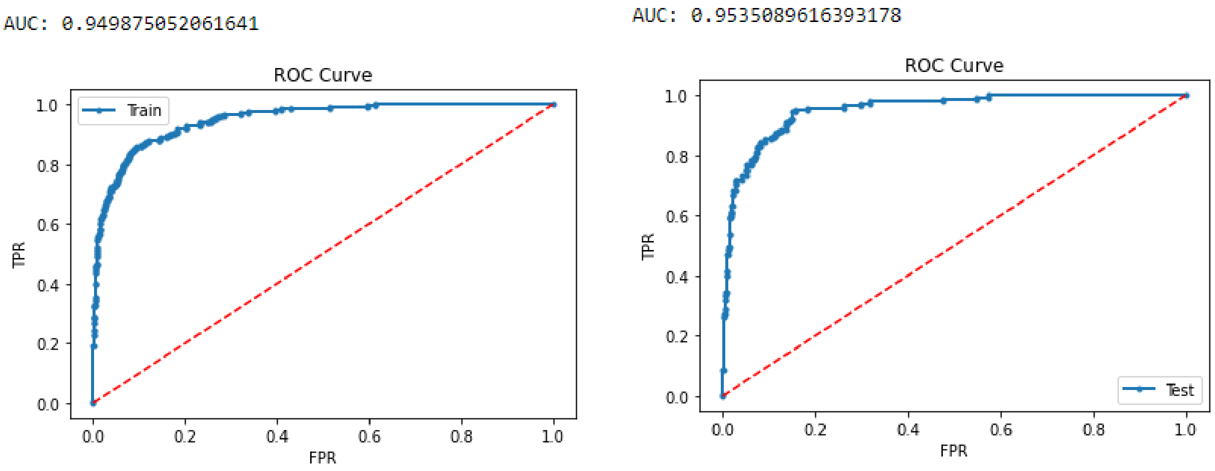
Now, based on the coefficients mentioned above, we can say that the "Debtors_Ratio_Latest" comes out to be the most important feature for prediction of companies which are going to default. Similarly, the "Book_Value_Adj_Unit_Curr" is most important feature for predicting the companies which are not going to default. Performance metrics for train and test set are given below:



```
              precision    recall  f1-score   support

           0       0.98      0.86      0.92      2156
           1       0.41      0.88      0.56       245

    accuracy                           0.86      2401
   macro avg       0.70      0.87      0.74      2401
weighted avg       0.93      0.86      0.88      2401
```

```
              precision    recall  f1-score   support

           0       0.99      0.83      0.90      1042
           1       0.43      0.95      0.59       142

    accuracy                           0.84      1184
   macro avg       0.71      0.89      0.75      1184
weighted avg       0.92      0.84      0.86      1184
```

Figure 5: Performance Metrics for LDA model after using RFE and balanced set (train set is on the left and test set performance is on the right)



```
AUC: 0.949875052061641
```

```
AUC: 0.9535089616393178
```

Figure 6: ROC Curves and AUC score for train (left) and test (right) set

Performance Comparison

Model	Accuracy		Recall		Precision		F1-score	
	Train	Test	Train	Test	Train	Test	Train	Test
LR (using RFE and Imbalanced Data)	96%	95%	67%	72%	92%	88%	78%	79%
LR (using RFE and balanced data)	94%	92%	88%	92%	65%	60%	75%	73%
RF Classifier (before tuning)	100%	98%	100%	89%	100%	95%	100%	95%
RF Classifier (before tuning)	98%	98%	85%	87%	96%	95%	90%	90%
RF Classifier (on balanced dataset after tuning)	95%	94%	96%	94%	70%	67%	81%	79%
LDA (on imbalanced data)	90%	88%	87%	87%	50%	49%	64%	63%
LDA (on balanced data)	89%	87%	90%	91%	49%	48%	63%	62%
LDA (using RFE on balanced data)	86%	84%	88%	95%	41%	43%	56%	59%

Table 1: Performance Comparison for the different models trained

Conclusion and Recommendations

Several models were trained for this classification problem in which we had to predict the companies which are going to default based on the balance sheet provided in the dataset. Following conclusions can be drawn based on the above analysis:

1. Random Forest Classifier after training on oversampled data and hyperparameter tuning, turns out to be the best model with **Recall (>94%)** for both train and test set.
2. Considering the interpretability of the model we can also go for logistic regression **Logit** at the cost of decreased precision value.
3. Although LDA model gave recall values very close to that of Logit for both train and test set. But, since the accuracy of the Logit model is much more **(8% train and 6% test set)**, Logit model should be picked between the two.

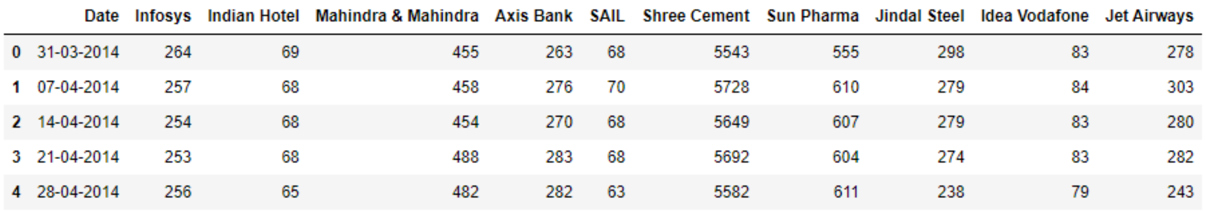
Problem 2

Stock Price Dataset Analysis

Information about the dataset

The dataset contains 6 years of information (weekly stock information) on the stock prices of 10 different Indian Stocks. We are assigned the task to calculate the mean and standard deviation on the stock returns and share insights.

Sample of the 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

Figure 7: Sample of Stock Price dataset

Stock Prices over time

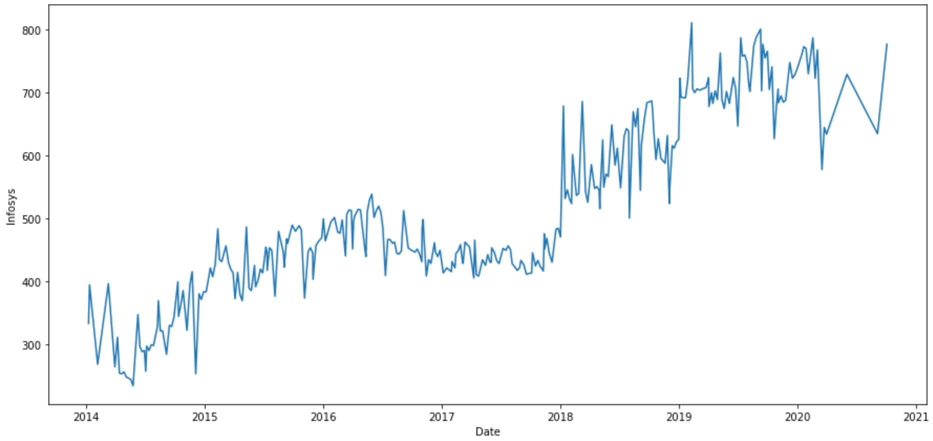


Figure 8: Stock Prices of Infosys

Observations:

1. An upward trend can be observed in Stock prices of Infosys.
2. Two sudden jumps in stock prices can observed in the year end of 2018 and 2019.
3. From 2017 to 2020 a similar pattern can be observed. Over a year, the stock prices do not vary much but in the year end there is a sudden rise in stock prices.
4. Variance in stock prices has increased in the past few years.

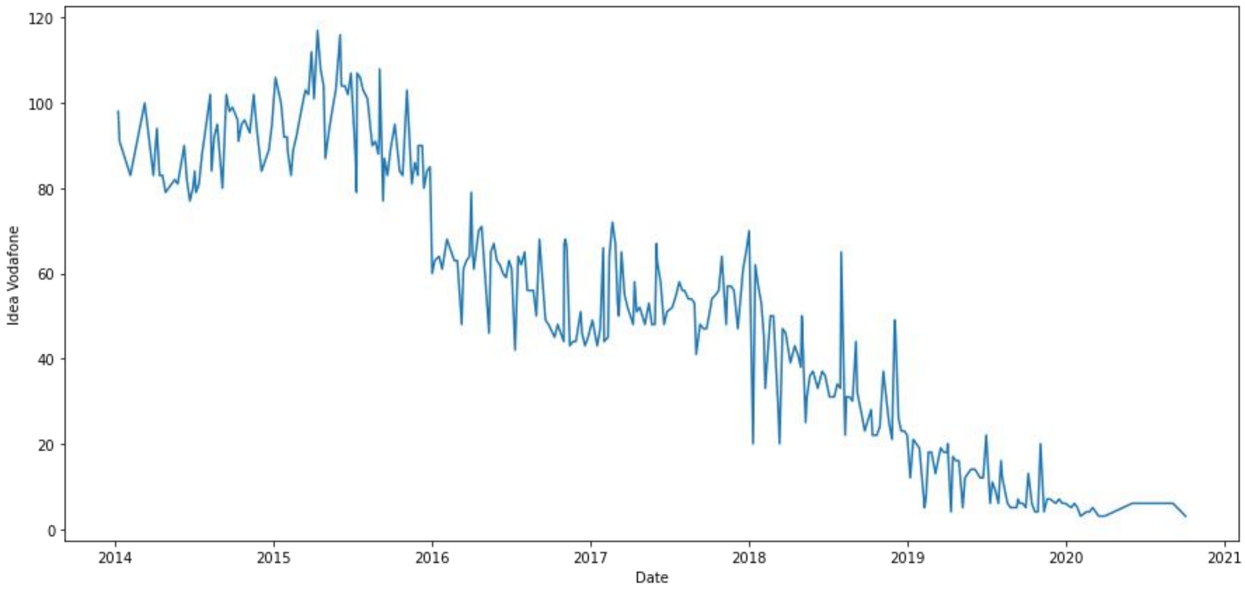


Figure 9: Stock Prices of Idea Vodafone

Observations:

1. A declination in the stock price over the years can be observed from the above plot.
2. Variance in stock prices has also decreased in the past few years.
3. A sudden decline in the stock prices can be observed in the start of the year 2016 **(right around the time when Reliance Jio announced their new plans i.e., 27 dec 2015).**

Returns Calculation

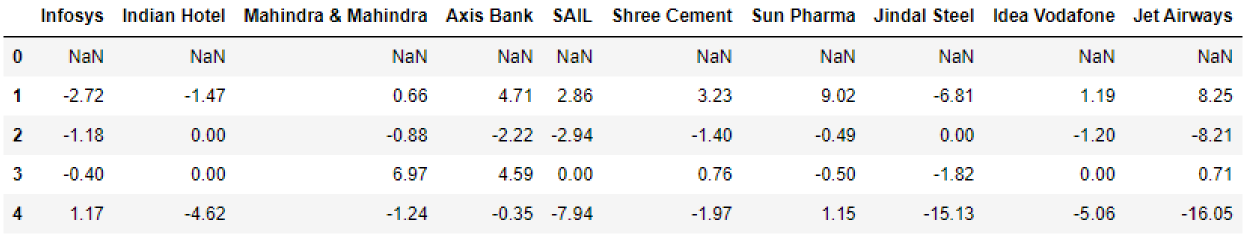
A return, also known as a financial return, in its simplest terms, is the money made or lost on an investment over some period of time. Return calculation can be done several ways. Simplest one can be done as per below mentioned formula:

$$\boxed{\boldsymbol{Return}\ (\boldsymbol{R_t}) = \frac{\boldsymbol{P_t} - \boldsymbol{P_{t-1}}}{\boldsymbol{P_t}}}$$

Where, $\boldsymbol{P_t}$: Price of stock at time t

$\boldsymbol{P_{t-1}}$: Price of stock at time $(t-1)$

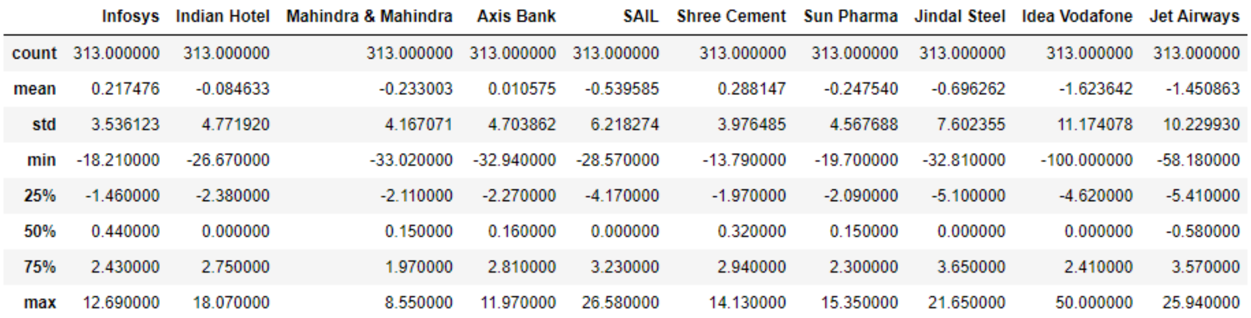
Using the above-mentioned formula, the returns were calculated. The sample of calculated return is given below:



	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	-2.72	-1.47	0.66	4.71	2.86	3.23	9.02	-6.81	1.19	8.25
2	-1.18	0.00	-0.88	-2.22	-2.94	-1.40	-0.49	0.00	-1.20	-8.21
3	-0.40	0.00	6.97	4.59	0.00	0.76	-0.50	-1.82	0.00	0.71
4	1.17	-4.62	-1.24	-0.35	-7.94	-1.97	1.15	-15.13	-5.06	-16.05

Figure 10: Returns (%) calculated for Stock Prices

Note that the first row is having nulls. That is because, the returns can be calculated only from the second timestamp mentioned in Stock Price dataset i.e., 07-04-2014. A continuously negative value of return represents a negative trend in Stock prices. Now, we can look at the statistical description of the returns to get some insights about the Stock price trends:

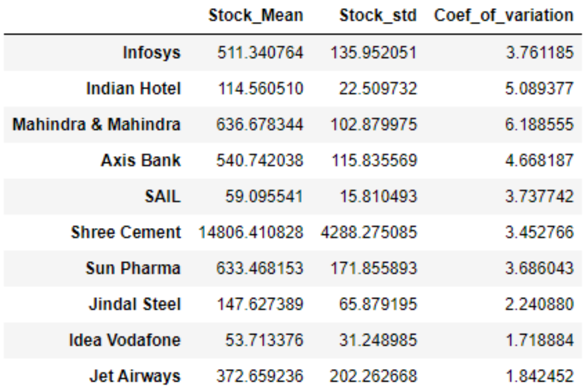


	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement	Sun Pharma	Jindal Steel	Idea Vodafone	Jet Airways
count	313.000000	313.000000	313.000000	313.000000	313.000000	313.000000	313.000000	313.000000	313.000000	313.000000
mean	0.217476	-0.084633	-0.233003	0.010575	-0.539585	0.288147	-0.247540	-0.696262	-1.623642	-1.450863
std	3.536123	4.771920	4.167071	4.703862	6.218274	3.976485	4.567688	7.602355	11.174078	10.229930
min	-18.210000	-26.670000	-33.020000	-32.940000	-28.570000	-13.790000	-19.700000	-32.810000	-100.000000	-58.180000
25%	-1.460000	-2.380000	-2.110000	-2.270000	-4.170000	-1.970000	-2.090000	-5.100000	-4.620000	-5.410000
50%	0.440000	0.000000	0.150000	0.160000	0.000000	0.320000	0.150000	0.000000	0.000000	-0.580000
75%	2.430000	2.750000	1.970000	2.810000	3.230000	2.940000	2.300000	3.650000	2.410000	3.570000
max	12.690000	18.070000	8.550000	11.970000	26.580000	14.130000	15.350000	21.650000	50.000000	25.940000

Figure 11: Statistical description of the Returns (%)

If we look at the mean values of the returns, we can observe ‘Shree Cement’ is doing best and ‘Idea Vodafone’ doing worst among the 10 companies.

Stock Means vs Standard Deviation



	Stock_Mean	Stock_std	Coef_of_variation
Infosys	511.340764	135.952051	3.761185
Indian Hotel	114.560510	22.509732	5.089377
Mahindra & Mahindra	636.678344	102.879975	6.188555
Axis Bank	540.742038	115.835569	4.668187
SAIL	59.095541	15.810493	3.737742
Shree Cement	14806.410828	4288.275085	3.452766
Sun Pharma	633.468153	171.855893	3.686043
Jindal Steel	147.627389	65.879195	2.240880
Idea Vodafone	53.713376	31.248985	1.718884
Jet Airways	372.659236	202.262668	1.842452

Figure 12: Stock Price Means and standard Deviation

Now, if we compare the coefficient of variation, we can observe that the value comes out to be highest for Mahindra & Mahindra which means the company is having less variation in stock prices compared to the rest of the stocks. Where as for ‘Idea Vodafone’ the value comes out to be lowest which means the stock price variation is more compared to the rest.

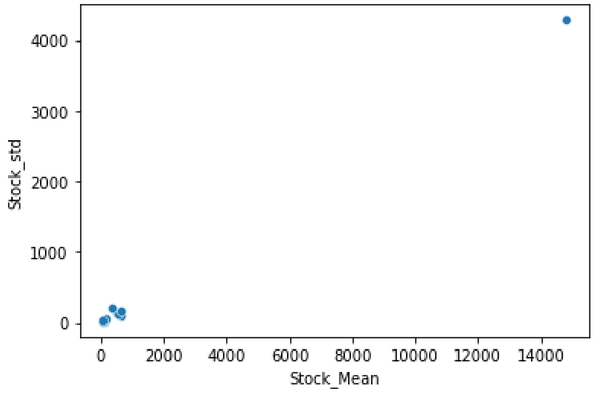


Figure 13: Stock Means vs Standard Deviation on Normal scale

Since, the value for “Shree Cement” is much higher than rest of the stocks, the rest of the data points get converged into very small. Hence, no insights can be generated by looking at this plot. Instead, we can plot the means and standard deviation on Logarithmic scale as followed:

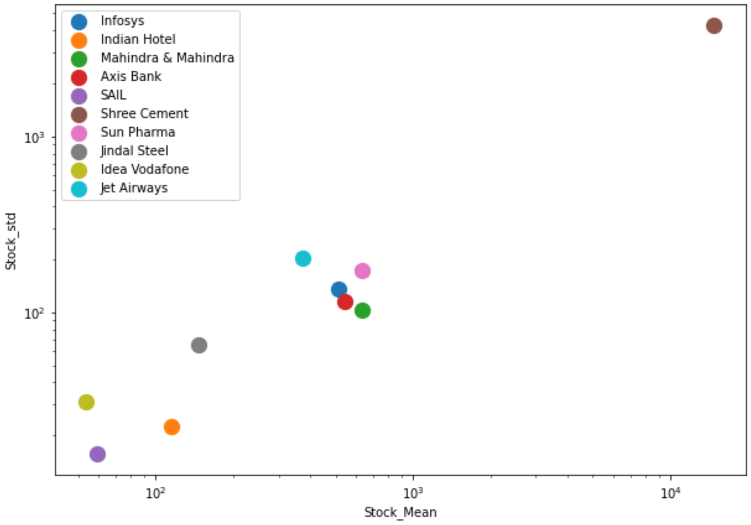


Figure 14: Stock Means vs Standard Deviation on Log scale

Now, the values which are far away from the apparent trend in the upward direction can be classified as ‘risky’ stocks as these stocks are having high stock price standard deviation compared to rest of the stocks. As per the observation from the plot we can say that ‘Idea Vodafone’ and ‘Jet Airways appears to be ‘risky’ stocks compared to the rest.

Conclusions and Recommendations

As per the simple analysis we performed, we can draw following conclusions:

1. **“Idea Vodafone”** and **“Jet Airways”** appears to be worst performing stocks compared to the rest.
2. Stock prices are significantly high for **“Shree Cement”.**
3. **“Mahindra & Mahindra”** appears to be safest stock to invest in, as the coefficient of variation came out to be the least for this stock. Though the returns are bad for this stock compared to rest.
4. Average returns are highest for **“Shree Cement”** with average value of **+0.28%.**

Now, as stated earlier the analysis which has been performed here is very simple one and should not be used for taking corporate decisions.