FRA MILESTONE 2 – MODEL COMPARISON AND MARKET RISK ANALYSIS

Akshaya Parthasarathy

PGP – DSBA Online June-E

Date: 27/06/2021

Table of Contents

Model comparisons 3

1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach 6

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 9

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, Random Forest and LDA models (include ROC Curve) 11

1.13 State Recommendations from the above models 12

Market Risk Analysis 13

2.1 Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference 14

2.2 Calculate Returns for all stocks with inference 16

2.3 Calculate Stock Means and Standard Deviation for all stocks with inference 17

2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference 18

2.5 Conclusion and Recommendations 19

Model comparisons

Problem statement:

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 labelled field.

Model comparisons based on Logistic Regression, LDA and Random Forest

Logistic Regression model results:

Model built:

Dep. Variable:	default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2390
Method:	MLE	**Df Model:**	11
Date:	Sun, 27 Jun 2021	**Pseudo R-squ.:**	0.5328
Time:	12:25:27	**Log-Likelihood:**	-384.74
converged:	True	**LL-Null:**	-823.47
Covariance Type:	nonrobust	**LLR p-value:**	4.349e-181

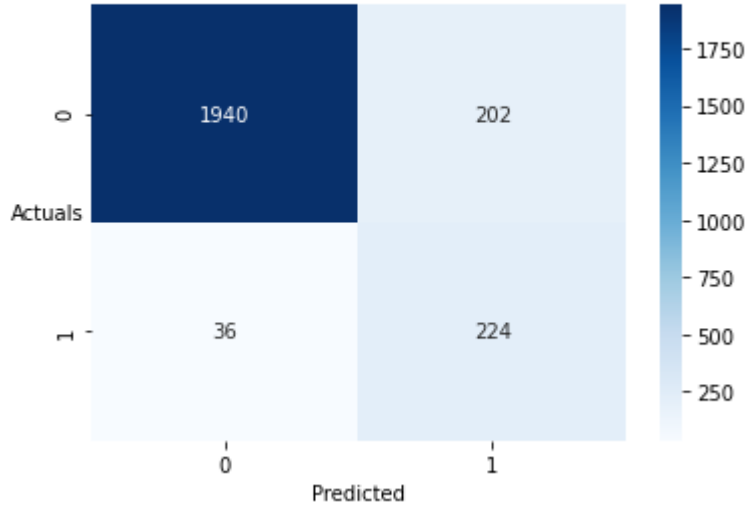
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0075	0.163	0.046	0.963	-0.312	0.327
Networth	-0.0161	0.003	-5.232	0.000	-0.022	-0.010
Gross_Block	0.0126	0.002	5.456	0.000	0.008	0.017
Value_Of_Output	-0.0214	0.004	-5.557	0.000	-0.029	-0.014
Cost_of_Prod	0.0278	0.005	5.807	0.000	0.018	0.037
Book_Value_Unit_Curr	-0.0350	0.012	-2.974	0.003	-0.058	-0.012
Book_Value_Adj_Unit_Curr	-0.0365	0.013	-2.782	0.005	-0.062	-0.011
ROG_Net_Worth_perc	-0.0290	0.012	-2.328	0.020	-0.053	-0.005
ROG_Cost_of_Prod_perc	-0.0155	0.005	-3.178	0.001	-0.025	-0.006
Current_Ratio_Latest	-1.1891	0.148	-8.053	0.000	-1.479	-0.900
Fixed_Assets_Ratio_Latest	-0.1900	0.078	-2.425	0.015	-0.344	-0.036
Interest_Cover_Ratio_Latest	-0.2831	0.065	-4.338	0.000	-0.411	-0.155

Identified 13 important variables using logistic regression (statsmodels). These variables have p-value less than 0.05 which shows their significance.

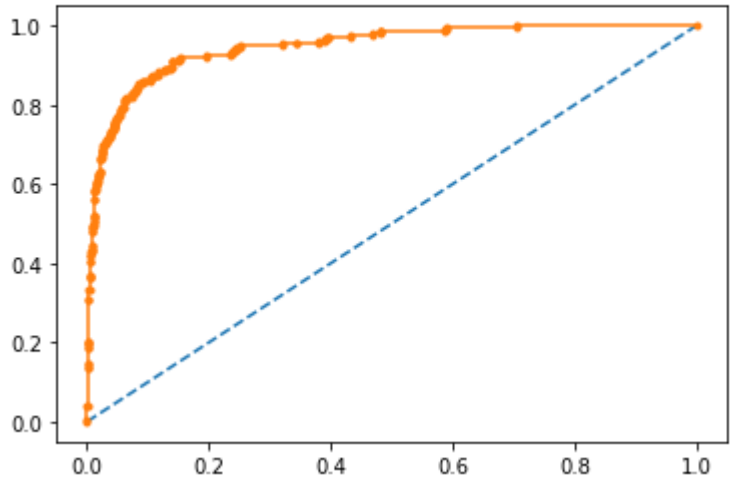
Performance metrics on train dataset:

```
Accuracy of the train dataset:  90.09
Classification Report:
              precision    recall  f1-score   support

           0      0.982     0.906     0.942      2142
           1      0.526     0.862     0.653       260

    accuracy                          0.901      2402
   macro avg      0.754     0.884     0.798      2402
weighted avg      0.932     0.901     0.911      2402
```

Confusion matrix:



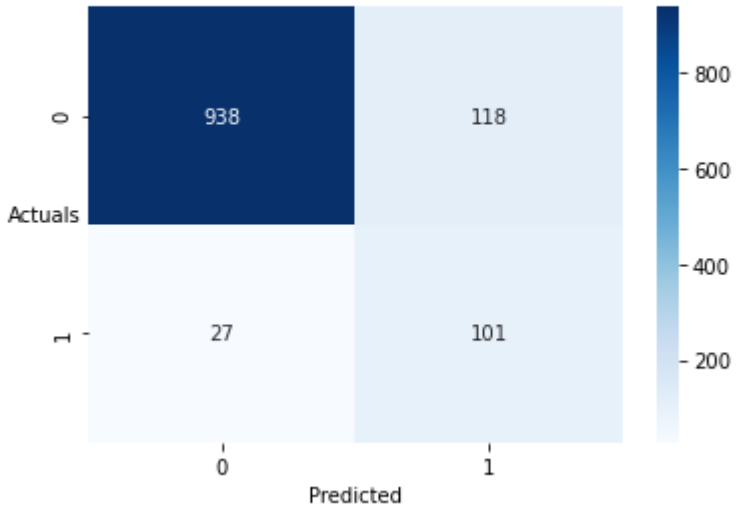
AUC for the Training Data: 0.948



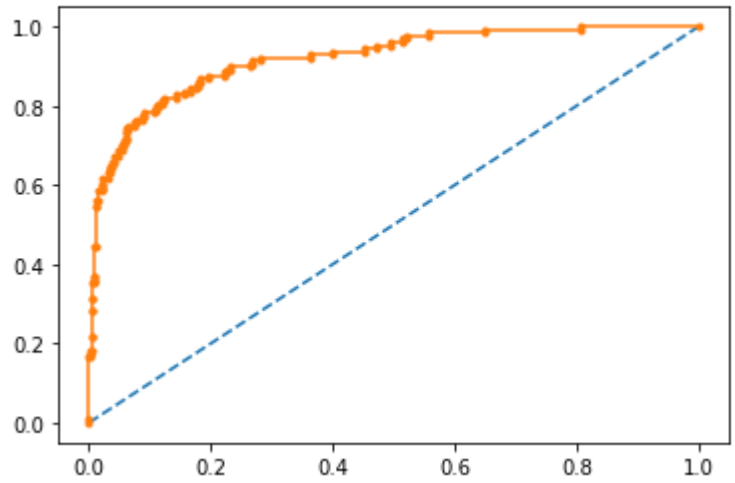
Performance metrics on test dataset:

```
Accuracy of the test dataset:  87.75
Classification Report:
              precision    recall  f1-score   support

           0      0.972     0.888     0.928      1056
           1      0.461     0.789     0.582       128

    accuracy                         0.878      1184
   macro avg      0.717     0.839     0.755      1184
weighted avg      0.917     0.878     0.891      1184
```

Confusion matrix:



AUC for the Testing Data: 0.920



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

Random forest is an ensemble technique that combines several base models (decision trees) in order to produce one optimal predictive model.

Building model using grid search cross validation method and tuning the hyper parameters in order to obtain a stable random forest.

Grid search parameters used:

```
GridSearchCV(cv=3, estimator=RandomForestClassifier(random_state=1),
             param_grid={'max_depth': [10, 15, 20],
                         'max_features': [5, 10, 15, 20],
                         'min_samples_leaf': [20, 30, 40],
                         'min_samples_split': [60, 90, 120],
                         'n_estimators': [100, 150, 200]})
```

Best model with pruning parameters:

```
{'max_depth': 10, 'max_features': 10, 'min_samples_leaf': 30, 'min_samples_split': 60, 'n_estimators': 150}
```

Variable importance:

Top 10 variables:

	Imp
Networth	0.304146
Book_Value_Unit_Curr	0.232806
Book_Value_Adj_Unit_Curr	0.201760
Current_Ratio_Latest	0.055493
Capital_Employed	0.047661
CEPS_annualised_Unit_Curr	0.019512
PBIDT	0.018357
CP	0.015574
Total_Asset_Turnover_Ratio_Latest	0.013238
PBDT	0.009857

Bottom 10 variables:

	Imp
ROG_PBIT_perc	0.000342
Cash_Flow_From_Investing_Activities	0.000331
Inventory_Ratio_Latest	0.000264
Value_of_Output_to_Total_Assets	0.000246
Selling_Cost	0.000224
ROG_CP_perc	0.000212
ROG_Capital_Employed_perc	0.000149
Revenue_exp_in_forex	0.000145
Revenue_earn_in_forex	0.000025
Capital_exp_in_forex	0.000000

Random forest classifier gives high importance to Net worth, Book_Value_Unit_Curr and Book_Value_Adj_Unit_Curr with more than 20% and rest of the variables have less than 5% importance. Moreover, in this model we could see that there are many variables with very less importance.

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

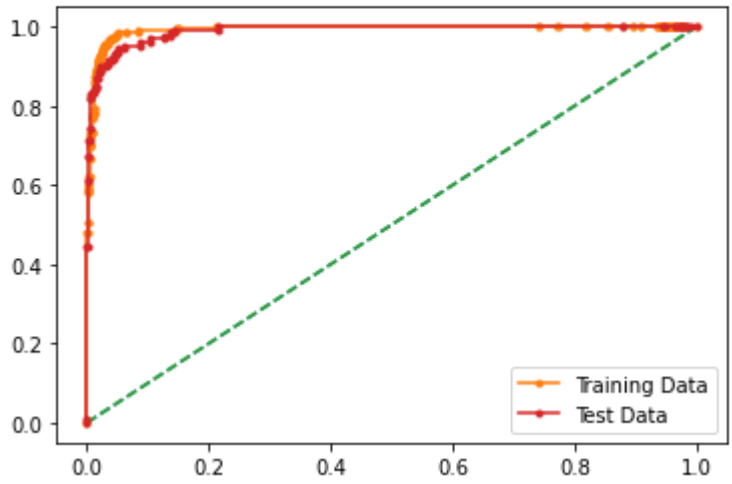
Accuracy of both train and test set

```
Accuracy of training data:  96.75
Accuracy of testing data:  96.62
```

Both the accuracy is approximately equal. There is no evidence of over or under fitting.

AUC – ROC curve for train and test dataset

```
AUC for the Training Data: 0.993
AUC for the Test Data: 0.989
```



Classification Report

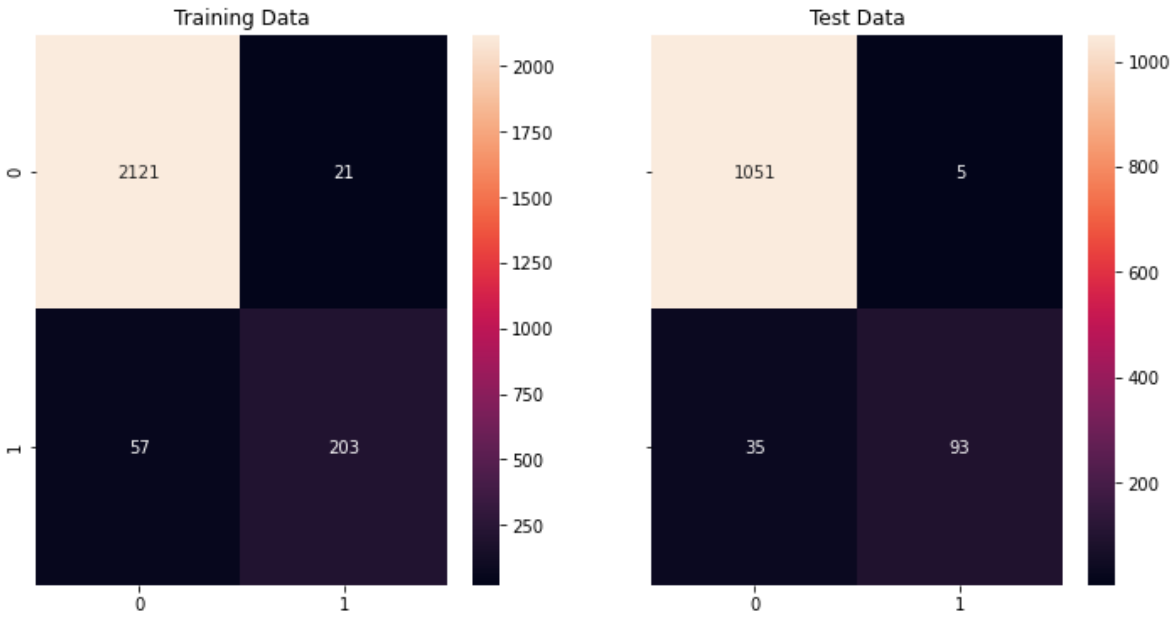
Classification Report of the training data:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	2142
1	0.91	0.78	0.84	260
accuracy			0.97	2402
macro avg	0.94	0.89	0.91	2402
weighted avg	0.97	0.97	0.97	2402

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	1056
1	0.95	0.73	0.82	128
accuracy			0.97	1184
macro avg	0.96	0.86	0.90	1184
weighted avg	0.97	0.97	0.96	1184

Confusion matrix



Training and testing set results are almost similar, with overall measures on a decent side.

Net worth, Book_Value_Unit_Curr and Book_Value_Adj_Unit_Curr are the most important variable for predicting claimed.

Recall of defaulters has come down with respect to the Logistic Regression model. Accuracy and precision of defaulter class have good values compared to Logistic Regression.

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

Linear Discriminant Analysis is a linear classification machine learning algorithm.

The algorithm involves developing a probabilistic model per class based on the specific distribution of observations for each input variable. A new data point is then classified by calculating the conditional probability of it belonging to each class and selecting the class with the highest probability.

This model is useful when we have independent variables are a clear distinguishers of target variable.

Grid search CV parameters used:

```
LinearDiscriminantAnalysis()

GridSearchCV(cv=3, estimator=LinearDiscriminantAnalysis(), n_jobs=-1,
             param_grid={'solver': ['svd', 'lsqr', 'eigen'],
                         'tol': [0.0001, 1e-05]})
```

Best parameters obtained:

```
{'solver': 'svd', 'tol': 0.0001}

LinearDiscriminantAnalysis()
```

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

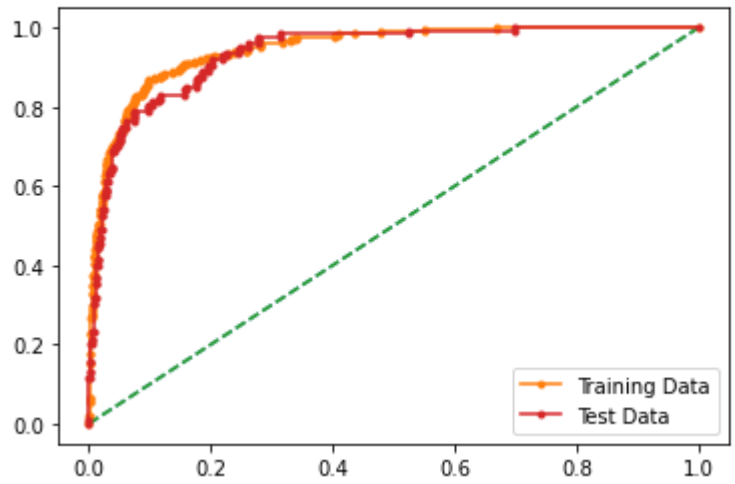
Accuracy of both train and test set

```
Accuracy of training data:  93.05
Accuracy of testing data:  92.4
```

Accuracy of test data is decreased. There is no evidence of over or under fitting. But overall LDA has the lower accuracy than random forest model but higher than logistic regression.

AUC – ROC curve for train and test dataset

```
AUC for the Training Data: 0.947
AUC for the Test Data: 0.938
```



Classification Report

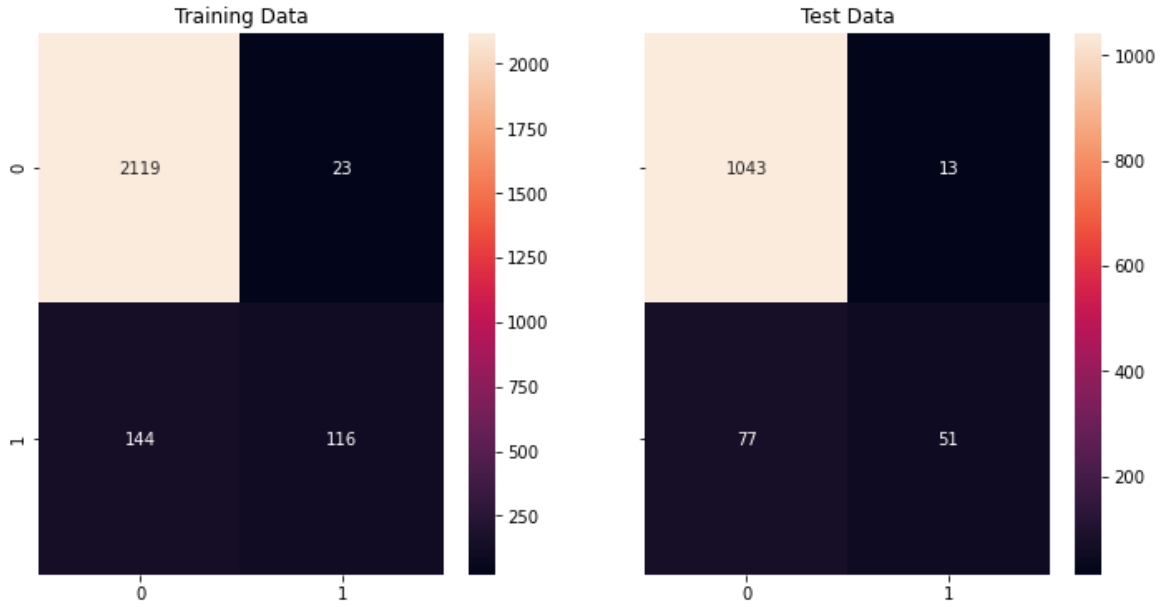
```
Classification Report of the training data:
```

	precision	recall	f1-score	support
0	0.94	0.99	0.96	2142
1	0.83	0.45	0.58	260
accuracy			0.93	2402
macro avg	0.89	0.72	0.77	2402
weighted avg	0.93	0.93	0.92	2402

```
Classification Report of the test data:
```

	precision	recall	f1-score	support
0	0.93	0.99	0.96	1056
1	0.80	0.40	0.53	128
accuracy			0.92	1184
macro avg	0.86	0.69	0.74	1184
weighted avg	0.92	0.92	0.91	1184

Confusion matrix



Training and testing set results are almost similar.

Recall of defaulters has come down with respect to the other models. Since in this problem recall is the important metric, LDA is not performing as well compared to others. Hence this model is not suitable.

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

Models are evaluated on the basis of the below techniques to see how good it will perform for future records.

Some of the model evaluation techniques are:

- Accuracy – how precisely the model classifies the data points.
- Confusion Matrix – 2 * 2 tabular structure reflecting the model performance in four blocks

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
Predicted Values	Negative (0)	FN	TN

- Receiver operating characteristics (ROC) curve – A technique to visualize classifier performance
- ROC_AUC score – Area under curve, which is by calculating the percentage area below the curve.

Best model for the given case study:

From the above results, we can say that Random Forest seem to be the optimized model for the given dataset. Since in this case study, classes are not well separated and there are lots of independent variables, hence LDA lacks the accuracy in discriminating between the class.

Moreover, the training and test set results are in line for Random Forest than other models.

1.13 State Recommendations from the above models

With increasing amount of data, companies and industries try to remain competitive by keeping themselves ahead of the curve. By analysing huge amounts of financial data, companies are able to obtain valuable information to determine their strategic plans such as risk control, crisis management or growth management. Logistic regression, Random Forest and LDA models have been employed in predicting the defaulters of companies.

Consolidated performance metrics (for the default class):

Models	**Logistic Regression**		**Random Forest**		**Linear Discriminant Analysis**	
Metrics	**Training set**	**Testing set**	**Training set**	**Testing set**	**Training set**	**Testing set**
Accuracy	90.97	88.85	96.75	96.62	93.05	92.4
Precision	0.55	0.49	0.91	0.95	0.83	0.80
Recall	0.88	0.82	0.78	0.73	0.45	0.40
F1 score	0.67	0.61	0.84	0.82	0.58	0.53

- **False positive (FP)** - Datapoints that are actually non-default but predicted as default. This is also known as type 1 error. In order to reduce the type 1 error, we have to increase the precision of the model (among the points identified as positives by the model how many are actually positive).

  Type 1 error in this case study means model has classified the data point as 1 instead of 0. The company which are identified wrongly as default might try to reassess the net worth, change in working capital or change in strategy as required. They might as well choose to sale their non-profitable investments. This might not be of priority for our case study, since predicting the actual non-defaulters as default might have minimal impact than false negatives tend to pose.

- **False negative (FN)** – Datapoints that are actually default but predicted as non-defaulters. This is known as type 2 error. In order to reduce to type 2 error. We have to increase recall (how many actual true data points are identified as true by the model)

  Type 2 error is that the model has classified the data point as 0 instead of 1. In which case, the company will continue their strategy which makes it risky investment for shareholders. Concentrating on this type of error is a priority.

Market Risk Analysis

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.

Exploratory data analysis

Head of the dataset (after renaming the column headers)

	Date	Infosys	Indian_Hotel	Mahindra_N_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

Shape of the dataset

```
The number of rows (observations) is 314
The number of columns (variables) is 11
```

Information of the dataset (after converting the data type of date column into 'datetime')

```
<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    datetime64[ns]
 1   Infosys              314 non-null    int64
 2   Indian_Hotel         314 non-null    int64
 3   Mahindra_N_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: datetime64[ns](1), int64(10)
memory usage: 27.1 KB
```

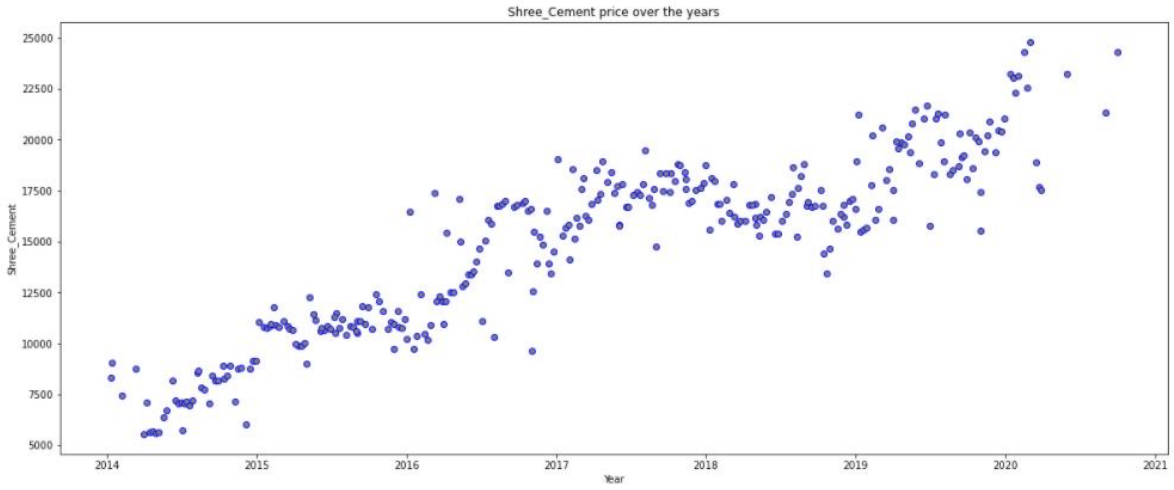
Summary statistics for price of all the stocks

	Infosys	Indian_Hotel	Mahindra_N_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways
count	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000
mean	511.340764	114.560510	636.678344	540.742038	59.095541	14806.410828	633.468153	147.627389	53.713376	372.659236
std	135.952051	22.509732	102.879975	115.835569	15.810493	4288.275085	171.855893	65.879195	31.248985	202.262668
min	234.000000	64.000000	284.000000	263.000000	21.000000	5543.000000	338.000000	53.000000	3.000000	14.000000
25%	424.000000	96.000000	572.000000	470.500000	47.000000	10952.250000	478.500000	88.250000	25.250000	243.250000
50%	466.500000	115.000000	625.000000	528.000000	57.000000	16018.500000	614.000000	142.500000	53.000000	376.000000
75%	630.750000	134.000000	678.000000	605.250000	71.750000	17773.250000	785.000000	182.750000	82.000000	534.000000
max	810.000000	157.000000	956.000000	808.000000	104.000000	24806.000000	1089.000000	338.000000	117.000000	871.000000

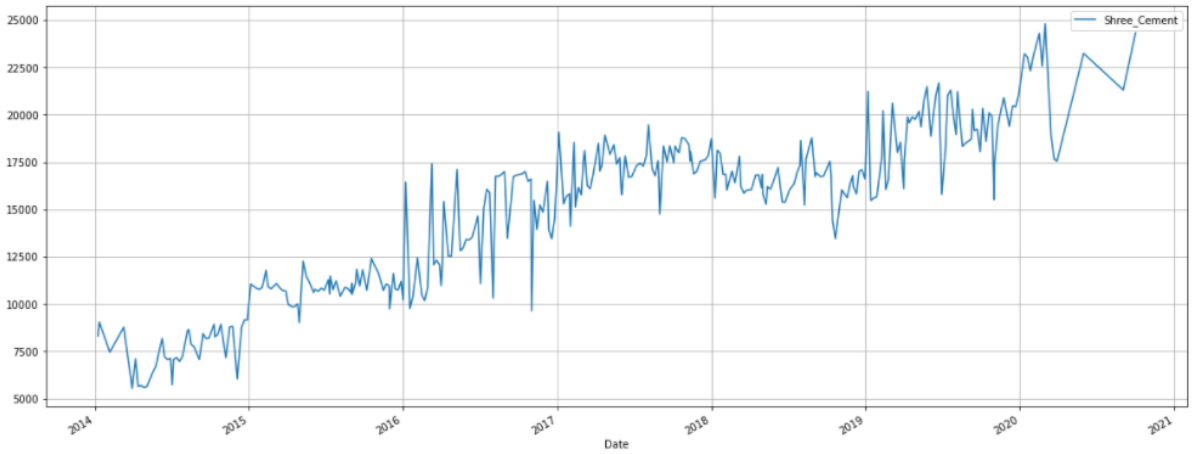
2.1 Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference

Taken two stocks 'Shree_Cement' and 'Idea_Vodafone' to explain the stock price graph.

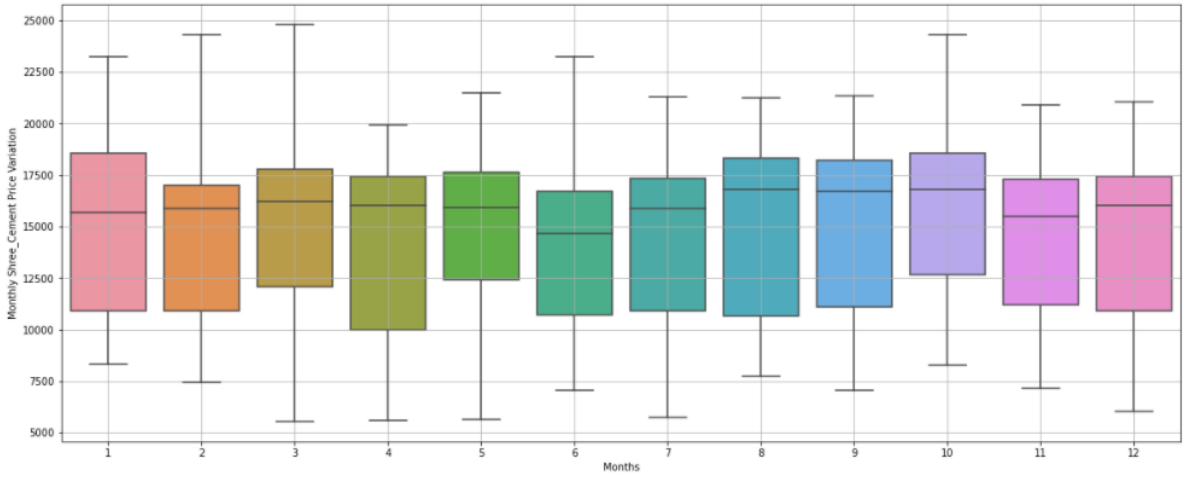
Shree cement price over time



This scatterplot shows the trend of Shree_Cement price over time. We can observe an upward trend over the years. Price was at 5000 during 2014 and it has increased to 25000 at 2020.

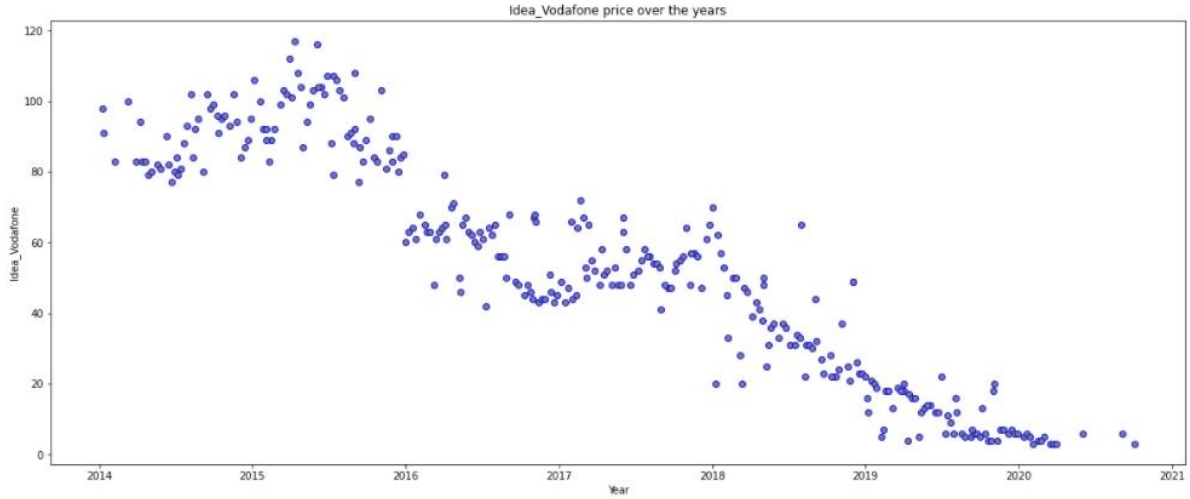


Month wise plot to check for seasonality

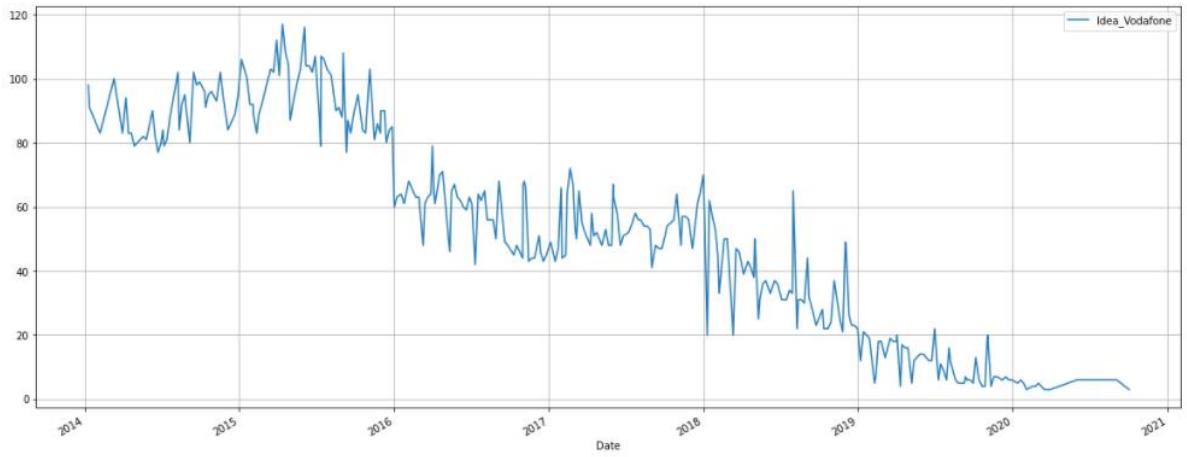


There is no visible seasonality present in the dataset for Shree_Cement price since the median is in and around 15000. But we can observe that the range of the prices are widely spread.

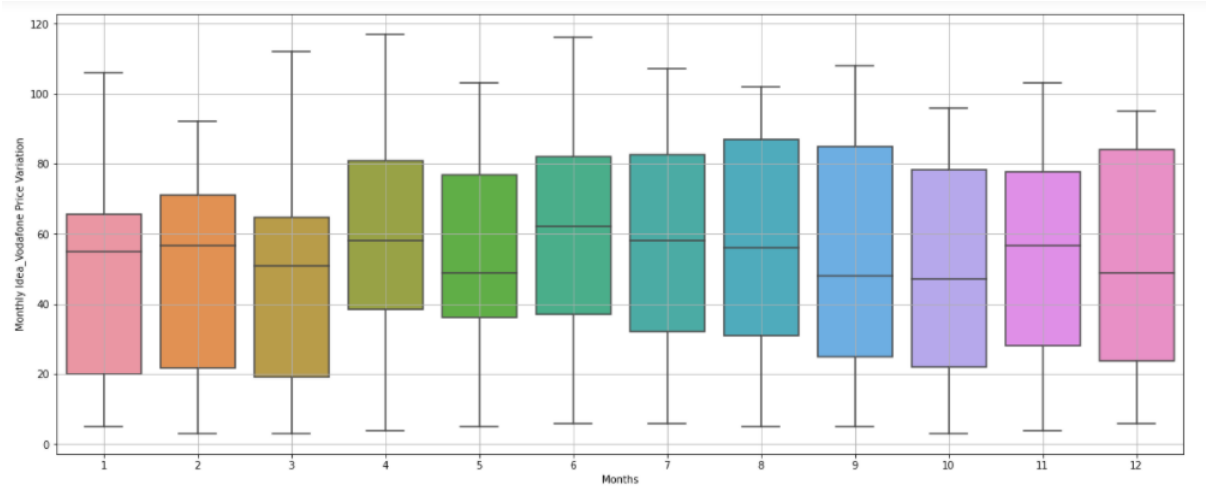
Idea_Vodafone price over time



This scatterplot shows the trend of Shree_Cement price over time. We can observe an downward trend over the years. Price was at 100 during 2014, peaks are available at 120 during 2015 end and it has decreased to 0 at 2019 itself.



Month wise plot to check for seasonality



There is no visible seasonality present in the dataset for Idea_Vodafone price since the median is between the range 50-60. Here also, we can observe that the range of the prices are widely spread.

2.2 Calculate Returns for all stocks with inference

Returns are the change in stock price as a proportion of what the stock price was in the earlier time period. Calculating returns by taking the log is preferred when we look at multiple time periods.

	Bharti_Airtel	DLF	ACC	BHEL	TCS	Maruti_Suzuki	Reliance	Dr_Reddy	ITC	TATA_Steel	Sensex
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	0.048949	0.025975	0.018762	0.077962	0.037945	0.083871	0.027239	0.019755	0.031416	0.031416	0.037893
2	0.003180	0.008511	-0.038656	-0.038221	-0.044411	0.022529	-0.025269	0.013401	-0.015585	0.063249	0.008215
3	0.031253	0.057629	0.020651	-0.026317	0.047951	-0.005792	-0.056695	-0.017117	-0.005249	-0.009725	-0.009001
4	-0.012384	-0.032523	-0.012186	0.013245	-0.025046	0.006617	-0.012579	-0.074473	-0.021277	-0.063887	-0.014877

A positive return is the profit, or money made, on the stock. Likewise, a negative return represents the loss or money lost on the stock. This is an important metric to calculate how well the stock has performed

Summary statistics of the returns dataset

	Infosys	Indian_Hotel	Mahindra_N_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	0.002794	0.000266	-0.001506	0.001167	-0.003463	0.003681	-0.001455	-0.004123	-0.010608	-0.009548
std	0.035070	0.047131	0.040169	0.045828	0.062188	0.039917	0.045033	0.075108	0.104315	0.097972
min	-0.167300	-0.236389	-0.285343	-0.284757	-0.251314	-0.129215	-0.179855	-0.283768	-0.693147	-0.458575
25%	-0.014514	-0.023530	-0.020884	-0.022473	-0.040822	-0.019546	-0.020699	-0.049700	-0.045120	-0.052644
50%	0.004376	0.000000	0.001526	0.001614	0.000000	0.003173	0.001530	0.000000	0.000000	-0.005780
75%	0.024553	0.027909	0.019894	0.028522	0.032790	0.029873	0.023257	0.037179	0.024391	0.036368
max	0.135666	0.199333	0.089407	0.127461	0.309005	0.152329	0.166604	0.243978	0.693147	0.300249

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

Stock mean – Average: The mean return of the selected stocks is intended to represent the behaviour of the market and to report the composite change in prices of the stocks.

```
Infosys              0.002794
Indian_Hotel         0.000266
Mahindra_N_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
dtype: float64
```

Each average reflects the general movement of each stock and serves as a benchmark for the performance of individual stocks in its sphere. Positive mean value is when we can see increase in stock prices than the initial and negative mean value is the decrease in stock prices.

Stock standard deviation – Volatility: It is a statistical measure of volatility, measuring how widely prices are dispersed from the average price. If prices trade in a narrow trading range, the standard deviation will return a low value that indicates low volatility. Conversely, if prices swing wildly up and down, then standard deviation returns a high value that indicates high volatility.

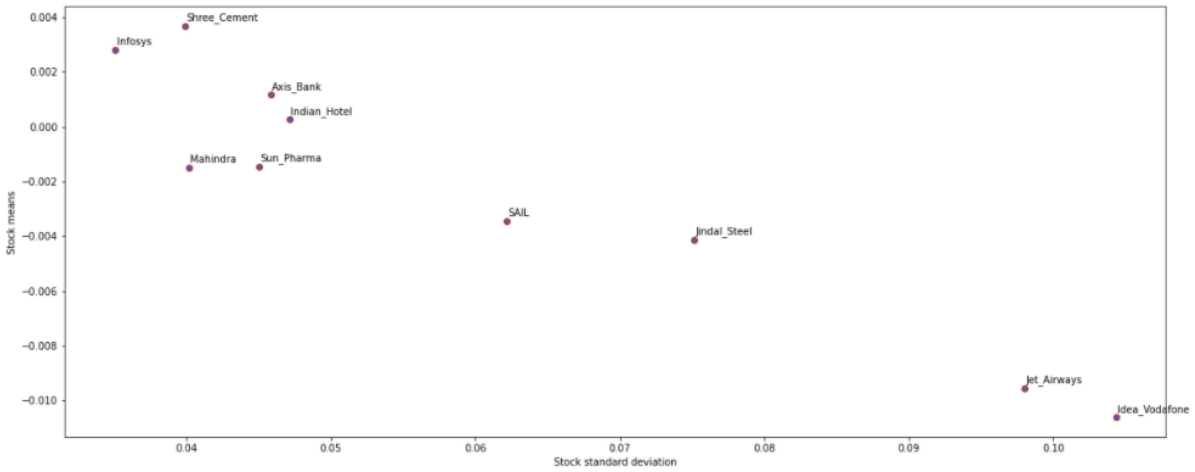
Highest standard deviation among the given companies is for Idea_Vodafone. This can be observed from the graphs before where can see a steep decrease in stock prices from 2018.

```
Infosys                 0.035070
Indian_Hotel            0.047131
Mahindra_N_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
dtype: float64
```

Created a dataframe with the stock mean values and stock standard deviation.

	Average	Volatility
Infosys	0.002794	0.035070
Indian_Hotel	0.000266	0.047131
Mahindra_N_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

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



Here is a combined plot of volatility vs average of all the given stocks.

Standard deviation is a measure of risk. It is used to capture the uncertainty of variables. If the value is high, then the returns are more uncertain and if the value is low, then the returns are less uncertain and risky.

From the above plot we can see that the stocks of Idea_Vodafone and Jet_Airways have the highest standard deviation and low mean values which means these stocks are a risky investment since they are in the declining stage.

The stocks of Shree_Cement, Infosys, Axis_Bank and Indian_Hotel have positive means which means the stock prices are increasing and they have correspondingly less standard deviation which represents that these stocks are less risky to invest in.

Comparing Shree_Cement and Mahindra which has comparative standard deviation values but the mean of Shree_Cement stocks are higher than that of Mahindra. At this point of time, Mahindra might be performing less as compared to other stock. Same goes with Mahindra and Sun pharma, these two stocks have same average stock prices but the volatility of Sun Pharma is higher than that of Mahindra, hence concluding that Sun Pharma might be a risky option.

2.5 Conclusion and Recommendations

Arranging the stocks by increasing order of volatility:

	Average	Volatility
Infosys	0.002794	0.035070
Shree_Cement	0.003681	0.039917
Mahindra_N_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

Simple definition of volatility is a reflection of the degree to which price moves A stock with a price that fluctuates wildly—hits new highs and lows or moves erratically—is considered highly volatile. A stock that maintains a relatively stable price has low volatility. A highly volatile stock is inherently riskier, but that risk cuts both ways. When investing in a volatile security, the chance for success is increased as much as the risk of failure. For this reason, we have to keep in mind the financial position of the company, performance and other metrics to determine what to invest in.

We can observe that Idea Vodafone and Jet Airways are the highly volatile stocks in the decreasing trend since the prices of these stocks have reduced drastically because of the greater number of selling movements involved rather than buying. As these companies are out of market now, they are not an ideal option for investing.

On the other hand, Infosys and Shree Cements are less volatile but their price is now showing increasing trend and selling of these stocks would be ideal since it can come down any time after this. Investing in a portfolio of first four stocks can be a better option since averaging out the prices gives the benefits of diversification.