Capstone Project-2

Yes bank stock closing price prediction

Ashfaque Sayyed

Contents

Problem statement

Yes Bank is a well-known bank in the Indian financial domain. Since 2018, it has been in the news because of the fraud case involving Rana Kapoor. Owing to this fact, it was interesting to see how that impacted the stock prices of the company and whether Time series models or any other predictive models can do justice to such situations. This dataset has monthly stock prices of the bank since its inception and includes closing, starting, highest, and lowest stock prices of every month. The main objective is to predict the stock's closing price of the month.

What is stock price

- The term stock price refers to the current price that a share of stock is trading for on the market.
- The price of a stock will go up and down in relation to a number of different factors, including changes within the economy as a whole, changes within industries, political events, war, and environmental changes.

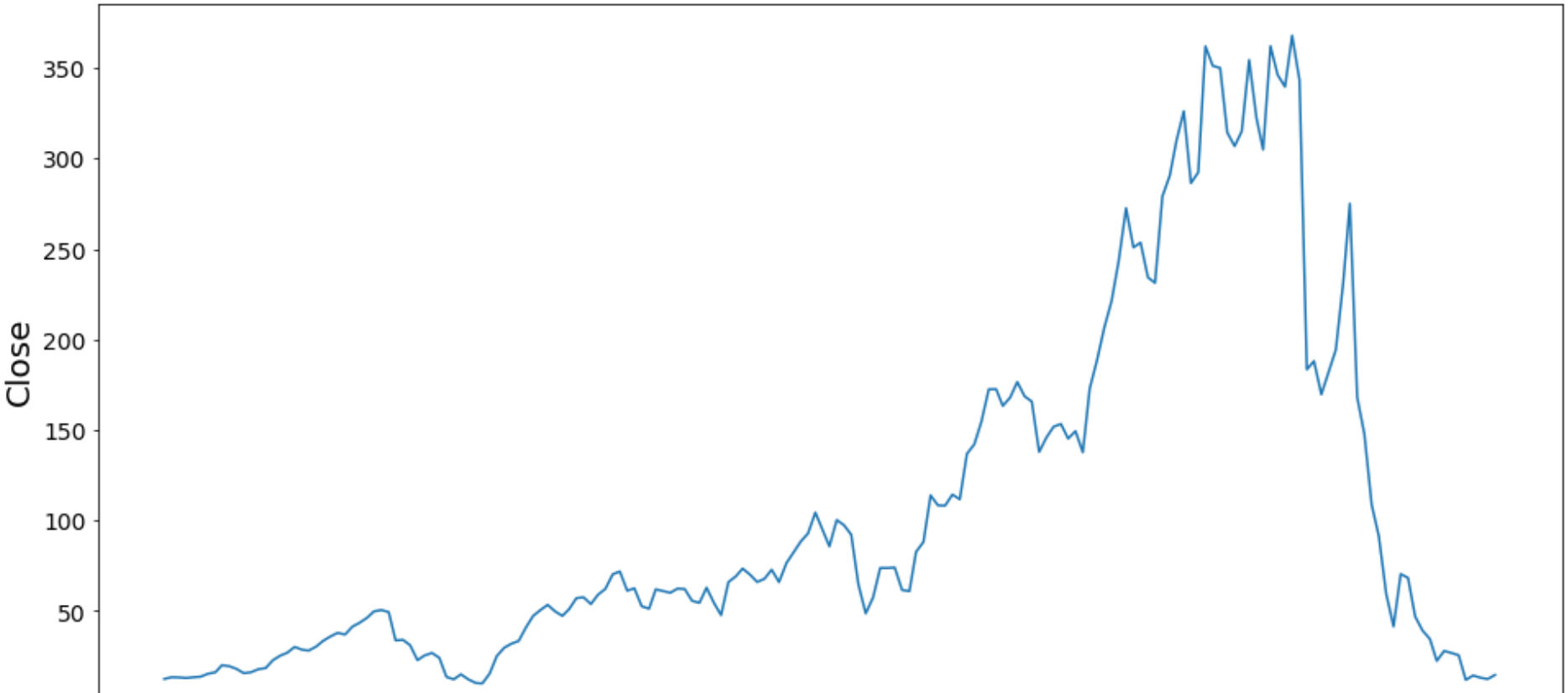
Data summary

Date: It specifies the date of particular stock price.
Open: It specifies the opening price of the stock in that month (Numeric).
High: It specifies the highest price of the stock in that month (Numeric).
Low:It specifies the Lowest price of the stock in that month (Numeric).
Close:It specifies the Closing price of the stock in that month (Numeric).

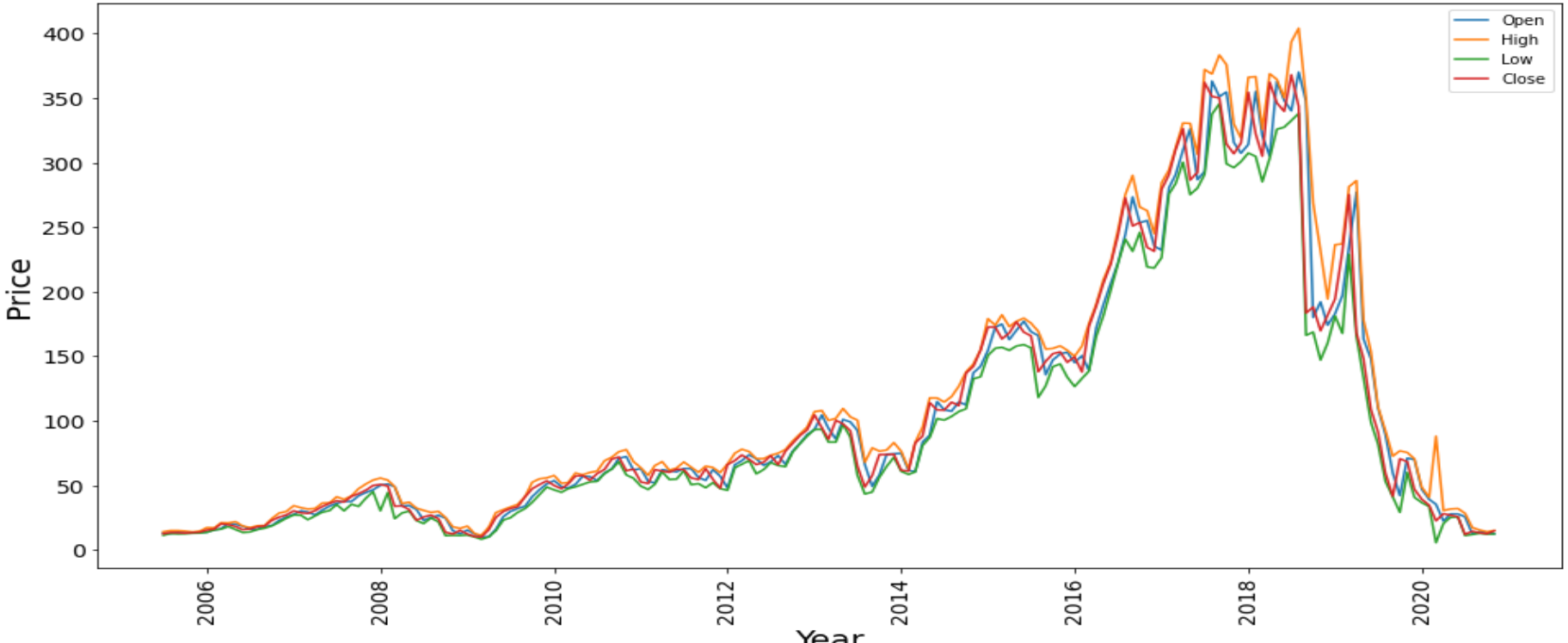
Exploratory data analysis

1. Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypothesis,
2. and to check assumptions with the help of summary statistics and graphical representations.
3. Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals

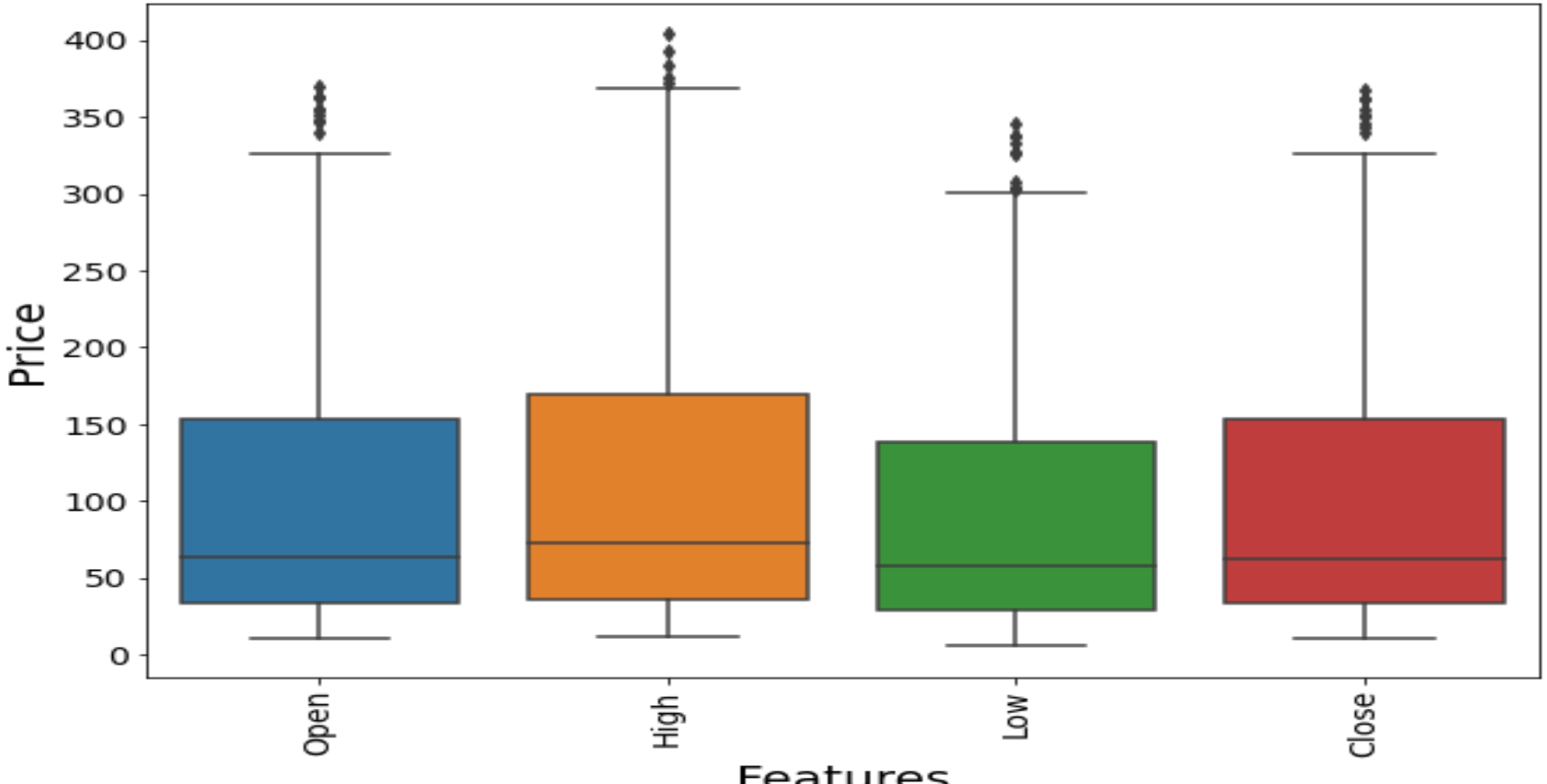
Plot of date VS closing price



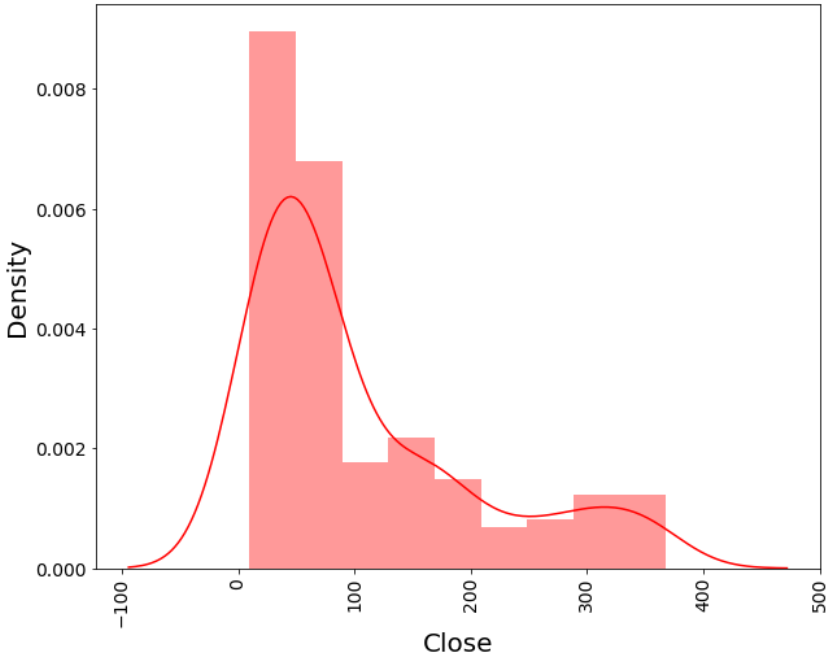
Plot of all prices against date

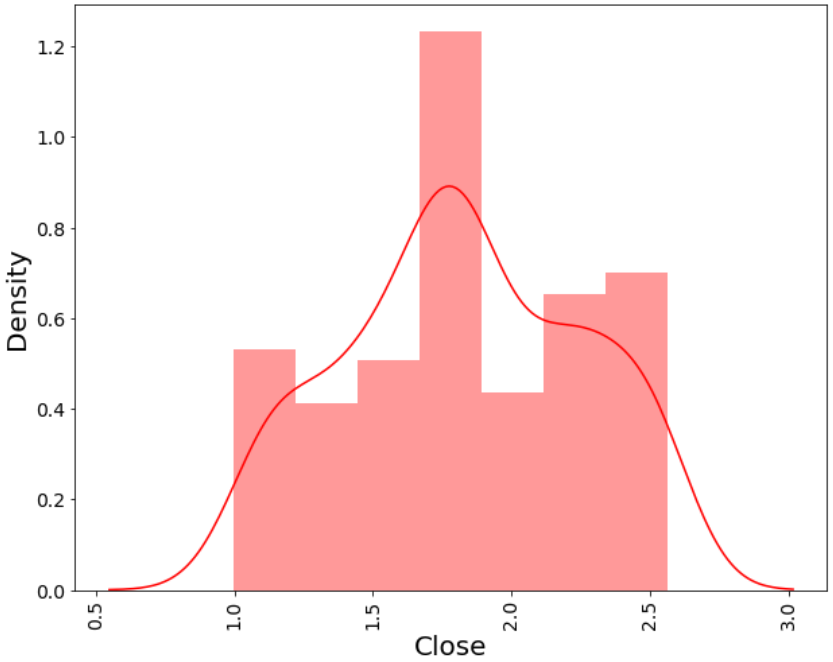


Boxplot

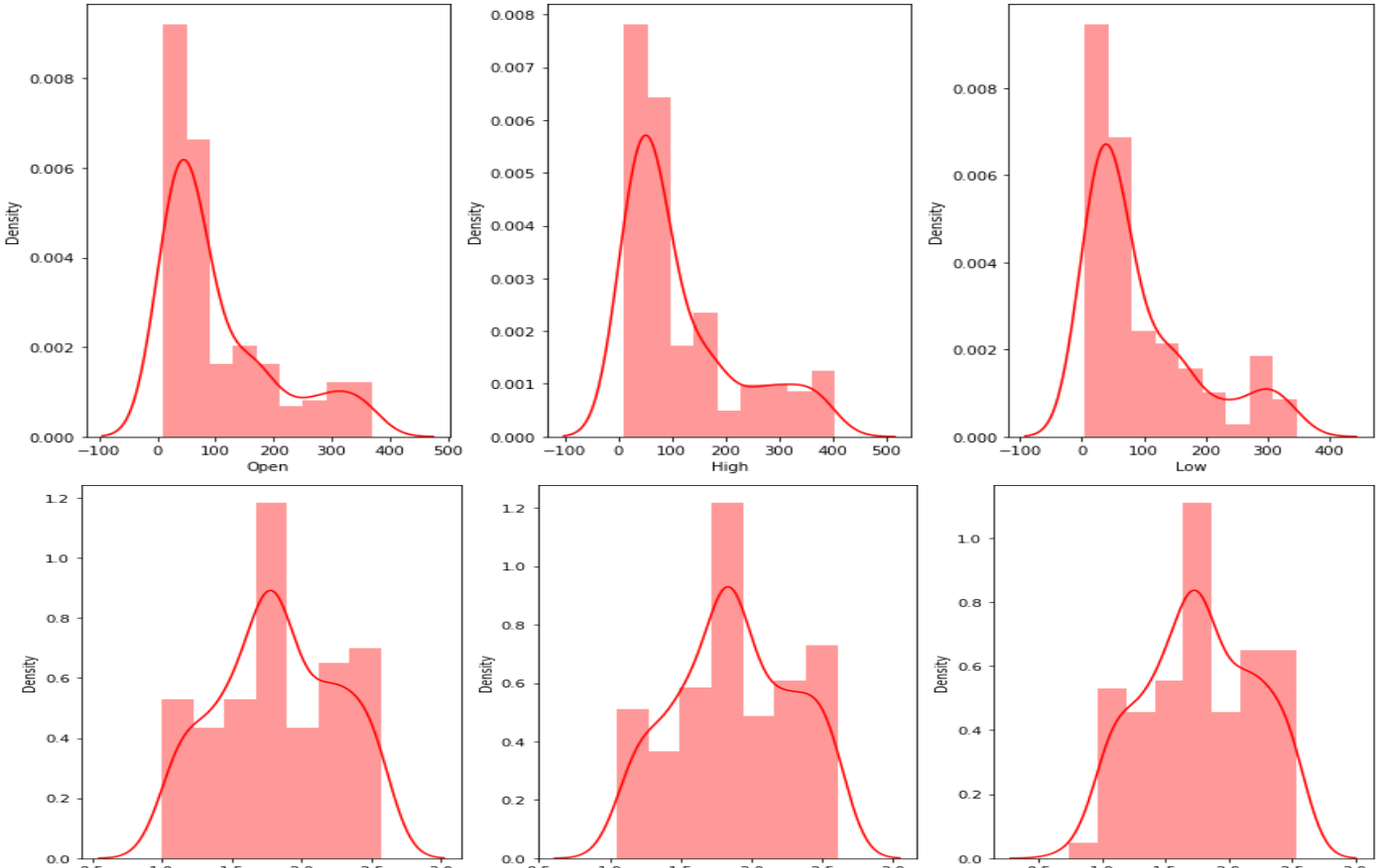

400
350
300
250
200
150
100
50
0
Price
Open
High
Low
Close
Features

Histogram of dependent variable

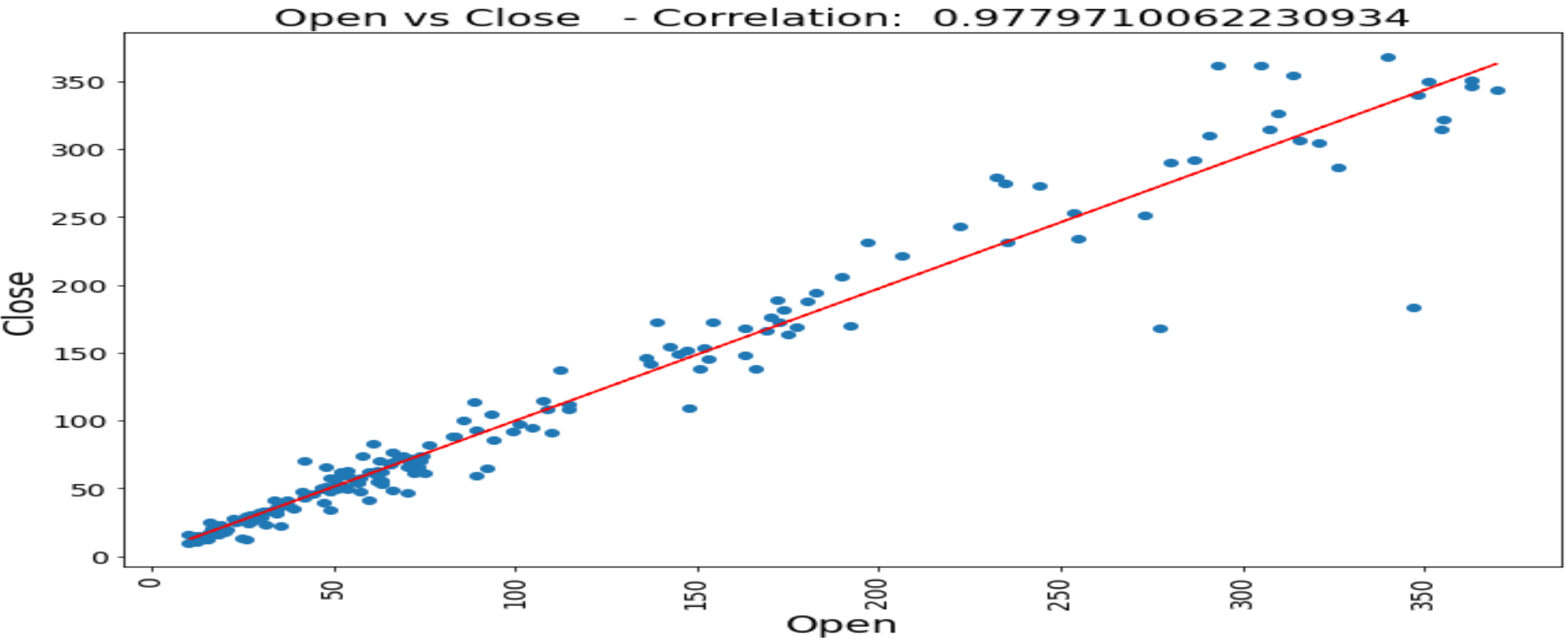


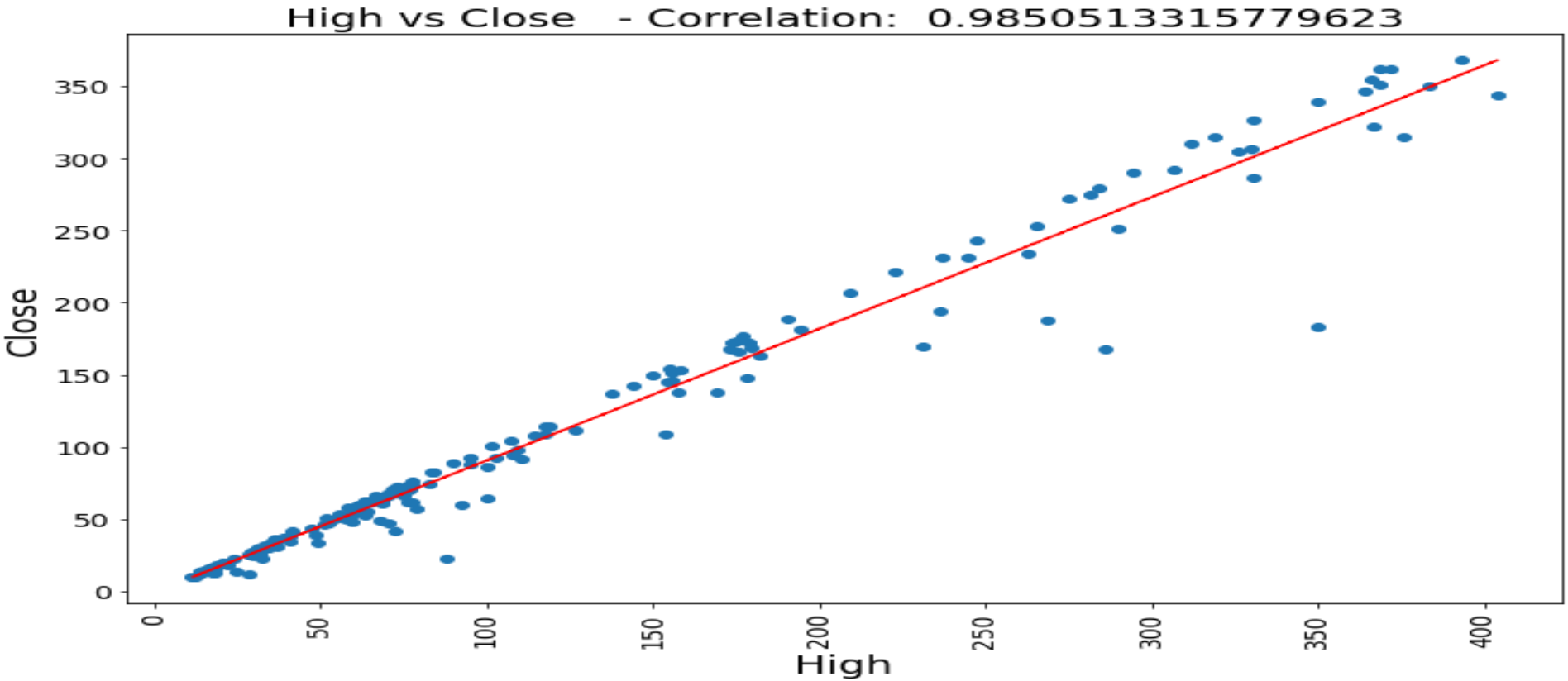


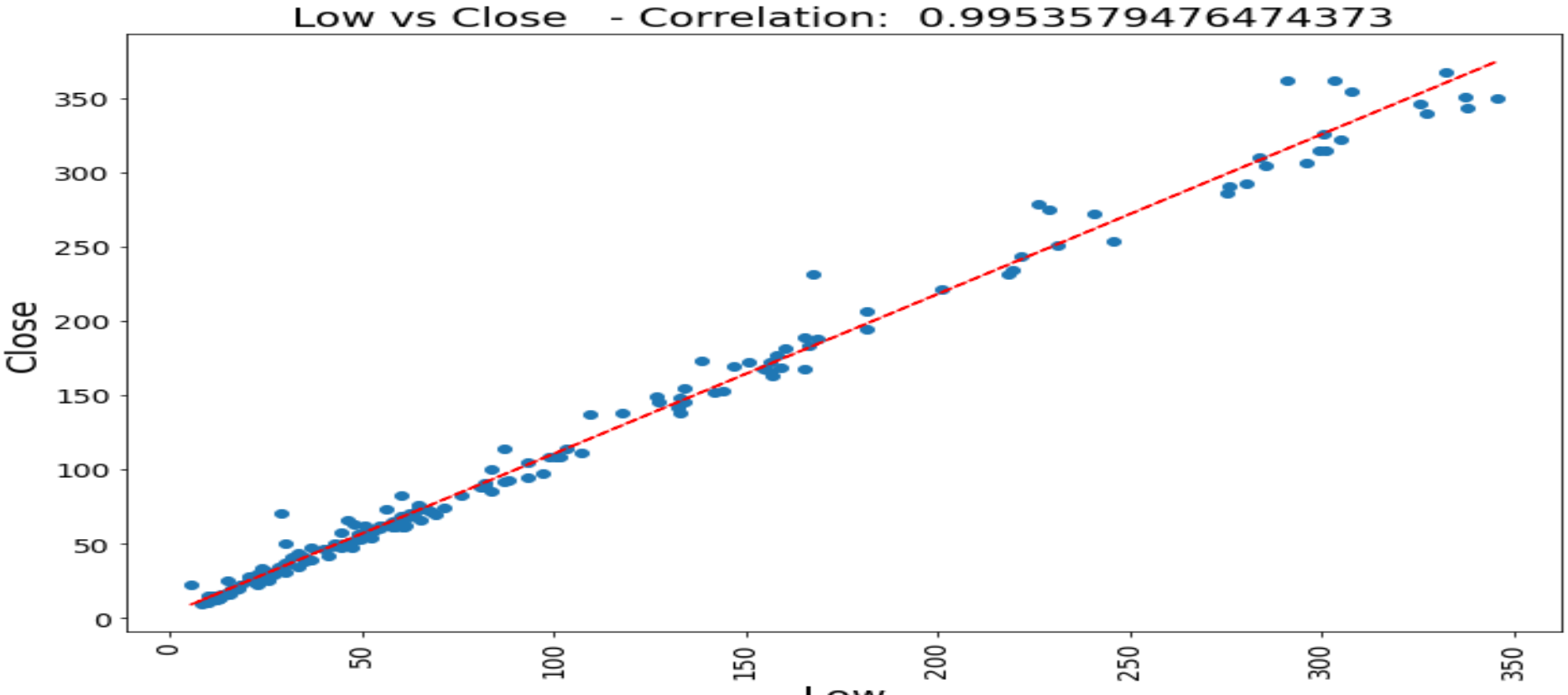
Histogram of Independent variable



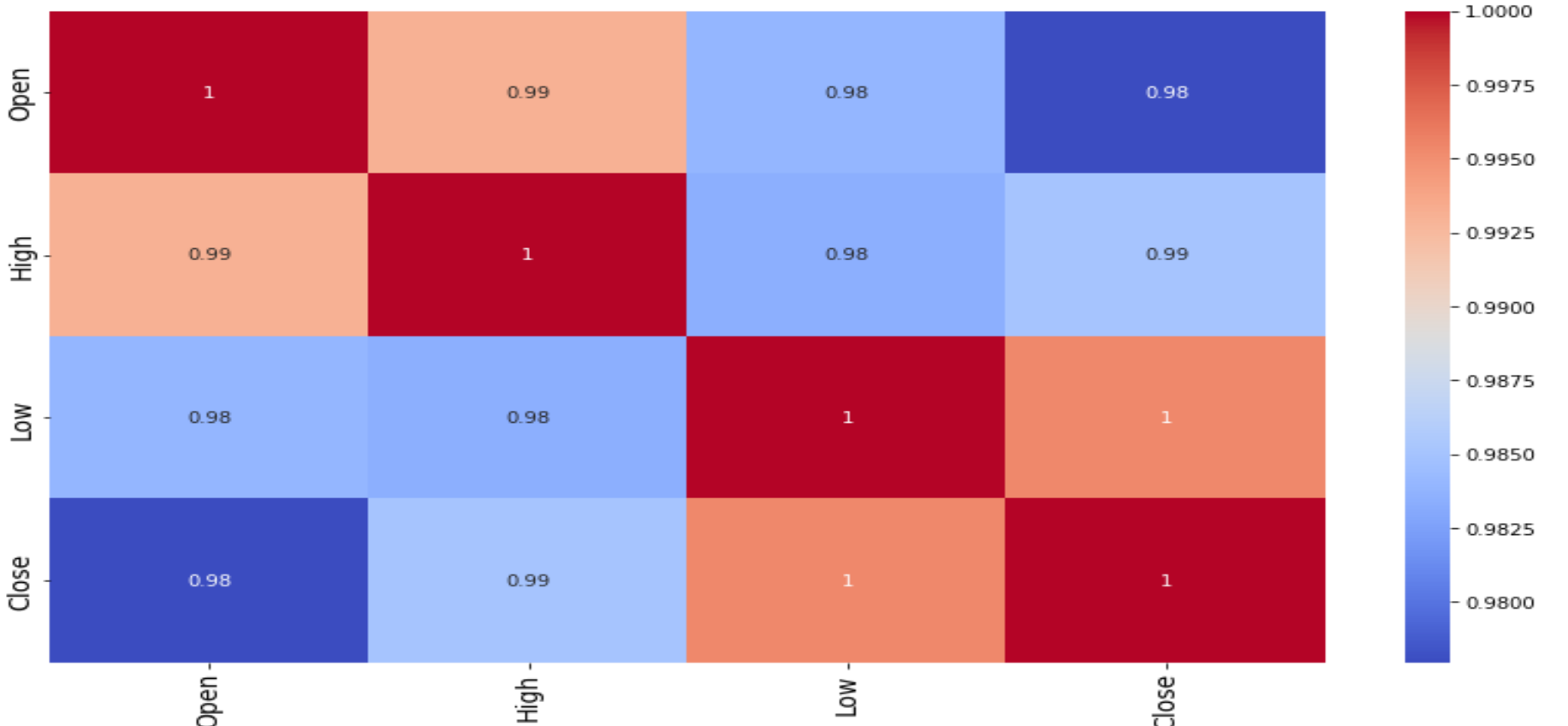
Scatter plot of independent variable vs dependent variable




High vs Close - Correlation: 0.985051331577962 3
Close
High
0
50
100
150
200
250
300
350
400


Low vs Close - Correlation: 0.995357947647437
Close
Low
0
50
100
150
200
250
300
350

Correlation



Train- test split

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

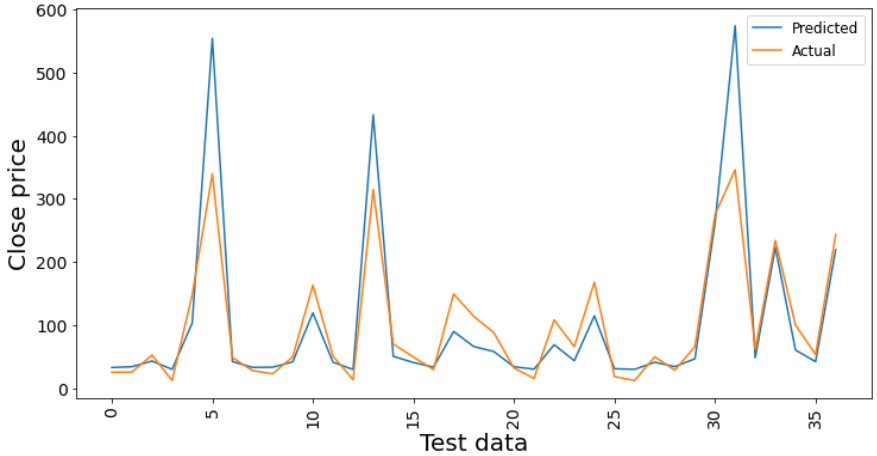
In this project we have used 80% data for training purpose and 20% data for test set.

The train-test procedure is appropriate when there is a sufficiently large dataset available.

Models

1.Linear regression

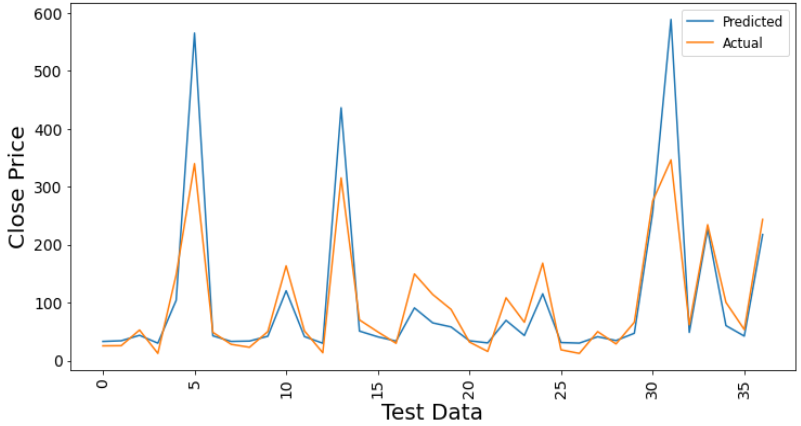
The basic aim of linear regression is to plot the best fit line between a dependent variable and an independent variable.



We can see the difference between the actual price and the predicted price.
Differences are comparatively high at peak points.

2. Lasso Regression

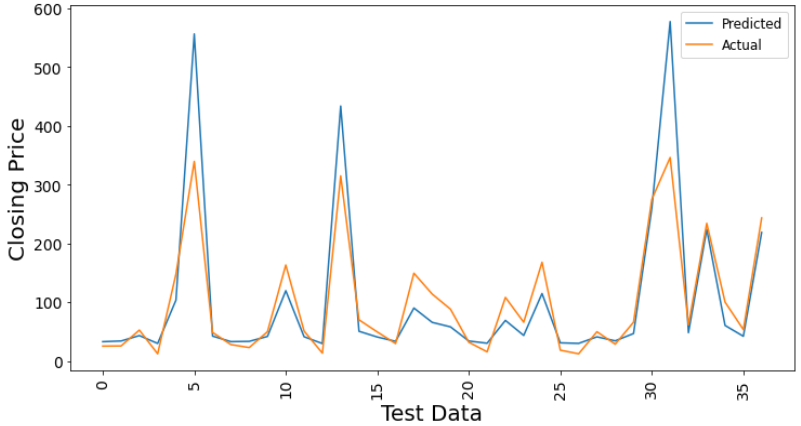
- Lasso: Least Absolute Shrinkage and Selection operator.
- It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.
- This method performs L1 regularization.



- The graph shows same characteristics that of a linear regression, there is a difference between actual and predicted values.
- Prediction has higher values than actual values.

3. Ridge Regression

• Ridge regression is a model tuning method that is used to analyses any data that suffers from multicollinearity.
• When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.
• This method performs L2 regularization.



- In ridge regression also values are not predicted correctly.

- At peak, prediction have higher values than actual values.
- Characteristics is similar as in linear and lasso regression.

Cross Validation and Hyperparameter Tuning

- Cross Validation is a technique using which Model is evaluated on the dataset on which it is not trained that is it can be a test data or can be another set as per availability or feasibility.
- Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of linear based models like Lasso and Ridge.
- We used Grid Search CV for hyperparameter tuning.
- Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance.

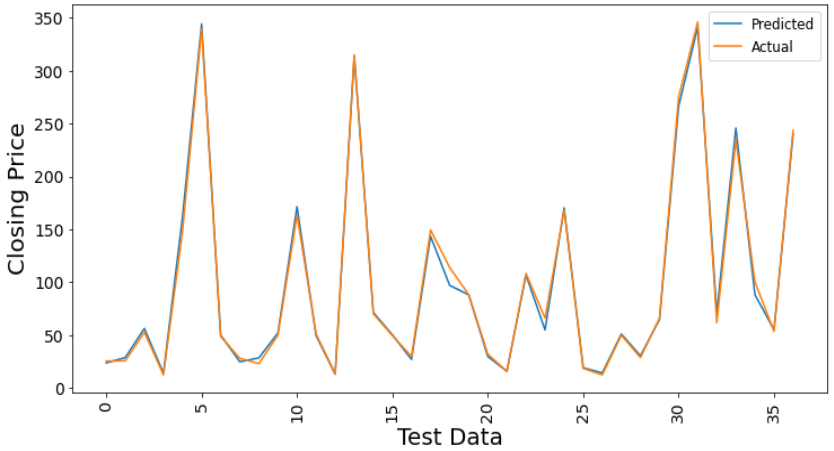
	Model	Train score	Test score	MSE	RMSE	MAE	MAPE	R2 Score
0	Linear regression	0.815	0.823	0.032	0.178	0.151	0.095	0.823
1	Lasso regression	0.815	0.821	0.032	0.179	0.152	0.096	0.821
2	Lasso after validation	NaN	NaN	0.032	0.180	0.153	0.097	0.819
3	Ridge regression	0.815	0.822	0.032	0.178	0.151	0.095	0.822
4	Ridge after validation	NaN	NaN	0.032	0.178	0.151	0.095	0.822

- Applying cross validation and hyperparameter tuning have not much effect on accuracy.
- The best Fit alpha value for lasso regression is : 0.01
- The negative mean squared error is : -0.035
- The Best Fit alpha value for ridge regression is : 10
- The negative mean squared error for is : -0.035

We applied 5 models out of 2 models gives best result

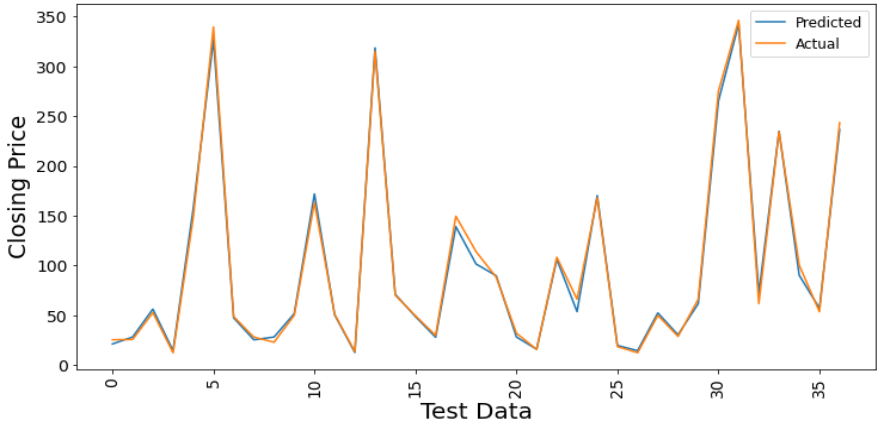
4. Random Forest

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting



5. XG boost

It is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. XG Boost is basically designed to enhance the performance and speed of a Machine Learning model.



- In random forest and XG boost accuracy is higher than other models.
- Prediction values approximately equal to actual values.

Models and their performance metrics

Random forest and XG boost model gives better R2 scores and least errors.

Model name	Train score	Test score	MSE	RMSE	MAE	MAPE	R2 score
Linear regression	0.815	0.823	0.032	0.178	0.151	0.095	0.823
Lasso regression	0.815	0.821	0.032	0.179	0.152	0.096	0.821
Ridge regression	0.815	0.822	0.032	0.178	0.151	0.095	0.822
Random forest	0.998	0.992	0.001	0.037	0.029	0.018	0.992
XG boost	0.999	0.991	0.002	0.039	0.030	0.020	0.991

Conclusion

- Features are multicollinear but can not drop the column because features are limited.
- High, low, open are directly correlated with the closing price of stocks.
- The test results of all the regression models are evaluated and compared. We checked performance metrics such as R2 score, Mean Square Error, and Root Mean Square Error etc.
- In linear, lasso and ridge accuracy are approximately equal even after applying cross validation.
- Results are not up to the mark with linear regression, ridge and lasso regression.
- Other models such as random forest and XG boost. With the help of this model we got better R2 scores and metrics.
- Out of all the model random forest and XG boost gives best result.

THANK YOU