A Supervised Machine learning Regression model on

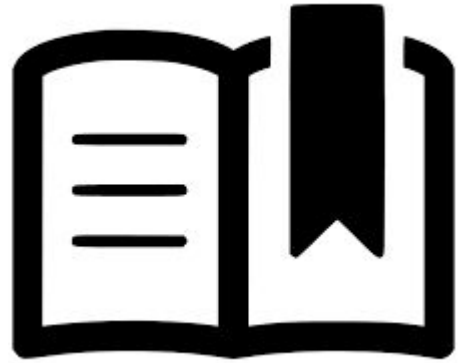
Predicting Yes Bank Closing Price

Capstone Project by :

Sanjeev Hegde

Data Science Trainee, AlmaBetter

Table of Contents:



Why Analyze Stock Closing Price..?

Problem Statement

Predicting Stock closing price is critical for making trade decisions. However, stock prices are governed by numerous factors such as

- Company financials
- News
- Global Market outlook
- Stock specific news
- Investor sentiment

All these factors impact the open, high, low and close of the stock

Let us build a model using machine learning to predict Yes Bank stock closing price in the upcoming sections

Introduction

Yes Bank is one of the prominent banks in the Indian Banking Sector. It was founded in the year 2004 and listed on exchange in 2005. It has witnessed highest profit of Rs.4,233 crore during FY 2018. However, in the year 2018, a fraud case was registered against Rana Kapoor who was the founder and then MD, CEO of the Yes Bank. Since then stock price has dropped significantly from its high.

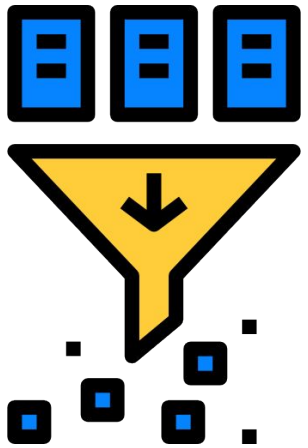
Yes Bank Data set

The Yes bank data set consists of 5 parameters:

- Date: The month against which Open, High, Low and Close prices of Yes Bank stock are plotted.
- Open: Opening price of the stock on the specified date. In this case, start of the month
- High: Highest price attained by the Yes Bank stock during the monthly period
- Low: Lowest price attained by the Yes Bank stock during the monthly period
- Close: Closing price of the Yes Bank stock at the end of the month and is also the dependent feature that needs to be predicted

Data Cleaning:

- There are total 185 entries in the data set starting from June 2005 to November 2020
- There are no null items in the data set
- All the dates to which data is given are unique, no duplicate date
- Date feature is assigned with ‘datetime64’ data type

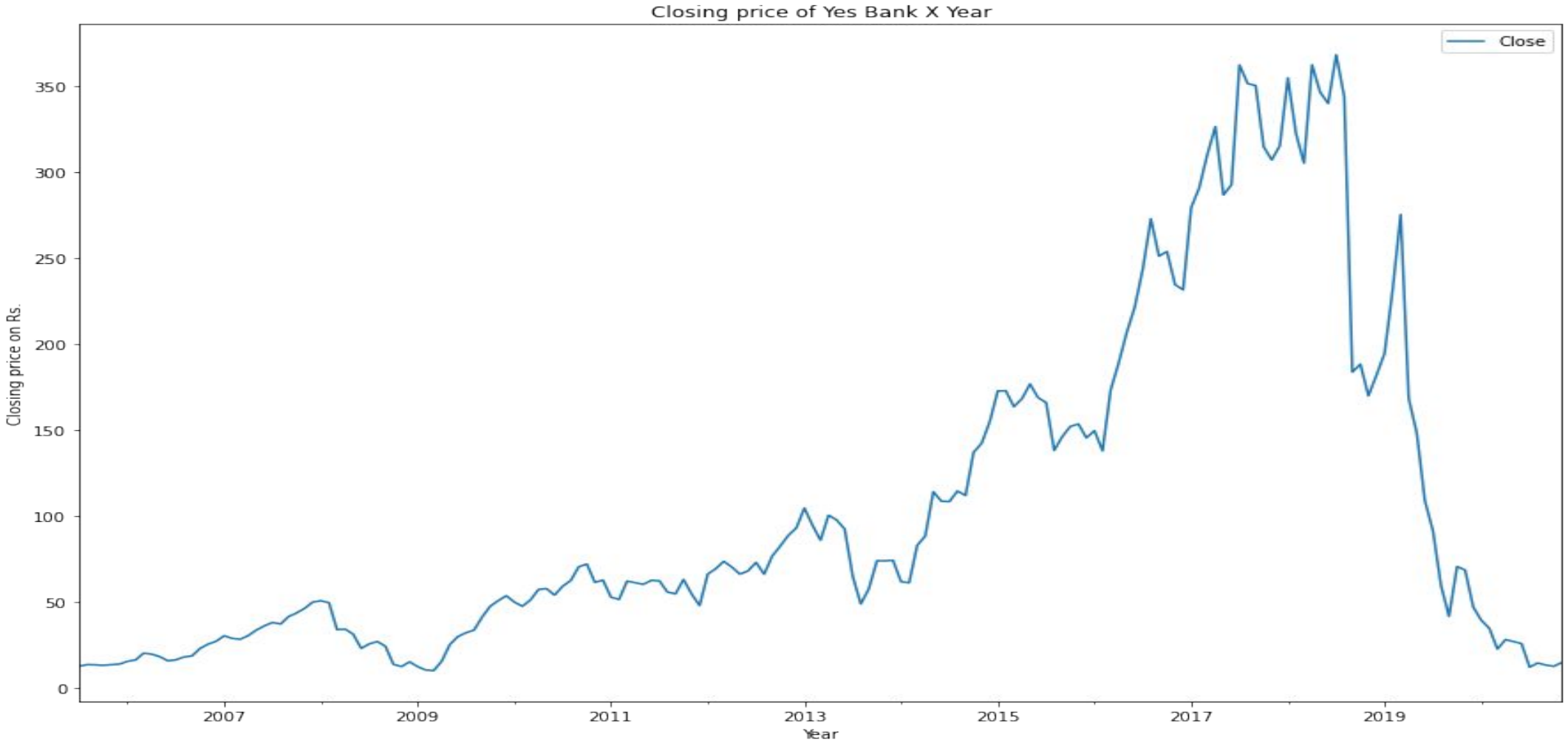


Statistical Overview:

- Highest price recorded by Yes Bank is Rs. 404
- Lowest price recorded by Yes Bank is Rs. 5.55
- Mean closing price of Yes Bank is Rs. 105.2
- Standard deviation about mean for closing price is Rs. 98.58 indicating violent moment in the stock
- Lowest opening recorded by Yes Bank is Rs. 10
- Lowest closing recorded by Yes Bank is Rs. 9.98

Data Visualization:

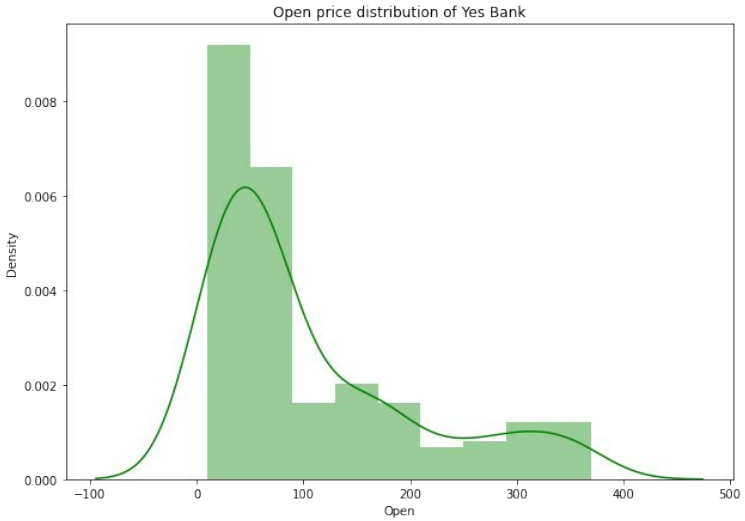
- A plot consisting of closing price of Yes Bank is plotted against timeline to understand the price moment and general trend. It can be easily figured out that Yes Bank shares plunged after fraud case against Rana Kapoor in 2018

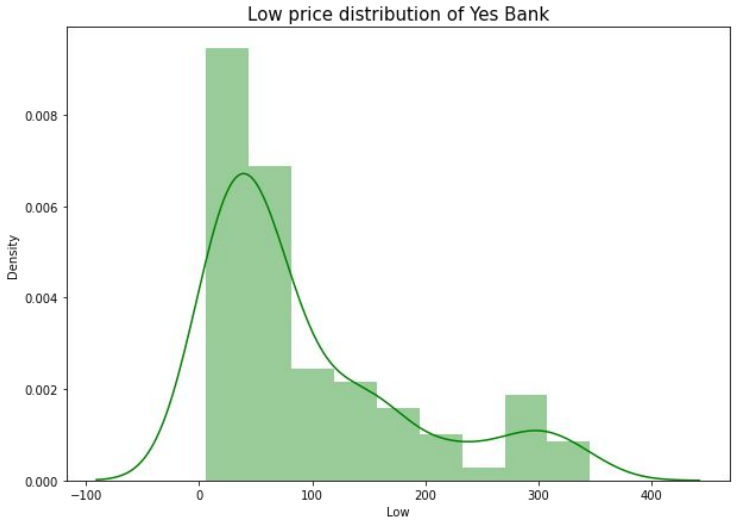


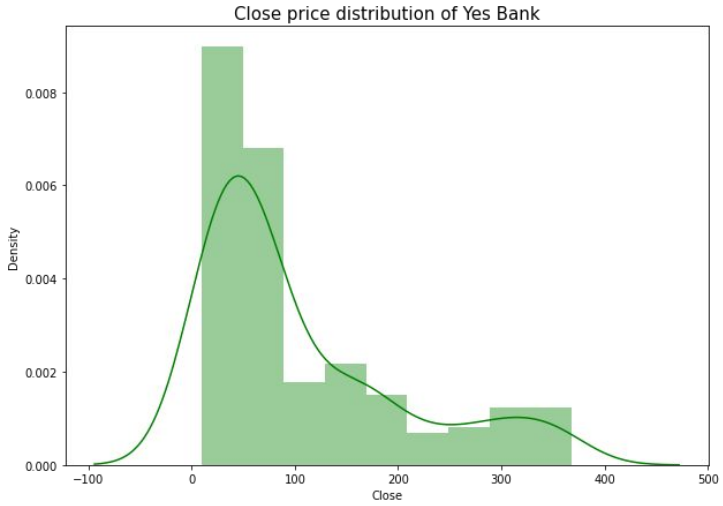
Data Visualization:

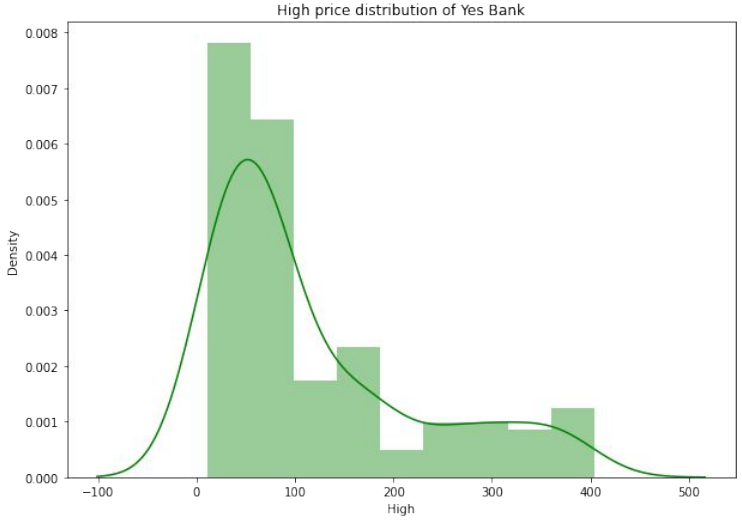
- **Skewness**

  Distribution of various features are plotted in histogram style to check for skewness in the data set and found to be right skewed





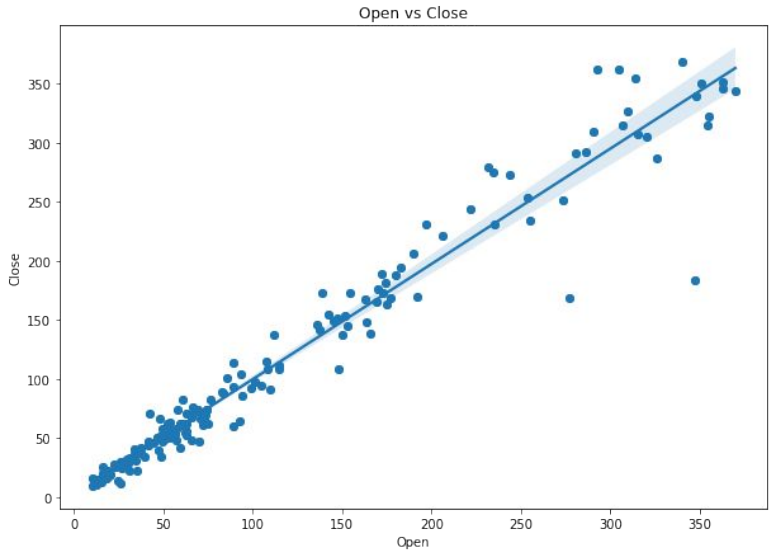


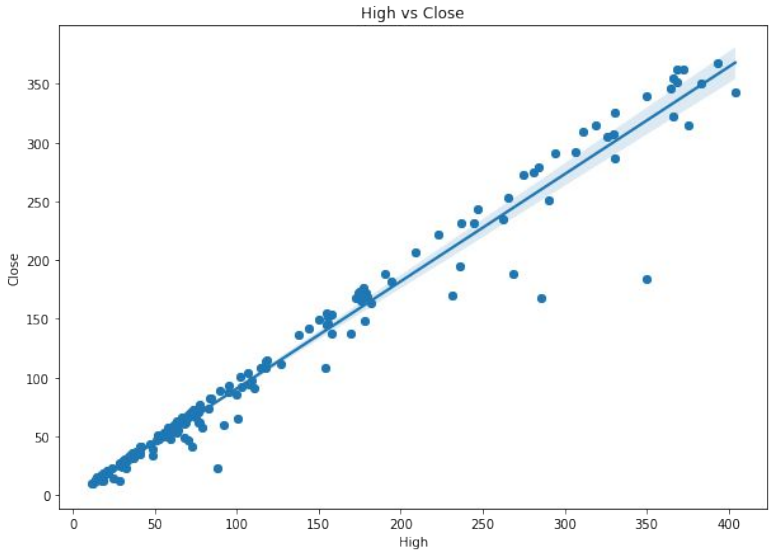


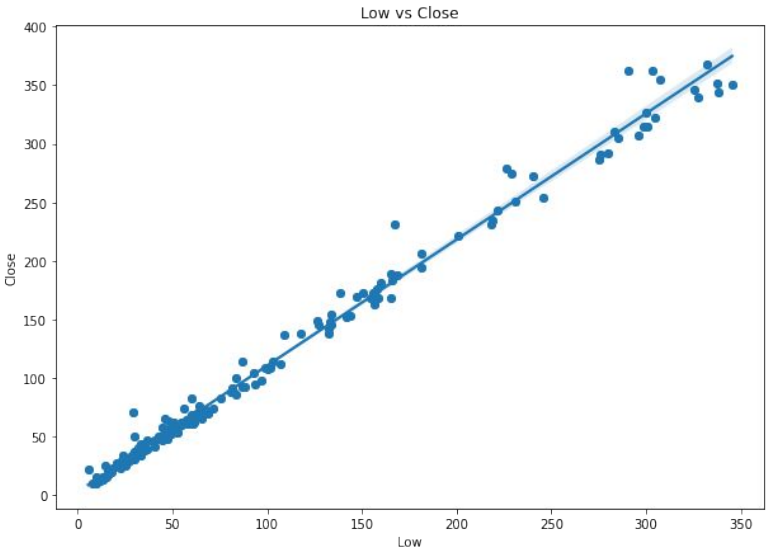
Data Visualization:

- **Relationship between Closing price and other independent features**

  Relationship between closing price and other independent variables is plotted along data distribution points. It is found that variables are linearly related.



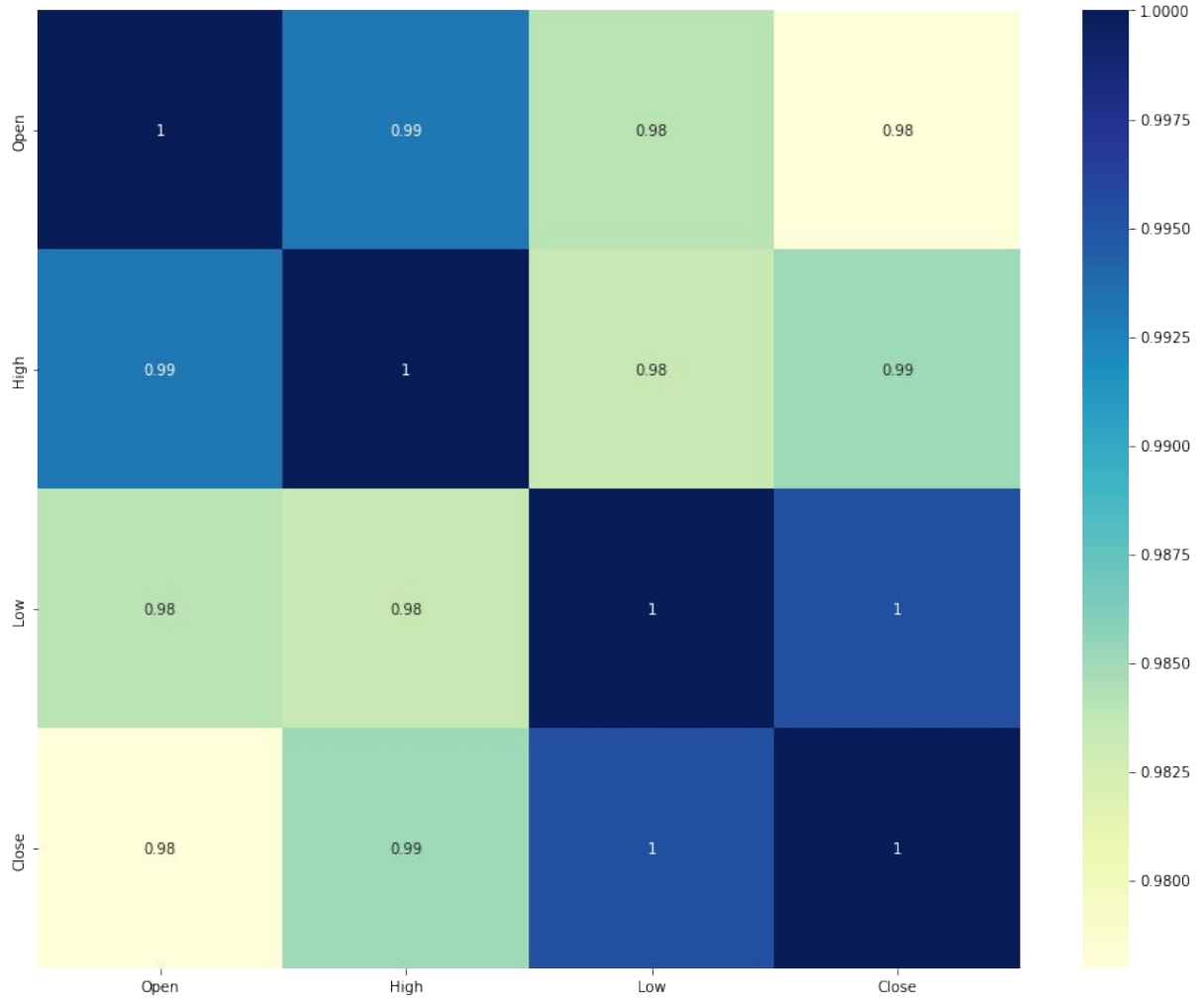




Data Visualization:

- **Correlation Heatmap**

  A Correlation Heatmap is plotted to check for relation between variables and are found to be highly correlated



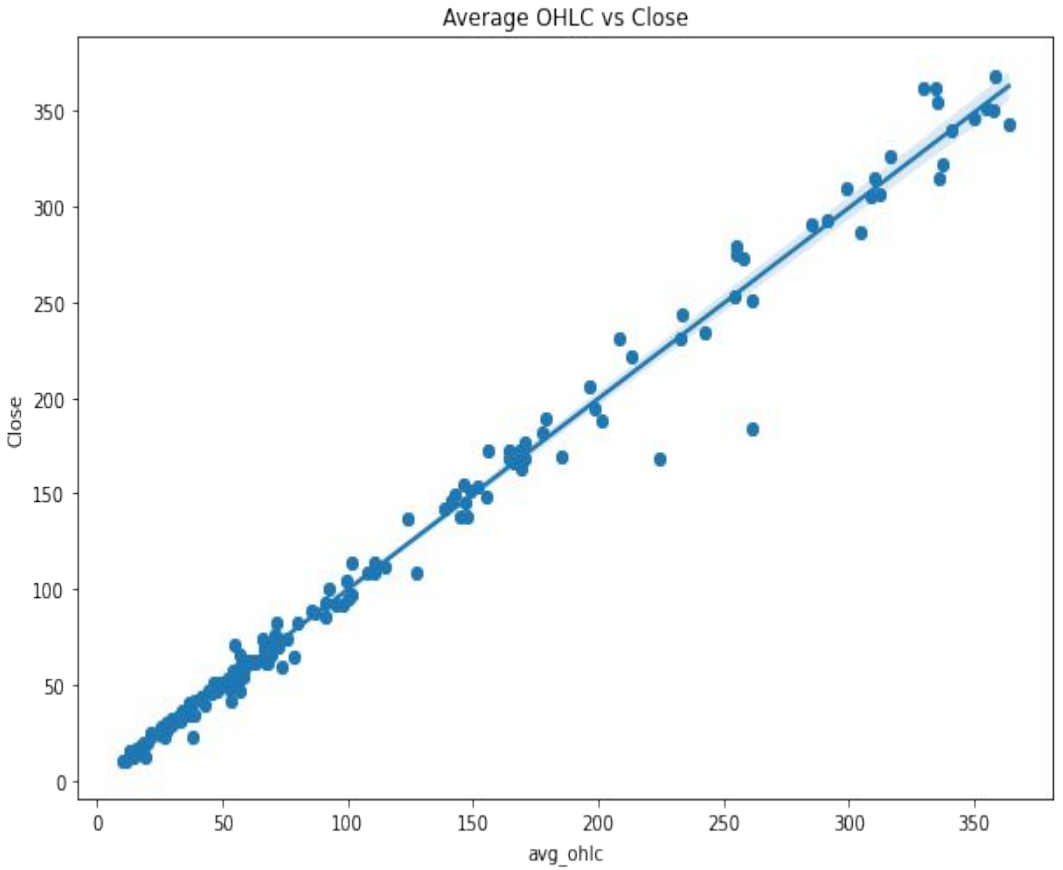
Feature Engineering:

- **New feature- average of OHLC, lags based on the average of OHLC**

  Stock market analysis techniques use the previous day's average of Open, High, Low and Close as pivots signifying the importance of previous day price of the stock. Further, technical analysis indicators such as moving averages, Relative Strength Index etc use OHLC data of previous 14 sessions to generate critical trade signals.

  Hence, knowing the importance of previous sessions' OHLC data, it is fair to incorporate lags that represent the average of previous 14 sessions' OHLC data.

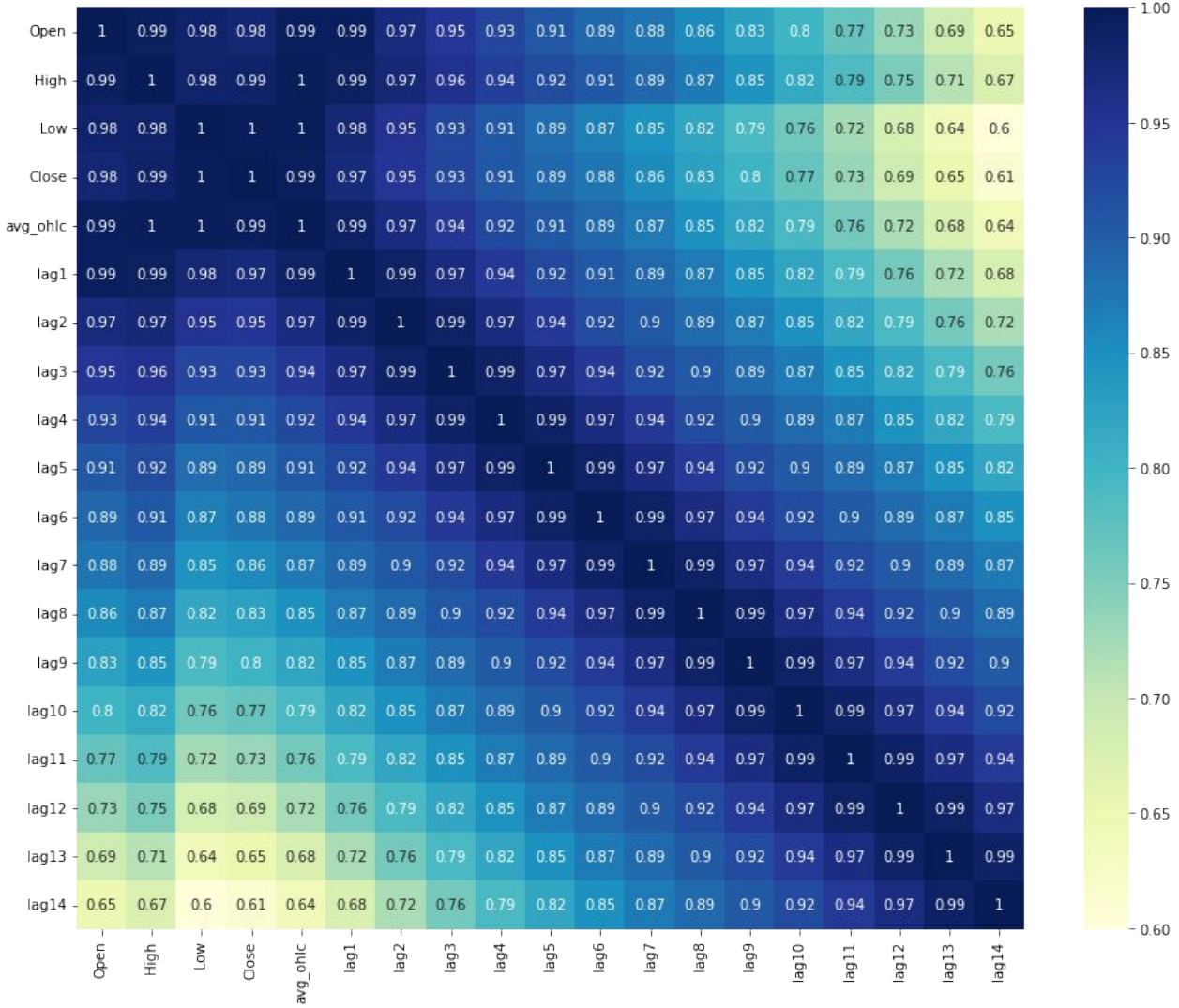
  Further, it is clear that similar to other independent variables, average of OHLC (avg_ohlc) exhibits linear relationship with closing price of the stock.



Feature Engineering:

- **Correlation Heatmap after incorporating lags**

  A Correlation Heatmap is plotted again. It is clear that there is a drop in the multicollinearity as compared to considering only Open, High, Low and Close



Splitting and Scaling Data:

- Independent variable is considered as monthly closing price
- Any entry having NaN after introducing Lags are dropped from the data set
- Lags are considered as independent variables representing previous 14 day OHLC data
- Standard scaler is used to scale the data
- Training data set size is set to be 0.7

Modularity in data handling:

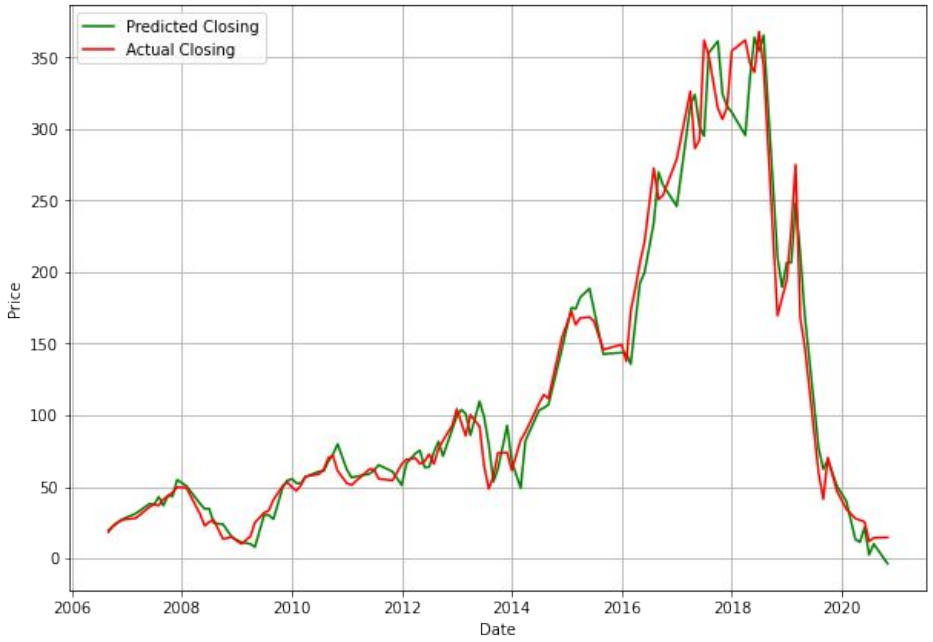
- Two different functions have been created for plotting graphs of stock price and evaluation metrics scores.
- Plotter: This function takes two series of price as input and plots line graph with both data in the same plane against another common variable (Timeframe in this case).
- Eval_metrics: This function takes two series of price values as input and prints evaluation metric scores after fitting machine learning models.
- Evaluation metrics used are: R2 score, Mean Squared Error(MSE), Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), Mean Absolute Percentage Error(MAPE)

Linear Regression:

Training Data Fit

- R2_score is: 0.9699
- MSE value is: 313.0597
- RMSE value is: 17.6935
- MAE value is: 11.9505
- MAPE value is: 0.1503

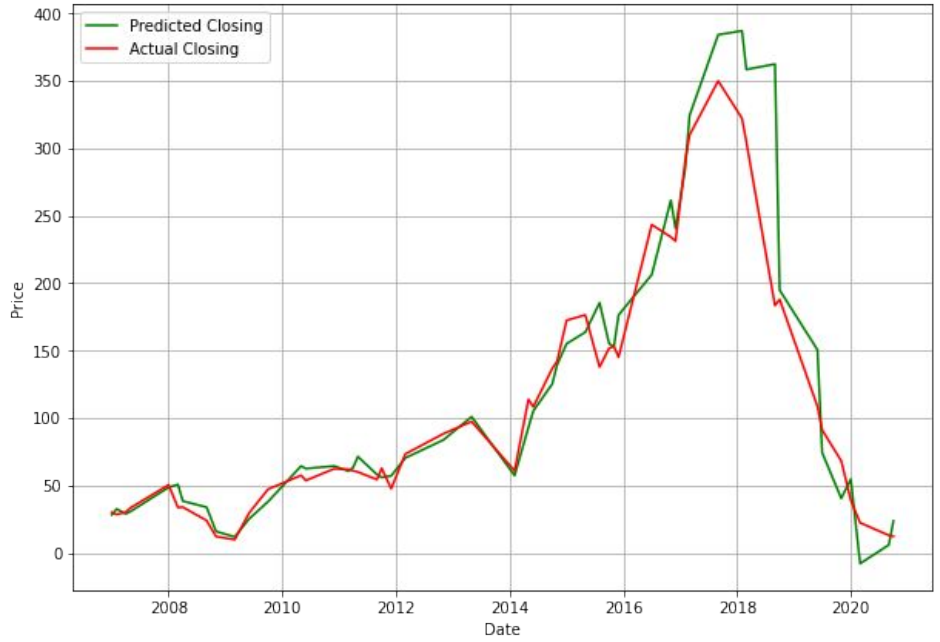
Predicted VS Actual Closing price



Testing Data fit

- R2_score is: 0.878
- MSE value is: 1005.4573
- RMSE value is: 31.7089
- MAE value is: 16.6157
- MAPE value is: 0.2019

Predicted VS Actual Closing price



Ridge Regression with 10 fold Cross Validation:

Training Data Fit

- r2_score is: 0.9614
- MSE value is: 401.472
- RMSE value is: 20.0368
- MAE value is: 13.5939
- MAPE value is: 0.1634

Alpha values:

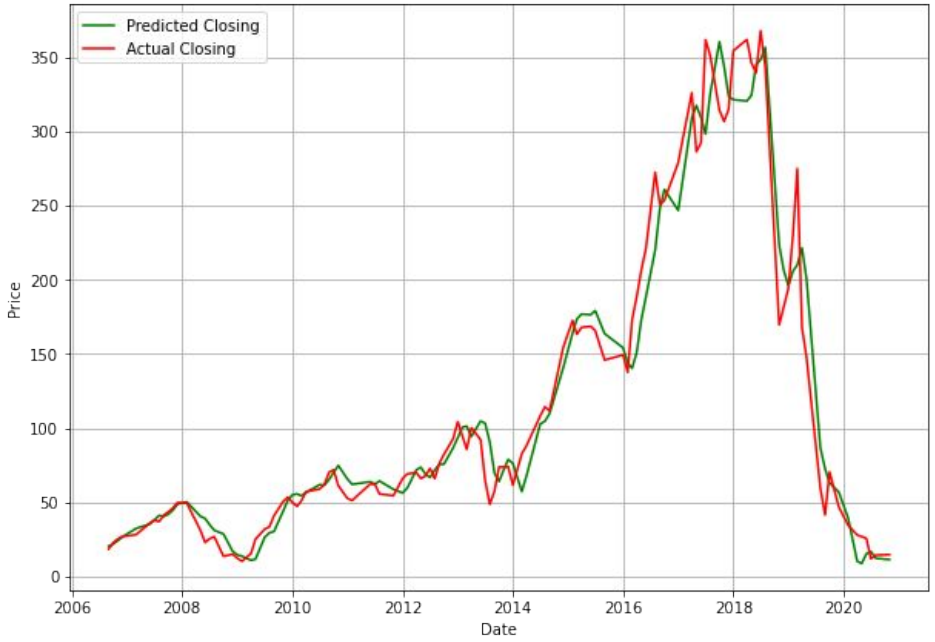
1e-15, 1e-13, 1e-10, 1e-8, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 2, 3, 5, 10, 20, 30, 40, 45, 50, 55, 60, 100

Best value: 2

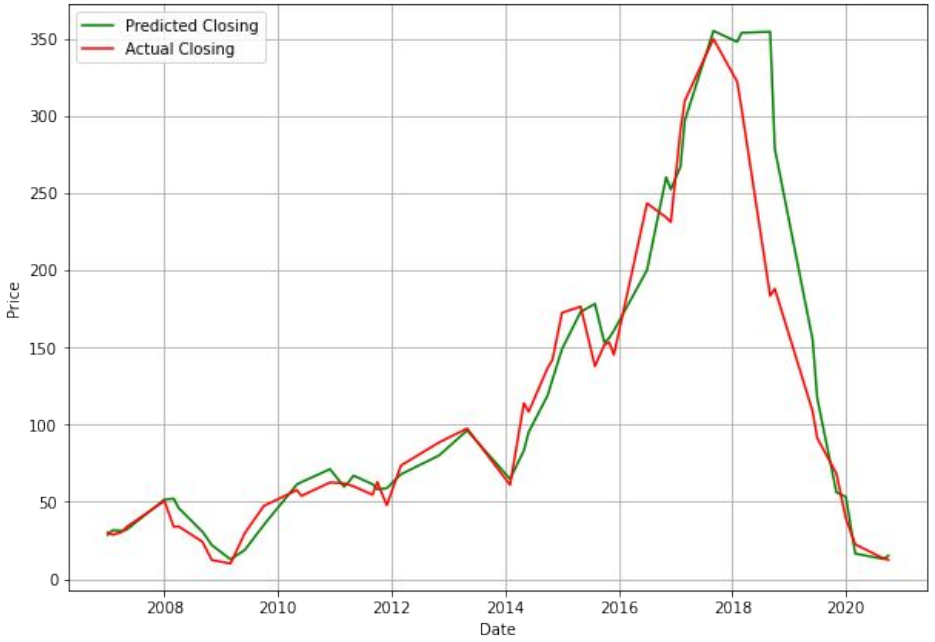
Testing Data fit

- r2_score is: 0.8762
- MSE value is: 1019.8925
- RMSE value is: 31.9358
- MAE value is: 17.1518
- MAPE value is: 0.1817

Predicted VS Actual Closing price



Predicted VS Actual Closing price



Lasso Regression with 10 fold Cross Validation:

Training Data Fit

- r2_score is: 0.9568
- MSE value is: 448.9065
- RMSE value is: 21.1874
- MAE value is: 13.295
- MAPE value is: 0.1641

Alpha values:

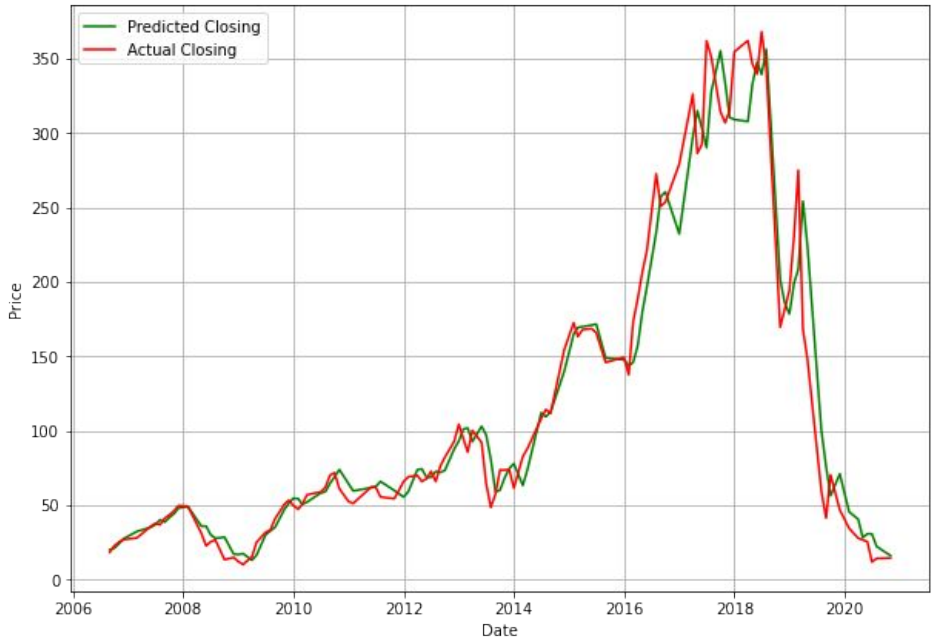
1e-15, 1e-13, 1e-10, 1e-8, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 2, 3, 5, 10, 20, 30, 40, 45, 50, 55, 60, 100

Best value: 3

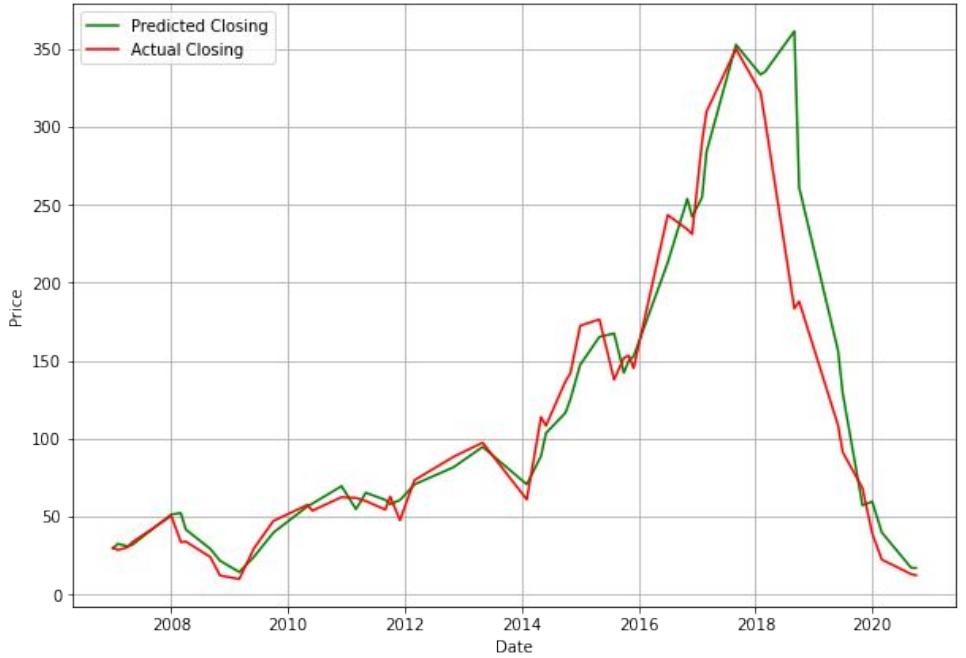
Testing Data fit

- r2_score is: 0.8825
- MSE value is: 967.9291
- RMSE value is: 31.1116
- MAE value is: 16.3884
- MAPE value is: 0.1933

Predicted VS Actual Closing price



Predicted VS Actual Closing price

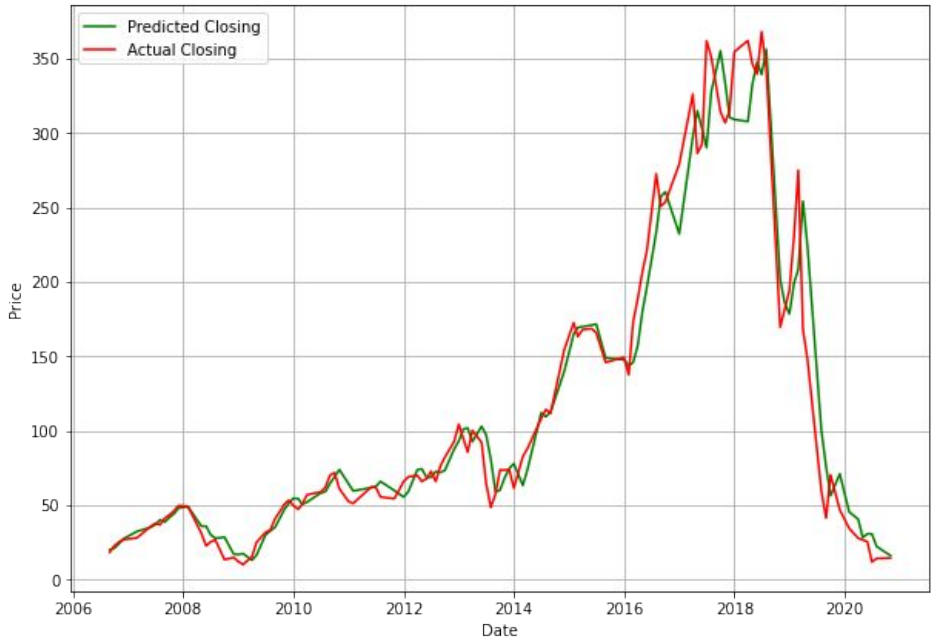


Elastic Net Regression with 10 fold Cross Validation:

Training Data Fit

- r2_score is: 0.9568
- MSE value is: 448.9065
- RMSE value is: 21.1874
- MAE value is: 13.295
- MAPE value is: 0.1641

Predicted VS Actual Closing price



Alpha values:

[1e-15,1e-13,1e-10,1e-8,1e-5,1e-4,1e-3,1e-2,1e-1,1,2,3,5,10,20,30,40,45,50,55,60,100]

L1 ratio:

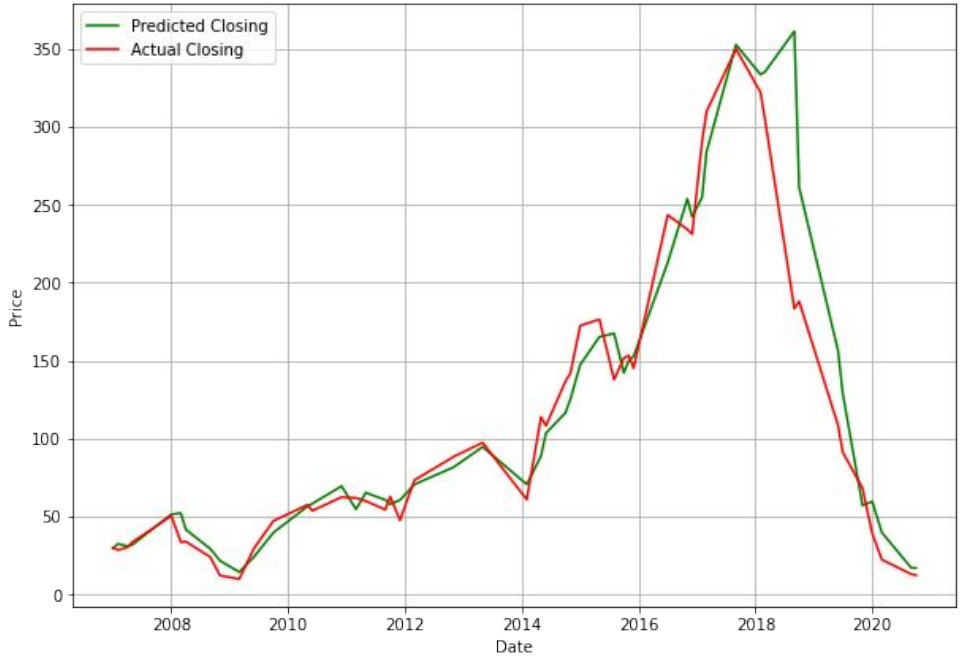
0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1,1.5,2

Best value alpha, L1 ratio: 3, 1

Testing Data fit

- r2_score is: 0.8825
- MSE value is: 967.9291
- RMSE value is: 31.1116
- MAE value is: 16.3884
- MAPE value is: 0.1933

Predicted VS Actual Closing price



XGBoost model with 10 fold Cross Validation:

Number of estimators:
400, 700, 1000
Max tree depth:
2,3,5,7,10
Regularization alpha:
1.1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 50

Regularization Lambda:
1.1, 1.2, 1.3
Subsample:
0.7, 0.8, 0.9
colsample_bytree:
0.3, 0.5, 0.7, 0.9, 1

XGBoost

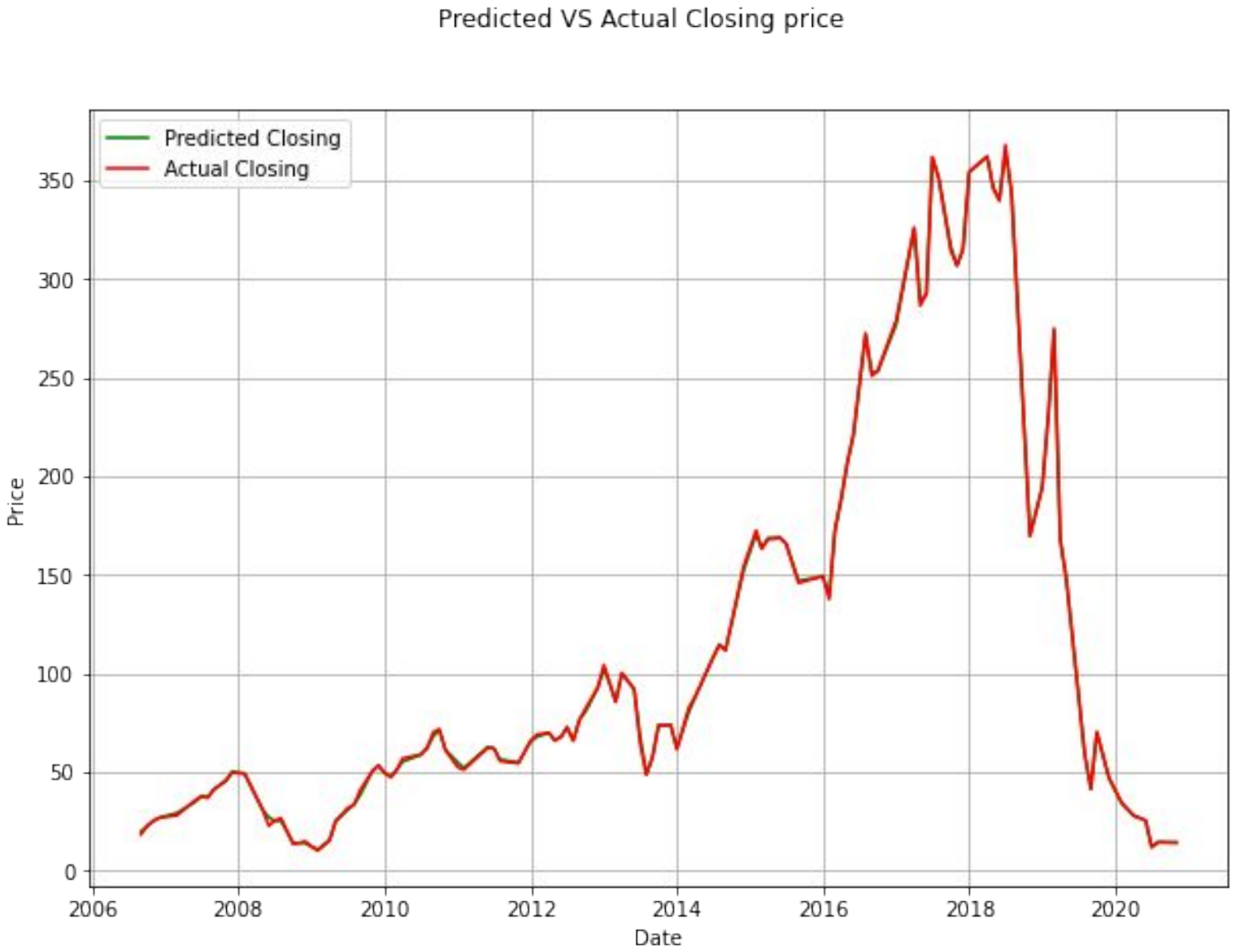
Best Values for:
Number of Estimators: **1000**
Max tree depth: **3**
regularization alpha: **7**
regularization lambda: **1.1**
subsample: **0.9**
colsample by tree: **1**

YES BANK

XGBoost model- Training Data:

Training Data Fit

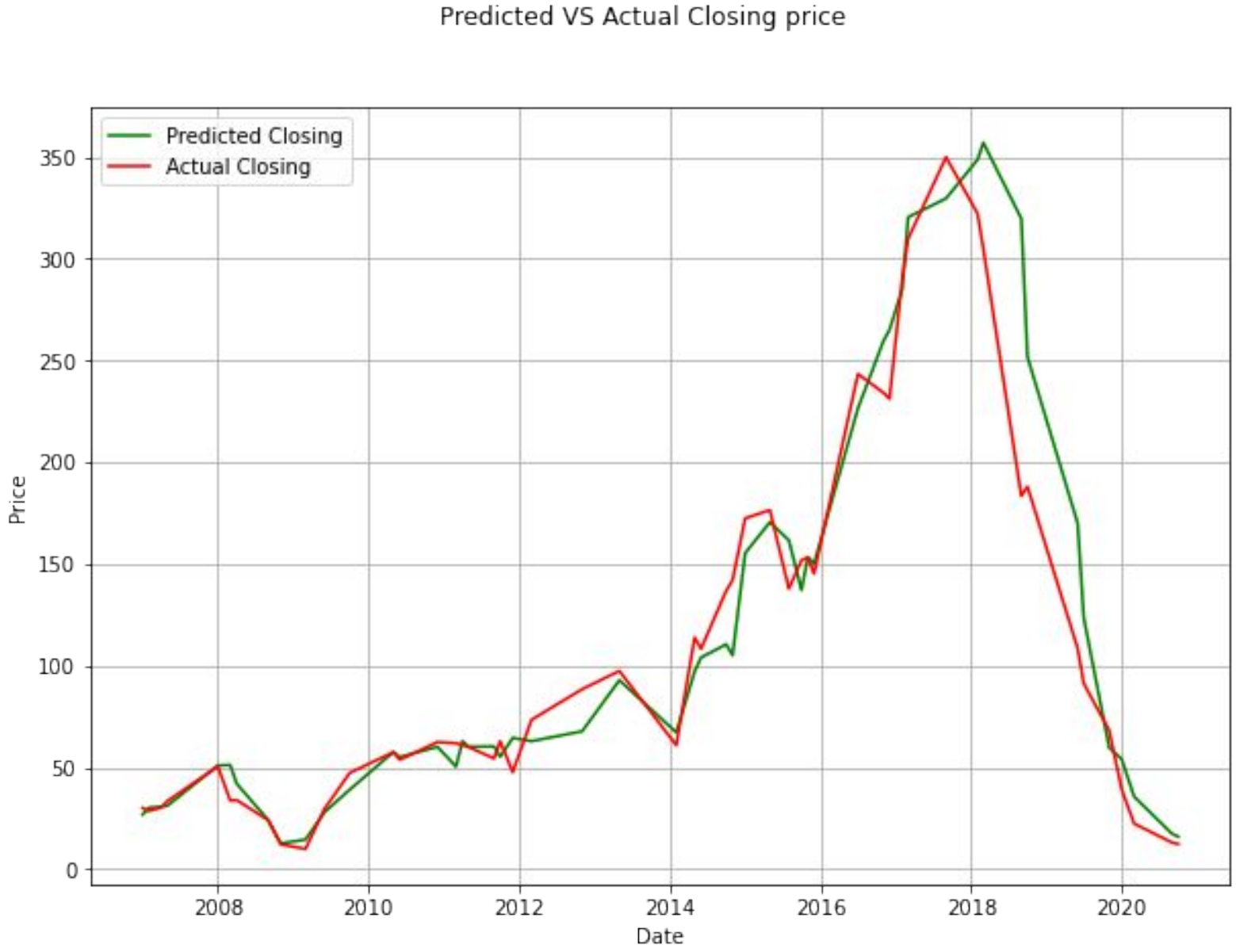
- r2_score is: 0.9999
- MSE value is: 1.0735
- RMSE value is: 1.0361
- MAE value is: 0.6516
- MAPE value is: 0.0118



XGBoost model- Testing Data:

Testing Data Fit

- r2_score is: 0.9087
- MSE value is: 751.9086
- RMSE value is: 27.421
- MAE value is: 15.7395
- MAPE value is: 0.1657



Summary and Conclusion

- With the given data at hand, an inspection of the same has been carried out visually and statistically for checking presence of any abnormalities in the data set.

- Average price of the Open, High, Low and Close were calculated and considered as independent variable to reduce multicollinearity.

- Further, Lags were introduced into the data set thereby essentially considering happenings of previous 14 sessions in the market and further eliminating multicollinearity.

- Simple Linear Regression model, Ridge and Lasso Regression models, XGBoost regression model were built based on training data set from the original data set and predictions were made for test data set along with evaluation metrics scores using previous 14 period lags.

- Hyperparameter tuning with 10 fold cross validation was performed on each regression model for optimizing loss function thereby successfully predicting closing prices for test data sets.

- It is found that XGBoost model was best in predicting closing prices for both training and testing data set with **0.9087 R2 score for test data**

Thank you