Yes Bank Closing Price Prediction

Sanjeev Hegde
Data science Trainee, AlmaBetter

***Abstract*:** Stock market is a legal platform for companies to raise funds for their operations from the public and investors by providing ownership in the company through company shares on Stock Exchanges. Each day millions of shares of hundreds of companies are traded between buyers and sellers on these exchanges driving the price of the stock higher or lower based on the prevailing supply or demand for the corresponding company shares. Bombay Stock Exchange(BSE) and National Stock Exchange(NSE) are the two market platforms in India wherein anyone can invest or trade contracts in the listed companies. Thus, to an investor or to a trader, predicting stock closing prices is of paramount importance in winning the trades or making significant gains from his investments. In this regard, an attempt to predict the closing price of Yes Bank stock has been successfully made using supervised machine learning. Regression models such as Linear regression, Regularised Linear Regression including Lasso, Ridge and ElasticNet, XGBoost Regression have been performed to predict the monthly closing price of Yes Bank. Amongst the various regression models, it was found that XGBoost was best performing with 0.9087 R2 score.

Keywords: Stock Exchange, Yes Bank, Exploratory Data Analysis (EDA), Feature Engineering, OHLC (Open, High, Low, Close), Linear Regression, Regularization, Lasso, Ridge, Elastic Net, Decision Trees, XGBoost, Data Visualization, Financial Year (FY), investment, trading, derivatives(financial), All Time High(ATH)

***Problem Statement*:** Predicting monthly closing price of Yes bank stock using historical OHLC data and regression models.

***Introduction*:** Yes Bank is one of the prominent banks in the Indian Banking sector. It was founded by Rana Kapoor and Ashok Kapoor in 2004 and got listed on the Indian Stock Exchanges in 2005. During its peak, Yes bank achieved profit of Rs. 4,233 crores in FY2018. However, later during the same calendar year a fraud case was registered against its Founder and MD cum CEO, Mr. Rana Kapoor. Following the charges against its founder, Yes bank share price has tanked more than 90 percent from its all time highs. Irrespective of the strategies followed by various investors and traders, such an event can never be foreseen by anyone except insiders. Further, factors such as the company's financials, global developments, global market outlook, stock specific events affect the price of the stock significantly on each trading day. All the movement in the stock price is therefore encompassed in its Open, High, Low and Closing price(OHLC). However, a sound regression model can successfully predict the due course of the share price considering the historical OHLC data. In this regard, as the objective of the project, various regression models have been developed for predicting the monthly closing price of the Yes bank stock considering 14 month historical OHLC data from the given data set. Further, the importance of opening price, high, low and closing price of any liquid financial instrument is clearly evident from the developed machine learning models.

1. ***Yes Bank Closing Price data set:*** The Data set contains following features:

- **Date:** The month against which OHLC price is given.
- **Open:** Opening price of Yes Bank on specified Date.
- **High:** Highest price achieved by the stock during the month.
- **Low:** Lowest price attained by the stock during the month.
- **Close:** Closing price of the stock during the end of the month.

2. ***Data Cleaning:*** Given Yes Bank Closing price data set was checked for abnormalities and following observations and modifications were made to the data set:

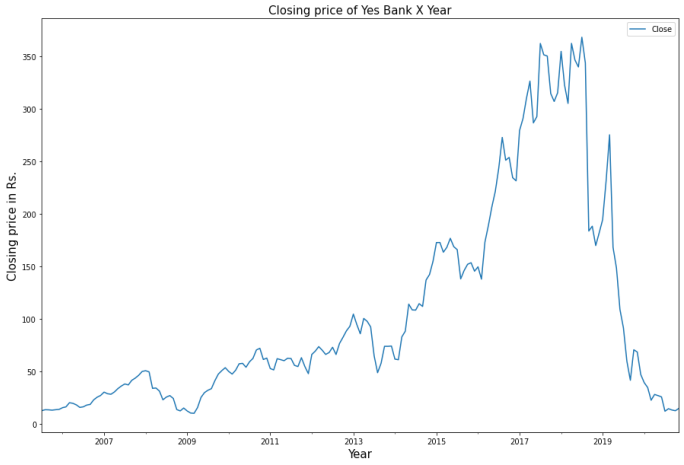
- There are no null values in the data set
- There are no duplicate entries in the data set
- For OHLC features, data type is float64 which is appropriate for further analysis
- Date feature data type has been converted to Datetime64 from object type
- There are a total of 185 entries in the data set to each feature.

3. ***Statistical Overview:*** Following observations were made regarding the features in the data during the period of collection of data i.e., till November 2020:

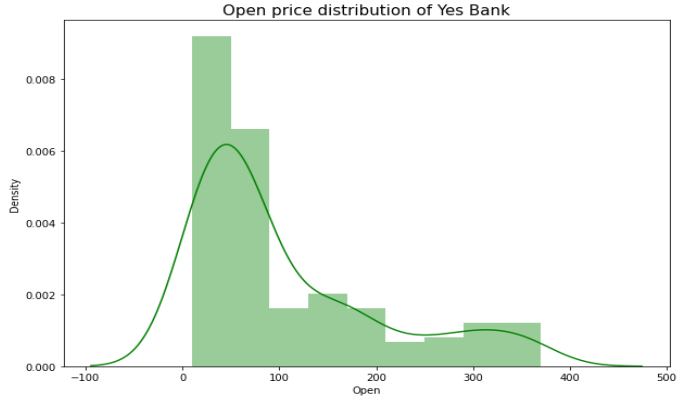
- Starting month from when OHLC data of Yes bank was recorded in the data set is July 2005.
- Highest price recorded by Yes Bank share is Rs. 404, while the lowest price is Rs. 5.55
- Mean closing price of Yes bank is Rs.105.2
- Standard deviation about mean for closing price is Rs. 105.2 indicating violent movement in the stock
- Lowest opening recorded by Yes bank is Rs. 10
- Lowest closing recorded by Yes Bank is Rs. 9.98

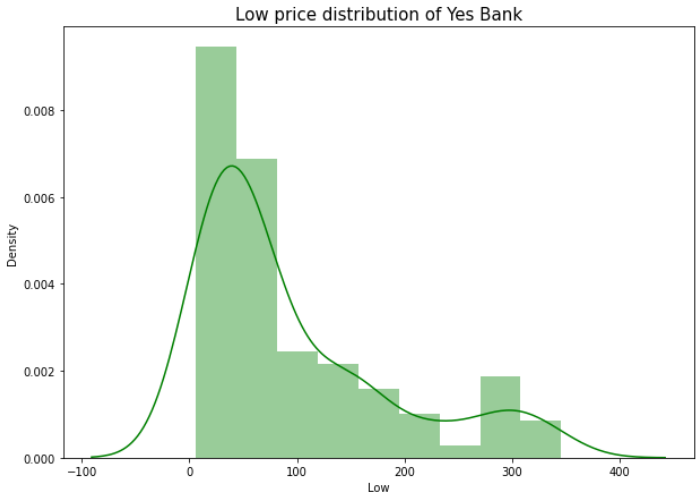
4. ***Exploratory Data Analysis:*** Visual representation of data is critical for understanding the intricacies of the data such as general trend in the data, inter feature relationships, skewness in the data etc. Same can be done using Matplotlib and Seaborn libraries on python. The following visual representations have been done for understanding the Yes Bank closing price data set better.

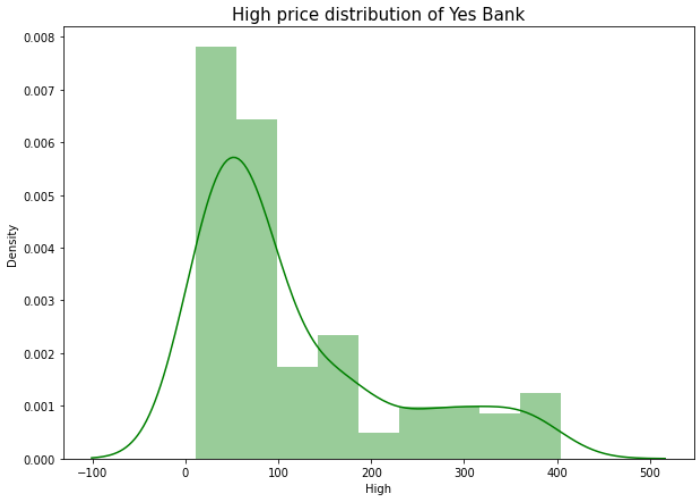
- **Yes Bank closing price plot against date:** The following visualization helps us understand the distribution of the Yes Bank share closing price during July 2005 to November 2020. It can be easily figured out that Yes Bank share price plunged heavily after the Rana Kapoor fraud case in 2018.

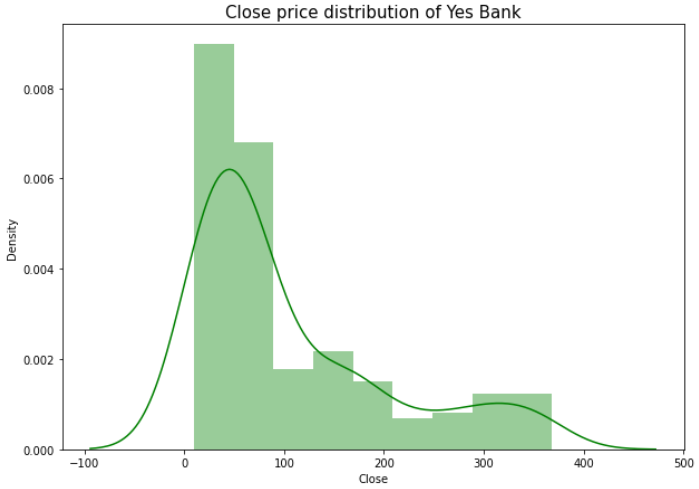


- **Skewness:** Distribution of various features are plotted in histogram style to check for skewness in the data..



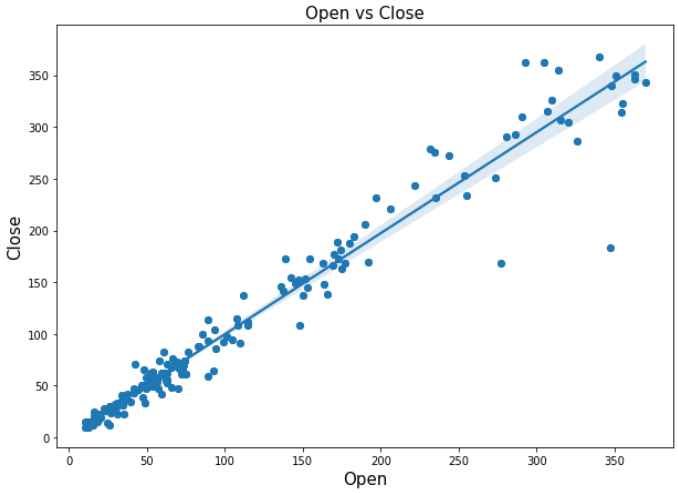


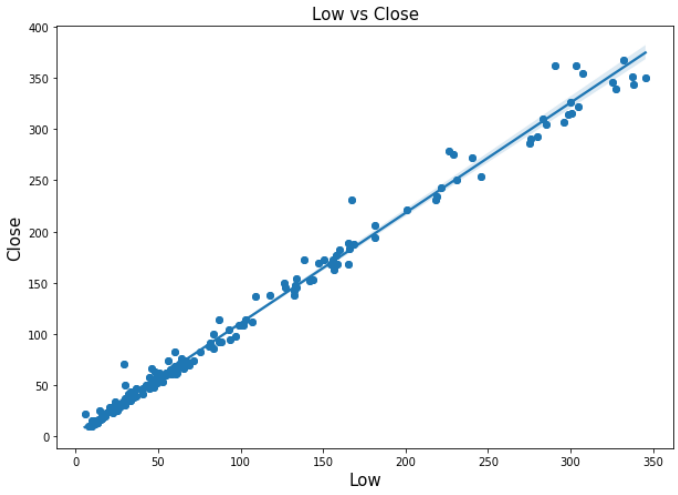


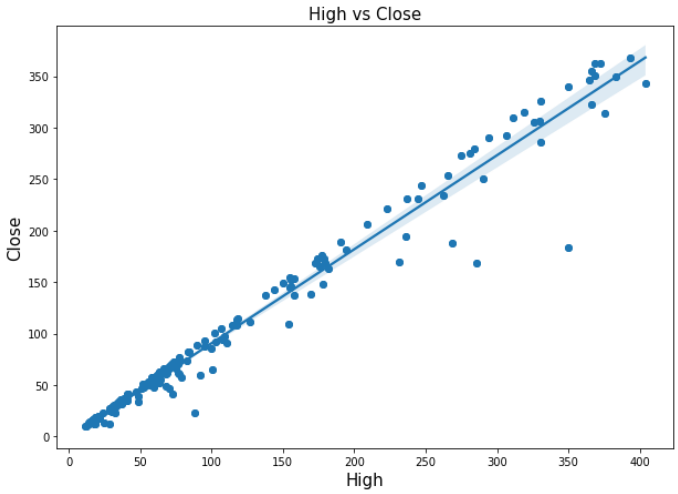


Thus, it can be seen that all the features in the data are right skewed (positively skewed) wherein mode> median>mean.

- **Closing Vs Other features:** All the numeric independent features i.e., Open, High and Low are plotted against Closing price to see the mathematical relationship between them.

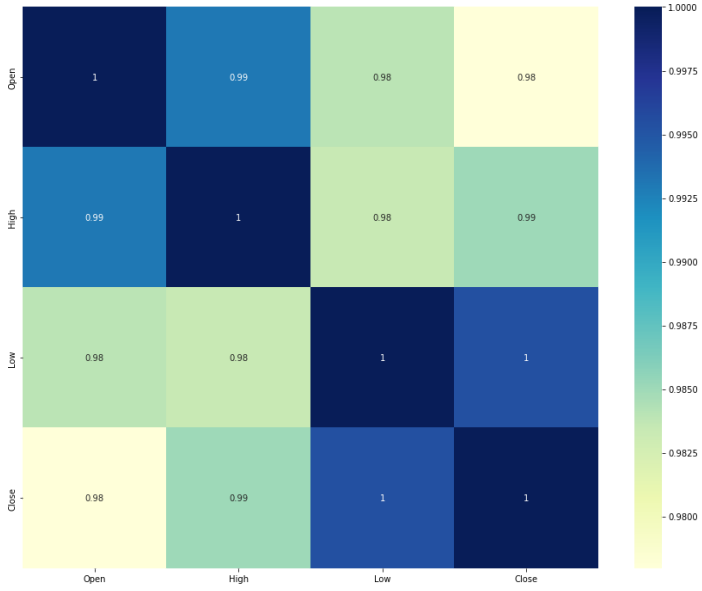






Hence, it can be seen that all the numeric independent features exhibit a linear relationship with the closing price.

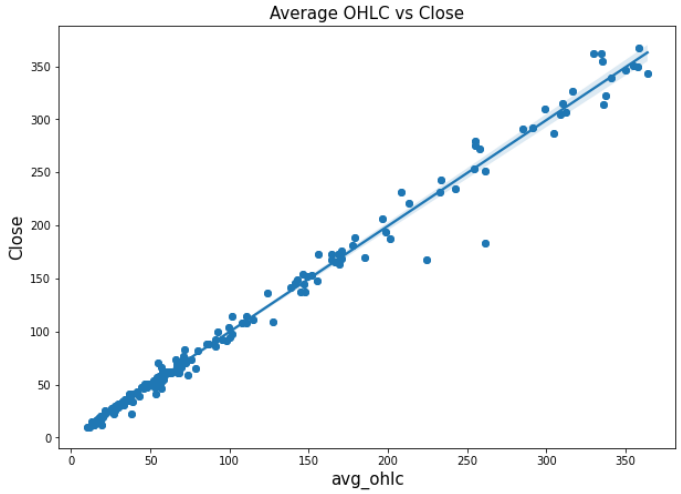
- **Correlation Heatmap:** A Correlation heatmap has been plotted between numeric features in the data to check for the extent of correlation.



Hence, it is evident that features in the data exhibit a high degree of correlation.

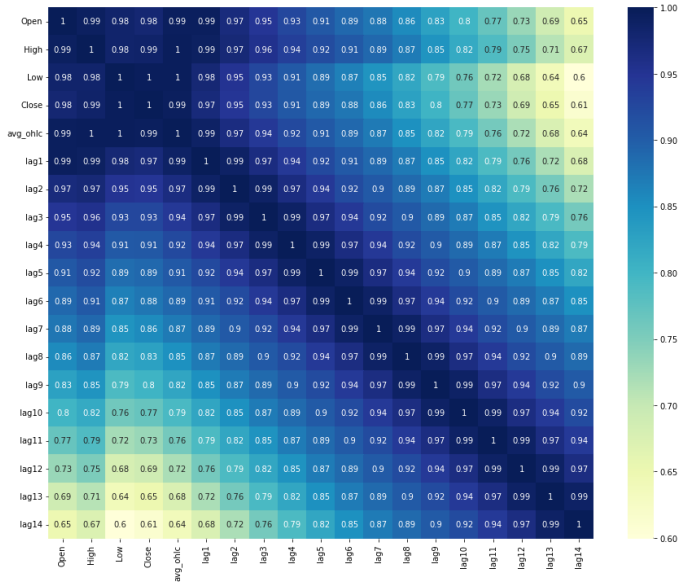
***5. Feature Engineering*:** An average of previous session's(Monthly in case of present Yes bank data set) OHLC values was added to the list of features(avg_ohlc). This enables us to consider the effect of the previous trading session, thereby eliminating the potential problems which could arise due to high correlation between the features. Further, such average values have been calculated for previous 14 sessions and introduced into the data set as lags, thereby reducing the multicollinearity even better.

- **Closing Vs Average OHLC of previous session:** A scatter plot has been plotted along with linear regression fit to visualize the mathematical relationship between above parameters.



It is found that the new feature i.e., average of OHLC values of the previous session(avg_ohlc) exhibits a linear relationship with closing price similar to Open, High and Low.

- **Correlation heatmap (lags):** A correlation heatmap has been plotted for lags introduced i.e., average values of OHLC for previous 14 sessions. It is found that multicollinearity is significantly lower between the lags.



Hence, it is ideal to consider the aforementioned lags as the appropriate features in the data set for building the machine learning models.

***6. Data Splitting and Scaling*:** Following steps were performed while splitting the data. Also, the standard scaler was used to scale the data.

- Monthly closing price is considered as the independent variable.
- All the entries with NaN values after introduction of lags were dropped from the data set.
- Lags were considered as the independent variables for performing the analysis.
- Training data size was 0.7 times the overall data

***7. Modularity in handling the data*:** Two different functions were built for better handling data in modular ways.

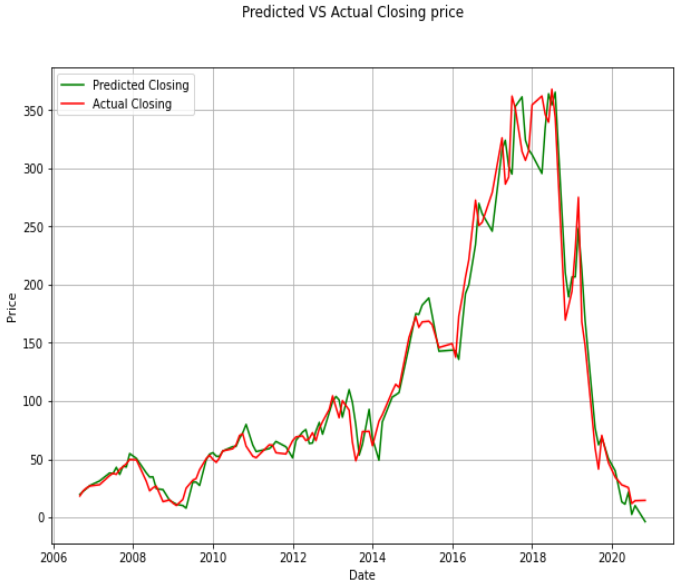
- **Plotter:** This function two series of price as the input arguments and plots line graph of both data in the same plane against another common variable(Timeframe of Yes bank price in this case)
- **Eval_metrics:** This function takes observed values and predicted values as input arguments and prints various evaluation matric scores for any regression model.
- R2 score, Mean Squared Error(MSE), Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), Mean Absolute Percentage Error(MAPE) are the various evaluation metric scores used in the Eval_metric function.

***8. Regression Models*:** Following Regression models were used with 10 fold cross validation for predicting the closing price of the Yes Bank:

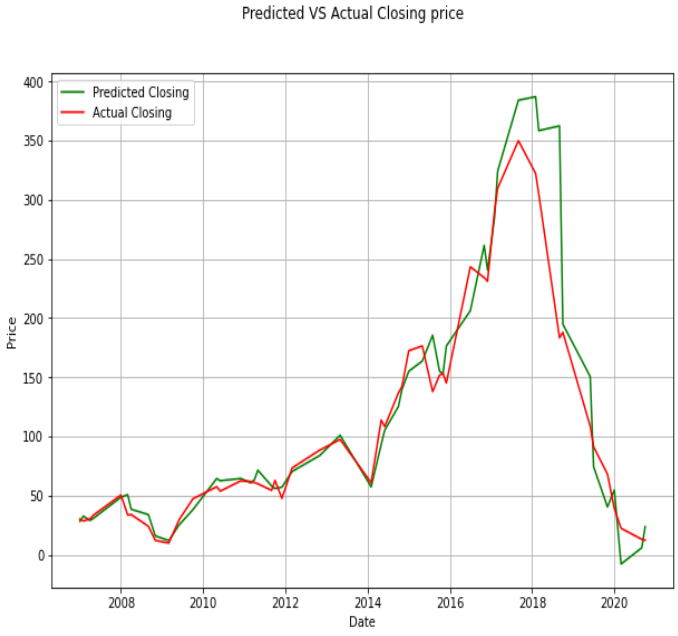
- **Linear Regression:** A linear regression was fit into the data with following scores and plots for training and testing data correspondingly.

Evaluation Metric Score	Training Data	Testing Data
R2_score	0.9699	0.878
MSE	313.0597	1005.4573
RMSE	17.6935	31.7089
MAE	11.9505	16.6157
MAPE	0.1503	0.2019

Predicted VS Actual Closing price for **training data** set after fitting linear regression.



Predicted VS Actual Closing price for **testing data** set after fitting linear regression.



- **Ridge Regression:** A Ridge Regression model was fitted onto the data with following values for Cross Validation parameter for making better predictions evident in the scores below.

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)

Evaluation Metric Score	Training Data	Testing Data
R2_score	0.9614	0.8762
MSE	401.472	1019.8925
RMSE	20.0368	31.9358
MAE	13.5939	17.1518
MAPE	0.1634	0.1817

Predicted VS Actual Closing price for **training data** set after fitting ridge regression.

Predicted VS Actual Closing price for **testing data** set after fitting ridge regression.

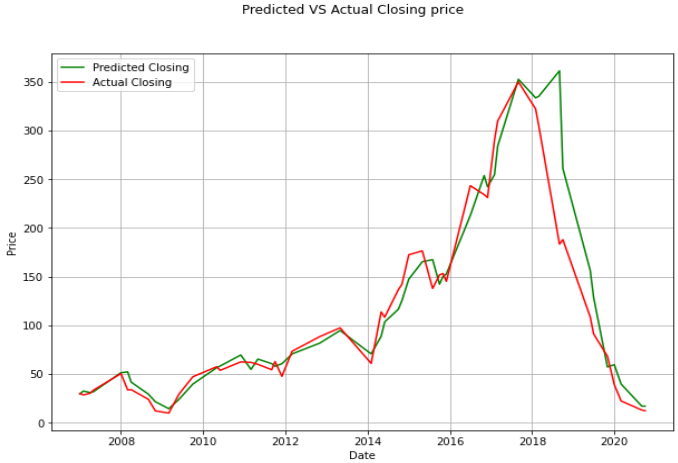
- **Lasso Regression:**A Lasso Regression model was fitted onto the data with following values for Cross Validation parameter for making better predictions evident in the scores below.

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)

Evaluation Metric Score	Training Data	Testing Data
R2_score	0.9568	0.8825
MSE	448.9065	967.9291
RMSE	21.1874	31.1116
MAE	13.295	16.3884
MAPE	0.1641	0.1933

Predicted VS Actual Closing price for **training data** set after fitting Lasso regression.

Predicted VS Actual Closing price for **testing data** set after fitting Lasso regression.



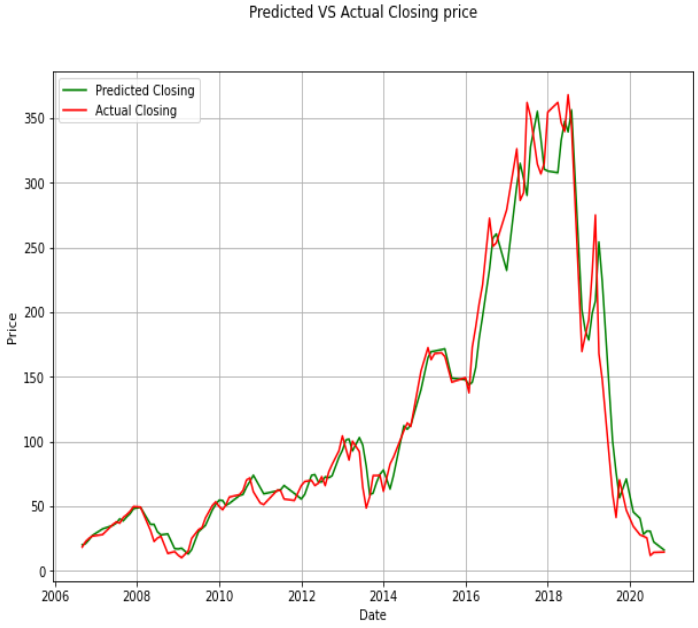
- **Elastic Net Regression:** An Elastic Net Regression model was fitted onto the data with following values for Cross Validation parameter for making better predictions evident in the scores below.

Alpha: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)

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: 1)

Evaluation Metric Score	Training Data	Testing Data
R2_score	0.9568	0.8825
MSE	448.9065	967.9291
RMSE	21.1874	31.1116
MAE	13.295	16.3884
MAPE	0.1641	0.1933

Predicted VS Actual Closing price for **training data** set after fitting elastic net regression.



Predicted VS Actual Closing price for **testing data** set after fitting Lasso regression.

- **XGBoost Model:** An XGBoost regression model was fitted onto the data sets along with 10 fold Cross Validation to predict the closing prices of Yes Bank. The Cross Validation Parameters and evaluation metric scores are given below:

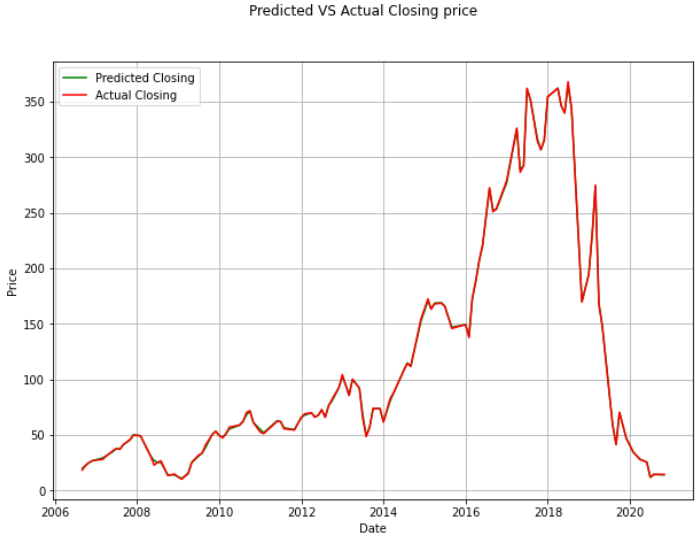
Number of estimators: 400, 700, 1000
Max tree depth: 2,3,5,7,10
Regularization alpha: 1.1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 50
Regularization lambda: 1.1, 1.2, 1.3
Subsample: 0.7, 0.8, 0.9
Colsample by tree: 0.3, 0.5, 0.7, 0.9, 1

CV Parameters	Best Value
Number of Estimators	1000
Max tree depth	3
regularization alpha	7
regularization lambda	1.1
subsample=	0.9
colsample by tree	1

Evaluation metric scores for XGBoost Regression Model:

Evaluation Metric Score	Training Data	Testing Data
R2_score	0.9999	0.9087
MSE	1.0735	751.9086
RMSE	1.0361	27.421
MAE	0.6516	15.7395
MAPE	0.0118	0.1657

Predicted VS Actual Closing price graph against timeline for **training data** set after fitting XGBoost regression.



Predicted VS Actual Closing price graph against timeline for **testing data** set after fitting XGBoost regression.

***9. Results and Conclusion*:** Following conclusions can be made after vigorous statistical and visual analysis along with feature engineering.

- Elimination of multicollinearity in a data set is critical for achieving good results in machine learning models.
- Importance of historical OHLC data is clear with the success of regression models built after considering 14 historical session average OHLC values as lags.
- A 10 fold hyper parameter tuning has been done for finding optimal values of the parameters as right predictions are critical in the stock market!
- It is found that the XGBoost model performed the best on both training and testing data, predicting the monthly closing price of Yes Bank with 0.9087 R2 score for testing data.

References:

- *https://github.com/*
- *https://stackoverflow.com/*
- *https://pandas.pydata.org/docs/*
- *https://www.almabetter.com/*
- *https://www.kaggle.com/*
- https://www.w3schools.com/
- https://www.geeksforgeeks.org/
- https://www.wikipedia.org/
- https://www.nseindia.com/