# Capstone Project-2

# Yes Bank Stock Closing Price Prediction

Zeeshan Ahmad

# 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 in 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.
- In this Project we have to analyze the closing price of stock market of Yes bank and predict the price.

# Data Summary

Date: It specifies the month and year of particular price.

Open: It specifies the opening price of the stock (Numeric).

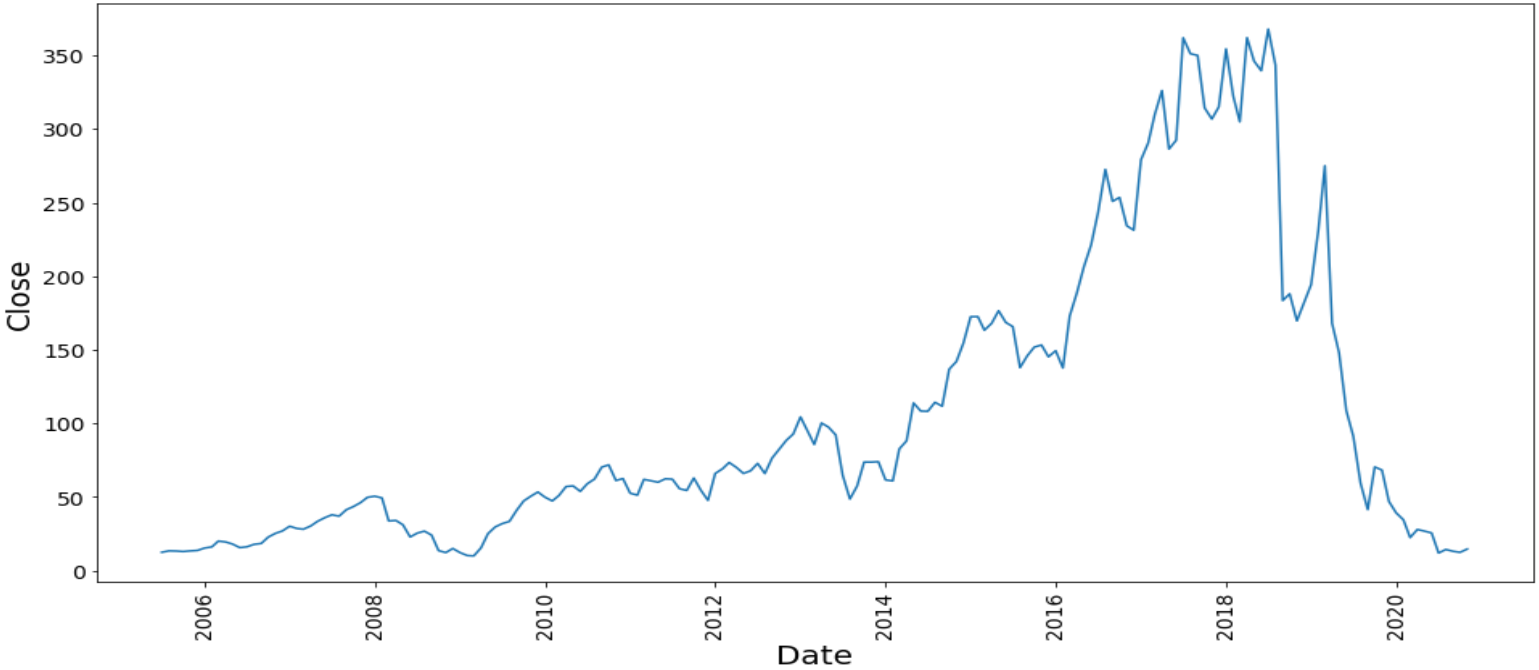
High: It specifies the highest price of the stock (Numeric).

Low: It specifies the Lowest price of the stock (Numeric).

Close: It specifies the Closing price of the stock (Numeric).

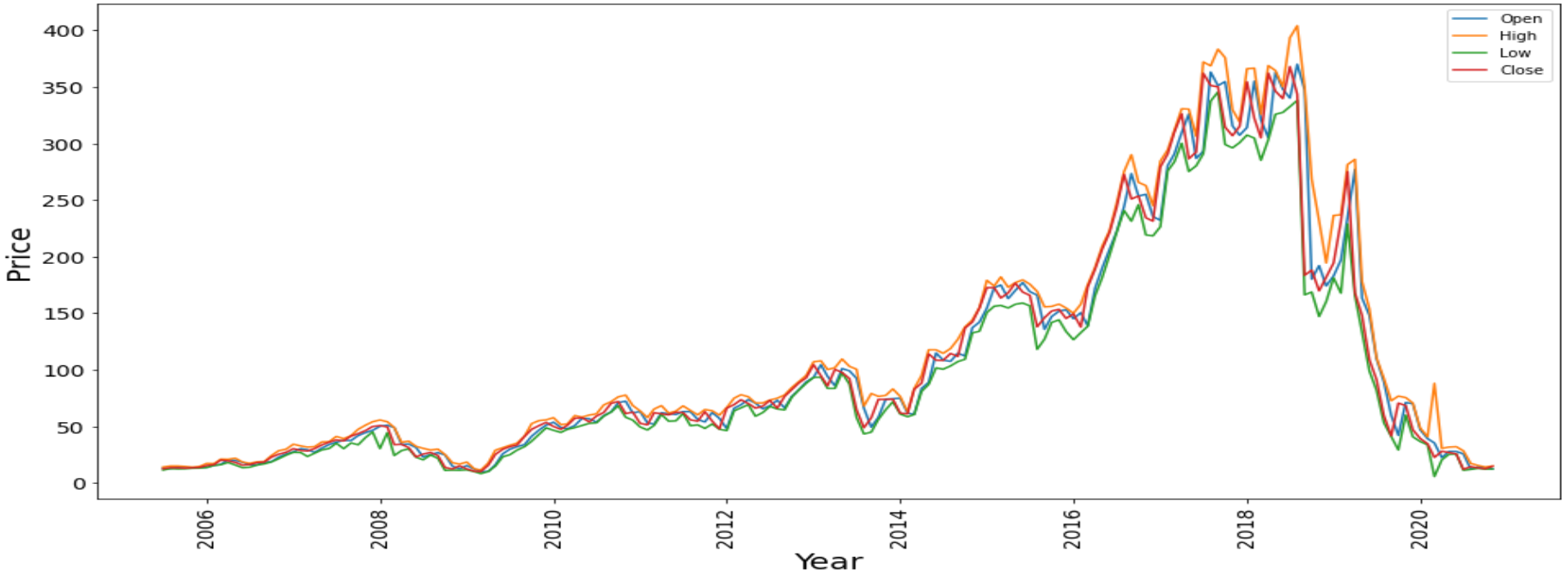
# Exploratory Data Analysis

## 1. Plot of Close Price




- We can see the peak is high in year between 2017 to 2019.
- Low is in start of year 2006, 2009 and after 2020.
- There is a sudden change after year of 2014.

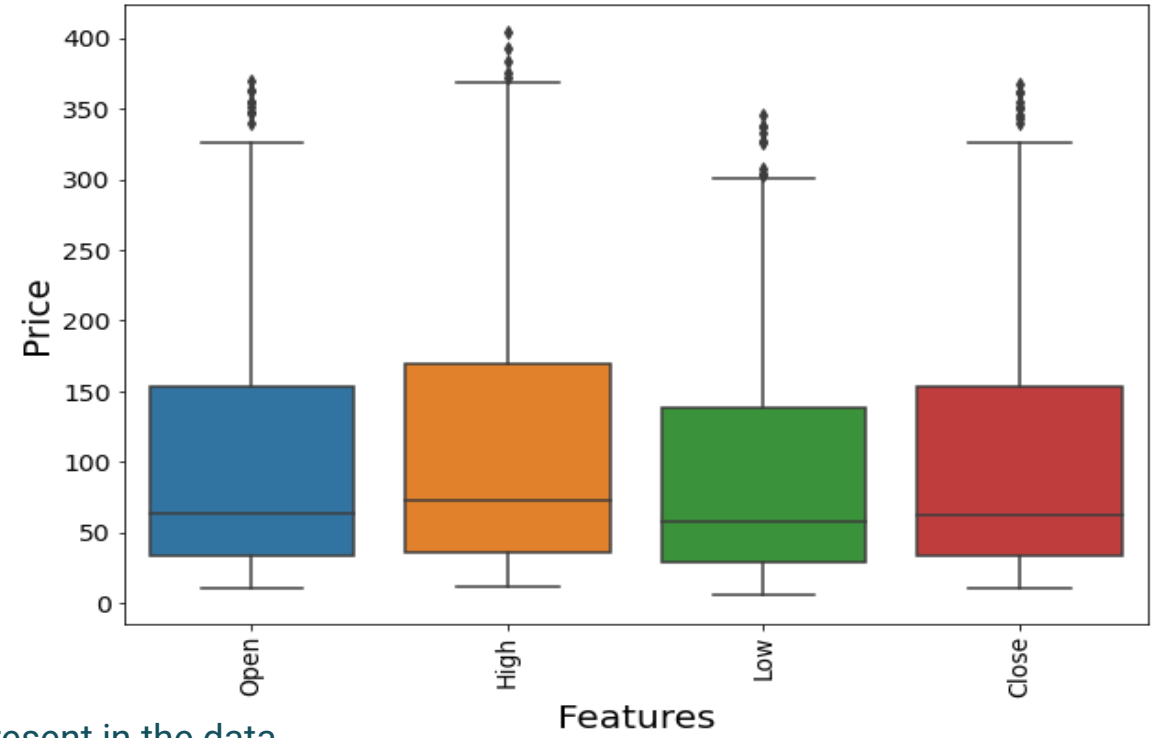
# 2. Plot of all features against Year




The above graph shows the variation of open, high, low and close of stock price.

All features have similar characteristics not much differences.

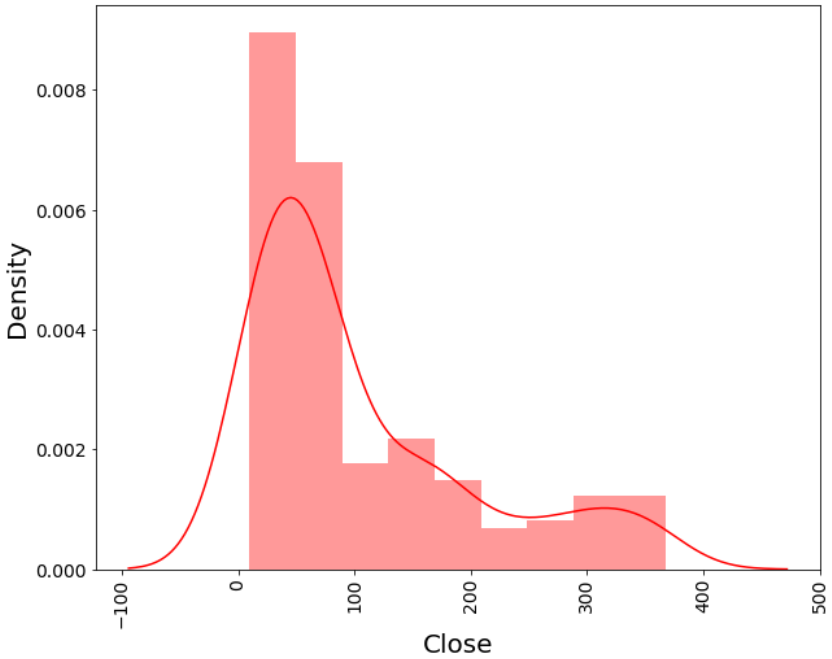
# 3. Features Outliers Detection



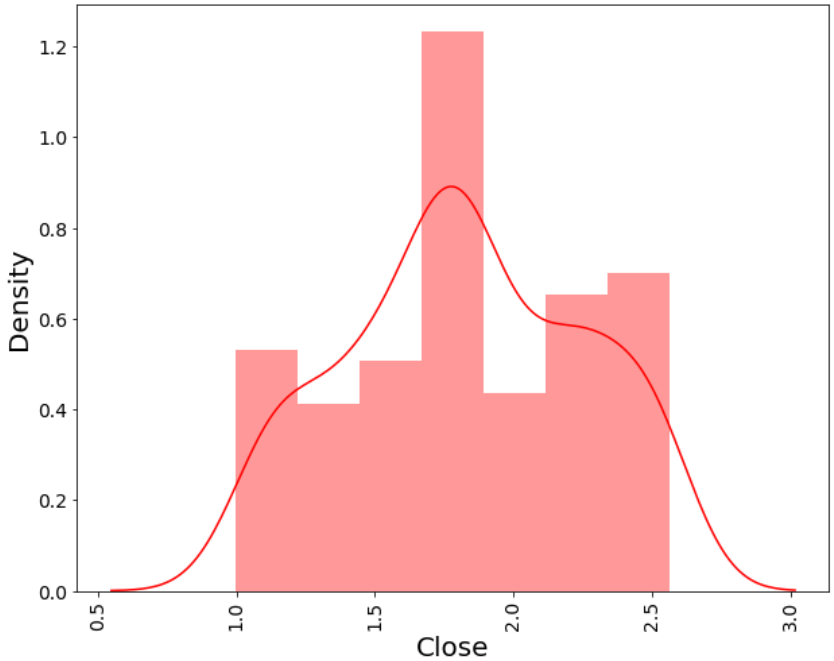

- Outliers are present in the data.
- As above boxplot shows outliers this is because of stock price fall from nearly around 400 to 20.This happen quick within very few months. That's why top value of stocks looks like outliers.
- Open and Close have very similar values.
- The median of the close price is around 60(approx.).

# 4. Distribution of Dependent Variable(Close Price)

Positively Skewed




After Log Transformation



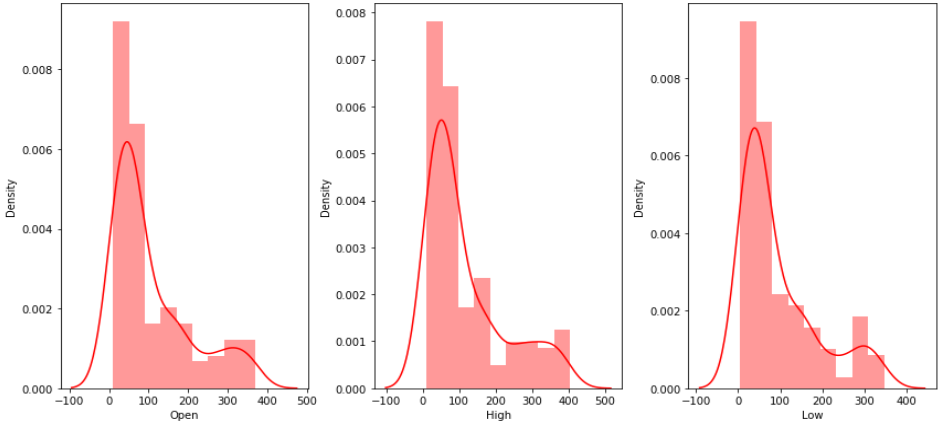

After applying log transformation. Data is normally distributed of close price.

# 5. Distribution of Independent Variables

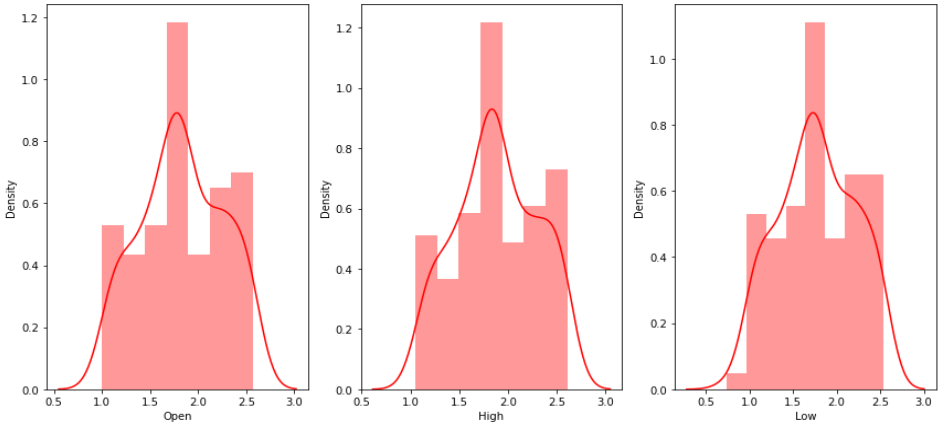
Other features i.e. independent variable are also positively skewed. The above histogram shows data distribution of open, high and low.

The independent data is normal distributed after applying log transformation.

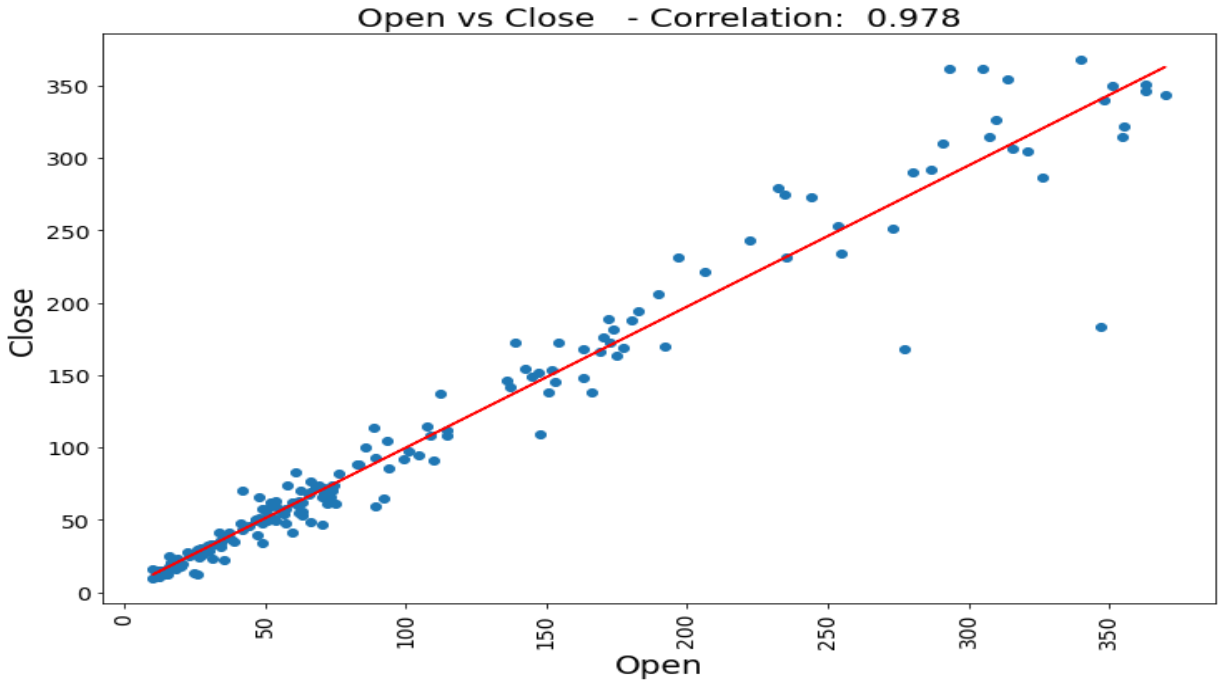
Other features are also positively skewed




After Log Transformation

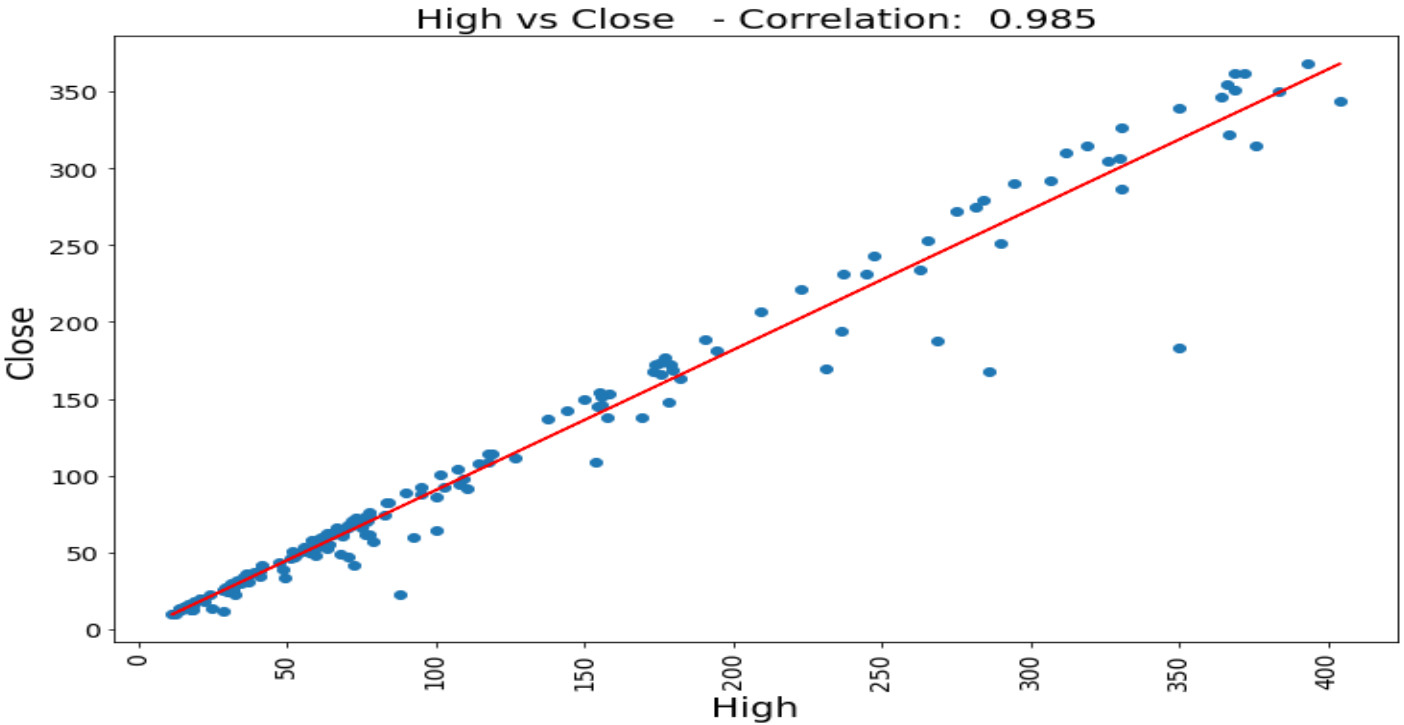


# 6. Scatter Plot of Close and Open Price



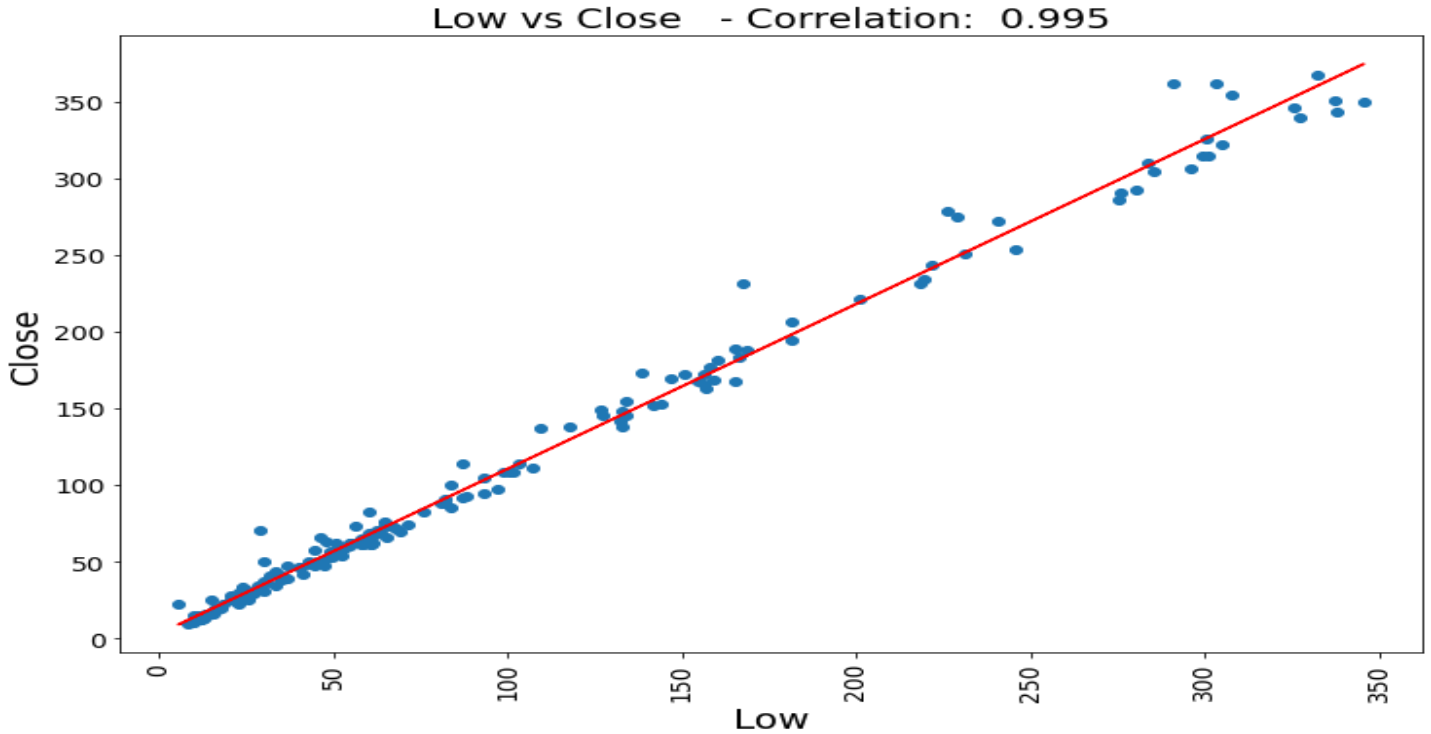

- 'Open price' is linearly correlated( correlation 0.978) with the dependent variable (close price).
- Small amount of Outliers are also present.

# 7. Scatter Plot of Close and High Price



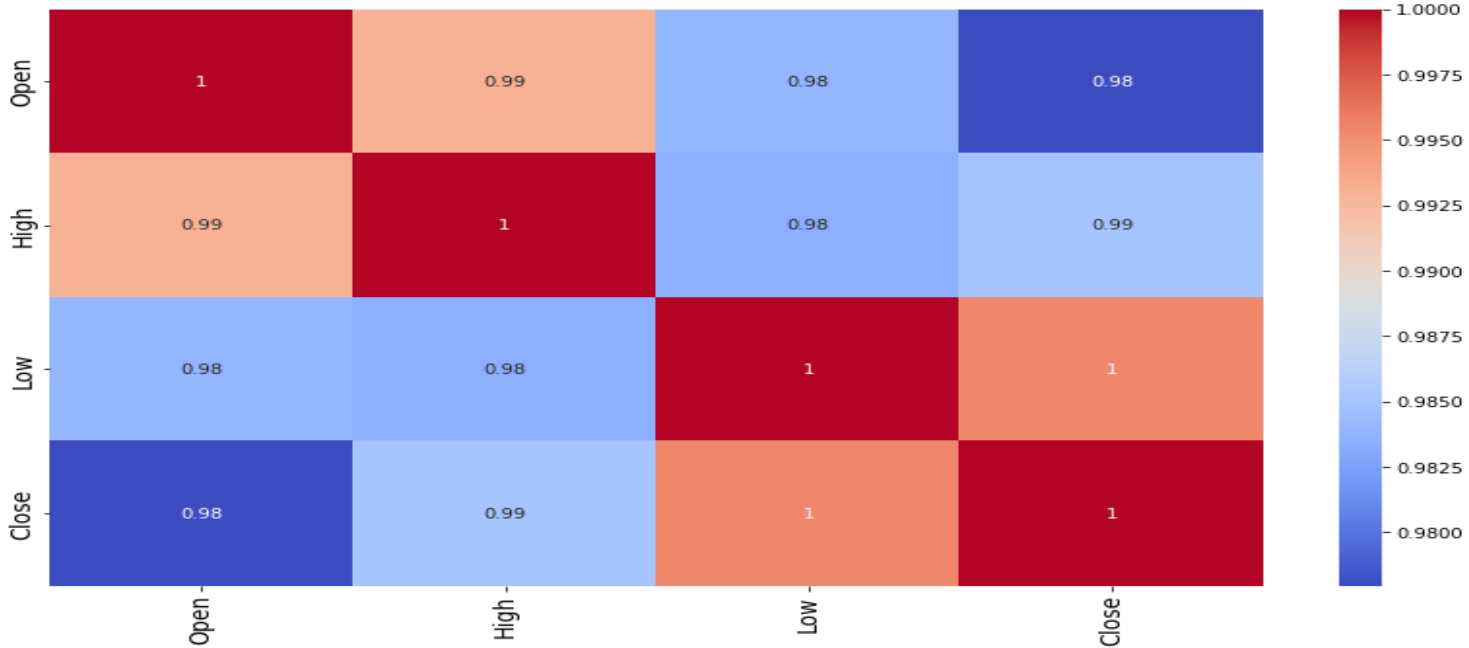

- The above scatter plot shows the variation of yes bank stock 'high price' against closing price.
- The 'High price' is linearly correlated (0.985) with the dependent variable (close price).

# 8. Scatter Plot of Close and Low Price




- The above scatter plot shows the variation of yes bank stock 'low price' against 'close price'.
- The 'Low price' is linearly correlated (0.995) with the dependent variable (close price).

# 9. Correlation between variables




- The above heatmap shows the high correlation between independent variables open, high and low with dependent variable close price.
- Data is highly correlated, multicollinearity also present.
- Even though we have high VIF in our dataset we can't perform feature engineering because features are important and limited.
- All this features are required to predict close price.

# Data Transformation

- To scale data into a uniform format that would allow us to utilize the data in a better way.
- For performing fitting and applying different algorithms to it.
- We are using Standardization (z- score) method which can be helpful in cases where the data follows a Gaussian distribution.
- We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.
- Standardization does not get affected by outliers because there is no predefined range of transformed features.

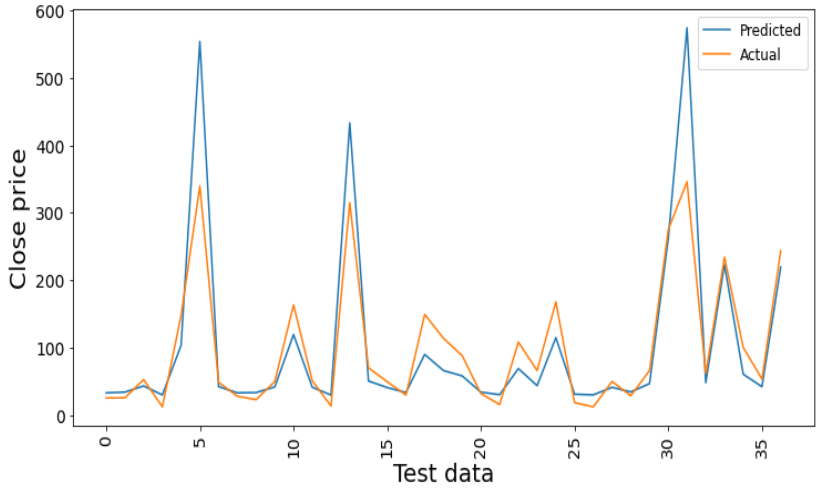
# Train- test split

- The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data.
- In this project we have used 80% data for training purpose and 20% data for test set.
- Training dataset is for making algorithm learn and train model.
- Test dataset is for testing the performance of train model.

# Models

## 1) Linear Regression:

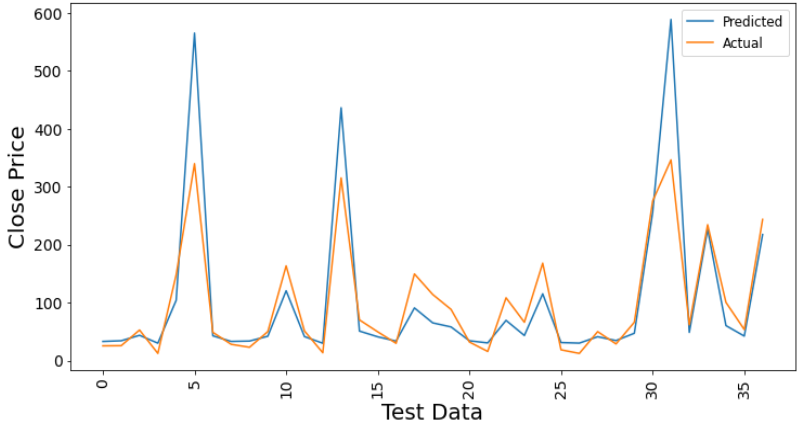
Linear regression algorithm shows a linear relationship between a dependent and independent variable; hence it is called as linear regression.




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

## 2) Lasso Regression

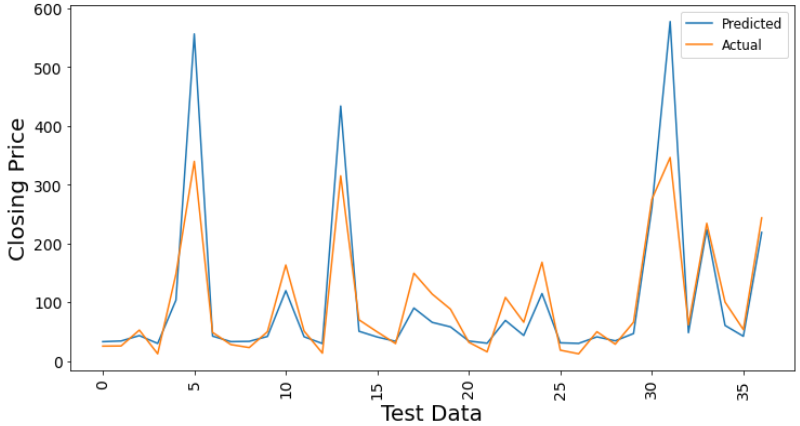
- Lasso: Least Absolute Shrinkage and Selection operator.
- It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.
- This method performs L1 regularization.




- 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

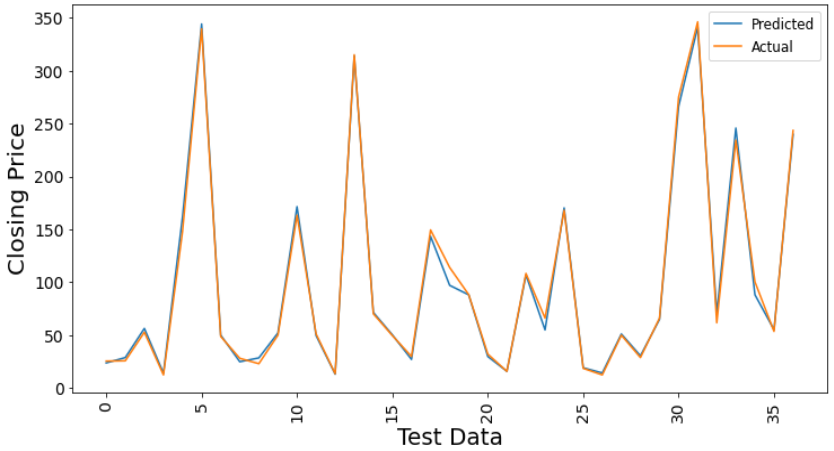
- Cross Validation is a technique using which Model is evaluated on the dataset on which it is not trained that is it can be a test data or can be another set as per availability or feasibility.
- Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid 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 is : -0.035

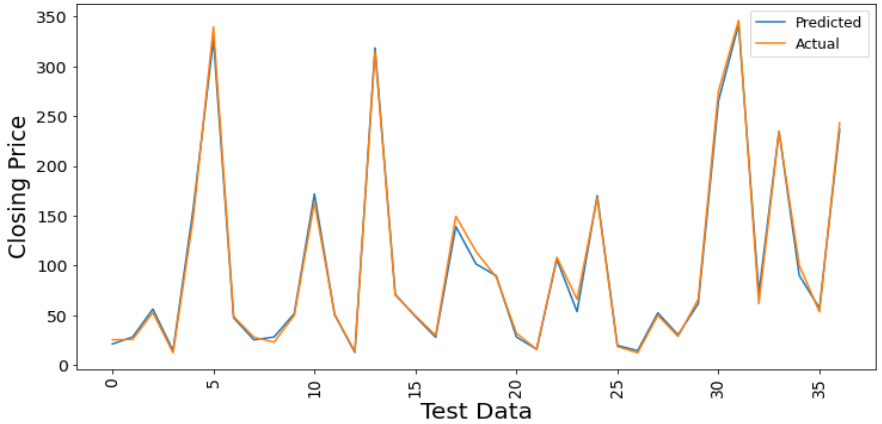
## 4) Random Forest

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.
- We applied 5 models out of 2 models gives best result

# 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

- High, low, open are directly correlate with the closing price of stocks.
- Features are multicollinear but can not drop the column because features are limited.
- The test results of all the regression models are evaluated and compared. We checked performance metrics such as R2 score, Mean Scored Error, and Root Mean Score 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, 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