Yes Bank Stock Closing Price Prediction

Team- Data Chronicle

Team members:

Kaushal Kumar Jha,
Shambhu Nath Jha,
Nimesh Thakur ,Asif PA,
Data Science Trainees,
AlmaBetter, Bangalore

Abstract:

In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. The paper focuses on the use of Regression based Machine learning to predict stock values.

Factors considered are open, close, low, and high and these factors are used to train the model and make future prediction.

Selection of best model is based on evaluation matrices mainly Root Means Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) which is used to predict forecast accuracy.

Model with lowest RMSE and MAPE will be considered as best model.

1.**Introduction**

Stock market is characterized as dynamic, unpredictable and non-linear in nature. Predicting stock prices is a challenging task as it depends on various factors including but not limited to political conditions, global economy, company's financial reports and performance etc. Thus, to maximize the profit and minimize the losses, techniques to predict values of the stock in advance by analyzing the trend over the last few years, could prove to be highly useful for making stock market movements.

Stock Price Prediction using machine learning helps you discover the future value of company stock and other financial assets traded on an exchange. The entire idea of predicting stock prices is to gain significant profits. Predicting how the stock market will perform is a hard task to do. There are other factors involved in the prediction, such as physical and psychological factors, rational and irrational behavior, and so on. All these factors combine to make share prices dynamic and volatile. This makes it very difficult to predict stock prices with high accuracy.

2. **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.
In given data set we have total 5 columns that we further divided into independent and dependent variable and used these variables to make future prediction.

We have performed regression analysis to make future prediction using various machine learning models and selected one model as best model after comparing them using evaluation matrices and this model will be going to use for future prediction.

3. Feature Description

The dataset of YES BANK has monthly stock prices of the bank since its inception and includes closing, starting, highest, and lowest stock prices of every month of around 180 observations.
It contains the following feature:

- Date: The date of record finalizes the transfer of the stock's ownership.
- Open: Open Price is the price at which the financial security opens in the market when trading begins. It may or may not be different from the previous day's closing price. The security may open at a higher price than the closing price due to excess demand of the security.
- High: High is the highest price at which a stock traded during a period.
- Low: Low is the highest price at which a stock traded during a period.
- Close: closing price of a stock is the price at which the share closes at the end of trading hours of the stock market.

4.Steps Involved

A. **Data Loading:** - First important step in any project, that is successfully loading of data because after loading data only we can make any decision. In this project first of all we have mounted the drive than we have loaded our dataset which was given in csv format.

B. **Understanding Of Data: -** In this step we have checked our dataset using head, tail, info() etc. methods to understand our dataset. After these methods we came to know that in our dataset 185 rows and 5 columns are there.

C. **Data Cleaning: -** Data cleaning is import to get best result from our dataset and in cleaning we check null values, duplicate data and correct data formats. In our dataset neither null values nor duplicate data was present, and we have changed the data column in correct format.

D. **Visualization: -** Visualization is important to understand dataset more clearly and for this we have used seaborn and matplotlib libraries. We will discuss more in details about visualization in Exploratory Data Analysis (EDA) part.

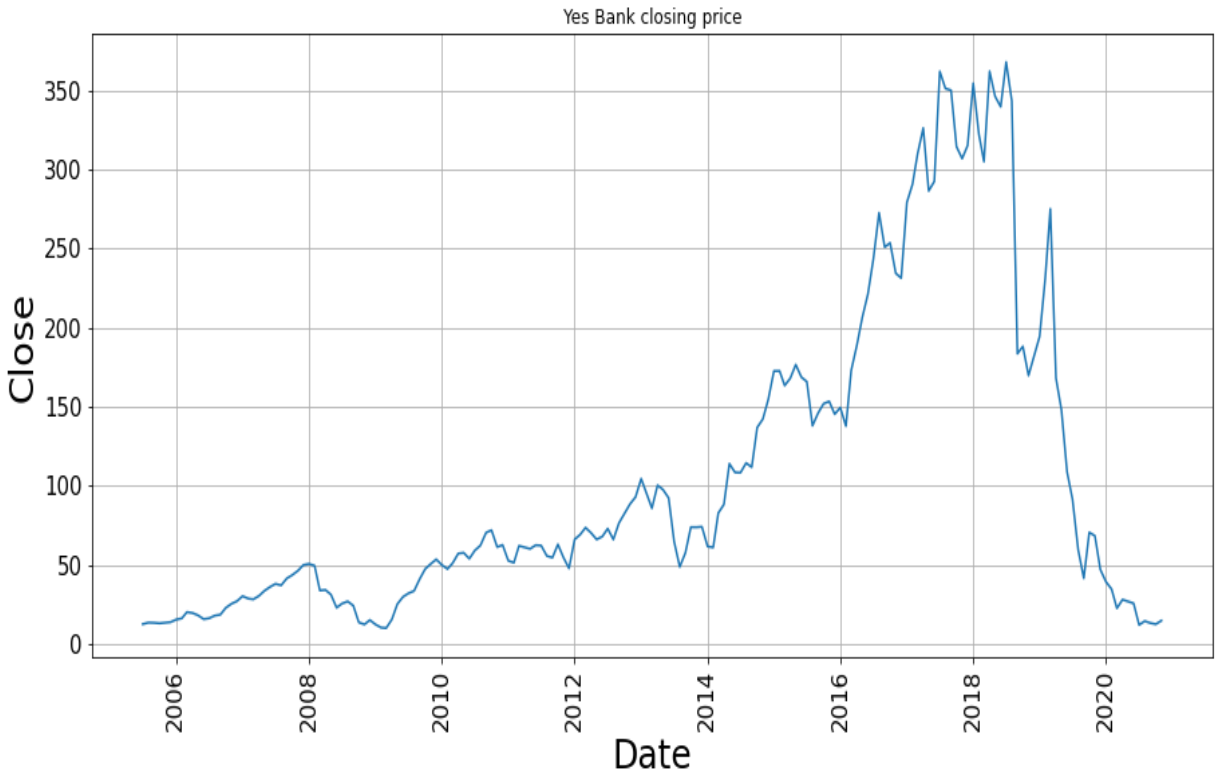
E. **Model Development: -** In this step first of all we have split the dataset into training and testing set and then made different models. We will see in details about models ahead in this.

EXPLORATORY DATA ANALYSIS(EDA)

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

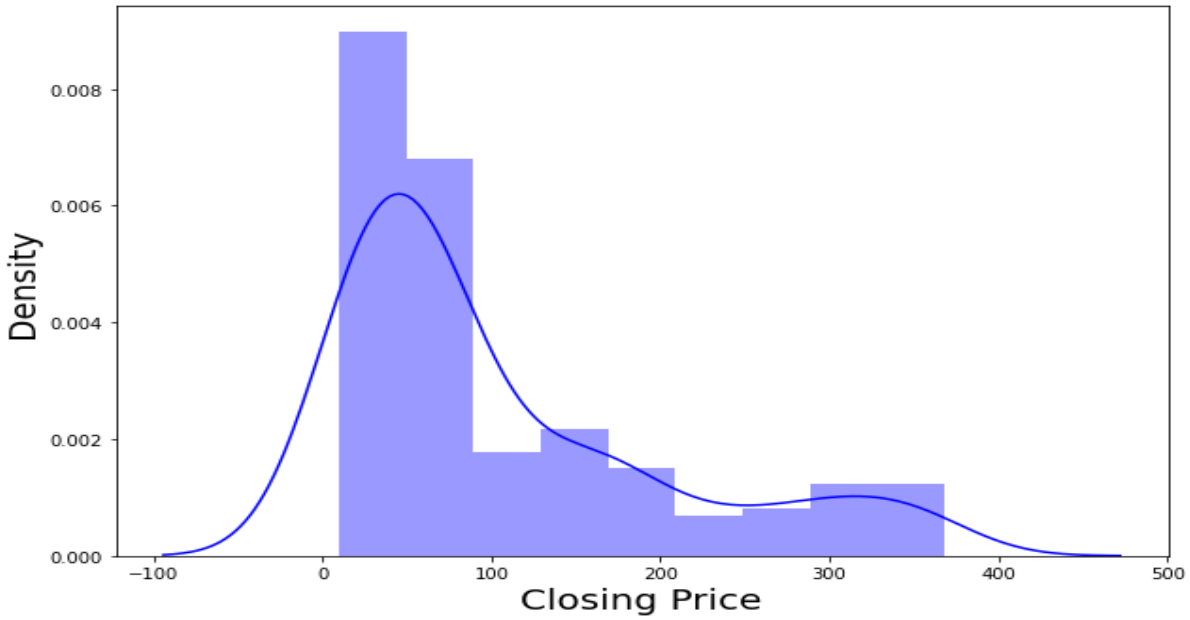
Univariate Analysis

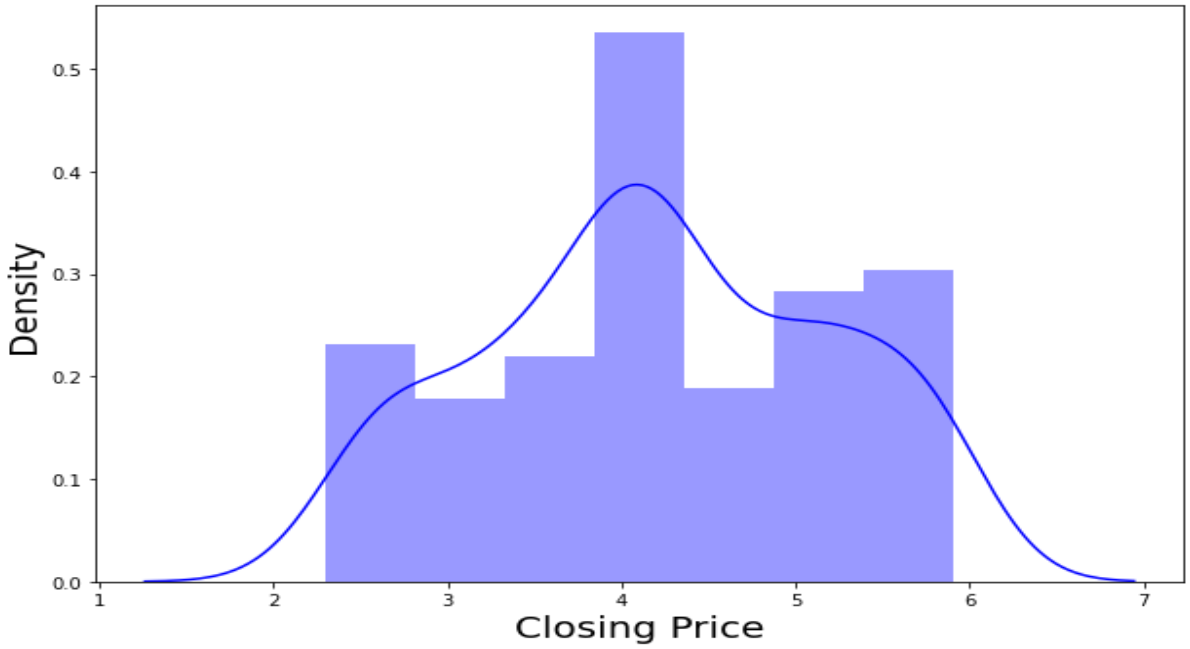
1) Closing Price Trend



From above graph it is clear that, trend was continuously increasing from 2009 till 2018. After 2018 it started to decrease due to fraud case of Rana Kapoor.

2)Distribution of Dependent/Target variable (Close)

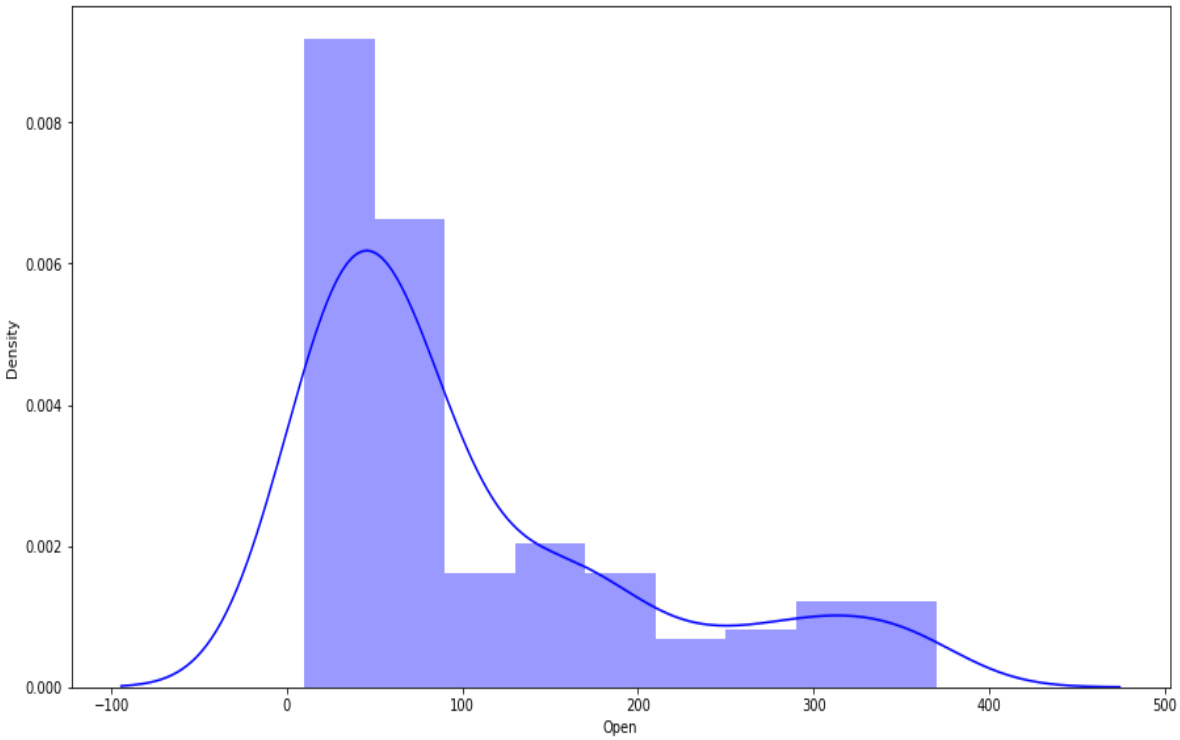


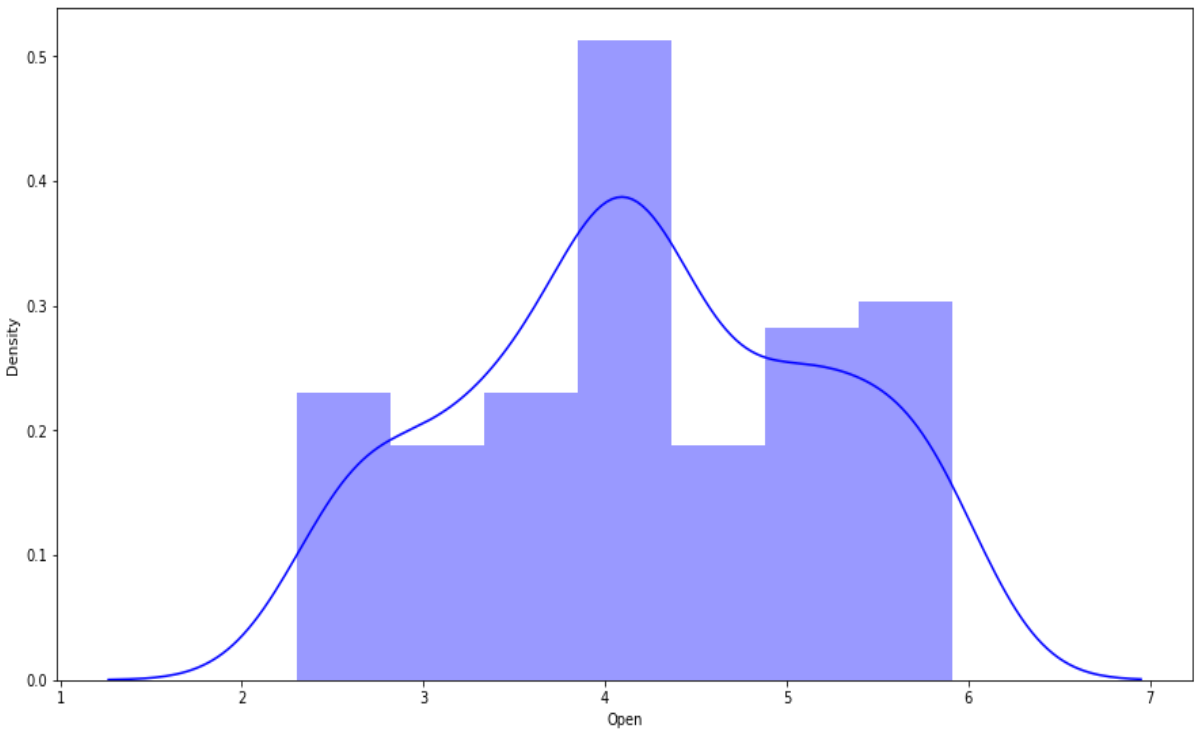


Distribution is **right skewed** so we have done **transformation** to make it **Normal** using **log transformation** method.

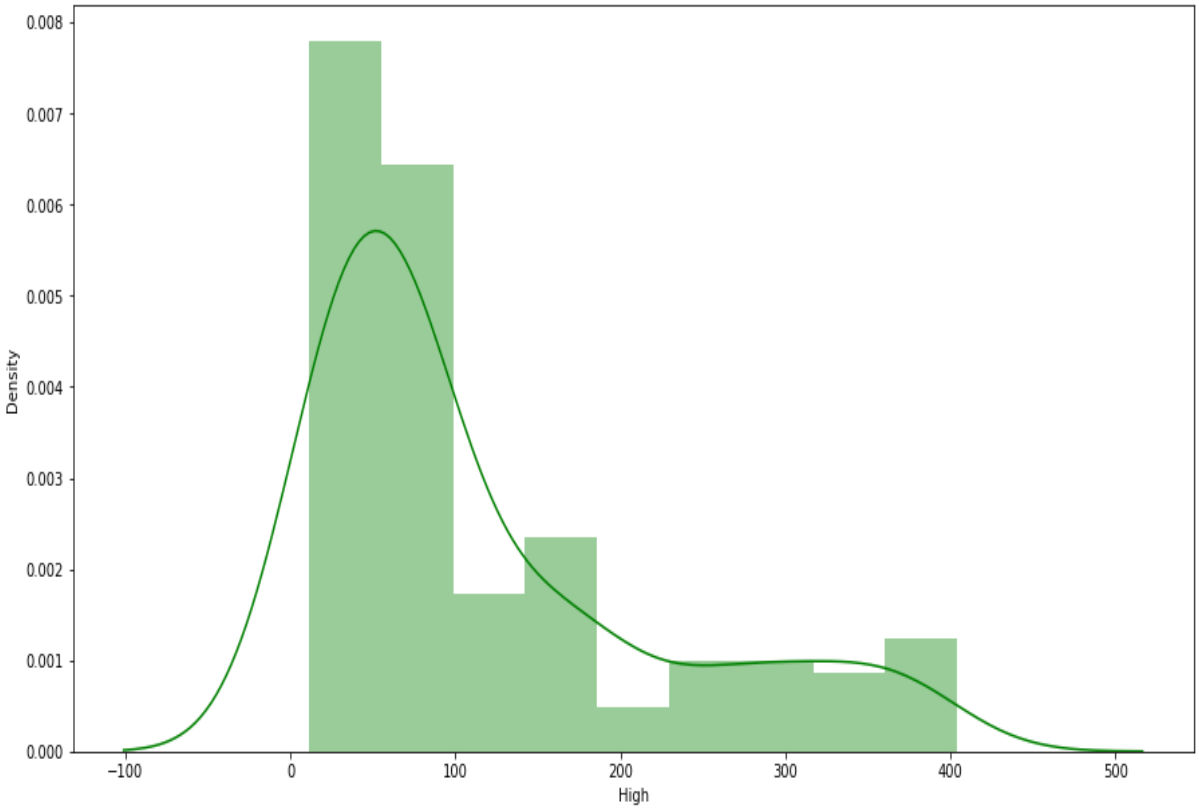
3.Distribution of all independent variable before and after transformation: -

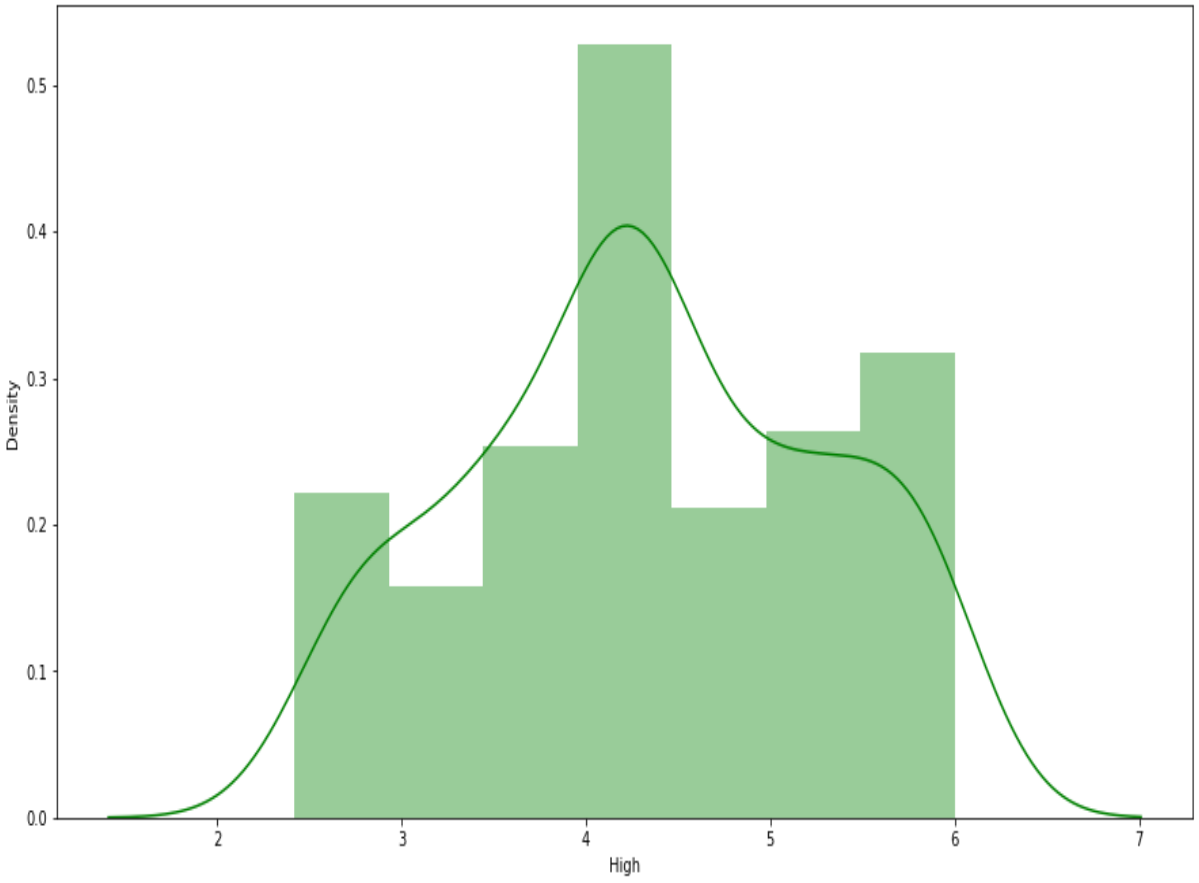
- **Open**



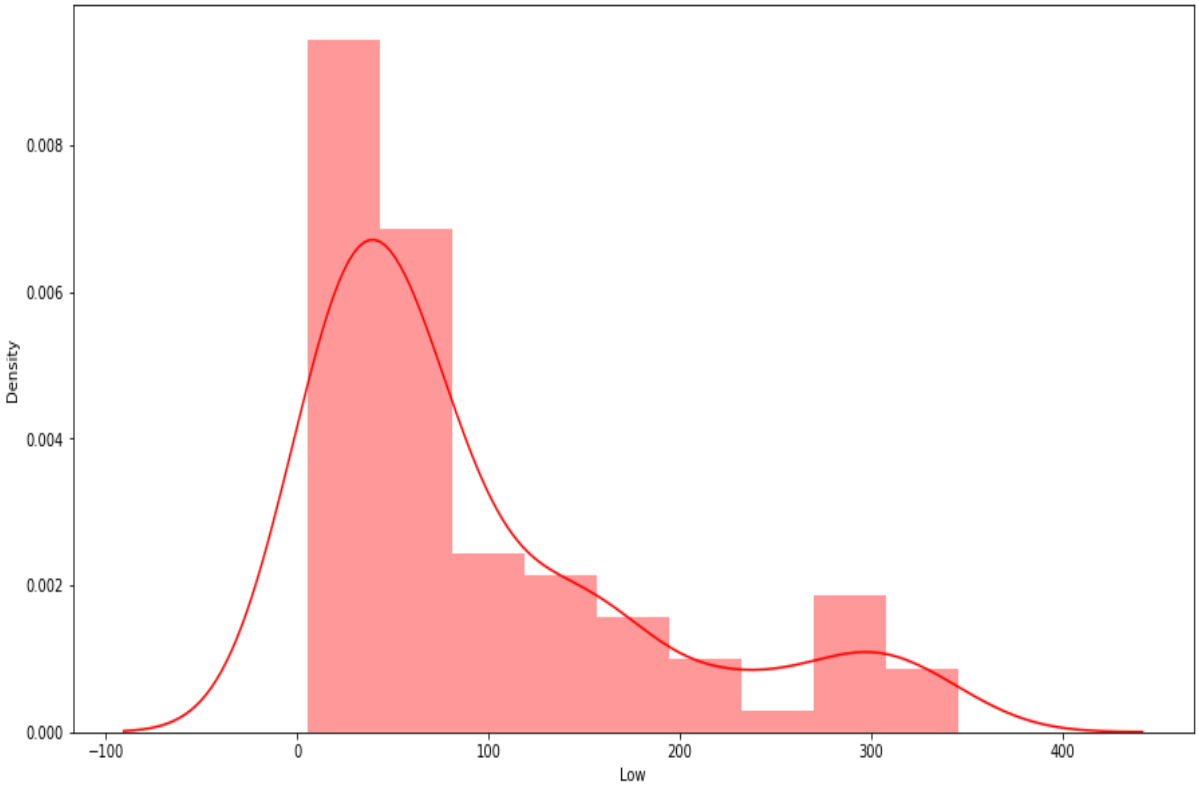


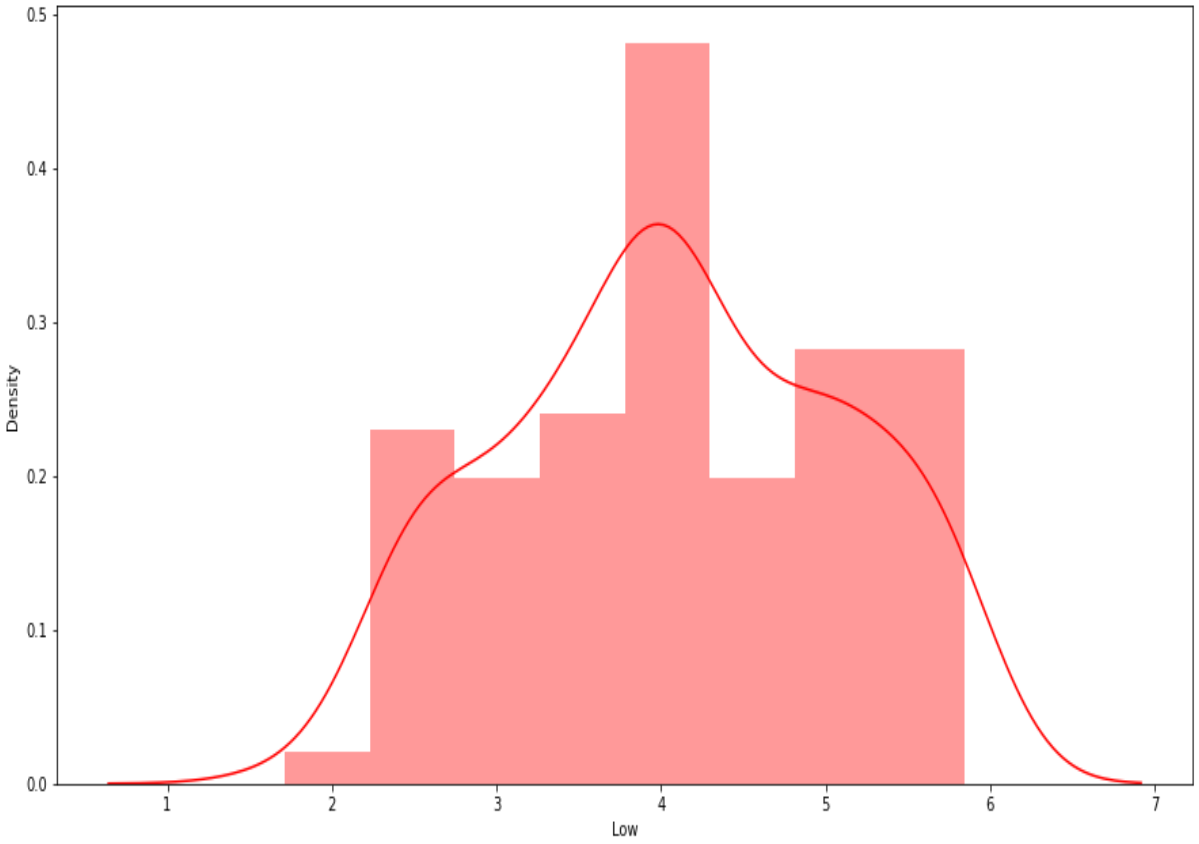
➢ **High**


0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
−100
0
100
200
300
400
500
High


0.5
0.4
0.3
0.2
0.1
0.0
Density
2
3
4
5
6
7
High

➢ **Low**



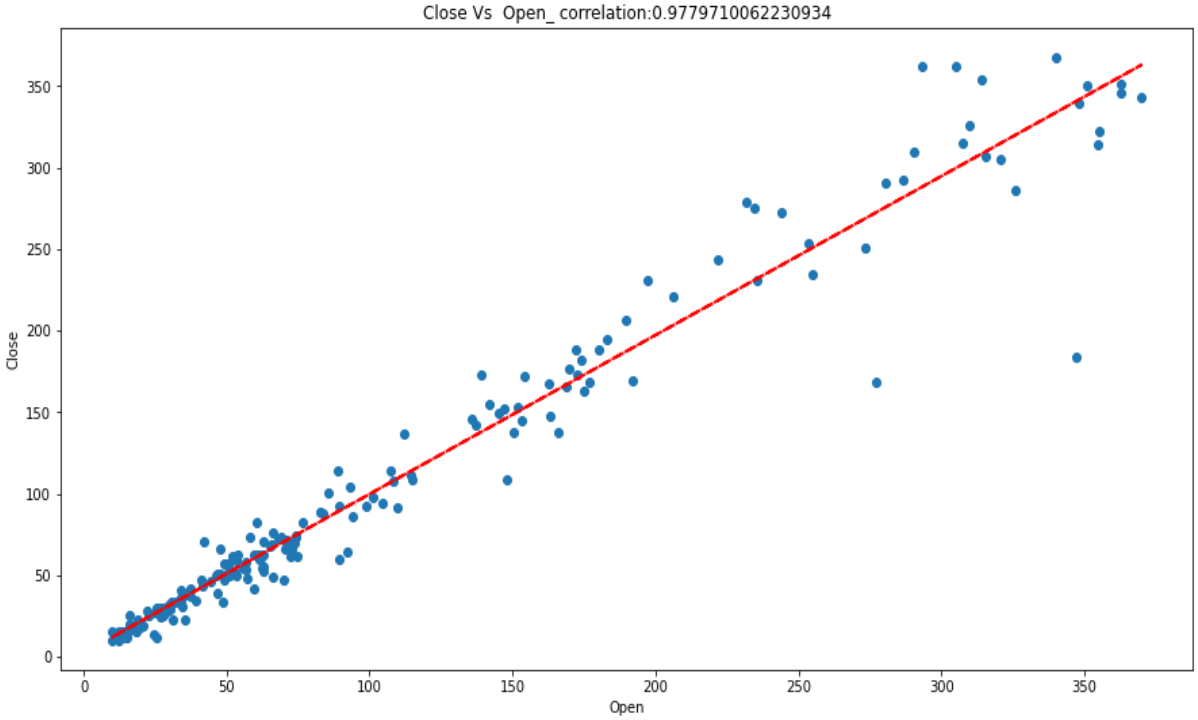


Distribution for all independent variables was right skewed after log transformation it became normal.

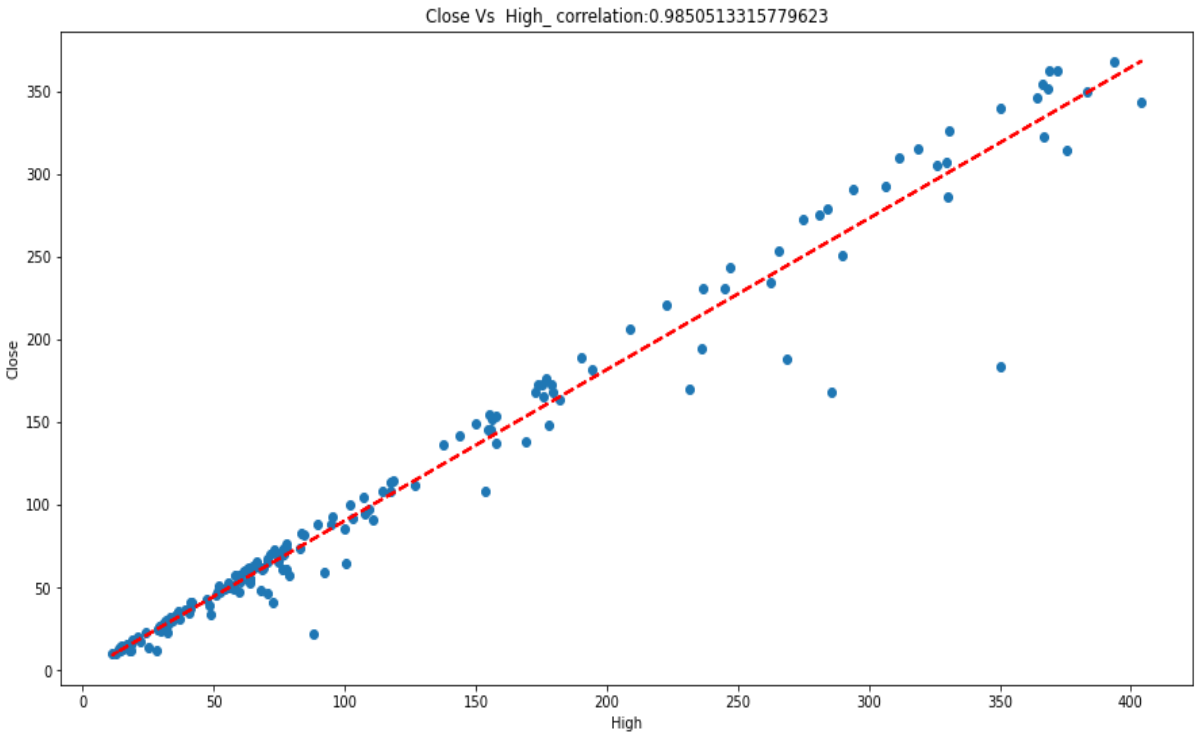
Bivariate Analysis.

In this we have checked the relation between independent and dependent variables.

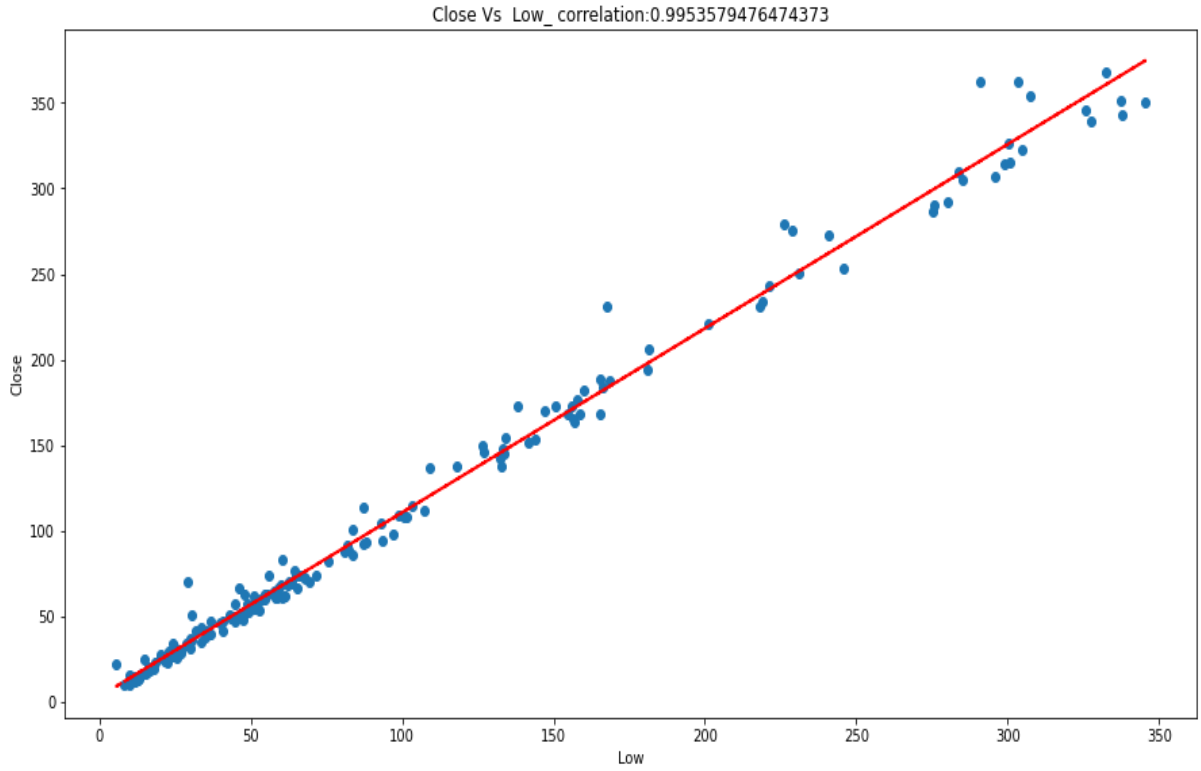
- Open Vs Close:



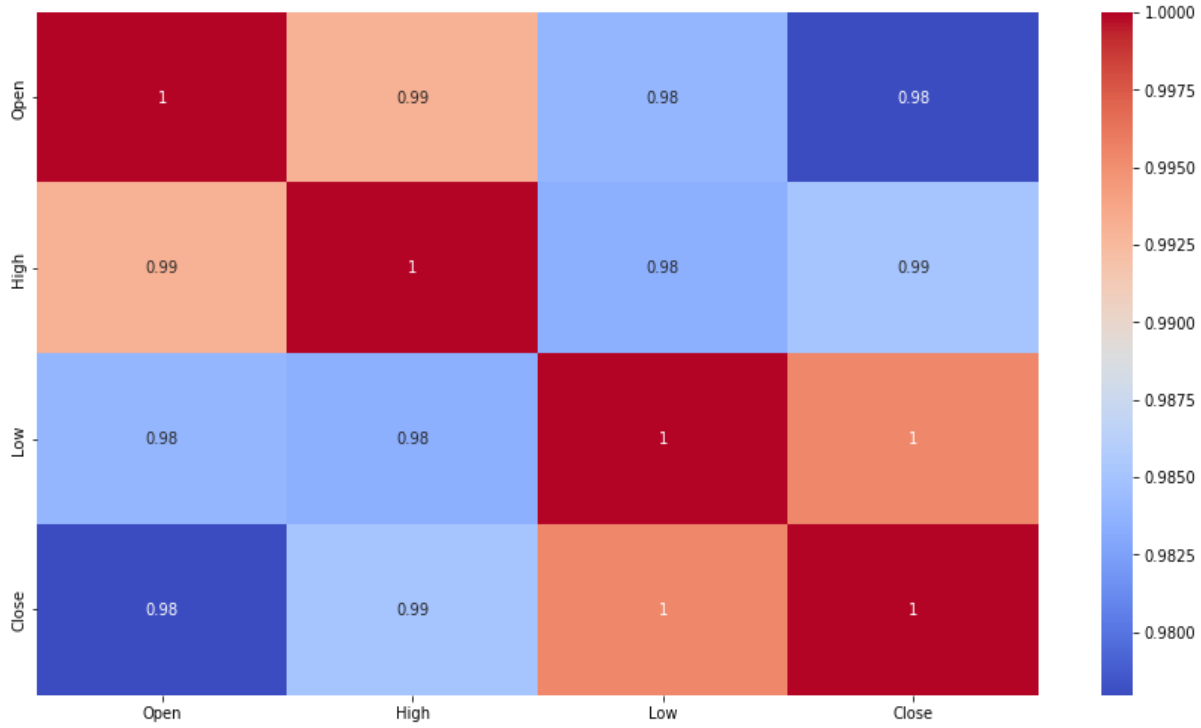
- **High Vs Close:**



➢ **Low Vs close:**



➢ **Correlation using Heatmap:**

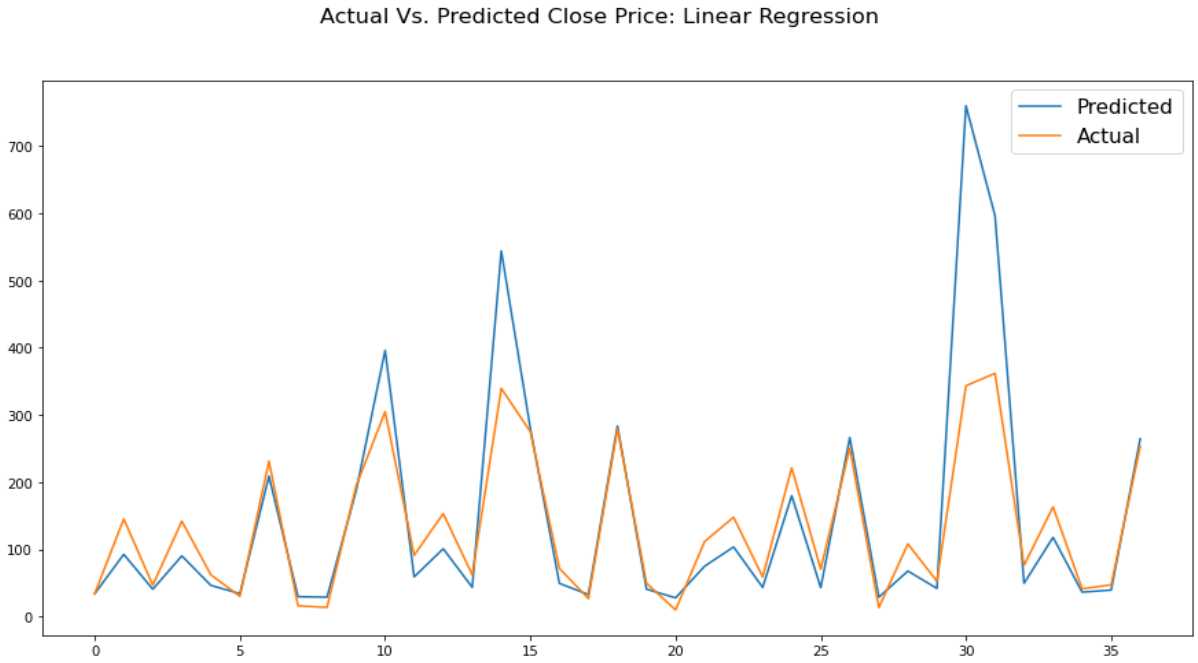


From above visualization it is clear that all independent variables are highly correlated with dependent variable.

Model Development

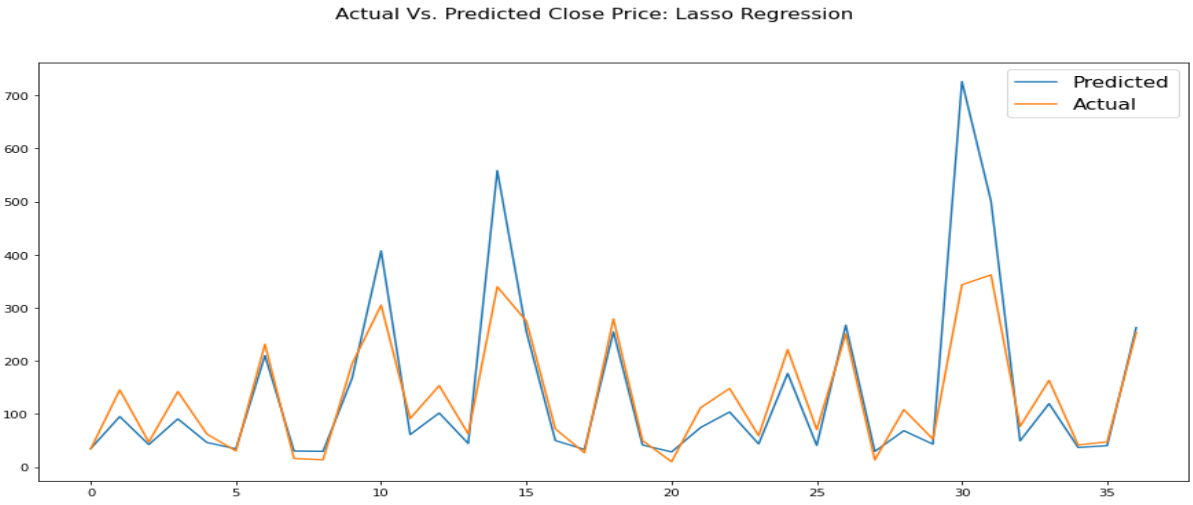
➢ Linear Regression:

Linear regression is a popular and uncomplicated algorithm used in data science and machine learning. It's a supervised learning algorithm and the simplest form of regression used to study the mathematical relationship between variables.



➢ Lasso Regression:

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients.

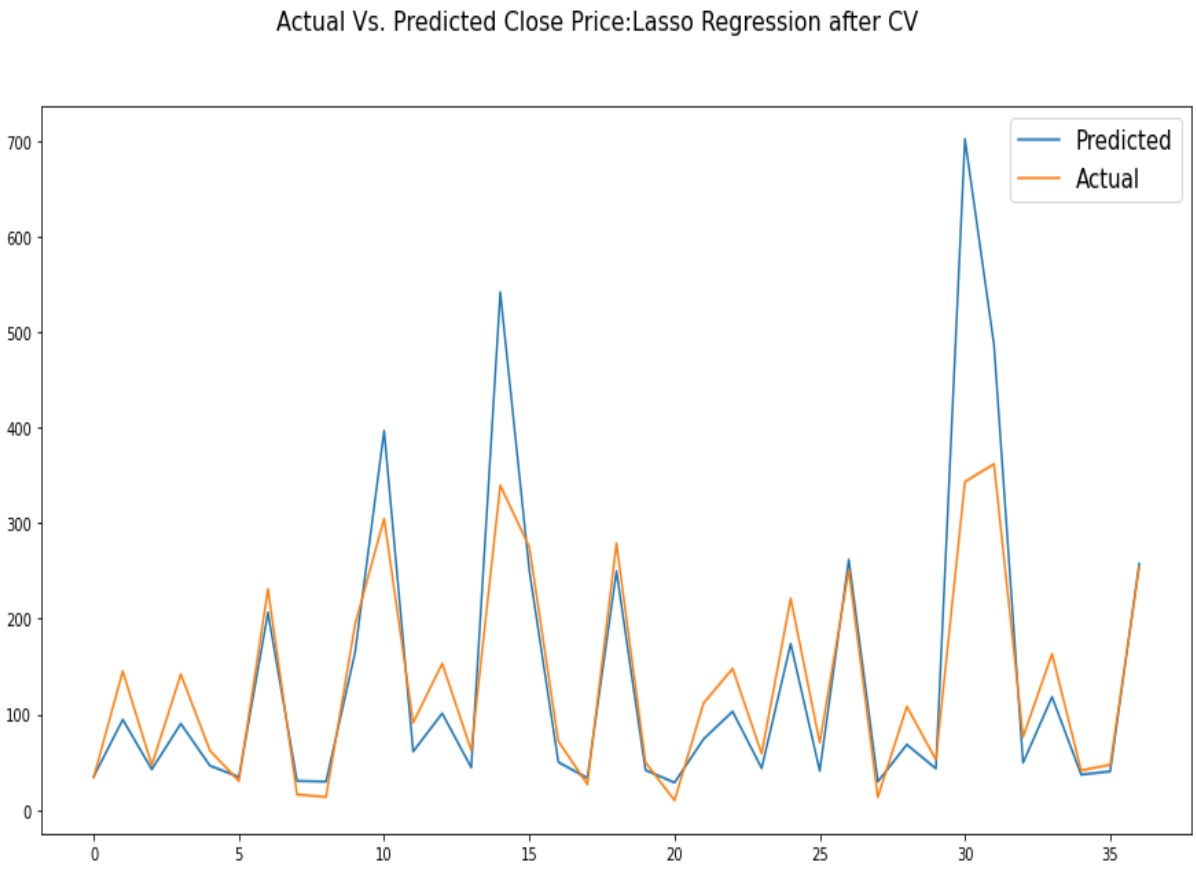


Cross Validation:

Cross validation (CV) is one of the techniques used to test the effectiveness of a machine learning models by dividing data into two segments: one used to learn or train a model and the other used to validate the model. If the model performs well over the test data and gives good accuracy, it means the model hasn't overfitted the training data and can be used for prediction. It is also a re-sampling procedure used to evaluate a model if we have a limited data.

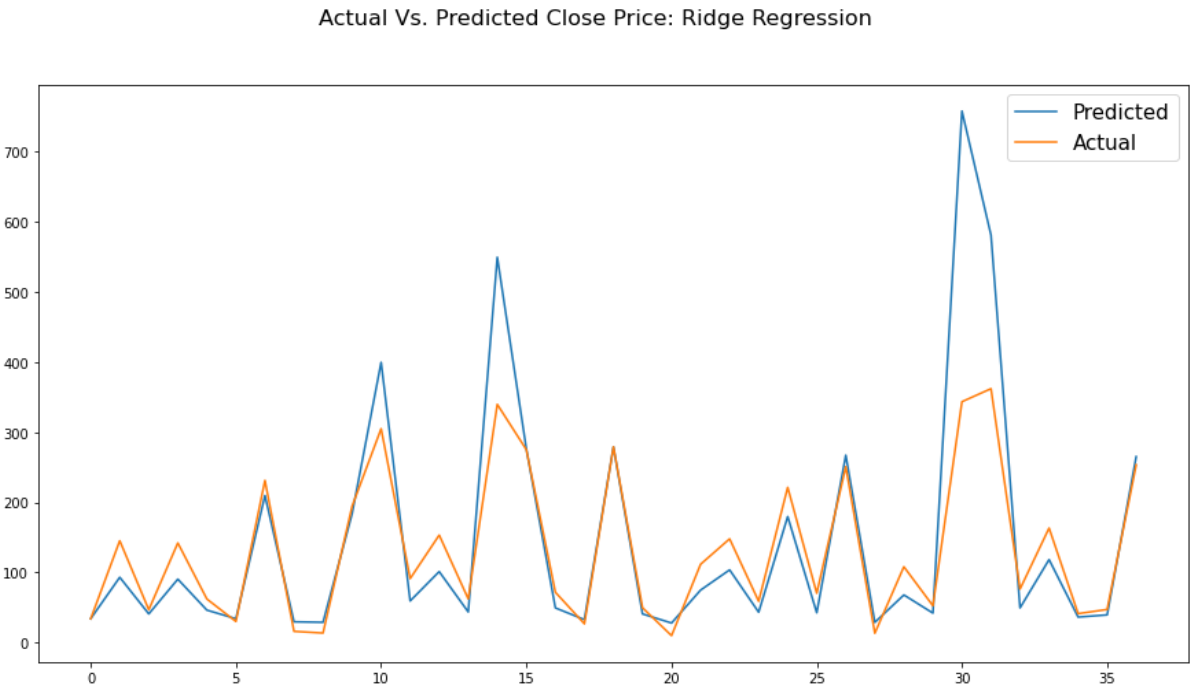
To reduce variability, in most methods' multiple rounds of cross validation is performed using different partitions, and the validation results are combined (e.g., averaged) over the rounds to give an estimate of the model's predictive performance.

Lasso Regression after cross validation.

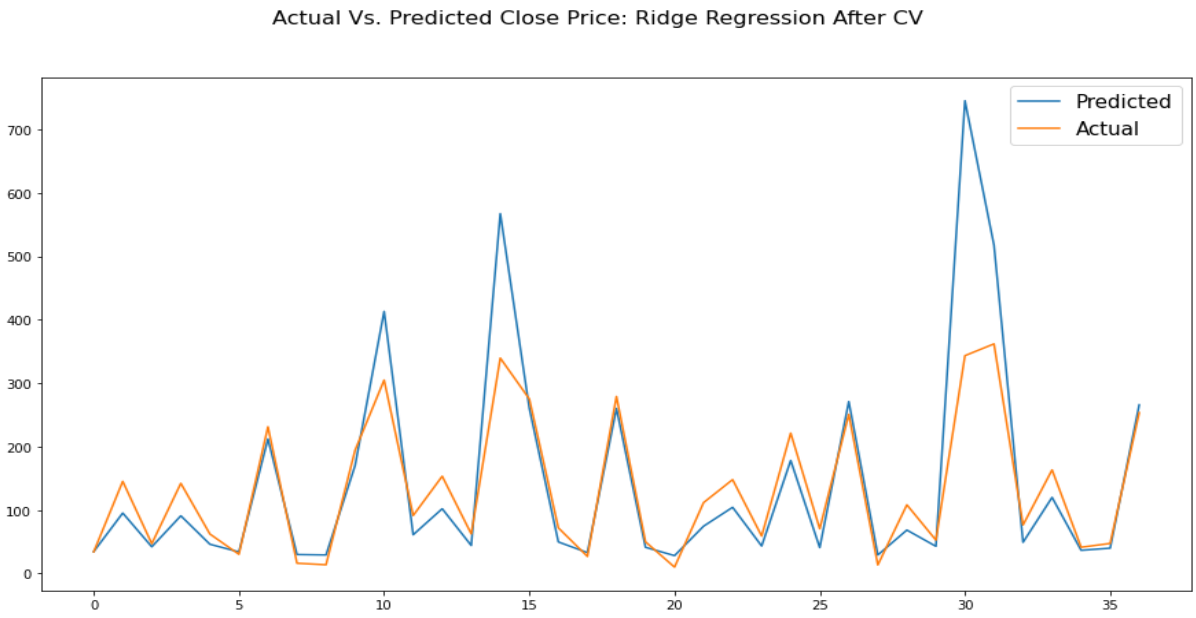


➢ Ridge Regression:

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

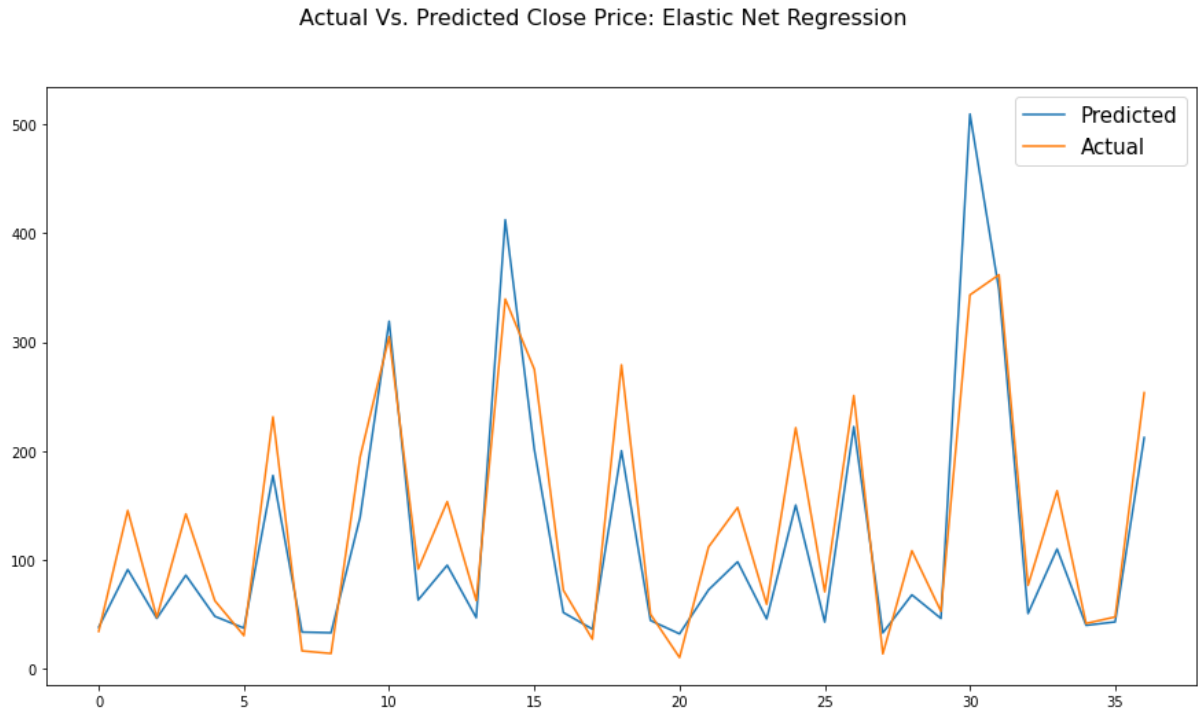


After Cross Validation:

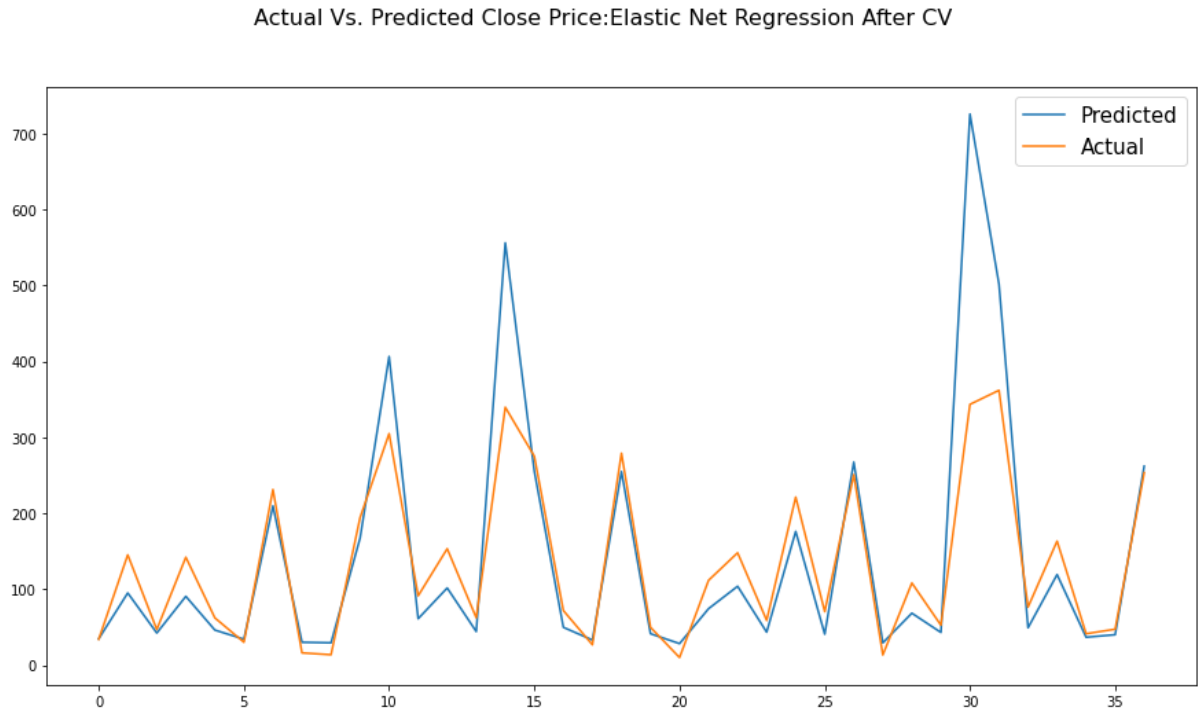


➢ Elatic Net Regression:

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

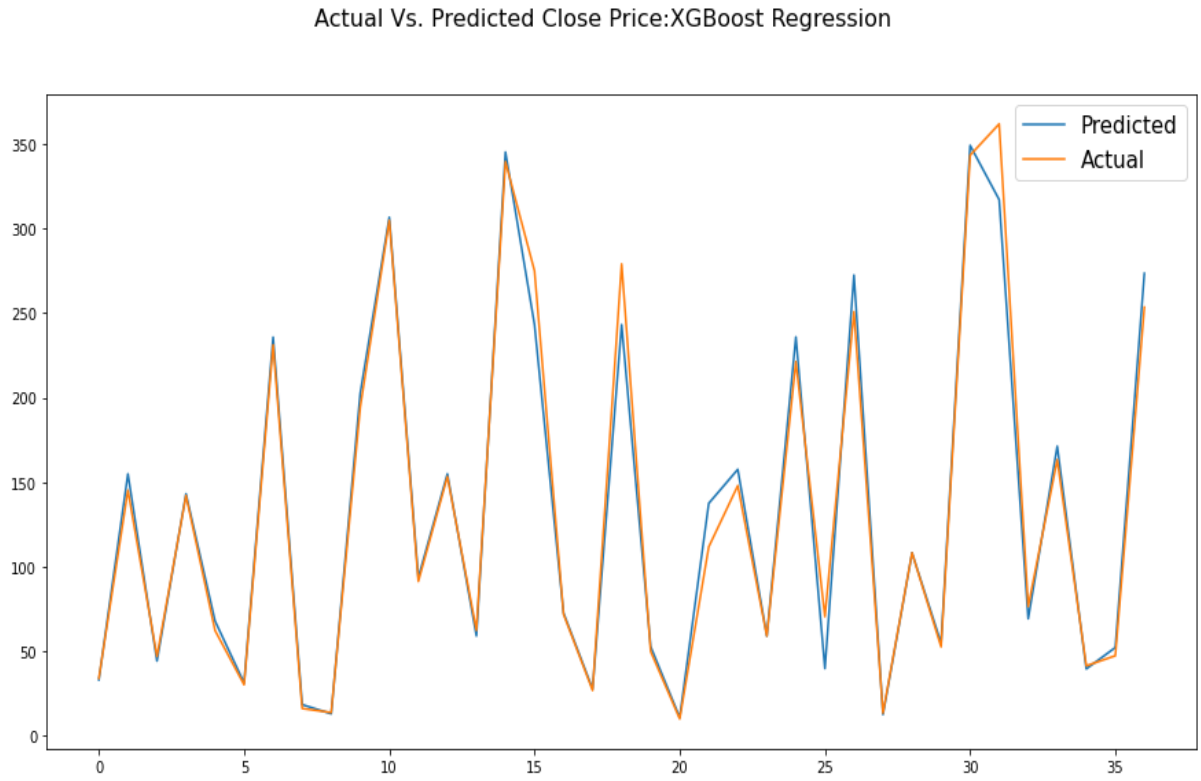


After Cross Validation:



➢ XG Boost Regression:

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.



Conclusions:

- Stock Price prediction with the help of Machine Learning models is less time consuming and also it gives good performance.
- Stock price was continuously increasing till 2018 after that it decreases due to fraud case of Rana Kapoor.
- All independent variables (Open, High& Low) are extremely correlated with dependent variable (Close).
- All independent variables are highly correlated with each other (Multicollinearity).
- Distribution of all independent and dependent variables was right skewed and after log transformation it became Normal.

- I have compared 5 models (**Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, and XG Boost Regression**) on the basis of **RMSE** and **MAPE**.
- **RMSE** and **MAPE** are mostly used as evaluation metrics to measure **forecast accuracy**.
- **XG Boost Regression** is best model among all five models with lowest **RMSE=0.0518**, and **MAPE=0.0174** than other models and also it has highest **r2 score** (**r2 score=0.9856**) than other models.