CAPSTONE PROJECT -2

YES BANK STOCK CLOSING PRICE PREDICTION

Presented By:

Minal Kharbade
Deveshya Gupta

CONTENT:

- Problem Statement
- Introduction
- Data cleaning
- Exploratory data analysis(EDA)
- Splitting different data
- Fitting different model
- Cross validation and hyperparameter
- conclusion

PROBLEM STATEMENT:

- Prediction of Yes bank stock closing price.
- Perform regression analysis using multiple models to predict the closing price of the stock and compare evaluation metrics for all of them to find best model.
- Getting accuracy of several machine learning model.

INTRODUCTION:

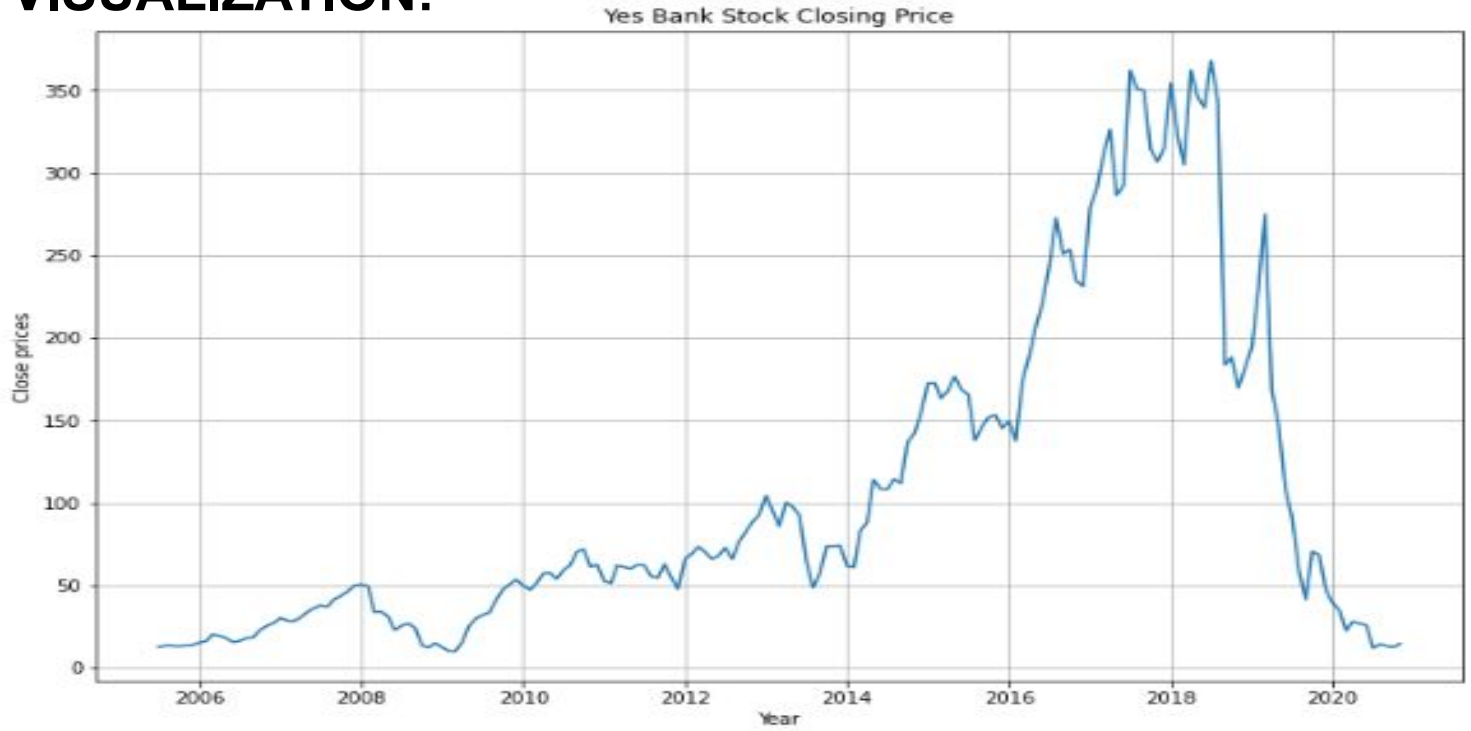
- Stock market is characterized as unpredictable and non linear in nature.
- Data set of yes bank stock prices contains observation regarding open ,close,high and low prices of yes bank stock from july 2005-november 2020.
- Date: monthly observation of stock prices since its inception.
- Open: The price of stock when stock market is open for the day.
- Close: The price of stock when stock market is close for the day.
- High: maximum price of stock maintained during given period of time.
- Low: minimum price of stock maintained during given period of time.

DATA CLEANING

- Dataset contains 5 column and 185 rows.
- Checking null values.
- Check duplicate values.
- Correction of datatype as per requirement.
- Checking outliers.

EXPLORATORY DATA ANALYSIS

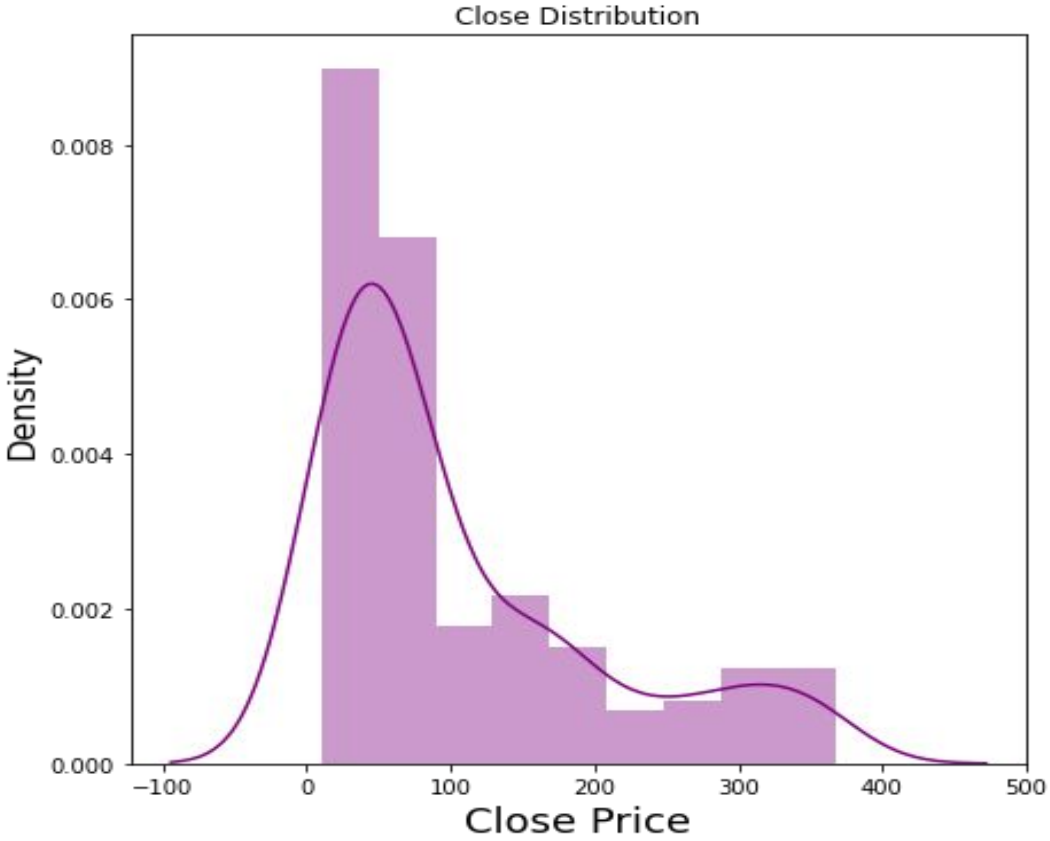
- **DATA VISUALIZATION:**



Since 2010 to 2018 the trend is increasing but after 2018 it decreases because of the fraud case involving Rana Kapoor.

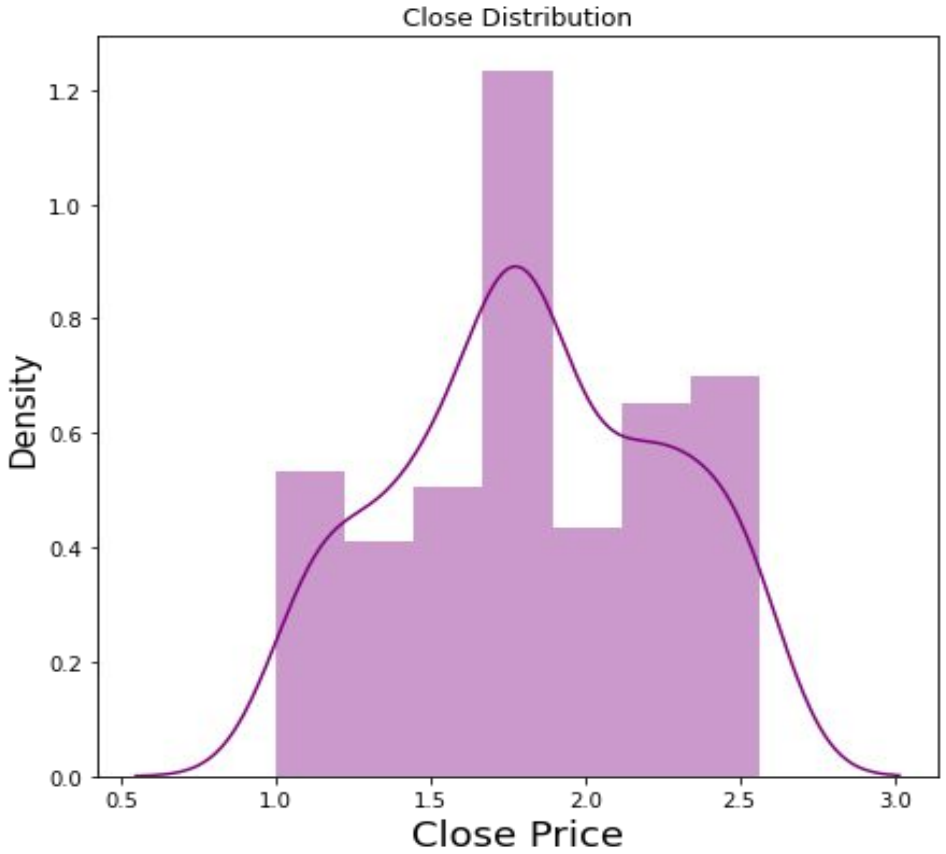
- **Distribution of closing price:**

- **We can see distribution of closing price is right skewed.**
- **We have to apply log algorithm to make it normal distribution.**

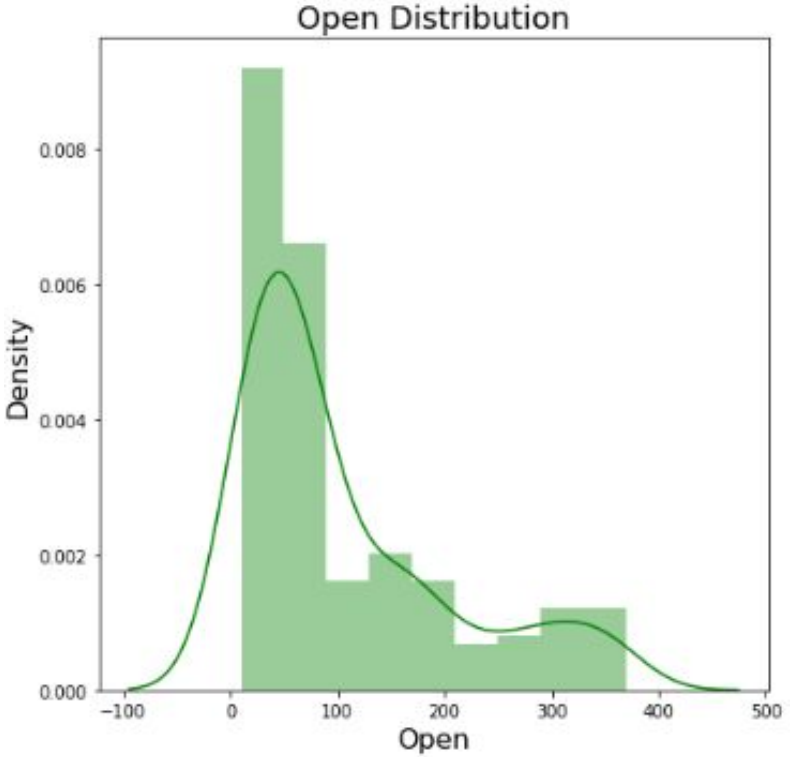


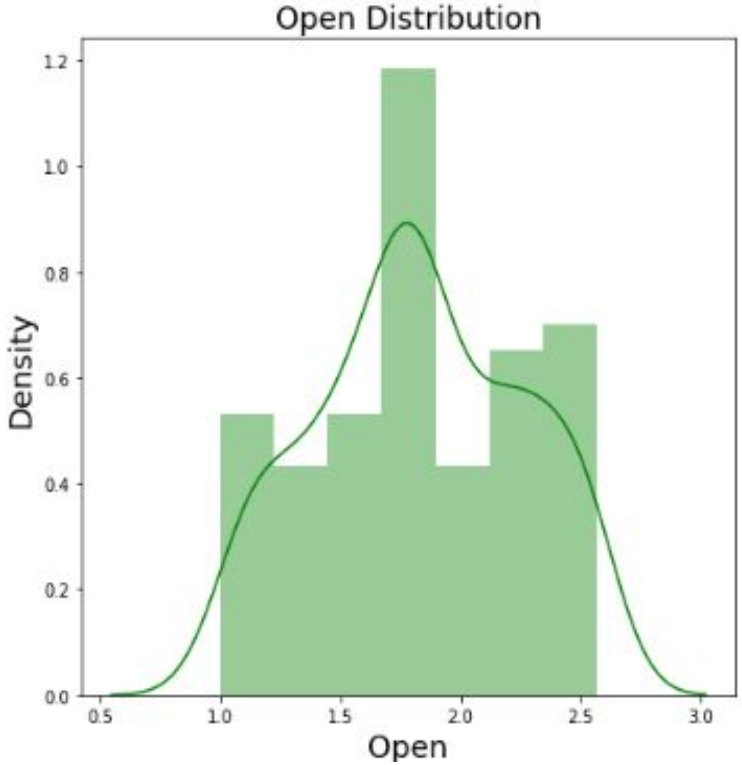
- **Distribution of closing price(<u>after log transformation</u>):**

- **After applying log algorithm distribution of closing price is normal distribution.**

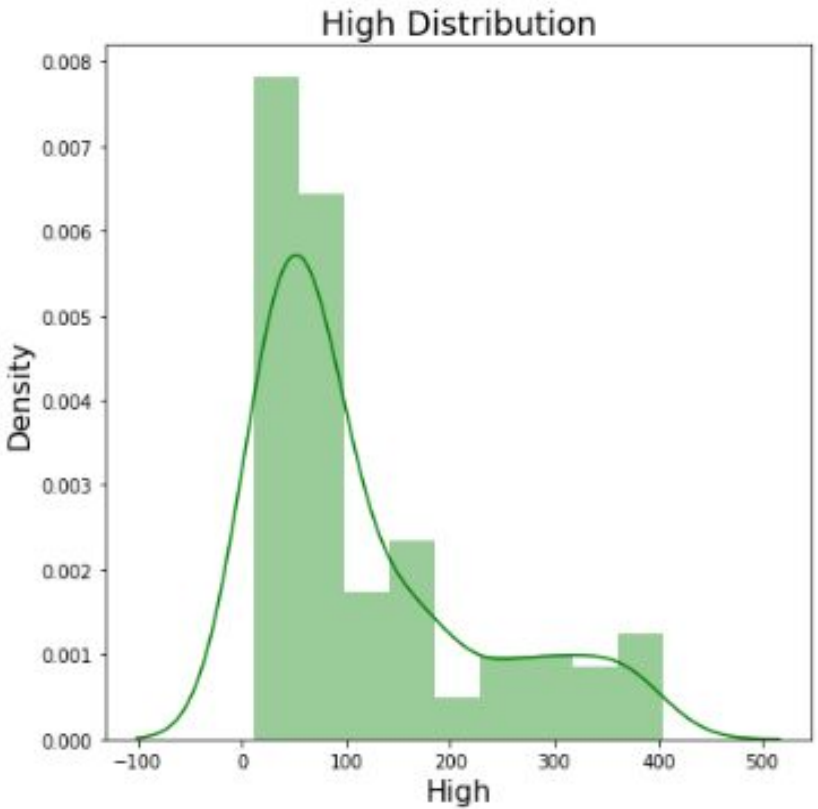


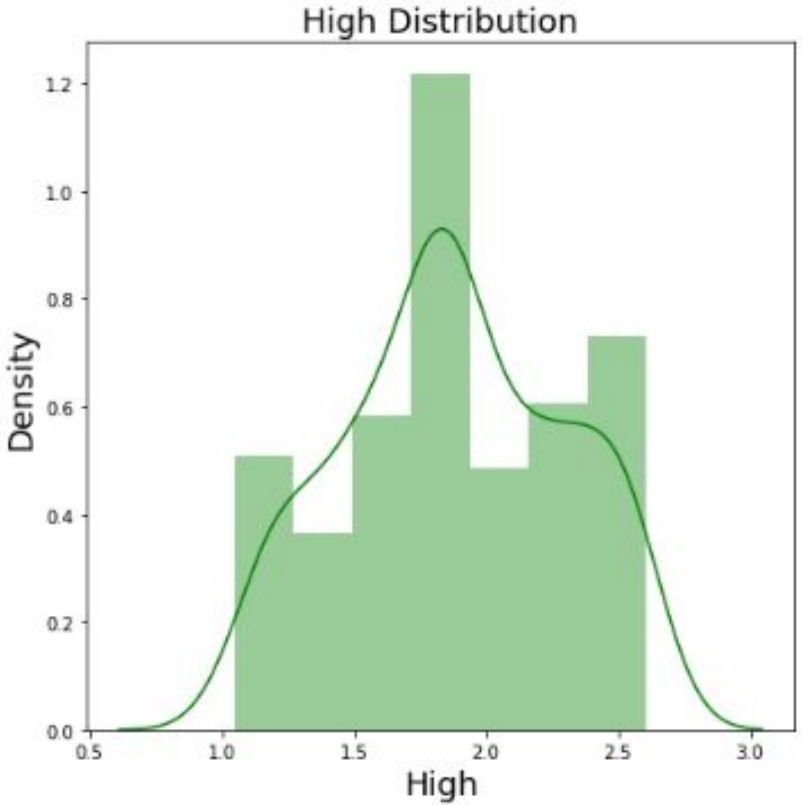
- **Distribution of open ,high and low price of stock:**
  (Before and after applying log algorithm)



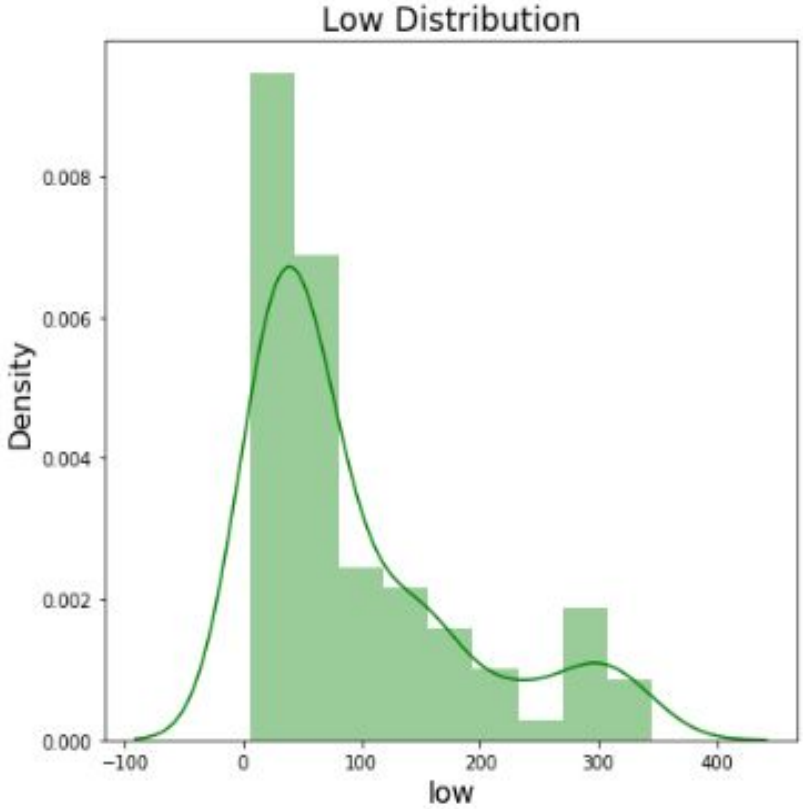


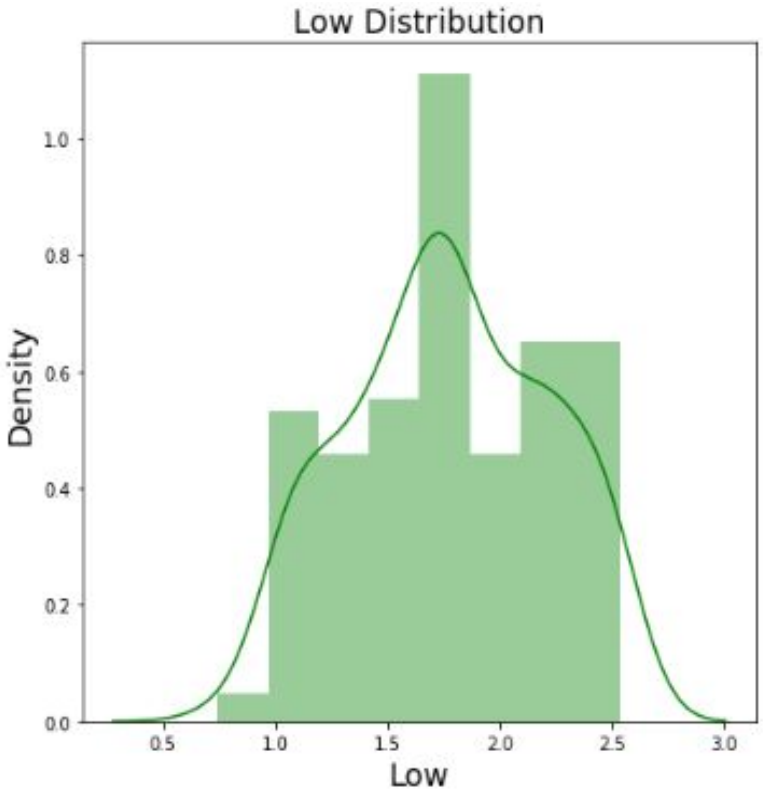
Distribution of opening price is also right skewed.after applying log algorithm distribution is normal.





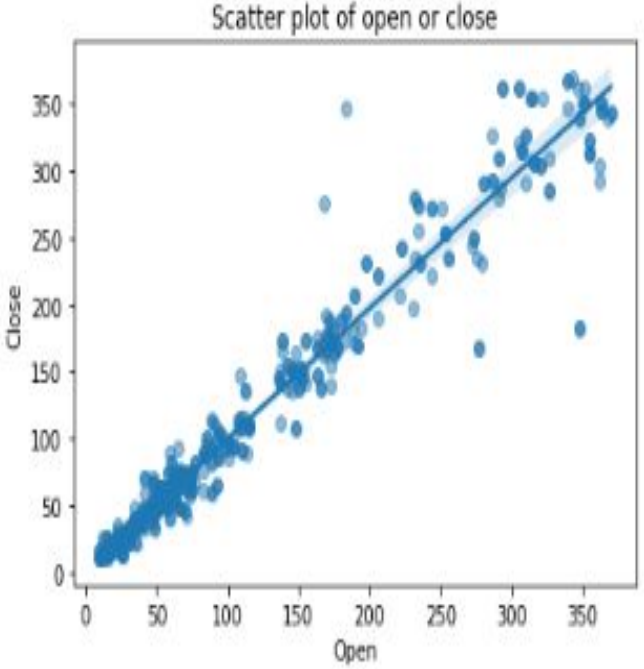
Distribution of high price is also right skewed.after applying log algorithm distribution is normal.

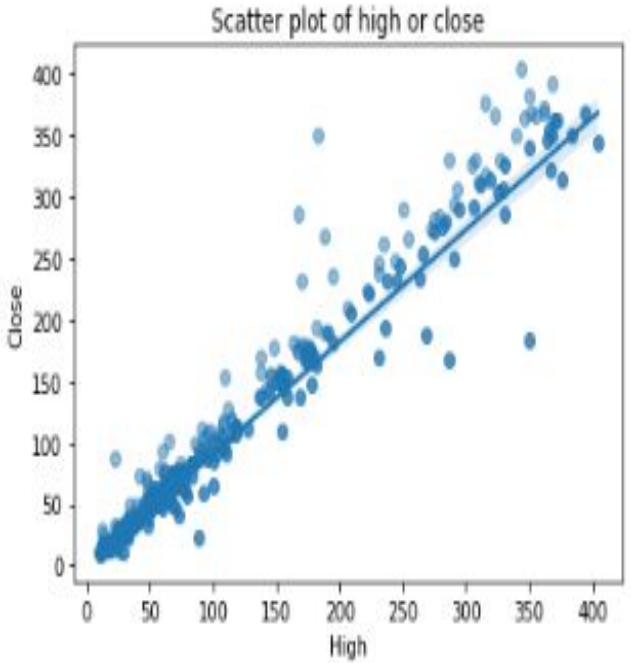


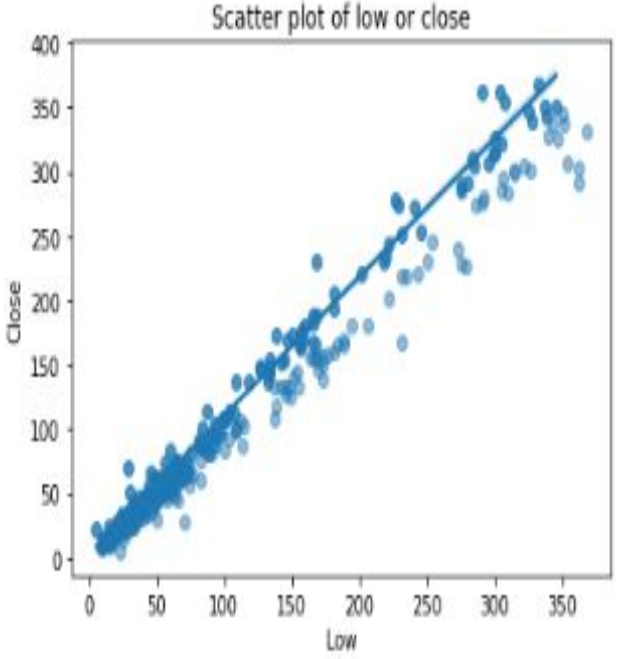


Distribution of Low price is also right skewed.after applying log algorithm distribution is normal.

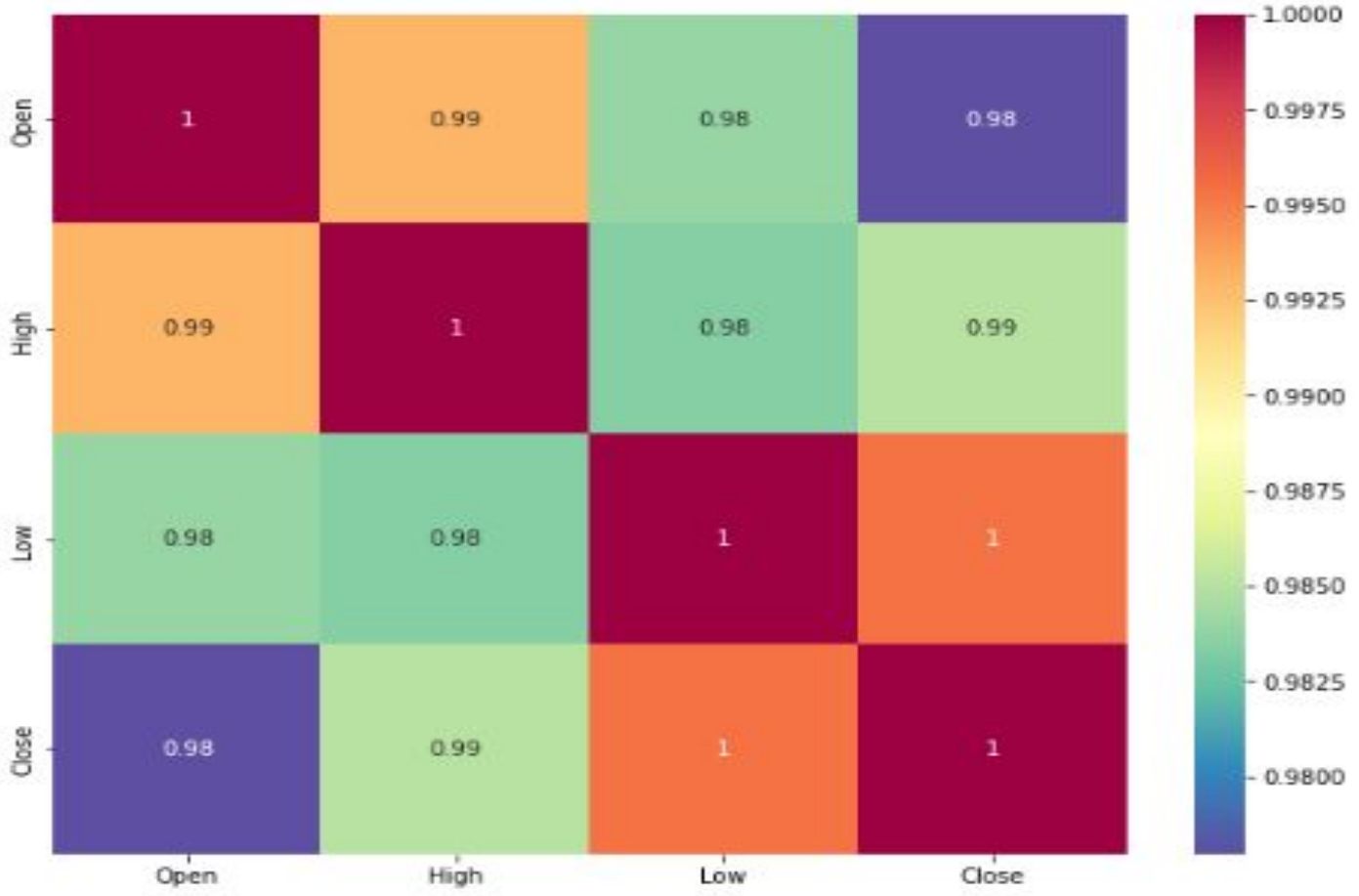
- **Scatterplot:**

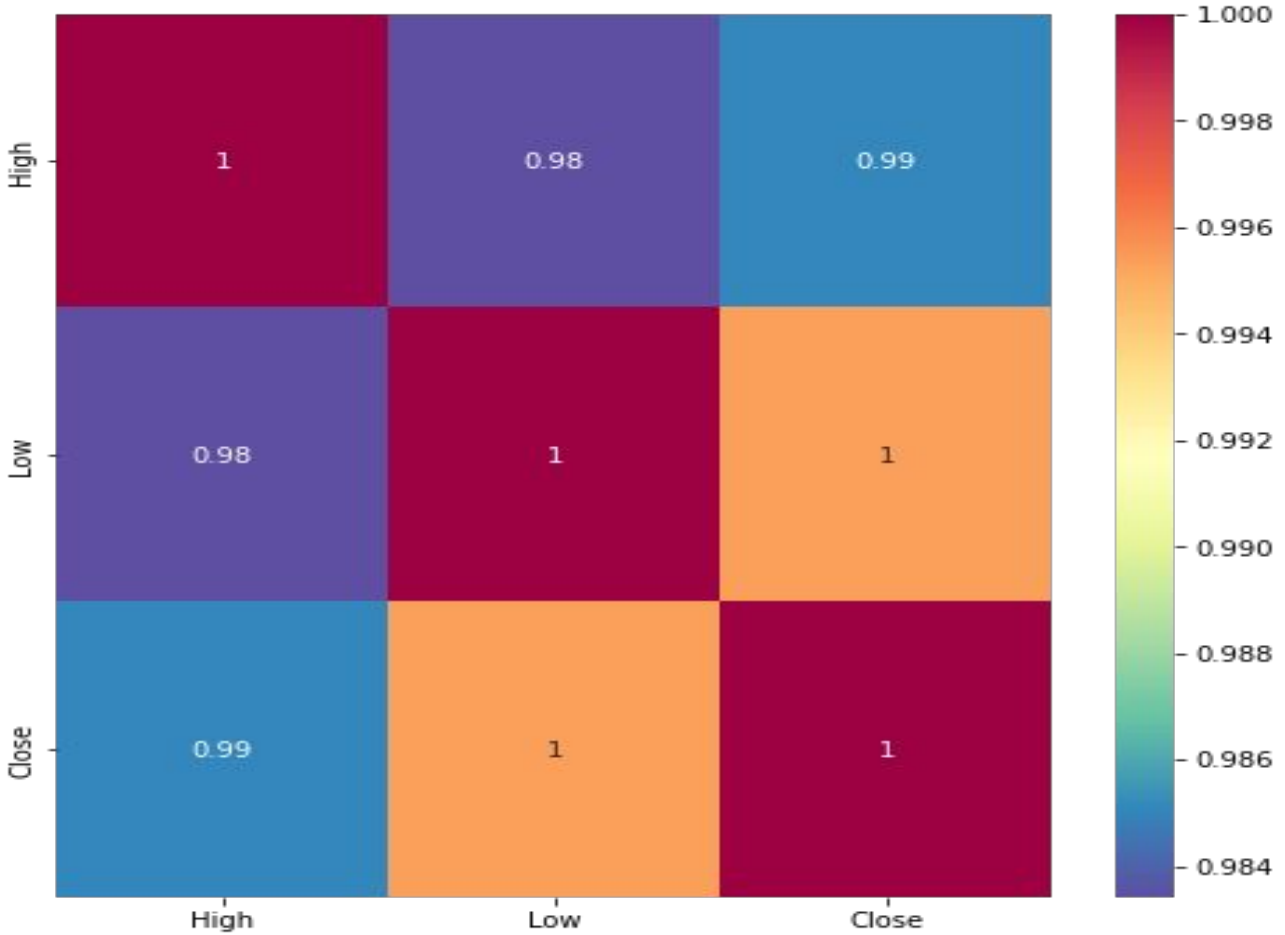






- **Correlation Analysis:**

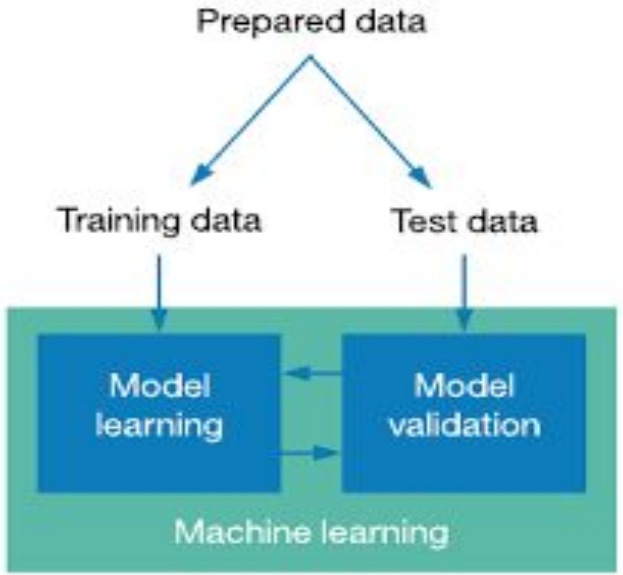




All the variables are highly correlated to each other

SPLITTING OF DATA

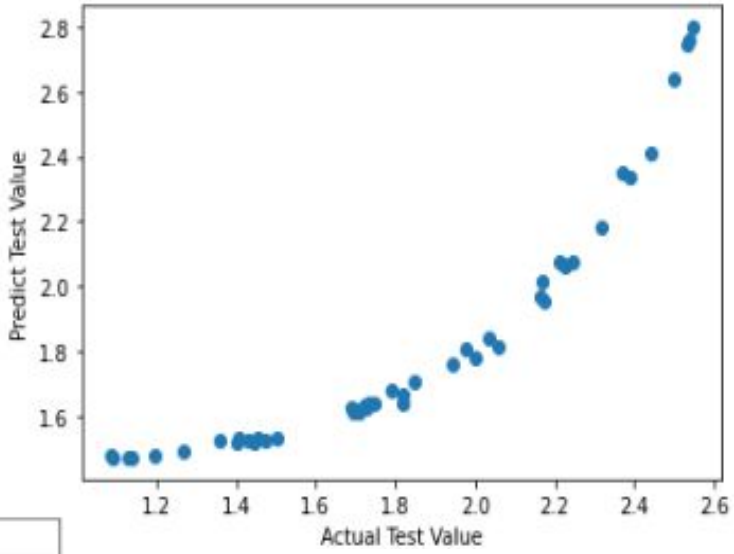
- We split our data for training and testing the model.
- Training dataset is for making algorithm learn and train model.
- Test dataset is for testing the performance of train model.
- Here 75% data is taken as training dataset and remaining 25% is taken for testing purpose.

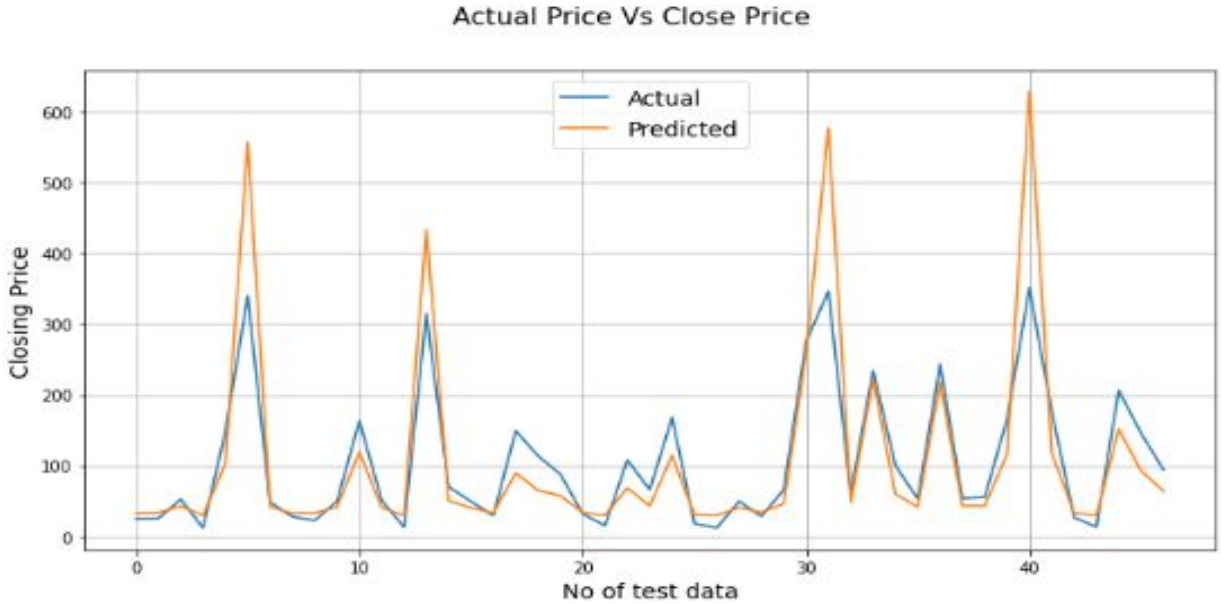


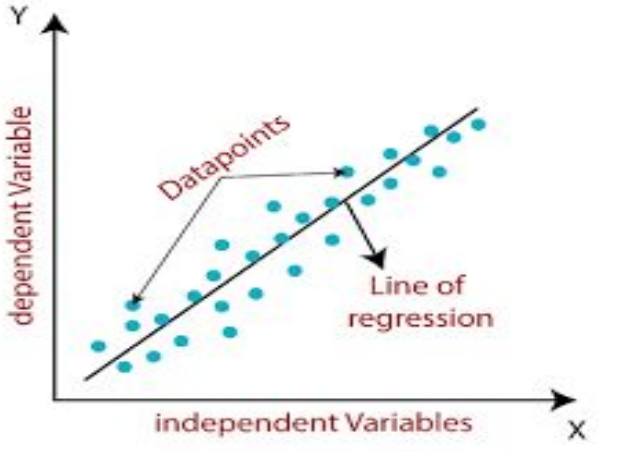
FITTING DIFFERENT MODEL

Linear Regression Model:

- Linear regression is one of the easiest and most popular machine learning algorithm.
- Linear regression is used to predict the value of variable on the value of another variable.

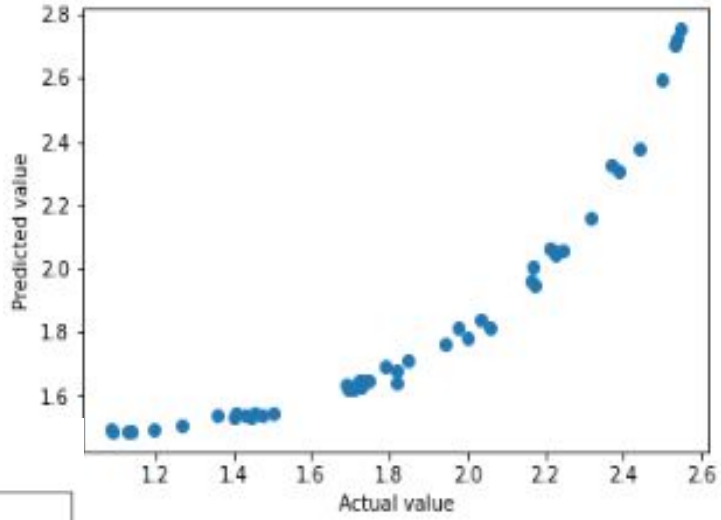


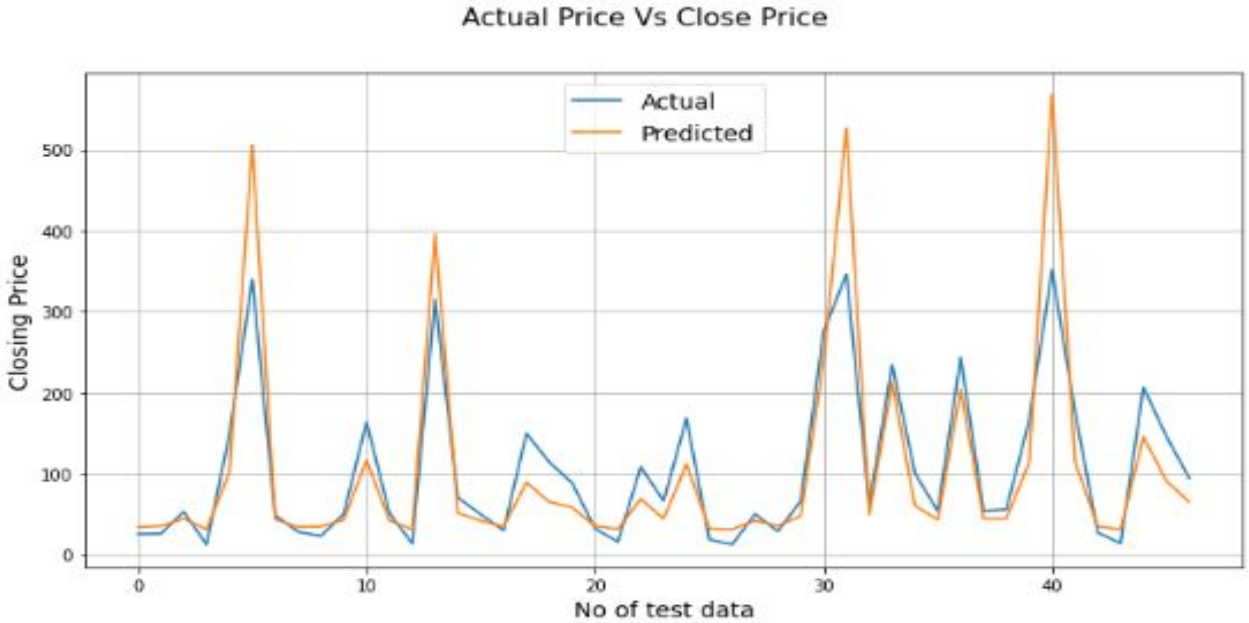




Lasso Regression Model:

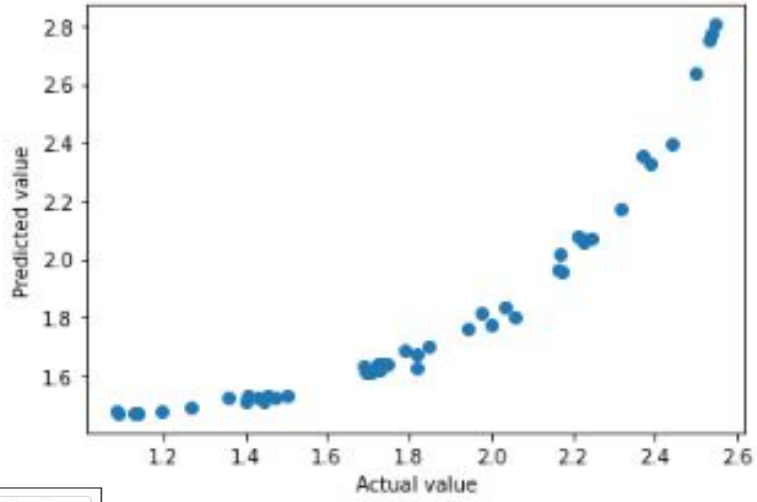
- Lasso means Least Absolute Shrinkage and Selection operator.
- Lasso regression is a type linear regression that usage shrinkage

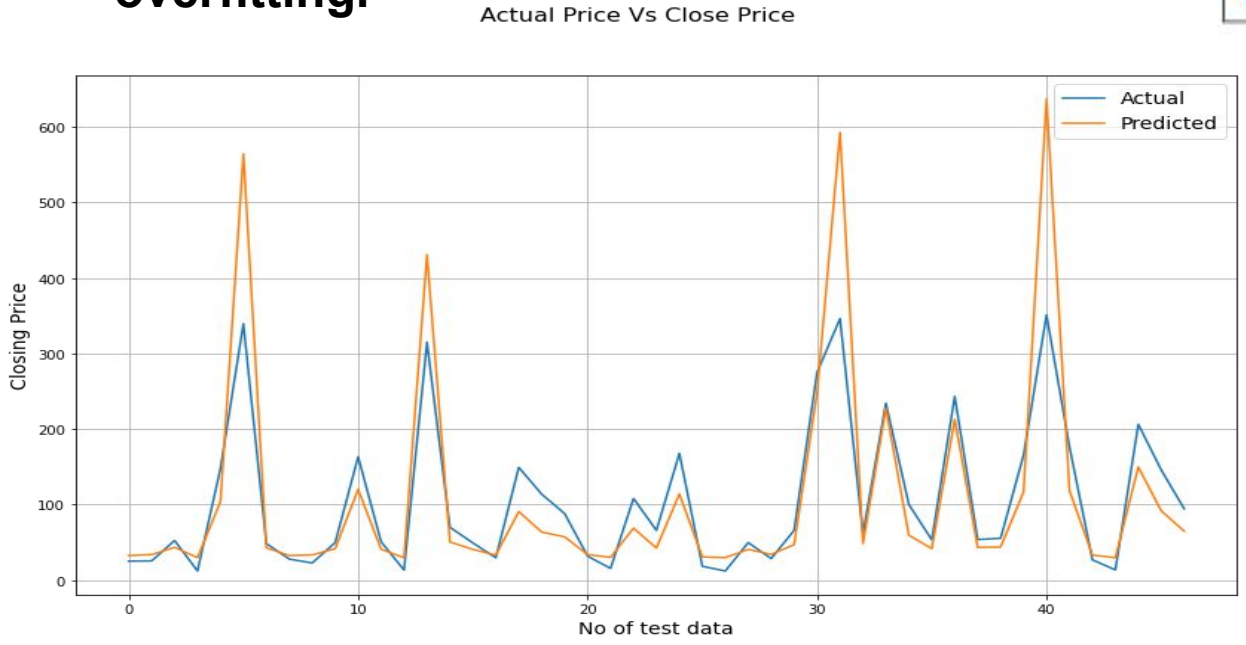




Ridge Regression Model:

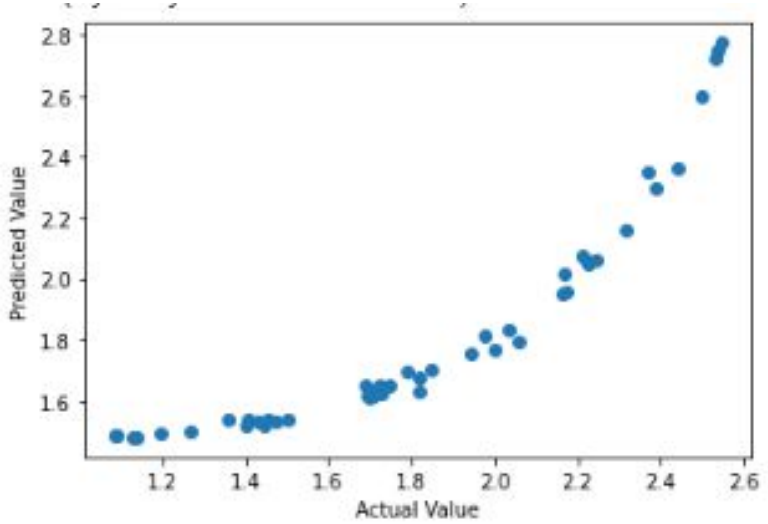
- Ridge regression is a method of estimating coefficients of multiple regression model in scenarios where linearly independent variables are highly correlated.
- Ridge regression solves the problem of overfitting.

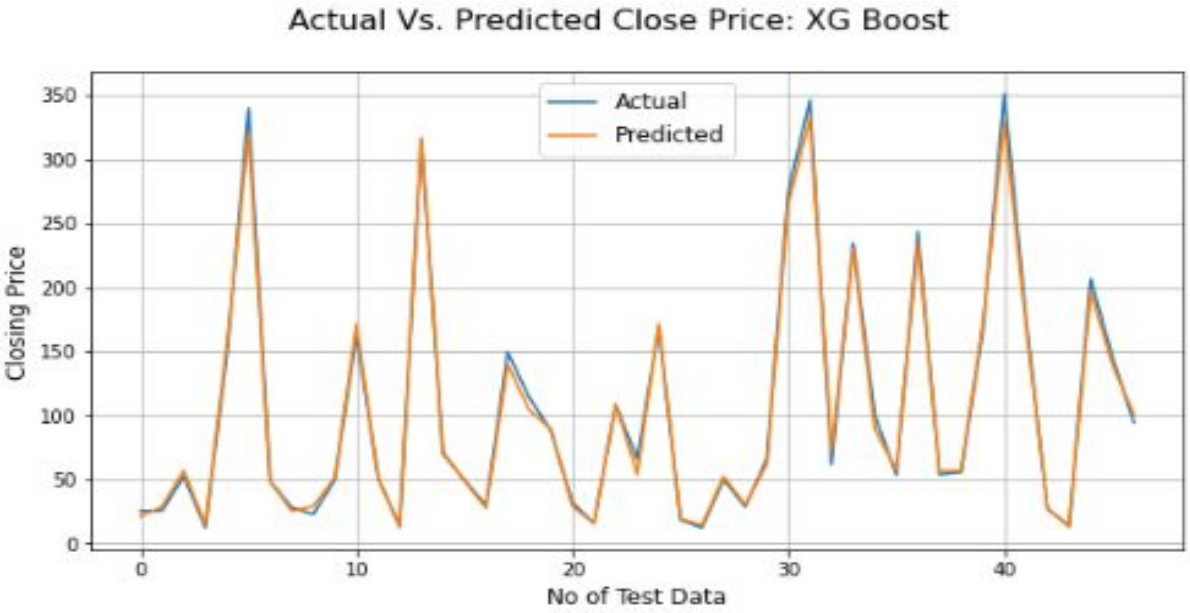




XG Boost Regressor:

- XG boost which stands for extreme gradient boosting, is scalable , distributed gradient boosted decision tree(GBDT) machine learning library.
- It provide parallel tree boosting to solve many data science in fast and accurate way.





Cross validation and tuning on XG boost regression

MSE	MSE	MAE	MAPE	R2
0.001	0.037	0.028	0.018	0.993

CROSS VALIDATION AND HYPERPARAMETER

- **Cross validation is the process of training learners one set of data and testing it using different set.**
- **Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.**

Evaluation Metric

Cross validation and tuning on Linear Regression:

MSE	MSE	MAE	MAPE	R2
0.032	0.18	.0156	0.096	0.821

Cross validation and tuning on Lasso regression:

MSE	MSE	MAE	MAPE	R2
0.033	0.181	0.157	0.096	0.819

Cross validation and tuning on Ridge regression:

MSE	MSE	MAE	MAPE	R2
0.033	0.182	0.158	0.097	0.816

CONCLUSION:

- The techniques we use for prediction is not only help researcher to predict the future stock closing price or and fraud but also helps investors or any person who dealing with the stock market in order to prediction of model with good accuracy.
- Xgboost regression is best model for yes bank stock closing price data this model use for further prediction.
- In this we use Linear Regression, Lasso Regression and Ridge Regression techniques and these gives us following result.
- Independent variables (High, Low and Open) are directly correlated with Dependent variable (Closing Price).

THANK YOU