Capstone Project

Yes bank Stock Closing Price Prediction

Presented By:
Aman Verma

Content

- 1. Problem Statement
- 2. Introduction
- 3. Data Cleaning
- 4. Exploratory Data Analysis (EDA)
- 5. Transforming Data
- 6. Splitting Data
- 7. Fitting Different Model
- 8. Cross Validation & Hyperparameter Tuning
- 9. Conclusion

Problem Statement

- Perform regression analysis using multiple models to predict the closing price of the stock and compare the evaluation metrics for all of them to find the best model.

- Prediction of Yes Bank stock closing price.

- Getting accuracy score of several machine learning model.

Introduction

1 Data set - data Yes Bank Stock Prices - contains observations regarding open,close, high and low prices of the yes bank stock from July 2005 - November 2020.

2 Date: Monthly observation of stock prices since its inception.

3 Open: The price of a stock when stock exchange market open for the day.

4 Close: The price of a stock when stock exchange market closed for the day.

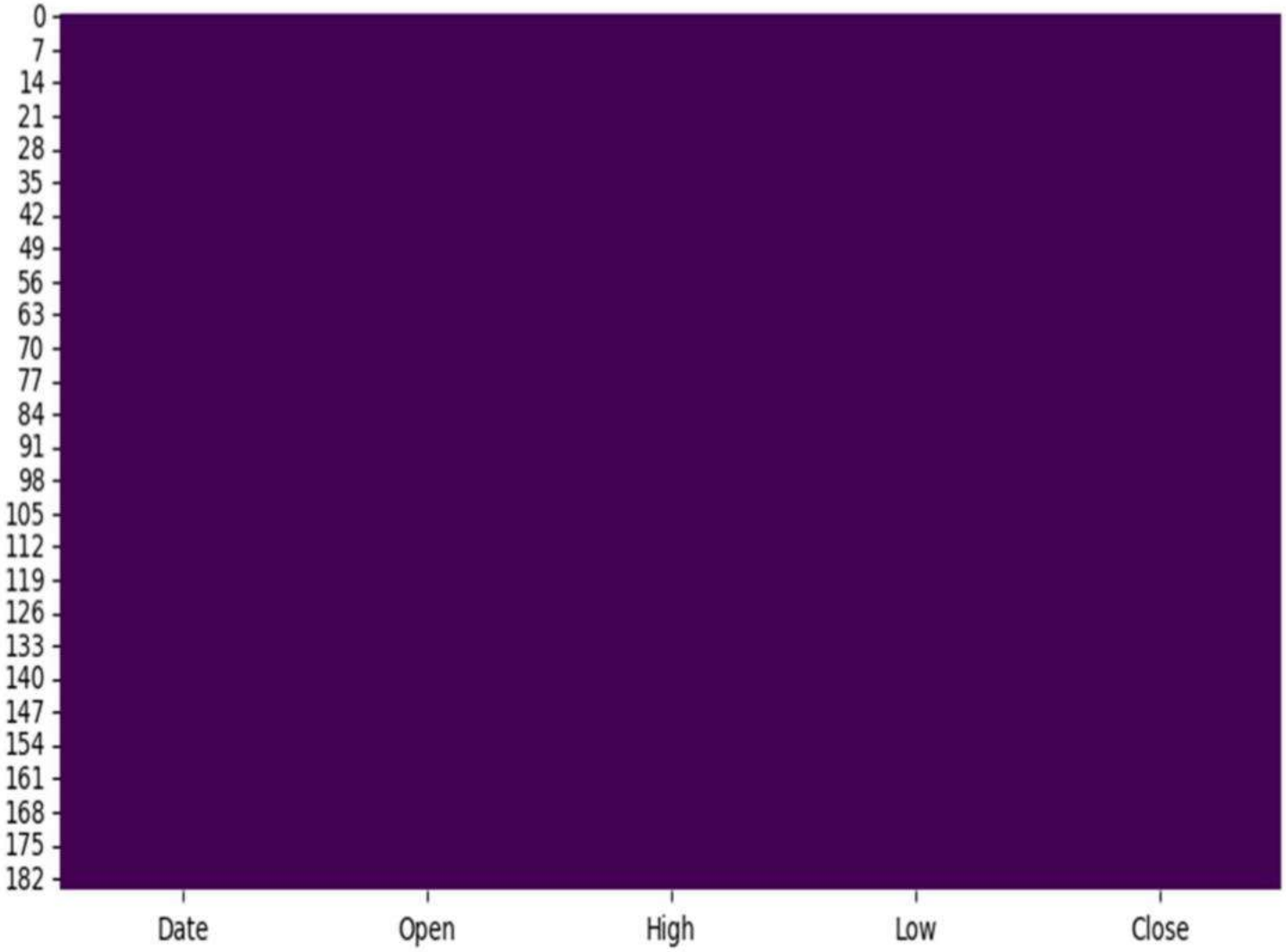
5 High: The maximum price of a stock attained during given period of time.

6 Low: The minimum price of a stock attained during given period of time.

Data Cleaning

- Null Values Treatment
- Duplicated Values Treatment
- Date Format Change
- Checking outliers
- So after successfully cleaning the dataset we have 185 columns and 5 rows

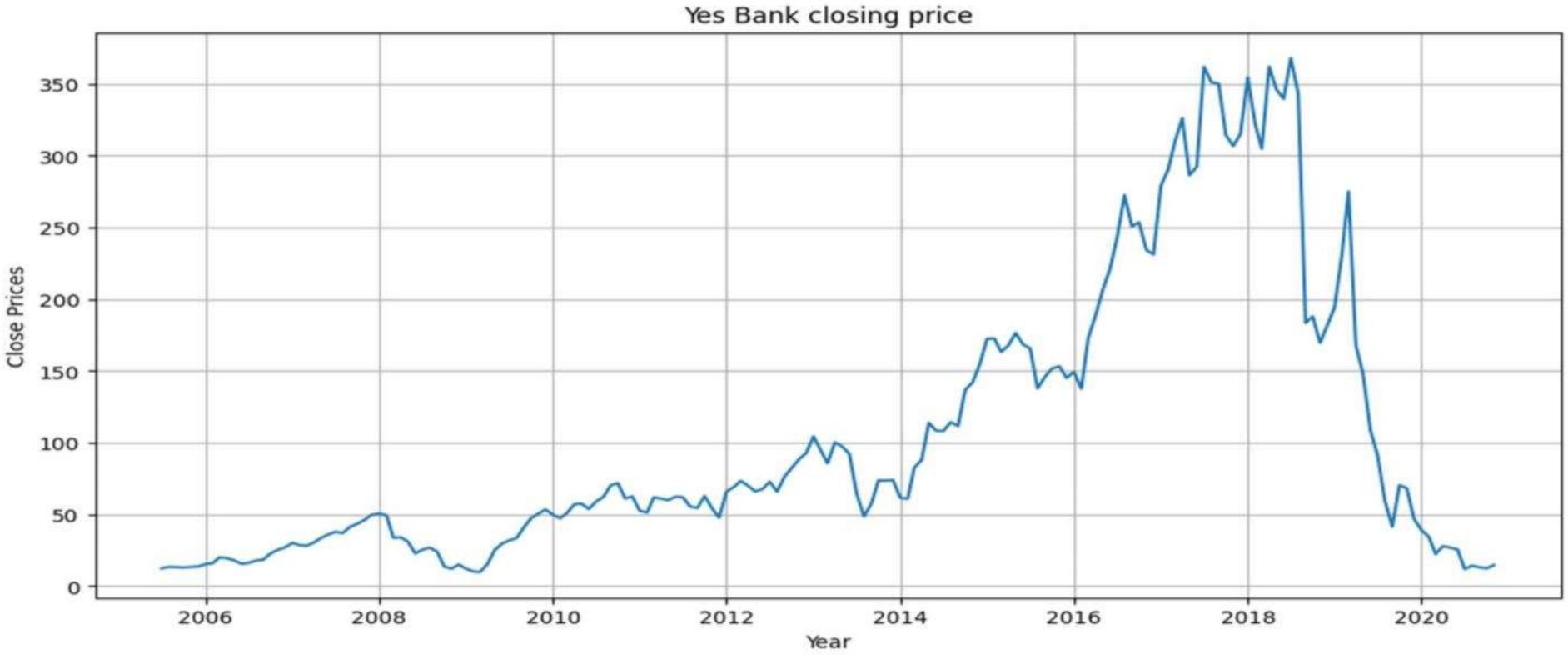
EDA (Checking NaN values)



No missing value present in our dataset.

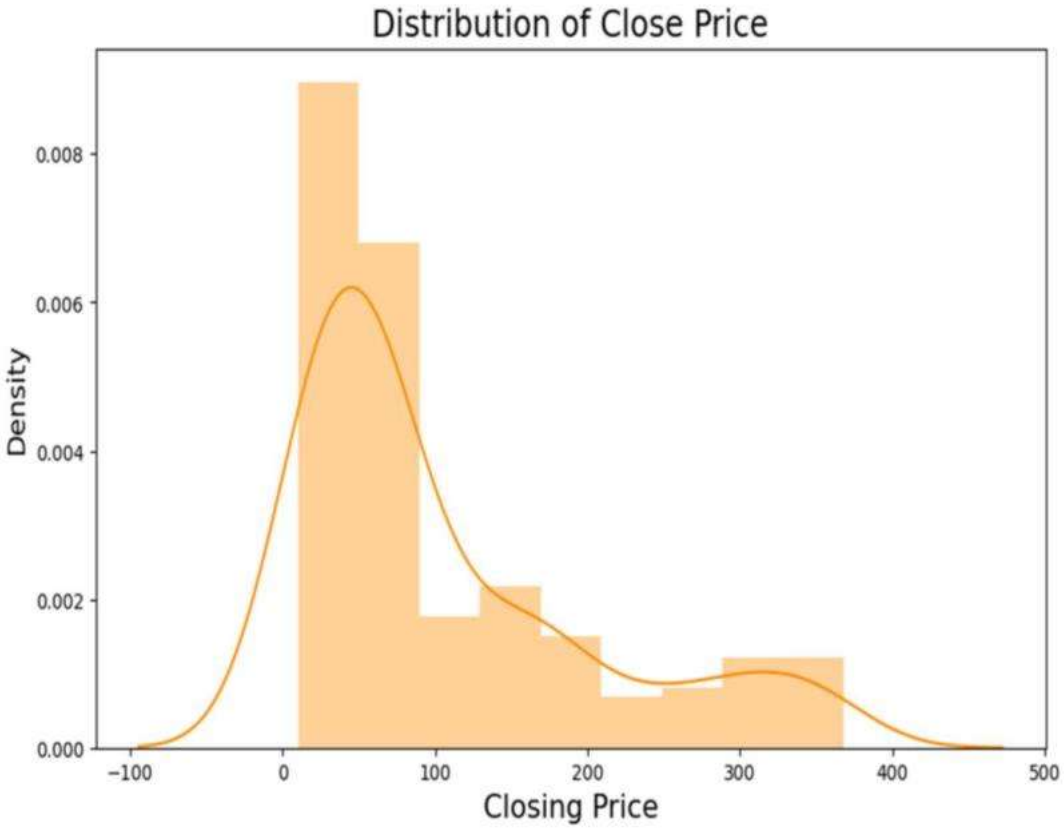
Exploratory Data Analysis

Visualising The Data



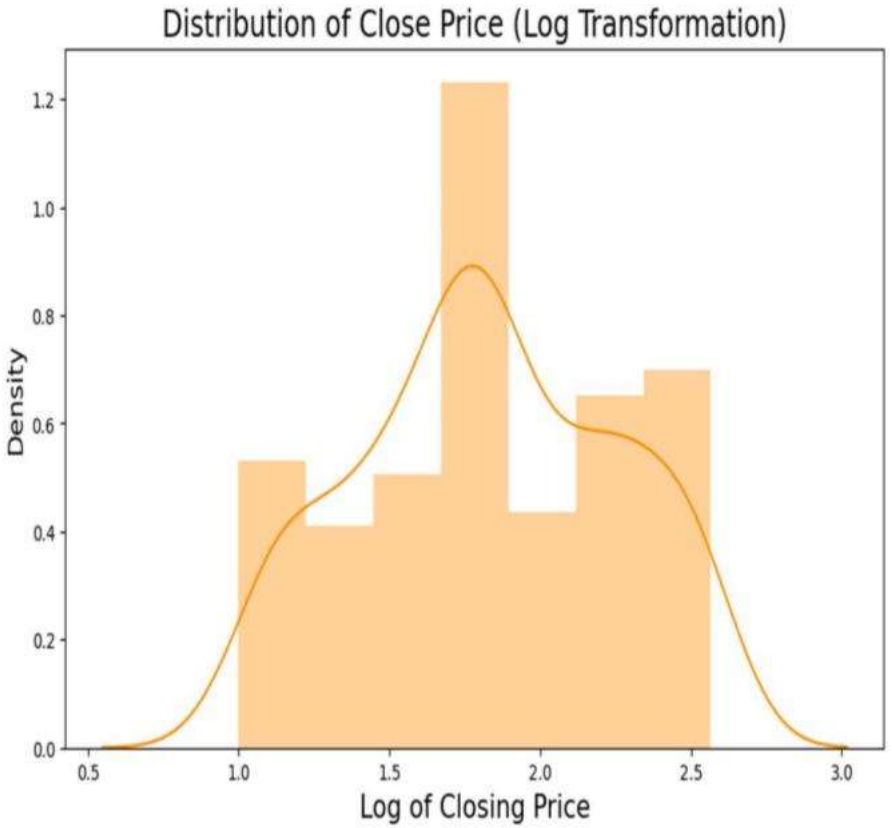
Here, it's clearly visible form the above plot that the stock prices saw a significant rise from year 2006 to 2018 . However, since 2018 the stock proces saw a major downfall and that is may be due to the fraud case.

Distribution of Closing Price



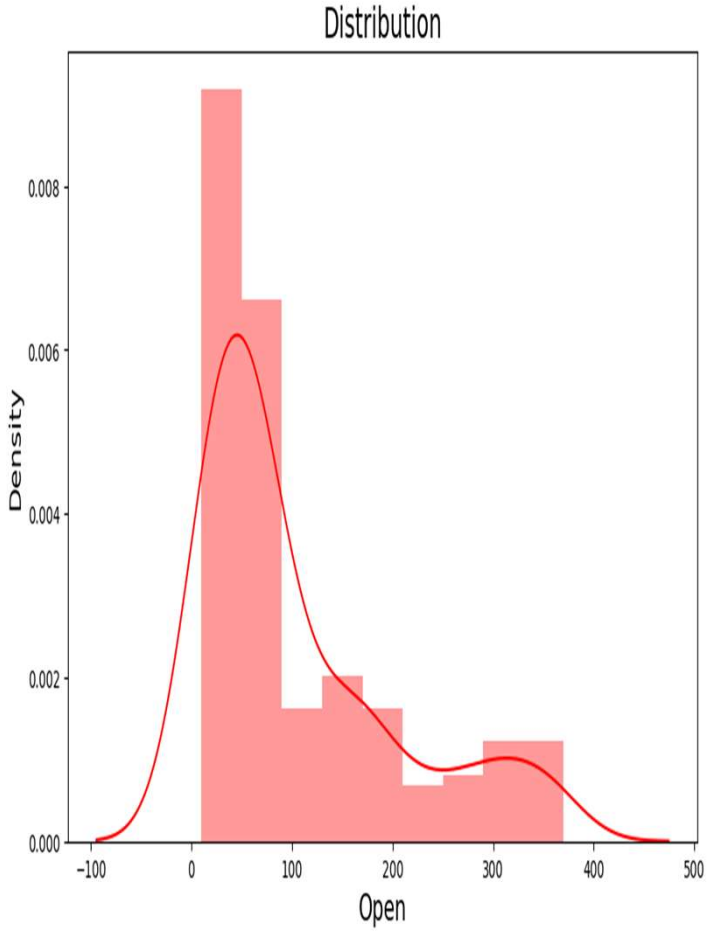
- Distribution of closing price is right skewed.
- We need this distribution to be normal distribution for training algorithm.

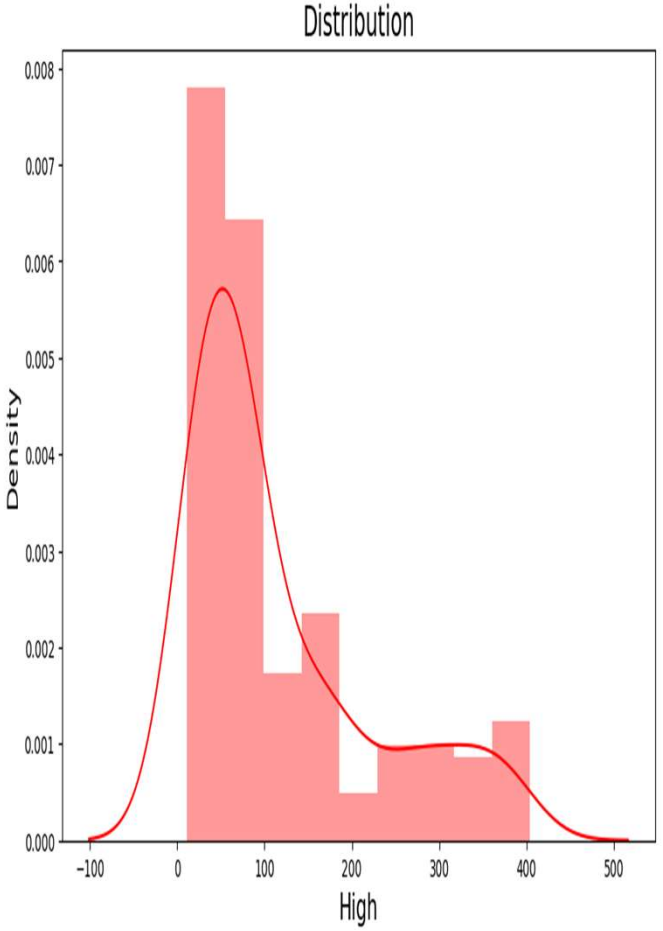
After Log Transformation

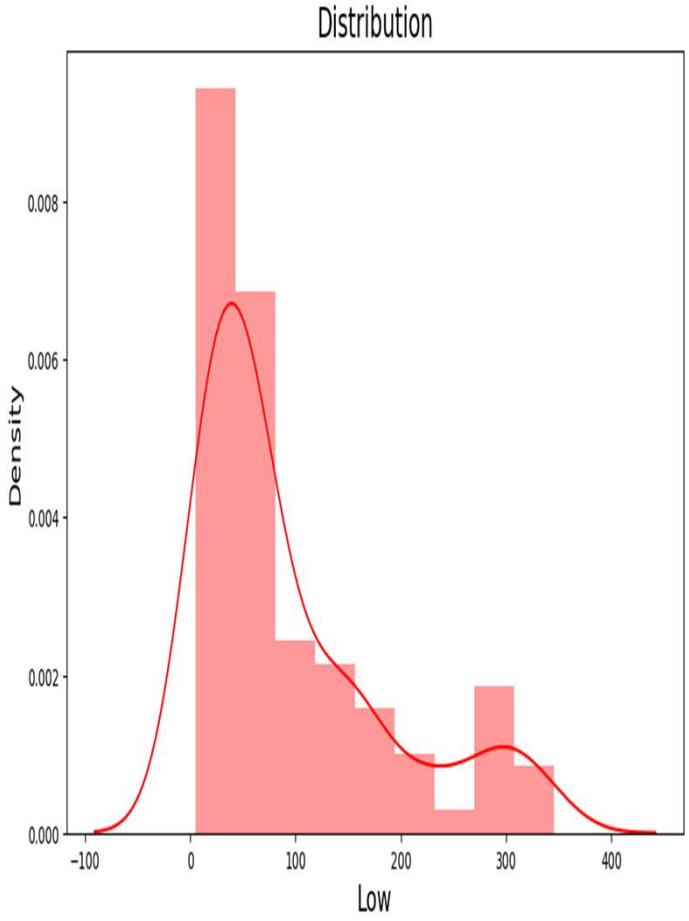


- Distribution of closing price is normal distribution.

Distribution of Open, High & Low Price of a stock

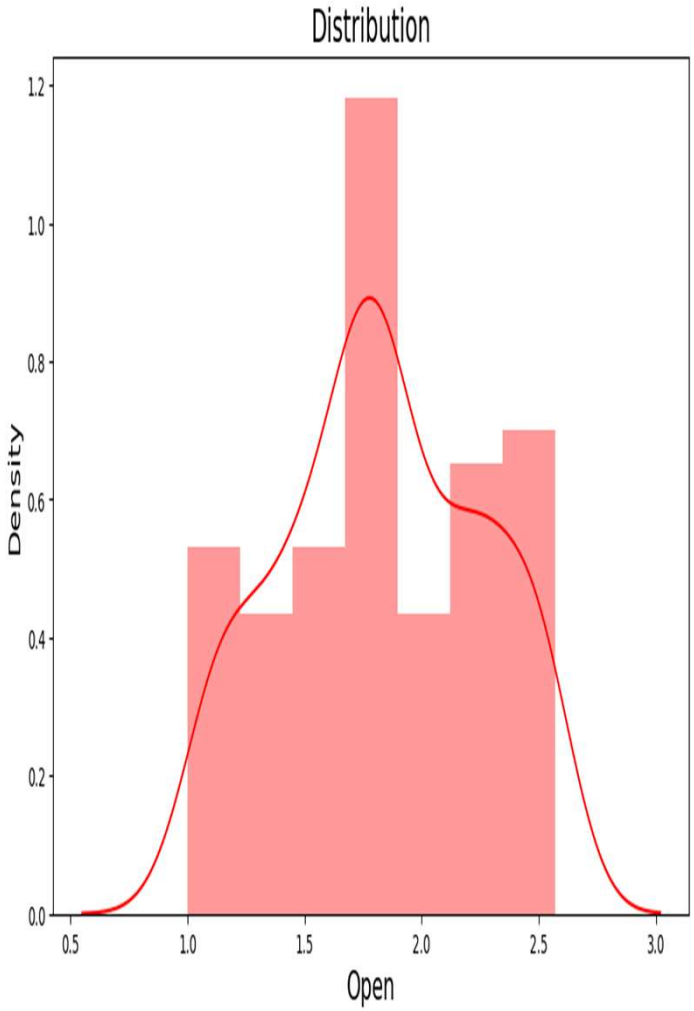


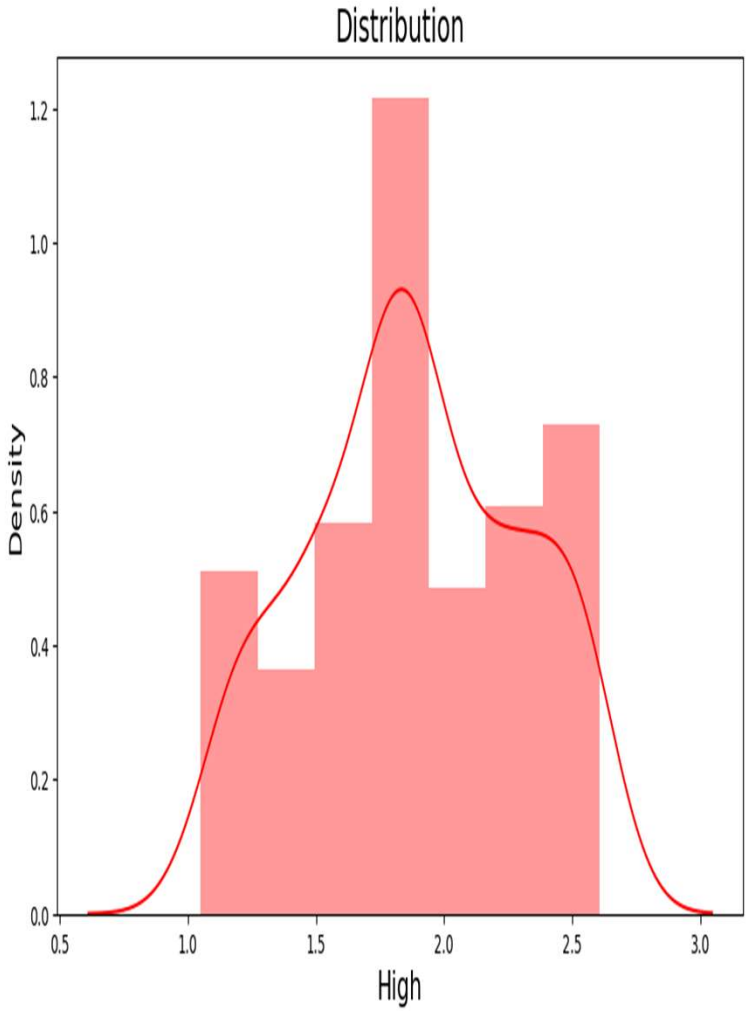


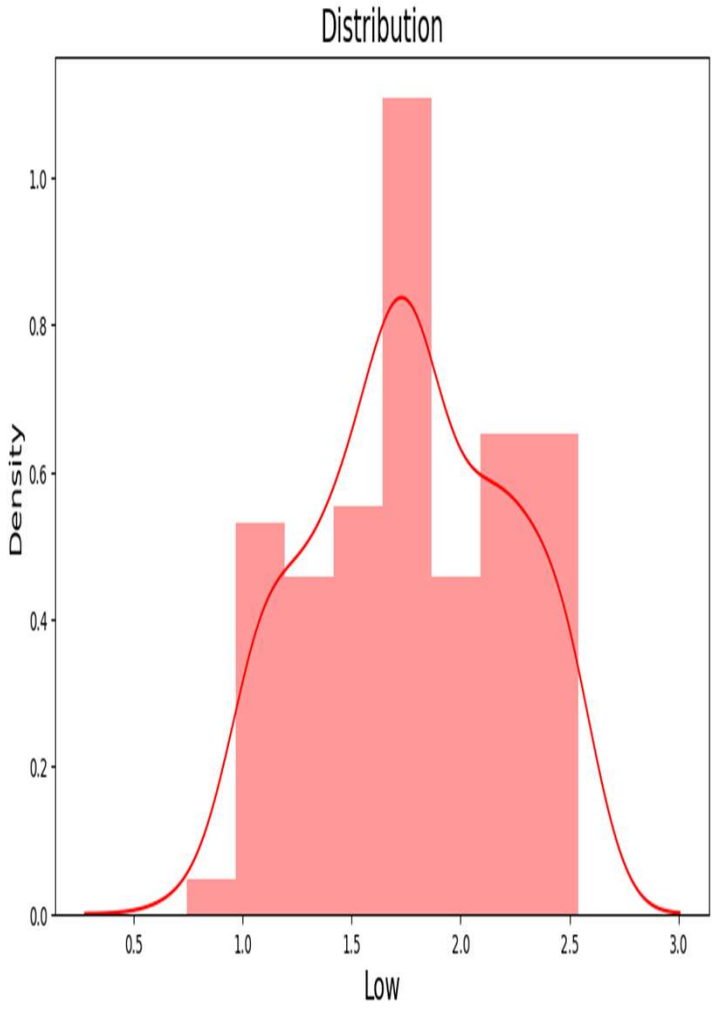


- Distribution of opening price, high price and low price are also right skewed.
- Log transformation applied to make this distribution normal.

Distribution of Open, High & Low Price of a stock after Log Transformation

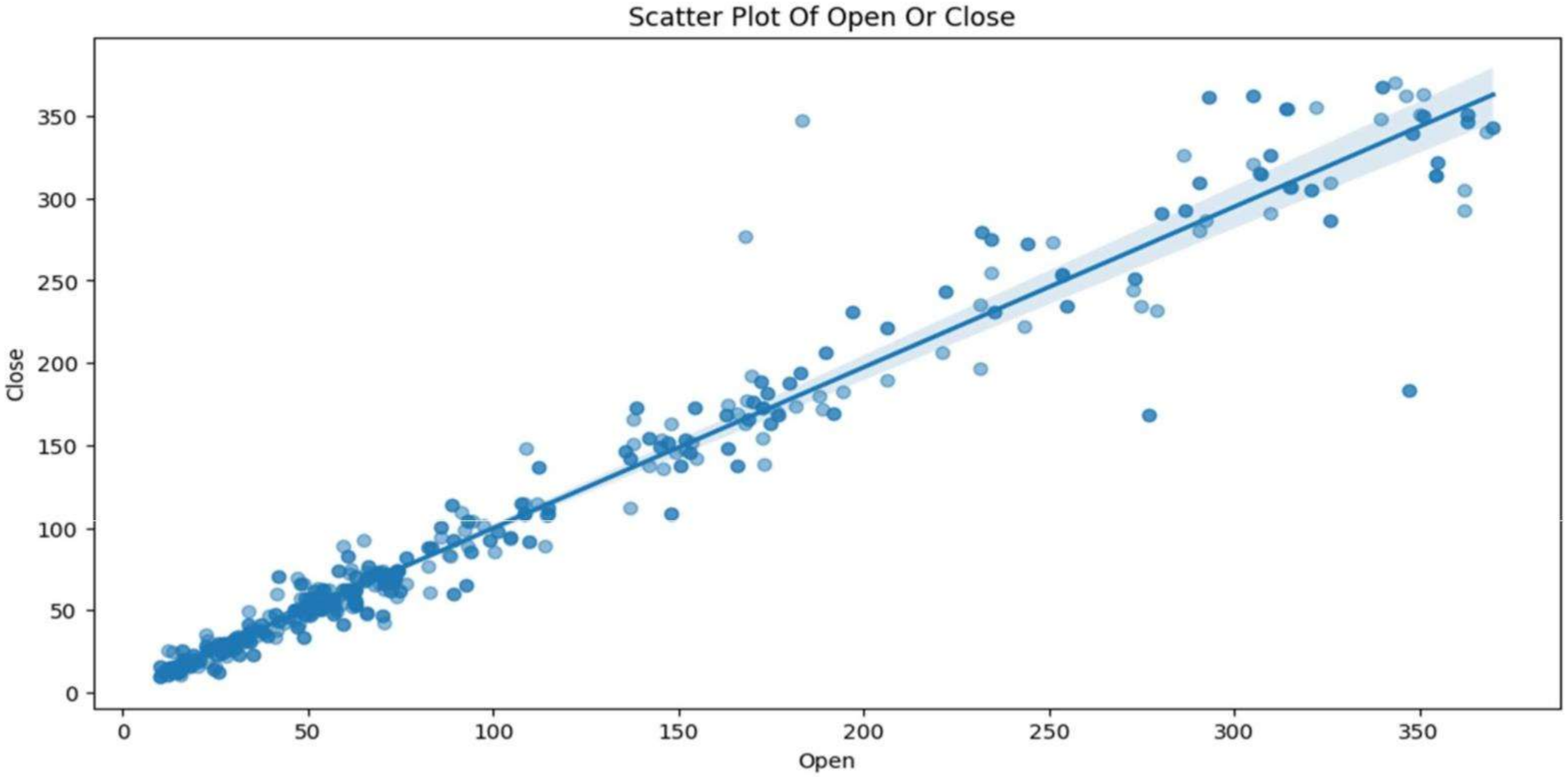


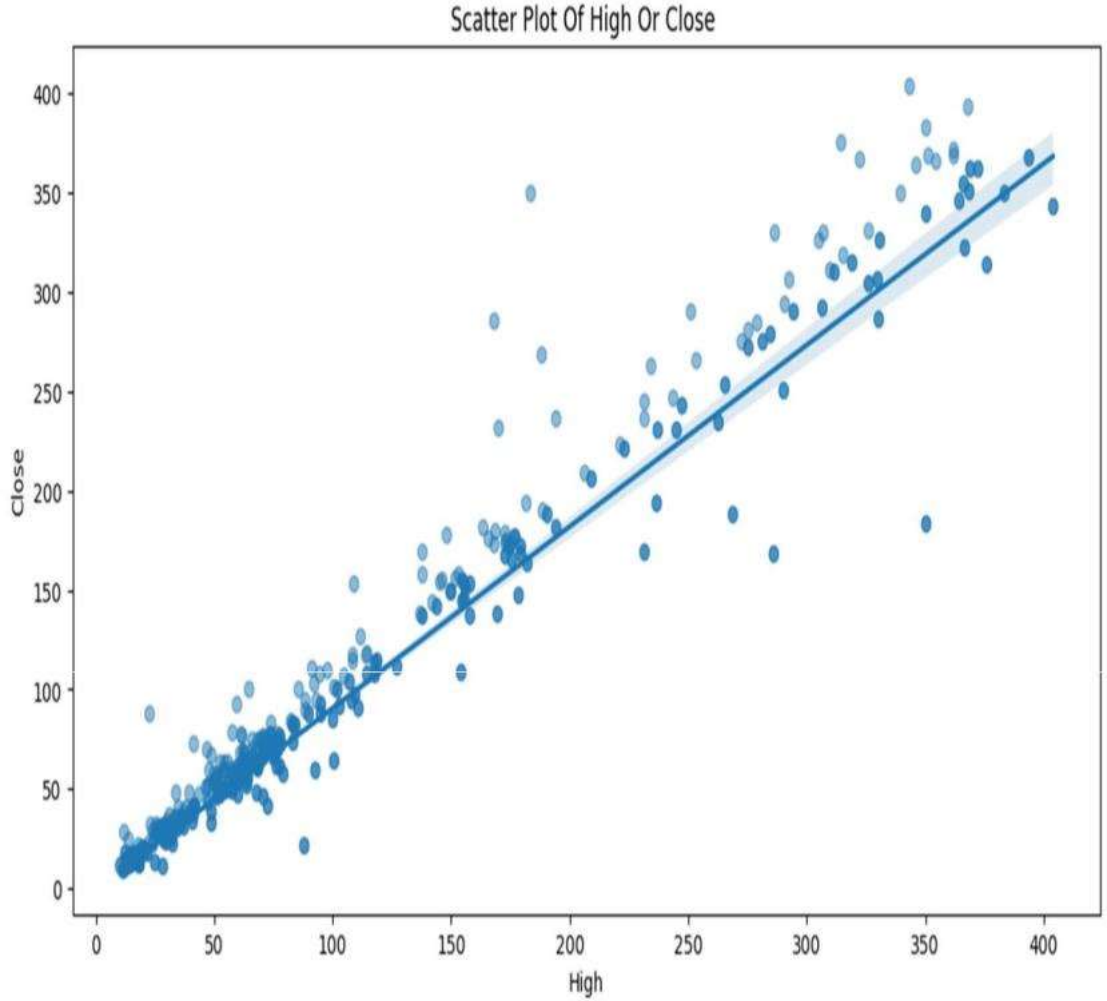


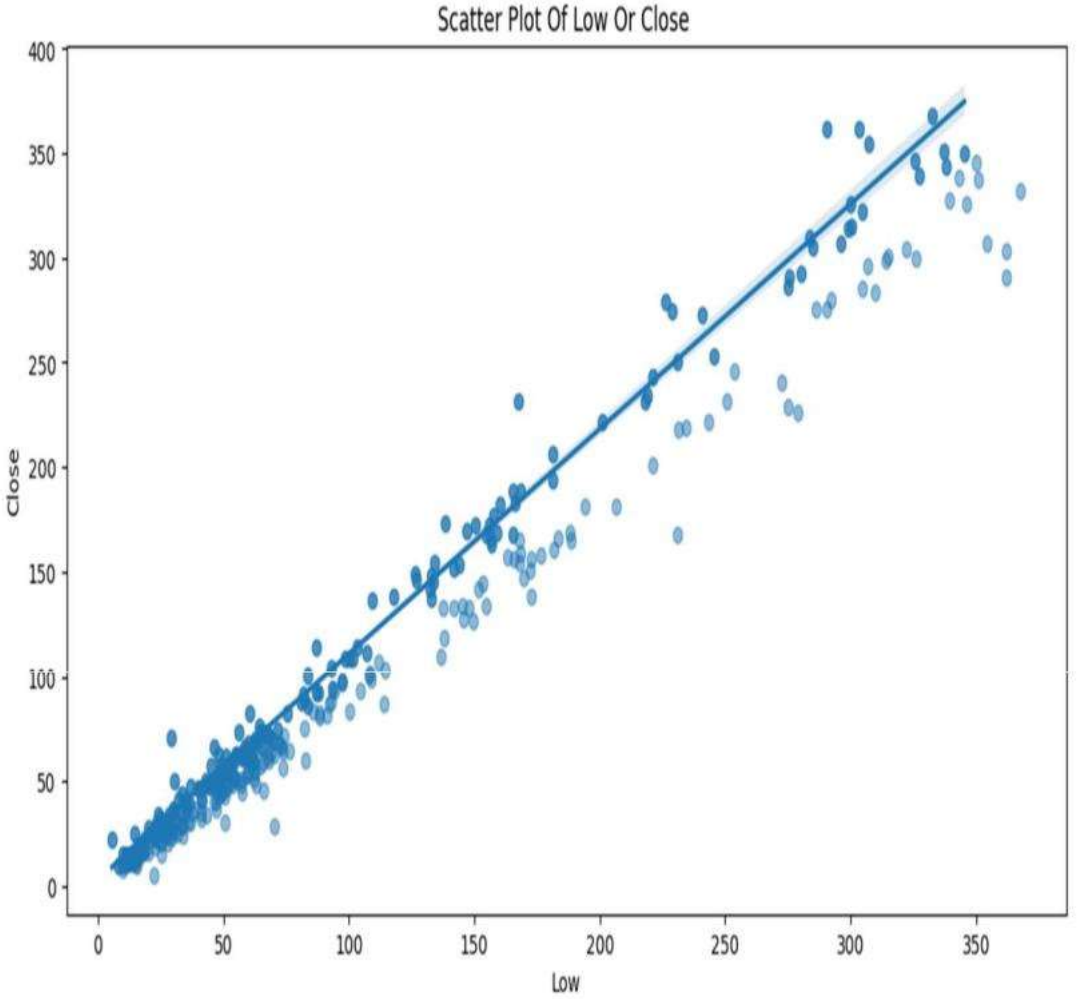


Distribution of opening price, high price and low price are now normal distribution.

Bivariate Analysis Plots



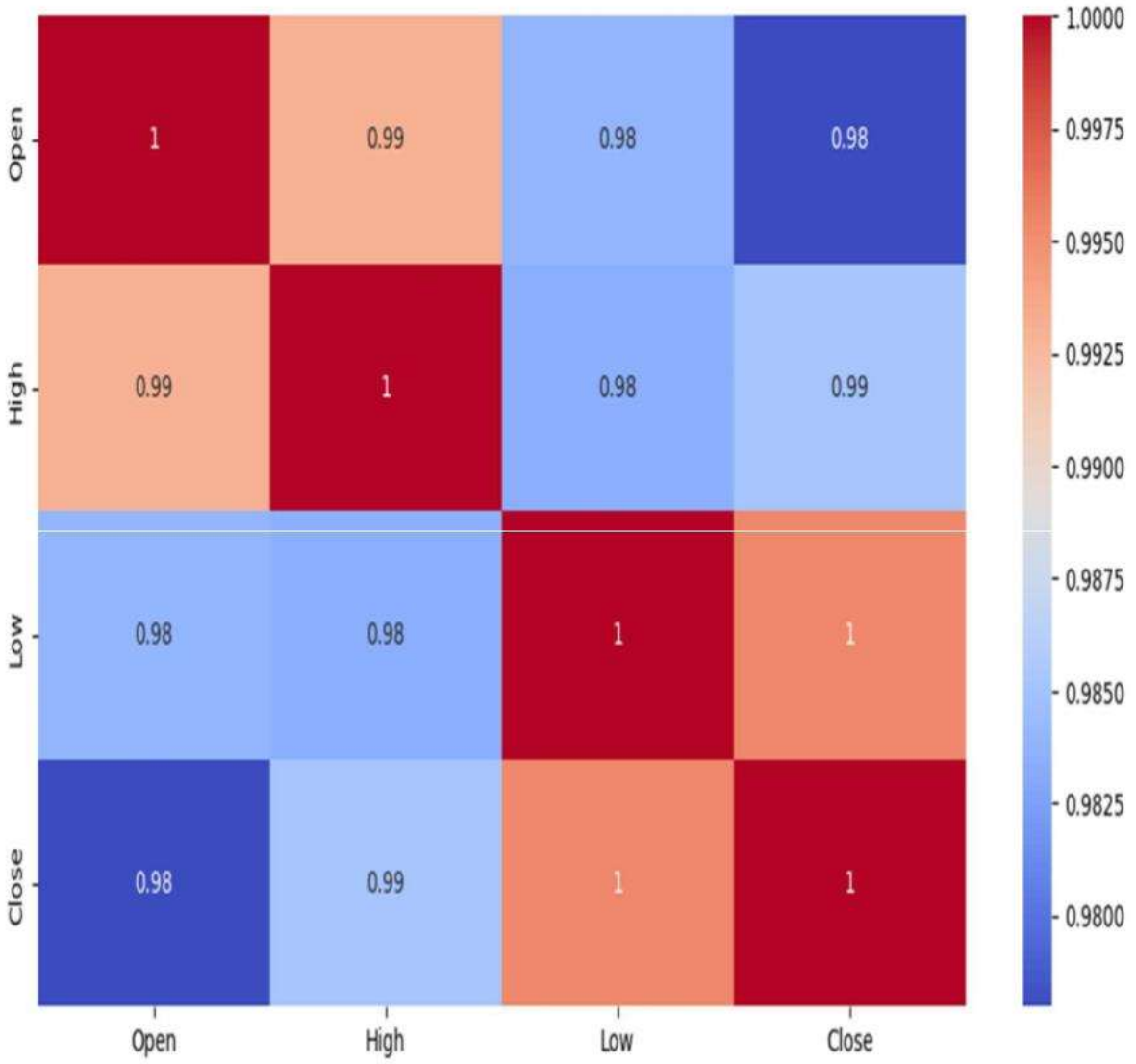




As we can see that there is linear relation and high correlation between each independent variables and our dependent variable.

Correlation Heatmap

- The correlation matrix helps us visualize the correlation of each parameter with respect to every other parameter.
- The colors changes from blue to red for highest to the lowest correlation values and vice versa.
- We can see in the heatmap on this slide that our dependent variable (close price) is highly correlated with all the other independent variables



Transformation of Data

- 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.

- The basic goal was to enforce a level of consistency or uniformity to dataset.

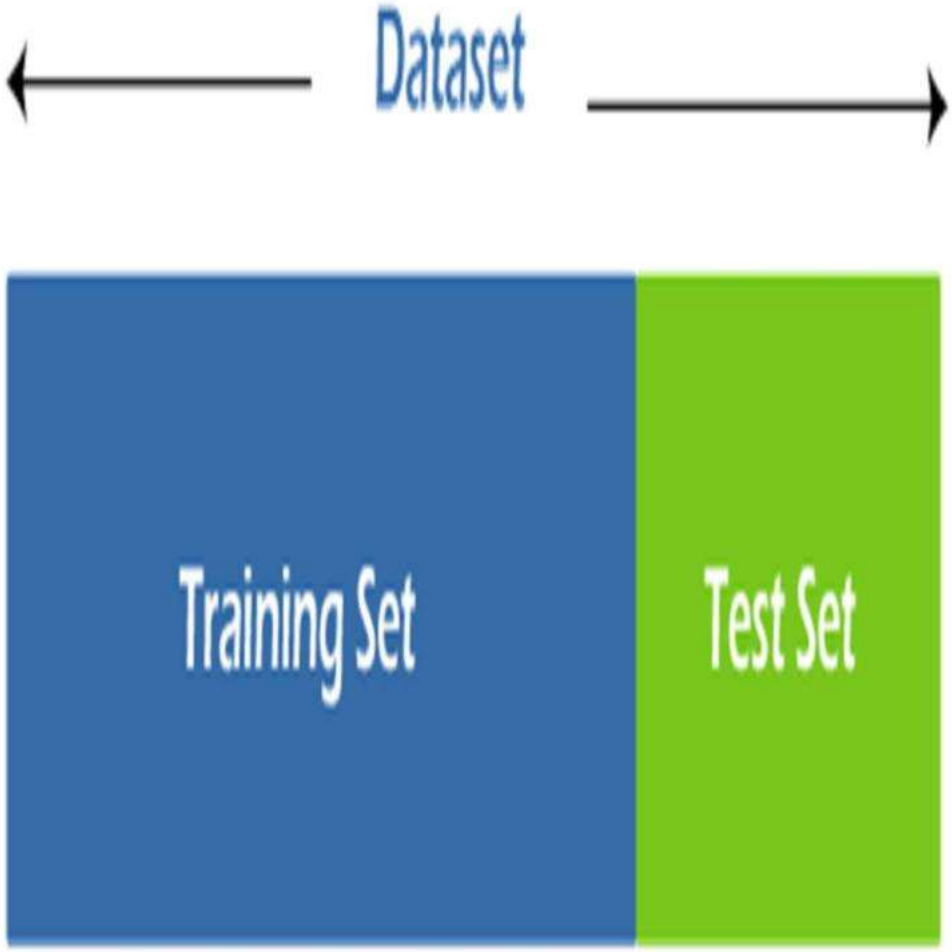
Splitting Data

- Data splits into training dataset and testing dataset.

- Training dataset is for making algorithm learn and train model.

- Test dataset is for testing the performance of train model.

- Here 80% of data taken as training dataset & remaining 20% of dataset used for testing purpose.

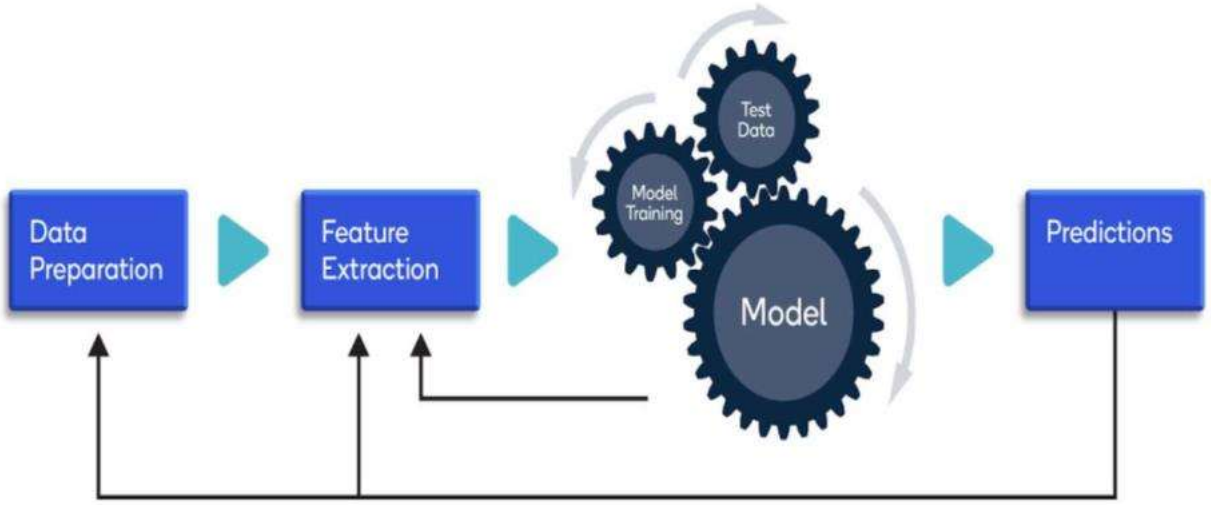


Model Implementation

Based on the linear relationship between the dependent and independent variables present in our data, we implemented following models on our data.

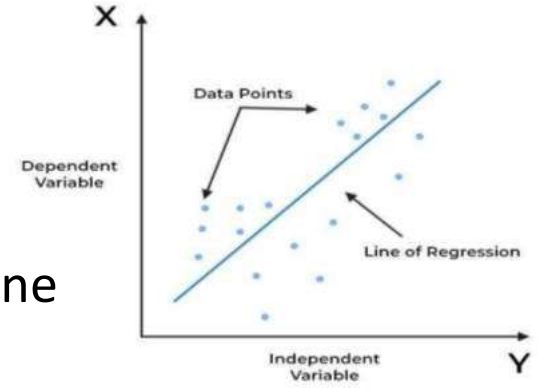
we build 5 regression models for our data.

- Linear Regression
- Lasso Rigression
- Ridge Rigression
- Elastic Net Regression
- XG Boost Regrression

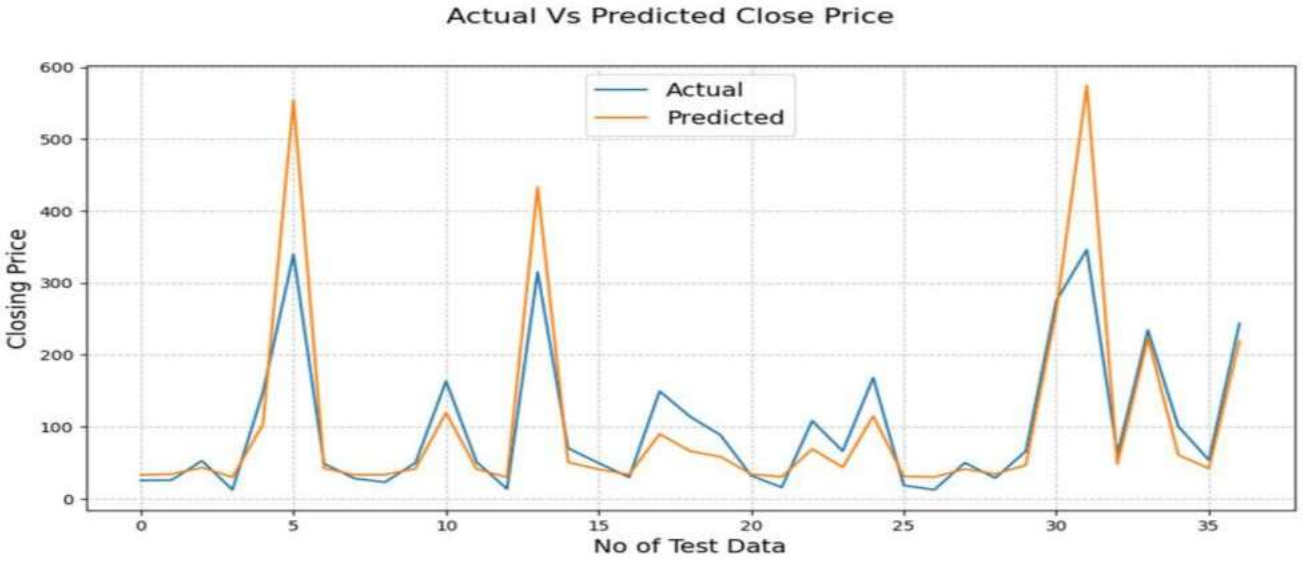


Fitting Different Model

Linear Regression



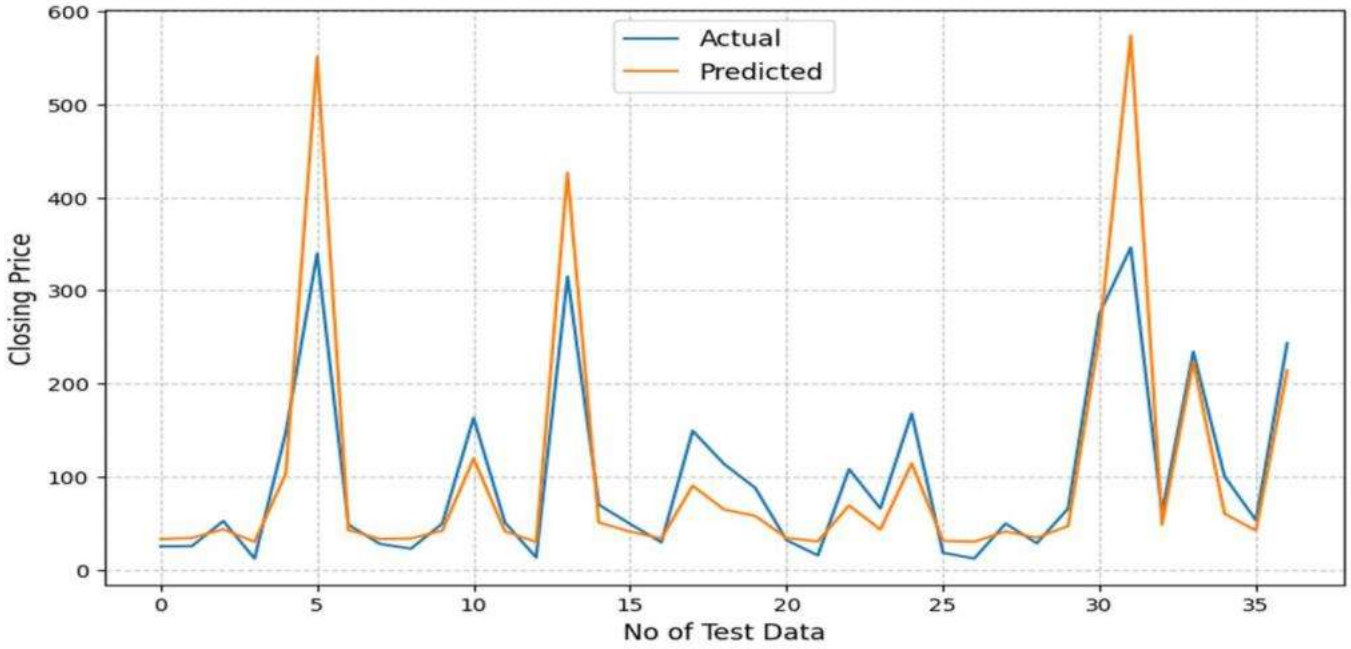
- Linear regression is one of the easiest and most popular Machine Learning algorithms.
- It is a statistical method that is used for predictive analysis.
- Linear regression algorithm shows a linear relationship between a dependent and independent variable; hence it is called as linear regression.



Evaluation Metrics: Linear Regression				
MSE	RMSE	MAE	MAPE	R2
0.032	0.179	0.1523	0.0962	0.82

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.



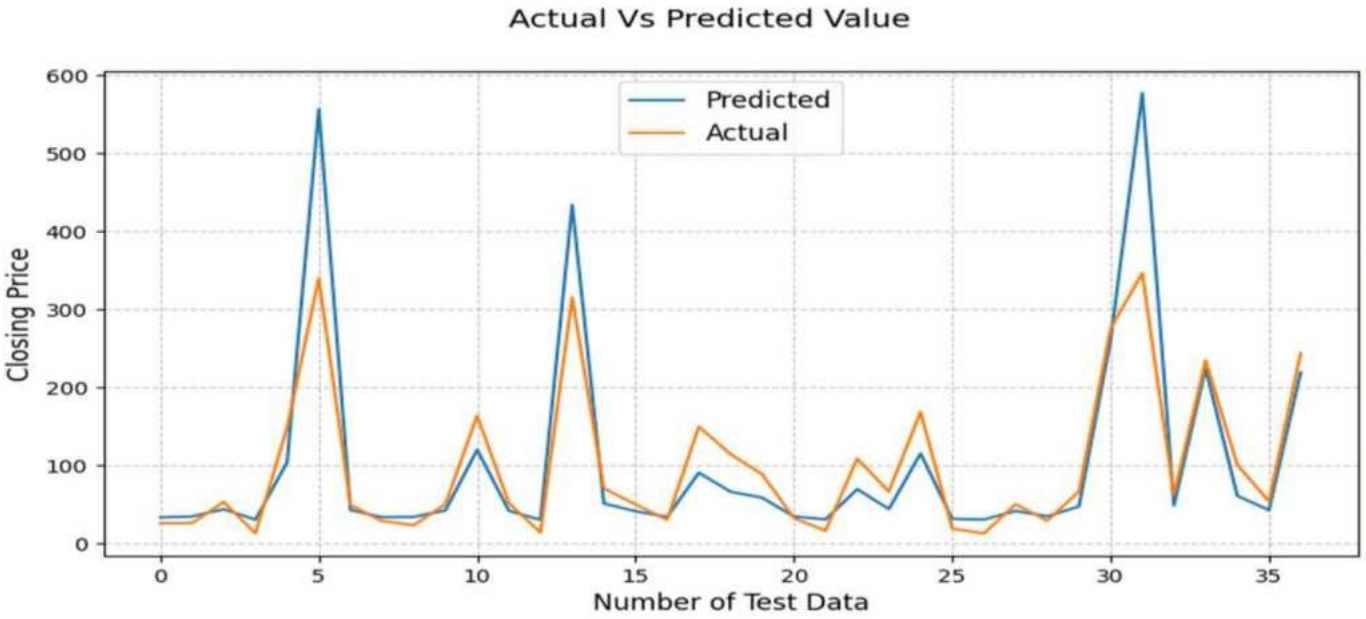
Evaluation Metrics: Lasso Regression				
MSE	RMSE	MAE	MAPE	R2
0.032	0.179	0.1523	0.996	0.82

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.

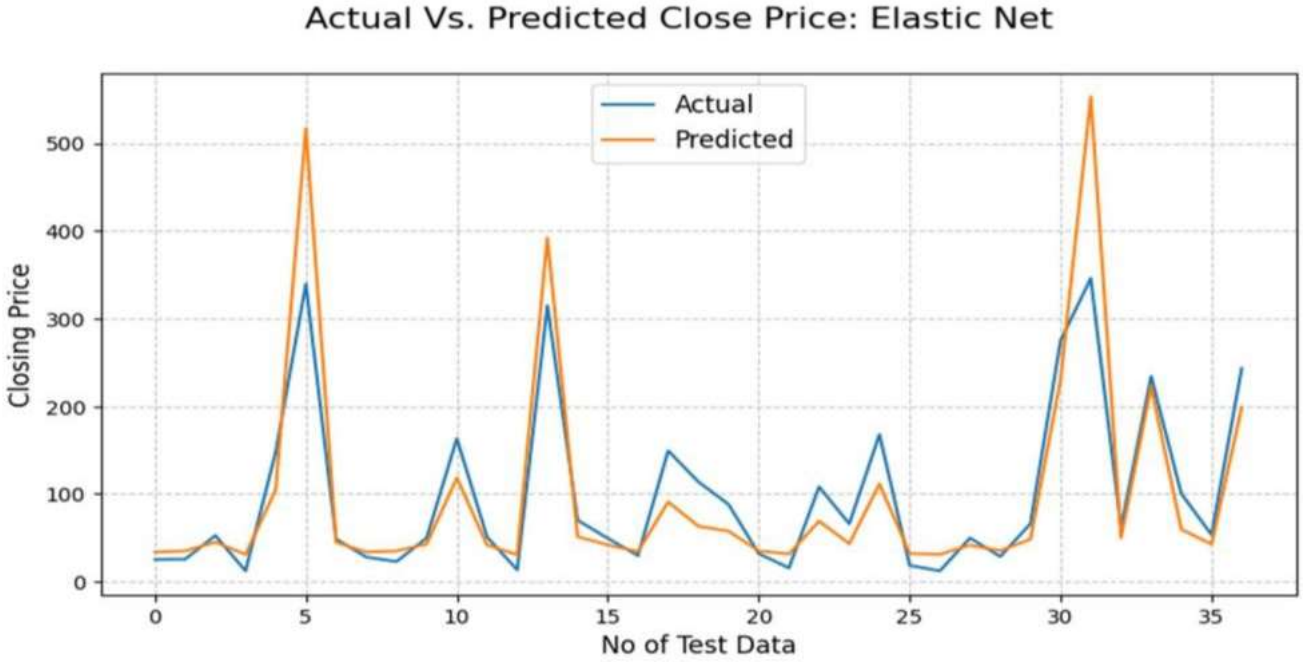
• This method performs L2 regularization.



Evaluation Metrics: Ridge Regression				
MSE	RMSE	MAE	MAPE	R2
0.0317	0.1779	0.1514	0.0955	0.8221

Elastic Net

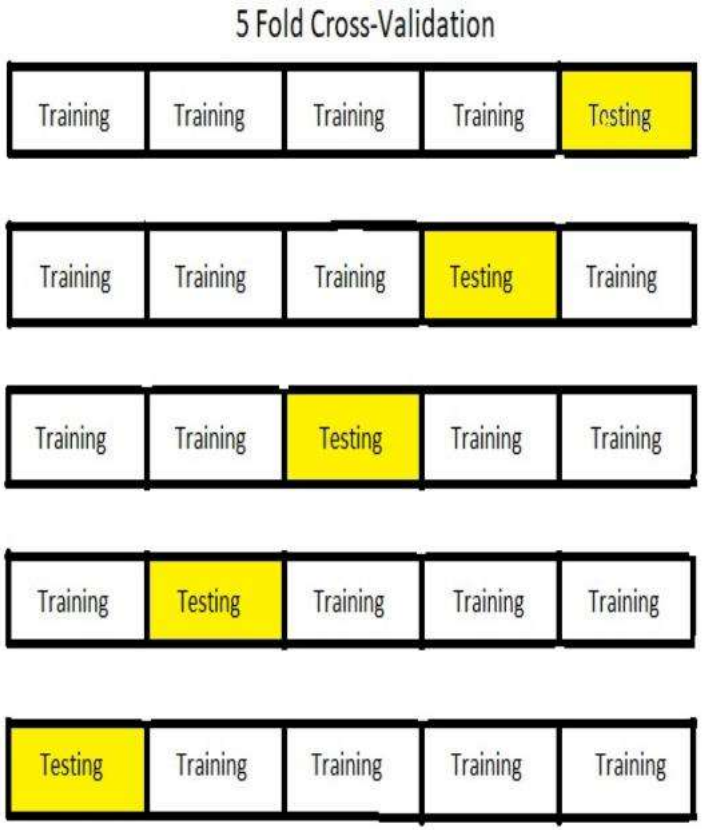
- Elastic net is a popular type of regularized linear regression that combines two popular penalties,specifically the L1 and L2 penalty functions.
- Elastic Net is an extension of linear regression that adds regularization penalties to the loss function during training.



Evaluation Metrics: Elastic Net				
MSE	RMSE	MAE	MAPE	R2
0.033	0.1815	0.154	0.0978	0.815

Cross Validation & Hyperparameter Tuning

- It is a resampling procedure used to evaluate machine learning models on a limited data sample.



- Basically, 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.

- Cross Validation & Hyperparameter tuning on Lasso Regression

Evaluation Metrics :- CV & tuning on Lasso Regression				
MSE	RMSE	MAE	MAPE	R2
0.032	0.179	0.1523	0.0962	0.82

- Cross Validation & Hyperparameter tuning on Ridge Regression

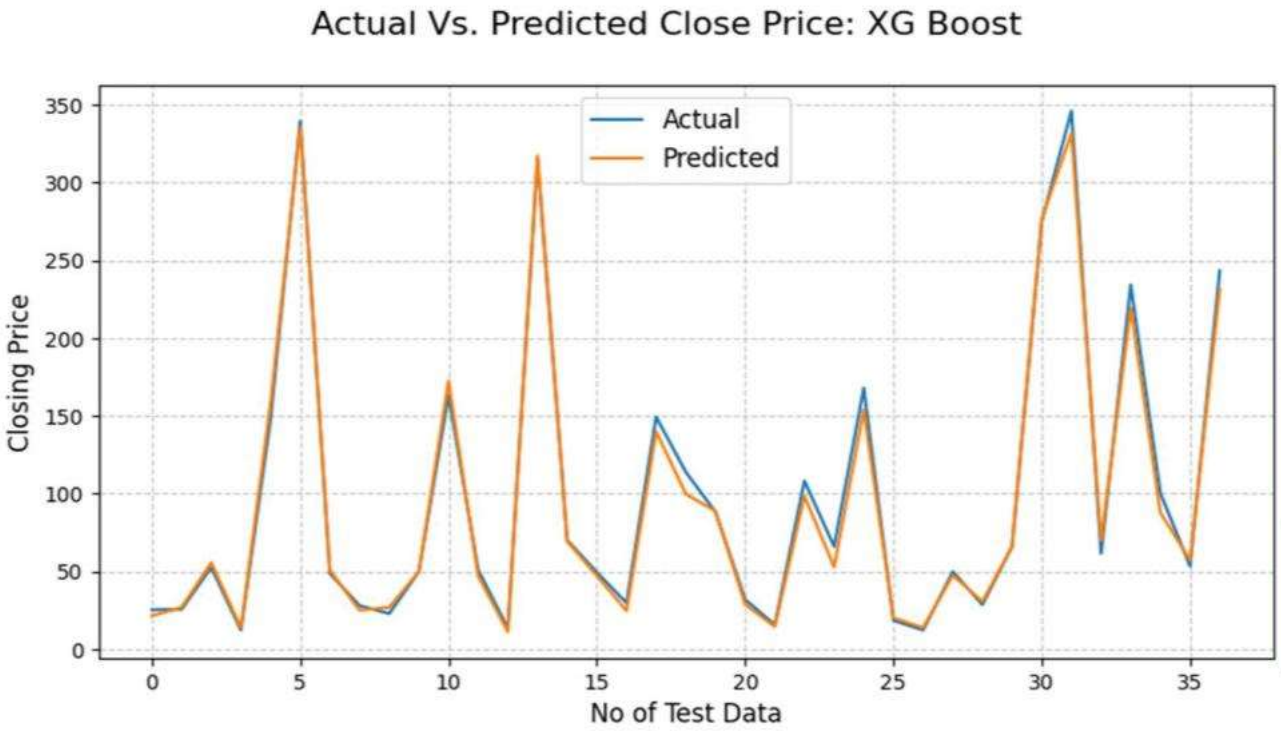
MSE	RMSE	MAE	MAPE	R2
0.0325	0.1804	0.1531	0.0968	0.8172

- Cross Validation & Hyperparameter tuning on Elastic Net

MSE	RMSE	MAE	MAPE	R2
0.0322	0.1795	0.1528	0.0968	0.819

XGBoost Regressor

XGBoost stands for “Extreme Gradient Boosting”.XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.



Evaluation Metrics: XGBoost Regression				
MSE	RSME	MAE	MAPE	R2

CONCLUSION -:

- The popularity of stock closing is growing extremely rapidly day by day which encourage researcher to find new methods if any fraud happens.
- This technique is used for prediction is not only helpful to researchers to predict future stock closing prices or any fraud happen or not but also helps investors or any person who dealing with the stock market in order to prediction of model with good accuracy.
- In this work we use linear regression technique, lasso regression, ridge regression, elastic net regression and XGBoost Regression technique these five models gives us the following results
- High, low, open are directly correlate with the closing price of stocks
- Target variable(dependent variable) strongly dependent on independent variables
- Xgboost regression is best model for yes bank stock closing price data this model use for further prediction

THANK YOU