CAPSTONE PROJECT - 2

YES BANK STOCK CLOSING PRICE PREDICTION

BY – SANKET KAMBLE

CONTENTS

- INTROCUTION
- OBJECTIVE
- PROBLEM STATEMENT
- DATA UNDERSTANDING
- EXPLORATORY DATA ANALYSIS (EDA)
- MODEL IMPLEMENTATION
- COMPARE ALL THE MODELS
- CONCLUSION

❖ INTRODUCTION

Yes Bank is a recognized bank in India's financial sector. Yes bank refers to Youth enterprise Scheme Bank. This bank listed in share market. The aim of this project is to construct a predictive model for close price prediction. The main point of the stock market is that people try to buy shares in a lower price and sell them when the price goes up, thereby making a profit. Any stock's price may vary depending on a number of variables. Events like the bank management personnel fraud case undoubtedly have a significant impact on stock prices.

Thus we are looking on such one case of Rana Kapoor yes bank fraud case. In order to forecast how other relevant features will affect the stock closing price of the bank, I have been given a dataset of Yes Bank stock prices. Several machine learning models have been applied to make a prediction.

❖ OBJECTIVE

A dataset including information on Yes Bank's monthly stock price has been made available to us. This project's goal was to apply several models to see if it was possible to predict stock prices and movement using various characteristics and/or previous performance.

By analyzing the links between the dataset's various properties, we can train the model and finally forecast the closing price using the trained data by appropriately passing the necessary parameters.

❖ 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

❖ DATA UNDERSTANDING

What did you know about your dataset?

While looking to the Downloaded data sets we can see they provided one data_YesBank_StockPrices.CSV file. In this csv file. details of all stock prices are given . We have been given access to a dataset that has 185 rows and 5 total columns of variables, including "date," "high," "low," "open," and "close ".

- Variables Description

Date : Date of record

Open : Opening Price

High : Highest price in the day

Low : Lowest price in the day

Close : Occupations of the speaker

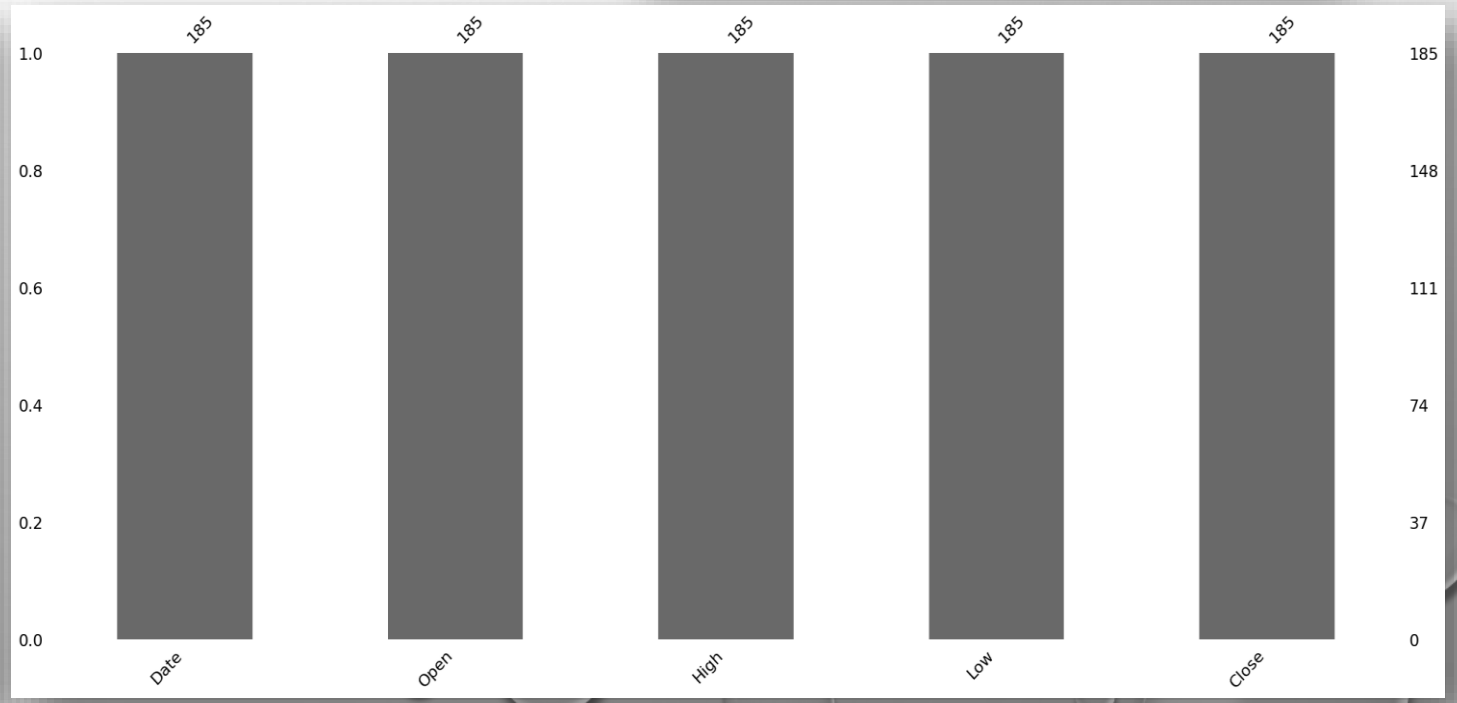
	Date	Open	High	Low	Close
0	Jul-05	13.00	14.00	11.25	12.46
1	Aug-05	12.58	14.88	12.55	13.42
2	Sep-05	13.48	14.87	12.27	13.30
3	Oct-05	13.20	14.47	12.40	12.99
4	Nov-05	13.35	13.88	12.88	13.41

❖ DATA UNDERSTANDING

There is no null and duplicate value in our data set
Data is in string object type.
Data need to convert in date time object.
Summary of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 185 entries, 0 to 184
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    185 non-null    object
 1   Open    185 non-null    float64
 2   High    185 non-null    float64
 3   Low     185 non-null    float64
 4   Close   185 non-null    float64
dtypes: float64(4), object(1)
memory usage: 7.4+ KB
```

	Open	High	Low	Close
count	185.000000	185.000000	185.000000	185.000000
mean	105.541405	116.104324	94.947838	105.204703
std	98.879850	106.333497	91.219415	98.583153
min	10.000000	11.240000	5.550000	9.980000
25%	33.800000	36.140000	28.510000	33.450000
50%	62.980000	72.550000	58.000000	62.540000
75%	153.000000	169.190000	138.350000	153.300000
max	369.950000	404.000000	345.500000	367.900000



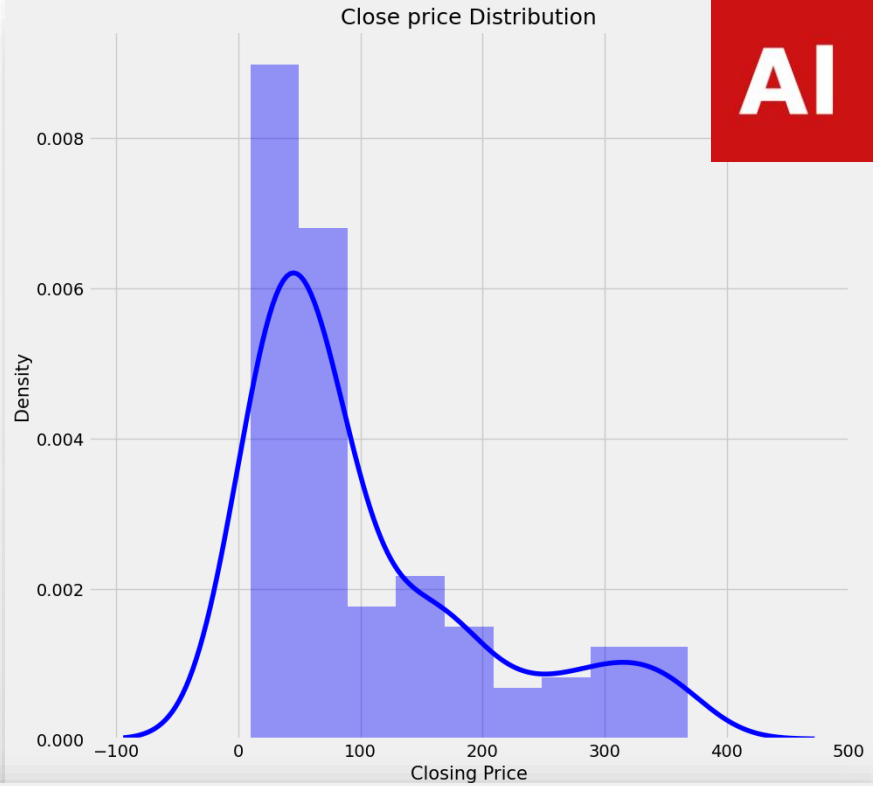
❖ EXPLORATORY DATA ANALYSIS (EDA)

CLOSE PRICE VS DATE GRAPH PLOT

As the graph clearly shows, the stock price was up from 2014-18. There is a sudden decrease in stocks after 2018 that justifies the effect of the fraud case against Rana Kapoor.

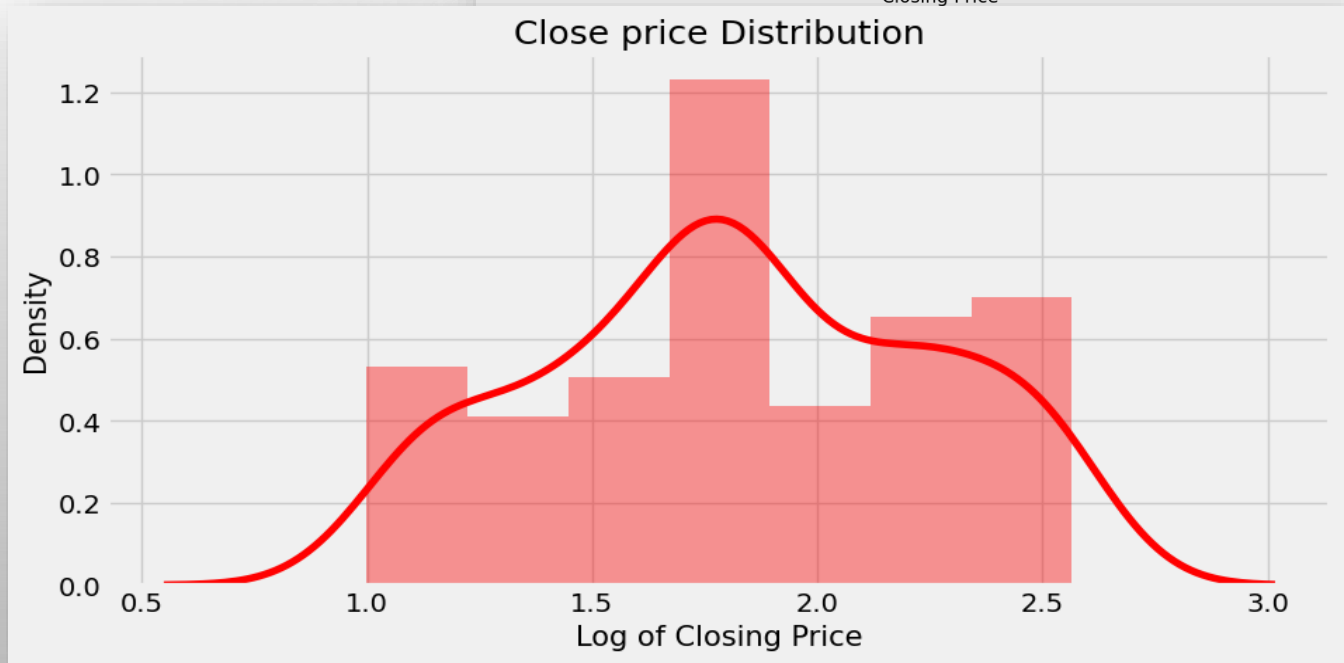
Data of closing price distribution plot is Right skewed
apply log transformation to make uniform distribution

distribution plot of Close price

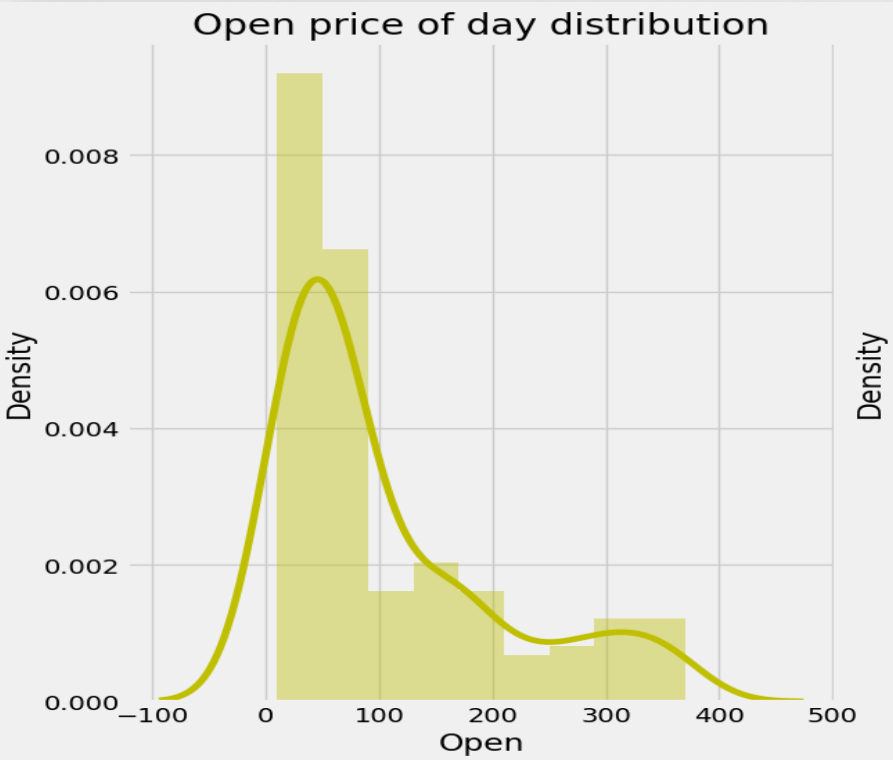


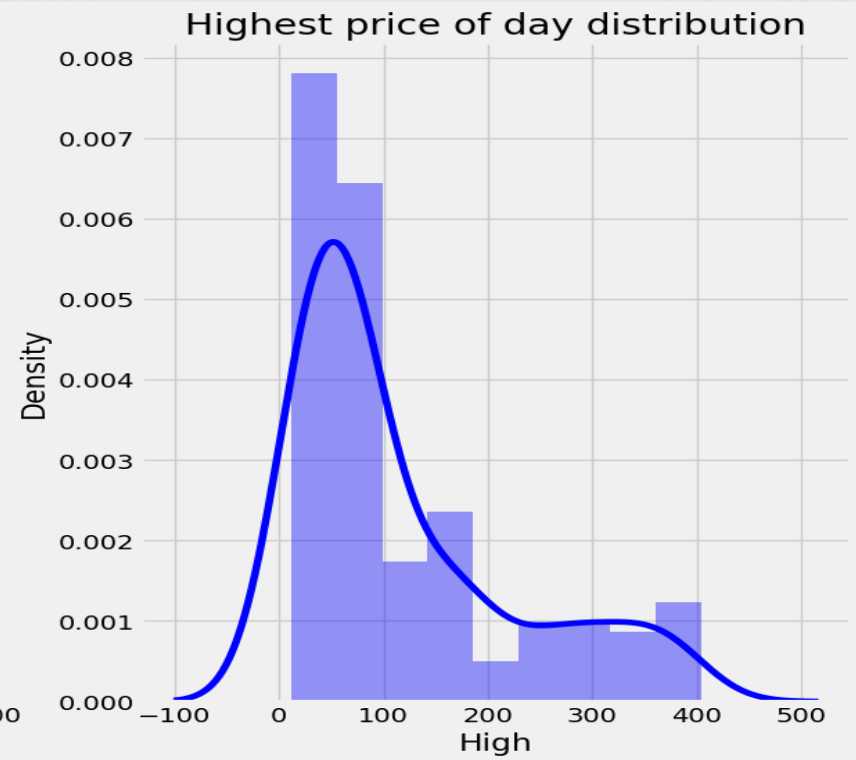
After the log transformation, the closing price distribution is more normal

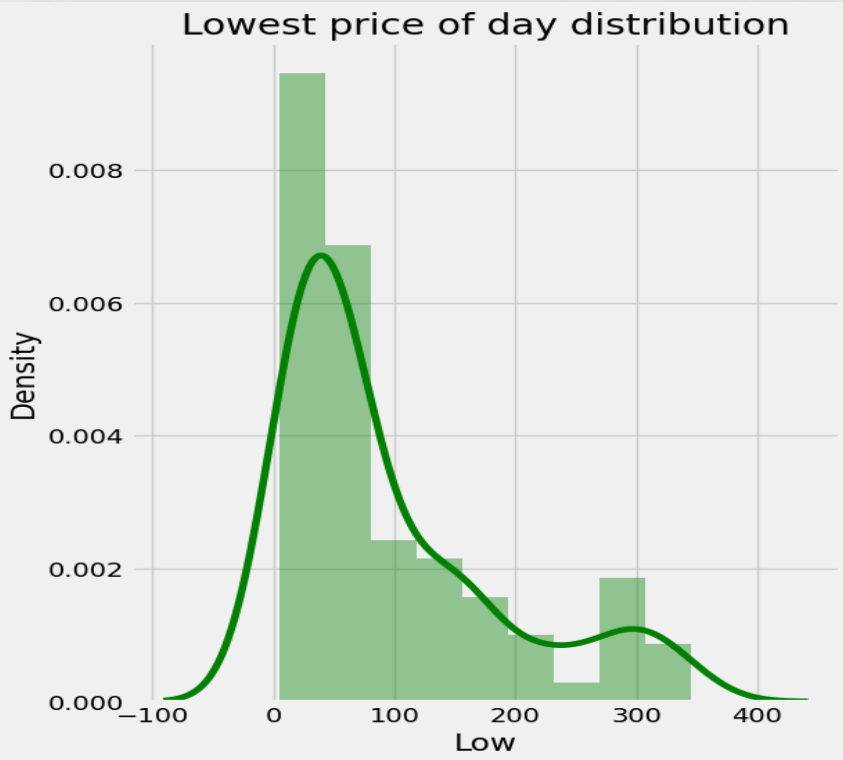
Applying log transformation to Close price distribution plot



Opening price, high price, and low price distribution are all right skewed.
To make them normally distributed apply log transformation

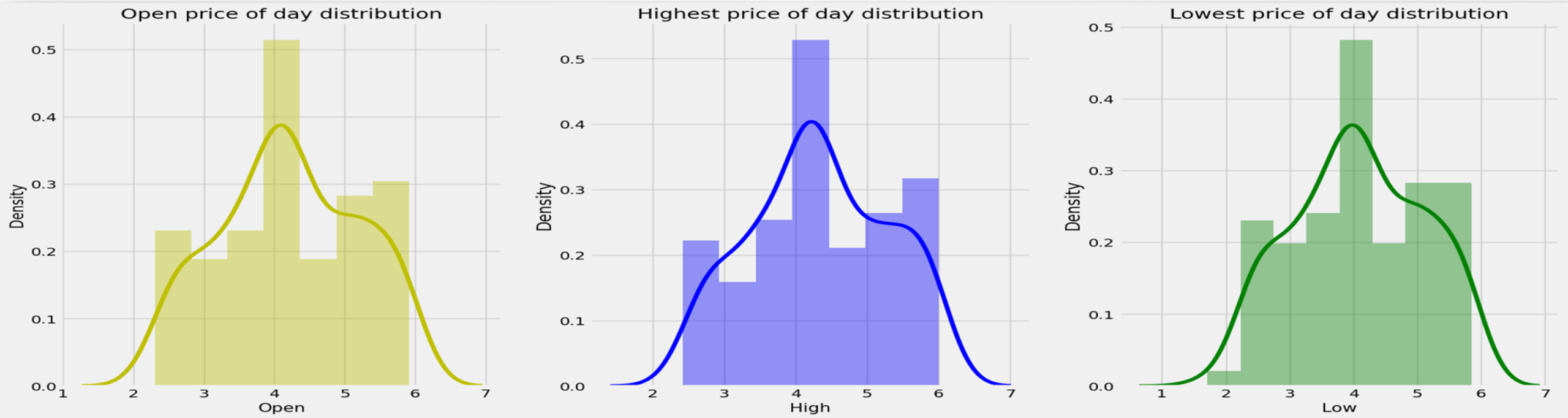






Plot distribution for Open , High and low.

The observed data was discovered to be skewed, as seen in the slides that came before. Before provide the data to our machine learning models, we will change it to make it uniform.

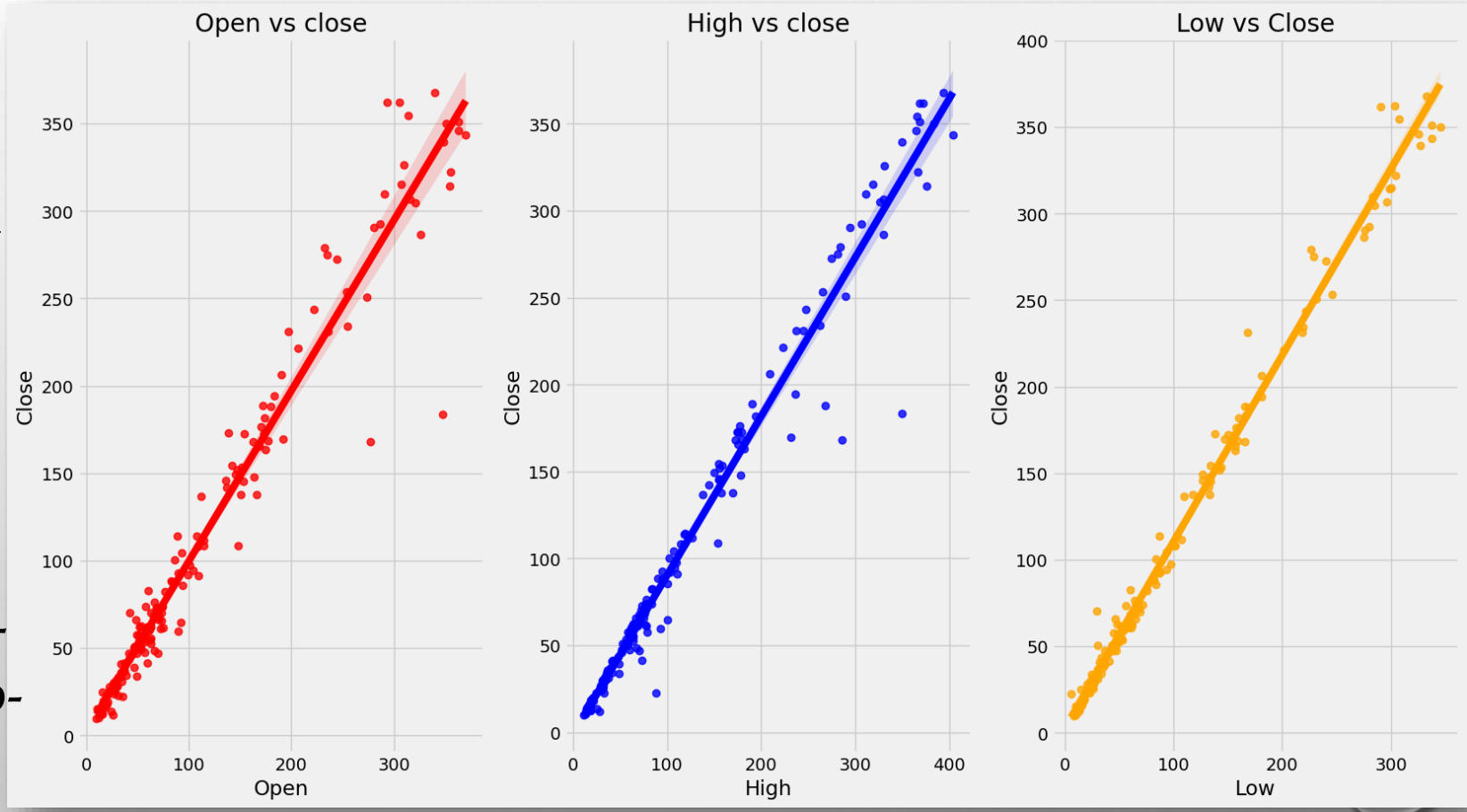


PLOT DISTRIBUTION FOR OPEN ,HIGH AND LOW

The Opening Price, High Price and Low Price distribution is now a normal distribution.

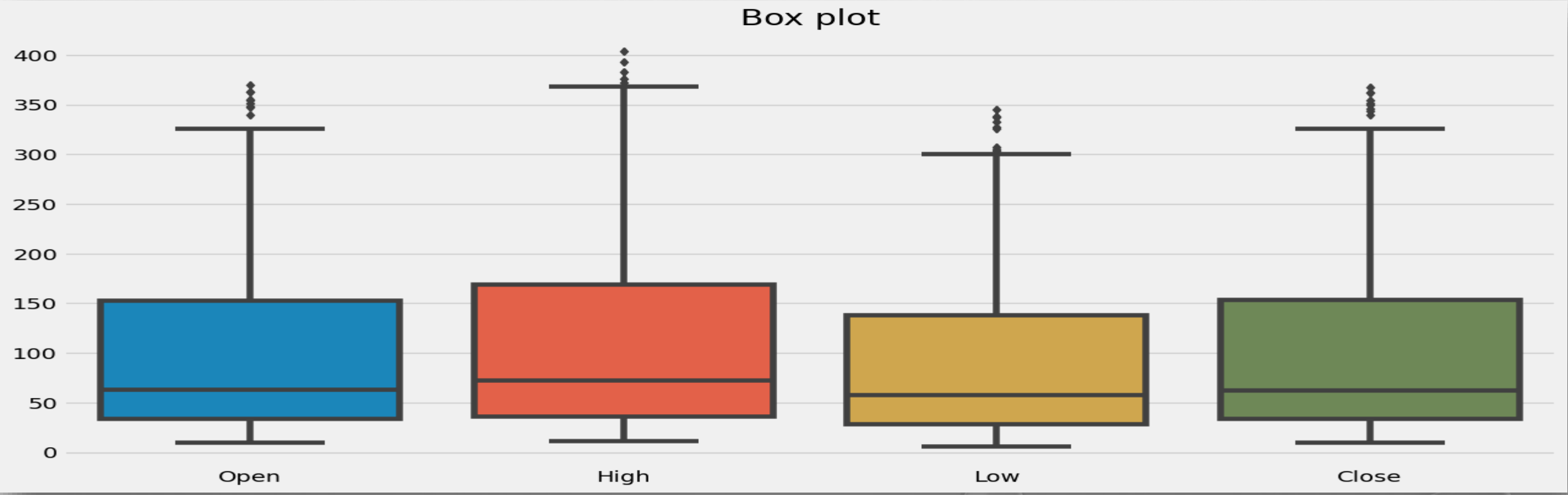
One of the most simple kinds of statistical analysis is bivariate analysis. Two variables are analyzed in order to figure out their empirical relationship with one another. Here we can see relationship between the Dependent Variable (close price) and Independent Variable
(open, High, Low price) .

In this entire scatter plot we can conclude that bivariate analysis shows high correlation of close price with other features, and other features also shows correlation between each Other. it appears that all indepe-ndent variables are directly prop-ortional to the target.



scatter plot to see the relationship between target dependent (closing price) and independent variables

Box plots are used to display the distributions of numerical data values, particularly when comparing them across various groups. The 5-number summary of the data is displayed using the box plot. They are designed to give high-level information at a glance and provide details. There are a few outliers. Every feature is extremely corelated with each other.
The median are close to each other.

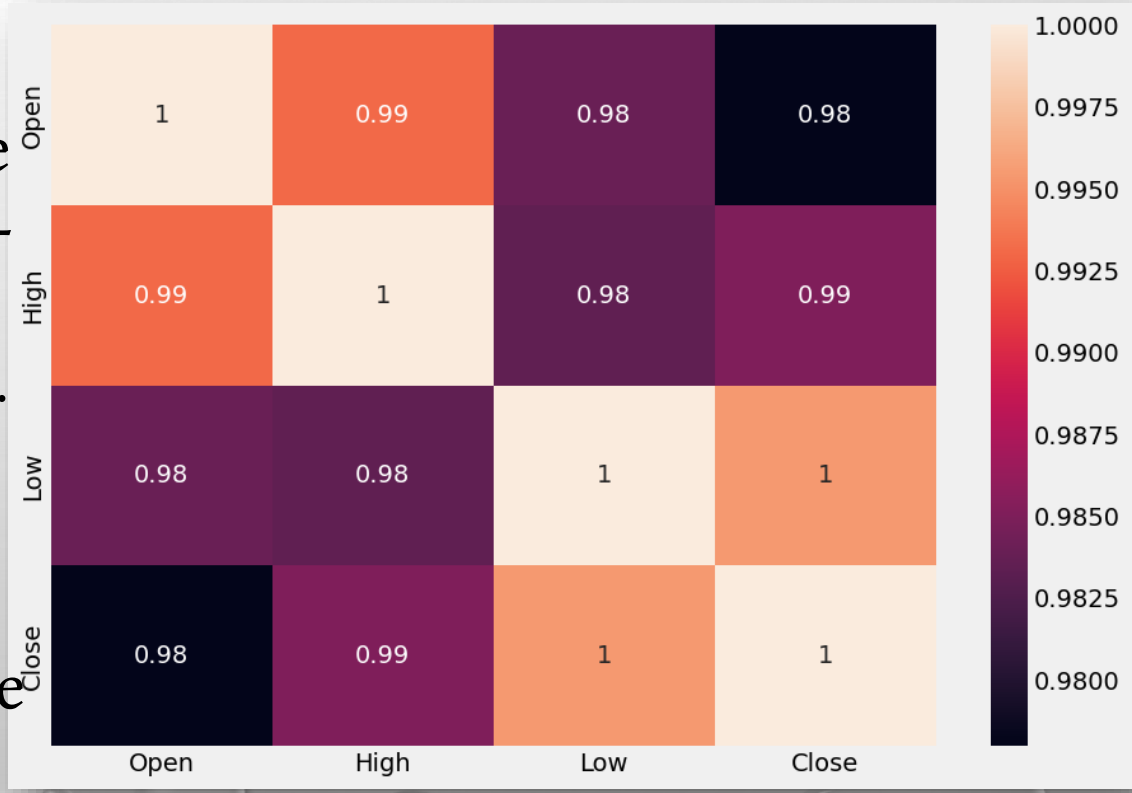


Box plot

A procedure for determining the connections between two variables is known as correlation.

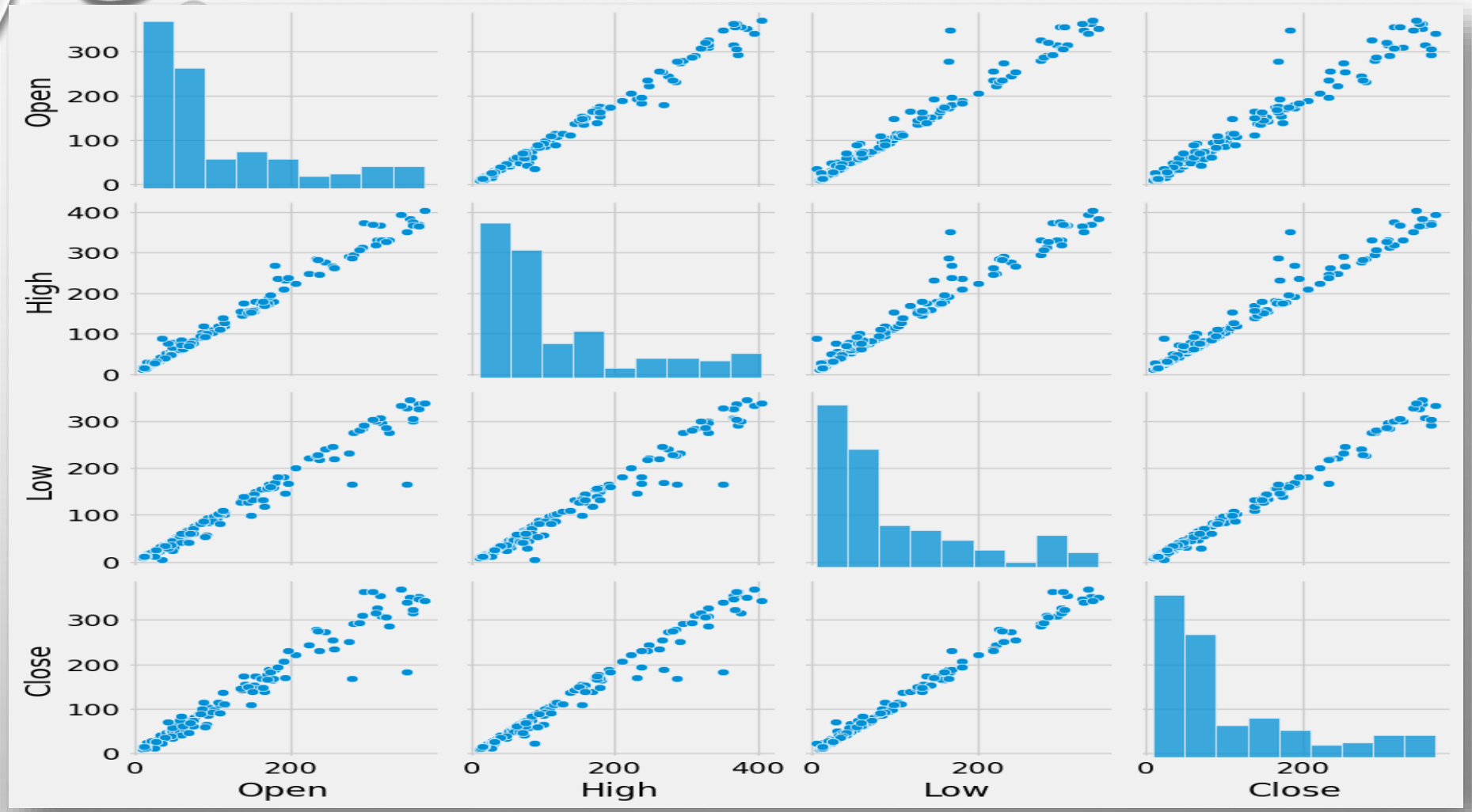
Heatmaps are useful for examining correlation. With one variable drawn on each axis, heatmaps are used to display relationships between two variables, therefore I utilised them.
Positive correlation: There would be 1 correlation. This indicates that the two variables changed in the same direction, either up or down. We have multic-ollinearity since there is a significant correlation between the independent variables in this situation.

High multicollinearity makes it difficult to fit models and make predictions since even small changes to just one independent variable can lead to wildly unanticipated outcomes.



Correlation analysis using Heatmap

Pair Plot



The pair chart helps us understand how the various pairs of variables vary with one another and how they are distributed.

The variables are dependent to each other . They change on a linear basis.

❖ MODEL IMPLEMENTATION

1. Linear Regression

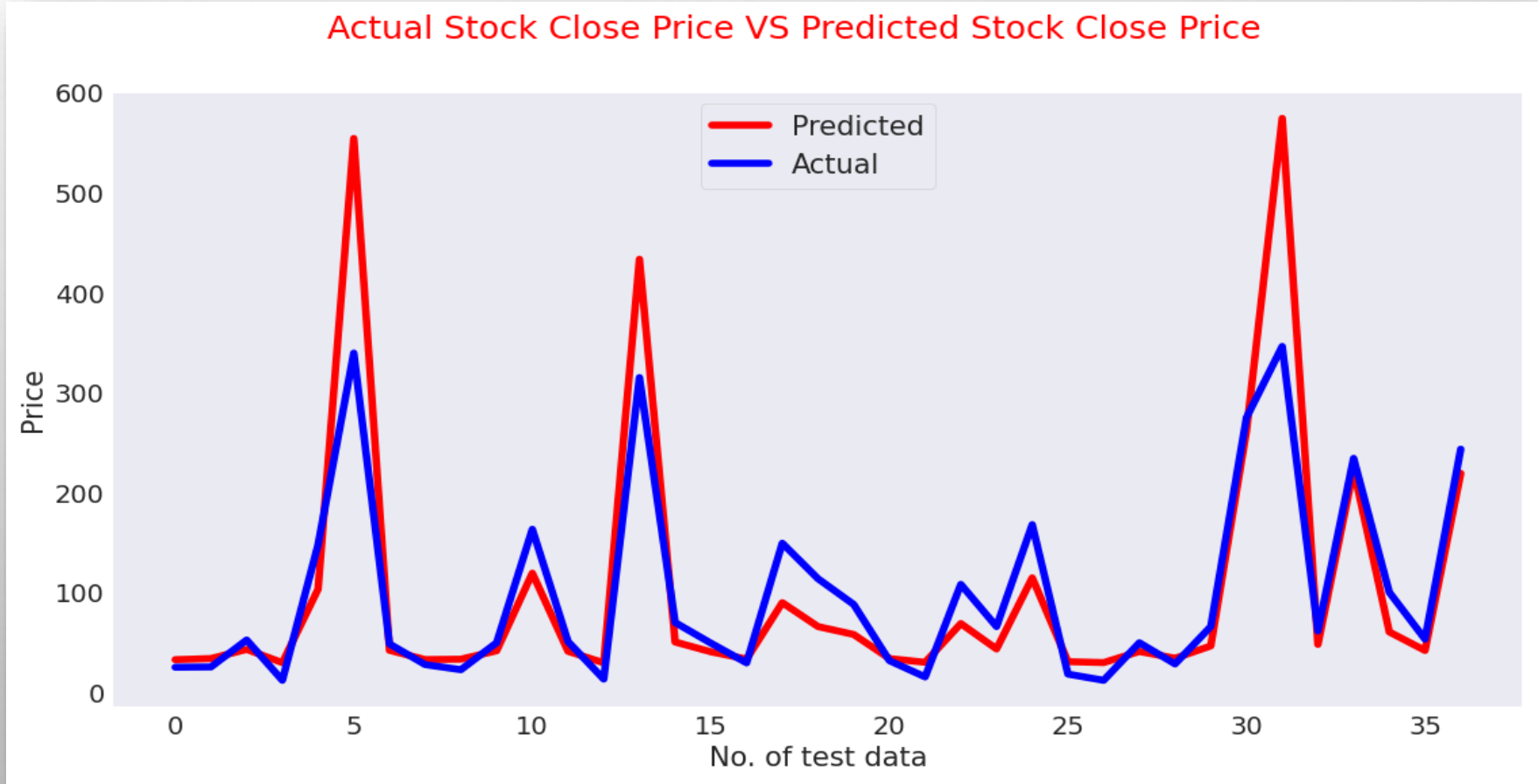
A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you're trying to predict. The differences between predicted and real output values are minimized by linear regression by fitting a straight line or surface.

- Explain each evaluation metric's indication towards business and the business impact pf the ML model used.

R2 evaluates how well a model fits the data. The R2 coefficient of determination in regression is a statistical indicator of how closely the regression predictions match the actual data points. When the R2 value is 1, the regression's predictions accurately reflect the data.

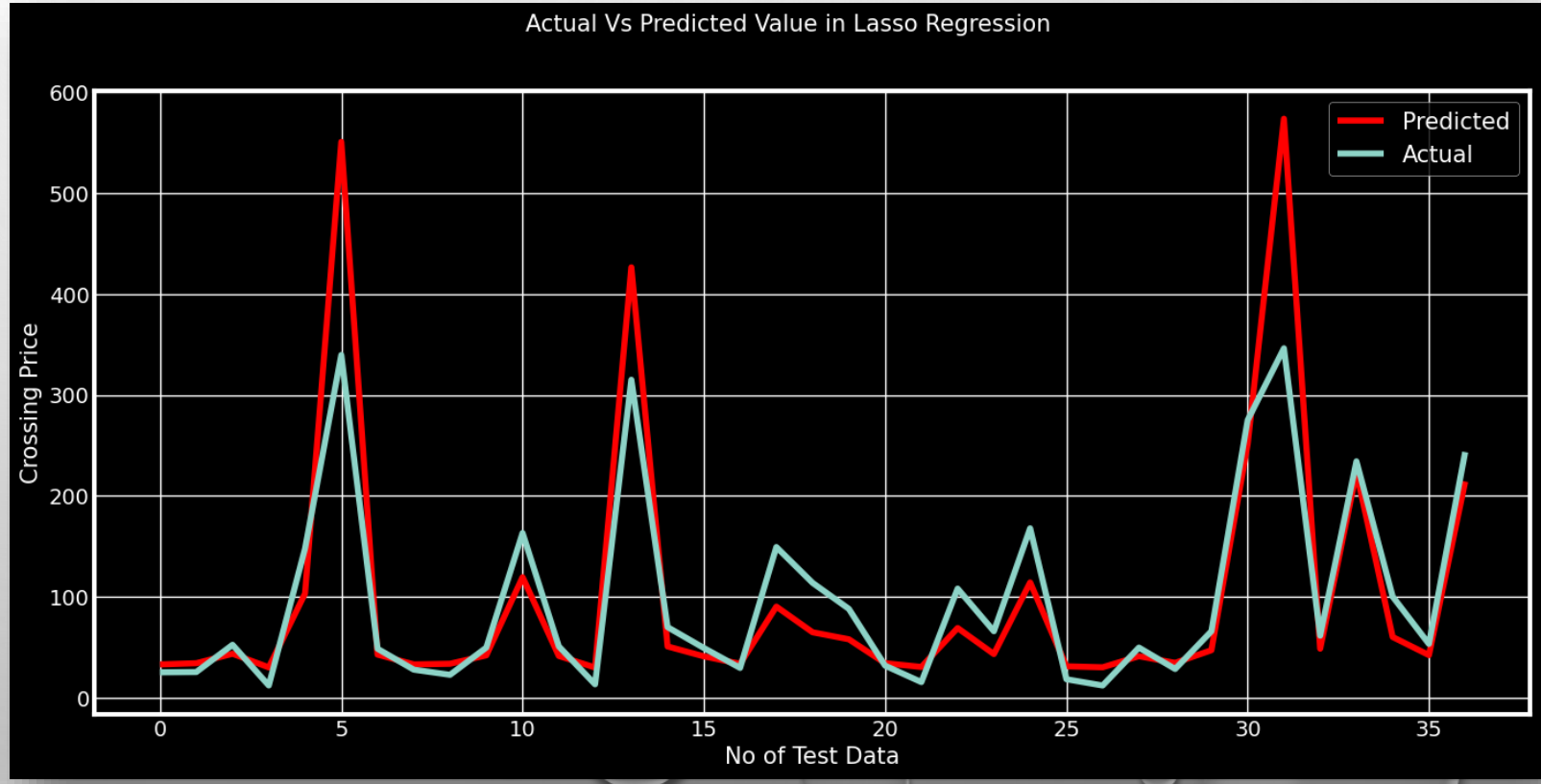
Testing Performance of Linear Regression Model

```
MSE : 0.032
RMSE : 0.178
MAE : 0.151
MAPE : 0.095
R2 : 0.823
```



2.Lasso Regression

Lasso regression is a linear regression, but it uses a "shrinkage" technique where the coefficients of determination shrink to towards zero. Lasso (least absolute shrinkage and selection operator) is regression analysis method that variable selection and regularization performs both in order to enhance the prediction accuracy and interpretability of the resulting statistical model. This method performs L1 regularization.



Evaluation Metrics of Lasso Regression

```
MSE score: 0.032
RMSE score: 0.179
MAE score: 0.152
MAPE score: 0.096
R2 score: 0.82
```

Cross- Validation

MSE score: 0.032 MAPE score: 0.097

RMSE score: 0.18 R2 score: 0.819

MAE score: 0.153

3.Ridge Regression

A regularized variation of linear least squares regression is ridge regression. It functions by reducing the weights or coefficients of the regression method. Ridge regression is a model adjustment method that is used to analyze any data that suffers from multi-llinearity. This method adjusts to L2. 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.

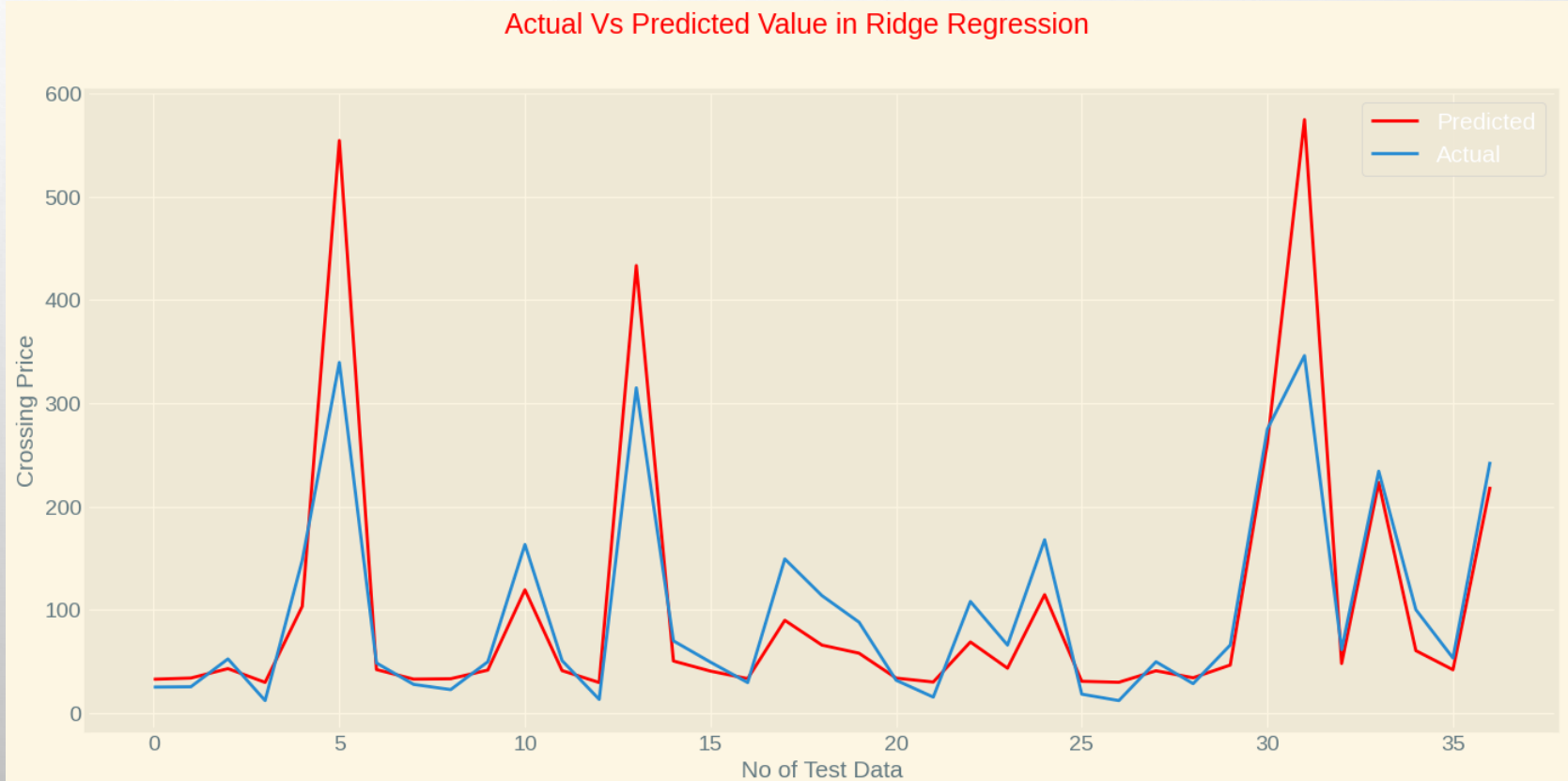
- Which hyperparameter optimization technique have you used and why?

In order to choose the optimal parameter, I utilised gridsearchcv.

Test Performance

```
MSE score: 0.0316
RMSE score: 0.1777
MAE score: 0.1513
MAPE score: 0.0954
R2 score: 0.8225
```

```
Cross- Validation
MSE:0.033 RMSE:0.18 MAE:0.153
MAPE: 0.097 R2: 0.817
```



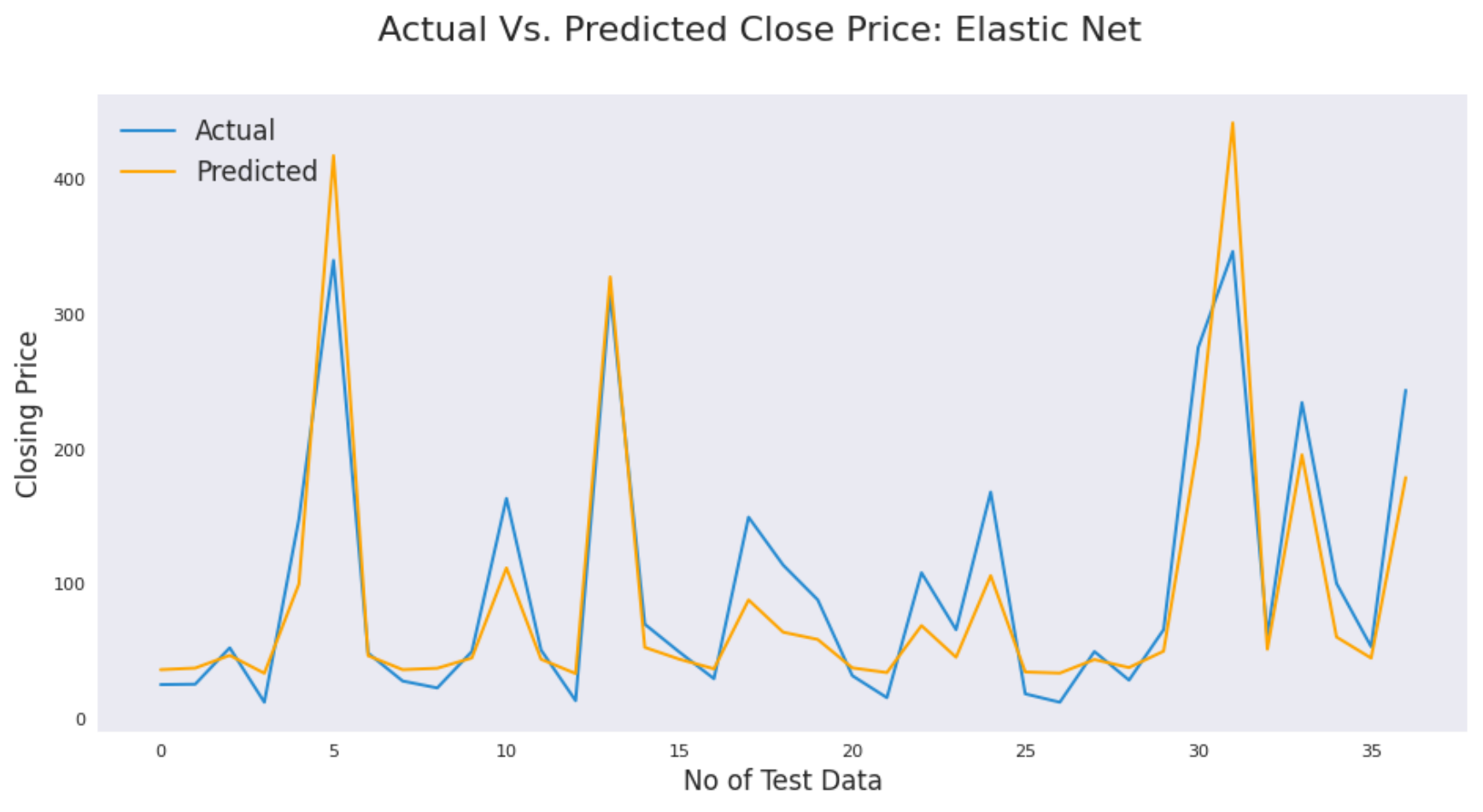
4.Elastic Net regression

- Explain the ML Model used and it's performance using Evaluation metric Score Chart.

The third type of regularization method is elastic net regression. It was created as a result of the Lasso regression's limitation. Lasso regression can’t take correct alpha and lambda values as per requirement of data. It combines two popular penalties, specifically the L1 and L2 penalty.

Test Performance

```
MSE : 0.036
RMSE : 0.191
MAE : 0.157
MAPE : 0.102
R2 : 0.796
```



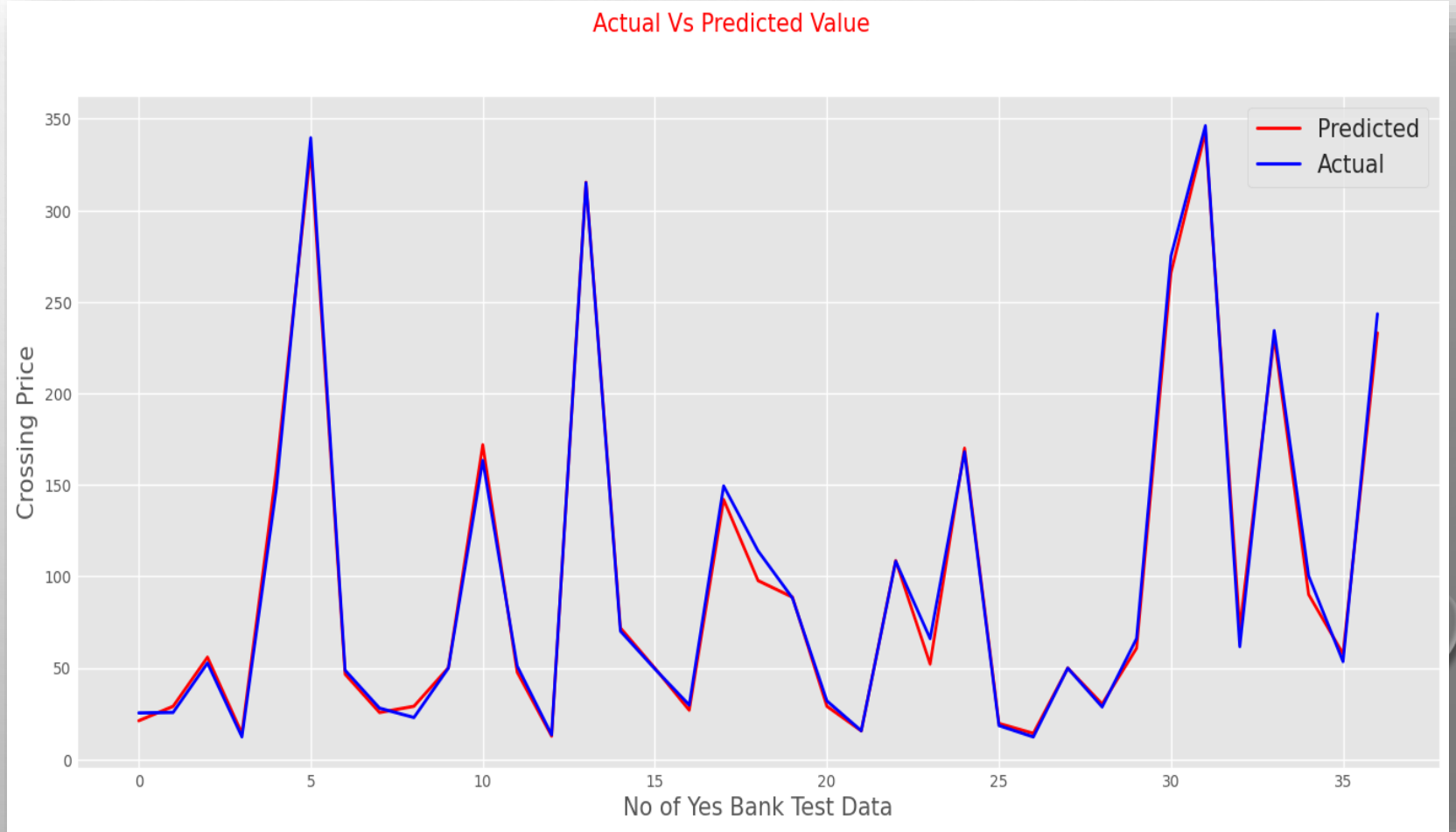
5.Gradient Boosting Regression

- Explain the ML Model used and it's performance using Evaluation metric Score Chart.

The difference between the present forecast and the known correct target value is calculated using gradient boosting regression. This variation is referred to as residual. After that, a weak model that maps features to that residual is trained using gradient boosting regression.

Test Performance

```
MSE : 0.002
RMSE : 0.041
MAE : 0.031
MAPE : 0.02
R2 : 0.99
```



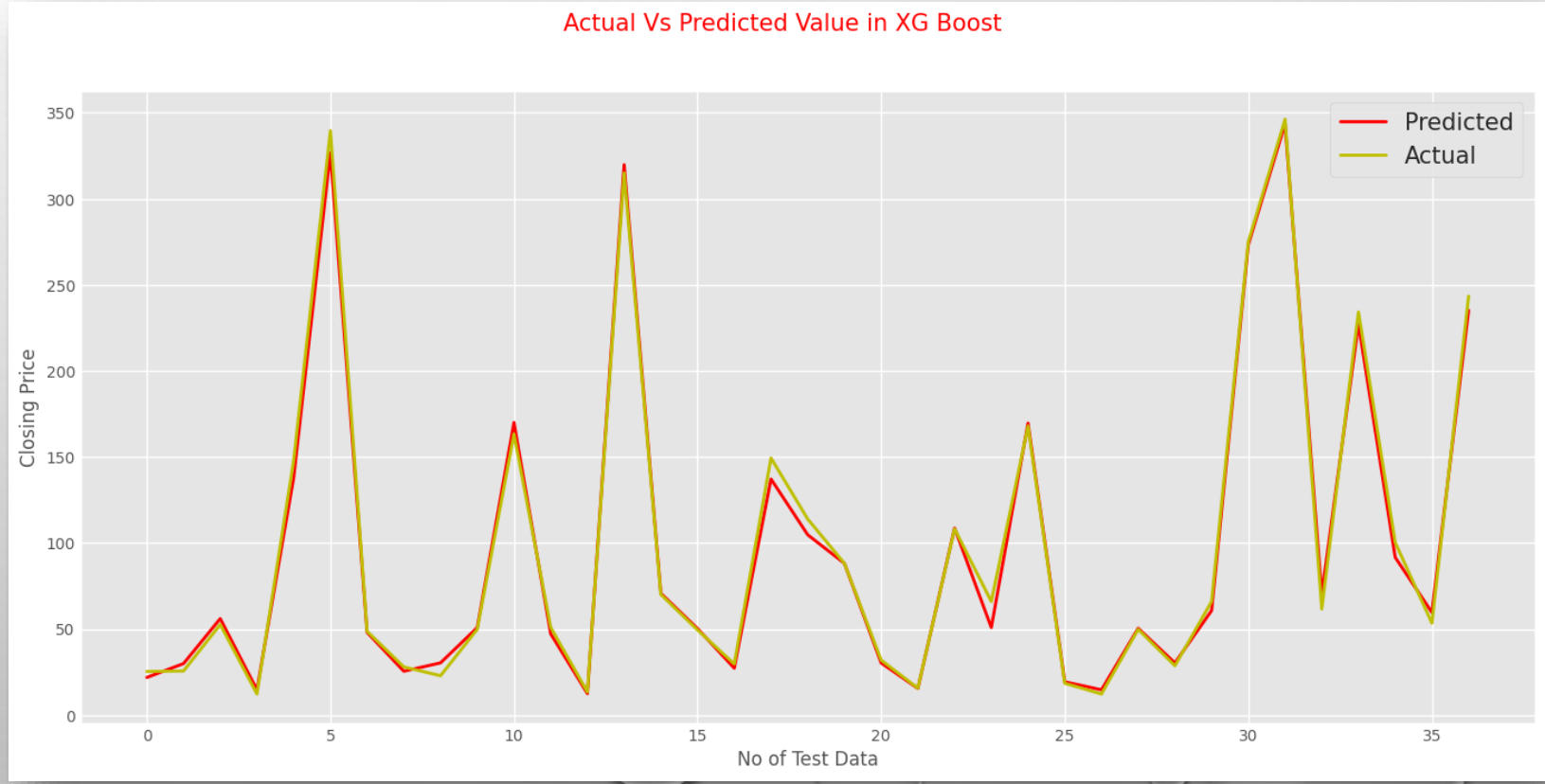
6.Extreme Gradient Boosting or XGBoost regression

- Explain the ML Model used and it's performance using Evaluation metric Score Chart

Each feature utilized for prediction is generally ranked in order of importance by the XGBoostRegression. Gradient boosting has the advantage that retrieving relevance ratings for each attribute is not too difficult after the boosted trees have been built. The two key benefits of XGBoost are model performance and execution speed.

Test Performance

MSE score: 0.002
RMSE score: 0.043
MAE score: 0.031
MAPE score: 0.02
R2 score: 0.989



- Which Evaluation metrics did you consider for a positive business impact and why?

Because we wanted to come as near to the close price as possible, here R2 score has received the most weight in my analysis. R2 evaluates how well a model fits the data. The R2 coefficient of determination in regression is a statistical indicator of how closely the regression predictions match the actual data points. When the R2 value is 1, the regression's predictions accurately reflect the data. In other words, the closer the model is to 1, the better it can forecast our dependent variable. R2 is the ideal evaluation indicator to provide a positive business impact given our desired objective because it will assist us assess how well our models are doing.I found this using gridsearchcv the perfect parameter.

❖ COMPARE ALL THE MODELS

❖ Which ML model did you choose from the above created models as your final prediction model and why?

Sr. No.	Name	MAE	MSE	RMSE	R2_score
1	Linear Regression	0.151	0.032	0.178	0.823
2	Lasso Regression	0.152	0.032	0.179	0.820
3	Lasso Regression cv	0.153	0.032	0.180	0.819
4	Ridge Regression	0.151	0.032	0.178	0.823
5	Ridge Regression cv	0.153	0.033	0.180	0.817
6	ElasticNet	0.157	0.036	0.191	0.796
7	Gradient Boosting Regression	0.030	0.002	0.041	0.991
8	XG Boost Regression	0.031	0.002	0.043	0.989

AI

I've decided to use Gradient Boosting Regression as my last prediction model. Although the other models also did well, the majority of them had a R square score of between 70% to 80%. The R-square score for We got 99% highest accuracy for Gradient Boosting Regression. If we look attentively at the graph, we can see that Gradient Boosting accurately captures the data trend, even when it goes around the corners. The other models weren't able to recognize the corner trend, which reduced their score. My second preference will be for XGBoost Regression reasons are the same above but we got second higest 98% accuracy.

For predicting continuous numeric values, XGBoost regression and Gradient Boosting regression are both effective machine learning models. If you need great accuracy and speed with a large and complicated dataset, XGBoost regression can be a better option. Since my dataset is relatively small has only has 185 rows and 5 columns, gradient boosting regression is a better option since it can provide a simpler and more interpretable model

AI

- Explain the model which you have used and the feature importance using any model explainability tool?

Gradient Boosting regression is a simpler ,easy to understood and more easily interpretable model which uses the same sequential forming approach to decision trees. Gradient boosting is a supervised learning algorithm that attempts to precisely predict a target variable by combining estimates from a set of simpler and weaker models.

One of the ensemble technique variations that uses numerous weak models combined for greater overall performance is known as gradient boosting. One of the most recognized machine learning techniques for tabular datasets is gradient boosting. It has excellent usability, can deal with missing values, outliers, and large cardinality categorical values on your features, and is strong enough to detect any nonlinear relationship between your model target and features.

❖ CONCLUSION

- Beginning from Exploratory Data Analysis we see the sudden change in stock price from 2014.
- Bank share price is at the highest in 2018-19.After there is a Sudden fall in price of stock .
- After that stock price start to increase again but price fall again.
- From the scatter plot we can see High,open,low price of share are directly correlate with the closing price of share.
- we applied following regression model on data set and result are evaluted and compared

1.LinearRegression

2.Lasso Regression

3.Ridge Regression

4.GradientBoostingRegressor

5. ElasticNetCV

6.XGBRegressor

- we check test performance such as mean absoulte error, mean squared error, root mean squared error, r2
- We got 99% highest accuracy for Gradient Boosting Regression and 98% for XGB Regressor
- we got almost simillar result for Linear Regression, Lasso Regression and Ridge Regression
- Cross Validation has been applied on various algorithms. However, the outcome is nearly identical.
- Using data on the closing price of Yes Bank's shares, Gradient BoostingRegressor and Xgboost regression is the best model to apply.

THANK YOU