Capstone Project-2

Yes Bank Stock Closing Price Prediction

Submitted by

Shubham Kadu

Data science trainee, Almabetter

Contents:

- Introduction
- Problem Statement
- Problem Objective
- Data Summary
- Exploratory data analysis
- ML-Model :
    1) Linear Regression
    2) Lasso Regression
    3) Ridge Regression
    4) ElasticNet Regression
- Conclusion

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 Information

```
<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
```

Data Summary:

We have the Yes Bank monthly stock price dataset.

Attribute Information:

1) Date: Date of the month of stock price.

2) Open: The opening price of the stock on a particular day

3) High: It's the highest price at which a stock traded during a period

4) Low: It's the lowest price at which stock traded during a period

5) Close: The closing price of a stock at the end of a Trading Day

Note:
The close feature is a dependent feature and others will be independent features.

Exploratory Data Analysis (EDA):

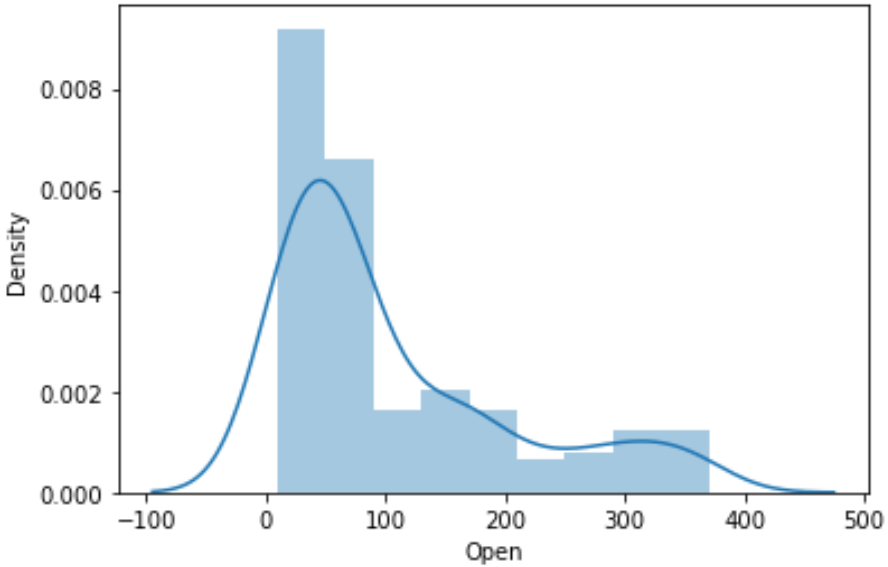
- The Visualize trend of dependent Variable

The above plot of Closing prices of different dates gives a very fluctuation in prices regarding different time-duration. After 2018 there is a sudden fall in the stock closing price.

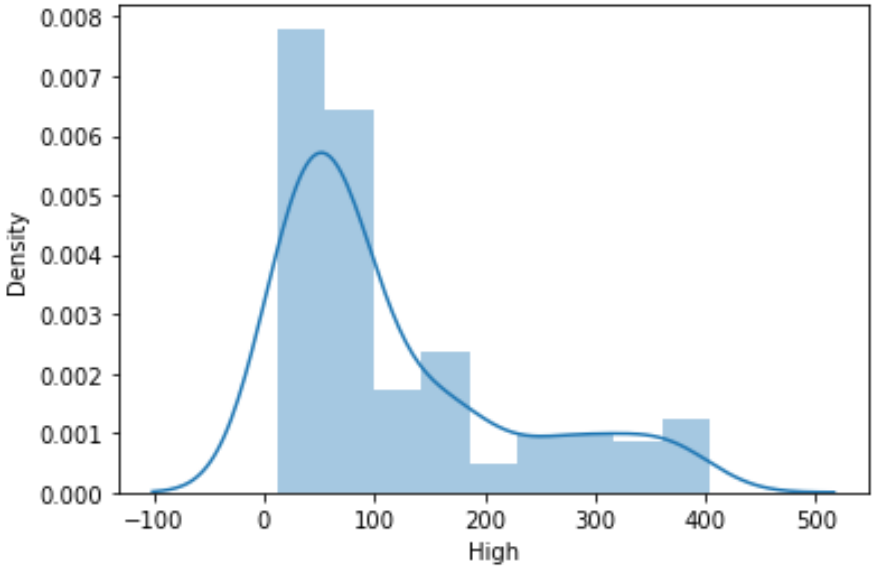
Univariate Analysis

❑ Distribution of Features

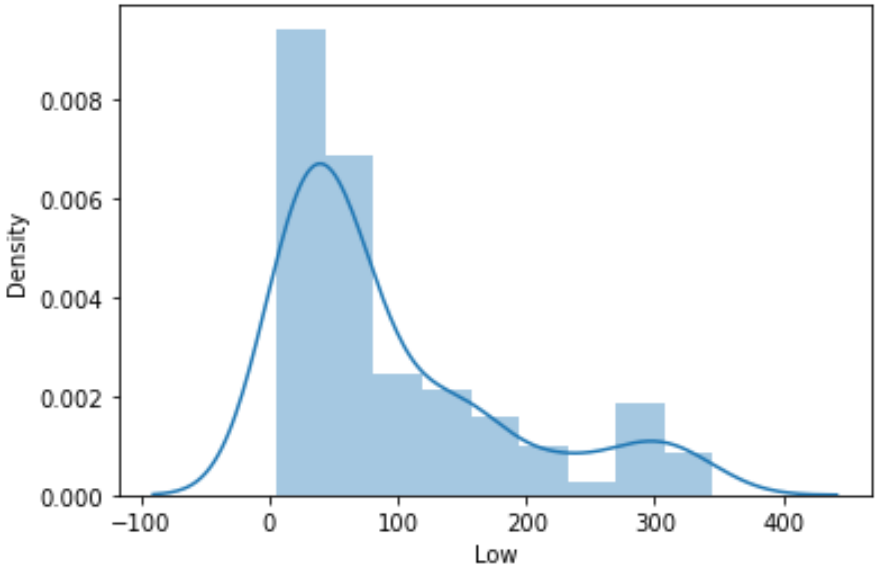
Distribution of Open



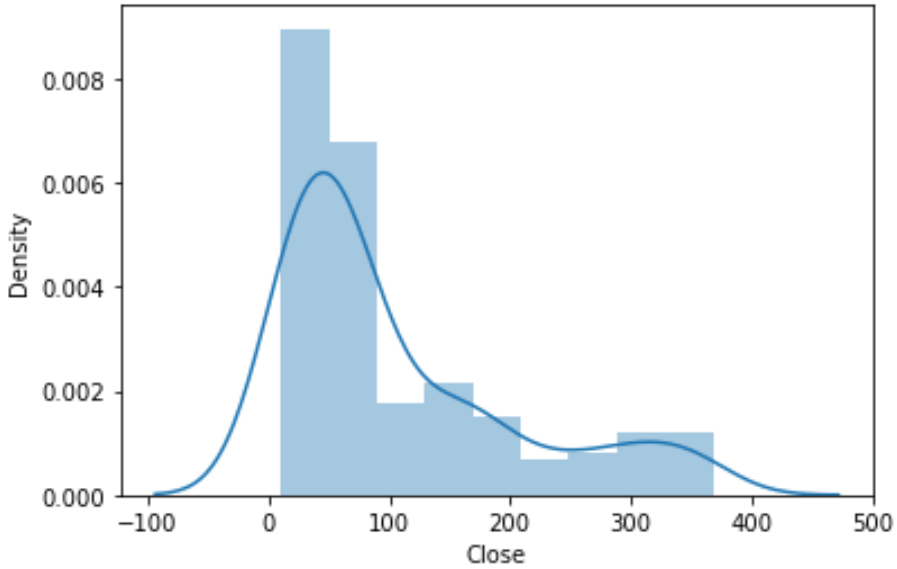
Distribution of High



Distribution of Low



Distribution of Close

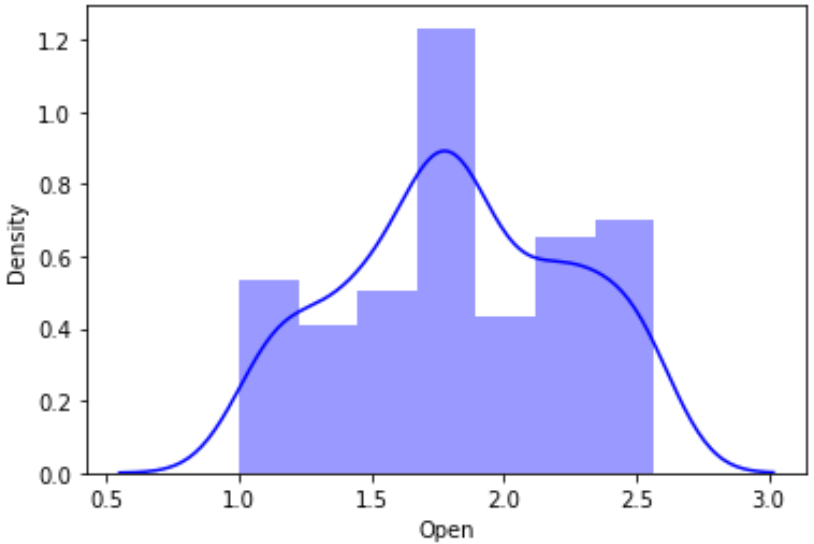


The above distribution of Stock Price is a positively right-skewed distribution. this is not a perfect normal distribution, so we have to apply some kind of transformation to see if it will look like a normal distribution or not. It can be corrected by applying Log Transformation then we'll have a look at how this data behave.

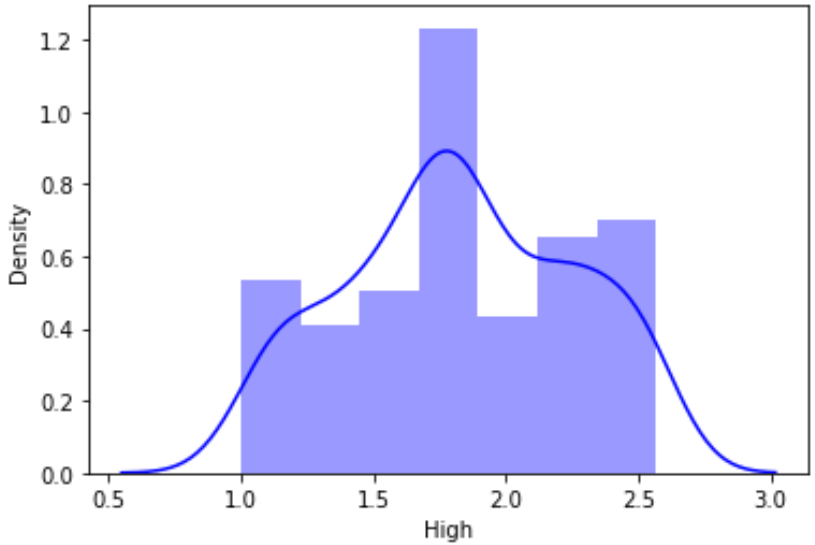
❑ Data Transformation

As observed in the preceding slides, the observed data was found to be positively skewed. We will transform the data to make it uniform before passing it into our machine learning models. for that, we applied a log transformation now Let's have a look at how they will look once the transformation is applied to them.

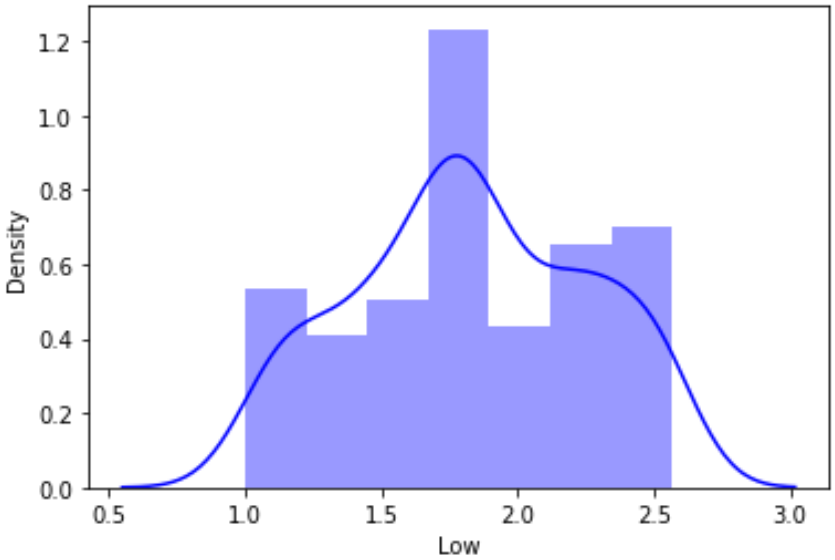
Distribution of Open after log transformation



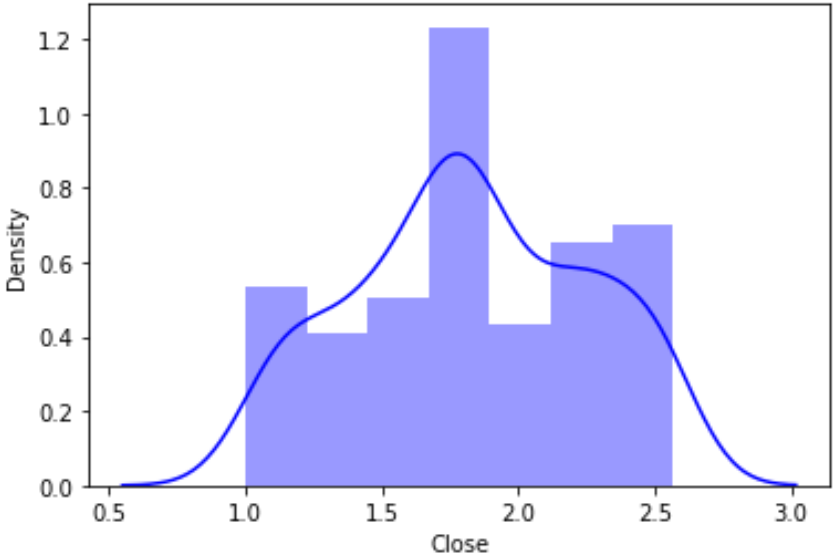
Distribution of High after log transformation



Distribution of Low after log transformation

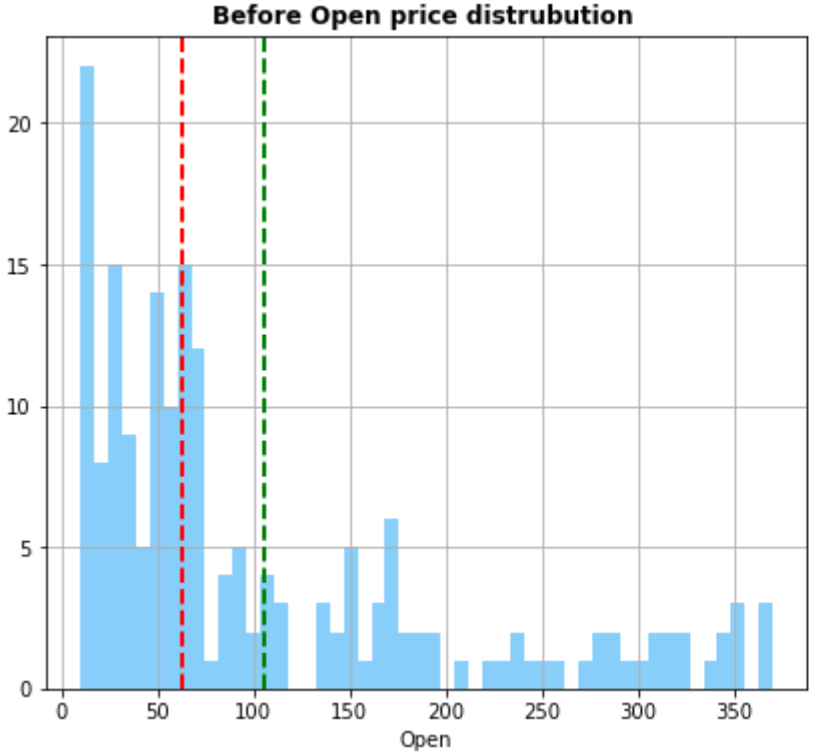


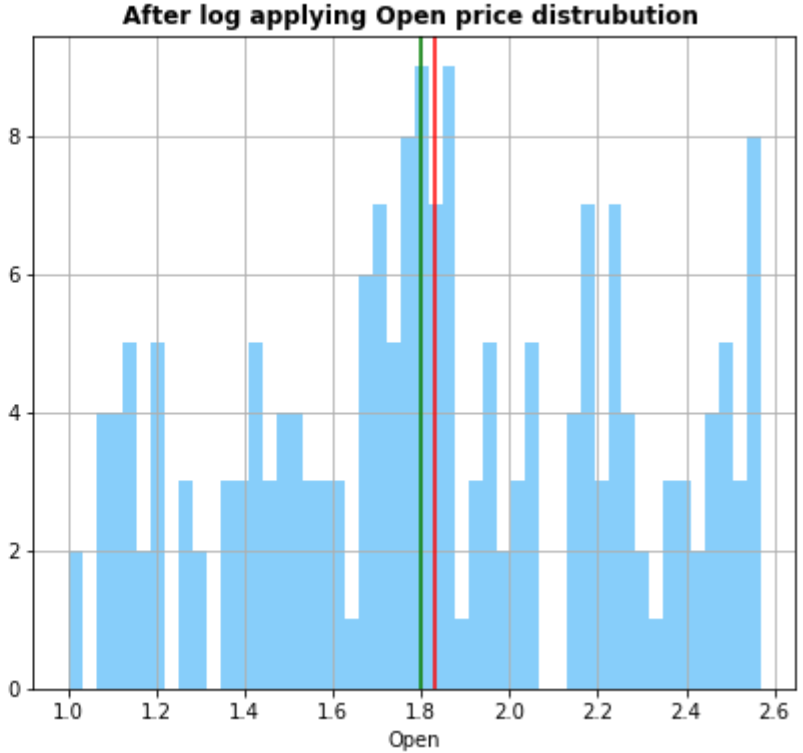
Distribution of Close after log transformation

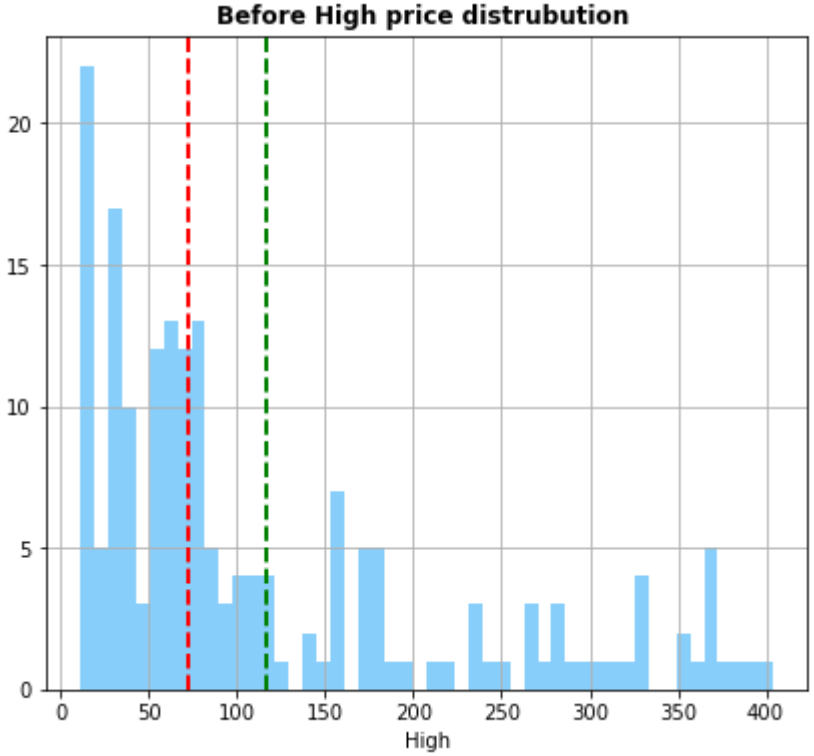


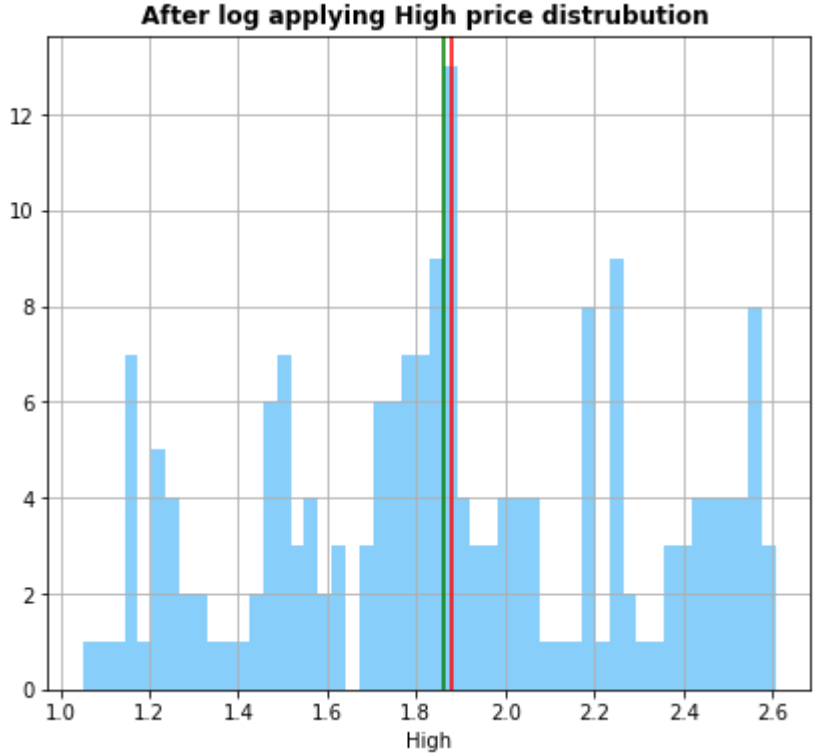
Now we can look at this distribution it is not a perfectly normal distribution but more or less its looking normal distribution.

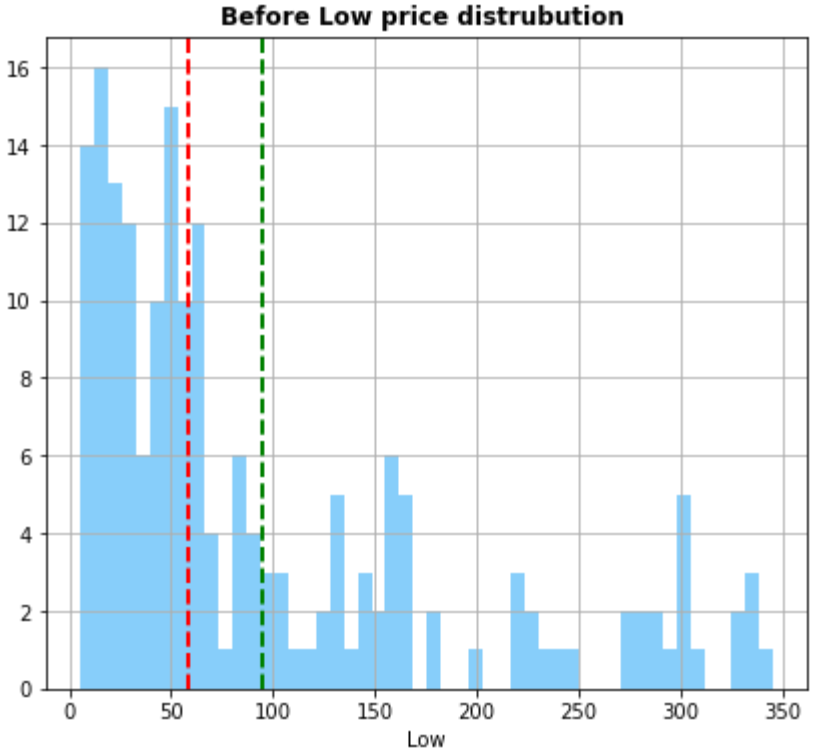
- Plotting histogram for each variable with mean and median of every single variable (before and after log transformation)

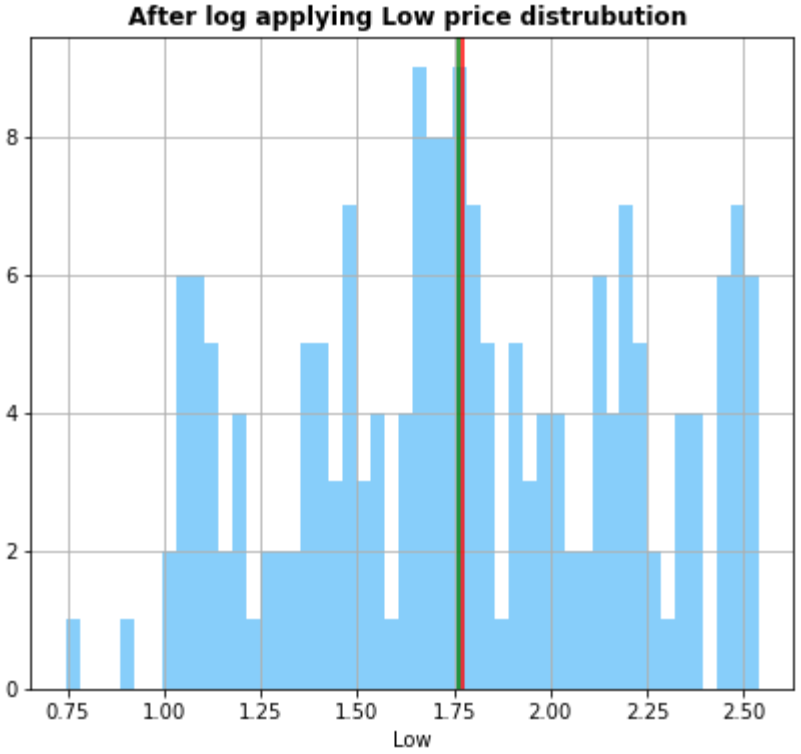


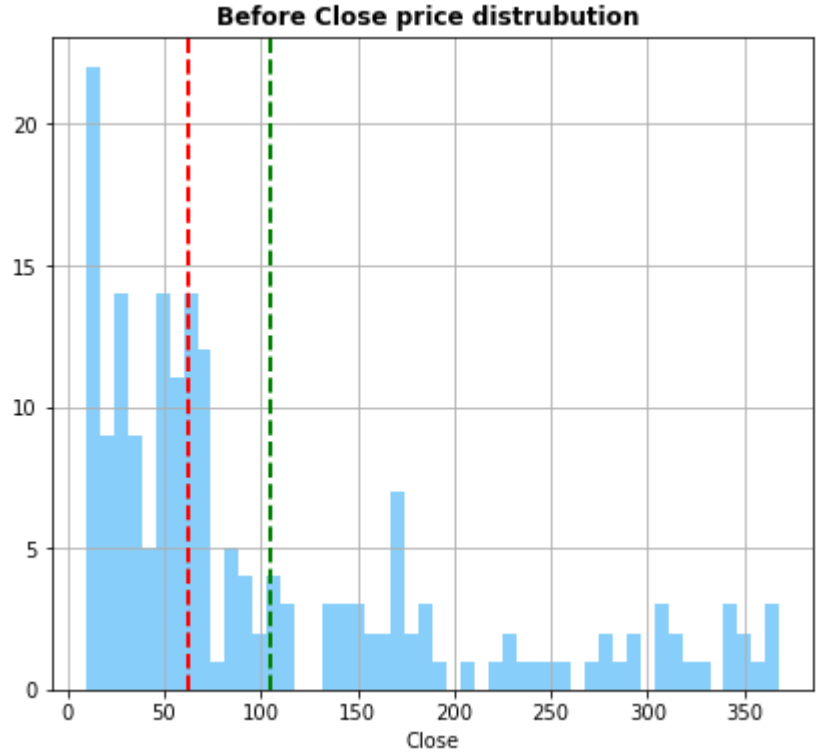


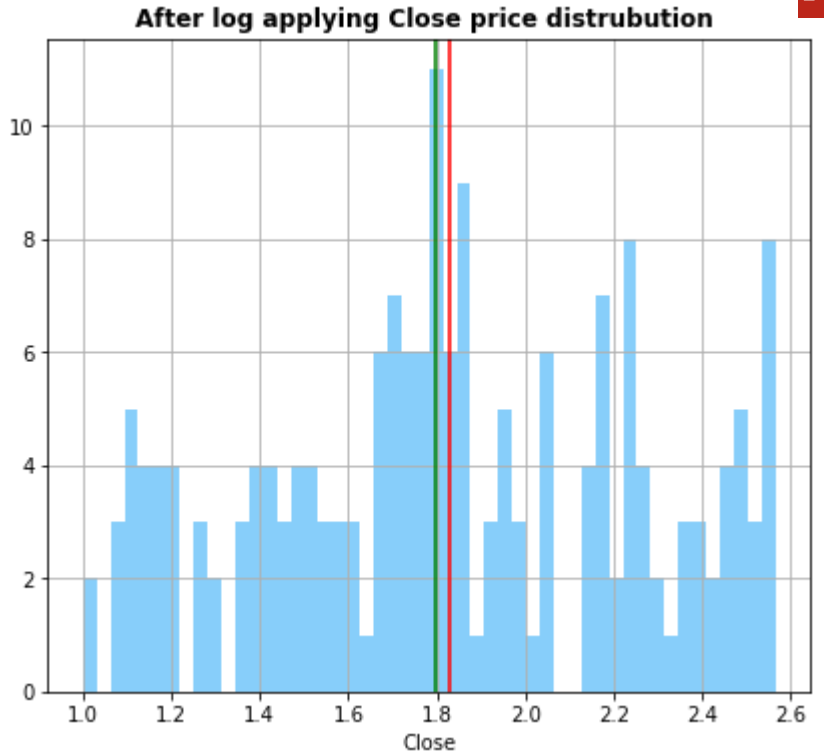

Before High price distrubution
20
15
10
5
0
0
50
100
150
200
250
300
350
400
High


After log applying High price distrubution
12
10
8
6
4
2
0
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
High


Before Low price distrubution
16
14
12
10
8
6
4
2
0
0
50
100
150
200
250
300
350
Low


After log applying Low price distrubution
8
6
4
2
0
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
Low

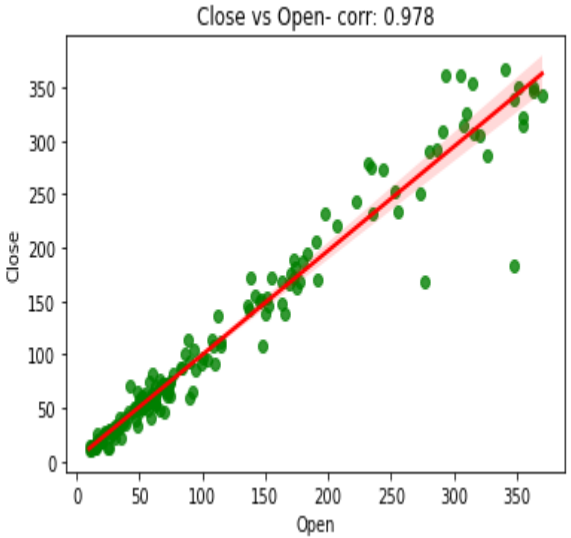


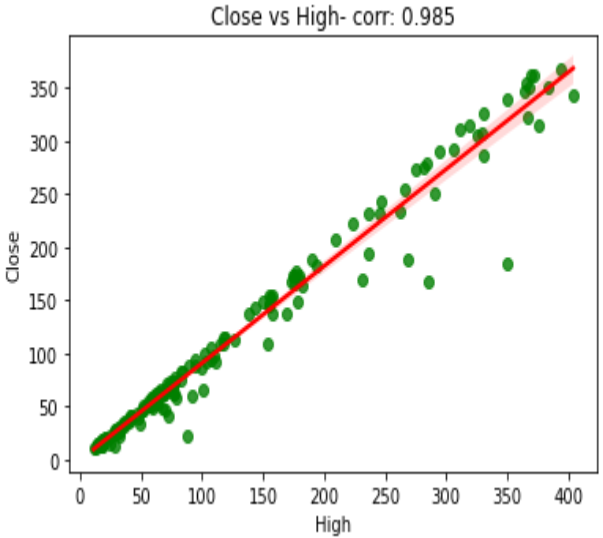


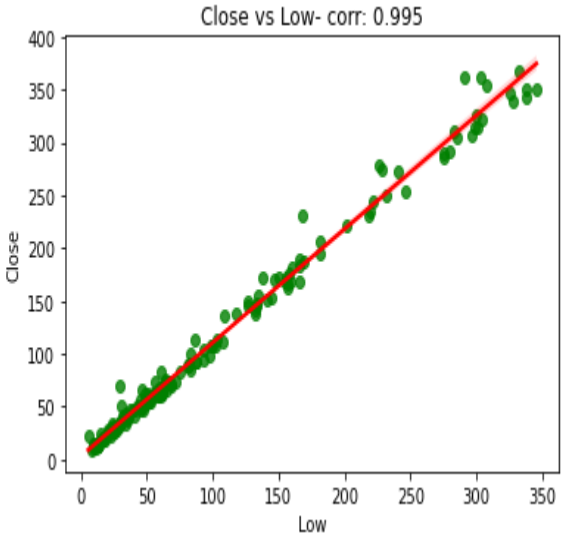
From the above plots, we can see that before applying the log our all features are positively skewed and After the log transformation, now it looks approximately normally distributed and if you'll observe statistically: the mean and median are always close to each other.

Bivariate Analysis

As we can see from the plots, we can conclude that the columns 'Open', 'High', and 'Low' these features are linear relations and high correlations between each independent and dependent variable. that means there is a strong correlation between all the independent variables.

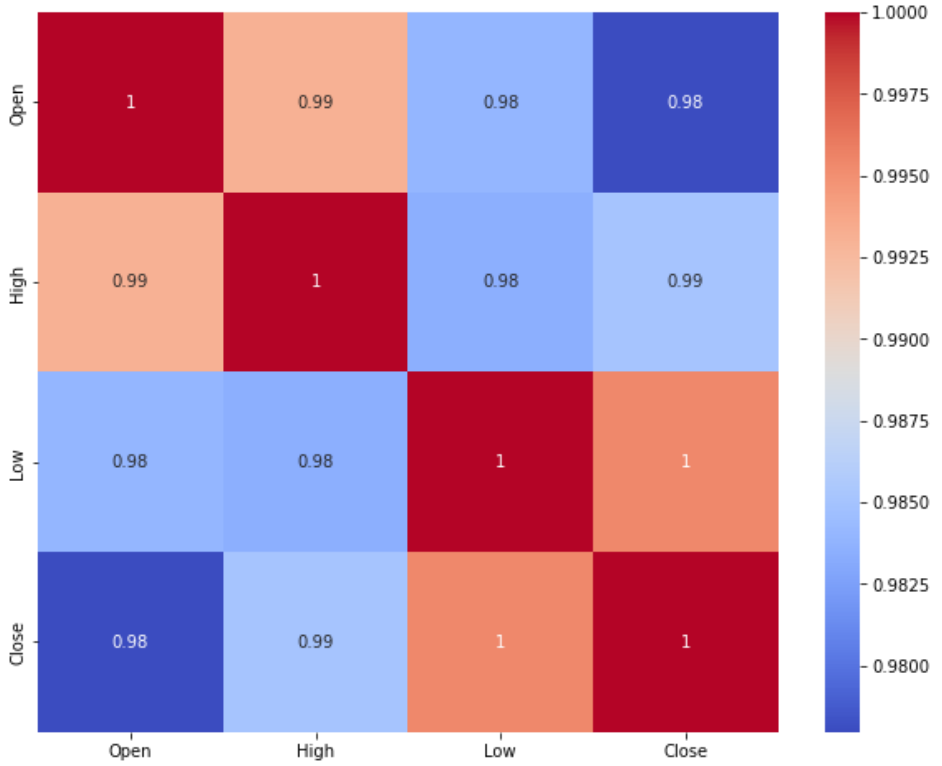






Correlation Analysis

- The heatmap shows some high correlations between variables and also us visualize the correlation of each parameter will respect to every other parameter.
- the heatmap, we can see that Every feature is extremely correlated with each other. This means there is high multicollinearity between each independent column.
- High multicollinearity is not good for fitting the model and prediction because a slight change in any independent variable will give very unpredictable results.
- so We have measured VIF scores in our dataset which means there is high multicollinearity between these variables., we have dropped one of them. then we fitted the model and make a prediction.



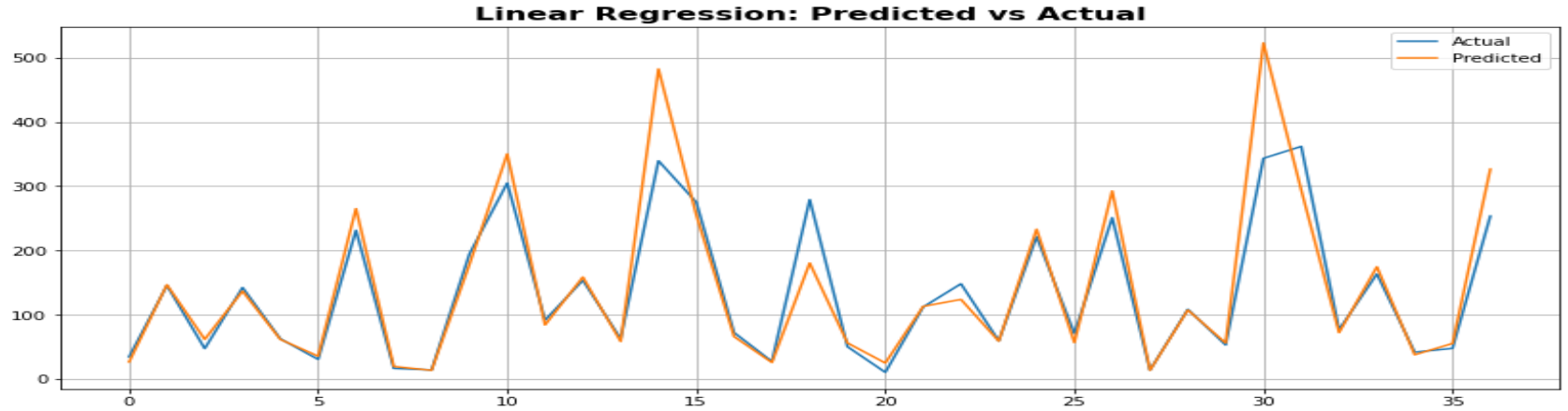
Machine Learning Model

❖ Linear Regression

- Model accuracy is moderate for training as well as test data. Therefore we can conclude that no overfitting.
- Since there is no overfitting.
- Our Linear Regression Model predicted the close price with a 0.064% Mean Absolute error.
- Our model Has a training accuracy of 94.58%.
- R2 value for both training and test data is moderate indicating that the model is fit well on both the datasets
- R2 value shows that our independence variable can describe 94.95% of our dependent variable.
- Adj R2 is about 90.43% .

Training Errors :
MSE: 0.0095
MAE: 0.0668
R2: 0.9457

Testing Error
MSE: 0.0094
MAE: 0.064
R2: 0.9495
Adj_R2: 0.9043

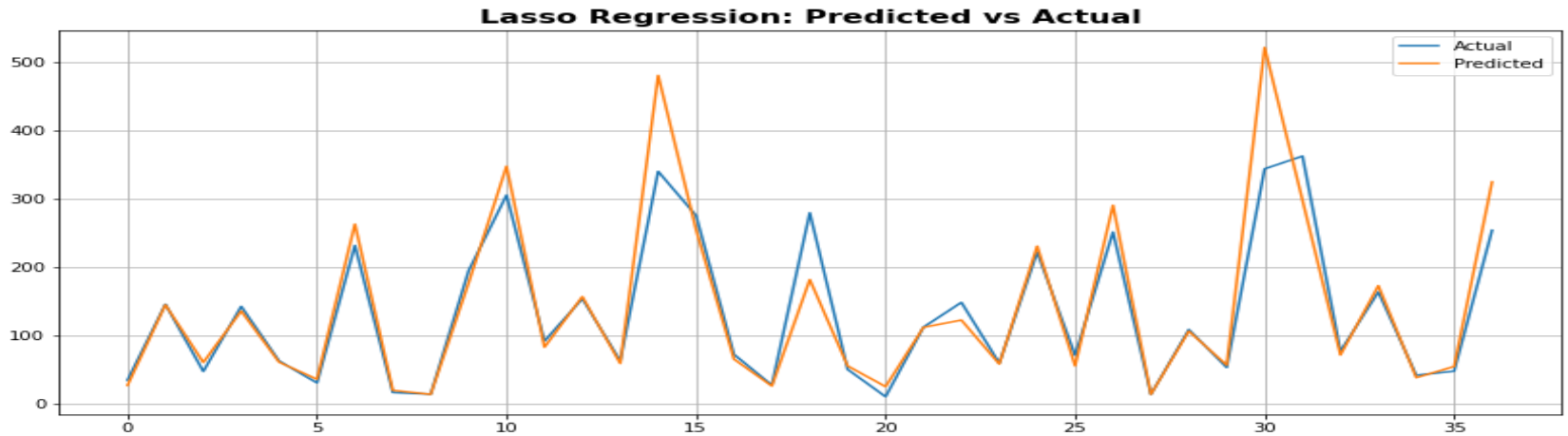


❖ Lasso Regression

- In the Lasso Regression Model accuracy is moderate for training as well as test data.
- Our lasso Regression Model predicted the close price with a 0.066% Mean Absolute error.
- Our model Has a training accuracy of 94.57%.
- R2 value for both training and test data is moderate indicating that the model is fit well on both the datasets
- Here, R2 is about 94.97% which means the model's independent features are able to describe our dependent variable and Adj R2 is about 90.48%.

Training Errors
MSE: 0.0095
MAE: 0.0668
R2: 0.9457

Testing Error
MSE: 0.0093
MAE: 0.0640
R2: 0.9497
Adj R2: 0.9048

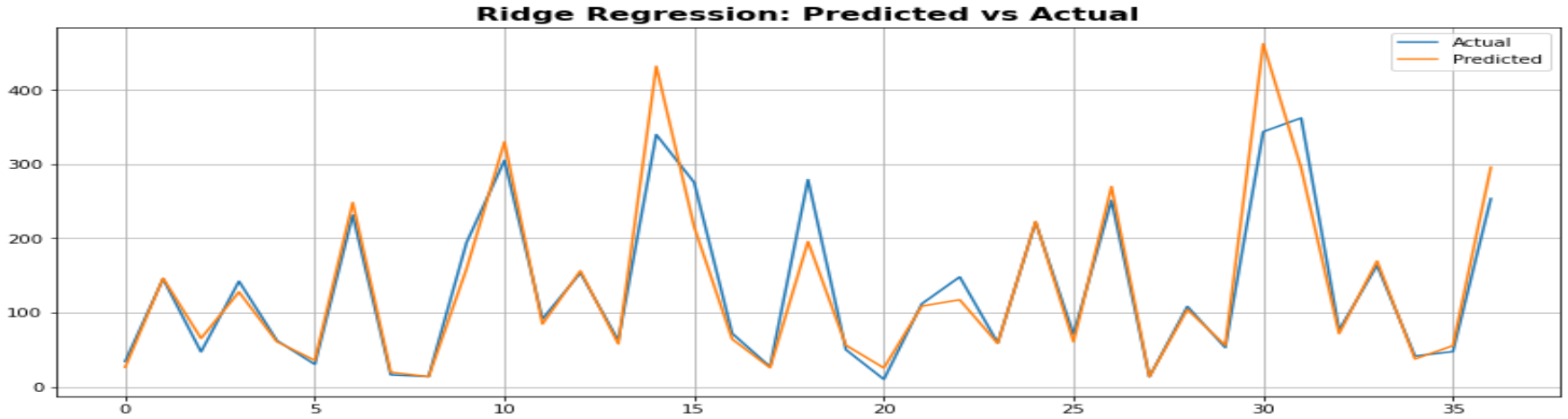


❖ Ridge Regression

- In the Ridge Regression Model accuracy is moderate for training as well as test data.
- Our Ridge Regression Model predicted the close price with a 0.066% Mean Absolute error.
- Our model Has a training accuracy of 94.57%.
- R2 value for both training and test data is moderate indicating that the model is fit well on both the datasets
- Here, R2 is about 95.25% which means the model's independent features are able to describe our dependent variable and Adj R2 is about 91.48%.

Training Errors
MSE: 0.0095
MAE: 0.0668
R2: 0.9457

Testing Error
MSE: 0.00884
MAE: 0.0620
R2: 0.9525
Adj R2: 0.9100

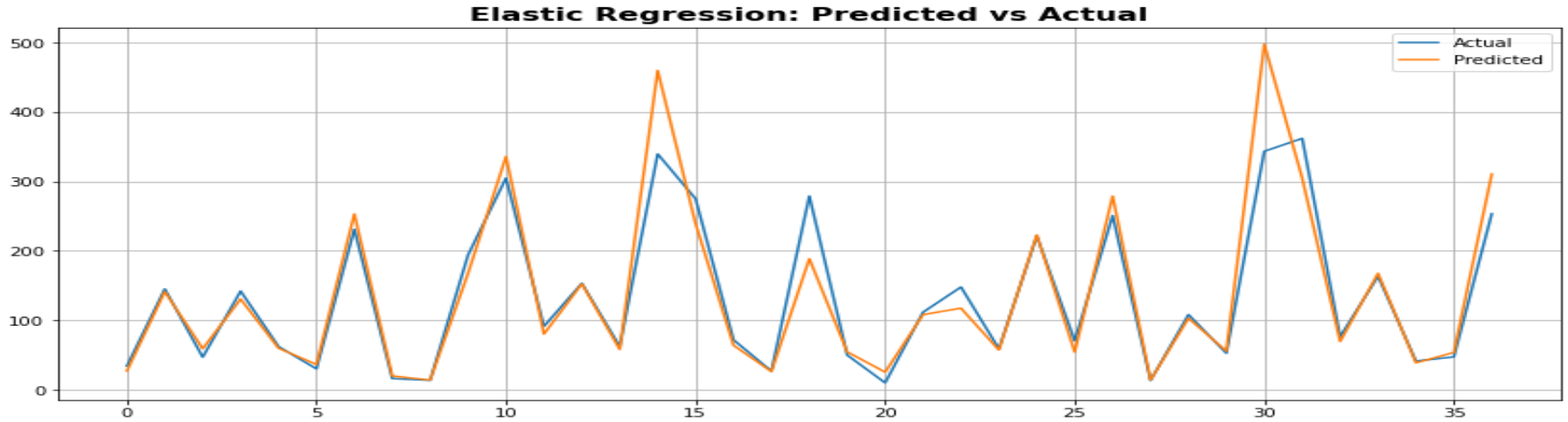


❖ ElasticNet Regression

- In the ElasticNet Regression Model accuracy is moderate for training as well as test data.
- Our ElasticNet Regression Model predicted the close price with a 0.063% Mean Absolute error.
- Our model Has a training accuracy of 82.31%.
- R2 value for both training and test data is moderate indicating that the model is fit well on both the datasets
- Here, R2 is about 95.05% which means the model's independent features are able to describe our dependent variable and Adj R2 is about 90.69%.

Training Errors
MSE: 0.0310
MAE: 0.1413
R2: 0.8231

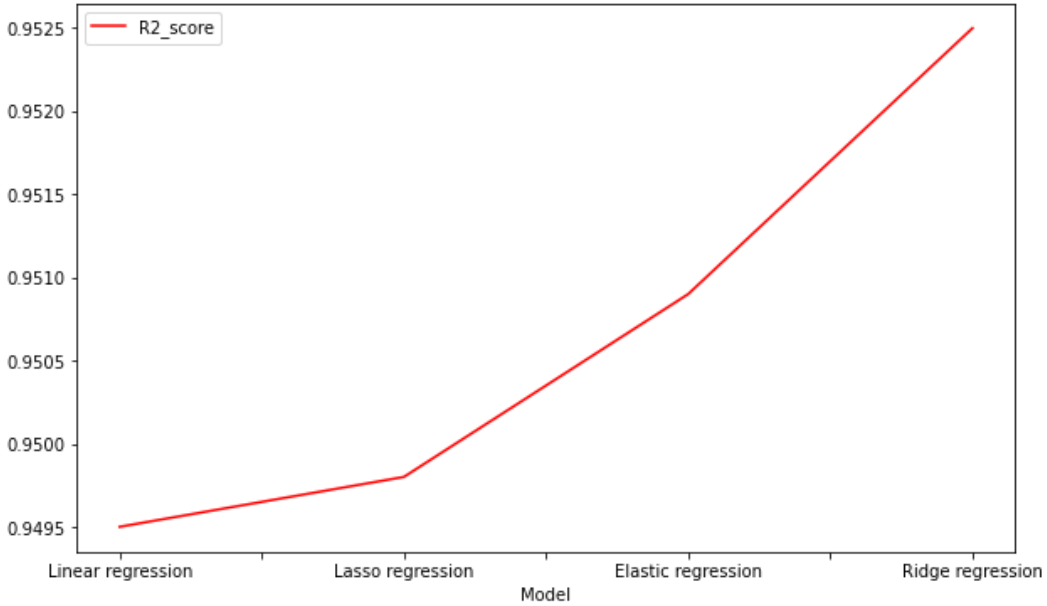
Testing Error
MSE: 0.0091
MAE: 0.0638
R2: 0.9508
Adj R2: 0.9069



❖ Over all Evaluation matrix

Model	MAE	MSE	RMSE	R2	Adj_R2	Training Score
Linear Regression	0.0648	0.0094	0.0970	0.9495	0.9043	0.9458
Lasso Regression	0.0640	0.0094	0.0967	0.9498	0.9048	0.9458
Ridge Regression	0.0621	0.0088	0.0941	0.9525	0.9100	0.9458
ElasticNet Regression	0.0638	0.0092	0.0957	0.9509	0.9069	0.8231

❖ Comparison among all models with R2 Score in one graph



Here we can see that Ridge Regression is giving the highest R2 score of on the Test dataset. Therefore we can say that Ridge regression is giving us optimal results in terms of the test dataset and is best for final prediction.

❑ Conclusion:

- first We started with data inspection and viewed the data distribution.
- With the help of visualization we checked that from 2018 onwards there is a sudden fall in the stock closing price.
- And Again With the help of a distribution plot we saw that our data is rightly skewed which doesn't look good in the viewing of the statistical hypothesis. So we applied some kind of transformation Log Transformation to convert it into a normal distribution.
- Target Variable is strongly dependent on Independent Variables.
- we have performed VIF to reduce multicollinearity
- Insights of all the models, A simple linear regression model was built and it was evaluated using accuracy, MSE, RMSE, r2_score, and Adj_R2, mean absolute percentage error.
- Linear Regression,, Lasso and Ridge are performing better than Elasticnet models with training accuracy of 94.58%, 94.58%, and 94.58% respectively.
- Apart from Linear Regression, Lasso, and Ridge, ElasticNet is also performing better but has less training accuracy.
- Ridge and ElasticNet have performed far much better after Cross-validation which is R2 is about 95.25% and 95.09% respectively.
- R2 and Adjusted R2 are around the range 95% and 91% in each model.

Thank you!