Capstone Project-2

Yes Bank Stock Closing Price Prediction

ABHISHEK SHUBHAM

CONTENTS

Problem Statement:

Perform regression analysis on Yes Bank Stock Price dataset using multiple models to predict the closing price of Yes Bank stock at the end of every month and compare the evaluation metrics for all of them to find the best model.

Data Pipeline

1. **Data Preprocessing**: At this stage, we check for missing values, duplicate values and treat them accordingly. Also reviewed the descriptive statistics of the numerical features. Furthermore, we check for the features present in our dataset and transform the columns if necessary.

2. **Exploratory Data Analysis:** In EDA we conducted Univariate and Multivariate Analysis of different features in order to better understand their spread, pattern and relationship with other features. It helps us to better understand our data and make inferences out of them.

3. **Feature Engineering:** At this stage we created new columns from the existing features which is helpful for model accuracy and prediction.

4. **Feature Selection:** At this stage, we did multicollinearity check and removed correlated features and selected features which would yield better result from our Regression models.

5. **Model Building:** In this part, we apply various Regression models to understand which one would give us best result.

Data Summary

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 with 185 observations. It contains the following features:

1. **Date**: It denotes date of investment done (in our case we have month and year).

2. **Open**: The opening price is the price at which a security first trades upon the opening of an exchange on a trading day i.e. buyers and sellers meet to make deals with the highest bidder, the opening price may not have to be the same as the last day's closing price.

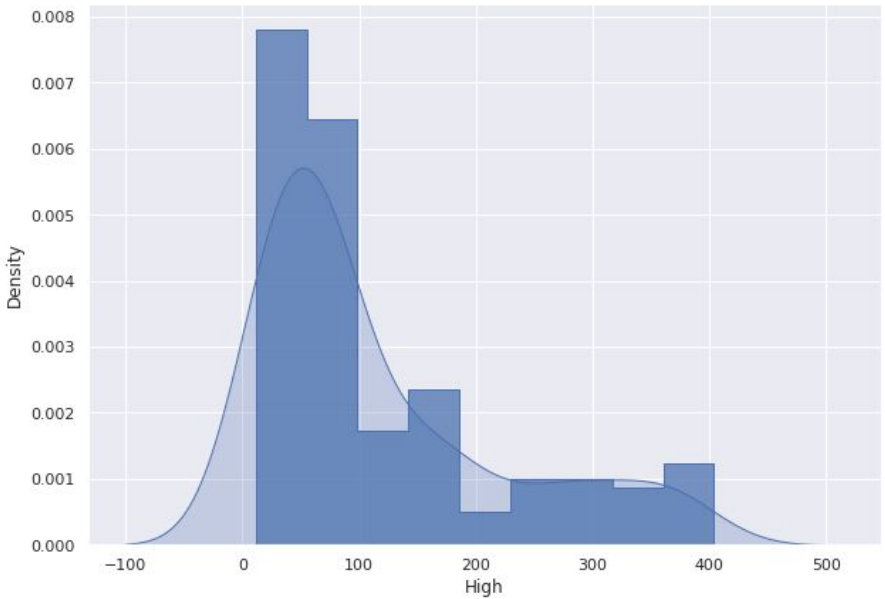
3. **High**: It’s the highest price at which a stock traded during a period.

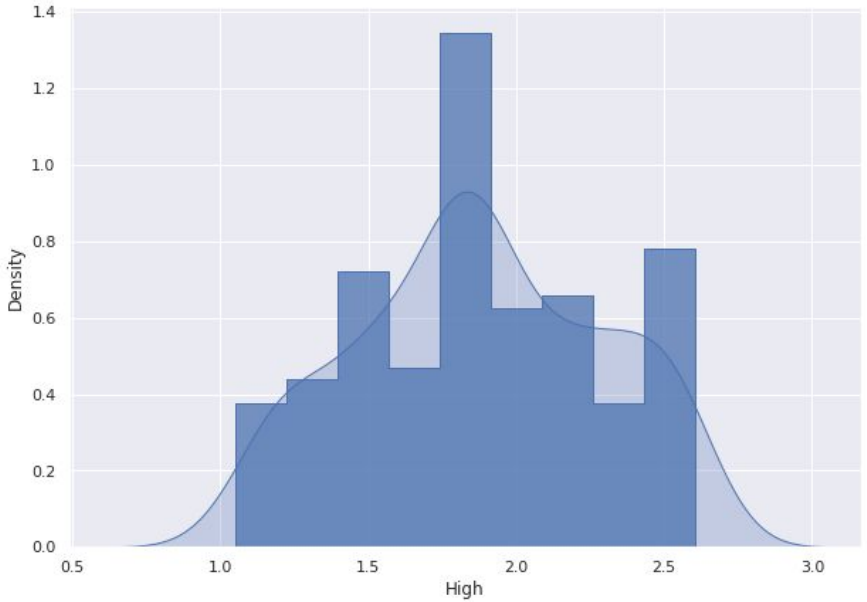
4. **Low**: It’s the lowest price at which a stock traded during a period.

5. **Close**: Close refers to the price of an individual stock when the stock exchange closed for the day.

EDA: (Univariate Analysis)

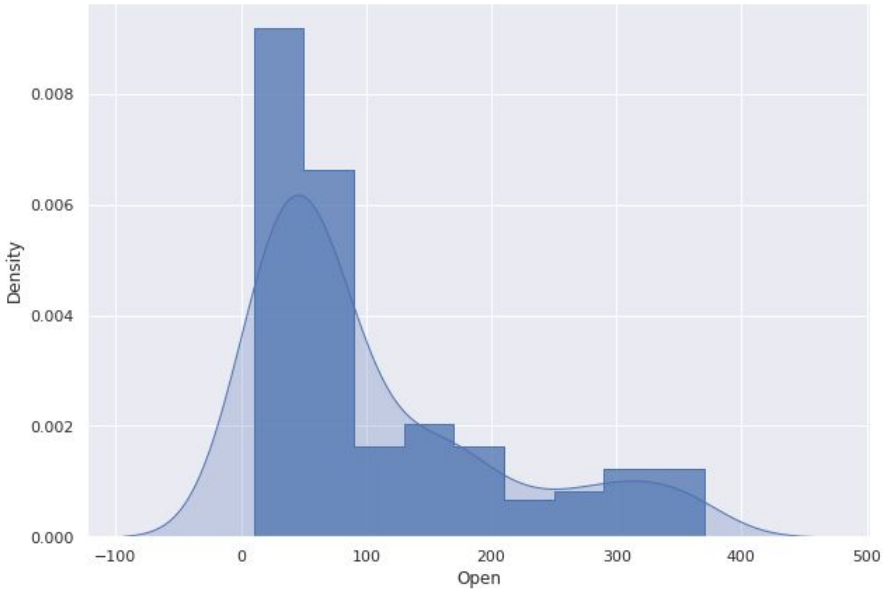
Distribution and Transformation of ‘High’

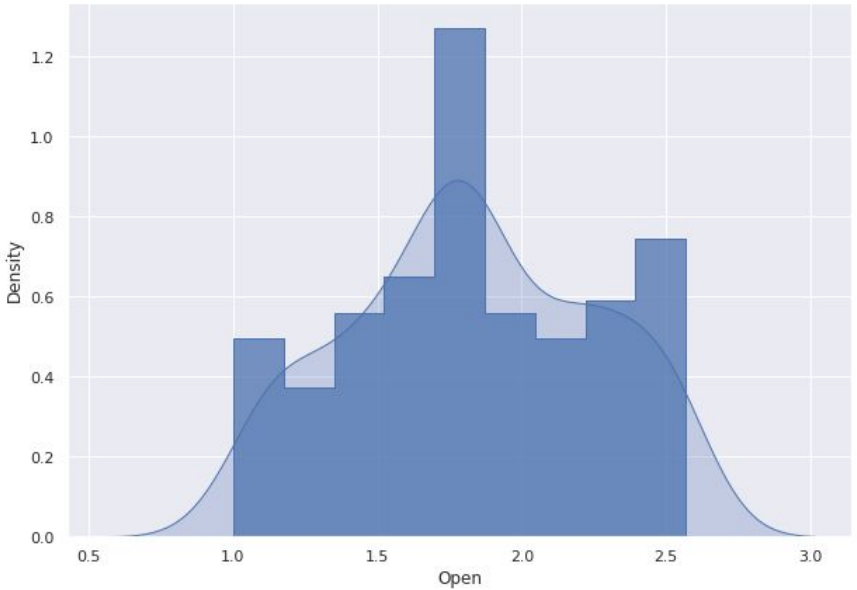




Univariate Analysis(cont.)

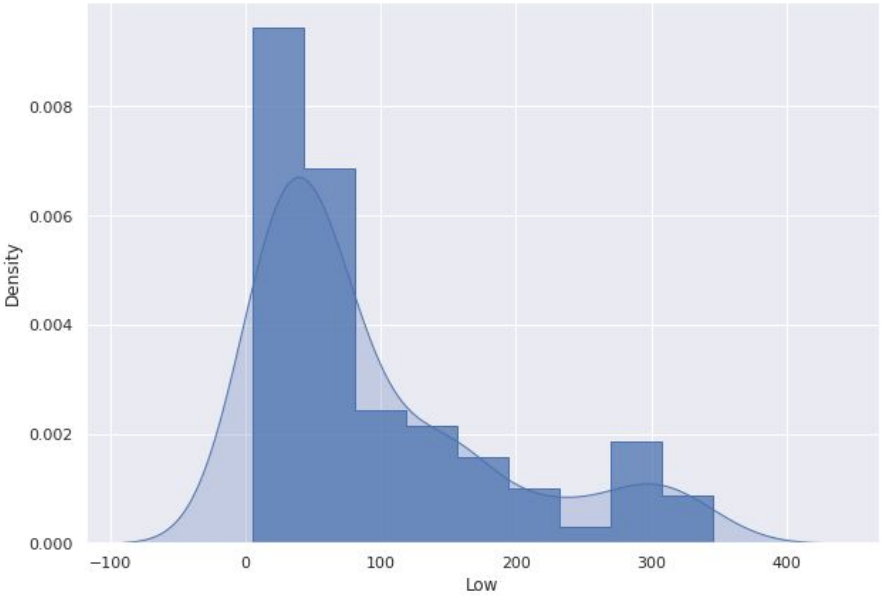
Distribution and Transformation of ‘Open’

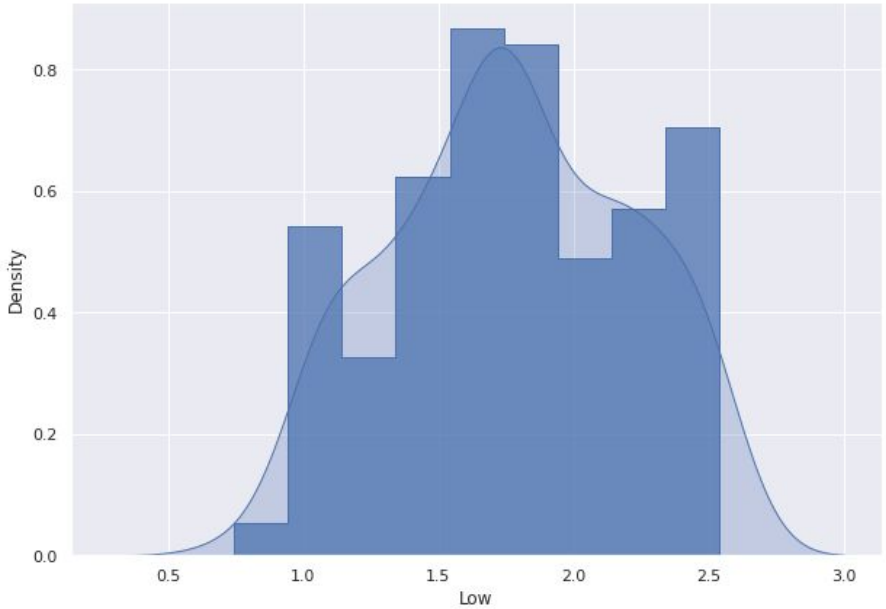




Univariate Analysis(cont.)

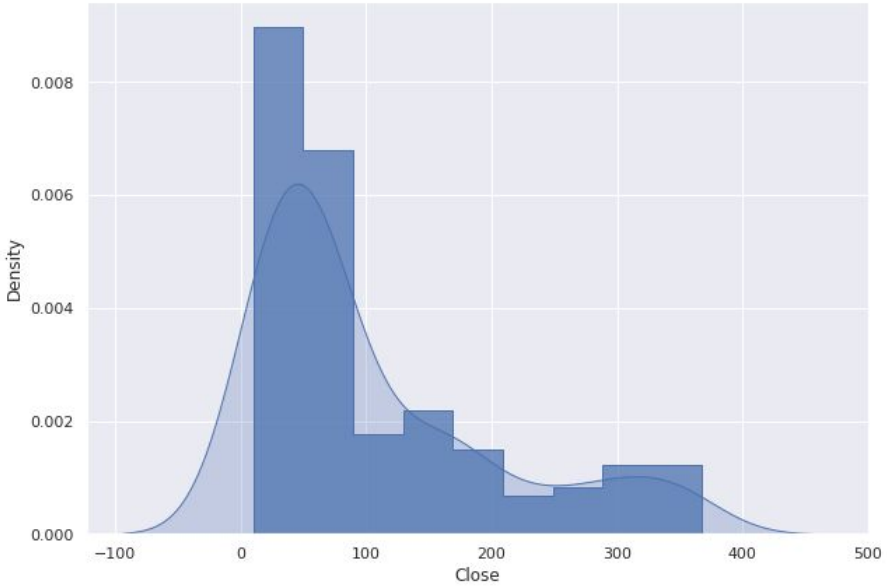
Distribution and Transformation of ‘Low’

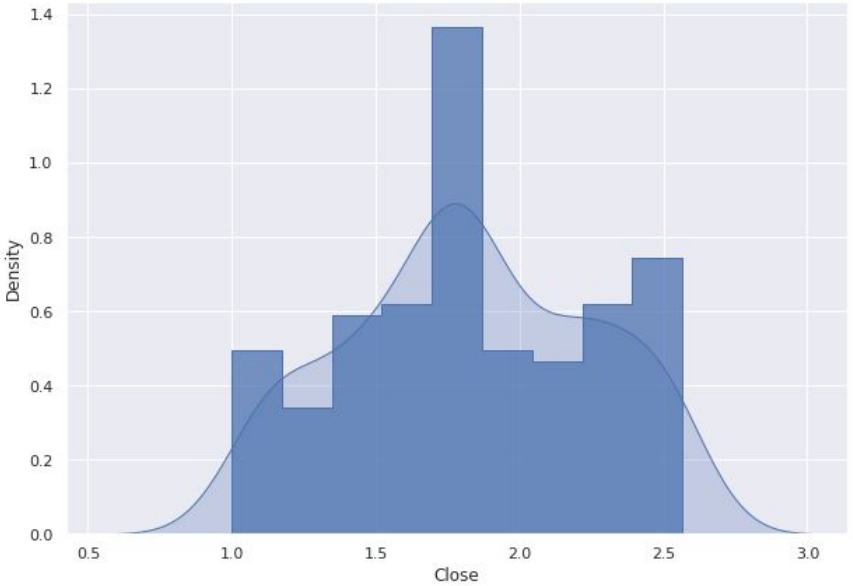




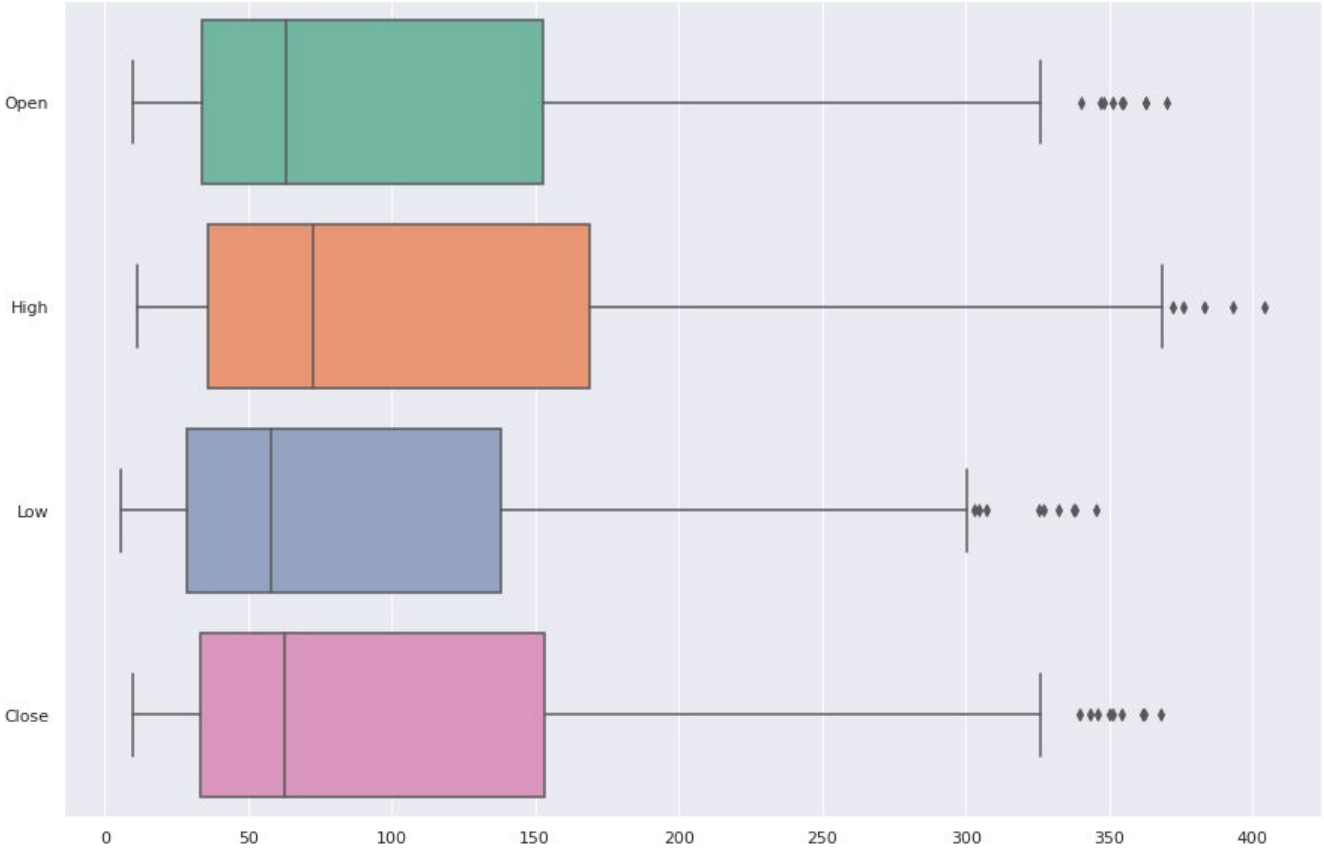
Univariate Analysis(cont.)

Distribution and Transformation of ‘Close’





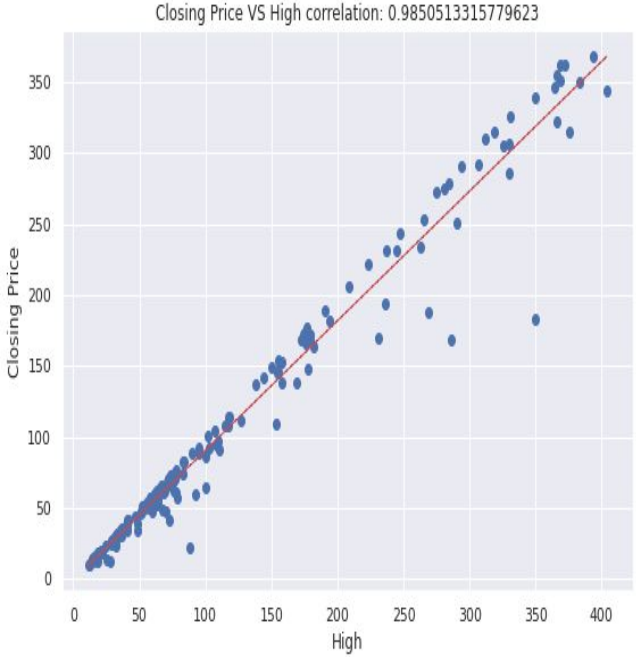
Univariate Analysis(cont.)

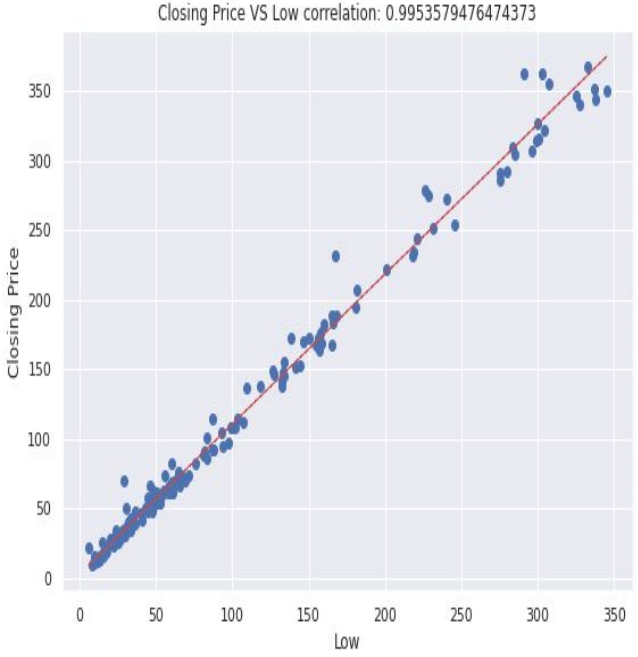


- From these Boxplot's it can be seen that the median values of Yes Bank Stock price for all the features lies between Rs 50 to Rs 80.
- There is a existence of outliers in all the features.

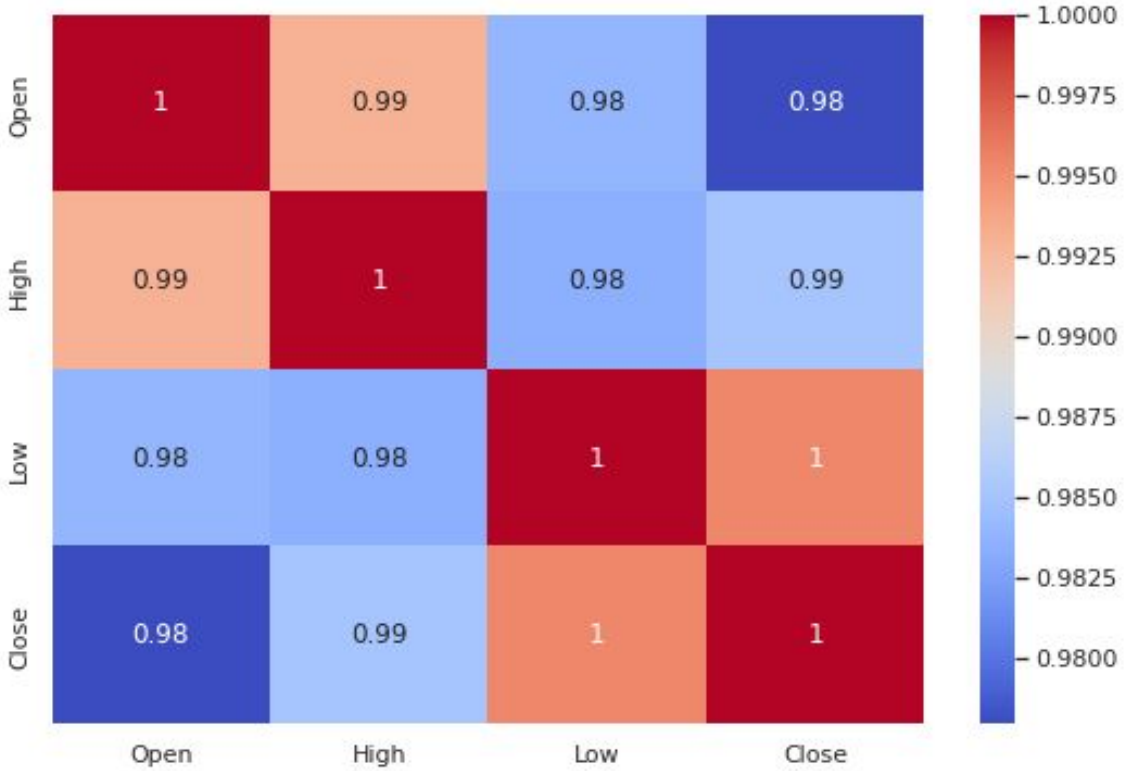
Bivariate Analysis

Bivariate Analysis



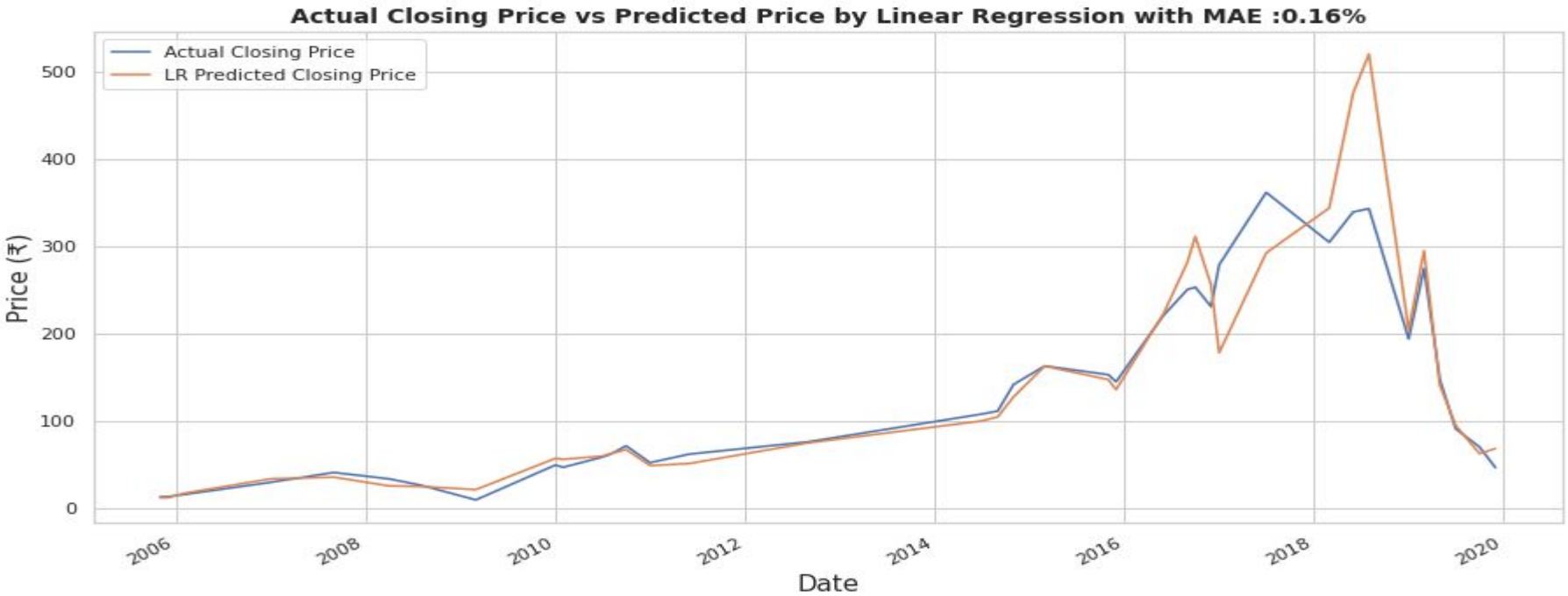


Correlation Matrix



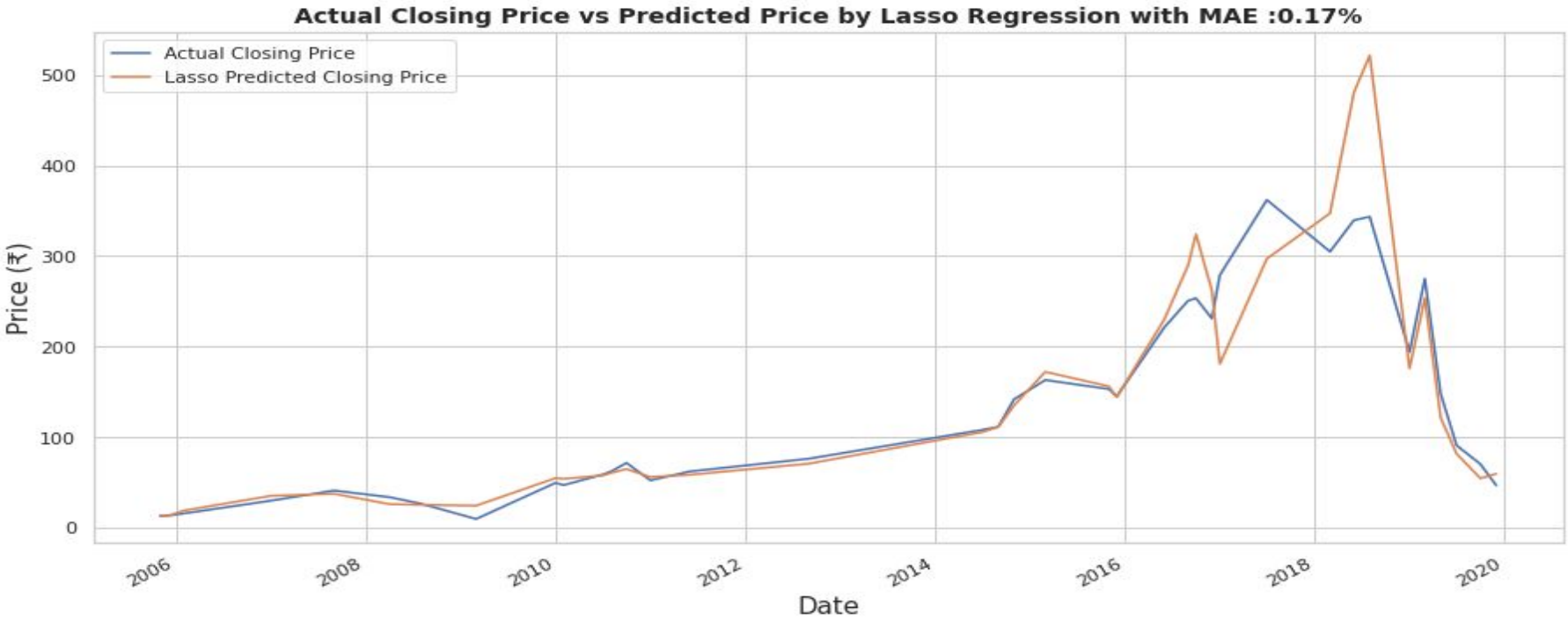
- Looking at the heatmap we can say that all the features are highly correlated.
- Feature 'Low' has highest correlation of 1 with dependent variable.
- There exists multicollinearity between Independent variables which means that we would need to remove one or more features from our dataset during feature selection.

Linear Regression



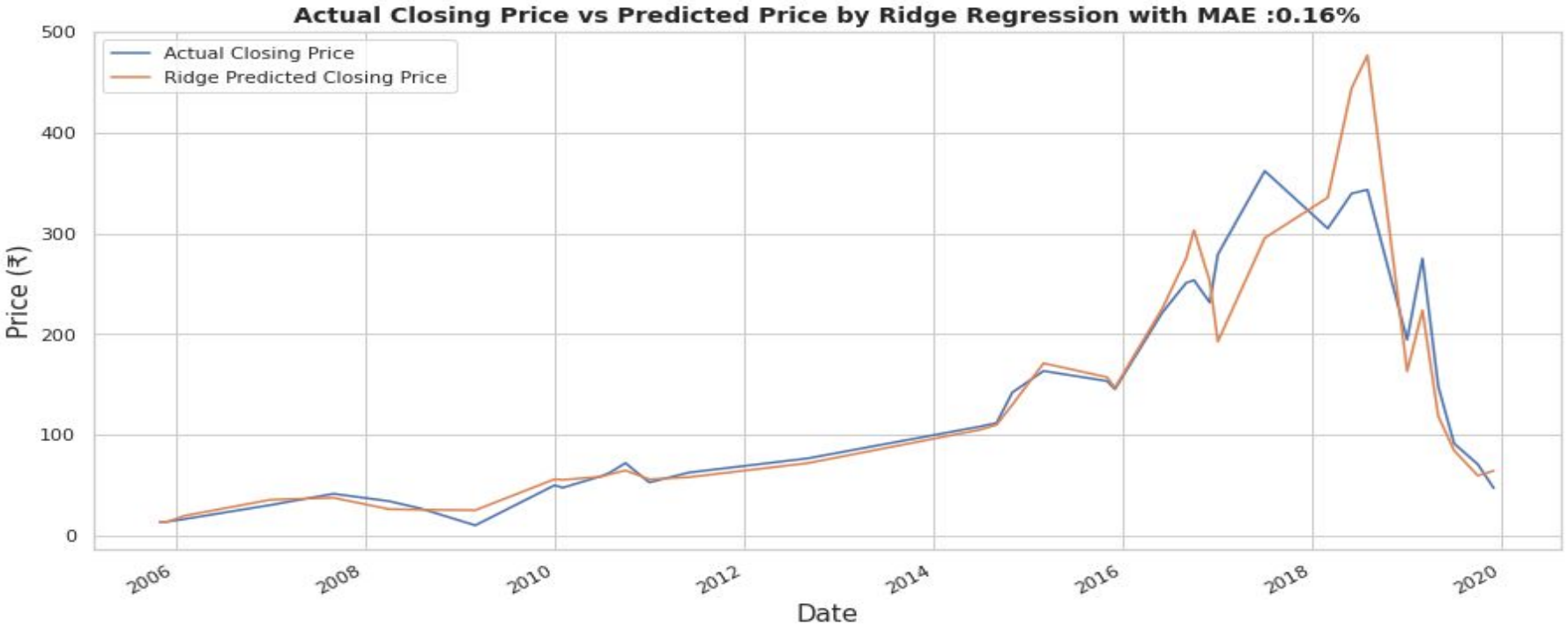
- Our Linear Model predicted the Close Price with 0.16% Mean Absolute Error.
- R2 score tells us that the selected independent features is able to describe 95.5 % of the dependent variable.
- Adjusted R2 score is 91.47%. This score is always less than or equal to R2 score.

Lasso Regression



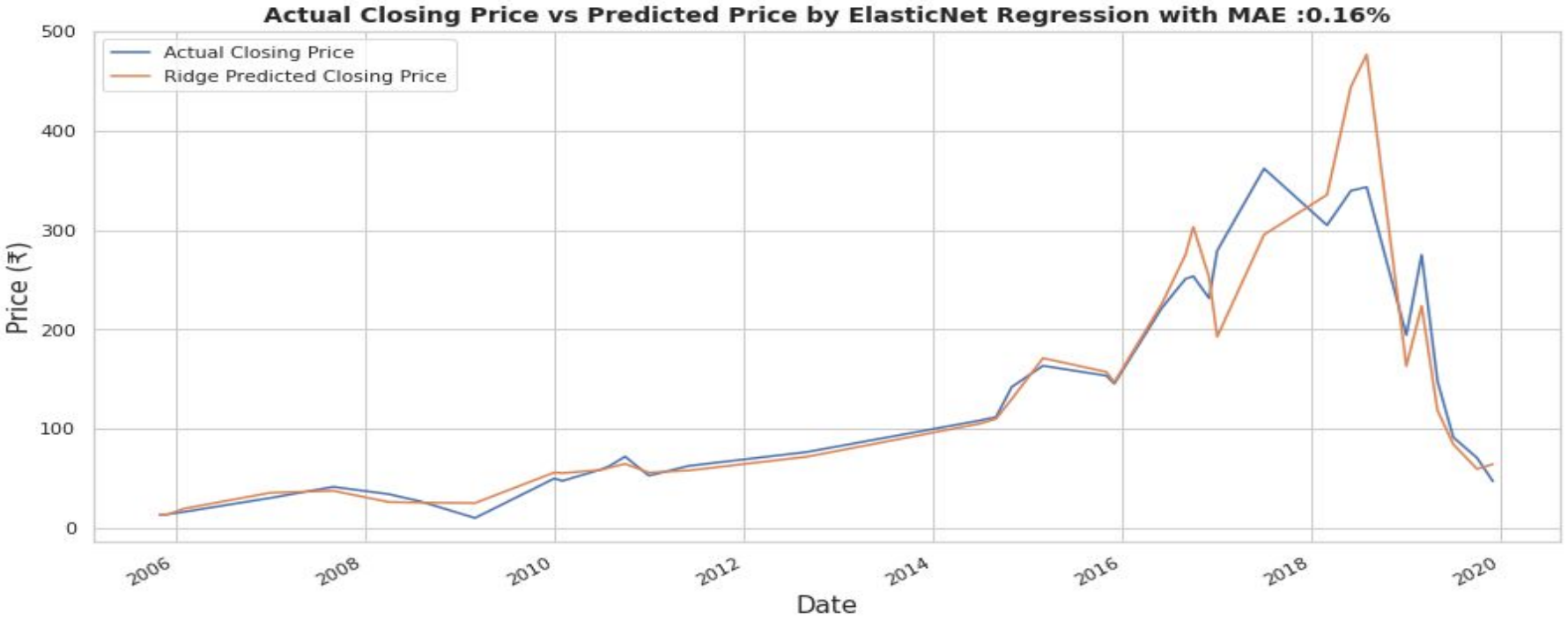
- Our Lasso model predicted the Closing Price with 0.17% Mean Absolute Error.
- R2 score is about 94.96%.
- Adjusted R2 score is 90.46%. We will consider adjusted R2 because we have too many independent features.

Ridge Regression



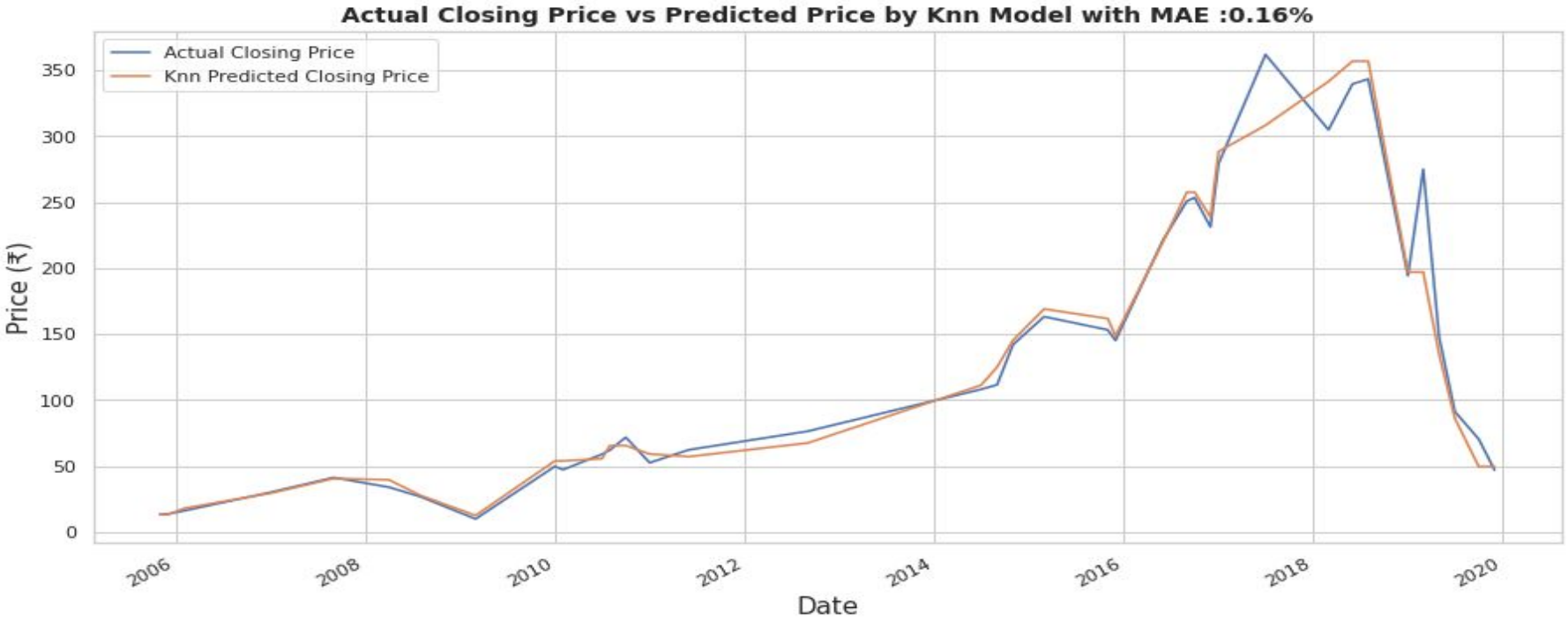
- Ridge model predicted the closing price with 0.16% MAE.
- R2 score for the model is about 95.25%.
- Adjusted R2 score is about 91%.

Elastic Net Regression



- Elastic Net model predicted the closing price with Mean Absolute Error 0.16%.
- R2 score is about 95.11%.
- Adjusted R2 score is about 90.74%.

K-Nearest Neighbours



- Knn model predicted the closing price with Mean Absolute Error 0.16%.
- R2 score is about 98.68%.
- Adjusted R2 score is about 97.51%.

Comparison of all Model Prediction in one graph

- All Models have good R2 and Adjusted R2 score.
- Knn Model has nearest prediction with actual closing price.

Model Comparison Matrix

	Linear Regression	Lasso Regression	Ridge Regression	Elastic Net Regression	K-nearest Neighbours
MSE	0.008379	0.009377	0.008848	0.009096	0.002445
RMSE	0.091535	0.096833	0.094061	0.095375	0.049452
R2	0.955021	0.949664	0.952505	0.951169	0.986872
Adjusted R2	0.914777	0.904627	0.910009	0.907478	0.975126

Conclusion

- **Target Variable is strongly correlated with all independent features.**
- **Low feature has the highest correlation with the dependent Variable.**
- **K-nearest Neighbours has best performance with R2 score 0.986872 followed by Linear Regression model with R2 score 0.955021.**
- **The accuracy for each models is more than 90%**

Thank you