Capstone Project-II

Linear Regression- Yes Bank Stock Closing Price Prediction

By- Mahima Phalkey
Sameer Satpute

CONTENTS:

1. Agenda
2. Problem Statement
3. Introduction
4. Data Summary
5. Basic Observations
6. EDA
7. Checking Multicollinearity
8. Data Modelling
9. Conclusion
10. Challenges
11. References

AGENDA

- We are having a dataset with the monthly stock price details for Yes Bank. The objective of this project has been to apply different models to check whether the prices/movement of the stock can be predicted using features and past performance by using Linear Regression

- Looking at the various features of the dataset, we can understand the relationships between them and accordingly pass the required parameters in the model to train it and ultimately predict the closing price.

PROBLEM STATEMENT

✓ Yes Bank Limited is an Indian private sector bank headquartered in Mumbai, India, and was founded by Rana Kapoor and Ashok Kapur in 2004.

✓ It offers a wide range of differentiated products for corporate and retail customers through retail banking and asset management services.

✓ On 5 March 2020, in an attempt to avoid the collapse of the bank, which had an excessive amount of bad loans, the Reserve Bank of India (RBI) took control of it.

✓ 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.

INTRODUCTION

YES BANK:

✓ Data on Yes Bank's monthly stock prices have been provided to us.

✓ This dataset is dated between July 2005 and November 2020.

✓ Due to the recent default fraud case of Rana Kapoor, the bank has been making headlines.

✓ By analyzing the dataset, we were able to predict the stock closing price for the bank based on other parameters.

DATA SUMMARY

YES BANK
YES BANK
YES BANK
YES BANK
YES BANK

Dataset Values

Date: In our data, it's a monthly observation of stock since it was listed.

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

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

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

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

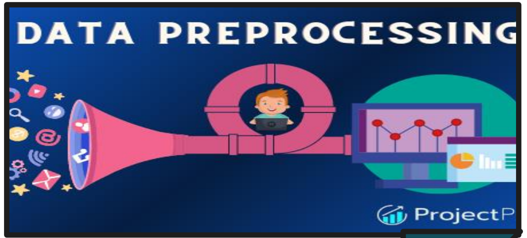
OBSERVATIONS

ne picture can't be displayed.

1) The shape of our dataset is 185 rows and 5 columns
2) Datatype of Date is given as object which we need to change that to Date Time
3) Yes bank stock listed on the month of July 2005. We have data available from July 2005 to November 2020
4) From the above statistical information we can see that it is not a normal distribution as mean and 50% values are having a lot of difference
5) There are no duplicates present
6) There are no null values present

EDA

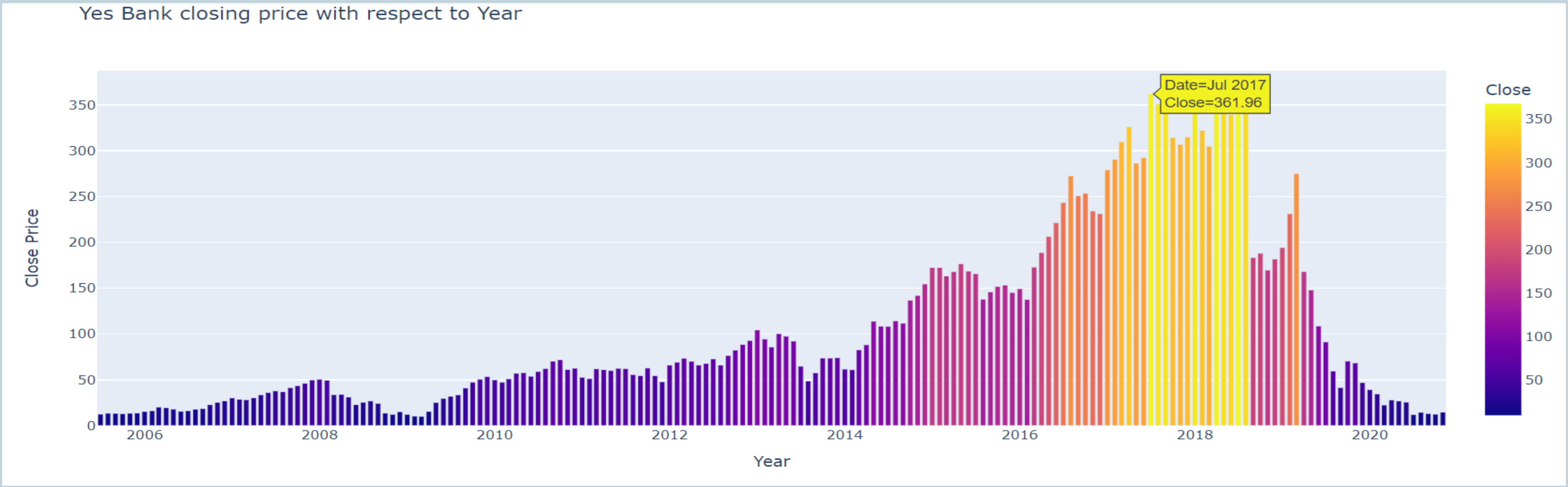
1)



The datatype of the Date column was given as an object which we need to change to Datetime datatype.

EDA (Visualizations)

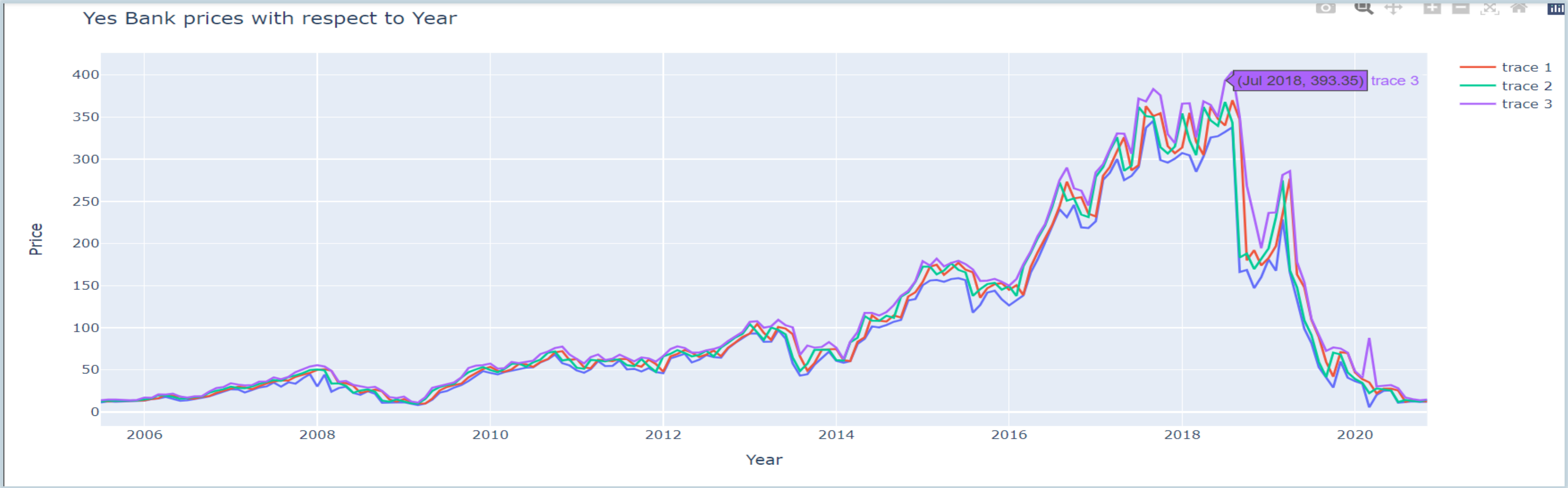
Yes Bank closing price with respect to Year



- Here we can see that the stocks were high from 2017 to 2018 but they dropped after 2018 because of a fraud case regarding Rana Kapoor.

EDA (Visualizations)

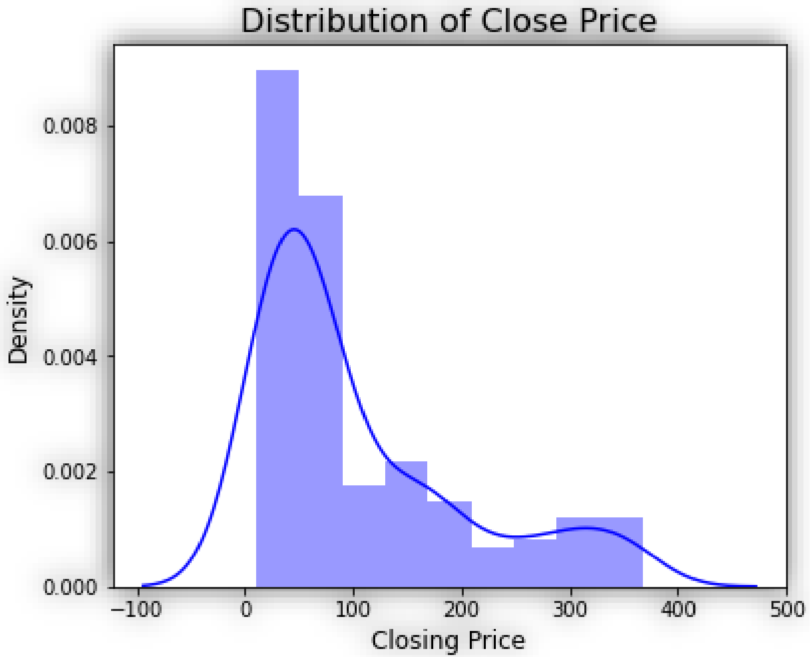
Yes Bank prices with respect to Year



- We can see in 2017 to 2019 there can be high action seen because of the difference in high and low lines.
- We can take the closing price of the stock as the dependent variable as it is the final price of that day.

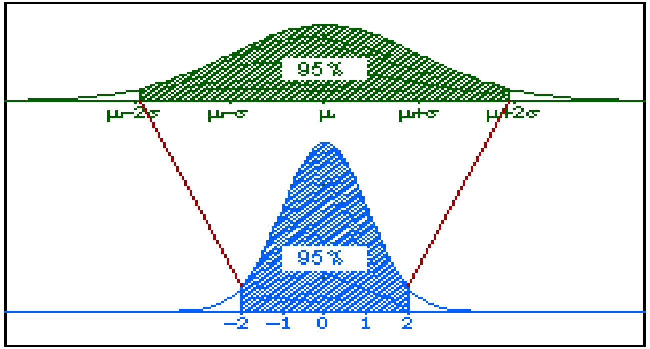
EDA (Visualizations)

Distribution of dependent variable Close Price of stock.



- It is a rightly skewed distribution.
- We need to do log transformation to make it normal distribution

TRANSFORMATIONS



- ✓ We know that in the regression analysis the response variable should be normally distributed to get better prediction results.
- ✓ Most of the data scientists claim they are getting more accurate results when they transform the independent variables too.
- ✓ It means skew correction for the independent variables.
- ✓ Lower the skewness better the result.

- ✓ Below are some types of methods or ways to deal above type of problem.

I. **<u>square-root for moderate skew:</u>** sqrt(x) for positively skewed data, sqrt(max(x+1) - x) for negatively skewed data

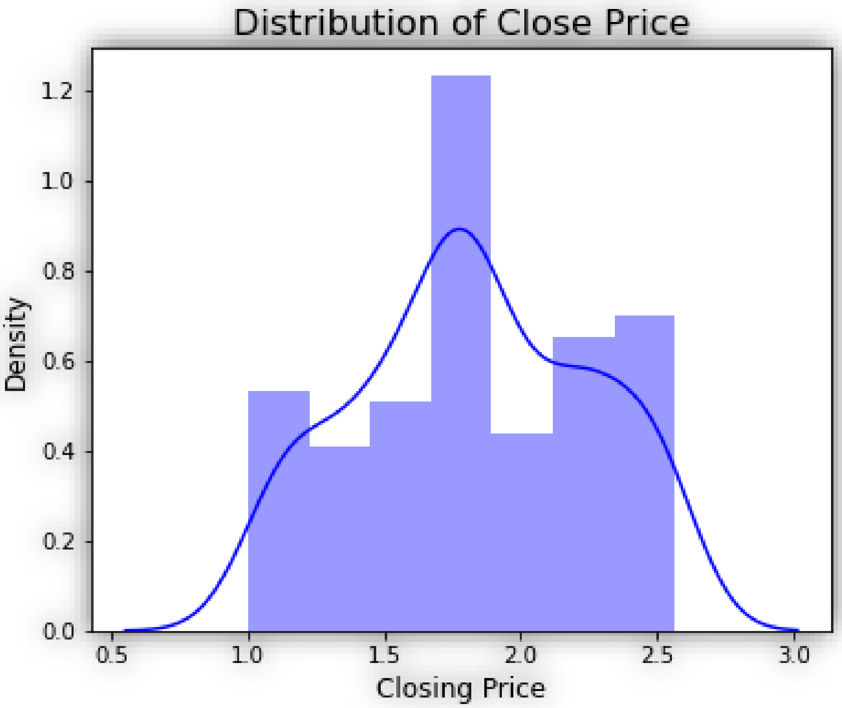
II. **<u>log for greater skew:</u>** log10(x) for positively skewed data, log10(max(x+1) - x) for negatively skewed data

III. **<u>inverse for severe skew:</u>** 1/x for positively skewed data 1/(max(x+1) - x) for negatively skewed data

IV. **<u>Linearity and heteroscedasticity:</u>** First try log transformation in a situation where the dependent variable starts to increase more rapidly with increasing independent variable values If your data does the opposite – dependent variable values decrease more rapidly with increasing independent variable values – you can first consider a square transformation.

EDA (Visualizations)

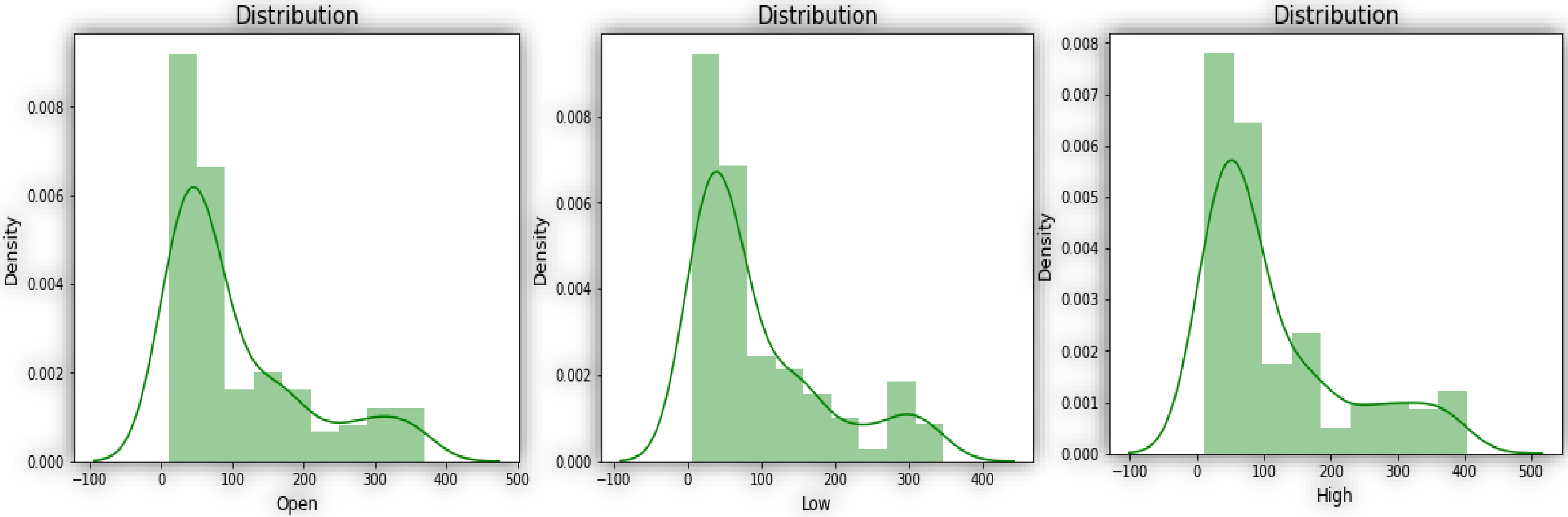
Distribution of dependent variable Close Price of stock(After Transformation).



- After using log transformation as it was positively skewed.
- Now it seems more normal

EDA (Visualizations)

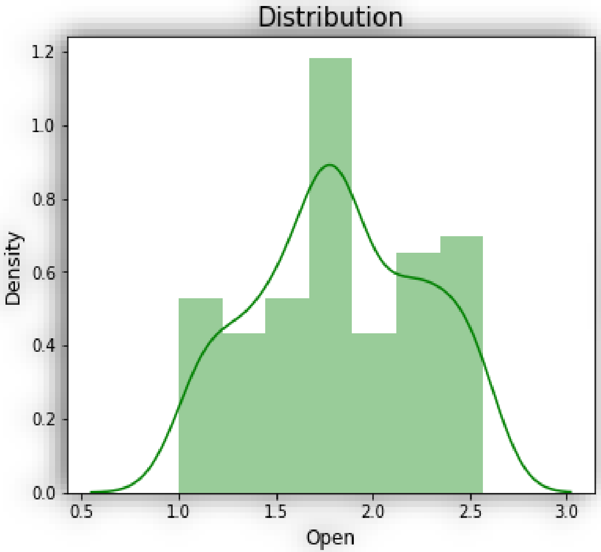
Distribution of numerical features High, Low and Open price of a stock

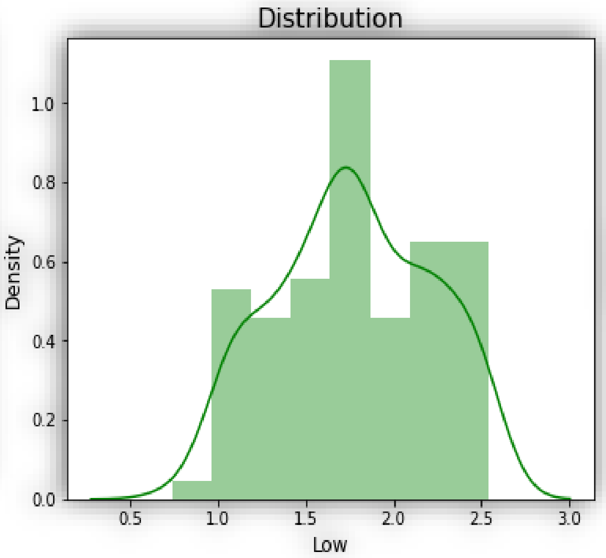


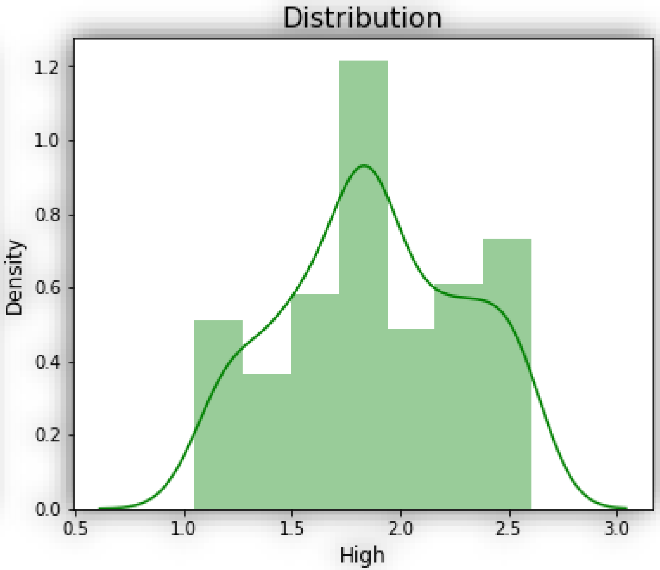
- It looks all numerical features are rightly skewed.
- Apply log transformation to make normal.

EDA (Visualizations)

Distribution of numerical features High, Low and Open price of a stock(After Transformation)



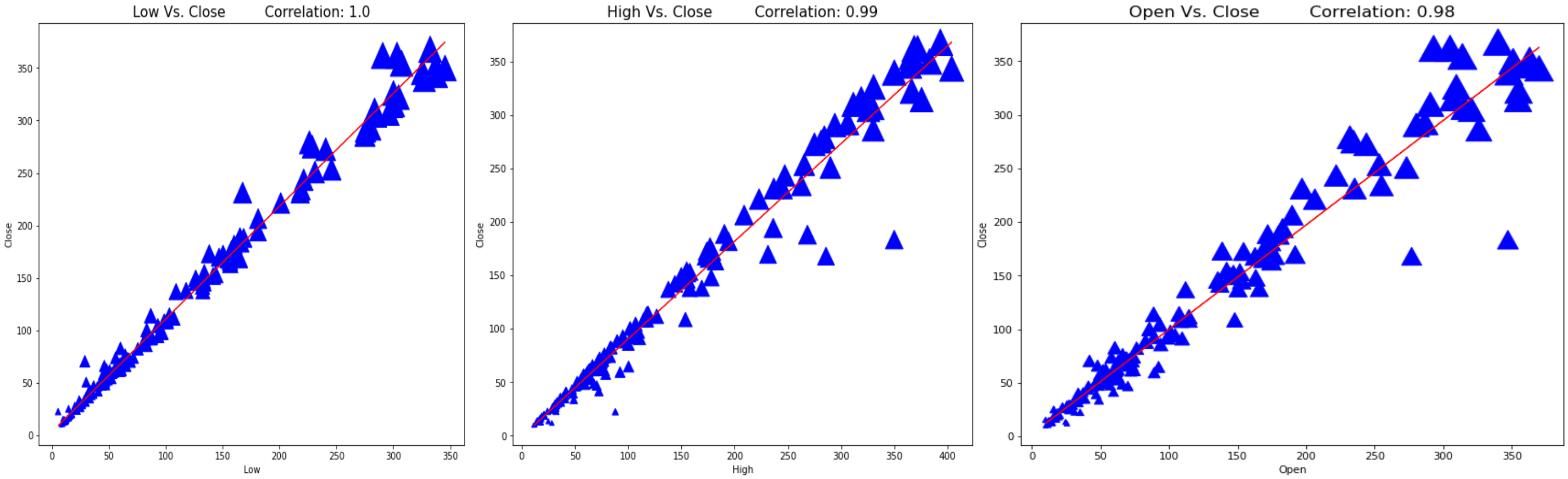




- After using log transformation as it was positively skewed.
- Now it seems more normal

EDA (Visualizations)

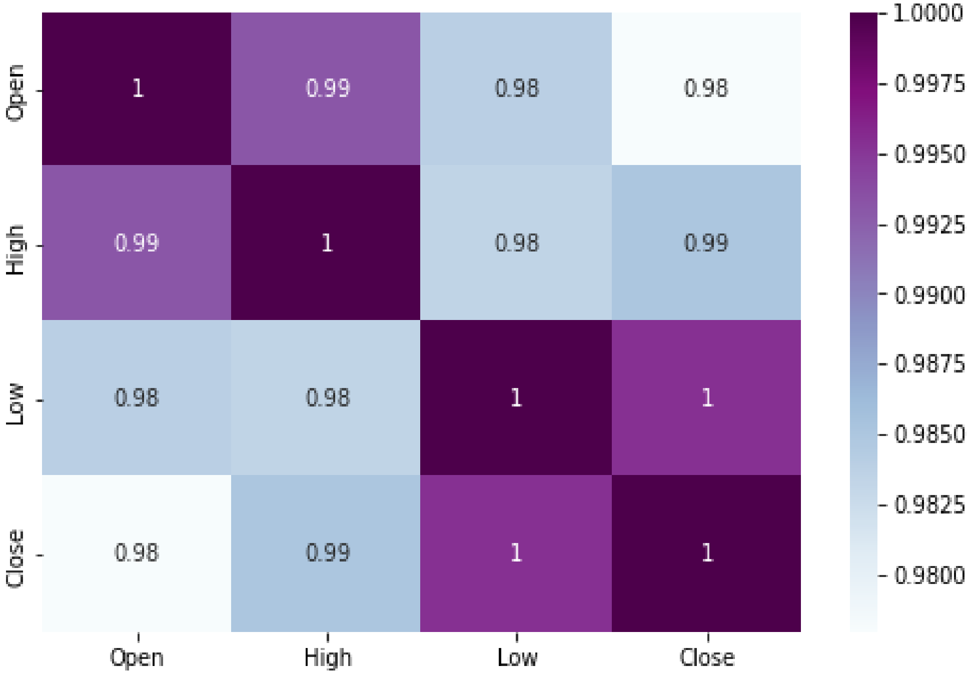
Scatter plot with best fit line



- Plotting a scatterplot with data points can help you to determine whether there’s a potential relationship between them.

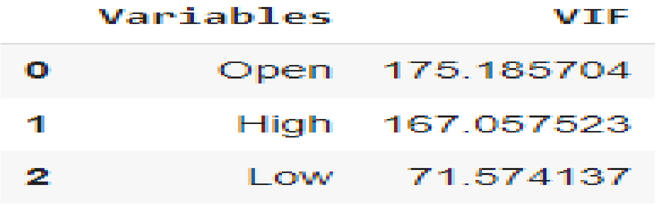
EDA (Visualizations)

Heat Map to see the correlation



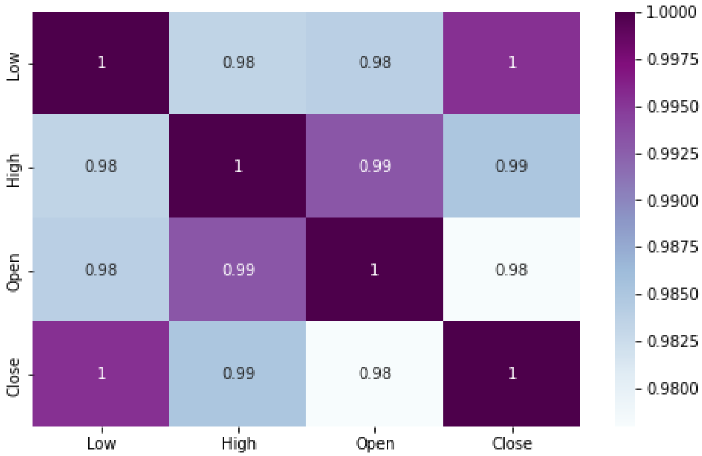
- ✓ There are very high correlations between independent variables which lead us to multicollinearity.
- ✓ High multicollinearity is not good for fitting models and prediction because a slight change in any independent variable will give very unpredictable results.

VIF(Variation Inflation Factor)



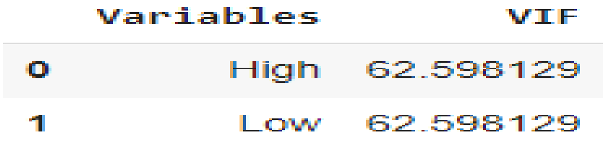
	Variables	VIF
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

- ✓ Variance inflation factor (VIF) is **a measure of the amount of multicollinearity in a set of multiple regression variables**

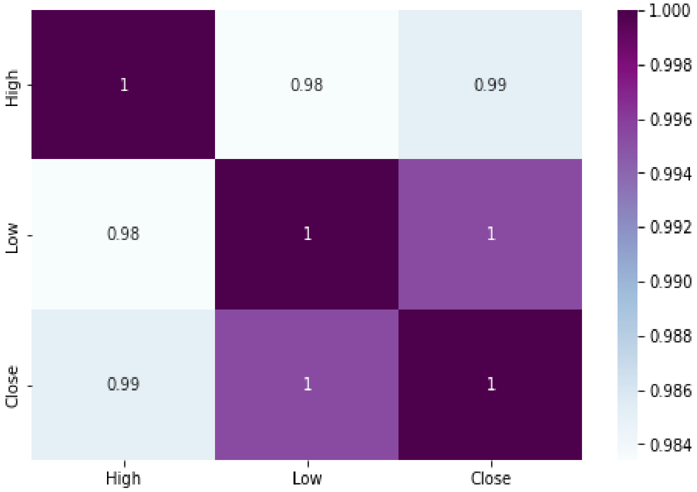


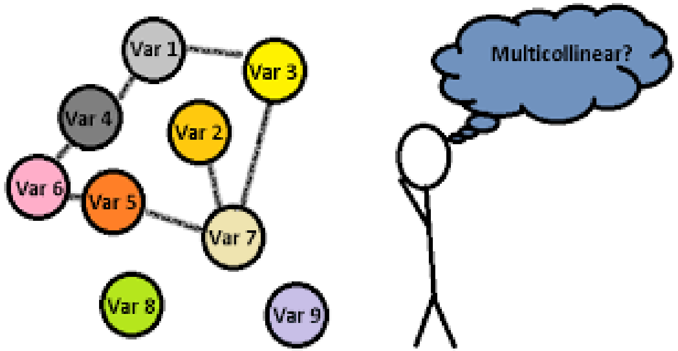
- ✓ We have very high VIF in our dataset so, we have to drop one of them which is least correlated with the dependent variable.
- ✓ I have decided to drop the open price as it is least correlated with the closing price.

VIF(Variation Inflation Factor)



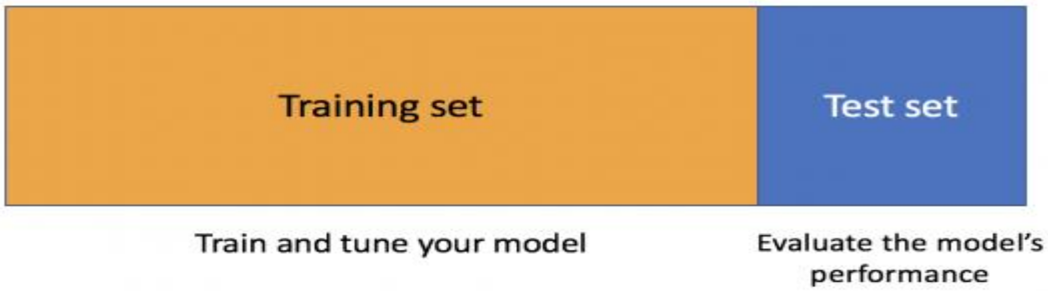
	Variables	VIF
0	High	62.598129
1	Low	62.598129





- ✓ Correlation after removing the Open column. We can see that VIF values are decreased.
- ✓ Further, we cannot remove any other feature as we are having limited features in our dataset.

Data Modelling



Splitting data

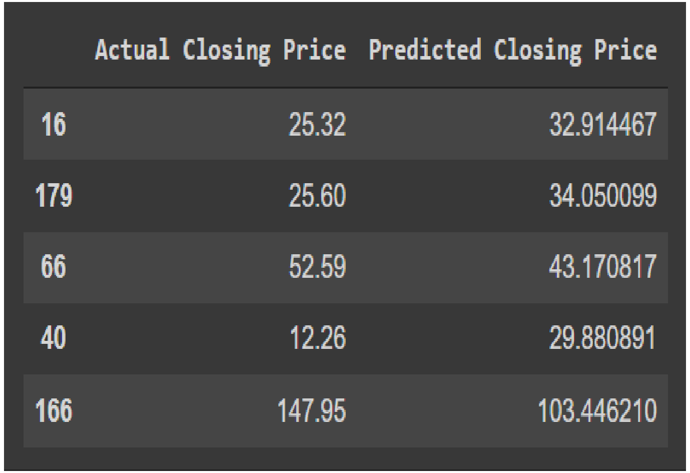
X = Independent variable(High, Low)

Y = Dependent variable(Close)

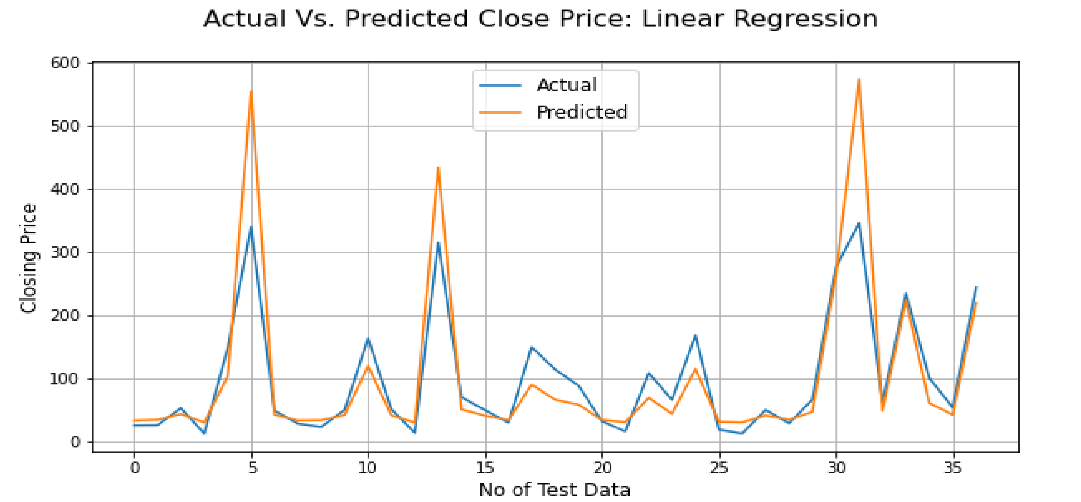
Splitting train-test data with 80-20

Data must be normally distributed before applying normalization. Normalization is one of the feature scaling techniques. We particularly apply normalization when the data is skewed on either axis i.e. when the data does not follow the Gaussian distribution. In normalization, we convert the data features of different scales to a common scale which further makes it easy for the data to be processed for modeling. Thus, all the data features(variables) tend to have a similar impact on the modeling portion. We have used zscore and log transformation.

Linear Regression

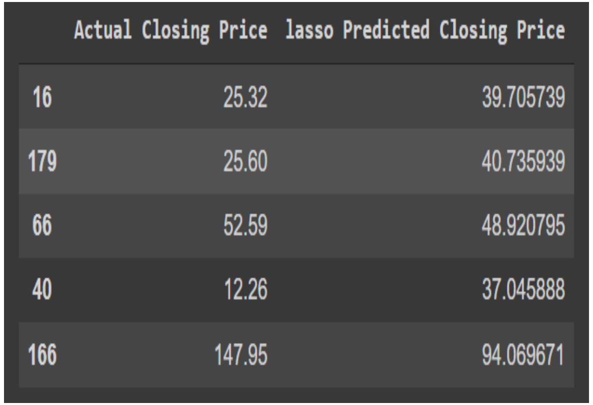


	Actual Closing Price	Predicted Closing Price
16	25.32	32.914467
179	25.60	34.050099
66	52.59	43.170817
40	12.26	29.880891
166	147.95	103.446210

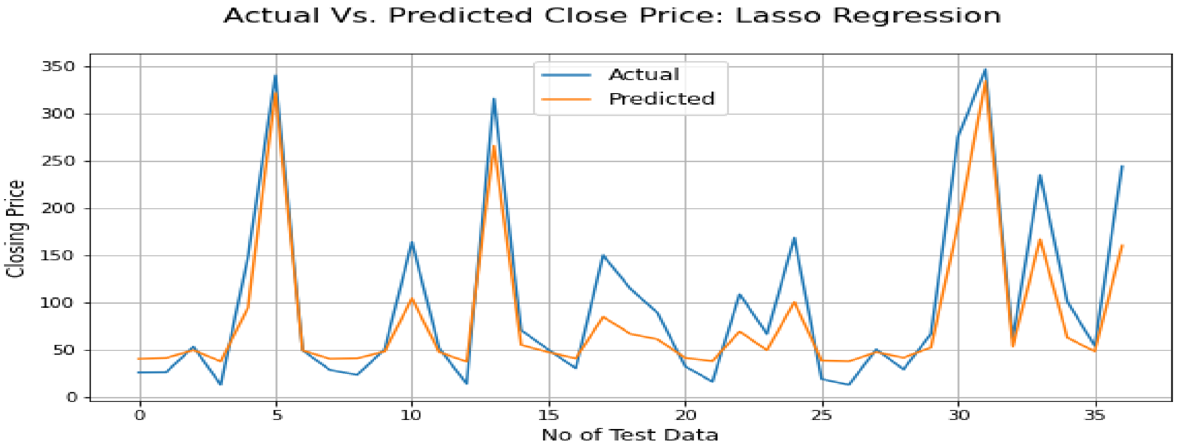


- ✓ Linear regression is one of the easy and popular Machine Learning algorithms.
- ✓ It is a statistical method that is used for predictive analysis.
- ✓ Linear regression makes predictions for continuous or numeric variables such as **sales, salary, age, product price,** etc.
- ✓ The linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression.
- ✓ Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Lasso Regression



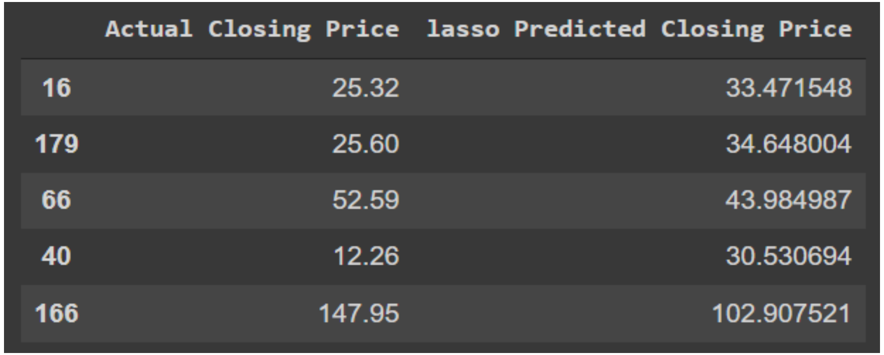
	Actual Closing Price	lasso Predicted Closing Price
16	25.32	39.705739
179	25.60	40.735939
66	52.59	48.920795
40	12.26	37.045888
166	147.95	94.069671



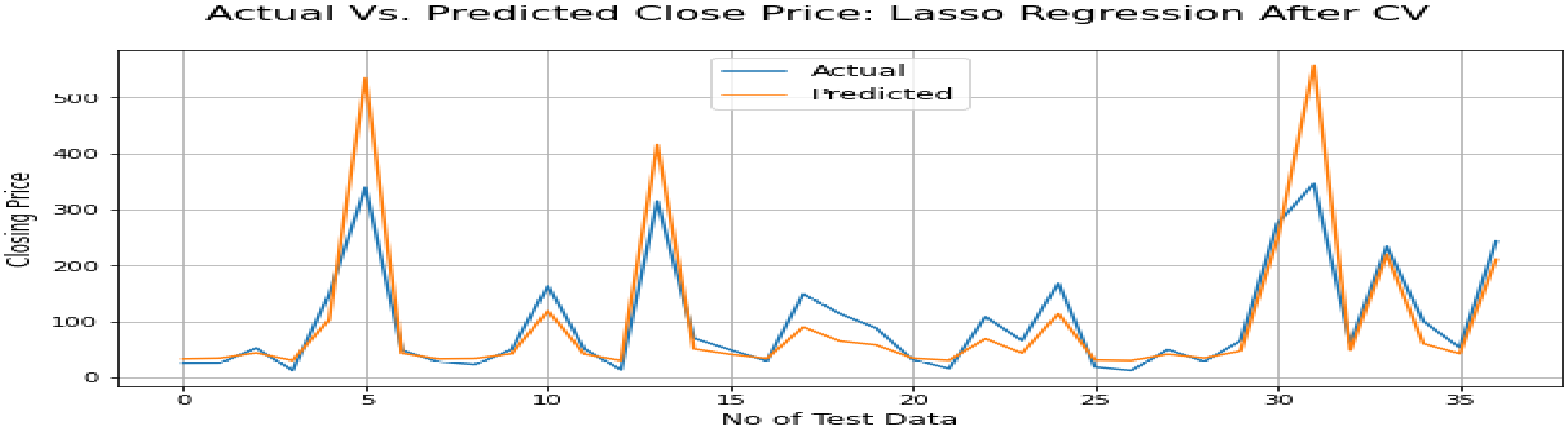
Actual Vs. Predicted Close Price: Lasso Regression

- ✓ Lasso regression is linear regression, but it uses a technique called **"shrinkage"** where the coefficients of determination shrink to towards **zero**.
- ✓ Linear regression gives you regression coefficients as observed in the dataset.
- ✓ The lasso regression allows to shrink or regularize coefficients to avoid overfitting and make them work better on different datasets.
- ✓ This type of regression is used when the dataset shows high multicollinearity or when you want to automate variable elimination and **feature selection**.

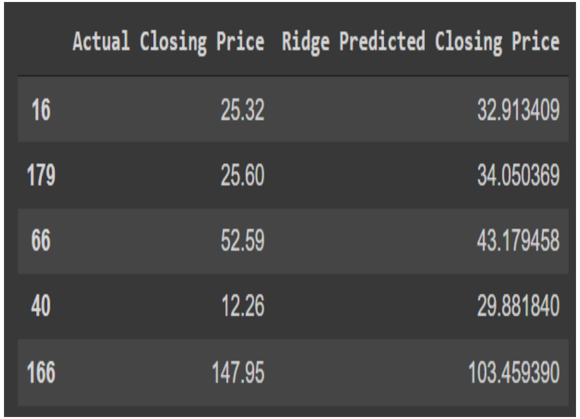
Cross Validation of Lasso Regression



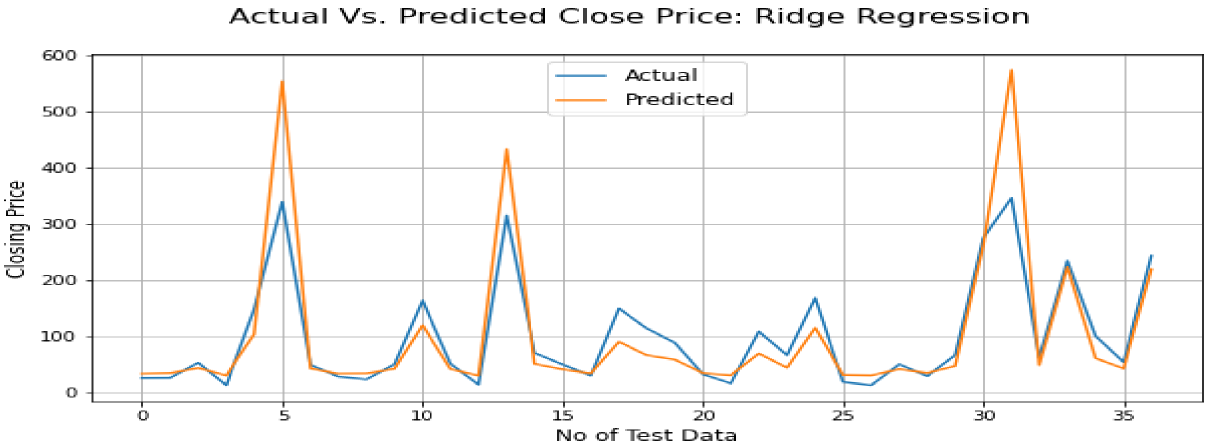
	Actual Closing Price	lasso Predicted Closing Price
16	25.32	33.471548
179	25.60	34.648004
66	52.59	43.984987
40	12.26	30.530694
166	147.95	102.907521



Ridge Regression



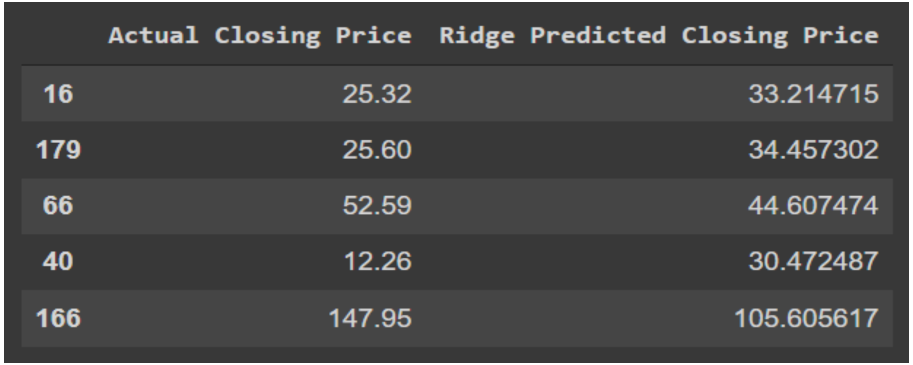
	Actual Closing Price	Ridge Predicted Closing Price
16	25.32	32.913409
179	25.60	34.050369
66	52.59	43.179458
40	12.26	29.881840
166	147.95	103.459390



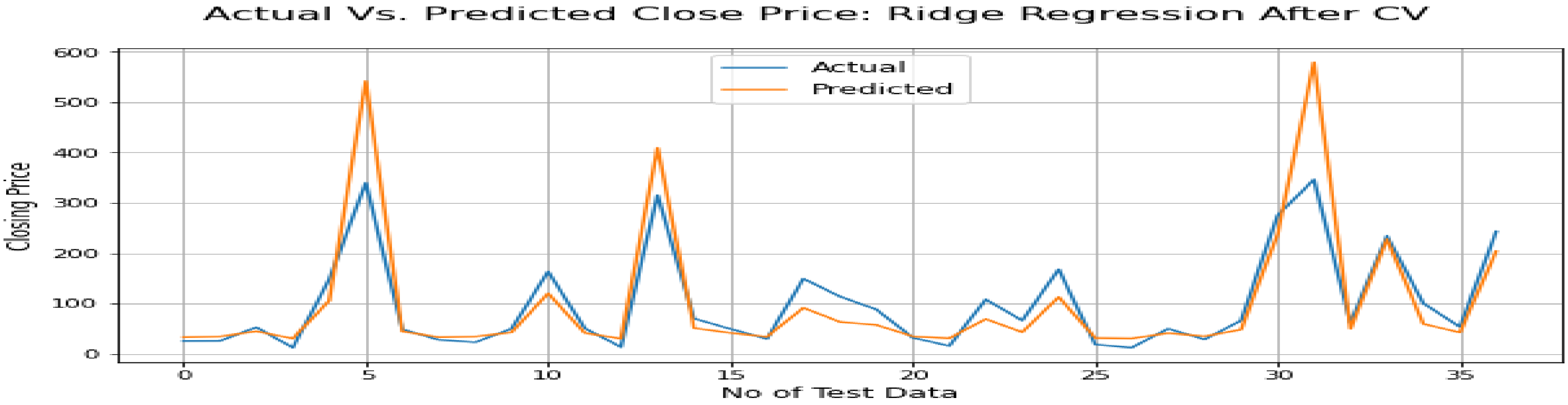
Actual Vs. Predicted Close Price: Ridge Regression

- ✓ Ridge regression is a regularized version of linear least squares regression.
- ✓ It works by shrinking the coefficients or weights of the regression model toward zero.
- ✓ This is achieved by imposing a squared penalty on their size.
- ✓ This is one of the methods of regularization techniques in which the data suffer from multicollinearity.
- ✓ In this multicollinearity, the least squares are unbiased and the variance is large and which deviates the predicted value from the actual value. Equations have an error term.

Cross Validation of Ridge Regression

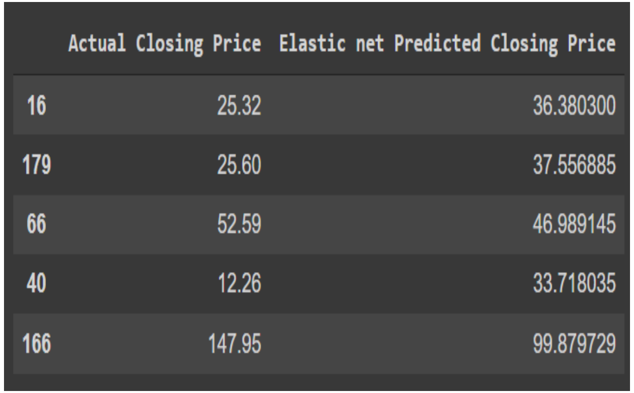


	Actual Closing Price	Ridge Predicted Closing Price
16	25.32	33.214715
179	25.60	34.457302
66	52.59	44.607474
40	12.26	30.472487
166	147.95	105.605617

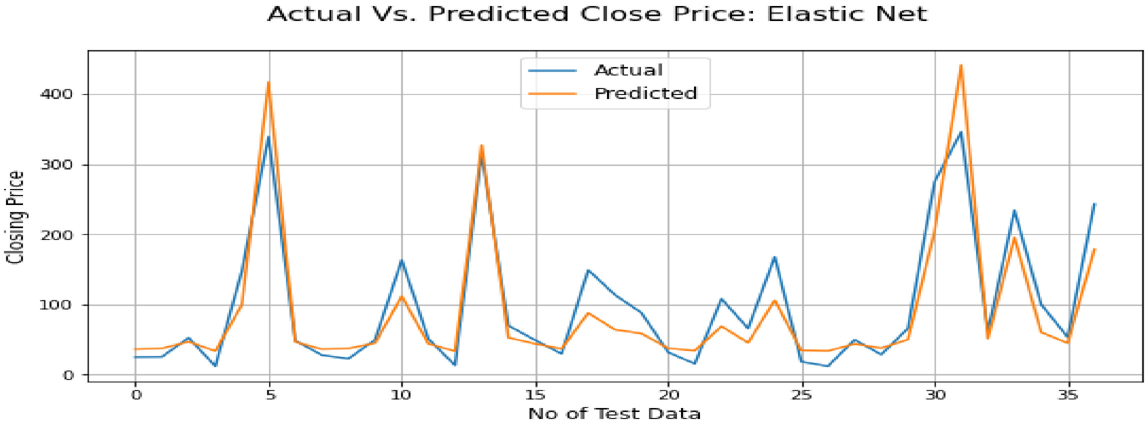


Actual Vs. Predicted Close Price: Ridge Regression After CV

Elastic Net Regression

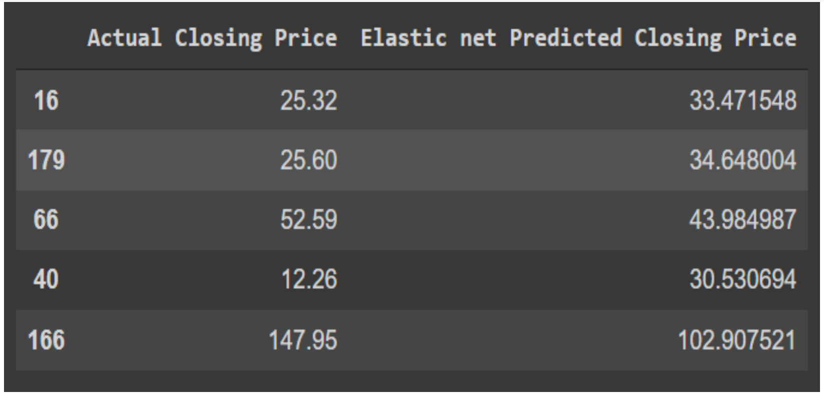


	Actual Closing Price	Elastic net Predicted Closing Price
16	25.32	36.380300
179	25.60	37.556885
66	52.59	46.989145
40	12.26	33.718035
166	147.95	99.879729

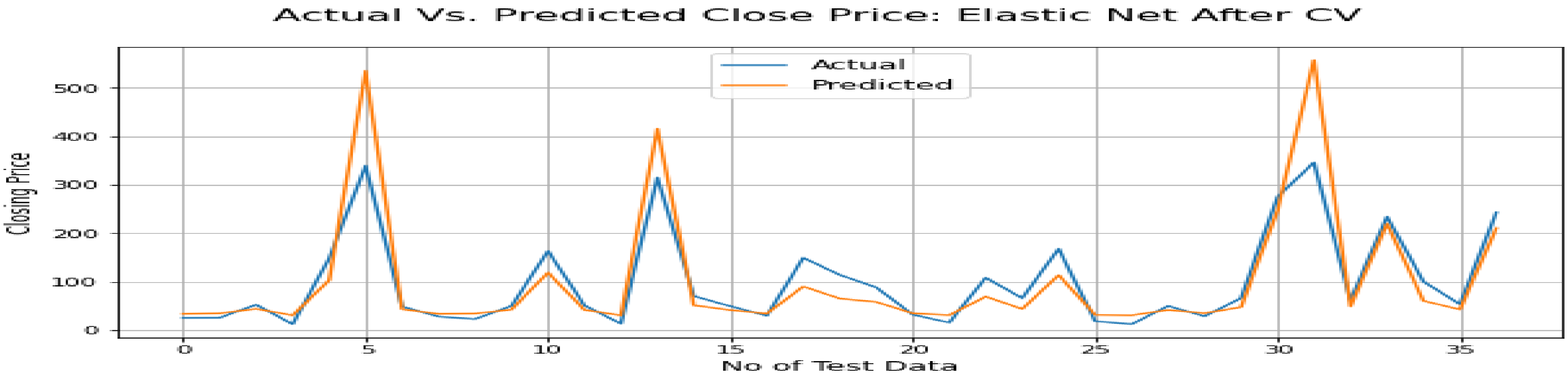


- ✓ Elastic Net Regression is the third type of Regularization technique.
- ✓ It came into existence due to the limitation of the Lasso regression.
- ✓ Lasso regression cannot take correct alpha and lambda values as per the requirement of the data.
- ✓ The solution for the problem is to combine the penalties of both ridge regression and lasso regression.

Cross Validation of Elastic Net



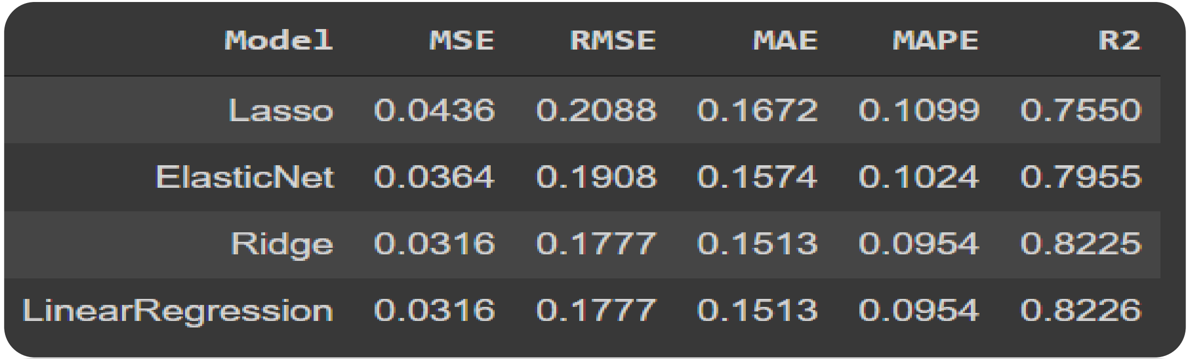
	Actual Closing Price	Elastic net Predicted Closing Price
16	25.32	33.471548
179	25.60	34.648004
66	52.59	43.984987
40	12.26	30.530694
166	147.95	102.907521



Challenges

- While deciding on the independent variables we faced difficulty as there were very limited parameters and they were having very high collinearity with the dependent variable.
- The disadvantage of Linear regression while predicting stock prices is that it is highly limited in its scope. Many predictors cannot be used, which is required to solve the stock price prediction problem.
- According to our observation we concluded that such problems can be better handled by using time series forecasting.

Conclusion



Model	MSE	RMSE	MAE	MAPE	R2
Lasso	0.0436	0.2088	0.1672	0.1099	0.7550
ElasticNet	0.0364	0.1908	0.1574	0.1024	0.7955
Ridge	0.0316	0.1777	0.1513	0.0954	0.8225
LinearRegression	0.0316	0.1777	0.1513	0.0954	0.8226

- Target variable(dependent variable) strongly dependent on independent variables
- We get maximum accuracy of 82%
- Linear regression and Ridge regression get almost the same R squared value
- Whereas the Lasso model shows the lowest R squared value and high MSE, RMSE, MAE, MAPE
- The target variable is highly dependent on input variables.
- Ridge regression shrunk the parameters to reduce complexity and multicollinearity but ended up affecting the evaluation metrics.
- Lasso regression did feature selection and ended up giving up worse results than ridge which again reflects the fact that each feature is important (as previously discussed).

References

I. Stack overflow

II. GeeksforGeeks

III. Jovian

IV. Research paper based on Stock price prediction using ANN

V. Analytics Vidhya

VI. Towards data science

THANK-YOU