Capstone project 2nd

Yes bank stock closing price prediction

By Aashish kumar from cohort Jerusalem

OUTLINE

- Overview and objectives
- Data Outline
- Exploratory data analysis
- Model Implementation
- Model Comparison
- Conclusion

Overview & Objective

Overview

- 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 any predictive models can do justice to such situations.

Objective

This dataset has monthly stock prices of YES 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 Outline

We have a dataset which contains monthly stock prices of Yes bank shares since the opening of the bank. It contains multiple features like:-

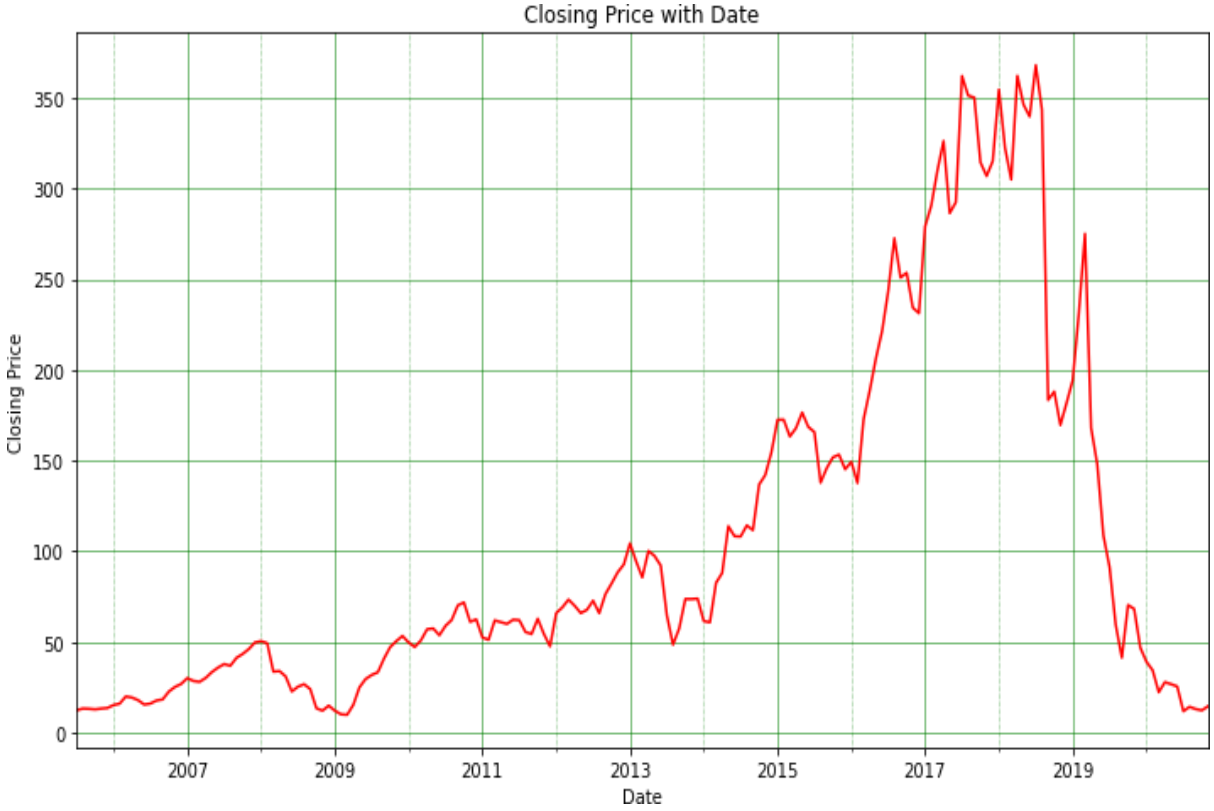
- Date :- denotes the date (so we can see the price at a given date.)
- Open :- denotes the price at which a stock started trading.
- High :- highest price at which a stock traded during a period.
- Low :- the minimum price at which a stock traded during a period.
- Close :- the closing price refers to a stock's trading price closed at the end.

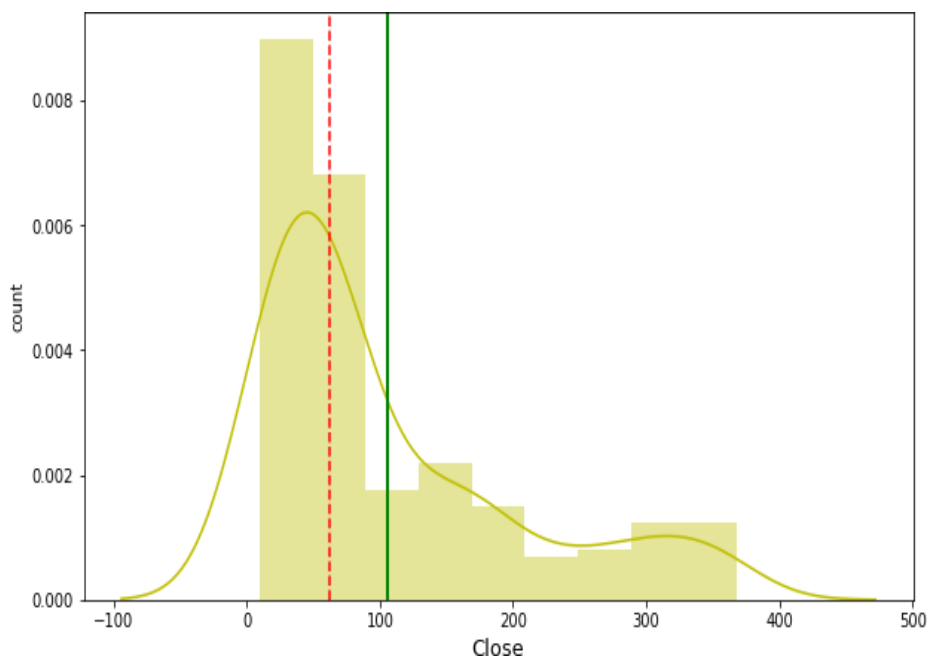
(It's a dependent variable which we need to predict using ML models. The closing price is the price of the stock at the end of the month or the time period in consideration.)

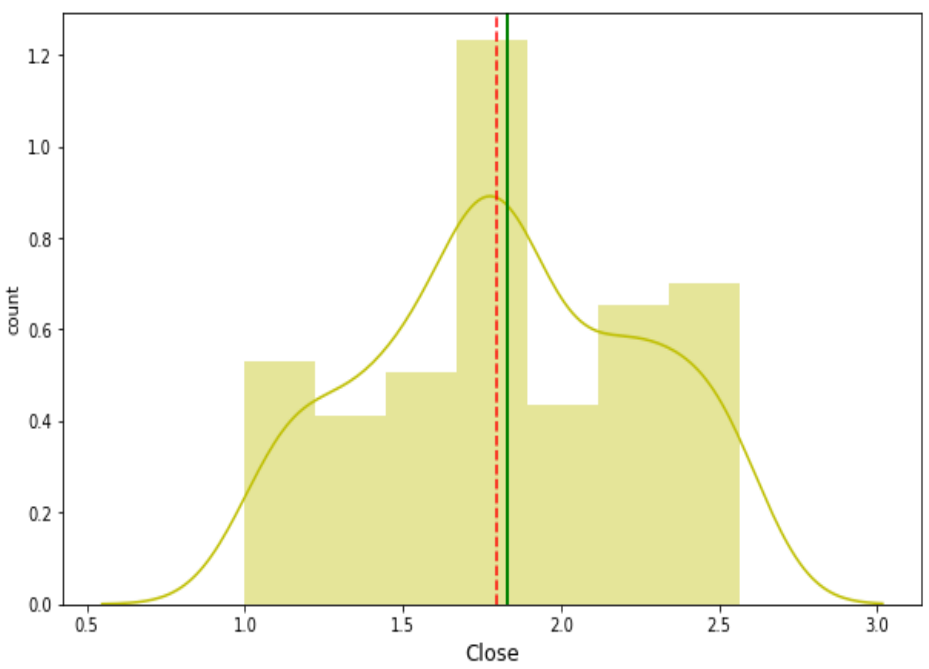
EDA : Visualizing our dependent variable.

- The graph demonstrates how closing price varies with each passing year.

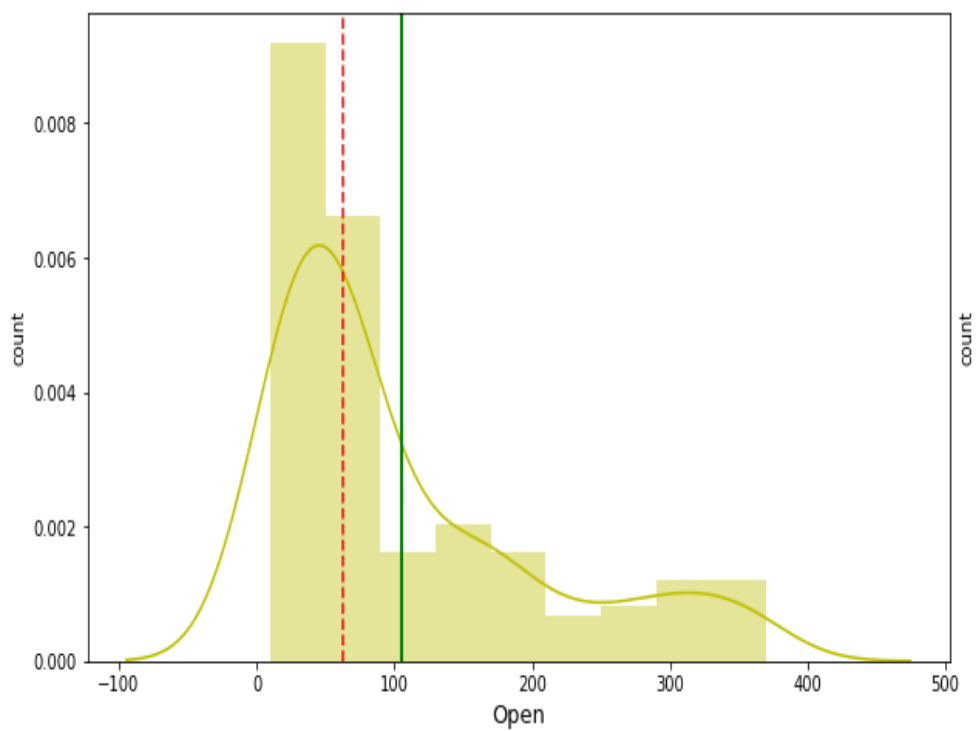
- We can clearly see from the graph that around 2018, when the fraud case involving Rana kapoor came to light, a clear significant dip can be seen in the stock price of Yes Bank data.

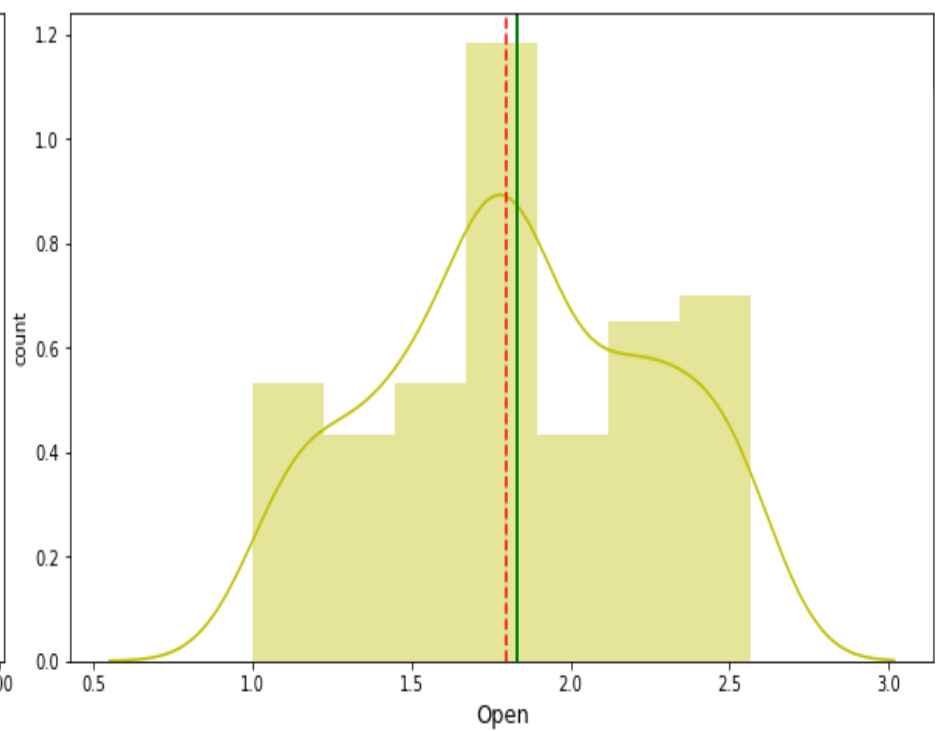




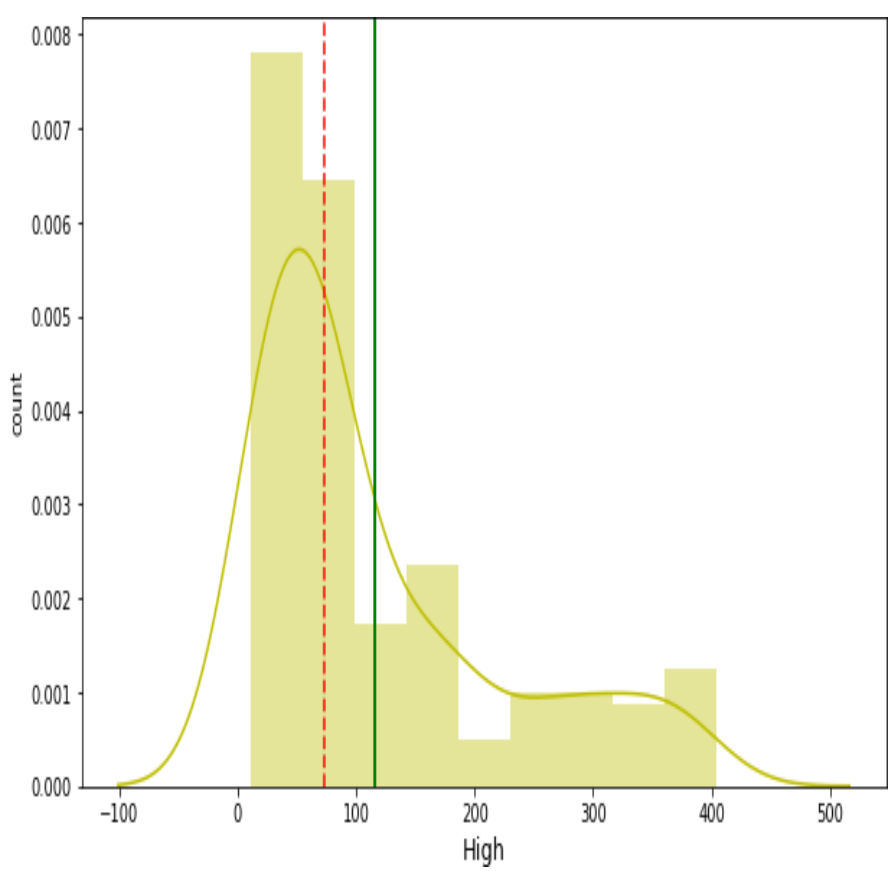


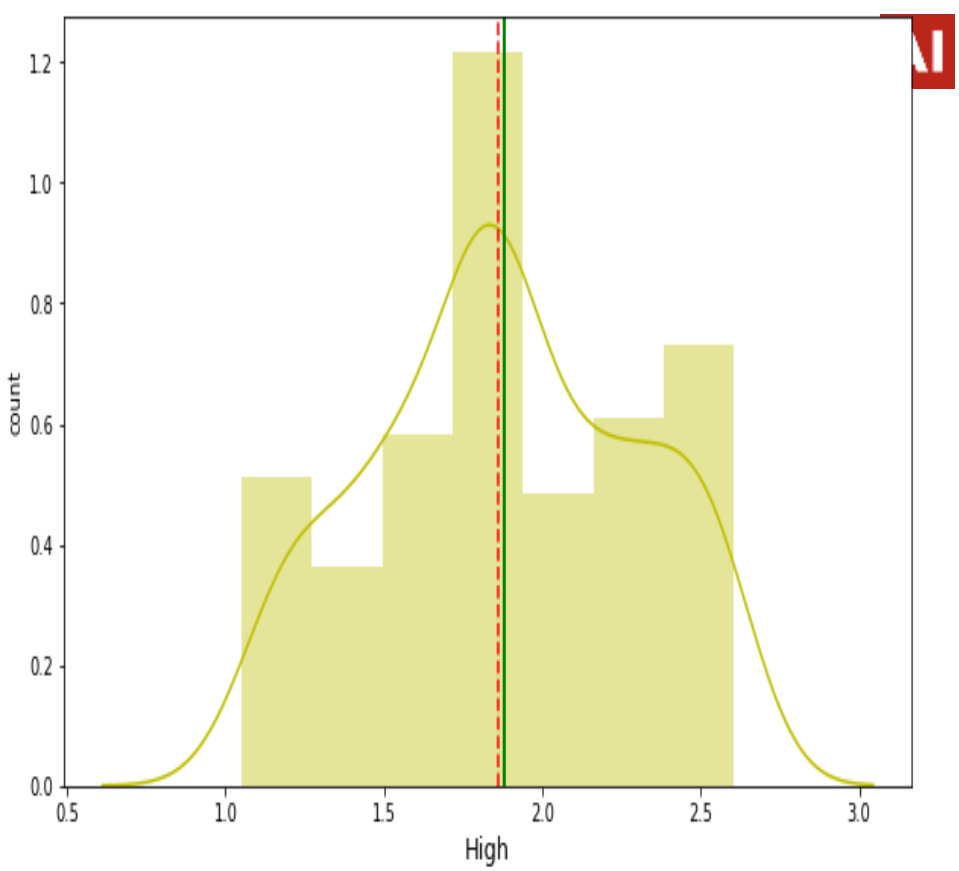
❑ Plotting the dependent variable. We can see that our dependent variable close is positively skewed (as seen on the left).
So we do a log transform on it and plot it as seen in the right chart. This makes it approximate normal distribution and is optimal for our model's performance. Now our mean and median are nearly equal.



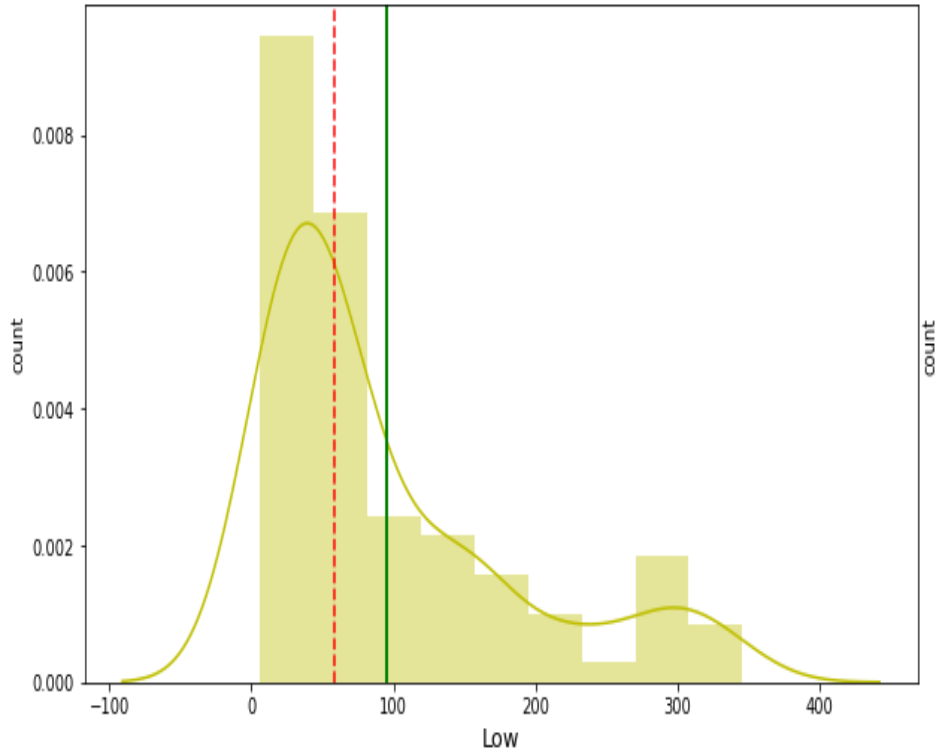


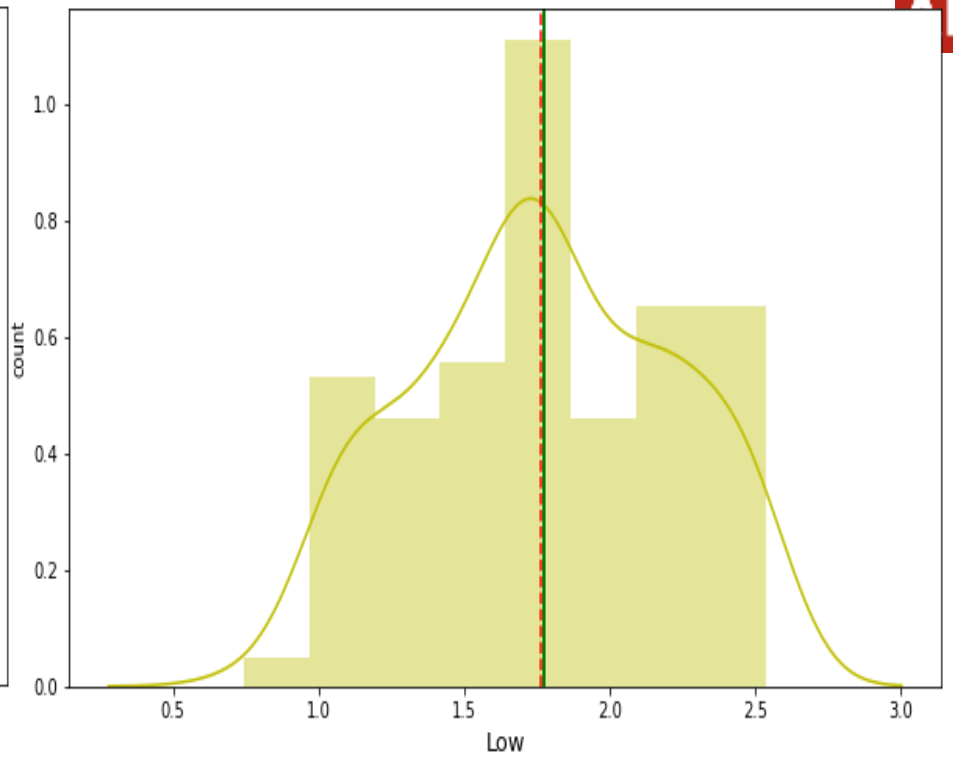
- Plotting the independent variables. As we see in the left chart, data is positively skewed, so we perform a log transform on it. In the right chart, we can see the transformed distribution which is similar to a normal distribution.





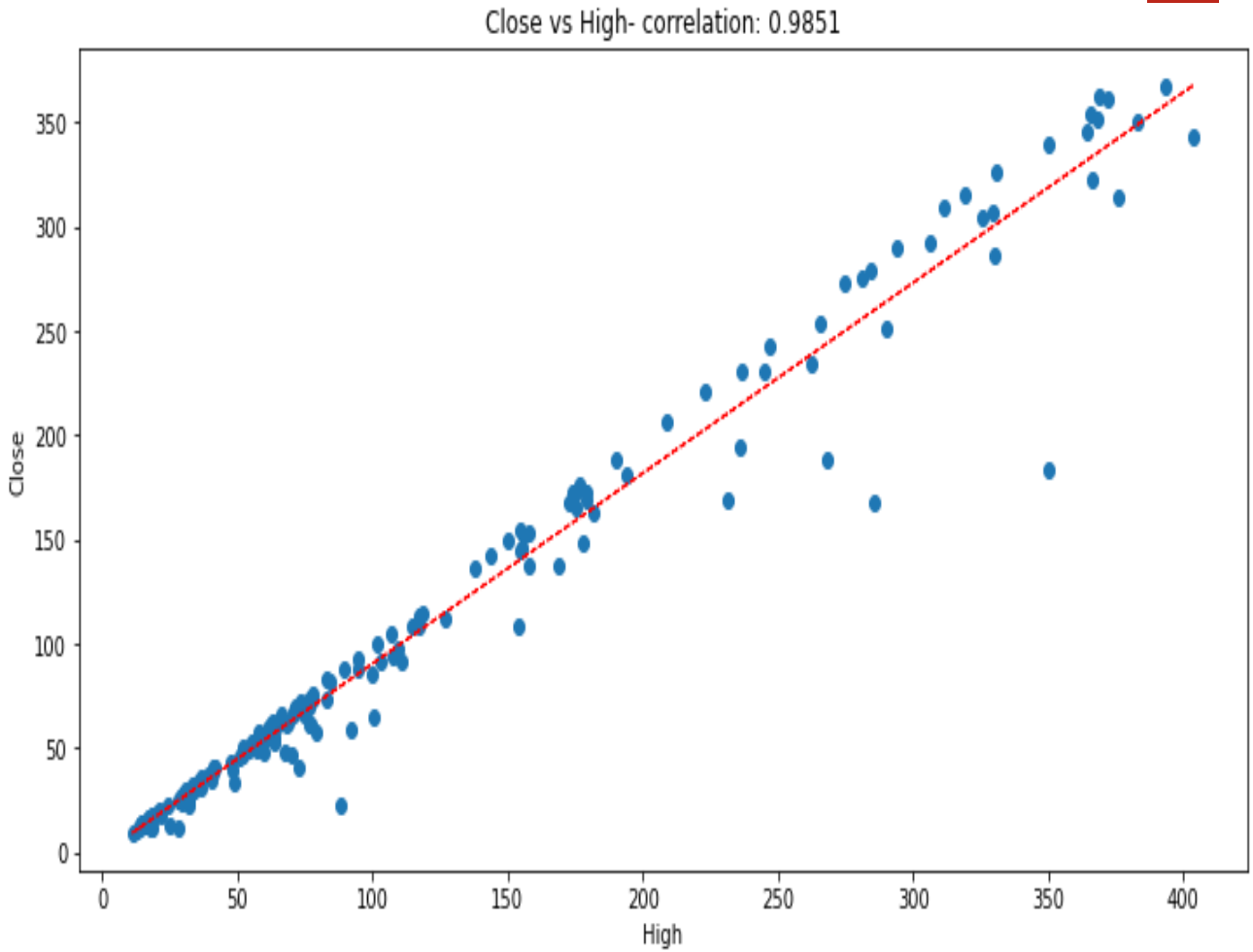
Distribution of dependent variable High before and after applying log transform.

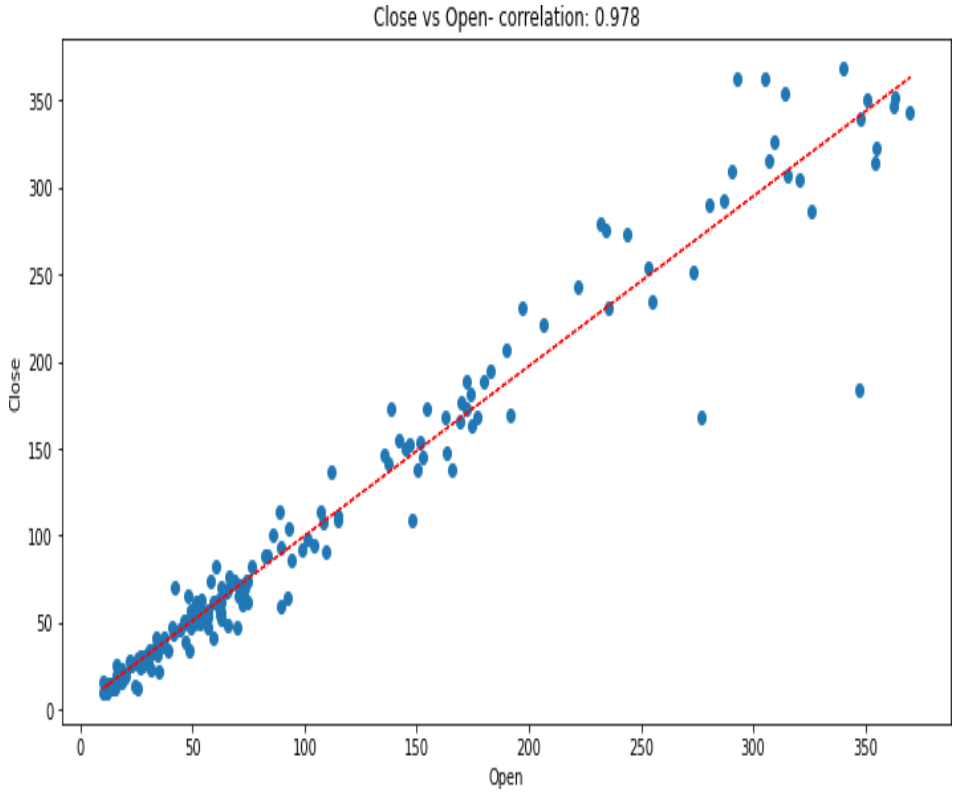


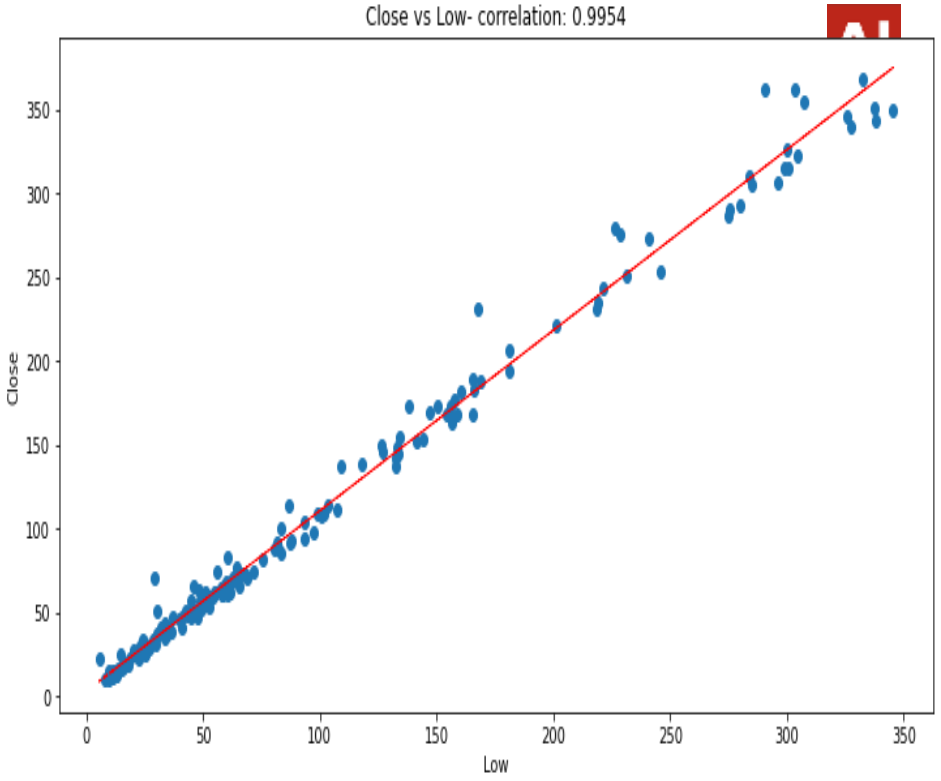


Distribution of dependent variable before and after log transformation.

- As we can see that there is linear relation and high correlation between each independent variables and our dependent variable.
- Also we can see that the value of correlation between dependent variable Close and feature High is 0.985



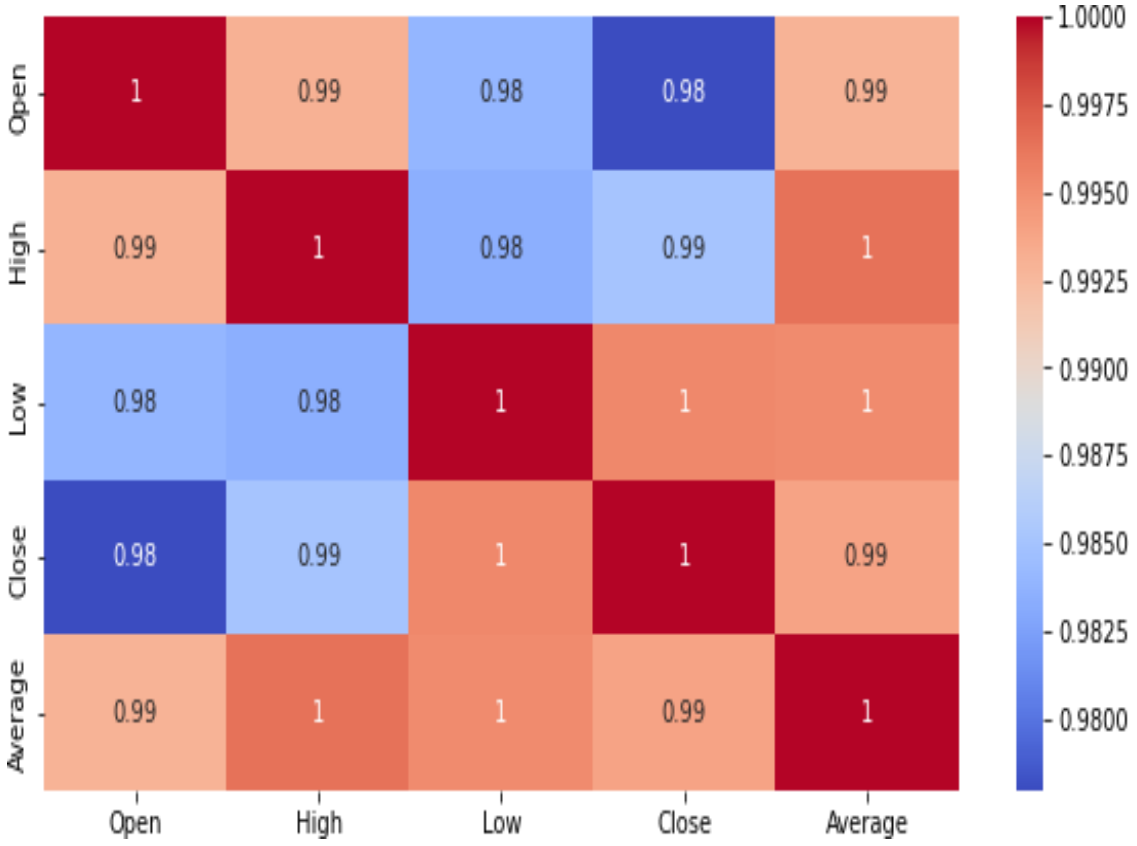




- As we can see that there is a linear relation and very high correlation between our dependent variable and independent variables. The value of correlation between Close and Open is 0.978 and b/w Close and Low is 0.9954.

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



Model Implementation

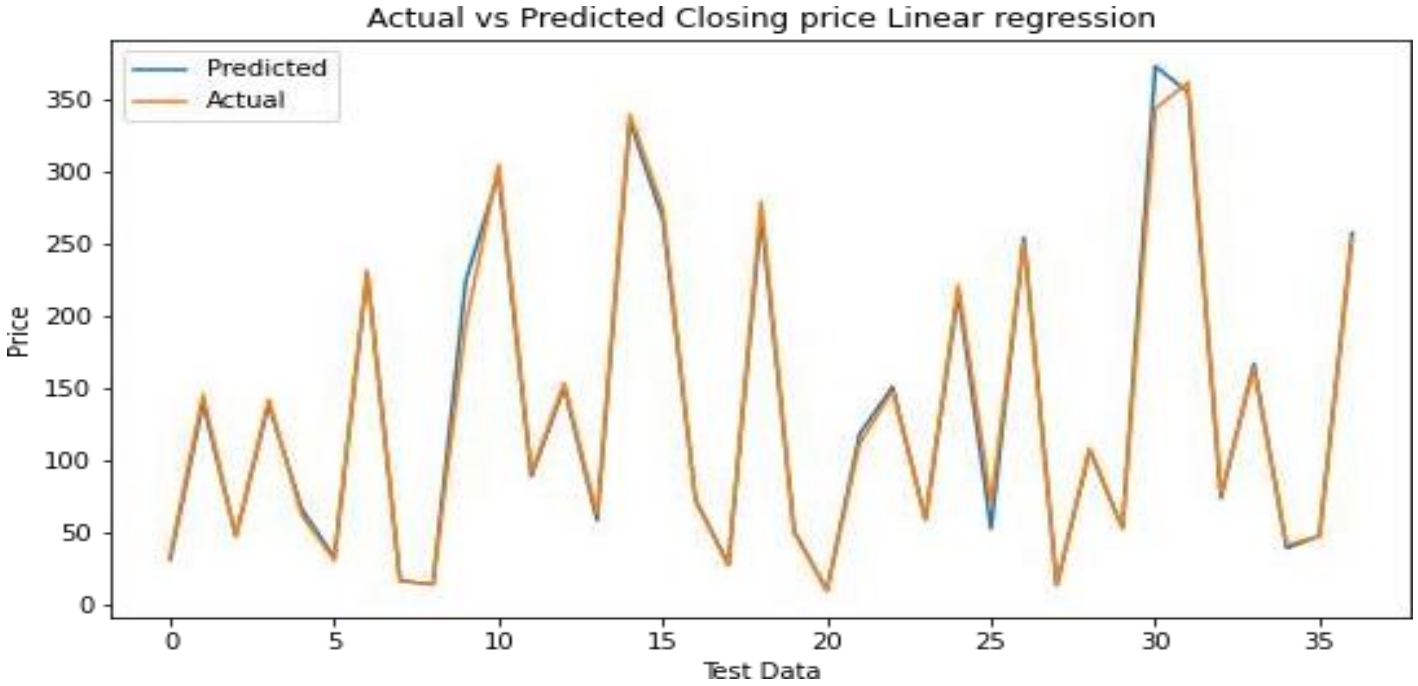
Based on the linear relationship between the dependent and independent variables present in our data, we implemented following models on our data.

- Linear Regression
- Lasso Regression with Cross-validation
- Ridge Regression with Cross-validation
- Elastic Net Regression with Cross-validation

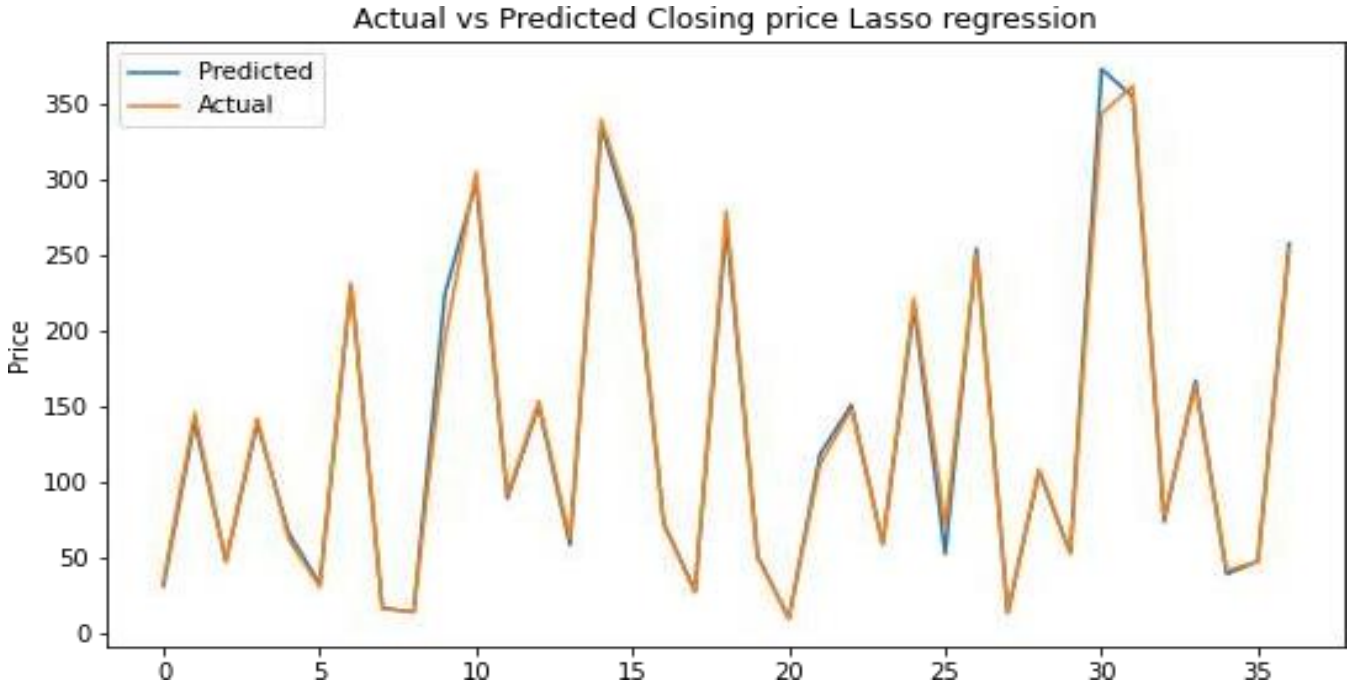
We fit these models on training data, learn the model parameters and then make predictions on test dataset. Then we check the performance of these models using various evaluation metrics such as :-

- Mean Absolute error.
- Mean squared error and RMSE
- R-squared and Adjusted R-squared

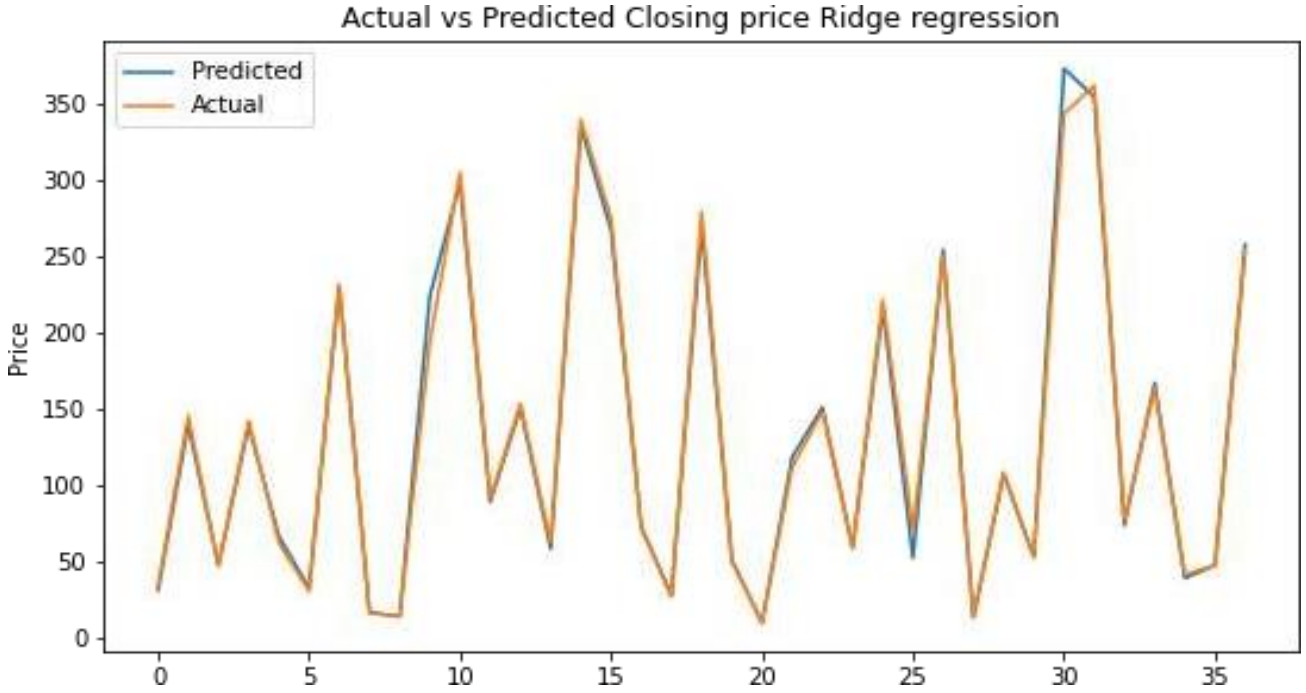
Finally, we select the best performing model based on these metrics.



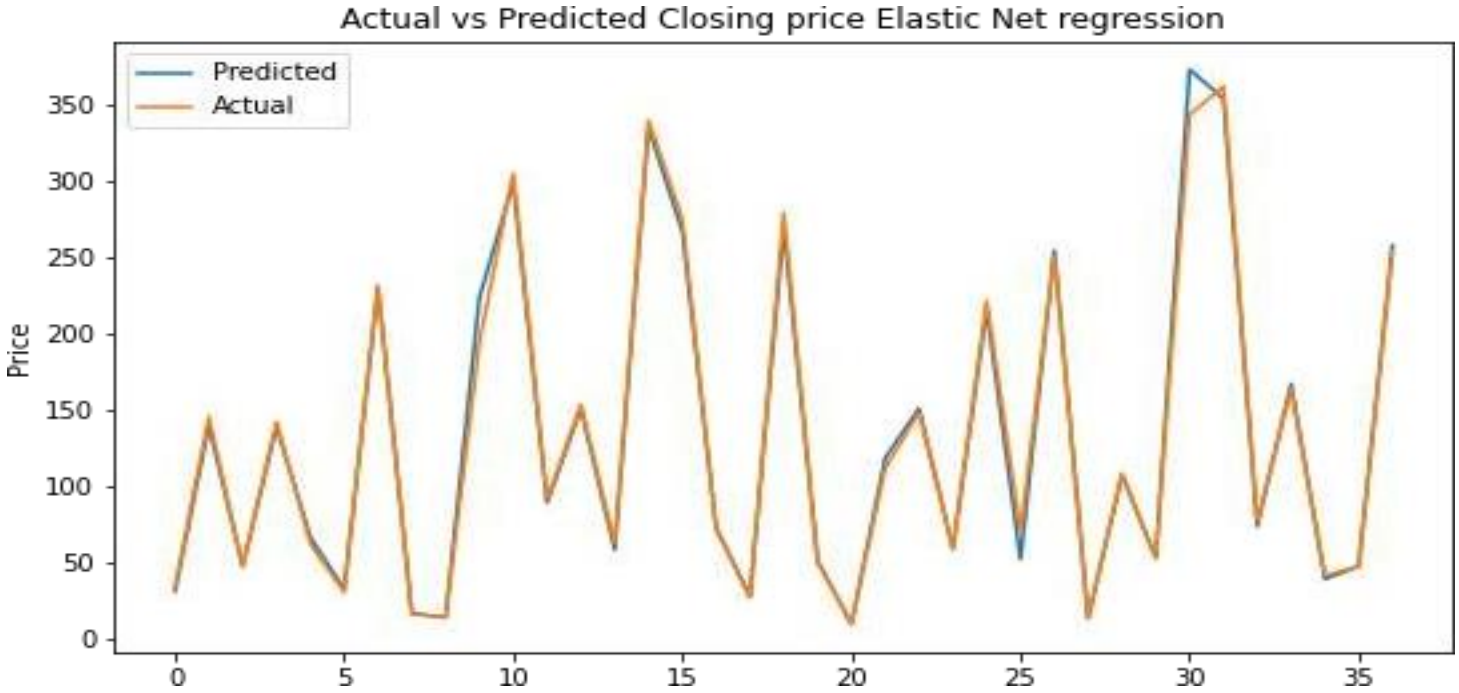
- Our **simple Linear Regression** Model predicted the closing price with Root Mean squared error(RMSE) of 8.3917
- R2 score of this model is 0.9937
- Adjusted R2 score has the value 0.9930 for this model. Which tells us that around 99.3 percent of the variance in our dependent variable is attributable to the independent variables.



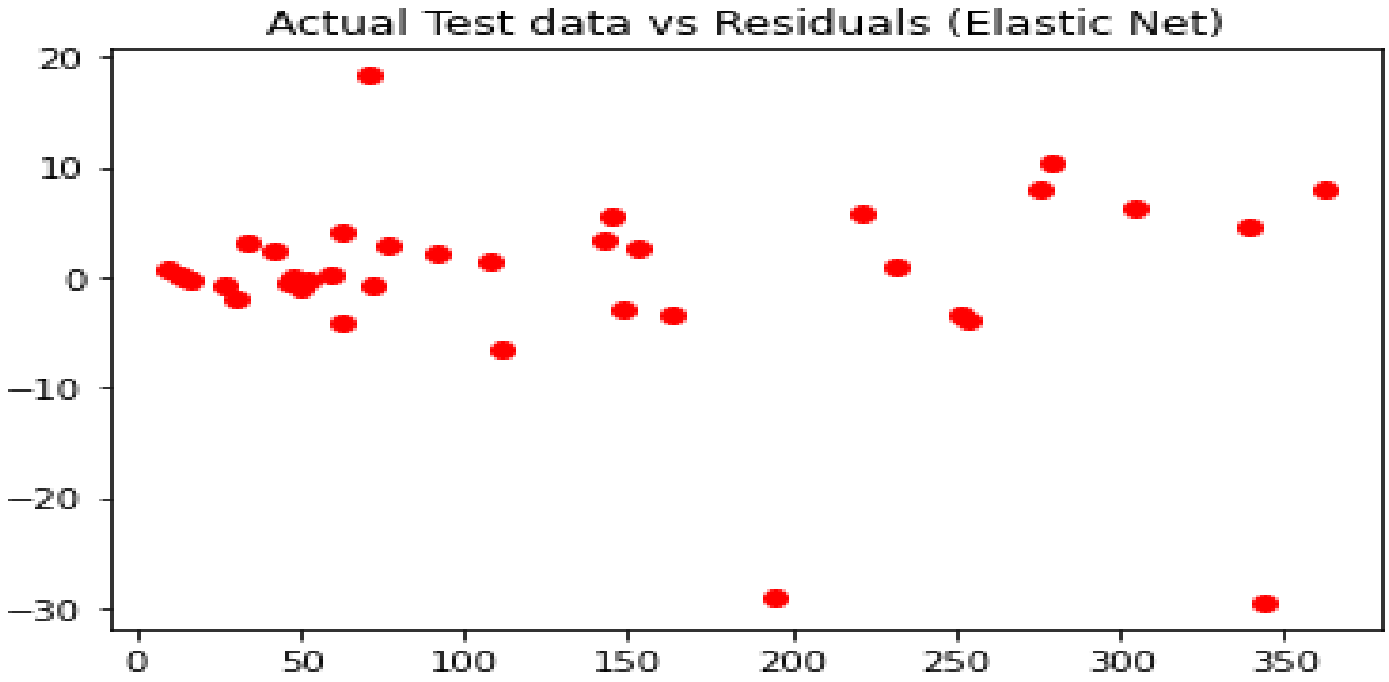
- Our **Lasso Regression** Model predicted the closing price with Root Mean squared error of 8.3864
- R2 score of this model is 0.9938
- Adjusted R2 score has the value 0.9932 for this model. Which tells us that around 99.32 percent of the variance in our dependent variable is attributable to the independent variables.



- Our **Ridge Regression** Model predicted the closing price with Root Mean squared error of 8.3824
- R2 score of this model is 0.9938
- Adjusted R2 score has the value 0.9932 for this model. Which tells us that around 99.32 percent of the variance in our dependent variable is attributable to the independent variables.



- Our **Elastic Net Regression** Model predicted the closing price with Root Mean squared error of 8.3760
- R2 score of this model is 0.9938
- Adjusted R2 score has the value 0.9932 for this model. Which tells us that around 99.32 percent of the variance in our dependent variable is attributable to the independent variables.



- In the above graph, I have plotted the residuals (actual value – predicted) against the predicted values of our best performing model – Elastic Net regression. This is to check whether **Heterodasceticity** is present in our data or not. Since the data is symmetrical around zero, we can safely say that there is no heterodasceticity in our data. Hence the assumption of linear regression is valid here.

Evaluation Metrics:

	Linear Regression	Ridge	Lasso	Elastic-Net
MAE	4.8168	4.8262	4.8334	4.8483
MSE	70.4204	70.3311	70.2641	70.1569
RMSE	8.3917	8.3864	8.3824	8.3760
R-square	0.9937	0.9938	0.9938	0.9938
Adjusted R-square	0.9930	0.9932	0.9932	0.9932

- We can clearly see from the table above that the best performing model is elastic net as it has higher accuracy and least error value.

Conclusions Drawn

- There is a high correlation between the dependent and independent variables. This is a good thing as we can make really accurate predictions using simple linear models.
- We implemented several models on our dataset in order to be able to predict the closing price and found that Elastic Net regressor is the best performing model with Adjusted R2 score value of 0.9932 and it scores well on all evaluation metrics.
- All of the models performed quite well on our data giving us the accuracy of over 99%..
- We found that there is a rather high correlation between our independent variables. This multicollinearity however is unavoidable here as the dataset is very small.
- We found that the distribution of all our variables is positively skewed. so we performed log transformation on them.
- Using data visualization on our target variable, we can clearly see the impact of 2018 fraud case involving Rana Kapoor as the stock prices decline dramatically during that period.
- With our model making predictions with such high accuracy even on unseen test data, we can confidently deploy this model for further predictive tasks using future real data.

THANK YOU