Capstone Project
Yes Bank Stock Price Prediction Technical Documentation

Abhishek Kumar
Abhishekasks151@gmail.com

Table of Content:-

Abstract :-

Yes Bank is a banking company that was founded in 2004 that offers a wide range of differentiated products for its corporate and retail customers through retail banking and asset management services. It is also a publically traded company. That provides an opportunity for anyone to invest in Yes bank and become a shareholder. But at the same time, it means that the valuation of the company is now in the hands of investors and speculators as share prices are often heavily impacted by public opinion.
We have used yes bank stock price data set. This dataset contains 5 different features that can be used for predicting close price prediction using machine learning. We have built machine learning regression model for price prediction. We have used some of best models.

Introduction:-

YES bank stands for Youth Enterprise Scheme Bank. Stock market is one of the major fields that attracts people, thus stock market price prediction is always a hot topic for researchers from both financial and technical domains. In our project our objective is to build a prediction model for close price prediction.

The entire idea of predicting stock prices is to gain significant profits. Predicting how the stock market will perform is a hard task to do. There are numerous other factors involved in the prediction, such as the psychological factor – namely crowd behavior etc. All these factors combine to make share prices very difficult to predict with high accuracy.

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 any 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 Description:-

Before performing any operation on the dataset, it is important to understand the data. After loading the data, we observed the dataset by checking a few of the first and last rows. We checked the shape of the dataset and there are 185 rows and 5 features columns in our dataset.
Let's understand the features present in our dataset.

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

• **Open:** Open means the price at which a stock started trading when the opening bell rang.

• **High:** High refer to the maximum prices in a given time period.

• **Low:** Low refer to the minimum prices in a given time period.

• **Close:** Close refers to the price of an individual stock at the end of the considered time period.

Exploratory Data Analysis:-

A) Data Cleaning: -

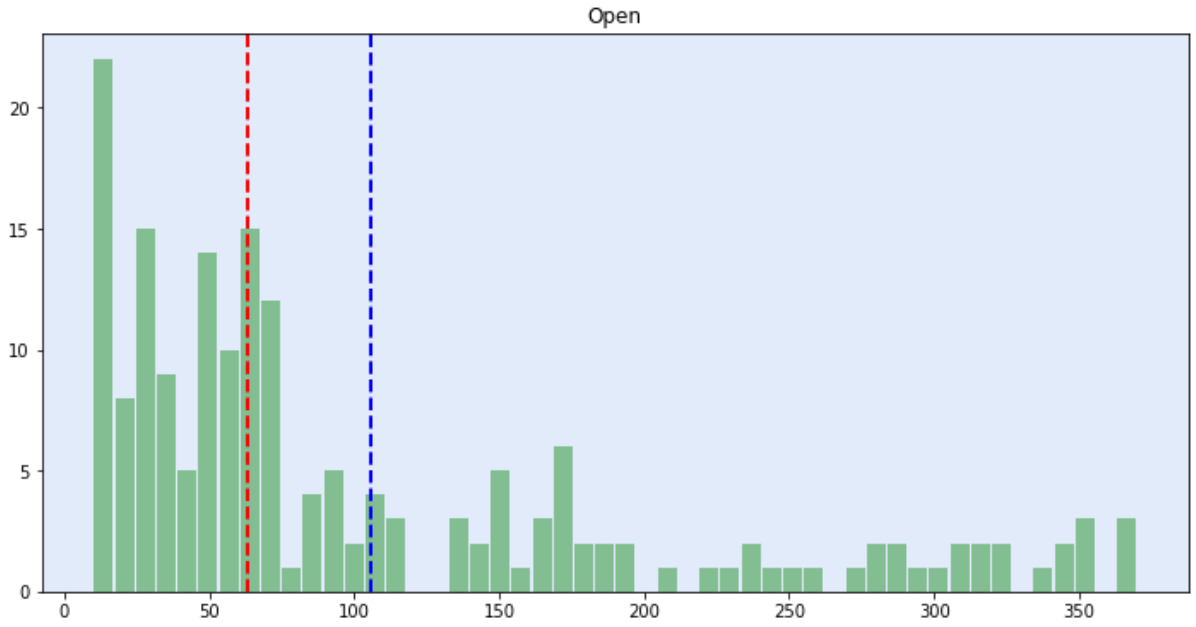
The Given Date in data is of Month-year format (mmm-yy) is converted to proper date of YYYY-MM-DD and given date column has dtype as object converting it into date time format.

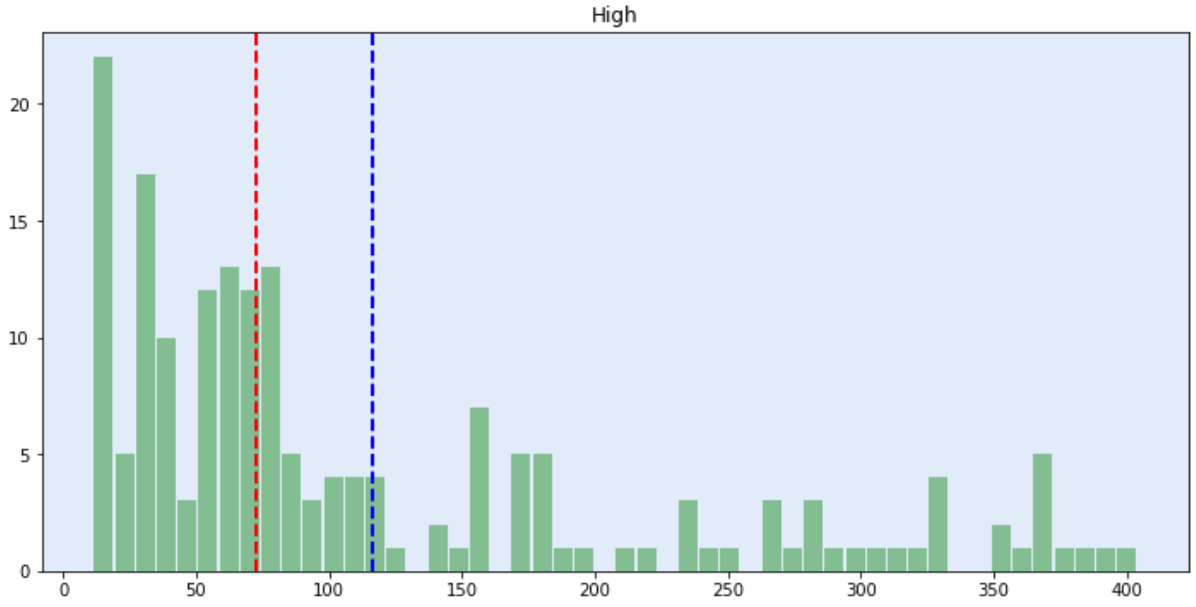
B) Null values Treatment:

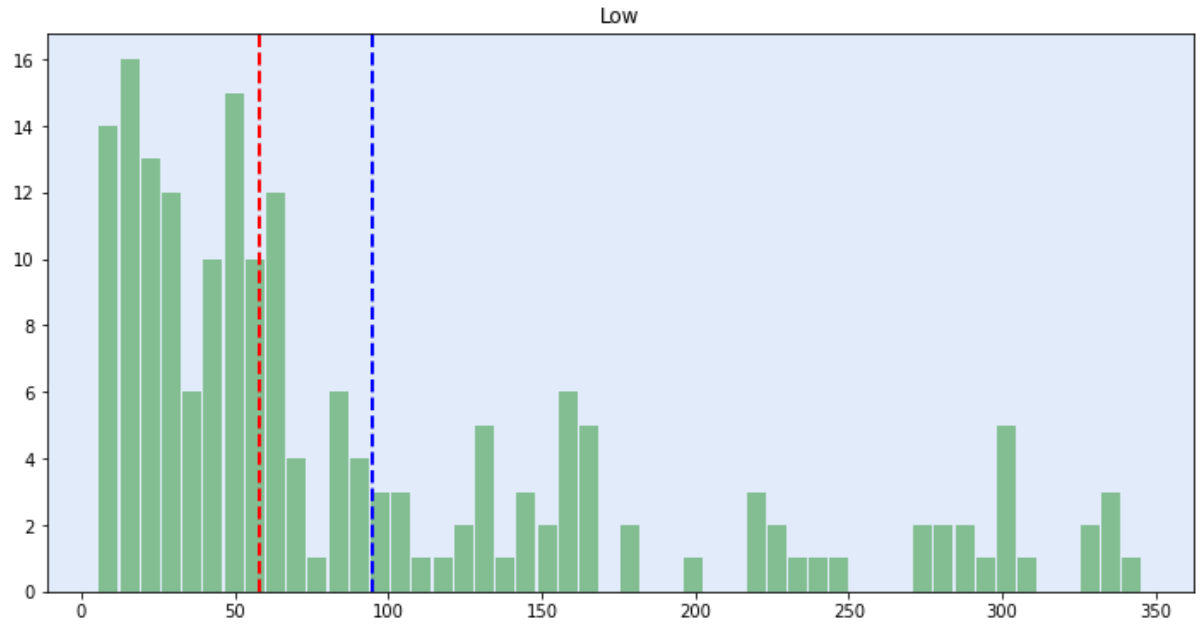
Our dataset does not contain null values which tend to affect our accuracy. If we had null values, we could drop them or imput them with mean or median depending on the situation.

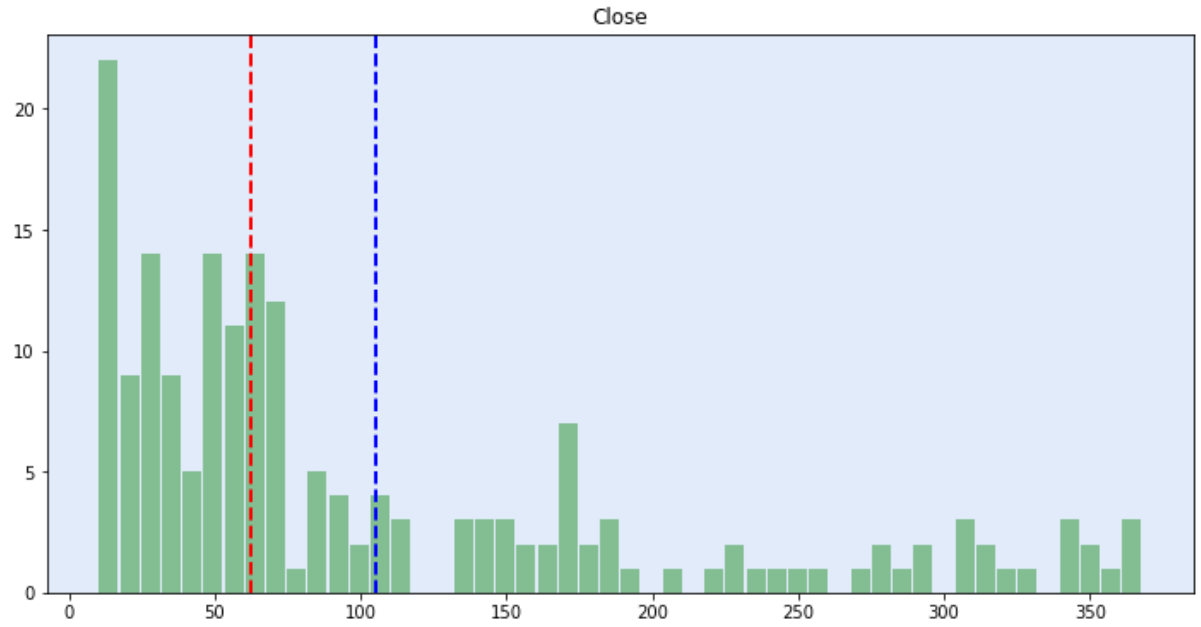
C) Data Visualization:

1. Univariate Analysis: In our yes bank stock market dataset all the features have positively skewed distributions.


Open
0
5
10
15
20
0
50
100
150
200
250
300
350

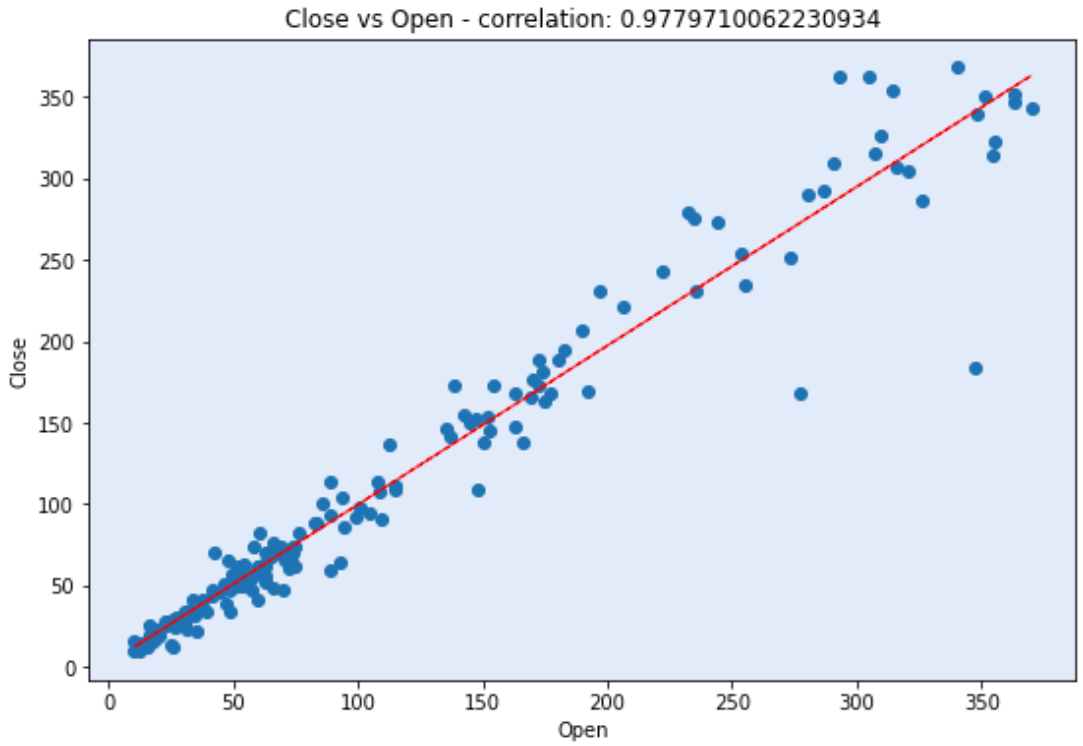

High
0
5
10
15
20
0
50
100
150
200
250
300
350
400

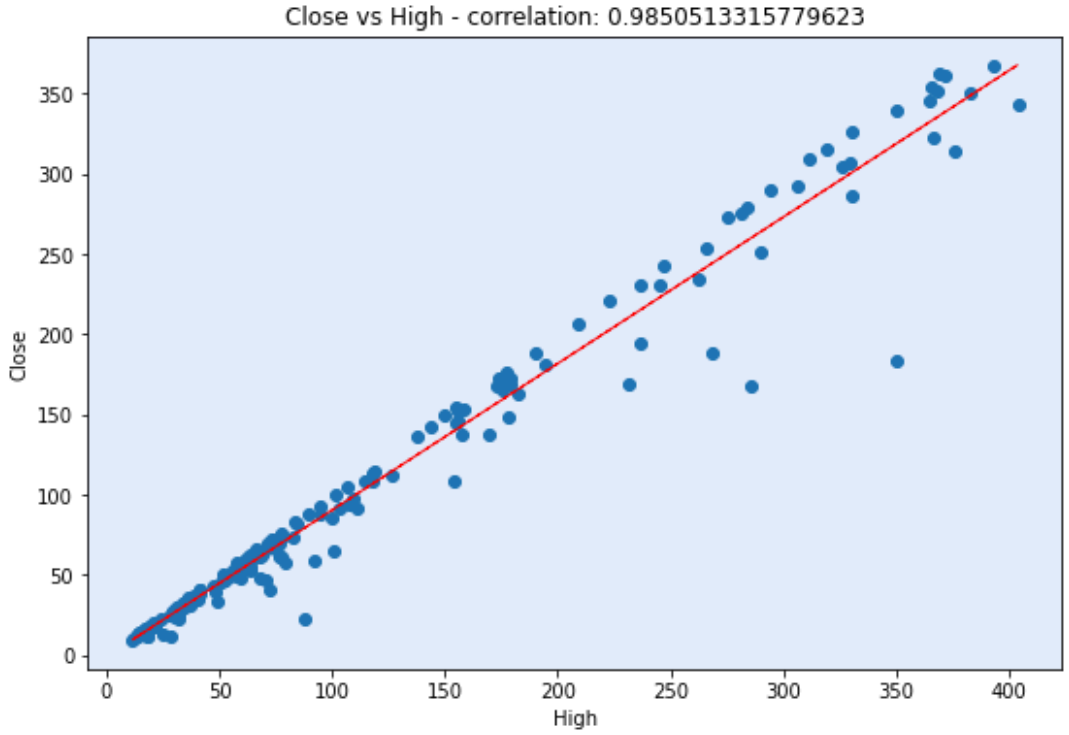


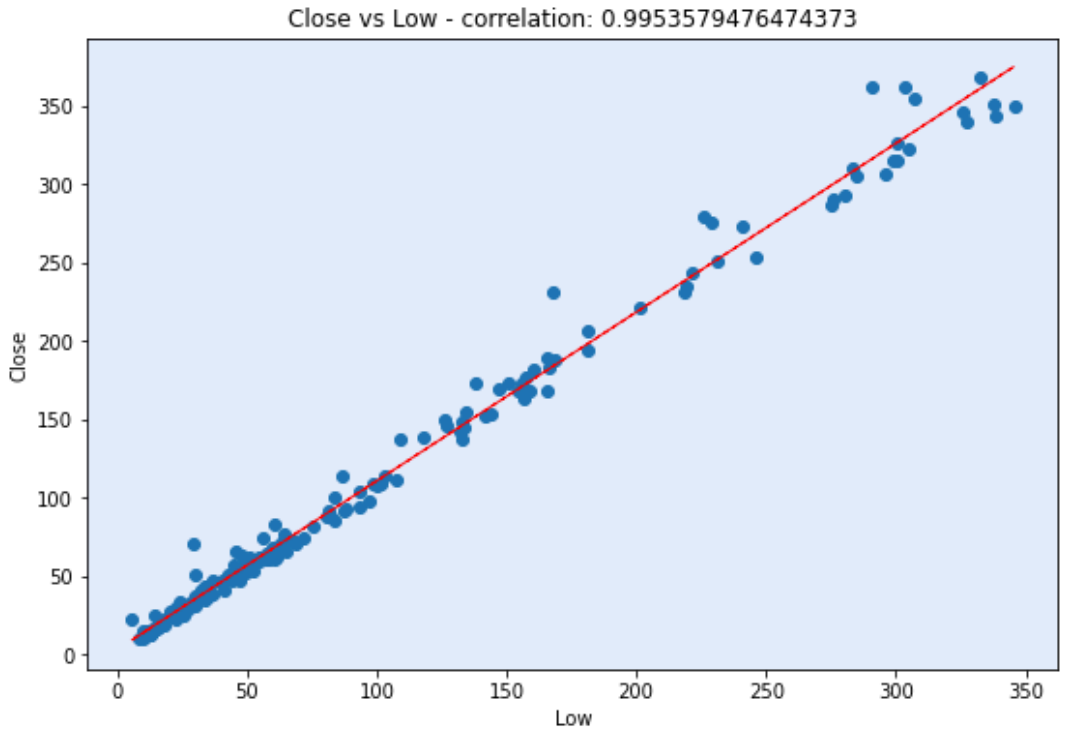


The above graph shows that they are not normally distributed. The mean and median should be equal for perfect normal distribution curve. So, we log transform all the features to normal distribution.

2. Bivariate Analysis: In the context of supervised learning, it can help determine the essential predictors when the bivariate analysis is done by plotting one variable against another.
The graphs below depict that there is high correlation between dependent (Close) and independent variable.


Close vs Open - correlation: 0.9779710062230934
Close
Open
0
50
100
150
200
250
300
350

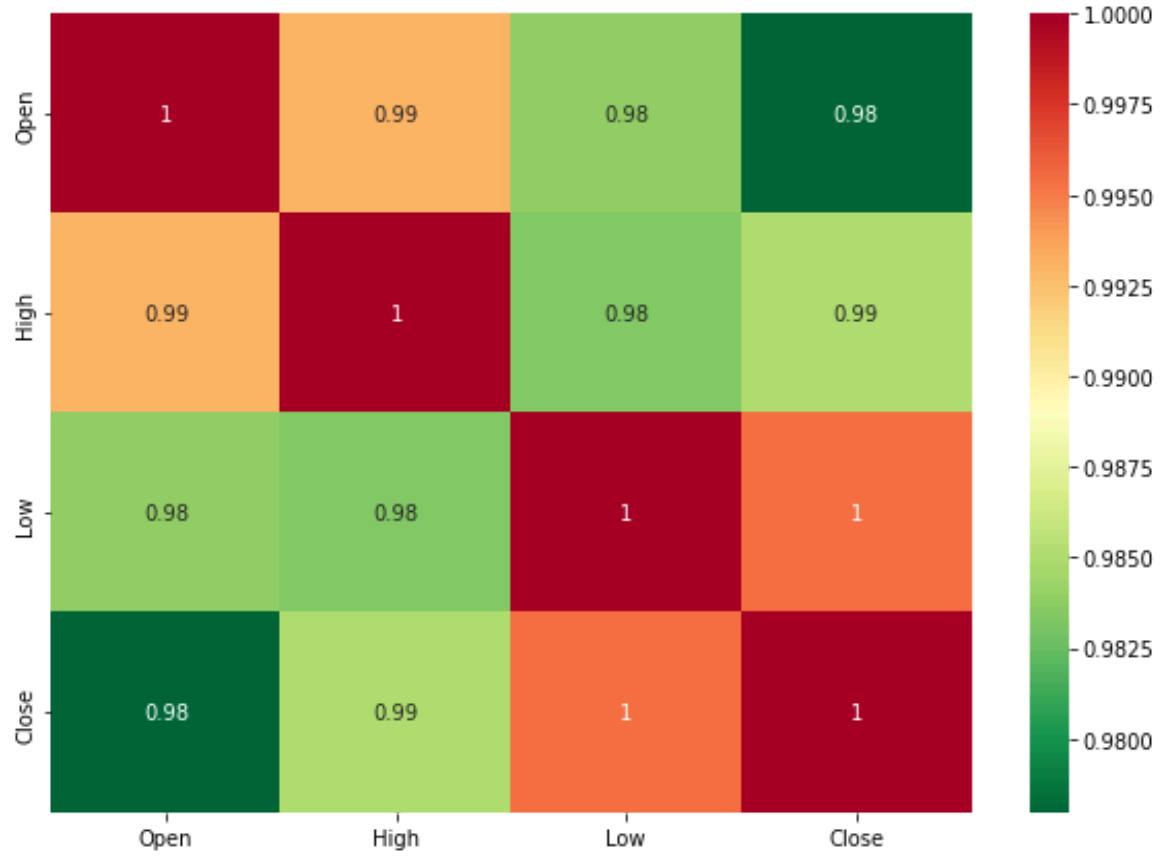

Close vs High - correlation: 0.9850513315779623
Close
High
0
50
100
150
200
250
300
350
400



3. Open price and Close Price:

From the following line plot, We conclude that the stock price is keep on increasing till 2018.But after 2018, the stock price is kept on decreasing due the fraud case involving Rana Kapoor.

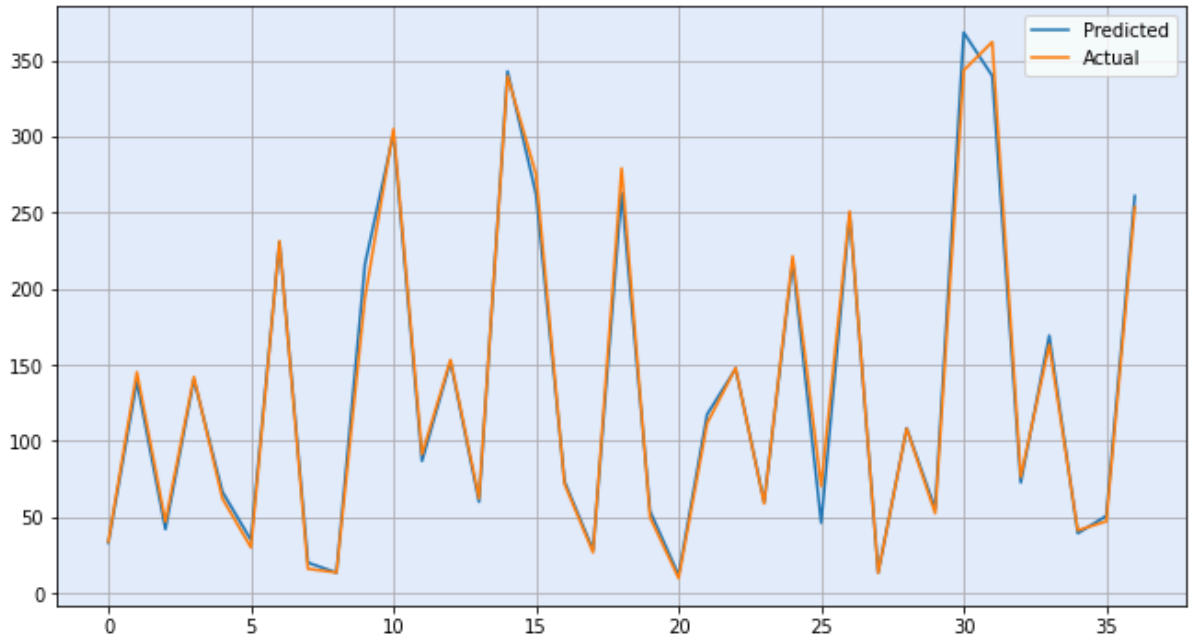
4. Correlation Analysis:
Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between numerical variables. This heatmap shows us the correlation between all numerical variables in our data.



We can see from above heatmap, that all our independent variables are highly correlated with one another. However due to this being a small dataset, we can do nothing to remedy this as removing these features or instances will lead to loss of information.

5. Modelling
A) Linear Regression:

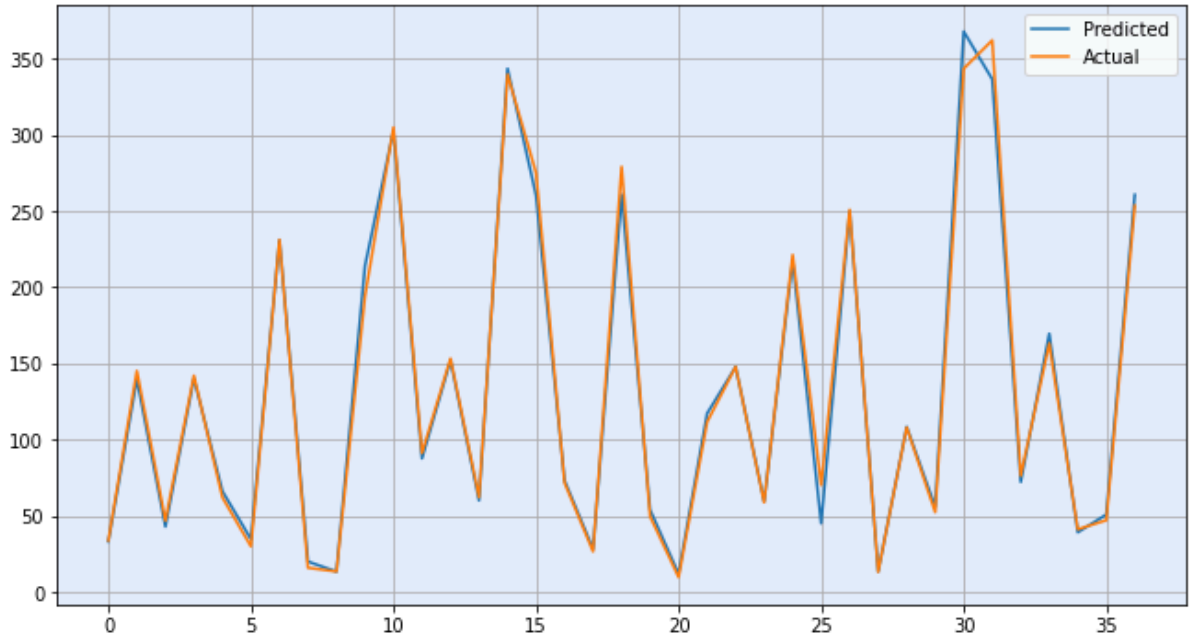
Linear regression is one of the easiest and most popular Machine Learning algorithms. It works best when there is a linear relationship between dependent and independent variables.



Conclusion:
Linear regression evaluation matrices:

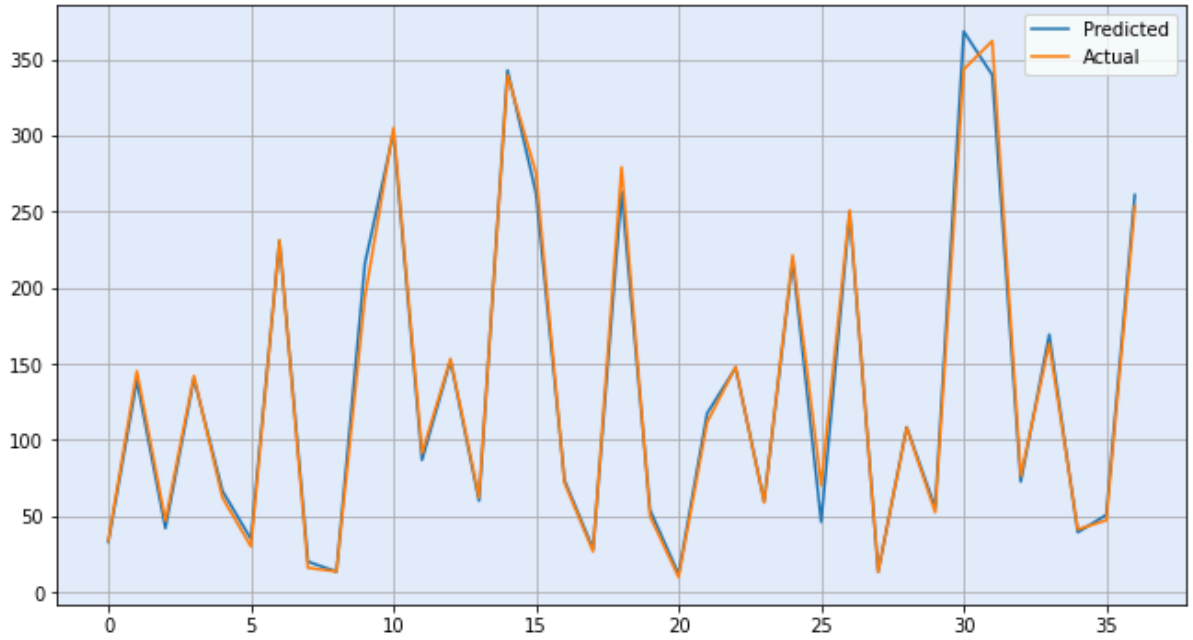
Mean Absolute Error On Testing Data : 4.020951233984054
Mean Squared Error On Testing Data : 80.56653535374932
Root Mean Squared Error On Testing Data : 8.975886326917767
R2 Score For Testing Data : 0.9928422855965229
Adjusted R2 Score for Testing Data : 0.9916878155314459

B) Ridge Regression with cross-validation:
Ridge regression is a regularized linear regression similar to lasso. However, it uses a different L2 penalty term for regularization.
It is used for regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.
The graph below shows the actual and predicted values of target variable as given by the model.



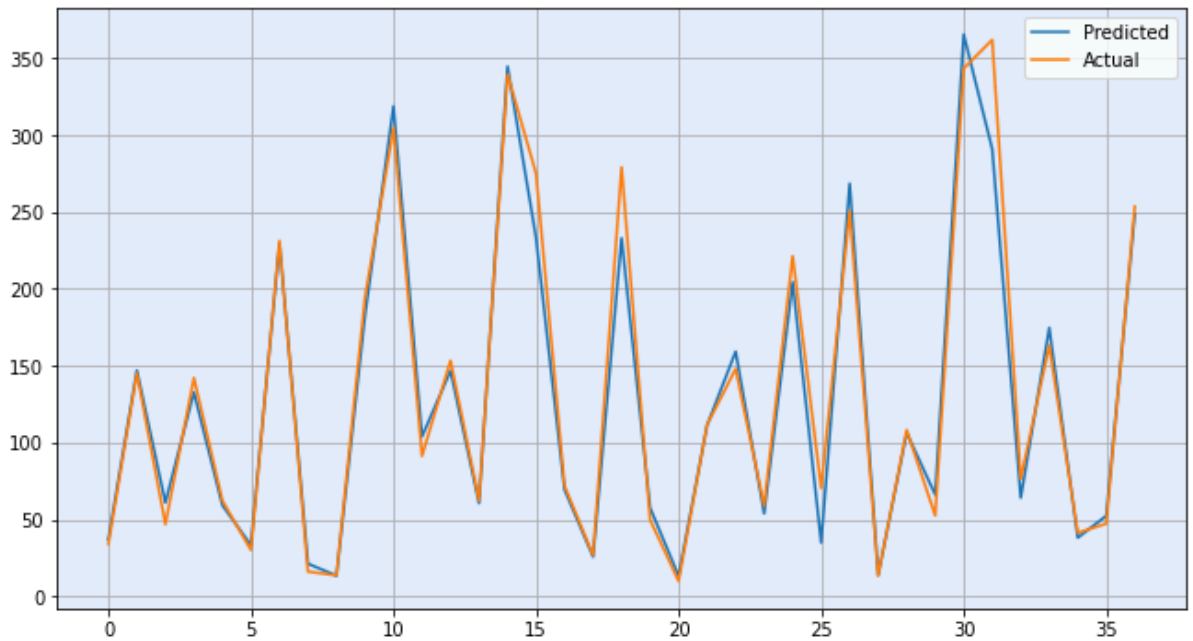
C) Lasso Regression (with cross-validation):

The goal of **lasso regression** is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. It does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.
Lasso performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.



D) Elastic Net Regression with cross-validation:

Elastic net regression works in a manner that takes the best of lasso and ridge regressions. It adds up the penalty terms for regularization in lasso and ridge(L1 and L2) and uses that for regularization. It is used for regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.



6. Final Conclusion

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.**

There is a high correlation between the dependent and independent variables. This is a signal that our dependent variable is highly dependent on our features and can be predicted accurately from them.

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 implemented models performed quite well on our data giving us the accuracy of over 99%.

We checked for presence of Hetero dasceticity in our dataset by plotting the residuals against the Elastic Net model predicted value and found that there is no Hetero dasceticity present. Our **model is performing well on all data- points**.

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.

There are some outliers in our features however this being a very small dataset, dropping those instances will lead to loss of information.

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.

.....