**Yes Bank Stock Closing Price Prediction**

**Sanket Kamble**

**AlmaBetter**

**ABSTRACT:**

Yes Bank is a recognized bank in India's financial sector. Yes bank refers to Youthenterprise Scheme Bank. This bank listed in share market. . The aim of this project is to construct a predictive model for close price prediction. The main point of the stock market is that people try to buy shares in a lower price and sell them when the price goes up, thereby making a profit. Any stock's price may vary depending on a number of variables. Events like the bank management personnel fraud case undoubtedly have a significant impact on stock prices.

Thus we are looking on such one case of Rana Kapoor yes bank fraud case. In order to forecast how other relevant features will affect the stock closing price of the bank, I have been given a dataset of Yes Bank stock prices. Several machine learning models have been applied to make a prediction.

**METHODOLOGY:**

We perform data analysis by doing the following: the data, data description, cleaning data Analyze and visualise the data while paying attention to null and missing values. then using various machine learning techniques. Starting from the dashboard of Almabetter, we must gather the data. Data cleaning and visualisation are then performed. The data is then analysed using various plots, and test results, including mean absolute error, mean squared error, root mean squared error, and r2, are checked to determine the reliability of the regression approach.

**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.

**DESCRIPTION OF DATA:**

While looking to the Downloaded data sets we can see they provided one .CSV file. In this csv file data\_YesBank\_StockPrices.csv details of all stock prices are given . There are five details of share are provided. In file data\_YesBank\_StockPrices data details of 5 columns is

* **Open:** The price at which a stock began trading when the opening bell rang is referred to as the "open."
* **High:** When we talk about high, we mean the maximum pricing at given time.
* **Low:** Low pricing are minimum prices in a given time period.
* **Close:** The closing price of a stock is its market value at the ending of a trading day.
* **Date:** It indicates the date of the investment .

**DATASETS BREAKDOWN:**

The next stage in data visualization is data processing.

* Import important libraries for analysis
* Mount Google drive and read csv files
* see all data information.
* Check the Null values
* Check the duplicate values
* Make the size column in one format
* Remove the special characters
* Change the data type where necessary

Then next steps are

* Exploratory Data Analysis ( EDA)
* Model Implementation
* Metric Evaluation
* Conclusions

**DATA ANALYSIS AND VISUALIZATION:**

**Case 1:**

Lets see year vs closing price of stock graph

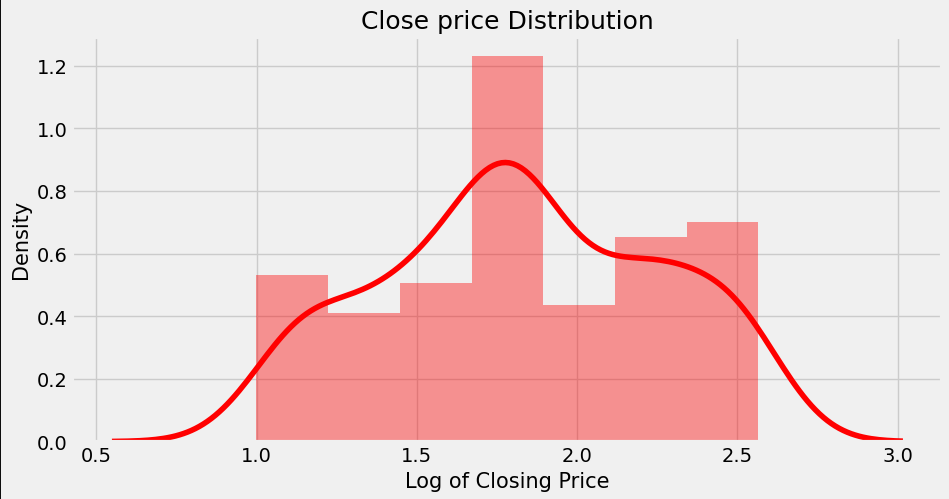
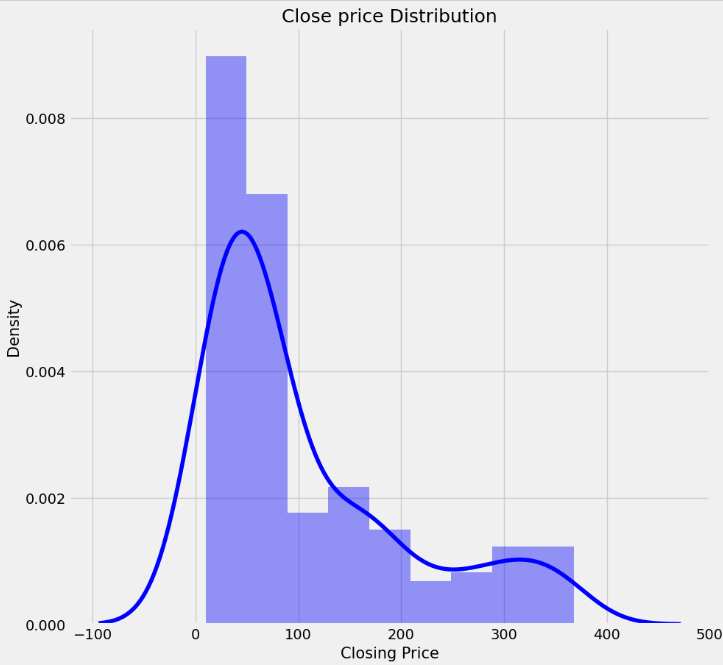


As the graph clearly shows, the stock price was up from 2014-18.

There is a sudden decrease in stocks after 2018 that justifies the effect of the fraud case against Rana Kapoor.

**Case 2:**

Close price distribution

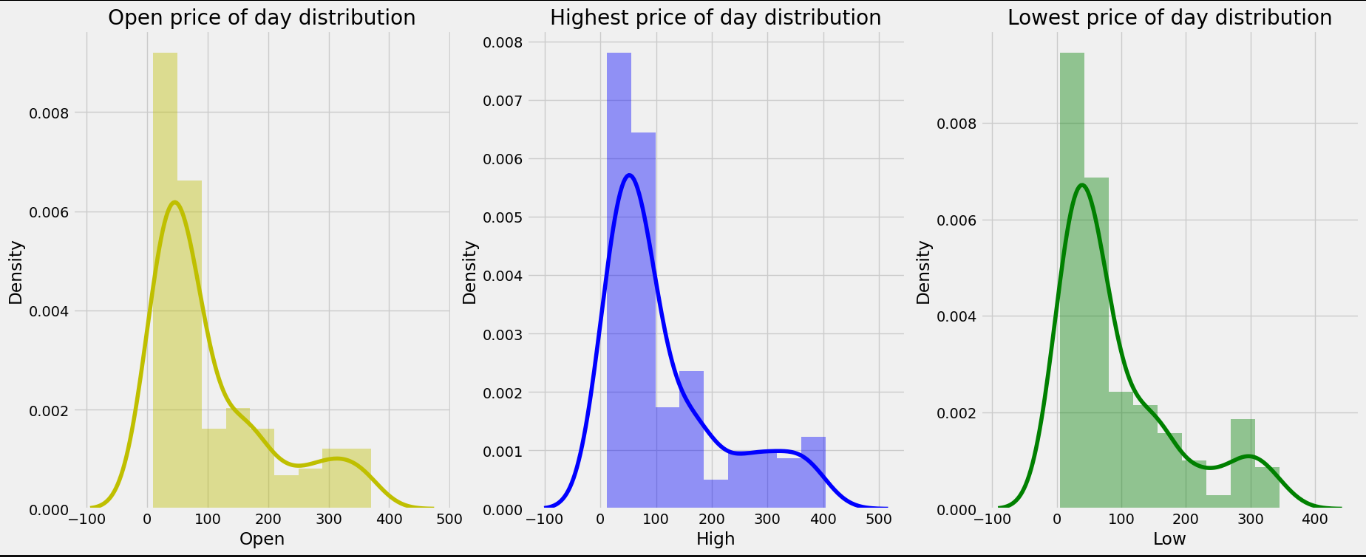


Data of closing price distribution plot is Right skewed

For the purposes of the training algorithm, this distribution must be normal. Closing price undergone a log transformation to make it more normal.

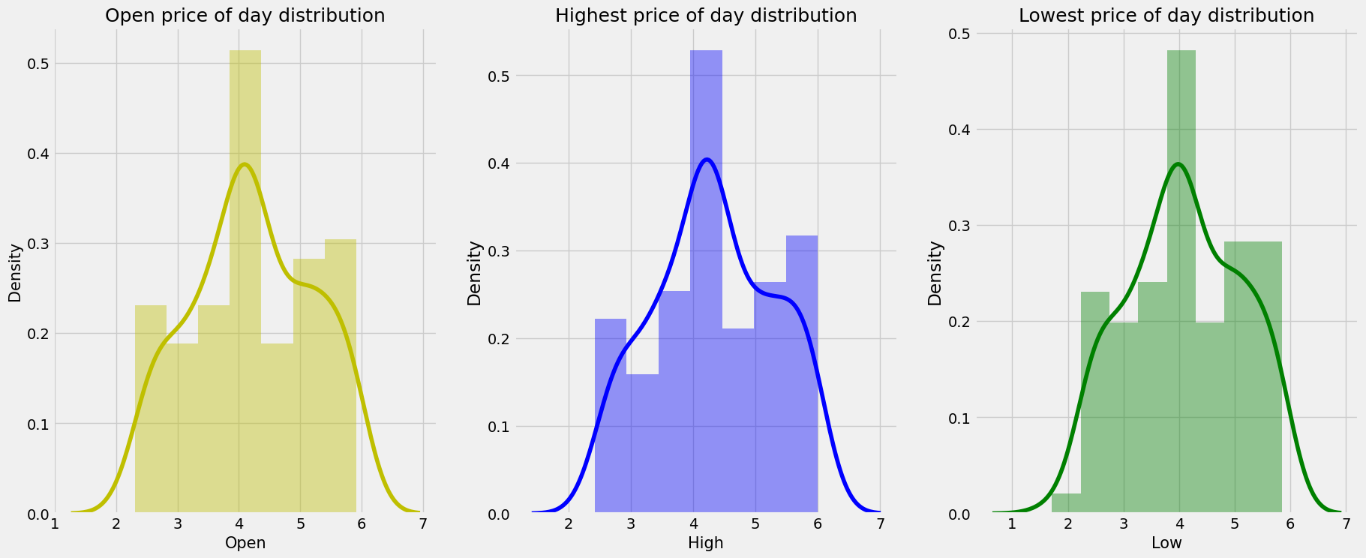
**Case 3:**

Plot the distribution for Open , High and low.



Opening price, high price, and low price distribution are all right skewed.

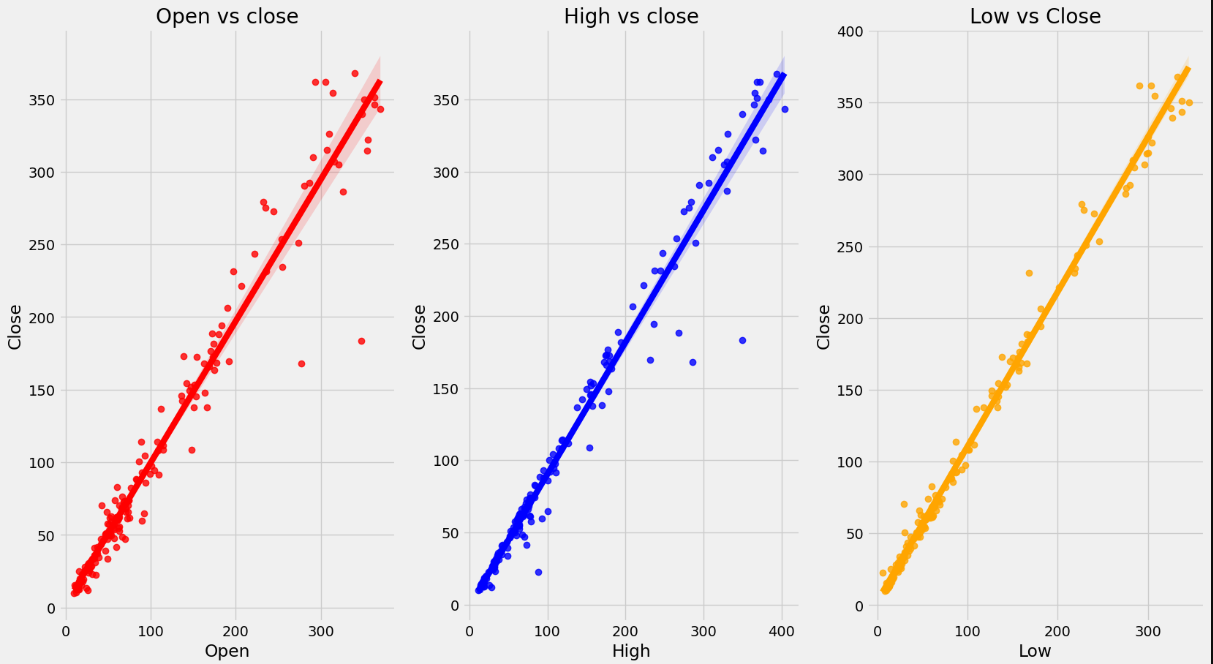
To make them normally distributed apply log transformation



The Opening Price, High Price and Low Price distribution is now a normal distribution.

**Case 4:**

Best fit line

****

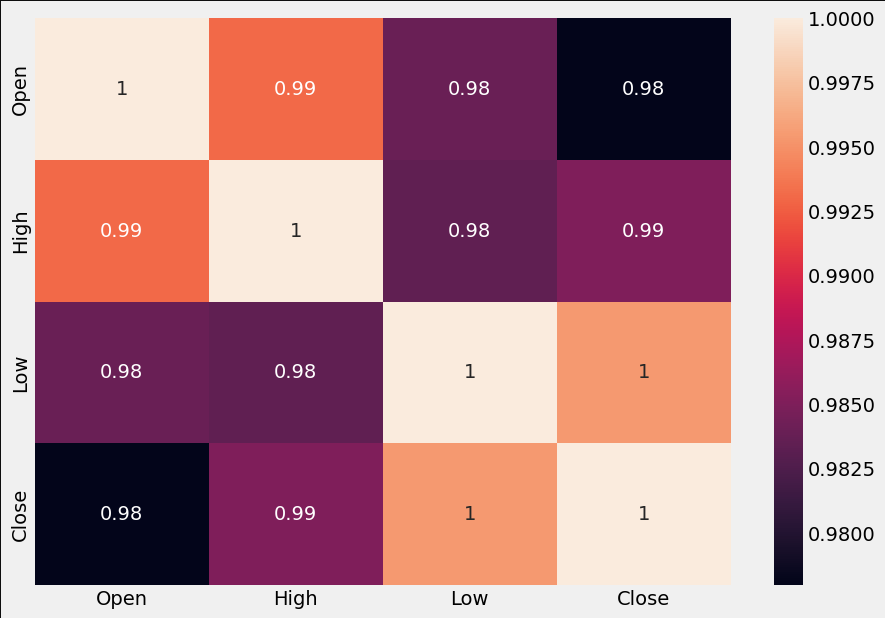
Here we can see relationship between the Dependent Variable (close price) and Independent Variable(open, High, Low price)

All variables vary with our dependent variable which is nearby.

It varies linearly with a few outliers.

**Case 5:**

Correlation analysis using Heatmap:



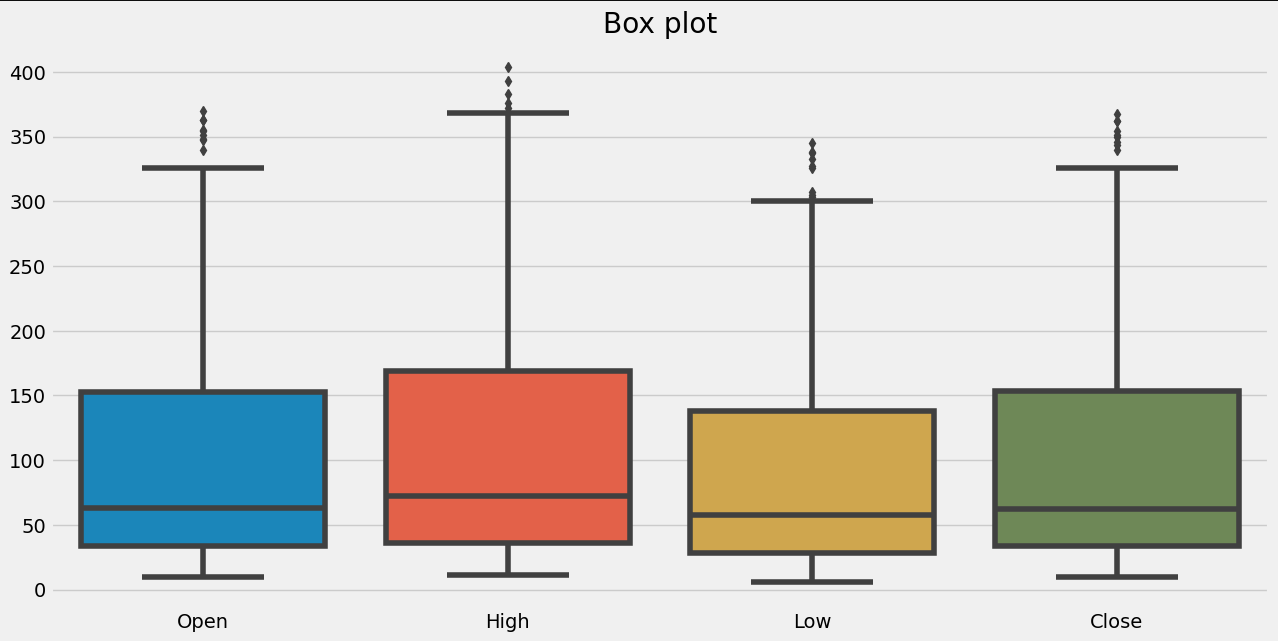
A procedure for determining the connections between two variables is known as correlation.

Positive correlation: There would be 1 correlation. This indicates that the two variables changed in the same direction, either up or down.

We have multicollinearity since there is a significant correlation between the independent variables in this situation. High multicollinearity makes it difficult to fit models and make predictions since even small changes to just one independent variable can lead to wildly unanticipated outcomes.

**Case 6:**

box plot



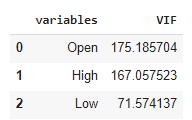
There are a few outliers. Every feature is extremely corelated with each other.

The median are close to each other.

**Multicollinearity**

Calculating VIF  (Variation Inflation Factor) will allow us to determine the presence of multicollinearity in our dataset and its extent.

Generally, any variable with VIF above 5  is regarded as multillinarian.

The VIF is very high. We are not removing any features from the dataset while we attempt to predict the outcome, assess the model's performance, and make adjustments as necessary. 

**Train & Test Split:**

X = Independent variable

Y = Dependent variable

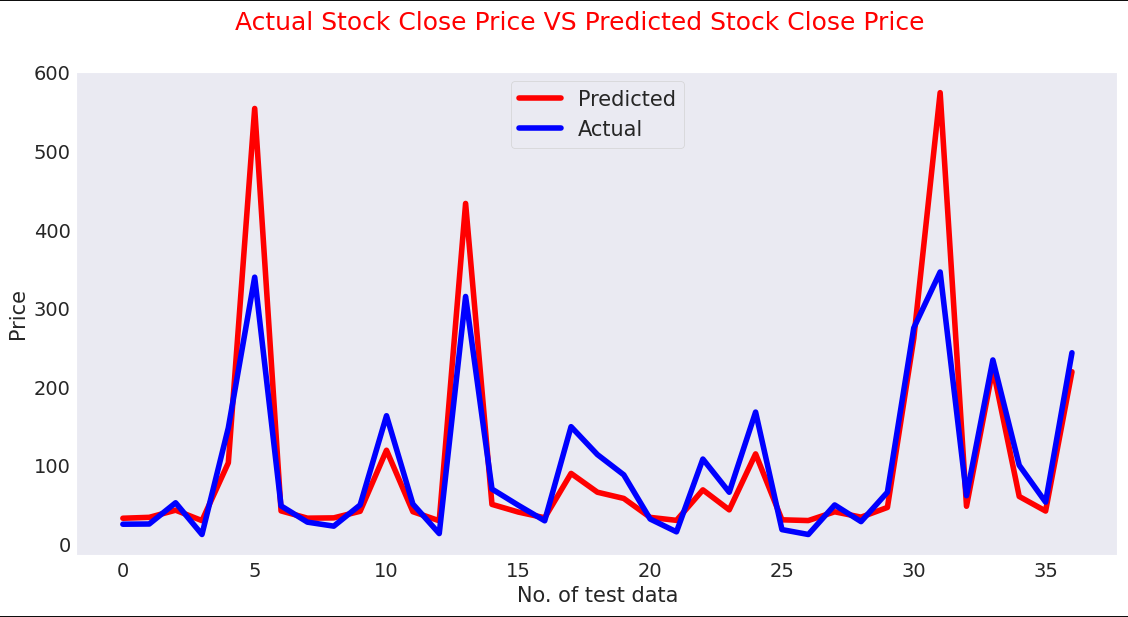
From dataset Training dataset and testing datasets are splited. Training datasets are used to train and optimise algorithms. The test dataset is used to evaluate the effectiveness of the train model. Here, 80% of the data are utilised for training, while the remaining 20% are used for testing.

**Algorithms:**

It is time to use the following models on the provided dataset.

* **Linear Regression**

A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you're trying to predict. The differences between predicted and real output values are minimized by linear regression by fitting a straight line or surface.

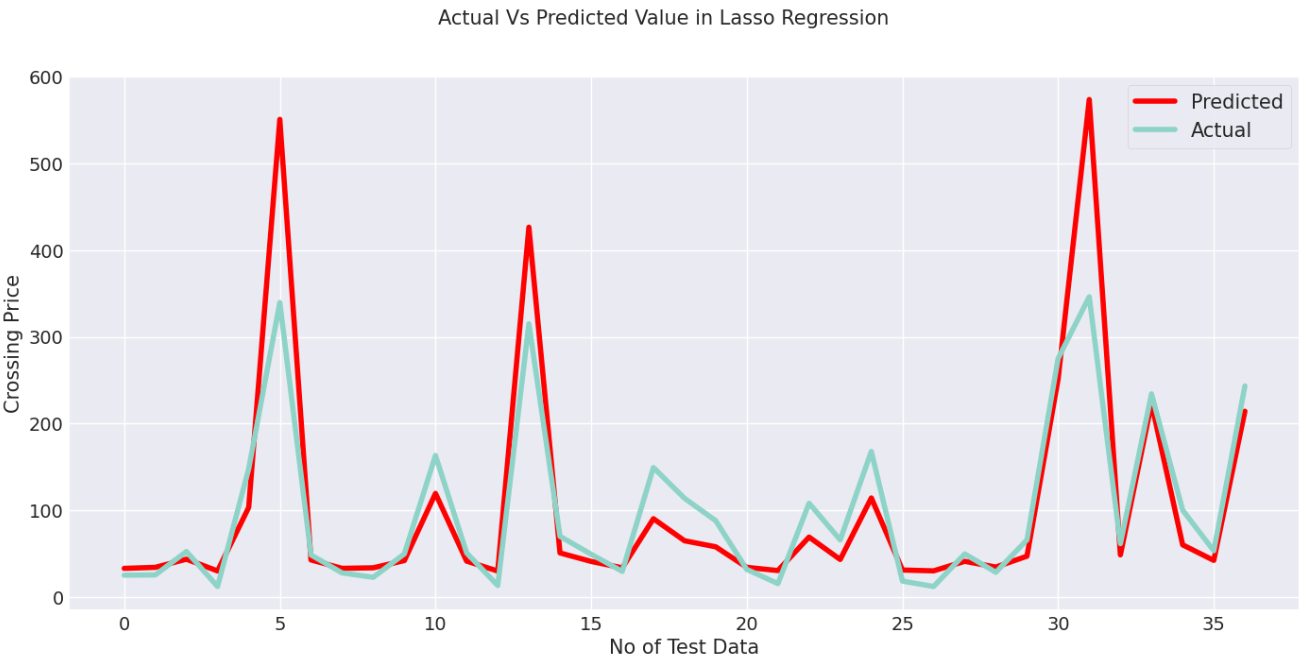


MSE: 0.032

R2: 0.823

* **Lasso Regression**

Lasso regression is a linear regression, but it uses a "shrinkage" technique where the coefficients of determination shrink to towards zero. Lasso (least absolute shrinkage and selection operator) is regression analysis method that variable selection and regularization performs both in order to enhance the prediction accuracy and interpretability of the resulting statistical model. This method performs L1 regularization.



MSE score: 0.032 Lasso Regression Cross Validation

R2 score: 0.82 MSE score: 0.0322 R2 score: 0.819

* **Ridge Regression**

A regularised variation of linear least squares regression is ridge regression. It functions by reducing the weights or coefficients of the regression method. Ridge regression is a model adjustment method that is used to analyze any data that suffers from multi-llinearity. This method adjusts to L2. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

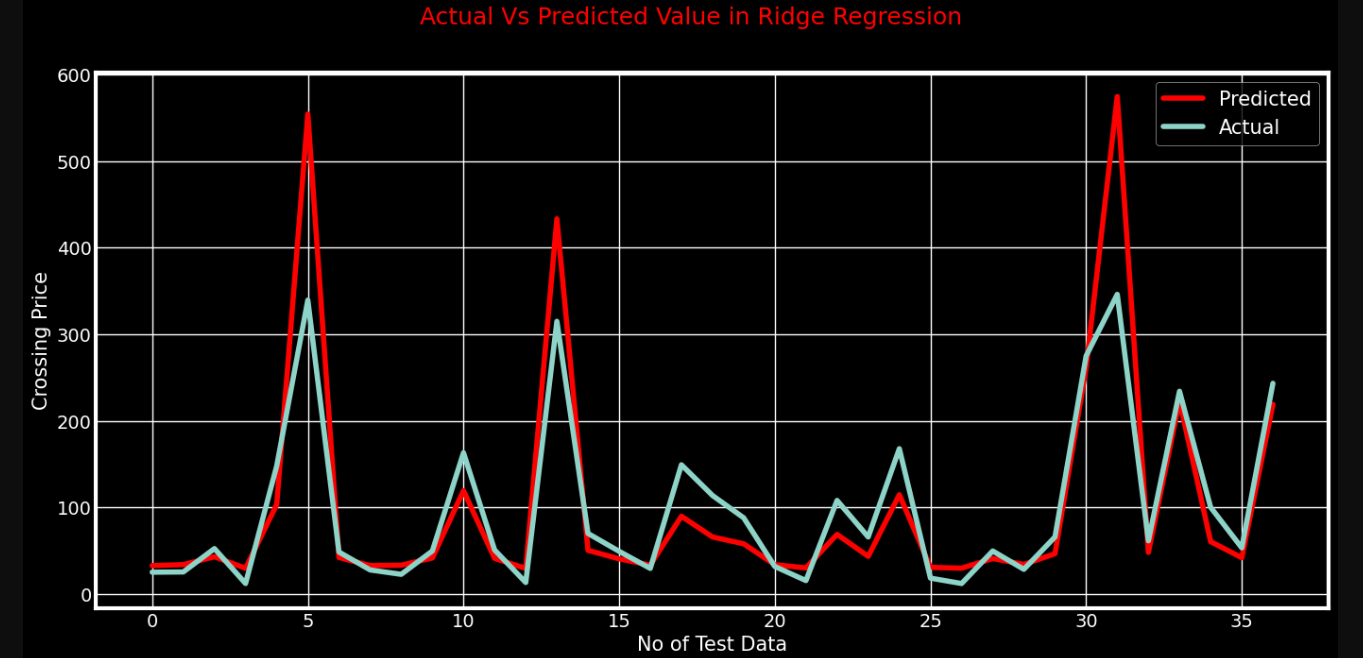
MSE score: 0.0316

R2 score: 0.8225

Cross Validation on Ridge Regression

MSE : 0.033

R2 : 0.817



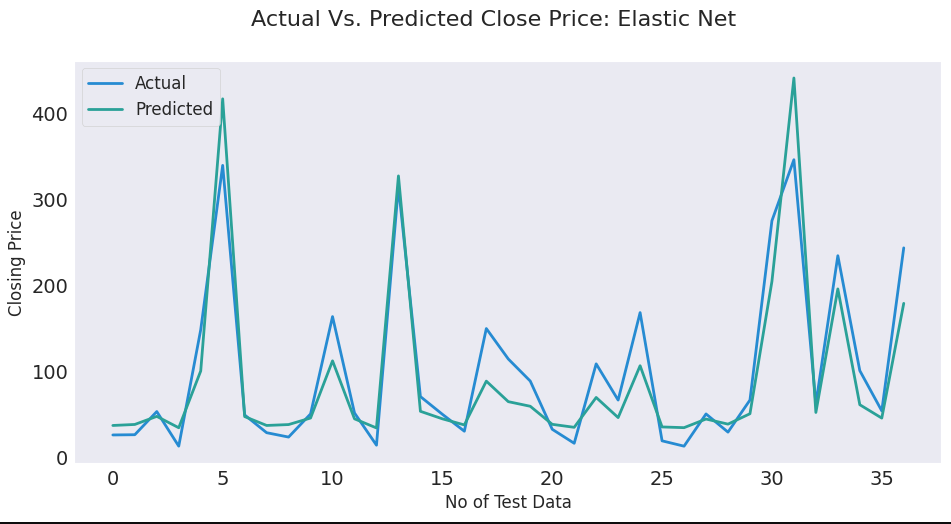
* **Elastic Net regression**

The third type of regularization method is elastic net regression. It was created as a result of the Lasso regression's limitation. Lasso regression can’t take correct alpha and lambda values as per requirement of data. It combines two popular penalties, specifically the L1 and L2 penalty.

Test Performance

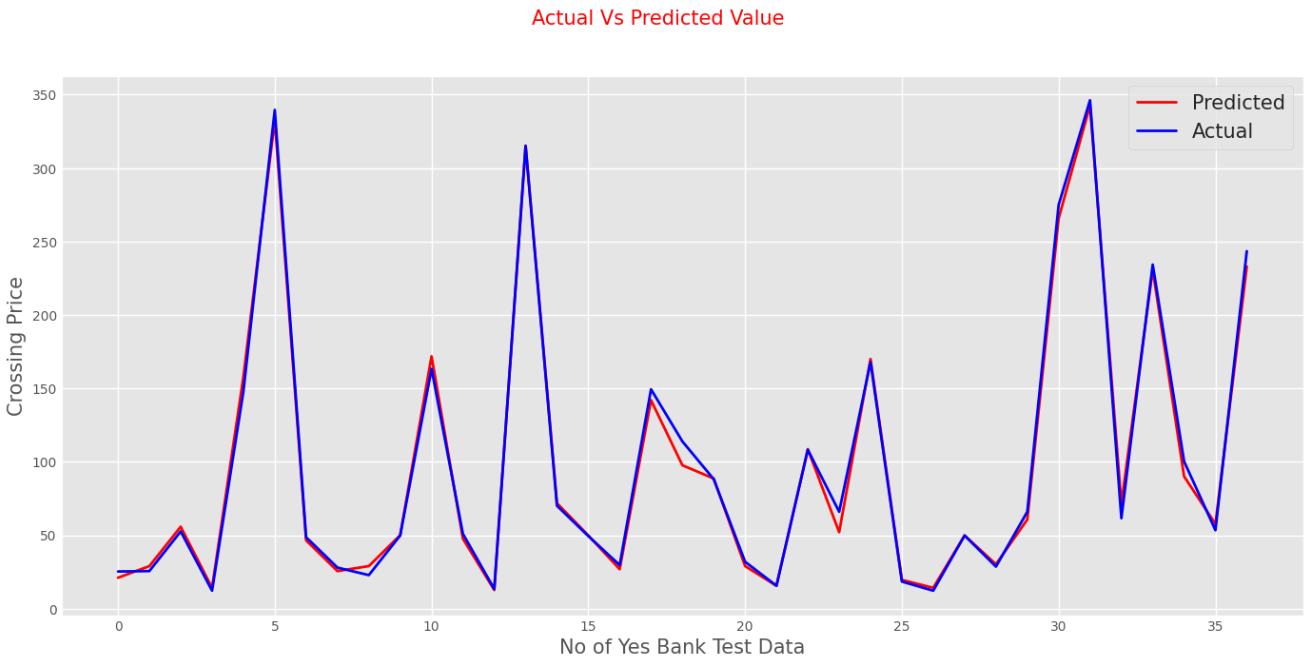
MSE : 0.036

R2 : 0.796



* **Gradient Boosting Regressor**

The difference between the present forecast and the known correct target value is calculated using gradient boosting regression. This variation is referred to as residual. After that, a weak model that maps features to that residual is trained using gradient boosting regression.



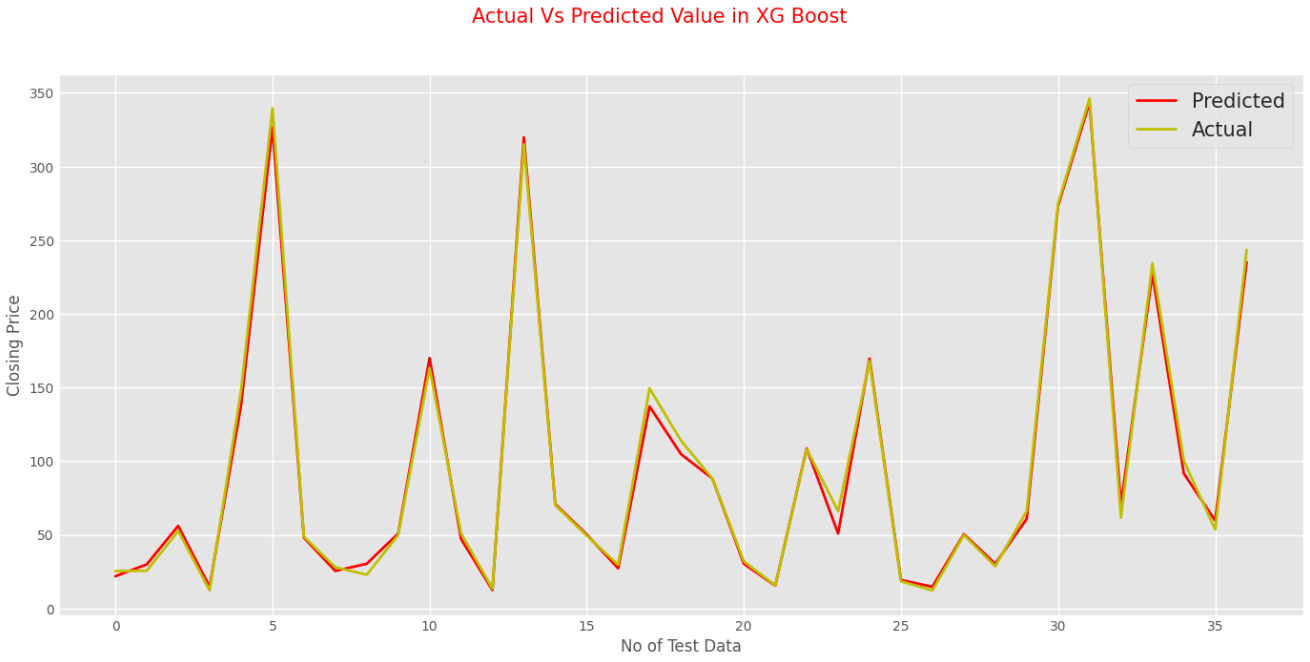
MSE : 0.002

R2 : 0.99

* **XGBRegressor**

Each feature utilised for prediction is generally ranked in order of importance by the XGBoostRegressor. Gradient boosting has the advantage that retrieving relevance ratings for each attribute is not too difficult after the boosted trees have been built.

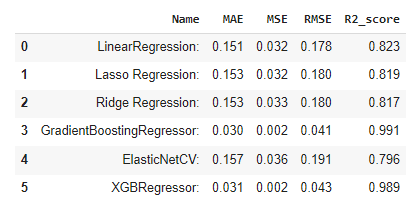
The two key benefits of XGBoost are model performance and execution speed.



MSE score: 0.0019

R2 score: 0.9894

**Compare all the Models:**



**Conclusion**

* Beginning from Exploratory Data Analysis we see the sudden change in stock price from 2014.
* Bank share price is at the highest in 2018-19.After there is a Sudden fall in price of stock .
* After that stock price start to increase again but price fall again.
* From the scatter plot we can see High,open,low price of share are directly correlate with the closing price of share.
* we applied following regression model on data set and result are evaluted and compared
* LinearRegression
* Lasso Regression
* Ridge Regression
* GradientBoostingRegressor
* ElasticNetCV
* XGBRegressor
* we check test performance such as mean absoulte error, mean squared error, root mean squared error, r2
* We got 99% highest accuracy for GradientBoosting Regression and 98% for XGB Regressor
* we got almost simillar result for LinearRegression,Lasso Regression and Ridge Regression
* Cross Validation has been applied on various algorithms. However, the outcome is nearly identical.
* Using data on the closing price of Yes Bank's shares, GradientBoostingRegressor and Xgboost regression is the best model to apply.

**References**

GeeksforGeeks

Alma Better

Wikipedia