YES BANK STOCK CLOSING PRICE PREDICTION

CAPSTONE PROJECT-II
ALMABETTER, BANGLORE

Kumar Vijay Mhaske

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 of around 185 observations

It contains the following features:

- **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.
- **High** - The high is the highest price at which a stock traded during a period.
- **Low** -The low is the highest price at which a stock traded during a period.
- **Close** -The closing price is a stock's trading price at the end of a trading day. This makes it the most recent price of a stock until the next trading session. The closing price is calculated as the weighted average price of the last 30 minutes, i.e., from 3:00 PM to 3:30 PM in case of equity.

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

INTRODUCTION:

Stock market is characterized as dynamic, unpredictable and non-linear in nature. Predicting stock prices is a challenging task as it depends on various factors including but not limited to political conditions, global economy, company's financial reports and performance etc. Thus, to maximize the profit and minimize the losses, techniques to predict values of the stock in advance by analyzing the trend over the last few years, could prove to be highly useful for making stock market movements. Traditionally, two main approaches have been proposed for predicting the stock price of an organization.

- ✓ Technical analysis method uses historical prices of stocks like closing and opening price, volume traded, adjacent close values etc. of the stock for predicting the future price of the stock.

- ✓ The second type of analysis is qualitative, which is performed on the basis of external factors like company profile, market situation, political and economic factors, textual information in the form of financial news articles, social media and even blogs by economic analysts.

- ✓ Nowadays, advanced intelligent techniques based on either technical or fundamental analysis are used for predicting stock prices. Particularly, for stock market analysis, the data size is huge and also non-linear.

- ✓ To deal with this variety of data an efficient model is needed that can identify the hidden patterns and complex relations in this large data set. Machine learning techniques in this area have proved to improve efficiencies by 60-80 percent as compared to the past methodology

STEPS INVOLVED:

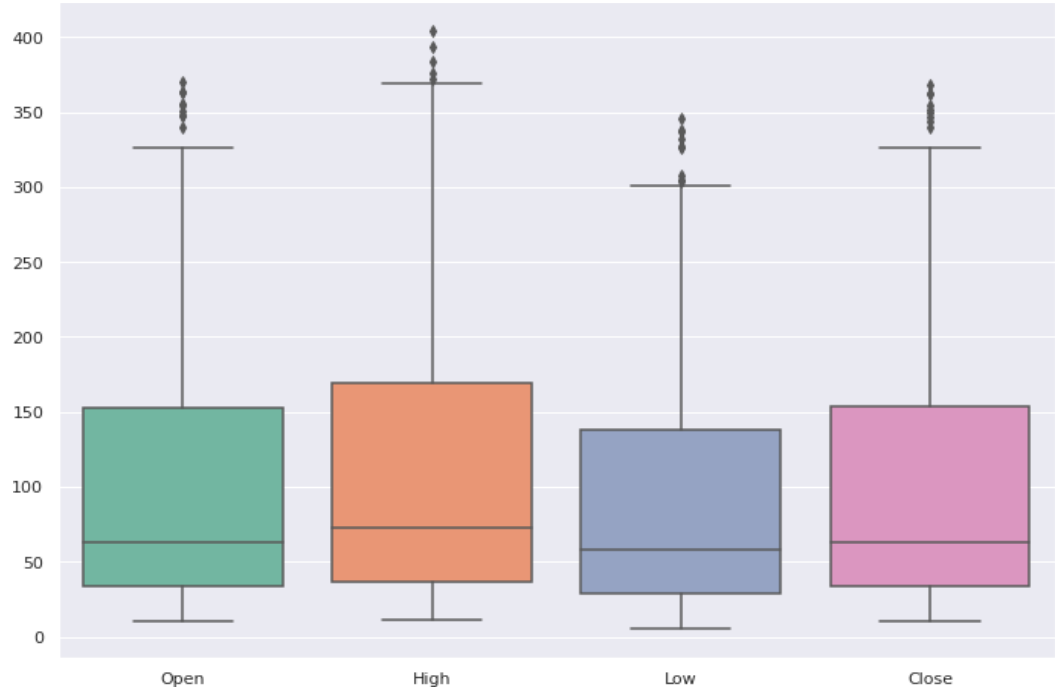
1. **Collection Of Data:** - Before building any machine learning model, it is vital to understand what the data is, and what are we trying to achieve. Data exploration reveals the hidden trends and insights and data preprocessing makes the data ready for use by ML algorithms. So, let's begin. . . To proceed with the problem dealing first we will load our dataset that is given to us in a csv file into a data frame. Mount the drive and load the csv file into a data frame

2. **Discussing Problem Statement:** - After analyzing the datasets we discussed with every single problem to overcome it. We all decided to divide our task and initialized it with our own problem statement. The problem statement was based on the target variable we took for analysis.

3. **Data cleaning:** - The next task was data cleaning which was easy with this dataset. As mentioned in the above points the data were float64 dtype, int64 dtype, object dtype, and datetime64.

4. **Exploratory Data Analysis:** - After data cleaning, it was sure to target some important columns for Exploratory Data Analysis. Matching the data with the correct suitable problem by python libraries to result from some insightful visualization was a great task. These also gives us a more information and graphs.

5. **Visualization of Analysis**: - The EDA parts make clear about data in a picture and graphical form. Mainly we perform matplotlib and seaborn libraries of python for the data analysis. The libraries help a lot with graphs.
    - **Missing values:** No missing values in dataset.
    - **<u>Checking Outliers</u>**:



EXPLORATORY DATA ANALYSIS(EDA)

The primary goal of EDA is to support the analysis of data prior to making any conclusions. Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization method.

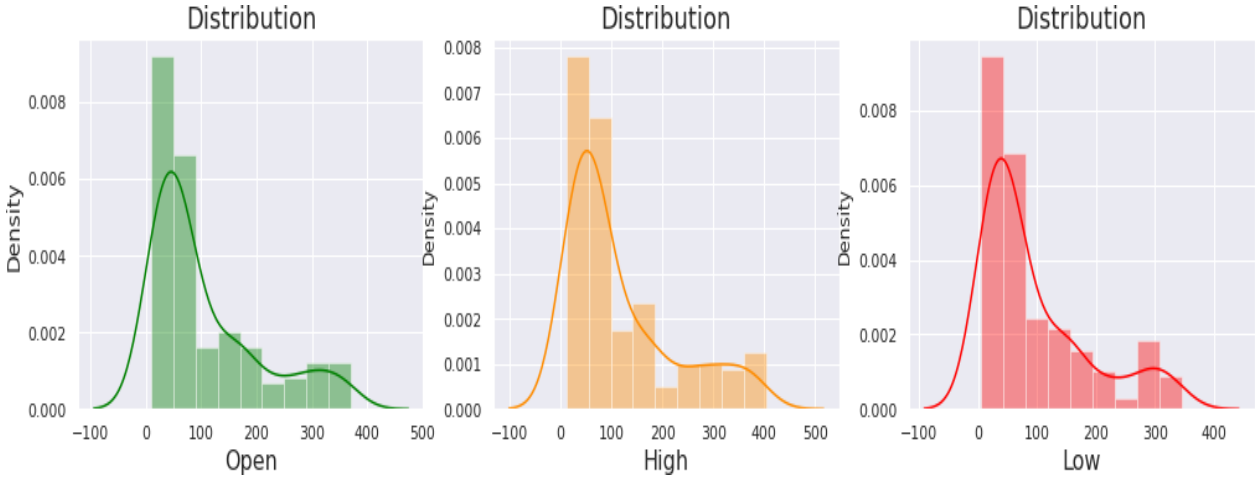
- **The trend of Close:**

- Now, we can easily see in above plot **the trend is increasing from 2009 to 2018** but after that the trend decreases. This is because of the fraud case ofinvolving ***Rana Kapoor.***

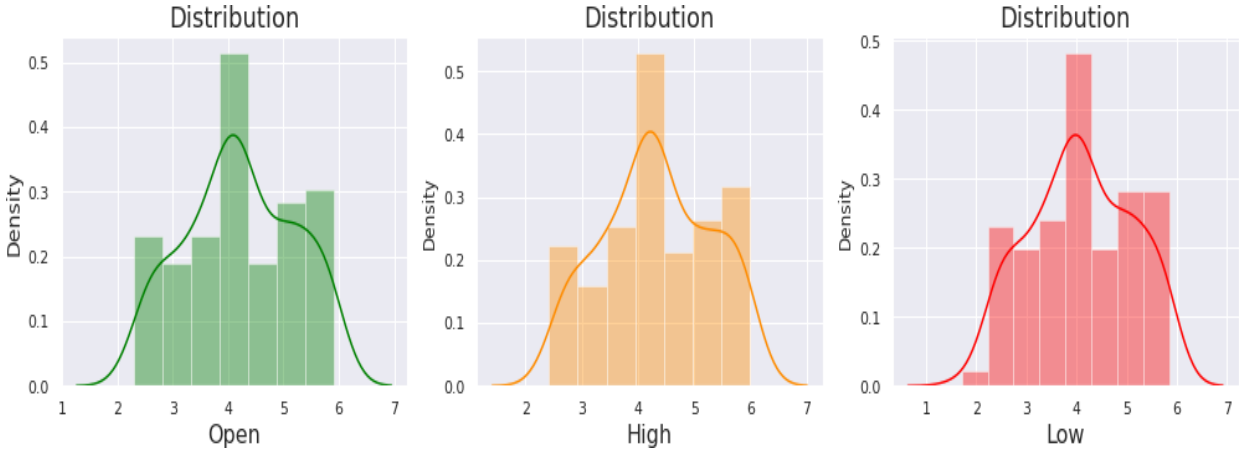
- **Dependent Variable of Close Price Of stock:**

- Now, distribution of closing price is more normal, after applying log transformation

- **Independent Variable of Open, High and Low Price Of stock:**

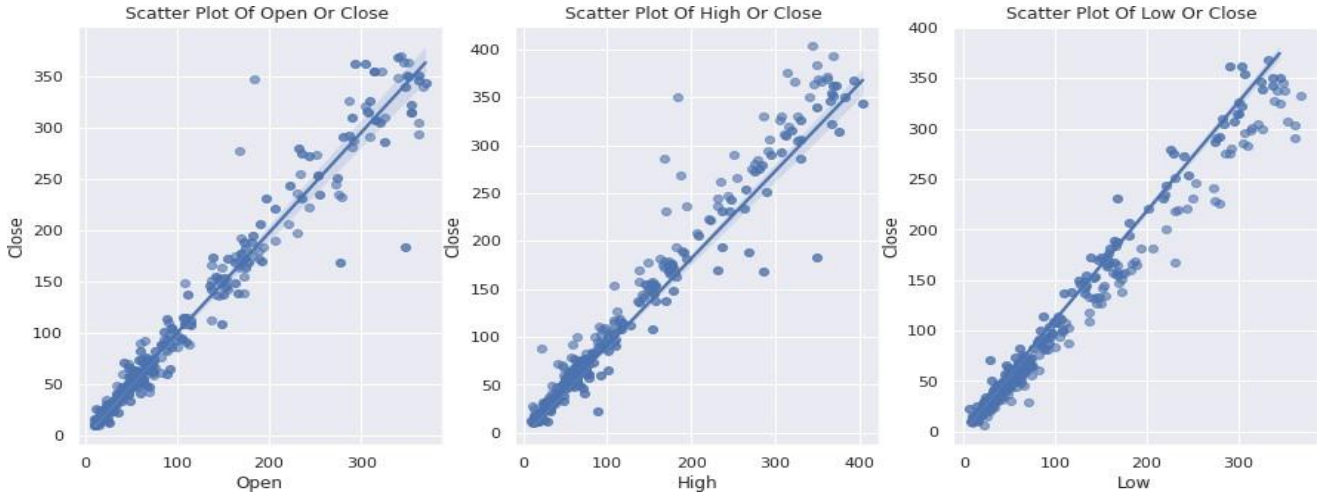


- It's looking like **rightly skewed** for all features.

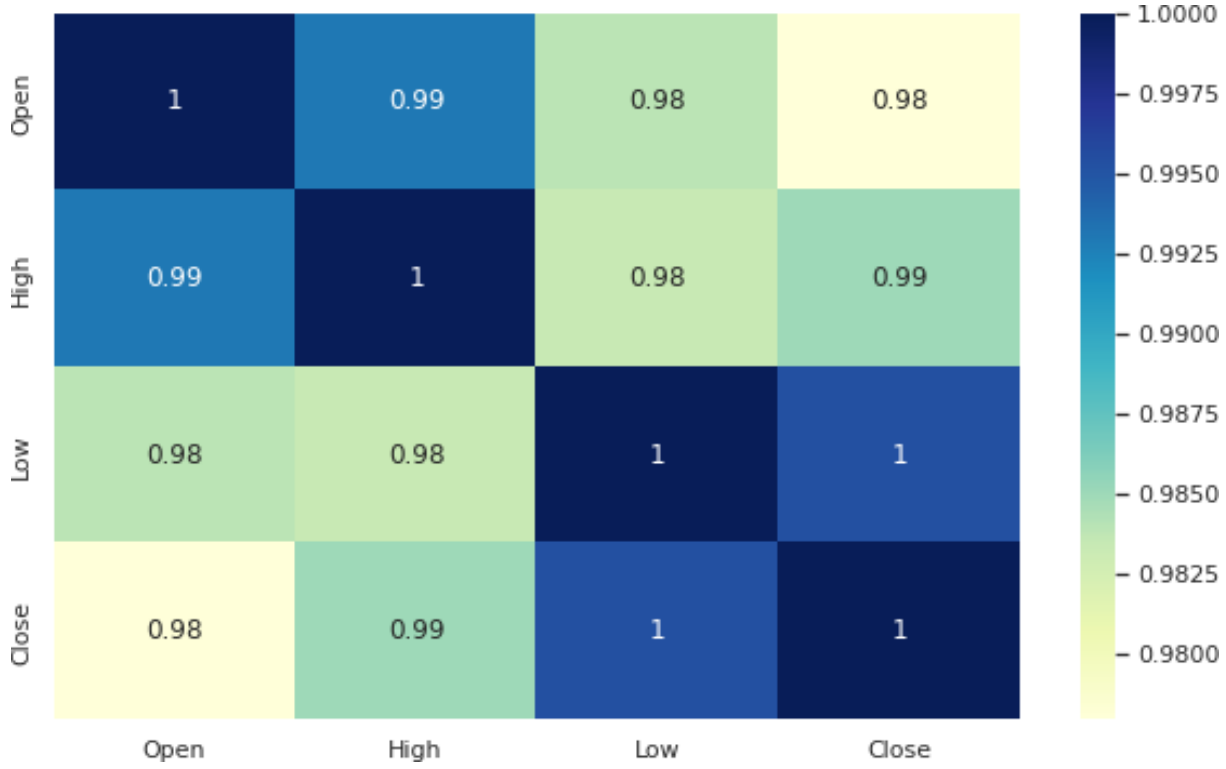


- After applying log transformation to make it normal distribution.

- **Relation between Dependent and Independent Variable:**

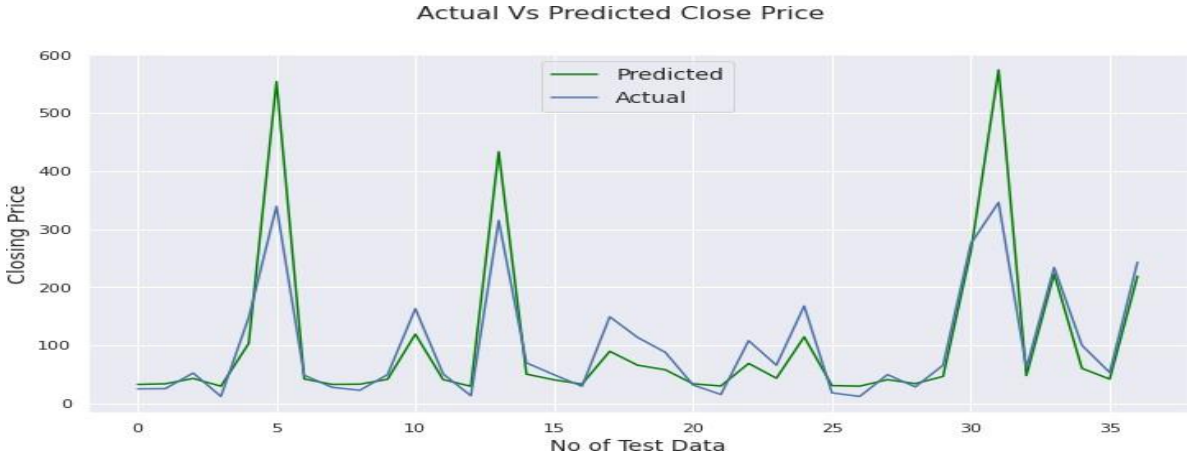


- **Correlation With Heatmap:**

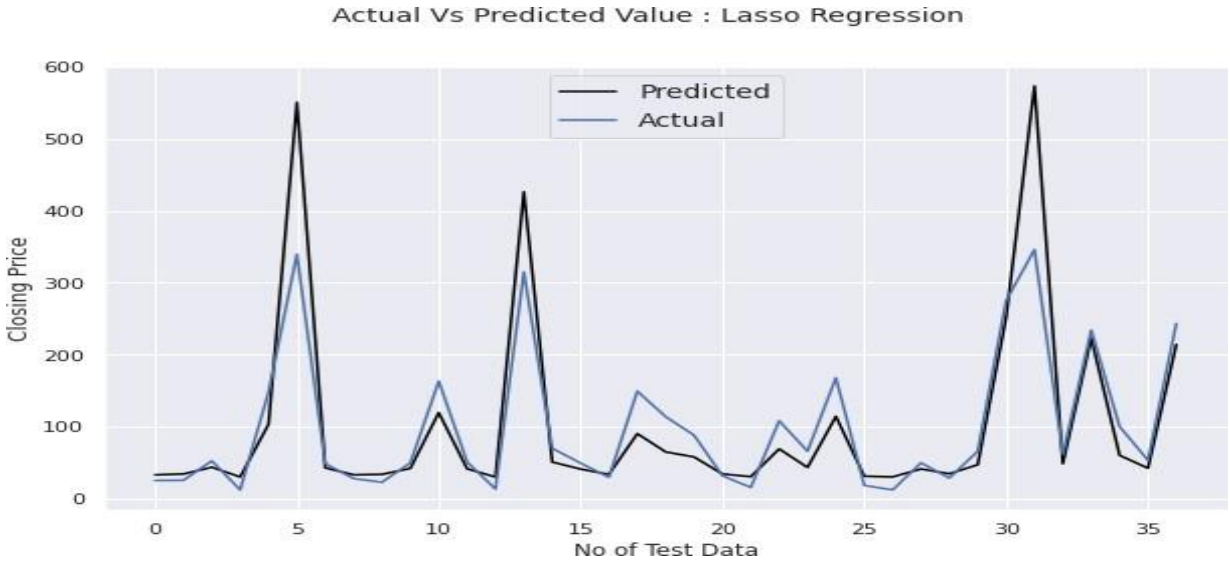


Model Development

1. **Linear Regression:** The most basic machine learning algorithm that can be implemented on this data is linear regression. The linear regression model returns an equation that determines the relationship between the independent variables and the dependent variable.

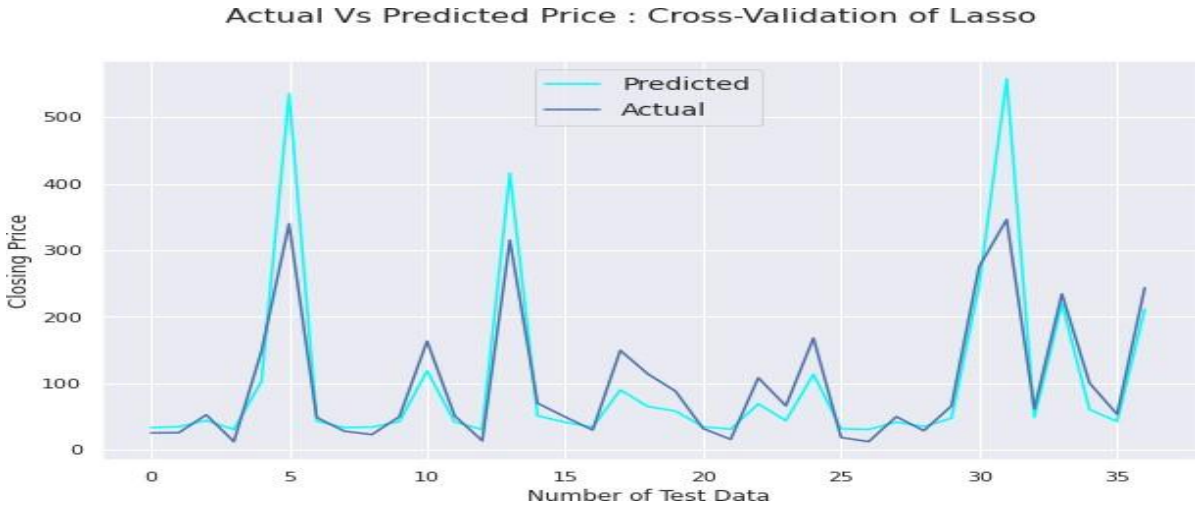


2. **Lasso Regression:** It is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e., Models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection elimination.

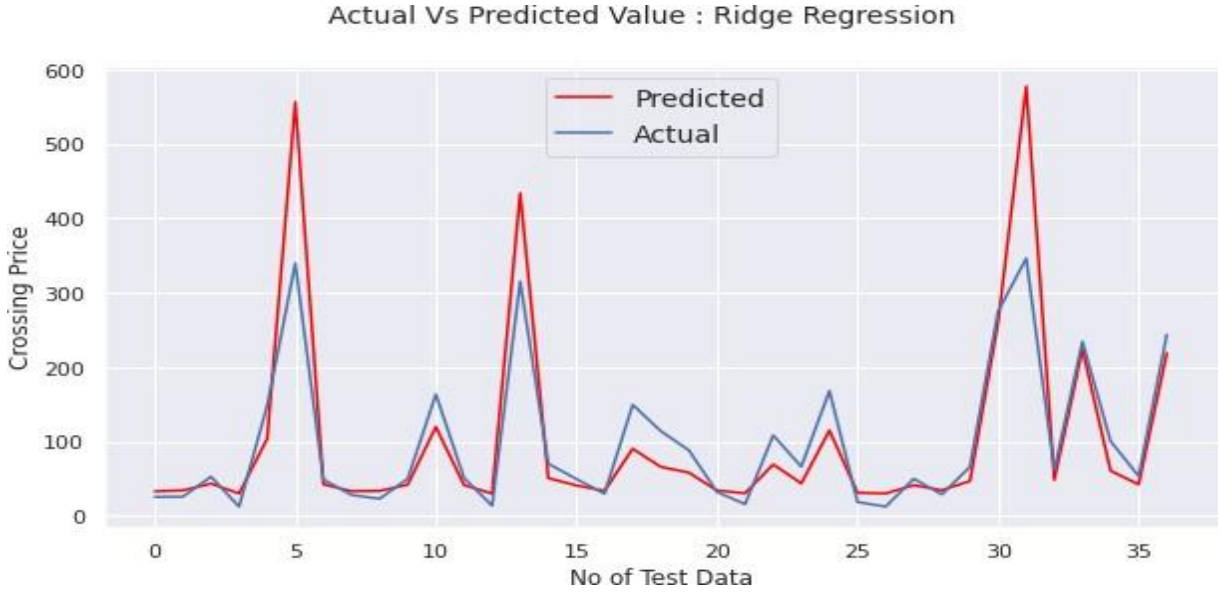


- Cross-validation, sometimes called rotation estimation or out of sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set).
- The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

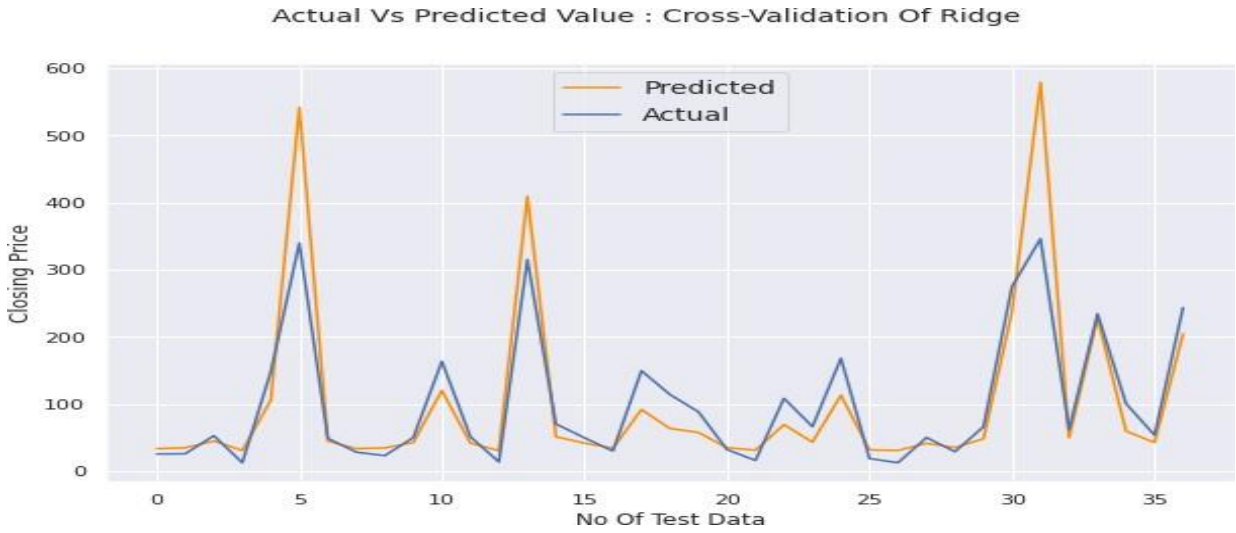
- To reduce variability, in most methods multiple rounds of cross validation are performed using different partitions, and the validation results are combined (e.g., averaged) over the rounds to give an estimate of the model's predictive performance.



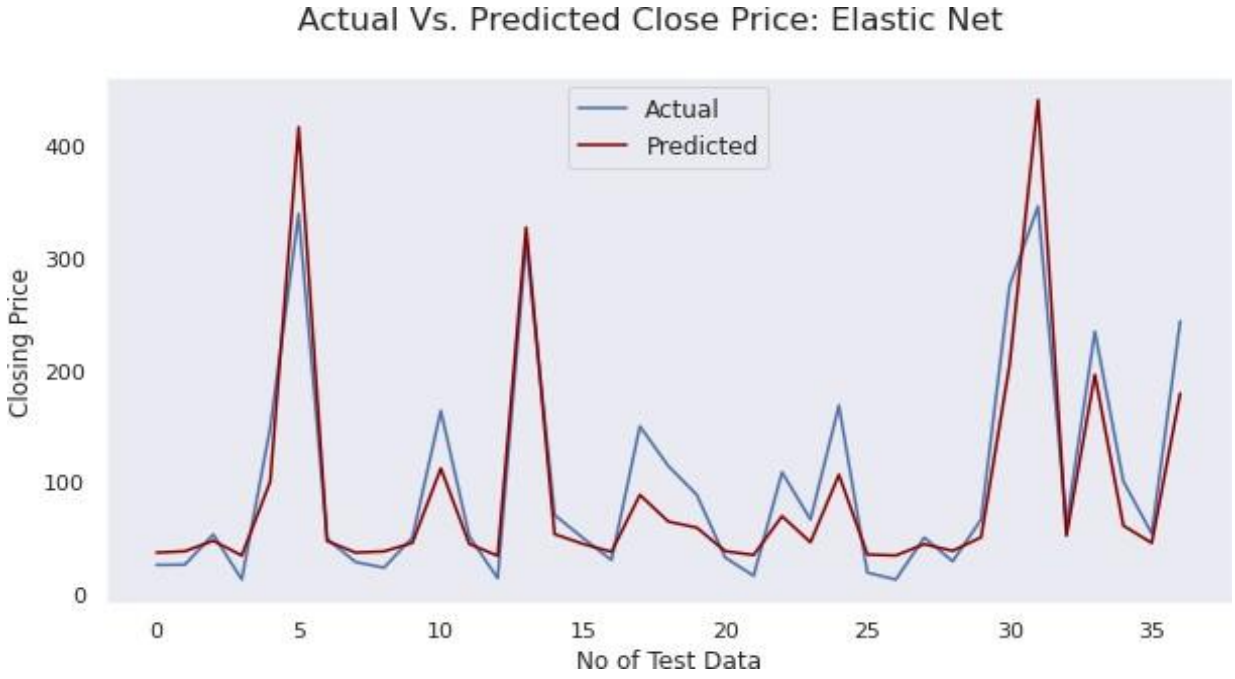
3. **Ridge Regression:** It is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.



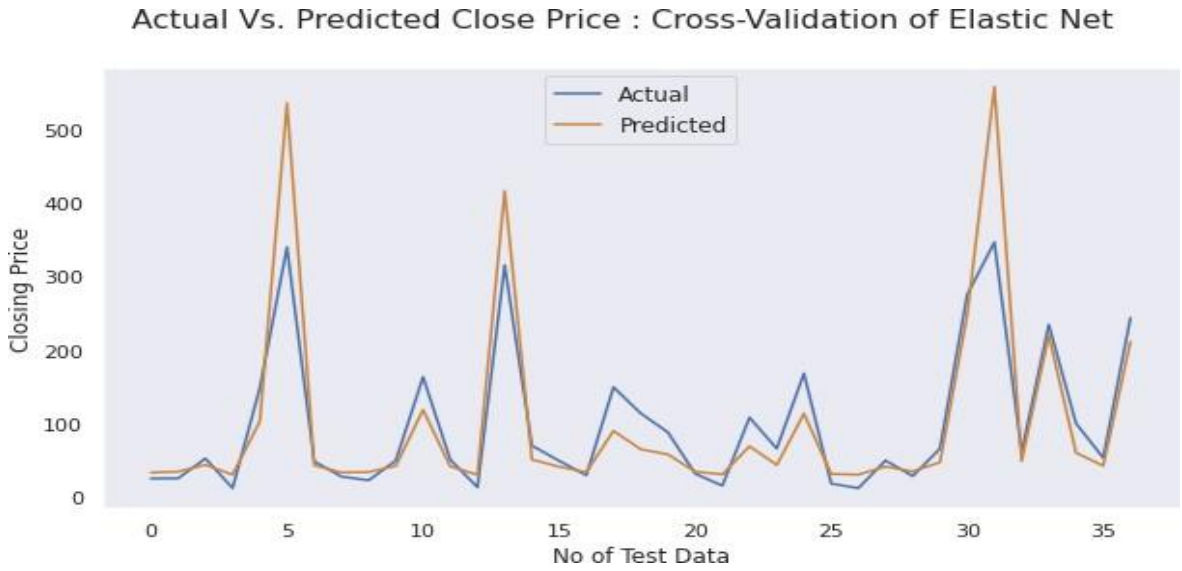
- After Implementing Cross Validation of Ridge:



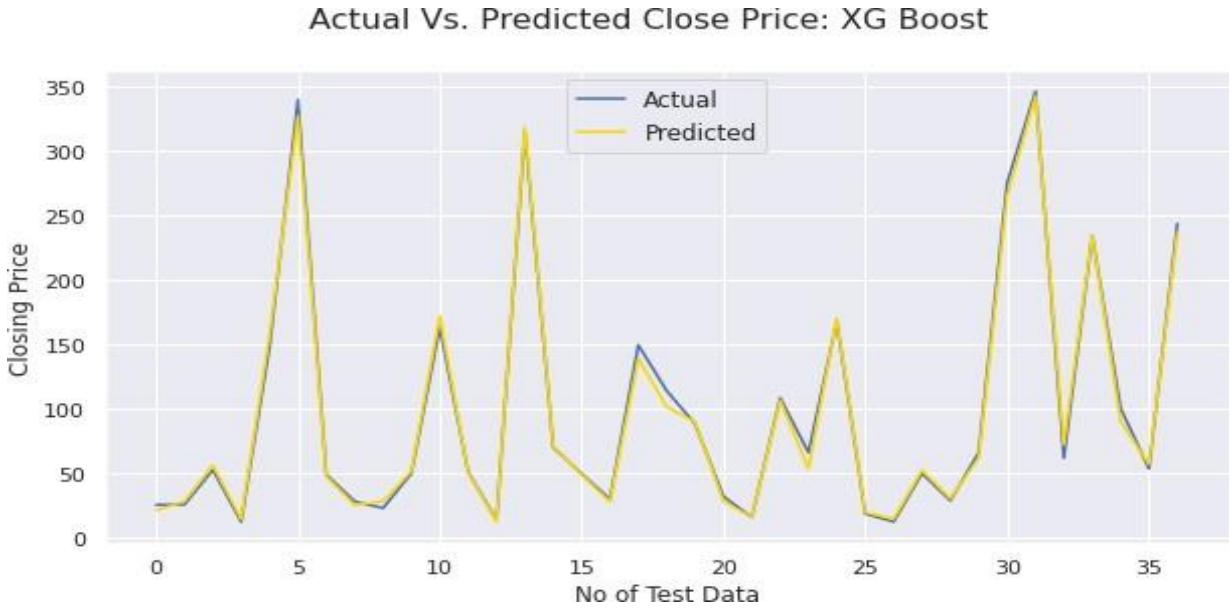
4. **Elastic Net Regression: -** Elastic Net is a regression method that performs variable selection and regularization both simultaneously. The term regularization is the main concept behind the elastic net. Regularization comes into the picture when the model is overfitted. Now we need to understand what overfitting means, so overfitting is a problem that occurs when the model is performing good with the training dataset, but with the test, the dataset model is giving errors; in this situation, the regularization is a technique to reduce the errors by fitting a function appropriately in the training dataset. These functions can be called penalties.



- After Implementing Cross Validation of Elastic Net:



5. **XG Boost Regression: -** It is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XGBoostis one of the ensembles learning methods. XGBoost expects to have the base learners which are uniformly bad at the remainder so that when allthe predictions are combined, bad predictions cancel out and better onesums up to form final good predictions.



Final View Point of all Modals

Model_Name	MAE	MSE	RMSE	MAPE	Rsquare
XGBRegressor	0.030	0.002	0.039	1.956	0.991
LinearRegression	0.151	0.032	0.178	9.543	0.823
Ridge	0.152	0.032	0.179	9.580	0.820
Lasso	0.152	0.032	0.179	9.623	0.820
ElasticNet	0.157	0.036	0.191	10.240	0.796

Conclusion:

- The popularity of stock closing is growing extremely rapidly day byday which encourage the researcher to find new methods if any fraud happens.
- This technique is used for prediction and is not only helpful to researchers to predict future stock closing prices or any fraud happen or not but also helps investors or any person who dealingwith the stock market in order to prediction of model with good accuracy.
- In this work we use the linear regression technique, lasso regression, ridge regression, elastic net regression, and XGBoostRegression technique. these five models give us the following results
- High, low, and open are directly correlate with the closing price ofstocks
- Target variable (dependent variable) is strongly dependent on independent variables
- Xgboost regression is the best model for yes bank stock closingprice data this model used for further prediction