Yes Bank Stock Closing Price Prediction

Technical Document

Submitted by:

Sananda Biswas Chatterjee

Data science trainee

AlmaBetter

Abstract

Predicting the future value of a company's financial stocks is the goal of stock market prediction. Applying "machine learning" to create predictions based on the values of current stock market indices by learning from their past values is a recent development in stock market prediction technology. To make predictions more accurate and simpler, machine learning uses a variety of models. In this case, we tried to predict the stock values by using regression to discover correlations between the features. Stock prices for Open, High, Low, and Close are taken into consideration when building the model and making predictions for the future.

Standard strategic metrics like Root Means Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R2 are used to assess the models. The low values of these indicators demonstrate how accurate the models are at predicting the stock closing prices.

Introduction

The stock market is renowned for its unpredictability, nonlinearity, and dynamic nature. Since stock values depend on a number of variables, including the political environment, the status of the global economy, the firm's financial performance, and others, it can be challenging to predict stock values. As a result, techniques for forecasting stock values by analysing the trend over the last few years may prove to be quite useful for making stock market moves, boosting profit, and minimising losses.

To predict stock values today, sophisticated intelligent methods based on either technical or fundamental analysis are applied. Particularly for stock market analysis, the data is very large and non-linear. To handle this variety of data, an efficient model that can uncover hidden patterns and complex interactions in this enormous data set is necessary.

Machine learning approaches, when compared to older methods, have the potential to unearth patterns and insights we hadn't previously discovered using information like a company's most recent announcements, their quarterly sales numbers, etc. These can be used to get extremely accurate forecasts.

Problem Statement

In the Indian financial industry, YES Bank is a well-known bank. Due to the Rana Kapoor fraud case, it has been in the headlines since 2018. This made it fascinating to investigate how it affected the company's stock prices and whether Time series models or other prediction models may be useful in such circumstances. Since the bank's inception, monthly stock prices have been collected in this dataset. Each month's closing, starting, highest, and lowest stock prices are included. The main goal is to predict the stock's monthly closing price. This dataset covers the closing, opening, highest, and lowest stock prices for each month since the bank's establishment.

Using a variety of machine learning models, we performed regression analysis to create predictions about the future. After comparing the models using evaluation matrices, we chose one as the best model, which we will use to make predictions about the future.

Feature Information

The YES BANK dataset contains closing, starting, maximum, and lowest stock values for each month over the course of 185 observations. It also includes monthly stock prices for the bank since its founding.

Here is the brief discussion about these features --

- **Date:** Monthly observation of stock prices since its inception.
- **Open:** The price of a stock when stock exchange market opens for the day.
- **Close:** The price of a stock when stock exchange market closed for the day.
- **High:** The maximum price of a stock attained during given period of time.
- **Low:** The minimum price of a stock attained during given period of time.

Steps Involved

Importing important libraries and dataset

Our first step is to import libraries. Libraries to assist us investigate the issue and carry out analysis to make judgments based on a set of data.

We are writing our script for this project using Google Collab. We used Yes Bank Stock Closing Price data that is freely available online under the Creative Commons License in order to obtain the information.

Cleaning the dataset

The data must be cleaned up in the following phase. The data we import frequently includes a variety of issues, including missing values, inaccurate data, etc. In order to be used for more thorough analysis, the data quality might be raised through cleaning. There were no duplicate or null values in our dataset, and we updated the data column to the proper format.

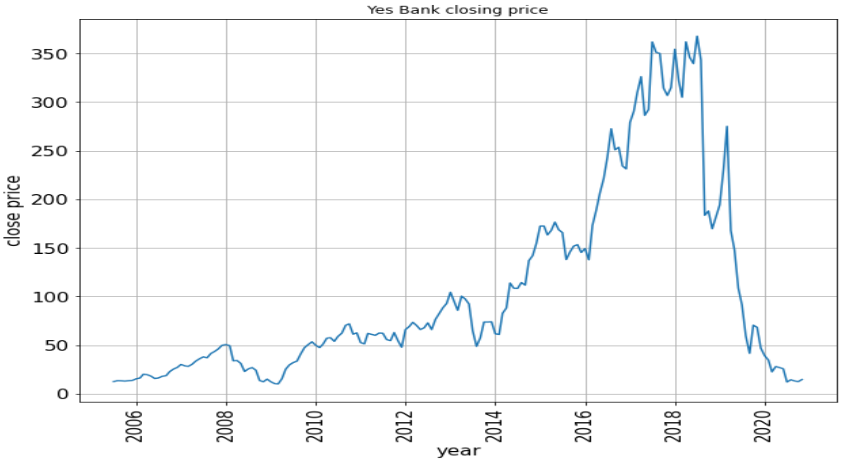
Exploratory Data Analysis and Data Visualization

After loading the dataset, we performed EDA to find the relationship between the data. EDA is a method for data analysis that utilizes visual methods. With the use of statistical summaries and graphical representations, it is used to identify trends, patterns, or to verify assumptions.

After importing the dataset, we used the procedure by contrasting the closing price of the stock, which is our target variable, with the stock's open, low, and high values. We were able to determine many aspects and connections between the target and the independent variables due to this process. We now have a better understanding of how each feature interacts with the target variable.

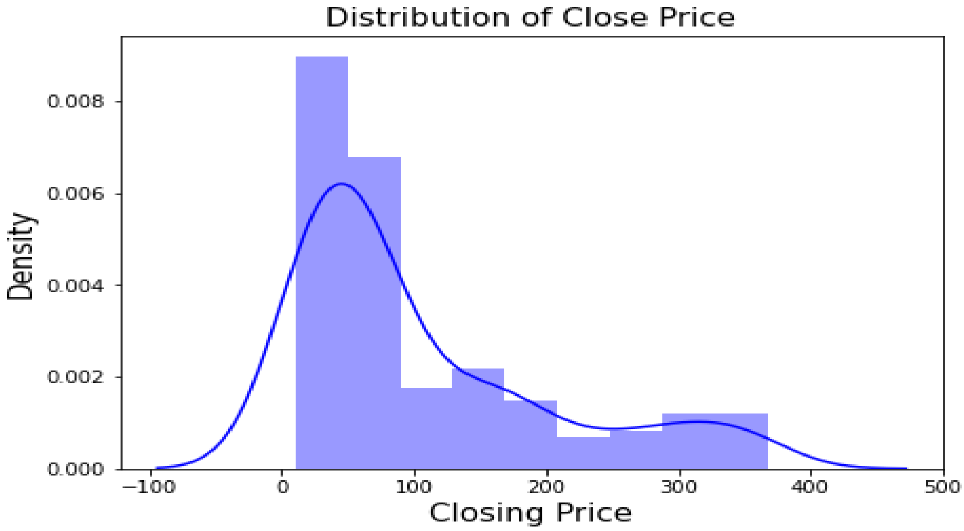
<u>Univariate Analysis</u>

- **Closing price plot**



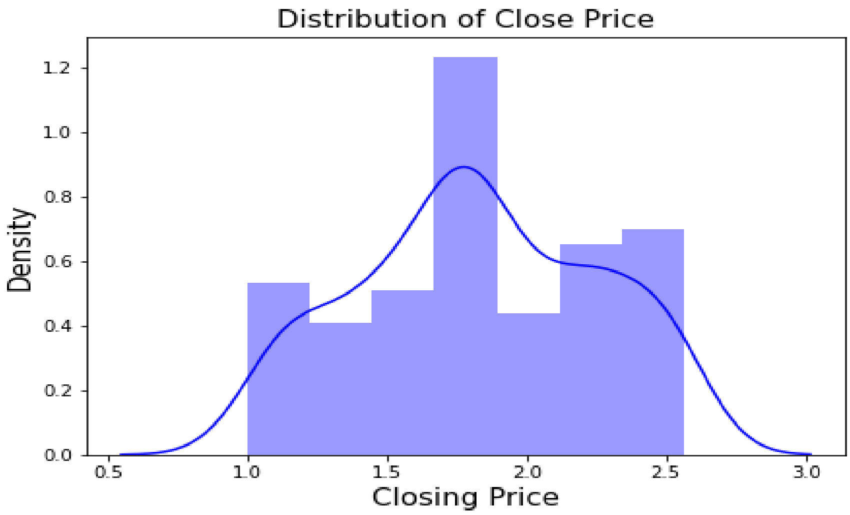
This plot is showing different scenario in different time-duration, we can clearly see that it was continuously increasing from 2009 till 2018. after 2018 there is a sudden fall in the stock closing price due to fraud case of Rana Kapoor.

- **distribution of dependent variable [close price of stock]**



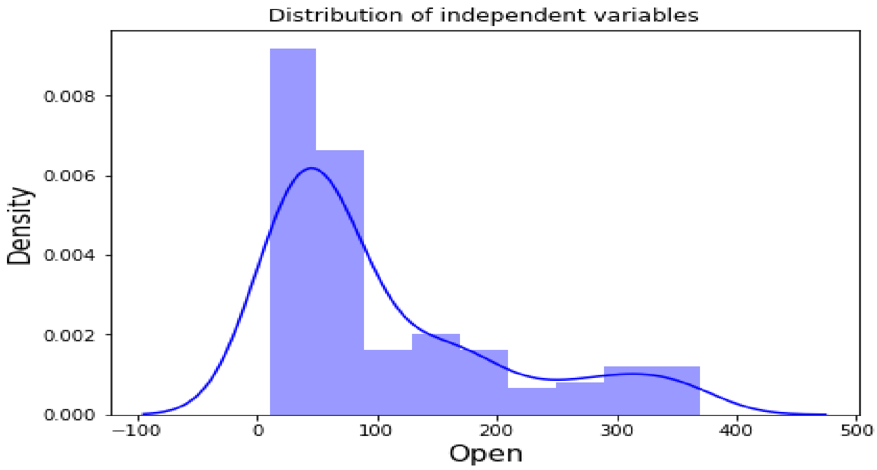
-

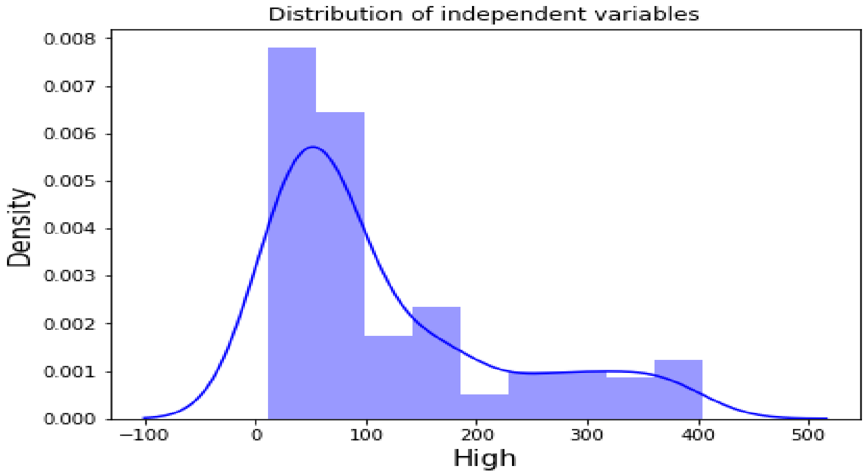
Here we can see that the distribution of stock closing price is rightly skewed distribution

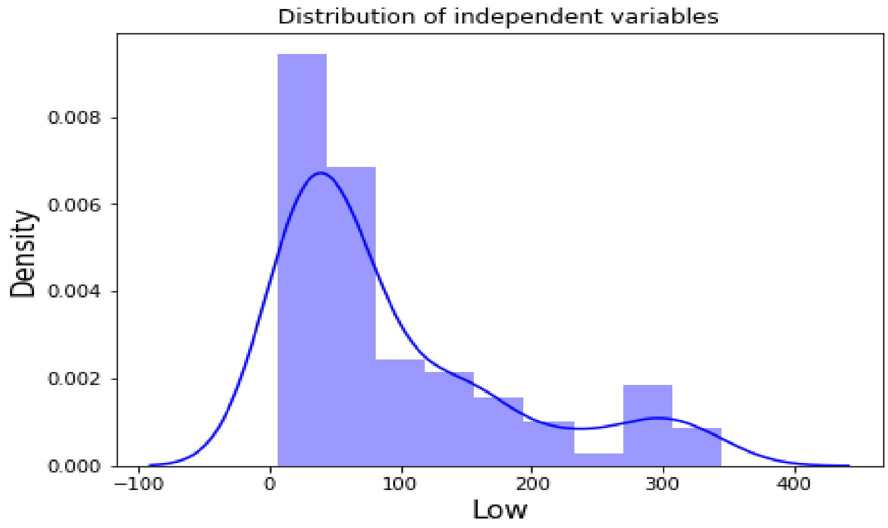


After applying log-transformation it looks like normally distributed.

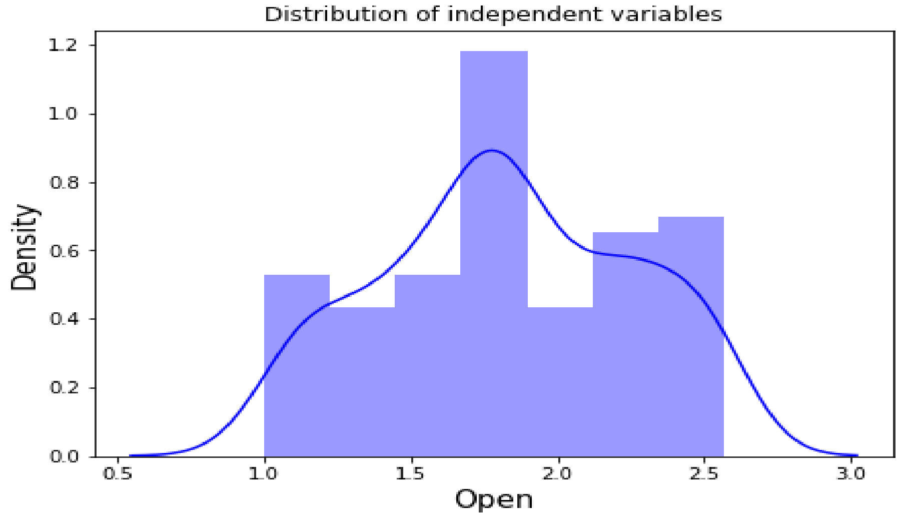
- **Distribution of independent variable [high, low, open]**

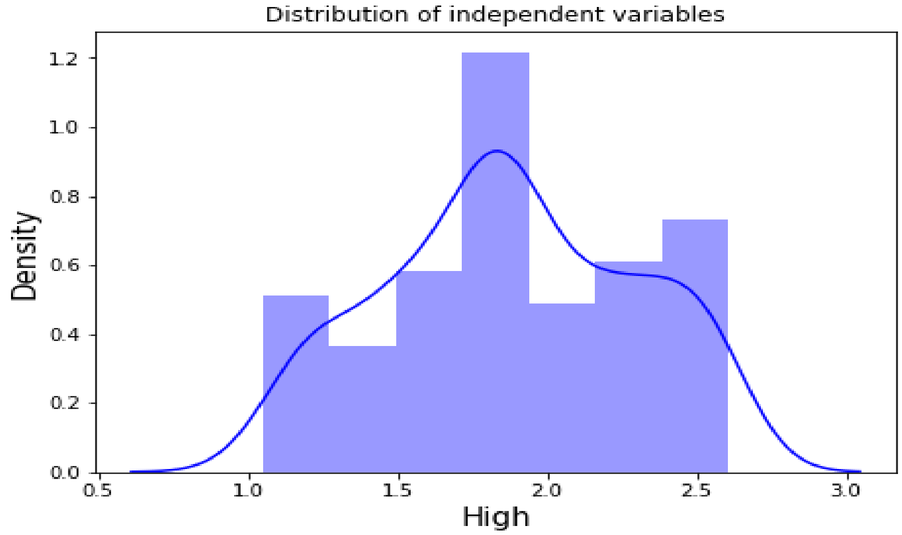


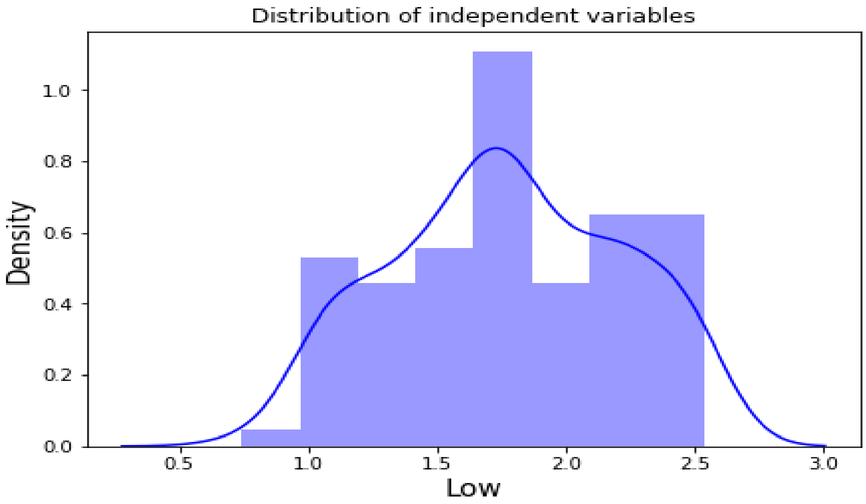




It seems independent variables are rightly skewed. Let’s use log transformation to make it normal





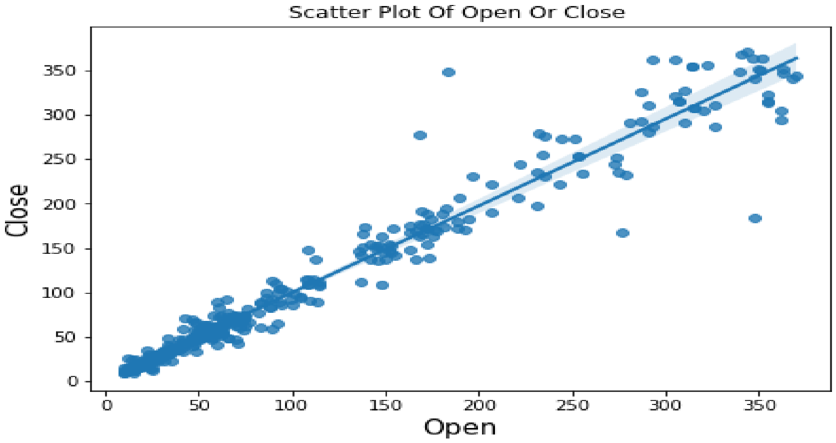


Distribution of independent variable looks normal after applying log transformation

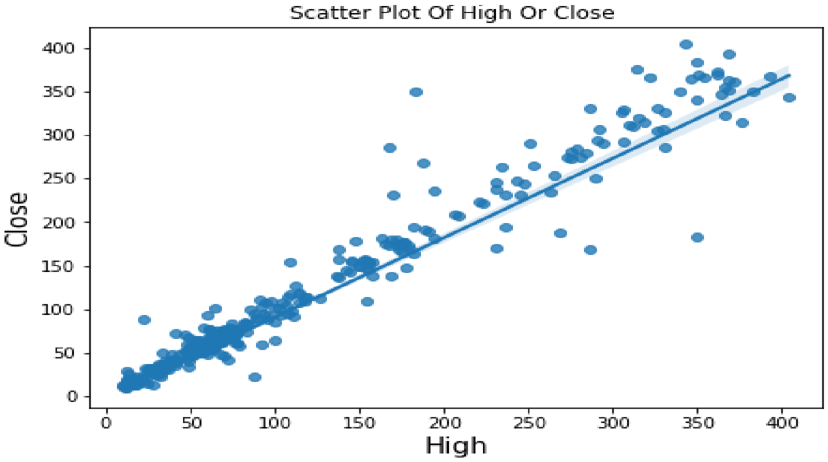
Bivariate Analysis

- **Distribution of dependent and independent variables**

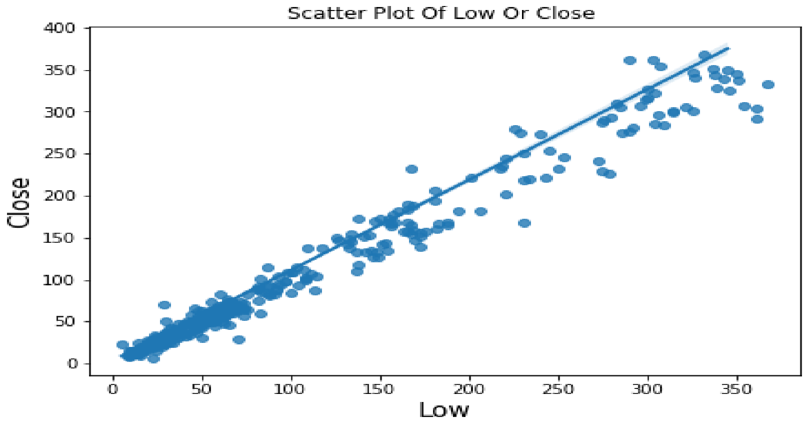
Open Vs Close:



High Vs Close:



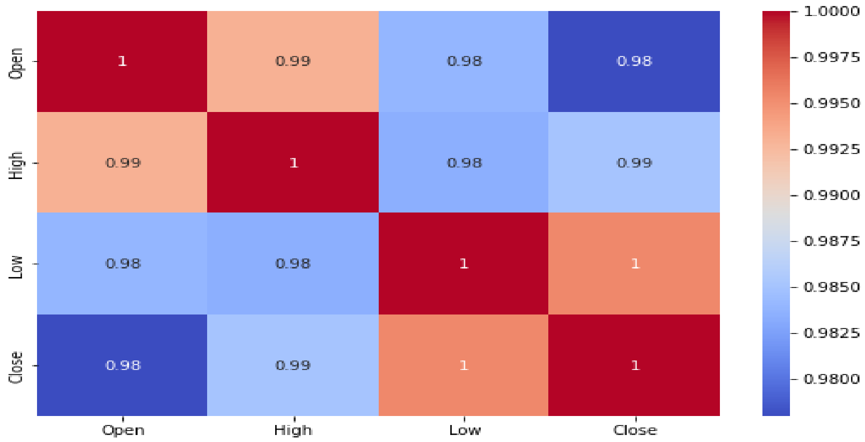
Low Vs Close:



From above visualization it is clear that all independent variables are highly correlated with dependent variable.

- **Correlation**

Correlation measures the relationship between two variables. A positive correlation should be 1. This indicates that the two variables changed in the same direction, either upwards or downwards.



From this plot we can see that there are very high correlation between independent variables which may lead us to multicollinearity. For model fitting and prediction, high multicollinearity is undesirable because even a small change in any independent variable can provide wildly unpredictable results.

Calculating the VIF (Variation Inflation Factor) will allow us to determine the amount of multicollinearity present in our dataset.

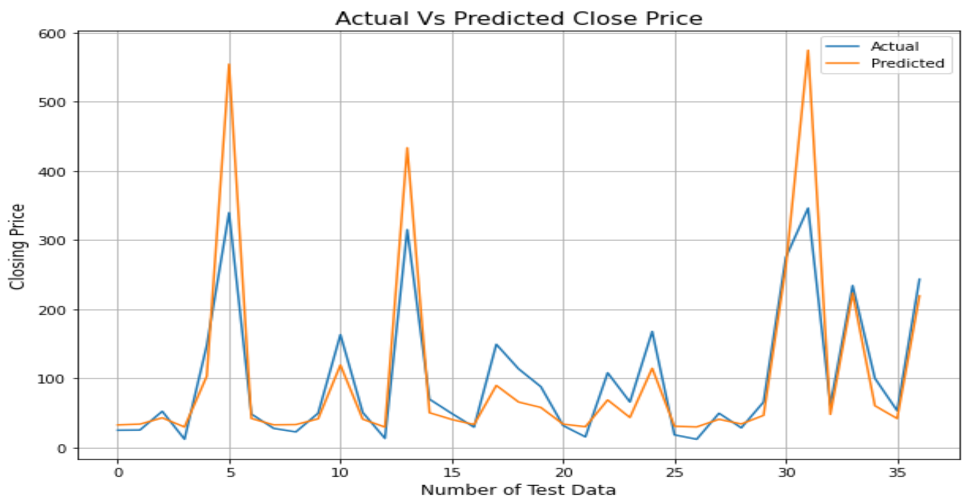
- **Variance Inflation Factor (VIF)**

  A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.
  Any variable with a VIF more than 5 is typically regarded as multicollinearity. The general rule is to drop the variable with the highest VIF, but you can choose the variable to be eliminated depending on business logic. In this case there are few options available and each feature is equally significant here.

Model Building

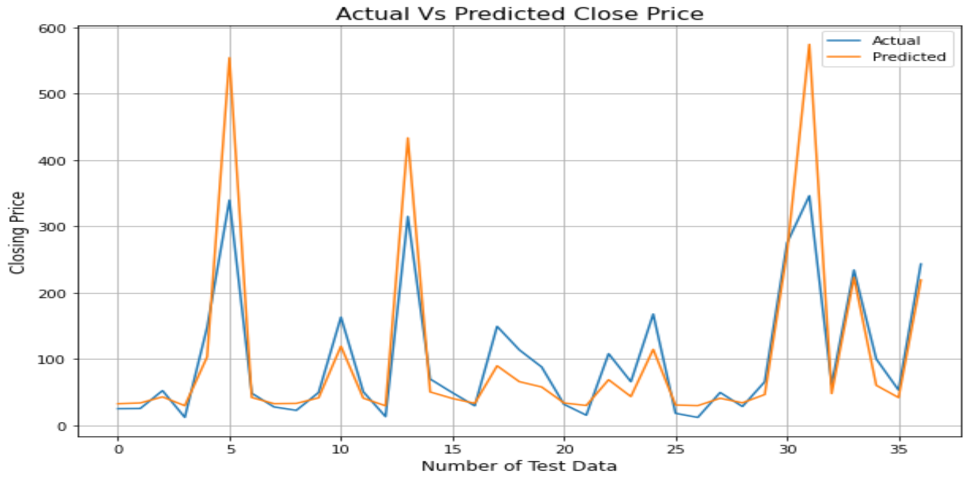
1) Linear Regression

In data science and machine learning, linear regression is a common and simple algorithm. The simplest type of regression, which is a supervised learning technique, is used to examine the mathematical relationship between variables.
The linear regression algorithm, demonstrates a linear relationship between a dependent (y) and one or more independent (y) variables. As a result of displaying a linear relationship, linear regression can be used to determine how the value of the dependent variable changes in response to the value of the independent variable.



2) **Lasso Regression**

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models. 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/parameter elimination.

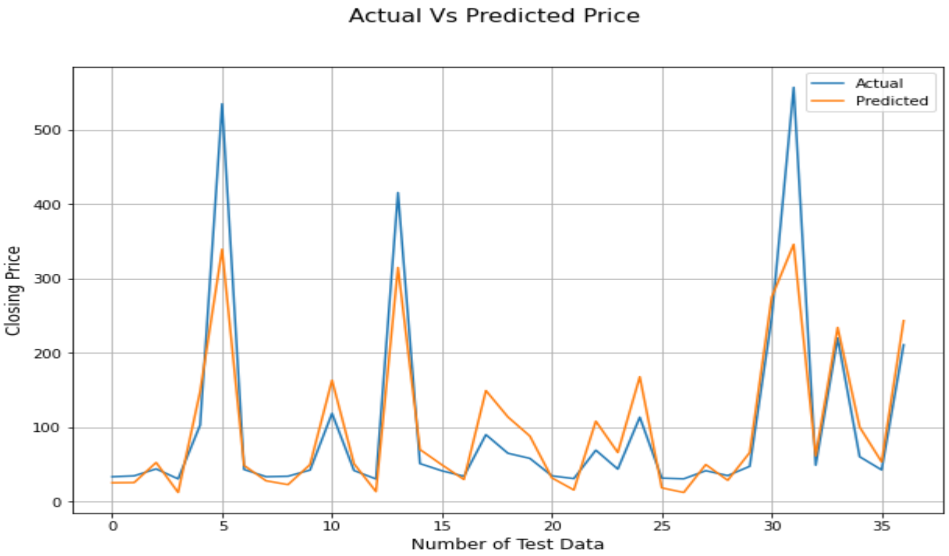


- ## Cross Validation

Cross-validation is a statistic method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

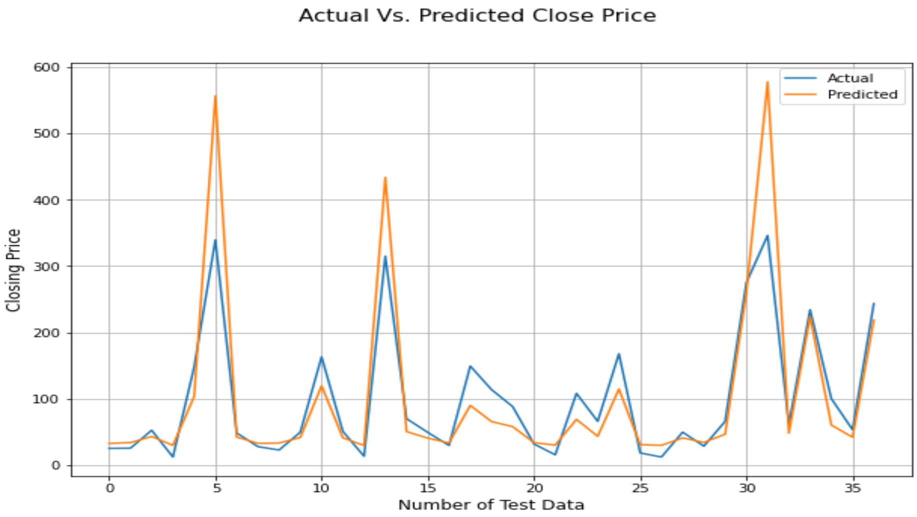
When dealing with a Machine Learning task, you have to properly identify the problem so that you can pick the most suitable algorithm which can give you the best score.

- **Cross Validation on Lasso Regression**

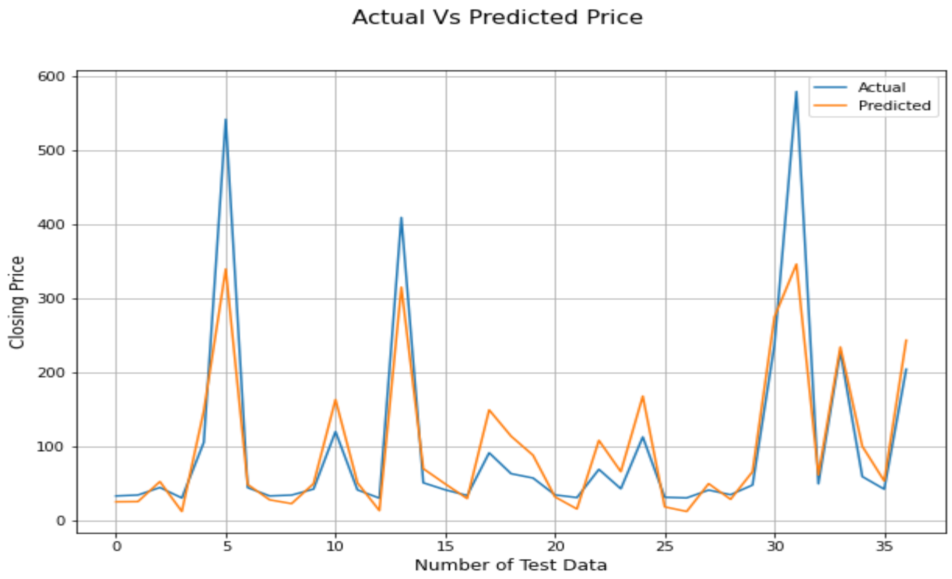


3) Ridge Regression

Ridge regression is the method used for the analysis of multicollinearity in multiple regression data. It is most suitable when a data set contains a higher number of predictor variables than the number of observations. The second-best scenario is when multicollinearity is experienced in a set. Multicollinearity happens when predictor variables exhibit a correlation among themselves. Ridge regression aims at reducing the standard error by adding some bias in the estimates of the regression. The reduction of the standard error in regression estimates significantly increases the reliability of the estimates.



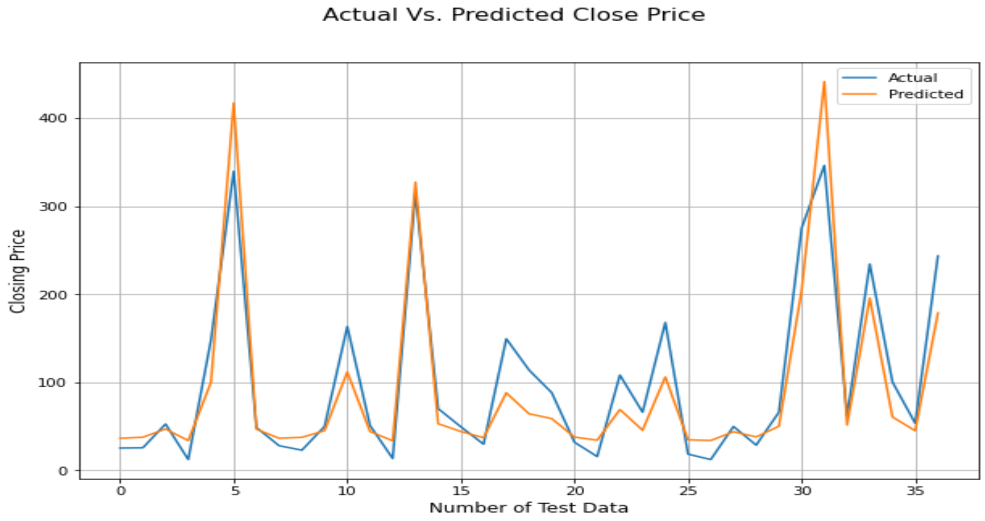
Cross validation on Ridge regression



4) Elastic Net Regression

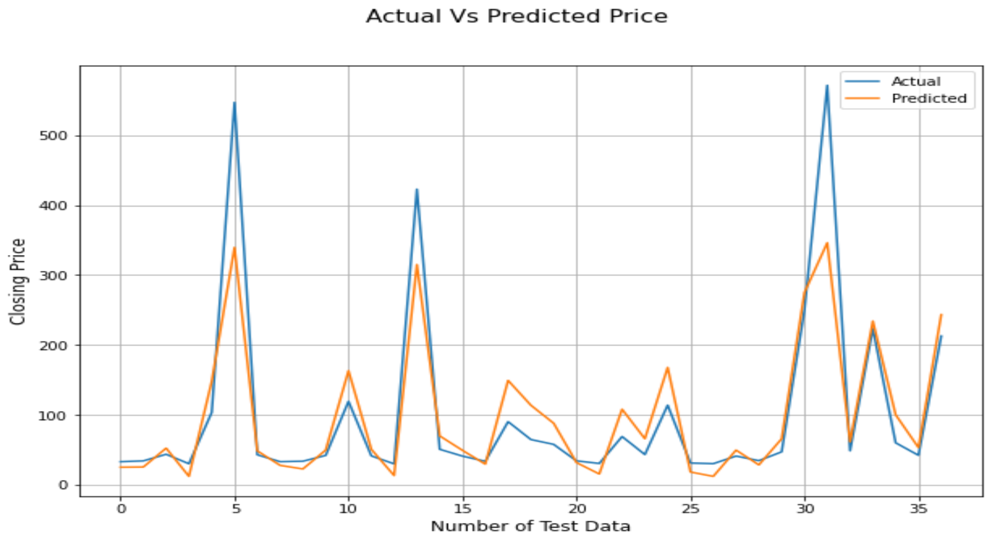
Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

The elastic net method performs variable selection and regularization simultaneously. And this technique is most appropriate where the dimensional data is greater than the number of samples used.



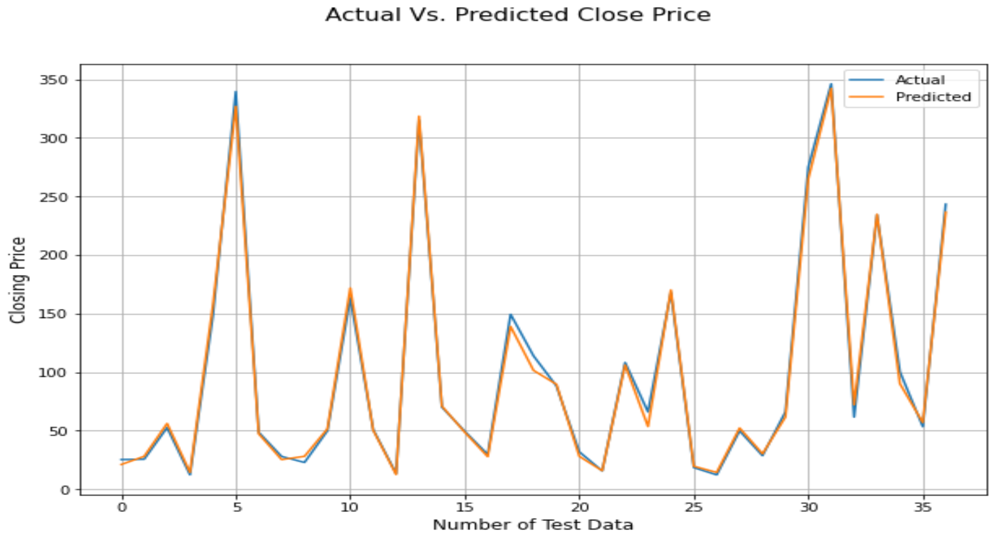
Cross validation on Elastic Net regression

- **Cross Validation on Elastic Net Regression**



5) XG Boost Regression

XG Boost is essentially same thing as gradient boosted trees algorithm. Gradient booting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler models. But the main difference here is how the residual trees are built. With XG Boost, the residual trees are built by calculating similarity scores between leaves and the preceding nodes to determine which variables are used as the roots and the nodes.



Conclusion

- The Close, Open, High, and Low values of Yes Bank's stock all tended to rise until 2018 before abruptly falling after Rana Kapoor's fraud case.
- Independent variables have a significant impact on the target variable (dependent variable).
- The relationship between the dependent and independent variables is linear.
- For linear, lasso, and ridge regressions, the R squared values are quite similar.
- The best model is XG Boost Regression, which has the highest r2 score (0.9913), the lowest RMSE (0.0394), and MAPE (0.0196) of the five models.