Capstone Project

Yes Bank Stock Closing Price Prediction

Done by:

Sananda Biswas Chatterjee

&

Amit Kundu

CONTENTS

INTRODUCTION

We have a dataset that belongs to the Yes Bank monthly stock prices from the months of July 2005 to November 2020. Due to Rana Kapoor's recent instance of default fraud, the bank has recently been in the news. We should be able to predict the closing price of the bank's stock using the dataset analysis and other factors.

The goal of this study is to test a variety of models to evaluate whether stock prices and movements can be predicted using features and historical data using regression technique. We must first understand the correlations between the various components of the dataset in order to provide the model with the appropriate parameters for training which eventually predict the closing price.

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.

FEATURE INFORMATION

The YES BANK dataset contains closing, opening, 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 it was listed.
- **Open**: The price of a stock when stock exchange market opens for the day.
- **Close**: The price of a stock when stock exchange market closes 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.

STEP TAKEN

Importing libraries and dataset

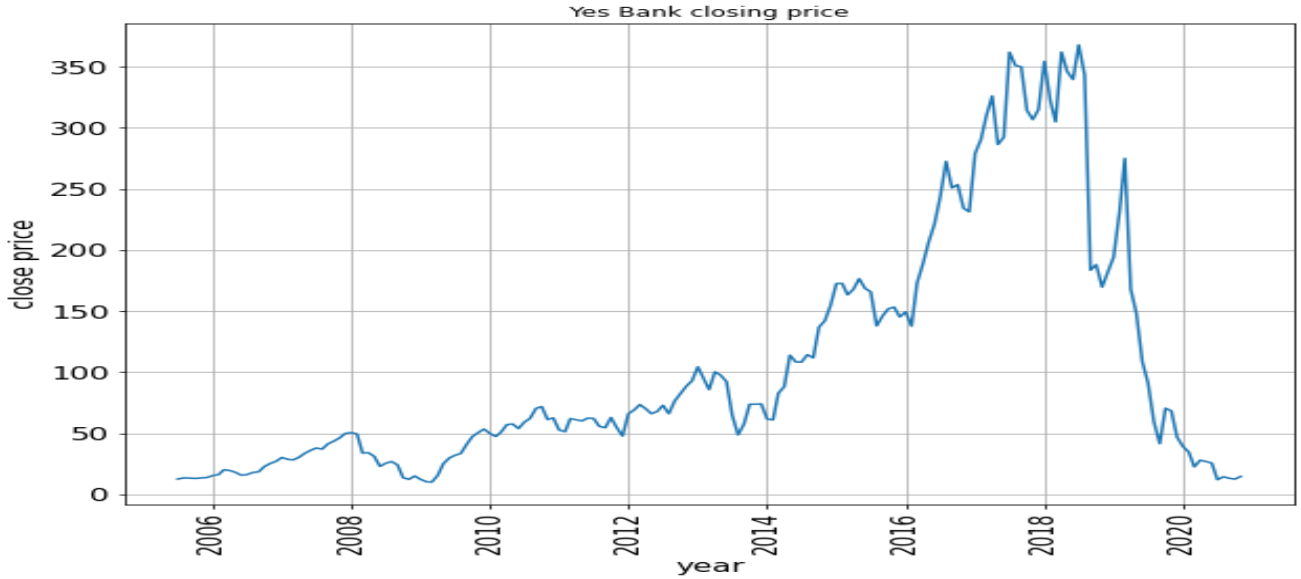
- Our first step is to import libraries to assist us in investigating the issues and carry out analysis to make judgments based on a set of data.
- We are writing our script for this project using Google Collab. We used Yes Bank Stock Closing Price data that is freely available online under the Creative Commons License in order to obtain the information.

STEP TAKEN

Cleaning the dataset

- ✓ The data we imported frequently includes a variety of issues, including missing and duplicate values, inaccurate data, etc.
- ✓ In order to be used for more thorough analysis, the data quality must be raised through cleaning.
- ✓ After cleaning the data we found no duplicate or null values in our dataset.
- ✓ We checked the outliers.
- ✓ We updated the date column to the proper format(i.e from Jul-05 to 2005-07-01).

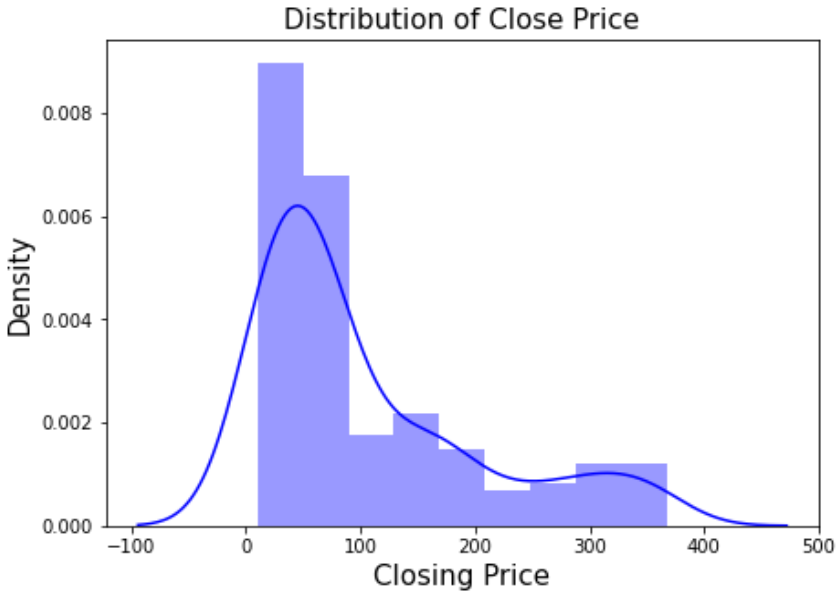
EXPLORATORY DATA ANALYSIS(EDA)



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.

Univariate Analysis

Distribution of dependent variable

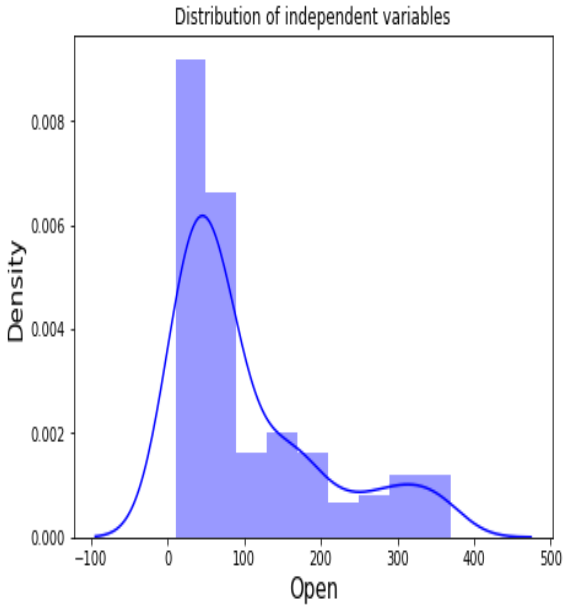


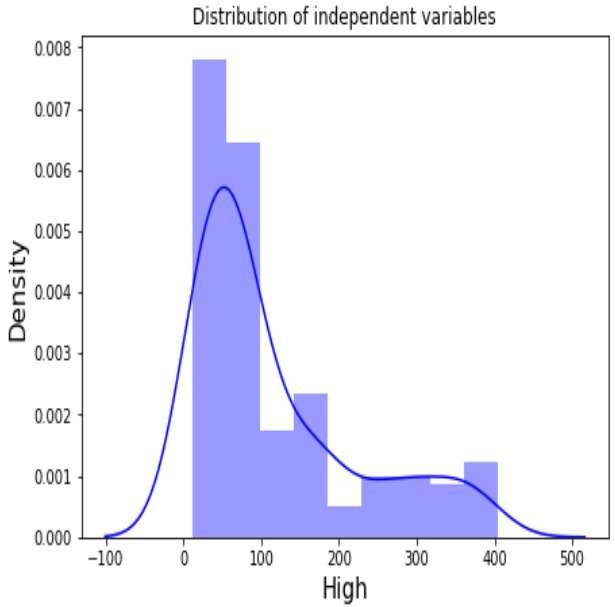
This chart makes it obvious that the distribution of stock closing prices is rightly skewed.

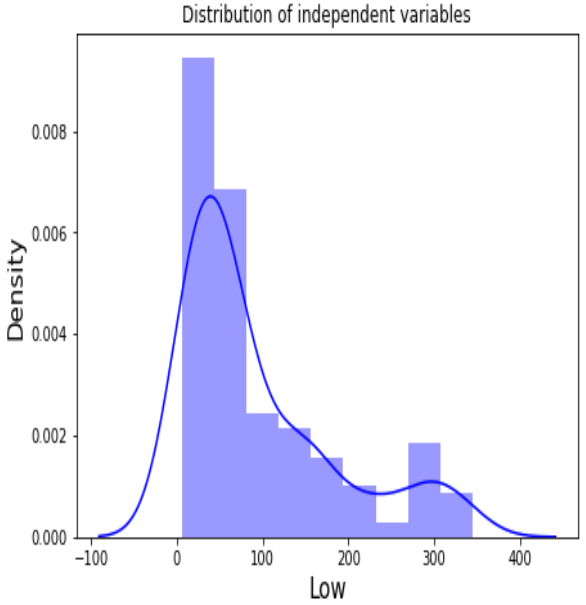
After Applying log transformation

To make it a normal distribution, we converted it using log transformation.

Distribution of Open, High & Low Price of a stock

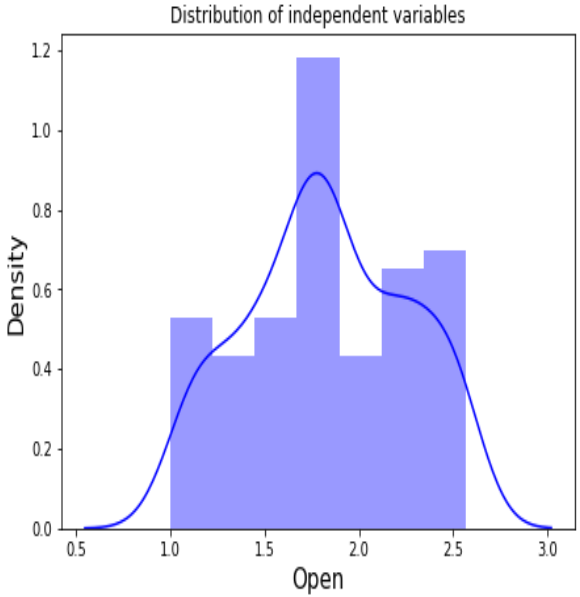


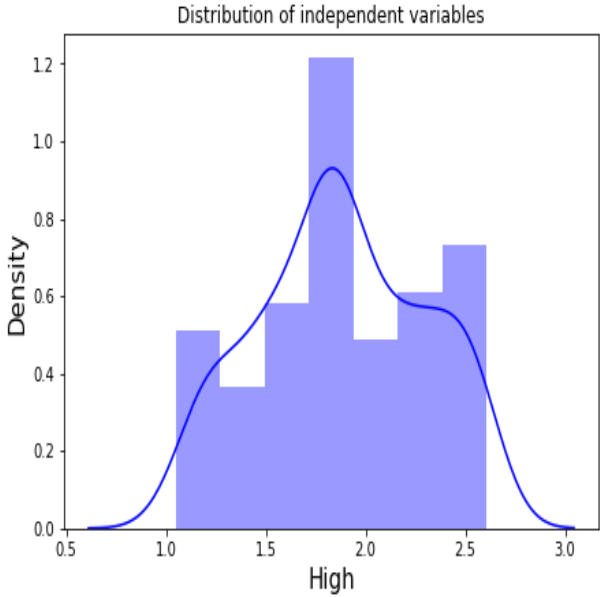


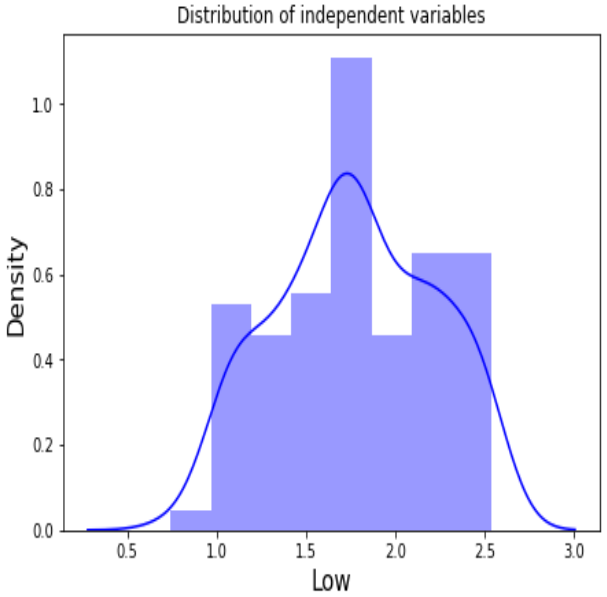


Opening price, high price, and low price distribution are all right skewed.

Distribution of Open, High & Low Price of a stock after Log Transformation

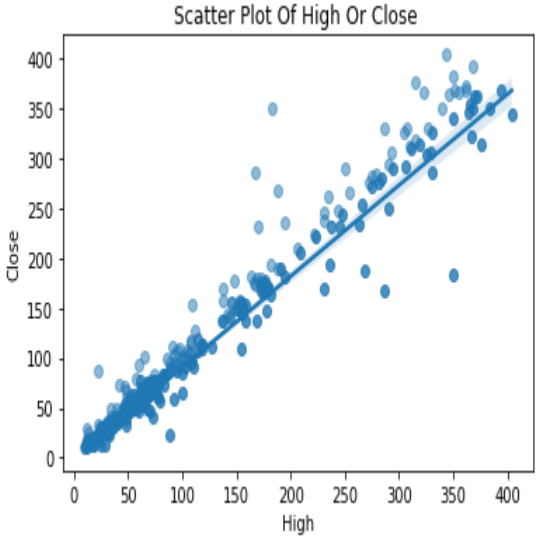


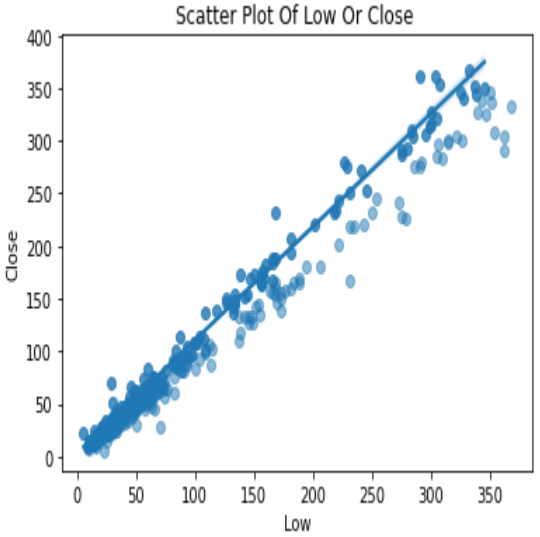




The log transformation was applied to all independent variable distributions to convert them to a nearly normal distribution.

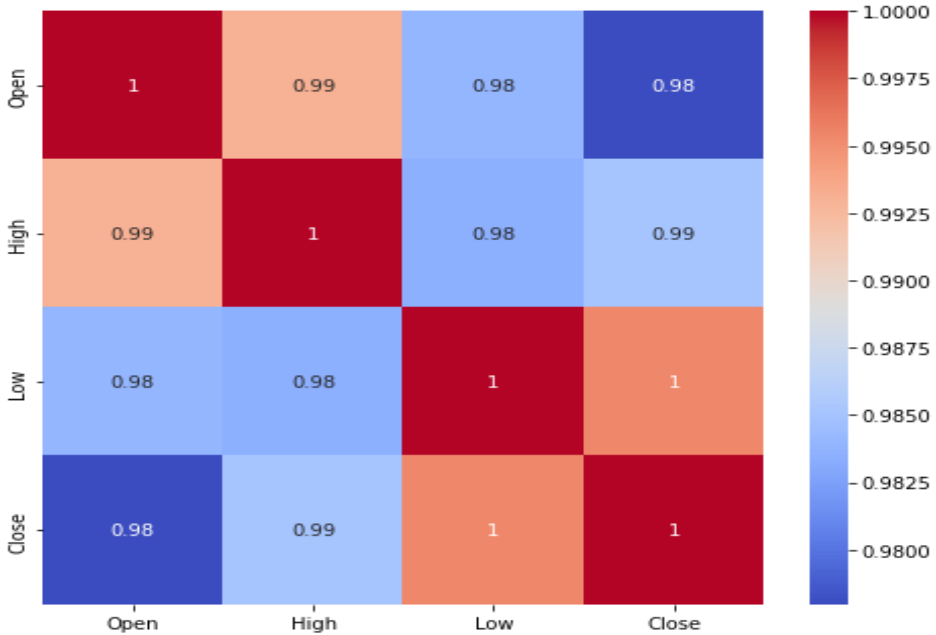
Bivariate Analysis Plots





The relationship between the dependent and independent variables can be observed to be highly correlated through the use of a scatter plot.

Correlation Between the Variables

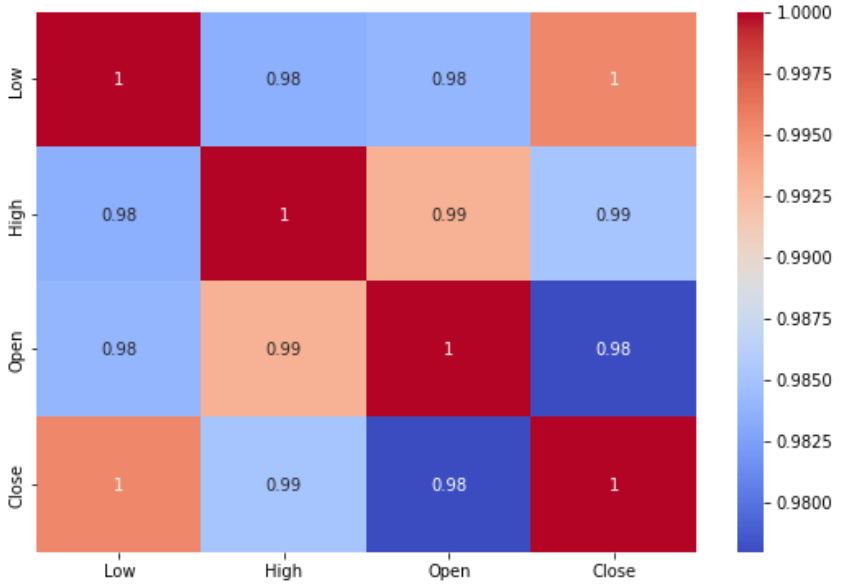


- The high correlation between the independent variables as found from above result, causes multicollinearity.
- For model fitting and prediction, high multicollinearity is undesirable because even a small change in any independent variable can produce wildly unpredictable results.

Correlation And VIF Analysis

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.

	Variables	VIF
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

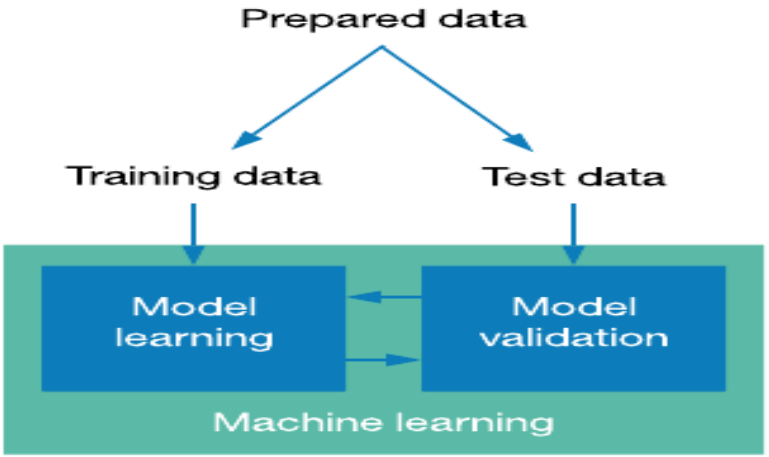


Any variable with a VIF more than 5 is typically regarded as multicollinear. The general rule is to drop the variable with the highest VIF, but you can choose the variable to be eliminated depending on business logic. In this case all the feature are equally significant here.

MODEL BUILDING

Splitting data in Train and Test set

- Data splits into training dataset and testing dataset.
- Training dataset is for making algorithm learn and train model.
- Test dataset is for testing the performance of train model.
- Here 80% of data taken as training dataset & remaining 20% of dataset used for testing purpose.



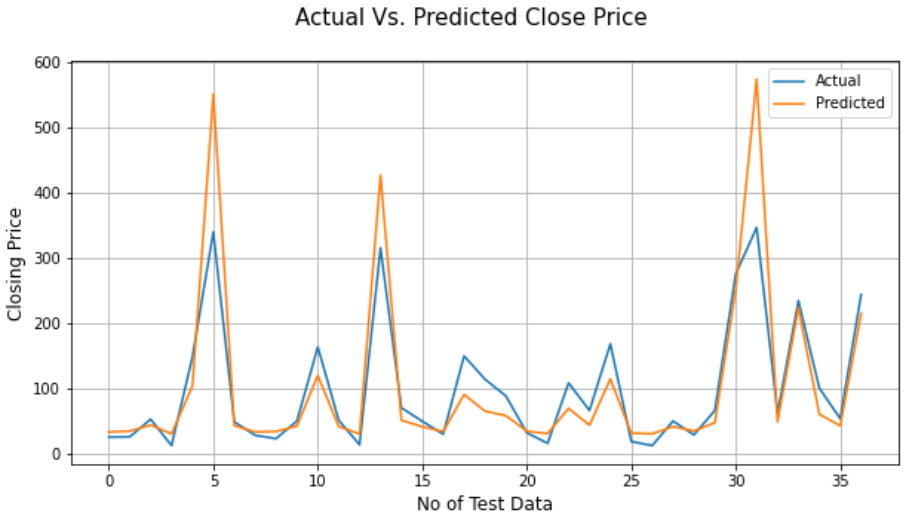
Linear Regression

- One of the simplest and most widely used Machine Learning techniques is linear regression.
- Linear regression is a statistical method that is used for predictive analysis. It is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.

EVALUATION METRICS				
MSE	RMSE	MAE	MAPE	R2 SCORE
0.0316	0.1777	0.1513	0.0954	0.8226

Lasso Regression

- Lasso regression is linear regression, but it uses a technique called **"shrinkage"** where the coefficients of determination shrink towards **zero**.
- It allows to shrink or regularize coefficients to avoid overfitting and make them work better on different datasets.
- This type of regression is used when the dataset shows high multicollinearity or when you want to automate variable elimination and **feature selection.** This method performs L1 regularization.



EVALUATION METRICS				
MSE	**RMSE**	**MAE**	**MAPE**	**R2 SCORE**
0.032	**0.179**	**0.1523**	**0.0962**	**0.82**

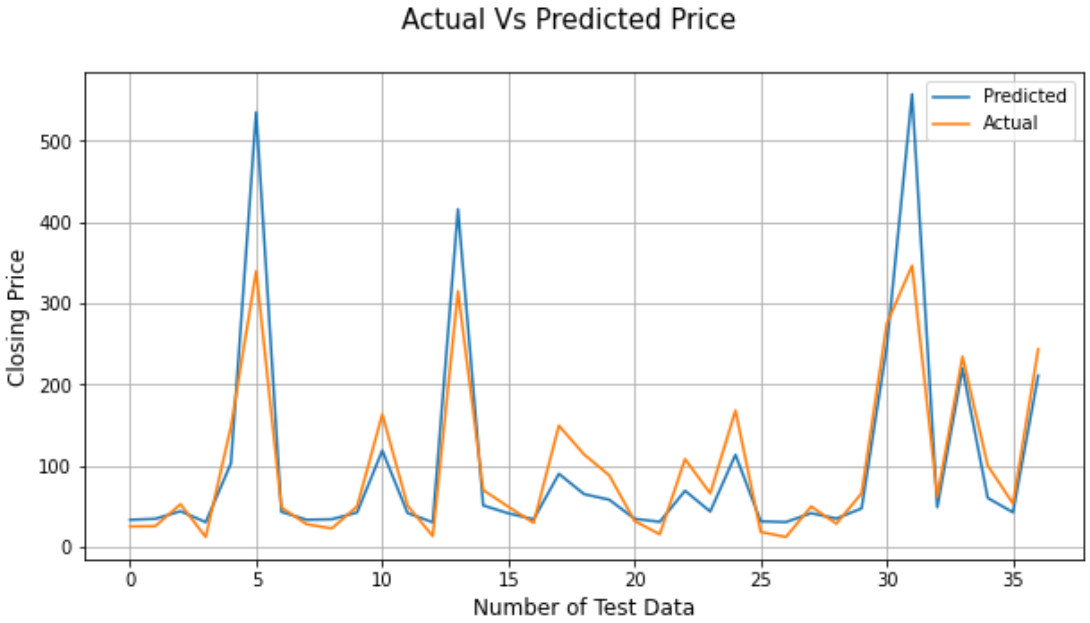
Cross Validation

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, we make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

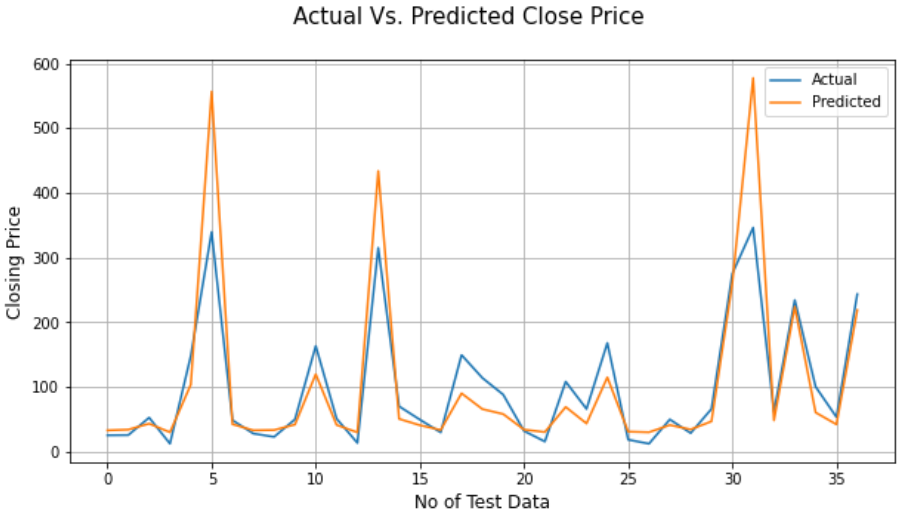
Lasso Regression(After Cross Validation)



EVALUATION METRICS				
MSE	RMSE	MAE	MAPE	R2 SCORE
0.0322	0.1795	0.1528	0.0968	0.819

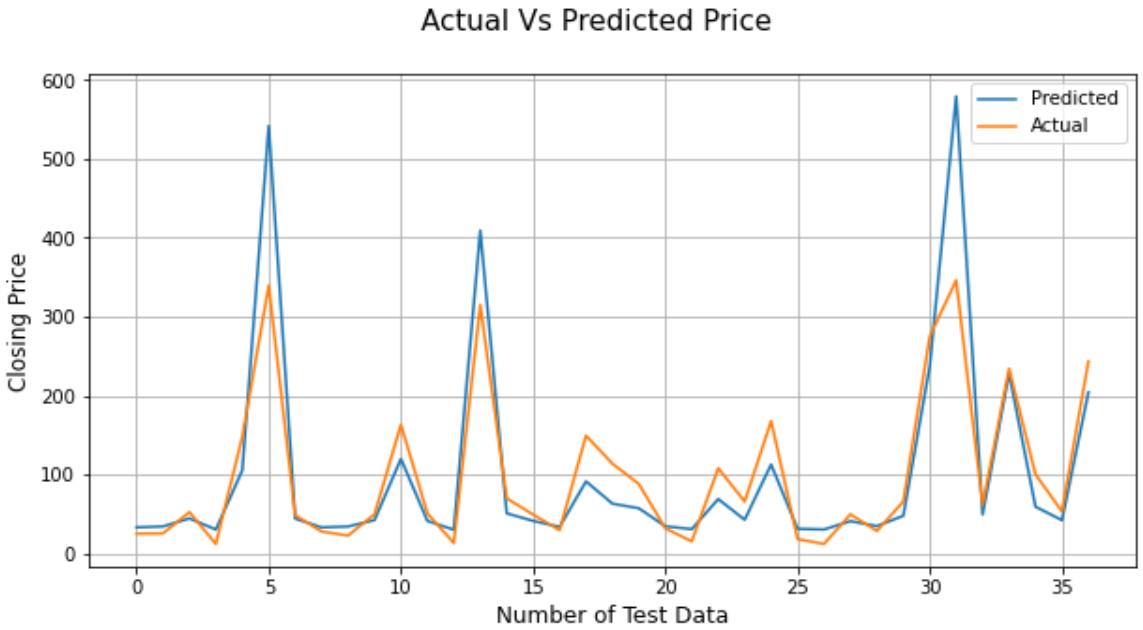
Ridge Regression

- Any data that exhibits multicollinearity can be analysed using the model tuning technique known as ridge regression.
- This technique carries out L2 regularisation.
- Predicted values differ much from real values when the problem of multicollinearity arises, least-squares are unbiased, and variances are significant.



EVALUATION METRICS				
MSE	RMSE	MAE	MAPE	R2 SCORE
0.0317	0.1779	0.1514	0.0955	0.8221

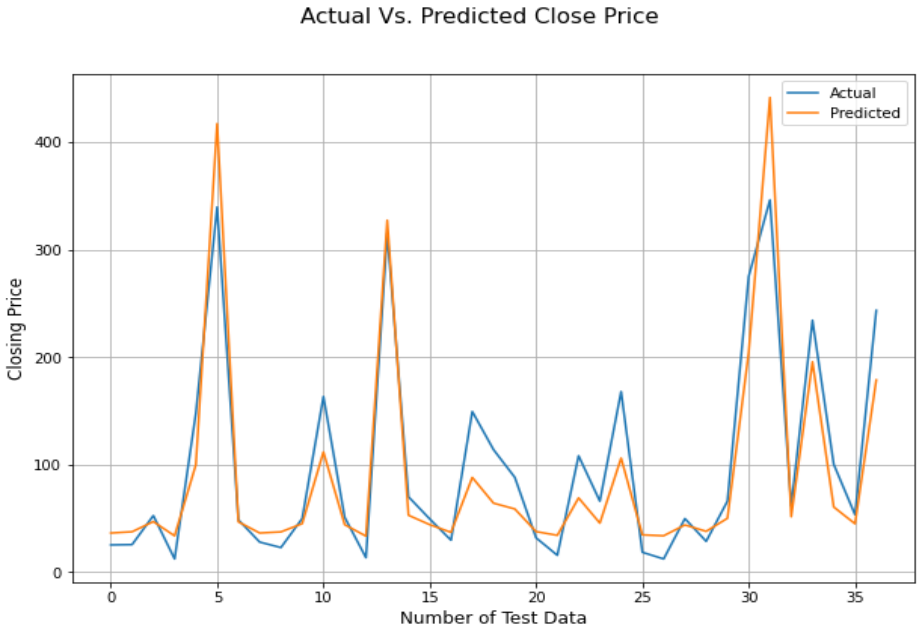
Ridge Regression(After Cross Validation)



EVALUATION METRICS				
MSE	RMSE	MAE	MAPE	R2 SCORE
0.0325	0.1804	0.1531	0.0968	0.8172

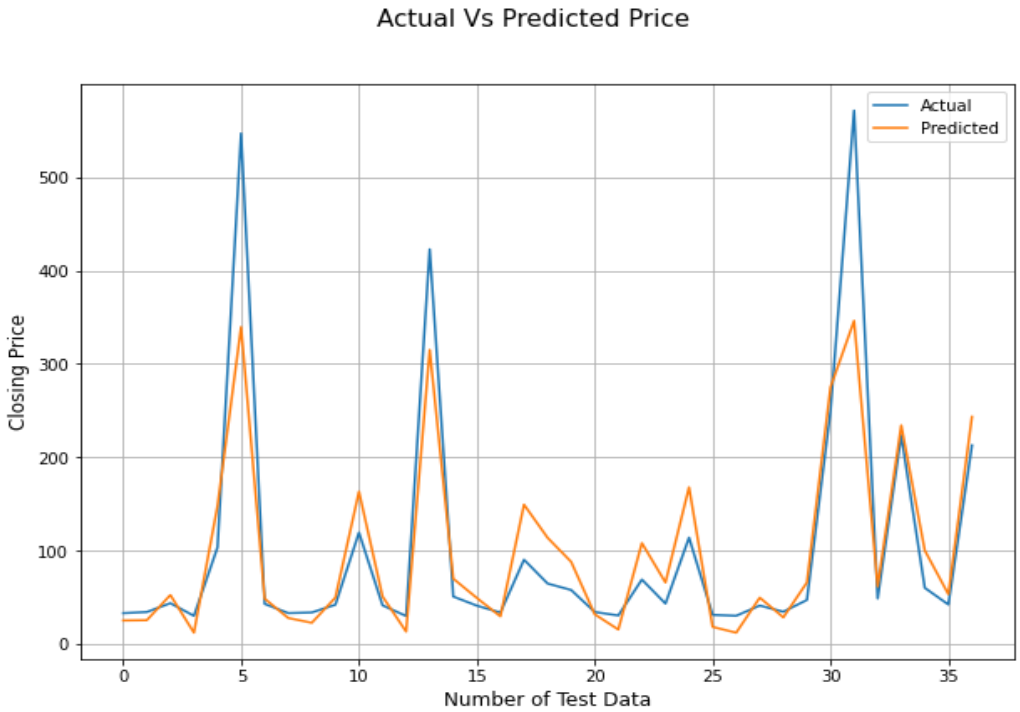
Elastic Net Regression

- 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..



EVALUATION METRICS				
MSE	RMSE	MAE	MAPE	R2 SCORE
0.0364	0.1908	0.1574	0.1024	0.7955

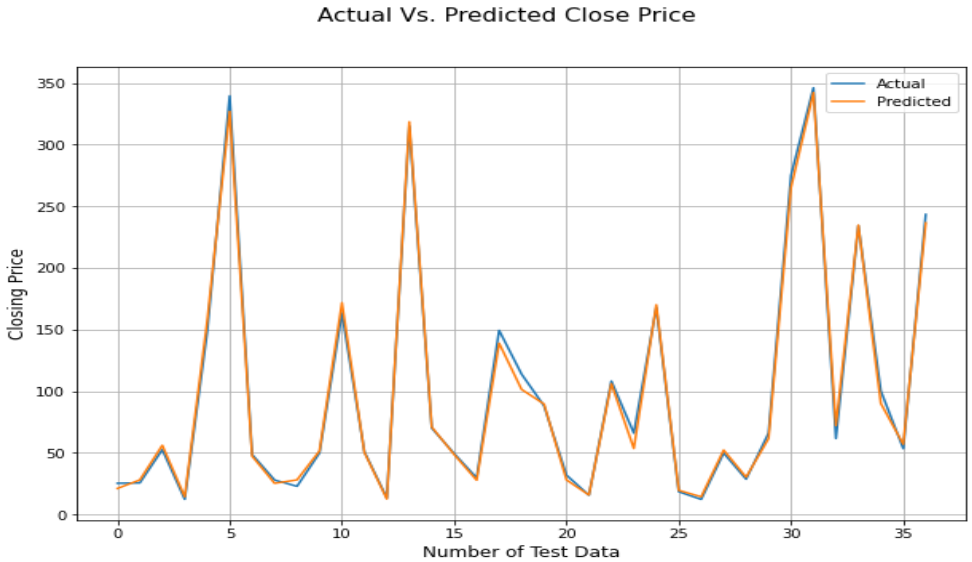
Elastic Net Regression(After Cross Validation)



EVALUATION METRICS				
MSE	RMSE	MAE	MAPE	R2 SCORE
0.0321	0.1791	0.1524	0.0963	0.8198

XG Boost Regression

- XG Boost stands for “Extreme Gradient Boosting”. XG Boost is a powerful technique for building supervised regression models.
- One can determine the accuracy of this statements (XG Boost) by understanding its objective function and base learners.
- The objective function includes a regularisation term and a loss function. It offers details on how far the model's predictions deviate from the actual values, or the difference between actual and predicted values.



EVALUATION METRICS				
MSE	RMSE	MAE	MAPE	R2 SCORE
0.0016	0.0394	0.0303	0.0196	0.9913

CONCLUSION

1.The tendency of Yes Bank's stock's Close, Open, High, and Low prices increased until 2018 and then unexpectedly decreased after fraud case of Rana Kapoor.

2. The target variable is highly dependent on input variables.

3. Each independent variable has a strong correlation with the others (Multicollinearity).

4. The R squared values for linear, lasso, and ridge regressions are nearly identical.

5. On the basis of RMSE(Root Mean Square Error) and MAPE(Mean Absolute Percentage Error), we compared 5 models (Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, and XG Boost Regression).

6. The dependent variable (Closing Price) and the independent variables (High, Low, and Open) have a direct correlation.

7. With the lowest RMSE (0.0394) and MAPE (0.0196) as well as the highest R2 score (0.9913), XG-Boost Regression is the best model.

THANK YOU