Capstone Project

Yes Bank Stock Price Prediction
Technical Documentation

Jyoti Patel
Agam Singh
Data science trainees
AlmaBetter

Table of Content:-

Abstract :

Yes Bank is a banking company 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 publicly 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 company's valuation is now in the hands of investors and speculators as share prices are often heavily impacted by public opinion.

We have used the Yes bank stock price data set. This dataset contains 5 features that can be used for close price prediction using machine learning. We have built a machine-learning regression model for price prediction. We have used some of the best models.

Introduction:

YES bank stands for Youth Enterprise Scheme Bank. The 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. A stock market is a public market where you can buy and sell shares for publicly listed companies. Stock Price Prediction using machine learning helps you get an estimate of the value of company stock going forward and other financial assets traded on an exchange. 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.

Data Description:

This dataset has 185 observations in it with 5 columns and it is a mix between categorical and numeric values.

● **Date:** This variable consists of categorical values and each value represents 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. This variable consists of numerical values.
● **High:** High refers to the maximum prices in a given time period.
● **Low:** Low refers 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.

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 for the month.

Problems we are dealing with in the project :

- The accuracy of predicting Yes Bank's stock price using traditional linear regression models may be limited, considering the complex and dynamic nature of financial markets.
- Multicollinearity among the independent variables in the dataset might hinder the accuracy of the linear regression models, requiring the implementation of regularization techniques like Lasso, Ridge, and Elastic Regression.
- The selection of the optimal set of features that have the most significant impact on Yes Bank's stock price prediction is crucial to improve the model's accuracy. This necessitates the use of feature selection methods like Lasso Regression.
- The fluctuating and volatile nature of stock prices makes it challenging to capture and predict the true underlying trends and patterns. The project needs to address this issue to enhance the accuracy of the stock price predictions.
- The availability and quality of historical data, financial indicators, and market data might pose a challenge, as they could be incomplete, unreliable, or require significant preprocessing to ensure accurate and meaningful analysis.
- The project requires efficient hyperparameter tuning for the regularization techniques (Lasso, Ridge, and Elastic Regression) to find the optimal balance between feature selection and addressing multicollinearity, thereby improving the accuracy of the stock price predictions.

- External factors, such as macroeconomic indicators, regulatory changes, or unforeseen events, may significantly impact stock prices, making it difficult to capture and incorporate these factors into the prediction models.
- The evaluation and comparison of the regression techniques need to be carefully conducted using appropriate performance metrics to ensure a fair and comprehensive assessment of their effectiveness in predicting Yes Bank's stock prices.

Steps involved:

Exploratory Data Analysis: After loading the dataset we performed this method by comparing our target variable which is Close column values with other independent variables. This process helped us figure out various aspects and relationships between the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

Data Exploration: Exploring the dataset to understand the structure of the data, identify any patterns or trends, and detect outliers.

Feature Engineering: Creating new features from the existing data to gain more insights and improve the predictive power of the model.

Data Visualization: Creating various charts and graphs to visualize the data and communicate the insights gained from the analysis.

Standardization of features: Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it. The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

Modeling: To evaluate our test data, we employed several regression models, including Linear Regression, Lasso Regression with cross-validation, Ridge Regression with cross-validation, and Elastic Net Regression with cross-validation. By leveraging these models, we aimed

to obtain comprehensive insights and make accurate predictions for our test dataset.

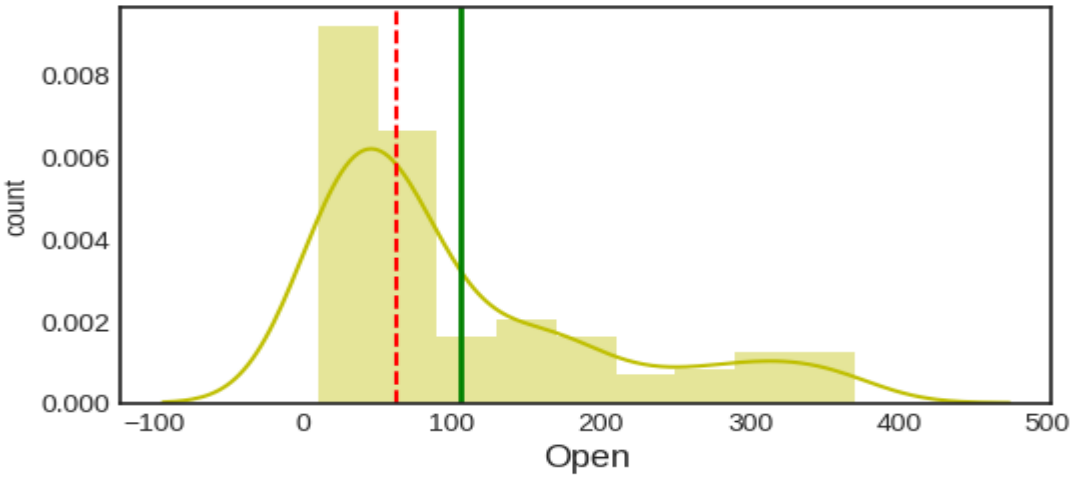
Exploratory Data Analysis:

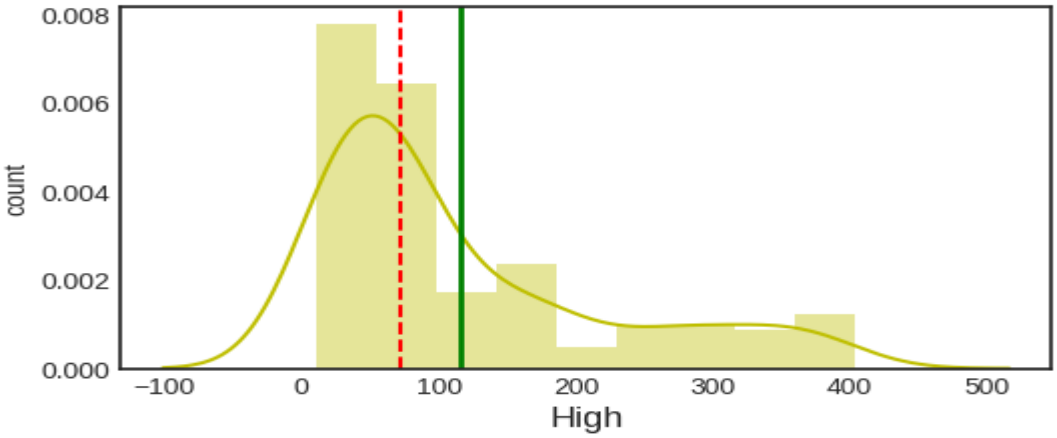
A) Data Cleaning: The Given Date in the data is in Month-year format (mmm-yy) and is converted to the proper date of YYYY-MM-DD and the given date column has dtype as an object converting it into date time format.

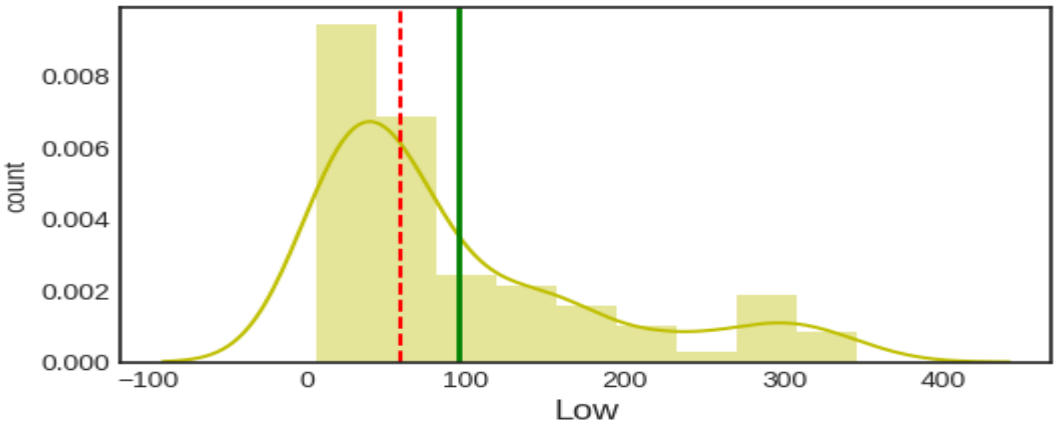
B) Null values Treatment: Our dataset does not contain null values which tend to affect our accuracy. If we had null values, we could drop them or input them with mean or median depending on the situation.

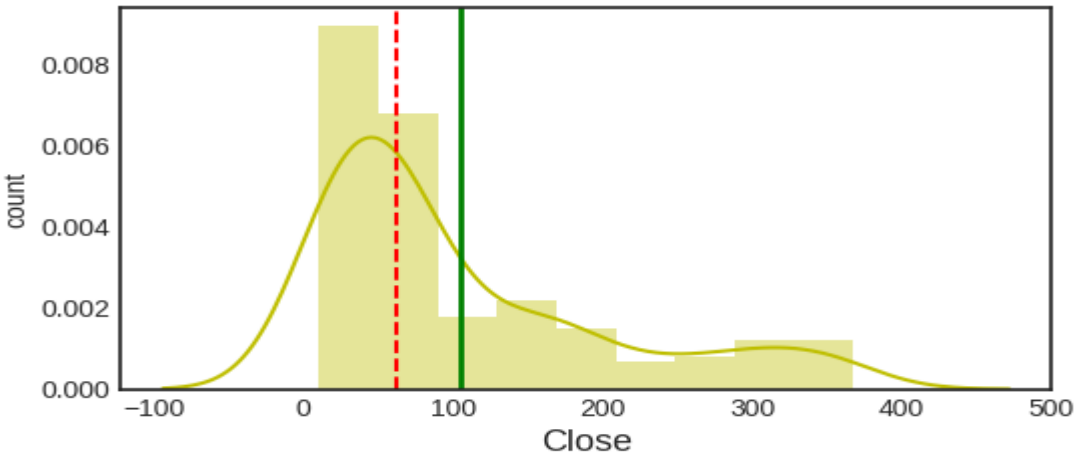
C) Data Visualization:

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





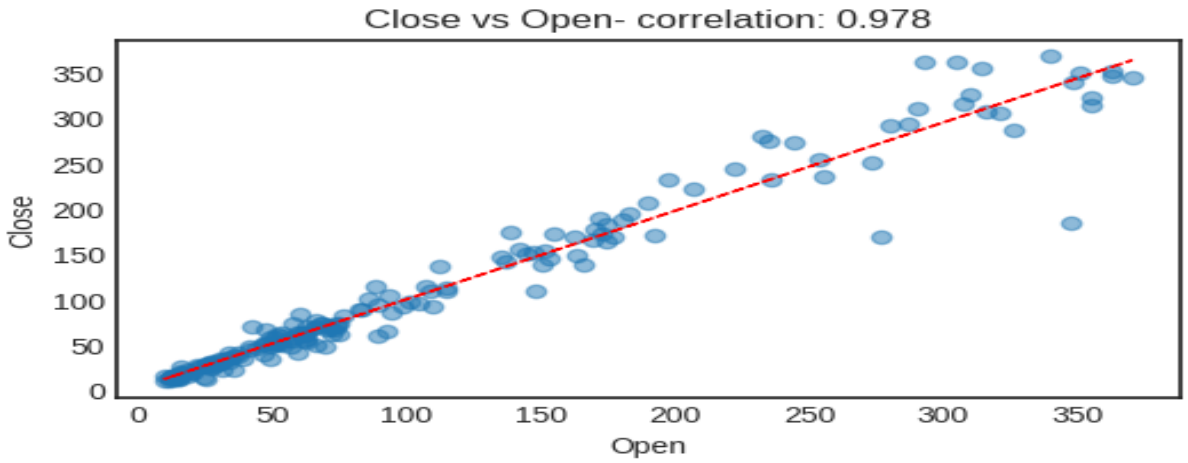


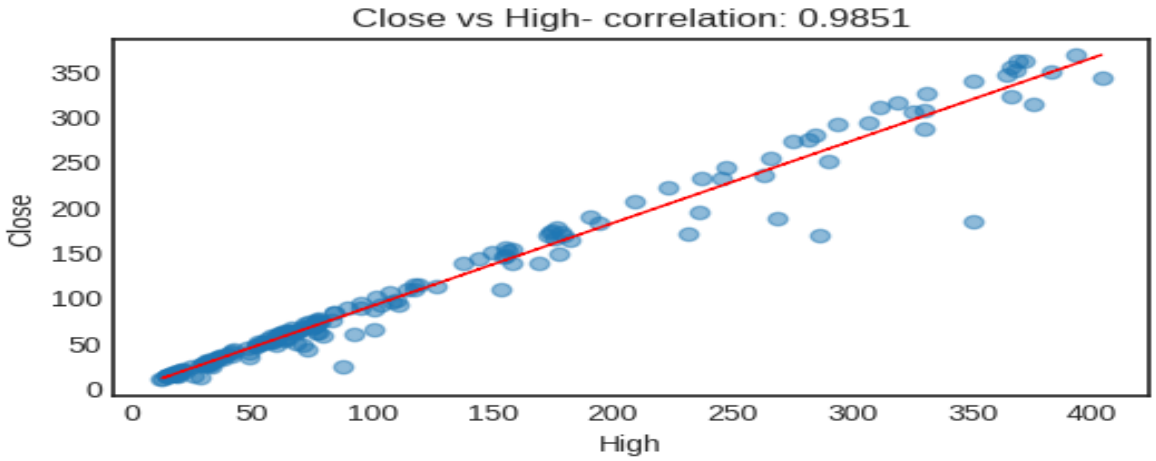


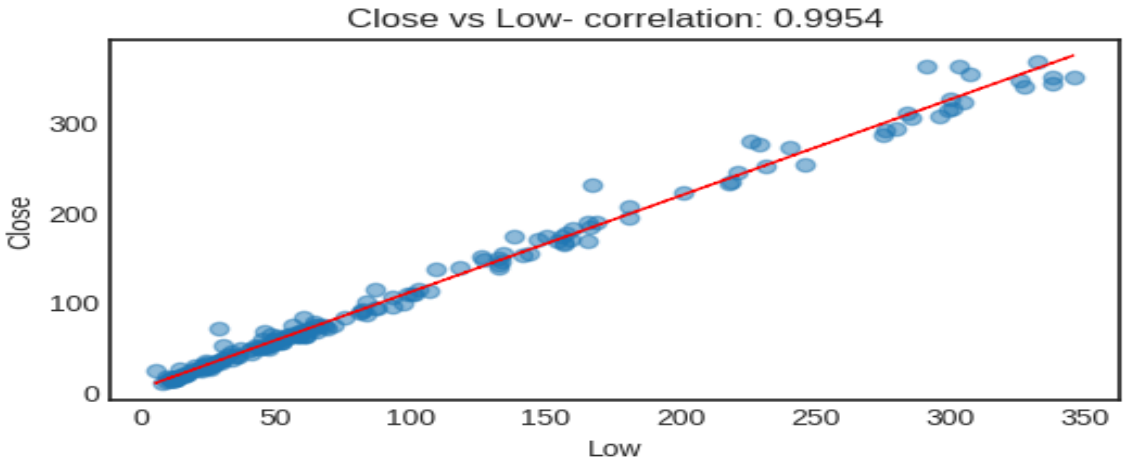
The above graph shows that they are not normally distributed. The mean and median should be equal for a perfect normal distribution curve. So we log transform all the features to normal distribution.

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 a high correlation between the dependent (Close) and independent variables.

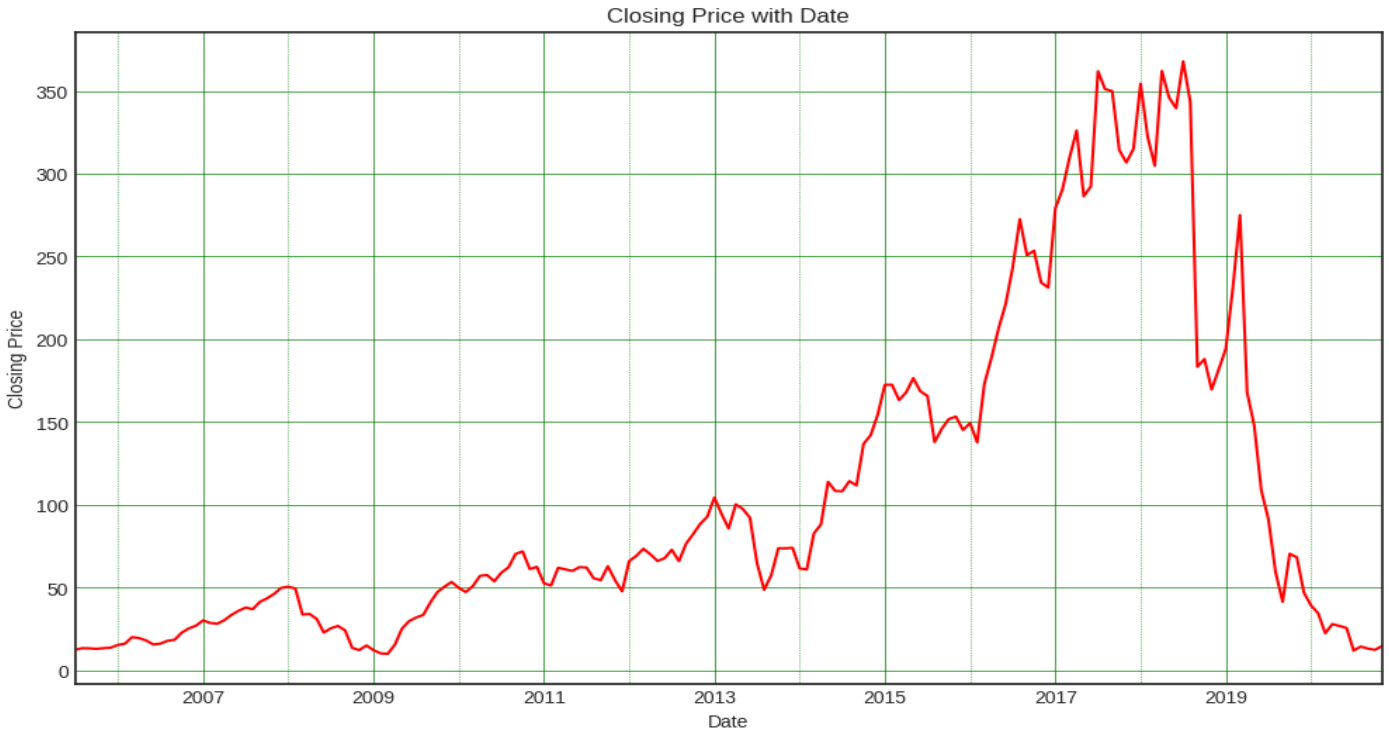






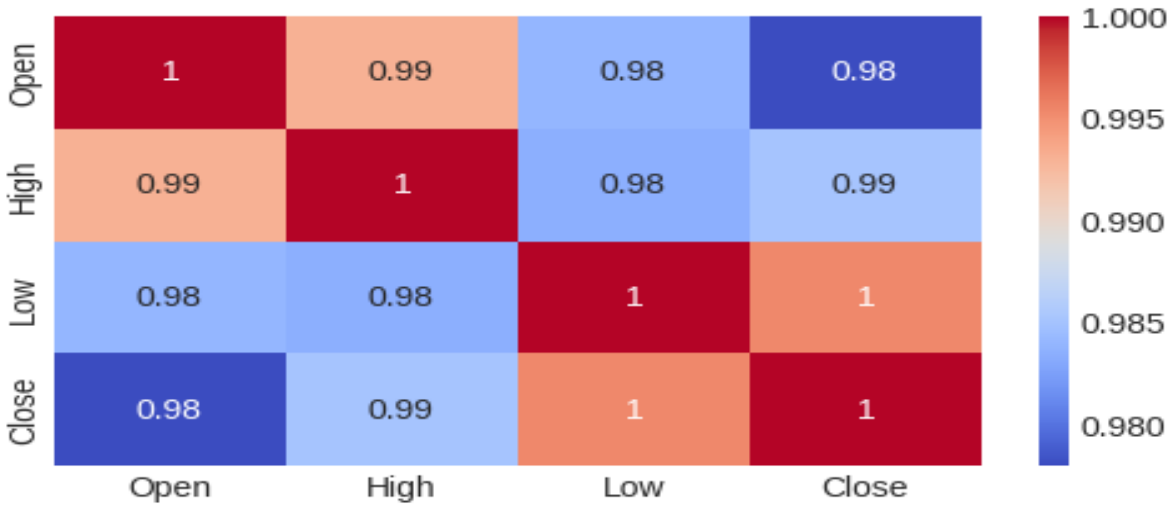
3. Open price and Close Price:

From the below line plot, We conclude that the stock price keeps on increasing till 2018. But after 2018, the stock price kept on decreasing.



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.

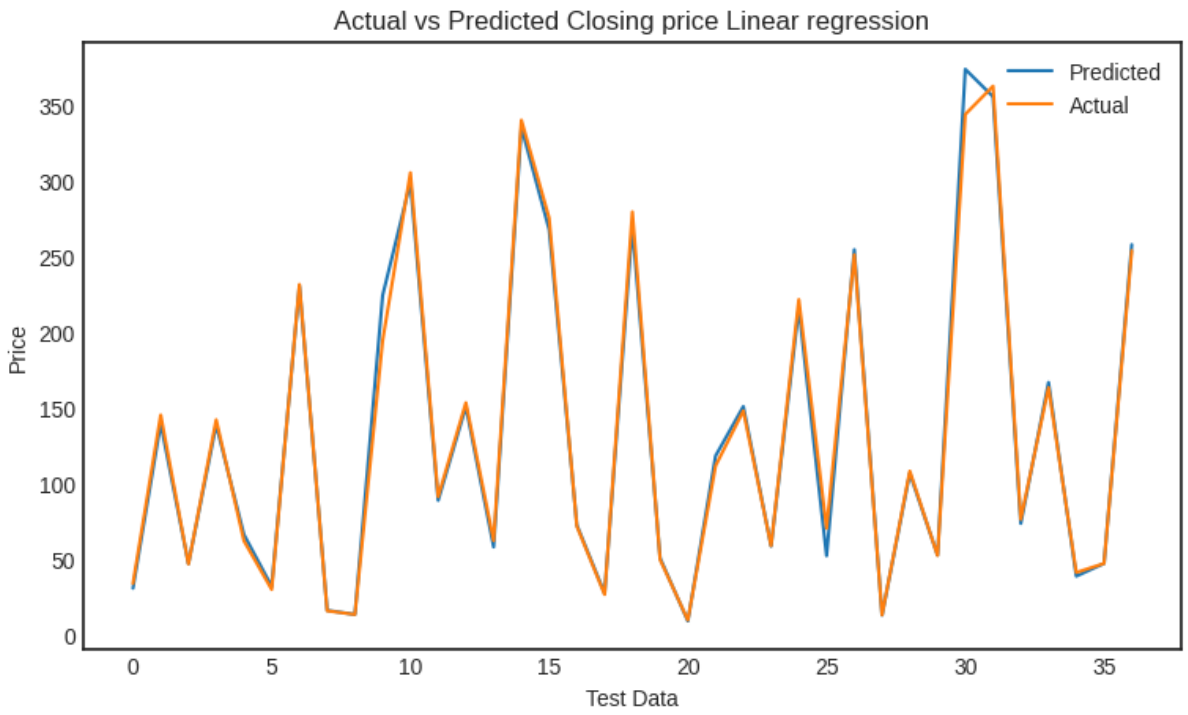


We can see from the above heatmap, that all our independent variables are highly correlated with one another. However, due to this being a small dataset, we can do nothing to remedy this as removing these features or instances will lead to loss of information.

Modeling

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.



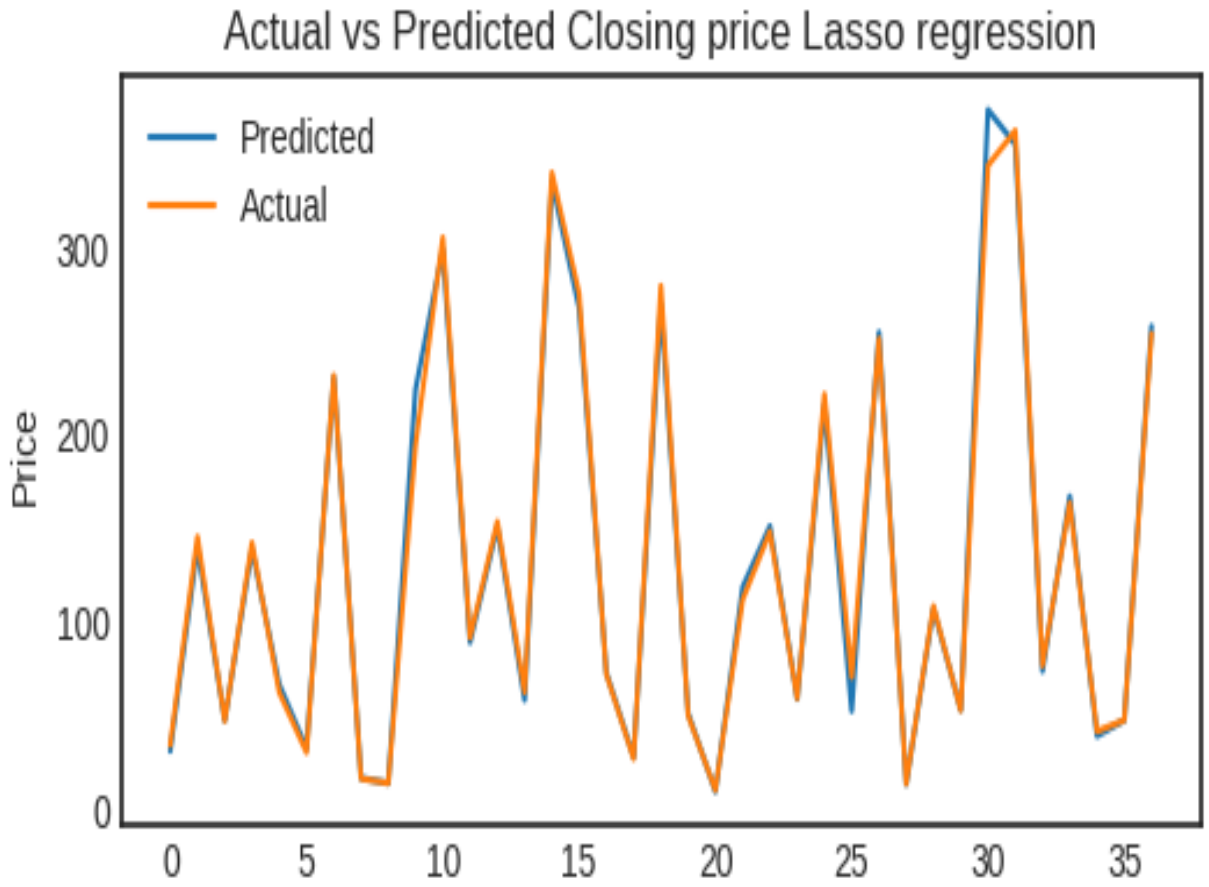
Results:

After implementing Linear Regression:

- Root Mean Square Error is approximately 8.3917
- Adjusted R Square is approximately 0.8212

B) 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.

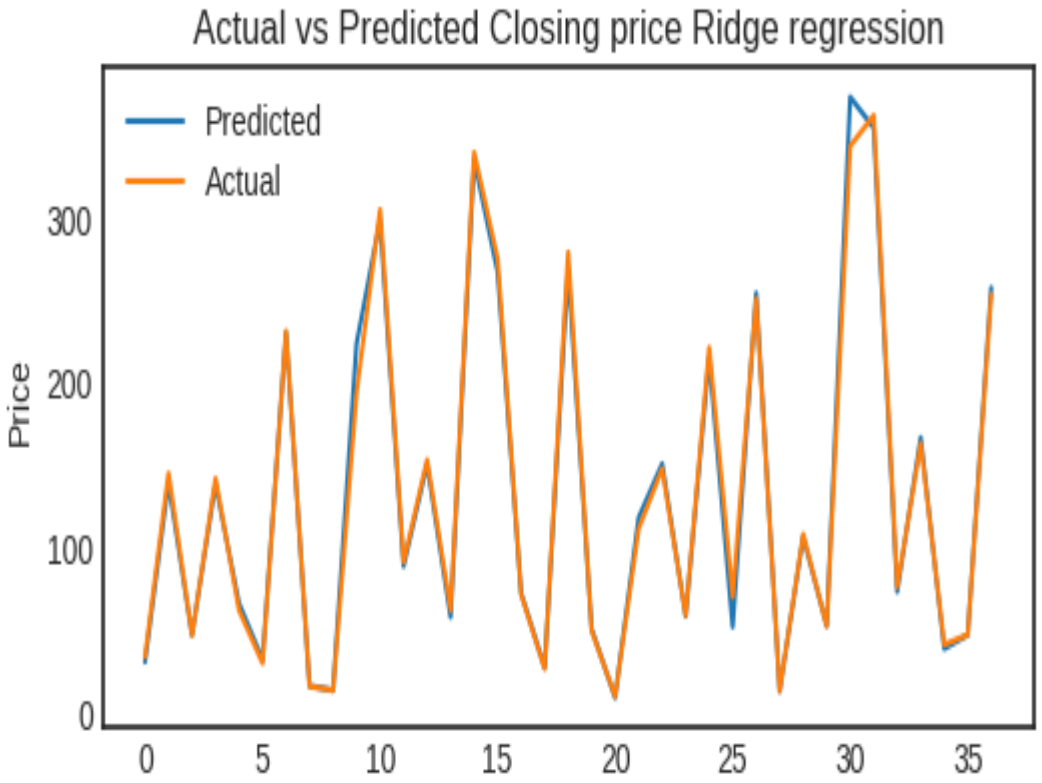


Results:

From the above Lasso Regression graph, we can see the difference between actual and predicted values. RMSE for lasso regression is 8.3864 and the adjusted R-squared value is 0.9932 We use a technique called Cross-validation using which we are able to find the optimal hyperparameter (regularization strength) for best performance.

C) 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 the target variable as given by the model.

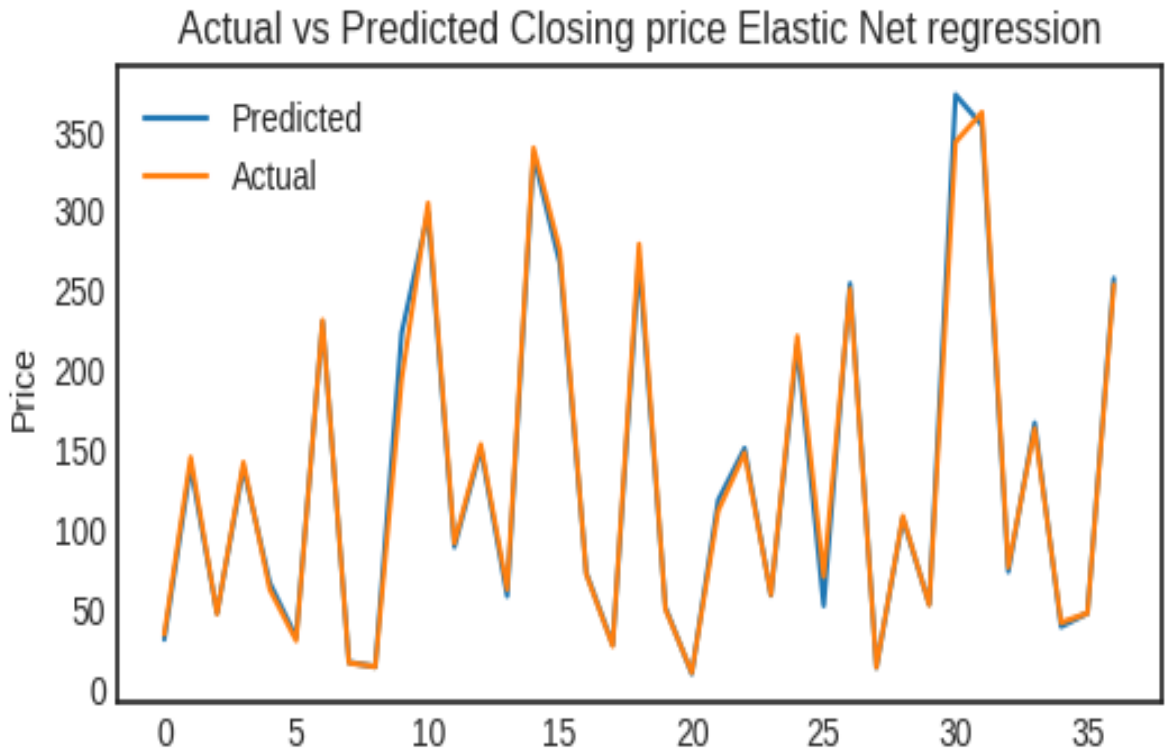


Results :

RMSE for Ridge regression is 8.3824 Adjusted R-squared value for Ridge regression is 0.9932 We use a technique called Cross-validation using which we are able to find the optimal hyperparameter (regularization strength) for best performance.

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.



Results :
RMSE for Elastic Net regression is 8.3760 Adjusted R-squared value for Elastic Net regression is 0.9932 We use a technique called Cross-validation using which we are able to find the optimal hyperparameter (regularization strength and L1 ratio) for best performance.

Conclusion:

Based on the analysis of the Yes Bank stock price prediction using regression analysis, including Linear Regression, Lasso Regression, Ridge Regression, and Elastic Net Regression, we can draw the following conclusions:

Evaluation Metrics: Among all the regression models, the Elastic Net Regression consistently outperforms the other models in every metric considered. The evaluation metrics, including Mean Absolute Error,

Mean Squared Error, Root Mean Squared Error, R2 score, and Adjusted R2 score, indicate that the Elastic Net Regression provides the best fit to the data.

Impact of Fraud Case: By visualizing the target variable, it becomes evident that the 2018 fraud case involving Rana Kapoor had a significant impact on the stock prices. The prices experienced a significant decline during that period.

Data Quality: The dataset used for the analysis does not contain any null values or duplicate data, ensuring the reliability of the results.

Outliers: Although some outliers were present in the dataset, removing them would result in a loss of valuable information due to the small size of the dataset. Hence, the outliers were retained in the analysis.

Skewness and Transformation: The distribution of all variables in the dataset was positively skewed. To address this, a log transformation was performed on the variables to achieve a more symmetric distribution.

High Correlation: The analysis revealed a high correlation between the dependent and independent variables, indicating that the target variable can be accurately predicted using the available features.

Multicollinearity: Due to the small size of the dataset, a high correlation was observed among the independent variables. Although multicollinearity is unavoidable in such cases, it does not significantly impact the performance of the models.

Model Performance: All implemented models performed exceptionally well, yielding an adjusted R2 score of over 99%. The Elastic Net Regression model achieved the highest adjusted R2 score of 0.9932 and consistently scored well in all evaluation metrics.

Heteroscedasticity: The presence of heteroscedasticity was assessed by plotting the residuals against the predicted values of the Elastic Net model. No heteroscedasticity was observed, indicating that the model performed well across all data points.

Deployment: Given the high accuracy of the model's predictions, it can be confidently deployed for future predictive tasks using new data.

The analysis of the Yes Bank stock price prediction using regression analysis demonstrates the effectiveness of the Elastic Net Regression model in accurately predicting the closing price. The model exhibits strong performance across various evaluation metrics and can be considered reliable for further use in predicting future stock prices.
