**Yes Bank Stock Prediction**

**Anushree V. Kajavadekar**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

## Fluctuations in stock prices are notoriously difficult to predict. For decades, economists have created complicated mathematical models that ultimately fail to describe share price movements. With the evolved technology we are now using Machine Learning algorithms to study the pattern and predict.

## Yes Bank is an Indian bank headquartered in Mumbai, India and was founded by Rana Kapoor and Ashok Kapoor in 2004. In June 2005 it went public with the issue price of ₹35. In September 2018, the then appointed CEO of Yes Bank, Rana Kapoor was alleged in a fraud case thus he was ordered by Kapoor to step down from his CEO position in January 2019.

## We are here trying to study the stock price pattern and did it get affected by this fraud case using our knowledge about Python and ML algorithms.

***Keywords: packages and libraries, exploratory data analysis, data visualization, Data Modelling***

**1.Problem Statement**

We have been given a dataset of 185 entries with 5 features which gives a study about this stock Pattern over 15 years. We need to identify the independent and the dependent features to conduct our study. Then using EDA and ML we will study the below: -

* Prediction of Yes bank Stock close Price
* Effect of the Rana Kapoor fraud case on the stock price
* Performing Regression Analysis using Multiple models to predict the Yes Bank Stock Close Price and compare the best fit model.

**Context**: The compact dataset given to us has monthly stock prices of the bank since its commenced and the features are closing, starting, highest and lowest prices of every month. The objective is to predict the dependent feature i.e stock’s closing price of the month. Dataset contains following features--

1. **Date**: contains the month and year of investment.
2. **Open**: the price at which a security first trades when an exchange opens for the day.
3. **Close**: the last price at which a stock trades during a regular trading session.
4. High: a data point on a stock chart that shows the highest value that a stock reached during a trading day
5. **Low**: the lowest price at which a specific stock trades over the course of a trading day.

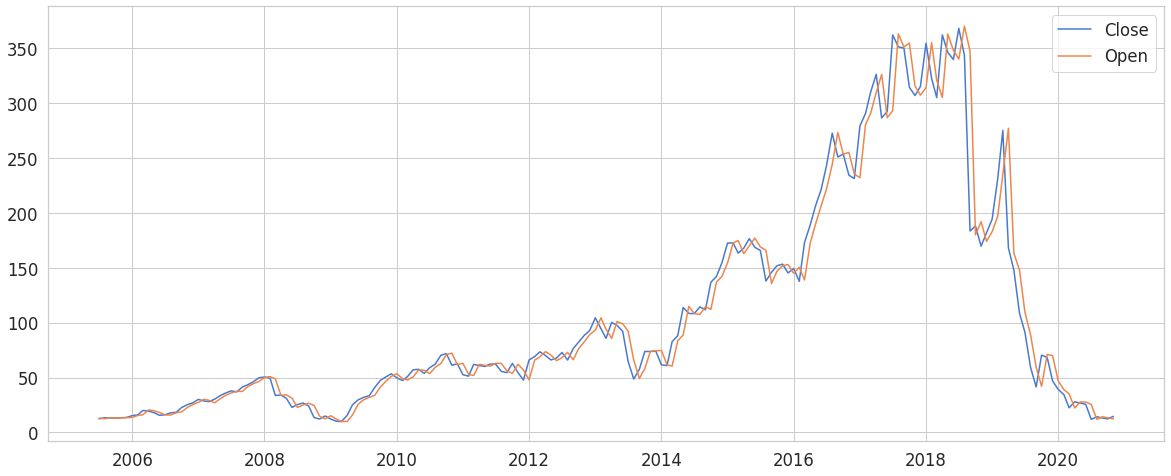
**2**. **Workflow**:

The solution code is divided into the following sections:

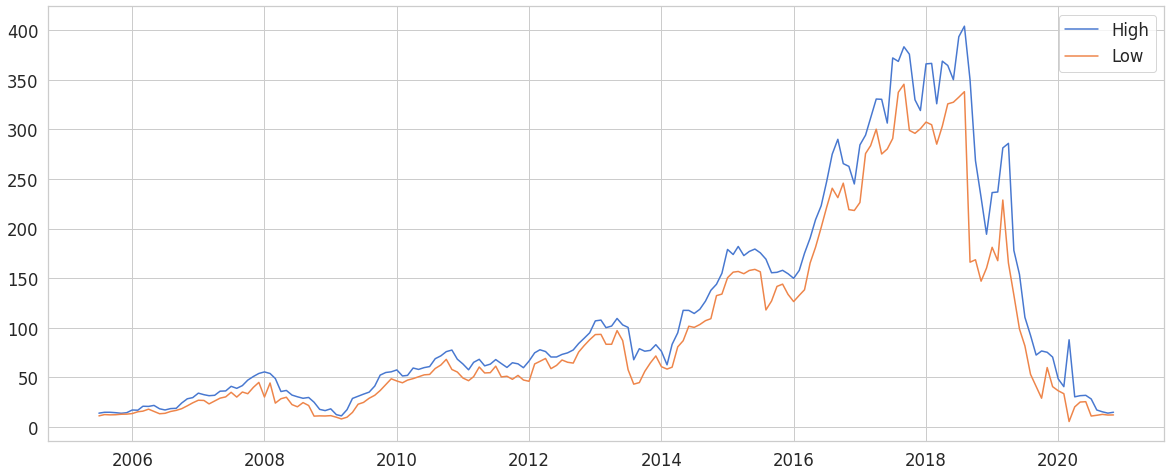
I. **Data Study**: First we will import multiple libraries required to perform the study and then load the CSV file. Later perform some operations to understand and analyze the data.

II. **Data Manipulation**: Since Dataset contains int64, float64, datetime64 type of data and no Null Values in the Dataset cleaning is not needed.

III. **EDA and DATA Visualization**:

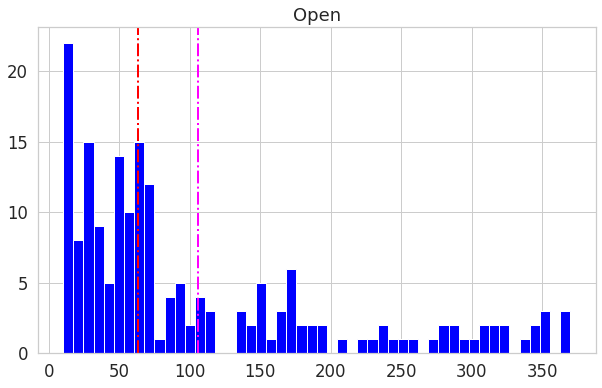
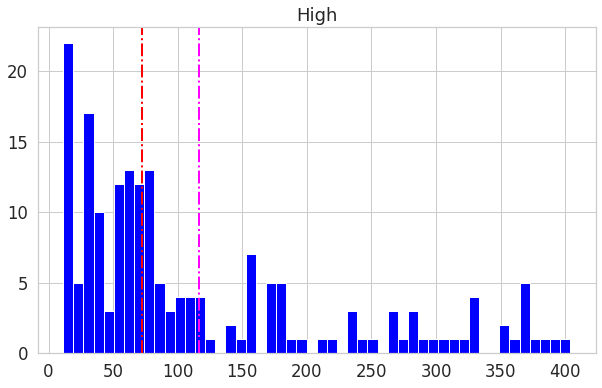
1.Year-Wise Study of Open and Close- 

This Data is between June 2005 to Nov 2020. Initially it opened around 20 till the peak value up to around 370. Then we can see after the fraud case in 2018 involving Rana Kapoor the Open value drastically drops with a small jump in 2019 and then further drops. Closing Value almost shadows the Open value.

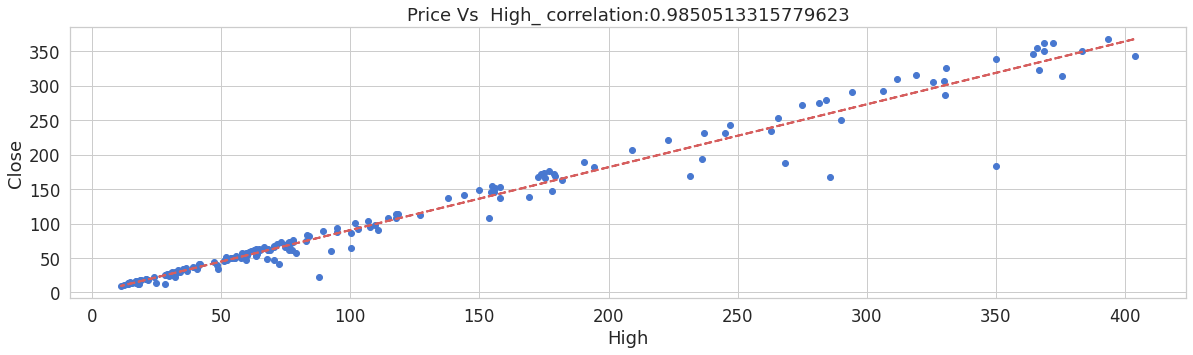
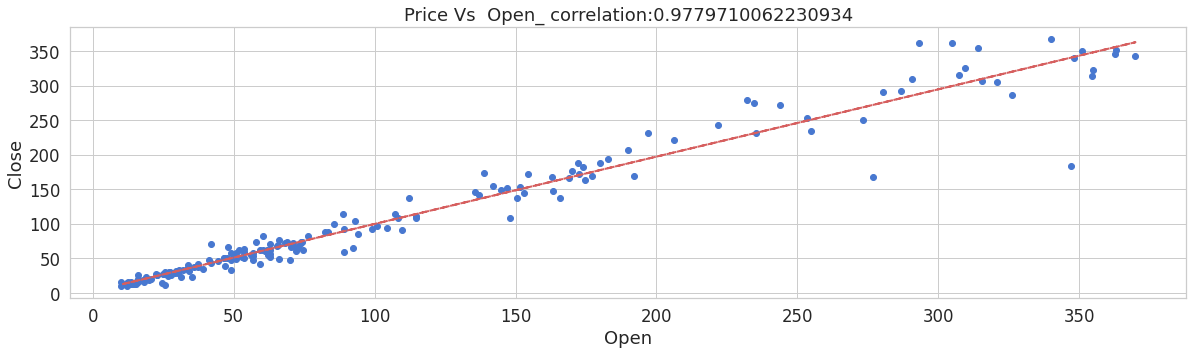
2.Year-Wise Study of High and Low-

3. Data Distribution:

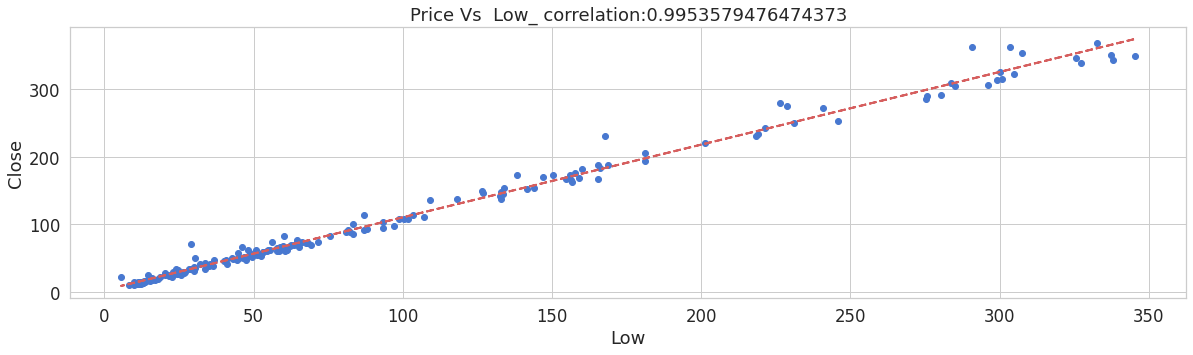
A Study of all possible values for a variable and also the relative frequency.

****

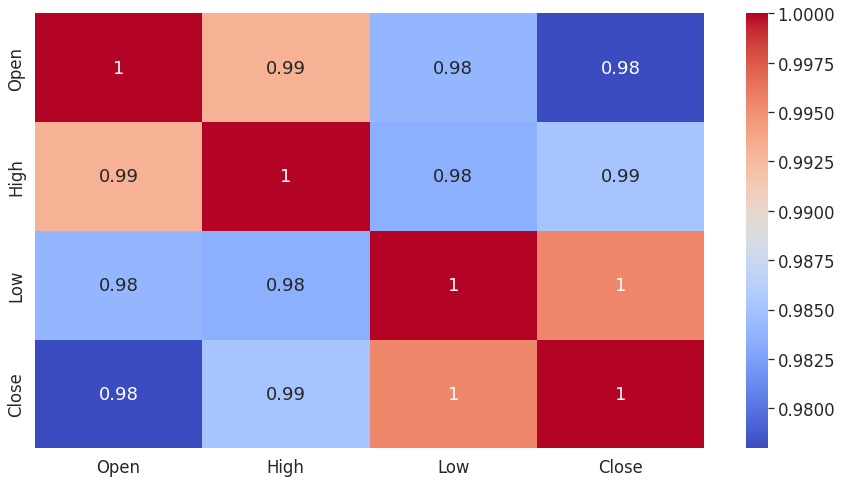
4. Correlation: 1. Price Vs. Open & Price Vs. High



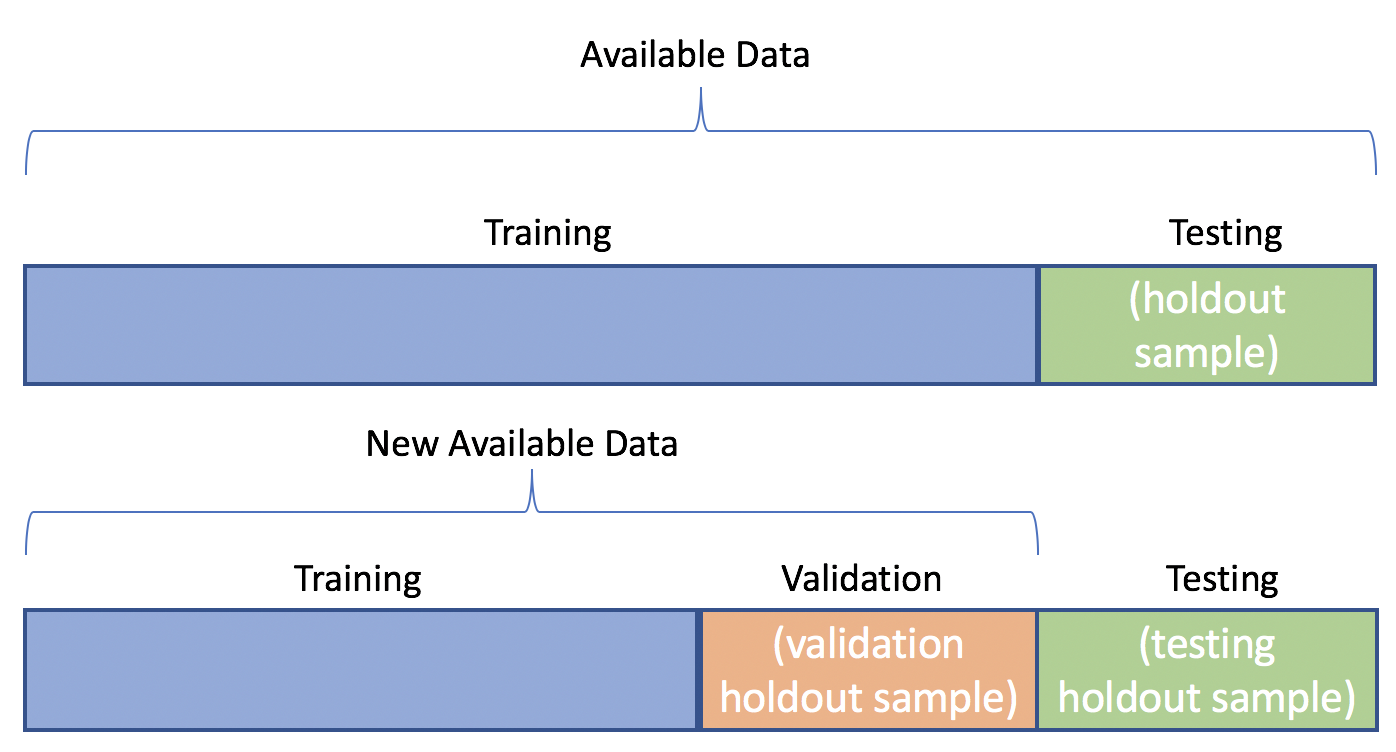
2. Price Vs. Low:



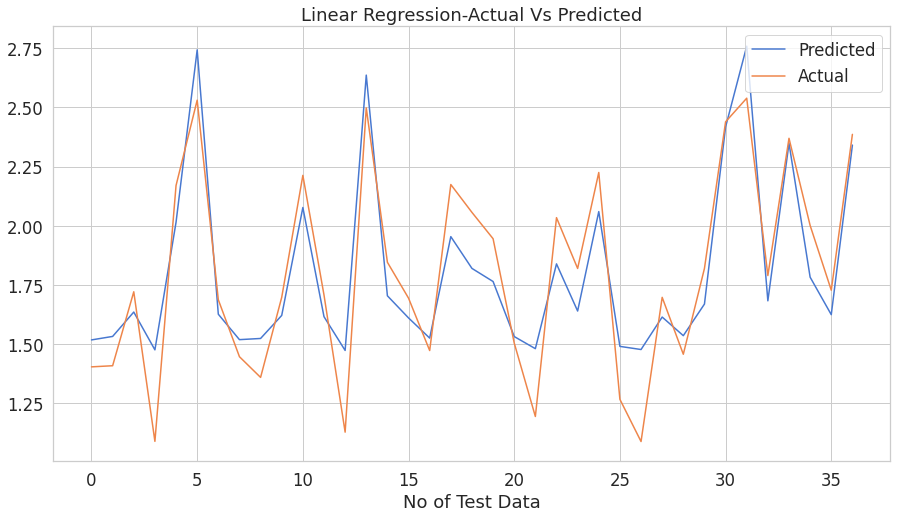
.

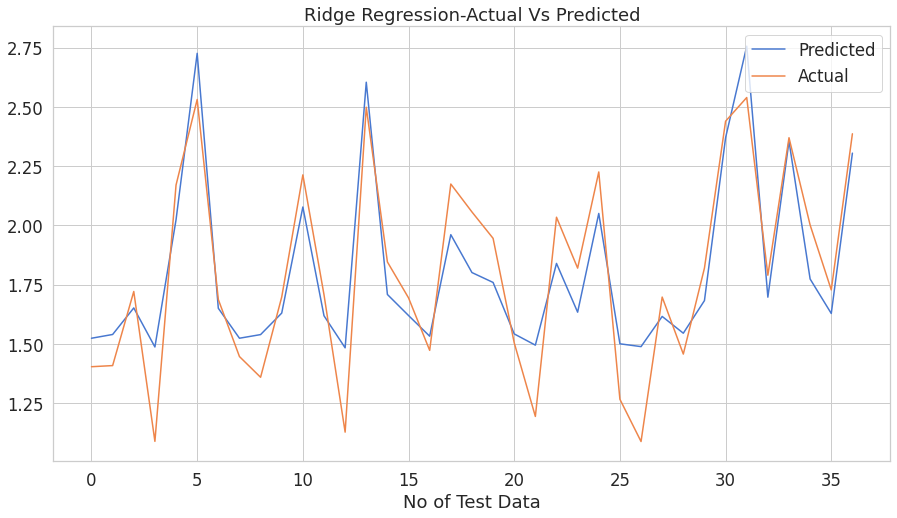
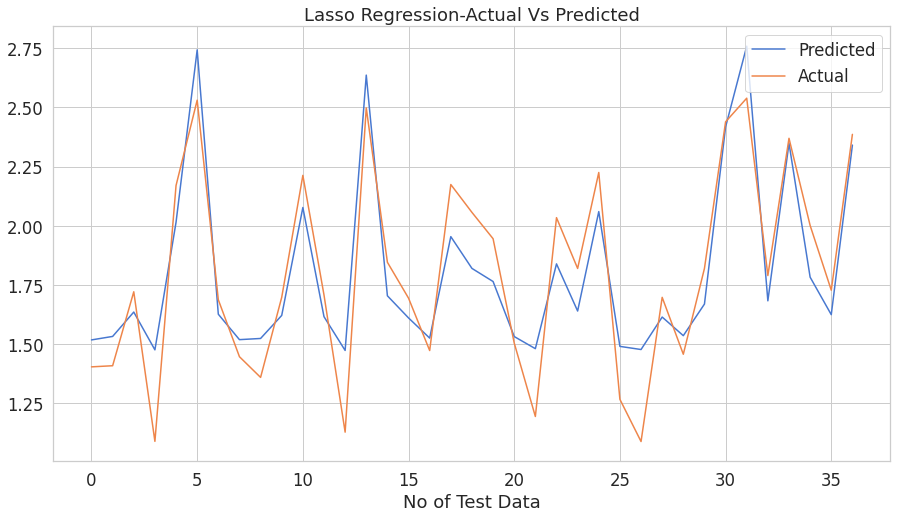
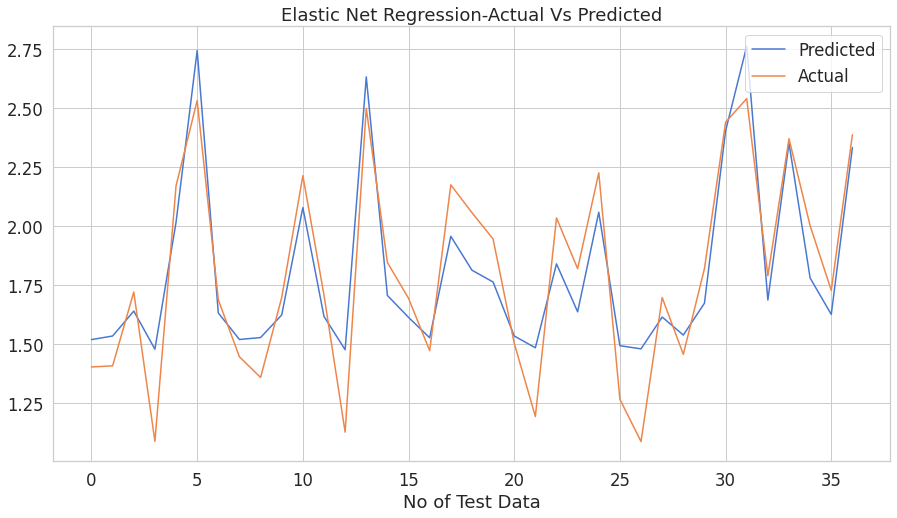
5.Heatmap: This shows the collective representation of correlation between all the features. As shown in the heatmap, we can see that there is a high correlation between each pair of dataset. While it is desirable for the dependent variable to be highly correlated with independent variables, the independent variables should ideally not have high correlation with one another.This causes an issue as multicollinearity is a problem for our models.

6. **Train/Test Model:** The library used to split the dataset: from sklearn.model\_selection import train\_test\_split Train : Test ratio : 80:20



7.**Model Training with cross validation and hyper parameter tuning**

1. Linear Regression: Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable. This calculator is built for simple linear regression, where only one predictor variable (X) and one response (Y) are used.

1. Lasso Regression: Lasso regression is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. 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.
2. Ridge Regression: Ridge regression is a type of linear regression technique that is used in machine learning to reduce the overfitting of linear models. It is a model tuning method that is used to analyze any data that suffers from multicollinearity.
3. ElasticNet 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.

**8. Metrics Comparison:**

| **Model Name** | **MSE** | **MAE** | **RMSE** | **R2** | **Adj.R2** |
| --- | --- | --- | --- | --- | --- |
| **Linear** | **0.03158** | **0.15128** | **0.1777** | **0.8225** | **0.8064** |
| **Ridge** | **0.031685** | **0.151477** | **0.178001** | **0.821997** | **0.805815** |
| **Elastic Net** | **0.031957** | **0.152095** | **0.178764** | **0.820468** | **0.804147** |
| **Lasso** | **0.32040** | **0.152317** | **0.178996** | **0.820001** | **0.803638** |

From the above study we can say that Linear regression has the best R2 and Adjusted R2 score hence Linear Regression is the best model.

**Conclusion:**

* After analyzing the given dataset for the stock, we can clearly see the impact of the 2018 fraud case involving Rana Kapoor as the stock prices decline dramatically during that period after 2018.
* There are no null values and also no duplicate values present in the data set.
* We found that the distribution of all our variables is positively skewed. so we performed log transformation on them.
* There is a high correlation between the dependent and independent variables. This is a signal that our dependent variable is highly dependent on our features and can be predicted accurately from them.
* I found that there is a high correlation between our independent variables. This multicollinearity is however unavoidable here as the dataset is very small.
* I implemented several models on our dataset in order to be able to predict the closing price and found that all our models are performing remarkably well.However, we can see that Linear regression is having the best R2 and Adjusted R2 score hence Linear Regression is the best model.
* Our model is performing well on all data-points.With our model making predictions with such high accuracy, we can confidently deploy this model for further predictive tasks using future data.

**References-**

1. Greeksforgeeks
2. <https://economictimes.indiatimes.com/yes-bank-ltd/stocks/companyid-1655>
3. https://www.javatpoint.com/regression-analysis-in-machine-learning