Yes Bank Stock Price Prediction

Meenakshi

Adityasingh Thakur

Tushar Wagh

Abstract: Yes Bank is an Indian bank headquartered in Mumbai, India and was founded by Rana Kapoor and Ashok Kapoor in 2004. It offers wide range of differentiated products for corporate and retail customers through retail banking and asset management services. We have used yes bank stock price data set. This dataset contains 5 different features that can be used for predicting close price prediction using machine learning. We have built a model which machine learning regression model for price prediction we used all models for prediction of stock price and compute the results and compare them for best accuracy and performance helps us to predict the future stock prices. We have used some of best models.

Keywords: Stock Price prediction, Linear regression, Lasso and Ridge regression, Elastic Net, Nearest neighbour, SVM

I. 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 Time series models or any other 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.

II. Data Description

Before performing any operation on the dataset, it is important to understand the data at a high level. Depending on size and type of data, understanding and interpreting data sets can be challenging. After loading data, we observed the dataset by checking a few of the first and last rows. We checked the shape of the dataset and identified that there are 185 rows and 5 features columns in our dataset. We observed that there are different types of data present in the dataset such as float, object.

- **Date:** It denotes 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.
- **High:** High refer to the maximum prices in a given time period.
- **Low:** Low refer to the minimum prices in a given time period.
- **Close:** Close refers to the price of an individual stock when the stock exchange closed for the day.

III. Introduction

YES bank stands for **Youth Enterprise Scheme Bank**. Stock market is one of the major fields that investors are dedicated to, 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, which focuses on short-term price prediction. A stock market is a public market where you can buy and sell shares for publicly listed companies. The stocks, also known as equities, represent ownership in the company. The stock exchange is the mediator that allows the buying and selling of shares.

Stock Price Prediction using machine learning helps you discover the future value of company stock 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 other factors involved in the prediction, such as physical and psychological factors, rational and irrational behaviour, and so on. All these factors combine to make share prices dynamic and volatile. This makes it very difficult to predict stock prices with high accuracy.

IV. Exploratory Data Analysis

A) Data Cleaning: -

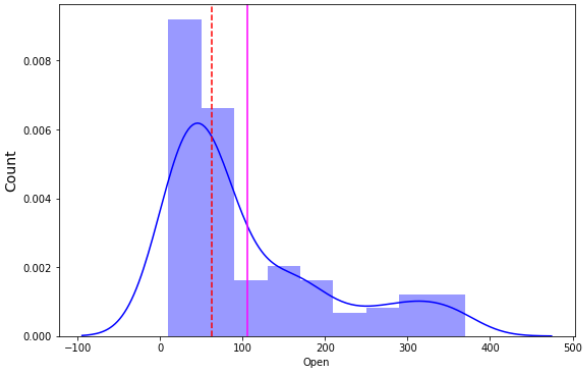
The Given Date in data is of format MMM-YY is converted to proper date of YYYY-MM-DD and given date column has dtype as object converting it into date time format.

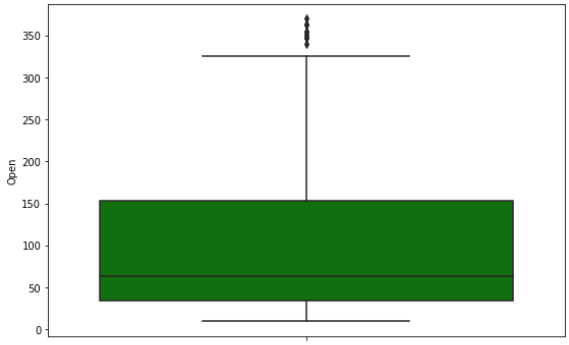
B) Null values Treatment:

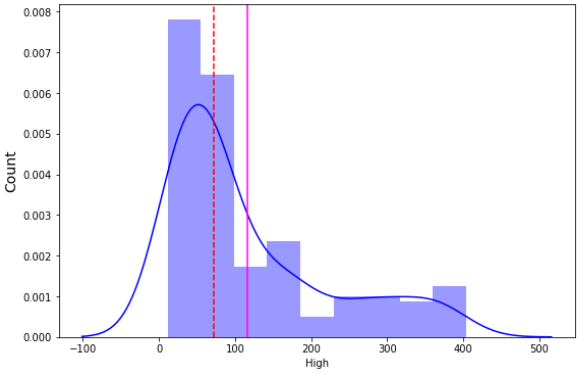
Our dataset does not contain null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result

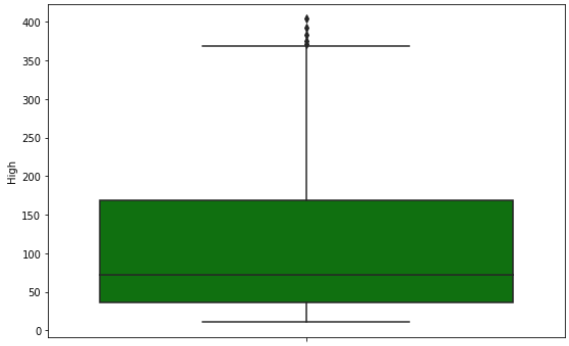
C) Data Visualization:

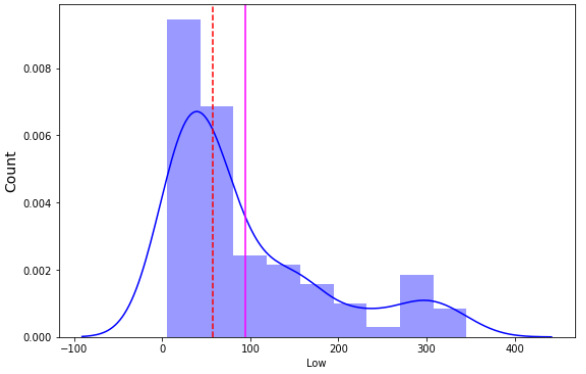
1. Univariate Analysis: In our yes bank stock market dataset all feature histogram are right skewed.

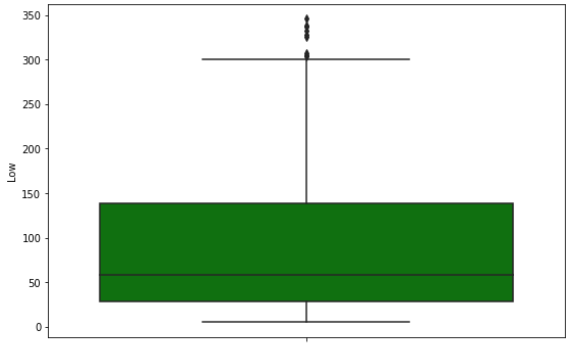


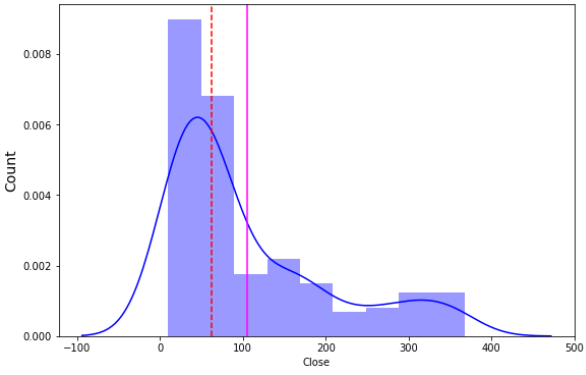


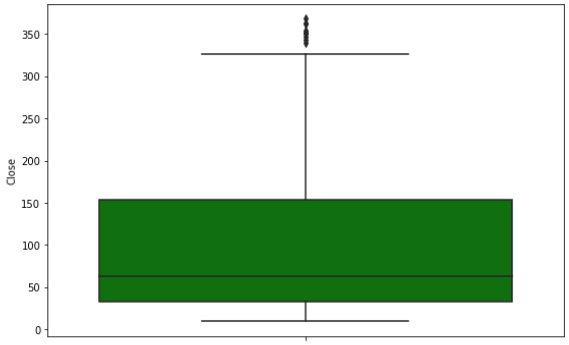












The above graph shows that they are not a normal distribution curve. The mean and median should be equal for perfect normal distribution curve. But, mean is not equal to median as there is not a perfect normal distribution curve. We need to convert all the features to normal distribution using log transformation.

Outliers are present in each column. By, converting our features to normal distribution using log transform. We can remove outliers from the dataset.

From the above box plot we can see that after applying np.log10() method with independent features "Open", "High", "Low" we get a normal distribution curve which helps to remove the outliers from the column "Open", "High", "Low".

From the box plot, we can also the **quartile (q1, q2, q3)**

We got the approximate result: -

For Feature **"Open": -**

- Lower Quartile (Q1): - 3.6
- Median (Q2): - 4.3
- Upper Quartile (Q3):- 5.0

For Feature **"High":-**

- Lower Quartile (Q1):- 3.6
- Median (Q2):- 4.4
- Upper Quartile (Q3):- 5.1

For Feature **"Low":-**

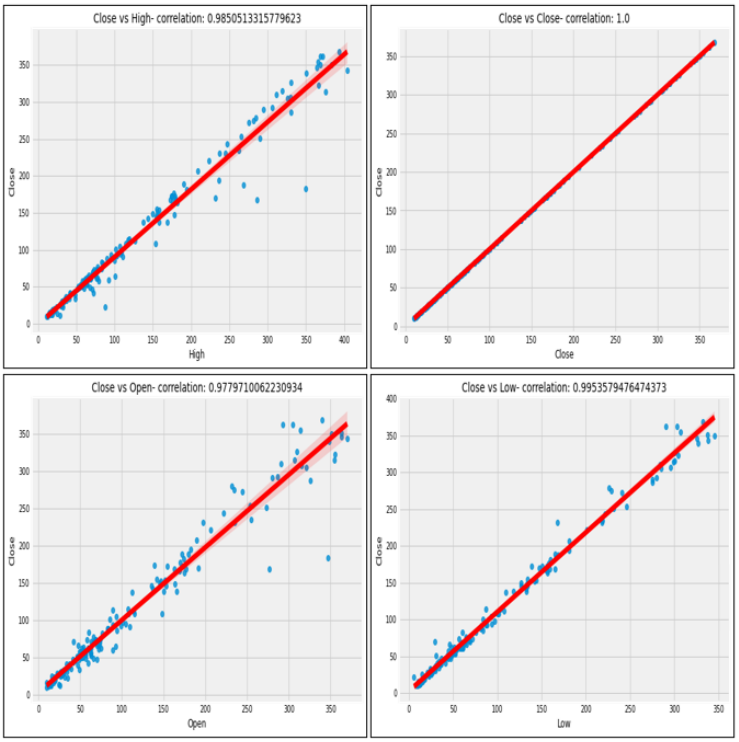
- Lower Quartile (Q1):- 3.3
- Median (Q2) :- 4.1
- Upper Quartile (Q3) :- 4.9

2. Bivariate Analysis:

In the context of supervised learning, it can help determine the essential predictors when the bivariate analysis is done keeping one of the variables as the dependent variable (Y) and the other ones as independent variables

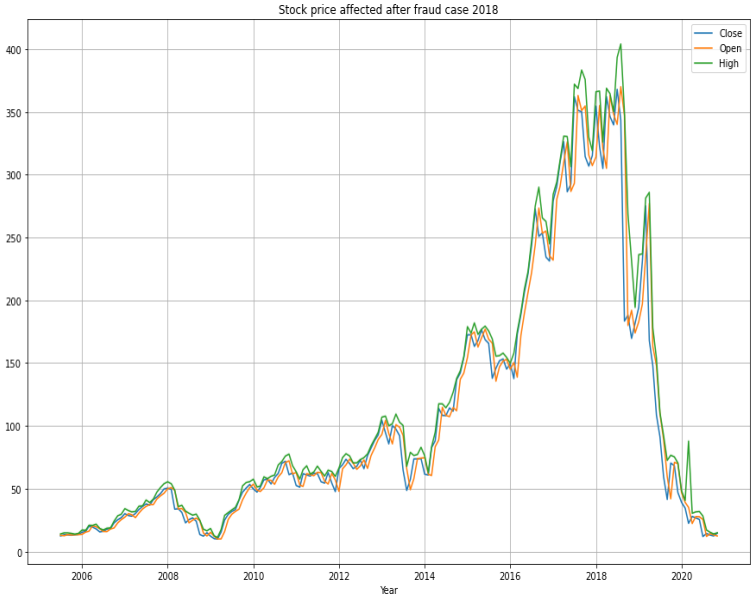
(X1, X2, … and so on) hence plot all Y, Xs pairs. So essentially, it is a way of feature selection and feature priority

The above graphs depicts that there is high correlation between dependent (Close) and independent (High, Low, Open) features. We try to reduce the correlation for better prediction of the model. We calculate the VIF factor to reduce the multicollinearity between independent variables.



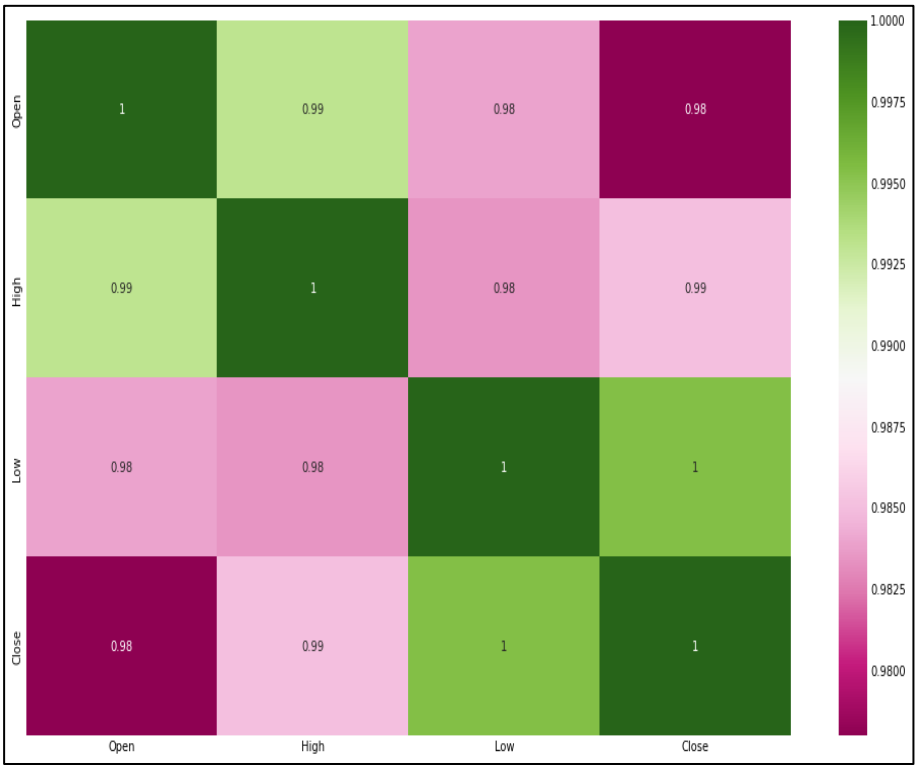
3. Open price and Close Price:

From the above line plot, We conclude that the stock price is keep on increasing till 2018.But after 2018, the stock price is kept on decreasing due the fraud case involving Rana Kapoor.



4. Correlation Analysis:

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables. In the above correlation n heatmap all variable shows highest correlation among them.



5. Multicollinearity:

	variables	VIF
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

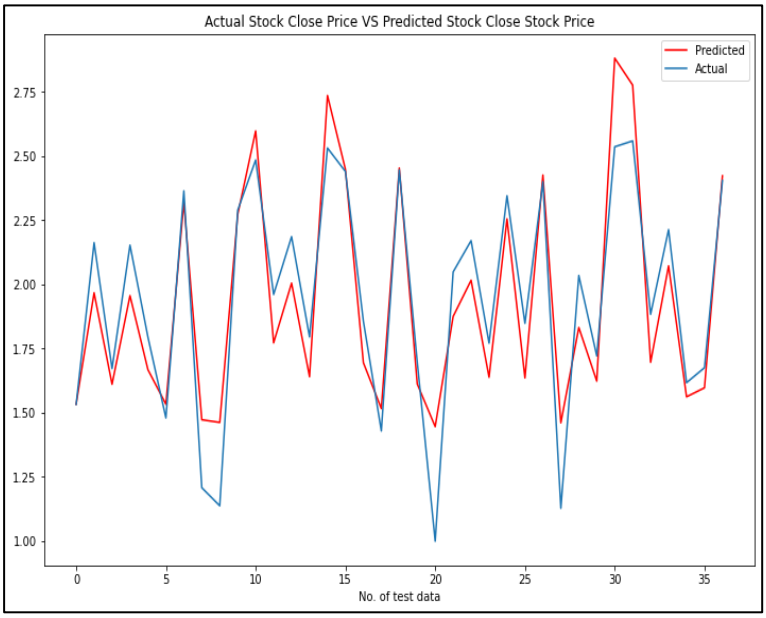
VIF scores are high so it implies that associated independent variables are highly collinear to each other in the dataset. As all the variables are equally important for closing stock price prediction, So we will not be performing any kind of feature engineering here.We are not removing any column because all the columns are equally important for prediction. Removing column lead to loss of valuable information (features) which are essential for accurate prediction for the model. It results in bad model.So, we are not deleting any features form the dataset and try to predict the result and see how the model performs with multicollinearity and evaluate the performance of the model.

V. Modelling

A) Linear Regression:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression algorithm shows a linear relationship between a dependent (y) (in our case is Close Price) and one or more independent (in our case Open, Low, high) variables, hence called as linear regression.

```
Mean Squared Error:  0.0319805266701623
Root Mean Squared Error:  0.1788310003052108
R2:  0.8283222778327901
Adjusted R2:  0.8127152121812256
Mean Absolute Percentage Error:  0.087  %
```



Conclusion:

After implementing Linear Regression:

- Mean Square Error is approximately 0.032
- Adjusted R Square is approximately 0.8212

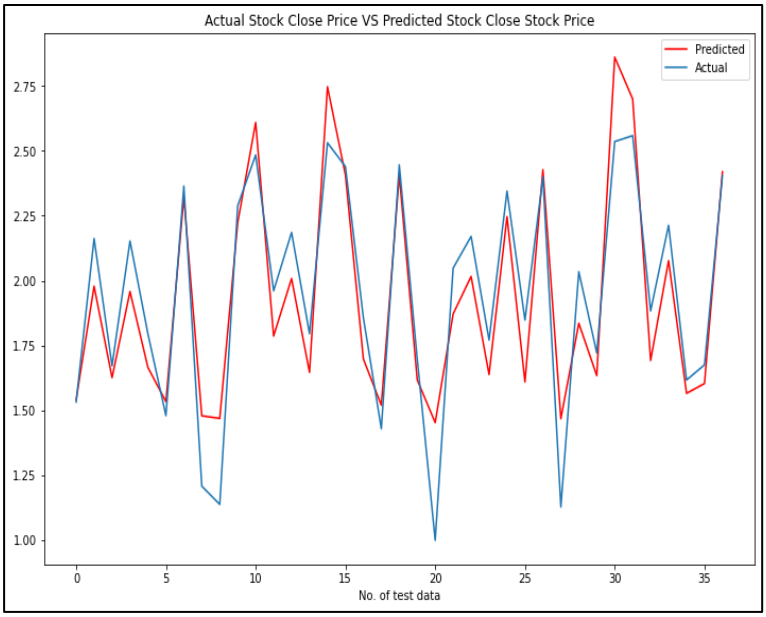
- Mean Absolute Percentage Error is 0.0918 %

B) Lasso Regression:

The goal of **lasso regression** is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. When we deployed the lasso regression model's result are given below table.

```
Mean Squared Error:  0.031621196970051925
Root Mean Squared Error:  0.17782349948769968
R2:  0.8302512299434985
Adjusted R2:  0.8148195235747256
Mean Absolute Percentage Error:  0.0876  %
```



Conclusion:

- From the above Lasso Regression graph, we can see that the No of Test Data is on the X-axis whereas the predicted values is being mapped on Y-axis.

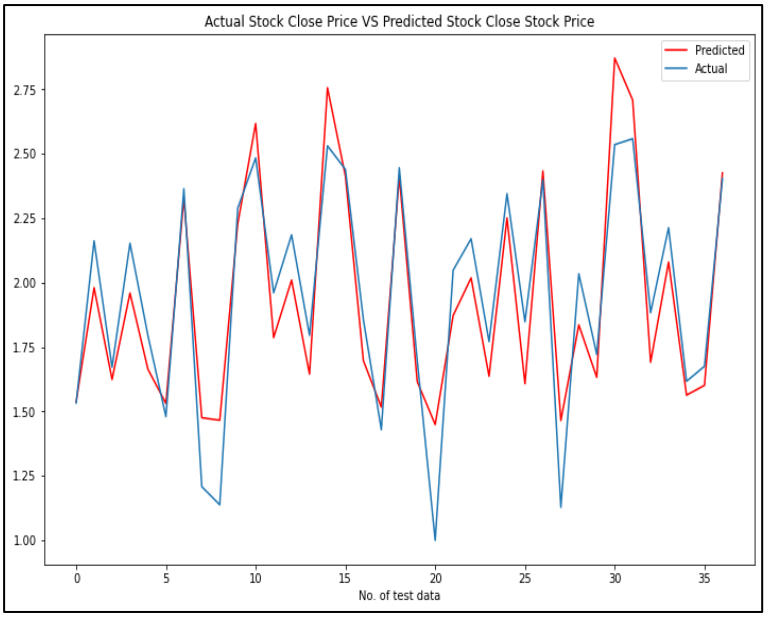
- In the Graph Predicted Value is indicated by Blue Color while the Actual Value is indicated by orange colour.
- When No of Test Data = 30 we can see that that the Predicted Value [More than 2.75] is much higher than the Actual Value [Approx. 2.52]. And in some cases, Actual Value is higher than Predicted Value.
- When No of Test Data = Between 5 To 10 range [Approx. 8], 17, 20, Between 25 to 30 range [Approx. 27] we can observe that the Predicted Value is less than the Actual Value.
- When No of Test Data = 6, 24, 26 there is less difference between the Actual Value and Predicted Value.

C)Cross validation in Lasso Regression: -

Cross Validation: In cross validation, we divide our dataset into 3 parts training, validation and testing. The testing data is only for the final check, train and validation is used for the hyper parameter tuning in order to avoid the data leakage.

Hyperparameters: are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. For eg, alpha, cv.

```
Mean Squared Error:  0.03173683133331823
Root Mean Squared Error:  0.17814834080989422
R2:  0.829630482064811
Adjusted R2:  0.814142344070703
Mean Absolute Percentage Error:  0.0875  %
```



Conclusion: -

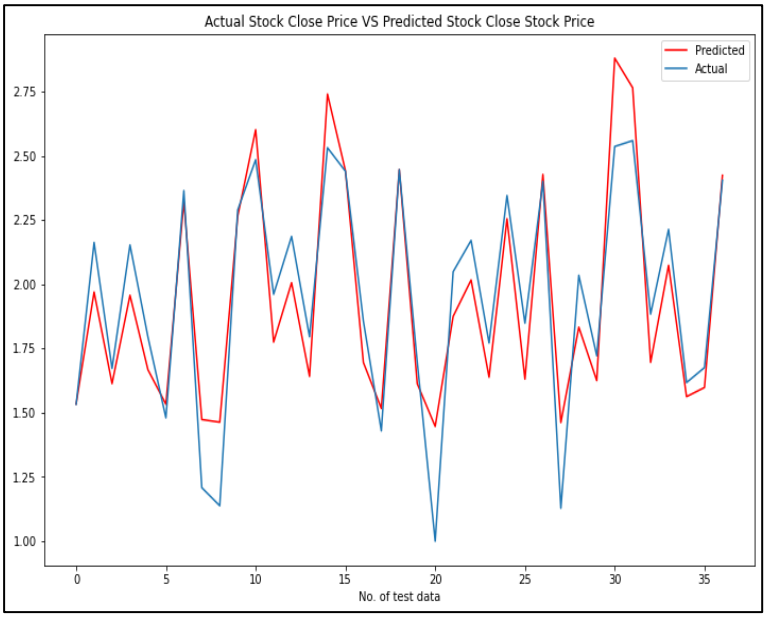
After implementing Lasso Regression with CV:

- Mean Square Error is approximately 0.032
- Adjusted RSquare is approximately 0.823
- Mean Absolute Percentage Error is 0.0923 %

D) Ridge Regression:

Ridge regression is a model tuning method that is used to analyses any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

```
Mean Squared Error:  0.03189782363389192
Root Mean Squared Error:  0.1785996182355716
R2:  0.8287662439071969
Adjusted R2:  0.8131995388078512
Mean Absolute Percentage Error:  0.0869  %
```



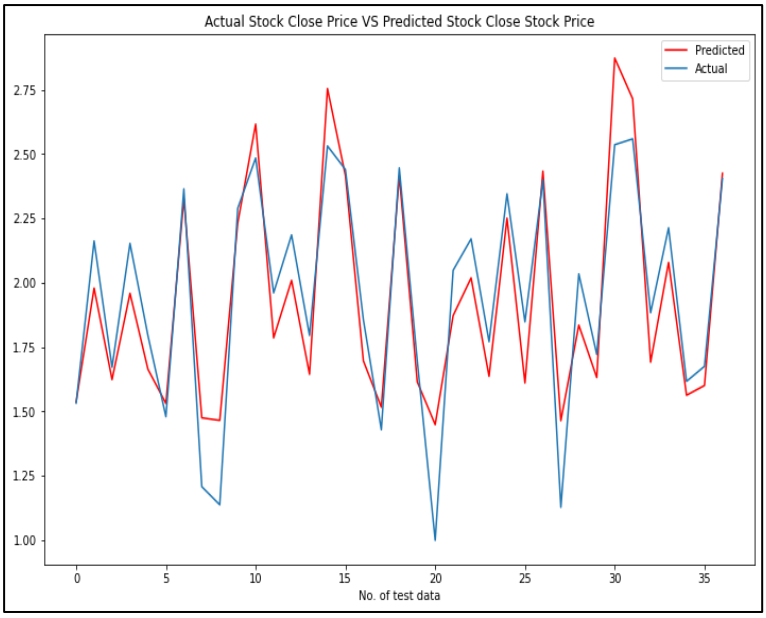
Conclusion

- From the above Ridge Regression graph, we can see that the No of Test Data is on the X-axis whereas the predicted values are being mapped on Y-axis.
- In the Graph Predicted Value is indicated by Blue Color while the Actual Value is indicated by Orange Colour.
- When No of Test Data = 30 we can see that that the Predicted Value [More than 2.75 [Approx.2.90]] is much higher than the Actual Value [Approx. 2.52] And in some cases Actual Value is Higher than Predicted Value.
- When No Of Test Data = Between 5 To 10 range [Approx. 8] , 17 , 20 , Between 25 to 30 range[Approx. 27] we can observe that the Predicted Value is less than the Actual Value.
- When No of Test Data = 6, 19, 24, 26 there is less difference between the Actual Value and Predicted Value.

E) Ridge Regression with Cross Validation

Ridge Regression is a technique for analysing multiple regression data that suffer from multicollinearity. It shrinks coefficients toward zero, but they rarely reach zero.

```
Mean Squared Error:  0.031714074234916886
Root Mean Squared Error:  0.1780844581509484
R2:  0.8297526466200409
Adjusted R2:  0.8142756144945901
Mean Absolute Percentage Error:  0.0874  %
```



Conclusion:

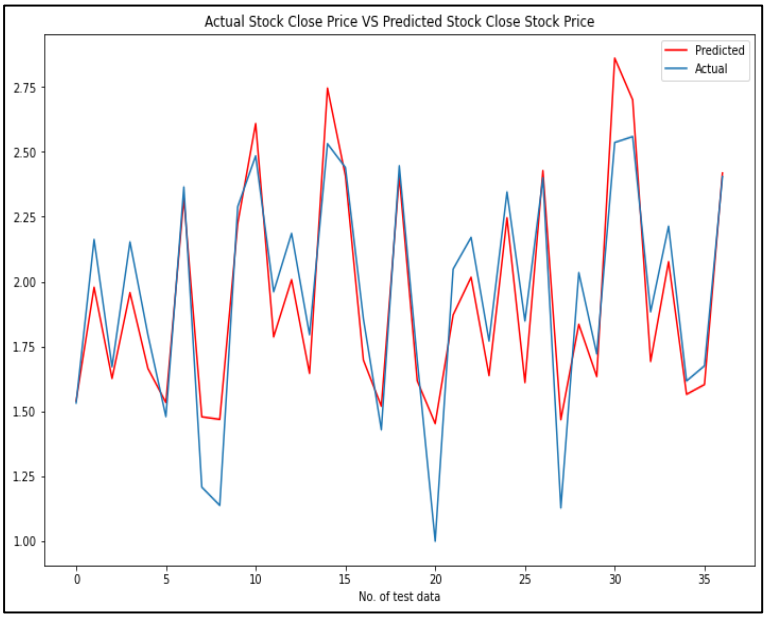
After implementing Ridge Regression with CV:

- Mean Square Error is approximately 0.0317
- Adjusted RSquare is approximately 0.814
- Mean Absolute Percentage Error is 0.092 %

F) Elastic Net Using Cross Validation

Elastic Net regression is a combination of Lasso regression and Ridge regression.

```
Mean Squared Error:  0.03159601336566866
Root Mean Squared Error:  0.17775267470749537
R2:  0.8303864204574344
Adjusted R2:  0.8149670041353829
Mean Absolute Percentage Error:  0.0876  %
```



Conclusion:

After implementing Elastic Net with CV:

- Mean Square Error is approximately 0.032
- Adjusted RSquare is approximately 0.824
- Mean Absolute Percentage Error is 0.0922

G) KNeighbour Regressor

KNN (Nearest neighbours) Regressor: - KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbours. Another approach uses an inverse distance weighted average of the K nearest neighborship KNN regression, the KNN algorithm is used for estimating continuous variables. One such algorithm uses a weighted average of the k nearest neighbours, weighted by the inverse of their distance. This algorithm works as follows:

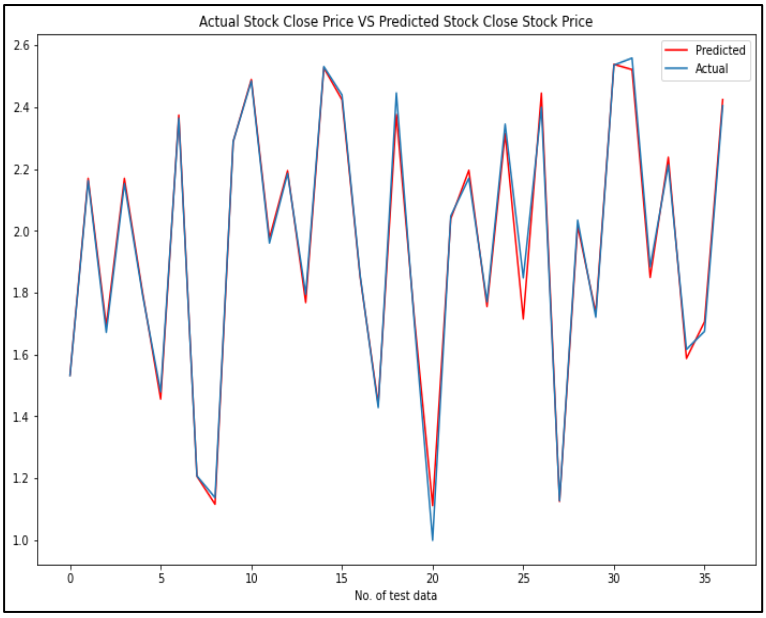
1. Compute the Euclidean from the query example to the labelled examples.

2. Order the labelled examples by increasing distance.

3. Find a heuristically optimal number k of nearest neighbours, based on RMSE. This is done using cross validation.

4. Calculate an inverse distance weighted average with the k-nearest multivariate neighbours.

```
Mean Squared Error:  0.0013166265949414065
Root Mean Squared Error:  0.03628534959100445
R2:  0.9929320909222173
Adjusted R2:  0.9922895537333281
Mean Absolute Percentage Error:  0.0136  %
```



Conclusion

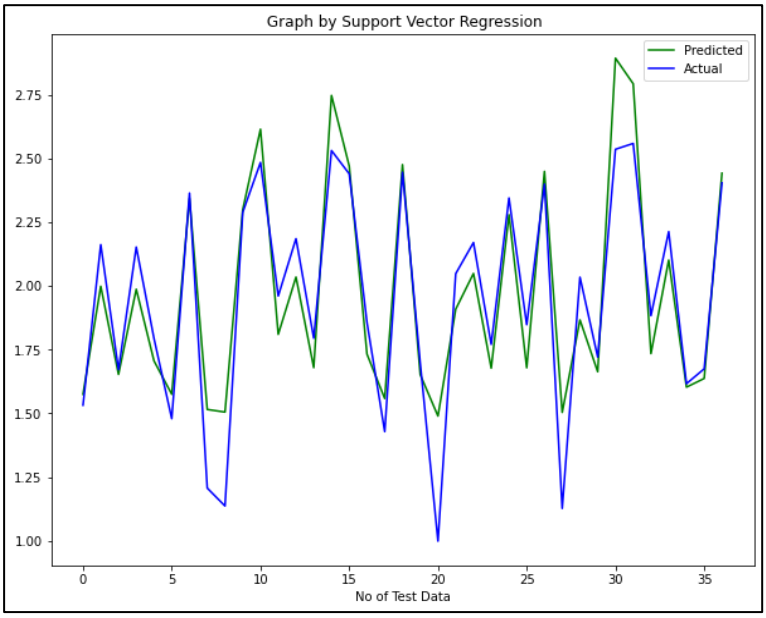
After implementing KNeighbour Regressor:

- Mean Square Error is approximately 0.002
- Adjusted RSquare is approximately 0.984
- Mean Absolute Percentage Error is 0.0213 %

H) SVR (Support Vector regressor):

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points. Support Vector Regression uses the same principle of Support Vector Machines. In other words, the approach of using SVMs to solve regression problems is called Support Vector Regression or SVR.

```
Mean Squared Error:  0.03175287822394554
Root Mean Squared Error:  0.17819337312017397
R2:  0.829544339217347
Adjusted R2:  0.8140483700552876
Mean Absolute Percentage Error:  0.0844  %
```



A kernel is a function which places a low dimensional plane to a higher dimensional space where it can be segmented using a plane. In other words, it transforms linearly inseparable data to separable data by adding more dimensions to it.

- Linear kernel: Dot product between two given observations
- Polynomial kernel: This allows curved lines in the input space
- Radial Basis Function (RBF): It creates complex regions within the feature space in our model we used the linear kernel.

Conclusion

- Actual values are shown by blue line and predicted values are shown by green color line in a graph above.
- SVR have less value of adjusted R^2 as compared to all other algorithms.
- It has mean error square of around 0.034.

VI. Evaluation Metric

Evaluating our models, we will consider the following metrics

1) MSE: Mean Squared Error is one of the most preferred metrics for a regression model. It is simply an averaged squared difference between the target value and value predicted by the regression model.

2) RMSE: Root mean squared value is the square root averaged squared difference between the target value and value predicted by the regression model.

3) MAPE: Mean Absolute Percentage Error is also known as mean absolute percentage deviation is a measure of prediction accuracy of forecasting methods in statistics.

4) R-Square: The metric that helps us to compare the current model with a constant model baseline and tell us how much our model is better.

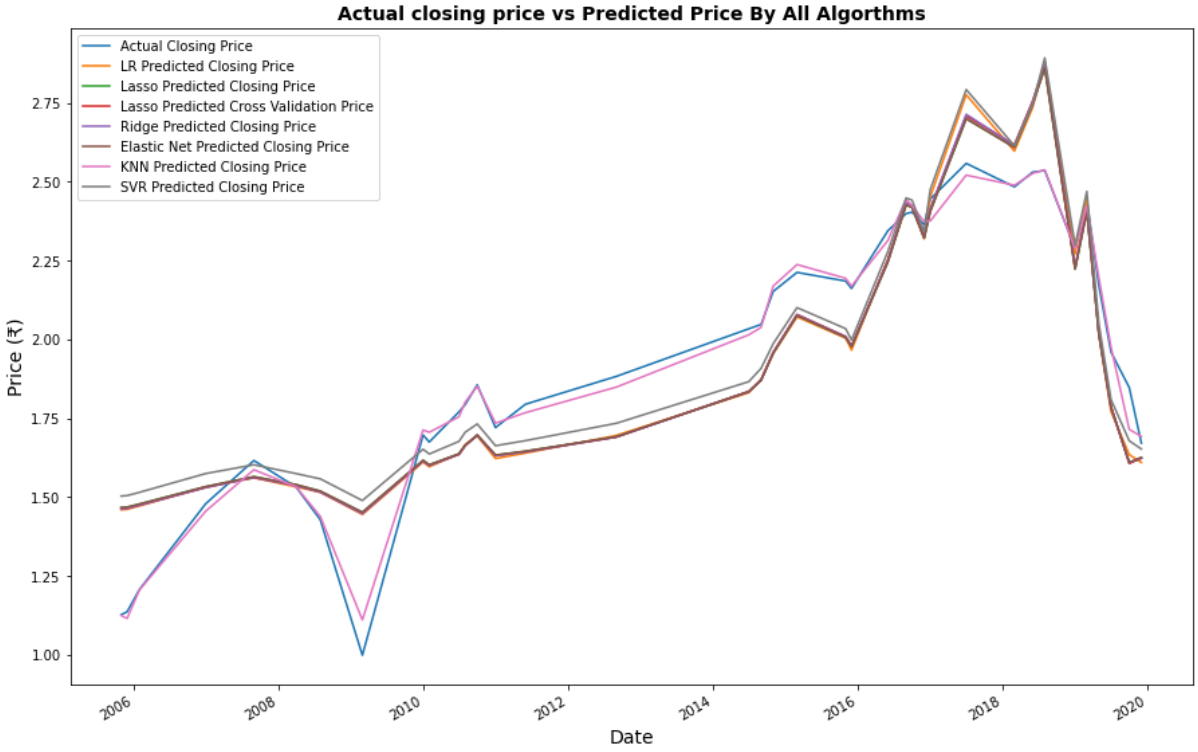
5)Adjusted R-square: Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.

	Model_Name	MSE	RMSE	R2	Adjusted R2	MAPE
0	Linear regression	0.0326	0.1806	0.8310	0.8213	0.0918
1	Lasso regression	0.0325	0.1803	0.8314	0.8217	0.0918
2	Lasso Regression CV	0.0321	0.1792	0.8335	0.8239	0.0923
3	Ridge Regression CV	0.0320	0.1789	0.8341	0.8245	0.0922
4	Elastiv Net CV	0.0321	0.1792	0.8336	0.8240	0.0922
7	K-Neighbor Regressor	0.0029	0.0539	0.9850	0.9841	0.0213
8	K-Neighbor Regressor CV	0.0015	0.0390	0.9921	0.9916	0.0158

Conclusion

KNN Regressor gives lowest MAE, MSE, RMSE, MAPE and best R^2 value.

Overall, we can say that KNN is the best model among all regression models which gives around 99.00% accuracy of predicting stock price of our dataset.



VII. Final Conclusion

- Target Variable is strongly dependent on Independent Variables.

- We have seen that there is neither null nor duplicate values. But Date feature have values of object data type. So, we converted it into proper date format YYYY-MM-DD.
- KNeighbour Regressor and KNeighbour Regressor CV performing better than other models with adjusted R^2 0.9841 and 0.9916 respectively.
- With the help of visualization, we have seen that from 2018 onwards there is sudden fall in the stock closing price. It makes sense how severly Rana Kapoor case fraud affected the price of Yes bank stocks.
- With the help of distribution plot, we see that our data is positively skewed.So, we apply some kind of transformation i.e., Log Transformation to convert it into a normal distribution.
- Lasso and Ridge regression models are giving the same result approximately.
- I have implements Cross Validation on different algorithm as CV performs better on small datasets. But the result is nearly same.
- In all the models except KNeigbour Regressor, the accuracy lies within the range of 81 to 83% and there is no such improvement in accuracy score even after hyperparameter tuning.
- Support Vector Regressor algorithm performs worst then other algorithm with accuracy of 81.2 %.
- KNR cross validation perform best with very less mean square error i.e., 0.015.