Capstone Project-II

Yes Bank Stock Closing Price Prediction

Team Members

Kaushal Kumar Jha
Shambhu Nath Jha
Nimesh Thakur
Asif PA

Points For Discussion

1. Problem Statement
2. Data Summary
3. Data cleaning
4. Exploratory Data Analysis
5. Correlation and VIF Analysis
6. Model Training
7. Conclusion

1. 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.
- We were interested to know that either ML model can predict the impact on stock price due to any fraud cases.
- So we have performed regression analysis to make future prediction using various machine learning models and selected one model as best model after comparing them using evaluation matrices and this model will be going to use for future prediction.

2.Data Summary

Data Set:- In Yes Bank Stock Price data set contains 185 rows and 5 columns which includes Date, Open, High, Low and Close.

Date:- The date of record finalizes the transfer of the stock's ownership

Open:- Open Price is the price at which the financial security opens in the market when trading begins. It may or may not be different from the previous day's closing price. The security may open at a higher price than the closing price due to excess demand of the security.

High:-High is the highest price at which a stock traded during a period.

Low:- Low is the lowest price at which stock traded during a period.

Close:- Closing price of stock is the price at which the share closes at the end of trading hours of the stock market.

3. Data Cleaning

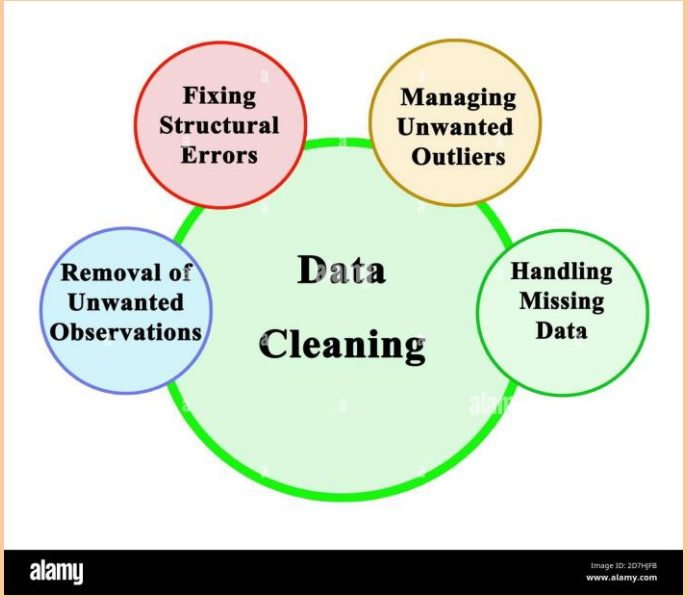
In this we performed:

1.Null value Treatment
2.Duplicate Treatment
3.Data Format change

1.Null Value Treatment : In given dataset no null value was there.

2. Duplicate Treatment :There was no duplicate data in dataset.

3.Data Format change: We have changed datatype of Date column from object to time64 format.



4.Exploratory Data Analysis(EDA)

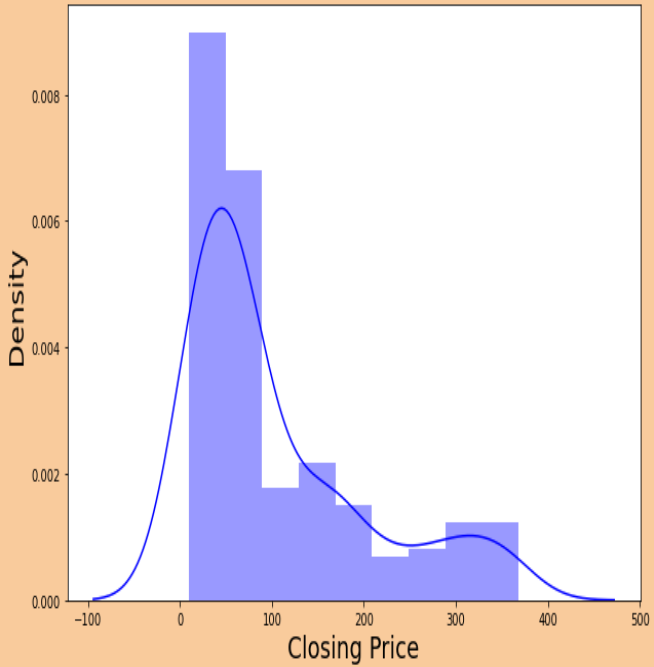
Univariate Analysis

From above graph it is clear that, trend was continuously increasing from 2009 till 2018. After 2018 it started to decrease due to fraud case of Rana Kapoor.

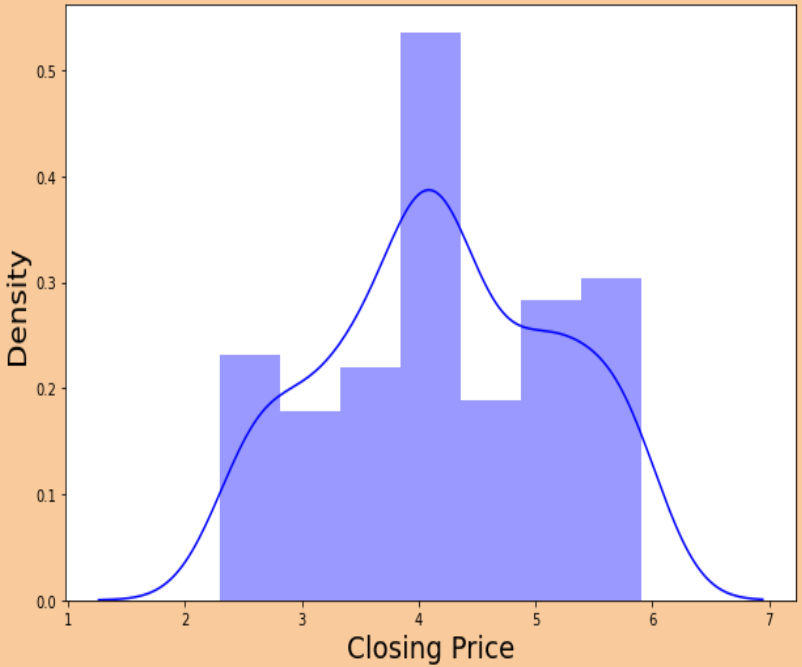
EDA continue..

Distribution of Close:- Distribution of target variable(Closing Price) was right skewed so we performed log transformation to make it normal

Before Log Transformation



After Transformation

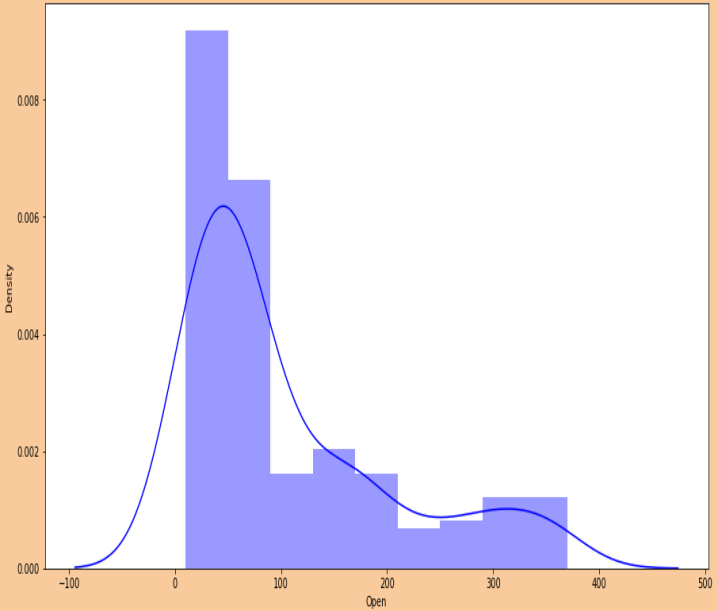


EDA Continue..

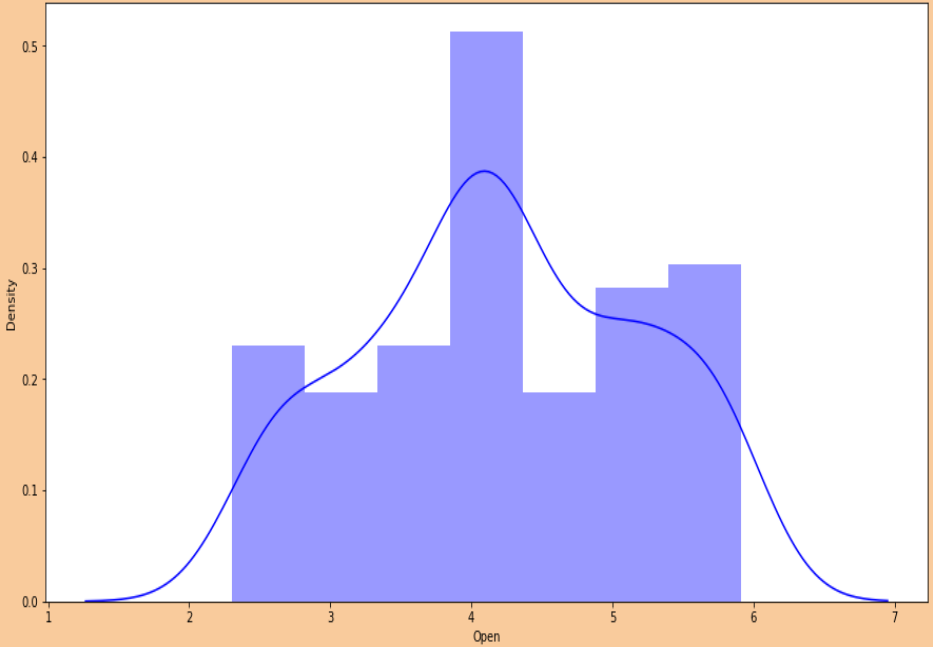
Distribution Of All Independent Variables(Open, High and Low)

Distribution of Open

Before Log Transformation



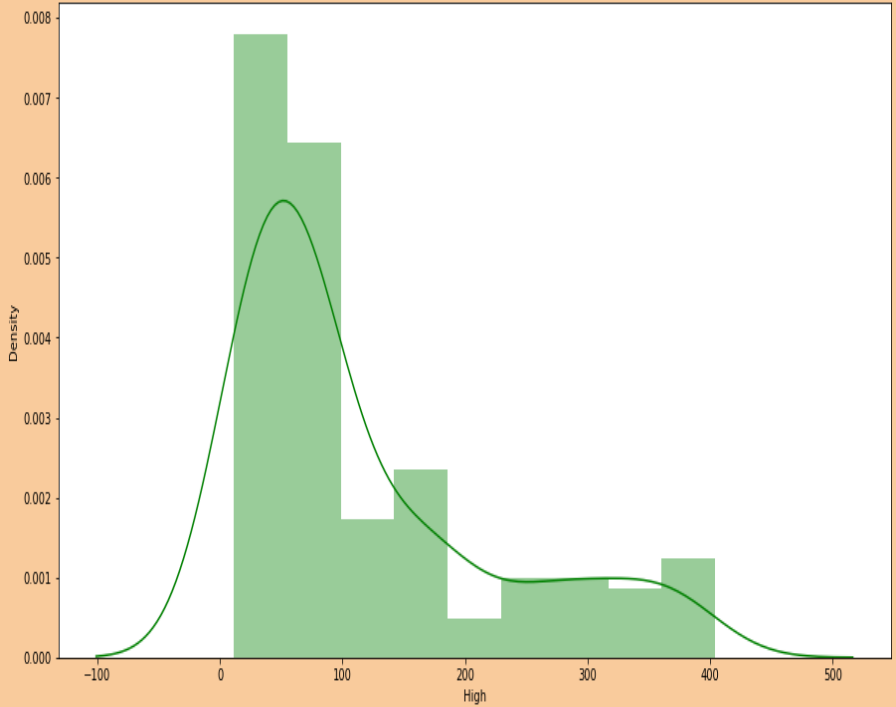
After Transformation



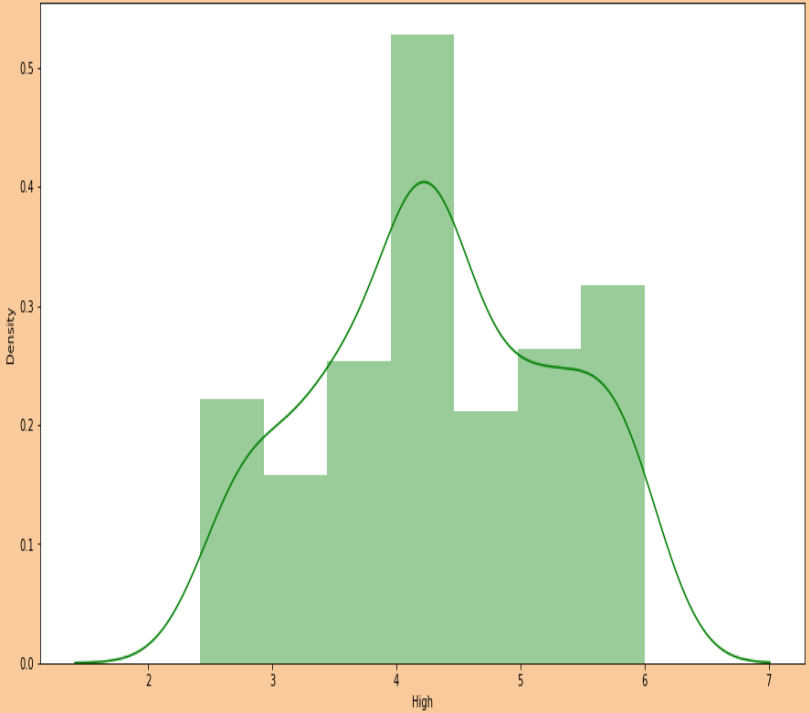
EDA continue…

Distribution of High

Before Transformation



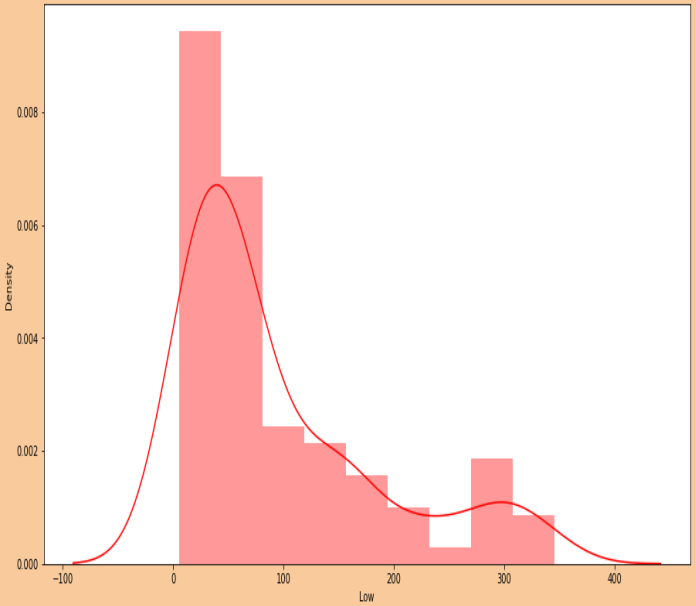
After Transformation



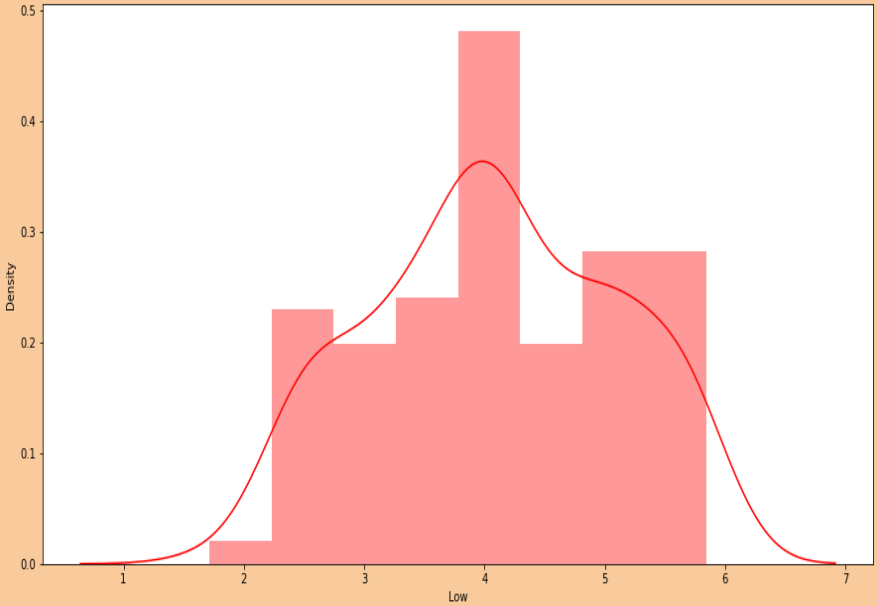
EDA continue…

Distribution of Low

Before Transformation



After Transformation

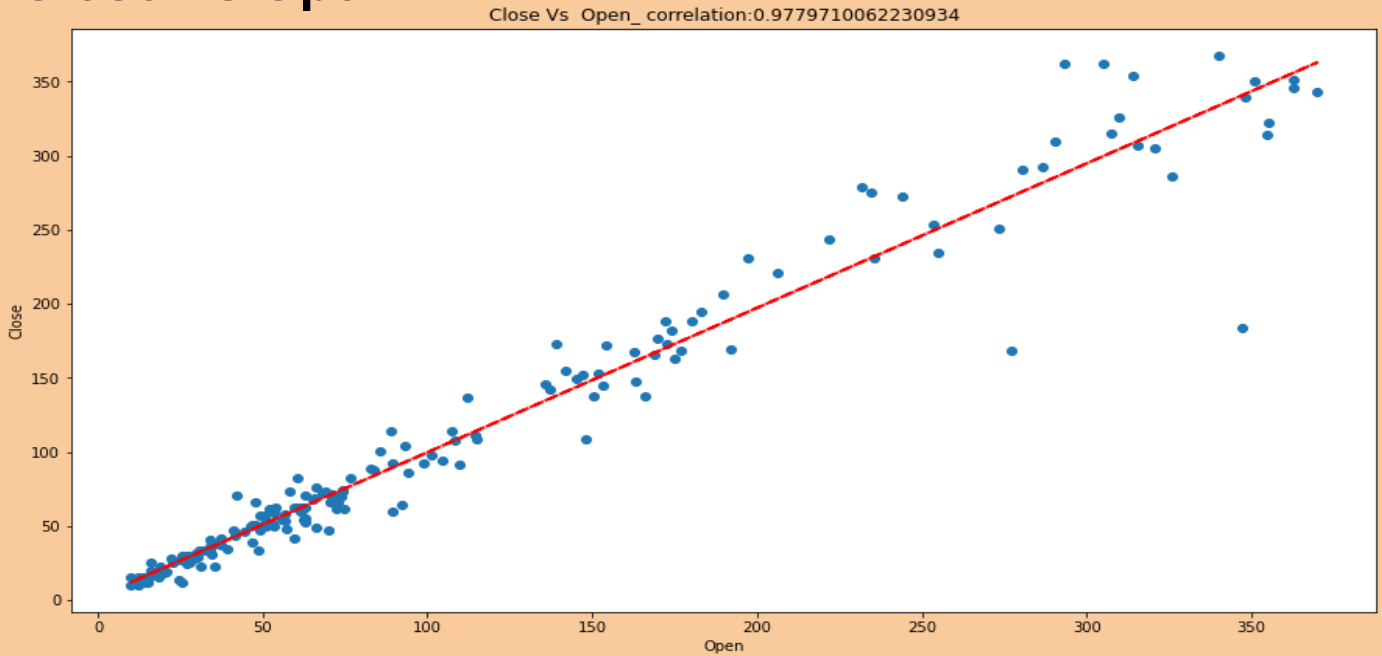


We have seen that all variables was right skewed so we have performed log transformation and after that all became normal.

EDA Continue…

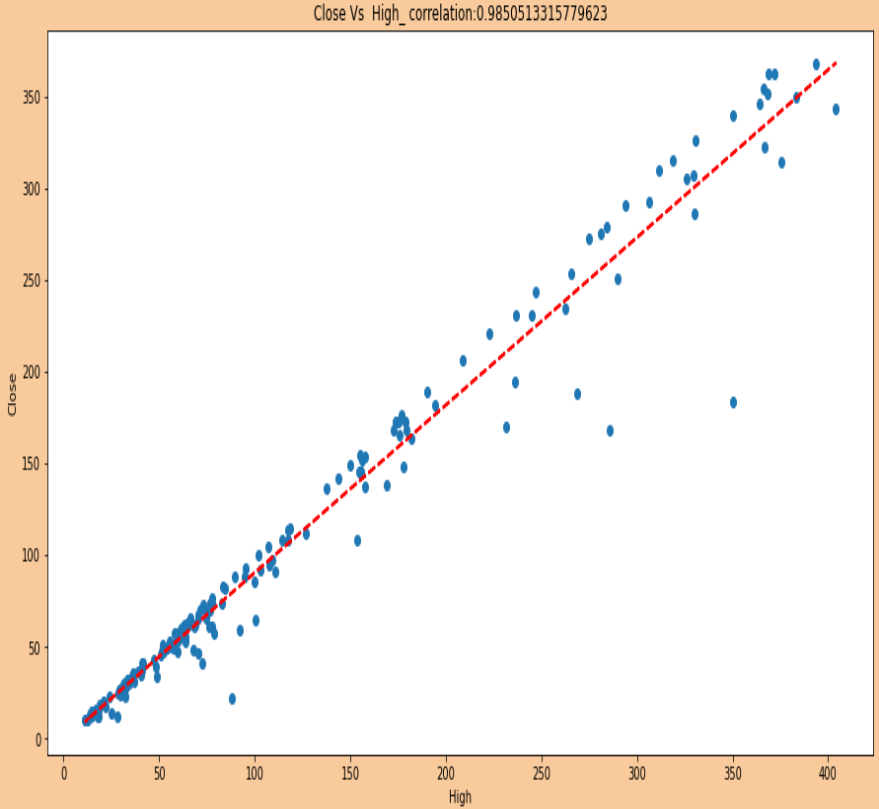
Bivariate Analysis: Now we will see the relationship between dependent and independent variable using scatter plot.

Close Vs Open

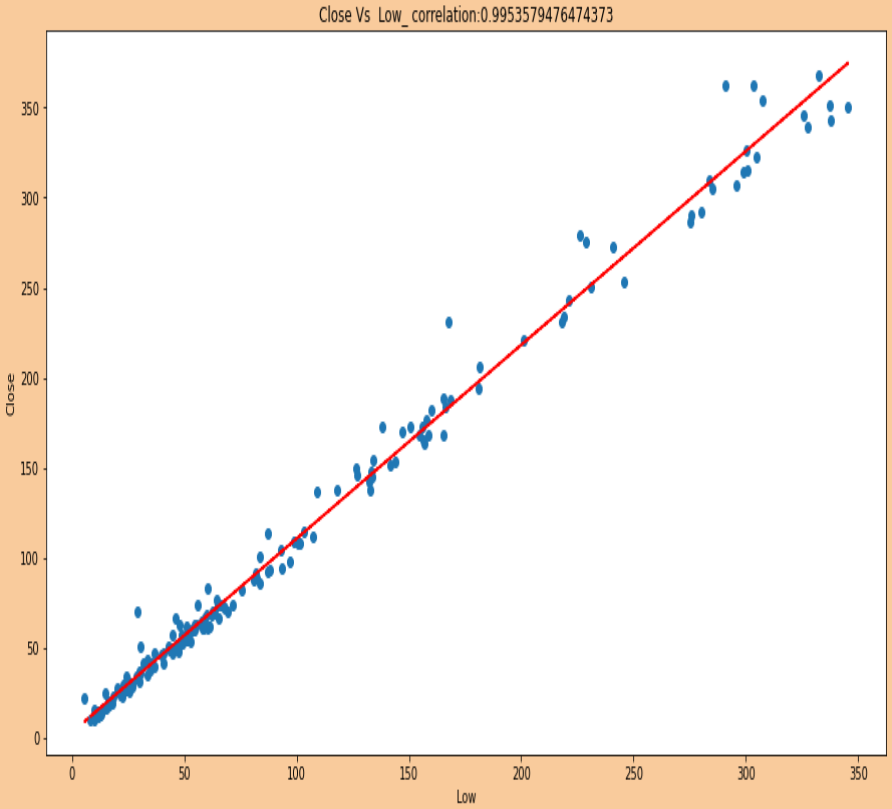


EDA Continue…

Close Vs High

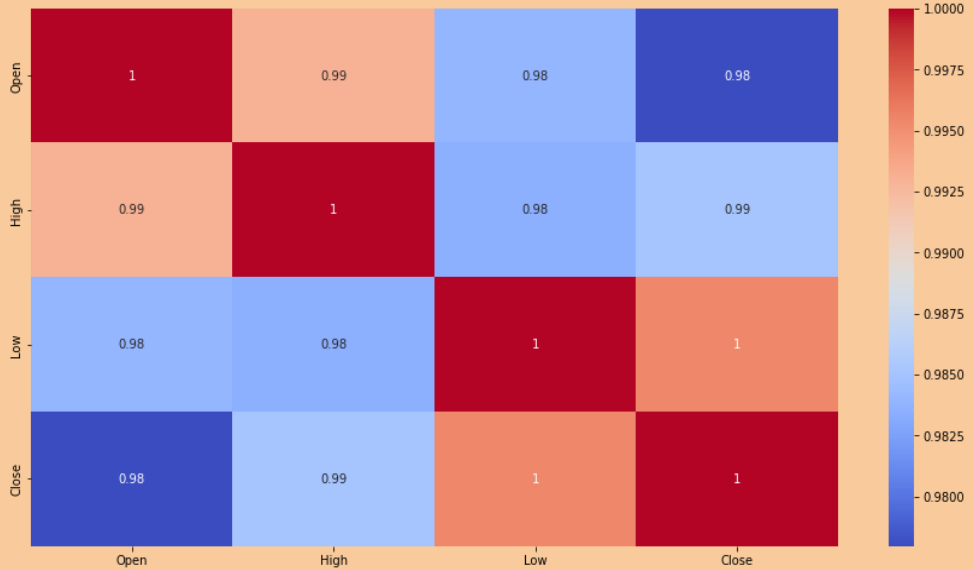


Close Vs Low



5.Correlation And VIF Analysis.

Correlation HeatMap



	variables	VIF
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

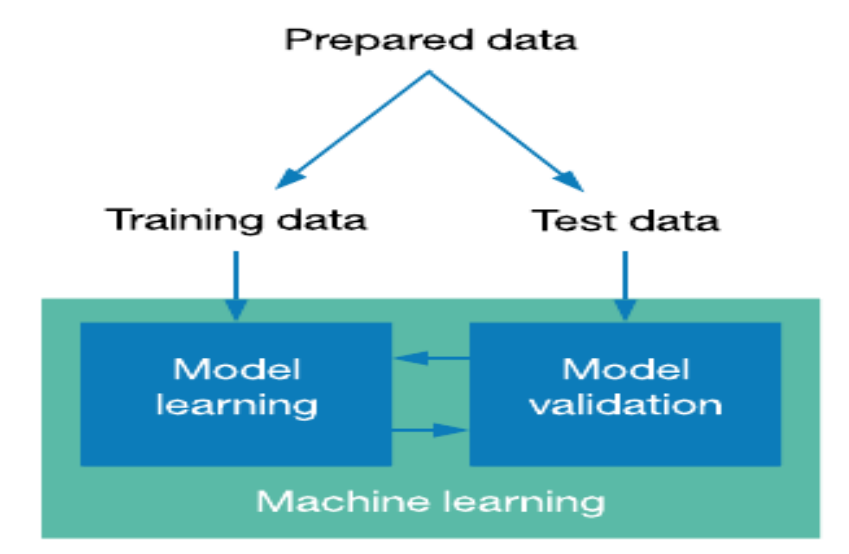
We can see that all variables are highly correlated and also we can see that VIF scores are very high for all our features which means there is a multicollinearity between our independent variables.

As we have limited features, dropping any one of them will result in loss of important data which are essential for accurate model prediction and it will result in bad model.

6.Model Training

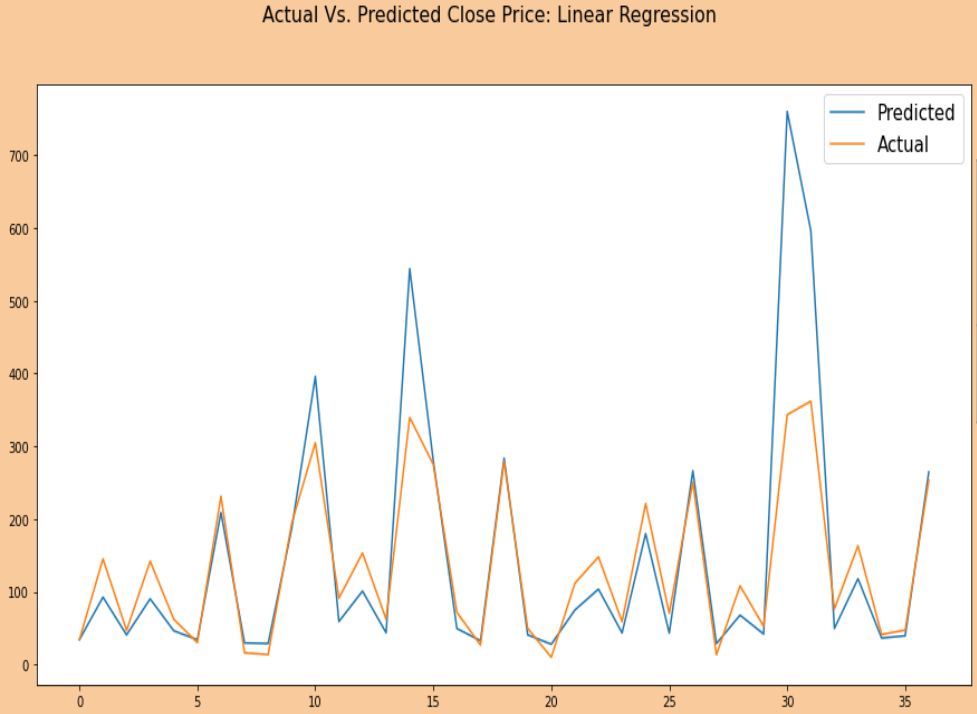
Train Test Split

- Train Test Split is used to split the dataset in Train and Test Sets.
- Train Set is used to train the model.
- Test set is used to test the performance of model.
- 80% of data is used to train the model and only 20% is used for testing purpose.



1.Linear Regression

- **Linear regression** is a popular and uncomplicated algorithm used in data science and machine learning.
- Linear regression shows a linear relationship between dependent and independent variables.

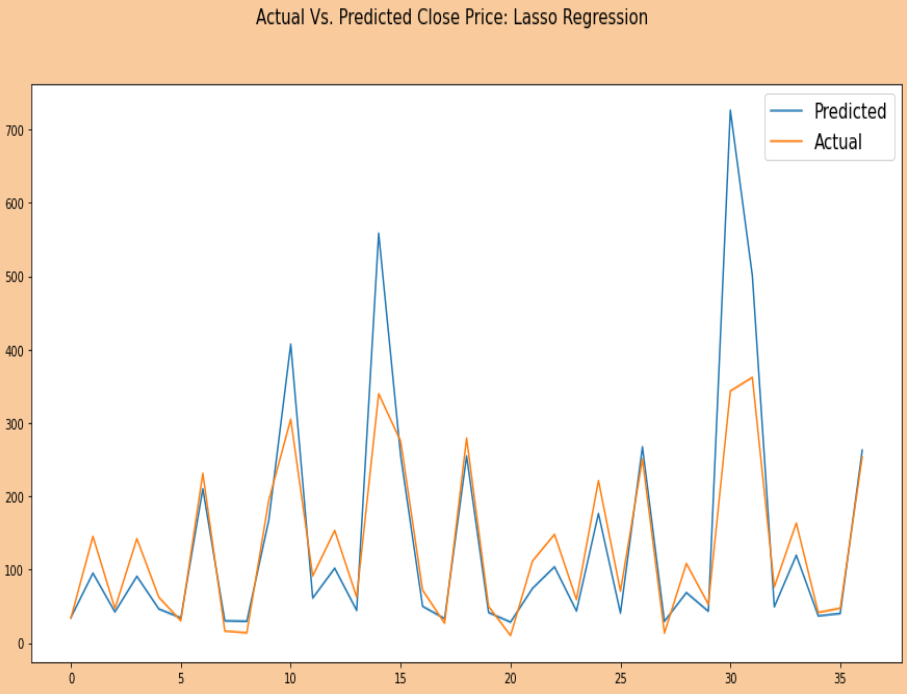


Evaluation Matric

MSE	RMSE	MAE	MAPE	R2 Score
0.032	0.1788	0.1457	0.087	0.8283

2. Lasso Regression

- It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of model.
- This regression perform L1 Regularization.



Evaluation Matrics

Before Cross Validation

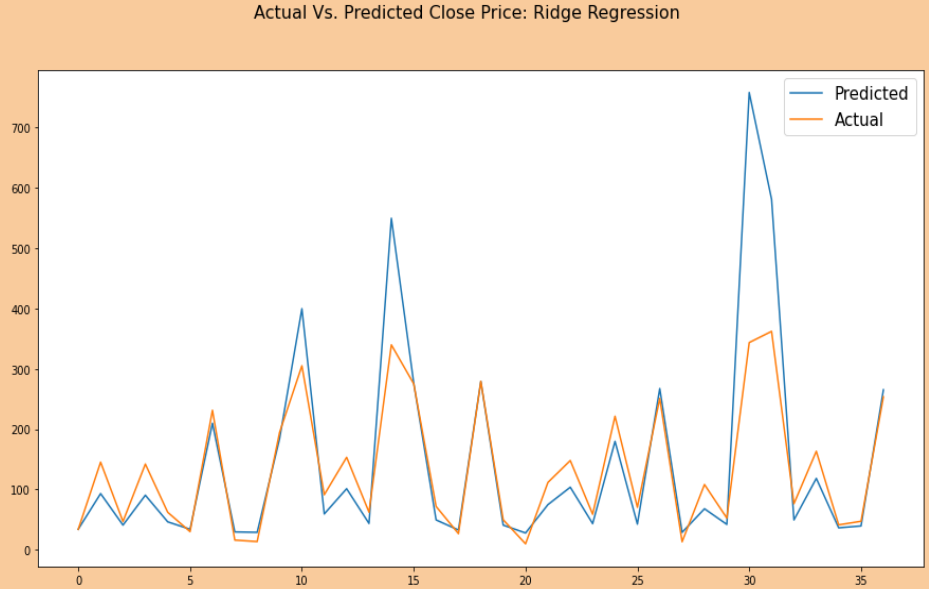
MSE	RMSE	MAE	MAPE	R2 Score
0.0316	0.1778	0.1463	0.0876	0.8303

After Cross Validation

MSE	RMSE	MAE	MAPE	R2 Score
0.0315	0.1775	0.1459	0.0877	0.8308

3.Ridge Regression

- Ridge regression is a model tuning method that is used to analyse 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 being far away from the actual values.



Evaluation Matrics

Before Cross Validation

MSE	RMSE	MAE	MAPE	R2 Score
0.0319	0.1786	0.1453	0.0869	0.8288

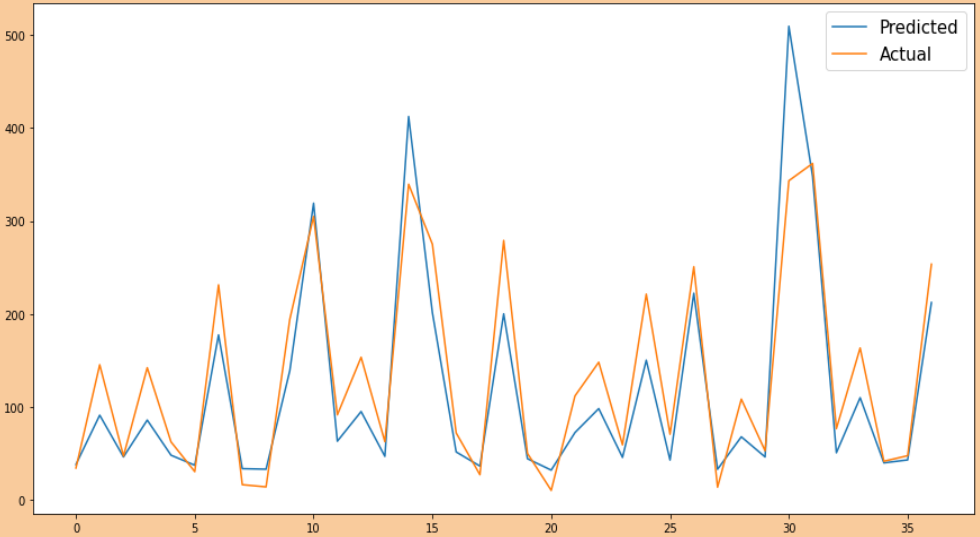
After Cross Validation

MSE	RMSE	MAE	MAPE	R2 Score
0.0317	0.1781	0.1464	0.0874	0.8298

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.

Actual Vs. Predicted Close Price: Elastic Net Regression



Evaluation Matrics

Before Cross Validation

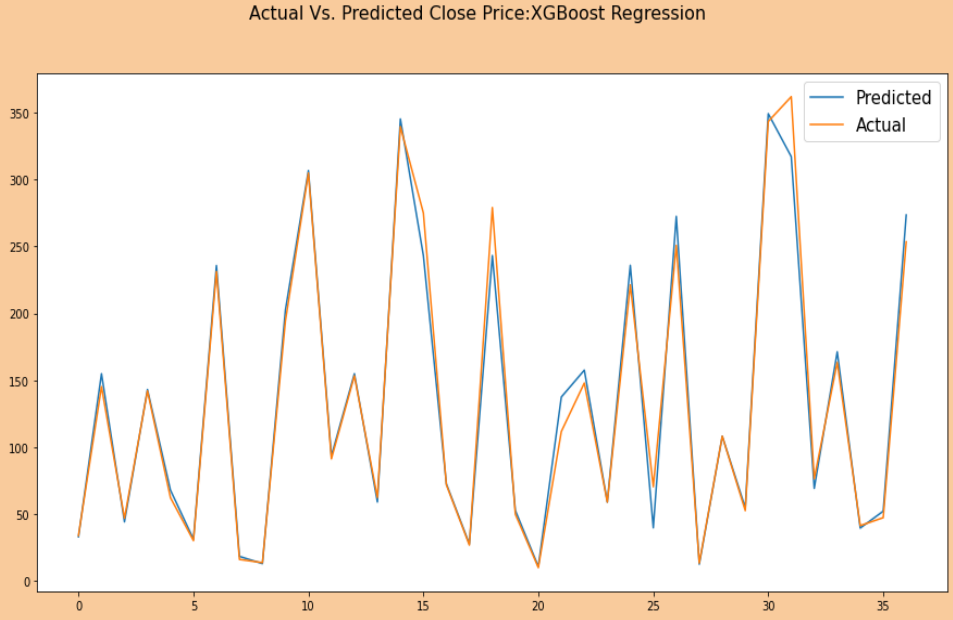
MSE	RMSE	MAE	MAPE	R2 Score
0.0344	0.1854	0.1514	0.0924	0.8155

After Cross Validation

MSE	RMSE	MAE	MAPE	R2 Score
0.0316	0.1778	0.1462	0.0876	0.8304

5.XG Boost Regression

- XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.
- It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.



EVALUATION MATRICS

MSE	RMSE	MAE	MAPE	R2 Score
0.0027	0.0518	0.0316	0.0174	0.9856

Evaluation Matric Of All Models

MODEL	MSE	RMSE	MAE	MAPE	R2 SCORE
Linear Regression	0.032	0.1788	0.1457	0.087	0.8283
Lasso Regression (CV)	0.0315	0.1775	0.1459	0.0877	0.8308
Ridge Regression (CV)	0.0317	0.1781	0.1464	0.0874	0.8298
Elastic Net Regression (CV)	0.0316	0.1778	0.1462	0.0876	0.8304
XG Boost Regression	0.0027	0.0518	0.0316	0.0174	0.9856

7.Conclusions

- ✓ Stock Price prediction with the help of Machine Learning models is less time consuming and also it gives good performance.
- ✓ Stock price was continuously increasing till 2018 after that it decreases due to fraud case of Rana Kapoor.
- ✓ All independent variables (Open, High& Low) are extremely correlated with dependent variable (Close).
- ✓ All independent variables are highly correlated with each other (Multicollinearity).
- ✓ Distribution of all independent and dependent variables was right skewed and after log transformation it became Normal.
- ✓ We have compared 5 models (**Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, and XG Boost Regression**) on the basis of **RMSE** and **MAPE**.
- ✓ **RMSE** and **MAPE** are mostly used as evaluation metrics to measure **forecast accuracy**.
- ✓ **XG Boost Regression** is best model among all five models with lowest **RMSE=0.0518**, and **MAPE=0.0174** than other models and also it has highest **r2 score** (**r2 score=0.9856**) than other models.


THANK YOU