Capstone Project

Yes Bank Stock Closing Price Prediction

by Sayan Bandopadhyay

Synopsis

'Yes Bank' is an Indian private sector bank headquartered in Mumbai, India and was founded in 2004. It offers a wide range of differentiated products for corporate and retail customers through retail banking and asset management services.

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.

So, this particular project is all about how different machine learning algorithm assisted to create models to predict the stock closing price of this particular bank.

Data Briefing

The dataset is comparatively a compact one in terms of size & also has fewer features or variables to work with.

- Dataset initially had 185 rows & 5 columns.

Now, regarding the extraction of the data information from the dataset, I used “info()” method which gave me the information about data types(dtypes), count of non-null values & memory usage by the dataset.

- There were 4 columns with float64 dtypes & 1 column with object dtype (object dtype was later converted into datetime dtype).
- No null values were present in the dataset.

To inspect the case of missing values in the dataset , I used “isna” method & found out that there were no missing values in the dataset.

Steps

Question framing (Based on problem statement)

↓

Data inspection & pre-processing

↓

Exploratory data analysis (EDA)

↓

Model implementation
(Linear Regression, Lasso Regression, Ridge Regression, KNN)

↓

Conclusions

Feature Overview

In this particular dataset, we have 5 features. They are :-

1. Date – Stores month & year data for each value of stock.
2. Open – Stores the price at which a stock started trading each month.
3. High – Stores the maximum price of the stock for each month.
4. Low – Stores the minimum price of the stock for each month.
5. Close – Stores the final trading price of the stocks for each month.

Independent or input variables → 'Open', 'High', 'Low'
Dependent or target variable → 'Close'

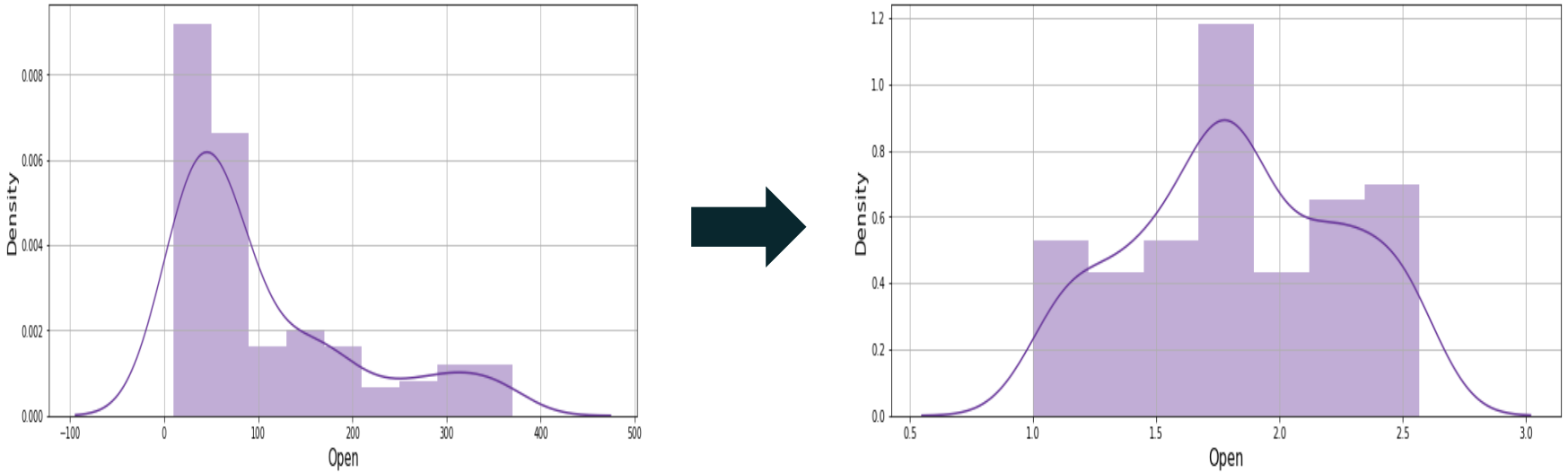
'Date' is only useful for EDA purpose & do not have any influence for closing price prediction.

Analysis Report
&
Model Performance

Stock Price Fluctuations

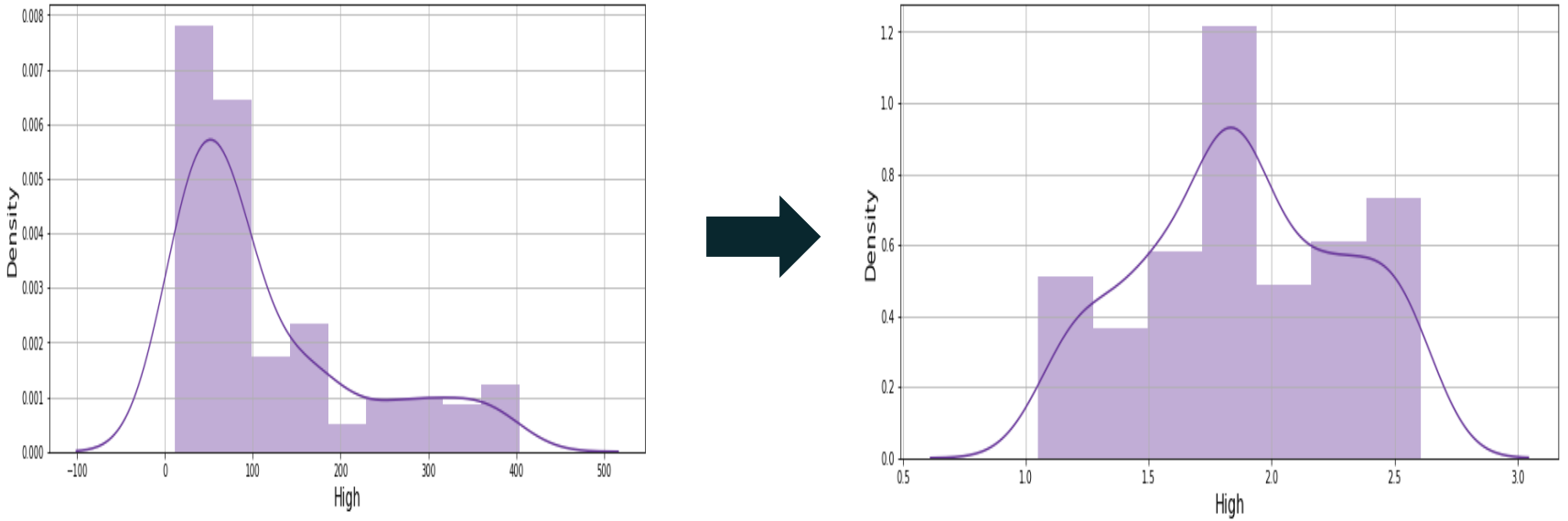
Here, its clearly visible from the above plot that the stock prices saw a significant rise from year 2006 to 2018 . However, since 2018 the stock prices saw a major downfall and that is may be due to the fraud case.

Univariate Analysis



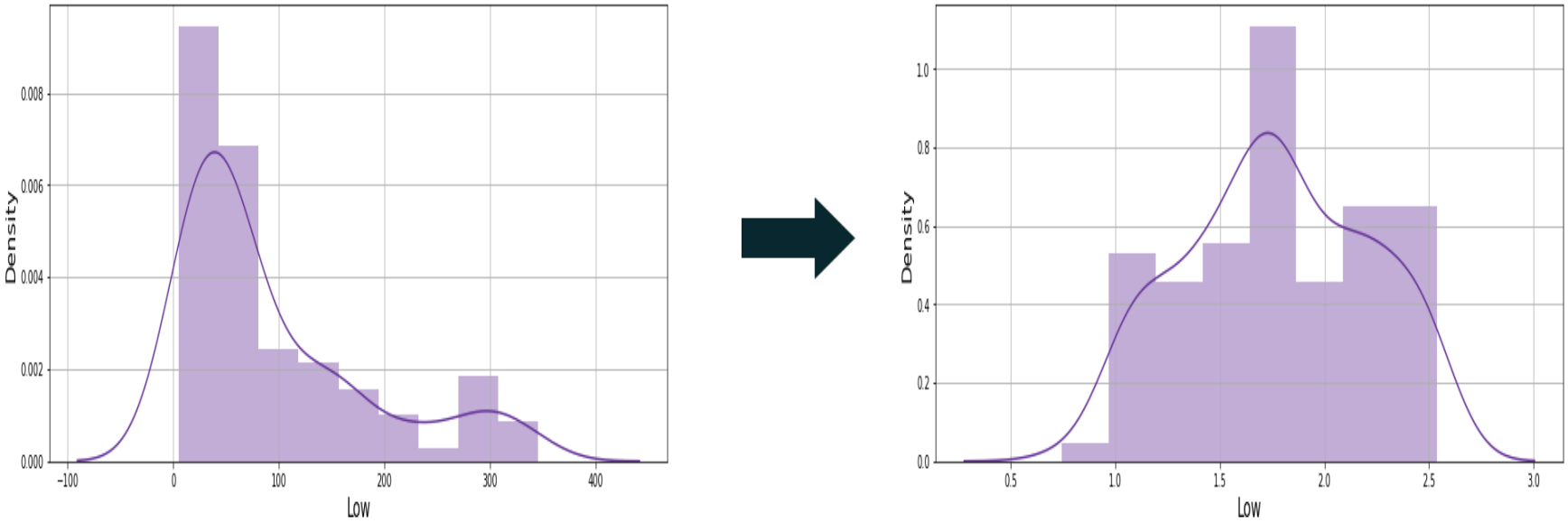
On visualising the first independent variable (Open) on the left hand side, I found out that its right skewed and upon applying log transformation, the plot is transformed into a normally distributed variable to some extent which is clearly visible in the right plot.

Continued....



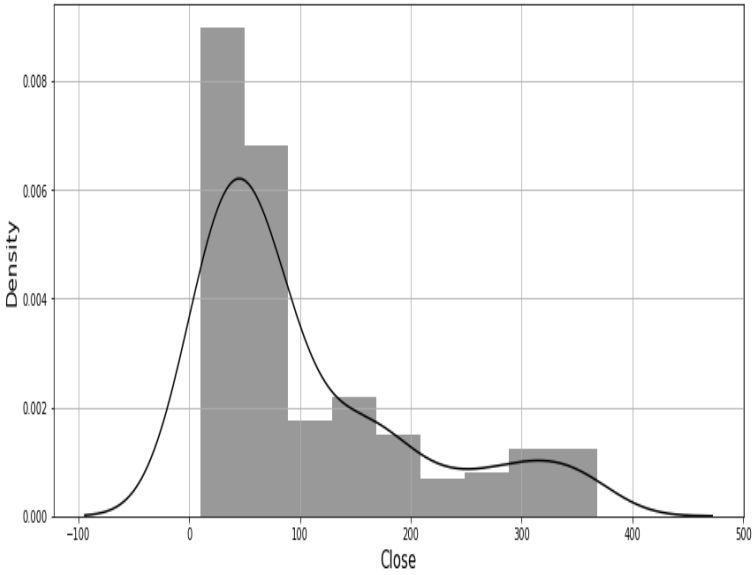
The second independent variable (High) was also right skewed which can be seen in the left hand plot and upon applying log transformation the variable got normally distributed to some extent which the right hand plot clearly depicts.

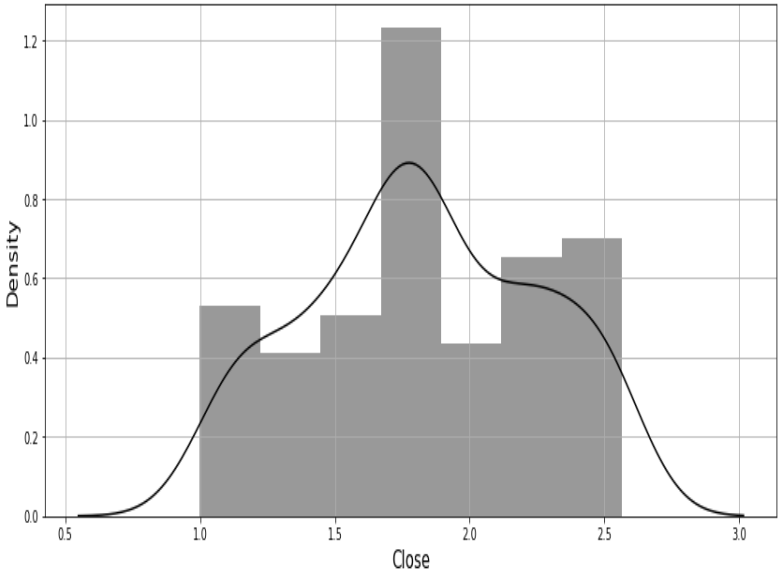
Continued....



Same as other two independent variables above, this independent variable (Low) was also right skewed and can be visualised by the left hand side plot. Applied log transformation and this particular variable also got normally distributed to some extent as shown in the right hand plot.

Continued....





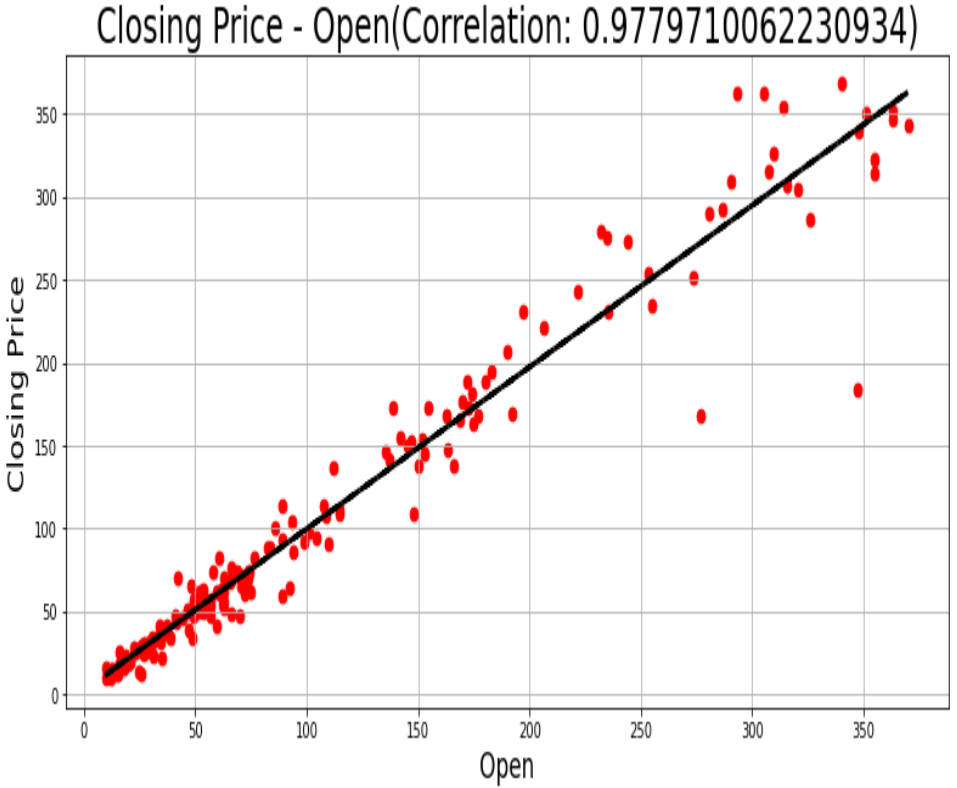
Same is the case for dependent variable (Close). Left hand side plot represents the right skewed dependent variable and right hand side plot represents normally distributed dependent variable after log transformation.

Bivariate Analysis

Plot on the right shows the correlation between stock opening price & stock closing price.

Here, a linear trend is clearly visible depicted by the black straight line which shows a positive correlation between two variables.

The two variables are almost 97% correlated to each other.

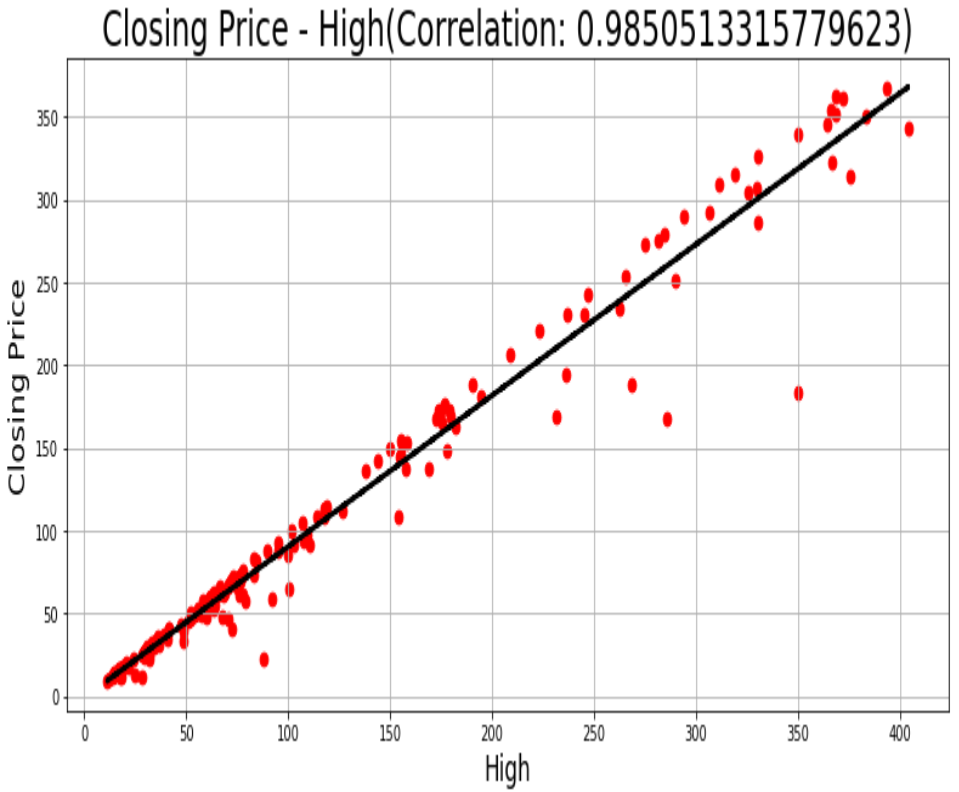


Continued....

This plot shows the correlation between high values of stocks & stock closing price.

Here also a linear trend is clearly visible between these two variables by the black straight line which shows a positive correlation between the two variables.

The two variables are almost 98% correlated to each other.

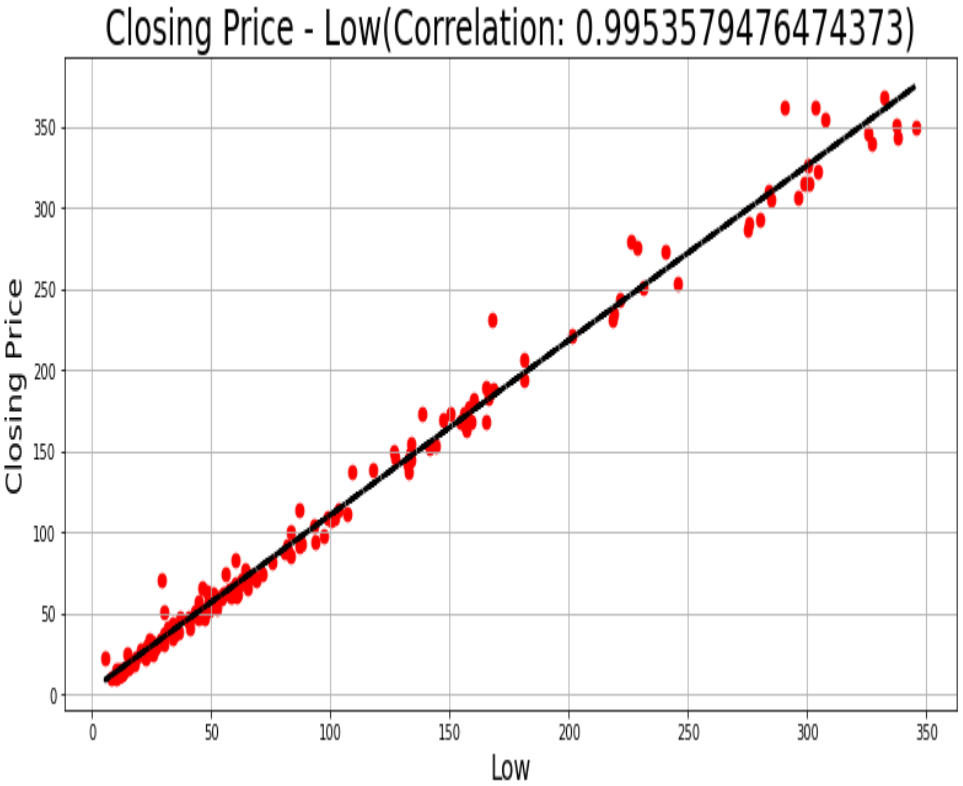


Continued....

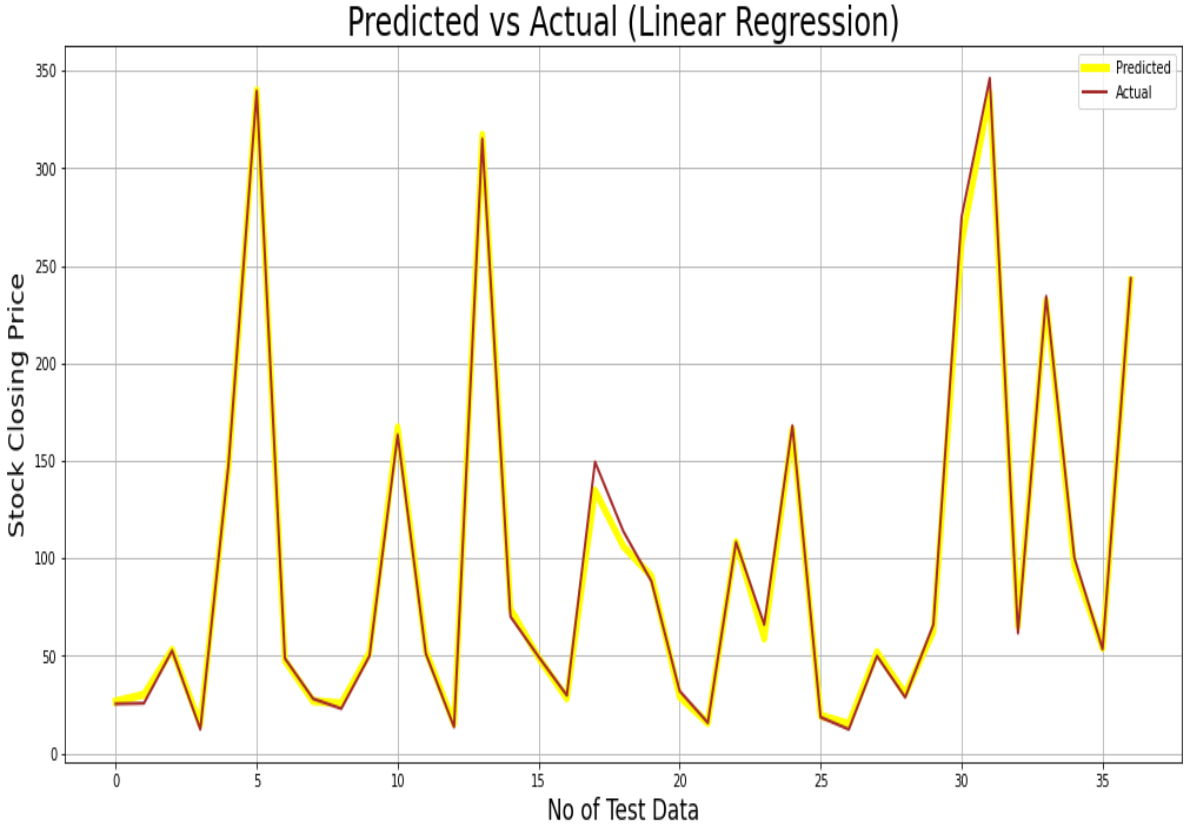
This plot shows the correlation between low values of stocks & stock closing price.

Same as above cases, a linear trend is visible by the black line which means that the two variables are positively correlated to each other.

The two variables are almost 99% correlated to each other.

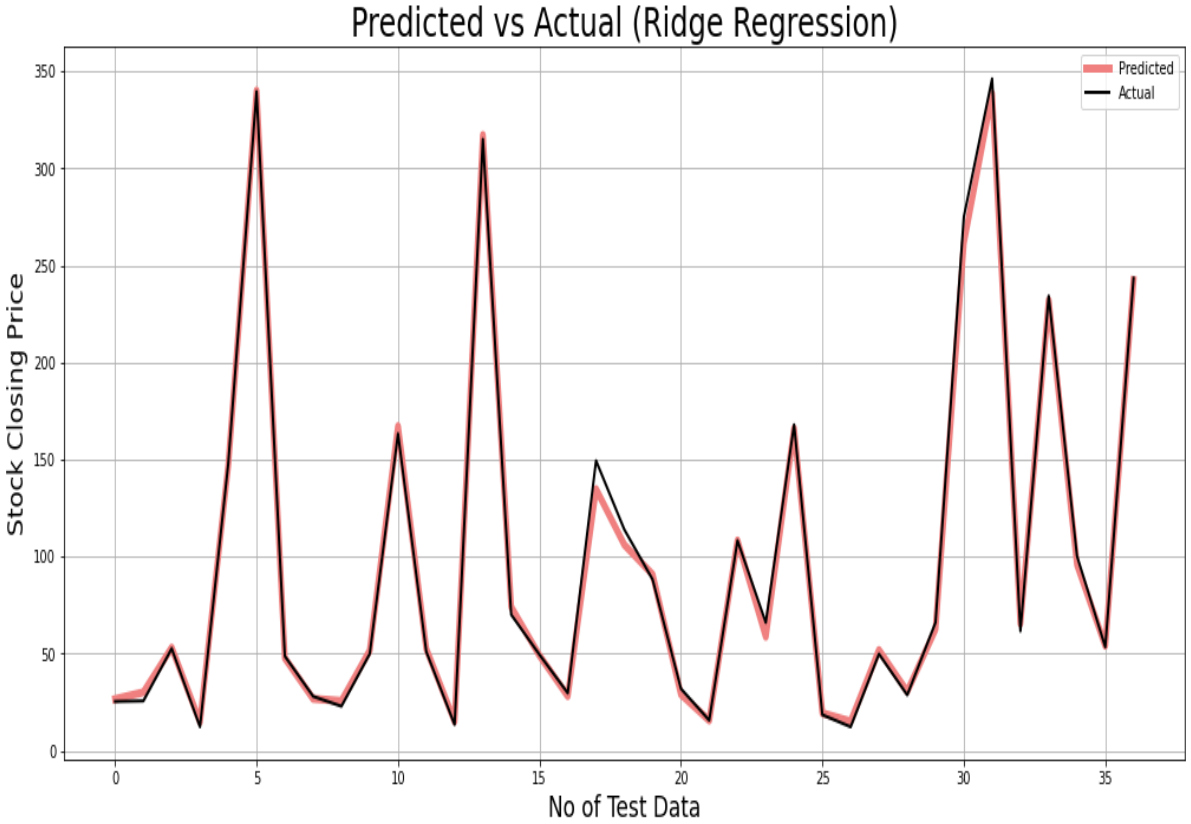


Linear Regression



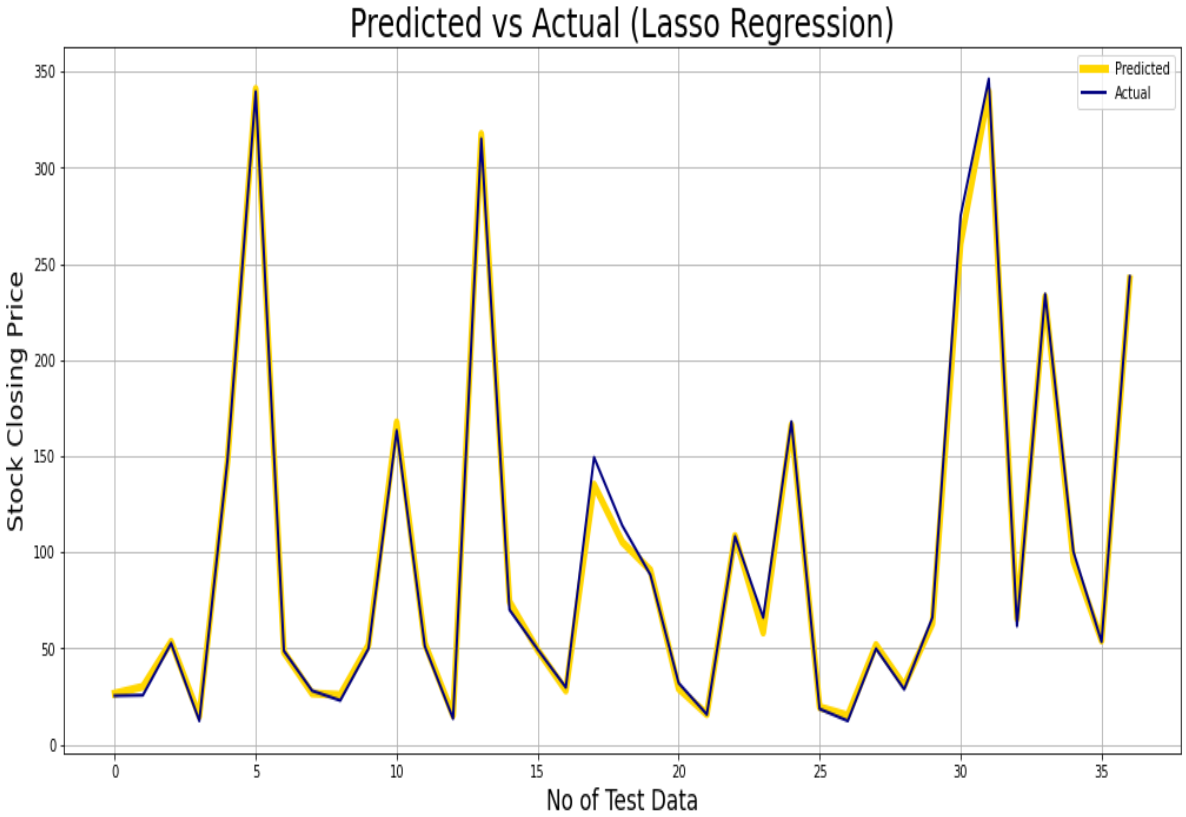
MAE : 3.0527647129646436
MSE : 19.988578593595
RMSE : 4.470858820584139
MAPE : 0.05404201340281821

Ridge Regression



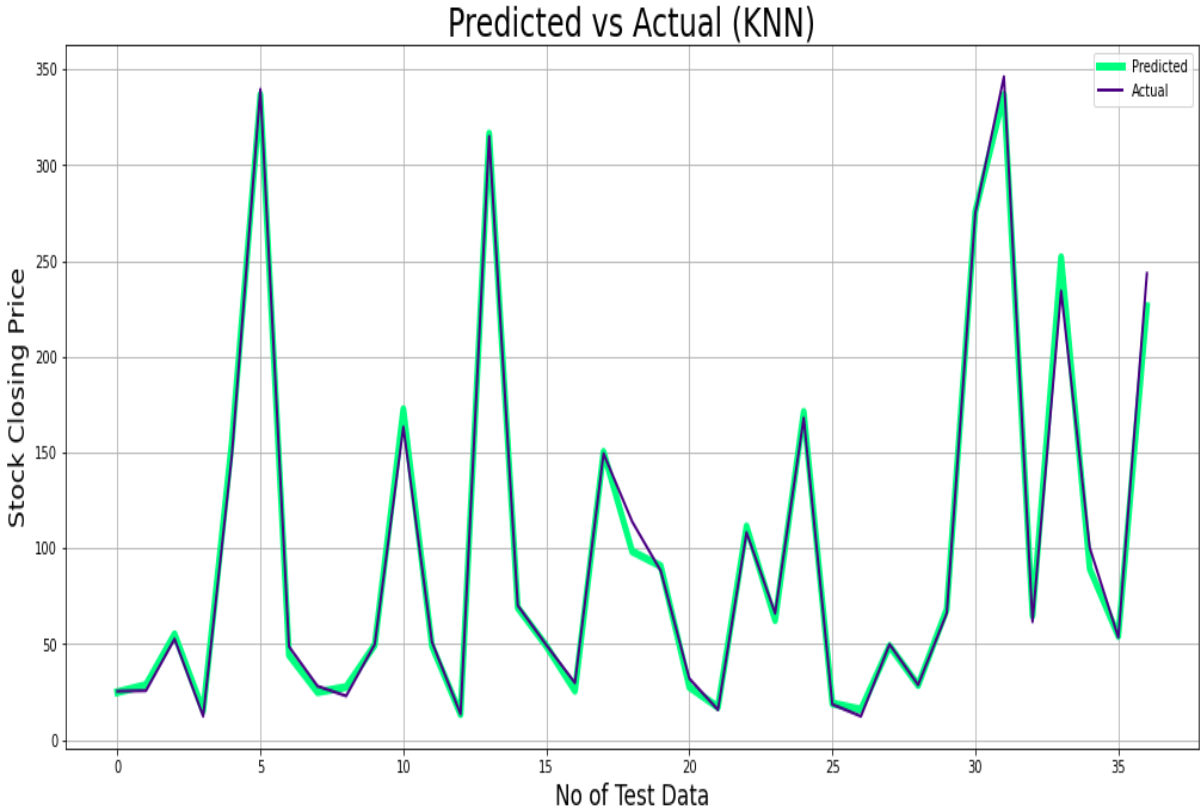
MAE : 3.0610224606861545
MSE : 20.095425485603723
RMSE : 4.482792152844444
MAPE : 0.0541673710982311

Lasso Regression



MAE : 3.132713292578181
MSE : 20.959149340321172
RMSE : 4.578116352859675
MAPE : 0.0553967482033884

KNN



MAE : 4.104324324324326
MSE : 37.72864864864864
RMSE : 6.142365069633084
MAPE : 0.06440972828840637

Metrics Comparison

Model Name	MAE	MSE	RMSE	MAPE
Linear regression	3.0528	19.9886	4.4709	0.0540
Ridge regression	3.0610	20.0954	4.4828	0.0542
Lasso regression	3.1327	20.9591	4.5781	0.0554
KNN	4.1043	37.7286	6.1424	0.0644

Above table contains the evaluation metrics for each model implemented in this project sorted in ascending order.

'Linear Regression' model has got lowest values for MAE, MSE, RMSE & MAPE.
'KNN' model has got the highest values for MAE, MSE, RMSE & MAPE.

Conclusions

- Independent variables(input variable) have a very high influence on dependent variable(target variable).
- Stock prices saw a downfall after 2018 due to the fraud case.
- The accuracy for each model is more than 95%.
- In this case, Linear Regression has given the best results with lowest MAE, MSE, RMSE and MAPE scores out of all.
- As we know that, lasso regression automatically selects only those features that are useful and hence discarded some features when applied in this case, whereas in this case all the features were important for prediction purpose so it ended up giving poor results.
- KNN performed the worst out of all.

Thank You