Capstone Project

Yes Bank Stock Closing Price Prediction

Submitted by - Swapnil Wankhede

Outline of presentation

- Problem Statement
- Introduction
- Exploratory Data Analysis
- Forming assumptions and obtaining insights
- Feature engineering and data pre-processing
- Model Implementation
- Comparison among implemented models using performance metrics
- Conclusion
- References

Problem Statement

Yes Bank is a well-known bank in the Indian financial domain, headquartered in Mumbai, India and was founded by Rana Kapoor and Ashok Kapoor in 2004. 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 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.

Introduction

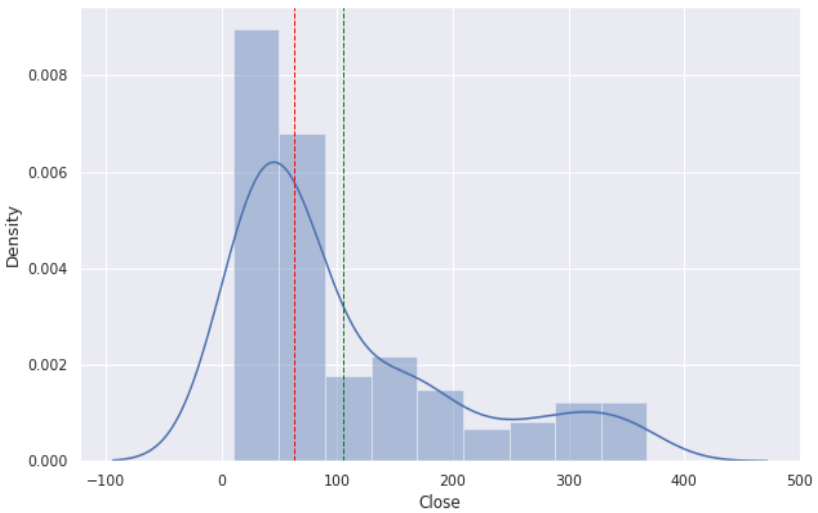
The given dataset has 185 observation with no missing values and no duplicates rows. The closing price is to be considered as dependent variable and rest of the variable are predictor variables. The objective is to predict the stock's closing price of the month and to deliver insights that-

- How news impacted the stock closing price by using line graph.
- Is there significant linear relationship between closing price of n^{th} day and opening price of $(n+1)^{th}$ day using parametric t-test.
- Which is the best implemented model by observing the performance metrics.

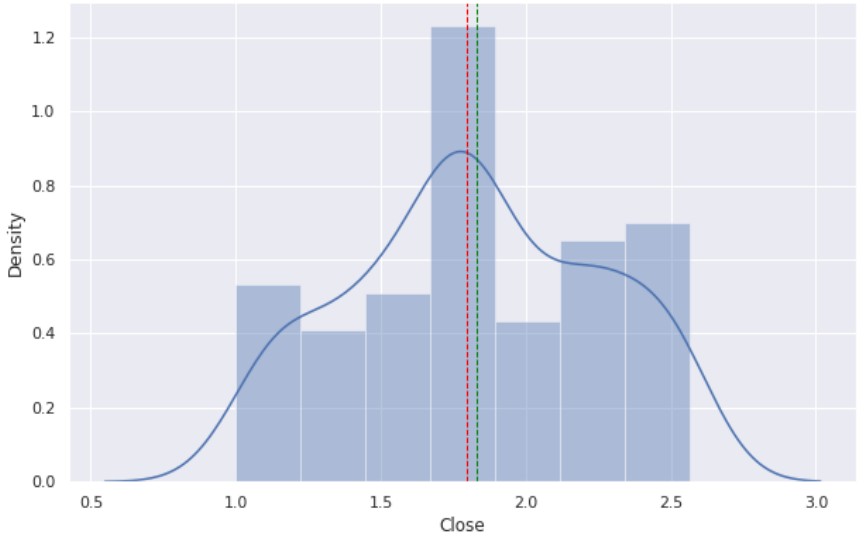
Many more things.

Exploratory Data Analysis:

Histogram with bell shaped curve



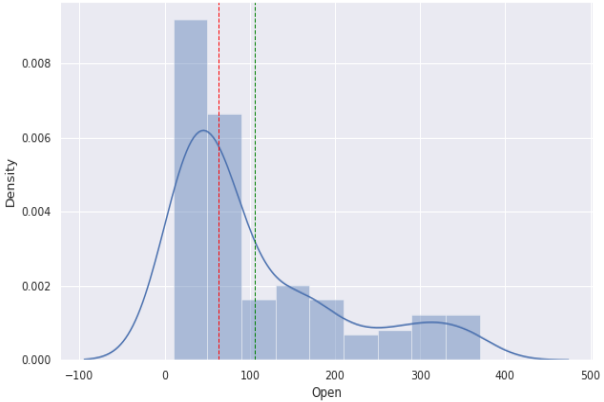
Before log Transformation

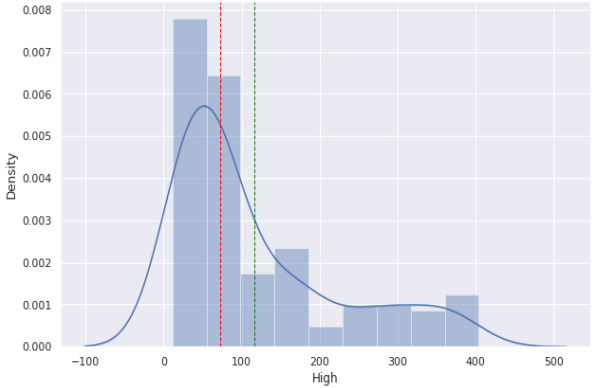


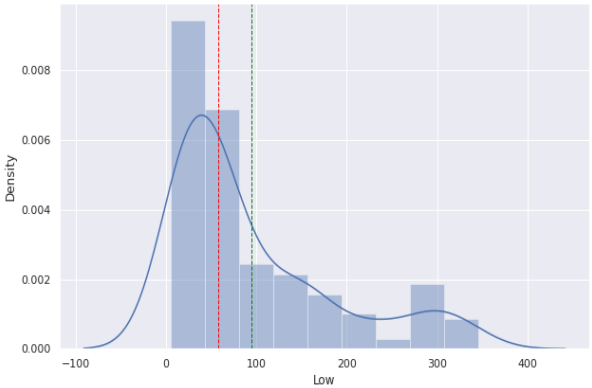
After log Transformation

It is observed that how mean is sensitive to outliers as the dependent variable is positively skewed.

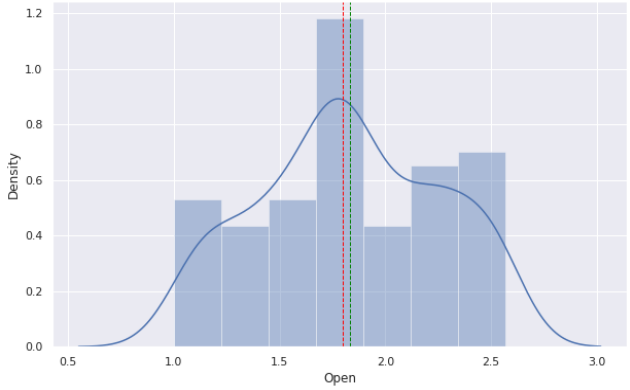
After applying log transformation, dependent variable looks normally distributed.

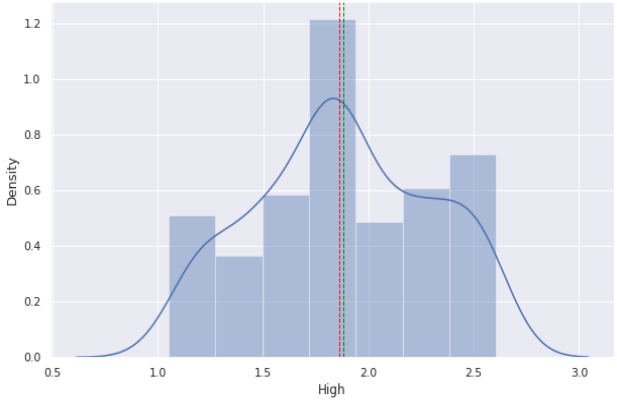


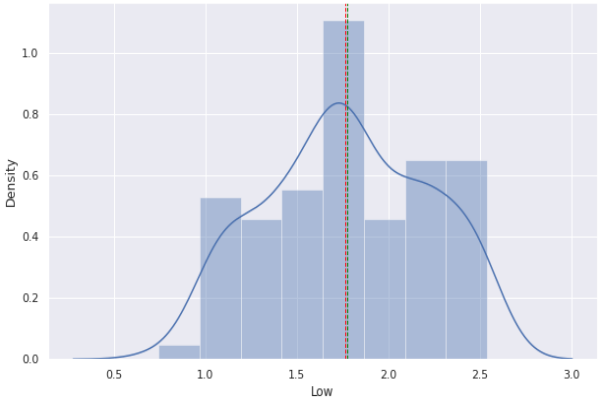




Before log Transformation

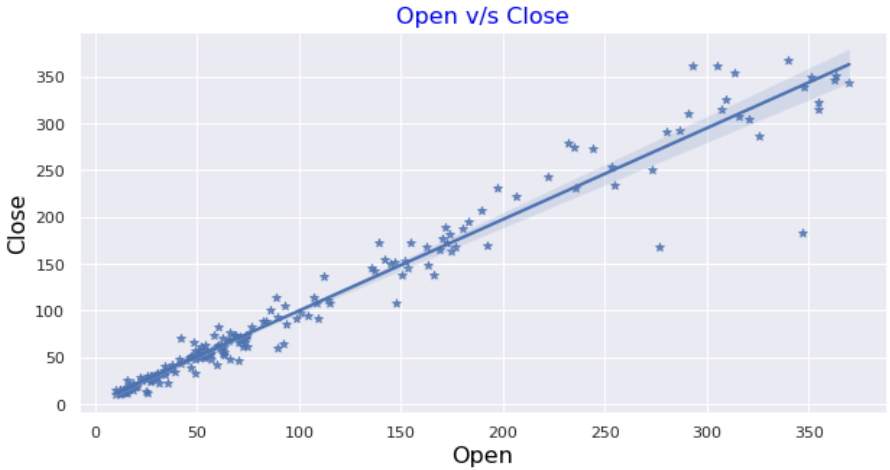






After log Transformation

Scatter plots:



Before log Transformation

After log Transformation

- Variance around the regression line is same at origin sided portion after that it is showing Heteroscedasticity.
- After the transformation, It means that the variance around the regression line is the same.

Forming assumptions and obtaining insights:

01) The visualization of dependent variable, ‘Closing price’ demonstrates how closing price varies with each passing year.

It is observed that since inception 2004, Closing price of the stock started smoothly increasing up to the year 2014. After this, it was observed that closing price started increasing exponentially up to the year 2018 which was the peak price but since 2018, it had been in the news because of the fraud case involving the co-founder Rana Kapoor. This news directly impacted the stock price as a result the closing price fall down rapidly and come to zero at year 2020.

02) Let's us perform the hypothesis test on the close pricing of n^{th} day and open pricing of $(n+1)^{th}$ day.

H_0: $\rho = 0$ v/s

H_1: $\rho \neq 0$ (two tailed hypo) Where, ρ is correlation coefficient. $-1 \leq \rho \leq 1$

The test statistic given that $\rho = 0.98$ from heatmap matrix using the following formula-

$t = \frac{\rho\sqrt{(n-2)}}{\sqrt{(1-\rho^2)}}$ = 66.6199 with 183 degree of freedom

we get the p- value = 0.0000.

It observed that p-value is small than 0.05 (level of significance). We can say that, we reject the null hypothesis.

There is sufficient statistical evidence at the 0.05 level of significance to conclude that there is a significant linear relationship between closing price n^{th} day and opening price of n+1 th day.

Feature engineering and data pre-processing:

- A Variance Inflation Factor (VIF) is measure of the amount of multicollinearity in regression analysis. It is found that VIF of independent features are very high. However, the given dataset contains three independent variables so we can ignore the multicollinearity as dropping or merging of variables leads to loss of information.
- Scaling the data is very important for us so as to avoid giving more importance to features with large values. This is achieved by normalization or standardization of the data.

Model Implementation:

01) Multiple Linear regression:

The linear regression model represents the response variable (y) as a function of open, high and low as are the predictor variables in given dataset is -

$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

β_1, β_2, β_3 are parameters to be estimated for X_1, X_2, and X_3 predictor variables.

All the assumption of multiple linear regression are satisfied. The hypothesis testing is applied to the significance of regression coefficients.

To test the significance of regression coefficients in linear regression models.

H_0: $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$ (Regression model does not exist)

H_1: $\hat{\beta}_1 \neq \hat{\beta}_2 \neq \hat{\beta}_3 \neq 0$ (Regression model does exist)

The F-statistics is defined as-

F = (SSR/DFssr) / (SSE/DFsse)

DFssr = Degree of freedom for regression model p = 4

DFsse = Degree of freedom for error; 181

The Tabulated F-Statistics of 185 observation is F_{tab} (3, 181) = 2.60 (from F_ Distribution tables used in analysis of variance).

F_{Cal} = 57.6666 (approx.)

Here, $F_{cal} > F_{tab}$ (3, 181) so we reject the null hypothesis with 5 % level of significance.

This model predicted the closing price with root mean squared error of 8.201. R^2 score of this model is 0.9935 which tells us that around 99% of variance in dependent variable is explained by independent variable. The graph is showing how the actual and predicted closing price of stock fit to testing data.

02) Ridge Regression

This model predicted the closing price with root mean squared error of 8.1815. R^2 score of this model is 0.9935 which tells us that around 99% of variance in dependent variable is explained by independent variable. The graph is showing how the actual and predicted closing price of stock fit to testing data.

03) LASSO Regression

Lasso regression model predicted the closing price with root mean squared error of 8.1901. R^2 score of this model is 0.9935 Which tells us that around 99% percent of the variance in dependent variable is explained by independent variable. Adjusted R^2 score has the value 0.993. The graph is showing how the actual and predicted closing price of stock fit to testing data.

04) Elastic Net Regression

This model predicted the closing price with root mean squared error of 8.1571. R^2 score of this model is 0.9935 which tells us that around 99% of variance in dependent variable is explained by independent variable. The graph is showing how the actual and predicted closing price of stock fit to testing data.

Comparison among implemented models using performance metrics

Let’s apply the test of goodness of fit.

To test the hypothesis,

H_0: The fit is good that means how well models fit a given testing set of data. v/s

H_1: The fit is not good that means models don't fit a given testing set of data.

It is observed that adjusted R^2 score of all models is greater than 0.5 (reference score)

so we accept the null hypothesis with 5% level of significance(LOS) i.e. the fit is good. It means that models fits a given testing set of data.

We came to conclusion that All models fits good. 99% variability in dependent variable "Close" explained by independent variables. The following graph is showing how the actual and predicted closing price of stock fit to testing data by implementing the various models.

All models fit good. However, It is found that Elastic net regression is the best performing model with adjusted R squared is 0.9931.

	Metric	Linear	LASSO	Ridge	Elastic Net
1	MSE	67.2576	67.087	66.938	66.641
2	RMSE	8.2010	8.1907	8.1815	8.1634
3	MAE	4.9777	4.9844	4.9906	5.0059
4	R^2score	0.9935	0.9935	0.9935	0.9935
5	Adj. R^2 score	0.9930	0.9930	0.9931	0.9931

Conclusion

1. Given dataset does not have missing values and duplicates values.

2. The closing price of stock is considered as dependent features whereas opening price, lowest price and highest price of stock are independent features.

3. Since 2018, yes bank had been in the news because of the fraud case involving the co-founder Rana Kapoor. This news directly impacted the stock price as a result the closing price fall down rapidly and come to zero at year 2020.

4. Independent variables such as Low, Open, High are shown the linearity with dependent variable Close.

5. There is sufficient statistical evidence at the $\alpha = 0.05$ level to conclude that there is a significant linear relationship between n^{th} day closing price and $(n+1)^{th}$ of the opening price.

6. All the features are positively skewed distributed. Mean is greater than Median i.e Mean > Media.

7. After the log transformation, Distribution of features are similar to normal distribution. The mean and median values are nearly same. It diminishes the outlier's effect and heteroscedasticity.

8. In multiple linear regression, $F_{cal} > F_{tab}(3,181)$ so we reject the null hypothesis with 5 % level of significance. we conclude that the close price of stock is dependent of Highest, Lowest and Opening price of stock i.e. $\hat{\beta}1 \neq \hat{\beta}2 \neq \hat{\beta}3 \neq 0$. It may be possible that estimate parameters are losing preciseness as multicollinearity is considered.

9. The fit is good at LOS 5%. It means that all models fits a given testing set of data.

10. Several models are implemented on the given dataset in order to predict the closing price and found that Elastic net regression is the best performing model with adjusted R squared is 0.9931.

References:

1. https://grow.almabetter.com/data-science/learn/full-stack-data-science/cp-supervised-ml-regression/Read

2. medium.com

3. towards.datascience.com

THANK YOU!