Capstone Project - 2

Yes Bank Stock Closing Price Prediction

Name: Rudrajit Bhattacharyya
Cohort: Zanskar Pro

What's inside?

1. Defining the problem statement
2. Defining the data
3. Data Cleaning/Pre-processing
4. EDA
    a. Trend of the Stock's closing and opening prices
    b. What's the all time high and all time low of the stock
    c. Distribution of all the variables
    d. How the features are related to each other
5. Data Transformation/Preparation
6. Defining Evaluation Metrics
7. Building Regression Models
8. Evaluating and Comparing all the Models
9. Conclusion

Defining the 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 185 records of 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.

What happened to Yes Bank?

Yes Bank is an Indian bank headquartered in Mumbai, India and was founded by Rana Kapoor and Ashok Kapur in 2004. It offers a wide range of differentiated products and services for corporate and retail customers. On 5th March 2020, in an attempt to avoid the collapse of the bank, which had an excessive amount of bad loans, the RBI took control of it and reconstructed the board with people from SBI, PNB etc.

The bank's management under the new leadership of Prashant Kumar, immediately repositioned itself and dealt with all internal and market related challenges to restore customer and depositor confidence.

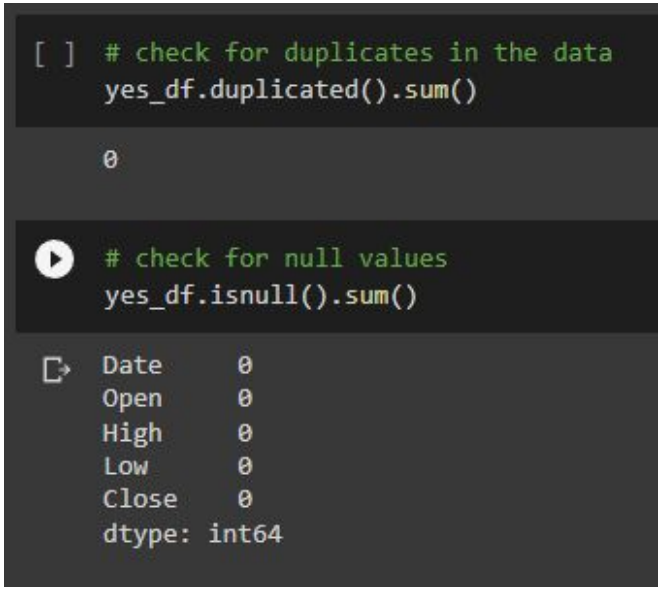
Data Summary

- **Date**: Contains the month and year information.
- **Open**: Contains information about the opening price on a particular month and year.
- **High**: High is the highest price at which a stock traded during the course of the trading day.
- **Low**: Low is the lowest price at which a stock traded during the course of the trading day.
- **Close**: Contains information about the closing price on a particular month and year.

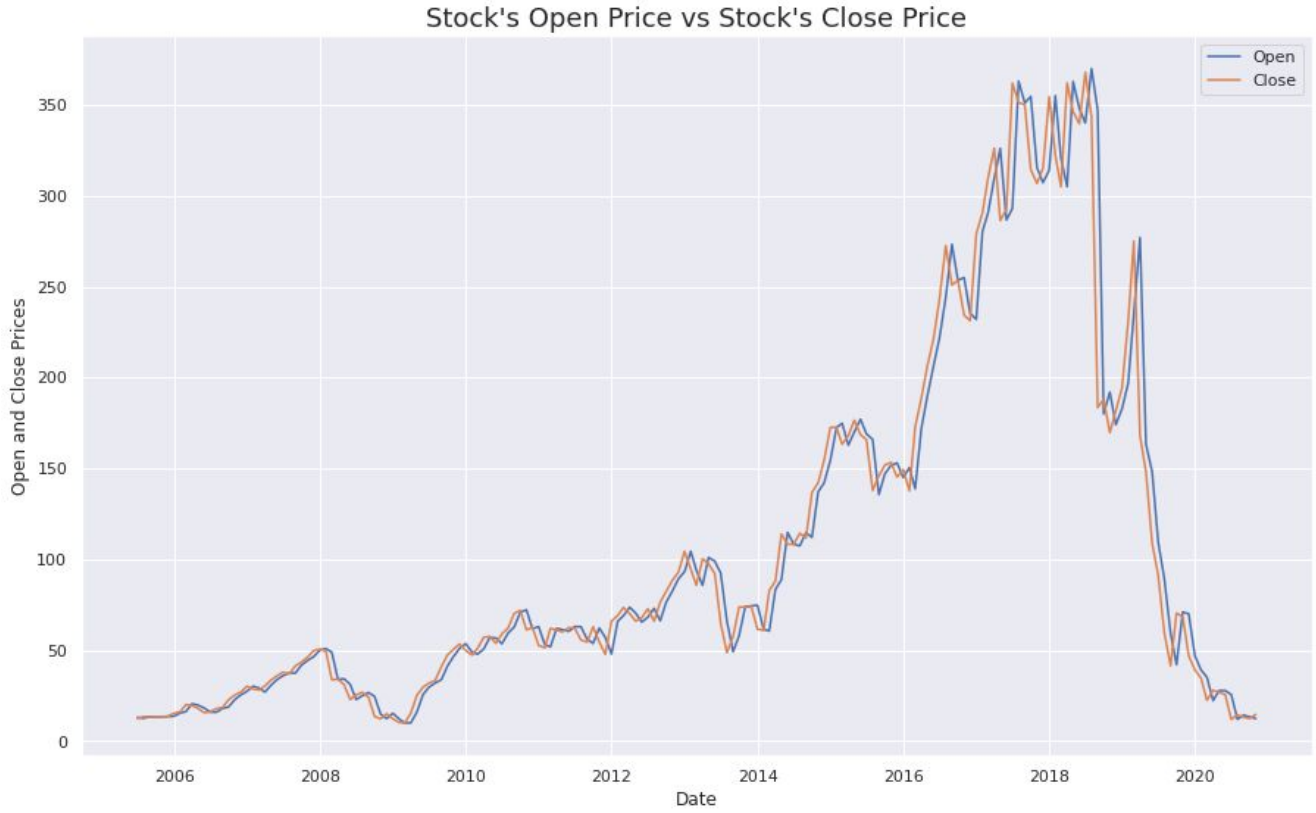
	Date	Open	High	Low	Close
0	Jul-05	13.00	14.00	11.25	12.46
1	Aug-05	12.58	14.88	12.55	13.42
2	Sep-05	13.48	14.87	12.27	13.30
3	Oct-05	13.20	14.47	12.40	12.99
4	Nov-05	13.35	13.88	12.88	13.41

Data Cleaning

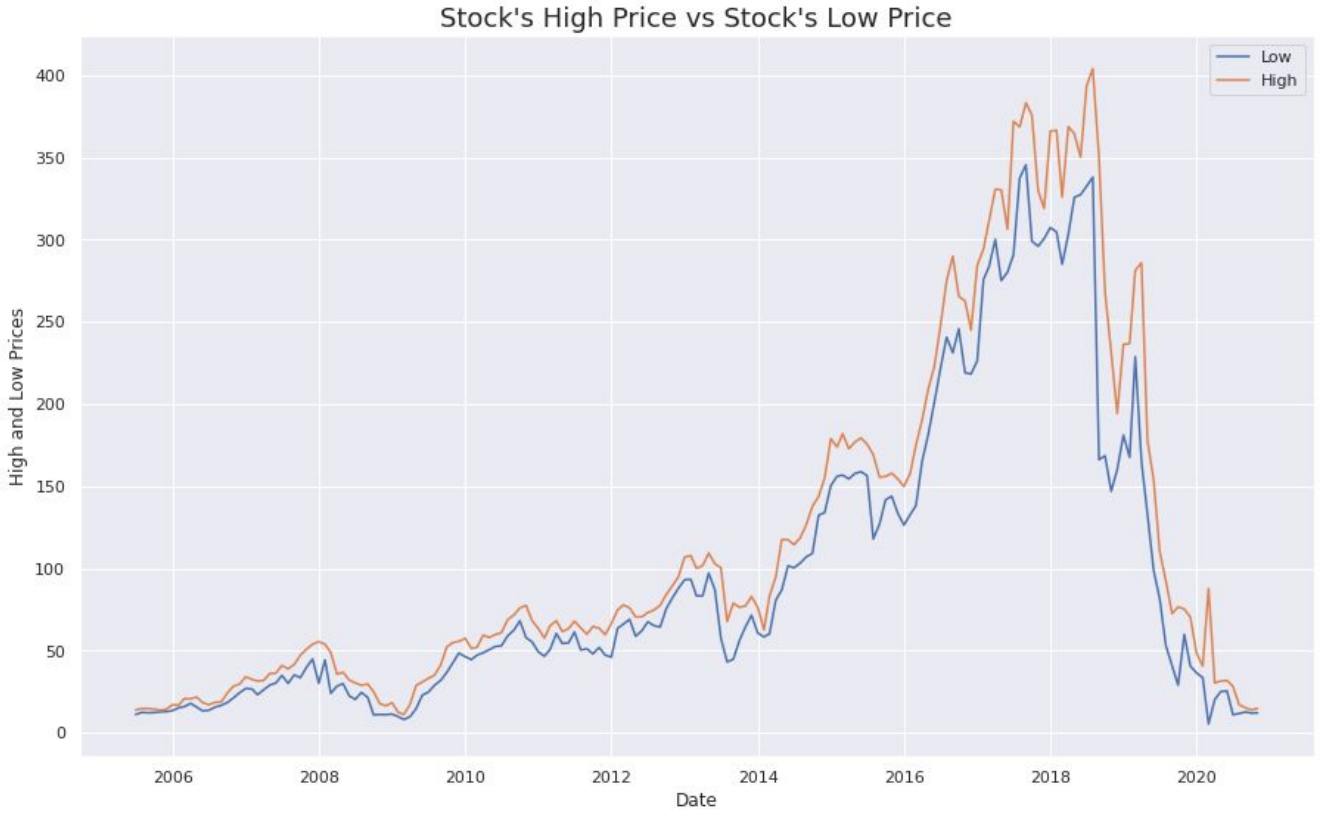
- When it comes to data, there are many different sorts of quality issues, which is why data cleaning is one of the most time consuming aspects of data analysis but thankfully this dataset was mostly a cleaned one with no duplicate records or null values present.
- The date column wasn't in the correct format and hence it was converted to a datetime object and in the format YYYY-MM-DD.



Exploratory Data Analysis



Exploratory Data Analysis

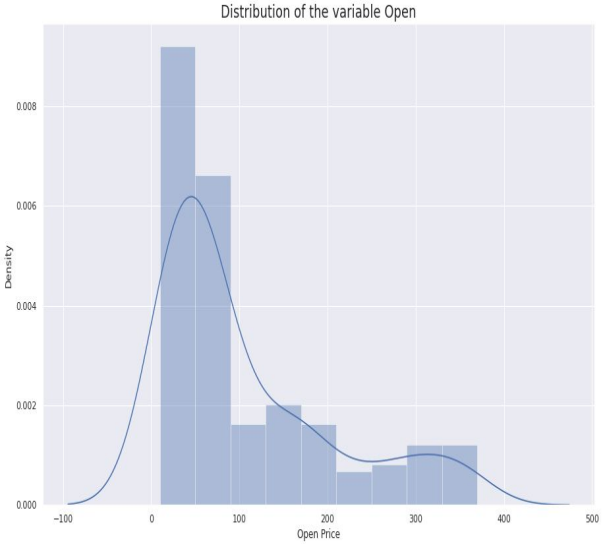


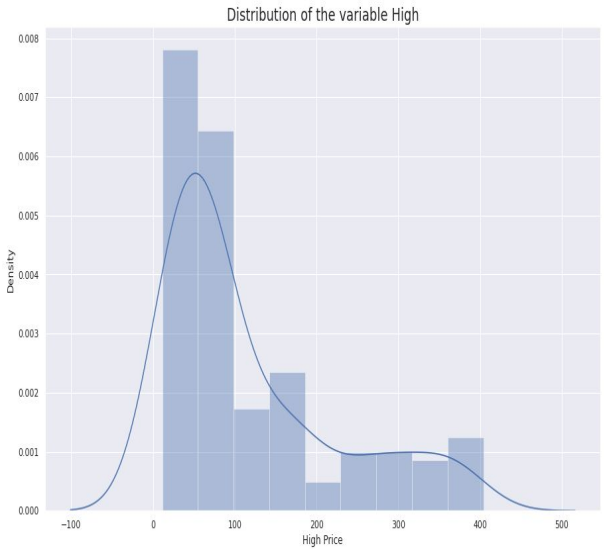
Exploratory Data Analysis

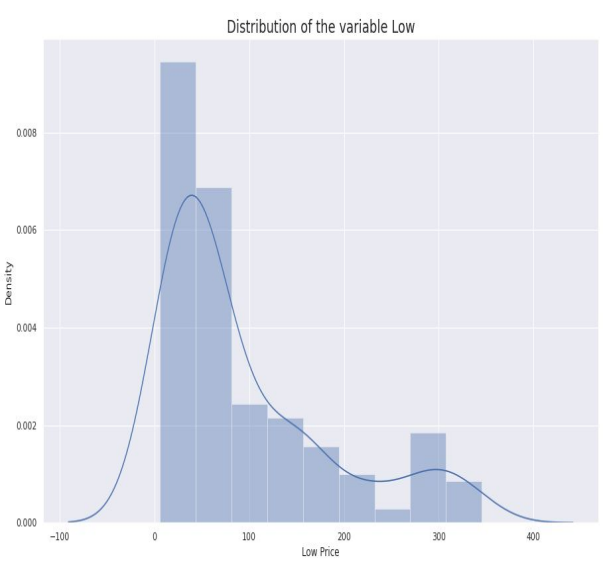
Distribution of the target variable

Applying log transformation

Exploratory Data Analysis

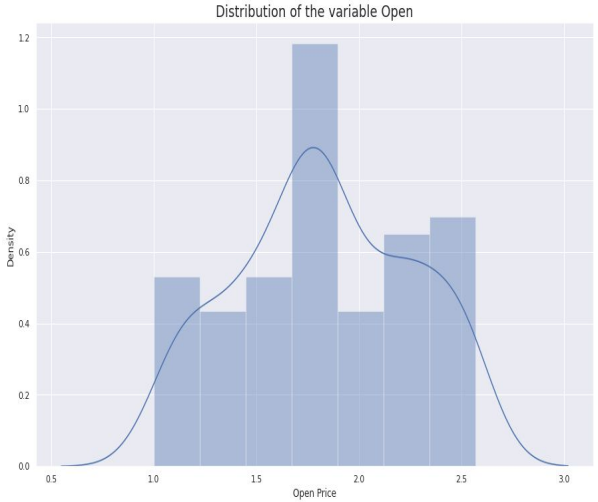


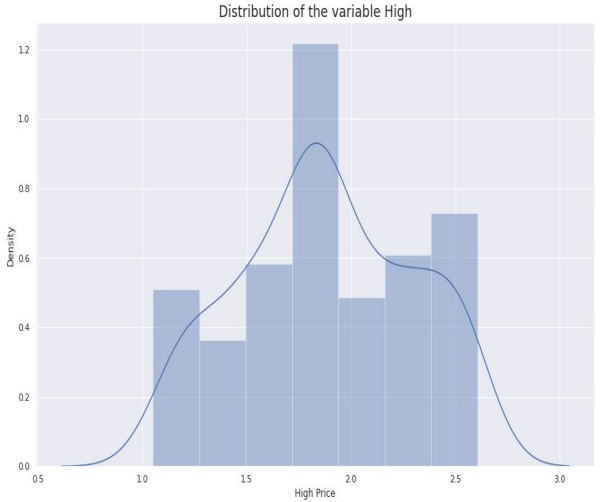


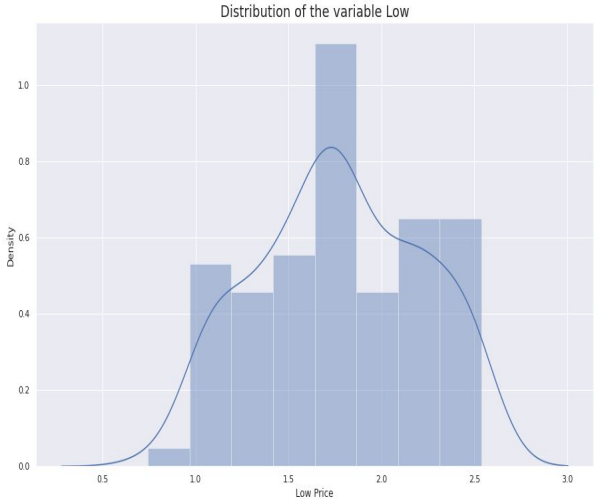


Distribution of the independent variables

Exploratory Data Analysis

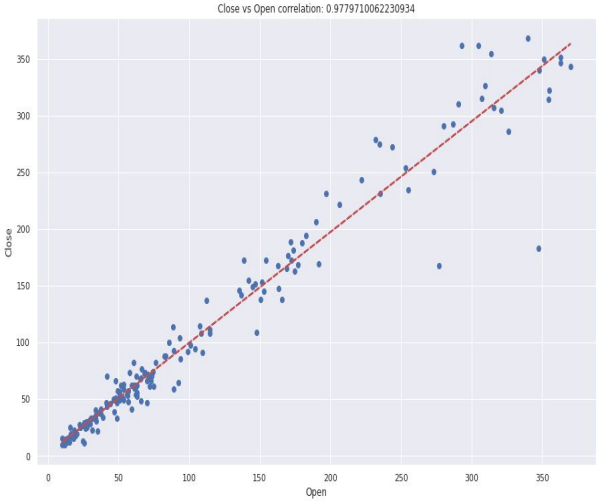


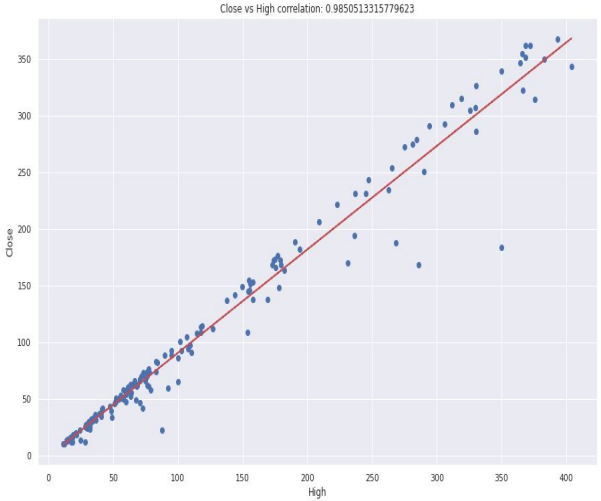


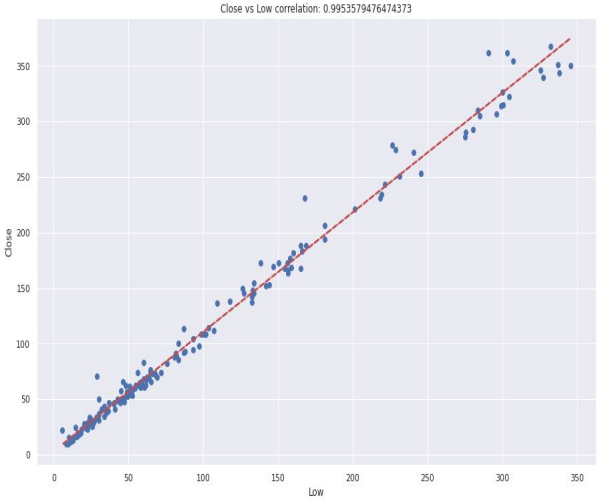


Applying log transformation to all the independent variables

Exploratory Data Analysis







Bivariate plots between the independent variables and the dependent variable

Exploratory Data Analysis

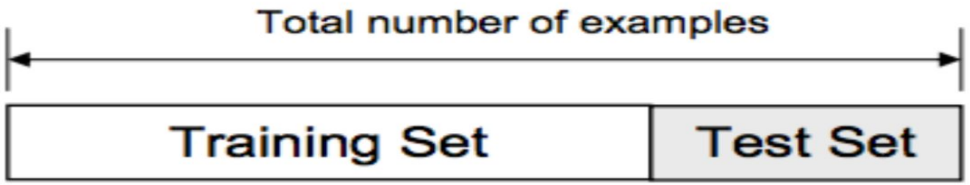
With the help of EDA, we can conclude that:

- There isn't much difference between the stock's opening price and the closing price on any given day. The best time to sell the shares was between 2016-2018 as the stock prices grew a lot between that period.
- It went on to touch an all time high of Rs 404 during Aug 2018 and then fell off drastically as soon as the Rana Kapoor fraud case came into the news. It went on to touch an all time low of Rs 5.55 during Mar 2020.
- It was an overall stable stock in the market till 2018.
- All the independent variables follow a linear relationship with the dependent variable and have a high positive correlation score.

Data Preparation

Train-Test Split:

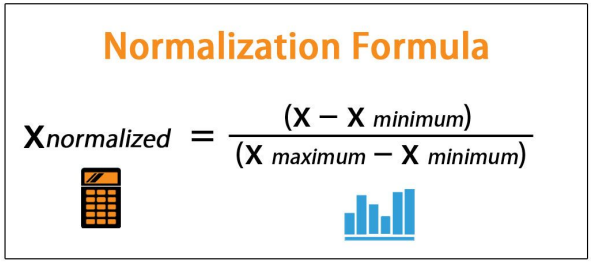
- The train test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.
- Cross validation has been used here while applying many algorithms except Linear Regression. It is a resampling method that uses different portions of the data to test and train a model on different iterations.



Data Transformation

Feature Scaling:

- Machine learning algorithms like linear regression, logistic regression, etc that use gradient descent as an optimization technique require data to be scaled. The difference in ranges of features will cause different step sizes for each feature which will make it difficult for gradient descent to move towards minima.
- Normalization is used here in which values are shifted and rescaled so that they end up ranging between 0 and 1.



Evaluation Metrics

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)}^{2}$$

The square of the difference between actual and predicted

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

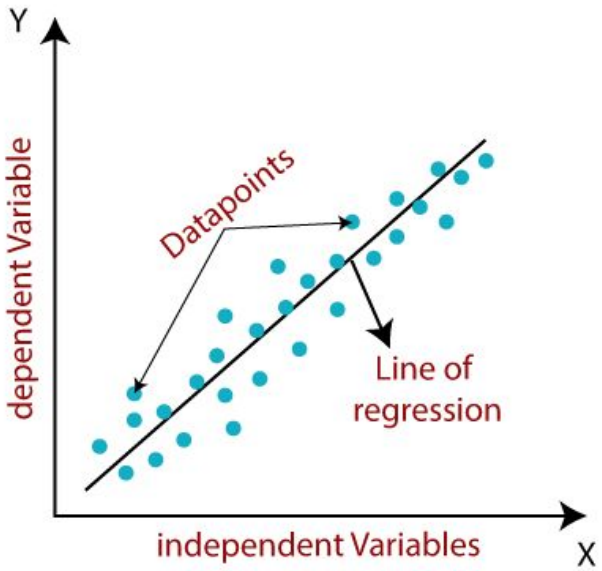
$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

$$\text{Adjusted } R^2 = 1 - \frac{SS_{residuals}/(n-K)}{SS_{total}/(n-1)}$$

Regression Analysis

Linear Regression:

- Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

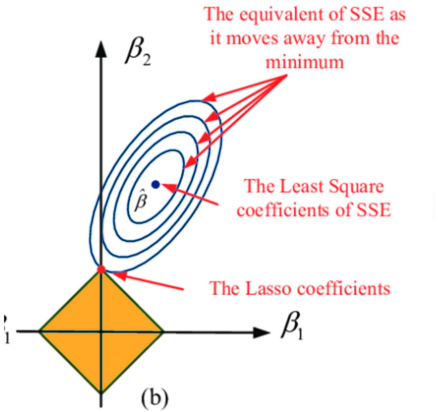


MSE	RMSE	R-Squared	Adjusted R-Squared
62.378016	7.897	0.994	0.994

Regression Analysis

Lasso Regression:

- Lasso regression is a popular type of regularized linear regression that includes an L1 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction.
- Lasso regression has performed almost similar to linear regression except a slight difference in the MSE metric.

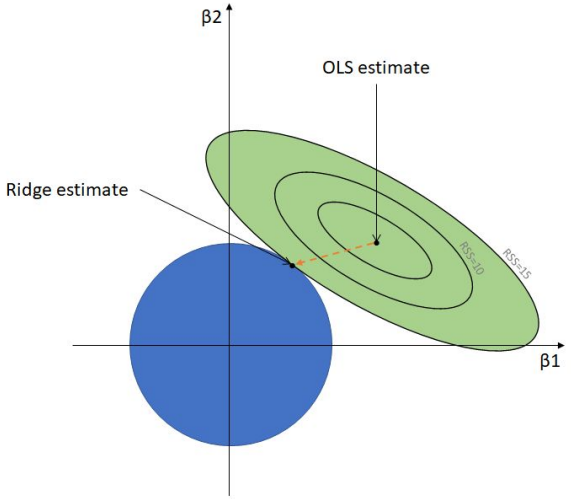


MSE	RMSE	R-Squared	Adjusted R-Squared
62.378020	7.897	0.994	0.994

Regression Analysis

Ridge Regression:

- Ridge regression is a popular type of regularized linear regression that includes an L2 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction.
- It is used to prevent multicollinearity and reduces the model complexity by coefficient shrinkage.

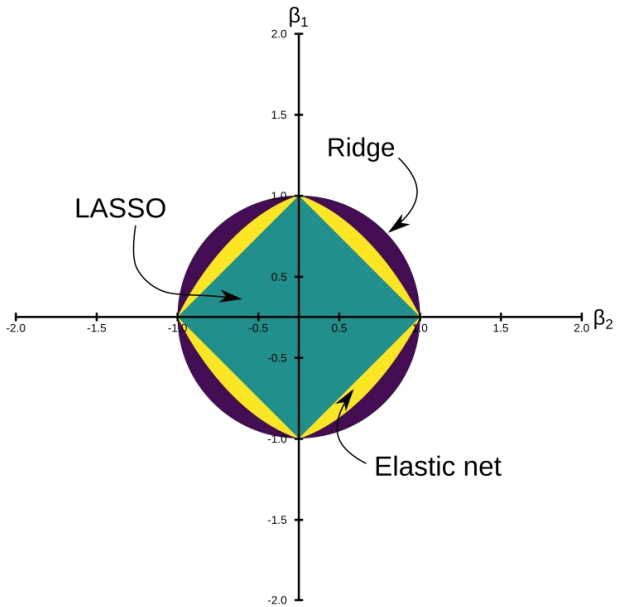


MSE	RMSE	R-Squared	Adjusted R-Squared
67.591	8.221	0.993	0.993

Regression Analysis

Elastic Net Regression:

- Elastic Net first emerged as a result of critique on lasso, whose variable selection can be too dependent on data and thus unstable. The solution is to combine the penalties of ridge regression and lasso to get the best of both the regularization techniques.



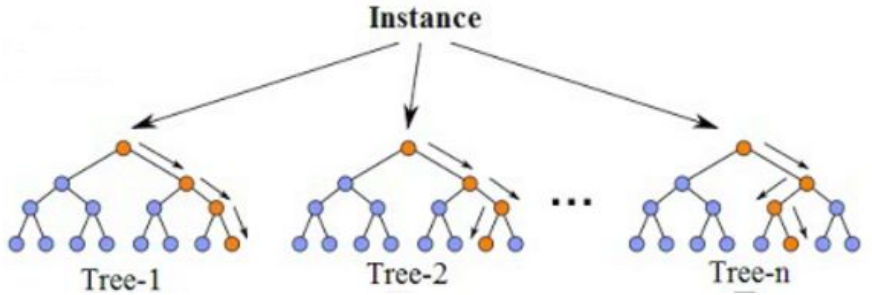
MSE	RMSE	R-Squared	Adjusted R-Squared
67.110	8.192	0.993	0.993

Regression Analysis

Random Forest Regression:

- Decision trees are great for obtaining non-linear relationships between input features and the target variable. The inner working of a decision tree can be thought of as a bunch of if-else conditions.
- Random forest is an ensemble of decision trees constructed in a certain random way.
- As our data is linearly related, this algorithm has performed poorly as compared to the linear models.

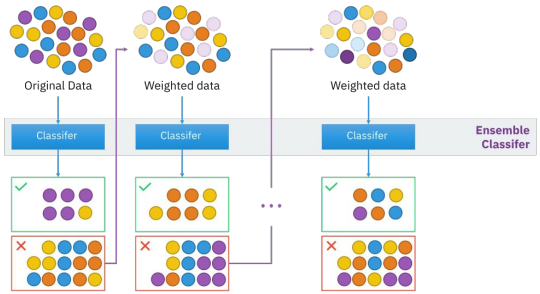
MSE	RMSE	R-Squared	Adjusted R-Squared
180.358	13.429	0.983	0.982



Regression Analysis

Gradient Boosting Regression:

- Gradient boosting is one of the variants of ensemble methods where we create multiple weak models and combine them to get better performance as a whole.
- It is powerful enough to find any non-linear relationship between the target variable and the input features and has great usability that can deal with missing values, outliers and high cardinality categorical values without any special treatment.



MSE	RMSE	R-Squared	Adjusted R-Squared
148.650	12.192	0.986	0.985

Model Performance & Comparison

Models	MSE	RMSE	R-Squared	Adjusted R-Squared
Linear Regression	62.378	7.897	0.994	0.994
Lasso	62.378	7.897	0.994	0.994
Elastic Net	67.110	8.192	0.993	0.993
Ridge	67.591	8.221	0.993	0.993
Gradient Boosting	148.650	12.192	0.986	0.985
Random Forest	180.358	13.429	0.983	0.982

Linear Regression performs better than any other model and this could be due to the simplicity of the data, cleanliness of the data and the linear relationship between the target variable and the input features.

Conclusion

We have reached the end of our analysis and prediction of Yes Bank's stock closing price. Let us now summarize few of the results we got:

- The stock's opening and closing price grew a lot between 2016 to 2018 and as soon as the Rana Kapoor fraud case occurred, it went down to the levels it was during its initial months.
- It touched an all time high during the period of Aug 2018 and an all time low during the period of Mar 2020.
- All the models used for prediction did a fairly good job and has achieved a R-squared score of around 98-99%.
- Linear Regression has performed the best out of all the models and this could be due to the simplicity of the data, cleanliness of the data, and the linear relationship between the independent variables and the dependent variable.
- Out of the 2 ensemble methods, gradient boosting has performed better than random forest in predicting the stock's closing price.

Thank You !