Capstone Project-II

Regression- Yes Bank Stock Closing Price Prediction

By- Ajay Tiwari

CONTENTS

1. Agenda
2. Problem Statement
3. Data Summary
4. Data observations & cleaning
5. EDA
6. Checking Multicollinearity
7. Data Modelling
8. Challenges
9. Conclusion

AGENDA

- We are having a dataset with the monthly stock price details for Yes Bank. The objective of this project has been to apply different models to check whether the prices/movement of the stock can be predicted using features and past performance by using Linear Regression

- Looking at the various features of the dataset, we can understand the relationships between them and accordingly pass the required parameters in the model to train it and ultimately predict the closing price.

PROBLEM STATEMENT

- ✔ Yes Bank Limited is an Indian private sector bank headquartered in Mumbai, India, and was founded by Rana Kapoor and Ashok Kapur in 2004.
- ✔ It offers a wide range of differentiated products for corporate and retail customers through retail banking and asset management services.
- ✔ On 5 March 2020, in an attempt to avoid the collapse of the bank, which had an excessive amount of bad loans, the Reserve Bank of India (RBI) took control of it.
- ✔ 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 linear regression models or any other predictive models can do justice to such situations.

DATA SUMMARY

Dataset Values

Date: In our data, it's a monthly observation of stock since it was listed.

Open: Open means the price at which a stock started trading when the opening bell rang.

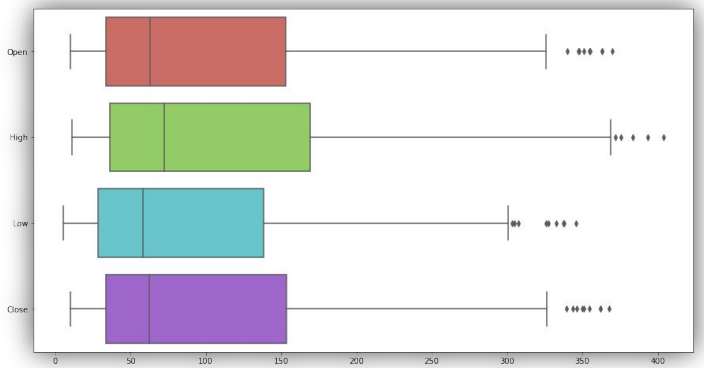
High: The maximum price of a stock attain during a given period of time.

Low: The minimum price of a stock attain during a given period of time.

Close: : Close refers to the price of an individual stock when the stock exchange closed for the day. .

DATA OBSERVATIONS & CLEANING

1) The shape of our dataset is 185 rows and 5 columns
2) Datatype of Date is given as object which we need to change that to Date Time
3) Yes bank stock listed on the month of July 2005. We have data available from July 2005 to November 2020
4) From the statistical information we can see that it is not a normal distribution as mean and 50% values are having a lot of difference
5) There are no duplicates present
6) There are no null values present

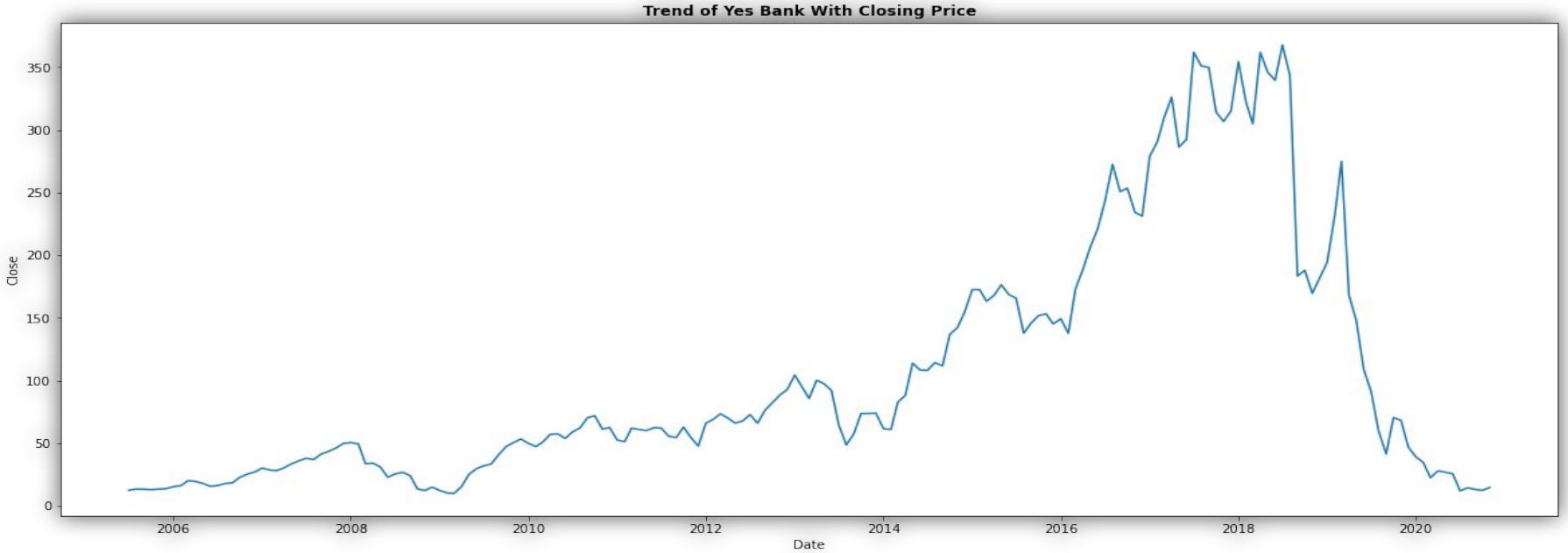


EDA

The datatype of the Date column was given as an object which we need to change to Datetime datatype.

EDA (Visualizations)

Yes Bank closing price with respect to Year

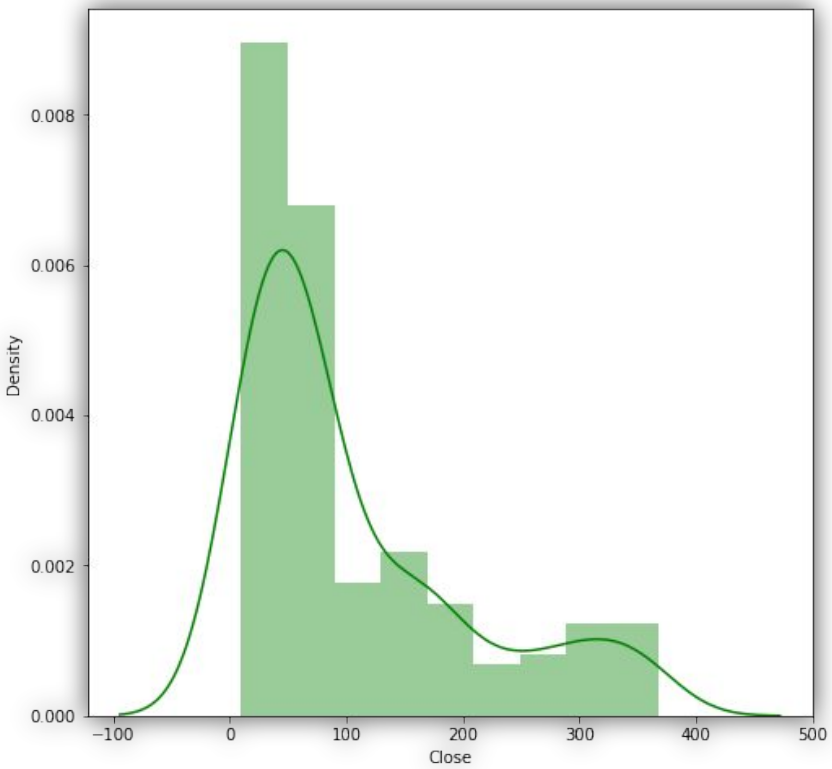


- Here we can see that the stocks were high from 2017 to 2018 but they dropped after 2018 because of a fraud case regarding Rana Kapoor.

EDA (Visualizations)

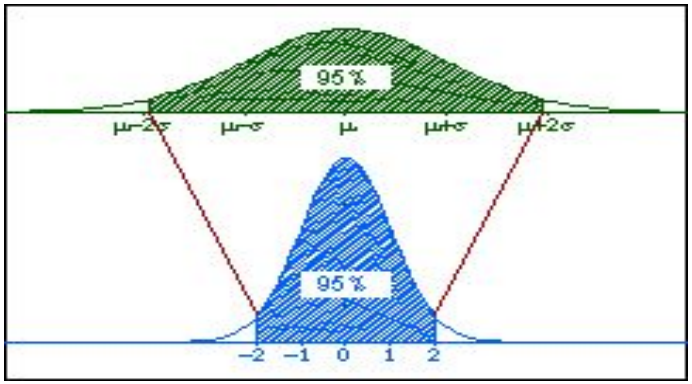
AI

Distribution of dependent variable Close Price of stock.



- It is a rightly skewed distribution.
- We need to do log transformation to make it normal distribution

TRANSFORMATIONS



- ✔ We know that in the regression analysis the response variable should be normally distributed to get better prediction results.
- ✔ Most of the data scientists claim they are getting more accurate results when they transform the independent variables too.
- ✔ It means skew correction for the independent variables.
- ✔ Lower the skewness better the result.

- ✔ Below are some types of methods or ways to deal above type of problem.

I. **square-root for moderate skew:** sqrt(x) for positively skewed data,
sqrt(max(x+1) - x) for negatively skewed data

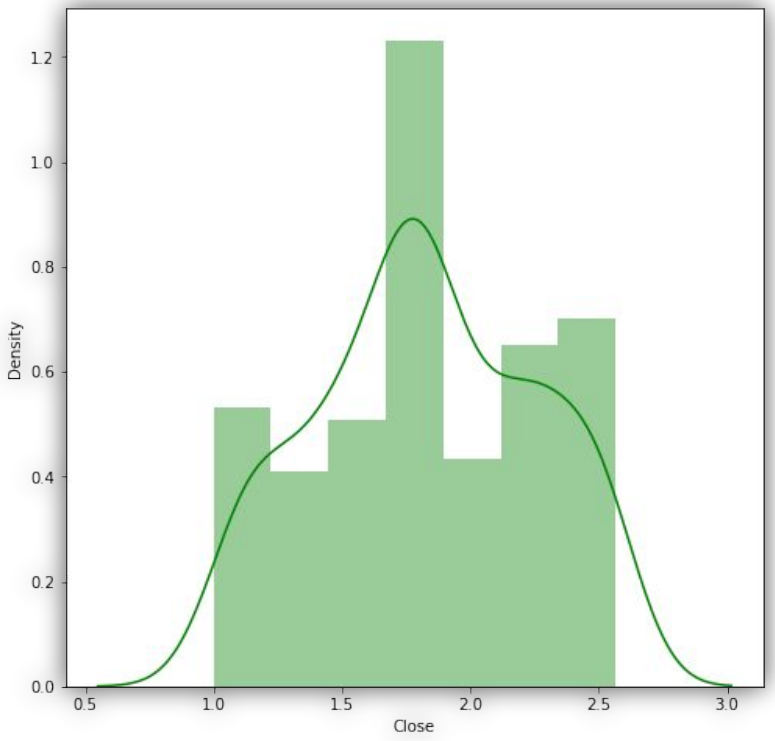
II. **log for greater skew:** log10(x) for positively skewed data, log10(max(x+1) - x) for negatively skewed data

III. **inverse for severe skew:** 1/x for positively skewed data
1/(max(x+1) - x) for negatively skewed data

IV. **Linearity and heteroscedasticity:** First try log transformation in a situation where the dependent variable starts to increase more rapidly with increasing independent variable values If your data does the opposite – dependent variable values decrease more rapidly with increasing independent variable values – you can first consider a square transformation.

EDA (Visualizations)

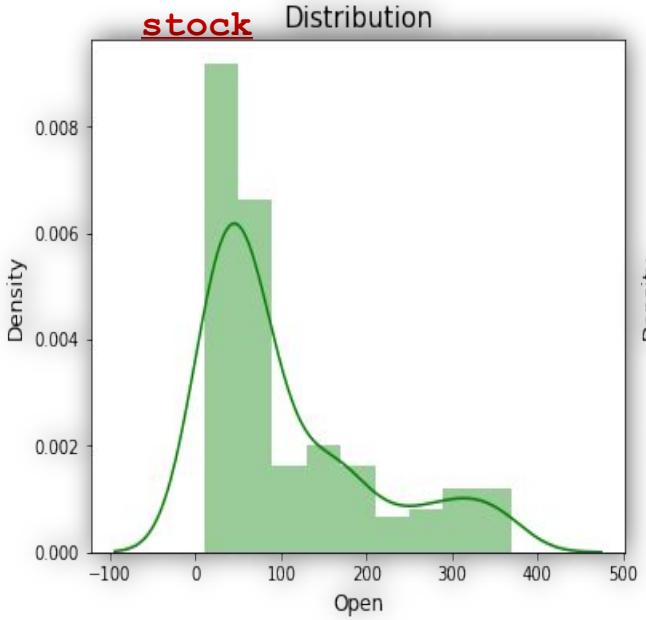
Distribution of dependent variable Close Price of stock(After Transformation).

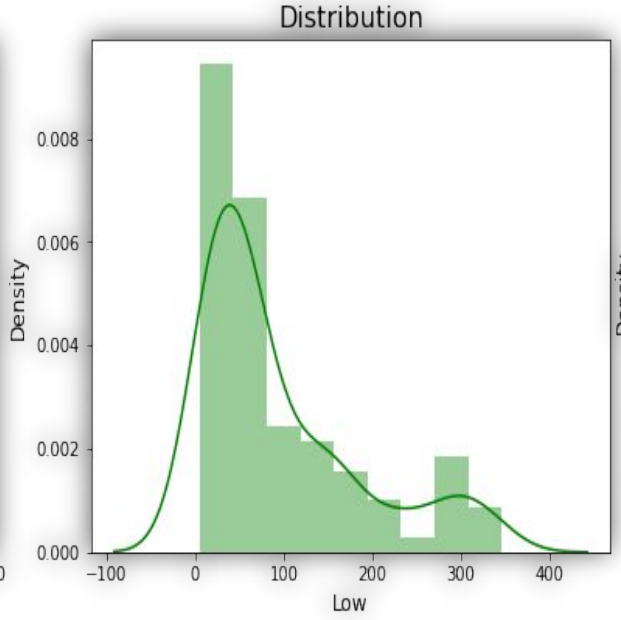


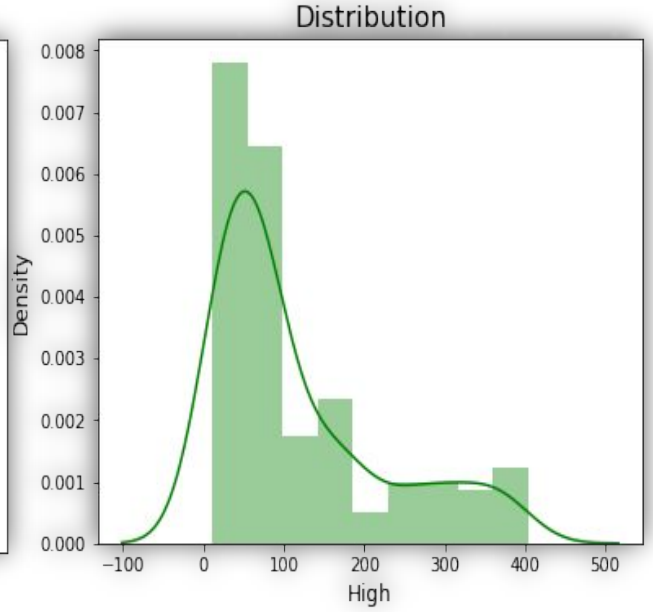
- After using log transformation as it was positively skewed.
- Now it seems more normal

EDA (Visualizations)

Distribution of numerical features High, Low and Open price of a stock



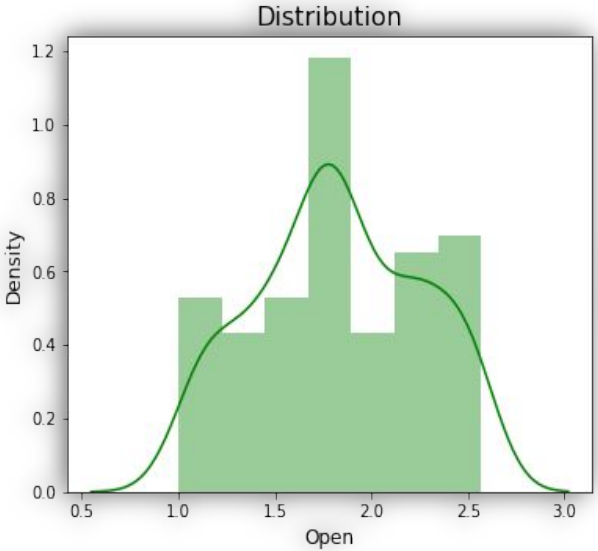


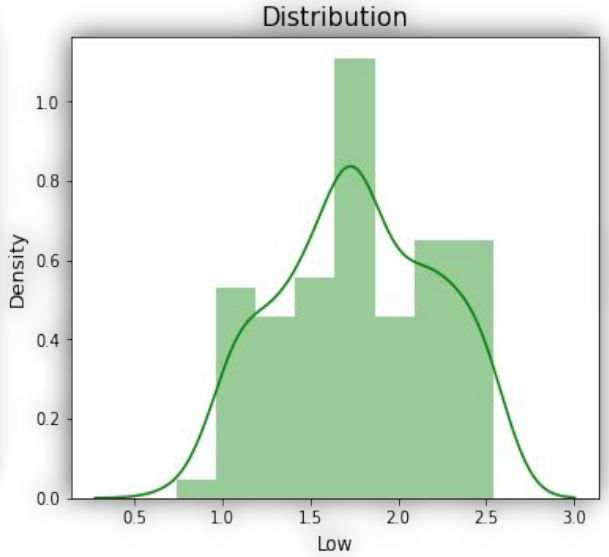


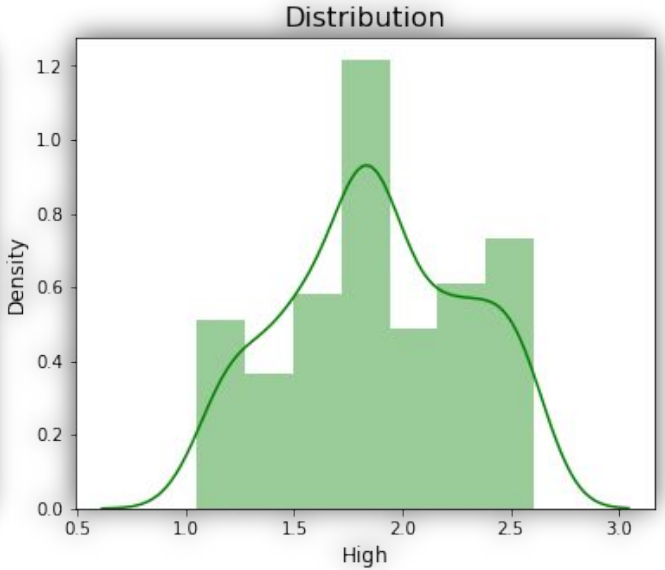
- It looks all numerical features are rightly skewed.
- Apply log transformation to make normal.

EDA (Visualizations)

Distribution of numerical features High, Low and Open price of a stock(After Transformation)



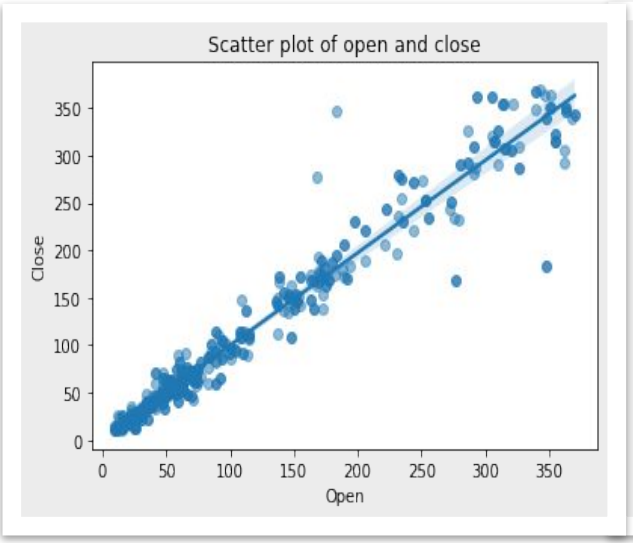


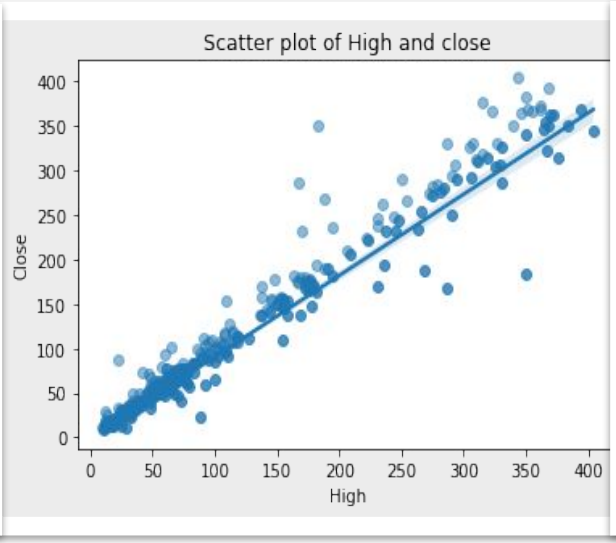


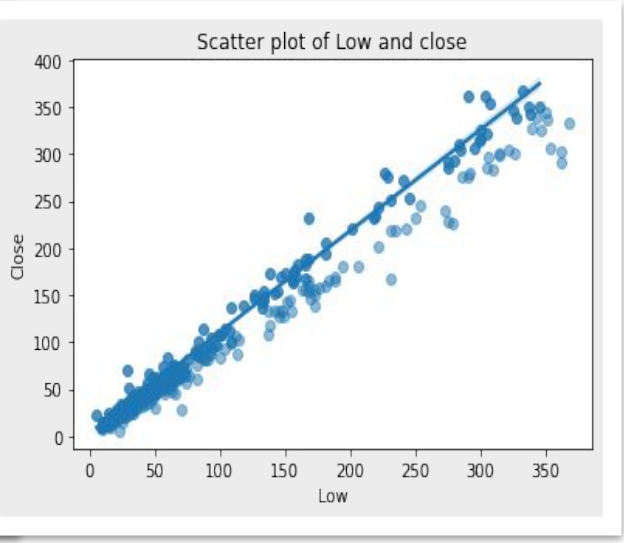
- After using log transformation as it was positively skewed.
- Now it seems more normal

EDA (Visualizations)

Scatter plot with best fit line





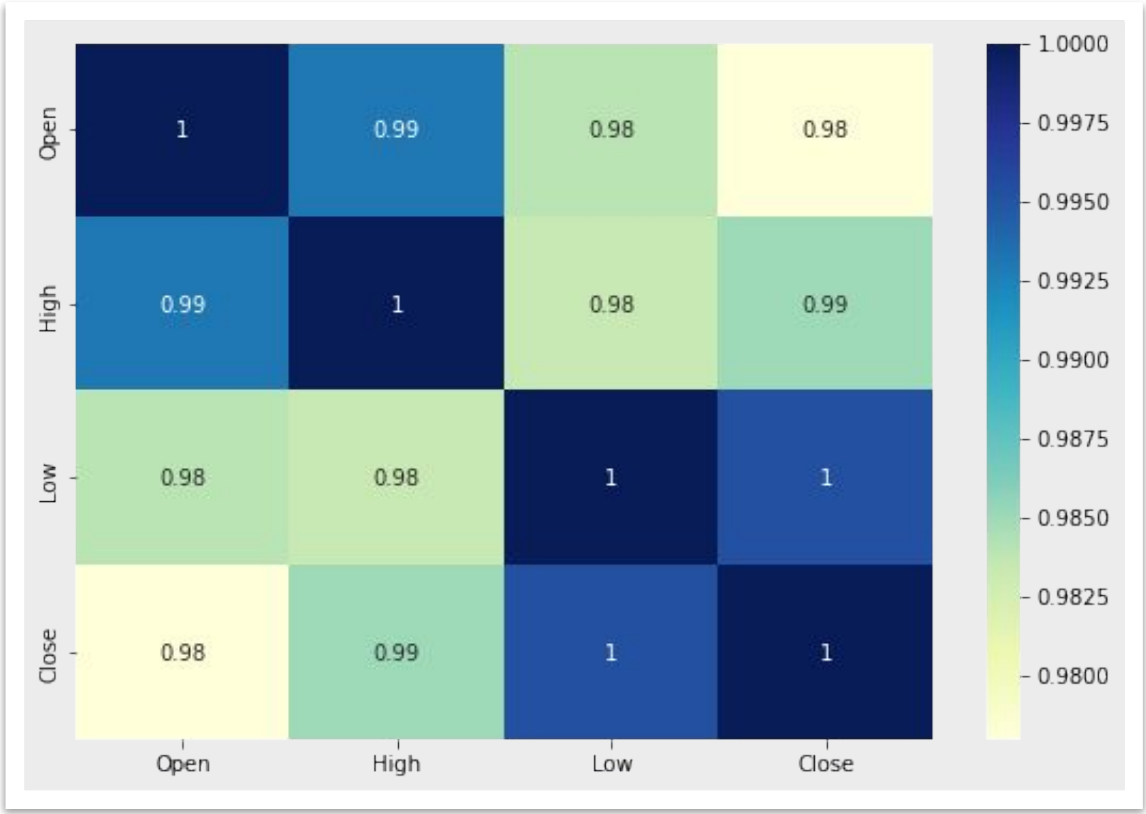


- Plotting a scatter plot with data points can help you to determine whether there's a potential relationship between them.

EDA (Visualizations)

AI

Heat Map to see the correlation



- ✔ There are very high correlations between independent variables which lead us to multicollinearity.
- ✔ High multicollinearity is not good for fitting models and prediction because a slight change in any independent variable will give very unpredictable results.

VIF(Variation Inflation Factor)

	Variables	VIF
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

- ✔ Variance inflation factor (VIF) is **a measure of the amount of multicollinearity in a set of multiple regression variables**

- ✔ In general case, Any variable having VIF above 5 is considered to be highly multicollinear.
- ✔ The thumb rule is to drop the highest VIF variable. However, you may choose to select the variable to be dropped based on business logic

Data Modelling

Splitting data

X = Independent variable(High, Low,Open)

Y = Dependent variable(Close)

Splitting train-test data with 80-20

Data must be normally distributed before applying normalization. Normalization is one of the feature scaling techniques. We particularly apply normalization when the data is skewed on either axis i.e. when the data does not follow the Gaussian distribution. In normalization, we convert the data features of different scales to a common scale which further makes it easy for the data to be processed for modeling. Thus, all the data features(variables) tend to have a similar impact on the modeling portion. We have used z score and log transformation.

Linear Regression

Date	Actual Closing Price	LR Predicted Closing Price
2008-04-01	34.06	26.036846
2015-12-01	145.23	136.250586
2019-12-01	46.95	68.780191
2014-11-01	142.08	127.936548
2010-08-01	62.22	62.863836

- ✔ Linear regression is one of the easy and popular Machine Learning algorithms.
- ✔ It is a statistical method that is used for predictive analysis.
- ✔ Linear regression makes predictions for continuous or numeric variables such as **sales, salary, age, product price,** etc.
- ✔ The linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression.
- ✔ Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
- ✔ We got a score of 95.50 for R squared value

Lasso Regression

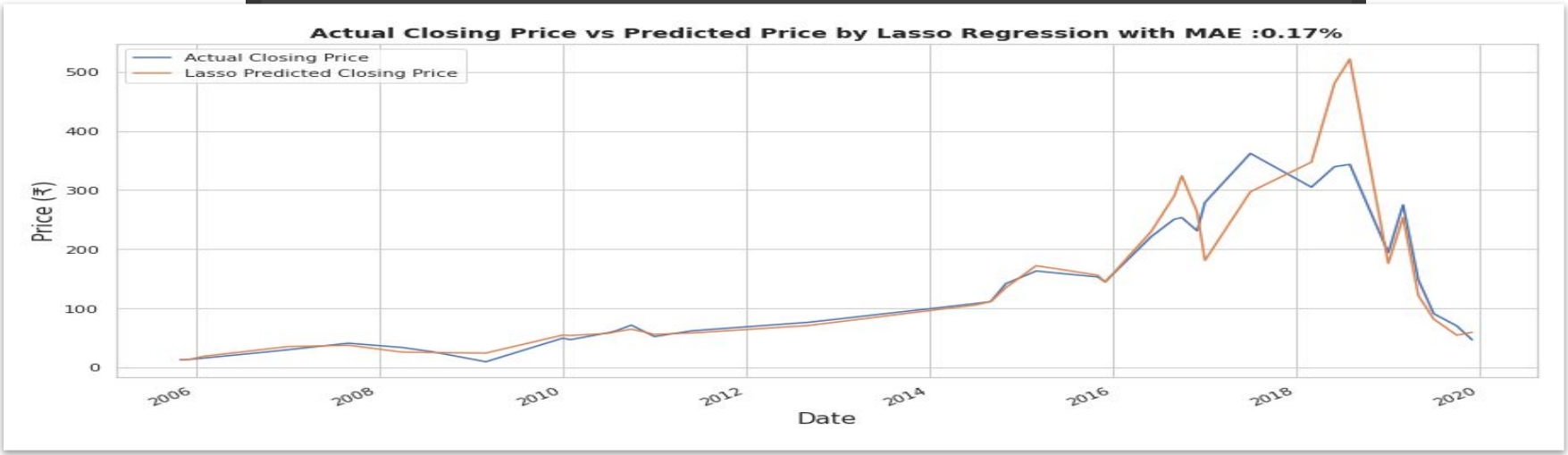
Date	Actual Closing Price	Lasso Predicted Closing Price
2008-04-01	34.06	26.365934
2015-12-01	145.23	144.313848
2019-12-01	46.95	59.849779
2014-11-01	142.08	135.358269
2010-08-01	62.22	60.649159

```
The model performance for test set:
--------------------------------------
Mean Squared Error:  0.009379605869086021
Root Mean Squared Error:  0.09684836534028864
R2:  0.9496484411579983
Adjusted R2:  0.9045970464046283
```

- ✔ Lasso regression is linear regression, but it uses a technique called **"shrinkage"** where the coefficients of determination shrink towards **zero**.
- ✔ Linear regression gives you regression coefficients as observed in the dataset.
- ✔ The lasso regression allows to shrink or regularize coefficients to avoid overfitting and make them work better on different datasets.
- ✔ This type of regression is used when the dataset shows high multicollinearity or when you want to automate variable elimination and **feature selection**.
- ✔ We got score of 94.96 for R squared value.

Cross Validation of Lasso Regression

Date	Actual Closing Price	Lasso Predicted Closing Price
2008-04-01	34.06	26.365934
2015-12-01	145.23	144.313848
2019-12-01	46.95	59.849779
2014-11-01	142.08	135.358269
2010-08-01	62.22	60.649159



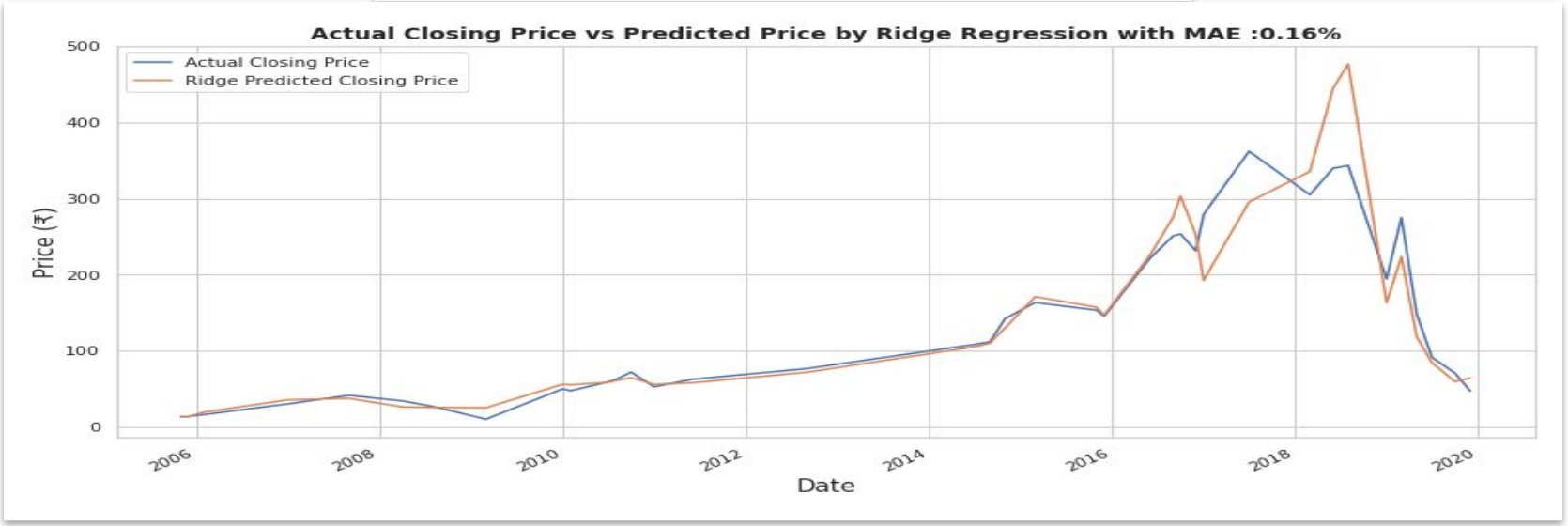
Ridge Regression

```
The model performance for test set:
--------------------------------------
Mean Squared Error:  0.009365359873489867
Root Mean Squared Error:  0.0967747894520565
R2:  0.9497249164487018
Adjusted R2:  0.9047419469554351
```

- ✔ Ridge regression is a regularized version of linear least squares regression.
- ✔ It works by shrinking the coefficients or weights of the regression model toward zero.
- ✔ This is achieved by imposing a squared penalty on their size.
- ✔ This is one of the methods of regularization techniques in which the data suffer from multicollinearity.
- ✔ In this multicollinearity, the least squares are unbiased and the variance is large and which deviates the predicted value from the actual value. Equations have an error term.
- ✔ We got a score of 94.97 as R squared value

Cross Validation of Ridge Regression

Date	Actual Closing Price	Ridge Predicted Closing Price
2008-04-01	34.06	26.142190
2015-12-01	145.23	146.673951
2019-12-01	46.95	64.154295
2014-11-01	142.08	129.931358
2010-08-01	62.22	60.731005



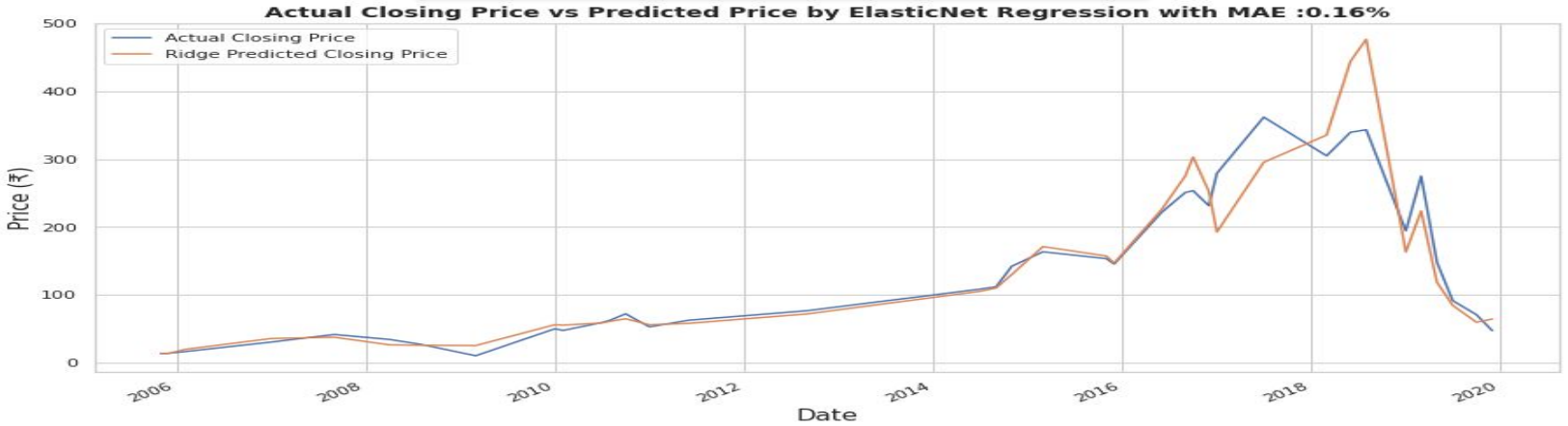
Elastic Net Regression

```
The model performance for test set:
--------------------------------------
Mean Squared Error:  0.00933354858491178
Root Mean Squared Error:  0.096610292352788O5
R2:  0.9498956856484749
Adjusted R2:  0.905065509649742
```

- ✔ Elastic Net Regression is the third type of Regularization technique.
- ✔ It came into existence due to the limitation of the Lasso regression.
- ✔ Lasso regression cannot take correct alpha and lambda values as per the requirement of the data.
- ✔ The solution for the problem is to combine the penalties of both ridge regression and lasso regression.
- ✔ We got a score of 94.98 as R squared value.

Cross Validation of Elastic Net

Date	Actual Closing Price	ElasticNet Predicted Closing Price
2008-04-01	34.06	26.763841
2015-12-01	145.23	142.542580
2019-12-01	46.95	59.677823
2014-11-01	142.08	130.999880
2010-08-01	62.22	60.162559



Challenges

- While deciding on the independent variables we faced difficulty as there were very limited parameters and they were having very high collinearity with the dependent variable.
- The disadvantage of Linear regression while predicting stock prices is that it is highly limited in its scope.
- Many predictors cannot be used, which is required to solve the stock price prediction problem.
- According to our observation we concluded that such problems can be better handled by using time series forecasting and K nearest neighbour also fbprophet.

Conclusion

1. There is increase in trend of Yes Bank's stock Close, Open.
2. High, Low price till 2018 and then sudden decrease.
3. We observed that open vs close price graph concluded that after 2018 yes bank's stock hitted drastically.
4. We saw Linear relation between the dependent and independent value.
5. There was a lot of multicollinearity present in data.
6. The target variable is highly dependent on input variables.

1. Ridge regression shrunk the parameters to reduce complexity and multicollinearity but ended up affecting the Linear Regression has given the best results with the lowest MSE, RMSE scores.evaluation metrics.

1. Lasso regression did feature selection and ended up giving up worse results than ridge which again reflects the fact that each feature is important (as previously discussed).
2. There is increase in trend of Yes Bank's stock .

	Linear Regression	Lasso	Ridge	ElasticNet
MSE	0.008379	0.009377	0.008848	0.009096
RMSE	0.091535	0.096833	0.094061	0.095375
R2	0.955021	0.949664	0.952505	0.951169
Adjusted_R2	0.914777	0.904627	0.910009	0.907478

References

I. Stack overflow
II. GeeksforGeeks
III. Analytics Vidhya
IV. Almabetter
V. github
VI. Towards data science


THANK-YOU