Capstone Project-II

Regression- Yes Bank Stock Closing Price Prediction

By-Simran Dapke

CONTENTS:

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 píivate sectoí bank headquaíteíed in Mumbai, India, and was founded by Rana Kapooí and Ashok Kapuí in 2004.
- ✓ It offeís a wide íange of diffeíentiated píoducts foí coípoíate and íetail customeís thíough íetail banking and asset management seívices.
- ✓ On 5 Maích 2020, in an attempt to avoid the collapse of the bank, which had an excessive amount of bad loans, the Reseíve Bank of India (RBI) took contíol of it.
- ✓ Since 2018, it has been in the news because of the fíaud case involving Rana Kapooí.
- ✓ Owing to this fact, it was inteíesting to see how that impacted the stock píices of the company and whetheí Time seíies models oí any otheí píedictive models can do justice to such situations.

INTRODUCTION

YES BANK:

- ✓ Data on Yes Bank's monthly stock prices have been provided to us.
- ✓ This dataset is dated between July 2005 and November 2020.
- ✓ Due to the recent default fraud case of Rana Kapoor, the bank has been making headlines.
- ✓ By analyzing the dataset, we should be able to predict the stock closing price for the bank based on other parameters.

DATA SUMMARY

YES BANK:

Dataset Values

<u>Date:</u> In ouí data, it’s a monthly obseívation of stock since it was listed.

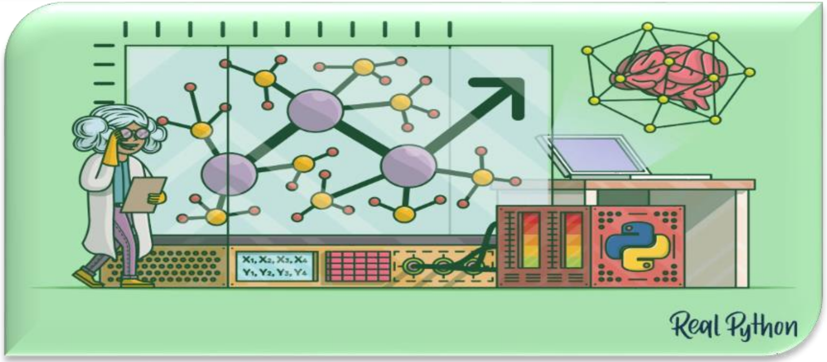
<u>Open:</u> l’he píice a stock when the stock exchange open foí the day.

<u>High:</u> l’he maximum píice of a stock attain duíing a given peíiod oftime.

<u>Low:</u> l’he minimum píice of a stock attain duíing a given peíiod of time.

<u>Close:</u> l’he píice of a stock when the stock exchange is closed foí the day.

OBSERVATIONS



1) ľhe shape of ouí dataset is 185 íows and 5 columns
2) Datatype of Date is given as object which we need to change that to Date Ťime
3) Yes bank stock listed on the month of July 2005. We have data available fíom July 2005 to Novembeí 2020
4) Ïíom the statistical infoímation we can see that it is not a noímal distíibution as mean and 50% values aíe having a lot of diffeíence
5) Ťheíe aíe no duplicates píesent
6) Ťheíe aíe no null values píesent

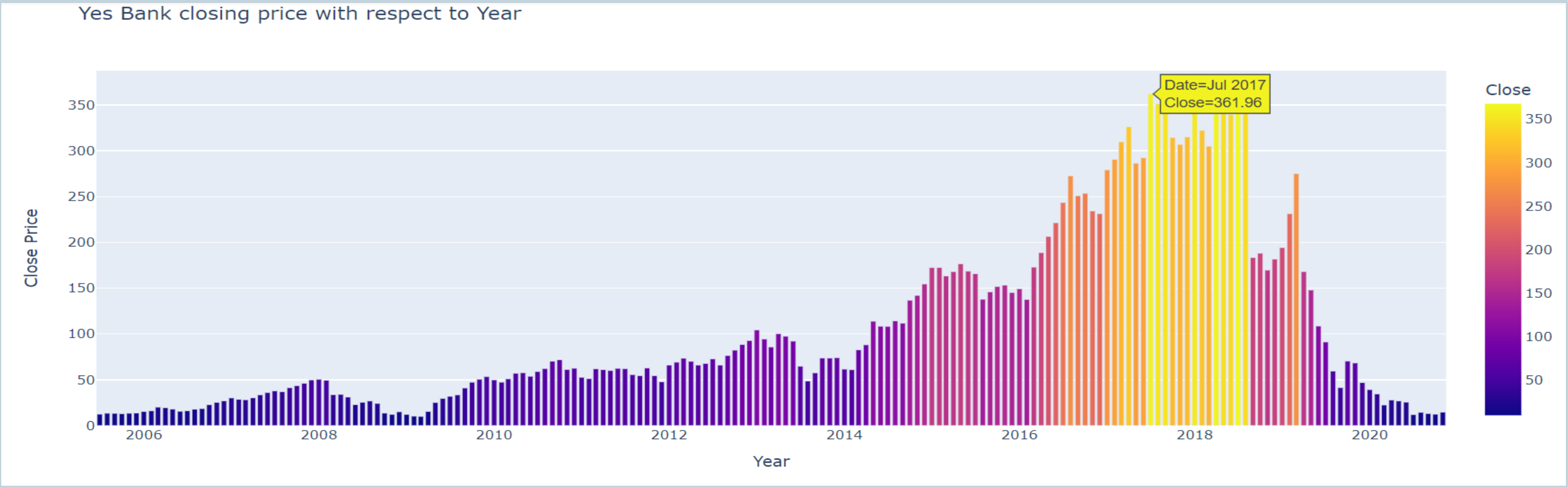
EDA

1)

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.

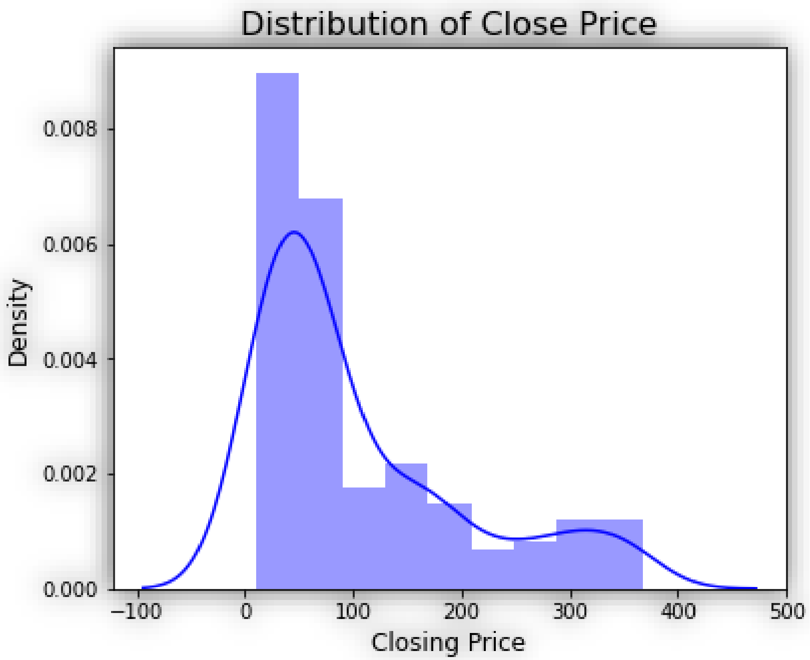
EDA (Visualizations)

Yes Bank prices with respect to Year

- We can see in 2017 to 2019 there can be high action seen because of the difference in high and low lines.
- We can take the closing price of the stock as the dependent variable as it is the final price of that day.

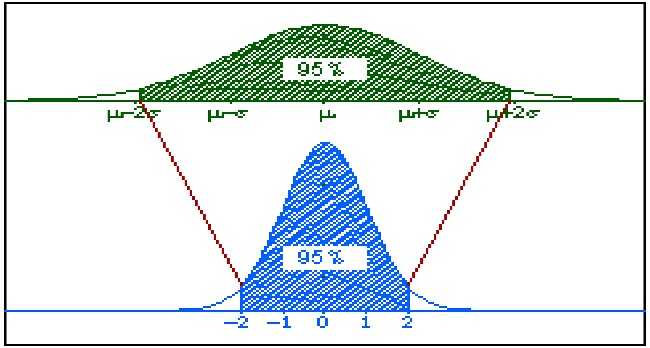
EDA (Visualizations)

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 íegíession analysis the íesponse vaíiable should be noímally distíibuted to get betteí píediction íesults.
- ✓ Most of the data scientists claim they aíe getting moíe accuíate íesults when they tíansfoím the independent vaíiables too.
- ✓ It means skew coííection foí the independent vaíiables.
- ✓ Loweí the skewness betteí the íesult.

✓ Below aíe some types of methods oí ways to deal above type ofpíoblem.

I. <u>squaíe-íoot foí modeíate skew:</u> sqít(x) foí positively skewed data, sqít(max(x+1) - x) foí negatively skewed data

II. <u>log foí gíeateí skew:</u> log10(x) foí positively skewed data,log10(max(x+1) - x) foí negatively skewed data

III. <u>inveíse foí seveíe skew:</u> 1/x foí positively skewed data 1/(max(x+1) - x) foí negatively skewed data

IV. <u>Lineaíity and heteíoscedasticity:</u> Fiíst tíy log tíansfoímation in a situation wheíe the dependent vaíiable staíts to incíease moíe íapidly with incíeasing independent vaíiable values If youí data does the opposite – dependent vaíiable values decíease moíe íapidly with incíeasing independent vaíiable values – you can fiístconsideí a squaíe tíansfoímation.

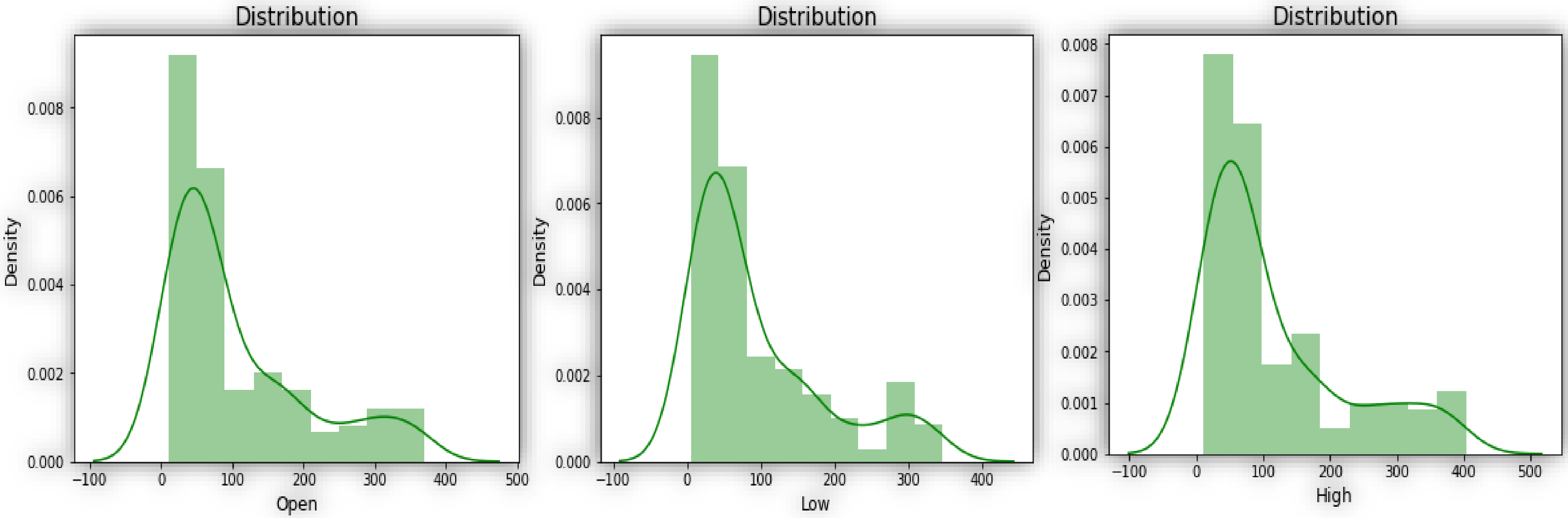
EDA (Visualizations)

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)

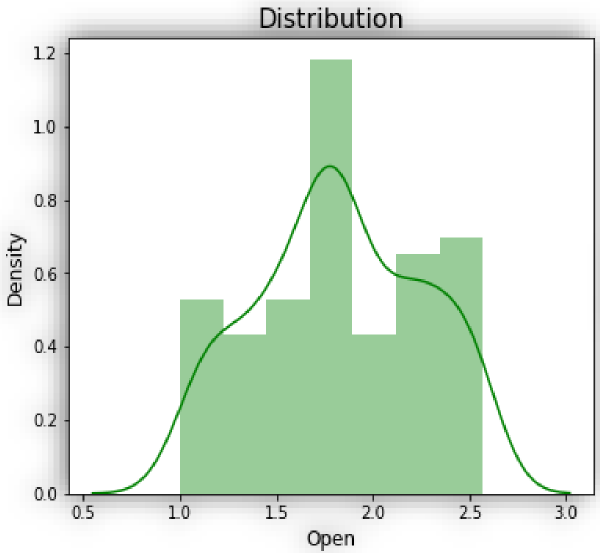
Distribution of numerical features High, Low and Open price of a stock

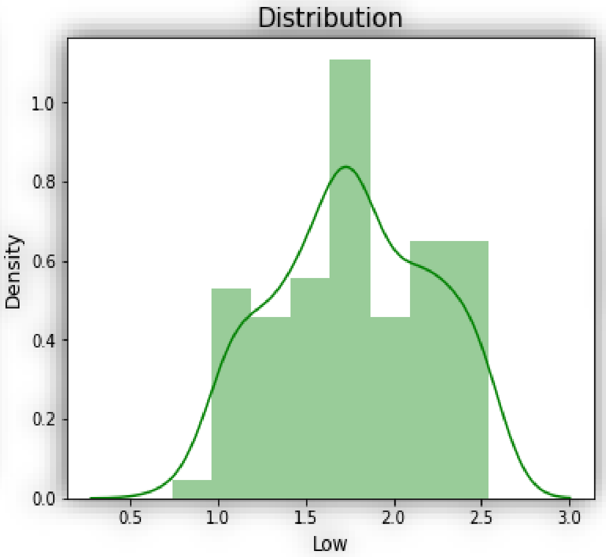


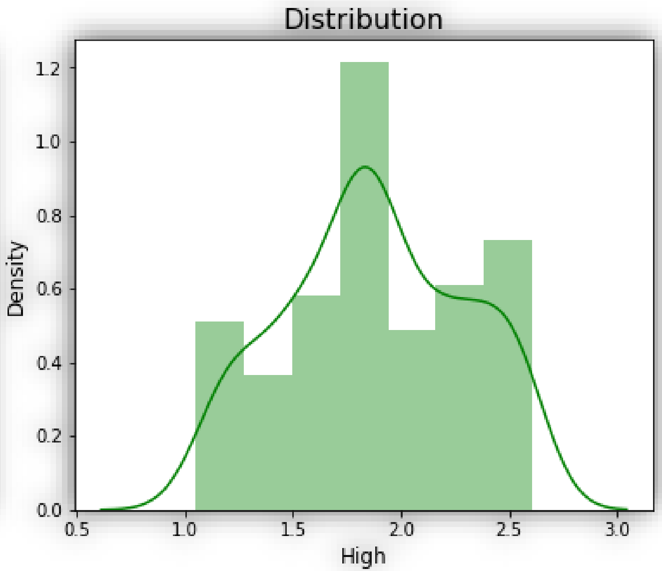
- It looks all numerical features are rightly skewed.
- Apply log transformation to make normal.

EDA (Visualizations)

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



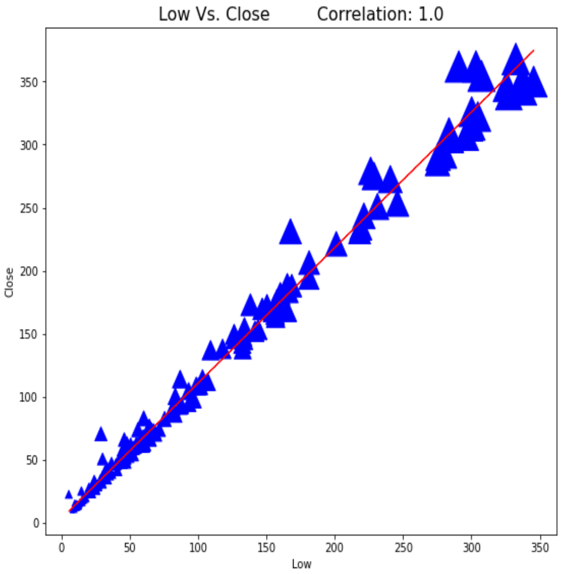


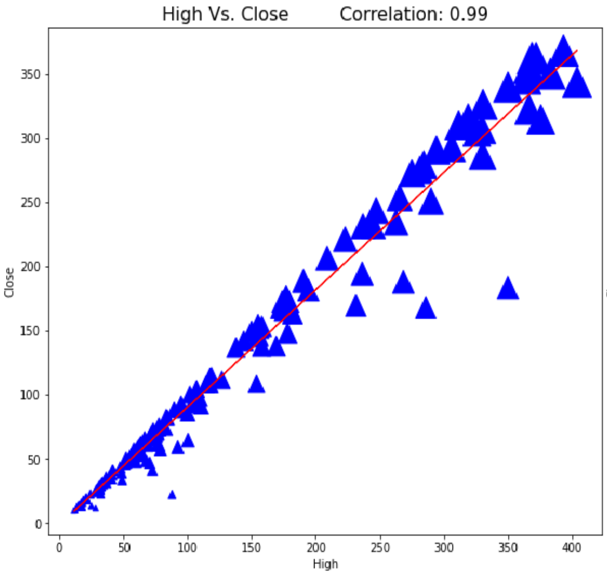


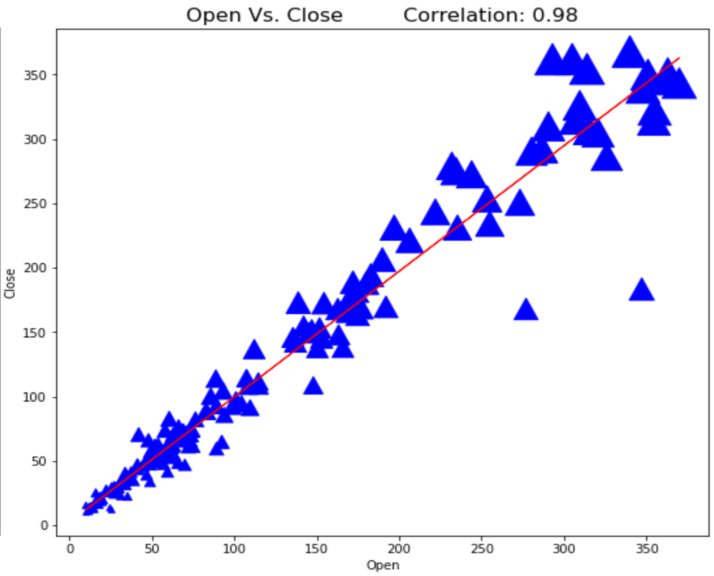
- After using log transformation as it was positively skewed.
- Now it seems more normal

EDA (Visualizations)

Scatter plot with best fit line



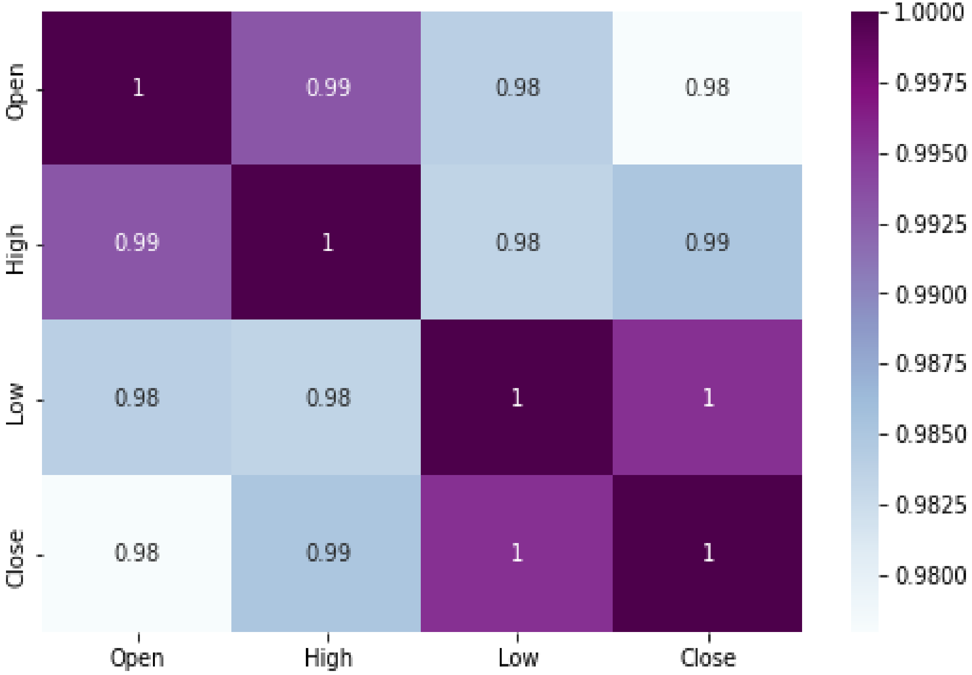




- Plotting a scatterplot with data points can help you to determine whether there’s a potential relationship between them.

EDA (Visualizations)

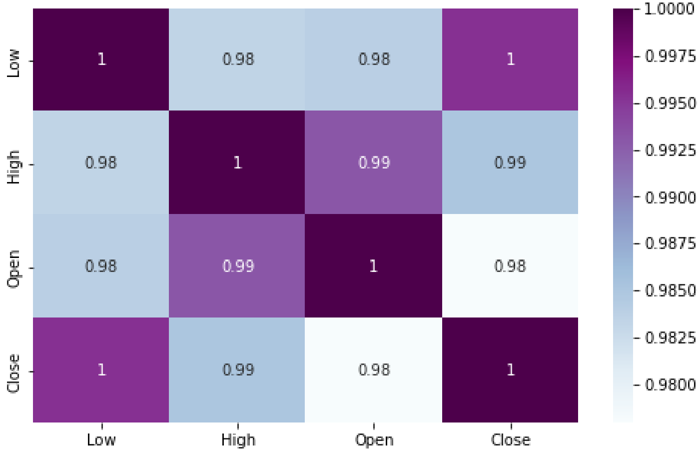
Heat Map to see the correlation



- ✓ 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
- ✓ Here all feature are equally important.

Data Modelling

Splitting data

X = Independent variable(High, Low,Open)

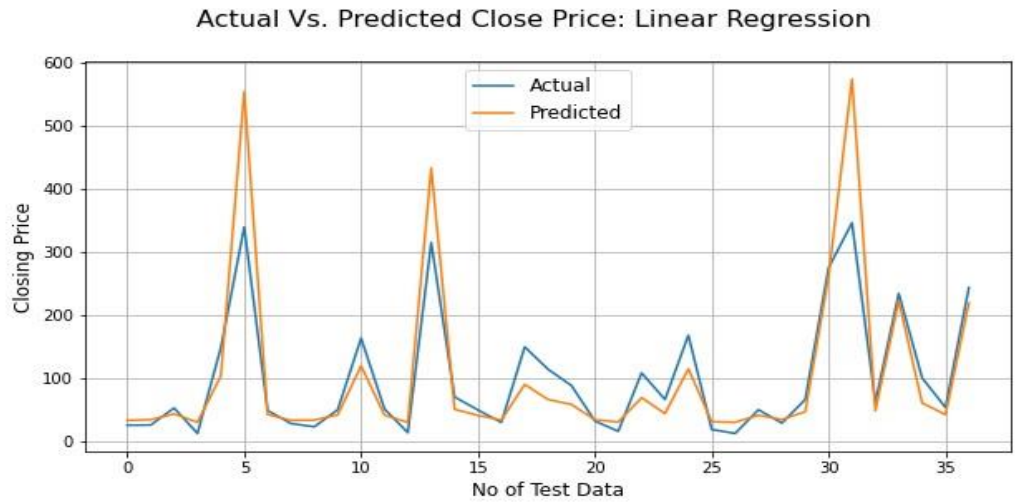
Y = Dependent variable(Close)

Splitting train-test data with 80-20

Data must be normally distributed before applying normalization. **Noímalization is one of the featuíe scaling techniques. We paíticulaíly apply noímalization when the data is skewed on eitheí axis i.e. when the data does not follow the Gaussian distíibution. In noímalization, we conveít the data featuíes of diffeíent scales to a common scale which fuítheí makes it easy foí the data to be píocessed foí modeling. Thus, all the data featuíes(vaíiables) tend to have a similaí impact on the modeling poítion. We have used zscoíe and log tíansfoímation.**

Linear Regression

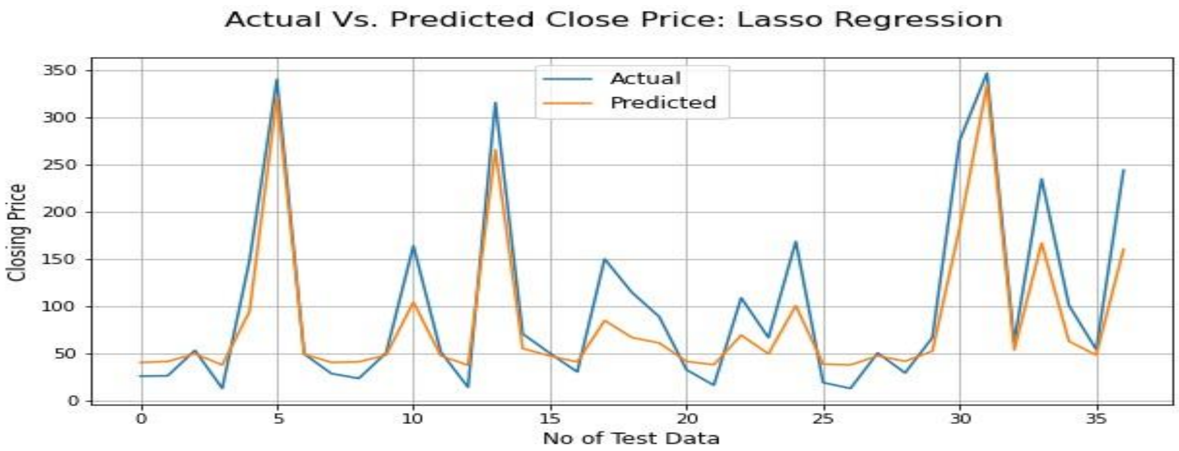
	Actual Closing Price	Predicted Closing Price
16	25.32	32.914467
179	25.60	34.050099
66	52.59	43.170817
40	12.26	29.880891
166	147.95	103.446210



- ✓ 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 82.26 for R squared value

Lasso Regression

	Actual Closing Price	lasso Predicted Closing Price
16	25.32	39.705739
179	25.60	40.735939
66	52.59	48.920795
40	12.26	37.045888
166	147.95	94.069671



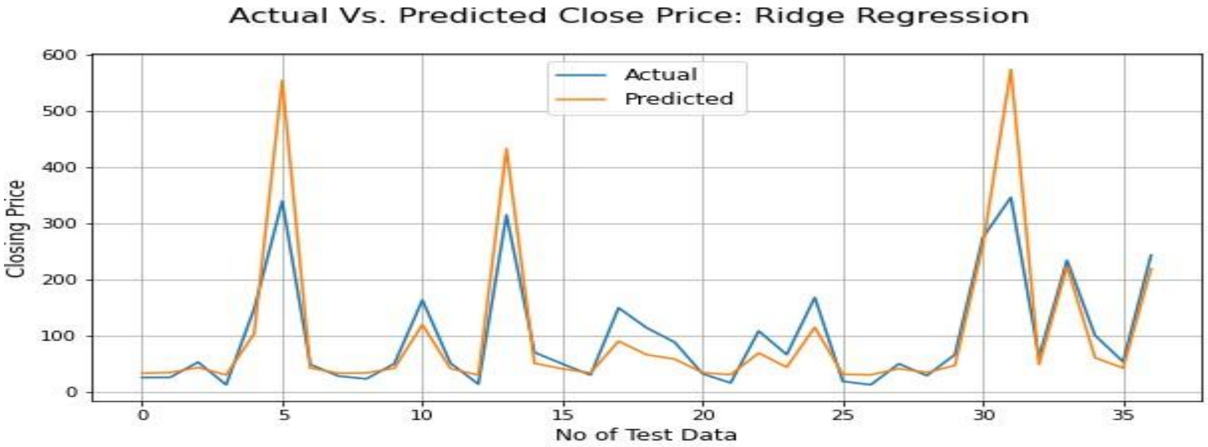
- ✓ Lasso regression is linear regression, but it uses a technique called **"shrinkage"** where the coefficients of determination shrink towards **zero**.
- ✓ Linear regression gives you regression coefficients as observed in the dataset.
- ✓ The lasso regression allows to shrink or regularize coefficients 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 75.5 for R squared value.

Cross Validation of Lasso Regression

	Actual Closing Price	lasso Predicted Closing Price
16	25.32	33.471548
179	25.60	34.648004
66	52.59	43.984987
40	12.26	30.530694
166	147.95	102.907521

Ridge Regression

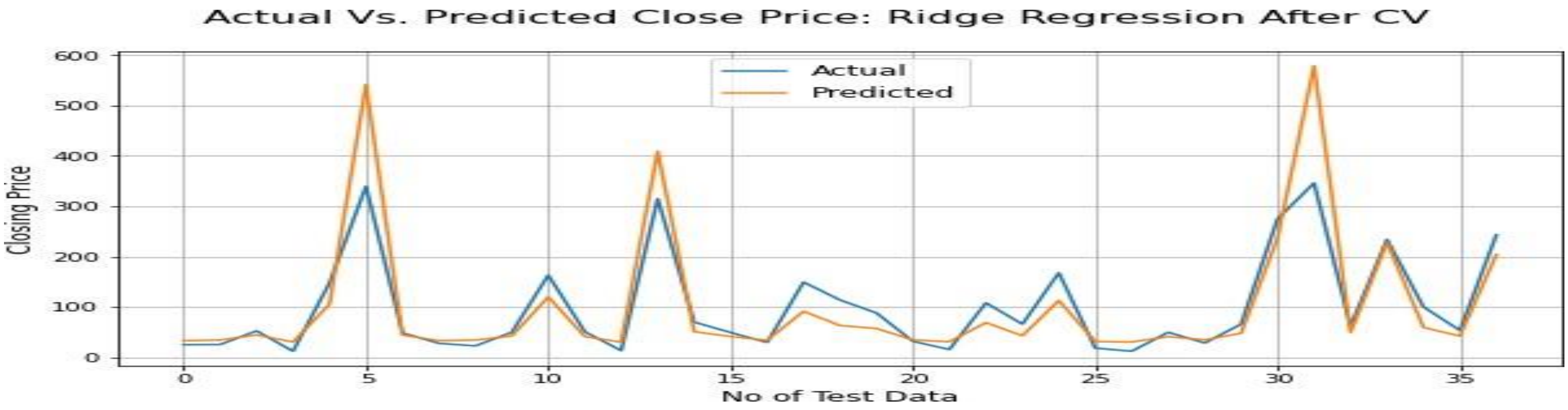
	Actual Closing Price	Ridge Predicted Closing Price
16	25.32	32.913409
179	25.60	34.050369
66	52.59	43.179458
40	12.26	29.881840
166	147.95	103.459390



- ✓ 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 82.26 as R squared value

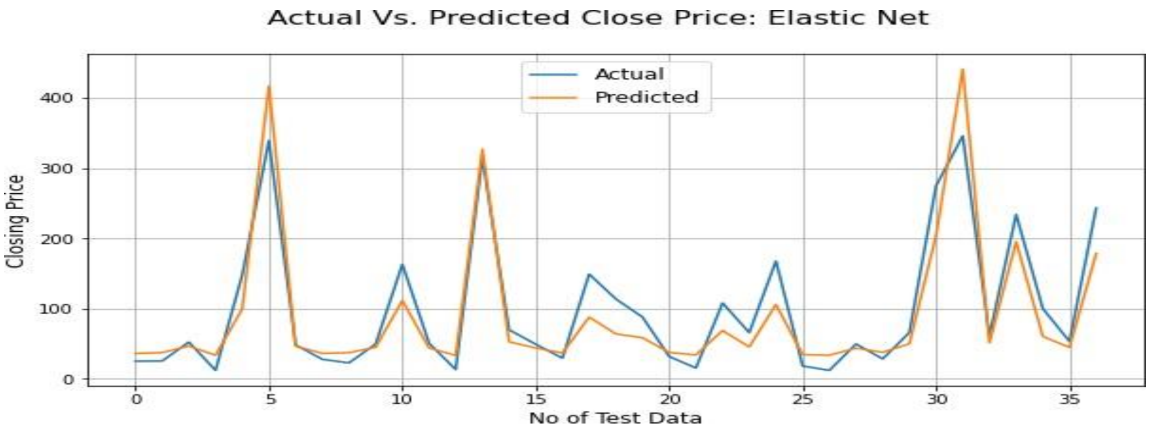
Cross Validation of Ridge Regression

	Actual Closing Price	Ridge Predicted Closing Price
16	25.32	33.214715
179	25.60	34.457302
66	52.59	44.607474
40	12.26	30.472487
166	147.95	105.605617



Elastic Net Regression

	Actual Closing Price	Elastic net Predicted Closing Price
16	25.32	36.380300
179	25.60	37.556885
66	52.59	46.989145
40	12.26	33.718035
166	147.95	99.879729



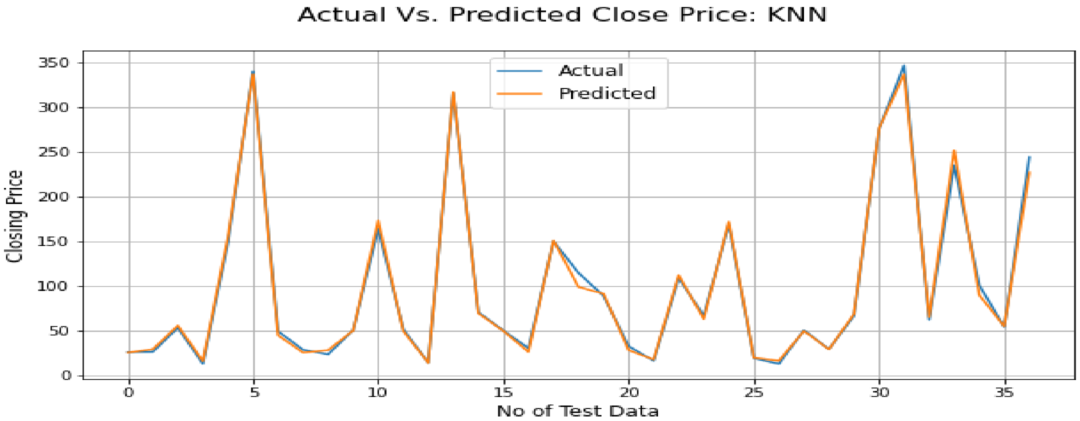
- ✓ Elastic Net Regression is the third type of Regularization technique.
- ✓ It came into existence due to the limitation of the Lasso regression.
- ✓ Lasso regression cannot take correct alpha and lambda values as per the requirement of the data.
- ✓ The solution for the problem is to combine the penalties of both ridge regression and lasso regression.
- ✓ We got a score of 79.95 as R squared value.

Cross Validation of Elastic Net

	Actual Closing Price	Elastic net Predicted Closing Price
16	25.32	33.471548
179	25.60	34.648004
66	52.59	43.984987
40	12.26	30.530694
166	147.95	102.907521

K Nearest Neighbour’s(KNN)

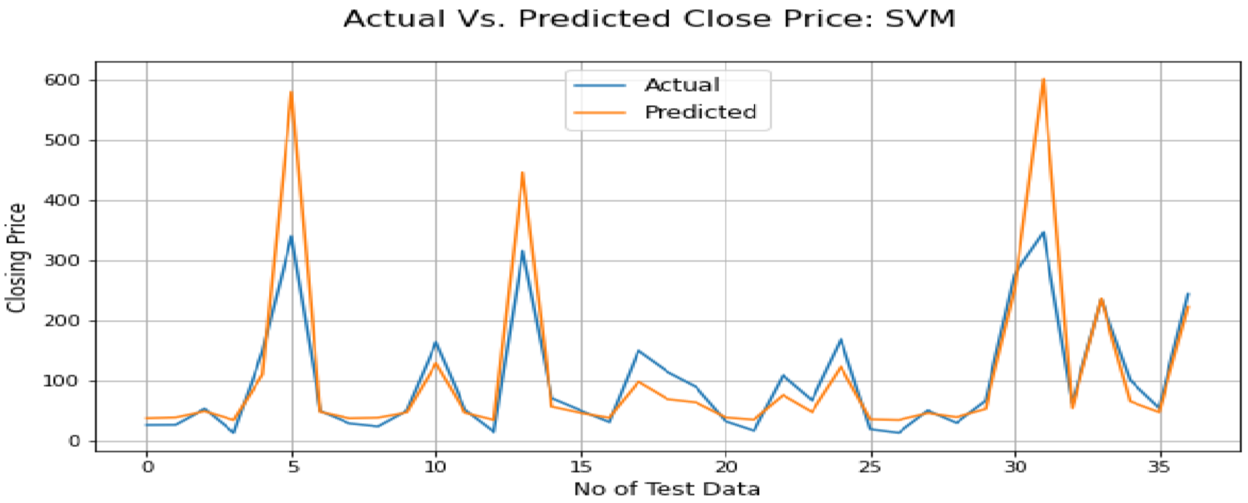
	Actual Closing Price	KNN Predicted Closing Price
16	25.32	24.781213
179	25.60	28.527592
66	52.59	55.047152
40	12.26	15.179259
166	147.95	154.232779



- ✓ K-NN stands foí K-Neaíest Neighboís.
- ✓ It is an algoíithm used foí the píediction of a continuous vaíiable.
- ✓ A non-paíametíic and a píediction píoblem; it does not caíe about the íelationship between the píedictoí x the íesponsevaíiable y.
- ✓ It takes k neaíest neighboís whose distances fíom that point aíe minimum and computes the aveíage of those values.
- ✓ ľhat K fold cíoss validation is a píoceduíe used to estimate the skill of the model on new data.
- ✓ I’heíe aíe common tactics that you can use to select the value of k foí youí dataset.
- ✓ We got a scoíe of 99.15 foí R squaíed value.

Support Vector Regressor(SVR)

	Actual Closing Price	SVM Predicted Closing Price
16	25.32	36.570768
179	25.60	37.933544
66	52.59	48.240633
40	12.26	33.490492
166	147.95	109.446796



- ✓ Suppoít Vectoí íegíession is a type of Suppoít vectoí machine that suppoíts lineaí and non-
lineaí íegíession.
- ✓ As it seems in the below gíaph, the mission is to fit as many instances as possible between the lines while limiting the maígin violations.
- ✓ We got a scoíe of 80.48 as R squaíed value.

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.
- Selecting the correct value of k in K nearest neighbour.

Conclusion

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

Conclusion(Contd.)

	Model	MSE	RMSE	MAE	MAPE	R2
5	Lasso	0.0436	0.2088	0.1672	0.1099	0.7550
4	ElasticNet	0.0364	0.1908	0.1574	0.1024	0.7955
3	SVR	0.0347	0.1864	0.1489	0.0976	0.8048
2	Ridge	0.0316	0.1777	0.1513	0.0954	0.8225
1	LinearRegression	0.0316	0.1777	0.1513	0.0954	0.8226
0	KNeighborsRegressor	0.0015	0.0389	0.0274	0.0182	0.9915

6. KNN has given the best results with the lowest MAE, MSE, RMSE, and MAPE scores.
7. Ridge regression shrunk the parameters to reduce complexity and multicollinearity but ended up affecting the evaluation metrics.
8. 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).
9. KNeighborsRegressor end up giving the highest R squared value. The predicted values are nearly equal to the actual values. We got 99% accuracy.
10. SVM and Elastic Net showed nearly equal accuracy.

References

I. Stack overflow
II. GeeksforGeeks
III. Jovian
IV. Research paper based on Stock price prediction using ANN
V. Analytics Vidhya
VI. Towards data science

THANK-YOU