Yes Bank Stock Closing Price Prediction

Swapnil Wankhede

Data Science, Trainee

AlmaBetter, Bangalore

Abstract

The main objective is to predict the stock closing price of the month of yes bank. Multiple Linear Regression, Ridge Regression, LASSO Regression and Elastic Net Regression are implemented on the dataset which contains 185 row labels and 5 features. Closing price is considered as dependent feature and opening price, lowest price and highest price are independent features. It is found that Elastic Net Regression is the best performing model with adjusted R squared is 0.9931. F-test is used to test the significance of regression. goodness of fit test is used to check how well the model fit to the data. t-test is used.

Keywords: stock prediction, predictive model, multiple regression, Ridge, LASSO, testing etc.

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 whether 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. The main objective is to predict the stock's closing price of the month.

Introduction

The objective of this project is to predict the stock's closing price of the month. Many more things have been explored and analysed with the help of the given dataset which contains 185 row labels with 5 variables which are numerical values except date variable including no missing and no duplicates values.

Pandas, NumPy, matplotlib, seaborn, sklearn, scipy are the libraries are used to explore, analyse and implement the model to given data. Histogram, Box Whisker plot, line plot, heatmap, etc used to visualize the data. Loading the data into the data frame, Exploratory data analysis, Feature engineering and data pre-processing, Forming assumptions and obtaining insights, various models' implementation are the roadmap was decided to come into proper conclusion.

Exploratory Data Analysis

EDA is an approach or philosophy for data analysis that employs a variety of techniques to uncover underline structure, extract important variables, detect outliers or anomalies, test underline assumptions, develop parsimonious model and determine optimal factor settings.

First, a directorial path for the dataset is created using pandas read function. Data has a shape (185, 5) it means 185 row labels and 32 features. It is found that data do not contain duplicate, missing values. It is converted date column to a proper datetime datatype and set as index as we need to track variation in stock price on different months.

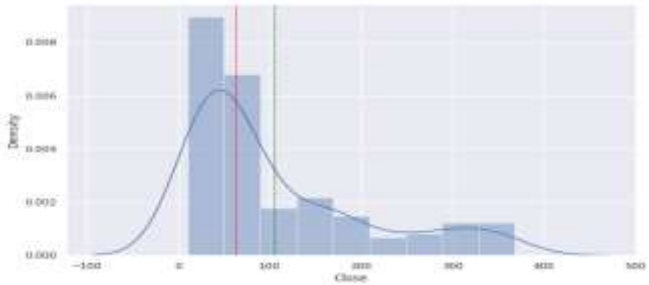
Following are variables within dataset have been elaborated with basic meanings:

Date	The date (Month and Year provided)
Open	The price of the stock at the beginning of a particular time period.
Close	The trading price at the end (in this case end of the month)
High	The Maximum price at which a stock traded during the period.
Low	The Lowest price at which a stock traded during the period

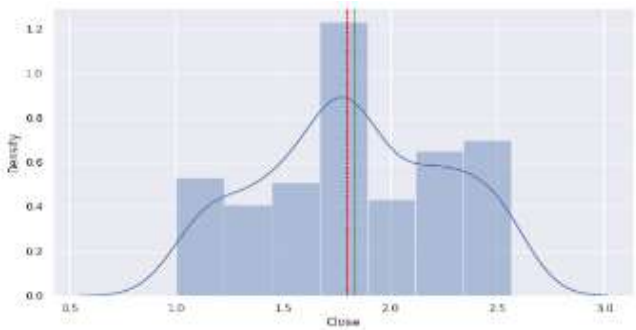
The objective of this project is to predict the stock's closing price of the month so the closing price of stock have to be considered as dependent feature whereas rest of the features are independent features for model implementation.

There are four components of EDA:

1) Resistance: Parametric tests are based on the mean estimation, which is sensitive to outliers or skewed distribution. Median is highly resistant. It is insensitive to change in extreme values while mean highly non-resistant. It is shown in histogram what the data tell us.



Before log Transformation



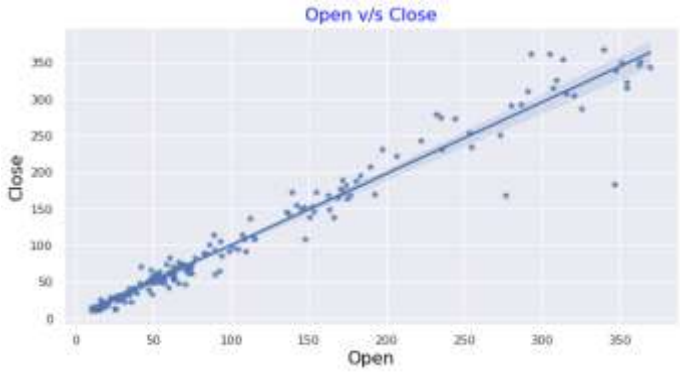
After log Transformation

From the graph, It is observed that how mean is sensitive to outliers as the given data is positive skewed. After applying the lots of transformation, It is found that log transformation is good one as a consequence, the given data looks normally distributed. Log transformation is applied to dependent and independent variable. Here just showed the graph of dependent variable.

2) Residual Analysis:

EDA follows the model that data
= model + residual

The residual or the error is the values that deviate from that expected value. By examining the residuals, we can assess the model adequacy. In this regression model analysis of set of data is not complete without careful examination of unusual behaviours in the data before transformation. Variance around the regression line is same at origin sided portion after that it is showing heteroscedasticity from the scatter plots as shown in



Before log Transformation

After the transformation, It means that the variance around the regression line is the same for all values of the predictor variables is shown by graph

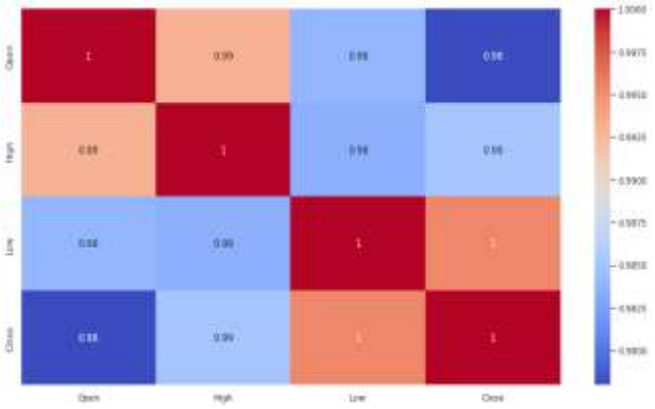
After log Transformation

Hence the assumption of linear regression is valid here.

3) Re-expression: It means proper choice of scale or simplifying the analysis of the data. the given data is not expressed properly. log-Transformation is applied to the data. It is come to conclusion that

- Distribution of features are similar to normal distribution. The mean and median values are nearly same.
- Log-transformation diminishes the outlier's effect and heteroscedasticity

4) Data Visualization: It is easier to detect a data pattern from a picture than from a numeric output. It is easy spot univariate outliers in one-dimensional charts such as a histogram, box-whisker plot. The correlation matrix helps to visualize the correlation among features.



It is found that the corelation between the variables is very high as shown in above graph.

Forming assumptions and obtaining insights:

01)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) Hypothesis Testing: Let's 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.

$-1 \leq \rho \leq 1$

we calculate the value of the test statistic given that $\rho = 0.98$ from heatmap matrix using the following formula-

$$t = \frac{\rho\sqrt{(n-2)}}{\sqrt{(1-\rho^2)}}$$

$t = 66.6199$

To obtain the p-value, we need to compare the test statistic to a t-distribution with 183 degree of freedom (since 185-2 = 183) we get the p- value is equal to 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:

Models are implemented on given dataset. By splitting the dataset into training and testing dataset. Models are fitted on training dataset, learn parameters and make the models predications on the test dataset. The performance of metrics:

Mean square error (MSE): It is defined as "the average of squared difference between the target value and the value predicted by the regression model."

Root Mean square error (RMSE): It is defined as "the square root of mean square error."

Mean Absolute error(MAE): It is defined as, " the absolute difference between the target value and the value predicted by the model and averaging it across the dataset.

R^2(R-squared): Coefficient of determination R^2 is defined as, " It is measure of variability in dependent variable explained by independent variables.

Adjusted R^2: It depicts the same meaning as R^2 but is an improvement of it. It adjusts for the increasing predictors and only shows improvement if there is a real improvement.

Following models are implemented on the given data.

1)Multiple linear regression

2)Ridge regression

3)LASSO regression

4)Elastic Net Regression

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.

Assumption:

1) the relation between the dependent and independent variables is linear as shown in scatter plot.

2) Mean of residual should be zero or close to o as much as possible.

3)variance around the regression line is the same for all values of the predictor variable.

4) there should not be multicollinearity in regression model.

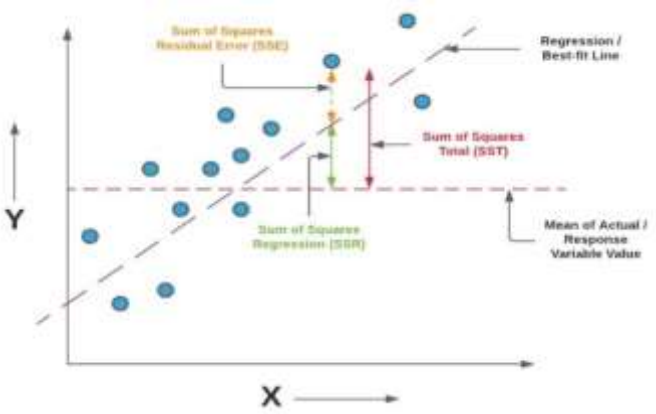
All the assumption of multiple linear regression are satisfied. The hypothesis testing is applied to the significance of regression coefficients and to test the goodness of fit in linear regression models

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

H_0: $\hat{}\beta1 = \hat{}\beta2 = \hat{}\beta3 = 0$ (Regression model does not exist)

H_1: $\hat{}\beta1 \neq \hat{}\beta2 \neq \hat{}\beta3 \neq 0$ (Regression model does not exist)

The above hypothesis can be tested using statistical test such as F-test.



In the above diagram, the variance explained by the regression model is represented using sum of squares regression (SSR). The variance not explained by the regression model is the sum of squares for error (SSE). The f-statistics is defined as-

$$f = (SSR/DFssr) / (SSE/DFsse)$$

DFssr = Degree of freedom for regression model; The value is equal to the number of parameters which is p = 4.

DFsse = Degree of freedom for error; The value is equal to the total number of records (N) minus the number of coefficients (p) equal to 181.

The Tabulated F-Statistics of 185 observation is F tab (3, 181) = 2.60 (from F_ Distribution tables used in analysis of variance) with 3 degree of freedom of sum of quare of regression and 181 degree of freedom of sum of square of residual which are the source of variation.

F_{Cal} = 57.6666 (approx.)

Here, $F_{cal} > F_{tab}$ (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. $\hat{}\beta1 \neq \hat{}\beta2 \neq \hat{}\beta3 \neq 0$

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 following graph is showing how the actual and predicted closing price of stock fit to testing data.

02) Ridge Regression:

It is regularized linear regression model is very similar to least square, except that the coefficients are estimated by minimizing a slightly different objective function. We minimize the sum of residual sum of square and a "penalty term" that penalizes coefficient size. The Ridge regression minimizes RSS + $\lambda \sum_{j=1}^{p} \beta j$. Where λ is a tuning parameter that seeks to balance between the fit of the model to the data and the magnitude of the model's coefficients.

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 following graph is showing how the actual and predicted closing price of stock fit to testing data.

03) LASSO Regression:

It is regularized linear regression model. It stands for least absolute shrinkage and selection operator. The LASSO regression minimizes RSS + $\lambda \sum_{j=1}^{p} |\beta j|$. 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 following graph is showing how the actual and predicted closing price of stock fit to testing data.

04) Elastic Net Regression:

It uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

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 following graph is showing how the actual and predicted closing price of stock fit to testing data.

Comparison Among Implemented Models

Several models are implemented on the given dataset in order to predict the closing price. The performance metric value of several models is in the table

	Metric	Linear Reg.	LASSO Reg.	Ridge Reg.	Elastic Net Reg
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

It is found that Elastic Net Regression is the best performing model with adjusted R^2 is o.9931.

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 fit a given testing set of data.

We came to conclusion that All models fit 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.

Conclusion

The dataset was used that contains data about Yes Bank Stock Prices. Exploratory Data Analysis, Feature Selection and Data Pre-processing and various models' implementation are done and came to conclusion that-

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 α = 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. $\hat{}\beta1 \neq \hat{}\beta2 \neq \hat{}\beta3 \neq 0$. 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 fit 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.

Challenges:

To make the hypothesis and use of proper parametric test were the challenges.

Future work:

Time Series model can be used as predictive model for predicating yes bank stock closing price.

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