Yes Bank Stock Closing Price Prediction

Sayan Bandopadhyay
Data science trainee,
AlmaBetter, Bangalore

Abstract

Closing stock is the amount of inventory that a business has on hand at the end of an accounting year. The amount of closing stock is to be ascertained by physically counting the inventory. This can also be determined by the perpetual inventory system to arrive at the end record of the number of closing stock or inventory.

Here, I am provided with a stock prices dataset of 'Yes Bank' so as to predict how the stock closing price of the bank gets influenced by other related features. For the purpose of prediction, different machine learning models has been used.

Keywords: machine learning, Yes Bank, stock prices, inventory

1.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. The 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.

2. Steps Involved

- **Framing Questions**: This is the very first step of the analysis process. Going through the problem statement deeply and to extricate some hidden questions that problem statement is throwing to us is the main objective here as based on which the analysis will choose its path.
- **Data Inspection**: Second step comes out to be the inspection of the dataset. Checking the structure and features of the dataset is the main objective here which includes checking of the dimension of the dataset, presence of null values and missing values and its treatment (if any) which is basically data cleaning.
- **Data Pre-processing**: Third step is data pre-processing where feature overview that is segregating independent and dependent variable and of course feature selection is the main objective here.
- **EDA**: Exploratory data analysis takes the fourth step where visualization methods are used to so as to check different basic insights that the dataset tells and also to check the distribution of all the features in the

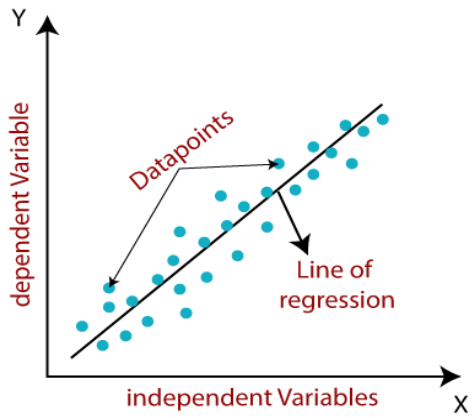
dataset. Here, the visualization techniques also helped so as to check correlation score between each of the variables.

- **Model Implementation**: Fifth is all about splitting data into training data and test data and fitting different machine learning algorithms for prediction of target variable or dependent variable.
- **Metric Evaluation**: This particular step has the objective of comparing the evaluation metrics of each of the models so as to find out the best fit model for this particular case.
- **Conclusions**: Last step is all about summing up the insights and model performances observed from analysis process.

3. Algorithms

In this case, four different supervised machine learning algorithms has been taken into consideration. They are –

➢ **Linear Regression**: It is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as '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. The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:

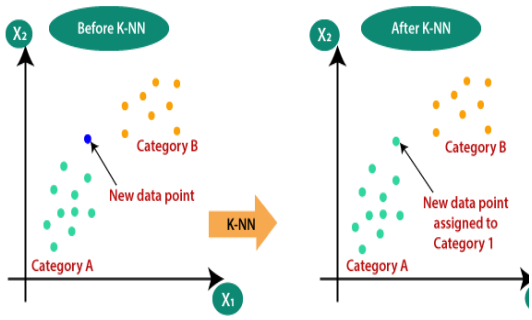


➢ **Ridge Regression**: Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

➢ **Lasso Regression**: The word "LASSO" stands for **L**east **A**bsolute **S**hrinkage & **S**election **O**perator. It is a statistical formula for the regularization of data models and feature selection.

- **KNN:** Stands for K-Nearest Neighbours is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It

assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.



4. Challenges Faced

- Data transformation took some more effort as all the variables in this dataset was right skewed and needs to be transformed into normally distributed variables.
- Choosing best models to implement and tuning those models was a challenge so as it get the best parameters for each model.
- Evaluation metrics for all models was very similar to each other which posed a challenge to choose the best one out of all the models implemented.

5. Conclusion

Going through all the predictive analysis process, it was concluded that the fraud case involving the bank affected its stock price fluctuations and based on model implementation 'linear regression' came out to be the best performing model in this case whereas 'KNN' performed the worst out of all.

References –

1. GeeksforGeeks
2. My Great Learning
3. Analytics Vidhya