VIT-AP
UNIVERSITY

STOCK MARKET PREDICTION USING RANDOM FOREST ALGORITHM IN R

GROUP-13

ATHRWA DESHMUKH - 19BCE7381

HARSHITHA KATA - 19BCE7377

K.TEJA SWAROOP - 19BCE7358

V.RAMAKRISHNA - 19BCE7378

CSE4027 - Data Analytics

Fall sem: 2021-2022

Submitted to:

Dr. SENTHIL MURUGAN

Submission

Date: 28 December 2021

ABSTRACT

Prediction of Stock price is nowadays an existing and interesting research area in financial and academic sectors to know the scale of economies. There did not exist any significant set of rules to estimate and predict the scale of share in the stock exchange. Many evolutionary technologies are existing such as technical, fundamental, time, statistical and series analysis which help us to attempt the prediction process, but none of the methods are proved as reliable and accurate tools to society in the estimation of stock exchange or share market scales. Here, we attempted to do innovative work through R programming to predict or sense the behavior tracking of the stock market sensex. Linear regression, Support Vector regression, Decision Tree, Random Forest Regressor and Extra Tree Regressor are implemented effectively in predicting the stock prices and define the activity between the exchanges and the securities between the buyers and sellers. We predicted the price of the stock based on the closing value and stock price. An algorithm with high accuracy we do the process of comparison for the accuracy of each of the models and finally is considered a better algorithm for predicting stock price. So, we used the Random forest algorithm. As the share market is a vague domain we cannot predict the conditions that occur, and also the share market can never be predicted. This job can be done easily and technically through this work and the main aim of this is to predict the stock prices.

INTRODUCTION

Stock Market forecasting is an act of analyzing and trying to diagnose the market worthiness of the stock existing in the company; it acts like a reliable instrument of financial growth of the company to trade on an exchange. Commerce and trade are the two economical elements that play a vital role in the evolution towards the economy of the nation in a wide range such as industry market and investors. In this process we can easily predict the value of the stock in case of price rise and praise fall at any period of time. Stock marketing is the primary source of any industry whether it is private sector or public sector to raise their funds for business expansions and also further growth of the company. The best actors are investors and industry involved in this stock exchange process of their securities. This is based on the pure concept of economic policy of demand and supply. For example if the demand of stock of the particular company decreases, there is always a fall in the price of the share of that particular company. Efficient market hypothesis is the technical theoretical and experimental challenges that are the motivational for getting efficient results of marketing. The stock prices completely reflect variable information about the constituents and the opportunities for earning abundant profits. The New York Stock exchange is one of the most successful stock exchanges worldwide. It is the world's No: 1 stock exchange that gives the reliable services those who seek. Many industries and companies nowadays are involved in this process. This contains huge sets of data which is difficult to extract, analyze and extract information. This is a big task to the users on the manual process. The market pattern and the prediction of time of purchase of stock is revealed by the analysis of the stock market. Significant profits can be achieved if there is a successful prediction process taken place. The historic data of the market represents varying conditions and helps in confirming the time series pattern has statistically significant predictive power and has high possibility of profitable trades and returns in the investment for competitive business.

PROBLEM STATEMENT

The stock market appears in the news every day. You hear about it every time it reaches a new high or a new low. There are just as many losses caused as the profit and it is difficult for beginners to invest in the stock market. The return of investment and business opportunities in the Stock market can increase if an efficient algorithm could be devised to predict whether the stock price will increase or decrease of an individual stock.

ABOUT THE DATASET

The dataset was downloaded from Kaggle, and it contains 3,40,744 training samples of which we used 1,00,000 training samples divided into train and test data in 80-20 ratio after removal of outliers.

Features in Dataset:

- Date
- Symbol
- Open
- Close
- Low
- High
- Volume
- Class

METHODOLOGY AND MODEL

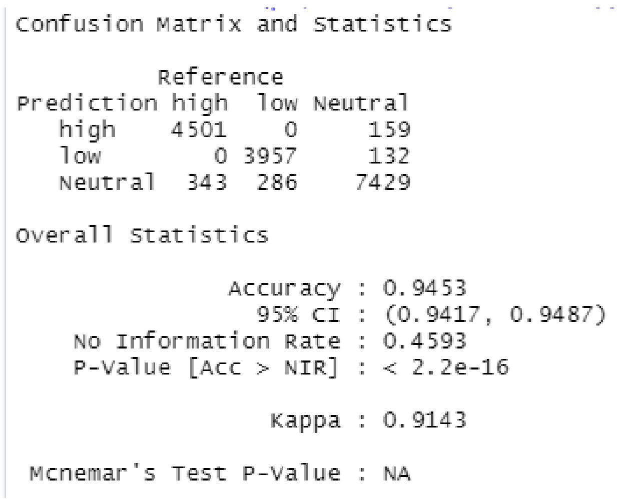
Decision trees are great for classification problems. The problem with using them is that they tend to overfit the training data, because if they are grown really deep they tend to learn the highly irregular patterns found in that particular data set. Thus, the Random Forest model is used instead, because it eliminates the problem of overfitting by training multiple decision trees on different subsamples of the feature space. The idea behind random forests is simple. First, the data is split into different partitions. Then a certain number of random features is used to create and train several decision trees. Each tree will then output a prediction. Each prediction will then be calculated for the number of votes, and the prediction with the highest number of votes will be the final prediction.

CODE DESCRIPTION

First the dataset is read into a data frame. Of the 3,40,744 rows present in the dataset we have used only the first 1,00,000 rows. We then checked for null values in the dataset. After that boxplot was used to detect the outliers and the outliers were then removed, reducing the training sample to 84,036 rows. We then confirmed that no null values or outliers were present. To have a better understanding of the dataset, data visualization was performed for various parameters. We then split the data into train and test in 80-20 ratio. The model was then trained using the train dataset and using 5 features overall (open, close, low, high, volume). The value of n was set to 100 i.e. 100 random decision trees were generated to train the model with 4 randomly selected attributes every time. The test data was then used for prediction of the class (high, low, or neutral) by the model and the confusion matrix was displayed along with the accuracy and several other parameters.

RESULTS

Based on our model and dataset, we got an accuracy of **94.53%** for the testing dataset.



```
Confusion Matrix and Statistics

          Reference
Prediction high  low Neutral
   high    4501    0     159
   low        0 3957     132
   Neutral  343  286    7429

Overall Statistics

               Accuracy : 0.9453
                 95% CI : (0.9417, 0.9487)
    No Information Rate : 0.4593
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9143

 Mcnemar's Test P-Value : NA
```

- **Outcome 1**

We tried including the outliers in the model and found out that we were getting comparatively low accuracy, but after removing the outlier and then training the model resulted in a very high accuracy.

- **Outcome 2**

We do not need to scale/normalize the data while using random forest and we could train the model in such a way as to predict whether the stock price will go higher, lower or remain neutral.

CONCLUSION

The main objective was to develop a prediction mechanism for the stock market prices in order to help with the decision making for investment. This objective has been accomplished.

Additionally data visualization provides us with a good understanding of the large data without really having to go through it.

REFERENCES

https://www.journaldev.com/47986/outlier-analysis-in-r

https://www.geeksforgeeks.org/data-visualization-in-r/

https://lamfo-unb.github.io/2017/07/22/intro-stock-analysis-1/

THANK YOU