ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/354085600

Stock Market Analysis Using Linear Regression and Decision Tree Regression

Conference Paper · August 2021

DOI: 10.1109/eSmarTA52612.2021.9515762

CITATIONS
13

READS
4,975

3 authors, including:

Rezaul Karim
Chongqing University
3 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Rezaul Hossain
Daffodil International University
1 PUBLICATION **12** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

ERP Solution View project

All content following this page was uploaded by Rezaul Karim on 14 October 2021.

The user has requested enhancement of the downloaded file.

2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA) | 978-1-6654-4078-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ESMARTA52612.2021.9515762

Stock Market Analysis Using Linear Regression and Decision Tree Regression

Rezaul Karim*
School of Big Data & Software engineering
Chongqing University
Chongqing, China
Rezaulkarim94@gmail.com

Md Khorshed Alam
School of Big Data & Software engineering
Chongqing University
Chongqing, China
khorshedalam49@gmail.com

Md Rezaul Hossain
Daffodil International University
Dhaka, Bangladesh
rezaulhossaincse@gmail.com

***Abstract*— In business, the Stock market or Share market is a more perplexing and sophisticated way to do business. Every business owner wants to reduce the risk and make an immense profit using an effective way. The bank sector, brokerage corporations, small ownerships, all depends on this very body to earn profit and reduce risks. However, using the machine learning algorithm of this paper to predict the future stock price and shuffle by using subsist algorithms and open source libraries to assist in inventing this unsure format of business to a bit more predictable. The proposed system of this paper works in two methods – Linear Regression and Decision Tree Regression. Two models like Linear Regression and Decision Tree Regression are applied for different sizes of a dataset for revealing the stock price forecast prediction accuracy. Moreover, the authors of this paper have revealed some development that could be the club to acquire better validity in these approaches.**

Keywords *—Data Analysis, Linear Regression, Decision Tree Regressor, Big Data, Stock Market Analysis, Supervised Machine Learning.*

I. INTRODUCTION

This is the prediction about stock market and we can solve it by using classification algorithm. So, in this research we use linear regression and Decision tree regression as a classification for prediction stock market.

STOCK MARKET is an ancient technique where people can easily do the trade stock and they can either lose or profit. People who have a misconception about the stock market are that they think it looks like gambling because they just know about profit or loss but nothing else. So the lacking of proper information and analyzing capability people think like that but the revolution of data science, big data, and awareness of the people are getting proper guidelines about the future and passed trading information. The stock market is related to the individual and national economy, so connection to the ethnic economy and it's vital for a nation because of the impact of each country's GDP.

A company sells their stock by the stock exchange then, they listed the price that is called IPO or initial public offering, and then the people buy that share from brokerages. Brokerages work like an intermediate medium between seller and buyer but for that brokerages charge the amount for doing this job [1]. It can be a bank, any small company including licenses, and so on.

Awareness is the main part of the stock market nowadays because needs to have a solid idea about trending stock rate, past and future also. This awareness could understand the upcoming rate of stock and be analyze the risk. Now a day, all of these could possible by the invention of data science, machine learning, and big data. So we can simply say that machine learning and data science is the best gift for the stock market and another relevant field.

The objective of this paper is to get a better decision using two supervised regression machine

978-1-6654-4078-3/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on August 29,2021 at 14:04:48 UTC from IEEE Xplore. Restrictions apply.

learning algorithms and the use of statistic formula gives us better accuracy of stock price predict. Actually, would discuss two regression models using different size of the dataset and getting the performance in a different aspect. So, the main concern about this paper is the machine learning algorithm, statistic, and a graphical view of data from a different outcome. Finally, compare the performance based on the size of the dataset and algorithmically.

II. LITERATURE REVIEW

Correct predict of the stock market is vital for the help of investors as it says that whether it would pay off or not. Many methods have been deployed for stock price prediction but Artificial Neural Network is the first method for predicting stock price trends [3]. In the last two decades, Kim [4] applied SVM to predict the stock market price.

Kavitha S. et al. (2016) [5] had compared historical data between linear Regression and Support Vector Regression. For the regression technique, they used the LeastMedSq function and SMOReg function respectively. LeastMedSq is a method of linear regression that minimizes the median of the squares of withdrawal from the regression line. SMOreg actualizes the support vector machine for regression. The boundaries can be mastered by utilizing different calculations. The calculation is chosen by setting the RegOptimizer. The most well-known calculation (RegSMOImproved) is because of Shevade, Keerthi, et al and this is the default RegOptimizer.

Verma R.et al. (2017) [6] had applied Artificial Neural Networks for prediction of the stock market and they also discussed theoretical knowledge about ANN and its features. For getting better results authors [7] had applied further updated and forward feed the values using backward propagation.

In this era, lots of work already done and ongoing about the stock market price that is referred to for capture optimal approach. Economic research encapsulates two-component trading philosophies [8]. Fundamental analysts practice all about from the overall financial and industry situations to the economical power and maintenance of isolated companies. Technical research differs from fundamental analysis, in that traders try to identify advantageous by searching at statistical trends, like intercourse in a stock's price and volume.

The primary research approach identifies expected stocks by analyzing their basic attributes. Statistics from economical reports like, cash books, balance sheets, and loss or profit statements are used to research the intrinsic worth of companies [9]. Economical radio statistics that subsume conduct corporate valuation, growth equilibrium, corporate valuation, corporate liquidity, and economic leverage form the basis of fundamental attributes [10].

Contrarian strategy is fundamental research of related strategy. This strategy attaches the factor of human feelings trend with fundamental research [11]. A contrarian investor takes in a contrary place in buying shares of stocks that proving poorly after then selling them when they perform better [12].

III. PROPOSED METHODOLOGY

The methodology of a paper is a vital and key feature because this stage discusses the whole process of functional activity. The machine learning algorithm mainly followed the four steps of methodology for analyzing the data and predicting the outcome and taking a decision. All steps are raw data, Data cleaning, and separate as a train and test dataset, fit train data in the model and evaluate the result using the test dataset.

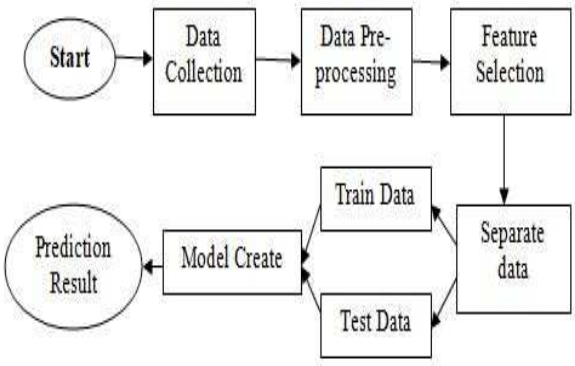

Figure 1. Flow chart of Methodology

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on August 29,2021 at 14:04:48 UTC from IEEE Xplore. Restrictions apply.

A. Dataset

In this stage, we look for collecting data from different sources, and finally, we choose a recent one-year dataset from Amazon [2] and separated into three parts of size (one year, six months, and three months). This dataset attribute is open, high, low, close and volume but we selected close as a label data and the rest of to extract the features that will help me predict the result.

B. Data Cleaning and Preprocessing

During this period, I checked whether data have a null value and unknown types of data that have been cleaned and filled up using a statistical formula. Like, when I founded a null value in a correspondent attribute then I checked types of values either discrete or classifier value. If, its classification value then calculating the median else calculated mean value and put it on the null places.

C. Separate Data into Train and Test Dataset

Prepared data separated into two part train and test the ratio of 80% and 20% respectively. The percentage of trains and tests would impact the accuracy of predicting the result. For three sizes of a dataset, I applied the same ratio for the train and test. At this stage what ration you want to choose for the train and test dataset it's up to you but if you take more train dataset compare to test then accuracy would be better. The general ratio for test and train dataset is 80% and 20% respectively [13].

D. Train Data Fit in Model

Choose an algorithm is vital for getting better predictions and selecting a proper algorithm based on the dataset. For this dataset, I selected Linear Regression and Decision Tree Regression to calculating the prediction of the diverse size of the dataset and train the model using the training dataset.

E. Test the Data

The test data set is 20% of the total and this learning approach applied the test dataset and gets the result then evaluates with the actual output.

IV. EXPECTED RESULT

A. For 100% Dataset (One year)

For machine learning, a comparative study given different predictions and accuracy is this paper. So I am going to show a resulted graph for a year dataset using Linear Regression as a classification and Decision Tree Classification. Using Linear Regression and Decision Tree accuracy becomes 0.9936 and 0.9936 respectively.

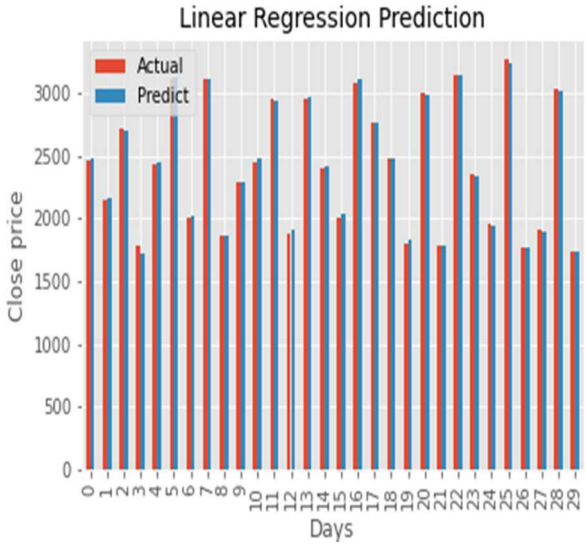

Figure 2. Simple Linear Regression

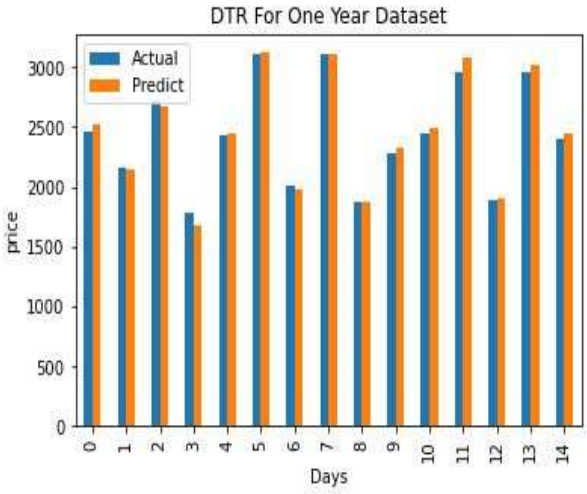

Figure 3. Decision Tree

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on August 29,2021 at 14:04:48 UTC from IEEE Xplore. Restrictions apply.

B. For 50% Dataset (Six Months)

When I divided data into two parts and the recent half part of the data used for both algorithms and showing prediction changed high but when it was about 100% dataset then the result was near both. The result was around 0.9796 and 0.9088 for Linear Regression and Decision Tree Regression respectively.

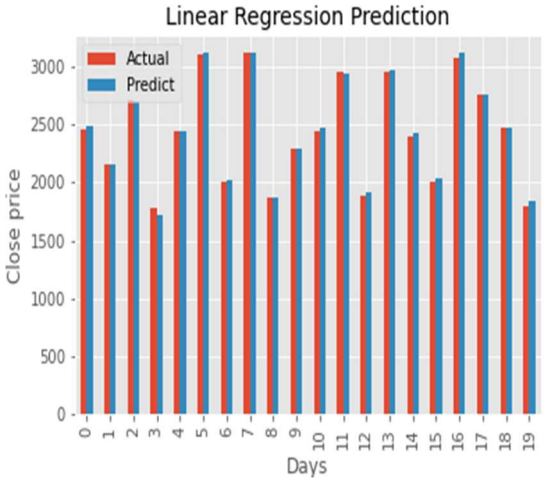

Figure 4. Linear Regression

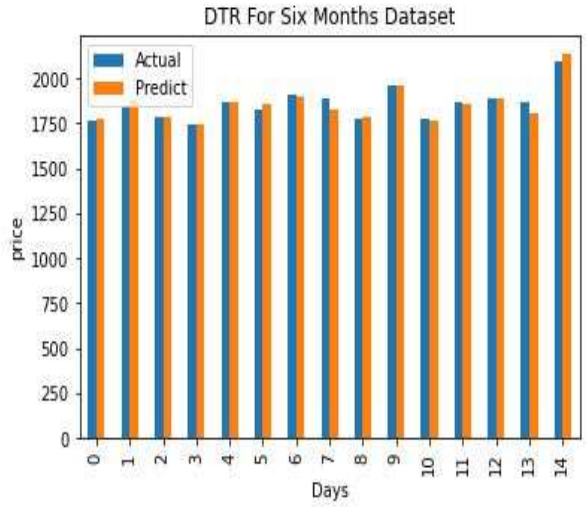

Figure 5. Decision Tree

C. For 25% Dataset (Three Months)

At this stage, I split data from the previous 50% of the dataset to the earlier 25% and applied the same approach, and getting interesting change from both algorithms. When Linear Regression gave me a prediction of 0.9723 then Decision Tree Classification was 0.8353 only.

Below the linear regression graph clearly showing that slightly change predicting data from actual data even though the amount of data has been reduced.

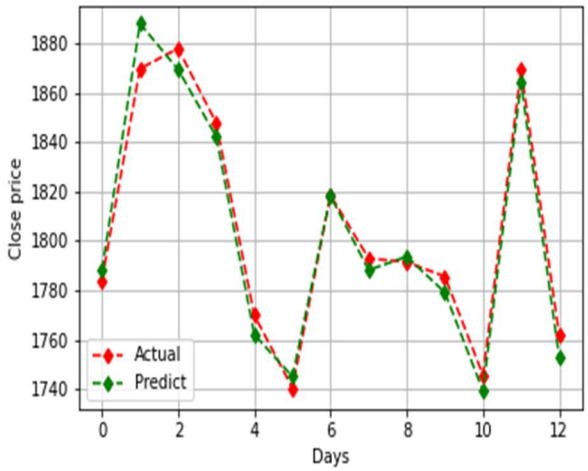

Figure 6. Linear Regression

Beneath the decision tree regression graph given worse accuracy compare to linear regression after reducing the size of the dataset.

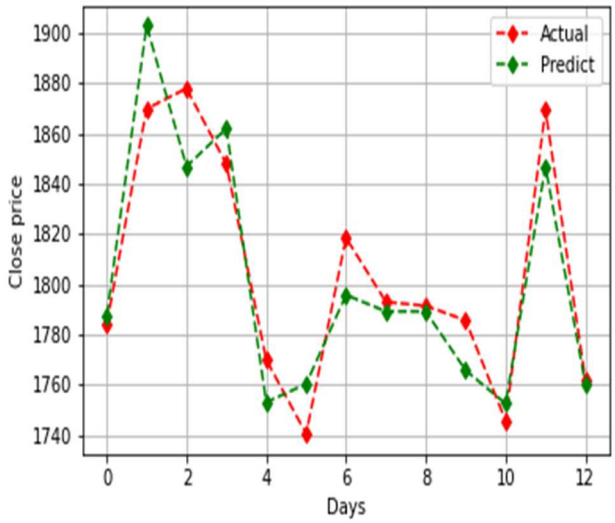

Figure 7. Decision Tree

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on August 29,2021 at 14:04:48 UTC from IEEE Xplore. Restrictions apply.

D. Comparison of those.

This comparative study say that Linear regression is given better prediction compare to Decision Tree Regression even if we change the size of the dataset but we can use both algorithms if we don't split the data into different size because base on the size Decision Tree Regression gives us a poor prediction. So for large and small datasets linear regression is preferable but for a small dataset, Decision Tree Regression gives us less prediction.

Figure 8. Accuracy Comparison

Above the figure, clearly showing that linear regression is given better prediction compare to Decision Tree Regression. While I changed the amount of data into three different parts of the dataset then the prediction of Linear Regression is changed slightly but Decision Tree Regression change is massive. So we could say that a supervised learning regression algorithm for the stock market predicts Linear Regression is the best option compare to Decision Tree Regression.

V. AWARENESS ABOUT MISTAKES

- Clear concept about the algorithm.
- Clean data using a proper statistical formula.
- Clear idea about the dataset.
- Great idea about the independent variable and dependent variable.
- Visual data using visual tools based on the dataset.

VI. FUTURE SCOPE

In this world, machine learning is vitally important for different aspects but among that important data science is very crucial one of them. If you see past one or two eras then data was not so much important compare to know a day. So it's very clear that the future world going to depending based on data science, machine learning, and statistic.

This paper would give so many directions for predict stock market price but we have known that following the linear regression model given to be better performance compare to Decision Tree Regression. The next way could be compared to unsupervised and reinforcement machine learning for the stock price market and hope would be to find great predict much better from the existing result.

VII. CONCLUSIONS

Machine learning has some great application and still, now it's a very popular tool and it's also depending much on data even it has evolved the future into a neural network and deep learning. All about this paper is stock market price predict using machine learning. There are various ways to implement the stock price prediction but applied only two regression algorithm.

The main aim of this paper is to use supervised machine learning algorithm Linear Regression and Decision Tree Regression have been applied for stock price prediction. The result reveals that Linear Regression is given better accuracy for both small and big datasets. On the other hand, Decision Tree Regression expresses the poor prediction price based on the size of the dataset.

REFERENCES

[1] Wikipedia Developer, "About Stock Market, Stock Brokerages and Stock Exchange", https://en.wikipedia.org/wiki/Stock_market

[2] Amazon,"Collected data from Amazon", https://finance.yahoo.com/quote/AMZN/history?p=AMZN

[3] G.Zhang, B.E. Patuwo, M.Y.Hu, Forecasting with artificial neural networks: the state of the art, Int.J.Forecasting14 (1998) 35-62.

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on August 29,2021 at 14:04:48 UTC from IEEE Xplore. Restrictions apply.

[4] K. Kim, Financial time series forecasting using support vector machines. Neurocomputing, 55, pp. 307–319, 2003.

[5] S. Kavitha, S. Varuna, and R. Ramya, "A Comparative Analysis on Linear Regression and Support Vector Regression," in Online International Conference on Green Engineering and Technologies (IC-GET), 2016.

[6] R. Verma, P. Choure and U. Singh, “Neural Networks through Stock Market Data Prediction,” in International Conference on Electronics, Communication and Aerospace Technology, 2017.

[7] G. V. Attigeri, M. P. M M, R. M. Pai, and A. Nayak, "Stock Market Prediction: A Big Data Approach," in TENCON 2015 - 2015 IEEE Region 10 Conference, Macao, China, 2015.

[8] Technical-Analysis, The Trader's Glossary of Technical Terms and Topics. 2005

[9] Graham, Benjamin; Dodd, David (December 10, 2004). Security Analysis. McGraw-Hill. ISBN 978-0071448208.

[10] Walsh, Ciaran (2003) Key Management Ratios, Third Edition, Prentice-Hall.

[11] Shefrin, Hersh (2002) Beyond Greed and Fear: Understanding behavioral finance and the psychology of investing. Oxford University Press.

[12] O'Shaughnessy, James (2009). Predicting the Markets of Tomorrow: A Contrarian Investment Strategy for the Next Twenty Years, Penguin Group. ISBN 1591841089.

[13] Sunny Srinidhi, “How to split your dataset to train and test datasets using SciKit Learn”, https://medium.com/@contactsunny/how-to-split-your-dataset-to-train-and-test-datasets-using-scikit-learn-e7cf6eb5e0d

Authorized licensed use limited to: CHONGQING UNIVERSITY. Downloaded on August 29,2021 at 14:04:48 UTC from IEEE Xplore. Restrictions apply.

View publication stats