International Core Journal of Engineering Volume 7 Issue 3, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0049

Exploring Machine Learning in Stock Prediction Using LSTM, Binary Tree, and Linear Regression Algorithms

Xinzhi Yan[1], Qian Cai[2], Shuhan Zhang[3], Tengkai Yu[4]

[1]University of California, Davis, Davis, California, 95616, USA;

[2]Cardinal Newman High School, Santa Rosa, CA, 95403, USA;

[3]Charlotte Catholic High School, Charlotte, North Carolina, 28226, USA;

[4]International School of Beijing, Beijing, 101318, China.

Abstract

The fluctuations of the stock market make a relatively accurate prediction method almost impossible to achieve. This research focused on the exploration of an accurate prediction model for stocks using trial and error. The research group constructed and applied three different deep learning sequential models, including the linear regression model, decision tree model, and long-short-term memory model (LSTM) to near-future stock-price-prediction. Out of the mentioned sequential models, researchers optimized and employed the best for each trading method. Based on the results, the research demonstrated that LSTM algorithms produced the most accurate predictions out of all three selected models. The work proved tremendous potential in the field of stock prediction using machine learning.

Keywords

Machine Learning; Stock Prediction; LSTM; Data Analysis; Decision Tree.

1. Introduction

Regarding the Stock Market, there are two prominent and well-known methods, namely fundamental analysis and quantitative trading. On the one hand, fundamental analysis is based on a subjective view of the industry or company's current true value and also one's own evaluation of this company's future value. It mostly relies on public information such as newsletters or the ideas behind the company. On the other, quantitative trading employs mathematical models to aid decision making, hence avoiding certain interruptions of human subjectivity and emotion. Different from fundamental analysis, quantitative trading is hard to form a general visualization from just data. Because the work focuses on writing a program that can aid people to make an informative decision. This research focuses on quantitative trading, utilizing three different models – linear regression model, decision tree model, and LSTM. With LSTM being the most advanced algorithm among the three and is designed for data prediction, it's chosen as the prediction method for the investigation's final results. [1]

2. Background and Related Works

Predicting stock market trends using machine learning algorithms has long been practiced by individuals and companies. The most prominent stock market prediction algorithm is the LSTM method, which was proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997. After years of development, the LSTM became one of the most accurate and successful methods. The use of the LSTM algorithms in the Indian Share market published in "Stock Price Prediction Using LSTM on

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0049

Indian Share Market" by Achyut Ghosh and his colleagues is one of the most recent publications involving the use of LSTM. In their paper, Ghosh and his group used variations of the LSTM method and produced a machine-learning model that takes in day-wise past stock information from Indian companies. The final model successfully produced stock predictions and growth of companies with extremely high accuracies. Moreover, improvements are also done on the LSTM methods. Recent research done by Taewook Kim and Ha Young Kim discovered the effectiveness of combining LSTM with other machine-learning algorithms. By combining LSTM and CNN, Taewook Kim and Ha Young Kim's fusion model takes in stock time series and stock chart images and produces results with greatly reduced prediction errors. [2, 3]

2.1 Theoretical Analysis of the Strategic Choice of Enterprise Investment under a Shared Economy

The data that the models utilized in training and testing included daily prices and volumes for every S&P 500 stock from 2013 to 2018. The data also included high and low prices on one day. After processing a large number of iterations in selecting the appropriate input value, the model chose the average of 'close' and 'high' of one day as the input date because the average best represented the fluctuation of that stock per day. The algorithms only analyzed three different companies in the S&P 500 via different methods. The research split the data using an approximate ratio of 4 to 1 for each stock in training and testing data. A sample piece of the data is shown in Figure 1 below.

	date	open	high	low	close	volume	Name
0	2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
1	2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL
2	2013-02-12	14.45	14.51	14.10	14.27	8126000	AAL
3	2013-02-13	14.30	14.94	14.25	14.66	10259500	AAL
4	2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL
5	2013-02-15	13.93	14.61	13.93	14.50	15628000	AAL
6	2013-02-19	14.33	14.56	14.08	14.26	11354400	AAL
7	2013-02-20	14.17	14.26	13.15	13.33	14725200	AAL
8	2013-02-21	13.62	13.95	12.90	13.37	11922100	AAL
9	2013-02-22	13.57	13.60	13.21	13.57	6071400	AAL

Figure.1 Data structure of American Airline from 2013 to 2018 [4]

2.2 Methods

In order to examine the impact of the predicting ability of three different methods in different time periods, the model predicted American Airlines stock in S&P 500 using each method.

Linear regression model: Linear regression is to model the relationship between two variables by fitting a linear equation to observed data. The simplest form of linear regression equation with one dependent and one independent variable is defined by the formula:

$$y = b * x + c$$

Where y is the estimated dependent variable score, c is the constant, be is the regression coefficient, and x is the score on the independent variable. The ease of implementation of this method is the motivation behind it.

Decision Tree Model: A decision tree is an algorithm that partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. There are three steps involved in the building of a decision tree, including splitting, pruning and tree selecting. Splitting is the process of partitioning the data into subsets. Pruning is to shorten the branches of the tree. Tree selection is the process of finding the smallest tree that fits the data. The attribute of this model is its easy visualization and comprehension is.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0049

LSTM: Short for "Long Short-Term Memory", LSTM is an appropriate algorithm to make predictions and processes based on time-series data. The stock market has enormously large historical data that varies with trade dates, which are time-series data. However, the LSTM model predicts the future price of stock within a short-time period with higher accuracy when the dataset has a huge amount of data. The motivation behind this model is that this model is a time series prediction, which makes it perfect for stock predictions. The model used the opening price and closing price from the previous day to predict the opening price of the next day. Then, the model used the predicted opening price to predict the closing price of that day. The process is repeated countlessly until it reaches the target number of days. [5]

2.3 Experiments and Results

The use of linear regression model produced results with an extremely high error rate. The research group started the experiment by inputting the daily prices and volumes for American Airlines from 2013 to 2018; then the research group ran the linear regression model and produced the results below. The graph's x-axis shows the number of days and the y-axis displays the close price of the stock in USD. The blue line presents the historical data; the purple line displays the prediction made by the model; the red line demonstrates the actual trend of the stock. The reason behind the high error rate lies in the fact that the linear regression model's inflexibility. The discontinuity happened because the program implemented tried to draw a best-fitted line through the trained data, and the predicted value started where the best-fitted line ended. Because the graph returned showed great variations from the future trend of the company's stock, the linear regression model was therefore proven to be Unusable.

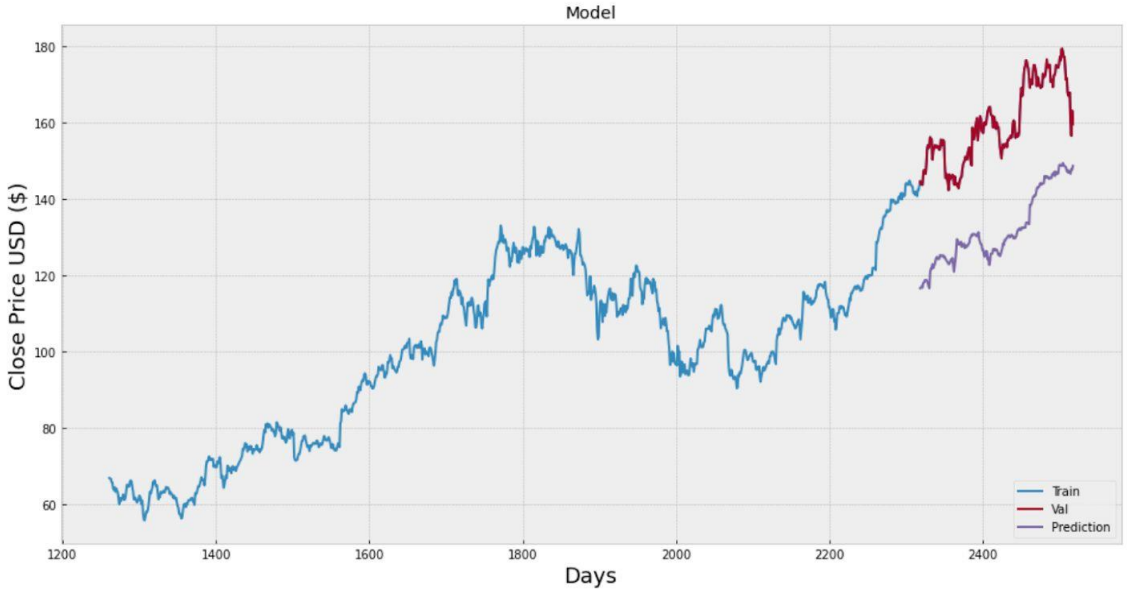

Figure 2. Stock prediction using Linear regression

The use of the decision tree model showed improvements from the linear regression model. However, the data produced showed huge fluctuations. The research group started the experiment by inputting the daily prices and volumes for American Airlines from 2013 to 2018, then the research group ran the decision tree model and produced the results below. The graph's x-axis shows the number of days and the y-axis displays the close price of the stock in USD. The blue line shows the historical data, the purple line shows the prediction made by the model, and the red line shows the actual trend of the stock. Although the prediction made by the decision tree model follows the actual trend of the stock, which is a huge improvement from the linear regression model, it still produced unusual fluctuations. The huge deviations from the actual data implied that the model was very inconsistent and therefore unsatisfactory for stock predictions. The reason behind this huge fluctuation is because when the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0049

program splits the dataset, a value in the dataset means to the computer that the stock could go up or down however sometimes that value had the wrong indication for the computer to understand thus giving some unusual fluctuations.

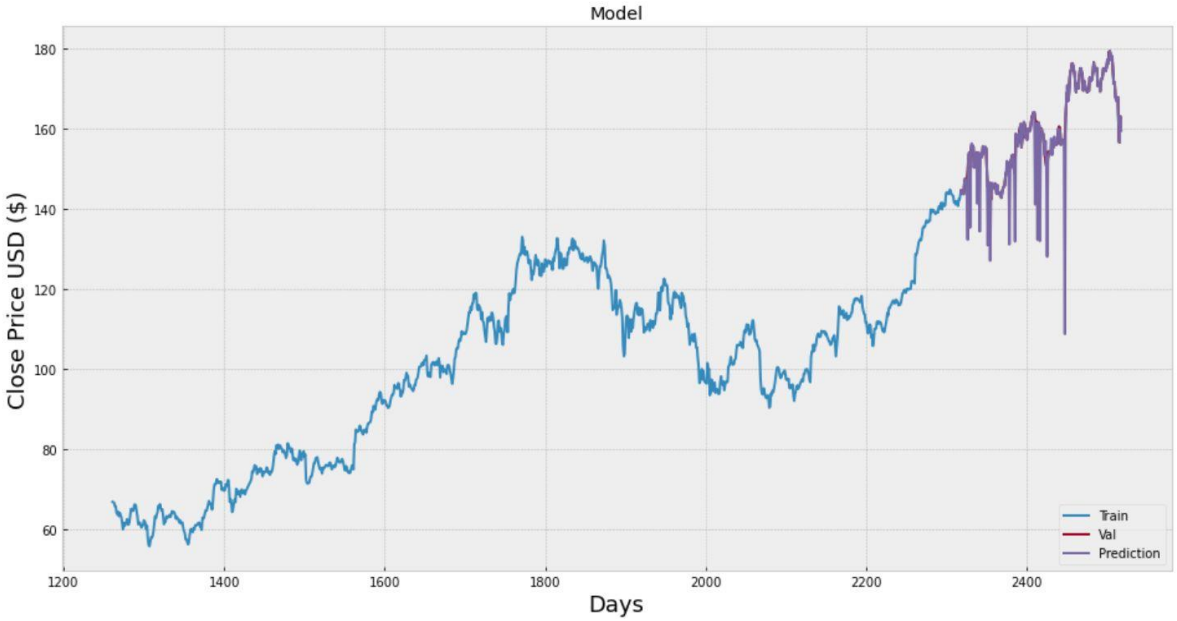

Figure 3. Stock prediction using decision tree

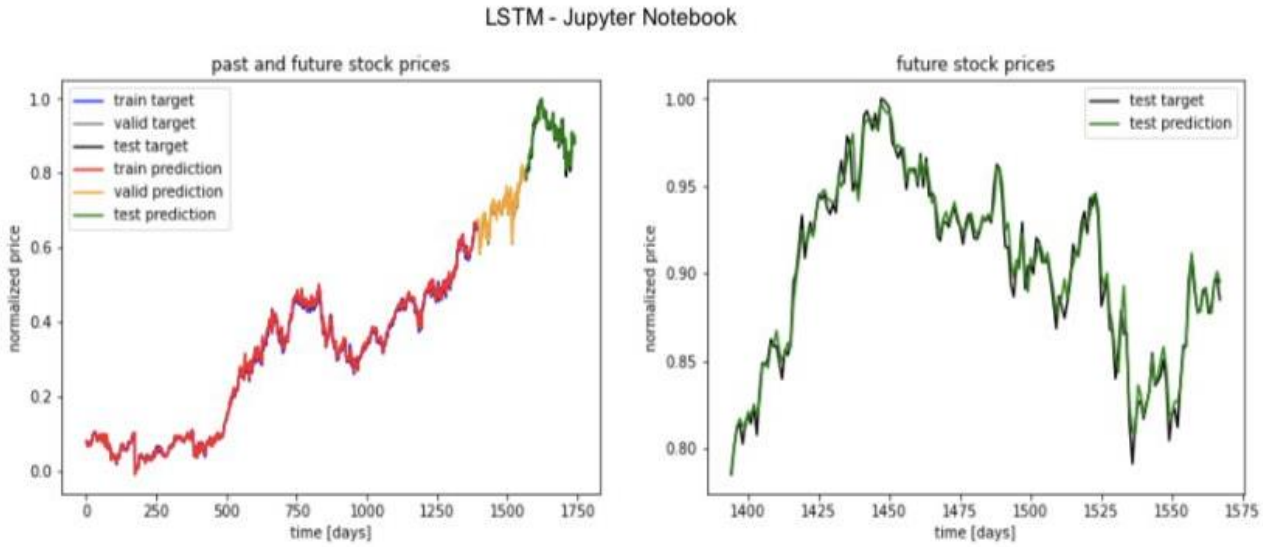

Figure 4. Stock prediction using LSTM

The implementation of the LSTM algorithm showed the highest accuracy rate compared to the previous two approaches. The research group started the experiment by inputting the daily prices and volumes for American Airlines from 2013 to 2018, then the research group ran the decision tree model and produced the results below. In Figure 4, the x-axis represents the number of days since the stock goes public and the y-axis shows the close price of the stock in USD. Figure 4 is made up of two graphs: the magnified image of the testing target on the right and the testing prediction on the left. The legends in Figure 4 include blue line, gray line, black line, red line, yellow line, and green line, which respectively stands for the train target, valid target, test target, train prediction, valid prediction, and test prediction. The input of the dataset is the price of the previous day opening and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0049

closing prices and output is the next day opening price. Then, the next input is the previous day closing price and next day opening price, the output is the next day closing price. The accuracy is on point thus the research conclude that LSTM is the best among all three methods attempted.

3. Conclusion

Through simulating the three potential stock prediction methods, linear regression, decision tree, and LSTM, we collected and visualized data for determining the most accurate model. Our results show, based on data, that the LSTM model is superior in terms of its accuracy in comparison with Linear Regression and Decision Tree. Therefore, we can conclude that a relatively accurate analysis and prediction of the stock market under a shared economy is achieved through employing LSTM.

References

[1] Shanthacumaran, Thenuja. “Stock Price Prediction Using LSTM (Long Short-Term Memory).” Data Science Central, 13 June 2020, 8:30, www.datasciencecentral.com/profiles/blogs/stock-price-prediction-using-lstm-long-short-term-memory.

[2] Ghosh, Achyut, et al. "Stock Price Prediction Using LSTM on Indian Share Market." Proceedings of 32nd International Conference on Computer Applications in Industry and Engineering, vol. 63, 2019, pp. 101-10, www.easychair.org/publications/download/LKgn. Accessed 20 Sept. 2020.

[3] Taewook, Kim. "Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data." PLOS ONE, 15 Feb. 2019, journals.plos.org/plosone/article?id=10.1371/ journal. pone. 0212320 #:~: text=We%20 propose% 20a% 20model%2C%20called,images%2C%20to %20predict%20stock%20prices. Accessed 20 Sept. 2020.

[4] Nugent, Cam. “S&P 500 Stock Data.” Kaggle, 10 Feb. 2018, www.kaggle.com/camnugent/sandp500.

[5] “Indexing and Selecting Data¶.” Indexing and Selecting Data - Pandas 1.1.2 Documentation, pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html.