LSTM-based Stock Prediction Modeling and Analysis

Ruobing Zhang[1,*]

[1]*Beijing University of Technology, BJUT*
**Ruobing Zhang.Email:1262385605@qq.com*

ABSTRACT
The stock market plays an important role in the economy of a country in terms of spending and investment. Predicting stock prices has been a difficult task for many researchers and analysts. Research in recent years has shown that Long Short-Term Memory (LSTM) network models perform well in stock price prediction, and it is considered one of the most precise prediction techniques, especially when it is applied to longer prediction ranges. In this paper, we set the prediction range of the LSTM network model to 1 to 10 days, push the data into the built LSTM network model after pre-processing operations such as normalization of data, and set the optimal values of epochs, batch_size, dropout, optimizer and other parameters through training and testing. By comparing with Linear Regression, eXtreme gradient boosting (XGBoost), Last Value and Moving Average, the results show that the LSTM network model does not perform better than other models when applied to a short forecasting horizon.

Keywords: *Long Short-Term Memory, Stock Market, forecasting, prediction.*

1.INTRODUCTION

With the continuous application and development of artificial intelligence technology and big data technology, along with the further improvement of the financial market and the strong demand of the financial service industry, stock market forecasting has attracted wide attention from the industry and academia[1]. Machine learning algorithms such as decision trees, genetic algorithms, support vector machines, logistic regression and deep learning network models have been applied to stock forecasting in research. In recent years, LSTM network models have become a hot research topic for researchers. The LSTM neural network is a type of realizable recurrent neural network model with selective memory and intra-temporal influence, which is very suitable for the stochastic non-stationary series of stock price time series. LSTM network is considered to be one of the most accurate forecasting techniques.

A large number of studies is currently active on the subject of LSTM neural network used in finance. Some studies used a various set of parameters with a different number of epochs to measure the RMSE of the LSTM neural network model so as to improve the accuracy of the model [2], others used ML algorithm based on LSTM RNN to do predictions and tried to find the best sets for bout data length and the number of training epochs that better suit the assets and maximize the accuracy of the predictions[3]. Another paper compared the evaluation indicators with the prediction results to find the appropriate number of LSTM layers and hidden neurons[4]. And some studies explored whether LSTM neural networks can be applied to the price trend prediction of individual stocks[5]. Also, somestudies employed tree-based models and neural networks (ANN, RNN, and LSTM) to correctly forecast the values of four stock market groups as a regression problem and to see which models are more precise[6]. Another paper did a comparison of ARIMA, ANN and LSTM for stock price prediction[7]. Many existing literatures study the application of LSTM neural network models to stock market forecasting, but most of them focus on applying LSTM to forecast a longer forecast range and improving the accuracy of the predictions. The performance of LSTM neural network models in this regard does outperform other machine learning methods. In addition to improving the prediction accuracy of LSTM neural network models by optimizing parameter settings and processing data, some studies have proposed more complex models based on LSTM models to improve the prediction accuracy. Some studies processed stock data through a wavelet transform and used an attention-based LSTM neural network to predict the stock[8]. Others proposed a multi-value associated network model of LSTM-based deep-recurrent neural network (Associated Net) to predict multiple prices of stock simultaneously[9].

Copyright © 2022 The Authors. Published by Atlantis Press International B.V.
This is an open access article distributed under the CC BY-NC 4.0 license -http://creativecommons.org/licenses/by-nc/4.0/.

However, there is a lack of research applying LSTM neural network models to predict the stock market in a short-term range. In this paper, we apply five models, LSTM, Linear Regression, XGboost, Moving Average, and Last Value, respectively, by setting the prediction scope and using the historical data of Vanguard Total Stock Market Index Fund ETF Shares (VTI) from 10/5/2018 to 10/4/2021, as the original data for short-term prediction. The forecasting performance of the models is evaluated using root mean square error (RMSE) and mean absolute percentage error (MAPE).

The main objective of this paper is to investigate the prediction accuracy of LSTM neural network models applied to short-term prediction ranges and to see whether the LSTM model shows certain advantages, compared to other machine learning algorithms.

2.LONG SHORT-TERM MEMORY (LSTM)

Long Short - Term Memory (LSTM) is a long and short-term memory network, which is a time-recursive neural network suitable for processing and predicting important events with relatively long intervals and delays in time series. The LSTM algorithm was first proposed by Sepp Hochreiter and Jurgen Schmidhuber in 1997 as a specific form of Recurrent neural network (RNN).

Long Short-Term Memory (LSTM) is one of many types of Recurrent Neural Network RNN, it's also capable of catching data from past stages and using it for future predictions [10]. In general, an Artificial Neural Network (ANN) consists of three layers: consists of three layers:

1) Input layer,

2) Hidden layers,

3) Output layer.

In LSTM, the early stages can be remembered by gating and joined along the memory line. The following Figure 1 depicts the composition of LSTM nodes.

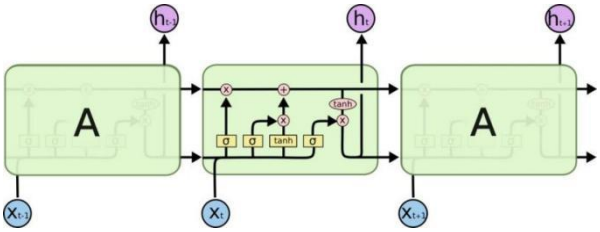

Figure 1. The internal structure of an LSTM [11].

The ability to memorize data sequences makes LSTMs a special kind of RNNs. Each LSTM node is mostly composed of a set of cells responsible for storing the passed data streams. The upstream connection model in each cell acts as a transport line to pass data from the past to the present, and the independence of the cells helps the model to dispose of filters that add values from one cell to another. Finally, the sigmoidal neural network layers consisting of gates drive the cells to an optimal value by disposing or letting the data pass. Each sigmoid layer has a binary value (0 or 1), where 0 means "doesn't let anything through"; and 1 means "lets everything through". The goal here is to control the state of each cell.

The gate is controlled in the following way:

- The Forget Gate outputs a number between 0 and 1, where 1 means "keep this completely"; and 0 means "ignore this completely".

- The Memory Gate selects which new data will be stored in the cell. First, a sigmoid layer "input gate layer" selects which values will be changed. Next, a tanh layer makes a vector of new candidate values that can be added to the state.

- The Output Gate determines what the output of each cell will be. The output value will be based on the state of the cell and the filtered freshest data[3].

3.METHODOLOGY AND DATA

Our system can be divided into 4 parts. These are raw data acquisition, date pre-processing, feature extraction and training of Neural Network.

3.1. Raw Data

The data used in this article is from Yahoo Finance and includes 7 features: Date, Open Price, High Price, Low Price, Close Price, Adjusted Closing Price and Volume. The data structure is shown in Table 1. For VIT, our data series cover the period going from 10/5/2018 to 10/4/2021.

3.2. Data Pre-processing

The pre-processing stage involves:

a) Data discretization: Part of data reduction but with particular importance, especially for numerical data

b) Data transformation: Normalization.

c) Data cleaning: Fill in missing values.

After the dataset is transformed into a clean dataset, the dataset is divided into training, testing and cross-validation set so as to evaluate. In order to build our model, we are going to use the LSTM RNN. Our model uses 60% of data for training, 20% of data for testing and 20% of data for cross-validation.

3.3. Feature Extraction

In this stage, only the features which are to be fed to the neural network are chosen. We will choose the feature from date, open, high, low, close, adj_close and volume.

In this paper, the feature adj_close is selected and its line graph is shown in Figure 2. It is clear that adjusted closing price rose considerably between 2018 and 2021. However, at the beginning of 2020, the entire economic market has been hit hard by the outbreak of the COVID-19 and there is a sudden decrease can be seen in this year, with the price of the stock dropped by roughly $60.

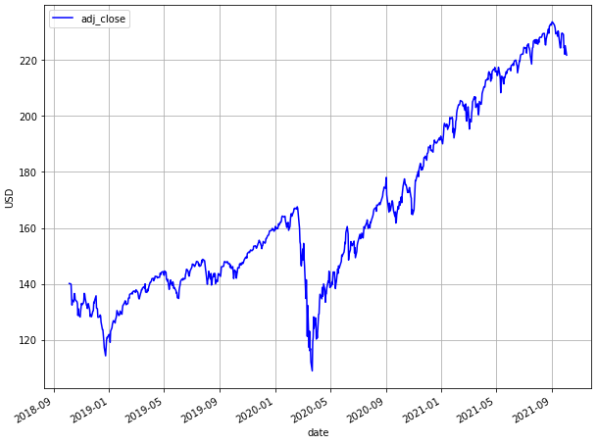

Figure 2. Adj_close

3.4. Training Neural Network

In this stage, the data is fed to the neural network and trained for prediction. For training, we use mean squared error to optimize our model and determine the optimal parameter settings:

Table 1.The structure of raw data

date	open	high	low	close	adj_close	volume
2018-10-05	148.619995	148.949997	146.729996	147.630005	140.132355	3197900
2018-10-08	147.279999	147.759995	146.360001	147.520004	140.027908	2515600
2018-10-09	147.250000	148.020004	147.020004	147.259995	139.781113	3690600
2018-10-10	146.929993	146.949997	142.410004	142.570007	135.329346	6890300
2018-10-11	142.020004	142.970001	138.750000	139.529999	132.443680	11707900
2018-10-12	142.000000	142.320007	139.490005	141.369995	134.190262	7981200
2018-10-15	141.160004	141.979996	140.500000	140.740005	133.592255	3386400
2018-10-16	141.750000	144.059998	141.360001	143.839996	136.534805	3990400
2018-10-17	143.910004	144.149994	142.330002	143.800003	136.496841	3148300
2018-10-18	143.309998	143.610001	141.000000	141.720001	134.522507	3783800

epochs = 50

batch_size = 8

lstm_units = 128

dropout = 0.2

The structure of our model is shown in the Table 2.

Table 2. the LSTM model summery

Layer(type)	Output Shape	Parameters
lstm_2(LSTM)	(None,9,50)	10400
dropout_2 (Dropout)	(None, 9, 50)	0
lstm_3 (LSTM)	(None, 50)	20200
dropout_3 (Dropout)	(None, 50)	0
dense_1 (Dense)	(None, 1)	51
Total params: 30,651 Trainable params: 30,651 Non-trainable params: 0		

4. RESULT AND DISCUSSION

In this experiment, four other methods were also used to make predictions in this paper separately and compared with the prediction results of applying the LSTM neural network model. The prediction results are shown in the following figures:

The result of using Last Value to predict is shown in the Figure 3. This is the most cost effective forecasting model. We can find that the forecast for each day (red cross) is simply the value of the previous day (green cross). Last Value is often used as a benchmark for comparing more complex models.

The result of using linear regression to predict is shown in the Figure 4. It can be found that this method does not capture changes in stock prices up or down well.

The prediction results of using Moving Average, XGBoost and LSTM can be shown in Figure 5, Figure 6 and Figure 7 respectively. Using the method of XGBoost is an iterative process of transforming weak learners into strong learners. In the moving average method, the predicted value will be the average of the first N values. As can be seen from the figures, there is not much difference between the prediction results of these three methods, but the prediction results obtained by using the LSTM and the XGBoost respectively are slightly better than using the Moving Average, with relatively small errors between the test and predicted values.

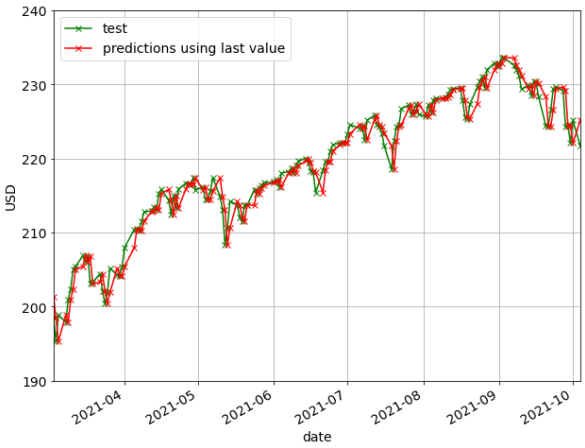

Figure 3. Forecast results of Last Value

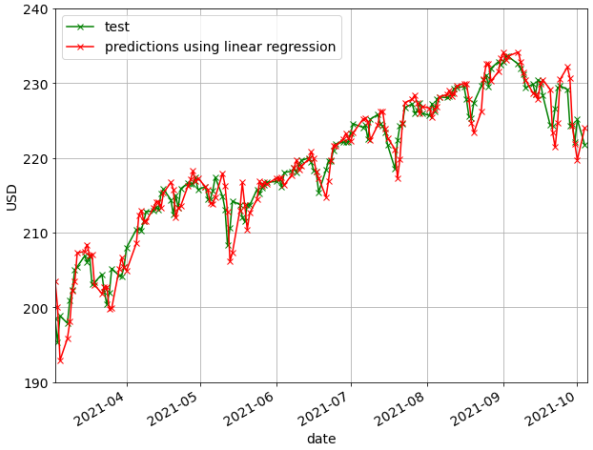

Figure 4. Forecast results of Linear Regression

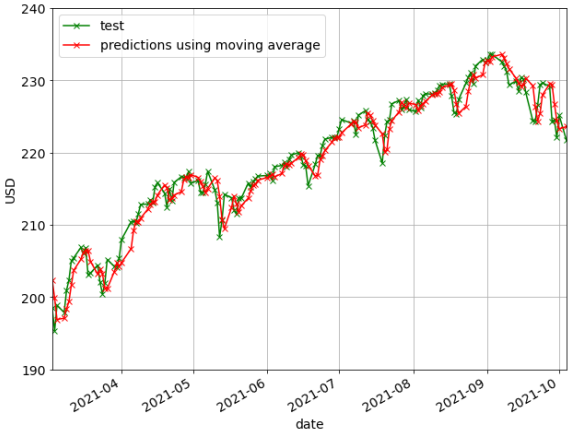

Figure 5. Forecast results of Moving Average

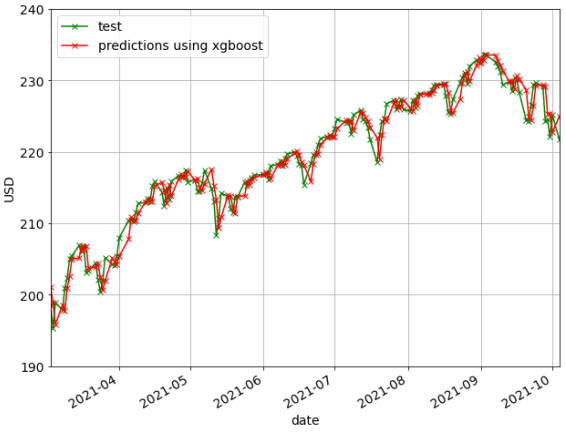

Figure 6. Forecast results of XGBoost

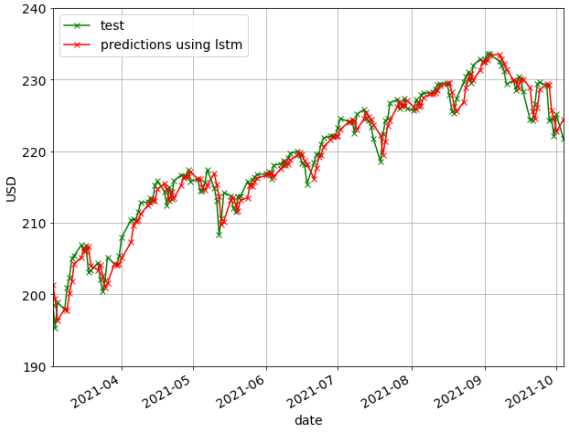

Figure 7. Forecast results of LSTM

The prediction results of the 5 models are integrated into the same graph for comparison, as shown in Figure 8, except for the poor prediction results of linear regression, the predictions of the other 4 are difficult to judge by direct observation.

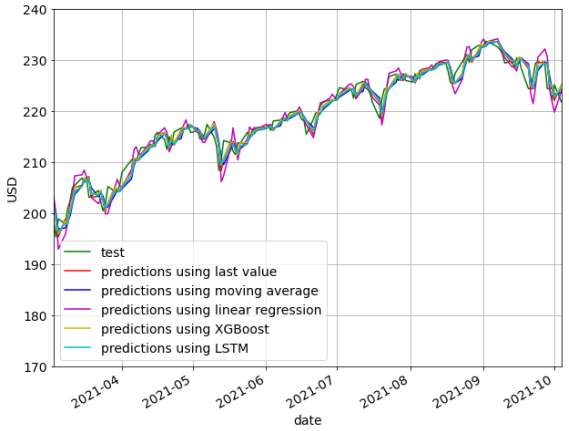

Figure 8. Forecast results of 5 methods

The article uses root mean square error (RMSE) and mean absolute percentage error (MAPE) to evaluate the prediction results.

The Root Mean Square Error (RMSE) is very sensitive to the very large or very small errors in a set of measurements, so the root-mean-square error can reflect the precision of the measurement very well. Equation (1) shows its formula, where y_i is the neural network output and y is the true value.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y)^2}{n}} \tag{1}$$

Mean Absolute Percentage Error (MAPE) is often employed to assess the performance of the prediction methods. MAPE is also a measure of prediction accuracy for forecasting methods in the machine learning area, it commonly presents accuracy as a percentage[6]. Equation (2) shows its formula, where y_i is the output value of the neural network and y is the true value.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y - y_i}{y}\right| \times 100 \tag{2}$$

The RMSE and MAPE values of the five methods are shown in Table 3.

Table 3. RMSE of five methods

Method	RMSE	MAPE(%)
Last Value	1.689	0.598
Moving Average	1.888	0.689
Linear Regression	2.270	0.780
XGBoost	1.647	0.593
LSTM	1.750	0.633

Comparing the RMSE and MAPE of the five methods, it can be seen from Table 3 that the two methods with the best prediction results are XGBoost and Last Value, which provide lower RMSE and MAPE. Last Value achieves better prediction results than the other more complex methods, especially outperforming the LSTM neural network model. This conclusion may be due to the fact that the prediction range is only 1. If the prediction range is longer, the other methods may be more effective than the Last Value. Additionally, there is the issue of data size. The LSTM is a neural network, and like any neural network requires a large amount of data to be trained on properly. The best predictions are obtained using XGBoost. The XGBoost performs second-order Taylor expansions on the loss function. This method introduces the second-order derivative both to increase accuracy and to allow customisation of the loss function. This largely avoids the problem of time lag, a drawback that is common in LSTM.

4. CONCLUSION

Trading in the stock market is growing rapidly and investors, analysts are eager to find a method and technique to effectively predict future stock market trends. In recent years, many studies have shown that LSTM neural network models are effective in predicting the stock market, and compared with other machine learning algorithms, LSTM neural network models perform very well when applied to longer prediction horizons.

In this paper, we compare the results of the LSTM neural network model in making short-term forecast range predictions with the results of the other four algorithms and find that the LSTM neural network model is not a perfect prediction method. However, it has to be admitted that the LSTM neural network model is better at capturing trends and seasonality in long-term forecast range prediction. This will encourage more researchers to use new techniques to find new forecasting methods that can be applied to more situations, thus helping investors, analysts or anyone interested in investing in the stock market by providing

them with a good knowledge of the future of the stock market.

REFERENCES

[1] Agrawal, J.G., Chourasia, V. and Mittra, A., 2013. State-of-the-art in stock prediction techniques. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(4), pp.1360-1366.

[2] Roondiwala, M., Patel, H. and Varma, S., 2017. Predicting stock prices using LSTM. International Journal of Science and Research (IJSR), 6(4), pp.1754-1756.

[3] Moghar, A. and Hamiche, M., 2020. Stock market prediction using LSTM recurrent neural network. Procedia Computer Science, 170, pp.1168-1173.

[4] PENG Yan, LIU Yuhong, ZHANG Rongfen, 2019. Modeling and analysis of stock price forecast based on LSTM. Computer Engineering and Applications, 55（11）, pp.209-212.

[5] Deng, Fengxin and Wang, Hongliang, 2018. Application of LSTM neural networks in stock price trend prediction - a study based on individual stock data from the US and Hong Kong stock markets. Collection, 14.

[6] Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A. and Salwana, E., 2020. Deep learning for stock market prediction. Entropy, 22(8), p.840.

[7] Ma, Q., 2020. Comparison of ARIMA, ANN and LSTM for stock price prediction. In E3S Web of Conferences (Vol. 218). EDP Sciences.

[8] Qiu, J., Wang, B. and Zhou, C., 2020. Forecasting stock prices with long-short term memory neural network based on attention mechanism. PloS one, 15(1), p.e0227222.

[9] Ding, G. and Qin, L., 2020. Study on the prediction of stock price based on the associated network model of LSTM. International Journal of Machine Learning and Cybernetics, 11(6), pp.1307-1317.

[10] Patterson, J. and Gibson, A., 2017. Deep learning: A practitioner's approach. " O'Reilly Media, Inc.".

[11] Olah, C. (2015). Understanding lstm networks–colah's blog. Colah. github. io.