Long Short-term Memory Applied on Amazon's Stock Prediction

Chenze Zhou[1, *]

[1] Ningbo University, Ningbo, China

* Corresponding Author Email: chenze.zhou@smu.ca

Abstract. More and more investors are paying attention to how to use data mining technology into stock investing decisions as a result of the introduction of big data and the quick expansion of financial markets. Machine learning can automatically apply complex mathematical calculations to big data repeatedly and faster. The machine model can analyze all the factors and indicators affecting stock price and achieve high efficiency. Based on the Amazon stock price published on Kaggle, this paper adopts the Long Short-term Memory (LSTM) method for model training. The Keras package in the Python program is used to normalize the data. The Sequence model in Keras establishes a two-layer LSTM network and a three-layer LSTM network to compare and analyze the fitting effect of the model on stock prices. By calculating RMSE and RMPE, the study found that the stock price prediction accuracy of two-layer LSTM is similar to that of three-layer LSTM. In terms of F-measure and Accuracy, the LSTM model of the three-layer network is significantly better than the LSTM model of the two-layer network layer. In general, the LSTM model can accurately predict stock price. Therefore, investors will know the upward or downward trend of stock prices in advance according to the prediction results of the model to make corresponding decisions.

Keywords: LSTM, stock price prediction, machine learning.

1. Introduction

Today's stock market has become increasingly complete, and many investment institutions and retail investors are also expected to buy stocks with the expectation of rising and gaining income. Stock investment is a high-risk and high-return industry, so investors must evaluate each stock through various means. The extensive data nature of the financial market determines the feasibility of introducing artificial intelligence technology to predict stock prices. The proper valuation and prediction of stocks can drive profits for investors, but it is also a difficult task. Among many machines learning models, LSTM is one of the valuable tools for stock price analysis and has been studied by many scholars. Bathla recently utilized LSTM to anticipate large fluctuations in stock prices.[1]. To identify and classify the emotion of the text, Ko Ching-ru et al. used the natural language processing tool BERT and combined it with LSTM for stock price analysis [2]. Cipiloglu et al. used LSTM to predict the returns of different investment portfolios and found that the construction of investment portfolios using predicted returns could significantly improve the performance of investment portfolios [3].

Inspired by prior researchers' academic study findings, this paper employs the LSTM model to examine Amazon's stock price trend. This experiment is completed in the language environment of Python3.10.6. To finish the data normalization, the sklearn package is utilized, and the Keras package is used to create the LSTM model. The evaluation metrics RMSE, RMPE, Accuracy, and F-measure are used to assess the performance.

This study begins by reviewing prior research findings. It includes the development history of RNN and LSTM and some applications of LSTM in stock prediction. Then, the experimental process of this paper is introduced. Data processing, model building, parameter adjustment, and evaluation index are involved in the experiment. Finally, it shows the research results of the experiment, that is, the comparison of the prediction performance of two-layer LSTM and three-layer LSTM on stock prices.

2. Related work

Neural networks are artificial intelligence technologies that educate computers to analyse data in ways similar to how the human brain processes data. It employs a hierarchical system of linked nodes, or neurons, analogous to the human brain. It is capable of developing adaptive systems that allow computers to learn from their mistakes and continue to improve. As a result, artificial neural networks can attempt to tackle complicated issues. The Hopfield network, the first-generation model of recurrent neural networks, was introduced by Saratha Sathasivam in 1982 [4]. Recurrent neural networks are capable of approximating nearly every nonlinear dynamic system with a high degree of accuracy when sufficient layers, units, and activation functions are used. They are very effective in extracting the temporal properties of sequence data. Recurrent neural networks are helpful tools used to predict data. They are also crucial in several fields, such as speech recognition and machine translation. Unlike a standard neural network, the hidden layer has linked nodes that may retain and use prior critical information to influence the output of subsequent nodes.

In the 1990s, academics began to use neural networks to make stock prediction. White applied neural network for the first time to the process the prediction of stock market time series and used the daily stock return of IBM company as the object of empirical study [5]. Kolarik and Rudorfer compared the ANN model with the ARIMA model and found that the ANN model had a better prediction effect [6]. Unlike many other prediction approaches, ANN does not limit the input variables (for example, how they are distributed). ANN can also better mimic homoscedasticity, that is, data with high volatility and unstable variances, since it can discover hidden relationships in the data without imposing any predefined associations.

However, traditional recurrent neural networks have the weakness of "gradient disappearance" and "gradient explosion." That is the inability to effectively exploit the relationships between distant inputs. Since the gradient of the front layer is slightly more significant due to training, the angle of the back layer increases exponentially. To solve the above two problems, Sepp Hochreit Jurgen Schmidhuber proposed the LSTM structure for the first time in 1997 [7].

Many scholars have used LSTM to estimate stock prices. Lu, Wenjie et al. suggested a CNN-LSTM-based prediction system. Individual prediction models such as CNN, RNN, are utilized to generate stock predictions. According to the experimental results, CNN-LSTM makes the smallest error of stock prediction [8]. Budiharto and Widodo proposed a data science model based on R language and statistical calculation of LSTM for stock price prediction in the Indonesia exchange, with an accuracy of 94.57% [9]. In most cases, the time series of stock prices are very complex. But this problem can be solved by breaking it down into subsequences. Yujun, Yang et al. used the LSTM and EMD method to train and predict each sub-series, and the experimental findings suggest that this strategy works well [10].

3. Methodology

The LSTM algorithm has been widely used in sequence prediction research and has proven to be quite successful. LSTM may create useful information while ignoring irrelevant data. An input gate, a forgetting gate, and an output gate comprise the LSTM. The input gate saves data to the cell that the model requires. When the model is running, the forgetting gate removes input that will not increase prediction accuracy. Information can be selected as an output via an output gate. LSTM provides a lot of benefits. First off, the input variable is unlimited. Second, to increase the degree of sample fitting, the complicated nonlinear connection between input variables is successfully fitted. In addition, overfitting is successfully avoided thanks to the notion of cyclic usage of neuron weights, which also greatly reduces the amount of neuron weights. Finally, the gradient explosion and gradient disappearance in ANN may be significantly resolved by the TANH activation function in DNN [7]. Based on the benefits of LSTM in data prediction, this research forecasts stock prices using the LSTM model. This experiment is done by downloading data, normalizing data, training a model, adjusting parameters, and forecasting stock.

3.1. Data download

The information in this article is derived from Amazon's stock price statistics on Kaggle. The time span is from May 1, 1997 to October 27, 2021, with a total of 6,155 data points. The experiment trains the model using the date and the matching opening price. The first 6000 data points serve as the training set, while the last 155 serve as the test set.

3.2. Data normalization

TANH function inside the LSTM makes the output range between -1 and 1, so the predicted value needs to be normalized to 0 and 1. It is vital to notice that even after the model has been completed, the values obtained remain between 0 and 1. As a result, we must unravel the standardization and compute the final evaluation index. This experiment uses Scikit-Learn's MinMaxScaler function to normalize the dataset to 0 and 1, base on Formula 1:

$$X\mathrm{i}' = \frac{Xi - X\min}{X\max - X\min} \tag{1}$$

Xi is the raw data that needs to be standardized. $X\max$ and $X\min$ represent the highest and lowest values in the raw data sequence requiring standardization processing. $X\mathrm{i}'$ is the target data obtained after standardization, and its value range is [0,1].

3.3. Model training

There are two kinds of deep learning models in Keras: the Sequential and the Model. The difference is in the topology. Models can be used to design very complex neural networks with arbitrary topologies, such as directed acyclic networks and shared layer networks. The Model can flexibly construct the network structure and set the relationship of each level. The Sequential of Keras package in Python is used for this paper’s model training. A Sequential is a stack of Sequential network layers. It has only one input and output, with only the adjacent relationships between the layers.

3.4. Parameter adjustment

Model parameters have a significant influence on the LSTM algorithm. For example, the change in the number of neuron layers and neuron units will exponentially increase the complexity of the model operation and affect the final prediction accuracy. The thousandth change in learning rate will significantly affect the efficiency and accuracy of the model in gradient descent. Therefore, the different LSTM model hyperparameters were set in this prediction experiment to compare the results. In this experiment, the time step=90, Batch size=64, and Epoch=10 were set unchanged, and the experimental results were observed by adjusting the number of neuron layers and units. Batch size refers to the number of training samples. Because machine learning techniques need a large volume of data, it is unrealistic to read the information entirely into the computer memory simultaneously. To solve this problem, the model need to divide the data into small pieces, read them into memory, and update the weight of neural network parameters after each amount of data is calculated. An epoch is the process of training all training samples at the same time. The model can better suit the actual scenario after numerous epochs. When a period's data is too enormous for a computer to process, it must be split into numerous batches for training. It should be noted that more epochs are not necessarily better. However, as the numbers increased, the models got more predictive. However, there is a risk of overfitting the model.

3.5. Stock prediction

The experimental outcomes are evaluated in two parts: prediction error and classification accuracy. To evaluate the model's stock prediction ability and compare experimental outcomes, the RMSE and MAPE are employed. The experiment can forecast how much the model's projected stock price differs from the actual stock price using these two indicators. The RMPE is a relative value that does not take into account the absolute value of the stock price. Formulas 2 and 3 are used to determine RMSE and MAPE:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(X_{prediction,t} - X_{real,t})^2} \tag{2}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{X_{prediction,t} - X_{real,t}}{X_{real}}\right| \tag{3}$$

In the real world, investors compare current and expected future stock prices. Investors buy stocks if they think they will go up in the future. Investors may want to sell if the stock price is expected to fall. Considering that accurate prediction of stock prices is a high requirement for the model. The accuracy of the model's categorization effect is the experiment's second metric. Investors can purchase or sell a stock if the model correctly forecasts whether the share price will climb or decline the next day. This evaluation approach is incapable of precisely predicting the size of stock price movements in order to optimize the protection of investors' profits. However, this strategy allows investors to assume less risk while ensuring that their stock returns are greater than zero. In addition, Accuracy and F-measure of model classification are used in the experiment. Precision and Recall should be calculated first to calculate these two evaluation statistics. Among all samples, the number of accurately classified samples corresponds to accuracy. The percentage of days on which the model accurately forecasts an increase or decrease in the share price is referred to as accuracy in this example. The Recall ratio is the percentage of positive class samples correctly categorized by the model divided by the total positive class samples. The precision of the model is the percentage of positive samples identified. In principle, the higher the Precision and Recall, the better, but they might be contradictory, and it is sometimes difficult to ensure that both are high. In this case, a new index F-measure is created to fully account for Precision and Recall. The F-measure is the Precision and Recall weighted harmonic average. It can comprehensively evaluate the model's fit impact. Formulas 4 to 9 are used to compute accuracy, recall, precision, and F-measure.:

$$Accuracy = \frac{TP + TN}{FP + FN + FP + TN} \tag{4}$$

$$Precision_{\text{positive}} = \frac{TP}{TP + FP} \tag{5}$$

$$Recall_{\text{positive}} = \frac{TP}{TP + FN} \tag{6}$$

$$Precision_{negative} = \frac{TN}{TN + FN} \tag{7}$$

$$Recall_{\text{nagative}} = \frac{TN}{TN + FP} \tag{8}$$

$$F_{\beta} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

$Precision$ is the weighted average of $Precision_{negative}$ and $Precision_{positive}$. $Recall$ is the weighted average of $Recall_{negative}$ and $Recall_{positive}$.

4. Results

After adjusting parameters such as epoch, batch size, and time step, two comparison tests were conducted. In the first experiment, two LSTM network layers were set, with one unit of neurons in each layer. In the second experiment, three LSTM network layers were set, and one unit of neurons was set in each layer. According to the results of stock prediction using the two-layer and three-layer LSTM model in Table 1, when the neural layer of the LSTM model is set as three layers, the RMSE result is smaller, which is 4.728774353. When the number of LSTM layers is 2, the RMSE result is 5.05089423 and the overall the gap is smaller. However, according to the evaluation of the two indicators of Accuracy and F-measure, the performance of the three-layer LSTM structure model semms to be significantly more accurate than that of the two-layer LSTM network. The Accuracy and F-measure of the three-layer LSTM structure model are approximately 5.8% and 6.6% higher than those of the two-layer network layer model, respectively.

Table 1. Comparison of experimental results error

test number	LSTM Layer	Units	RMSE	RMPE	Accuracy	F measure
1	2	1	5.05089423	0.014056661	0.480519481	0.473647131
2	3	1	4.728774353	0.012914561	0.538961039	0.539366578

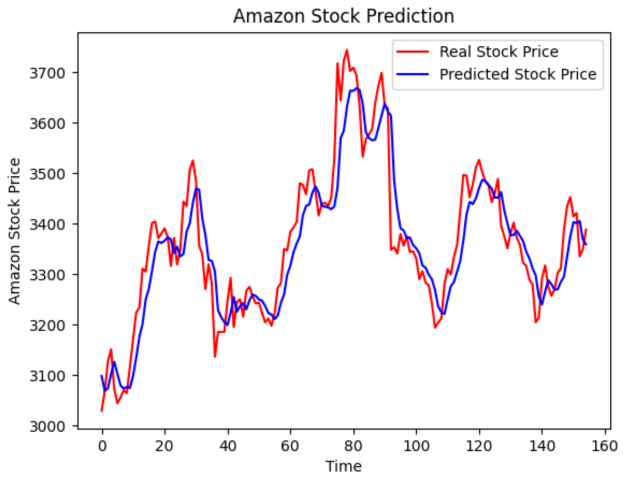

Figure 1. Amazon Stock Prediction when LSTM Layer = 2

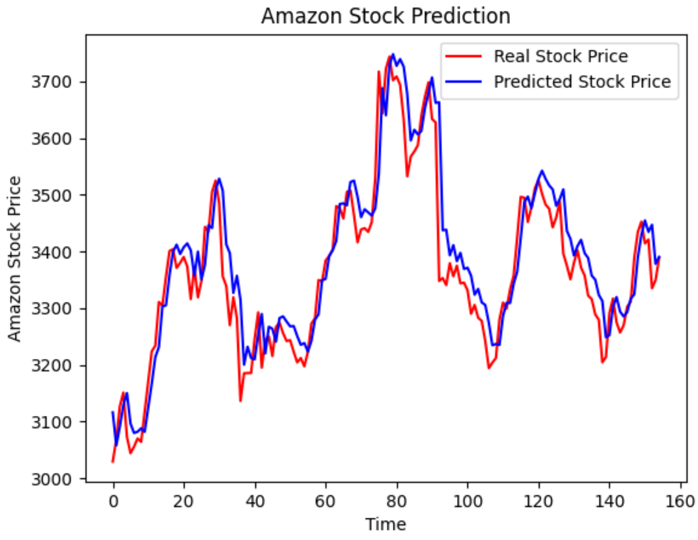

Figure 2. Amazon Stock Prediction when LSTM Layer = 3

5. Conclusions

Predicting stock prices is essential in both academia and business. Using machine learning to predict stock prices has recently been a hot topic. The computer can replace the human brain for many efficient data processing and analysis. Among many machines learning models, this paper uses the LSTM model to make the stock prediction of Amazon. The model established in this experiment can accurately predict the stock's opening price on the second day. The model adopts the LSTM algothrim to train and predict the stock data of Amazon for 6155 trading days from May 15, 1997, to October 27, 2021. After normalizing the stock data, the fitting effects of two-layer LSTM and three-layer LSTM are compared in the LSTM network model. The prediction of the three-layer LSTM model is slightly better than those of the two-layer LSTM model in terms of RMSE and RMPE. However, the prediction performance of the three-layer model is much superior than that of the two-layer LSRM in Accuracy and F-measure. Although there are some errors in specific stock prices, according to the comparison of the real and the predicted stock price, the model can reasonably predict the rising or falling trend of stock prices. This prediction is beneficial for investors to buy and sell operations. Investors can use the model's prediction of the trend of the stock price to decide to buy or sell the stock. In the future, macro-economic indicators, financial news and comments, market indexes, and other features can be further considered to optimize the model to enhance the overall forecast accuracy performance further. The LSTM model can be combined with ARIMA, attention mechanism and other technologies to improve the accuracy of forecasting ability.

References

[1] Bathla, Rani, R., & Aggarwal, H. (2022). Stocks of year 2020: prediction of high variations in stock prices using LSTM. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-022-12390-5

[2] Ko, & Chang, H.-T. (2021). LSTM-based sentiment analysis for stock price forecast. PeerJ. Computer Science, 7, e408–e408. https://doi.org/10.7717/peerj-cs.408

[3] Cipiloglu Yildiz, & Yildiz, S. B. (2022). A portfolio construction framework using LSTM-based stock markets forecasting. International Journal of Finance and Economics, 27(2), 2356–2366. https://doi.org/10.1002/ijfe.2277

[4] Sathasivam, S., & Abdullah, W. A. T. W. (2008). Logic learning in Hopfield networks. arXiv preprint arXiv:0804.4075.

[5] White, "Economic prediction using neural networks: the case of IBM daily stock returns," IEEE 1988 International Conference on Neural Networks, 1988, pp. 451-458 vol.2, doi: 10.1109/ICNN.1988.23959.

[6] Kolarik, T., & Rudorfer, G. (1994). Time series forecasting using neural networks. ACM Sigapl Apl Quote Quad, 25(1), 86-94.

[7] Hochreiter, & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Lu, Li, J., Li, Y., Sun, A., & Wang, J. (2020). A CNN-LSTM-Based Model to Forecast Stock Prices. Complexity (New York, N.Y.), 2020, 1–10. https://doi.org/10.1155/2020/6622927

[9] Budiharto. (2021). Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). Journal of Big Data, 8(1), 47–47. https://doi.org/10.1186/s40537-021-00430-0

[10] Yujun, Yimei, Y., & Jianhua, X. (2020). A Hybrid Prediction Method for Stock Price Using LSTM and Ensemble EMD. Complexity (New York, N.Y.), 2020, 1–16. https://doi.org/10.1155/2020/6431712