Stock Predicting based on LSTM and ARIMA

Huizi Qian[1,*]

[1] *Department of Industrial Economics, University of Chinese Academy of Social Sciences, 102445, Beijing, China*
[*] *Corresponding author. Email: qianhuizish@163.com*

ABSTRACT

With the application of artificial intelligence algorithm in the financial field, it soon becomes an interesting issue and a research hotspot to predict stock price. In this paper, LSTM and ARIMA models are adopted to explore the attracting stock price prediction. Besides, forecasting accuracy is comprehensively compared by several statistic indicators, i.e., MSE, MAE and RMSE. Based on the historical closing price collected from the Yahoo Finance, the above models are constructed. The prediction results show that the LSTM algorithm has a smaller MSE, MAE and RMSE, than the alternative ARIMA. The results in this paper may be beneficial to investors in the capital market when forecasting the future prices.

Keywords: *Stock price, Predicting, LSTM, ARIMA.*

1. INTRODUCTION

Currently, big data becomes an interesting topic, investors all over the world are trying to use various big data-based methods to predict the fluctuation of stock price. The traditional model cannot adapt to the increasingly huge stock market. With the vigorous development of neural network, its powerful information processing ability has more and more applications in real stock price prediction.

Stock price prediction methods mainly include time series prediction method, technical index analysis method and artificial intelligence method. Time series analysis method predict the future data through the historical information of time series, mainly including moving average (MA), auto regressive conditional heteroscedasticity (ARCH), auto regressive moving average (ARMA) and auto regressive integrated moving average (ARIMA). MA model is the simplest prediction model. The principle is to take the average value of past variables as the predicted value of the next cycle and reduce the influence of past extreme values through averaging, so as to achieve the effect of smoothing noise. Moving average method can predict time series data simply and quickly, but its model parameters are often set subjectively, which is difficult to analyze and accurately predict the actual situation. Compared with MA model, ARMA model and arch model can better predict financial time series. In addition, academia is also improving ARMA model and arch model. Technical index analysis method is widely used in the field of quantitative investment. Its prediction of stock price is mainly qualitative rather than quantitative, that is, it mainly predicts the future change trend of stock price. Another method which is deep learning algorithm models mainly include back propagation neural network (BPNN), recurrent neural network (RNN), long-term and short-term memory network (LSTM), convolutional neural network (CNN). Different types of neural networks are suitable for analyzing different types of data, while the model suitable for time series data such as stock price is RNN. Although RNN can learn sequence relationship, it has the problem of long-term dependence. Scholars use LSTM to improve it. The gating mechanism in each unit can adjust the long-term memory in time according to current input and historical information.

This paper designs two predictive models, i.e., LSTM and ARIMA, and applies the models to Google price prediction. Due to the fact that many factors will affect the prediction results, this paper pays attention to optimize data selection and process, parameter adjustment and the whole framework. Besides, this paper exhibits the future predictions. In this paper, we make some in-depth investigations by calculating three commonly used indicators, i.e., MSE, MAE and RMSE to evaluate the forecast performance of the two selected model, the results show that LSET beats the ARIMA no matter which indicator is chosen.

The rest of this paper is organized as below. Section 2 depicts the data, Section 3 summarize the methods,

Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 485–490, 2022.
https://doi.org/10.2991/978-94-6463-036-7_72

Section 4 presents the empirical results and Section 5 is the conclusion.

2. LITERATURE REVIEW

Prediction has been a hot research topic in stock market for a long time. In 1988, White, H. first studied the neural network model of stock price prediction. His prediction model was based on stock IBM, and the learning and prediction results were good. Since then, large quantities of studies were done to apply neural network when predicting the future stock prices [1]. Based on the capital market data collected in Korean, a comparative study between the neural network and traditional time series prediction was implemented by Lee et al. The authors found that the model beats the commonly ANN model when trying to predict stock price [2]. Merh et al. combined the forward neural network model and ARIMA to predict the future prices of certain stock. The authors argued that ARIMA model outperformed the ANN model regarding the forecasting accuracy [3]. Khashei et al. found that the performance of neural network was not applicable to some real time sequences. Therefore, they extend the model to a new hybrid artificial neural network. Compared with the neural network model, the model provides better prediction for 3 independent actual data sets [4]. Based on the variant of moving average model in time series analysis, Hansun, S did a similar investigation. Jakarta Stock Exchange (JKSE) composite index data was also selected for investigations regarding futures predictions [5]. Sreelekshmy Selvin et al. compared 3 types of deep learning architectures, including CNN, RNN and LSTM sliding window model, to predict NSEI member stocks. They found that CNN architecture could identify changes in stock trends and was superior to other models[6].Liu, S. et al. used CNN to formulate large number of stock selection strategy, and then used LSTM to formulate quantitative timing strategy to improve profits[7].Kim, H. Y. et al. proposed a hybrid method, which combined GARCH and LSTM model, and the result had lower prediction error [8].Lin, Y. et al. used the mixed model of long and short memory (LSTM) and fully integrated empirical mode decomposition combined with adaptive noise to predict the stock index prices of Standard & Poor's 500 index(S&P500) and China Securities 300 index (CSI300). Ceemdan decomposed the original data to obtain multiple IMF and a residual [9]. Gao, Y. et al. designed a model to optimize stock forecasting. They created a series of technical indicators, including investor financial data and sentiment indicators, and used deep learning algorithms and principal component analysis to reduce the dimensions of many influencing factors of the stock price. In addition, they also compared the stock market prediction performance of LSTM and GRU under different parameters [10].

3. Algorithm

3.1 LSTM

LSTM (Long and short term memory) was widely used in the field deep learning. [11-13] The LSTM algorithm is an artificial recurrent neural network (RNN) structure. Compared with classic network, LSTM adds the feedback mechanism, which makes the model different.

Due the feedback mechanism, the LSTM is able to dispose the full sample as well as a certain data point. For instance, LSTM is suitable for speech recognition, non-segmented, connected handwriting recognition, as well as network traffic or anomaly detection in intrusion detection system.

The concise form of the forward passage equation of LSTM unit with forgetting gate is given below [11-14]:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$$

$$O_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$$

$$\widetilde{C_t} = \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$

$$C_t = f_t \circ c_{t-1} + i_t \circ \widetilde{C_t}$$

$$h_t = o_t \circ \sigma_h(c_t) \quad (1)$$

Input, output and forgetting gate composed the common LSTM unit. The device will record this value at any interval time, the information inside and outside the three gate control devices. In many applications, compared with sequence learning methods such as RNN and hidden Markov model, LSTM is relatively insensitive to interval length[15].

Hidden Markov model (X_n, Y_n) :

$$P(Y_n \in A | X_1 = x_1, \ldots, X_n = x_n) = P(Y_n \in A | X_n = x_n)$$

$$P(Y_{t_0} \in A | \{X_t \in B_t\}_{t \le t0}) = P(Y_{t_0} \in A | X_{t0} \in B_{t0}) \quad (2)$$

Most stock price prediction problems using LSTM seem to be established in this way, that is, a model uses only isolated stock prices as the only source of information for prediction. LSTM attempts to understand any form of structural and sequential dependence, only in the context of past prices related to themselves. However, theoretically, if the daily stock market prices follow random walk, they will be completely independent of each other. Obviously, it is completely futile to try to use LSTM to learn any type of structure.

Random walk formula:

$$E(S_n) = \sum_{j=1}^{n} E(Z_j) = 0$$

$$E(S_n^2) = \sum_{i=1}^{n} E(Z_i^2) + 2 \sum_{1 \le i < j \le n} E(Z_i Z_j) = n$$

$$\lim_{n\to\infty} \frac{E(|S_n|)}{\sqrt{n}} = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \quad (3)$$

3.2 ARIMA

Autoregression model (AR) is a linear and easy-understanding method in forecasting technologies, as the model assumes that future price can be fully reflected by the past information. However, empirical investigation may not always show excellent results. Thus, a ARIMA (auto regressive integrated moving average) [16] was introduced to forecasting field, which is a natural extension of the AR model. Generally, three parameters (P, D, q) form the ARIMA model. P represents for the number of lag length in the auto regressive terms; q is a parameter for the moving average terms while D is to ensure a stationary sequence. A standard formation of ARIMA is shown as follows.

$$(1 - \sum_{i=1}^{p} \varphi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \quad (4)$$

Simple moving average (SMA) equations is given bellow [17]:

$$SMA_k = \frac{p_{n-k+1} + p_{n-k+2} \cdots + p_n}{k}$$

$$= \frac{1}{k} \sum_{i=n-k+1}^{n} P_i \quad (5)$$

$$SMA_{k,next} = \frac{1}{k} \sum_{i=n-k+2}^{n+1} p_i$$

$$= \frac{1}{k}(p_{n-k+2} + p_{n-k+3} + \cdots + p_n + p_{n+1} + p_{n-k+1} - p_{n-k+1})$$

$$= \frac{1}{k}(p_{n-k+1} + p_{n-k+2} + \cdots + p_n) - \frac{p_{n-k+1}}{k} + \frac{p_{n+1}}{k}$$

$$= SMA_{k,prev} + \frac{1}{k}(p_{n+1} - p_{n-k+1}) \quad (6)$$

3.3 Evaluation algorithms

Statistically, forecast error refers to difference between actual value and the forecasted values. Commonly, the difference can be measured by several indicators, i.e., mean absolute error (MAE), Mean Square Error (MSE), root mean squared error (RMSE) [17-18]. In this paper, these indicators are also adopted and is summarized as follows.

Root mean square error (RMSE) is a common measure of the difference between the predicted value of the model or estimator, that is, between the sample value or population value and the observed value.

$$RMSE = \left(\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}\right)^{\frac{1}{2}} \quad (7)$$

Mean square error (MSE) is the average of the square of the measurement error, that is, the average square difference between the estimated value and the actual value.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n} \quad (8)$$

Mean absolute error (MAE) is a measure of the error between pairs of observations representing the same phenomenon.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} e_i}{n} \quad (9)$$

4. EXPERIMENT

4.1 Dataset

The data in this article is derived from Yahoo finance (https://finance.yahoo.com/). The Google stock price from April 30th, 2017, to April 30th, 2022, is selected from Yahoo finance. The data contains High, Low, Open, Close, Volume, company_name etc. As shown below, the difference between minimum and maximum values are comparatively high which may reflect a fact that the company shows a stable development tendency. The values in Table 1 refer to the dollars per unit in the stock market.

Table 1 Descriptive Statistics of the Selected Stock

Stock	Min	Max	Mean	Variance
GOOGLE	898.7000	3014.1799	1564.1706	415923.2058

4.2 Methodology

This paper study the data from the stock market, especially some technology stocks, use pandas to obtain stock information and visualize all aspects of the stock. Finally, this paper study several methods to analyze stock risk based on the previous performance history of the stock and predict the future stock price through LSTM method.

Since the historical data set available on the company's website contains only a few features, such as high and low stock prices, open and close, stock trading volume, etc., this is not enough. In order to obtain higher forecast price accuracy, new variables are created using existing variables.

Before the LSTM model experiment of stock price, 95% of the total data of each stock is used as the training set data to train the parameters of the model, while the rest is the test set data. The prediction results of the model on the test set verify the advantages and disadvantages of the model. Then, build the LSTM network architecture: add six LSTM layers and several dropout layers to prevent over fitting. 64 units are the dimensions of the output and parameter return_sequences represents

whether to return the last output in the output sequence or the complete sequence, parameter input_ shape as the shape of the training set. The drain layer specified as 0.2 will be discarded and the dense layer specified as one unit output will be added. Finally, the model is suitable for 100 times with a batch size of 32. This paper use a batch of short sub sequence randomly selected from the training data instead of training RNN on the complete observation sequence.

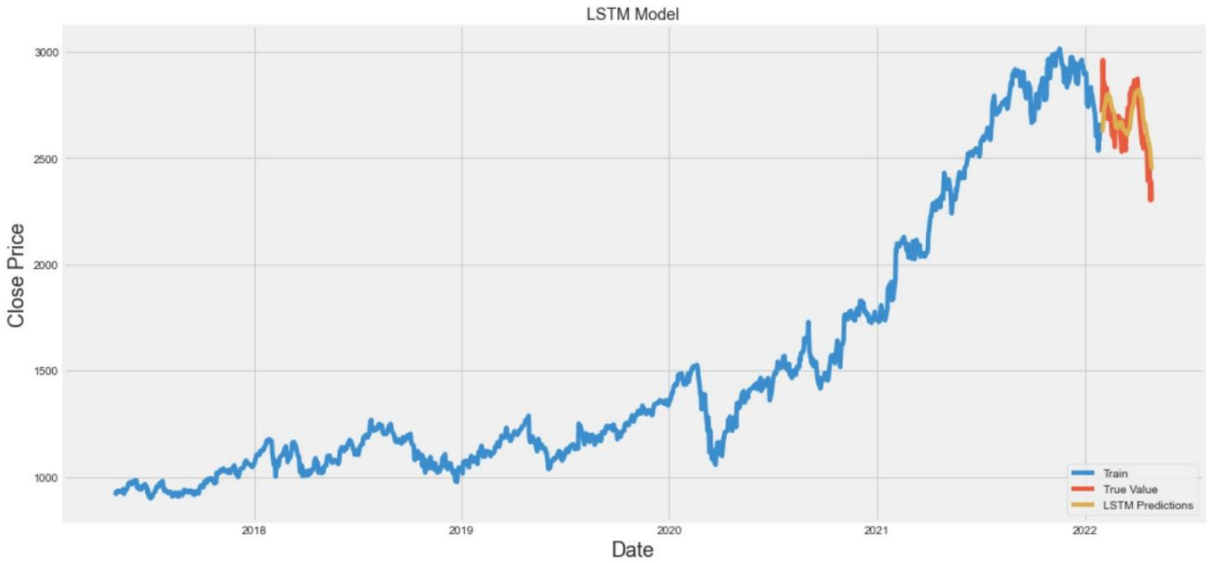

Figure 1 Google Stock Price Predicting by LSTM

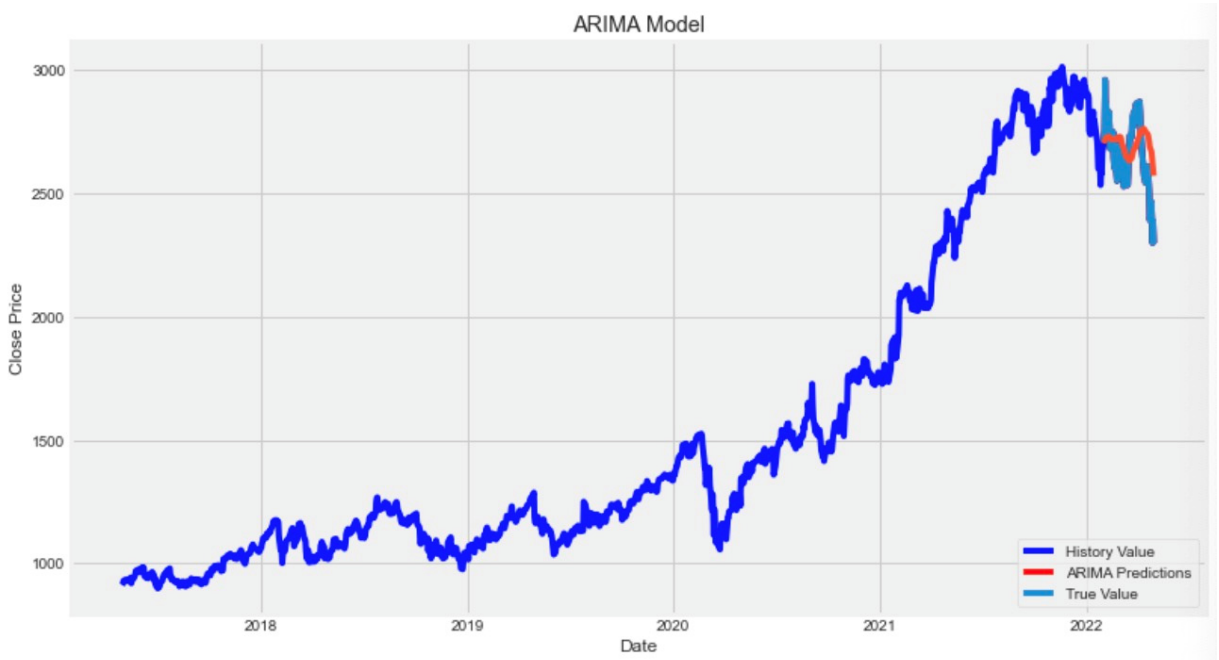

Figure 2 Google Stock Price Predicting by ARIMA

Moving average forecast is like finding the mean of the last N values. Use df[col]. rolling () creates a window of size N from the first entry of the column. This window is such that it returns t, t -1 ···, t-(N-1) rows for timestamp t (if N rows are not possible, it gives the maximum possible). That solves the problem of creating the window. Then predict the value at timestamp t. This means that the algorithm shifts the values by one row by append NaN at the start. Finally, this paper predicts and evaluates the above predictions using the metrics.

From figure 1 and figure 2, it can be seen that the predicting price trend is well fit the exact price, although there are some deviations. From Table 2, it can be concluded that LSTM prediction is better than ARIMA prediction.

Table 2. Evaluation results

Model Evaluation	LSTM	ARIMA
MAE	9.5921	154.5234
MSE	156.9601	38849.0341
RMSE	12.5283	197.1015

5. CONCLUSION

Currently, artificial intelligence algorithm is booming to apply in diverse aspects of social life. Undoubtedly, financial area is involved. Specifically, the algorithms are widely used to predict futures stock prices. In this paper, based on the artificial intelligence algorithm, LSTM and ARIMA methods are comprehensively compared. After comparing several statistical indicators of MAE, MSE and RMSE, it is found that LSTM beats the ARIMA method regarding the sample data used in the paper.

However, deficiencies exist. For example, when predicting stock price, numerous mathematical and statistical models are suitable, adopting alternative methods except for the LSTM and MA deserve in-depth investigation.

REFERENCES

[1] White, H. (1988, July). Economic prediction using neural networks: The case of IBM daily stock returns. In ICNN (Vol. 2, pp. 451-458).DOI: https://doi.org/10.1109/5.771073

[2] Lee, C. K., Sehwan, Y., & Jongdae, J. (2007). Neural network model versus SARIMA model in forecasting Korean stock price index (KOSPI). Issues in Information System, 8(2), 372-378.DOI:https://doi.org/10.48009/2_iis_2007_372-378

[3] Merh, N., Saxena, V. P., & Pardasani, K. R. (2010). A comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting. Business Intelligence Journal, 3(2), 23-43.This paper can be achieved at: https://www.researchgate.net/profile/Kamal-Pardasani/publication/45602108_A_comparison_between_Hybrid_Approaches_of_ANN_and_ARIMA_for_Indian_Stock_Trend_Forecasting/links/5417162d0cf2f48c74a3f030/A-comparison-between-Hybrid-Approaches-of-ANN-and-ARIMA-for-Indian-Stock-Trend-Forecasting.pdf

[4] Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. Expert Systems with applications, 37(1), 479-489.DOI:https://doi.org/10.1016/j.eswa.2009.05.044

[5] Hansun, S. (2013, November). A new approach of moving average method in time series analysis. In 2013 conference on new media studies (CoNMedia) (pp. 1-4). IEEE.DOI:https://doi.org/10.1109/conmedia.2013.6708545

[6] Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017, September). Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) (pp. 1643-1647).IEEE.DOI: https://doi.org/10.1109/ICACCI.2017.8126078

[7] Liu, S., Zhang, C., & Ma, J. (2017, November). CNN-LSTM neural network model for quantitative strategy analysis in stock markets. In international conference on neural information processing (pp. 198-206). Springer, Cham.DOI:https://doi.org/10.1007/978-3-319-70096-0_21

[8] Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems with Applications, 103, 25-37.DOI:https://doi.org/10.1016/j.eswa.2018.03.002

[9] Lin, Y., Yan, Y., Xu, J., Liao, Y., & Ma, F. (2021). Forecasting stock index price using the CEEMDAN-LSTM model. The North American Journal of Economics and Finance, 57, 101421.DOI:https://doi.org/10.1016/j.najef.2021.101421

[10] Gao, Y., Wang, R., & Zhou, E. (2021). Stock Prediction Based on Optimized LSTM and GRU Models. Scientific Programming, 2021.DOI:https://doi.org/10.1155/2021/4055281

[11] Hochreiter, S., & Schmidhuber, J. (1996). LSTM can solve hard long time lag problems. Advances in neural information processing systems, 9. This paper can be achieved at: https://proceedings.neurips.cc/paper/1996/file/a4d2f0d23dcc84ce983ff9157f8b7f88-Paper.pdf

[12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.This paper can be achieved at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.676.4320&rep=rep1&type=pdf

[13] Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.This paper can be achieved at: https://www.researchgate.net/profile/Y-

Bengio/publication/2839938_Gradient_Flow_in_Recurrent_Nets_the_Difficulty_of_Learning_Long-Term_Dependencies/links/546cd26e0cf2193b94c577c2/Gradient-Flow-in-Recurrent-Nets-the-Difficulty-of-Learning-Long-Term-Dependencies.pdf

[14] Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In Thirteenth annual conference of the international speech communication association.This paper can be achieved at: http://www-i6.informatik.rwth-aachen.de/publications/download/820/Sundermeyer-2012.pdf

[15] He, T., & Droppo, J. (2016, March). Exploiting LSTM structure in deep neural networks for speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5445-5449). IEEE.DOI:https://doi.org/10.1109/ICASSP.2016.7472718.

[16]Kalpakis, K., Gada, D., & Puttagunta, V. (2001, November). Distance measures for effective clustering of ARIMA time-series. In Proceedings 2001 IEEE international conference on data mining (pp. 273-280). IEEE.This paper can be achieved at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.5139&rep=rep1&type=pdf

[17] Chai, T., & Draxler, R. R. (2014). Arguments against avoiding RMSE in the literature. Geoscientific model.This paper can be achieved at: https://gmd.copernicus.org/articles/7/1247/2014/gmd-7-1247-2014.pdf

[18] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.DOI:https://doi.org/10.3354/cr030079