Proceedings of ICFTBA 2025 Symposium: Data-Driven Decision Making in Business and Economics
DOI: 10.54254/2754-1169/2025.BL28061

Stock Price Prediction Based on Machine Learning

Yuhao Jiang

School of Cyberspace Security, University of Science and Technology of China, Hefei, China
Corresponding author: jyh20050117@mail.ustc.edu.cn

Abstract. Predicting the stock price has been a hot topic among both investors and scholars. Researchers have conducted many studies to improve the prediction efficiency by using different models. This article aims to find a suitable model based on machine learning for the stock price prediction because some traditional statistical models do not perform well in this area according to past studies. After collecting the stock prices of Ping An Bank and China Merchants Bank during the period from 2021/7/1 to 2024/7/1 in the CSMAR and AKShare libraries, the study uses these data to train the Autoregressive Integrated Moving Average (ARIMA) model and Long Short-Term Memory (LSTM) model, obtaining the prediction results. Ultimately, by comparing the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to evaluate these two models, it could be concluded that the LSTM model outperforms the ARIMA model in predicting the stock price, which is because the machine learning model is capable of handling nonlinear relationships and long-term dependency.

Keywords: stock price, machine learning, ARIMA model, LSTM model

1. Introduction

The stock price prediction has always attracted widespread attention from investors and traders. Organizations and individuals can easily earn giant profits in the stock market if they can predict the stock price accurately. However, anticipating the stock price is a difficult problem for both financial experts and traders due to various and sophisticated factors that are related to stock market. In recent years, with the rapid development of Artificial Intelligence, more and more people have begun to utilize modern technologies (such as machine learning and deep learning) to assist in stock price prediction. Compared with traditional predicting methods, which mainly analyze past economic data and provide limited information, the methods to predict stock prices based on machine learning can improve the accuracy and efficiency [1]. There are many kinds of effective models based on machine learning, contributing to the stock price prediction, such as Long Short-Term Memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), and so on.

A number of professors have done some research on machine learning-based stock price prediction. Ni conducted a comparative analysis between the ARIMA model and the LSTM model based on gold stock price prediction. The result shows that the Root Mean Square Error (RMSE) of ARIMA is 2.62, and the Mean Absolute Error (MAE) is 2.33, while the two metrics of LSTM are 2.194 and 1.789, respectively. It demonstrated the Superiority of LSTM over ARIMA in complex

DOI: 10.54254/2754-1169/2025.BL28061

forecasting scenarios [2]. Siripurapu used an LSTM model to predict the stock price for the next 5 minutes based on historical data from the past 30 minutes. However, the result showed that the prediction accuracy is slightly lower than expected [3]. Zhang utilized the superior feature extraction capabilities of Convolutional Neural Network (CNN), combined it with Support Vector Machine (SVM) to predict stock price and some relevant indicators. After replacing the Sigmoid classifier with CNN, the study found a significant improvement in prediction accuracy; its RMSE decreased by 50% [4]. Additionally, to solve existing problems such as overfitting and low data quality, Zheng also proposed some coping strategies and methods in his article [5]. All in all, it can be concluded that machine learning models have unique advantages over past analysis methods. However, many scholars found that using some models alone achieves limited predictive accuracy. Meanwhile, it should be noted that deep learning methods such as LSTM still face many challenges, like balancing prediction accuracy and efficiency, processing multimodal data, and the influence of complex factors [5].

In this article, a typical machine learning-based model, LSTM, is used in predicting stock prices. Meanwhile, this article compares this model with the traditional statistical model Autoregressive Integrated Moving Average (ARIMA) to obtain a better model.

2. Method

2.1. Data sources and samples

For the sake of ensuring the availability, reliability, and standardization of the data, this study focuses on researching the daily adjusted close price deriving from Ping An Bank and China Merchants Bank in the China Stock Market & Accounting Research Database (CSMAR), during the period from 2021-07-01 to 2024-07-01 [6]. Ping An Bank and China Merchants Bank are two large-cap stocks, with active trading and high-quality data, making it suitable for model training. In addition, selecting large samples over the past three years ensures timeliness and avoids short-term randomness. This article utilizes daily adjusted closing prices to represent stock price.

2.2. Introduction of ARIMA model

2.2.1. Fundamental principle and structure

ARIMA is a model that is widely used in time series analysis and it plays an imperative role across many kinds of fields, especially in stock price prediction. The ARIMA model consists of three basic parts: Autoregressive Model (AR), Integration(I) and Moving Average Model (MA).

AR is a statistical model that is used for time series analysis and forecasting. It assumes the future values are determined by a linear combination of past values, which means a value only relates to a certain number of its previous values. An p-th order AR model, denoted as AR(p), can be expressed as below:

$$X_t = C + \sum_{i=1}^{p} \Phi_i X_{t-i} + \varepsilon_t \quad (1)$$

X_t is the observed value at time point t. ε_t is the error term at time t, assumed to be independently and identically distributed with a mean of zero and constant variance σ2.

MA is also a statistic model that is used in time series analysis and forecasting. It assumes that the current observed value is a linear combination of past random error terms. A q-th order MA model.

DOI: 10.54254/2754-1169/2025.BL28061

Denoted as MA(q), can be expressed as below:

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-q} \tag{2}$$

μ is the mean of this time series, which is the same for all the time points. "I" denotes differencing, which is used to transform non-stationary time series into stationarity. Through first-order (I=1) or second-order (I=2) differencing, it can eliminate trends and seasonal factors from the time series.

Compared with the original series, the differenced series reduce trends and seasonality. Typically, applying first-order differencing shortens the sequence by 1 observation. Actually, the order of differencing can be any positive integer as long as it can obtain a stationary sequence.

2.2.2. Research procedure

This article collects the stock price from 6 different stocks during the past 4 years in CSMAR. It needs to be stated that ARIMA requires the time series to be stationary, in other words, its mean and variance should be constant. Therefore, before fitting the ARIMA model with these data, it is necessary to do stationarity testing for time series. This study adopts Augmented Dickey-Fuller (ADF) Test, which is used to test the unit root.

After testing, if the time series is non-stationary, it needs to do some additional data processing—differencing (I=1 or I=2), until the series becomes stationary.

Subsequently, based on the close price of stocks that are collected, this study uses databases from Python, utilizes stock prices to train the model, determines optimal parameters, and ultimately completes model construction.

2.3. Introduction to LSTM model

2.3.1. Introduction to neutral network

Neutral network is a simulation of human brain [7]. With the constant development of neutral network, its application expands across various fields, such as image recognition and medical diagnostics [7]. It has powerful simulation abilities with minimal error. Usually, it can be divided into three layers: Input layer, Hidden layer and Output layer. The data is fed through the Input layer and processed by the Hidden layer and ultimately, produces results that meet requirements with minimal error.

2.3.2. Fundamental principles of LSTM

The research found that LSTM, a variant of RNN models, demonstrates the best results for stock price prediction among all the neural network [8]. Hochreiter and Schmidhuber in 1997 proposed the LSTM network to address the data persistence requirements of RNN. This model introduced gate mechanisms to construct specialized memory cells, thereby solving gradient instability problem [9]. By introducing gating mechanisms to selectively remain or forget information, it is suitable for long-sequence model. A LSTM unit consists of several core elements: Forget Gate, Input Gate, Output Gate and Cell State. The mathematic principles are as followed.

Forget Gate:

$$f_t = \sigma\left(W_f \bullet [h_{t-1}, x_t] + b_f\right) \tag{3}$$

DOI: 10.54254/2754-1169/2025.BL28061

f_t is the output of forget gate(a vector with values between 0 and 1) σ is the sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$). h_{t-1} is the hidden state from the previous timestep b_f is the bias term of the forget gate

After calculating the forget gate values, feed the result data into model to update the new information.

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\widetilde{C_t} = tanh(W_C \bullet [h_{t-1}, x_t] + b_C) \tag{5}$$

t is the candidate new memory (using tanh activation, range [-1, 1]). tanh is the hyperbolic tangent (tanh) function which is used to compress the candidate memory into the interval [-1, 1]. Combine the results of input gate and forget gate, update the cell state C_t .

Finally, activate the output gate and generate the hidden state ht.

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \odot tanh(C_t) \tag{7}$$

In this model, tanh is used as the activation function to introduce nonlinear transformation, enabling the model to learn sophisticated patterns. Additionally, the output of forget gate ft determines how much historical information to discard from memory cell and the output of input gate it regulates the amount of new information written into the memory cell. Therefore, this article adopts LSTM model and it achieves more accurate predictions than other models.

2.3.3. Research procedure

The procedure for stock price prediction by using Model LSTM consists of five important steps: select relevant financial stock data from CSMAR, preprocess the data such as Normalization and ADF test, train the model with Python Programmatically, and fine-tuned the hyperparameters to obtain a suitable model, and finally predict the stock price. The specific steps are as Figure 1.

DOI: 10.54254/2754-1169/2025.BL28061

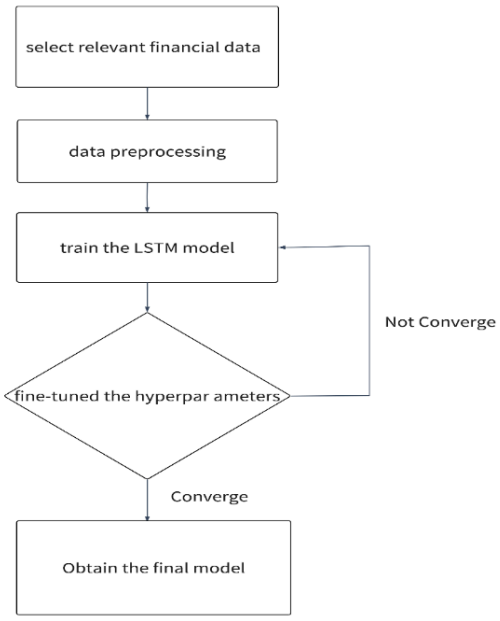

Figure 1. The procedure of training the model (picture credit: original)

3. Results and discussion

3.1. Prediction results of the ARIMA model

After fitting the ARIMA model using the price trend of two stocks during the period from 2021-07-01 to 2024-07-01, this article shows the model performance of 2 different stocks.

Figure 2 represents the fitting result of Ping An Bank and Figure 3 represents the fitting result of China Merchants Bank. Note: The x-axis range (0–720) corresponds to the three years between July 1, 2021, and July 1, 2024.

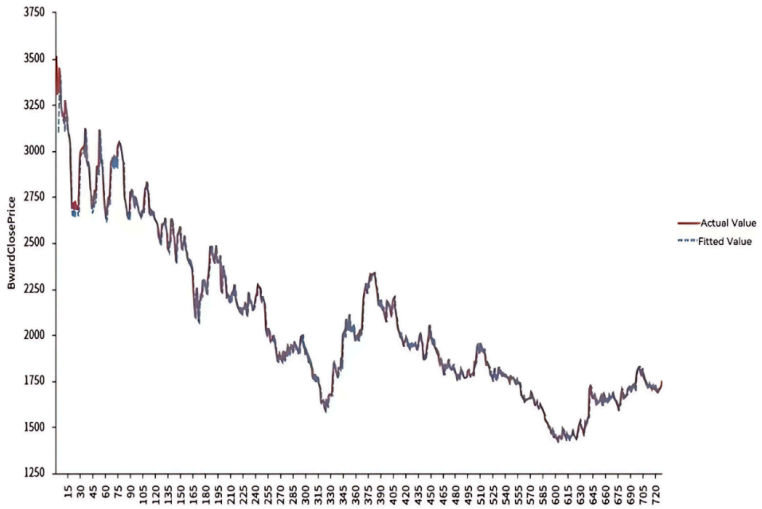

Figure 2. ARIMA stock price prediction of ping an bank (picture credit: original)

DOI: 10.54254/2754-1169/2025.BL28061

Through systematic parameter grid search with AIC as the evaluation metric, the research selects ARIMA (1,2,1) as the statistical model of Ping An Bank, which yields the lowest AIC among all the test combinations. After fitting the parameters, the research obtained the model equation as: $y(t) = 0.012 + 0.040 \times y(t-1) - 0.999\varepsilon(t-1)$. The p-value of the Q-statistic (Q6) exceeds 0.1, indicating that the model residuals exhibit white noise properties. Therefore, the model meets the requirement.

After calculating the error of actual value and predicted value, the MAE is 28.4563, and the RMSE is 48.2360.

Figure 3. ARIMA stock price prediction of china merchants bank (picture credit: original)

The fitting method is the same as above. This research selects ARIMA (2,1,2) as the statistical model. After fitting the parameters, this article obtained the model equation as: $y(t) = -0.101 + 0.562 \times y(t-1) - 0.992 \times y(t-2) - 0.583\varepsilon(t-1) + 0.990\varepsilon(t-2)$. The p-value of the Q-statistic (Q6) exceeds 0.1, indicating that the model residuals exhibit white noise properties. Therefore, the model meets the requirement.

After calculating the error of the actual value and the predicted value, the MAE is 2.7638, and the RMSE is 3.9074.

3.2. Prediction results of LSTM model

For the sake of handling nonlinear relationships and long-term dependency to improve the prediction accuracy, the research selects LSTM as another model, which is based on machine-learning method. This study utilizes the stock price of Ping An Bank and China Merchants Bank during 2021/7/1 to 2024/7/1 to train the LSTM model. For convenience, the research read stock price data from the AKShare library, which can be directly imported in Python. Figure 4 represents the fitting result of Ping An Bank and Figure 5 represents the fitting result of China Merchants Bank.

DOI: 10.54254/2754-1169/2025.BL28061

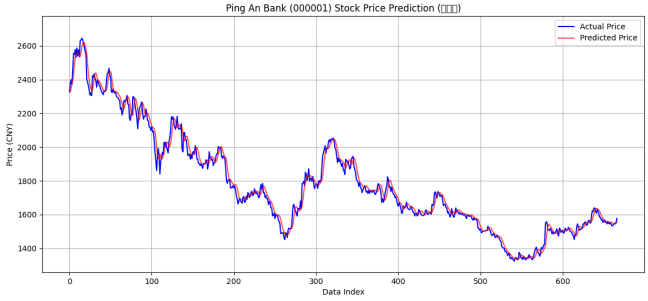

Figure 4. LSTM stock price prediction for ping an bank (picture credit: original)

After training an LSTM model using the stock price of Ping An Bank, the model demonstrates an MAE of 23.52 and an RMSE of 34.80.

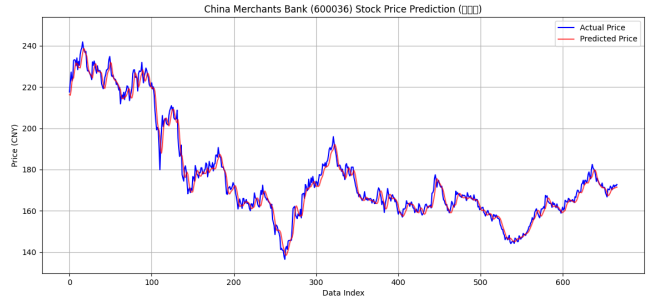

Figure 5. LSTM stock price prediction for china merchants bank (picture credit: original)

After training an LSTM model using the stock price of Merchants Bank, the model demonstrates an MAE of 2.09 and an RMSE of 2.91.

Through comparing, it can be found that both MAE and RMSE of the LSTM model are lower than those of the ARIMA model, which means the prediction results of the LSTM are more accurate than ARIMA.

3.3. Discussion of results

ARIMA is a traditional statistical model and it needs data to be stationary. LSTM is a nonlinear neural network model and it is suitable for dealing with nonlinear relationships. After fitting the data, the result indicates that LSTM performs well in obtaining long-term dependencies in time series. Therefore, the machine learning-based model consistently outperforms ARIMA in stock price prediction accuracy. The possible reasons are that the ARIMA model fails to capture complex nonlinear patterns in stock prices, but LSTM is capable of learning complex stock price dynamics and excels in making long-term forecasts based on machine learning.

However, using the LSTM model requires a large amount of data to prevent overfitting and its training speed is very slow. Therefore, when the available dataset is small, the ARIMA model is more suitable for predicting the stock price due to their quickness and simplicity compared to complex machine learning methods. What's more, ARIMA model demonstrates robust performance with small sample sizes.

While the prediction results are as to the true value, the LSTM model can still be strengthened. The experimental conditions in this research are limited, with better conditions, the model can be

DOI: 10.54254/2754-1169/2025.BL28061

adjusted to further reduce the error value and enhance its prediction accuracy, such as using some multi-model fusion methods and so on. Yuan proposed a model named “VMD (Variational Mode Decomposition)- PSO (Particle Swarm Optimization)- LSTM model” [10]. He took the stock data of Kweichow Moutai as an example and found that the MAE can decrease by 40% and RMSE decrease by 60% compared with using the LSTM model alone [10]. Wu et al. designed an Event-Aware LSTM Model [11]. This model did well in Handling aperiodic events and outperformed than normal LSTM model in stock price prediction [11]. All in all, in future researches, the LSTM model still requires further improvement.

4. Conclusion

The study utilizes a machine-learning-based model LSTM and a traditional statistic model ARIMA to anticipate future stock prices. The method used in this case is to feed processed data into the model to train it, and then obtain the suitable parameter and model. After the work, this article conducts a comparison between the actual value and predicted value to analyze models. The result demonstrates that LSTM is more accurate in predicting the movement of stock prices than ARIMA over a long period. The investors can use the prediction results obtained from the research to invest with more accuracy in the stock market, which improves investment efficiency significantly. All in all, machine learning method plays an imperative role in the field of stock price forecasting. It is worth noting that the model can still be improved by using some Multi-Model fusion methods, which can further enhance the stock price prediction accuracy in future studies.

In the future, with the rapid development of AI, machine learning will significantly affect the way that investors make financial decisions, helping them make more beneficial investments and attracting more people to get involved. However, although the machine learning model has shown reliable results in stock price forecasting, the stock market is very sophisticated and uncontrollable, which might lead to a sharp decline or an increase due to a single factor, such as a policy change or natural disasters. It is impossible to predict the stock price 100 % accurately. Therefore, investors can not rely solely on these models for investment decisions. There is still a lot of room for these models to develop in the future.

References

[1] Kong, T. (2023). Machine learning in finance: A case study on forecasting Google's stock price. In Proceedings of the 2nd International Conference on Financial Technology and Business Analysis (Part 4) (pp. 448–452). Ed. Smith School of Business, Queen's University.
[2] Ni, W., Tian, M., Jiao, Y., et al. (2025). Machine learning-based multi-model prediction of gold stock prices. Journal Name, 38(2), 169–172.
[3] Siripurapu, A. (2014). Convolutional networks for stock trading. Stanford University Department of Computer Science, 1(2), 1–6.
[4] Zhang, G. (2016). Application of improved convolutional neural networks in financial forecasting (Master's thesis). Zhengzhou University.
[5] Zheng, Z. (2023). A review of stock price prediction based on LSTM and TCN methods. In Proceedings of the 2nd International Conference on Financial Technology and Business Analysis (Part 1) (pp. 132–138). Ed. Stony Brook Institute, Anhui University.
[6] China Stock Market & Accounting Research Database. (2025). Data retrieved from https: //data.csmar.com (Data coverage: 2024-07-01 to 2025-07-30).
[7] Jiang, S. (2025). Stock price prediction based on LSTM model. Jiangsu Commercial Forum, (1), 83–86.
[8] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669.
[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

DOI: 10.54254/2754-1169/2025.BL28061

[10] Yuan, H., Song, Q., Zhou, Y., et al. (2025). Research on stock price prediction based on VMD-PSO-LSTM multi-scale hybrid model. Journal of Kashgar University, 46(3), 26–31.
[11] Wu, X., Zhao, G., Liu, H., et al. (2025). Stock trend prediction based on event-LSTM model. Computer Engineering and Applications. Advance online publication.