Time Series Forecasting Based on ARIMA and LSTM

Peiqi Liu[1,*]

[1] *Department of Business Management, United International College, Zhuhai, 519000, China*
[*] *Corresponding author.Email: p930006120@mail.uic.edu.cn*

ABSTRACT
With the increasing demand for designing a future strategy to minimize risk and make a benefit. The time series analysis becomes an essential tool in social science, engineering, and finance. Therefore, investors and researchers endeavor to investigate kinds of models to improve the accuracy of the forecasting result. Originally, Autoregression (AR) and Moving average (MA) model was developed to forecast next period data. Moreover, ARIMA was built to solve the non-stationarity of data. Besides, ARCH and GARCH were built to capture the volatility of data. Later, the neural network model in deep learning gets its popularity with higher accuracy for prediction. ANN model and LSTM model are widely used in time series analysis for the different research areas. By evaluating the performance between the traditional ARIMA model and burgeoning LSTM model on stock price prediction, the paper could guide investors to manage their assets with time series forecasting tools.

Keywords: *Time series forecasting. Stock price forecasting. Deep Learning. ARIMA. LSTM*

1. INTRODUCTION

1.1 Background

Forecasting the price movements of the stock plays a crucial role in the financial market. Investors who accurately predict the moving direction of the stock price can take optimal strategy in trading which makes great benefit. In contrast, investors who wrongly forecast the price will suffer from loss. Therefore, investors who look for maximizing their profit are eager to search for a method that can correctly predict stock price. As the financial market grows fast, the increasing availability of historical data gives usability to Time Series Forecasting (TSF), which is an important area of machine learning with a sequence of time components. The dataset used in TSF includes points preserved in time order, which is the sequence of discrete-time equally space in time, where the movement of stock prices will be forecasted by analyzing observed points in such series. TSF can be an effective tool for investors to analyze future stock prices and further to create economic value. Therefore, in the past several decades, a lot of models and techniques with time series have been developed for stock prices forecasting.

From a traditional perspective, the Box-Jenkins Method is a linear model extensively used which includes the auto-regression model (AR), the moving average model (MA), the auto-regressive moving average model (ARMA), and auto-regressive integrated moving average model (ARIMA). However, with the rapid and sustainable development of the economy and the increase of risk factors in the financial market, most of the sequences are nonlinear, and the traditional ARMA model of the financial market cannot describe the characteristics of the distribution sequence of risk factors more completely and accurately. Therefore, time-varying conditional variance with Autoregressive Conditional Heteroskedasticity (ARCH) model arises. The ARCH process is another model used to forecast the volatility of stock prices. Volatility represents a degree of fluctuation in stock prices. As for volatility, it is essential to detect the complicated behavior of conditional variance. As the result, the Generalized ARCH model (GARCH), a high order ARCH model is developed based on infinite ARCH specifications which reduce the number of estimated parameters from infinity to two.

From an innovative perspective, a variety of neural network models are widely applied in different areas. The artificial Neural Network (ANN) model gain its favor for its ability to learn patterns from the nonlinear, non-stationary, and high-noise time series data of stock prices. The development of the ANN model is inspired by the animal's brain which can process complex information by pattern learning [1]. As an artificial intelligence method, ANN has great ability in model stationarity, fault tolerance, and data processing [2]. Recurrent Neural Network (RNN), a model that is more complicated model than ANN, becomes popular since it makes information flow in different directions in its layers. Moreover, Long Short-Term Memory (LSTM), an improvement for Recurrent Neural Network (RNN), has performed well in time series analysis in recent years. The disadvantages of gradient disappearance and gradient explosion in the RNN model can be removed during the LSTM training process of long sequences since LSTM has the characteristic of expanding according to the sequence of time [3].

1.2 Related Research

ARIMA model was first developed by George Box and Gwilym in their textbook Time Series Analysis: Forecasting and Control to engage in Time Series Analysis (TSA) in 1970. They proposed to use ARMA when the sequence is stationary and use ARIMA when the sequence is non-stationary [4]. Many investors then apply the ARIMA model when analyzing their trading strategy and some researchers have found that ARIMA is effective in forecasting stock prices. Nau believes that the ARIMA model is a relatively sophisticated and accurate algorithm for time series forecasting [5]. Zumbo et al. researched that ARIMA is a good method for non-stationary time series prediction that is composed of an autoregressive and a moving average model and was successfully utilized for time series prediction in different areas which includes financial markets [6].

Bollerslev developed the GARCH model which is based on the ARCH model constructed by Robert Engle in 1982 [7]. Shengtantu applied the GARCH model in financial markets and found that it not only accurately depicts and describes whether the impact of positive and negative financial market shocks on stock prices is asymmetric, but also has obvious significance and role in the study of the symmetrical relationship between expected operating returns and expected economic risks of the financial market [8]. Assous et al. researched that the GARCH model can effectively solve the volatility problem of time series since it can accurately describe the basic characteristics of the "thick tail" of stock prices in financial markets [9].

However, some researchers argued that the accuracy of classic TSF models is not satisfactory. Yang and Wang questioned the accuracy of the ARIMA and GARCH models because financial data have high noise and dynamic characteristics. The flexible relationship between the dependent variable and independent variable limits the further application and expansion of traditional TSF. They proposed that LSTM is a better forecasting method for stock prices [10]. By comparing prediction results of stock prices under the ANN model and the ARIMA model, Milad and Seyed researched that the ANN model has higher accuracy than ARIMA Model [11]. Sima et al. compare the ARIMA model with the LSTM model, arguing that the LSTM model can achieve higher accuracy than the ARIMA model [12]. Besides, Lu et al. applied different models to forecast stock prices and the result shows that the CNN-LSTM model has the highest accuracy for the next day stock price [13].

As the reviewed papers above, it can be inferred that studies on forecasting stock prices are still being raised among researchers and it seems that newly proposed models are more advanced than the classic one. However, there still exists debate between different models.

1.3 Objective

This paper aims to investigate stock prices trend forecasting models by comparing the accuracy of prediction between classic models (ARIMA) and burgeoning models (LSTM). By applying these models to the stock prices of Walmart (WMT), which is one of the world's biggest retailers as well as the world's largest corporations, it can provide a clear understanding of time series analysis as well as make a benefit. Its stock price could be an effective indicator of the market situation. Therefore, it is essential to have an accurate prediction as it reflects great interest to investors and assets managers. Using a proper model can help increase the accuracy of the prediction result and thus ensure a safe investment.

2. METHODOLOGY

2.1 Datasets

To investigate the patterns of the stock prices, the two-year daily data of Walmart was collected from 2 January 2020 to 31 December 2021, which is publicly available on the website Yahoo! Finance [14]. The dataset is composed of one column of time sequence as well as the other five columns: open stock price, highest stock price, lowest stock price, close stock price, adjusted closed price, and trade volume during the day. The adjusted close stock price with time sequence data is kept since the stock value can be more accurately reflected after the adjustment and can be more precise for TSF. ARIMA model and Multi-Step LSTM model will be applied to these data to forecast future stock price trends. Determination of the model accuracy will be based on indicators of Autocorrelation Functions (ACF), Partial Autocorrelation Function (PACF), and Root Mean Score Error (RMSE).

2.2 ARIMA Model

ARIMA(p,d,q) is a model that can be used to predict future values if given past value with time sequence, which combined AR, MA, and Integrated. Therefore, p represents the order of the AR term, q represents the order of the MA term and d is the number of differences required to make the time series stationary. The reason for stationarity is that ARIMA can only be applied to stock price prediction if the time series is not white noise and not seasonal [15]. Since the future value of the AR model is only depends on its lags and that of the MA model only depends on the lagged forecast error, the future value of a variable in the ARIMA model is a linear combination that consists of past series value and random noise which expressed as follows:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q} \quad (1)$$

Where Y_i is the value of the variable at time t, α is a constant, β_i is the coefficient for autoregression, ϕ_i is the coefficient for moving average and ε_i is the random error at time t.

Therefore, the general steps to predict future stock price includes identifying stationarity of time series, if it is non-stationary, transferring it to stationary by differencing. Then, the ARIMA model is built by determining proper p, d, and q to forecast the future stock price. The last step is to evaluate the accuracy of the model.

2.3 LSTM Model

LSTM is an improvement of recurrent neural networks (RNN) in complex deep neural networks. RNNs are a kind of deep network allowing information to be passed from one step of the network to the next with its feedback loops. However, traditional RNNs face the challenge to deal with long-term dependence especially when the information lack of obvious connection. Moreover, the problem of vanishing or exploding gradient caused during the updating weight recurring training process makes RNNs less applicable. Fortunately, LSTM can eliminate both problems and hence has become popular in modeling complex sequential data.

The designation of the LSTM model enables it to deal with long-term dependence and gradient problems. LSTM model consists of cell states and three kinds of the gate. The cell state is used to store past information and gates are used to manage the flow of information through these cells. Forget gate, update gate, and output gate are three kinds of gates in LSTM. Forget gate controls the flow of the information with sigmoid function, which outputs a number between 0 and 1 during the learning process. Zero implies completely forgetting the information, thus, nothing will go through the cell while one represents completely keeping the information to go through the cell. After keeping or forgetting information, the input gate helps to decide the new data to be stored in the cell state by the input gate layer and a tanh layer. Finally, the output gate determines the output value that can combine the cell state and the newly arriving data.

3. EXPERIMENTS

3.1 Forecasting by ARIMA

Investigating the WMT stock price data from 2 January 2020 to 31 December 2021, the time series presents a non-stationary characteristic in Figure 1. The ACF and PACF measurements for stationarity also indicate that the time series is not stationary as shown in Figure 2 and Figure 3. Hence, the ARIMA model cannot be applied directly until it becomes stationary. After having one time difference to time series, the new times series is more stationary as shown in Figure 4. The ACF and PACF for the new time series in Figure 5 and Figure 6 prove its stationarity. Consequently, d for ARIMA will equal 1 since the time series become stationary after the first difference.

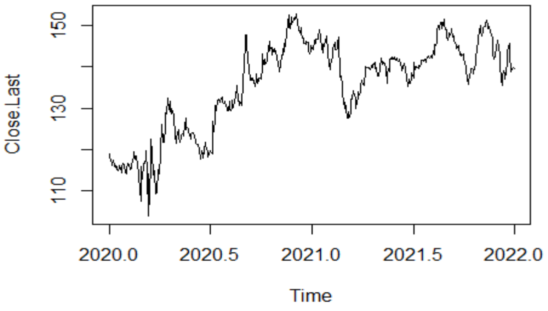

Figure 1 Time series for WMT

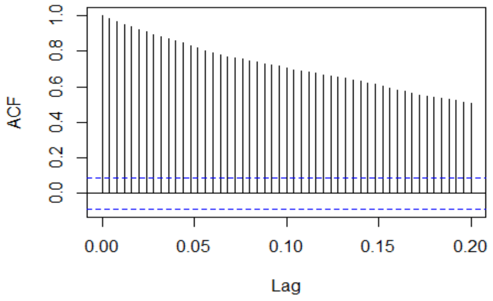

Figure 2 ACF before differencing

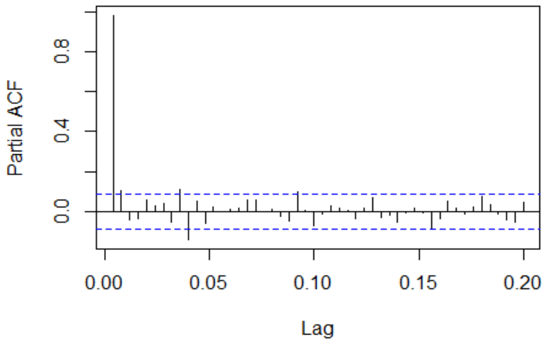

Figure 3 PACF before differencing

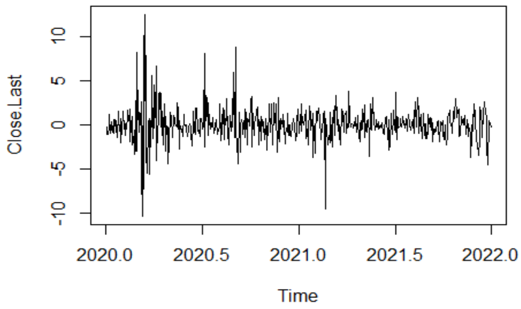

Figure 4 Time series for WMT after differencing

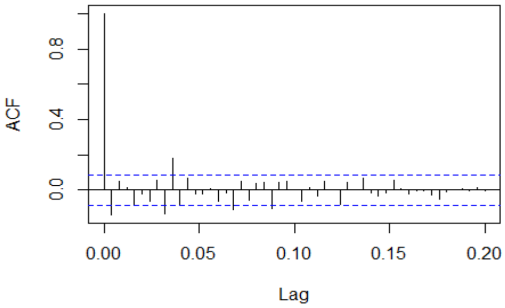

Figure 5 ACF after differencing

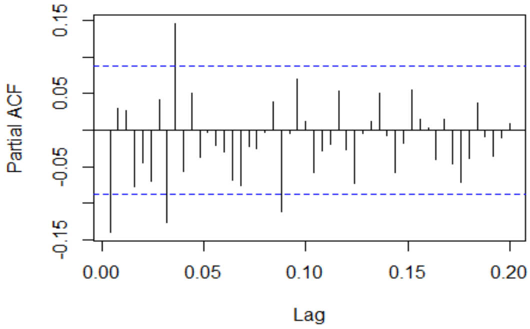

Figure 6 PACF after differencing

The Auto-ARIMA model gives (1, 1, 0) as parameters of ARIMA. However, ARIMA (1, 1, 0) will be less effective for forecasting stock price since it has little information which is not sufficient to make a complex prediction. The prediction result for the next 5 days is shown in Figure 5, where only a horizontal line represents the future price. As the result, customized model ARIMA (1, 1, 10) will be applied to the time series and the forecasting result is presented in Figure 7. Moreover, Figure 8 demonstrates the comparison between prediction results and actual stock prices for two different ARIMA models.

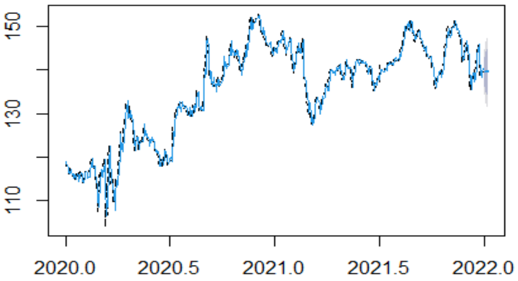

Figure 7 Forecasting based on ARIMA (1, 1, 0)

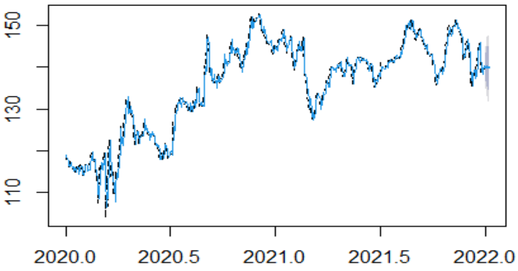

Figure 8 Forecasting based on ARIMA (1, 1, 10)

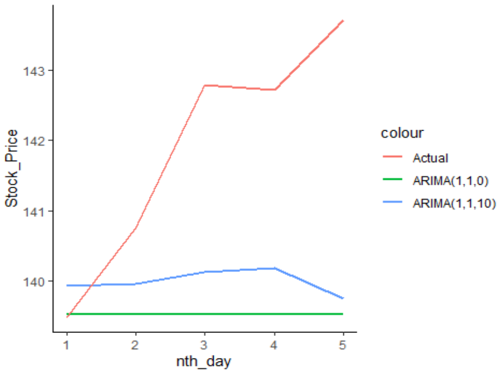

Figure 9 Results Comparison for different ARIMA models

3.2 Forecasting by LSTM

Applying the basic LSTM model to the dataset, the first 75% of data is selected as a training set and the rest of 25% is selected as the test set. Both training set and test set are normalized by the Min-Max Normalization method as shown in Equation (2). Thus, the periodic information and range information are kept by normalization. Furthermore, normalization can improve the speed to solve the optimal solution and enhance training efficiency. The variable of LSTM is defined and the LSTM model is built after normalization. As shown in Figure 10, the loss function of train data and that of test data converge well which means that the model has high robustness. Therefore, the basic LSTM model can be applied to the WMT stock basic to obtain prediction results which are shown in Figure 11.

$$x' = \frac{x-\min(x)}{\max(x)-min(x)} \qquad (2)$$

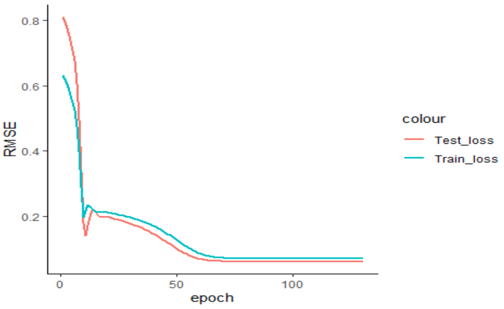

Figure 10 Loss Function

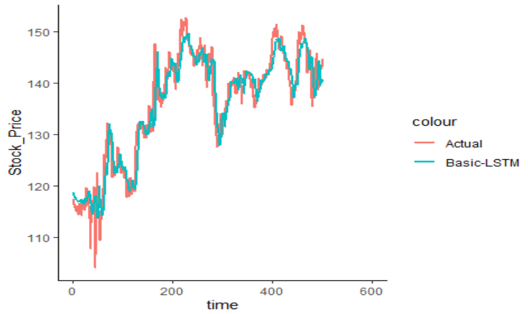

Figure 11 Forecasting based on basic LSTM

However, even the results indicate that the basic LSTM model has relatively high accuracy in the prediction of stock price, the performance can be viewed as overfitting which lacks useful information in making a real trading strategy. Thus, the multi-step LSTM model is applied to the dataset to investigate a more pragmatic prediction. The next five days' stock price is predicted step by step and the comparison between actual result and prediction is shown in Figure 12. Even though the multi-step LSTM model underestimates the stock price overall but accurately predicts the future trend of the stock price for the next five days.

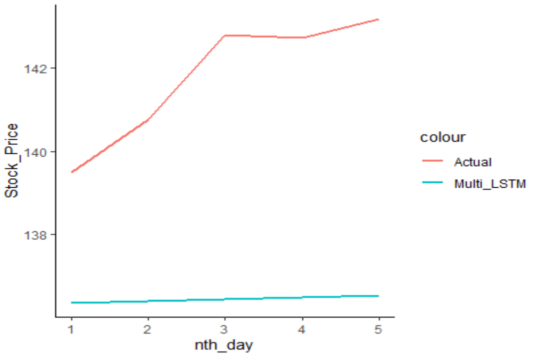

Figure 12 Forecasting based on Multi-Step LSTM

3.3 Comparison between ARIMA and LSTM

Root Mean Squared Error (RMSE) will be used to measure the accuracy of different time series models on the stock price prediction. The calculation process is presented in Equation (3) where m stands for several estimators, yi represented the actual value of the stock and $\hat{y}_i$ is the estimated value by models.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=0}^{m}(yi - \hat{y}_i)^2} \qquad (3)$$

The comparison result is shown in Table 1 where ARIMA (1,1,10) has the lowest RMSE of 1.915 while Multi-Step LSTM has the highest RMSE of 5.503. ARIMA (1,1,0) has a relatively low RMSE of 1.968 while basic LSTM has an RMSE of 3.022.

Table 1. Accuracy of different models

Forecasting Models	RMSE	Rank
Auto ARIMA (1,1,0)	1.967541	2
Customized ARIMA (1,1,10)	1.915135	1
Basic LSTM	3.021648	3
Multi-step LSTM	5.503046	4

4. CONCLUSION

Traditional time series forecasting and deep learning forecasting can be two categories of financial time series analysis. This study compares the performance of the traditional ARIMA model and deep learning LSTM model on predicting the stock price of WMT, the world's largest retailer business, from 2 January 2020 to 31 December 2021. The result shows that customized ARIMA (1,1,10) gains a higher accuracy than others in stock price prediction. Considering the problem of overfitting, the model built by LSTM lacks availability for planning the investment strategy. As the result, the prediction achieved by the ARIMA model may be more applicable. Besides, it should be stressed that the LSTM model is much more complex than the ARIMA model and lacks interpretability. Nonetheless, the ARIMA model fails to capture the stock price trend for the next 5 days while multi-step LSTM accurately predicts the upward stock price trend for the next 5 days even underestimates the price overall. Therefore, it may indicate that the ARIMA model could be used for forecasting very short-term stock prices such as one day while the multi-step LSTM model is suitable for predicting relatively long-term stock price trends such as a week.

REFERENCES

[1] M. S. Mhatre, F. Siddiqui, M. Dongre, Paramjit Thakur, A Review Paper on Artificial Neural Network: A Prediction Technique, International Journal of Scientific & Engineering Research, vol. 16, no. 12, 2015, pp. 161-163.

[2] B. W. Wanjawa, L. Muchemi, ANN Model to Predict Stock Prices at Stock Exchange Markets, 2015, pp. 3-20.

[3] K. Zhou, W. Y. Wang, T. Hu, C. H. Wu, Comparison of Time Series Forecasting Based on Statistical ARIMA Model and LSTM with Attention Mechanism, Journal of Physics: Conference Series, vol. 294, no. 1, 2019, pp. 3. DOI : https://doi.org/10.1088/1755-1315/294/1/012033

[4] G. E. P. Box, G. M. Jenkins, Time series analysis: forecasting and control. Journal of Time, vol. 31, no.

3, 2010, pp. 190-200. DOI: https://doi.org/ 10.1111/j.1467-9892.2009.00643.x

[5] R. Nau, Mathematical structure of ARIMA models, 2014, pp.1–8.

[6] B. D. Zumbo, E. Kroc, Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS [Heteroscedasticidad en análisis de regresión múltiple: qué es, cómo detectarlo y cómo resolverlo con aplicaciones en R y], The Journal of Modern Applied Statistical Methods, vol. 17, 2019. DOI: https://doi.org/ 10.22237/jmasm/1555355848

[7] T. Bollerslev, Generalized Autoregressive Conditional Heteroskedasticity, Journal of Econometrics, 1986, pp. 307–328. DOI: https://doi.org/10.1016/0304-4076(86)90063-1

[8] S. T. T. Wang, Research on the Volatility of BYD Stocks Price Based on GARCH Family Model, 1994, pp. 280–284. DOI: https://doi.org/10.14018/j.cnki.cn131085/n.2020.09.118

[9] H. F. Assous, N. Al-Rousan, D. Al-Najjar, H. Al-Najjar, Can international market indices estimate TASI’s movements? The ARIMA model. Journal of Open Innovation: Technology, Market, and Complexity, 2020, pp. 1–17. DOI: https://doi.org/10.3390/joitmc6020027

[10] Q. Yang, C. Wang, A study on forecast of global stock indices based on deep LSTM neural network, Statistical Research, 2019, vol. 36, no. 6, pp. 65–77.

[11] S. F. Milad, S. H. R. Hajiagha, Forecasting Stock Price Using Integrated Artificial Neural Network and Metaheuristic Algorithms Compared to Time Series Models, Soft Computing, vol. 25, no. 13, 2021, pp. 483–513. DOI: https://doi.org/10.1007/s00500021057755

[12] S. Siami-Namini, N. Tavakoli, A.S. Namin, A Comparison of ARIMA and LSTM in Forecasting Time Series, 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018, pp. 1394–1401. DOI: https://doi.org/10.1109/ICMLA.2018.00227

[13] W.J. Lu, J. Z. Li, Y. F. Li, A. J. Sun, J. Y. Wang, A CNN-LSTM-Based Model to Forecast Stock Prices, Introduction, 2020, pp. 11. DOI: https://doi.org/10.1155/2020/6622927

[14] DOI:https://www.yahoo.com/author/yahoo-finance

[15] S. Khan, H. Alghulaiakh, ARIMA Model for Accurate Time Series Stocks Forecasting, International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, 2020. DOI: https://doi.org/10.14569/IJACSA.2020.0110765