

Application of ARIMA and LSTM Model to the Forecast of CSI 300 Close Price

Zebang Zhang *

School of Marine Sciences, Sun Yat-sen University, Zhuhai, China

* Corresponding Author Email: zhangzb25@mail2.sysu.edu.cn

Abstract. This article uses two time-series analysis model, ARIMA model and Long Short-Term Memory neuron network model to predict the close price of CSI 300 Index. The ARIMA model only uses the past close price values to predict future close price while the LSTM model combine multiple variables into the prediction of future close price. The methods of fitting data with these two models are introduced. The predictive performance of the two fitted models is evaluated by calculating the Root Mean Squared Error, which is an indicator of the difference between predicted values and real values. This article then analyzes the prediction results and proposes the respective advantages and limitations of the two models. It is found that the ARIMA model performs better when predicting close price in a shorter period but may not perform well when doing long-term prediction. The LSTM model can predict a similar series trend with multiple variables and its parameters can be adjusted to get smaller RMSE and improve prediction performance.

Keywords: ARIMA model; long short-term memory model; root mean squared error; prediction.

1. Introduction

In securities investment, stocks are a very important investment object. For market investors, the volatility of stock prices is often unpredictable, resulting in high investment risks. For listed companies, changes in stock prices are very crucial because the rise in stock prices can bring more financial support to them and promote their development. For a country, the healthy development of the stock market is conducive to guiding the rational allocation of market funds and resources [1]. Therefore, in order to better predict stock price, it is meaningful and necessary to explore reasonable and effective prediction methods. However, the stock market is not a simple linear market, it is a nonlinear market influenced by multiple, external factors, which makes it hard to predict stock price [2].

The rapid development of machine learning has driven the development and application of neural networks, and a large amount of research has confirmed that using neural networks can make better predictions for stock price [3]. At present, comparing the predictive ability of traditional time series analysis models such as the ARIMA model and neural network models such as Long Short-Term Memory Network, finding the optimal prediction model has become a worthwhile research topic.

The ARIMA model is a common time series model that comprehensively considers the relationship between financial markets and stock market indices in economic forecasting. It is a good tool for predicting short-term stock market trends. At the same time, the ARIMA model is becoming increasingly sophisticated in technology, using a finite parameter linear model to illustrate the autocorrelation characteristics of time series for statistical analysis and mathematical calculations. It has properties such as simplicity, feasibility, and sensitivity, and is currently the most common model that can be used to fit time stationary sequences [4]. LSTM neural network can handle long-term dependency problems. It has been used to make forecast for many time series successfully. The LSTM neural network has a special information storage memory unit. Simultaneously, Forget Gates, Input Gates, and Output Gates were constructed to control the memory unit [3]. Forget gate decides how much unit information should be abandoned, Input gate updates the state, and Output gate decides gives the output of the unit. Other studies have shown that specific parameters of LSTM can be adjusted to achieve higher levels of prediction accuracy [5]. Neural networks are capable of dealing with multi-variable problems. After combining multiple data variables, the conclusion is that the

multi-feature LSTM model has better prediction performance and can obtain more accurate stock price data trends than LSTM model with only one feature [6].

The research and prediction object of this article is the sample data of the daily CSI 300 Index, covering a total of 261 days from July 1, 2022, to July 26, 2023, including the open price, high price, low price, close price and trading volume of each day. In terms of prediction methods, this article adopts the ARIMA model and LSTM model. When predicting the close price of CSI 300 Index, the open price, high price, low price, close price and trading volume will also be used as the multiple input features of LSTM neural network. This article adopts the Root Mean Squared Error (RMSE) to define the prediction ability of each model. Finally, after successfully fitting the two models and using the fitted models to make forecasts, this article will analyze the prediction results and conclude the respective advantages and limitations of the two models. For the LSTM neural network model, this article will continuously adjust and optimize the parameters of the model, such as the parameter of the Dropout layer and the number of hidden units on the LSTM layer, to obtain the better LSTM model [7].

2. Methods

2.1. Data Source and Description

The data used in this paper is the daily CSI 300 index, from July 1, 2022, to July 26, 2023, totally 261 days. Data on each day include the open price, high price, low price, close price, adjusted close price and volume. The adjusted close price is abandoned in this paper. The dataset can be downloaded from the Yahoo Finance website.

2.2. ARIMA Model

The ARIMA model stands for Autoregressive Integrated Moving Average model (p, d, q) of a time series, y_t . In the general ARIMA model, y_t is expressed as the following equation:

$$y_t = c + \varphi_1 y_{d_{t-1}} + \cdots + \varphi_p y_{d_{t-p}} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t \quad (1)$$

In the equation, y_{d_t} represents the differenced, stationary time series as differencing is helpful when converting a non-stationary time series, y_t to a stationary one. The value of d stands for how much times the original time series is differenced. φ represents the coefficient of the differenced time series lagged value, and the number of φ is expressed as p. Similarly, θ represents the coefficient of lagged error terms, and the number of θ is expressed as q.

2.3. Long Short Term Memory Network

The open price, high price, low price, trading volume and close price of the last day is set as the input features, and the close price of the day is set as the output label of LSTM network. In this paper, time step is set as 1, which means that the five variables on the last day are used as the input features for the prediction of today's close price. To accomplish this task, a LSTM model is built based on the deep learning framework Keras in Python and several steps should be taken.

2.3.1 Data preprocessing

This includes data importing and data normalization and transformation. After importing the sample data, by using the following equation, the data will be readjusted to the range of 0 to 1. Normalized data will help avoid the gradient error values in the LSTM model [8].

$$X_{scaled} = (X - X_{min}) / (X_{max} - X_{min}) \quad (2)$$

Then, the normalized dataset needs to be transformed into a supervised learning dataset, which means to create labels for input features. In this study, open, high, low, volume, close price (5 features)

of the last day is used to predict today's close price, so today's close price should be the label for the feature values.

2.3.2 Dividing and reshaping data

80% of the preprocessed data is then divided into the training data set while the rest 20% of the preprocessed data is divided into the test data set. After that, both data sets are reshaped to become suitable for the LSTM model according to its size, the chosen time step and the number of features [8,9].

2.3.3 Building LSTM model

First, in the established sequential model, this paper first adds a LSTM layer to it. In this layer, the parameters needed to be set include number of neurons, time steps and the number of feature variables. Loss function is chosen as Mean Absolute Error (MAE). The optimization algorithm uses Adam.

Then another Dropout layer that randomly sets a portion of the input units as 0 to prevent overfitting is added. Finally, this paper adds a Fully Connected layer to the model to output the prediction result.

2.3.4 Model training and forecasting

While using the training data set to train the data, the loss values are recorded. After the network is trained, it is fitted to the test data set and make its prediction. The predictive performance is evaluated through the comparison between predicted values and real values using RMSE, where y_i represent real series values and $\hat{y}_i$ represent the predicted values [10].

$$RMSE = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n} \quad (3)$$

3. Results and Discussion

3.1. ARIMA Model Results

The series of close price from July 1, 2022 to July 26, 2023, in total 261 days is plotted, and Augmented Dickey-Fuller Test (ADF test) is used to test if it is stationary (Table 1), then the results indicates that it's a non-stationary series. Therefore, the series is differenced for one time and the differenced close price series is plotted again, which seem to be stationary [11]. Then the ADF test is applied to the differenced close price series, and it proves that the differenced series is stationary (Figure 1 and 2).



Fig. 1 CSI 300 close price

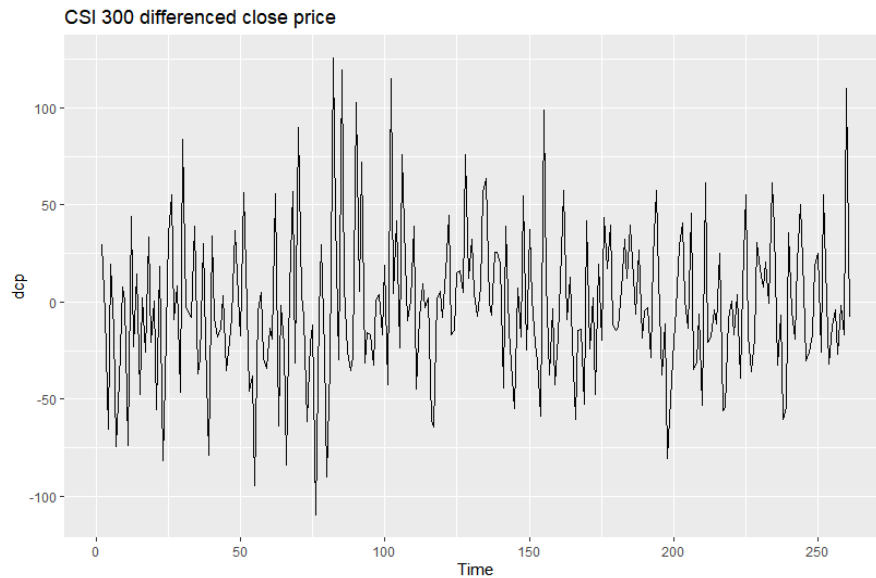


Fig. 2 CSI 300 differenced close price.

Table 1. Results of Augmented Dickey-Fuller Test

Series	P-value	Conclusion
Close price	0.42(>0.05)	Non-stationary
Differenced close price	0.01(<0.05)	stationary

The ACF and PACF of the differenced close price series are plotted to help decide the value of p and q , which are the parameters of the ARIMA model. From the Figure 3 below, it is obvious that ACF is relatively large at lag 5, while PACF is quite significant at lag 5. Moreover, the AIC criterion tells lower AIC value represent the better fit of a model. The AIC values of different ARIMA ($p, 1, q$) models are also calculated. The conclusion of the best p and q values which give the lowest AIC value are consistent with the observation from the ACF and PACF plots. Therefore, ARIMA (5,1,5) are chosen to fit the close price series [12].

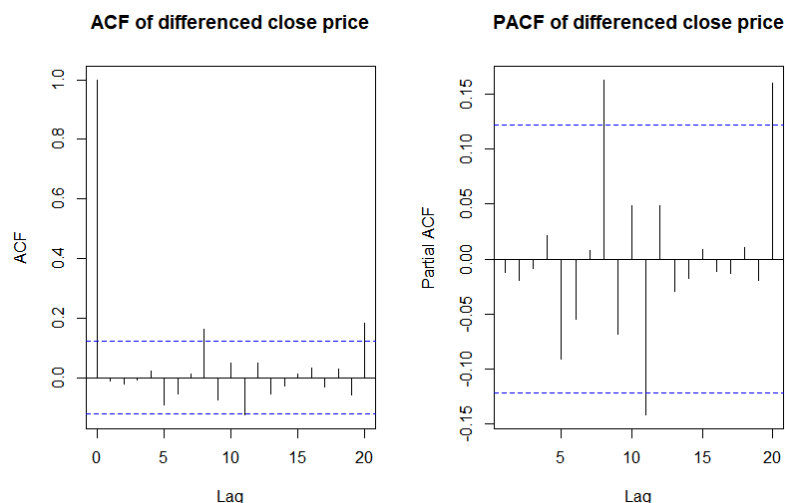


Fig. 3 ACF and PACF of differenced close price

The coefficients of the ARIMA (5,1,5) model are shown in the table 2 below.

Table 2. Parameters of ARIMA (5,1,5) model

	c	φ_1	φ_2	φ_3	φ_4	φ_5	θ_1	θ_2	θ_3	θ_4	θ_5
value	0	0.118	-0.852	-0.861	0.173	-0.933	-0.100	0.894	0.899	-0.120	0.927

The first ARIMA (5,1,5) model is fitted by the close price of the first 240 days and will be used to forecast those of the next 21 days. Another ARIMA (5,1,5) model is fitted by the close price of the first 216 days and will be used to forecast those of the next 45 days. The residuals of each ARIMA (5,1,5) model are checked to ensure that they are consistent with white noise because only when they are consistent with white noise can it be concluded that the model is valid. Ljung-Box test is applied to the residuals (Table 3).

Table 3. Results of Ljung-Box test

Days of forecast	P-value	Conclusion
21	0.15(>0.05)	White noise
45	0.21(>0.05)	White noise

After ensuring the models are valid, the models are used for prediction. To better visualize the difference between real values and predicted values, they are plotted in the same figure. The predicted values are drawn as the red curve in Figure 4 and 5 below and the real values are drawn as the white curve.

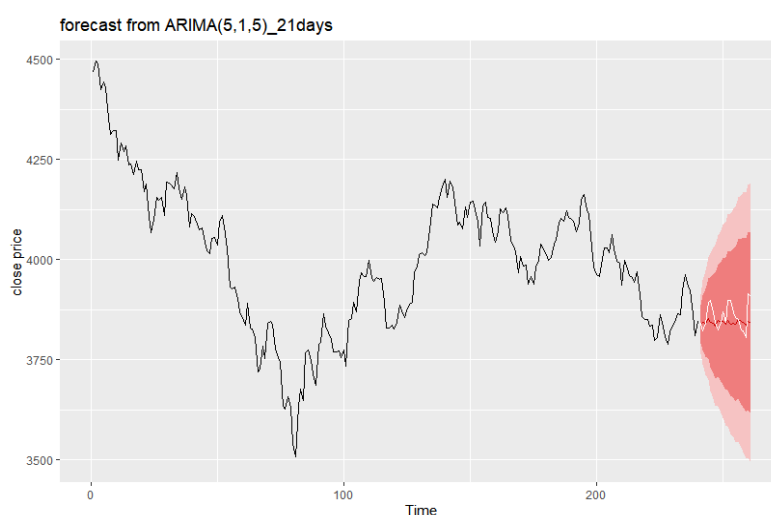


Fig. 4 Forecast from ARIMA (5,1,5) _21days

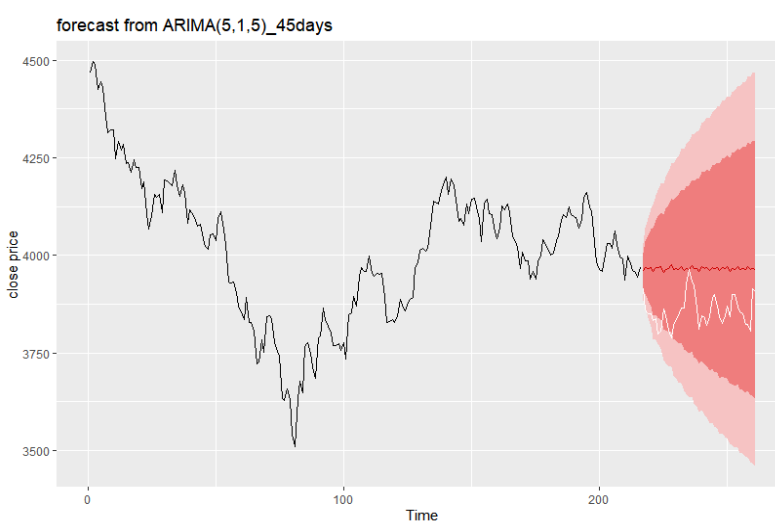


Fig. 5 Forecast from ARIMA (5,1,5) _45days

The metric, RMSE is calculated in each case (Table 4). Obviously, the RMSE of forecasting 21 days is greatly smaller than that of forecasting 45 days, indicating that the prediction is much better when forecasting 21 days. Moreover, according to Figure 5, the real values differ much with the predicted values, indicating that the forecast is unsatisfactory.

Table 4. Metric of RMSE

Days of forecast	RMSE
21	26.48
45	124.06

3.2. LSTM Results

The close price series is first imported, then normalized, transformed into a supervised learning dataset and finally divided into two sets. Moreover, inside the two sets, the input features and output label are separated again. Generally, when using the training set to train the model, the values of the loss function will gradually decrease and then remain at a low level as the epoch number increases [13].

After trying different parameters of the models and testing the LSTM model for some times, it is found that RMSE is relatively small when the number of units on the LSTM layer is set as 100 and the parameter of the dropout layer is set as 20% to prevent overfitting.

The two sets are given to the adjusted LSTM model for training, the loss curve is shown in the figure below. When the number of epochs nearly reaches 50, the loss value decreases very slowly, and the loss value of the test set is smaller than that of train set, so 50 epochs will be proper (Figure 6).

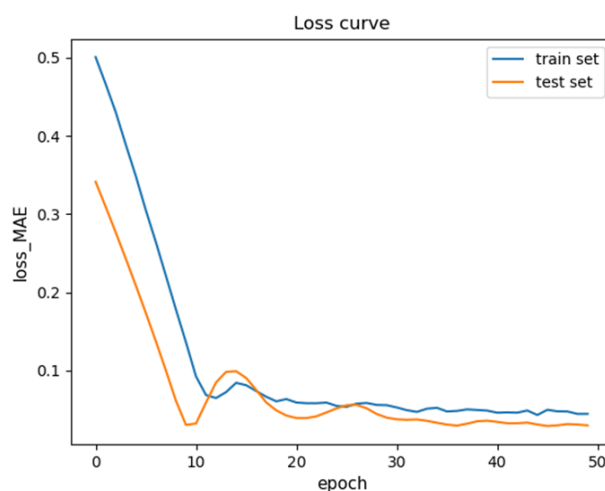


Fig. 6 Loss curve of LSTM training process

After the model is trained, the test dataset is used for prediction and evaluation of the predictive performance. The test data includes the input features' values and real close price of 52 days, the predicted values and real values are plotted below (Figure 7).

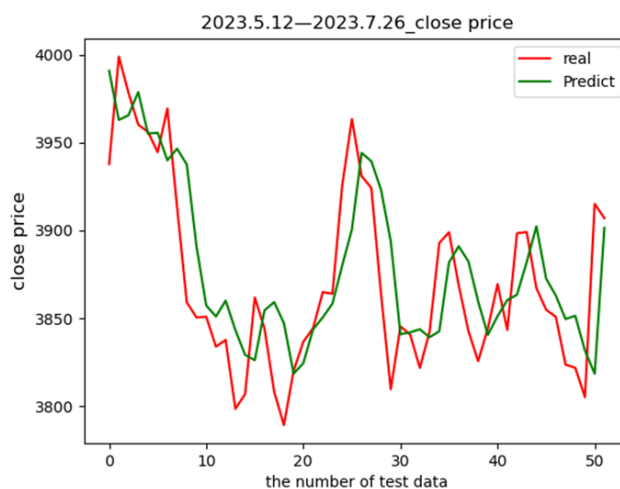


Fig. 7 Forecast result from LSTM model_52days

It is clear that predicted value curve has very similar trend with the real value curve. Moreover, the volatility of the two curves do not differ much. This proves that the LSTM model is capable of using multiple variables to make predictions. Then the selected metric, RMSE, is calculated. In fact, the parameters of the LSTM model, which are dropout and units of LSTM layer, are changed and adjusted for some times to find the model with better prediction performance. The results are that when the two parameters change, the changes in the loss curve and the forecast result are hard to observe on plots because they are very slight. However, the metric, RMSE, will change value, indicating that the model with smallest RMSE has the best prediction performance among others (Table 5).

Table 5. Different model parameters and metrics

Numbler	dropout	Units of LSTM layer	RMSE
1	0.2	100	35.24
2	0.2	64	39.69
3	0.5	100	37.45

4. Conclusion

Both the ARIMA model and the LSTM model can be used as a tool to forecast CSI 300 close price. The method of fitting them to the time series is different and their prediction performance is different. The ARIMA model has relatively small Root Mean Squared Error when predicting CSI 300 Index close price in a short period. However, when ARIMA model is used for long period prediction, the RMSE may significantly increase, and the prediction result may be unsatisfactory as the volatility of stock price may be much greater than that of the predicted values. Therefore, the ARIMA model has the advantage of predicting short-term stock price in some cases but has the limitation of only being able to capture linear relationships, not being able to capture nonlinear relationships.

The Long Short-Term Memory model has the quite outstanding advantage of being capable of handling problems with multiple variables. This means that when doing stock price prediction, other features like trading volume and high price can be added as the input to enhance the accuracy of prediction. Besides, various model parameters like the parameter of dropout layer, the number of units on LSTM layer can be continuously adjusted to get smaller RMSE. If the model and its parameters are suitable, it can predict a relatively accurate trend of stock price and obtain similar price volatility. However, the LSTM model is more complex than many other models and training the model with large data sometime takes time.

References

- [1] Chen Zejun. Research on the Prediction of S&P 500 Stock Returns Based on LSTM Neural Network. Harbin Institute of Technology, 2021.
- [2] Shi Min. Predicting Stock Prices Based on GARCH Model and BP Neural Network. Nanjing University, 2018.
- [3] Sui Jincheng. Research on Stock Forecasting Based on LSTM Neural Network. Qingdao University of Science and Technology, 2021.
- [4] Zhao Ning. Short term prediction of Vanke stock price based on ARIMA model. Financial Journal, 2023, (01): 34-36.
- [5] Ding Wenjuan. Comparison of ARIMA Model and LSTM Model Based on Stock Prediction. Industrial Control Computer, 2021, 34 (07): 109-112+116.
- [6] Xi Xiaoya, Qin Hebin, Lu Zhijuan. Prediction of Stock Price Changes Based on LSTM Neural Network Model: Taking Baidu Stock Price as an Example. National Circulation Economy, 2022, (16): 102-105.
- [7] Yang Chuncan. Research on Stock Forecasting Based on LSTM. Northeastern University of Finance and Economics, 2022.

- [8] Abdul Quadir Md, et al. Novel optimization approach for stock price forecasting using multi-layered sequential LSTM. *Applied Soft Computing*, 2023.
- [9] Moghar A, Hamiche M. Stock Market Prediction Using LSTM Recurrent Neural Network. *Procedia Computer Science*, 2020, 170: 1168-1173.
- [10] Hum Nath Bhandari, et al. Predicting stock market index using LSTM. *Machine Learning with Applications*, 2022.
- [11] Huang Mengting. Empirical Study on Stock Price Prediction Based on ARIMA Model. *Neijiang Technology*, 2023, 44(03): 61-62
- [12] Shi Linfeng. Research on the closing price of Vanke Company's stock based on ARIMA model. *Business Exhibition Economics*, 2021, (03): 50-52.
- [13] Rather A M. LSTM-based Deep Learning Model for Stock Prediction and Predictive Optimization Model. *EURO Journal on Decision Processes*, 2023.