Comparison of ARIMA and LSTM for Stock Price Prediction

Haoran Wu[1,a,*], Shuqi Chen[1,b], Yicheng Ding[2,c]

[1]*Wenzhou-Kean University, No. 88, University Road, Li' ao Street, Ouhai District, Wenzhou, Zhejiang, China*
[2]*College of Business and Public Management, Wenzhou, China*
[a]*wuhaor@kean.edu, [b]chenshuq@kean.edu, [c]dingyih@kean.edu*
[*]Corresponding author

Abstract: The prediction of time series is an extremely important and challenging part, and due to some other unavoidable external factors, it can also affect its prediction results. Therefore, in order to compare which model can produce better predictions, this paper explores the following two models. First of all, the more traditional and classical model is the "Autoregressive Integrated Moving Average" (ARIMA). Moreover, according to the continuous exploration and development over the years, more variables have evolved, such as SRIMA (seasonal ARIMA), etc. This model shows considerable advantages in short-term prediction but has more disadvantages in the long term, which is not the best choice. Another model is closer to today's emerging technologies and more dependent on artificial intelligence data analysis, such as "Convolutional Neural Network" (CNN) and "Recurrent Neural Network" (RNN), among which there are many variants. This paper focuses on a special variant, Long Short-Term Memory (LSTM) neural network, which can learn from past data and relate to current data. Furthermore, it is found from the data in this paper that LSTM's prediction of data is better than ARIMA model.

1. Introduction

Apple released the first iPhone on January 9, 2007. So far, the iPhone has sold more than 2 billion units, making it the most popular mobile phone in the world. According to market research firm Counterpoint [1], from 2007 to 2022, Apple sold 2,167.39 million phones. At the same time, it has a very high market share of nearly 20 percent worldwide.

Apple doesn't just own the phone business. Its products and services include iPhones, Macs, iPods, Air Pods, Apple TV, Apple Watch, Beats products, Apple Care, iCloud, digital content store, streaming, and licensing services. Apple's product innovations have always led to changes in the business technology world. Like Apple's Air Pods, the first truly wireless headphones, Apple's M series chips, which are equipped with Mac and iPad, are a new era of ARM architecture chips, which also announced that Apple's self-developed chips (Apple Silicon) could cover most of its products and devices. Its innovation and product lines bring it significant revenue. According to Apple Reports

Third Quarter Results [2], Apple's fiscal third-quarter net income was $82.959 billion, and net profit was $19.442 billion. Net income from Macs was $7.382 billion, iPads $7.224 billion, and wearables, home products, and accessories $8.084 billion. This is a revenue level unmatched by any technology company in the world.

Apple's technology products not only change the lifestyle of consumers but also play a vital role in the development of the general direction of the technology field and the technological upgrading at the source of the supply chain. Therefore, understanding Apple's stock price has become an important issue, and using the appropriate time series method to predict it is crucial. For example, the accurate prediction of stock prices will provide investors with helpful decision-making information to grasp the right direction of technology development investment.

"Changes in stock prices have substantial explanatory power for U.S. investment, especially for long-term samples, and even in the presence of cash flow variables. [3]" For investors, stocks make sense. Stocks are profitable. When investors buy stocks, they get dividends for the growth of the company. They can also speculate by buying stocks. Buy low, sell high and then gain the difference. According to Lee et al. [4], to design an optimal investment strategy, it is important to accurately predict the market by understanding its characteristics.

This study uses various machine learning methods to predict the closing price of Apple. Collecting 1188 historical data (from 1/4/2016 to 9/21/2020) and using two models to predict the closing price of Apple. These models are"Auto-regressive Integrated Moving Average" (ARIMA) and "Long Short-Term Memory" (LSTM). By comparing the two models' training and testing, the one that can predict the closing price of Apple is found.

2. Literature review

Prediction is an indispensable and challenging part of time series data analysis. Meanwhile, the data type and underlying context of the time series will affect the performance and accuracy of the time series data analysis and prediction technology adopted. Other factors, such as seasonality, economic shocks, unexpected events, and other internal changes can also affect the forecast [5].

This paper aims to use two kinds of time series prediction models to compare the data errors and accuracy in the prediction of Apple stock data, so as to select the more advantageous time series model. Besides, there are several reasons why this paper chooses the data of Apple stock to analyze: first, Apple is a company with a global reputation, and the data is reliable and representative to a certain extent; second, Apple has a long history, and the changes in its data also reflect the economic development trend at different times, which can provide a certain reference for the prediction of future economic changes.

Therefore, the first model to be explored in this paper is the "Auto-regressive Integrated Moving Average", also known as ARIMA [6]. Compared with other models, this is a more traditional and classical time series model. This technology is generally completed through the following steps: First, linear regression is used for model fitting, and then moving average is used for prediction. This linear regression-based approach has been developed over the years, and many variants of the model have been developed, such as SARIMA (seasonal ARIMA) and ARIMAX (ARIMA with explanatory variables). These models perform fairly well in the short term but have a lot of weaknesses in the long term [7].

Another model, which is closer to today's emerging technology, is machine learning, especially the method based on deep learning, which is based on artificial intelligence data analysis, and will elevate the data analysis process to zero level, in which the model built is data-driven rather than model-driven. In addition, for the underlying application field, the best learning model can be trained. For example, "Convolution Neural Network" (CNN) is suitable for image recognition and other

problems, and "Recurrent Neural Network" (RNN)[8] is more suitable for modeling problems, such as time series data and analysis. However, there are many variants of RNN-based models. Most of these RNN-based models differ mainly in their ability to remember input data. Another special type of RNN model that this article focuses on is the "Long Short-Term Memory" (LSTM) network. It models relationships between longer input and output data. This RNN-based model such as LSTM, known as a feedback-based model, can learn from past data and uses gates like a grid structure to remember past data and build a model of past and current data so that the input data is traversed only once [9].

However, according to most experimental studies, it can be found that the deep learning-based model is superior to the traditional Arima-based model in predicting time series, especially for long-term prediction. Therefore, in order to prove this point, this paper proves that LSTM model is superior to ARIMA model by collecting the same data and applying it to two different models [10].

3. Methodology

3.1. Auto-regressive Integrated Moving Average (ARIMA)

The ARIMA (p, d, q) model is a linear analysis model, that uses time series data to make predictions. It is characterized by 3 terms: *p*, *d*, *q* where *p* represents the order of AR term, a simple form of an AR model of order p, i.e., AR(p), can be written as a linear process given by:

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t \tag{1}$$

where x_t is the stationary variable, c is constant, the terms in ϕ_i are autocorrelation coefficients at lags 1, 2, p and ϵ_t, the residuals, are the Gaussian white noise series with mean zero and variance σ_ϵ^2. q represents the order of the MA term and *d* represents the order of differencing to transform a non-stationary series to stationery. An MA model of order q, i.e., MA(q), can be written in the form:

$$x_t = \mu + \sum_{i=0}^{q} \theta_i \epsilon_{t-i} \tag{2}$$

where μ is the expectation of x_t (usually assumed equal to zero), the θ_i terms are the weights applied to the current and prior values of a stochastic term in the time series, and θ_0=1. We assume that ϵ_t is a Gaussian white noise series with mean zero and variance σ_ϵ^2. We can combine these two models by adding them together and form an ARIMA model of order (p, q):

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t + \sum_{i=0}^{q} \theta_i \epsilon_{t-i} \tag{3}$$

The methodology starts with identifying models based on autocorrelation (ACF) behavior and partial autocorrelation (PACF) plots. Following the initial nonseasonal differencing, the autocorrelation (ACF) and partial autocorrelation (PACF) plots for the training data are shown in the figures below. Because a phenomenon of the white noise process is present in both plots, a random walk model and a random walk with drift are constructed based on the behavior of the ACF and PACF plots.

A one-step out-of-sample prediction is therefore the most effective for predicting. The technique used in this case does multi-step out-of-sample forecasting with re-estimate, which means that every time the model is re-fitted, the best estimation model is constructed. The method takes the historical data set as input, creates a forecast model, and then outputs the prediction's RMSE value. The method creates two data structures to store the constantly projected values for the test data sets, called "prediction," and the continually added training data set, called "history," at each iteration.

3.2. Long Short-Term Memory (LSTM)

Unlike ARIMA models, the LSTM model can distinguish between short-term price spikes or falls and long-term trend reversals with the use of indicators like MACD and RSI when used to forecast stock prices. This will enable us to differentiate between actual price trends and market oddities. Additionally, the LSTM model is a type of Recurrent Neural Network (RNN) that can store and absorb knowledge from a protracted series of observations. A multi-step univariate forecast technique was used to construct the algorithm. The entire process' mathematical connection can be stated as follows:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{4}$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{5}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{6}$$

$$\overrightarrow{C_t} = tan(W_c h_{t-1} + U_c x_t + b_c) \tag{7}$$

$$C_t = C_{t-1} \odot f_t + i_t \odot \overrightarrow{C_t} \tag{8}$$
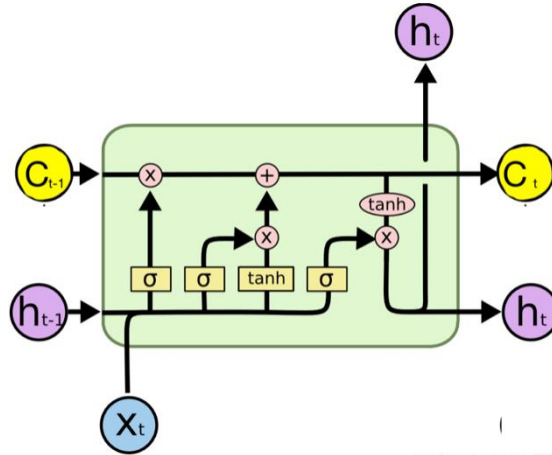
$$h_t = o_t \odot tanh(C_t) \tag{9}$$

Figure 1: The principle of LSTM model.

where W_i, W_i, W_o, W_c are the weights of h_{t-1}. U_i, U_f, U_o, U_c are the weights of x_t. b_i, b_f, b_o, b_c are bias factors. σ is the sigmoid function making sure that the values of i_t, f_t, o_t are between 0 and 1. tanh is the hyperbolic tangent function. It can be conducted that the information contained in the concealed state at the previous instant and the present cell state C_t both have an impact on the magnitude of h_{t-1}. Additionally, W_c, the primary factor that caused the gradient to vanish, has no bearing on the computation of the present cell state. Therefore, the problem of gradient disappearance during the training process is successfully addressed by the addition of a gating structure, and the precision of model prediction is also enhanced. In order to solve the problem of short-term memory of RNN and make it effectively use long-term timing information, the LSTM model introduces three logic control units: Input Gate, Output Gate, and Forget Gate. And they are connected to a multiplicative element. The input and output of information flow and the state of the Memory cell are controlled by setting the weight at the edge of the memory unit of the neural network connecting with other parts. In the training process of LSTM, the data features at time t are first input to the input layer, and the results are output through the activation function. Input the Output result, the hidden layer output at time t-1, and the information stored in the cell unit at time t-1 into the node

of the LSTM structure and output the data to the next hidden layer or output layer through the processing of Input Gate, Output Gate, Forget Gate and cell unit. Output the results of LSTM structure nodes to the neurons of the output layer, calculate the backpropagation error and update each weight. LSTM model's specific structure is shown in Figure 1.

4. Results

In the beginning, we extracted historical time-series Apple stock data from Yahoo Finance which is from 1/4/2016 to 9/21/2020. Each data set features a number of variables such as Open, High, Low, Close, and Adjusted Close. We use the "Close" variable as the only feature to be fed into the ARIMA and LSTM models. Each data set was split into two subsets: training and test datasets where the last 60 observations of the dataset was used for testing the accuracy of the model and the remaining dataset was used for training.

We use root mean square error to assess each model's effectiveness in forecasting Apple's closing price (RMSE). The root mean square error (RMSE) is a popular metric for analyzing a model's prediction accuracy. It computes the differences or residuals between the observed and predicted values. The measure compares prediction errors of multiple models for a same dataset, rather than between datasets. The indicator equation is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \tag{10}$$

where Y_i is the actual observations, $\hat{Y}_i$ is the predicted value obtained proposed forecasting model and N is the number of forecast values. The best forecasting model will be the one with the lowest RMSE.

We begin with the stationarity test of ARIMA model-ADF test. First, perform regression on the equation and check the ADF test to see if the hypothesis $\delta = 0$ can be rejected. The ADF test statistics is -0.104 while the p-value is 0.949, there is weak evidence against the null hypothesis(H0), so we accept the null hypothesis that the data is not stationary. To deal with the non-stationarity, the first difference in Apple stock price is used to convert the non-stationary data to stationary data. The ADF test statistics after the first difference is -6.131 while the p-value is 8.401. There is strong evidence against the null hypothesis(H0), so we reject the null hypothesis,i.e. data is stationary. The parameter d is the order of difference frequency changing from non-stationary time series to stationary time series.

Then we identify models based on the behavior of autocorrelation (ACF) and partial autocorrelation (PACF) plots. After that, we can determine the preliminary values of autoregressive order p, the order of differencing d, and the moving average order q, which is (1, 1, 1) for the parameter of the ARIMA model.

Table I presents the findings. According to the data, the average Rooted Mean Squared Error (RMSE) for ARIMA and LSTM models is 3.261 and 0.237, respectively. This indicates that LSTM significantly reduces error rates. According to the RMSE values, LSTM-based models perform better than ARIMA-based models.

In Figure 2(a), This is Apple's stock price after a first-order difference, with the y-axis representing the price and the x-axis representing the date. The blue line represents the training data, the green line is the predicted Apple stock price of the ARIMA model, and the red line represents the actual price. Figure 2(b) shows Apple's closing price over date. The blue line represents the true stock price, the red line represents the test data set, and the yellow line represents the predict value of the LSTM model. From Figure 2. we can conclude that comparing ARIMA and LSTM model, the prediction accuracy of the LSTM is superior to ARIMA based on Apple stock price. By observing the results,

we can also find that the ARIMA model's prediction is directional, which comes from its model assumption liner, while the LSTM model has a better ability the reflect the fluctuations of the stock.
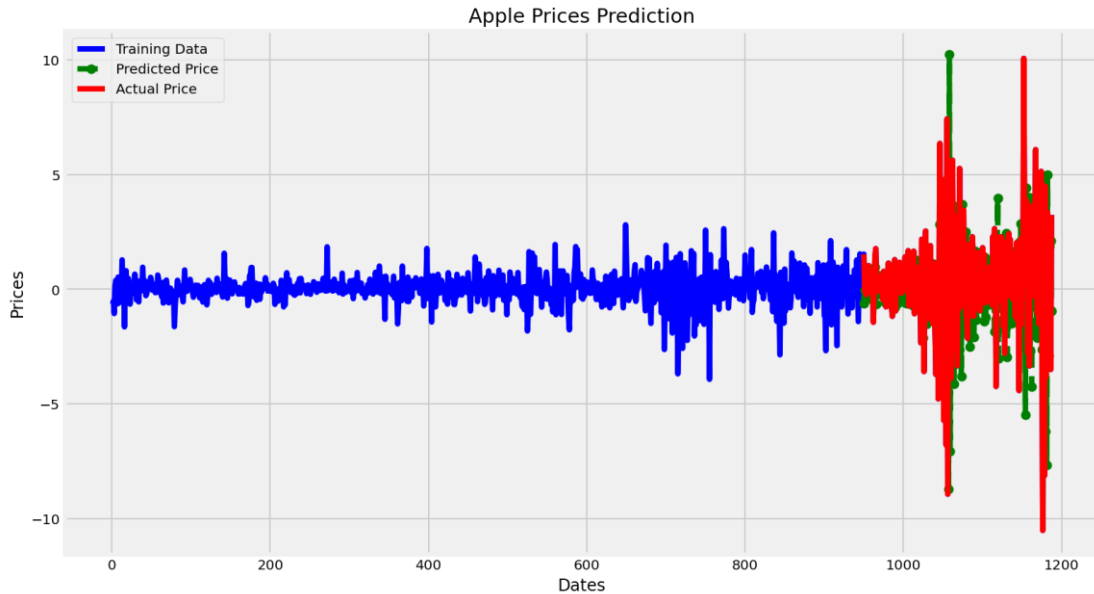
(a) APPLE price prediction using ARIMA Model
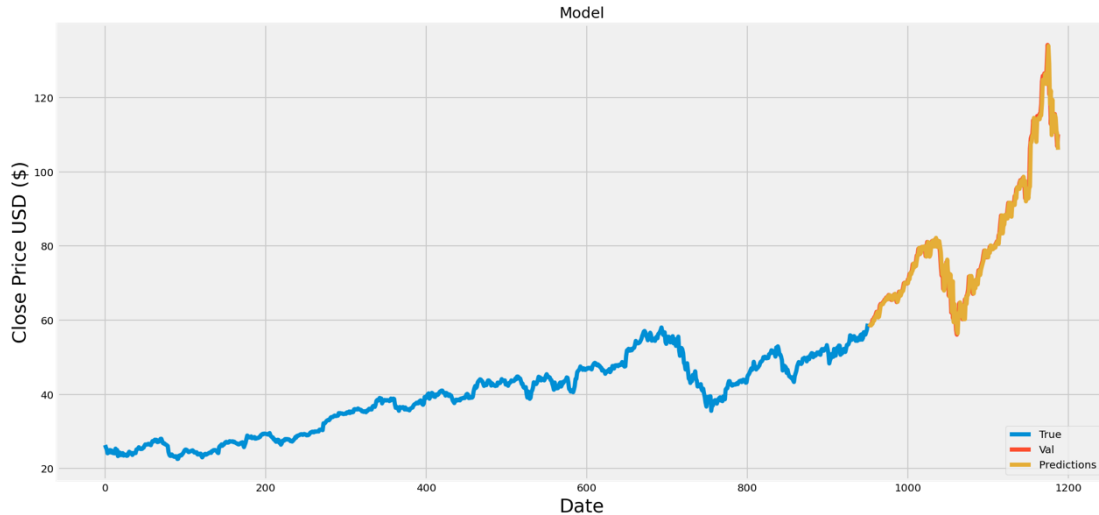
(b) APPLE price prediction using LSTM Model

Figure 2: Apple price prediction using different model

Table 1: The RMSEs of ARIMA and LSTM models.

	ARIMA	LSTM
RMSE	3.261	0.237

5. Conclusion

This paper uses two forecasting machine learning models, ARIMA and LSTM models to predict Apple closing stock prices from 1/4/2016 to 9/21/2020. In order to determine the best machine learning model, RMSE is chosen. Of the two models, the LSTM model has the better performance in Apple stock price prediction because it has smaller RMSE values. With recent advances in the development of techniques based on complex machine learning, especially deep learning algorithms,

the shortcomings of both models have come to the light for instance, both the LSTM model and the ARIMA model essentially exploit relationships that may exist in time series without considering other external factors. The fluctuation of the stock price is not only related to the change of time but also affected by many factors including market factors, political factors, macroeconomic factors, industrial factors, and other external influences of the enterprise, as well as the management ability and organizational structure of the company. Although the LSTM model can introduce indicators like RSI, MSCD to better judge market fluctuations and sudden changes, it is suggested to combine the LSTM model with other models for better prediction.

References

[1] TEAM COUNTERPOINT. (2021, February 18). Apple iPhone Market Share: By Quarter. Counterpoint Research. https://www.counterpointresearch.com/apple-iphone-market-share-quarter/

[2] Apple Reports Third Quarter Results. (2022, July 28). Apple Newsroom. https:// www. apple. com/ newsroom/ 2022/07/ apple-reports-third-quarter-results/

[3] Barro, R. J. (1990). The Stock Market and Investment. Review of Financial Studies, 3(1), 115–131. https:// doi. org/ 10. 1093/rfs/3.1.115

[4] Lee, T. K., Cho, J. H., Kwon, D. S., & Sohn, S. Y. (2019). Global stock market investment strategies based on financial network indicators using machine learning techniques. Expert Systems with Applications, 117(228-242), 228–242. https://doi.org/10.1016/j.eswa.2018.09.005

[5] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). A comparative analysis of forecasting financial time series using arima, lstm, and bilstm. arXiv preprint arXiv:1911.09512.

[6] Qin, J., Tao, Z., Huang, S., & Gupta, G. (2021, March). Stock price forecast based on ARIMA model and BP neural network model. In 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) (pp. 426-430). IEEE.

[7] Hu, Y., Tao, Z., Xing, D., Pan, Z., Zhao, J., & Chen, X. (2020, August). Research on stock returns forecast of the four major banks based on ARMA and GARCH model. In Journal of Physics: Conference Series (Vol. 1616, No. 1, p. 012075). IOP Publishing.

[8] Cao, Z., Ding, X., & Tao, Z. (2022, September). News Detection for Recurrent Neural Network Approach. In 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI) (pp. 794-797). IEEE.

[9] Stevenson, S. (2007). A comparison of the forecasting ability of ARIMA models. Journal of Property Investment & Finance.

[10] Zhu, T., Liao, Y., & Tao, Z. (2022). Predicting Google's Stock Price with LSTM Model. Proceedings of Business and Economic Studies, 5(5), 82-87.