Stock Price Prediction Based on Machine Learning

Ke Huang

Jinan University-University of Birmingham Joint Institute, Jinan University,Guangdong,Guangzhou, 510000,China
kxh118@stu2021.jnu.edu.cn

Abstract. Due to the rapid development of the internet and the financial industry, stock price prediction has become a widely discussed topic. Government departments and regulatory agencies use stock price forecasts to understand the market's health, which helps formulate economic strategies and prevent systemic financial risks. By predicting their stock price fluctuations and those of their competitors, companies can better allocate resources and make decisions for the future. For financial academia and professional analysts, stock price forecasting is an important tool for studying market behavior and economic trends. Based on existing machine learning research, this paper aims to analyze the effectiveness of the Long-Short Term Memory (LSTM) networks and the Autoregressive Integrated Moving Average (ARIMA) model in predicting Apple's stock price comparatively. Additionally, it seeks to summarize the advantages and disadvantages of each model. The experimental results indicate that the LSTM model fits the test set more closely to the actual values, and its Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) values are all relatively smaller. This suggests that the LSTM model exhibits slightly better accuracy and evaluation metrics for long-term predictions than the ARIMA model. Overall, in the research on improving long-term and large-scale prediction results, scholars can focus more on optimizing and enhancing LSTM models in the future compared to ARIMA models. In contrast, improving the accuracy and credibility of LSTM model predictions has more potential for application and academic value.

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

1 Introduction

As machine learning technology advances quickly and the Internet becomes more widely used, individuals are depending more and more on prediction models and data-driven decision-making to direct their work in a variety of fields. Stocks are an integral part of the vast financial market. In the financial market, stock price fluctuations have always been a focal point of widespread attention. Accurate predictions of future stock trends are crucial for investors and market regulators. Stock price prediction involves simulating possible future trends based on available data and available information.

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_11

Predicting stock prices to prevent them from collapsing helps stabilize the financial market and helps investors make better decisions [1,2].

The correct stock price prediction method can be executed at a lower cost and obtain a more accurate prediction of stock prices. However, traditional stock price prediction methods are often limited by information lag, subjective judgment, and model overfitting, resulting in a certain degree of uncertainty and error in their prediction results. The emergence of machine learning methods can address many of the shortcomings associated with traditional forecasting methods. As scholars Yuqing Dai and Yuning Zhang have noted, these methods demonstrate significant advantages over conventional approaches in dynamic markets [3]. For example, machine learning models can continuously improve their performance by learning newly acquired data, an adaptive ability that traditional methods lack. Secondly, machine learning models can quickly process large-scale data, which is more accurate and efficient than traditional processing methods [4]. Furthermore, machine learning models excel at uncovering deep intrinsic relationships within large-scale data. For example, scholars such as Qimiao Qian have conducted a comparative analysis of the Random Forest model and the Gradient Boosted Decision Tree (GBDT) model to predict the stock price of Contemporary Amperex Technology Co. Limited (CATL). They found that the GBDT model provides more accurate predictions and has a broader range of data utilization [5].

Although machine learning models possess numerous well-established algorithms, there is a lack of in-depth analysis and comparative studies on the effectiveness of different algorithmic models. Therefore, this article aims to predict Apple's stock based on machine learning models and explore the performance of different algorithm models in stock price prediction. Firstly, the dataset is preprocessed, and then the preprocessed data is separately input into the Autoregressive Integrated Moving Average (ARIMA) model and the Long Short-Term Memory (LSTM) networks. This paper analyze which algorithm yields better predictive performance by comparing the data predicted by these two models.

2 Data and Methods

2.1 Data source and data preprocessing

This paper selects stock data collected from Kaggle as the dataset, ultimately choosing Apple's stock data from 1980 to 2021 extracted from Yahoo Finance as the dataset. The selected features for data collection include trading date, opening price, closing price, highest price, lowest price, trading volume, and adjusted closing price, resulting in a total of 10,468 data points. As one of the highest market capitalization companies globally, Apple's performance in the stock market is often regarded as a barometer for tech stocks and even the entire market. Its stock price fluctuations can potentially influence the volatility of the entire market. This dataset has the characteristics of a large quantity. Most of the return values conform to the DataFrame type, easy processing, complete data, and no missing values.

Compared to the highest and lowest prices of the stock on a given day, the closing price, which is the price at the end of the trading day, is more reliable. The highest and lowest prices can experience significant fluctuations due to various influencing factors. Therefore, this paper chooses to use the closing price as the feature column for data input, thereby reducing the dimensionality of the input data. To facilitate faster convergence of the data, this paper employs the ‘MinMaxScaler‘ to scale the data to a range between 0 and 1. The date column, 'Date', in the dataset is converted to the 'datetime' type in Pandas and set as the index of the DataFrame. To adjust model parameters and prevent overfitting, the dataset is ultimately divided into a training set and a testing set.

This paper selected a subset of the dataset, specifically the last 100 data periods, for display, as shown in Fig 1. From Fig 1, it can be observed that the stock prices in these previous 100 periods generally exhibit a downward trend. The overall range of fluctuation in trading volume is insignificant, indicating relative stability.

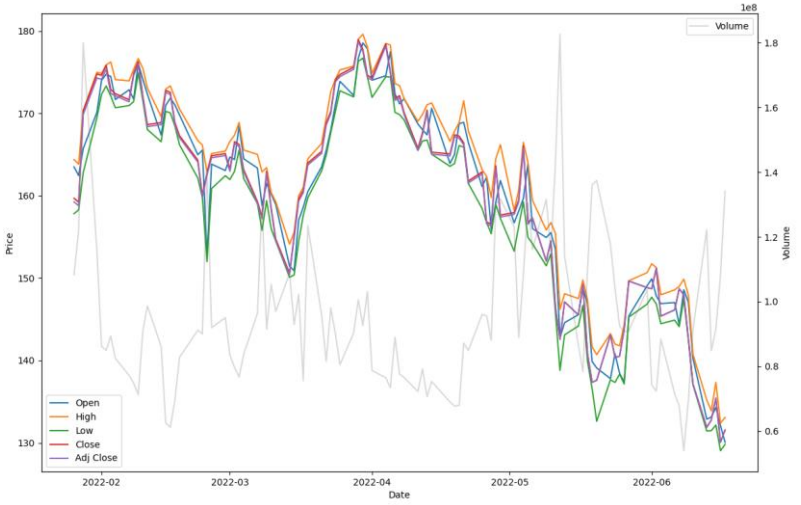

Fig. 1. Stock price and Volume

2.2 Introduction to Models and Algorithms

ARIMA model. The ARIMA model combines the characteristics of the Autoregressive (AR) model, differencing part(I), and Moving Average (MA) model. It is suitable for analyzing data that has a strong temporal association. This subsection mainly explains the concepts and principles of each component and will conclude with the construction of the ARIMA model:

The Autoregressive (AR) model primarily applies a model that utilizes the relationship between the current observation and its own previous observations. In this

model, the main parameter is p. p represents the order of the autoregressive model, which indicates the linear correlation between the current value and the preceding p observations. The mathematical expression is:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t \quad (1)$$

Where c represents the constant term, ϕ_1, ϕ_2, $\cdots$, ϕ_p are the parameters, and ϵ_t represents the error term.

The primary purpose of differencing (I) is to balance the non-stationary parts of the dataset, helping to eliminate trends and seasonality in the data. The key parameter for this part is 'd', which indicates the number of times differencing is needed. The mathematical expression is:

First order:

$$\nabla X_t = X_t - X_{t-1} \quad (2)$$

Second order:

$$\nabla^2 X_t = (\mathrm{X_t} - X_{t-1}) - (\mathrm{X_{t-1}} - X_{t-2}) \quad (3)$$

Accordingly.

The Moving Average (MA) model is a time series model that describes the relationship between the current observation and past observation errors. The primary parameter in the model is 'q', which indicates the linear relationship between the current observation and the past 'q' observation errors. The mathematical expression is:

$$X_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (4)$$

Where μ is the mean of the series，θ_1、θ_2、$\cdots$ θ_q are the parameters，ϵ_t is the error term.

By merging the autoregressive (AR) model, the differencing (I) part, and the moving average (MA) model, the ARIMA(p,d,q) model can be obtained. Generally, the mathematical expression can be constructed as follows:

$$X_t = \Sigma_i^p \phi_i X_{t-i} + \Sigma_j^q \theta_j \epsilon_{t-j} \quad (5)$$

Where ϕ_i represents the coefficient preceding the i-th observation，θ_j represents the coefficient preceding the j-th observation error.

The ARIMA model can adapt to various types of time series by adjusting its parameters (p, d, q), whether they are short-term, long-term, seasonal, or non-seasonal.

LSTM model. The LSTM model is a special Recurrent Neural Network (RNN) type. Compared to a regular RNN, an LSTM can remember input values from much further back in time, making it well-suited for processing long sequence data. Its structure mainly includes three special "gates": the input gate, the forget gate, and the output gate. This subsection will describe in detail the functions and operations of these three gates:

The role of the input gate is to decide which new information needs to be written into the cell state. First, it generates an input gate i_t, and then it generates a candidate cell state $\tilde{C}_t$. The mathematical expressions are as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

Here, σ denotes the Sigmoid function, tanh represents the hyperbolic tangent function, W_i and W_C are the weight parameters, b_i and b_C are the bias parameters, h_{t-1} represents the hidden state from the previous time step, and x_t represents the current input value.

The function of the forget gate is to determine which information needs to be discarded. The principle is that for the hidden state h_{t-1} from the previous time step and the current input x_t, the forget gate will compute a value f_t between 0 and 1. This value will be used to adjust the hidden state from the previous time step h_{t-1}. The mathematical expression is as follows:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \quad (8)$$

Where W_f represents the weight parameter and b_f represents the bias parameter.

After the input information passes through the input gate and the forget gate, the cell state C_t of the LSTM will be updated. The mathematical expression is as follows:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (9)$$

The duty of the output gate is to determine the extent to which the current cell state influences the output. It generates an output gate o_t and computes the hidden state h_t at present. The mathematical expression is:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (11)$$

Where W_o is the weight parameter and b_o is the bias parameter.

These gates control the flow of information, allowing LSTM to retain crucial information in sequential data while forgetting unimportant information.

Evaluation metrics. The text introduces multiple metrics as evaluation indicators for the model, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). This section will respectively explain the roles and principles of each metric.

MSE is calculated by determining the squared differences between each predicted value and its corresponding actual value, and then taking the mean of these squared differences. It measures the deviation between estimated and empirical values. Therefore, a lower MSE indicates closer predictions to the actual values, implying better model performance. The mathematical expression is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (12)$$

Where $\hat{y}_i$ represents the predicted value and y_i represents the actual value.

Comparable to MSE, the RMSE measures the deviation between predicted and actual values. Still, it scales the result by taking the square root of MSE, aligning the metric's unit with the original data. RMSE is more sensitive to large errors compared to MSE. A lower RMSE indicates more minor deviations between predicted and actual values, signifying better predictive performance. The mathematical expression is:

$$RMSE = \sqrt{MSE} \tag{13}$$

MAE is obtained by taking the average of the absolute differences between predicted and true values. It measures the average level of prediction errors and provides a measure of error magnitude that can be directly observed. The mathematical expression is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{14}$$

3 Results

3.1 Model execution results

In this subsection, the running results of the constructed models will be presented separately, as shown in Figs 2 and 3. The real data for both exhibits minimal fluctuation before the year 2000, remaining relatively stable. Between 2000 and 2021, there is a gradual increase followed by a rapid ascent in the actual data for both models. Moving forward in time, there is a noticeable decline in the real data. Regarding the validation of the test set, the LSTM model depicted in Fig 2 closely approximates the real data in both predicted values and trend fluctuations, demonstrating a good predictive capability. Alternatively, observing the ARIMA model in Fig 3, despite the overall trend of the prediction aligns roughly with the accurate data, there is a considerable discrepancy between predicted and actual values, indicating a relatively modest predictive performance.

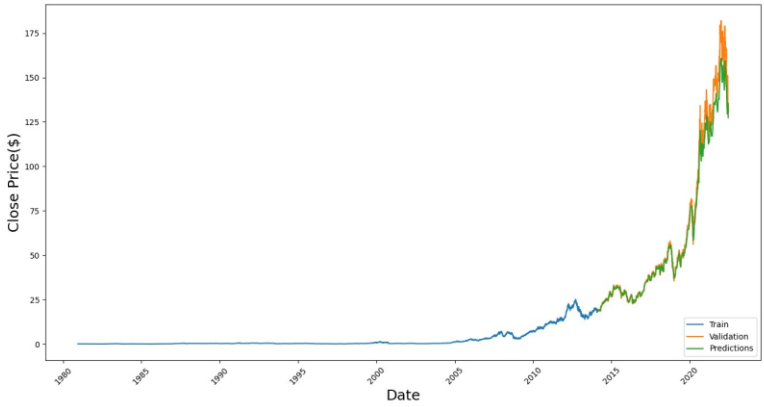

Fig. 2. LSTM model

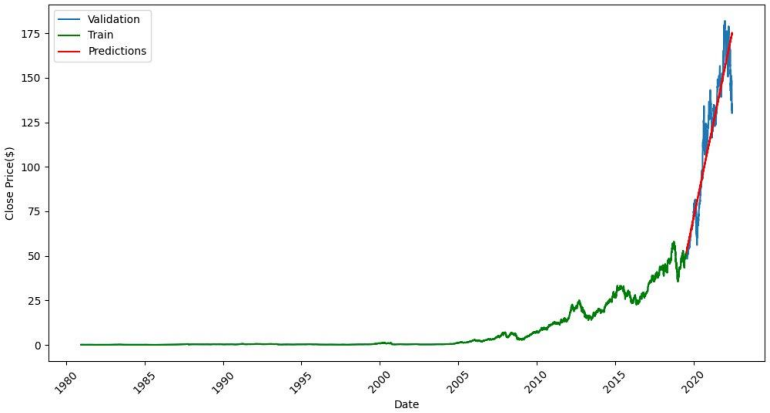

Fig. 3. ARIMA model

Table 1. Model evaluation metrics

Model	***MSE***	***RMSE***	***MAE***
LSTM	36.705924	6.058541	3.404632
ARIMA	140.505721	11.853510	8.880392

For the LSTM model's running results, the model evaluation metrics are shown in Table 1. The MSE is 36.705924, the RMSE is 6.058541, and the MAE is 3.404632. The model exhibits a good fitting effect. Fig 1 shows that the overlap between the yellow part of the real data and the green part of the model-fitted data is high, indicating a good fitting effect of the model.

For the running results of the ARIMA model, as shown in Table 1, the evaluation results indicate that the MSE is 140.505721, the RMSE is 11.853510, and the MAE is 8.880392. Observing Fig 2, it is noticeable that the model's fitting effect is not significant. There is a relatively high resemblance between the real and fitted parts, but the trend of the predicted part appears close to linear.

3.2 Prediction results

This subsection will present the prediction results of the closing prices of stocks for the next 1000 periods using different models, as shown in Figs 4 and 5. In the graphs illustrating the forecast outcomes of the next 1000 closing prices of Apple Inc. stocks by different models, both models exhibit roughly similar trends in the initial portion of the data. However, as time progresses, significant discrepancies appear between the two models' predictions. Fig 4 shows that the LSTM model's prediction results exhibit nonlinear characteristics, with the overall data displaying an approximate left-skewed distribution. The closing price predicted by the LSTM model fluctuates greatly but does not fall below $0, which aligns with the basic logic of stock prices. Conversely, upon

examining Fig 5, the results predicted by the ARIMA model appear to be linear, indicating a continuous decline in the future closing prices of Apple Inc. stocks, falling below $0. This outcome is clearly unreasonable, as actual stock prices cannot plummet to negative values. This discrepancy may arise because the ARIMA statistical model is based on linear assumptions, lacking the built-in linear constraint of non-negativity. Therefore, when dealing with inherently non-negative data, such as stock prices, it is possible to encounter some unrealistic prediction results.

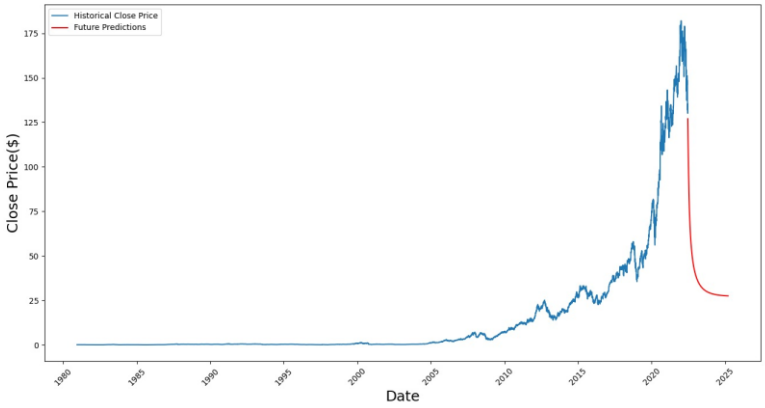

Fig. 4. The prediction results of LSTM model

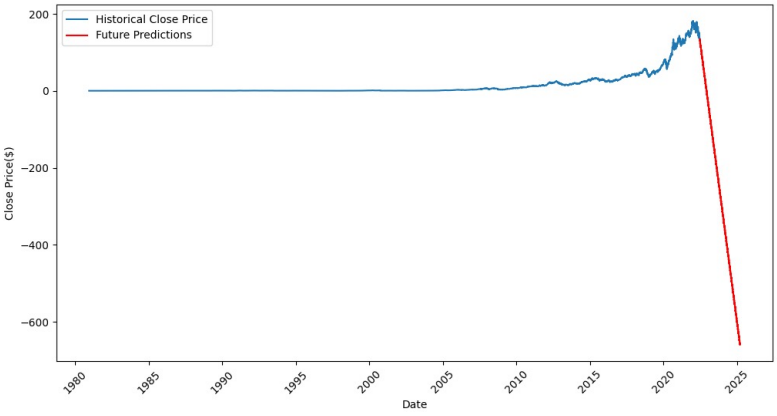

Fig. 5. The prediction results of ARIMA model

4 Suggestions

Although this study achieved predictive results, the findings still have many limitations. Firstly, the dataset used in this study might still be relatively small for machine learning models. For ARIMA models, shorter prediction intervals may yield better results. However, the time period required for the predicted results in this study is relatively long, resulting in less significant predictive and fitting effects. Additionally, the tool in statsmodels for automatically finding optimal ARIMA model parameters (p, d, q) was utilized in constructing the ARIMA model code. This step might differ from manually selecting the optimal parameters, indirectly leading to decreased model fitting effectiveness.

When using the ARIMA model, it is advisable to choose the differencing order for non-stationary data carefully. When selecting model parameters, it is essential to compare the effects of different parameters as much as possible to achieve better model performance. For instance, as noted by scholar Xiaoli Zhao, optimizing ARIMA model parameters based on time granularity can significantly improve the accuracy of model predictions [6]. Another approach to optimize the forecasting performance of the ARIMA model is to introduce new mechanisms. Xuan Chen, Yu Liao, and others have proposed an ARIMA model based on the Time-Varing Filtering Empirical Mode Decomposition (TVF-EMD) method, significantly improving prediction accuracy compared to a single ARIMA model [7]. Zixia Weng utilized the ARIMA model in stock price analysis, using Construction Bank as an example. She pointed out that the ARIMA model can achieve stationary fitting of non-stationary time series through differencing and exhibits high accuracy in short-term forecasting [8]. Therefore, in the future, researchers can focus on how to utilize the ARIMA model to predict continuous long-term periods and obtain a large number of desired results with relatively high accuracy. Despite achieving relatively good results, there is still room for improvement in LSTM models, as there remains a gap between the predicted and actual values. Zhangyuan Zhou and Xiaoling He introduced an attention mechanism to achieve a trading strategy based on the Principal Component Analysis-Attention-LSTM (PCA-Attention-LSTM) model, simplifying the overall network structure while also achieving higher prediction accuracy [9]. In the future, researchers can explore and introduce more mechanisms to optimize LSTM models further. Additionally, machine learning models can be increasingly utilized to handle big data in ecological environments, thereby advancing the construction of environmental governance systems [10].

5 Conclusion

This study uses machine learning methods to analyze and observe the predictive performance of LSTM and ARIMA models in forecasting the closing prices of Apple Inc. stocks for the next 1000 periods. The study concludes that the results of the LSTM and ARIMA models are similar in short-term predictions. However, the LSTM model outperforms the ARIMA model for long-term overall predictions in terms of both

model evaluation metrics and overall prediction results. Despite the good performance of the LSTM model in this study, there are still discrepancies between its predicted results and the actual values. Future improvements could involve introducing or combining existing mechanisms to optimize the performance of the LSTM model and make its predictions more convincing in various scenarios. As for the relatively less effective ARIMA model, future optimization could focus on improving the method for selecting model parameters to forecast relatively short-term results. Similar approaches could be applied to enhance the model's overall performance by introducing or combining new mechanisms, thus improving the credibility of the predictions. In the future, machine learning models could also be further explored for applications beyond stock price prediction, such as handling big data in ecological environments, as mentioned in this study, to better assist in the development of environmental governance systems.

References

1. Peng, Y.C., Ni, X.R., & Shen, J. Enterprise "Deserting the Real Economy for the Virtual Economy" and Financial Market Stability: From the Perspective of Stock Price Crash Risk. Economic Research Journal, 10, 50-66. (2018)
2. Jiang, X.Y., & Xu, N.X. Corporate Overinvestment and Stock Price Crash Risk. Journal of Financial Research, (8), 141-158. (2015)
3. Dai, Y., & Zhang, Y. Machine Learning in Stock Price Trend Forecasting. Stanford University. Retrieved from http://cs229.stanford.edu/proj2013/DaiZhang-MachineLearningInStockPriceTrendForecasting.pdf. Accessed on June 21, 2021. (2013)
4. Hu, M. Comparative Analysis of Software Vulnerability Detection Methods Based on Deep Learning and Traditional Detection Models. Mobile Information, 45(10), 149-151. (2023)
5. Qian, Q.M., Zhang, D., Wang, Y.M., Liu, R., & Cai, F.J. Application of Machine Learning in Stock Price Prediction. China Market, (21), 7-10. doi:10.13939/j.cnki.zgsc.2022.21.007. (2022)
6. Zhao, X.L. Application of ARIMA Model in Short-term Passenger Flow Forecasting of Urban Rail Transit. Modern Urban Rail Transit, (8), 77-82. (2023)
7. Chen, X., Kang, J., Zhang, W.X., Xiang, H.R., Liao, Y., & Liao, M.Y. Short-term Traffic Volume Forecasting Based on TVF-EMD and ARIMA Models. Transportation Technology, 12(3), 188-195. (2023)
8. Weng, Z.X. Stock Price Analysis and Forecasting Based on ARIMA Model: A Case Study of China Construction Bank. Modern Information Technology, 7(14), 137-141. (2023)
9. Zhou, Z.Y., & He, X.L. Stock Price Prediction Method Based on Optimized LSTM Model. Statistics and Decision, (6), 143-148. (2023)
10. Xu, C.Y. Application of Machine Learning in Big Data of Ecological Environment. Chemical Design Communications, 49(8), 177-179. (2023).