Proceedings of Business and Economic Studies, 2022, Volume 5, Issue 5
http://ojs.bbwpublisher.com/index.php/PBES
ISSN Online: 2209-265X
ISSN Print: 2209-2641

Prediction of Amazon's Stock Price Based on ARIMA, XGBoost, and LSTM Models

Zhe Zhu[1], Kexin He[2]*

[1]College of Business and Public Management, Wenzhou-Kean University, Wenzhou 325060, Zhejiang Province, China

[2]College of Faculty of Liberal Arts, Wenzhou-Kean University, Wenzhou 325060, Zhejiang Province, China

****Corresponding author:*** Kexin He, hekex@kean.edu

Abstract: Finding the best model to predict the trend of stock prices is an issue that has always garnered attention, and it is also closely related to investors' investment dynamics. Even the commonly used autoregressive integrated moving average (ARIMA), extreme gradient boosting (XGBoost), and long short-term memory (LSTM) have their own advantages and disadvantages. We use mean squared error (MSE) to judge the most suitable model for predicting Amazon's stock price from many aspects and find that LSTM is the model with the best fitting effect and the closest to the real curve. However, the LSTM model still needs to improve in terms of performance so as to reduce the bias. We anticipate the discovery of more models that are apt for predicting stocks in the future.

Keywords: Amazon; ARIMA; XGBoost; LSTM

Online publication: October 21, 2022

1. Introduction

Stocks are indeed fascinating. According to Asness, "A stock, also known as equity, is a security that represents the ownership of a fraction of the issuing corporation [1]." The ups and downs of stocks reflect the operations of the capital market. However, there is a misconception that the stock market is the capital market. In addition, the rise and fall of the stock market do not equate with the rise and fall of the real economy. This misunderstanding leads to the idea that the capital market can function better when stocks rise and real enterprises can receive sufficient support to develop. In fact, the stock market is only a part of the capital market; that is to say, the role of the stock market cannot be exaggerated or overlooked. Afterall, it is the most active part of the capital market. The economic dynamism of the market is boosted when the most dynamic forces come into play. The sharp rise in the stock market proves that the status of the capital market in the financial market is constantly rising.

Besides, the stock market is in a constant change on a daily basis, and the ups and downs of stocks affect the investors. The answer to which why people choose to invest in stocks is as follows: "Stocks offer investors the greatest potential for growth (capital appreciation) over the long haul [2]." It is precisely because of the high rate of return that investors increase their investment in stocks. There are different benefits when investing in different stocks, and different investment groups have different definitions of stock returns. Only by being careful in each step of the investment can the investor expect a high return.

Analytically speaking, Amazon's distinct position in the international market is quite representative. The ever-increasing penetration rate of the internet has resulted in an ever-increasing demand for e-commerce. As an early developing e-commerce company, Amazon occupies a dominant position in the global e-commerce market. Amazon has achieved a monopoly in the industry, and it is often the choice of companies that wish to publish and sell products online. Amazon has received many positive and negative reviews. Some people think that Amazon is a great place to shop online; it has everything one could possibly need; the price is reasonable, and there is quality to the items sold; Amazon is the perfect place to shop online! However, some people think that Amazon has poor ethics in the sense that the goods received do not conform to the advertised ones. The different experiences of shoppers and the comments made on public platforms can affect a company's image, which in turn affects the company's stock.

We first use the most non-complex Fourier transform and average autoregression to observe the basic information of Amazon's stock, followed by autoregressive integrated moving average (ARIMA), extreme gradient boosting (XGBoost), and long short-term memory (LSTM) to estimate the stock. The objective is to determine the best model for predicting Amazon's stock price. We discuss our point of view in two parts: the first part is a review of literature, in which we find the best model by reviewing some relevant information and looking for models that study the stock market for comparison; the second part focuses on describing the experimental results through continuous analysis and comparison of data, concluding and indicating the expectations of the future stock market.

2. Literature review

In order to find the most suitable model to predict Amazon's stock price, we use moving average (MA) and Fourier transform to observe, followed by ARIMA, LSTM, and XGBoost to predict. In 1998, Ho and Xie [3] studied the approach to repairable system reliability forecasting based on ARIMA. They concluded that the ARIMA model is an alternative that provides satisfactory results for forecasting. In 2019, there was a global outbreak of the coronavirus (COVID-19); Benvenuto *et al.* [4] used a simple econometric model – ARIMA – to forecast the spread of COVID-19 and the trend of COVID-19's incidence and prevalence. Zhang [5] proposed an approach that combined both ARIMA and artificial neural network (ANN) models to improve data accuracy, taking advantage of the unique strength of ARIMA and ANN models in linear and nonlinear modeling. Contreras [6] used ARIMA to predict the next-day electricity prices for maximum benefits. Ariyo [7] presented the extensive process of the stock price predictive model by using the ARIMA model to analyze the New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE); the results showed that the ARIMA model has a strong potential for short-term forecast. Smagulova [8] investigated an emerging topic that uses a memristor circuit to achieve the hardware acceleration of LSTM because LSTM is a recurrent neural network with a state memory and multilayer cell structure. Gers [9] focused on LSTM because it has a shorter time lag than RNNs on tasks, making it an accurate measurement or generation of time intervals. Graves [10] studied the TIMIT speech corpus with bidirectional and unidirectional LSTM networks and found that LSTM outperforms RNNs. Chen [11] forecasted China's stock returns using the LSTM model, improving its accuracy from 14.3% to 27.2%, and reflecting the potential of the LSTM model for stock price prediction. Fu [12] used the LSTM and gated recurrent units (GRU) neural network (NN) method to forecast real-time traffic flow and control traffic. As for the XGBoost model, Liao [13] constructed a dynamic weighting multi-factor stock selection strategy established on the XGBoost model. He used the XGboost machine learning method to forecast the information coefficient (IC) of components, revealing that the XGBoost model is useful for IC coefficient prediction. Basak [14] used two algorithms, random forests and gradient boosted decision trees, to facilitate the connection of the concerning – whether the stock price will change with respect to the existing price n days earlier. Kumar [15] used data from Yahoo Finance and several algorithms based on seasonal (S)ARIMA and XGBoost to estimate the value. Gumelar [16] experimented with about 25 companies to forecast the closing price by using two highly accurate analyses, LSTM and XGBoost, and found that XGBoost has a 99% prediction accuracy. In this study, we use Fourier transform and MA to observe the basic information of Amazon, followed by ARIMA, LSTM, and XGBoost to estimate the stock price, aiming to determine the most suitable model for predicting Amazon's stock price.

3. Method

3.1. Autoregressive integrated moving average

The ARIMA model was introduced by Box and Jenkins in 1970 [7]. It is one of the most popular methods used for prediction. Also known as the Box-Jenkins method, the ARIMA model consists of a set of activities for identifying, estimating, and diagnosing ARIMA models using time series data. The ARIMA model has shown to be effective for accurate short-term prediction. Its performance in short-term prediction outperforms complex structural models. In the ARIMA model, the value of a variable in the future is a linear combination of past values and past errors, expressed as follows:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{1}$$

where Y_t is the actual value and ε_t is the random error at t; ϕ_i and θ_j are the coefficients, while p and q are integers that are often referred to as autoregressive and moving average, respectively.

3.2. Long short-term memory

At time t, x_t is the input data of the LSTM cell, h_{t-1} is the output of the LSTM cell at the previous moment, c_t is the value of the memory cell, and h_t is the output of the LSTM cell. The calculation process of the LSTM unit can be divided into several steps [17].

(1) First, calculate the value of the candidate memory cell $\tilde{c}_t$; W_c is the weight matrix, and b_c is the bias.

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2}$$

(2) Calculate the value of the input gate i_t; the input gate controls the update of the current input data to the state value of the memory cell, σ is the sigmoid function, W_i is the weight matrix, and b_i is the bias.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i \tag{3}$$

(3) Calculate the value of the forget gate f_t; the forget gate controls the update of the historical data to the state value of the memory cell, W_f is the weight matrix, and b_f is the bias.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{4}$$

(4) Calculate the value of the current moment memory cell c_t; c_{t-1} is the state value of the last LSTM unit.

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{5}$$

where $*$ represents the dot product. The update of memory cell depends on the state value of the last cell and the candidate cell, and it is controlled by the input gate and forget gate.

(5) Calculate the value of the output gate o_t; the output gate controls the output of the state value of the memory cell, W_o is the weight matrix, and b_o is the bias.

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

(6) Finally, calculate the output of the LSTM unit h_t.

$$h_t = o_t * \tanh c_t \tag{7}$$

Benefitting from the three control gates and memory cell, LSTM store, read, reset, and update long time information easily. It is important to note that the dimensions of the output can be controlled by setting the dimensions of the weight matrix due to the sharing mechanism of the LSTM internal parameters. LSTM establishes a long time-delay between input and feedback. The gradient will neither explode nor disappear because the internal state of the memory cell in this architecture maintains a continuous error flow [18].

3.3. Extreme gradient boosting

The XGBoost algorithm is based on gradient boosting decision tree (GBDT) [19]. Compared with GBDT, the advantage of XGBoost is that it supports linear classifiers and performs Taylor expansion for the cost function by introducing a second derivative to ensure more accurate results. There principles of XGBoost are discussed below.

The XGBoost model uses additive training method to optimize the objective function, which means the optimization process of the latter step relies on the result of its previous step. The t-th objective function of the model can be expressed as follows:

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) + constant \tag{8}$$

where l represents the loss term of the t-th round, $constant$ represents a constant term, and Ω is the regularization term of the model, shown as follows:

$$\Omega(f_t) = \Upsilon \cdot T_t + \lambda \frac{1}{2} \sum_{j=1}^{T} w_j^2 \tag{9}$$

where both Υ and λ are customization parameters. Generally, the larger these two values are, the simpler the structure of the tree is, and the overfitting problem can be solved effectively. Performing a second-order Taylor expansion on (8), the process is given by

$$obj^{(t)} = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + constant \tag{10}$$

where g is the first derivative, and h is the second derivative. They can be described as follows:

$$g_i = \partial_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{t-1}) \tag{11}$$

$$h_i = \partial^2_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{t-1}) \tag{12}$$

Substitute (9), (11), and (12) into (10), and take the derivative. Then, solutions can be obtained from (13) and (14) as follows:

$$w_j^* = -\frac{\sum g_i}{\sum h_i + \lambda} \tag{13}$$

$$obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum g_i)^2}{\sum h_i + \lambda} + \gamma \cdot T \tag{14}$$

where obj^* represents the score of loss function; the smaller the score, the better the structure of the tree. w_j^* refers to the solution of weights.

4. Experiment and analysis

With the goal of finding the best model for predicting Amazon's stock price, MA and Fourier transform were first applied to observe the basic information of Amazon.

We extracted yearly historical Amazon stock price from 2010 to 2019 from Yahoo Finance, including its open price, high price, low price, closing price, *etc*. Then, we selected its closing price and drew a broken line chart to observe its rising and falling trends. In **Figure 1**, we can see that the closing price of Amazon's stock remained at about 250 USD and gradually rose with small fluctuations from 2010 to 2013. Amazon's "online retail" has achieved success at that time, and its profit growth and valuation have continued to improve since then. Since 2014, its closing price has seen a rapid increase, reaching nearly 1,000 USD in 2018, but with significant fluctuations. In less than a year from 2018, the closing price of Amazon's stock jumped from less than 1,000 USD to about 2,000 USD; it then fell sharply to about 1,500 USD in 2019 after reaching a peak. Amazon is the largest online e-commerce company in the United States. With the increase in demand for online shopping in the United States and the good e-commerce shopping experience provided by Amazon, the development of its e-commerce scale still maintains a relatively high growth. Additionally, its business scale continues to expand; this is one of the main reasons for the rise of its stock price. Taking April 20, 2016, as the boundary, the data is divided into two subsets: training and test datasets (**Figure 1**).

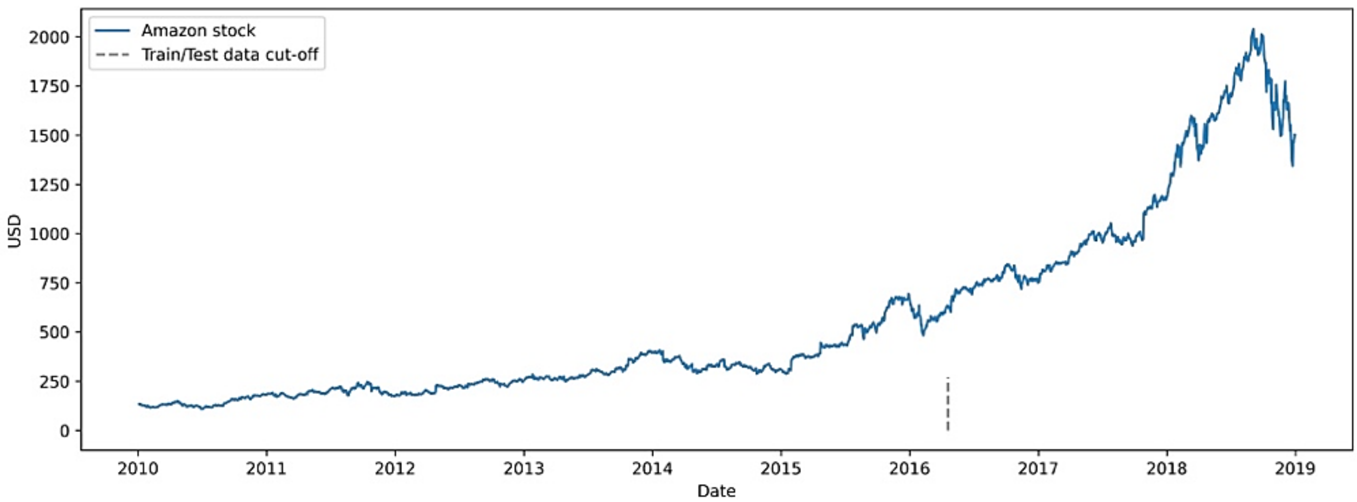

Figure 1. Amazon's stock price

MA7, MA21, Closing Price, Upper Band, and Lower Band are used as technical indicators for stock forecasting from 1850 to 2250, as shown in **Figure 2**. We set two moving averages with different calculation days, 7 days and 21 days, to understand the overall operating trend of the stock price from different periods. From **Figure 2**, we can see that the trend of MA7 and MA21 is consistent with the Closing Price, staggered with it many times, and fluctuates between the Upper Band and Lower Band. From 1970 to around 2200, the Closing Price, MA7, and MA21 increased significantly, but decreased after reaching a peak in 2200.

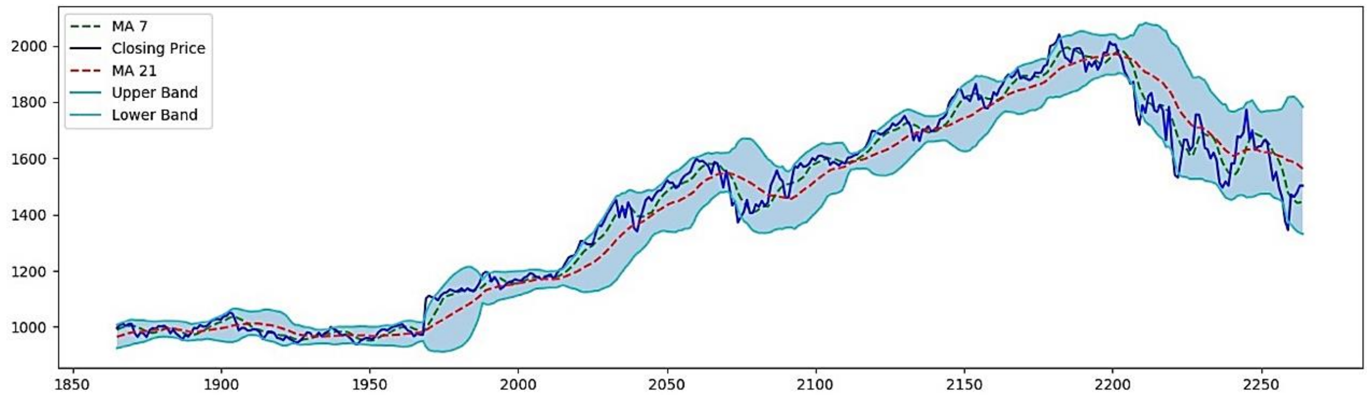

Figure 2. Technical indicators for Amazon – last 400 days

Using momentum and moving average convergence/divergence (MACD), we measured the rate of change of Amazon's stock price during this period and reflected it directly in **Figure 3**. Two green horizontal dotted lines are marked in **Figure 3** at positions 15 and -15. It can be seen that the trend of Momentum is stable, fluctuating up and down around the green horizontal dotted line at 15. Momentum can be seen to be slightly lower than the green horizontal dotted line at 15 between 1850–2050. It then overlaps with the green horizontal dotted line at 15 between 2050. After 2110, Momentum can be seen to be slightly above the green horizontal dotted line at 15 before overlapping the line again around 2210. It can be seen that MACD (green dash-dotted line) fluctuates greatly between the green horizontal dotted line at 15 and -15 in 1850–1970, with a relatively stable trend. The majority of MACD are above the green horizontal dotted line at 15 in 1970–2200, resulting in several rises with large fluctuations, but there is a lower trend from 2200–2260.

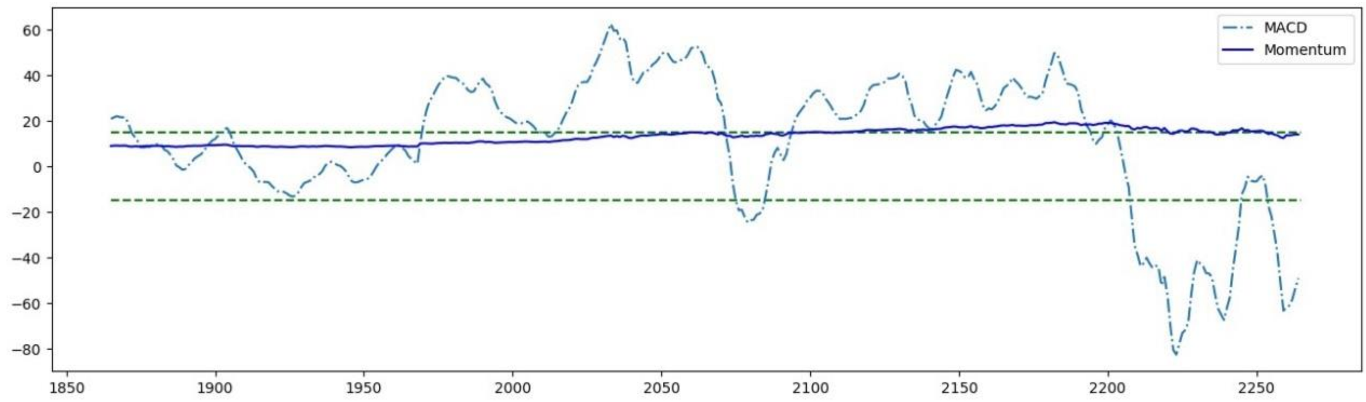

Figure 3. MACD

Historical stock price is a large number of noisy data. Stock price changes on a daily basis, and not every indicator can be used to predict the stock market. Therefore, we use Fourier transform to remove the noise in the original data. In order to see the denoising effect from an intuitive perspective, we collected 2,000 days of Amazon's daily closing price data, as shown by the purple curve in **Figure 4**.

We then use Fourier transform with 3, 6, 9, and 100 components and compare the curve with the real price, as shown in **Figure 4**. With the increase of components, we can see that the curve, after noise filtering, is more consistent with the real data trend, such as the Fourier transform with 100 components represented by the red curve.

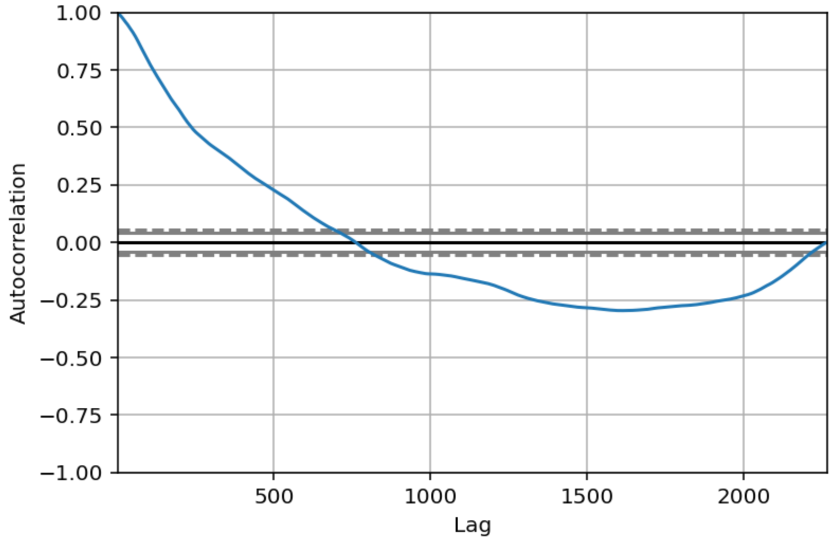

Figure 4. Amazon's (closing) stock price and Fourier transforms

In the experiment, we use three prediction methods: ARIMA, XGBoost, and LSTM. By comparing the MSE obtained, we can judge the accuracy of the model in predicting Amazon's stock price.

We match ARIMA to the entire dataset and check the residuals. Built in Panda, according to its autocorrelation diagram, as shown in **Figure 5**, we can see a positive correlation with the first 500–700 lags and a negative correlation with the first 700–2000 lags. Therefore, a good starting point for the model parameter AR can be 5. Here, ARIMA(5,1,0) model is applied. This sets the lag value to 5 for AR, utilizes a difference order of 1 to make the time series stationary, and uses a MA model of 0. When fitting the model, it provides a lot of debugging information about the fitting of the linear regression model. Turn it off by setting the DISP parameter to 0.

Briefly, we fit the ARIMA(5, 1, 0) model, with conditional sum-of-squares (CSS) as the fitting method, and Akaike information criterion (AIC) and Bayesian information criterion (BIC) as the model parameters.

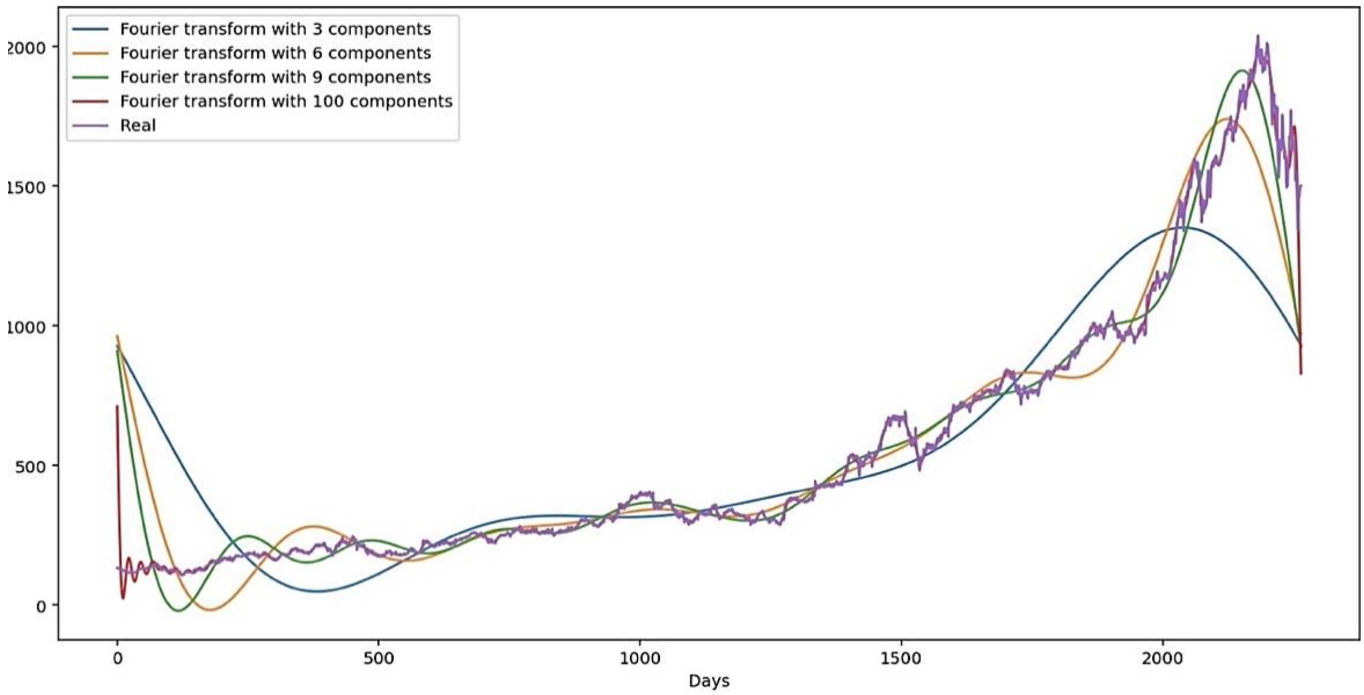

Figure 5. Autocorrelation of Amazon's closing stock price

The algorithm forecasts the expected value (yhat), adds yhat to the prediction data structure, and then adds the actual value to the test set for model refining and re-fitting. Eventually, having built the prediction and history data structures, the final output test MSE value is 557.865. Compared with the actual price, the predicted line chart of Amazon's stock price is drawn, as shown in **Figure 6**. We can see that the predicted value of the model is close to the actual value, and the model fitting effect is good.

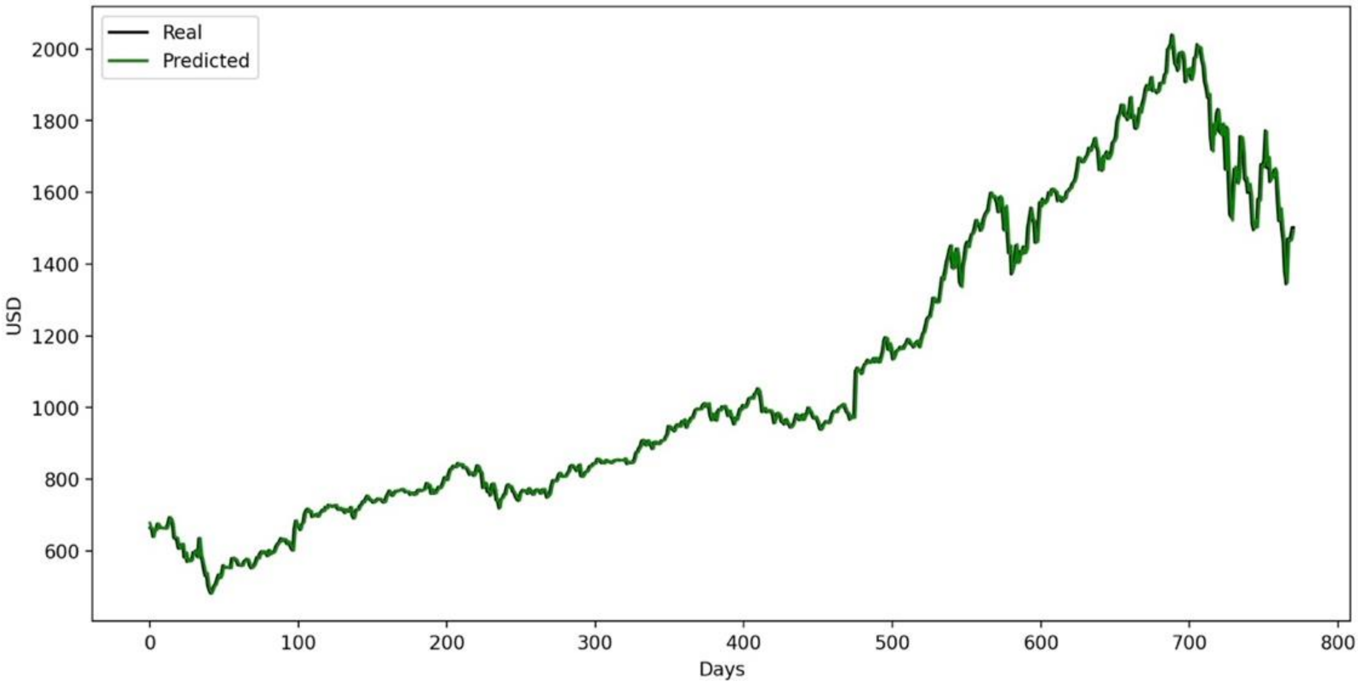

Figure 6. ARIMA model on Amazon's stock

In order to evaluate the model performance accurately, the data is divided into a training set and a validation set. By calculating the training set and the validation set, the data respective are obtained. The XGBoost parameters are as follows: gamma = 0.0, n_estimators = 200, base_score = 0.7, colsample_bytree = 1, and learning _ rate = 0.05. We then obtain the root mean squared error (RMSE) of the training set and validation set, respectively. With the increasing number of iterations, the RMSE of the training set gradually approaches 0, while the RMSE of the validation set gradually approaches 600.

We applied a stacked LSTM model. The loss of the model training process (on the training set) is continuously optimized as the training process advances. The initial optimization and loss decrease rapidly, and then gradually stabilize. From the running results of the code, the MSE of the training set reduced to 0.0524, while that of the verification set reduced to 0.0514, as shown in **Table 1**.

Table 1. MSE in LSTM model

Epochs	MSE training/validation
10	0.0614 / 0.0598
30	0.0591 / 0.0586
60	0.0569 / 0.0560
80	0.0546 / 0.0533
97	0.0524 / 0.0523
99	0.0529 / 0.0514

By observing the MSE, we can see the performance of the LSTM model. Compared with the training set, the effect on the test set is slightly biased, but the overall trend is still well predicted.

Three prediction methods are used in total, and the MSE obtained is summarized in **Table 2**. It is evident that MSE is much smaller than ARIMA and XGBoost when the LSTM model is used to predict Amazon's stock price. It is clear that the LSTM model has obvious advantages in meeting our demand for forecasting Amazon's stock price.

Table 2. MSE of three models

Model	MSE
ARIMA	557.865
XGBoost	360,000
LSTM	0.0514

5. Conclusion

By looking at Amazon's stock using a simple data model and comparing the three models, the following conclusions can be drawn.

First of all, the stock trends and closing price curves of MA7 and MA21 learned in two-calculation days are basically the same. Secondly, using the stock change rate observed by Momentum and MACD, it is found that Momentum tends to be stable but MACD has several large fluctuations. Third, after denoising with Fourier transform, with the increase of components, the denoised curve is more consistent with the curve of the real data. By comparing the three models – ARIMA, XGBoost, and LSTM – the MSE is obtained to judge the stock price accuracy. The final comparison shows that LSTM is the best model for predicting Amazon's stock price. In order to evaluate the MSE model performance, we split the data into a training set and a validation set. When using the LSTM model to predict Amazon's stock price, MSE is much smaller than ARIMA and XGBoost.

When looking at the performance of the LSTM model, although the impact on the test set is slightly skewed compared to the training set, the overall trend is still very predictable. In the future, this model can be further studied, which would be a good research direction for predicting stock prices and understanding the stock market.

Disclosure statement

The authors declare no conflict of interest.

Author contributions

Z.Z. conceived the idea of the study, and K.H. performed the experiments.

References

[1] Asness CS, Frazzini A, Pedersen LH, 2012, Leverage Aversion and Risk Parity. Financial Analysts Journal, 68(1): 47–59.

[2] Qin J, Tao Z, Huang S, et al., 2021, Stock Price Forecast Based on ARIMA Model and BP Neural Network Model. 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), IEEE, 426–430.

[3] Ho SL, Xie M, 1998, The Use of ARIMA Models for Reliability Forecasting and Analysis. Computers &

Industrial Engineering, 35(1–2): 213–216.

[4] Benvenuto D, Giovanetti M, Vassallo L, et al., 2020, Application of the ARIMA Model on the COVID-2019 Epidemic Dataset. Data in Brief, 29: 105340.

[5] Zhang GP, 2003, Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. Neurocomputing, 50: 159–175.

[6] Contreras J, Espinola R, Nogales FJ, et al., 2003, ARIMA Models to Predict Next-Day Electricity Prices. IEEE Transactions on Power Systems, 18(3): 1014–1020.

[7] Ariyo AA, Adewumi AO, Ayo CK, 2014, Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, IEEE, 106–112.

[8] Smagulova K, James AP, 2019, A Survey on LSTM Memristive Neural Network Architectures and Applications. The European Physical Journal Special Topics, 228(10): 2313–2324.

[9] Gers FA, Schraudolph NN, Schmidhuber J, 2002, Learning Precise Timing with LSTM Recurrent Networks. Journal of Machine Learning Research, 3: 115–143.

[10] Graves A, Fernandez S, Schmidhuber J, 2005, Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. International Conference on Artificial Neural Networks, Springer, 799–804.

[11] Chen K, Zhou Y, Dai F, 2015, A LSTM-Based Method for Stock Returns Prediction: A Case Study of China Stock Market. 2015 IEEE International Conference on Big Data (Big Data), IEEE, 2823–2824.

[12] Fu R, Zhang Z, Li L, 2016, Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction. 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, 324–328.

[13] Liao J, Zhang R, 2018, Dynamic Weighting Multi Factor Stock Selection Strategy Based on XGBoost Machine Learning Algorithm. 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), IEEE, 868–872.

[14] Basak S, Kar S, Saha S, et al., 2019, Predicting the Direction of Stock Market Prices Using Tree-Based Classifiers. The North American Journal of Economics and Finance, 47: 552–567.

[15] Kumar DS, Thiruvarangan BC, Vishnu A, et al., 2022, Analysis and Prediction of Stock Price Using Hybridization of SARIMA and XGBoost. 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), IEEE, 1–4.

[16] Gumelar AB, Setyorini H, Adi DP, et al., 2020, Boosting the Accuracy of Stock Market Prediction using XGBoost and Long Short-Tergumem Memory. 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), IEEE, 609–613.

[17] Cao J, Li Z, Li J, 2019, Financial Time Series Forecasting Model Based on CEEMDAN and LSTM. Physica A: Statistical Mechanics and Its Applications, 519: 127–139.

[18] Jiang H, He Z, Ye G, et al., 2020, Network Intrusion Detection Based on PSO-XGBoost Model. IEEE Access, 8: 58392–58401.

[19] Ma X, Sha J, Wang D, et al., 2018, Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGBoost Algorithms According to Different High Dimensional Data Cleaning. Electron Commerce Res Appl, 31: 24–39.