WEP Warwick Evans Publishing

Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 14
AFTEM 2024

Application of Stock Price Prediction during the "Double 11" Period Based on the LSTM Model

Haozhe Wu [1, *, #], Zhenshen Pan [2, #], Jiaying Zhai [3, #]

[1] School of Mathematics and Statistics, Wuhan University, Wuhan, China, 430072

[2] School of Management, Guangdong University of Science and Technology, Dongguan, China, 523083

[3] School of International Economics, China Foreign Affairs University, Beijing, China, 102206

* Corresponding author: 2021312011099@whu.edu.cn

[#] These authors contributed equally.

Abstract. Studying stock price changes during special periods such as the "Double 11" shopping festival using the Internet E-commerce Industry Index can aid investors and e-commerce companies in making informed decisions. This paper focuses on stock price fluctuations during the "Double 11" period, using the "Internet E-commerce" Industry Index from January 20, 2016, to August 16, 2024, as the research object. The study employs ARIMA, LR, LGBM, and LSTM models to predict stock prices in the "Internet E-commerce" sector and uses RMSE, MAE, and R^2 as evaluation metrics to compare and assess model performance. The results confirm that the LSTM model demonstrates superior effectiveness and accuracy in predicting nonlinear stock price fluctuations, especially during special periods like "Double 11," compared to other models. Unlike traditional models, LSTM excels in capturing long-term dependencies and complex patterns in time series data. Its ability to adapt to nonlinear relationships and temporal variations makes it a more robust tool for stock price forecasting, particularly in volatile periods such as "Double 11".

Keywords: LSTM Model; Stock Price Prediction; "Double 11" Shopping Festival; Time Series Analysis; Neural Network.

1. Introduction

In today's financial markets, stock price fluctuations have become a focal point for investors and researchers. With the rapid development of the e-commerce industry, "Double 11" has evolved from the largest commercial event in China's internet sector into a globally recognized shopping festival, with an increasingly significant impact on stock prices. However, as financial markets become more complex, stock price prediction has also become increasingly challenging, particularly the fluctuations in stock prices during the "Double 11" period, which pose significant challenges to accurate forecasting.

Stock price prediction has long been a crucial issue in academic research. However, in the current field, there is relatively little research focusing on stock price prediction under special circumstances like "Double 11." Events related to e-commerce, such as "Double 11," can significantly stimulate consumer purchasing behavior, potentially having a substantial impact on the prices of related stocks. Due to the irregular volatility, uncertainty, and complexity inherent in stock price data, accurate prediction remains challenging. While many methods for predicting financial data have been developed, providing a solid foundation for stock price prediction, they still struggle to capture the high level of uncertainty and complexity in the market accurately. Traditional models, such as ARIMA and Linear Regression, often fall short when dealing with nonlinear patterns and long-term dependencies inherent in stock market data. In contrast, Long Short-Term Memory (LSTM) models are specifically designed to address these challenges by capturing both short-term fluctuations and long-term trends in time series data. LSTM's unique architecture, including its memory cell and gating mechanisms, allows it to effectively handle complex temporal dependencies, making it particularly well-suited for forecasting during highly volatile periods like "Double 11".

This study aims to explore stock price prediction during "Double 11" using the Long Short-Term Memory (LSTM) model. By collecting and processing relevant data, the research evaluates the effectiveness and accuracy of the LSTM model in stock price forecasting. This approach not only highlights the LSTM model's strengths but also introduces an innovative method for addressing the challenges of stock price forecasting during periods of high volatility.This approach can help investors seize opportunities, mitigate risks, and make it easier for business decision-makers to realign corporate strategies and develop correct financial plans.

2. Literature Review

Starting from the influence of emotions on stock prices, XiaoZhe Gong proposed a stock price prediction model based on the Autoregressive Integrated Moving Average (ARIMA) model and Backpropagation Neural Network (BPNN), combined with sentiment analysis of news tendencies. This model demonstrated good effectiveness in stock prediction through comparative analysis [1]. Similarly, Wang Yuhan and Zhang Yameng proposed a stock price prediction model based on BP neural networks and High-Order Fuzzy Cognitive Maps, which also contributed significantly to the field of stock price prediction. However, BPNN as an underlying model still has certain deficiencies in stock price forecasting [2].

Peipei Wang et al. introduced a multi-feature fusion model for stock price prediction that combines the ARIMA model with the Dual Features Attention-Over-Attention-BERT (DFAOA-BERT) model. While ARIMA still has limitations in predicting irregular stock movements, this method provided new insights into stock price prediction [3]. In contrast, Jiajian Zheng et al. proposed a U.S. stock market trend prediction method based on a Random Forest model, which exhibited higher accuracy when the prediction window exceeded 60 days, although its performance in short-term (e.g., 30 days) prediction was relatively weak [4].

In response, Suman Saha et al. innovatively applied graph-based methods to stock market analysis and prediction. Although graph-based models are not yet mature and have significant limitations regarding computational performance, they have also effectively advanced the field of stock price prediction [5]. Some researchers have sought to identify more effective prediction methods for irregular data through performance comparisons. For example, G. Wang and Chuanjun Zhao found that LSTM outperforms traditional models such as ARIMA, General Regression Neural Network (GRNN), Exponential Smoothing (ES), Autoregressive Moving Average (ARMA), Bayesian Networks, Wavelet Transform, Support Vector Machines, and ordinary Recurrent Neural Networks in handling long-term dependencies and complex patterns in time series data [6] [7].

Houjian Li used the LSTM model to predict China's carbon prices and found that multivariate LSTM models have higher accuracy and stability in carbon price prediction through comparisons with different models [8]. Ya Li and Zhanguo Wei further validated the accuracy of this model in predicting regional logistics demand [9]. Building on this, Jian Chen et al. demonstrated the model's efficiency in rapid urban flood risk prediction [10]. These studies provide a feasible approach for using LSTM in predicting stock prices of the internet e-commerce sector during the "Double 11" shopping festival.

In our work, considering the significant impact of the "Double 11" shopping festival on the internet e-commerce sector, this sector was chosen as the representative for analysis. Historical stock price data of the internet e-commerce sector were analyzed using the ARIMA, Linear Regression (LR), Light Gradient Boosting Machine (LGBM), and Long Short-Term Memory (LSTM) models. The performance of these models was evaluated by comparing the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

The results indicated that the LSTM model exhibited superior performance in predicting stock prices of the internet e-commerce sector during the "Double 11" period. Consequently, the LSTM model was determined to be the optimal choice for forecasting future stock price trends.

3. Model Construction

3.1. Overview of the LSTM Model

LSTM (Long Short-Term Memory) was first introduced by Hochreiter and Schmidhuber in 1997, primarily to address the issue of long-term dependencies in the processing of long time series data. Traditional Recurrent Neural Networks (RNNs) often encounter problems such as vanishing or exploding gradients when dealing with long time series, making it difficult to effectively learn features that rely on long-term dependencies. The structural diagram is shown in Figure 1.

In time series data prediction, the LSTM model introduces three gating units to control the flow of data, namely the forget gate, input gate, and output gate. Additionally, a memory cell is introduced to store information. The functions of these three gating units are implemented through the sigmoid function and element-wise multiplication. These gating units do not provide additional information; rather, they are responsible for selecting the amount of information to pass through, thereby determining how information is retained, updated, and output at each time step.

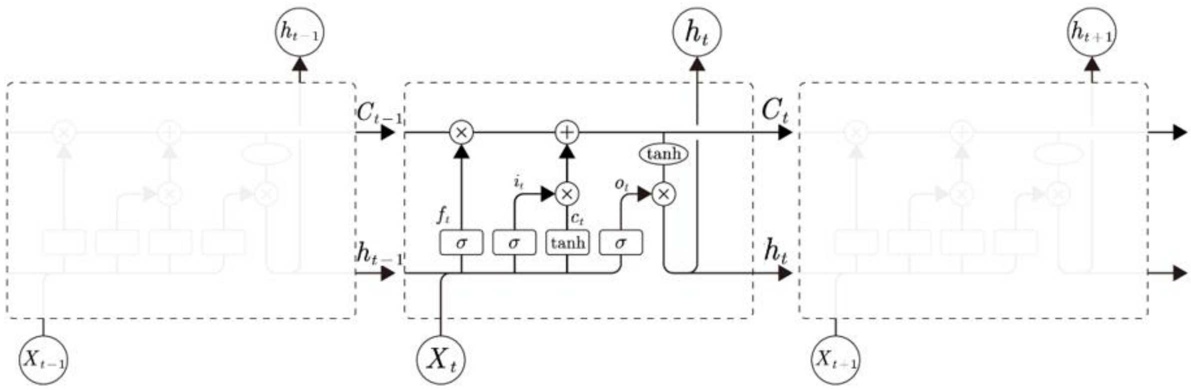

Figure 1. LSTM Structural Diagram

The execution process can be summarized as follows:

Firstly, the forget gate determines which information from the previous time step's memory cell should be forgotten by taking as input the hidden state from the previous time step and the sequence data from the current time step.

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

In this formula, W_f represents the weight matrix for the forget gate, b_f is the bias term, h_{t-1} is the hidden state from the previous time step, x_t is the input at the current time step, and σ denotes the sigmoid activation function.

Next, the input gate determines how the current time step's input X_t and the previous time step's hidden state h_{t-1} influence the new memory cell C_t. Meanwhile, the candidate memory cell $\tilde{C}_t$ is generated through the tanh function, which represents the new information that the current input can provide. The LSTM updates the memory cell C_t by combining the effects of the forget gate and the input gate.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (4)$$

In these formulas, W_i represents the weight matrix of the input gate, b_i denotes the bias term, and tanh is the hyperbolic tangent activation function.

Finally, the output gate controls the output of the current time step, which is determined by the output gate and the value of the current memory cell C_t, and the output is adjusted by applying the tanh function to the memory cell.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

In these formulas, W_o represents the weight matrix for the output gate, b_o is the bias term, C_t is the updated memory cell.

3.2. Model Architecture Design

To effectively capture the temporal characteristics of the A-share e-commerce index, this study employs the Long Short-Term Memory (LSTM) network as the core predictive tool. The LSTM model, with its unique gating structures and memory cells, performs exceptionally well in time series forecasting. Based on this, our process design is illustrated in Figure 2.

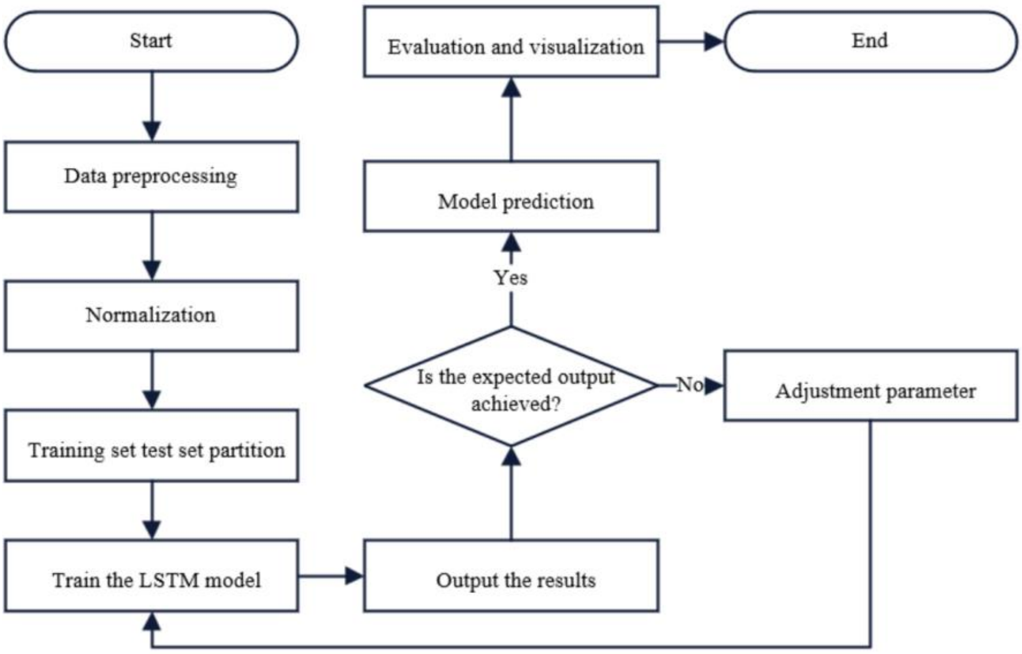

Figure 2. Model Architecture Flowchart

During the data processing stage, the raw data is initially cleaned to eliminate erroneous data and irrelevant information. Additionally, interpolation methods are employed to address any missing data, ensuring data completeness and consistency. Subsequently, the data is normalized to remove the influence of differing magnitudes, thereby providing high-quality input for subsequent model training and prediction.

To optimize the LSTM hyperparameters, we employed a grid search approach, evaluating a range of configurations including LSTM units (100, 200, 300), dropout rates (0.1, 0.2, 0.3), and learning rates (0.001, 0.01). Each configuration was trained on the dataset, and performance was monitored using validation loss. The ultimate goal was to minimize validation loss while maintaining generalization.The final choice of hyperparameters was based on the best balance between model performance and generalization.

During the model training stage, the dataset was divided into training and testing sets to ensure that the model performs well on unseen data. For network initialization, we selected LSTM layers with 200 units and incorporated appropriate Dropout layers to prevent overfitting. The model was optimized using the RMSprop optimizer with a learning rate set at 0.001. The loss function employed was Mean Squared Error (MSE) to ensure stability in the regression task.

During the training process, the model parameters were iteratively updated, with changes in the loss value serving as the criterion for assessing the model's convergence. The training utilized an Early Stopping strategy, where the process was halted if the model's performance on the validation set ceased to improve over several consecutive epochs. After each epoch, the model's performance was evaluated based on its validation set performance. If the validation loss did not show significant improvement over 5 consecutive epochs, the training was prematurely stopped to avoid overfitting and excessive training time.

After the model training is completed, the model's predictive performance is validated using the test set. The model's output is compared with the actual data to calculate errors, and the model's performance is evaluated using relevant metrics. If the predictive results do not meet expectations, model parameters such as learning rate, the number of hidden units, and the number of layers need to be readjusted, followed by retraining and testing until the model's predictions meet the desired accuracy. The final stopping criterion for the model training is when the validation loss no longer shows significant improvement or when a predetermined number of training epochs is reached, resulting in a stable model ready for real-world predictions.

4. Empirical Analysis and Solution

4.1. Data Preprocessing

In financial time series forecasting, it is crucial to ensure sample quality while fully considering the diversity and representativeness of the subject. To ensure the applicability and accuracy of the time series forecasting model in the e-commerce industry, we selected the A-share "Internet E-commerce" sector index (code: 881177) as the data source, representing the overall performance of China's online retail market.

The sample data covers all trading periods from January 20, 2016, to August 16, 2024, with data sourced from the Tonghua Shun client.

To enhance the model's prediction accuracy for the Chinese Internet E-commerce sector index, we conducted comprehensive feature extraction and indicator construction based on the original data. From dimensions such as trading volume and total transaction amount, 13 indicators were selected as feature input vectors. The LSTM model was employed to thoroughly explore the deep-seated information within these features that influence stock index movements, aiming for accurate prediction of the e-commerce index's closing price. The feature vectors and corresponding indicators of the A-share E-commerce Index are shown in Table 1.

Table 1. Feature Vector and Corresponding Indicators

Feature Vector	Indicator
x_1	Time
x_2	Opening Price
x_3	Highest Price
x_4	Lowest Price
x_5	Closing Price
x_6	Price Increase (%)
x_7	Amplitude (%)
x_8	Trading Volume
x_9	Transaction Amount
x_{10}	Transaction Amount/Trading Volume
x_{11}	Highest - Lowest
x_{12}	Trading Volume * Amplitude
x_{13}	Opening Price - Closing Price

In deep learning, sample data is typically divided into training, validation, and test sets, as illustrated in the following figure. The training set is used to estimate model parameters (such as weight matrix W), the validation set is used to fine-tune the neural network architecture (such as the number of hidden layers and hidden units), and the test set is used to evaluate the generalization ability of the trained model (i.e., its ability to predict on out-of-sample time series data). To ensure that the model adheres to the chronological nature of stock market data, we carefully structured the dataset such that the training set always consists of earlier time periods than the test set. This prevents any potential leakage of future information into the model's training process, which would otherwise distort the evaluation of the model's predictive power. Specifically, the training set was comprised of stock data from January 20, 2016, to November 29, 2022, while the test set included data from November 30, 2022, to August 16, 2024,as shown in Table 2.

Table 2. E-Commerce Index Dataset Split

Dataset	Trading Days	Time Period
Training Set	1668	2016/01/20 - 2022/11/29
Validation Set	417	2022/11/30 - 2024/08/16

Additionally, when dealing with time series data, particularly in the training of deep learning models, normalization is a critical step to ensure model convergence and enhance prediction performance. For a given time series dataset x_1, x_2, x_3,......,x_t,the normalization formula is as follows:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)},\ i = 1,2,\dots,t \tag{7}$$

In this formula, $\min(x)$ and $\max(x)$represent the minimum and maximum values of the time series data, respectively. This process scales the data to a uniform numerical range, reducing the sensitivity of feature values to different magnitudes. Consequently, it better controls gradient changes during model training and accelerates model convergence.

4.2. Model Performance Evaluation Metrics

To provide a more objective evaluation of the model, this study selects three metrics to assess the prediction results from different perspectives: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{10}$$

In these formulas, y_i represents the actual value of the i-th sample, $\hat{y}_i$ represents the predicted value of the i-th sample, $\bar{y}$ is the mean of the samples, and n is the total number of samples.

4.3. Empirical Results Analysis

The empirical analysis results are shown in Figure 3, where the LSTM model demonstrates significant advantages in the e-commerce dataset. From the LSTM prediction results, it is evident that the model can effectively capture the fluctuations in the closing price. Its prediction curve closely matches the actual values in the validation set, maintaining high predictive accuracy even during periods of significant market volatility. This suggests that the LSTM-based prediction model is capable of handling complex time series characteristics and is suitable for forecasting the closing price of China's internet e-commerce industry index.

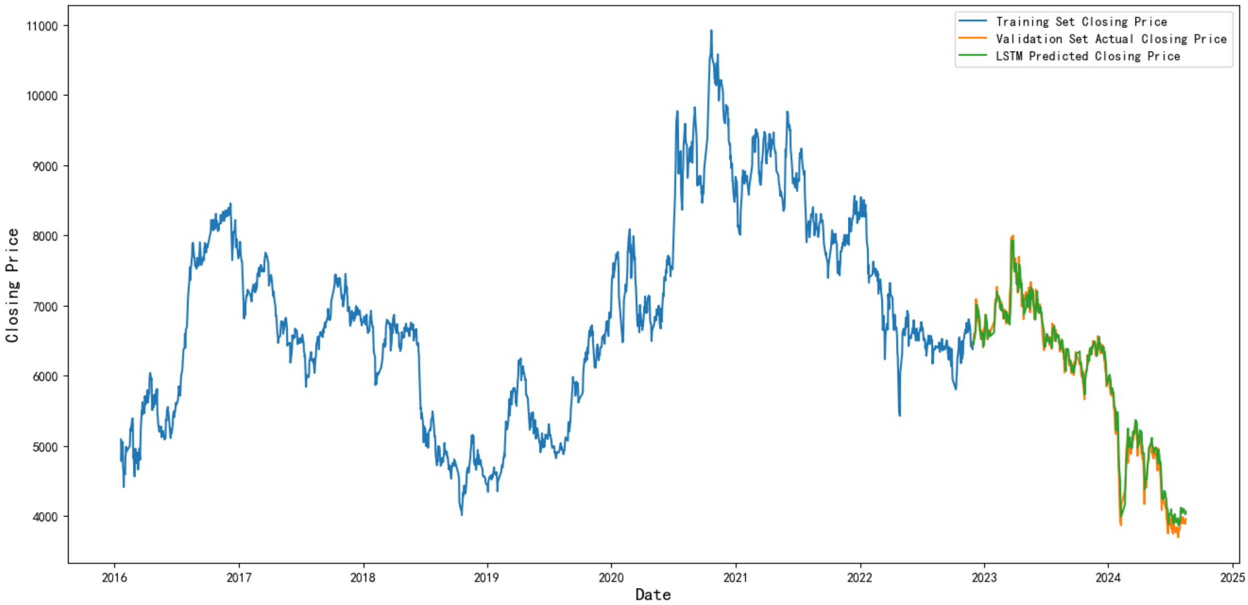

Figure 3. LSTM Model Closing Price Prediction

In contrast, the LGBM and Linear Regression models performed significantly worse in the predictions, as shown in Figures 4 and 5. Although the LGBM model is widely used in machine learning, its prediction curve in this empirical analysis deviated significantly from the actual closing prices, resulting in a large error and failing to demonstrate its advantages. The Linear Regression model performed even worse, with its predictions showing a substantial deviation and the prediction curve exhibiting high instability. This indicates that the Linear Regression model struggles to handle the complex nonlinearities and time dependencies inherent in stock market data.

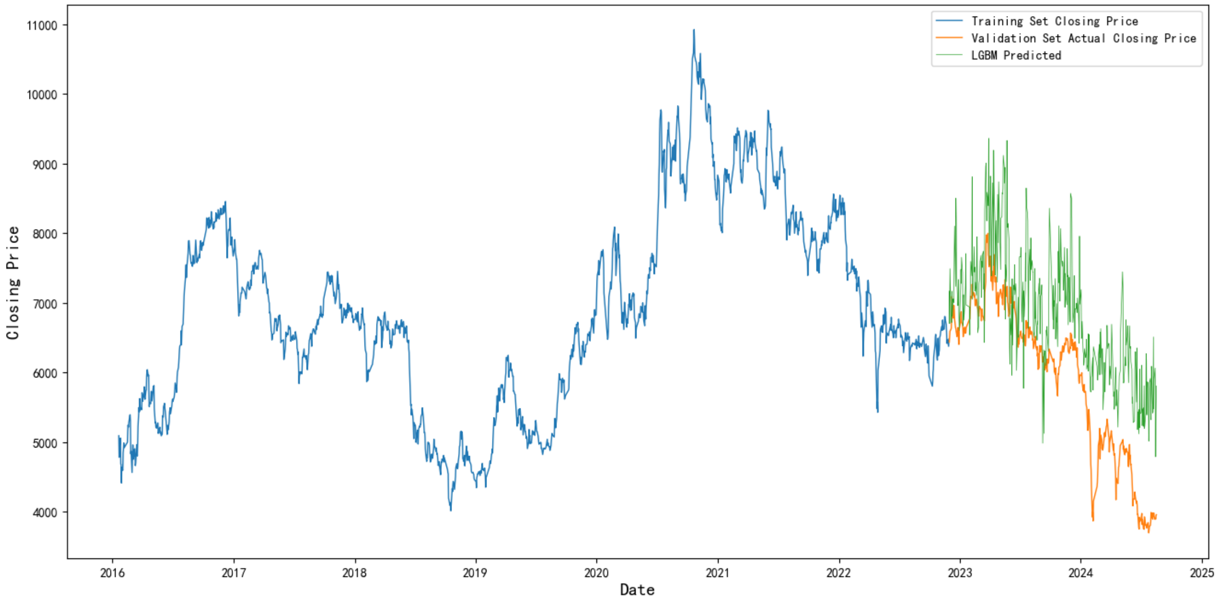

Figure 4. LGBM Model Closing Price Prediction

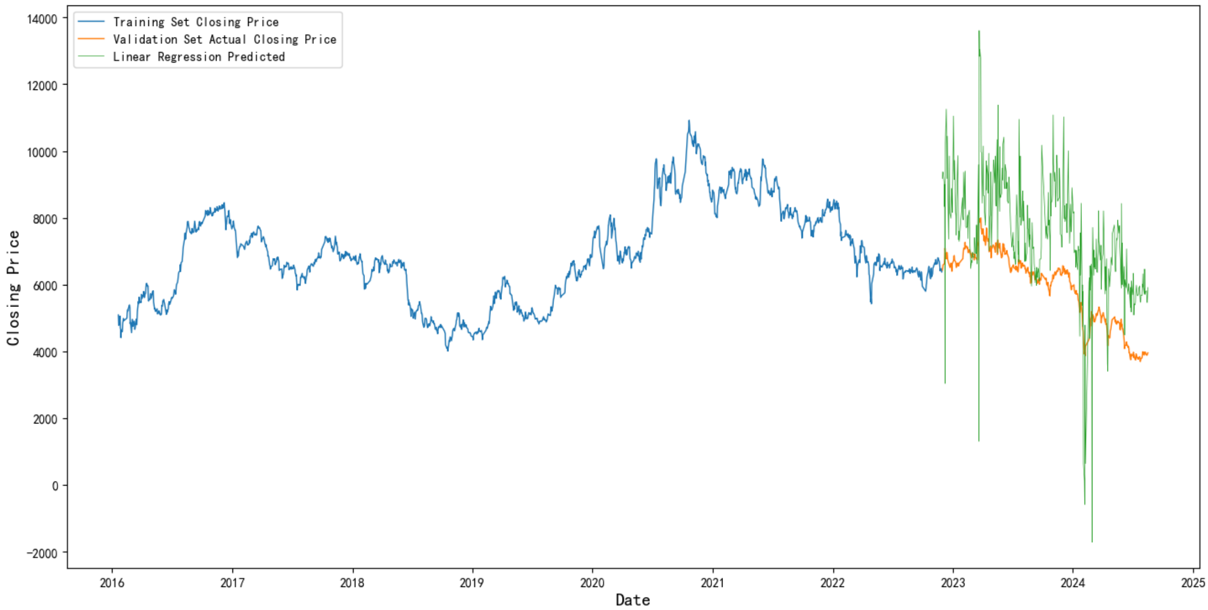

Figure 5. Linear Regression Model Closing Price Prediction

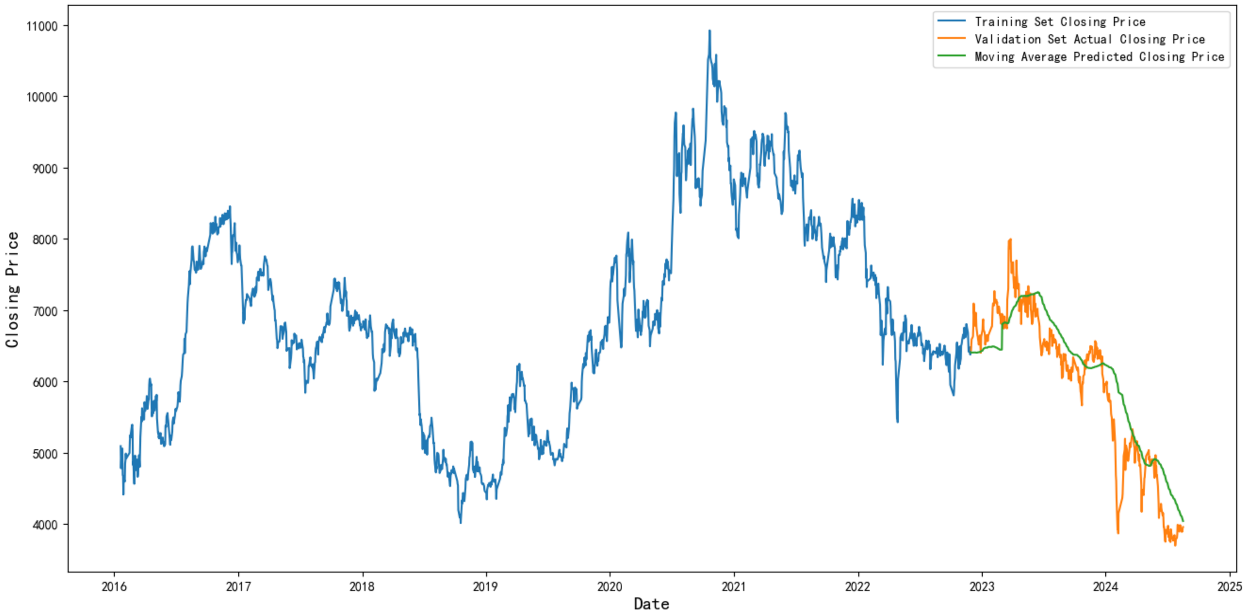

Figure 6. Moving Average Method Closing Price Prediction

Finally, the Moving Average method, as a traditional time series analysis technique, exhibited prediction performance that fell between the other models, as shown in Figure 6. While it provides relatively smooth predictions during stable periods, its prediction lag becomes quite apparent in the face of significant market fluctuations, making it unable to promptly reflect the market's dramatic changes.

4.4. Experimental Results

Table 3. Performance Evaluation Metrics for Different Models

	RMSE	MAE	R^2
LSTM	142.15	108.18	0.98
MA	480.78	368.71	0.81
LGBM	1135.63	979.75	-0.04
LinearRegression	2021.30	1737.78	-2.29

The final conclusion is clearly demonstrated through the comparative analysis of the graphs and various evaluation metrics, highlighting the performance of different models in forecasting. As shown by the evaluation metrics in Table 3, the LSTM model excelled across all measures, with an RMSE of 142.15, MAE of 108.18, and an R^2 as high as 0.98. This indicates a high level of predictive accuracy and generalization ability, significantly outperforming the other models. Although the traditional Moving Average method is simple, it performed reasonably well in capturing long-term trends, with an R^2 of 0.81. However, due to its inability to handle short-term volatility, the RMSE reached 480.78 and the MAE 368.71, which still lags behind the LSTM model. Among the machine learning models, both LGBM and Linear Regression did not perform well, with Linear Regression particularly underperforming, as indicated by an RMSE of 2021.30, MAE of 1737.78, and an R^2 that even showed a negative value of -2.29. This result implies that the model's fit to the data was extremely poor, unable to capture the trends in stock prices. The LGBM model performed slightly better, but its R^2 was still negative (-0.04), indicating severe underfitting issues.

To ensure the robustness and reliability of the results, we conducted multiple experiments, each with different data splits and random initializations. In total, more than three rounds of experiments were performed, with the best results presented in Table 3. These results represent the optimal balance between training efficiency and model performance. By averaging the results across multiple trials, we minimized the potential impact of any single split or initialization on the evaluation metrics, further reinforcing the conclusion that the LSTM model consistently outperforms the other models in both accuracy and generalization.

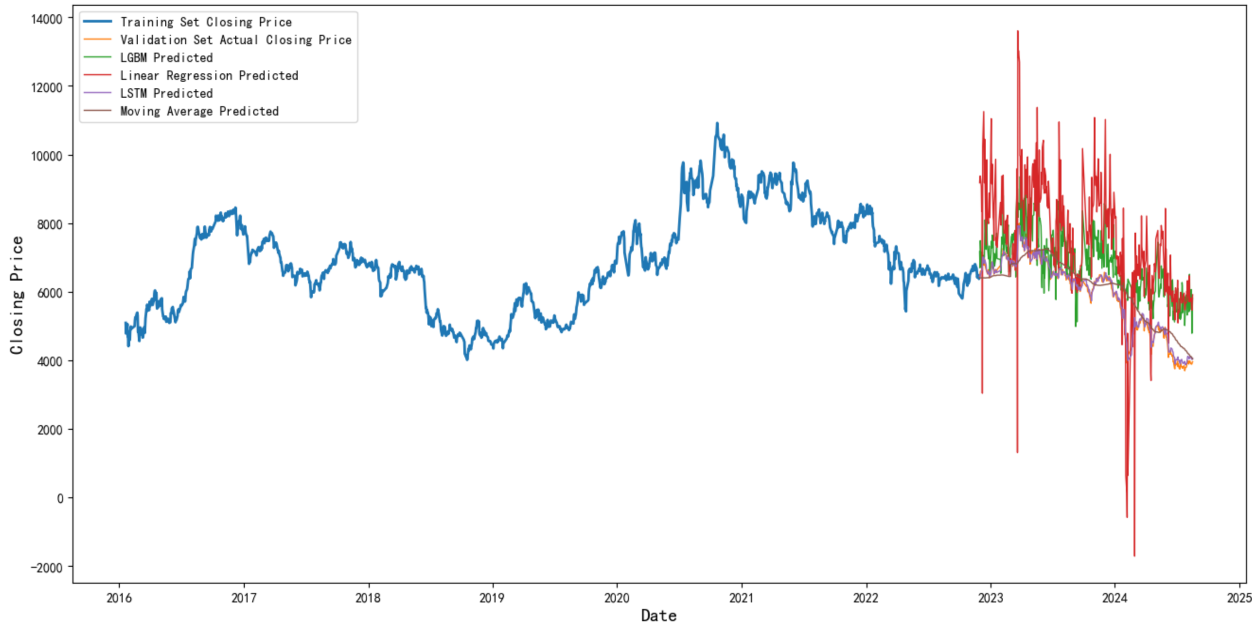

Figure 7. Comparison of Predictions from LSTM, LGBM, Linear Regression, and Moving Average

In the comparison shown in Figure 7, the LSTM prediction curve almost completely overlaps with the actual closing price curve, whereas the prediction curves of the other models show significant deviations, especially those of the Linear Regression and LGBM models, highlighting their inadequacy in capturing price fluctuations.

Overall, the LSTM model stands out among various models due to its strong learning capability, demonstrating its superiority in forecasting the internet e-commerce industry index.

5. Conclusion

Given the importance and difficulty of predicting stock prices in the context of "Double 11," this paper proposes a method for forecasting stock price fluctuations in the internet e-commerce sector during the "Double 11" shopping festival based on the LSTM model. The experiment validates that the LSTM network model performs superiorly in stock price analysis and prediction by comparing the results of different forecasting methods under the "Double 11" context.

This study applies deep learning technology to stock price forecasting in the e-commerce industry during the "Double 11" period, providing a new decision support tool for investors and e-commerce enterprises. To further enhance the model's prediction accuracy and generalization capability, future research will explore the application of the LSTM model in different industries and market environments. By incorporating other financial factors, such as sentiment analysis, the accuracy of predictions can be improved, offering more valuable references for investors and enterprises. Additionally, extending static data to real-time data streams could enable dynamic stock price forecasting, providing investors with more timely decision support.

References

[1] Gong, X. (2024). ARIMA-BPNN based stock price prediction model based on fusion news sentiment analysis. *Scalable Computing: Practice and Experience*, *25* (3), 2062 - 2075.

[2] Wang, Y., Zhang, Y., & Wei, G. (2023). Stock price prediction based on BP neural network and higher-order fuzzy cognitive map. *Intelligent Computing and Applications*, *13* (08), 100 - 106+113. (In Chinese)

[3] Wang, P., Zhou, X., & Chen, X. (2024). Stock price and sentiment multi-feature nested attention fusion stock price prediction research. *Systems Engineering*, 1-14. Retrieved August 21, 2024. (In Chinese).

[4] Zheng, J., Xin, D., Cheng, Q., Tian, M., & Yang, L. (2024, February 27). The random forest model for analyzing and forecasting the US stock market in the context of smart finance. *arXiv preprint* arXiv:2402.17194.

[5] Saha, S., Gao, J., & Gerlach, R. (2022). A survey of the application of graph-based approaches in stock market analysis and prediction. *International Journal of Data Science and Analytics*, *14* (1), 1 - 15.

[6] Wang, G., et al. (2019). Application of a long short-term memory neural network: A burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China. *Epidemiology and Infection*, *147*, e194, 1 - 7.

[7] Zhao, C., Wu, M., Liu, J., Duan, Z., Shen, L., Shangguan, X., Liu, D., & Wang, Y. (2023). Progress and prospects of data-driven stock price forecasting research. *International Journal of Cognitive Computing in Engineering*, *4*, 100 - 108.

[8] Li, H., Huang, X., Zhou, D., Cao, A., Su, M., Wang, Y., & Guo, L. (2022). Forecasting carbon price in China: A multimodel comparison. *International Journal of Environmental Research and Public Health*, *19*, 6217.

[9] Li, Y., & Wei, Z. (2022). Regional logistics demand prediction: A long short-term memory network method. *Sustainability*, *14* (13478).

[10] Chen, J., Li, Y., Zhang, C., Tian, Y., & Guo, Z. (2023). Urban flooding prediction method based on the combination of LSTM neural network and numerical model. *International Journal of Environmental Research and Public Health*, *20*, 1043.

[11] Shahrel, M. Z., Mutalib, S., & Abdul-Rahman, S. (2021). PriceCop–Price monitor and prediction using linear regression and LSVM-ABC methods for e-commerce platform. *International Journal of Information Engineering and Electronic Business (IJIEEB)*, *13* (1), 1 - 14.

[12] Ji, C., Zhang, C., Hua, L., Ma, H., Nazir, M. S., & Peng, T. (2022). A multi-scale evolutionary deep learning model based on CEEMDAN, improved whale optimization algorithm, regularized extreme learning machine and LSTM for AQI prediction. *Environmental Research*, *215* (Pt 1), 114228.

[13] Wang, W., Shao, J., & Jumahong, H. (2023). Fuzzy inference-based LSTM for long-term time series prediction. *Scientific Reports*, *13* (1), 20359.

[14] Salehi, W., Baglat, P., Gupta, G., Khan, S. B., Almusharraf, A., Alqahtani, A., & Kumar, A. (2023). An approach to binary classification of Alzheimer's disease using LSTM. *Bioengineering (Basel)*, *10* (8), 950.

[15] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *12* (9), 1735 - 1780.