Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 1
AIEA 2023

Stock Price Prediction based on LSTM and XGBoost Combination Model

Yiming Zhu

School of International Trade and Economics, University of International Business and Economics, Beijing, 100105, China

Abstract. In recent years, many machine learning and deep learning algorithms have been applied to stock prediction, providing a reference basis for stock trading, and LSTM neural network and XGBoost algorithm are two typical representatives, each with advantages and disadvantages in prediction. In view of this, we propose a combination model based on LSTM and XGBoost, which combines the advantages of LSTM in processing time series data and the ability of XGBoost to evaluate the importance of features. The combination model first selects feature variables with high importance through XGBoost, performs data dimensionality reduction, and then uses LSTM to make predictions. In order to verify the feasibility of the combination model, we built XGBoost, LSTM and LSTM-XGBoost models, and carried out experiments on three data sets of China Eastern Airlines, China Merchants Bank and Kweichow Moutai respectively. Finally, we concluded that the proposed LSTM-XGBoost model has good feasibility and universality in stock price prediction by comparing the accuracy of the predicted images and their performance in RMSE, RMAE, and MAPE indicators.

Keywords: Stock Price Prediction; Machine Learning; LSTM Neural Network; XGBoost Ensemble Learning; Combination Model.

1. Introduction

The stock market is a vital part of the global economy, reflecting the value and future growth potential of enterprises. The fluctuation of stock prices is affected by many factors, such as market demand and supply, company financial performance, macroeconomic policies, social events, investor sentiment, etc. Therefore, predicting stock prices is a difficult and worthwhile task that can assist investors in making rational investment decisions, enhancing returns, and lowering risks.

With the advancement of the Internet and big data technology, the data volume in the stock market is growing exponentially, providing a rich and diverse source of information for stock price forecasting. However, traditional stock price prediction methods, such as fundamental analysis, technical analysis, statistical models, often fail to fully exploit these data and cannot effectively deal with the complexity, nonlinearity, and dynamism of the data. Therefore, in recent years, more and more researchers have started to investigate the feasibility of using machine learning algorithms for stock price prediction.

Machine learning is a technique that enables intelligent behavior by learning rules and patterns from data. Machine learning algorithms can automatically extract features from large-scale data, build models, and make predictions or decisions based on new input data. Machine learning algorithms have strong generalization and adaptability, and can cope with nonlinear, high-dimensional, and dynamic data, making them suitable for stock price prediction.

Deep learning techniques based on neural networks, such as long short-term memory (LSTM), have been widely used in the research of stock price prediction in recent years. These techniques can effectively process time series data and capture the dynamic changes and long-term dependencies of stock prices, thanks to the strong memory ability of neural networks. However, deep learning techniques also have some drawbacks, such as high requirements for computing resources, easy overfitting, and difficult to explain internal mechanisms (LSTM is a black box model). Therefore, traditional machine learning algorithms, such as extreme gradient boosting (XGBoost), still play an important role in the field of stock prediction. The XGBoost algorithm has many advantages over deep learning techniques. First, it has a wide range of applications and can handle various structured data, such as tabular data and numerical features. Second, it can automatically handle the interaction between features, simplifying the need for manual feature engineering. Third, it is highly optimized, using technologies such as multi-threading and parallel computing, and has excellent performance. Fourth, it is highly interpretable and can provide feature importance assessments, which help explain the decision-making process of the model.

This article attempts to construct a combination model based on LSTM and XGBoost. Firstly, XGBoost is used for feature selection, and then LSTM is used for prediction. This paper applies the combination model to three stock price data sets of China Eastern Airlines, Kweichow Moutai and China Merchants Bank, and compares the prediction results with the single XGBoost and LSTM models. It is found that the combination model has further improved the prediction effect, and has significantly improved the prediction accuracy and hysteresis.

The article is organized as follows: In section two, we introduce the principles and modeling process of XGBoost and LSTM; In section three, we introduce the construction principle and model evaluation indicators of the LSTM-XGBoost combination model; In section four, we elaborate on the relevant processes of data obtain, dataset partitioning and data processing; In section five, we present the experimental results and conduct a specific analysis and the conclusions are given in section six.

2. Related Work

2.1. XGBoost Principles and Modeling

XGBoost is an improved algorithm based on gradient boosting decision tree (GBDT). Different from traditional GBDT, XGBoost supports linear classifiers, including logistic regression classifiers and linear regression classifiers. In addition, XGBoost uses the second-order Taylor expansion when processing the loss function, so that the input of its weak classifier includes the second-order derivative in addition to the residual (first-order derivative). And XGBoost adds a regularization penalty term to the loss function of each weak classifier. Its function is to control the number of leaf nodes in the model, so as to limit the complexity and avoid overfitting of the model. The derivation process of the objective function of XGBoost is as follows.

The initial objective function is

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega\,(f_i). \tag{1}$$

Among them, $l(y_i, \hat{y}_i^{(t)})$ represents the difference between the predicted value and the real value, and $\Omega(f_i)$ is a regular term, expressed in the form of

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda, \tag{2}$$

where ω_j is the weight item of the jth leaf node.

Then rewrite $l(y_i, \hat{y}_i^{(t)})$ into the form of $l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$, where $\hat{y}_i^{(t-1)}$ means that the model prediction of the previous $t-1$ round is retained, $f_t(x_i)$is the newly added model in the t_{th} round, so the objective function can be rewritten into the form of

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + C, \tag{3}$$

where C is a constant term. The goal is to find f_t to optimize the objective function to be as small as possible.

Use Taylor expansion to approximate the original objective function, and define two variables g_i and h_i, as shown in

$$g_i = \partial_{\hat{y}(t-1)} l\left(y_i, \hat{y}^{(t-1)}\right) \tag{4}$$

$$and\ h_i = \partial^2_{\hat{y}(t-1)} l\left(y_i, \hat{y}^{(t-1)}\right). \tag{5}$$

At this time, the objective function can be rewritten in the form of

$$Obj^{(t)} \approx \sum_{i=1}^{n} [\, l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + C. \tag{6}$$

Since the predicted value and the real value of the previous t-1 round model are fixed values when the model is constructed in the t-th round, $l(y_i, \hat{y}_i^{(t-1)})$ can be used as a constant term, and at the same time Since the constant term has no effect on the optimization solution, it is removed. The objective function at this time can be rewritten in the form of

$$\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(ft). \tag{7}$$

Then rewrite the objective function from traversal on the sample to the form of traversal on the leaf nodes, and expand the regular term. Now, the objective function can be expressed by

$$Obj^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T. \tag{8}$$

For convenience, two variables are defined, as shown in

$$G_j = \sum_{i \in I_j} g_i \tag{9}$$

$$and\ \ H_j = \sum_{i \in I_j} h_i\,. \tag{10}$$

Substituting it into the objective function, the new objective function is

$$Obj^{(t)} = \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2\right] + \gamma T. \tag{11}$$

In order to solve the weight of the leaf node j when minimizing the objective function, take the partial derivative of the objective function and make it equal to 0, as shown in

$$\frac{\partial J(f_t)}{\partial \omega_j} = G_j + \left(H_j + \lambda\right) w_j = 0. \tag{12}$$

Solving equation (12) can solve the weight ω_j of leaf node j,

$$\omega_j = -\frac{G_j}{H_j + \lambda}. \tag{13}$$

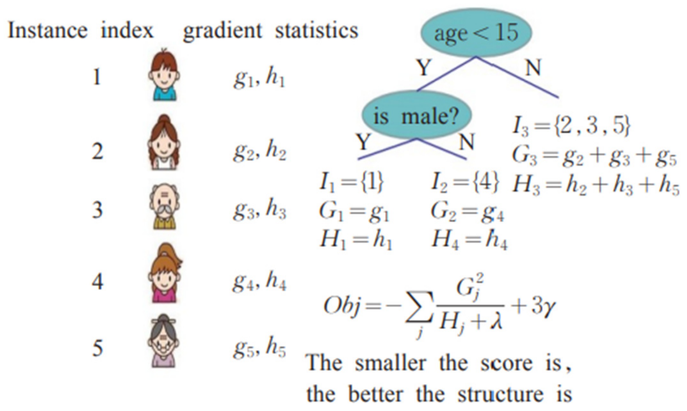


Figure 1. Schematic diagram of XGBoost decision tree effect evaluation

(Chen, T. (2014). Introduction to Boosted Trees [PowerPoint slides]. [6]. Image on slide 28.)

Finally, use Figure 1 to illustrate how to use the objective function to evaluate the effect of the decision tree.

Bring it back to the original objective function and we can get

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T. \tag{14}$$

2.2. LSTM Principles and Modeling

The LSTM network is a structural variant based on the improvement of the Recurrent Neural Network (RNN). LSTM contains three gating mechanisms, namely, forgetting gate, input gate and output gate. These gating mechanisms enable LSTM to automatically filter, forget and output input data, thereby effectively learning important patterns and laws in the data. . In addition, the key idea of LSTM is to introduce a memory unit called "cell state", which can store long-term information and can selectively forget and update information. This improvement enables LSTM to better capture long-term dependencies when processing data with time series properties, and overcomes the shortcomings of traditional RNNs that are prone to gradient disappearance and gradient explosion problems in long sequence data. The overall structure of LSTM is introduced below.

First, through the forget gate, as shown in Figure 2, the process can be expressed in the form of

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{15}$$

where x_t is the current data and h_{t-1} is all hidden information in the previous stage.

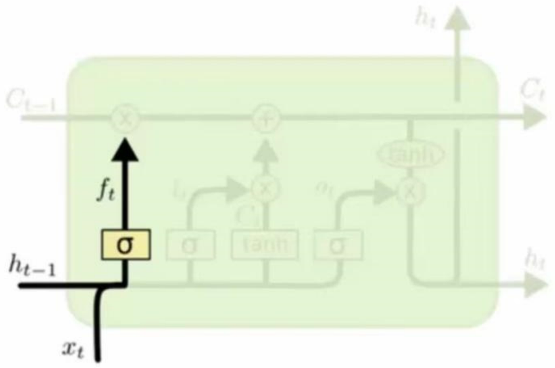


Figure 2. Schematic diagram of LSTM forgetting gate

σ is the Sigmoid function, as shown in Figure 3. The Sigmoid layer outputs a value between 0 and 1, describing how much each part can pass. 0 means "do not allow any amount to pass", 1 means "allow any amount to pass". The Sigmoid function determines which information will be retained through a set of weight parameters.

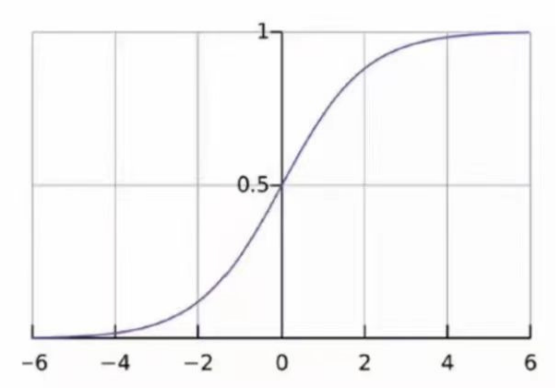


Figure 3. Schematic diagram of the sigmoid function

Then multiply the calculated f_t with the control parameter c_{t-1} of the previous stage to determine which information to discard.

Then, through the input gate, the updated information is determined, as shown in Figure 4. The process can be expressed in the form of

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{16}$$

$$and\ \tilde{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \tag{17}$$

Among them, i_t is the new information to be preserved, and $\tilde{C_t}$ is the control parameter for the formation of new data.

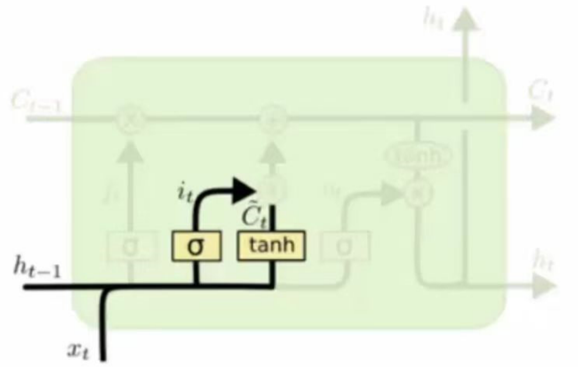


Figure 4. Schematic diagram of LSTM input gate

Secondly, update the cell state, as shown in Figure 5, the process can be expressed in the form of

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t}. \tag{18}$$

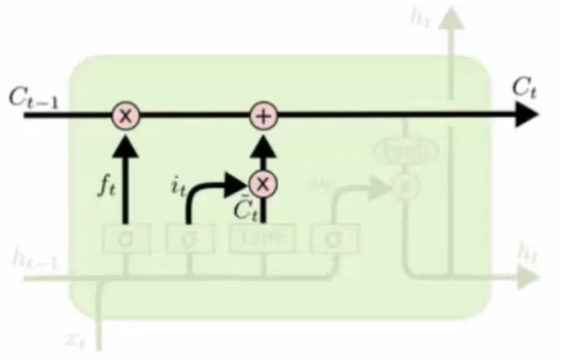


Figure 5. Schematic diagram of LSTM cell state update

Finally, the information is output by the output gate, as shown in Figure 6, and the process can be expressed by

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{19}$$

$$and\ h_t = o_t * tanh(C_t). \tag{20}$$

Where the output is produced using the new control parameters.

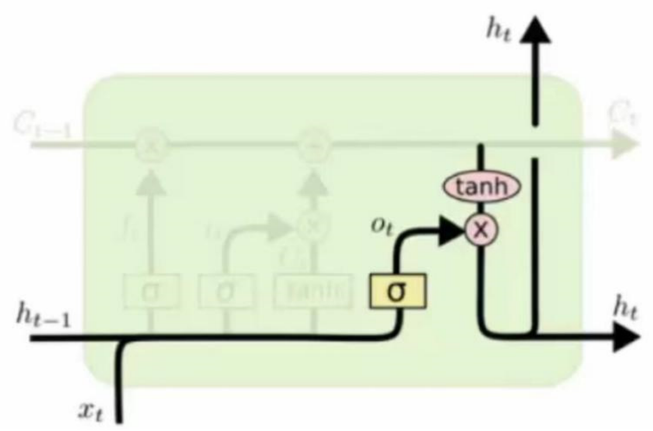
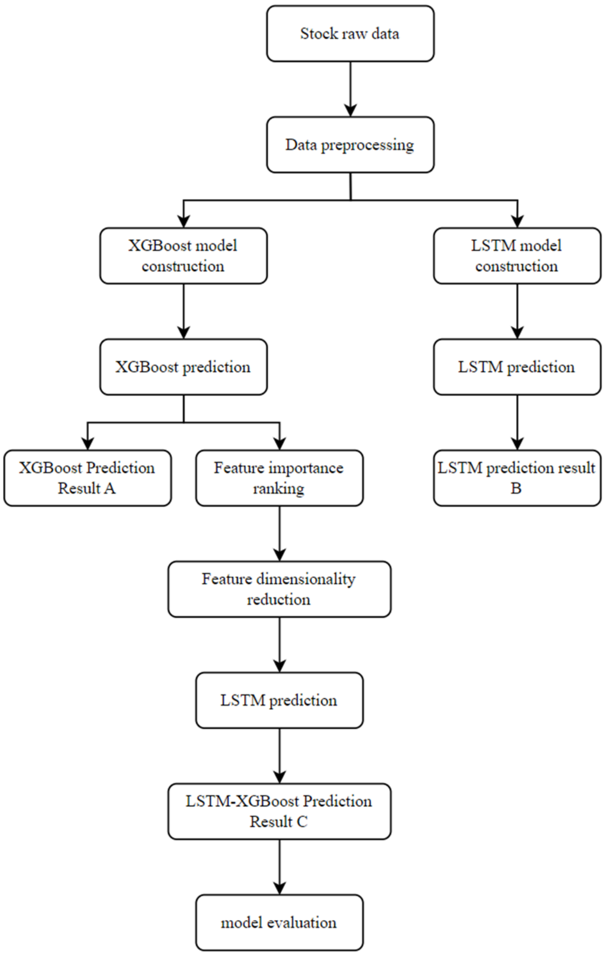


Figure 6. Schematic diagram of LSTM output gate

3. Construction of LSTM-XGBoost Combination Model

3.1. Principles of the Combination Model



Stock raw data
Data preprocessing
XGBoost model construction
LSTM model construction
XGBoost prediction
LSTM prediction
XGBoost Prediction Result A
Feature importance ranking
LSTM prediction result B
Feature dimensionality reduction
LSTM prediction
LSTM-XGBoost Prediction Result C
model evaluation

Figure 7. Experimental process

The LSTM model performs well in processing time series data due to its memory ability, long-term dependence, and gating mechanism. Therefore, we will use the LSTM model for the final stock price prediction. However, the LSTM model also puts forward higher requirements for feature variable selection. Due to the relatively complex neural network structure of the LSTM model, excessively high data dimensions will increase the computational complexity and memory consumption of the model, reducing the training efficiency of the model. At the same time, high-dimensional data is often accompanied by relatively more redundant information and noise, which increases the risk of overfitting in the model and reduces its generalization ability. Therefore, it is particularly important to perform dimensionality reduction on the dataset before inputting data into the model.

At the same time, we noticed that XGBoost calculates the importance score of features during the training process, which can effectively help us understand which features contribute the most to the predictive performance of the model. XGBoost has three ways to calculate feature importance, namely weight, gain, and coverage. Among them, the weight form is the default method, which represents the number of times a feature variable is used as a split node. Secondly, the gain form refers to the information gain brought about by each feature segmentation during the process of constructing the decision tree, which is the reduction in the value of the objective function. Finally, the coverage form is the number of samples covered by the leaf nodes represented by the feature variable divided by the number of times the feature variable is used for splitting. The greater the coverage, the closer a feature variable is to the root of the tree and the higher its importance.

Given the advantages of XGBoost in feature importance assessment, we applied XGBoost to feature engineering and constructed the LSTM-XGBoost model. The specific process is shown in Figure 7. In the first stage, the original dataset is input into the XGBoost model to obtain scores on the importance of each feature.

In the second stage, based on score ranking, select the top five feature variables with importance ranking and record them.

In the third stage, the filtered data set is input into the LSTM model, and the parameters of the model are adjusted to obtain the optimal prediction results.

3.2. Model Evaluation Indicators

We introduced three performance indicators to evaluate the predictive performance of combination models and individual models, namely root mean square error (RMSE), root mean square absolute error (RMAE), and mean absolute percentage error (MAPE). RMSE and RMAE measure the absolute magnitude of the deviation between the predicted value and the true value, expressed in the form of formulas (21) and (22), respectively. The difference is that RMSE is more sensitive to outlier data and can sensitively capture error values with large deviations, while RMAE is less susceptible to extreme values and can accurately reflect the overall situation of errors. MAPE measures the relative magnitude of the deviation between the predicted value and the true value, which is expressed in the form of formula (23). The measurement standard for MAPE is percentage, with a value of 0% indicating a perfect model and a value greater than 100% indicating a poor quality model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{R}}_i - \boldsymbol{R}_i)^2} \tag{21}$$

$$RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\boldsymbol{R}}_i - \boldsymbol{R}_i\right|} \tag{22}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{R}_i - R_i}{R_i}\right| \times 100\% \tag{23}$$

4. Data Representation and Processing

4.1. Obtaining Data

This article selects the next day's closing price of stocks as the model output, which is the indicator we want to predict. To achieve this goal, we selected basic stock market indicators, technical indicators, and financial indicators as input variables. Specifically, the indicators selected in this article are as follows:

(1) Basic market indicators: including open opening price, high highest price, low lowest price, close closing price of the day, volume of trading, amount of transaction volume, turnover rate turn, and percentage increase/decrease (pctChg). Such market indicators can reflect investors' trading situation of the stock, and are therefore closely related to stock price fluctuations.

(2) Technical indicators: including 9-day EMA of index moving average_ SMA for 5, 10, 15, and 30 cycles of simple moving averages_ 5, SMA_ 10, SMA_ 15, SMA_ 30. Moving average convergence divergence MACD, moving average signal line of fast line, relative strength index RSI, and Wilhelm index WPR.

(3) Financial indicators: including rolling P/E ratio peTTM, rolling P/S ratio psTTM, and rolling P/L ratio pcfNcfTTM. Investors will evaluate a company's financial health, profitability, and risk level by analyzing such financial indicators, and then make trading decisions.

The specific meanings of technical and financial indicators are shown in Table 1:

Table 1. Technical index description

Indicator Name	Indicator Meaning
EMA_9	a more sensitive moving average that can quickly reflect the trend of stock price changes
SMA_5	a commonly used short-term moving average that can display the short-term trend of stock prices
SMA_10	a commonly used medium-term moving average that can display the mid-term trend of stock prices
SMA_15	a commonly used medium to long term moving average that can display the medium to long term trend of stock prices
SMA_30	a commonly used long-term moving average that can display the long-term trend of stock prices
MACD	an indicator that measures stock price trends and momentum, consisting of a fast line (DIF) and a slow line (DEA)
Signal line	a slow line in the MACD indicator, used to determine the buying and selling signals of the MACD indicator
RSI	an indicator that measures the magnitude and speed of stock price fluctuations, used to determine whether the stock price is in an overbought or oversold state
WPR	an indicator used to analyze short-term market trends and predict high and low points during the cycle period
peTTM	a measure of the ratio of stock price to earnings per share, used to determine whether a stock is overvalued or undervalued
psTTM	a measure of the ratio of stock price to sales per share, used to determine whether a stock has growth potential
pcfNcfTTM	a measure of the ratio of stock price to cash flow per share, used to determine whether a stock is stable

The experimental data in this article was collected from the open-source data interface provided by the Baostock Securities Data Platform, covering a total of 1355 trading days from January 1, 2018 to August 1, 2023. Taking the research object Eastern Airlines (securities code 600115) as an example, its closing price situation during the time span we studied is shown in Figure 8.

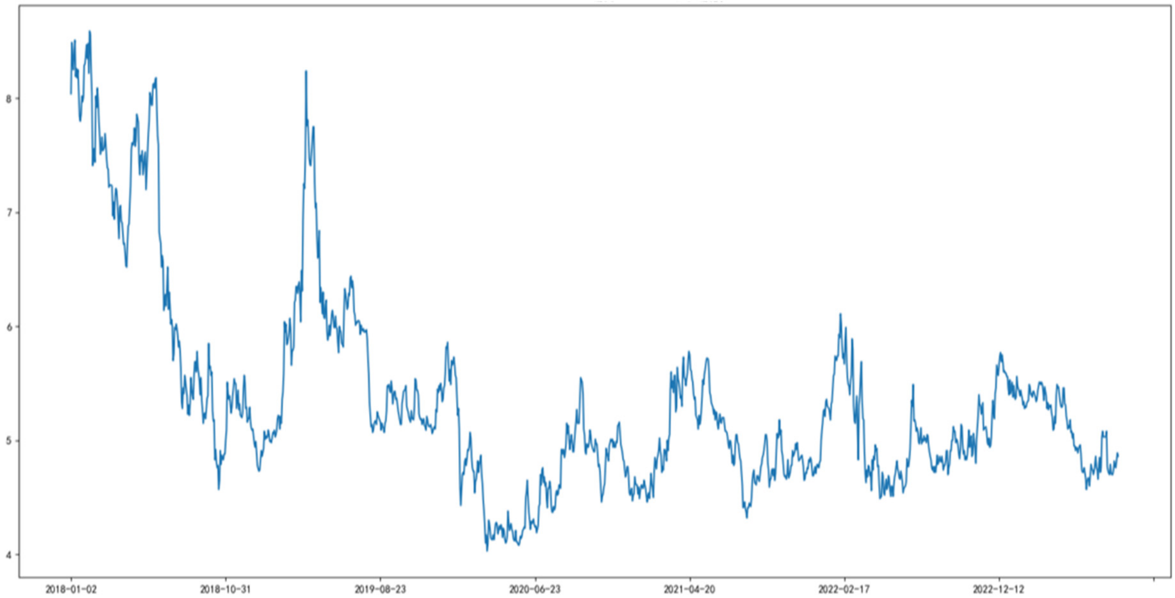


Figure 8. Time series of daily closing prices of China Eastern Airlines (sh.600115)

4.2. Dataset Partitioning

We divided the experimental dataset into two parts: a training set and a testing set, with the training set used for model training, covering the period from January 1, 2018 to November 26, 2021, accounting for 70% of the entire dataset. The test set is used to evaluate model performance, with a range of November 29, 2021 to August 1, 2023, accounting for 30% of the entire dataset. The composition of the dataset is shown in Table 2.

Table 2. Dataset division

data set	proportion	Days	period of time
Training set	0.7	948	20180101~20211126
Test set	0.3	407	20211129~20230801

4.3. Data Processing

Due to the differences in dimensions and units of stock market data, in order to eliminate the adverse effects of different dimensions on prediction, it is necessary to normalize the relevant data. This article uses the method of deleting the mean and reducing it to unit variance to process the data, as shown in formula (24):

$$X' = \frac{X - mean}{std} \tag{24}$$

The original array is of the same order of magnitude after normalization, without losing its original feature attributes, so it can be used comprehensively.

5. Experimental Results and Discussion

In order to verify the feasibility of the LSTM-XGBoost model, we established three models, namely, XGBoost, LSTM and LSTM-XGBoost, and conducted experiments on three datasets, namely, China Eastern Airlines, China Merchants Bank and Kweichow Moutai.

5.1. Experimental Results on the Eastern Airlines Dataset

Figure 9 shows the comparison between the predicted and true values of the XGBoost model on the Eastern Airlines test set. From the graph, it can be seen that the predicted value of XGBoost is relatively close to the true value in the overall trend, and has a small lag, which can better reflect the overall trend of the stock price. However, its performance in characterizing local subtle changes in stock prices is relatively poor, and there is a problem of overreaction or underreaction to changes in stock prices during certain periods of time.

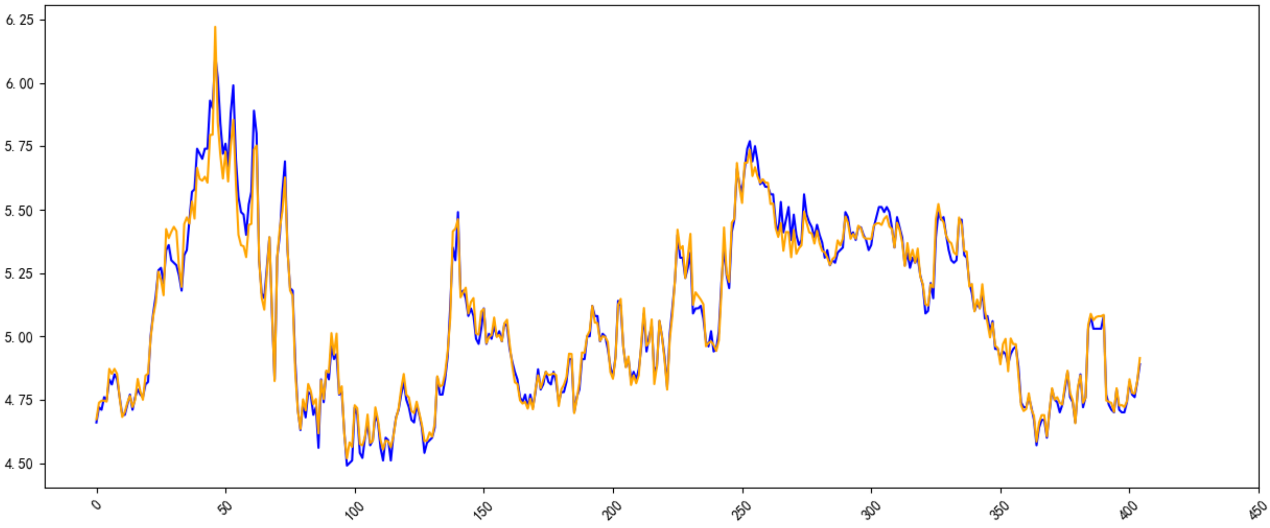


Figure 9. Predicted and true values of XGBoost model

In addition, we obtained the importance ranking of input features through the XGBoost model, as shown in Figure 10. We selected the top five important features from them, which are the closing price (173), the highest price (86), the William index (59), the rise and fall range (43), and the opening price (39).

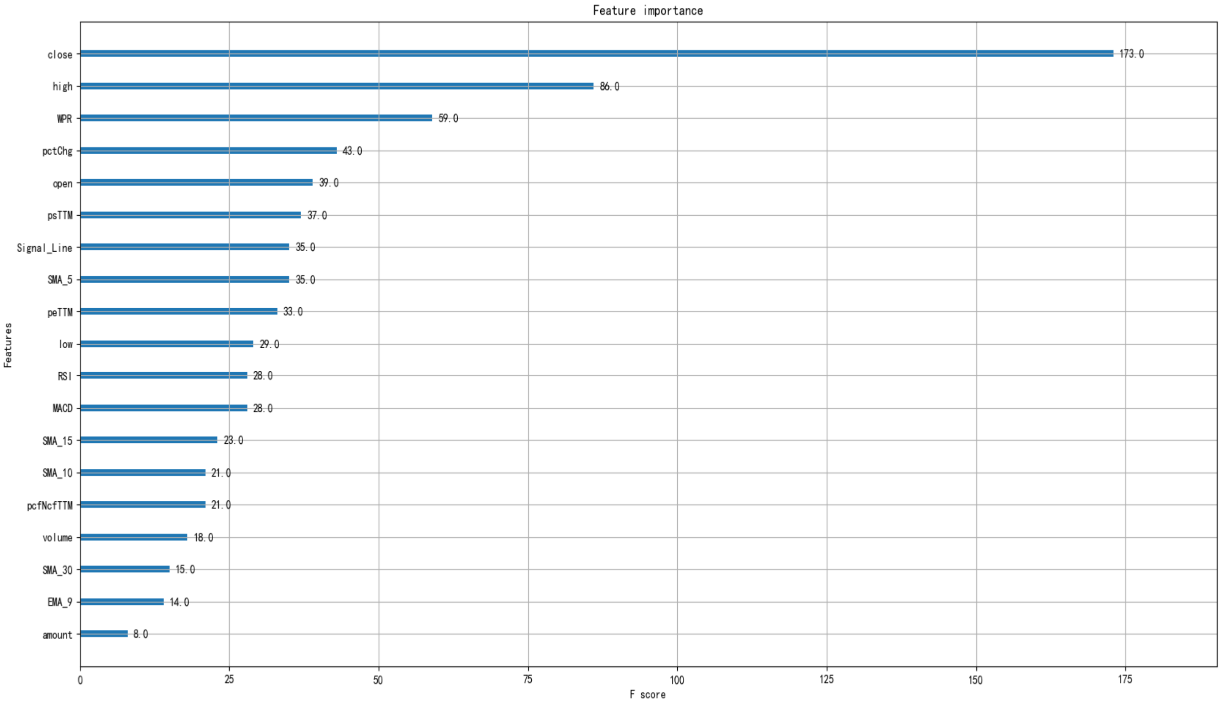
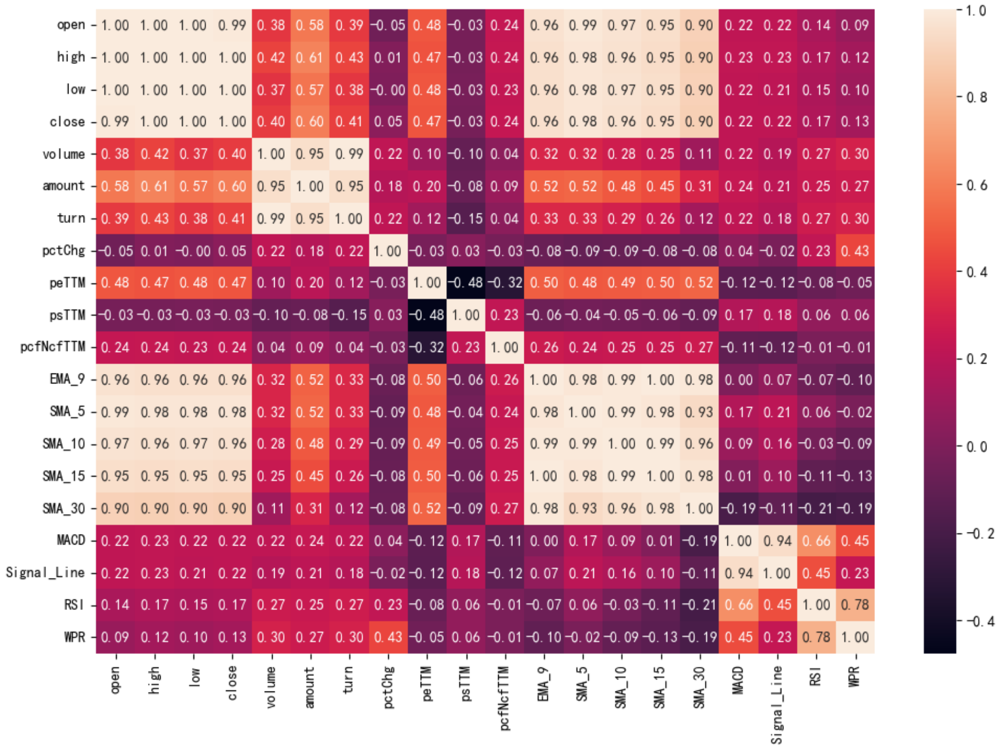


Figure 10. Ranking of feature importance



	open	high	low	close	volume	amount	turn	pctChg	peTTM	psTTM	pcfNcfTTM	EMA_9	SMA_5	SMA_10	SMA_15	SMA_30	MACD	Signal_Line	RSI	WPR
open	1.00	1.00	1.00	0.99	0.38	0.58	0.39	-0.05	0.48	-0.03	0.24	0.96	0.99	0.97	0.95	0.90	0.22	0.22	0.14	0.09
high	1.00	1.00	1.00	1.00	0.42	0.61	0.43	0.01	0.47	-0.03	0.24	0.96	0.98	0.96	0.95	0.90	0.23	0.23	0.17	0.12
low	1.00	1.00	1.00	1.00	0.37	0.57	0.38	-0.00	0.48	-0.03	0.23	0.96	0.98	0.97	0.95	0.90	0.22	0.21	0.15	0.10
close	0.99	1.00	1.00	1.00	0.40	0.60	0.41	0.05	0.47	-0.03	0.24	0.96	0.98	0.96	0.95	0.90	0.22	0.22	0.17	0.13
volume	0.38	0.42	0.37	0.40	1.00	0.95	0.99	0.22	0.10	-0.10	0.04	0.32	0.32	0.28	0.25	0.11	0.22	0.19	0.27	0.30
amount	0.58	0.61	0.57	0.60	0.95	1.00	0.95	0.18	0.20	-0.08	0.09	0.52	0.52	0.48	0.45	0.31	0.24	0.21	0.25	0.27
turn	0.39	0.43	0.38	0.41	0.99	0.95	1.00	0.22	0.12	-0.15	0.04	0.33	0.33	0.29	0.26	0.12	0.22	0.18	0.27	0.30
pctChg	-0.05	0.01	-0.00	0.05	0.22	0.18	0.22	1.00	-0.03	0.03	-0.03	-0.08	-0.09	-0.09	-0.08	-0.08	0.04	-0.02	0.23	0.43
peTTM	0.48	0.47	0.48	0.47	0.10	0.20	0.12	-0.03	1.00	-0.48	-0.32	0.50	0.48	0.49	0.50	0.52	-0.12	-0.12	-0.08	-0.05
psTTM	-0.03	-0.03	-0.03	-0.03	-0.10	-0.08	-0.15	0.03	-0.48	1.00	0.23	-0.06	-0.04	-0.05	-0.06	-0.09	0.17	0.18	0.06	0.06
pcfNcfTTM	0.24	0.24	0.23	0.24	0.04	0.09	0.04	-0.03	-0.32	0.23	1.00	0.26	0.24	0.25	0.25	0.27	-0.11	-0.12	-0.01	-0.01
EMA_9	0.96	0.96	0.96	0.96	0.32	0.52	0.33	-0.08	0.50	-0.06	0.26	1.00	0.98	0.99	1.00	0.98	0.00	0.07	-0.07	-0.10
SMA_5	0.99	0.98	0.98	0.98	0.32	0.52	0.33	-0.09	0.48	-0.04	0.24	0.98	1.00	0.99	0.98	0.93	0.17	0.21	0.06	-0.02
SMA_10	0.97	0.96	0.97	0.96	0.28	0.48	0.29	-0.09	0.49	-0.05	0.25	0.99	0.99	1.00	0.99	0.96	0.09	0.16	-0.03	-0.09
SMA_15	0.95	0.95	0.95	0.95	0.25	0.45	0.26	-0.08	0.50	-0.06	0.25	1.00	0.98	0.99	1.00	0.98	0.01	0.10	-0.11	-0.13
SMA_30	0.90	0.90	0.90	0.90	0.11	0.31	0.12	-0.08	0.52	-0.09	0.27	0.98	0.93	0.96	0.98	1.00	-0.19	-0.11	-0.21	-0.19
MACD	0.22	0.23	0.22	0.22	0.22	0.24	0.22	0.04	-0.12	0.17	-0.11	0.00	0.17	0.09	0.01	-0.19	1.00	0.94	0.66	0.45
Signal_Line	0.22	0.23	0.21	0.22	0.19	0.21	0.18	-0.02	-0.12	0.18	-0.12	0.07	0.21	0.16	0.10	-0.11	0.94	1.00	0.45	0.23
RSI	0.14	0.17	0.15	0.17	0.27	0.25	0.27	0.23	-0.08	0.06	-0.01	-0.07	0.06	-0.03	-0.11	-0.21	0.66	0.45	1.00	0.78
WPR	0.09	0.12	0.10	0.13	0.30	0.27	0.30	0.43	-0.05	0.06	-0.01	-0.10	-0.02	-0.09	-0.13	-0.19	0.45	0.23	0.78	1.00

Figure 11. Thermodynamic diagram of correlation coefficients of characteristic variables

Figure 11 shows the thermodynamic diagram of the correlation coefficients of 20 feature variables selected from the initial dataset. When using the LSTM model alone for prediction, we first use the correlation coefficient method for dimensionality reduction, which selects features based on the strength of the correlation between feature variables and label values. We selected five characteristic variables that have the strongest correlation with the next day's closing price, namely the closing price of the day (0.991), the highest price (0.989), the lowest price (0.988), the opening price (0.984), and the 5-day moving average of the simple moving average (0.969).

Figure 12 shows the variation curve of the loss function value during the training process of the LSTM model. It can be seen that the overall loss value of the training set (loss) and the overall loss value of the test set (val_loss) decrease synchronously, indicating that the model training is normal. Finally, loss and val_loss stabilized above 0 and around 0.2, respectively.

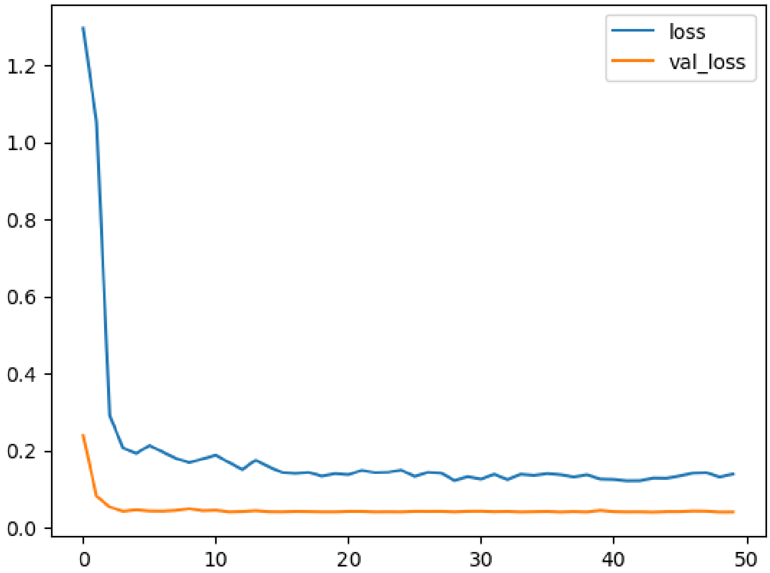


Figure 12. LSTM loss function value change curve

Figure 13 shows the comparison between the predicted and true values of the LSTM model on the test set. It can be seen from the figure that the LSTM model performs well in predicting both the overall trend of stock prices and local subtle changes, with better accuracy than the XGBoost model. However, its disadvantage is that the prediction lag is relatively high.

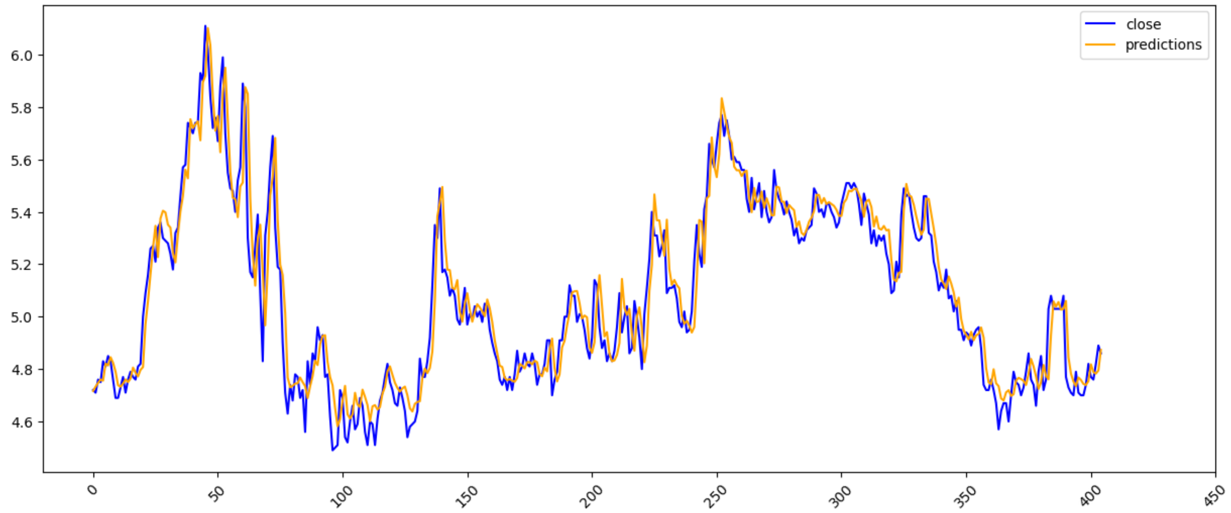


Figure 13. Predicted and true values of LSTM model

Figure 14 shows the variation curve of the loss function of the LSTM-XGBoost model during the training process. It can be seen that the overall loss value of the training set (loss) and the overall loss value of the test set (val_loss) decrease synchronously, indicating that the model training is normal. Finally, loss and val_loss is stable above 0 and around 0.1, respectively. Among them, the overall loss value of the test set of the combination model is smaller than that of the independent LSTM model, indicating that the former has a relatively better fitting effect on the test set.

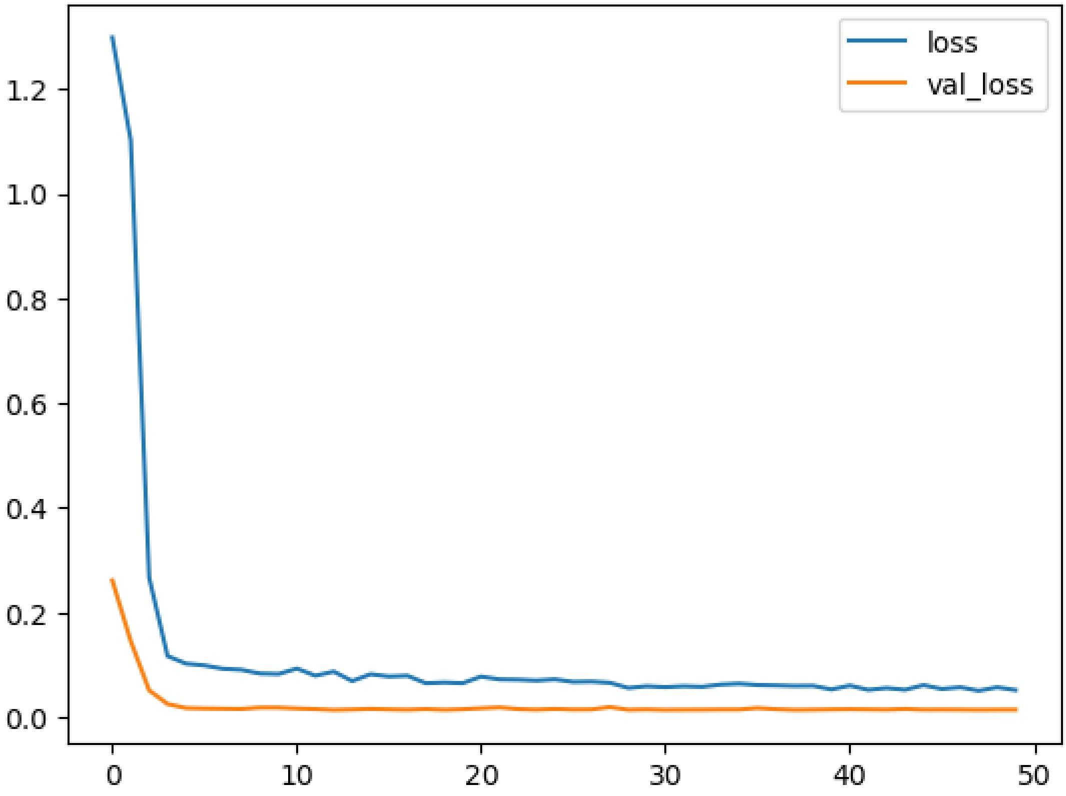


Figure 14. LSTM-XGBoost loss function value change curve

Figure 15 shows the comparison between the predicted and true values of the LSTM-XGBoost model on the test set. Similar to independent LSTM models, the combination model performs well in depicting overall trends and local changes. At the same time, the combination model also combines the advantage of low lag of XGBoost model prediction results, which has higher accuracy.

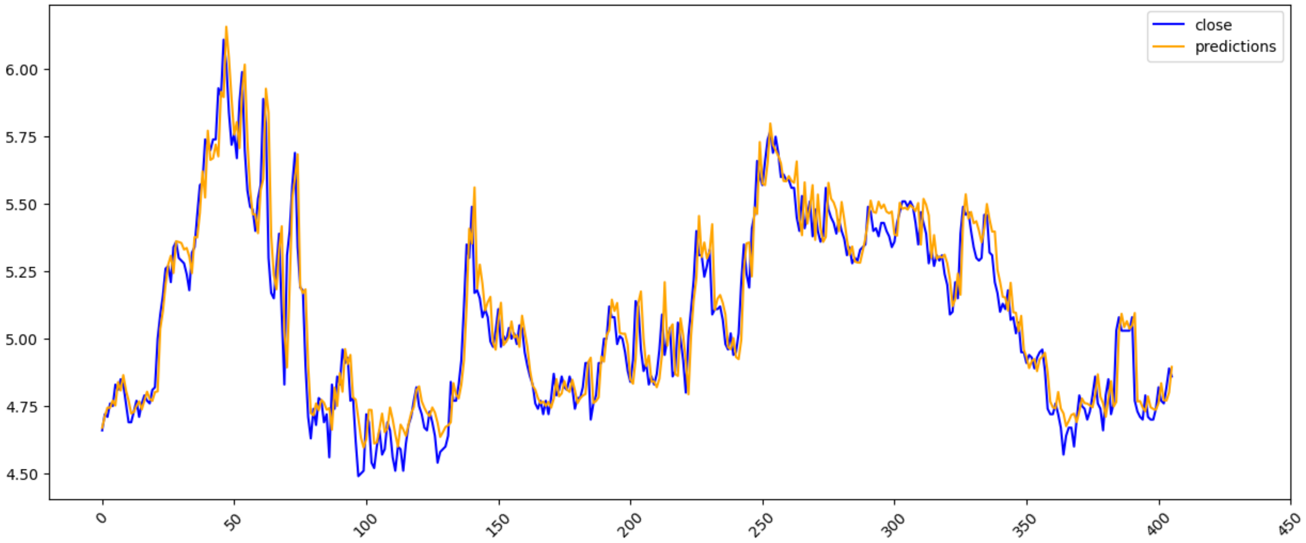


Figure 15. Predicted and True Values of LSTM-XGBoost Model

Table 3 lists the RMSE, RMAE, and MAPE of the three models, which can intuitively reflect the prediction quality of each model.

Table 3. Prediction performance of different models on the Eastern Airlines dataset

Model	RMSE	RMAE	MAPE%
XGBoost	0.1104	0.3316	2.31
LSTM	0.1094	0.2826	1.57
LSTM-XGBoost	0.1053	0.2762	1.49

According to the data in the table, the order of performance of the three models in prediction is as follows: LSTM-XGBoost model, LSTM model, and XGBoost model.

The LSTM-XGBoost model performs best, outperforming the XGBoost model and LSTM model in RMSE, RMAE, and MAPE. Compared to the XGBoost model and LSTM model, its RMSE index is 4.62% and 3.75% lower, its RMAE index is 16.71% and 2.26% lower, and its MAPE index is 0.82% and 0.08% lower. This also reflects the feasibility of the combination model proposed in this article.

The LSTM model ranks second in performance and outperforms the XGBoost model in all three indicators, demonstrating the superiority of the LSTM model in processing time series data and indirectly verifying the rationality of the combination model proposed in this paper.

5.2. Experimental Results on Kweichow Moutai Dataset

The experimental results obtained on a single data set may be random. In order to better study the performance of various models in stock price forecasting and verify the universality of portfolio models, we conducted experiments on the Kweichow Moutai data set. Figures 16 to 18 show the experimental results.

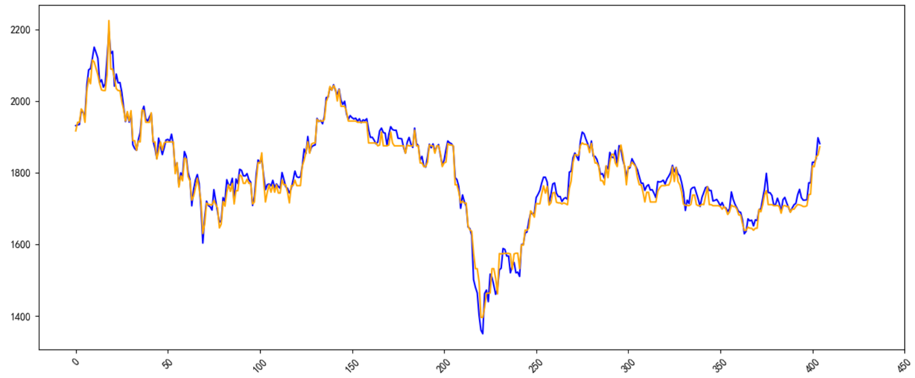


Figure 16. Predicted and true values of XGBoost model

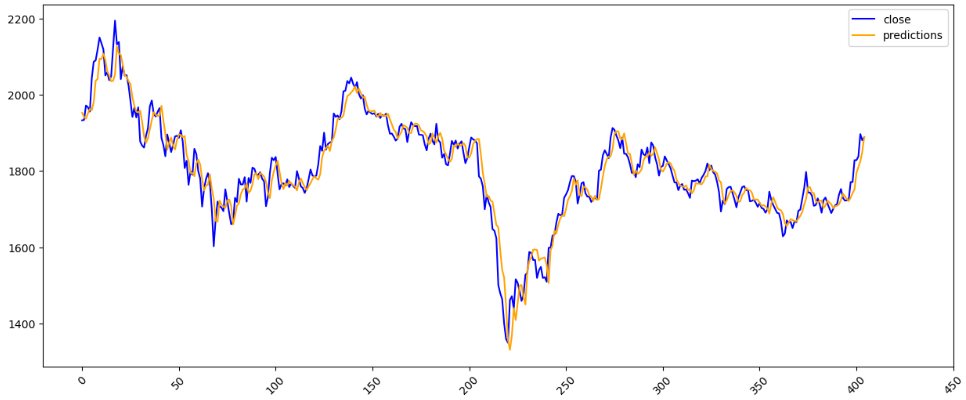


Figure 17. Predicted and true values of LSTM model

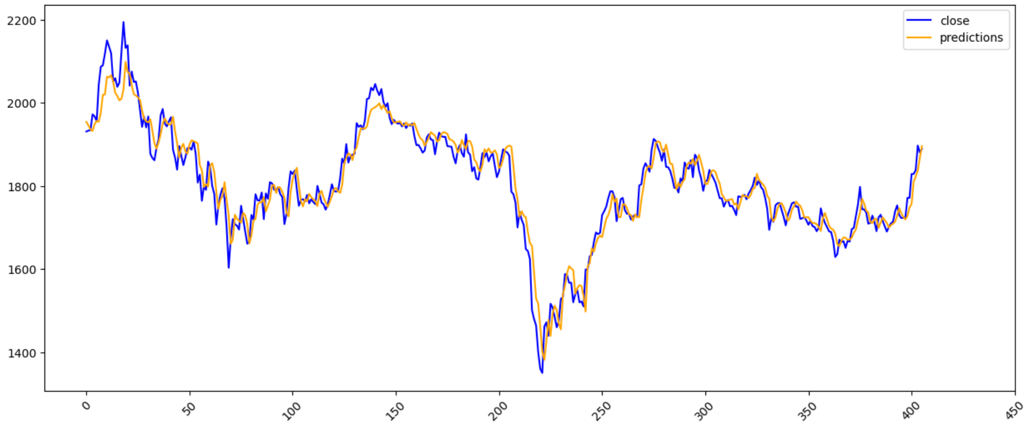


Figure 18. Predicted and true values of LSTM-XGBoost model

It can be seen from Figure 16 to Figure 18 that XGBoost's overall trend in Kweichow Moutai dataset is still good compared with the data set of China Eastern Airlines, but its local performance has deteriorated to a greater extent. LSTM still performs well in terms of overall trends and local performance, but the lag phenomenon is still evident. LSTM-XGBoost significantly optimizes the prediction time delay while maintaining both good overall and local performance, resulting in optimal prediction accuracy.

Table 4 lists the RMSE, RMAE, and MAPE of the three models, which can more intuitively reflect the prediction quality of each model.

Similar to the results on the Eastern Airlines dataset, the order of performance of the three models in prediction is still LSTM-XGBoost model, LSTM model, and XGBoost model. Among them, the RMSE index of the LSTM-XGBoost combination model is 9.83% and 9.45% lower than that of the XGBoost model and LSTM model, respectively. The RMAE index is 18.43% and 7.91% lower, while the MAPE index is 0.58% and 0.26% lower.

Table 4. Prediction performance of different models on Kweichow Moutai dataset

Model	RMSE	RMAE	MAPE%
XGBoost	41.2263	6.4130	2.12
LSTM	41.0517	5.6808	1.80
LSTM-XGBoost	37.1719	5.2313	1.54

5.3. Experimental Results on the China Merchants Bank Dataset

The experimental results of the three models on the China Merchants Bank dataset are shown in Figures 19 to 21.

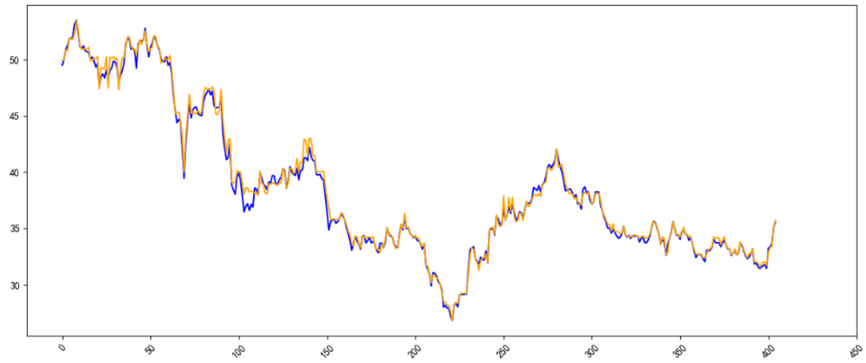


Figure 19. Predicted and true values of XGBoost model

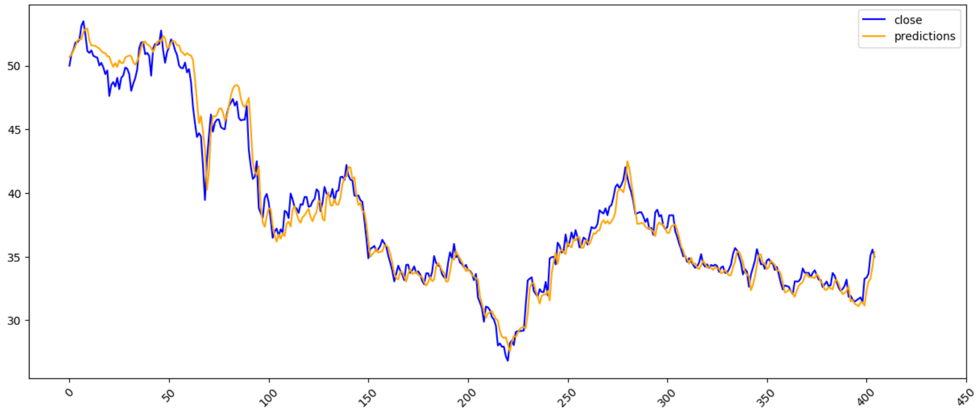


Figure 20. Predicted and true values of LSTM model

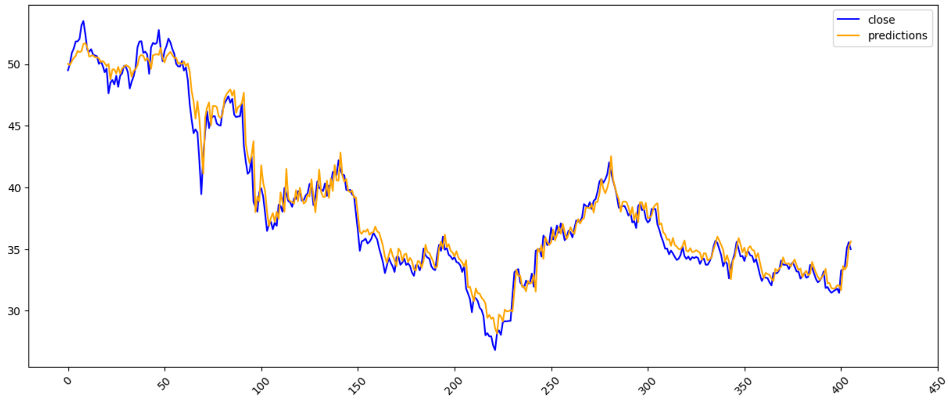


Figure 21. Predicted and true values of LSTM-XGBoost model

As shown in Figures 19 to 21, the performance of the three models on the data set of China Merchants Bank is similar to that on the data sets of China Eastern Airlines and Kweichow Moutai, among which the combination model still shows good prediction performance.

At the same time, observing Table 5, it can be intuitively seen that the order of advantages and disadvantages of the three models is still LSTM-XGBoost model, LSTM model, and XGBoost model. Among them, the RMSE index of the LSTM-XGBoost combination model is 5.09% and 2.38% lower than that of the XGBoost model and LSTM model, respectively. The RMAE index is 5.08% and 1.01% lower, while the MAPE index is 0.09% and 0.02% lower.

Table 5. Prediction performance of different models on China Merchants Bank dataset

Model	RMSE	RMAE	MAPE%
XGBoost	0.9297	0.8497	1.79
LSTM	0.9039	0.8147	1.72
LSTM-XGBoost	0.8824	0.8065	1.70

6. Conclusion

The LSTM-XGBoost combination model proposed in this article combines the advantages of LSTM in processing time series data and the advantages of XGBoost in evaluating the importance of features. In the actual prediction, we applied the combination model and the separate XGBoost and LSTM models to the three data sets of China Eastern Airlines, Kweichow Moutai and China Merchants Bank, and found that compared with the separate XGBoost and LSTM models, the combination model has better performance in predicting the overall trend, local changes and hysteresis. Meanwhile, the relatively uniform performance of the combined model on the three datasets also reflects the strong robustness of the model. At the same time, the usage scenario of this model is not limited to stock price prediction, but is applicable to various types of data that involve multiple characteristic variables and have time series characteristics.

In addition, we should note that when stocks experience significant fluctuations, there is still a significant gap between the predicted results of the model and the true values. This may be because the price fluctuations of stocks are not only related to market indicators, financial indicators, and technical indicators, but also closely related to current affairs news, investor sentiment, and related stock trends. Therefore, more relevant factors can be quantified and input into the model in the future, which may improve the prediction accuracy of the model.

References

[1] Chen, T. and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 785-794.

[2] Dezhkam, A. and Manzuri, M. T., 2023. Forecasting stock market for an efficient portfolio by combining XGBoost and Hilbert–Huang transform. Engineering Applications of Artificial Intelligence, vol. 118, pp. 1–13.

[3] Fang, Y., Lu, Z. and Ge, Y., 2022. Stock price prediction for the LSTM-CNN model with joint RMSE losses. Computer Engineering and Applications, 58(9), pp. 294-302.

[4] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp. 1735-1780.