BYD Stock Price Prediction Based on LSTM and XGBoost Models

Sihang Chen

{a1988911@adelaide.edu.au}

Department of Commerce, Adelaide University, Adelaide, SA 5000, Australia

Abstract. The stock performance of BYD (002594.SZ) is attracting increasing attention. This research aims to use eXtreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) to predict BYD's next trading day's closing price and compare their performances. The research used BYD stock data from 2015 to 2025. The research constructed close, volume, 5-day and 20-day simple moving averages (MA5, MA20) and Relative Strength Index (RSI) as lagged features, and used Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R2 as evaluation indicators. The result was the LSTM model had better performance compare to the XGBoost model on the dataset from 2015 to 2025 and the dataset from 2015 to 2024. The conclusion is XGBoost trains faster and has stable performance while LSTM has better generalization stability and usually performs better. The framework of this research could also be used to other stocks' prediction.

Keywords: LSTM, XGBoost, Stock prediction, Machine learning, Financial time series.

1 Introduction

Stock markets are full of noise and uncertainty, but innumerable people want to figure out the markets fluctuation pattern and predict the ups and downs of stock prices. Stock prediction is a hard problem in machine learning because stock prices are influenced by many factors and stock markets are driven by unpredictable news, sentiment shifts, and exogenous shocks. Stock prices may even be influenced by prediction itself. Many researchers have already done research into the problem of forecasting stock price. They have developed many different methods, including LSTM, Bi-LSTM, XGBoost, Light Gradient Boosting Machine (LightGBM), Large Language Model (LLM), Convolutional Neural Network (CNN) and etc [1, 2, 3]. Zhang and Fan used XGBoost model to sort the features and help build LSTM model [3]. And Hu and Ye scored stocks by combining the predicted price change rate in a week with the predicted price change rate for the next trading day [4]. In Usmani and Shamsi's research, LSTM based Weighted and Categorized News Stock prediction model (WCN-LSTM) is proposed and the performance is better than traditional models [5]. Liu et al.'s article raised LSTMs with an attention mechanism to analyze the fundamental financial ratios of companies and the MAE of the proposed LSTM with attention model outperforms other methods [6]. Baek's study proposes a Genetic Algorithm optimization-based deep learning technique

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365107

(CNN-LSTM) that predicts the next day's closing price based on an artificial intelligence model to more accurately predict future stock prices [7]. López and Marcos’s empirical research indicates that in financial forecasting, modern Gradient Boosting Decision Tree (GBDT) models such as XGBoost and LightGBM are comprehensively superior to AdaBoost [8]. BYD is a world-leading Chinese new energy vehicle company, and it became the world's top-selling electric vehicle brand in 2025. BYD stock is attracting increasing attention and the stock price has also risen sharply in recent years. Although BYD stock price is supported by fundamentals of BYD’s actual business operations, BYD stock price still fluctuates sharply. Sometimes the ups and downs of the stock price can be very confusing, so this research aims to use XGBoost and LSTM models to forecast the next trading day’s closing price of BYD (002594.SZ) and compare their performances. XGBoost and LSTM had been chosen because they are popular and efficient ways to predict the stock prices. Both models are good at predicting time series data. This research provides a specific example of stock price prediction. The conclusion of this research can help identify the strengths and weaknesses of XGBoost and LSTM when using them to predict stock prices, and help improve investors return rate and reduce the risks.

2 Methodology

2.1 Dataset Introduction

The BYD stock dataset was obtained from akshare using python code. The dataset contains 7 columns (date, open, close, high, low, volume, amount), it covers 2664 days stock data of BYD (002594.SZ) from Jan 1, 2015 to Dec 31, 2025. The research was based on Windows 11 and Python 3.12 using Jupyter Notebook 7.4.5. The origin volume column is measured in lots (100 shares), which was transformed into shares as the unit.

2.2 Models Introduction

LSTM. LSTM is a specialized Recurrent Neural Network (RNN) architecture, first proposed by Sepp and Jürgen in 1997 [9]. It is designed to address the vanishing gradient problem and exploding gradient problem that plague traditional RNNs when processing long sequences of data.

Unlike standard RNNs that rely on simple recurrent connections, LSTMs incorporate a unique gated cell structure to control the flow of information. The core components of an LSTM unit include Forget Gate, Input Gate, Output Gate and Cell State.

LSTMs excel at processing and predicting data with long-range temporal dependencies. They have become the backbone of many sequence-based tasks in Natural Language Processing (NLP) and time series analysis, such as machine translation, text generation, speech recognition, stock price forecasting, and sentiment analysis. Compared to traditional RNNs, LSTMs can effectively learn patterns from sequences with hundreds or even thousands of time steps, making them one of the most widely used deep learning models for sequential data.

XGBoost. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.

As an extension of the gradient boosting framework, XGBoost introduces several key enhancements to address the limitations of traditional Gradient Boosting Machines (GBMs). Its core principle is to build an ensemble of weak prediction models (typically decision trees) in a sequential manner, where each new tree corrects the errors of the previous ones by minimizing a regularized loss function.

XGBoost has become one of the most popular machine learning algorithms in data science competitions (such as Kaggle) due to its high performance and scalability. It is widely used in real-world scenarios, including credit risk assessment, click-through rate (CTR) prediction, fraud detection, recommendation systems, and time series forecasting.

XGBoost is faster in training on small and medium-sized structured data, more interpretable, and less prone to overfitting, while LSTM is good at capture long-term dependencies on large-scale sequence data, but it has high training costs and is a "black box" model.

2.3 Evaluating Indicators

The research used MAE (see equation (1)), RMSE (see equation (2)) and R^2 (see equation (3)) as evaluating indicators.

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

MAE measures the average of the absolute differences between actual and predicted values [10].

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

RMSE is the square root of Mean Square Error (MSE) (the average of the square differences between actual and predicted values), representing the error in the same units as the actual observations, making it easier to interpret [10].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \ . \tag{3}$$

R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is often used to evaluate the goodness of fit of a model. The closer the R^2 value is to1, the more effectively the model can predict the data [10].

2.4 Process

The first step is dataset importing, which involves loading the dataset into the notebook. The second step is Exploratory Data Analysis (EDA). This research drew Closing Price and Volume Time Series Plot (Figure 1), Daily Log Returns Time Series (Figure 2), and Feature Correlation Heatmap (Figure 3).

The upper graph shows the trend of BYD's closing price from 2015 to 2025. From 2015 to 2020, the price kept in a low level, generally not exceeding 20 yuan. And after 2020, the price rose quickly and fluctuated from lower than 50 yuan to higher than 130 yuan. The lower graph shows the trading volume of BYD stock. The overall pattern presents a low baseline + high peak pulse trend. Transaction activity has significantly increased since 2020, and there was an extreme surge in volume from 2025 to 2026, indicating a sharp rise in market attention.

Fig. 1. Closing Price and Volume Time Series Plot. (Picture credit: Original)

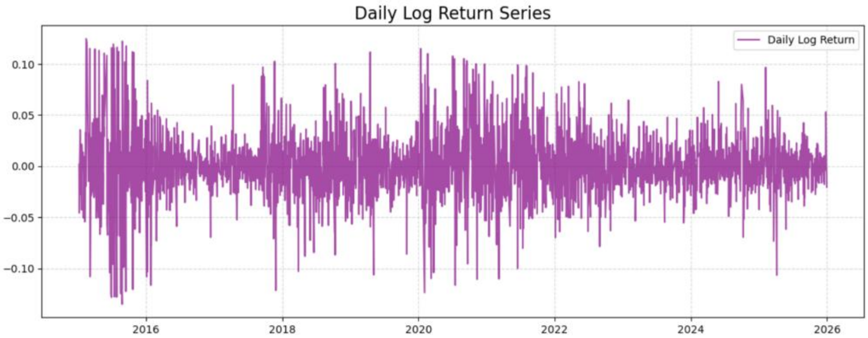

Fig. 2. Daily Log Returns Time Series. (Picture credit: Original)

Figure 2 demonstrates the daily return rate of BYD stock. The overall pattern shows profit always comes with risks and the average daily return rate approaches zero.

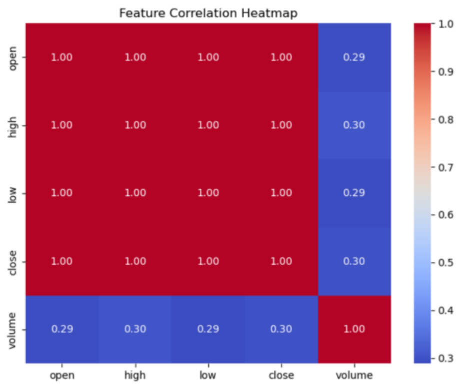

Fig. 3. Feature Correlation Heatmap. (Picture credit: Original)

The third step is feature engineering. This research constructed 5 widely used technical indicators as lagged features, including close, volume, MA5, MA20 and RSI. MA5 and MA20 are calculated as the arithmetic mean of the closing prices over the respective windows. RSI is computed following Wilder's research with a 14-day lookback period [11]. Feature Correlation Heatmap for Lagged Features is shown on Figure 4.

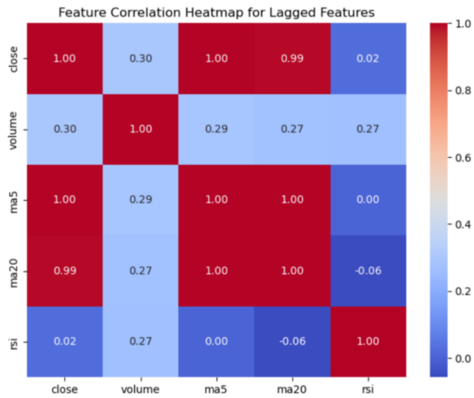

Fig. 4. Feature Correlation Heatmap for Lagged Features. (Picture credit: Original)

The fourth step is training and evaluating the XGBoost model. The procedure includes constructing lagged features, splitting training/test set by time (80%:20%), training the model, applying the model to predict next trading day's closing price, evaluating the performance, and tuning parameters.

The fifth step is training and evaluating the LSTM model. The procedure includes constructing lagged features, splitting training/test set by time (80%:20%), normalization, constructing time series sequences with a sliding window equals to 20 days, building LSTM model, training the model, applying the model to predict next trading day's closing price, denormalization, evaluating the performance, and tuning parameters.

The sixth step is result comparison and visualization. This research drew the Model Performance Comparison Table (Table1), XGBoost: Predicted vs True Closing Price on Training Set (Figure 5), LSTM: Predicted vs True Closing Price on Training Set (Figure 6), Model Predictions vs True Closing Price (last 200 Trading days) (Figure 7), Model Predictions vs True Closing Price (Full Test Set) (Figure 8), XGBoost Feature Importance (Gain)(Top 10) (Figure 9).

Table 1. Model performance comparison.

Model	Dataset	MAE	RMSE	R^2
XGBoost	Train	0.8585	1.3042	0.9984
XGBoost	Test	4.0501	6.3319	0.8975
LSTM	Train	1.3951	2.2182	0.9953
LSTM	Test	2.7644	3.8495	0.9626

Fig. 5. XGBoost: Predicted vs True Closing Price on Training Set. (Picture credit: Original)

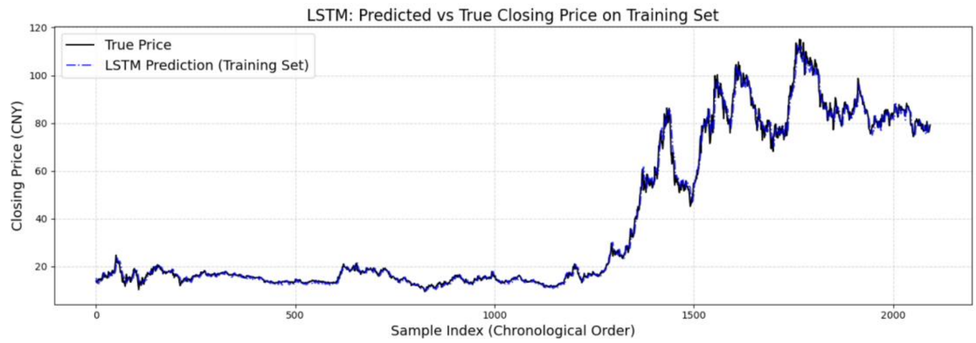

Fig. 6. LSTM: Predicted vs True Closing Price on Training Set. (Picture credit: Original)

The two models both performed well on the training set, the XGBoost model performed a bit better.

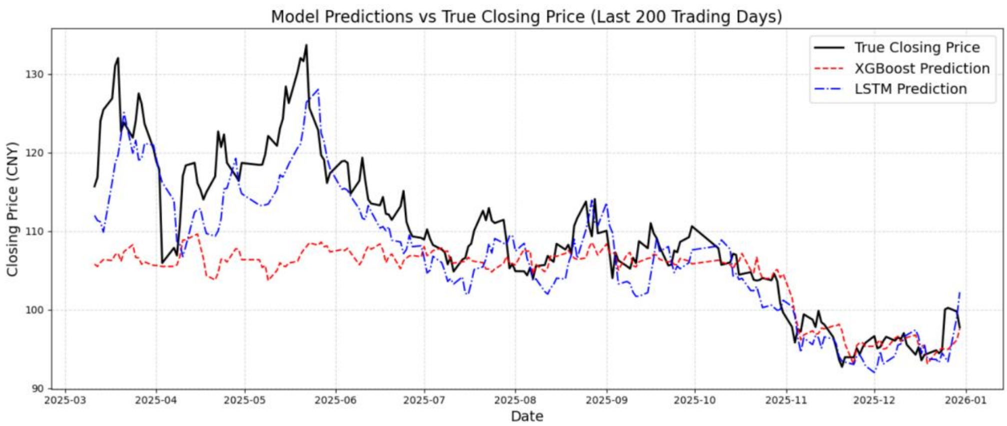

Fig. 7. Model Predictions vs True Closing Prices (last 200 Trading days). (Picture credit: Original)

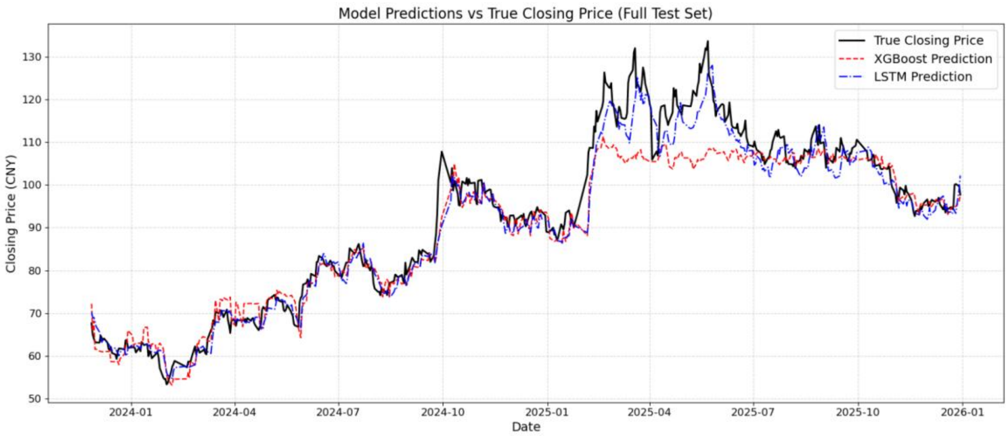

Fig. 8. Model Predictions vs True Closing Prices (Full Test Set). (Picture credit: Original)

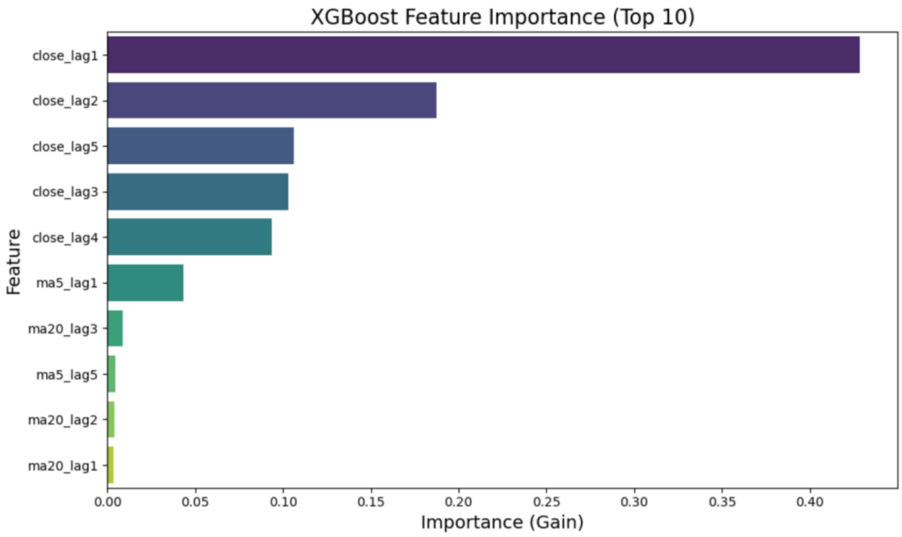

Fig. 9. XGBoost Feature Importance (Gain) (Top 10). (Picture credit: Original)

Figure 9 indicates that closing prices are the most powerful signal and short-term lags are more important than long-term lags.

3 Results

The XGBoost model achieved lower MAE, RMSE and higher R2 on the training set compare to the LSTM model, but the LSTM model got lower MAE, RMSE and higher R2 on the test set. That means although the XGBoost model performs better on the training set, the LSTM model performs better on the test set. The LSTM model seems performed well on the full test set, while the XGBoost model seems worked well most of the time but performed bad from March 2025 to September 2025 on the test set. When the proportion of the training set changed or more features were brought into the models, like MACD, the result had barely improved.

When the stock data of 2025 was removed, the same models showed a similar performance result. the XGBoost model performs better on the training set while the LSTM model performs better on the test set. Thus, the LSTM model had better performance than the XGBoost model no matter the stock data of 2025 was removed or not. The Model Performance Comparison and visualization are shown in Table 2 and Figure 10 respectively.

Table 2. Model performance comparison (2015-2024).

Model	Dataset	MAE	RMSE	R^2
XGBoost	Train	0.7751	1.1913	0.9985
XGBoost	Test	2.3506	3.0746	0.9187
LSTM	Train	1.6402	2.5032	0.9933
LSTM	Test	2.1463	2.8529	0.9289

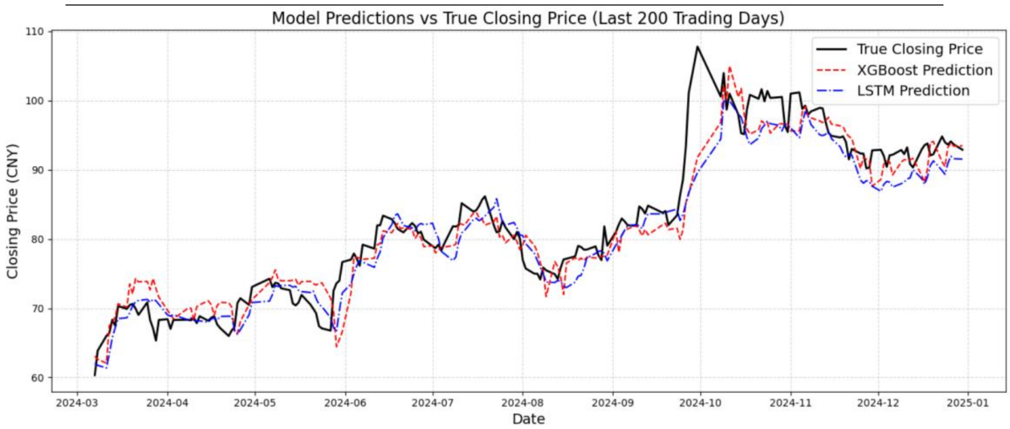

Fig. 10. Model Predictions vs True Closing Prices (last 200 Trading days) (2015-2024). (Picture credit: Original)

4 Conclusion

This research compared XGBoost and LSTM and the results showed LSTM has better generalization stability and usually performed better, but XGBoost trains faster and the performance is more stable, the performances between multiple runs keep the same. Although both models have the ability to predict future stock prices, they still could not perform as well as on the training set, and sometimes had a delayed reaction compare to the true closing price. This delay precisely reflects the inherent difficulty of stock price prediction. Because models can only react to past information, not anticipate the future. Although this research used and tuned LSTM and XGBoost to predict BYD stock price and achieved some results, but the research only used the stock data to predict the next trading day’s closing price, lack of financial data and market sentiments. In the future, the directions can be importing more types of data, continuing to optimize the models, and validating the models on other stocks.

References

[1] Chen, L. J., Pan, Z. D., Chen, Q.: Research on Stock Price Prediction Framework Based on Large Language Model. Practice and Cognition of Mathematics, 55(12): 159-171. (2025) DOI: 10.20266/j.math.24-1224.

[2] Liu, C., Feng, G.: CNN-BiLSTM and LightGBM Stock Prediction Based on Dual Attention Mechanism. Computer and Modernization, (09): 97-103. (2025)

[3] Zhang, X. T., Fan, Y. S.: Stock Trend Evaluation and Analysis Using LSTM Model Based on XGBoost Feature Selection. Computer Knowledge and Technology, 19(09): 91-94+97. (2023) DOI: 10.14004/j.cnki.ckt.2023.0403.

[4] Hu, Z. B., Ye, Z., Ge, J.: Research on financial time series prediction based on multi-task learning. Journal of South-Central University for Nationalities (Natural Science Edition), 1-9. (2025) DOI: 10.20056/j.cnki.ZNMDZK.20250849.

[5] Usmani, S. S., & Shamsi, J. A.: LSTM based stock prediction using weighted and categorized financial news. PLOS ONE, 18(3), e0282234. (2023) DOI: 10.1371/journal.pone.0282234.

[6] Liu, K.-H., Lin, Y.-S., Liu, T.-J., & Tai, W.-S.: StockQM: A Cross-Frequency Dataset for Stock Prediction and a New Stock Prediction Model. 2024 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), Taichung, Taiwan, pp. 275-276. (2024) DOI: 10.1109/ICCE-Taiwan62264.2024.10674208.

[7] Baek, H.: A CNN-LSTM Stock Prediction Model Based on Genetic Algorithm Optimization. Asia-Pacific Financial Markets, 31(2), 205–220. (2024) DOI: 10.1007/s10690-023-09412-z.

[8] López de Prado, M.: Advances in Financial Machine Learning. Wiley. (2018)

[9] Hochreiter, S. & Schmidhuber, J.: Long Short-Term Memory. Neural Computation. (1997) 9. 1735-1780. 10.1162/neco.1997.9.8.1735.

[10] Bao, W. Z., Cao, Y. T., Yang, Y., Che, H. J., Huang, J. J., & Wen, S. P.: Data-driven stock forecasting models based on neural networks: A review. Information Fusion, 113, 102616. (2025) DOI: 10.1016/j.inffus.2024.102616.

[11] Wilder, J. W.: New Concepts in Technical Trading Systems. Greensboro: Trend Research. (1978)