Empirical Study on BYD Stock Price Prediction Based on LSTM Model

Xiangrui Cheng[1,a,*]

[1]Tiangong University, No. 399 Binshui Road, Tianjin, China
a. 2966384497@qq.com
**corresponding author*

Abstract: This study employs Long Short-Term Memory (LSTM) neural network models for stock price prediction. LSTM models excel in managing nonlinearities and time-series dependencies in data, thereby enhancing the accuracy of stock price forecasts, reducing investment risks, and increasing returns. Investors can reasonably use the LSTM model to predict stock prices and obtain greater returns. The research utilizes eight years of historical data from BYD, including opening prices, closing prices, highest prices, lowest prices, and trading volumes, as training data. By adjusting various hyperparameters such as epochs, learning rates, and training-test ratios, the study analyzes their impact on the predictive accuracy of the model and identifies optimal configurations through comparative experimental methods. We found that the best parameters were when epoch number was 25, training ratio was 90:10 (minimum mean square error was 0.00175572) and 4 years of training data was selected as the training set (minimum mean square error was 0.00146477).

Keywords: LSTM, stock price prediction, RNN, neural networks.

1. Introduction

With the rapid development of China's financial market, stocks have become a "barometer" of the market economy, attracting significant attention from investors and researchers. The volatility of stock prices makes accurate predictions challenging, and traditional statistical models have limitations. In recent years, with advances in artificial intelligence and machine learning, neural network models have emerged as a prominent research focus in the financial sector. This study aims to explore the application of the Long Short-Term Memory (LSTM) neural network model for stock price prediction.

The LSTM model excels in processing time series data and can achieve higher accuracy in handling nonlinear and serial correlation problems. This addresses the limitations of traditional machine learning algorithms in dealing with the time series correlations of stock data, making LSTM particularly advantageous for stock price prediction. By building and training LSTM models, we can accurately predict stock prices and provide valuable insights for investors, aiding them in making informed investment decisions, reducing risks, and enhancing returns.

The main body of this paper is divided into six parts. The first part outlines the background of stock price prediction, emphasizing the unique value and significance of the LSTM model for this task. The second part reviews various studies on stock price forecasting using the LSTM model, detailing their methodologies, subjects, and conclusions. The third part describes the experimental design, including data sources, preprocessing, model evaluation criteria, and LSTM structural analysis. The fourth part presents the experimental results using tables and figures. The fifth part discusses the experimental results and the conclusions drawn from the study. The final part summarizes the overall conclusions, addresses the limitations of the research, and proposes potential solutions.

2. Literature review

In recent years, the application of the Long Short-Term Memory (LSTM) network in predicting stock prices has garnered considerable attention. In 2018, Wang Jun et al. introduced an attention mechanism integrated into the traditional Seq2Seq model and conducted a 5-day prediction experiment using data from 50 constituent stocks of the Shanghai Stock Exchange spanning 2015 to 2017 [1]. The experimental findings demonstrated that this method enhances prediction accuracy by at least 3 percentage points compared to alternative approaches, particularly for predicting uncertain long-term series.

Subsequently, in 2021, Huang Yucheng et al. further refined the traditional Seq2Seq model by varying time series lengths and utilized the dataset of stocks from the Shanghai Stock Exchange, code-named 002, covering January 5, 1998, to June 2, 2020 [2]. Their investigation identified optimal prediction outcomes when the time series length parameter approached i = 0.01n, highlighting the LSTM model's superior predictive capabilities in their experimentation.

In 2022, Lin Xin et al. integrated an Attention mechanism into the LSTM model, employing the Shanghai Stock Exchange Industrial Index and the Shanghai Stock Exchange Environmental Protection Index as representative datasets. They conducted daily predictions from January 2, 2014, to September 22, 2020, and compared the performance against MLP, RNN, and traditional LSTM models, affirming that the AM-LSTM model achieved the highest prediction accuracy [3].

Shortly thereafter, Tao Yongkang et al. proposed an LSTM model incorporating an attention mechanism in 2023, focusing on stock forecasting within the banking sector. Through their research, they determined an optimal time step of 10 and verified through comparative analysis with standard LSTM models that LSTM-ATT exhibited superior predictive accuracy [4].

Huang Chaobin et al. utilized the LSTM neural network to predict the Shanghai Composite Index, benchmarking results against BP neural networks, CNN models, RNN models, and GRU neural networks. Their study, based on over 7,000 sample data points from specific stocks within the Shanghai Composite Index spanning 1990 to 2019, and utilizing 11 feature dimensions, split the dataset into a training set and a test set at a ratio of 7:3. Their findings underscored LSTM's robust predictive performance [5].

Additionally, Li Liping et al. compared LSTM, BP, and Elman models, examining the influence of neural network hidden layers. Their experimental outcomes demonstrated that LSTM neural networks achieved smaller MAE and RMSE values compared to the other models, affirming superior prediction accuracy [6].

Zhu Wenchao's comparison of RNN, LSTM, and GRU models utilized data from Changjiang Securities spanning January 1, 2008, to December 31, 2021, across 9 variables. Their study, under conditions of 3 hidden layers and 1 fully connected layer, revealed that the GRU model exhibited the most favorable fitting effects [7].

Exploring the combination of breadth learning (BLS) and deep learning, Han Ying et al. introduced complementary ensemble empirical mode decomposition (CEEMD) for noise reduction, addressing the volatility inherent in stock sequences. Their study focused on agricultural, forestry, animal husbandry, and fishery stock prices, demonstrating substantial improvements over baseline and existing models in multiple accuracy metrics [8]. Specifically, their model effectively mitigated issues such as poor fitting and time lag during periods of significant data fluctuation.

3. Methods

3.1. Study Design

This study employs the Long Short-Term Memory (LSTM) network model to predict stock price changes. By adjusting various hyperparameters such as epochs, learning rate, and training ratio, the study evaluates their impact on the model's prediction accuracy. An experimental comparison method is used to assess model performance under different configurations and identify the optimal setup.

3.2. Data Collection

The dataset comprises historical trading data for BYD (002594) stocks on the Chinese mainland stock exchange, spanning from November 2016 to November 2023. This includes daily opening price, closing price, highest price, lowest price, and trading volume, covering an eight-year period.

3.3. Data Preprocessing and Fundamental Settings

Data preprocessing involves several key steps:

3.3.1. Handling Missing Values

Missing values are filled using interpolation methods.

3.3.2. Feature Selection

Features with high correlation, such as opening price, closing price, and trading volume, are selected.

3.3.3. Data Normalization

Data is normalized to the range [0, 1] to expedite model convergence.

Fundamental settings include:

Loss Function: Mean square error (MSE) is used as the loss function.

Optimizer: The Adam optimizer is employed.

Hyperparameter tuning includes:

Epochs: Set to 20, 25, 30, 40, and 50 for experimental comparison.

Learning Rate: Fixed at 0.001.

Training Ratio: Varied at 75:25, 80:20, 85:15, and 90:10 for experimental comparison.

3.4. Model Evaluation

Root mean square error (RMSE) is used as the evaluation metric.

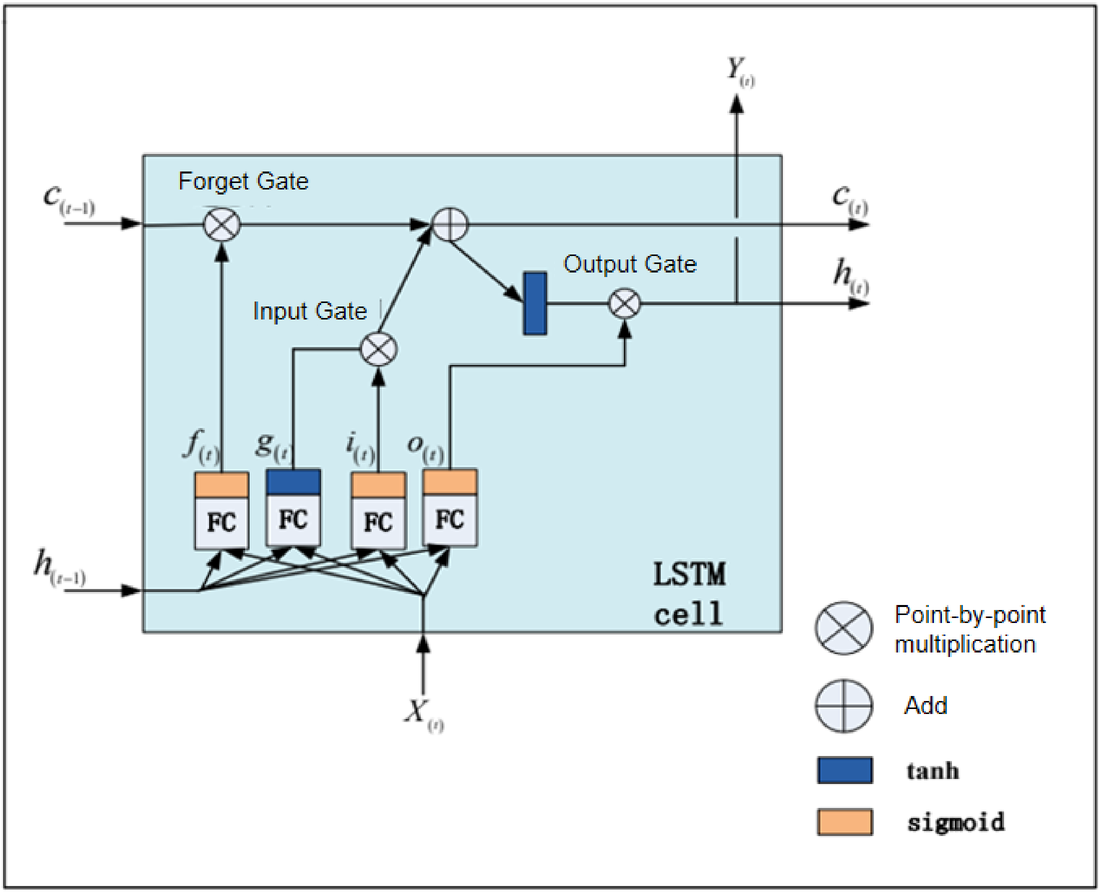


Figure 1: Memory cell of LSTM

3.5. LSTM structure analysis

3.5.1. Input data

The input to the LSTM unit includes the current time step's input data and the previous time step's output value.

3.5.2. Forget Gate

The forget gate determines the proportion of information from the previous time step's memory unit state that needs to be forgotten. It calculates a forget ratio based on the current time step's input data and the previous time step's output value. This ratio is applied to the previous time step's memory unit state to decide which information to retain and discard.

3.5.3. Input Gate

The input gate determines the amount of new information from the current time step to be stored in the memory unit. It calculates an input ratio using the current time step's input data and the previous time step's output value. Additionally, the LSTM unit generates a new candidate memory state, which, together with the input gate's output, determines the new information to be added to the memory unit.

3.5.4. Cell State Update

The current time step's cell state comprises two parts: information from the previous time step's cell state adjusted by the forget gate, and new information from the current time step adjusted by the input gate. These parts combine to form the current time step's cell state.

3.5.5. Output Gate

The output gate determines the amount of information from the current time step's cell state to be output. It calculates an output ratio based on the current time step's input data and the previous time step's output value. This ratio is applied to the current time step's cell state to decide which information will be passed to the output value.

3.5.6. Output Value

The current time step's output value is derived from the output gate and the current time step's cell state. The output gate adjusts the cell state information to ensure that the final output value includes the necessary information for the current time step.

4. Results

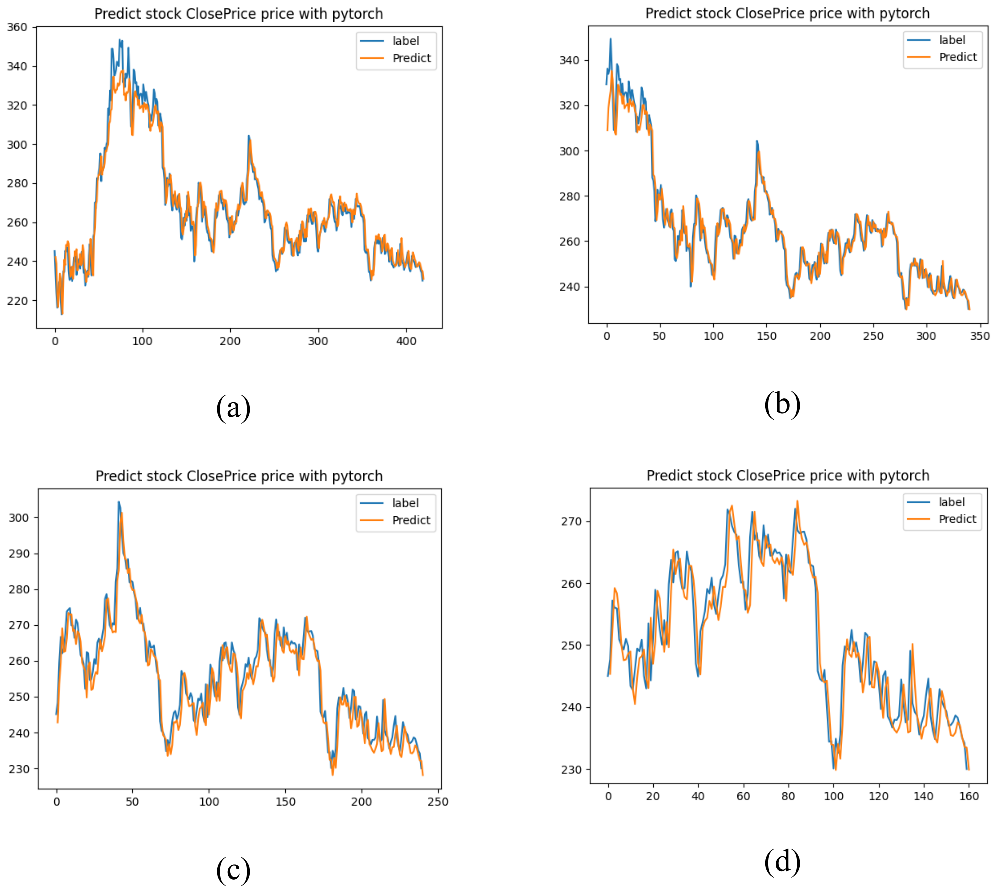
Based on the aforementioned experimental conditions, this study designed and executed three sets of control experiments.

Firstly, utilizing the LSTM model, adjustments were made to the number of training epochs to determine the point at which the model achieves saturation in sample detection. For this experiment, epoch numbers were tested within the range of 20 to 40. It was determined that the optimal training effect occurred at 25 epochs.

Table 1: presents the Mean Squared Error (MSE) of the model across various epochs.

Epoch	MSE
20	0.00173421
25	0.00164962
30	0.00168809
40	0.00167847

Secondly, adjustments were made to the training-test sample ratio, and prediction experiments were conducted under four conditions: 75:25, 80:20, 85:15, and 90:10. The results are illustrated in the subsequent figure.



(a)

(b)

(c)

(d)

Figure 2: Comparison of model predictions for different training-test sample ratios 75:25; (b)80:20; (c)85:15; (d)90:10.

The errors of the four control experiments were recorded as shown in the following table

Table 2: The MSE size of the model for different training-test sample ratios

Train-test ratio	75:25	80:20	85:15	90:10
RMSE	0.00447932	0.0033114	0.00215872	0.00175572

Finally, using the last quarter of the control sample set as the test set, several past years were used as training sets in eight control experiments. The numerical results are presented in the following table.

Table 3: MSE size of the model when the training set span is different

Training set span (unit: year)	1	2	3	4	5	6	7	8
RMSE	0.0176122	0.00368225	0.00148657	0.00146477	0.0015021	0.00158986	0.0015021	0.00153486

5. Discussion

Observing the results from the initial group of control experiments reveals that the predictive performance of the model varies significantly with the number of training epochs, peaking at 25 epochs. This phenomenon arises because insufficient training samples prevent model saturation, thereby limiting its effectiveness. Conversely, an excess of training samples may lead to overfitting, thereby diminishing performance. Thus, it is crucial during research to meticulously adjust both training samples and epochs to optimize model performance before proceeding with subsequent experiments.

Examining the second set of control experiments demonstrates that the model's predictive efficacy varies with the training-test ratio. Across ratios tested from 75:25 to 90:10, superior performance is consistently observed with higher training ratios, reaching optimal effectiveness at 90:10.

Analysis of the outcomes from the third set of control experiments highlights the impact of training data age on model performance. Older training data compromises sample reliability, thereby diminishing predictive accuracy. Optimal results are achieved when utilizing a training dataset spanning four years.

6. Conclusion

Through systematic adjustments of influential variables such as epochs, training-test ratios, and training dataset spans, this study employed comparative experimental methods to identify optimal configurations that maximize model predictive accuracy. Specifically, 25 epochs, a training ratio of 90:10 yielding a minimal Mean Squared Error (MSE) of 0.00175572, and a training dataset spanning four years resulting in the lowest MSE of 0.00146477 were determined as optimal parameters.

However, the experimental design of this study reveals limitations in the granularity of control group division, particularly highlighted in the third experiment where smaller training samples hindered short-term predictive capabilities. Addressing these limitations, future enhancements could include integrating dynamic adaptive training models and augmenting data using reinforcement learning techniques for improved short-term predictions. Additionally, broader investigations into market behavior dynamics could provide further insights into optimizing model performance across diverse scenarios.

References

[1] *Wang Jun, Zhang Peng, & Yuan Shuai. (2018). A Comparison of Seq2Seq RNN and LSTM Models for Stock Prediction. Times Finance (35), 3.*

[2] *Huang Yucheng, & Fang Weiwei. (2021). Research on Stock Price Prediction Based on LSTM Network. Modern Computers, 27(34), 6.*

[3] *Lin Xin, & Zhu Xiaodong. (2022). LSTM Stock Price Prediction Model Based on Attention Mechanism. Journal of Chongqing Technology and Business University (Natural Science Edition) (002), 039.*

[4] *Tao Yongkang, Zhang Guangqiang, & Li Peng. (2023). Research on Stock Prediction Based on LSTM Model with Attention Mechanism. Journal of Lanzhou University of Arts and Science (Natural Science Edition) 37(2), 49-54.*

[5] *Huang Chaobin, & Cheng Ximing. (2021). Research on Stock Price Prediction Based on LSTM Neural Network. Journal of Beijing Information Science and Technology University (Natural Science Edition), 036(001), 79-83.*

[6] *Li Liping, Zeng Lifang, Jiang Shaoping, & He Wenqian. (2023). Stock Price Prediction Based on LSTM Neural Network. Journal of Yunnan Minzu University: Natural Science Edition, 32(4), 528-532.*

[7] *Zhu Wenchao. (2023). Stock Price Prediction—Financial Time Series Modeling and Decision Making Based on LSTM. Modern Marketing: Lower (3), 39-41.*

[8] *Han Ying, Zhang Dong, Sun Kaiqiang, Tan Haoran, & Lu Chao. (2023). Research on a New Stock Prediction Model Combining Long Short-Term Memory Network and Width Learning. Operations Research and Management, 32(8), 187.*