Proceedings of Business and Economic Studies, 2024, Volume 7, Issue 5
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Deep Learning-Based Stock Price Prediction Using LSTM Model

Jiayi Mao[1], Zhiyong Wang[2]*

[1]Chengdu Shude High School, Chengdu 610400, China
[2]University of Electronic Science and Technology of China, Chengdu 610400, China

***Corresponding author:** Zhiyong Wang, 15618517686@163.com

Abstract: The stock market is a vital component of the broader financial system, with its dynamics closely linked to economic growth. The challenges associated with analyzing and forecasting stock prices have persisted since the inception of financial markets. By examining historical transaction data, latent opportunities for profit can be uncovered, providing valuable insights for both institutional and individual investors to make more informed decisions. This study focuses on analyzing historical transaction data from four banks to predict closing price trends. Various models, including decision trees, random forests, and Long Short-Term Memory (LSTM) networks, are employed to forecast stock price movements. Historical stock transaction data serves as the input for training these models, which are then used to predict upward or downward stock price trends. The study's empirical results indicate that these methods are effective to a degree in predicting stock price movements. The LSTM-based deep neural network model, in particular, demonstrates a commendable level of predictive accuracy. This conclusion is reached following a thorough evaluation of model performance, highlighting the potential of LSTM models in stock market forecasting. The findings offer significant implications for advancing financial forecasting approaches, thereby improving the decision-making capabilities of investors and financial institutions.

Keywords: Autoregressive integrated moving average (ARIMA) model; Long Short-Term Memory (LSTM) network; Forecasting; Stock market

Online publication: October 22, 2024

1. Introduction

Globally, there are 60 recognized stock exchanges, presenting financial experts with the considerable task of identifying investment opportunities within a vast array of financial instruments. Wealthier investors, in particular, seek advanced models for budgeting and stock price forecasting. To address this need, financial experts must have a profound understanding of market dynamics and the ability to make independent decisions. As economies evolve, the demand for accurate stock predictions grows, intensifying interest in the analysis of stock trends. However, the inherent volatility of stock prices poses a continuous challenge to reliable forecasting, leading to ongoing debates among experts.

The introduction of the Efficient Market Hypothesis in 1956 provided a theoretical framework for stock price prediction, proposing that stock prices reflect all available market information, including potential events and their associated probabilities. This concept has shaped subsequent efforts in stock price forecasting, acknowledging the complexity and probabilistic nature of market movements.

Moreover, stock price movements are inherently time series data. Time series forecasting is a prominent area of research, with applications ranging from stock price prediction to business planning, weather forecasting, and resource allocation. While forecasting can be viewed as a subset of supervised regression problems, time series data exhibit unique characteristics, such as strong auto-correlation. Common models for time series prediction include autoregressive (AR), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) models [1-3]. Stock market analysis is often discussed as a specific type of time series due to its unique attributes. Given the complexity of time series models, traditional linear regression approaches frequently fail to capture significant patterns, as many time series exhibit non-linearity. Without robust non-linear models, it is challenging to predict or estimate the intricate and dynamic shifts within the financial sector. The non-linear nature of financial time series data demands the development of advanced models that can better account for the complexity and variability present in financial markets, thus improving the accuracy of stock market forecasts [4,5].

Deep neural networks enable the prediction of future stock prices in the presence of anomalous non-linearity. These networks attempt to map and amplify the information required for modeling a function, leading to improved predictive outputs. Neural networks consist of neurons interconnected through weighted inputs, with activation functions introducing non-linearity into the network. This non-linear feature is then propagated across different neurons. Neural networks typically operate via backpropagation, utilizing gradient descent to minimize errors from the output layer to the input layer [6]. In various time series forecasting tasks, deep learning methods outperform other models in predicting non-linear data. This study focuses on the Long Short-Term Memory (LSTM) model, which is well-suited for time series prediction, and compares its performance with other widely used models.

2. Methodology

2.1. Common methods for current stock prediction

2.1.1. Linear regression

Linear regression assumes a linear relationship between variables, suggesting that future stock prices can be predicted through a linear combination of past stock prices. However, this linear assumption limits the model's predictive performance to a certain extent.

2.1.2. Support vector regression

Support vector regression (SVR) is a machine learning algorithm derived from the support vector machine (SVM) framework, extended for regression tasks. The core principle of SVR involves finding a hyperplane that best fits the data while maximizing the margin—the distance between the hyperplane and the nearest data points from both classes. These nearest points, known as support vectors, are crucial in determining the position and orientation of the hyperplane.

The algorithm aims to minimize prediction errors while balancing model complexity with the fit to training data. It achieves this by employing a loss function, commonly epsilon-insensitive loss, which permits some deviation in individual predictions without incurring significant penalties. This feature makes SVR particularly effective in handling non-linear and high-dimensional data, enabling the model to capture complex relationships while maintaining generalization to new data.

2.1.3. Decision tree regression

The decision tree regression algorithm operates by constructing a hierarchical model that recursively partitions the input space based on feature values, with the goal of minimizing the variance of the target variable at each step.

Starting from a root node, the data is split according to the feature that maximizes the reduction in variance, forming a tree structure where internal nodes represent feature-based decisions, and leaf nodes represent predicted values. The algorithm iteratively selects the best feature for splitting until a predefined stopping criterion is met, such as reaching a certain depth or a minimum sample size at a node. The final model consists of a tree where each path from root to leaf corresponds to a series of decisions based on input features, culminating in a prediction, typically the mean of the target variable's values within the leaf node's data subset. While this approach provides interpretability and flexibility, caution is required to avoid overfitting, often addressed through pruning.

2.1.4. Random forest regression

Random forest regression is an ensemble learning method that builds multiple decision trees during training and outputs the average prediction of these individual trees. The core principle involves creating a forest of decision trees, where each tree is trained on a different random subset of the data and features, introducing diversity to reduce overfitting and improve generalization. During prediction, each tree in the forest votes on the output, and the average of these votes is taken as the final prediction. This ensemble approach generally results in a more accurate and robust model compared to a single decision tree.

2.2. Long Short-Term Memory model

The LSTM model, initially proposed by Hochreiter and Schmidhuber in 1997, has gained widespread popularity for addressing time series prediction problems [7]. As an advanced form of Recurrent Neural Networks (RNNs), LSTM has demonstrated strong performance across various tasks and is now extensively used [8]. It resolves the challenge of retaining information over extended periods by integrating gated units and memory cells into its architecture. The memory cells store data from recent inputs, and their state is updated whenever new information arrives. When processing information, the cell state is refreshed in conjunction with new data.

LSTM's primary objective is to maintain long-term dependencies, effectively tackling the vanishing gradient problem commonly encountered in traditional RNNs. This is achieved through gating mechanisms that regulate the flow of information into, within, and out of the memory cells, allowing the network to learn and retain relevant information over long sequences.

In all recurrent neural networks, the modules of the neural system are arranged in a chain-like structure. While standard RNNs possess a basic modular structure, LSTM networks have a distinct architecture. In this design, gates control the flow of data into the cell state. A gate consists of a sigmoid function combined with element-wise multiplication. The sigmoid function produces values between 0 and 1, where 0 signifies “no data passage” and 1 indicates “passage of all data.” The controlled flow of information through these gates allows LSTM networks to process sequential data effectively. The essential components and functions of LSTM networks are as follows:

(1) Memory cells: The core units of LSTM that store information over time. Unlike traditional RNNs, LSTM’s memory cells can retain data for longer periods, crucial for recognizing temporal patterns.

(2) Input gate: This gate determines how much of the new input information should be stored in the memory cell. It uses a sigmoid function to select important values to be updated, allowing the network to selectively incorporate new information.

(3) Forget gate: The forget gate decides what information from the memory cell should be retained or discarded. It uses a sigmoid function to output a value between 0 (completely forget) and 1 (fully retain).

(4) Output gate: This gate controls when and what information from the memory cell is output, determining the appropriate representation for the next sequence step or final output.

(5) Cell state: The cell state serves as the core memory storage, updated by the input and forget gates, and influences the output gate’s decision-making process.

(6) Activation functions: LSTM networks typically use non-linear activation functions, such as tanh for cell state transformation and sigmoid for gate regulation, which introduce non-linearity crucial for learning complex patterns.

(7) Hidden state: The hidden state, derived from the cell state, is output at each time step, carrying learned information forward in the sequence for predictions or decisions.

3. Study setup and result analysis

3.1. Datasets

Four banking sector stocks from the Shanghai Stock Exchange were selected for this study, covering the period from March 1, 2019, to March 1, 2024. The dataset features include the opening price, highest price, lowest price, and closing price for each day. The primary focus of this study is on predicting the closing prices.

3.2. Descriptive statistical analysis

Initially, the data for the four selected stocks over the past five years were visualized, and their moving averages were plotted, as shown in **Figure 1**. Preliminary observation suggests that the stock price trends for Shanghai Pudong Development Bank, China Minsheng Banking Corp., and Huaxia Bank are largely consistent. In contrast, China Merchants Bank exhibits a significantly divergent trend. This observation implies that while companies within the same industry may share similarities, they also exhibit unique characteristics, highlighting the nuanced differences among institutions within the same sector.

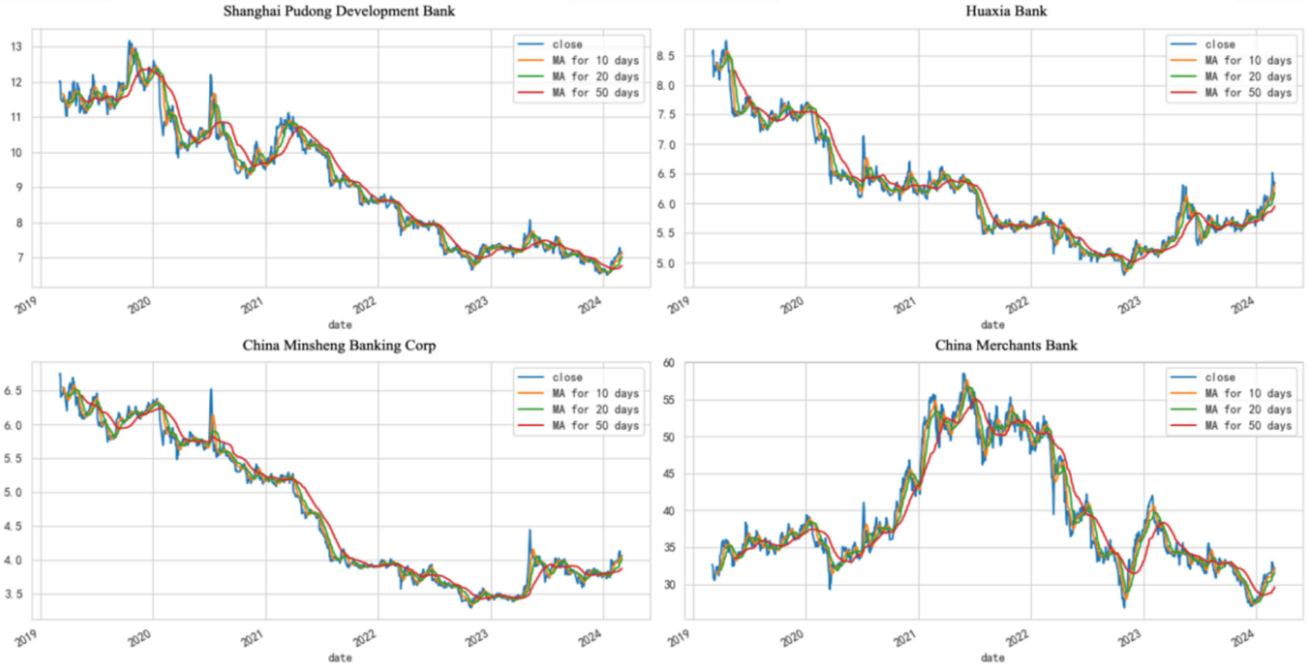

Figure 1. Visualization of the stocks and their corresponding moving averages

In stock analysis, the rate of stock return is a pivotal indicator for evaluating the performance of investments. It is often used to represent the fluctuation in stock prices over a defined period and the returns accrued by investors. Several formulas exist for calculating stock returns, including Simple Return, Logarithmic Return, Arithmetic Mean Return, and Geometric Mean Return. For this study, the Logarithmic Return, also known as Continuous Return, was chosen as it encapsulates the continuous compounding of stock price movements and possesses favorable mathematical properties. The calculation formula is as follows:

$$R = ln(\frac{P_{end}}{P_{begin}}) \tag{1}$$

Where P_{begin} represents the stock price at the beginning of the period, and P_{end} denotes the stock price at the end of the period.

A histogram of the returns was constructed, as depicted in **Figure 2**. It can be observed that the mean returns of the four banks are concentrated around zero, with China Merchants Bank exhibiting the most dispersed returns, indicating the highest variance.

The Mean-Variance Model, a foundational theoretical framework for evaluating and selecting portfolios, was proposed by Harry Markowitz in 1952 and underpins modern portfolio theory. This model posits that, for a given level of risk, investors will choose portfolios that maximize expected returns, or alternatively, select the portfolio with the least risk for a given expected return. A single stock can be considered a simple portfolio, where the mean represents the expected return and the variance indicates the risk. The computation and visualization of the mean and variance of the returns for the four stocks are shown in **Figure 3**.

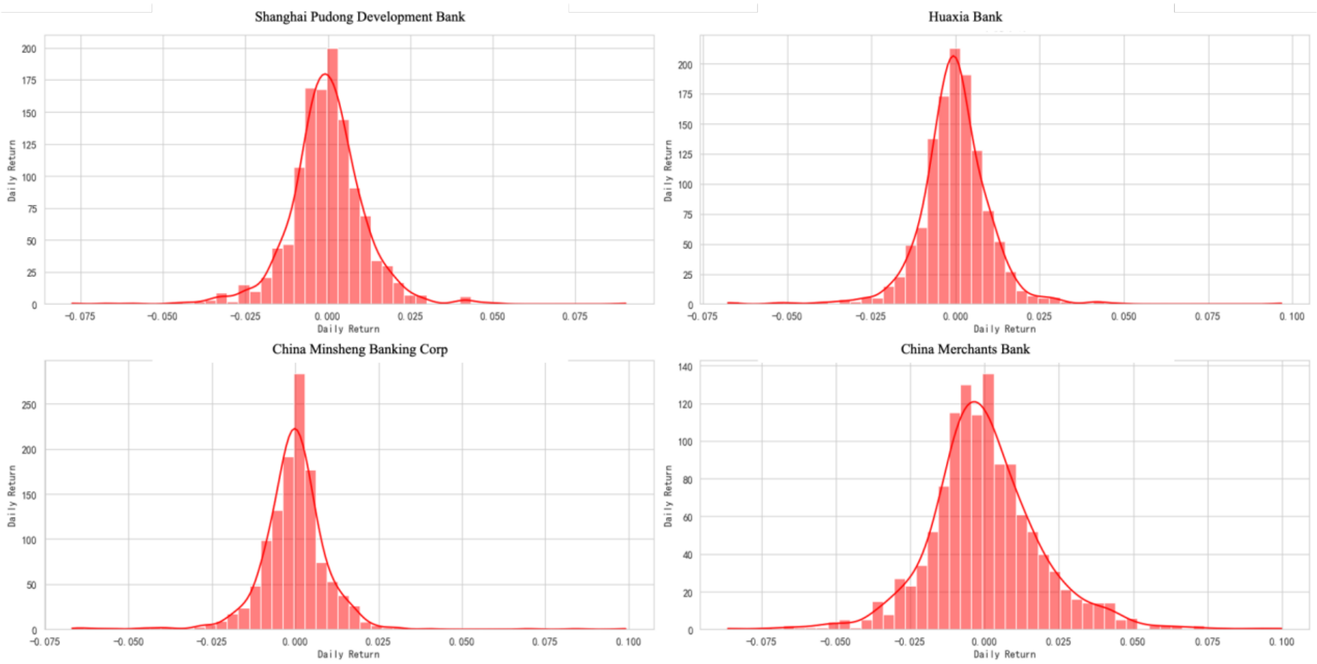

Figure 2. Histogram of stock returns

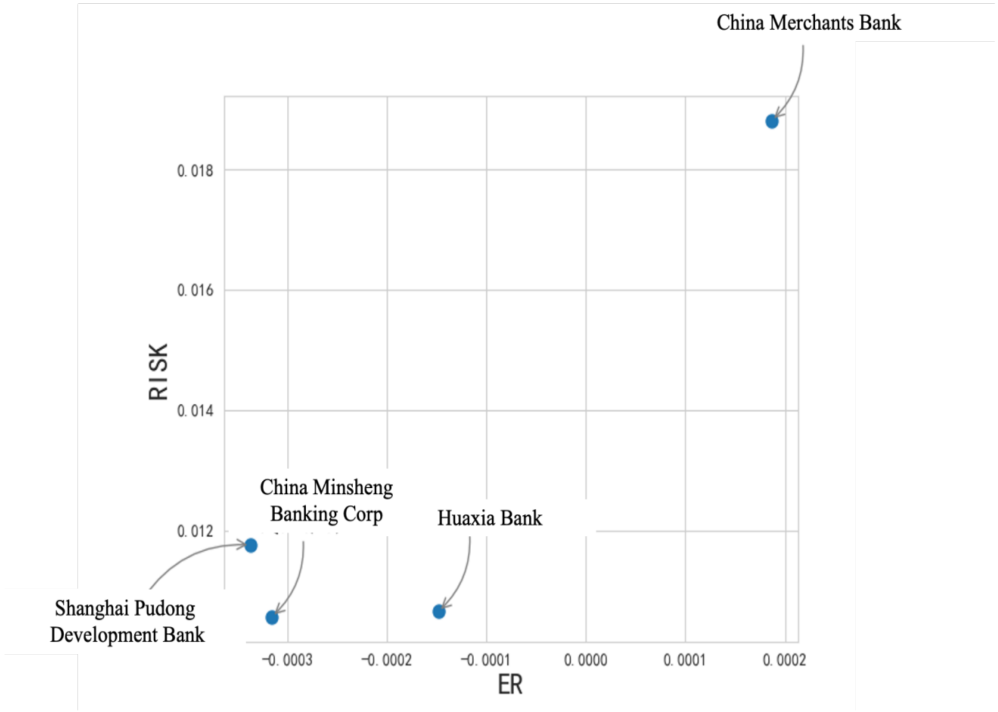

Figure 3. The mean and variance of the returns for the four stocks

From **Figure 3**, it is evident that China Merchants Bank exhibits the highest variance in returns, as well as the highest return rates. In contrast, Shanghai Pudong Development Bank shows lower returns and higher variance compared to Minsheng Bank. Rational investors would likely avoid investing in Shanghai Pudong Development Bank. Regarding Minsheng Bank, Huaxia Bank, and China Merchants Bank, investment decisions would depend on the trade-off between risk and return. Investors with a higher risk tolerance might opt for China Merchants Bank due to the potential for higher returns, while risk-averse investors may prefer Minsheng Bank for more stable returns with lower risk.

Since the four stocks belong to the same industry, the correlation coefficients of their stock prices and return rates were calculated, as shown in **Figure 4**. It is evident that Minsheng Bank, Shanghai Pudong Development Bank, and Huaxia Bank exhibit strong correlations in their stock price series, indicating similar price trends. In contrast, China Merchants Bank deviates significantly from the others. In terms of return rates, all four stocks show a considerable degree of similarity, with positive correlation coefficients suggesting that the return rate trends are relatively aligned.

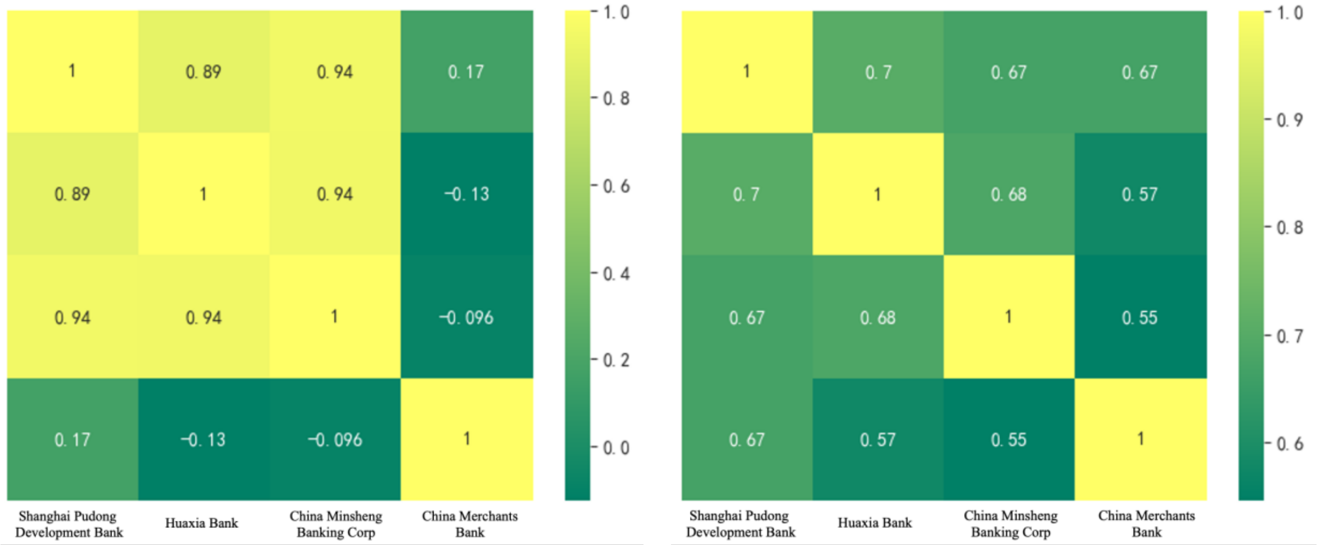

Figure 4. The correlation between the stock price series (left) and the return series (right) of the four stocks

3.3. Evaluation metrics

Several evaluation metrics are available to assess the accuracy of predictive models. The Root Mean Square Error (RMSE) is one such widely used metric. In this study, the original values and predicted values are represented by y_i and , respectively, with n denoting the total number of data points. This RMSE is calculated as follows:

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

3.4. Experimental procedure and results

Initially, the data were normalized using the Min-Max method, which performs a linear transformation to map the values to the interval [0,1]. The specific expression is as follows:

$$y_i = \frac{x_i - min\{x_j\}}{max\{x_j\} - min\{x_j\}} \tag{3}$$

Subsequently, training and testing samples were constructed. A training period of 60 days was selected, where the sample features consisted of stock prices from the preceding 60 days, and the sample label corresponded to the stock price on the 61st day. Thus, the 60-day stock prices were used to predict the price on the 61st day. The dataset was split into training and testing sets, with the most recent 10% of data designated as the testing set, and the previous 90% as the training set.

Table 1. LSTM experimental results

No. of Epochs	RMSE	Time(s)
1	0.1683	45
10	0.0679	255
100	0.0356	2499

The fitting results and actual outcomes are shown in **Figure 5**. The predicted values generated by LSTM closely align with the actual values, demonstrating high efficacy in forecasting stock price fluctuations.

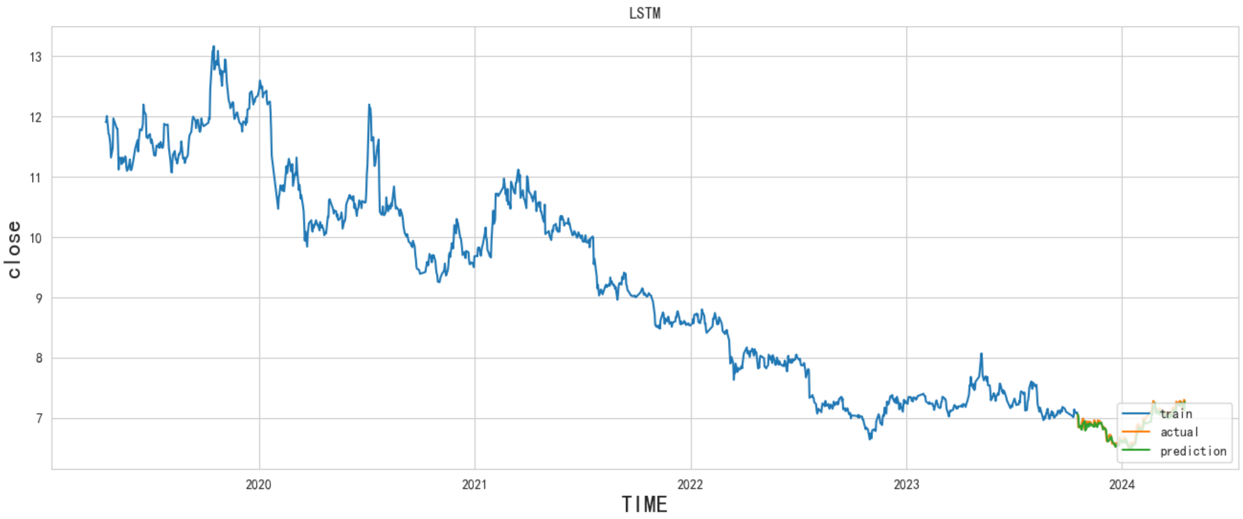

Figure 5. Comparison of LSTM predicted values and actual values

LSTM was then compared with other forecasting methods. Specifically, the predicted values from linear regression were plotted against the actual values, as shown in **Figure 6**. A significant discrepancy can be observed, with linear regression performing poorly in comparison to LSTM.

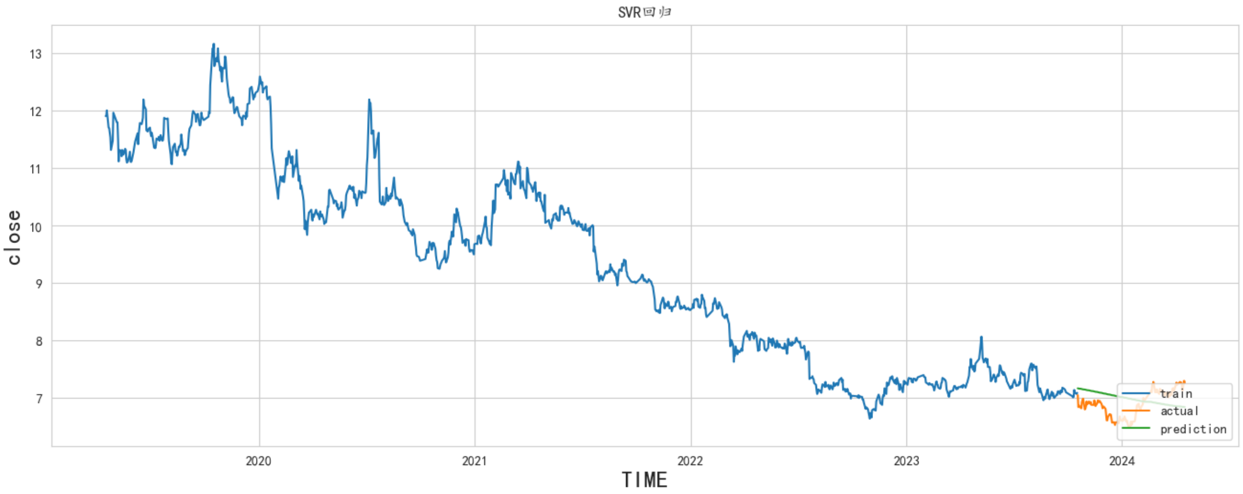

Figure 6. Comparison of linear regression predicted values and actual values

The specific model comparison results are presented in **Table 2**. It is clear from the table that traditional machine learning approaches, such as linear regression and support vector machines, do not perform as effectively as deep learning methods like LSTM.

Table 2. Model comparison

Model	RMSE
Linear regression	0.7951
SVR	0.2673
Decision tree	0.2997
Random forest	0.2056
LSTM (100 epochs)	0.0356

4. Conclusion

Through the aforementioned research, it is evident that deep learning algorithms demonstrate a significant performance improvement compared to conventional machine learning algorithms, exerting a profound impact on modern technology, particularly in the development of various time series-based predictive models. In the context of stock price prediction, they achieve the highest level of accuracy when compared to other regression models. LSTM-based forecasting models can be employed by individuals and enterprises for stock market prediction, thereby assisting investors and financial institutions in gaining greater economic benefits.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Kılıç DK, Uğur Ö, 2018, Multiresolution Analysis of S&P500 Time Series. Annals of Operations Research, 260(1): 197–216. https://doi.org/10.1007/s10479-016-2215-3

[2] Li P, Jing C, Liang T, et al., 2015, 2015 2nd International Conference on Information Technology, Computer, and Electrical Engineering, October 16–18, 2015: Autoregressive Moving Average Modeling in the Financial Sector. ICITACEE, Semarang, 68–71. https://doi.org/10.1109/ICITACEE.2015.7437772

[3] Zhang G, Zhang X, Feng H, 2016, Forecasting Financial Time Series Using A Methodology Based on Autoregressive Integrated Moving Average and Taylor Expansion. Expert Systems, 33(5): 501–506. https://doi.org/10.1111/exsy.12164

[4] Kaastra I, Boyd M, 1996, Designing A Neural Network for Forecasting Financial and Economic Time Series. Neurocomputing, 10(3): 215–236. https://doi.org/10.1016/0925-2312(95)00039-9

[5] Lendasse A, de Bodt E, Wertz V, et al., 2000, Non-Linear Financial Time Series Forecasting – Application to the Bel 20 Stock Market Index. European Journal of Economic and Social Systems, 14(1): 81–91. https://doi.org/10.1051/ejess:2000110

[6] Mandic DP, Chambers JA, 2001, Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and

Stability. John Wiley & Sons, Inc., Hoboken (NJ).

[7] Hochreiter S, Schmidhuber J, 1997, Long Short-Term Memory. Neural Computation, 9(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Chen K, Zhou Y, Dai F, 2015, 2015 3rd IEEE International Conference on Big Data, October 29 – November 1, 2015: A LSTM-Based Method for Stock Returns Prediction: A Case Study of China Stock Market. IEEE, Santa Clara, 2823–2824.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.