Research on Stock Price Prediction based on LSTM Modeling-A Stock Market as a Case Study

Yuxin Wang*

School of business, Chengdu University of Technology, Sichuan, China

*Corresponding author: wang.yuxin@student.zy.cdut.edu.cn

Abstract. This research intends to predict the closing values of the Shanghai Stock Exchange 50 Index using an enhanced Long Short-Term Memory (LSTM) neural network approach. The investigation leverages a comprehensive data set from August 2013 to August 2023, encompassing various market conditions. Alongside the primary LSTM analysis, the study introduces supplementary methodologies, specifically the Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF), to validate and enrich the time-series analysis. This study verified the non-stationary of the original data and processed them with method. 80%of the data was allocated for training, while the subsequent 20% was reserved for testing, the LSTM model, with its intricate architecture, exhibits superior predictive capabilities, particularly in stable market conditions. However, its performance wanes during volatile market periods, suggesting further refinements are necessary. This research contributes to the existing literature by applying machine learning methodologies to the turbulent and complex A-share market, offering invaluable insights for both academic researchers and market investors.

Keywords: LSTM; SSE 50 Index; stock price prediction; time series analysis.

1. Introduction

In recent years, due to the tense international situation, as well as the Federal Reserve continued to raise interest rates and global capital flows and other reasons, the A-share market has been more volatile, a number of brokerage firms believe that the current A-share market shows more will be the bottom of the characteristics, but with the policy, easing is expected to enhance the market is expected to recover. All of the volatility of the A-share market and the existing situation on the one hand, increase the possibility of investors investing in the later stage of the potential return, on the one hand, also increase the risk associated with investment is very important to help investors make reasonable investment decisions. On the one hand, also increases the risk associated with investment, so helping investors make reasonable investment decisions is very important.

At the same time, as artificial intelligence continues to evolve within the realm of big data, stock market forecasting has garnered significant interest from both the commercial sector and scholarly communities, given the heightened demand for enhanced services in the financial market [1]. Forecasting equity values stands as a paramount obstacle encountered by banking entities, corporations, and solo market participants [2]. Currently, the mainstream models for stock prediction are ARIMA, GRU, and LSTM [3]. Different prediction methods have different tendencies, advantages, and disadvantages, and from Chen's study, it becomes evident that when evaluating the efficacy of the LSTM approach vis-a-vis 18 traditional forecasting methods for stock volatility, LSTM emerges superior in its predictive accuracy. Moreover, as historical stock data accumulates, the predictive outcomes from the LSTM approach exhibit increasing stability. Based on this, this paper selects the LSTM model as the basis for analyzing and studying the China stock market and further proves the stock price prediction in practice.

Long Short-Term Memory (LSTM), is a novel cyclic network architecture combined with an appropriate gradient-based learning algorithm, inputs can be analyzed using time series [4]. Based on the time characteristics of stocks and the LSTM neural network algorithm, the corresponding stock trading prediction model is established [5].

LSTM has also spawned several innovative models, Naik et al. focused on examining stock return forecasts employing neural networks that have LSTM characteristics. They explored an RNN equipped with both long and proximate temporal memory capabilities to project forthcoming equity yields [6]. To enhance the precision of forecasting stock market indices, Wu et al. proposed a new framework structure that integrates CNN with LSTM, considering the unique attributes of financial temporal data for the objective of forecasting. This combination aims to refine the accuracy of stock price predictions [7]. Also, in the research of CNN and LSTM, researchers have proposed a set of deep learning regression models for stock price prediction based on CNN and LSTM. This framework amalgamates various statistical, machine learning, and deep neural approaches, yielding remarkable precision in forecasting equity values." [8]. Also, existing studies have compared the LSTM model with other models, evaluating advanced deep learning techniques, including LSTM, against conventional machine learning approaches like Random Forest and ARIMA, the aim was to determine the superior predictive method [9]. To achieve the aforementioned objective, the team extensively reviewed many scholarly articles centered on various computational learning methods with diverse input parameters. They observed that, while the regression's RMSE and the classification models' precision were indicators, the performance metrics exhibited significant variations, the LSTM approach outperformed other examined computational and profound learning techniques in terms of precision [10].

From the existing research, LSTM model-related topics have a lot of empirical research and analysis and have a certain superiority in stock price prediction models reflecting higher accuracy, currently in the volatility of the A-share market, there is a need to use LSTM-related models to carry out further research, analyze and predict the stock price, and further help investors analyze the market and make better investment decisions to get the best results. market and make better investment decisions to get higher returns.

2. Methods

2.1. Data Source and Description

SSE 50 Index, officially released on January 2, 2004, is a sample of 50 stocks selected as the most representative stocks in the Shanghai Stock Exchange with large market size and good liquidity. SSE 50 is one of the most representative blue-chip indices in the Shanghai market and is also the tracking underlying of the first domestically traded open-ended index fund (ETF). As of July 2023, the distribution of its top ten industries is: food and beverage, banking, non-banking financial, pharmaceutical and biological, power equipment, non-ferrous metals, utilities, construction and decoration, petroleum and petrochemical, electronics, with traditional industries such as finance and consumption as the main industries, and industry concentration is very high, with the weight of the financial sector close to 60%. A comprehensive response to the Shanghai Stock Exchange has the most market influence of many leading companies in the overall picture. The current total market capitalization of shares is more than 100 billion, responding to the Shanghai market for the best market influence of the leading companies in the overall situation, is one of the important indicators of the A-share market. The data used is the stock data of SSE 50 (000016) from August 1, 2013, to August 1, 2023, for the past ten years, the source of the data is Yahoo Finance.

2.2. Indicator Selection and Description

In most of the stock price prediction studies, basic trading indicators such as opening price, minimum price, maximum price, closing price, and volume are frequently used, and these five indicators are also used in this paper, without adding additional technical indicators, to predict the closing price of the SSE 50 index value on the next day.

2.3. Introduction of the Method

In this revised analysis, we employed a Long Short-Term Memory (LSTM) neural network to predict the closing prices of the SSE 50 Index, focusing on data from August 2013 to August 2023. A significant enhancement was made by applying first-order differencing to the data, following which a stationary test confirmed the transformed data to be stationary. The prepossessed data was then normalized using the MinMaxScaler to ensure better model performance. The data divided into a training set and a test set. The model consists of two LSTM layers, a fully connected layer and an output layer. And was trained using the Adam optimizer and the mean squared error loss function, and finally, the results were visualized to fit the stock price trend.

3. Results and Discussion

3.1. Descriptive Statistics

The data utilized in this analysis are sourced from Yahoo Finance. To forecast the closing price of the SSE 50 Index, a Long Short-Term Memory (LSTM) neural network is employed, a specialized variant of Recurrent Neural Networks (RNNs). LSTMs are noted for their effectiveness in capturing long-term dependencies in time-series data, making them particularly well-suited for tasks requiring a high degree of serial dependence, such as stock market price forecasting.

Initially, closing price information from the historical data of the Shanghai 50 Index, spanning from August 2013 to August 2023, is extracted. Recognizing the non-stationary inherent in the data, first-order differencing is applied to the closing prices, thereby achieving a stationary time series. Following this, MinMaxScaler is utilized to normalize the differenced data, converting all values to a scale between 0 and 1. This normalization step is instrumental in aiding model convergence and enhancing predictive accuracy.

A sliding window technique is employed for feature engineering, taking the closing prices from the previous 60 days to forecast the subsequent closing value. In terms of model architecture, the LSTM network consists of two LSTM layers, each with 50 units, an entirely interconnected layer, followed by a final output layer. The Relu activation function is employed within each LSTM layer. It is compiled using the Adam optimizer and trained using the mean square error (MSE) loss function, with batch sizes of 1 and 5 training cycles (epochs). Upon completion of the training phase, the model's performance is assessed using the test subset.

3.2. Analysis and Discussion of Results

3.2.1 Achieving stationary through first-order differencing

Therefore first-order differencing was performed on the data and after the differencing, the smoothing test was performed again, from the graph after the first order differencing we can see that there is no longer a clear trend or periodicity in the data (Fig. 1). The ADF test graph (Table 1) shows the p-values of the Augmented Dickey-Fuller (ADF) test before and after the first-order differencing was performed. From the depicted illustration, the p-value prior to differencing approaches 0.20., which means that the data is non-stationary.

The p-value after performing first-order differencing decreases significantly and is almost close to 0, which indicates that the data is smooth after differencing. It is proved that first-order differencing is an effective method to make non-smooth time series smooth and provides a basis for subsequent time series analysis and modeling work.

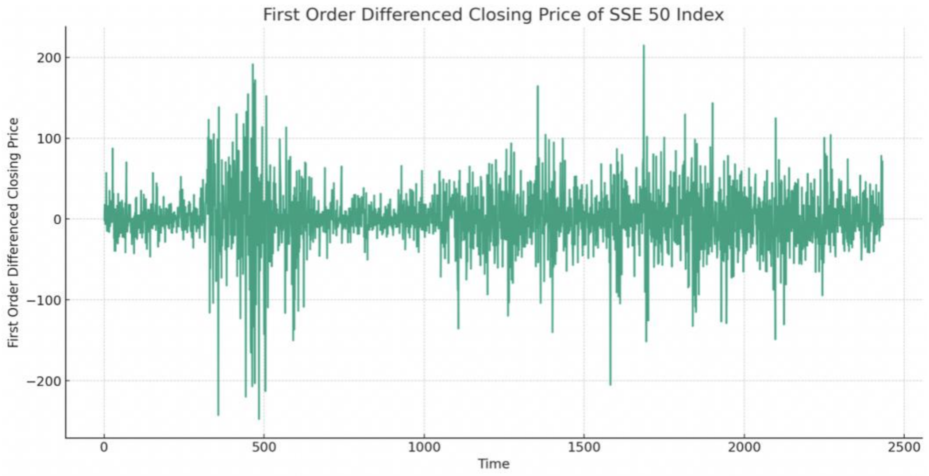

Fig. 1 Smoothness test results (After first-order differencing)

Table 1. ADF Inspection Form

Variables	Order of difference	t	p	AIC	Critical value		
					1%	5%	10%
Closing	0	-2.207	0.204	24278.597	-3.433	-2.863	-2.567
	1	-8.465	0.000***	24268.631	-3.433	-2.863	-2.567

3.2.2 Supplemental analysis with ACF and PACF to validate LSTM model

To complement our primary examination with LSTM neural networks, the study incorporated auto-correlation function (ACF) and partial auto-correlation function (PACF) analyses. These methods are commonly used in time series analysis to identify any patterns or cycles in the data that might not be immediately obvious. Figures 2 and 3 present the ACF and PACF plots, respectively.

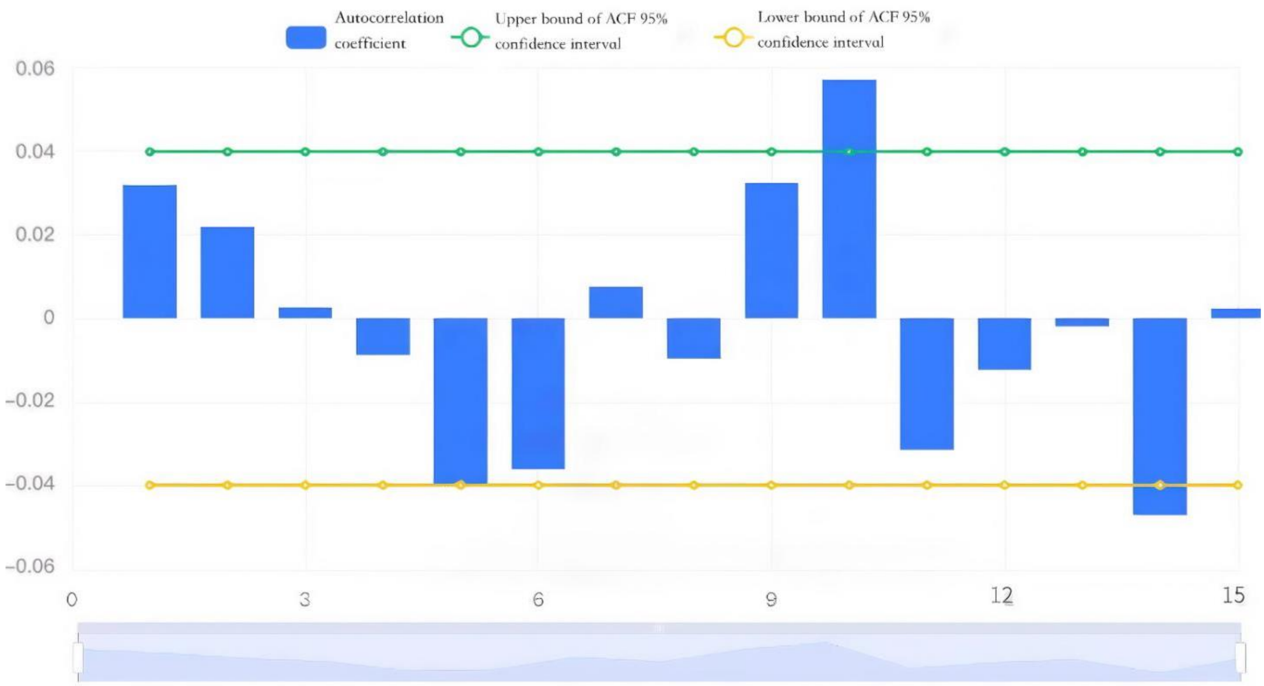

Fig. 2 ACF Chart

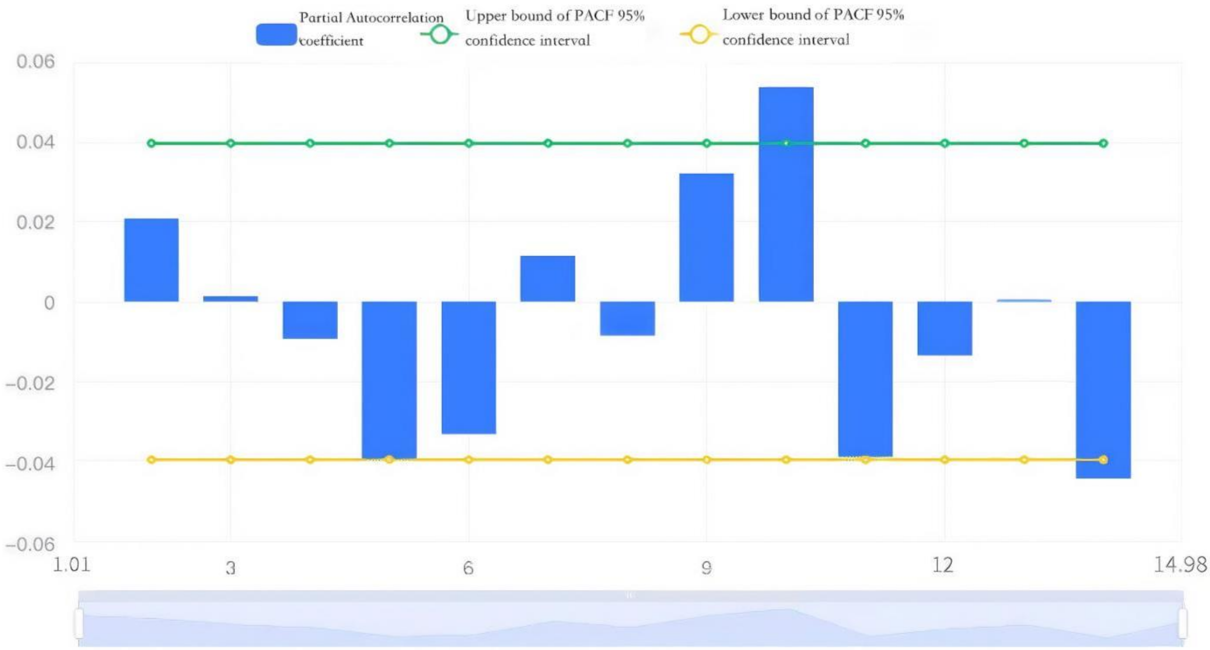

Fig. 3 PACF Chart

The ACF plot (Figure 2) reveals significant autocorrelations at various lags, suggesting that past values of the stock price have a substantial influence on its future values. This confirms the necessity of using models like LSTM that can capture long-term dependencies. The PACF plot (Figure 3), on the other hand, shows significant spikes at specific lags, indicating that these lags have a more direct influence on the stock price. These findings corroborate the choice of hyperparameters in our LSTM model, particularly the lagged values used as input features. The insights gained from the ACF and PACF analyses align well with the findings from our LSTM model. The autocorrelation patterns validate the model's ability to predict stock prices with reasonable accuracy, thereby confirming its robustness. The PACF results further justify our choice of model architecture and hyper-parameters.

3.2.3 LSTM forecast

After applying first-order differencing to the data, the predictions of the model are very close to the actual closing prices of the test set, because first-order differencing usually helps to reduce the effect of noise and outliers in the data, making the model more focused on the main dynamic characteristics of the data. It also further validates the accuracy and reliability of LSTM in stock price prediction. The resulting graphs show that the predicted and actual prices almost overlap, especially when the market trend is clear. However, we also notice that the prediction accuracy of the model decreases during periods of increased market volatility. This may be due to the fact that the LSTM model does not perform well when dealing with highly volatile data (Fig. 4).

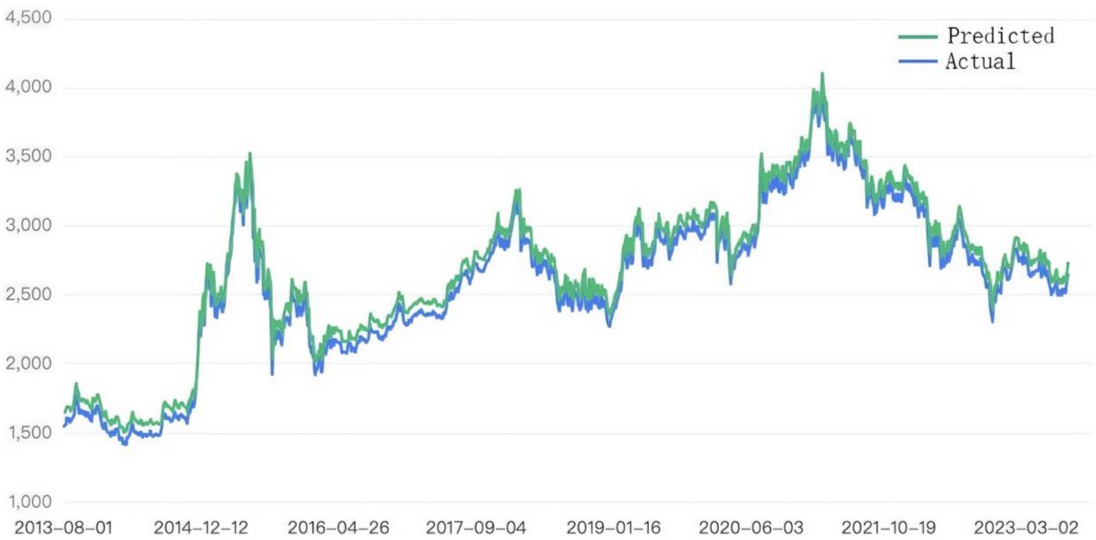

Fig. 4 LSTM forecast results

3.3. Suggestions and Deficiencies

Although the overall performance of the model is satisfactory, there is still room for improvement. Currently, the model only uses the closing price as an input feature adding more features such as trading volume, opening price, etc. may improve the predictive ability of the model. Model tuning can also be performed to adjust the hyperparameters of the model through methods such as grid search or random search, which may further improve the accuracy of the model. Moreover, more complex model architectures can be used to improve the accuracy of the model. Finally, it is also possible to improve prediction accuracy by combining multiple models using integrated learning techniques such as bagging or boosting. To further improve the prediction accuracy of the closing price of the SSE 50 Index by the above means.

4. Conclusion

This research marks a significant advancement in the realm of stock price prediction, specifically targeting the Shanghai Stock Exchange 50 Index through a LSTM neural network model. The incorporation of additional time-series analysis methods, such as the Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF), fortifies the robustness of the LSTM model. These supplementary analyses confirm the model's ability to capture the intricate, long-term dependencies inherent in stock prices. A noteworthy contribution of this study lies in its methodological rigor, employing first-order differencing to transform non-stationary data into a stationary format, thereby enhancing the reliability of the LSTM model.

Despite its successes, the model's efficiency diminishes in volatile market conditions, indicating scopes for improvements. Potential enhancements include feature expansion beyond closing prices, hyper-parameter tuning, and the utilization of ensemble learning techniques. By synergizing machine learning algorithms and financial forecasting, this study bridges a significant academic gap and equips investors with reliable, data-driven tools for market analysis. This body of work stands as a compelling testimony to the growing capabilities of machine learning algorithms, especially LSTM, in decoding the complexities of financial markets, setting the stage for future research and real-world applications.

References

[1] Agrawal J G, Chourasia D V S, Mittra D A K. State-of-threat in stock prediction techniques. International Journal of Advanced Research in Electrical Electronics & Instrumentation Engineering, 2013, 2(4).

[2] Kao L J, Chiu C C, Lu C J, Yang J L. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. Neurocomputing, 2013, 534-542.

[3] Alkhatib K. et al. A New Stock Price Forecasting Method Using Active Deep Learning Approach. Journal of Open Innovation: Technology, Market, and Complexity, 2022, 8(2): 2199-8531.

[4] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9(8).

[5] Liu S, Liao G and Ding Y. Stock transaction prediction modeling and analysis. 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018, 2787-2790.

[6] Naik N, Mohan B R. Study of Stock Return Predictions Using Recurrent Neural Networks with LSTM, International Conference on Engineering Applications of Neural Networks. 2019.

[7] Wu. et al. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. Multimedia Systems, 2021, 29(3).

[8] Mehtab S and Sen J. Stock Price Prediction Using CNN and LSTM-Based Deep Learning Models. Proceedings of the 4th IEEE International Conference on Electronics, Communication, and Aerospace Technology, 2020, 1481-1486.

[9] Millikan E, Subramanian Pand Joseph M H. Bitcoin Vision: Using Machine Learning and Data Mining to Predict the Short-Term and Long-Term Price of Bitcoin. Webology, 2021, 18: 751-760.

[10] Prasad A and Seetharaman A. Importance of Machine Learning in Making Investment Decision in Stock Market. Vikalpa, 2021, 46(4): 209–222.