Trend Prediction Analysis of Shanghai Composite Index Based on LSTM Neural Network

Rongze Gao

Department of Business, Hangzhou City University, Hangzhou, China

32111230@stu.hzcu.edu.cn

Abstract. The prediction of stock indices, particularly the Shanghai Composite Index (SCI), is crucial for understanding the health and direction of China's financial market in the post-pandemic era. Given the challenges of traditional forecasting methods due to the unpredictable nature of stock prices, this study explores the application of the Long Short-Term Memory (LSTM) neural network, complemented by Monte Carlo simulations and regularization with the daily trading data from September 21 2020, to August 13, 2023, to forecast the SCI's price trends for Q4 2023. The findings suggest a trajectory characterized by an initial decline, succeeded by a steady upward trend throughout the entirety of the quarter. Notably, the results have adeptly encapsulated both the lingering effects of the pandemic and the long-term rising trend of the Shanghai Composite Index (SCI). This research accentuates the potential of advanced neural network models in deciphering complex stock market behaviors, offering a groundbreaking perspective for market participants in navigating future investment decisions.

Keywords: LSTM Model, Shanghai Composite Index, Forecasts.

1. Introduction

The prediction of stock indices holds paramount significance in the realms of business and finance. Stock indices, representing a collection of individual stocks, encapsulate the overall health and trajectory of a market. However, the inherent nature of financial assets, characterized by their random walk behavior, makes the prediction of stock indices an exceptionally challenging endeavor. This unpredictability stems from the myriad of factors influencing stock prices, from global economic shifts to company-specific news, rendering traditional forecasting methods often inadequate.

The Shanghai Composite Index (SCI) stands as a barometer for China's financial markets, analogous to the role played by the S&P 500 in the United States. Comprising a blend of various stocks from different sectors, the SCI offers a comprehensive representation of the market dynamics and economic pulse of the nation. Predicting its trends has perennially been at the epicenter of concerns for market participants. A correct prognosis of the index's trajectory not only aids investors in making informed decisions but also fosters effective capital allocation and augments the efficiency of the market. In the aftermath of the pandemic, the Chinese stock market has witnessed a period of stagnation, making the accurate forecasting of the Shanghai Composite Index even more pivotal. Such predictions serve as a guiding light, reflecting the potential economic trajectory of China in these uncertain times.

This paper delves into the application of the Long Short-Term Memory (LSTM) model to predict the price trends of the Shanghai Composite Index for the fourth quarter of 2023. Leveraging the capabilities of LSTM, known for its proficiency in handling sequential data, we aim to capture the intricate patterns and dependencies in historical data. To enhance the robustness of our predictions, multiple Monte Carlo simulations were conducted, with the regression results serving as the foundation for the trend forecasts. Furthermore, a comprehensive analysis is provided, shedding light on the potential underlying reasons for the predicted trends, offering readers a holistic understanding of the anticipated market movements and their implications.

Through this research, hoping it could aspire to contribute to the ongoing discourse on financial forecasting, emphasizing the potential of advanced neural network models in navigating the complexities of stock market predictions.

2. Literature Review

Financial forecasting, especially in the stock market, has been extensively studied. These studies can be broadly categorized into three main approaches: traditional time series models, machine learning algorithms based on data mining techniques, and neural network models.

In the realm of traditional time series forecasting models, the ARMA model stands out as a prominent method. Li Min and colleagues effectively predicted the short-term future prices of the Shanghai Composite Index using the ARMA model as early as 2000 [1]. Furthermore, Feng Pan and Cao Xianbing validated the efficacy of the ARIMA model in forecasting the opening prices of China Merchants Bank (600036) stocks in 2011[2].

Neural network models have also gained traction in financial forecasting. In 1991, Matsuba introduced the application of neural networks for long-term stock price prediction and simulated stock price fluctuations using these models [3]. In 1990, Kimoto and Asakawa employed neural network methods to forecast the Japanese stock market, offering a novel solution to stock price prediction challenges [4]. Additionally, G.Peter Zhang, in 2004, conducted a comparative analysis between neural networks and the ARIMA model for time series forecasting, with empirical results indicating superior predictive capabilities of neural networks[5].

Recurrent Neural Networks (RNNs) inherently excel at handling sequential data due to their iterative nature. However, challenges such as the vanishing and exploding gradient problems, as highlighted by Bengio and others, have hindered the widespread application of standard RNNs [6]. To address these issues, Hochreiter and Schmidhuber introduced the Long Short-Term Memory (LSTM) model in 1997[7]. The LSTM, with its gating mechanisms, effectively tackles the long-term dependency issues faced by RNNs. Research by Li Jie and Lin Yongfeng in 2018 also emphasized that RNNs tend to forget previous state information over time, leading to the introduction of LSTMs [8]. Validating the efficacy of LSTMs in financial forecasting, Chen and colleagues modeled and predicted returns in the Chinese stock market using the LSTM model, demonstrating its effectiveness [9].

3. Dataset and Preprocessing

The data used comes from Yahoo Finance, representing the Shanghai Composite Index (000001.SS) daily trading data from January 1, 1992, to August 13, 2023. Fragments of this data are shown in Table 1.

Table 1. Shanghai Composite Index (000001.SS) daily trading data

Date	Open	High	Low	Close	Adj Close	Volume
2023/7/31	3287.209	3322.129	3281.013	3291.04	3291.04	459000
2023/8/1	3288.76	3305.34	3279.566	3290.949	3290.949	459900
2023/8/2	3281.858	3290.82	3252.304	3261.69	3261.69	405300
2023/8/3	3254.565	3280.864	3247.272	3280.463	3280.463	384700
2023/8/4	3296.094	3315.049	3281.734	3288.084	3288.084	470900
2023/8/7	3276.776	3276.776	3258.537	3268.831	3268.831	360300
2023/8/8	3260.72	3274.52	3246.67	3260.62	3260.62	316500
2023/8/9	3251.15	3257.08	3240.29	3244.49	3244.49	271200
2023/8/10	3242.21	3257.03	3234.12	3254.56	3254.56	287400
2023/8/11	3254.48	3254.48	3189.25	3189.25	3189.25	330100

Due to the potential large fluctuations in stock prices, using the raw data directly might make model training challenging. To reduce data volatility and make it more stable, we first applied a logarithmic transformation. Logarithmic transformations can help linearize exponential relationships and stabilize the variance in the data.

$$y_t = log(x_t) \tag{1}$$

Where y_t represents the log-transformed value, and x_t is the original data. To stabilize the time series data further and eliminate potential seasonality and trends, we took the first-order difference of the log-transformed data.

$$d_t = y_t - y_{t-1} \tag{2}$$

Where $d_t$is the result of the first-order difference, and y_t and y_{t-1} are the log-transformed data for the current and previous days, respectively.

To adapt the data for machine learning model inputs, we reshaped the data using the reshape function. Considering the impact of data magnitude on model training, we normalized the data using the MinMaxScaler function. This ensures that all data values lie between 0 and 1, helping to enhance the training effectiveness of the model.

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{3}$$

Where z represents the normalized data, and x is the differenced data.

4. LSTM Methodology and Implementation

4.1. LSTM Model Structure and Principles

Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) architecture designed to handle long-term dependencies. Traditional RNNs struggle with long-term dependencies due to a tendency to lose long-term information as iterations progress. LSTMs effectively counteract this through their unique gated structure.

LSTM is designed around the idea that the system should have the ability to selectively remember or forget information. To achieve this, LSTMs introduce the concept of "gates" that allow information to be selectively entered, retained, or exited. The principle of each control gate is as follows:

Forget Gate:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \tag{4}$$

Here, the forget gate's role is to determine which portions of the cell state are retained or dismissed. In this equation: f_t signifies the output of the forget gate at timestep t, which decides the fraction of the preceding cell state C_{t-1} that should be retained. W_f, h_{t-1}, x_t, and $b_f$are the weight matrix, the LSTM's prior timestep output, the current timestep's input vector, and the bias term, respectively. The Sigmoid activation function, represented by σ, constrains values between 0 and 1, denoting the retention fraction.

Input Gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$

The input gate is responsible for deciding updates to the cell values. Here: $i_t$determines the magnitude of new candidate values to be assimilated into the cell state. W_i, b_i, $\tilde{C_t}$, W_C, and b_C represent the weight matrix, bias for the input gate, new candidate values, weight matrix for these values, and their bias term, respectively. The hyperbolic tangent function (tanh) maps values within a -1 to 1 range.

Output Gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t \times \tanh(C_t) \tag{8}$$

The output gate's responsibility lies in determining which portions of the cell state are communicated outward. It operates by: Using a sigmoid layer to decide the cell state segments to relay. Applying the tanh function on the cell state and then multiplying the result with the sigmoid output. In this context: o_t determines the fraction of the cell state considered for the LSTM's output. W_o, b_o, h_t , and C_t denote the weight matrix, bias term, LSTM's output, and the current cell state, respectively.

Cell State Update:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{9}$$

Regarded as the "memory" segment of the LSTM, the cell state undergoes updates by combining the outcomes from the forget gate with fresh candidate values. In this equation: C_t is the current cell state. f_t, C_{t-1}, i_t, and $\tilde{C}_t$ are outputs from the forget and input gates, preceding cell state, and new candidate values, respectively.

4.2. LSTM Model Architecture and Parameter Selection

Utilizing the PyTorch framework, our LSTM Model is an advanced LSTM-based neural network tailored for predicting financial market dynamics. The model's architecture begins with an input layer designed to handle data from 700 historical trading days. This feeds into an intricate LSTM layer composed of eight sub-layers, each having 384 neurons, with the default tanh activation function used for gate operations and cell states. To combat overfitting, a 20% dropout rate is applied both within the LSTM layer and in a subsequent dedicated dropout layer. The model then transitions through two fully connected layers: the first transforming the 384-dimensional output from the LSTM to a 768-dimensional vector, and the second condensing it back to 384 dimensions. Ultimately, the output layer projects a 100-dimensional vector, designed to forecast values for 100 future trading intervals with a linear activation function. Using Adam optimizer with an optimal learning rate for training.

4.3. Loss function

In order to enhance the model's capability to capture the fluctuations inherent in the time series data and mitigate the phenomenon of overfitting, a regularization term was introduced to the loss function, rather than solely relying on the Mean Squared Error (MSE) as the objective criterion.

$$L = m - \lambda \times \sigma^2 \tag{10}$$

In the formula, m is the Mean square error (MSE), λ is the is a hyperparameter that represents the weight of the volatility penalty, σ^2 is the variance between model outputs. The model training was executed over 70 epochs.

5. Experimental and Analysis

The primary forecast results for the Shanghai Composite Index are illustrated in Figure 1. The x-axis represents the dates of trading days, while the y-axis denotes the Shanghai Composite Index's prices. The blue segment depicts data from 700 historical trading days utilized for prediction, and the red segment showcases the model's forecast for the upcoming 90 trading days' index trends.

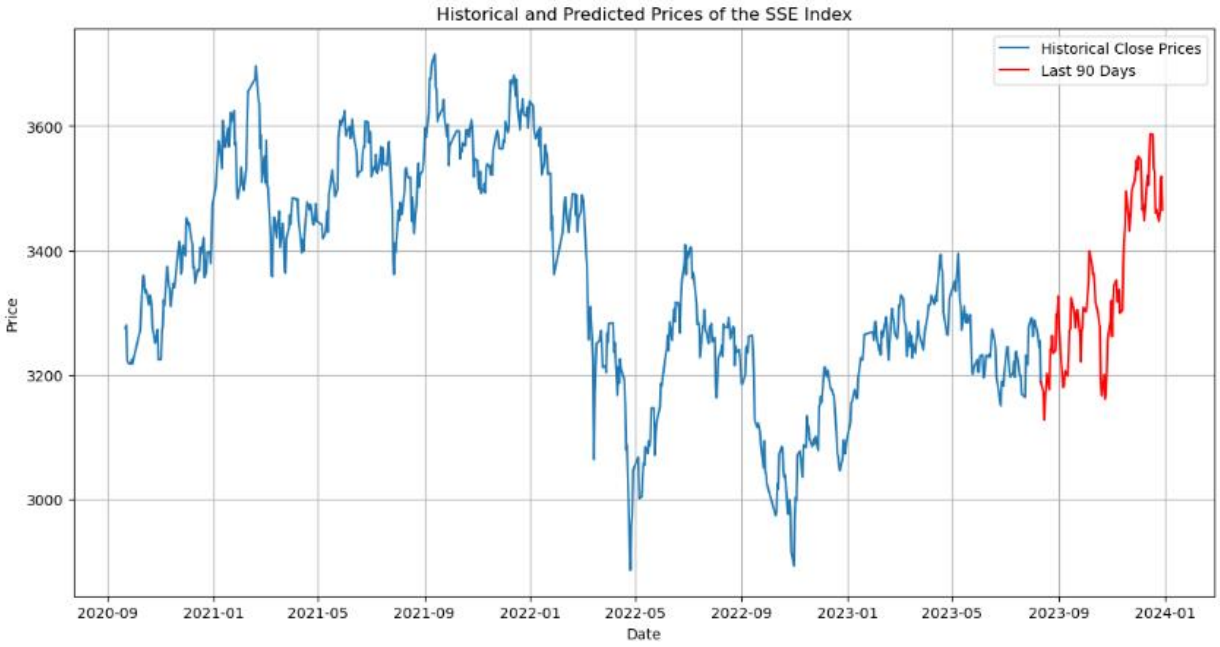


Fig. 1 The primary forecast results

In addition to the loss function with volatility incentive, the model captures the volatility and periodicity of historical data in figure 1. As the prediction results and data set placing on the one Q-Q plot(in Figure 2), the prediction results and data almost follow the same distribution.

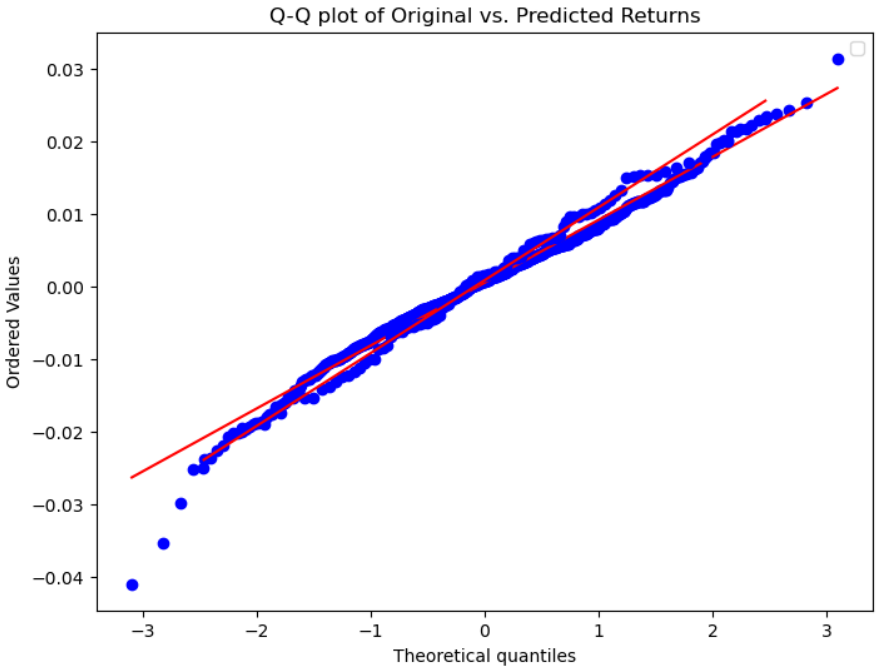


Fig. 2 Q-Q plot of Original and Predicted Returns

In the realm of financial markets, accurate prediction of specific asset prices often proves elusive due to the inherent complexities and volatile nature of such markets. However, a more macroscopic view, focusing on market trends rather than pinpoint asset prices, is frequently of greater practical importance to investors. This is primarily attributed to the prevalent understanding that asset price movements, over certain time horizons, typically exhibit stochastic behaviors that can be characterized as random walks following a log-normal distribution. In essence, future prices, while unpredictable, are conceived to be based on present-day information augmented by random perturbations. Such randomness in the context of log returns translates to them adhering to a normal distribution, leading to the perception of log prices as a form of random walk.

To account for the uncertainty inherent in model predictions within this environment, a Monte Carlo simulation was employed. This numerical technique, known for its strength in estimating a model's potential outputs by repeatedly sampling from probabilistic distributions, proved invaluable. In our context, by leveraging Monte Carlo simulations, not only was a singular predictive value obtained, but a confidence interval was established as well. This was achieved by generating 1,000 samples, thus providing deeper insights into the range of potential outcomes, affording decision-makers more information and, consequently, enhanced flexibility. Such an approach, considering the unpredictability and intricacy of financial markets, is especially salient.

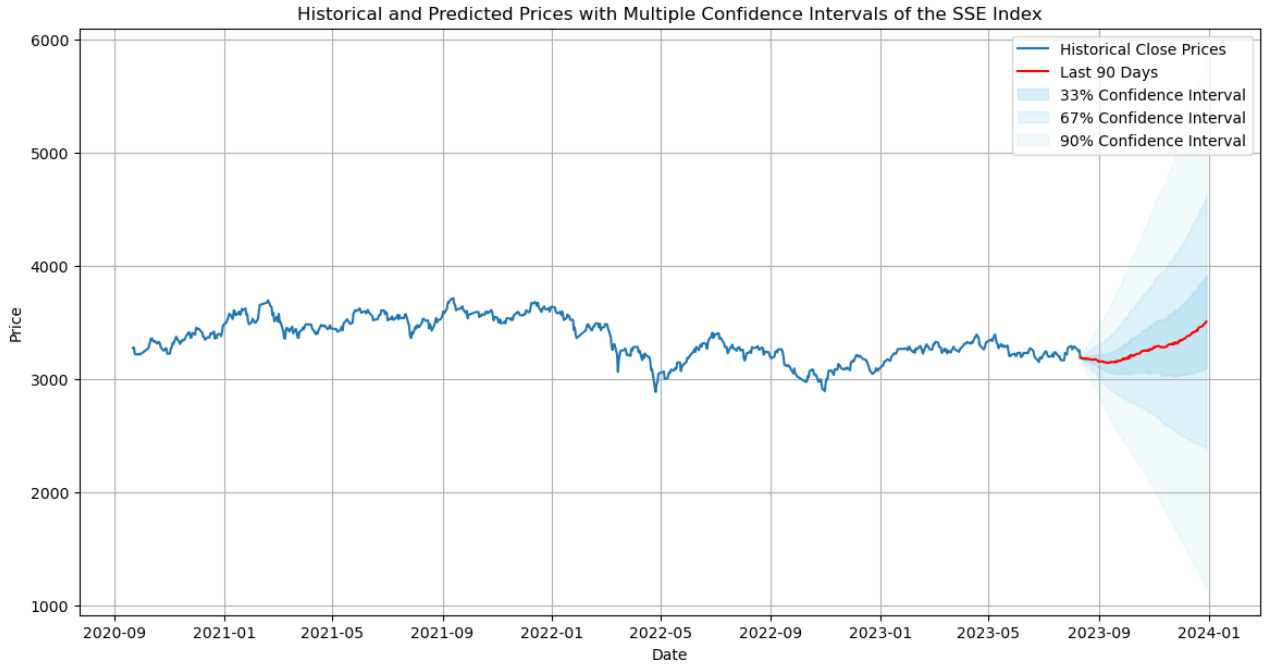


Fig. 3 Historical and Predicted Prices with Multiple Confidence Intervals of the SSE Index

Figure 3 shows the forecast data after 2000 Monte Carlo simulations, in which the prediction part in red is the mean reversion of 2000 predictions, and the sky blue part is the confidence interval of the prediction under Monte Carlo simulation. Different transparency represents different confidence intervals

the LSTM-based predictions presented a notably expansive confidence interval. The expansive nature of this interval arises primarily from a significant variance in the prediction outcomes. During the training phase, a deliberate emphasis was placed on capturing the historical volatility and cyclical tendencies inherent in the data. This was achieved by introducing a regularization term to the loss function to incentivize the volatility component. While this aided in closely mirroring the historical fluctuations, it simultaneously increased the variance in the predicted outcomes. Moreover, it's important to highlight that the return predictions didn't conform to a standard normal distribution. Instead, they exhibited characteristics of a distribution with "fat tails" or, more formally, leptokurtic tendencies. Such distributions indicate a higher probability of extreme values than what one might expect under a standard normal distribution. This not only points to more frequent and significant market anomalies but also further accentuates the uncertainty surrounding future predictions.

Within the Monte Carlo simulation framework, a mean reversion approach was adopted. By utilizing the expected value from the predictive simulations to represent the daily asset price, certain inherent complexities arose. Specifically, the volatility induced by the incentivization mechanism during training was notably attenuated. This dampening of volatility can be attributed to the Central Limit Theorem, which emerges naturally when taking the expected value across numerous simulated outcomes. Consequently, this mean-reverted result, with diminished volatility, proves particularly apt for trend analysis in financial contexts.

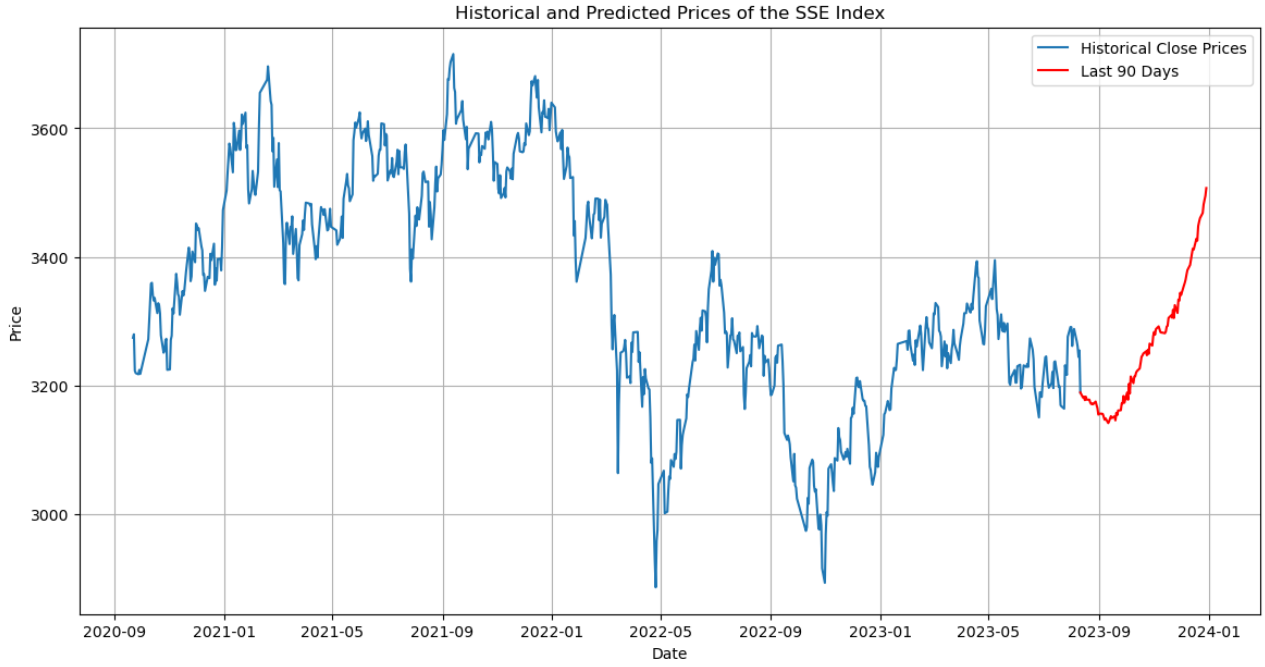


Fig. 4 Historical and Predicted Prices of the SSE Index

From Figure 4, it is evident that the model's forecast suggests a potential trajectory for the Shanghai Composite Index characterized by an initial decline, followed by an upward trend. This anticipated ascent is projected to span throughout the entirety of the Q4 quarter. Preliminary predictions anticipate a fluctuating downturn, settling within the 3,200 to 3,100 range prior to the onset of the Q4 period in 2023. Thereafter, a resurgence is expected, with the possibility of the index rebounding to a high of 3,500 points before the culmination of 2024.

On one hand, the model adeptly captures the lingering downtrend of the Shanghai Composite Index. It probably the residue from the economic volatilities induced by the COVID-19 pandemic. As early as 2020, in the early days of COVID-19 outbreak, Chen Lin and Qu Xiaohui had already discussed the impact of COVID-19 on China's stock market [10,11]. Now that the negative effects of COVID-19 are fading, it will take time for investors to regain confidence in the markets.

On one hand, the model has adeptly captured the long-term rising trend of the Shanghai Composite Index. This is evident in the Q4 predictions, where the index's trajectory first indicates a bottoming out followed by an upward movement. From a statistical perspective based on historical data, the long-term expected return of stock indices, due to the advancement of societal productive forces, tends to be slightly above zero. Since the cessation of the pandemic, the Shanghai Composite Index has consistently remained undervalued. As the repercussions of the pandemic wane, investor confidence is gradually restored and the economy steadily recovers. It is a logical progression for the Shanghai Composite Index to rebound. Furthermore, to stimulate the market economy and bolster market sentiment, the Chinese government has introduced pertinent policies. Notably, the People's Bank of China implemented a cut in the reserve requirement ratio (RRR) for financial institutions by 0.25 percentage points on March 27, 2023. This excludes financial institutions that already operate with a 5% deposit reserve requirement. Following this adjustment, the weighted average deposit reserve ratio for financial institutions is approximately 7.6% [12].

6. Conclusion

In an ever-evolving and intricate financial landscape, the accurate forecasting of stock indices is imperative, particularly as these indices encapsulate the holistic health and momentum of a nation's economic engine. This research was primarily instigated by the quest to adeptly predict the trend of the Shanghai Composite Index (SCI) for the fourth quarter of 2023, especially against the backdrop of economic perturbations spurred by the COVID-19 pandemic.

With a holistic approach, our study employed the Long Short-Term Memory (LSTM) model, renowned for its prowess in deciphering sequential data. Bolstering its robustness, multiple Monte Carlo simulations were undertaken to afford a spectrum of potential outcomes, thereby widening the horizon for market participants and decision-makers. The empirical findings suggested a probable dip in the SCI in the initial phase of Q4 2023, succeeded by a promising uptrend.

Significantly, this study serves as a bellwether for various stakeholders in the financial domain. Investors, policymakers, and market strategists alike can harness the insights from our predictions, establishing informed strategies for the imminent quarter. Furthermore, our research accentuates the unmatched capability of advanced neural network models like LSTM in navigating the intricacies of stock market forecasting.

However, like every academic endeavor, our study is not devoid of limitations. Firstly, the nature of our prediction, distinct from conventional test-set experiments, lacks the validation from real-time backtesting with contemporary data. While this current absence is a limitation, it paves the way for a subsequent validation in 2024, juxtaposing our predictions against the actual market performance. Secondly, the post-pandemic era, replete with the repercussions of COVID-19, remains sparsely researched. The scarcity of comprehensive discussions on the lingering economic implications of the pandemic aftermath hampers a holistic interpretation of our findings. Nevertheless, as more studies emerge illuminating the post-pandemic economic landscape, a re-evaluation in 2024 could potentially offer more nuanced insights.

In conclusion, this study signifies a crucial stride in financial forecasting, merging advanced neural networks with traditional financial paradigms. Despite its limitations, it offers a promising trajectory for the SCI, culminating 2023 on a hopeful note. Looking ahead, future research endeavors could delve deeper into post-pandemic economic dynamics, integrating them with advanced predictive models, offering an even more refined outlook for financial markets worldwide.

References

[1] Li Min, Zou Jiezhong, Li Junping, Liang Jianwu. Prediction of Shenzhen and Shanghai stock markets using ARMA model. Journal of Changsha Railway Institute, 2000(01): 78-84.

[2] Feng Pan, Cao Xianbing. Empirical research on stock price analysis and prediction based on ARMA model. Mathematics in Practice and Understanding, 2011, 41(22): 84-90.

[3] Matsuba, Ikuo. Application of neural sequential associator to long-term stock price prediction. IEEE International Joint Conference on Neural Networks. IEEE, 1991: 1476-1479.

[4] Kimoto T, Asakawa K, Yoda M, et al. Stock market prediction system with modular neural networks. Neural Networks, 1990., 1990 IJCNN International Joint Conference on. IEEE 1990: 1-6 vol.1.

[5] G. Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 2003(50): 159-175.

[6] Y. Bengio, P. Simard and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166, March 1994.

[7] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780.

[8] Li Jie, Lin Yongfeng. Time-series data prediction based on multi-time scale RNN. Computer Applications and Software, 2018, 35(7): 33-37.

[9] Chen K, Zhou Y, Dai F. A LSTM-based method for stock returns prediction: A case study of China stock market. IEEE International Conference on Big Data. IEEE, 2015: 2823-2824.

[10] Xu Hong, & Pu Hongxia. The impact of the COVID-19 pandemic on China's stock market: A study based on the event research method. Financial Forum, 2021, 26(7): 11.

[11] Chen Lin, & Qu Xiaohui. Market reaction to infectious public health events: A study on the impact of the COVID-19 epidemic on China's stock market. Financial Forum, 2020, 25(7): 10.

[12] The People's Bank of China. The People's Bank of China announces the decision to reduce the reserve requirement ratio for financial institutions on March 27, 2023. URL: http://www.pbc.gov.cn/goutongjiaoliu/113456/113469/4821841/index.html , last accessed 2023/8/27.