Research Article

Research on Shanghai Stock Exchange 50 Index Forecast Based on Deep Learning

Yiling Ding, Ning Sun, Jiahao Xu, Pengyan Li, Jiaxin Wu, and Sai Tang

1Harbin Normal University, Harbin, China
2Harbin Institute of Technology, Harbin, China

Correspondence should be addressed to Sai Tang; tangsai86@126.com

Received 26 January 2022; Revised 8 March 2022; Accepted 12 March 2022; Published 30 March 2022

Copyright © 2022 Yiling Ding et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

After decades of advance development, China’s stock market has gradually arisen into one of the world’s most important capital markets. The stock price index can well reflect the health status and macro change trend of a country’s economic development, which can be said to be a barometer of the country’s economic development. Studying the stock price index forecast is of great significance to the entire national economy and to each investor. Using 2 tools, Python and EViews8.0, and taking the Shanghai Stock Exchange 50 index as an example, the long short-term memory (LSTM) model in deep learning (DL) and the Autoregressive Integrated Moving Average (ARIMA) model are selected for fitting and prediction. The research results explain that the Root Mean Squared Error (RMSE) of LSTM model is lower, and the model based on DL method has stronger prediction ability on stock price index than traditional stock prediction model. This model is an effective stock prediction method.

1. Introduction

The trend of a country’s stock market can profoundly reflect the country’s economic development. The forecast of the stock market can increase the understanding of the development trend of the economic market, so that correct measures can be made on time for market changes. For financial market managers, the risks and abnormalities of stock market changes can be discovered in time through stock forecasting. When the stock forecast fluctuates abnormally, managers can pay attention to it in time and take targeted measures, which effectively reduce stock market losses and narrow the scope of risk. In March 2020, U.S. stocks fuse four times, causing huge losses to U.S. stocks. Good forecasting methods can reduce this risk to a certain extent.

The paper is organized as follows:

In Section 2, we provide a brief discussion of the current literature. Section 3 presents the DL Neural Network. Section 4 offers the proposed network architecture and the statistical model. Section 5 contains the results of the numerical experiments. And Section 6 concludes the paper.

2. Literature

In the early traditional financial data analysis, researchers used to apply statistical knowledge to it. For example, models that are widely used in time series analysis are as follows: differential autoregressive moving average model (ARIMA), autoregressive conditional heteroscedasticity model (ARCH), and exponential smoothing model (Exponential Smoothing). Farzaneh Nassir Zadeh used the ARMA model to predict the SP500 index and found that the prediction effect was average [1]. These statistical models all use historical time series data to fit the internal relations between the data as much as possible, so as to achieve the purpose of predicting financial data information. However, these models are not a good fit for the nonlinear and noisy stock market, and these models also have relatively high requirements for the input data, making predictions difficult and the prediction accuracy is low.

The rapid development of computer technology has made it possible for traditional machine learning models such as support vector machines to be widely used in the research of financial data, and good results have been
achieved. Can et al. used Support Vector Machine (SVM) to predict stock prices [2]. With the rapid development of the financial systems of various countries in the world, the complexity and diversity of the world's financial systems have greatly increased. Therefore, simple machine learning models have been difficult to accurately present effective prediction results.

Artificial intelligence (AI) technology has developed rapidly in recent years, and DL, that is, deep neural networks, has achieved great results in visual processing and voice applications. Stock market researchers also want to apply this novel technology to stock price forecasts to improve the accuracy of stock price forecasts. After combing through relevant foreign literature, we can find that, in the research of applying DL to stock price prediction, the more common ones are BP neural network, convolutional neural network, cyclic neural network, and long- and short-term memory network.

In 1988, White first applied the BP neural network to the analysis and prediction of stock market data. He took the stocks of IBM as the research object. However, due to the ubiquitous gradient disappearance problem of the BP neural model, the calculation effect was not good [3]. Hinton et al. proposed the Deep Belief Network (DBN) in 2006, which broke through the shackles of the development of neural networks. Since then, deep neural networks have developed rapidly [4]. Palangi pointed out that the cyclic neural network RNN can process time series information, but because the network has the problems of gradient disappearance and gradient explosion, it is difficult to analyze long-period time series data [5]. Compared with the previous recurrent neural network model, Persio et al. established the Wavelet-RNN model to predict stock prices. They took the SP500 index as the research object and obtained better prediction performance [6]. At the same time, in order to solve the problems of the above-mentioned neural network, the long- and short-term memory network came into being. It eliminated the problems of ordinary recurrent neural networks by introducing the concept of cells, thereby achieving more accurate time series and predictions in a larger range. Therefore, it is very suitable for the price prediction of financial products such as stocks and funds.

Zeng proposed an improved financial time series data modeling and analysis method based on the Deep Belief Network (DBN) decision algorithm, using the advantages of DL to process unstructured data to study the financial time series data of the Shanghai and Shenzhen stock markets. The accuracy rate can reach 90.5442% [7]. Lin et al. established a prediction model based on convolutional neural network and BP neural network to predict Shanghai zinc futures prices. Empirical evidence shows that this model has a relatively high accuracy rate for futures price expectations [8]. Liu et al. constructed an improved multifactor model based on recurrent neural network, through which the deep characteristics of several factors of stock price were extracted, and the stock price was predicted. The results of the prediction showed that this stock price prediction method is relatively accurate [9]. Han and others established a multilayer neural sensor neural network model to predict and analyze Apple's stocks, and the results show that the multilayer neural model is more accurate [10]. Chen et al. made predictions on the Shanghai and Shenzhen 300 Index based on the data of DL and the number of posts posted by stock bars. They compared the prediction results with 19 volatility prediction models and empirically proved that the prediction effect of DL is better than other general models [11]. Feng of Beijing Technology and Business University and others used the long- and short-term memory network model for the trend study of stock indexes, compared it with models such as SVR, and found that the DL model has better predictive effects [12]. Chen used the BP neural model to predict Baidu's stock and then used the ARIMA model to predict the stock price of Ali. Practice has proved that the short-term forecast using the BP neural network has good accuracy [13].

In summary, the research done by predecessors mostly used traditional investment analysis methods and modern statistical methods and could not cope with the non-stationarity and loud noise of the stock market. The research on AI in the field of stock price prediction was not perfect. Moreover, when using AI methods for analysis, simpler machine learning methods and single-layer neural systems are often used. There are problems such as rapid reduction and sharp increase of neural network gradients, and the prediction accuracy when predicting long-term time series data is low. Based on previous research, this paper uses the LSTM long-term memory network in DL, selects the Shanghai 50 index as a sample, overcomes various types of gradient problems of the recurrent neural network, and then predicts the stock price and compares it with the ARIMA model. The comparison can well supplement the previous research deficiencies and has a relatively large research significance.

3. DL Neural Network

It is an artificial neural network that has many layers between input and output layers. Each layer consists of neurons, synapses, biases, and functions. This section also defines this network completely. Firstly, define DL, which defines as computers learn things similar to human minds. Secondly, it describes long short-term memory (LSTM) neural network thoroughly.

3.1. DL. Humans have relatively strong learning capabilities. For example, we can teach children which are cars and which are bicycles. After a period of time, they can naturally classify new samples. It can be seen that humans have the ability to generalize from a certain scale of data. Machine learning usually first makes a hypothesis about a problem and then uses a computer to train the data. After training and learning the parameters of the model, the original data is finally predicted and researched.

DL network is a key side branch of the machine learning discipline. It is a more complex network system on top of machine learning. The characteristic of DL is that it is generally a multilayer neural network structure with
multiple hidden layers. Its effect is that it can analyze the underlying features to form more conceptual features, which can then be used for classification and other processing.

The network structure of the DL network is actually multilayered, and this neural network is actually inspired by the human brain neural network. As early as 1904, biologists have understood the structure of neurons. As can be seen from Figure 1, a human brain neuron generally has multiple dendrites, and dendrites are mainly used to receive input information. Then, there is an axon. There are many terminal synapses at the end of the axon, and these synapses are connected to the dendrites of other neurons. This mechanism of message transmission and processing inspired the construction of artificial neural networks. As shown in Figure 2, it is a multilayer neural network structure.

3.2. LSTM Neural Networks. Recurrent Neural Networks (RNN) are often used in the modeling and prediction of sequence data, such as the trend of stock prices over time and the sequence of words that constitute sentences in natural language. Why do recurrent neural networks have this ability? This is because it has a special structure. When observing the general neural network model, it can be found that, from the data entry layer to the hidden layer, and then from the hidden layer to the data output layer, there are connections between data layers of different natures, but the nodes of the data layers of different natures are connected. There is no connection between them, as shown in Figure 2. The difference of the cyclic neural network is that the nodes of its hidden layer are connected with each other. So, it is a neural network that is better at classifying and predicting data. As shown in Figure 3, due to this special structure, in the cyclic neural network, the input to the latent layer has both the output of the entry layer and the output of the latent layer. This is also the origin of the name of the recurrent neural network.

Long short-term memory network is a relatively common extended model of cyclic neural network. The standard RNN can handle relatively long intervals of related information, but when faced with sequence data with a relatively long-time span, it will cause problems; that is, the current situation is easily affected by the previous situation. LSTM can use its distinctive structure to solve this problem. In Figure 4, you can see the overall framework of the long- and short-term memory model. The calculation process of LSTM can be divided into three steps in general. The first step is to throw away certain information from the long-term state. The detailed task process is forgetting. The input of the gate layer ft is ht-1 and xt, and each element in the output matrix is the value of (0,1) and is calculated with each corresponding position element in the matrix Ct-1. The second step of LSTM is to store new information in a long-term state. The detailed task process consists of three parts: a tanh layer creates a new vector, a sigmoid layer controls the update of the elements of the candidate vector, and finally input the new data into the state Go in. The last step of LSTM is how to get the output information ht. The detailed task process is to use the input gate layer to select the output element and get the ht to be output.

4. Empirical Research

4.1. Empirical Analysis of LSTM Model. The raw data uses the daily data of the Shanghai Stock Exchange 50 Index from January 6, 2020, to January 6, 2021, with a total of 244 rows...
of data. The basic characteristic value involved in these data is the closing price variable. The source of these data is the tushare database. The normal data standardization method is used in the process of standardizing the original data set. The standardization process of this method is based on the average and standard deviation of the sample data. The sequence data processed by this method will conform to the standard normal distribution, which will facilitate our subsequent data processing and prediction.

After data standardization, 75% of the data set is used as the data set for training to generate model parameters for training, and 25% is used as the data set for testing, which is selected as the result of prediction. At the same time, it can be referred to as training set and test set for short. Using 244 rows of data for training and testing and predicting the feature of the closing price on the second day based on the information from the previous day, the closing price is the key to whether each investor makes a profit. It is an important data for studying the rise and fall of the stock market [14], so the closing price at \( T \) is used to predict the closing price at \( T+1 \).

The training environment of the model is as follows: the hardware equipment is Intel(R) Core(TM)i5-6200U CPU@ 2.30 GHz, 2.40 GHz, 64-bit operating system, 8 GB memory; the software environment is Windows 10 system, python 3.8, Anaconda 3-4.2.0-Windows-x86_64. In the LSTM model, in order to have more accurate prediction results, two hidden layers are selected. Batch_size, that is, the number of samples in each batch of training, is set to 60. Hidden layer units, that is, the number of hidden layer units, are set to 10. Time_step, that is, the size of the time step, is set to 1. Both the input layer dimension and the output layer dimension are set to 2 layers. Using Adam dynamic learning rate, as the error decreases, the learning step difference is dynamically adjusted to calculate the optimal value at a faster speed.

After defining the neural network variables, set the weights of the input layer and output layer, and then bias. When defining the LSTM neural network, the tensor is converted into two dimensions for calculation, and the calculated data enters the hidden layer. Convert the tensor to 3 dimensions, as the input of the LSTM cell, and the final result as the output of the output layer.

After defining the training model, train the model. As the number of iterations is larger, the prediction effect will be better, but it will take longer. After considering the hardware and software environment, the number of iterations is 1000. After the model is trained, the model is predicted, and finally a line graph is used to indicate the predicted result.

4.2. Empirical Analysis of ARIMA Model. The data of the ARIMA model also selects the daily closing price of the Shanghai 50 Index from January 6, 2020, to January 6, 2021, with a total of 244 samples, which basically meet the modeling requirements of the time series.

According to the above data, the changes in the Shanghai 50 Index have a certain time trend, so we can preliminarily judge that this series of data is a nonstationary time series.

We still need to be more rigorous in judging the stationarity of the original time series, so the unit root test is indispensable. The results of the test are shown in Table 1. Observing from Table 1, it can be found that the value of the t-Statistic is greater than 0.8, which exceeds the three negative values of the test level, so the null hypothesis is rejected, which means that the series we test is not stable.

Taking into account the instability of this time series, but also in order to reduce the sequence data error, we need to perform a first-order difference transformation on this time series data. After the transformation is completed, we need to do an ADF test on this sequence. The effect of the transformation is shown in Table 2. The observation results show that, after the first-order difference transformation, this sequence, the trend of the data is almost nonexistent, and it has changed from unstable data to stationary series data, so we can be sure to establish a moving average model with \( d = 1 \).

The most powerful and common method to identify ARIMA models is autocorrelation and partial autocorrelation functions. In EViews8.0, the autocorrelation and partial autocorrelation analysis graphs of samples are usually used for model identification and order determination.

After several experiments and adjustments, it can be concluded that when \( p = 1, d = 1, \) and \( q = 0 \), the model is relatively significant, and the fitting error is relatively small. So, finally, choose to use ARIMA (1, 1, 0) as the model frame.

Table 3 is some test results of the established model. Observing Table 3, we can find that the R2 value of the model is close to 1, the F-statistic value is also in line with expectations, and the AIC value is relatively small. This series of data can reflect that the model we built is remarkable and usable.

It is not enough that the model is significant as a whole. We also need to determine whether the residual value of the model estimation result is random, that is, whether it is white noise. We can find that the autocorrelation coefficients of the data are in the 0.95 confidence region, and the P value is greater than the test level of 0.05. Therefore, the residuals of the model we built are random. Use this model to conduct empirical research on the Shanghai 50 Index. The forecast is reasonable and feasible.

5. Forecast Results and Analysis

5.1. LSTM Model Prediction Results. Figure 5 shows the prediction effect of using the long and short-term memory network model to predict the Shanghai 50 Index. From an overall point of view, it can be found that the two curves of the true value of the index and the predicted value mostly overlap, especially the prediction effect between the 100th to the 125th data is relatively good, and the values are basically consistent. The trend from the first data to the 50th data is basically the same, but there is still a large deviation in the value. After the 200th data, the predicted trend is basically consistent, but the numerical deviation is relatively large, and there is a tendency for the deviation to increase. At the same time, it can be found that the accuracy of the constructed model in predicting the downward trend is greater.
Table 1: Unit root inspection.

Null hypothesis: CLOSING_PRICE has a unit root

Exogenous: None
Lag length: 0 (automatic - based on SIC, maxlag = 14)

<table>
<thead>
<tr>
<th>t-Statistic</th>
<th>Prob.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.854117</td>
<td>0.9031</td>
</tr>
</tbody>
</table>

Test critical values:
- 1% level: -2.835152
- 5% level: -1.891315
- 10% level: -1.597296

Table 2: Unit root test of the first-order difference sequence.

Null hypothesis: D(CLOSING_PRICE) has a unit root

Exogenous: None
Lag length: 0 (automatic - based on SIC, maxlag = 14)

Augmented dickey-fuller test statistic: -13.78892 (Prob.* = 0.0000)

Test critical values:
- 1% level: -2.612921
- 5% level: -1.982723
- 10% level: -1.590394

Table 3: ARIMA model estimation results.

Method: Least squares
Included observations: 242 after adjustments
Convergence achieved after 4 iterations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2989.475</td>
<td>183.0214</td>
<td>14.21287</td>
<td>0.0000</td>
</tr>
<tr>
<td>AR (1)</td>
<td>0.9920331</td>
<td>0.012132</td>
<td>79.38719</td>
<td>0.0000</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.967580</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.967432</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>38.29695</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>322664.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-1123.272</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-statistic</td>
<td>6565.879</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prob. (F-statistic)</td>
<td>0.000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Forecast results.
than the accuracy of predicting the upward trend. The reason for this phenomenon is that most of the training data selected when training the LSTM model is based on a continuous decline and a slow upward trend. Mainly, there is less learning about data with a large upward trend. Finally, from a long-term perspective, the stock price prediction index of the constructed LSTM neural network model is overall higher than the real situation.

5.2. ARIMA Model Prediction Results. The static prediction will substitute the true value into the model for each prediction and then continue to predict the future. The model that we have built is used to predict the index, and the prediction effect obtained is shown in the discounted chart 6. The blue solid line in the line chart shows the predicted value of the closing price, and the red dashed line above and below the solid line represents the meaning of 2 times the standard deviation. It can be seen that the predicted value basically conforms to the trend of actual value. According to the calculation, it can be concluded that the RMSE is 9.83991, indicating that the root mean square is relatively small; the inequality coefficient is 0.008613, which indicates that the error of the established model is relatively small.

At the same time, in order to further reflect the static forecasting ability of the ARIMA model, the static forecast predicted value and the actual observation value are placed in the same graph for comparison and observation. The comparison result is shown in Figure 6. It can be clearly seen from the figure that, overall, the predicted value of the model and the actual closing price are roughly the same, and the prediction effect is average. However, observing the specific prediction values can be found to be inaccurate, and the prediction values are relatively rough; especially when there are large fluctuations, the prediction effect will be greatly reduced.

Then, use the dynamic prediction method to make predictions, and the prediction results are shown in Figure 7. The meaning of the solid line and the dashed line is the same as the above static prediction chart. It can be seen that the fluctuation of the predicted value is relatively gentle. According to the calculation, the RMSE is 65.23719, indicating that the root mean square is relatively large; at the same time, the inequality coefficient is equal to 0.029731, which explains the general prediction effect of the dynamic prediction.

At the same time, in order to further observe the dynamic prediction ability of the ARIMA model, the predicted value of the dynamic prediction and the actual observation value are placed in the same graph for comparison and observation. The comparison result is shown in Figure 8. It can be clearly seen from the figure that the predicted value of the dynamic forecast can only roughly reflect the trend of the closing price of the stock and cannot be a good forecast. The dynamic prediction method has large errors, and the prediction effect is not good.

According to the prediction results of the four Figures 6–9, the static prediction effect of the SSE 50 Index is better than the dynamic prediction. In terms of the overall situation of the forecast, the short-term forecast effect is relatively ideal. The static prediction uses the sample data as an information set, and the prediction of the future situation is only based on the sample information set, while the dynamic prediction is based on the sample information set. Therefore, for time series data such as the closing price of stocks that are affected by various factors, the difficulty of dynamic forecasting is relatively large, and the accuracy of forecasting is relatively low compared to static forecasts that fully use existing information to predict. Of.

5.3. Comparative Analysis of Forecast Results. Prediction is an estimate of the future situation, so it is inevitable that there is a certain deviation between the predicted value and the true value. The quality of a model is often calculated by the prediction deviation. The size of the prediction deviation determines the size of the prediction accuracy. Generally speaking, the larger the error value, the lower the prediction accuracy of the model. Commonly used error evaluation indicators are RMSE, R2, etc. Here, considering the simplicity of the calculation and the clarity of the indicators, I chose to use the indicator RMSE to compare and analyze the prediction results.

Formula (1) shows the details of the RMSE formula. The square error can reflect the sum of the squares of the difference between each predicted value and the predicted value. After averaging them, the root sign can make the unit and the estimated value fall in the same order of magnitude for better estimation. The error is described, and the smaller its value, the higher the accuracy.

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_{true,i} - Y_{predict,i})^2}. \tag{1}
\]

Use the long- and short-term memory network model to obtain the prediction result, calculate its RMSE result, and compare it with the RMSE value of the ARIMA model prediction result. The comparison results obtained are shown in Table 4. It can be seen from the figure that the RMSE of the LSTM model is relatively small, which shows that the reference value of this model in the stock index prediction is relatively large. The RMSE of the two forecasts of the ARIMA model is relatively large, which reflects that the effect of this model on the stock index forecast is not very good. At the same time, the RMSE of the prediction result of the long- and short-term memory network model is significantly lower than the error value of the ARIMA model, indicating that the prediction result of the LSTM model is closer to the true value, the prediction effect is more accurate, and the information of the feature value is better captured. It can provide forecasters with more meaningful information.

Statistical models such as the ARIMA model all use historical time series data to fit the internal relationship between the data as much as possible, so as to achieve the
purpose of predicting financial data information. However, these models are not a good fit for the nonlinear and noisy stock market, and these models also have relatively high requirements for the input data, making predictions difficult, and the prediction accuracy is low. Financial markets, especially stock market data, are usually nonlinear data sets, which have the characteristics of nonstationarity, nonlinearity, high noise, and high instability. Traditional statistical models are difficult to use.

The LSTM model in DL eliminates the problems of ordinary recurrent neural networks by introducing the concept of cells, thereby achieving more accurate time series and predictions in a larger range, so it is more suitable for price predictions of financial products such as stocks and funds. By building a multilevel neural network, it converts shallow features into high-level features and thus has a relatively good predictive ability, which is very suitable for current stock market forecasts.
Figure 8: Dynamic forecast and actual value.

Figure 9: Static prediction results.

Table 4: Comparison of RMSE.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>1.319</td>
</tr>
<tr>
<td>ARIMA static prediction</td>
<td>9.838</td>
</tr>
<tr>
<td>ARIMA dynamic prediction</td>
<td>65.237</td>
</tr>
</tbody>
</table>
6. Conclusion

Stock market forecasting has always been the frontier of economics and finance in these years. The stock index forecasting model established in this article can predict the closing price of stocks on the second day more accurately, whether for the country, stock market managers, or stock investors. All have relatively large actual value. By establishing the LSTM model and the ARIMA model at the same time to predict the same data set, and then calculating the RMSE of different models, it is concluded that the prediction error of the LSTM model is much smaller than the ARIMA model, which gives the value of the DL prediction model constructed in this paper. At the same time, it reflects the advantages of DL in the direction of stock prediction from the perspective of quantitative analysis.

Although this article uses the DL method to carry out the above-mentioned research and analysis, there are still some deficiencies in the research of applying LSTM model to stock index prediction. This paper only constructs the long- and short-term memory network model in DL for prediction. In the future, LSTM model can be combined with CNN, BP model, etc. to further improve the algorithm advantages of LSTM. The input data of the LSTM model in this article only uses basic market data. In the future, stocks can be analyzed from multiple dimensions, such as inputting technical indicator data and more financial information data. This paper only uses fewer data sets and simpler LSTM functions. In the future, you can try to improve the preprocessing function, activation function, training function, etc. in the model, increase the number of data sets input to the model, and expand the time span and geography of the data set Breadth, thereby improving the predictive ability of the model.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors’ Contributions

Yiling Ding and Ning Sun contributed to this article equally.

Acknowledgments

This research was supported by the Fundamental Research Funds for the Central Universities (Grant Nos. HIT.HSS.201851 and HIT.HSS.202118).

References