Proceedings of ICFTBA 2024 Workshop: Finance in the Age of Environmental Risks and Sustainability
DOI: 10.54254/2754-1169/98/2024OX0186

A Study of Chinese Stock Price Prediction Based on LSTM and Time Series Linear Regression Model

Shuting Xiao[1,a,*]

[1]*Business School, King's College London, London, WC2R 2LS, United Kingdom*
a. shuting999@ldy.edu.rs
**corresponding author*

Abstract: With the advancement of technology and the increasing popularity of machine learning, new opportunities have emerged across various industries. In the financial sector, an increasing number of individuals are attempting to utilize machine learning to enhance the accuracy of stock price predictions. This paper will also endeavor to apply the LSTM model for predicting the stock prices of companies listed on the main boards of the Chinese stock market, while simultaneously comparing it with traditional time series linear regression model. Against the highly complex backdrop of the stock market, this paper aims to explore whether machine learning can surpass traditional models in achieving superior predictive results. However, in this research, the result did not show that LSTM overperform significantly linear regression model due to completely different economic and political background and limited parameters used. In order to improve the accuracy and stability of LSTM, it might be considered to incorporate additional influencing factors and combining with other models for market prediction.

Keywords: Stock Price Prediction, Machine Learning, Long Short-Term Memory, Time Series Linear Regression Model.

1. Introduction

In recent years, machine learning has experienced explosive growth, with varying degrees of application in different industries. Stock price prediction has long been a challenge in the financial industry, and it is also a topic of hot discussion. However, due to the strong volatility and high uncertainty of stock prices, as well as the influence of various factors from different aspect, there is considerable noise in the data, making accurate prediction of stock prices difficult. In order to continuously improve the accuracy of stock price prediction, scholars from both domestic and international communities have experimented with various prediction methods. Before the current technological advancements, stock price prediction was mainly divided into technical analysis and fundamental analysis methods. With the development of statistics, people began to utilize time series analysis for stock price forecasting, but for nonlinear and non-stationary series such as stock prices, the prediction accuracy was not high. With the popularity of machine learning, neural networks have gradually gained recognition and been applied in research, Studies have found that neural networks are highly suitable for analyzing non-stationary, random, and nonlinear problems. Therefore, many scholars have begun experimenting with neural networks for stock prediction, and the results have shown significant improvements in the accuracy of stock price prediction. However, there are still

DOI: 10.54254/2754-1169/98/2024OX0186

some intractable issues with commonly used neural networks that hinder the continuous improvement of prediction accuracy. For instance, BP(Back Propagation) neural networks are unable to capture the temporal information of stocks, while RNN (Recurrent Neural Network) networks only possess short-term memory. During model optimization, issues such as gradient vanishing or gradient explosion often occur, leading to a decrease in model accuracy. LSTM(Long Short-term Memory), on the other hand, can effectively address these issues. Therefore, this paper will attempt to utilize LSTM for stock prediction.

This paper will select two stocks from the Chinese stock market for prediction. These two stocks are listed respectively on the Shanghai and Shenzhen Main Boards and have been listed for over ten years, with relatively stable company conditions, to avoid abnormal fluctuations caused by special situations such as corporate bankruptcy. LSTM and Linear regression models will be separately used to predict the stock prices. The primary objective is to investigate whether the prediction accuracy of LSTM indeed outperforms traditional Linear regression models in the heavily regulated financial market of China. Furthermore, it will try to explore the advantages and disadvantages of LSTM in stock prediction through the experimental process, as well as potential methods for improvement, thereby contributing new empirical results to the application of machine learning in stock prediction.

2. Literature Review

2.1. Stock Price Prediction

The development of stock prediction has spanned a long period, evolving from traditional time series models to machine learning models, and now to deep learning neural network models. Stock price prediction has accumulated ample experimental and theoretical support, and its prediction accuracy has been constantly improving.

As early as 1933, Cowles et al. analyzed thousands of stocks over a five-year period but found that the stock market was unpredictable [1]. Then, in 1970, Fama proposed the Efficient Market Hypothesis, arguing that all information in the market is already reflected in stock price movements, and investors cannot obtain excess returns from historical analysis [2]. However, research on stock price prediction did not disappear. Scholars began to experiment with simple linear models to process stock data. In the 1980s, Shiller et al. studied the linearization of the rational expectations present value model of stock prices and found that it could achieve good results in predicting stock returns [3]. In 1982, Engle innovatively proposed the Autoregressive Conditional Heteroskedasticity model (ARCH model), which effectively simulated the clustering and time-varying nature of stock volatility [4]. Liu established an ARMA model to predict the stock price of Ansteel after first-order decomposition of its price series, achieving good but still not ideal prediction results [5]. Tang et al. combined the ARMA and GARCH models to predict stock prices and also used the ARMA model alone [6]. They found that the combined model outperformed the individual models. While scholars' research has shown that stock prices can be predicted using historical data, due to the large amount of noise and nonlinear fluctuations in stock data, as well as the increasing duration of predictions, linear models have consistently struggled to achieve ideal prediction results.

With the rise of machine learning, scholars began to experiment with nonlinear models for stock price prediction. They introduced machine learning models such as Support Vector Machines (SVM), Random Forests, and Gradient Boosting Trees and successfully applied them to stock prediction. Zhang used SVM with technical indicators to accomplish stock prediction, achieving approximately 65% accuracy through experimentation [7]. However, it was found that as the training sample size increases, SVM encounters greater difficulty in solving quadratic optimization problems. Zhang et al. used a Random Forest model with 16 technical indicators to predict stock market conditions [8]. Backtesting results showed that the model could handle large amounts of data while maintaining high

DOI: 10.54254/2754-1169/98/2024OX0186

prediction accuracy. Krauss et al. predicted the S&P 500 index using Random Forests, Gradient Boosting Trees, and deep learning, and found that the stock selection model constructed using Gradient Boosting Trees outperformed the other two methods [9]. Zhang et al. combined technical indicators with a Gradient Boosting Tree model to predict the CSI 300 index, and the results showed significantly higher accuracy compared to linear regression models and Random Forest models [10]. However, traditional machine learning prediction methods require preprocessing and feature engineering for complex and large amounts of prediction data.

With the recent advancements in computer hardware and the emergence of dataset sharing, deep learning has experienced rapid development, and stock price prediction research has gradually expanded to deep learning neural networks. Through extensive research, scholars have begun to introduce various neural networks such as CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), and LSTM (Long Short-Term Memory) to predict stock prices, aiming to further enhance prediction accuracy.

2.2. Neutral Network Development and Application in Stock Prediction

Neural networks are designed based on the biomimetic principle of the neural network structure of the human brain. They consist of numerous interconnected neuron nodes, where weights are assigned and adjusted. These networks utilize specific activation functions to activate the cell states and generate outputs. Due to this design, neural networks possess exceptional nonlinear capabilities and powerful learning abilities.

Kimoto et al. established the TPOIX model based on neural networks to predict the weighted average index of the Tokyo Stock Exchange in Japan [11]. The experimental results showed that the prediction performance of this model was superior to the standard weighted average index model. Zhang et al. found in their research that the accuracy of neural network models in predicting nonlinear time series data is far higher than that of ARIMA models [12]. Shinkai-ouchi et al. conducted stock market predictions using Bayesian estimation and neural networks separately, and the neural network model still achieved better prediction results [13]. Zhang et al. compared the stock price prediction performance of logistic regression and LSTM neural network models, and LSTM demonstrated better prediction results with a lower MSE [14]. Huang et al. simultaneously utilized BP, LSTM, CNN, RNN, and GRU neural network models for prediction and compared the loss functions MAE, MSE, and MAPE [15]. Among them, the LSTM model still exhibited the smallest loss.

The utilization of neural networks for stock market prediction has become increasingly prevalent, with diverse exploration directions. However, overall, prediction models employing neural networks generally outperform traditional prediction models. Among these, the LSTM model has emerged as a standout among numerous neural network algorithms.

2.3. Summary

The advent of the artificial intelligence era has indeed broken the limitations of traditional prediction models. Based on the aforementioned research, neural network algorithms, particularly the LSTM model, have indeed improved the accuracy of model predictions. However, their application in the Chinese financial market remains relatively scarce. The Chinese stock market possesses unique characteristics such as the price limit system and the T+1 trading system. Under these specific conditions, it remains to be verified whether LSTM can maintain a high level of prediction accuracy and whether it can achieve significantly better accuracy compared to traditional linear regression models. This paper will utilize Python to apply LSTM and linear regression models to predict stocks in the Chinese market and compare the differences between them.

DOI: 10.54254/2754-1169/98/2024OX0186

3. Methodology

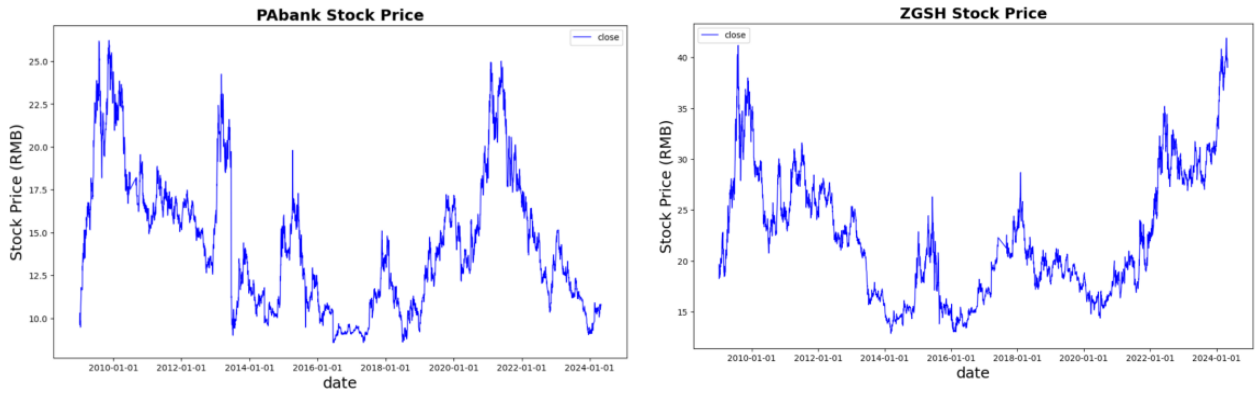
3.1. Data Selection and Processing

This paper will select two stocks listed on the Shanghai Stock Exchange and Shenzhen Stock Exchange respectively - Ping An Bank Co. Ltd (000001, named PABank in the rest paper) and China ShenHua Energy Co. Ltd (601088, named ZGSH in the rest paper), as shown in Table 1. Daily data for all trading days from January 1, 2009, to April 31, 2024, of these two stocks will be obtained from the official website of Shenzhen Securities Information Co., Ltd. for model training and testing, as Figure 1shown. The original data includes daily stock opening prices, closing prices, high prices, low prices, trading volumes, and transaction amounts. Only closing prices will be used to predicted in this research.

Due to the presence of exceptional suspensions and other situations in the acquired data, the stock prices for some dates are missing. Therefore, before model training and backtesting, the corresponding dates with missing data are excluded, and then the overall sample is divided into training set and test set. The first 80% of the data for the two stocks is used as the training set to train the model, and the remaining 20% of the data is used as the test set to verify the generalization ability of the model.

Table 1: Dataset selected.

Stock	Total volume	Volume without null	Training set	Testing set
PABank (000001)	3723	3641	2897	724
ZGSH (601088)	3723	3655	2908	727



(a) PABank

(b) ZGSH

Figure 1: Stock price trend.

3.2. Linear Regression Model

In this article, Python is selected as the programming language for implementation, and the Linear Regression Model from sklearn is introduced to predict the closing price of the 21st day based on the closing prices of the past 20 days. The regression equation is as follows:

$$ClosePrice_{21}(y) = \beta_1 * ClosePrice_1 + \beta_2 * ClosePrice_2 + \ldots + \beta_{20} * ClosePrice_{20} \quad (1)$$

As this article only utilizes the closing price as a feature for price prediction, it needs to prepare time series data using historical data. In the original data, it will add a lagged column representing

DOI: 10.54254/2754-1169/98/2024OX0186

the data for the 21st day to construct the time series data. It finally got 3620 and 3634 sets of data respective for PABank and ZGSH.

3.3. Long Short-Term Memory Model

3.3.1. The Principal of LSTM

LSTM, or Long Short-Term Memory, is a type of time-recursive neural network that evolved from RNN and was first introduced by Hochreiter and Schmidhuber in 1997. LSTM was primarily proposed to address the short-term memory issues of RNNs as well as the problems of gradient vanishing and gradient explosion that often arise during training.

The network structure of LSTM consists of three gates and a memory cell. These three gates are the forget gate, the input gate, and the output gate, while the memory cell retains historical memory information. Below are the computational principles and relevant formulas for the three gates.

The forget gate determines which information to discard from the cell state. It uses the sigmoid activation function to output a value between 0 and 1 for each element in the cell state. The mathematical formula for the forget gate is as follows:

$$f_t = \sigma(W_f * [h_{t-1}, X_t] + b_f) \tag{2}$$

In the input gate section, it is necessary to determine the information to be added to the cell state. This involves two parts: the input gate, which uses the sigmoid activation function to determine the values to be updated, and a tanh layer, which creates new values to be added to the cell state. The mathematical formulas for these are as follows:

$$i_t = \sigma(W_i * [h_{t-1}, X_t] + b_i) \tag{3}$$

$$C_t = tanh(W_C * [h_{t-1}, X_t] + b_C) \tag{4}$$

Next, using the above results can calculate the updated cell state C_t' at time t. The formula can be expressed as follows:

$$C_t' = f_t * C_{t-1}' + i_t * C_t \tag{5}$$

Finally, the value of the output gate could be determined. The formula can be expressed as follows:

$$O_t = \sigma(W_O * [h_{t-1}, X_t] + b_O) \tag{6}$$

$$h_t = O_t * \tanh(C_t') \tag{7}$$

where, σ: Sigmoid activation function; W_f: Weight matrix for the forget gate; W_i: Weight matrix for the input gate; W_c: Weight matrix for the cell state update; W_o: Weight matrix for the output gate; b_f: Bias for the forget gate; b_i: Bias for the input gate; b_c: Bias for the cell state update; b_o: Bias for the output gate; h_t: Output at time t; C_t: Updated cell state at time t.

In order to gain better understanding of LSTM, Figure 2 below is a diagram illustrating the structure of the LSTM model.

DOI: 10.54254/2754-1169/98/2024OX0186

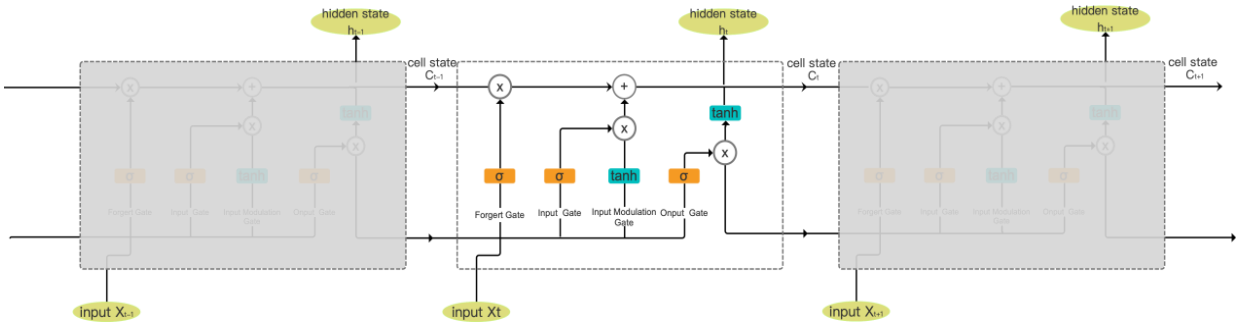


Figure 2: LSTM basic structure (Photo/Picture credit: Original).

3.3.2. Design and Implementation of LSTM

This paper utilizes Python language and the PyTorch deep learning framework to implement the LSTM model. The experimental process includes the following key steps, as Figure 3 shown. Firstly, the data is normalized to enable the model to converge better. The model is then trained using the aforementioned partitioned training data, and the training loss is observed to confirm whether the model has achieved effective convergence. If not, the parameters are fine-tuned to achieve better convergence. After the model achieves a satisfactory convergence effect, testing is performed using the test set.

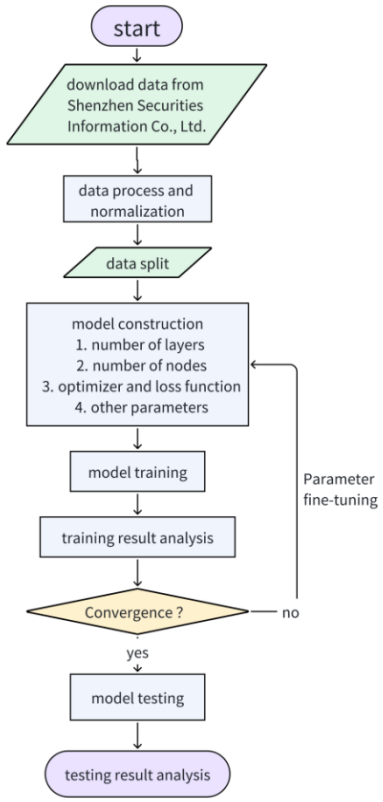


Figure 3: LSTM process (Photo/Picture credit: Original).

The main adjustable parameters involved in the entire process are as follows: the number of neurons in the hidden layer, the number of hidden layers, the number of training epochs, and the learning rate. During the parameter selection process, the number of layers and neurons are crucial.

DOI: 10.54254/2754-1169/98/2024OX0186

When the number of neurons or layers is too small, underfitting may occur; while when the number of neurons or layers is excessive, overfitting may arise. After multiple experiments, as for PABank, it finally achieved the best results when selecting 64 neurons in the hidden layer and two hidden layers. In the model training, it is allowed the model to undergo 310 training epochs, and set the learning rate to 0.01. In terms of ZGSH, the best result occurs when selecting 64 neurons in the hidden layer and two hidden layers with 230 training epochs and same learning rate.

3.4. Model Evaluation

After completing the model construction and testing, MSE, RMSE, and R^2 will be used to evaluate the regression performance of the model. MSE measures the degree of fit of the model by calculating the mean of the squared differences between the predicted values and the actual observed values. A smaller MSE value indicates better model prediction performance. RMSE is the square root of MSE and is more sensitive to larger errors. When dealing with problems with larger numerical values or larger datasets, RMSE can better reflect the accuracy of predictions. R^2, also known as the goodness of fit, represents the degree of correspondence between the predicted results and the actual occurrences. The closer R^2 is to 1, the better the fit.

$$MSE = \frac{1}{N} * \sum_{n-1}^{N} (y_p - y_n)^2 \tag{8}$$

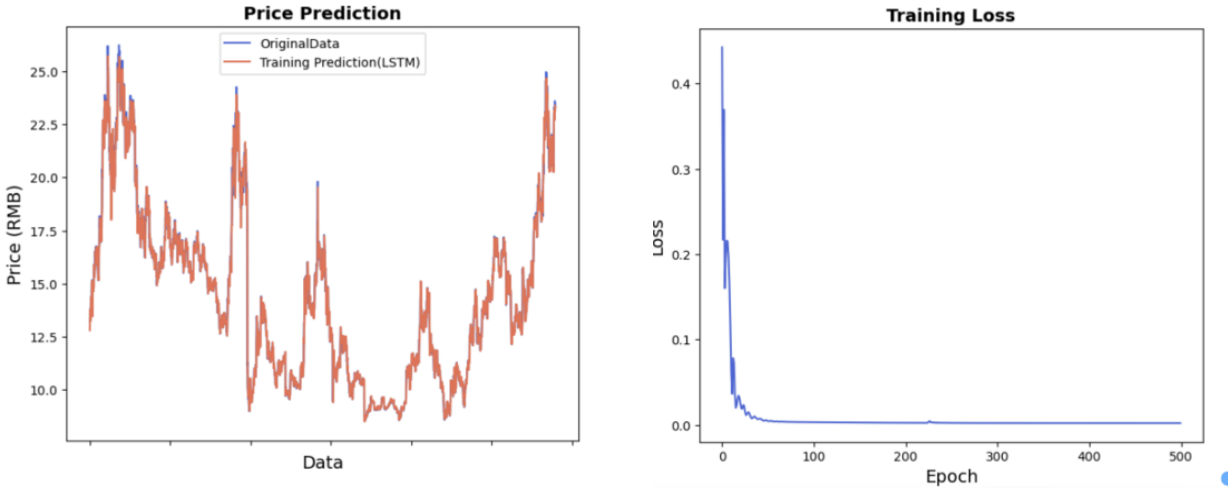
$$R^2 = \frac{\sum_{n-1}^{N} (y_p - y_m)^2}{\sum_{n-1}^{N} (y_n - y_m)^2} \tag{9}$$

Where y_p represents the predicted value, y_n represents the actual value, y_m represents the mean value of the actual values, and N represents the number of data points.

4. Results

4.1. Description

The parameters selected by LSTM after multiple experiments have converged well, and its training loss is as shown in Figure 4 and Figure 5.

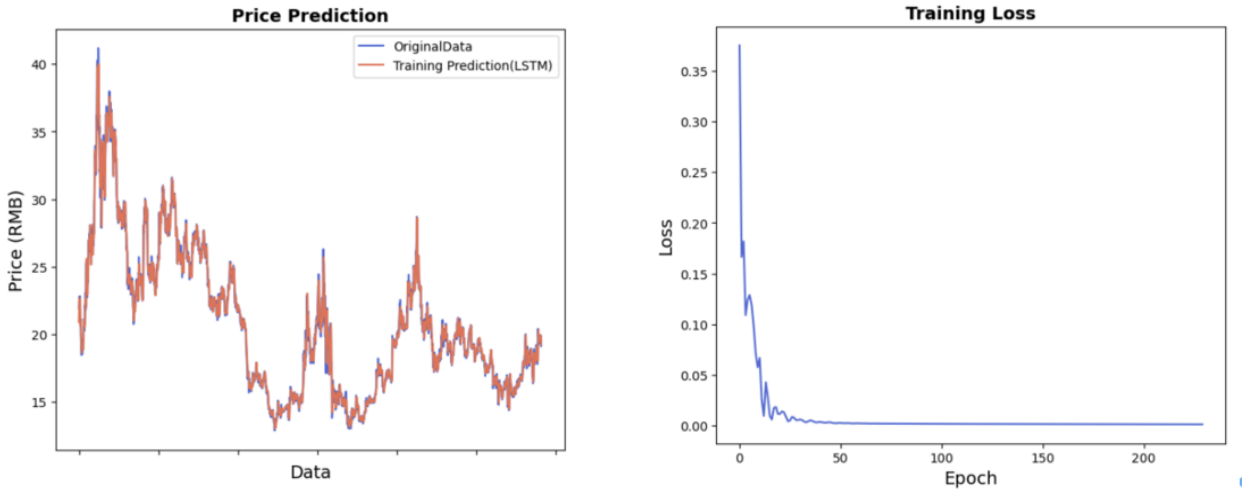


(a) Price prediction

(b) Training loss

Figure 4: PABank (LSTM)'s price prediction and training loss.

DOI: 10.54254/2754-1169/98/2024OX0186



(a) Price prediction (b) Training loss

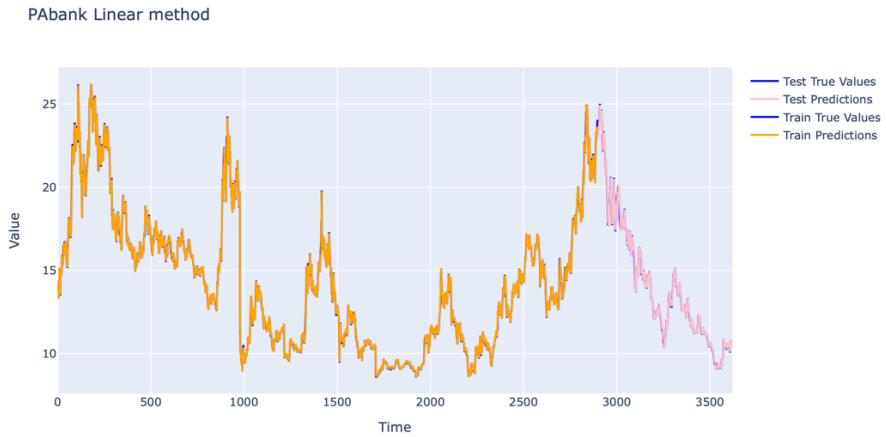
Figure 5: ZGSH (LSTM)'s price prediction and training loss.

Table 2 shows the results of error values and fitted values after conducting linear regression and LSTM model predictions on two separate stocks. As evident from the data in Table 2, both models achieved similar results, indicating that LSTM did not significantly outperform the linear regression model. However, both models still achieved a satisfactory level of fitting.

Table 2: Prediction results.

STOCK	PABank		ZGSH	
MODEL	LINEAR	LSTM	LINEAR	LSTM
Train MSE	0.1531	0.1578	0.2247	0.2919
Test MSE	0.0870	0.0894	0.3797	0.4441
Train RMSE	0.3913	0.3972	0.4741	0.5403
Testing RMSE	0.2950	0.2990	0.6162	0.6664
Training R^2	0.9908	0.9905	0.9921	0.9897
Testing R^2	0.9938	0.9937	0.9873	0.9851

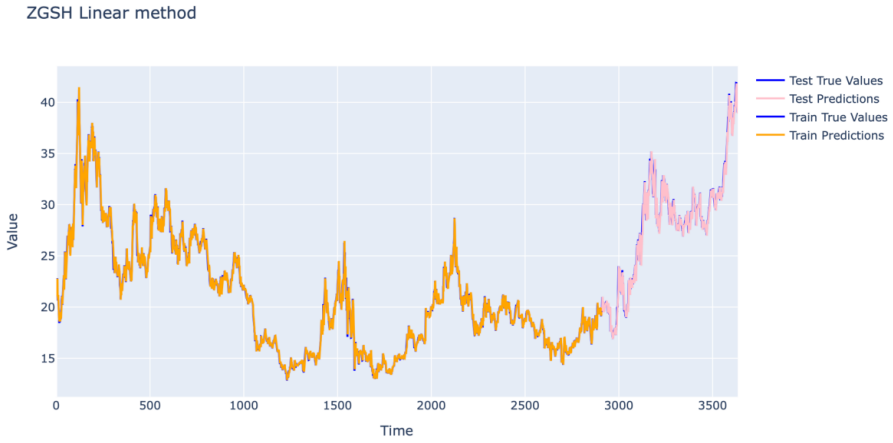
Figure 6 shows the fitting images of two stocks using two different methods.



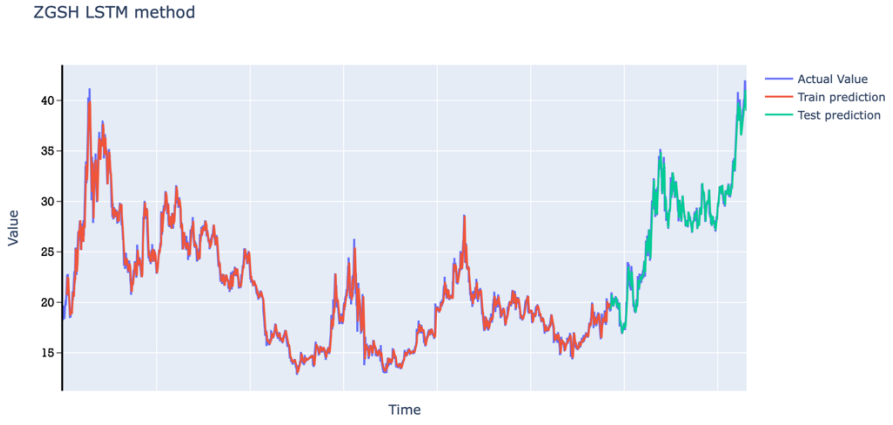
(a) PABank Linear method

DOI: 10.54254/2754-1169/98/2024OX0186

(b) PABank LSTM method



(c) ZGSH Linear method

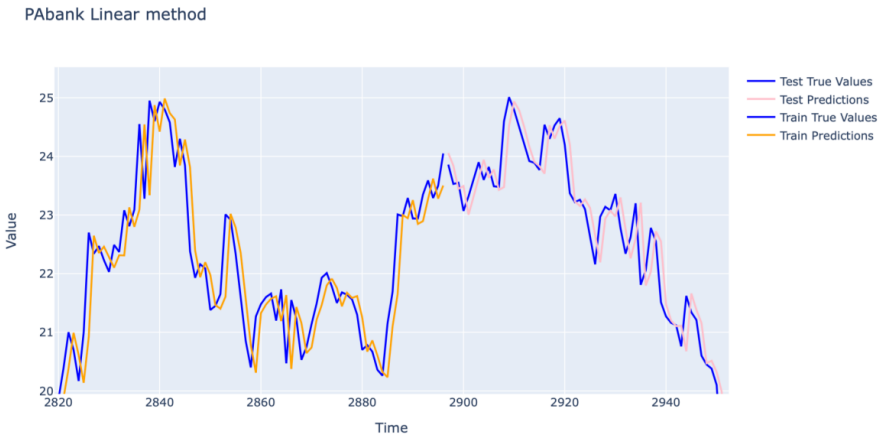


(d) ZGSH LSTM method

Figure 6: Prediction results of different methods.

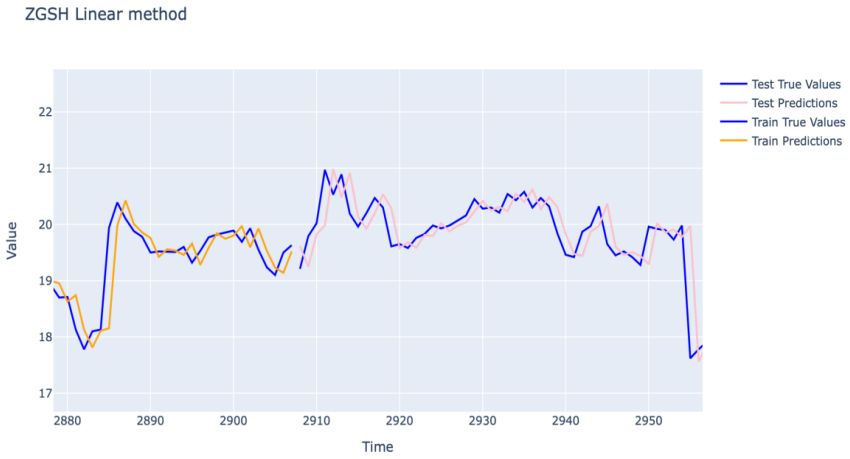
From Figure 7, it can be observed that both prediction methods are able to relatively well fit the future trends of the stocks. However, upon closer inspection by zooming in on local trends to gain a clearer understanding of their specific prediction situations, a certain degree of lag is evident in both prediction methods. Additionally, LSTM exhibits a significant discrepancy from the actual values in peak prediction.

DOI: 10.54254/2754-1169/98/2024OX0186



(a) PABank Linear method (part)

(b) PABank LSTM method (part)



(c) ZGSH Linear method (part)

DOI: 10.54254/2754-1169/98/2024OX0186

(d) ZGSH LSTM method (part)

Figure 7: Prediction results of different methods.

4.2. Analysis

Based on the experimental results, both prediction methods achieved satisfactory outcomes, not only exhibiting good performance in terms of error and fitted values, but also demonstrating a close alignment between the predicted trends and the actual trends in the final prediction charts. This may be attributed to the fact that both stocks, being stocks with relatively stable fundamentals and consistently good performance in the Shanghai and Shenzhen stock markets over a long period, exhibit fewer abnormal fluctuations within a short period of time (one month).

However, in predicting the two stocks, the performance of LSTM did not significantly outperform the linear regression model. There are potentially two reasons for this. Firstly, this study only utilized the closing price as a feature for predicting stock prices, which contains limited predictive information. As a result, it was difficult for LSTM to demonstrate a significant advantage over the linear model. Additionally, the daily price fluctuation limits in the Chinese market differ from those in European and American markets. the existence of these limits means that it is difficult for stock prices in the Chinese market to experience significant fluctuations in a short period of time. This may also indicate that linear characteristics are more prominent in the short term. Sun also illustrated that LSTM outperform in Chinese Market due to lack of sufficient parameters [16]. many other researchers, like Bao(2020),Chen(2019),et.al, suggested that it is helpful for LSTM to combine with other methods or increase more factors to improve its accuracy and stability. Therefore, these might be reasons why there is no huge difference between LSTM and Linear model.

Furthermore, both methods exhibited a certain degree of lag in fitting the stock price prediction curves compared to the actual values, and there were significant lags and numerical differences in depicting the peaks. This could be attributed to the fact that this study only utilized the closing price to predict stock prices, whereas there are numerous factors that influence stock market trends in the market, especially in the Chinese market. The Chinese market is heavily influenced by policies, and the majority of investors in this market are individual investors. These factors contribute to the difficulty in accurately predicting the price of a specific day in the future solely based on historical prices.

5. Conclusion

The stock price system itself is an extremely complex market influenced by multiple factors, and predicting stock prices is a dynamic, irregular, and complex process. While LSTM shows a certain

DOI: 10.54254/2754-1169/98/2024OX0186

degree of predictability in the Chinese market for forecasting the overall trend of stock prices, whether its performance is significantly superior to the linear regression model still requires further investigation. To further enhance the predictive capabilities of LSTM or the linear regression model for the Chinese market, consider including other more influential features for prediction, such as sentiment factors, fundamental information, technical indicators, and so on. Additionally, for the LSTM model, it can be combined with other methods, such as principal component analysis (PCA) or k-means, for optimization and prediction to achieve better results.

References

[1] *Cowles 3rd, A. (1933). Can Stock Market Forecasters Forecast? Econometrica, 1(3), 309–324.*

[2] *Fama, E. F. (1970). Efficient capital markets. Journal of finance, 25(2), 383-417.*

[3] *Campbell, J. Y., & Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. The review of financial studies, 1(3), 195-228.*

[4] *Eagle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica, 50(4), 987-1007.*

[5] *Liu, H.M. (2008). Application of ARIMA Model in Stock Price Prediction. Guangxi Journal of Light Industry, 24(6), 92-93.*

[6] *Tang, H., Chiu, K. C., & Xu, L. (2003). Finite mixture of ARMA-GARCH model for stock price prediction. In Proceedings of the Third International Workshop on Computational Intelligence in Economics and Finance (CIEF'2003), North Carolina, USA (pp. 1112-1119).*

[7] *Zhang, Y. C., & Zhang, Z. Q. (2007). Application of Support Vector Machine in Stock Price Prediction. Journal of Beijing Jiaotong University, 31(6), 73-76.*

[8] *Zhang, X., & Wei, Z. X. (2018). Application of Random Forest in Stock Trend Prediction. China Management Informationization, 21(3), 120-123.*

[9] *Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689-702.*

[10] *Zhang, X., Wei, Z. X., & Yang, T. S. (2018). Application of GBDT Ensemble Model in Stock Prediction. Journal of Hainan Normal University: Natural Science Edition, 31(1), 73-80.*

[11] *Kimoto, T., Asakawa, K., Yoda, M., & Takeoka, M. (1990, June). Stock market prediction system with modular neural networks. In 1990 IJCNN international joint conference on neural networks (pp. 1-6). IEEE.*

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

[13] *Shinkai-Ouchi, F., Koyama, S., Ono, Y., Hata, S., Ojima, K., Shindo, M., ... & Sorimachi, H. (2016). Predictions of cleavability of calpain proteolysis by quantitative structure-activity relationship analysis using newly determined cleavage sites and catalytic efficiencies of an oligopeptide array. Molecular & Cellular Proteomics, 15(4), 1262-1280.*

[14] *Zhang, X. C., Xu, X. P., & Wei, S. L. (2020). Application of LSTM Neural Network in Stock Price Prediction. Computer Knowledge and Technology, 16(28), 39-43.*

[15] *Huang, C. B., & Cheng, X. M. (2021). 'Research on Stock Price Prediction Based on LSTM Neural Network'. Journal of Beijing Information Science and Technology University (Natural Science Edition), 36(1), 79-83.*

[16] *Sun, R. (2016). Research on the prediction model of American stock index trend based on LSTM neural network, Capital University of Economics and Business.*