



Application of LSTM Models in Stock Return Forecasting

Yanning Li 

Department of Management, Beijing Institute of Technology, Beijing 100000, China
1120233232@bit.edu.cn

Abstract. With the accelerated volatility of the stock market and significant differentiation of individual stock returns in recent years, accurate prediction of individual stock returns has become increasingly crucial for financial investment. However, practical issues remain regarding the stability of existing prediction models, necessitating further exploration of reliable return prediction tools to meet market needs. This paper focuses on the application of LSTM models in individual stock return prediction, aiming to capture nonlinear characteristics for more accurate forecasts. It first defines concepts of LSTM, ARIMA, and other models, and reviews their existing applications as a theoretical basis. Subsequently, the LSTM model is constructed, explanatory variables are selected, and test results are visually presented. Finally, robustness evaluation indicators such as RMSE and MAE are adopted to analyze the model's fitting effect. The study finds that LSTM can fully account for nonlinear time-series dependencies in individual stock returns and achieve certain results in capturing complex market patterns, but it also has shortcomings, including low direction prediction accuracy and training efficiency. Thus, increasing model capacity and selecting more characteristic data is recommended to improve prediction accuracy.

Keywords: Stock Return, Earnings Forecasts, LSTM Model.

1 Introduction

In recent years, the development of science and technology has made it an important driving force of the stock market, and the stock market has shown the characteristics of structural differentiation, which has aggravated the overall volatility of stock returns. At the same time, the public's interest in the stock market has risen sharply, and the participation of more investors has led to a demand for more accurate stock market predictions to achieve higher-than-average returns. As one of the challenging key contents in financial research, the prediction of the stock market is the key sector to improve the yield. With advances in machine learning and algorithmic technology, numerous studies have examined stock forecasting from different perspectives using various forecasting models [1].

Previous studies usually used time series models based on historical data, such as the autoregressive average moving model, the autoregressive conditional

heteroscedasticity model, etc [2]. This method requires high data smoothness, which does not conform to the natural characteristics of stock market data. On this basis, the improved recurrent neural network overcomes this problem to some extent, but it is prone to gradient disappearance and unable to remember information long ago [3]. As an advanced version of RNN, the long-term and short-term memory network forms a time-series model of controllable memory through its internal core structure: forgetting gate, input gate, and output gate, and has stronger long-term information processing ability [4]. Due to its natural time series processing ability and the automatic learning nonlinear characteristics that fit the background of the digital and intelligent era, LSTM matches the demand of the individual stock return prediction model characterized by time series, long-term dependence, and nonlinear prediction, and is a good choice of return prediction model. Existing studies have confirmed that the LSTM model can effectively capture the price fluctuation trend of the stock market [5]. Relevant research has verified and proposed that a single-layer LSTM can achieve more accurate model fitting by predicting the closing price through a multivariate input characteristic series [6].

At present, the prediction of the LSTM model has projected the target to the stock market, but the prediction of individual stock returns is still lacking. In the selection of model variables, the selection of data lacks the discussion of special cases, and does not take into account external factors such as market sentiment and industry linkage. Taking the above research gaps into account, this study follows the research method of data preprocessing, model building, model training, and prediction evaluation. In terms of environment configuration. This study builds an LSTM model and obtains data from the CSMAR database. This paper aims to predict the return rate of individual stocks in the stock market through the single-layer LSTM model, and evaluate the fitting effect by using the *t*-test.

2 Relevant Theory and Model Basis

Individual stock returns measure the level of profits investors get from their investments and are affected by their own trading characteristics and market sentiment fluctuations. Among them, its own trading characteristics also refer to basic trading indicators, including core stock trading data such as daily opening price, daily maximum price, and daily minimum price, which reflect the mixed daily limit fluctuation range of the stock pricing benchmark. Market sentiment corresponds to the activity of funds in the stock market and directly reflects the attention of market funds to individual stocks, which is expressed as daily trading volume [7]. Due to the common influence of many factors, investor sentiment usually fluctuates, making it difficult to accurately quantify. Therefore, the nonlinear relationship between the stock return as the explained variable and the explanatory variables affecting the return rate is the difficulty of accurately predicting the return. The core of long-term and short-term memory is to regulate the flow of information into and out of memory units through a set of gates, to realize effective learning of long-term relationships and effective prediction [8].

In terms of data sources, the study extracted typical individual stocks from the CSMAR database. As the largest securities company in China, CITIC Securities Co., Ltd. has a stable market value at the top of the securities business sector, high trading activity, and earnings fluctuations are not only affected by the cycle of the securities industry, but also related to the trend of the capital market. With complete data, it can ideally reflect the nonlinear time characteristics of individual A-share earnings, and is selected as the stock that fits the LSTM model. The data time range is defined as 2020-2024, and transaction information is collected. By filtering the core fields and eliminating the missing values, the data is preliminarily cleaned. The 3σ criterion and box plot method are used to identify abnormal values, and the Z-score standardization method is used to standardize the dimensions of explanatory variables to ensure the consistency of data and the matching with the model.

3 Construction and Implementation of an LSTM Model

3.1 Core Elements of Model Construction

Data preprocessing is required before training the LSTM model for prediction. The first core check is to check whether the data information is empty, filter the core column through the mapping code, retain the key column, and delete the null value and non-numeric rows of the date column. For the null value of the characteristic column, this study adopts a combination of linear interpolation and elimination for missing values in feature columns. For the processed value, calculate the logarithmic yield and check the data again. The second part of the core verification is the processing of data outliers. First, define the `process_outliers` function to clean up extreme outliers, such as negative prices. Then, the global and local outliers are identified by using 3σ and a box graph, respectively, and the dual validity test is carried out. In addition, multiple collinearity tests are added in this study. To facilitate the subsequent model fitting, the processed data set adopts the Z-score standard. According to LSTM's input requirements, time-series samples are structured as three-dimensional data and split into training, validation, and test sets at a ratio of 7:2:1. Considering that there may be errors in the model, exception capture is added during the call to facilitate the positioning problem.

After data processing, the LSTM model is built using TensorFlow and the Keras library. In order to ensure the fitting accuracy of the verification results, this study uses the `build_lstm_model` to define the single-layer LSTM architecture. The input layer is set with dimensions of (60,6), i.e., 60-time steps and 6 characteristic numbers. Using the characteristics of 6 types of explanatory variables in the past 60 trading days, the log return on the 61st trading day is predicted. The single hidden layer contains 32 LSTM neurons, which activate the tanh function to capture the long-term and short-term dependencies in time series data. Add a dropout layer after the hidden layer, determine 50 as the maximum training round, and set the early stop mechanism. Randomly discarding 20% of neurons is used to inhibit the over-fitting of the model and reduce the redundant association between neurons. The output layer is a linear activation function containing one neuron and outputs the predicted logarithmic rate

of return. According to the structural characteristics of the LSTM model, the key to the construction is the `create_sequences` function, which converts the characteristic variables into the three-dimensional time series samples required by the LSTM, and predicts and evaluates the training set, validation set, and test set, respectively.

3.2 Implementation Characteristics and Verification Ideas

The study's implementation process focuses on the core goal of individual stock return prediction: prediction accuracy and timeliness. Therefore, it is necessary to achieve the deep integration of LSTM models. This study has achieved this goal from three aspects: data selection, code programming, and model building. Firstly, the selection of data covers six types of explanatory variables, and through data cleaning to eliminate dimensional differences and data noise, three-dimensional time series samples that meet the input requirements of LSTM are constructed. Secondly, Python imports the library adapted to LSTM, splits the code according to different steps, realizes logical progression, avoids syntax errors, and adds exception handling to the code, providing support for the screening of LSTM training effect. Finally, the model uses a simplified single-layer LSTM and sets early stopping to avoid overfitting.

Mean Absolute Error and mean square error, as common indicators of evaluation models, are widely used to evaluate the fitting effect of regression analysis. Their advantages lie in clearly reflecting the regression model's performance regardless of the range and distribution of true values. The determination coefficient R^2 achieves the performance of effectively indicates the accuracy of the model. Therefore, considering the advantages and disadvantages of different methods, this study adopts MSE, Mae, and R^2 tests to comprehensively evaluate the prediction results of the model and ensure the effectiveness of the test results.

4 Analysis of model prediction effect

4.1 Model Fitting Effect

In order to ensure the accuracy of the conclusion, the model fitting effect is analyzed first. According to the training loss curve of Fig. 1, with the increase of training rounds, the training loss shows an overall continuous downward trend, and gradually converges and stabilizes in the range of about 0.0003-0.0004. This trend shows that in the process of training time series data, the relationship between the fluctuation law of daily logarithmic yield of CITIC Securities and time series data is gradually captured through parameter optimization within the model, which verifies the effectiveness of the training process. The validation loss curve performs better than the training loss curve. On the basis that the basic trend of overall decline to stability remains unchanged, as the number of training increases, the model tends to improve, and the verification loss curve shows a much lower loss rate than the training loss curve from the initial stage, and rapidly decreases with the number of training rounds. In the middle stage, the decline rate of verification loss has slowed down and gradually stabilized. After 25 rounds, the verification loss remained basically unchanged.

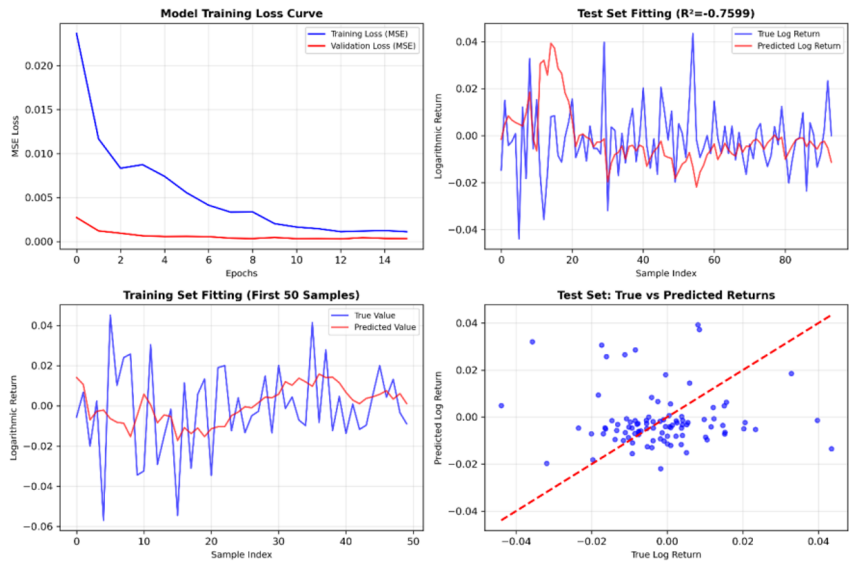


Fig. 1. Model effect diagram (Picture credit: Original).

In order to avoid invalid training and over-fitting, an early stopping mechanism is pre-set in the code; the model stops training when validation loss does not decrease for five consecutive epochs through `val_loss` monitoring. This process avoids the loss of time series data caused by invalid training and ensures the convergence of the model. Finally, the difference between the stable value of training loss and verification loss is small, and there is no obvious contradiction between the two curves, which proves that the model does not have over fitting or under fitting in the training process, and achieves the ideal training effect.

4.2 Validation of Prediction Results

To enhance fitting accuracy, the first two periods of CITIC Securities' 2026 data are added to the test set to verify the LSTM model's prediction effect. Root Mean Square Error and Mean Absolute Error are selected as evaluation indicators to measure the overall error range and prediction stability, while the fitting line reflects the linear relationship between predicted and true values. Test results show the test set's RMSE 0.0167 and MAE 0.0124 are lower than the training set's, RMSE=0.0208, MAE=0.0150, indicating better generalization ability. RMSE remains at a low level, confirming predicted values are close to the real return average. Residuals are concentrated in [-0.02, 0.02] and symmetrically distributed, proving no systematic deviation, though some samples have large prediction deviations and extreme residuals. Robustness analysis shows the test set's lower error than the training set, verifying error stability across data subsets and effective data segmentation. The low proportion of extreme residuals confirms the model's stability for most samples and the robustness of the residual distribution.

Compared with traditional methods like ARIMA and ARMA, which are confined to linear autocorrelation, LSTM can capture nonlinear time-series momentum in individual stock returns, directly integrate multi-dimensional features like price and trading volume, and achieve multi-feature fusion for prediction. With sufficient data and rich features, LSTM inherently has the potential to break through average prediction limitations, capture more complex market patterns, and maintain long-term prediction stability. However, LSTM's innovations also bring certain limitations: visual prediction results show low direction prediction accuracy, failing to highlight its nonlinear advantages. This is because LSTM cannot clarify autocorrelation order as clearly as ARIMA and requires substantial high-quality data and features to support fitting for more accurate results. These requirements consume significant prediction time and require collecting more parameters, making it less efficient than ARIMA. At present, whether through simple linear fitting or LSTM models, breaking through the white noise limit of stock return data remains challenging.

In order to further improve the lack of direction prediction of LSTM, the following suggestions for model optimization are proposed. At the model level, feasible solutions include increasing the number of LSTM layers, increasing the number of units, and increasing the model capacity, so as to obtain more powerful data support. Adjust the time window step size to capture more accurate time sequence rules and enhance the accuracy of prediction direction. In terms of data selection, macro factors including inflation rate, market sentiment and industry index can be incorporated to enrich parameter characteristics, and market periods with greater earnings volatility can be focused on to improve the effectiveness of information collection.

4.3 Analysis of Practical Application Value

According to the above analysis results, it can be seen that the LSTM model has the advantages of low risk and low error for mean value prediction, and has the limitation of inaccurate capture of time series data. Considering the return on investment and risk control, this paper discusses the strategic reference that the LSTM model can provide for different types of investors.

For individual investors with a small amount of capital and moderate risk tolerance, the core application scenarios of LSTM focus on the auxiliary tools for long-term holding, such as risk prediction and position control. As the predicted value of the model is close to the average value, the short-term yield fluctuation shows a convergence trend, which can better reflect the average signal of the market and provide an effective reference for the control of positions. However, the effect of the model on direction prediction is weak, and short-term transactions and time prediction should be avoided when using it. For institutional investors with a large amount of funds and strict risk control, if RMSE is considered to be included in the weight calculation, the LSTM model can realize the portfolio prediction with minimum variance, so as to optimize the quantitative portfolio. The specific implementation strategy is to carry out risk stratification for individual stocks, reduce the individual stocks with large forecast volatility, improve the weight of stable individual stocks, and incorporate the model forecast volatility as a risk factor into the prediction of position allocation. To sum up, the core application value of LSTM lies in robust risk

control rather than accurate time series prediction, which has higher potential and value in long-term configuration and quantitative strategy.

5 Conclusion

Using CSMAR stock data from 2015 to 2025 and incorporating the first two periods of 2026 samples, this paper constructs an LSTM model to predict logarithmic returns. The results show the model converges stably: training loss decreases from 0.0069 to 0.0011, validation loss is consistently lower than training loss without an upward trend, and there is no overfitting but slight underfitting. The test set's RMSE and MAE are lower than the training set's, indicating good generalization performance. However, the low slope of the test set's fitting line reflects weak direction prediction accuracy, limiting the model's practical timing value. Residuals are symmetrically concentrated in $[-0.02, 0.02]$ without systematic deviation, showing strong prediction stability, though extreme residuals pose risks for some samples. This mainly stems from individual stock log returns being similar to white noise with scarce effective time-series signals; despite LSTM's nonlinear modeling ability, it cannot break through financial data's stochastic nature. Adding 2026 data reduces the error range but fails to improve direction prediction, confirming the inherent difficulty of short-term return forecasting. Financial investors are advised not to rely solely on LSTM results but to use the model as a risk control auxiliary, focusing on long-term portfolio optimization based on comprehensive fundamental and technical judgments.

This paper's main contribution is to systematically verify LSTM's performance in individual stock return prediction. Compared with traditional methods such as Arima, it is clear that LSTM has the advantages of multi-feature fusion, error stability, and generalization robustness. This is conducive to providing a more realistic empirical basis for financial time series prediction. It puts forward suggestions on differentiation strategies from the perspective of investors, makes up for the gap between model research and practical application, and provides a reference framework for risk control for individual and institutional investors. There are still some deficiencies in the current research, including the mild under-fitting caused by the limited model capacity, and the feature dimension is still relatively single; The direction prediction ability did not break through the white noise limit, and the ability to capture extreme markets was insufficient. Future research should focus on introducing an attention mechanism to enhance the capture of key time-series information and fusing multi-dimensional feature data to construct a richer feature system.

References

1. Kumbure, M.M., Lohrmann, C., Luukka, P., Porras, J.: Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications* 197, 116659 (2022).

2. Efendi, R., Arbaay, N., Deris, M.M.: A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. *Information Sciences* 441, 113-132 (2018).
3. Wen, X., Li, W.: Time series prediction based on LSTM-attention-LSTM model. *IEEE access* 11, 48322-48331 (2023).
4. Al-Selwi, S.M., Hassan, M.F., Abdulkadir, S.J., Muneer, A., Sumica, E.H., Alqushaibi, A., Ragab, M.G.: RNN-LSTM: From applications to modeling techniques and beyond—Systematic review. *Journal of King Saud University-Computer and Information Sciences* 36(5), 102068 (2024).
5. Li, Z., Yu, H., Xu, J., Liu, J., Mo, Y.: Stock market analysis and prediction using LSTM: A case study on technology stocks. *Innovations in Applied Engineering and Technology*, 1-6 (2023).
6. Bhandari, H.N., Rimal, B., Pokhrel, N.R., Rimal, R., Dahal, K.R., Khatri, R.K.: Predicting stock market index using LSTM. *Machine Learning with Applications* 9, 100320 (2022).
7. Muneer, A., Ali, R.F., Almaghthawi, A., Taib, S.M., Alghamdi, A., Ghaleb, E.A.A.: Short term residential load forecasting using long short-term memory recurrent neural network. *International Journal of Electrical & Computer Engineering* (2088-8708) 12(5) (2022).
8. Hodson, T.O.: Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, 1-10 (2022).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

