Think Twice Before Plugging Variables into Model

Xing You Li[1,2*]

[1]School of International Business, Zhejiang International Studies University, Hangzhou, Zhejiang China
[2]Department of Economics, Ghent University, Ghent, Belgium

*lininglixingyou@qq.com

Abstract. With the development of artificial intelligence, an increasing number of AI models are being applied in the financial sector. The Long Short-Term Memory (LSTM) model, as an AI model for processing time-series data, has achieved promising results in the investment field. Currently, many studies use LSTM models with inputs mainly consisting of variables such as prices, returns, and volatility, while some studies also include additional variables to improve prediction accuracy. However, these studies lack sufficient discourse on why these variables are chosen and what variables should be inputted. This is due to the lack of interpretability of the relationships between variables in AI models, resulting in a decreasing emphasis on the theoretical connection between input data and prediction results. In this study, we use LSTM models to predict stock returns, with both return and price-to-earnings ratio (P/E ratio) sequences as inputs. Based on the change in LSTM model prediction accuracy resulting from different input data, we suggest that providing more variables without selection may not necessarily lead to better prediction results. For the LSTM model, the momentum effect of the input variable sequence is related to its prediction accuracy, and grouping stocks according to P/E ratio indicators can improve the predictive performance of the LSTM model.

Keywords: LSTM, P/E ratio, momentum effect

1 Introduction

Stock price prediction is an attractive yet challenging field in quantitative finance and time series analysis, as evidenced by the studies of Hu, Zhao, and Khushi (2021) [1] and Hewamalage et al. (2021) [2]. Many factors, such as inflation, seasonality, economic policies, company performance, economic shocks, and political unrest, can affect stock prices, thereby reducing the accuracy of any prediction system. However, reliable stock price prediction can bring significant benefits to companies, shareholders, and investors, and can serve as a key indicator for guiding the formulation of economic policies. Various methods, including traditional time series analysis, machine learning, and deep learning, have been proposed in the stock price prediction field over the past few decades. To design an accurate stock price prediction system, some fundamental

© The Author(s) 2024
J. Liao et al. (eds.), *Proceedings of the 2024 2nd International Conference on Digital Economy and Management Science (CDEMS 2024)*, Advances in Economics, Business and Management Research 292,
https://doi.org/10.2991/978-94-6463-488-4_48

issues must be deeply considered, such as feature selection, model fitting, and prediction.

In the past, linear models such as the ARIMA model were commonly used to capture various features of time series. Although linear models can be effective for short-term forecasting, their regression-based approach may not be suitable for nonlinear problems and may be less effective for long-term forecasting. In Ballings et al.'s paper (2015) [3], ensemble methods such as random forest and AdaBoost were compared with other classifiers such as neural networks, logistic regression, support vector machines, and k-nearest neighbors. The conclusion showed that random forest had the highest accuracy for predicting stock price changes. Chen et al.'s (2018) [4] study proposed an RNN Boost model to predict the prices of the Chinese stock market by combining RNN and AdaBoost models, which achieved better accuracy than the baseline RNN model.

LSTM models are widely used in the financial field because they are good at handling time-series data. Siami-Namini et al. (2018) [5] compared the predictive performance of ARIMA and LSTM models for time-series data. In Mehtab et al.'s study (2021) [6], the LSTM regression model was used to predict stock price data from the Indian NIFTY 50 index, and the results showed that the LSTM model was more effective than traditional machine learning methods. Siami-Namini et al. (2019) [7] compared the effects of ARIMA, LSTM, and bidirectional LSTM (BiLSTM) models on predicting financial time series data, and found that the BiLSTM model achieved the best performance. Baek et al. [8] introduced a new framework called ModAugNet, which includes two LSTM modules: a preventive overfitting LSTM and a prediction LSTM. The authors found that the ModAugNet model was significantly better than the baseline model.

2 Research Design and Empirical Study

2.1 LSTM and P/E Ratio

Long Short-Term Memory (LSTM) models are a type of Recurrent Neural Network (RNN) that are commonly used for time series analysis and prediction. They aim to capture long-term dependencies and overcome the problem of vanishing gradients that is often encountered in traditional RNNs. LSTM models have several variants, including standard LSTM, Gated Recurrent Units (GRU), and LSTM with peephole connections. These models differ in their architecture and the way they process information flow and memory units.

LSTM is able to better handle the long-term dependency problem because it has two important components: memory cells and gate units. The input to the LSTM model includes the current input data and the previous state information (composed of the hidden state and memory cell from the previous LSTM output), while the output is the current prediction and the current state information (new hidden state and memory cell). The overall structure consists of multiple LSTM layers, each containing one memory cell and three gate units: input gate, forget gate, and output gate. The input gate determines which parts of the current input need to update the memory cell, the forget gate determines which information needs to be forgotten, and the output gate determines

which information needs to be output to the next layer. The parameters of these gate units are learned through training, allowing the LSTM model to adaptively learn the features and patterns of input data, and thus achieve time series data prediction.

The price-to-earnings ratio (PE) is a widely used indicator in stock market analysis. It is obtained by dividing a company's stock price by its earnings per share (EPS), and is commonly used as a valuation indicator for stocks. The PE ratio can also serve as a tool for predicting future stock returns. Typically, stocks with a lower PE ratio are considered undervalued, while those with a higher PE ratio are viewed as overvalued. As such, investors may use the PE ratio to identify potential undervalued or overvalued stocks and adjust their investment portfolios accordingly.

Numerous studies have found a close relationship between PE and stock returns. For example, Al-Mwalla and Al-Omari (2010) [9] suggested that there exists a long-term cointegrating relationship between PE and stock returns. However, there are differing views on the relationship between PE and stock returns. Weigand and Irons (2007) [10] proposed that when the overall stock market is in a high PE phase, stocks tend to exhibit higher returns. Therefore, the PE ratio's ability to predict future stock prices may also be affected by multiple factors, which researchers or investors need to take into account.

2.2 Research Design

This paper aims to address the following questions: Can using the price-to-earnings ratio (PE) as an additional input variable improve the accuracy of LSTM models trained on stock return sequences? What is the theoretical basis for the relationship between input variables and prediction results? Are there alternative ways to incorporate PE as a return factor into LSTM models for better performance? To achieve these objectives, the following procedure is designed:

Step 1: Two benchmarks are established to evaluate the accuracy of LSTM model predictions: (1) the average return rate of the CSI 380 Index and (2) a portfolio consisting of the top 25% stocks with the highest predicted return rates based on the previous day's stock returns.

Step 2: The effectiveness of using stock returns as input variables is verified by inputting stock return sequences into the LSTM model, selecting the top 25% of stocks based on the LSTM model's output, and calculating the portfolio's return rate.

Step 3: The predictive power of PE on stock returns is tested using the following three methods: (1) predicting stock returns based on PE, building a portfolio based on the prediction results, and calculating its return rate; (2) using PE sequences as the sole input variable for the LSTM model, building a portfolio based on the prediction results, and calculating its return rate; and (3) inputting both PE and return sequences into the LSTM model, building a portfolio based on the prediction results, and calculating its return rate.

Step 4: Based on the results of Step 3, the paper proposes using PE as a metric to group stock data, using only return sequences as the input variable for LSTM models in different groups, building portfolios based on the LSTM model's output, and comparing the difference in return rates.

Overall, this research seeks to investigate the impact of using PE as an additional input variable in LSTM models trained on stock return sequences and explore different ways to incorporate PE as a return factor for better performance.

2.3 Experimental Results and Analysis

The experiments in this study are based on the constituent stocks of the "CSI 380 Index" in the Chinese stock market. Stocks with severe data missing are removed, leaving a total of 172 stocks. The time span covers all trading days between January 1, 2020, and May 31, 2021. The price-to-earnings ratio (PE) is selected as the factor indicator for the rate of return. The model construction is based on the LSTM model-related functions provided by the Matlab toolbox. Figure 1 presents a schematic diagram of the LSTM model:

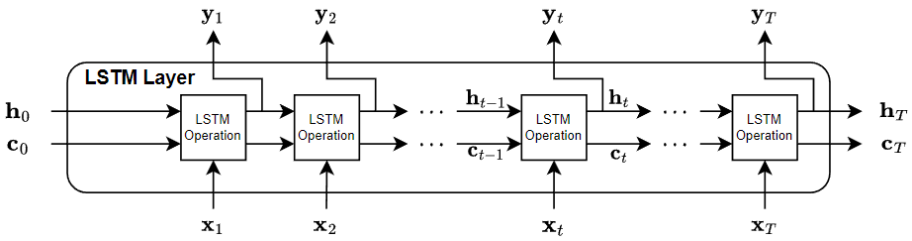

Fig. 1. LSTM model architecture.

All data is standardized to facilitate model convergence. After referring to previous experience and gradually tuning the parameters, the main model parameters in this study have been set as follows to achieve better results: due to the small number of input and output variables, in order to avoid overfitting, the number of hidden units (numHiddenUnits) is set to 30, and the total number of training epochs (MaxEpochs) is set to 250. The initial learning rate (InitialLearnRate) is set to 0.005, and the learning rate is allowed to change after a certain number of epochs to prevent it from falling into a local extreme value, so the drop factor (LearnRateDropFactor) is set to 0.2, and the remaining parameters are not specially adjusted. All the aforementioned parameters and model settings can be visually configured using MATLAB's built-in Neural Network Toolbox, as illustrated in Figure 2:

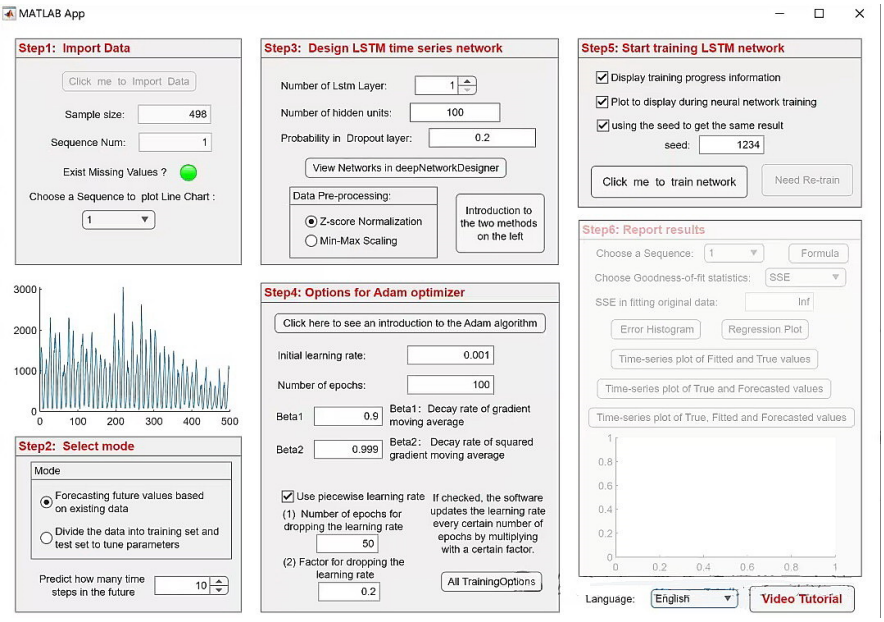

Fig. 2. MATLAB Neural Network Toolbox Configuration

In the calculation of investment portfolio returns, considering the difficulty for non-institutional investors to short sell stocks in the Chinese stock market, the rate of return calculation only considers the profit brought by the rise of stocks and does not involve the profit of short selling falling stocks. Five trading days are selected as the input sequence length for training data, and the LSTM model provides the predicted rate of return for the sixth day. The top 25% stocks with the highest predicted rate of return are selected to form an investment portfolio, and the stocks in the portfolio have the same weight. The data between January 1, 2020 and December 31, 2020 were used as training data, while the data between January 1, 2021 and May 31, 2021 were used as validation data.

According to Table 1 below, the average return of the "CSI 380 Index" from January 1, 2021 to May 31, 2021 was 4.8%, with a Sharpe ratio of 1.8 (annualized, same below). The investment portfolio constructed using the "previous day's return as the prediction result" method only achieved a return rate of 1.7%, with a Sharpe ratio of only 0.58. However, the investment portfolio established based on the LSTM model prediction results using the return rate sequence as input achieved a return rate of 9.3%, with a Sharpe ratio of 3.2. This result far exceeds the two benchmark indicators used as evaluation standards, indicating that the combination of the LSTM model and the return rate sequence can achieve good results.

Table 1. Comparison of Investment Portfolio Performance

Investment Portfolio	Return Rate	Sharpe Ratio
CSI 380 Index	4.8%	1.8

Previous Day's Return	1.7%	0.58
LSTM Model Prediction (stock returns)	9.3%	3.2

Before using PE as an input variable, its predictive power for stock returns was tested. A scoring method commonly used in quantitative investment was employed to select high-scoring stocks based on their PE ratio as the scoring indicator. Table 2 shows that the investment portfolio constructed based on PE achieved a return of 5.83% with a Sharpe ratio of 1.8, outperforming the average return of the Shanghai Stock Exchange 380 Index and achieving the same Sharpe ratio. This indicates that the PE ratio is effective in predicting future stock returns.

Table 2. Benchmark Yield and PE Investment Portfolio Yield

nvestment Portfolio	Return Rate	Sharpe Ratio
CSI 380 Index	4.8%	1.8
Portfolio based on PE (>75%)	5.83%	1.8

Many studies have included various indicators as inputs for LSTM models, such as the opening price, closing price, highest price, lowest price, etc., as used in the study by Ding and Qin (2019) [11]. Wu and Li (2021) [12] selected the same input variables for a CNN-LSTM model. Heiden and Parpinelli (2021) [13] went further and added indicators of investor sentiment to the LSTM model. However, it is unclear whether including these indicators will lead to better results. As shown in the table above, PE achieved a certain level of effectiveness in traditional methods. The next step is to investigate whether PE can maintain this effectiveness in LSTM models. The table 3 below shows the results of constructing an investment portfolio based only on the PE sequence as the input variable.

Table 3. Benchmark Yield and LSTM Model Based on PE Sequence

Investment Portfolio	Return Rate	Sharpe Ratio
CSI 380 Index	4.8%	1.8
LSTM Model Prediction (stock returns)	9.3%	3.2
LSTM Model Prediction (PE)	5.8%	1.8
Portfolio based on PE (>75%)	5.83%	1.8

The above results indicate that compared to the performance obtained using the PE scoring method, PE did not show stronger performance in LSTM. Therefore, can PE provide additional information when combined with the stock return series as input variables in the LSTM model to improve the originally excellent prediction results? The specific results are shown in the table 4 below.

Table 4. Benchmark Yield and LSTM Portfolio Yield

Investment Portfolio	Return Rate	Sharpe Ratio
CSI 380 Index	4.8%	1.8
Previous Day's Return	1.7%	0.58
LSTM Model Prediction (stock returns)	9.3%	3.2
LSTM Model Prediction (stock returns and PE)	7.5%	2.4

It can be seen that compared with the LSTM prediction results with only the stock return series as input variables, the output results of the LSTM model with both the PE sequence and the stock return series as input variables did not effectively improve the portfolio's return rate and Sharpe ratio, but rather showed a certain degree of decline, although it was still higher than the benchmark index. This result shows that simply putting PE into the model may not necessarily achieve better results, and the prediction accuracy has not improved further on the original basis. This paper believes that the reason for this may be that the LSTM model focuses on the analysis and prediction of time series data, while the PE data is calculated based on stock prices and company earnings data. The company's earnings data is quarterly data, which is only announced when listed companies release quarterly and annual reports. Therefore, from a time series perspective, the information content of the PE data may be relatively low and may not provide sufficient additional help to the return series.

Based on the results of the previous section, this paper proposes: first, there is a certain correlation between the momentum effect of variables and the prediction accuracy of the LSTM model. Second, since the calculation of the price-to-earnings ratio is based on the stock price, changes in the stock price will also be reflected in the volatility of its price-to-earnings ratio, so it is possible to consider grouping stocks based on PE. Then, the effect of the LSTM model may differ in each group. Next, we will do the following work: first, divide the training set and validation set according to the high and low PE values, and divide the original data into three groups: the lowest 25% of stocks by PE value, the highest 25% of stocks by PE value, and a mix of the two. Second, test the difference in momentum effects among the three groups. Finally, train the LSTM model with three sets of data and test their effects.

The Information Coefficient (IC) of the factor is used here to test the momentum effect, and the calculation method of the momentum effect: Selecting five trading days as the time window, the return rate is calculated based on the prices of the previous four days, and the fifth day is excluded to minimize the impact of other factors such as reversal effect. Then, the correlation coefficient between the return rates of the first four days and the sixth day is calculated. The specific results are shown in the table 5 below.

Table 5. Factor IC Value Test Results

Validation set data	stock returns	PE
≤25%PE	0.0453	0.0337
≥75%PE	0.0195	-0.0269
≤25% and ≥75%	0.0223	-0.0214

Based on the above table, it can be observed that on the one hand, the momentum effect of PE is significantly lower than that of the return sequence in all three grouped datasets, which may be the reason why the predictive accuracy of the LSTM model deteriorates after adding PE as an additional variable, consistent with the hypothesis proposed earlier. On the other hand, there is a difference in the momentum effect IC value of factors in different groups of the same variable, indicating that grouping stocks according to PE is feasible, and the stocks in the groups have certain differences in momentum characteristics.

So, how does the LSTM model perform in these groups? The returns of each investment portfolio are shown in the table 6 below:

Table 6. LSTM Portfolio Yield Based on Grouped Data and Benchmark Yield

Investment Portfolio	Return Rate	Sharpe Ratio
CSI 380 Index	4.8%	1.8
LSTM Model Prediction (stock returns)	9.3%	3.2
Average return rate of ≤25% PE stock groups	4.37%	1.829
Average return rate of ≥75% PE stock groups	3.58%	0.922
Average return rate of ≤25% and ≥75% PE stock groups	4.06%	1.45
LSTM portfolio returns for ≤25% PE stock groups	10.47%	3.066
LSTM portfolio returns for ≥75% PE stock groups	-1.71%	-
LSTM portfolio returns for ≤25% and ≥75% PE stock groups	-1.26%	-

Combined with the table above, it can be seen that first, stocks in the lower PE (≤25%) group have higher momentum effects, and the LSTM model in this group achieved better results, with the investment portfolio obtaining the highest return of 10.47%, and its Sharpe ratio (3.066) only slightly lower than that of the LSTM model in the entire 380 index stocks (3.2). Second, the momentum effect in the other two groups (≥75% PE and mixed groups) decreased, and the LSTM model's predictions did not perform ideally in these two groups, with the investment portfolio obtaining negative returns, far below the benchmark index.

3 Conclusions

The experimental results presented above demonstrate that: firstly, there is a certain correlation between the momentum effects of input variables and the prediction accuracy of the LSTM model. For stock data, input variables with significant momentum effects tend to produce better prediction results with the LSTM model. When there are multiple input variables, the differences in momentum effects between variables can also affect the model's prediction accuracy. Secondly, the comparison of different uses of PE reveals that in order to achieve good results with new technologies such as artificial intelligence, one cannot ignore traditional financial theory. Only by combining traditional financial theory with appropriate usage scenarios can new technologies and models achieve better results. Finally, as demonstrated by this study, although artificial intelligence technologies represented by LSTM models have achieved remarkable results in the financial field, there are still areas for improvement in the technical details of their application. Due to the black-box nature of artificial intelligence models in data processing, some researchers believe that the more variables the model is fed, the better the prediction accuracy will be, while ignoring the importance of causal relationships between input variables and output results, as well as the internal logic of the interactions between economic variables in the model. The experiments presented in this paper highlight the need to pay more attention to the relationship between new technologies and traditional knowledge, and that the invention and advancement of new technologies

is not to replace traditional knowledge and theory, but rather to build upon them in order to achieve even better performance.

References

1. Z. Hu, Y. Zhao and M. Khushi. (2010) A survey of Forex and stock price prediction using deep learning. J. Applied System Innovation, vol. 4, no. 9, pp. 1–30.
2. H. Hewamalage, C. Bergmeir and K. Bandara. (2021) Recurrent neural networks for time series forecasting: Current status and future directions.J. International Journal of Forecasting, vol. 37, no. 1, pp. 388–427.
3. M. Ballings, D. V. D. Poel, N. Hespeels and R. Gry. (2015) Evaluating multiple classifiers for stock price direction prediction.J. Expert Systems with Applications, vol. 42, no. 20, pp. 7046–7056.
4. W. Chen, C. K. Yeo, C. T. Lau and B. S. Lee. (2018) Leveraging social media news to predict stock index movement using RNN-boost.J. Data & Knowledge Engineering, vol. 118, pp. 14–24.
5. S. Siami-Namini, N. Tavakoli and A. S. Namin. (2018) A comparison of ARIMA and LSTM in forecasting time series. in 17th IEEE Int. Conf. on Machine Learning and Applications (ICMLA), Orlando. pp. 1394–1401.
6. S. Mehtab, J. Sen and A. Dutta. (2021) Stock price prediction using machine learning and LSTM-based deep learning models. Communications in Computer and Information Science, vol. 1366, pp. 88–106.
7. S. Siami-Namini, N. Tavakoli and A. S. Namin. (2019) A comparative analysis of forecasting financial time series using ARIMA, LSTM, and BiLSTM. arXiv preprint arXiv:1911.09512.
8. Y. Baek and H. Y. Kim. (2018) ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module. J. Expert Systems with Applications, vol. 113, pp. 457–480.
9. RA Weigand, R Irons. (2007) The Market P/E Ratio, Earnings Trends, and Stock Return Forecasts. J. The Journal of Portfolio Management, 33 (4) :87-101
10. [10] Shen P. (2000) The P/E ratio and stock market performance. J. Economic Review, 2000, 85(16):23-36.
11. Ding G, Qin L. (2020) Study on the prediction of stock price based on the associated network model of LSTM. J. International Journal of Machine Learning and Cybernetics, 11: 1307-1317.
12. Wu M T, Li Z, Herencsar N, et al. (2021) A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. J. Multimedia Systems :1-20.
13. Heiden Alexandre and Rafael Stubs Parpinelli. (2021) Applying LSTM for Stock Price Prediction with Sentiment Analysis. 15th Brazilian Congress of Computational Intelligence.

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.