Analysis of the Difference in Stock Price Between A-shares and American Stocks in Machine Learning

Jing Cao[1,†] and *Xuanze* Sun[2,†]

[1]School of accounting, Guangdong Baiyun University, 519000 Guangzhou, China
[2]School of international Education, Henan University of Animal Husbandry and Economy, 450000 Zhengzhou, China
[†]These authors contributed equally.

Abstract. Contemporarily, stock market is the most representative financial investment tool in the world. The application of machine learning has had a significant impact on the development of society and economy as well as productivity, and has also been inextricably linked to the securities market. This study will analyse and compare the technological development of machine learning in the last five years, as well as the stock value data and stock price fluctuations of A-shares and American stocks in the field of machine learning. In this way, the machine learning technology may change the global stock market in the future, and the prospect of this technology in the future. This paper introduces three forecasting models, namely Light Gradient Boosting Machine (lightGBM) model, Convolutional Neural Networks (CNN) model and Long short-term memory (LSTM) model, and studies their influence on stocks and forecasting accuracy. Applying machine learning to financial investment is a two-edged sword, with advantages and disadvantages, opportunities and challenges, depending on whether and the measure to implement it.

1 Introduction

Stocks are the most representative and common financial investment instrument internationally. With the rapid development of the world economy, the scale of the securities market is also constantly expanding. With the development of technology, artificial intelligence technology is constantly maturing. The application of artificial intelligence has had a significant impact on the development of society, economy and productivity. It has the potential to become a hot spot for future technological development. At the same time, it also has an inseparable connection with the securities market.

Yang et al. incorporate stock prediction into stock selection to specifically capture the future features of stock markets, thereby forming a novel hybrid (two-step) stock selection method (involving stock prediction and stock scoring) [1]. Li et al. represent articles of textual news by means of vectors of feeling through the analysis of feelings; put in place a deep layered learning model to learn sequential information in a series of market snapshots that is built by technical indicators and news feelings; put in place a fully connected neural network to make inventory forecasts [2]. Another group uses LSTM model in deep learning to learn and forecast the stock market valuation indicator, price-earnings ratio (P/E ratio) [3]. Liu et al. study research on stock price prediction method based on deep learning. The LSTM neural network model is applied to market price prediction, and the outcome of the prediction is compared to the RNN model [4]. The subject of Damrongsakmethee et al. is to develop a Deep Learning model to forecast the stock price market, by using the variant of Long Short-Term Memory (Deep LSTM) [5]. Leippold et al. analyze a comprehensive set of return prediction factors using various machine learning algorithms [6]. When writing computation-intensive tasks, such as image recognition programs based on deep learning He et. analyze a concise algorithm that uses a combination of multi-process and signal mechanisms, and uses queues to share data between processes, thereby making full use of CPU and GPU performance [7]. The primary purpose of Aldhyani et al. was to develop an intelligent framework with the capability of predicting the direction in which stock market prices will move based on financial time series as inputs [8].

With the continuous development of the world economy, artificial intelligence technology is constantly evolving and improving,gradually entering the public life. It has played a great role in promoting human development and economy. In general, it may become the mainstream in the near future, and artificial intelligence technology requires a large number of human resources and data optimization to improve, so as to have a great impact on the global financial and stock markets, with many ups and downs.

This paper will analyze and compare the technological development of machine learning in the last five years, as well as the stock price valuation data of A-shares and American stocks in the field of machine learning and the

* Corresponding author: Yuhao.Liu@calhoun.edu

ups and downs of stock prices. In this way, machine learning technology can predict the changes of the global stock market in the future,draw a conclusion and explain its limitations. The principles of machine learning are shown in Fig. 1.

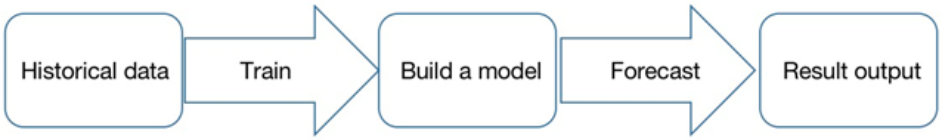

Fig. 1. Principles of machine learning.

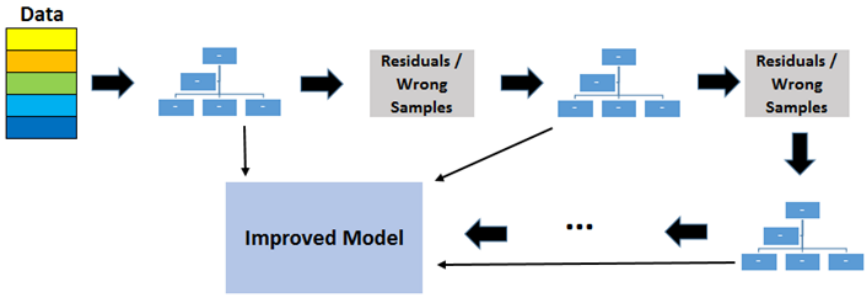

Fig. 2. Gradient boosted decision tree [9].

2 LightGBM

LightGBM is an open source machine learning library developed by Microsoft that uses gradient enhanced decision trees to build highly accurate predictive models, the principle of which is shown in Fig. 2. It is suitable for a variety of applications, including regression, classification, sorting, etc. LightGBM has significant advantages in efficiency, accuracy, and scalability, and has achieved state-of-the-art results in many machine learning benchmarks and Kaggle competitions, providing a powerful tool for investors in the U.S. stock market.

LightGBM and other models can be applied to the U.S. stock market and individual stocks. Empirical studies show that these models can build price prediction models for U.S. stock indexes and individual stocks, and achieve decent prediction accuracy and investment returns. For example, Sadia et al. compared several machine learning algorithms in predicting S&P 500 component stocks and found that LightGBM model could achieve an annualized return of up to 15.7% [10]. Shen et al. found that LightGBM model could achieve an accuracy of 77.6% in predicting the rise and fall of 7 U.S. stocks such as Tesla, Apple and Netflix 3 days later [11].

LightGBM also supports stock trend classification and can predict the "bull market", "bear market" or "consolidation" stages of U.S. stocks. Ranco et al. found that using 5 types to classify and predict the S&P 500 index, the LightGBM model achieved an accuracy of 76% [12]. This can help investors determine market timing and adopt appropriate trading strategies. In addition, LightGBM can comprehensively consider the impact of macroeconomic factors, policy changes, geopolitical events, etc. On the U.S. stock market, it can learn the interaction between different influencing factors and establish a more comprehensive and accurate predictive model [13]. This diversified predictive model is more suitable for long-term value investment and asset allocation.

LightGBM can also be used to predict the prices of stock categories or industries. By analyzing the data of many stocks, it can identify the common trends and drivers of a particular category or industry and build a more accurate predictive model. This can help investors make more prudent decisions in stock selection and asset allocation. In addition to historical stock price data, LightGBM can comprehensively analyze other influencing factors such as monetary policy, macroeconomic data, news events and build a more comprehensive and accurate stock prediction model. This is conducive to long-term value investing. By analyzing a large amount of historical stock index data and influencing factors, LightGBM can identify the cyclical changes and transmission mechanisms of the stock market, provide references for investors to judge the current market stage (bull market, bear market) and future trends [14].

In summary, LightGBM is an advanced and practical machine learning algorithm that can be widely applied in the U.S. stock investment field, especially in stock price prediction and market trend judgment. However,its limitations should also be recognized, and it needs to be combined with artificial intelligence and expertise, not completely dependent on it.

3 CNN

Previous section describes the impact of the LightGBM on US stocks, but it is not only applicable to US stocks, but also to A-shares. Song and Zhuo used LightGBM model in an article to compare various indicators of state-

owned and private enterprises [15]. The results shown in Table.1.

In addition to this,other models like the Convolutional Neural Network (CNN) are also applicable to the A-share market. CNN, a type of feedforward neural network, leverages convolutional operations to extract data features. It can be utilized to predict stock markets and can also be integrated with other machine learning techniques. Ma and Yan used a CNN model to predict A-shares in a study [16]. The researchers first chose the two most representative indices, the CSI 300 and the SSE Composite, as their research samples. During the research, they found the prediction accuracy for the CSI 300 was 70.67%, while for the SSE Composite, it was 68.11%. Moreover, when compared with the components of the CSI 300, the CNN model had a mean predictive accuracy of 69.89%. Furthermore, they studied the effect of feeling on the model's predictive performance. The results suggested that social media sentiment did not significantly impact the former, but did somewhat influence the latter. They also examined the effect of the pandemic on model performance. Prior to the pandemic outbreak, the CNN model had an average prediction accuracy of 70.11%, but this fell to 66.67% after the outbreak, a significant reduction. Moreover, social media sentiment significantly improved for both indices after the outbreak.

Table 1. Evaluation of the group model of state-owned enterprises and private enterprises.

Evaluation indicators	precision	recall	F1 Score
state-owned enterprise	0．906	0．707	0．795
private enterprise	0．832	0．782	0．806

According to the results, it is evident that the CNN model has a clear advantage in the A-share market. However, CNN also has its limitations. Firstly, the CNN might perform well on the training data but falter on the test data, a phenomenon known as overfitting. This may be due to the CNN learning specific patterns in the training data that are not applicable to the test data. Secondly, the CNN model might be affected by uncertainty or non-linear factors in the stock market, such as sudden events or policy changes. These factors might prevent the CNN from accurately capturing market characteristics. In addition, the CNN may overlook crucial information such as the correlation among stocks. In summary, CNN model have their strengths and weaknesses. The appropriate application of these models is paramount.

4 LSTM

The stock market has a long and short-term memory effect and non-linear features that cannot be expressed through the traditional linear pricing model. Therefore, we need to use LSTM model (neural network model for both long-term and short-term memory) to capture the nonlinear price structure between the five stock market factors (equity portfolio performance, market value, book/market ratio, income and Chinese equity investment), reflect the non-linear relationship between the factors, and improve the out-of-sample fit degree, the performance of lengthy and short strategies and prediction accuracy. Pan et al. found that LSTM model has many advantages compared with traditional model, which can adapt to the changing environment in the stock market and improve the forecasting ability and investment effect of the pricing model [17].

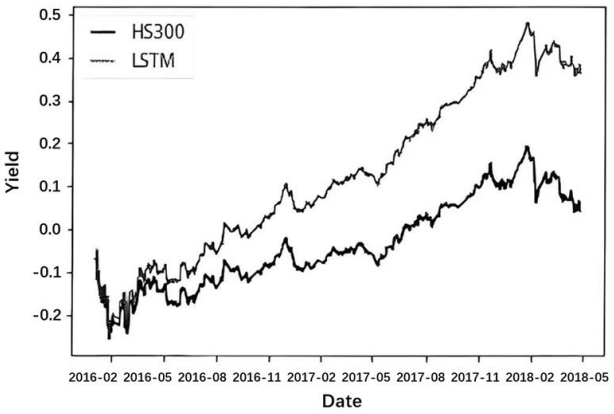

Fig. 3. Backtest result.

For the purpose of verifying the predictive capability of the LSTM neural network model, the author chooses to make reference to the CSI 300 Index. By constructing a strategy combination consistent with the weight of the constituent stocks of CSI 300 Index, the author carries out full position operation according to the corresponding weight at the beginning, and then through the historical performance data, LSTM is used to predict the rise and fall of each constituent stock of CSI 300 index, in weekly units. If it is predicted that the yield of the constituent stock will rise in the next week compared with that of the day, continue to hold or buy (if the weight of the subject is 0), if it is predicted that the yield of the constituent stock will fall in the next week compared with that of the day, keep watching or sell(if the target weight is not 0). Based on the above strategy logic, the backtest trading strategy is constructed, the initial adjustment frequency is daily, the daily forecast of the rise and fall trend of all constituent stocks in the next week is adjusted accordingly, and the simulated backtest data from January 2016 to May 2018 is conducted, as shown in Fig. 3 [18].

From the results of the backtest, the cumulative return of the constructed LSTM neural network model is higher than that of the CSI 300 index in the market in the long run. At the same time, the backtest results of LSTM strategy show that the accuracy of LSTM in predicting the upward trend is higher than that of predicting the downward trend, which is largely due to the fact that in order to facilitate the comparison with the benchmark return, the constructed strategy is held according to the weight of the constituent stocks of the benchmark index. Another reason is that most of the training data in the early stage of the LSTM model are dominated by the trend of

continuous rise and continuous fall, and there is a lack of abnormal plunge data samples.

According to the characteristics of LSTM neural network model in deep learning algorithm, this paper jumps out from the thought of traditional research methods, and enriches the content and ideas of the research on expected return. However, the application of LSTM algorithm to the research of expected returns still has some shortcomings. In the in-depth research of price or return prediction, although the predictive ability of the model has certain advantages, the maximum performance of the deep learning algorithm is still not brought into play. Hence, there is a lot of room for improvement in forecasting capabilities.

5 Limitations and prospects

Financial markets are constantly changing, and the distribution of data in the future and present are inconsistent, making it difficult for machine learning models to adapt to market dynamics and to evaluate model effectiveness and generalization capabilities. The data available for financial tasks is limited, especially for low-frequency or extreme events (e.g., financial crises, black swan events) which may only have a small amount of data or even no data samples for model learning. A small sample size makes the application of automated learning methods very challenging. Besides, the financial market is influenced by many complex factors, including political, social, technological, natural disasters, and other aspects. These factors are often difficult to express numerically and to convert into forms understandable by algorithms. Additionally, there are some factors that are hard to quantify, e.g., market conditions and investor confidence. These factors can influence market trends to some extent but are difficult to express numerically. Moreover, financial forecasting problems involve many different influencing factors. These factors can affect market price changes on various time scales and dimensions and usually interact with each other. To build an effective model, it requires the use of professional techniques and a large amount of data to capture these influencing factors and build accurate prediction models.

The future may see an increased reliance on online or incremental learning methods, allowing financial models to continuously update their parameters based on new data. This would facilitate real-time and dynamic financial investments that are not reliant on outdated or static data and strategies. On this basis, we would be better equipped to respond to market changes, seize opportunities, mitigate risks, and improve returns. In addition, transfer learning or meta-learning methods could be utilized to leverage data and knowledge from other domains to aid models in learning with limited samples. This would allow us to achieve cross-domain and cross-task financial investments, not restricted to a single or limited dataset or knowledge base. For instance, we could transfer knowledge from fields like neuroscience, psychology, and biology to financial markets, thereby enhancing the model's generalization capability and unearthing more potential value and innovative points. Natural language processing or computer vision offers promising solutions for leveraging unstructured data, like text or images, to extract useful information. This would enable multi-source and multi-dimensional financial investments, rather than relying solely on structured or numeric data. On this basis, we could better interpret the implications and significance of data, enriching the input dimension and output results of models. Given the complexity and diversity of influential factors, deep neural networks or reinforcement learning methods could be combined to capture intricate relationships and interaction effects using multi-layered network structures or dynamic strategy updates. This would enable more profound and intelligent financial investments, rather than merely simplistic or manually created rules or assumptions. In this case, one could better comprehend the essence and laws of the market, optimizing the model's decision-making process and outcomes.

6 Conclusion

To sum up, this study mainly analyzes the US stock and A-shares based on LightGBM, CNN and LSTM model, as well as investigates their impact and predictable accuracy on stocks. Applying machine learning to financial investment is a two-edged sword, it has advantages and disadvantages, as well as challenges. It depends on whether and how to apply it. To leverage the role of machine learning in the field of financial investment, it is necessary to combine professional knowledge and experience, while selecting data and algorithms for adjustment, building effective models and making appropriate strategies, and continuously updating and adjusting.Although small frequency financial events will make it very challenging for the model to learn data samples, it can be seen from the analysis of machine learning in stocks that this technology will have very good application and results in the future.

Author Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. F. Yang, Z. Chen, J. Li, L. Tang, Appl. Soft Comp., **80**, 820-831 (2019).
2. X. Li, P. Wu, W. Wang, Infor. Proc. & Mana., **57(5)**, 102212 (2020).
3. G. Li, M. Xiao, Y. Guo, In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) 551-554 (2019).
4. D. Liu, A. Chen, J. Wu, In 2020 2nd International Conference on Information Technology and Computer Application (ITCA) 69-72 (2020).

5. T. Damrongsakmethee, V. E. Neagoe, In 2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI) 1-6 (2020).
6. M. Leippold, Q. Wang, W. Zhou, J. of Fin. Eco., **145(2)**, 64-82 (2022).
7. F. He, X. Hu, S. Liu, T. Li, K. Zhu, X. Bao, C. Jiang, In 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) **4**, 775-779 (2021).
8. T. H. Aldhyani, A. Alzahrani, A, Elec., **11(19)**, 3149 (2022).
9. K. H. Sadia, A. Sharma, A. Paul, S. Padhi, S. Sanyal, Int. J. Eng. Adv. Technol, **8(4)**, 25-31 (2019).
10. J. Shen, M. O. Shafiq, J. of big Data, **7(1)**, 1-33 (2020).
11. G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, I. Mozetič, PloS one, **10(9)**, e0138441 (2015).
12. M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of machine learning* (MIT press 2018).
13. A. Muneer, S. M. Fati, Fut. Inter., **12(11)**, 187 (2020).
14. C. Ma, S. Yan, Fin. Res. Lett., **49**, 103025 (2022).
15. M. Song, J. Zhu, Shanghai Mana, **42(3)**, 35-39 (2020).
16. C. Ma, S. Yan, Fin. Res. Lett., **49**, 103025 (2022).
17. S. Pan, S. Long, Y. Wang, Y. Xie, Inter. Rev. of Final. Ana., **87**, 102627 (2023).
18. J. Ji, Quantitative investment strategy based on LSTM (Master thesis, Huazhong University of Science and Technology 2020).