Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Stock price prediction model based on machine learning

Jiajun Sun

Department of Artificial Intelligence and Software, Software Engineering Institute of Guangzhou, Guangzhou, Guangdong, China

chana@ldy.edu.rs

Abstract. This paper reviews the problem of stock price prediction and discusses the current application status of traditional statistical models, machine learning and deep learning methods in financial data processing. By analyzing the limitations of traditional methods such as ARIMA and regression models in capturing market nonlinear characteristics and noise processing, it is pointed out that these methods have deviations in actual predictions; while models such as support vector regression, random forest, and LSTM can automatically extract features using historical data to improve prediction accuracy. The article introduces in detail the improvement ideas based on Bayesian optimization and hybrid models, and discusses how to use attention mechanisms, multimodal data fusion, and real-time dynamic update mechanisms to further improve the stock price prediction system. Finally, the shortcomings of existing methods in terms of data quality, model interpretability, and cross-market adaptability are analyzed, and the future development direction of combining cutting-edge technologies such as reinforcement learning is prospected to provide theoretical support and practical guidance for financial decision-making and risk management.

Keywords: Stock prediction; Machine learning; Deep learning; Hybrid model.

1. Introduction

The stock market is an important part of the financial system. Its operation not only affects the development of the macroeconomy, but also has a profound impact on individual investors and corporate decision-making. With the deepening of global economic integration and the promotion of technological progress, the volatility and complexity of the stock market are increasing, making stock market prediction a research issue that has attracted much attention.

The price trend of the stock market is affected by many factors, including macroeconomic indicators, corporate financial conditions, market sentiment, policy changes, etc. How to accurately predict stock prices is of great significance to investors' asset allocation, corporate financing decisions, and financial risk management. Taking the Ghana stock market as an example, Queku's research shows that there is a close connection between the performance of the Ghana stock market and the macroeconomic situation, which further highlights the importance of accurately predicting stock prices for grasping macroeconomic trends and maintaining economic stability [1]. Therefore, researching and developing effective stock prediction methods has always been the focus of academic and industrial circles.

At present, the methods of stock prediction mainly include traditional statistical methods, intelligent prediction methods based on machine learning, and hybrid methods combining multiple data sources. Among them, time series analysis, regression models, neural networks, deep learning and other technologies have shown certain prediction capabilities in different scenarios. However, due to the high nonlinearity of the stock market, the interference of noisy data, and the impact of black swan events, existing methods still face many challenges, and the accuracy and stability of stock predictions still need to be further improved.

Based on the above background, this paper aims to review the research related to stock price prediction from the perspectives of prediction methods, existing limitations, and future prospects. By combining the prediction methods of machine learning and deep learning, the characteristics and effects of each method are analyzed in detail; at the same time, the shortcomings of current research

at the technical and theoretical levels are deeply analyzed, and possible future development directions are prospected. It is hoped that this review can provide a comprehensive reference for subsequent stock prediction research and further promote innovation and practice in this field.

2. Prediction Approaches

2.1. Machine learning Approaches

Machine learning methods are a type of technology widely used in forecasting tasks. They use historical data to build models and then make predictions in new data. This method often has low computational complexity and fast training speed. Because it can automatically learn patterns in data, machine learning methods can make predictions without explicit programming rules, which is suitable for stock forecasting problems.

Dong et al. used the closing stock prices of Ford Motor Company over the past 50 years as the research object. They first used Augmented Dickey-Fulle (ADF) to analyze the stationarity of the original sequence. The test results showed that the sequence rejected the non-stationary null hypothesis at the 1% significance level, indicating that the original data met the stationarity requirements and did not need to be processed by difference. The study constructed an ARIMA model and determined the optimal parameter combination to be ARIMA (3,0,2) based on the minimum information criterion. The model validation results showed that the root mean square error (RMSE) of the forecast period was $4.12 and the mean absolute percentage error (MAPE) was 3.27%, indicating that the model has high forecasting accuracy [2]. The study [3] used Amazon stock as the object and predicted the recent daily return of the stock by using a simple linear regression model. The study showed that the simple linear regression model can also provide investors with information about stock price trends and returns, which helps investors make decisions to buy and sell stocks. Sricharan and Joshi used the support vector regression (SVR) method to predict the Indian stock market. SVR is a regression model based on a support vector machine (SVM) that controls the sparsity and parameters of the decision process by selecting different kernel functions. In order to improve the prediction effect, they used a powerful feature extractor to extract trend parameters in financial data and accelerated model training in a GPU environment. The results show that SVR significantly improves the prediction performance in processing financial data and can build an accurate stock market prediction model. In addition, Liu et al. compared Apple's stock price using decision tree regression, linear regression, random forest regression and support vector regression (SVR), and used mean square error for evaluation, and found that the prediction effect of SVR was significantly better than other models [4].

Stock price prediction has always been an important research topic in the financial field. Due to the nonlinearity and complexity of stock data, a single model often cannot achieve ideal prediction results. In recent years, researchers have tried to combine the random forest (RF) model with other methods to improve the accuracy of stock prediction. Zhang used the random forest method based on Bayesian optimization for stock prediction. Bayesian optimization prevents the model from falling into the dilemma of the local optimal solution. To demonstrate the optimization effect, they compared the support vector machine, the original random forest and the LGBM model, and performed Bayesian optimization on the latter two, achieving better results. Experiments show that the optimized model performs well in stock return prediction, with RMSE and MAE reduced by 0.01, MAPE reduced by 0.4, and R2 value increased by 0.02. The random forest model based on Bayesian optimization has the lowest prediction error on different data sets and has wide applicability [5].

2.2. Deep learning Approaches

Deep learning methods automatically learn features and make predictions by simulating the structure of human brain neural networks. Deep learning can extract complex high-level features from raw data without explicit feature engineering.

Stock market data is highly volatile, and traditional prediction methods often have difficulty accurately capturing this volatility. Traditional statistical models (such as exponential smoothing) can smooth data fluctuations to a certain extent by weighted averaging historical data. The hybrid system combining exponential smoothing and long short-term memory network (LSTM) proposed in [6] can effectively handle highly volatile data by dynamically decomposing key components of time series such as trend and seasonality. "At the same time, the hybrid system combining exponential smoothing and long short-term memory cells (LSTM) proposed in this paper can predict time series outputs through weighted averaging. ETS helps to dynamically decompose the key components of each individual time series, enabling the model to learn how to represent them and thus better handle high volatility data [6].

Long short-term memory network (LSTM) plays a very important role in the field of stock prediction. Since stock market data is usually highly nonlinear, noisy and has complex time series characteristics, traditional time series models often find it difficult to capture long-term dependencies. LSTM can effectively retain important historical information and alleviate the gradient vanishing problem through its unique memory unit and gating mechanism (including forget gate, input gate and output gate), thereby achieving better results in predicting future stock price trends. Qi et al. used the LSTM method to predict stocks during COVID-19. They used LSTM to predict the price of a single stock and combined it with historical indexes as empirical data. In addition, they collected a large number of tweets and conducted quantitative analysis of the tweets through deep learning methods. In order to optimize LSTM In order to improve the prediction effect, they introduced an attention mechanism to extract key factors and optimize the weights between historical indexes and public sentiment. Experimental results show that the predicted price curve of the model almost coincides with the real price curve, and the performance is relatively good [7].

The rapid development of machine learning and deep learning technologies (such as SVR, random forests, and LSTM) in recent years has provided new possibilities for overcoming the shortcomings of traditional methods. Huang et al. pointed out in their review that traditional statistical models (such as ARIMA) have significant deviations in nonlinear financial time series prediction, while hybrid models can reduce prediction errors by 15% to 20% by combining statistical methods with deep learning technologies [8].

3. Existing Limitations and Future Prospect

First, from the perspective of data quality, the noise problem inherent in high-frequency trading data is still very significant. For example, the tweet sentiment data mentioned in the literature [6] is often interfered with by noise, which to some extent weakens the explanatory power of sentiment indicators on stock price fluctuations. In addition, there is a significant time lag effect between macroeconomic indicators and stock prices, and their quantitative analysis faces great challenges, such as the 6 to 8 month time lag problem observed in the Nepal case, which makes it difficult to directly convert the predictive effectiveness of economic indicators into instant indicators for stock price prediction.

The small sample problem is particularly prominent in emerging markets (such as Southeast Asia and Africa). For example, only 12% of the companies in the Indonesian G20 Index have a continuous trading record of more than 5 years, and traditional models are prone to overfitting in small sample scenarios [9].

Second, in terms of model construction, traditional statistical models (such as ARIMA) make it difficult to capture the nonlinear relationships that are prevalent in the market due to their inherent linear assumptions. Related research (see literature [8]) points out that their prediction errors may be as high as 15-20%. At the same time, although deep learning models have shown great potential in processing complex data relationships, they have also been proven to be insufficiently adaptable when facing extreme events (such as the COVID-19 epidemic impact) (for example, reference [10]). In addition, most existing models rely mainly on historical data and lack a real-time update mechanism

for dynamic changes in data, such as the quarterly update frequency used in the Indonesian market case, which limits the application effect of the model in real-time prediction.

From a theoretical perspective, the differences in market mechanisms between different countries and regions make the model less universal across markets. For example, there are significant differences in factor composition between the Ghanaian and Indian markets, which directly leads to the difficulty of seamless model migration. On the other hand, due to its "black box" characteristics, the internal decision-making process of the machine learning model is difficult to fully explain, thereby limiting its application in financial theory verification. The economic implications of Bayesian optimization in reference [5] have not been fully revealed, further exacerbating this problem.

From a financial theory perspective, existing models have not effectively reconciled the contradiction between the efficient market hypothesis (EMH) and behavioral finance. For example, the SVR model's ability to fit technical indicators nonlinearly can test whether there are predictable short-term trends in the market, thereby challenging the strong validity assumption of the EMH [11].

To address the above limitations, future research can be carried out from the aspects of hybrid model optimization, multimodal data fusion, interpretability enhancement, and real-time dynamic system construction, that is, developing a three-layer hybrid architecture that integrates statistical models, deep learning, and reinforcement learning (for example, reference [8]). At the same time, attention mechanisms can be introduced to process multi-source data (for example, reference [6]), thereby improving the model's prediction accuracy. In addition, integrating satellite images, news texts, and social media data to construct a cross-domain knowledge graph to reveal the potential relationship between factors not only helps to improve data interpretation capabilities, but also provides support for mining new predictive factors. Furthermore, the interpretability of the model can be enhanced by combining the SHAP value or LIME method to conduct in-depth analysis of the model decision logic (refer to the analysis of the importance of SVR model features in reference [9]) and developing a visual interface to display the prediction process. Finally, a real-time prediction platform is built based on a stream data processing framework such as Flink or Kafka, and an adaptive learning rate mechanism is designed to cope with sudden changes in market structure, thereby ensuring that the model maintains efficient operation in a dynamic market environment.

4. Conclusions

This paper reviews the current status of the application of machine learning and deep learning in stock price prediction, and systematically analyzes the advantages and disadvantages of various methods and their performance on different data sets. The study found that although traditional models have advantages in stability and theoretical explanation, their prediction accuracy is often unsatisfactory when faced with highly nonlinear and noisy problems in market data. On the contrary, prediction models based on machine learning and deep learning, especially hybrid models, show higher potential in capturing complex data characteristics and reducing prediction errors. However, current research still has problems such as data quality, model interpretability, and cross-market adaptability. In the future, combining multimodal data, real-time dynamic update mechanisms, and interpretability enhancement techniques will be an important direction to improve the level of stock price prediction. Overall, this review not only provides a comprehensive theoretical basis for the field of financial forecasting, but also provides a useful reference for investment decision-making and risk management practice.

References

[1] Queku, I. C., & Carsamer, E. (2016). Stock market and macroeconomic performance in Ghana: New evidence. International Journal of Social Science and Humanities Research, 4(4), 436–447.

[2] Can Dong. (2022). Stock Trend Forecasting Using the ARIMA Model. Highlights in Science, Engineering and Technology.

[3] Xiaoyu Ma. (2023). Analysis of Amazon Stock Using Simple Linear Regression and Time Series ARIMA Model. Highlights in Science, Engineering and Technology.

[4] Vuong, P. H., Phu, L. H., Van Nguyen, T. H., Duy, L. N., Bao, P. T., & Trinh, T. D. (2024). A bibliometric literature review of stock price forecasting: from statistical model to deep learning approach. Science Progress, 107(1), 00368504241236557.

[5] Zhang, Y., Zheng, X., Yang, S., Meng, S., Yang, Z. Y., & Fei, X. (2024, May). A Random Forest Stock Prediction Model Based on Bayesian Optimization. In 2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 42-46). IEEE.

[6] Koushik, C., Pranav, M. V., Arjun, R. K., & Shridevi, S. (2022). Hybrid Exponential Smoothing-LSTM-Based Univariate Stock Market Prediction for Financial Sectors in NIFTY50. In Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022 (pp. 357-368). Singapore: Springer Nature Singapore.

[7] Qi, Y., Yu, W., & Deng, Y. (2021, April). Stock prediction under COVID-19 based on LSTM. In 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) (pp. 93-98). IEEE.

[8] Huang, K. (2024, October). Stock Price Prediction Based on Machine Learning. In Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (Vol. 115, p. 87). Springer Nature.

[9] Aït-Sahalia, Y., & Jacod, J. (2014). High-frequency financial econometrics. In High-Frequency Financial Econometrics. Princeton University Press.

[10] Sricharan, S., & Joshi, V. (2022, July). Comparison of SVR Techniques for Stock Market Predictions. In 2022 IEEE International Conference on Data Science and Information System (ICDSIS) (pp. 1-6). IEEE.

[11] Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. Handbook of the Economics of Finance, 1, 1053-1128.