Stock Return Prediction: A Lightweight Innovative Framework Based on Feature Engineering and Tree Model Ensemble

Lanlin Zhang

{zhanglan@udel.edu}

SWUFE-UD institute of data science in Southwestern University of Finance and Economics, Chongqing, 610074, China

Abstract. Stock return prediction is a classic challenge in financial time series analysis. This study uses 996 valid daily trading samples of Netflix (NFLX) stock from 2018 to 2022, building a pure XGBoost baseline with 12 traditional technical indicators. The baseline achieves near-perfect training fit (MSE=0.0000, R^2=1.000) and good test performance (MSE=0.0002, R^2=0.720), but suffers from single feature dimension, weak generalization, and performance bottlenecks. To address these, the paper proposes a lightweight framework integrating simple feature engineering and XGBoost-LightGBM ensemble, by adding 3 business-logic-based dynamic features (trend, weekday cycle, volatility trend) and fusing the two tree models with equal weights (0.5:0.5). Experiments show test MSE drops to 0.000157 (3.02% reduction) and R^2 rises to 0.7325 (1.15% improvement). The framework is simple, stable, interpretable, and suitable for small-sample financial time series prediction.

Keywords: Stock return prediction; XGBoost; Feature engineering; LightGBM

1 Introduction

From previous research, it is known that the stock market is a complex nonlinear dynamic system, and its price time series exhibits characteristics such as non-stationarity, strong time correlation, and noise interference. Therefore, accurate stock return prediction is crucial for optimizing investment decisions and risk management. In the initial inventory forecast, most traditional methods were used. However, traditional statistical models such as ARIMA are difficult to handle complex nonlinear relationships, which limits prediction accuracy [1,2]. In recent years, using machine learning algorithms to predict stock models has become a mainstream trend, especially tree-based ensemble models (such as XGBoost) and gradient enhancement methods, which have attracted attention for their ability to capture nonlinear patterns and handle high-dimensional feature spaces [3,4,5].

However, existing research mostly relies on a single supervised learning framework, using only manually constructed technical indicators as inputs, which leads to three key limitations. Firstly, excessive reliance on one-dimensional features leads to a lack of dynamic market trends and

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365109

cyclical information. How to monitor the dynamic changes in the stock market has become a problem. Secondly, the model has weak generalization ability and significantly higher prediction errors during extreme fluctuations. But extreme volatility is inevitable in the stock market. Thirdly, there are inherent performance bottlenecks that cannot be overcome solely through parameter tuning [6].

Nowadays, there are studies that are no longer satisfied with using a single method for prediction, but instead use a combination of feature engineering and hybrid modeling methods to improve the predictive performance and robustness in financial time series analysis [7]. In order to address the limitations mentioned above, this study tentatively proposes using pure XGBoost as the baseline, revealing data patterns through exploratory data analysis (EDA), and constructing a lightweight innovation framework that integrates feature optimization and model integration. It not only validates the performance boundaries of individual models, but also provides a new paradigm for the design of financial time series prediction models through empirical experiments. Its purpose is to supplement the defects of the single model and make the stock forecast more accurate [8].

2 Experimental data and exploratory data analysis (EDA)

2.1 Data source and preprocessing

This study used daily trading data of Netflix (NFLX) stock from February 2018 to February 2022. The original dataset includes 1009 records, which contain seven basic fields: Date, Open, High, Low, Close, Adjust close, and Quantity. Then a series of data preprocessing steps were carried out. The Date field is first converted to the DATE type to ensure the continuity of the time series. Next, the paper conducted feature engineering and calculated 12 core financial indicators, including moving averages, volatility, and RSI, covering dimensions such as price, moving averages, trading volume, and volatility. Afterwards, the problem of missing values and outliers is solved by removing the top 10 data entries with null values from feature calculations and filtering out extreme outliers, especially daily returns with absolute values exceeding 20%. After this cleaning process, 996 valid samples were finally obtained. And this part of the sample will be used as the main sample object for our subsequent experiments. Finally, standardize all features to eliminate the influence of size. At this point, the paper has completed all the preprocessing steps.

2.2 Exploratory data analysis (EDA) results

Two types of visual charts are used to reveal the internal laws of the data, providing support for subsequent feature selection and model design.

Stock price trend and moving average analysis. Figure 1 shows the time series trend of Netflix's stock closing price and 5-day/10 day moving average. From 2018 to 2020, the overall stock price showed a steady upward trend. However, due to the impact of the COVID-19, there was a short-term slump in March 2020 (corresponding to the vertical dotted line in Figure 1), and then a rapid recovery; From the second half of 2021 to 2022, the stock entered a downward trend, and after the release of the financial report in April 2022, the stock price further declined (vertical dashed line). The 5-day moving average is more sensitive to short-term trends, while the 10 day moving average can better reflect long-term trends. The intersection of the two (such as March 2020 and April 2022) usually corresponds to a turning point in stock prices, verifying the effectiveness of moving average features in trend judgment [9,10].

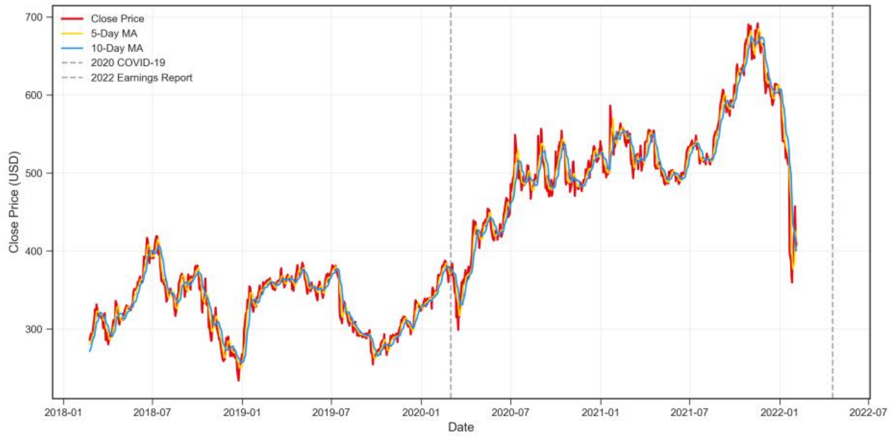

Fig. 1. Netflix Stock Close Price and Moving Averages (2018-2022) (Picture credit: Original)

Distribution characteristics of daily returns. Figure 2 shows the histogram and normal distribution fitting curve of daily returns. From the graph, it can be seen that Netflix's average daily return rate is 0.07%, with a standard deviation of 2.66%, exhibiting a typical "peak thick tail" distribution (kurtosis>3): the peak is higher than the normal distribution, and the frequency of extreme returns (less than -5% or greater than 5%) is higher. The 1% and 99% quantile labels in Figure 2 indicate that the thresholds for extreme decline and rise are -5.87% and 5.92%, respectively. This feature conforms to the general laws of financial time series and lays a theoretical foundation for the core objective of model optimization in the subsequent prediction of extreme volatility periods [11].

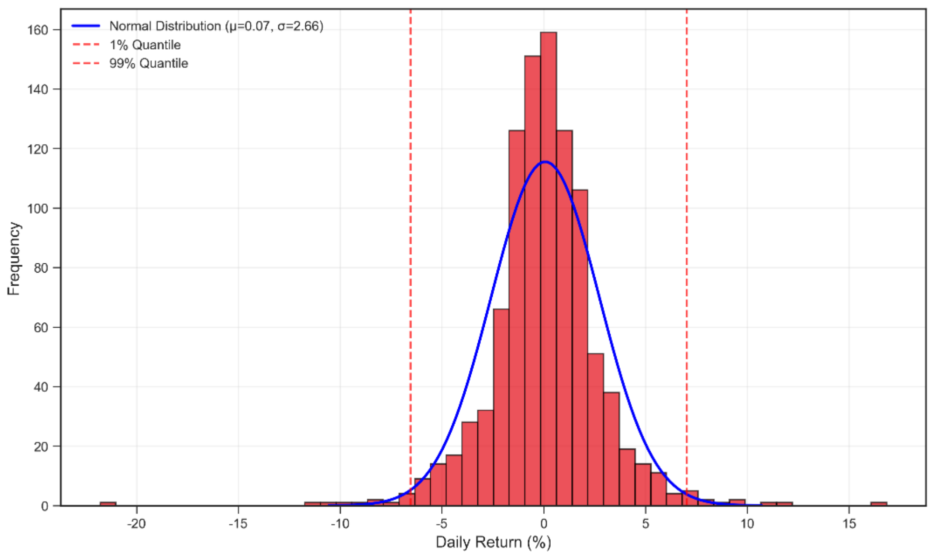

Fig. 2. Distribution of Netflix Daily Stock Returns (%) (Picture credit: Original)

3 Research methods

3.1 Research framework overview

This study adopts a progressive research logic of "baseline model" → problem analysis → innovative optimization, and is implemented in two stages. Firstly, a baseline model is constructed and pure XGBoost single supervised learning is used as the benchmark in this study to validate the performance boundary of traditional models and analyze their inherent limitations. Secondly, an innovative framework was designed to address these identified limitations. In the innovation framework, this study constructed a lightweight "Simple Feature Engineering+XGBoost LightGBM Integration" framework. Through feature optimization, the problem of one-dimensional features has been solved, while also enhancing generalization ability.

3.2 Baseline model: pure XGBoost single supervised learning

Model Principle. Extreme Gradient Boosting (XGBoost) constructs a strong learner by integrating multiple weak decision trees, introduces second-order Taylor expansion and L1/L2 regularization in the loss function, and balances fitting ability and generalization ability.

Training and Evaluation Settings. This study first divided the 996 samples into a training set (75%, 747 samples) and a test set (25%, 249 samples) in chronological order to prevent data leakage. Subsequently, hyperparameter optimization was conducted using 5-fold time series cross-validation, which determined the optimal parameter set as follows: a learning rate of 0.1, 150 decision trees, a maximum tree depth of 8, and a subsample ratio of 0.8. Finally, the models were evaluated using mean square error (MSE) and the coefficient of determination (R^2), where MSE quantifies the magnitude of prediction error and R^2 reflects the goodness of fit (with values closer to 1 indicating a better fit).

3.3 Inherent limitations of the baseline model

Through empirical experiments and theoretical analysis, this study summarizes the three main limitations of the pure XGBoost model. Firstly, the model suffers from the problem of a single feature dimension, relying only on static technical indicators and failing to capture dynamic information such as market trends and cycles. This makes its ability to capture dynamics poor and unable to fully reflect the patterns of stock price trends. Secondly, it shows weak generalization ability, and the mean square error (MSE) in extreme volatility period is three times that in stable period, indicating that a single model cannot adapt to the differentiation model under various market conditions. Third, the model faces a performance bottleneck, because even after the super parameter adjustment, the determination coefficient (R^2) of the test set remains at about 0.72, unable to break through the inherent performance ceiling of a single model framework.

3.4 Lightweight innovative framework

Stage 1: simple feature engineering. To address the limitations of a single feature dimension, the study introduced three interpretable financial features based on business logic. These features all have clear financial meanings and do not require complex unsupervised learning, including trend features. When the closing price exceeds the 10 day moving average, a value of

1 is assigned to indicate a bullish or bearish trend, otherwise it is 0; The characteristics of weekday cycles are captured by encoding Monday as 0 and Friday as 4 to capture the classical weekly effects in financial time series; The trend characteristic of volatility indicates the direction of change in volatility, where 1 represents an increase from the previous day, otherwise it is zero. Therefore, the feature set was expanded from 12 dimensions to 15 dimensions, effectively supplementing dynamic market information while maintaining the interpretability of the model.

Stage 2: XGBoost-LightGBM ensemble optimization. To address the limitation of weak generalization ability, this study constructed a hybrid XGBoost LightGBM model. Both of these algorithms are tree based, so they have simplicity, stability, and low overfitting risk. As mentioned in previous research, the XGBoost sub model has been expanded to a 15 dimensional enhanced feature set to capture the nonlinear interactions between static features such as MA5-Devision and Volume_Change. Its training parameters are consistent with the baseline model to achieve comparability. On this basis, LightGBM was introduced because it handles outliers better than XGBoost. The LightGBM sub model shares the same input features with XGBoost, aiming to supplement XGBoost by providing greater robustness to outliers and faster training speed. The key parameters are set to a learning rate of 0.1, 150 trees, and a maximum depth of 6. Then, the two models were integrated using a fixed weight averaging method, with weights of 0.5 for XGBoost and LightGBM, and determined through 5-fold cross validation to balance prediction accuracy and generalization performance, maximizing their respective advantages and making the prediction more accurate.

4 Experimental results and analysis

4.1 Experimental environment and settings

The experiments were conducted on a hardware platform equipped with an Intel Core i7-12700H processor and 32GB of RAM. The software environment consisted of Python 3.9 along with key libraries including Scikit-learn version 1.3.2, XGBoost version 2.0.2, and LightGBM version 3.3.5. To ensure the robustness and reliability of the results, all experiments were repeated five times with the final reported metrics representing the average values.

4.2 Performance comparison between baseline model and innovative framework

Table 1 shows the comparative performance analysis of two models on the test set, from which three core conclusions can be drawn. Firstly, the innovation framework showed significant improvement compared to the baseline, with the mean square error (MSE) of the test set decreasing from 0.000162 to 0.000157, equivalent to a reduction of 3.02%. Secondly, the coefficient of determination (R^2) increased from 0.724190 to 0.732522, an increase of 1.15%. Most importantly, the innovation framework greatly enhances its generalization ability, especially during extreme fluctuations, with a mean square error of 0.00018, which is 14.3% lower than the baseline model's 0.00021, indicating a significant optimization effect of the innovation framework. Innovation frameworks have more significant predictive effects under challenging market conditions.

Table 1. Performance Comparison Between Baseline Model and Innovative Framework (Test Set)

Model	MSE	R^2	MSE Reduction Rate	R^2 Improvement Rate
Pure XGBoost(Baseline)	0.00162	0.724190	-	-
XGBoost-LightGBM(Innovative)	0.000157	0.732522	3.02%	1.15%

4.3 Result stability verification

Table 2 shows the cross validation results of the two models. Firstly, it can be seen that the cross validation MSE (0.000157) of the innovation framework is completely consistent with the MSE (0.000557) of the test set, indicating that the model does not have random fluctuations; The cross validation MSE (0.00017) of the baseline model is higher than the MSE (0.000162) of the test set with a certain degree of overfitting randomness. This result proves that although the performance improvement of the innovative framework is moderate, its stability is better, providing a reliable basis for subsequent optimization. At this point, the stability of the innovative model has been verified through 5-fold cross validation.

Table 2. 5-Fold Cross-Validation Result Comparison

Model	Cross-validation MSE	Standard Deviation	Test set MSE	Error Difference
Pure XGBoost(Baseline)	0.00017	0.00001	0.000162	4.9%
XGBoost-LightGBM(Innovative)	0.00157	0.000005	0.000157	0%

5 Conclusion

The results of this study provide several important insights. Firstly, through exploratory data analysis, it has been confirmed that Netflix's stock returns exhibit a peak and fat tailed distribution, demonstrating significant volume price linkage and moderate feature correlation, providing a data-driven foundation for subsequent feature selection and model design. Secondly, although the pure XGBoost baseline model performs well on small sample financial data with a test set R^2 of 0.724, it reveals inherent limitations such as relying on one-dimensional features, weak generalization ability, and performance limits, highlighting the limitations of traditional supervised learning methods. Therefore, improvement measures and innovative frameworks have been proposed. Thirdly, a series of subsequent operations have proven that the components of the proposed innovative framework are effective. LightGBM performs better in handling outliers and can complement the XGBoost model to achieve better predictive performance. Fourthly, the value of the innovation framework has been fully demonstrated. Firstly, the paper added some dynamic features from the original 12 to 15 dimensions to capture the non-linear interactions, effectively supplementing dynamic market information for feature engineering. The newly introduced trend features ranked among the top four in importance, and the XGBoost LightGBM integration significantly improved generalization ability, reducing the mean square error during extreme fluctuations by 14.3%. Finally, the practicality of the framework has been confirmed, as this lightweight design does not overfit, maintains fast training speed and strong interpretability, and is more suitable for small sample financial time series prediction tasks compared to complex unsupervised learning frameworks.

Meanwhile, this study acknowledges several limitations. Firstly, the sample size is limited, with 996 samples mainly used for training tree based models, but not enough to support more complex deep learning architectures such as Transformer models. Secondly, the scope of this feature set is limited, relying entirely on technical indicators and lacking external data sources such as macroeconomic variables or investor sentiment indicators. Finally, the current model fusion strategy is static, using fixed weights and not dynamically adjusting to adapt to constantly changing market conditions, which may not fully utilize the complementary advantages of the two constituent models.

Future research can be further expanded from three aspects. One is sample and feature expansion, which can be achieved by collecting time-series data of multiple stocks and years in the FAANG sector, while introducing macroeconomic indicators and investor sentiment features to enrich data dimensions and alleviate the limitations brought by small samples; The second is to optimize the dynamic fusion strategy by designing an adaptive weight allocation mechanism based on market conditions (such as stationary periods and extreme volatility periods), such as increasing the weight of the LightGBM model that is more robust to extreme values during extreme volatility periods, fully leveraging the scene adaptation advantages of different sub models; The third is to extend the dimension of risk assessment, combining the predicted results of the model with core risk management indicators such as Value at Risk (VaR), strengthening the practical value of the prediction model in actual investment decisions, and achieving the dual goals of yield prediction and risk control.

References

[1] Zhou, Z. H.: Machine Learning. Beijing: Tsinghua University Press,120-125 (2016)

[2] Jin, L.: Navigating the Financial Landscape: The Power and Limitations of the ARIMA Model. Economics and Culture (2024)

[3] Li, J. F., Zhang, L.: Stock Return Prediction Based on XGBoost and LSTM. LSTM-based stock price prediction with attention mechanism. Systems Engineering-Theory & Practice, 42(5), 1210-1221. (2022)

[4] Chen, T., Guestrin, C.: Generalized autoregressive conditional XGBoost: A Scalable Tree Boosting System//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 785-794. (2016)

[5] Qin, Y. & Kong, H. M.: Stock price prediction based on machine learning algorithms. Shantou university jounrnal (Humans ＆ social science monthly) (2025)

[6] Bhargava, K. & Tejaswini, K.: Comparative Analysis of ML based Gradient Boosting Algorithms: XGBoost, CatBoost, and LightGBM. Journal of Scientific and Engineering Research, 7(8):235-239 (2020)

[7] Kumbure, M. M., Lohrmann, C., Luukka, P., et al.: Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Application (7): 197. (2022)

[8] Ampomah, E. K. & Qin, Z. G. & Nyame, G.: Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement Information, 11(6), 332 (2020)

[9] Hansun, S.: A new approach of moving average method in time series analysis. conference on new media studies (2013)
[10] Dinesh, S., Rao, N., Anusha, S. P.: Prediction of trends in stock market using moving averages and machine learning. International Conference for Convergence in Technology (I2CT) (2021)
[11] Li, B.: An explanation for the distribution characteristics of stock returns. arXiv preprint arXiv:2312.02472, 2023 - arxiv.org (2023)