Application of AdaBoost in Predicting Individual Stock Returns

Shuotong Xing

School of Finance, Central University of Finance and Economics, Beijing 102206, China
2024310519@email.cufe.edu.cn

Abstract. Since the stock market is affected by multiple factors, stock return series exhibit significant nonlinear and nonstationary characteristics, and traditional linear models struggle to effectively fit their volatility patterns. This paper takes the daily return of a share as the research object, constructs a simplified factor system covering volume and price, technology, momentum, and nonlinear interaction factors, and constructs three models of linear regression, random forest, and AdaBoost for comparative analysis. Focusing on the problem that the AdaBoost model is easily disturbed by redundant factors and has weak generalization ability in regression tasks, special optimization is carried out from three aspects: factor screening, parameter adjustment, and outlier processing. The empirical results show that the optimized AdaBoost model is significantly better than linear regression in the test set, and is close to the performance level of a random forest. Factor importance analysis shows that the momentum factor and the volatility factor contribute the most to yield forecasting, and the nonlinear interaction factor also plays a positive role. The research shows that through extreme factor reduction and parameter optimization, the AdaBoost model can effectively improve the prediction accuracy of individual stock returns, which can provide a reference for quantitative investment decision-making.

Keywords: AdaBoost, Individual Stock Returns, Integrated Learning.

1 Introduction

With the continuous development of China's capital market, the prediction of individual stock returns has become a core research topic in the fields of quantitative investment, risk management, and asset pricing. The proliferation of high-frequency trading data and the rapid application of machine learning techniques have provided new research pathways for stock return forecasting [1]. However, stock prices are influenced by multiple factors, including market conditions, firm fundamentals, and investor behavior, exhibiting significant nonlinearity and non-stationarity. Traditional linear models struggle to fully capture these underlying patterns, resulting in limited prediction accuracy [2].

In the field of individual stock return prediction, scholars have conducted extensive research from various perspectives. At the methodological level, existing reviews have systematically examined the application of machine learning in stock prediction,

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_46

categorizing the current research into single models, improved models, and ensemble models [3, 5]. In empirical studies, the random forest model has been applied to predict the returns of CSI 500 constituent stocks, demonstrating a strong capability in fitting relative stock returns [4]; further research has confirmed the effectiveness of random forest in quantitative stock selection [7]. For the AdaBoost model, one study combined it with a time-weighted mechanism to construct a dynamic multi-factor stock selection model, validating its effectiveness in A-share markets [6]; another study improved the AdaBoost algorithm specifically for stock selection tasks [8]. Additionally, Boosting algorithm fusion models under the Stacking framework have been applied to quantitative investment strategy design [9], while multi-factor models combined with machine learning have been extensively explored for quantitative investment [10]. Deep learning-based approaches have also been applied to multi-factor stock price prediction, further enriching the methodological toolkit in this domain [11]. These studies provide valuable references for the application of machine learning in stock return prediction. However, there remains a lack of systematic optimization schemes addressing the specific shortcomings of the AdaBoost model in regression tasks—namely, its susceptibility to redundant factor interference and its limited generalization ability.

Existing literature has analyzed the forecasting and pricing of financial derivatives from the aspects of option pricing model construction, parameter estimation, numerical solution and volatility modeling, and some studies have focused on the application of machine learning in stock yield forecasting, but there are still research gaps in the extreme factor reduction and optimization of AdaBoost model, the adaptability comparison of different integrated learning models in stock yield forecasting and the multi-dimensional verification of model robustness. This paper takes the daily return of A-share stocks as the research object, uses CSMAR stock market database data, adopts data preprocessing, factor reduction, model construction and special optimization, and obtains the conclusion that AdaBoost model can approach random forest performance after factor and parameter optimization, and that nonlinear factors play an important role in yield prediction by analyzing the difference in prediction performance between linear regression, random forest and optimized AdaBoost model. In theory, it clarifies the optimization effect of simplified nonlinear factors on the AdaBoost model. In practice, it can provide efficient and reliable technical support for quantitative investment decision-making and stock risk management, and help investors improve the accuracy of yield forecasting.

2 Theory and Model

This paper mainly involves two core theories: one is the deep learning prediction theory, which takes data as the core. It captures the complex nonlinear relationship of data by building a multi-level learning structure, breaks the limitations of the traditional linear model, and provides theoretical support for the application of the machine learning model in the prediction of individual stock returns. Its core idea is highly consistent with the integrated learning logic of the AdaBoost model, emphasizing that

the prediction accuracy can be improved through multi-level and multi-dimensional learning. Second is the stock return volatility theory, which holds that the volatility of individual stock returns is not random but is jointly affected by internal laws and external factors and is thus predictable. Its characteristics, including volatility clustering and long memory, provide a theoretical basis for factor engineering construction and model optimization in this paper. In addition, combined with the efficient market hypothesis and behavioral finance theory, this paper further explains the internal logic of the volatility of individual stock returns, and shows that by capturing the correlation between market trading behavior and factors, the accuracy of yield forecasting can be improved, which provides theoretical support for the research ideas and methods of this paper.

The AdaBoost model is a classic serial integrated learning model. Its core principle is to train multiple weak learners iteratively, dynamically adjust the sample weight and learner weight according to the prediction error of each weak learner, and combine multiple weak learners to form a strong learner. Finally, high-precision prediction is realized. Its core structure can be divided into input layer, hidden layer and output layer: the input layer is responsible for receiving the characteristic data after preprocessing and factor optimization, and completing the input and preliminary transmission of data. The hidden layer is composed of several weak learners, each of which is trained independently, focusing on capturing the local nonlinear law in the data, focusing on the samples with large prediction errors through sample weight adjustment, and gradually correcting the prediction deviation. The output layer integrates the prediction results of all weak learners through weighted voting to output the final prediction value of individual stock returns. The AdaBoost model demonstrates strong adaptability in nonlinear time series prediction. Its serial iterative error correction mechanism can effectively capture the nonlinear fluctuation patterns of individual stock returns, requires no preset data distribution assumptions, flexibly adapts to the nonstationarity and volatility clustering of returns, and effectively reduces overfitting risk through the integration of weak learners.

3 Data Processing and Model Building

The data used in this paper comes from the CSMAR stock market database. The original data are daily trading data of individual stocks, with approximately 6.02 million observations, covering basic indicators including stock code, trading date, closing price, trading volume, turnover and tradable market value, as well as technical indicators including MA5 and RSI14. The target variable is the daily log return. The preprocessing process includes: first, sample screening. Stocks with complete trading data were retained, and 35700 samples were obtained, taking into account training efficiency and data quality. Second, outlier processing. The extreme values of the target variable are truncated using the 3σ rule to prevent outliers from interfering with AdaBoost's error correction. Third, missing value processing. Observations with missing values for target variables and core factors are deleted, and the remaining factors are imputed by

stock group averages to ensure data completeness. Fourth, standardize all characteristics according to stock grouping to eliminate dimensional differences.

The volatility of individual stock yields, options and futures prices is affected by multiple factors, and the core can be divided into three categories: first, macroeconomic factors, including interest rates, inflation rates and GDP growth, directly affect market capital supply and investor expectations. Second, the company's fundamental factors, such as profitability, asset liability ratio, industry status and so on, determine the intrinsic value of stocks. Third, market trading factors, covering turnover, turnover rate, technical indicators, etc., reflect the impact of short-term market trading behavior on prices. From the perspective of time series characteristics, the yield of individual stocks belongs to a typical non-linear time series, which has the characteristics of nonstationary, volatility aggregation, long memory, and peak thick tail, and its volatility law is difficult to be accurately described by a linear relationship. Given the characteristics of AdaBoost, which is sensitive to redundant features and requires nonlinear inputs, this study constructs a 4-category factor system consisting of 9 characteristic factors. The basic volume price factor includes closing price, trading volume, and turnover. Technical indicators include the 5-day moving average and the relative strength index. Return-related factors include lag phase 1, yield, 5-day momentum factor, and 10-day reversal factor. The nonlinear interaction factor includes the price-volume interaction factor. Delete redundant indicators, and finally retain 35700 data.

Based on the simplified screened data, three types of models are constructed for comparison. First, linear regression, default parameter training, is used to verify the added value of nonlinear factors. Secondly, as a comparison model, random forest is a parallel integration model, which is suitable for capturing complex nonlinear relations by setting n_estimators=100, max_depth=10, and min_samples_split=3. Thirdly, as the core of optimization, the AdaBoost integration model is serial, and the base learner is a decision tree.

In view of its shortcomings, the above extreme reduction factor system is adopted at the factor level, and the nonlinear interaction factor is introduced to activate the nonlinear fitting ability. The learning rate dropped from 0.1 to 0.05, the number of weak learners increased from 100 to 200, and the loss function was square, balanced fitting, and generalization. Strengthen the truncation of 3σ outliers, ensure that there is no lack of data, improve the stability of training, and obtain an AdaBoost model after optimization. In chronological order, 35700 samples were divided into 25000 training sets, 3570 verification sets, and 7140 test sets, with a ratio of about 7:1:2. The training set is used for model training, the verification set is used for parameter tuning, and the test set is used for final performance evaluation. The mean square error, mean absolute error, and coefficient of determination were used to evaluate the performance of the model.

4 Model Prediction Effect

4.1 Model Fitting Effect

This paper selects the last 50 trading days of the sample stocks in the test set to draw a comparison chart between their real returns and the predicted values of each model. As shown in Fig. 1, the thick blue line represents the actual log return, the red dashed line represents the predicted value of the optimized AdaBoost model, and the green dashed line represents the predicted value of linear regression. The AdaBoost model can better follow the real value in trend turning points and fluctuation ranges, especially in some local fluctuations, showing stronger nonlinear capture ability, while the linear regression prediction value is relatively smooth, and the fitting ability of nonlinear laws is obviously insufficient. This intuitively confirms that the fitting effect of the AdaBoost model on the complex model of individual stock returns has been significantly improved through factor reduction and parameter optimization.

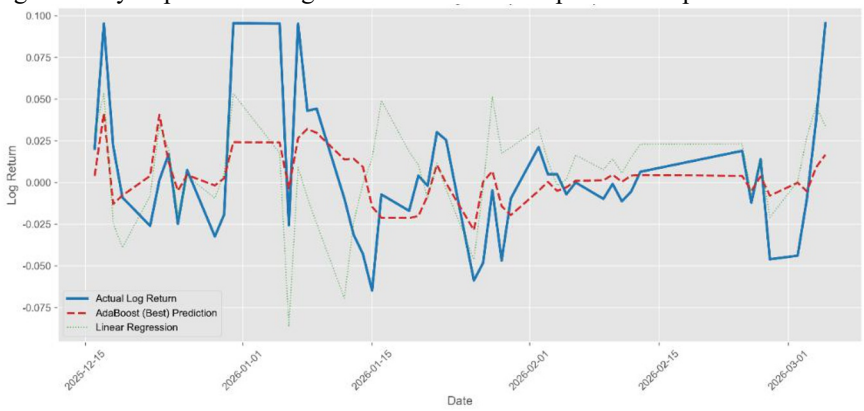


Fig. 1. Comparison of actual and predicted log returns (Picture credit: Original).

4.2 Verification of Prediction Effect

The prediction performance of the model is evaluated by three core indicators: mean square error, mean absolute error, and coefficient of determination. The final evaluation results of the three models on the test set are shown in Table 1. The random forest model achieves the best performance, with an R^2 value of 0.502 and the lowest MSE. After optimization, the AdaBoost model yields an R^2 of 0.40 and an MSE of 0.0007, which significantly outperforms linear regression and approaches the performance level of the random forest model. This result is in line with the characteristics of the model-the parallel integration mechanism of random forest has a higher fault tolerance rate for continuous value prediction, while AdaBoost, as a serial integration model, has a naturally weaker generalization ability in regression tasks than random forest, but its performance has been greatly improved through extreme factor reduction and parameter optimization. AdaBoost R^2 is close to random forest, and MSE is also at a low level, which verifies the effectiveness of the optimization strategy.

Table 1. Model evaluation results.

Model	R^2	MSE	MAE
Linear regression	0.28	0.0008	0.0221
Random forest	0.50	0.0006	0.0183
AdaBoost (Best)	0.40	0.0007	0.0198

4.3 Factor Importance

As shown in Fig. 2, this study further analyzes the factor importance of the AdaBoost model and finds that the momentum factor MOM5 and the price moving average gap MA5_gap contribute the most, followed by the volatility factor VOLAT5 and RSI14. This shows that short-term trends and volatility characteristics are the most critical to yield forecasting, and the constructed non-linear interaction factors also play a role, indicating that the model successfully uses the non-linear relationship.

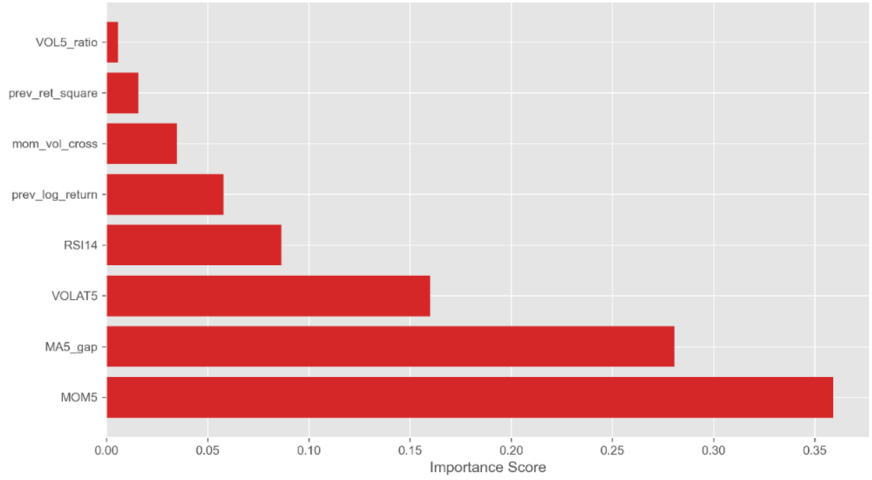


Fig. 2. Factor importance of the optimized AdaBoost model (Picture credit: Original).

Based on the yield forecast results of the model, a simple investment strategy can be constructed: for example, when the log yield of the forecast next day is higher than a certain threshold, buy, and when it is lower than a certain threshold, sell or short position. The strategy is tested back on the test set, and indicators such as strategy return, maximum withdrawal, and Sharpe ratio can be calculated. From the perspective of prediction accuracy, the optimized AdaBoost model has a certain timing ability, and its prediction direction accuracy is about 55%-60%, which can provide reference signals for quantitative transactions according to rough estimates of R-squared and MSE.

5 Conclusion

Taking the daily returns of A-share stocks as the research object, this paper constructs a simplified factor system including volume-price indicators, technical indicators, momentum factors and nonlinear interaction factors. Linear regression, random forest, and AdaBoost models are used to predict and compare the rate of return, respectively. In view of the shortcomings of the AdaBoost model in the task of return, special optimization is carried out from three aspects: factor screening, parameter adjustment, and outlier processing. The empirical results show that the R^2 of the optimized AdaBoost model on the test set is 0.40 and the MSE is 0.0007, which is significantly better than the linear regression model and close to the performance level of a random forest. Factor importance analysis shows that momentum factor MOM5 and price average difference Ma5_gap contribute the most to the forecast, while volatility factor and RSI14 also play an important role, indicating that short-term trends and volatility characteristics occupy a core position in the forecast of individual stock returns. From the research results, the AdaBoost model can effectively improve the fitting ability of nonlinear characteristics after extreme factor reduction and parameter optimization, and verify the applicability of ensemble learning in the prediction of individual stock returns. Although its prediction accuracy is still slightly lower than that of random forest, it has unique advantages in factor sensitivity and error correction mechanism, and is suitable for prediction scenarios with high requirements for non-linear relations. The research also shows that the quality of factor engineering directly affects the performance of the model, and the elimination of redundant factors and the introduction of nonlinear factors are the key links to improve the generalization ability of the model.

The main contributions of this paper are reflected in three aspects. First, this study systematically compares the performance of linear regression, random forest and AdaBoost in individual stock return prediction and reveals the impact of different ensemble mechanisms on prediction accuracy. Secondly, aiming at the shortcomings of AdaBoost model in regression task, a set of optimization strategies considering factor screening and parameter optimization are proposed, which effectively improves its prediction performance. Thirdly, through factor importance analysis, the core role of short-term momentum and volatility characteristics in the prediction of individual stock returns is clarified, which provides an empirical basis for the construction of follow-up factors. On the whole, this paper makes up for the shortcomings of existing studies in AdaBoost factor optimization and multi-model comparison, and provides a reusable methodological reference for research in related fields. However, this study still has certain limitations. For instance, model training relies on historical data and assumes that market patterns maintain certain continuity, and the prediction performance may decline under the impact of structural changes or emergencies. In addition, this paper does not conduct a complete strategic back test of the forecast signal, and in order to consider the transaction cost factor, its practical applicability still needs to be further verified. Future research can try to combine AdaBoost with a deep learning method to build a hybrid model, taking into account both nonlinear fitting ability and generalization ability.

References

1. Li X.Y., Li Z.F., Li Q.Q., et al. A Review of Research on Asset Return Prediction Based on Machine Learning[J]. *China Management Science*, 2025, 33(01): 311-322.
2. Mu Y., Li D.Q. Research on Stock Price Forecasting Based on Markov Chain Method[J]. *Financial Development Research*, 2024(08).
3. A Review of Stock Prediction Research in Quantitative Investment Based on Models and Algorithms[J]. *Computer Engineering and Applications*, 2025(19).
4. Application of Machine Learning in Stock Prediction[J]. *Information Systems Engineering*, 2024(04).
5. Yang Y.Y., Lu T.Y., Ren T.T., et al. Dynamic Multi-factor Stock Selection Model Based on Time Weighting and AdaBoost Ensemble[J]. *Systems Engineering*, 2025(01).
6. Zhou L. Research on Stock Multi-factor Investment Based on Random Forest Model[J]. *Financial Theory and Practice*, 2021(07).
7. Xu Y.E., Luo Y. A Review of Stock Prediction Research Based on Combination Models[J]. *Science Technology and Industry*, 2022(11).
8. Wang S.Y., Cao Z.F., Chen M.Z. Research on Application of Random Forest in Quantitative Stock Selection[J]. *Operations Research and Management*, 2016(03): 163-168.
9. He C., Wu F., He Y., et al. Stock Selection Model Based on Improved AdaBoost Algorithm[J]. *Intelligent Computer and Applications*, 2020(02).
10. Chen C.L., Zou X.N. Design and Application of Quantitative Investment Strategy Based on Boosting Algorithm Fusion Model under the Stacking Framework[J]. *Journal of Econometrics*, 2024, 4(2): 425-441.
11. Zou L.X. Research on Quantitative Investment Based on Multi-factor Model[J]. *China Business Review*, 2021(24): 100-103.

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.

