Original Research Article

Fund price prediction and simulated trading based on machine learning

Jingxiong Gao
School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, Anhui, 233000, China

Abstract: This paper explores fund return prediction and the performance of long-short strategies via machine learning, taking Chinese open-end equity and partial equity hybrid funds as samples. We construct 13 fund characteristic anomaly factors and verify their explanatory power for fund returns using Fama-MacBeth regression. Adopting the LightGBM model with empirical hyperparameters and a 12-month rolling window framework, we predict fund returns, sort funds into ten groups by predicted returns, and build a long-short portfolio by going long on the top group and short on the bottom one. Empirical results reveal a significant return gradient effect across the ten groups. The study verifies the feasibility and profitability of the LightGBM-based long-short strategy, and provides targeted investment implications for different types of investors, highlighting the value of fund characteristic anomaly factors in portfolio optimization.

Keywords: return prediction; anomaly factors; lightGBM; rolling window; machine learning

1. Introduction

With the sustained and steady development of China's economy, household wealth has accumulated rapidly. the fund market is undergoing profound changes and vigorous development. According to data from the Asset Management Association of China (AMAC), as of the end of July 2025, there were 164 public fund management institutions in China, with a total net asset value of publicly offered funds reaching 35.08 trillion yuan, breaking through the 35-trillion-yuan mark for the first time. To effectively meet investors' core demand for asset value preservation and appreciation, the theoretical value and practical significance of fund research have become increasingly prominent. In investment practice, selecting high-quality fund products and constructing more stable and sustainable fund portfolios can strongly support households and institutional investors in achieving asset appreciation.

2. Theoretical foundation and literature review

Since the last century, scholars have conducted extensive research on the driving factors of fund returns. Research methods have evolved from early single-factor models to multi-factor models, and in recent years, to the deep integration of machine learning, deep learning and multi-factor pricing. Research methods have been continuously innovated, and research perspectives have been continuously expanded. In the field of asset pricing, multi-factor pricing models serve as the theoretical foundation for return prediction.

2.1. Fund characteristic anomaly factors

As key tools for capturing return drivers, fund characteristic factors have long been the focus of academic research in both theoretical development and empirical application. Anomaly factors refer to characteristic indicators that can predict asset returns and cannot be fully explained by traditional pricing models.

Net value characteristic factors are mainly constructed based on fund historical returns, volatility, risk-adjusted returns and other net-value-related indicators, representing the most basic and widely used factors in fund return prediction.Grinblatt et al. (1995)[1] studied U.S. mutual funds and found that the momentum effect has significant predictive power, and funds with higher momentum factors achieve higher future excess returns. Fama & French (2010)[2] studied U.S. mutual funds and found that although most funds do not generate significant excess returns, the historical excess returns (alpha) of some funds exhibit persistence and can predict future performance.

2.2. Tests of anomaly factor pricing power

With the proliferation of anomaly factors, scholars have focused on the robustness and determinants of predictive power to distinguish genuine pricing factors from statistical false discoveries. Many scholars have systematically re-examined anomalies via portfolio sorts and FamaMacBeth regressions, finding that only some factors sustainably deliver excess returns.

For example, Hou et al. (2017)[3] used a uniform portfolio sorting method to reexamine 447 anomaly factors documented in the literature. Such studies revalidate firm characteristics with independent pricing power, standardize factors as model inputs to predict oneperiodahead returns, and construct longshort portfolios to backtest profitability, verifying the pricing ability of characteristic factors.Harvey et al. (2016)[4] criticized the "factor zoo" phenomenon in finance, arguing that many new factors may be false discoveries from data mining and require stricter statistical thresholds.

2.3. Application of machine learning algorithms

Research on fund return prediction using anomaly factors has advanced substantially in depth and scope, shifting from single-factor to multi-factor frameworks, from linear to nonlinear relationships, and from U.S. to global markets. Methodological innovation has deepened understanding of fund return predictability.

Linear machine learning algorithms handle high-dimensional data, avoid overfitting and improve out-of-sample accuracy through regularization and variable selection. Representative models include Ridge Regression, LASSO (Least Absolute Shrinkage and Selection Operator), and Elastic Net Regression.Feng et al. (2020)[5] used LASSO to select hundreds of fund characteristics and found that profitability and investment factors are the most statistically significant predictors.

DeMiguel et al. (2023)[6] showed that the Gradient Boosting Decision Tree (GBDT) can extract effective information from conventional fund characteristics. Long-only portfolios formed by predicting top-performing funds via GBDT yield significant risk-adjusted returns.Leippold et al. (2022)[7] applied the Random Forest model to fund return prediction in the Chinese stock market, capturing nonlinear relationships between stock characteristics and future returns and achieving strong predictive performance.

3. Data preprocessing

This paper selects open-end funds of ordinary equity type and partial equity mixed type from the CSMAR Public Fund Database. For pricing factors, this paper first selects 14 fund performance characteristic variables to proxy anomaly factors, as shown in the table below:

Table 1. Characteristic factors.

Characteristic Factors	Abbreviation
Average Return	AvgReturn
Standard Deviation of Returns	StdReturn
Sharpe Ratio	Sharpe
Average Return of Market Portfolio (%)	MktPortfolioAvgReturn
Beta Coefficient	Beta
Jensen's Alpha	Alpha
Treynor Ratio (%)	Treynor
Market Timing Gamma	TMTiming
ly Stock Selection Alpha	TMStkSlct
ly Bear Market Timing Gamma1	CLBearTiming
ly Bull Market Timing Gamma2	CLBullTiming
ly Market Timing Gamma	CLTiming
ly Stock Selection Alpha	CLStkSlct
Holding Period Return	RangeReturn

Using the Fama-MacBeth regression,The regression sample covers a time dimension of 121 months and a cross-sectional dimension of 1,509 funds.

Table 2. Fama-MacBeth regression.

Statistical Indicators	Value
Overall R-squared	0.54
Between R-squared	0.34
Within R-squared	0.54

In terms of overall fitting performance, the model achieves an Overall R-squared of 0.54, implying that the 13 characteristic factors jointly explain 54% of the variation in fund monthly returns, which is moderate in asset pricing regressions. The Within R-squared is also 0.54, indicating that the factors have strong explanatory power for the time-series fluctuations of fund returns and can effectively capture the dynamic patterns of fund returns across periods. In contrast, the Between R-squared stands at 0.34, which is relatively low. This suggests that these factors have limited explanatory power for cross-sectional return differences among funds, and there may exist unobserved fund-specific fixed characteristics (such as fund size, investment style differences, etc.) that affect cross-sectional returns.

4. Empirical analysis

According to studies by Li Bin (2020)[9], gradient boosting decision tree algorithms perform well in pricing research using characteristic anomaly factors with machine learning methods. Therefore, this paper adopts the LightGBM method as a representative machine learning approach for theoretical return analysis.

4.1. Model architecture design

The LightGBM model involves multiple hyperparameters that directly affect model training and prediction accuracy. The specific parameters are set as follows: n_estimators (number of decision trees) = 100, which is widely used in financial forecasting to balance fitting and generalization.learning_rate = 0.05 to improve stability and reduce overfitting.max_depth = 5 to limit model complexity and avoid overfitting to training data.random_state = 42 to ensure reproducibility.verbose = -1 to disable training logs and improve efficiency.n_jobs = -1 to use all available CPU threads for faster computation on largescale data.

this study uses empirically common parameters in financial machine learning to ensure robustness and practicality. The rationality of parameter settings is indirectly verified by the significant and persistent returns of the longshort portfolio.

In each forecasting round of the rolling window, the trained LightGBM model predicts the nextmonth fund returns. Funds in the test set are then sorted by predicted returns and divided into 10 equalsized groups. When the total number of funds cannot be evenly divided by 10, the first 9 groups maintain equal size, and the remaining samples are assigned to the 10th group to minimize imbalance.

The average monthly return of each group is calculated as the group return. The longshort portfolio return is defined as the difference between the average return of Group 10 (highest predicted return) and Group 1 (lowest predicted return), which hedges against systematic risk.

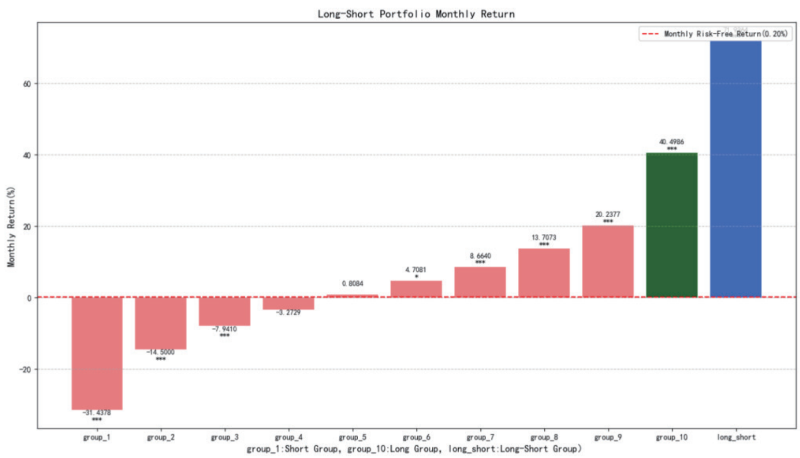

Figure 1. Long-short portfolio monthly return.

4.2. Portfolio performance analysis

As shown in the tables and figures, the long-short returns present a significant return gradient effect. From Group 1 to Group 10, the average monthly return increases gradually, accompanied by improved statistical significance and risk-adjusted returns, which strongly supports the predictability of the LightGBM model and the effectiveness of the long-short strategy.

Group 1 (the lowest-expected-return group) achieves an average monthly return of −31.4378%, significant at the 1% level, with an annualized Sharpe ratio of −3.3658. Groups 2 and 3 also show significantly negative returns, but the negative magnitude gradually narrows. Groups 4 to 6 display a transition from negative to positive returns, with weak significance. From Group 7 to Group 10, returns become significantly positive and rise sharply. Group 10 (the highest-expected-return group) reaches an average monthly return of 40.4986%, significant at the 1% level, with an annualized Sharpe ratio of 3.9395.

The return range from Group 1 to Group 10 is 71.9364%, which fully validates that the LightGBM model can effectively distinguish funds with different future performance. The significant gradient and stable monotonic pattern confirm that the model can accurately capture cross-sectional differences in fund returns and support profitable long-short strategies.

5. Conclusion

Investors should scientifically apply fund characteristic anomaly factors and model tools to optimize their portfolios based on their risk appetite, capital scale, and practical capabilities. The LightGBM model is recommended as the core screening tool, focusing on funds with high expected returns identified by the model to construct a dynamically adjusted portfolio with monthly rebalancing.

In asset allocation, investors may refer to the key effective factors captured by the model: historical excess return, short-term return momentum, market linkage, and risk-adjusted efficiency.Funds with strong performance in these four dimensions should be prioritized.Meanwhile, equal-weight allocation can be adopted to diversify single-fund risk and reduce unsystematic risk caused by high concentration.

Considering transaction costs and constraints in practice, investors should reasonably control rebalancing frequency, avoid illiquid niche funds, reduce the erosion of returns by transaction costs, and allocate a small proportion of low-correlation assets to hedge systematic market risk and further improve risk-adjusted returns.

References

[1] Grinblatt, M., S. Titman and R. Wermers, Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. American Economic Review, 1995. 85(5): p. 1088-1105.

[2] Fama, E.F. and K.R. French, Luck versus skill in the cross-section of mutual fund returns. The journal of finance, 2010. 65(5): p. 1915-1947.

[3] Hou, K., C. Xue and L. Zhang, Replicating anomalies. 2017, National Bureau of Economic Research.

[4] Harvey, C.R., Y. Liu and H. Zhu, … and the cross-section of expected returns. The Review of Financial Studies, 2016. 29(1): p. 5-68.

[5] Feng, G., S. Giglio and D. Xiu, Taming the factor zoo: A test of new factors. The Journal of Finance, 2020. 75(3): p. 1327-1370.

[6] DeMiguel, V., et al., Machine learning and fund characteristics help to select mutual funds with positive alpha. Journal of Financial Economics, 2023. 150(3): p. 103737.

[7] Leippold, M., Q. Wang and W. Zhou, Machine learning in the Chinese stock market. Journal of Financial Economics, 2022. 145(2): p. 64-82.

[8] Li, B. and A.G. Rossi, Selecting mutual funds from the stocks they hold: A machine learning approach. Available at SSRN 3737667, 2020.