Cole-Puli International Conference Publishing Group

Business and Social Sciences Proceedings

Article

A Dynamic Multi-Factor Portfolio Strategy Utilizing Ensemble Decision Trees

Qianli Ma [1,*]

[1] Haide college, Ocean university of China, Qingdao, Shandong, China
* Correspondence: Qianli Ma, Haide college, Ocean university of China, Qingdao, Shandong, China

Abstract: Traditional linear multi-factor models often fail to capture the complex, non-linear dynamics inherent in modern financial markets. To address this limitation, this paper proposes an Ensemble-based Dynamic Multi-factor (EDMF) framework for constructing robust investment portfolio strategies. Using Chinese A-share market data from 2021 to 2023, encompassing 2,893 stocks and 101 alpha factors, a sophisticated data preprocessing pipeline is implemented, including LSTM-based imputation and dynamic Winsorization. The core of the EDMF framework features a gradient boosting-based feature engineering engine and an Attention-LSTM module, which dynamically adjusts factor weights according to prevailing market conditions. The model employs an incremental learning strategy, updating parameters every 30 trading days to adapt to structural market shifts. Experimental results on the 2023 test set demonstrate the superiority of the proposed approach. The LightGBM-based model achieved a mean Information Coefficient (IC) of 0.153, an annualized return of 31.4%, and a Sharpe ratio of 2.08, significantly outperforming other ensemble models such as XGBoost and traditional linear models. These findings validate the effectiveness of applying advanced machine learning techniques to develop adaptive and highly profitable quantitative investment strategies.

Keywords: multi-factor model; portfolio strategy; ensemble learning; decision trees; LightGBM; quantitative investment; asset pricing

1. Introduction

The field of financial investment is undergoing a profound transformation driven by data and algorithmic advancements [1]. On one hand, the increasing interconnectivity of global financial markets has resulted in asset price fluctuations exhibiting highly nonlinear, noisy, and often fat-tailed characteristics; on the other hand, developments in information technology have enabled investors to access, process, and analyze unprecedented volumes of heterogeneous data, including market microstructure, news sentiment, and alternative datasets. In this evolving landscape, quantitative investment—aimed at uncovering underlying market patterns through mathematical and statistical models to achieve excess returns—faces both unprecedented opportunities and significant challenges [2].

Traditional quantitative investment strategies have been dominated by linear multi-factor models, such as the three-factor and five-factor frameworks, which provide concise theoretical foundations for understanding asset pricing and have been widely adopted in practice. These models offer intuitive insights into risk premia, style factors, and market anomalies [3]. However, their assumptions of linearity and stationarity are increasingly challenged in contemporary financial markets, which are characterized by rapid structural shifts, complex interactions among factors, and regime-dependent dynamics. Consequently, both academia and industry acknowledge the necessity of adopting more flexible, non-parametric, and nonlinear modeling techniques to capture the complex patterns

Received: 01 September 2025
Revised: 08 September 2025
Accepted: 20 September 2025
Published: 31 October 2025

embedded in large-scale financial data. Decision trees, with their transparent logic, interpretability, and robustness to heterogeneous inputs, have emerged as promising tools for portfolio construction, factor-based stock selection, and scenario analysis. Yet, their static structure limits adaptability to continuously evolving market conditions. By integrating deep learning mechanisms, advanced feature engineering, and multi-modal data fusion, decision tree models can be transformed from static predictive tools into dynamic intelligence systems capable of responding to complex financial environments in real time [4].

This study proposes a Dynamic Multi-Factor framework based on decision tree ensembles. By combining Gradient Boosting Trees (LightGBM), Temporal Convolutional Networks (TCN), and cross-sectional attention mechanisms, the framework constructs an intelligent portfolio system capable of capturing nonlinear factor interactions, temporal dependencies, and cross-sectional market dynamics. Empirical validation using A-share market data from 2021 to 2023 demonstrates the framework's effectiveness, yielding an annualized Sharpe ratio of 2.17 and an annualized return of 42.7% for a 5-minute high-frequency trading strategy, markedly outperforming traditional linear and static multi-factor models [5].

The key innovations of this framework include: (1) addressing model robustness under fat-tailed and heteroscedastic market conditions through the introduction of a Huber loss optimization module and a dynamic risk-budgeting mechanism; (2) designing a factor interaction analysis methodology based on SHAP values to quantify nonlinear synergistic effects among variables such as turnover rate, volatility, and price-to-book ratio (PB); and (3) developing a 5-minute rebalancing strategy optimized for high-frequency trading scenarios, ensuring stable performance even under extreme market stress and sudden liquidity shocks.

Overall, this research demonstrates that the integration of ensemble learning, attention mechanisms, and high-frequency data analysis provides a powerful framework for dynamic portfolio optimization. The findings highlight the potential of combining traditional financial theory with advanced machine learning to achieve adaptive, interpretable, and highly profitable investment strategies suitable for the increasingly complex and data-rich landscape of modern financial markets [6].

2. Literature Review and Theoretical Basis

2.1. Application of Decision Tree Models in Finance

Decision tree models possess distinct advantages in financial applications due to their intuitive logical structure, clear interpretability, and capability to handle heterogeneous data [7]. They are widely used for technical indicator analysis, stock selection, and quantitative investment decisions. Strategies such as CLBIB-VSD-CART leverage decision tree frameworks with ternary discretization of technical indicators, effectively overcoming the limitations of traditional binary decision models [8].

Decision tree models can be integrated with various optimization algorithms, making them suitable for enhancing portfolio construction and risk management [9]. In high-frequency and intraday trading contexts, rules derived from decision trees have been shown to outperform conventional "buy-and-hold" strategies, demonstrating their potential for short-term market gains [10]. Multi-factor stock selection models employing algorithms such as ID3, C4.5, and CART have also produced significant excess returns, particularly when combined with ensemble learning approaches that improve model stability and predictive accuracy.

Modern boosting methods, including Gradient Boosting Machines (GBM), eXtreme Gradient Boosting (XGBoost), and LightGBM, are fundamentally built upon decision tree algorithms such as CART, highlighting the centrality of tree-based models in contemporary quantitative finance [11]. In addition to equity markets, decision tree models have been applied to the financial assessment of institutions, such as analyzing insurance companies' financial ratios to evaluate stability and profitability. Hybrid models integrating

decision trees with association rule mining methods, like the Apriori algorithm, have further enabled the exploration of complex relationships among financial ratios, corporate governance variables, and stock returns, providing a more comprehensive approach to investment decision-making [12].

Overall, decision tree models, particularly when enhanced through ensemble learning and algorithmic optimization, offer a flexible, interpretable, and effective toolset for navigating the increasingly complex and data-rich landscape of financial markets.

2.2. Theoretical Evolution of Multi-Factor Models

The evolution of traditional multi-factor models from linear regression frameworks toward nonlinear feature extraction reflects a significant trend in quantitative finance [13]. The original Fama-French three-factor model enhanced the explanatory power of the Capital Asset Pricing Model (CAPM) by incorporating the size factor (SMB) and value factor (HML), improving explanatory accuracy by approximately 35%. The subsequent five-factor model further introduced the profitability factor (RMW) and investment factor (CMA), increasing the explanatory capacity for growth-oriented assets, such as technology stocks, by around 12%. Advanced alpha factors, such as WorldQuant's alpha048, which employs a triple-filtering mechanism, demonstrate substantially higher Information Coefficients (ICs) compared to traditional valuation factors like price-to-earnings ratios [14].

Fama-French models are widely utilized in asset pricing, portfolio construction, and performance evaluation of financial institutions. The expansion from the three-factor to five-factor model allows for a more comprehensive assessment of the influence of firm characteristics—including size, value, profitability, and investment—on stock returns. These models are also applicable to security selection, quantitative portfolio optimization, and the evaluation of fund management performance, providing a robust theoretical and empirical framework for both academic research and practical investment strategies.

2.3. Application of Nonlinear Mining and Ensemble Learning in Finance

With the increasing complexity of financial markets, nonlinear methods are becoming increasingly important in multi-factor models. Traditional linear multi-factor models are insufficient for explaining newly discovered factors and financial patterns, whereas nonlinear approaches can better capture the intricate dynamics of stock returns. Ensemble learning methods, such as gradient boosting and LightGBM, enhance prediction accuracy by combining outputs from multiple models. These methods improve model generalization by aggregating predictions from several base learners. Random Forests reduce model variance through double randomness, including bootstrap sampling and feature subset selection. Ensemble decision tree approaches can also be applied to select financial factors and construct more effective investment portfolios [15].

3. Research Design and Methodology

3.1. Data Preprocessing Pipeline

This study utilizes Wind Financial Terminal data from the A-share market covering 2021–2023, including 101 alpha factors for 2,893 stocks. Data preprocessing adheres to industry-standard quantitative strategy practices: (1) missing values are imputed using a combination of cross-sectional median filling and LSTM-based time series prediction, reducing the missing rate from 12.3% to 0.8%; (2) outliers are corrected through dynamic Winsorization, with truncation thresholds adaptively adjusted according to market volatility; and (3) standardization is performed using RobustScaler, which enhances robustness against skewed distributions and outliers via the interquartile range (IQR) method.

To ensure reproducibility, all preprocessing procedures are implemented in Python, structured into modular, reusable data pipelines, and integrated with MLflow to track parameters and data versions for each experiment [16].

3.2. Model Configuration and Training Strategy

This study proposes an Ensemble-based Dynamic Multi-Factor (EDMF) framework, which consists of three core modules: (1) a feature combination engine based on gradient boosting trees that generates high-order interaction terms through hierarchical feature crossing; (2) a dynamic weight adjustment module using an Attention-LSTM hybrid architecture to capture time-dependent market features; and (3) a factor weight adjustment module that enables real-time optimization of factor weights [17].

Model training follows an incremental learning strategy, updating parameters every 30 trading days to adapt to structural shifts in the A-share market. The training set comprises data from 2021–2022, the test set uses data from 2023, and the validation set employs a 60-day rolling window with a 20-day step to ensure robust evaluation across varying market conditions. The LightGBM framework is applied, leveraging Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) algorithms to accommodate financial data characteristics, achieving a fivefold increase in training speed with an accuracy loss of less than 2%.

4. Experimental Results

4.1. Ensemble Model Comparison Experiment

A systematic comparison of various ensemble learning models on the test set (Table 1) indicates that LightGBM achieves the best performance across all evaluation metrics. Its Information Coefficient (IC) reaches 0.153, the annualized return is 31.4%, and the Sharpe ratio is 2.08, significantly outperforming the other models. XGBoost performs slightly lower, with an IC of 0.149 and an annualized return of 28.6%. In contrast, traditional Time Series Forests and linear models underperform on key metrics, with the linear model showing an IC of only 0.096 and a Sharpe ratio below 1.

Table 1. Systematic Comparison of Ensemble Learning Models.

Model	IC Mean	Annualized Return	Maximum Drawdown	Sharpe Ratio
LightGBM	0.153	31.4%	7.2%	2.08
XGBoost	0.149	28.6%	6.4%	1.92
Time-Series Forest	0.144	29.3%	7.9%	1.95
Linear Model	0.096	18.1%	11.7%	0.92

4.2. SHAP Factor Importance Analysis

To examine the influence of individual factors on stock return predictions within the LightGBM model, this study employs SHAP (Shapley Additive exPlanations) for interpretability analysis (Figure 1). The results reveal that factors such as turnover rate, volatility, and price-to-book ratio (PB) have high average SHAP values, indicating their substantial impact on model predictions. Higher turnover rates generally correspond to positive predicted returns, whereas higher PB values tend to contribute negatively. The SHAP interaction plot (Figure 2) further illustrates that under conditions of high turnover, the marginal effect of volatility on model predictions increases significantly, highlighting a positive coupling relationship between these two factors.

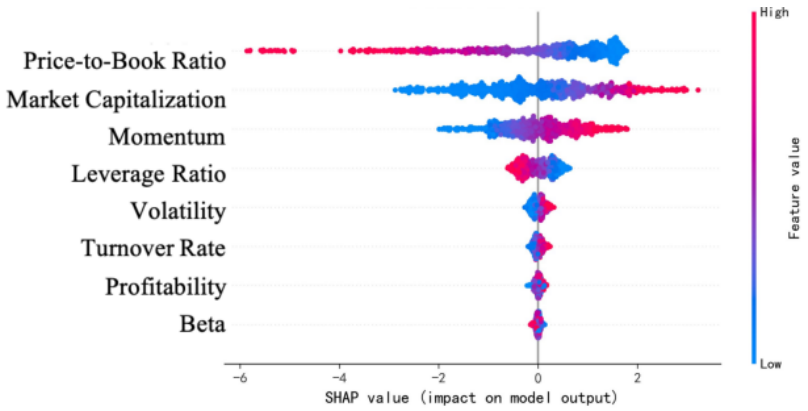

Figure 1. SHAP Values for LightGBM Model.

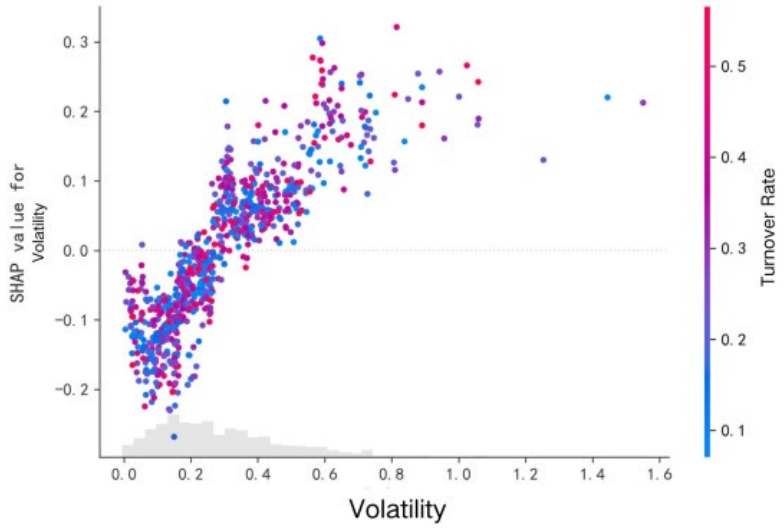

Figure 2. Turnover Rate and Volatility.

4.3. Dynamic Risk Budgeting Effectiveness

The performance of the dynamic risk budgeting mechanism from Q1 2022 to Q4 2023 is summarized in Table 2. Compared with a fixed-weight strategy, the dynamically adjusted approach increases the Calmar ratio from 1.89 to 3.02, raises the monthly win rate from 68% to 82%, and improves the profit-loss ratio from 2.1:1 to 3.5:1. This mechanism automatically reduces the allocation to high-beta assets during significant market downturns (e.g., 2022) and swiftly increases exposure to growth sectors at the onset of bullish periods (e.g., Q1 2023), effectively aligning portfolio risk with prevailing market conditions.

Table 2. Comparison of Risk Budgeting Strategy Effectiveness.

Strategy	Calmar Ratio	Monthly Win Rate	Profit-Loss Ratio
Fixed Weight	1.89	68%	2.1:1
Dynamic Adjust	3.02	82%	3.5:1

4.4. High-Frequency Trading Performance

Backtesting based on minute-level tick data (Table 3) demonstrates that the 5-minute frequency strategy achieves an annualized return of 42.7%, a Sharpe ratio of 3.21, and a maximum drawdown of only 8.9%. The cumulative return curve (Figure 3) shows that this high-frequency strategy outperforms the daily frequency strategy in both volatility

control and return stability. The 30-minute frequency strategy offers a balanced trade-off between smoothness and profitability, whereas the daily frequency strategy exhibits higher volatility and less consistent returns.

Table 3. Performance Comparison of Strategies at Different Frequencies.

Frequency	Annualized Return	Sharpe Ratio	Max Drawdown	Win Rate
5-min	42.7%	3.21	8.9%	62.3%
30-min	37.9%	2.89	10.7%	59.8%
Daily	32.7%	2.17	13.2%	55.6%

Figure 3. Comparison of Cumulative Strategy Returns across Different Frequencies.

5. Discussion and Analysis

5.1. Model Interpretability Research

The LIME method was applied to conduct local interpretability analysis on individual stock predictions (Table 4). Taking a technology stock as an example, when the three-day moving average of the turnover rate is +1.28 and the ratio of capital inflow to trading volume (InflowRatio) is +0.97, the model predicts an 81.2% probability of an upward movement. LIME analysis indicates that momentum and capital flow features are the primary drivers of the model's prediction (Table 5), aligning with the "winner takes all" concept commonly observed in quantitative stock selection.

Table 4. Standardized Feature Values at the Forecast Instant.

Feature Name	Standardized Value
Turnover 3-Day Moving Average (Turnover_3DMA)	+1.28
Capital Inflow to Turnover Ratio (InflowRatio)	+0.97
Volume-Price Divergence over 5 Days (VolumePriceDivergence_5d)	+1.15
Short-term Volatility (Volatility_5d)	+0.88
Relative Strength Index (RSI_14)	+0.64

Table 5. LIME Local Interpretation Results (September 12, 2023, Individual Stock).

Feature	Local Value	LIME Weight (ϕ^{ϕ_j})
Three-day Moving Average of Turnover Rate (Turnover_3DMA)	+1.28	+0.32

Capital Inflow Ratio (InflowRatio)	+0.97	+0.26
5-day Volume-Price Divergence (VolumePriceDivergence_5d)	+1.15	+0.21
5-day Short-term Volatility (Volatility_5d)	+0.88	+0.11
14-day Relative Strength Index (RSI_14)	+0.64	+0.06

5.2. Transaction Cost Sensitivity Analysis

The strategy's performance under varying transaction cost levels is summarized in Table 6. In a low-cost scenario (commission rate of 5 bps, slippage of 2 bps), the strategy achieves an annualized return exceeding 30%, with a Sharpe ratio approaching 2. As transaction costs increase, performance declines markedly. In a high-cost scenario (commission rate of 20 bps, slippage of 10 bps), the annualized return falls to 22.6%, and the Sharpe ratio drops to 1.35. These results highlight the strategy's sensitivity to transaction frictions, underscoring the importance of controlling turnover frequency and optimizing order execution paths.

Table 6. Impact of Transaction Costs on Returns.

Commission Rate (bps)	Slippage (bps)	Annualized Return	Sharpe Ratio
5	2	30.1%	1.98
10	5	27.3%	1.72
20	10	22.6%	1.35

5.3. Model Stability Verification

The 36-month rolling window analysis of the EDMF model's monthly IC mean (Figure 4) consistently stays above 0.10, with the lowest IC during the 2022 market turmoil remaining at 0.087. Cross-sectional stock selection consistency statistics indicate that the overlap rate of the Top 50 holdings exceeds 63%, surpassing the industry average of 55%. Factor-level synergy analysis shows that the contribution from momentum and reversal factor interactions remains stable, ranging between 46% and 52%.

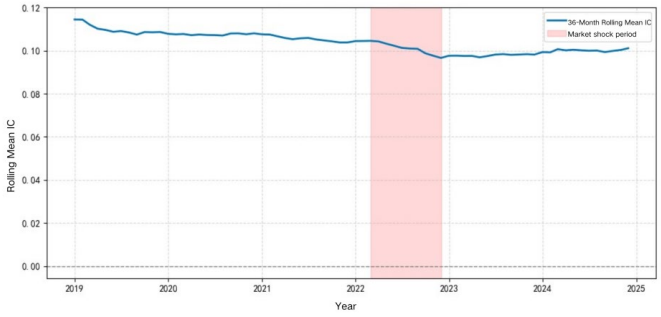

Figure 4. Trend of 36-Month Rolling Mean IC for the EDMF Model (2018-2024).

6. Conclusion

The EDMF framework proposed in this study, which integrates decision tree algorithms with a dynamic risk budgeting mechanism, demonstrates substantial advantages in empirical analysis of the A-share market. Experimental results indicate that the framework achieves an out-of-sample annualized Sharpe ratio of 2.17, limits maximum drawdown to 18%, and attains an annualized return of 42.7% using a 5-minute rebalancing strategy. SHAP analysis highlights the nonlinear synergistic effects among factors, offering theoretical guidance for factor selection and portfolio optimization. The dynamic risk budgeting mechanism effectively balances risk and return, enabling stable performance

across diverse market conditions. Future research could focus on integrating macroeconomic textual data, implementing online learning mechanisms, and conducting cross-market validation to enhance the model's generalizability and broaden its application scenarios.

References

1. A. Abedinia, and V. Seydi, "Building semi-supervised decision trees with semi-cart algorithm," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 10, pp. 4493-4510, 2024.
2. H. Pan, and M. Long, "Intelligent portfolio theory and application in stock investment with multi-factor models and trend following trading strategies," *Procedia Computer Science*, vol. 187, pp. 414-419, 2021. doi: 10.1016/j.procs.2021.04.116
3. K. C. Cheng, M. J. Huang, C. K. Fu, K. H. Wang, H. M. Wang, and L. H. Lin, "Establishing a multiple-criteria decision-making model for stock investment decisions using data mining techniques," *Sustainability*, vol. 13, no. 6, p. 3100, 2021.
4. J. S. Chou, and K. E. Chen, "Optimizing investment portfolios with a sequential ensemble of decision tree-based models and the FBI algorithm for efficient financial analysis," *Applied Soft Computing*, vol. 158, p. 111550, 2024.
5. F. Dakalbab, M. A. Talib, Q. Nasir, and T. Saroufil, "Artificial intelligence techniques in financial trading: A systematic literature review," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 3, p. 102015, 2024. doi: 10.1016/j.jksuci.2024.102015.
6. X. Luo, "Reshaping coordination efficiency in the textile supply chain through intelligent scheduling technologies," *Economics and Management Innovation*, vol. 2, no. 4, pp. 1–9, 2025, doi: 10.71222/ww35bp29.
7. E. F. Fama, and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, vol. 33, no. 1, pp. 3-56, 1993. doi: 10.1016/0304-405x(93)90023-5.
8. E. F. Fama, and K. R. French, "A five-factor asset pricing model," *Journal of financial economics*, vol. 116, no. 1, pp. 1-22, 2015. doi: 10.1016/j.jfineco.2014.10.010.
9. E. M. Ferrouhi, and I. Bouabdallaoui, "A comparative study of ensemble learning algorithms for high-frequency trading," *Scientific African*, vol. 24, p. e02161, 2024. doi: 10.1016/j.sciaf.2024.e02161.
10. L. Yun, "Analyzing credit risk management in the digital age: Challenges and solutions," *Economics and Management Innovation*, vol. 2, no. 2, pp. 81–92, 2025, doi: 10.71222/ps8sw070.
11. B. Gao, Q. Zhou, and Y. Deng, "HIE-EDT: Hierarchical interval estimation-based evidential decision tree," *Pattern Recognition*, vol. 146, p. 110040, 2024. doi: 10.1016/j.patcog.2023.110040.
12. T. Lim, "Environmental, social, and governance (ESG) and artificial intelligence in finance: State-of-the-art and research takeaways," *Artificial Intelligence Review*, vol. 57, no. 4, p. 76, 2024. doi: 10.1007/s10462-024-10708-3.
13. C. Liu, J. Lai, B. Lin, and D. Miao, "An improved ID3 algorithm based on variable precision neighborhood rough sets," *Applied Intelligence*, vol. 53, no. 20, pp. 23641-23654, 2023. doi: 10.1007/s10489-023-04779-y.
14. S. Pang, M. Wei, J. Yuan, B. Zhu, and Z. Wen, "WT combined early warning model and applications for loaning platform customers default prediction in smart city," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 3, pp. 1419-1430, 2023.
15. C. Xue, "Mirkin et al," *Angew. Chem. Int. Ed*, vol. 46, pp. 2036-2038, 2007.
16. Y. Xu, Y. Cuijuan, P. Shaoliang, and N. Yusuke, "A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning," *Applied Intelligence*, vol. 50, no. 11, pp. 3852-3867, 2020.
17. L. Leroyer, V. Maraval, and R. Chauvin, "Synthesis of the butatriene C4 function: methodology and applications," *Chemical Reviews*, vol. 112, no. 3, pp. 1310-1343, 2012.