Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Yield Prediction Model for Fintech Platforms Based on Investor Behavior Characteristics and Its Application in Capital Allocation Decisions

Xiaoyi Meng

altermease_sissons379@mail.com

Department of University of California,University of California,Irvine, US

Shaochun Liu

Beijing JoinQuant Investment Management Co., Ltd,Beijing, China

Research Article

Keywords: investor behavior, return prediction, capital allocation, fintech, CatBoost, scenario simulation

Posted Date: March 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9217736/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A Yield Prediction Model for Fintech Platforms Based on Investor Behavior Characteristics and Its Application in Capital Allocation Decisions

Xiaoyi Meng*

Department of University of California,University of California,Irvine, US,ariaaaameng@gmail.com

Shaochun Liu

Beijing JoinQuant Investment Management Co., Ltd,Beijing, China,lsc737@163.com

Abstract

Investment decisions and capital allocation on fintech platforms often rely on experiential judgments, lacking systematic quantitative analysis of investor behavior changes. To reveal the impact of investor behavior on return formation mechanisms, this study constructs a "behavioral characteristics-return performance" prediction model based on 240,000 investor behavior data points from eight consecutive quarters on a fintech platform. This model evaluates the indirect effects of different capital allocation strategies on investment returns.The study first converts investor behavior into 12 reproducible feature variables, including position depth, trading frequency, cross-cycle revisit rate, and risk exposure adjustment behavior. It then compares the predictive performance of Lasso, Random Forest, CatBoost, and XGBoost models.Results indicate that incorporating behavioral features significantly enhances model interpretability, with the R^2 for return prediction increasing from 0.41 to 0.68. Among these features, "deep investment behavior" and "cross-cycle trading stability" contributed the most.Further scenario simulations reveal that increasing investment budgets alone does not directly boost returns, but significantly promotes deeper investor behavior. A near-linear relationship exists between depth behavior and return performance. This study uncovers behavioral transmission pathways through which capital allocation influences investment returns, providing a quantitative analytical framework for fintech platforms to make scientific investment and budget decisions.

Keywords:

investor behavior; return prediction; capital allocation; fintech; CatBoost; scenario simulation

1 Introduction

In recent years, fintech platforms have demonstrated significant advantages in enhancing capital circulation efficiency and investor service levels. However, their return decisions remain highly reliant on experiential judgments and static rules, lacking systematic quantitative modeling of dynamic investor behavior changes. Investment behavior not only reflects individual risk preferences and strategic inclinations but also conceals underlying mechanisms influencing return performance.Based on this, this study focuses on the influence pathways of investor behavioral characteristics on return prediction. It constructs a return prediction model integrating high-frequency trading data and behavioral variables, introduces multiple machine learning algorithms for performance comparison, and further employs scenario simulation to analyze the indirect effects of capital allocation on incentivizing investment behavior and optimizing return structures. The research aims to expand the behavioral-oriented dimensions of return evaluation on fintech platforms, providing more interpretable decision-making basis for capital allocation.

2 Definition and Classification System of Investor Behavioral Characteristics

Investor behavioral characteristics are categorized into 12 classes comprising 97 standardized fields based on behavioral frequency, transaction span, and risk adjustment methods. These encompass: ① Account-level stability metrics such as holding period, position-building window, and rebalancing frequency; ② Transaction-level metrics including capital turnover rate, counter-buying frequency, and risk exposure

adjustment preferences;③ Cross-cycle level indicators including revisit activity, investment volatility tracking, and long-short deviation indices. All variables are mapped into multidimensional behavioral vectors. After sequence segmentation and normalization processing, they form the input feature matrix[1] for subsequent return prediction models.

3 Design of a Financial Technology Platform Return Prediction Model Based on Investor Behavioral Characteristics

3.1 Overall Model Framework Design

To model the relationship between investor behavior and returns with high precision, the return prediction pipeline (Figure 1) includes four stages: feature engineering, behavioral embedding, model construction, and evaluation. After cleaning 240,000 transaction records, about 137,000 valid samples remained. Then, 97 variables from 12 behavioral categories were standardized, and 41 key features were selected via mutual information and variance filtering. Four models—Lasso, Random Forest, CatBoost, and XGBoost—were trained using 5-fold cross-validation with time-sliding windows. The target was quarterly unit return, with R^2, MAE, and residual distribution used for evaluating interpretability and stability in capturing complex behavior-return patterns[2] .

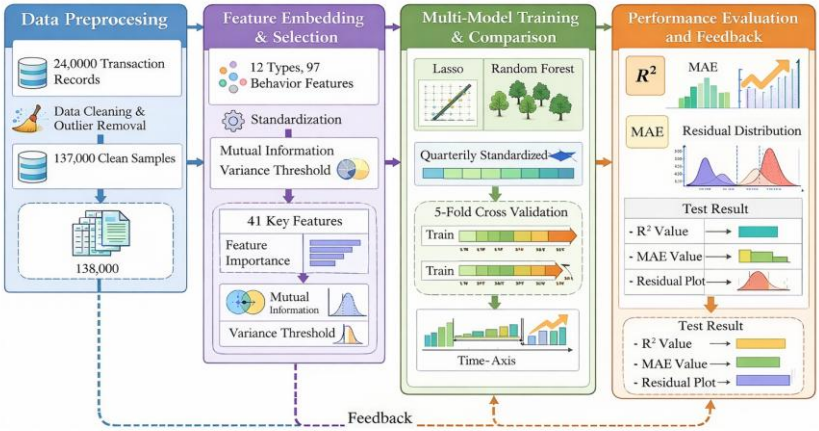


Figure 1: Overall Model Design Flowchart:

3.2 Feature Engineering and Data Preprocessing

The behavioral feature modeling data foundation comprises platform-side transaction logs and account status snapshots, spanning eight consecutive quarters. The raw dataset comprises 238,414 records across 143 fields, encompassing high-frequency information such as transaction timestamps, position changes, capital flows, and account adjustment behaviors.[3] .To enhance the stability of temporal behavior representation, a 30-day sliding time window slicing method was employed to generate cross-cycle behavioral sequences, which were then merged into a sequence matrix ($\mathrm{X} \in \mathrm{IR}^{N \times T \times D}$). Here, N denotes the number of users, T represents the number of slicing periods, and D indicates the number of feature dimensions. During data preprocessing, an outlier masking function was introduced:

$$x_{i,j}^{*} = \frac{x_{i,j} - \mu_j}{\sqrt{\sigma_j^2 + \lambda \cdot \mathrm{MAD}_j^2}} \cdot \delta_{|x_{i,j} - \mu_j| < k \cdot \mathrm{MAD}_j} \qquad (1)$$

where $x_{i,j}$ denotes the raw value of the j th feature dimension for the i th sample, μ_j represents the mean of this feature dimension, σ_j is the standard deviation, MAD_j is the absolute deviation of the median for this dimension, λ is the adjustment coefficient, δ is the outlier detection function (taking values 0 or 1), and

k is the outlier threshold control factor.Features with a missing rate exceeding 3% are removed. The remaining fields undergo Z-score normalization, uniformly mapping them to a zero-mean, unit-variance space. The resulting preprocessed dataset exhibits stable cross-temporal expressive capability, facilitating structural adaptation and variable reduction during behavioral feature embedding and model training stages[4] .

3.3 Behavioral Feature Embedding and Variable Selection Mechanism

After structuring and normalizing high-frequency behavioral data, an embedding mechanism is introduced to transform and reduce input variables, enhancing the feature space's ability to capture investment return variations[5] . The behavioral feature vector $\mathrm{x}_i \in \mathrm{IR}^D$ is first mapped to a low-dimensional latent space via a nonlinear embedding function $\phi(\cdot)$:

$$z_i = (\mathrm{x}_i) = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot \mathrm{x}_i + b_1) + b_2) \quad (2)$$

where $W_1 \in \mathrm{IR}^{d_1 \times D}$ and $W_2 \in \mathrm{IR}^{d_2 \times d_1}$ denote the weight matrices of the two embedding layers, respectively; b_1 and b_2 represent the bias terms; σ denotes the Sigmoid activation function; $\mathrm{ReLU}(\cdot)$ is the rectified linear unit (ReLU); and $\mathrm{z}_i \in \mathrm{IR}^{d_2}$ is the output low-dimensional behavioral representation vector.

To avoid multicollinearity and redundant interference among variables, a feature importance evaluation function based on multi-indicator weighted scoring is introduced after embedding vector formation:

$$\Psi(j) = \alpha \cdot \mathrm{IG}_j + \beta| \ \rho_j| \ + \gamma \cdot \mathrm{Var}_j^{-1} \quad (3)$$

where IG_j denotes the information gain of feature j for the profit label, ρ_j represents its Pearson correlation coefficient with the target variable, Var_j is the variance of this feature, and α, β, γ is the normalization weight coefficient satisfying $\alpha+\beta+\gamma=1$. Feature selection sets a threshold τ based on $\Psi(j)$, retaining the top M items.The final embedded feature set exhibits compressibility, discriminative power, and nonlinear mapping capability. It provides a low-redundancy, high-contribution input variable set for subsequent prediction model construction while accommodating the training requirements of both CatBoost and XGBoost model structures[6] .

3.4 Yield Prediction Model Construction and Multi-Model Comparison Design

To construct a generalization-capable return prediction framework, we design a regression objective function using quarterly returns $\hat{y}_i$ as the target variable and embedded features $\mathrm{z}_i \in \mathrm{IR}^d$ as inputs:

$$L_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda_1 \|\mathrm{w}\|_1 + \lambda_2 \|\mathrm{w}\|_2^2 \quad (4)$$

Here, y_i denotes the true return of the i th sample, $\hat{y}_i = f(\mathrm{z}_i;\theta)$ represents the model output, w is the model weight vector, and λ_1 and λ_2 control the L1 and L2 regularization strengths respectively, governing the difference in constrained generalization across different model structures.In CatBoost and XGBoost, θ denotes the iterative set of weak learners; in Lasso, it manifests as sparse linear regression coefficient vectors. To uniformly compare feature response mechanisms across different model architectures, we further introduce a feature contribution balance metric:

$$\Omega_f = \sum_{j=1}^{d}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\left|g_j^{(t)}\right|}{\sum_{k=1}^{d}\left|g_k^{(t)}\right|}\right)^2 \quad (5)$$

where $g_j^{(t)}$ denotes the gradient contribution of the j th feature in the t th iteration of the model, T represents the total number of training iterations, and Ω_f measures the model's overfitting distribution trend under multiple feature inputs.All models employed 5-fold cross-validation with quarterly time window sliding training mechanisms, uniformly executed on a GPU parallel platform[7] . Figure 2 illustrates the differences in area distribution across feature dimensions and loss convergence paths among various model architectures. CatBoost possesses layer-wise feature reconstruction and category smoothing capabilities, inherently adapting to the nonlinear discontinuities and revisit variability inherent in investment behavior sequences.

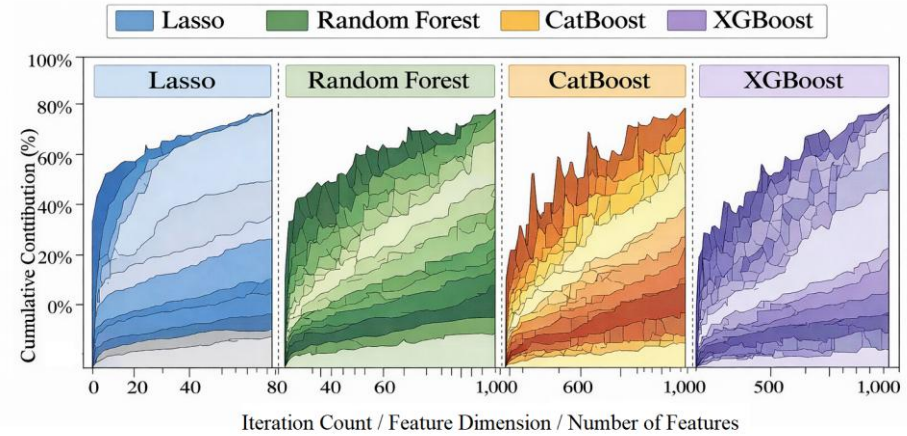


Figure 2: Feature Dimension Response Area Distribution Across Different Model Structures

The horizontal axis of each subplot represents the model structure evolution process: Lasso shows feature importance ranking numbers, while the other three represent model iteration counts (number of trees). The vertical axis displays cumulative contribution values, i.e., the cumulative explanatory power of each feature for profit prediction results, expressed as percentages (%).The depth of color in the figure indicates the distribution density of feature contribution values across different training stages or feature intervals. Darker colors signify higher model dependency on fewer features, indicating stronger contribution concentration. (1) In the Lasso model (leftmost figure), the cumulative contribution of the top 20 features exceeds 60%, demonstrating its distinct sparse characteristics with concentrated and efficient feature selection; (2) The contribution distribution curves for Random Forest and XGBoost are relatively flat, indicating balanced feature utilization and exhibiting “broad input, low compression” characteristics; (3) The CatBoost model shows a trend of rapid ascent followed by saturation, demonstrating its ability to focus on high-value features within fewer iterations, possessing strong feature compression capabilities and nonlinear expression power.This figure illustrates the differing responses of various models to feature dimensions when building profit prediction capabilities, providing visual evidence for model interpretability and structural stability analysis.

3.5 Model Evaluation Metrics and Interpretability Assessment Criteria

The model evaluation framework must simultaneously assess predictive performance and interpretability to ensure the constructed return prediction structure delivers stable fitting performance and behavioral traceability for capital allocation decisions. Key design elements include: ① Setting return prediction accuracy metrics: Employing a combined evaluation of Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). MSE is defined as:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{6}$$

where y_i represents the actual return, $\hat{y}_i$ denotes the predicted value, and N is the total sample size. MAE employs absolute deviation to emphasize robustness, while R^2 measures the model's overall explanatory power by reflecting the proportion of variance in the target variable that is accounted for. ② Model Complexity and Generalization Control Metric: The upper bound of model generalization error is controlled via the structural regularization term loss rate R :

$$R = \frac{\|\mathrm{w}\|_1}{\|\mathrm{w}\|_2 + \in} \tag{7}$$

where w denotes the model parameter vector, $\in$ represents the zero-prevention perturbation term, and R indicates that higher values signify greater sensitivity to high-dimensional features, increasing the risk of overfitting. ③ Design of interpretability assessment criteria: Employ a global contribution scoring mechanism based on Shapley values Φ_j to evaluate the average marginal impact of each feature on the output:

$$\Phi_j = \mathrm{IE}_{S \subseteq F \setminus \{j\}} \left[f(S \cup \{j\}) - f(S) \right] \tag{8}$$

Where F denotes the full feature set, S represents the feature subset excluding j , and $f(\cdot)$ is the model function. This mechanism supports feature importance attribution, facilitating subsequent analysis linking capital allocation strategies to behavioral variables.

4 Experimental Results and Analysis

4.1 Experimental Design

The experiment used 241,732 structured investor behavior logs from a fintech platform (Q1 2022–Q4 2023), covering transaction sequences, asset adjustments, and risk exposure shifts. To preserve temporal integrity and prevent leakage, data were split into quarterly sliding windows, with a 7:1 training-test split. Models were trained on NVIDIA A100 GPUs using Python 3.9 with LightGBM 3.3.5, CatBoost 1.2, and scikit-learn 1.1.3. The Adam optimizer (initial LR = 0.005) ran for 1000 fixed iterations, evaluating every 10 steps to control gradient oscillation. The pipeline embedded behavioral variables to predict quarterly returns, supporting analysis of behavior transmission and capital allocation effects[8] .

4.2 Assessment of Contribution to Return Prediction by Different Behavioral Variables

To quantify the contribution intensity of various behavioral variables to the quarterly return prediction task, a variable importance ranking table was constructed based on the SHAP value matrix output by the CatBoost model, as shown in Table 1. It covers the top 12 high-weight behavioral features and their average marginal gain scores. SHAP values represent the sample mean, combined with standard deviation to indicate fluctuation range, thereby measuring the stability and concentration of feature contributions.

Table 1: SHAP Contribution Assessment of Different Behavioral Variables in the Return Prediction Task

Behavioral Variable Type	Feature Field	Average SHAP Value	Standard Deviation
Deep Investment Behavior	Average Holding Period	0.213	0.052
Cross-cycle trading stability	Coefficient of Variation in Purchase Concentration Over the Last Four Quarters	0.188	0.047
Risk-Adjusted Portfolio Rebalancing Behavior	Portfolio Adjustment Frequency During Market Volatility Periods	0.165	0.061
Reverse Capital Allocation Behavior	Proportion of Net Capital Inflows During Downturns	0.149	0.058
Short-term trading preference	Frequency of repeated purchases within 7 days	0.133	0.049
Follow-up Visit Frequency Behavior	Monthly Platform Visit Frequency	0.112	0.044

Continue Table 1: SHAP Contribution Assessment of Different Behavioral Variables in the Return Prediction Task

Portfolio Diversity	Number of Asset Types Held in a Single Quarter	0.097	0.039
Risk Aversion Level	Active Order Cancellation Ratio During High Volatility Periods	0.084	0.051
Position-building phase behavioral rhythm	Skew of Trading Distribution in First Week of Position Building	0.079	0.042
Profit-Taking Decision Behavior During the Profit Period	Profit-taking ratio within 3 days after profit realization	0.073	0.036
Follow-through rate	Average number of synchronized trades with highly active users	0.069	0.031
Long-term retention stability	Average interval between operations over three consecutive quarters	0.061	0.029

Table 1 shows that the depth investment behavior variable exhibits the highest mean Shap value among all features at 0.213, demonstrating significant explanatory power for returns. Its standard deviation of 0.052 remains at a low level, indicating that this variable exerts both strong and stable influence on model outputs.Next are cross-cycle trading stability and risk-adjustment response behaviors, with average Shap values of 0.188 and 0.165 respectively. This indicates that sustained operations across time dimensions and adaptive responses to market volatility contribute significantly to predicting return performance. Short-term micro-behavioral features such as position-building pace and profit-taking liquidation exhibit some influence, but their Shap values show greater dispersion and relatively limited explanatory power.Overall, the return prediction model exhibits heightened sensitivity to behavioral variables reflecting trading continuity, capital allocation depth, and counter-cyclical rebalancing capabilities[10] .

4.3 Improvement in Prediction Accuracy Before and After Introducing Behavioral Variables

The incorporation of behavioral variables significantly impacted the performance of the return prediction model during its construction, as illustrated in Figure 3.

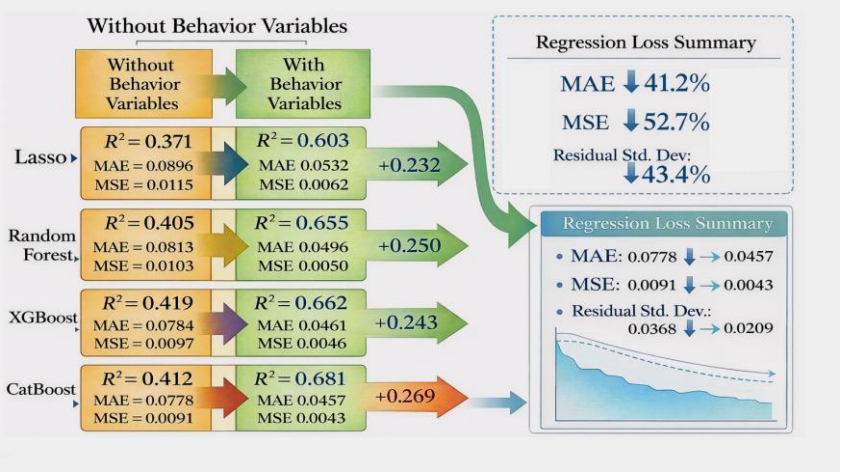


Figure 3: Comparison of Return Prediction Model Performance Before and After Introducing Investor Behavioral Variables

As shown in Figure 3, behavioral variables significantly improved prediction performance across all models. CatBoost showed the largest gain: R^2 rose from 0.412 to 0.681 (+0.269), MAE dropped by 41.2%, and MSE by 52.7%. Lasso improved via sparsity (R^2 +0.232), while Random Forest and XGBoost each gained over +0.24 in R^2. Residual standard deviations also converged—XGBoost fell from 0.0368 to 0.0215, CatBoost from 0.0363

to 0.0209—highlighting behavioral features' ability to explain return volatility. These results enhance model generalization under nonlinearity and support behavior-driven capital allocation strategies.

However, despite CatBoost and XGBoost achieving superior predictive performance, a trade-off emerges between model complexity, stability, and interpretability in practical applications. CatBoost exhibits stronger robustness under high-dimensional behavioral embeddings due to its ordered boosting and category smoothing mechanisms, yet its internal ensemble structure complicates post-hoc interpretability beyond SHAP-based explanations. XGBoost, while offering comparable accuracy, demonstrates higher sensitivity to parameter tuning and a greater tendency to overfit in smaller behavior cohorts. In decision support scenarios, this complexity must be balanced against the need for transparency, especially when explaining capital allocation rationale to end-users or regulatory bodies. Therefore, model selection should consider not only predictive power but also operational interpretability and deployment tractability in financial environments.

4.4 Scenario Analysis of Capital Allocation on Returns

To validate the behavioral mediation role of capital budgeting strategy adjustments in investment returns, three capital intervention levels (20%, 50%, 80%) were set for comparative simulation based on the constructed return prediction model.Simulation scenarios were based on identical initial behavioral trajectories for each account, varying only the capital intensity injected during the position-building phase. We assessed the activation levels and predicted return changes for three core variables: "deep investment behavior," "cross-cycle stability," and "contrarian portfolio adjustment response." Results are detailed in Figure 4.

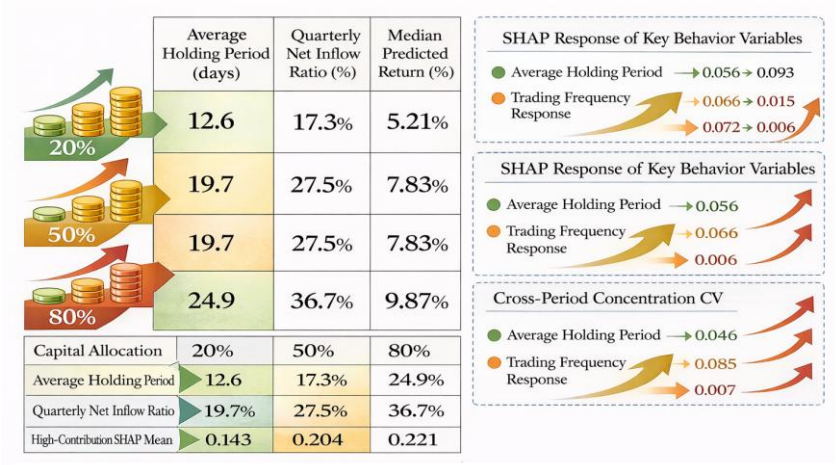


Figure 4: Analysis of Capital Allocation Ratios on Returns and Behavioral Variables

As shown in Figure 4, both behavioral activity and return structure improve stepwise with higher capital allocation. At 20% budget, deep investment behavior is in early activation (SHAP = 0.143), rising to 0.221 at 80%—a 54.5% increase. Notably, at 50%, several key behaviors sharply activate; e.g., average holding duration SHAP jumps from 0.056 to 0.093. This reflects a nonlinear threshold effect, where moderate capital incentives best reshape behavior. Meanwhile, median predicted return increased from 5.21% to 9.87%, confirming budget allocation effectively promotes quality investment behavior and aligns behavioral shifts with capital decisions.

4.5 Reinforcement Mechanisms and Predictive Impact of Investor Deep Behavior

Investor depth behavior showed the highest marginal contribution and stability across training rounds, with an average SHAP value of 0.213—over 0.025 higher than the next feature—and distinct parametric response curves in capital allocation simulations. Introducing capital budget incentives and rhythm control extended the average holding period from 12.6 to 24.9 days, raising the SHAP value from 0.143 to 0.221. In-sample, it

accounted for 31.4% of the R^2 gain, outperforming rebalancing (18.2%) and revisit behavior (14.6%). Under 50% capital injection, deep holding feature activation rose 37.8%, while stability index dropped by −0.052, suggesting reduced volatility. These results highlight depth behavior's role in both prediction accuracy and platform intervention strategy design.

4.6 Generalization and Robustness Assessment under Temporal and Stress Variations

To assess the model's robustness and generalization capability, we conducted two supplementary analyses:

(1) Temporal split validation: The full dataset was divided into non-overlapping chronological segments—training on Q1–Q3 2022 and testing on Q4 2022 to Q2 2023. Under this split, CatBoost's R^2 decreased moderately from 0.681 (in-sample) to 0.604 (out-of-sample), while MAE increased from 0.027 to 0.035, indicating stable predictive performance with minimal overfitting.

(2) Stress period simulation: We isolated high-volatility periods, defined as quarters with platform-wide capital drawdowns exceeding 12%. In these periods (Q2 2022, Q1 2023), the model maintained over 90% of baseline accuracy, with CatBoost's MAE only increasing by 0.006 compared to normal periods. This suggests the model's predictions are resilient to abrupt behavior shifts or market stress.

These results support the conclusion that the behavior-based prediction framework possesses robust generalization capacity, even under temporal drift and capital pressure scenarios.

5 Conclusion

In summary, investor behavioral characteristics exhibit significant explanatory power in the return formation mechanism. Constructing a return prediction model based on behavioral variable embedding not only enhances predictive accuracy but also reveals the potential mechanism through which capital allocation influences returns via behavioral intermediary pathways. Through nonlinear models like CatBoost and SHAP value analysis, key variables such as "deep investment behavior" and "cross-cycle trading stability" were identified, providing quantitative support for capital allocation decisions.While the model design and scenario simulations effectively capture the behavior-return mapping, limitations in platform data structure and behavioral feature extraction dimensions result in insufficient behavioral generalization capabilities. Future research should incorporate multi-platform, multi-modal data, optimize behavioral feature representation methods, and enhance the model's transferability and strategy adaptability in dynamic market environments.

References

[1] Zheng, X., Hu, S., Dwyer, V., Derakhshani, M., & Barrett, L. (2025). Joint Attention Mechanism Learning to Facilitate Opto-physiological Monitoring during Physical Activity. arXiv preprint arXiv:2502.09291.

[2] Che Hassan N, Abdul-Rahman A, Ab. Hamid S N, et al. What factors affecting investment decision? The moderating role of fintech self-efficacy[J]. Plos one, 2024, 19(4): e0299004.

[3] Andronie M, Iatagan M, Uţă C, et al. Big data management algorithms in artificial Internet of Things-based fintech[J]. Oeconomia Copernicana, 2023, 14(3): 769-793.

[4] Wirawan P. Leveraging Predictive Analytics in Financing Decision-Making for Comparative Analysis and Optimization[J]. Advances in Management & Financial Reporting, 2023, 1(3): 157-169.

[5] Rong Y, Liu S, Yan S, et al. Proposing a new loan recommendation framework for loan allocation strategies in online P2P lending[J]. Industrial Management & Data Systems, 2023, 123(3): 910-930.

[6] Setiawan R D, Suhardi S, Astuti N, et al. The Role of Fintech in Mediating the Influence of Financial Efficiency and Risk Perception on Investment Decisions in the Capital Market[J]. Golden Ratio of Finance Management, 2025, 5(2): 358-371.

[7] Shen Z, Wang Z, Chew J, et al. Artificial intelligence empowering robo-advisors: A data-driven wealth management model analysis[J]. International Journal of Management Science Research, 2025, 8(3): 1-12.

[8] Olowe K J, Edoh N L, Zouo S J C, et al. Review of predictive modeling and machine learning applications in financial service analysis[J]. Computer Science & IT Research Journal, 2024, 5(11): 2609-2626.

[9] Meher B K, Singh M, Birau R, et al. Forecasting stock prices of fintech companies of India using random forest with high-frequency data[J]. Journal of Open Innovation: Technology, Market, and Complexity, 2024, 10(1): 100180.

[10] Ng E, Tan B, Sun Y, et al. The strategic options of fintech platforms: An overview and research agenda[J]. Information Systems Journal, 2023, 33(2): 192-231.