Incremental Value of Conditional Volatility in Extreme Gradient Boosting Stock Selection: Evidence from Dynamic China Securities Index 300 Constituents

Siya Chen
School of Data Sciences, Zhejiang University of Finance and Economics, Zhejiang Province, 310000, China
2254715909@qq.com

ABSTRACT

This paper tests whether stock-level GARCH conditional volatility adds predictive and economic value to an XGBoost stock-selection model in China's A-share market. Using a point-in-time panel of 553 historical CSI 300 constituents from 2018 to 2025, the conventional, GARCH, and GARCH-plus-market models attain ten-day AUC values of 0.531, 0.534, and 0.534. With next-open execution and a 20-basis-point one-way cost, their top-decile annualized net returns are 10.93%, 12.14%, and 11.32%. The GARCH signal is visible across several checks, but increments are statistically weak and sensitive to implementation. The evidence supports a marginal, conditional, short-horizon role for conditional volatility rather than a robust standalone trading advantage.

Keywords: A-Share Market; Conditional Volatility; Extreme Gradient Boosting; Generalized Autoregressive Conditional Heteroskedasticity; Quantitative Stock Selection

1. INTRODUCTION

Machine-learning models are increasingly used in empirical asset pricing because they capture nonlinear relations among firm characteristics. Gradient-boosted trees are well suited to structured equity panels in which valuation, size, momentum, liquidity, and risk variables interact [3][4]. Recent evidence also supports complexity and tree-based structures under disciplined out-of-sample evaluation [10][11]. These strengths make leakage control, chronological validation, transaction-cost assumptions, and dependence-robust inference essential.

Volatility is a natural candidate for enrichment. ARCH and GARCH models describe conditional heteroskedasticity through past shocks and conditional variance [1][2]. Unlike rolling standard deviation, GARCH recursively updates a latent risk state. Prior asset-pricing and Chinese stock-selection studies document the relevance of volatility, momentum, liquidity, and valuation characteristics [4][5], but few isolate a stock-level GARCH feature inside the same boosted-tree model for a dynamic Chinese index universe.

This study asks whether rolling GARCH (1,1) volatility improves out-of-sample ranking and implementable portfolio performance beyond conventional factors. Three nested XGBoost models are estimated: M1 uses conventional factors, M2 adds stock-level GARCH volatility, and M3 adds both GARCH volatility and the contemporaneously available CSI 300 return. The design contributes a point-in-time constituent panel, constant labels and trading rules across nested models, and paired inference that separates positive point estimates from statistically reliable increments.

The main finding is deliberately bounded. GARCH-augmented models show small ten-day gains, and trees use the GARCH feature nonlinearly. Yet the increments over M1 are not statistically significant and do not generalize uniformly to 20- and 60-day horizons.

2. LITERATURE REVIEW AND HYPOTHESES

2.1 Machine Learning and Cross-Sectional Stock Selection

XGBoost combines additive regression trees, shrinkage, subsampling, and explicit complexity penalties [3]. These properties make it effective for tabular data with nonlinear thresholds and high-order interactions. Gu, Kelly, and Xiu [4] show that nonlinear machine-learning methods can improve empirical asset-pricing predictions, with price trends, liquidity,

volatility, and valuation ratios among the strongest signals. Evidence from Chinese multifactor stock-selection applications further supports portfolio-level evaluation under realistic out-of-sample settings [5].

2.2 Conditional Volatility as an Incremental Feature

The GARCH (1,1) model extends the ARCH framework by allowing current conditional variance to depend on both the preceding squared innovation and preceding conditional variance [1][2]. This persistence and adaptive recursion distinguish it from a fixed-window historical standard deviation. However, volatility forecasting and cross-sectional return ranking are different tasks. The relevant question here is not whether GARCH estimates volatility accurately in isolation, but whether its conditional state contains incremental information after historical volatility and conventional stock characteristics are already available to XGBoost.

2.3 Backtest Discipline, Dynamic Universes, and Replication

The empirical setting requires careful treatment of sampling and backtest bias. Survivorship and delisting-related selection can overstate performance [8]. Machine-learning backtests also face repeated-trial and model-selection risks, motivating chronological validation, transaction costs, and auditable execution assumptions [9][14]. Replication evidence further cautions that predictors often weaken under stricter samples [15].

2.4 Hypotheses

H1 predicts that M2 improves ten-day out-of-sample discrimination or ranking relative to M1, as reflected in AUC, IC, rank IC, or the top-minus-bottom return spread. H2 predicts that M2 or M3 improves net portfolio return or risk-adjusted performance under a common execution and cost convention. Nonlinearity, horizon dependence, and market-state dependence are treated as exploratory mechanisms rather than substitutes for these tests.

3. RESEARCH DESIGN

3.1 Data, Universe, and Predictors

The sample covers 2018-2025, with 2017 used only to initialize rolling features. Monthly historical CSI 300 constituent snapshots, equity prices, valuation data, index returns, and Shenwan industries are frozen locally. Features are computed for the 553-stock historical union before date-level membership filtering, yielding 574,475 observations across 1,923 trading days.

Conventional predictors include size, valuation ratios, turnover, momentum, reversal, historical volatility, and industry indicators. Missing values are filled within date-industry and then date; variables are winsorized and date-standardized. Stock-level GARCH volatility is estimated with a 252-day rolling window, at least 200 observations, normal innovations, and 20-day refitting; the value dated t uses only returns through t-1.

$$Z_{i,t} = \frac{W(X_{i,t}) - \mu_t}{\sigma_t} \tag{1}$$

Equations (1)-(3) summarize date-local standardization and rolling GARCH (1,1), and Table 1 reports the principal empirical design. The transformations keep inputs comparable across dates while preventing future information from entering the feature set.

$$r_{i,t} = \mu_i + \varepsilon_{i,t}, \quad \varepsilon_{i,t} = \sqrt{h_{i,t}} \cdot z_{i,t}, \quad z_{i,t} \sim N(0,1) \tag{2}$$

$$h_{i,t} = \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{i,t-1} \tag{3}$$

Table 1. Principal Empirical Design.

Component	Specification
Universe	Historical dynamic CSI 300; 553-stock union
Panel	574,475 stock-days; 2018-01-29 to 2025-12-31
Target	Top/bottom 30% of ten-day forward return
Split	2018-2021 train; 2022 validation; 2023-2025 test
Models	M1 conventional; M2 + GARCH; M3 + GARCH + market
Execution	Signal after close; next-open entry; ten-day holds
Cost	20 bps one-way; 0/10/50 bps sensitivity

The dynamic-universe design is central: entries and exits are respected before feature construction and ranking, so the test does not condition on later index survival. Omitting disappearing or excluded securities ex post can bias prediction and portfolio performance upward [8].

3.2 Labels, Models, and Validation

Within each date, stocks in the top and bottom 30% of ten-day forward returns define binary labels; the middle 40% is excluded from fitting but retained for ranking and portfolios. Data are split chronologically with horizon embargoes. Hyperparameters are chosen by validation AUC across M1-M3, and final ten-day models use 300 trees, depth three, learning rate 0.05, 0.8 row and feature subsampling, L1 penalty 0.1, L2 penalty 1.0, and seed 42.

$$R_{i,t}^{(\tau)} = \frac{P_{i,t+\tau}}{P_{i,t}} - 1, \quad \tau \in \{10,20,60\} \tag{4}$$

$$Y_{i,t} = \begin{cases} 1, R_{i,t}^{(10)} \geq Q_{0.70,t}, \\ 0, R_{i,t}^{(10)} \leq Q_{0.30,t}. \end{cases} \tag{5}$$

Equations (4) and (5) define forward returns and binary training labels: $P_{i,t}$ is the adjusted price, τ is the forward holding horizon, $R_{i,t}^{(\tau)}$ is the tau-day forward return, $Q_{0.70,t}$ and $Q_{0.30,t}$ are same-date return quantiles, and $Y_{i,t}$ is used only when a stock belongs to the top or bottom 30% of the cross-section.

$$\hat{p}_{i,t} = sigmoid\left(\sum\nolimits_k f_k\left(x_{i,t}\right)\right) \tag{6}$$

Equation (6) summarizes the XGBoost classifier, where $x_{i,t}$ is the feature vector, $f_k(.)$ is the kth regression tree, and sigmoid(.) is the logistic link. The predicted probability $\hat{p}_{i,t}$ ranks the full test cross-section before portfolio formation.

3.3 Backtest and Inference

Every ten prediction trading days, each model forms equal-weight top-10% and top-20% portfolios. Signals are observed after the close, executed at the next open, and held for ten trading days, producing 72 non-overlapping periods. Costs apply to total traded notional. Predictive metrics use Newey-West HAC standard errors with lag 10 [7]; portfolio increments use paired holding-period and HAC tests, and Sharpe differences use a paired moving-block bootstrap.

$$\text{Turnover}_t = 0.5 \times \sum\nolimits_i \left|w_{i,t} - w_{i,t-1}\right| \tag{7}$$

$$r_t^{\text{net}} = r_t^{\text{gross}} - c \cdot N_t \tag{8}$$

$$\text{Sharpe} = \frac{\text{mean}(r_t^{\text{net}})}{\text{sd}(r_t^{\text{net}})} \times \sqrt{25.2} \tag{9}$$

Equations (7)-(9) define the trading-cost and risk-adjusted-return measures. Here $w_{i,t}$ is portfolio weight, c is the one-way cost rate, N_t is total traded notional including entry and liquidation, and 25.2 annualizes non-overlapping ten-trading-day holding returns.

3.4 Extended Mechanism, Stability, and Implementation Tests

Supplementary analyses test whether the bounded result depends on market state, random seed, feature composition, or implementation. Market states use signal-date information only: 20-day CSI 300 volatility and 60-day trailing trend. Daily state differences use HAC (10), and portfolio differences use the 72 holding periods.

Stability checks repeat seeds 42, 7, 2024, 2025, and 2026 and perturb tree count, learning rate, and depth. Ablations remove historical volatility, GARCH volatility, or the CSI 300 return from M3 while retaining the same chronological split.

Implementation checks vary portfolio width, rebalancing frequency, and costs. Concentration is summarized by Herfindahl indices, and industry exposure is tested through within-industry ranking and a 30% single-industry cap.

$$\mathrm{HHI}_t = \sum_g s_{g,t}^2, \quad N_t^{\mathrm{eff}} = \frac{1}{\mathrm{HHI}_t} \tag{10}$$

In (10), $s_{g,t}$ is the portfolio weight of industry g on date t. For equal-weight stocks, the same expression measures stock concentration. A smaller HHI and larger effective count indicate broader diversification.

3.5 Additional P2 Robustness Design

Two P2 checks keep the main split and trading rule unchanged. A low-complexity L2 logistic-regression benchmark tests whether the GARCH signal depends entirely on XGBoost's nonlinear structure. A second check replaces GARCH with lagged 252-day downside volatility, processed with the same filling, winsorization, and z-scoring rules.

4. EMPIRICAL RESULTS

4.1 Descriptive Evidence

The mean daily cross-section contains 298.74 stocks. Missingness is low, and preprocessing produces complete model inputs. GARCH volatility is strongly but imperfectly correlated with 20- and 60-day historical volatility (0.786 and 0.829) and has pooled within-stock lag-one correlation of 0.941, so it is best interpreted as a related persistent risk state.

Table 2. Principal Ten-Day Results.

Model	AUC	IC	RankIC	Ann. return	Sharpe	MDD
M1	0.531	0.025	0.043	10.93%	0.696	-18.16%
M2	0.534	0.028	0.047	12.14%	0.770	-18.06%
M3	0.534	0.029	0.047	11.32%	0.732	-17.72%

4.2 Prediction and Portfolio Performance

Table 2 reports that M1, M2, and M3 attain ten-day AUC values of 0.531, 0.534, and 0.534 and rank IC values of 0.043, 0.047, and 0.047. M2 has the largest top-minus-bottom decile spread (0.366%), while M3 has the highest IC (0.029). At the model level, all AUC means exceed 0.5 and all rank IC means exceed zero at the 1% level under HAC inference; only M3's IC is significant at 5%, and no model's return spread is significant.

At a 20-bps one-way cost, the top-decile M1, M2, and M3 portfolios earn 10.93%, 12.14%, and 11.32% annualized net returns against a 6.02% benchmark. M2 has the highest net Sharpe ratio (0.770), whereas M3 has the smallest maximum drawdown (-17.72%). The top-quintile results are lower and closer together: annualized returns are 7.73%, 8.21%, and 8.18%. Figure 1 plots the corresponding net asset values and drawdowns. These are favorable economic point estimates, but they must be assessed with paired incremental tests.

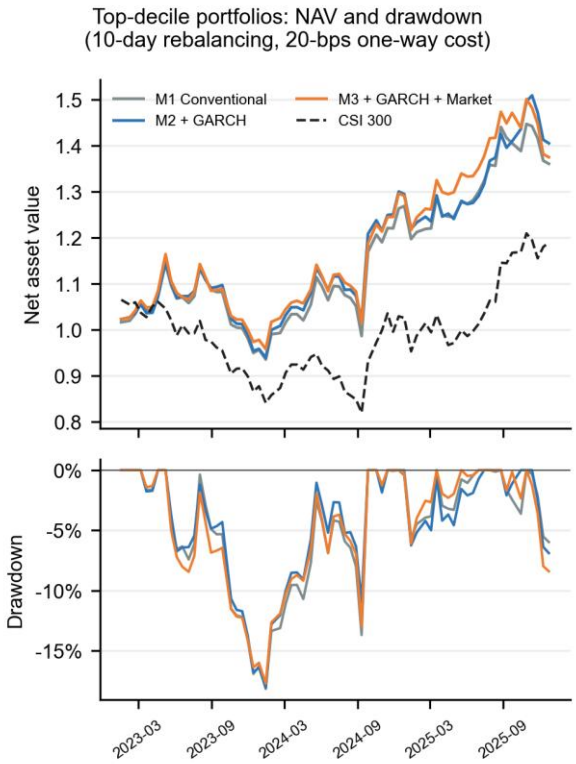

Figure 1. Net asset values and drawdowns of the top-decile portfolios.

4.3 Incremental Statistical Evidence

Table 3. Paired Holding-Period Incremental Estimates And P-Values.

Metric	M2-M1 increment (p)	M3-M1 increment (p)
AUC	0.0019 (0.321)	0.0028 (0.156)
IC	0.0025 (0.306)	0.0036 (0.128)
RankIC	0.0032 (0.265)	0.0035 (0.195)
Top-Bottom10	0.0010 (0.175)	0.0007 (0.408)
Top10 ann. return (arith.)	1.08 pp (0.589)	0.30 pp (0.842)
Top10 Sharpe	0.075 (0.564)	0.036 (0.672)

Notes: Return increments in Table 3 are paired non-overlapping ten-trading-day holding-period differences multiplied by 25.2. They are arithmetic annualizations of paired period differences, not differences between the CAGR-style annualized returns reported in Table 2.

As shown in Table 3, all ten-day M2-M1 predictive increments are positive, but p-values for AUC, IC, rank IC, and Top-Bottom10 remain above conventional thresholds. M3-M1 results are similar, and all 95% confidence intervals cross zero. Top-decile return and Sharpe increments are also insignificant, so H1 and H2 are not supported despite favorable point estimates.

4.4 Interpretation and Robustness

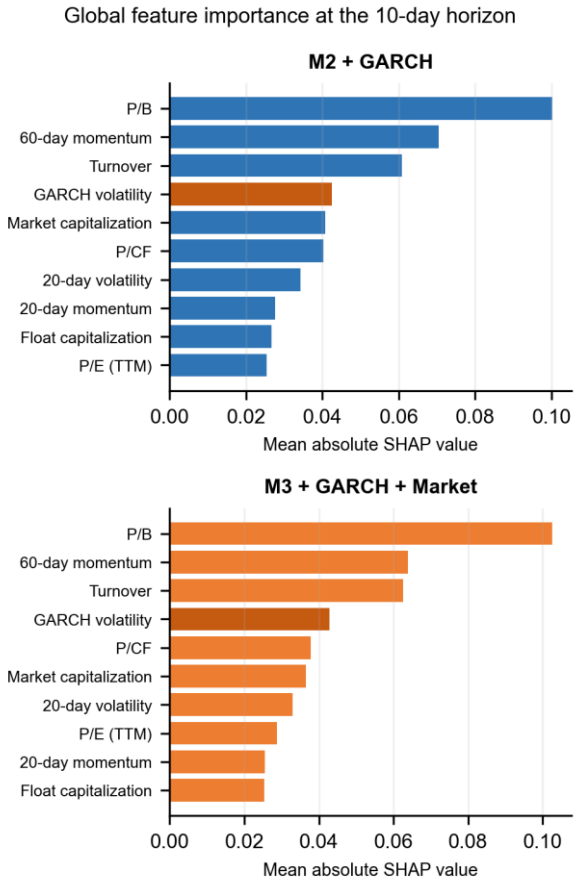

Figure 2. Global SHAP importance in the augmented ten-day models.

As shown in Figure 2, GARCH volatility ranks fourth among reported non-industry features in M2 and M3, with normalized SHAP shares of 7.22% and 7.41%. Figure 3 shows that low relative GARCH volatility contributes positively while high volatility contributes negatively. SHAP confirms model use of GARCH as a risk-state variable, but importance is not evidence of causal or statistically significant incremental value [6].

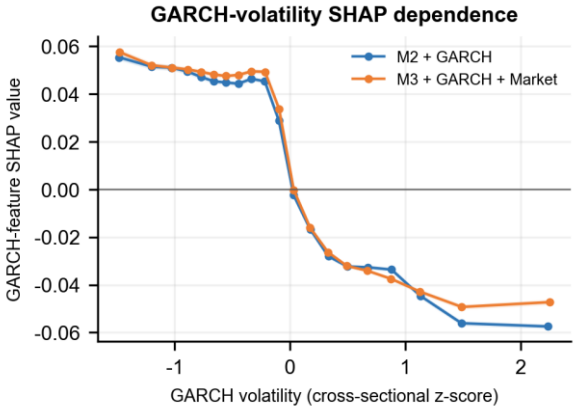

Figure 3. Binned SHAP dependence of standardized GARCH volatility.

Table 4. Robustness Summary.

Design	Metric	M1	M2	M3
20-day prediction	AUC	0.535	0.531	0.533
60-day prediction	AUC	0.505	0.508	0.509
CSI 500 Top10	Ann. return	7.69%	8.78%	8.27%
CSI 300 Top10, 50 bps	Ann. return	6.55%	7.74%	7.28%

Table 4 summarizes robustness tests that reinforce the bounded interpretation. At 20 days, M1 leads on AUC, IC, and rank IC; at 60 days, prediction is weak. In the dynamic CSI 500 universe and at 50-bps costs, augmented models retain some favorable point estimates, but economic margins narrow and increments remain insignificant.

4.5 P1 Extended Evidence

4.5.1 Market-State Heterogeneity

Table 5. Exploratory Market-State Heterogeneity of M3-M1.

State type	State	IC Δ	p	Top10 return Δ	p
volatility	Low	0.0051	0.052	-0.12 pp	0.945
volatility	High	-0.0016	0.747	1.93 pp	0.707
trend	Bear	0.0039	0.270	2.85 pp	0.121
trend	Bull	0.0032	0.227	-2.39 pp	0.279

Notes: State variables use signal-date information only. Return differences are arithmetic annualizations. The analysis is exploratory and p-values are not adjusted for multiple comparisons.

Table 5 reports market-state results that are suggestive only. M3's IC increment is larger in low-volatility markets, while Top10 return increments are more favorable in weak-trend periods; neither state contrast is statistically established. Figure 4 visualizes the IC and rank-IC increments across these ex ante states.

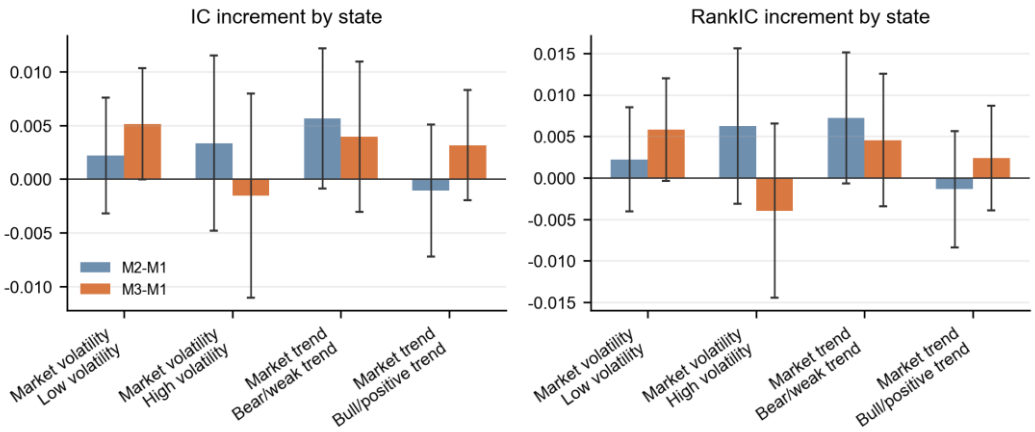

Figure 4. IC and rank-IC increments across ex ante market states.

4.5.2 Random-Seed Stability and Feature Ablation

Table 6. Stability Across Five Random Seeds, Mean (SD).

Model	AUC	IC	RankIC	Top10 ann. return
M1	0.53122 (0.00093)	0.02761 (0.00203)	0.04352 (0.00165)	12.00% (0.91 pp)
M2	0.53250 (0.00117)	0.02875 (0.00163)	0.04512 (0.00160)	12.71% (1.36 pp)
M3	0.53360 (0.00086)	0.02941 (0.00100)	0.04671 (0.00130)	12.27% (1.27 pp)

Notes: The five seeds are computational repetitions, not independent economic samples.

Table 6 shows that prediction rankings are stable across seeds, with mean Spearman correlation near 0.96 and Top10 holding Jaccard similarity around 0.75-0.80. Figure 5 plots the seed-level test metrics and Top10 returns. Portfolio-return increments vary more, showing that small ranking changes near the cutoff can materially affect realized returns.

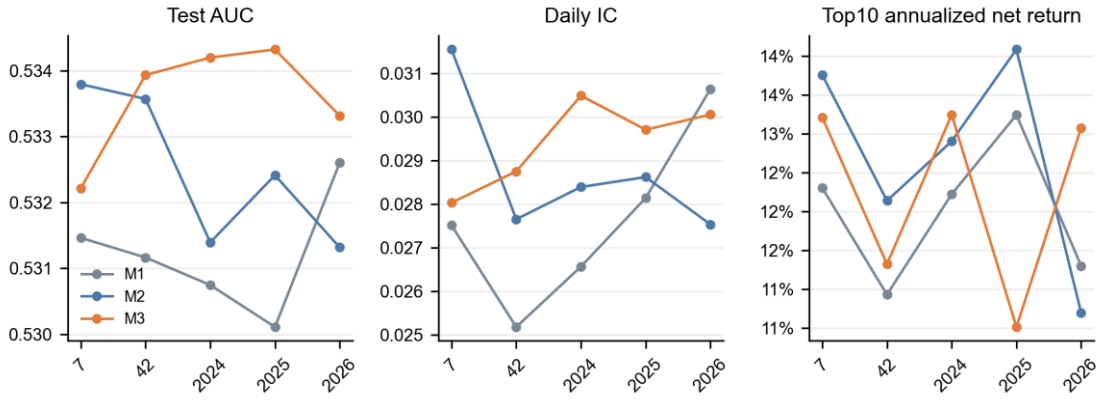

Figure 5. Test metrics and Top10 returns across random seeds.

Table 7. Deletion-Based Feature Ablation.

Variant	AUC	IC	RankIC	Top10 return	Sharpe
Full M3	0.5339	0.0288	0.0469	11.32%	0.732
No historical volatility	0.5315	0.0239	0.0430	11.71%	0.754
No GARCH	0.5311	0.0264	0.0430	13.45%	0.825
No market return	0.5336	0.0277	0.0466	12.14%	0.770

Table 7 shows that historical volatility, GARCH, and market return each contribute to fitted prediction, while Figure 6 summarizes the corresponding prediction and portfolio effects. Removing GARCH or the market factor can raise Top10 annualized return, so predictive contribution does not translate monotonically into economic gain.

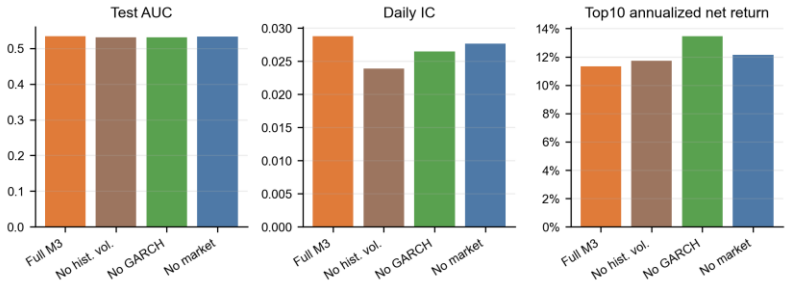

Figure 6. Prediction and portfolio effects of deleting volatility components.

4.5.3 Portfolio Implementation Robustness

Table 8. Portfolio Width, Performance, And Concentration.

Portfolio	Model	Stocks	Ann. return	Sharpe	Turnover	Largest industry
Top5%	M1	14.8	14.52%	0.872	31.8%	57.8%
Top5%	M2	14.8	16.04%	0.958	30.1%	59.2%
Top5%	M3	14.8	15.43%	0.931	31.1%	63.8%
Top10%	M1	29.8	10.93%	0.696	27.8%	41.7%
Top10%	M2	29.8	12.14%	0.770	27.6%	41.1%
Top10%	M3	29.8	11.32%	0.732	25.5%	44.8%
Top20%	M1	59.8	7.73%	0.546	26.6%	26.5%
Top20%	M2	59.8	8.21%	0.578	26.5%	26.5%
Top20%	M3	59.8	8.18%	0.585	26.9%	27.1%
Top30%	M1	89.8	8.19%	0.594	26.6%	19.8%
Top30%	M2	89.8	7.47%	0.559	27.0%	19.7%
Top30%	M3	89.8	7.96%	0.593	27.0%	20.0%

Notes: Ten-day rebalancing and 20-bps one-way cost. Largest industry is the time-series average of each signal date's single largest industry weight, computed with point-in-time Shenwan industry history; it is not the maximum of average industry weights.

Table 8 shows that Top5 portfolios have the highest returns but only about 15 stocks and heavy industry concentration. Top10 reduces concentration while preserving positive augmented-model point estimates; Top30 and month-end implementations remove the advantage. Cost increases lower all returns, though M2 remains highest at Top10. Figure 7 visualizes the portfolio-width tradeoff between annualized return and industry concentration.

Table 9. Industry-Exposure Control Backtest.

Strategy	Model	Ann. return	Sharpe	Max DD	Largest ind.	Eff. ind.
Original	M1	10.93%	0.696	-18.2%	41.7%	4.2
Original	M2	12.14%	0.770	-18.1%	41.1%	4.2
Original	M3	11.32%	0.732	-17.7%	44.8%	3.9
Within-ind.	M1	5.24%	0.382	-23.4%	8.6%	24.5
Within-ind.	M2	3.24%	0.274	-24.6%	8.6%	24.5
Within-ind.	M3	5.72%	0.417	-23.5%	8.6%	24.5
30% cap	M1	6.55%	0.451	-20.8%	29.2%	5.7
30% cap	M2	7.98%	0.541	-18.8%	28.9%	5.6
30% cap	M3	7.35%	0.493	-21.8%	29.2%	5.7

Notes: Top10 follows the project's top-decile convention. Largest ind. uses the same time-series average of daily largest industry weight as Table 8. Within-industry portfolios select the top 10% inside each industry; the constrained portfolio caps any one industry at 30%. Ten-day rebalancing and 20-bps one-way cost are retained.

Table 9 reports the industry-control backtests and confirms that raw returns contain substantial sector exposure. Within-industry ranking sharply reduces concentration but lowers annualized returns, while a 30% cap leaves M2 and M3 above M1. The economic ranking therefore depends on portfolio construction rather than a universal trading edge.

Top10 and Top20 backtests contain no missing entry opens, and prior-close exit valuation is needed for fewer than 0.05% of selected positions. Exact A-share price-limit execution is not claimed because the daily data lack ST status, board-specific limits, queues, and intraday fills.

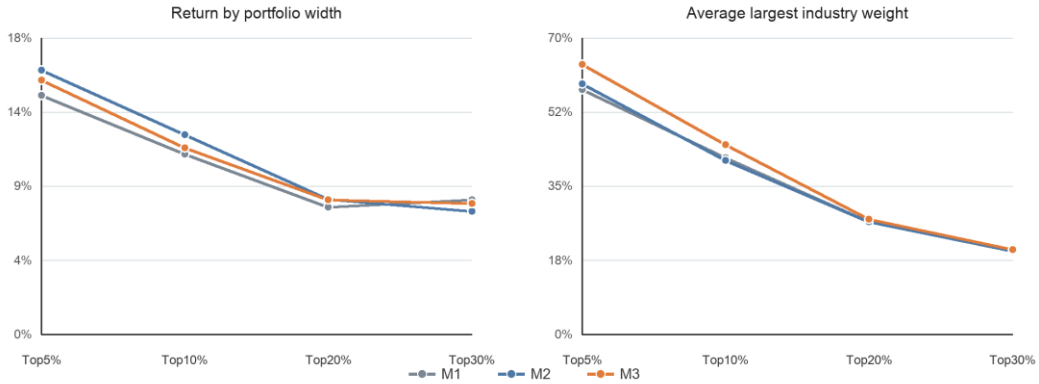

Figure 7. Portfolio width, annualized return, and industry concentration.

4.6 P2 Additional Robustness Evidence

Table 10. Xgboost Versus Logistic Baseline.

Algorithm	Model	AUC	IC	RankIC	Top10 return	Sharpe	MDD
Logit	M1	0.521	0.011	0.030	-2.74%	-0.093	-26.5%
Logit	M2	0.524	0.014	0.035	-0.77%	0.028	-23.6%
Logit	M3	0.524	0.014	0.035	-0.77%	0.028	-23.6%
XGBoost	M1	0.531	0.025	0.043	10.93%	0.696	-18.2%
XGBoost	M2	0.534	0.028	0.047	12.14%	0.770	-18.1%
XGBoost	M3	0.534	0.029	0.047	11.32%	0.732	-17.7%

Notes: Both algorithms use the same dynamic CSI 300 panel, M1/M2/M3 definitions, ten-day label, train-validation-test split, and strict Top10 backtest convention.

Table 10 shows that the logistic baseline preserves a weak directional GARCH pattern: AUC and IC rise from M1 to M2/M3, but Top10 returns remain far below XGBoost and return increments are insignificant. Thus, the signal is not exclusive to trees, while economic conversion still depends on model form.

Table 11. Garch Versus Downside-Volatility Specification.

Specification	AUC	IC	RankIC	Top10 return	Sharpe	MDD
M1 conventional	0.531	0.025	0.043	10.93%	0.696	-18.2%
M2 GARCH	0.534	0.028	0.047	12.14%	0.770	-18.1%
M3 GARCH + market	0.534	0.029	0.047	11.32%	0.732	-17.7%
M2 downside-vol	0.533	0.027	0.046	11.63%	0.739	-19.1%
M3 downside-vol + market	0.532	0.025	0.044	10.33%	0.675	-18.9%

Notes: The downside-volatility variants replace GARCH volatility with a strictly lagged 252-day downside-volatility factor. All other split, label, XGBoost, and backtest settings are held fixed.

Table 11 reports that the downside-volatility alternative produces AUC values close to the main GARCH models, and its M2 Top10 return remains above M1. However, the return increments are small and insignificant. Volatility-state information is therefore not an artifact of one estimator, but alternative definitions do not establish a stronger edge.

5. DISCUSSION

The evidence is consistent with conditional volatility acting as a short-lived state variable rather than a general return predictor. GARCH updates recent shocks and persistent variance, so its marginal content is most plausible at the ten-day horizon. SHAP patterns suggest that the model favors low relative conditional volatility and penalizes high volatility after controlling for other characteristics.

Three distinctions limit the claim. Significant AUC or rank IC for a model does not imply a significant nested-model increment; SHAP importance explains fitted predictions rather than independent economic reliability; and small backtest differences share market exposure and correlated holdings. The results therefore fit the broader machine-learning asset-pricing literature while showing that an intuitive feature can be used without producing a stable advantage.

P1 checks sharpen this distinction. GARCH and market information generate small ranking improvements that persist across seeds and decline under deletion, yet portfolio returns vary because changes near the Top10 cutoff alter holdings.

Market-state and implementation results also reject a universal mechanism. IC gains appear more visible in low-volatility states, while favorable portfolio differences appear more in weak trends; industry controls, Top30 portfolios, and month-end rebalancing weaken the advantage. The defensible interpretation is conditional use of volatility information under transparent implementation.

The literature supports this caution. GARCH (1,1) remains a natural volatility benchmark [12], but evaluation depends on imperfect proxies and loss functions [13]. Backtest-overfitting and factor-replication studies likewise warn against treating selected return improvements as stable discoveries [14][15].

Several limitations remain. Constituent membership is monthly rather than daily, valuation data have limited frequency, and the 2023-2025 test period yields only 72 non-overlapping ten-day portfolio observations. Exact price-limit execution cannot be reconstructed from daily data. Robustness repetitions help triangulate the result, but future work should use longer point-in-time samples, daily constituent events, limit-aware execution, and additional volatility specifications before claiming a stable premium.

6. CONCLUSION

This study evaluates rolling stock-level GARCH volatility in an XGBoost selection model using a leakage-audited dynamic CSI 300 panel. Ten-day GARCH models improve several point estimates, and the signal remains visible across seeds, ablations, a logistic baseline, and a downside-volatility alternative. Nevertheless, paired tests do not establish significant improvement over M1, and gains vary with implementation and exposure control. Conditional-volatility information therefore has marginal, short-horizon predictive value, but the evidence does not prove a robust trading advantage.

REFERENCES

[1] Engle, R. F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4): 987-1007. DOI: 10.2307/1912773.
[2] Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3): 307-327. DOI: 10.1016/0304-4076(86)90063-1.
[3] Chen, T., and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785-794. DOI: 10.1145/2939672.2939785.
[4] Gu, S., Kelly, B., and Xiu, D. (2020) Empirical asset pricing via machine learning. Review of Financial Studies, 33(5): 2223-2273. DOI: 10.1093/rfs/hhaa009.
[5] Gao, J., Guo, H., and Xu, X. (2022) Multifactor stock selection strategy based on machine learning: Evidence from China. Complexity, 2022: 7447229. DOI: 10.1155/2022/7447229.

[6] Lundberg, S. M., and Lee, S.-I. (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, 30.
[7] Newey, W. K., and West, K. D. (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3): 703-708.
[8] Shumway, T. (1997) The delisting bias in CRSP data. The Journal of Finance, 52(1): 327-340. DOI: 10.1111/j.1540-6261. 1997. Tb 03818.x.
[9] Arnott, R., Harvey, C. R., and Markowitz, H. (2019) A backtesting protocol in the era of machine learning. The Journal of Financial Data Science, 1(1): 64-74. DOI: 10.3905/jfds.2019.1.064.
[10] Kelly, B., Malamud, S., and Zhou, K. (2024) The virtue of complexity in return prediction. The Journal of Finance, 79(1): 459-503. DOI: 10.1111/jofi.13298.
[11] Bryzgalova, S., Pelger, M., and Zhu, J. (2025) Forest through the trees: Building cross-sections of stock returns. The Journal of Finance, 80(5): 2447-2506. DOI: 10.1111/jofi.13477.
[12] Hansen, P. R., and Lunde, A. (2005) A forecast comparison of volatility models: Does anything beat a GARCH (1,1)? Journal of Applied Econometrics, 20(7): 873-889. DOI: 10.1002/jae.800.
[13] Patton, A. J. (2011) Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics, 160(1): 246-256. DOI: 10.1016/j.jeconom.2010.03.034.
[14] Bailey, D., Borwein, J., Lopez de Prado, M., and Zhu, Q. J. (2016) The probability of backtest overfitting. The Journal of Computational Finance. DOI: 10.21314/jcf.2016.322.
[15] Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023) Is there a replication crisis in finance? The Journal of Finance, 78(5): 2465-2518. DOI: 10.1111/jofi.13249.