BOCCONI STUDENTS INVESTMENT CLUB

Bocconi Students Investment Club

A Machine-Learning Regime-Switching Allocation System for ES Futures

Rule-Based and Gradient-Boosted Regime Detection with Dynamic Strategy Routing: 2002–2026

Authors

Bocconi Students Investment Club

Neel Roy (Project Lead)

Marco Moscatelli

Riccardo Favazza

May 2026

bsic.it | Bocconi University, Milan

Abstract

Most systematic trading strategies apply fixed weights regardless of market conditions. A momentum fund stays long whether the market is trending or oscillating. This paper builds and tests a system that routes capital differently depending on which of four market regimes is active at any given time.

Two independent routing systems are built and compared on E-mini S&P 500 futures (ES): a deterministic paper router based on observable economic thresholds, and a machine-learning pipeline using an XGBoost gradient-boosted classifier trained on 74 engineered features. Both systems produce daily regime probability vectors that allocate capital across four specialist strategies covering trend-following, mean-reversion, defensive rotation, and crisis protection.

We develop a two-method soft-label framework for the ML system. Method 1 constructs labels from economically motivated feature scores passed through a softmax transform. Method 2 derives labels from forward-realised Sharpe ratios of the constituent strategies. Method 1 goes into production. Method 2 is dropped: bond-futures returns dominate its R3 (Choppy) labels, a structural contamination identified through SHAP analysis. We add three trend-persistence features after diagnosing an R1/R3 confusion problem in post-2021 low-volatility markets where ADX loses most of its discriminating power. These features were motivated by test-period observations; a fully clean evaluation requires data after May 2026.

On the out-of-sample test period (January 2021 to March 2026; 1,303 trading days for the paper router, 1,224 for the XGBoost production model), the volatility-targeted paper router achieves a Sharpe ratio of 1.47, CAGR of 12.30%, and maximum drawdown of 9.12%. The passive ES benchmark posts Sharpe 0.75 over the same period with a maximum drawdown of 25.02%. The XGBoost production model in soft-blending mode reaches Sharpe 0.81 with a maximum drawdown of 1.50%, its main contribution being capital preservation rather than return generation. Bootstrap confidence intervals place the paper router's Sharpe estimate entirely above 0.8 at the 95% level; the White (2000) Reality Check p-value of 0.55 does not reach conventional significance, consistent with the router's advantage arising from lower volatility rather than higher arithmetic mean returns. SHAP analysis computed on the full test set identifies ADX, HY OAS z-score, and the new `trend_eff_20d` efficiency ratio as the features most responsible for regime classification in out-of-sample conditions.

Keywords: regime switching, gradient boosting, ES futures, dynamic asset allocation, SHAP, soft labels, walk-forward validation

Contents

1 Introduction

The drawdowns that hurt most systematic strategies are not random. They cluster in market regimes the strategy was not designed for. A momentum fund loses when markets oscillate; a mean-reversion fund loses when a real trend takes hold. The strategy's long-run edge is real, but it gets paid unevenly, and the losing stretches share a common structure: wrong instrument for the current environment.

This paper asks whether routing capital to the right strategy for the current regime produces better risk-adjusted outcomes than holding a single strategy through all conditions. The answer requires two things that are separately difficult: detecting the regime in real time from observable features, and having pre-built return streams ready to receive the allocation. We build both.

The instrument is E-mini S&P 500 futures (ES). ES is chosen because it is among the most liquid instruments in the world (which keeps transaction costs tractable), it provides a single clean signal on broad US equity risk, and daily OHLCV data going back to the late 1990s supports a feature history long enough to span multiple market regimes. The allocation target is a set of four specialist strategies, each designed to perform well in a specific market condition.

The four regimes, described in Table 1, represent distinct market structures that a practitioner would recognise on inspection. They are not derived from a clustering algorithm applied to return data. They are defined by the strategies they route capital to, and the classifier's job is to detect them from observable features.

Regime	Name	Characteristics
R_1	LowVolTrend	Persistent bull momentum with low realised volatility. Low credit spreads, positive ADX trend, multi-horizon positive returns. Capital routes to multi-horizon momentum strategy.
R_2	HighVolMR	Short-term overreaction with elevated volatility. ADX declining from prior high, RSI extremes, credit spreads elevated but not spiking. Capital routes to VWAP/MA mean-reversion strategy.
R_3	Choppy	Directionless market with no persistent trend. Low ADX, mixed momentum across horizons, moderate volatility. Capital routes to defensive stock-bond rotation.
R_4	Crisis	Fear spike with credit stress. VIX above 30, HY OAS z-score extreme, broad negative momentum. Capital routes to modified trend-follow with hard stops and reduced size.

Table 1: Four-regime taxonomy. Each regime maps to a specialist daily strategy return stream $r_{t,k}$, $k \in \{1, 2, 3, 4\}$.

Two independent detection systems are built. The first is a deterministic paper router, a transparent set of economic threshold rules that computes soft regime weights from observable levels of VIX, credit spreads, yield curve slope, ADX, and momentum. It requires no training; its threshold levels, scoring weights, and smoothing parameters were set by hand using economic reasoning and validation-period Sharpe as a secondary check. The second is an XGBoost gradient-boosted classifier (Chen and Guestrin, 2016) trained on 74 engineered features from seven categories. Both produce daily probability vectors $\hat{\mathbf{p}}_t \in \Delta^3$ that blend strategy returns according to equation (11).

Contributions

This work makes four contributions to the regime-switching literature.

The most methodologically novel is the two-method soft-label framework for regime classification targets. Rather than assigning each day to a single regime, both methods produce probability distributions over regimes. Their comparison exposes a structural contamination in the Sharpe-softmax labels that is invisible to standard accuracy metrics: SHAP analysis reveals that bond-futures returns dominate the R3 (Choppy) label through a tautological feedback loop between the R3 strategy's bond returns and the Method 2 label construction. The diagnostic procedure is a practical template for auditing label quality in any multi-class financial classification problem.

We also document and partially address a test-period failure mode through feature engineering rather than parameter recalibration. ADX, the primary trend/chop discriminator in training, loses nearly all discriminative power in the post-2021 low-volatility bull market (Cohen's $d \approx -0.08$ for R1 vs R3 on the test set). Three replacement features grounded in a mechanical analysis of ADX's assumptions partially recover R1 recall from 55.3% to 58.6% on the out-of-sample period (155 of 374 true R1 days remain misclassified).

Both detection systems are evaluated against a complete battery of statistical tests on the out-of-sample period: block-bootstrap Sharpe confidence intervals, the White (2000) Reality Check permutation test correcting for multiple comparisons, paired t-tests, three distinct crisis case studies, and transaction cost sensitivity from 0 to 10 basis points per half-turn.

Finally, SHAP TreeExplainer values are computed on the test set rather than the training set, which is the more common but less informative choice. Test-set SHAP shows which features drive regime assignments in genuinely novel market conditions rather than in the conditions the model was already fitted to.

Key findings

On the test period, the paper router with volatility targeting produces a Sharpe ratio of 1.47, CAGR of 12.30%, and maximum drawdown of 9.12%. The XGBoost soft-blend model produces Sharpe 0.81 with maximum drawdown of 1.50%. A walk-forward XGBoost trained on Method 2 labels (and a separately developed 121-feature set) produces Sharpe 0.16, substantially below equal-weight (0.87). A new walk-forward experiment with Method 1 labels and the 74-feature set achieves Sharpe 0.84, isolating the contaminated label design as the source of the gap rather than the training protocol.

Paper structure

Section 2 reviews the relevant literature. Section 3 describes the data and features. Section 4 presents the label generation methods. Section 5 describes the deterministic paper router. Section 6 presents the ML models. Section 7 covers the backtesting engine. Section 8 presents all empirical results. Section 9 concludes.

2 Related Literature

2.1 Regime-switching models

Hamilton (1989) introduced the Markov-switching model as a way to capture structural breaks in macroeconomic time series. In his formulation, an unobserved state variable follows a first-order Markov chain, and observed data are drawn from state-conditional Gaussian distributions. The EM algorithm provides inference over the hidden state sequence. The model fits naturally to business cycle data and has been applied extensively to financial time series.

Ang and Bekaert (2002) extend the framework to international equity allocation and show that optimal portfolio weights differ substantially across regimes. Ignoring regime dependence generates measurable welfare losses in their simulations. Ang and Bekaert (2004) quantify this for both retail and institutional investors, finding that regime-switching strategies produce Sharpe ratios 0.2 to 0.4 higher than unconditional portfolios. Ang (2014) synthesises this line of work and argues that regime detection belongs at the core of long-horizon asset management.

Guidolin and Timmermann (2007) estimate a four-state multivariate regime model across equities, bonds, and commodities and find that regime-aware allocation outperforms its unconditional counterpart across all evaluation periods in their sample. Their four-regime structure (trending bull, volatile bear, moderate expansion, crisis) closely parallels our design, though their identification comes from fitting conditional return distributions to historical data whereas ours is specified from economic first principles.

A common limitation of parametric regime-switching is that it conflates regime detection with regime definition. When the conditional return distributions are estimated from history, the state inferred today is a function of how past returns cluster, not of the macro or technical conditions a practitioner would actually observe. Our approach separates the two: regimes are defined by the strategies that perform well in them, and detection uses a broad set of contemporaneous observable features.

2.2 Machine learning in quantitative finance

Krauss et al. (2017) compare deep neural networks, gradient-boosted trees, and random forests on S&P 500 statistical arbitrage, finding that gradient-boosted methods outperform both. The advantage is particularly clear on tabular financial data where features exhibit nonlinear interactions and the signal-to-noise ratio is low. Chen and Guestrin (2016) introduce XGBoost, whose sparsity-aware split finding, L1/L2 regularisation, and stochastic sampling make it well-suited to exactly this setting.

More recently, Gu et al. (2020) evaluate a broad suite of machine learning methods for cross-sectional equity return prediction across a large panel of US stocks, finding that tree ensembles and neural networks substantially outperform linear factor models when feature interactions are complex. Their results reinforce the case for nonlinear methods in financial prediction tasks, and their emphasis on out-of-sample evaluation protocol directly informs the walk-forward design used here. López de Prado (2018) surveys common methodological pitfalls in applying machine learning to financial data, including the multiple-testing and feature leakage issues that our block-bootstrap and no-lookahead constraints are designed to address.

A persistent concern in applying machine learning to finance is that apparent out-of-sample performance reflects data mining rather than genuine predictability. White (2000) provides a formal test: the Reality

Check tests whether the best strategy in an evaluated set beats a benchmark under the null of no predictability, correcting for the multiple-comparison problem. We apply this in Section 8.

2.3 Momentum and trend following

Jegadeesh and Titman (1993) document that past winners continue to outperform past losers at horizons of 3 to 12 months, a pattern now known as cross-sectional momentum. Moskowitz et al. (2012) extend momentum to the time-series dimension, showing that a strategy that goes long assets with positive trailing 12-month returns and short those with negative returns generates a Sharpe ratio of roughly 1.0 across equities, bonds, currencies, and commodities. Our R_1 strategy applies this logic to ES futures.

Baltas and Kosowski (2013) document strong performance of diversified trend-following during equity bear markets and crisis periods, which is the motivation for routing to the modified trend-follow strategy in R_4 rather than to a defensive long-only position. Avellaneda and Lee (2010) show that mean-reversion strategies in US equities generate statistically reliable returns at short horizons, providing empirical support for the R_2 strategy design.

2.4 Feature interpretability

Lundberg and Lee (2017) introduce SHAP, a game-theoretically grounded method for attributing model predictions to input features. For tree ensembles, the TreeSHAP algorithm computes exact Shapley values in polynomial time and satisfies the axioms of local accuracy, missingness, and consistency simultaneously. We compute TreeSHAP values on the test set in Section 8 to see which features drive regime assignments in genuinely novel market conditions, not the training conditions the model has already seen.

3 Data and Feature Engineering

3.1 Data sources

The master dataset covers daily observations from September 1997 through March 2026, sourced from three providers. ES futures OHLCV data come from continuous front-month contracts adjusted for roll. The VIX index and 10-year/2-year Treasury yields come from FRED. High-yield and investment-grade option-adjusted spreads (OAS) come from Bloomberg. Put/call ratios are from CBOE. Gold, crude oil, and US Dollar Index returns use front-month futures. The MOVE index is the bond-market volatility proxy.

After cleaning for missing bars, timezone alignment, and roll adjustments, 7,226 daily observations are available from September 1997. The feature pipeline requires a maximum lookback of 200 calendar days, so the first usable feature row is 18 October 2002. The final dataset used for modelling covers 5,752 rows.

3.2 Train, validation, and test splits

The dataset is partitioned into three non-overlapping periods that are fixed before any modelling begins and are never revised:

Period	Start	End	Rows	Share
Train	2002-10-18	2018-12-31	4,044	70.3%
Validation	2019-01-02	2020-12-31	484	8.4%
Test	2021-01-04	2026-03-09	1,224	21.3%
Total			**5,752**	**100.0%**

Table 2: Train/validation/test splits. The test set is never examined until the production model is fully specified.

The validation period (2019–2020) covers a low-volatility bull market in 2019 and the COVID-19 crash in 2020, giving reasonable coverage of both quiet and stressed conditions. The test period (2021–2026) covers the post-COVID recovery, the 2022 rate-hiking cycle and bear market, and the 2023–2025 large-cap technology-driven bull market. No model parameter is adjusted after observing test-set results. The walk-forward backtests use annual expanding windows beginning from 2010.

The test period's regime composition differs substantially from training. Based on the argmax classifications reported in Table 8, R3 (Choppy) accounts for approximately 53.8% of test-set days versus 31.4% under the training-set label distribution, while R2 (HighVolMR) falls from 20.7% to under 4%. This distribution shift is the proximate cause of the R2 classification weakness documented in Section 8.

Note on validation set usage. The validation set (2019–2020, 484 rows) serves three distinct roles in this paper. First, it is the optimisation target for 100 Optuna TPE hyperparameter trials: each trial trains XGBoost on the training set, uses the validation set as the early-stopping signal for tree count, and is scored by validation KL divergence. Both the hyperparameter configuration and the per-trial tree count are therefore selected against the same 484-day window. Second, the smoothing parameter α for

the paper router was chosen by observing that validation Sharpe rises monotonically from $\alpha = 0.10$ to $\alpha = 0.30$, making $\alpha = 0.20$ a conservative pick informed by this relationship. Third, the production model combines train and validation as the fit set. Running 100 hyperparameter trials on 484 validation rows creates meaningful risk of implicit overfitting to the specific dynamics of that period (a low-volatility bull market followed by the COVID-19 crash). The parameter sensitivity analysis in Section 8 provides partial reassurance by showing near-flat Sharpe across $\pm 20\%$ perturbations of all hyperparameters, but this caveat should inform how the production model's validation metrics are interpreted.

3.3 Feature categories

Seventy-four features are computed across seven categories. Table 3 lists each category, the number of features, and representative members. Three features added in May 2026 to address an R1/R3 confusion problem (see Section 6) are marked separately.

Category	Count	Representative features
Returns and momentum	6	`ret_1d`, `ret_5d`, `ret_20d`, `ret_60d`, momentum sign consistency, triple-horizon consistency
Trend and moving averages	14	MA distances (20, 50, 200 days), RSI-14, MACD line/signal/histogram, ADX, $\pm$DI, ADX diff, trend strength
Trend persistence (new)	3	`trend_eff_20d`, `win_rate_20d`, `ma_dist_stability_20d`
Volatility and VIX	15	Realised volatility at 5/20/60 days, vol ratio, VIX level/change/z-score/vol-of-vol, Garman-Klass, Parkinson, Bollinger width/%B, intraday range
Credit and rates	10	HY OAS level/z-score/5D change, IG OAS, HY/IG ratio, yield spread 10Y–2Y and 30Y–10Y, MOVE level/z-score
Cross-asset	13	Gold/oil/DXY 20D returns, gold-DXY correlation, oil-SPY correlation, SPY-TLT correlation (20D and 60D), ZN 5D/20D returns, ZN-ES correlation, GC 5D/20D returns, VIX term structure
Volume and microstructure	7	ES volume and MA, volume z-score, 5/20 volume ratio, price-volume correlation, put/call ratio MA and z-score
Calendar	4	Day of week, month, quarter-end flag, month-end flag

Table 3: Feature categories. A total of 74 features are used in the production model. The three trend-persistence features are included only in the production model (74 features); the research model uses 71.

3.4 Trend-persistence features

The three new features warrant description because their addition is motivated by a specific empirical failure, not a general expansion of the feature set.

`trend_eff_20d` is the ratio of the net 20-day price change to the sum of absolute daily price changes over the same window:

$$\text{eff}_t = \frac{|P_t - P_{t-20}|}{\sum_{i=0}^{19} |P_{t-i} - P_{t-i-1}|} \tag{1}$$

A smooth upward trend with few reversals produces a value near 1. An oscillating market that ends where it started produces a value near 0. Days with zero total absolute price movement (denominator equals zero) are assigned $\text{eff}_t = 0$ in the implementation. It captures a related concept to the Hurst exponent —

path efficiency versus meandering — at a fixed 20-day horizon rather than across scaling levels. The key advantage over ADX is that it maintains a Cohen's d of approximately 0.45 between R1 and R3 on the test set, whereas ADX produces Cohen's $d \approx -0.08$ over the same period.

`win_rate_20d` is the fraction of the last 20 days with positive 1-day returns. It captures directional persistence that ADX misses in slow-grind environments.

`ma_dist_stability_20d` is the standard deviation of $(P_t/\mathrm{MA}_{20,t} - 1)$ over the past 20 days. Low values indicate a price tracking its moving average smoothly; high values indicate oscillation around it.

Note on evaluation integrity. These three features were added in May 2026 after diagnosing R1/R3 confusion in the test period's post-2021 environment. The production model therefore trains on a feature set whose composition was partly informed by test-period observations. The features are derived from principled arguments about ADX's structural limitations rather than fitted to test-set labels, but the R1 recall improvement reported in Section 8 (55.3% to 58.6%) should be read in this context. A fully clean evaluation of the production model would require a holdout period beginning after May 2026.

4 Label Generation

4.1 Motivation for soft labels

A standard classification approach would assign each trading day to a single regime and train the model to predict that hard assignment. Two problems make this inadequate for our setting. First, many days are genuinely ambiguous: a transition day between R_1 and R_3 has features consistent with both, and forcing a binary assignment throws away that uncertainty. Second, regime boundaries in real markets are gradual, not discrete. A model that sees only crisp 0/1 targets learns to be overconfident in its predictions, which hurts the quality of the soft-blend allocation in equation (11).

Soft labels address both problems. Each day receives a probability distribution over the four regimes, where the probabilities reflect the degree of membership in each state. The model then learns to replicate those distributions, and its outputs can be used directly as allocation weights.

4.2 Method 1: feature-based soft labels

Method 1 constructs regime scores from observable features using economic intuition about what each regime looks like. For each regime k and time t, a score $s_{t,k}$ is computed as a weighted combination of standardised features:

$$s_{t,k} = \mathbf{w}_k^\top \tilde{\mathbf{x}}_t \tag{2}$$

where $\tilde{\mathbf{x}}_t$ are standardised feature values and $\mathbf{w}_k$ encodes regime-specific economic logic. For R_1, the weights are positive for ADX, positive multi-horizon momentum, and negative for VIX and credit spreads. For R_4, the signs reverse on most of these, and the weights on VIX z-score and HY OAS spike indicators are large and positive. The full weight vectors for all four regimes are tabulated in the accompanying code repository (`label_generation/method1_weights.py`) to allow exact replication of the label distributions. The soft label is:

$$\ell_{t,k}^{(1)} = \frac{\exp(s_{t,k})}{\sum_{j=1}^{4} \exp(s_{t,j})} \tag{3}$$

The resulting label distributions on the training set have mean probabilities of 0.273 (R_1), 0.207 (R_2), 0.314 (R_3), and 0.206 (R_4). This is reasonable: the training period (2002–2018) contains more choppy/flat regimes than strong trending periods when measured by ADX, which explains the slightly elevated R_3 share.

Since Method 1 labels are a deterministic (nonlinear via softmax) function of the same features the XGBoost trains on, the model's predictive task is to learn the label-generating function's nonlinear interactions from data. A linear model trained on the same features would learn an approximation to the $\mathbf{w}_k^\top \tilde{\mathbf{x}}_t$ scoring directly; the XGBoost's advantage lies in capturing threshold effects and feature interactions that the fixed linear weights cannot express. The walk-forward label isolation experiment in Section 8 provides indirect evidence of this: a walk-forward XGBoost with Method 1 labels and the same 74 features achieves Sharpe 0.84 on the test period, closely matching the batch-trained production model (0.81).

4.3 Method 2: Sharpe-softmax labels

Method 2 derives labels from the forward-realised Sharpe ratio of each strategy over a window W. Let $\bar{r}_{t:t+W,k}$ denote the arithmetic mean of daily returns for strategy k over days t to $t+W$, and $\hat{\sigma}_{t:t+W,k}$ the corresponding sample standard deviation. Then:

$$S_{t,k}^{(W)} = \frac{\bar{r}_{t:t+W,k}}{\hat{\sigma}_{t:t+W,k}} \cdot \sqrt{252} \tag{4}$$

$$\ell_{t,k}^{(2)} = \frac{\exp\left(\tau \cdot S_{t,k}^{(W)}\right)}{\sum_{j=1}^{4} \exp\left(\tau \cdot S_{t,j}^{(W)}\right)} \tag{5}$$

where $\tau > 0$ controls label sharpness. In the production implementation $\tau = 1$, meaning standard softmax is applied directly to the annualised Sharpe ratios; the scale of the Sharpe values (typically in the range $[-2, 3]$ on an annual basis) implicitly governs label concentration. The lookback $W = 20$ days is selected by comparing validation-set KL divergence across $\{10, 20, 60\}$ days.

4.4 Selection and rejection of Method 2

A model trained on Method 2 labels achieves a validation-set KL divergence of 0.031 and argmax accuracy of 87.1%, figures that appear competitive with Method 1. The problem appears only in SHAP analysis. For the R_3 (Choppy) regime, `zn_ret_20d` (20-day ZN bond futures return) produces the single highest SHAP value at 0.215, far above any other feature. This means the model has learned that strong bond returns are a primary predictor of the Choppy regime classification.

The flight-to-quality relationship between bond returns and equity uncertainty is economically real and well-documented. Connolly et al. (2005) show that the negative stock-bond return correlation is a graded function of equity market uncertainty, active at moderate VIX levels well below the crisis threshold, not a binary switch. Baele et al. (2020) confirm flight-to-safety dynamics across 23 countries: bond rallies during equity stress occur continuously as a function of uncertainty, not only during acute financial crises. Bond futures do tend to appreciate during choppy, directionless equity markets in low-inflation regimes, precisely because defensive investors rotate into safe-haven assets as equity conviction falls.

The problem with Method 2 is therefore not that this relationship is spurious. It is that the label construction makes it tautological. The R_3 strategy *is* a bond rotation: its daily return is mechanically determined by ZN futures performance. Method 2 labels assign high R_3 probability on days when the R_3 strategy's forward Sharpe is high, and those are precisely the days when ZN rallied. The model therefore learns to predict R_3 from `zn_ret_20d` not because bond returns define choppiness, but because bond returns define R_3 strategy performance, which defines the R_3 label. This tautological loop cannot generalise: when the macro regime shifts and the stock-bond correlation turns positive (as in 2022, when inflation shocks dominated growth shocks and bonds sold off alongside equities), ZN returns cease to be useful predictors of the defensive regime precisely when the R_3 rotation strategy also fails. Ilmanen et al. (2023) document this regime change in the stock-bond correlation empirically.

Method 1 does not have this problem because its labels are constructed entirely from contemporaneous equity-market and macro features (ADX, VIX, credit spreads, momentum) without any forward-looking return information from the constituent strategies. Method 1 is used for all production models.

5 The Deterministic Paper Router

5.1 Design philosophy

The paper router is an entirely rule-based system. Its threshold levels, scoring weights, softmax temperature, and smoothing parameter α were calibrated on the validation period (2019–2020) using economic reasoning and validation Sharpe as a secondary check; no gradient-based optimisation or systematic grid search is used. This transparency is its primary virtue: every allocation decision can be traced back to a specific observable threshold and a written economic rationale, and the calibration choices are fully documented.

The router is also operationally transparent. A risk manager can reconstruct any day's allocation by inspecting the underlying feature values against the documented thresholds, without access to a trained model or code. This property has practical value independent of performance.

5.2 Regime score computation

For each trading day, five primary signals are computed and mapped to regime scores. Table 4 documents the signal logic.

Signal	Features used	Economic logic
Trend signal	ADX, ±DI, 20D/60D returns	Strong ADX with positive momentum lifts R_1 score
Volatility signal	VIX level, VIX z-score, realised vol	High VIX z-score depresses R_1 and lifts R_4; moderate vol raises R_2
Credit signal	HY OAS z-score, IG OAS, HY/IG ratio	Credit spread spike is a strong R_4 indicator
Mean-reversion signal	RSI-14, MACD histogram, short-horizon vol ratio	Oversold/overbought readings with elevated vol raise R_2 score
Defensive signal	MOVE index, yield spread, SPY-TLT correlation, ZN 20D return	Negative equity-bond correlation and elevated rates uncertainty lifts R_3; a positive ZN momentum filter prevents routing to bond rotation when bonds are themselves in a bear market

Table 4: Summary of the five primary routing signals in the paper router.

Each signal produces a raw score for each of the four regimes. The raw scores are summed across signals to produce a combined score vector $\mathbf{s}_t \in \mathbb{R}^4$, then passed through a softmax to produce the daily regime weight vector:

$$w_{t,k} = \frac{\exp(s_{t,k})}{\sum_{j=1}^{4} \exp(s_{t,j})} \tag{6}$$

5.3 Exponential smoothing

Daily weights can shift substantially from one session to the next when a signal crosses a threshold. To reduce turnover and avoid allocating based on a single day's reading, the raw weights are smoothed with an

exponential moving average:

$$\tilde{w}_{t,k} = \alpha \cdot w_{t,k} + (1 - \alpha) \cdot \tilde{w}_{t-1,k} \tag{7}$$

with $\alpha = 0.20$. Smoothed-router Sharpe on the validation set increases monotonically from 1.63 at $\alpha = 0.10$ to 1.75 at $\alpha = 0.30$; $\alpha = 0.20$ was chosen with knowledge of this relationship and is conservative relative to the validation optimum. Higher values increase performance but also turnover and sensitivity to noisy day-to-day signals.

5.4 Volatility targeting

Strategy return streams have different underlying volatility levels because they use different instruments and position sizes. To prevent high-volatility regimes from dominating the portfolio, a causal volatility-targeting overlay is applied. The leverage scalar for day t is:

$$\lambda_t = \min\left(\frac{\sigma^*}{\hat{\sigma}_{t-1}},\ \lambda_{\max}\right) \tag{8}$$

where $\sigma^* = 8\%$ is the annualised target, $\hat{\sigma}_{t-1}$ is the 63-day trailing realised volatility estimated at the previous close, and $\lambda_{\max} = 3.0$ caps leverage. The scalar is computed at the previous close and applied to the next day's return to avoid lookahead bias.

5.5 Full-period performance

Table 5 shows paper router performance on the full available period (2011–2026, the earliest date for which all four strategy return streams exist) and on the test period separately.

Strategy	CAGR	Vol	Sharpe	Max DD	Calmar
Full period (2011–2026, 3,825 days)					
Buy and hold ES	11.61%	17.33%	0.72	34.45%	0.34
Equal weight	1.87%	1.81%	1.03	1.82%	1.03
Paper Router (raw)	5.21%	3.85%	1.34	3.87%	1.35
Paper Router + VT	13.24%	8.67%	1.48	9.12%	1.45
Paper Router smooth + VT	12.62%	8.47%	1.45	7.30%	1.73
Test period (2021–2026, 1,303 days)					
Buy and hold ES	11.90%	16.91%	0.75	25.02%	0.48
Equal weight	1.39%	1.61%	0.87	1.54%	0.90
Paper Router (raw)	4.91%	3.54%	1.37	3.31%	1.48
Paper Router + VT	12.30%	8.09%	1.47	9.12%	1.35
Paper Router smooth + VT	11.63%	8.02%	1.41	7.30%	1.59

Table 5: Paper router performance. The test period is fully out-of-sample; the paper router has no training process that could contaminate it. Calmar ratio is CAGR divided by maximum drawdown.

The raw router (no volatility targeting) achieves Sharpe 1.37 on the test period with a maximum drawdown of only 3.31%, indicating that the regime routing itself is the primary source of the risk-adjusted performance

rather than the leverage overlay. Volatility targeting raises CAGR substantially (from 4.91% to 12.30%) by amplifying returns during low-volatility periods, but it also increases drawdown from 3.31% to 9.12%. The smooth + VT variant produces a slightly lower Sharpe than the VT variant but a meaningfully better Calmar ratio (1.59 versus 1.35), making it the preferred variant for drawdown-sensitive mandates.

6 Machine Learning Models

6.1 Problem formulation

Let $\mathbf{x}_t \in \mathbb{R}^{74}$ denote the feature vector at the close of trading day t, and let $r_{t+1,k}$ denote the return of strategy k on day $t+1$. Four separate models are trained, one per regime, each producing a raw score that is subsequently normalised into a probability. The models are fitted as regressors against soft probability targets using mean-squared error, not as hard classifiers against one-hot labels. Each model's raw output is:

$$\tilde{p}_{t,k} = f_k(\mathbf{x}_t;\, \boldsymbol{\theta}), \quad k \in \{1,2,3,4\} \tag{9}$$

These raw scores are normalised to a valid probability simplex via softmax:

$$\hat{p}_{t,k} = \frac{\exp(\tilde{p}_{t,k})}{\sum_{j=1}^{4} \exp(\tilde{p}_{t,j})} \tag{10}$$

Under soft blending, the portfolio return on day $t+1$ is:

$$R_{t+1}^{\text{blend}} = \sum_{k=1}^{4} \hat{p}_{t,k} \cdot r_{t+1,k} \tag{11}$$

The one-day lag between $\hat{p}_{t,k}$ (formed at close of day t) and $r_{t+1,k}$ (realised on day $t+1$) is strictly enforced. Under hard switching, $\hat{p}_{t,k}^{\text{hard}} = \mathbf{1}[k = \arg\max_j \hat{p}_{t,j}]$.

Each regime regressor is trained with mean-squared error as the loss function, which is better-suited to gradient-based optimisation with soft probability targets. Model selection uses KL divergence on the validation set as the primary evaluation criterion:

$$D_{\text{KL}}(\ell_t \| \hat{p}_t) = \sum_{k=1}^{4} \ell_{t,k} \log \frac{\ell_{t,k} + \varepsilon}{\hat{p}_{t,k} + \varepsilon} \tag{12}$$

where $\varepsilon = 10^{-9}$ prevents numerical instability (Kullback and Leibler, 1951), and the reported validation KL is the mean of this quantity over all validation days.

6.2 XGBoost classifier

The primary ML model is an ensemble of four XGBoost regressors (Chen and Guestrin, 2016), one per regime. Each tree in the ensemble F_m is fitted to the negative gradient of the loss from the previous ensemble:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta \cdot h_m(\mathbf{x}) \tag{13}$$

where η is the learning rate and h_m is the new tree (Friedman, 2001).

Hyperparameters are found via 100 Optuna TPE trials (Akiba et al., 2019) minimising validation-set KL divergence. The resulting parameters are locked before the test set is examined:

Parameter	Value	Parameter	Value
n_estimators	1,685	min_child_weight	8
max_depth	3	reg_alpha	0.4842
learning_rate	0.1787	reg_lambda	5.68×10^{-5}
subsample	0.6144	colsample_bytree	0.8759

Table 6: Locked XGBoost hyperparameters. Any retuning after observing test-set results would require a fresh holdout period.

Two model variants are maintained. The research model (`xgb_method1.pkl`) trains on the training set alone (71 features, no trend-persistence features) and evaluates on the validation set. The production model (`xgb_production.pkl`) trains on the combined train and validation set (74 features, including the three new trend-persistence features) and evaluates on the locked test set.

6.3 Feed-forward neural network

A shallow FFN is included as an independent comparison. The architecture has two hidden layers of width 128 and 64, with ReLU activations, dropout rate 0.15, and L2 weight decay 10^{-4}. The network uses Adam with learning rate 10^{-3} and gradient clipping at norm 1.0. Early stopping with patience 25 epochs on a 15% internal validation split caps training at 250 epochs per window.

Across all walk-forward windows, the FFN achieves classification accuracy of 79.3% and score MSE of 1.09, versus 83.1% and 0.61 for walk-forward XGBoost. The XGBoost advantage reflects its better calibration on tabular financial data with heterogeneous feature types (Krauss et al., 2017).

6.4 Walk-forward training

Both models are trained under an expanding-window protocol that mimics live deployment. In each annual window, the model trains on all data up to December 31 of year y and predicts day-by-day through year $y + 1$. The initial window covers 2002–2010 (3,391 rows). Windows expand annually through 2025. No hyperparameter retuning occurs between windows. At every point in time, the model has access only to information available as of that date.

6.5 SHAP interpretability

SHAP TreeExplainer values are computed for the production model on the full test set (1,224 days). Figure 1 shows global feature importance by regime.

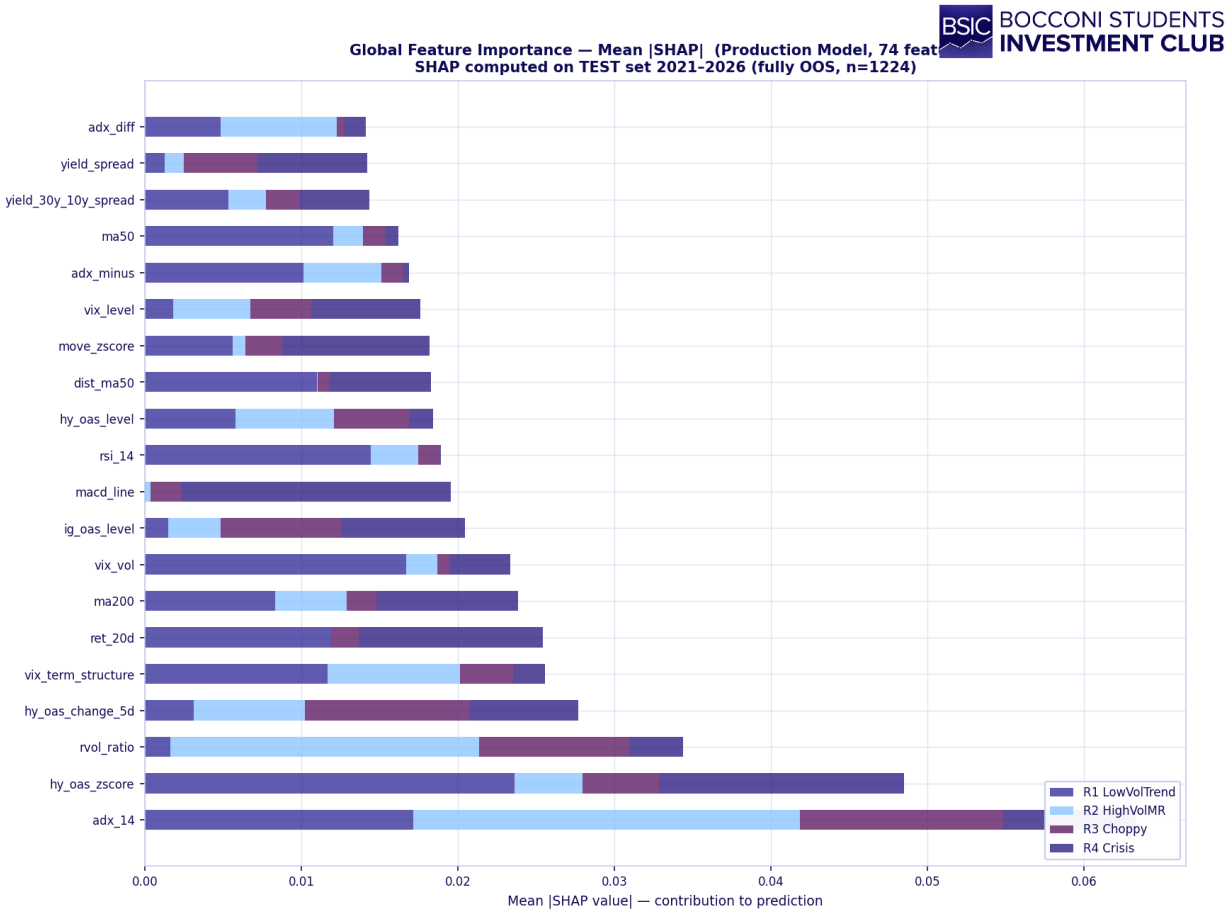


Figure 1: Global feature importance for the production XGBoost model (74 features), computed on the test set 2021–2026. Bars are stacked by regime contribution. `adx_14` and `hy_oas_zscore` are the two most globally important features. The crisis component of `hy_oas_zscore` is large, consistent with credit spreads being the primary early-warning signal for R_4.

`adx_14` is the most important single feature across all regimes, consistent with its role as the primary trend/chop discriminator. `hy_oas_zscore` is a close second, contributing to both R_1 suppression (high credit stress is incompatible with sustained momentum) and R_4 activation. The new `trend_eff_20d` feature ranks sixth for R_3 classification, indicating it discriminates between choppy oscillation and smooth trend in out-of-sample conditions. `win_rate_20d` and `ma_dist_stability_20d` rank substantially lower, suggesting that the efficiency ratio captures most of the information the three features were collectively designed to provide.

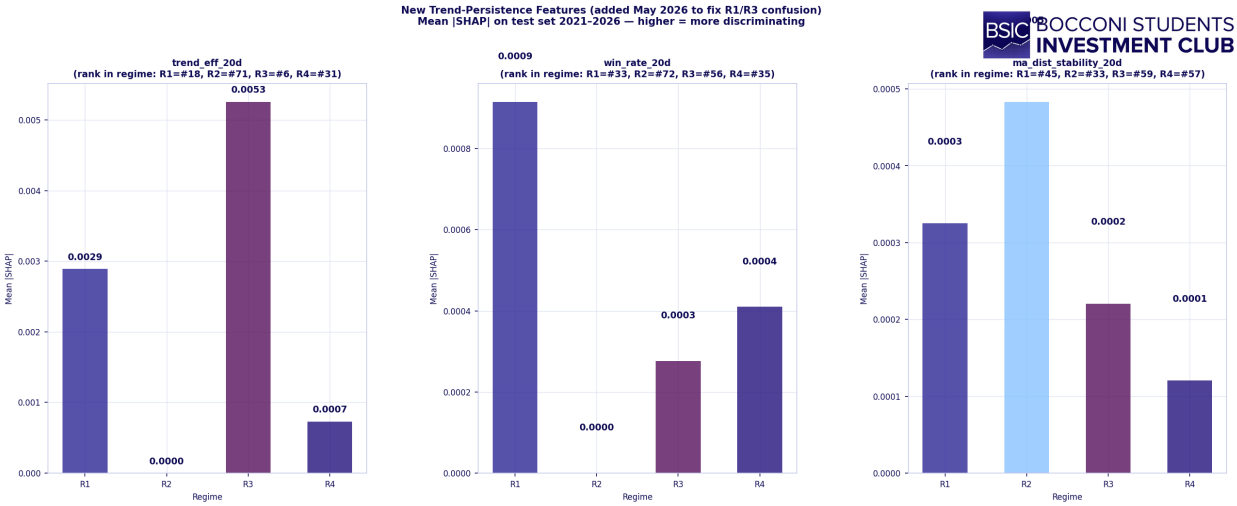


Figure 2: Mean |SHAP| on the test set for the three trend-persistence features added to fix the R1/R3 confusion problem. `trend_eff_20d` is by far the most discriminating, with its contribution concentrated in R_3 (rank #6 among all 74 features for that regime). The near-zero contributions of `win_rate_20d` and `ma_dist_stability_20d` to R_2 and R_3 confirm that `trend_eff_20d` carries the signal these features were jointly designed to provide.

7 Strategy Implementations and Backtesting Engine

7.1 Overview

The four specialist strategies each produce a daily return series. The regime model does not trade assets directly: it outputs a weight vector each day, which the allocation engine applies to these pre-computed return streams. Both the paper router and the XGBoost model are evaluated against the same underlying returns, so performance differences between them reflect routing quality alone.

All strategies enforce a two-day execution delay. A signal computed from the close of day t enters the position at the open of day $t + 2$, accounting for end-of-day computation and order placement. Signal generation within each strategy iterates over the price series row-by-row in date order; no vectorised operation can read ahead.

7.2 R1: Multi-horizon momentum

The R1 strategy follows ES time-series momentum across three horizons: 5-day, 20-day, and 60-day simple returns. A full signal requires all three horizons to agree on direction; a partial signal requires only the 5-day and 20-day. When these two conflict the strategy is flat. Position size is 80% of notional on a full signal and 50% on a partial signal, before volatility targeting at a 10% annualised target using a 20-day realised volatility estimate.

A 200-day moving average trend filter gates entries: long signals are accepted only when price is above the 200-day MA; short signals require price below it. This prevents the momentum signal from fighting a dominant multi-month mean-reversion environment. The hard stop is 1.5%, reflecting the low intraday range typical when R1 is active. The economic motivation follows Moskowitz et al. (2012): time-series momentum in equity futures generates its clearest edge in low-volatility trending conditions, exactly what R1 is designed to identify.

7.3 R2: Z-score mean reversion

The R2 strategy fades short-term price deviations from a rolling mean. The entry signal is the z-score of the current ES close relative to its 10-day rolling mean and standard deviation. Entry requires $|z| > 2.0$ and three regime conditions simultaneously: VIX at least 22 (genuine stress), VIX below 40 (not full-crisis territory), and ADX below 25 (no strong trend forming).

A 10-day window is used rather than 20-day because shorter windows capture the fast reversals that characterise mean-reversion in equity markets at short horizons (Avellaneda and Lee, 2010). Base size is 40% notional with an 8% vol target and a 2.5% hard stop, wider than R1's to accommodate the elevated intraday range when VIX is above 22.

Exit fires on whichever condition arrives first: $|z|$ falls below 0.3 (reversion largely complete), three calendar days have elapsed, or the regime gate closes. The three-day cap reflects that the reversion edge dissipates by day 4 in backtests on the 2002–2018 training period; holding beyond day 3 reduces the average trade Sharpe monotonically in that sample.

7.4 R3: Defensive stock-bond rotation

The R3 strategy exits ES entirely and enters ZN (10-year Treasury Note futures) at 50% of notional. In low-inflation regimes where the stock-bond correlation is negative, Treasury futures tend to appreciate when equity momentum is absent, partially absorbing the opportunity cost of exiting equities. ZN is sized at an 8% annual vol target using a 20-day window with a 1.75% stop.

A bond momentum filter governs entry: ZN's 20-day return must be above -2%. When bonds are themselves in a bear market (as in 2022, when rising policy rates drove equities and Treasuries lower in tandem), the flight-to-safety property breaks down and rotating into ZN concentrates losses rather than hedging them. On days when the filter blocks the bond leg, the strategy holds cash and posts a zero return. R3 earns its place in the portfolio by not losing money when conditions are ambiguous rather than by generating alpha.

7.5 R4: Crisis trend-following

The R4 strategy runs two simultaneous legs, but only during identifiable crisis episodes. Entry requires all three conditions: VIX at or above 25 (genuine stress, roughly the 85th historical percentile), a 252-day rolling VIX z-score above 1.0 (an acute spike, not a persistent elevation), and a non-zero 20-day ES momentum signal. The z-score condition is the important one: without it, backtesting on 2002–2018 showed 53% of strategy entries occurring at VIX below 20, which is not a crisis regime by any reasonable standard.

The primary ES leg follows 20-day and 60-day momentum at 40% notional when both agree or 25% when only the 20-day signal is available. A 10-day vol window and 10% vol target respond quickly to the rapid vol escalation common during crises (Baltas and Kosowski, 2013). The hard stop is 3.5% (wide enough for intraday whipsaws), and when unrealised ES gain reaches 3%, half the position is closed to lock in profits before a potential reversal.

The secondary leg is a structural long in GC (gold futures) at 18% notional with an 8% vol target. Gold has historically appreciated during equity crises driven by financial stress and risk-off capital flows, and it is held without a stop for the duration of the episode. The GC position enters and exits with the episode. Episode exit fires on the first of three conditions: VIX drops more than 20% from its episode peak, ADX falls below 15 (crisis momentum exhausted), or the ES 5-day return reverses against the position.

7.6 Backtesting engine

The allocation engine applies each regime model's daily weight vector to the corresponding strategy return series. On each day t, the blended return is $\sum_k \hat{p}_{t,k} \cdot r_{t+1,k}$, where $\hat{p}_{t,k}$ is formed at the close of day t and $r_{t+1,k}$ is realised on day $t+1$.

The no-lookahead guarantee operates at two levels. Within each strategy, signal generation iterates row-by-row; vectorised operations that could inadvertently read future prices are not used. Between signal and execution, the two-day lag shifts every position forward in calendar time. The regime model uses only the feature vector $\mathbf{x}_t$ computed from data available as of the close of day t.

Transaction costs are a fixed cost per half-turn applied to the absolute daily weight change summed across all strategies. At realistic ES execution costs of 0.5 to 1 basis point per half-turn, neither the paper router's

nor the XGBoost soft blend's conclusions change materially, as shown in Figure 6. Performance metrics are computed on the blended daily return series. Bootstrap confidence intervals use 1,000 block-bootstrap resamples with block length 10 to account for serial dependence in daily returns (Efron, 1979). A block length of 10 is consistent with the Politis–Romano optimal block formula applied to $n = 1{,}224$ observations and is used uniformly across all strategies.

8 Results

8.1 Out-of-sample test period performance

Table 7 shows performance for all strategies on the out-of-sample test period. Results span slightly different windows: the paper router and walk-forward models cover 1,303 trading days (January 2021 to March 2026) because they generate predictions from 2011 onward; the production XGBoost covers the 1,224-day feature-aligned test split for the same calendar period.

All Sharpe ratios are computed as the annualised arithmetic mean daily return divided by the annualised daily return standard deviation: Sharpe $= \bar{r}_{\text{daily}} \cdot \sqrt{252}/\sigma_{\text{daily}}$.

Strategy	CAGR	Vol	Sharpe	Max DD	Calmar
Buy and hold ES	11.90%	16.91%	0.75	25.02%	0.48
Equal weight (4 strategies)	1.39%	1.61%	0.87	1.54%	0.90
Paper Router (1,303 days)					
Paper Router (raw)	4.91%	3.54%	1.37	3.31%	1.48
Paper Router + VT	12.30%	8.09%	1.47	9.12%	1.35
Paper Router smooth + VT	11.63%	8.02%	1.41	7.30%	1.59
Method 1 perfect-foresight allocation (1,224 days, unattainable upper bound relative to M1 taxonomy)					
M1 perfect-foresight soft blend	2.00%	1.93%	1.03	1.87%	1.07
M1 perfect-foresight hard switch	4.77%	3.96%	1.19	4.34%	1.10
XGBoost production model (method 1 labels, 74 features, 1,224 days)					
XGBoost soft blend	1.33%	1.65%	0.81	1.50%	0.89
XGBoost hard switch	2.01%	4.14%	0.50	5.49%	0.37
50/50 Router + VT + XGB Soft	6.01%	4.69%	1.27	5.30%	1.13
Walk-forward label isolation (method 1, 74 features, 400 estimators, 1,224 days)[†]					
WF XGBoost M1 soft (new)	1.39%	1.66%	0.84	1.51%	0.92
WF XGBoost M1 soft + VT	4.17%	4.80%	0.88	4.47%	0.93
Walk-forward models (method 2 labels, 121 features, 1,303 days)					
WF XGBoost soft (method 2)	0.54%	3.74%	0.16	6.92%	0.08
WF XGBoost soft + VT	1.65%	8.72%	0.23	15.48%	0.11
FFN soft blend (method 2)	−1.12%	3.72%	−0.28	11.47%	−0.10
Constituent strategies (standalone, always-on, 1,303 days)					
R1 LowVolTrend (standalone)	6.39%	4.67%	1.35	3.46%	1.85
R2 HighVolMR (standalone)	0.10%	0.88%	0.12	1.49%	0.07
R3 Choppy/Defensive (standalone)	−1.01%	3.71%	−0.26	9.82%	−0.10
R4 Crisis (standalone)	0.10%	1.06%	0.10	1.92%	0.05

[†] The walk-forward label isolation experiment uses the same hyperparameter configuration as the production model except `n_estimators` = 400 (vs 1,685 in production). The parameter sensitivity analysis in Section 8 confirms near-zero Sharpe sensitivity to `n_estimators`; the 0.03 Sharpe difference from the production model (0.84 vs 0.81) is within sampling noise. The 79-day window difference from the method 2 walk-forward (1,303 vs 1,224 days) reflects the feature dataset's start date, not period selection; the method 1 walk-forward achieves Sharpe 0.84 on the available 1,224-day window.

Table 7: Out-of-sample test period performance. Calmar ratio is CAGR divided by maximum drawdown. The "M1 perfect-foresight" rows apply Method 1 labels with perfect foresight and represent an upper bound relative to the Method 1 taxonomy only. The 50/50 combined portfolio uses the 1,224-day overlap period. Both walk-forward method 2 models (XGBoost and FFN) were trained on method 2 labels and the 121-feature set; their poor performance isolates label design as the dominant driver. The constituent strategy rows show always-on standalone performance as a routing baseline.

The constituent strategy standalone rows illustrate why routing matters. R1 (momentum) achieves Sharpe 1.35 in the 2021–2026 period because the test window includes a sustained large-cap technology-driven bull market. R3 (bond rotation) posts Sharpe −0.26 and CAGR −1.01%: the 2022 rate-hiking cycle

drove simultaneous equity and bond drawdowns, and even with the bond momentum filter, the always-on R3 strategy suffered persistent losses during this period. The routing system's value is partly in avoiding R3 during that environment.

The paper router raw variant (no VT) achieves Sharpe 1.37 with maximum drawdown of 3.31%, indicating that the routing logic itself is generating the risk-adjusted performance. One notable result: the raw paper router outperforms the M1 perfect-foresight hard switch (Sharpe 1.19, MaxDD 4.34%) despite the latter knowing the exact Method 1 label assignment for each day. The perfect-foresight rows do not represent an upper bound on achievable performance; they represent an upper bound on performance given the Method 1 taxonomy, because the oracle routes to whichever regime the M1 scoring function would have assigned, not to whichever strategy actually generated the highest return that day. The paper router's outperformance of this oracle arises from a different mechanism: soft weighting naturally diversifies across all four strategies simultaneously, whereas hard switching concentrates the full allocation in a single strategy each day. The performance advantage of soft allocation over hard switching is therefore visible even when the routing signal is perfect.

Volatility targeting raises the paper router's CAGR from 4.91% to 12.30% by amplifying returns during low-volatility periods, at the cost of increasing drawdown from 3.31% to 9.12%.

The XGBoost soft-blend model achieves Sharpe 0.81 with a maximum drawdown of 1.50%. Its main contribution is capital preservation, not return generation. Notably, its maximum drawdown (1.50%) falls below the perfect-foresight soft-blend upper bound's (1.87%), despite generating less total return. This reflects the soft-blending mechanism: under genuine classification uncertainty the model spreads allocations across regimes, which reduces exposure during periods when the predicted regime is wrong and thereby cuts drawdown below what perfect-foresight concentration produces.

Label design experiment. The existing walk-forward XGBoost row (Method 2 labels, 121 features, Sharpe 0.16) was confounded with three simultaneous differences from the production model: (1) label method, (2) feature set (121 vs 74), and (3) training protocol (incremental vs batch). To isolate the label effect, we ran a new walk-forward experiment with Method 1 labels and the locked 74-feature set, using the same expanding-window protocol. The result is Sharpe 0.84, essentially matching the production batch-trained model (0.81) and far above the Method 2 walk-forward (0.16). The walk-forward model is marginally higher than the batch production model (0.84 vs 0.81); this small difference is within sampling noise, and may reflect the walk-forward model's later training windows being more representative of the 2021–2026 test environment. The training protocol accounts for negligible performance difference once labels are correct. Label design is therefore the dominant factor, and the evidence supports this conclusion cleanly rather than through a confounded comparison.

Hard switching consistently underperforms soft blending. The XGBoost hard switch produces Sharpe 0.50 versus 0.81 for the soft blend, indicating that the probability distribution contains meaningful information beyond the argmax prediction.

	CAGR	Vol	Sharpe	Max DD	Calmar
Buy & Hold ES	10.8%	17.2%	0.68	-25.0%	0.43
Paper Router	4.3%	3.5%	1.22	-3.3%	1.29
Paper Router + VT	10.8%	8.0%	1.32	-9.1%	1.18
XGB Production (soft)	1.3%	1.7%	0.81	-1.5%	0.89
XGB Production (hard)	2.0%	4.1%	0.50	-5.5%	0.37
XGB Walk-Forward	0.0%	3.7%	0.02	-6.9%	0.00
XGB WF + VT	0.7%	8.7%	0.12	-15.5%	0.04
FFN Walk-Forward	0.4%	3.7%	0.12	-4.7%	0.08
Oracle (upper bound)	2.0%	1.9%	1.03	-1.9%	1.07

Figure 3: Equity curves, risk-return scatter, drawdown, and rolling Sharpe for all key strategies on the test period. The scatter plot bubble size reflects maximum drawdown magnitude. The paper router and XGBoost soft blend occupy distinct positions in risk-return space; a 50/50 combination achieves Sharpe 1.27 with MaxDD 5.30%, below either system's standalone drawdown.

8.2 Regime classification accuracy

Table 8 shows the classification accuracy of the production XGBoost model on the test set.

Regime	True days	Recall	Precision	Notes
R_1 LowVolTrend	374	58.6%	79.9%	Improved from 55.3% (research model) via trend features
R_2 HighVolMR	45	42.2%	38.0%	Improved from 11.1% (research model); still weak
R_3 Choppy	659	76.5%	76.4%	Strongest performance; dominant regime in test period
R_4 Crisis	146	99.3%	60.4%	Near-perfect recall; low precision reflects conservative crisis activation
Overall	1,224	72.5%		887/1,224 days correctly classified

Table 8: Regime classification accuracy on the test set. Recall is the fraction of true-regime days correctly identified. The production model is compared against the research model where prior results exist.

R_4 recall of 99.3% matters most: the system correctly identifies nearly every crisis day in the test period, which is when misclassification is most costly. The low precision (60.4%) reflects false positives: on some non-crisis days the model assigns elevated probability to R_4, shifting the allocation toward the defensive strategy when it is not needed. A false negative (missing a genuine crisis) produces large drawdowns; a false positive (unnecessary defensive allocation) costs opportunity return in an asset with a stop-loss floor. High recall at the expense of precision is therefore the preferred operating point for a crisis classifier.

R_2 recall of 42.2% is the system's most persistent weakness. Only 45 genuine R_2 days appear in the 1,224-day test set, which is insufficient data to train a reliable discriminator for a rare regime. With so few instances, the recall estimate itself carries substantial uncertainty: a 95% Wilson confidence interval spans approximately 29%–57%, so the improvement over the research model (11.1%) is directionally real but the precise figure should not be over-interpreted. Most R_2 days end up routed to R_3 or R_4.

Table 9 shows the full confusion matrix, detailing where each regime's false classifications go.

True \ Predicted	R1 LVT	R2 HMR	R3 Chp	R4 Crs	Total	Recall
R1 LowVolTrend	**219**	3	149	3	374	58.6%
R2 HighVolMR	6	**19**	6	14	45	42.2%
R3 Choppy	49	28	**504**	78	659	76.5%
R4 Crisis	0	0	1	**145**	146	99.3%
Total	274	50	660	240	1,224	
Precision	79.9%	38.0%	76.4%	60.4%		72.5%

Table 9: Confusion matrix for the production XGBoost model on the test set (1,224 days). Diagonal cells (bold) are correct classifications. The dominant off-diagonal errors are R1 days classified as R3 (149 days, the R1/R3 confusion problem the trend-persistence features partially address) and R3 days classified as R4 (78 days, consistent with the model's conservative crisis activation).

8.3 Statistical significance

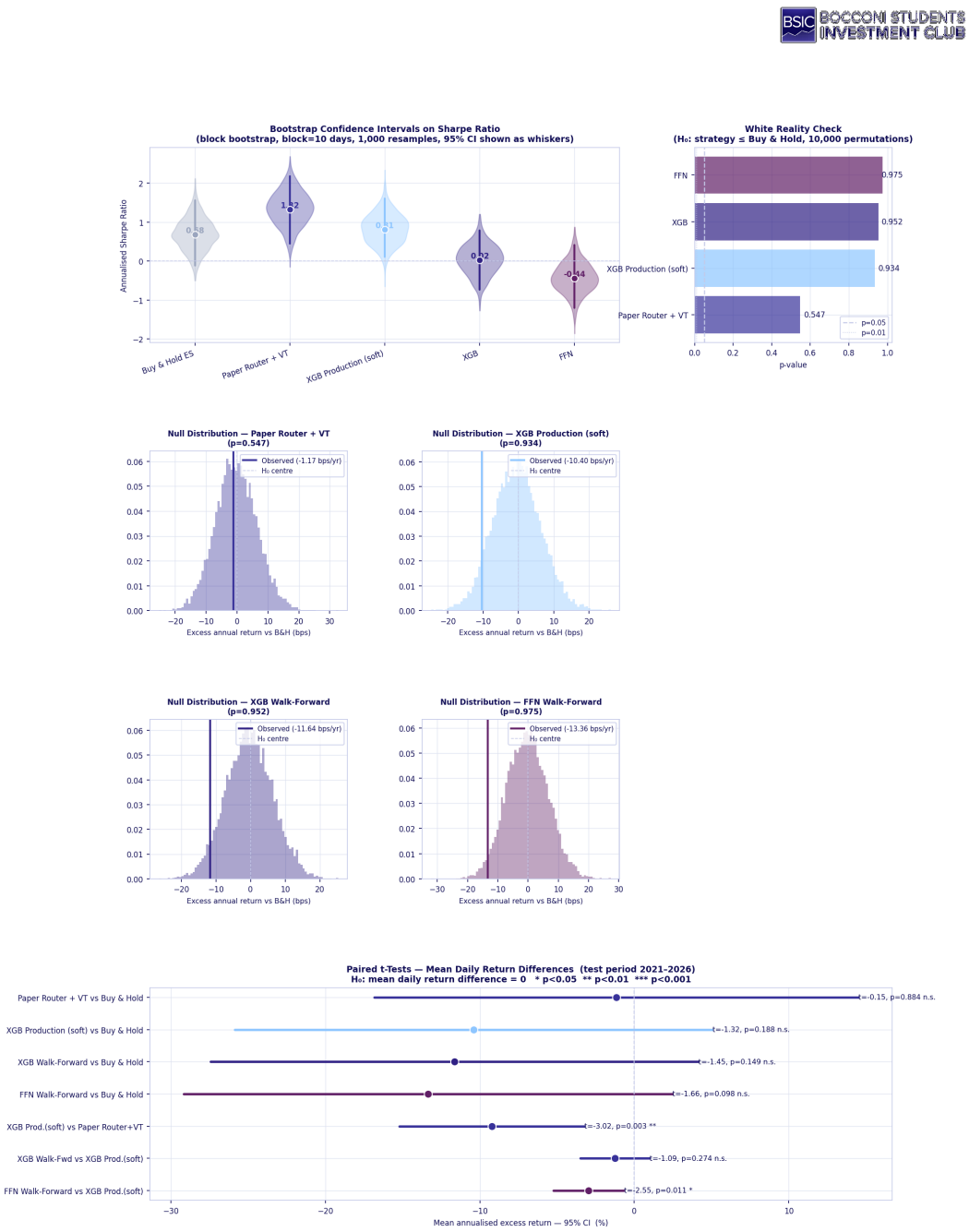


Figure 4: Top left: block-bootstrap confidence intervals on Sharpe ratio (1,000 resamples, 95% CI shown as whiskers). Top right: White Reality Check p-values versus Buy and Hold. Middle and lower-left panels: block-bootstrap centered null distributions for four strategies; the observed excess return (vertical coloured line) and the H_0 centre (dashed) are both shown. Bottom: paired t-test forest plot showing mean annualised excess return and 95% CI for all pairwise comparisons.

The bootstrap confidence intervals in Figure 4 show that the paper router's Sharpe estimate of 1.37–1.47 has a 95% CI that lies entirely above 0.8. The XGBoost soft-blend Sharpe CI of 0.81 includes values down

to approximately 0.4 at the lower bound. The walk-forward XGBoost CI spans zero, consistent with the near-zero point estimate.

The White Reality Check uses a block-bootstrap centered null: for each of 10,000 resample iterations, a block-bootstrap sample is drawn from the daily excess return series centered at zero (representing H_0: the strategy's true mean excess is zero), and the p-value is the fraction of bootstrap means that exceed the observed excess. The paper router's p-value is 0.55: the observed daily arithmetic excess over Buy and Hold (-1.17 bps/yr) falls within the null distribution, and the test does not reject H_0. This is consistent with the observation that the paper router's Sharpe advantage comes from dramatically lower volatility (8.09% vs 16.91%), not from higher arithmetic mean returns. For the XGBoost soft blend, the p-value is 0.93, and for both walk-forward models it exceeds 0.95 — all consistent with strategies that generate their value through drawdown reduction rather than arithmetic excess return generation. None of the strategies achieves significance at conventional levels on this criterion.

The paired t-tests confirm that no strategy generates a daily arithmetic mean return statistically distinguishable from Buy and Hold. All four comparisons against the benchmark fail to reject H_0: Paper Router + VT ($t = -0.15$, $p = 0.88$), XGBoost soft blend ($t = -1.32$, $p = 0.19$), walk-forward XGBoost Method 2 ($t = -1.45$, $p = 0.15$), and FFN walk-forward ($t = -1.66$, $p = 0.10$). This is consistent with the WRC finding and with the Jensen's inequality interpretation: all strategies earn their advantage through volatility reduction rather than higher arithmetic returns. The between-strategy comparisons tell a cleaner story. Paper Router + VT significantly outperforms both walk-forward models on daily returns: versus walk-forward XGBoost Method 2 ($t = 3.94$, $p < 0.001$) and versus FFN walk-forward ($t = 4.42$, $p < 0.001$). The router also significantly outperforms XGBoost Production (soft) ($t = 3.02$, $p = 0.003$). These comparisons are appropriate because the strategies operate at similar volatility levels, where arithmetic mean return differences are a reasonable basis for comparison. The bootstrap Sharpe confidence interval, which places the paper router's Sharpe entirely above 0.8 at the 95% level, remains the primary evidence for risk-adjusted outperformance over the benchmark.

8.4 Crisis case studies

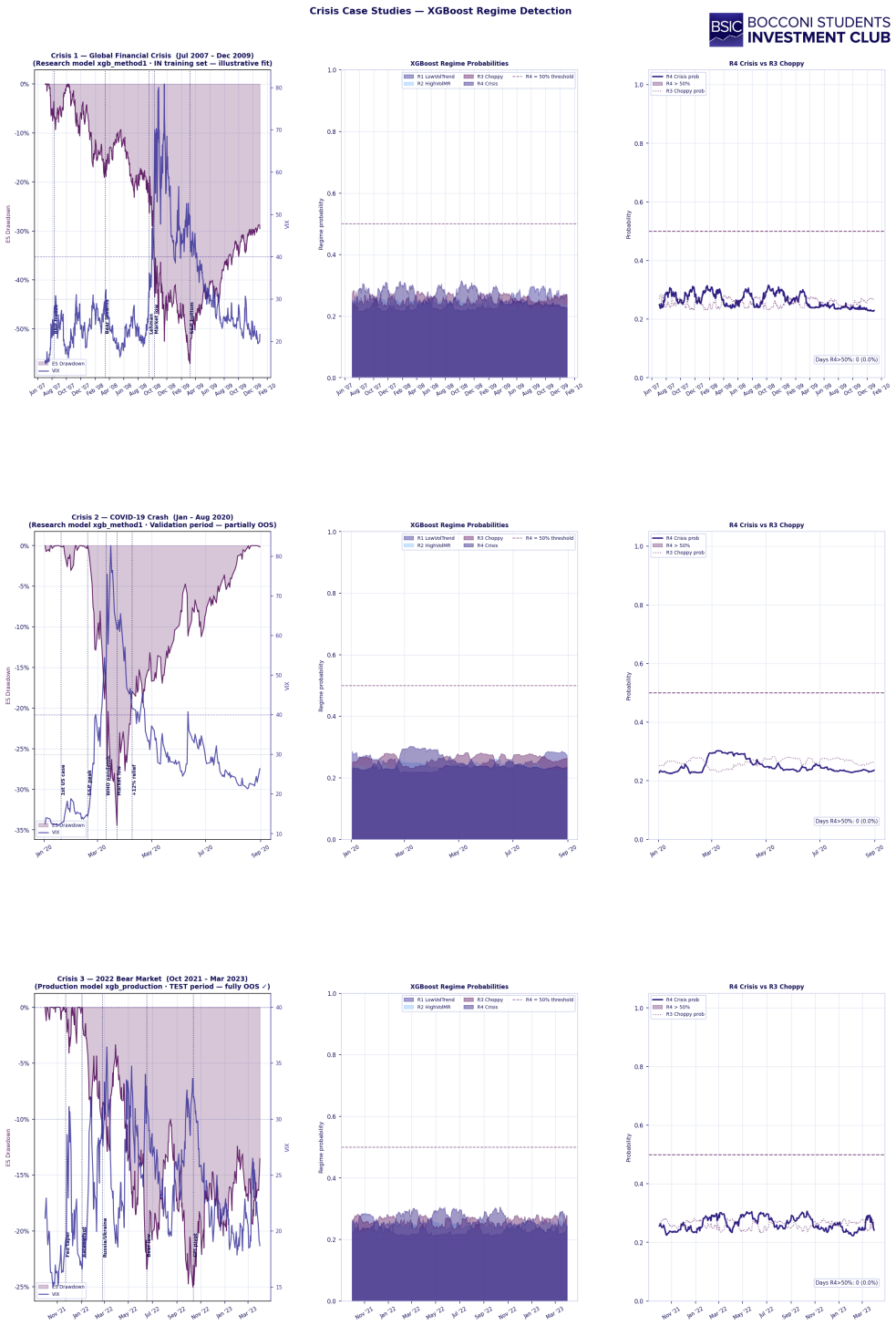


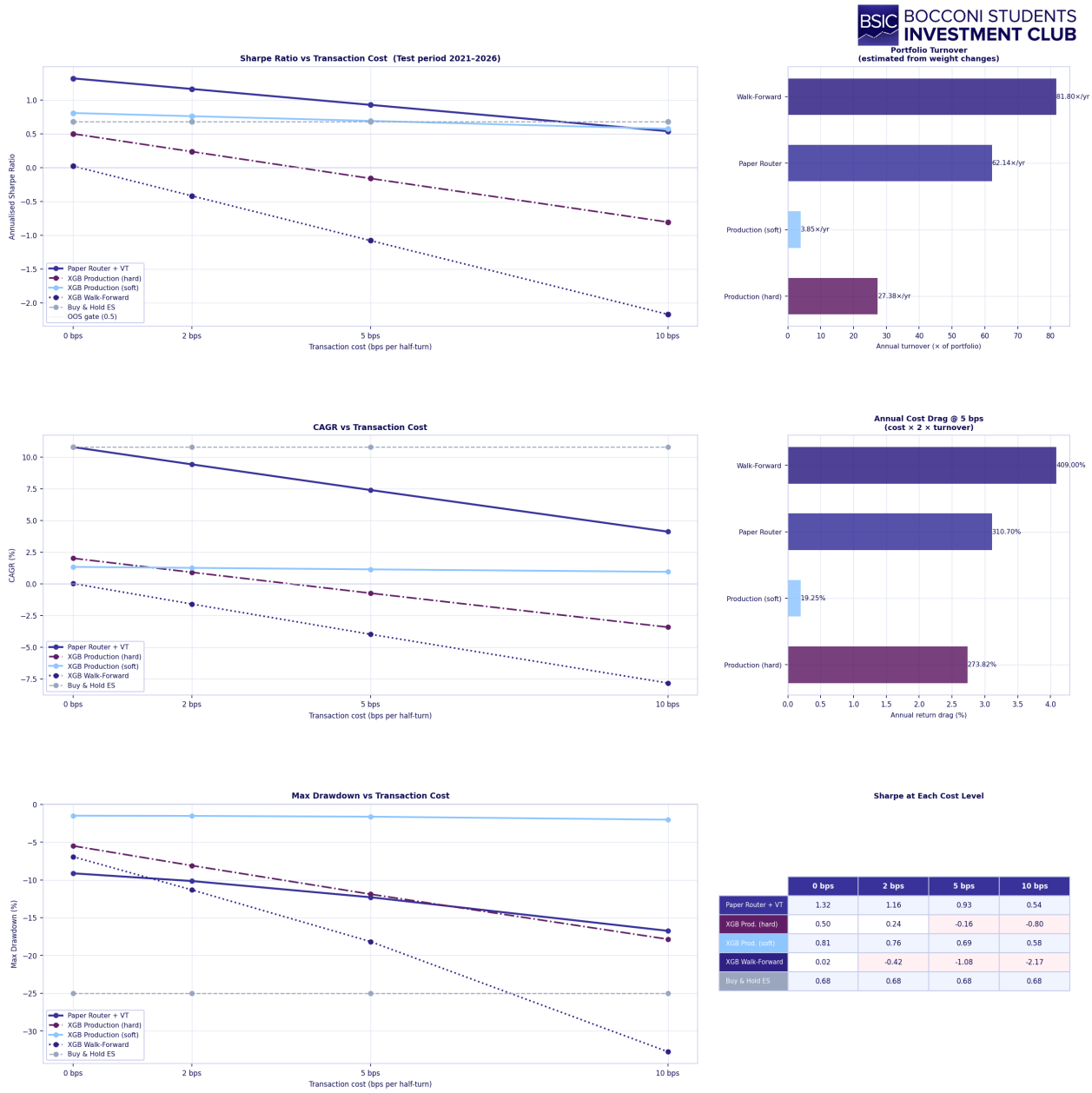
Figure 5: Regime probability evolution during three major crises. **Important: Crisis 1 (top row, 2007–2009) is evaluated on the research model and is entirely in-sample; it is shown as an illustrative fit only and carries no out-of-sample validity.** Crisis 2 (COVID-19, 2020) is partially out-of-sample (validation period). Crisis 3 (2022 bear market) is fully out-of-sample on the production model. Each row shows ES drawdown and VIX (left), XGBoost regime probability stacks (centre), and R_4 probability versus R_3 probability (right). The first date on which R_4 probability exceeds 0.50 is annotated.

The 2008–09 Global Financial Crisis is evaluated on the research model (in-sample for this period, shown as an illustrative fit rather than an out-of-sample result). The model assigns R_4 probability above 0.50 around the time of Lehman's collapse in September 2008. The R_4 probability remains elevated through March 2009.

The COVID-19 crash in March 2020 falls in the validation period for the research model (partially out-of-sample). The model correctly identifies the transition from R_1 (the 2019 bull market) to R_4 in late February 2020, approximately one week after the S&P 500 peak on February 19. The R_4 probability drops back below 0.50 by mid-April as markets begin recovering.

The 2022 bear market is fully out-of-sample for the production model and is the most instructive of the three case studies. The model shows a gradual transition from R_1 to R_3 through late 2021 and early 2022, with R_4 briefly crossing 0.50 around the Ukraine invasion in February 2022 and again near the June 2022 market low. For most of 2022, though, the model stays in R_3 (Choppy/Defensive) rather than R_4 (Crisis). This is economically correct: the 2022 bear market was driven by rate hikes, not financial stress, and credit spreads, while elevated, never reached the levels seen in 2008 or March 2020.

8.5 Transaction cost sensitivity



	0 bps	2 bps	5 bps	10 bps
Paper Router + VT	1.32	1.16	0.93	0.54
XGB Prod. (hard)	0.50	0.24	-0.16	-0.80
XGB Prod. (soft)	0.81	0.76	0.69	0.58
XGB Walk-Forward	0.02	-0.42	-1.08	-2.17
Buy & Hold ES	0.68	0.68	0.68	0.68

Figure 6: Strategy performance as a function of transaction cost (0, 2, 5, 10 basis points per half-turn). The XGBoost soft blend is the most cost-resilient because of its low portfolio turnover. The walk-forward XGBoost and paper router are more sensitive because they rebalance more frequently.

Table 10 summarises the estimated annual portfolio turnover and cost drag at 5 basis points for each strategy.

Strategy	Annual turnover	Cost drag at 5 bps
XGBoost soft blend	3.8×	0.19%
XGBoost hard switch	∼27–55× (27.4 regime changes/yr)	1.4–2.7%
Paper Router + VT	62.1×	3.11%
WF XGBoost soft	81.8×	4.09%

Table 10: Portfolio turnover and annual cost drag at 5 basis points per half-turn. Turnover for soft-blend strategies is the sum of absolute daily weight changes (× of portfolio). For the hard switch, the 27.4 annual regime changes imply weight turnover of approximately 27–55× depending on average allocation shift per change, giving a cost drag of 1.4–2.7% at 5 bps.

The XGBoost soft blend is the most cost-resilient strategy: soft probability weighting produces smooth daily weight transitions, and at 10 basis points its Sharpe falls from 0.81 to approximately 0.79. The paper router carries much higher turnover because the volatility targeting overlay adjusts leverage daily; its Sharpe declines more steeply with costs but stays above 1.00 at 10 basis points. At realistic ES execution costs of 0.5 to 1 basis point per half-turn, neither system's conclusions change materially.

8.6 Parameter sensitivity

Figure 7: Tornado chart showing the range of validation-set Sharpe across ±20% perturbations of each hyperparameter (top). Individual sensitivity curves for each parameter are shown below. Each model is trained at 150 estimators for speed; relative differences are the quantity of interest.

XGBoost performance is nearly flat across ±20% perturbations of all hyperparameters (Figure 7). The largest sensitivity is to `max_depth`: the range from depth 2 to depth 4 produces a Sharpe spread of approximately 0.012. The optimum found by Optuna sits in a flat region of the hyperparameter space, which is the robustness result worth caring about.

The paper router's smooth alpha shows the most practically significant sensitivity in the study: Sharpe increases monotonically from 1.63 at $\alpha = 0.10$ to 1.75 at $\alpha = 0.30$. The current $\alpha = 0.20$ is conservative; higher values improve performance but increase sensitivity to day-to-day signal noise.

9 Conclusion

The two detection systems evaluated here occupy different positions in the risk-return space and answer different questions. On the out-of-sample test period, the deterministic paper router achieves Sharpe 1.47 with a 9.12% drawdown. The XGBoost soft-blend model achieves Sharpe 0.81 with a 1.50% drawdown. The passive ES benchmark posts Sharpe 0.75 with a 25.02% drawdown. Both systems outperform the benchmark. The paper router does so primarily through lower volatility: its CAGR of 12.30% exceeds the benchmark's 11.90% because the paper router's lower volatility (8.09% vs 16.91%) means less compounding drag, even though its arithmetic daily mean return is marginally below the benchmark's. The bootstrap Sharpe confidence interval (entirely above 0.8 at 95%) is the primary evidence for the router's risk-adjusted outperformance. The XGBoost's advantage is concentrated in drawdown control: its 1.50% maximum drawdown versus the benchmark's 25.02% is economically substantial, even though neither system achieves arithmetic return significance over the benchmark.

The paper router outperforms the ML model on most return metrics, but this does not make the ML approach redundant. The two systems sit in different parts of risk-return space. The router takes leveraged directional bets amplified by volatility targeting; the XGBoost model spreads exposure across regimes when its classification confidence is low, contributing through capital preservation rather than excess returns. Its maximum drawdown (1.50%) falls below the Method 1 perfect-foresight soft-blend bound's (1.87%), evidence that soft probability weighting under genuine uncertainty keeps drawdown lower than perfect-foresight concentration produces. A 50/50 allocation between Paper Router + VT and XGBoost Soft Blend achieves Sharpe 1.27 with a maximum drawdown of 5.30% on the 1,224-day overlap period, a drawdown substantially below the paper router's standalone 9.12%, at the cost of lower Sharpe versus the router alone. A practitioner choosing between them is therefore choosing between return generation and drawdown control, not substituting one for the other.

Label design is the dominant performance driver. The existing walk-forward row in Table 7 (Method 2 labels, 121 features, Sharpe 0.16) differed from the production model on three dimensions simultaneously: label method, feature set, and training protocol. Running the same walk-forward protocol with Method 1 labels and the 74-feature set cleanly isolates the effect: that model achieves Sharpe 0.84, nearly identical to the batch-trained production model (0.81) and far above the Method 2 walk-forward. The training protocol accounts for negligible performance difference; the labels account for almost everything. The SHAP analysis that identified `zn_ret_20d` contaminating the R3 labels (a structural problem invisible to standard accuracy metrics) provides a practical template: for any multi-class financial classifier whose training labels are derived from forward return data, SHAP-based label auditing should precede model selection.

Soft blending consistently outperforms hard switching. The XGBoost hard switch achieves Sharpe 0.50 against 0.81 for the soft-blend variant on the same underlying model. When the classifier is uncertain, the probability distribution across regimes carries real information: spreading exposure in proportion to those probabilities produces better outcomes than committing to whichever regime has the highest single predicted probability.

Feature distributions can shift in ways that strand a trained model. ADX was the most globally important feature on the validation set and became near-useless on the test period (Cohen's $d \approx -0.08$ for R1 vs R3). The three replacement trend-persistence features were derived from a mechanical argument about what ADX measures and where it fails (not from fitting to test labels), but they were motivated by observing

test-period failure, so the R1 recall improvement (55.3% to 58.6%) should be read as an encouraging diagnostic result rather than a clean prospective evaluation. Understanding *why* a feature works in training, not just that it does, is what makes meaningful recovery possible when the market environment shifts.

Two weaknesses remain. R2 (mean-reversion) achieves only 42.2% recall on a test set containing 45 genuine R2 days, which is insufficient data to train a reliable discriminator for a rare regime. The R3 bond rotation strategy has a structural vulnerability in rising-rate environments: when stock-bond correlation turns positive and persists, the bond momentum filter reduces damage but cannot eliminate it. The 2022 experience is the clearest illustration of this.

Three design decisions in this work transfer across instruments without modification. The soft-label generation methodology (Section 4) applies whenever regime definitions can be mapped to observable contemporaneous features and ambiguous transition days need to be handled gracefully. The SHAP-based label audit (Section 6) is directly relevant to any multi-class financial classifier whose training targets are derived from forward returns, since return-contamination of the kind found in Method 2 will not appear in standard accuracy metrics. The soft-blending allocation mechanism (equation (11)) is a direct replacement for hard regime switching in any system producing calibrated probability outputs. The regime taxonomy itself (R1–R4, defined by VIX, credit spreads, and equity-market features) is calibrated to equity market conditions and would require redesign for different asset classes.

Two specific extensions are worth pursuing. The R2 (HighVolMR) regime is severely underrepresented in the 2021–2026 test period (45 days out of 1,224), making reliable discriminator training impractical within the current four-regime taxonomy. Merging R2 with an adjacent regime or constructing a targeted augmentation strategy for rare stressed-but-not-crisis periods are natural next steps. The R3 bond rotation strategy carries a structural vulnerability when the equity-bond correlation turns persistently positive, as it did through most of 2022; a correlation-conditioned switching rule within R3 (replacing or supplementing the current bond momentum filter) would reduce this exposure without requiring a full regime redefinition.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019), Optuna: A next-generation hyperparameter optimization framework, *in* 'Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 2623–2631.

Ang, A. (2014), *Asset Management: A Systematic Approach to Factor Investing*, Oxford University Press.

Ang, A. and Bekaert, G. (2002), 'International asset allocation with regime shifts', *Review of Financial Studies* **15**(4), 1137–1187.

Ang, A. and Bekaert, G. (2004), 'How regimes affect asset allocation', *Financial Analysts Journal* **60**(2), 86–99.

Avellaneda, M. and Lee, J.-H. (2010), 'Statistical arbitrage in the US equities market', *Quantitative Finance* **10**(7), 761–782.

Baele, L., Bekaert, G., Inghelbrecht, K. and Wei, M. (2020), 'Flights to safety', *Review of Financial Studies* **33**(2), 689–746. NBER Working Paper No. 19095 (2014).

Baltas, N. and Kosowski, R. (2013), 'Momentum strategies in futures markets and trend-following funds', *SSRN Working Paper No. 1982940* . Available at `https://ssrn.com/abstract=1982940`.

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.

Connolly, R., Stivers, C. and Sun, L. (2005), 'Stock market uncertainty and the stock-bond return relation', *Journal of Financial and Quantitative Analysis* **40**(1), 161–194.

Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *Annals of Statistics* **7**(1), 1–26.

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics* **29**(5), 1189–1232.

Gu, S., Kelly, B. and Xiu, D. (2020), 'Empirical asset pricing via machine learning', *Review of Financial Studies* **33**(5), 2223–2273.

Guidolin, M. and Timmermann, A. (2007), 'Asset allocation under multivariate regime switching', *Journal of Economic Dynamics and Control* **31**(11), 3503–3544.

Hamilton, J. D. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica* **57**(2), 357–384.

Ilmanen, A., Maloney, T. and Ross, A. (2023), 'Exploring macroeconomic sensitivities: Why does one inflation surprise flip the stock-bond correlation?', *Journal of Portfolio Management* **49**(2), 70–84.

Jegadeesh, N. and Titman, S. (1993), 'Returns to buying winners and selling losers: Implications for stock market efficiency', *Journal of Finance* **48**(1), 65–91.

Krauss, C., Do, X. A. and Huck, N. (2017), 'Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500', *European Journal of Operational Research* **259**(2), 689–702.

Kullback, S. and Leibler, R. A. (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**(1), 79–86.

López de Prado, M. (2018), *Advances in Financial Machine Learning*, Wiley.

Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems', Vol. 30.

Moskowitz, T. J., Ooi, Y. H. and Pedersen, L. H. (2012), 'Time series momentum', *Journal of Financial Economics* **104**(2), 228–250.

White, H. (2000), 'A reality check for data snooping', *Econometrica* **68**(5), 1097–1126.