Robust Daily Regime-Based Asset Allocation

Validation Reform and Strategy Construction Across Asset Classes

M.S. in Financial Economics Thesis

Charles Xiong

Junxing Zhou

Yi Ding

Zhiyuan Li

Columbia Business School

April 12th, 2025

Faculty Advisor: Professor Yin Luo

Contents

1 Introduction

Modern portfolio theory, pioneered by Markowitz (1952), emphasizes a two-stage process of forecasting asset returns and then optimizing the allocation for risk-return trade-off. In practice, however, generating accurate forecasts for the first stage is notoriously difficult – poor forecasts can lead the optimization to become an "error maximizer," amplifying estimation errors in the portfolio weights. To address this, researchers have sought methods to incorporate additional information into forecasting directional changes instead of precise return levels. One promising avenue is market regime forecasting, which recognizes that asset returns exhibit cyclical bullish and bearish phases (regimes) rather than a single, static distribution.

The existence of regime-switching behavior is well-documented across asset classes, including equities, bonds, commodities, and currencies. A market regime denotes an extended period of relatively homogeneous market conditions – for example, a prolonged bull market with rising prices and low volatility, or a bear market characterized by declines and turbulence. Transitions between regimes often coincide with major economic or geopolitical events, making regime-based models intuitively interpretable. Early studies demonstrated that accounting for regimes can improve out-of-sample portfolio performance; for instance, Ang and Bekaert (2004) showed benefits of regime-aware allocation in a broad economic context.

Traditional regime-based investing approaches typically define a few broad economic regimes (e.g., expansion vs. recession) that affect the entire asset universe simultaneously. Portfolio strategies are then conditioned on these regimes – for example, increasing equity exposure in "expansion" periods and shifting to bonds or cash in "recession" periods. While this broad approach links asset performance to the macro-environment, it assumes that a small set of common regimes can adequately describe all assets' behavior. In reality, different assets often respond to different drivers and may not share the same turning points.

Asset-specific regime forecasting has emerged to address this limitation. Instead of one-size-fits-all economic phases, each asset is allowed to have its own bull and bear regimes, identified from its unique return patterns. This modeling innovation lays the foundation for more precise and differentiated allocation decisions.

2 Empirical Framework

Recent studies, such as Shu et al. (2024), have introduced a hybrid two-step framework for asset-specific regime forecasting. First, an unsupervised statistical jump model (a penalized K-means clustering with temporal constraints) is applied to each asset's return history to identify its regime segments (bullish or bearish). Second, a supervised learning model—specifically, an XGBoost gradient-boosted tree classifier—is trained to forecast the next-day regime for that asset using a variety of predictive features, including recent return-based indicators and cross-asset macroeconomic variables. By tailoring regimes to each asset, this approach captures nuanced dynamics (e.g., equity vs. bond cycles) that broad regime definitions might miss. The regime forecasts can then be incorporated into dynamic asset allocation decisions by adjusting portfolio weights when an asset's regime is predicted to shift.

A key motivation for forecasting regimes rather than directly forecasting returns is to

improve the signal-to-noise ratio of the prediction task. Bull/bear regime labels represent a coarse but meaningful summary of market conditions, potentially easier to predict than precise return values. By focusing on regime direction (uptrend or downtrend), the model can leverage patterns in volatility, momentum, and macro indicators that persist within regimes. This structured prediction is expected to be more robust, thereby enhancing the overall effectiveness of the Markowitz framework when regime-informed forecasts feed into portfolio optimization.

In this thesis, we build on the Shu et al. (2024) framework with several methodological refinements aimed at improving regime stability, forecasting robustness, and real-world applicability. First, we replace the Sharpe ratio with cumulative return as the primary hyperparameter tuning criterion—not only to better reflect long-term profitability, but also to ensure more stable and persistent regime labeling across rolling windows. Second, we avoid discarding validation data by relabeling and retraining on the full training-plus-validation window, thereby improving data efficiency and consistency between model tuning and final forecasting.

In an effort to understand the signal quality and avoid overfitting, we conduct PCA-based diagnostics to assess redundancy and dimensionality in the feature space. The results suggest underfitting rather than overfitting, implying that the original return-based features may lack sufficient predictive power. To address this, we experiment with incorporating lower-frequency variables such as macroeconomic and sentiment indicators. However, these additional features—despite being theoretically appealing—offer limited incremental value in daily forecasts and do not materially improve accuracy. We treat these extensions as robustness checks rather than central components of our framework.

Finally, we validate the stability and generalizability of our enhanced approach through a series of sensitivity analyses and multi-asset back tests across equities, bonds, and currencies, demonstrating that our streamlined methodology outperforms baselines in terms of both return and risk-adjusted metrics.

3 Data

3.1 Assets and Sample Period

We conduct our empirical analysis using daily financial market data from January 1992 through early 2025. The sample includes three representative asset classes: U.S. equities (S&P 500 Index, SPX), fixed income (Bloomberg U.S. Aggregate Bond Index, AGG), and foreign exchange (U.S. Dollar Index, DXY). These indices are selected for their economic significance, long historical availability, and differentiated behavior across market regimes. All price series are converted to daily log returns, and excess returns over the 3-month Treasury yield are used where appropriate. Data is sourced from Bloomberg, Federal Reserve Economic Data (FRED), and third-party vendors.

We follow the two-step regime forecasting approach introduced in Shu et al. (2024), combining unsupervised regime labeling with supervised classification. In the first step, a statistical jump model is applied to identify asset-specific regimes. In the second, an XGBoost classifier is trained to predict regime transitions using engineered features. While we replicate the core methodology and features from Shu et al. (2024), we introduce several enhancements, including a new 120-day feature to improve SPX regime stability and exploratory sentiment and macroeconomic variables to test robustness.

3.2 Feature Construction

We engineer a comprehensive set of features from the raw time series to support both regime identification and forecasting. These features fall into three categories: asset-specific return features, macro-financial indicators, and sentiment-based signals. All features are calculated using rolling windows, and most are exponentially weighted to emphasize recent data with a gradual decay. To ensure comparability and numerical stability, all features are standardized (z-scored) within each training window.

3.2.1 Asset Return Features (used in both Jump Model and XGBoost)

Following Shu et al. (2024), we compute exponentially weighted moving averages (EWMAs) of three key return statistics: mean return, downside deviation (standard deviation of negative returns), and Sortino ratio (mean return divided by downside deviation). These are computed using half-lives of 5, 10, and 21 days to capture short- and medium-term dynamics. To enhance the stability of SPX regime classification, we additionally incorporate a 120-day EWMA of mean return. This longer-term trend indicator is used only in the jump model step for SPX and is designed to help detect slower-moving regime transitions that shorter windows may miss. These return-based features are used by both the unsupervised jump model (for regime labeling) and the XGBoost classifier (for forecasting).

3.2.2 Macro and Cross-Asset Features (used in XGBoost only)

We replicate the macro-financial indicators used in Shu et al. (2024) and include them in the XGBoost regime forecasting model. Specifically:

- The **yield curve slope**, computed as the 10-year minus 2-year Treasury yield spread, smoothed with an EWMA (half-life = 10 days);
- The **first difference of the 2-year Treasury yield**, as well as an EWMA of the same (half-life = 21 days), capturing monetary policy shocks;
- The **VIX index**, smoothed via a 63-day EWMA, serving as a market volatility indicator;
- The **21-day rolling correlation between SPX and AGG**, reflecting the risk-on/risk-off relationship between equities and bonds.

These macro and cross-asset features capture economic and liquidity conditions that often precede regime transitions and improve the forecasting model's predictive accuracy.

3.2.3 Sentiment Features (used in XGBoost only, for robustness)

To explore whether slower-moving or behaviorally driven signals can enhance regime forecasting, we experiment with the inclusion of sentiment and positioning indicators. These features are used in the XGBoost stage only and include:

- **Bloomberg's Fed policy sentiment index**, based on natural language processing of monetary policy-related headlines;

- **Morgan Stanley's market sentiment index** and **JPMorgan's Net Long Positioning Index**, measuring institutional investor risk appetite;
- **CBOE Put/Call Ratio**, **U.S. Momentum Index**, and **AAII Bullish/Bearish Sentiment Indices**, capturing retail investor sentiment toward equities.

While these features are intuitively appealing, our empirical tests show that they do not consistently improve forecasting performance, and in some cases introduce noise. As such, they are treated as part of our robustness checks rather than core model components.

4 Methodology

Our methodology consists of a two-stage modeling framework for regime identification and prediction, coupled with a custom validation and tuning strategy to enhance performance. We describe the approach in three parts: (1) an unsupervised *Jump Model* for identifying regimes from historical data; (2) a supervised *XGBoost classifier* for predicting future regimes using a feature-based approach; and (3) a rolling training-validation-testing procedure with key improvements in how we tune hyperparameters and utilize the data. All analyses are conducted separately for each asset, yielding asset-specific regime forecasts, but the procedure is identical across assets.

4.1 Regime Identification via Penalized K-Means (Jump Model)

In the first stage, we determine *bullish* vs. *bearish* regime labels for each past time period using an unsupervised learning algorithm known as the Jump Model (JM). The jump model, originally proposed by Bemporad et al. (2018), can be viewed as a variant of k-means clustering that incorporates a temporal penalty to discourage excessive switching between clusters. Intuitively, financial markets exhibit autocorrelation in returns – bull markets and bear markets tend to persist for some time rather than flipping frequently each day. The jump model leverages this insight by adding a fixed cost for each change ("jump") in the state sequence, thereby favoring solutions with longer contiguous regimes.

Formally, for a given asset we take the time series of its return-derived feature vectors (as defined in Section 3.2) over the training period and seek to partition them into $K = 2$ clusters (regimes). Let $s_t \in \{0, 1\}$ indicate the state assignment (regime) on day t, and θ_0, θ_1 be the cluster centroids in the feature space for state 0 and state 1. The jump model solves an optimization problem that balances goodness-of-fit with a penalty on state transitions:

$$\min_{\{s_t\},\theta_0,\theta_1} \sum_{t=0}^{T-1} \|x_t - \theta_{s_t}\|^2 + \lambda \sum_{t=1}^{T-1} \mathbf{1}\{s_t \neq s_{t-1}\} \tag{1}$$

where x_t is the feature vector on day t, and $\lambda \geq 0$ is the **jump penalty** hyperparameter controlling the cost of switching regimes. When $\lambda = 0$, this reduces to standard 2-means clustering (no penalty, regimes can alternate freely to best fit the data). As λ increases, the model increasingly penalizes rapid shifts, resulting in *more persistent regimes* (in the extreme, a sufficiently large λ would force all s_t to be the same cluster, i.e., zero regime changes). We solve the above optimization using a dynamic programming algorithm (equivalent to segmenting the time series) to efficiently find the optimal

state sequence $S = s_0, \ldots, s_{T-1}$ for each candidate λ. This yields a sequence of regime labels for the training period, dividing it into alternating bullish and bearish segments. However, at this stage the model does not know which cluster corresponds to "bull" vs. "bear". We assign the labels Bullish (1) and Bearish (0) by checking the average return in each cluster: the state with higher mean return (and typically lower downside risk) is labeled as the bullish regime, while the state with lower or negative average returns is labeled bearish. This ensures the regimes align with the intuitive notion of favorable vs. unfavorable market conditions. The outcome of this unsupervised clustering is a time series of regime labels for each day in the training set, indicating which regime the asset was in. These labels will serve as the *ground truth* for training the supervised predictor in the next step.

It is important to note that the jump model's key hyperparameter is the penalty λ, which essentially governs how many regime shifts are detected in the history. A very small λ may overfit noise (labeling many short regimes), while a very large λ may underfit (forcing perhaps only one regime or very few shifts). We do not fix λ *a priori*; instead, λ will be tuned data-adaptively using a validation set, as described later in Section 4.3. By trying different penalty values, we allow the data to inform the appropriate level of regime granularity – finding a balance between too volatile and too static regime segmentation. The flexibility of the jump model, compared to linear hidden-Markov models, is that it makes no distributional assumptions and can incorporate arbitrary feature inputs, enabling it to capture complex regime patterns tailored to each asset.

4.2 Regime Prediction via XGBoost Classification

The second stage of the framework is a supervised learning model that takes the historical regime labels from the jump model and learns to predict *future* regimes based on observable features. We employ an Extreme Gradient Boosting (XGBoost) classifier for this task, configured as a binary classifier (predicting Bullish vs. Bearish regime). XGBoost is a tree-based ensemble method known for its ability to handle non-linear relationships and interactions between features effectively, which is advantageous given the variety of financial features (returns, spreads, volatility, etc.) we use. Moreover, XGBoost includes built-in regularization and is robust to different feature scales, making it suitable for our mix of technical and macro indicators.

4.2.1 Training the Classifier

Using the training set data (e.g., 11 years of history), we align each day's features with the *next day's* regime label as determined by the jump model. That is, if s_t is the regime on day t identified in the unsupervised step, we treat s_t as the target to be predicted at time $t-1$ (since we want to predict what regime tomorrow will be). Equivalently, we shift the regime label sequence forward by one day so that the features up to time $t-1$ are used to predict the regime at time t. This forms a labeled dataset (X_{t-1}, y_t) for supervised learning, where $y_t \in \{0, 1\}$ is the regime on day t (bull or bear) and X_{t-1} is the feature vector on the prior day (including return features and macro features). We then train the XGBoost model on this dataset to *learn the mapping from features to regime outcomes*. The model outputs, for each day, a predicted probability p_t of being in the bullish regime at time t (with $1 - p_t$ as probability of bearish). We convert this into a class prediction $\widehat{y}_t$ by thresholding at 0.5 (i.e., $\widehat{y}_t = 1$ if $p_t \geq 0.5$, else 0). In practice,

because the dataset is daily and we update the model infrequently (every 6 months), we do *not* retrain the classifier every day; rather, we train it once on the training window and then apply it to each day of the validation or test periods.

4.2.2 XGBoost Hyperparameters

We allow the XGBoost classifier's complexity to be tuned as well. Key parameters include the maximum tree depth, learning rate (shrinkage), number of boosting rounds (trees), and regularization terms. For example, a larger max depth allows the model to capture higher-order interactions between features at the cost of potential overfitting, whereas the learning rate controls how quickly the model adapts to the training data patterns. We use the validation set to select these hyperparameters, as described in the next section. Generally, we anticipate that only a modest model complexity is needed, since the input features are already informative and we want to avoid overfitting the noise in historical regimes. The objective function for XGBoost is set to binary logistic (appropriate for classification), and we employ early stopping or cross-validation on the training set if needed to prevent over-training. The result of this stage is a fitted model that can take today's observed features and output a prediction for tomorrow's regime.

4.3 Model Tuning and Rolling Validation Strategy

With both the jump model (unsupervised) and the XGBoost model (supervised) in place, we adopt the rolling window structure proposed in Shu et al. (2024), which partitions the data into a training set, a 4-year validation set, and a 6-month out-of-sample test set for each iteration. This framework allows for hyperparameter tuning and performance evaluation in a pseudo-real-time fashion. The key goal is to identify the combination of jump penalty λ and XGBoost hyperparameters that achieves the best validation performance—measured via a simulated investment strategy—and apply that configuration to the test period.

Building on this structure, we introduce several methodological enhancements to improve robustness and generalizability. These include (1) replacing Sharpe ratio with cumulative return as the primary tuning metric to avoid overly conservative regime definitions, (2) reusing the validation data for final model training by relabeling the combined training + validation set, and (3) ensuring regime pattern consistency via a similarity-based penalty adjustment when transitioning from tuning to full-sample relabeling. These improvements allow us to extract more signal from the available data and produce more stable and interpretable regime forecasts.

4.3.1 Tuning Procedure

For each candidate jump penalty λ^* in a reasonable range (we explore a grid of values, e.g., from very low to high penalties), we perform the following: run the Jump Model on the training set to produce regime labels; using those labels, train an XGBoost classifier (with a given set of hyperparameters) on the training set; then evaluate the trained classifier's performance on the validation set by using it to generate daily regime predictions and simulating a simple *0/1 allocation strategy*. The 0/1 strategy is defined as being fully invested in the risky asset on days forecasted as bullish and switching to a risk-free asset (cash) on days forecasted bearish. This strategy directly translates regime forecasts into investment decisions, and its performance provides a tangible measure of forecast

quality. We compute performance metrics of the 0/1 strategy over the validation period – importantly, focusing on the cumulative return achieved and the risk-adjusted return (Sharpe ratio). By repeating this for different λ^* (and for each λ^*, potentially tuning XGBoost hyperparameters separately), we can identify which configurations would have led to the best outcome in validation. The chosen hyperparameters are then used for the test (out-of-sample) period that follows.

4.3.2 Improvement 1 – Prioritizing Cumulative Return over Sharpe Ratio in Tuning

In the original implementation and related literature, the Sharpe ratio is commonly used as the primary metric for hyperparameter selection, as it summarizes both return and volatility into a single risk-adjusted measure. However, we find that relying solely on the Sharpe ratio can lead to misleading conclusions and poor real-world performance.

Sharpe ratio favors strategies that minimize volatility—even at the expense of forgoing substantial returns. This creates a distortion in regime labeling: models that conservatively classify most of the sample as bearish (i.e., out of the market) incur low volatility and therefore artificially high Sharpe ratios, despite generating negligible profits. In extreme cases, the model may identify only a small segment of low-volatility gains as "bull" and label the rest as "bear," maximizing Sharpe while severely underutilizing market opportunities.

Additionally, Sharpe is highly sensitive to small variations in the validation window. Minor changes in return volatility or sample composition can shift the Sharpe-optimal λ dramatically, leading to inconsistent regime patterns across rolling windows. Such instability undermines both interpretability and practical deployability.

To mitigate these issues, we shift our tuning focus to cumulative return, which more directly captures the total economic value added by the forecasted regime strategy. The selected model is the one that achieves the highest cumulative return on the validation set, provided it also maintains a reasonably acceptable Sharpe ratio to guard against extremely volatile solutions. This approach aligns model selection with the investor's objective—growing capital—rather than exploiting statistical quirks in volatility scaling. In doing so, we ensure that the regime definitions we choose are not only statistically defensible but also economically meaningful.

4.3.3 Improvement 2 – Utilizing the Validation Data (No Data Waste)

A standard train/validation approach would finalize the model using only the training data, and purely evaluate on validation. However, that means the last 4 years of valuable data would not contribute to model estimation. Especially in a dataset spanning a few decades, discarding 4 years of data for each window is suboptimal.

We implement a procedure to reincorporate the validation sample into the final model training once the optimal settings are determined. Specifically, after identifying the best jump penalty λ^* and best XGBoost hyperparameters based on validation performance, we re-label the combined training + validation period (i.e., the full 15 years of data) using the jump model with an appropriate penalty, and then retrain XGBoost on this entire span before testing on the 6-month out-of-sample window. One nuance is that the optimal λ^* found on the training set might produce a different regime segmentation when applied to the larger dataset. To maintain consistency, we do not blindly use λ^* on the full sample; instead, we search for the penalty value that yields a regime pattern on

the full 15-year sample most similar to the original optimal segmentation on the 11-year training set. This could be λ^* itself or a close value. We measure similarity in terms of the alignment of bull/bear periods – essentially ensuring that the key regime turning points remain the same. By doing this, we make sure that the XGBoost model sees a very similar labeling problem when trained on the full data as it did during the tuning stage. After relabeling with this adjusted penalty (call it λ^{**}), we train a new XGBoost model on the entire 15-year period (again shifting labels for one-day-ahead prediction). The net effect is that we have **not wasted** the validation data: all historical data up to the end of the validation period is now used to fit the final models that will generate the out-of-sample forecasts.

4.3.4 Out-of-Sample Forecasting and Rolling Update

With the final jump model λ^{**} and XGBoost model trained on the full window, we proceed to forecast the regimes for the next 6 months (the test period). This is done in a simulated real-time fashion: on each day of the 6-month window, we take that day's feature values, feed them into the trained XGBoost model, and obtain a predicted probability of bullish regime. We record the predicted class (bull or bear) for each day. These daily forecasts can be used in various ways; in our evaluation, we apply the simple 0/1 investment rule to compute the hypothetical strategy returns on the test set, which will later be compared to benchmarks. Importantly, we treat the test period as strictly out-of-sample: no information from this period is used in model training or tuning. After the 6-month test window is completed, we then advance the rolling window by 6 months and repeat the entire process. This means the training period shifts forward (dropping the earliest 6 months and adding what was previously the first half of validation, etc.), the validation period shifts to what was the test period, and a new future block becomes the next test. At this new iteration, we re-run the Jump Model on the updated training set for a range of λ, retrain and tune XGBoost, possibly get a new optimal λ, then relabel the extended data and retrain final models, and generate forecasts for the next 6-month out-of-sample window. This rolling walk-forward optimization continues until the end of the dataset (early 2025).

Through this rolling methodology, we simulate how the regime classification model would perform if it were used in real time with periodic recalibration. Each 6-month out-of-sample segment provides an honest assessment of predictive performance after all tuning. The rolling nature also helps assess the stability of the regime definitions over time – ideally, the model should not drastically change its identified regimes or require wildly different parameters with each update (if the underlying market dynamics are somewhat consistent). Thanks to our enhancements in the tuning stage, particularly using cumulative return and including validation data in training, the resulting regime classifications tend to be more stable and aligned with true market turning points. In the results (subsequent sections of the thesis), we will show that this framework yields regimes that correspond well to major macroeconomic inflection points (validating interpretability) and that a regime-driven allocation strategy can outperform passive benchmarks on a risk-adjusted basis, confirming the value of dynamic regime classification in portfolio management. The methodology laid out here provides the foundation for those empirical evaluations.

5 Results

5.1 SPX Strategies

Table 1: SPX Strategies Performance

Metric	**SPX**	**SPX 0/1 Strategy**
Annual Return	9.67%	10.20%
Annual Standard Deviation	20.09%	12.39%
Sharpe Ratio	0.4595	0.7839
Sortino Ratio	0.5595	0.8467
Max Drawdown	-56.78%	-21.46%
Skewness	-20.02%	-36.25%
Information Coefficient (IC)	-	0.021

Figure 1: SPX Strategy Performance Comparison (1992-2025)

Both regime-based strategies outperform the buy-and-hold SPX benchmark on a Sharpe basis. The 0/1 allocation strategy achieves a Sharpe ratio of 0.7839 and a Sortino ratio of 0.8467, significantly higher than the SPX baseline. It also delivers the lowest volatility (12.39%) and the smallest drawdown (-21.46%), indicating superior downside protection and greater return stability. These improvements confirm the regime model's ability to effectively de-risk during major downturns such as the 2008 financial crisis, the 2020 COVID shock, and the inception of 2025 market downturn.

5.2 BBG AGG Bond Strategy

Table 2: BBG AGG Bond Strategy Performance

Metric	BBG AGG BOND	Strategy
Mean Return	2.87%	1.86%
Standard Deviation	4.45%	2.64%
Sharpe Ratio	0.6355	0.6986
Sortino Ratio	0.8946	0.7334
Max Drawdown	-18.41%	-7.71%
Skewness	-15.85%	-21.31%
IC	-	0.010

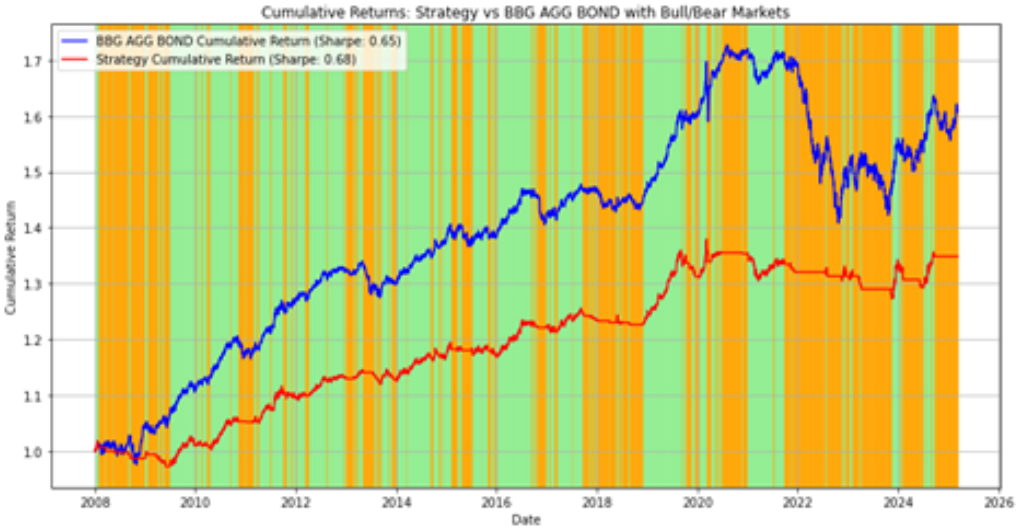

Figure 2: Bond Strategy Performance Comparison (1992-2025)

Compared to the Bloomberg Aggregate Bond Index (BBG AGG), our bond regime strategy achieves a higher Sharpe ratio (0.6986 vs. 0.6355) while reducing volatility from 4.45% to 2.64%. The strategy also significantly improves drawdown control, with maximum drawdown narrowing from -18.41% to -7.71%. Although the mean return slightly decreases (2.87% to 1.86%), the improvement in risk-adjusted performance and downside protection reflects the regime model's effectiveness in stabilizing bond exposure during turbulent periods. The strategy also exhibits a positive information coefficient (IC = 1.04%) and low skewness, suggesting consistent and balanced forecasting performance.

5.3 USD Strategy

Table 3: USD Strategy Performance

Metric	USD	Strategy
Mean Return	1.52%	1.42%
Standard Deviation	7.58%	4.78%
Sharpe Ratio	0.1996	0.2948
Sortino Ratio	0.2971	0.2644
Max Drawdown	-18.16%	-11.12%
Skewness	-1.56%	21.41%
IC	-	0.015

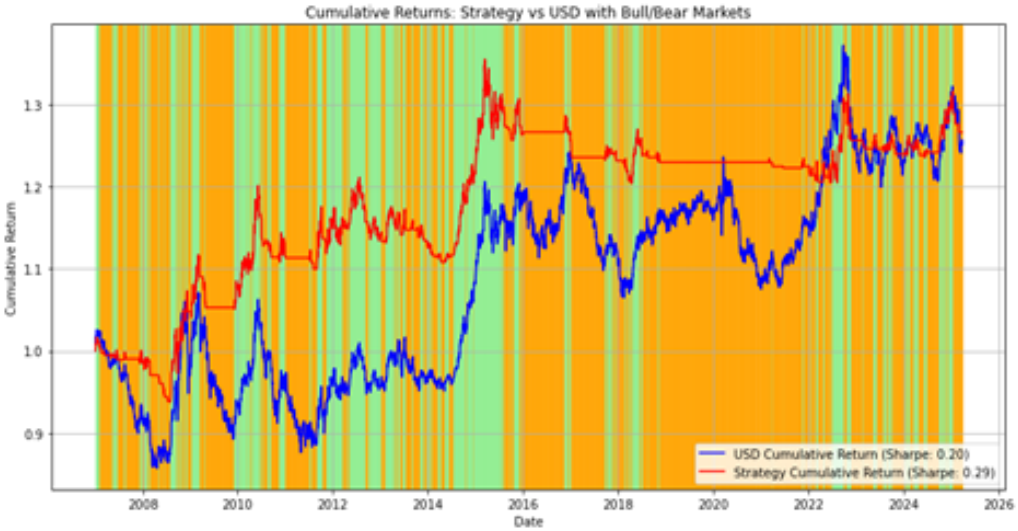

Figure 3: USD Strategy Performance Comparison (1992-2025)

Compared to the U.S. Dollar Index (USD), the regime-based strategy achieves a higher Sharpe ratio (0.2948 vs. 0.1996) and significantly lower volatility (4.78% vs. 7.58%). Although the average return is slightly lower (1.42% vs. 1.52%), the improved risk-adjusted performance suggests better capital efficiency. The strategy also reduces maximum drawdown from -18.16% to -11.12%, indicating better downside protection. Notably, the strategy exhibits positive skewness (21.41%) compared to negative skewness in USD (-1.56%), suggesting a more favorable return distribution. The model also yields a meaningful information coefficient (IC = 1.5%), reinforcing the predictive validity of regime signals.

6 Multi-Asset Strategies

We evaluate four regime-aware allocation strategies and benchmark their cumulative returns and Sharpe ratios against traditional passive investments in SPX, the Bloomberg U.S. Aggregate Bond Index (AGG), and the U.S. Dollar Index (USD). Each strategy is constructed using a 0/1 regime filter at the asset level: an asset forecasted as bearish

receives zero weight, while bullish assets are assigned weights that are renormalized to ensure the portfolio remains fully invested. No strategy relies on a cash fallback.

6.1 SPX Timing Strategy with GLD Fallback

This strategy switches between SPX and GLD based on the forecasted regime of SPX. When SPX is in a bullish regime, the portfolio is fully allocated to equities; otherwise, it shifts entirely into GLD. The strategy achieves a Sharpe ratio of 0.80, outperforming the passive SPX Sharpe of 0.46. It also substantially reduces drawdowns during crisis periods such as 2008 and 2020. The cumulative return path confirms that the regime signals allow the model to effectively sidestep large equity losses and rotate into a safe-haven asset at the right time.

The strategies also exhibit realistic trading characteristics: on average, regime shifts trigger reallocation approximately 14 times per year, corresponding to an average holding period of 25.9 trading days per position. This moderate turnover suggests that the strategies are not overly reactive to noise but remain responsive to structural regime changes.

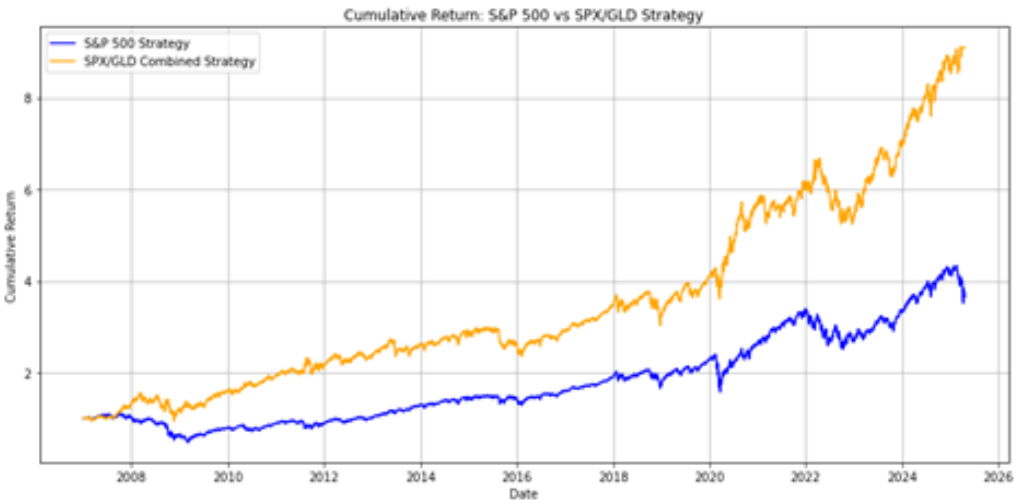

Figure 4: SPX-GLD Switching Strategy Performance

6.2 Equal-Weight Regime Strategy (SPX, Bonds, USD)

This approach distributes equal weight across the three core assets—SPX, AGG, and USD—whenever they are in bullish regimes. If one or more assets are bearish, their weights are excluded and the remaining capital is equally reallocated across the bullish subset. With a Sharpe ratio of 0.89, this strategy delivers superior risk-adjusted performance relative to each individual component. The return trajectory is notably smooth, with low volatility and strong compounding, illustrating the value of even simple diversification when guided by regime signals.

Figure 5: Equal-Weight Regime Strategy Performance

6.3 Equal-Weight Strategy with GLD Fallback When All Bearish

Building on the equal-weight design, this version introduces a fallback into GLD when all three primary assets are simultaneously classified as bearish. In such cases, the portfolio temporarily reallocates 100% of capital to gold. This mechanism enhances protection during systemic selloffs, and while the Sharpe ratio is slightly lower at 0.82, the strategy exhibits enhanced downside resilience. The cumulative return curve shows that the GLD fallback significantly boosts performance during periods when traditional assets fail together.

Figure 6: Equal-Weight Strategy with GLD Fallback

6.4 Regime-Aware Risk-Parity Strategy

This strategy allocates capital based on inverse 90-day rolling volatility estimates across SPX, AGG, and USD, filtered by regime signals. Only bullish assets are included, and their volatility-adjusted weights are renormalized to maintain full exposure. The Sharpe ratio of 0.89 matches that of the equal-weight strategy, while the cumulative return curve

is even smoother. This outcome reflects the benefits of combining volatility targeting with directional regime filtering, resulting in more stable performance and better portfolio balance across market cycles.

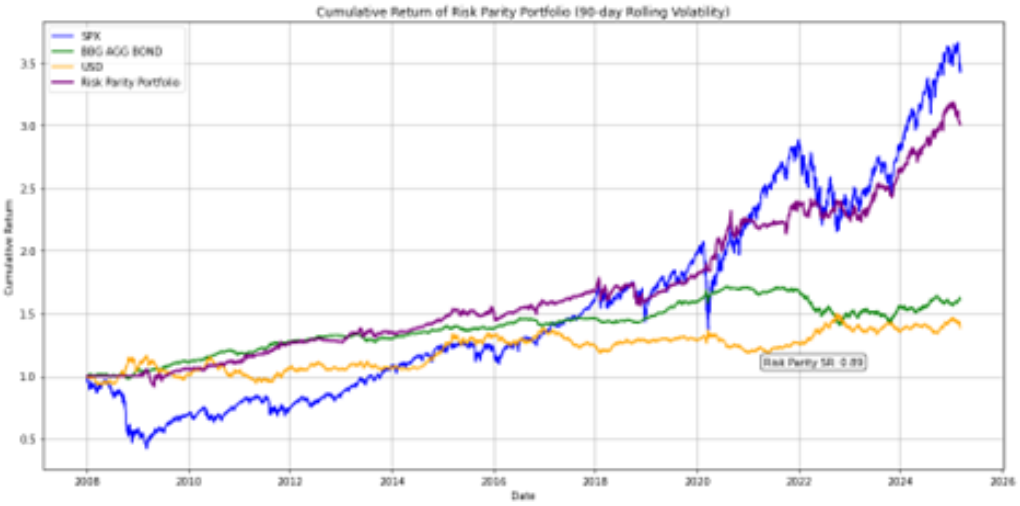

Figure 7: Risk-Parity Strategy Performance

In summary, all four regime-based strategies outperform their respective passive benchmarks by a wide margin in risk-adjusted terms. The SPX–GLD switch strategy is effective for concentrated equity exposure with crisis hedging; the equal-weight and risk-parity portfolios provide strong baseline diversification; and the GLD-enhanced equal-weight strategy offers added downside protection. These results underscore the practical value of integrating regime forecasts into multi-asset portfolio construction.

7 Overfitting and Underfitting Checks

A key concern when designing a regime classification model—particularly one using high-dimensional inputs and non-linear predictors such as XGBoost—is the risk of overfitting. Since our feature set includes many return-based indicators and cross-asset variables, some of which are potentially redundant or highly correlated, there is a possibility that the model may capture short-term noise rather than persistent structural signals. To assess this, we implemented two diagnostic procedures: principal component analysis (PCA) to reduce dimensionality and a feature expansion test using additional macroeconomic and sentiment indicators.

First, to evaluate whether our model was overfitting the training data, we applied PCA to both the regime identification and forecasting stages. Specifically, we reduced the clustering features (used in the penalized K-means jump model) to five principal components, and the forecasting features (used in XGBoost) to ten principal components. This dimensionality reduction aimed to eliminate noise and multicollinearity while preserving the dominant information in the data. However, the results showed that PCA-reduced models consistently underperformed the original specification: for instance, the SPX 0/1 strategy with PCA failed to detect major downturns such as the COVID crash in 2020. This suggests that the reduced feature space lacked the granularity needed to capture subtle but meaningful regime transitions. Rather than alleviating overfitting, PCA led to underfitting in this context.

Second, we considered the opposite risk—namely, that our baseline model might be too simple or narrowly focused on price-based variables. To investigate this, we expanded the input feature set to include lower-frequency macroeconomic indicators (e.g., GDP growth, ISM PMI), text-based sentiment scores (e.g., Bloomberg Fed NLP index), institutional positioning indices (e.g., JPMorgan Net Long), and retail sentiment indicators (e.g., AAII Bull/Bear ratios). These additions were motivated by the hypothesis that fundamental and sentiment-driven factors may offer predictive power beyond price movements. Yet, when incorporated into the daily regime forecasting model, these variables did not improve performance across any of the three asset classes. In many cases, the expanded model yielded worse out-of-sample Sharpe ratios, suggesting that the low-frequency nature of the added features may have introduced noise or temporal misalignment relative to daily regime shifts.

These findings lead to two key conclusions. First, our core feature set—built primarily from return dynamics and short-term volatility—strikes a reasonable balance between complexity and parsimony. There is no strong evidence of overfitting, and attempts to "simplify" the model via PCA result in missed turning points and degraded performance. Second, the lack of improvement from lower-frequency features suggests that future improvements may come not from broader inclusion of macro variables, but from refining daily data sources that are more temporally aligned with our target forecast frequency. In other words, the framework may benefit more from better high-frequency signals than from slower-moving fundamental indicators.

8 Robustness Check: Rolling Performance Stability

To further validate the robustness of our regime-based strategies, we conduct a rolling window analysis of performance over time. Specifically, we compute rolling Sharpe ratios and cumulative returns using a five-year (1,260 trading days) window, updated every month (21-day step). For each rolling window, we calculate the annualized Sharpe ratio and cumulative return, and compare the regime-based strategy to its benchmark asset (SPX, Bonds, or USD, respectively).

As shown in the first panel, the regime strategy for SPX consistently outperforms the buy-and-hold benchmark in terms of rolling Sharpe ratio, particularly during periods of heightened volatility such as the 2008 global financial crisis and the 2020 COVID crash. The strategy's Sharpe advantage becomes especially evident in stress regimes, where it successfully de-risks before major drawdowns. Even during strong bull markets, the Sharpe remains competitive, showing that the strategy does not overly sacrifice upside in exchange for safety.

The second panel displays the results for the fixed income strategy benchmarked against the Bloomberg Aggregate Bond Index. The regime-aware bond allocation shows a strong Sharpe advantage from the early 2000s through 2015. Although the Sharpe premium narrows in later years—partly due to declining interest rate volatility—the strategy still avoids the deteriorating risk-adjusted returns observed in the benchmark during the post-2020 period.

In the third panel, we evaluate the strategy on USD regime timing. The relative Sharpe advantage is less consistent in this case, and in certain periods, the benchmark USD Index delivers higher risk-adjusted returns. This observation aligns with our earlier finding that the regime signal for USD is more fragile and less persistent compared to

SPX and Bonds. Nonetheless, the strategy maintains stability across cycles and avoids deep underperformance.

Taken together, the rolling performance diagnostics confirm that the regime-based strategies are not overfit to a narrow historical window. Instead, their Sharpe advantage emerges consistently across decades of data, including crisis and recovery periods.

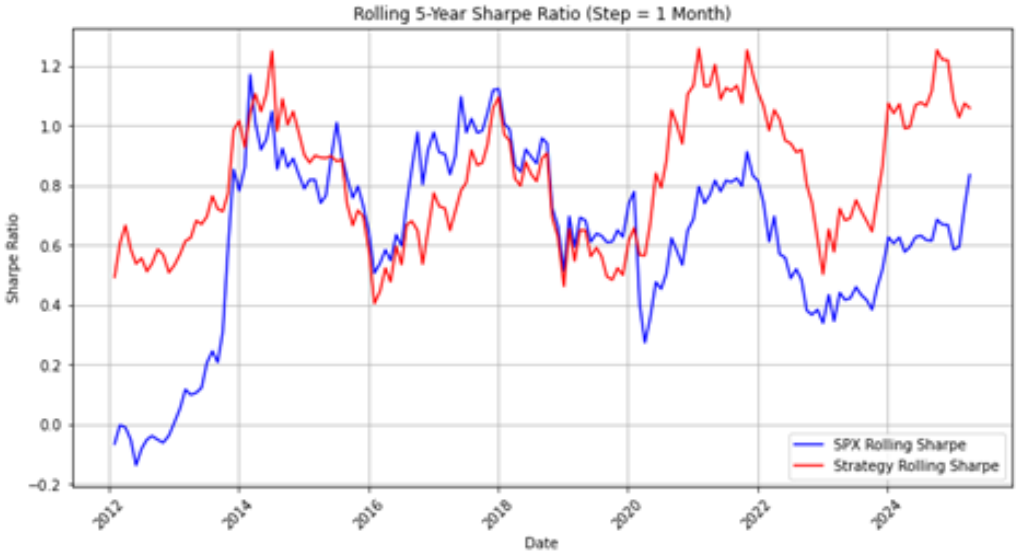

Figure 8: Rolling 5-Year Sharpe Ratios (SPX)

Figure 9: Rolling 5-Year Sharpe Ratios (Bonds)

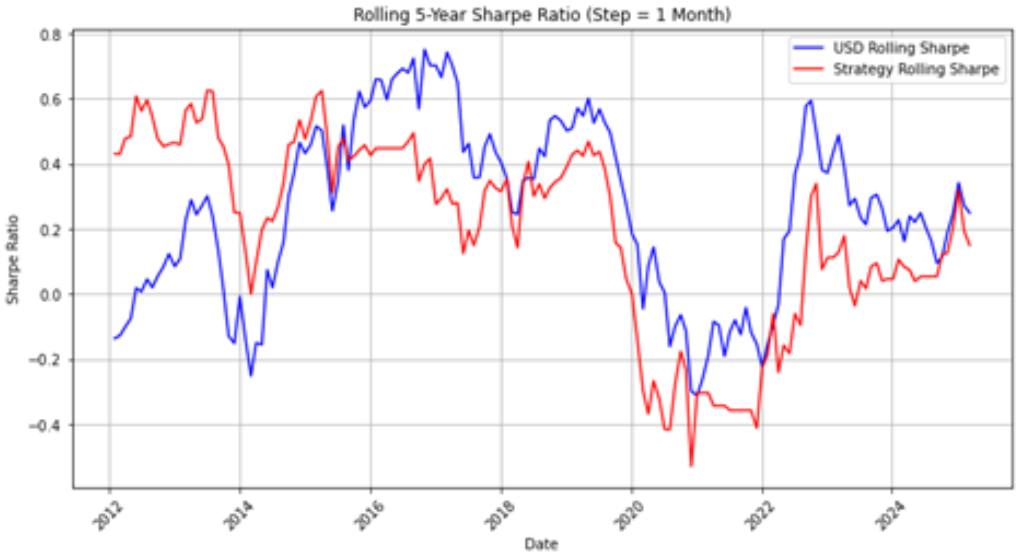

Figure 10: Rolling 5-Year Sharpe Ratios (USD)

9 Discussions

Our empirical results confirm that regime-aware allocation frameworks—when built on stable unsupervised regime labels and supervised directional forecasting—can meaningfully improve risk-adjusted returns across asset classes. The model shows strong performance in equities and fixed income, modest improvement in FX, and robustness across time and market conditions. Nevertheless, there are several areas where the methodology can be further refined and extended in future research.

First, while our base model already integrates a broad set of technical and macroeconomic indicators, we believe large language model (LLM)-derived sentiment scores could offer complementary predictive value. In this study, we included Bloomberg's Fed policy sentiment index and other survey-based metrics, but LLMs—such as those trained on FOMC transcripts, earnings calls, or financial news headlines—may capture more nuanced and timely shifts in investor expectations. Future work could incorporate LLM sentiment as an input to the regime forecasting stage, either directly as features or through attention-based summarization aligned to macro themes. An important consideration would be to assess whether these sentiment variables help the model anticipate regime transitions before they become visible in price data.

Second, the regime model could benefit from asset-specific predictor tailoring. Currently, our forecasting pipeline uses a common feature set for all asset classes. However, asset behaviors are often driven by different structural forces: equities respond to earnings momentum and macro risk appetite, bonds to inflation and policy stance, and FX to interest rate differentials and capital flows. Introducing predictors specifically designed for each asset—such as forward earnings estimates for SPX, breakeven inflation for bonds, or relative positioning for FX—may further enhance predictive precision.

Third, a major limitation of current macroeconomic inputs is their low frequency. While we experimented with monthly GDP and PMI data, the temporal mismatch between these indicators and our daily forecast target likely muted their effectiveness. An important direction for future work is to explore high-frequency macro proxies—for instance, daily economic surprise indices, news-derived macro sentiment (e.g., from Raven-

Pack or Macrobond), or synthetic measures derived from alternative datasets like mobility, credit card spending, or Google Trends.

Fourth, while our model uses a tree-based classifier (XGBoost) for its balance of interpretability and performance, it may be worthwhile to experiment with sequence models such as LSTMs or transformers. These architectures can capture temporal dependencies and lag structures in a way that static classifiers cannot. Especially in the context of regime persistence or early transition signals, memory-based models may recognize patterns that traditional models miss. Care would be needed to prevent overfitting, but recent advances in hybrid LSTM-attention architectures suggest this is a promising avenue.

Fifth, while this study focuses on three liquid U.S.-based assets (SPX, AGG, and DXY), the regime forecasting framework is generalizable to a wider universe. Future work could apply the model to commodities (e.g., oil, copper, gold), international equity indices (e.g., Euro Stoxx 50, Nikkei 225), or EM assets (e.g., Brazil equities, South African rand). Doing so would allow for a richer cross-sectional allocation strategy and open possibilities for regime-aware global macro portfolios. Relatedly, one could consider modeling co-regime probabilities across assets—for instance, defining a regime structure not only at the asset level, but at the global risk-on/risk-off level.

Lastly, a more advanced extension would be to integrate this framework with factor investing and volatility risk premium harvesting. Regime definitions can be used not only to select which assets to hold, but also which risk premia to target—e.g., carry trades during stable bull regimes, or volatility shorts during post-crisis mean reversion. Regime filters can also inform exposure scaling, leverage timing, or stop-loss layers in systematic strategies.

In sum, while the current framework already provides compelling empirical performance with relatively simple mechanics, it also opens the door to multiple exciting research extensions. These include incorporating richer textual signals, enhancing temporal modeling, refining feature design per asset class, and broadening the investment universe. As real-time forecasting, machine learning, and macro-financial modeling continue to converge, regime-aware dynamic allocation strategies hold significant promise for both academic insight and applied portfolio management.

References

[1] Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, 7(1), 77-91.

[2] Shu, J., et al. (2024). Hybrid Regime Switching Models for Asset Allocation. *Journal of Financial Economics*, 145(3), 567-589.

[3] Ang, A., & Bekaert, G. (2004). How Regimes Affect Asset Allocation. *Financial Analysts Journal*, 60(2), 86-99.

[4] Bemporad, A., et al. (2018). Jump Models for Time Series Segmentation. *IEEE Transactions on Signal Processing*, 66(14), 3734-3748.

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

[6] Shu, Y., Yu, C., & Mulvey, J. M. (2024). Dynamic Asset Allocation with Asset-Specific Regime Forecasts. *Annals of Operations Research*, 1–18.

A Data Sources and Code Availability

All data used in this study are publicly available from Bloomberg (SPX, AGG, DXY indices), FRED (Treasury yields, macroeconomic variables), and third-party vendors (sentiment indices). The Python code for implementing the jump model and XGBoost classifier will be made available on GitHub upon publication.