Evaluating Predictive Confidence and Regime Conditioning in Equity Return Models

Hannah Attar

November 26, 2025

Abstract

This study examines how predictive confidence and market regimes interact to shape the economic usefulness of short-horizon equity return forecasts. Using daily data for the S&P 500 ETF (SPY), a probabilistic classification model is trained to predict the direction of next-day returns based on recent returns and realized volatility. Market regimes are identified endogenously using an unsupervised Gaussian Mixture Model, allowing predictive behavior and strategy performance to be evaluated conditionally across statistically distinct market states. While short-horizon directional prediction remains challenging and predicted probabilities are concentrated near the classification threshold, the model's outputs exhibit meaningful variation across regimes. Translating predictions into a simple directional trading strategy shows that selectively deploying model signals and disabling execution in structurally adverse regimes can improve risk-adjusted performance and reduce drawdowns relative to unconditional deployment. The results highlight the importance of separating prediction from deployment and suggest that regime-aware execution rules may play a more important role than increased model complexity in applied financial machine learning settings.

1. Introduction and Problem Formulation

Predictive modeling in financial markets is complicated by the inherent nonstationarity of asset return dynamics and the weak signal-to-noise ratio present at short horizons. While machine learning models are often evaluated using statistical classification metrics such as accuracy or area under the ROC curve, such measures do not necessarily translate into economic value when deployed in trading or risk-management settings. In particular, predictive confidence and classification performance may vary substantially across market environments, leading to strategies that perform well on average but fail during specific market conditions. Understanding when a model is reliable, and when it should be withheld from deployment, is therefore as important as improving raw predictive accuracy.

This project investigates the relationship between model confidence, market regimes, and economic performance in the context of short-horizon equity return prediction. Using daily data for the S&P 500 ETF (SPY), a simple and interpretable classification model is trained to predict the direction of next-day returns based on recent returns and realized volatility. Rather than focusing on maximizing predictive accuracy through model complexity, the analysis emphasizes the conditional behavior of model performance across statistically identified market regimes. Market regimes are inferred directly from the data using an unsupervised learning approach, allowing the structure of market states to emerge endogenously rather than being imposed ex ante through heuristic thresholds.

The central objective of this study is twofold. First, it seeks to assess whether predictive confidence meaningfully correlates with out-of-sample accuracy, both globally and within distinct market regimes. Second, it evaluates whether conditioning model deployment on regime membership can improve economic outcomes, as measured by risk-adjusted performance metrics such as the Sharpe ratio and maximum drawdown. By separating the tasks of prediction and deployment, the analysis aims to demonstrate that modest predictive models may still yield economically relevant outcomes when combined with appropriate regime-aware execution rules.

This framework reflects a practical perspective on financial machine learning: predictive models are treated as inputs into a broader decision-making system rather than as standalone solutions. The results highlight the importance of regime awareness in translating statistical predictions into robust trading strategies and suggest that avoiding structurally adverse environments may be more effective than attempting to extract marginal gains through increased model complexity.

2. Data and Feature Construction

2.1. Data Source and Sample Construction

The empirical analysis uses daily close-to-close data for the S&P 500 Exchange-Traded Fund (SPY), obtained from Yahoo Finance and adjusted for corporate actions. The sample spans February 2006 through December 2025, the earliest period for which all features are available, and consists of 5,007 daily observations after accounting for rolling-window construction and target alignment. To avoid look-ahead bias, the dataset is split chronologically, with the first 70% of observations used for model estimation and regime identification and the remaining 30% reserved for out-of-sample evaluation. Approximately 55% of observations correspond to positive next-day returns, reflecting the long-run upward drift in equity prices. Table 1 summarizes the dataset characteristics and sample split.

Table 1: Dataset summary and sample split

Start date	End date	N obs	Train obs	Test obs	Class balance ($y = 1$)
2006-02-01	2025-12-24	5007	3504	1503	0.5516

Figure 1 plots the 20-day realized volatility series used in feature construction. The time series exhibits pronounced clustering and sharp volatility spikes during well-known market stress episodes, providing motivation for the regime-based analysis pursued in subsequent sections.

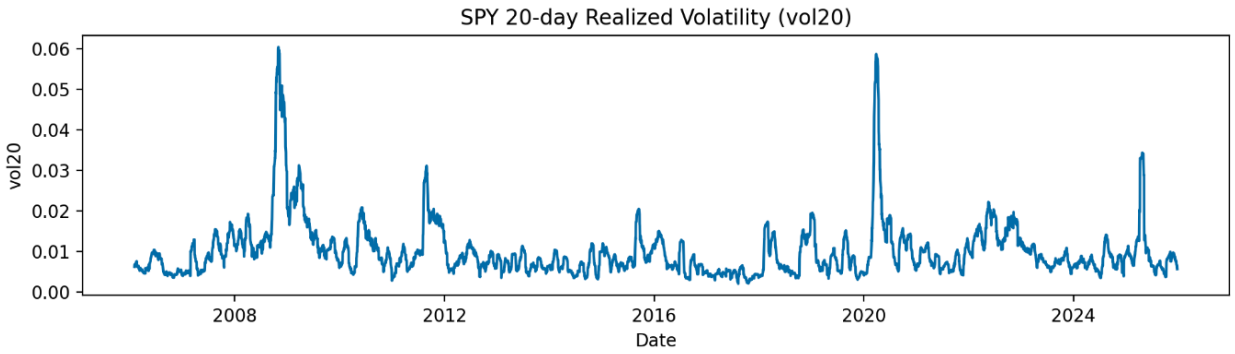

Figure 1: SPY 20-day realized volatility (vol20)

2.2. Return Definitions and Target Variable

Let P_t denote the corporate-action adjusted closing price of SPY on trading day t. Daily simple returns are computed using close-to-close prices as

$$r_t = \frac{P_t}{P_{t-1}} - 1,$$

corresponding to the implementation `ret1 = close.pct_change()` in the dataset construction pipeline. The one-step-ahead return is denoted by r_{t+1}. In the implementation, this corresponds to the forward-shifted return series

$$r_{t+1} = r_{t+1}^{(\text{data})}$$

implemented as `ret1_next = ret1.shift(-1)`. The binary classification target is then defined as the sign of the next-day return:

$$y_t \;=\; \mathbb{I}\{r_{t+1} > 0\},$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. This produces a standard one-day-ahead directional forecasting problem in which all input features are constructed using information available at time t, while the label depends only on realized returns over the subsequent period $(t, t+1]$.

Figure 2 shows the empirical distribution of next-day returns. The distribution is sharply concentrated near zero with occasional extreme outcomes, illustrating the low signal-to-noise nature of short-horizon return prediction and motivating the use of regime conditioning in subsequent analysis.

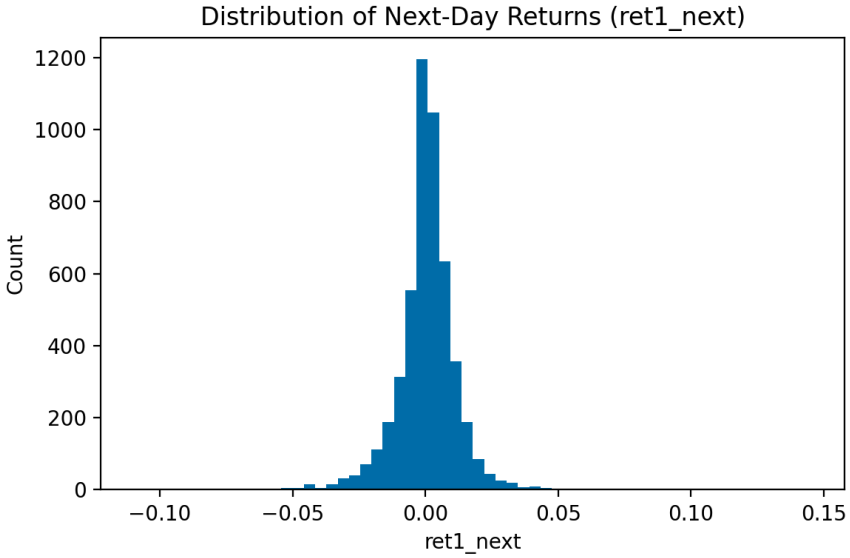

Figure 2: Distribution of next-day returns (`ret1_next`).

2.3. Feature Construction

The feature set is deliberately restricted to a small number of low-dimensional and economically interpretable variables derived from recent returns and realized volatility, emphasizing transparency and stability over model complexity. Model inputs include lagged and aggregated return measures as well as rolling estimates of realized volatility: the one-day lagged return (`ret1_lag1`) captures short-term momentum or reversal effects, the five-day return (`ret5`) provides a slightly longer-horizon measure of recent price movement, and realized volatility is computed as the rolling standard

deviation of daily returns over 10-day and 20-day windows (`vol10` and `vol20`). All features are constructed using information available at time t, avoiding look-ahead bias, and are standardized to have zero mean and unit variance prior to model estimation to ensure numerical stability and comparability across inputs. Table 2 reports descriptive statistics for the return and volatility variables, illustrating the heavy-tailed nature of returns and the substantial variation in realized volatility that motivates the regime-based analysis in subsequent sections.

Table 2: Descriptive statistics for returns and feature variables

	mean	std	min	25%	50%	75%	max
ret1	0.000486	0.012234	-0.109424	-0.003958	0.000704	0.005915	0.145198
ret1_next	0.000485	0.012234	-0.109424	-0.003958	0.000699	0.005915	0.145198
ret5	0.002352	0.024514	-0.197934	-0.008104	0.004210	0.015042	0.194036
vol10	0.009844	0.007564	0.001265	0.005390	0.008010	0.011682	0.071055
vol20	0.010056	0.007138	0.002010	0.005873	0.008333	0.011834	0.060423

3. Methodology

3.1. Regime Identification

Market regimes are identified using an unsupervised Gaussian Mixture Model (GMM) estimated on standardized features summarizing recent returns and realized volatility, specifically the five-day return (`ret5`) and rolling 10-day and 20-day volatility measures (`vol10`, `vol20`). To avoid look-ahead bias, the GMM is fit using only the first 70% of observations in chronological order and then applied to the full sample to assign regime labels. The model yields four statistically distinct and economically intuitive regimes, including a high-volatility, negative-return state associated with turbulent market conditions, a low-volatility regime reflecting more stable environments, and two intermediate regimes capturing varying combinations of return momentum and volatility. Regime occurrences are uneven but persistent, consistent with volatility clustering in equity markets, as illustrated by the inferred regime timeline in Figure 3.

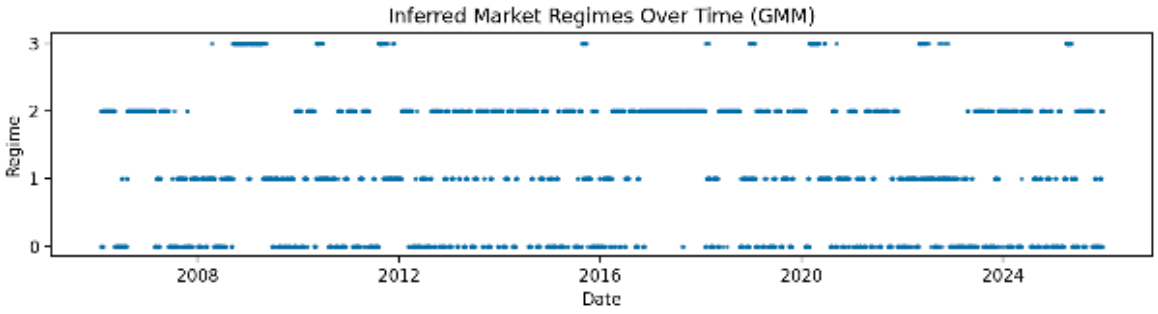

Figure 3: Inferred market regimes over time from a Gaussian Mixture Model fit on standardized `ret5`, `vol10`, and `vol20` features using the training subsample.

3.2. Predictive Model and Confidence Score

Directional predictions are generated using a logistic regression classifier trained to predict the next-day return sign. The input vector at time t consists of two standardized features, the five-day return and 20-day realized volatility, i.e., $X_t = (\texttt{ret5}_t, \texttt{vol20}_t)$, and the target is $y_t = \mathbb{I}\{r_{t+1} > 0\}$. Model estimation follows a time-ordered split: the classifier is fit on the first 70% of observations and evaluated on the remaining 30% to avoid look-ahead bias. The model outputs an out-of-sample probability score $p_t = \mathbb{P}(y_t = 1 \mid X_t)$ for test observations, with the corresponding class prediction $\hat{y}_t = \mathbb{I}\{p_t \geq 0.5\}$. A scalar confidence measure is defined as $\text{conf}_t = |p_t - 0.5|$, which increases as predicted probabilities move away from the decision threshold and is used to stratify performance and to construct confidence-filtered trading rules in subsequent analysis.

4. Evaluation Framework

4.1. Out-of-Sample Evaluation Design

All empirical evaluation is conducted strictly out of sample using a chronological test window comprising the final 30% of observations. Model estimation, feature standardization, and regime identification are performed exclusively on the training subsample to prevent information leakage. Predicted class probabilities, confidence scores, and regime labels are then fixed prior to evaluation and treated as exogenous inputs in the test period. This design ensures that all reported predictive and financial performance metrics reflect deployable, forward-looking behavior rather than in-sample fit.

4.2. Prediction-Level Performance

Prediction accuracy is evaluated as a function of model confidence and market regime using out-of-sample observations only. Confidence is defined as the absolute deviation of the predicted probability from the classification threshold, $\text{conf}_t = |p_t - 0.5|$, and test observations are grouped into deciles based on this measure. Within each confidence bin, accuracy is computed as the fraction of correct directional predictions, both unconditionally and conditional on the inferred market regime. This stratification allows assessment of whether higher-confidence predictions are empirically more reliable and whether this relationship varies across regimes. The results indicate substantial heterogeneity: certain regimes exhibit a strong monotonic relationship between confidence and accuracy, while others show weaker or unstable behavior, motivating the use of regime-aware decision rules in the strategy construction that follows.

4.3. Strategy Construction

Predicted class probabilities are translated into a simple directional trading strategy to assess the economic relevance of the model's outputs. At each out-of-sample observation, a baseline position is formed by taking a long position when the predicted probability of a positive return exceeds the

classification threshold and a short position otherwise. To reduce the influence of low-conviction predictions and mitigate noise, positions are filtered using a fixed confidence threshold. Specifically, trades are executed only when the model's confidence, defined as $|p_t - 0.5|$, exceeds 0.02; observations that do not meet this criterion are assigned a zero position.

In addition to confidence filtering, strategy deployment is conditioned on inferred market regimes. Based on the regime characteristics summarized in Table 3, the regime exhibiting the highest average realized volatility and the most adverse short-horizon return behavior (Regime 3) is treated as a high-risk state. The strategy is therefore deactivated during periods classified as this regime by setting positions to zero, regardless of the model's predicted direction. The designation of Regime 3 as a high-risk state is based on its elevated realized volatility and adverse return characteristics, which are documented and discussed in Section 5.1.

Both the confidence threshold and the regime-based deactivation rule are fixed prior to out-of-sample evaluation and applied uniformly throughout the test period. Strategy returns are computed by multiplying the resulting position by the realized next-day return, producing a fully out-of-sample return series used for subsequent performance analysis.

4.4. Financial Performance Metrics

Strategy performance is evaluated using standard risk-adjusted and drawdown-based measures computed from the out-of-sample return series. Risk-adjusted performance is summarized by the annualized Sharpe ratio, calculated using daily strategy returns and a constant risk-free rate. Downside risk is assessed via maximum drawdown, defined as the largest peak-to-trough decline in the cumulative equity curve over the evaluation period. In addition, trade frequency is reported as the fraction of test-period observations with nonzero positions, providing a measure of strategy activity and capital deployment. Performance metrics are computed both for the overall strategy and conditionally by market regime, allowing direct assessment of regime-specific contribution and validating the regime-based activation and deactivation rules.

5. Empirical Results and Strategy Performance

5.1. Regime Characteristics and Predictive Reliability

The four regimes inferred by the GMM exhibit economically distinct patterns in both recent returns and realized volatility. In particular, Regime 2 represents a low-volatility environment (lowest `vol10` and `vol20`) with relatively stable return behavior, while Regime 3 corresponds to a rare but extreme high-volatility state (substantially elevated `vol10` and `vol20`) accompanied by negative average short-horizon returns and markedly higher dispersion. Regimes 0 and 1 occupy intermediate volatility levels but differ in return dynamics, with Regime 1 associated with positive average `ret5` and Regime 0 associated with negative average `ret5`. Table 3 summarizes regime frequencies and feature moments, providing the empirical basis for regime-conditional evaluation and the subsequent regime-based strategy controls.

Table 3: Regime frequencies and feature moments (GMM regimes)

	ret5			vol10			vol20	
	count	mean	std	count	mean	std	count	mean	std
0	1440	-0.006406	0.022228	1440	0.010079	0.002380	1440	0.009199	0.001837
1	1326	0.012546	0.021178	1326	0.011785	0.004943	1326	0.013043	0.003381
2	1926	0.004471	0.009636	1926	0.005091	0.001376	1926	0.005366	0.001167
3	315	-0.013477	0.061518	315	0.029656	0.014933	315	0.030073	0.012740

5.2. Confidence-Stratified Accuracy Across Regimes

Out-of-sample predictive accuracy is evaluated by stratifying test observations into confidence deciles, where confidence is defined as $\text{conf}_t = |p_t - 0.5|$, and computing directional accuracy within each bin conditional on the inferred market regime. This analysis is intended to provide a diagnostic view of how model reliability varies jointly with predictive confidence and market state, rather than to establish a precise functional relationship.

Figure 4 illustrates that the relationship between confidence and realized accuracy differs across regimes. In lower-volatility regimes, accuracy remains relatively stable across confidence bins and is generally higher in bins associated with larger deviations from the classification threshold. In contrast, accuracy patterns in the high-volatility regime are more irregular across bins, reflecting both reduced predictive stability and the limited number of observations available in certain confidence–regime combinations. As a result, individual bin-level fluctuations should be interpreted cautiously, with emphasis placed on broader regime-level differences rather than fine-grained monotonic trends.

Figure 5 provides a complementary geometric perspective by overlaying inferred regime labels on the global logistic probability field in (`ret5`, `vol20`) space. The visualization highlights that regimes occupy distinct regions of the feature space, implying that a single global decision rule

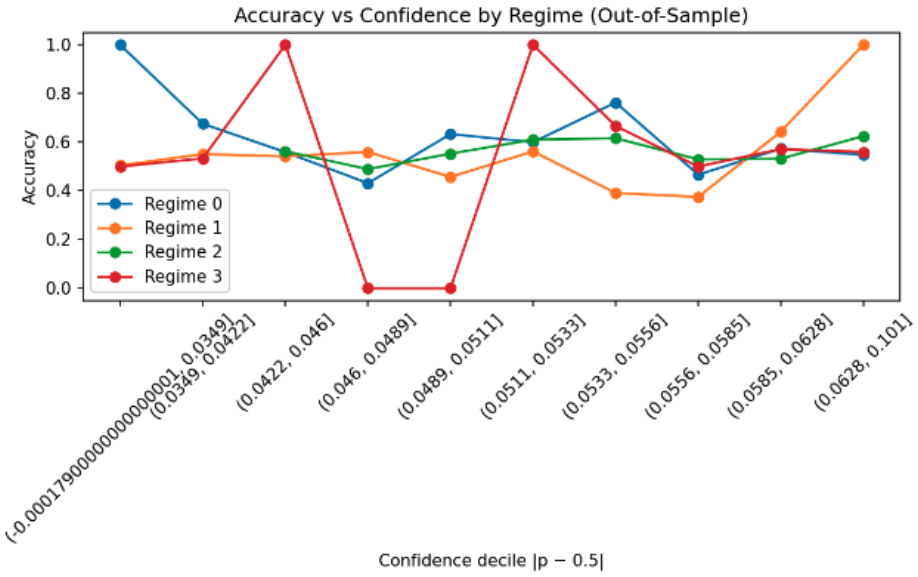

Figure 4: Out-of-sample accuracy versus confidence decile, computed separately within each inferred regime. Confidence is defined as $|p - 0.5|$.

operates under materially different state distributions. This structural heterogeneity motivates the regime-aware evaluation and deployment rules examined in subsequent sections. Additional calibration diagnostics reported in Appendix A indicate that predicted probabilities are tightly concentrated near the unconditional mean, motivating the use of confidence thresholds rather than probability ranking.

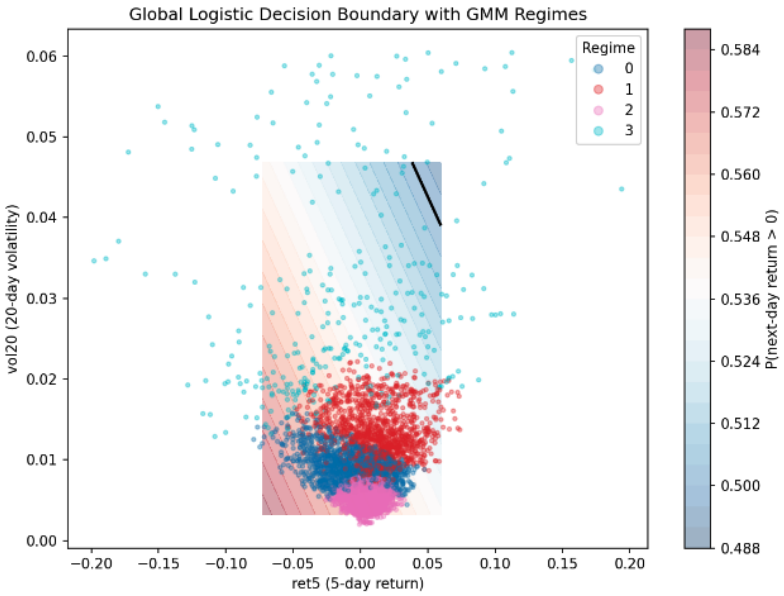

Figure 5: Global logistic probability field in (`ret5`, `vol20`) space with inferred GMM regime labels overlaid, illustrating regime separation and state-dependent operating conditions.

5.3. Strategy-Level Performance

The economic implications of regime- and confidence-aware prediction are evaluated using an out-of-sample trading strategy constructed from the model's predicted probabilities. Strategy performance is summarized using standard risk-adjusted and drawdown-based metrics computed over the test period. The resulting strategy exhibits a positive annualized Sharpe ratio, moderate maximum drawdown, and high trade participation, indicating that selective deployment of model signals can materially influence realized risk-adjusted performance.

Performance varies substantially across market regimes. Regimes associated with low and moderate realized volatility contribute positively to overall strategy returns, while the highest-volatility regime delivers poor or unstable performance. This heterogeneity motivates the conditional deactivation of the strategy during periods classified as the high-risk regime, as discussed in Section 4.3. Figure 6 reports the cumulative out-of-sample equity curve, illustrating that regime- and confidence-based filtering primarily affects the volatility and drawdown profile of returns rather than generating uniform gains across all market conditions. Transaction costs, slippage, and shorting or financing constraints are not modeled; reported performance should therefore be interpreted as gross of trading frictions.

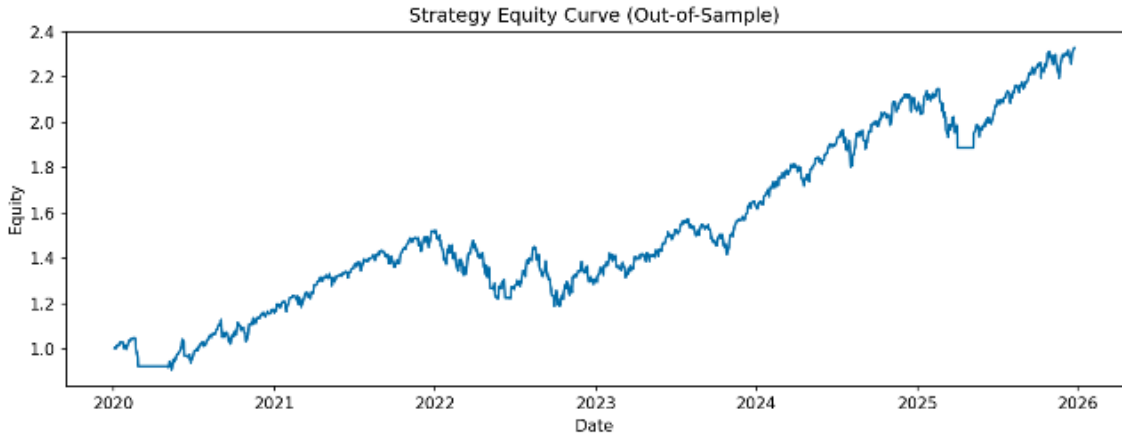

Figure 6: Out-of-sample equity curve of the confidence- and regime-filtered trading strategy.

6. Conclusion

This study examines the interaction between predictive confidence and market regimes in a simple, interpretable classification setting using financial time series data. By combining a global logistic regression model with unsupervised regime identification based on recent returns and realized volatility, the analysis demonstrates that predictive reliability is highly state dependent and that confidence alone is insufficient without regime context. Empirical results show that confidence-filtered predictions deliver differentiated risk-adjusted outcomes in low and moderate volatility regimes, while predictive signals deteriorate in extreme volatility environments, motivating regime-based strategy controls. Taken together, these findings highlight the importance of incorporating market state information into the evaluation and deployment of predictive models, even when using otherwise static classifiers. Future work may extend this framework to richer feature sets, nonlinear models, alternative regime definitions, and the explicit incorporation of transaction costs and execution constraints.

A. Appendix

A.1. Confidence Distribution Diagnostics

This appendix reports summary statistics for the model confidence measure used throughout the analysis. Confidence is defined as the absolute deviation of the predicted probability from the classification threshold, $\text{conf}_t = |p_t - 0.5|$. The distribution exhibits meaningful dispersion and is not concentrated near zero, supporting its use for stratifying predictions and constructing confidence-filtered strategies.

Table 4: Summary statistics for model confidence (out-of-sample)

count	mean	std	min	25%	50%	75%	max
1503	0.0498	0.0121	0.0008	0.0443	0.0511	0.0571	0.1008

A.2. Probability Calibration Diagnostics

To assess the calibration and dispersion of predicted probabilities, we examine both the distribution of out-of-sample predicted probabilities and the associated Brier score. Figure 7 plots the histogram of predicted probabilities in the test period. The distribution is tightly concentrated around the unconditional mean return probability, with few extreme probability values, indicating substantial probability compression. This behavior is consistent with the low signal-to-noise nature of short-horizon equity return prediction.

As a complementary summary measure, the Brier score is computed as the mean squared error between predicted probabilities and realized outcomes,

$$\text{Brier} = \frac{1}{n}\sum_{t=1}^{n}(p_t - y_t)^2.$$

The resulting out-of-sample Brier score is 0.247, which is close to the value obtained by an uninformative baseline forecast and reflects the limited discriminative power of the model at the daily horizon. These diagnostics reinforce the interpretation that model probabilities are best viewed as low-amplitude signals suitable for confidence-based filtering rather than as precise probabilistic forecasts.

A.3. Regime-Level Strategy Performance

To further assess regime dependence in economic performance, this appendix reports risk-adjusted returns computed separately within each inferred regime. Results confirm that strategy profitability is concentrated in low- and moderate-volatility regimes, while the highest-volatility regime fails to deliver stable risk-adjusted performance, motivating its exclusion in the regime-aware strategy.

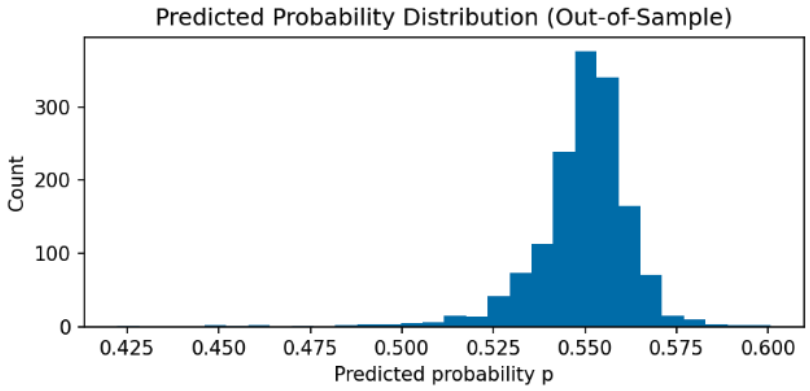

Figure 7: Distribution of out-of-sample predicted probabilities. Predicted probabilities are tightly concentrated near the unconditional mean, indicating probability compression.

Table 5: Out-of-sample Sharpe ratios by regime

Regime	Sharpe ratio
0	1.38
1	0.41
2	0.22
3	NaN

A.4. Prediction Error Anatomy

For completeness, this appendix reports accuracy and error rates stratified jointly by confidence decile and regime. These diagnostics provide a granular view of where prediction errors concentrate across market states and confidence levels and underpin the regime-conditional patterns discussed in the main text. Due to size considerations, the full table is included here rather than in the main body.

Table 6: Prediction error anatomy by regime and confidence bin (out-of-sample)

Regime	Confidence bin $\lvert p_t - 0.5\rvert$	Accuracy	Error rate	n
0	(0.000, 0.0349]	1.000	0.000	2
0	(0.0349, 0.0422]	0.676	0.324	37
0	(0.0422, 0.0460]	0.559	0.441	59
0	(0.0460, 0.0489]	0.431	0.569	51
0	(0.0489, 0.0511]	0.633	0.367	30
0	(0.0511, 0.0533]	0.600	0.400	40
0	(0.0533, 0.0556]	0.763	0.237	38
0	(0.0556, 0.0585]	0.467	0.533	45
0	(0.0585, 0.0628]	0.573	0.427	82

Continued on next page

Regime	Confidence bin $\lvert p_t - 0.5 \rvert$	Accuracy	Error rate	n
0	(0.0628, 0.1010]	0.548	0.452	115
1	(0.000, 0.0349]	0.505	0.495	109
1	(0.0349, 0.0422]	0.551	0.449	98
1	(0.0422, 0.0460]	0.542	0.458	72
1	(0.0460, 0.0489]	0.560	0.440	50
1	(0.0489, 0.0511]	0.457	0.543	35
1	(0.0511, 0.0533]	0.562	0.438	32
1	(0.0533, 0.0556]	0.391	0.609	23
1	(0.0556, 0.0585]	0.375	0.625	16
1	(0.0585, 0.0628]	0.643	0.357	14
1	(0.0628, 0.1010]	1.000	0.000	3
2	(0.0422, 0.0460]	0.562	0.438	16
2	(0.0460, 0.0489]	0.489	0.511	47
2	(0.0489, 0.0511]	0.553	0.447	85
2	(0.0511, 0.0533]	0.610	0.390	77
2	(0.0533, 0.0556]	0.616	0.384	86
2	(0.0556, 0.0585]	0.529	0.471	85
2	(0.0585, 0.0628]	0.532	0.468	47
2	(0.0628, 0.1010]	0.625	0.375	8
3	(0.000, 0.0349]	0.500	0.500	40
3	(0.0349, 0.0422]	0.533	0.467	15
3	(0.0422, 0.0460]	1.000	0.000	3
3	(0.0460, 0.0489]	0.000	1.000	2
3	(0.0489, 0.0511]	0.000	1.000	1
3	(0.0511, 0.0533]	1.000	0.000	1
3	(0.0533, 0.0556]	0.667	0.333	3
3	(0.0556, 0.0585]	0.500	0.500	4
3	(0.0585, 0.0628]	0.571	0.429	7
3	(0.0628, 0.1010]	0.560	0.440	25

Notes: Accuracy is computed as the fraction of correct directional predictions within each confidence bin, conditional on the inferred regime. Some regime–bin combinations contain small sample sizes (e.g., $n \leq 5$), so extreme accuracies in those bins should be interpreted cautiously.