Regime-Dependent Structure in Silver Pricing: An Explainable Machine Learning Approach

A Case Study on AI-Driven Commodity Market Decision-Making

End-Term Project

Machine Learning in Finance

Param Shah

Khushi Khanna

M.S. Financial Engineering

New York University (Tandon School of Engineering)

December 19, 2025

Contents

1 Executive Summary

This project develops a regime-dependent machine learning framework to analyze short-horizon predictability and factor relevance in the silver market. The business objective is not to claim that silver prices are easily predictable, but to answer a more operationally useful question for commodity desks:

> **Under which market conditions do economic drivers produce repeatable structure in silver returns, and which drivers matter most in each regime?**

Silver is a hybrid commodity: it behaves partly like a precious metal (investment/ monetary hedge) and partly like an industrial input (growth/manufacturing cycle sensitivity). This dual role makes its dynamics highly state-dependent. Therefore, a single global model is expected to fail; a regime-aware and explainable approach is required.

We define regimes using two dimensions:

(a) **Volatility state**: high vs low realized volatility (20-day realized volatility relative to a rolling median threshold)

(b) **Trend state**: uptrend vs downtrend (200-day trend filter: price relative to moving average)

This produces four regimes: HighVol_UpTrend, HighVol_DownTrend, LowVol_UpTrend, and LowVol_DownTrend. Models are trained and evaluated via strict out-of-sample walk-forward splits (rolling 5-year training window; testing year-by-year) to replicate real trading deployment.

Because short-horizon return magnitude predictability is weak (a stylized fact in liquid markets), we prioritize finance-standard evaluation metrics:

- **Spearman Information Coefficient (IC)**: rank correlation between predicted and realized returns; standard measure of signal quality in factor research
- **Hit rate**: sign accuracy (directional correctness)
- **Block-bootstrap IC t-stat**: statistical robustness under autocorrelation (block resampling)
- **Classification metrics** for an alternate target: predicting $P(r_{t,t+5} > 0)$ using AUC, log loss, and Brier score

We benchmark three models (Ridge, Random Forest, XGBoost) to avoid single-model storytelling and to test robustness across linear and nonlinear hypothesis classes. Finally, we apply time-series-safe **block permutation importance** as an explainability method: importance is measured by the drop in IC when a feature is permuted in blocks, preserving autocorrelation.

Key empirical findings:

- **Predictability is regime-dependent.** Uptrend regimes show materially higher IC than downtrends.
- **HighVol_UpTrend is the most robust regime:** IC is positive in a high fraction of years for Ridge and Random Forest, suggesting persistent ranking structure during volatile rallies.
- **Downtrends are structurally difficult:** results are unstable and often negative, consistent with liquidation, funding constraints, and macro shock dominance.
- **Economic drivers differ by regime:**
  - In volatile uptrends: gold–silver relative valuation, volatility state, and equity risk-on proxies (SPX) matter most.
  - In stable uptrends: volatility, rates, momentum, and risk sentiment (VIX) become central.
  - In downtrends: long trend state and rate shocks dominate, consistent with deleveraging and USD/liquidity regimes.

From a business perspective, this pipeline supports:

- regime-conditioned signal deployment / throttling,
- risk-aware hedging and scenario analysis,
- factor monitoring and stress-testing,
- explainable communication to risk committees and stakeholders.

2 Business Problem in the Commodities Sector

2.1 Why Silver is a Difficult Commodity to Model

Silver's price is driven by overlapping channels:

1. **Investment channel (precious metal behavior):** monetary policy expectations, real rates, USD funding conditions, safe-haven flows, inflation hedging narratives.

2. **Industrial channel:** manufacturing cycle, electronics and solar demand, global growth expectations, energy/metal complex co-movement.
3. **Cross-asset positioning channel:** risk-on/off shifts, equity volatility regimes, CTA / momentum flows, commodity index rebalancing, liquidity shocks.

These channels dominate at different times, causing **structural breaks** and **state dependence**. In practice, commodity desks must answer: *Is the current market environment one where our signals should work?* This is exactly what regime-conditioned ML is designed to test.

2.2 Business Use Cases of This AI System

This project maps directly to real commodity workflows:

- **Systematic trading / signal gating:** activate signals only in regimes where IC is historically positive and stable; reduce exposure when regimes are historically noisy or adverse.
- **Risk management / stress testing:** identify which macro drivers matter in drawdowns vs rallies; informs hedging design (e.g., rates hedge vs USD hedge).
- **Portfolio allocation:** understand when silver behaves closer to gold (monetary) versus closer to industrial metals (growth), supporting diversification and hedging decisions.
- **Explainability and governance:** block-permutation importance provides a defensible model interpretation framework that is more appropriate than standard feature importance in time series.

3 Dataset:Sources and Preprocessing

3.1 Data Description

The dataset contains daily observations including:

- Silver spot price (primary series)
- S&P 500 index (risk-on proxy)
- Gold price (precious metals anchor)
- Oil (growth / inflation / commodity cycle proxy)
- Copper (industrial cycle / China-sensitive proxy)

- DXY (USD funding conditions)
- VIX (risk sentiment / volatility pricing)
- U.S. 10-year yield (macro rates regime proxy)
- Gold/Silver ratio (relative valuation measure)

The sample includes thousands of daily points (approximately 7,000+ after feature creation and NA drops), spanning multiple macro regimes: pre-2008 expansion, the global financial crisis, QE era, the 2011 precious metals spike, COVID shock, 2022 inflation/rate shock, and the recent post-2023 rate normalization phase.

3.2 Cleaning and Data Integrity Controls

Key preprocessing steps:

- Removal of Excel artifact columns
- Strict chronological sorting by date
- Conversion to a time-series index
- Drop of NA rows created by rolling windows (e.g., 200-day MA, 252-day z-scores)

3.3 Stationarity and Feature Engineering

Using raw price levels would induce spurious relationships (non-stationary series). Therefore, we engineer primarily stationary predictors:

- **Log returns** for all major price series:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right)$$

- **Momentum (EWMA 60-day)**:

$$\mathrm{mom}_{60}(t) = \mathrm{EWMA}_{60}(r_{t-1})$$

- **Realized volatility (20-day)**:

$$\mathrm{rvol}_{20}(t) = \sqrt{\sum_{i=1}^{20} r_{t-i}^2}$$

- **Gold–silver valuation z-score (252-day)**:

$$z_{GS}(t) = \frac{GS(t) - \mu_{252}(t-1)}{\sigma_{252}(t-1)}$$

- **Rate shock**:

$$d10y(t) = y_{10}(t) - y_{10}(t-1)$$

- **Long trend (200-day)**:

$$\text{trend}_{200}(t) = \frac{S_t}{MA_{200}(t-1)} - 1$$

Critically, all rolling features are lagged (using shift(1)) to ensure no look-ahead bias.

4 Targets: Regression and Classification

4.1 Regression Target: 5-day Forward Return

We define the primary predictive label as:

$$y_t = \log\left(\frac{S_{t+5}}{S_t}\right)$$

This is the standard forward return formulation for commodity forecasting and aligns with weekly horizon trading decisions.

4.2 Why Add a Classification Target

Return magnitude is extremely noisy. Many realistic trading systems are based on ranking or directional probability rather than magnitude forecasts. Therefore, we define:

$$\text{Up}_t = \mathbf{1}(y_t > 0)$$

and train probabilistic models to estimate $P(\text{Up}_t = 1)$.

This target is operationally interpretable:

- risk managers can map probabilities into position sizing,
- desks can set thresholds (e.g., trade only if $P > 0.55$),
- probabilities allow calibration checks (Brier, log loss).

5 Regime Framework:Economic Interpretation

5.1 Regime Construction

Regimes are built using:

- realized volatility relative to a rolling median threshold,
- trend relative to a 200-day moving average filter.

The four regimes correspond to distinct macro/market environments:

5.2 Economic Intuition Behind Volatility-Trend Regimes

1) HighVol_UpTrend (volatile rally) Often observed during:

- crisis-policy response regimes (QE, liquidity injections),
- inflation hedging narratives,
- speculative precious metals flows,
- risk-on rebounds where leverage returns quickly.

Mechanism: volatility is elevated but the trend is positive, so the market experiences large swings with upward drift. This regime is where cross-asset positioning and relative value signals often become informative.

2) LowVol_UpTrend (stable expansion) Often observed during:

- steady growth expansions,
- declining risk premia (low VIX),
- stable funding conditions.

Mechanism: macro conditions are stable; systematic strategies (trend/momentum) often perform better, and linear relationships can generalize.

3) HighVol_DownTrend (crisis liquidation) Often observed during:

- sudden tightening of financial conditions,
- margin calls and deleveraging,
- USD liquidity shocks,
- rapid repricing of rates expectations.

Mechanism: correlations shift, liquidity dominates fundamentals, and short-horizon returns become less learnable. This is the regime where model failure is expected and itself becomes an important risk signal.

4) LowVol_DownTrend (slow grind / regime decay) Often observed during:

- persistent tightening cycles,
- slow demand deterioration,
- gradual USD appreciation environments.

Mechanism: trends persist but signals can be mixed; valuation and macro factors may matter more than pure momentum.

5.3 Regime Smoothing

Short-lived regime flips create noise and artificially inflate backtest instability. We smooth regime labels by forcing minimum run length, aligning with the idea that macro regimes do not meaningfully change day-to-day.

Figure 1: Silver Price with Structural Market Regimes

Note on regime persistence. Regime labels in Figure 1 are subject to a minimum persistence constraint of 20 trading days. Short-lived regime flips below this threshold are suppressed and absorbed into the surrounding dominant regime. Without this constraint, regime classification becomes excessively fragmented, producing a highly chopped visualization that is economically uninterpretable and inconsistent with the notion that macro and volatility regimes evolve gradually rather than day-to-day.

Visual Regime Validation

Figure 1 provides a time-series visualization of silver prices overlaid with the inferred volatility–trend regimes. Several economically meaningful patterns emerge. Sustained price rallies (e.g., mid-2000s, post-2020) coincide with extended HighVol_UpTrend and LowVol_UpTrend phases, while sharp drawdowns and crisis episodes (e.g., 2008, 2013, 2022) are dominated by HighVol_DownTrend regimes. The imposed minimum regime persistence ensures that regime transitions reflect genuine shifts in market state rather than transient noise, reinforcing the interpretation of regimes as structural market environments rather than short-term technical signals. This visualization motivates the regime-conditioned modeling approach used in subsequent sections.

6 AI Techniques Used

6.1 Predictive Models

We use three ML methods as alternative hypothesis spaces:

Ridge Regression (L2-regularized linear model) Chosen for stability, interpretability, and robust generalization under correlated macro features.

Random Forest Regressor Captures nonlinear interactions (e.g., volatility interacting with valuation) with relatively robust behavior under noise.

XGBoost Regressor Powerful boosted tree method, but often more sensitive to overfitting under low signal-to-noise time series.

For the classification extension: **Logistic Regression (probabilistic baseline)** and **Random Forest Classifier** are used to estimate probability of positive returns.

6.2 Out-of-Sample Walk-Forward Evaluation

We apply rolling-window walk-forward splits:

- Train on the past 5 years only
- Test on the next calendar year
- Repeat through the dataset

This mimics production deployment: a model trained only on information available at the time.

6.3 Explainability: Block Permutation Importance (IC-drop)

Standard feature importance is inappropriate in time series because it breaks autocorrelation and can create misleading attributions. We use **block permutation**:

1. predict OOS with full features,

2. permute one feature in contiguous blocks (preserving time structure),
3. recompute IC,
4. define importance as:

$$\Delta IC = IC_{\text{base}} - IC_{\text{permuted}}$$

Positive ΔIC means that feature contributes to ranking skill; negative values suggest it may harm generalization.

7 Main Results: Regression (Return Ranking Skill)

7.1 IC by Regime (Core Regime-Dependence Result)

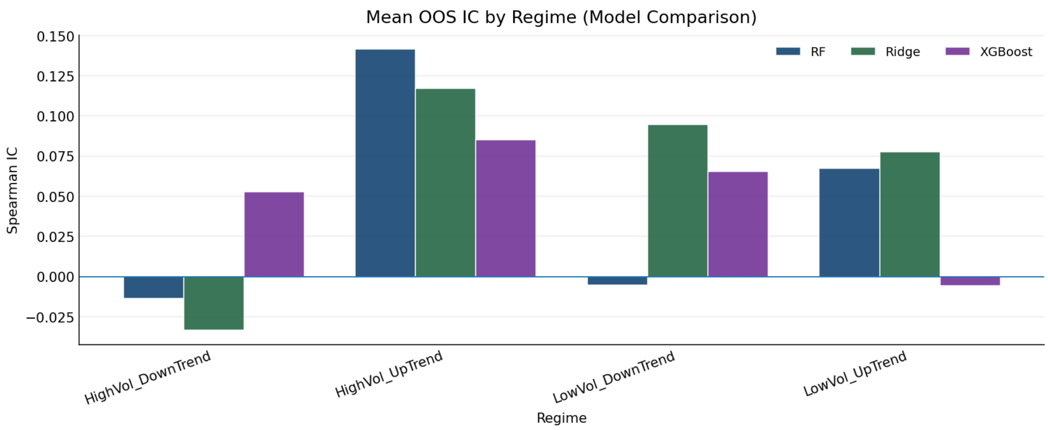

Figure 2: Mean Out-of-Sample Spearman IC by Regime (Model Comparison)

Interpretation: The most important conclusion is that the signal is not global. Predictability concentrates in uptrend regimes, especially HighVol_UpTrend, where all three models show positive IC on average. Downtrend regimes show near-zero or negative IC for Ridge and RF, consistent with the idea that drawdowns are dominated by risk-off liquidation and funding constraints.

7.2 Block-Bootstrap IC t-statistics (Robustness under Autocorrelation)

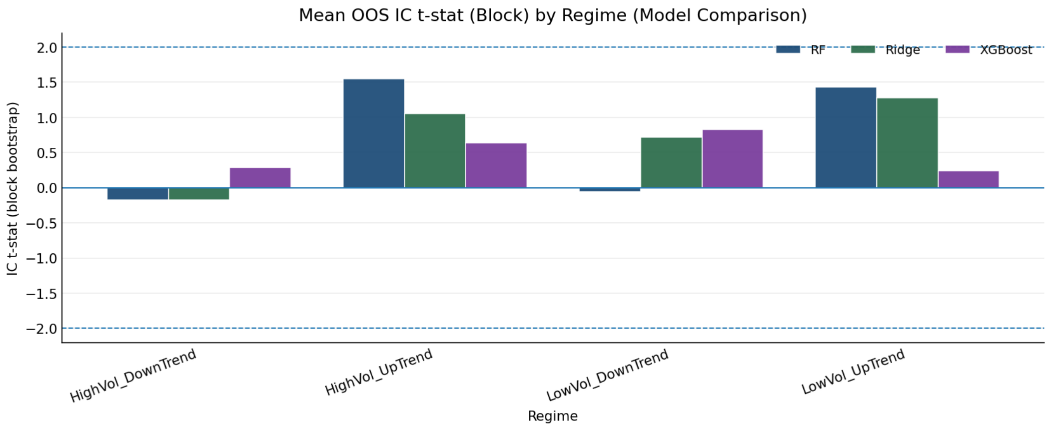

Figure 3: Mean Block-Bootstrap IC t-stat by Regime

Interpretation: The block-bootstrap t-stat summarizes statistical reliability when returns are autocorrelated. Most t-stats lie below 2, which is expected given weak return predictability at 5-day horizons. Nonetheless, the uptrend regimes deliver systematically higher t-stats than downtrends, supporting the claim that regime segmentation recovers meaningful structure.

7.3 Heatmap Summary (Compact View Across Regimes)

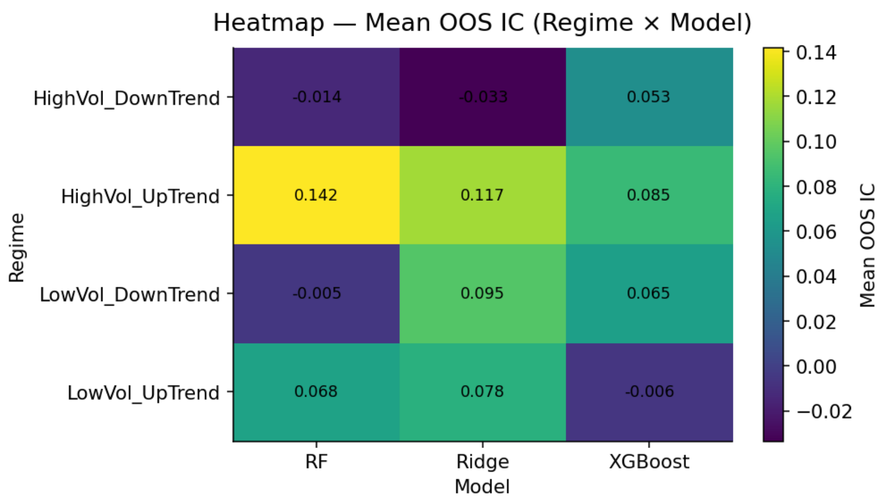

Figure 4: Heatmap — Mean OOS IC (Regime × Model)

Interpretation: No model dominates universally. Ridge and RF are consistently strong in HighVol_UpTrend. In downtrends, performance is unstable, and XGBoost sometimes shows positive IC where others fail, likely reflecting nonlinear episodic relationships rather than stable structure.

8 YoY Regime Diagnostics

A critical risk in commodity ML is that results are driven by one crisis year. Therefore, we evaluate stability across years.

8.1 HighVol_UpTrend (Most Robust Regime)

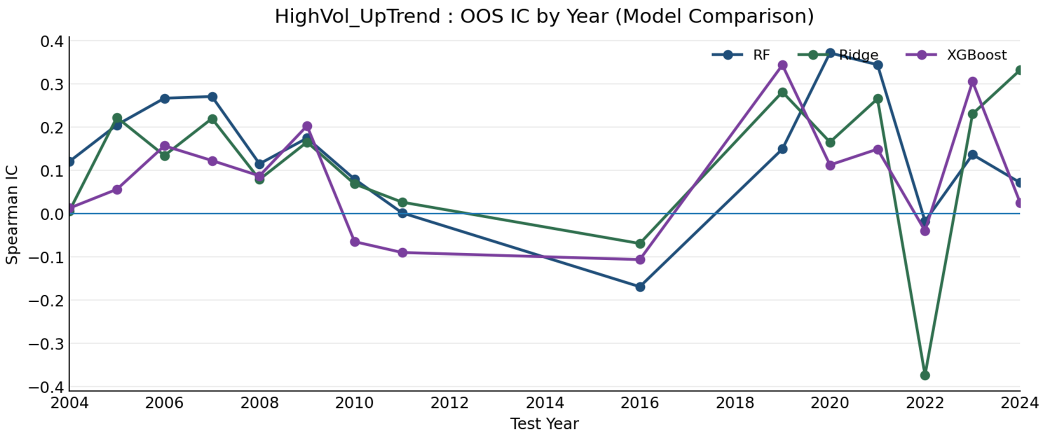

Figure 5: HighVol_UpTrend: Year-by-Year OOS IC (Model Comparison)

Finance interpretation: HighVol_UpTrend corresponds to volatile rallies often driven by policy/liquidity and positioning. Across years, RF and Ridge show a high fraction of positive IC values, suggesting persistent ranking structure. However, performance dips during major structural breaks (e.g., periods where rate shock dominates all assets), illustrating that even regime-aware models can fail when the macro driver changes.

8.2 HighVol_DownTrend (Liquidation Regime)

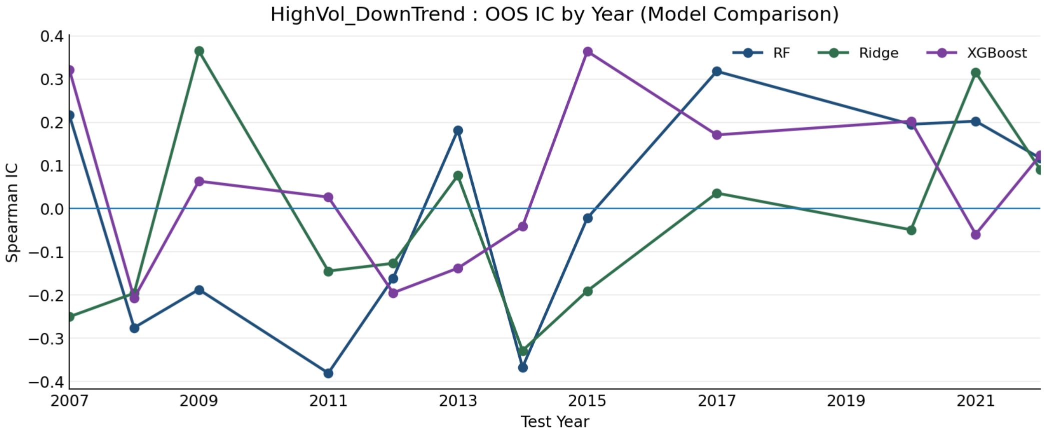

Figure 6: HighVol_DownTrend: Year-by-Year OOS IC (Model Comparison)

Finance interpretation: This is the regime of margin calls, forced selling, and rapid tightening in risk appetite. In such environments, relationships between macro variables and silver returns are unstable, and correlations often converge ("everything sells off"), destroying model signal. The instability observed here is therefore consistent with market microstructure and funding risk dynamics.

8.3 LowVol_DownTrend and LowVol_UpTrend

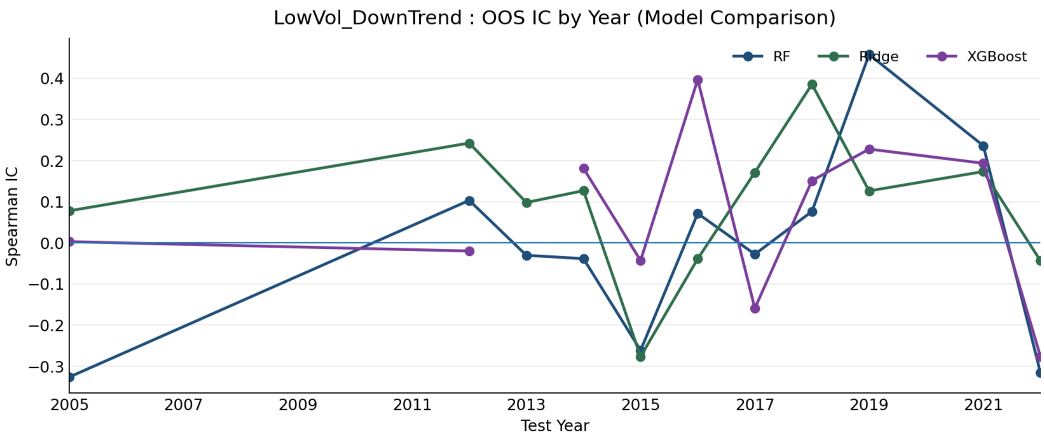

Figure 7: LowVol_DownTrend: Year-by-Year OOS IC (Model Comparison)

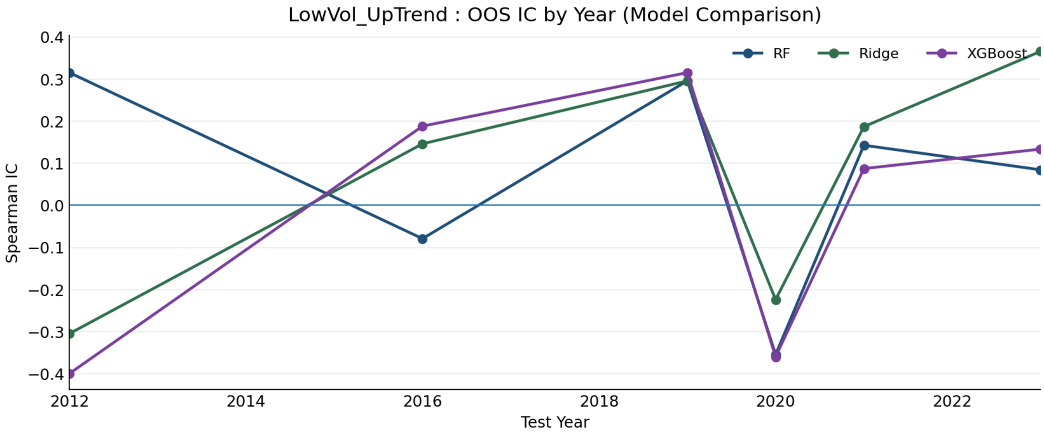

Figure 8: LowVol_UpTrend: Year-by-Year OOS IC (Model Comparison)

Interpretation: Low volatility regimes show fewer samples per year and therefore higher estimation noise. Still, Ridge tends to remain more stable than nonlinear models in LowVol_UpTrend, consistent with the idea that stable expansions allow linear macro relationships to generalize.

9 Classification Extension: Predicting $P(r_{t,t+5} > 0)$

9.1 AUC by Regime

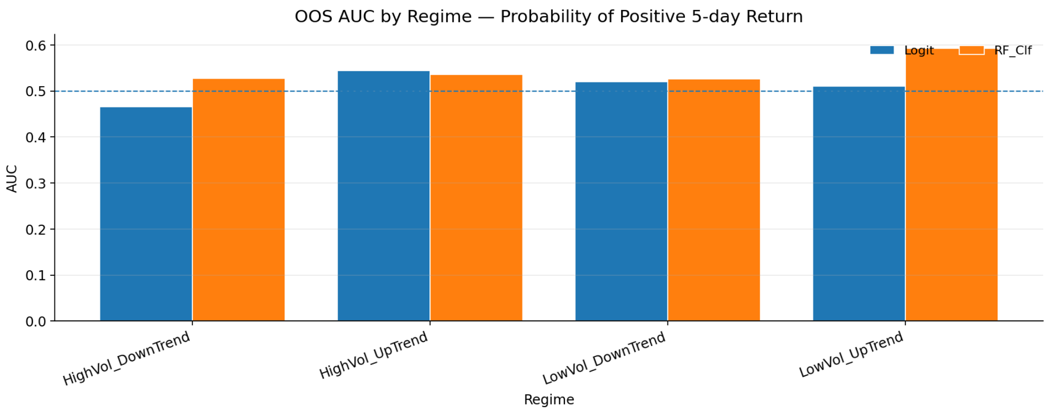

Figure 9: OOS AUC by Regime — Probability of Positive 5-day Return

Interpretation: AUC results reinforce the main conclusion: predictive structure depends on market state. RF classifier performs strongest in LowVol_UpTrend, indicating that stable conditions allow more learnable directional probability structure. In HighVol_DownTrend, AUC declines toward (or below) 0.5, highlighting the difficulty of forecasting sign during liquidation regimes.

10 Explainability Results: Regime-Dependent Drivers

We interpret block-permutation importance as regime-specific economic attribution.

10.1 HighVol_UpTrend Drivers (Volatile Rally)

Top drivers include:

- **Gold–Silver z-score** (relative valuation / precious metals linkage)
- **Realized volatility** (state variable capturing market turbulence)
- **Long trend filter** (macro positioning / CTA effects)
- **SPX returns** (risk-on spillover)

Macro rationale: In volatile rallies, silver frequently trades as a leveraged expression of precious-metals sentiment. Relative valuation versus gold becomes informative when investors rotate within the precious complex. Equity risk sentiment and volatility state jointly determine whether flow-driven continuation dominates or whether reversal risk is high.

10.2 LowVol_UpTrend Drivers (Stable Expansion)

Top drivers include:

- **Volatility + VIX** (risk pricing even when realized vol is low)
- **Rate shocks (d10y)** (real-rate expectations channel)
- **Momentum** (trend persistence)
- **Oil** (reflation / growth proxy impacting industrial demand narrative)

Macro rationale: Stable expansions strengthen cross-asset macro linkages. Rates anchor discounting and USD funding, while oil and copper act as growth-cycle proxies. Momentum persists because liquidity is available and positioning is not forced.

10.3 Downtrend Drivers (Risk-Off and Funding Stress)

Downtrends show importance concentrated in:

- **trend_200** (state variable capturing prolonged weakness)

- **d10y** (tightening shocks and real-rate repricing)

Macro rationale: During drawdowns, silver is influenced by the cost of capital, USD liquidity, and forced de-risking. Industrial narratives weaken; macro tightening and deleveraging dominate. Hence model signal deteriorates — an expected and meaningful finding.

11 Expected Benefits (Business Value)

This AI pipeline provides value in multiple ways:

11.1 Regime-Aware Signal Deployment

Instead of trading signals blindly, desks can:

- increase reliance on models in HighVol_UpTrend where IC is historically stable,
- reduce exposure or widen risk limits in HighVol_DownTrend where models become unreliable.

11.2 Risk Management and Hedging Insights

By identifying which macro variables matter in each regime, the system supports:

- hedge design (rates-driven vs USD-driven vs equity-risk driven),
- scenario analysis (e.g., "if rates shock increases, which regime likely emerges?"),
- stress-testing of factor exposures.

11.3 Explainability and Governance

The block-permutation IC-drop method gives a defensible and time-series-appropriate attribution mechanism suitable for communicating with:

- risk committees,
- model governance teams,
- non-technical stakeholders.

12 Risks, Model Limitations, and Challenges

12.1 Model Risk and Overfitting

Returns have low signal-to-noise. Complex models (e.g., boosting) can overfit, especially when regimes reduce sample sizes. This is why:

- walk-forward OOS evaluation is mandatory,
- multiple model classes are compared,
- block bootstrap is used for realistic t-stats.

12.2 Regime Misclassification Risk

Regimes are defined mechanically; threshold choice and smoothing parameters may affect results. In production, regime definitions should be stress-tested and possibly enhanced using macro regime indicators (policy rate regime, inflation regime, liquidity indices).

12.3 Structural Breaks and Market Evolution

Events such as COVID, 2022 inflation shock, or major policy regime changes alter relationships between features and silver returns. This motivates:

- rolling retraining,
- stability reporting,
- explicit monitoring for performance decay.

12.4 Data Quality and Availability

Commodity datasets often involve missing values, asynchronous updates, and vendor-specific revisions. Robust pipelines must include:

- validation checks,
- missing data handling policies,
- reproducible transformations.

13 Conclusion and Next Steps

This case study shows that machine learning can be used effectively in commodities not as a naive predictor of returns, but as a disciplined, regime-aware diagnostic system. Silver's predictability is highly state-dependent: uptrends show repeatable structure, while downtrends are dominated by liquidity and macro shocks.

The most credible contribution of this work is the combination of:

- regime-aware evaluation,
- finance-appropriate metrics (IC and block-bootstrap robustness),
- time-series-safe explainability (block permutation IC-drop),
- business-ready interpretation aligned with real market mechanisms.

14 References

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

[2] Kelly, B., and Pruitt, S. (2020). Forecasting with many predictors. *Annual Review of Financial Economics*, 12, 209–224.

[3] Biau, G., and Scornet, E. (2016). A random forest guided tour. *TEST*, 25, 197–227.

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