

Background & Motivation

Cost 1 builds a prediction and explainability pipeline **from scratch on public data**: 12 large-cap US stocks and their 4 sector ETFs, 7 years of daily prices and volume, earnings and FOMC calendars, the VIX, and RavenPack news sentiment (WRDS). We first fixed the plan (a 5-day forward target, XGBoost with TreeSHAP, faithfulness and stability tests), then built the data and features, and ran a first purged walk-forward XGBoost that trailed a linear baseline. This poster completes the plan: the out-of-sample predictions compared with realised returns, the ticker and sector encoding test, matched refit cadence and tuning, TreeSHAP, and the evaluation of the explanations.

Data & Feature Engineering

- **Universe**: AAPL, NVDA, MU (XLK) · JPM, BAC, BRK-B (XLF) · JNJ, UNH, LLY (XLV) · XOM, CVX, COP (XLE); 2019-08 → 2026-08, 21,024 stock-days. Per-ticker z-scores for event flags (pooled z-scores over-flag volatile names).
- **News**: 290M RavenPack rows → relevance ≥ 50 , de-duplicated, daily sentiment + article count per ticker (70–98% coverage); scheduled vs. unscheduled event features.
- **Features (18)**: three factor families (raw, sector-relative, rolling-relative) pruned to a frozen set with no $|\text{corr}| > 0.5$ — rel. momentum 5d/21d/12–1m, 63d beta: rel. vol, vol ratio, Δ VIX, volume z, days to earnings, days since FOMC — plus 8 news features (sentiment daily/3d/21d/momentum, news-volume z, pending-event repricing, unscheduled shock).

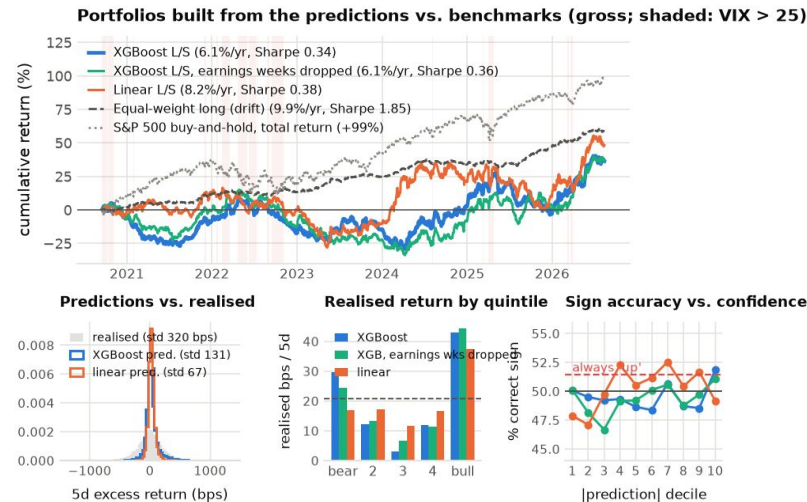
Feature	Definition
rel_mom_5d / 21d	stock minus sector-ETF return over 5 / 21 days
rel_mom_12_1	12-month sector-relative momentum, skipping the last month
beta_63d	63-day rolling beta of daily returns to the sector ETF
rel_vol_21d	21-day realised vol of the sector-relative return
vol_ratio_5_21	5-day over 21-day realised vol of the stock
vix_chg_5d	5-day change in the VIX
volu_z_63d	log-volume z-score over 63 days
days_to_earn	trading days to the next scheduled earnings print
days_since_fomc	trading days since the last FOMC decision
avg_sentiment, news_volume	RavenPack daily mean sentiment, article count
sent_rol_3d / 21d	rolling means of daily sentiment
sent_momentum	3-day minus 21-day rolling sentiment
news_volume_z	article-count z-score over 63 days
pending_event_repricing	5-day change in realised vol within 10 days of a print
unscheduled_shock_dummy	abnormal volume not explained by earnings or FOMC

Target & Walk-Forward Protocol

- **Target**: y = 5-day forward return of the stock minus its sector ETF (continuous).
- **Protocol**: trailing 252-day window, 6-day purge so no training label overlaps a test label; refit monthly (71 refits, 17,760 OOS stock-days) or daily. XGBoost: 200 trees, depth 4, lr 0.05, 80% subsampling. Linear baseline: daily-refit regression on the 10 frozen features.
- **Evaluation**: predictions vs. realised returns out of sample — RMSE/MAE against naive forecasts (zero, each stock's trailing mean), sign accuracy against always-'up', calibration slope, correlation, realised return by prediction quintile; a long-top-3/short-bottom-3 portfolio as a secondary view.

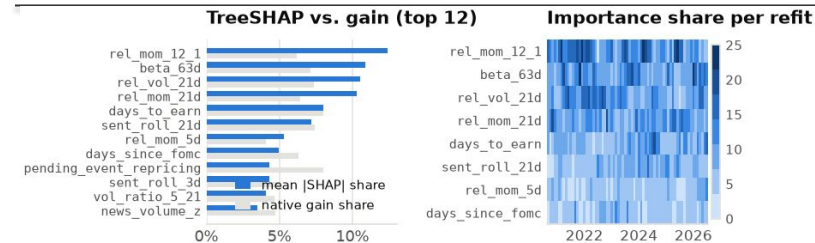
Out-of-Sample Predictions vs. Realised Returns (2020-09 to 2026-08)

Forecast	RMSE bps	R ² vs 0	Sign acc.	corr	Calib. slope
XGBoost (monthly refit)	342	-0.136	49.5%	0.033	0.08
XGBoost (daily refit)	343	-0.136	49.4%	0.048	0.11
XGBoost (earnings weeks dropped)	336	-0.097	49.4%	0.052	0.14
Linear (daily refit)	325	-0.023	50.2%	0.039	0.19
Naive: predict 0	321	0.000	—	—	—
Naive: stock's trailing mean	321	-0.002	50.7%	0.065	0.39
Naive: always 'up'	—	—	51.5%	—	—

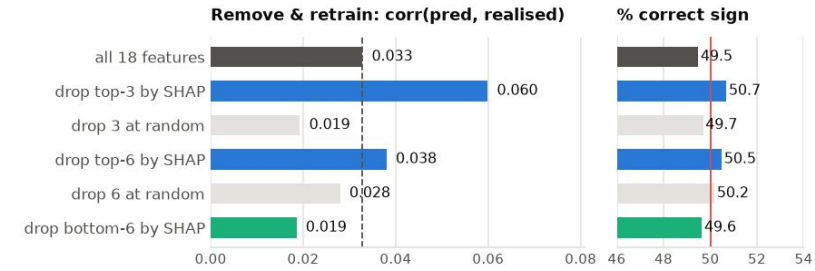


- **Ordering, yes; accuracy, no**. Predicted and realised returns correlate 0.033 and the top quintile realises ~43 bps/5d vs a 21 bps mean, but the forecasts are ten times too large (realised moves 8 bps per 100 predicted): RMSE 342 bps is worse than predicting zero (321) and sign accuracy 49.5% is below always-'up' (51.5%). Each stock's trailing mean alone correlates 0.065 with outcomes: drift.
- **Variants**. Daily refit raises the correlation to 0.048 and dropping earnings weeks from training to 0.052, neither improving RMSE or sign accuracy. Telling the model the ticker overfits (corr 0.017); sector dummies change nothing (0.031); a 36-point tuning grid has negative correlation everywhere on 2020–mid-2023.
- **What it misses**. The largest errors are the jumps: MU's +24–29% weeks (Mar 2024, May 2026) and UNH's –24–28% weeks (Apr 2025, May 2025) against predictions of a few hundred bps; 35% of the largest 1% of errors have an earnings print in the window (8% of all stock-days). By stock, the correlation runs from +0.08 (CVX, MU) to -0.06 (BAC, UNH). Per-stock and largest-miss plots are in the paper.

Explainability: TreeSHAP on 71 Walk-Forward Refits



- **Flat profile, moderate agreement with gain**. rel_mom_12_1 carries 12% of attribution, the top six ≈ 59%, news 28%. Spearman(SHAP, gain) = 0.74: gain ranks the 87%-missing pending_event_repricing joint-first, SHAP ninth. The top feature changes with the training window (same top-1 in 35% of refits).
- **Dependence**. 12-month relative momentum → reversal at the negative end; beta → sharply positive above 1.5 (the NVDA/MU regime): the model learned the sample's outliers and winners.



- **Faithfulness (remove and retrain)**. Dropping the top-3 SHAP features raises the correlation with realised returns from 0.033 to 0.060 and sign accuracy from 49.5% to 50.7%; dropping the bottom-6 lowers the correlation to 0.019. SHAP faithfully describes a model that fitted tails and drift; the out-of-sample information sits in the features it ranks lowest.
- **Stability**. Six seeds: Spearman 0.996, identical top-3. Input noise at 0.1 / 0.25 / 0.5 sd: Spearman 0.98 / 0.96 / 0.94. Across consecutive refits only 0.84: time, not randomness, changes the story.

Diagnostics & Conclusions

- **Why the forecasts look like this**. The target is a Student-t with $v \approx 2.5$ (top 5% of stock-days = 55% of variance), its std spans XOM 1.5% → MU 6.1%, consecutive rows are near-copies ($N_{\text{eff}} \approx N/4.9$; Newey–West SE of a correlation 0.03–0.04), and NVDA/MU/LLY carry the +21 bps mean. Squared-error learners fit tails and drift; no comparison on 12 names is statistically decisive.
- **What the explanations show**. TreeSHAP is faithful to the model and stable under seeds and noise, yet its top features hurt out of sample: importance describes what was fitted; remove-and-retrain tells signal from noise.
- **Next**. Wider universe; evaluation against naive and drift benchmarks by default; LSTM / Chronos-2 with integrated gradients; remove-and-retrain as the primary XAI test.