Stock Predictability around S&P 500 Rebalancing Events

Leo Li Alice He

Under supervision of: Ali Hirsa, Miao Wang, and Bolaji Eniwaye at BlackRock

Background & Objectives

Passive index rebalances create short-lived, systematic microstructure pressure (volume spikes, auction slippage, VWAP drift). We ask whether there is **hour-level** predictability around the window and—crucially—whether it **generalizes across events**.

Scope & disclaimers — Headline metrics are **pre-cost**. MSCI constituents/weights are pending; **S&P 500 is used as a methodological proxy**.

Objectives

- Detect weak but consistent signals robust to class imbalance.
- Validate temporal generalization: train on past rebalances → test on the most recent (*event-level* splits for headline claims).
- Produce calibrated probabilities and practical decision thresholds.
- Align with supervisor guidance: build features from announcement → $T-2$, extend coverage to $T-2 \dots T+2$, expand to ≥ 5 events.

Contributions

- Multi-event design (LOEO & rolling OOS) + **hour-wise threshold calibration**.
- Reproducible pipeline with config/seed snapshots and event-level leakage control.

Data & Split

Events (rebalance day): 2024Q3 (2024-09-20), 2024Q4 (2024-12-20), 2025Q1 (2025-03-21); plan: add ≥ 2 more.
Universe: S&P 500; *label* Added/Deleted via inter-event changes. For full Deleted coverage, **recommend** fetching with $(\text{prev} \cup \text{curr})$ snapshots (*planned*).
Quotes: Alpaca minute OHLCV → hourly aggregation; drop 09:30–10:00 noise.
Targets: v1 uses 14:00/15:00/16:00 next-hour direction; v2 will cover $T-2 \dots T+2$.
Split: *Cross-rebalance* for headline; 80/20 ticker split kept as a quick diagnostic.

Sample sizes per hour (N tickers): Q3: $N = 503$ each; Q4: $N = 501$; Q1: $N = 501$.

```
2024Q3 {'14:00': 503, '15:00': 503, '16:00': 503}
2024Q4 {'14:00': 501, '15:00': 501, '16:00': 501}
2025Q1 {'14:00': 501, '15:00': 501, '16:00': 501}

2024Q3 label balance:
  14:00: N=503, positive_rate=0.565
  15:00: N=503, positive_rate=0.256
  16:00: N=503, positive_rate=0.660

2024Q4 label balance:
  14:00: N=501, positive_rate=0.058
  15:00: N=501, positive_rate=0.515
  16:00: N=501, positive_rate=0.415

2025Q1 label balance:
  14:00: N=501, positive_rate=0.549
  15:00: N=501, positive_rate=0.277
  16:00: N=501, positive_rate=0.657
```

Label balance by hour. 2024Q4 14:00 positive rate $\approx 0.058 \Rightarrow$ naive Accuracy is misleading; use BalAcc/AUROC.

Methods & Features (v1 → v2)

Validation — LOEO & rolling OOS: Q3→Q4; Q3+Q4→Q1. **Models** — **pooled cross-ticker LNN baseline**; other models (e.g., GBDT/MLP) optional for screening. **Loss/metrics** — BCE / weighted BCE; Accuracy, **BalancedAcc**, **AUROC**; ranking-by-bucket (planned).

Feature set (v1)

- Prices/volume: `close_last`, `VWAP_ratio`, `logVolume`, `DollarVolume`, `CumReturn`.
- Event/diurnal: `Was_Added`, `Was_Deleted`, `Minute_Sin/Cos`.
- Microstructure/scale: `LiquidityStress` (median $|\text{ret}|/\text{DollarVol}$), `Implied_AUM` (rebalance-day vol / 7-day median).

Shift to v2 (aligned with supervisor)

- Build features using **announcement** → $T-2$ data only; add **DayRel** one-hot for $T-2 \dots T+2$.
- Normalize `DollarVolume` by **mean/median** to avoid sum-scale bias.
- Add **sector** conditioning; **enforce non-overlap** between feature and prediction windows to avoid endogeneity.

Leakage & governance

- Scale with *train-fold* stats only; compute per-event stats independently (no cross-event leakage).
- Save `config.yaml`/seeds; export `metrics.json`, reliability plots (planned); maintain `SOURCES.md`.

Calibration & Diagnostics

Implemented — **hour-wise decision-threshold calibration**: learn $\tau_{14}, \tau_{15}, \tau_{16}$ on train-event validation folds and *apply unchanged* to the held-out event.

Planned — **temperature scaling** (binary: $\tilde{p} = \sigma(z/T)$) to improve probability calibration; **no-trade band** (skip trades if $\max_c \tilde{p}(c \mid x) < \tau$); **Brier/ECE** and reliability diagrams.

	Event	RDay	NumTrainEvents	TrainSamples	TestSamples	Group	Accuracy	BalancedAcc	F1	AUROC
0	2024Q4	2024-12-20	1	1509	1503	Added	0.533333	0.500000	0.695652	0.464286
1	2024Q4	2024-12-20	1	1509	1503	Deleted	NaN	NaN	NaN	NaN
2	2024Q4	2024-12-20	1	1509	1503	Hour=14:00	0.069860	0.457810	0.100386	0.542519
3	2024Q4	2024-12-20	1	1509	1503	Hour=15:00	0.495010	0.490310	0.568995	0.508549
4	2024Q4	2024-12-20	1	1509	1503	Hour=16:00	0.423154	0.504734	0.586552	0.439732

Rolling OOS log: device/setup, learned per-hour thresholds, and group/hour tables.

Holdout	TrainSamples	TestSamples
2024Q4	1509	1503
2025Q1	3012	1503

Unsuccessful Attempts & Lessons

- **5-bin ordinal target**: unstable at current data scale (macro-F1 low, AUROC ≈ 0.5); **binary + calibration** is more robust.
- **Per-ticker LNN**: marginal gains are small and highly volatile across tickers; maintenance cost is high.
- **80/20 ticker split (53.8% Acc.)**: for quick smoke tests only; *does not* justify cross-event claims (headlines use event-based splits).

Headline Results & Slices

Cross-event summary (headline)

Event	Accuracy	BalancedAcc	AUROC
2024Q4	0.330	0.500	**0.615**
2025Q1	0.490	0.490	0.490

- 2024Q4 shows ranking skill despite skew; 2025Q1 is near-random ⇒ regime shift / threshold migration issues.
- Small-N slices (e.g., *Added*/*Deleted*) have wide CIs; always report N with metrics.

	Accuracy	BalancedAcc	F1	AUROC	Event
0	0.330007	0.499468	0.494732	0.614632	2024Q4
1	0.485695	0.485020	0.449822	0.487444	2025Q1

Q4 per-hour (holdout)

Hour	Acc.	BalAcc	AUROC	Pos. rate
14:00	0.07	0.46	0.54	5.8%
15:00	0.50	0.49	0.51	51.5%
16:00	0.42	0.50	0.44	41.5%
Overall	0.33	0.50	**0.61**	–

Threats to Validity

- **Label skew & threshold bias**: extreme class imbalance at 14:00 depresses raw Accuracy; AUROC better reflects ranking value.
- **Deleted coverage gap**: minutes for *Deleted* names require $(\text{prev} \cup \text{curr})$ snapshots (*planned fix*).
- **Proxy bias**: using S&P as a proxy for MSCI may dilute liquidity/metadata signals.
- **Small-N slices**: *Added*/*Deleted* groups are small; report N and uncertainty intervals.
- **Regime shift / portability**: 2025Q1 cross-event performance ≈ random, suggesting threshold/probability non-portability.
- **Architecture sensitivity**: pooled cross-ticker vs per-ticker models react differently to structural mismatch.

Key Takeaways & Next Steps

Takeaways

- Ranking signal (AUROC) often emerges before Accuracy; handle hour-wise imbalance (esp. 14:00).
- **Binary + threshold calibration** > coarse 5-bin ordinal at current data scale.
- Event-based splits (LOEO / rolling OOS) are mandatory for credible claims; ticker 80/20 is diagnostic only.

Next (aligned with supervisor)

- Expand to ≥ 5 rebalances; features from announcement → $T-2$; add **sector** conditioning.
- Report **Brier/ECE** and *cost-aware* outcomes (commission, half-spread/impact, participation caps; auction vs. continuous).
- **Integrate newly acquired Alpaca minute OHLCV (2019Q3–2025Q1, [-15,0] days)** for multi-event training (currently not integrated).
- Enforce strict **non-overlap** between feature and prediction windows; evaluate threshold robustness and portability.
- Ship shared folder + `SOURCES.md`; keep config/seed snapshots for audit.

Limitations: S&P proxy while MSCI access pending; Deleted coverage depends on snapshot union (planned fix); thresholds/calibration do not migrate well into 2025Q1.