

How Much Evidence Does a Backtest Really Have?

Gene Yu September 2026

ABSTRACT

Can technical features predict ETF returns three to six months ahead? A random forest plus XGBoost ensemble is trained on 162 ETFs, stocks and indices in eight factor categories, with a shared calendar split with 126-session embargoes and an evaluation protocol fixed in code before the final test. Out-of-sample ranking looks strong, with ROC-AUCs up to 0.90. Yet only two of eight categories produce a decision rule on validation, and neither beats the alternative of ignoring the model on test: the edges are +1.71 and +1.34 block-bootstrap standard errors against a bar of 2. The limit is argued to be the amount of independent evidence, not the model. Correlated tickers leave each category with only 1.0 to 2.4 effective independent series, and a multi-month hold leaves roughly 53 to 99 independent outcomes per category across the whole price history. Four ways a backtest overstates its evidence are measured: per-ticker splits that leak up to 52% of test rows, naive standard errors that are 2–3 times too small, a peak label that yields 2.3–3.1 times more positives, and longer holds that halve the evidence. One result crossed the bar only because a threshold was rounded. A pre-registered forward test is proposed in closing.

0.90 best out-of-sample ROC-AUC	2 of 8 categories that produce a trading rule	+1.71 / +1.34 test edges, in standard errors, against a bar of 2	53–99 independent outcomes per category, whole history
--	--	--	--

1 Introduction

A backtest with thousands of trades can hold only a few dozen independent observations.

Trades that overlap in time share most of their outcome window and tickers in the same category move together, both of which make a result look more certain than it is, and neither shows up in a trade count or a win rate.

This paper measures that gap on one system. The model was built to predict whether an ETF would be materially higher three to six months after entry. It was then tested under progressively stricter checks until nothing survived.

The model is a random forest plus XGBoost ensemble, trained on technical features across eight factor categories and 162 symbols. Although early versions looked strong, with high win rates, large compounded returns, and out-of-sample ROC-AUCs up to 0.90, under a protocol fixed before the final test, two of eight categories produced a decision rule, and neither beat the simple alternative of ignoring the model (+1.71 and +1.34 standard errors against a bar of 2).

The binding constraint is the amount of independent evidence, not the model. The contributions are four measured quantities:

- Splitting each ticker's history separately put up to 52% of test rows on dates that another ticker in the same category had trained on. A shared calendar split removes this entirely.
- Treating overlapping, correlated trades as independent understates the standard error by about 2 to 3 times for small-cap ETFs, but only about 1.1 times for international and emerging-market ETFs.
- The whole price history holds roughly 680 to 1,250 independent outcomes per category at a 3–10 session hold, 53 to 99 at 3–6 months, and 26 to 49 at 6–12 months.
- Copying one probability floor to six decimal places instead of deriving it exactly moved an out-of-sample result from +1.34 to +2.33 standard errors, across the bar.

The rest of the paper is organized as follows. Section 2 places this work among existing tools for dependent financial data. Sections 3 and 4 describe the data and the evaluation protocol. Section 5, the core, measures four ways the evidence gets inflated. Section 6 reports results, Section 7 discusses why the limit is structural, and Section 9 proposes a pre-registered forward test.

2 Background and related work

Every tool this paper uses already exists. The contribution is to apply them together to one system and report how much each one changes the answer.

López de Prado (2018, ch. 7) shows that standard k-fold cross-validation leaks information when labels span time. He proposes purging, which drops training observations whose label window overlaps the test period, and an embargo, which drops a buffer of observations after each test fold. The 126-session gaps between splits are an embargo of exactly one label window. Section 5.1 measures a related leak across correlated tickers on the same calendar date, which purging within one series does not catch.

Bailey, Borwein, López de Prado and Zhu (2017) define the probability of backtest overfitting as the chance that the configuration chosen as best in-sample ranks below the median out of sample, and estimate it with combinatorially symmetric cross-validation. Bailey and López de Prado (2014) propose the deflated Sharpe ratio, which corrects a Sharpe ratio for selection among many trials and for non-normal returns. Both tools address the risk this paper studies, but neither statistic is computed here. Instead, their shared premise, that the number of configurations tried must enter the significance bar, is adopted and applied through a permutation test over candidate floors (Section 4.3).

Harvey, Liu and Zhu (2016) catalogue hundreds of published return factors and argue that a new factor should clear a t-ratio above 3.0, not the conventional 2.0. The bar of 2 standard errors used here is therefore lenient by their standard. A Bonferroni bar of 2.73 across the eight categories is reported as a descriptive reference only.

The block bootstrap of Künsch (1989) resamples blocks of consecutive observations so that dependence within a block survives resampling. Politis and Romano (1994) randomize the block length to make the resampled series stationary. Fixed blocks of 182 calendar days are used here, one holding period, so that trades whose outcome windows overlap fall in the same block.

Kish (1965, pp. 161–162, 258) formalized the design effect, the ratio of a design's variance to that of a simple random sample of the same size. For clusters of b elements with intraclass correlation ρ , it equals $1 + (b - 1)\rho$, and the sample size divided by it gives an effective sample size. The formula is borrowed here to count how many independent series a category of n correlated tickers holds (Section 3). This is an analogy: Kish's ρ is a within-cluster correlation of a sampled quantity, while the one used here is the mean absolute pairwise correlation of daily returns.

Niculescu-Mizil and Caruana (2005) show that boosted trees and random forests often produce distorted probabilities, and that Platt scaling or isotonic regression can correct them. They also find that isotonic regression overfits when the calibration set is small, roughly below 200 to 1,000 cases, where Platt scaling does better. That finding matters here, because several of the calibration slices in this study hold only tens to a few hundred positive examples (Section 6.1).

This paper proposes no new method. It is a measured case study of one system, one frozen dataset, and the size of each correction on it.

3 Data

The 162 symbols in eight categories hold between 1.04 and 2.35 effective independent series each. That range, not the symbol count, sets how much evidence the data can give.

Each category is a factor exposure: credit conditions, energy and commodities, growth and technology, inflation and safe havens, international and emerging markets, broad market beta, rates and recession, and small caps. ETFs are used wherever possible. Single stocks bring survivorship bias, because today's listed names are the ones that survived. Three categories depart from this. `growth_tech` holds 16 single stocks beside 11 ETFs. `energy_commodity` holds one stock (CVX). `market_beta` and `small_cap` include three price indices (GSPC, DJI and RUT), which give long histories but cannot be traded directly. Appendix B lists every symbol.

Prices are daily, adjusted, and frozen in a snapshot downloaded on 2026-09-04. A manifest records the SHA-256 of all 484 data files; re-downloading does not reproduce them, because adjusted prices are revised after each dividend. Illiquid ETFs were removed on the grounds that returns measured on an untradable fund are not returns anyone could have earned. The screen was applied by hand-editing the ticker lists, not in code, so it

cannot be reproduced from the repository. It removed 36 ETFs using a liquidity filter of at least \$5M in average daily dollar volume.

The three price indices were kept and were in effect exempt from the screen. Their reported volume is an exchange-wide aggregate rather than trading in the index, so dollar volume cannot be measured for them, and the screen's own rationale would exclude them. Section 8 discusses the consequences.

All eight categories share one pair of calendar cutoffs. Train runs from each category's first date to 2017-01-24, validation from 2017-07-26 to 2021-11-09, and test from 2022-05-12 to 2026-09-03. Each gap is 126 sessions, one full label window, and rows inside it are dropped because their labels would read across the seam.

Table 1. Universe, splits and effective series (T1, T2).

Category	Symbols	Train starts	Train rows	Validation rows	Test rows	Mean pairwise corr.	Mean absolute corr.	Effective series
credit_conditions	22	2002-07-30	39,078	23,739	23,804	0.58	0.58	1.68
energy_commodity	20	1962-01-02	59,213	21,640	21,640	0.50	0.50	1.92
growth_tech	27	1962-01-02	152,378	29,214	29,214	0.54	0.54	1.80
inflation_safe_haven	13	2003-12-05	29,355	13,614	14,066	0.26	0.38	2.35
international_emerging	27	1996-03-18	108,670	29,214	29,214	0.69	0.69	1.42
market_beta	24	1927-12-30	90,861	25,742	25,968	0.79	0.79	1.24
rates_recession	14	2002-07-30	33,036	15,148	15,148	0.21	0.56	1.68
small_cap	15	1987-09-10	54,672	16,230	16,230	0.95	0.95	1.04

Rows are symbol-days summed across the category. Train start is the earliest symbol in the category; most symbols start much later. Effective series uses the mean absolute correlation.

Effective series. Symbols in a category move together, so they are not independent draws. Following Kish's design effect, effective series are counted as

$$n_{eff} = \frac{n}{1 + (n-1)\bar{\rho}}$$

where n is the number of symbols and ρ is the mean absolute pairwise correlation of daily returns. Returns are computed within each split file, so none spans an embargo gap. `small_cap`'s 15 symbols have a mean correlation of 0.95 and amount to 1.04 series: in effect, one.

The absolute value matters for categories that hold inverse funds. `rates_recession`'s 14 symbols range from T-bills to leveraged inverse Treasuries (TBT, TMV). A fund short the Treasuries beside it is the same bet with its sign flipped, but its negative correlation pulls the signed mean down to 0.21, which would count the category as 3.70 series. In absolute value the mean is 0.56, and the category amounts to 1.68. `inflation_safe_haven`, which holds the dollar fund UUP beside foreign-currency funds, falls from 3.13 to 2.35 in the same way. Three daily measures were computed: signed, absolute, and an eigenvalue effective rank. The rule, fixed before any of them was computed, was to report the most conservative, which is the absolute one.

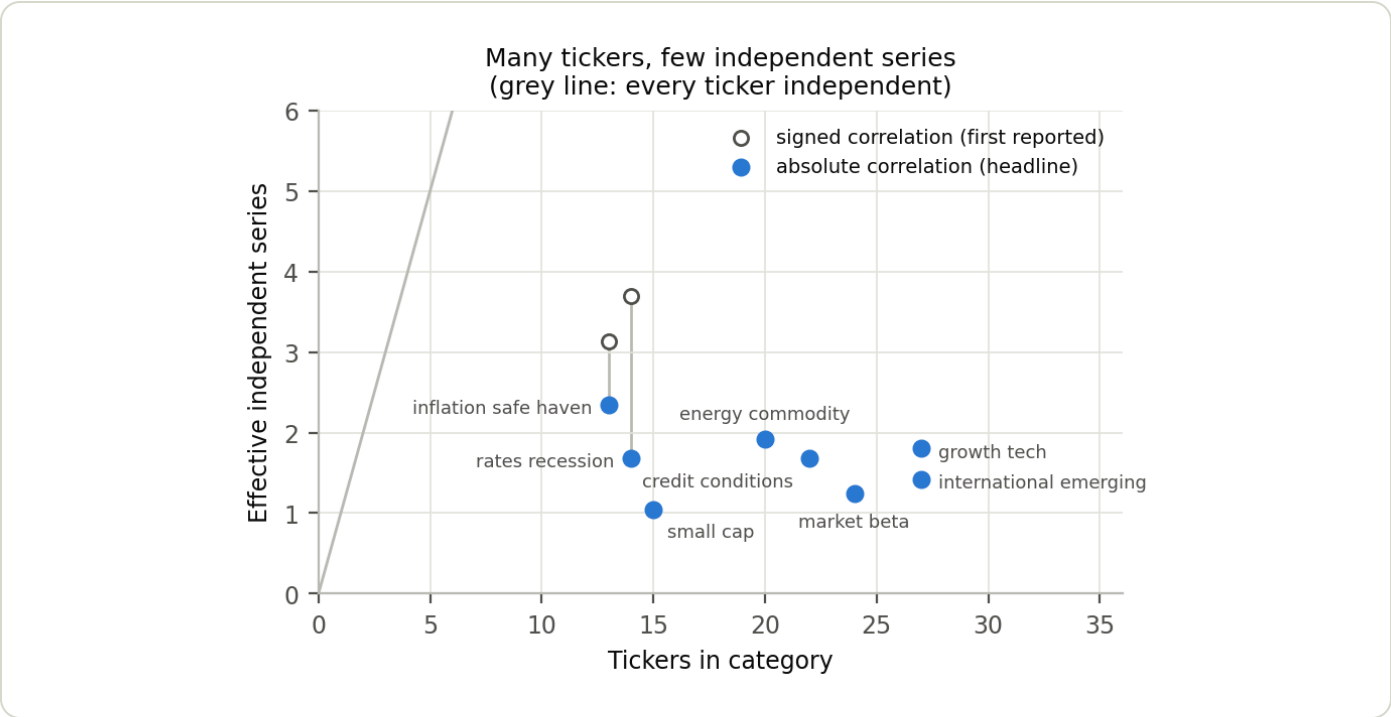


Figure 1. Symbols per category against effective independent series. Filled circles use absolute correlation (the headline); open circles show the signed measure where it differs. Adding correlated symbols barely moves the count.

4 Method

The model is ordinary; the evaluation protocol is the point. Every rule below was fixed in code before the run that produced this paper's tables scored the test split.

4.1 Label

A row is positive if the median close over sessions 63 to 126 after entry exceeds the entry close by at least the category's swing threshold. Thresholds range from 10% (credit_conditions) to 50% (growth_tech) and were set so that roughly 6–7% of training rows are positive (Table 4).

The more common peak label, which asks whether the maximum high in the window crosses the threshold, was rejected. A peak can last a single day. A strategy with a fixed holding period cannot reliably sell into it, so the peak label rewards moves the strategy does not capture. It also produces 2.3 to 3.1 times as many positives at the same threshold (Section 5.3). The median over the second half of the window asks a question closer to the trade: is the price still up when the trade is allowed to exit?

4.2 Model

Each category gets its own ensemble of a random forest and an XGBoost classifier, trained on technical features that do not depend on price level, plus market-wide context. Seeds are fixed (42 for both learners).

Raw tree-ensemble scores are not probabilities (Niculescu-Mizil and Caruana, 2005), so each model is calibrated on a held-out slice of the training period. Isotonic regression is used when the slice holds enough positives, Platt (sigmoid) scaling when it holds fewer (energy_commodity, 44 positives), and no calibration when it holds too few to fit either (credit_conditions with 7, inflation_safe_haven with 0). An uncalibrated category has no probabilities, so no probability floor can be read from it.

4.3 Evaluation protocol

The protocol turns a ranking into a trading rule on validation, then scores that rule once on test.

1. Every trade the model would open with no threshold is binned by predicted probability. A bin must hold at least 25 trades. Each bin boundary is a candidate floor. For each candidate, the separation is the mean return of trades above the floor minus the mean return of trades below it.
2. Trades are grouped into blocks of 182 calendar days by entry date, one holding period, and blocks are resampled 2,000 times. Trades whose outcome windows overlap mostly fall in the same block. With fewer than 6 blocks, no standard error is reported.
3. Picking the best of several candidate floors inflates its t-statistic. Returns are permuted across blocks 2,000 times, the best-of-candidates statistic is recorded each time, and the chosen floor is required to beat its 95th percentile. A floor ships only if it clears that bar.
4. The shipped floor is applied to test once. The null runs the same entry and exit machinery with the floor at zero, i.e. every trade the model would open if its ranking were ignored. The edge is mean return per trade, model minus null.
5. The two-sided 5% Bonferroni bar is also reported across eight categories, 2.73 standard errors, as a descriptive reference. It decides nothing.

Floors are derived exactly by solve_threshold and read from the run's own output, never retyped. Section 7.2 shows why that rule exists.

5 Four ways the evidence gets inflated

Each of the four mistakes below makes a backtest look better supported than it is. Each subsection gives the mistake, the measured size of its effect on this system, and the fix.

5.1 Split leakage

In an earlier version of this project, each symbol's own history was split separately, with its oldest 55% of rows assigned to train and its newest 30% to test. Symbols start on different dates, so one symbol's test period overlapped another symbol's training period. Because symbols in a category are highly correlated, the model had in effect already seen those test days.

Table 2 gives the share of test rows dated on a calendar day that some other symbol in the same category was trained on. It reaches 52.3% for market_beta and 49.7% for growth_tech, and is lowest for rates_recession at 4.5%.

Table 2. Share of test rows on dates another symbol trained on (T3).

Category	Per-symbol split	Shared calendar split
market_beta	52.30%	0.00%
growth_tech	49.68%	0.00%
inflation_safe_haven	42.94%	0.00%
small_cap	25.03%	0.00%
international_emerging	24.75%	0.00%
credit_conditions	23.49%	0.00%
energy_commodity	17.06%	0.00%
rates_recession	4.51%	0.00%

This leakage is removed by using one pair of calendar cutoffs per category, shared by every symbol, with a 126-session embargo at each seam. Leakage falls to 0% by construction. These figures are measured on the current universe and data; the original project measured 39–49% on an older, smaller universe.

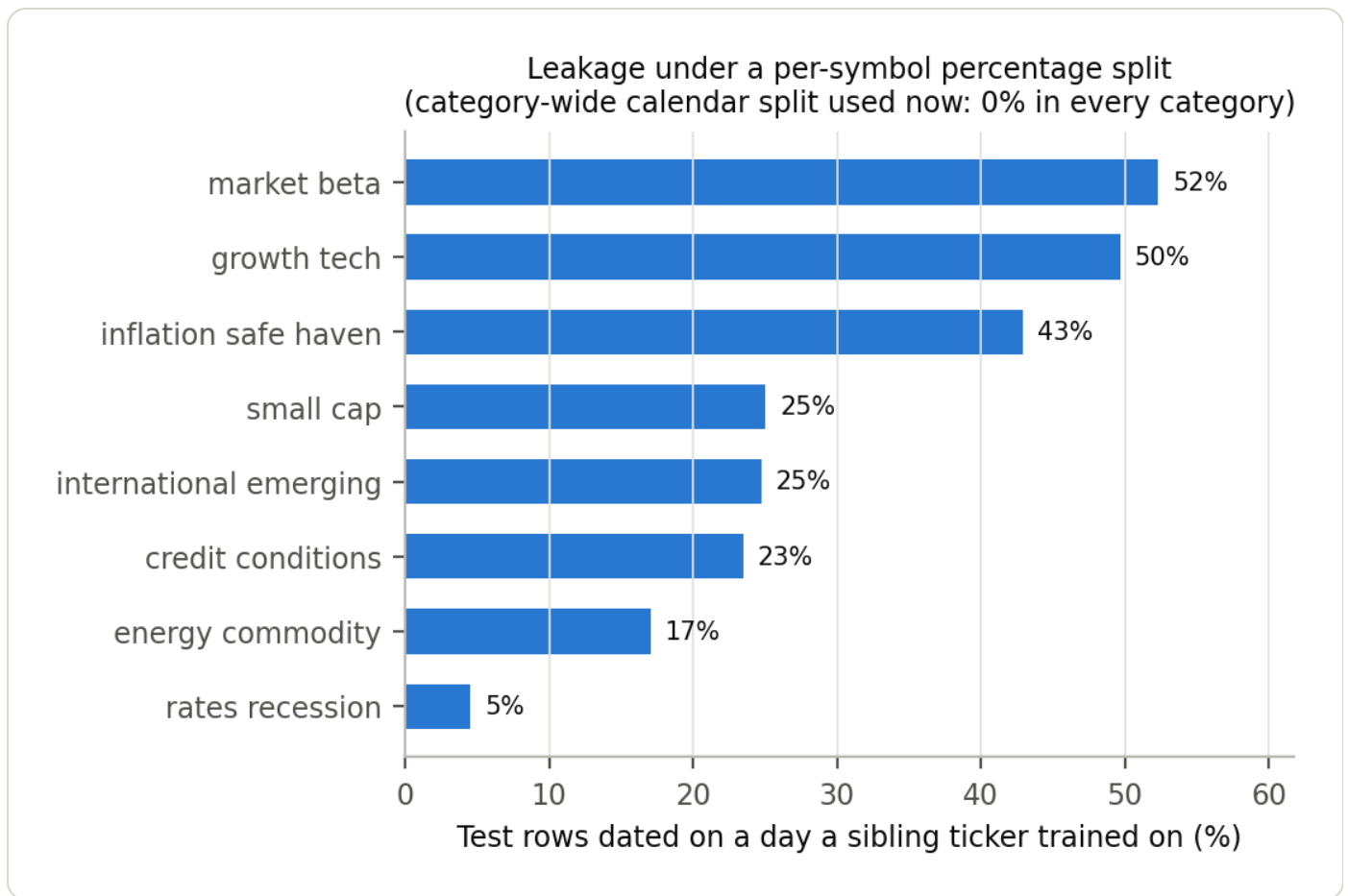


Figure 2. Test-row leakage under the per-symbol split (old) and the shared calendar split (current).

5.2 Correlated, overlapping trades

Standard errors are often computed as if every trade were independent. Two things break that assumption: consecutive entries in the same symbol share most of a 63–126 session outcome window, and entries in correlated symbols on the same day share the same market move.

Table 3 compares the naive standard error with the block-bootstrap one for the two categories that reach a test. For `small_cap`, the naive error is about a third of the real one on validation and half of it on test. For `international_emerging`, the two differ by only 6–8%.

Table 3. Naive vs block-bootstrap standard errors (T7, T8).

Category	Stage	Naive SE	Block SE	Ratio	t if trades were independent	t with block SE
small_cap	Validation separation	1.90%	5.74%	3.0×	8.3	2.75
small_cap	Test edge	1.34%	2.75%	2.1×	2.74	1.34
international_emerging	Validation separation	2.73%	2.89%	1.06×	3.2	3.01
international_emerging	Test edge	1.70%	1.84%	1.08×	1.85	1.71

The unevenness has a visible cause. `small_cap`'s 119 validation trades fall on just 27 entry dates, and its 15 symbols amount to 1.04 effective series, so most trades are copies of one bet. `international_emerging`'s 197 trades spread over 75 entry dates and 1.42 effective series. The practical consequence is stark: `small_cap`'s test edge would read +2.74 standard errors under the naive error, just above even the Bonferroni bar of 2.73. With the block error it is +1.34.

This is corrected by resampling whole 182-day blocks of entry dates, so that overlapping and same-day trades move together.

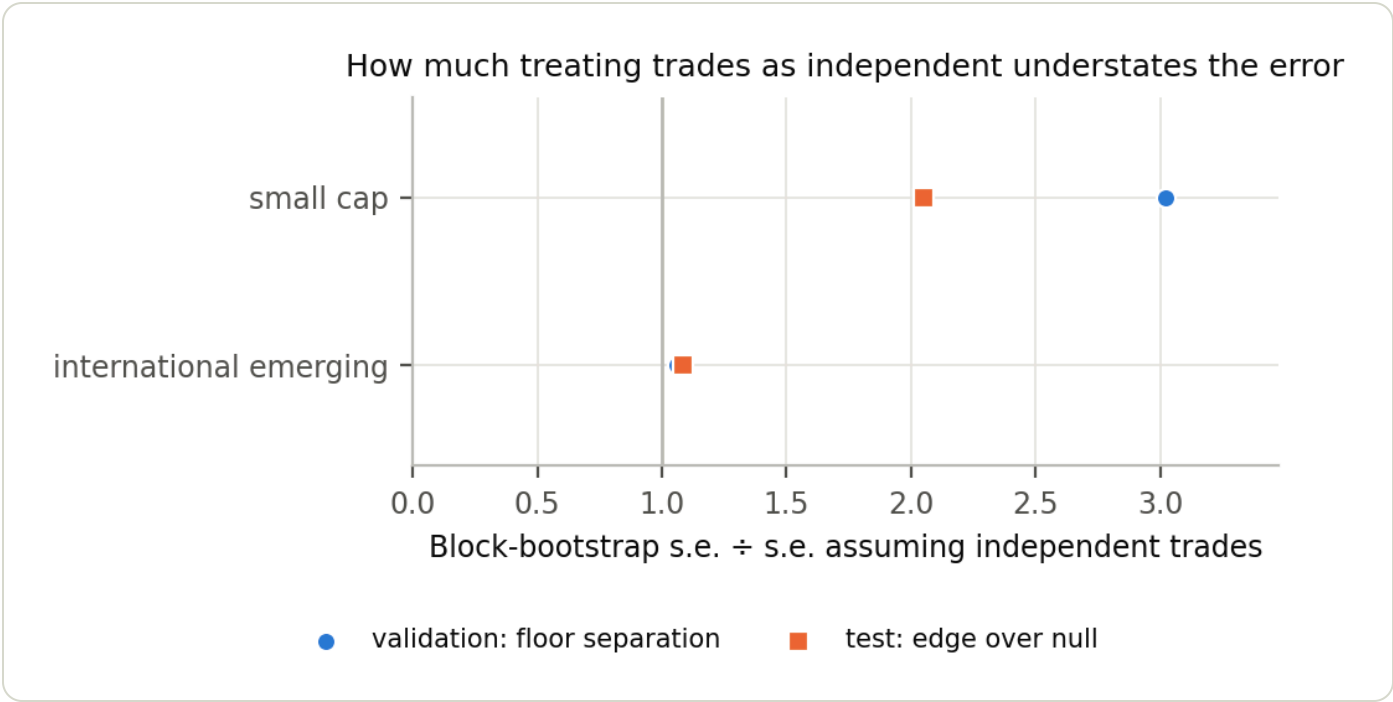


Figure 3. Ratio of block-bootstrap to naive standard error.

5.3 Label definition

A trade is often labelled a success if the price touched the target at any point in the window (the peak label), when the strategy can only exit at the end of a fixed hold.

At the same threshold and horizon, the peak label marks 2.3 to 3.1 times as many training rows positive as the terminal label (Table 4). A model trained on it learns to predict spikes the strategy cannot sell into, and its hit rate overstates what trades would earn.

Table 4. Positive-label rate on train, terminal vs peak, 63–126 sessions (T4).

Category	Threshold	Terminal	Peak	Peak / terminal
small_cap	20%	5.92%	18.11%	3.06
energy_commodity	25%	6.90%	20.27%	2.94
market_beta	15%	7.54%	21.66%	2.87
international_emerging	25%	6.64%	17.85%	2.69
credit_conditions	10%	5.62%	15.06%	2.68
rates_recession	12%	6.15%	16.31%	2.65
inflation_safe_haven	15%	7.45%	17.43%	2.34
growth_tech	50%	7.44%	17.12%	2.30

At a 126–252 session horizon the ratio is smaller, 1.90 to 2.47, because a longer window gives the terminal price more time to catch up with the peak.

The terminal label of Section 4.1 is used instead, since it asks whether the price is still above the threshold during the period in which the trade can exit.

5.4 Holding horizon

Holding periods are sometimes lengthened without the evidence being recounted. A longer hold means fewer non-overlapping trades in the same history, and correlation across symbols cannot make up the difference.

Multiplying non-overlapping holds by effective series gives a rough count of independent outcomes in each category's whole history, before any split (Table 5). Moving from a 3–10 session hold to a 3–6 month hold cuts it by about 13 times; moving to 6–12 months halves it again.

Table 5. Independent outcomes over the whole history, by holding horizon (T5).

Category	3–10 sessions	63–126 sessions (3–6 months)	126–252 sessions (6–12 months)
growth_tech	1,254	99	49
inflation_safe_haven	1,152	89	45
energy_commodity	948	75	36
international_emerging	941	74	37
market_beta	809	63	31
rates_recession	786	62	30 (cannot be split)
credit_conditions	732	57	28 (cannot be split)
small_cap	680	53	26

At 6–12 months, credit_conditions and rates_recession do not even have enough history for the project's train, validation and test sizing rule: they need 4,884 sessions and have 4,374 and 4,680.

No fix exists inside the data, only a choice. At 63–126 sessions every category still fits the three-way split; at 126–252 two do not. The count of independent outcomes should be reported beside every result, so a reader can see how little a given horizon allows.

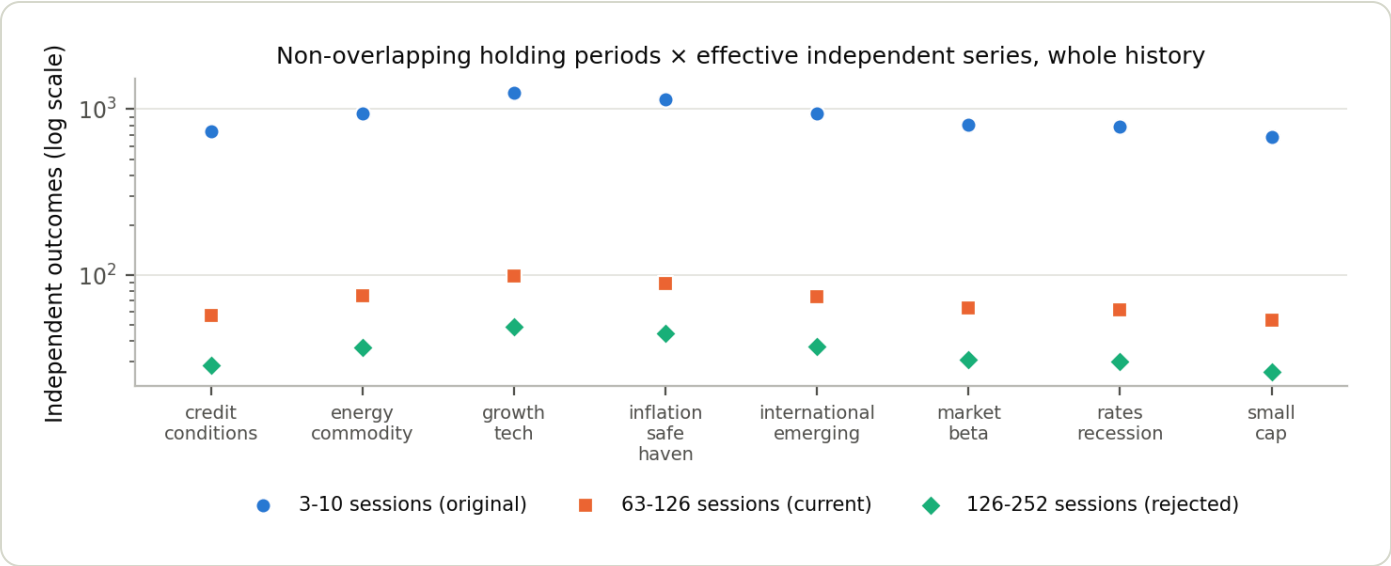


Figure 4. Independent outcomes available at each holding horizon, by category.

6 Results

The models rank well out of sample, only two of eight categories yield a trading rule, and neither rule beats ignoring the model by 2 standard errors on test.

6.1 Calibration and ranking

credit_conditions and inflation_safe_haven cannot be calibrated: their calibration slices hold 7 and 0 positive examples. Their scores still rank (test ROC-AUC 0.889 and 0.826), but they are not probabilities, so no floor can be read from them.

The other six have test ROC-AUCs from 0.557 to 0.897 (Table 6). Five of the six are above 0.70, which looks strong for a return-prediction model.

Table 6. Calibration and ranking on the test split (T6).

Category	Calibration	Calibration positives	Test base rate	ROC-AUC	PR-AUC lift	ECE
small_cap	isotonic	303	3.87%	0.897	7.43	0.036
credit_conditions	none	7	3.00%	0.889	5.42	0.011
rates_recession	isotonic	181	4.81%	0.887	4.71	0.021
inflation_safe_haven	none	0	14.52%	0.826	2.93	0.037
growth_tech	isotonic	513	7.12%	0.823	3.71	0.042
international_emerging	isotonic	287	3.62%	0.774	2.84	0.027
market_beta	isotonic	846	10.37%	0.708	2.58	0.066
energy_commodity	sigmoid	44	14.83%	0.557	1.13	0.132

PR-AUC lift is PR-AUC divided by the base rate; 1.0 is random ranking.

The calibrated models under-predict, all six of them (Figure 5). Every reliability curve lies above the diagonal. The clearest case is small_cap: its top decile of test rows had a mean predicted probability of 1.5%, and 24.2% of those 1,165 rows turned out positive. energy_commodity predicts 0.2–12% across its deciles while 7–18% came true, with almost no slope.

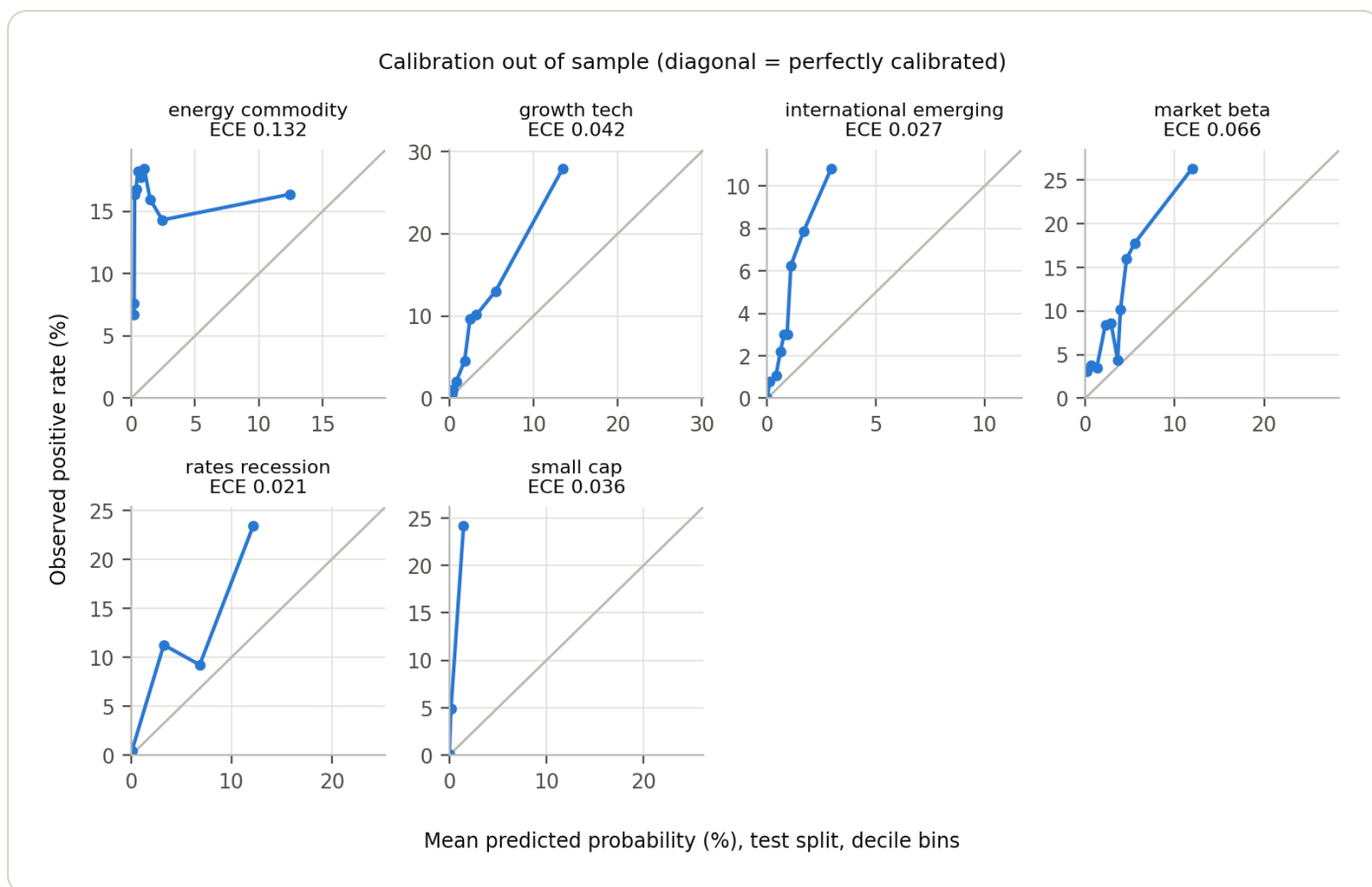


Figure 5. Out-of-sample calibration, test split, decile bins. The diagonal is perfect calibration.

These AUCs deserve less weight than they appear to. Test rows overlap by up to 99% of their outcome window, so each category's test set is one heavily overlapping path through 2022–2026, not a sample of many. No standard errors are attached to them for that reason.

6.2 Floor search on validation

Four of the six calibrated categories cannot produce a floor at all. Their calibrated probabilities pile up on a few plateaus, so no more than one probability bin reaches 25 trades, and a marginal-return curve needs at least two (Table 7). This is a shortage of distinct predictions, not of trades as such: `small_cap` produced a floor from 119 trades while `growth_tech` could not from 181.

Table 7. Floor search on validation (T7).

Category	Trades	Entry dates	Bins ≥ 25 trades	Candidates	Floor	Separation	t (block SE)	Bar after search
international_emerging	197	75	≥ 2	2	1.875%	+8.69%	3.01	2.60
small_cap	119	27	≥ 2	1	0.129%	+15.80%	2.75	2.53
growth_tech	181	59	1	0	–	–	–	–
market_beta	180	79	1	0	–	–	–	–
energy_commodity	174	104	0	0	–	–	–	–
rates_recession	105	45	1	0	–	–	–	–

The two survivors both clear their search-corrected bar on validation. At international_emerging's floor, 50 trades above it earned +9.40% on average and 147 below it earned +0.71%. At small_cap's floor, exactly 1/776, 73 trades above earned +11.18% and 46 below lost 4.62%. Both rest on only 6 date blocks.

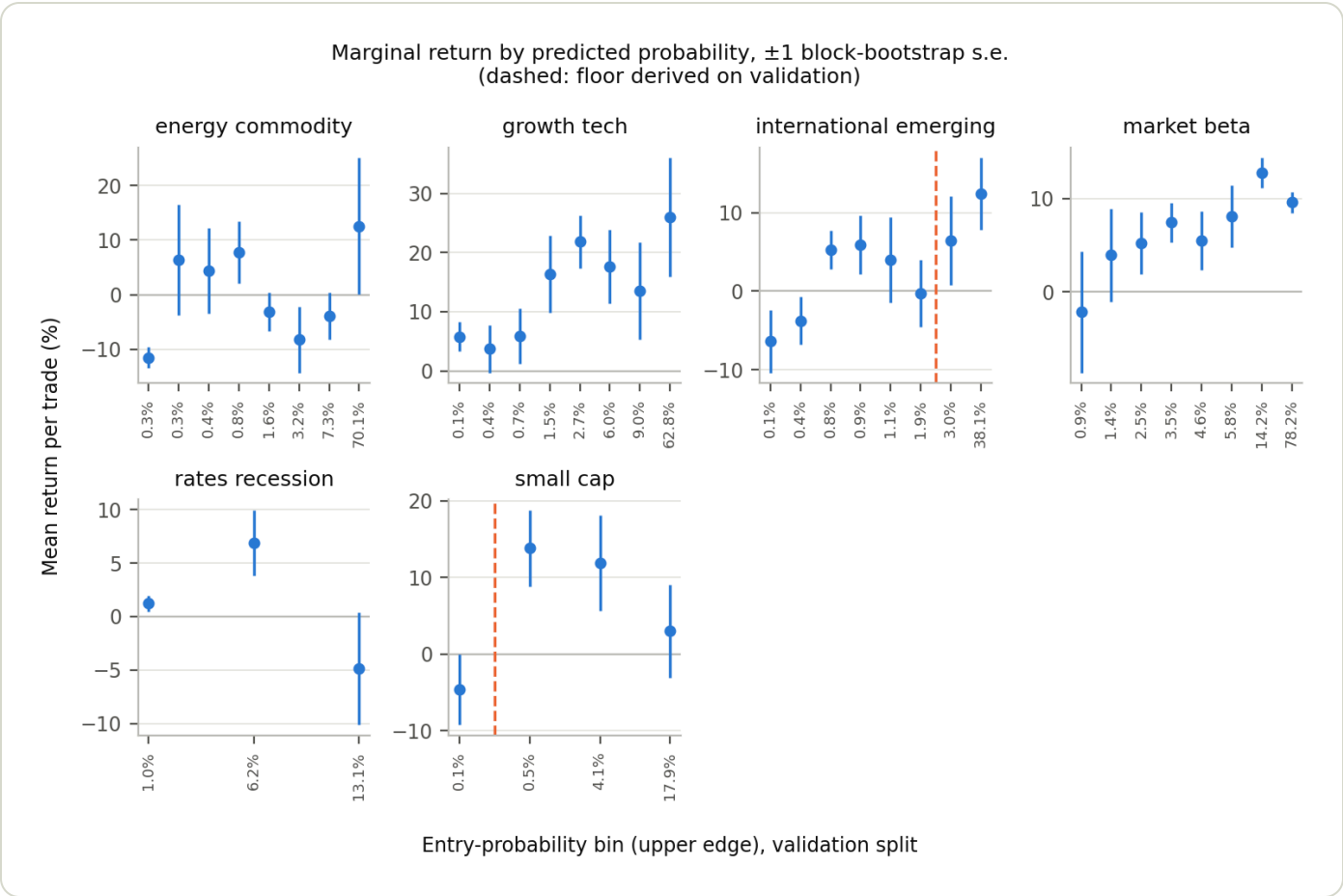


Figure 6. Marginal return by predicted-probability bin, validation split, with block-bootstrap error bars.

6.3 Out of sample

Neither rule clears the bar on test (Table 8). international_emerging's trades above the floor earned +10.77% against +7.62% for the model-off null, an edge of +3.14% per trade at +1.71 standard errors. small_cap's earned +8.90% against +5.22%, an edge of +3.68% at +1.34 standard errors.

Table 8. Test: model at its validation floor vs the model-off null (T8).

Category	Model trades (entry dates)	Model mean	Null trades	Null mean	Edge	Block SE	t	p (one-sided)	Usable blocks
international_emerging	102 (79)	+10.77%	178	+7.62%	+3.14%	1.84%	+1.71	0.044	7
small_cap	97 (31)	+8.90%	90	+5.22%	+3.68%	2.75%	+1.34	0.091	6

Both edges are positive, and both are well short of 2 standard errors, let alone the Bonferroni reference of 2.73. The null itself earned 5–8% per trade: 2022–2026 rewarded simply holding these ETFs, and the model's selection adds a small, uncertain amount on top. The win rates confirm it: 78.4% for the international_emerging model against 79.8% for its null.

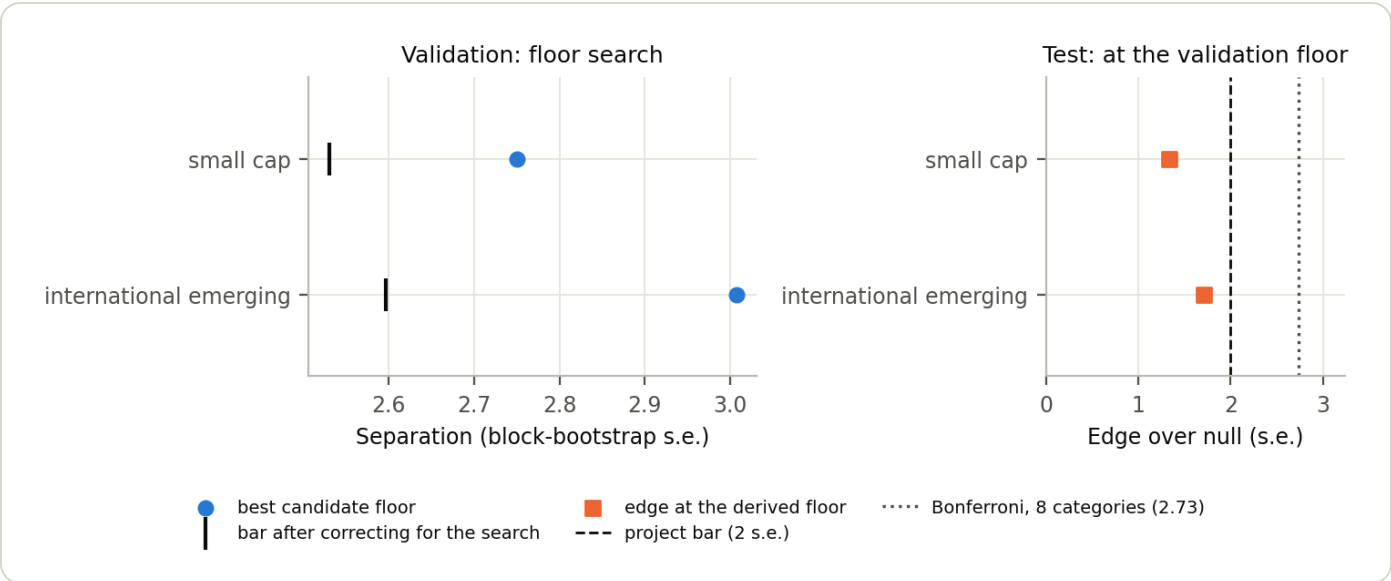


Figure 7. Left: validation separation against the search-corrected bar. Right: test edge against the project bar (2) and the Bonferroni reference (2.73).

6.4 Stability

Walk-forward cross-validation on pre-test data shows how much ranking quality moves between periods (Table 9). Each of five expanding folds trains on everything before its cutoff and scores the next, embargoed block; test data is never read.

Table 9. Walk-forward ranking quality, 5 folds (T9).

Category	ROC-AUC mean	ROC-AUC SD	Folds with ROC-AUC < 0.5	PR-AUC lift, min–max
rates_recession	0.843	0.087	0	1.52–10.46
credit_conditions	0.831	0.073	0	2.57–12.34
growth_tech	0.793	0.073	0	1.67–5.80
small_cap	0.750	0.077	0	1.54–4.71
inflation_safe_haven	0.745	0.114	0	1.20–3.27
market_beta	0.704	0.156	1	0.91–3.00
energy_commodity	0.668	0.084	0	1.02–3.05
international_emerging	0.620	0.173	2	0.93–3.01

international_emerging, the category with the stronger test edge, is the least stable ranker: its ROC-AUC falls below 0.5, worse than random, in 2 of 5 folds. Fold-level trading edges are not reported, because each fold holds too few non-overlapping holding periods to resample.

6.5 Timing or sorting?

A ROC-AUC compares every pair of one positive and one negative row and asks how often the positive is ranked higher. On the test split, 93% to 97% of those pairs come from two different symbols (Table 10). A model can win such a pair by knowing which symbols tend to make large moves. Only pairs drawn from the same symbol measure timing.

When it is split this way, the categories separate. The same-symbol AUC for small_cap is 0.896, almost exactly its pooled 0.897, so its ranking is timing. credit_conditions (0.800), rates_recession (0.778) and market_beta (0.720) keep most of theirs. growth_tech, international_emerging and inflation_safe_haven fall to 0.60–0.65, and energy_commodity to 0.497, no better than chance.

Table 10. Test ROC-AUC split into pairs from different symbols and pairs from the same symbol (T10).

Category	Pooled AUC	Pairs from different symbols	AUC, different-symbol pairs	AUC, same-symbol pairs	Symbols with no test positives
small_cap	0.897	93.36%	0.897	0.896	0
credit_conditions	0.889	95.93%	0.893	0.800	8
rates_recession	0.887	94.21%	0.893	0.778	8
inflation_safe_haven	0.826	93.91%	0.841	0.597	7
growth_tech	0.823	96.83%	0.829	0.646	10
international_emerging	0.774	96.82%	0.778	0.644	3
market_beta	0.708	96.10%	0.708	0.720	2
energy_commodity	0.557	95.56%	0.560	0.497	2

Same test rows as Table 6. The pooled AUC is the pair-weighted average of the two parts. Every pair involving a symbol with no test positives is a different-symbol pair.

Table 11 asks the same question from the other side. Logistic-regression baselines are fitted on train with the same labels, using only the symbol's identity, only its 126-session realized volatility, or both. In several categories that simple information already does most of the work. For rates_recession, volatility alone scores 0.933 against the model's 0.887. For international_emerging, ticker and volatility together score 0.797 against 0.774, and for energy_commodity every baseline beats the model. Counting only pairs that fall in the same decile of the ticker-and-volatility score, where the baseline has little to say, the model still ranks at 0.887 for small_cap and 0.808 for rates_recession, but at 0.614 for international_emerging and 0.516 for energy_commodity.

Table 11. Test ROC-AUC of the model against ticker and volatility baselines (T11).

Category	Model	Ticker only	Volatility only	Ticker + volatility	Model, within baseline deciles
small_cap	0.897	0.519	0.617	0.596	0.887
credit_conditions	0.889	0.737	0.852	0.802	0.768
rates_recession	0.887	0.826	0.933	0.801	0.808
inflation_safe_haven	0.826	0.815	0.787	0.814	0.602
growth_tech	0.823	0.734	0.835	0.799	0.672
international_emerging	0.774	0.762	0.785	0.797	0.614
market_beta	0.708	0.466	0.636	0.512	0.708
energy_commodity	0.557	0.671	0.680	0.647	0.516

Highlighted: a baseline that outranks the model. The last column counts only pairs whose rows share a decile of the ticker-and-volatility score; 0.5 means the model adds nothing beyond the baseline.

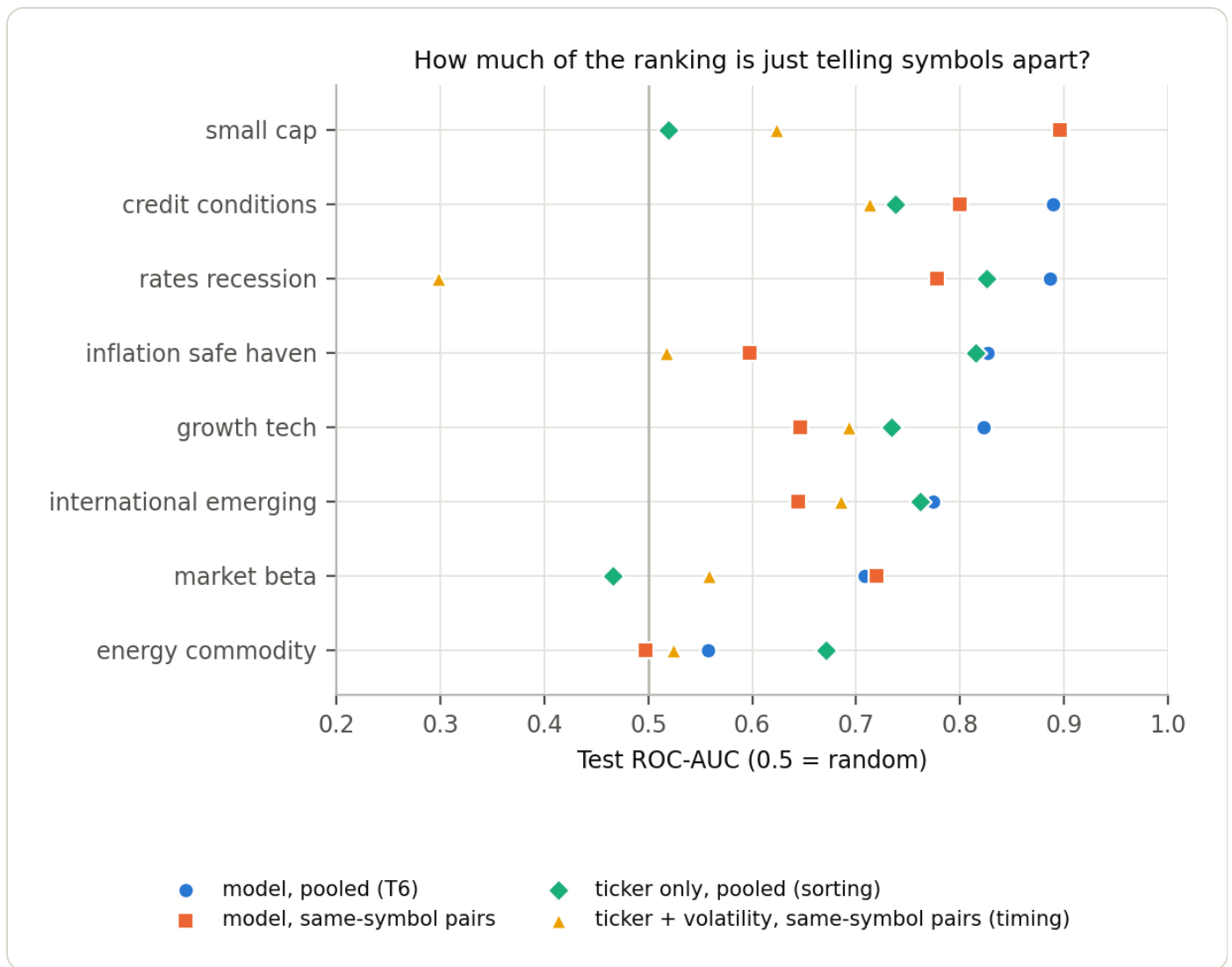


Figure 8. Where the ranking comes from: the model pooled (Table 6) and on same-symbol pairs only, a ticker-only baseline (pure sorting), and the ticker-and-volatility baseline on same-symbol pairs (timing from volatility alone). 0.5 is random.

Two cautions apply. These diagnostics were added after a review, with the test window already seen, so they are reported as descriptions, not as tests; only the first, unweighted version of the same-symbol AUC was specified before it was computed. Like Table 6, they carry no standard errors, because test rows overlap by up to 99% of their outcome window. The two categories that reached a test in Section 6.3 sit at opposite ends. *small_cap*'s ranking is almost entirely timing, while *international_emerging*'s same-symbol AUC is 0.644 and a ticker-and-volatility baseline outranks it.

7 Discussion

7.1 Why the limit is structural

A better model would not fix this. The number of independent outcomes is set by the data, and the two obvious ways to get more do not work. Adding tickers to a category adds rows but almost no evidence when the tickers are correlated: small_cap's 15 symbols are 1.04 effective series (Figure 1). Lengthening the hold, which is what a multi-month strategy requires, removes evidence: each doubling of the horizon halves the non-overlapping outcomes (Figure 4). At 3–6 months, each category's entire history holds roughly 53 to 99 independent outcomes, before any of it is set aside for validation and test. The test split covers about 4.3 years on disk, but only about 3 of those years can be traded. Each split file is scored on its own, so its first 204 rows go to feature warm-up and its last 126 have no complete outcome window. That leaves 752 scored rows per symbol, from 2023-03-07 to 2026-03-05, or just under 6 non-overlapping holds per effective series.

7.2 The rounding episode

Derived floor $1/776 = 0.0012886598 \rightarrow +1.34 \text{ SE.}$

Rounded floor $0.001289 \rightarrow +2.33 \text{ SE.}$

A difference of 0.00000034 , and the result crosses the bar.

The most fragile result in the project turned on the sixth decimal place. Validation derived small_cap's floor as exactly $1/776 = 0.0012886598$. The configuration file carried it as 0.001289 , a difference of 0.00000034 . Both print as 0.129% . Isotonic calibration maps many rows to the same probability, so rows sit on plateaus, and one plateau lies between the two values, 35 of the 119 validation entries and 2,252 test bars. The rounded value was therefore a different rule, one that validation never chose. It opened 77 test trades instead of 97 and scored $+4.61\%$ per trade at $+2.33$ standard errors, clear of the bar. The derived rule scored $+1.34$. Earlier write-ups of this project quoted the $+2.33$. A result that crosses the significance bar or not depending on how a number is copied is not evidence of an edge, but evidence that the sample is too small to tell. Since then, floors are read from the run's own output and never retyped. The result of the rounded rule is kept for the record in the tables of the first clean paper run (commit add7648), which is where the figures in this subsection come from.

7.3 Accuracy is not profit

High ranking quality and no trading edge are both true here, and they are not in conflict. ROC-AUC scores every row, including the many rows the strategy never trades, and it rewards ordering, not magnitude. The trading test asks something narrower: among the rows the strategy would actually enter, does the model's selection earn more than the unselected set? In 2022–2026 the unselected set already earned 5–8% per trade, so the model had to beat a high bar with few independent bets. A test ROC-AUC of 0.897 for small_cap sits beside a test edge of $+1.34$ standard errors.

7.4 The path taken

This paper's protocol is the end of a longer path, and several choices changed along the way. The project began with a 3–10 session hold, a label that counted any 15% move inside the window, a smaller universe, and a split made separately for each symbol. Decision thresholds were at one point chosen by requiring agreement between validation and test, which uses the test set for selection. The holding horizon then moved to 63–126 sessions, a 126–252 session horizon was tried and rejected, the peak label gave way to the terminal label, and the split became a shared calendar split. Each change had a reason, documented in Section 5, but each was also a fork in what Gelman and Loken call the garden of forking paths: an analysis shaped by the data it is then tested on.

Two consequences follow. First, the 2022–2026 test window is not a pristine holdout. Earlier versions of the project evaluated overlapping periods, and the rounded small_cap result was seen before the protocol was frozen. Second, that is why the rules were then fixed in code, the dataset hashed, and the final run made from a clean commit, with the test scored once. Freezing the rules limits further forking; it cannot undo earlier looks. Only data that did not exist when the rules were frozen can do that (Section 9).

8 Limitations

- One asset class (US-listed ETFs, some stocks and indices), one model family (tree ensembles on technical features), and one data vendor. The measured quantities in Section 5 are properties of this data; other universes will give other sizes.
- Growth_tech's 16 single stocks and energy_commodity's one are today's survivors, which biases their history upward. Three price indices (GSPC, DJI, RUT) extend history but cannot be traded directly. Their series are likely price-only, while the ETF series are dividend-adjusted, so their returns are not on the same footing as the rest of the universe. Each has a tradable near-twin in the universe (DIA for DJI, IWM and VTWO for RUT, and RSP, OEF, IWB and VONE as close substitutes for GSPC), so a robustness run without the indices is feasible.
- Every standard error in Tables 7 and 8 rests on 6 or 7 blocks of 182 days. A block bootstrap with so few blocks gives a rough standard error, and the standard error of that standard error is itself large.
- Test rows overlap by up to 99% of their outcome window, and no trustworthy resampling scheme for them was found. They describe one path through 2022–2026.
- The paper run regenerates its label rates and evidence counts (Tables 4 and 5), but not model or trading results at that horizon. Those come from earlier project runs on different data and splits, and are not reported as results here.
- Earlier iterations of the project looked at overlapping periods. (Section 7.4)

9 Conclusion and a pre-registered forward test

Technical features rank multi-month ETF outcomes well out of sample, but under a protocol fixed in advance, no category's trading rule beats ignoring the model by 2 standard errors. The binding limit is evidence: 53 to 99 independent outcomes per category across the whole history, too few to separate a modest edge from luck.

More history cannot be manufactured, but new history accumulates. A pre-registered forward test is therefore proposed:

1. Freeze the two surviving rules now: `international_emerging` at a floor of 0.018749 and `small_cap` at exactly 1/776, with the models, features, code commit and data pipeline as of this paper.
2. Register the test before any new data arrives: the null (the same models with the floor at zero), the statistic (edge per trade, block-bootstrap standard error with 182-day blocks), and the bar (2 standard errors, one-sided).
3. Paper-trade every signal from the freeze date, recording entries and exits as they happen.
4. Score once at a date fixed in advance.

The timeline is set by the block count. The protocol needs at least 6 blocks of 182 days to report a standard error, so the earliest possible scoring date is about 3 years after the freeze. Even then, the test may not be decisive. The test split's own scored window held about 3 years of entries (Section 7.1), so a 3-year forward test would hold about as much evidence as the test did. If the true edges equal those seen on test, it would be expected to show about +1.7 standard errors for `international_emerging` and +1.3 for `small_cap`. Since standard errors shrink with the square root of time, reaching 2 would take roughly 4 and 7 years of entries respectively. A 3-year test can still falsify the rules, by showing a negative edge; confirming them needs longer, or a larger true edge than observed.

References

- Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. *Journal of Computational Finance*, 20(4), 39–69. doi:10.21314/JCF.2016.322 · author copy
- Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5), 94–107. author copy · SSRN
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465. link
- Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68. publisher · NBER w20592
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons. Design effect and intraclass correlation, Section 5.4, pp. 161–162; definition of Deff, Section 8.2, p. 258.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3), 1217–1241. doi:10.1214/aos/1176347265

López de Prado, M. (2018). *Advances in Financial Machine Learning*. Hoboken, NJ: John Wiley & Sons. Ch. 7, Cross-validation in finance. publisher

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). author copy

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313. doi:10.1080/01621459.1994.10476870

Appendix A. Reproducibility

Every table and figure is produced by one script from one frozen dataset; nothing is edited by hand.

Code commit	c85ea0b322418807a235bb8bd56232f49c96fe0a (clean working tree)
Command	python scripts/paper_run.py --require-clean
Run	2026-09-30, 12:54–13:12 UTC, about 18 minutes
Data	paper/data_manifest.json: SHA-256 of 484 files, snapshot of 2026-09-04; manifest hash 7c37520f...a6fc87
Environment	Python 3.11.3; numpy 1.26.4, pandas 2.0.3, scikit-learn 1.4.2, xgboost 3.2.0, ta 0.11.0
Seeds	random forest 42, XGBoost 42, bootstrap and permutation 20260830
Resampling	2,000 bootstrap replicates, 2,000 permutations, 182-day blocks, minimum 6 blocks
Reproduction check	refit probabilities match the saved models to within 8.9×10^{-14}

The script checks every data file against the manifest and stops on any mismatch. `--require-clean` refuses to run with uncommitted changes, so the recorded commit is the code that produced the numbers. Re-downloading prices does not reproduce the dataset, because adjusted prices are revised after every dividend; the archived snapshot is required.

Appendix B. Symbols per category

Category	n	Symbols
credit_conditions	22	ANGL, BKLN, CWB, EMB, FALN, HYG, HYLB, IGIB, IGSB, JNK, LQD, PCY, PFF, PGX, SHYG, SJNK, SPIB, SPLB, USHY, VCIT, VCLT, VCSH
energy_commodity	20	AMLP, BNO, COPX, CVX*, DBA, DBC, DBO, FCG, GDX, GDXJ, IEO, MLPX, OIH, PICK, SIL, UNG, USO, XES, XLE, XOP
growth_tech	27	AAPL*, ADBE*, AMAT*, AMD*, AMZN*, ARKK, AVGO*, CRM*, CSCO*, GOOGL*, HACK, IBM*, IGV, INTC*, META*, MSFT*, MU*, NVDA*, ORCL*, QCOM*, QQQ, SKYY, SMH, SOXX, TXN*, VGT, XLK
inflation_safe_haven	13	FXE, FXY, GLD, IAU, IVOL, PALL, PPLT, SCHP, SGOL, SLV, SPIP, TIP, UUP
international_emerging	27	ACWX, EEM, EFA, EIDO, EWA, EWC, EWD, EWG, EWH, EWI, EWJ, EWL, EWP, EWQ, EWS, EWT, EWU, EWW, EWY, EWZ, EZA, FXI, IEMG, INDIA, TUR, VEU, VWO
market_beta	24	DIA, DJI†, GSPC†, IWB, IWV, MTUM, OEF, QUAL, RSP, SCHX, SPLV, SPYD, USMV, VLUE, VONE, XLB, XLC, XLF, XLI, XLP, XLRE, XLU, XLV, XLY
rates_recession	14	BIL, EDV, GOVT, IEF, IEI, SHV, SHY, STIP, TBT, TLH, TLT, TMV, VTIP, ZROZ
small_cap	15	FNDA, IJR, IJS, IJT, IWM, IWN, IWO, RUT†, SCHX, SLYG, SLYV, VB, VBK, VBR, VTWO

* single stock · † price index, not directly tradable. All others are ETFs. Symbols as listed in the data manifest.