

UNDER WORK IN PROGRESS!! paper is not finished, check back never!!

Short-Horizon Price Direction Is Not Predictable: A Negative Result on 258,737 Real Minute Bars

Cooper E.

Independent researcher · jcooperkai@gmail.com

ABSTRACT

A widespread assumption among retail participants is that short-horizon price direction carries exploitable signal. We test that assumption directly on 258,737 real one-minute bars spanning roughly 180 days, with expanding-window walk-forward validation, purging to prevent overlap, and metrics implemented from scratch rather than inherited from a library. Direction over a fifteen-minute horizon is not predictable: walk-forward AUC is 0.5105 against a 0.5 baseline, and the Brier score improves on a coin flip by 0.00004 across 11,497 out-of-sample windows. The model is well calibrated -- expected calibration error 0.0096 -- but it is calibrated at a coin flip, and calibration without discrimination is not signal. We then show where the remaining structure actually lives: the driftless binary fair value of the instrument is highly informative intra-window, reaching AUC 0.977 one minute from settlement, but this is the option's delta rather than alpha, since any efficient quote prices it identically. The only tradeable residual is quote lag and volatility mispricing, two quantities that must be measured live and cannot be recovered from historical bars. Finally, we describe two defects in our own evaluation harness that produced large fictitious profits and were caught only by insisting that a zero-edge control return exactly zero.

1. Introduction

Claims of short-horizon predictability are common and rarely accompanied by a falsifiable test. The usual failure is not fraud but methodology: overlapping windows, leakage across the train/test boundary, metrics computed with library defaults that quietly differ from what the author assumes, and the absence of any configuration in which the reported profit would be impossible.

This note tests the assumption under conditions designed to make a positive result difficult to fake. Every design choice below was fixed before the result was known.

2. Data and protocol

The dataset is 258,737 one-minute bars of real market data covering roughly 180 days. Prediction windows are fifteen minutes long and non-overlapping. The label is whether the close exceeded the open of the window -- a plain binary outcome with no threshold to tune.

Thirteen features are computed strictly from history preceding the window: multi-scale momentum over 1, 5, 15, 30 and 60 minutes; realised volatility at two scales; sign persistence; a volume trend ratio; time-of-day as sine and cosine; and a sixty-minute drift term.

Validation is expanding-window walk-forward across six folds. Each fold trains only on windows strictly preceding its test block, and the boundary is purged by one window so that no training sample can overlap a tested window. The model is logistic regression implemented from scratch, as are Brier score, AUC and expected calibration error, so that no library convention is silently assumed.

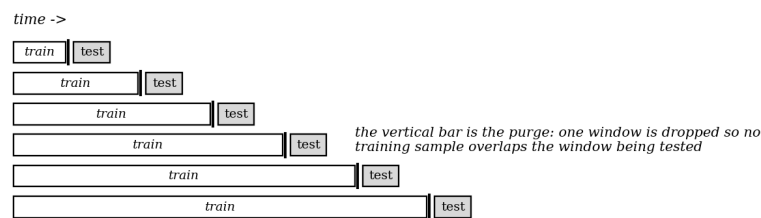


Figure 1. Walk-forward protocol. Training expands, testing moves forward, and the vertical bar marks the purge: one window is discarded at each boundary so no training sample overlaps the window being tested. Without this, a fifteen-minute label computed from overlapping data leaks future information into training.

3. Result: no directional skill

Metric	Model	Baseline	Difference
Brier score	0.24996	0.25000	+0.00004
AUC	0.5105	0.5	+0.0105
Expected calibration error	0.0096	--	--

Table 1. Out-of-sample performance across 11,497 windows. The Brier improvement over a coin flip is four parts in one hundred thousand.

The calibration figure deserves emphasis because it is the most likely to be misread. A model with ECE 0.0096 is well calibrated: when it says 52 percent, the event occurs about 52 percent of the time. That sounds like competence. It is not. The model is calibrated *at a coin flip* -- its probabilities cluster tightly around 0.5 and it never expresses a confident view that turns out to be right more often than chance. Calibration measures honesty about uncertainty; discrimination measures whether there is any information. Only the second is tradeable.

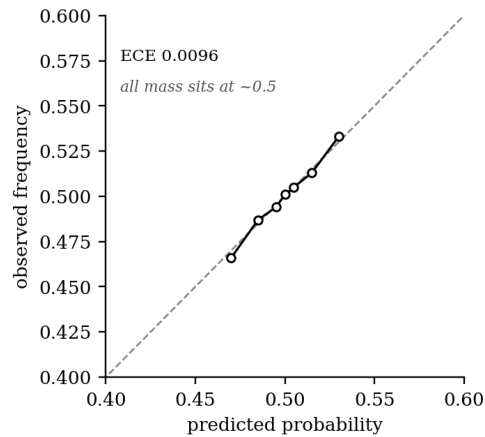


Figure 2. Reliability diagram. Points lie on the diagonal -- the model is calibrated -- but the entire predicted range spans roughly 0.47 to 0.53. There is nothing to trade.

4. Where the structure actually is

A negative result is more useful when it also says where to stop looking. The instrument is a binary option struck at the window-open price K . At minute t within the window, its driftless fair value is

$$f(t) = \Phi(\ln(S_t / K) / (\sigma \cdot \sqrt{\tau}))$$

with σ a trailing one-minute realised volatility estimate and τ the remaining time. Measured against realised settlement, this quantity is calibrated (ECE approximately 0.02) and becomes decisive as τ falls.

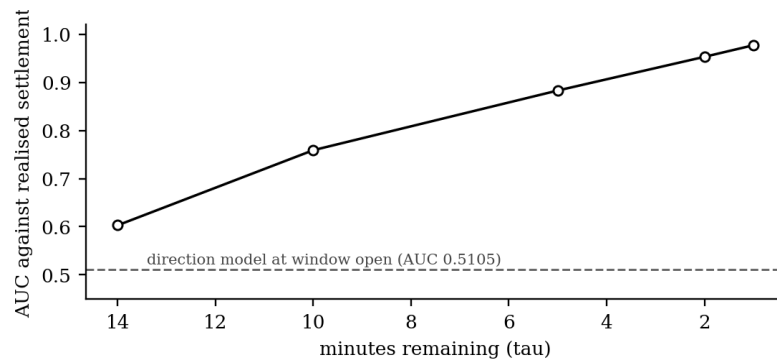


Figure 3. Fair-value discrimination against realised settlement as time runs out, compared with the directional model at window open. The contrast is the point of the paper: predicting direction at the open is impossible, while the same instrument becomes almost fully determined minutes later. That information is a consequence of observed price, not a forecast of it.

This is not alpha. It is the option's delta. Any quote observing the same spot prices it identically, so the fair value is visible to the market as soon as it is visible to us. The only tradeable residual is the extent to which a quoted mid *lags* this fair value, or misprices volatility.

5. Bounding the residual, with a control that must fail

Model a quoted mid as the fair value lagged by L and shrunk toward 0.5 by a favourite discount d , then trade only when the disagreement exceeds the half-spread. This produces an edge surface in two live-measurable parameters rather than free ones.

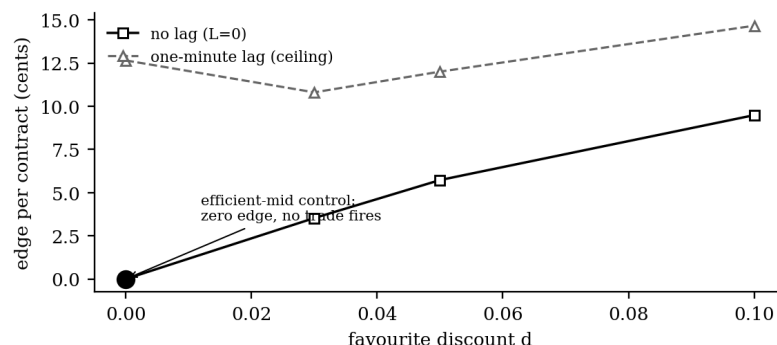


Figure 4. Modelled edge per contract. The filled point at the origin is the control: with no lag and no discount -- an efficient quote -- the edge is exactly zero and no trade fires. This is the most important point on the chart, because it demonstrates the measurement cannot manufacture profit from a fair quote. Any harness lacking such a point should not be believed.

Both L and d are properties of a live venue, not of history. Establishing them requires recording quotes against spot over real elapsed time, which historical bars cannot supply. We therefore make no claim about whether the residual is exploitable, only that it is the sole place remaining to look and that it is measurable in principle.

6. Two harness defects, and the control that caught them

Both were discovered by requiring the zero-edge configuration to return exactly zero, and neither announced itself as an error.

Clamp asymmetry. Fair value was left unclamped while the modelled mid was clamped to $[0.02, 0.98]$. On near-certain contracts this manufactured a spurious one-to-two cent disagreement out of nothing. The efficient-mid control, which must return zero, instead reported turning fifty units into 535,827. Clamping both quantities identically removed the artefact.

Unbounded compounding. Reinvesting a growing bankroll across tens of thousands of sequential windows produced returns of order 10^{250} . The arithmetic was correct; the scenario was not, since venue liquidity caps position size long before that. Adding an absolute per-trade cap restored economically meaningful figures.

The general lesson is procedural rather than statistical. **Every backtest should contain a configuration in which profit is impossible, and that configuration should be checked before any other number is read.** Both defects were invisible to inspection, produced plausible-looking equity curves, and would have survived ordinary review.

7. What this rules out

Within the tested horizon and feature class, direction is unpredictable at the level of measurement available here. This does not prove that no directional signal exists at any horizon under any feature set. It does mean that the specific strategy family most commonly attempted by retail participants -- predict the next fifteen minutes from recent price and volume history -- has no measurable edge on 180 days of real data, and that a practitioner should require evidence at this standard before deploying capital against it.

We report the null because nulls of this kind are rarely published, and their absence is part of why the assumption persists.

References

1. G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1-3, 1950.
2. F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637-654, 1973.
3. M. Lopez de Prado. *Advances in Financial Machine Learning*. Wiley, 2018.
4. A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. *ICML*, 2005.