Price Path Information and Football Match Prediction in Polymarket

Junze Yuan

April 25, 2026

Abstract

This report asks whether first-half *price path* information—aggregate signed change and interval-based signed or absolute changes—improves probabilistic forecasts of football match outcomes beyond the pre-match Polymarket price P_0. A central secondary question is whether path features add value beyond simply observing the halftime level P_{HT} as a summary of first-half updating.

Main finding. Path features do improve on P_0 on all five cross-validation folds, but P_{HT} alone achieves the lowest mean CV log loss (0.534) of any specification tested—beating the best explicit path model ($P_0 + \Delta_{\mathrm{HT}}$, CV log loss 0.540) by 0.006 on average. In other words, *where* the price ends up at halftime is more informative than *how* it got there. Interval-based path families do not outperform this benchmark on average; richer engineering is therefore kept in the appendix. Results are predictive (not causal); the sample is modest ($n = 927$ conditions, 309 events) and fold-level gains are heterogeneous (gain range: 0.02–0.08 log loss units across folds).

1 Introduction

1.1 Background and motivation

Prediction-market prices behave as probability-like signals that update continuously as information arrives. Football matches provide a natural laboratory: the first half generates a sequence of market revisions that may contain information not captured by the opening price P_0 alone. Halftime is a natural cutoff, but the substantive question is whether the *route* to halftime matters—not only the endpoint.

1.2 Research question and hypothesis

This report studies whether first-half market price-path and price-volatility features, such as aggregate signed change and interval-based signed or absolute changes, improve prediction of the final match outcome relative to relying only on the pre-match market price. A secondary comparison asks whether those path-based features add value beyond simply observing the halftime level.

1.3 Defining information gain

In this report, information gain refers to an improvement in out-of-sample probabilistic prediction after adding first-half market information to the pre-match baseline P_0. Operationally, first-half information gain is measured as a reduction in `log_loss` or Brier score relative to a baseline model that uses only P_0. This is a predictive definition, not a causal claim and not an information-theoretic entropy measure.

- Baseline model: uses only P_0.
- Updated model: adds first-half information features such as Δ_{HT}, interval signed changes, and interval absolute-change proxies for volatility. These features capture the price path and price volatility of the first half.
- Positive information gain means lower out-of-sample `log_loss` or Brier score than the P_0 baseline.
- P_{HT} can be used as an endpoint benchmark, but it is not the definition of information gain.

1.4 Why probabilistic accuracy matters

Classification accuracy alone can miss meaningful changes in forecast quality. Metrics such as `log_loss` and Brier score better reflect calibration and the quality of predicted probabilities.

2 Data

2.1 Data source and structure

Data come from Polymarket football markets. Each row is one **condition** (`condition_id`) nested under an event (`event_id`) identified by `kickoff_utc`. The binary target y is the realised outcome for that contract side. Because each match generates multiple conditions (e.g. home-win and away-win sides), conditions within the same event are dependent; all splits therefore assign *entire events* to a single fold.

Table 1 shows three conditions from the same event. This small example illustrates both the row-level unit of analysis and why event-level grouping is required: the three rows represent different contract sides from one match and therefore cannot be split across train and validation folds.

Table 1: Example price-snapshot rows before feature expansion.

Event	Kickoff (UTC)	P_0	P_{15}	P_{30}	P_{HT}	y
42403	2025-09-20 16:30	0.355	0.885	0.860	0.965	1
42403	2025-09-20 16:30	0.385	0.025	0.040	0.012	0
42403	2025-09-20 16:30	0.255	0.140	0.115	0.035	0

Table 2 shows the corresponding research-ready features used by the logistic models. Signed changes preserve direction, while absolute changes isolate volatility regardless of whether the market moved toward or away from a contract side.

Table 2: Example research-ready path features and binary target.

Event	P_0	Δ_{HT}	$\Delta_{0\text{–}15}$	$\Delta_{15\text{–}30}$	$\Delta_{30\text{–}\mathrm{HT}}$	$\lvert\Delta_{\mathrm{HT}}\rvert$	y
42403	0.355	0.610	0.530	-0.025	0.105	0.610	1
42403	0.385	-0.374	-0.360	0.015	-0.029	0.374	0
42403	0.255	-0.220	-0.115	-0.025	-0.080	0.220	0

2.2 Price snapshots and path features

Four price snapshots are collected per condition: P_0 (kick-off), P_{15} (15 min), P_{30} (30 min), and P_{HT} (halftime). Derived path features are then computed as signed or absolute changes between consecutive snapshots (see Section 3.1).

2.3 Sample summary

Table 3 reports the key sample statistics. With 927 conditions across 309 events, the dataset spans roughly seven months of football markets. The positive-outcome rate of $\approx 33\%$ reflects the three-way (home/draw/away) contract structure.

Table 3: Sample summary (all splits combined).

Quantity	Value
Total rows (conditions)	927
Unique events (matches)	309
Train / validation / test rows	510 / 210 / 207
Development sample for CV (train + val)	720
Held-out test set	207
Mean outcome y (positive rate)	≈ 0.333
Kickoff date range (UTC)	2025-08-15 to 2026-03-22

3 Methodology

3.1 Feature construction

Figure 1 illustrates the timeline of price observations and the features derived from them.

The five main specifications are:

- **A – Baseline:** P_0 only.

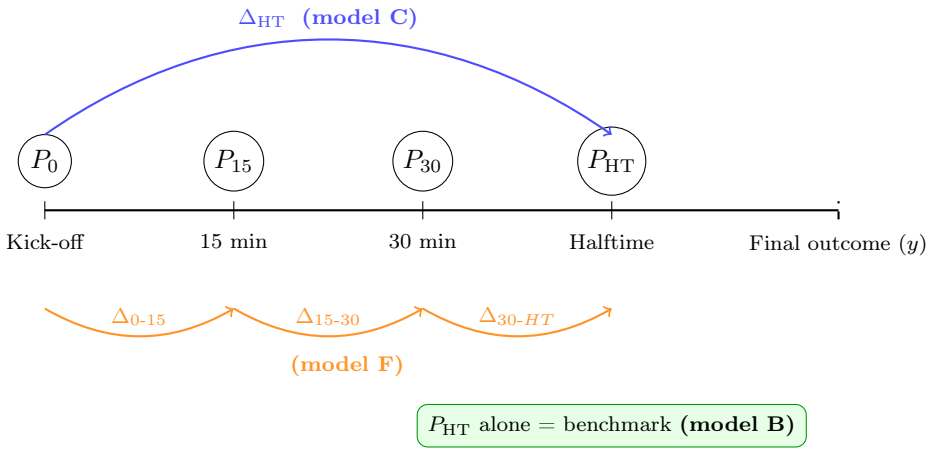


Figure 1: Price snapshot timeline and feature groupings. Model C uses only the aggregate change Δ_{HT}; model F decomposes it into three interval changes; the benchmark B uses neither—only the endpoint level P_{HT}.

- **B – Halftime benchmark:** P_{HT} only. Uses no path information; answers how much first-half information a single endpoint price captures.
- **C – Aggregate path:** P_0 and Δ_{HT}. Adds the total signed first-half change to the baseline.
- **F – Interval signed path:** P_0 with $\Delta_{0–15}$, $\Delta_{15–30}$, $\Delta_{30–HT}$. Tests whether *timing* of price movements within the half matters.
- **G – Interval absolute path:** P_0 with absolute interval changes $|\Delta_{0–15}|$, $|\Delta_{15–30}|$, $|\Delta_{30–HT}|$. Tests whether *volatility* (regardless of direction) is informative.

Additional exploratory specifications (D, E, H, I) are listed in Appendix A.1.

3.2 Model

All specifications use logistic regression (`sklearn`, `max_iter`=5000) on raw, bounded features without standardisation, matching the evaluation protocol.

3.3 Evaluation

Cross-validation. Grouped chronological 5-fold CV orders entire events by kick-off date and trains on past blocks to score later matches, preserving temporal realism and preventing data leakage within events.

Test confirmation. The final model is refit on all development data ($n = 720$) and evaluated once on the held-out test set ($n = 207$). Test metrics supplement but do not override CV conclusions, given the single-split uncertainty.

Primary metrics. Mean and standard deviation of log loss and Brier score across folds; count of folds beating baseline A; held-out test metrics. The estimand is incremental predictive value over P_0, not causal feature importance.

4 Results

4.1 Main model comparison

Table 4 reports cross-validated and held-out test performance for the five main specifications. Two findings are immediately visible:

1. **All first-half models beat P_0.** Models B, C, F, and G each win all five CV folds on log loss. The baseline P_0 CV log loss of 0.598 is the worst score in the table.

2. **The halftime level benchmark B is best overall.** P_{HT} alone achieves the lowest mean CV log loss (0.534) and lowest mean CV Brier (0.179). Crucially, it outperforms the best explicit path model (C, CV LL = 0.540) by 0.006—despite using *no* information about how the price arrived at halftime.

The "vs. P_0" column (absolute CV log loss improvement) makes these magnitudes direct: the benchmark gains 0.064 over P_0; model C gains 0.058; model F gains 0.044; model G gains barely 0.003. Interval decomposition (F) is consistently worse than the aggregate change (C), and G (absolute intervals) adds almost nothing over the raw P_0 baseline.

Table 4: Main comparison: baseline, path families, and halftime benchmark. CV = chronological 5-fold on $n = 720$; test = single holdout on $n = 207$. The "vs. P_0" column shows absolute CV log loss reduction relative to model A; higher is better. The benchmark row (B) is highlighted.

Specification	CV log loss		CV Brier		CV folds vs. P_0	vs. P_0 (CV LL ↓)	Test LL	Test Brier
	mean	std	mean	std				
A: P_0 only	0.598	0.029	0.205	0.013	0/5	—	0.620	0.215
C: $P_0 + \Delta_{\text{HT}}$	0.540	0.045	0.180	0.020	5/5	−0.058	0.586	0.200
F: P_0 + interval Δ's	0.554	0.038	0.186	0.017	5/5	−0.044	0.586	0.200
G: P_0 + interval $\lvert\Delta\rvert$'s	0.595	0.029	0.204	0.013	5/5	−0.003	0.617	0.213
B: P_{HT} only *(benchmark)*	**0.534**	0.056	**0.179**	0.024	5/5	**−0.064**	0.594	0.202

Note on the test set. On the held-out test set, models C and F achieve log loss 0.586, marginally *below* the benchmark B (0.594)—the reverse of the CV ranking. This single-split reversal does not overturn the CV-based conclusion; it does indicate that the benchmark's average advantage is modest in magnitude (≈0.006–0.008 in CV log loss) and should not be overstated. The conservative interpretation is that P_{HT} is at least as good as explicit path decompositions, and likely better on average.

Figure 2 visualises the mean CV errors with fold-level standard deviations. The benchmark (green) achieves the lowest bar in both panels, while the aggregate path (blue) and interval-signed path (purple) trail it closely. Model G is omitted from this summary plot because its mean CV error is nearly identical to model A and is instead shown directly in Figure 3.

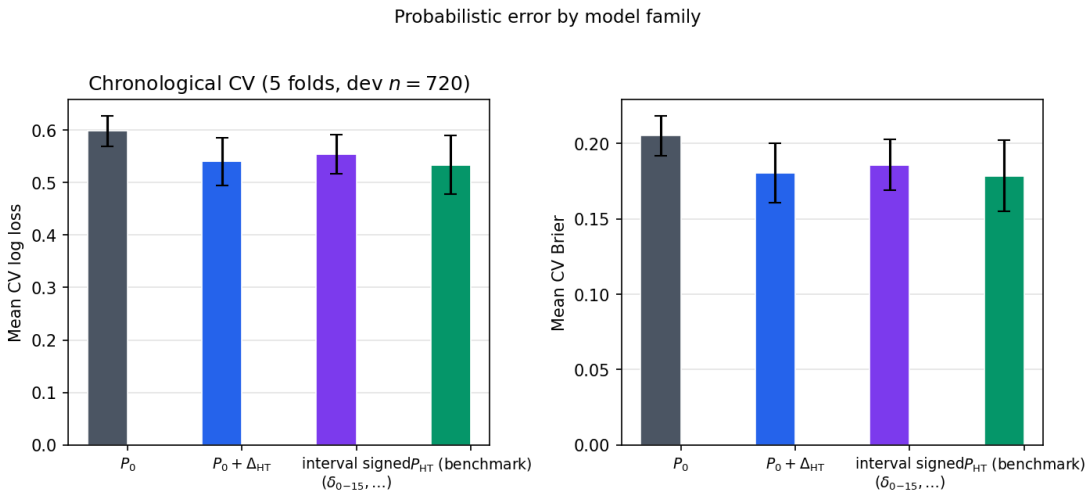


Figure 2: Mean CV log loss (left) and Brier score (right), with ±1 std across folds. The halftime benchmark P_{HT} achieves the lowest mean error in both panels despite using no path information, indicating that the halftime price level alone captures most of the usable first-half signal.

4.2 Interpretation of the main result

Figure 3 plots per-fold log loss gains (relative to P_0) for models B, C, F, and G. Gains are **positive in every fold** for all four models, confirming that first-half information consistently helps. The benchmark B leads in three of five folds (1, 2, 3); model C edges ahead in fold 5, while B and C are effectively tied in fold 4. **Fold 3 shows the largest gains** (≈0.08 log loss units); fold 5 shows the smallest gain for the benchmark (<0.02). Model G, the volatility-only interval specification, is positive but very small in every fold (≈0.001–0.006 log loss units). This shows that absolute first-half movement by itself does not add meaningful signal beyond the pre-match baseline.

Figure 4 shows reliability curves for P_0, P_{HT}, and $P_0+\Delta_{\text{HT}}$ on the validation set. P_0 shows visible miscalibration in the 0.2–0.4 predicted probability range; P_{HT} and $P_0 + \Delta_{\text{HT}}$ track the diagonal more closely across mid-range probabilities. All three diverge from the diagonal in the tails (below 0.15 and above 0.65), reflecting sparse calibration data in extreme-probability bins. Lower mean log loss therefore reflects improved mid-range calibration rather than uniformly better estimates across the whole $[0, 1]$ range.

4.3 Robustness and secondary comparisons

Table 5 summarises four checks that stress-test the main conclusions without reopening broad feature search.

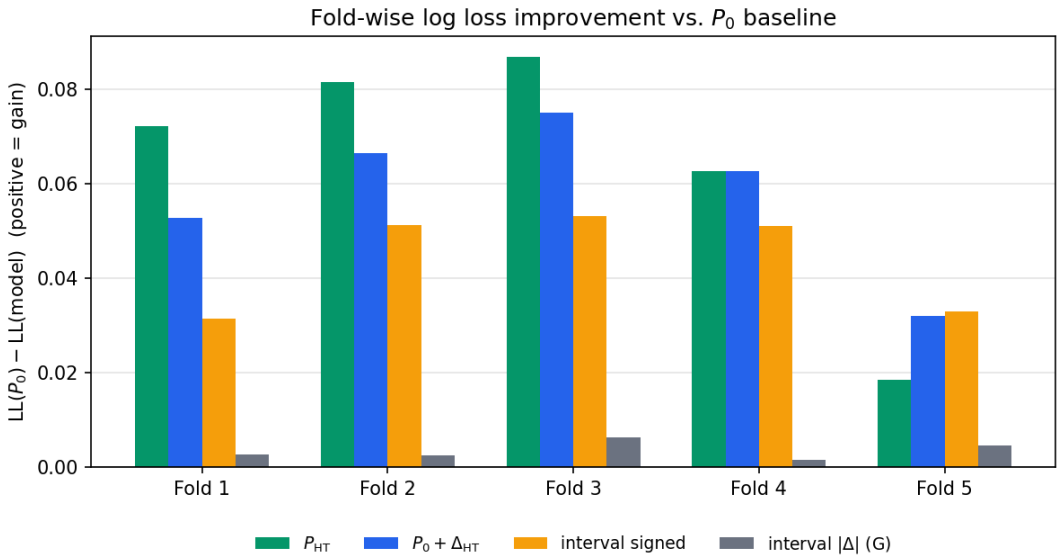


Figure 3: Per-fold log loss gain over P_0 baseline: $\mathrm{LL}(P_0) - \mathrm{LL}(\text{model})$; positive values indicate lower error than P_0. Benchmark, aggregate signed-change, and interval signed-change gains vary substantially across folds. Model G adds interval absolute changes to capture volatility, but its bars remain close to zero, indicating little incremental predictive value from volatility alone.

Table 5: Robustness checks (development sample). All checks support the benchmark conclusion; none reverses the main finding.

Check	Comparison	Finding	Verdict	Interpretation
P_{30} vs. P_{HT}	Single-feature logits	CV LL lower for P_{HT}	Supports	The 30-min price is weaker than the halftime price; using an earlier snapshot is not equivalent.
Winsorized Δ_{HT}	P_0+ winsorized Δ_{HT} vs. P_{HT}	CV LL lower for P_{HT}	Supports	Benchmark advantage is not driven solely by extreme halftime moves.
Interval vs. aggregate	F vs. C	Mean CV LL lower for C	Supports	Breaking the aggregate change into three subintervals hurts on average.
Stratified CV	P_{HT} vs. P_0, by stratum	P_{HT} wins in all strata	Supports	Holds when stratifying on $P_0(1-P_0)$ and $\lvert\Delta_{\mathrm{HT}}\rvert$; no volume stratification available.

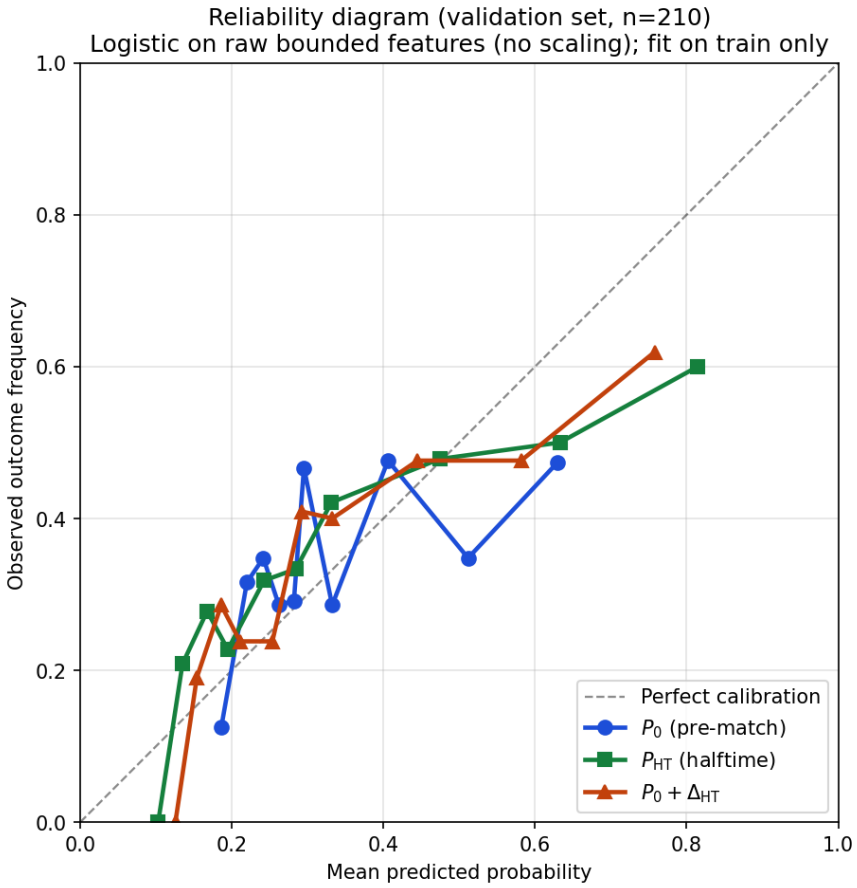


Figure 4: Reliability diagram (validation set, $n = 210$). Models fit on train only. P_0 shows systematic miscalibration in the 0.2–0.4 range; P_{HT} and $P_0 + \Delta_{\text{HT}}$ are closer to the diagonal across mid-range probabilities, but all three diverge in the tails where calibration data are sparse.

5 Discussion

First-half price movements contain genuine predictive signal beyond the pre-match price P_0: every path model beats the baseline on all five CV folds, and the aggregate signed change Δ_{HT} accounts for most of that gain. This confirms the primary research question—first-half path information does help.

The more nuanced result concerns the *level–path* contrast. The halftime price P_{HT} is the endpoint that path features collectively approximate, and it outperforms explicit path parameterisations on mean CV log loss. This is consistent with market prices functioning as near-sufficient statistics for first-half information: by halftime, goal events, momentum shifts, and tactical developments are already encoded in the price level. Decomposing the path into sub-interval changes (model F) or using absolute changes (model G) does not recover additional signal; it appears to add noise instead.

The central question was whether the *route* to halftime matters, not just the endpoint. Based on this sample, the answer is: mostly no. The halftime level captures most of the usable first-half information for parsimonious logistic regression. However, the margin is small—0.006 in CV log loss—and on the held-out test set the direction reverses. The honest conclusion is that path and level features are roughly comparable in this setting; neither dominates decisively.

Path features could still add value in richer settings. With liquidity or trading-volume data, path information could be weighted by market depth: a large move on thin volume carries different information than the same move with heavy activity. Higher-frequency snapshots (e.g. every 5 minutes) might reveal non-linear within-half dynamics that four snapshots cannot capture. Both avenues require data not available in the current extract.

The following constraints bound the conclusions. The sample covers only 309 events; fold-level gains range from 0.02 to 0.08 log loss units, indicating that the average effect may not represent any particular game subset reliably. Multiple conditions per event share outcomes; splits are assigned at the event level to prevent leakage, but within-event dependence reduces effective sample size. No liquidity or order-flow data are available, so path features cannot be weighted by trading intensity. Lower average log loss reflects improved mid-range calibration; tail behaviour remains imprecise at this sample size. All results are predictive associations only; no causal claim is made.

6 Conclusion

First-half Polymarket price paths improve probabilistic match forecasts relative to the pre-match price P_0—all path models win all five CV folds over the baseline. However, a single number—the halftime price level P_{HT}—matches or exceeds the predictive accuracy of every explicit path construction tested on mean cross-validated log loss (0.534 vs. 0.540 for the best path model).

The practical implication is that, for parsimonious forecasting under the current data and model class, recording P_{HT} alone at halftime suffices. Richer path engineering belongs in appendix-level exploration unless future data (higher-frequency quotes, liquidity signals, or a larger sample) stabilises its advantage. All claims are predictive associations; the sample is modest; and fold-level

gains are heterogeneous.

7 References

Github repository is available at `https://github.com/rmz59/Polymarket`.

A Appendix

A.1 Exploratory model families and ROC curves

Table 6 lists all nine specifications (A–I) sorted by mean CV log loss. Several multi-feature rows achieve competitive *test* log loss (H at 0.553, I at 0.538), which is why the main text emphasises CV averages and parsimony. Added complexity did not deliver a clearly better answer on the primary CV metric.

Table 6: All models sorted by mean CV log loss. Benchmark B highlighted; path models C and I are the closest competitors.

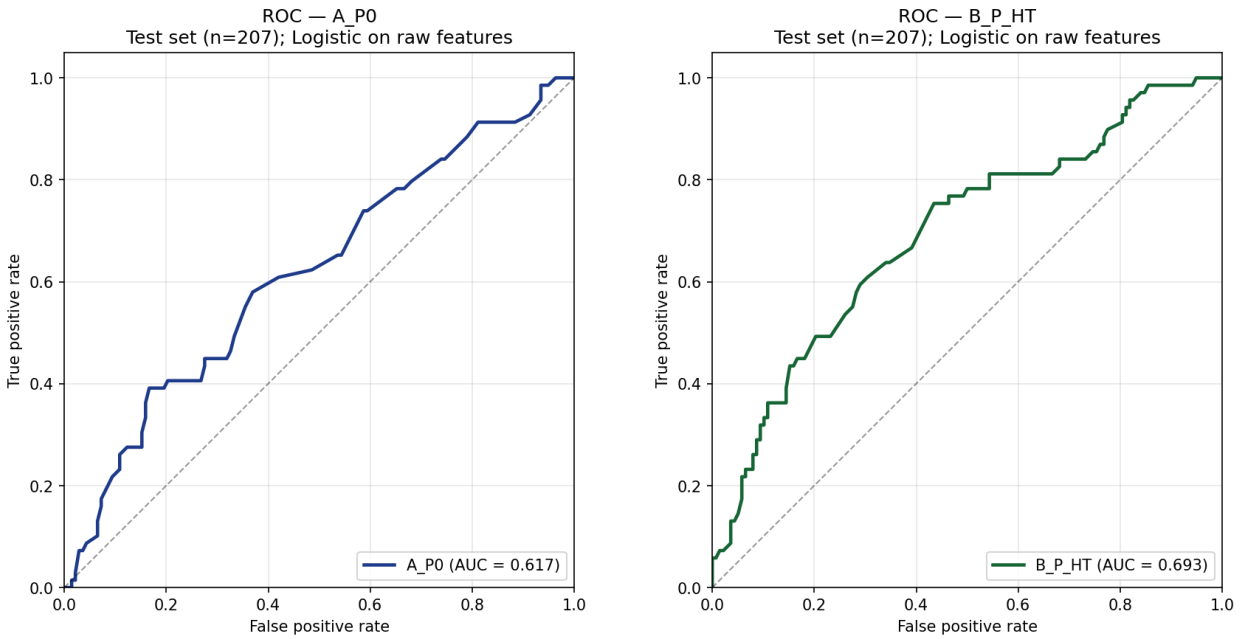
ID	Features	Mean CV log loss	Mean CV Brier
B	P_{HT} only *(benchmark)*	**0.534**	**0.179**
C	P_0, Δ_{HT}	0.540	0.180
I	Prior benchmark	0.538	0.181
E	P_0, Δ_{HT}, $\lvert\Delta_{\text{HT}}\rvert$	0.541	0.181
H	P_0, interval Δ's + $\lvert\Delta\rvert$'s	0.553	0.186
F	P_0, interval Δ's	0.554	0.186
D	P_0, $\lvert\Delta_{\text{HT}}\rvert$	0.594	0.203
G	P_0, interval $\lvert\Delta\rvert$'s	0.595	0.204
A	P_0 only *(baseline)*	0.598	0.205

ROC curves on the test set ($n = 207$) complement the log loss results. AUC rankings are broadly consistent with CV log loss: B (0.693), C (0.696, marginally above B on this split), A (0.617). AUC measures discrimination only and does not capture calibration; it is secondary to probability-based scores throughout.

A.2 Distribution of first-half path features

This appendix provides context on the magnitude and heterogeneity of the first-half price changes used as features. On the training split, $|\Delta_{\text{HT}}|$ has median ≈ 0.10, 90th percentile ≈ 0.28, and maximum ≈ 0.61, indicating that price paths vary substantially across matches. The robustness checks in Table 5—winsorization and stratified CV—confirm that the main results are not driven solely by a handful of extreme-move events.

Future versions could include compact histograms of Δ_{HT}, $\Delta_{30\text{–}HT}$, and $|\Delta_{30\text{–}HT}|$ generated directly from `new_sum_table.csv` without rerunning any model.



(a) Baseline P_0 (AUC = 0.617).

(b) Halftime benchmark P_{HT} (AUC = 0.693).

Figure 5: ROC curves for baseline and benchmark on the test set. The benchmark's larger AUC is consistent with its lower CV log loss.

A.3 Large first-half movement examples

Markets with $|\Delta_{\mathrm{HT}}| > 0.28$ (above the 90th percentile) represent the most volatile first halves in the sample. The winsorization check in Table 5 finds that the benchmark conclusion holds after capping extreme moves, confirming these events do not drive the main result.

Detailed match-level case tables (sourced from `halftime_large_move_observations.csv`) can be appended if raw event identifiers are cleared for publication.

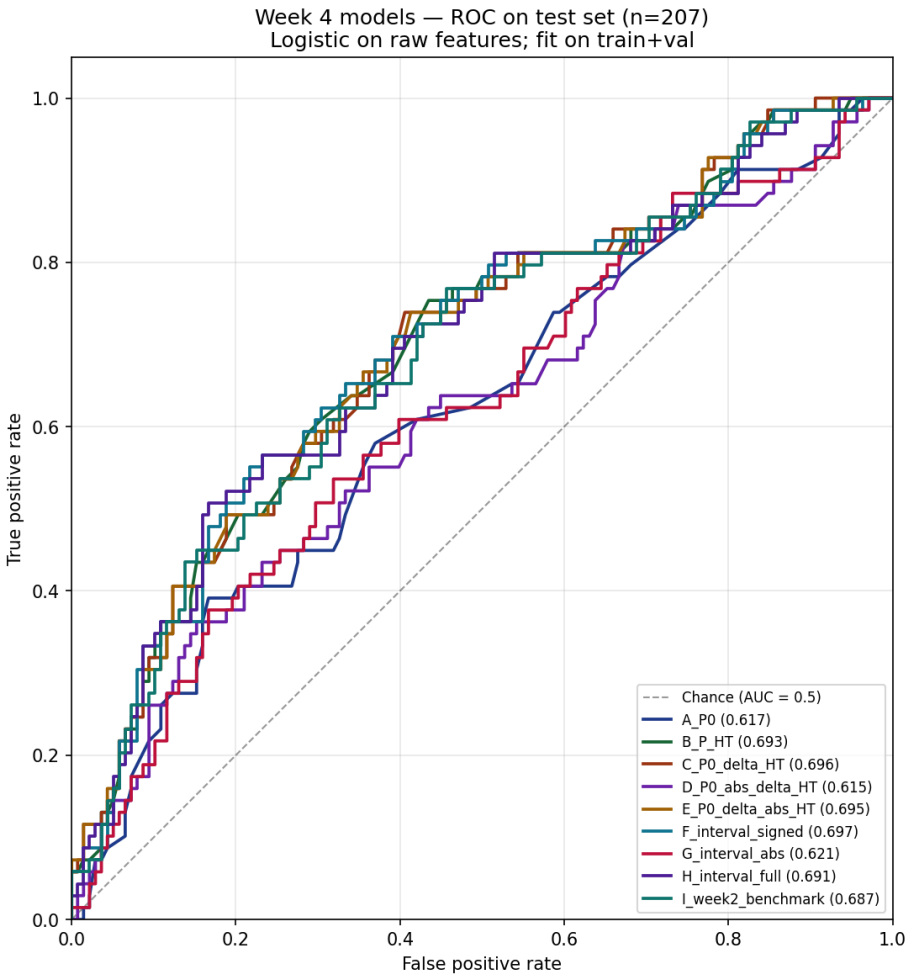


Figure 6: ROC overlay for all model IDs on the test set. Models C, E, F, and H cluster near AUC ≈ 0.695–0.697; G and D trail near the P_0 baseline.

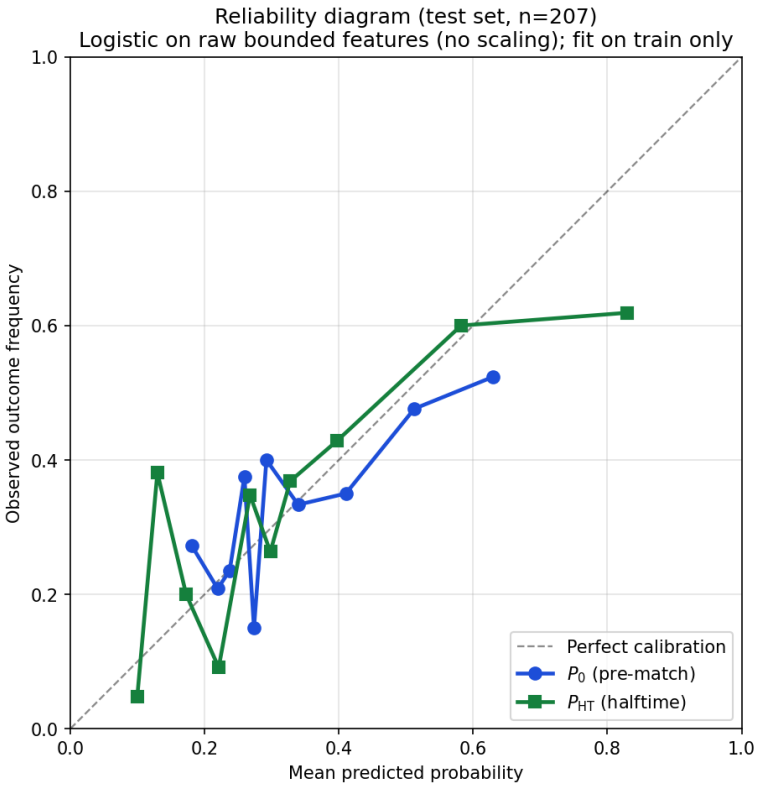


Figure 7: Reliability diagram on the test set ($n = 207$). P_{HT} (green) tracks the diagonal more closely than P_0 (blue) in the 0.2–0.7 range; tail miscalibration remains for both.