SportsLabs Research
Technical Report SLR-2026-01 | October 2026

Pricing, Mispricing and Staking in Sports Betting Markets

Empirical foundations for a positive-expected-value analytics engine, with original backtests on 30,632 European football matches (2012/13–2024/25)

Abstract. We provide the scientific foundation for the SportsLabs pricing engine. Using football-data.co.uk odds for six European leagues over thirteen seasons (30,632 matches), we test six pre-registered hypotheses. (i) Pinnacle closing odds are close to calibrated; the power and Shin de-vig methods reduce the favourite–longshot bias left by multiplicative normalisation (calibration slope 1.013 and 1.023 vs. 1.049, the last z=3.46), but log-loss differences between methods are economically tiny ($< 1.5 \times 10^{-4}$ nats) and not significant. (ii) Closing prices beat pre-closing prices in every season (log-loss gain 0.0024, t=6.0). (iii) A consensus-mispricing strategy in the spirit of Kaunitz et al. (2017) is profitable *only* when the consensus is the sharp book: betting when the best market price exceeds de-vigged Pinnacle pre-closing fair odds by $\geq 2\%$ yields 11,937 bets, ROI +4.7% (95% cluster-bootstrap CI $[1.6, 8.0]$), mean closing-line value (CLV) +3.4%, whereas the same rule built on the multiplicatively de-vigged market average loses money (−2.7%). No configuration survives a Holm correction across 75 tests (minimum adjusted p=0.15), realised ROI decays after 2018 and is not significant in walk-forward out-of-sample tests, yet CLV stays positive (+2.6 to +3.3%) in every sub-period. (iv) CLV predicts realised ROI one-for-one (slope 0.91, SE 0.09) and detects an edge with roughly 300 times fewer bets than ROI. (v) Dynamic Dixon–Coles and Elo models are 0.021 nats worse than the close and add no out-of-sample information. (vi) The Poisson "at least one" transform that underpins anytime-scorer pricing is well calibrated but slightly under-predicts scoring (71.6% vs. 72.7%). We translate these results into an engine specification—sharp-consensus de-vig, EV thresholds, quarter-Kelly with exposure caps, and CLV as the primary performance metric—and develop the player-prop methodology (rate×opportunity, empirical-Bayes shrinkage, correlated same-game parlays) that the engine applies where markets are thinner.

Reproducibility. All code, data and intermediate results are in `research/` (Appendix B). **Disclaimer.** Research only; not financial advice. Gambling involves risk of loss.

Contents

1 Introduction **3**

2 Literature review **3**

3 Mathematical framework **4**
3.1 Implied probability and overround 4
3.2 De-vig methods 5
3.3 Expected value, CLV and Kelly 5
3.4 Variance, power and sample size 5
3.5 Rating models 5

4 Sport-by-sport multifactor specifications **6**

5 Data and methodology **6**
5.1 Data 6
5.2 Methodology 7

6 Results **7**
6.1 Calibration and choice of de-vig method (H1) 7
6.2 Pre-closing versus closing accuracy (H2) 9
6.3 Consensus-mispricing strategy (H3) 9
6.4 CLV as the performance metric (H4) 12
6.5 Staking: flat versus Kelly (Monte Carlo) 13
6.6 Do ratings models add information? (H5) 14
6.7 The Poisson "at least one" transform (H6) 14

7 Player-prop modelling methodology **15**
7.1 Why prop markets are softer than sides and totals 15
7.2 Rate × opportunity decomposition 15
7.3 Empirical-Bayes shrinkage 16
7.4 Correlation and same-game parlays 16
7.5 Validation protocol for props 17

8 Live betting and latency **17**

9 Opportunity set ranked by durability **17**

10 Engine specification derived from the findings **17**

11 Limitations, risk, account limiting and responsible gambling **18**

A Formulas and numerical details **20**

B Code and data pointers **20**

List of Figures

1 Reliability of de-vigged Pinnacle closing probabilities (all outcomes stacked, 20 quantile bins). Right: observed minus predicted with 95% binomial intervals. . . 8

2 ROI from betting every outcome by decimal-odds band at three price sources, 2012/13–2024/25 (bars: 95% CI). 9
3 ROI by edge threshold. (a) Consensus definitions at the market-maximum price. (b) Pinnacle consensus at increasingly executable prices; wide second-best-book intervals at $\tau \geq 3\%$ are clipped at the axis limits. 10
4 Cumulative flat-stake profit for four strategies. 11
5 Season-by-season realised ROI and mean CLV, Pinnacle consensus, Max price, $\tau = 2\%$. 12
6 Realised ROI against CLV decile (91,896 priced outcomes). 12
7 Left: realised bankroll paths (log scale). Right: Monte Carlo distribution of maximum drawdown for fractional Kelly. 13
8 Out-of-sample log-loss of market and model forecasts. 14
9 Calibration of $1 - e^{-\lambda}$ for "team scores at least once" with market-implied λ. . . 15

List of Tables

1 Factor sets and model forms by sport. 6
2 Forecast accuracy of de-vigged Pinnacle closing odds by method (30,632 matches). ΔLL is the paired log-loss difference vs. multiplicative with 95% bootstrap CI. FLB slope b from the stacked logistic calibration regression ($b = 1$: no residual bias). 8
3 Consensus-mispricing backtest, 2012/13–2024/25, unit stakes. CI: 95% match-cluster bootstrap. p: one-sided bootstrap for ROI≤ 0; p_{Holm} over all 75 configurations. 10
4 Era split for Pinnacle-consensus rules ($\tau = 2\%$). 95% match-cluster bootstrap CIs. 11
5 Monte Carlo staking outcomes, closing line as truth (1,000 paths, 5,929 bets). "Ruin": bankroll falls below 20% of initial at any time. 13
6 Out-of-sample forecast accuracy, 2014/15–2024/25 (25,881 matches). 14
7 Opportunity classes, ranked from most to least durable edge. 17
8 Default engine parameters and their empirical justification. 18

1 Introduction

A sports-betting "edge" is a claim that a price o offered by some bookmaker exceeds the reciprocal of the true outcome probability, $o\,p > 1$. Two facts make such claims hard to establish. First, bookmakers are skilled forecasters who shade prices strategically (Levitt, 2004), and the closing prices of high-limit books aggregate the information of the whole market. Second, betting returns are extraordinarily noisy: at decimal odds of 3.0 a single bet's profit has standard deviation ≈ 1.4 stakes, so a genuine 2% edge needs roughly 31,000 bets to be detected by realised return alone (Section 3.4). Practitioner claims are therefore usually under-powered, over-fitted, or both (Winkelmann et al., 2024).

This report has two goals. The first is empirical: to measure, on a large public dataset and with explicit statistical testing, which of the ideas a +EV engine relies on actually hold. The second is constructive: to derive from those measurements a concrete specification for the SportsLabs engine, including the player-prop markets (NHL and soccer anytime goalscorer, NFL anytime touchdown, NBA player points) it will price.

We pre-specify six hypotheses, each tested in Section 6:

Hypothesis H1 (De-vig) *Methods that model the favourite–longshot bias (power, Shin, additive) yield better-calibrated probabilities from closing odds than multiplicative normalisation.*

Hypothesis H2 (Information flow) *Closing prices are more accurate than pre-closing prices.*

Hypothesis H3 (Consensus mispricing) *Betting the best available price when it exceeds a de-vigged consensus by a threshold τ earns a positive return, out of sample and net of realistic execution haircuts.*

Hypothesis H4 (CLV) *Closing-line value is an unbiased predictor of realised return (slope 1) and a far more powerful test statistic.*

Hypothesis H5 (Models) *A team-ratings model (Elo or Dixon–Coles) adds information beyond market prices.*

Hypothesis H6 (Poisson props) *With a market-implied scoring rate λ, $\Pr(N \geq 1) = 1 - e^{-\lambda}$ is calibrated.*

Summary of findings. H1 is supported for calibration bias but not for scoring rules; H2 is strongly supported; H3 is supported in-sample only for a sharp-book consensus and is not robust to multiplicity correction, era, or walk-forward selection; H4 is supported; H5 is rejected; H6 is supported up to a small, correctable bias. The engine design in Section 10 follows directly.

2 Literature review

Market structure. Levitt (2004) shows, with proprietary wager data from an NFL handicapping contest, that bookmakers do not simply balance books: they set prices that deviate from market-clearing levels to exploit predictable bettor biases, earning more than the commission alone. The implication for an engine is that soft-book prices embed deliberate, side-specific shading, not just noise.

Efficiency of football odds. Elaad et al. (2020) study the prices of fifty-one online bookmakers for over 16,000 English matches since 2010. Aggregate odds show no favourite–longshot or outcome-type bias, but individual bookmakers are inefficient: their prices do not fully use the information in competitors' prices. This is exactly the mechanism exploited by consensus strategies. Kaunitz et al. (2017) operationalise it: they bet when a bookmaker's odds exceed those implied by the cross-bookmaker average, report profits in a ten-year historical simulation

on closing odds, a six-month minute-level simulation and five months of live betting, and document that winning accounts were restricted. Franck et al. (2010) find betting-exchange prices more accurate than bookmakers' across the "Big Five" leagues, and Croxson and Reade (2014) show that exchange prices update "swiftly and fully" when goals are scored just before half-time.

Persistence. Winkelmann et al. (2024) combine simulation and fourteen seasons of real data for the five major European leagues. Inefficiencies appear in single seasons but are not persistent or systematic, and statistically significant single-season effects are expected even under full efficiency. Their warning about multiple testing and short samples motivates our use of Holm corrections and walk-forward selection.

From odds to probabilities. Shin (1993) derives bookmaker prices in the presence of a fraction z of insider money, yielding a de-vig method that removes more margin from longshots. Štrumbelj (2014) finds Shin probabilities more accurate than basic normalisation or regression approaches across sports and bookmakers. Clarke et al. (2017) propose the power method, show it universally outperforms multiplicative normalisation and outperforms or matches Shin on three large datasets, and show that Shin and additive coincide for two-outcome markets.

Models and evaluation. Maher (1982) introduced independent Poisson goals with team attack/defence strengths; Dixon and Coles (1997) added a low-score dependence correction and time-decay weighting, and reported positive returns in 1995/96. Hvattum and Arntzen (2010) use Elo differences in ordered-logit models and find betting odds superior to their rating-based forecasts. Hubáček et al. (2019) improve profitability by de-correlating model forecasts from bookmaker odds. Constantinou and Fenton (2012) argue for the ranked probability score (RPS) for ordinal football outcomes; we report RPS alongside log-loss and Brier score.

Derivative markets. Newall (2015) documents that UK bookmakers' heavily-advertised complex bets carry far higher margins: over the 2014 World Cup, overrounds averaged 0.281 on correct-score bets and 0.478 on first-goalscorer bets. Reade et al. (2020) find a significant favourite–longshot bias in exact-scoreline odds that is not present in match-result odds, and Kain and Logan (2014) find that over/under totals are a much poorer predictor of total points than the spread is of the margin in US sports. Ramirez et al. (2023) report mispricing linked to pre-match attention in tennis.

Staking. Kelly (1956) shows that maximising the expected logarithm of wealth maximises the long-run growth rate; Whitrow (2007) generalises Kelly staking to many simultaneous bets. For NFL margins, Stern (1991) shows the margin relative to the spread is approximately normal with standard deviation slightly under 14 points (13.86 in his sample), the basis of normal margin models. For shrinkage of noisy player rates we rely on the empirical-Bayes framework of Efron and Morris (1975).

3 Mathematical framework

3.1 Implied probability and overround

For an n-way market with decimal odds $o_1, \ldots, o_n$, the raw implied probabilities are $q_i = 1/o_i$ and the booksum is $B = \sum_i q_i$. The overround is $B - 1$; the bookmaker's expected margin on a balanced book is $1 - 1/B$. A de-vig method is a map $(q_1, \ldots, q_n) \mapsto (\pi_1, \ldots, \pi_n)$ with $\sum_i \pi_i = 1$.

3.2 De-vig methods

$$\text{Multiplicative:} \quad \pi_i = q_i / B, \tag{1}$$

$$\text{Additive:} \quad \pi_i = q_i - (B-1)/n, \tag{2}$$

$$\text{Power:} \quad \pi_i = q_i^k, \;\; k > 1 \text{ solves } \textstyle\sum_i q_i^k = 1, \tag{3}$$

$$\text{Shin:} \quad \pi_i = \frac{\sqrt{z^2 + 4(1-z)\, q_i^2 / B} - z}{2(1-z)}, \;\; z \in [0,1) \text{ solves } \textstyle\sum_i \pi_i = 1. \tag{4}$$

Multiplicative scaling removes margin proportionally and therefore leaves any favourite–longshot bias (FLB) in place. Additive, power and Shin all remove proportionally more margin from longshots. In Shin's model z is the share of insider money; we solve (3) and (4) by vectorised bisection (Appendix A).

3.3 Expected value, CLV and Kelly

Let p be the true probability and o the price taken. The expected profit per unit stake is $\mathrm{EV} = o\,p - 1$. With a fair-probability estimate $\hat{p}$ the *edge* is $\hat{e} = o\hat{p} - 1$. Let π^c be the de-vigged closing probability of a sharp book. The closing-line value of a bet is

$$\mathrm{CLV} = o\,\pi^c - 1, \tag{5}$$

i.e. the bet's EV if the close is the truth. If closing prices are calibrated, $\mathbb{E}[\text{profit} \mid \mathrm{CLV}] = \mathrm{CLV}$, which H4 tests.

For a single binary bet with net odds $b = o - 1$ the Kelly fraction maximising $\mathbb{E}\log W$ is

$$f^* = \frac{b\,p - (1-p)}{b} = \frac{o\,p - 1}{o - 1}, \tag{6}$$

with growth rate $g(f) = p\log(1 + fb) + (1-p)\log(1-f)$. Fractional Kelly λf^* ($\lambda = \frac{1}{2}, \frac{1}{4}$) trades growth for variance: to second order $g(\lambda f^*) \approx (2\lambda - \lambda^2) g(f^*)$, so half-Kelly keeps about 75% of the growth with half the volatility. For m simultaneous independent bets the exact solution maximises $\mathbb{E}\log\left(1 + \sum_j f_j R_j\right)$ over the 2^m outcome states (Whitrow, 2007); we use the common approximation of scaling single-bet fractions proportionally whenever their sum exceeds an exposure cap C (here $C = 0.25$ per day).

3.4 Variance, power and sample size

A unit bet at odds o with win probability p has profit variance $\sigma^2 = o^2 p(1-p)$. To detect a true ROI μ with one-sided size α and power $1-\beta$, one needs

$$n \approx \left(\frac{(z_{1-\alpha} + z_{1-\beta})\,\sigma}{\mu} \right)^2. \tag{7}$$

With $\alpha = 0.05$, $1-\beta = 0.8$ and fair odds of 2.0 ($\sigma = 1$), $n \approx 15{,}456$ bets for $\mu = 2\%$; at odds 5.0 ($\sigma = 2$), $n \approx 61{,}825$. Replacing realised profit by CLV, whose per-bet standard deviation is an order of magnitude smaller, shrinks n by the squared ratio (Section 6.4).

3.5 Rating models

Elo. Home rating advantage H; expected score $E = 1/(1 + 10^{-(R_h + H - R_a)/400})$; update $R_h \leftarrow R_h + K\,m(g)\,(S - E)$ with $S \in \{1, \frac{1}{2}, 0\}$ and goal-difference multiplier $m(g)$. Ratings map to three-way probabilities via an ordered logit, $\Pr(Y \le j) = \Lambda(c_j - \beta\,\Delta R)$ (Hvattum and Arntzen, 2010).

Poisson / Dixon–Coles. Goals $X \sim \text{Pois}(\lambda)$, $Y \sim \text{Pois}(\mu)$ with $\log\lambda = \mu_0 + \gamma + \alpha_h + \delta_a$, $\log\mu = \mu_0 + \alpha_a + \delta_h$, and joint pmf multiplied by $\tau_\rho(x,y)$ for $x,y \le 1$: $\tau(0,0) = 1-\lambda\mu\rho$, $\tau(0,1) = 1+\lambda\rho$, $\tau(1,0) = 1+\mu\rho$, $\tau(1,1) = 1-\rho$ (Dixon and Coles, 1997). Parameters are fitted by time-weighted likelihood with weights $e^{-\xi\Delta t}$.

Normal margin models. For high-scoring sports, the final margin M is modelled as $M \sim \mathcal{N}(s, \sigma^2_{\text{sport}})$ around the market spread s, so $\Pr(\text{win}) = \Phi(s/\sigma)$ and $\Pr(\text{cover } \ell) = \Phi((s-\ell)/\sigma)$; Stern (1991) estimates $\sigma \approx 13.86$ for the NFL. The engine estimates σ per sport and season as the residual standard deviation of $M - s$ on its own closing-line data; key-number effects (NFL margins of 3 and 7) require a discrete correction to the normal.

Live win probability. If a fraction f of the game has elapsed with lead L, and pre-game expected margin is s, then under a Brownian approximation $M \mid L \sim \mathcal{N}\big(L + s(1-f),\, \sigma^2(1-f)\big)$ and $\text{WP} = \Phi\big((L+s(1-f))/(\sigma\sqrt{1-f})\big)$. For soccer the remaining goals are Poisson with rates $\lambda(1-f)$, $\mu(1-f)$ adjusted for game state. Bayesian updating of team strength uses the pre-game market as the prior and in-game scoring rates as the likelihood; the posterior weight on in-game evidence grows with f.

4 Sport-by-sport multifactor specifications

Table 1 lists, for each sport the engine covers, the factor set and model form that the literature and our results support. The common principle (Section 6.6) is that a fundamentals model is used as a *prior* or as a pricing device for thin markets, never to overrule a liquid sharp close.

Table 1: Factor sets and model forms by sport.

Sport	Core factors	Model form
NFL	QB-adjusted efficiency (EPA/play), pace, injuries, rest/travel, weather (wind, precipitation), home field	Normal margin around market spread; totals from pace×efficiency; key-number discrete adjustment
NBA	Net rating with lineup/minutes adjustment, rest and back-to-backs, pace, travel, injuries/load management	Normal margin; possessions×points-per-possession for totals; minutes projections drive player markets
MLB	Starting pitcher (strikeout/walk rates, contact quality), bullpen fatigue, lineup handedness splits, park and weather	Run-scoring Poisson/negative binomial by half-inning; moneyline via simulation
NHL	Shot-quality (expected goals), goaltender save-rate above expected (shrunk), special teams, back-to-backs	Poisson/Skellam goals with empty-net and overtime adjustments
Soccer	Expected goals for/against, time-decayed attack/defence, lineups, schedule congestion	Dixon–Coles bivariate Poisson; market-implied λ from 1X2 and totals (Section 7)
College	NCAAF/NCAAB: opponent-adjusted efficiency, roster turnover, heavy shrinkage early season, home court	As NFL/NBA with larger σ and stronger priors

5 Data and methodology

5.1 Data

We downloaded the free CSV files of football-data.co.uk for England (Premier League E0, Championship E1), Germany (D1), Spain (SP1), Italy (I1) and France (F1), seasons 2012/13–2024/25 (78 files; 30,665 matches). After requiring complete Pinnacle pre-closing, Pinnacle

closing, market-maximum and market-average 1X2 odds, 30,632 matches remain. The files contain: Pinnacle *pre-closing* odds (PSH/PSD/PSA) and *closing* odds (PSCH/PSCD/PSCA); market maximum and average pre-closing odds (BetBrain aggregates BbMx/BbAv until 2018/19, "market" Max/Avg from 2019/20); closing market maximum and average (MaxC/AvgC) for 2019/20 onward (14,017 matches); and between three and seven individual bookmakers (Bet365, Bet&Win, Interwetten, William Hill, VC Bet, Ladbrokes, Pinnacle; the count declines to ≈ 3.4 named books in 2024/25). The site's notes describe the main odds columns as pre-closing odds; exact timestamps are not provided, so "pre-closing" should be read as an earlier, bettable snapshot rather than a true opening line.

5.2 Methodology

Scoring. We report mean log-loss $-\log \hat{p}_y$, multi-class Brier score $\sum_j (\hat{p}_j - \mathbb{1}\{y = j\})^2$ and RPS. Differences between forecasters are tested by paired bootstrap over matches (1,000–4,000 resamples) and paired t-statistics.

Calibration and FLB. Stacking the three outcomes of every match, we regress the outcome indicator on $\text{logit}(\pi)$ in a logistic model with match-clustered standard errors. A slope $b > 1$ indicates that longshots are over-priced relative to favourites (FLB remaining after de-vig).

Betting backtests. Bets are unit stakes on single outcomes. ROI confidence intervals use a *cluster bootstrap by match* (2,000 resamples), since two outcomes of the same match can both be bet. One-sided p-values for H_0: ROI ≤ 0 use the null-centred bootstrap distribution. All 75 configurations of the consensus grid (3 consensus definitions $\times$ 5 price definitions $\times$ 5 thresholds) are corrected jointly with Holm (1979). *Walk-forward*: for each test season $s \geq 2014/15$ we choose the (consensus, τ) pair with the highest ROI on all seasons before s (minimum 300 bets) and apply it, unchanged, to season s.

Execution realism. The market maximum is an upper bound on what is bettable: it can be stale, an error, or limited to trivial stakes. We therefore report four alternatives: maximum with 1% and 2% haircuts on net odds, $1 + (o - 1)(1 - h)$; the best of the named mainstream books; and the *second*-best named book.

Models. Elo (K=20, H=60, goal-difference multiplier, 20% regression to 1500 between seasons, promoted teams at 1400) with an ordered logit refitted walk-forward. Dynamic Poisson ratings are refitted weekly per league on the previous 730 days with decay $\xi = 0.0019$/day (half-life ≈ 1 year), a weak prior that shrinks teams with little history (promoted sides) toward below-average strength, and a Dixon–Coles ρ estimated on the 2012/13–2013/14 burn-in ($\hat{\rho} = -0.02$). Models are evaluated on 2014/15–2024/25 (25,881 matches). Encompassing is tested by log-linear pooling $\hat{p} \propto \pi_{\text{mkt}}^{a}\, \hat{p}_{\text{model}}^{b}$ with a likelihood-ratio test of $b = 0$, and by a walk-forward blend whose weights are fitted only on past seasons.

6 Results

6.1 Calibration and choice of de-vig method (H1)

Mean overround is 2.46% for Pinnacle at the close, 2.64% for Pinnacle pre-close and 5.21% for the market average; the market-maximum booksum is only 0.19% above one, i.e. near-arbitrage is common at the best prices. Pinnacle's closing overround rose from 2.12% in 2012/13 to 3.12% in 2024/25.

Table 2: Forecast accuracy of de-vigged Pinnacle closing odds by method (30,632 matches). ΔLL is the paired log-loss difference vs. multiplicative with 95% bootstrap CI. FLB slope b from the stacked logistic calibration regression ($b = 1$: no residual bias).

Method	Log-loss	Brier	RPS	ΔLL ($\times 10^{-4}$) [95% CI]	FLB slope b (SE)
Multiplicative	0.97988	0.58360	0.19837	—	1.049 (0.014)†
Additive	0.97975	0.58351	0.19833	−1.30 [−2.96, 0.39]	—
Power	**0.97973**	**0.58350**	**0.19833**	−1.42 [−3.01, 0.42]	1.013 (0.014)
Shin	0.97976	0.58352	0.19834	−1.20 [−2.43, 0.02]	1.023 (0.014)

† H_0: $b = 1$ rejected, $z = 3.46$. Shin: $z = 1.68$; power: $z = 0.93$. Mean Shin insider share at the close $\bar{z} = 0.0124$.

The power method attains the best value on all three scores, but all differences are below 1.5×10^{-4} nats and none is significant at 5% (Shin is closest, $t = -1.89$). The calibration regression is more discriminating: multiplicative normalisation leaves a significant FLB ($b = 1.049$), Shin halves it and power removes it ($b = 1.013$, not significantly different from 1). Expected calibration error over 20 bins falls from 0.0068 (multiplicative) to 0.0054 (Shin). *Verdict on H1:* supported for bias, not for aggregate accuracy. The engine uses the power method by default, with Shin as a robustness check (Section 10).

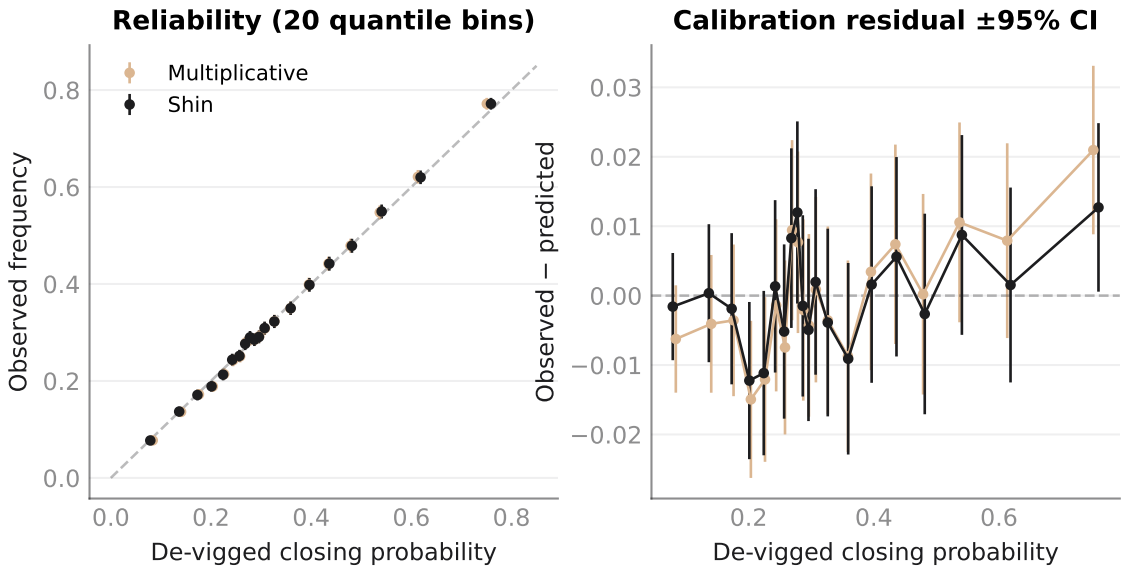


Figure 1: Reliability of de-vigged Pinnacle closing probabilities (all outcomes stacked, 20 quantile bins). Right: observed minus predicted with 95% binomial intervals.

Favourite–longshot bias in raw prices. Betting every outcome at a fixed price source shows the classic FLB (Figure 2): at Pinnacle's close ROI falls from +0.2% (odds 1.0–1.5) to −3.7% (3–5), −5.0% (5–10) and −15.4% (10+); at the market average from −0.5% to −20.7%. At the market *maximum*, short prices return +1.7% (1.0–1.5) and +1.9% (1.5–2), significant at roughly 2 standard errors each, a first indication that best-price shopping alone is near break-even on favourites.

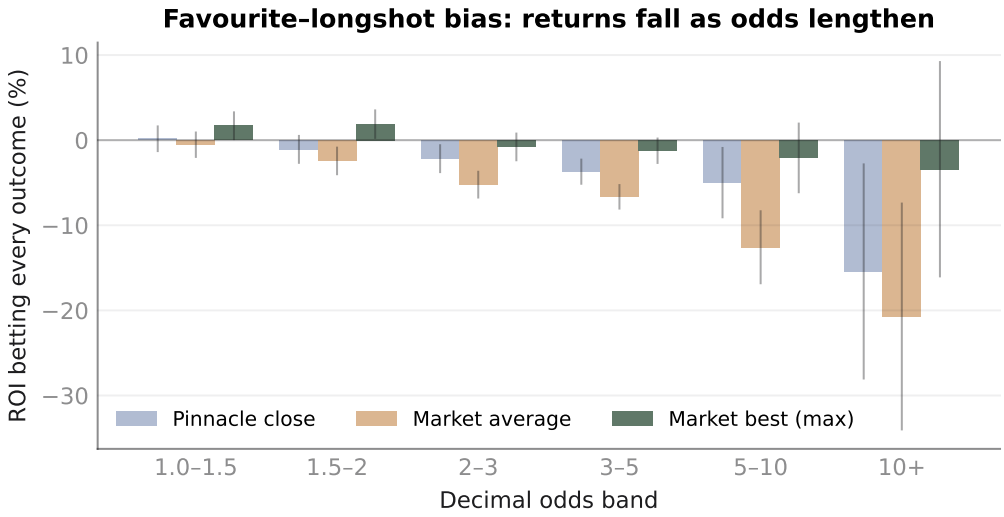


Figure 2: ROI from betting every outcome by decimal-odds band at three price sources, 2012/13–2024/25 (bars: 95% CI).

6.2 Pre-closing versus closing accuracy (H2)

Closing Pinnacle probabilities have lower log-loss than pre-closing ones: the paired difference is 0.00242 nats (95% CI $[0.00169, 0.00321]$, $t = 5.98$; Shin de-vig; multiplicative and power give $t = 6.17$ and 5.92). The close is better in 13 of 13 seasons. Late information (team news, sharp money) is incorporated between the two snapshots, and it is this movement that the CLV metric captures. *H2 is supported.*

6.3 Consensus-mispricing strategy (H3)

Table 3 reports the core grid. Three findings stand out.

(a) The consensus definition matters more than the threshold. With the market average de-vigged multiplicatively, the strategy loses at every threshold (-1.4% to -2.7%; mean CLV negative). The cause is the FLB documented above: proportional de-vig of a 5%-margin average leaves longshots looking under-priced, so the rule systematically buys over-priced longshots. Power de-vig of the same average is roughly break-even ($+0.9\%$ at $\tau = 0$, positive CLV $+1.0\%$). Using Pinnacle's own pre-closing price as the consensus produces the strongest and most monotone results.

(b) Edges at the Pinnacle consensus are real in CLV terms. At $\tau = 2\%$: 11,937 bets, ROI $+4.72\%$ (CI $[1.61, 8.03]$), mean CLV $+3.41\%$, and 69% of bets have positive CLV. Haircuts of 1% and 2% *raise* measured ROI ($+5.9\%$, $+6.4\%$) because they remove the marginal bets whose edge was within the haircut.

(c) Executable prices erode the edge. Restricted to the best of the named mainstream books, ROI at $\tau = 2\%$ falls to $+2.6\%$ (CI $[-2.5, 7.8]$) on 4,961 bets; at the second-best named book it is indistinguishable from zero (CLV $+0.2$ to $+0.7\%$). Much of the measured edge therefore sits in prices at the extreme of the market distribution, precisely the prices that are most often stale or limit-restricted.

Table 3: Consensus-mispricing backtest, 2012/13–2024/25, unit stakes. CI: 95% match-cluster bootstrap. p: one-sided bootstrap for ROI≤ 0; p_{Holm} over all 75 configurations.

Consensus	Price	τ	Bets	ROI %	95% CI	CLV %	p / p_{Holm}
Avg (mult.)	Max	0%	44,076	−1.39	[−3.10, 0.25]	−0.53	.95 / 1
Avg (mult.)	Max	2%	21,893	−2.72	[−5.79, 0.45]	−0.44	.96 / 1
Avg (power)	Max	0%	35,109	0.86	[−0.45, 2.21]	1.04	.11 / 1
Avg (power)	Max	1%	17,341	1.74	[−0.41, 3.96]	2.03	.07 / 1
Avg (power)	Max	2%	7,971	−0.24	[−4.12, 3.79]	3.27	.55 / 1
Pinnacle (power)	Max	0%	34,864	1.30	[−0.20, 2.70]	1.60	.04 / 1
Pinnacle (power)	Max	1%	21,210	2.44	[0.37, 4.56]	2.42	.01 / .70
Pinnacle (power)	Max	2%	11,937	**4.72**	[1.61, 8.03]	3.41	.002 / .15
Pinnacle (power)	Max	3%	6,750	6.98	[2.30, 11.91]	4.43	.002 / .15
Pinnacle (power)	Max	5%	2,519	9.00	[−1.03, 19.57]	6.59	.05 / 1
Pinnacle (power)	Max−1%	2%	8,431	5.86	[1.98, 9.89]	3.37	.002 / .15
Pinnacle (power)	Max−2%	1%	10,988	4.57	[1.52, 7.78]	2.29	.002 / .15
Pinnacle (power)	Max−2%	2%	5,929	6.37	[1.59, 11.19]	3.35	.008 / .54
Pinnacle (power)	Best named	0%	14,522	1.60	[−0.78, 4.08]	1.38	.10 / 1
Pinnacle (power)	Best named	2%	4,961	2.64	[−2.51, 7.81]	3.38	.15 / 1
Pinnacle (power)	Best named	3%	3,015	4.81	[−2.27, 11.91]	4.46	.10 / 1
Pinnacle (power)	2nd named	0%	5,257	0.64	[−2.97, 4.30]	0.23	.36 / 1
Pinnacle (power)	2nd named	1%	2,260	2.63	[−3.86, 9.35]	0.67	.23 / 1

Full 75-row grid in `results/a3_consensus.json`. The smallest bootstrap p-values are at the resolution floor of 2,000 resamples.

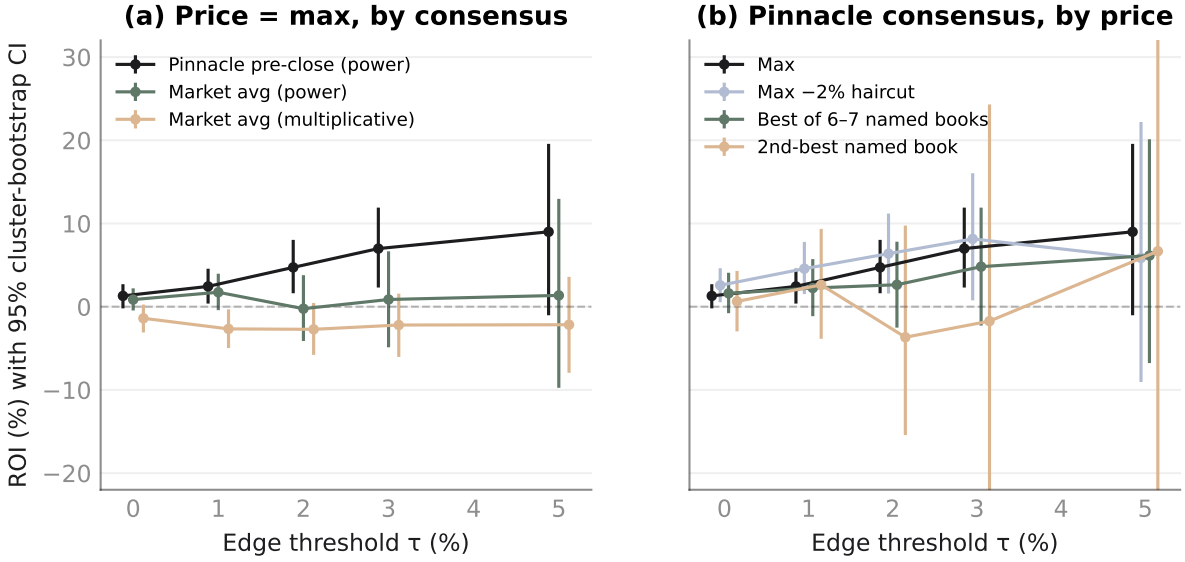


Figure 3: ROI by edge threshold. (a) Consensus definitions at the market-maximum price. (b) Pinnacle consensus at increasingly executable prices; wide second-best-book intervals at $\tau \geq 3\%$ are clipped at the axis limits.

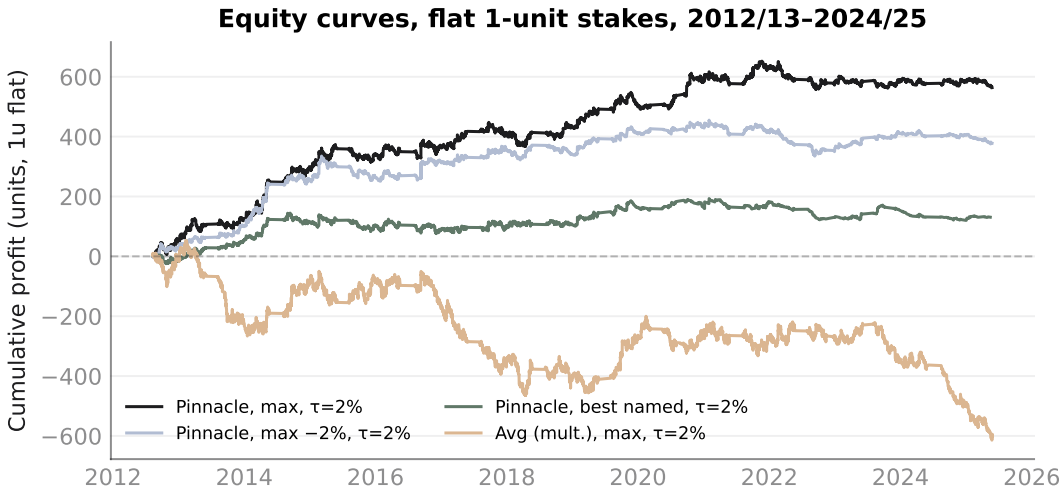


Figure 4: Cumulative flat-stake profit for four strategies.

Multiplicity, persistence and out-of-sample performance. No configuration survives Holm correction (minimum $p_{\text{Holm}} = 0.15$). Figure 5 and Table 4 show why caution is warranted. For the headline rule (Pinnacle, Max, $\tau = 2\%$), ROI was +6.3% (CI $[2.4, 10.5]$) in 2012/13–2018/19 but +1.7% ($[-3.7, 7.4]$) in 2019/20–2024/25 and −1.8% ($[-10.0, 6.0]$) in 2022/23–2024/25, while bet volume fell from $\approx$ 1,100 to $\approx$ 530 per season. In contrast, mean CLV barely moved: +3.6%, +3.0% and +3.3%. Walk-forward selection always picked the Pinnacle consensus (with τ = 3–5%), but its out-of-sample ROI, +3.6% on 2,041 bets at Max (CI $[-7.7, 14.9]$) and +3.9% on 2,784 bets at Max−2% ($[-4.3, 12.2]$), is not significant; at the best named book it is −0.1% ($[-9.4, 9.8]$) despite CLV of +5.2%. This pattern, persistent CLV with noisy and decaying realised ROI, is consistent both with Winkelmann et al. (2024) and with the power calculation in Section 6.4: a few thousand bets cannot confirm a 3% edge by ROI alone.

Table 4: Era split for Pinnacle-consensus rules ($\tau = 2\%$). 95% match-cluster bootstrap CIs.

	2012/13–2018/19			2019/20–2024/25		
Price	Bets	ROI % [CI]	CLV %	Bets	ROI % [CI]	CLV %
Max	7,817	6.29 [2.4, 10.5]	3.62	4,120	1.74 [−3.7, 7.4]	3.01
Max−2%	4,150	9.47 [3.3, 15.9]	3.69	1,779	−0.85 [−9.4, 8.2]	2.55
Best named	4,148	3.49 [−1.6, 8.6]	3.61	813	−1.73 [−16.0, 13.4]	2.25

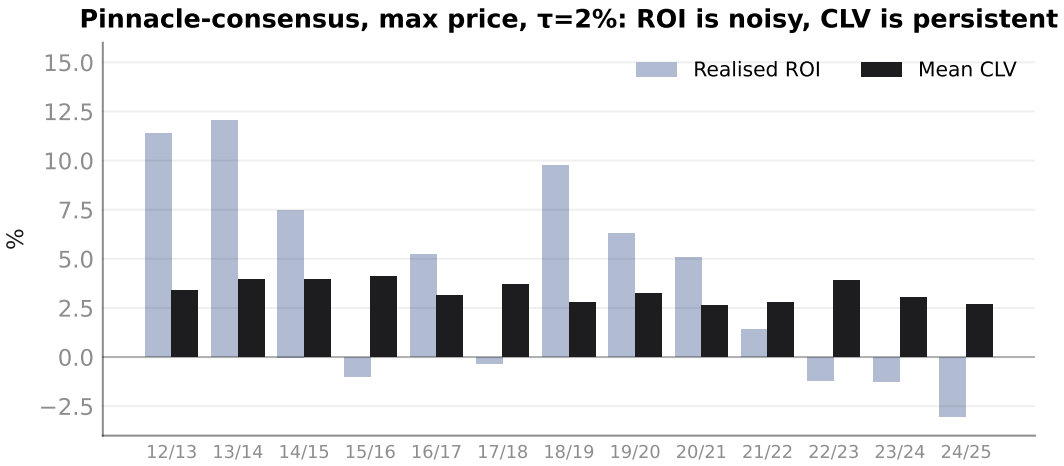


Figure 5: Season-by-season realised ROI and mean CLV, Pinnacle consensus, Max price, $\tau = 2\%$.

Closing snapshot. For 2019/20–2024/25 (14,017 matches) we repeat the test with closing Max odds against the de-vigged Pinnacle close. At $\tau = 1\%$, 14,260 bets return $+2.97\%$ (CI $[0.05, 5.69]$); at $\tau = 3\%$, $+4.85\%$ ($[-0.03, 9.32]$). Because the closing maximum is the price most likely to be stale or unavailable, we treat this as an upper bound rather than evidence of exploitability. *Verdict on H3:* supported in sample, for a sharp consensus, before execution frictions; not established out of sample or after multiplicity correction.

6.4 CLV as the performance metric (H4)

Across all 91,896 outcome-level prices at the pre-closing maximum, grouping by CLV decile reproduces realised ROI almost exactly (Figure 6): the bottom decile (mean CLV -16.2%) returns -16.0% and the top decile ($+15.1\%$) returns $+14.3\%$. A regression of per-bet profit on CLV with match-clustered errors gives slope 0.909 (SE 0.092; H_0: $b = 1$ not rejected) and intercept -0.34% (SE 0.32%). Within the headline strategy, the 8,216 bets with positive CLV returned $+9.4\%$ (SE 2.0) and the 3,721 with non-positive CLV returned -5.5% (SE 3.1). Pre-closing edge predicts CLV with slope 0.75 ($R^2 = 0.07$): a 1% pre-closing edge is worth about 0.75% at the close.

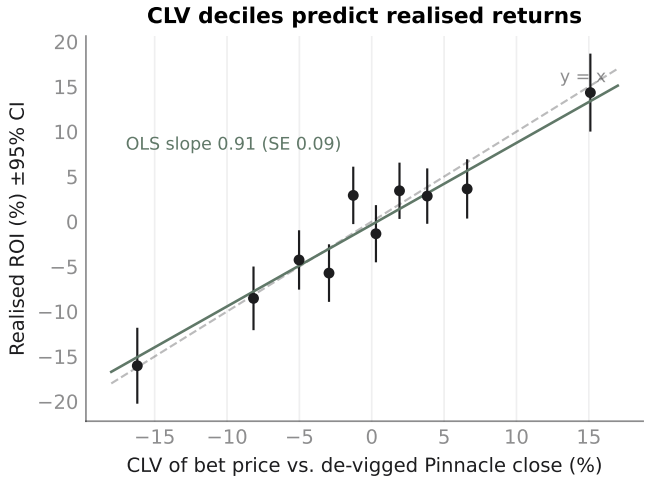


Figure 6: Realised ROI against CLV decile (91,896 priced outcomes).

Power analysis. For the headline strategy (mean odds 4.45) the per-bet standard deviation of profit is 1.83 stakes, whereas that of CLV is 0.0996. By (7), confirming an ROI of 1/2/3/5%

requires about 206,960 / 51,740 / 22,995 / 8,278 bets, whereas confirming a CLV of the same size requires 614 / 154 / 69 / 25 bets, a reduction of roughly 330 times. *H4 is supported*: CLV against a sharp close is the right primary metric, with realised ROI as a slow-moving check.

6.5 Staking: flat versus Kelly (Monte Carlo)

We stake the Pinnacle-consensus, Max−2%, $\tau = 2\%$ stream (5,929 bets over 1,712 match-days) with flat stakes of 1% of *initial* bankroll, and with full, half and quarter Kelly computed from the Pinnacle pre-closing fair probability, capping total daily exposure at 25% of bankroll. Realised paths (Figure 7) end at 4.8× (flat), 161.6× (full), 53.0× (half) and 9.5× (quarter), with maximum drawdowns of 21%, 88%, 63% and 38%. These terminal values are inflated by a realised ROI above its expectation and by the assumption of unlimited liquidity at the maximum price; they are not forecasts.

The Monte Carlo analysis is more informative. Holding the bet sequence fixed and drawing outcomes from the de-vigged Pinnacle close (expected ROI +3.35%), 1,000 simulations give Table 5. Full Kelly has the highest median outcome but a 100% chance of a drawdown of at least 50% and a 17% chance of losing 80% of the bankroll; quarter Kelly loses money in only 1.1% of paths and never suffers an 80% drawdown.

Table 5: Monte Carlo staking outcomes, closing line as truth (1,000 paths, 5,929 bets). "Ruin": bankroll falls below 20% of initial at any time.

Staking	Median final	5th pct	P(loss)	P(ruin)	P(DD≥50%)	Median MDD	95th pct MDD
Flat 1% of initial	2.98	0.70	8.3%	20.9%	57.3%	54.6%	100%
Full Kelly	21.47	0.36	9.9%	16.6%	100%	85.4%	96.8%
Half Kelly	12.50	1.25	3.5%	0.9%	78.0%	58.0%	79.2%
Quarter Kelly	4.75	1.59	1.1%	0.0%	7.5%	33.8%	52.0%

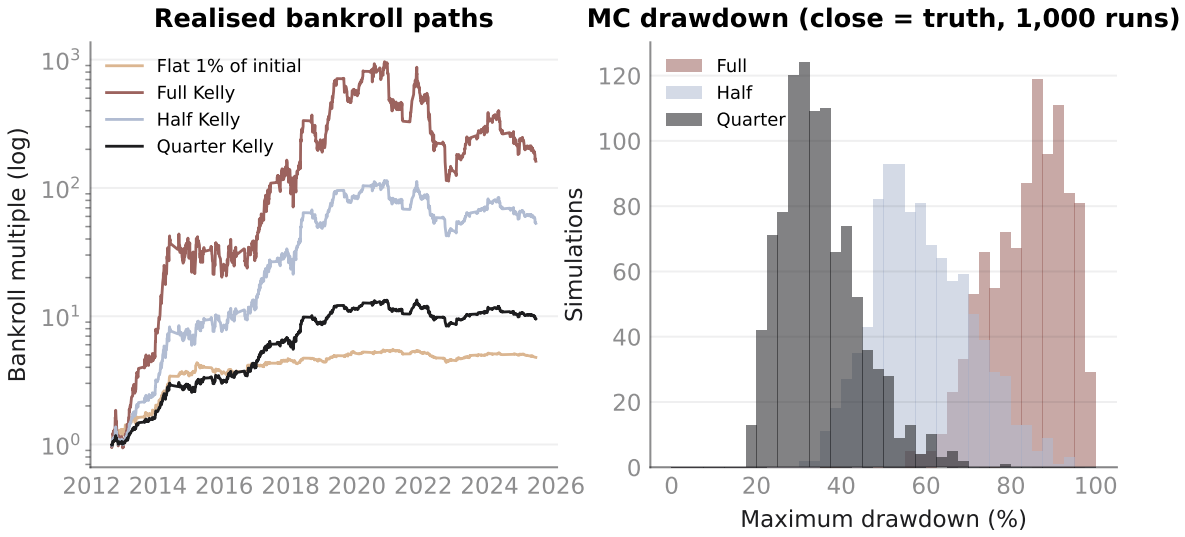


Figure 7: Left: realised bankroll paths (log scale). Right: Monte Carlo distribution of maximum drawdown for fractional Kelly.

Note that "flat 1% of initial bankroll" is *not* conservative here: the stream turns over 59 initial bankrolls, so fixed stakes do not shrink after losses. Proportional staking (Kelly fractions) is what controls ruin.

6.6 Do ratings models add information? (H5)

Table 6 and Figure 8 compare forecasts over 2014/15–2024/25 (25,881 matches). Both models clearly beat base rates but are about 0.021 nats per match worse than the Pinnacle close (DC − close: +0.0207, CI [0.0184, 0.0229]). In-sample pooling assigns the models a small *negative* weight ($b = -0.10$ for DC, -0.14 for Elo, alongside market exponents $a \approx 1.1$). The likelihood-ratio tests are significant ($p = 0.008$ and $p < 0.001$), but this reflects the pooled forecast sharpening the market rather than the models contributing positive information. Out of sample, walk-forward blends change log-loss by between -1.1×10^{-4} and $+0.7 \times 10^{-4}$ nats, all with CIs spanning zero. Betting the models' own edges at the market maximum loses: DC at $\tau = 2\%$, 32,301 bets, -3.2% (CI $[-5.3, -1.3]$), mean CLV -1.0%; Elo at $\tau = 5\%$, -4.4%. *H5 is rejected* for liquid 1X2 markets.

Table 6: Out-of-sample forecast accuracy, 2014/15–2024/25 (25,881 matches).

Forecaster	Log-loss	Brier	RPS
Pinnacle close (power)	**0.98003**	**0.58377**	**0.19828**
Pinnacle pre-close (power)	0.98232	0.58532	0.19898
Dynamic Dixon–Coles	1.00074	0.59796	0.20467
Elo + ordered logit	1.00234	0.59906	0.20521
Base rates	1.07267	0.64894	0.22936

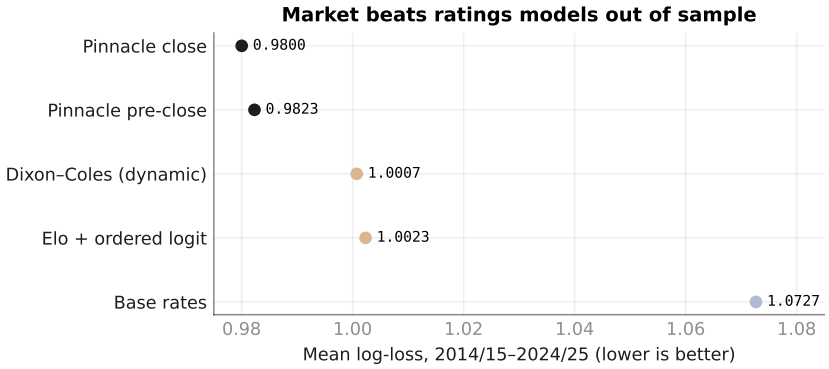


Figure 8: Out-of-sample log-loss of market and model forecasts.

6.7 The Poisson "at least one" transform (H6)

Anytime-scorer pricing rests on $\Pr(N \geq 1) = 1 - e^{-\lambda}$. We test it at team level, where outcomes are plentiful. For each match we infer the total goal rate λ_T from de-vigged over/under 2.5 odds ($\Pr(\text{Pois}(\lambda_T) \geq 3) = \pi_{\text{over}}$) and split it into λ_h, λ_a so that independent Poissons reproduce the de-vigged Pinnacle closing home-win probability. Over 30,596 matches (61,192 team-games) the implied rates are unbiased for goals (mean $\hat{\lambda} = 1.365$ vs. 1.358 observed) and $1 - e^{-\hat{\lambda}}$ is well calibrated across deciles (Figure 9; Brier 0.1855 vs. 0.1986 for a constant). It slightly under-predicts scoring, 71.6% vs. 72.7%, because goals are mildly *under*-dispersed relative to Poisson (variance/mean of residuals 0.93), so the zero class is over-predicted (28.4% vs. 27.3%). A logistic recalibration has intercept 0.044 (SE 0.017) and slope 1.016 (SE 0.017). *H6 is supported* up to a small, correctable bias: the engine adds a fitted intercept (or uses a slightly under-dispersed count distribution) when converting player rates to "anytime" probabilities.

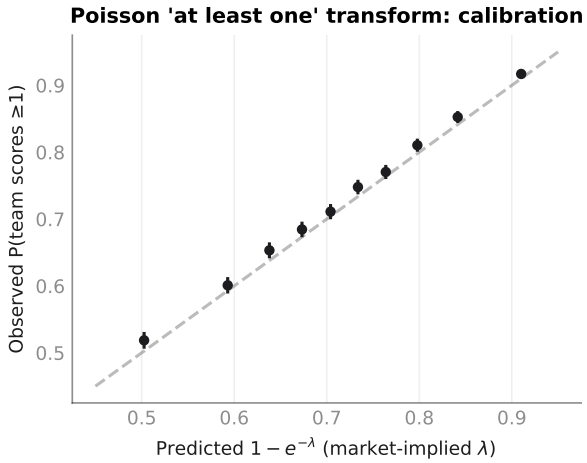


Figure 9: Calibration of $1 - e^{-\lambda}$ for "team scores at least once" with market-implied λ.

7 Player-prop modelling methodology

The engine prices four prop families: NHL anytime goalscorer, NFL anytime touchdown, NBA player points and soccer anytime goalscorer. The results above dictate the architecture: in liquid markets the sharp price is the best estimate available and models do not beat it (H5), so models earn their keep where no sharp consensus exists. Props are such markets.

7.1 Why prop markets are softer than sides and totals

The peer-reviewed evidence is indirect but consistent. Margins rise with market complexity: first-goalscorer overrounds averaged 0.478 at the 2014 World Cup against 0.281 for correct scores (Newall, 2015). Derivative markets show biases that the main market does not: exact-scoreline odds exhibit a significant favourite–longshot bias absent from match-result odds (Reade et al., 2020), and totals are much poorer predictors than spreads (Kain and Logan, 2014). Bookmakers also shade prices toward recreational preferences (Levitt, 2004), and recreational demand for props is concentrated on "overs" and stars. Structurally, props have lower limits and therefore less sharp money moving them, and their prices depend on inputs that are revealed late (lineups, minutes, power-play units). We found no peer-reviewed study that directly measures the efficiency of US player-prop prices, so the claim that props are *exploitable* (as opposed to merely higher-margin) is a hypothesis the engine must test with its own CLV tracking, not an established fact.

7.2 Rate × opportunity decomposition

Every prop is modelled as an *opportunity* (minutes, time on ice, snaps, carries plus targets, shots) multiplied by a *rate* per opportunity, and both are anchored to the game environment implied by the market:

$$\lambda_i = \underbrace{\mathbb{E}[\text{opportunity}_i]}_{\text{role, minutes, game script}} \times \underbrace{r_i}_{\text{shrunk skill rate}} \times \underbrace{\kappa_{\text{game}}}_{\text{market-implied team environment}} . \tag{8}$$

- **Soccer anytime scorer.** $\lambda_i = \lambda_{\text{team}} \cdot s_i \cdot m_i/90$, where λ_{team} is market-implied as in H6, s_i is player i's shrunk share of team expected goals (non-penalty xG share plus penalty-taker share × penalty rate), and m_i is expected minutes (starter probability × minutes if starting, plus substitute minutes). Then $\Pr = 1 - e^{-\lambda_i}$, recalibrated as in H6.

- **NHL anytime goalscorer.** $\lambda_i = \left(\mathrm{TOI}_i^{\mathrm{EV}}\, g_i^{\mathrm{EV}} + \mathrm{TOI}_i^{\mathrm{PP}}\, g_i^{\mathrm{PP}}\right) \cdot \kappa$, with goal rates per 60 minutes decomposed as shots/60 $\times$ shrunk shooting percentage, κ the ratio of market-implied team goals to the team baseline, and an empty-net term for trailing/leading scenarios. Then $\Pr = 1 - e^{-\lambda_i}$.
- **NFL anytime touchdown.** Team expected touchdowns follow from market-implied points, $\mathrm{Pts}_{\mathrm{home}} = (T + S_h)/2$ for total T and home spread margin S_h (positive if home favoured), times touchdown share of points. Player $\lambda_i = \mathbb{E}[\mathrm{TD}_{\mathrm{team}}] \cdot (\text{rush share}_i\, \phi^{\mathrm{rush}} + \text{target share}_i\, \phi^{\mathrm{rec}})$, with red-zone opportunity shares, and $\Pr = 1 - e^{-\lambda_i}$ (plus return/defensive TD contributions for eligible players).
- **NBA player points.** Points = minutes $\times$ points per minute. Minutes $M_i \sim \mathcal{N}(\mu_M, \sigma_M^2)$, truncated, with μ_M from rotation, injuries, rest and blow-out risk (a function of the spread); conditional on M, points $\sim \mathrm{NB}(\mu = M r_i \kappa_{\mathrm{pace}},\ k)$ with dispersion k estimated per player archetype. The over/under probability integrates over minutes, $\Pr(P > \ell) = \int \Pr(P > \ell \mid M)\, dF(M)$, by quadrature or simulation; a normal approximation with $\mathrm{Var}(P) = \mathbb{E}[M]\, r\kappa + \mathrm{Var}(M)(r\kappa)^2 +$ (NB extra) is adequate for high-usage players.

7.3 Empirical-Bayes shrinkage

Player rates are noisy. Following Efron and Morris (1975), we shrink each observed rate toward a position- or role-level prior, with the prior strength estimated from the population:

$$\text{Gamma–Poisson (counts per exposure):} \quad \hat{r}_i = \frac{\alpha + y_i}{\beta + t_i}, \qquad r \sim \Gamma(\alpha, \beta), \tag{9}$$

$$\text{Beta–binomial (conversion rates):} \quad \hat{\theta}_i = \frac{a + x_i}{a + b + n_i}, \qquad \theta \sim \mathrm{Beta}(a, b), \tag{10}$$

with (α, β) and (a, b) fitted by marginal maximum likelihood or method of moments over all players at the position, so that Var of observed rates = prior variance + sampling variance. Shooting percentage (NHL), goals per shot (soccer) and red-zone touchdown rate (NFL) need heavy shrinkage. Shot and target volume stabilise much faster and need light shrinkage. Time decay (exponential weights) handles role changes, and a role-change flag resets the prior when a player's usage shifts (trade, promotion to PP1, injury to a teammate).

7.4 Correlation and same-game parlays

For binary legs A, B with marginals p_A, p_B and phi correlation ϕ,

$$\Pr(A \cap B) = p_A p_B + \phi \sqrt{p_A(1 - p_A)\, p_B(1 - p_B)}, \tag{11}$$

so ignoring a positive ϕ (a striker scoring and his team winning; a quarterback's passing yards and his receiver's touchdown) understates the joint probability. The engine prices same-game parlays (SGPs) by *joint simulation* of the game state (team scoring, then player allocation conditional on team totals). This preserves the structural correlations, including negative ones such as teammates competing for the same touchdowns or shots. A Gaussian copula on the simulated marginals is a fallback when simulation is too slow. Sportsbooks apply their own correlation adjustments and larger margins to SGPs, so a +EV SGP requires the book's correlation adjustment to be wrong by more than its extra margin. The engine therefore reports $\mathrm{EV} = o_{\mathrm{SGP}}\, \hat{p}_{\mathrm{joint}} - 1$ and flags legs whose correlation is unusually priced, rather than assuming SGPs are favourable.

7.5 Validation protocol for props

Each prop model is validated (i) on calibration and log-loss against outcomes, (ii) against the de-vigged consensus of the sharpest available prop books (an encompassing test as in H5), and (iii) prospectively by CLV against the prop's own closing consensus. A prop model is promoted to live recommendation only when its prospective CLV over at least several hundred bets has a 95% CI excluding zero. By Section 6.4, that is about 150 bets for a 2% edge.

8 Live betting and latency

In-play prices update quickly and fully after major news (Croxson and Reade, 2014), although Angelini et al. (2022) find pre-match and in-play mispricing on a betting exchange that is explained by a reverse favourite–longshot (favourite) bias and grows after surprising goals. The engine's live layer therefore (i) uses the Bayesian win-probability models of Section 3 with the pre-game close as the prior; (ii) treats any edge that depends on reacting faster than the book to a scoring event as infeasible, given that books suspend markets and that our ingestion latency is measured in seconds; and (iii) focuses on slower-moving state mispricing, such as red cards, game-state effects on totals and player minutes in blow-outs, where models carry information that a quickly re-opened price may not. Latency budget, quote age and suspension status are logged with every live recommendation, and a quote older than its staleness threshold is never shown as +EV.

9 Opportunity set ranked by durability

Table 7 ranks opportunity classes by how durable we expect the edge to be. The ranking combines our evidence (H3, H4) with the cited literature, and is a judgement rather than a test.

Table 7: Opportunity classes, ranked from most to least durable edge.

Class	Mechanism and key formula	Main risk
Promotions and bet insurance	Free bet (stake not returned) at back odds b, hedged by laying at l with commission c: guaranteed profit $S(b-1)(1-c)/(l-c)$. Insurance refunding a losing stake as a free bet: $\mathrm{EV} = S[(o\,p-1)+(1-p)\,\rho]$, ρ = free-bet conversion rate	Promo terms, account flags
Line shopping vs sharp consensus (+EV)	$o\,\pi^{\text{sharp}} - 1 > \tau$; validated by CLV (H3, H4)	Limits, stale prices (Sec. 11)
Niche props	Model prices in thin markets (Sec. 7)	Low limits, model error
Arbitrage	$\sum_i 1/o_i^{\max} < 1$; stakes $s_i \propto 1/o_i$ guarantee $1/\sum_i 1/o_i^{\max} - 1$	Palps, voids, limits, account closure
Middles	Under the normal margin model, $\Pr(\text{middle}) = \Phi\left(\frac{b-s}{\sigma}\right) - \Phi\left(\frac{a-s}{\sigma}\right)$ for lines $a < b$	Small probability, needs volume
SGP correlation	Joint simulation vs book's correlated price	Higher margins, opaque pricing
Exchanges / prediction markets	Lowest margins; price-taker vs maker economics, commission	Liquidity, fees, regulation

10 Engine specification derived from the findings

1. **Odds ingest.** Poll all available books; store every quote with source, timestamp, limit (if known) and suspension status. Retain full history to compute closing lines.

2. **Normalisation.** Map events, markets and selections to canonical IDs; convert American/fractional to decimal; reject quotes older than a staleness threshold, and quotes more than a set multiple of the market dispersion away from consensus (likely errors).
3. **De-vig consensus.** Use the sharpest liquid book (Pinnacle-type) where it exists, de-vigged by the *power* method (Table 2: no residual FLB). Fallback: a weighted consensus of several books, also power de-vigged. Never use a multiplicatively de-vigged average of soft books (Table 3: -2.7%).
4. **Model blend.** For liquid sides and totals, weight = 0 on team-ratings models (H5). For props and thin markets, the model is the prior and the consensus of available prop books the likelihood; blend weights are fitted walk-forward by log-linear pooling.
5. **EV and stake.** Recommend when $\hat{e} = o\,\hat{p} - 1 \geq \tau$ with $\tau = 2\%$ by default (the threshold where CLV is $\approx 3.4\%$ and ROI was strongest in-sample). Raise to 3% for stale-prone books and use 1–2% only for books with verified fills. Stake quarter Kelly (Table 5: P(ruin)$=0$, P(DD$\geq$50%)$=7.5\%$), capped at 25% total daily exposure, with per-bet caps set by book limits.
6. **CLV tracker.** For every recommendation, record the bet price, the closing sharp price and CLV by (5). Report rolling mean CLV with a bootstrap CI and the implied expected ROI. Demote a market, book or prop model whose 500-bet rolling CLV CI includes zero. Realised ROI is reported, but not used for promotion decisions (Section 6.4).

Table 8: Default engine parameters and their empirical justification.

Parameter	Default	Justification
De-vig method	Power (Shin check)	Best scores; FLB slope 1.013 vs 1.049 multiplicative
Consensus source	Sharp book	Avg-mult consensus loses 1.4–2.7%; Pinnacle consensus earns +1.3 to +7.0% in-sample
Edge threshold τ	2% (1–3%)	CLV rises monotonically with τ (1.6→6.6%) while volume falls; 2–3% is the best trade-off in-sample
Execution haircut	1–2% on net odds	ROI positive after 2% haircut; second-best price $\approx$ break-even
Staking	0.25 Kelly, 25% daily cap	Lowest drawdown risk with high growth (Table 5)
Primary KPI	CLV vs sharp close	Slope 0.91 (SE 0.09); $\sim$330$\times$ fewer bets to detect edge
Props conversion	$1 - e^{-\lambda}$ + intercept	Calibrated; +1.1 pp under-prediction corrected

11 Limitations, risk, account limiting and responsible gambling

Data. One sport (European football), six leagues, 1X2 markets only. US data could not be obtained from a free, scriptable source during this study (the sportsbookreviewsonline archive returned HTTP 404), so the sport-specific sections for the NFL, NBA, MLB, NHL and college sports are methodological, not empirically validated here. Snapshot timestamps are not published, and the aggregate maximum mixes bookmakers whose identity, limits and staleness are unknown. The provider changed from BetBrain aggregates to "market" Max/Avg in 2019/20, and the number of named books fell, which confounds the era comparison.

Statistical. The grid contains 75 correlated tests; after Holm correction nothing is significant at 5%. Walk-forward ROI is not significant. Realised Kelly paths are single draws and are inflated by above-expectation outcomes. Monte Carlo results assume the close is the truth and that every bet is filled at the quoted price.

Account limiting. Kaunitz et al. (2017) were restricted within months of winning, and the strongest edges in our data sit at extreme prices that are the first to be limited. Any deployment must assume stake limits that shrink as CLV accumulates. The engine therefore tracks per-book fill rates and maximum accepted stakes, and its expected-value reporting should be net of them.

Model risk. Prop models depend on minutes and role projections that can change at lineup release; recommendations must be re-priced after news. SGP pricing depends on correlation estimates that are hard to validate.

Responsible gambling. Even a correctly identified edge leaves a substantial chance of long drawdowns: half-Kelly suffered a 50% drawdown in 78% of simulations. The product should present EV and CLV as estimates with uncertainty, never as guaranteed profit. It should support deposit and loss limits and self-exclusion, and should not target users showing signs of problem gambling. Sports betting is legal only in some jurisdictions and only for adults; the product must enforce both.

References

Angelini, G., De Angelis, L. and Singleton, C. (2022). Informational efficiency and behaviour within in-play prediction markets. *International Journal of Forecasting*. doi:10.1016/j.ijforecast.2021.05.012.

Clarke, S., Kovalchik, S. and Ingram, M. (2017). Adjusting bookmaker's odds to allow for overround. *American Journal of Sports Science* 5(6), 45–49. doi:10.11648/j.ajss.20170506.12.

Constantinou, A. C. and Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports* 8(1). doi:10.1515/1559-0410.1418.

Croxson, K. and Reade, J. J. (2014). Information and efficiency: goal arrival in soccer betting. *The Economic Journal* 124(575). doi:10.1111/ecoj.12033.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C* 46(2), 265–280. doi:10.1111/1467-9876.00065.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* 70(350), 311–319. doi:10.1080/01621459.1975.10479864.

Elaad, G., Reade, J. J. and Singleton, C. (2020). Information, prices and efficiency in an online betting market. *Finance Research Letters* 35, 101291. doi:10.1016/j.frl.2019.09.006.

Franck, E., Verbeek, E. and Nüesch, S. (2010). Prediction accuracy of different market structures—bookmakers versus a betting exchange. *International Journal of Forecasting* 26(3), 448–459. doi:10.1016/j.ijforecast.2010.01.004.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2), 65–70.

Hubáček, O., Šourek, G. and Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting* 35(2), 783–796. doi:10.1016/j.ijforecast.2019.01.001.

Hvattum, L. M. and Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting* 26(3), 460–470. doi:10.1016/j.ijforecast.2009.10.002.

Kain, K. J. and Logan, T. D. (2014). Are sports betting markets prediction markets? Evidence from a new test. *Journal of Sports Economics*. doi:10.1177/1527002512437744.

Kaunitz, L., Zhong, S. and Kreiner, J. (2017). Beating the bookies with their own numbers—and how the online sports betting market is rigged. arXiv:1710.02824. arxiv.org/abs/1710.02824.

Kelly, J. L., Jr. (1956). A new interpretation of information rate. *Bell System Technical Journal* 35(4), 917–926. doi:10.1002/j.1538-7305.1956.tb03809.x.

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets? *The Economic Journal* 114(495), 223–246. doi:10.1111/j.1468-0297.2004.00207.x.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica* 36(3), 109–118. doi:10.1111/j.1467-9574.1982.tb00782.x.

Newall, P. W. S. (2015). How bookies make your money. *Judgment and Decision Making* 10(3), 225–231. doi:10.1017/s1930297500004630.

Ramirez, P., Reade, J. J. and Singleton, C. (2023). Betting on a buzz: mispricing and inefficiency in online sportsbooks. *International Journal of Forecasting* 39(3), 1413–1423. doi:10.1016/j.ijforecast.2022.07.011.

Reade, J. J., Singleton, C. and Vaughan Williams, L. (2020). Betting markets for English Premier League results and scorelines: evaluating a forecasting model. *Economic Issues* 25(1), 87–106.

Shin, H. S. (1993). Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal* 103(420), 1141–1153. doi:10.2307/2234240.

Stern, H. (1991). On the probability of winning a football game. *The American Statistician* 45(3), 179–183. doi:10.1080/00031305.1991.10475798.

Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International Journal of Forecasting* 30(4), 934–943. doi:10.1016/j.ijforecast.2014.02.008.

Whitrow, C. (2007). Algorithms for optimal allocation of bets on many simultaneous events. *Journal of the Royal Statistical Society: Series C* 56(5), 607–623. doi:10.1111/j.1467-9876.2007.00594.x.

Winkelmann, D., Ötting, M., Deutscher, C. and Makarewicz, T. (2024). Are betting markets inefficient? Evidence from simulations and real data. *Journal of Sports Economics* 25(1), 54–97. doi:10.1177/15270025231204997.

football-data.co.uk (2026). Historical football results and betting odds data, and notes. `https://www.football-data.co.uk/` (accessed 18 October 2026).

A Formulas and numerical details

Solving the power and Shin de-vigs. For power, $h(k) = \sum_i q_i^k - 1$ is strictly decreasing in k for $q_i < 1$, with $h(1) = B - 1 > 0$, so bisection on $[1, 3]$ converges. For Shin, $h(z) = \sum_i \pi_i(z) - 1$ is decreasing on $[0, 0.4]$; 60 bisection iterations give machine precision. Both are vectorised over matches.

Market-implied Poisson rates (H6). λ_T solves $1 - F_{\text{Pois}}(2; \lambda_T) = \pi_{\text{over}}$; the split $s \in (0, 1)$ solves $\sum_{x>y} \text{Pois}(x; s\lambda_T)\,\text{Pois}(y; (1-s)\lambda_T) = \pi_H$ (Brent's method).

Cluster bootstrap. Let P_m and C_m be the summed profit and bet count of match m. A bootstrap replicate draws matches with replacement and computes $\sum P_{m^*} / \sum C_{m^*}$.

Holm correction. Ordered $p_{(1)} \le \cdots \le p_{(M)}$; adjusted $\tilde{p}_{(j)} = \max_{i \le j} \min\{1, (M - i + 1)p_{(i)}\}$, $M = 75$.

Kelly with daily cap. For bets j on day d, $f_j = \lambda \max\{0, (o_j \hat{p}_j - 1)/(o_j - 1)\}$; if $\sum_j f_j > C$, scale all f_j by $C / \sum_j f_j$; stakes are $f_j W_d$, settled at day end.

B Code and data pointers

All paths are relative to

- `data/*.csv` (78 football-data.co.uk files), `data/all.pkl` (merged); `code/download.py`, `code/load.py`.
- `code/common.py`: de-vig functions, scoring rules.
- `code/a1_calibration.py` → `results/a1_calibration.json` (H1, H2, FLB).

- `code/a3_consensus.py`, `code/a3b_era.py` → `results/a3_consensus.json`, `a3b_era.json`, `bets.pkl` (H3).
- `code/a45_staking_clv.py` → `results/a45_staking_clv.json` (staking, CLV, power).
- `code/a6_models.py` → `results/a6_models.json` (H5); `code/a7_poisson_props.py` → `results/a7_poisson_props.json` (H6).
- `code/figs.py` → `figs/`; `paper/main.tex` → `paper/SportsLabs_Research_EV_Engine.pdf`.

Random seeds are fixed in each script. A full re-run takes under two minutes on a single CPU.