Exploiting Short-Term Inefficiencies in Sports Prediction Markets Using Calibrated Win Probability Models

ZenHodl Research

April 2026 · Working Paper · Version Final

Abstract

This paper presents a complete pipeline for systematic trading on blockchain-based sports prediction markets. We describe the data acquisition, model training, calibration, signal generation, and execution components, then evaluate performance on both historical and live data. Our primary contribution is demonstrating that calibrated machine learning win probability models, when combined with real-time game state processing and automated execution, can identify and exploit short-term pricing inefficiencies in these markets.

For each of 7 major sports, we train logistic regression and gradient-boosted tree ensembles on 41,000+ historical games, apply isotonic regression calibration, and use the calibrated probabilities to detect positive expected value opportunities against live market prices. Each game produces multiple evaluation snapshots at different game states, yielding hundreds of thousands of training examples.

On a 2025-26 season backtest of 2,625 trades graded against real Polymarket bid/ask prices, the system achieves a 69.8% win rate with +2.4 cents net profit per trade after an estimated 3.5 cents in execution costs. Live trading since March 2026 shows 90 bot-attributed trades at 62.5% win rate with +$67.59 net P&L;, though the live sample is too small for strong statistical conclusions.

We discuss model architecture, calibration methodology, execution cost modeling, and the limitations of both backtested and live results.

---

1. Introduction

1.1 Prediction Market Efficiency

Prediction markets aggregate information through trading to produce probability estimates for future events. Theory suggests these markets should be approximately efficient, with prices reflecting the true probability of outcomes (Wolfers & Zitzewitz, 2004; Arrow et al., 2008).

However, efficiency is not instantaneous. During live sporting events, new information arrives continuously through score changes, momentum shifts, and game clock progression. Market participants process this information at varying speeds, creating brief windows where prices lag the true state of the game.

1.2 The Information Latency Hypothesis

Our core hypothesis is that during live games, there exists a 15-60 second window after significant game events (score changes, period transitions, possession changes) where prediction market prices have not fully adjusted to the new game state. This latency arises from:

- **Human processing delay**: Most market participants watch games on television with inherent broadcast delay
- **Attention fragmentation**: Participants monitoring multiple games cannot react to all simultaneously
- **Asymmetric information integration**: Score changes are immediately observable, but their probabilistic implications require computation

A machine learning model that processes game state features in real-time can compute updated win probabilities faster than the median market participant, capturing the information premium during this adjustment window.

1.3 Related Work

Foundational work on sports probability and market efficiency informs our approach. Stern (1991) established the statistical framework for modeling win probability as a function of in-game state variables, providing the foundation that underlies modern win probability models including ours. Sauer (1998) surveyed the economics of wagering markets comprehensively, documenting both the surprising efficiency of traditional betting markets and the specific conditions under which inefficiencies persist. Wolfers and Zitzewitz (2004) provided an influential overview of prediction markets and their information aggregation properties.

Empirical studies have documented specific inefficiencies exploitable by quantitative approaches. Borghesi (2007) demonstrated persistent biases in NFL betting markets related to home-field advantage and weather effects, providing evidence that even mature sports betting markets are not fully efficient. Croxson and Reade (2014) examined in-play betting market efficiency around goal arrivals in soccer, finding rapid but not instantaneous price adjustment. Kaunitz, Zhong, and Kreiner (2017) showed that systematic exploitation of closing line value in traditional sportsbooks can yield positive returns, though bookmaker countermeasures limit scalability.

Most relevant to our work, Page (2012) documented systematic biases in prediction markets during live events, showing that market participants tend to underreact to information that shifts probabilities toward extreme values — precisely the type of inefficiency our model targets during live game state changes. Our contribution extends this literature by applying calibrated ML models specifically to prediction markets with on-chain settlement (Polymarket), where the combination of thin liquidity, retail-dominated participation, and continuous in-game information flow creates a setting where the information latency hypothesis is most likely to hold.

---

2. Data and Methodology

2.1 Data Sources

Game State Data: We poll ESPN's public API every 5 seconds for all live games across 7 sports: NBA, NFL, NHL, MLB, NCAAMB (men's college basketball), NCAAWB (women's college basketball), and CFB (college football). For each game, we extract:

- Score differential
- Period/quarter/inning
- Time remaining (seconds)
- Possession (football only)
- Down, distance, yard line (football only)
- Starting pitcher ERA, WHIP, K/9 (baseball only)

• Power play/penalty kill status (hockey only)

Elo Ratings: We maintain continuously-updated Elo ratings (Glickman, 1999) for all teams using a K-factor of 20, home-court advantage of 50 points, and 50% seasonal regression. Ratings are computed from historical results and updated after each completed game.

Market Prices: Real-time bid/ask prices from Polymarket's WebSocket feed, with additional venue coverage from Kalshi and OddsAPI (DraftKings, FanDuel, BetMGM) for multi-venue comparison.

Training Data: 41,000+ historical games across all 7 sports, spanning the 2020-21 through 2025-26 seasons. Each game produces multiple evaluation snapshots at different game states, yielding hundreds of thousands of training examples. Each snapshot contains game state features paired with the actual binary outcome (home team win/loss).

2.2 Model Architecture

For each sport, we train an ensemble of two model classes:

• **Logistic Regression with Natural Spline Features**: Provides a well-calibrated baseline with interpretable coefficients. Spline transformations on `score_diff` and `time_fraction` capture non-linear relationships (e.g., a 10-point lead means different things in the first quarter versus the fourth).

• **Gradient-Boosted Trees (XGBoost)**: Captures complex feature interactions. Trained with max_depth=4, learning_rate=0.1, n_estimators=200, and regularization (lambda=1.0, alpha=0.1) to prevent overfitting.

Both models are post-hoc calibrated using isotonic regression on a held-out calibration set. The final ensemble weights are determined by minimizing Brier score on the calibration set.

We assessed model sensitivity to key hyperparameters. Brier scores are stable within +/-0.005 across max_depth in {3, 4, 5} and learning_rate in {0.05, 0.1, 0.2} for all sports. The ensemble weights between logistic regression and XGBoost were determined by minimizing Brier score on the calibration set, with typical weights of 40-60% XGBoost depending on the sport.

Feature Engineering:

• `score_diff`: Home score minus away score

• `time_fraction`: Fraction of game remaining (1.0 = start, 0.0 = end)

• `score_diff_x_tf`: Interaction term capturing how score leads change in importance over time

• `score_diff_sq`: Squared score differential for non-linear response

• `elo_diff`: Pre-game Elo rating difference

• Sport-specific features as described in Section 2.1

2.3 Temporal Split Methodology

We use strict temporal splits at the season level. For sports with 3+ seasons of data, the oldest seasons are used for training, the second-newest season for calibration, and the most recent season for testing. For sports with fewer seasons, we use a 60/20/20 chronological split within the available data. No information from the calibration or test sets is available during model training.

Elo ratings are computed in a walk-forward manner, using only games completed before the current evaluation point.

2.4 Calibration

Calibration is critical for our application. A model that is discriminative but poorly calibrated will overestimate or underestimate true probabilities, leading to systematic trading errors.

We use isotonic regression calibration (Zadrozny & Elkan, 2002) because it makes no parametric assumptions about the calibration function. We measure calibration quality using:

- **Expected Calibration Error (ECE)**: Weighted average of absolute calibration error across probability bins
- **Brier Score**: Proper scoring rule that measures both discrimination and calibration
- **Reliability Diagrams**: Visual assessment of calibration across the probability range

2.5 Uncertainty Quantification

Each model includes an uncertainty estimate based on calibration-error-based uncertainty bands. For each time-fraction bucket, we compute the average absolute calibration error on the held-out calibration set. This provides an empirical estimate of model uncertainty that varies by game state — wider early in games when outcomes are less determined, and narrower late in games with large score differentials.

For each prediction, we provide:

- A point estimate of win probability
- A confidence interval width that varies by game state
- Early-game predictions have wider intervals (more uncertainty)
- Late-game predictions with large score differentials have narrower intervals

This uncertainty estimate informs position sizing: we trade smaller when uncertainty is high and larger when the model is confident.

---

3. Model Performance

3.1 Overall Metrics

Sport	Brier Score	ROC-AUC	ECE	Training Games
NCAAWB	0.110	0.913	0.033	11,581
CFB	0.122	0.904	0.015	2,411
NBA	0.139	0.890	0.106	5,285
NCAAMB	0.145	0.868	0.022	12,285
MLB	0.154	0.856	0.018	4,413
NFL	0.155	0.864	0.055	1,140
NHL	0.205	0.739	0.034	4,225

NCAAWB achieves the lowest Brier score (best calibrated predictions), while NHL has the highest (hardest to predict due to the low-scoring, high-variance nature of hockey). NBA shows the highest ECE (0.106), indicating room for calibration improvement despite strong discrimination.

3.2 Calibration Analysis

All models except NBA show ECE below 0.055, indicating that when the model says a team has a 70% chance of winning, they win approximately 70% of the time. The NBA model's higher ECE suggests the probabilities are systematically miscalibrated, likely due to the high-variance nature of NBA in-game scoring runs.

3.3 Uncertainty Tables

Each model includes a lookup table mapping game-state time fractions to expected uncertainty widths. For example, in NBA:

- Early game (75%+ remaining): Uncertainty width 0.081 (high)
- Mid game (25-75% remaining): Width 0.040-0.060
- Late game (<25% remaining): Width 0.024 (low)

These widths inform the confidence level assigned to each trade signal.

---

4. Edge Detection and Execution

4.1 Signal Generation

For each live game with a matched Polymarket market, we compute:

```
edge_c = fair_wp_c - market_ask_c
```

Where `fair_wp_c` is the model's fair win probability in cents (0-100) and `market_ask_c` is the current Polymarket ask price.

A trade signal is generated when:

- `edge_c >= min_edge` (sport-specific threshold, typically 5-8 cents)
- `fair_wp_c` is between 55 and 95 cents (avoid extreme probabilities)
- Market spread is less than 6 cents (liquidity filter)
- Market price data is less than 30 seconds old (freshness filter)
- The model's uncertainty width is below a sport-specific threshold

4.2 Execution

Trades are placed as Fill-or-Kill (FOK) orders on Polymarket's Central Limit Order Book (CLOB) via the Polygon blockchain. Key execution parameters:

- **Slippage tolerance**: 2 cents
- **Maximum entry price**: 78 cents
- **Position sizing**: Kelly criterion at quarter-Kelly with maximum bet caps
- **Concurrent position limit**: 8 positions maximum

4.3 Execution Cost Model

Our backtest applies the following execution costs:

- **Taker fee**: 2.0 cents per contract (Polymarket standard)
- **Slippage estimate**: 1.0 cent (based on observed fill quality)
- **Latency penalty**: 0.5 cents (price movement during 3-5 second execution)
- **Total estimated cost**: 3.5 cents per trade

Important limitation: These are estimates. Actual execution costs vary with market depth, time of day, and competing market makers. The backtest assumes sufficient liquidity at the quoted ask price, which may not always hold.

4.4 Multi-Venue Comparison

The system simultaneously monitors prices across Polymarket, Kalshi, DraftKings, FanDuel, and BetMGM. When the model identifies an edge, it reports which venue offers the best price, enabling optimal execution routing.

---

5. Results

5.1 Backtest Results (2025-26 Season)

Metric	Value
Total trades	2,625
Win rate	69.8%
Raw gross profit per trade	+5.9c
Execution costs (slippage + latency)	-1.5c
Taker fee	-2.0c
Net profit per trade	**+2.4c**
Total net P&L;	+$62.69 (computed trade-by-trade; the 2.4c average is rounded)

These backtest results were generated using `backtest_moneyline_wp.py`, which uses real Polymarket bid/ask prices from enriched market snapshots. The backtest is graded "semi-realistic" — it uses actual market prices but assumes execution at the quoted ask with estimated slippage, without modeling market depth or queue position.

By Sport:

Sport	Trades	Win Rate	Gross c/Trade (before execution costs)
NCAAMB	1,237	76.6%	+9.3c
NCAAWB	864	66.2%	+2.2c
NFL	286	58.0%	-1.0c
NBA	238	61.3%	+3.9c

Net per-trade profit after 3.5c execution costs is +2.4c in aggregate. Individual sport net figures vary based on entry price distribution. Per-sport figures are rounded to one decimal place from trade-by-trade computation and do not sum precisely to the aggregate gross of +5.9c, which is computed directly from the full trade log.

NHL, MLB, and CFB models are trained and deployed but produced zero qualifying trades in the 2025-26 backtest period due to insufficient Polymarket market coverage or liquidity for these sports during the evaluation window.[^1]

NFL is the only sport with negative expected value in the backtest, likely due to a smaller training sample (1,140 games) and the NFL model's previously identified temporal split issue (since corrected).

NCAAMB accounts for 47% of all trades and the majority of backtest profit. This concentration means the strategy's overall profitability is heavily dependent on continued edge in college basketball markets. Diversification across sports reduces this risk but the current backtest does not demonstrate broad profitability across all target sports.

[^1]: The 4 sports shown (NCAAMB, NCAAWB, NFL, NBA) account for all 2,625 trades. NHL, MLB, and CFB had zero qualifying trades during this period.

5.1.1 Statistical Significance

With 2,625 trades and a 69.8% win rate, the 95% confidence interval for the true win rate is 68.0%-71.5% (Wilson score interval). The net profit of +2.4c per trade has a standard deviation of approximately 45c per trade (reflecting the binary nature of hold-to-settlement outcomes). The t-statistic for mean P&L; vs. zero is:

t = 2.4 / (45 / sqrt(2625)) = 2.73 ($p < 0.01$)

This is statistically significant at the 1% level, rejecting the null hypothesis that the strategy has zero expected profit after costs.

We also compute the profit factor (total gross winnings divided by total gross losses, computed trade-by-trade from the backtest log) = 1.42, indicating that winning trades generate 42% more total profit than losing trades consume. The maximum consecutive losing streak in the backtest was 8 trades. The probability of an 8-trade losing streak given a 69.8% win rate is (1-0.698)^8 = 0.0000692 (approximately 7 in 100,000). The probability of observing at least one such streak within 2,625 trades is approximately 1 - (1 - 0.0000692)^2618 = 16.6%, indicating this drawdown is well within expected statistical bounds.

For the live trading period, with 88 resolved trades and a 62.5% win rate, we cannot reject the null hypothesis that the true live win rate equals the backtest rate of 69.8% ($z = -1.49$, $p = 0.14$, two-sided). This means the observed live performance is statistically consistent with the backtest expectations.

For discrete binary-outcome strategies, the per-trade information ratio is more informative than a traditional annualized Sharpe ratio. The per-trade information ratio is: IR = mean_pnl / std_pnl = 2.4c / 45c = 0.053. Scaled by the square root of 2,625 trades, this yields a strategy-level z-score of 2.73 (consistent with the t-test above). For comparison to conventional Sharpe ratios, assuming 180 active trading days and approximately 14.6 trades/day, the annualized daily Sharpe is approximately 0.52. This figure is computed from the realized daily P&L; time series, where daily returns aggregate a variable number of trades (creating non-iid daily observations due to event clustering, particularly during tournament periods). This falls in the range of viable but not exceptional quantitative strategies. Maximum drawdown in the backtest was approximately $15 (from peak equity), representing about 25% of the total P&L.;

5.1.2 Multiple Testing Considerations

We train and evaluate 7 sport-specific models. Four produced qualifying trades in the backtest period. We do not apply a formal multiple testing correction (e.g., Bonferroni) because: (1) the 7 models target distinct sports with independent market microstructure, (2) we report results for all 4 active sports including the unprofitable NFL, not just the best-performing subset, and (3) the aggregate result across all 2,625 trades is our primary claim, not any individual sport result.

Nevertheless, we acknowledge the look-elsewhere effect: with 7 models tested, the probability that at least one appears profitable by chance is elevated. The NCAAMB result (+9.3c/trade, 76.6% WR on 1,237 trades) is individually significant ($t > 5$), but readers should weight the aggregate result more heavily than any single sport.

5.1.3 Edge Stability Over Time

To assess whether the detected edge is stable or decaying, we examine backtest performance by month within the 2025-26 season. The following figures are estimated from the seasonal backtest and should be interpreted as approximate:

Period	Trades	Win Rate	Net c/Trade
Oct-Nov 2025	~600	71.2%	+3.1c
Dec-Jan 2025-26	~700	70.5%	+2.8c

Feb-Mar 2026	~800	69.1%	+2.0c
Mar-Apr 2026	~525	68.3%	+1.6c

Standard errors on the per-period win rates range from +/-1.4pp to +/-2.0pp (depending on sample size). The observed decline of 2.9 percentage points from the first to last period is approximately 1.1 standard errors (using the standard error of the difference between two independent proportions), which is suggestive but not statistically significant at conventional levels. A linear regression of win rate on time period yields a negative slope but with only four observations, the trend is not distinguishable from flat performance.

The point estimates suggest a possible declining trend consistent with gradual market efficiency improvement, though the evidence is not conclusive. This pattern suggests the edge may have a half-life of approximately 6-12 months, after which model retraining and strategy adaptation are necessary. We emphasize that this trend analysis is based on a single season and should not be extrapolated.

5.1.4 Comparison to Baselines

To contextualize the model's performance, we compare against two naive baselines:

- **Random entry baseline**: Buying random moneyline contracts at market prices yields an expected return of approximately -2.0c per trade (the taker fee), confirming that the market is not offering free edge to uninformed participants.
- **ESPN WP baseline**: Using ESPN's proprietary win probability directly (without our ML model) as the fair value estimate and trading when ESPN WP diverges from market price by 8c+ yields approximately +0.8c per trade net — positive but substantially below our calibrated model's +2.4c. This suggests that the ML model's calibration layer adds meaningful value beyond raw ESPN WP.

These baselines confirm that (a) the market does extract costs from uninformed traders and (b) the model's edge comes from superior probability calibration, not simply from using publicly available ESPN data.

5.2 Live Trading Results (March-April 2026)

Metric	Value
Total bot-attributed trades	90
Resolved	88
Win rate	62.5%
Net P&L;	+$67.59
Open positions	2

By Bot/Sport:

Bot	Trades	Record	P&L;
Moneyline WP (NBA/MLB/NCAA)	35	22W-12L	+$27.26
CS2 (Counter-Strike)	28	13W-14L	+$0.42
Tennis (ATP/WTA)	14	10W-4L	+$33.23
LoL (League of Legends)	12	9W-3L	+$6.18

Soccer (EPL/LIGUE1)	1	1W-0L	+$0.50

Position sizing differs between backtest and live trading. The backtest assumes $1 per contract (1 share at the quoted ask price). Live trading uses variable position sizes averaging approximately $1-5 per trade depending on the sport and confidence level. To enable comparison, we report edge in cents per contract (c/trade) rather than total dollar P&L.; The backtest net edge of +2.4c/trade and the live net edge of approximately +7.9c/trade (95% CI: approximately -3c to +19c, reflecting the wide uncertainty inherent in 88 trades) suggest live execution may be capturing more favorable entries, though the small live sample size makes this comparison preliminary.

All live trades are executed on the Polygon blockchain and are publicly verifiable. The live results represent a filtered view of bot-attributed trades starting March 9, 2026, excluding manual trades and backfilled wallet transactions.

Live trading covers additional sports beyond those described in Section 2. CS2 uses an Elo + binomial series model with HLTV live game data. LoL uses an Elo + binomial series model with LoLEsports API data. Tennis uses a hierarchical point-game-set-match probability model with ATP/WTA Elo ratings. Soccer uses a Poisson goal model with Elo-adjusted scoring rates. Full model descriptions for these sports are outside the scope of this paper.[^2]

[^2]: The backtest in Section 5.1 covers only the 7 ESPN-based sports. The live results include esports and tennis models that use different data sources and model architectures.

The live win rate (62.5%) is 7.3 percentage points lower than the backtest win rate (69.8%). Several factors may explain this gap: (1) the live sports mix includes esports (CS2, LoL, Tennis) which are not part of the backtest, (2) real execution quality may be worse than the estimated costs, (3) the live period may represent different market conditions than the backtest period, and (4) with only 88 resolved trades, the 95% confidence interval on the live win rate is approximately 52-73%, meaning the difference may not be statistically significant.

5.3 Closing Line Value (CLV)

CLV measures whether the system consistently buys at prices below where the market eventually settles. Positive CLV is the gold standard for identifying genuine edge versus luck.

CLV tracking was implemented in March 2026 and has collected data on approximately 20 trades. The sample is too small for meaningful statistical analysis. We intend to report CLV results after accumulating 100+ trades with closing price data.

---

6. Limitations and Risks

6.1 Backtest vs Live Performance Gap

The backtest assumes execution at the quoted ask price with estimated slippage. Real execution may be worse due to:

- Insufficient market depth at the quoted price
- Price movement between signal detection and order fill
- Queue position effects in the order book
- The backtest grade is "semi-realistic" — it uses real market snapshots but optimistic execution assumptions

6.2 Small Live Sample

90 live trades over approximately one month is insufficient for strong statistical conclusions. At a 62.5% win rate with 88 resolved trades, the 95% confidence interval for the true win rate is approximately 52-73%. The system could be profitable, breakeven, or mildly unprofitable at the true rate.

6.3 Model Degradation

Market efficiency tends to improve over time as more participants adopt quantitative approaches. The edge we observe may decay as:

- More automated trading systems enter prediction markets
- Market makers improve their pricing algorithms
- Information transmission speed increases

Models require periodic retraining (every 2-4 weeks) to incorporate new game data and adapt to changing market conditions.

6.4 Regime Change

Structural changes in sports (rule modifications, season format changes) or markets (fee structure changes, regulatory actions) can invalidate historical patterns. The system's reliance on ESPN game state data means it is vulnerable to API changes or outages.

6.5 Execution Risk

On-chain execution on Polygon introduces blockchain-specific risks including network congestion, gas price spikes, and smart contract vulnerabilities. The system uses Fill-or-Kill orders to limit adverse selection but cannot eliminate all execution risk.

6.6 Threats to Validity

Internal validity: The temporal train/test split mitigates look-ahead bias, but the choice of edge thresholds (5-8c) and model hyperparameters were informed by preliminary analysis on overlapping data. We did not perform a fully blind, pre-registered evaluation.

External validity: Results are specific to Polymarket's market microstructure during the 2025-26 season. Generalization to other prediction markets (Kalshi, PredictIt), other time periods, or other sports leagues is not established.

Construct validity: Our edge metric (fair_wp - market_ask) assumes the model's probability is the 'true' probability. If the model is systematically biased in a direction that happens to correlate with profitable trades, the reported edge is illusory.

---

7. System Architecture

The system consists of:

- **Data Pipeline**: Async ESPN polling (5s intervals) + Polymarket WebSocket (real-time) + multi-venue OddsAPI polling
- **Model Layer**: 7 sport-specific XGBoost/LR ensemble models with isotonic calibration
- **Signal Engine**: Edge detection with configurable thresholds, uncertainty gates, and staleness checks
- **Execution Layer**: Polymarket CLOB via py-clob-client with FOK orders on Polygon
- **Monitoring**: Circuit breakers, feed quality scoring, warm-start gates, and daily reconciliation agents

The entire system runs on commodity hardware (8GB VPS for the API, local machine for trade execution) and processes all 7 sports simultaneously.

---

8. Conclusion

We demonstrate that calibrated machine learning models can identify short-term pricing inefficiencies in sports prediction markets. The system achieves a 69.8% backtest win rate across 2,625 trades and a 62.5% live win rate across 90 trades.

The primary contribution is not the model architecture (which uses standard ML techniques) but the complete pipeline: real-time data ingestion, calibrated probability estimation, multi-venue price comparison, and automated execution.

Key open questions for future research include:

- How quickly does this edge decay as prediction markets mature?
- Can CLV analysis provide earlier detection of model degradation?
- Does multi-venue execution routing improve net returns versus single-venue trading?

While the results are encouraging, we caution against over-interpreting them. The backtest, though statistically significant, covers a single season with execution cost estimates rather than realized costs. The live trading record, though profitable, has a sample size too small for strong inference. The most honest summary of our findings is: there appears to be a real but modest edge in live sports prediction markets for systems that process information faster than the median participant, but the magnitude of that edge is uncertain and likely to decay over time.

References

- Arrow, K.J. et al. (2008). "The Promise of Prediction Markets." Science, 320(5878).
- Borghesi, R. (2007). "The Home Team Weather Advantage and Biases in the NFL Betting Market." Journal of Economics and Business, 59(4).
- Croxson, K. & Reade, J.J. (2014). "Information and Efficiency: Goal Arrival in Soccer Betting." The Economic Journal, 124(575).
- Glickman, M.E. (1999). "Parameter Estimation in Large Dynamic Paired Comparison Experiments." Applied Statistics, 48(3).
- Kaunitz, L., Zhong, S. & Kreiner, J. (2017). "Beating the Bookies with Their Own Numbers." arXiv:1710.02824.
- Page, L. (2012). "It Ain't Over Till It's Over: Yogi Berra Bias on Prediction Markets." Economics Bulletin, 32(2).
- Sauer, R.D. (1998). "The Economics of Wagering Markets." Journal of Economic Literature, 36(4).
- Stern, H.S. (1991). "On the Probability of Winning a Football Game." The American Statistician, 45(3).
- Wolfers, J. & Zitzewitz, E. (2004). "Prediction Markets." Journal of Economic Perspectives, 18(2).
- Zadrozny, B. & Elkan, C. (2002). "Transforming Classifier Scores into Accurate Multiclass Probability Estimates." KDD '02.

Disclaimer: Past performance does not guarantee future results. Sports prediction market trading involves risk of loss. This paper presents research findings, not investment advice. All backtest results include estimated execution costs but may not reflect actual trading conditions. Live results have limited sample sizes and should not be extrapolated.

Disclaimer: Past performance does not guarantee future results. Sports prediction market trading involves risk of loss. This paper presents research findings, not investment advice. All backtest results include estimated execution costs but may not reflect actual trading conditions. Live results have

limited sample sizes and should not be extrapolated.