

Peer Review

Review of: "Accuracy Without Profit: A Statistical Evaluation of Machine Learning Profitability in the English Premier League"

Elia Yathie Matsumoto¹

1. Sao Paulo School of Economics, Fundação Getúlio Vargas, São Paulo, Brazil

- **RELEVANCE:** Excellent, Good, Acceptable, Little, None
- **TECHNICAL/SCIENTIFIC RIGOUR:** Excellent, Solid, Valid, Marginal, Questionable
- **QUALITY OF PRESENTATION:** Excellent, Well Written, Readable, Revision needed, Unacceptable
- **NOVELTY/ORIGINALITY:** Pioneering, Significant, Interesting, Minor, None
- **OVERALL:** Strongly Acceptable, Accepted, Neutral, Rejected, Strongly rejected

Main paper contributions and Positive aspects

This paper provides a demonstration of why using just predictive accuracy can be a misleading metric for profitability in efficient markets like sports betting. By analyzing English Premier League data, the author showed that while machine learning models can achieve statistically distinct forecasts, they fail to generate consistent profits, primarily due to significant calibration errors that cause overconfident bets. This insight positively shifts the focus from merely optimizing accuracy to prioritizing well-calibrated probability estimates—an important lesson for applying predictive models in any real-world, adversarial decision-making context.

Positive aspects:

- The text makes a balanced discussion about machine learning techniques' outcomes and the nuances of human-driven markets.
- The feature engineering process is well described, detailing how rolling metrics like "Last 5 Games" transform raw match data into actionable trends.

- Methodologically, the study seems cautious, employing strict walk-forward validation to definitively eliminate look-ahead bias, thereby ensuring real-world applicability.
- The application of the Diebold-Mariano test provided statistical evidence that the model outputs are distinct from market odds, confirming they extract unique signals rather than merely replicating consensus.
- Most importantly, when faced with the apparent paradox of decent accuracy (~51%) coupled with poor profitability, the author delivers a diagnostic: the “models’ risk-averse” avoidance of high-value draws and the critical phenomenon of “Alpha Decay” as the market adapted post-2015.

Observed deficiencies and Suggestions for improvements

This section 2 (Related Work and Background) provides a conceptual foundation but could be strengthened by broadening its references. The literature cited is somewhat limited, particularly regarding recent advances in probability calibration techniques (e.g., modern post-hoc calibration methods) and a wider array of contemporary predictive models used in sports analytics. Expanding this review would better contextualize the study within current methodological discussions.

While the study provides good analysis using established tree-based ensembles, its methodological scope presents opportunities for enhancement. The exclusive reliance on Random Forest, XGBoost, and LightGBM means the input space is consistently explored through similar, non-linear splitting mechanisms, potentially missing insights from alternative model families.

Suggestions:

- A possible addition would be to include a Generalized Linear Model (GLM), such as multinomial logistic regression, as a benchmark; its linear assumptions might yield surprisingly robust or interpretable baselines, clarifying whether complexity is necessary.
- Furthermore, evaluating performance via a weighted accuracy metric could better reflect the economic value of correctly predicting rare, high-odds outcomes like draws.
- Lastly, to improve clarity, separating the currently combined Figure 1(a) and 1(b) into two distinct figures would enhance readability (confusion matrices are unreadable).

These changes could strengthen the experimental design and the communicative power of the results.

Declarations

Potential competing interests: No potential competing interests to declare.