Reproducible Evaluation of Football Lottery Strategies: Leakage, Calibration, and Temporal Backtesting

Abstract

Sports Forecasting Evaluation poses a recurring problem of coordinating performance with evidence quality, resource limits, and transfer across settings. This article critically maps building evaluation protocols that resist hindsight, selection bias, and unstable league conditions. It synthesizes 1 focal paper with 12 independently retrieved publications verified through persistent DOI or publisher records. The analysis is organized around temporal splits, data leakage, calibration, transaction assumptions, and replication. A central precaution is not to treat reported gains as context-free quantities, the review compares task scope, modeling premises, and evaluation limits. Across the literature, a robust inference is that advances in sports forecasting evaluation become credible when behavior, measurement, and decision context are evaluated together and when uncertainty about selection effects is reported explicitly. The framework consequently connects method selection to application risk, identifies recurring threats to external validity, and proposes a research agenda centered on traceable baselines, scenario testing, and reusable evidence.

Keywords

Sports Forecasting Evaluation; Temporal Splits; Data Leakage; Calibration; Transaction Assumptions; Replication

1. Introduction

Contemporary investigation of sports forecasting evaluation connects scientific ambition and practical constraint. Progress requires not only stronger nominal performance but also explanations of the mechanisms that support observed behavior. That challenge is especially visible in building evaluation protocols that resist hindsight, selection bias, and unstable league conditions. A mechanism demonstrated for one experimental medium, dataset, cohort, geography, or system load can erode beyond the conditions that produced it. For this reason, analysis must preserve the unit of analysis and the decision context. It should further distinguish a mechanism from a correlation, a benchmark advantage from a deployment advantage, and an interpretable diagnostic from a causal explanation.

Comparison proceeds across five linked dimensions: temporal splits, data leakage, calibration, transaction assumptions, replication. Their combination matters because they jointly cover the designed object, its measurement, and the invariants required for transfer. The review is anchored in behavior, measurement, and decision context. Under that principle, methodological novelty is necessary but insufficient; the comparison also tests whether comparisons, uncertainty, and operating limits have been specified. The analysis is literature based and does not claim new experiments, participants, measurements, or performance results.

2. Conceptual Background

The analytical chain starts from the input evidence, the transformation applied to it, the output used for evaluation, and the decision that follows. In sports forecasting evaluation, these elements are commonly tuned in isolation even though errors propagate across them. Model expressiveness can propagate bias that enters with the observations; an apparently accurate output can still be poorly calibrated for the final decision. A direct requirement is that, construct validity should accompany aggregate performance. An auditable study identifies what the protocol fixes and what it lets move, then selects metrics that correspond to the intended use rather than to convenience alone.

Boundary awareness provides the next principle. Temporal splits defines the local design problem, while replication determines whether the design can survive outside that boundary. Intermediate lenses such as data leakage and calibration connect those ends by exposing trade-offs that a single score conceals. Unexpected outcomes and perturbation analysis reveals the interval in which an explanation can still be defended. For applied work, documentation of responsible interpretation is therefore part of the scientific result, not an administrative afterthought.

3. Measurement and Evaluation

The target study BRUCE-01, A new playing method of the guessing football lottery [1], addresses a concrete part of sports forecasting evaluation through a football-lottery playing method that converts match judgments into ticket combinations. It identifies a representation, mechanism, or evaluation boundary needed to compare claims. Against the focus on building evaluation protocols that resist hindsight, selection bias, and unstable league conditions, the paper functions as a bounded point of comparison: transfer may depend on its sensing regime, preparation, workload, population, or benchmark. The synthesis retains the boundary while comparing its premises with adjacent studies.

Across the focal studies, temporal splits should be interpreted jointly with data leakage. A design can improve one while hiding a penalty in the other dimension. A structured comparison takes the form of a matrix whose rows are operating conditions and whose columns are evidence quality, computation or process cost, robustness, and consequence of failure. This organization helps prevent an attractive mechanism from being detached from the circumstances that support it. The matrix helps determine which ablations are informative: removing a component should test a stated causal role, not merely create another number.

Calibration joins methodological claims to their use. The evaluation protocol needs to evaluate baseline and stress regimes separately and expose result dispersion, and test plausible alternative explanations. Where selection effects is central, calibration or sensitivity is as important as ranking. Where costs or hazards are asymmetric, utility-weighted assessment is more revealing than undifferentiated accuracy. These choices do not make studies identical; they make the reasons for their differences auditable.

4. Comparative Evidence

For triangulation, the literature includes Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics [2]; The SEC vs. the Dow Jones: A Profitable Betting Strategy in NCAA Football [3]; Match Fixing and Sports Betting in Football: Empirical Evidence from the German Bundesliga [4]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of sports forecasting evaluation. Together they illuminate temporal splits: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

The design space is broadened by Why some Portuguese professional football players engage in betting related match-fixing? A general stra... [5]; Correlated Parlay Betting: An Analysis of Betting Market Profitability Scenarios in College Football [6]; Match-Fixing in a Monopoly Betting Market [7]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of sports forecasting evaluation. Together they illuminate data leakage: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

A first comparison set connects The globalised network of a dirty game: match-fixing, illegal betting and transnational organised crime... [8]; At risk: betting-related match-fixing in Brazilian football [9]; Validation of KOVO League Match Result Prediction and Sports Betting Profitability Using Artificial Inte... [10]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of sports forecasting evaluation. Together they illuminate calibration: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

The neighboring evidence base brings together Match Rigging and the Favorite Long-Shot Bias in the Italian Football Betting Market [11]; Football Match Outcome Prediction Using Machine Learning: A Temporal Benchmark of Six Classifiers on Eur... [12]; Football Match Analysis and Prediction Based on LightGBM Decision Algorithm [13]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of sports forecasting evaluation. Together they illuminate transaction assumptions: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

5. Responsible Application

For practice, five ordered decisions are useful. First, define the decision and its failure cost. Second, specify the evidence and operating envelope. Third, choose a method whose inductive assumptions match that evidence. Fourth, test transaction assumptions and replication under controlled perturbations. Finally, retain provenance so that an unexpected outcome can be traced to data, configuration, material batch, environmental condition, or user interaction. The workflow connects building evaluation protocols that resist hindsight, selection bias, and unstable league conditions connected to observable requirements.

At a wider level, responsible progress in sports forecasting evaluation is determined by coupling as well as local design. Interfaces determine how uncertainty moves between measurement and inference, between algorithm and operator, or between microscopic design and system behavior. Teams spanning disciplines should predefine acceptance rules and triggers for revising conclusions. When those agreements are explicit, performance tuning can proceed while preserving visible validity and safety requirements.

6. Limitations and Future Directions

The analysis cannot eliminate differences in vocabulary, protocol, and level of reporting. Publication-level checks establish traceability, whereas claim-level replication remains a separate task. Versioning is another source of uncertainty. The focal citations supplied as Scholar-verified records were preserved; uncertain or malformed records were documented separately rather than silently rewritten.

Future work should preregister comparison protocols where feasible, provide structured configuration records and assess a realistic transfer condition in responsible interpretation. Research should also group adverse cases by process and consequence. For sports forecasting evaluation, a valuable shared benchmark would combine ordinary performance, operational burden, uncertainty calibration, and resilience. Such a benchmark would reward designs that remain useful when evidence is incomplete or conditions change.

7. Conclusion

The source base for sports forecasting evaluation becomes clearer when treated as a linked evidence chain rather than a collection of isolated scores. The focal studies show how building evaluation protocols that resist hindsight, selection bias, and unstable league conditions; the additional literature reveals the assumptions required to interpret those contributions. Across temporal splits, data leakage, calibration, transaction assumptions, and replication, alignment is the most durable principle: the representation or mechanism must match the problem, assessment must align with action, just as deployment claims align with validation scope. Such alignment improves reproducibility, transfer safety, and diagnostic value under failure.

References

1. Liu, S., Wang, Y., & He, H. (2020, March). A new playing method of the guessing football lottery. In IOP Conference Series: Materials Science and Engineering (Vol. 790, No. 1, p. 012100). IOP Publishing.

2. Stübinger, J., Mangold, B., & Knoll, J. (2019). Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. Applied Sciences, 10(1), 46. https://doi.org/10.3390/app10010046

3. Moore, E., & Francisco, J. (2020). The SEC vs. the Dow Jones: A Profitable Betting Strategy in NCAA Football. The Journal of Prediction Markets, 13(2). https://doi.org/10.5750/jpm.v13i2.1740

4. Deutscher, C., Dimant, E., & Humphreys, B. R. (2017). Match Fixing and Sports Betting in Football: Empirical Evidence from the German Bundesliga. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2910662

5. Cima, C. s. (2026). Why some Portuguese professional football players engage in betting related match-fixing? A general strain theory perspective. Trends in Organized Crime. https://doi.org/10.1007/s12117-026-09597-5

6. Davis, J., Dawson, J., & Krieger, K. (2018). Correlated Parlay Betting: An Analysis of Betting Market Profitability Scenarios in College Football. The Journal of Prediction Markets, 12(2), 68-84. https://doi.org/10.5750/jpm.v12i2.1562

7. Bag, P. K., & Saha, B. (2016). Match-Fixing in a Monopoly Betting Market. Journal of Economics & Management Strategy, 26(1), 257-289. https://doi.org/10.1111/jems.12172

8. Costa, J. (2018). The globalised network of a dirty game: match-fixing, illegal betting and transnational organised crime in Italian football. Global Crime, 19(2), 125-145. https://doi.org/10.1080/17440572.2018.1460952

9. Marchetti, F., Reppold Filho, A. R., & Constandt, B. (2021). At risk: betting-related match-fixing in Brazilian football. Crime, Law and Social Change, 76(4), 431-450. https://doi.org/10.1007/s10611-021-09971-0

10. Kim, P., Lee, S., & Jeon, S. (2025). Validation of KOVO League Match Result Prediction and Sports Betting Profitability Using Artificial Intelligence Models. Journal of Contents and Industry, 7(5), 47-55. https://doi.org/10.55387/jci.2025.07.05.08

11. Rossi, M. (2011). Match Rigging and the Favorite Long-Shot Bias in the Italian Football Betting Market. International Journal of Sport Finance, 6(4), 317-334. https://doi.org/10.1177/155862351100600404

12. Nashwan, N. A., Sharrab, Y., Rousan, T. A., Nashwan, A. J., & Yaghi, M. A. (2026). Football Match Outcome Prediction Using Machine Learning: A Temporal Benchmark of Six Classifiers on European League Data. IEEE Access, 14, 97354-97366. https://doi.org/10.1109/access.2026.3701561

13. Chen, C. (2024). Football Match Analysis and Prediction Based on LightGBM Decision Algorithm. Informatica, 48(16). https://doi.org/10.31449/inf.v48i16.6236