Football Match Result Prediction Using Machine Learning and Binary Model Fusion

Douglas da S. Souza[1], Leandro S. de Araújo[1]

[1] Institute of Computing – Universidade Federal Fluminense – RJ, Brazil

douglas.souza@id.uff.br, leandro@ic.uff.br

Abstract. *This paper investigates English Premier League match result prediction using complementary machine learning approaches: direct multiclass classification, temporal goal forecasting, Isolation Forest-based binary anomaly detection, and simple binary decision-level fusion. Experiments were conducted on EPL data from the 2006–07 to the 2024–25 seasons using an expanding-window walk-forward validation protocol. Results show that betting odds concentrate most of the predictive information in direct classification, while temporal goal forecasting provides useful numerical estimates but does not necessarily improve outcome reconstruction. Isolation Forest underperforms supervised models in the binary setting, and binary fusion does not outperform the best individual model.*

1. Introduction

The prediction of football match outcomes is a challenging task due to the high uncertainty and variability involved in the sport. Match results depend on technical, tactical, physical, and contextual factors, such as home advantage, recent team performance, squad quality, injuries, suspensions, schedule effects, and random events during the match. Even with the increasing availability of historical data, the outcome of a football match remains noisy and difficult to predict [Berrar et al. 2019].

Machine learning methods have been widely explored as tools for sports analytics, including match outcome prediction, performance analysis, and decision support [Berrar et al. 2019]. In football, most studies formulate the problem as a supervised classification task, in which the model predicts one of three possible outcomes: home win, draw, or away win. However, this formulation is particularly difficult because draws are often harder to identify and because the information available before the match may already be partially reflected in betting odds.

Betting odds are especially relevant in this context because they summarize market expectations about the possible outcomes of a match [Igiri and Nwachukwu 2014]. They may incorporate several sources of information, such as team strength, recent form, home advantage, injuries, and public information available before the game. Therefore, evaluating models with and without odds is important to understand whether additional features provide predictive information beyond what is already captured by the market.

This work investigates the prediction of English Premier League (EPL) match results using complementary machine learning approaches. First, the problem is formulated as a direct multiclass classification task, comparing classical machine learning models. Second, the problem is reformulated as a temporal goal prediction task, in which home and away goals are estimated separately and then converted into a final match outcome. Third, Isolation Forest is evaluated in a binary formulation that distinguishes home wins from non-home wins. Finally, simple binary decision-level fusion strategies are applied to combine selected models from the previous stages, including configurations with and without Isolation Forest.

The main contributions of this paper are threefold. First, we compare direct multiclass classification models for EPL match outcome prediction using different feature sets and classical machine learning algorithms. Second, we evaluate alternative formulations of the problem, including temporal goal forecasting and an Isolation Forest-based binary anomaly-detection approach for distinguishing home wins from non-home wins. Third, we assess simple binary decision-level fusion strategies in an H versus non-H formulation, combining selected models from the previous stages with and without Isolation Forest under a temporal walk-forward validation protocol.

2. Related Work

Football match outcome prediction has been studied using machine learning and statistical approaches based on historical team statistics, home advantage, recent performance, contextual variables, and betting odds. Igiri and Nwachukwu [Igiri and Nwachukwu 2014] investigated EPL result prediction using data mining techniques, artificial neural networks, and logistic regression. Berrar et al. [Berrar et al. 2019] emphasized the importance of domain knowledge and feature engineering for soccer outcome prediction.

Another important line of research models football scores directly. Maher [Maher 1982] proposed a statistical model for football scores based on goal-count modeling, while Dixon and Coles [Dixon and Coles 1997] analyzed football scores and betting-market inefficiencies. Goddard [Goddard 2005] also investigated regression models for forecasting goals and match results.

Model combination has also been explored in football prediction. Zhang et al. [Zhang et al. 2019] proposed a fusion-based approach combining classifiers such as Support Vector Machine, Random Forest, and Naive Bayes. In contrast, this paper compares multiple EPL prediction formulations, including direct classification, temporal goal forecasting, Isolation Forest-based binary modeling, and simple binary decision-level fusion under an expanding-window walk-forward protocol.

3. Methodology

3.1. Dataset

The dataset contains English Premier League (EPL) matches from the 2006–07 to the 2024–25 seasons. Match results, betting odds and match statistics were obtained from Football-Data.co.uk, while aggregated team attributes were obtained from Sofifa. For direct classification, each instance represents one match and the target is the full-time result: home win (H), draw (D), or away win (A). After preprocessing and removal of matches without sufficient rolling history, this dataset contained 5,279 matches.

Four feature sets were evaluated (Section 4.1): `odds_only`, with only Bet365 pre-match odds; `odds_sofifa`, combining odds and Sofifa attributes; `sofifa_rolling`, combining Sofifa attributes and rolling statistics without odds; and `full`, combining all previous groups of variables. For goal prediction (Section 4.2), a separate dataset was built by organizing matches chronologically by team and season and computing rolling averages from previous matches, resulting in 6,257 matches. Unlike the direct classification stage, the regression experiments were evaluated on a single feature representation.

3.2. Predictive Formulations

Three main predictive formulations were evaluated. First, direct multiclass classification was performed using classical machine learning models, including Logistic Regression, Random Forest, XGBoost, LightGBM, Support Vector Machines, K-Nearest Neighbors and Gaussian Naive Bayes. Second, temporal goal prediction was formulated as a regression task, training independent models for full-time home goals (*FTHG*) and full-time away goals (*FTAG*) using Poisson Regression, Decision Tree, Random Forest, XGBoost and LightGBM. Predicted goals were converted into H/D/A outcomes using a decision margin, defined as a threshold applied to the predicted goal difference between the home and away teams. If the predicted goal difference was greater than the margin, the match was classified as H; if it was lower than the negative margin, it was classified as A; otherwise, it was classified as D. This margin was treated as a post-processing hyperparameter and optimized over values between 0.15 and 0.35 during the initial five-season training window to maximize the macro F1-score. Third, Isolation Forest [Liu et al. 2008] was evaluated in a binary setting, where home wins were treated as the normal pattern and draws or away wins were grouped as non-H. This approach investigates whether home wins exhibit a sufficiently distinct pattern to be separated from other outcomes.

In the Isolation Forest formulation, home wins were used as the reference class because they corresponded to the most frequent individual outcome in the original multiclass distribution. In the complete dataset, home wins represented 46.01% of the matches, compared with 24.02% draws and 29.97% away wins. Therefore, the objective was not to assume that home wins are normal in an absolute sense, but to investigate whether this frequent outcome forms a sufficiently identifiable pattern when compared with draws and away wins grouped as non-H.

For the fusion stage, all model outputs were converted to a common binary representation, H versus non-H. This choice was necessary because Isolation Forest was evaluated as a one-class/binary formulation, while the direct classification and goal forecasting models originally produced H/D/A predictions. The fusion was implemented as a simple and interpretable decision-level strategy. In configurations with three models, combining direct classification, goal prediction, and Isolation Forest, the final output was H when at least two models predicted H; otherwise, it was non-H. In configurations without Isolation Forest, the final output was H only when both remaining models predicted H.

3.3. Evaluation Protocol

All experiments followed an expanding-window walk-forward protocol. Models were initially trained on seasons 2006–07 to 2010–11 and tested on 2011–12; then the test season was incorporated into the training set, and the process was repeated until the final season. Hyperparameter tuning was restricted to the initial historical window and then kept fixed for the remaining splits.

Multiclass models were evaluated using accuracy, macro precision, macro recall, macro F1, log loss and computational time. Regression models were evaluated using MAE and RMSE, and their derived H/D/A predictions were also evaluated with classification metrics. Binary models and fusion were evaluated using accuracy, balanced accuracy, precision, recall, F1 and macro F1.

4. Experiments and Results

4.1. Classification Model Results

Table 1 summarizes the classification results obtained across four feature sets: `odds_only`, `odds_sofifa`, `sofifa_rolling`, and `full`. For the supervised classification models, the task was formulated as a direct multiclass prediction problem, where each model predicted one of three possible outcomes: home win (H), draw (D), or away win (A). In contrast, Isolation Forest was evaluated separately as a binary formulation, distinguishing home wins from non-home wins. Therefore, its results are not directly equivalent to the multiclass models, but are included in the table to summarize the performance of the classification-based approaches evaluated in this work.

The reported values correspond to the average performance across the walk-forward test seasons, rather than to a single train-test split. The best accuracy-oriented multiclass model was XGBoost with the `odds_only` feature set, reaching 55.28% accuracy. Among the feature sets, `odds_only` and `odds_sofifa` generally achieved competitive results, while `sofifa_rolling`, which excludes betting odds, tended to present lower performance. This suggests that betting odds concentrated a large portion of the predictive information available before the match.

Table 1. Results of direct classification models and Isolation Forest across feature sets.

Model	`odds_only`		`odds_sofifa`		`sofifa_rolling`		`full`	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
XGBoost	0.5528	0.4109	0.5492	0.4096	0.5287	0.3848	0.5508	0.4159
LightGBM	0.5469	0.4190	0.5480	0.3994	0.5259	0.3771	0.5474	0.4094
SVM RBF	0.5482	0.4006	0.5441	0.3898	0.5361	0.3854	0.5452	0.3902
Log. Reg.	0.5500	0.3974	0.5420	0.3907	0.5392	0.3957	0.5480	0.4041
Random Forest	0.4789	0.4222	0.5156	0.4335	0.5162	0.4229	0.5341	0.4419
Gaussian NB	0.5107	0.4453	0.5028	0.4451	0.5216	0.4282	0.5070	0.4603
KNN	0.4725	0.4267	0.4923	0.4370	0.4760	0.4180	0.4921	0.4322
Isolation Forest	0.5281	0.4912	0.5276	0.5248	0.5235	0.5220	0.5186	0.5172

4.2. Goal Prediction Results

Table 2 presents the results of the temporal goal prediction models. In this formulation, the models output continuous predictions for home and away goals. These predicted scores were then converted into match outcomes using a decision margin: if the predicted home-goal advantage was greater than the margin, the match was classified as a home win (H); if the predicted away-goal advantage was greater than the margin, it was classified as an away win (A); otherwise, the match was classified as a draw (D). Therefore, the accuracy and macro F1 values reported in the table were computed from the derived H/D/A outcomes, while MAE and RMSE were computed from the continuous goal predictions.

The margin was treated as a post-processing hyperparameter selected during the initial training window.

Table 2. Results of goal prediction models.

Model	Margin	Acc.	F1 macro	MAE	RMSE
LightGBM	0.20	0.4724	0.3984	0.9546	1.2189
Poisson Reg.	0.25	0.4498	0.3655	0.9531	1.2187
Random Forest	0.35	0.4399	0.3723	0.9531	1.2185
XGBoost	0.35	0.4475	0.4042	0.9591	1.2252
Decision Tree	0.30	0.4394	0.3917	0.9833	1.2577

4.3. Fusion Strategy Results

Table 3 presents the binary fusion results obtained by combining models from the three predictive formulations. In the direct classification stage, the XGBoost and Gaussian Naive Bayes models trained with the `full` feature set were used. In the goal prediction stage, Poisson Regression and LightGBM were evaluated, while the Isolation Forest model was trained with the `odds_sofifa` feature set. All model outputs were converted to the binary H versus non-H representation before applying the fusion strategy.

Table 3. Fusion strategy results.

Model 1	Model 2	Model 3	Acc.	F1 macro
XGBoost full	Poisson Reg.	Isolation Forest	0.6101	0.6004
XGBoost full	Poisson Reg.	-	0.6281	0.6280
Gaussian NB full	Poisson Reg.	Isolation Forest	0.6191	0.6177
Gaussian NB full	Poisson Reg.	-	0.6550	0.6436
XGBoost full	LightGBM	Isolation Forest	0.6054	0.5954
XGBoost full	LightGBM	-	0.6242	0.6242
Gaussian NB full	LightGBM	Isolation Forest	0.6155	0.6136
Gaussian NB full	LightGBM	-	0.6507	0.6381

The results show that removing Isolation Forest consistently improved fusion performance. For instance, the combination of Gaussian Naive Bayes with Poisson Regression increased from 61.91% accuracy and 61.77% macro F1 with Isolation Forest to 65.50% accuracy and 64.36% macro F1 without it. A similar pattern was observed for the LightGBM-based combinations, indicating that the Isolation Forest predictions were not sufficiently complementary to those of the supervised models in the final binary decision.

Among the evaluated fusion strategies, the best result was achieved by combining Gaussian Naive Bayes trained on the `full` feature set with Poisson Regression. Although this produced a competitive binary predictor, it did not outperform the best individual binary classifier. Therefore, the main conclusion of the fusion experiment is not that simple voting improves performance, but that fusion gains depend on the complementarity and quality of the combined models. Since betting odds already concentrate a large portion of the predictive signal, additional model sources did not consistently provide enough marginal information to change the final decision. Moreover, the binary H versus non-H formulation enables the inclusion of Isolation Forest, but does not directly solve the original multiclass H/D/A problem. More advanced fusion approaches, such as weighted voting, stacking, or meta-learning, could better exploit complementary errors.

5. Conclusion

This paper investigated English Premier League match result prediction using complementary machine learning formulations, including direct multiclass classification, temporal goal forecasting, Isolation Forest-based binary anomaly detection, and simple binary decision-level fusion under an expanding-window walk-forward protocol. The results showed that feature sets containing betting odds generally provided the strongest predictive signal, while additional Sofifa attributes and rolling statistics did not consistently improve performance. Temporal goal forecasting provided useful numerical estimates, with Poisson Regression achieving the lowest overall MAE, although this did not necessarily translate into better final outcome reconstruction. The best fusion strategies were obtained without Isolation Forest; however, they still did not surpass the best individual binary classifier, suggesting that simple voting or consensus rules are limited when the combined models make similar errors. As limitations, the fusion stage was evaluated in a binary H versus non-H setting, while the original task is multiclass, and no statistical significance test was applied to compare small differences between models. Future work may explore additional leagues, alternative temporal features, different rolling-window sizes, statistical significance analysis, and more advanced fusion strategies, such as weighted voting, stacking, or meta-learning.

Declaration on Generative AI Use

Generative AI tools were used for linguistic revision, textual cohesion, and manuscript structuring support. All experiments, results, analyses, and final decisions were reviewed and validated by the authors.

Acknowledgements

This research was funded by CNPq (grant 307770/2025-7).

References

Berrar, D., Lopes, P., and Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 108:97–126.

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 (Applied Statistics)*, 46(2):265–280.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2):331–340.

Igiri, C. P. and Nwachukwu, E. O. (2014). An improved prediction system for football match result. *IOSR Journal of Engineering*, 4(12):12–20.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

Zhang, Q., Xu, H., Wei, L., and Zhou, L. (2019). Prediction of football match results based on model fusion. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, pages 198–202.