The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Predicting English Premier League Final Rankings Using Contextual Machine Learning and Monte Carlo Simulation

Rushil Reddy Yaragonda (author)[1]

[1]UIA International School of Tokyo, Tokyo, Japan

Abstract

Forecasting the final English Premier League (EPL) table is challenging due to football's low-scoring nature, outcome variability, and evolving team strength during the season. This research presents a pipeline that predicts match outcomes using contextual features and simulates the entire season to generate probabilistic final rankings, following the season-simulation tradition in football forecasting. Historical EPL match data from 2019 to 2024 are used to train probabilistic classifiers, specifically Random Forest and Gradient Boosting, implemented using scikit-learn, to estimate the probabilities of home win, draw, and away win. For the 2025–26 forecast, the model relies solely on the published fixture list and pre-season team context, including summer transfer-window net spending, without incorporating any match results from the target season. Match probabilities are converted into a distribution over final tables using Monte Carlo. The results include the average and median finishing positions for each club, as well as the probabilities of finishing in the top four and of relegation.

Keywords: Premier League, sports analytics, machine learning, Monte Carlo simulation, ranking prediction.

Introduction

Sports analytics has developed from a niche interest into a discipline that informs critical decisions in recruitment, tactical planning, and risk management. Football is central to this transformation due to its extensive data availability and commercial significance. Nevertheless, accurate forecasting in football remains difficult. The Premier League, in particular, serves as a rigorous testbed because of its competitive balance, diverse playing styles, and the influence of factors such as injuries and congested schedules on match outcomes.

Season-level prediction is addressed by considering both individual match outcomes and the forecast of the complete league table at the end of the season. This task is inherently more complex than match-level forecasting, as errors can accumulate over 38 games, newly promoted teams often lack sufficient top-flight data, and contextual factors can significantly influence outcomes beyond team strength. Methods that disregard variables such as recent form, squad availability, or home advantage often fail to produce realistic season simulations.

Previous research frequently employs statistical goal models, particularly Poisson-based approaches, due to their mathematical tractability and ability to simulate results via sampling from the expected goals (Gao, 2017; Louzada et al., 2014). While these methods serve as effective baselines, their reliance on distributional assumptions restricts their ability to capture the complexities of football [2]. Alternative studies have implemented machine learning techniques, such as artificial neural networks, to identify patterns in match statistics (Aka et al., 2021). Although these methods can enhance predictive accuracy, they often lack interpretability and may not explicitly incorporate contextual signals relevant to practitioners and analysts (Aka et al., 2021).

A practical compromise involves the use of tree-based machine learning models, specifically Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001), to predict match outcomes using contextual features. Monte Carlo simulations are then applied to the fixture list to generate a distribution of final league tables, following established season-simulation forecasting approaches (Gao, 2017; Louzada et al., 2014). The primary advantage of this approach is its flexibility, as tree-based models can incorporate variables such as recent form, home and away effects, and transfer spending without imposing rigid assumptions about goal distributions. The simulation process transforms match-level probabilities into season-level predictions, providing uncertainty estimates instead of a single deterministic outcome.

In addition to generating a ranked list, this approach addresses interpretable season-level questions, such as identifying clubs with the highest probability of finishing in the top four and those most at risk of relegation (Louzada et al., 2014). This probabilistic framing is valuable because football decisions are typically made under uncertainty, and probability distributions offer more actionable insights than single-point predictions.

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Background

Sports analytics is a rapidly expanding discipline with applications that include player performance evaluation and result forecasting across multiple sports. Predictive modelling is already integrated into decision-making processes in sports such as basketball and baseball. Football, however, presents unique challenges due to its low-scoring nature, high variability, and numerous contextual factors that influence outcomes. The English Premier League (EPL), recognised as one of the most competitive and widely followed leagues globally, provides an ideal yet demanding environment for predictive modelling. Prediction methods often use statistical models, such as the Poisson distribution, that assume goals are independent and identically distributed (Gao, 2017; Louzada et al., 2014). Although Poisson-based models provide a useful framework for season simulation, they are limited in their ability to incorporate essential contextual factors, such as player form, team momentum, injuries, and home/away differences (Gao, 2017; Louzada et al., 2014). These limitations constrain their flexibility and diminish their realism when applied to the complexities of football.

Other research has adopted machine learning techniques to better capture these dynamics. Neural network approaches can utilize comprehensive match statistics to learn non-linear relationships and achieve strong predictive performance (Aka et al., 2021). Hybrid methods that combine classical statistical modelling with additional contextual variables can further enhance realism (Louzada et al., 2014). However, these approaches involve trade-offs: neural networks may function as black boxes with limited interpretability (Aka et al., 2021), while regression-based models often remain constrained by rigid distributional assumptions (Gao, 2017; Louzada et al., 2014).

These considerations collectively highlight the need for methods that can address non-linear relationships, integrate contextual information, and maintain sufficient interpretability to justify predictions. The present work addresses this gap by introducing a season-ranking prediction pipeline that leverages contextual features with tree-based models such as Random Forest and Gradient Boosting (Breiman, 2001; Friedman, 2001), and converts match probabilities into comprehensive league-table forecasts through simulation (Gao, 2017; Louzada et al., 2014).

Method

The proposed system has four components: (1) data ingestion (historical match results, fixtures, and transfer spending), using publicly available EPL match archives and external transfer summaries (football-data.co.uk, n.d.), (Transfermarkt, n.d.), (2) feature engineering to create contextual pre-match variables, (3) a supervised learning model that outputs match outcome probabilities, and (4) a Monte Carlo simulation engine that samples match results across the full fixture list to generate a distribution of final league tables, following simulation-based season forecasting approaches (Gao, 2017; Louzada et al., 2014).

Each match is represented by a feature vector x and a categorical target $y \in \{H, D, A\}$ corresponding to home win, draw, or away win. The model learns $P(y \mid x)$. The season forecast is generated by simulating all fixtures $M = \{1, \ldots, 380\}$ and converting simulated results into points and ranks, consistent with common season simulation methods in football prediction (Gao, 2017; Louzada et al., 2014).

The core idea is to capture context that Poisson goal models do not naturally express (Gao, 2017; Louzada et al., 2014). A main feature is the recent points-based form over a rolling window w matches. For team *i*, define points from match t as $p_i(t) \in \{0,1,3\}$. Rolling form is defined as:

$$\mathrm{FormPts}_i(t) = \sum_{k=t-w}^{t-1} p_i(k).$$

For a fixture at time t between home team h and away team a, key derived features include:

$$\mathrm{HomeForm}(t) = \mathrm{FormPts}_h(t), \quad \mathrm{AwayForm}(t) = \mathrm{FormPts}_a(t)$$

$$\mathrm{FormDiff}(t) = \mathrm{HomeForm}(t) - \mathrm{AwayForm}(t)$$

Home advantage is represented as an indicator $I_{home}=1$. Transfer spending is incorporated as a pre-season team attribute. Let $\mathrm{NetSpend}_i$ denote net transfer spend for team *i*, sourced from transfer summaries (Transfermarkt, n.d.). For each fixture:

$$\mathrm{HomeNet} = \mathrm{NetSpend}_h, \quad \mathrm{AwayNet} = \mathrm{NetSpend}_a$$

$$\mathrm{SpendDiff} = \mathrm{HomeNet} - \mathrm{AwayNet}.$$

Two tree-based classifiers are trained and compared. In Random Forest (RF), class probabilities are averaged across T decision trees (Breiman, 2001):

$$P(y = c \mid x) = \frac{1}{T} \sum_{t=1}^{T} P_t(y = c \mid x),$$

where $c \in \{H, D, A\}$. In Gradient Boosting (GB), an additive model is learned and converted into probabilities via SoftMax (Friedman, 2001):

$$F_c(x) = \sum_{m=1}^{M} \gamma_{m,c}\, f_m(x), \qquad P(y = c \mid x) = \frac{e^{F_c(x)}}{\sum_{c'} e^{F_{c'}(x)}}.$$

Given a fixture list for the target season, the model produces a probability vector for each match j:

$$\pi_j = (P(H \mid x_j), P(D \mid x_j), P(A \mid x_j))$$

A simulated result is sampled as:

$$y_j \sim \text{Categorical}(\pi_j).$$

Points are then assigned:

$$\text{PtsHome}(y_j) = \begin{cases} 3 & y_j = H \\ 1 & y_j = D \\ 0 & y_j = A \end{cases} \quad \text{PtsAway}(y_j) = \begin{cases} 0 & y_j = H \\ 1 & y_j = D \\ 3 & y_j = A \end{cases}$$

When multiple teams finish a simulated season with the same number of total points, a tie-breaking mechanism is required to assign final ranks. In this study, ties are resolved via a random draw rather than goal difference, as the current match-level classifiers focus on outcome probabilities (H, D, A) rather than specific scorelines. While this differs from the official Premier League tie-breaking criteria, it ensures that the simulation remains grounded in the probabilistic outputs of the tree-based models without introducing external assumptions about scoring margins that the models were not explicitly trained to predict.

After all fixtures are simulated, teams are ranked by total points. Repeating this process S times yields an empirical distribution over final ranks, from which mean rank, median rank, top 4 probability, and relegation probability are estimated (Gao, 2017; Louzada et al., 2014).

Evaluation/validation

Historical EPL match data from seasons 2019/20 onwards is used for training and validation, with strict season-by-season separation to prevent leakage, using publicly available EPL match archives (football-data.co.uk, n.d.). Pre-processing includes consistent team naming across sources (results, fixtures, transfers) and construction of rolling features using only matches available prior to each fixture. Transfer spending is treated as a season-level covariate and merged into each team's match record using transfer summaries (Transfermarkt, n.d.).

Models are evaluated on a held-out season using standard probabilistic classification metrics. Accuracy is reported for interpretability, but probabilistic metrics are prioritized because the simulation step uses calibrated probabilities and the final objective is season-level simulation rather than single-match classification. Furthermore, to fulfil the objective of maintaining model interpretability, feature importance rankings are extracted from the Random Forest and Gradient Boosting models. This allows for a quantitative assessment of how heavily the pipeline weighs contextual signals, such as summer transfer spending and rolling form, relative to historical performance when predicting match outcomes. The main metrics are log loss and Brier score (Hastie et al., 2009):

$$\mathrm{LL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c}\mathbf{1}(y_i = c)\log P(y_i = c \mid x_i), \qquad \mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c}\big(\mathbf{1}(y_i = c) - P(y_i = c \mid x_i)\big)^2.$$

To validate the season ranking prediction, the simulation engine is run on the held-out season fixtures, producing a distribution of final tables using the season-simulation framework described in prior EPL forecasting work (Gao, 2017; Louzada et al., 2014). The predicted table is compared with the actual final standings using Spearman's rank correlation, mean absolute rank error, and the overlap between predicted high-probability teams and realised top four and relegation outcomes (Louzada et al., 2014).

Using the selected model and S=3000 simulations (seeded for reproducibility), the system produces the following forecast summary:

Table 1

Predicted premier league table

Position	Club	Average rank	Median rank	Probability (top 4)	Probability (relegation)
1	Arsenal	2.354000	2.0	0.895667	0.000000
2	Manchester City	2.779000	2.0	0.839333	0.000000
3	Liverpool	3.886000	4.0	0.666333	0.000000
4	Chelsea	4.518333	4.0	0.565333	0.000000
5	Aston Villa	5.594333	5.0	0.383333	0.001333

6	Newcastle United	6.784667	6.0	0.262000	0.002333
7	Manchester United	9.112333	9.0	0.096667	0.011333
8	Everton	9.293667	9.0	0.082333	0.012000
9	Crystal Palace	9.560333	9.0	0.069667	0.017333
10	Bournemouth	10.256667	10.0	0.044000	0.020333
11	Brentford	10.860000	11.0	0.039667	0.031000
12	Wolverhampton Wanderers	11.464333	12.0	0.030000	0.046667
13	West Ham United	11.792000	12.0	0.015000	0.052333
14	Brighton and Hove Albion	12.772333	13.0	0.004333	0.058333
15	Fulham	13.555667	14.0	0.005333	0.115000
16	Nottingham Forest	15.641667	16.0	0.000667	0.289667
17	Sunderland	16.277333	17.0	0.000333	0.365000
18	Tottenham	17.621000	18.0	0.000000	0.575000
19	Burnley	17.722333	18.0	0.000000	0.597000
20	Leeds United	19.163000	19.0	0.000000	0.816667

Related Work

Statistical goal models commonly use Poisson and Poisson-regression variants to model goals, then simulate match results and league tables (Gao, 2017; Louzada et al., 2014). The appeal is interpretability and a clean generative story, but these models can be constrained by assumptions about independence and stable scoring rates [2].

Machine learning approaches for football prediction have been explored to better capture non-linear effects, including work using neural networks to estimate rankings based on match statistics (Aka et al., 2021). These models can capture complex patterns but may require careful feature design and can be harder to interpret (Aka et al., 2021).

Hybrid approaches augment statistical baselines with contextual variables such as home advantage and indicators of team instability. This highlights the importance of context, but it often still inherits rigid modelling choices from the underlying statistical framework (Louzada et al., 2014). The approach in this paper retains the season-simulation idea while replacing strong distributional assumptions with ensemble-based contextual outcome classification using Random Forest and Gradient Boosting (Breiman, 2001; Friedman, 2001).

Conclusion

This paper introduced a pipeline for predicting Premier League season rankings by combining contextual, pre-match features with tree-based machine learning models and Monte Carlo season simulation (Gao, 2017; Louzada et al., 2014). Rather than relying on Poisson goal assumptions, the approach learns match outcome probabilities from signals such as recent points form, home advantage, and transfer spending (Gao, 2017; Transfermarkt, n.d.), and then converts those probabilities into full-table forecasts with uncertainty estimates through repeated season simulation (Gao, 2017; Louzada et al., 2014). The output includes not only a predicted final ranking but also actionable probabilities such as top four and relegation risk, consistent with season-level probability forecasting in prior EPL work (Louzada et al., 2014). Future work may extend this framework by incorporating richer squad-level features, such as injuries, minutes played, and expected goals, and by improving tie-breaking realism through joint modelling of goals and outcomes.

References

Aka, H., Aktuğ, Z. B., & Kılıç, F. (2021). Estimating the England Premier League ranking with artificial neural network. Applied Artificial Intelligence, 35(5), 393–402.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

football-data.co.uk. (n.d.). England: Premier League (E0) results and odds datasets (CSV archives). Retrieved September 11, 2025, from https://www.football-data.co.uk/englandm.php

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

Gao, J. (2017). Predicting Premier League final points and rank using Poisson regression and simulation (Project report). [Institution/University if known].

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Louzada, F., Suzuki, A. K., & Salasar, L. E. B. (2014). Predicting match outcomes in the English Premier League: Which will be the final rank? Journal of Data Science, 12(2), 235–254.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Transfermarkt. (n.d.). Premier League club transfer spending / transfer balance (club totals by season). Retrieved September 11, 2025, from https://www.transfermarkt.com