MACHINE LEARNING AND SPORTS ANALYTICS

A Reproducible and Explainable Machine Learning Framework for Football Player Performance Prediction Using Historical Data

Shivendra Singh Gour, Umesh Kumar

Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, India

Published November 2026

Abstract

The prediction of football players' performance is a significant machine learning problem since clubs, analysts and players engaged in fantasy leagues have to evaluate how a player will perform before a game. However, the usefulness of such a model is determined not only by its predictive accuracy on a certain data set, but also by the question whether its features are based only on historical data, whether its predictions take into account positions of players and whether they can be interpreted by the coaching staff and analysts. The review of literature summarizes the research achievements in the area of predicting score goals, expected goals models, estimation of fantasy points, outcome forecasting of matches, assessment of players, workload-related injury predictions. However, all these aspects in the literature work independently, while the current paper offers a comprehensive, reproducible and understandable framework for predicting players' performance based on past experience This research relies on statistics from participating players in various seasons obtained from publicly available football databases. The study consists of data preprocessing techniques, exploratory data analysis, lag feature engineering methods, the process of target skew and position differences, and finally, the application of the models. Linear Regression, Decision Tree, Random Forest and XGBoost regressors are trained, compared and discussed in terms of MAE, RMSE and R^2, and, for the best-performing model, player-level explanations are produced with SHAP values. A method called chronological train-test split was used in which future seasons could not get into training. Among various things highlighted in the literature review was the issue with the size of players' samples being low, while the desirable outcomes like goals had a large number of zeros. Also, it was marked that usage of random splits can lead to overestimation of accuracy. However, since the computational part of the project, which would yield the numerical results for the metrics, is not included in the provided materials, this study does not make any claims regarding the results. Instead, it describes the evaluation protocol and the format of the results that will be obtained and used for the preparation of the final report.

Keywords: Football analytics; player performance prediction; machine learning; historical data; XGBoost; Random Forest; SHAP; explainable AI; sports analytics

1 Introduction

Player performance prediction refers to estimating how a footballer will perform in a future match or season from what is known about the player's past. Typical targets include goals, assists, goal contributions per 90 minutes, match ratings or fantasy-football points. These forecasts support recruitment, squad planning, contract decisions and fantasy-football selection, so the prediction problem can be framed as a supervised regression task in which historical statistics are used to estimate a numerical performance outcome.

Machine learning has become a primary approach to this problem because player records combine many correlated variables such as minutes played, shots, passes, defensive actions, expected goals, age and position. Linear models provide an informative baseline, while tree-based models are able to capture non-linear interactions between variables, for example between age and minutes played. Ensemble methods such as Random Forest and gradient boosting are generally highly attractive for structured/tabular data. Research has also progressed towards expected-goals modelling, position-specific forecasting, player valuation and explainable artificial intelligence.

The key practical challenge is that football data is small, skewed and time-ordered. Markopoulou, Papageorgiou and Tjortjis forecast goals for players across four top European leagues, but had to restrict their sample to players present in all six seasons, which left roughly 100 players per league, and some of their models showed signs of overfitting. Their results also showed that previous goals and expected goals were among the strongest predictors [1].

In a similar vein, Huang and Zhang compared six ensemble regressors on about 12,000 players and found that gradient-boosted trees gave the best test R^2 of 0.901 when predicting market value, while SHAP showed which player attributes drove the output [5]. These insights underline the present study's choice to treat time-aware validation and explanation as key components rather than after-thoughts.

1.1 Research objectives

1. To develop a consistent machine-learning pipeline for predicting a football player's success from data.
2. To compare the work of the Linear Regression, Decision Trees, Random Forests and XGBoost qualities.
3. To measure how target skewness and positions affect the errors in predictions and use proper ways of dealing with them.
4. To employ MAE, RMSE and R^2 metrics for the evaluation of the models working under chronological (season-based) splitting.
5. To choose the best performing model regarding the prediction accuracy and ability to work with various playing positions.
6. To utilize SHAP-based explanations to discover global and player-level drivers of predicted performance.

1.2 Research questions

1. RQ1: How do commonly-used regression algorithms differ in predicting next-season player performance from historical statistics?
2. RQ2: How do target skew and differences between playing positions affect prediction error, and which treatments help?
3. RQ3: Which evaluation metrics and validation schemes offer a more honest perspective on forecasting accuracy than a random split?
4. RQ4: Which historical features (for example previous output, minutes played, age and expected goals) make the largest contribution to the model's predictions?

2 Literature review

The literature search for the topic of football player performance prediction has focused on peer-reviewed journals and recent preprints covering goal-scoring forecasts, expected-goals models, fantasy-points forecasting, match-outcome prediction, player valuation and injury modelling. Table 1 summarises eight studies that were selected for their relevance to the proposed framework.

Table 1. Selected journal studies and their relevance to the proposed framework.

Study	Journal	Approach	Key finding	Limitation	Relevance
[1] Markopoulou et al. (2024)	Machine Learning and Knowledge Extraction	LR, ridge, RF, GB, XGBoost, MLP; four leagues, six seasons	XGBoost best in most scenarios; previous goals and xG most important	Small samples; some overfitting	Season-lag player regression
[2] Groos (2025)	arXiv preprint	Position-specific XGBoost + RF ensembles for fantasy points	Comparable to a commercial forecaster; 5–34% lower RMSE than last-5 baseline	No team news or minutes forecast; short test window	Position-wise rolling features
[3] Hewitt & Karakuş (2023)	Franklin Open	LR and gradient boosting xG on 15,575 shots	Position-adjusted and player-adjusted xG models	Shot-level only; one event-data source	Shot-quality features
[4] Beal et al. (2021)	AAAI Conference	RF combining text, Dixon–Coles and bookmaker odds	63.19% accuracy vs 59.11% for Dixon–Coles	One news source; dataset not exhaustive	Match-level ensemble prediction
[5] Huang & Zhang (2023)	arXiv preprint	Boruta selection; six regressors; SHAP on ≈12,000 players	GBDT best with test R^2 of 0.901	Underestimates superstar values; public data less detailed	Tree ensembles + SHAP
[6] van Arem et al. (2025)	Applied Sciences	Explainable ML on player histories; one-year-ahead forecasts	Random forest most suitable; built-in uncertainty estimates	One-year horizon only	History-to-future forecasting
[7] Freitas et al. (2025)	PLoS ONE	SVM, neural network, AdaBoost on GPS data of 34 players	SVM reached AUC 0.85 for noncontact injury	Only 18 injuries; one team, one season	Small-sample risk
[8] Saberisani et al. (2025)	Frontiers in Sports and Active Living	Decision tree on GPS workload ratios, 25 players	Decelerations gave highest AUC (0.91); none passed the 0.95 threshold	One team, one model, scarce injuries	Workload variables

2.1 Base papers for the present study

There are two pieces of research that serve as the primary methodological basis for the present study. First, Markopoulou et al [1]. provide an example of a complete player-level forecasting workflow, which includes season-lag features, several regression algorithms, grid-search tuning and error metrics across four leagues. On the other hand, Huang and Zhang [5] provide a valuable foundation for integrating tree ensembles with feature selection and interpretation based on SHAP in the data on players. The ongoing research utilizes the empirical multi-model comparison of the previous research with the focus on the interpretability of the latter along with a strictly chronological validation design.

2.2 Thematic synthesis

There are three key points to note in the literature analyzed. There is a significant correlation between the predictive capacity and the player history characteristics: previous performance, expected goals, and rolling averages are identified to be the most beneficial predictors [1,2]. Secondly, tree-based machine-learning methods such as XGBoost, random forests, and gradient boosting tend to yield good performance on the player tabulated data, though the leading method varies across leagues, positions, and goals [1,2,5,6]. Thirdly, interpretability is becoming extremely important since analysts want to understand which characteristics impact players' predications despite the actual number presented [5,6]. Hence, the analyzed literature supports an integrative approach that combines historical feature engineering, multi-model benchmarking, and interpretability. The proposed study is intentionally very practical one: it does not seek to replace specialized research on event-level tracking data, injury risk or market valuation, but rather establishes a reproducible baseline that can be built upon later.

2.3 Research gap

There is plenty of research mentioned in the literature on goal prediction, expected goals, fantasy points and match result prediction, valuation and injury risk. However, these areas are mainly discussed separately in limited availability studies limited to several or only one league/season. Small sample sizes are addressed in study [7] and study [8], which cover injury cases from one team and several dozens of players, and also in study [1], which analyzes seasonal stats of those players only. In order to offer something useful to an undergraduate researcher, a pipeline for public multi-season player statistics needs to be created that would validate forecasts in time and explain predictions individually.

As such, the contribution of this work is framed around providing an end-to-end solution that can be used as a reference for similar tasks. This approach does not claim to offer groundbreaking novelties in each of the aspects but focuses on combining them into a coherent whole. The contribution is showcased through a combination of preprocessing, EDA, lag-based feature engineering, handling of target skew and positional differences, comparison of ML performance, and SHAP explanations.

3 Research methodology

3.1 Research design

This study will adopt the design of an experimental machine learning analysis. The main objective of my research is to perform regression to predict a player's performance in the next season from historical statistics. In particular, I will use multi-season player statistics from public football data sources such as Sports Reference, which was also the source in [1], as the initial data for investigation. The models will be implemented in Python with scikit-learn [13]. I will ensure that the data preprocessing steps are fitted on the training seasons and applied to the held-out season to avoid leaking information.

3.2 Dataset

The dataset will consist of player-season records from the top European leagues, including appearances, minutes played, goals, assists, shots, passes, defensive actions, expected goals (xG), age and position. The target variable is the player's performance in the following season, defined as goal

contributions (goals plus assists) per 90 minutes, or goals for goal-focused experiments. The final source will be a public database; however, the paper does not include a local copy of the raw files. This is why the statistical description of the records, including the number of players after data cleaning, missing values and the target distribution, is not invented here.

3.3 Data preprocessing

- Examine data types, missing values, duplicates, and inconsistencies in player names and season labels.
- Encode categorical variables such as position and league into numerical representations.
- Keep only players with sufficient minutes played, and deal with missing or invalid entries (for example, unavailable xG in early seasons).
- Split variables into predictors (historical seasons) and the target (next-season performance).
- Split the data chronologically, training on earlier seasons and testing on the most recent season, so that no future information is used.
- Fit scalers and encoders on the training seasons only and use them to transform the test season. Do not touch the test set.

3.4 Exploratory data analysis

EDA will explore the distribution of the target and its relationship with age, minutes played, position, previous-season output, shots and expected goals. The emphasis will be on information that can be derived from plots rather than information that could be incorrectly inferred. EDA will also address data-quality issues, the skew towards zero in goal-based targets, and possible feature engineering.

3.5 Feature engineering

- Encoding of categorical variables such as position and league
- Lag features from the previous one to three seasons, and per-90-minute rates instead of raw totals
- Retention of essential continuous variables such as age, minutes played and expected goals
- Creation of easily interpretable derived variables (for example, goal contributions per 90) only when having a meaningful analytic interpretation
- Elimination of identifiers and checking for highly correlated features where required

3.6 Target skew and position handling

Markopoulou et al. recognized that there should also be an evaluation of a particular subset of elite goalscorers [1]. The first study involves scores for each subject in the sample. For the second condition, it introduces a mitigating factor making use of position-specific models, minimal minute filter, or the transformed variable but applies it only to the training sample. Depending on the empirical analysis that will exclude the assumptions about superior performance of one of the methods, the procedure will be selected.

3.7 Machine learning models

Table 2 presents the machine learning models included in the proposed experiment.

Table 2. Machine learning models included in the proposed experiment.

Model	Role	Reason for inclusion
Linear Regression	Linear baseline	Interpretable benchmark for next-season output
Decision Tree	Non-linear rule-based regressor	Captures interactions in an interpretable structure
Random Forest	Bagged tree ensemble	Robust tabular baseline with non-linear capacity [10]
XGBoost	Gradient-boosted trees	Strong tabular benchmark and leading candidate [9]

3.8 Evaluation metrics

Measurement of accuracy cannot be achieved in a regression problem, so the evaluation process is based on metrics that give error proportions. Mean Absolute Error (MAE) gives a mean error in the target's units whereas Root Mean Squared Error (RMSE) works as a penalty for a substantial miss, for example, a star player having a very unexpected season. R^2 indicates how much variance in the target can be explained by the model compared to a very simple prediction model. Reporting how MAE compares to RMSE is also useful since the values do not necessarily go in the same direction. The results are also given in terms of positions so that good overall results are not misleading about the relatively poor forecasts for a certain position. The naive approach is taken into consideration, too, in order to be able to compare the results.

The metrics are defined in Equations (1)–(3). Lower MAE and RMSE and higher R^2 indicate better forecasts. Because the test season is later than all training seasons, these scores estimate forward-looking accuracy rather than in-sample fit.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where y_i is the observed value, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean of the observed values and n the number of test records.

3.9 Explainability with SHAP

SHAP (SHapley Additive exPlanations) [11] will be used to interpret the results from the final selected model. The global SHAP summary can help prioritize features in terms of their contribution to the predicted performance for the test season, while the local SHAP explanation can provide insights into the model behavior for a specific player, such as a forward whose forecast is much higher or lower than his actual output. The interpretations will differentiate between the model's association and a causal relationship: a feature with a high SHAP value indicates that the feature is highly associated with the model output but does not necessarily mean that changing this feature causes a change in a player's performance.

3.10 Proposed system flow

The proposed system flow is: raw data → preprocessing → EDA → lag feature engineering → chronological split → skew/position handling → LR / DT / RF / XGBoost → MAE / RMSE / R^2 → best model → SHAP → performance insight.

4 Results and discussion

4.1 Status of experimental results

This paper does not contain any executed experiments. Some figures from the literature are useful as context only: for example, Markopoulou et al. report a best MAE of 1.29 goals for XGBoost on Serie A data [1], and Huang and Zhang report a test R^2 of 0.901 for market-value prediction [5]. These values belong to different datasets and targets, so they are expectations for orientation, not results, and cannot be compared directly with this study. The actual result should be represented by the error scores that can only be obtained by running the proposed pipeline on the selected dataset.

4.2 Result reporting framework

Table 3 presents the final experimental comparison table to be populated from the executed model.

Table 3. Final experimental comparison table to be populated from the executed model.

Model	MAE	RMSE	R^2	MAE (Forwards)	MAE (Midfielders)	MAE (Defenders)
Logistic Regression	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
Decision Tree	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
Random Forest	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
XGBoost	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured

4.3 Expected analytical interpretation

According to what has been written previously, tree ensembles and boosted trees are good algorithms when dealing with tabular data of players, but final results may depend on multiple factors related to the league, player position and feature set [1,6]. Thus, a model that has higher performance than a naive baseline defined by the performance of the previous season (although maybe by a small margin) can still be helpful whereas a model that simply follows the naive baseline may not be beneficial at all. A small MAE value with a bad R^2 score may also indicate that the model predicts most average players well. Also, the cross-study analysis demonstrates the importance of representation and validation of data. On the one hand, Markopoulou et al. highlight that previous goals and expected goals are one of the most powerful predictors, while Beal et al. prove that combining textual data with Dixon–Coles and bookmaker predictions proves more productive than any of the methods used alone (63.19% accuracy vs. 59.11% for Dixon-Coles alone) [1,4,12]. Therefore, individual studies should include a complete description of the research design rather than simply reporting the error score.

4.4 Explainability analysis

Once the final model has been chosen, the SHAP analysis will indicate which features have had the most substantial average influence on the forecast relative performance and how their values either positively or negatively affect the predicted values. For instance, if the previous-year performance and expected goals have been credited with the most significant rankings, which is the case in Oenning (2022), this will align well with the existing literature. Local analysis results will be presented for a number of players whose predictions produced good results as well as for a number of players whose predictions were far from the reality. Local explanation refers to the model explanation rather than football causation.

4.5 Practical interpretation

The performance forecast model is evaluated based on its efficiency in decision-making. In recruiting players and forming a team, forecasts should be interpreted with their possible error in mind, and lists of players should reflect this by presenting them as a shortlist instead of a rating. While applying the model to fantasy football situations, as in [2], the advantage comes from spotting high-yielding players where forecasting error is the biggest. A position-by-position outcome can be used, along with SHAP explanations, to evaluate the forecast reliability.

4.6 Limitations

- The analysis done at the league level is not as extensive as the tracking and event data of clubs, and might not cover all leagues.
- Information at the season level does not capture the form, injury situation, tactical modifications or transfers between certain matches that have an impact on performance next season.
- The effectiveness of the model is determined by the preprocessing of data, filters for a minimum length of play, the features selected for modelling, and hyper-parameter optimization.
- SHAP values explain model behavior, but should not be interpreted as causal effects on player performance.
- Conclusions based on final, empirical analysis cannot be drawn until the pipeline is executed and results are reported.

5 Conclusion and future scope

5.1 Conclusion

This paper provides an explainable and reproducible framework for football player performance prediction using machine learning and historical data. It combines data preprocessing, lag-based feature engineering, chronological validation, handling of target skew and positional differences, comparison of Linear Regression, Decision Tree, Random Forest and XGBoost regressors, and SHAP-based explanation. The framework is grounded in studies on goal forecasting, fantasy-points prediction, player valuation and expected-goals modelling.

The main contribution of this research is therefore an integrated and transparent baseline rather than a set of reported accuracy figures. Numerical conclusions will be drawn only after the pipeline is executed on the selected dataset and the results are reported together with the full evaluation protocol.

5.2 Future scope

- Add match-level and event or tracking data to capture form within a season.
- Compare position-specific models with a single global model.
- Add prediction intervals or uncertainty estimates for individual forecasts, as in [6].
- Include injury and workload information where available [7,8].
- Test additional algorithms such as LightGBM, CatBoost and neural networks.
- Extend the framework to additional leagues and to women's football.
- Design a Streamlit-based interface for analysts and scouts to explore forecasts and explanations.
- Study the concepts of model drift and retraining as new seasons are added.

References

[1] Markopoulou C, Papageorgiou G, Tjortjis C. Diverse machine learning for forecasting goal-scoring likelihood in elite football leagues. *Mach Learn Knowl Extr*. 2024;6(3):1762–1781. https://doi.org/10.3390/make6030086

[2] Groos D. OpenFPL: An open-source forecasting method rivaling state-of-the-art Fantasy Premier League services. *arXiv*:2508.09992 [Preprint]. 2025.

[3] Hewitt JH, Karakuş O. A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*. 2023;4:100034. https://doi.org/10.1016/j.fraope.2023.100034

[4] Beal R, Middleton SE, Norman TJ, Ramchurn SD. Combining machine learning and human experts to predict match outcomes in football: A baseline model. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2021. (arXiv:2012.04380)

[5] Huang C, Zhang S. Explainable artificial intelligence model for identifying market value in professional soccer players. *arXiv*:2311.04599 [Preprint]. 2023.

[6] van Arem KW, Goes-Smit F, Söhl J. Forecasting the future development in quality and value of professional football players. *Appl Sci*. 2025;15:8916. (arXiv:2502.07528)

[7] Freitas DN, et al. Predicting noncontact injuries of professional football players using machine learning. *PLoS One*. 2025;20(1):e0315481. https://doi.org/10.1371/journal.pone.0315481

[8] Saberisani R, et al. Prediction of football injuries using GPS-based data in Iranian professional football players: A machine learning approach. *Front Sports Act Living*. 2025;7:1425180. https://doi.org/10.3389/fspor.2025.1425180

[9] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785–794.

[10] Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.

[11] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30*; 2017. p. 4765–4774.

[12] Dixon MJ, Coles SG. Modelling association football scores and inefficiencies in the football betting market. *J R Stat Soc Ser C Appl Stat*. 1997;46(2):265–280.

[13] Pedregosa F, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.