

Comparative Analysis of Machine Learning Models for Predicting Top Goal Scorer and Goalkeeper Performance in Football

Tanish Sharma
Department of CSE
AUUP.
tanishsharma2003@gmail.com

Parv Bagga
Department of CSE
AUUP
parvbunny@gmail.com

Kinshuk Ahuja
Department of CSE
AUUP
kinshukahuja20@gmail.com

Seema Sharma
Department of CSE
AUUP
seema.modgil@gmail.com

Abstract— The use of football analytics has transformed how players are evaluated, strategies are developed, and recruitment processes are planned. Machine learning (ML) has become a vital tool in analyzing player performance, especially in identifying leading goal scorers and highly effective goalkeepers. This research focuses on understanding how team statistics influence goals and saves, examining not only individual player metrics but also the relationship between team attacks, team goals, and the impact of each player's threatening attacks on their goal-scoring output. Additionally, metrics related to goals conceded by the team were analyzed to establish any correlation with the number of saves made by goalkeepers. The study aims to predict the top goal scorer and the best goalkeeper using seven ML models: Random Forest, XG Boost, Linear Regression, K-Nearest Neighbor (KNN), Decision Trees, Support Vector Regression (SVR), and Gradient Boosting. Data was gathered through web scraping and API integration from five major leagues: La Liga, Bundesliga, Premier League, Ligue 1, and Serie A covering the seasons from 2017/2018 to 2022/2023. The models' performance was assessed using statistical measures like Mean Squared Error (MSE), Mean Absolute Error (MAE), R-Squared (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Findings indicate that XG Boost performed best for predicting goal scorers, while Linear Regression was most effective for forecasting goalkeeper saves. This research enhances decision-making for managers and analysts by providing data-driven insights that improve strategic planning and player evaluation.

Keywords—Machine Learning Models, Performance Metrics, API, Web Scraping, Sports Analytics, Z-Score Normalization, Encoding, Feature Engineering, Hyperparameter Optimization

I. INTRODUCTION

Soccer is a widely played sport globally, with experts continuously analyzing millions of players to enhance their performance. Earlier assessments of footballers were based on goals, assists, and clean sheets however these methods were shallow and would not be reflective of the true value of the player.

Research in modern football analytics utilizes advanced methodologies for data-driven performance evaluations through stats like expected goals, assists, and other defensive measures. The exponential growth of football data has given clubs, analysts, and league managers the

motivation to utilize machine learning for predicting future performances, optimizing team tactics, and scouting potential players.

II. PROBLEM STATEMENT

Though the usage of machine learning model for player performance assessment has already been used but there is a need to compare multiple Machine Learning models for both goal-scoring ability and goalkeeper performance. An example of this is the many questions that need to be addressed, such as:

How would the model's performance be affected if it prioritized team metrics over individual metrics?

Which ML model gave the best prediction for goal scorers?

How do different machine learning models perform in predicting goalkeeper performance?

What preprocessing steps would enhance model performance?

III. LITERATURE REVIEW

Our study focuses on two main modules: predicting the top goal scorer and evaluating the best goalkeeper. From the literature it has been observed that there is a significant gap for the data collection methods, coverage of seasons of various leagues and the prediction models that predict the top goal scorer and best goalkeeper. Player position, shot conversions, and other offensive stats are the main emphasis of top goal scorer forecasts [17]. Numerous research have used ensemble machine learning models for this purpose, like SVM, which is ideal for capturing complex relationships between player attributes including form, fitness, and game context [9]. The author focused on different player attributes and game situations for better understanding of offensive play but did not predict the best goalkeeper performance.

On the other hand, determining the best goalkeeper necessitates a different methodology, which includes examining defensive metrics like save percentage, goals allowed, and reaction time in addition to taking shot distance, angle, and defensive pressure into consideration [13].

Different feature extraction techniques are required since the data in this module must evaluate a goalkeeper's capacity to react in high-pressure situations and limit scoring possibilities. The other author captured the ambiguity and unpredictability of goalkeeper performance under various match conditions using Bayesian Networks (BN) and Ensemble Methods are frequently employed in this field [11]. But the author did not predict the top goal scorer performance.

Although machine learning models are necessary for both modules, the selection of models is characterized using properties of the available data. The authors emphasized on offensive measures for goal scorer forecasts. The authors used the ML models like SVM and XGBoost because of their superior predictive performance and capacity to handle complicated, high-dimensional data [9]. However, here the author used Bayesian networks and regression models for goalkeeper performance evaluation. The author concluded that these models are good at handling the probabilistic aspect of goalkeeper performance, especially when saving shots [15].

Additionally, the significance of statistical modeling in sports forecasting extends to player performance and goal-scoring probability, where machine learning techniques are essential. According to recent studies, various ML models, including those built on tree-based algorithms, have demonstrated efficacy in predicting player ranks and goal-scoring probabilities [16]. These models can be more effective in assessing player performance by utilizing machine learning, providing insightful information about player performance that would be difficult for conventional approaches to obtain.

In conclusion, there are notable differences between the two modules best goalkeeper rating and top goal scorer prediction in terms of feature extraction, data gathering, and ML techniques. Bayesian Networks and ensemble methods are excellent at capturing goalkeeper performance under various circumstances, whereas models such as SVM and XGBoost are good at forecasting offensive performance [5]. This comparative study emphasizes how important customized machine learning methods are for performance analysis and sports prediction [2].

IV. METHODOLOGY

ML models are utilized to forecast the top goal scorer and the best goalkeeper for a given competition. Figure 1 shows flowchart of the given work. It includes data gathering, pre processing, model training, and evaluation using suitable performance metrics.

V. DATA COLLECTION

To construct an optimal machine-learning model, this research used two different sources from which player and team performance data was collected. The dataset includes historical player statistics, such as goal-scoring metrics and goalkeeper performance indicators [12]. Data were also collected through web scraping techniques as well as API integration across five leagues from the seasons of 2017/2018 until 2022/2023 [8]. The different data collection methods used for the proposed model are discussed below:

1) Sportsmonks Football API: Sportsmonks provides reliable and detailed statistics for players and teams. In our case, the relevant dataset was constituted by 1468 instances and 14 features for different players. The data was collected or the players based on maximum goals, moderate goals, and least goals in a single season. As shown in TABLE I, the considered feature introduced diversity into the data, such as players playing at different positions with varying game time, among others [1].

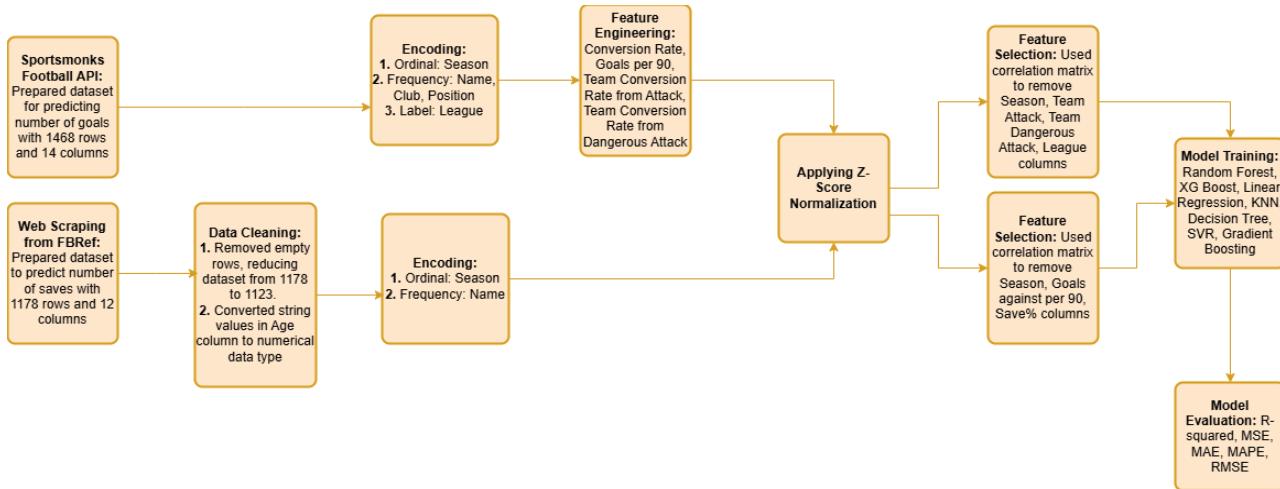
TABLE I. DATASET FEATURES FOR GOALS

Feature	Description
Season	The football season during which the data was recorded
Name	Player Name
Age	Player Age
Club	The football club they played for during that season
Position	Player position
Minutes Played	Total playing time of the player during a season in minutes
Goals	Number of goals in a season by a player
Shots	Total shots taken by a player in a season.
Team Attack	Total number of attacking plays made by a footballers team
Team Attack Average	Average number of attacking plays per match by a footballers team
Team Dangerous Attack	Total number of dangerous attacking plays made by a footballers team
Team Dangerous Attack Average	Average number of dangerous attacking plays per match by a footballers team
Team Goals	Goals scored by a footballers team during a season
League	Name of the league in which the player's club is competing

2) FBRef: It is a renowned organization known for offering comprehensive data coverage across the wide range of sports. Using this source, we built a dataset comprising 1178 entries. instances and 12 features that included players who saved the least goals, moderate goals, and maximum goals in a particular season. As shown in TABLE II, all features here also contribute to the diversity of the data.

TABLE II. DATASET FEATURES FOR SAVES

Feature	Description
Name	Full name of the goalkeeper
Age	Player Age
Minutes Played	Total playing time of the player during a season in minutes
Goals Against	Goals conceded in a season by a goalkeeper
Goals Against per 90	The average amount of goals conceded by a goalkeeper in 90 minutes
Shots on Target Against	Shots on target a goalkeeper faced
Saves	Saves a goalkeeper made during a season
Save%	Percentage of saves
Clean Sheets	Matches in which the goalkeeper conceded no goals
Clean Sheet%	Percentage of clean sheets
Penalty Kicks Saved	Total penalty kicks blocked by the goalkeeper
Season	The football season during which the data was recorded



3) *Feature Engineering:* For forecasting top goal scorer, in addition to extracting data from the API, four additional features were created for accurate predictions. This changed the tally of number of features for top goal scorer from 14 to 18 features. These features are discussed below:

a) *Conversion Rate:* The percentage of shots taken by the player that resulted in goals is given by equation 1.

$$\frac{\text{Goals}}{\text{Shots}} \times 100 \quad (1)$$

b) *Goals per 90:* Equation 2 gives the mean amount of goals scored by the player per 90 minutes of play.

$$\text{Goals}/(\text{Minutes Played}) \times 90 \quad (2)$$

c) *Team Conversion Rate from Attack:* The percentage of total team attacks that resulted in goals is given by equation 3.

$$\frac{\text{Team Goals}}{\text{Team Attack}} \times 100 \quad (3)$$

d) *Team Conversion Rate from Dangerous Attack:* Equation 4 gives the percentage of dangerous team attacks that resulted in goals.

$$\frac{\text{Team Goals}}{\text{Team Dangerous Attack}} \times 100 \quad (4)$$

VI. DATA PRE-PROCESSING

Data Cleaning: The top goal scorer dataset did not have any missing values but there were some missing values in the datasets related to goalkeepers. Given that the percentage of missing values was negligible, the individuals were dropped from the data matrix, leaving a total number of rows reduced from 1178 to 1123. The Age column also had some string values that were converted into numerical datatype.

Encoding: Since both the datasets contained text values it was important to encode them into numeric values using appropriate encoding techniques as shown in TABLE III and TABLE IV [18].

TABLE III. ENCODING FOR GOALS DATASET

Feature	Encoding Type
Season	Ordinal Encoding
Name	Frequency Encoding
Club	Frequency Encoding
Position	Frequency Encoding
League	Label Encoding

TABLE IV. ENCODING FOR SAVES DATASET

Feature	Encoding Type
Season	Ordinal Encoding
Name	Frequency Encoding

Normalization: Z-Score Normalization was used on the features of the datasets except for Season, Name, League, and Position in the goal scorer dataset, and the Season and Name in the goalkeeper dataset. This method normalizes the data by centering it around a mean of zero with a standard deviation of one. It ensures that all numeric columns are scaled uniformly, prohibiting any sole feature from dominating the model's learning process due to differences in magnitude. This standardization significantly improved model convergence and accuracy during training.

Feature Selection: A correlation matrix was constructed to eliminate features with minimal correlation to the response variables, thus retaining only the most influential attributes in training the models [3]. Least correlated columns were removed from the goal-scoring dataset like Season, Team Attack, Team Dangerous Attack, and League data had negligible effect on the number of goals scored. Similar to goalkeeper dataset, columns like Season, Goals Against per 90, and Save% were removed due to least correlation with blocks made by a goalkeeper. This reduced the amount of columns from 18 to 14 and from 12 to 9, respectively. This feature selection process helped in dimensionality reduction, improving model interpretability, and enhancing computational efficiency while properly capturing the predictive power.

VI. MODEL TRAINING

Seven ML models were trained for purposes of predicting the number of goals for players and saves made by the goalkeepers. This work compared Random Forest, XGBoost, Linear Regression, KNN, Decision Tree, SVR, and Gradient boosting [15]. Moreover, grid search CV was performed over all applicable algorithms for hyperparameter optimization as shown in TABLE V, thus deriving the extremely best combination of values applicable to each model. Grid search CV was also used with 5 folds for cross-validation. This ensured robust evaluation of the models.

TABLE V. BEST HYPERPARAMETERS

Model	Best Parameters (Goals)	Best Parameters (Saves)

Random Forest	Maximum Tree Depth : 20 Minimum Sample per Leaf : 1 Minimum Sample per Split : 2 Number of Estimator : 100	Maximum Tree Depth : None Minimum Sample per Leaf : 1 Minimum Sample per Split : 2 Number of Estimator : 200
XG Boost	Column Sample Rate per Tree : 1.0 Learning Rate : 0.1 Maximum Tree Depth : 3 Number of Estimator : 200 Sub-sample Ratio : 0.8	Column Sample Rate per Tree : 1.0 Learning Rate : 0.2 Maximum Tree Depth : 3 Number of Estimator : 200 Sub-sample Ratio : 0.8
KNN	Number of Neighbors : 7 Weighting Method: Distance	Number of Neighbors : 10 Weighting Method: Distance
Decision Tree	Maximum Tree Depth : 10 Minimum Samples per Leaf : 1 Minimum Samples per Split : 5	Maximum Tree Depth : 10 Minimum Samples per Leaf : 4 Minimum Samples per Split : 2

VII. EVALUATION METRICS

The model is evaluated via several statistical metrics like as MSE, MAE, Coefficient of Determination, RMSE, and MAPE. MSE determines mean squared differences between the actual and predicted values, thus allocating a bigger weight to larger inaccuracies. Hence, this statistic is more sensitive to significant deviations, yet helps evaluate model performance while granting large errors a larger impact on the final measure. MAE computes mean of the absolute error between actual and forecasted values. MAE has a simple interpretation of the mean magnitude of the error interpreted equally, allowing it to allow the researcher to assess importance when all deviations are considered equally important. The R-squared estimates the extent to which the non dependent variable explains variation in the dependent or target variable between zero and one. RMSE assesses the mean magnitude of the error between the actual and predicted values measured in the same unit as the dependent or target variable, penalizing bigger errors more than small ones, useful in measuring the accuracy of the model. MAPE provides the average amount by which a given prediction underestimates the target. MAPE provides a dimensionless measure of error, making its use useful in model comparisons across different datasets.

VIII. RESULTS AND DISCUSSIONS

A. Goal-Scoring

An analysis of the statistical performance metrics were carried for the seven different ML based models [15], shown in TABLE VI. Model evaluation is determined through five different performance metrics also shown in TABLE VI. Such measures provide a detailed assessment of every model regarding precision, reliability, and generalization capability towards predicting the top goal scorer.

TABLE VI. PERFORMANCE METRICS FOR PLAYER GOALS

Model	MSE	MAE	R-Squared	RMSE	MAPE (in %)
Random Forest	0.02	0.05	0.98	0.13	14.91
XG Boost	0.01	0.05	0.99	0.10	16.75
Linear Regression	0.20	0.34	0.80	0.45	90.59
KNN	0.12	0.22	0.89	0.34	62.37
Decision Tree	0.03	0.07	0.97	0.18	36.63

SVR	0.03	0.09	0.97	0.17	19.59
Gradient Boosting	0.01	0.06	0.99	0.12	20.82

The analysis from TABLE VI and Figure 2 indicates that some algorithms like ensemble methods Random Forest, XG Boost, and Gradient Boosting [15] fit the underlying patterns of the dataset well .

Among the various models evaluated, XG Boost was able to perform the best on this task, balancing predictive accuracy and error reduction. It was able to generalize well and performed consistently across metrics. Random Forest, and Gradient Boosting gave equally good results and, therefore, may be good alternatives for the task [6]. On the contrary, models like Linear Regression struggled with predictability, making them unsuitable for handling complex datasets.

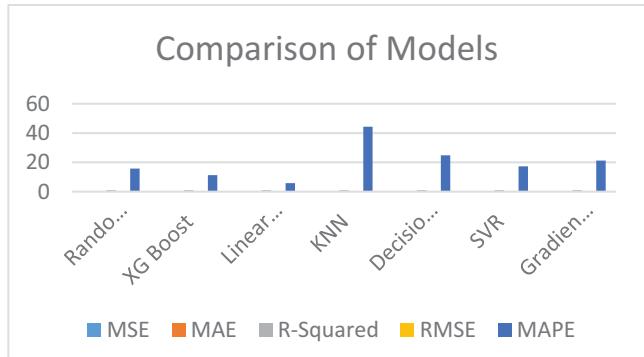
B. Goalkeeper Saves

Seven machine learning models, specifically Random Forest, XGBoost, linear regression, KNN, Decision Tree, SVM, and Gradient Boosting [15], were trained for the prediction of goalkeeper saves. The results were analyzed with respect to standard performance metrics, namely MSE, MAE, R squared, RMSE, and MAPE.

TABLE VII. PERFORMANCE METRICS FOR GOALKEEPER SAVES

Model	MSE	MAE	R-Squared	RMSE	MAPE (in %)
Random Forest	0.04	0.55	0.96	0.20	10.53
XG Boost	0.03	0.45	0.97	0.17	7.92
Linear Regression	0.02	0.20	0.98	0.14	5.86
KNN	0.12	0.95	0.85	0.35	32.90
Decision Tree	0.07	0.79	0.93	0.26	18.24
SVR	0.08	2.83	0.91	0.28	21.67
Gradient Boosting	0.05	0.74	0.95	0.22	13.48

The findings from TABLE VII and Figure 3 depicts the performance of models. The analysis of results indicates that Linear Regression, along with ensemble methods such as Random Forest and XGBoost, demonstrate strong performance in capturing the underlying structure of the datasets [4]. Whereas KNN offered a comparatively less accurate prediction along with higher error rates [10].



Linear Regression exhibited high accuracy and was identified as the best model for predicting goalkeeper saves. It achieved the highest accuracy with the lowest error rates, making it the most suitable model for this prediction task. Ensemble methods such as Random Forest and XG Boost

were good alternatives that yielded good performance as well. In contrast, models like KNN were found to be inadequate in handling the input variations present in the dataset.

IX. CONCLUSION

When the focus shifts from individual metrics to team metrics to evaluate players performance using machine learning models, it provides a holistic view of a player's contribution within the team, ensuring more context-aware and balanced performance evaluation. This approach enhances predictive accuracy and improves decision making in football analytics. Emphasizing team based data such as collaborative actions and group strategies, models would depict patterns that would otherwise be missed when dealing with a single performance metric. This entails good decision-making and formulation of strategy, adding greater predictive capacity to the model in forecasting match outcomes.

For predicting the top goal scorer, Random Forest, Gradient Boosting, XGBoost, have demonstrated strong predictive performance by effectively capturing the nonlinear relationships between player abilities, playing conditions, and tactical factors. The models utilize several features, such as shots, team attacks, and goals per 90 minutes, which are helpful in predicting scoring potential with high predictive power.

For predicting the best goalkeeper, Random Forest and Linear Regression excel by assessing both individual capabilities and team dynamics, providing a more accurate evaluation of goalkeeper performance throughout the match.

To improve the model accuracy, essential preprocessing steps were taken that includes handling missing data, normalizing or standardizing features, encoding categorical variables, and ensuring data consistency. Feature engineering, such as creating interaction terms or extracting meaningful temporal features, can further enhance model performance.

To conclude, this work compared the seven different machine learning models for predicting top goal scorers and best goalkeepers by using both individual and team based statistics.

REFERENCES

- [1] Maystre, Lucas, et al. "The player kernel: learning team strengths based on implicit player contributions." arXiv preprint arXiv:1609.01176 (2016).
- [2] GUDMUNDSSON, J. and HORTON, M., 2016. Spatio-Temporal Analysis of Team Sports-A Survey. arXiv Prepr. arXiv1602. 06994.
- [3] Seidenschwarz, P.G., 2021. Data-Driven Analytics for Decision Making in Game Sports (Doctoral dissertation, University_of_Basel).
- [4] Igiri, C.P. and Nwachukwu, E.O., 2014. An improved prediction system for football a match result. IOSR journal of Engineering, 4(12), pp.12-20.
- [5] Alfredo, Y.F. and Isa, S.M., 2019. Football match prediction with tree based model classification. International Journal of Intelligent Systems and Applications, 11(7), pp.20-28.
- [6] Razali, N., Mustapha, A., Yatim, F.A. and Ab Aziz, R., 2017, August. Predicting football matches results using Bayesian networks for English Premier League (EPL). In Iop conference series: Materials science and engineering (Vol. 226, No. 1, p. 012099). IOP Publishing.
- [7] Rotstein, A.P., Posner, M. and Rakityanskaya, A.B., 2005. Football predictions based on a fuzzy model with genetic and neural tuning. Cybernetics and Systems Analysis, 41, pp.619-630.
- [8] Rodrigues, F. and Pinto, Â., 2022. Prediction of football match results with Machine Learning. Procedia Computer Science, 204, pp.463- 470.
- [9] Lutz, R., 2015. Fantasy football prediction. arXiv preprint arXiv:1505.06918.
- [10] Langseth, H., 2013. Beating the bookie: A look at statistical models for prediction of football matches. In Twelfth Scandinavian Conference on Artificial Intelligence (pp. 165-174). IOS Press.
- [11] Goes, F.R., Meerhoff, L.A., Bueno, M.J.O., Rodrigues, D.M., Moura, F.A., Brink, M.S., Elferink-Gemser, M.T., Knobbe, A.J., Cunha, S.A., Torres, R.S. and Lemmink, K.A.P.M., 2021. Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. European Journal of Sport Science, 21(4), pp.481-496.
- [12] Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D. and Giannotti, F., 2019. PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Transactions on Intelligent Systems and Technology (TIST), 10(5), pp.1-27.
- [13] Anzer, G. and Bauer, P., 2021. A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). Frontiers in sports and active living, 3, p.624475.
- [14] Wang, Z., Veličković, P., Hennes, D., Tomašev, N., Prince, L., Kaisers, M., Bachrach, Y., Elie, R., Wenliang, L.K., Piccinini, F. and Spearman, W., 2024. TacticAI: an AI assistant for football tactics. Nature communications, 15(1), p.1906.
- [15] Diverse Machine Learning for Forecasting Goal-Scoring Likelihood in Elite Football Leagues Christina Markopoulou, George Papageorgiou and Christos Tjortjis School of Science and Technology, International Hellenic University, 57001 Thessaloniki,Greece
- [16] Huang, Q.; Mao, J.; Liu, Y. An Improved Grid Search Algorithm of SVR Parameters Optimization. In Proceedings of the 2012 IEEE 14th International Conference on Communication Technology, Chengdu, China, 9–11 November 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1022–1026.
- [17] Zeng, Z.; Pan, B. A Machine Learning Model to Predict Players Positions Based on Performance. In Proceedings of the 9th International Conference on Sport Sciences Research and Technology Support, Online, 28–29 October 2021; SCITEPRESS—Science and Technology Publications: Setúbal, Portugal, 2021; pp. 36–42.
- [18] Chaudhary, R., & Verma, H. (2020). Model evaluation metrics in machine learning. Journal of Computational Sports Analytics, 8 (2), 113–127