machine learning & knowledge extraction

MDPI

Article

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; cmarkopoulou@ihu.edu.gr (C.M.); gpapageorgiou2@ihu.edu.gr (G.P.)
* Correspondence: c.tjortjis@ihu.edu.gr

Abstract: The field of sports analytics has grown rapidly, with a primary focus on performance forecasting, enhancing the understanding of player capabilities, and indirectly benefiting team strategies and player development. This work aims to forecast and comparatively evaluate players' goal-scoring likelihood in four elite football leagues (Premier League, Bundesliga, La Liga, and Serie A) by mining advanced statistics from 2017 to 2023. Six types of machine learning (ML) models were developed and tested individually through experiments on the comprehensive datasets collected for these leagues. We also tested the upper 30th percentile of the best-performing players based on their performance in the last season, with varied features evaluated to enhance prediction accuracy in distinct scenarios. The results offer insights into the forecasting abilities of those leagues, identifying the best forecasting methodologies and the factors that most significantly contribute to the prediction of players' goal-scoring. XGBoost consistently outperformed other models in most experiments, yielding the most accurate results and leading to a well-generalized model. Notably, when applied to Serie A, it achieved a mean absolute error (MAE) of 1.29. This study provides insights into ML-based performance prediction, advancing the field of player performance forecasting.

Keywords: sport analytics; performance prediction; machine learning (ML); data analytics; football

Citation: Markopoulou, C.; Papageorgiou, G.; Tjortjis, C. Diverse Machine Learning for Forecasting Goal-Scoring Likelihood in Elite Football Leagues. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1762–1781. https://doi.org/10.3390/make6030086

Academic Editor: Abdulhamit Subasi

Received: 24 April 2024
Revised: 10 July 2024
Accepted: 25 July 2024
Published: 28 July 2024

1. Introduction

In the realm of sports, a big change occurred due to sports analytics. The use of advanced tracking technologies not only offers organizations and coaches vital insights regarding athlete performance but also generates a wealth of data [1]. The extensive data produced serve as a catalyst, empowering coaches to refine their decision-making and strategic approaches [2]. This insight extends beyond shaping roster composition, cost reduction, and increasing team value [3,4]. Moreover, this wave of innovation not only amplifies team competitiveness but also injects a new level of excitement into sports for fans. Real-time access to detailed statistical information improves the fan experience by providing a stronger connection between spectators and the complicated nature of the game.

Football, as the most generally recognized and followed sport worldwide, provides an ideal arena for the application and research of sports analytics. The game's complicated design, combined with its massive global fan base, provides a rich tapestry for the analysis of advanced analytical methods. In this context, this research focuses on forecasting a player's performance by predicting the number of goals a player is likely to achieve in the upcoming season based on historical data.

Noteworthy studies in the field of sports analytics are mentioned below. The authors in [5] conducted two experiments related to football, focusing on team and player performance prediction. In the first experiment, they employed two tactics. The primary objective of the first approach was to forecast whether a team would secure a better position in the table for the 2017–2018 season compared to the previous two seasons. Using the random

forest algorithm, this method achieved an accuracy of 70%. The second strategy involved simulating football matches for the 2018–2019 season to categorize results as home victories, away wins, or draws. The English Premier League exhibited the highest match outcome accuracy at 57%, while the Spanish La Liga had the lowest root mean squared error (RMSE). In their second experiment, the researchers explored the characteristics and moves during a game that could impact a defender's rating. The dataset included 59 central defenders from the English Premier League during the 2016–2017 season. They employed the multiple linear regression model with backward elimination, achieving an R-squared metric of 0.867. In our study, we focus on player performance by considering all the various positions that players occupy on the field, aiming to predict the total number of goals scored.

In [6], the researchers utilized the Wyscout public dataset to forecast player positions using sports performance and psychological attributes. Six key indicators, encompassing accuracy of shot, accuracy of simple pass, accuracy of glb (ground loose ball), accuracy of defending duel, accuracy of air duel, and accuracy of attacking duel, were selected as input variables to train a BP neural network. The model's hyperparameter combinations were evaluated using k-fold cross-validation. Ultimately, the model attained an accuracy rate of 77%. Compared with this study, our research advances by using player positions, along with other variables, to enhance the prediction of the total number of goals scored.

Furthermore, injuries in sports pose a threat for both individuals and teams, with possible long-term consequences for players' careers and the overall effectiveness and achievements of sports clubs. These injuries frequently necessitate extensive times, affecting team performance and match outcomes. Thus, injuries are of great importance in the world of sports.

In 2020, a study [7] aimed to investigate the effectiveness of machine learning (ML) in detecting injury risk factors among elite male youth footballers. The research involved analyzing 355 athletes who underwent a series of neuromuscular tests (anthropometric measurements, single leg countermovement jump, tuck jump assessments). The results highlighted various factors associated with injury risk. The most common were asymmetry in a single-leg countermovement jump (SLCMJ), 75% hop, Y-balance, tuck jump knee valgus, and anthropometrics measures.

Additionally, in 2022, researchers conducted a study focusing on predicting injury risk in professional football players using body composition parameters and physical fitness evaluations. Their research, which comprised 36 male players from the First Portuguese Soccer League during the 2020–2021 season, looked at 22 distinct characteristics. Sectorial postures, body height, sit-and-reach performance, one-minute push-up count, handgrip strength, and 35-min linear speed were all found to be the most important variables in predicting injury risk for elite football players, using net elastic analysis. Notably, ridge regression was the most accurate model, with an RMSE of 0.591 for predicting the frequency of potential injury occurrences [8]. This study differs in focus from our research; however, both studies utilize regression models, among other techniques, to predict their target variables.

Football teams are also using wearable gadgets during training and matches to track players' physical abilities. These devices help experts analyze data and provide useful insights to clubs for better player management and strategic planning. The rising use of wearable technology highlights its growing importance in influencing football-related decisions.

Specifically, in 2022, researchers in [9] attempted to construct a model for predicting lower-body injuries in male footballers resulting from over- or undertraining leveraging wearable technology. It is widely recognized that predicting injuries remains challenging due to individual biological variations and players' psychophysical conditions. The study utilized Catapult wearable global positioning trackers to gather data during both training sessions and matches. Among the algorithms, XGBoost produced the highest accuracy, reaching 90%. The utilization of wearable devices will improve player performance analysis by delivering real-time data on metrics like heart rate, movement patterns, etc. This information will be essential to having more accurate results.

1.1. Related Work

Understanding and predicting football players' performance is an important aspect of sports analytics. Extensive research has been conducted in this area, with the goal of uncovering crucial findings that will benefit the broader field of football analytics. This subsection provides a brief overview of the related work that has influenced our understanding of predicting a player's performance.

In [10], the researchers undertook a study on predicting football player performance, specifically focusing on overall performance value. They developed separate models based on player position, leading to a linear regression algorithm with an accuracy of 84.34%. Additionally, when predicting a player's future market value based on the performance values of the first model, the algorithm demonstrated 91% accuracy. With this approach, coaches should be able to identify football potential without bias stemming from factors such as team budget or league competitiveness.

In 2018, a study was developed with the goal of predicting English Premier League football outcomes [11]. The dataset covered a period of 11 seasons, with the training phase comprising 9 seasons (from 2005 to 2014), followed by two seasons of testing (from 2014 to 2016). The home/away attribute emerged as one of the most important features. This attribute depicts whether a team plays at its home stadium or not. Predicting football outcomes posed challenges, notably due to the substantial occurrence of draws, which constitute 25% of the testing dataset. Various models, including Gaussian naïve Bayes, support vector machine, random forest, and gradient boosting, were evaluated during the experimentation. The best model was gradient boosting, which achieved a ranked probability score (RPS) of 0.2158 from weeks 6 to 38 in the English Premier League over the 2 seasons.

Different research examines how situational variables and performance indicators affect match outcomes in the English Premier League during the 2017–2018 season. Using decision trees, it was discovered that scoring first was the most important factor. Clearance, show, and possession percentage have varying importance depending on the opponent's quality. The findings can assist coaches and managers in setting goals for players and teams during training and games [12].

1.2. Research Overview

This dissertation delves into football, a globally known sport. Its aim is to predict a player's performance in terms of goals using historical data from the preceding four seasons (2018–2019 to 2021–2022) and conduct the evaluation in the final season (2022–2023). Specifically, this study includes players from four leagues: Bundesliga, Premier League, La Liga, and Serie A. Additionally, a dataset comprising players for all leagues was implemented.

Data collection relied on a reliable source, Sports Reference. Data were collected from seasons 2017–2018 to 2022–2023 with more than 5000 players. Furthermore, preprocessing and feature engineering were necessary to format the dataset appropriately. As part of the process were the transformation of data to historical (season lag features) and the division of the dataset, focusing on players within the top 30% in terms of scoring performance. Subsequently, each version was subdivided into three cases based on the attributes utilized in the training phase, as detailed in the subsequent Section 2.2.1. Data Collection.

Various ML algorithms were evaluated, including linear and ridge regression, random forest, gradient boosting, XGBoost, and multilayer perceptron. The effectiveness of the models was measured using metrics like mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), mean absolute percentage error (MAPE), and R-squared.

1.3. Contributions

This paper presents a comparative study of the four major European leagues: Bundesliga, Premier League, La Liga, and Serie A. This comparison underscores the strengths and weaknesses of various ML models, providing insights into their effectiveness. Our

findings suggest that XGBoost should be considered a strong candidate for predicting the total number of goals for datasets structured similarly to those in this study.

Additionally, our research identifies the league-specific datasets that yield the most effective performance prediction outcomes. By analyzing attributes such as player positions, historical performance metrics, and other relevant variables, we pinpoint the key factors that contribute to accurate goal prediction across different leagues.

2. Materials and Methods

This section outlines the processes and complications involved in the methodology. It examines the entire data collection process, starting from scraping data to cleansing and feature engineering. The goal is to illustrate the modifications made to the dataset before implementing ML algorithms. Alongside, the research hypotheses that guide our investigation are presented.

Model comparisons are assessed using metrics such as MAE, MSE, RMSE, MAPE, and R-squared. These metrics provide a comprehensive evaluation of the performance of our predictive models.

2.1. Research Questions/Hypothesis

1. How do different ML models, including linear regression, ridge regression, random forest, gradient boosting, XGBoost, and multilayer perceptron (MLP), compare in predicting football player performance?
2. Which league-specific dataset demonstrates the most effective performance prediction outcomes, and based on what attributes?

2.2. Methodology

This section provides an in-depth exploration of the procedures that contribute to the effectiveness of the analytical process.

The initial step involves data collection through scraping from Sports Reference, a platform offering athlete statistics across various sports. The dataset includes football players from the 2017–2018 to 2022–2023 seasons, exceeding 5000 players with a total of 35 features. The dataset, comprising a diverse range of football players representing various nations, teams, and leagues, has been narrowed down to exclusively include players from four leagues: Bundesliga, Premier League, La Liga, and Serie A. The ML algorithms are trained using data covering the 2018–2019 to 2021–2022 seasons, with the subsequent 2022–2023 season employed as the test dataset.

During the initial phases of data preprocessing, the dataset was refined to include players participating in all seasons, resulting in a significant reduction in data. Moreover, two versions of the dataset were created, one that contained all the players and another with players who are in the top 30% quartile based on goal performance. Finally, three different cases were developed regarding the training features. Case 1 considered the features most strongly correlated with the target variable 'Goals'; case 2 involved the removal of one attribute from highly correlated pairs; and case 3 retained all available columns.

To enhance realism in evaluating football player performance predictions, we converted the dataset to include past statistics. Season lag features from previous seasons were introduced, allowing models to forecast season goals using data from preceding seasons. Additionally, we introduced a 'Previous_Gls' column, indicating the player's goal count in the prior season.

To fulfill the primary objective of this study, various ML models are used, including linear regression, ridge regression, random forest, gradient boosting, XGBoost, and multilinear regression.

For this study, several libraries were used, including Pandas for data manipulation and analysis and Matplotlib for data visualization. Furthermore, sklearn was utilized to implement and evaluate the ML models. Ultimately, the assessment of the outcomes was conducted using three metrics: MAE, MSE, RMSE, MAPE, and R-squared. These

metrics provide valuable insights into the dependability and effectiveness of the models, and their analysis is presented in the following section. Figure 1. presents a flowchart for the proposed methodology.

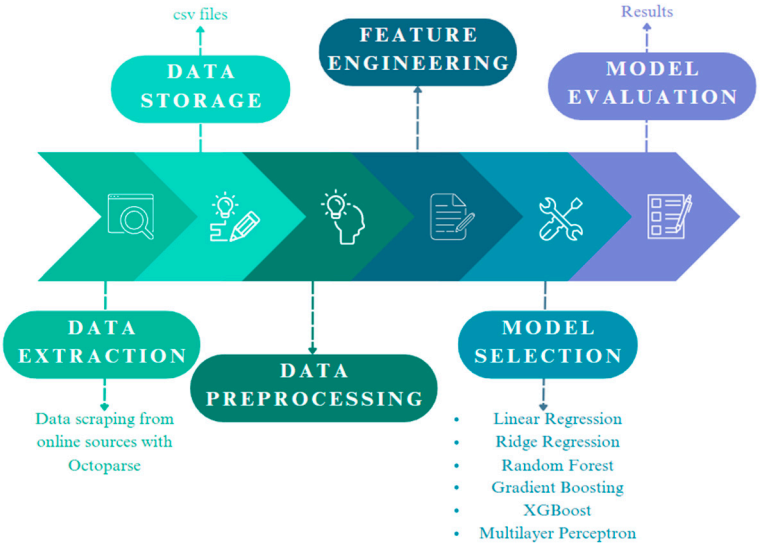

Figure 1. Flowchart for the proposed methodology.

2.2.1. Data Collection

The central focus of this paper revolves around the process of data collection. In the realm of football statistics, a plethora of websites offer information on clubs and players. Consequently, ensuring the legitimacy of the acquired data becomes critical, as any inaccuracies could jeopardize the precision of the results.

Specifically, this study's dataset was obtained from Sports Reference [13], a renowned organization that provides significant data coverage across a wide range of sports. To execute the data collection procedure, the scraping tool that was used was Octoparse.

Table 1 presents information about the number of records and features in the initial scraped dataset per league, and Table 2 provides feature descriptions.

Table 1. Dataset dimensions.

Leagues	N. of Rows	N. of Columns
Bundesliga	3000	35
Premier League	3207	35
La Liga	3482	35
Serie A	3576	35

The dataset incorporates a set of features detailed below. Nevertheless, it is important to note that the dataset employed for the ML algorithms underwent significant transformations, resulting in a format distinct from the one described above. A detailed analysis of these changes is presented in the Section 2.2.2. Pre-processing and Section 2.2.3. Feature Engineering subsections.

Table 2. Dataset description.

Feature	Description
Player	Name of the player
Nation	Nationality of the player
Pos	Position most played by the player

Table 2. *Cont.*

Feature	Description
Squad	Club the player is currently playing
Age	Age of the player at season's start
Born	Player's year of birth
MP	Matches played by the player
Starts	Game or games started by the player
Min	Minutes played by the player
90s	Minutes played divided by 90
GLs	Goals scored or allowed
Ast	Assists
G+A	Goals and assists
G-PK	Non-penalty goals
PK	Penalty Kicks made
Pkatt	Penalty Kicks attempted
CrdY	Yellow cards
CrdR	Red cards
xG	Expected goals
npxG	Non-Penalty Expected Goals
xAG	Expected Assisted Goals
npxG+xAG	Non-Penalty Expected Goals plus Assisted Goals
PrgC	Progressive Carries
PrgP	Progressive Passes
PrgR	Progressive Passes Rec
GLs per 90′	Goals scored per 90 min
Ast per 90′	Assists per 90 min
G+A per 90′	Goals and assists per 90 min
G+A-PK per 90′	Goals plus Assists minus Penalty Kicks made per 90 min
xG per 90′	Expected Goals per 90 min
xAG per 90′	Expected Assisted Goals per 90 min
xG+xAG per 90′	Expected Goals plus Assisted Goals per 90 min
npxG per 90′	Non-Penalty Expected Goals per 90 min
npxG+xAG per 90′	Non-Penalty Expected Goals plus Assisted Goals per 90 min

2.2.2. Pre-Processing

First, a series of modifications were applied to the dataset to improve its suitability for the prediction models. An initial adjustment involved converting object-type columns into strings. Furthermore, attributes 'Rank' and '90s-Minutes played divided by 90' were eliminated due to their lack of meaningful information.

Another observation revealed examples of players who played for multiple football clubs during the same season. As a result, the decision was made to calculate the average value for players in such situations, specifically for arithmetic columns. A composite string name was generated in the 'Squad' column, concatenating team names for these players.

A key criterion in this phase was the inclusion of players who participated in all six seasons, leading to the removal of those who did not meet this criterion. Consequently, the dataset underwent a significant reduction. Bundesliga experienced a reduction from 1185 distinct players to 109 players, while the Premier League saw a decrease from 1298 unique players to 112 players. Likewise, La Liga witnessed a decline from 1431 individual players to 97 players, and Serie A had a decrease from 1441 unique players to 106. Additionally, a supplementary dataset was introduced, encompassing players from all leagues (424 players in total). To distinguish players and their respective leagues, a new 'League' column was introduced, featuring numerical codes (e.g., League = 1 for Bundesliga, League = 2 for Premier League, League = 3 for La Liga, League = 4 for Serie A).

A more advanced distinction was made, focusing on players' goal performance during the most recent season (2022–2023). The dataset was divided into two distinct subsets: one containing all players and another containing only those ranked in the top 30% quartile based on their goal achievements in the last season.

Subsequently, the focus shifted towards determining the features to be included in the algorithms, a critical process known as dimensionality reduction. Dimensionality reduction decreases the total number of input variables in a dataset [14].

Case 1 contained the 10 columns with the highest correlation to the target variable 'Goals'. The selection criteria were based on the Pearson correlation coefficient. In Case 2, a distinctive approach involved calculating the percentage of correlation for each pair of attributes. As a result, one column from each highly associated pair was kept. Finally, case 3 included all available columns from the dataset. Notably, in the dataset containing the total number of players across all leagues, the column 'League' was introduced to distinguish the players. Table 3 presents the features for Case 1, Case 2, and Case 3.

Table 3. Features for each case.

Cases	Features
Case 1	xG, npXG, npXG+xAG, xG per 90′, Previous Goals, npxG per 90′, xG+xAG per 90′, G+A, PrgR
Case 2	Nation, Pos, Squad, Age, MP, Ast, PK, CrdY, CrdR, PrgC, PrgP, PrgR, Previous Goals
Case 3	Nation, Pos, Squad, Age, Born, MP, Starts, Min, Ast, G+A, G-PK, PK, PKatt, CrdY, CrdR, xG, npxG, xAG, npxG+xAG, PrgC, PrgP, PrgR, Gls per 90, Ast per 90, G+A per 90, G-PK per 90, G+A-PK per 90, xG per 90, xAG per 90, xG+xAG per 90, npxG per 90, npxG+xAG per 90, Previous Goals.

2.2.3. Feature Engineering

As outlined earlier, the initial objective was to train the ML algorithms using the dataset of the first four seasons (2018–2019 to 2021–2022) and subsequently evaluate their performance on the test set from the last season (2022–2023). Nonetheless, since key statistics such as predicted goals and assists are included, using this approach could produce results that are too optimistic and do not reflect realistic outcomes.

To address this concern, an alternative methodology was implemented. To avoid reliance on current-season statistics, the dataset underwent transformation to incorporate historical data. Each row displayed past statistics, enabling the algorithm to predict a player's goal count for the 2018–2019 season using data from the previous season (2017–2018). Additionally, a new column, 'Previous Goals,' was introduced, denoting the player's goal for the 2017–2018 season, while the 'Goals' column indicated the goals for the subsequent season (2018–2019). Therefore, in the final dataset, each row depicts the seasonal performance statistics of each player from the last season, aiming to forecast the upcoming season's goals.

The primary goal was to anticipate how many goals a player would score in the 2022–2023 season using data from the previous season (2021–2022). This strategy, known as season lag features, uses historical data to identify patterns that contribute to accurate predictions.

2.2.4. Modeling

Six different ML algorithms were used to predict the number of goals the player will achieve in the 2022–2023 season. These were: linear regression, ridge regression, random forest, gradient boosting, XGBoost, and multilayer perceptron algorithm.

Linear regression is a statistical approach for modeling the relationship between a dependent variable and one or more independent variables by fitting a straight line through the data points. The goal is to select the best-fitting line that minimizes the discrepancy between observed and anticipated values [15]. Linear regression was chosen for its simplicity, providing a strong baseline for comparison. On the other hand, ridge regression is a statistical technique that reduces the multicollinearity in linear regression, which arises when independent variables are strongly correlated [16]. A ridge regression

model estimates coefficients using a biased estimator instead of ordinary least squares (OLS), resulting in lower variance and reduced standard error, making it useful for addressing multicollinearity issues [2].

Random forest is an ensemble method that can be used for both regression and classification problems. It constructs many decision trees during training and returns the average prediction (regression) or most frequent class (classification) of the individual trees. It is robust, scalable, and good at handling complicated datasets while minimizing overfitting [17].

Another algorithm is gradient boosting. It is a model that combines an ensemble of weak learners, most commonly decision trees. It works by fitting each new tree to the residual errors of the preceding ones, progressively increasing the model's prediction accuracy [18]. The XGBoost algorithm is an optimized implementation of gradient boosting. It integrates advanced features such as regularization and tree pruning techniques [19]. These models were chosen because of their ability to handle different data distributions through ensemble techniques.

Lastly, MLP is an artificial neural network with multiple layers of neurons. It includes an input layer, one or more hidden layers, and an output layer. Except for the input layer, each employs nonlinear activation functions to capture complex data relationships. MLP is good for capturing intricate patterns in data [20].

Furthermore, grid search was employed for all algorithms to optimize hyperparameters, aiming to find the most effective combination of values for each model [21]. The hyperparameters table with the values for the best models of each scenario is available in Appendix A, in Tables A1–A5. Additionally, feature importance was performed to determine the impact of input variables on model prediction. All algorithms produced feature importance scores, except for MLP. Furthermore, predictions were rounded to integers to ensure compatibility with the discrete structure of goal counts. Finally, metrics were calculated for both training and testing datasets to enable thorough evaluation and comparison.

3. Results

This section evaluates the outcomes of each algorithm across the different datasets, followed by a comparative analysis to determine the best performer. The primary metric for assessment is MAE. MAE indicates the proximity of predictions to actual values. The influence of the remaining metrics is similarly crucial. Nevertheless, MAPE encounters inaccuracies in cases with null values presented in the target variable; hence, it will only be included in the results of the top 30% player dataset.

3.1. Bundesliga's Player Performance

This section shows the results of the Bundesliga. This dataset consists of players who played in Bundesliga teams during the full six-season duration, from 2017–2018 to 2022–2023.

3.1.1. All Players' Analysis for Bundesliga

It is noteworthy that the standard deviation for this dataset was computed to 3.24 goals, reflecting the extent of variability in goal-scoring performances. Moreover, the aggregate count of players amounts to 109.

Table 4 presents the best models per case and the corresponding metrics.

The error values presented are relatively small, indicating satisfactory performance of our models overall. Notably, the XGBoost algorithm, utilizing features from case 2, appears to outperform others. The most influential predictor for XGBoost is the number of previous goals. Additionally, the consistency between MAE and RMSE values in both training and testing results suggests that the models avoid overfitting. Table 5 presents feature importance for the most accurate models for all players per case, where bold indicates the three most important features.

Table 4. Most accurate models for all players per case for Bundesliga.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	R^2 Test/R^2 Train
Case 1	Ridge Regression	1.80/1.50	6.37/4.99	2.52/2.23	0.41/0.52
Case 2	XGBoost	1.71/1.03	5.41/2.28	2.33/1.51	0.50/0.78
Case 3	Ridge Regression	1.80/1. 48	6.77/4.77	2.60/2.18	0.38/0.54

Table 5. Feature importance for the most accurate models for all players per case for Bundesliga.

Cases	Best Model/ Features	xG per 90′	npxG per 90′	xG+xAG per 90′	Previous Goals	PrgR	CrdY
Case 1	Ridge Regression	**6.06**	**4.61**	**1.75**	0.29	0.00	-
Case 2	XGBoost	-	-	-	**0.41**	**0.09**	**0.08**
Case 3	Ridge Regression	**3.19**	**2.37**	**3.21**	0.04	0.00	−0.17

3.1.2. Elite Players' Analysis for Bundesliga

The initial dataset underwent a reduction from 109 to 34 players, with the requirement that the players must have scored at least 3 goals to be included. The implementation also introduced the mean absolute percentage error (MAPE) metric, where lower values indicate better performance. The dataset's standard deviation, measured at 3.79 goals, provides insight into data variability. Table 6 presents the most accurate models per case for the top 30% of elite players.

Table 6. Most accurate models for top 30% elite players per case for Bundesliga.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	MAPE Test
Case 1	Ridge Regression	2.27/1.97	8.85/7.24	2.86/2.69	0.38
Case 2	Random Forest	1.71/1.02	4.38/1.73	2.09/1.31	0.29
Case 3	Gradient Boosting	2.17/1.60	8.60/4.18	2.93/2.05	0.30

Results indicate the consistent superiority of the random forest algorithm. The metrics of this model have persistently better values across the error parameters. Specifically, a MAPE of 0.29 implies that, on average, model predictions differ from actual values by around 29%. Generally, lower MAPE values typically signify enhanced accuracy.

It is observed that feature previous goals is again the most important factor for the best model, while expected plus assisted goals per 90 min significantly impacts the other 2 models. Finally, the absence of a significant difference between train and test measures shows that overfitting did not occur. Table 7 presents the feature importance for the best models for the top 30% of elite players per case, where bold indicates the three most important features.

Table 7. Feature importance for the most accurate models for top 30% elite players per case for Bundesliga.

Cases	Best Model/ Features	xG+xAG per 90′	xG per 90′	npxG per 90′	Previous Goals	Age	MP	GLs per 90′	PrgR
Case 1	Ridge Regression	**6.06**	**3.68**	**3.24**	0.20	-	-	-	0.00
Case 2	Random Forest	-	-	-	**0.39**	**0.09**	**0.08**	-	0.07
Case 3	Gradient Boosting	**0.51**	**2.37**	**3.21**	0.03	0.00	0.02	0.06	0.04

3.2. Premier League's Player Performance

This section presents the findings for Premier League, offering a thorough examination of the outcomes. The dataset comprises solely players who participated in a Premier League team for the entire six seasons.

3.2.1. All Players' Analysis for Premier League

The dataset's standard deviation was determined to be 4.93 goals, covering a total of 112 players. Among the three cases, XGBoost consistently emerged as the best performing model, demonstrating the lowest MAE when utilizing features from case 1. Table 8 highlights the best models per case.

Table 8. Most accurate models for all players per case for Premier League.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	R^2 Test/R^2 Train
Case 1	XGBoost	1.93/1.66	10.35/6.17	3.22/2.48	0.53/0.75
Case 2	XGBoost	1.93/1.75	9.88/6.73	3.14/2.59	0.55/0.73
Case 3	XGBoost	1.95/1.59	10.60/5.57	3.26/2.36	0.52/0.78

Significantly, there are minimal disparities between the training and test values for MAE and RMSE, although slightly larger variations are observed for MSE.

Table 9 presents feature importance for the most accurate models for all players' analysis per case, where bold indicates the three most important features. It highlights the importance of the 'Previous Goals' feature across all cases, while expected goals emerged as the most influential feature in the top-performing model.

Table 9. Feature importance for the most accurate models for all players per case for Premier League.

Cases	Best Model/ Features	xG	Previous Goals	npxG	Pos	PK
Case 1	XGBoost	**0.42**	**0.15**	**0.12**	-	-
Case 2	XGBoost	-	**0.45**	-	**0.09**	**0.06**
Case 3	XGBoost	**0.25**	**0.08**	**0.07**	0.02	0.02

3.2.2. Elite Players' Analysis for Premier League

The initial dataset of 112 players underwent refinement to encompass only 35 players meeting the goal criteria. The refinement revealed that players have to accomplish at least 3 goals to qualify for inclusion in the final dataset. Standard deviation for this dataset was calculated to be 6.04 goals. Table 10 presents the most accurate models per case for the top 30% of elite players.

Table 10. Most accurate models for top 30% elite players per case for Premier League.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	MAPE Test
Case 1	XGBoost	3.67/2.29	27.34/9.74	5.23/3.12	0.49
Case 2	Gradient Boosting	3.27/2.23	20.38/7.48	4.51/2.73	0.48
Case 3	Random Forest	3.61/2.26	25.39/8.04	5.04/2.83	0.50

Across all cases, gradient boosting yielded the lowest MAE for both training and test sets. Upon examination of the error metric values for all models, signs of overfitting become apparent. Additionally, gradient boosting uses 'Previous Goals' as the most important feature of goal-scoring. This observation is evident in Table 11, illustrating feature importance for the top 30% of elite players per case, with bold highlighting the three most significant features.

Table 11. Feature importance for the most accurate models for top 30% elite players per case for Premier League.

Cases	Best Model/ Features	xG	Previous Goals	G+A	PK	PrgP	Gls per 90′
Case 1	XGBoost	**0.38**	**0.15**	**0.08**	-	-	-
Case 2	Gradient Boosting	-	**0.65**	-	**0.06**	**0.04**	-
Case 3	Random Forest	**0.44**	**0.17**	0.04	0.03	0.00	**0.06**

3.3. La Liga's Player Performance

The results and discoveries from La Liga are presented below. This league comprises players who have exclusively competed for La Liga teams for the six-season period.

3.3.1. All Players' Analysis for La Liga

The dataset encompasses 97 players, with a standard deviation of 4.35 goals. The small differences between training and testing values suggest minimal overfitting. Table 12 depicts the most accurate models per case and the corresponding metrics. In particular, MLP results as the preferred algorithm based on MAE, using the features of case 1.

Table 12. Most accurate models for all players per case for La Liga.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	R^2 Test/R^2 Train
Case 1	MLP	1.72/1.88	6.91/8.66	2.63/2.94	0.48/0.57
Case 2	XGBoost	1.89/1.65	7.49/6.33	2.74/2.52	0.43/0.69
Case 3	XGBoost	1.78/1.48	6.94/4.91	2.63/2.21	0.48/0.76

Table 13 presents feature importance for the most accurate models for all players per case, where bold indicates the three most important features. It is worth noting that no feature importance values are available for MLP, which was the best model in this analysis. However, for XGBoost, the primary feature is previous goals.

Table 13. Feature importance for the most accurate models for all players per case for La Liga.

Cases	Best Model/ Features	Previous Goals	PK	PrgP	xG	Gls per 90′	xG+xAG per 90′
Case 1	MLP	-	-	-	-	-	-
Case 2	XGBoost	**0.45**	**0.11**	**0.08**	-	-	-
Case 3	XGBoost	0.03	0.00	0.01	**0.18**	**0.11**	**0.08**

3.3.2. Elite Players' Analysis for La Liga

The dataset decreased from 97 to 32 players. The results of this selection process revealed that players needed to achieve at least two goals to be included in the final dataset. The dataset's variance is demonstrated by a calculated standard deviation of 5.55. Furthermore, the proximity of values between MAE and RMSE for both training and testing sets remains below 1.00, indicating minimal overfitting in the data.

Table 14 highlights the best models per case for the top 30% of elite players. Random forest in case 1 demonstrates superior performance compared to the other algorithms, particularly in terms of MAE, whereas random forest in case 2 has better results in MSE and RMSE.

Moreover, based on the feature importance values, the most important feature for random forest is non-penalty expected plus assisted goals. Table 15 presents feature importance for the best models per case for the top 30% of elite players, where bold indicates the three most important features.

Table 14. Most accurate models for top 30% elite players per case for La Liga.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	MAPE Test
Case 1	Random Forest	2.37/2.78	10.99/12.92	3.31/3.59	0.50
Case 2	Random Forest	2.39/3.05	10.67/15.10	3.27/3.89	0.53
Case 3	Ridge Regression	2.42/3.20	11.17/17.74	3.34/4.21	0.53

Table 15. Feature importance for the most accurate models for top 30% elite players per case for La Liga.

Cases	Best Model/ Features	npxG+xAG	G+A	xG per 90′	Previous Goals	PrgR	PK
Case 1	Random Forest	**0.31**	**0.16**	**0.13**	0.02	0.02	-
Case 2	Random Forest	-	-	-	**0.68**	**0.13**	**0.07**
Case 3	Ridge Regression	**0.09**	**0.12**	0.00	**0.09**	0.00	0.02

3.4. Serie A's Player Performance

In this section, the outcomes of Serie A are examined. The data were narrowed down to exclusively include players from Serie A teams from 2017–2018 to the 2022–2023 season.

3.4.1. All Players' Analysis for Serie A

The dataset encompasses a total of 106 players. Additionally, the calculated standard deviation of 4.35 goals provides valuable information into the variability of goal-scoring performance.

Table 16 highlights the best models per case for all players. Analyzing the metrics table reveals an overall satisfactory performance. The minor differences between the two values imply limited overfitting. Eventually, the XGBoost algorithm, using the feature of case 1, emerges as the favored based on MAE.

Table 16. Most accurate models for all players per case for Serie A.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	R^2 Test/R^2 Train
Case 1	XGBoost	1.29/1.39	3.96/4.68	1.99/2.16	0.48/0.76
Case 2	XGBoost	1.40/1.49	3.66/5.59	1.91/2.36	0.52/0.71
Case 3	XGBoost	1.33/1.32	4.12/4.23	2.03/2.06	0.46/0.78

The feature with the biggest influence is the expected goals, followed by expected plus assisted goals per 90 min. Table 17 presents feature importance for the best models for all players per case, where bold indicates the three most important features.

Table 17. Feature importance for the most accurate models for all players per case for Serie A.

Cases	Best Model/ Features	xG	xG+xAG per 90′	npxG	Previous Goals	PK	PrgR	npxG+xAG
Case 1	XGBoost	**0.48**	**0.12**	**0.10**	**0.02**	-	0.04	-
Case 2	XGBoost	-	-	-	**0.32**	**0.15**	**0.09**	-
Case 3	XGBoost	**0.21**	**0.08**	0.05	0.03	0.04	0.01	**0.11**

3.4.2. Elite Players' Analysis for Serie A

Out of the original dataset, only 38 players were chosen based on meeting the minimum goal requirement. Players needed to score at least 2 goals to qualify in the final dataset. The dataset's variability is indicated by its standard deviation, computed at 5.08 goals. Table 18 illustrates the best models per case for the top 30% of elite players and the corresponding metrics.

Table 18. Most accurate models for top 30% elite players per case for Serie A.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	MAPE Test
Case 1	MLP	1.69/2.90	5.98/17.25	2.45/4.15	0.45
Case 2	Gradient Boosting	1.66/2.00	4.88/6.57	2.21/2.56	0.47
Case 3	XGBoost	1.83/2.02	7.65/9.97	2.77/3.16	0.47

Referring to the provided table, gradient boosting results as the top-performing algorithm across all metrics. The strong consistency between the training and testing values indicates the algorithm's proficiency in handling unseen data. Additionally, it is noteworthy that the 'Previous Goals' feature emerged as the most influential, underscoring its importance in forecasting future performance. Feature importance values are not available for the MLP model. Table 19 depicts the feature importance for the most accurate models for the top 30% of elite players, where bold indicates the three most important features.

Table 19. Feature importance for the most accurate models for top 30% elite players per case for Serie A.

Cases	Best Model/ Features	Previous Goals	Nation	PrgP	xG	npxG	npxG+xAG
Case 1	MLP	-	-	-	-	-	-
Case 2	Gradient Boosting	**0.44**	**0.20**	**0.12**	-	-	-
Case 3	XGBoost	0.03	0.02	0.01	**0.13**	**0.12**	**0.09**

3.5. All Players Dataset Player Performance

This case contains all players from the four leagues. An extra column named 'League', has been added to reflect each player's original league.

3.5.1. All Players' Analysis for All 4 Leagues

The dataset consists of 424 players, with a computed standard deviation of 4.23 goals. Reflecting the dataset's diversity, the results indicate excellent performance by the models. XGBoost emerges as the best-performing algorithm, demonstrating the lowest values across all metrics, using all features. Table 20 presents the best models per case for all players' analysis.

Table 20. Most accurate models for all players per case for all four leagues.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	R^2 Test/R^2 Train
Case 1	XGBoost	1.69/1.66	6.68/6.63	2.58/2.57	0.51/0.65
Case 2	Random Forest	1.78/1.58	6.76/5.52	2.60/2.35	0.50/0.71
Case 3	XGBoost	1.67/1.51	6.48/5.19	2.55/2.28	0.52/0.72

Key influential features include expected goals and previous goals. Table 21 depicts feature importance for the most accurate models for all players per case, where bold indicates the three most important features.

Table 21. Feature importance for the most accurate models for all players per case for all four leagues.

Cases	Best Model/ Features	xG	xG per 90′	Previous Goals	PK	Pos
Case 1	MLP	**0.46**	**0.13**	**0.12**	-	-
Case 2	XGBoost	-	-	**0.56**	**0.08**	**0.07**
Case 3	XGBoost	**0.34**	**0.09**	**0.11**	0.02	0.01

3.5.2. Elite Players' Analysis for All 4 Leagues

In this case, the dataset was reduced to 157 players. This selection process revealed that players needed to score at least 2 goals to qualify for inclusion in the final dataset. The calculated standard deviation of the dataset is 5.18.

Table 22 illustrates the best models per case for the top 30% of elite players. The gradient boosting algorithm emerges as the top performer across all cases, with the best-performing algorithm being the one utilizing all features. Notably, the values of the training and testing datasets exhibit close alignment, indicating the absence of overfitting.

Table 22. Most accurate models for top 30% elite players per case for all four leagues.

Cases	Best Model/ Metrics	MAE Test/MAE Train	MSE Test/MSE Train	RMSE Test/RMSE Train	MAPE Test
Case 1	Gradient Boosting	2.30/2.44	11.90/10.76	3.45/3.28	0.47
Case 2	Gradient Boosting	2.31/2.47	11.04/10.83	3.32/3.29	0.48
Case 3	Gradient Boosting	2.28/2.32	11.17/9.42	3.34/3.07	0.48

The top three crucial features for case 3 include expected goals, non-penalty expected plus assisted goals, and expected goals per 90 min. Table 23 depicts feature importance for the best models for the top 30% of elite players, where bold indicates the three most important features.

Table 23. Feature importance for the most accurate models for top 30% elite players per case for all four leagues.

Cases	Best Model/ Features	xG	xG per 90′	npxG+xAG	Previous Goals	PK	PrgR
Case 1	Gradient Boosting	**0.39**	**0.18**	**0.16**	0.09	-	0.04
Case 2	Gradient Boosting	-	-	-	**0.65**	**0.08**	**0.07**
Case 3	Gradient Boosting	**0.36**	**0.10**	**0.16**	0.05	0.03	0.00

4. Discussion

This section comprehensively examines the outcomes of each league and its cases. The final section offers an in-depth comparison of these findings, providing valuable insights into the performance across the different datasets. Specifically, the XGBoost algorithm performed best with Serie A's dataset and attributes from case 1, featuring the 10 most correlated features related to the target variable 'Goals', with a MAE of 1.29. Next, we address the feature importance conclusions and the threats to the validity of our research.

4.1. Implications

Random forest proves to be the most effective algorithm within the Bundesliga dataset when utilizing the elite players dataset in case 2. Contrary to expectations, an algorithm trained on a significantly reduced dataset demonstrates superior results. Attributes in case 2 were selected by evaluating the correlation for each pair of features, with one feature chosen from each highly correlated pair. Notably, the feature with the highest importance value is 'Previous goals', indicating its critical role in predicting future outcomes. Additionally, 'Age' emerges as another significant feature, suggesting potential variations in player performance based on age. Another observation arises from the standard deviation calculation. With a standard deviation of 3.79 goals for the reduced dataset and a MAE of 1.71, the prediction range for a player scoring 10 goals would be 10+/-3 with an error margin of 1.71, indicating a good outcome.

In the analysis of the Premier League dataset encompassing all players, superior performance was observed in case 1 with the XGBoost algorithm. Features such as expected goals, non-penalty expected goals, and previous goals emerged as the most influential factors, highlighting their significant impact on the target variable. Expected goals represent

a statistical metric in football used to assess the likelihood of a goal being scored from a given shot. Furthermore, the disparity between the train and test metrics values suggests a minimal occurrence of overfitting.

Case 1, along with its associated attributes, was utilized to achieve the best results, employing the entire La Liga's dataset with MLP.

Once again, in Serie A league, the XGBoost algorithm in case 1 demonstrated the most optimal performance. It is worth mentioning that in this league, the values between the training and testing sets are closely related to those in other leagues.

Lastly, XGBoost emerged as the top-performing algorithm in case 3, including data from all four leagues. Conversely, for the reduced dataset, gradient boosting yielded the best metric errors. Expected goals played a pivotal role in both scenarios. As anticipated, the dataset containing all players exhibited superior results.

To better understand their distinction, we will provide an illustration involving one player from each league and their performances. These players were T. Müller (Bundesliga), D. Welbeck (Premier League), K. Benzema (La Liga), and N. Barella (Serie A).

Typically, using center backs or defenders as examples yields more precise predictions compared to forwards. This happens because defenders usually do not score in a season. Figure 2 summarizes the performance predictions for players from all leagues.

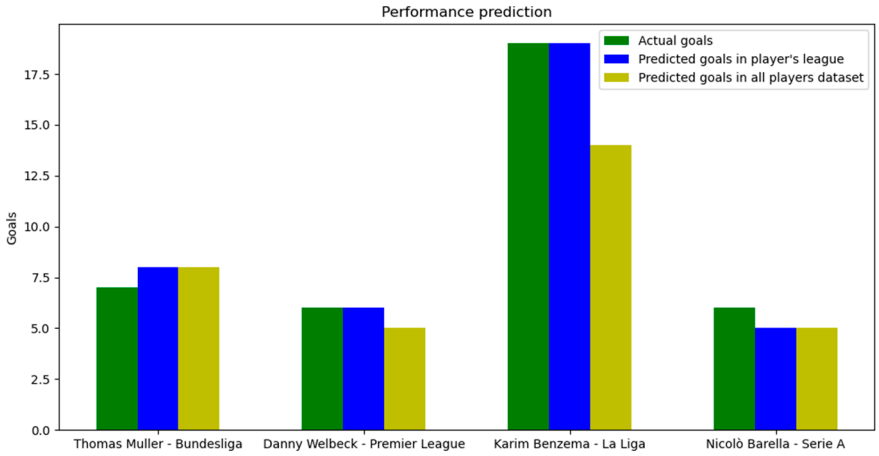

Figure 2. Performance prediction of players from all leagues.

T. Müller, a forward, scored a total of 7 goals in the 2023–2023 season. The random forest algorithm predicted 8 goals when applied to the elite players dataset, resulting in a 1-goal discrepancy. Similarly, the XGBoost algorithm, using data from all four leagues, also predicted 8 goals for Müller, aligning closely with the random forest prediction. For D. Welbeck, who is a midfielder, XGBoost precisely predicted the number of goals he scored in the final season using the league-specific dataset. In contrast, XGBoost, employing data from all four leagues, predicted 5 goals. In the instance of Karim Benzema, a striker who scored 19 goals, the MLP model accurately predicted the actual goals in the La Liga's dataset, while the XGBoost algorithm in the extended dataset predicted 14 goals, deviating from the actual count by 5 goals. Lastly, N. Barella, a skillful midfielder in Serie A, contributed 6 goals during the 2022–2023 season. XGBoost was the best algorithm in both datasets, predicting 5 goals.

Certainly, there are cases where the actual goals of players perfectly match the predictions made by all models. Conversely, there are also instances where the predicted goals for players show significant deviations from their actual achievements.

4.2. Comparative Insights across Leagues

As already mentioned, the comparison of algorithms across the different cases is a crucial aspect of this study. The results are summarized in Table 24, unveiling interesting insights.

Table 24. Performance results from all scenarios—comparative analysis.

League	Best Model/ Metrics	MAE	MSE	RMSE	MAPE	R^2
Bundesliga	Random Forest (case 2—30%)	1.71	4.38	2.09	0.29	-
Premier League	XGBoost (case 1—all)	1.93	10.35	3.22	-	0.53
La Liga	MLP (case 1—all)	1.72	6.91	2.63	-	0.48
Serie A	XGBoost (case 1—all)	1.29	3.96	1.99	-	0.48
All dataset	XGBoost (case 3—all)	1.67	6.48	2.55	-	0.52

Serie's A dataset depicted the lowest metric values compared to another league or the combined dataset. Among these metrics, MAE stands out as the most crucial metric, indicating the proximity of predictions to the actual number of goals. For instance, with XGBoost, MAE was recorded at 1.29, suggesting that if a player scored 10 goals, the prediction would fall within the range of 9 to 11 goals. While not all predictions achieve perfect accuracy, overall, they demonstrate high efficacy.

Moreover, it is noteworthy that only in the Bundesliga did an algorithm utilizing the reduced dataset yield superior results. Specifically, when the XGBoost algorithm was tested on the entire dataset across all leagues, MAE reached its second-lowest value. This underscores the notion that, despite cultural and gameplay differences among leagues, the comprehensive dataset generally produces more accurate predictions. Nevertheless, it is essential to recognize that the other error metrics do not exhibit similar trends.

Most significantly, the XGBoost algorithm emerges as the overall victor in three out of five scenarios, indicating its effectiveness across diverse datasets. Consequently, data scientists are advised to prioritize this algorithm when dealing with similar datasets. Furthermore, we validate key factors for each league regarding feature importance that contribute to achieving more accurate results.

Additionally, researchers must acknowledge the significance of their studies, as the results obtained surpass many previous endeavors. Although some studies may report marginally superior error metrics, it is crucial to acknowledge the challenge of comparing results across different datasets, given the substantial variations present in different sports and their dynamics.

4.3. Feature Importance

We discuss here the key features that consistently demonstrated high importance values and significantly contributed to the accuracy of our goal prediction models. Previous research [22] identified expected goals and previous goals as the most influential features in goal prediction.

Our feature analysis concluded that a player's goal-scoring performance is significantly related to previous goals, expected goals, and expected goals per 90 min. Previous goals represent a player's historical scoring performance, which is a reliable predictor of future goal-scoring potential. Expected goals (xG) provide a player's probability of scoring based on the opportunities presented to them. Finally, expected goals per 90 min (xG per 90′) is a normalized metric allowing for a fair comparison of players with variable minutes of presence on the pitch.

4.4. Threats to Validity

As previously mentioned, our experiments yielded high accuracy and good results. However, it is important to note several potential threats to the validity of this research.

One of the initial assumptions was to divide the dataset by picking the top 30% of athletes based on goal scoring. This judgment sought to assess the performance of algorithms on players whose actual goals were not zero. Positions like goalkeepers and defenders often have few scoring opportunities, making it easier for algorithms to anticipate

their scored goals for the 2022-23 season. By focusing on players with non-zero goals, we hoped to generate a more challenging and informative evaluation of the prediction models.

Another potential threat was the selection of variables. We considered three different scenarios regarding feature selection. In the first scenario, we selected the 10 features with the highest Pearson correlation to the target variable 'Goals'. In the second scenario, we calculated the correlation percentage for each pair of attributes and retained only one column from each highly correlated pair, resulting in a total of 13 features. The third scenario included all available features from the dataset.

5. Conclusions and Future Work

In this study, our primary objective was to predict the scoring performance of football players, meaning the total goals, using historical data. Data were scraped from 4 leagues: Bundesliga, Premier League, La Liga, and Serie A, reaching more than 5000 players originally for six seasons. Seasons 2018–2019 to 2021–2022 were used to train the models, while season 2022–2023 was used as the testing dataset.

We assessed the performance of six ML algorithms: linear regression, ridge regression, random forest, gradient boosting, XGBoost, and multilayer perceptron. We employed two versions of each algorithm, one using the entire dataset and another using the elite players (top 30% quartile). A further division was conducted based on the features utilized for training. The effectiveness of each model was evaluated through various metrics such as MAE, MSE, RMSE, MAPE, and R-squared.

The findings revealed that the XGBoost algorithms in 3 out of 5 categories outperformed other models and demonstrated higher accuracy. Specifically, the best results were found in Serie's A dataset, where the MAE was 1.29. It is evident that sports analytics will play a crucial role in the future, driven by the large volumes of data. Sport clubs will progressively have more data scientists to optimize player performance across metrics like physical fitness, technical skills like striking accuracy, and other aspects essential for maximizing on-field contributions.

In summary, this study provides significant knowledge for football clubs, managers, and coaches. It allows them to make better decisions and predict player performance, resulting in overall team improvement. Our research findings indicate the feasibility of accurately predicting a player's performance in the upcoming season based on historical data. However, further improvements should be made to obtain greater precision and efficacy.

Data scientists can explore various ways to improve their work, including leveraging more advanced and complex statistics or including statistics and using player statistics for every match of the season to enrich the dataset utilized for model training. Furthermore, insights from our study can be used to estimate a team's total goals by aggregating individual player performance. This will potentially offer information on the team's ranking prospects.

As discussed in another section, wearable devices or cameras can provide valuable insights into players' physical movements and conditions. These gadgets track a range of statistics, including heart rate and breathing patterns, which can improve goal-scoring performance analysis [23]. Additionally, analyzing Twitter data using sentiment analysis offers a novel approach to understanding the psychological factors influencing player performance [24]. This assists coaches and teams in decision-making and morale management [25]. Finally, injury analytics play a crucial role in optimizing player performance and reducing injury risks. Teams can leverage data on player fitness and movement patterns to enhance player well-being and maintain peak physical condition throughout the season [26].

Author Contributions: Conceptualization: C.M.; methodology: C.M. and G.P.; software: C.M. and G.P.; validation: C.M., C.T. and G.P.; formal analysis: C.M. and G.P.; investigation: C.M. and G.P.; resources: C.M., C.T. and G.P.; data curation: C.M. and G.P.; writing—original draft preparation: C.M. and G.P.; writing—review and editing: C.M., C.T. and G.P.; visualization: C.M.; supervision: C.T.; project administration: C.T. and C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available on GitHub at https://github.com/christinamarkopoulou/Diverse-Machine-Learning-for-Forecasting-Goal-Scoring-Likelihood-in-Elite-Football-Leagues (created on 9 July 2024).

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This manuscript is according to the guidelines and complies with the Ethical Standards.

Appendix A

Table A1. Hyperparameters of Bundesliga for the best models.

Algorithms/ Scenario	Case 2—All Players	Case 2—Elite Players
Random Forest	{'max_depth': 5, 'n_estimators': 300}	-
XGBoost	-	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.8}

Table A2. Hyperparameters of Premier League for the best models.

Algorithms/ Scenario	Case 2—All Players	Case 2—Elite Players
Gradient Boosting		{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200}
XGBoost	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200, 'subsample': 0.8}	-

Table A3. Hyperparameters of La Liga for the best models.

Algorithms/ Scenario	Case 1—All Players	Case 1—Elite Players
Random Forest	-	{'max_depth': 2, 'n_estimators': 100}
MLP	{'activation': 'identity', 'alpha': 5e-05, 'hidden_layer_sizes': (50,), 'solver': 'adam'}	-

Table A4. Hyperparameters of Serie A for the best models.

Algorithms/ Scenario	Case 1—All Players	Case 2—Elite Players
Gradient Boosting	-	{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 100}
XGBoost	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200, 'subsample': 1.0}	-

Table A5. Hyperparameters of All dataset for the best models.

Algorithms/ Scenario	Case 3—All Players	Case 3—Elite Players
Gradient Boosting	-	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200}
XGBoost	{'learning_rate': 0.01, 'max_depth': 4, 'n_estimators': 200, 'subsample': 0.9}	-

References

1. Morgulev, E.; Azar, O.H.; Lidor, R. Sports Analytics and the Big-Data Era. *Int. J. Data Sci. Anal.* **2018**, *5*, 213–222. [CrossRef]
2. Papageorgiou, G.; Sarlis, V.; Tjortjis, C. Evaluating the Effectiveness of Machine Learning Models for Performance Forecasting in Basketball: A Comparative Study. *Knowl. Inf. Syst.* **2024**, *66*, 4333–4375. [CrossRef]
3. Haq, N.F.; Onik, A.R.; Hridoy, M.A.K.; Rafni, M.; Shah, F.M.; Farid, D.M. Application of Machine Learning Approaches in Intrusion Detection System: A Survey. *Int. J. Adv. Res. Artif. Intell.* **2015**, *4*, 9–18. [CrossRef]
4. Papageorgiou, G.; Sarlis, V.; Tjortjis, C. An Innovative Method for Accurate NBA Player Performance Forecasting and Line-up Optimization in Daily Fantasy Sports. *Int. J. Data Sci. Anal.* **2024**. [CrossRef]
5. Pantzalis, V.C.; Tjortjis, C. Sports Analytics for Football League Table and Player Performance Prediction. In Proceedings of the 2020 11th International Conference on Information, Intelligence, Systems and Applications, Piraeus, Greece, 15–17 July 2020; pp. 1–8.
6. Zeng, Z.; Pan, B. A Machine Learning Model to Predict Player's 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.
7. Oliver, J.L.; Ayala, F.; De Ste Croix, M.B.A.; Lloyd, R.S.; Myer, G.D.; Read, P.J. Using Machine Learning to Improve Our Understanding of Injury Risk and Prediction in Elite Male Youth Football Players. *J. Sci. Med. Sport* **2020**, *23*, 1044–1048. [CrossRef] [PubMed]
8. Martins, F.; Przednowek, K.; França, C.; Lopes, H.; de Maio Nascimento, M.; Sarmento, H.; Marques, A.; Ihle, A.; Henriques, R.; Gouveia, É.R. Predictive Modeling of Injury Risk Based on Body Composition and Selected Physical Fitness Tests for Elite Football Players. *J. Clin. Med.* **2022**, *11*, 4923. [CrossRef] [PubMed]
9. Majumdar, A.; Bakirov, R.; Hodges, D.; Scott, S.; Rees, T. Machine Learning for Understanding and Predicting Injuries in Football. *Sports Med. Open* **2022**, *8*, 73. [CrossRef] [PubMed]
10. Pariath, R.; Shah, S.; Surve, A.; Mittal, J. Player Performance Prediction in Football Game. In Proceedings of the 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1148–1153.
11. Baboota, R.; Kaur, H. Predictive Analysis and Modelling Football Results Using Machine Learning Approach for English Premier League. *Int. J. Forecast.* **2019**, *35*, 741–755. [CrossRef]
12. Stübinger, J.; Mangold, B.; Knoll, J. Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. *Appl. Sci.* **2019**, *10*, 46. [CrossRef]
13. Sports Reference. Sports Reference. 2023. Available online: https://www.sports-reference.com (accessed on 1 September 2023).
14. Chatzilygeroudis, K.; Hatzilygeroudis, I.; Perikos, I. Machine Learning Basics. In *Intelligent Computing for Interactive System Design*; ACM: New York, NY, USA, 2021; pp. 143–193.
15. Poole, M.A.; O'farrell, P.N. The Assumptions of the Linear Regression Model. *Trans. Inst. Br. Geogr.* **1971**, 145–158. [CrossRef]
16. McDonald, G.C. Ridge Regression. *WIREs Comput. Stat.* **2009**, *1*, 93–100. [CrossRef]
17. Parmar, A.; Katariya, R.; Patel, V. A Review on Random Forest: An Ensemble Classifier. In Proceedings of the International conference on intelligent data communication technologies and internet of things (ICICI), Coimbatore, India, 7–8 August 2018; Springer: Cham, Switzerland, 2019; pp. 758–763.
18. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A Comparative Analysis of Gradient Boosting Algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [CrossRef]
19. Mo, H.; Sun, H.; Liu, J.; Wei, S. Developing Window Behavior Models for Residential Buildings Using XGBoost Algorithm. *Energy Build.* **2019**, *205*, 109564. [CrossRef]
20. Devadoss, A.V.; Ligori, T.A.A. Forecasting of Stock Prices Using Multi Layer Perceptron. *Int. J. Comput. Algorithm* **2013**, *2*, 440–449.
21. 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.
22. Giannakoulas, N.; Papageorgiou, G.; Tjortjis, C. Forecasting Goal Performance for Top League Football Players: A Comparative Study. In Proceedings of the Artificial Intelligence Applications and Innovations, León, Spain, 14–17 June 2023; Maglogiannis, I., Iliadis, L., MacIntyre, J., Dominguez, M., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 304–315.

23. Li, R.T.; Kling, S.R.; Salata, M.J.; Cupp, S.A.; Sheehan, J.; Voos, J.E. Wearable Performance Devices in Sports Medicine. *Sports Health Multidiscip. Approach* **2016**, *8*, 74–78. [CrossRef] [PubMed]
24. Chen, Z.; Kwak, D.H. It's Okay to Be Not Okay: An Analysis of Twitter Responses to Naomi Osaka's Withdrawal Due to Mental Health Concerns. *Commun. Sport* **2023**, *11*, 439–461. [CrossRef]
25. Dreyer, F.; Greif, J.; Günther, K.; Spiliopoulou, M.; Niemann, U. Data-Driven Prediction of Athletes' Performance Based on Their Social Media Presence. In Proceedings of the Discovery Science (DS), Montpellier, France, 10–12 October 2022; pp. 197–211.
26. Hecksteden, A.; Schmartz, G.P.; Egyptien, Y.; Aus der Fünten, K.; Keller, A.; Meyer, T. Forecasting Football Injuries by Combining Screening, Monitoring and Machine Learning. *Sci. Med. Footb.* **2023**, *7*, 214–228. [CrossRef] [PubMed]