Soccer Defender Analytics: A Comprehensive Framework for Quantitative Evaluation

Robert Velez, Joshua Brown, and Jonathan Locala

Johns Hopkins University Sports Analytics Research Group

Advised by Dr. Anton Dahbura and Tad Berkery

Abstract

Soccer, recognized globally as "The Beautiful Game," is celebrated by billions of people worldwide. Despite its popularity, soccer remains analytically underdeveloped due to inherent challenges such as low scoring frequency and limited availability of individual performance statistics. Traditional qualitative evaluations often overshadow data-driven insights, particularly regarding defensive roles, leading to their frequent undervaluation, compared to offensive counterparts, whose contributions are more visibly reflected in statistics. This imbalance is illustrated historically, with only five defenders ever receiving the Ballon d'Or. This project addresses the analytical disparity by developing a robust, data-driven framework to quantitatively rate and rank defenders. The approach incorporates advanced metrics, extensive feature engineering, and sophisticated machine learning techniques, aiming to provide a scalable and accurate evaluation system for player scouting and development. Several deliverables and insights result from this study, including an XGBoost model trained to deterministically assign ratings to defenders purely on data, a predictive model predicting defensive ratings for future seasons of defenders based on their previous performance and seasonal trends, and a redistribution of defender ratings aiming to make comparisons between defensive player ratings and non-defensive players fairer.

Materials, Methods, and Data

The project utilized a comprehensive dataset from FBRef encompassing defensive metrics for nearly 2,000 defenders from Europe's top five leagues. Metrics included tackles won, aerial duels percentage, ball recoveries, interceptions, blocks, errors leading to goals, and discipline-related statistics. To further see how some of these statistics interact with each other, we implemented feature engineering to further quantify the value of defenders. Feature engineering involved constructing new metrics such as Defensive Efficiency per 90 Minutes (sum of Tackles, Blocks, and Interceptions per 90 minutes), Tackle Contribution (Possession-winning Tackles added to the product of Tackle Success Rate and Total Tackles), Offensive Contribution (sum of Attacking Third Involvement, Crosses, and Shots per 90 minutes), Win Ratio, and Penalty Risk (sum of Yellow Cards, Red Cards, and Penalties per 90 minutes). Feature engineering allowed us to force the model to consider the intersection between these metrics, rather than to wait and "see if it occurred" when training it only on raw metrics. Additionally, considering metrics such as "Offensive Contribution" helped to distinguish defenders from midfield players in a way that promoted defensive behavior. As the aim of the model was to create a distribution of ratings where better performing defenders were rated higher than worse performing defenders, we did not deem it fair to allow the model to favor defenders who played in a more "offensive manner" and earned higher ratings as a result. While the value of specifically "defensive behavior" to a team's performance is debatable, more information about this is included in the "Discussion" section of this paper.

Data preprocessing included standardizing statistics relative to average playing time to mitigate bias toward frequently playing defenders. Defenders with fewer than five full 90-minute appearances per season were filtered out to enhance data reliability. One-hot encoding was implemented for categorical variables such as player positions, leagues, and squad information, enabling the model to effectively utilize categorical distinctions without assuming numeric relationships. However, information regarding squad information was later removed entirely to prevent overfitting due to the fact that each squad has a limited number of defensive players.

Principal Component Analysis (PCA) was initially employed to reduce the dimensionality of the dataset, simplifying the feature space and highlighting the most impactful metrics. This provided important insights into what features distinguished between defensive players most strongly. Analyzing these components further allowed us to also filter out over-identifying information from the dataset (such as squads). Additionally, an advanced scaling adjustment technique equalized rating distributions between defenders and other positional groups, increasing the number of valid training samples significantly from 1,977 to 3,521. This involved viewing non-defensive and defensive ratings both as normal distributions (by applying the Central Limit Theorem, as our sample size was large enough), determining the mean and variance of both distributions, and scaling defensive ratings to have the same mean and variance as non-defensive ratings. No adjustment was done to the mean (due to the distributions already having a statistically insignificant mean difference of 0.01), but the standard deviation of defenders was scaled from 0.21 to 0.28. This was done by calculating z-scores of each defender and multiplying them by 1.30 before recalculating the rating based on the mean and new z-score. We called this rating our “adjusted rating”. Our justification for this was that we are not treating this as a “reassigned rating”, but instead letting this number represent a defender’s rating if outliers were recognized to the same extent offensive outliers are. We recognize that they are not, and are not promoting these “adjusted ratings” for official use in any capacity. These are merely hypothetical ratings useful for analysis. By combining the similarly distributed datasets into one for training, we were able to then have data demonstrating high defensive performance and lower defensive performance (due to being offensive players), allowing the model to gain a better sense of what statistics represent a strong defensive playstyle, and to better assign ratings indicating such performance.

Machine learning models tested included Linear Regression, Ridge and Lasso regressions, Decision Trees, Random Forest, and Gradient Boosting (both a generic model and an XGBoost model). Hyperparameter optimization involved a comprehensive approach utilizing Grid Search and the Optuna library, systematically refining key parameters such as tree depth, alpha (L1 regularization), lambda (L2 regularization), number of estimators, subsample ratios, and column sampling per tree. A rigorous cross-validation strategy (80%-10%-10% training-validation-test splits) was applied to ensure generalizability and robustness. L1 and L2 regularization were both heavily emphasized in training to avoid aimlessly reducing the amount

of parameters considered (which occurred when only L1 regularization was present), rather than meaningfully limiting parameter weight to work against overfitting.

While the choice of using an XGBoost model was not made until testing, it has important parameters that should be explained here rather than in the "Results" section of this paper. The objective function that the model aims to minimize in training is regularized squared error. This means that the function minimizes the sum of 1) the average difference between expected and assigned ratings from the model, squared, 2) the quantity of the parameter λ_1 (lambda-one) multiplied by the sum of the absolute value of the weights of each feature, and 3) the quantity of the parameter λ_2 (lambda-two) multiplied by the sum of the square of the weights of each feature. The hyperparameter of max-depth refers to how many layers each tree involved in the XGBoost model could have. Learning rate refers to how much the model adjusted its parameters based on fitting to the training data after each iteration of training. Subsample value refers to the fraction of training samples that are used when training each tree. Colsample-by-tree value refers to the fraction of features that are used when training each tree. Regularization alpha refers to the L1 regularization parameter (λ_1) used in the equation for regularized squared error, and regularization lambda refers to the L2 regularization parameter (λ_2) used in the same equation. Minimum child weight refers to the minimum sum of weights of features allowed when determining the lowest node in a tree. This information comes from the XGBoost documentation page.

The primary measurement of model accuracy was R^2, but mean absolute error (MAE) was also calculated for each metric. This means that despite the model's goal being to assign its own ratings that would be "better" than existing ratings, accuracies relative to existing ratings were still calculated for reference. The justification for this seemingly contradictory method was to have confirmation that the model shared some degree of correlation to existing metrics, but did not copy them entirely. More explanation of this justification is provided in the "Discussion" section of this paper. The model's performance was also evaluated with a classification task. Defender ratings were split using the KBinsDiscretizer library, with a quantile split being directly employed to create 3 groups (which were not perfectly even due to defenders having ties in their ratings) based on ratings. High ratings ranged from 6.71 to 7.87, medium ratings ranged from 6.47 to 6.70, and low ratings ranged from 5.83 to 6.46. Note that the defender ratings used were the ratings resulting from the distributional scaling described earlier in this section. (Though this should not affect classification results, as the scaling only increased the spread of the ratings and maintained order, so the quantiles would have been the same.) From this, the model trained on player statistics and their assigned rating group, and made predictions for the rating groups of unseen players in testing.

The resulting adjustments and predictions derived from the Jupyter notebooks were plugged into the Tableau platform in order to create interactive, effective, and impactful visuals for slideshow, poster, and demo purposes. It is important to mention that clever usage and

leveraging of built-in Tableau features such as Calculation Fields, Parameters, and Filters made the visuals possible and interactive. Calculation Fields provided the basis for consistent sorting and a max number of players displayed, irrespective of season. Parameters were very influential toward the interactivity of the visuals, allowing for the smooth toggling between pre and post adjustment ratings. Finally, filters allowed for the seamless transition between each season and the aggregation of all seasons.

The visuals we decided to move forward with were horizontal bar charts, box plots, vertical bar charts, and scatter plots. The horizontal bar charts were key in representing dynamic visuals for the top 20 players (and defenders) pre and post rating adjustment for each of the 2021-2022, 2022-2023, and 2023-2024 seasons. The horizontal bar charts were also helpful in visualizing average defensive players ratings per season for each of the top five European leagues, which painted an interesting picture of a possible bias towards and against certain leagues. Moreover, the box plots were crucial in helping visualize the distributions per grouped position for the past three seasons. The box plots were especially effective at showing the change in outliers for defenders pre and post adjustment and how they more closely mimic those of forwards. Vertical bar charts were effective in showing the adjusted ratings against the predicted adjusted ratings for the top 10 defenders and midfielders. Finally, scatter plots were incredibly important in portraying the overall spread of the seasonal ratings of players per grouped position. The scatter plot was especially effective when toggling between pre and post adjustment ratings.

Results

The finalized descriptive XGBoost model achieved a cross-validated R^2 value of 0.51, demonstrating strong predictive capability and reliability. Tableau visualizations effectively illustrated defender rating distributions across three seasons (2021-2024), capturing consistency trends and performance fluctuations. The predictive XGBoost model using similar architecture as the descriptive model achieved a cross-validating R^2 value of 0.46, although this value is subject to further testing. (Another R^2 value of 0.59 was also found, but the potential source of overfitting for this value has not yet been determined, and more testing is required.) When using the descriptive model for a classification task (splitting the defender ratings into high, medium, and low groups), the associated R^2 was 0.62.

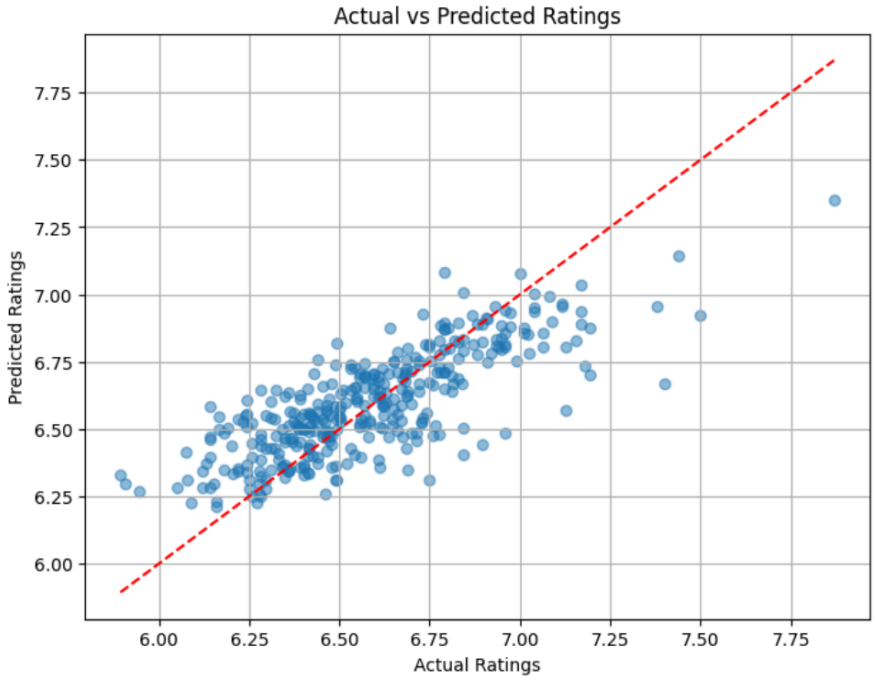

Figure 1: The Descriptive Model's Predicted Ratings Compared to Actual (Scaled) Ratings

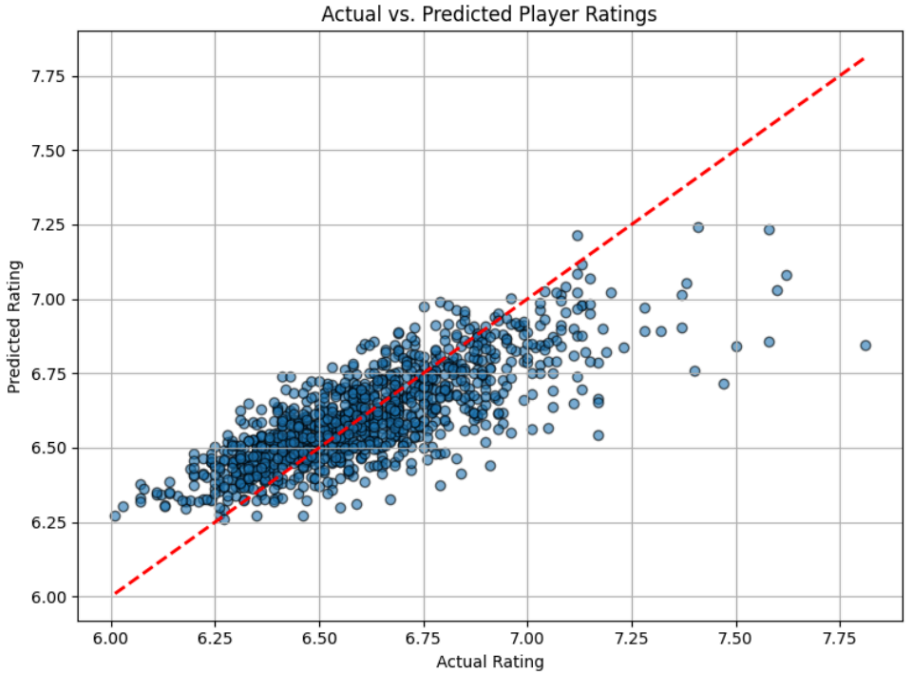

Figure 2: The Predictive Model's Predicted Ratings Compared to Actual (Scaled) Ratings

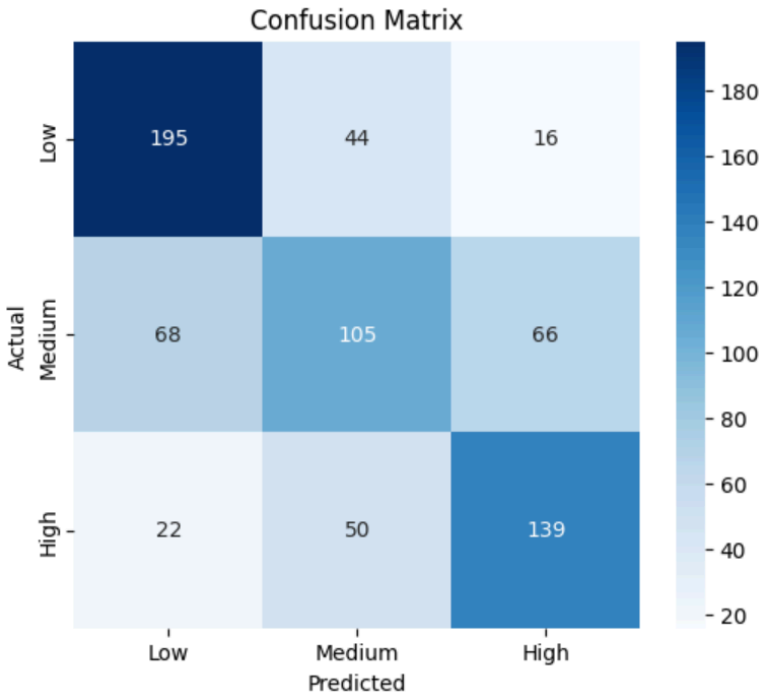

Figure 3: Confusion Matrix of Player Rating Classification of Descriptive Model

The optimal XGBoost parameters were minimization based on regularized square error, maximum tree depth of 3, learning rate of 0.061, subsample value of 0.618, colsample-by-tree value of 0.652, alpha regularization value of 0.630, lambda regularization value of 0.024, and a minimum child weight of 8. These values are explained in the "Materials and Methods" section of this paper. These parameters were determined when testing the descriptive model. However, when determining hyperparameters for the predictive model, these hyperparameters were (relatively) optimal, as when running the hyperparameter tuning process using the Optuna library, cross validation across varying different hyperparameter sets resulted in R^2 values as large as 0.46, which was the same R^2 the model reached when set up with the descriptive model's hyperparameters and trained on seasonal player data from 2021 to 2023, with testing done on data from 2023 to 2024.

Principal component analysis and clustering techniques provided deeper analytical insights, revealing distinct clusters of defenders characterized primarily by metrics such as clearances, tackles won, and interceptions. Feature importance analyses from the gradient boosting model further underscored the substantial predictive power of engineered metrics including Defensive Efficiency, Penalty Risk, and Tackle Contribution. The strongest indicators of higher defensive ratings were pass interceptions, tackles in the attacking third of the field, and number of tackles. Statistics involving tackles (e.g., percent of tackles that moved the ball away from the attacker, percent of tackles leading to a change in possession, etc.) tended to be the most distinguishing feature, as seen by the principal component analysis.

Discussion

The analytical framework effectively addresses the undervaluation of defenders by attempting to quantitatively capture their contributions. Traditional defensive metrics, focused on discrete, easily recorded actions, may inadequately represent overall effectiveness due to oversimplification of the events in a game. The metrics introduced here seek to bridge this gap by combining multiple traditional metrics into more holistic measures.

The descriptive model's regression R^2 of 0.51 was sufficient for the proceeding analysis of the model. An R^2 above 0.7 would be ineffective, as at that point, the model would be mimicking existing ratings rather than showing a moderate correlation, negating the intended effect of modeling ratings using machine learning techniques. An R^2 below 0.3 would also have been ineffective due to showing little enough correlation to existing ratings that no amount of analysis would suffice for proving the validity of the model's assigned ratings. While the predictive model's regression R^2 of 0.46 was slightly lower, it was sufficient for the same reasons, and also expected due to biases and non centered ratings particular to 2023-2024 ratings being unable to be accounted for.

Model training revealed that the XGBoost model would be the most promising for assigning our own ratings. It combined L1 and L2 regularization in a way that we were looking for, while also being very compatible with grid searching and hyperparameter optimization techniques. Its initial R^2 was also the initial highest of the different descriptive models we trained (starting at 0.45).

Outlier analysis highlighted that defenders typically exhibit more consistent, low-impact actions compared to attackers, whose contributions are generally more directly tied to goal-scoring and game-changing events. A defender may complete ten passes in a short period of time while building out from the back, whereas an attacker may go thirty minutes without touching the ball. However, each of these completed passes are low-impact events, whereas the one time an attacker touches the ball they may end with a shot on goal, significantly altering their ratings. Moreover, a turnover by a defensive player is considered game-altering because it gives the other team a direct chance to score, whereas an attacker turning the ball over does not hurt a team as much. These points emphasize the necessity of sophisticated analytical techniques and nuanced metrics for accurately evaluating defensive contributions, and explain why attacker ratings have significantly more outliers.

Questions regarding significance of certain statistics beyond a numeric value are significant, and were unable to be accounted for in the model directly. For example, a tackle leading to a change in possession that led to a goal compared to a tackle leading to a change in possession that did not lead to a goal would be viewed the same statistically based on inputs to the model. When training, the model may have seen the WhoScored metric that may have been slightly higher for the defender whose tackle led to a goal, but in testing, there is no way for the

model to examine the tackle's effectiveness beyond if it led to a goal or not. This problem can be attributed to the lack of data that is available for defensive statistics, and would be critical to address in future continuation of the research on the model.

Additionally, feature engineering involving calculated metrics such as "Offensive Contribution" are a potential source of bias. While initially added to improve the R^2 between WhoScored ratings and the model's assigned ratings, we recognize that this can be a potential bias towards certain playstyles that do not necessarily correlate to significant defensive performance. Players such as Trent Alexander-Arnold, who may have higher offensive contribution due to his style of playing a right-back position involving significant actions from the attacking third of the field, would be placed lower due to having less particular "defensive contribution", despite this significantly supporting his team. Again, having more specific statistics that address direct contribution of actions beyond possession changes would be necessary to try to correct for this source of bias, and this information is not available to us at this time. A future pursuit of research involving further development of the model with more advanced data would likely be needed to properly address this issue.

Additionally, it is important to discuss the relevance of the model at this point, as one could ask what gives our model any weight over existing rating methods, and for an explanation of what specifically is the motivation in our model's methods for assigning ratings. There are two primary justifications for our model's relevance. The first is that existing methods can be very subjective and can be based on how important the analyzers "feel" certain moves are. WhoScored starts with each player having a 6.0 rating and slightly increases or decreases (by an unknown amount) based on actions that WhoScored is relatively private about. By forming a model that predicts ratings based purely on known numbers, we decrease subjectivity so that we know what set of input defensive statistics lead to a particular rating output.

The second justification of relevance is that our model focuses more on individual statistics rather than team statistics. If a team performs extremely well against another team due to having a much stronger offensive line, it is unfavorable for defensive ratings to be extra high due to "effective actions seeming even better coming from a higher performing team", which is arguably an extension of the psychological winner effect (Heiss). It is also unfavorable for defensive ratings to be low because defenders are not acting very often if the ball is not in their half of the field for the majority of the game. So by purely looking at their stats on a measure of success in what they have accomplished (and comparing it against what defenders have attempted and how long they have played to try to mitigate the problem described in the previous sentence), our model tries to avoid bias from being put on well/poorly performing teams.

Conclusion

This research effectively develops and validates a comprehensive analytical framework for quantitatively assessing soccer defenders from both descriptive and predictive lenses. The

resulting defensive model aims to capture defender impacts and provide valuable insights applicable to understanding strong defensive performance, and the predictive model extends this analysis to aid in defensive scouting and talent identification for future seasons. Additionally, the distributional scaling of defensive players allows for improved visibility of defensive success, providing an interpretation allowing fans and coaches alike to better identify particularly high-performing defenders. Future work should focus on integrating tracking data, refining off-ball defensive metrics, and further developing predictive modeling capabilities, thereby strengthening and expanding the analytical foundations laid by this study.

References

"2023-2024 Big 5 European Leagues Defensive Action Stats (Players)." *FBref.com*, fbref.com/en/comps/Big5/2023-2024/defense/players/2023-2024-Big-5-European-Leagues-Stats. Accessed 19 Apr. 2025.

"2023-2024 Big 5 European Leagues Miscellaneous Stats (Players)." *FBref.com*, fbref.com/en/comps/Big5/2023-2024/misc/players/2023-2024-Big-5-European-Leagues-Stats. Accessed 19 Apr. 2025.

"2022-2023 Big 5 European Leagues Defensive Action Stats (Players)." *FBref.com*, fbref.com/en/comps/Big5/2022-2023/defense/players/2022-2023-Big-5-European-Leagues-Stats. Accessed 19 Apr. 2025.

"2022-2023 Big 5 European Leagues Miscellaneous Stats (Players)." *FBref.com*, fbref.com/en/comps/Big5/2022-2023/misc/players/2022-2023-Big-5-European-Leagues-Stats. Accessed 19 Apr. 2025.

"2021-2022 Big 5 European Leagues Defensive Action Stats (Players)." *FBref.com*, fbref.com/en/comps/Big5/2021-2022/defense/players/2021-2022-Big-5-European-Leagues-Stats. Accessed 19 Apr. 2025.

"2021-2022 Big 5 European Leagues Miscellaneous Stats (Players)." *FBref.com*, fbref.com/en/comps/Big5/2021-2022/misc/players/2021-2022-Big-5-European-Leagues-Stats. Accessed 19 Apr. 2025.

Bundesliga Player Statistics | Whoscored.Com, www.whoscored.com/Regions/81/Tournaments/3/Seasons/9649/Stages/22128/PlayerStatistics/Germany-Bundesliga-2023-2024. Accessed 19 Apr. 2025.

Bundesliga Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/81/tournaments/3/seasons/9120/stages/21026/playerstatistics/germany-bundesliga-2022-2023. Accessed 19 Apr. 2025.

Bundesliga Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/81/tournaments/3/seasons/8667/stages/19862/playerstatistics/germany-bundesliga-2021-2022. Accessed 19 Apr. 2025.

Heiss, Rebecca. "The Winner Effect: The Science of Success and How to Use It." *Dr. Rebecca Heiss*, 7 Feb. 2022, rebeccaheiss.com/the-winner-effect-the-science-of-success-and-how-to-use-it/.

LaLiga Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/206/tournaments/4/seasons/9682/stages/22176/playerstatistics/spain-laliga-2023-2024. Accessed 19 Apr. 2025.

LaLiga Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/206/tournaments/4/seasons/9149/stages/21073/playerstatistics/spain-laliga-2022-2023. Accessed 19 Apr. 2025.

LaLiga Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/206/tournaments/4/seasons/8681/stages/19895/playerstatistics/spain-laliga-2021-2022. Accessed 19 Apr. 2025.

Ligue 1 Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/74/tournaments/22/seasons/9635/stages/22105/playerstatistics/france-ligue-1-2023-2024. Accessed 19 Apr. 2025.

Ligue 1 Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/74/tournaments/22/seasons/9129/stages/21037/playerstatistics/france-ligue-1-2022-2023. Accessed 19 Apr. 2025.

Ligue 1 Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/74/tournaments/22/seasons/8671/stages/19866/playerstatistics/france-ligue-1-2021-2022. Accessed 19 Apr. 2025.

Premier League Player Statistics | Whoscored.Com, www.whoscored.com/regions/252/tournaments/2/seasons/9618/stages/22076/playerstatistics/england-premier-league-2023-2024. Accessed 19 Apr. 2025.

Premier League Player Statistics | Whoscored.Com, www.whoscored.com/regions/252/tournaments/2/seasons/9075/stages/20934/playerstatistics/england-premier-league-2022-2023. Accessed 19 Apr. 2025.

Premier League Player Statistics | Whoscored.Com, www.whoscored.com/regions/252/tournaments/2/seasons/8618/stages/19793/playerstatistics/england-premier-league-2021-2022. Accessed 19 Apr. 2025.

Serie A Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/108/tournaments/5/seasons/9659/stages/22143/playerstatistics/italy-serie-a-2023-2024. Accessed 19 Apr. 2025.

Serie A Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/108/tournaments/5/seasons/9159/stages/21087/playerstatistics/italy-serie-a-2022-2023. Accessed 19 Apr. 2025.

Serie A Player Statistics | Whoscored.Com, https://www.whoscored.com/regions/108/tournaments/5/seasons/8735/stages/19982/playerstatistics/italy-serie-a-2021-2022. Accessed 19 Apr. 2025.

XGBoost Parameters | XGBoost Documentation, https://xgboost.readthedocs.io/en/stable/parameter.html. Accessed 3 March 2025.