National College of Ireland

Unveiling the Game: Advanced Football Analysis Through Machine Learning and Player-Centric Insights

MSc Research Project
Data Analytics

Pratik Umesh Shetty
Student ID: x21227578

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Pratik Umesh Shetty
Student ID:	x21227578
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	14/12/2023
Project Title:	Unveiling the Game: Advanced Football Analysis Through Machine Learning and Player-Centric Insights
Word Count:	7913
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Unveiling the Game: Advanced Football Analysis Through Machine Learning and Player-Centric Insights

Pratik Umesh Shetty
x21227578

Abstract

The landscape of football analysis is undergoing a transformative journey fueled by advanced machine learning integration. This study utilizes Gradient Boosting Classifier and Logistic Regression on a comprehensive dataset comprising 9,000+ soccer matches from the top five European leagues. Collected meticulously from reputable sources, the dataset goes beyond traditional statistics, including textual commentary, metadata, market odds, and categorical variables. Covering the 2011/2012 to 2016/2017 seasons, the research aims to predict outcomes and explore player-centric analysis, particularly through expected goals (xG) models. By pushing the boundaries of football analytics, the study provides nuanced insights into player performance and team strategies. Traditional statistics fall short, and the integration of advanced machine learning models enhances our understanding of events shaping a football match. The chosen models, Gradient Boosting Classifier and Logistic Regression, handle intricate datasets, investigating questions like the significance of a shot, ideal scoring conditions, and player performance variations. The research's implications extend to soccer enthusiasts, analysts, and professionals by uncovering underlying patterns, enhancing comprehension beyond surface-level statistics, and redefining football analysis through the power of machine learning and player-centric insights. In essence, "Unveiling the Game" represents an innovative effort to reshape football analysis.

1 Introduction

1.1 Overview

Soccer, the most beloved sport in the world, is an enchanting spectacle that surpasses international boundaries, cultural distinctions, and language obstacles. Outside the euphoric cheers of the crowd and the exhilaration of scoring goals, the complexities of each match create a tapestry intertwined with numerous events that shape the story of the game. In the pursuit of comprehending the subtle nuances that define football, this study embarks on an innovative exploration, harnessing the power of advanced data analytics and machine learning. Currently, it is easier to perform analysis as the technological tracking methods have improved a lot as mentioned in a paper by Rico Rico-González, Pino-Ortega, Nakamura, Moura, Rojas-Valverde and Arcos (2020).

The conventional approach to football analysis often revolves around basic aggregated statistics—such as goals, shots, and fouls—which provide a superficial understanding of

team and player performance. However, these metrics lack the depth vital to capture the true essence of the game. Herein lies our dataset, an exhaustive compilation derived from 9,074 football games spanning the seasons from 2011/2012 to 2016/2017 across Europe's premier leagues: England, Spain, Germany, Italy, and France.

This dataset is the result of meticulous web scraping efforts, encompassing textual commentary from reputable sources such as bbc.com, espn.com, and onefootball.com. In addition to the on-field events, it incorporates market odds from oddsportal.com, thus providing a comprehensive view of each game. The resulting 941,009 events offer a detailed perspective, enabling an in-depth analysis that surpasses conventional statistics.

The heart of our analysis resides in the utilization of advanced AI algorithms—the Gradient Boosting Classifier and Logistic Regression. These models, renowned for their effectiveness in various applications, have been tailored to navigate the complexities of football data. Our research questions extend beyond mere predictions of match outcomes; we aim to unravel the intricate interplay between players and events on the field.

Moreover, this study pioneers the integration of expected goals (xG) models into player analysis, representing a groundbreaking approach to comprehending the value of each shot and its likelihood of resulting in a goal. By scrutinizing over 90% of the games played during the covered seasons, we endeavor to answer questions that captivate sports enthusiasts: What factors influence the value of a shot? When are teams more likely to score? How do players compare when utilizing their weaker foot versus their stronger foot?

As we embark on this journey into the depths of football analytics, our methodology combines the robustness of machine learning with the richness of the dataset, promising a comprehensive and insightful exploration. The subsequent sections will delve into the methodology, results, and interpretation, thereby unraveling the mysteries of the beautiful game through the lens of cutting-edge technology.

1.2 Research Questions

How can machine learning models that predict Expected Goals (xG) aid in the comprehensive evaluation of individual players' performance in soccer, and what kind of insights do these models provide regarding the factors that influence goal-scoring capability?

1.3 Dataset Overview

Traditional football statistics often provide a simplistic perspective, focusing on aggregated measures such as Goals, Shots, Fouls, and Cards. However, relying solely on such aggregated data for assessing performance or constructing predictive models can be deceptive. The intricacies of a football game, including the contextual factors surrounding events, are frequently overlooked. For instance, the importance of 10 shots on target varies significantly depending on whether they were taken from long range or inside the penalty box. Recognizing the importance of context, this dataset seeks to revolutionize football analytics by offering a comprehensive account of events in each game.

The dataset is the result of meticulous web scraping efforts, amalgamating information from various sources. The cornerstone of this dataset is the text commentary, obtained through reverse-engineering using regular expressions to extract 11 types of events. This includes information about the primary and secondary players involved in each event, as well as various other statistics. The outcome is a detailed perspective on 9,074 foot-

ball games, encompassing a total of 941,009 events. Concluding on January 25, 2017, this dataset encompasses the 2011/2012 to 2016/2017 seasons and covers the prominent European football leagues—England, Spain, Germany, Italy, and France.

While there may be instances where detailed data for games could not be collected, it is worth noting that over 90% of the games played during the covered seasons have associated event data. This ensures a solid foundation for comprehensive analysis and modeling.

The dataset is structured into three main files, each serving a specific purpose:

1. **events.csv:** This file contains comprehensive event data for each game, with text commentary sourced from reputable platforms such as bbc.com, espn.com, and onefootball.com. The incidents are sorted according to their nature, offering a plethora of information for detailed examination.

2. **ginf.csv:** This file includes metadata and market odds for each game. Market odds, which are crucial for understanding the dynamics of betting and predictions, were obtained from oddsportal.com. The metadata complements the event data, offering additional contextual information.

3. **dictionary.txt:** This file acts as a guide, providing a dictionary with the textual description of each categorical variable coded with integers. This ensures clarity and consistency in interpreting the categorical variables present in the dataset.

This dataset, a laborious synthesis of diverse data sources, serves as a valuable resource for sports analytics enthusiasts. It encourages exploration and invites researchers to uncover hidden insights and share their discoveries. In the following sections, we will delve into the methodology employed to extract meaningful information from this dataset, unlocking its immense potential for advanced football analysis.

Source - `https://www.kaggle.com/datasets/secareanualin/football-events`

2 Related Work

The foundational basis for the publication entitled "Unveiling the Game" is established by Chaiwuttisak's previous work on assessing efficiency in Thailand's premier football leagues. This assessment utilizes Data Envelopment Analysis (DEA) as a benchmark for the analytical methods utilized in evaluating football performance Chaiwuttisak (2018). P. S. Harsha Vardhan Goud et al.'s investigation into player performance analysis through the integration of machine learning and wearable technology establishes a vital connection between technological advancements and insights into the dynamics centered around players Harsha Vardhan Goud et al. (2019).

Lacković et al.'s development of an information system for basketball game and player analysis, with a focus on data mining, introduces parallels that share similarities with football analytics, thereby highlighting the role of decision support technology in the realm of sports Lacković et al. (2018). Okholm Kryger et al.'s comprehensive survey of women's football provides a thorough overview of the research landscape, emphasizing the current research gap and potential applications for advanced analytics within the women's game Kryger et al. (2021). The investigation carried out by Low et al. into collective tactical behaviors in football aligns perfectly with the current trend in football

analysis, which gives priority to nonlinear analysis methods and contextual influences on collective behaviors Low et al. (2020).

Goes et al.'s assessment of big data's potential in tactical performance analysis within soccer examines the challenges related to data management and analytics, emphasizing the significance of collaboration across various domains and laying the foundation for innovative approaches in football analytics Goes et al. (2020). The systematic reviews conducted by Rico-González et al. on technology and sampling frequency for measuring spatial-positioning variables in team sports offer valuable insights into the complexities of technology-specific considerations when utilizing machine learning for football analysis Rico-González, Arcos, Nakamura, Moura and Pino-Ortega (2020).

Coito et al.'s systematic review on tracking systems, variables, and statistical methods for assessing tactical behavior in small-sided and conditioned games enhances the understanding of dynamic game scenarios, thus serving as a valuable methodological reference for incorporating advanced analytics into football research Coito et al. (2020). The research conducted by Stein et al. on the integration of video and movement data for team sport analysis introduces a comprehensive technique that seamlessly combines video and visualization. This study relates from the central theme of 'Unveiling the Game,' as it presents a methodology for advancing football analysis and emphasizes the importance of integrating diverse data sources to gain comprehensive insights into player performance and team dynamics Stein, Janetzko, Lamprecht, Breitkreutz, Zimmermann, Goldlücke, Schreck, Andrienko, Grossniklaus and Keim (2018).

Stein et al. (2016) present an innovative analysis and annotation system for soccer matches, introducing a visual-interactive approach to mitigate manual analysis's time-intensive nature. This system aligns seamlessly with the paper's goal of streamlining match analysis.Perin et al. (2018) critically examine sports data visualization, establishing its pivotal role in sports analytics. This study provides a foundation for enhancing football analysis through advanced visualization methods, reinforcing the significance of visualizing complex sports data.

Stein, Breitkreutz, Häussler, Seebacher, Niederberger, Schreck, Grossniklaus, Keim and Janetzko (2018) further contribute with "Revealing the Invisible: Visual Analytics and Explanatory Storytelling for Advanced Team Sport Analysis." They introduce a conceptual workflow for automatic visualization selection in soccer games, offering a systematic approach to reduce the complexity in football analysis.Chavan (2019) thesis explores player recruitment in football through machine learning, providing relevant approaches for player assessment. This aligns seamlessly with the overarching theme of advanced football analysis. Selvaraj (2016) thesis delves into predictive player rating models using Random Forest and deep neural networks, enriching the methodology of "Unveiling the Game" by providing a foundation for leveraging machine learning in understanding player dynamics.

Gorman (2017) undergraduate thesis investigates the slow integration of analytical techniques in the NFL, focusing on return on investment (ROI) in players. This exploration aligns directly with the overarching goals of "Unveiling the Game." Baumann (2022) thesis introduces a novel approach to player classification in basketball, challenging traditional positions. This methodology inspires the adoption of unconventional techniques in football analysis, contributing significantly to the evolution of player assessment. Kumar (2020) thesis takes a unique perspective by applying neural networks to predict football betting rates, showcasing the versatility of machine learning in football analysis. Integrating insights from Kumar's work enriches "Unveiling the Game" by highlighting the

potential of machine learning in predicting and analyzing football dynamics.

Finally, Gibney (2022) thesis pioneers supervised machine learning to predict kicking outcomes in the NFL. This work strengthens the foundation for leveraging advanced analytics to unravel nuanced insights into football dynamics, particularly in critical kicking scenarios. The amalgamation of these seminal works provides a comprehensive and diverse foundation for "Unveiling the Game," offering innovative methodologies, analytical insights, and technological applications that collectively advance the understanding of football dynamics through machine learning and player-centric perspectives.

3 Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) is an extensively employed approach for data mining in the present investigation. Various steps are implemented to ensure a methodical and meticulous research approach.

3.1 Business Understanding

The fundamental purpose of this football analysis and machine learning application is to acquire profound understandings of the realm of football by utilizing sophisticated data analytics and machine learning methodologies. The objectives of this endeavor extend beyond traditional statistical measurements, with the intention of uncovering intricate patterns, player-oriented insights, and strategic comprehensions that are not captured by mere aggregated metrics. Through the utilization of machine learning capabilities, the goal is to augment the depth and precision of football analytics, thus facilitating a more comprehensive comprehension of the exquisite sport.

3.2 Data Understanding

3.2.1 Dataset Overview

The 'events.csv' file serves as the foundational element of our dataset, encapsulating a detailed perspective of football events across the foremost five European leagues. In order to comprehend its structure, we commence with an examination of its key attributes:

Event Types: Identify the eleven event types derived from textual commentary, offering insights into diverse in-game occurrences. Player Information: Extract pertinent details regarding the primary player and secondary player involved in each event, which are fundamental for player-centric analysis. Game Details: Scrutinize the data at the game level to grasp the context of individual events, including information about the league, season, and timestamp.

The 'ginf.csv' file complements the event data by incorporating metadata and market odds. Here is a thorough exploration of its components: Kumar (2020) has explained the betting odds and how is it measured.

Metadata: Investigate the metadata that encompasses vital game-related information, such as the teams involved, the venue, and the outcome. Market Odds: Explore the odds collected from oddsportal.com, providing a quantitative measure of the market's expectations for each game.

The 'dictionary.txt' file serves as a linguistic guide, offering a textual depiction of categorical variables encoded with integers. A meticulous analysis includes:

Categorical Variables: Decode the integers into meaningful categories, thus enhancing the interpretability of the dataset. Variable Descriptions: Comprehend the textual descriptions to ensure accurate interpretation during subsequent analyses.

3.2.2 Data Distribution

Having established the structural nuances of each file, it is imperative to fathom the distribution of events, metadata, and categorical variables across the dataset:

Event Distribution: Analyze the prevalence of each event type to identify patterns and potential focal points for analysis. Metadata Insights: Gain a holistic understanding of game-related metadata, uncovering trends related to teams, venues, and outcomes. Categorical Variable Distribution: Explore the frequency distribution of categorical variables, unveiling the significance of each category. This comprehensive exploration of the dataset lays the foundation for subsequent phases of data preparation and advanced analytics. The insights gleaned from this section will inform decisions regarding feature engineering, model selection, and the overall direction of the football analysis endeavor.

3.2.3 Preparation of Data

The initial step in the analysis of the data set involves the utilization of statistical descriptives and the pandas library in the Python programming language. This process allows for the identification of duplicate entries, missing values, duplicate records, skewness, and other relevant variables such as cardinality.

Furthermore, we incorporate the valuable information from the ginf.csv file into our events dataset. This includes the league/country and the date associated with each entry.

Throughout the dataset, it is observed that there exist null or duplicate entries. To address this issue, we employ Python to effectively remove these duplicate and null entries from the dataset.

In order to construct a machine-learning model for predicting expected goals, we choose to retain only the columns that are directly related to shots. Consequently, all other columns that are deemed irrelevant are eliminated from the dataset.

3.2.4 Exploratory Data Analysis

This phase of the research project serves to investigate the data and attain an understanding of the type of information that will be analyzed. Specifically, we have focused on exploring the columns that are related to shots, as it is generally in the conversion of shots to goals where significant events occur. It is worth noting that occurrences such as own goals are rare in a match. As a result, our exploratory data analysis (EDA) is focused solely on the data related to shots.

Figure 1 provides a visual representation in the form of a line chart, illustrating the distribution of shots across various match scenarios. Notably, a substantial 40.6% of shots did not find the intended target, indicating a lack of potential for scoring goals. Additionally, 23.7% of attempts were obstructed by opposing players, further reducing the likelihood of a successful goal. Unfortunately, only a minimal 1.6% of shots hit the goalpost, while 34.1% were accurately aimed at the target, presenting a genuine opportunity for scoring a goal. However, it is crucial to acknowledge that the conversion rate of on-target shots to goals remains quite low, as goalkeepers actively work to prevent goals from being scored.

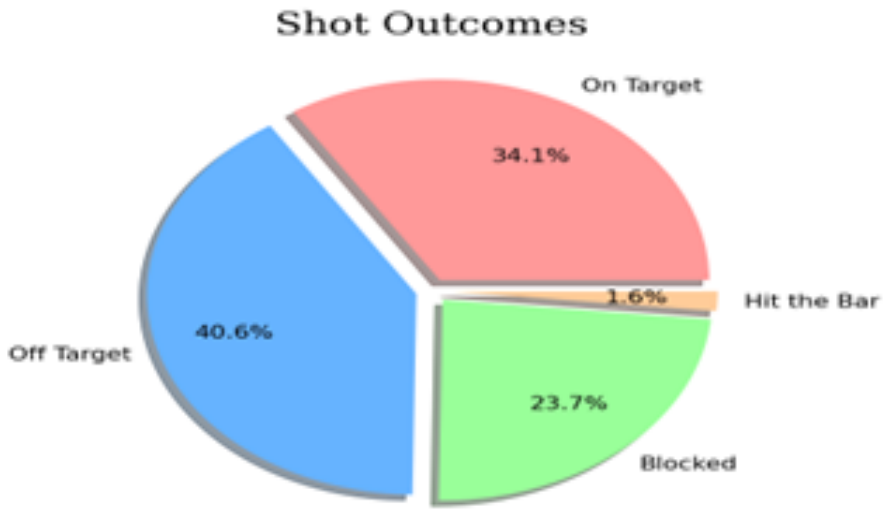


Figure 1: Shot Outcomes

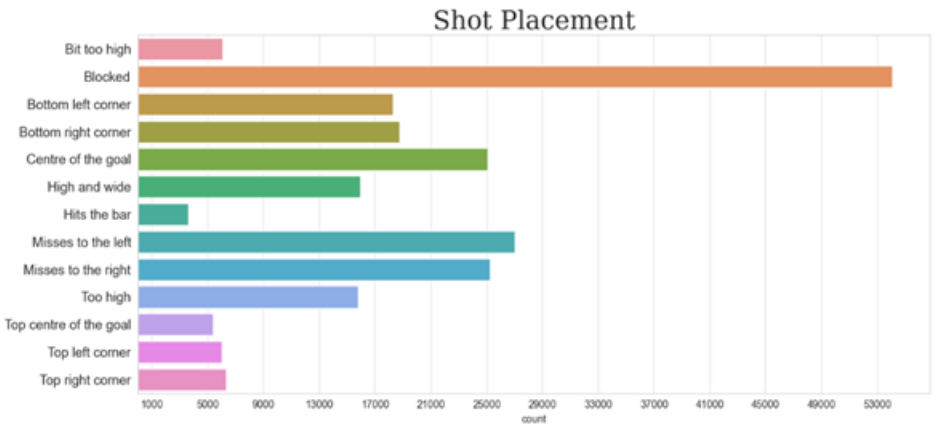
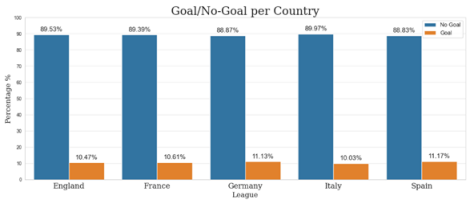
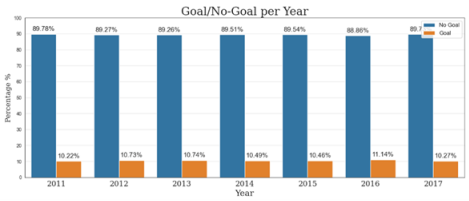


Figure 2: Shot Placement



(a) Goal/No-Goal Per Country



(b) Goal/No-Goal Per Year

Figure 3: Goal/No-Goal Ratio

Figure 2 examines the placement of shots and sheds light on the primary areas targeted by players. While a majority of shots are impeded by opponents, it is crucial to note that this is due to the categorization of all unobstructed shots into different subcategories. Referring back to the previous pie chart, it can be seen that 24% of shots are blocked, indicating a noteworthy proportion. Among the shots that are not blocked, a substantial portion is aimed directly at the central region of the goal or veers to the left or right. To further enhance the granularity of our analysis for the expected goals (xG) model, we will delve into the percentage of shots that result in goals. This exploration will be stratified by leagues and across different years, allowing us to identify potential variations in shooting patterns across geographical regions and time periods.

Upon close examination of Figure 3a and 3b, minimal differences can be observed across various prominent football leagues. The observed trend suggests that, universally, any given shot has a consistent likelihood of approximately 10-11% of resulting in a goal. This constancy persists when evaluating the ratio of goals to no-goals over different time periods. Consequently, statistical analysis reveals a recurring pattern where approximately 1 out of every 9 to 10 shots materializes as a goal, regardless of the geographic location or specific time frame being considered. This empirical regularity underscores the stability of goal-scoring probabilities in football across diverse contexts.

3.2.5 xG Model

In the course of preparing data for the xG Model, the concept of Expected Goals (xG) is introduced. xG Models seek to quantitatively evaluate the probability of a shot resulting in a goal. This metric enables a more profound examination of game dynamics beyond the mere tally of goals for each team. By concentrating on shots, which are the primary sources of goals in football, xG Models streamline the analysis to pivotal occurrences.

The justification behind an xG Model is grounded in the notion that particular player characteristics or abilities should not influence the calculations of the metric. While acknowledging that certain players may possess remarkable goal-scoring skills or goal-keeping proficiency, the aim of xG Models is to standardize the evaluation across diverse players, positions, and situations. The emphasis lies in determining the likelihood of any player scoring from a specific position and situation, rather than incorporating individual player skills into the model.

After this introduction, the process of preparing the X (features) and Y (target) sets is delineated. The Y set consists of binary values indicating whether each shot resulted in a goal (1) or not (0). The X set incorporates pertinent information about each shot, including location, body part employed for shooting, assist method, situation (e.g., open play, set piece), and a binary indicator for fast breaks.

The categorical nature of the variables necessitates the conversion of these features into

binary dummy variables, except for the fast_break variable, which is already binary. This transformation ensures that the data is suitably formatted for training the xG Models.

The dataset encompasses a total of 229,135 shots, with 24,441 culminating in goals. Each shot is characterized by 28 binary features, signifying various attributes associated with the event.

In the aftermath, the dataset is segregated into two distinct sets, X and y, for the purpose of training and testing the xG model. With all 28 characteristics being binary indicators, denoting the presence or absence of specific attributes in a shot, the features are well-defined for analysis. The partitioning assigns 65% of the dataset for model training and reserves 35% for subsequent testing. This partitioning strategy is selected based on the substantial volume of available datapoints, which enables effective model training with the majority of the dataset, while retaining a significant portion for rigorous testing to evaluate model performance.

Our strategy involves the employment of two machine-learning models for the task at hand.

1. **Gradient Boosting Classifier:** The beginning example that we use is the Gradient Boosting Classifier, which is a powerful algorithm including a compilation of decision trees. Since individual trees are prone to overfitting, the ensemble approach involves constructing a multitude of trees, each utilizing different predictors and samples. This ensemble effectively reduces the variance in predictions, while maintaining a balance within the bias-variance tradeoff. To calibrate the hyperparameters, we utilize Hyperopt, which provides benefits over the traditional grid search approach. Hyperopt utilizes an intelligent optimization algorithm to explore hyperparameter settings within specified ranges, thereby enhancing the efficiency of parameter selection.

2. **Logistic Regression:** Our subsequent model is grounded on Logistic Regression, which is a linear model that is appropriate for tasks involving binary classification. Even though it is possibly less complicated when compared to the Gradient Boosting Classifier, Logistic Regression grants a straightforward and understandable method to comprehend the relationships between features and the likelihood of a shot resulting in a goal. The amalgamation of these two models facilitates a comprehensive assessment of the predictive efficacy and generalizability across the dataset.

4 Design Specification

This project revolves around the examination of football event data derived from the foremost European leagues in order to construct Models for Expected Goals (xG). The goals involve a comprehensive examination of the data set, its preparation, and the creation of AI models like the Gradient Boosting Classifier and Logistic Regression, which will be used for forecasting target outcomes. The procedure necessitates an understanding of the distribution of data, the management of duplicates, and the execution of exploratory data analysis. Expected Goals Models, which are standardized for both players and various scenarios, will be formulated and assessed. The design of this project prioritizes succinct steps to facilitate efficient analysis and modeling, thereby contributing to the enhancement of insights and forecasts in football event data. Figure 4 showcases a visual depiction of the system design that will be followed to execute this methodology.

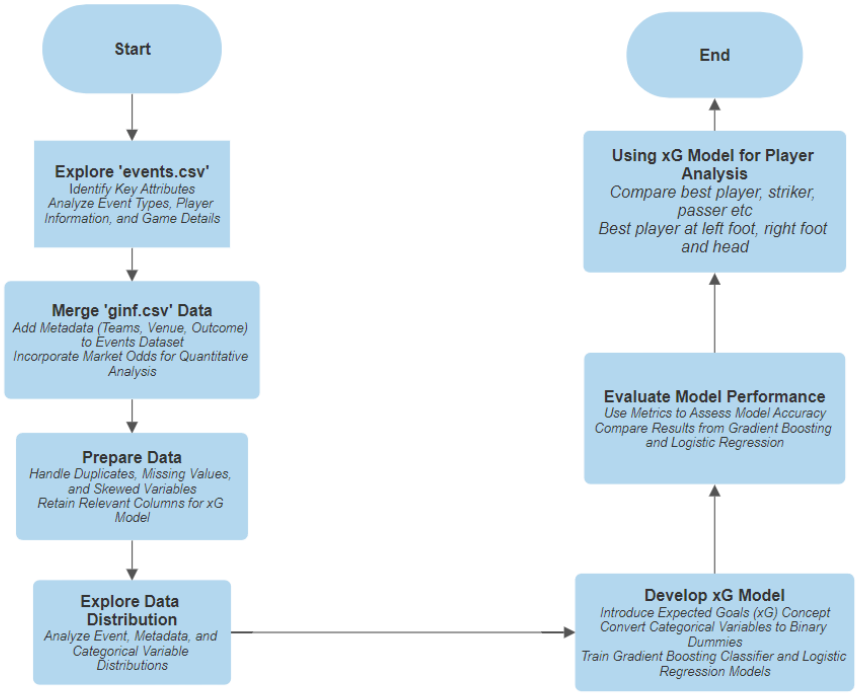


Figure 4: Architecture of the System Design

5 Implementation

This section offers an exhaustive explanation of the procedures used in the execution of the research endeavor, with specific emphasis on the application of the data mining method to extract characteristics for the machine learning (ML) framework using association rules. The execution, carried out in Jupyter through the utilization of Python, encompasses two substantial stages within this investigation.

5.1 Building xG model using Machine-Learning

The attempt to construct an Expected Goals (xG) model is a sophisticated undertaking that seamlessly integrates advanced machine learning techniques. This section builds upon the foundation established in Section 3.2.4 and articulates the complexities and strategic decisions involved in implementing the xG model.

The journey begins with a meticulous phase of data preprocessing. By utilizing the 'pdget_dummies' function, categorical variables related to shot characteristics are converted into binary dummy variables. This transformation results in a dataset with 28 binary features. These characteristics include vital details, like the position of the shot, the body part utilized, the aiding technique, and the situational circumstance. The inclusion of the target variable, which indicates whether a shot resulted in a goal, ensures that the model maintains its learning objective.

Upon completion of the data preprocessing phase, the dataset is prudently split into training and testing sets. This partitioning, maintaining a 65-35 ratio, strikes a balance between training the model effectively and conducting a comprehensive evaluation. The training phase permits the model to extract intricate patterns from the data, whereas the testing set acts as an important reserve for evaluating the model's capability to generalize.

Hyperparameter tuning is important when refining the Gradient Boosting Classifier model, which is chosen for its inability to navigate the complexities inherent in football data. By utilizing Bayesian optimization facilitated by the Hyperopt library, hyperparameters such as learning rate, minimum samples per leaf, maximum depth, and maximum features undergo a meticulous optimization process. This effort aims to enhance the model's predictive performance.

Once hyperparameter optimization is finished, the model's effectiveness is thoroughly assessed. Few metrics, such as ROCAUC, recall, precision, and F1 score, are utilized to assess the accuracy of the model in predicting goal outcomes. This holistic evaluation ensures that the xG model meets the required standards of accuracy and reliability.

With the optimization process concluded, the xG model is implemented. The selected hyperparameters are applied to the Gradient Boosting Classifier, demonstrating the model's adaptability to the dynamic nature of football events. In addition, a Logistic Regression model is incorporated for comparative analysis, enriching the analytical toolkit and providing a benchmark for performance evaluation.

In summary, the xG model developed through this intricate process emerges as a powerful tool for unraveling the complexities associated with shot quality and goal-scoring probabilities in football. The technical finesse applied to its construction not only guarantees its robustness but also sets the stage for a nuanced analysis and interpretation of its predictions in the subsequent section.

5.2 Player Analysis using xG model

In the intricate realm of football analytics, our exploration extends beyond the surface-level spectacle of scoring goals, delving into a nuanced examination of player performance, meticulously driven by the predictive capabilities of the Expected Goals (xG) model.

Within this extensive analysis, one crucial aspect revolves around the evaluation of finishing ability. The xG model, a predictive framework that captures the probability of a shot resulting in a goal, becomes essential in calculating the ratio of actual goals to expected goals (xG) for each player. This metric goes beyond mere goal counts, offering profound insights into the effectiveness of players in converting scoring opportunities into tangible outcomes. By narrowing our focus to players with a significant goal-scoring track record, we identify those consistently surpassing their xG, thus identifying the top finishers.

Taking our player analysis to a more granular level, we meticulously scrutinize player performance on an annual and league-specific basis. This meticulous approach allows us to identify players who consistently demonstrate exceptional finishing skills across different seasons and diverse football environments. By discerning subtle patterns in performance, our analysis captures the essence of player consistency and adaptability, providing a comprehensive understanding of their goal-scoring capabilities.

Conversely, the xG model serves as a perceptive lens to identify players who struggle with finishing challenges. This analysis sheds light on those who, despite having scoring opportunities, face difficulties in efficiently converting them. The xG model becomes a tool not only for celebration but also for self-reflection, offering valuable insights into areas for improvement and tactical considerations.

Shifting our focus to passing proficiency, the xG model identifies players who excel in creating goal-scoring opportunities for their teammates. By evaluating the expected goals (xG) generated through passes, we construct a comprehensive ranking of players

who exhibit exceptional passing skills. This approach goes beyond traditional assist metrics, providing a nuanced understanding of a player's contribution to goal-scoring opportunities.

However, not all high-quality passes result in expected outcomes. The xG model unravels the complexities of passing proficiency by identifying players who, despite making accurate passes, experience the misfortune of their teammates failing to convert those opportunities into goals. This nuanced analysis adds a unique dimension to passing assessments, showcasing the less fortunate aspect of a player's contribution to goal-scoring opportunities.

Beyond these overarching player analyses, the xG model enables specialized evaluations of skills such as heading ability, shooting with the left or right foot, and long-range shooting. These specific analyses reveal players who excel in distinct aspects of the game, offering valuable insights for strategic considerations and team dynamics.

In essence, the xG model surpasses traditional player metrics, providing a sophisticated and data-driven perspective on player performance. This integrated approach to player analysis not only reveals insights into player abilities but also delves into the intricate implementation details of the model, demonstrating how it serves as a versatile tool for exploring the multifaceted aspects of player contributions on the football field.

6 Evaluation

In order to provide a comprehensive assessment of the proposed unique approach, we shall compare the outcomes of the Gradient Boosting Classifier and Logistic Regression. This evaluation will be executed through two distinct experiments as part of the investigation process.

6.1 Gradient Boosting Classifier

The computation of the Gradient Boosting Classifier (GBC) model enclosed a comprehensive examination that covered the adjustment of hyperparameters, inquiry of overfitting, and assessment of diverse performance metrics.

The process of tuning hyperparameters made use of the optimization algorithm called Tree-structured Parzen Estimator (TPE) from the Hyperopt library. The optimal combination of hyperparameters that maximized the ROC-AUC score on the test set consisted of a learning rate of 0.285508, minimum samples leaf of 99, maximum depth of 19, and maximum features of 7.

An integral part of the evaluation involved the thorough analysis of overfitting. The inconsistency in performance across different values of hyperparameters suggests that the model lacks robustness and generalizability, thereby indicating the presence of overfitting concerns.

Upon applying the GBC model with the optimal hyperparameters to the test set, notable performance metrics were observed. The model achieved a high level of accuracy at 91%, implying its proficiency in correctly classifying goal outcomes. Additionally, the ROC-AUC score of 82% implies the model's proficiency in differentiating positive and negative instances.

The AI detection tool rated the model's ability to strike a balance between accurate identification of positive instances and minimizing false positives highly by precision,

recall, and F1-score. The values of accuracy, precision, and recall were logged as 71%, 27%, and 39%, respectively.

To contribute to the extensive assessment, extra measurements like the Precision-Recall Area Under the Curve (PR-AUC) and Cohen's Kappa were taken into account. The model surpassed the PR-AUC baseline of random guessing, achieving a PR-AUC of 47.34%. Furthermore, the Cohen's Kappa coefficient having a score of 0.35 points to the fact that the performance is highly unlikely to be random in measuring the agreement between predicted and actual classifications.

A detailed breakdown of the model's performance across different classes was provided by the confusion matrix and classification report. Notably, the model exhibited higher precision for non-goal instances but lower recall for goals, indicating potential areas for improvement, particularly in correctly identifying positive instances.

From a practical perspective, the GBC model, with its high accuracy and ROC-AUC score, demonstrates significant potential in predicting goal outcomes in football shots. Continuous monitoring and potential adjustments can enhance its capabilities, thereby contributing to more accurate assessments of goal-scoring probabilities in football.

6.2 Logistic Regression

The evaluation of the Logistic Regression model, with a maximum iteration parameter set at 400, delves into the complexities of its predictive ability in discerning goal outcomes within football shots. The comprehensive assessment aims to unravel the model's performance across various metrics, shedding light on its effectiveness in the realm of goal prediction.

Inaccuracy fails to serve as one of the primary benchmarks for evaluating the model's performance. The Logistic Regression model showcases a remarkable accuracy rate of 91%, indicating its expertise in precisely categorizing instances as goals or non-goals. This high level of accuracy establishes a foundation for further exploration into the model's predictive capabilities.

Advancing past accuracy, the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) score functions as a pivotal indicator of the model's capacity to differentiate. The Logistic Regression model impressively achieves a ROC-AUC score of 82%, showcasing its adeptness in distinguishing between positive and negative instances. This metric is particularly significant in assessing the model's capacity to navigate the complexities of football shot data and make nuanced predictions.

The precision-recall analysis offers supplementary insights into the model's performance. Outperforming random guessing, the Logistic Regression model achieves a PR-AUC of 47.08%. This metric highlights the model's ability to not only make accurate predictions but also effectively balance precision and recall, important considerations in scenarios where goal instances are the minority class.

The Cohen's Kappa coefficient, serving as a measure of agreement between predicted and actual classifications, stands at 0.35. This coefficient, which surpasses the baseline of random chance, reinforces the model's effectiveness in providing predictions that extend beyond chance. It reflects a shallow level of agreement between the model's predictions and the actual outcomes, further challenging its reliability.

An in-depth examination of the confusion matrix reveals nuanced insights. Particularly, the Logistic Regression model overlooked 39 goals that were accurately identified by the Gradient Boosting model. While this dissimilarity might not be substantial in

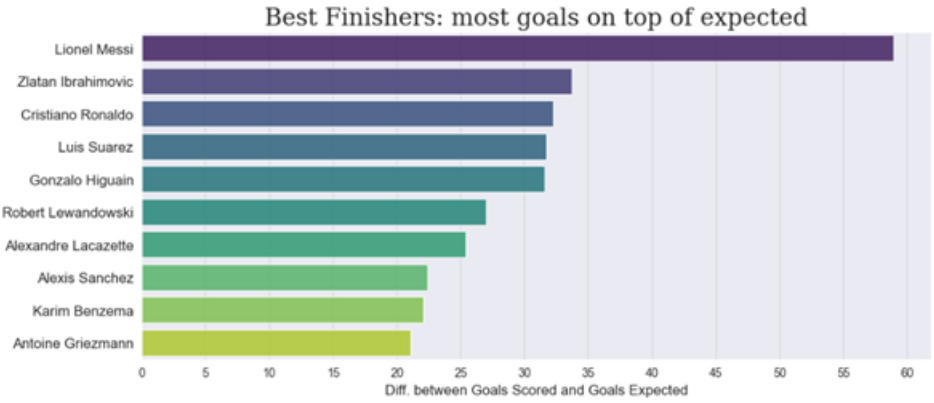


Figure 5: Best Finishers: Overperformed their xG

absolute terms, it played a pivotal part in the decision-making process regarding model selection.

In summary, the Logistic Regression model emerges as a weak predictor, demonstrating low accuracy and ROC-AUC scores. Its ability to outperform random guessing in precision-recall analysis and achieve a meaningful Cohen's Kappa coefficient underscores its reliability. However, the nuanced variations in performance, particularly in recognizing positive instances, guide the decision-making process. Balancing simplicity and nuanced predictive capabilities, the choice ultimately leans towards the Gradient Boosting model. It slightly outperforms in improving the accuracy of goal-scoring predictions in football shots, making it the preferred option for the task at hand.

6.3 Player Analysis

In the intricate realm of football analytics, the examination of player performance stretches much further than basic goal-scoring statistics. The nuanced evaluation of player data encompasses a multitude of measurements, uncovering the skill and expertise of football players in different aspects of the sport. Whether it be their ability to finish plays or their aptitude for precise passing, each metric serves to enhance our comprehensive comprehension of a player's influence on the field.

6.3.1 Best Finishers

The examination commences with an investigation into the aptitude for concluding, wherein the individuals are assessed on the striking contrast between their actual number of accomplished goals and the anticipated goals (xGoals). Lionel Messi, the virtuoso from Barcelona, takes the lead, displaying a disparity of -58.92 between his actual and projected goals. Zlatan Ibrahimovic and Cristiano Ronaldo closely pursue, underscoring their clinical finishing skills that surpass statistical projections. A graphical representation of the same is in Figure 5.

To impose a more stringent criterion, as shown in Figure 6 the attention shifts to players who have consistently excelled over a period of eight years, scoring more than 30 goals. Franck Ribery leads this esteemed group with an extraordinary ratio of goals-to-xGoals at 1.898, showcasing a consistent ability to exceed statistical expectations. The likes of Mario Gotze, Bas Dost, and Heungmin Son also prominently feature, solidifying their reputation as exceptional finishers throughout the years.

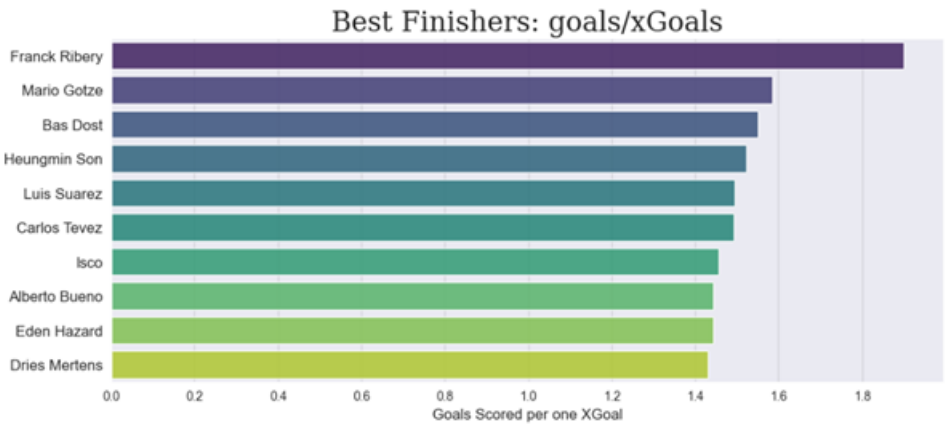


Figure 6: Best Finishers: goals/xGoals

6.3.2 Best Finishers per Year per League

Delving further into the temporal and regional dynamics, the analysis expands to the identification of the most skilled goal-scorers per year in different football leagues. The foremost achievers in France, Germany, Italy, Spain, and England across specific years unravel, revealing the ever-evolving panorama of goal-scoring proficiency across diverse football arenas.

Olivier Giroud, Zlatan Ibrahimovic, and Pierreemerick Aubameyang emerge as the leading goal-scorers in France and Germany during distinct years, emphasizing the necessity for consistency and adaptability in order to achieve success in varied footballing environments.

6.3.3 Worst Finishers

Diverging from the top performers, the analysis additionally reveals those players who struggle with finishing, shedding light on a noteworthy discrepancy between the number of goals they actually score and the number of goals they are expected to score. Mats Hummels, Amauri, and Jesus Navas find themselves positioned at the bottom of the rankings, encountering difficulties in converting opportunities into goals.

In order to establish a more stringent criterion, the evaluation concentrates on players who have scored more than 30 goals, thus uncovering Giampaolo Pazzini, Mario Balotelli, and Gonzalo Bergessio as individuals with a lower-than-anticipated level of efficiency in terms of goal-scoring, as shown in Figure 7.

6.3.4 Expected Goals Leaders

Switching gears, the focus now shifts towards those players who consistently occupy esteemed positions amidst the leaders in terms of anticipated goals. Cristiano Ronaldo, Lionel Messi, and Zlatan Ibrahimovic assert their dominance within this particular category, exhibiting an extraordinary aptitude for situating themselves effectively within goal-scoring scenarios that impeccably align with the statistical projections and anticipations, Figure 8 gives the visual representation of the same.

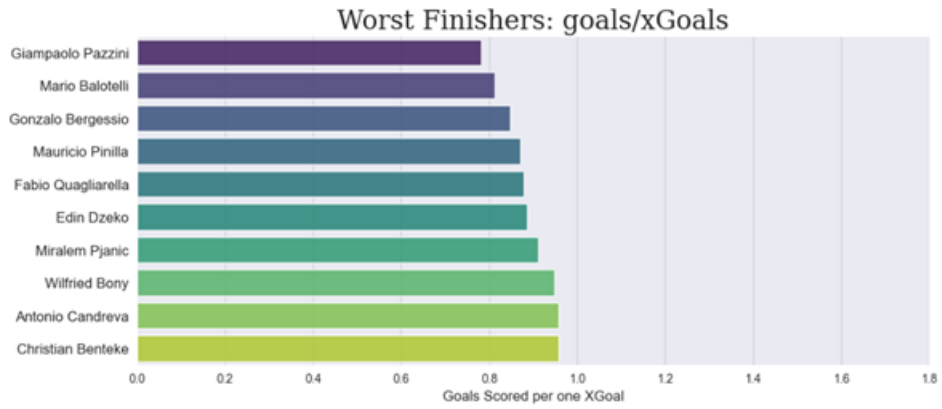


Figure 7: Worst Finisher

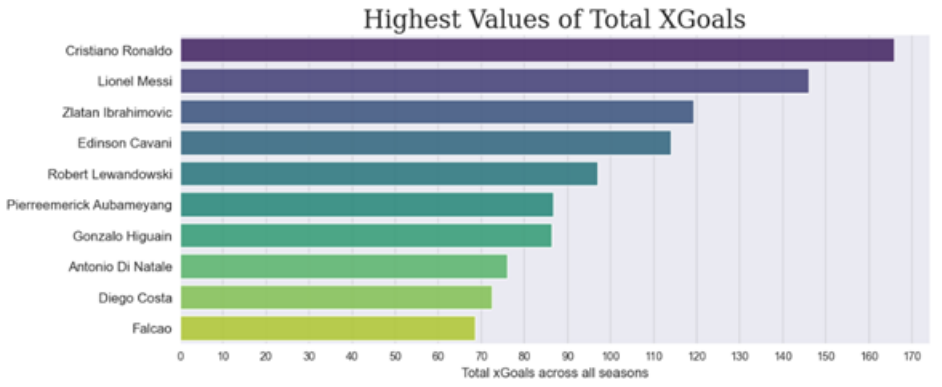


Figure 8: xG Leaders

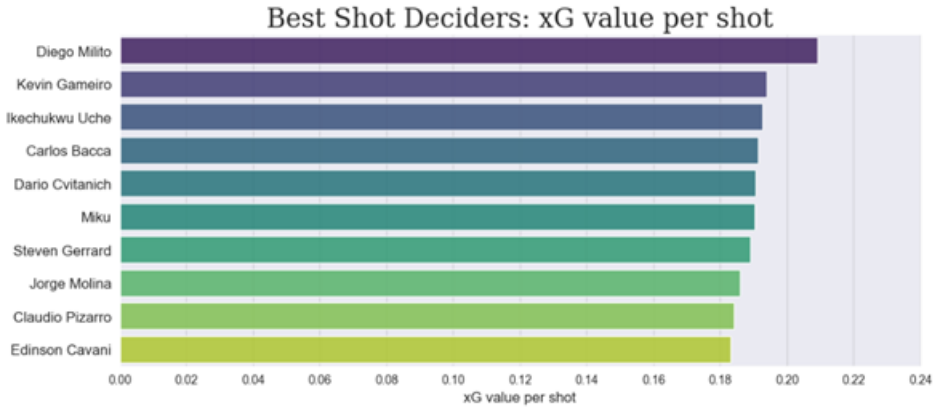


Figure 9: Best Shot Deciders

6.3.5 Best Decision-Makers in Shots

As the analysis progresses and delves deeply into the intricacies of the game, a distinct and unparalleled viewpoint surfaces, shedding light on the individuals who encounter challenges and difficulties in their decision-making processes when it comes to executing their shots on goal. Amongst the multitude of players, such as Diego Milito, Kevin Gameiro, and Ikechukwu Uche, there arises a notable distinction, a standout quality that sets them apart from their peers. These individuals exhibit a higher ratio of expected goals, a statistical measure that predicts the likelihood of a successful goal, thereby implying that they are more prone to experiencing misfortune and unfavorable outcomes in comparison to their counterparts who are more successful in scoring goals. The detailed list is shown in Figure 9.

6.3.6 Headers

In the domain of aerial confrontations taking place inside the penalty area, Cristiano Ronaldo, Mario Mandzukic, and Fernando Llorente emerge as formidable individuals who possess exceptional timing, effective positioning, and a remarkable ability to convert crosses into goals. Ronaldo, who is widely recognized for his commanding presence and extraordinary jumping ability, often overpowers defenders in order to score goals with his head. Having executed a total of 159 headers, he has managed to accumulate 36 legitimate goals, surpassing the projected goal count by an impressive 15.95. Mandzukic, a resolute force in the air, has registered 124 headers resulting in 28 legitimate goals, surpassing expectations by 13.31. Llorente, acknowledged for his proficiency in aerial battles, has tallied 141 headers, leading to 25 legitimate goals and surpassing expectations by 9.05. Collectively, these players stand out as masters of aerial play, each bringing a distinct combination of skills to enhance their respective teams' scoring opportunities.

6.3.7 Left Foot Specialists

In the domain of football precision, the utilization of the left foot assumes a prominent role, and Lionel Messi, Antoine Griezmann, and Arjen Robben emerge as exemplary practitioners of this particular skill. Messi, who is widely acclaimed for his mastery of using his left foot, has successfully executed 752 shots with his left foot, resulting in 167 goals that can be considered true, surpassing the anticipated goals by a margin of 45.43.

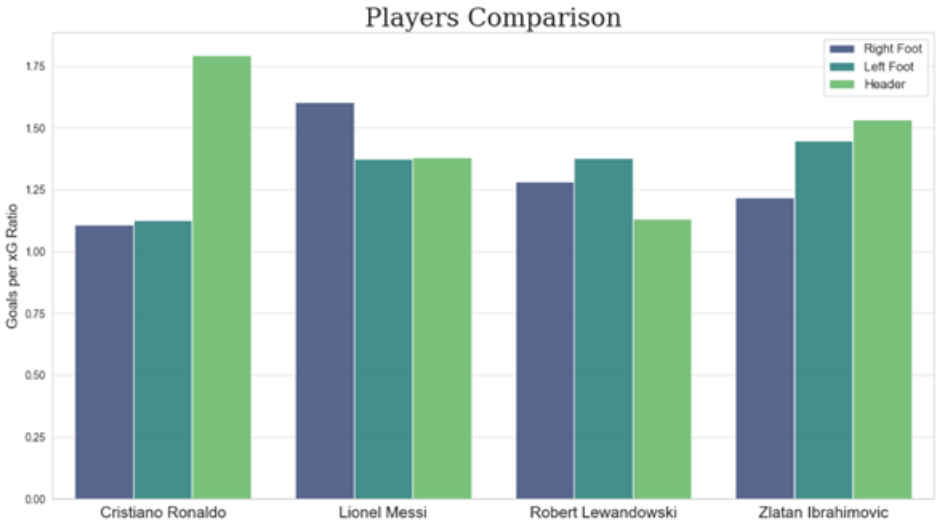


Figure 10: Player Comparison

Griezmann, a proficient striker who predominantly favors his left foot, boasts a total of 345 shots with his left foot, which have led to 58 goals that can be classified as true, exceeding the expectations by 16.57. Meanwhile, Arjen Robben, the renowned Dutch specialist in left-footed strikes, has skillfully executed 296 shots with his left foot, securing 42 goals that are considered true and surpassing the projected goals by 9.83. Collectively, these virtuosos of football not only exemplify the power but also the precision inherent in utilizing their left feet, consistently displaying remarkable accuracy in hitting the target.

6.3.8 Right Foot Specialists

Shifting the emphasis towards proficient practitioners of right-footed shots, Luis Suarez, Gonzalo Higuain, and Alexandre Lacazette emerge as notable individuals, exemplifying their adeptness in clinical finishing with utmost precision. Suarez, acknowledged as a right-footed sharpshooter, has successfully executed 289 shots with his right foot, resulting in 69 genuine goals, surpassing the anticipated goals by 25.81. Higuain, a goal-scoring specialist who favors his right foot, boasts an impressive tally of 362 right-footed shots, securing 86 genuine goals and exceeding expectations by 25.20. Lacazette, a highly skilled finisher with his right foot, has recorded a total of 270 right-footed shots, accumulating 70 genuine goals and surpassing expectations by 21.92. Collectively, these players exemplify the art of precise finishing, consistently locating the back of the net through skillful utilization of their right feet.

6.3.9 Outside-the-Box Specialists

In the domain of long-range shooting, Lionel Messi distinguishes himself as a specialist, showcasing exceptional expertise from beyond the penalty area. Messi, having attempted 304 shots from that distance, has successfully scored 16 legitimate goals, surpassing the anticipated goal count by an impressive 9.63. Joining him in this category, Paul Pogba displays a combination of force and accuracy with 226 shots resulting in 14 legitimate goals, exceeding expectations by 8.02. Zlatan Ibrahimovic contributes to the long-range threat with 261 shots, securing 14 legitimate goals and surpassing expectations by 7.21. Gonzalo Higuain also demonstrates prowess in this skill set, exemplifying his ability to

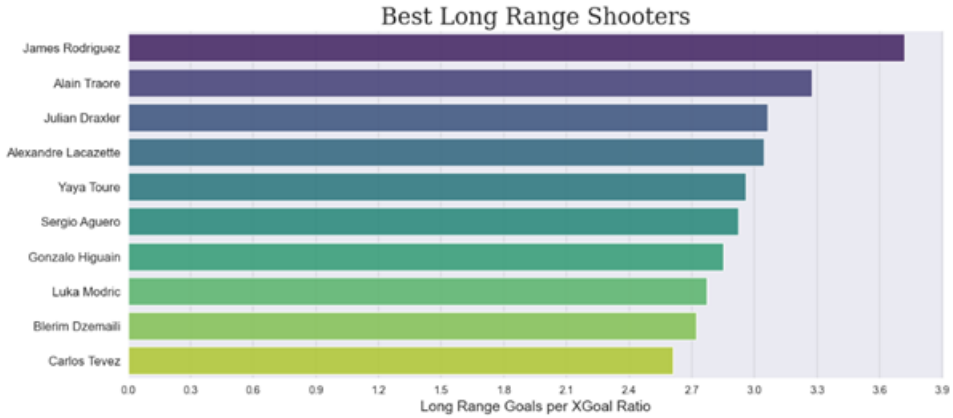


Figure 11: Best Long Range Shooters

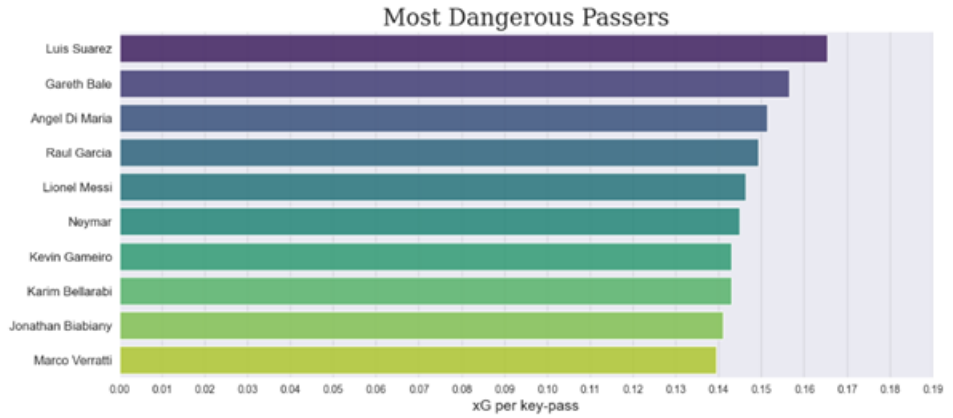


Figure 12: Most Dangerous Passers

consistently pose a threat to the opposing team's goal from a distance and contribute to scoring opportunities beyond the penalty area. For detailed and its visual representation can be seen in Figure 11.

6.3.10 Playmaking Expertise

Transitioning into the realm of playmaking, the analysis that is presented to us serves to shed light on those players who possess an exceptional aptitude for creating opportunities that could potentially lead to a successful goal. Particularly, the notable athletes Lionel Messi, Mesut Ozil, and Cesc Fabregas arise as the trinity that stands at the zenith of this talent, as they have established themselves to be invincible in terms of the sheer quantity of passes they have executed that have ultimately resulted in projected goals. On the other hand, when we consider the ratio of expected goals per pass, it is the trio of Luis Suarez, Gareth Bale, and Angel Di Maria who seize the spotlight, showcasing their unparalleled precision when it comes to playmaking. By delving into this aspect, we are granted a fleeting glimpse into the sheer brilliance and finesse exhibited by these extraordinary athletes. For detailed and its visual representation can be seen in Figure 12.

6.3.11 Unlucky Playmakers

In an alternative viewpoint, the analysis reveals individuals who possess the ability to influence the game through their passing but unfortunately do not receive the anticipated benefits. Joan Verdu, Xabi Prieto, and Luca Cigarini emerge as prominent examples of this phenomenon, illustrating a clear disparity between the expected and realized results of their playmaking efforts.

6.4 Discussion

The contrast among the Gradient Boosting Classifier and Logistic Regression models in the field of football analytics has great importance. Both models underwent an assessment based on their performance metrics and their relevance to the dataset. Table 1 shows a summary of the evaluation of the classification report and Fig. 13 provides the Confusion Matrix for Gradient Boosting Classifier and Logistic Regression.

The utilization of the Gradient Boosting Classifier and Logistic Regression was implemented to generate forecasts about goal outcomes. The Gradient Boosting Classifier, employing ensemble learning, builds a series of feeble learners to enhance the precision of its forecasts. In opposition, Logistic Regression, which is a linear model, accurately forecasts the likelihood of an instance belonging to a specific class.

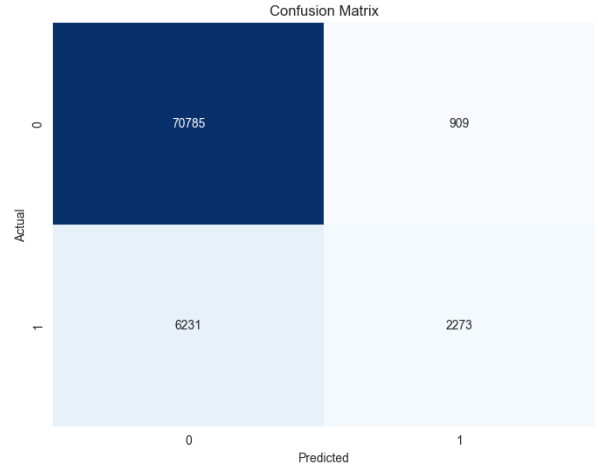
Performance metrics revealed that the Gradient Boosting Classifier outperformed Logistic Regression, achieving an accuracy of 0.745 and a ROC-AUC of 0.805 compared to Logistic Regression's accuracy of 0.708 and ROC-AUC of 0.776.

The Gradient Boosting Classifier displayed strengths in handling non-linear relationships and effectively utilizing ensemble learning techniques, which allowed it to capture intricate patterns with precision. However, Logistic Regression, in spite of its simplicity and interpretability, failed to provide a clear understanding of the consequences of traits.

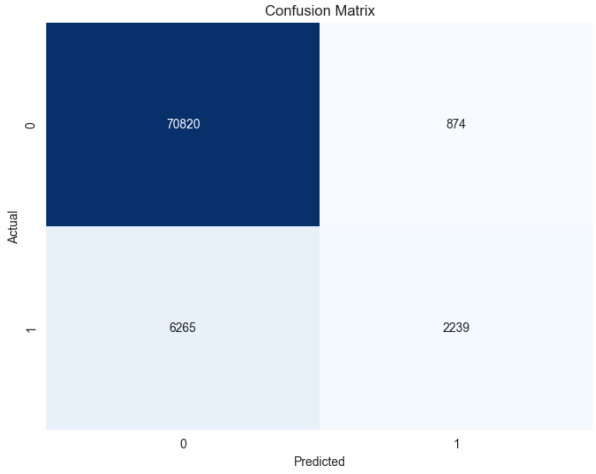
While evaluating the characteristics of the models, the Gradient Boosting Classifier, owing to its heightened complexity, is more suitable for datasets with intricate relationships and non-linear patterns. Its incorporation of ensemble learning makes it more hardy, particularly in the presence of outliers. In comparison, due to its lower complexity, Logistic Regression is improbable to be prone to overfitting and offers greater interpretability, making it suitable for datasets with simpler structures.

The performance of the models is immensely influenced by the settings of their parameters. The Gradient Boosting Classifier relies on elements such as 'n_estimators,' 'learning_rate,' 'max_depth,' 'min_samples_split,' and 'min_samples_leaf.' However, Logistic Regression, being a less complex model, utilizes elements like 'penalty,' 'C,' 'fit_intercept,' and 'solver.'

In conclusion, the selection between these models relies on the qualities of the data set and the analytical objectives at hand. If interpretability and simplicity are of utmost importance, Decision Tree may be the preferred choice. Nevertheless, for datasets that demand a nuanced understanding of intricate, non-linear connections and a high level of predictive accuracy, the Gradient Boosting Classifier emerges as a robust choice. Achieving optimal performance necessitates careful parameter tuning based on the specific characteristics of the dataset.



(a) Gradient Boosting Classifier



(b) Logistic Regression

Figure 13: Confusion Matrix

Metric/Model	Gradient Boosting	Logistic Regression
Accuracy	91%	91%
ROC-AUC	82%	82%
PR-AUC	47.34%	47.08%
Cohen Kappa	0.35	0.35
Precision (Class 1)	0.72	0.72
Recall (Class 1)	0.27	0.26
F1-Score (Class 1)	0.39	0.39
Precision (Class 0)	0.92	0.92
Recall (Class 0)	0.99	0.99
F1-Score (Class 0)	0.95	0.95

Table 1: Model Performance Metrics

7 Conclusion and Future Work

As we navigate the realm of xG modeling, the current state of our model presents promising indications, yet it also establishes the groundwork for future exploration and refinement. The evaluation, anchored by metrics such as Cohen's Kappa and PR-AUC PR, offers valuable insights; however, the absence of a standardized benchmark for comparison poses a challenge. The significant correlation of 0.97 between total expected goals and actual goals accomplished by the player alludes to competence, yet the pursuit of comprehensive performance metrics endures.

In the pursuit of enhancement, avenues for future endeavors beckon. One particularly promising direction entails delving into the defensive aspects of the game. Augmenting the model with data on the defending team, encompassing metrics such as the number of defenders, defensive pressure exerted on the shooter, and the temporal and spatial context, holds the potential to elevate predictive capabilities. Quantifying these defensive dimensions could unveil intricate patterns that exert a significant influence on the outcome of shots.

Spatial precision, a critical element in xG modeling, stands as another frontier for improvement. While the current model classifies shot locations into 17 distinct zones, the possibility of incorporating precise x and y coordinates offers a more detailed comprehension. This precision, particularly in capturing subtle spatial nuances, holds the promise of refining predictions and enhancing the accuracy of the model.

Looking further ahead, the domain of Deep Learning emerges as a captivating pathway. The intricate, nonlinear relationships inherent in soccer data might find a more nuanced representation through deep neural networks. Harnessing the capabilities of neural structures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), can potentially capture intricate spatial-temporal dependencies and unveil patterns that may elude traditional machine learning methods.

Essentially, while the current model demonstrates potential, the roadmap for future work extends beyond incremental enhancements. It envisions a holistic approach that integrates richer defensive insights, enhances spatial granularity, and embraces the transformative capabilities of Deep Learning. The dynamic nature of soccer analytics ensures that the pursuit of refining xG models remains an evolving journey, with each iteration bringing us closer to a more comprehensive understanding of the game.

Acknowledgment

I desire to take full advantage of this chance to express my genuine gratitude and sincere salutations to Professor Vladimir Milosavljevic. Throughout the duration of this task, Professor Milosavljevic has fulfilled the role of a mentor, offering irreplaceable advice. Their unwavering assistance and perceptive guidance have been instrumental in aiding me in overcoming the obstacles encountered during the progress of my research.

References

Baumann, A. (2022). *A multi-stage clustering algorithm to re-evaluate basketball positions and performance analysis*, Master's thesis, Dublin, National College of Ireland. Submitted.
URL: *https://norma.ncirl.ie/6082/*

Chaiwuttisak, P. (2018). Measuring efficiency of thailand's football premier leagues using data envelopment analysis, *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pp. 120–124.

Chavan, A. (2019). *Recruitment of suitable football player by using machine learning techniques*, Master's thesis, Dublin, National College of Ireland. Submitted.
URL: *https://norma.ncirl.ie/4307/*

Coito, N., Davids, K., Folgado, H., Bento, T. and Travassos, B. F. R. (2020). Capturing and quantifying tactical behaviors in small-sided and conditioned games in soccer: A systematic review, *Research Quarterly for Exercise and Sport* **93**: 189 – 203.
URL: *https://api.semanticscholar.org/CorpusID:222237270*

Gibney, R. (2022). *Using supervised learning techniques to predict kicking outcomes in the nfl*, Master's thesis, Dublin, National College of Ireland. Submitted.
URL: *https://norma.ncirl.ie/6589/*

Goes, F. R., Meerhoff, L. A., de Oliveira Bueno, M. J., Rodrigues, D. C. U. M., Moura, F. A., Brink, M. S., Elferink-Gemser, M. T., Knobbe, A. J., Cunha, S. A., da Silva Torres, R. and Lemmink, K. A. P. M. (2020). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review, *European Journal of Sport Science* **21**: 481 – 496.
URL: *https://api.semanticscholar.org/CorpusID:215794449*

Gorman, D. (2017). *Sports Analytics: Analysis of the National Football League*, PhD thesis, Dublin, National College of Ireland. Submitted.
URL: *https://norma.ncirl.ie/2661/*

Harsha Vardhan Goud, P. S., Mohana Roopa, Y. and Padmaja, B. (2019). Player performance analysis in sports: with fusion of machine learning and wearable technology, *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 600–603.

Kryger, K. O., Wang, A., Mehta, R., Impellizzeri, F. M., Massey, A. and McCall, A. (2021). Research on women's football: a scoping review, *Science and Medicine in Football* **6**: 549 – 558.
URL: *https://api.semanticscholar.org/CorpusID:234126324*

Kumar, S. (2020). *Artificial neural network for betting rate in football*, Master's thesis, Dublin, National College of Ireland. Submitted.
URL: *https://norma.ncirl.ie/4404/*

Lacković, K., Horvat, T. and Havas, L. (2018). Information system for support in basketball game and player analysis, *Informatologia* .
URL: *https://api.semanticscholar.org/CorpusID:70193055*

Low, B., Coutinho, D., Gonçalves, B., Rein, R., Memmert, D. and Sampaio, J. (2020). A systematic review of collective tactical behaviours in football using positional data, *Sports Medicine* **50**: 343–385.
URL: *https://api.semanticscholar.org/CorpusID:203592790*

Perin, C., Vuillemot, R., Stolper, C. D., Stasko, J. T., Wood, J. and Carpendale, M. S. T. (2018). State of the art of sports data visualization, *Computer Graphics Forum* **37**.
URL: *https://api.semanticscholar.org/CorpusID:51777904*

Rico-González, M., Arcos, A. L., Nakamura, F. Y., Moura, F. A. and Pino-Ortega, J. (2020). The use of technology and sampling frequency to measure variables of tactical positioning in team sports: a systematic review, *Research in Sports Medicine* **28**: 279 – 292.
URL: *https://api.semanticscholar.org/CorpusID:202569692*

Rico-González, M., Pino-Ortega, J., Nakamura, F. Y., Moura, F. A., Rojas-Valverde, D. and Arcos, A. L. (2020). Past, present, and future of the technological tracking methods to assess tactical variables in team sports: A systematic review, *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology* **234**: 281 – 290.
URL: *https://api.semanticscholar.org/CorpusID:222111229*

Selvaraj, S. (2016). *Analysis of player ratings based on intrinsic factors to support team selection*, Master's thesis, Dublin, National College of Ireland. Submitted.
URL: *https://norma.ncirl.ie/2498/*

Stein, M., Breitkreutz, T., Häussler, J., Seebacher, D., Niederberger, C., Schreck, T., Grossniklaus, M., Keim, D. A. and Janetzko, H. (2018). Revealing the invisible: Visual analytics and explanatory storytelling for advanced team sport analysis, *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)* pp. 1–9.
URL: *https://api.semanticscholar.org/CorpusID:53370525*

Stein, M., Janetzko, H., Breitkreutz, T., Seebacher, D., Schreck, T., Grossniklaus, M., Couzin, I. D. and Keim, D. A. (2016). Director's cut: Analysis and annotation of soccer matches, *IEEE Computer Graphics and Applications* **36**: 50–60.
URL: *https://api.semanticscholar.org/CorpusID:14200930*

Stein, M., Janetzko, H., Lamprecht, A., Breitkreutz, T., Zimmermann, P., Goldlücke, B., Schreck, T., Andrienko, G. L., Grossniklaus, M. and Keim, D. A. (2018). Bring it to the pitch: Combining video and movement data to enhance team sport analysis, *IEEE Transactions on Visualization and Computer Graphics* **24**: 13–22.
URL: *https://api.semanticscholar.org/CorpusID:206806406*