UNIVERSITY OF SUSSEX

SCHOOL OF MATHEMATICAL AND PHYSICAL SCIENCES

DEPARTMENT OF MATHEMATICS

Expected Goals (xG) Prediction in Football: Examining the Effect of Features on the Performance of the xG

Student:

Kamaldeen Ekundayo Aiyeleso

260792

Supervisor:

Dr. Kate Shaw

Submitted in partial fulfilment of the requirements for the MSc degree in Data Science at the University of Sussex

Abstract

Expected goal, shortened as xG in sports football, is a concept that has gained continuous attention from sports organisations and is used to estimate the goal-scoring chances that a team has had, different from the total goals they scored. To estimate the xG value of any given shot, several pieces of information known as shot features are modelled using statistical and machine learning approaches. In this project, machine learning is introduced to build a statistical model to estimate the xG value of a given shot. The shot features are also explored to understand the effect of the features on the xG estimation.

The project employed basic features commonly used to describe a shot, such as distance, angle, body part, and technique, to develop a basic xG model. The basic xG is compared with an advanced xG model built using additional features. During the training and evaluation, models such as logistic regression, random forest, decision tree, and XGBoost are introduced to train the xG model and evaluated using the log loss metric, which estimates the efficiency and reliability of the model in calculating the xG value of a given shot.

To have an empirical understanding of the effect of each feature on the xG estimation using the basic and advanced features, the project introduces a concept known as the Shapley additive value to calculate the contribution of each feature. This gives a better understanding of the most important and least important features; however, all features must be considered as they give the model enough context to make the xG estimation reliable and efficient.

Hence, the project reviews the performance differences between the basic and advanced xG and evaluates the contribution of each feature. This is done using a quantitative and comparative analysis of the xG models' accuracies using the log loss values from both basic and advanced xG models, demonstrating the impact of the input parameters on the reliability of the xG model.

Acknowledgement

I would like to express my special gratitude to my supervisor, Dr. Kate Shaw, for her immense support throughout the project timeline, contributing to the successful completion of this project.

I would also like to thank my parents, Mr. Abubakar and Mrs. Haolat Aiyeleso, for their endless support and encouragement, ensuring I had a focused mind.

My final appreciation goes to my fiance, Raheemah Azeez, who has been a huge part of my journey towards completing this project.

Table of Contents

List of Figures

List of Tables

1.0 Introduction

1.1 Background Study

Football has been around for many years and is regarded as one of the world's most famous and widely practised sports. Football is the most popular sport globally. As acknowledged by FIFA, a widely recognised sport's governing body, the viewership of the FIFA 2018 and 2022 World Cup was estimated to reach approximately half of the world's total population. Football in the context of European sport can be dated back to the mid-nineteenth century (Wood, 2017) and has enjoyed a spread out across all world regions. Football organisations have made substantial investments over the years, establishing clubs that garner extensive support from passionate followers known as fans. These dedicated fans exhibit unwavering enthusiasm and eagerly anticipate outstanding performances from their beloved clubs in any competition. Their team's goal is to secure victories and claim trophies, allowing fans to assert their superiority over rival clubs proudly. Winning matches and securing victories in football brings increased profitability through revenue sharing (Özaydin and Donduran, 2019), attracts more fans, and lures talented players to contribute to the club's success.

Beyond the scope of business investment in football, individual participation has been attributed to other benefits, including promoting physical fitness and general well-being. It involves aerobic and anaerobic activities, improving cardiovascular health, strength, endurance and coordination (Bangsbo et al., 2006). There are also opportunities for personal development and teaching players essential life skills such as teamwork, communication, discipline and leadership. It can also be attributed to giving a platform for social interaction and inclusion, providing equal opportunities for people of different ages, races, backgrounds and genders to participate and excel (Tacon, 2007).

While we have understood the great benefits of football across several instances, it is important to know how the success of football is being measured. Winning is an important part of football, usually determined by scoring more goals than the opposing team. In football, goals are determined when a football has been played into the goalpost of a second team and are usually awarded to the other team. Sometimes, no goals are scored during a football game which brings to question whether there could have been the possibility of goals being scored. In football, 10% of total shots played are usually converted to goals (Pollard and Reep, 1997; Tenga et al., 2010; Lucey et al., 2015, cited in Anzer and Bauer, 2021), making football a low-scoring sport. Also, a goal is counted as one, and to win a match, such a team will have to outscore their opponent.

Due to the low-scoring nature of football sports, teams try to be as effective as possible with their football players when taking crucial shots to score goals. Most of the shot attempts are either off-target or saved by the goalkeeper. To understand how effective the shots are, scoring chances measurement and quantification have been introduced and have led to a significant development in the analysis of football and summarised as expected goals (xG). The xG is the measured probability that a shot played towards an opponent's goalpost will result in a goal (Hewitt and Karakuş, 2023). Simply put, what is the percentage success of such a shot resulting in a goal? The xG is usually based on several factors, such as the distance to the goal, angle (Lucey et al., 2015), body parts used and technique. These factors are described as the shot features. The shot features are measurable attributes or contextual information about the shot.

The xG concept created a better understanding that not all shots have the same goal scoring probability. It may then be assigned a probability value between 0 and 1 for each shot, known as the xG value and this can be measured using a combination of the contextual or attributed features of the shot. This xG value helps football analysts better understand how efficient a football team is in their shot outputs and making recommendations. The idea is to come up with a model that can be used to measure goal-scoring opportunities across teams, players and within football match events.

In determining the xG, the machine learning technique has been introduced. Machine learning is a robust field applied to several domains, including sports analytics (Chmait and Westerbeek, 2021). According to IBM, Machine learning is a branch of computer science that focuses on using historical data and algorithms, a set of statistical rules to understand how humans and systems learn. Machine learning extracts knowledge from data and processes it for predictive, prescriptive or descriptive purposes. It is an intersection of statistics, computer science and artificial intelligence. In this research project, the historical data used is football data (Muller & Guildo, 2017).

Sports analytics is a fast-growing market area with an estimated global amount of 2.98bn USD in 2022, according to a report published by Fortune Business Insights (2023). This is projected to grow from 3.78bn USD in 2023 to 22.13bn USD by 2030. This suggests that it is a thriving market, and sports organisations continue to explore its benefits, especially in helping their coaches and management officials enhance their performance. Sports, in general can benefit mainly from the introduction of machine learning concepts to address and solve their business problems. Football organisations may also be able to analyse players' effectiveness beyond the general team performance.

1.2 Problem Statement

Developing an accurate and reliable expected goal xG) model is of great importance to sports organisations. As discussed earlier, football clubs use the xG model to scout undervalued players, demonstrating the importance of having access to a reliable xG model. Also, football coaches can use xG output of their team to assess performance and provide coaching advice for areas of improvement. Despite the wide use and application of the xG model in football, significant challenges are associated with developing a highly accurate and reliable xG model.

One major challenge is the lack of freely accessible data (Hewitt & Karakus, 2023). Although organisations like StatsBomb have gathered a large amount of data which covers over 12 years of football match events. This data can serve as a reliable source and will be the sole data source for this project thesis. While it is recognised that this data source may not contain up-to-date event information of recent football matches, it is still reliable and provides enough foundation for developing the model. However, individual football clubs keep many advanced data and techniques private. This ensures that rival clubs do not gain a competitive advantage by having access to their xG models (Hewitt & Karakus, 2023).

Developing an accurate and reliable xG model is very crucial to sports organisations. Having discussed how football clubs use the concept of xG to scout undervalued players, these clubs will ultimately benefit from using a near accurate xG model. Also, considering that the xG model can be used to make coaching advice and improvement, coaches will be looking to review what worked well during and after a football match, having examined their xG values. While we have understood the relevant and wide usage and application of the xG model, we should realise the problems and difficulties of developing a very accurate and reliable xG model. The xG model can be very challenging due to the current lack of freely accessible data (Hewitt & Karakus, 2023), and this has motivated the need to examine the effect of the features used in developing a reliable xG model. This will aim towards providing recommendations on xG model factors that are deemed highly relevant and can help football organisations focus on improving the quality of data to be gathered, taking into consideration the factors or features that will be revealed in the investigation.

Sports body like StatsBomb gathers thousands of events data and can be used in developing the xG model.

1.3 Goals and Objectives

The goals of this project have been divided into several stages and are listed as the following:

- Develop the basic xG model using basic features. The basic features will include features that are not considered advanced and are widely used across many literature reviews. The basic xG model will also serve as the baseline for performance evaluation and validation.

- Identify the advanced features, combined with the basic features to develop a second xG model known as advanced xG. During this stage, the model will benefit from both the basic features and the proposed advanced features.
- Examine the individual importance of the features from the most reliable model.

1.4 Scope and Limitations

1.4.1 Scope

The xG model proposed in this project will use a publicly sourced dataset from the StatsBomb official data repository downloaded from GitHub online repository. While the data source provider is reliable, it should be recognised that limitations are within the different leagues covered by the dataset. The StatsBomb data source provides detailed event data, which covers the comprehensive event-level information allowing analysis of each shot event, which will be used to develop the xG model. It also covers player-specific information, which gives the opportunity to put into context information such as the playing position and individual location of the player allowing examining the impact of players on the outcome of shots.

1.4.2 Limitations

Considering the number of leagues and seasons covered by the StatsBomb dataset, it may have an inherent bias which will only consider few types of leagues, reducing its ability to generalise across uncovered leagues in the world. It should also be noted that there is evidence of incomplete data in some new techniques that were not employed more than a decade ago. This will instead lead to a reduction in the available data to train the model as uniformity must be ensured in the instances and null values removed.

Football is a constantly evolving sport with changes in the style of play, tactics and rules adaptation making it challenging to develop an xG model that contextualises the most recent rules and techniques. The use of the StatsBomb dataset is characterised by having a specific time period, although the available dataset provides context up to 2021, which is reliable.

Some external factors which could also be added to part of the features used in building the xG model are not present in the data. These factors include player fatigue level, pitch and weather conditions. It will also be limited to using the StatsBomb dataset as efforts towards combining data from other sources will lead to different data structures due to the apparent difference in how football companies gather their football event dataset.

2.0 Literature Review

2.1 Expected Goal (xG) Model

The xG is applied by many footballing organisations. It is the probability measure of any shot taken by a football player resulting in a goal. This can be measured basically by using information such as the distance and angle of the shot from the goal post, giving a perspective of difficulty in shots using factors such as the distance or angle. This concept of xG is demonstrated to quantify the effectiveness of the shot aiming to be a goal and helps football teams, and coaches understand differences in chances of shots resulting in a goal.

Mead O'Hare (2023) defined xG models as statistical models that estimate the probability of a shot resulting in a goal based on several factors or features such as the shot distance, shot angle, shot technique, defender proximity and several other features. Since the xG uses the concept of probability, we would expect the outcome of the xG to always be between 0 and 1. 0 indicates no possibility of a goal being scored from such a shot and 1 indicates the likelihood of a goal. Mathematically, the xG probability involves using statistical or machine learning modelling techniques to use the features described earlier to compute the probability of the shot resulting in a goal. Similarly, these models represent valuable tools used to evaluate the efficiency and effectiveness of football teams, players and playing strategies in football matches.

The xG models, which have now evolved from using traditional features to evaluate them, as Lucey et al. (2015) described, now use much more sophisticated features and factors to determine the xG value of shots. This has gone beyond the use of shot location, which closely demonstrates the traditional technique applied to the development of xG. Since it is now understood that the xG model uses more advanced techniques, such as machine learning modelling, more features can be used to examine the development of xG models beyond the basic features and develop algorithms for deriving the goal probability.

The shot location was one of the critical parameters examined in creating xG, although initially represented by zones on the pitch. The recent shot location being collected can give information on the exact shot distance measured in yards and angle measured in degrees from the goal post. Rathke (2017) discussed the significance of shot location in xG models and how it affects scoring probability. While these are considered basic features, advanced features such as defence density (the number of defenders directly impacting the path of the shot), shot technique, shot type (i.e., header, penalty, freekick), and player under pressure have recently been used to improve the xG model's performance. These features provide a more comprehensive understanding of the situation of the shot and its probability towards resulting in a goal.

To develop a reliable xG model, adequate and rich data is highly required. In the past, successful xG models have been facilitated by the presence of such data. StatsBomb, a popular football data provider, has been widely used and will be adopted in this project as the only data source to train and evaluate the xG model. Some research articles, such as Hewitt & Karakus (2023), used the StatsBomb data to build their xG model, demonstrating the reliability of the StatsBomb data provider, which is open source.

Finally, it can be understood that Expected Goals (xG) models provide a better understanding of their use in sports analytics, especially football. This can be used to measure the quality of a shot and the scoring chances or probability, hence, providing comprehensive performance analysis, evaluating a player’s ability and making strategic decisions. The xG models offer rich and useful insights into understanding the difficult goal-scoring situation in football.

2.2 Shot Event Features

In this project, the shot event features represent the factors or properties of the shot, which are used in developing the xG model. These features have been gathered using the StatsBomb data representing event and match data. They define the conditions and properties of each shot used to develop the xG models.

The shot distance is a common feature used in the development of xG model. This has constantly been used in previous xG models, as seen in the section 2.3. It is a basic feature used in determining xG of a goal. Similarly, the angle of the shot relative to the goal posts is used to express chances of goal-scoring opportunity. This will be considered as another basic feature that will be used to develop the basic xG model.

Other features that will be considered in this project and discussed include the shot technique, shot type, defence density, under pressure and body part. Combining all the above features can be used to describe a shot and measure the scoring probability.

The features will be categorised into two groups later in this work. The first group will represent the basic features commonly used to develop xG models. At the same time, the second group will consider all features from the first group and new advanced features. By defining two groups for the shot event features, the project seeks to evaluate the importance of the features on the performance of the xG models across the basic and advanced features; it will seek to answer this question by building a basic and advanced xG model which relies on the use of the basic features and addition of advanced features.

2.3 Related Works

In 2013, Michael Caley presented a blog article on using shot location to determine the odds of scoring a goal. This article explained that shot locations can be broadly divided into eight zones, with 13% of shots from outside the box usually being converted to the goal against 35% of shots from within the box. Using the 2009/2010 English Premier League football statistics till 2013, the article analysed the shot data and quantified the effect of shot location on goal outcome. Zone 1, the closest to the goalpost, recorded 43% goal of shots played in that position. Compared with zones further from the goalpost, a decline in the percentage of goals was recorded. This article used fewer features and only determined the probability of the shots resulting in a goal based on the recorded goals scored from each zone calculated from the total shots played from the zones.

Lucey et al. (2014) presented an academic article on expected goals where the value of a shot is quantified using player tracking data. The paper examines strategic features such as proximity of the defender, speed of play, space and spatial location of the shot. In the article, defender proximity is defined as the number of defenders between the shot and the goal, while space describes the distance between the shot and the next defender. With the traditional method of estimating shots resulting in a goal, the prediction error which calculates the percentage of wrongly classified shot outcomes would be a value of 17.45%. To significantly reduce the prediction error, the spatial location of the shots was suggested and saw the prediction error reduced to 16.62%. This spatial location gave a coordinate location of the player and helped generate more accurate predictions. The paper examines the prediction error reduction by comparing the xG model using different features such as only shot context (spatial location and angle), shot context combined with defending and shot context with defending and attacking. The analysis proved that prediction error could be significantly reduced using a combination of all strategic features proposed, as they play an essential role in accurately estimating the likelihood of a shot resulting in a goal.

Eggels et al., (2016) presented a paper that explains the goal-scoring opportunities using a predictive analytics model developed using historical football event data. The model aimed at defining the probabilistic estimates of scoring chances and quantifying the value of a shot and goal-scoring opportunity. In addition to using features previously discussed, player and goalkeeper quality was considered in the xG model built in this paper. It aimed to have low variance, error and high generalisation ability and improved business value. This paper tested four different models for performance comparison and evaluation and suggested a calibration technique to evaluate the Brier score, explaining the alignment between the predicted probabilities and goal rates.

Rathke, (2017) examined goal scoring in European football leagues and built an xG model focusing on forward football players. The approach improved the work carried out by Carley, 2014. Zones were used to define Shot distances and angles determined from the goalposts in each zone. The paper evaluated the use of Shot distance alone to measure xG compared with combining shot angle features. In conclusion, the shot distance was an important feature in developing xG. However, it should be combined with the shot angle to create a more effective and reliable xG model. It should be noted that the paper also recognised that xG value can be impacted by the team's quality, as xG does not measure well against the top league teams. When this paper was released, xG is still in its early stage in football and hadn't benefitted greatly from detailed research.

Kharrat et al., (2017) introduced goalkeeping skills and big chances created in the xG model proposed in their paper. Four specialist models were developed using data from different shot types based on freekick, penalty, open play and header. It was recognised that the situations will be different and appropriate to treat separately. It also fits their main objective of developing the xG model for plus-minus player rating against the xG outcome of shots. This means xG is studied by players' actions which affect the net expected goals of the team. It is important to note that the features used in this model are basic, as spatiotemporal data of player position are not introduced to determine the shot distance or angle. Although, the xG models benefitted from features like goal difference and time of the match.

A study conducted by Spearman (2018) attempted to quantify shots as a result of an off-ball scoring opportunity while making some references to expected goals. Although this was not primarily developed to build an xG model, it still featured the use of strategic features such as shot location to find the probability of a shot resulting in a goal. This aspect of the paper is of interest to the current project work. It gives an academic point of view of understanding reported basic and advanced features that have been considered in the development of expected goals models in recent times.

In another paper by Anzer & Bauer, (2021), they introduced the use of positional and event data trained using supervised machine learning models with all the features discussed so far, including some new features. This work is very similar to the features proposed in the current project. Although, this considered the use of specifying freekick and after freekick events. Some additional features include the speed of the player taking the shot, the goalkeeper's position and pressure on the player. Five machine learning models were used to train the features and calibrate the probabilities, in which the final evaluation demonstrated the gradient boosting model having the highest AUC score. The xG model in this paper outperformed models developed in previous approaches like Rathke (2017) and Lucey et al. (2014) with an improved average prediction error of 9.28% as it considered newer features. Although, it is important to state that the model also benefitted from a larger amount of training shot data of more than 100,000 instances.

Two more academic papers like this project will be discussed. The work carried out by Mead et al., (2023) can be closely linked to the proposed idea in this project. Beyond incorporating more advanced features, Mead et al. evaluated the importance of the features used in building the xG model. Some new features formed by modelling features from existing event data include the Elo ratings, team form and match importance. In estimating the performance of the classification models adopted for building the xG model, the log loss value of the classification model was first tested. The lower the log loss, the more reliable the classification algorithm is in estimating the probability value of the goals from shots. The paper demonstrated the importance of the features using the model feature importance library, which ranks predictors according to how much information they contributed to determining the prediction outcome.

Finally, the work carried out by Hewitt & Karakus (2023) did great justice to extending the use of player position to adjust the developed xG model. Ideally, this paper recognised that football players who are not naturally forwarders might skew the prediction using models built with natural forward football event shot data. Also, among some natural forwards, highly gifted and technical players may skew the goal-scoring ability creating a bias when such a model is used to test football players who are less superior in their footballing abilities. Hence, the author proposed a player and position-adjusted xG model.

3.0 Methodology

This chapter will introduce the football event data sourced to develop the proposed expected goal model and explain the machine learning techniques used in generating the model. The section will discuss the data source, give a detailed description of the dataset being used, provide a breakdown of the shot features and further explain the modelling of derived features using existing features. Overall, the final aim of the project is to develop two types of xG models, which are basic and advanced xG models. The following describes the outline of the methodology section.

- Section 3.1 Project Design provides a general structure of how the xG model will be developed and describes the overall analysis that will be carried out.
- Section 3.2 Data Collection explains the data provider and gives a general dataset description.
- Section 3.3 Data Preparation explains the cleaning techniques, exploration and development of new columns from the existing columns in the dataset.
- Section 3.4 Shot Features cover the description of all event features that will be used to train both basic and advanced expected goals models. Rich diagrams and mathematical explanations will be included where necessary.
- Section 3.5 Grouping describes features that will be used for the basic and advanced models separately.
- Section 3.6 Model framework is a general description of how the expected goal model will be approached and the problem statement this project is trying to address.
- Section 3.7 Model training introduces the machine-learning techniques that have been adopted for this project.
- Section 3.8 Performance evaluation seeks to clarify the results obtained from the expected goals model developed and validation of the model to meet the clear goals.

3.1 Project Design

A supervised machine learning approach has been suggested for the development of the xG models and can be termed predictive analytics. This is in combination with several exploratory analysis that will be carried out to describe the dataset, parameters that are present and the features used in building the model.

Machine learning approach is more suitable as it allows the modelling of dataset which are much larger and handling of any form of complexity. The patterns are recognisable easily and a suitable model is developed. Subsequently, the classification machine learning model will provide the estimated probability values of the input features. These values are interpreted as the xG value. The xG value measures the probability assigned to the input feature.

3.1 Data Collection

In developing the expected goals model, several data sources have been identified with a preference for relatively available data while also considering rich information from such data. Football event data, which contains spatiotemporal information, can be a bit tedious to collect as advanced machines and accurate human evaluation and validation of such data is required. Considering that the expected goal (xG) model seeks to include advanced features, having access to updated data is deemed important.

3.1.1 StatsBomb Data

One data source will be implemented and used throughout this project: the StatsBomb open data downloaded from the GitHub repository. StatsBomb is an organisation which helps collect rich football data across La Liga, Fifa and many other competitive leagues. For this project, the data has been made available from GitHub and stored across many directories. StatsBomb open data is considered reliable as it contains granular event information required to develop the xG model. StatsBomb collects event information which highlights different actions that have been carried out during a match. It averages more than 2000 events per match consisting of passes made, shots played, throwing, fouls, end of a half time and many more events actions that have occurred during the football match. The xG model will largely benefit from the events data related to shots. Shots in this project refer to actions taken when kicking a football aimed at a goalpost. In the football data used, about 6 professional leagues have been covered. They include La Liga (2010 – 2021) which consists of matches played by FC Barcelona against other teams in the league, FIFA World Cup (2018, 2022), Indian Super League (2021/2022), FA Women's Super League (2018 – 2021), Women's World Cup (2019) and Champions League (2017). The project also recognises the varying difference in the technicality of different leagues especially men's versus women's league. The xG model will benefit from learning both competition type and generalise appropriately.

Although these events may not be current with the most recent football matches but can be sufficient to cover over 12 years of football events. While this data can be reliable for building an xG model, much of the techniques and richer data used are privatised by most football clubs. The knowledge and technique are limited to the owning companies to ensure rival clubs do not gain any competitive advantage against them (Hewitt & Karakus, 2023).

A total of 963 matches were obtained from the selected filters of data described above relating to the type of football matches. This contained a total of 24,495 events which are directly related to a shot.

3.2 Preparation of Data

The event data comes as JSON files and contains all features used to measure an action. For the purpose of this project, shot actions have been prioritised and relevant features from the events are extracted. Each event is a separate JSON file, which is saved with a unique ID and later interpreted as the match ID. It is important to understand that the events do not contain information about the name of the match, competition and year of the match. To access the match information which will be used to identify and carry out statistical analysis on specific match and competition, separate competition and match information JSON files have been provided by StatsBomb. Recall that each event is a JSON file and relates to a single match. Within the matches directory, there are information about the event ID containing match ID, date of the match, competition information and season information. Once the table had been created from the matches JSON file, it was joined with the event data using the match id information as the unique key. Finally, this gave a table containing all features and match information. The final output of the table saved as a CSV contains important columns such as season name which identifies the year of the match, competition name, event type relating to the shot, measured xG value by StatsBomb known as Stats xG and the outcome of the shot event.

In the next section, the creation of new features from existing features will be explained in detail. Some of the approaches taken include applying mathematical concepts to obtain new features such as the Shot distance and Shot angle. This is because the shot location feature has only been provided in the form of a x and y coordinates. To also obtain the defence density, an advanced technique known as point-in-triangle test had been introduced.

3.3 Shot Features

This section introduces the detailed discussion on the features used to describe the shot. It will cover the information on how the features are measured and as well as mathematical formulars used to express the feature. Both basic and advanced features are covered in this section.

Shot Distance: The shot distance d describes the distance between the shot taker and the centre of the goalposts. It is measured in yards and stored as a float data type. The shot distance is obtained from the shot location column in the StatsBomb dataset. The shot location stores data in list format containing information using coordinate values of x and y. The coordinates x and y are represented on the pitch having a size of 120 by 80 yards on the x and y axes respectively. The x measures between 0 and 120 yards, while the y measures between 0 and 80 yards. Using the Euclidian formula, the given coordinate of the shot taker can be recorded as point p, and the coordinate of the centre of the goal post, which will have a fixed value of 120 on the x-axes and 40 on the y-axes represented as point q. Distance from point p and point q is obtained using the Euclidian distance shown in I below.

$$d_{(p,q)} = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$

Equation I: The expression for calculating Euclidian distance between two points.

The above equation represents the Euclidian distance of a two-dimensional plane calculated using Pythagoras' theorem and measured in yards. Point p and q coordinates define the shot location and centre of the goalpost on the pitch.

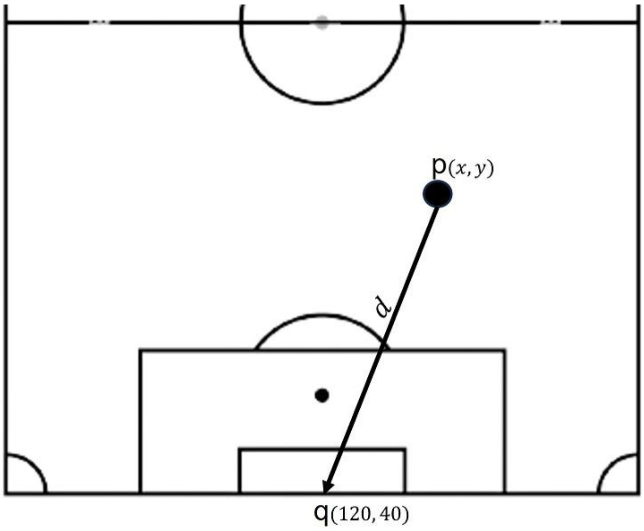

Figure 1 Image showing the feature 'Shot distance' measured from the shot taker p to the centre of the goalposts q.

Shot Angle: Shot angle is described as the angle between the two lines connecting the goal post ends and the location of where the shot is taken from. The angle of the shot is another feature used to express the difficulty of a taken shot (Galbraith & Lockwood, 2010), described as θ and measured in degrees. Using the shot location coordinates, the shot angle can be calculated using the cosine angle formula. This is because the triangle view that's formed between the shot taker's location and the goalpost is easily obtainable. The goalpost has a fixed coordinate of 120 by 36 and 120 by 44 on both ends of B and C shown in Figure 2 below. These coordinates have been validated from the documentation provided by StatsBomb labelling the coordinates of important locations on the football pitch. In Figure 2 below, consider point A as the shot taker coordinates, which form an angle ABC with the goalpost.

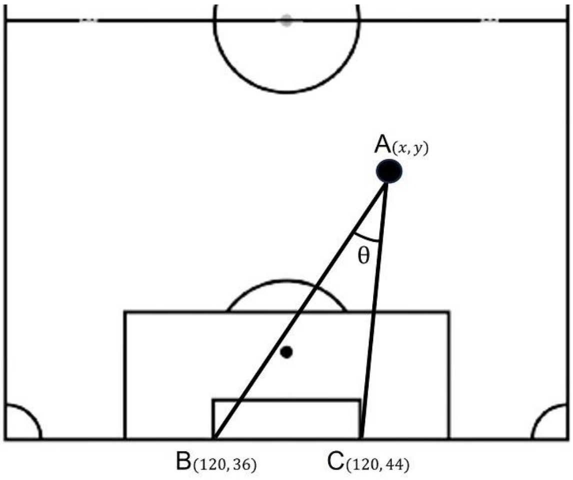

Figure 2 Description of the shot angle formed with the goal posts at BC from location A of the player taking the shot.

The angle BAC, otherwise represented as θ in Figure 2 above is calculated using equation II below. It is measured in degrees and can have a value between 0 and 180. Shots that are taken from a corner kick and directed towards the goalpost will form a zero angle and shot taken directly on the goal line will form a straight angle of 180 degrees.

$$cos^{-1} = \frac{AB^2 + AC^2 - BC^2}{2 * AB * AC}$$

Equation II Expression for calculating cosine angle of a triangle.

Defence Density: Defence density information has been introduced into the xG model and the concept seeks to examine the impact of players who are present within the shot angle. This has been limited to the opposing players and any teammate that is within the shot angle is excluded. The number of opponent players can be between 0 and 11 and represents the defence density. To determine the defence density, the point-in-triangle test has been introduced. From Figure 3 below, we can see the triangle formed by a shot taker A with respect to the goalposts B and C and the opponent in the angle of view represented by P. Using the Equation III below, the total area of the triangle ABC is obtained. The equation is extended for use to calculate the areas of triangles: PAB, PBC and PCA. These triangles are formed by connecting point P with each of the vertices of triangle ABC. The sum of the areas of the triangles PAB, PBC and PCA must be equal to the area of the triangle ABC obtained previously. This confirms the point P is within the triangle ABC. The technique is applied to all opposing players coordinate position when a shot is taken and total opponent within the shot angle can be counted and represented as defence density. It is unitless as it only represents total defenders within the shot angle area.

$$area = \frac{1}{2} \times \left| A_x \times \left(B_y - C_y\right) + B_x \times \left(C_y - A_y\right) + C_x \times \left(A_y - B_y\right) \right|$$

Equation III Expression for calculating area of a triangle using coordinates.

From the dataset, a freeze frame data information has been provided which contains player location coordinates at the time of taking the shot. In Figure 3 below, we can see the location of the player present within the shot angle. The player taking the shot forms an angle of view with the goalposts which has 1 opposing player present within the angle.

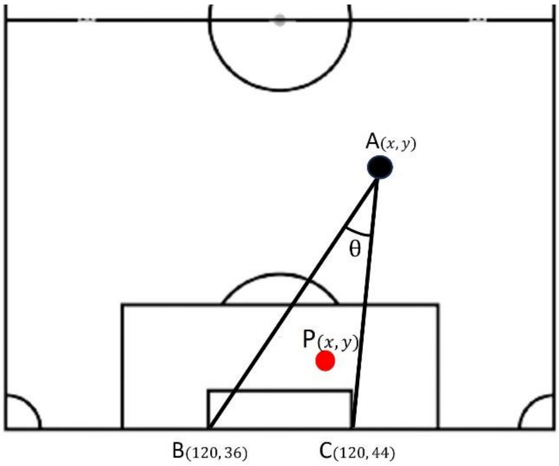

Figure 3 Image description of defence density showing opponent player.

Shot Type: The Shot Type is described with three broad situations on the pitch. These situations include free-kick, open play and corner kicks. The free-kick Shot type is a shot taken outside the 18-yard box when an opponent player fouls an attacking player. The free-kick Shot type usually involves defenders positioning themselves at strategic locations, as the play is brought to a temporal pause. The positioning seeks to reduce the possibility of a shot reaching the goalpost which could result in a goal. The open play Shot type is taken with an ongoing action in football that doesn’t require the game paused. This implies defending players do not have time to position themselves strategically but must react to the kind of effort the attacker makes to move the ball past them. A corner kick Shot Type usually comes from the edges of the opponent's pitch side. They are usually played with the intention of setting up an attacking teammate who can apply suitable techniques in directing the ball towards a goal attempt. However, this will result in an open play shot and does not define the property of the corner kick shot type the xG model in this project will be using. The use of penalty shots has been excluded from the xG model. This is to avoid penalty shots skewing the data and making the model oversimplified. According to Hewitt & Karakus 2023, penalty shots have an industry-accepted xG value of 0.776. Subsequently, the xG value that will be computed with the model will be regarded as a non-penalty xG value.

Shot Technique: During the shot attempt, attackers will apply varying skills and techniques to the shot they are taking. The dataset provides 7 shot techniques which are normal, lob, volley, half volley, back heel, diving header and overhead kick. Based on the scenario and situation, footballers can attempt to take their shots using a suitable Shot technique. Using the normal Shot technique, the player directly shoots the ball towards the goalpost with as much power as possible. The backheel Shot technique means the attacker played the shot with the heel. The diving header technique demonstrates a shot attempted with the head, with the player diving to get to the ball. The half volley Shot technique is applied to a ball that bounces off the ground while the lob Shot technique tries to go over opposition players. The half volley Shot technique is the opposite of the volley technique which does not require the ball to bounce off from the ground. The shot is taken usually when the ball is in the air without touching the ground. It forms a high arc in its trajectory movement. An overhead kick is a shot technique attempted when the player is having their back towards the goal. In conclusion, the shot technique will be described as a categorical variable within the model and represented as the string data type.

Body Part: The body part defines any allowable part of the body, that has been used to play the shot. The allowable body parts are not penalised for a foul attempt. They include the head, right or left foot and any other part of the body which does not include the two full arms. The body part column is stored as strings in the dataset and treated as a categorical column. The categorical options are left foot, right foot, head and other.

Under Pressure: An attacker is considered under pressure if constantly defended at a close range by a defending player when an attempt on shot is being made. However, the project does not specify the number of defenders on the player but only considers the situation to be true or false. The column is expressed in a Boolean data type and is false for a player freely taking shots without disturbance from any defending opponent.

Position Name: Football players have their designated positions and roles on the football pitch. For the purpose of simplifying the dataset, 4 broad positions will be utilized and formed from the existing information provided. The position will be classified into attacker, midfielder, defender and goalkeeper and the datatype is represented as a string and stored as a categorical variable.

The above features have been described and are required to be separated into what is termed the basic and advanced proposed xG model. Table 1 below explains the features that will be used to train the basic and advanced xG models.

Basic xG Model Features	Advanced xG Model Features
Shot Distance	Shot Distance
Shot Angle	Shot Angle
Shot Type	Shot Type
Body Part	Body Part
	Shot Technique
	Position Name
	Defence Density
	Under Pressure

Table 1 Features used to describe and build the basic and advanced xG models.

3.4 Model Design

The xG model development will consider the use of a classification machine learning approach and is considered binary classification problem. The classifier machine learning models are trained using shot features mapped to an outcome of goal or no goal and compute a probability value for the outcome usually between a value of 0 and 1. The aspect of the classifier models which predicts the shot outcome is not a focal point of the project due to the nature of expected goal values which is a measure of the predicted probability of the shot features. However, the model will be trained with the goal outcome feature which it uses to estimate the prediction probability. Hence, the model focuses on outputting the probability value of the shot features and interpreted as the xG value for that shot. The prediction probability of the supervised classifier machine learning models is usually estimated by the probability value assigned to the outcome close to being in a positive class, in this case, goal outcome class. This technique explains what determines the predicted class in the case of a binary outcome such as the xG model.

The framework of the expected goal model is demonstrated below in Figure 4. Each shot instance contains the shot features that are mapped to an actual outcome that is known. The outcome can be “goal” or “no goal”.

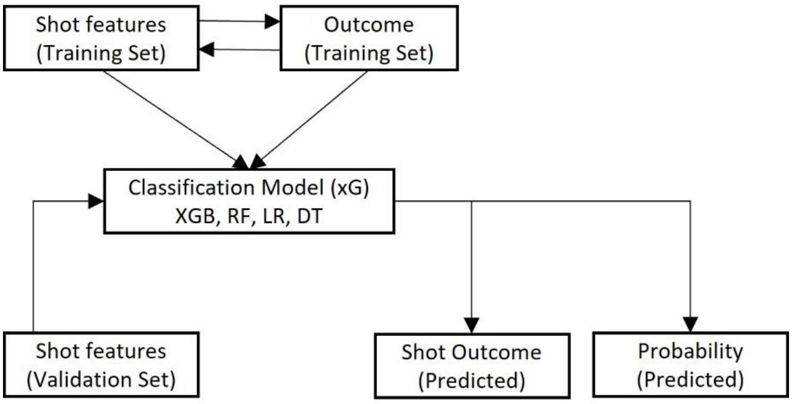

Figure 4 xG model framework describing the shot features and shot outcome probability.

Figure 4 above shows the xG model framework. The shot features is mapped to an actual outcome of the shot (goal, no goal), and fed into 4 different classifier machine learning models namely XGboost (XGB), RandomForest (RF), LogisticRegression (LR) and DecisionTree (DT). The model developed is termed xG model and using the validation set which contains shot instances that are unseen to thew model, an outcome and goal probability can be predicted and estimated. The design of the model will detail the machine learning approach that has been introduced in developing the xG model. The probability outcome of each machine learning model represents the xG value of each shot instance that has been fed into the model. The framework is designed for both the proposed basic and advanced xG models. The consideration for the basic xG model is the basic model features described in Table 1. The advanced xG model makes use of the advanced model features on the right side of Table 1 during training and evaluation.

Several classifier models have been proposed for each instance of the basic and advanced xG model. This is to compare the performance of the classifier models and recommend the most suitable and reliable model for prediction probability which estimates the xG value of each shot. In the later part of the project, the evaluation techniques will be discussed in detail. To answer the final question proposed in the problem statement, the feature importance is obtained from the models and used to explain how the contribution of the introduced advanced features has impacted the performance of the xG model adopted. Since the shot outcome variable is non-continuous, the use of correlation between each feature and the outcome is technically difficult and not recommended. Hence, the solution which has been proposed is to use a technique known as SHAP value discussed in section 3.5, to find the contribution and importance of each feature in the machine learning model adopted. This feature importance ranks the order contribution of each feature to the prediction output of the model.

During the model training, pre-processing techniques will be introduced such as hyperparameters tunning for the purpose of improving the performance of the model.

Hyperparameters are values which a model uses to find the best setting for an introduced dataset. The aim is to model the data in a way that the highest performance can be obtained. To achieve this, the adopted evaluation technique for the performance of the model is introduced as it is very important. Also, scaling of the dataset is highly recommended and will be applied as part of the data pre-processing for modelling.

In the next section, the machine learning classifier models which have been proposed are explained and described. The techniques used within each model to develop the xG model is discussed.

3.4.1 Logistic Regression (LR)

The logistic regression model is a type of discriminative classifier model, used for classifying observation having two classes or multiple classes. This model is discriminative as it tries to learn to distinguish the classes (Jurafsky & Martin, 2023) which is the shot outcome in the proposed xG model. The algorithm is used to detect the relationship between the shot features and the shot outcome and is expressed as a probabilistic classifier which uses supervised machine learning.

The LR model uses a sigmoid function to obtain the predicted probability of the shot features resulting in a goal which is between 0 and 1. A decision boundary which has a threshold value of 0.5 is usually set and determines the predicted class using the probabilistic values obtained from the model. When the value is greater than the threshold value, this indicates a positive prediction outcome or less than the threshold, indicates a negative prediction outcome (Muller & Guido, 2016).

3.4.2 Decision Tree (DT)

The decision tree model is a supervised machine learning model which is also discriminative in nature and can be used to predict classes associated with the data. This model can also measure the probability values which represent the xG value of shot features used to train the model. Although, the DT is not naturally probabilistic in nature compared to the LR without calibrating (Niculescu-Mizil and Caruana, 2005). However, a calibration technique is introduced and discussed in section 3.6. This explains how the DT model has been adjusted to produce a more calibrated probability estimate.

During the training of the decision tree model, the shot features are divided into nodes which output the corresponding class in the shot outcome. During this stage, the outcome of each feature node is regarded as the tree. This goes on recursively mapping the correct leaf to each new node that is formed, and essentially, describes the tree structure. The end of the node will always output a predicted class of goal or no goal (Kotsiantis, 2011).

To determine the xG value of each combined feature at a given node before calibration, the probability value is obtained by estimating the fraction of data points in that leaf node that belong to the corresponding class.

3.4.3 Random Forest Classifier (RF)

The RF model is an ensemble type of supervised machine learning model which uses multiple decision trees (Breiman, 2001) data modelling to create an improved prediction and is now applied in sports analytics to develop an xG model. This technique can create many trees and uses the averaging or voting technique to determine its prediction outcome (Dietterich, 2000). RF is known for its ability to handle high imbalanced (Kaur, Pannu and Malhi, 2019) and high dimensional datasets, making it a good choice for one of the proposed machine learning models which have been adopted to develop the xG model due to football sports low goal to shots ratio.

The RF model will be trained with the shot features using the basic and advanced features. Within the RF model architecture, the algorithm can produce a reliable analysis of the feature importance, which helps identify the influence of each feature proposed to build the xG model. The feature importance analysis will be used to understand the shot features which impact the probability of a shot resulting in a goal.

3.4.4 XGBoost Model (XGB)

XGBoost also known as extreme gradient boosting is an example of tree gradient boosting algorithm, known for its computational speed and model performance (Malik, Harode and Singh, 2020). In the gradient boosting, errors are minimized by ensuring that new weak learners usually decision trees, are corrected by learning from the mistakes of previous learners (Friedman, 2001).

The XGBoost is also known for its ability to handle high-dimensional datasets and can compute the feature importance of the features used to develop the xG model. This answers the question on finding the importance of the features used to develop the xG model.

In conclusion, the proposed machine learning models are relatively powerful and will be used to train and validate the xG model. In the section 3.6, an important concept known as model calibration is introduced and discussed. This calibration is to ensure the suitability of the proposed machine learning frameworks in outputting the predicted probability of the shots known as the adopted xG value.

3.5 Evaluation Metrics

To evaluate the model and examine the feature importance of the shot features, 2 techniques have been introduced namely logarithmic loss (log loss) and Shapley Additive Explanations (SHAP) value.

Logarithmic Loss: The log loss function is a common metric which is used to assess the distance between the predicted distribution and true labels (Aggarwal et al., 2020). This metric evaluates the accuracy of the predicted probabilities when compared with the actual class labels of the data. The log loss metric is an evaluation technique used to identify a perfectly calibrated probability estimate algorithm (Tacon, 2007). The lower the log loss value, the more accurate the model performance.

During the log loss evaluation, the predicted probability is usually referenced towards the positive true label. Meaning, for negative prediction, the probability is considered very small and close to 0. The higher the value of the probability estimate towards 1, the more likelihood of the shot resulting in a goal.

To better understand the interpretation of the log loss metric, the log loss is usually ranging from 0 to a higher number. When the log loss score is 0, it indicates a perfect probabilistic model which can present probability values which are same as the true outcome of the shot. When the log loss has a higher value, it shows that the predicted probabilities are moving away from the true outcome of the shot and producing errors in its predictions.

Consider a sample y with output of 0 and 1, the probability estimate P is expressed as;

$$P(y = 1)$$

The log loss is expressed as.

$$Logloss = -\frac{1}{N}\sum y * log(p) + (1 - y) * \log{(1 - p)}$$

Equation IV

In the equation IV above, N is represented as the total shots in the dataset, y is the true label of the shot outcome (0 or 1), and p is expressed as the predicted probability of the shot outcome.

The final log loss value is between 0 and 1, with 0 being a perfect log loss and signifying a perfect probability model adopted for the xG model. The log loss value obtained is compared with previous work by Mead O’Hare (2023) discussed in the section 2.3 and used to assess the performance of the proposed xG model.

Feature Importance: The use of SHAP interprets the machine learning model and has been introduced to examine the importance and contribution of the features to the model output. According to Miller (2017), interpretability is the level to which a human can understand what has caused a decision. With the machine learning model outputting a probability value for a given shot features instance, it is important to understand the factors which contributed to the outputted probability value. Shapley assigns values that quantify each feature's contribution to the model's probability prediction and uses a cooperative game theory (Molnar, 2020). The Shapley value can have negative or positive impact, with negative Shapley value indicating contribution towards the model outputting reduced probability value, and positive Shapley value contributes towards the model outputting highest probability value of 1. The total Shapley value is the total sum contribution of each instance from the features and this ranges from 0 to a maximum total contribution. It can be described as the marginal contribution of each feature across all possible coalitions. The higher the total Shapley value, the more the feature has contributed to the model's prediction. The lower the total Shapley value of a feature, the less the feature has contributed to the model's prediction.

3.6 Model Calibration

Probability calibration has been introduced to ensure the suitability of all models adopted in outputting a reliable probability value. It limits the effect of bias using boosted tree models where they push the probability values away from 0 and 1. According to (Niculescu-Mizil and Caruana, 2005), boosted trees need to be better calibrated and, as such, will benefit greatly from calibration. Logistic regression is good pre-calibration and may not need further calibration.

When a model is well calibrated, the probability value of a shot resulting in a goal can best be relied upon. Similar shots within the dataset are always expected to have closely related probability values. To examine the calibration of the models, a reliability graph showing a fraction of positive predictions is plotted against the mean predicted value. In a perfectly calibrated model, the line plot demonstrates equal values between the points on the x-axis and y-axis. The calibration plot shown below is an example from the popular scikit-learn documentation website in Figure 5 below. In the plot, Logistic regression is well calibrated as it was closer to the dotted lines, followed by Random Forest.

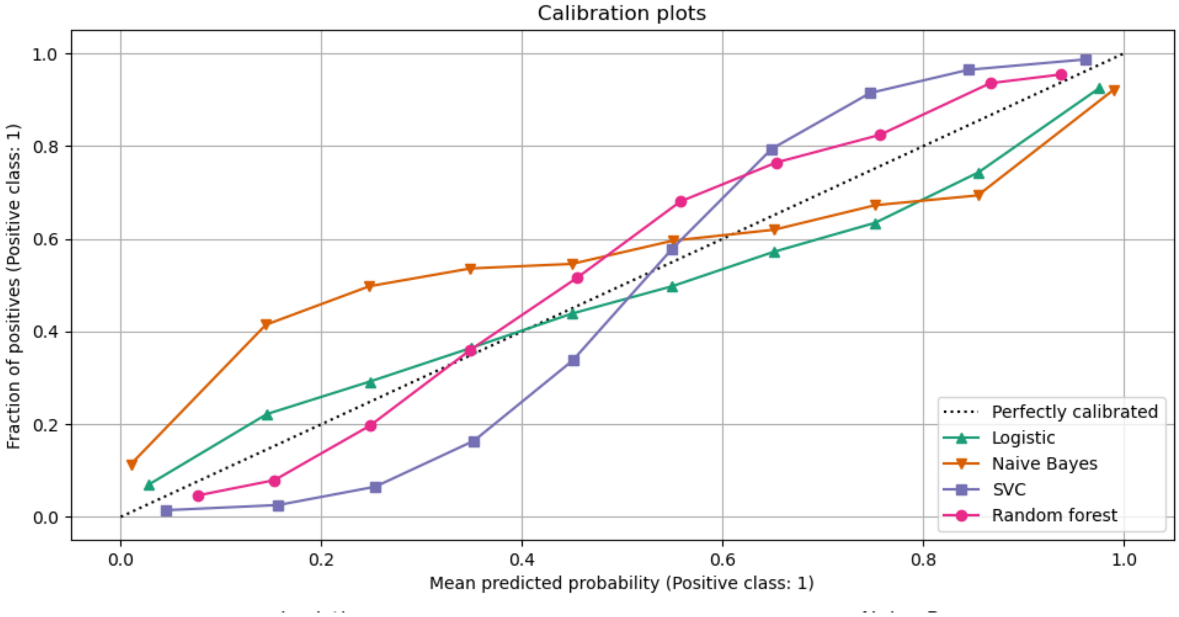

Figure 5 Calibration plot showing the performance of different machine learning model on a binary class task (scikit-learn, 2023).

With the aim of improving the probability prediction performance of the models used in the xG model, two techniques widely used in calibration have been proposed - Platt scaling and Isotonic regression technique. The Platt scaling technique is used for the logistics regression and XGBoost model while the isotonic regression is selected for non-parametric models such as decision tree. Once the model has been trained, a test dataset separate from the training dataset and validation dataset is used to calibrate the model. The validation set is iteratively used to test both calibrated and non-calibrated models on the basic and advanced xG model.

4.0 Experiments and Results

In this section, the exploration of the features will be carried out to understand the relationship between the features and the xG values provided by the StatsBomb data. The exploration will also seek to answer fundamental questions on impact of each feature with respect to scoring chances and goals scored. Next will be to examine the performance of the model and discuss the results. This involves the use of defined metrics such as the log loss and SHAP value to evaluate the accuracy of the model and determine the impact of each feature. The objective is to produce a xG model that has very low log loss value and be able to explain the importance of each feature.

In the last part of this section, the model with the best performance will be used to experiment and test football matches based on La Liga matches. A team-based total xG value will be predicted for 10 seasons between 2010 and 2021. These experimental results will compare the performance of the model with actual xG value and goals scored by the teams.

4.1 Shot Features Analysis

From the analysis conducted so far, a combination of data visualization and modelling has been used to understand the impact of the features on the xG values. In exploring the shot event data, the shot distance was closely analysed with the shot angle. The La Liga matches played between 2018 and 2021 were analysed to understand the effect of Shot distance and angle on Stats xG values. Stats xG value represents the probability of scoring goals assigned to the shots by the StatsBomb xG model and comes with the StatsBomb dataset.

The plots shown in Figure 6 describe the relationship between goal-scoring chance using the Stats xG value, with the corresponding shot distance and angle. The plot shows that smaller values of Shot distance have higher Stats xG value. Looking at the Shot Angle vs Stats xG value, it shows a direct relationship of increase in Shot angles leading to higher Stats xG values.

In conclusion, the plot shows that shots played from close distances, which also form high shot angles will have higher Stats xG values assigned. When shots are played from a far distance, they will typically have smaller angles and result in very small Stats xG values.

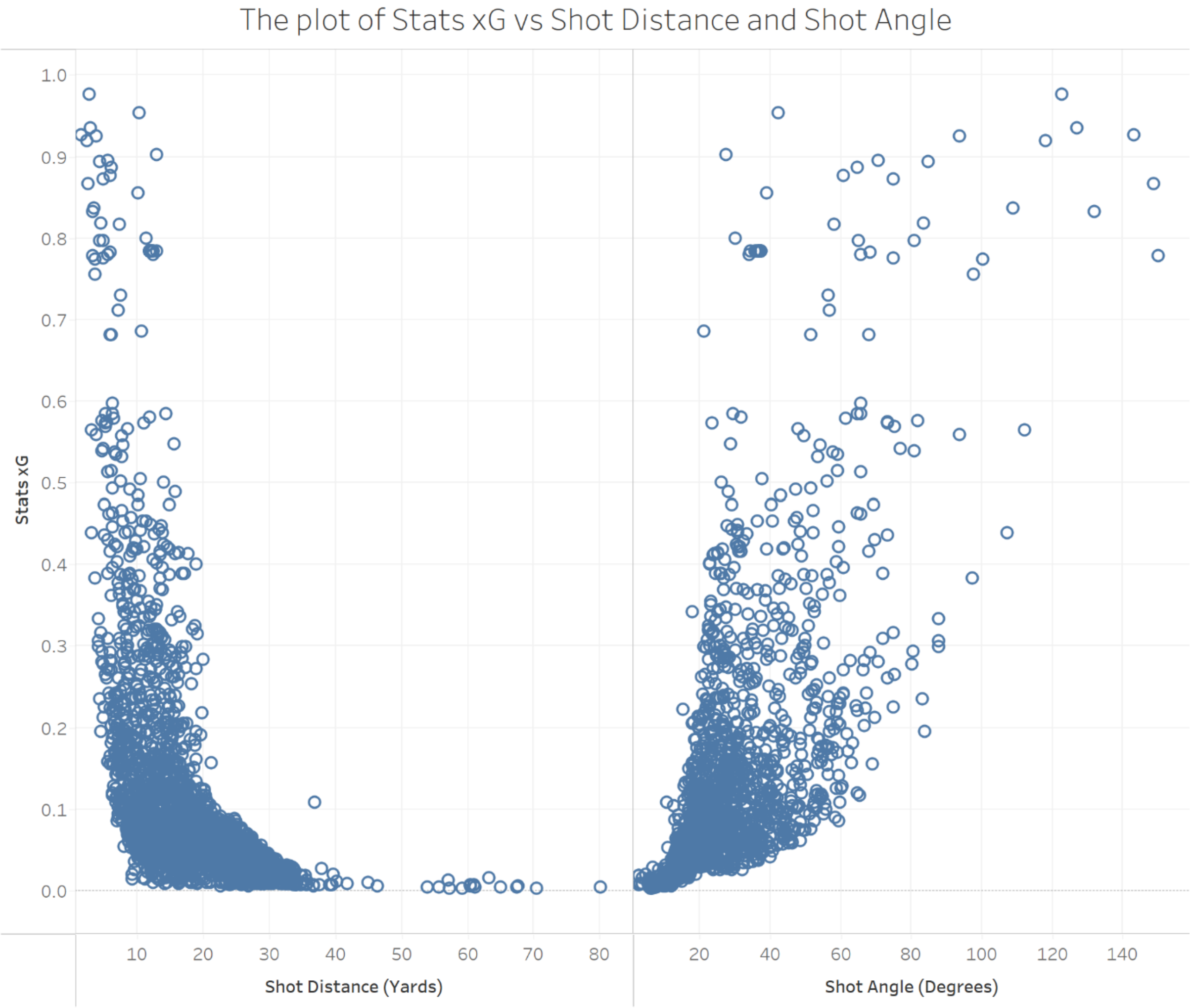

Figure 6 Plot of StatsxG values shown against shot distance and shot angles.

In the Figure 6 above, smaller angles with higher distance have lower xG value. Indicating that to have higher goal scoring chance, the player will need to take the shot from a distance close to the goalposts. On the left, the Shot distance expressed in yards is examined against the Stats xG value. On the right, the shot angle expressed in degrees is observed against Stats xG.

In Figure 7 below, the total number of goals scored and missed when under pressure is shown. This feature was used as part of the advanced xG model. 86.3% of the total goals scored as shown in the plot is when a player is not under pressure. This means there is a higher chance that a shot will result in a goal if there is no pressure from the opponent.

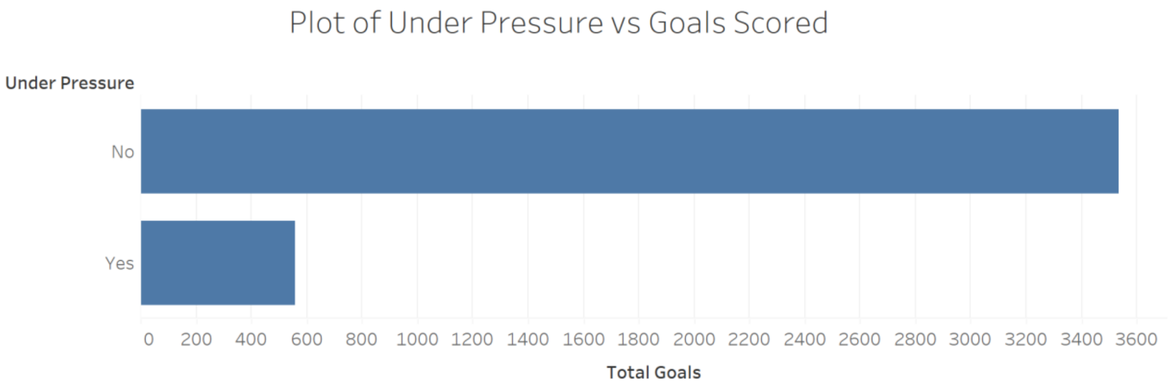

Figure 7 Plot of total goals scored when under pressure and no pressure.

Shot techniques are examined against the total goals scored. In the Figure 8 below, normal shot technique accounted for 78.4% of the total goals scored across all shot techniques. Figure 8 below describes the distribution of goals scored by techniques.

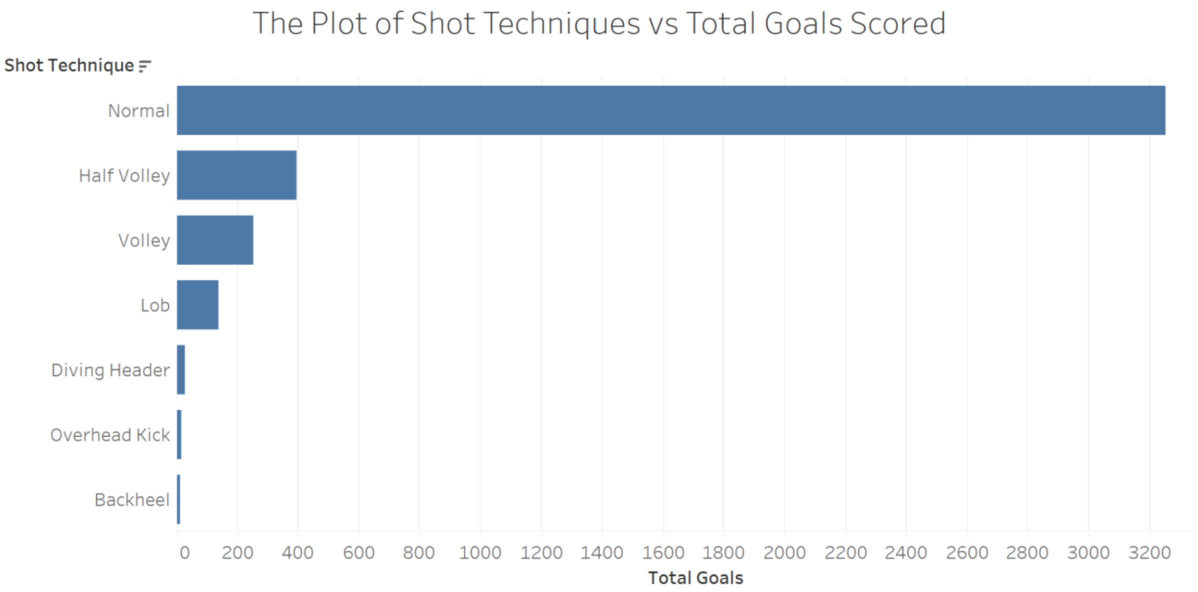

Figure 8 Plot of total goals scored across all shot techniques

4.2 Model Results

4.2.1 Performance Analysis of the Models

In this section, an explanation of the performance of the basic and advanced xG models developed is discussed using the log loss values. Considering different supervised machine learning models used, the observations are outlined and the pre-calibration and post-calibration stages results.

Starting with the basic xG model, the LR, RF, DT and XGBoost supervised machine learning models were used to build the xG model and consistently, logistic regression produced the best log loss value. This was pre-calibration and without the shot events scaled. The FIFA

World Cup 2022, FA Women's Super League 2019 and La Liga 2020/2021 football events were set aside to represent the validation data. The rest of the events contained all leagues mentioned in chapter 3 and between 2010 and 2022. The validation data contained 5,468 instances while the remaining dataset contains 18,572 which was eventually divided into 70% training size and 30% test or calibration size.

It is worthy to note that, onehotencoding transformation technique (Potdar, Pardawala and Pai, 2017) was used on the categorical datasets to improve information gain and make the data more useful for the model. The transformation technique makes categorical data which are not numerical, interpretable for the model. The Figure 8 demonstrates the linearity between the shot distance, shot angle and the Stats xG value. The presence of linearity may have contributed to the LR performance and its probabilistic nature, making it easy to compute goal probability values of each shot. The model was tested on the validation set which is completely unseen during calibration and before calibration of the models. The XGB model gave a promising performance but not as compared to the logistics regression. A look at the shot distance and angle versus the goal outcome indicates a linear relationship, which suggests LR benefitting most from being able to interpret the model better than XGB, RF and DT.

	Pre-Calibration				**Post-Calibration**			
	LR	**RF**	**XGB**	**DT**	**LR**	**RF**	**XGB**	**DT**
Basic xG	0.306	0.777	0.323	5.663	0.306	0.331	0.320	0.345
Advanced xG	**0.293**	0.655	0.310	5.941	0.294	0.314	0.308	0.343

Table 2 Summary of model results using log loss evaluation.

In Table 2, the log loss scores are computed for each machine learning model Random Forest (RF), Logistic Regression (LR), XGBoost (XGB) and Decision Tree (DT) using the basic xG and advanced xG shot features. We can see the change in the log loss value across the models used for the basic xG when not calibrated and calibrated. Before the models are calibrated, RF, XGB and DT performed poorly by having higher log loss values than the LR. Once the models were calibrated, the log loss values reduced as shown in the right side of Table 2. The LR model reduced in performance by having an increased log loss value when calibrated. This is as a result of overfitting and won't be needing a calibration in the subsequent analysis. RF and DT significantly improved in log loss value but didn't outperform the LR and the XGB models. Figure 9 below shows the reliability curve for the machine learning models developed in the project in the basic xG. The reliability curve gave a better understanding of the model that calibrated better to the proposed xG by having its curve near and close to the perfectly calibrated line and can be used to compute the xG values of matches played in the season and competition identified in the validation data.

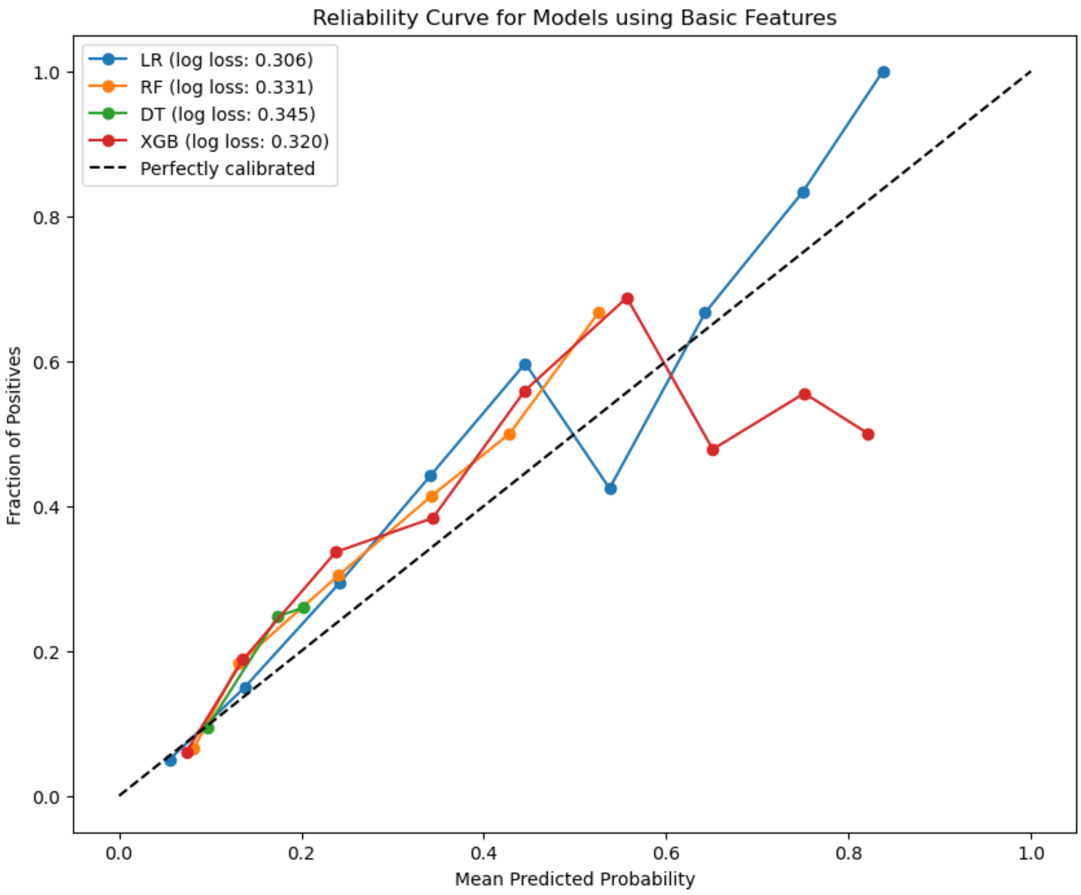

Figure 9 Reliability curve of the machine learning models used to develop the basic xG model.

Discussion on Performance with Advanced xG

The log loss performance seen across all machine learning models in the basic xG developed was not promising. The LR model produced the best log loss value in both basic and advanced xG model, having a log loss value of 0.306 and 0.292 respectively. When compared with related work by (Mead & O’Hare, 2023), their xG model using LR achieved a log loss value of 0.285. Although, it should be noted that fewer data have been used to develop this project’s xG model with majority of La Liga and Women’s football leagues data.

As seen in Table 2, the change in the log loss values of the basic and advanced xG model across all the tested machine learning models indicates that that adding advanced features reduced the log loss by 4.8%. The Figure 10 also shows a well calibrated LR, when compared with the Figure 9 and demonstrating a better xG model. The reliability curve of the advanced xG model is presented and as seen, LR still maintained its highest performance of all models employed. The inference that has so far been picked suggests the huge presence of linearity between the predictors and the outcome and LR is probabilistic in nature. This explains the reason for no improvement when calibrated as compared to other models that suffered poorly in trying to output reliable probability values which is interpreted as the xG value.

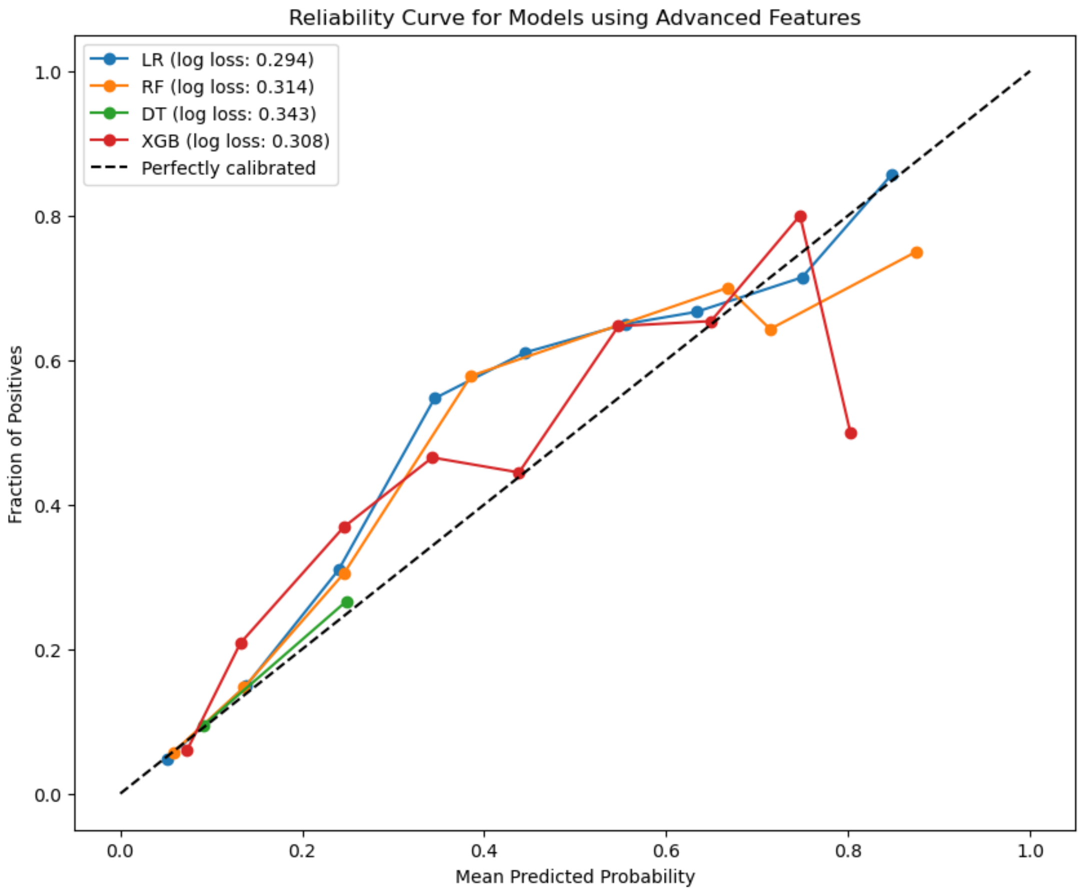

Figure 10 Reliability curve of the machine learning models used to develop the advanced xG model.

From the two model results described using the reliability curve and the log loss value, the advanced xG is more promising and can be tested across the competition introduced in this project.

In the next section, the feature importance of the advanced xG models is discussed which answers the final question of the project, giving an understanding of how important each of the features are to the model and how much contribution each have made.

4.3 Impact of Feature Selection

The impact of the shot features is examined using the SHAP value. A beeswarm plot has been proposed to view the positive and negative impact towards the xG values prediction by each feature. This SHAP value represents the contribution each feature has made to the overall probability prediction outcome.

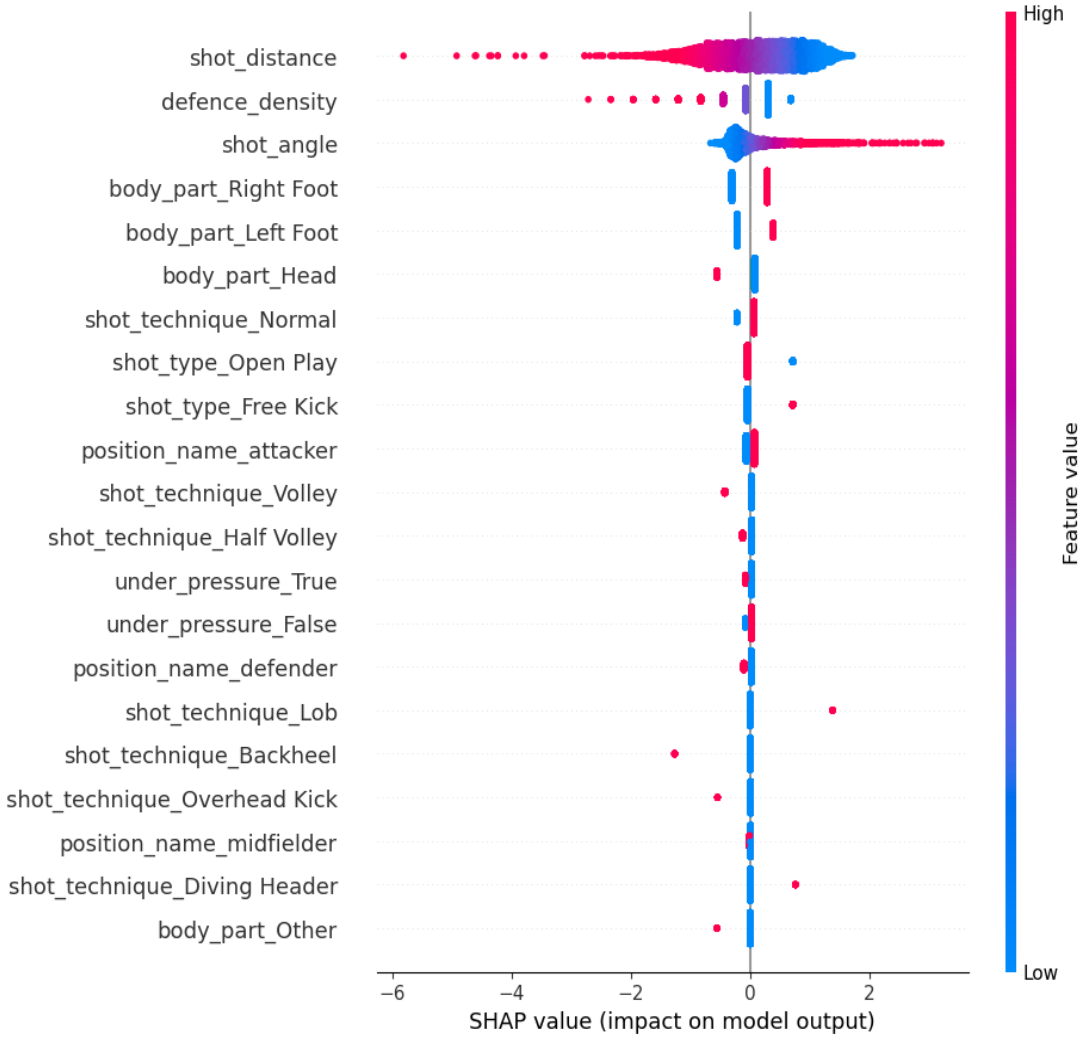

Figure 11 Beeswarm plot showing the impact of features on xG model output using the LR model.

A look at the Shapley values of the most outstanding model (LR) using the advanced features explains the contribution of the features to the model. Figure 11 shows the top features for determining xG and are sorted in order of high to low importance. The plot shows the features value, and impact on the model. The Figure 12 shows the total Shapley value of each feature. The shot distance has a 24.5% contribution to the model output, followed by the Defence density and Shot angle having 12.24% and 11.42% contribution to the model output respectively. These features also have a wide spread of the SHAP value indicating the variations in how they affect the outcome of the xG prediction. To better understand the contribution of each, shot distance is seen to contribute to reduced xG value when the distance is high. A small value of Shot distance contributes towards a higher xG value. The Shot angle is the direct opposite of the Shot distance. Increased Shot angle contributes positively to the xG outcome, increasing the chances of goal scoring. A look at defence density indicates that more opponent players directly in front of the player taking the shot will reduce the chances of goal scoring and looking at the reduced value of defence density, this has contributed to the increased prediction of the xG value. The body part right foot and left foot had 10.6% and 10.2% contribution to the model outcome. While for the head body part, this reduced greatly to 6.1%. The right foot when used indicate the feature value of 1 and shows a positive contribution to the prediction outcome which increases the xG value, although, this may not be the case for players who have weak right foot.

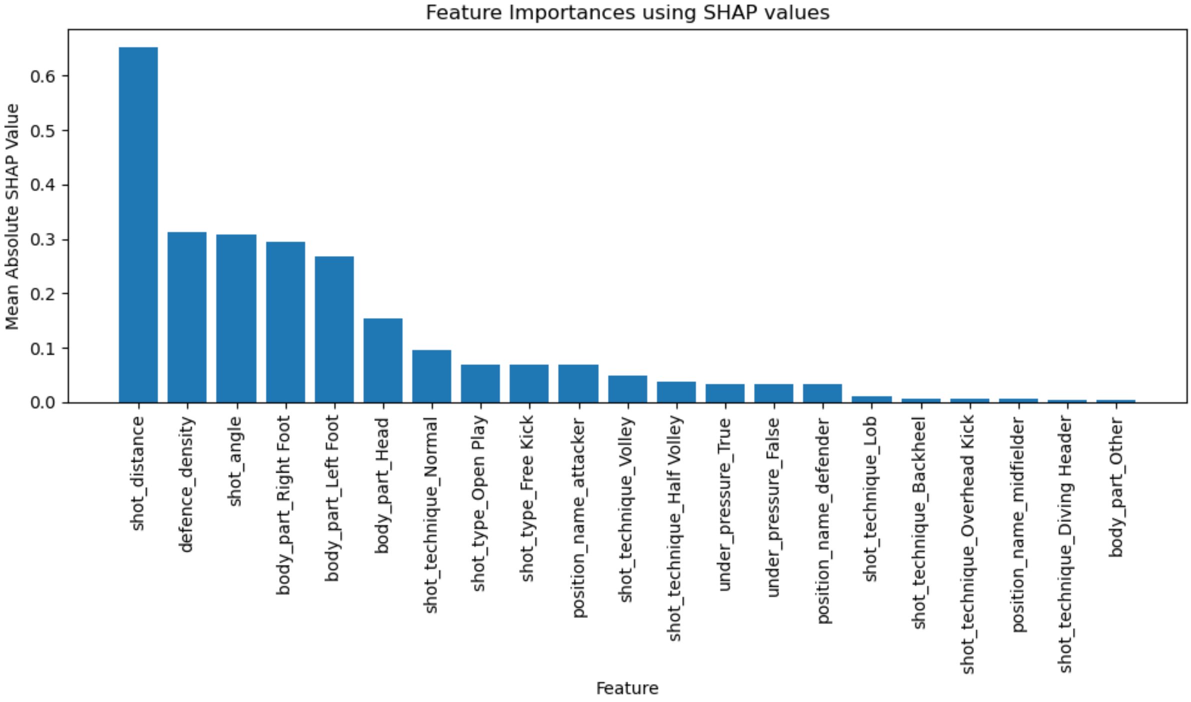

Figure 12 A simplified plot showing the contributions of each feature.

The rest of the features indicate very little contribution to the prediction output. Although, it can be seen in the basic xG model that not using enough features could impact the performance of the xG model. While it is important to recognise the addition of the position of the player, the attacking player contributes positively towards the outcome of xG prediction as seen in the SHAP plot. From the dataset, this is because more attackers take more shots than the rest of the classified position and have a higher probability of scoring.

4.4 Team-Based Results of xG

The total xG values of teams are obtained using the advanced xG model, which sums up each xG per shot played by the teams. This total xG value is used in comparison with the actual goals scored by the team. Important to note that due to the limitation of data availability, the La Liga shots contain matches played by Barcelona against other La Liga teams between 2010 and 2020. The total xG values are obtained and compared with the actual goals scored by the top 10 teams. In addition to the La Liga shots, the FIFA World Cup 2018 and 2022 are evaluated with the advanced xG model and presented in Table 4.

Team Name	Matches	Shots	Model xG	Goals	Goal/xG
Barcelona	341	5198	799.99	850	1.06
Real Madrid	19	266	31.90	22	0.69
Valencia	19	180	22.85	20	0.87
Real Sociedad	20	199	23.27	19	0.81
Real Betis	15	137	15.79	17	1.07
Sevilla	19	198	24.74	17	0.68
Celta Vigo	14	144	19.64	15	0.76
Villareal	16	156	16.5	15	0.90
Atletico Madrid	20	169	16.86	11	0.65
Deportivo La Coruna	11	115	12.93	10	0.77

Table 3 xG values using LR and advanced features, compared with actual goals and goals to xG.

In Table 3, the LR model adopted predicted an xG value of 799.99 from which a total goal of 850 was scored. In interpreting the xG outcome, when the total xG value of a team is less than the total goals scored by the team, it indicates excellent finishing from the team as they are getting very good results such as goal scoring from the outcome of the shots. When the xG value is higher than the total goals scored, this means most of the shot had a good chance of being converted to a goal but that hasn't been the case. A simplified explanation is the goals to xG ratio. Teams who are effective in utilising their chances have a Goal/xG value higher than 1. It demonstrates more goals compared to their expected goals.

The xG values for teams that participated in the 2018 and 2022 World Cup is also evaluated and computed. In Table 4, the FIFA World Cup data shows a breakdown of reported xG values from the shots played and total goals scored by the participating countries.

Team Name	Matches	Shots	Model xG	Goals	Goal/xG
France	14	179	18.39	23	1.25
England	12	154	16.75	21	1.25
Croatia	14	191	19.07	20	1.04
Argentina	11	152	17.20	17	0.98
Belgium	10	142	17.69	15	0.84
Brazil	10	196	25.35	15	0.59
Portugal	9	119	10.01	15	1.49
Spain	8	119	14.76	14	0.98
Netherlands	5	43	5.34	10	1.86
Japan	8	87	9.92	10	1.00

Table 4 xG values of countries that participated in FIFA World Cup 2018 and 2022.

It can be inferred that the advanced xG model was consistent in estimating a lower xG for the teams in comparison to the goals scored except for Argentina, Belgium and Brazil. This means the teams with higher xG value create more chances but scores less goals. When we put this into context, teams like Argentina and Brazil are heavy on attack, but we can see that the xG model estimates that they do not take all their chances well enough.

5.0 Conclusions and Future Works

5.1 Conclusions

In this project, the performance of xG models built using basic and advanced features was evaluated using the log loss metric. In the basic xG model, features such as shot distance, shot angle, shot type and shot technique were used to train and test the model with LR, RF, DT and XGB machine learning models. The LR model performed best by having the lowest log loss value. The performance of models such as DT and RF were considerably poor by having very high log loss value. This led to the introduction of a calibration concept which is used to train models that are not probabilistic in nature. The calibration of the models improved the log loss score of the RF, XGB and DT models. Hyperparameter tuning was introduced to the DT and RF with the aim of reducing the log loss value, the final log loss score was still higher, with LR maintaining a log loss score of 0.294 with the advanced xG model. Hence, the final model adopted was the LR, which didn’t require hyperparameter tuning and calibration, as shown in Table 4.1.

The impact of adding advanced features is seen in the model by having an improved log loss score, which was reduced by 4% from 0.306 to 0.293, although when compared with related work by (Mead O’Hare 2023), their xG model recorded a better log loss score of 0.281. The noticeable difference in this project and their project is their use of dataset with more features and covering top professional leagues like English Premier League, Serie A league and Bundesliga. The log loss value obtained in this project demonstrate that the predicted xG values aligns closely with the true shot outcome and the use of the advanced features improved the reliability of the model to predict xG values. Though the advanced xG model did not surpass the Mead and O’Hare xG model, the advanced model is more reliable than the basic xG model.

Beyond improving the performance of the xG model by adding advanced features, the impact of each feature is analysed using the Shapley value. This aspect provides clarity on understanding the importance and contribution of the features to the model output. The shot distance feature accounts for more than 20% of the total contribution to the model output, followed by the defence density and shot angle. With the defence density having the second highest contribution to the model output, it can be concluded that the addition of the feature is very impactful. This is also clear from the change in the basic and advanced xG model performance.

The project also discussed the application of the xG model to calculate xG values of teams across different competition and used to evaluate the total xG values, which is compared to the total goals scored. This was applied in the context of team performance and player position performance.

5.2 Future Works

While the project has been able to present an xG model built using selected features which showed promising results during its evaluation, it is important to note that xG model can still benefit from further studies and improvement of the existing techniques reviewed within this project. Some of the limitations identified within the project, such as data availability, may have impacted the possible overall log loss score observed. As such, it will be necessary to retrain the model using datasets across the top professional competitions to build a more robust xG model that is able to generalise well.

Information such as player quality should be factored in and used to test the impact it would have on the outcome of xG prediction. Also, factors like match state could describe the current mentality of the team especially one describing if the team is losing or winning and if the team is attempting to overturn a high goal deficit should be included and tested.

Additionally, the use of a neural-network model was not tested in this project. Neural network models are very promising in the contest of probability estimation projects to which the xG is equivalent. Although, the neural network is best suited when there is a considerably large amount of data. To effectively test the outcome of xG prediction using the neural network, more high-quality data will be required.

References

Aggarwal, A., Xu, Z., Feyisetan, O. and Teissier, N. (2020). On Primes, Log-Loss Scores and (No) Privacy. *arXiv.org*. [online] doi:https://doi.org/10.48550/arXiv.2009.08559.

Bangsbo, J., Mohr, M. and Krustrup, P. (2006). Physical and metabolic demands of training and match-play in the elite football player. *Journal of Sports Sciences*, 24(7), pp.665–674.

Breiman, L. (2001). Random Forests. *Machine Learning*, [online] 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems*. [online] pp.1–15. doi:https://doi.org/10.1007/3-540-45014-9_1.

Fortune Business Insights (2023). *Sports Analytics Market Size, Share, Industry & Forecast 2029*. [online] www.fortunebusinessinsights.com. Available at: https://www.fortunebusinessinsights.com/sports-analytics-market-102217.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, [online] 29(5), pp.1189–1232. doi:https://doi.org/10.1214/aos/1013203451.

Hewitt, J.H. and Karakuş, O. (2023). A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer). doi:https://doi.org/10.48550/arxiv.2301.13052.

Kaur, H., Pannu, H.S. and Malhi, A.K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Computing Surveys*, 52(4), pp.1–36. doi:https://doi.org/10.1145/3343440.

Kotsiantis, S.B. (2011). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), pp.261–283. doi:https://doi.org/10.1007/s10462-011-9272-4.

Madrero Pardo, P. (2020). *Creating a Model for Expected Goals in Football Using Qualitative Player Information*. [online] *upcommons.upc.edu*. Available at: https://upcommons.upc.edu/handle/2117/328922 [Accessed 29 May 2023].

Malik, S. (2017). XGBoost - A Deep dive into Gradient Boosting (Introduction Documentation).

Malik, S., Harode, R. and Singh, A. (2020). XGBoost: a Deep Dive into Boosting (Introduction Documentation). doi:https://doi.org/10.13140/RG.2.2.15243.64803.

Mead, J., O’Hare, A. and McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, 18(4), pp.e0282295–e0282295. doi:https://doi.org/10.1371/journal.pone.0282295.

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. doi:https://doi.org/10.48550/arxiv.1706.07269.

Molnar, C. (2022). *Interpretable machine learning: a guide for making black box models explainable*. Munich, Germany: Christoph Molnar.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*. doi:https://doi.org/10.1145/1102351.1102430.

Özaydin, S. and Donduran, M. (2019). An Empirical Study of Revenue Generation and Competitive Balance Relationship in European Football. *Eurasian Journal of Business and Economics*, 12(24), pp.17–44. doi:https://doi.org/10.17015/ejbe.2019.024.02.

Potdar, K., Pardawala, T. and Pai, C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 175(4), pp.7–9. doi:https://doi.org/10.5120/ijca2017915495.

scikit-learn. (2023). *1.16. Probability calibration*. [online] Available at: https://scikit-learn.org/stable/modules/calibration.html [Accessed 27 Jul. 2023].

Tacon, R. (2007). Football and social inclusion: Evaluating social policy. *Managing Leisure*, 12(1), pp.1–23. doi:https://doi.org/10.1080/13606710601056422.

Vovk, V. (2015). The fundamental nature of the log loss function. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.1502.06254.

Wood, D. (2017). *Football and Literature in South America*. Routledge.