**SHOOT YOUR SHOT**

Player Recruitment Strategy based on Expected Goals Modelling

A Project Report submitted in fulfilment of the requirement for the award of the degree of

Bachelor of Science in Economics

Submitted By,   
**MR. SAHIL PAWAR**

(BECO20039)

Under the supervision of

**Dr ASHISH KULKARNI**Assistant Professor

Gokhale Institute of Politics and Economics

**Gokhale Institute of Politics and Economics Pune 411004**

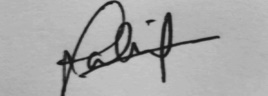
**Declaration**

I, Mr. Sahil Pawar hereby, declares that the dissertation entitled is the outcome of my own study undertaken under the guidance and supervision of Dr Ashish Kulkarni, Gokhale Institute of Politics & Economics, Deemed University, Pune. It has not previously formed the basis for the award of any degree, diploma or certificate of this Institute or University. I have duly acknowledged all the sources used by me in the preparation of this dissertation.

Name

(Sahil Pawar)

(Signature)



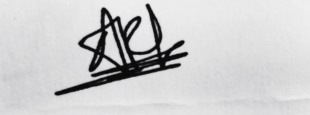
**Certificate**

This is to certify that the dissertation entitled, is the record of the original work done by Sahil Pawar bearing PNR/ registration under my guidance and supervision. The result of the research presented in this project report has not previously formed the basis for the award of any degree, diploma, or certificate of this Institute or of any other Institute or University.

Name of Faculty

Dr. Ashish Kulkarni

Signature of Faculty



Gokhale Institute of Politics and Economics, Pune, 411004

**Acknowledgement**

I would like to include a couple of expressions of gratitude for the people who have helped me right from the beginning of this project. Throughout the written work of this thesis, I have confronted a lot of difficulties and without the support, persistence and direction of these people included, this work would not have been finished. It is to them I owe my most profound appreciation.

.

The accomplishment of this work is an after-effect of sheer diligent work, and assurance put in by me with the guidance of my professor and thesis guide, Prof Dr Ashish Kulkarni. I am extremely grateful for his support and encouragement from the inception of this thesis. I’d also like to thank Mr Udayan Rathore (Visiting Professor at Gokhale Institute) for the initial brainstorming and conceptualisation of this idea, David Sumpter (Professor at Uppsala University, Sweden) for his book‘ Soccermatics: Mathematical Adventures in the Beautiful Game’ and the subsequent online resources, Ram Srinivas (Head of MRKT Insights, United Kingdom) for helping me with retrieving useful football information and for his direction, which immensely helped in understanding the topic at hand and Jayati Sharma (Research Analyst at CPC Analytics, Pune) for the constant inspiration and support throughout the project. I would also like to express my gratitude to friends and family who helped in small but very significant ways throughout the project.

# Contents

[Abstract 7](#_heading=h.30j0zll)

[Glossary 8](#_heading=h.1fob9te)

[List of Tables 9](#_heading=h.3znysh7)

[List of Figures 10](#_heading=h.2et92p0)

[1. Introduction 11](#_heading=h.tyjcwt)

[1.1 Contextualisation 11](#_heading=h.3dy6vkm)

[1.2 Goals and Motivation 13](#_heading=h.4d34og8)

[1.3 Introduction to Expected Goals (xG) 13](#_heading=h.2s8eyo1)

[2. Research Objectives and Research Questions 17](#_heading=h.17dp8vu)

[2.1 Research Objectives and research questions: 17](#_heading=h.3rdcrjn)

[2.2 Research Question 17](#_heading=h.26in1rg)

[2.3 Structure of the Research 17](#_heading=h.35nkun2)

[3. Literature Review 19](#_heading=h.1ksv4uv)

[4. Research Methodology 23](#_heading=h.44sinio)

[4.1 Data Description 24](#_heading=h.2jxsxqh)

[4.1.1 Data Acquisition 24](#_heading=h.z337ya)

[4.1.2 Data Exploration: 25](#_heading=h.4i7ojhp)

[4.1.3 Data Specifications: 30](#_heading=h.1ci93xb)

[Independent variables: 31](#_heading=h.2bn6wsx)

[4.2 Model Description 35](#_heading=h.qsh70q)

[4.2.1 Model Preparation 35](#_heading=h.3as4poj)

[4.2.2 Model Design 37](#_heading=h.1pxezwc)

[5. Model Results and Evaluation: 39](#_heading=h.49x2ik5)

[5.1 Classification Report 39](#_heading=h.147n2zr)

[5.2 Confusion Matrix 40](#_heading=h.3o7alnk)

[5.3 ROC Curve 41](#_heading=h.23ckvvd)

[5.4 PR-AUC Curve 42](#_heading=h.ihv636)

[6. Model Interpretation: 44](#_heading=h.32hioqz)

[6.1 Model Coefficient 44](#_heading=h.41mghml)

[6.2 Correlation Matrix 47](#_heading=h.2grqrue)

[7. Experiments and Discussion 49](#_heading=h.3fwokq0)

[7.1 Descriptive Shot Analysis: 49](#_heading=h.1v1yuxt)

[7.2 Player Recruitment Analysis: 53](#_heading=h.2u6wntf)

[7.2.1 Case Study: Roberto Firmino 56](#_heading=h.19c6y18)

[7.2.2 Case Study: Antoine Griezmann 58](#_heading=h.28h4qwu)

[7.2.3 Case Study: Brentford FC 59](#_heading=h.nmf14n)

[8. Conclusion 63](#_heading=h.37m2jsg)

[8.1 Limitations: 63](#_heading=h.1mrcu09)

[8.2 Future Work 64](#_heading=h.46r0co2)

[Bibliography 65](#_heading=h.2lwamvv)

[Plagiarism Report (Turnitin) 73](#_heading=h.3l18frh)

# Abstract

Shooting in association football is a critical aspect of the game, and its impact on the final scoreline has long been recognized. However, there is much more to shooting than just scoring goals. In this research paper, the author explores the potential applications of Expected Goals (xG) models for analyzing player performances for talent recruitment in football.

The author creates a binary classifier-based xG model using logistic regression, entirely coded in Python using Pandas and Matplotlib and uses historical data from Wyscout for five seasons across Europe’s top 5 domestic competitions to determine the likelihood of a shot being a goal. By analyzing the characteristics of a shot, such as its location, the type of shot and the context of the game, the model calculates the probability of the shot resulting in a goal. The binary classifier allows for a more straightforward interpretation of the model's output, with each shot classified as either a goal or not a goal.

Using the xG model, the author examines the performances of individual players and teams. The model allows for a more in-depth analysis of a player's contribution to the team, beyond just the number of goals they score. Additionally, the author demonstrates how football clubs can use the xG model for talent recruitment, identifying promising prospects who may have been overlooked by traditional scouting methods. Further the research consists of 3 real life case studies of players and clubs that showcase how the methods discussed in this thesis can be applied and validated.

Overall, this research highlights the importance of shooting in football and demonstrates the potential of data-driven analysis in the sport. This research has important implications for coaches, players, and clubs looking to improve their performance on the pitch.

# Glossary

* **Goal**: A goal occurs when a shot taken by a player goes inside the goalposts.
* **Assist**: An assist is a final pass which is made by a teammate before the shooter takes a shot which leads to a goal.
* **Event Data:** Data which includes a record of every event that happens on a football field. Position of players, data regarding a shot, pass or a goal is included in Event Data.
* **Tracking Data:** Tracking data always gives us detailed insight into every player’s positioning and movement on the pitch. This includes gauging the movement of every player throughout the game, irrespective of whether they have the ball at their feet or not.
* **Header**: A header in football is a goal that is scored by the head
* **On the ball events**: “On the ball” refers to any action that takes place when the player has the football at his feet
* **Finisher:**  The person who scores the goal
* **Dead ball Situation:** a dead ball situation refers to a moment during the game when the ball is out of play, and play has stopped. This can occur for a variety of reasons, including when the ball has gone out of bounds, when a foul has been committed, or when the referee has blown the whistle to indicate a stoppage in play.
* **Open Play:** Open Play refers to the phase of the game where both teams are in possession of the ball and are actively trying to create scoring opportunities. Open play are events in football which are not preceded by a dead ball situation.

# List of Tables

|  |  |
| --- | --- |
| Table Number | Table Name |
| 1 | Example of xG use-case in match outcomes |
| 2 | Data description for the research |
| 3 | List of Variables |
| 4 | Classification Report |
| 5 | Model Coefficients |
| 6 | Example of correlation between variables |
| 7 | Best Finishers: based on the difference of Actual Goals and xG |
| 8 | Best Finishers: based on highest ratio |
| 9 | Top Young Players based on the xG Model |
| 10 | Roberto Firmino at Hoffenheim |
| 11 | Roberto Firmino at Liverpool |
| 12 | Antoinne Griezmann at Real Sociedad and Athletico Madrid |
| 13 | Brentford Player Sales (Last 6 Years) |

# List of Figures

|  |  |
| --- | --- |
| Figure Number | Name of the Figure |
| 1 | Outcome of Shots |
| 2 | Distribution of shots |
| 3 | Proportion of Shots resulting in a goal |
| 4 | Goals/no goals based on country |
| 5 | Goals/no goals based on the season (year) |
| 6 | Angle of the shot from goal |
| 7 | Angle and distance of shot from goal |
| 8 | Confusion Matrix |
| 9 | ROC Curve |
| 10 | PR Curve |
| 11 | Correlation Matrix |
| 12 | Scatter Plot for best finishers in Europe |

# 1. Introduction

## 1.1 Contextualisation

“You have got to shoot otherwise, you can’t score”

* Johan Cruyff, former professional footballer, and coach

Association Football, like basketball and hockey, is an invasion sport. Two teams go against each other and invade the opponent’s territory to score points or goals. The team with the most goals at the end wins the game. This is the understanding that any random individual has about sports. Sometimes you would look at the scoreline at the end of the game to know which team won. Yes, the final scoreline tells you which team won, but it does not necessarily tell you which team played ‘*better*’. One could apply the same logic to gauge player performances as well. For example, at the end of an *‘El Clasico*’ game, people would ask, “How many goals did Messi score?” and “How many goals did Ronaldo score?” and conclude that the player who scored the most goals of the two had a better game. Until the last decade, only basic football metrics were used to determine and measure performances and match outcomes to explain what happened in a match or, as mentioned in the above example, to gauge individual performances.

The advent of advanced technologies has allowed for the quantification of athletic actions in sports, leading to the rise of sports analytics. Despite being the largest sport globally, association football has yet to be at the forefront of this data revolution, lagging behind other major sports in analytical modelling quality. The reliance on recent results and emotions, rather than data-driven insights, remains a significant factor in decision-making in football.

Football's complexity, characterised by its low-scoring and dynamic nature, presents a significant challenge in modelling efforts. At present, the most utilised statistics for evaluating team performance and match events are basic measures such as shots, shots on goal, ball possession, and the number of passes (Brooks, Kerr, and Guttag 2016). These metrics, however, provide an incomplete and narrow view of a game and are insufficient in accurately explaining its outcome. The sophisticated nature of football calls for both quantitative and qualitative performance indicators. (Tiippana Tuomas, 2020).

Several metrics, including ball possession percentage, shots taken, and player distance covered, are commonly utilized, particularly by sports media outlets. Even though these metrics do give a great account of what happened in the football game, reliance on only them can often be misleading as they do not tell the full story and just tell numbers which project the face value of the game. For example, at the end of a game, you could wrongly assume that a team (Team X) with a much higher number of shots than the opponent (Team Y) dominated the game offensively compared to their opponents. But it could so happen that Team X took most of these shots from positions on the football field, from where the chance of scoring was close to zero. Of course, on paper, it would say that Team X had an overall better game since they took more shots than Team Y, but as mentioned, that isn’t necessarily true. The main idea to understand is that chance plays a greater role in determining the number of goals scored by each team compared to events such as shots and corners. These latter events may be more indicative of the overall quality of the teams' performances. (Wheatcroft, 2020).

The evolution of analytical metrics has aided in the generation of increasingly accurate insights from data. The book "Moneyball: The Art of Winning the Unfair Game" by Lewis (2004) and the subsequent movie, which depicted the story of an underdog baseball club in the 2000s, brought advanced analytics in sports to the mainstream. In his account, Lewis describes how the underdogs managed to qualify for the playoffs two years in succession, despite operating with a limited budget. The team achieved this feat by exploiting statistical metrics that were validated through analytical methods, which were previously disregarded by the conventional baseball community. This methodology was effective, cost-efficient, and progressive at the same time and gave Oakland an edge that other baseball clubs lacked.

The use of data analytics found its way to football soon after. Although the application is newer in football, there have been significant successful developments in football analytics. With time, an increasingly high amount of football clubs have started implementing this in their systems.

One of the major developments in football analytics has been the creation of the Expected Goals (commonly known as xG) metric (Lucey et al., 2014). In his article titled “Assessing the performance of Premier League goal scorers”, Sam Green (2012) created an analytical framework known as Expected Goals (xG) that involves estimating the likelihood of a goal being scored from a given goal-scoring opportunity.

The xG value of any given shot is the probability of that shot leading to a goal. In simpler terms, Expected Goals (xG) is a metric to measure the quality of any given shot. xG allows us to look at the game from a qualitative measure, and not only from a quantitative measure (which would have been the case if we only considered the number of shots).

## 1.2 Goals and Motivation

Over the course of this research thesis, I bring to light how, just like in baseball, this advanced data analytics approach has revolutionised football. To accomplish this, I explore the xG metric in-depth and go on to develop my own Expected Goals model using binary classification, to provide a more precise and accurate way to assess football games which will give us a framework to effectively assess individual players based on their performances and future potential of their growth as well as analysing team performances.

In order to achieve this objective, I undertake an in-depth analysis of the xG metric and develop my own Expected Goals model through binary classification. The proposed model aims to provide a more precise and accurate approach to evaluating football games, offering a framework for the effective assessment of individual players based on their performances and potential for growth, which is vital for player scouting and recruitment.

## 1.3 Introduction to Expected Goals (xG)

Expected goals (xG) is a statistic used in football to evaluate the quality of a scoring chance. It is based on a prediction algorithm that quantifies the probability that a shot will result in a goal, based on a number of factors such as the distance and angle to the goal, the body part used to take the shot (e.g. header, foot, etc.), state of the game and the type of assist leading up to the shot amongst other variables. The idea behind xG is to provide a more accurate and detailed analysis of a team's attacking and defensive performance, beyond just the number of goals scored or conceded.

The xG model has three major use-cases in Association Football (Tiippana Tuomas, 2020). They are

1. Analysing performance of playing teams
2. Analysing the performance of players
3. Identifying prospective talents to recruit in order to strengthen the football club

First is to assess the performance of teams playing. Since football is a low-scoring sport, randomness or luck often plays a significant role in the outcome of a game. Because of this, it could so happen that a team plays very well, creates many goal-scoring opportunities and were “the better team” in that game in all aspects but still end up losing. The expected goals model provides us with a deeper insight into the game beyond the superficial scoreline.

One of the main advantages of xG is that it considers the difficulty of the chances that a team creates and faces. For example, a team that consistently creates high-quality chances (i.e. high xG) but fails to convert them may still be performing well, while a team that scores a lot of goals from low-quality chances may not be as effective in the long run. Thus, in layman's terms, the expected goals model tells us “which team was more likely to win the game in the light of created goal scoring opportunities, and second of all, how well could the teams convert these opportunities into actual goals.” (Tiippana Tuomas, 2020)

xG has been widely adopted in football analytics in recent years and has been shown to be a reliable predictor of future performance. Studies (Zhao & Zhang, 2019) have found that teams with high xG ratios (i.e. the ratio of a team's xG to its actual goals scored) tend to overperform in the short term, while teams with low xG ratios tend to underperform. This suggests that xG can be useful for identifying teams that may be due for a regression to the mean, or for identifying teams that are outperforming their underlying statistics.

For example: On Jan 27th, 2021 Burnley FC played against Aston Villa. Burnley won the match by a scoreline of 3-2. In table 1, the observed and expected goals of Burnley FC and Aston Villa FC are shown. As we can see, Burnley did not produce as many expected goals as Aston Villa, but they were able to score and win the game by capitalising on their opportunities. Considering the difference in the expected goals, it could be said that Burnley was lucky to win the match. This shows us that, although on the scoreline it says that Burnley won the game, which will be great news for Burnley supporters, Burnley must not get carried away by this result as in reality, they have been very poor in their performance in this match and have the factors of randomness and luck to thank for winning the game. In Aston Villa’s case, although they lost the game they performed very well. Of course, simply playing well on xG does not win you the game, so Aston Villa must now focus on making sure that they convert the chances that they create, so that they can match their good performance with equally good results.

|  |  |  |
| --- | --- | --- |
| x | Burnley FC | Aston Villa |
| Actual Score | 3 | 2 |
| Expected Goals | 0.5 | 2.6 |

**Table 1: Example of xG use-case in match outcomes**

The second aim of the expected goals framework is to evaluate the individual players' performances. Each player is assigned a value for expected goals and expected assists, which enables the analysis of three aspects of their performance. Firstly, the model assesses the player's contribution to their team's goal-scoring probability, which is indicated by the value of their expected goals. Secondly, the model measures the player's ability to convert their expected goals into actual goals, which is computed by subtracting the actual number of goals scored from their goal expectancy. Finally, the model also demonstrates the player's ability to help their teammates create better opportunities for scoring goals. This is represented by the expected assists metric, which is derived from the expected goals model.

The third and final application of the expected goals model is for talent identification and recruitment, in order to enhance the overall performance of the football club. In modern football, xG is widely used as a tool to evaluate the quality of a player's attacking contributions. By quantifying the probability of a shot resulting in a goal, xG can provide a more detailed and accurate analysis of a player's scoring or conceding record, beyond just the number of goals they have scored or conceded.

It can be used to evaluate players in several ways, including scouting. One way expected goals can be used in player scouting is to identify players who are creating high-quality scoring chances for themselves and their teammates. By looking at a player's xG, scouts can get a sense of how effective they are at creating chances and how likely they are to score in a given situation. This can be especially useful for identifying players who are underperforming in terms of their actual goals scored compared to their expected goals, as it may indicate that they are due for positive regression in the future.

For example, a player who consistently creates high-quality chances (i.e. high xG) for their team may be more valuable to recruiters, even if they have not scored a large number of goals. Similarly, a player who consistently concedes low-xG chances may be more valuable to a recruiter looking for a strong defensive player.

However, xG should not be used in isolation when evaluating players. Other factors, such as a player's physical and technical abilities, their tactical understanding, and their team and opponents, can also have a significant impact on their performance. Therefore, although xG provides a very important metric of understanding the player, recruiters consider xG as just one part of a broader analysis of a player's potential value.

For the purpose of this research thesis, we will be only focusing on the use cases of analysing the performances of players and identifying them as prospective talents to effectively recruit them. More on this is discussed in Chapter 2.4 (Structure of the Research)

# 2. Research Objectives and Research Questions

## 2.1 Research Objectives and research questions:

1. To examine and quantify the quality of shots taken in top-flight football competitions and create an xG (expected goals) model for the same to help eliminate randomness in quantifying a football game.
2. To determine the factors that influence the quality of a shot that results in a goal
3. To conduct a player recruitment analysis based on the output of the model

## 2.2 Research Question

How can we quantify the quality of shots taken during a football game and use them for analysing player performances and recruiting football players more efficiently?

## 2.3 Structure of the Research

This thesis comprises nine chapters, which present a comprehensive analysis of the use of Expected Goals (xG) in player recruitment strategies.

Chapter 1 introduces the topic at hand and provides a foundational understanding of football’s relation with data analytics and how the Expected Goals model actually works. Chapter 2 comprises the research objectives and questions being discussed in this research thesis.

Chapter 3 provides a detailed review of the relevant academic literature and related works in the field of xG and player recruitment.

In Chapter 4, the research methodology is extensively discussed, focusing on the theoretical framework and rationale for the use of a binary classification logit model. This chapter is further divided into two sections: the Data Description section, which outlines the sources of data, dataset exploration, and methods of data acquisition, and the Model Description section, which elaborates on the model preparation and design.

Chapter 5 showcases the results of the model, providing a thorough discussion of the classification reports and other model evaluation methods. In Chapter 6, the findings are analyzed in greater detail, and interpretations are made based on correlation, coefficients, and feature analysis.

Chapter 7 explores the experiments and analyses conducted based on the xG model developed. This section expands on the model's analysis and presents three case studies demonstrating the efficiency of the model in player scouting.

Lastly, the study's limitations are discussed, and future research topics are suggested in the conclusion of this research thesis. Overall, this thesis contributes to the advancement of xG modelling in football and sheds light on its potential to revolutionize player recruitment strategies.

# 3. Literature Review

Football has a low-scoring nature, which means that randomness and luck often have a big impact on the scoreline and the end result (Lucey, Carr , Monfort, Bialkowski, & Matthews, 2015). As noted previously, the outcome of a football match does not always reflect the true performance of the competing teams. Randomness or chance, as indicated by Noordman (2019), can exert a significant influence on the result. In the longer run, luck or randomness cancels out which exposes gaps and holes in a team’s performance which were not reflected in the results in the shorter term. Because of this, a number of metrics (Fernandez and Bornn, 2018; Hamilton, 2017) have been developed over the last decade to try and eliminate this randomness, or at least minimise it.

Numerous studies have been conducted to examine and explore the problem at hand. Since shots are a more common event than goals, shot-based metrics are less based on randomness and hence considered as a better predictor than goal-based metrics (Goodman, 2018). To achieve the objective of this project, we will estimate the probability of a shot leading to a goal. One can also see this as a way of quantifying the efforts of the player on that specific action (Routley, 2015).

The concept of performance analysis, which is the root of the xG metric, is deeply studied in the paper by [Reep and Benjamin, 1968] where they discussed the concepts of skill and chance in shots leading to goal. Although there was no explicit mention of the term ‘Expected Goals’, this lay the foundation for an area of study which was later again worked on by Reep and a research fellow in their paper ‘Measuring the Effectiveness of playing strategies at Soccer’ (Pollard and Reep, 2002). In their paper, they discuss how using a hypothetical system which records on-the-ball events can be used to disassemble the game into a series of team possessions events, which can be represented by a quantitative variable that represents the probability of a goal being scored. This variable can be used to evaluate both the expected outcome of a team possession originating in a given situation, as well as the actual outcome of the possession (Brechot, 2020). In this paper, they outlined a number of factors like the distance from the goal, angle, space enjoyed by the striker, number of touches before taking the shot, type of build-up play, and location of the assist. These factors were used in their model to evaluate the quality of a goalscoring chance.

Much academic research in performance analysis was conducted for football and other invasion sports for decades to come, with the first known academic publication of the Expected Goals model to be published by [Lucey et al., 2014]. Lucey built a univariate logistic regression model to analyse the shots taken based on binary classifications such as location, distance, and context of the game.

It is noteworthy to mention the use of various football recording systems throughout history. Despite the abundance of data available, the application of a statistical methodology to analyse such data has been sparse. In the early 1980s, a system which directly records every action of a game was developed in Canada (Franks et al., 1983), with subsequent analysis and consideration of relevant statistical aspects (Franks, 1988). Similar game recording methods were subsequently adopted in England (Hughes, 1988), Finland (Paukku, 1994), the USA, and other countries. The match analysis section of the abstracts and proceedings from the World Congress of Science and Football in the 1990s provides a comprehensive overview of the diverse attempts made to analyse team performance (Reilly et al., 1988, 1993).

A similar solution to this problem was made by [Eggels, 2016]. Eggels considered the use of predictive modelling to explain the outcome of a match based on the available data from the match (rather than trying to predict the outcome of the game before the game starts). Instead of just using the classical logistic regression to develop a solution, Eggels used 3 different tree-based classifiers along with the logistic regression to test out which classification gave the best results. They also made use of the individual skills of the shooting player and goalkeeper which he obtained from EA Sports’ FIFA video game database. Developments were made in testing out alternate variables to further make the model more accurate in its prediction.

An astonishingly accurate model created by (Rathke, 2017) predicted that shot likelihoods of Bundesliga (German domestic league) and Premier League (English domestic league) events only took distance and angle from goal as their key features (Raudonius L. et al., 2022). Besides this, one such contribution was made by [Pollard, Ensum, Taylor; 2004] with their study that found distance to the goal and to the nearest defender to be the most important variables that help predict whether a shot will be successful.

Van den Hoek (2019) in his paper created an xG model based on player positional data to assess the expected outcome of a match. Den Hoek employs an extensive array of Poisson distribution simulations to determine the most likely number of goals that will be scored by each team. The resulting values are used to predict the expected outcome of the match, which can then be compared to the actual outcome to evaluate its accuracy. The research findings indicate that a single-match outcome frequently deviates from the projected outcome based on the scoring opportunities presented by both teams, as discussed in den Hoek's paper.

The study by Trainor & Chappas (2013) on Goal Expectation and Efficiency in Soccer defines shooting efficiency as the ratio of actual goals to expected goals. In other words, shooting efficiency is calculated by dividing the number of actual goals scored by the number of expected goals that a player or team was predicted to score based on the quality of the chances created. The study examines how shooting efficiency can be used to evaluate player and team performance in football.

Some novel research has been done to expand the metric of Expected Goals beyond the variable of shooting. In recent years research has been done on taking into account the context of how and why the shot was taken, notable in assessing the values of passes by (Power et al. 2017) where he introduces a supervised approach using hand-crafted features to measure the risk and reward associated to a pass and also the entire attacking phase even if it does not result in a goal-scoring attempt (Seidenschwarz, Link, and Lang, 2016; Bornn, Cervone and Fernández, 2019; Spearman 2018). The xG value is the central component of all the statistical models mentioned, and they are all designed to a significant extent around the probabilities of scoring goals.

In the last few years, the concept of spatiotemporal data has begun to take shape in expected goals modelling. Spatio-temporal data refers to data that captures not only the movements and actions of the player in possession of the ball but also those of all players on the football field throughout the match. Lucey et al. (2015) presented a shot prediction model based on spatiotemporal data to obtain highly strategic results using variables like how close the defender was, the playing chemistry between the surrounding players and the speed of play during different parts of the game. These features contributed to an enhanced assessment of the probability of a shot culminating in a goal. A few years later, football researcher W. Spearman (2018) wrote a paper where he used spatiotemporal data to analyse the off-the-ball movement of players which immensely helped in calculating off-the-ball scoring opportunities, hence enriching the expected goals values. To explain this with an example, imagine Player A passes a 20-yard pass inside the penalty area, which Player B scores with a header. Of course, Player B scored a goal, but what is important to note is how Player B ran inside the penalty area, made space for himself against the opponent's defenders and made sure he was in a position where Player A could pass the ball, before himself scoring the goal. His off-the-ball movement to be in the right place at the right time contributed to the quality of the chance (improving his xG).

Over the last couple of years, important decisions such as player recruitment (Burn-Murdoch, 2018) or match preparation (Noordman, 2019) are based on data instead of intuition. In the realm of professional team sports, the primary objective is to effectively translate financial resources into on-field success, specifically converting wages into victories. When faced with limited wage budgets, teams must strive for increased efficiency in how they allocate their resources to compete with better-resourced opponents. This requires identifying players who offer the most value for their wages, namely those who can contribute the most to winning in relation to the amount of compensation they receive.

The xG framework has demonstrated its utilitarianism in the football domain, particularly in Danish clubs. FC Midtjylland succeeded in securing their first Danish league championship by using this approach for player recruitment (de Hoog, 2015). Although they kept their algorithm and computations undisclosed for understandable reasons, this outcome indicates that statistical methods can be effective in improving the prospects of football clubs by offering coaches and clubs such analytical tools.

Thus, we can see that clubs’ decision-makers are likely to exhibit an outcome bias, where they underestimate the role of randomness in match outcomes and assign too much weight to the observed outcomes in their performance evaluation (Baron & Hershey, 1988; Gauriot & Page, 2019; Brechot 2020). As a result, decision-makers tend to not make necessary adjustments after lucky wins and overreact after unlucky losses.

# 4. Research Methodology

Classification is the task of learning a target function (*f)* that maps each attribute set x to one of the predefined class labels (y) (Rokach, 2009; Pollard, 2002). In the case of the expected goal model, the class labels are goal or no goal.

Therefore,

Y ∈ {0,1} (where 0 represents no goal and 1 represents a goal)

The dataset is denoted by D = {(*xi, yi*)} n i=1 where each sample (*xi , yi*) is independently sampled from the joint distribution with density p(x, y) which includes an instance *xi* ∈ X and a label *yi* ∈ Y . The goal of a binary classifier is to train an optimal mapping function.

In the context of classification problems, there does not exist a singular, optimal classification algorithm that surpasses all other available algorithms. To determine the most appropriate algorithm for a specific problem, one approach is to utilize cross-validation to determine the most effective classification algorithm.

Based on these metrics, the best model can then be selected. For this study, a binary classifier is needed to create our model. Anticipating the optimal classifier for our model which will produce the best results for our model can be challenging. My choice for the machine learning algorithms for our xG model is based on previous research by (Van den Hoek, 2019) and (Eggels, 2016). For the sake of simplicity and reliability, I chose to build a logistic regression model for this project. It is a good fit especially for this project, as it works well with linear dependencies.

I will further be describing my research methodology in depth in the next two subsections. In subsection 4.1 I discuss the description of my data ie. how the data was acquired, the scope of the dataset and the independent variables I will be using for my model. In subsection 4.2, I discuss the description of my xG model ie. the design of the model, how it was prepared etc.

## 4.1 Data Description

### 4.1.1 Data Acquisition

In the context of football analytics, tracking data and event data are two primary sources of information that are used to extract insights about the performance of players and teams. Tracking data provides live positioning information about players during specific passages of play, while event data capture discrete actions that occur during a match, such as a pass, shot, or tackle. While tracking data has the potential to provide more detailed and nuanced insights into player and team behaviour, it is often difficult to collect and organize. As a result, models using tracking data may have limited numbers of shots, which serve as observations in statistical models.

Given these challenges, for the current project, we chose to work with event data, which is more abundant and more minutely monitored. While this type of data may not provide as much detail as tracking data, it is still a valuable source of information that can be used to identify key patterns and trends in player and team performance. By focusing on event data, we aimed to develop a more accessible and practical model for evaluating the quality of shots taken during a football game and to use this information for analysing player performances and recruiting football players more efficiently.

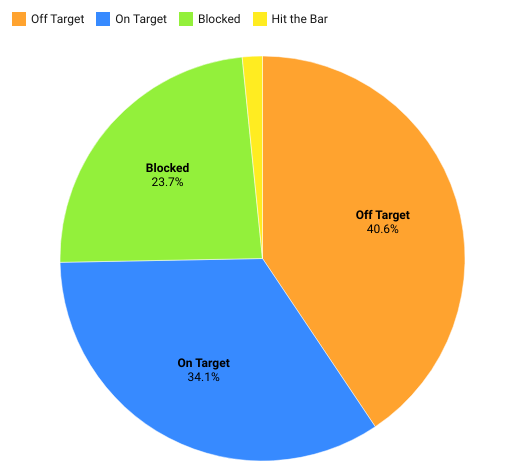
For the sake of this project, I approached Wyscout for their event data dataset. Wyscout is among the preeminent global companies offering event data services in association football. They analyse over 250 leagues and competitions worldwide using their video-based analysis technology and collect data which is widely used in the world of professional football. The dataset includes a detailed view of 9,074 games, totalling 9,41,009 events and 2,28,498 shots from the five biggest football competitions: Bundesliga (Germany), Serie A (Italy), La Liga (Spain), Ligue 1 (France) and Premier League (England) from 2011/2012 season to 2016/2017 season.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Games | Events | Shots | Goals |
| Event Data | 9074 | 9,41,009 | 2,28,498 | 24,446 |

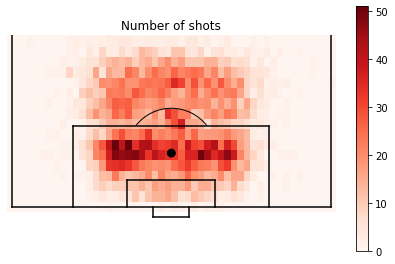
**Table 2: Data description for the research**

### 4.1.2 Data Exploration:

Since our model is associated exclusively with shots and the characteristics associated with a shot leading to the goal, let us explore the data and see the distribution of how different possible outcomes a shot can have.

**Figure 1: Outcomes of Shots**

As we can see from the pie chart above, most of the shots are off-target, resulting in more than 40 per cent of them. Intuitively, this makes sense as the number of shots taken in a game is generally much higher than the goals scored in a game. For context, the average number of shots taken in a Premier League match is 16.4 shots per game (from both teams) whereas the average number of goals scored per game is 2.87 only. It is also seen in the graph that 34.1 per cent of shots are on target. Of course, the ‘number of shots on target’ does not translate into the ‘number of goals scored’, because it could be so that the shot was saved by the goalkeeper, deflected by a defender and so on.

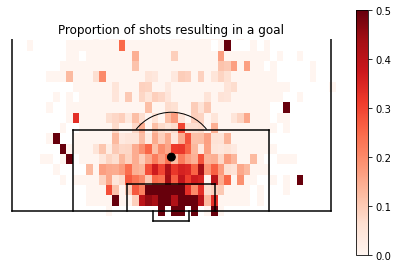


**Figure 2: Distribution of shots**

For further understanding and exploration of the depth of our dataset, I created a visualisation of the shots taken in a football game using Python programming language, Matplotlib and the FCPython's soccer package. This visualisation provides a graphical representation of the distribution of shots in a football game by highlighting the areas of the field from where the shots (in the dataset) were taken. The visualisation serves as a valuable tool for analysing the shot distribution and identifying key areas on the field from where the shots are taken frequently.

We can infer from this that most of the shots taken by football players are through the centre of the pitch, especially in the penalty box. Players tend to take a shot when they are in a ‘goal-scoring position’, which is an area of the field from where they are most likely to convert the shot into a goal. As obvious, the closer the player is to goal and the smaller the angle of the ball from the goal, the more chance the shot has of being a goal. This can be seen in the graph whereby the volume of shots taken inside the penalty box is significantly higher compared to the rest of the field. Long-range shots from outside the box are also fairly common, especially in teams which cannot penetrate the defending team’s defence and resort to taking a shot from long range to try out their luck.

The visualisation in this study provides further insights into the shot distribution patterns in football. Our analysis indicates that the volume of shots taken extremely close to the goal is relatively low in comparison to other areas within the penalty box. This apparent paradox can be explained by the fact that it is rare for an attacking player to reach such a position and take a shot, due to the presence of opposing defenders who obstruct their path. Additionally, this area is typically occupied by the goalkeeper, who can effectively collect the ball before a shot is taken, thus reducing the probability of a successful shot. These observations further emphasise the importance of understanding the complexities of shot distribution in football and the challenges faced by attacking players in creating goal-scoring opportunities.



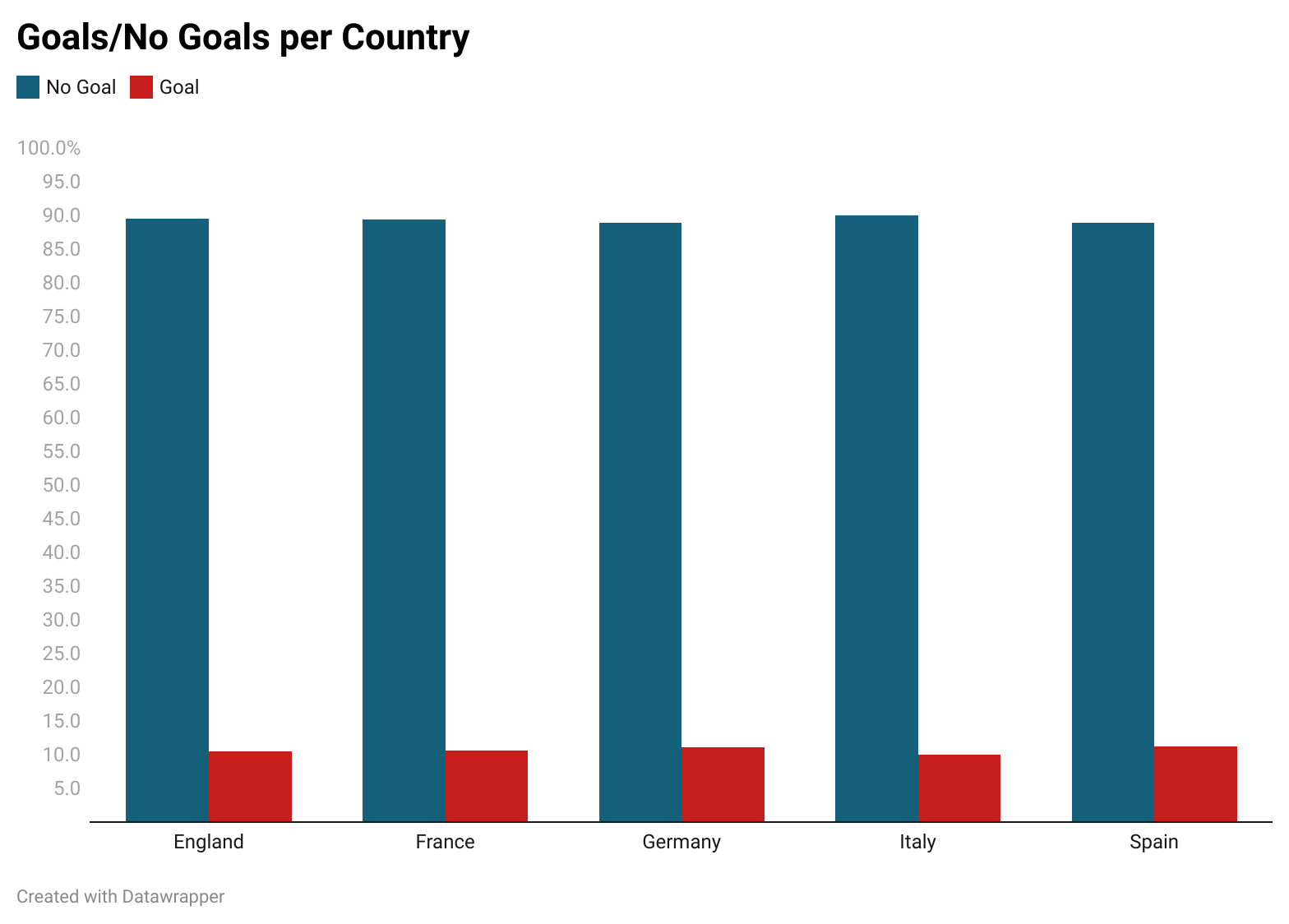
**Figure 3: Proportion of shots resulting in a goal**

The visual representation of Figure 3 presents an analysis of the areas of the pitch from which shots were taken and ultimately resulted in a goal. The shade of the colour used in the visualisation represents the frequency at which shots resulting in goals were taken. A substantial proportion of goals were scored within the penalty box, with the concentration of such goals decreasing as we move further away from the goal. Consequently, it is inferred that there is a proportional decrease in the probability of scoring as the distance from the goal increases.

Furthermore, it is apparent that the majority of goals were scored centrally, where the angle from the goal is smaller in comparison to areas where the angle from the goal is greater. This serves as evidence that the probability of converting a shot into a goal is inversely proportional to the angle from the goal.

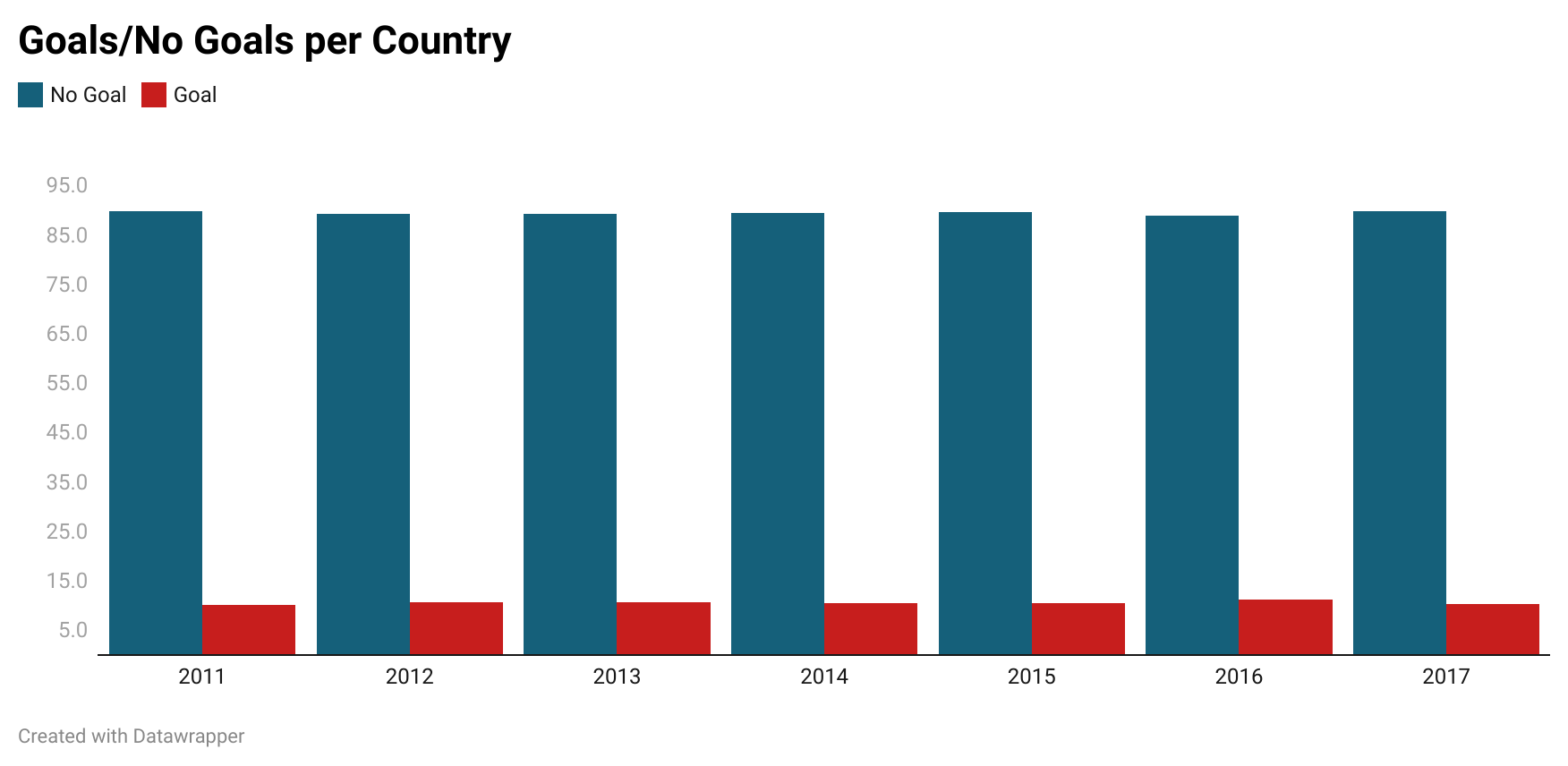
However, it must be noted that there exist outliers in the visual representation. Such instances occur when a player attempts a shot on a goal from an area with a low probability of scoring and still manages to score a goal. The reason for this is that there is a relatively low volume of shots taken from such areas, hence if a few of these shots (observations) are scored, the data will indicate a higher probability. This phenomenon highlights why such events are considered outliers, as the likelihood of such an occurrence in a football match is low and thus not a reliable way of scoring goals. Attacking teams are aware of this and usually aim to bring the ball as close to the goal as possible, rather than attempting a shot from a far distance with a low probability of scoring.

Now, we will delve into the examination of the percentage of shots that led to a goal. This analysis holds significant importance for our Expected Goals model. This analysis will be conducted in two forms: one across all five domestic leagues, and the other across years within our designated time frame. Through this analysis, we will be able to identify patterns and trends in shots resulting in goals across the world and through time.



**Figure 4: Goals/No Goals based on Countries**

In Figure 4, we can see that the chance of a goal being scored from a potential shot is only 10 per cent on average. And this is being reflected across all the top 5 leagues in Europe. This is empirically proven by (Pollard and Reep, 1997; Tenga et al., 2010; Lucey et al., 2014), which claim that 1 per cent of all attacking plays lead to a goal and 10 per cent of all shots lead to a goal Thus we can conclude that on average, from a potential 10 shots, 9 are missed and 1 of them results in a goal. Thus, the probability of a goal being scored across the top domestic leagues is 0.1.



**Figure 5: Goals/No Goals based on Season (Year)**

In Figure 5, we analyse the shots across five seasons of football to find a pattern in goal scoring across time. The results of this analysis are very similar to that for the domestic leagues, and we can conclude that even across time, for 10 potential shots taken 9 are missed and one leads to a goal. Thus, the probability of a goal being scored across the five seasons of football in Europe is 0.1.

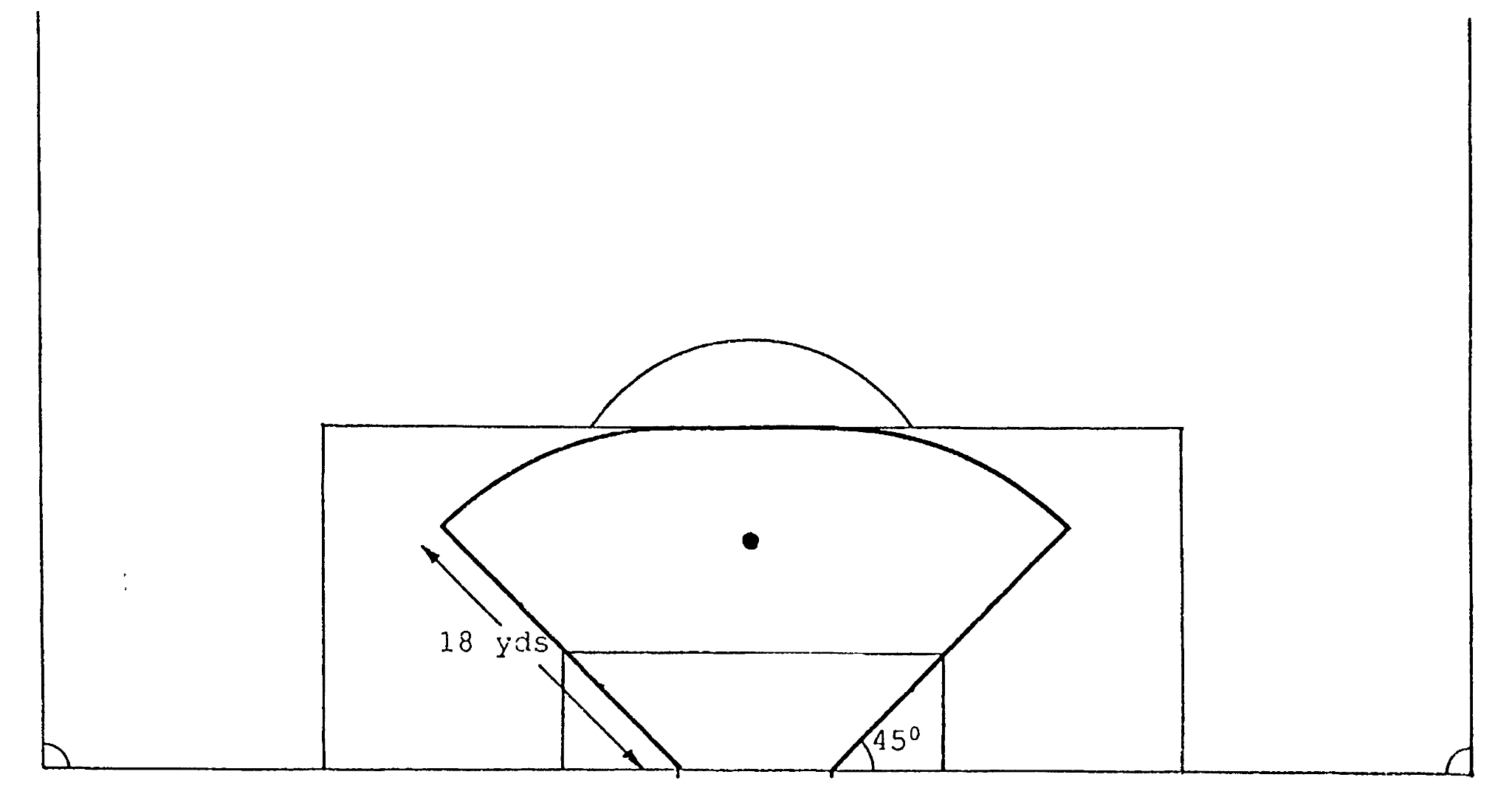
### 4.1.3 Data Specifications:

Now that we have explored the available data, we will prepare the dataset for building our model. We are working with event data and will be using it for our model. For doing so, we will start by creating X and Y sets. Y will be a set of every shot (observation) that has been included in our database and will also tell the result of that shot (0 for no goal, 1 for goal). Y acts as our dependent variable. X will be a set including all the meaningful and relevant information of shots that are available in the data and will act as our dependent variables.

#### Independent variables:

1. **Location**:

It is common knowledge that, if a person was given a football and asked to shoot it into a goal, they would try to go as close to the goal as possible before taking the shot. As demonstrated in the visualisation (Figure 6), this is because the closer the distance is from the goal the better the chances that the shot will be on target, and thus increases the chances of it being a goal. This idea of proximity to goal is talked about in-depth in (Pollard, 2004), whereby the author brings to light the idea of proximity. Pollard suggests that, if we know the exact location on the pitch from where the player shoots the ball, we can calculate the probability of it leading to a goal. For this, he used two main independent variables, namely: the distance between the point where the shot was initiated and the goal, measured using Euclidean metric, along with the angle at which the shot was taken relative to the goal. The same has been visually depicted in Fig 6, which has an arc inside the penalty box from which most goals are scored. The probability of scoring from inside the arc was 0.189 compared to 0.014 from outside it (Pollard and Reep, 1997). Figures 6 and 7 are obtained from the mentioned papers by Pollard (1997, 2004)

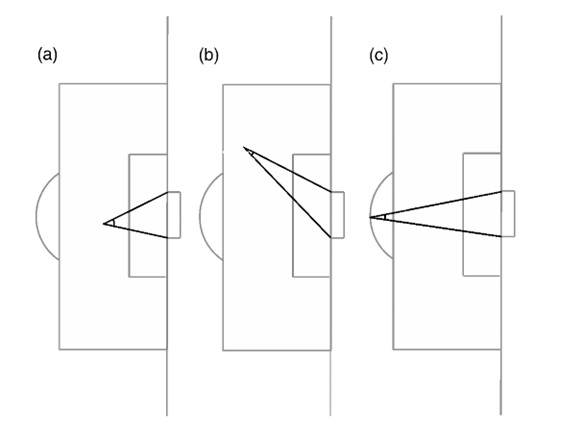


**Figure 6: Angle of shot from goal**

Along with distance, the angle from the goal also plays a major role in determining the outcome of the shot. The lesser the θ (angle from the two goal posts to the point from which the shot has been taken), the greater chance of scoring a goal.

As seen in FIGURE 7, Point (a) is taken inside the penalty box, very close to the goal and has a θ of 33 degrees. Point B has a θ of 17 degrees but is shot from inside the penalty box whereas Point (c) also has a θ of 17 degrees but is shot from outside the penalty box.

We can thus conclude that as distance, θ decreases. The best way to score a goal is to be as close to the goal in distance and as central to the goal in terms of shot angle.



**Figure 7: Angle and distance of shot from goal**

To simplify it, *the more of the goal you can see when you shoot, the better your chance of scoring.* (Sumpter, 2017)

The event dataset that I am using for this model did not have accurate availability of (x,y) coordinates of where the shot has been taken on the football pitch. To solve this, I used the conceptualisation done by Alex Rathke in (Rathke, 2017), whereby he split the pitch into eight zones and trained a model for them, indirectly taking shot location and angle into consideration.

For my own model, I divided the football pitch into 15 locational zones, based on the ideas of proximity as we just discussed and indirectly taking into account shot location and angle. This has been further elaborated on in Chapter 5.2.

1. **Body part:**

In football, the attacking team is allowed to use any part of their bodies other than their hands for scoring a goal. Players usually take shots with their feet and at times with their heads as well. So feet and head are the two most common body parts for scoring a goal. Each football player has a preferred dominant foot from which they take most of their shots. However, at times the player does not have the time or space to move the ball on their dominant foot and take a shot, and thus resorts to taking a shot from their weaker foot instead. Based on historical data, the probability of a shot taken from the weaker foot being converted into a goal is lower than a shot taken from the preferred dominant foot.

Headers are an interesting case. A header is usually taken when the ball is at a height that is high enough that the attacking player can’t use his legs to shoot. Headers are commonly seen during corners or freekicks, but also during open play. According to [G Anzer, 2021], the probability of scoring from a header is very less compared to using a foot. But Anzer also points out that in a situation where the attacking player needs to use his head to score if he used his foot instead the probability of scoring would be lower than the probability of scoring with the head. To explain this using an example, if Neymar passed the ball to Messi in the air and the ball rose at chest height when it came near Messi, Messi has a higher chance of scoring with his head than with his foot. Another pointer about headed shots is that they largely also depend on how far the header was taken from the goal and also the method of assist (how the assist was made, how fast the ball arrived and at what height). This is because an attacking player by themselves cannot generate a massive force on the ball with their head while shooting towards the goal, the assist is equally important.

Thus, the body part used to take a shot is an important independent variable while building our xG Model

1. **Situation of Play:**

The surrounding situation of play or the context as to how and when the attacking player took the shot is also an important feature when it comes to analysing our shot. The contextual factors surrounding a shot can vary significantly depending on preceding actions. A header resulting from a corner, for instance, will typically pose greater difficulty due to the increased concentration of players in the area than a header from a cross in open play. The probability of a shot leading to a goal differs drastically when we take context into account. According to [Innerdrive](https://blog.innerdrive.co.uk/sports/the-psychology-of-perfect-penalties#:~:text=Research%20shows%20that%20if%20a,the%20outcome%20of%20the%20match.), there is an 86 per cent chance that a penalty kick is converted into a goal. This means that for every 10 penalty kicks, at least 8 of them are scored. We can thus conclude that if an attacking team gets a penalty kick, there is a high probability of that shot leading to a goal. Thus the xG of a shot taken from a penalty kick is relatively higher.

Now if we were to consider the same for a shot taken from open play or a shot taken from a corner or a free-kick, the probability of the shot being converted into a goal would vary. Earlier In the data exploration section we concluded that approximately 10 per cent of the shots taken from open play result in a goal. Whereas for a corner kick, that number drops to 4 per cent and the same drops to 2 per cent for a free-kick. Thus, the situation of play is an important variable while building our xG model.

1. **Assist Method:**

As mentioned in the Body-Part section, the assist method is an important factor when considering the quality of chance. An assist in football is the pass made right before the shot is taken. So for example, if Player A passes the ball to Player B and player B scores a goal, Player A is said to have assisted that goal. The method of assist can either be a normal pass, a through ball, a cross, or a headed pass. At times it could be so that Player B himself created the chance and no one actually assisted him, thus it is possible to have no assist for a particular goal.

Each method of assist has varying degrees of accuracy and this largely depends on the quality of the player who is assisting the ball. Generally, a through ball or a cross gives better chances of scoring as compared to a headed pass as they get the attacking player in positions from where scoring a goal is highly likely. The method of assisting thus, is an important factor while considering our model.

1. **Counter-Attack**

A counterattack is when Team A tries to score a goal but fails to do so and Team B (which was defending when Team A was attacking) swiftly tries to attack Team A by taking advantage of the fact that Team A’s defensive players are not in structure as they were just trying to attack. Counter-attacking is a popular tactic for scoring goals as it finds defending players in compromising positions as compared to other times during the game and attacking players can easily exploit this.

As a counterattack gives an advantage to the attacking team, it also increases the chance of a shot taken leading to a goal as the defending players are not 100 per cent capable of defending the goal as they would have been otherwise. Hence, the factor of a counterattack is also a factor we will be considering in our xG model. In the model, if a shot taken was a result of a counterattack it will be denoted by a 1 and if it was not it will be denoted by a 0.

## 4.2 Model Description

### 4.2.1 Model Preparation

For the purpose of our model, we will need to incorporate a binary classifier. Given the choice of a logit model, tree-boosting system, and artificial neural networks, the decision has been made to employ a Logistic Regression model for its simplicity, widespread usage, and compatibility with linear dependencies.

The variables we discussed in the Data Description section are our independent variables and what we will be using to build our xG model. The variables are all categorical in nature and hence will be converted into dummy variables to make them binary in nature (this is for all variables other than counterattack, which is already binary in nature). We have a total of 28 binary variables that describe each shot taken in our dataset. Table 3 summarises all the variables in our dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| x | Variable | Variable Type | Description |
| 1 | location\_centre\_box | Location | Centre of the penalty box |
| 2 | location\_diff\_angle\_lr | Location | Difficult angle and long range |
| 3 | difficult\_angle\_left | Location | Difficult angle on the left of field |
| 4 | difficult\_angle\_right | Location | Difficult angle on the right of field |
| 5 | left\_side\_penalty\_box | Location | Left side of the penalty box |
| 6 | left\_side\_6yard\_box | Location | Left side of the six-yard box |
| 7 | right\_side\_penalty\_box | Location | Right side of the penalty box |
| 8 | right\_side\_6yard\_box | Location | Right side of the six-yard box |
| 9 | close\_range\_shot | Location | Very close range to goal |
| 10 | Penalty\_spot | Location | Penalty spot |
| 11 | outside\_penalty\_box | Location | Outside the penalty box |
| 12 | long\_range\_shot | Location | Long range |
| 13 | more\_35yards | Location | More than 35 yards |
| 14 | more\_40yards | Location | More than 40 yards |
| 15 | not\_recorded | Location | Not recorded |
| 16 | left\_foot | Body Part | Left foot |
| 17 | right\_foot | Body Part | Right foot |
| 18 | header | Body Part | Header |
| 19 | no\_assist | Assist Method | No Assist |
| 20 | assist\_pass | Assist Method | Pass |
| 21 | assist\_cross | Assist Method | Cross |
| 22 | assist\_header | Assist Method | Headed Pass |
| 23 | assist\_through\_ball | Assist Method | Through Ball |
| 24 | set\_piece | Situation of Play | Set Piece |
| 25 | corner | Situation of Play | Corner |
| 26 | free\_kick | Situation of Play | Free Kick |
| 27 | open\_play | Situation of Play | Open Play |
| 28 | fast\_break | Counter Attack | Counter Attack |

**Table 3: List of Variables**

In our dataset, we have a total of 2,29,135 shots taken, of which 24,441 resulted in a goal. As seen in Table 3, each of these 2,29,135 shots is described by 28 distinct characteristics. All of these characteristics are binary in their classification nature; hence they only indicate a YES or a NO (1 or 0) to any of the characteristics.

### 4.2.2 Model Design

It is typical in a Machine Learning Model to split the data into two sets: the training set and the testing set. The objective of the learning model is to achieve good performance on the testing set, which includes data that the model has not encountered before during the training phase. This helps evaluate the model's generalisation ability and ensure that it is not overfitting the training data. Non-sample data and if a model is trained on all the data there is a high chance of overfitting, which is not ideal (Chollet, 2017).

I divided the database of X and Y sets into training data and testing data. I used 65 per cent of my database for the purpose of training our xG model and the remaining 35 per cent of it for testing it. I chose to use 65 per cent of the set for training data because we have enough data observations in this 65 per cent to train a good xG model. Moreover, it is preferred to have more shots for pure testing of the xG model. Hence, I will go ahead with a 65/35 distribution for my training and testing data respectively.

# 5. Model Results and Evaluation:

After running the logistic regression model on the testing data, we come to the following concluding results:

## 5.1 Classification Report

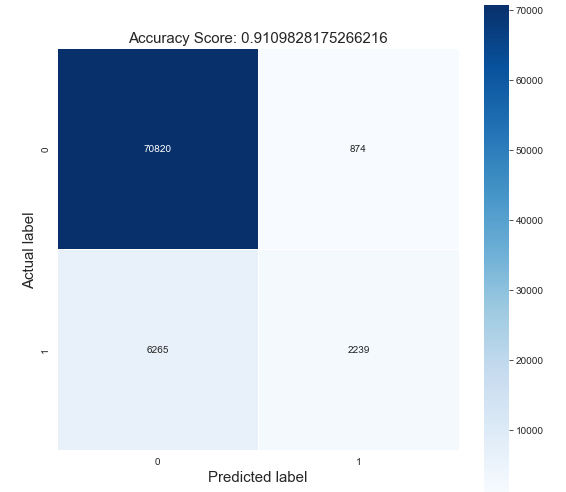
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1 Score | Support |
| 0 | 0.92 | 0.99 | 0.95 | 71694 |
| 1 | 0.72 | 0.26 | 0.39 | 8504 |
|  |  |  |  |  |
| Accuracy |  |  | 0.91 | 80198 |
| Macro Avg | 0.82 | 0.63 | 0.67 | 80198 |
| Weighted Avg | 0.90 | 0.91 | 0.89 | 80198 |

**Table 4: Classification report**

Based on the classification report, our xG model has demonstrated an accuracy of 91 per cent in predicting the likelihood of a player's shot resulting in a goal or not. The evaluation suggests that the model performs exceptionally well in predicting class 0 (No-Goal), with a precision of 92 and a recall of 99, resulting in an F1-score of 0.95.

However, the model's performance in predicting class 1 (Goals) is not as impressive. The precision and recall scores for class 1 are recorded as 72 per cent and 26 per cent, respectively, which translates to an F1 score of 0.39. The relatively poor performance of the model in classifying shots as goals can be attributed to the inherent challenge of accurately predicting which shots will result in goals. As previously discussed, only approximately 1 out of 10 shots lead to a goal. So, although the F1 score for class 1 is only 0.39, that is bound to be the case because of the nature of the algorithm we have created.

## 5.2 Confusion Matrix



**Figure 8: Confusion Matrix**

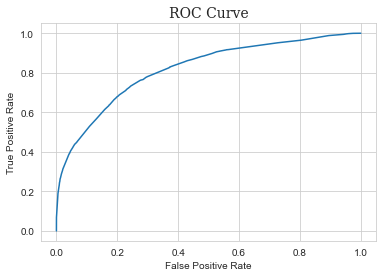
The confusion matrix provides a visual representation of the model's performance in correctly classifying the target variable. The matrix displays the number of true positive, true negative, false positive, and false negative predictions made by the model. By analysing the values within the matrix, we can determine the accuracy, precision, recall, and other performance metrics of the logit model.

In this case, a high number of true positive and true negative predictions would indicate a well-performing model, while a high number of false positive and false negative predictions would suggest the need for further optimization or improvement.

For our model, the confusion matrix reveals that the model accurately classified a significant proportion of the shots as no-goals. We see that 70,820 shots that were not goals were correctly classified by the model as no-goals true negative predictions. However, there were 6,265 instances where the model incorrectly classified shots as no-goals, when it was in fact, successful goal.

Furthermore, the model accurately identified 874 shots as goals, but it also missed 2,239 successful shots that were wrongly classified as non-goals. The high number of missed shots could suggest that the model needs further improvement in predicting successful shots or that there are underlying factors that need to be considered when constructing the model.

## 5.3 ROC Curve

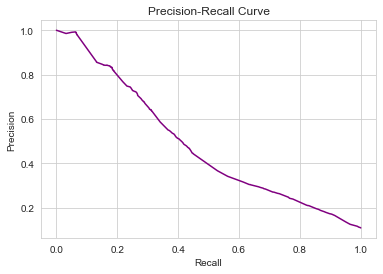


**Figure 9: ROC Curve**

Our model gives an excellent ROC-AUC score of 82 per cent which illustrates the diagnostic ability of our model at various threshold settings.

ROC AUC, although a widely popular metric for evaluating a logit model, is not without its issues. For imbalanced classification with a severe skew and few examples of the minority class, the ROC AUC can be misleading. In our situation, there are far more shots that result in goals than shots that do not, which causes an imbalance in our data. As a result, if we were to correctly anticipate that a shot wouldn't always result in a goal, we would already have an accuracy of 82 per cent. This highlights the limitations of ROC AUC and underscores the need for additional metrics that can better handle imbalanced data.

## 5.4 PR-AUC Curve



**Figure 10: Precision Recall-Curve**

The PR-AUC curve, a commonly used performance metric in machine learning, is slated to be used in the evaluation of our model. The PR curve is an effective diagnostic for imbalanced binary classification models because of its emphasis on the minority class. It plots the precision-recall trade-off for various classification thresholds and offers a more nuanced view of model performance compared to metrics such as accuracy which are less informative in the context of imbalanced data.

As for our specific xG model, the baseline precision for PR-AUC is 0.11 per cent, a value obtained by randomly guessing the outcome of each shot. This serves as a benchmark for assessing the model's predictive power. Our model, however, achieves an impressive PR-AUC score of 47.37 per cent. This performance is considered exceptional as it demonstrates the model's ability to effectively distinguish between the positive and negative classes.

# 6. Model Interpretation:

Although it is a necessity that our model produces accurate predictions which are reliable, it is also important to effectively communicate these predicted results so that they are easily interpretable. Our model will be later used by football clubs, coaches, and football analysts as a vital source of information for player recruitment, improving individual player performance and having an overall data-centric grip over the whereabouts of what happens beyond just the football match.

Complicated classifier-based Machine Learning models sometimes act as a black box and their predictions thus are more difficult to interpret. This, therefore, results in an inverse relationship between accuracy and interpretability, as per [Lundberg and Lee, 2012].

### 6.1 Model Coefficient

In a logistic regression model, the coefficients serve as estimators for the regression parameters and provide valuable information regarding the relative importance of each independent variable in predicting the outcome variable.

The sign of the coefficient indicates the direction of the relationship between the predictor variable and the outcome variable. A positive coefficient suggests that an increase in the predictor variable is associated with an increase in the likelihood of the event occurring, while a negative coefficient suggests that an increase in the predictor variable is associated with a decrease in the likelihood of the event occurring. The magnitude of the coefficient indicates the strength of the relationship between the predictor variable and the outcome variable, with larger coefficients indicating a stronger relationship. Table 5 displays the independent variables in our model and the corresponding coefficients of each variable.

|  |  |
| --- | --- |
| Variable Names | Corresponding Coefficients |
| fast\_break | 1.7161 |
| location\_centre\_penalty\_box | 0.324 |
| location\_diff\_angle\_lr | -1.614 |
| difficult\_angle\_left | -0.698 |
| difficult\_angle\_right' | -0.847 |
| left\_side\_penalty\_box | -1.023 |
| left\_side\_6yard\_box | 0.663 |
| right\_side\_penalty\_box | -0.943 |
| right\_side\_6yard\_box | 0.654 |
| close\_range\_shot | 2.1 |
| penalty | 2.341 |
| outside\_box | -1.72 |
| long\_range | -2.124 |
| more\_35yards | -2.406 |
| more\_40yards | -1.926 |
| not\_recorded | 7.195 |
| right\_foot | 0.223 |
| left\_foot | 0.161 |
| headers | -0.408 |
| no\_assist | -0.091 |
| assist\_pass | -0.063 |
| assist\_cross | -0.436 |
| assist\_header | -0.32 |
| assist\_through\_ball | 0.887 |
| open\_play | 0.115 |
| set\_piece | 0.87 |
| corners | 0.586 |
| free\_kick | 0.595 |

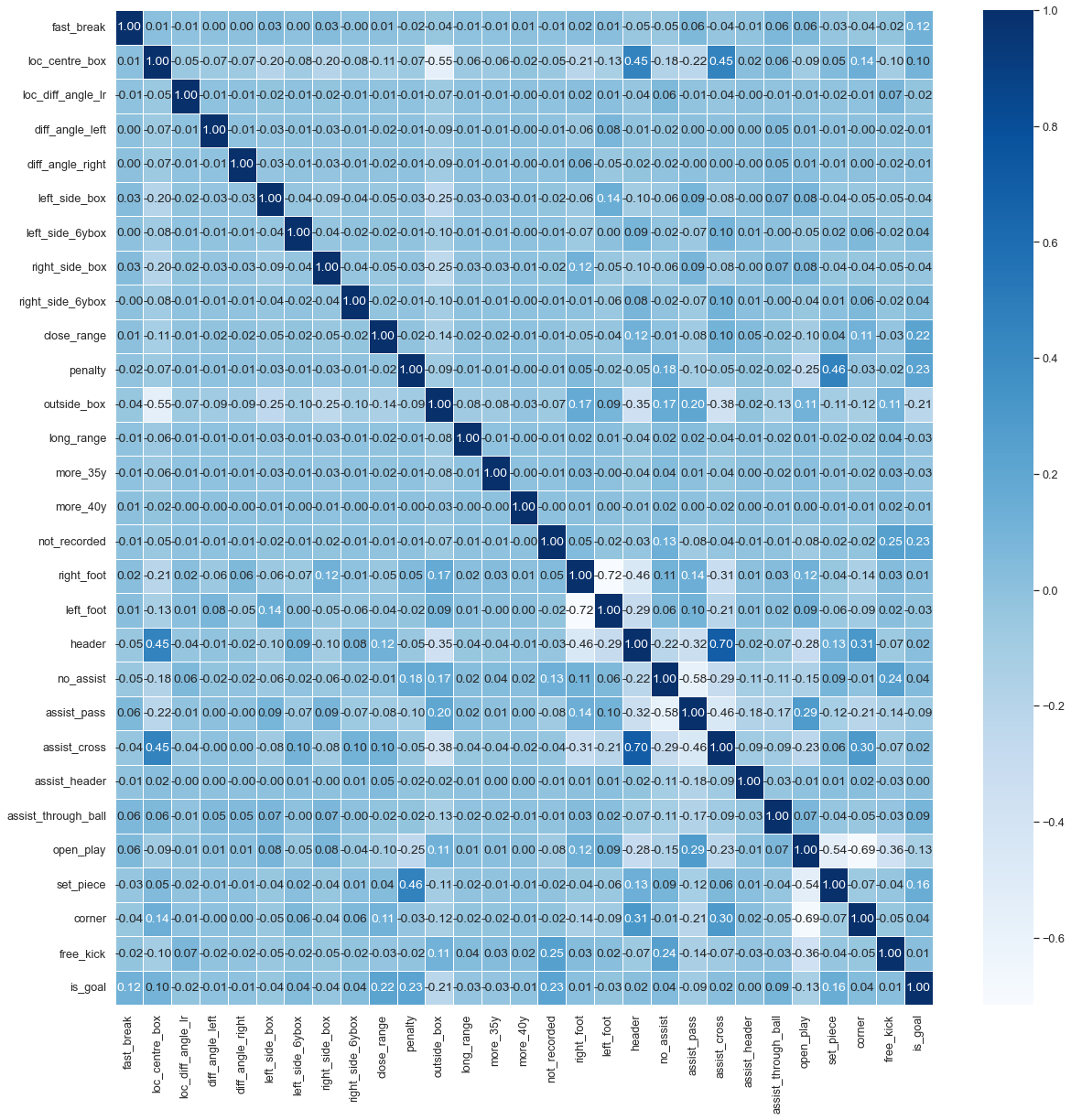
**Table 5: Model Coefficients**

Positive coefficients indicate that an increase in the corresponding independent variable is associated with an increase in the log-odds of the dependent variable occurring. For example, "fastbreak" has a coefficient of 1.7161, which suggests that the occurrence of a "fast break" event is positively associated with the outcome. Shots taken from "penalty" spots and "close-range" have large positive coefficients, suggesting that they are highly associated with the outcome or likelihood of the event occurring. This is in line with our intuition that shots taken from closer range or from a penalty spot are more likely to result in a goal.

Negative coefficients indicate that an increase in the corresponding independent variable is associated with a decrease in the log-odds of the dependent variable occurring. For example, "outside box" has a coefficient of -1.72, which suggests that shots taken from outside the box are negatively associated with the outcome.

Shots taken from longer distances such as "more\_35y" and "more\_40y" have large negative coefficients, suggesting that they are negatively associated with the outcome or likelihood of the event occurring. This also makes sense, as shots taken from longer distances are generally more difficult to convert into goals.

### 6.2 Correlation Matrix



**Figure 11: Correlation Matrix**

To further strengthen our interpretation of the model and to find relationships within our independent variables, we have plotted out a Correlation Matrix which does the job of showing the correlation values, which measure the degree of the linear relationship between each possible pair of coefficients.

One of the major interpretations of the above Correlation Matrix is the relationship between Assist Method - Cross, a Header and Loc\_Centre\_Box. In simple terms, it means that an attacking Player A will cross the ball inside the penalty area and Player B, who is located centrally in the penalty area close to the goal, will head the ball inside the goal. It could also be so that the assist has been crossed from a corner.

| Location on the Pitch | Assist Method | Body Part | Situation of Play |
| --- | --- | --- | --- |
| Centre Box | Cross | Header | Corner |

**Table 6: Example of Correlation between variables**

We can see in the matrix that according to our model if a header (Header) is taken in an area centrally close to the goal (Loc\_Centre\_Box), it has a correlation of 0.45. Similarly, when an assisted cross is headed to the goal, it has a correlation of 0.70 between them. It could also be so that the assist has been crossed from a corner and in this case a correlation of 0.31 is obtained.

# 7. Experiments and Discussion

Now that we have our Expected Goals model built, tested and validated it is time to put it to use and analyse the shots model in depth. Over the course of this chapter, I will be discussing the essence and fundamentals of this model by analysing players in our dataset against it.

We will discuss the players with the highest xG, the players with the best ratio and also discuss the top young players in the world at the time, who showed promising potential and talent. This analysis will be done in the subchapter of Shots Analysis

Followed by this we will discuss how we can use our model for efficient player recruitment and further use two real-life player case studies to emphasise and showcase how our model actually accurately works.

One advantage we have in this study is that we are using football data from the last decade (2011-2017), and since our study is based on efficient player recruiting, we can validate if the players whom we claim to be ‘promising talents with great potential’ (in that time period) actually lived up to the expectations. We can do this by comparing our data to where their careers are today, in 2023.

## 7.1 Descriptive Shot Analysis:

| x | Player | Difference | True Goals | Expected Goals (xG) |
| --- | --- | --- | --- | --- |
| 1 | Lionel Messi | -58.48 | 205 | 146.52 |
| 2 | Zlatan Ibrahimovic | -32.99 | 153 | 120.01 |
| 3 | Gonzalo Higuain | -32.38 | 118 | 85.62 |
| 4 | Luis Suarez | -31.70 | 96 | 64.32 |
| 5 | Cristiano Ronaldo | -30.62 | 198 | 167.38 |
| 6 | Robert Lewandowski | -25.89 | 124 | 98.11 |
| 7 | Alexandre Lacazette | -25.88 | 88 | 62.12 |
| 8 | Alexis Sanchez | -22.88 | 80 | 57.12 |
| 9 | Diego Costa | -21.16 | 93 | 71.84 |
| 10 | Karim Benzema | -20.81 | 85 | 64.19 |
| 11 | Antoine Griezmann | -20.75 | 80 | 59.25 |
| 12 | Eden Hazard | -19.01 | 62 | 42.99 |

**Table 7: Best Finishers: based on the difference of TrueGoals and xG**

Table 5 shows the top 12 players with the highest difference in xG and actual goals. This means that these are players that have consistently outperformed their xG in the time period of the data we have available. Thus we can say that these players consistently score more than they are supposed to based on the chances they get. One could question if this is purely based on luck or actual footballing brilliance, but given that we have this table based on five seasons' worth of data, we can safely say that this isn't down to just luck.

We can see that Lionel Messi is ranked first in this table with a difference of -58.49 goals. This means that Messi has scored around 58 more goals than an average person with the same quality of goal-scoring chances would have scored.

Cristiano Ronaldo is found 5th on this table. That is slightly lower for a player of his calibre, but we must note that although Ronaldo’s difference between xG and Actual goals isn’t as high as Messi’s, Ronaldo has a higher xG (167.38) than Messi’s xG of 146.52. From this, we can infer that although Messi is better at converting and consistently outperforming xG, Ronaldo is better at creating goal-scoring opportunities for himself or taking higher-quality of shots on goal, which is reflected by his high xG.

One important thing to note here is the statistics of Antoinne Griezmann, whom I have written about in a detailed player recruitment case study in Chapter 7.2.2.

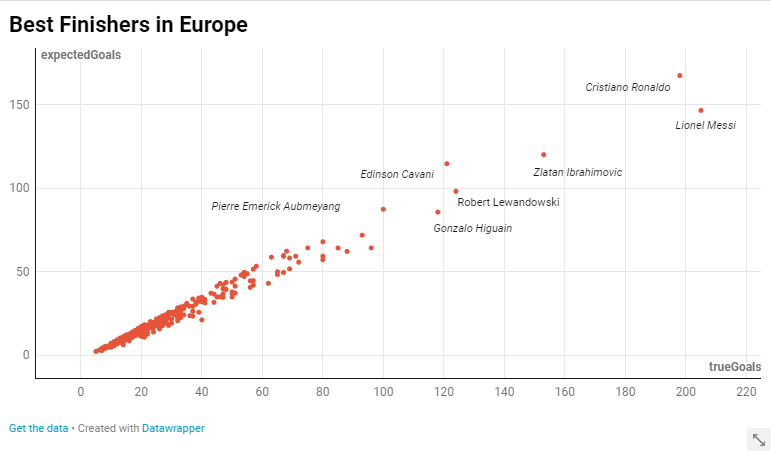
| Rank | Name | trueGoals | expectedGoals | Ratio |
| --- | --- | --- | --- | --- |
| 1 | Franck Ribery | 40 | 21.12 | 1.893939 |
| 2 | Mario Gotze | 37 | 23.39 | 1.581873 |
| 3 | H.M Son | 32 | 20.63 | 1.551139 |
| 4 | Bas Dost | 36 | 23.62 | 1.524132 |
| 5 | Carlos Tevez | 39 | 25.70 | 1.5117510 |
| 6 | Luis Suarez | 96 | 64.24 | 1.494396 |
| 7 | Alberto Bueno | 32 | 21.57 | 1.483542 |
| 8 | Isco | 33 | 22.44 | 1.470588 |
| 9 | Eden Hazard | 62 | 42.99 | 1.442196 |
| 10 | Alberto Bueno |  |  |  |

**Table 8: Best Finishers: Based on Ratio**

Table 6 provides an alternative lens to look at our dataset by examining the top 10 players based on the best trueGoals to xG ratio, (for players having a minimum of 30 goals). We can see that Frank Ribery is at the top of the table with a ratio of 1.89. From this we can infer that for every shot that Ribery was expected to score, he scored 1.89 goals. This shows that he was a player who was objectively very good at creating goal-scoring opportunities for himself and then actually scoring them as well.

From this table, a similar thing can be said about HM Son, who at that time was a young talent and had just started showing signs of his potential. Between 2011 and 2015, Son played in the German league as a winger and scored 32 goals from an xG of 20.63. What is impressive about his statistics is that his goal-to-xG ratio was 1.55 at the time. This stat, for a player as young as him at the time, was a remarkable underlying statistic and he was eventually sold to a big English club, Tottenham Hotspurs for $22 million, becoming the most expensive Asian player in history.

Figure 13. is a graphical representation of all the players playing in the top 5 leagues of Europe between 2011-12 to 2016-17, highlighting the trueGoals they have scored and the expectedGoals the quality of chances created.



**Figure 12: Scatter Plot of Best Finishers in Europe**

From the graph we can clearly see that Cristiano Ronaldo and Leonel Messi are very high up in the graph, with a considerably big gap between them and third-placed Zlatan Ibrahimovic. This is a testament to the fact that they both are not just in theory but also statistically two of the greatest football players.

## 7.2 Player Recruitment Analysis:

The main objective of this project is to use the xG model that we have created and tested to evaluate players’ performances and develop a framework for clubs to scout and recruit players more effectively.

In today’s landscape, when football is very competitive and big clubs with rich owners spend hundreds of millions to buy players and improve their squads, it can be unsustainable for other clubs to actually compete. Moreover, because of the media-hype players get instantaneously based on singular performances, it is common for these players to be tagged with over-inflated price tags. Thus because of this, football clubs fail to compete in the player market to buy these top talents and thus end up failing to compete in the tournaments as well. This idea is deeply elaborated in “*Playbooks and Checkbooks: An Introduction to the Economics of Modern Sports*” by (Szymanski, 2009).

The xG metric, along with many other advanced football analytics tools gives a platform for clubs to find talent not based on hype and superficial attention, but rather spot players based on underlying numbers which show that they can be as good a player they wish to buy at a much cheaper rate. By comparing a player's actual goals scored to their expected goals, one can get a better sense of whether a player is over-performing or underperforming relative to the quality of shots they are taking.

For example, let us say we have two players (Player A and Player B). We assume that both players are similar in their playing style and physical attributes and have taken the same number of shots in a season. Player A has scored 5 goals and Player B has scored 15 goals. From this, we can infer that Player B seems like a much better player than Player A and thus will be in more demand, which will make him more expensive. But then, if we take expected goals into account, we find that Player A has scored 5 goals from an xG of 12 and Player B has scored 15 goals from an xG of 10.

From this, we can say that for the quality of chances Player A had and the type of shots he took, he should have scored more goals than he actually did and hence was unlucky. He is thus “underperforming his potential”. ‘With the right type of training and development, this player in the years to come could fulfil this potential of scoring the goals he was expected to score. Thus, for a layman, it looks like Player B is a much better player as he has more goals and thus his market valuation will be higher, but we can buy Player A, who would be a young, talented player with a very high ceiling of potential and improvement. And most importantly, he will not come at an overly exaggerated price tag.

This is not to say that Player B is not a good player. What we can infer from this is that Player B has scored more goals than he was expected to score. He has thus overperformed. This is common for players who have had a breakthrough season and reached their true potential as they begin scoring more goals than they should be.

xG can also be used to evaluate a player's ability to create high-quality chances for themselves or their teammates. A player with a high xG but a low actual goal total may be considered a good playmaker who is creating high-quality chances for his teammates, even if he is not scoring many goals himself. Additionally, xG can be used to evaluate a player's shooting ability, as a player with a high xG may be considered a good finisher, even if he is not scoring many goals.

We must note that, when clubs use data analytics to recruit players, they are very specific about the tactical system in which they play their football. What this means is that, if a club wants to buy a Striker, they won’t just take any striker, however good he may be. They will recruit players who have a similar style of play, which aligns with the tactical system that they have created as the blueprint of the club. Players that blend well in the system becomes an integral part of the club and those who don’t are often sold to other clubs are high prices which bring the club large returns on investment (because they know that the player they are signing has high potential).

Table 7 provides further elaboration on how xG helps in better and efficient player recruitment by displaying the top 12 players under the age of 25 (at the end of the 2016-17 season) who have scored a minimum of 30 goals. By doing this we can filter out the young promising talents from our large database of player information.

| x | Name | Age | Position | trueGoals | xG | difference | ratio |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Alex Lacazette | 25 | Centre Forward | 88 | 62.12 | -25.88 | 1.42 |
| 2 | Antoine Griezmann | 25 | Centre Forward | 82 | 61.25 | -20.75 | 1.34 |
| 3 | Mauro Icardi | 24 | Centre Forward | 72 | 55.64 | -16.36 | 1.30 |
| 4 | Neymar | 25 | Right Winger | 68 | 53.21 | -4.79 | 1.09 |
| 4 | Harry Kane | 25 | Centre Forward | 65 | 49.97 | -15.03 | 1.30 |
| 5 | Eden Hazard | 22 | Right Winger | 62 | 42.99 | -19.01 | 1.44 |
| 9 | Roberto Firmino | 25 | Centre Forward | 59 | 43.77 | -6.23 | 1.14 |
| 7 | Romelu Lukaku | 24 | Centre Forward | 53 | 47.92 | -5.08 | 1.10 |
| 8 | Mattia Destro | 22 | Centre Forward | 51 | 45.55 | -5.45 | 1.12 |
| 10 | Ciro Immobile | 25 | Centre Forward | 48 | 43.52 | -4.48 | 1.10 |
| 11 | Stevan Jovetic | 25 | Centre Forward | 41 | 33.42 | -7.58 | 1.23 |
| 12 | Domenico Berardi | 22 | Centre Forward | 40 | 32.09 | -7.91 | 1.25 |

**Table 9: Top young players based on the xG model**

To explore the use case of Expected Goals on Player Recruitment a bit more deeply, we have created two case studies on real life examples of players who have been efficiently recruited by clubs with the use of advanced analytical metrics like Expected Goals. The trueGoals data used in the case studies is obtained from [Understat](http://understat.com) and the expectedGoals data is obtained from our own xG Model. The two players we will be discussing are Roberto Firmino and Antoinne Griezmann along with a club named Brentford FC.

### 7.2.1 Case Study: Roberto Firmino

Roberto Firmino plays for Liverpool Football Club in the English Premier League. He is one of the key players in the team and has helped change the fortunes of the club by winning multiple trophies during their time at Liverpool, most notably the English Premier League and the European Champions League.

Roberto Firmino was bought by Hoffenheim from his boyhood club Figueirense in 2011 for €3million at the age of 20. Lutz Pfannenstiel, Hoffenheim’s Club Director, revealed that he found out about Firmino while playing a football simulator game called Football Manager, which relies on advanced football analytics for player statistics. Thus, based on pure statistical and analytical understanding, Hoffenheim paid €3million for an unknown player.

The transfer proved to be a masterstroke as the Brazilian had a fabulous four-year spell with the German club. He made 153 appearances for the club piling up 49 goals and 36 assists. He also won the Bundesliga Breakthrough Player of the season award at the end of the 2013-14 campaign. For Table 8, data for “Matches Played” and “TrueGoals” have been taken from [Understat](https://understat.com/). Expected Goals metric is calculated from our model.

| Season | Matches Played | TrueGoals | ExpectedGoals |
| --- | --- | --- | --- |
| 2010-11 | 11 | 3 | 16.92 |
| 2011-12 | 30 | 7 |
| 2012-13 | 33 | 5 |
| 2013-14 | 33 | 16 | 17.51 |
| 2014-15 | 33 | 7 | 8.34 |

**Table 10: Roberto Firmino at Hoffenheim**

During his first three seasons at Hoffenheim, Firmino scored 15 goals in total. According to our model, however, Firmino had an xG of 24 in the first three seasons combined. What this means is that Firmino was underperforming to his true potential as a striker in the first 3 seasons. This isn't necessarily a bad thing just because he was “underperforming”. What it means is that from the chances of scoring goals that Firmino got and for the shots that he took over the course of these 3 seasons, he scored fewer goals than he should have. Since Firmino was just 22 at that time (which is a young age for a professional footballer), this meant that Firmino had the potential of contributing much more than he had done by then if he improved his game in the few areas, he was lacking in.

Firmino did exactly that, and over the next 2 seasons scored 10 and 11 goals respectively with an xG of XYZ and XYZ in each season respectively. From the numbers of the last 2 seasons, we can see that Firmino started getting much better and reaching the goals he was ‘Expected’ to score based on the goal-scoring chances he created and shots he took. Although much better from the first three seasons, the numbers were still a little underperforming for his xG.

In 2015, Hoffenheim sold Roberto Firmino to Liverpool FC for a grand total of €30 million. To put it in perspective, that is 10 times the cost at which they initially bought him. The reason why Liverpool was willing to pay such a large sum of money was not just because Firmino was a young talented striker who could score goals but also because he had not yet reached his true potential as a striker and the time he would reach it; he would be much more valuable than the €30 million spent on him.

Firmino went on to score 107 goals for Liverpool, being one of their main players and winning many of the major honours a professional football player can win. Although Firmino hasn’t done exceedingly well to “overpower” his xG or even score as many goals as a traditional striker score, his contribution to the team as not just a scorer of goals but a creator of goals has been equally impressive.

|  |  |  |  |
| --- | --- | --- | --- |
| Season | Matches Played | TrueGoals | Expected Goals |
| 2015-16 | 31 | 10 | 9.53 |
| 2016-17 | 35 | 11 | 9.70 |

**Table 11: Roberto Firmino at Liverpool**

For Table 9, data for “Matches Played” and “TrueGoals'' have been taken from [Understat](https://understat.com/). Expected Goals metric is calculated from our model

### 7.2.2 Case Study: Antoine Griezmann

Antoine Griezmann is a French professional footballer who began his professional career with the Spanish club Real Sociedad in 2009. As a youth player, Antoine Griezmann played for Real Sociedad Deportiva and was Real Sociedad in 2009, for a reported transfer fee of €4 million. The transfer was considered a bargain at the time as Griezmann was a relatively unknown young player and the fee was considered low for a player of his potential. During his time at Real Sociedad, Griezmann established himself as one of the top young players in the Spanish League (La Liga) and was touted to have a high ceiling of development.

Real Sociedad was playing in the Second Division of Spanish Football when Griezmann broke through into their first team in 2009. He spearheaded their attack with 7 goals that season and won Real Sociedad a well-deserved promotion to the First Division in Spain (La Liga). Between 2010 and 2014, Griezmann scored 22 goals in 94 appearances for Real Sociedad, while his expected goals (xG) total was 19.3. He was thus “underpowering his xG” and this suggests that although he was consistently creating high-quality chances and converting them at a high rate, he had the potential of being much better with the right development. In this period, Griezmann was relatively unknown, but his xG numbers were a good indication of his potential and he was able to make a big impact on the team.

In 2014, Griezmann moved to Atlético Madrid, for a reported transfer fee of €30 million. This was considered a high transfer fee at the time, but Griezmann had established himself as a top player in La Liga and was widely considered to be one of the best young talents in Europe. Atlético Madrid saw him as a valuable addition to their team, and the transfer fee reflected this. Playing as a striker, Griezmann scored 60 goals in La Liga for Atlético Madrid between 2014-2017 with an xG of just 42.27. This means that for the goal-scoring chances that Griezmann got and the quality of shots he took, he scored more goals than he was expected to score on a regular basis. This statistic clearly shows how Griezmann consistently overpowered his xG during these seasons.

| Season(s) | Club | Matches Played | True Goals | Expected Goals |
| --- | --- | --- | --- | --- |
| 2010-2014 | Real Sociedad | 94 | 22 | 19.3 |
| 2014/15 | Atlético Madrid | 37 | 22 | 14.71 |
| 2015/16 | Atlético Madrid | 38 | 22 | 16.04 |
| 2016/17 | Atlético Madrid | 36 | 16 | 11.20 |

**Table 12: Antoinne Griezmann at Real Sociedad and Atlético Madrid**

One important thing to note is that Griezmann played in a conservative Atlético Madrid side that tactically was not very attack-minded. This meant that Griezmann was not permitted to play free-flowing and attacking football as much as a striker of his calibre might want to play to reach his potential. Despite this, Griezmann rose to be one of the leading goal scorers in La Liga in our given time period, winning the Best Forward award and the Most Valuable Player award in 2016.

After 5 years at Atlético Madrid, he was sold to FC Barcelona in a transfer worth €120 million, becoming the fifth-most expensive player of all time. For perspective, he was sold at 4x the price he was bought at, with Atlético Madrid earning a direct profit of €90 million from this sale.

Thus, from this example, we can see how Atlético Madrid took the gamble of paying €30 million for an 18-year-old player who had a high ceiling of development based on his Expected Goals metric, but it actually paid off as he scored over 90 goals for them and was later sold at a profit of €90 million. Overall, Antoine Griezmann's career has been a success story, his early years at Real Sociedad, his Expected Goals numbers were a good indication of his potential, and he was able to become one of the best players in Europe, and a key player in the French national team that won the World Cup in 2018.

### 7.2.3 Case Study: Brentford FC

Brentford FC's approach to data-driven player recruitment has been one of the key factors in the club's recent success. In 2012, the club was bought by Matthew Benham, a statistician and former professional gambler, who was one of the first in football to adopt a statistical-based approach to player recruitment and squad management in his club. He brought in Rasmus Ankersen, a Danish football analytics expert, to help implement this strategy at the club. Ankersen was of the belief that expected goals analysis can help clubs identify players who are more likely to score goals and contribute to the team's success. He argues that this approach can help clubs avoid wasting resources on players who may not perform as well as expected, and instead focus on recruiting players who are more likely to deliver results (Ankersen, 2012).

In the ten-year timespan from 2012 (the year he bought the club) to 2022, Benham along with Ankersen took Brentford from League Two (4th tier of English Football) to The Premier League (first tier) for the first time in 74 years.

This data-driven approach has helped Brentford find and sign talented players who were overlooked by other clubs. For example, in the 2019-2020 season, Brentford had players such as Ollie Watkins and Said Benrahma, who were relatively unknown but had excellent underlying statistics. These players went on to have a major impact on Brentford and were eventually sold to Premier League clubs for large fees.

This approach has also helped the club achieve success on the pitch with a relatively low wage bill. In the 2019-2020 season, Brentford had the third-lowest wage bill in the English Championship (2nd tier) and finished third in the league, missing out on promotion to the Premier League by just one point. This success was built on the back of an efficient attack that had an xG per game of 1.58 (fourth-best in the league) and a strong defence that had an xGA per game of 1.10 (best in the league).

In terms of player recruitment, Brentford has a well-established scouting network that covers both domestic and international leagues. The club's scouts provide detailed reports on players and their performance data, which is then analysed by the club's data analytics team to identify potential targets. Brentford also has a unique player development model that focuses on improving players' individual skills and tactical understanding, which has helped the club produce several talented players who have gone on to play for other teams.

Some exponents of this data-driven recruitment process include players like Said Benrahma, Ollie Watkins and Neal Maupay, who were bought by Brentford for very low costs (relatively, were developed by Brentford to reach their potential as top talents and later either incorporated in the first-team or sold to another club for a much higher price (since the player was then more developed at his skill and has experience playing at a higher-level). In simpler terms, Brentford buys undervalued players and sells overvalued players.

In order to provide an understanding of the utilisation of Expected Goals (xG) within Brentford's recruitment process, it is imperative to examine the statistical data and figures that demonstrate the significance of the club's xG-based methodology. Over the years, the implementation of xG in Brentford's recruitment approach has had a notable impact, both on the field as well as financially. Table 11 provides a study of the top 10 players whom Brentford signed for low costs and eventually sold years later as fully developed players for higher transfer fees, thus generating a sustainable income from the efficient business. It is important to note that after the players were bought for Brentford, they played on average 3 seasons of football for them before being sold off. That means that Brentford’s model is not merely based on generating financial profits as they recognized and utilised the player’s potential by playing them during their time at the club, and then sold them when their market valuation had risen.

| Player | Price he was bought (£) | Price he was sold (£) | Net Difference (£) |
| --- | --- | --- | --- |
| Ollie Watkins | 6.3 million | 30.6 million | 24.3 million |
| Said Benrahma | 1.5 million | 20.8 million | 19.3 million |
| Neil Maupay | 1.8 million | 20 million | 18.2 million |
| Ivan Toney | 5 million | still plays for the club | still plays for the club |
| Chris Mepham | Free | 12.2 million | 12.2 million |
| Andre Grey | 0.6 million | 11.2 million | 10.6 million |
| Ezri Konsa | 2.6 million | 12 million | 9.4 million |
| Scott Hogan | 0.9 million | 9.5 million | 8.6 million |
| Ryan Woods | 1.2 million | 6.5 million | 5.3 million |
| Nico Tenaris | 0.1 million | 5 million | 4.9 million |

**Table 13: Brentford Player Sales (Last 6 Years)**

These statistics show that Brentford's expected goals-based approach has been successful in helping the club identify and sign talented players who have gone on to perform well for the club. By using xG as a key metric in their recruitment process, Brentford has been able to achieve success with limited resources and build a competitive squad which works on a sustainable business model, driven entirely by data analytics.

# 8. Conclusion

In this research thesis I aimed to quantify the quality of shots taken during a football game and use them for analysing player performances and recruiting football players more efficiently. For this purpose, a binary-classifier-based Expected Goals (xG) model was created using Logistic Regression, coded entirely in Python. This model uses historical data related to shots across 5 seasons of football and uses its predictive algorithm to determine whether the shot will result in a goal or not a goal. The objective of this thesis is to use this model for effective and efficient player recruitment in professional football clubs which has been effectively done in the Player Recruitment Analysis section (7.2) where I present 3 case studies (Antoine Griezmann, Roberto Firmino and Brentford FC ) where I compare their Actual Statistics to the statistics predicted by our model and this showcase and validate how the xG model we have created is influential in player recruitment. The model has a predictive accuracy of 91 per cent, a precision of 91 per cent, a recall of 99 per cent and an F1 score of 0.95. These metrics show that our model is substantially effective. A detailed literature review is done at the beginning of the thesis which encompasses all the previously done work in the field of football analytics, Expected Goals modelling and player recruitment in football. Below, I elaborate on the limitations of the research thesis and the scope of further research.

## 8.1 Limitations:

Firstly, if we had event data for a few more seasons and across more competitions, the model would be richer in terms of its performance and predictive power. Along with this, apart from event data if we had tracking data at our disposal, it would give an added dimension to our model as we would be able to improve the location-based variables, helping in increasing the accuracy.

Another limitation of our model is that the location-based information available to us is zone-based (Rathke, 2017). It would have been ideal if instead of zonal locations we had information in x and y coordinates whereby we could measure ‘distance from goal’ and ‘angle from goal’ as separate entities. Availability of location in the form of coordinates would increase the of our data and this directly will influence the predictive accuracy of our xG model.

Additionally, it would be great if information existed about the defending team as well. Information about the number of defenders between the location of the shot and goal, their positioning, as well as the positioning of the goalkeeper, would immensely improve the dataset. Along with this, xG does not take into account the game state, i.e.. if the goal was scored when the scoring team was leading the game or not (according to the score).

## 8.2 Future Work

As mentioned above the limitations, both in terms of data as well as spatiotemporally can significantly make the research richer. Along with this, it would be interesting to see how an XG model is created using more advanced machine learning techniques like XGBoosting, Artificial Neural Networks and Random Forests. I shall try and expand on this thesis in the coming years to further enrich my research and increase its credibility.

# Bibliography

1. *Antoine Griezmann*. (2023, February 4). Wikipedia. https://en.wikipedia.org/wiki/Antoine\_Griezmann
2. Anzer, G., & Bauer, P. (2021). A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer). *Frontiers in Sports and Active Living*, *3*. https://doi.org/10.3389/fspor.2021.624475
3. Baron, J., & Hershey, J. (1988). Outcome Bias in Decision Evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569–579. https://www.sas.upenn.edu/~baron/papers/outcomebias.pdf
4. Bialkowski, A., Lucey, P., Carr, P., Matthews, I., Sridharan, S., & Fookes, C. (2016). Discovering Team Structures in Soccer from Spatiotemporal Data. *IEEE Transactions on Knowledge and Data Engineering*, *28*(10), 2596–2605. https://doi.org/10.1109/tkde.2016.2581158
5. Brechot, M., & Flepp, R. (2020). Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. *Journal of Sports Economics*, *21*(4), 335–362. https://doi.org/10.1177/1527002519897962
6. Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(5), 338–349. https://doi.org/10.1002/sam.11318
7. Brownlee, J. (2020, January 5). *ROC Curves and Precision-Recall Curves for Imbalanced Classification*. Machine Learning Mastery. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/
8. Burn-Murdoch, J. (2018, November). *How data analysis helps football clubs make better signings*. Financial Times; Financial Times. https://www.ft.com/content/84aa8b5e-c1a9-11e8-84cd-9e601db069b8
9. Cavus, M., & Biecek, P. (2022). Explainable expected goal models for performance analysis in football analytics. *ArXiv:2206.07212 [Cs]*. https://arxiv.org/abs/2206.07212
10. Chappas, C. (2013, August 6). *Goal Expectation and Efficiency* (C. Trainor, Ed.). StatsBomb | Data Champions. http://statsbomb.com/2013/08/08/goal-expectation-and-efficiency/
11. Chollet, F. (2017). *Deep Learning with Python*. Manning, Cop.
12. Clarys, J., Reilly, T., & Stibbe, A. (2003). *Science and Football II*. Taylor And Francis, An Imprint Of Taylor & Francis.
13. de Hoog, M. (2015, March 24). *How data, not people, call the shots in Denmark*. The Correspondent. https://thecorrespondent.com/2607/how-data-not-people-call-the-shots-in-denmark/230219386155-d2948861
14. Eggels, H. (2016). *Expected goals in soccer explaining match results using predictive analytics*. https://pure.tue.nl/ws/files/46945853/855660-1.pdf
15. Fairchild, A., Pelechrinis, K., & Kokkodis, M. (2018). Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality. *Journal of Sports Analytics*, *4*(3), 165–174. https://doi.org/10.3233/jsa-170207
16. Fernandez, J., Barcelona, F., & Bornn, L. (n.d.). *Wide Open Spaces: A statistical technique for measuring space creation in professional soccer*. http://www.lukebornn.com/papers/fernandez\_ssac\_2018.pdf
17. Fernández, J., Barcelona, F., Bornn, L., Los, D., & Dodgers, A. (2019). *Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer*. http://www.lukebornn.com/papers/fernandez\_ssac\_2019.pdf
18. Franks, I. M. (1988). Analysis of association football. *Soccer Journal*, *33*(5), 35–43.
19. Gauriot, R., & Page, L. (2019). Fooled by Performance Randomness: Overrewarding Luck. *The Review of Economics and Statistics*, *101*(4), 658–666. https://ideas.repec.org/a/tpr/restat/v101y2019i4p658-666.html
20. Gerrard, B. (2017, March 16). *The role of analytics in assessing playing talent* (J. Baker, S. Cobley, J. Schorer, & N. Wattie, Eds.). Eprints.whiterose.ac.uk; Routledge. https://eprints.whiterose.ac.uk/98808/
21. Goodman, M. (2018, December 28). *Goodman: How to evaluate the Premier League title race after half a season*. The Athletic. https://theathletic.com/736588/2018/12/27/goodman-how-to-evaluate-the-premier-league-title-race-after-half-a-season/
22. Green, S. (2023). *Assessing The Performance of Premier League Goalscorers*. Stats Perform. https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/
23. Hamilton, H. (2017). *Expected saves: An inverse of expected goal?*
24. Harm Eggels, Ruud van Elk, & Mykola Pechenizkiy. (2016). *Explaining Soccer Match Outcomes with Goal Scoring Opportunities Predictive Analytics*. MLSA@PKDD/ECML. https://www.semanticscholar.org/paper/Explaining-Soccer-Match-Outcomes-with-Goal-Scoring-Eggels-Elk/549b3dc2b9601dbde0807c35d629b6c3ddb40489
25. Herbinet, C. (2018). *Predicting Football Results Using Machine Learning Techniques*. https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf
26. Hughes, M. (1988). Computerized notation analysis in field games. *Ergonomics*, *31*(11), 1585–1592. https://doi.org/10.1080/00140138808966808
27. InnerDrive. (n.d.). *The Psychology of Perfect Penalties*. Blog.innerdrive.co.uk. https://blog.innerdrive.co.uk/sports/the-psychology-of-perfect-penalties#
28. Jeggo, L. (2019, May 28). *Brentford: Leading the way in recruitment strategy*. Football Bloody Hell. https://footballbh.net/2019/05/28/brentford-championship-tactical-analysis-statistics-analysis/
29. Kharrat, T., McHale, I., & Peña, J. L. (2017). Plus-minus player ratings for soccer. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2019.11.026
30. Kollár, A. (2021). Football analytics for better betting: Pitch partitioning, possession sequences, expected goal model and player evaluation on Dawson model. *SSRN Electronic Journal*. https://doi.org/10.14293/s2199-1006.1.sor-.pptaukv.v1
31. Lago Peñas, C. (2022, April 12). *How to increase the probability of scoring in penalties shoot-outs*. Barça Innovation Hub. https://barcainnovationhub.com/how-to-increase-the-probability-of-scoring-in-penalties-shoot-outs/
32. Lago-Peñas, C., Lago-Ballesteros, J., Dellal, A., & Gómez, M. (2010). Game-Related Statistics that Discriminated Winning, Drawing and Losing Teams from the Spanish Soccer League. *Journal of Sports Science & Medicine*, *9*(2), 288–293. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3761743/
33. Leenders, J. (2018, October 22). *xG: evaluating the quality of goal-scoring chances in football*. SciSports. https://www.scisports.com/expected-goals-evaluating-the-quality-of-goal-scoring-chances-in-football/
34. Lefgren, L., Platt, B., & Price, J. (2015). Sticking with What (Barely) Worked: A Test of Outcome Bias. *Management Science*, *61*(5), 1121–1136. https://doi.org/10.1287/mnsc.2014.1966
35. Lewis, M. (2004). *Moneyball : The Art of Winning an Unfair Game*. W.W. Norton.
36. Link, D., Lang, S., & Seidenschwarz, P. (2016). Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data. *PLOS ONE*, *11*(12), e0168768. https://doi.org/10.1371/journal.pone.0168768
37. Mewis, J. (2016, November 17). *Liverpool star was discovered by scout on Football Manager*. Mirror. https://www.mirror.co.uk/sport/football/news/roberto-firmino-discovered-hoffenheim-director-9280292
38. Morales, C. A. (2016). A mathematics-based new penalty area in football: tackling diving. *Journal of Sports Sciences*, *34*(24), 2233–2237. https://doi.org/10.1080/02640414.2016.1177657
39. Nadeau, L., Richard, J.-F., & Godbout, P. (2008). The validity and reliability of a performance assessment procedure in ice hockey. *Physical Education & Sport Pedagogy*, *13*(1), 65–83. https://doi.org/10.1080/17408980701444718
40. Noordman, R. (2019). Improving the estimation of outcome probabilities of football matches using in-game information [Master’s Thesis]. In *Amsterdam School of Economics, Faculty of Economics and Business*. https://www.scisports.com/wp-content/uploads/2019/10/Noordman-Rogier-12366315-MSc-ETRICS.pdf
41. Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank. *ACM Transactions on Intelligent Systems and Technology*, *10*(5), 1–27. https://doi.org/10.1145/3343172
42. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, *6*(1). https://doi.org/10.1038/s41597-019-0247-7
43. Pardo, P. M. (2020). *Creating a Model for Expected Goals in Football using Qualitative Player Information* [Masters Thesis]. https://upcommons.upc.edu/bitstream/handle/2117/328922/147841.pdf?sequence=1&isAllowed=y
44. Partida, A., Martinez, A., Durrer, C., Gutierrez, O., & Posta, F. (2021). Modeling of Football Match Outcomes with Expected Goals Statistic. *Journal of Student Research*, *10*(1). https://doi.org/10.47611/jsr.v10i1.1116
45. Pollard, R. (2004a). *(PDF) Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space*. ResearchGate. https://www.researchgate.net/publication/240641737\_Estimating\_the\_probability\_of\_a\_shot\_resulting\_in\_a\_goal\_The\_effects\_of\_distance\_angle\_and\_space
46. Pollard, R. (2004b). Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space. *International Journal of Soccer and Science*, *2*(1). https://www.researchgate.net/publication/240641737\_Estimating\_the\_probability\_of\_a\_shot\_resulting\_in\_a\_goal\_The\_effects\_of\_distance\_angle\_and\_space
47. Pollard, R., & Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *46*(4), 541–550. https://doi.org/10.1111/1467-9884.00108
48. Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017). Not All Passes Are Created Equal. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3097983.3098051
49. Rasmus Ankersen. (2012). *The Gold Mine Effect*. Icon Books Ltd.
50. Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, *12*(Proc2). https://doi.org/10.14198/jhse.2017.12.proc2.05
51. Raudonius, L., & Seidl, T. (n.d.). *Shot analysis in different levels of German football using Expected Goals*. Retrieved February 16, 2023, from https://dtai.cs.kuleuven.be/events/MLSA22/papers/MLSA22\_paper\_5416.pdf
52. Reep, C., & Benjamin, B. (1968). Skill and Chance in Association Football. *Journal of the Royal Statistical Society. Series a (General)*, *131*(4), 581–585. https://doi.org/10.2307/2343726
53. Reilly, T., Cabri, J., & Duarte Araújo. (2005). *Science and football V : the proceedings of the Fifth World Congress on Science and Football*. Routledge.
54. Reilly, T., Murphy, A., Lees, A., & Davids, K. (1988). *Science and football : proceedings of the first World Congress of Science and Football*. Routledge.
55. Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. https://doi.org/10.1007/s10462-009-9124-7
56. Sarkar, S., & Kamath, S. (2021). Does luck play a role in the determination of the rank positions in football leagues? A study of Europe’s “big five.” *Annals of Operations Research*, *Nov 2021*. https://doi.org/10.1007/s10479-021-04369-6
57. Secareanu, A. (n.d.). *Football Events*. Www.kaggle.com. Retrieved February 16, 2023, from https://www.kaggle.com/datasets/secareanualin/football-events
58. Spearman, W. (2018). Beyond Expected Goals. In *Research Gate*. MIT Sloan Sports Analytics Conference. https://www.researchgate.net/publication/327139841\_Beyond\_Expected\_Goals
59. StatsBomb. (2014, February 18). *Comparing actual and expected goals*. StatsBomb | Data Champions. https://statsbomb.com/articles/soccer/comparing-actual-and-expected-goals/
60. Steinbach, M., Kumar, V., & Tan, P.-N. (2006). *Introduction to Data Mining Instructor’s Solution Manual*. https://www-users.cse.umn.edu/~kumar001/dmbook/sol.pdf
61. Sumpter, D. (2023). *The Geometry of Shooting*. https://soccermatics.medium.com/the-geometry-of-shooting-ae7a67fdf760
62. Szymanski, S. (2009). *Playbooks and Checkbooks An Introduction to the Economics of Modern Sports*. Princeton University Press. https://muse.jhu.edu/book/36256
63. T Paukku. (1994). *And it’s another goal . . .: Deep in the heart of Finland, a former footballer turned professor is pioneering a computer program that he hopes will help make the game more popular. Timo Paukku went to meet him*. New Scientist. https://www.newscientist.com/article/mg14219294-000/
64. Tiippana, T. (2020). *How Accurately Does the Expected Goals Model Reflect Goalscoring and Success in Football?* https://aaltodoc.aalto.fi/bitstream/handle/123456789/99805/bachelor\_Tiippana\_Tuomas\_2020.pdf?sequence=1&isAllowed=y
65. Van den Hoek, N. (2019). Improving expected-goals models: Towards more accurate values for individual shots by considering more detailed information. *Master Thesis, Jheronimus Academy of Data Science*.
66. Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.11.001
67. Wheatcroft, E., & Sienkiewicz, E. (2021). *A Probabilistic Model for Predicting Shot Success in Football*. Www.semanticscholar.org. https://www.semanticscholar.org/paper/A-Probabilistic-Model-for-Predicting-Shot-Success-Wheatcroft-Sienkiewicz/6bb2dc204021d4906f306ef8124edf2d839f400a
68. Williams, A. (2020, April 7). *The roots of Expected Goals (xG) and its journey from “nerd nonsense” to the mainstream*. These Football Times. https://thesefootballtimes.co/2020/04/08/the-roots-of-expected-goals-xg-and-its-journey-from-nerd-nonsense-to-the-mainstream/
69. *xG Explained*. (n.d.). FBref.com. https://fbref.com/en/expected-goals-model-explained/
70. Zhao, Y., & Zhang, H. (2019). Analysis of goals in the English Premier League. *International Journal of Performance Analysis in Sport*, *Volume 19, 2019*(Issue 5), 820–831. https://doi.org/10.1080/24748668.2019.1661613

# Plagiarism Report (Turnitin)

