Expected Value Metrics in Women's Soccer

by

Darby Goodall

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Master of Science

Department of Statistics
University of Manitoba
Winnipeg

Copyright © 2025 by Darby Goodall

Abstract

With advancing technology, the collection and analysis of sports data has become more comprehensive and sophisticated. One of the latest developments in Sports Analytics has resulted in publicly available tracking data. In this research, we will first introduce a model based on dependent and sequential logistic regression that uses tracking data to estimate the probability of a shot, a shot on target and a goal occurring throughout a soccer match. Secondly, and most importantly, we will discuss how the above models can be used to construct expected value metrics such as the expected number of shots, shots on target and goals over a whole game. While previous research on expected value metrics focuses on men's soccer, we concentrate our analysis on women's soccer to close the research gap between the two genders. This research analyzes the 2022 Union of the European Football Association (UEFA) European Women's Championship. Finally, we will conclude by discussing our findings and the performance of our models.

Acknowledgment's

I want to thank Dr. Alexandre Leblanc for his guidance and feedback throughout this research. His passion for Statistics in sports is an inspiration, and his continuous support has made this work enjoyable. I would also like to thank the members of my evaluation committee for providing knowledgeable insights and guiding me along this journey.

I want to thank CANSSI, NSERC, the University of Manitoba Faculty of Graduate Studies, and Faculty of Science for providing me with financial support throughout this journey.

Finally, I want to thank my friends and family for believing in me even when I did not believe in myself. Their support and encouragement have meant the world to me, and I could not have done it without them.

Contents

Contents ii

List of Tables ix

List of Figures xii

1 Introduction 1

1.1 Literature Review 2

1.2 Objective 6

1.3 Outline 8

2 Data Exploration 11

2.1 Overview 11

2.2 Cleaning 16

2.3 Feature Engineering 22

3 Modelling Soccer Matches **47**

3.1 Dependent Logistic Regression 47

3.1.1 Model and Definitions 48

3.1.2 Fitted Values and Expected Value Metric 55

3.1.3 Practical Considerations 58

3.2 Sequential Logistic Regression 63

3.2.1 Models and Definitions 63

3.2.2 Fitted Values and Expected Value Metrics 69

3.2.3 Practical Considerations 79

4 Results **83**

4.1 Reduced Model 83

4.1.1 Model Selection 84

4.1.2 Shots Model 85

4.1.3 Shots On Target Model 90

4.1.4 Goals Model 95

4.2 Expected Value Metrics in Men's Soccer 102

5 Discussion **109**

5.1 Data Challenges 109

5.2 Applications 111

5.3 Future Work 114

5.4 Conclusion 118

Bibliography **121**

List of Tables

2.1 Description of the quantitative variables 23

2.2 Description of the indicator variables 24

2.3 Description of the indicator variables (continued from Table 2.2) 25

3.1 Estimated probabilities vs. observed frequency of shots using the full model 56

3.2 Estimated probabilities vs. observed frequency of shots on target using the full model 72

3.3 Estimated probabilities vs. observed frequency of shots on target given that a shot occurred using the full model 73

3.4 Estimated probabilities vs. observed frequency of goals using the full model 76

3.5 Overview of expected value metrics and observed values 78

3.6 Overview of expected value metrics and observed values (continued from Table 3.5) 79

3.7 Likelihood ratio test results for inclusion of the distance from the goal line to the goalkeeper 81

4.1 Parameter estimates when modelling the probability of a shot using the reduced model 85

4.2 Likelihood ratio test results for inclusion of the distance between the ball and the net 88

4.3 Estimated probabilities vs. observed frequency of shots using the reduced model 89

4.4 Parameter estimates when modelling the probability of a shot on target given that a shot occurred using the reduced model . 91

4.5 Estimated probabilities vs. observed frequency of shots on target using the reduced model 93

4.6 Parameter estimates when modelling the probability of a goal given that a shot on target occurred using the reduced model . 96

4.7 Estimated probabilities vs. observed frequency of goals using the reduced model 98

4.8 Estimated probabilities vs. observed frequency of goals using the full model on the men's tournaments 104

4.9 Example of expected value metrics and observed values in the Men's UEFA European Football Championships 105

4.10 Parameter estimates when modelling the probability of a goal given that a shot on target occurred for the Men's UEFA European Football Championship 106

List of Figures

2.1 Visualization of freeze frame data 15

2.2 Visualization of before and after the frame orientation was corrected 17

2.3 Representation of two consecutive frames with identical coordinates but different teams 21

2.4 Visualization of the cone created between the ball location and the net 29

2.5 Visualization of the angle and distance between the ball and the net 33

2.6 Visualization of the distance to the nearest defender outside the cone and the amount of pressure being applied from inside the cone 35

2.7 Depth of five query points relative to the opposite team using various measures of data depth 40

3.1 Expected Shot Value (ESV) vs. observed number of shots . . . 58

3.2 Tree diagram 64

3.3 Expected Shot on Target Value (ETV) vs. observed number of shots on target 71

3.4 Expected Goal Value (EGV) vs. observed number of goals . . 74

4.1 Expected Shot Value (ESV) produced by the reduced model vs. observed number of shots 90

4.2 Expected Shot on Target Value (ETV) produced by the reduced model vs. observed number of shots on target 94

4.3 Expected Goals Value (EGV) produced by the reduced model vs. observed number of shots on target 99

4.4 Expected value metrics produced by the full model vs. the reduced model 101

4.5 Expected Goals Value (EGV) vs. the observed number of goals produced by the full model for men's and women's data 103

Chapter 1

Introduction

In every corner of the globe, you can find a soccer fan base with enough passion to fill a stadium. For many fans, soccer is more than just a game. It provides a sense of belonging and connection. This passion is seen around the world and often passed down for generations. Due to its popularity, soccer has a sizeable economic impact, which opens doors for advances in data analytics that can change how this popular game is played. Over the years, data analysis has improved player and team performance, tactical decision-making, player recruitment and transfer decisions. With advancing technology, the collection and analysis of sports data has become more comprehensive and sophisticated. For instance, one of the latest developments in Sports Analytics has resulted in publicly available tracking data. This new type of data has been used in numerous studies to evaluate player and team performance. Unfortunately, the majority of this research focuses on analyzing men's sports. Because of this, we are interested in analyzing women's soccer data to help close the gender

gap in Sports Analytics and expand our understanding of how data analytics can impact the popular game of soccer.

This research uses event and tracking data and logistic regression to estimate the probability of a shot, a shot on target and a goal. With these estimated probabilities, we then develop expected value metrics associated with these events. As discussed below, our research will potentially impact decision-making in soccer and other sports.

1.1 Literature Review

Sports Analytics has existed for a long time, but has received added visibility since the advent of the book and movie Moneyball (Lewis, 2003). The most recent change in Sports Analytics has been the introduction of tracking data. Traditionally, player and team performance has been evaluated based on specific in-game events such as passes, shots, and goals. With the availability of tracking data, it is now possible to include player position and movement to the above when assessing performance.

An important part of the most recent literature focuses on expected goals metrics. For example, Lucey et al. (2015) used spatiotemporal data and an expected goal value to improve shot prediction. The authors discussed attributes that impact expected goals, including defender proximity, defence formation and attacking features such as crosses, long passes and player speed. They discovered that, in combination with shot location, features including speed of

play, interaction of surrounding players, defender proximity and game phase considerably impact the chances of a team scoring. Mead et al. (2023) developed expected goals models using player/team ability and incorporated psychological effects through goal differential. They found that when their developed player value (which represents the player's ability) and psychological effects increased, the probability of a goal occurring also increased. Rathke (2017) concluded that distance from the goal and shot angle significantly impact expected goals. To prove this, they divided the attacking half into eight zones and estimated the probability of scoring given that a shot occurred from each zone. Scholtes and Karakuş (2024) used hierarchical logistic regression models to determine if there is a player or position effect when estimating the probability of a goal given that a shot occurred. They found that attacking midfielders and forwards are expected to score more using a model that includes distance and angle to the goal. They also discovered that, as the number of predictors in the model increases, the positional effect reduces. Anzer and Bauer (2021) introduced an expected goals model that used shot and tracking data. They incorporated the goalkeeper position, pressure metrics, the event that initiated the sequence and defending player positions in their models. Their models outperformed those of Lucey et al. (2015) and Rathke (2017). They concluded that the model improvement is likely due to increased dataset size and the selected features (Anzer and Bauer, 2021). More recently, Bandara et al. (2024) implemented expected goals models that incorporate events

before the shot to improve the expected goals metric. For instance, they found that attacks initiated from the sides of the 18-yard box, shots taken in front of the goal within the 18-yard box, and shots created from passes to the far post all seem to play a role in predicting whether or not a shot is likely to lead to a goal.

Another topic in Sports Analytics that has undergone numerous advances in recent years is the concept of expected contribution. Metrics focusing on expected contribution typically allow researchers to measure an action's contribution to an observed outcome. For example, Bransen et al. (2019) introduced an expected contribution metric that evaluates a pass based on whether it increases or decreases the chances of scoring. The authors compared pass execution to historical data. Using this information, they assigned each pass a score based on how likely it was to lead to a goal. From this, the authors ranked players based on their expected contribution to determine who contributes the most to goal-scoring opportunities. Similarly, Rein et al. (2017) proposed a Voronoi-diagram approach to assess the effectiveness of a pass by evaluating the attacking space dominance and the number of defenders between the ball carrier and the goal. The authors showed that both above-mentioned factors significantly impact the number of goals scored. More generally, Delcroos et al. (2019) introduced a framework for evaluating player actions in terms of their impact on game outcomes. Their proposed framework combines the change in the probability of scoring and the change in the probability of conceding a goal into an action rating. One benefit of this approach is that it allows

us to evaluate various actions, including crosses, passes, dribbling and shots. The authors found that their proposed framework outperforms other player performance metrics, such as the number of goals and assists per game.

Finally, a vast majority of Sports Analytics research discusses men's sports. However, the growing popularity of women's sports and the increased interest in Sports Analytics have led to more freely available data for women's sports. This advancement has allowed researchers to compare results obtained from men's soccer to women's. For example, Bradley et al. (2014) studied the differences between men and women in the UEFA (Union of European Football Associations) Champions League. In particular, they found that men cover more distance at higher speeds than women. They also found that women have lower pass completion rates. More recently, Bransen and Davis (2021) analyzed women's tracking data. They compared differences in men's and women's soccer and found that women typically have a higher shot conversion rate, tend to head the ball closer to the goal and that draws comparatively are rare in women's soccer. They also compared expected goals models between men's and women's soccer and found that meaningful features for modelling shots in men's soccer differ from those in women's soccer. This conclusion is unsurprising as the playing styles vary significantly between men's and women's soccer. Finally, Heiner et al. (2014) implemented a Markov chain model for evaluating Division One women's soccer in the National Collegiate Athletics Association. They studied the impacts of various offensive skills including shots, passes and

controlling the ball, along with a specified state that indicates how the skill was performed. For example, a pass can be classified as moving toward a player's own goal, moving toward the goal being attacked, or under another one of the categories mentioned in the paper. The authors found that passing the ball into the 18-yard box and taking corner kicks increases the probability of scoring. They also note that their results cannot necessarily be generalized beyond the teams included in their research. This issue is common when analyzing women's sports due to the limited amount of publicly available data, and it demonstrates the importance of collecting data on and researching women's soccer separately from men's.

1.2 Objective

Research on expected goals and contributions in men's soccer has been extensive. To address the research gap between the two genders, we concentrate our analysis on women's soccer. In general, soccer is known to be a low-scoring game. Because of this, estimating the probability of shots, shots on target, and goals comes with challenges. It is common for a soccer game to end with no goals and very few shots. As a result, it seems desirable to develop different metrics for evaluating team performance rather than simply relying on the observed number of shots and goals.

To improve our understanding of women's soccer, we are interested in using dependent and sequential logistic regression to model the probability of a shot,

a shot on target, and a goal occurring. We also aim to develop expected value metrics using the models we implement. These expected value metrics are useful for evaluating team performance. For instance, the number of goals scored in a match is not always representative of the events during a game. A team could have the ball for the majority of the game but end up losing because they couldn't get the ball in the back of the net, and their opponent scored a lucky goal. In these situations, the observed number of goals does not accurately represent the dominating team's performance. To avoid this issue, expected value metrics are developed based on scoring opportunities, allowing them to capture events throughout the game better.

The potential applications of this research are extensive. It can help us determine factors that lead to a team's success and impact how coaches and players make decisions. For example, it can potentially influence how coaches and players prepare for specific opponents: by analyzing data and identifying team strategies, teams can prepare to handle their opponent's strengths and capitalize on their weaknesses. In addition, our research may also impact the video game and sports betting industries, as it could help produce more realistic game simulations and better predictions of game outcomes. For instance, Mead et al. (2023) showed that expected goals are a better predictor of a team's future performance than observed goals.

1.3 Outline

This thesis will proceed as follows. In Chapter 2, we discuss our data exploration. We focus on data-related issues and how they were addressed. We also discuss some of the key variables used throughout our analysis and give a brief overview of the data structure and contents.

In Chapter 3, we implement dependent logistic regression to estimate the probability of a shot occurring. Subsequently, we define the Expected Shot Value (ESV) metric using the fitted values obtained from this model. We then discuss sequential logistic regression to estimate the probability of a shot on target and a goal occurring. Furthermore, we use these models to obtain the Expected Shot on Target Value (ETV) and Goal Value (EGV) metrics for each team in each match. Following the above, we evaluate the performance of our developed models and metrics.

In Chapter 4, we develop reduced models for our sequential logistic regression that estimate our probabilities of interest using the most influential variables. Additionally, we compare model outcomes and the expected value metrics produced by the reduced models with those produced by the full models. Furthermore, we expand upon our current models by exploring the differences between men's and women's soccer. We develop expected value metrics for the men's 2020 and 2024 UEFA European Football Championships, and we determine which of the considered covariates are important for estimating the probability of a shot, a shot on target and a goal.

Finally, Chapter 5 discusses challenges faced throughout our research, applications of our metrics and future research plans.

Chapter 2

Data Exploration

In this chapter, we provide an overview of the data to be used in the later chapters, including a discussion of the tournament from which it was collected and important attributes that will be used throughout our analysis. Finally, we also outline the procedure used to clean the data and discuss how the provided variables were transformed into more useful ones for our analysis.

2.1 Overview

The data used for this research contains play-by-play information and player positions throughout the matches for the 2022 Union of European Football Associations (UEFA) Women's Championship Tournament. The play-by-play and tracking data used for this analysis were obtained from Hudl Statsbomb (2025a). Play-by-play data provides information on each event during the game with exact time stamps, while tracking data provides the position of the

players on the pitch at these times. Combining these two data types yields a more complete picture of the events throughout the game and allows for a more thorough analysis.

The UEFA tournament was held in ten different stadiums across England over three-and-a-half weeks in July 2022. The tournament was structured as follows. The first stage is the group stage, where the qualified teams are separated into groups. Each team plays against every other team within their group, with the top two teams of each group advancing to the knockout stage. During the knockout phase, the winning teams advance until they reach the final. Out of the 31 games in the UEFA tournament, we selected 21 for our analysis. The remaining ten games were excluded because of what we deemed to be major data-related issues. A data frame is associated with each match, containing between 2000 and 4000 rows representing events in the game. The event data contains information on offensive and defensive plays, including but not limited to passes, shots, carries, pressure, and interceptions. These data also provide information about each event, such as the time, outcome, players involved, location of the event, and position of players on the pitch at those times. To analyze this data, we combined the 21 data sets generated from each game into one data set.

An important feature obtained from these data is the timestamp, which indicates when each event was performed to the millisecond. This timestamp is divided into three columns representing minutes, seconds, and milliseconds. We used these columns to create a time variable in minutes through unit

conversion. The resulting variable is a continuous value representing the time throughout the game. There are a few important considerations involving our time variable that require a brief discussion. First, it is important to note that extra time impacts this time variable. At the end of each half, there is an opportunity for referees to add additional time to the period based on the length of stoppages in play. The extra time in our data ranges from zero to nine minutes per half, with an average of three minutes. The timestamp for additional time is formatted such that extra time from one period overlaps with the times at the start of the next period. For example, if there are three minutes of added time in the first half of the game, then the first half will end at 48 minutes, but since the second half starts at 45 minutes, there is an overlap in the times. To distinguish between these variables, we use the provided period variable, which indicates what period the event occurred in. We also implemented a frame index variable that guarantees that the frames are placed in the correct order, even though time may overlap as per the above description. Second, the number of frames per second is commonly used in Sports Analytics to describe data. The more frames available per second, the better the data frequency. Using the developed time variable, we determined that there are 0.046 minutes (or 2.75 seconds) between frames on average. It is essential to acknowledge that the clock continues to operate during stoppages in soccer games. As a result, we sometimes see longer times between frames. In our data, these breaks are up to 5.3 minutes long and most often occur when a player commits a foul and a free kick is earned or when a player is

injured. Although these long breaks initially looked like outliers, it is clear that they are a unique feature in soccer games. Finally, our data has the added restriction that games played in the knockout stage cannot end in a tie. In the event of a tie after 90 minutes of regular play and the associated extra time, the teams compete in two 15-minute overtime periods. Then, if there is still a tie after overtime, the teams will participate in a shootout.

Another piece of information obtained from these data is a series of freeze frames that display the location of each player within the visible area, defined as the area of the field visible on broadcasts. The freeze frame data includes x and y coordinates of each visible player (ie. players in the visible area), indicates which player is performing the action and whether the other players are teammates of the actor or not. Figure 2.1 displays an example of a freeze frame including the visible area, player, and ball locations. In our data, the x-coordinates range from zero to 120 yards and the y-coordinates range from zero to 80 yards, where the origin is located in the bottom left corner of Figure 2.1. Unfortunately, various challenges exist because we can only see a portion of the field and only have access to coordinate data for visible players. Additionally, the freeze frame data does not include player identification for all the visible players, instead, it only contains identification for the acting player. This limited our analysis to team strategy and made any individual player analysis impossible.

The raw data contains 14.2 player coordinates per frame on average. In some frames, we have as few as one visible player, and in others, we have

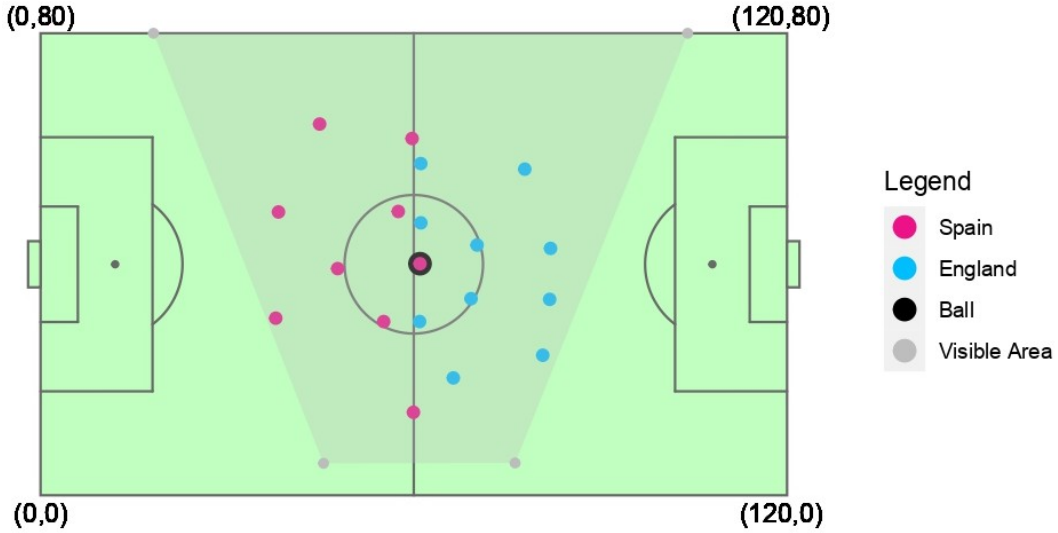


Figure 2.1: Information provided in the first frame of the Spain vs. England game. The coloured points represent the players on each team. The black point identifies the player in control of the ball and, the grey region represents the visible area. This region is created using the grey points, which we refer to as the limits of our bounding box. Note that a soccer team typically has 11 players, and this freeze frame only displays ten players from England and nine players from Spain. This occurs because we only have access to player coordinates when they are located inside the bounding box. In this particular frame, we are missing the goalkeeper coordinates for both teams and one player coordinate for Spain.

all 22 players. This analysis aims to study how various predictor variables impact a player's chance of shooting, hitting the target, and scoring a goal. While exploring numerous predictor variables, we concluded that many of these predictors require a minimum number of player coordinates per frame to be measured accurately. Because of this, we restricted our analysis to frames with at least three players per team. This adjustment removed roughly 2.6%

of the total frames available for each game. This value is small and, hopefully, won't have a major impact on our analysis.

Finally, the data set contains the locations on the pitch where each action took place. From this information, we obtained the approximate location of the ball for each frame. This is crucial as ball location is not directly available in the data. In cases where the action was ball-related, we assigned the coordinates of the ball to be the same as those of the action. On the other hand, when the action was defensive, we assigned the ball coordinates to the opposing team player nearest to the defensive action. For example, if a player was applying pressure, we assigned the ball to the player under pressure. Finally, if the action was not ball-related, we did not assign coordinates for the ball as we did not have ball location information at that time.

2.2 Cleaning

While processing the data to create animations of the games, problems arose that required further examination. The created animations revealed that game play was sometimes suddenly flipping to the opposite side of the field and back between a few frames. This was flagged as an issue because the sudden movement was not aligned with how soccer games typically progress. The left-hand side of Figure 2.2 displays an example of an incorrect series of frames. We can see that the middle frame is incorrectly oriented compared to the frames above and below. After further research, we found that Hudl StatsBomb had

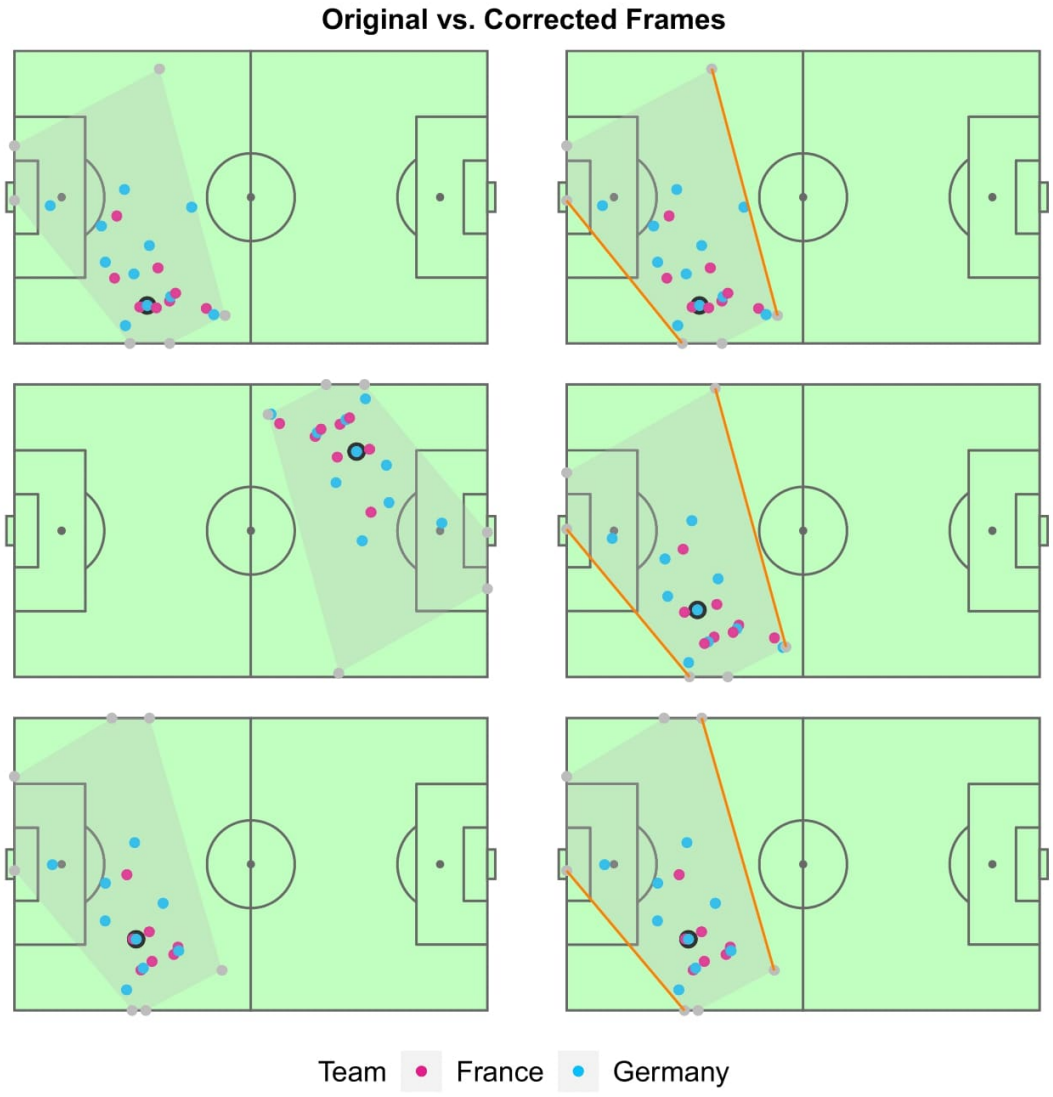


Figure 2.2: **(Left Column)** Three consecutive frames where the middle frame is incorrect. **(Right Column)** The same three frames after implementing our flip detection procedure, the middle frame is now in the correct direction. Additionally, the frames on the right contain orange lines that represent the cone used to detect the incorrectly flipped frames.

converted the data to ensure that both teams attacked the goal on the 120-yard line (ie. the right side of the pitch on these frames). Various online resources indicated that Hudl StatsBomb implemented these changes using the team identification variable, which indicates the team that performs the action at each time point along with some exceptions. Using this variable, we flagged all events conducted by the team that should be attacking the goal at the zero-yard line and corrected their orientation. After adjusting the orientation of these frames, the remaining incorrectly flipped frames were associated with the exceptions used to create the raw data. We found no available online resources that identify the exceptions utilized when flipping the coordinates. Consequently, we analyzed the frames and developed a list of actions associated with these exceptions. This list includes "fouls won with an advantage", "lost 50/50 battles", "lost aerials", and "incomplete dribbles". Once this list was created, we verified which frames contained actions on the exception list and corrected their orientation. Through this process, we rectified a majority of the incorrectly flipped frames. Unfortunately, with this process, we did not capture all the incorrectly flipped frames. This is likely caused by the fact that we could not find an official list of exceptions, and we had to create one through visual inspection.

To resolve this issue, we created a method that automatically identifies flips to address the remaining incorrect frames. We started by analyzing the average position of the players and significant changes in the visible area to detect any flips. For instance, if the average of the x-coordinates for all players

shifted by more than 60 yards within a single frame, we marked it as potentially incorrect. While this method successfully identified most incorrectly oriented frames, we wanted to implement a more dependable approach.

Therefore, we utilized the coordinates of the vertices within the visible area to identify four vertices and two sides of a cone, with the extrapolated origin of this cone indicating the camera's approximate location. When multiple intersections were possible, we compared the current frames with earlier frames to determine which cone origin from the previous frame was nearest to the current one. Upon locating this intersection, we evaluated whether the frame was flipped incorrectly; if so, we flagged it as an issue. For this new method to be effective, we operate under the assumption that the camera's location remains constant throughout the game. After implementing this revised approach, the number of identified errors increased significantly. This new procedure allowed us to detect and rectify the remaining incorrect frames. The right side of Figure 2.2 shows the same sequence of frames with the corrected frame in the middle. Additionally, the orange lines represent the cone used in the flip detection process. This figure demonstrates that the original data issues have been resolved appropriately.

Another issue with the data was that multiple events occurred simultaneously with identical player positions. After further investigation, we determined that these situations occurred when a player received, recovered or intercepted the ball and when an opponent applied pressure. For example, if a player received the ball at time t and made a one-touch pass, the raw data

would contain two frames at time t. One row provides information associated with receiving the ball, and the other contains information about the pass. In this research, we are more interested in the ball's location at each time point than in the event that occurred. Consequently, the duplicate rows did not provide any additional information, and therefore, we removed the duplicate entries. In particular, specific event types held more importance in our analysis; hence, we removed the events associated with types that were not of interest. This included, but was not limited to ball receipt, recovery and pressure. This adjustment does not create an issue for our analysis since we are primarily interested in the ball's location throughout the game.

The final major challenge we faced while cleaning the data was that some frames were exact duplicates of each other, with opposite team identification variables. We utilize the team identification variable when creating plots and determining which team has the ball; therefore, this inconsistency raises concerns. Figure 2.3 shows an instance of such data duplication where the x and y player coordinates are identical in consecutive frames, yet the players switch from one team to another. This is evident from the point colours and the unchanged legend. Given that the frames are identical, they do not provide any additional information. Consequently, we decided to remove the duplicate frames from the data set. Upon further investigation, we found that this issue arises when two frames have the same timestamp and one of the actions is incomplete. For instance, if a player is about to receive the ball but does not, or if a player attempts to win a 50/50 ball but fails. We proceeded with our

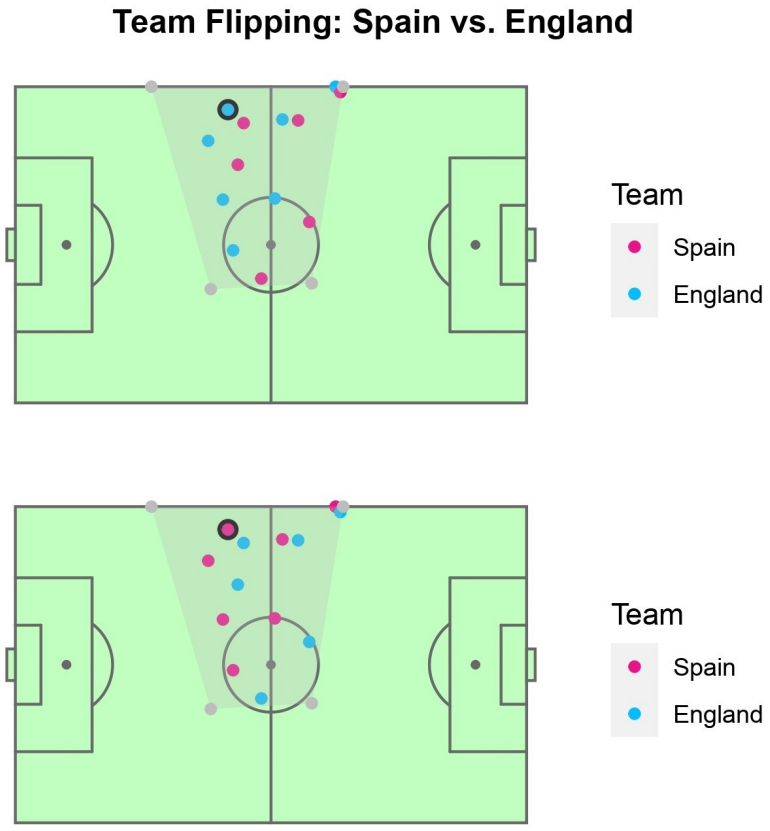


Figure 2.3: Two consecutive frames with identical coordinates but different teams. Note that the legend is identical between the two frames, indicating an issue.

analysis once the incomplete actions were removed from the data.

In summary, we have corrected flipped frames and resolved issues with duplicated frames and opposite team identification variables. As a result, we obtained clean data to be used for our analysis and to create smooth-flowing animations of the games with minimal errors. We will now proceed with a discussion of the variables used in our analysis and highlight their role in estimating the probability of a shot, a shot on target and a goal.

2.3 Feature Engineering

This section discusses how we transformed the variables provided in the raw data into more informative variables for our analysis. An overview of the quantitative variables discussed in this section can be found in Table 2.1. Additionally, details regarding indicator variables used in our analysis can be found in Tables 2.2 and 2.3.

To begin, the raw data contained a column of acting player positions with 25 categories, ranging from "Left Centre Forward" to "Right Back". These positions explain where players start the game, but due to the fluidity of soccer, player positions often shift throughout the game. Because of this, we did not want to measure the effect of each position individually; instead, we grouped the various positions into three common categories: defenders, midfielders, and forwards. We expect athletes who play midfield and forward positions to have a higher chance of shooting, shooting on target, and perhaps scoring, since

Table 2.1: Description of the quantitative variables.

Category	**Description**	**Variable Name(s)**
Distance	The distance between the ball carrier and the goal line of the net being attacked. Measured in yards.	distToAttackingNet
Depth	The integrated rank-weighted depth of the ball carriers' coordinates relative to the opposing team's coordinates, excluding the goalkeeper.	depthNotInPos.NG.IRW
Pressure	The distance between the ball carrier and the nearest defender outside the cone. Measured in yards. The distance between the ball carrier and the nearest defender inside the cone. Measured in yards.	closestOut closestIn
Interaction Terms	The interaction between the distance from the ball to the goal line and the number of defenders inside the cone. The pressure being applied to the ball carrier.	dInCone_dist pressure

Table 2.2: Description of the indicator variables.

Category	**Description**	**Variable Name(s)**
Defenders Inside the Cone	The number of defenders with coordinates inside the cone formed with the ball carrier and the goal posts of the net being attacked.	oneDInCone twoDInCone threePlusDInCone
Defenders Between the Ball and the Net	The number of defenders with an x-coordinate between the ball carriers x-coordinate and the net being attacked.	dBwBallAndNet_2to4 dBwBallAndNet_5to7 dBwBallAndNet_8Plus
Score Difference	The number of goals the team in possession at time t is ahead by.	scoreDiff1Ahead scoreDiff2Ahead scoreDiff3PlusAhead scoreDiffAhead
	The number of goals the team in possession at time t is behind by.	scoreDiff1Behind scoreDiff2Behind scoreDiff3PlusBehind scoreDiffBehind
Pressure	There is a defender within ten yards of the ball carrier and inside the cone.	pressureInCone.10
	There is a defender within ten yards of the ball carrier and outside the cone.	pressureOut.10
Play Pattern	The event at time t is part of the sequence of events that was initiated by a specified event.	corner freekick counter
Possession	The number of passes made in the current possession up to time t is in the specified range.	medPos (4 - 15) longPos ($\geq$ 16)
Position	The position of the player performing the event at time t.	midfield forward
Angle	The angle between the ball carrier and the net being attacked. Measured in degrees.	medAngle (30 - 60) largeAngle (60 - 90)

Table 2.3: Description of the indicator variables (continued from Table 2.2).

Category	**Description**	**Variable Name(s)**
Body Part	The event at time t is executed with the player's head.	header
Offensive Team	The team that is playing and in possession of the ball at time t.	O_England O_Austria O_Norway O_Northern_Ireland O_Spain O_WNT_Finland O_Germany O_Denmark O_Portugal O_Switzerland O_Netherlands O_Sweden O_France O_Italy O_Belgium O_Iceland
Defensive Team	The team that is playing and not in possession of the ball at time t.	D_England D_Austria D_Norway D_Northern_Ireland D_Spain D_WNT_Finland D_Germany D_Denmark D_Portugal D_Switzerland D_Netherlands D_Sweden D_France D_Italy D_Belgium D_Iceland

they play more offensively than defenders. It is also worth noting that the player positions provided in the raw data are related to the player performing the action, whether it is a ball-related action or not. For example, when an opponent applies pressure to the ball carrier, the position identification variable provides the position of the player applying pressure, not the player with the ball. In these situations, we determined the position of the ball carrier by using the last known ball carrier position. Determining the ball carrier position in this way assumes that the ball carrier remains unchanged from the previous time point. Since we are working with event data, any change in possession would be explicitly captured through a relevant event, thus justifying the assumption. As a result, we created indicator variables for the forward and midfield positions that take on a value of one when the ball carrier belongs to that position group and zero otherwise. These variables are designed to measure the effect of the specified position relative to defenders. With this, the impact that a defender has on the probability of an event occurring will be captured in the model's intercept.

Next, the score difference during a game can influence how aggressively a team attacks or how compactly they defend. Because of this, we wanted to capture the effect of the score difference throughout the game. We developed a current score variable for each team since the raw data did not provide one. It is worth noting that own goals are labelled differently in the data, and therefore, an additional check was required to capture them. To verify our results, we cross-checked the produced final scores with the final scores

given in the Hudl StatsBomb game outcomes data set. Once verified, we determined the goal difference at each time point in the game and developed indicator variables that can account for being ahead or behind and by how many goals. We expect to see a relationship between how often shots take place and the score difference, because if a team is behind, they may become more desperate, and we would expect them to perhaps take more shots.

Another variable in the original data describes the event that initiated the sequence of events up to the current time. This variable consists of nine categories: "regular play", "from a corner", "from a free kick", "from a throw-in", "from a counter", "from a goal kick", "from the goalkeeper", "from a kickoff" and "other". From now on, we will refer to this variable as the play pattern. Our primary interest is in events that may lead to a higher chance of scoring; therefore, we focus on sequences initiated by corners, free kicks and counterattacks. We used the following system to create indicator variables for these events. If an event is part of the sequence following a corner, then the corner variable is set to one, and it is set to zero otherwise. The same idea applies to the free kick and counterattack variables. Constructing indicator variables in this way, then implies that the effect of all the remaining play patterns is captured in the intercept of the model. We expect these newly created variables to impact the probability of a shot, a shot on target, and a goal occurring. For example, corner kicks happen close to the opposing net and often involve a cross into the 18-yard box, which can lead to a goal-scoring opportunity. Additionally, counterattacks are plays in which the defending team gets the

ball and quickly attacks the opponent's net. This can catch the competitor off guard and may result in a scoring opportunity before the opponents can recover their defensive shape. Finally, when a defending player commits a foul, a free kick is awarded and can result in a goal-scoring opportunity. Due to the nature of free kicks and since they can occur anywhere on the pitch, it seems likely that this variable will be less significant in estimating the probability of a shot. Nevertheless, we still include this variable in our analysis to account for its potential impact.

Next, the number of defenders between the ball carrier and the goal being attacked can influence whether a shot will take place or not. To capture this impact, we created multiple variables that have the potential to provide different insights. The first variable we developed counts the number of defending players, excluding the goalkeeper, between the ball and the goal line. It is important to note that this variable only considers the x coordinates for the defensive players. Figure 2.4 displays a freeze frame with seven defensive players, excluding the goalkeeper, between the ball carrier and the goal line. It is worth noting that, in this frame, one player from the defending team is missing since there are typically 11 players on each team. This missing player may be beyond the visible area in the lower left corner of the frame. If this is the case, the number of players counted between the ball and the goal would be incorrect. Conversely, if the missing player is situated to the right of the visible area, then there is no impact on the number of defenders. We proceed with our analysis, understanding that this may impact our results. From the

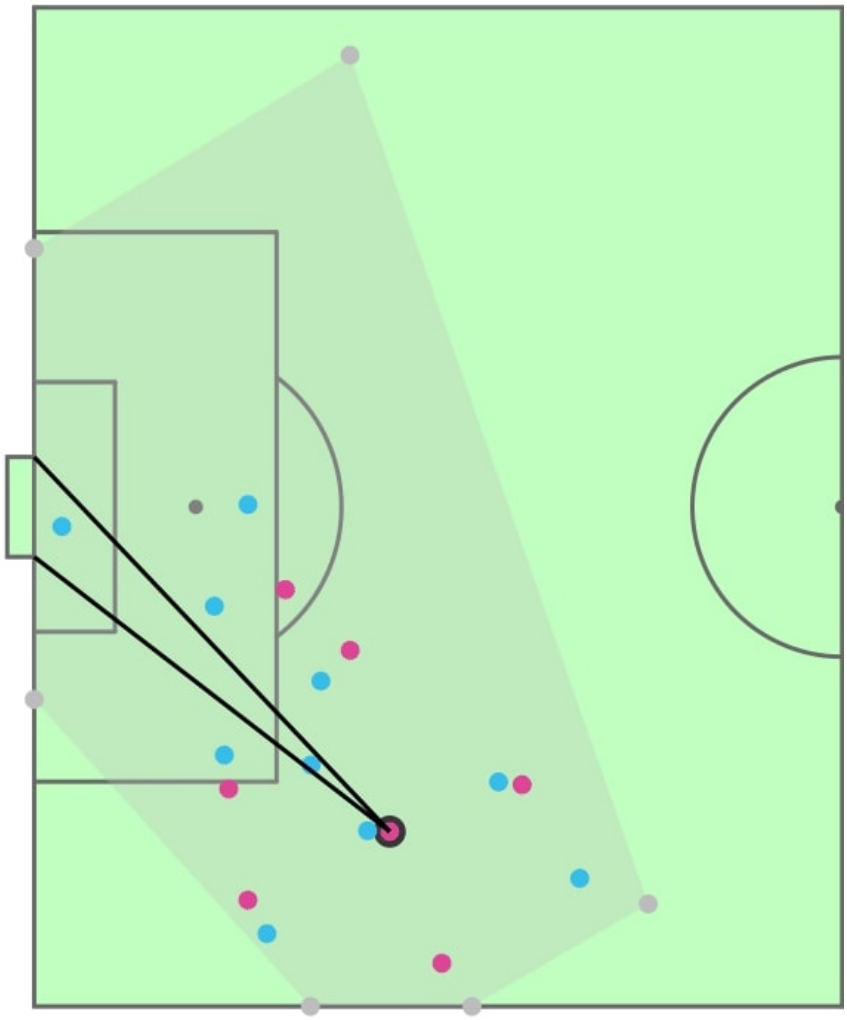


Figure 2.4: The cone created between the ball location and the net is denoted by the two black lines. In this example, there is one defender in the cone, excluding the goalkeeper, and seven defenders in the visible area between the ball and the goal line, excluding the goalkeeper.

figure, we can see that the defending players are not necessarily close to the ball in terms of their y coordinates. This implies that the number of defending players between the ball and the net captures the impact of players who may not directly impact a player's ability to shoot but may have more of an effect on a player's ability to pass or create plays toward the net. To capture this influence, we created three categories: "between two and four defenders", "between five and seven defenders" and "more than eight defenders". Then, each of these categories is associated with an indicator variable. This also implies that the effect of having fewer than two defenders between the ball and the goal line is captured in the intercept. Utilizing categories instead of a single indicator variable for each potential number of defenders simplifies model interpretation, provides more flexibility for the selected coefficients and reduces the number of covariates. If we had instead used a single variable containing the number of defenders between the ball and the net, we would be assuming that the relationship between the different values is linear. This would imply that having two defenders between the ball and the net has twice the impact on the probability of a shot as having only one defender. Since we feel that we cannot make this assumption, we decided to use indicator variables instead.

The second variable developed to account for the defending player positions counts the number of defending players, excluding the goalkeeper, within the cone, which is defined as the area of the field between the ball and the two goalposts of the net being attacked. Figure 2.4 demonstrates how the cone is created. By counting the number of players on the opposite team inside the

cone, we can determine the value used for this variable in this frame. One issue with this variable is that it does not capture defensive pressure when the attacking player is far from the net. This is because the cone is narrower and unlikely to include many defending players. To capture the impact of the number of players in the cone, we created three categories: "one defender", "two defenders" and "more than two defenders". Each category is associated with an indicator variable that takes on a value of one if the number of defenders in the cone matches the variable's value and zero otherwise. Constructing indicators in this way then implies that the effect of having no defenders in the cone is captured in the intercept. We implement indicator variables for the number of defenders in the cone for the same reasons mentioned above in the case of the number of defenders between the ball and the goal line.

In general, these variables have the potential to play different roles when attempting to capture an attacking player's shooting ability. As the number of defenders in the cone and between the ball and the net increases, we believe the probability of a shot will decrease. It is worth noting that the two abovementioned variables exclude the goalkeeper from their counts. It is clear that the goalkeeper being inside the cone or between the ball and the net will impact the probability of a shot occurring. However, since we are working with broadcast data, we do not always have access to the goalkeeper's positioning. Therefore, including the goalkeeper in the abovementioned counts would result in missing information for some frames. To avoid this inconsistency, we exclude the goalkeeper from the counts.

The ball carrier's location relative to the net can also impact whether the player will shoot, shoot on target, and score. As the ball approaches the corners of the field, the angle to the net becomes increasingly acute, and the goalkeeper can better protect the net. To capture this effect, we created an angle variable, which takes on values near 90 degrees when the player is in front of the net and values near zero when the player approaches the corners. This variable was available in the original data when a shot occurred. Therefore, we calculated its values at all time points throughout the match. Figure 2.5 displays a freeze frame where the angle to the net is 45.53 degrees. Using this angle variable directly in our models amounts to assuming that the relationship between its values and the logit of the considered probability is linear, which is not necessarily true. Therefore, we created indicator variables related to the angle between the ball and the net to increase the interpretability of our results. The two categories associated with our angle variable are "between 30 and 60 degrees" and "between 60 and 90 degrees". When the angle falls in the specified range, that variable is set to one; otherwise, it is zero. This implies that when the angle is less than 30 degrees, its effect on the probability of an event occurring will be captured in the intercept of the model.

Based on previous research on expected value metrics in soccer, the distance to the opponent's net is expected to affect the probability of our events of interest (Rathke, 2017). This distance is measured between the player with the ball and the centre of the net being attacked. Similar to the angle variable, the distance to the goal was present in the original data only when a shot occurred,

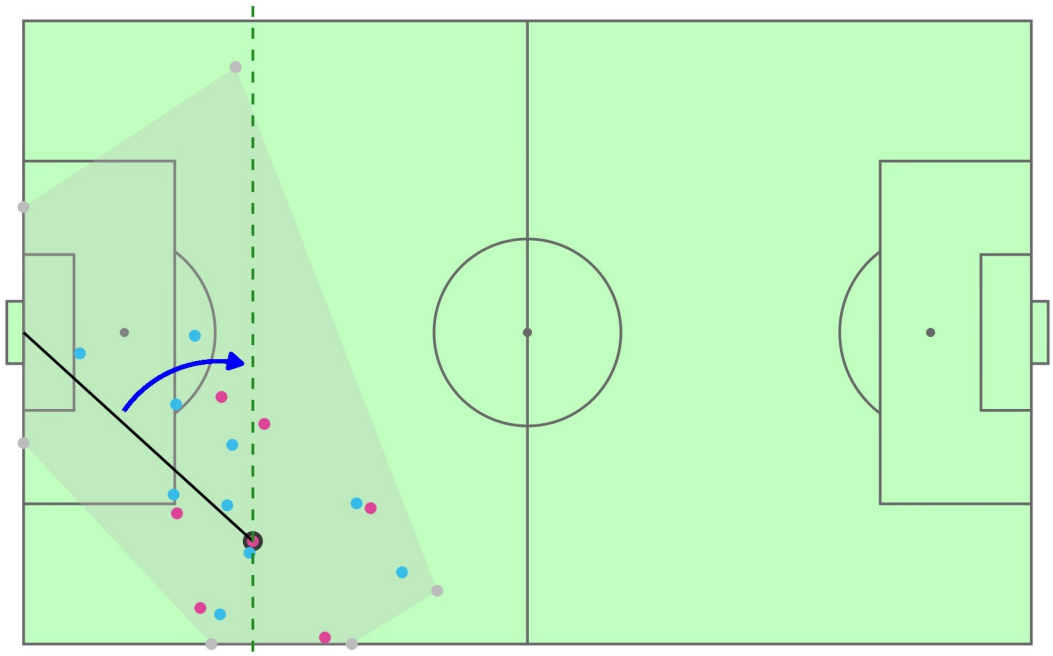


Figure 2.5: Visualization of the angle and distance between the ball's location and the net. In this figure, the angle indicated by the blue arrow is 45.53 degrees, and the distance associated with the length of the black line is 38.26 yards.

therefore, we calculated its value at all time points. A representation of this variable can be seen in Figure 2.5. In general, if the ball carrier is far from the opponent's net, it is unlikely that a shot will take place, but as the player approaches the net, it becomes more likely. Based on this relationship, we expect the probability of a shot to increase as the distance decreases.

Another variable we constructed and used measures the pressure applied within the cone created with the ball location and the goal posts. This variable takes on a value of one if a defender is within a specified distance of the ball carrier and inside the cone. Otherwise, it takes on a value of zero. To

avoid missing values, we implemented an indicator variable rather than using the distance between the ball carrier and the nearest defender in the cone. If we had instead used the distance for this variable, any frames with no defenders in the cone would result in a missing value. We experimented with various distances to determine which seemed the most useful in estimating the probability of a shot on target and a goal. We concluded that ten yards was an effective choice for this distance. In Figure 2.6, the green circles surrounding the ball carrier represent the five- and ten-yard radii. Additionally, the black lines represent the cone formed between the ball carrier and the opponent's net. In addition to pressure applied within the cone, we expect the amount of pressure applied outside the cone to have an impact on the probability of a shot as well. Unlike the pressure applied from within the cone, using indicator variables in this case did not greatly impact the probabilities of interest. Instead, we found that the distance between the ball carrier and the nearest defender outside the cone seemed the most useful. Figure 2.6 displays the distance to the closest defender outside the cone, represented by the pink line. Combining the two variables mentioned above allows us to capture the overall pressure on the ball carrier and hopefully improve our ability to estimate the probability of a shot, a shot on target and a goal.

We originally considered two variables that are related to the current possession in the game. Hudl Statsbomb (2019) defines possession as a period where the ball remains in play, and one team continuously has ball control. A new possession starts when the opposing team establishes control of the ball

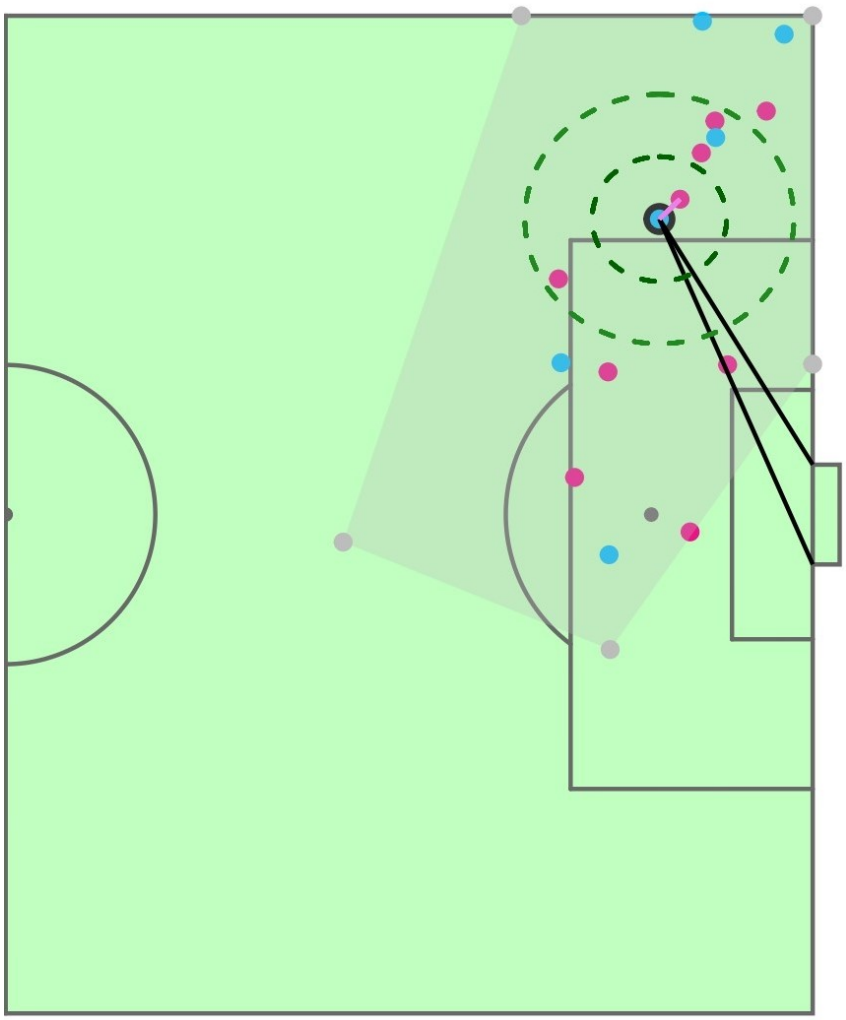


Figure 2.6: The cone is represented by the two black lines connecting the ball carrier and the goal line. The green dashed circles around the ball represent the five and ten-yard radii around the ball. These regions are used to create indicator variables related to pressure applied from inside and outside of the cone. Also depicted in the figure above is the distance to the nearest defender outside the cone, associated with the length of the pink line.

or when the ball goes out of bounds. The first variable we are interested in is the number of passes in the current possession. To incorporate the effect of the number of passes into our model, we developed indicator variables for short, medium and long possessions using the number of passes made up to the current time point. In the 2022 UEFA European Women's Championship, we found that 50% of all possessions had fewer than four passes, and only 25% of all possessions had more than six passes. We also found that the maximum number of passes in a single possession was 39. With this in mind, we used the following conditions to develop our indicator variables. If the number of passes up to the current time point in the current possession was between four and 15, we set the indicator variable representing a medium possession to one and zero otherwise. Additionally, the long possession was assigned a value of one when the number of passes was at least 16. Constructing indicator variables in this way implies that the effect of a possession having fewer than four passes will be captured in the intercept of our model. We used indicator variables instead of the number of passes to avoid making a linearity assumption, as discussed above with the other variables. We also decided to implement possession categories because it would be unreasonable to interpret indicator variables defined for each number of passes. These indicators may impact the probability of a shot since more passes could indicate that a team is in control of the game and is waiting for an ideal moment to attack. The second possession variable we were originally interested in is the elapsed time within the possession. This variable restarts at zero whenever there is a new

possession and counts the length of time from the start of the possession till the current time point. Similar to the variables representing the number of passes, a considerable amount of time in possession may represent control of the game. It is worth noting that this variable and the variables representing the number of passes measure similar features. This similarity is expected to lead to multicollinearity, and therefore, we decided to include only the indicators representing the number of passes and omit the variable accounting for the elapsed time.

Another variable we are interested in studying that other works on expected value metrics have not considered is the depth of the ball relative to the players on the field. Data depth measures the centrality and outlyingness of data objects within a group of similar objects. Typically, depth measurements allow for a central-outward ordering of points by assigning the most central points a high depth and a decreasing depth as points move outwards (Serfling and Zuo, 2000). Using the cleaned data, we looked at the depth of the player with the ball relative to the players on the same team and relative to the players on the opposite team as two potentially interesting variables. These measurements provide information on the team's strategy when attacking the opponent's net. After reviewing various depth measures, we concluded that integrated rank-weighted (IRW) depth was the most relevant and useful in our context.

The IRW depth is defined in Ramsay et al. (2019) as follows. Consider a query point y and a set of n unique univariate points. We denote this

set by $X = (X_1, \ldots, X_n)'$ and the empirical cumulative distribution function associated with it by F_n. Then, the univariate depth of y with respect to the set X is defined by

$$D(y, X) = \min(r(y) - 1, n - r(y)),$$

where $r(y) = nF_n(y) + 1$ is the rank of y within X. Now, consider a query point y and a sample X in $\mathbb{R}^2$, and let the sample projected onto a unit vector $u \in \mathbb{R}^2$ be denoted by $X(u) = (u'X_1, \ldots, u'X_n)'$. The IRW depth of y with respect to X is defined as

$$D_{IRW}(y, X) = \frac{1}{\pi} \int_S D(u'y, X(u))du,$$

where S represents the unit circle. This definition implies that the IRW depth of y is the average univariate depth of y over all possible univariate projections onto a unit vector from $\mathbb{R}^2$. Finally, since the univariate depth is constant over certain sections of S, we can simplify the above expression and write

$$D_{IRW}(y, X) = \sum_{k=1}^{\lceil n/2 \rceil} w_k(y, X)k,$$

where $w_k(y, X)$ is the proportion of projections for which y has a univariate depth of k. This simplified equation shows that we can obtain the IRW depth using a weighted average based on how often our query point is at the specified depth among all possible univariate projections of the query point and sample.

IRW depth has numerous properties of interest. In particular, IRW depth reaches its maximum in the centre of a data cloud and gradually approaches zero as the query point grows without bounds (Ramsay et al., 2019). This is important for our analysis since the ball carrier is often beyond the perimeter of the data cloud. IRW depth produces an interpretable value that is sensitive to how far beyond the data cloud the query point is. By comparison, more common measures of depth, such as Tukey and Simplicial depth, will give a depth value of zero when the ball is outside the data cloud, regardless of its distance from the cloud. This can be seen in Figure 2.7 where the depth relative to the players from the opposite team is calculated for five query points using Simplical, Tukey and IRW depth. From these plots, we can see that once the query point lies beyond the cloud of data points, the only measure of depth that gives a meaningful value is the IRW depth. As the query point gets further from the data, we see a gradual decrease in the IRW depth value. On the other hand, Tukey and Simplicial depth take on a value of zero whenever the query point is outside the data cloud. Additionally, IRW depth is a continuous function of the query point, which is crucial in our analysis. In contrast, Simplicial and Tukey depths result in discontinuous step functions that change substantially and suddenly. In Figure 2.7, we see that the change in depth values is faster in the Tukey and Simplicial depths; conversely, the IRW depth changes gradually, which seems desirable for our analysis. This property is important for our research because we expect the probability of an event occurring to vary continuously with the ball position. Because of

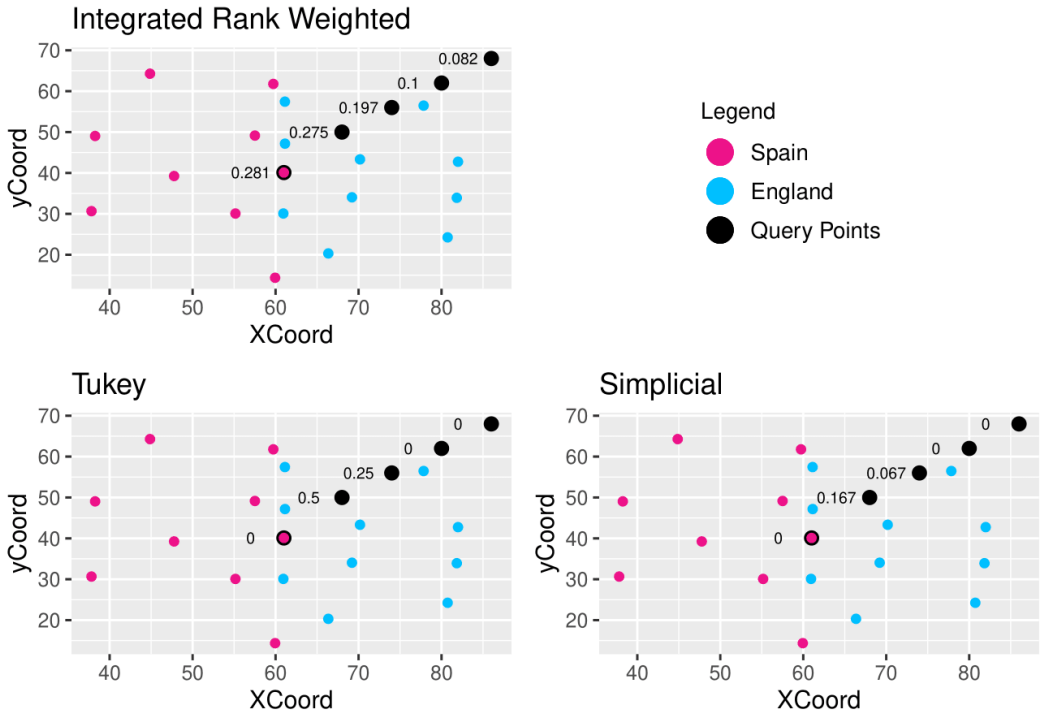


Figure 2.7: Depth of five query points relative to the opposite team using Integrated Rank-Weighted (IRW), Simplicial and Tukey depth, scaled such that all depth measures range from 0 to 0.5.

this relationship, it seems natural to think that implementing a measure of depth with similar properties will better capture these changes. Therefore, considering the above-mentioned properties, we concluded that IRW depth was the best option for our research.

Various challenges arose while working with the data that required further investigation. First, since our data is collected via broadcast footage, we only have access to the player positions within the camera range. This suggests

that the depth of the ball is measured relative to the visible players instead of all players, which will impact the calculated depth values. A caveat described in Hudl Statsbomb (2021) is that in some cases, the goalkeeper is on the same team as the actor, but they will not be listed as a teammate. As a result, the goalkeeper will be counted as a member of the incorrect team when using the teammate attribute to assign players to teams. We excluded the goalkeeper from all depth calculations to avoid issues with this caveat.

Another variable that may impact the probability of a shot, a shot on target and a goal is the body part of the player attempting the shot that makes contact with the ball. To account for this potential effect, we developed an indicator variable that takes on a value of one when contact with the ball is made using the head and is zero otherwise. We expect this variable to impact the probability of a shot, a shot on target and a goal. For instance, a common play in soccer is a cross into the 18-yard box. In this situation, headers are very likely; therefore, details about the body part used can provide insights into the outcome of the play and whether the team in possession has a shooting opportunity.

Interaction terms account for how variables interact and how that relationship impacts the response. In our data, it is not unreasonable to assume that many of the variables could interact with each other. As a result, we developed various interaction-like variables to capture these joint effects. The first of these uses two of the previously discussed variables. The two variables we use are the distance between the ball location and the goal line and the

number of defenders in the cone. We define this interaction-like variable as,

$$\text{dInCone_dist} = \begin{cases} \dfrac{(\text{numDInCone} + 1)}{\sqrt{\text{distToAttackingNet}}} & \text{if distToAttackingNet} \\ & \text{is not NA,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where numDInCone is an integer representing the number of defenders in the cone produced by the ball carrier and the goalposts of the net being attacked, as seen in Figure 2.4. These two variables are naturally related, and therefore we are interested in investigating how they will jointly impact the probability of a shot, a shot on target and a goal. For instance, as the shape and size of the cone change, we expect the number of defenders in the cone to change as well. From (2.1), it is evident that an increase in the number of defenders in the cone or the distance to the net will result in an increased interaction term. This, in turn, would likely decrease the probability that a player will shoot, shoot on target and score.

The second interaction-like variable developed for our model uses the distance from the ball location to the nearest defender in the cone and the number of defenders in the cone. This variable is intended to measure the pressure being applied to the player with the ball. This pressure term is zero when there are no defenders in the cone. Otherwise, when defenders are in the cone, we use the distance between the closest defender and the ball to obtain the

pressure term. This interaction term is defined as,

$$\text{pressure} = \begin{cases} 0 & \text{if numDInCone is zero,} \\ e^{-\sqrt{\text{closestIn}}} & \text{otherwise,} \end{cases}$$

where closestIn represents the distance to the nearest defender in the cone when one is present. Based on this equation, the smaller the distance between the defender and the ball, the higher the value of the pressure term will be. Additionally, the pressure decreases very rapidly as this distance increases.

Additionally, soccer involves various team strategies that may impact the probability of a shot, a shot on target and a goal. For instance, Spain is renowned for their sequences of passes while waiting for the perfect opportunity. Germany typically adopts a very tactical approach, rarely deviating from their game plan. Conversely, England generally emphasizes advancing the ball up the field with minimal passes. The diversity in playing styles and overall strength of teams is essential when modelling shots. One significant challenge is that a team's performance varies considerably depending on the opponent. For instance, when England plays against a strong opponent such as Spain, they may struggle to get good shooting opportunities. Conversely, in a game against a weaker team, such as Northern Ireland, England may dominate the possession and have many shooting opportunities. To capture this impact, we developed offensive and defensive covariates for each team in the tournament except for Spain, whose offensive and defensive impact will be captured in the intercept of the model. A team's offensive covariate takes on a value of one

when playing and in possession of the ball, otherwise, it is zero. Additionally, the team's defensive covariate takes on a value of one when they are playing but not in possession of the ball. Implementing our team variables in this way will potentially allow us to better understand how each team attacks and defends. This will also hopefully demonstrate how much of an impact the opponent has on a team's performance.

It is worth noting that team strategy can significantly impact the probability of a shot, a shot on target and a goal. Although not directly used in this research, team formation could potentially be used to capture the impact of team strategy. For instance, a team could play in a more defensive formation, such as four defenders, five midfielders and one forward or a more offensive formation with three defenders, four midfielders and three forwards. In general, some formations will be more effective against different types of teams. The effects of this would depend significantly on the opponents' skill and their formation. Due to our limited data, we do not include any covariates that capture this effect, but it would be a factor worth considering in future research.

Finally, we note that a key feature in modelling outcomes in soccer is the home team advantage, as shown by Nevill and Holder (1999) and Goumas (2017) who studied the importance of home team advantage across various sports. It is common to believe that a team playing at home has the advantage of being comfortable in their surroundings and has a fan base to support them. Since we are working with tournament data, this attribute is not included in

our model. All the games in the tournament are played in England, which implies that the home team advantage is confounded with England's team effect. As a result, the home team effect is not estimable in our situation.

In this section, we provided a detailed explanation of the variables that are potentially important for estimating the probability that one of the specified events occurs. In the following chapter, we proceed with a discussion of how we implement dependent and sequential logistic regression to estimate the probability of a shot, a shot on target and a goal.

Chapter 3

Modelling Soccer Matches

In this chapter, we discuss the implementation of dependent logistic regression to estimate the probability of an open play shot occurring at any point during a game. Furthermore, we discuss the application of sequential logistic regression to estimate the probability of a shot on target and a goal occurring in open play. Finally, we define metrics associated with expected open play shots, shots on target, and goals throughout the game.

3.1 Dependent Logistic Regression

In this section, we will explore the model used to assess the probability of an open play shot and outline the expected value metric related to the number of such shots. Lastly, we will discuss our model's fitted values and expected value metrics.

3.1.1 Model and Definitions

We employ dependent logistic regression to estimate the probability of a shot occurring at any moment. Our motivation for using dependent logistic regression is as follows. First, logistic regression requires that the response be a binary variable, taking on a value of one when an event occurs and zero otherwise. In our research, shots represent successes, meaning that our response variable will take on a value of one when a shot occurs at time t and zero otherwise, thereby satisfying the first assumption. Secondly, logistic regression requires that the response variable be independent of its previous values. Since we are modelling whether a shot will occur at some specific time in a soccer match, we can confidently assert that this assumption is not satisfied. In soccer, events that occur at the current time point are potentially highly dependent on prior events. Therefore, we focus on dependent logistic regression, which allows for dependence between response variables. With these considerations, we have confirmed the applicability of dependent logistic regression, and we proceed with a more formal discussion of the concepts.

To begin, Hudl Statsbomb (2019) defines a shot as a goal-scoring attempt using any legal body part. We work from that definition and further define a shot on target as a shot that resulted in one of the following outcomes: "Goal", "Saved", "Post", "Saved to Post" or "Blocked". Finally, a goal is defined in Hudl Statsbomb (2019) as a shot that crossed the goal line.

We aim to use the information available at each time point to estimate the

probability of a shot being taken by the team controlling the ball. For this, let S_t denote the event that a shot is taken in open play at time t and consider the binary response variable

$$Y_{1t} = \begin{cases} 1 & \text{if a shot occurs in open play at time } t, \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

for times $t = 1, 2, \ldots, n$, where n is the number of recorded events in a game. It is important to note that we are only interested in shots that occur during open play. In our research, this excludes penalty kicks, direct free kicks and direct corner kicks. We excluded these events because they have unique features that may impact the results. Our response variable Y_{1t} has a Bernoulli distribution, and we expect neighbouring variables to be dependent. Therefore, we perform dependent logistic regression with shots representing successes. We write

$$P_1(t) = \mathbb{P}(S_t|I_{1t}) = \mathbb{P}(Y_{1t} = 1|I_{1t}),$$

where I_{1t} is the information set available at time t and is defined by

$$I_{1t} = \{y_{t-i}, x_{1jt} \ : \ 1 \leq i \leq l, \, 1 \leq j \leq p_1\}.$$

This set includes the values of the p_1 covariates at time t, represented by x_{1jt} and past values of the response variable, represented by y_{t-i}. The number of lagged values, considered to model dependence, is represented by l. This information set can be viewed as the event where $X_{1jt} = x_{1jt}$ and $Y_{t-i} = y_{t-i}$, hence we will treat it as an event and make use of common set notation. It

is understood that this may be seen as a slight abuse of notation; however, it avoids the use of more cumbersome notation. The probability of a shot occurring at a given time point is then modelled by,

$$P_1(t) = \mathbb{P}(S_t|I_{1t}) = g\left(\beta_{10} + \sum_{j=1}^{p_1} \beta_{1j} x_{1jt} + \sum_{i=1}^{l} \alpha_{1i} y_{t-i}\right), \tag{3.2}$$

where

$$g(x) = \frac{e^x}{1+e^x}. \tag{3.3}$$

The terms $\beta_{11}, \beta_{12}, \ldots, \beta_{1p_1}$ are the coefficients associated with each of the p_1 covariates in our model and β_{10} represents the intercept. Some variables of interest include distance to the net, angle to the net, number of passes in the current possession, defender locations and body part in contact with the ball. Finally, $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1l}$ are the coefficients associated with the lagged response variables. Other forms of dependence could be considered, but we restricted our attention to Autoregressive models in this work.

Using the above, we can define the Expected Shot Value (ESV) metric for team A in a single game as,

$$\text{ESV}_A = \sum_{t \in C_A} \widehat{\mathbb{E}}(Y_{1t}|I_{1t}) = \sum_{t \in C_A} \widehat{P}_1(t),$$

where $C_A \subset \{1, 2, \ldots, n\}$ is the set of times where team A has possession of the ball and $\widehat{P}_1(t)$ corresponds to the fitted values from our dependent logistic

regression model. Similarly, we can define the ESV metric for team B in a single game as,

$$\text{ESV}_B = \sum_{t \in C_B} \widehat{\mathbb{E}}(Y_{1t}|I_{1t}) = \sum_{t \in C_B} \widehat{P}_1(t),$$

where C_B is the set of times where team B has possession of the ball. Then, from this, the combined ESV metric for a single game is defined as

$$\text{ESV} = \text{ESV}_A + \text{ESV}_B. \tag{3.4}$$

Based on these definitions, we see that our ESV metrics depend on the information available at each time t by conditioning on the information sets $I_{11}, I_{12}, \ldots I_{1n}$. Hence, our ESV metrics are calculated as the game evolves, capturing in-game fluctuations with the goal of better understanding how teams perform throughout the game. To accomplish this, we implement our ESV metrics using conditional probabilities. Finally, to avoid cumbersome notation, we will proceed with the understanding that when we refer to the ESV metric, we are usually referring to the ESV metric of one team in one game; this should be clear based on context.

Another method of measuring team performance is to simply use the observed number of open play shots, that is, to consider the total number of shots taken but excluding penalty shots, direct free kicks and direct corner kicks. One major disadvantage of this metric is that it does not account for specific events throughout the match. For example, some playing styles involve waiting for the perfect shooting opportunity. Because of this, a team may have

the ball for long periods near the goal but wait until they have the ideal opportunity to shoot. In this case, the observed number of shots will perhaps report that one shot was taken, whereas our ESV metric will indicate that a typical team in this situation would have generated more shooting opportunities based on how they controlled the ball. Comparing these two approaches, we expect our ESV metric to better capture the game's progression than the observed number of shots.

To better appreciate our suggested definition of ESV given in (3.4), consider calculating the expected number of shots in open play for a game. To start, since the probability of a shot occurring depends on the values of the covariates, we need knowledge of the covariate distributions at all times during the game to carry out the calculations outlined below. The following derivation demonstrates how $\mathbb{E}(S)$, the expected number of shots (before the match begins), could be approximated if these distributions were available and illustrates how this expectation is related to our ESV metric. First, using the Law of Total Expectation, we write the total number of open play shots for a game as $S = \sum_{t=1}^{n} Y_{1t}$, and write its expectation as

$$\begin{aligned}\mathbb{E}(S) &= \sum_{t=1}^{n} \mathbb{E}(Y_{1t}) \\ &= \sum_{t=1}^{n} \mathbb{E}\big[\mathbb{E}(Y_{1t}|X_t)\big],\end{aligned}$$

where Y_{1t} is defined in (3.1) and X_t denotes the vector of covariates at time t to avoid more cumbersome notation. Then, from the definition of an expectation, we can write

$$\mathbb{E}(S) = \sum_{t=1}^{n} \int_{R_t} \mathbb{E}(Y_{1t}|X_t = x) f_{X_t}(x)\, dx, \tag{3.5}$$

where $f_{X_t}(x)$ is the joint probability density function for the covariates at time t and R_t is the range of these covariates. Depending on covariate types, the generic integral in (3.5) may be multivariate and include a combination of summations and integrals. For continuous covariates, it incorporates an integral and the probability density function; for discrete covariates, it uses a summation and the probability mass function.

Next, using the definition of a conditional expectation, we can show that the conditional expectation in (3.5) simplifies in the following way;

$$\begin{aligned}\mathbb{E}(Y_{1t}|X_t = x) &= \sum_{y_{1t}=0}^{1} y_{1t}\mathbb{P}(Y_{1t} = y_{1t}|X_t = x) \\ &= \mathbb{P}(Y_{1t} = 1|X_t = x),\end{aligned}$$

since Y_{1t} is a Bernoulli random variable. Hence (3.5) becomes

$$\mathbb{E}(S) = \sum_{t=1}^{n} \int_{R_t} \mathbb{P}(Y_{1t} = 1|X_t = x) f_{X_t}(x)\, dx. \tag{3.6}$$

Now, assuming the covariate distributions are known, each integral in (3.6)

can be approximated through a Monte Carlo simulation as

$$\int_{R_t} \mathbb{P}(Y_{1t} = 1|X_t = x) f_{X_t}(x) dx \simeq \frac{1}{B} \sum_{j=1}^{B} \mathbb{P}(Y_{1t} = 1|X_t = x_{jt}),$$

where B is a suitably large number and $x_{1t}, x_{2t}, \ldots, x_{Bt}$ are independent and generated from f_{X_t}. Here, $x_j = (x_{j1}, \ldots, x_{jn})'$ can be thought of as replicates obtained from simulating different games, numbered from 1 to B. From this, we obtain that (3.6) could be approximated by

$$\begin{aligned} \mathbb{E}(S) &\simeq \sum_{t=1}^{n} \left[\frac{1}{B} \sum_{j=1}^{B} \mathbb{P}(Y_{1t} = 1|X_t = x_{jt}) \right] \\ &= \frac{1}{B} \sum_{j=1}^{B} \left[\sum_{t=1}^{n} \mathbb{P}(Y_{1t} = 1|X_t = x_{jt}) \right]. \end{aligned}$$

Finally, $\sum_{t=1}^{n} \mathbb{P}(Y_{1t} = 1|X_t = x_{jt})$ is the ESV metric for game j when parameters are known in this logistic regression model. Based on the above derivations, we conclude that

$$\mathbb{E}(S) \simeq \frac{1}{B} \sum_{j=1}^{B} ESV_j,$$

is the average ESV computed over many simulated games. This derivation demonstrates that $\mathbb{E}(S)$ is incompatible with our goal of developing a metric that evolves with the match and gives a match-specific measure of performance. As we have seen, the calculation and/or approximation of $\mathbb{E}(S)$ requires the

Law of Total Expectation, which in turn requires knowledge of the covariate distributions. These distributions are challenging to obtain and assuming them to be known would be unreasonable in our research. Additionally, $\mathbb{E}(S)$ provides a metric that is averaged over many games, but we are primarily interested in obtaining a metric defined for each specific game to compare games played by different teams.

Considering these ideas, we determined that creating an ESV metric that is game-specific, accounts for the events of each game and avoids assumptions about covariate distributions aligns more closely with our objectives. From this, we will proceed with our defined ESV metric. The following section will examine statistical techniques to determine which variables seem most relevant to our analysis and the results obtained when fitting our models to our soccer data.

3.1.2 Fitted Values and Expected Value Metric

In this section, we analyze the results obtained from fitting our model to the UEFA 2022 data. In general, the primary goodness-of-fit test used for binary logistic regression is the Hosmer-Lemeshow test. This test assesses whether the observed number of events matches the expected number of events based on population subgroups (Nattino et al., 2018). Unfortunately, in its current form, this test requires that the data be independent, and since our goal is to test the fit of a dependent logistic regression model, we cannot use it directly. We

have not found a goodness-of-fit test directly applicable to dependent logistic regression models, so we rely on various sanity checks to verify the fit of our model. We begin by checking how often our observed shot frequencies coincide with the expected proportions obtained from fitting our model. To accomplish this, we divided our range of fitted probabilities into ten subgroups, with the i^{th} subgroup range being denoted by $\hat{p}_{i,\min}$ and $\hat{p}_{i,\max}$, as shown in Table 3.1. We denote the number of times our estimated probabilities landed within the specified range by N_i and the number of times our estimated probability landed within the specified range and a shot was taken by s_i. Finally, the observed frequencies are denoted by $f_i = \frac{s_i}{N_i}$ and given in the last column.

Table 3.1: Estimated probabilities vs. observed frequency of open play shots using the full model. For example, from the first row of the table, we see there were 40775 instances of game play with an estimated shot probability between 0 and 0.05. Among these instances, a total of 72 shots occurred, resulting in an observed frequency of 0.0018.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	f_i
1	0.00	0.05	40775	72	0.0018
2	0.05	0.10	814	87	0.1069
3	0.10	0.20	673	117	0.1738
4	0.20	0.30	340	69	0.2029
5	0.30	0.40	156	48	0.3077
6	0.40	0.50	99	35	0.3535
7	0.50	0.60	51	19	0.3725
8	0.60	0.70	36	19	0.5278
9	0.70	0.80	46	36	0.7826
10	0.80	1.00	72	65	0.9028

Our goal is to check whether f_i falls between $\hat{p}_{i,\min}$ and $\hat{p}_{i,\max}$. If it does, this indicates that our model is performing as expected. In Table 3.1, as the row index increases, N_i and s_i both decrease, a result that is in line with the fact that good shooting opportunities tend to be rare in soccer. We also note that some observed frequencies do not fully align with the expected probabilities. For example, we obtain an observed frequency of 0.3535 in the sixth row. This value is below $\hat{p}_{i,\min} = 0.40$, but this result is not entirely surprising, as there are only 99 occasions during gameplay that correspond to this range of shot probabilities.

Next, in Figure 3.1 we have displayed a plot of the ESV metric obtained from fitting our model to the UEFA 2022 data and the observed number of shots in each game. The $y = x$ line represents the values where the observed number of shots and the ESV metric are equal. Using this as a reference, we see that most points are appropriately scattered about the $y = x$ line. We note, however, that the points on the right-hand side of the plot are more sparse. This is unsurprising as soccer is usually a low-scoring game, and this region indicates a higher number of observed shots and a higher ESV metric. Based on the produced figure and table, it seems like our model for estimating the probability of a shot is performing reasonably well.

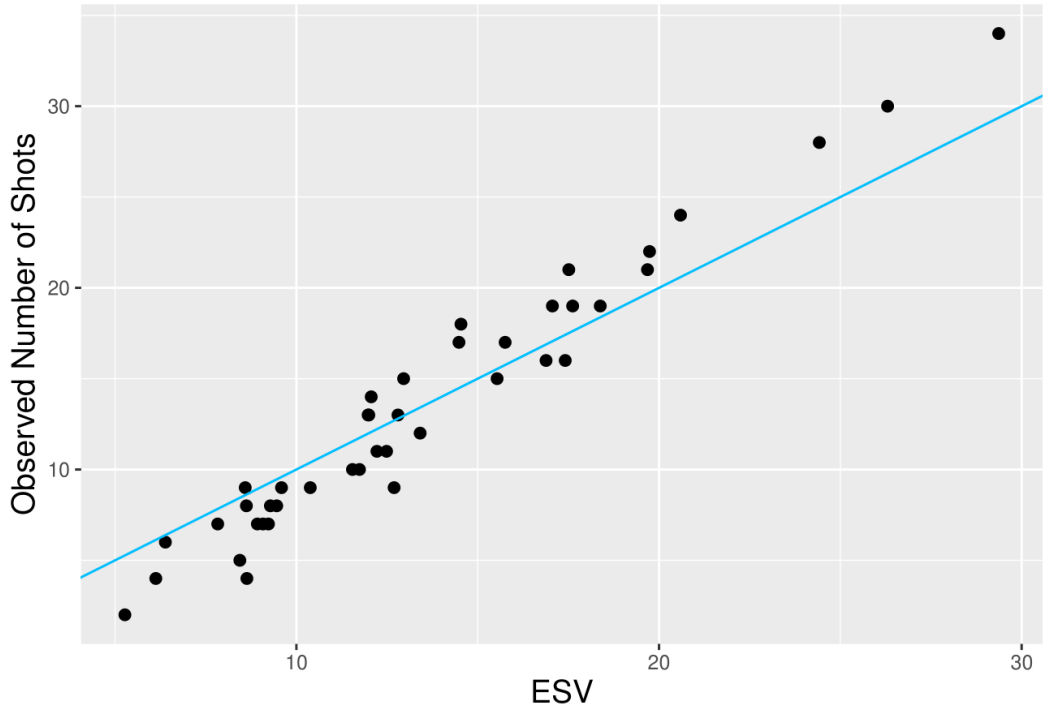


Figure 3.1: Plot comparing the Expected Shot Value (ESV) and the observed number of open play shots for each game and team. The blue line represents the $y = x$ line, and each point represents one team's performance in a game.

3.1.3 Practical Considerations

To perform dependent logistic regression in R, we use the `logitx` function from the `gets` package. This function estimates an autoregressive logit model by maximizing the logit likelihood (Sucarrat, 2024). The `logitx` function was selected over the commonly used `glm` function since the `logitx` function allows for the seamless incorporation of lagged response variable values. We compared the results obtained from the `glm` and `logitx` functions for a simple logistic

regression problem to guarantee that this function performs as expected. We believe the `logitx` function is reliable for our analysis since the two functions performed similarly when lagged values were included (or not) in our models. This being said, while reviewing the performance of the `logitx` function, we found that the p-values associated with the estimated coefficients in our dependent logistic regression model were incorrect. After further investigation, we found that the `logitx` function was producing a one-sided p-value instead of a two-sided p-value in its summary table. We contacted the package's author, and they updated their package that is available on their github repository. The update will be made in the Comprehensive R Archive Network repository as soon as possible. For the remainder of this research, we implement our dependent logistic regression with the updated package obtained from the author, as it produces the correct p-values based on our goals.

To capture the dependence of time points in our model, we use an autoregressive (AR) term. The AR term in a dependent logistic regression model allows current values to be predicted based on a linear combination of past response variable values. This implies that the earlier time points are valuable in predicting the current ones. This is applicable in soccer because when a shot, a shot on target or a goal occurs at time $t-1$, it could impact the chances of another event happening at time t. To select the number of lagged values used in our model, we reviewed the model fitting results using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These are model selection procedures that can assist in the lagged value selection process. AIC

uses a maximum likelihood approach, as discussed in Akaike (1974), and BIC implements model selection using Bayes estimators as discussed in Schwarz (1978). Based on our tests, the most informative model was the second-order lagged AR model, which we refer to as AR(2). Going forward, this is the only order that was considered for our analyses.

Now, to implement dependent logistic regression using the `logitx` function, we combined the data frame from each of the 21 games into one large data frame. The resulting data frame has more than 45000 rows, and each row represents an event in one of the 21 games, starting with the first event in the first game and continuing till the last event in the 21^{st} game. Because the data is combined into a single data frame in this manner, there are situations in our dependent logistic regression where the lagged response variable used in our model comes from the end of the previous game. For instance, when using an AR(2) model, the first two frames of game two will use a lagged response variable pulled from the end of game one. With this, there are a total of 40 incorrectly pulled lagged values (out of more than 45000 entries in the likelihood function). To verify that these lagged values do not significantly impact the results, we reversed the order of the games in the combined data frame, thus changing the lagged values. We found no significant difference in the results, demonstrating that the small number of lagged values pulled from previous games does not impact the results.

The abovementioned sanity check is a reasonable approach to ensure that

the incorrectly pulled lagged response variables do not impact our model's results. However, we can further solidify this statement by modelling the games separately and ensuring that no incorrectly lagged values are used in the models. Constructing our own log-likelihood function in the process, this log-likelihood was then maximized using the `optim` function, which we provided with an initial vector of coefficient values obtained from the `logitx` function output. We implemented this alternate method for fitting our model to the soccer matches and found results similar to the ones originally obtained with the `logitx` function. Having verified that the issue around lagged responses has no discernible impact on model fitting, we proceed by implementing and fitting our models with one data frame and the `logitx` function.

In general, it is important to proceed cautiously when interpreting the coefficient values produced by fitting our models. Each coefficient represents the variable's impact on the independent variable, conditionally on the other variables already being used in the model. In particular, when we first included the team variables, the coefficients were not producing the expected results. Instead, the coefficients for our offensive team variables had what looked like the incorrect sign based on our understanding of soccer. To investigate this potential discrepancy, we estimated the probability of a shot occurring using only information about the teams playing. This produced coefficients that aligned with our expectations. For instance, the strongest offensive teams had positive offensive strength parameters. Next, we added individual variables to our model to determine if the relationship between variables affected the sign

of our coefficients. We determined that the coefficients for the offensive team changed when our model included the distance to the net being attacked. The distance variable plays a significant role in estimating the probability of a shot, and after further investigation, we believe this relationship can be explained as follows. It is clear that as teams approach the opponent's net, the probability of a shot occurring is likely to increase. However, it is also true that stronger teams will be closer to their opponents net more often than weaker teams. For instance, strong teams are more likely to wait for the perfect opportunity to shoot. They will pass the ball around the opponent's net until they can almost guarantee that they will get a shot on target. Conversely, weaker teams won't get as many good opportunities and will take advantage of any chance they get. For instance, consider an offensively strong team such as Spain. When we fit the model using only team variables, most of the other team's estimated coefficients are smaller than Spain's, which is expected. However, when the distance variable is added to the model, most of the other teams have larger estimated coefficients than Spain. This result may seem counterintuitive, but upon further investigation, it can be explained as follows. Spain is a dominant team known for their patience and willingness to pass the ball until the perfect chance arises. Because of this, Spain spends a lot of time near the net and doesn't shoot at every opportunity. The weaker teams in the tournament are expected to shoot more than Spain, when we know they are close to the net, since they are more likely to take advantage of the chances they get. This explains why the coefficients of the offensive team variables have the opposite

sign when distance is included in our model. This investigation demonstrates the importance of being cautious when making inferences based on models produced with many covariates.

3.2 Sequential Logistic Regression

3.2.1 Models and Definitions

Sequential logistic regression involves performing binary logistic regression in multiple successive steps, where the later variables are modelled conditionally on the previous ones (Kauppi and Saikkonen, 2008). As a first step, we perform the dependent logistic regression discussed in Section 3.1. In Figure 3.2, this step refers to the first node of the tree. We discussed the use of dependent logistic regression to model the probability of a shot occurring. Recall that our response was a binary variable denoting whether or not a shot occurred at time t, and we introduced the covariates to be used to predict this binary variable in Section 2.3.

Next, we reduce the data set to contain only rows where a shot took place and perform logistic regression to estimate the probability of a shot on target, conditionally on a shot being taken. For this, let T_t denote the event that a shot on target is taken in open play at time t. This step refers to the second node in Figure 3.2. The number of shots that took place is denoted by

$$n_1 = \sum_{t=1}^{n} \mathbb{I}(S_t) = \sum_{t=1}^{n} Y_{1t}, \tag{3.7}$$

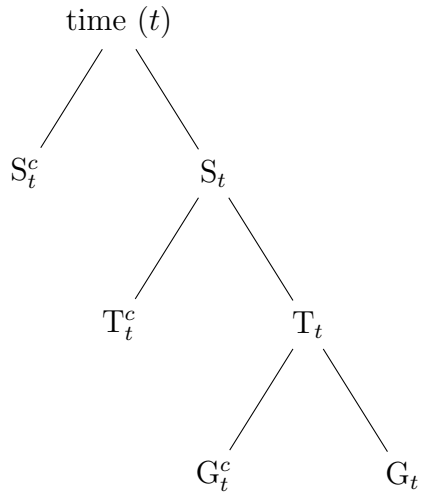


Figure 3.2: Tree diagram showing shots (S), shots on target (T) and goals (G) at some time t during a game.

where n is the number of frames in the game and $\mathbb{I}(A)$ stands for the indicator function that event A occurs. The response variable in the second model is defined as,

$$Y_{2t} = \begin{cases} 1 & \text{if a shot on target occurs in open play at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

where $t = 1, 2, \ldots, n$. From this, we can model the probability of a shot on target occurring at time t by,

$$\begin{aligned} P_2(t) &= \mathbb{P}(T_t | I_{1t} \cap I_{2t}) \\ &= P_1(t)\mathbb{P}(T_t | S_t \cap I_{2t}), \end{aligned} \tag{3.8}$$

where $P_1(t) = \mathbb{P}(S_t|I_{1t})$ is defined in (3.2),

$$\mathbb{P}(T_t|S_t \cap I_{2t}) = g\left(\beta_{20} + \sum_{j=1}^{p_2} \beta_{2j} x_{2jt}\right), \tag{3.9}$$

with $g(x)$ defined in (3.3) and the information set I_{2t} is defined as

$$I_{2t} = \{x_{2jt} \ : \ 1 \leq j \leq p_2\}.$$

With this, we assume that, conditionally on a shot being taken, a shot will be on target or not, independently of all previous shots. Since we are not implementing dependent logistic regression at this stage, the information set does not include lagged response variables. The terms $\beta_{21}, \beta_{22}, \ldots, \beta_{2p_2}$ are the coefficients associated with each of the p_2 covariates in the second stage model and β_{20} is the intercept for that model. It is worth noting that these coefficients and their associated covariates are not necessarily the same as those used in the first stage, and that we can have a different number of covariates used between the first two stages (ie. p_1 does not necessarily equal p_2). Additionally, the variables in the second stage are labelled from 1 to p_2; however, this does not necessarily imply that the j^{th} variable will be the same in the two models. Hence, we refer to these covariates as x_{2jt} in the second model and x_{1jt} in the first.

Let us briefly examine and justify that the intersection notation used in (3.8) for information sets (which we see as events, see Section 3.1.1) is

appropriate in this context. For this, we consider a simple example. Let the event A be defined as

$$A = \{X = 2\} = \{X = 2,\ Y \in \mathbb{R}\},$$

and event B be defined as

$$B = \{Y = 3\} = \{X \in \mathbb{R},\ Y = 3\}.$$

From this, it is clear that

$$A \cap B = \{X = 2,\ Y = 3\}.$$

The above illustrates how using $I_{1t} \cap I_{2t}$, the intersection of our information sets, allows us to condition on all values of the covariates included in these sets and available at time t.

It is also important to note that in (3.9), the shots on target are modelled conditionally on a shot being taken at time t; hence, estimation is only based on n_1 frames, with n_1 defined in (3.7). Using the above, we can define the Expected Shot on Target Value (ETV) metric for team A in a single game as,

$$\text{ETV}_A = \sum_{t \in C_A} \widehat{P}_2(t) = \sum_{t \in C_A} \widehat{\mathbb{P}}(S_t | I_{1t}) \widehat{\mathbb{P}}(T_t | S_t \cap I_{2t}),$$

where $C_A \subset \{1, 2, \ldots, n\}$ is the set of times where team A has possession of the ball and $\widehat{\mathbb{P}}(S_t | I_{1t})$ and $\widehat{\mathbb{P}}(T_t | S_t \cap I_{2t})$ correspond to the fitted values from the first two stages of our sequential logistic regression model. Then, we can

define the ETV metric for team B in the same game in a similar manner, and the combined ETV metric for that game as

$$\text{ETV} = \text{ETV}_A + \text{ETV}_B.$$

Similar to the ESV metric, it is crucial to note that our ETV metrics depend on the events in the game leading up to each time t. This metric is not equivalent to the expected number of shots on target throughout the game when calculated before the start of the game. To avoid cumbersome notation, we will proceed with the understanding that when we refer to the ETV metric, we are usually referring to the ETV metric of one team in one game; this should be clear based on context.

Finally, we reduce the data set again to include only rows containing a shot on target and perform logistic regression to model the probability of a goal being scored conditionally on there being a shot on target. This final step refers to the third node in Figure 3.2 and is modelled conditionally on a shot on target having been made at time t. The number of shots on target that took place is denoted by

$$n_2 = \sum_{t=1}^{n} \mathbb{I}(T_t) = \sum_{t=1}^{n} Y_{2t} \tag{3.10}$$

and uses, as its response variable,

$$Y_{3t} = \begin{cases} 1 & \text{if a goal occurs in open play at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

where $t = 1, 2, \ldots, n$. From this, the probability of G_t, the event that a goal occurs in open play at time t, is

$$\begin{aligned} P_3(t) &= \mathbb{P}(G_t | I_{1t} \cap I_{2t} \cap I_{3t}) \\ &= P_2(t)\mathbb{P}(G_t | T_t \cap I_{3t}) \\ &= \mathbb{P}(S_t | I_{1t})\mathbb{P}(T_t | S_t \cap I_{2t})\mathbb{P}(G_t | T_t \cap I_{3t}), \end{aligned}$$

with I_{3t} now defined as

$$I_{3t} = \{x_{3jt} \ : \ 1 \leq j \leq p_3\}.$$

Additionally, the conditional probability of a goal occurring at time t given that a shot on target occurred at time t is modelled as,

$$\mathbb{P}(G_t | T_t \cap I_{3t}) = g\left(\beta_{30} + \sum_{j=1}^{p_3} \beta_{3j} x_{3jt}\right),$$

with the function $g(x)$ established in (3.3). The terms $\beta_{31}, \beta_{32}, \ldots, \beta_{3p_3}$ are the coefficients associated with each of the p_3 covariates in our model and β_{30} is the intercept for the third model. It is worth noting, again, that the covariates used here are not necessarily the same as the ones used in the first two stages of sequential logistic regression, and that p_3 can be different from p_1 and p_2. It is also important to note that the goals are modelled conditionally on a shot on target occurring at time t; hence, estimation is based on n_2 frames, with n_2 defined in (3.10). Using the above, we can define the Expected Goal Value

(EGV) metric for team A in a single game as,

$$\text{EGV}_A = \sum_{t \in C_A} \widehat{P}_3(t) = \sum_{t \in C_A} \widehat{\mathbb{P}}(S_t | I_{1t}) \widehat{\mathbb{P}}(T_t | S_t \cap I_{2t}) \widehat{\mathbb{P}}(G_t | T_t \cap I_{3t}),$$

where $C_A \subset \{1, 2, \ldots, n\}$ is the set of times where team A has possession of the ball and $\widehat{\mathbb{P}}(S_t | I_{1t})$, $\widehat{\mathbb{P}}(T_t | S_t \cap I_{2t})$ and $\widehat{\mathbb{P}}(G_t | T_t \cap I_{3t})$ correspond to the fitted values from our sequential logistic regression models. Then, we can define the EGV metric for team B in the same game in a similar manner, and the combined EGV metric for that game as

$$\text{EGV} = \text{EGV}_A + \text{EGV}_B.$$

Similar to ESV and ETV, it is vital to note that our EGV metrics depend on the events in the game leading up to each time t. As with previous metrics, this metric is not equivalent to the expected number of goals that could be calculated for a game before it begins. Finally, to avoid cumbersome notation, we will proceed with the understanding that when we refer to the EGV metric, we are referring to the EGV metric of one team in one game; this should be clear based on context.

3.2.2 Fitted Values and Expected Value Metrics

In general, estimating the probability of a shot on target or a goal occurring at some time t is more challenging than estimating the probability of a shot occurring at that time. In particular, goals are challenging to predict because

of their rarity. Soccer is generally a low-scoring game, resulting in less available data to study the impact of covariates on the final results. In this section, we examine the results produced by our sequential logistic regression models when fitted to the UEFA 2022 data to gain insight into shots on target and goals.

In sequential logistic regression, the second stage we implemented estimated the probability of a shot on target conditionally on a shot being taken and produced the ETV metric. To assess the fit of this model, we go through various sanity checks. First, Figure 3.3 compares the ETV produced by our model and the observed number of open play shots on target. Most of the points in this figure are appropriately scattered about the $y = x$ line. An interesting observation can be made about the point located in the top right corner of Figure 3.3, which represents Spain's performance against Finland. In this game, Spain had possession of the ball for 65 minutes, whereas Finland only had possession for 28 minutes, demonstrating that Spain dominated the play and suggesting many goal-scoring opportunities could have occurred. Our ETV, for this game, seems to have captured Spain's domination of the game, even though it is not common to observe this many shots on target in a game.

Next, similar to our shots model, we produced a table that displays how often the observed frequency of shots on target falls within the expected probability range. In Table 3.2, N_i represents the number of fitted values that fall between $\hat{p}_{i,\min}$ and $\hat{p}_{i,\max}$ and s_i denotes the number of times a shot on target occurred while the fitted probability was within the specified range. Finally, f_i is calculated using $\frac{s_i}{N_i}$. If our model performs well, we expect f_i to

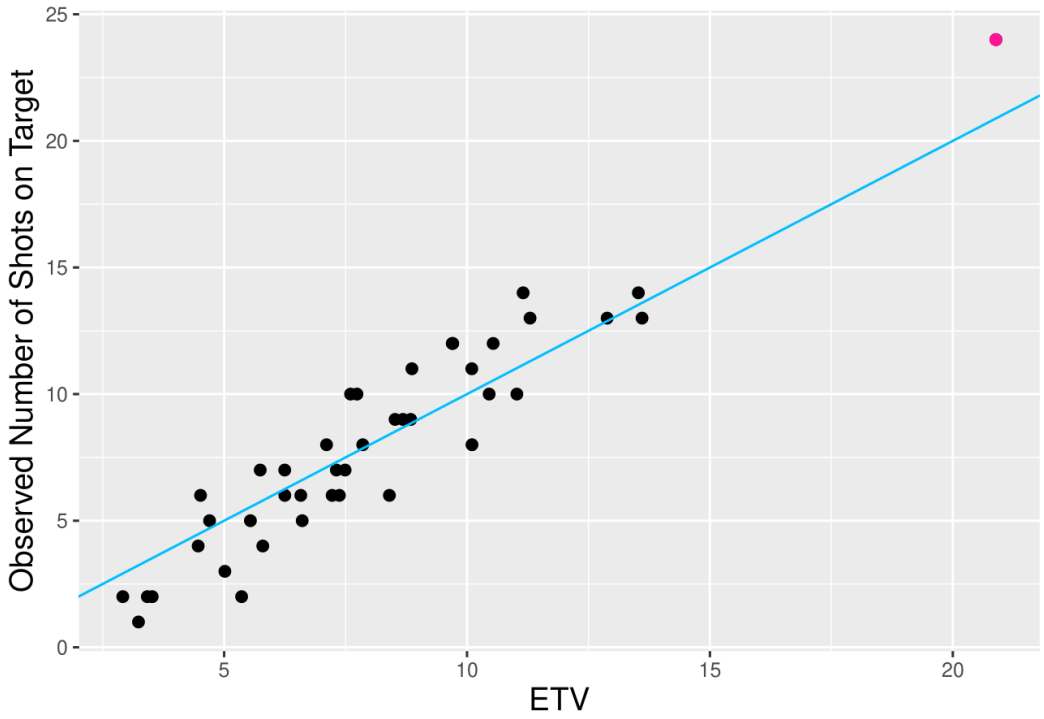


Figure 3.3: Plot comparing the Expected Shot on Target Value (ETV) and observed number of open play shots on target for each game and team. The blue line represents the $y = x$ line, and each point represents one team's performance in a game. For instance, the pink point (●) represents Spain's performance against Finland.

fall within or near the specified range of probabilities. Examining this table, it is evident that as N_i decreases, our model struggles to produce observed frequencies within the specified range of fitted probabilities. Beyond the sixth row, our observed frequencies no longer fall within the expected probabilities, which is likely caused by the reduced amount of available data. If we had access to more data, we would, in turn, have more data associated with good shooting chances and would expect to see better observed frequencies. This

being said, this table and the figure above indicate that our model performs well in estimating the probability of a shot on target when N_i is large, and that our model produces usable ETV metrics.

One of the main focuses in this section is to verify that the sequential logistic regression model estimates the probability of a shot on target reasonably well. With that said, it remains interesting to verify the fit of the model used to estimate the probability of a shot on target, given that a shot occurred. Similar to the previous models, we constructed a table that displays how often the observed frequency of shots on target given a shot falls within the expected probability range. Table 3.3 demonstrates that our model to estimate the probability of a shot on target given that a goal occurred is performing reasonably well. In particular, beyond the third row of the table, all observed frequencies are falling within or close to the specified ranges. The first three

Table 3.2: Estimated probabilities vs. observed frequency of open play shots on target using the full model.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	f_i
1	0.00	0.05	41445	68	0.0016
2	0.05	0.10	628	67	0.1067
3	0.10	0.20	510	70	0.1373
4	0.20	0.30	237	48	0.2025
5	0.30	0.40	120	38	0.3167
6	0.40	0.50	69	23	0.3333
7	0.50	0.60	29	12	0.4138
8	0.60	0.70	16	9	0.5625
9	0.70	1.00	8	4	0.5000

rows are not performing as well, but this is unsurprising as the overall number of events within these ranges is quite small, being a reflection of the fact that team do not generally take bad shots.

The third stage of sequential logistic regression estimates the probability of a goal, conditionally on there being a shot on target. We will examine the outcomes from this model and follow the same approach as we did for the previous two models.

First, Figure 3.4 displays the EGV and the observed number of goals scored across all games. It is challenging to compare this figure to that of the ETV and ESV because the scale varies significantly. For instance, shots range from zero to 34, but goals only range from zero to seven. With this in mind, we see that most of the points are appropriately scattered about the $y = x$ line. In particular, the pink-highlighted point in Figure 3.4 represents England's

Table 3.3: Estimated probabilities vs. observed frequency of open play shots on target given that a shot occurred using the full model.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	f_i
1	0.00	0.10	3	1	0.3333
2	0.10	0.20	13	1	0.0769
3	0.20	0.30	23	8	0.3478
4	0.30	0.40	60	20	0.3333
5	0.40	0.50	73	30	0.4110
6	0.50	0.60	86	51	0.5930
7	0.60	0.70	123	82	0.6667
8	0.70	0.80	92	63	0.6848
9	0.80	1.00	94	83	0.8830

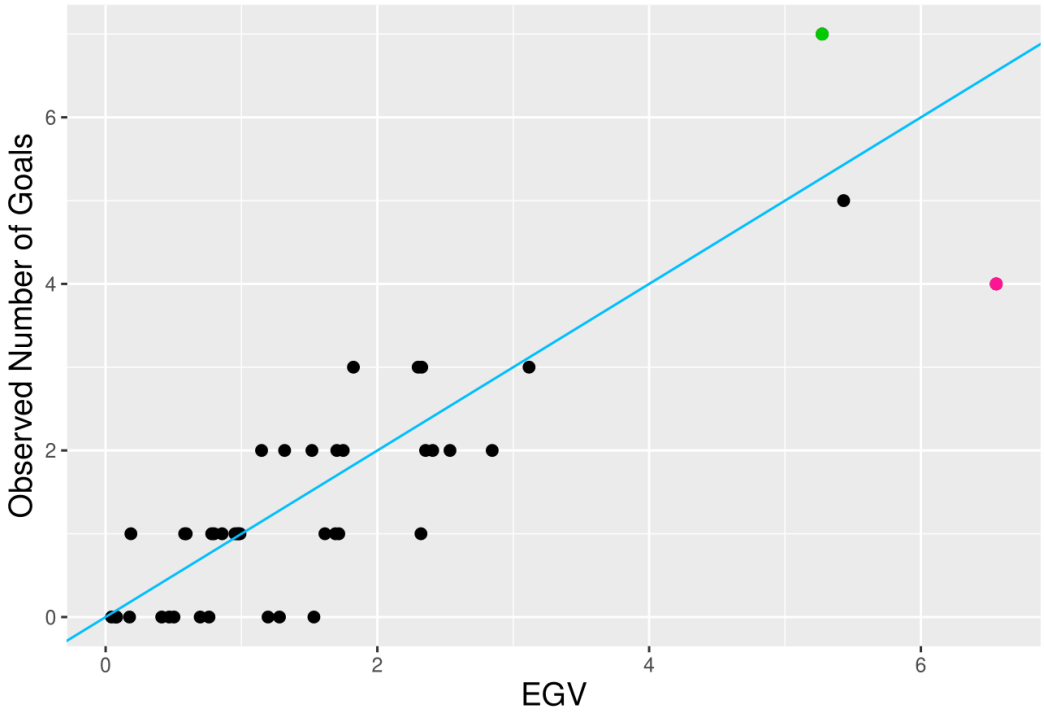


Figure 3.4: Plot comparing the Expected Goal Value (EGV) and observed number of open play goals for each game and team. The blue line represents the $y = x$ line, and each point represents one team's performance in a game. For instance, the green point (●) represents England's performance against Norway, and the pink point (●) represents England's performance against Northern Ireland.

performance against Northern Ireland. England had 28 shots and only managed to find the back of the net four times. In this game, England also had possession of the ball for approximately 53 minutes, whereas Northern Ireland only had possession for 36 minutes. Interestingly, when we compare the EGV metric and the observed number of goals, we see that the EGV is higher than the observed number of goals. This difference can be explained in a few ways.

First, England was in the lead for the entire game; therefore, it would not be surprising if they attacked the opponent's net less aggressively towards the end of the match. This could potentially lead to many chances that don't result in goals and therefore would increase the EGV, while not contributing to any observed goals. Secondly, it is also possible that England was more patient in this game than in their other games. For instance, England may have passed up good opportunities in hopes that better ones would come. If these better chances didn't eventually come, England's EGV would again increase while the observed number of goals would remain unchanged. Finally, at this point in the tournament, England had won two games and had essentially guaranteed their spot in the knockout stage. As a result, it would not be surprising if England decided to rest some of their stronger players. In turn, this could result in a smaller number of chances converted into goals. In general, our EGV metric attempts to capture scoring opportunities; therefore, given the point presented above, it is unsurprising that our EGV metric for England in this game is higher than the observed number of goals.

The green-highlighted point in Figure 3.4 represents England's performance against Norway. Comparing the observed number of shots, shots on target, goals and the amount of time in possession, it is clear that England dominated this game. Additionally, based on Figure 3.4, England scored more goals than expected in this match, but we can see that our EGV metric is still reasonably close to the observed number of open play goals. An EGV that is smaller than the observed number of goals indicates that England was able to capitalize

on opportunities that a typical team in a similar situation would not be able to. This is not a surprising result, as England was a dominant team in this tournament.

Next, Table 3.4 displays the results of our second sanity check for the EGV. As discussed previously, the model performs well if the observed frequency, denoted by f_i, falls between or near $\hat{p}_{i,\min}$ and $\hat{p}_{i,\max}$. From this table, we notice that our model behaves reasonably well in the first four rows, but as N_i decreases, our model struggles to produce observed frequencies within the specified range. This is an expected result since there is less data available to examine. As discussed previously, goals are rare events, and therefore, it is expected that large goal probability estimates will not be frequent. In general, this table and the figure above indicate that our model performs reasonably well at estimating the probability that a goal will occur when N_i is large. Evidently, with more and better quality data, we could hope to evaluate the

Table 3.4: Estimated probabilities vs. observed frequency of open play goals using the full model.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	f_i
1	0.00	0.05	42730	17	0.0004
2	0.05	0.10	142	9	0.0634
3	0.10	0.20	100	8	0.0800
4	0.20	0.30	45	14	0.3111
5	0.30	0.40	24	6	0.2500
6	0.40	0.50	17	3	0.1765
7	0.50	1.00	4	3	0.7500

performance of our metrics better. In addition, we also separately verified the fit of the model used to estimate the probability of a goal occurring given that a shot on target occurred (that is, of the third stage model). We omit the table of frequencies in this case, but confirmed that our model performs reasonably well.

In general, our dependent and sequential logistic regression models perform reasonably well at estimating the probability of a shot, a shot on target and a goal. Additionally, our expected value metrics provide valuable insights into the events throughout the game. Tables 3.5 and 3.6 display an overview of the expected value metrics produced by our models for each game and team. In these tables, we have also included the time of possession of each team (in minutes) and the observed number of shots, shots on target, and goals. These additional values allow us to compare our expected value metrics to values that describe the evolution and outcome of the games.

Our primary goal when developing expected value metrics was to capture game evolution and get a better understanding of domination. Looking at these tables, we can see that our expected value metrics are similar to the observed number of events in most cases. We also see differences between our expected value metrics and the observed number of events in some matches. These differences are a potential indication that the observed outcome of the match was not an accurate reflection of the events that took place during the game.

Table 3.5: Overview of expected value metrics and observed values for games one to 15.

Game	Team	minInPos	ESV	Shots	ETV	OnT	EGV	Goals
1	AUT	43.18	9.59	9	5.74	7	0.08	0
1	ENG	48.82	17.41	16	8.84	9	0.98	1
2	ENG	52.35	19.74	22	11.15	14	5.27	7
2	NOR	40.57	6.12	4	3.24	1	0.47	0
3	ENG	61.45	11.74	10	6.61	5	2.85	2
3	ESP	62.52	14.54	18	7.11	8	1.61	1
4	ENG	53.80	24.42	28	13.53	14	6.55	4
4	NIR	36.31	8.44	5	5.54	5	0.76	0
5	FRA	46.46	12.00	13	8.68	9	0.50	0
5	GER	48.61	9.28	8	7.37	6	1.70	2
6	FRA	53.97	16.88	16	11.03	10	5.43	5
6	ITA	42.07	11.98	13	7.73	10	1.72	1
7	FRA	61.59	26.30	30	13.60	13	1.28	0
7	NED	60.56	12.69	9	8.40	6	1.53	0
8	GER	36.07	8.92	7	4.70	5	1.32	2
8	ESP	56.91	12.80	13	6.25	7	0.41	0
9	SWE	48.14	12.96	15	8.86	11	1.15	2
9	SUI	48.30	6.39	6	4.46	4	0.59	1
10	NED	45.09	11.55	10	6.25	6	0.95	1
10	SWE	48.69	12.22	11	6.58	6	0.78	1
11	NED	56.46	15.76	17	8.52	9	1.82	3
11	POR	38.62	8.59	9	4.51	6	0.80	1
12	BEL	29.73	5.27	2	2.91	2	0.19	1
12	FRA	65.62	20.59	24	9.71	12	1.52	2
13	POR	42.54	7.83	7	3.52	2	1.20	0
13	SWE	51.45	12.06	14	7.60	10	2.33	3
14	BEL	43.13	9.46	8	5.79	4	2.32	1
14	ITA	54.38	19.68	21	12.88	13	0.70	0
15	AUT	44.77	12.49	11	7.49	7	0.04	0
15	GER	47.88	17.51	21	10.10	11	1.75	2

3.2.3 Practical Considerations

One of the downfalls of estimating the probability of a goal occurring through sequential logistic regression is that any variables used in the later models must be available at all time points. For instance, if we include a variable in our model for estimating the probability of a goal that only has a value when a shot occurs, then we can only estimate the probability of a goal at those time points. This presents an issue for our analysis since we are working with broadcast data and cannot always obtain the variables needed. For example, previous literature by Anzer and Bauer (2021) demonstrated that the distance between the goalkeeper and the goal line was an important factor in estimating the probability of scoring. Unfortunately, in our data, the goalkeeper's distance

Table 3.6: Overview of expected value metrics and observed values for games 16 to 21 (continued from Table 3.5).

Game	Team	minInPos	ESV	Shots	ETV	OnT	EGV	Goals
16	NIR	39.59	9.08	7	5.01	3	1.69	1
16	NOR	55.46	17.06	19	9.69	12	2.41	2
17	NED	46.29	14.48	17	10.54	12	2.30	3
17	SUI	52.18	13.41	12	10.45	10	0.97	1
18	POR	54.56	15.53	15	10.10	8	2.53	2
18	SUI	41.67	9.23	7	7.22	6	2.35	2
19	DEN	52.60	18.38	19	11.30	13	0.86	1
19	FIN	41.55	8.62	8	3.42	2	0.07	0
20	BEL	48.60	10.38	9	7.31	7	0.18	0
20	ISL	45.44	17.62	19	7.85	8	0.99	1
21	ESP	65.75	29.37	34	20.89	24	3.12	3
21	FIN	28.33	8.63	4	5.36	2	0.58	1

from the goal line is only available when the goalkeeper is within the bounds of the visible area or when a shot occurs. This leads to complications when using sequential logistic regression to estimate the probability of a goal at any time point. We considered implementing imputation methods to handle the missing data, but since there are a lot of frames with missing data, this was deemed too challenging to accomplish in the current context with data from only 21 games. The distance between the goalkeeper and the goal may not be eligible for inclusion in our sequential logistic regression model to produce expected value metrics. However, it is still worth investigating the findings of Anzer and Bauer (2021) in the current context.

Using the likelihood ratio goodness-of-fit test, we can determine how much a specified variable or set of variables contributes to the log-likelihood. In this context, the null hypothesis H_0 states that the specified variable or group of variables does not add anything to the model. The alternative hypothesis H_1 states that the specified variable or group of variables is important for the model. This test requires that both models have the same independent variables and that the dependent variables of one model must be a subset of the dependent variables of the other model. Any variables excluded from one of the two models are tested for their contribution. The likelihood ratio test statistic is calculated by

$$\text{LR} = 2(l_1 - l_0), \tag{3.11}$$

where l_i represents the logarithm of the maximized likelihood function for model i, where $i = 0, 1$. The test statistic in (3.11) asymptotically follows a chi-

squared distribution with degrees of freedom equal to the number of variables whose contribution will be tested for. Finally, the p-value is calculated using this test statistic and the chi-squared distribution. If the produced p-value is less than the significance level, then we reject H_0. This means there is sufficient evidence that the variable is important and should be included in our model.

Table 3.7 displays the results obtained from the likelihood ratio test when evaluating the contribution of the distance from the goalkeeper to the goal in our model for estimating a goal given that a shot on target occurs. The produced p-value denoted by $\mathbb{P}(\geq \text{LR})$ is small, suggesting that the distance between the goalkeeper and the goal line would be beneficial in our model. Unfortunately, in the data obtained from Hudl Statsbomb (2025a), the distance between the goalkeeper and the goal line is only available when a shot occurs. Using the tracking data, we can obtain the desired distance when the goalkeeper is within the visible area; however, on average, this can only be calculated in around 37% of frames per game. This restricted amount of data would result in numerous missing values when estimating the probability of

Table 3.7: Results from the likelihood ratio test when we include all the variables in the full model and are interested in testing the contribution of the distance from the goal line to the goalkeeper given all other variables are considered to be in the model.

i	df_i	l_i	$\text{df}_1 - \text{df}_0$	$\text{LR} = 2(l_1 - l_0)$	$\mathbb{P}(\geq \text{LR})$
0	58	-85.861			
1	59	-83.423	1	4.8762	0.02723

our desired events, and therefore, was excluded from our analysis. With more complete data, we could include this variable, which would likely increase the accuracy of our models.

Additionally, we would be interested in considering other variables that cannot be found in our data, especially information related to the specific players involved in the action. Firstly, a player's dominant foot influences whether a shot is on target and leads to a goal. For instance, if a right-footed player attempts a shot with their left foot, there could be a noticeable difference in shot quality. While this difference might be subtle at elite soccer levels, we anticipate a difference when comparing shots taken with a player's dominant versus non-dominant foot. Secondly, the probability of a shot on target or a goal occurring certainly depends on the shooter. For example, experienced players with extensive goal-scoring records should have higher scoring probabilities than novice athletes playing in their first international competition. We also expect to see a difference in shot probabilities based on each player's playing style. For instance, aggressive players are likely to have higher shooting probabilities than passive players. Similarly, it would also be advantageous to consider goalkeeper quality. While all top-level goalkeepers possess remarkable skills, variations in their playing styles can exist. Including a variable to reflect these differences could significantly affect goal probabilities and assist teams in adjusting to the specific style of the goalkeeper they will play against.

Chapter 4

Results

We now move to discuss various aspects of our proposed models. In particular, we discuss model reduction to determine which covariates are the most important in estimating the probability of a shot, a shot on target and a goal. Additionally, we discuss the similarities and differences when estimating the probability of these events and producing expected value metrics in men's soccer.

4.1 Reduced Model

In real-world applications, clients focus on producing accurate expected value metrics. As a result, models used in industry and, in particular, in Sports Analytics often include variables that might not be statistically significant. Although it may not be necessary in such contexts, it remains interesting to identify variables that play a bigger role in estimating the probability of a shot,

a shot on target, and a goal. Therefore, this section compares the impacts of including variables in our models and develops reduced models that include only statistically significant variables.

4.1.1 Model Selection

The primary goal of model reduction is to simplify model complexity while capturing the maximum amount of information. The first model reduction method we implement to identify important covariates for our model is investigating the p-values and standard errors. Ideally, we aim for the variables included in our model to have a low standard error and a p-value below the significance level, usually defined as 0.05. We can also verify the inclusion of variables in our models using the likelihood ratio test, which was described in detail in the previous section. Finally, we also evaluate the AIC of the models as discussed in Section 2.3 to assist in the model selection procedure.

In the following three sections, we implement the likelihood ratio test, evaluate the models' AIC, and evaluate p-values and standard errors of parameters associated with potential covariates to select a more parsimonious model for estimating the probability of a shot, a shot on target, and a goal. Additionally, we compare the expected value metrics produced by the full model with the ones obtained by the reduced model to improve our understanding of the results.

4.1.2 Shots Model

Implementing the methods discussed above, we determined that the covariates displayed in Table 4.1 are the most relevant when estimating the probability of a shot. It is important to proceed with caution when making inferences using these parameter estimates because a covariate's impact on the probability of

Table 4.1: Parameter estimates from the reduced model used to estimate the probability of a shot. See Tables 2.1, 2.2 and 2.3 for variable descriptions.

	coef	std.error	p-value	signif
(Intercept)	-1.2813	0.2788	4.30e-06	***
ar2	-1.5814	0.4497	0.00044	***
depthNotInPos.NG.IRW	-1.3110	0.6246	0.03584	*
distToAttackingNet	-0.1788	0.0083	<2.2e-16	***
midAngle	1.2274	0.2343	1.63e-07	***
largeAngle	1.6351	0.2397	9.07e-12	***
dBwBallAndNet_2to4	0.3583	0.1743	0.03981	*
dBwBallAndNet_5to7	0.8746	0.1931	5.93e-06	***
dBwBallAndNet_8Plus	1.4468	0.2426	2.48e-09	***
twoDInCone	-0.6933	0.2422	0.00421	**
threePlusDInCone	-1.2550	0.4101	0.00221	**
pressureInCone.10	0.3783	0.1904	0.04696	*
corner	-0.6465	0.1619	6.54e-05	***
counter	1.8939	0.2644	8.08e-13	***
inter_pressure	-2.1841	0.7575	0.00394	**
inter_dInCone_dist	0.1007	0.0291	0.00054	***
midfield	0.5635	0.1223	4.08e-06	***
forward	0.5263	0.1563	0.00076	***
head	2.4138	0.1920	<2.2e-16	***
longPos	-0.4482	0.2179	0.03970	*

Note: $* \; p < 0.05$; $** \;\; p < 0.01$; $*** \; p < 0.001$

a shot is measured conditionally on the other variables included in the model.

That said, the estimated coefficient for the IRW depth relative to the team without the ball is negative. This indicates that an increase in the depth of the ball will lead to a lower chance of shooting, conditionally on the other variables in the model. This is aligned with our understanding of soccer. An increase in the depth of the ball relative to the opponents implies that defenders surround the ball carrier, and is perhaps less likely to have a good shooting opportunity. Additionally, the standard error associated with the IRW depth relative to the team that does not have the ball is large. This is unsurprising and likely caused by the restricted amount of data we have access to.

We also observe that distance and angle have low standard errors and small p-values, indicating that these variables are useful for modelling shots. This outcome aligns with research conducted previously by others, including Rathke (2017). In addition, distance has a negative coefficient, implying that, conditional on the other covariates, the probability of scoring is expected to decrease as the distance to the net increases. Conversely, both indicator variables representing the medium and large angles have a positive coefficient, indicating that the chance of scoring should increase, conditionally on the other variables, when the ball is in the specified range. Comparing the two angle coefficients, we see a larger positive value associated with the large angle indicator, which is expected, as this indicates that the player is almost directly in front of the net (see Figure 2.5).

Additionally, the two indicator variables that specify if a player is a midfielder or forward have positive coefficients, low standard errors and small p-values. This suggests that the probability of a shot occurring increases when midfielders and forwards have the ball, conditionally on the other covariates in the model.

The event that initiates a sequence, the defender locations, and the body part used also influence the probability of a shot occurring. For instance, a play that originates from a counterattack has a positive coefficient, therefore implying that the probability of a shot occurring is expected to increase during plays that initiate from counterattacks, given that the effects of all other variables are already accounted for in the model. We also see a positive coefficient when a player can head the ball. Headers are rare in soccer games, and they occur most commonly when a cross ends in the 18-yard box, and an attacking player can head it. Because of this specific sequence of events involving headers, if we know that a header occurs, we expect the probability of a shot to increase significantly.

To illustrate that the `logitx` function produces results that are consistent with the literature, we conducted likelihood ratio tests for all variables. The likelihood ratio test statistic was defined in (3.11) and can be used to test the significance of specific variables in our models. Using the p-value produced by the likelihood ratio test, we discovered an issue with the `logitx` function that was discussed in Section 3.1.3 and resolved after contacting the author. This emphasizes the importance of conducting various checks when perform-

ing statistical analysis. For example, Table 4.2 displays the results from the likelihood ratio test when excluding the distance from the goal line to the ball. We see a large change in the log-likelihood and a highly significant p-value. This indicates that distance to the goal is an important variable when estimating the probability of a shot, which aligns with past research on this topic. The likelihood ratio test was performed on all variables to determine their impact on estimating the probability of a shot occurring. After reviewing various combinations using the likelihood ratio test, AIC, and examining the estimated coefficients, we concluded that the covariates displayed in Table 4.1 are the most relevant when estimating the probability of a shot.

Once we selected the covariates for the reduced model, we implemented the sanity checks discussed in Chapter 3. First, we verified how often our observed shot frequencies coincide with the expected proportions. Table 4.3 displays the number of times our fitted values landed within the specified range, denoted by N_i, and the number of times our fitted values landed within the specified range and a shot was taken, denoted by s_i, and the observed frequencies obtained

Table 4.2: Results from the likelihood ratio test when we include all the variables listed in Table 4.1, but exclude the distance between the ball carrier and the net being attacked.

i	df_i	l_i	$\text{df}_1 - \text{df}_0$	$\text{LR} = 2(l_1 - l_0)$	$\mathbb{P}(\geq \text{LR})$
0	19	-1813.8			
1	20	-1421.5	1	784.52	<2.2e-16

from the reduced model, denoted by $f_{i,R}$. Our reduced model is performing as expected when the observed frequency falls within or near $\hat{p}_{i,\min}$ and $\hat{p}_{i,\max}$. For comparison, the observed frequencies obtained from our full model in Table 3.1, denoted by $f_{i,F}$, are also included in this table. The two models perform similarly, however, the full model seems to have a slightly better fit based on the observed frequencies.

Figure 4.1 displays the ESV metric from the reduced model versus the observed number of shots. In general, if we compare this to Figure 3.1, we can see that the values produced from the reduced model are more dispersed around the $y = x$ line. This indicates that our reduced model is not capturing the same information as the full model. The reduced model does not include all the covariates available in the full model, therefore, this outcome

Table 4.3: Estimated probabilities vs. observed frequency of open play shots using the reduced model compared to the full model calculated previously.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	$f_{i,R}$	$f_{i,F}$
1	0.00	0.05	40789	74	0.0018	0.0018
2	0.05	0.10	815	107	0.1313	0.1069
3	0.10	0.20	672	106	0.1577	0.1738
4	0.20	0.30	316	67	0.2120	0.2029
5	0.30	0.40	184	47	0.2554	0.3077
6	0.40	0.50	100	35	0.3500	0.3535
7	0.50	0.60	43	13	0.3023	0.3725
8	0.60	0.70	28	17	0.6071	0.5278
9	0.70	0.80	44	36	0.8182	0.7826
10	0.80	1.00	71	65	0.9154	0.9028

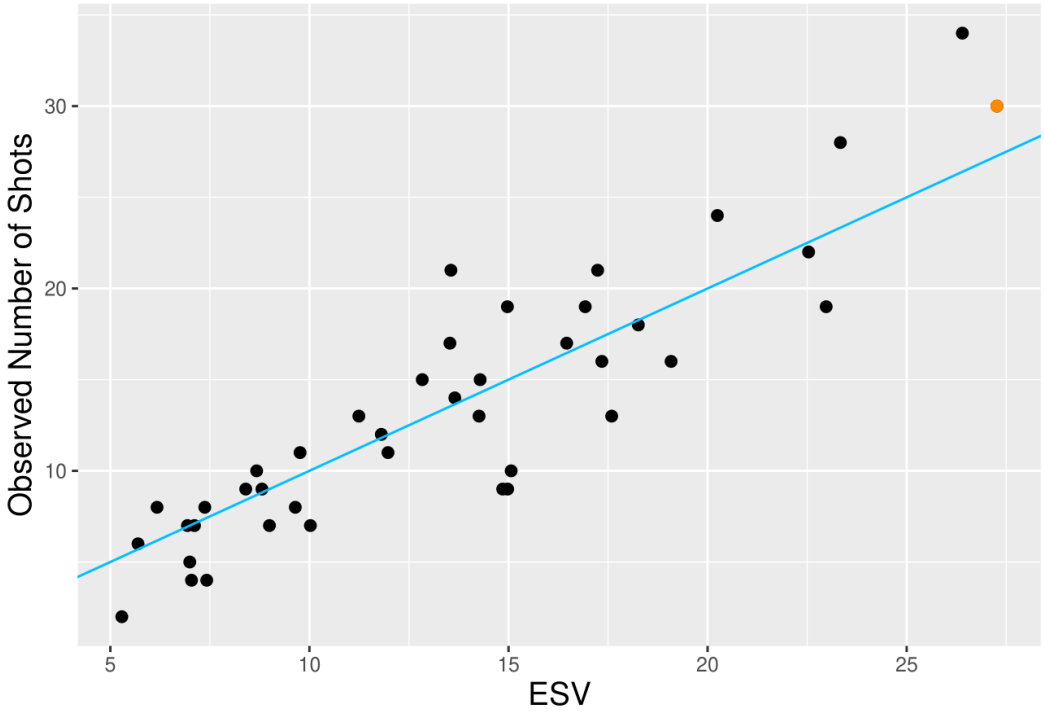


Figure 4.1: Plot comparing the Expected Shot Value (ESV) produced by the reduced model to the observed number of shots. The blue line represents the $y = x$ line, and each point represents one team's performance in a game. For instance, the orange point (●) represents France's performance against the Netherlands.

was anticipated. While these excluded covariates are not deemed statistically significant, they could still influence the ESV metric.

4.1.3 Shots On Target Model

The second level of our sequential logistic regression model focused on estimating the conditional probability of a shot on target given that a shot was

taken and producing an ETV metric. The likelihood ratio test, AIC, standard errors and p-values revealed that the covariates listed in Table 4.4 are the most important for estimating the conditional probability of a shot on target. For instance, the distance to the net being attacked has a low p-value, low standard error, and the likelihood ratio test indicates that it significantly reduces the log-likelihood of our model. Moreover, we observe that the coefficient for distance is negative, indicating that as a player gets closer to the goal and the distance decreases, the conditional probability of a shot on target occurring is expected to increase, conditionally on the other variables included in the model. We see a similar relationship with the indicator variables representing corners and headers. For these covariates, the negative coefficients imply that when a header or a corner occurs, we anticipate a decrease in the conditional

Table 4.4: Parameter estimates from the reduced model used to estimate the probability of a shot on target given that a shot occurred. See Tables 2.1, 2.2 and 2.3 for variable descriptions.

	coef	**std.error**	**p-value**	**signif**
(Intercept)	0.5758	0.4584	0.20910	
distToAttackingNet	-0.0523	0.0137	0.00013	$***$
midAngle	1.1494	0.4373	0.00857	$**$
largeAngle	1.1922	0.4352	0.00616	$**$
corner	-0.4215	0.2311	0.06814	$\bullet$
scoreDiff1Behind	0.5423	0.2496	0.02984	$*$
head	-0.9239	0.2640	0.00017	$***$
oneDInCone	-0.3341	0.1953	0.08718	$\bullet$
twoDInCone	-0.5597	0.3308	0.09063	$\bullet$

Note: $\bullet$ $p < 0.1$; $*$ $p < 0.05$; $**$ $p < 0.01$; $***$ $p < 0.001$

probability of a shot on target, conditionally on the other variables included in the model. These parameter estimates are interesting compared to the coefficients obtained from our shots model. For example, when estimating the probability of a shot, we observed that headers have a positive coefficient. It is important to recall that the produced parameter values are intended to be interpreted conditionally on the other variables in the model. With this said, there are a few potential explanations for the difference in signs between the two models. To begin, headers most often occur when a pass is made into the 18-yard box and a forward deflects it towards the net. Since this is one of the most common scenarios in which headers occur, the knowledge that an event is executed with the head makes it highly probable that the action will be a shot, explaining why the model for estimating the probability of a shot produces a positive coefficient. On the other hand, headers are much harder to control and aim than shots made with the foot. Because of this, headers do not play as much of a role in estimating the conditional probability of a shot on target, and can make it less likely for a shot to be on target. With this, we can see why the model for estimating the conditional probability of a shot on target produced a negative coefficient for the header variable.

In general, comparing the p-values and standard errors in Table 4.4 with the model used to estimate the probability of a shot in Table 4.1, we can see that this model has slightly higher standard errors and p-values. This result is unsurprising because this model uses a reduced amount of data, and shots on target are rarer than shots, making the model fitting step more challenging

and leading to inferential results that are less precise.

Similar to the previous model, Table 4.5 allows us to compare fitted probabilities with observed frequencies. Recall that N_i is the number of fitted probabilities that fall within the specified range, and s_i counts the number of shots on target that occurred when the fitted probability was in the defined range. Additionally, $f_{i,R}$ represents the observed frequency produced when fitting the reduced model and $f_{i,F}$ represents the frequency obtained from fitting the full model. We see very similar values between the two models in the low ranges. Furthermore, as the fitted probabilities increase and the value of N_i decreases, the full model produces slightly better outcomes. This result is unsurprising as the full model captures more information.

Our main objective in this research was to develop effective expected value metrics for shots, shots on target, and goals. Therefore, we compare the ETV

Table 4.5: Estimated probabilities vs. observed frequency of an open play shot on target using the reduced model compared to the full model.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	$f_{i,R}$	$f_{i,F}$
1	0.00	0.05	41403	82	0.0020	0.0016
2	0.05	0.10	626	56	0.0895	0.1067
3	0.10	0.20	545	74	0.1358	0.1373
4	0.20	0.30	264	56	0.2121	0.2025
5	0.30	0.40	129	30	0.2326	0.3167
6	0.40	0.50	65	27	0.4154	0.3333
7	0.50	0.60	23	10	0.4348	0.4138
8	0.60	1.00	7	4	0.5714	0.5417

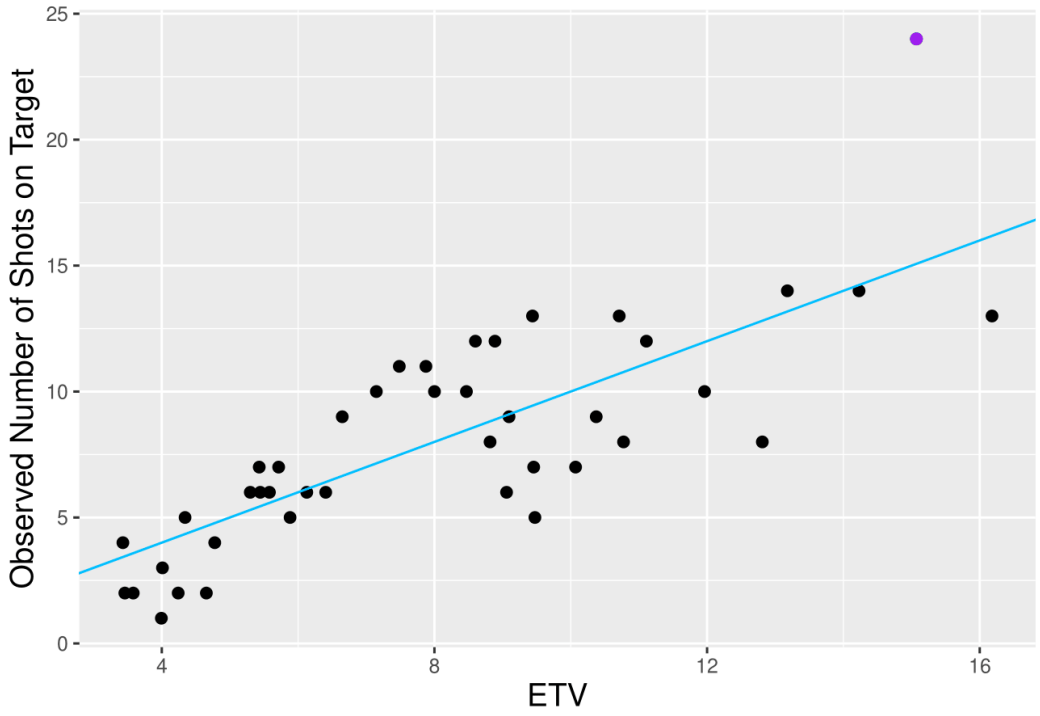


Figure 4.2: Plot comparing the Expected Shot on Target Value (ETV) produced by the reduced model and observed number of shots on target. The blue line represents the $y = x$ line, and each point represents one team's performance in a game. For instance, the purple point (•) represents Spain's performance against Finland.

metric generated by the full model, perhaps more likely to be adopted in industry, with the model simplified through statistical methods. Figure 4.2 illustrates the ETV metric from the reduced model alongside the observed number of shots on target. Compared to Figure 3.3, there is considerably more variation around the $y = x$ line. This discrepancy is anticipated, given the reduced number of covariates informing our model.

In particular, the purple-highlighted point represents Spain's performance against Finland. As discussed in the previous chapter, Spain dominated this game based on possession and scoring chances. Our full model provided an ETV close to the observed number of shots on target, but in Figure 4.2, our reduced model is not performing as well. This indicates that our reduced model is not capturing the expected number of shots on target as well as the full model in this case, which is likely caused by the reduced number of covariates.

Overall, we have shown that the variables displayed in Table 4.4 are the most important for estimating the probability of a shot on target. This model does not perform as well as the full model, but it still provides vital insights about the covariates that have the greatest impact. With this, we proceed to the final model in our sequential logistic regression to estimate the probability of a goal occurring given a shot on target is observed, using a reduced model.

4.1.4 Goals Model

Finally, we proceed similarly to the previous two sections by discussing the differences between the full and reduced models for estimating the probability of a goal given a shot on target has occurred. Table 4.6 displays the variables retained in the reduced model, their coefficients, standard errors and p-values.

Similar to the previous models, the coefficient for the distance between the ball carrier and the net is negative and important for estimating the probability

of a goal given that a shot on target occurred. Another key variable for estimating the conditional probability of a goal is the distance between the ball carrier and the nearest defender outside the cone. This variable has a positive coefficient, a low p-value, and a low standard error. The positive coefficient associated with this covariate indicates that, as the defender closes in on the ball carrier, reducing the distance between them, we expect the probability of a goal to decrease, conditionally on the other variables in the model. On the other hand, as the distance between the nearest defender outside the cone and the ball increases, we expect the conditional probability of a goal to increase, conditionally on the other variables in the model. This is unsurprising since the ball carrier is less likely to get a clear shot at the goal when under pressure and, the more space a player has, the more likely they will be able to set themselves up to take a good shot.

Additionally, the IRW depth of the ball, relative to the non-possession

Table 4.6: Parameter estimates from the reduced model used to estimate the probability of a goal given that a shot on target occurred. See Tables 2.1, 2.2 and 2.3 for variable descriptions.

	coef	**std.error**	**p-value**	**signif**
(Intercept)	-0.3094	0.6167	0.61609	
distToAttackingNet	-0.1824	0.0289	2.91e-10	$***$
closestOut	0.3089	0.0885	0.00049	$***$
inter_pressure	-4.6475	1.6943	0.00609	$**$
depthNotInPos.IRW	3.2046	1.7017	0.05967	$\bullet$

Note: $\bullet$ $p < 0.1$ $*$ $p < 0.05$; $**$ $p < 0.01$; $***$ $p < 0.001$

team, has a low p-value and a high standard error. Although a high standard error is undesirable, the ratio of the coefficient to the standard error remains comparable to other covariates. This indicates that the IRW depth remains significant for our model. In particular, the coefficient produced for the IRW depth of the ball relative to the non-possession team is positive. This implies that as the depth of the ball increases, meaning the ball carrier is becoming more central to the opposing team, the probability of scoring is expected to increase, conditionally on the other variables in the model. To explain this concept, it is important to consider how soccer matches usually evolve. Typically, the defending team will become more compact while the offensive team tries to attack their net. This results in the attacking team often having space and freedom to move around the outside of the pitch, but as they move towards the net, that space becomes limited. As a result, when the ball carrier moves towards the net, the IRW depth associated with the ball is likely to increase. The only instance where we would expect the ball carrier to have a low depth at the time of the shot is if the player is on a breakaway and behind the final defensive line.

Lastly, the pressure indicator variable, which measures the pressure exerted within the cone up to ten yards from the ball carrier, produced a negative coefficient. This outcome indicates that an increase in the pressure applied inside the cone is expected to reduce the probability of a goal occurring, conditionally on the other variables in the model. We also note that the p-value is below the established significance level, and the standard error is small, demonstrating

the importance of this variable.

Similar to the previous two models, we used Table 4.7 to assess our model results. This table also includes the observed frequency found using our full model from Table 3.4 for comparison. There are no major differences in the frequencies when N_i is large, however, we see that the full model generally has a better fit based on the observed frequencies.

Next, we compared the EGV metrics produced by the reduced and full models. Figure 4.3 displays the EGV produced by the reduced model compared to the observed number of goals scored. Compared to Figure 3.4, we see a significant increase in the variation around the $y = x$ line. This result is again expected because the reduced model accounts for fewer covariates than the full model. In particular, the orange highlighted point in Figure 4.3 represents France's performance against the Netherlands. This point falls below the $y = x$ line, indicating that France scored fewer goals than expected. This result is similar to the one discussed when implementing the full model, but the effect

Table 4.7: Estimated probabilities vs. observed frequency of open play goals using the reduced model compared to the full model calculated previously.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	$f_{i,R}$	$f_{i,F}$
1	0.00	0.05	42667	20	0.0005	0.0004
2	0.05	0.10	211	12	0.0569	0.0634
3	0.10	0.20	137	13	0.0949	0.0800
4	0.20	0.30	38	13	0.3421	0.3111
5	0.30	0.40	8	2	0.2500	0.2500
6	0.40	1.00	1	0	0.0000	0.2857

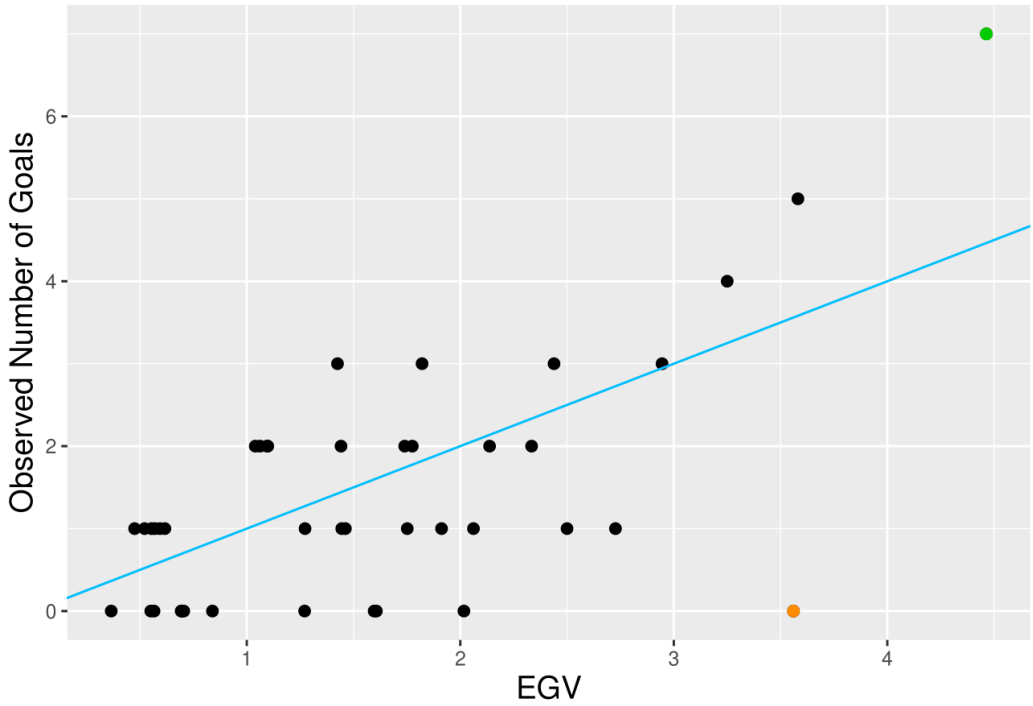


Figure 4.3: Plot comparing the Expected Goals Value (EGV) produced by the reduced model and observed number of shots on target. The blue line represents the $y = x$ line, and each point represents one team's performance in a game. For instance, the green point (●) represents England's performance against Norway, and the orange point (●) represents France's performance against the Netherlands.

is accentuated here.

Additionally, the green-highlighted point represents England's performance against Norway. The observed number of goals England scored in this game exceeds the EGV. England achieved an eight-nothing victory in this match, which is extremely rare in soccer. This kind of lop-sided match is expected in tournaments such as the UEFA European Women's Championship because

qualified teams often have a wide range of skill sets. This is especially true when studying women's soccer, as the tournament qualification is, in general, less competitive than the men's. These differences can lead to high-scoring games whose outcomes our reduced models may struggle to capture. One explanation for this particular game is that England dominated the game and capitalized on more opportunities than a typical team would. This result was not unexpected, as England went on to win this tournament.

Overall, the variables selected as meaningful for estimating the conditional probability of a goal are the distance to the goal, the distance to the closest defender outside the cone, whether there is pressure in the cone within ten yards, and the IRW depth of the ball relative to the opposing team. These variables are meaningful because they significantly reduce the log-likelihood according to the likelihood ratio test. The effect of these covariates and associated interpretations have been discussed above.

Finally, we compared the EGV metric between the full and reduced models to check for stability in our approach. The left-hand side of Figure 4.4 displays the relationship between the ESV metric produced by the full and the reduced model. The plot shows a strong positive correlation, which indicates that our ESV metric is reasonably stable. In contrast, the right-hand side of Figure 4.4 displays the relationship between the EGV metric produced by the full model and the reduced model. Compared to the ESV metric, the EGV metric has more variation about the $y = x$ line, which indicates that our EGV metric is less stable. There is still a positive relationship between the full and

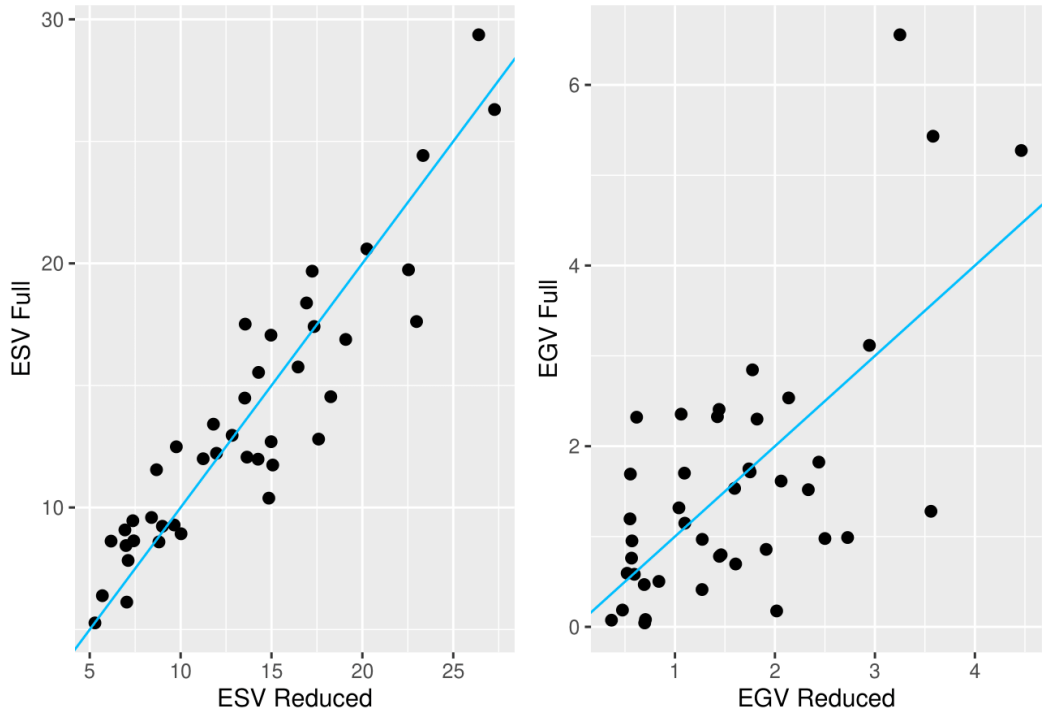


Figure 4.4: **(Right)** Plot comparing the Expected Shot Value (ESV) metric between the full and reduced models for women's soccer. **(Left)** Plot comparing the Expected Goal Value (EGV) metric between the full and reduced models.

reduced model EGV metrics; however, the two models produce different results in some cases. Additionally, we found that the reduced model produced EGV metrics further from the observed number of goals than the full model. Overall, the inconsistency between the full and reduced model EGV metrics is not surprising, as we are developing expected value metrics with a relatively small data set. Moreover, the increased variation in the EGV metrics is likely caused by the rarity of goals, which makes it quite challenging to estimate the

probability of a goal occurring.

In general, identifying the importance of covariates is more challenging in the shots on target and goals models. This is likely because we have less data at these stages. We also found that the expected value metrics associated with the reduced models had more variation than the ones associated with the full models. This difference was anticipated and is likely caused by the smaller number of covariates incorporated in the reduced models.

4.2 Expected Value Metrics in Men's Soccer

Most Sports Analytics research focuses on men's sports, therefore, it is worth testing our methods on men's data. The data sets selected for this comparison were the 2020 and 2024 UEFA European Championships. These data sets were also obtained from Hudl Statsbomb (2025b). These tournaments are structured similarly to the UEFA European Women's Championship Tournament. As a result, we expect our analysis to produce similar results, although previous work by Bransen and Davis (2021) and Bradley et al. (2014) has suggested that important structural differences do exist between men's and women's soccer.

Using the same model as we implemented with the women's data, we constructed expected value metrics associated with the number of shots, shots on target and goals using the men's data. One interesting difference between the men's and women's data is the variability in the observed number of shots,

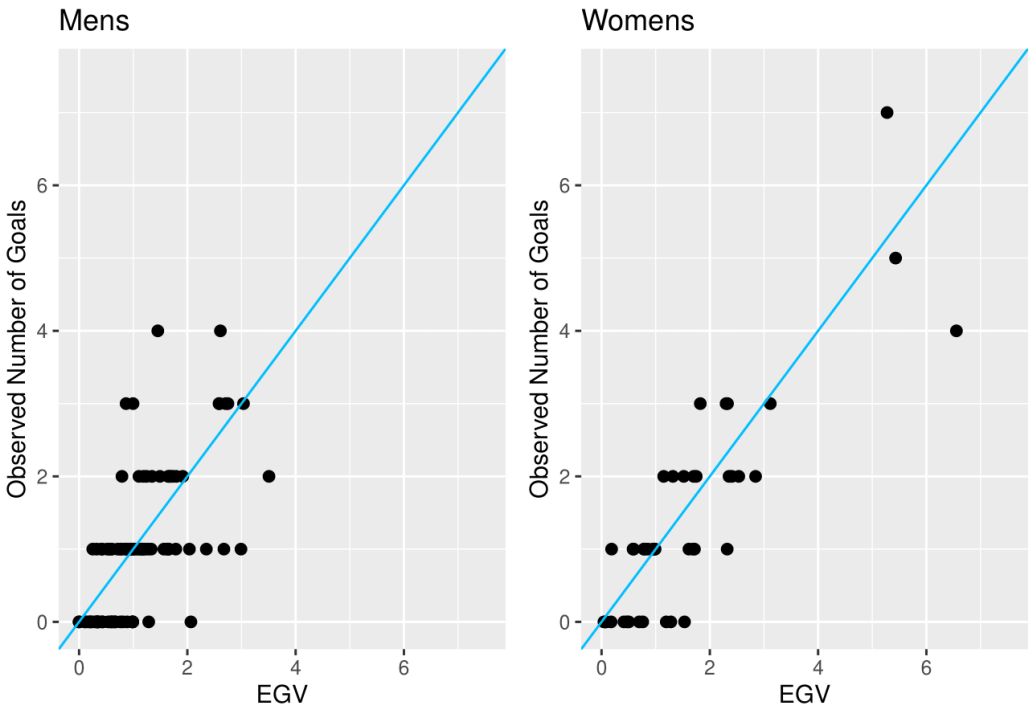


Figure 4.5: Plot comparing the Expected Goals Value (EGV) and the observed number of goals produced by the full models for men's and women's data. The blue lines represent the $y = x$ line, and each point represents one team's performance in a game.

shots on target and goals. Figure 4.5 displays the EGV and the observed number of goals scored in the men's and women's tournaments. Comparing these two plots, we see that all of the games in the men's tournament have four or fewer goals per team. Conversely, the women's plot displays that most of the games had fewer than three goals, except for three of the games. One explanation for this difference is that there are, in general, more men's teams that are competitive at the UEFA championships. This implies that the games have

the potential to be closer in score and more challenging to dominate. Conversely, the women's tournaments often have more variation in the skill level of teams. As a result, it is more common to see games with more goals in the women's tournament compared to the men's.

Table 4.8 displays the observed frequency of goals and the expected probability ranges. Our model performs reasonably well when the observed frequency of goals, denoted by f_i, falls within or near the specified range of probabilities. Looking at Table 4.8, we see that beyond the fourth row, we have observed frequencies outside the desired range. This is not surprising, since there are fewer events in these specified ranges. With this, we can conclude that our model performs reasonably well at estimating the probability of a goal occurring. This also suggests that, with more data, we could likely also see improved fits with the women's data.

Table 4.8: Estimated probabilities vs. observed frequency of open play goals using the full model in the Men's UEFA European Football 2020 and 2024 Championships.

i	$\hat{p}_{i,\min}$	$\hat{p}_{i,\max}$	N_i	s_i	f_i
1	0.00	0.05	90502	36	0.0004
2	0.05	0.10	249	13	0.0522
3	0.10	0.20	189	24	0.1270
4	0.20	0.30	42	9	0.2143
5	0.30	0.40	15	4	0.2667
6	0.40	0.50	6	2	0.3333
7	0.50	1.00	6	4	0.6667

Using the full models, we produced expected value metrics associated with shots, shots on target and goals in the Men's 2020 and 2024 UEFA European Championship. We omit the full overview of the observed events and expected value metrics in this case, as the produced figures and tables are similar to those obtained previously for the women. We would, however, like to highlight an interesting example. Table 4.9 displays a game between Spain and Switzerland in the Men's 2020 UEFA European Championship. In this match, Spain dominated over Switzerland, having possession of the ball for the majority of the game and having numerous scoring chances. Spain scored just eight minutes into this match and was in the lead till the 68^{th} minute, when a mistake by the defenders led to a tying goal for Switzerland. This goal pushed the game into overtime and eventually penalty shots, where Spain ultimately won. This example is important for us to consider, as the observed number of goals does not provide an accurate representation of the events throughout the game, but our EGV does. In particular, we see that Spain has an EGV of 2.07 in this game, and Switzerland only has an EGV of 0.76. This highlights the usefulness of our metrics, as they provide a more informative summary of

Table 4.9: Interesting example of expected value metrics and observed values in the Men's UEFA European Football Championships.

Team	**minInPos**	**ESV**	**Shots**	**ETV**	**OnT**	**EGV**	**Goals**
SPN	84.08	21.13	25	15.52	20	2.07	1
SUI	36.88	11.78	8	6.34	4	0.76	1

how the game progressed than the observed number of goals.

As with the women's data, we were interested in determining what covariates are statistically significant when estimating the probability of our events of interest. Table 4.10 displays the parameter estimates produced by the model used to estimate the probability of a goal given that a shot on target occurred. In this table, the distance from the attacking net and the angle indicator representing shots in front of the net have a large impact. We also see that the depth of the ball and sequences that are initiated by corner kicks can influence the conditional probability of a goal occurring. Finally, when there are five to seven players between the ball and the net, we expect the conditional probability of a goal occurring to decrease, conditionally on the other variables in the model.

Overall, our models also perform well when estimating the probability of

Table 4.10: Parameter estimates from the reduced model used to estimate the probability of a goal given that a shot on target occurred in the Men's UEFA European Football Championship. See Tables 2.1, 2.2 and 2.3 for variable descriptions.

	coef	**std.error**	**p-value**	**signif**
(Intercept)	0.4233	0.4435	0.33984	
distToAttackingNet	-0.1188	0.0188	2.76e-10	$***$
largeAngle	1.0225	0.2695	0.00015	$***$
corner	-0.8486	0.3973	0.03268	$*$
depthNotInPos.NG.IRW	-2.6081	1.3298	0.04984	$*$
dBwBallAndNet_5to7	-0.4746	0.2846	0.09536	$\bullet$

Note: $\bullet$ $p < 0.1$; $*$ $p < 0.05$; $**$ $p < 0.01$; $***$ $p < 0.001$

events in men's soccer. As noted in previous literature, we found that covariates that are important in estimating the probability of events in men's soccer are not necessarily the same as those deemed important in women's soccer. Additionally, the emphasis on each of the covariates differs. For instance, the distance to the attacking net, player position and number of players between the ball and the net greatly impact the probability of a shot in both men's and women's soccer. In contrast, we found a reduced emphasis on the possession length, the number of defenders in the cone and the shot angle in men's soccer compared to women's. Additionally, in the model used to estimate the probability of a goal, we found that corner kicks, shots taken from large angles, and times when there are five to seven defenders between the ball and the goal all have a greater impact in men's soccer compared to women's. Further, the impact of distance on the probability of a goal occurring is similar in both models. These results align with those produced by Bransen and Davis (2021). Finally, we also developed expected value metrics for our men's data and discovered that our metrics indeed captured the scoring opportunities produced by teams. We also confirmed that our methodology can be applied in different contexts and provides interpretable results.

Chapter 5

Discussion

In this last chapter, we begin by discussing the challenges associated with the data used for this research. Next, we review applications of expected value metrics in Sports Analytics. Finally, we discuss goals for future research in this area.

5.1 Data Challenges

In Statistics, it is well-known that our results are only as good as our data. This concept is felt throughout this research and created challenges for our analysis.

One major challenge with the Hudl StatsBomb data is that it was collected using broadcast footage. This restricts our data to the players within the visible area, resulting in missing information that likely has an impact on our analysis. For example, our IRW depth variable measures the depth of the ball

location relative to the players in the visible area instead of all players. We developed our models with this restriction in mind, but we would expect the impact of our covariates to change if we had access to all player locations at all times. In addition, working with broadcast data also restricted our selection of covariates to variables with a small number of missing values. For instance, including covariates such as the distance between the goalkeeper and the net would benefit our analysis. Unfortunately, since the goalkeeper is often outside the visible area, there are currently too many missing values to incorporate this variable into our models.

Secondly, the freeze frame data from Hudl Statsbomb did not include player identification for all the visible players. The lack of identification limited our analysis to team strategy and made any individual player analysis impossible. If we had access to data with player identification, it would be interesting to track individual players and would perhaps allow us to discover player-specific trends and/or actions that lead to goal-scoring opportunities. For example, it is common to see players move in and out of high-density areas to make plays. Using data depth, we can potentially measure this movement and increase our understanding of how these plays impact goal-scoring opportunities.

Finally, the data used in this research provides positional information at the time of each event. Although this produces sufficient detail to perform our analysis, it would be beneficial to have more frequently collected positional data. More frequent data would allow for a more accurate analysis of the trends in the lead-up to a shot. For example, when reviewing the depth

of the ball during the lead-up to a shot, we found very sudden and drastic changes in depth values. Obtaining more frequent event data would produce smoother depth curves and improve our ability to measure the ball's path as it approaches the goal.

In general, the Hudl Statsbomb data provides sufficient information to conduct a statistical analysis. However, a much more detailed analysis could be conducted, and our results would potentially be improved if we had access to positional data for all players, had player identification information and had more frequently collected data. With these restrictions in mind, this research accomplished our main goal of developing expected value metrics associated with the number of shots, shots on target and goals in women's soccer matches. In the following section, we will proceed with a discussion of some applications of this research.

5.2 Applications

The applications of this research are extensive. In particular, expected value metrics can be used in industry and future research to increase our understanding of the dynamics of soccer matches. Similar to other research in Sports Analytics, this research can improve decision-making before, during and after matches to optimize team and, eventually, player performance.

To begin, our research can assist in identifying team strategies and help coaches and players prepare for different types of opponents. For example, we

can potentially help teams determine which plays result in a higher chance of scoring for themselves and their opponent. With this, coaches can better prepare their teams for competition. Furthermore, soccer is a very low-scoring game with most high-level soccer games ending with fewer than four goals and sometimes with no goals. In these situations, analyzing games based on observed events is challenging. Consequently, new methods of quantifying performance in games are required. Our developed expected value metrics fill this research gap by analyzing opportunities over observed events and by adapting to the events happening throughout the game. This adjustment creates a complete picture of how games evolved, even when no goals occur. With this, our expected value metrics can provide coaches and players with a more in-depth understanding of what actions lead to good scoring opportunities. This information can then be used to optimize team and player performance.

In addition, our research has the potential to assist in determining what type of strengths to look for when scouting new players. For instance, coaches can gain insights on which players to scout relative to their current ones and how new players could potentially impact the team. This will result in improved scouting and could drastically affect the team's performance in the long run. Additionally, if our data included player identification, we could potentially evaluate the contributions of each player to our expected value metrics. For instance, we could decompose our expected value metrics into a sum of player-specific terms. This could assist in streamlining the process of trading players and helping organizations make informed decisions.

Our research will also have vast impacts outside of team performance. For example, Mead et al. (2023) showed that expected goal statistics are better at predicting team performance than traditional metrics, including the number of shots and goals. For instance, tracking a team's expected goals metric throughout a season is a better predictor of how well a team performs than the observed number of goals or the overall game outcomes. With this improved understanding of team performance, the sports betting industry will potentially see improved game outcome predictions.

Additionally, this work may also influence the video game industry as it can help develop more realistic game simulations. For instance, incorporating our estimated probabilities and expected value metrics could result in more realistic simulated team and individual player behaviours throughout the game. Overall, this could improve user experience and likely improve the sports video game industry.

In recent years, we have seen increased visibility in women's sports. In particular, as the Professional Women's Hockey League, National Soccer League in the United States and the Northern Super League in Canada expand, we will hopefully see improved data accessibility that can be used for research in women's sports. In general, our goal in this research was to expand the existing knowledge and encourage others to pursue research in women's sports. With more interest in women's sports-related research, the accessibility to data and resources will grow significantly.

5.3 Future Work

Building on the work done in this thesis, our future research will focus on understanding actions and strategies that are hard to measure in soccer and other sports. For instance, quantifying a team's defensive ability has not been explored in detail, but it may reveal important characteristics of soccer matches. One idea that is promising is to measure defensive ability using our expected value metrics. Studying the impacts of events on the number of conceded goals can inform coaches and players how frequently they allow opponents into goal-scoring positions. In turn, this will help teams adjust their defensive strategies to minimize such opportunities.

Another concept in Sports Analytics that is challenging to quantify is the amount of pressure applied to an individual player. A commonly used approach to quantifying pressure in soccer involves a binary indicator variable that accounts for whether the ball carrier was under pressure. One downfall with this approach is that the direction and distance of the pressure are not accounted for. Another approach to capture pressure is by using the distance between the ball carrier and the player applying pressure. But again, this fails to capture the direction from which the pressure is applied. To fill this gap, Andrienko et al. (2017) proposed an approach to capture the pressure applied to the ball using the direction that the pressure comes from. They considered that pressure can be applied from multiple sources simultaneously, and that this pressure depends on the distance between the presser and the ball carrier

and the direction relative to the net being attacked. Furthermore, Gu et al. (2024) proposed a novel approach using the ball carrier's body position and current pitch control. They found that incorporating three-dimensional body position improved their model's ability to quantify the amount of pressure a player is under. Comparing our pressure interaction term with previously implemented techniques, it is clear that our pressure term is a simplified version. Our covariate accounts for the pressure direction by restricting our analysis to the defenders in the cone and includes the distance to that player via our closestIn variable. It would be interesting to implement a more complex pressure term in future research, as incorporating player speed, direction, body positioning and pitch control would likely improve our model and produce better expected value metrics.

Another interesting extension of our current work involves calculating the expected number of goals given that a shot occurred. We refer to this as the Expected Goal Upon Shot Taken (EGUST) metric. We can calculate the total EGUST in a game for team A using,

$$\begin{aligned}\text{Total EGUST}_A &= \sum_{t \in \{t:\mathbb{I}(S_t)=1\} \cap C_A} \widehat{\mathbb{P}}(G_t | S_t \cap I_{2t} \cap I_{3t}) \\ &= \sum_{t \in \{t:\mathbb{I}(S_t)=1\} \cap C_A} \widehat{\mathbb{P}}(G_t | T_t \cap I_{3t}) \widehat{\mathbb{P}}(T_t | S_t \cap I_{2t})\end{aligned}$$

where $\widehat{\mathbb{P}}(T_t | S_t \cap I_{2t})$ and $\widehat{\mathbb{P}}(G_t | T_t \cap I_{3t})$ correspond to the fitted values from our sequential logistic regression models. The total EGUST for team B can then be calculated similarly, and the combined total EGUST for both teams

is found using,

$$\text{Total EGUST} = \text{Total EGUST}_A + \text{Total EGUST}_B. \tag{5.1}$$

This metric differs from our previously developed EGV metric because Total EGUST only incorporates the fitted probabilities from the shots on target and goals models at the time of a shot and leaves out all opportunities where a shot was not taken. Conversely, the EGV calculation involves the probability of a goal being scored at all time points. The average EGUST based on all shots can then be calculated using

$$\text{Average EGUST} = \frac{\text{Total EGUST}}{n_1},$$

where total EGUST is defined in (5.1) and n_1 corresponds to the total number of shots taken over a game. The average EGUST can potentially measure the quality of the shots taken by a team by scaling the total EGUST for that team relative to the total number of shots. For instance, the average EGUST for team A in a given match is calculated by

$$\text{Average EGUST}_A = \frac{\text{Total EGUST}_A}{n_A},$$

where n_A is the number of shots taken by team A over that match. This metric may indicate how likely it is for shots to end up as goals and could be calculated based on a specific game, tournament, or league. In general, this new metric then measures the quality of shots taken by a team during a game

and could potentially provide new insights about the expected number of goals given that a shot occurred and increase our understanding of soccer matches.

Additionally, we would like to expand the concepts used in this thesis to more complex objects. For example, this research used the depth of a point within a sample of points to extract positional information, but we would like to extend this and consider using the depth of more complex objects. One promising idea is to examine the depth of a set of points among many similar sets of points. For example, in soccer, we could study the depth of one team's positioning relative to the opposite team at a specific time point, or perhaps relative to the positions of that same team at other time points in the match. This allows us to identify trends that lead to goal-scoring opportunities and study the differences in playing styles, but requires us to create new depth measures and concepts.

Another promising idea is to consider player trajectories as they move on the field, using functional data approaches. In particular, we could incorporate player/ball trajectories as covariates in a functional version of logistic regression. We could also look at the depth of a trajectory among many other trajectories. Specifically, we could use such a depth measure to follow the path of the ball leading up to a shot for multiple shots, and then we could compare the trajectories by studying the depth of the curves associated with these trajectories using a functional version of data depth. This research also applies to sports such as football and hockey. For instance, in American and Canadian football, we could analyze receiver routes that lead to completed

passes by comparing numerous paths.

One major challenge with publicly available tracking data is that the availability of player coordinates often depends on the area visible on broadcasts. As a result, developing techniques that account for the missing players seems like a logical and potentially impactful next step. For instance, this research implements modelling of soccer matches with the understanding that some player locations are missing. Adapting our techniques to account for missing players (for instance, using imputation techniques on the missing locations) could improve our models, reveal key characteristics of soccer matches that would otherwise be lost, and open the door to methodological advancements. In particular, there are currently no established approaches for handling missing data in the context of data depth. Introducing such methods could lead to broader applications in statistical methodology beyond Sports Analytics.

5.4 Conclusion

In conclusion, this thesis has examined the application of dependent and sequential logistic regression to estimate the probability of a shot, a shot on target, and a goal occurring at any time point during a game. Furthermore, we employed these models to construct expected value metrics associated with these events. Our developed expected value metrics provide a deeper understanding of each team's scoring opportunities throughout the game. This is

accomplished by incorporating lagged response variables and numerous covariates that impact the probability of a shot, a shot on target and a goal.

Our research focused primarily on women's soccer to fill the gap in statistical analysis of women's sports. Our expected value metrics highlight important characteristics in matches that would otherwise not be captured when considering the observed number of shots, shots on target and goals. Additionally, we developed reduced models to further understand what covariates play significant roles in estimating the probability of a shot, a shot on target and a goal.

Additionally, we compared the expected value metrics and reduced models produced for women's soccer data to men's. In particular, we found that covariates that are important in estimating the probability of events in men's soccer are not necessarily the same as those deemed important in women's soccer. Further, we determined that our expected value metrics perform reasonably well with men's data, highlighting that our methodology seems useful across a variety of contexts and seems to lead to improved performance as more data becomes available.

Finally, with the increased interest in Sports Analytics, we are seeing more leagues and teams incorporate statistical analysis into their everyday operations. This change is drastically impacting the quality of decision-making in these organizations and leading to a better fan experience. In this thesis, we expanded upon previously researched topics by diving deeper into women's soccer. We hope that this encourages other researchers to study women's sports

and to continue uncovering the unique and similar aspects that women's sports have to offer.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19(6)*, 716–723. 60

Andrienko, G., N. Andrienko, G. Budziak, J. Dykes, G. Fuchs, T. von Landesberger, and H. Weber (2017). Visual analysis of pressure in football. *Data Mining and Knowledge Discovery 31(6)*, 1793–1839. 114

Anzer, G. and P. Bauer (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*, 624475–624475. 3, 79, 80

Bandara, I., S. Shelyag, S. Rajasegarar, D. Dwyer, E.-j. Kim, and M. Angelova (2024). Predicting goal probabilities with improved xg models using event sequences in association football. *PloS One 19(10)*, e0312278. 3

Bradley, P., A. Dellal, M. Mohr, J. Castellano, and A. Wilkie (2014). Gender differences in match performance characteristics of soccer players competing in the uefa champions league. *Human Movement Science 3*, 159–171. 5, 102

Bransen, L. and J. Davis (2021). Women's football analyzed: interpretable expected goals models for women. In *Proceedings of the AI for Sports Analytics (AISA) Workshop at IJCAI 2021*. 5, 102, 107

Bransen, L., J. Van Haaren, and M. van de Velden (2019, February). Measuring soccer players' contributions to chance creation by valuing their passes. *Journal of Quantitative Analysis in Sports 15(2)*, 97–116. 4

Delcroos, T., L. Bransen, J. Van Haaren, and J. Davis (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1851–1861. 4

Goumas, C. (2017, July). Modelling home advantage for individual teams in uefa champions league football. *Journal of Sport and Health Science 6*, 321–326. 44

Gu, C., J. Na, Y. Pei, and V. De Silva (2024). Player pressure map – a novel representation of pressure in soccer for evaluating player performance in different game contexts. Preprint at: https://arxiv.org/abs/2401.16235. 115

Heiner, M., G. W Fellingham, and C. Thomas (2014). Skill importance in women's soccer. *Journal of Quantitative Analysis in Sports 10(2)*, 287–302. 5

Hudl Statsbomb (2019, May). *StatsBomb Open Events Structure and Data Specification v4.0.0*. Hudl Statsbomb. https://github.com/statsbomb/open-data/tree/master/doc. 34, 48

Hudl Statsbomb (2021, November). *Open Data 360 Frames Specification v1.0.0*. Hudl Statsbomb. https://github.com/statsbomb/open-data/tree/master/doc. 41

Hudl Statsbomb (2025a). hudlstatsbomb. 2022 UEFA European Women's Championship data last accessed on April 2, 2025. https://statsbomb.com/. 11, 81

Hudl Statsbomb (2025b). hudlstatsbomb. 2020 and 2024 UEFA European Men's Championships data last accessed on April 21, 2025. https://statsbomb.com/. 102

Kauppi, H. and P. Saikkonen (2008). Predicting u.s recessions with dynamic binary response models. *The Review of Economics and Statistics 90(4)*, 777–791. 63

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company. 2

Lucey, P., A. Bialkowski, M. Monfort, P. Carr, and I. Matthews (2015, February). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proceedings of 9th Annual MIT Sloan Sports Analytics Conference*. 2, 3

Mead, J., A. O'Hare, and P. McMenemy (2023). Expected goals in football: Improving model performance and demonstrating value. *PloS ONE 18(4)*, e0282295. 3, 7, 113

Nattino, G., M. L. Pennel, and S. Lemeshow (2018). Assessing the goodness of fit of logistic regression models in large samples: A modification of the hosmer-lemeshow test. *Biometrics 76(2)*, 549–560. 55

Nevill, A. and R. Holder (1999). Home advantage in sport. an overview of studies on the advantage of playing at home. *Sports Medicine (Auckland) 28(4)*, 221–236. 44

Ramsay, K., S. Durocher, and A. Leblanc (2019). Integrated rank-weighted depth. *Journal of Multivariate Analysis 173*, 51–69. 37, 39

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise 12(2proc)*, 514–529. 3, 32, 86

Rein, R., D. Raabe, and D. Memmert (2017, October). "which pass is better?" novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science 55*, 172–181. 4

Scholtes, A. and O. Karakuş (2024). Bayes-xg: player and position correction on expected goals (xg) using bayesian hierarchical approach. *Frontiers in Sports and Active Living 6*, 1348983. 3

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6(2)*, 461–464. 60

Serfling, R. and Y. Zuo (2000). General notions of statistical depth function. *The Annals of Statistics (28)2*, 461–482. 37

Sucarrat, G. (2024). *logitx: Estimate an autoregressive logit model with covariates*. Rdocumentation. gets (version 0.38). https://www.rdocumentation.org/packages/gets/versions/0.38/topics/logitx. 58