Quantification and Analysis of Second Balls in Soccer

by

Jackson Sears

A thesis submitted to the
School of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Master of Science in **Computer Science**

Faculty of Business and Information Technology
University of Ontario Institute of Technology (Ontario Tech University)
Oshawa, Ontario, Canada
September 2025

THESIS EXAMINATION INFORMATION

Submitted by: Jackson Sears

Master of Science in Computer Science

Thesis Title: **Quantification and Analysis of Second Balls in Soccer**

An oral defense of this thesis took place on *September 3, 2025* in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Richard W. Pazzi
Research Supervisor	Patrick Hung
Research Co-supervisor	Jayshiro Tashiro
Examining Committee Member	Gabby Resch
Thesis Examiner	Nick Wattie

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

In soccer, second balls are crucial to control possession and create attacking chances, but have remained largely unexplored. In this thesis, a mathematical framework is created to identify, classify, and extract second balls from data. Building on this foundation, the novel Expected Second Ball Value (xSBV) model uses machine learning and Markov chains to estimate both the probability of winning a second ball and the likelihood that the following possession leads to a goal. Predictive models achieved a top-3 accuracy of 60% for second ball location and an ROC-AUC score of 0.79 for predicting the winning team. The key results highlighted specific areas to target for higher success rates and produced a ranking of players based on their second-ball winning ability. This thesis extends existing literature for second ball analysis, offering valuable applications for player evaluation and tactical decision-making.

Keywords: Soccer; Football; Second balls; Sport Analytics; Performance evaluation

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I authorize the Ontario Tech University to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the Ontario Tech University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Jackson Sears

Statement of Contributions

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am sole source of the creative works and/or inventive knowledge described in this thesis.

Acknowledgements

I would like to express my gratitude to everyone who supported me throughout my Master's here at Ontario Tech.

First, I am thankful to my supervisor, Professor Patrick Hung, for his support, kindness, and encouragement. Without it, I would never have had the opportunity to combine my love for the sport of soccer with my academic career. Next, I am grateful to Professor Jay Tashiro for the countless virtual meetings in which he challenged me to think about my research in new and creative ways, and for the invaluable feedback and support he provided me with for this thesis. Finally, to my family — thank you for your continued support throughout my academic and athletic career, as without it, I would not be standing here today at the conclusion of this journey, excited for what is ahead.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The game of soccer has advanced far beyond its traditional roots, and data analytics play an increasingly important role in the development of strategies, the evaluation of player performance, injury prevention, fan engagement, and overall team performance [1]. Soccer analytics involves the analysis of data to discover patterns, trends, and insights that can give teams a competitive edge. In this thesis, soccer analytics are investigated, specifically player and team evaluation, and a model is created and analyzed to advance the state of the field.

1.1 Motivation

The evaluation of the in-game actions of players to assess their impact on the outcome of a game is an important sub-section of soccer analytics [2]. Some examples of in-game actions are shots, passes, dribbles/carries, tackles, and even runs. Unlike traditional statistics that focus on goals and assists, evaluating other actions that players perform delves deeper into the nuances of player behaviour and decision-making. Given the nature of soccer as a low-scoring game, where goals occur in scarcity compared to more common actions such as passing, dribbling, and tackling, creating metrics to evaluate the level at which players perform said actions can help give better insight into predicting team

success.

An action that often goes unnoticed is the ability to win second balls [3]. A second ball is a possession gain following an intervention from an attempted long pass. Simply put, a second ball occurs after the ball is kicked long and neither team clearly controls it. Instead, the ball bounces, gets deflected, or is contested, and then the players rush to try and take control of the ball. Second balls are defined in detail with examples in Section 2.2. Second balls occur frequently in soccer matches and are often unpredictable in terms of where they occur. Unpredictability often leads to attacking chances and control of possession by the team that is winning them. A quote from Pep Guardiola, one of the most successful soccer managers, states the importance of winning second balls, "The main thing in English football is to control the second ball. Without that, you cannot survive." [4].

Despite the importance of winning second balls, there is a lack of research quantifying and analyzing the value of winning second balls and how to do so successfully. Although a recent paper has begun to address this gap [5], the amount of academic research on second balls is limited. Therefore, this thesis serves as some of the first academic research to explore and quantify the impact of second balls on a soccer game. To understand the impact of second balls, their influence on goal-scoring opportunities is analyzed, which will be explored in later sections.

1.2 Research Questions

This thesis aims to develop a model that can identify and predict the attacking threat associated with second ball opportunities during a soccer match. Various machine learning techniques will be used to understand second balls. This thesis aims to answer the following questions.

Research Questions

1. Can second balls be quantified using available soccer data?
2. Can second ball wins be accurately predicted?
3. What are the key factors influencing second ball wins?
4. Can the potential risk of second balls be predicted?

Through analysis, answering the above questions will lead us to important insights to help managers and teams make informed decisions about player selection and team strategies.

1.3 Contributions

This thesis makes several novel contributions to the study of second balls in soccer:

- **A mathematical framework for second ball identification:** Second ball sequences and possession chains are defined formally, allowing systematic extraction and analysis from event data.
- **Second ball win finding algorithms:** Functions are created in Python that detect and extract second ball wins from the StatsBomb event data.
- **Introduction of xSBV (expected Second Ball Value):** A model is proposed that combines spatial prediction, possession outcome prediction, and expected value modelling to quantify the impact of second ball wins.
- **Tactical insights for players and teams:** Actionable results are presented at the player and team level that highlight effective second ball winners, strategies, and areas for performance improvement.

1.4 Outline

The remainder of this thesis is organized as follows.

- **Chapter 2** provides background on the data used, introduces second ball definitions, and discusses their tactical importance. It also reviews related work and ideas in soccer analytics.
- **Chapter 3** describes the methodology used to build and implement the xSBV model. It includes the mathematical definition of second balls, data processing, feature engineering, model design, and evaluation metrics.
- **Chapter 4** presents the results of the model. This includes model performance, tactical insights, and use cases at the player and team level.
- **Chapter 5** summarizes the key findings, outlines the practical implications for users, and discusses the limitations and directions for future work.

Chapter 2

Background

Before looking into the existing literature on this research topic, it is important to discuss some relevant background information necessary to understand soccer analytics. This section is organized into three parts: Data, Second Balls, and Literature Review.

2.1 Data

Soccer analytics relies heavily on two main types of data that are often mentioned in the literature: event data and tracking data. Event data deals with contextual information, where tracking data deals with spatio-temporal information. Each type offers unique insights into player and team performance. Although event and tracking data can be used separately, their combination can provide a more comprehensive understanding of the game.

2.1.1 Event Data

Event data records discrete occurrences or actions during a match, such as passes, shots, tackles, and dribbles [6]. Aside from actions with the ball, event data also records other events like fouls and bookings. Each event is timestamped and contains the specific

location on the pitch where the event took place, as well as other contextual information about who acted, the outcome of the action, and more. Event data is typically collected and annotated by humans who manually watch game film. Figure 2.1 shows an example of event data and the information it can give about a game. The positions of the players are not provided, and the event data solely capture how the ball is moved across the field. Table 2.1 shows a sample of the data set. The column headings may differ depending on the provider of the event data, but they typically follow a similar structure.

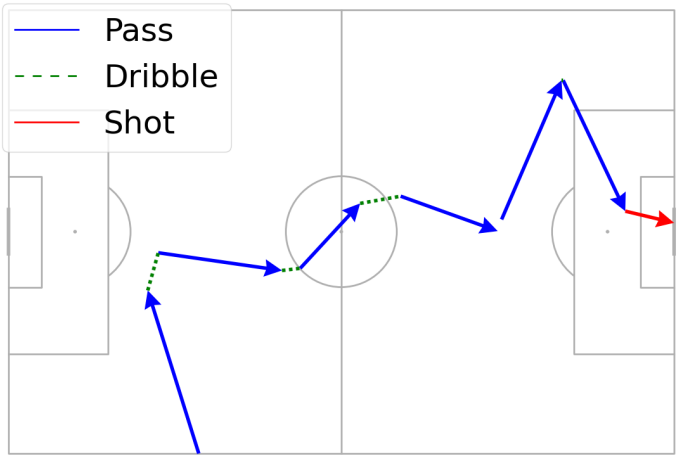


Figure 2.1: Event data visualization of the game winning goal scored by Mikel Oyarzabal to give Spain a 2-1 advantage over England in the 2024 UEFA European Championship.

Index	Type	Player	Team	Outcome	Location
1	Pass	Dani Carvajal	Spain	Success	(35, 80)
2	Carry	Aymeric Laporte	Spain	–	(25, 50)
3	Pass	Aymeric Laporte	Spain	Success	(30, 45)
4	Carry	Fabian Ruiz	Spain	–	(50, 53)
5	Pass	Fabian Ruiz	Spain	Success	(55, 52)
6	Carry	Dani Olmo	Spain	–	(63, 36)
7	Pass	Dani Olmo	Spain	Success	(70, 35)
8	Pass	Mikel Oyarzabal	Spain	Success	(90, 40)
9	Pass	Marc Cucurella	Spain	Success	(103, 15)
10	Shot	Mikel Oyarzabal	Spain	Goal	(110, 35)

Table 2.1: Example of what event data looks like in tabular form.

2.1.2 Tracking Data

While event data only focuses on what is happening with the ball, tracking data records the real-time positions of all players and the ball throughout a soccer match [7]. Typically recorded at a high frequency of around 25 Hz, tracking data offers detailed information about player movements and team shape, enabling deeper insights than event data alone. Tracking data is collected using advanced technology, which is expensive and is typically utilized only by private companies and professional clubs. Due to the need to maintain a competitive advantage, tracking data is not commonly accessible to the public.

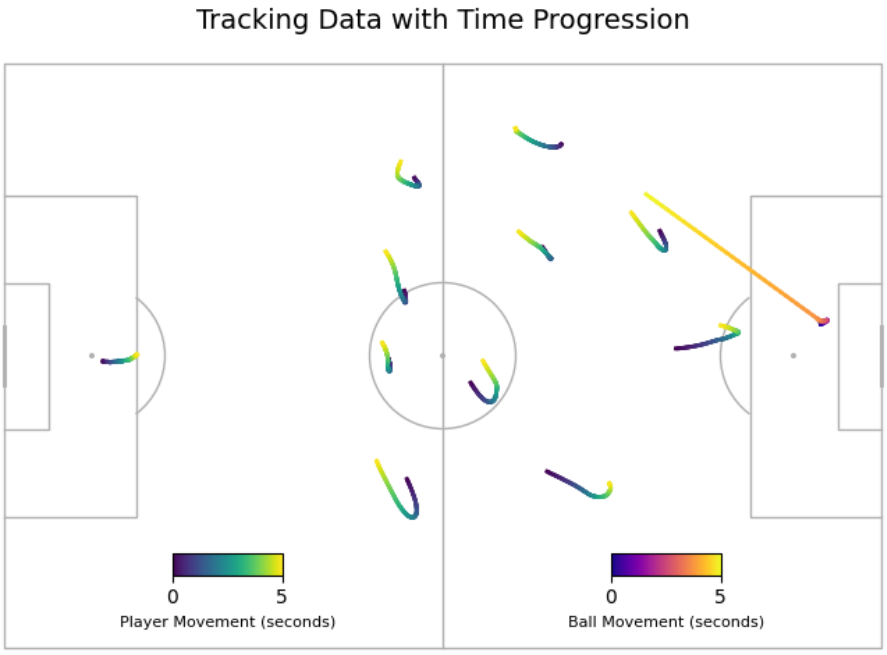


Figure 2.2: Tracking data of 11 players and the ball over a 5 second time frame with position updates every 0.04 seconds.

Figure 2.2 illustrates tracking data and its ability to capture the spatio-temporal aspect of players and the ball entirely. When tracking data is coupled with event data a complete picture of what is happening on and off the ball at all times can be achieved.

2.1.3 StatsBomb 360 Data

StatsBomb 360 is a type of data that supplements regular event data with parts of tracking data [8]. Although StatsBomb 360 data is not as powerful as full event and tracking data, it still offers more detail than just event data. Tracking data records the position and movement of all players on the field at high frequencies, providing a continuous, real-time view of the game. In contrast, StatsBomb 360 captures player positions only at the time of specific events, such as passes, shots, or tackles. Also, StatsBomb 360 data does not always capture the position of all 22 players on the field. StatsBomb 360 is less comprehensive than tracking data, which can limit the depth of analysis. Since it can be hard to find full tracking data, StatsBomb 360 data is a great alternative. It provides enhanced context and spatial insights beyond what event data alone can offer, making it a valuable tool for analysis in situations where tracking data is unavailable. In Figure 2.3, at the time a pass is made, the positions of the surrounding players are shown. Again, note that not all player positions are available.

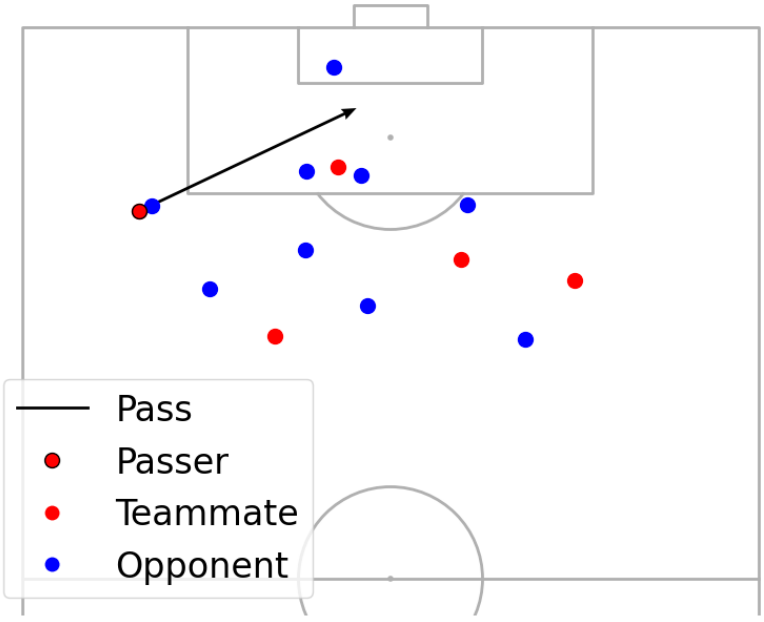


Figure 2.3: StatsBomb 360 data added to the assist by Marc Cucurella to Spain's game winning goal from Figure 2.1.

2.2 Second Ball Definitions

To understand this thesis, a working definition of second balls and second ball wins must be established. With little to no specific academic research found on second balls and wins, the inspiration for the definition comes from a blog post written by Chun Hang titled, "Quantifying Second-Ball Wins" [3]. The following definitions have been slightly modified, but remain largely unchanged as defined in the original blog post.

Definition 1 (Second Ball) *In soccer, a "second ball" refers to the loose ball that results from a contest between players, often after an aerial duel, clearance, or miscontrol.*

Definition 2 (Second Ball Win) *A possession gain or regain following an intervention from a long pass.*

1. ***Long Pass (long ball:*** *A pass with a length of 20 metres or greater that any player meets, successful or not.*
2. ***Interventions:*** *The first point of contact with the long ball must be an intervention. This intervention can be in the form of an aerial duel between players, a clearance attempt, or miscontrol. The key part of the intervention is that it leads to a second ball.*
3. ***Possession Gain/Regain:*** *A player controls the second ball gaining or regaining possession for their team. Including the player with the ball, their team must complete two successful actions to be considered a second ball win.*

2.2.1 Practical Example

To build a better understanding of what a second ball is, let us take a look at a few screenshots of a real in-game second ball win example. The goal is to help visualize what

is happening when talking about second balls. Figure 2.4 shows a second ball win with three snapshots, one for each part of the second ball win. In order, the highlighted player denotes the long ball by Spain, the clearance by England, and the second ball win by Spain.

Figure 2.4: Second ball win example taken from in-game footage of Spain vs England in EURO 2024. Red boxes are used to show the area of interest. From top to bottom the pictures show the following: long pass, clearance, regain.

2.3 Importance of Second Balls

While the definition of second balls and the data used to investigate them has been established, their importance remains unexplored. To justify a deeper investigation, their impact on match outcomes and season-long performance is quantified. The following analysis examines the relationship between second ball success, goal difference, and overall team performance to provide evidence for further investigation.

2.3.1 League Performance

To establish a link between second ball wins and team success, the relationship between goal difference (GD) and total points earned in a season is examined. Using data from the top five European leagues [9], a linear regression is performed, where the dependent variable is points and the independent variable is GD.

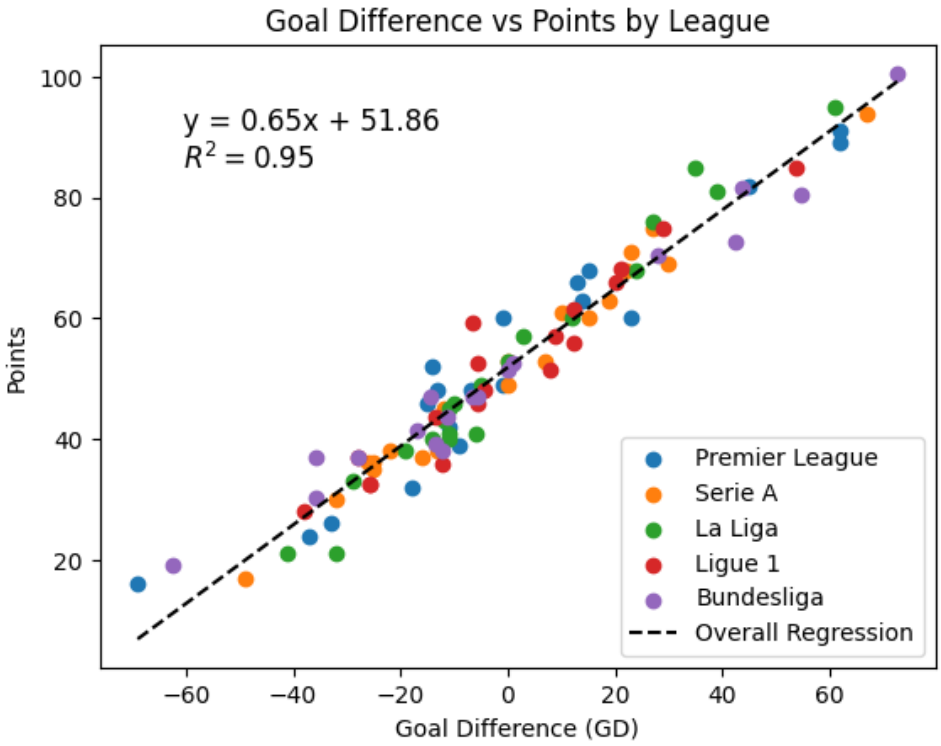


Figure 2.5: Linear regression of points vs goal difference from the top 5 leagues in the 2015/2016 season.

In Figure 2.5, an R^2 value of 0.95 demonstrates a strong linear relationship between variables. Based on the regression, an increase in GD by 1 corresponds to an average increase of 0.65 points. Potentially, even marginal improvements in GD can affect a team's league standing.

2.3.2 Shot Creation

Next, the role of second ball wins in generating goal-scoring opportunities is analyzed. The following information is from the data set.

- 19.38% of the second ball wins result in a shot within the following possession.
- The average expected goals (xG) per shot is 0.10098 or 10.1%.
- On average, teams contest approximately 24 second balls per game.

The above list suggests that improving second ball success could lead to a higher volume of shot opportunities. Also, expected goals (xG) is a metric that gives the probability that a shot results in a goal. xG will be discussed more in Section 2.4.

2.3.3 Impact of Winning More Second Balls

To quantify the potential impact of second ball wins, the effect of a team winning five additional second balls per match is estimated. Based on the data set:

- Winning 5 extra second balls per game leads to approximately 0.969 additional shots per game, or 36.81 additional shots per season (for a standard 38 game season).
- Given an average xG per shot of 0.10098, this translates to 3.72 expected goals gained per season.
- Conversely, by preventing the opponent from winning these second balls, a team can reduce shots conceded, preventing 3.72 expected goals conceded per season.

Thus, by improving second ball efficiency, a team could see a net gain of 7.44 GD per season.

2.3.4 Translating Second Ball Success to League Points

Finally, the findings are integrated into the GD vs. Points regression model. If a team gains 3.72 expected goals during a season and prevents 3.72 expected goals, their goal difference increases by 7.44. This improvement corresponds to 4.84 additional points over a season using the established regression. It is important to emphasize that this estimate is based on model assumptions and averages. The potential points gained should be interpreted as an indicative projection rather than a guaranteed outcome. The analysis illustrates the *potential* value of second ball wins, while acknowledging that in practice the actual impact may be higher or lower depending on team strategy, player execution, and variance in outcomes.

Promotion/relegation is a system where the best-performing teams move up a division and the worst-performing teams move down. There is no promotion in the top divisions; teams simply play to become champions. The financial impact of being promoted or relegated has a lasting impact on a club. Promotion is worth between $238 and $280 million over 7 years, where the cost of relegation is between $225 and $262 million over 7 years [10]. These numbers alone provide clubs with a high degree of interest in doing everything they can to get promoted and avoid relegation wherever necessary. Looking at Table 2.2, the difference between winning the league and avoiding relegation can be very few points. In this example, with three extra points, Newcastle United could have avoided relegation. Based on my analysis, less than five extra second ball wins per game could have potentially saved Newcastle. Southampton missed out on European qualification by three points, which means they missed out on potentially large amounts of money. My analysis suggests that providing tactical insight into winning more second balls could significantly improve teams' chances of a successful season.

Rank	Team	Points
1	Leicester City	81
2	Arsenal	71
3	Tottenham	70
4	Manchester City	66
5	Manchester United	66
6	Southampton	63
7	West Ham	62
8	Liverpool	60
9	Stoke City	51
10	Chelsea	50
11	Everton	47
12	Swansea City	47
13	Watford	45
14	West Brom	43
15	Crystal Palace	42
16	Bournemouth	42
17	Sunderland	39
18	Newcastle	37
19	Norwich City	34
20	Aston Villa	17

Legend:

Champions and European Qualification

European Qualification

Relegation zone

Table 2.2: Final league standings for the 2015/2016 English Premier League with legend.

2.4 Literature Review

Now that we have established a solid background in soccer analytics research, we can begin the literature review section of the thesis. To my knowledge, only one academic paper has directly addressed the concept of second balls [5]. This notable gap in the literature led to the exploration of related player and team evaluation research to understand how existing methodologies can be adapted to analyze second balls. Finally, an in-depth summary of the work done on second balls will be provided, highlighting the strengths and limitations of the current approaches to quantifying second balls.

2.4.1 Expected Goals Models

A good place to start is the Expected Goals (xG) model. In 2012, Sam Green [11] laid the foundation for xG as a metric to quantify the quality of scoring chances in a soccer

game. By analyzing shot data, Green developed a probabilistic model that estimates the likelihood of a goal being scored from any given shot by using circumstance specific information. Consider two forwards, Player A and Player B, who scored 5 goals in the same number of games. Their stats may seem similar, but in reality, Player A's xG was 3.5 goals, whereas Player B's xG was 7 goals. The low xG value for Player A means that they did not have as many chances as Player B, but were very clinical and highly skilled (or lucky). On the other hand, Player B's high xG value meant they had many chances, but were wasteful and not as skilled (or unlucky). A metric like xG can provide deeper insight about players, helping professional teams make better decisions that best suit their needs, which is ultimately one of the main goals of soccer analytics. Green's model represented a significant advancement in evaluating player performance and team strategies, providing a better understanding of match outcomes and shooting efficiency outside of simple goal counts and shot statistics.

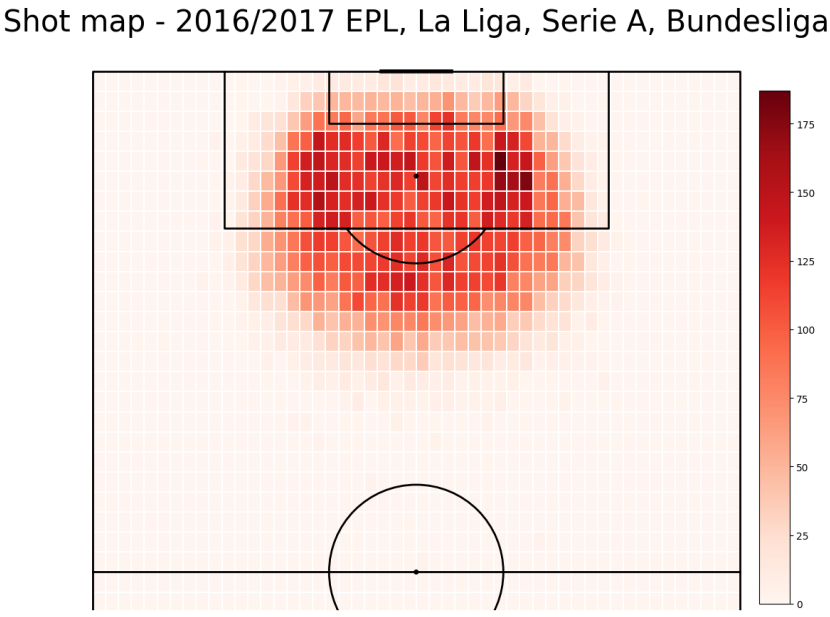


Figure 2.6: Heat map of shots taken from different areas of the pitch across the English Premier League, La Liga, Serie A, and Bundesliga in the 2016/2017 season.

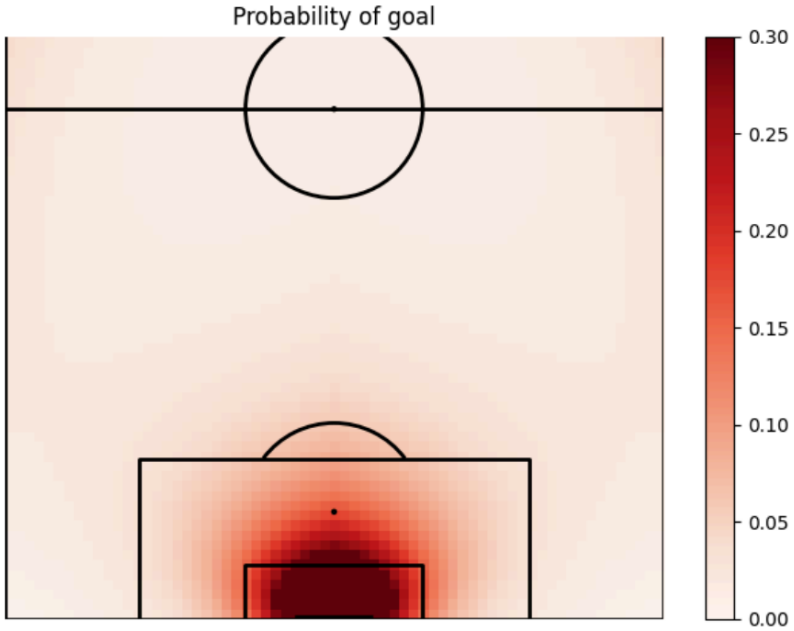


Figure 2.7: Heat map of the probability of scoring based on shot location and angle to the goal. This figure represents an xG model's values depending on shot location.

In Figures 2.6 and 2.7, the difference between the quantity and quality of the shots is shown. Figure 2.6 shows the number of shots from different locations on the pitch, with many shots occurring close to and far from the goal. Figure 2.7 shows the xG values for each area of the pitch, which gives insight that a player is much more likely to score when he is closer to the goal and in the middle of the net. These are simple examples, just taking into account distance and angle from the goal, when in reality many other factors like pressure, type of shot, and goalkeeper positioning affect the shooter's chances.

xG models have since evolved, becoming one of the most extensively researched soccer analytics metrics. The first important area to study was understanding which features affected the accuracy of an xG model the most. A feature is a parameter given to the machine learning algorithm that provides information to help predict the desired results [12]. Initially, distance alone was thought to be a strong enough feature to produce accurate xG models [13]. However, logistic regression performed on a season that combined distance and angle to the goal worked better [14]. Rathke also argues that expected goals

are a reliable predictor of where teams will finish in their league. The problem with using a metric like expected goals to predict team placement is that stronger teams with more money and better players will score more goals and have higher expected goals. So, even though it could be used as a reliable predictor of the league table, many other reasons explain a team's goal success. It is argued that xG models have better applications in evaluating players.

Another topic of research about xG models is how the models are trained. One paper looked to improve model performance by adding features such as the current game week of the season, the goal difference of the game, the age of the shot taker, and the length of possession before the shot occurred [15]. For example, a team losing by a goal with five minutes left in a game was more likely to convert a given chance than if it had been earlier in the game. The psychological factor of "needing" a goal to avoid the negative outcome of losing helps a player to score. In contrast, Robberechts and Davis investigated xG models that use only factors on the field, such as the location of the shot, the body part used, and the speed of the shot [16]. It is interesting as both works achieved a similar area under the Receiver Operating Characteristic curve of 0.8, indicating that the model is efficient in distinguishing between shot outcomes [17]. Both papers also had Brier scores close to 0.08, indicating good model accuracy [18]. Although xG models do not need to be explicitly trained with psychological factors or the shooter's age to produce good results; they can be used to investigate the effects of those factors. Investigating psychological and age factors can lead to a wide variety of new research questions. How does the age of a player impact their goal output? How does team dynamics affect the outcome of a player's shooting ability? The point is that since xG models are well developed, they can test the correlation of other external and performance factors, giving a reason to extend the research to incorporate new ideas.

It is also worth noting that most xG models from earlier years used only event data due to ease of use and availability. As early as 2015, a considerable shortcoming of xG models

that did not incorporate tracking data was discussed in blog posts [19]. As a result, new avenues for research that included tracking data were created. The absence of tracking data, a considerable issue, was formalized in a paper about defensive pressure. The notion of an opposition getting close to the shooter with the ball significantly decreases the chance that a shot will reach its intended target [20]. The probability decreases because the shot could be blocked, or the attacker feels rushed to shoot. The lack of synchronization between the event and tracking data was a problem that increased the difficulty of accurately using the two types of data together. Humans record event data, introducing variability in the temporal alignment between the event and tracking data that should be, in fact, synchronous. For example, if a model uses the event data timestamps, the tracking data used to describe the position of the opposition might be at a different moment in the game, leading to inaccuracy. Anzer and Bauer found that without synchronization, event timestamps differ by 1.82 (+-4.06) s, whereas with synchronization, they only differed by 0.23 (+-0.49) s [21]. The work by Anzer and Bauer bridged the gap between using event and tracking data in tandem, and now many models are incorporating these ideas into new work.

When originally wanting to create a thesis on improving xG models, this review of the literature clarified the abundance of research and improvements that had already been performed. Instead the focus is shifted to using xG to help evaluate second ball wins and the attacking chances created from the resulting second ball possessions.

2.4.2 Other Expected Models

xG models are just one piece of the puzzle. On average, there are only nine shots on target between both teams during a 90-minute game [22]. Compared to the average of 905 passes between teams during a match in the English Premier League 2019-20 [23], intuitively there would seem to be more information to understand how more frequent actions such as passes and dribbles affect player performance. From the idea of using

historical data to make predictions about current data, many different expected models similar to xG have been created.

Expected Threat (xT) is a player evaluation metric that quantifies the potential impact or "threat" that player actions have on increasing their team's likelihood of scoring a goal [24]. Unlike xG which evaluates shots, xT evaluates how passes and dribbles contribute to advancing the ball into more dangerous attacking areas. Although Singh created this metric [24], Sarah Rudd provided the mathematical explanation for it 7 years earlier in 2011 [25], using Markov chains. A Markov chain model assumes that a change between states is independent of earlier events. Regardless of whether a team has had the ball for sustained possession or just won it, the model only looks at the current state. The pitch is divided into zones and historical data is used to determine the probabilities that the ball will move to and from each zone. Since second balls occur after a long ball and can be considered an independent event, a similar approach using Markov chains could be used to model the probability of where a second ball could land.

The Expected Pass Turnovers (xPT) is another metric that can be used to evaluate player performance. xPT is a logistic mixed-effects model that calculates the probability that any given pass results in a turnover [26]. For example, a player who has an xPT value of 20/100 is expected to turnover 20 out of 100 passes, but in reality turned over 40/100, it can be said that they are doing a below average job of retaining the ball. Since pass attempts often follow second balls, xPT can potentially be used to analyze the risk associated with these passes.

Other models have been created to assess the creativity of passes. un-xPass values the creativity, usefulness, and originality of passes by training an XGBoost model based on hand-crafted features and a deep learning model called SoccerMap [27]. The model predicted the selected receiver of passes to understand the most typical passes seen in a game; however, their selection model only achieved an accuracy of around 54%. Compared to other research such as [28], where an F1 score of 0.95 is achieved for a similar task

using a Temporal Graph Network and [**xPass**], where an accuracy of 72% is achieved for the intended receiver portion of an expected pass model, there are different approaches that have achieved higher levels of accuracy.

Becoming familiar with expected models in soccer analytics is important, as they are some of the simpler concepts to understand while still having important underpinnings in the research field. Using historical data to build these predictive models, information about certain parts of the game can be found. Although expected models usually focus on one part of the game such as shooting [11], [13], [14], [15], [16], passing [24], [25], [27], [**xPass**], or turnovers [26], there are more advanced models that try to evaluate player performance as a whole, which are discussed next.

2.4.3 Advanced Performance Models

Advanced performance models offer a more complete evaluation of player performance, not just in a limited scope as in the previous section. These advanced models use player actions throughout a match to assess their overall influence on team performance [2]. By considering contributions to both attack and defence, these models provide an understanding of individual player performance and how it contributes to the team's success. Although these models provide insight into overall player performance, they often lack interpretability in specific areas of the game, unlike the models discussed in the previous section.

In 2016, a metric was introduced called Dangerousity [29]. Dangerousity quantifies the attacking performance of teams, specifically by calculating the probability of a goal being scored for every point in time when a player has the ball. Using event and tracking data, functions are created to model the position of the ball, the level of ball control, the amount of pressure applied to a player, and the density of opponents in front of the goal. Soccer experts calibrated the functions based on the dangerousity scores outputted by the model and what the experts thought the values should be. To this extent, it is an

early model without the use of machine learning that allows for the quantification of a goal being scored or conceded in soccer.

Valuing Actions by Estimating Probabilities (VAEP) is a model to evaluate overall player and team performance [30]. The framework was built using only event data. The model can predict the probability of any given action resulting in a goal. For example, a pass that improves the position of the team might have a value of 0.05 or a five percent increase in the probability of a goal. In contrast, a turnover might result in a negative score, as it increases the probability that the opposition scores.

(c) Top-10 players in terms of goals + assists per 90 minutes (g+a/90)

R_{g+a}	Player	g+a/90	R_{vaep}	Market Value
1	S. Agüero	1.235	14	€ 75m
2	M. Salah	1.232	2	€ 150m
3	P. Aubameyang	1.191	42	€ 75m
4	P. Coutinho	1.049	1	€ 140m
5	R. Sterling	0.972	7	€ 120m
6	H. Kane	0.905	9	€ 150m
7	L. Sané	0.853	47	€ 100m
8	G. Jesus	0.808	204	€ 70m
9	A. Martial	0.795	6	€ 60m
10	O. Niasse	0.749	17	€ 7m

Figure 2.8: Top players in terms of goals and assists per 90 minutes (g+a/90) in the 2017/2018 English Premier League.

(d) Top-10 players in terms of our VAEP player ratings

R_{vaep}	Player	Rating	R_g	R_a	R_{g+a}	Market Value
1	P. Coutinho	0.899	10	2	4	€ 140m
2	M. Salah	0.817	1	23	2	€ 150m
3	K. De Bruyne	0.641	72	4	15	€ 150m
4	E. Hazard	0.636	21	122	34	€ 150m
5	R. Mahrez	0.635	34	11	16	€ 60m
6	A. Martial	0.607	13	13	9	€ 60m
7	R. Sterling	0.579	7	6	5	€ 120m
8	P. Pogba	0.549	55	9	28	€ 80m
9	H. Kane	0.545	4	140	6	€ 150m
10	S. Heung-Min	0.539	19	36	17	€ 50m

Figure 2.9: Top players in terms of VAEP player ratings in the 2017/2018 English Premier League.

Figures 2.8 and 2.9 are taken from [30] and compare the goal and assist ratings with the VAEP ratings. The results quantify players who only occasionally appear on mainstream stat tables. For example, Kevin De Bruyne, Riyad Mahrez, and Paul Pogba do not regularly rank among the world's top 10 combined goals and assists. However, if you talked to experts, coaches, players, and fans, the overwhelming feelings about these players' abilities would be very high. Not only does VAEP shed light on excellent players who do not rank highly in the conventional goals and assists charts, but it also uncovers potentially undervalued players. Models that give insight into players who are not 'superstars' are valuable, as teams can find players who provide great value to their team that would otherwise go unnoticed.

The most significant criticism of this work is that only on-the-ball actions are considered. Since the model only uses event data it does not incorporate the positions of other players. This is important especially for second balls. When considering my problem, the VAEP framework would provide a probability change based on the winner of a second ball. However, it does not account for the positions of players around the ball, which likely has a significant impact. Also, VAEP cannot understand the likelihood of who will win the second ball. Although it might be able to predict the value of the increase in chance of scoring from the second ball, VAEP is not designed to predict who will most likely win a second ball and why. Understanding the reasoning behind how second balls emerge is an important point of this thesis, as it can potentially lead to insights about how a team can improve their chances at winning games through increased second ball wins.

Further analysis [31] compared VAEP and xT and showed a few important conclusions. First, VAEP favours goal-scoring actions, whereas xT favours key passes and dribbles. Also, VAEP does a better job at capturing the risk-reward tradeoff of actions. If an action would increase a team's chances of scoring by 30% but increase the opponent's chances of scoring by 40% if unsuccessful, VAEP captures this within its model,

while xT does not. However, the robustness of VAEP was much worse than that of xT. Player ratings across seasons were split into samples, and the consistency of the ratings using both metrics were compared. VAEP scored much worse because goals contributed significantly to the overall ratings; therefore, the results can be skewed by a few extra goals included or excluded from the samples. xT, which focuses on ball-progressing actions, had a much higher Pearson correlation of 0.89 (best case was 0.59 for VAEP).

The next important player evaluation metric is called Expected Possession Value (EPV). Similar to dangerousity [29], EPV models the likelihood of a team scoring or conceding the next goal at any point in time [32]. The model uses event and tracking data and is one of the most complete and commonly used player evaluation metrics today. The following equation represents the EPV model, which uses T_t for time, and further breaks down to separate parts for each of the three different actions (A): passes (ρ), shots (ζ), and dribbles (∂).

$$EPV(t) = \mathrm{E}[X|T_t] = \underbrace{\mathrm{E}[X|A=\rho]}_{\substack{\text{Passing value}\\ \text{surface}\\ \text{dependent on}\\ \text{destination}\\ \text{location}}} \overbrace{\mathrm{P}(A=\rho)}^{\text{Action likelihood model}} + \underbrace{\mathrm{E}[X|A=\varsigma]}_{\substack{\text{Expected}\\ \text{goals}\\ \text{model}}} \overbrace{\mathrm{P}(A=\varsigma)}^{\text{Action likelihood model}} + \underbrace{\mathrm{E}[X|A=\delta]}_{\substack{\text{Drive value}\\ \text{surface}\\ \text{dependent on}\\ \text{destination}\\ \text{location}}} \overbrace{\mathrm{P}(A=\delta)}^{\text{Action likelihood model}}$$

Figure 2.10: Expected Possession Value (EPV) equation used to model the probability of a team scoring or conceding the next goal at a given point in time.

The above equation shows the main components of the EPV model and is taken from [32]. The model includes the value gained from each action as well as the likelihood that each action will occur. The added requirement of incorporating the likelihood of something happening, not just the value if that something happens, is key to creating a metric that can predict and assign value to second balls. The probability of an action taking place is a step beyond what most simple expected value models quantify. However,

it is a requirement for my second ball model since the probability of different players receiving the ball and whether they will retain, lose, or create danger with the ball is crucial to gain a deeper insight into second balls.

A criticism of the original EPV model was that the results were hard to interpret for non-analysts, which the majority of soccer coaches and staff are. It is crucial for a coach to understand the technology to build trust and make informed decisions that prioritize team success. To combat said criticism, a decomposed EPV model was created [33] that breaks the original model down into various sub-components, each of which is estimated separately. Not only does this make results easier to interpret, but it also allows for various new practical applications.

$$EPV_t = (\sum_{l \in L} \overbrace{\mathbb{E}[G|A=\rho, D_t=l, T_t]}^{\text{Joint expected value surface of passes}} \overbrace{\mathbb{P}(D_t=l|A=\rho, T_t))}^{\text{Pass selection probability}} \mathbb{P}(A=\rho|T_t)$$
$$+ \overbrace{\mathbb{E}[G|A=\delta, T_t]}^{\text{Expected value of ball drives}} \mathbb{P}(A=\delta|T_t)$$
$$+ \overbrace{\mathbb{E}[G|A=\varsigma, T_t]}^{\text{Expected value from shots}} \mathbb{P}(A=\varsigma|T_t)$$

$$\mathbb{E}[G|A=\rho, D_t, T_t] = \overbrace{\mathbb{E}[G|A=\rho, O_\rho=1, D_t, T_t]}^{\text{Expected value of successful/missed passes}} \overbrace{\mathbb{P}(O_\rho=1|A=\rho, D_t, T_t)}^{\text{Probability of successful/missed passes}}$$
$$+ \mathbb{E}[G|A=\rho, O_\rho=0, D_t, T_t]\, \mathbb{P}(O_\rho=0|A=\rho, D_t, T_t)$$

$$\mathbb{E}[G|A=\delta, T_t] = \overbrace{\mathbb{E}[G|A=\delta, O_\delta=1, T_t]}^{\text{Expected value of successful/missed ball drives}} \overbrace{\mathbb{P}(O_\delta=1|A=\delta, T_t)}^{\text{Probability of successful/missed ball drives}}$$
$$+ \mathbb{E}[G|A=\delta, O_\delta=0, T_t]\, \mathbb{P}(O_\delta=0|A=\delta, T_t)$$

Figure 2.11: EPV equations being broken down into components to provide better understanding.

Taken from [33], Figure 2.11 shows the original EPV equation broken down to include the expected value of successful/missed passes and the probability of said passes, as well as the expected value of successful/missed dribbles and the probability of said dribbles. It is noted that a broken-down shots equation is not included, as it is simply an xG model, which is a common and well-understood metric. Customized convolutional neural networks are used for training the components involving passing, such as pass success probability, expected pass value, and expected pass selection. Standard shallow neural networks are used for dribble probability, expected dribble and shot value, and the action selection probability components. The adaptive moment estimation algorithm [34], a first-order gradient-based optimization of stochastic objective functions, is used to train each component of the model.

Since EPV is one of the most common player and team evaluation models, it is important to develop a good understanding of it, especially since many of the broken down components from [33] can be used or modified to second balls. Passing probabilities, expected value, and expected pass selection are all components that can be incorporated when analyzing immediate options and the best options available after a second ball occurs. A similar idea can be used to create components that quantify the second ball win probability, expected value of the second ball, and the probability of where the second ball will land.

There has been other research on quantifying player decision-making and the value of passes. One model evaluates the risk of losing the ball and the gain in attacking value when passing using the EPV [35]. The work incorporates all available passing options when a player has the ball to understand the differences between safe and risky passes. There is a strong correlation of $R^2 = 0.754$ between the two variables of closeness to the opponent's goal and risk when passing, meaning that defenders play the safest passes. In contrast, when you move towards the attackers, riskier passes are made. The latter point is trivial, as you must be safe to limit the possibility of turnovers near your goal

to avoid conceding. You must be creative and play riskier passes in the attacking end to beat defences and score goals. Another interesting takeaway is that the model shows the risk assumed when receiving the ball, which sheds light on how players can retain and possess the ball under adverse conditions. Adapting this idea to second balls and finding which players are good at dealing with the quick and often chaotic scenarios of second balls could be useful for analysis.

Anzer and Bauer created a binary classification model that incorporates different features, such as the probability that a pass is intercepted or blocked, to determine the difficulty of a pass by predicting the receiver of a pass with 93.0% accuracy [**xPass**]. A similar result was obtained by [28], where a Temporal Graph Network predicted pass receivers and success rates that achieved an F_1 score of 0.95 and an AUC score of 0.92.

Utilizing where the ball could potentially go led to a model to quantify the quality of off-ball positioning of players before shots occur [36]. The model uses pitch control, a metric that will be described in the next section, to quantify the probability of scoring at the next on-ball event. Spatial probability densities are created to visualize which areas are most likely to have goals occur from, which can be spatially integrated to give a probability.

The advanced performance models that have been discussed thus far give a complete idea of the difficulty of problems that can be tackled using more in-depth machine learning and analytical methods. As seen in [28], [**xPass**], [29], [35], and [36], using not only the expected value of actions, but also the probability that specific actions occur, a more holistic evaluation model can be made. Examining the likelihood of events occurring, regardless of whether they do, is an important takeaway for building my second ball model. Other approaches, such as [30] and [32], evaluate every action that occurs on the field. However, analysis with second balls has not been conducted, and whether these models fully encapsulate second balls is unknown. For these reasons, there are plenty of ideas that we can adapt to help us understand second balls.

2.4.4 Ball Recovery Metrics

In the final section of the literature review, ball recovery metrics are discussed, as they share many similarities with second balls. Ball recovery metrics are those that quantify parts of the game after the ball changes possession. Some examples of these instances are turnovers, second balls, counterattacks, pressing, and repressing. All of these share a key factor: transition moments. Part of the reason second balls are so dangerous is that they result in a transitional moment. When the team in possession suddenly loses the ball, they must reorganize to defend. During this re-organizational period, there can be an opportunity for swift attacks to take advantage and create scoring threats.

Pitch control is defined as the probability that a team will have the ball if it reaches a specific location on the field [37]. Using a physics-based approach that incorporates the velocities, location, and time it takes a player to control the ball, a model is created that visualizes and helps to understand possession.

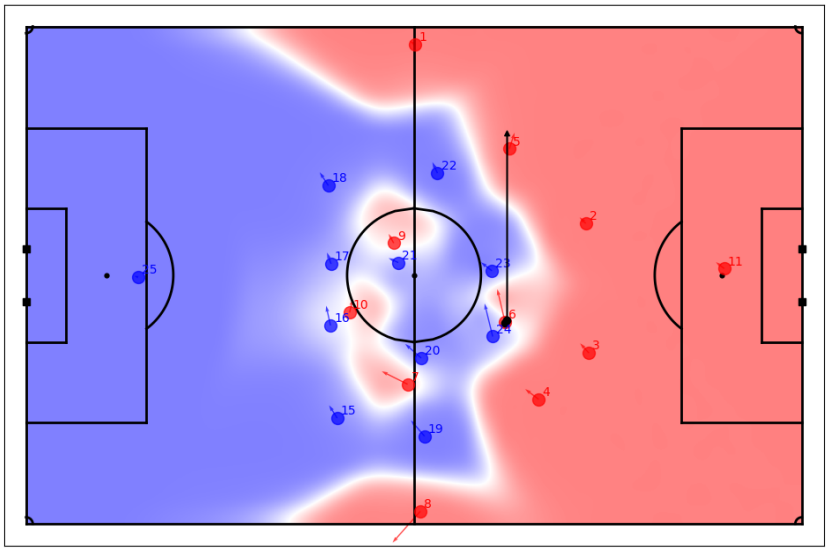


Figure 2.12: Pitch Control model visualized to show the probability that either team possesses the ball given it travels to a point on the pitch. The figure shows all players from both teams, their respective velocities, and where the ball is moved to in the next action (black).

Figure 2.12 is adapted from online tutorials [38] and shows how pitch control quantifies

the space that is controlled by either team. With players on each team denoted by blue or red, the model shows which colour team would most likely control the ball if the ball were to travel to any given point on the field. Many models incorporate pitch control into them as it can help to give insight about which passes will be successful, where attacking threats are, and what areas of the pitch specific team play styles might benefit from playing in. Second balls can benefit significantly from such a metric. Using pitch control to understand who controls what space around a second ball will potentially help strengthen the accuracy of the proposed predictive models.

Another model incorporates pitch control to quantify pressure in soccer [39]. Pressure is when an opponent closes the space and time that a player with the ball has. It is an attempt to force a turnover or make it difficult for the ball player to make a good pass. A graph neural network is trained to predict a possession outcome model, representing the probability that the team in possession loses the ball. A player pressure map is created using pitch control to quantify the pressure on the player with the ball in six directions. By incorporating pressure maps and using pitch control, the accuracy of predicting when a player will lose the ball due to pressure increases from 55.8% to 75.2%. Pressure is an important concept, as a second ball occurrence could potentially yield significant results regarding the pressure applied to a player after winning a second ball. For example, if it is not possible to win a second ball, then quick pressure could lead to the least dangerous scenarios for the defending team.

Since pressure is important in limiting the attacking options for a team, the ability to find and create space for a player is the goal of the team in possession. A method for quantifying spatial value occupation and generation during open play was presented in [40]. The player's influence on controlling the ball at a given point is modelled using a bivariate normal distribution. The team's control can be calculated by accounting for the influence of every player on the team. Then, by using the observation that most defenders are positioned normally in high-threat areas of the field, their influence can be

used to calculate the value of space. The metric shows space gained by players, which could be a useful tool for second balls.

Valuing Pressure Decisions by Estimating Probabilities (VPEP) is a metric to value pressing actions performed by players [41]. Similar to VAEP [30], a probabilistic classifier, in this case XGBoost, is used to train the probability that the defending team will recover the ball ($P_{recovery}$) and the probability that the defending team will concede a goal-scoring opportunity shortly (P_{attack}). Using only event data, an ROC AUC of 0.902 for $P_{recovery}$ and 0.806 for P_{attack} is obtained, as well as the following calibration curves and histograms.

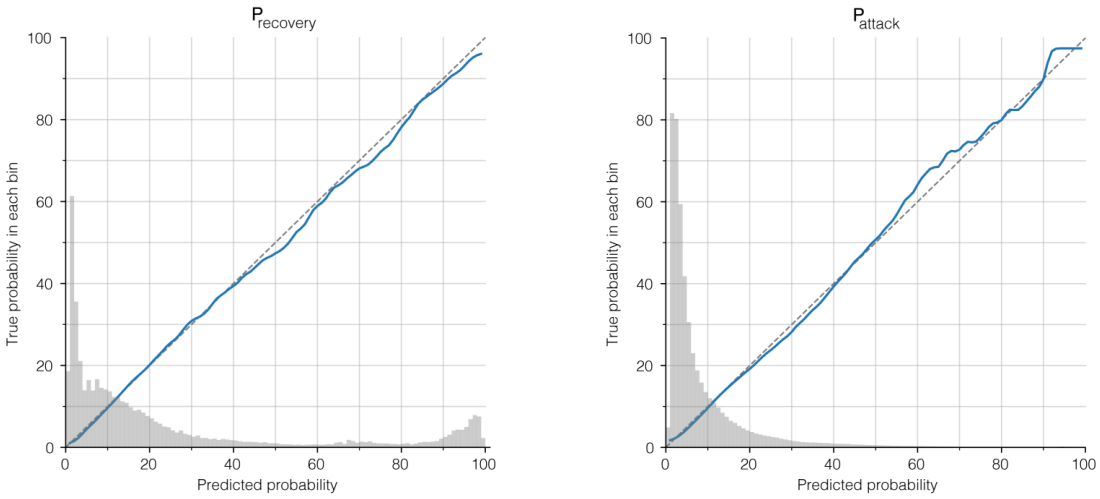


Figure 2.13: Probability calibration curves and histograms of the predicted probabilities.

Figure 2.13 is taken from [41]. It is noted that the curves fit well, which indicates a good calibration. The technique used in Figure 2.13 is a great way to visualize the prediction accuracy of the model vs. the accuracy of the ground truth. The work of VPEP was extended [42] to include tracking data. Following a similar approach, but incorporating pass selection probabilities, expected value of ball recoveries, and expected threat value provides more insight into the risk and rewards of pressing. Figure 2.14 is taken from [42] and improves the interpretability of the original VPEP model since examples show where the risk and reward come from. With many less common or abstract ideas, such as pressing, counter-pressing, and, in my case, second balls, being able to

visualize results to understand the metric is very important for coaches and managers to build trust with the model.

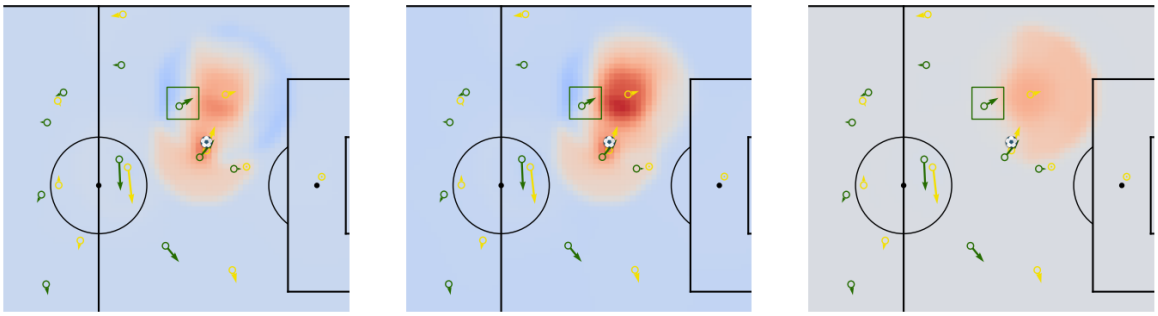


Figure 2.14: Metric can identify where pressing effectiveness can be maximized. In this example, the player who is boxed can improve their pressing effectiveness by pressing more aggressively. Higher and lower values for the effectiveness, risk, and reward are plotted in red and blue, with the team in green attacking towards the right.

Counterattacks are the last topic we will cover in this section. Counterattacks are a fast attack that happens after winning the ball from the opponent [43]. They are similar to second balls, since teams can be unprepared to defend following a change in possession, which can lead to goal-scoring opportunities. In [43], counterattacks are quantified through a framework that allows understanding of such sequences. The goal was to predict counterattacks from in-game sequences before they happen. A logistic regression model achieved an accuracy of 64%. Conclusions such as having more players behind the ball, results in less dangerous counterattacks, but there is a notable gap in quantifying the danger of these types of scenarios. Other papers have analyzed counterattacks, such as in Major League Soccer (MLS) [44], and the physical demands of a counterattacking vs ball possession playing style [45]. However, little research has been done to quantify the value of counterattacks and what makes them successful.

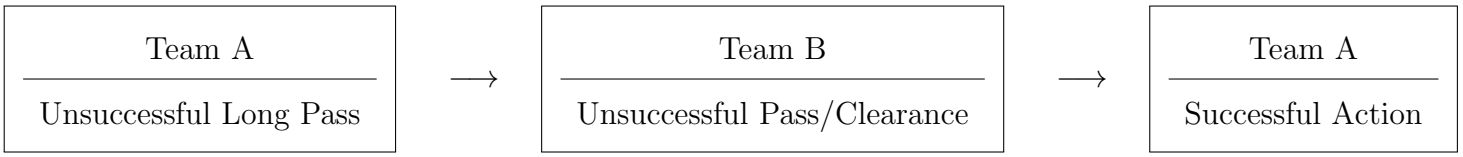
2.4.5 Second Balls

Finally, there is a lack of academic research on second balls. To date, only one research paper has directly analyzed second balls [5], while a small number of blogs and articles have offered exploratory analyses [3][46]. Despite these contributions, second balls remain largely an unexplored area of research. This section reviews the academic study and other sources to summarize the current state of research on second balls and to identify areas for future research.

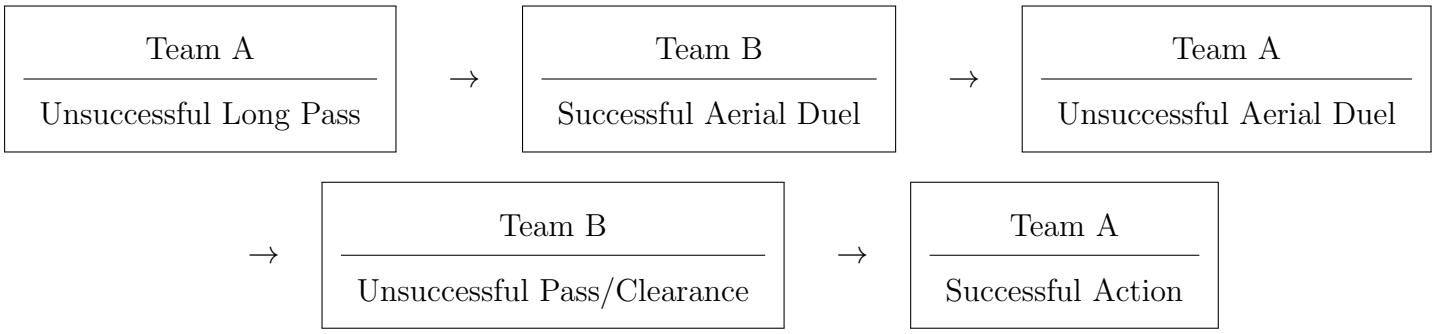
Sunjic et al. [5] analyzed how second balls affect technical performance. Technical performance is divided into variables such as shots, passes, duels, and dribbles. Using linear mixed models, the researchers found that second ball wins in the first and final third are positively correlated with technical performance. However, second ball wins in the second third showed no significant effect. The authors suggest the lack of influence from the second third may be due to the higher player density which makes it more difficult to successfully complete passes, dribbles, and ball receptions. While plausible, it can also be argued that identifying effective passing or dribbling options could increase the likelihood of successful outcomes, leading to a potential influence. One limitation of the study is that second balls are not rigorously defined and only referred to as a possession gain after an aerial or ground duel. It is unclear what constitutes a possession gain and whether soccer practitioners would agree that second balls follow a ground duel. However, the paper provides evidence of the importance of second ball wins in team success and helps motivate the research in this thesis.

Hang [3] uses the same definitions for second balls and second ball wins as Definition 1 and Definition 2. Hang then creates the novel idea of second ball chains. Hang describes a second ball chain as the events/actions leading to the second ball. That is, the long ball, intervention, and possession gain/regain. The following are the original chains created in [3]:

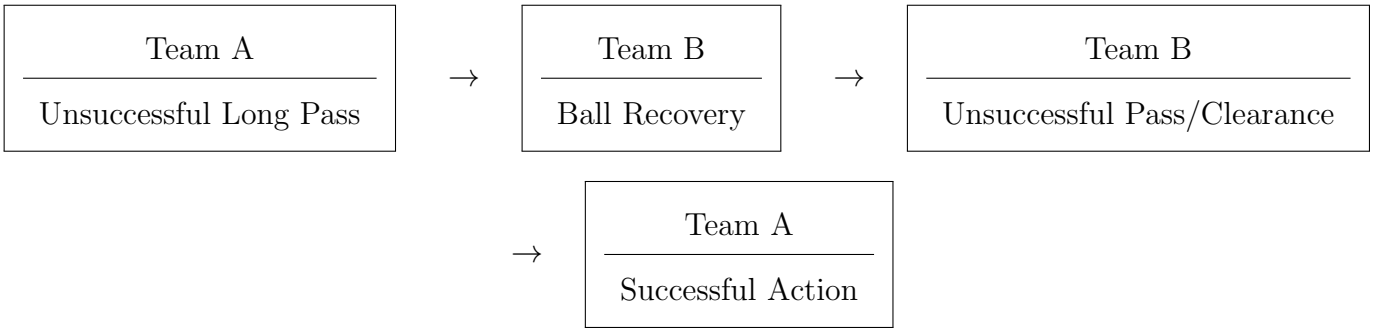
Chain ABA



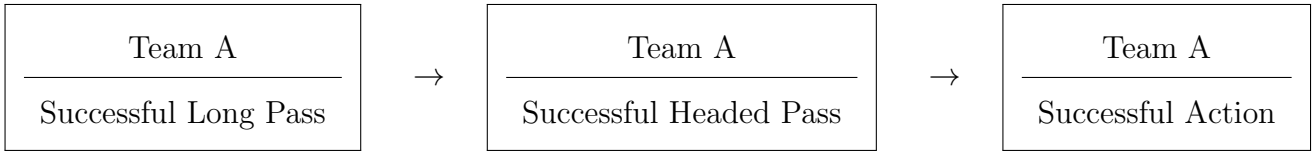
Chain ABABA



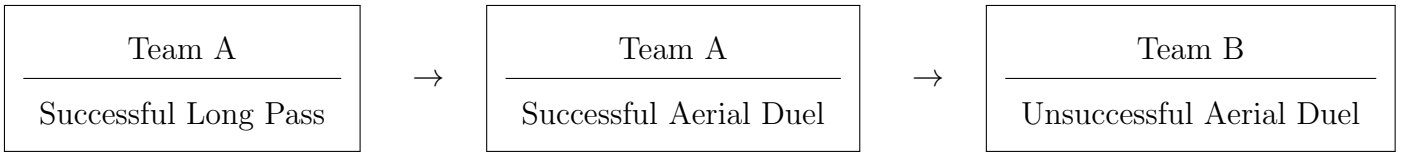
Chain ABBA

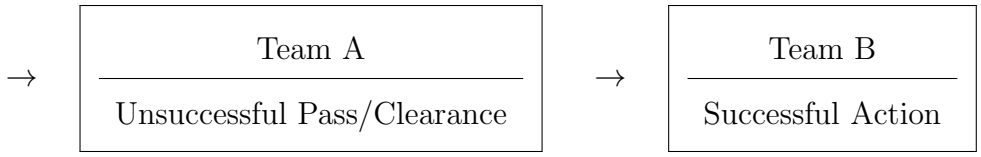


Chain AAA

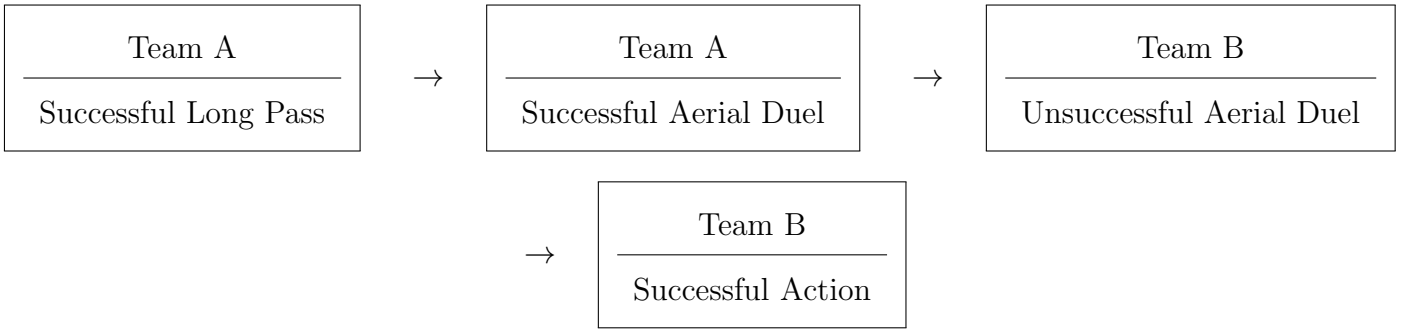


Chain AABAB

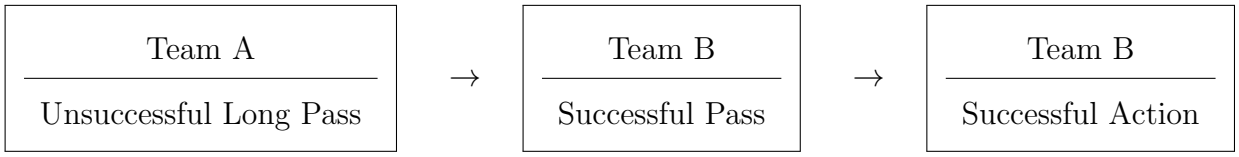




Chain AABB



Chain ABB



Hang's framework effectively captures how second balls occur by documenting the actions leading to a possession gain. As a result, these structured chains enable algorithmic extraction from data. Hang further classifies these seven chains into open-play and set-piece scenarios, doubling the total to 14 second ball chains. Also, it is important to note that Hang counts a second ball win only if the player who wins the second ball completes a successful action. In my model, to analyze possession meaningfully, we will count a second ball win if the player's action and subsequent action are both successful. In Hang's case [3], he does not analyze what happens after a second ball win. His framework stops at possession gain, meaning the resulting possession and potential attacking opportunities remain unexplored. Additionally, Hang acknowledges the importance of third-balls – a loose ball after two consecutive aerial contests – but again has not explored them. This work aims to build upon Hang's foundation by examining both the

outcomes of second ball wins and incorporating third ball, fourth ball, and subsequent scenarios.

2.5 Summary

This chapter began by exploring the primary data sources used in soccer analytics - event and tracking data - which serve as the cornerstone for modern performance analysis. Second balls were defined, emphasizing their importance in possession and goal-scoring scenarios. However, second balls must formally be defined to quantify and extract them from the data. Finally, the literature review provided a comprehensive examination of existing research, highlighting many different player and team evaluation metrics and the gap in research related to second balls. The next chapter will introduce the methodology and modelling approaches used to quantify and analyze second balls.

Chapter 3

Methodology

This section details the approach used to quantify and analyze second balls. First, the current definition of second balls is updated to address the previous limitation. Next, a mathematical framework is introduced to formally define second ball chains and the subsequent possessions. Finally, the methodology details the data processing techniques, modelling approaches, and evaluation criteria used to assess second ball chains and possessions, as well as their predictive factors.

3.1 Second Ball Chains

In this section, we update the second ball chains in [3] to address three main problems.

1. Second Ball Wins
2. Naming Conventions
3. Higher-Order Balls

3.1.1 Second Ball Wins

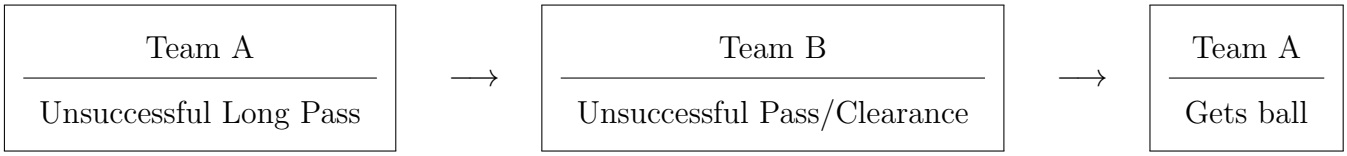
In [3], a second ball win occurs when the player who gets the second ball performs a *successful* action with it. That is, a completed pass, dribble, shot, or drawing a foul. The

problem with this definition is that it can be overly simplistic. In practice, the context around second ball situations is often chaotic and rushed. A player may complete a pass under pressure that does not meaningfully retain possession—for example, a pass that leaves the recipient in a poor position or immediately under pressure, leading to a turnover. In such cases, awarding a second ball win based solely on one successful action overstates the actual control or advantage gained. To address this, we define a second ball win as requiring not just one, but two consecutive successful actions by the team that gains the second ball. The second ball aligns with StatsBomb's definition of a possession being established after two successful actions. By doing so, it is hypothesized that my framework offers a more robust and meaningful control measure following a second ball situation, filtering out cases where an initial successful action does not truly lead to sustained possession. To clarify, my definition has already been defined in Definition 2, this just provides understanding as to why we have changed it from its original definition in [3].

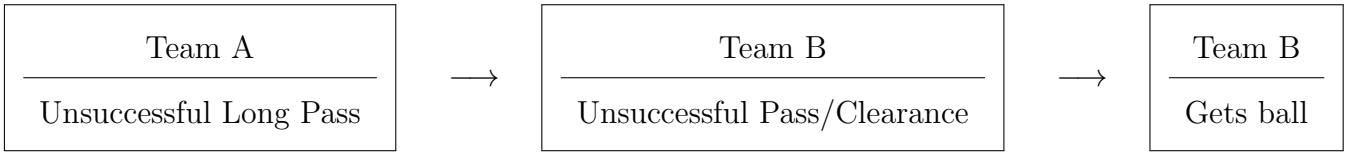
3.1.2 Naming Conventions

In the original chains, a duel between players would be in two separate blocks; that is, one block of the chain would be the successful team, and the next would be the unsuccessful team. The graphic of the chains makes it seem these events happen one after the other. However, the binary action of a duel, one team successful, one unsuccessful, occurs at the same time. To solve this, we create one block called Duel, with the winning team included. Also, we aim to simplify the readability, as chain ABABA is the opposite of chain AABAB, yet it is not clear by the naming convention. The following are the six new second ball chains we will use in this thesis:

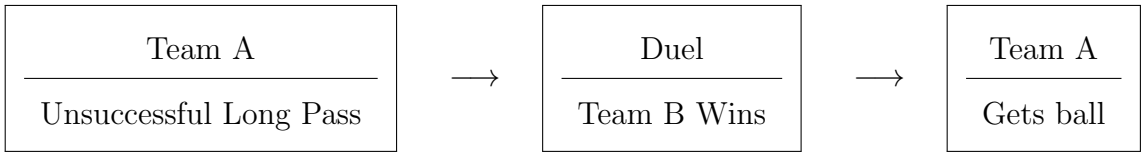
Chain ABA



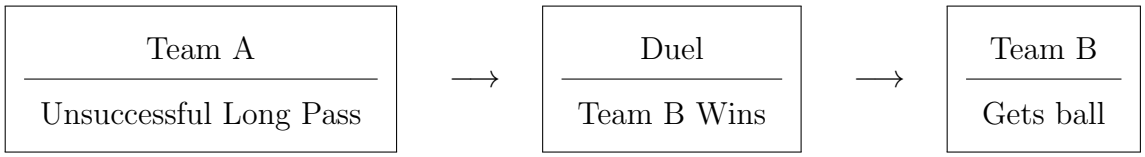
Chain ABB



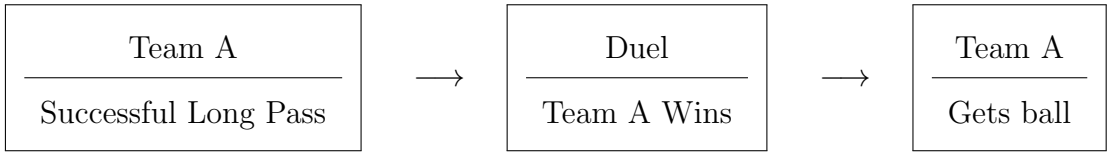
Chain ADBA



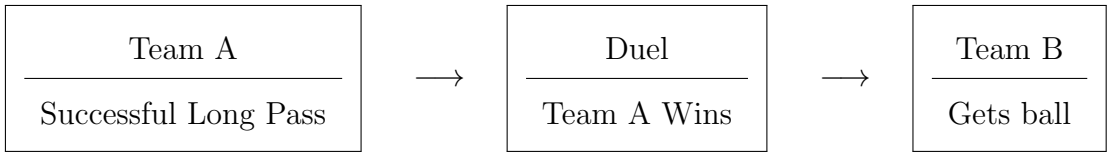
Chain ADBB



Chain ADAA



Chain ADAB



It is important to note that a second ball chain is just the events that lead up to the second ball. It does not include the winner of the second ball. These chains are designed to help identify second ball occurrences within a dataset. The other reason for the change in the naming convention is that it is now clear when a second ball contains a duel. For example, ABA remains unchanged in both versions. Team A long ball, Team B pass/clearance, and Team A collects the second ball. Comparing this to the ABABA chain from [3], logically it would make sense to think the following: Team A long ball, Team B pass/clearance, Team A pass/clearance, Team B pass/clearance, and Team A collects the second ball. However, this is wrong since the ABABA chain involves a duel between opponents. To solve this problem, a 'D' for duel is included into the chain naming. Whatever letter follows the 'D', is the team that wins the duel. For example, Chain ADAB is read as, Team A long ball, Duel won by Team A, Team B collects the second ball. The newly created naming conventions make it easier to understand the second ball chains.

3.1.3 Higher-Order Balls

The final point addressed from the work in [3] is that it only follows a set of very strict chains and leaves out other scenarios that might be of interest to my analysis. For example, chain ADBA means the following:

1. Team A long ball
2. Duel, that is won by Team B
3. Team A gets the second ball

We also know that if Team A gets the second ball and completes two successful actions, they are rewarded with a second ball win. However, what if upon getting the second ball, Team A receives the second ball, makes a completed pass, however, Team B applies pressure immediately, forcing Team A to turn the ball over, resulting in Team B completing

two consecutive successful actions and gaining possession. This situation is not accounted for, along with many other scenarios where the ball hops back and forth between teams until someone establishes possession. Although these scenarios are not exactly second balls, we still want to include them in the analysis since they serve the same purpose; helping to understand the relationship between chaotic interactions after a long ball and how they affect goal-scoring opportunities, possession, and game dynamics. To account for these newly introduced scenarios, we must define them in a way that represents their nature. We call them Higher-Order Balls and use the following definitions to describe them:

Definition 3 (Transition Window) *A fixed time period (e.g., 5 seconds) following when a second ball occurs, during which both teams have the opportunity to establish possession.*

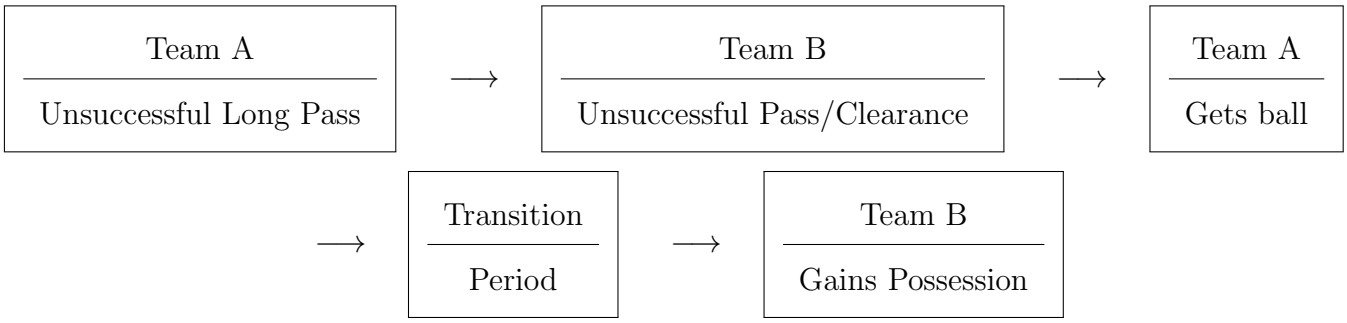
Definition 4 (Higher-Order Balls) *A higher order ball refers to a scenario where a team initially gains the second ball but fails to establish immediate possession (i.e., does not complete two consecutive successful actions right away).*

Definition 5 (Higher-Order Ball Win) *A Higher-Order Ball Win occurs when a team is able to establish possession during the transition window after a second ball occurs.*

The terms are defined since, in theory, you can have a third-ball, fourth-ball, etc.; however, due to the nature of the analysis done in this thesis, higher-order balls and wins are referred to as second balls and second ball wins. Second balls and higher-order balls are formally defined in the next section, but a way to differentiate between a second ball win and a higher-order ball win is still required. Twelve new chains are introduced, the same as the original six, but each with the suffix 'TA' or 'TB'. These suffixes read 'T' to indicate a transition window, and 'A' or 'B' depending on which team establishes

possession before the end of the window. For example, the following chain represents ABATB:

Chain ABATB



These new kinds of chains, which include the transition period, are applied twice to each of the existing chains and are shown in Figure 3.1. Therefore, there are 12 new chains, 18 in total.

Chain Type	Description
Second Ball Chains	
ABA	Team A long ball, Team B pass/clearance, Team A gets ball
ABB	Team A long ball, Team B pass/clearance, Team B gets ball
ADBA	Team A long ball, Team B wins duel, Team A gets ball
ADBB	Team A long ball, Team B wins duel, Team B gets ball
ADAA	Team A long ball, Team A wins duel, Team A gets ball
ADAB	Team A long ball, Team A wins duel, Team B gets ball
Higher-Order Chains (with Transition)	
ABATA	ABA, transition period, Team A establishes possession
ABATB	ABA, transition period, Team B establishes possession
ABBTA	ABB, transition period, Team A establishes possession
ABBTB	ABB, transition period, Team B establishes possession
ADBATA	ADBA, transition period, Team A establishes possession
ADBATB	ADBA, transition period, Team B establishes possession
ADBBTA	ADBB, transition period, Team A establishes possession
ADBBTB	ADBB, transition period, Team B establishes possession
ADAATA	ADAA, transition period, Team A establishes possession
ADAATB	ADAA, transition period, Team B establishes possession
ADABTA	ADAB, transition period, Team A establishes possession
ADABTB	ADAB, transition period, Team B establishes possession

Table 3.1: Second Ball and Higher-Order Chains and their descriptions.

The final clarification that will give us a complete working terminology is to differentiate between second balls and second ball wins. We can add the suffix 'w' to a second ball chain to signify that a team establishes possession. For example, ABAw means, Team A long ball, Team B pass/clearance, Team A receives ball *and* establishes possession. Note for the higher-order chains, due to scope, I am only looking at the scenarios where possession is established. Therefore, the last letter in the chain tells which team establishes possession.

3.2 Second Ball Possessions

Now that the newly updated second ball chains are established, what happens after a second ball win: a *second ball possession*, must be introduced. Since one of my goals is not only to quantify second balls, but also analyze how they affect possession and goal-scoring chances, what a team does with the ball after winning the second ball must be tracked. The following is defined:

Definition 6 (Second Ball Possession) *The sequence of events after a second ball win until there is a stoppage in play or the opposing team establishes possession.*

Keeping track of possession after the second ball win is straightforward, as a simple check to see if a goal, foul, out-of-bounds, stoppage, or possession change occurs. StatsBomb has contextual information on each of these options for every event. With the understanding of second balls redefined and the concepts of higher-order balls and second-ball possessions introduced, the framework can now be formally defined.

3.3 Mathematical Definition of Second Balls

To build towards a model to analyze second ball situations, it is important to define them mathematically. Defining second balls rigorously enables their frequency to be quantified, their outcomes to be tracked, and predictive models to be developed.

3.3.1 Formal Definition

Let's define a second ball sequence S that begins at time t_0 when a second ball occurs. Let T be a team and W_s be a second ball win.

$$W_s = \begin{cases} 1, \text{ if T completes 2 consecutive passes within } t_0 + \tau \\ 0, \text{otherwise} \end{cases} \tag{3.1}$$

where τ is the transition window. The transition period is defined as:

$$T_{trans} = [t_o, t_f] \tag{3.2}$$

where $t_f = min(t_{S_A=2}, t_{S_B=2}, t_0 + \tau)$ and $t_{S_x=y}$ is the time t, at which team x makes y successful actions.

Once t_f is reached, the possession of the winning team is tracked, let's say P_T, until one of the following:

- Possession loss, that is, the opposing team completes two consecutive successful actions, a foul, an out of bounds, or any other stoppage.
- Goal

Formally:

$$P_T = \{a_{T_1}, a_{T_2}, \ldots, a_{T_n}\} \tag{3.3}$$

where a_{T_i} is the ith successful action by team T. Therefore, a_{T_n} is a possession loss or a goal.

3.3.2 Second Ball Win and Possession Example

Let us look at an example to visualize the mathematical definition;

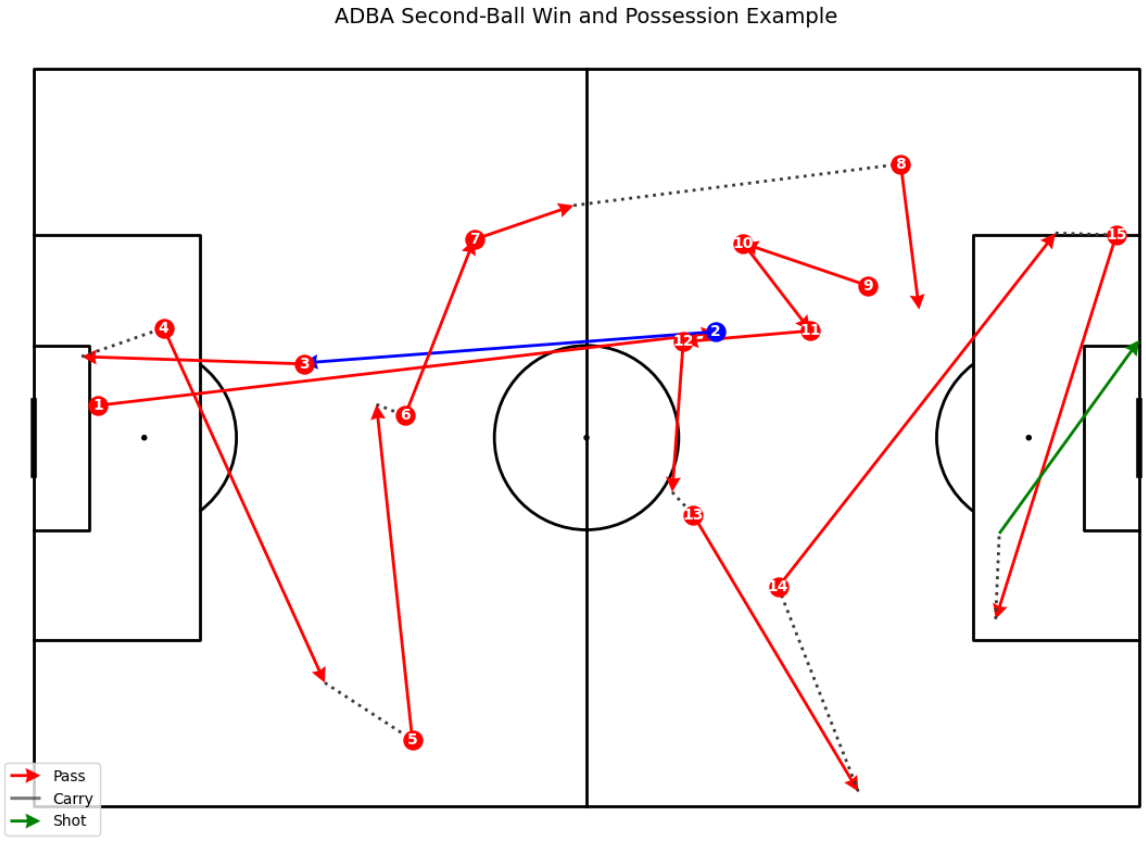


Figure 3.1: A passing map that shows an ADBA second ball win (1-3) followed by the resulting possession (4-15), which ends with a shot off target.

Figure 3.1 visualizes just one example of the endless second ball win and possession possibilities. It can be seen that Team A (red) clears the ball, most likely by goal-kick, where Team B (blue) wins the duel and sends the ball back towards Team A. Team A then collects the second ball and proceeds to maintain possession up the field where the possession ends with a shot that is off target. The example is included to help provide clarity to the reader.

3.4 Model Overview

To answer the research questions, a model is proposed to investigate key factors and predictors of second ball wins and the following possessions. To achieve an explainable and holistic model, the model is split into three components. First, the location prediction component predicts where the second ball is most likely to occur given a list of spatial and contextual features. Next, the second ball winning team prediction component predicts which team will win the second ball given a list of spatial and contextual features. Last, Markov chains are used to analyze the probability of scoring from the second ball win. The final metric, Expected Second Ball Value (xSBV), combines these components into a model that can be used to analyze second ball wins. The xSBV is defined as the product of these three components:

$$\text{xSBV} = \underbrace{P(L)}_{\substack{\text{Location} \\ \text{Probability}}} \times \underbrace{P(W \mid L, \mathbf{X})}_{\substack{\text{Win} \\ \text{Probability}}} \times \underbrace{\Delta_{i,j} P(G_H \mid W)}_{\substack{\text{Gain} \\ \text{Difference}}} \tag{3.4}$$

where:

- L is the predicted location of the second ball,
- W denotes the winning team (binary outcome: $W = 1$ for Team A, $W = 0$ for Team B),
- $\mathbf{X}$ represents contextual features (e.g., player positions, duel type),
- $\Delta_{i,j} P(G_H \mid W)$ is the difference in probability of a goal occurring within H transitions after a second ball win, W, from location i to j.

3.4.1 Pitch Discretization

To simplify modelling and interpretability, the pitch is partitioned into a 4x6 grid. A 4x6 grid was chosen to find a middle ground between spatial detail and computational

simplicity. Dividing the pitch into four rows provides a basic distinction between wide and central areas. Dividing the field into six columns allows for analysis across the defensive, middle, and attacking thirds, while providing more detail than a simpler three-column division.

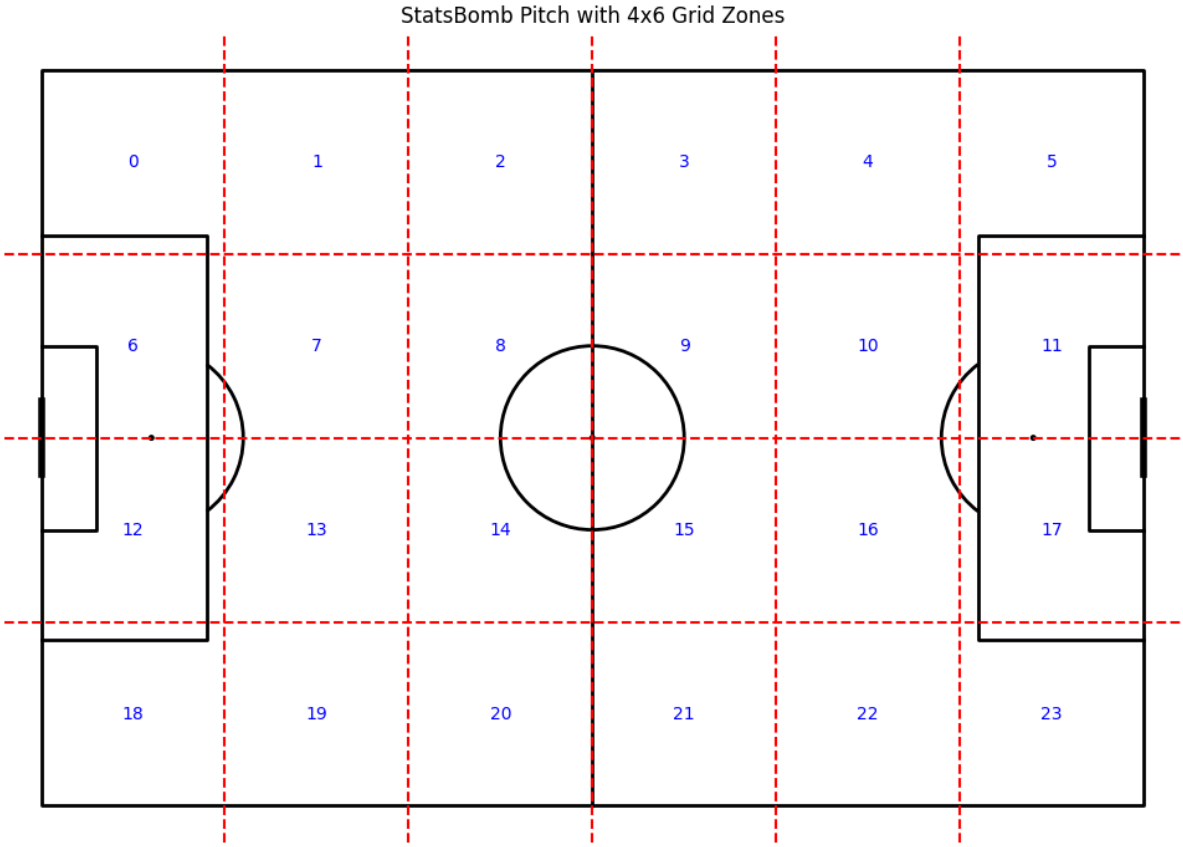


Figure 3.2: 4x6 grid used to partition the soccer pitch into zones for simpler modelling and analysis.

3.5 Data Preparation and Features

3.5.1 Datasets

This thesis makes use of multiple datasets from StatsBomb's open data [8]. The datasets used are the 2024 UEFA Men's European Championship dataset, which includes both standard event data and StatsBomb 360 data, and the 2015/2016 league data from the

top five European leagues, which contains only event data. The top five leagues are the English Premier League, La Liga (Spain), Bundesliga (Germany), Serie A (Italy), and Ligue 1 (France). Including 360 data allows for a deeper understanding of certain second ball scenarios. However, since 360 data is relatively scarce, especially across full seasons, the much larger 2015/2016 dataset provides the bulk of the data used for large-scale modelling and analysis.

Country/Region	**Competition**	**Season**	**Event data**	**360 Data**	**Games**
Europe	UEFA European Championship	2024	Yes	Yes	51
England	English Premier League	2015/2016	Yes	No	380
Spain	La Liga	2015/2016	Yes	No	380
Germany	Bundesliga	2015/2016	Yes	No	306
Italy	Serie A	2015/2016	Yes	No	380
France	Ligue 1	2015/2016	Yes	No	380

Table 3.2: Summary of Datasets Used

3.5.2 Data Processing

To prepare the dataset for modelling, a custom second ball extraction pipeline was created based on both theoretical definitions and empirical observations, detailing the process. To begin the process, around 20 matches were manually watched and annotated to identify clear second ball situations. These annotated sequences were used to derive consistent patterns in event data (e.g., long ball → duel → recovery), which informed the design of automated extraction rules. Based on the patterns, custom functions were implemented to scan the event data for second ball sequences matching this logic.

Long balls were defined as passes or clearances that travelled more than 20 yards in the air. Clearance distance was not included in the data, so the Euclidean distance between locations of the clearance event and the following event was calculated and used. Following a long ball, subsequent events were scanned for second ball patterns

(e.g., ABA, ADBA, etc.) using a five-second transition window, with possession defined as completing at least two successful actions. Only cases where a team successfully established possession were included.

One problem faced was that whichever team performs an action with the ball, the coordinate system is defined so that the team performing the action is attacking from left to right. This is problematic when analyzing second balls since the ball can go back and forth between teams many times in a short period, causing the locations of some events to not be relative to the locations of other events. To ensure directional consistency across samples, coordinates were flipped so that the team initiating the long ball always attacked from left to right. This flipping simplifies model training and spatial interpretation and can be described with the algorithm below.

Algorithm 1 Normalize Event Coordinates to Left-to-Right Attacking Direction

```
procedure FlipCoordinates(events)
    for each event in events do
        if event.team ≠ team_a then
            event.x ← 120 − event.x
            event.y ← 80 − event.y
    return events
```

For reproducibility, the full code used for extracting second ball wins is included in Appendix A.

3.5.3 Data Visualization

To support modelling of second balls, it was important to understand the spatial characteristics of the data. Visualizing the distributions of long ball and second ball locations allowed for an understanding of where these events typically occur on the pitch. Location distribution plots guided modelling choices, such as pitch discretization and feature

engineering. These figures not only aid in validating the data pipeline but also provide intuitive context for the probabilistic outputs of the models.

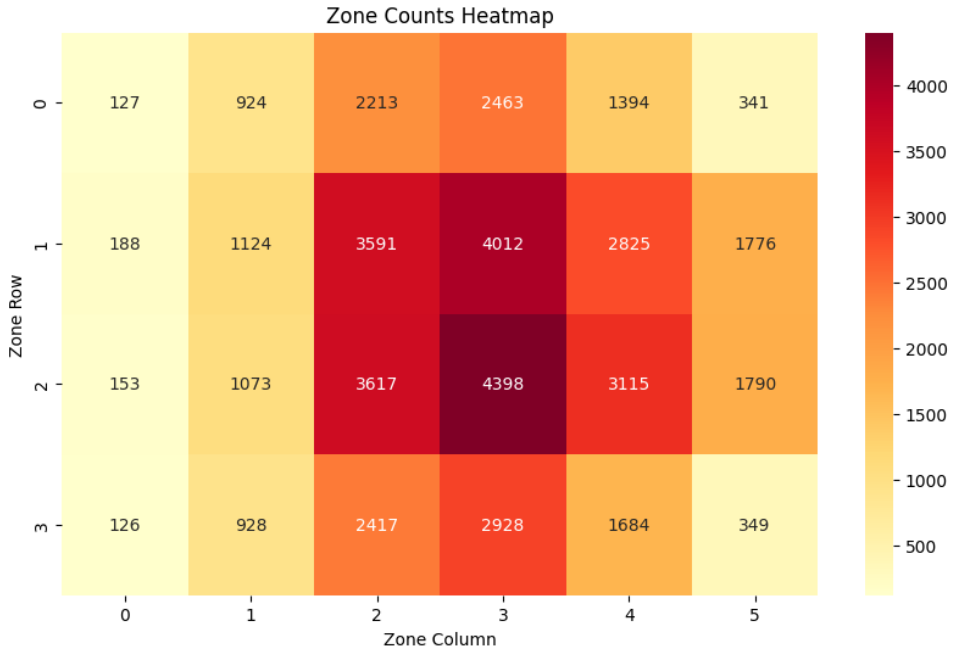


Figure 3.3: Distribution of where second balls occur by zone. Visualization to understand where on the field high risk vs low risk zones occur.

Visualizing the second ball locations using a heat map makes it clear to understand the distribution of where second balls occur in soccer. The following statements can be observed:

1. Second balls in the defensive third are low when compared to other thirds.
2. The middle third dominates the locations of second balls.
3. Second balls are less likely to occur in wide areas.

3.5.4 Feature Engineering

Thirteen features were initially created to capture the spatial, temporal, and contextual dynamics of second ball situations. These features were derived from the event data,

and were designed to provide relevant information for each of the three components of the xSBV model. The features include variables related to ball position and trajectory, player proximity, field zones (based on pitch discretization), and other game contextual information. Each feature was selected to contribute to one or more of the components, while maintaining generalizability across game contexts. The complete feature set was standardized or encoded as appropriate and was used consistently across model training and evaluation unless explicitly excluded in ablation testing.

Feature Name	**Description**
pass_start_x, pass_start_y	x and y coordinates of the starting location of the initial long ball.
pass_end_x, pass_end_y	x and y coordinates of the ending location of the initial long ball.
pass_distance	Length of the initial long ball.
pass_angle	Angle in radians of the initial long ball to the goal.
dx, dy	Change in x and y coordinates of the start and end location of the initial long ball.
dist_to_centre	Distance of the end location of the initial long ball to the centre of the field.
is_defensive, is_midfield, is_attacking	Boolean flags to indicate what third of the field the end location of the long ball occurs in.
is_contested	Boolean flag to indicate if the initial long ball is met with a duel, e.g. ADBA, ADAA, ADAB, or ADBB.
second_x, second_y	x and y coordinates of the second ball location.
has_transition	Boolean flag to indicate if there is a transition period in the second ball win or not.

Table 3.3: Features used in the xSBV model, with corresponding feature names and descriptions.

3.6 Location Prediction

Objective: Predict the zonal location of where a second ball occurs given a list of features.

3.6.1 Model Formulation

Let p_0 be the initial duel location and $X_{context}$ be a list of features. The second ball location, L, is modelled as a stochastic outcome:

$$P(L_z|p_0, X_{context})$$

where L_z is the corresponding zone that location L belongs to and $X_{context} = \{$pass_start_x, pass_start_y, pass_distance, pass_angle, dx, dy, dist_to_centre, is_defensive, is_midfield, is_attacking, is_contested$\}$

3.6.2 Class Imbalance

One challenge in predicting second ball locations is the imbalance in the distribution of second ball locations across the dataset. Certain zones, particularly central and middle-third areas, are far more likely to have second balls occur than wide or extreme defensive/offensive zones. The imbalance can bias machine learning toward overpredicting high-frequency zones and underrepresenting less common but tactically important areas. To address this potential bias, class weights are used during model training to penalize errors on underrepresented classes more heavily. The weight for each class c was calculated using the inverse frequency formula [47]:

$$w_c = \frac{N}{C \cdot N_c}$$

where N is the total number of training samples, C is the number of classes (24

zones), and N_c is the number of samples in class c. This weighting ensures that the loss function reflects a more balanced treatment of each class during training. Although applying these weights can lead to a decrease in overall accuracy, it improves fairness across zones by encouraging the model to consider the full spatial distribution of second balls.

3.6.3 Modelling Techniques

The location prediction task is a multi-class classification problem [48], where the goal is to predict the pitch zone where the ball is most likely to land after a contested aerial action or clearance. Each instance is assigned a class label based on the observed ball destination. To model this, the primary technique used is Extreme Gradient Boosting (XGBoost), a powerful gradient-boosted decision tree algorithm [49]. Due to the multi-class nature of the task, the model is trained using softmax objective with log-loss as the optimization criterion. Model performance is evaluated using top-k accuracy metrics [50], specifically top-1 and top-3, to assess how often the true location falls within the top predicted zones.

3.7 Second Ball Winning Team Prediction

Objective: Estimate the probability that a given team wins the second ball at location L.

3.7.1 Model Formulation

A binary classifier predicts:

$$P(W = Team_A | L, X_{context})$$

where W denotes a second ball win and $X_{context}$ = {pass_start_x, pass_start_y, pass_distance, pass_angle, dist_to_centre, is_defensive, is_midfield, is_attacking, is_contested, second_x, second_y, has_transition}

3.7.2 Class Distribution

Once again, looking at the class distribution can help reveal imbalances in the underlying data.

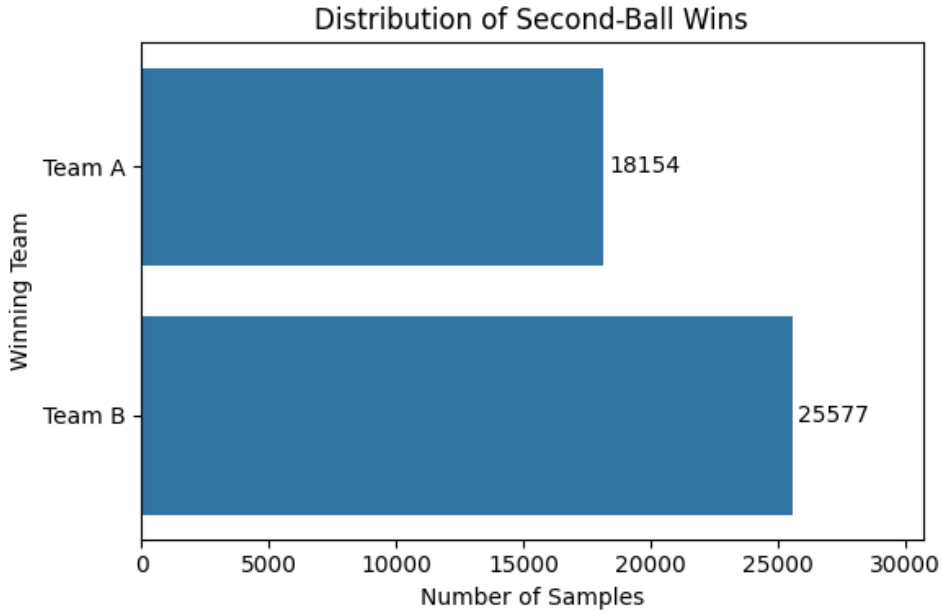


Figure 3.4: Distribution of second ball wins. Remember that Team A always initiates the long ball that leads to the second ball.

In Figure 3.4, Team B is more likely to win the second ball. Since the imbalance is not large and the balancing techniques for binary classification hardly impact accuracy [51], the model is trained and tested without balancing.

3.7.3 Modelling Techniques

The task of predicting which team will win a second ball is formulated as a binary classification problem [52], where the model must determine whether the team initiating

the long ball or the opposing team will gain possession following the contest. Three machine learning models are used to address this problem: XGBoost, Random Forest, and Logistic Regression. Random Forest is an ensemble method that combines multiple decision trees to reach a single result, offering resistance to overfitting, especially when the feature space is large and noisy [53]. Logistic Regression provides a simple and interpretable model, relying on a linear combination of input features passed through a sigmoid function to estimate the probability of each class [54]. Comparing these models allows for an evaluation of accuracy and how each handles second balls differently.

3.8 Gain

To quantify the probability that a goal occurs in the possession after a second ball win, a Markov model called *gain* is proposed. This approach is similar to [25], [24], and [32], but adjusted to my problem, which captures both immediate and transitional danger of second balls.

3.8.1 Markov Chains for Possession Transitions

A Markov chain is a probabilistic model that represents a system transitioning between discrete states, where the probability of moving to the next state depends only on the current state and not the sequence of events that precedes it [55]. The assumption that the next state depends only on the current state, known as the Markov property, is a simplification that enables efficient modelling of second ball possessions. In reality, past actions may influence future outcomes, but this assumption is reasonable in a soccer context. For instance, there are many ways a team might advance the ball into the attacking third. However, once the ball is in the attacking third, the likelihood of scoring or losing possession is mainly determined by the current position and state of play, rather than the actions that led there. Absorbing states are such that once transitioned into,

they cannot be left. In my context, there are two absorbing states: a goal and the end-of-possession. Transitions represent the movement of the ball between states. A transition matrix is constructed using observed transitions from the data, and convergence is used to estimate the long-term trends of the absorbing states.

$$P = \begin{bmatrix} p_{z_0,z_0} & p_{z_0,z_1} & \cdots & p_{z_0,z_{23}} & p_{z_0,z_g} & p_{z_0,z_{eop}} \\ p_{z_1,z_0} & p_{z_1,z_1} & \cdots & p_{z_1,z_{23}} & p_{z_1,z_g} & p_{z_1,z_{eop}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p_{z_{23},z_0} & p_{z_{23},z_1} & \cdots & p_{z_{23},z_{23}} & p_{z_{23},z_g} & p_{z_{23},z_{eop}} \\ 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \tag{3.5}$$

The Markov chain framework is well-suited to this problem due to its interpretability, simplicity, and ability to capture the stochastic nature of ball progression in soccer. Formally, the possession evolution is modelled as a Markov process where:

- **Transient States**: The 24 grid zones $\{z_1, \ldots, z_{24}\}$
- **Absorbing States:** Goal and End-of-Possession $\{z_g, z_{eop}\}$
- **Transitions**: Probabilities $P(z_j|z_i, W)$ are estimated from historical data, conditioned on the winning team W.

3.8.2 Convergence of the Transition Matrix

Each multiplication of the transition matrix P by itself represents an additional step in the possession sequence. Specifically, P^n describes the probabilities of reaching any given state after n transitions [56]. For second ball possessions, the probability of a possession eventually reaching one of the absorbing states is of interest. By repeatedly multiplying the matrix by itself, these probabilities change and eventually stabilize. This process is known as *convergence*. In practice, the matrix is repeatedly multiplied by itself until

there is a small enough state difference between iterations. The difference threshold value is arbitrarily chosen, and in this work 0.025 is used. Convergence allows for the long-term distribution of second ball possessions to be determined.

3.8.3 Model Formulation

Horizon-Limited Markov Chain

Let $\mathbf{P_t} \in \mathbb{R}^{24\times24}$ denote the probability transition matrix between transient states, and let $\mathbf{P}_g \in \mathbb{R}^{24\times1}$ denote the column vector of transition probabilities from each transient state into the absorbing state "Goal". For a fixed horizon H, the cumulative probability of absorption in the Goal state within H steps, or *gain*, is

$$P(G_H) = (\sum_{h=1}^{H} P_t^h)P_g \tag{3.6}$$

Where:

- $\mathbf{P_t^h}$ gives the distribution of transient states after exactly h steps,
- multiplying by $\mathbf{P}_g$ projects these distributions onto the probability of transitioning to Goal,
- summing over h accounts for scoring at any step up to horizon H.

Gain Difference

The *gain difference* $\Delta_{i,j}P(G_H)$ is defined as the difference between the cumulative probability of scoring a goal within H steps of zones:

$$\Delta_{i,j}P(G_H) = P(G_H)[j] - P(G_H)[i], \tag{3.7}$$

where $P(G_H)[n]$ is the nth element of $P(G_H)$, i.e., the probability of scoring within H steps when starting in zone z_n.

Expected Gain

Let $\mathbf{P_{first}} \in \mathbb{R}^{24\times24}$ be the empirical distribution of the *first successful team action* destination conditioned on the start zone. Multiplying the first-action matrix with the expected scoring value and then subtracting the probability of scoring in H steps from starting zone z_i gives the expected gain for the first action following a second ball win.

$$xP(G_H, z_i) = P_{first}P(G_H) - P(G_H)[i] \tag{3.8}$$

Comparing expected gain to the actual gain players receive from second ball wins provides insight into which players underperform or overperform relative to expectations.

3.9 Summary

This chapter detailed the methodology for quantifying and analyzing second ball situations in soccer through the novel xSBV (Expected Second Ball Value) model. The section began by updating the previous definitions of second balls, as well as introducing second ball possessions. Next, second balls were formally defined to allow for extraction throughout the data sets. The xSBV model was introduced as a three-component pipeline: (1) a location predictor to estimate where the second ball will land, (2) a model to predict which team is likely to win the second ball, and (3) a Markov chain model to evaluate the gain from the first pass after a second ball win. Each component was mathematically formulated and trained using a set of features derived from the event data. This framework lays the foundation for the performance assessments and tactical insights presented in the following chapter.

Chapter 4

Results

This chapter presents the evaluation of the proposed xSBV framework. The goal is to assess the performance and interpretability of the three components of the model: the second ball location predictor, the winning team prediction model, and the gain estimator that uses Markov chains. The results are structured as follows. First, the predictive performance of each component using standard classification metrics is reported, including accuracy and area under the receiver operating characteristic (AUC-ROC) scores. Ablation testing is then performed to analyze the contribution of individual feature sets and which help the models perform most efficiently. Next, the internal logic of the models is explored using SHapley Additive exPlanations (SHAP) to identify the most influential features and understand how they affect predictions. Finally, player and team-level analysis is provided, along with tactical insights that emerge from the xSBV outputs, highlighting practical value and use cases of the framework for performance evaluation and player identification in soccer. This chapter is aimed to demonstrate not only the predictive quality of the components but also their interpretability for understanding the role of second balls in a soccer.

4.1 Component Performance

4.1.1 Location Prediction

Top-1, Top-3 and Log Loss

Rather than relying solely on the top-1 predicted location for second balls, this work adopts a top-3 prediction strategy. In the context of soccer, the landing zone of a second ball is uncertain due to chaotic nature. By considering the three most probable zones, a more realistic representation of how the model is thinking and where it is predicting the second balls to land is captured. Additionally, top-3 evaluation reduces the harshness of strict classification accuracy and better reflects the model's value in practical settings where anticipating a region, rather than a single zone, can inform decision-making.

In addition to accuracy-based metrics, logarithmic loss (log loss) helps evaluate the model's predictions. Log loss penalizes incorrect predictions based on their confidence [57]. A model that assigns high probability to the correct class will receive a lower (better) log loss score, whereas overconfident incorrect predictions are heavily penalized. This metric provides a more nuanced view of model performance, where each prediction involves 24 possible zones.

Model	Top-1 Accuracy	Top-3 Accuracy	Log Loss
XGBoost before ablation	0.2119	0.5937	2.6328
XGBoost after ablation	0.2382	0.6022	2.6712
Naive Baseline	0.1063	0.2849	-

Table 4.1: Performance comparison between XGBoost model and a naive baseline for second ball location prediction.

The pre-ablation XGBoost model achieved top-1 accuracy of 21.2% and top-3 accuracy of 59.4%. After ablation testing, accuracy improved further to 23.8% top-1 and 60.2% top-3. While these numbers indicate low prediction accuracy, they are notable given the complexity and novelty of the task. A naive baseline approach, simply pre-

dicting the most common zones, only has an accuracy of 10.6% top-1 and 28.5% top-3 accuracy. Significantly outperforming the naive baseline suggests the model is not random and has some ability to learn spatial patterns. Despite room for improvement, these results mark an important step forward in quantifying and modelling second ball outcomes in soccer.

Also, the model's performance was evaluated using log loss. The pre-ablation model had a log loss of 2.63, which increased to 2.67 after ablation testing. Although accuracy increased, so did log loss. One explanation could be that the model correctly classified more second ball locations, but also got worse at classifying already incorrect situations. Nevertheless, the log loss remains lower than what would be expected from random guessing, reinforcing that the model is capturing valuable information about predicting second ball locations. As a novel use of machine learning for second ball prediction, this approach establishes a foundation for future work. Further gains through improved data and modelling techniques would be a natural extension.

Confusion Matrix

To better understand how the model performs across individual zones, a confusion matrix is used. A confusion matrix visualizes the distribution of correct and incorrect predictions [58]. The confusion matrix is particularly useful in a multi-class setting like second ball location prediction, where certain zones may be more prone to misclassification due to spatial similarity or class imbalance. By analyzing the confusion matrix, systematic patterns in the model's predictions and insight into areas where the model struggles can be identified, helping to inform both model refinement and tactical interpretation. Figure 4.1 shows the number of times a particular zone is predicted by the model compared to the true zone of where the second ball occurred. A perfect model would only have numbers on the left to right downwards diagonal, that is, where the predicted zone is the correct zone of the second ball.

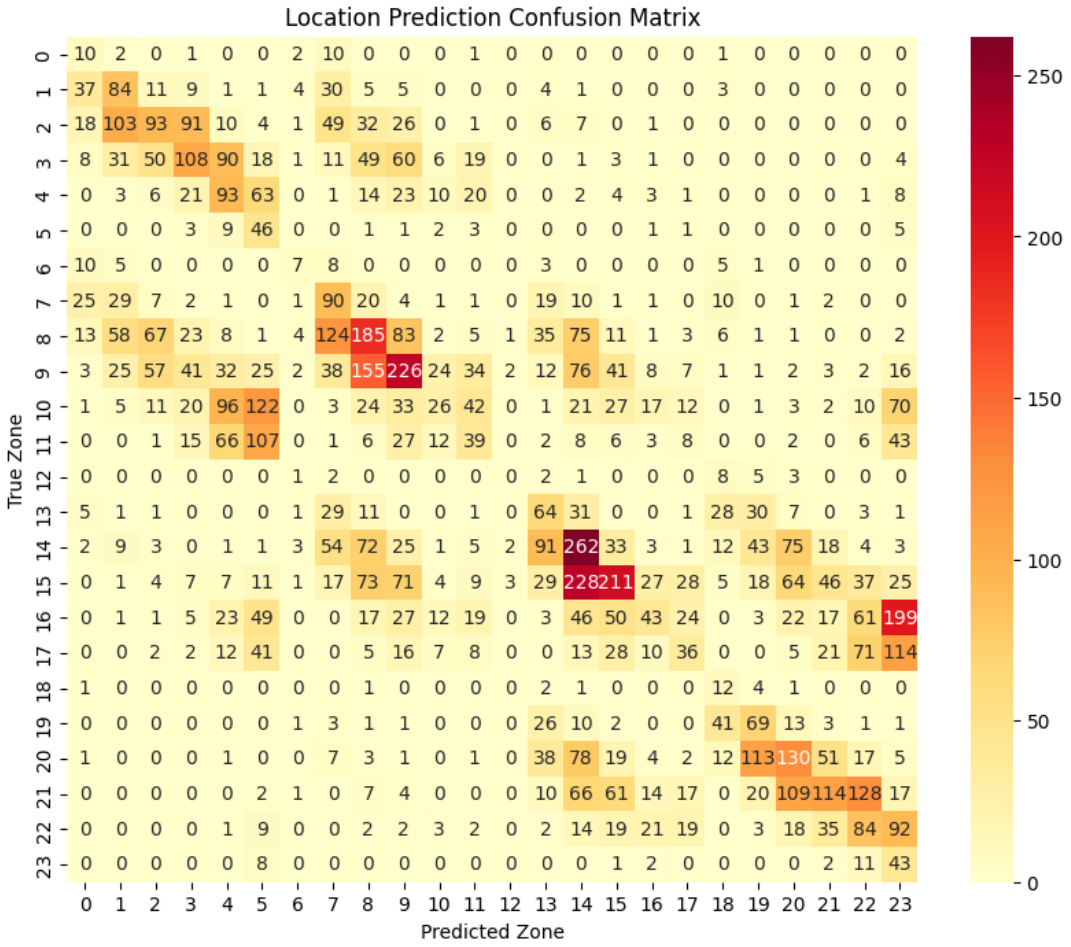


Figure 4.1: Heat map of the confusion matrix associated with the location prediction component.

The confusion matrix in Figure 4.1 is used to analyze the incorrect predictions and reveals some interesting insight about the location prediction component. The first is that the model can learn spatial relationships for second balls. It is intuitive that if the model predicts zone 14, then zones 13 and 15 also have larger-sized counts for the true zone, since geometrically these three zones are in a row. Less intuitive, but if you look up or down around six zones from the predicted/true zone diagonal for any given zone, you will see higher than normal counts as well. Based on the discretization of the pitch, a plus or minus six value is required to transition from one column to a different row on the field. This indicates that the model's most common mispredictions are in neighbouring zones, meaning the model has some understanding of spatial relationships.

Class-wise precision/recall/F1

Accuracy metrics such as class-wise precision, recall, and F1 scores were calculated for each zone. Precision measures how often the model is correct when it predicts a specific zone, while recall assesses how well the model identifies all instances of that zone [59]. The F1 score is the harmonic mean of precision and recall. Analyzing these scores across all zones allows us to pinpoint which areas of the pitch the model handles reliably and which zones present challenges, shedding light on both tactical and modelling implications.

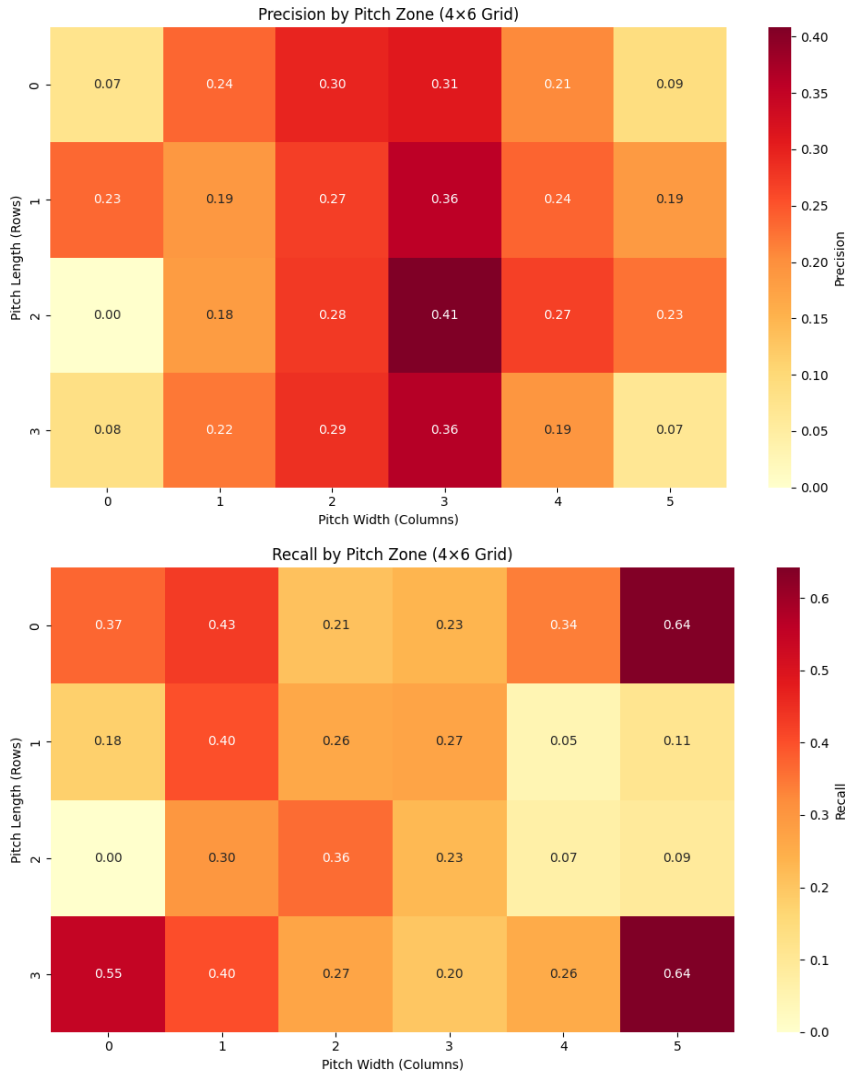


Figure 4.2: Precision and Recall of the final XGBoost model for the location prediction component. Top image shows the precision and bottom image shows the recall.

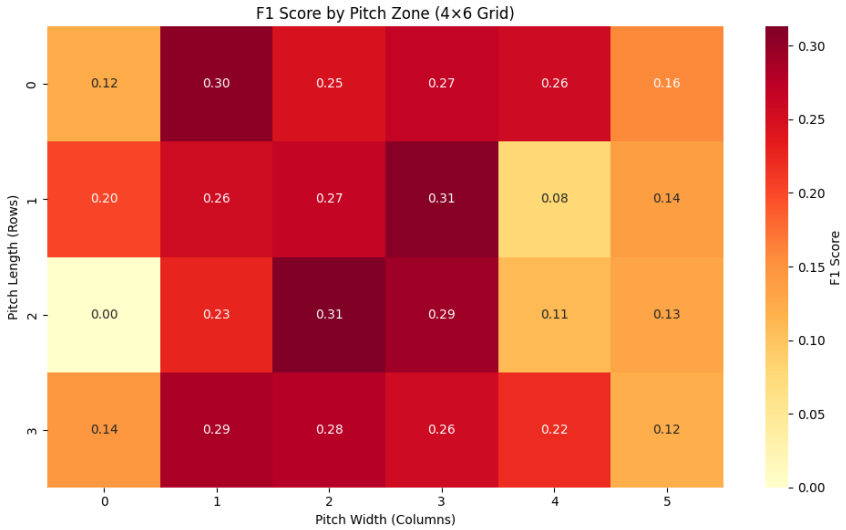


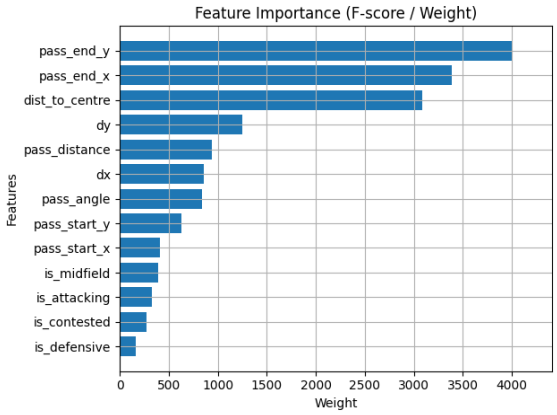
Figure 4.3: F1 score by zone showing the mean of precision and recall for each.

Figure 4.3 provides a visual to understand which zones the model is better and worse at predicting. When trying to understand zones where there is low performance, it is important to look at the precision and recall scores individually. Looking at both scores is important because if F1 is low, precision or recall could still be high. For example, when looking at Figure 4.3, it can be seen that zones 10 and 16 have low F1 scores. Comparing this to figure 4.2, it can be seen that the precision for these zones is higher, whereas the recall is very low. This means that when the model predicted zones 10 and 16, it had higher accuracy. However, the model incorrectly predicted another class many times when the true class was one of these zones. This suggests that to capture the whole relationship of second balls in central attacking areas, new features or model architecture must be advanced to improve the model's performance.

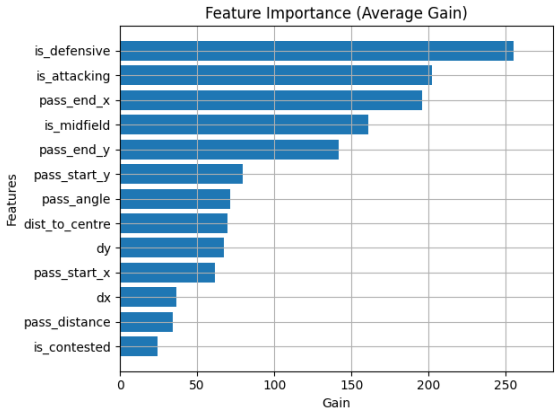
Interestingly, wide attacking positions such as zones 5 and 23 have low precision and high recall, meaning the model often identifies these zones, but frequently is incorrect in doing so. The change in high and low precision and recall in attacking zones suggests that the model lacks sufficient information based on the given features to predict these locations accurately. It sets up a direction for future work to investigate the nuances of second ball location prediction in more advanced chaotic zones of the field.

Ablation Testing

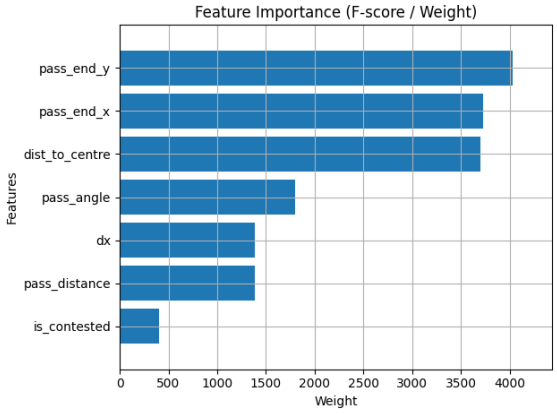
The contribution of each input feature to the model's performance is analyzed with ablation testing. Backward sequential feature selection systematically removes features one at a time and evaluates the model's performance without them, identifying the subset of features that yields the highest accuracy [60]. The testing focused on top-1 accuracy as the scoring metric, with the model retrained at each step to assess the impact of feature removal. Ablation testing helps eliminate redundant or noisy variables. The final reduced feature set produced a model with slightly higher accuracy, indicating that some features may have introduced unnecessary complexity.



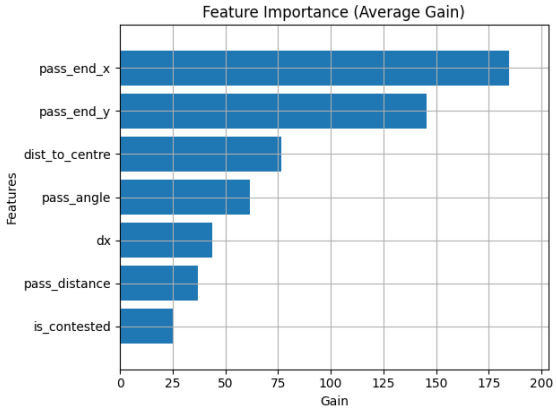
(a) F-score (All Features)



(b) Gain (All Features)



(c) F-score (After Ablation)



(d) Gain (After Ablation)

Figure 4.4: Feature importance for the location prediction component. Top row shows results with all features included; bottom row shows results after ablation testing. Left column uses F-score, right column uses average gain.

Figure 4.4 uses the following [61] to explain the feature importance:

- **F-Score (Frequency Score):** refers to the number of times a feature is used to split the data across all decision trees in the model. It gives a sense of how often a feature is considered useful by the model, but it does not account for the quality or usefulness of the split.

- **Gain:** measures the improvement in the model's performance relative to a feature.

Intuitively, the location of the initial duel contributes the most to the prediction ability of the model. This is an expected result since it is assumed the second ball location would be nearby. It is interesting to note that many features that the original model used and ranked highly, such as the boolean indicators for which third of the pitch the initial duel occurs in, are not used by the improved model. This could suggest that the final model only needs a few features to begin to understand the spatial relationship between zones. The next steps are to include other contextual factors such as tracking data to look at the distribution of players near the initial duel or even what team wins the duel to give more information as to where the ball will potentially go.

SHAP Analysis

The internal logic of how features contribute to the location prediction model is analyzed using SHAP (SHapley Additive exPlanations). SHAP offers a game-theoretic approach to explain the output of machine learning models by attributing a value to each feature's contribution [62]. Unlike feature importance rankings, SHAP provides local interpretability, allowing insight to not only feature importance, but also how each feature influences the predictions of the model. For second ball location prediction, SHAP helps uncover the spatial and tactical cues the model relies on to determine where the ball is likely to land.

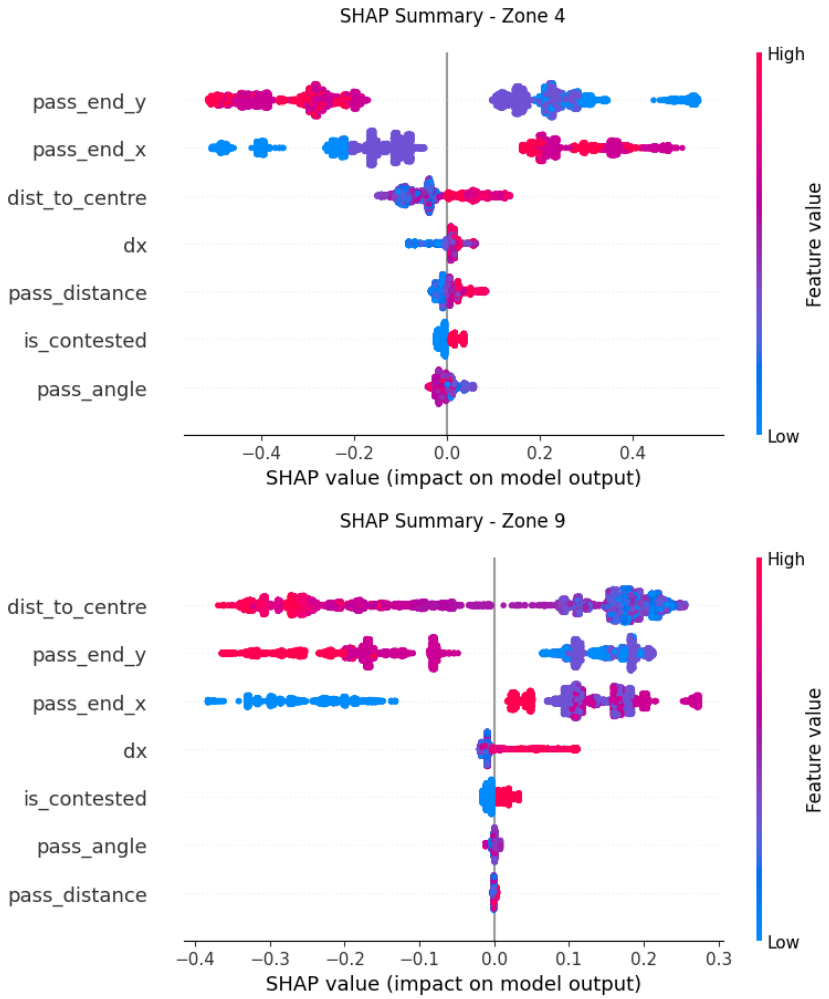


Figure 4.5: SHAP graphs for zone 4 (top) and zone 9 (bottom) describing feature impact on the model output.

Figure 4.5 shows how the model predicts zones. Zone 4 is a wide attacking zone that relies mainly on the x and y coordinates of the initial long pass. The larger the x coordinate and the smaller the y coordinate, the more the model is impacted, indicating that the second ball will occur in zone 4. Zone 9 follows the same logic, but given that it is more central, the model largely uses the distance to centre feature to help differentiate between the wide and central zones. SHAP summaries have been made for every zone, helping to get an idea of how features impact the model's predictions. However, only these two are included to build an understanding of how they work, without being too overwhelming.

4.1.2 Winning Team Prediction

Accuracy, AUC-ROC, and Log Loss Scores

To evaluate the performance of the team winning prediction component, three core metrics are reported: accuracy, log loss, and the area under the receiver operating characteristic curve (AUC-ROC). AUC-ROC offers a more nuanced evaluation by measuring the model's ability to distinguish between positive and negative classes [63]. An AUC-ROC score of 0.5 indicates random performance, whereas a score closer to 1.0 reflects strong discriminatory power. Log loss and accuracy are once again included to further help analyze the model performance.

Model	Accuracy	AUC-ROC Score	Log Loss
Logistic Regression	0.704	0.759	-
Random Forest	0.711	0.778	-
XGBoost	0.713	0.789	-
Logistic Regression after ablation	0.704	0.761	0.584
Random Forest after ablation	0.718	0.790	0.554
XGBoost after ablation	0.722	0.799	0.548
Naive Baseline	0.585	0.5	-

Table 4.2: Accuracy, AUC-ROC, and Log Loss scores for different model variants.

In Table 4.2, it is shown that although all three models perform similarly before and after ablation, XGBoost is marginally the best. It is more important that the naive baseline of always predicting team B as the winner of the second ball is outperformed. With a slightly imbalanced dataset, predicting team B every time results in an accuracy of 0.585, significantly lower than the best accuracy of 0.722. Another important note is the AUC-ROC scores of the final models. All models are upwards of 0.75, with the best being 0.799. These scores provide a good indication that the models are reliable and meaningfully confident at discerning between winning and losing outcomes.

Confusion Matrix

To get a better idea of the errors the model is making, we can investigate the confusion matrix for the model.

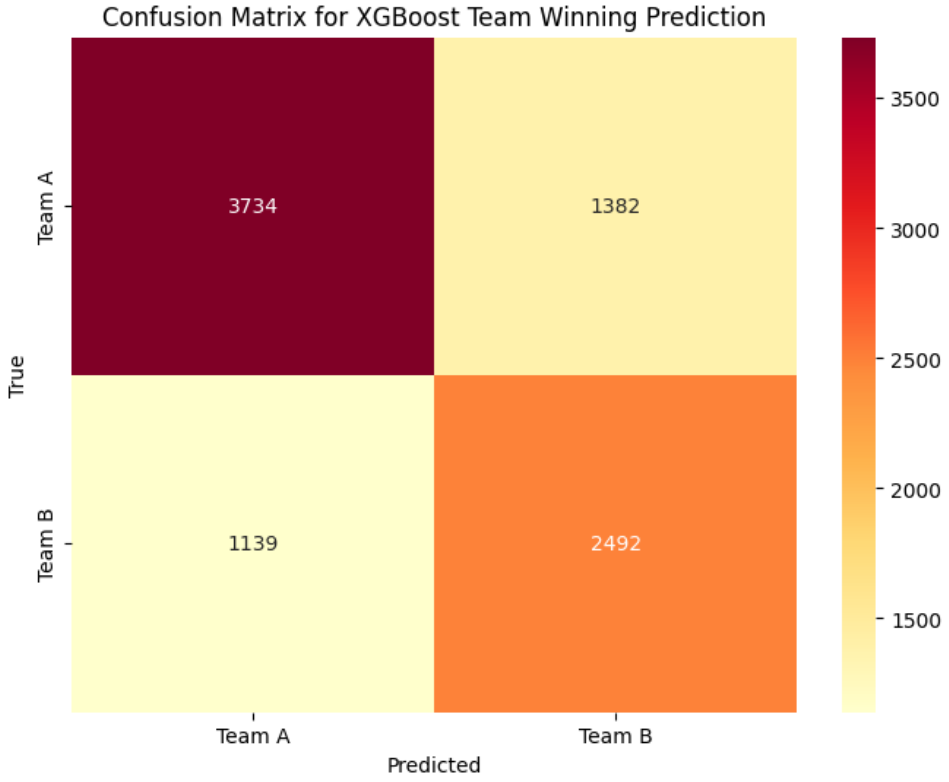
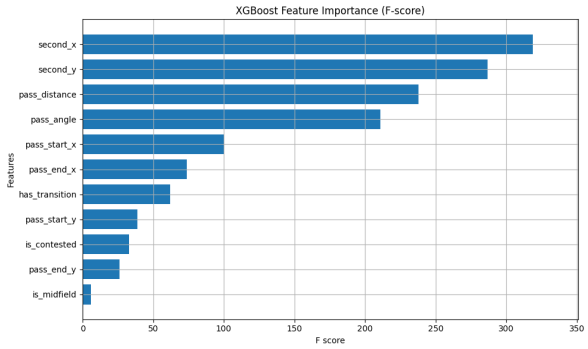


Figure 4.6: Visualization of the confusion matrix associated with the final XGBoost model for the team winning prediction.

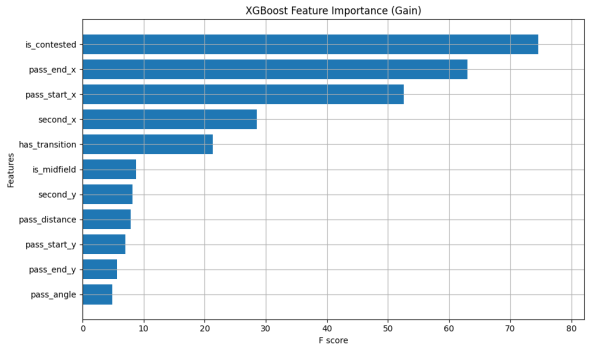
Figure 4.6 provides more insight into the types of errors that the model makes. Interestingly enough, the incorrect classifications are similar in numbers, with incorrectly predicting team B, a false-negative, being slightly larger. The difference may be due to the slight dataset imbalance, but it is not too large to raise significant concern. Instead, future work should focus on improving the accuracy of the model, or lowering both the false-positive and false-negative rather than solely worrying about one.

Ablation Testing

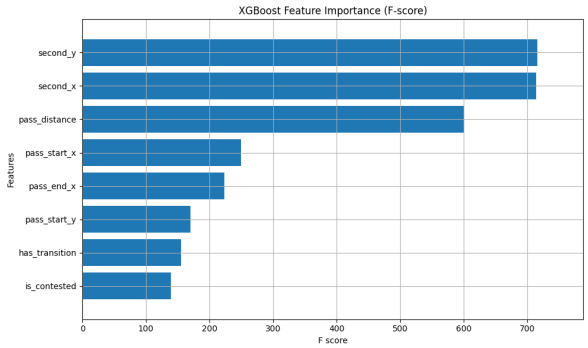
Once again, the ablation results are investigated.



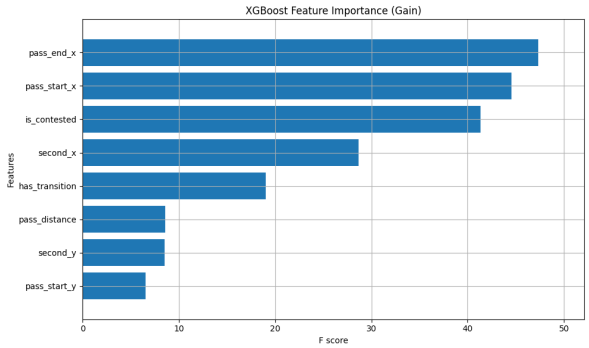
(a) F-score of XGBoost model for team winning prediction before ablation testing.



(b) Gain of XGBoost model for team winning prediction before ablation testing.



(c) F-score of XGBoost model for team winning prediction after ablation testing.



(d) Gain of XGBoost model for team winning prediction after ablation testing.

Figure 4.7: Feature importance using F-score and Gain for the XGBoost team winning prediction model before and after ablation testing.

In Figure 4.7, the feature importance for the model before and after ablation testing is observed. An interesting point is how is_contested is one of the lowest F-score features after ablation, but also one of the highest gain features. The change in score means that is_contested is not frequently being used to split trees, but when it is being used, it often has a significant impact on the decision of the model. This will be investigated more with SHAP analysis, but suggests that is_contested works in tandem with other features to reach a verdict on prediction.

It is also no surprise that the x coordinate of the second ball location is the most important predictor of which team wins the second ball. Anyone with soccer experience would say it makes sense because the closer you are to the opponent's goal, the more

likely the opponent is to win the second ball. This is because the defence has a larger need to win the second ball since they would be in imminent danger of being scored if it was lost. The x coordinate is just one example of how the feature importance agrees with practical soccer knowledge. However, to reach more informed conclusions, SHAP analysis must be performed to understand how each feature impacts the model.

SHAP Analysis

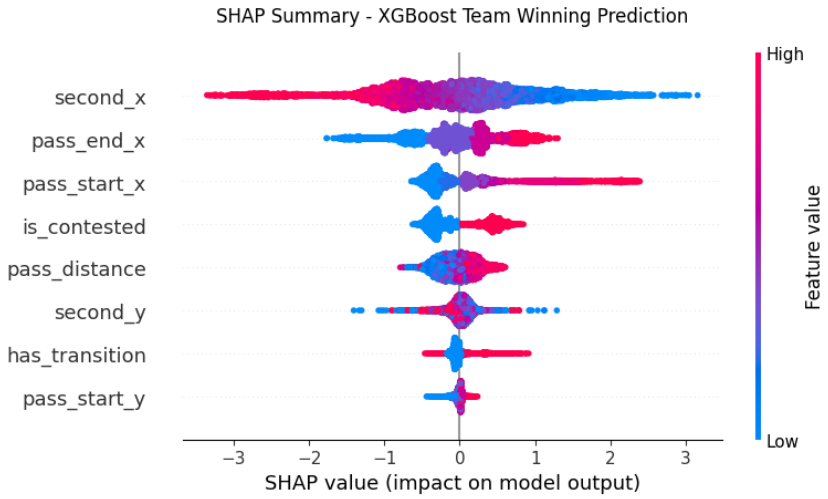


Figure 4.8: SHAP Analysis for the winning team prediction component after ablation testing.

After analyzing the SHAP plot, the previous suspicions from Section 4.1.2 are confirmed. The lower the x coordinate of the second ball location, the more likely team A is to win. Conversely, the higher the coordinate, the more likely team B is to win. In the middle of the field, the SHAP value is closer to 0, meaning it could go either way. Patterns emerge when analyzing the SHAP values for the winning team prediction model. One such pattern involves the feature second_ball_x. Here, lower x-values (i.e., the left side of the pitch) are associated with a positive SHAP impact for team A wins, while higher x-values tend to favour team B. Conversely, the feature pass_end_x, which captures the final horizontal destination of the original long ball or clearance, shows the opposite pattern: higher x-values (i.e., passes ending deeper into the attacking half) have

a stronger positive contribution toward predicting a team A win. This contrast might suggest that team A is more likely to win second balls when the initial pass ends further forward, possibly because the defending team will clear the ball further away from their goal, meaning back towards team A's end.

Another key to the previous point is the SHAP of is_contested. A high value, meaning the initial long ball is contested, influences the model towards a team A win. Combining the pass_end_x and is_contested means that passes further into the opponent's half only increase team A's chances of winning the second ball if they can contest the long ball. So, team A must proactively choose a target where they have a player to contest the long ball. An interesting thought would be to include player heights and/or aerial win rates in the features. Targeting a forward who can contest long balls against a lesser aerial threat defender may provide a higher probability of winning the second ball.

4.1.3 Gain

Visualizing Convergence

To get an intuition for how the long-term distribution of second ball possessions emerge, the heat maps of the absorbing states can be observed. Figure 4.9 shows the probability of the possession ending after 20 transitions. There is a clear pattern that the closer the team in possession gets to the opponent's goal, the more likely the possession will end. Again, this is trivial, but now we have evidence to back up the claim. Figure 4.9 also shows the probability of a goal after 20 transitions. Again, there is a clear pattern that the closer the team in possession gets to the opponent's goal, the more likely they are to score.

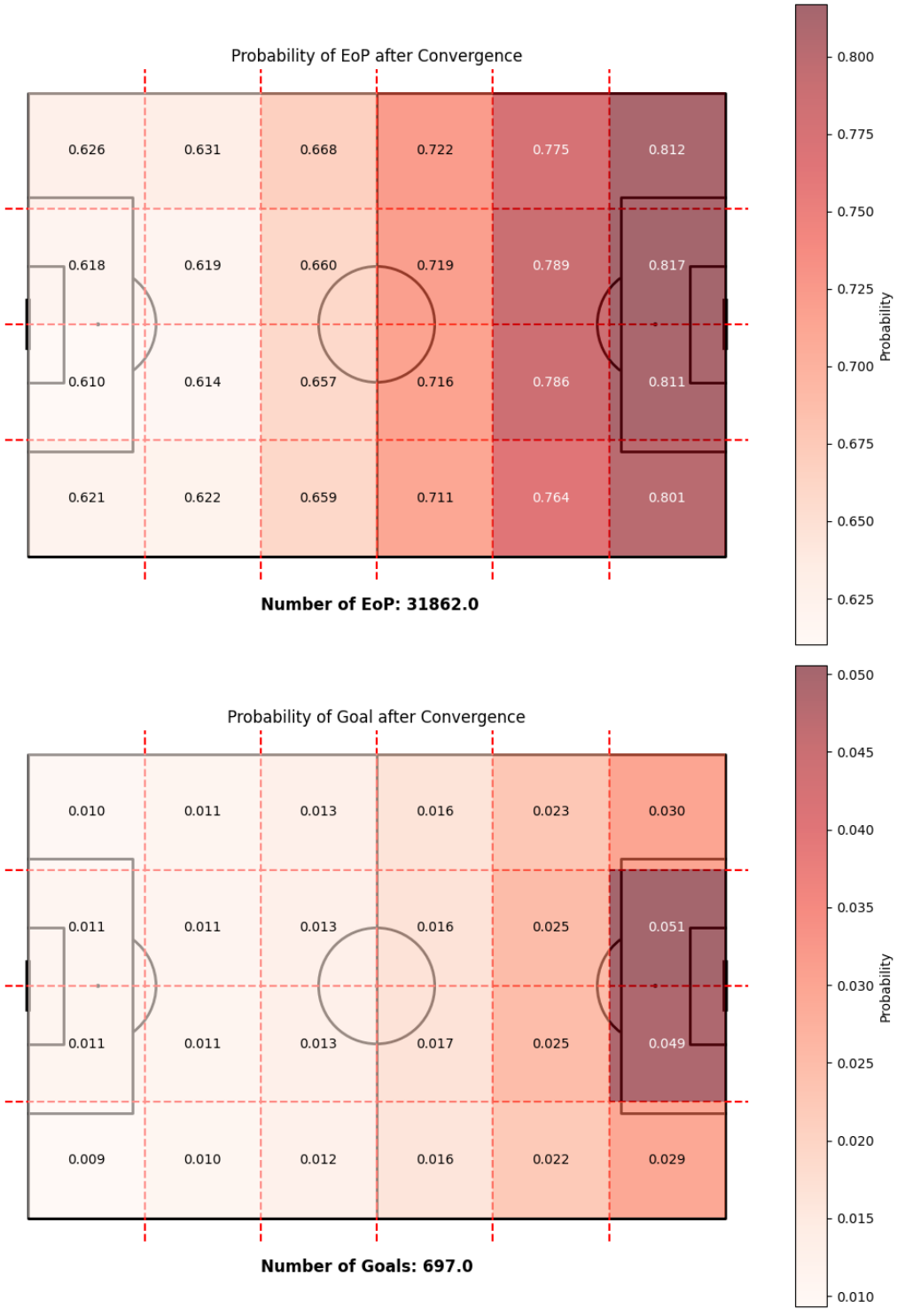


Figure 4.9: Heat map of the probability of absorption states based on zone after convergence of transition matrix. End-of-possession on top and goals on bottom.

Zone Valuation

Visualizing the gain per zone can provide insight into how areas of the pitch are valued. Unless otherwise stated, the analysis will be conducted using $H = 5$.

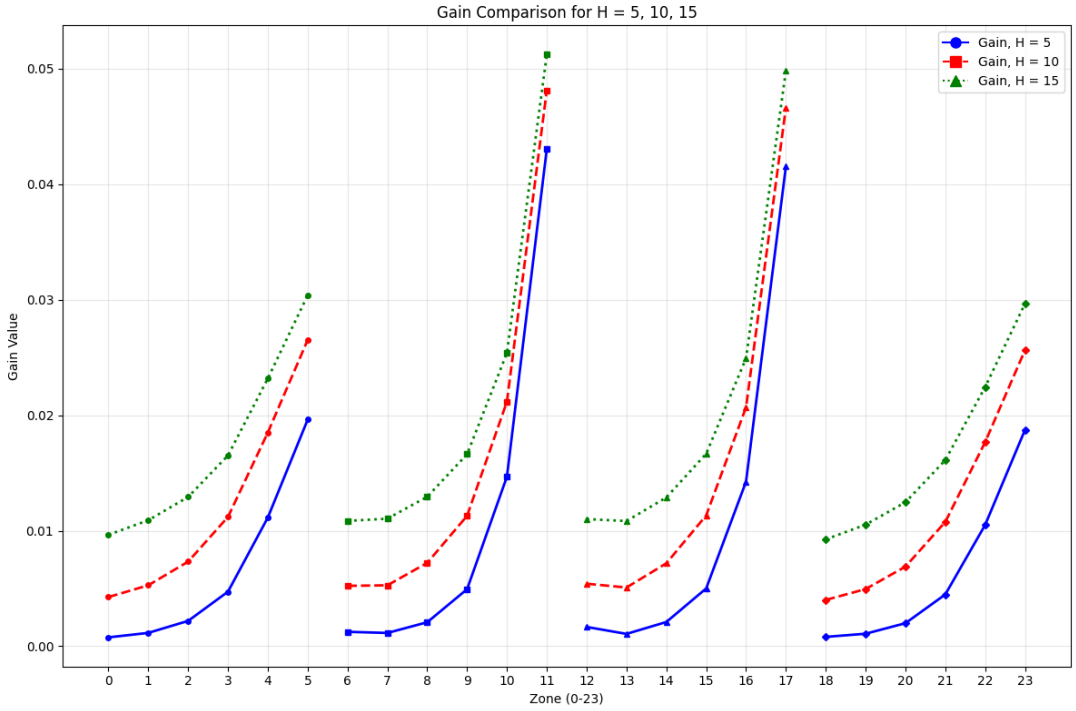


Figure 4.10: Scatter plot for the gain value per zone for different values of H.

As expected, the cumulative probability that a goal occurs in the next H actions increases as H increases. Similarly to Figure 4.9, we see that for $H = 5, 10, 15$, the closer and narrower a player is to the opponent's goal the more likely a goal will occur.

The Bootstrap

To quantify the uncertainty in the gain component, the bootstrap, a resampling technique, is used. Bootstrapping uses an observed sample to construct a statistic's sampling distribution [64]. The observed median from the original sample of N scores is calculated. An empirical sampling distribution is created by repeatedly drawing N random samples with replacement from the dataset of second ball possessions. Drawing N random samples is considered one bootstrap sample. This process is repeated $B = 1000$ times. The

bootstrap distribution is formed by calculating the median for each bootstrap sample.

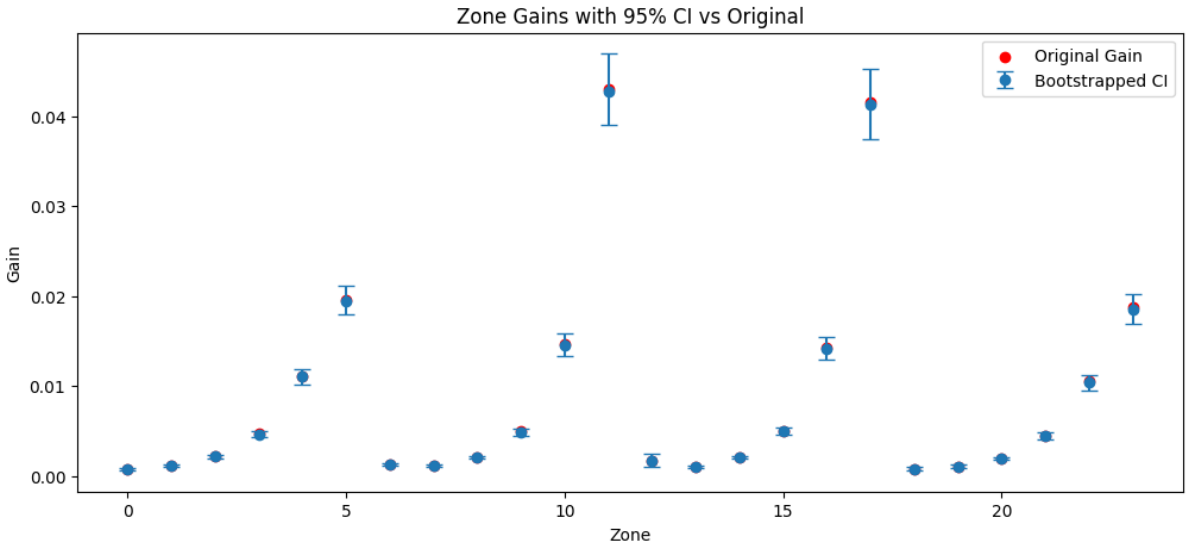
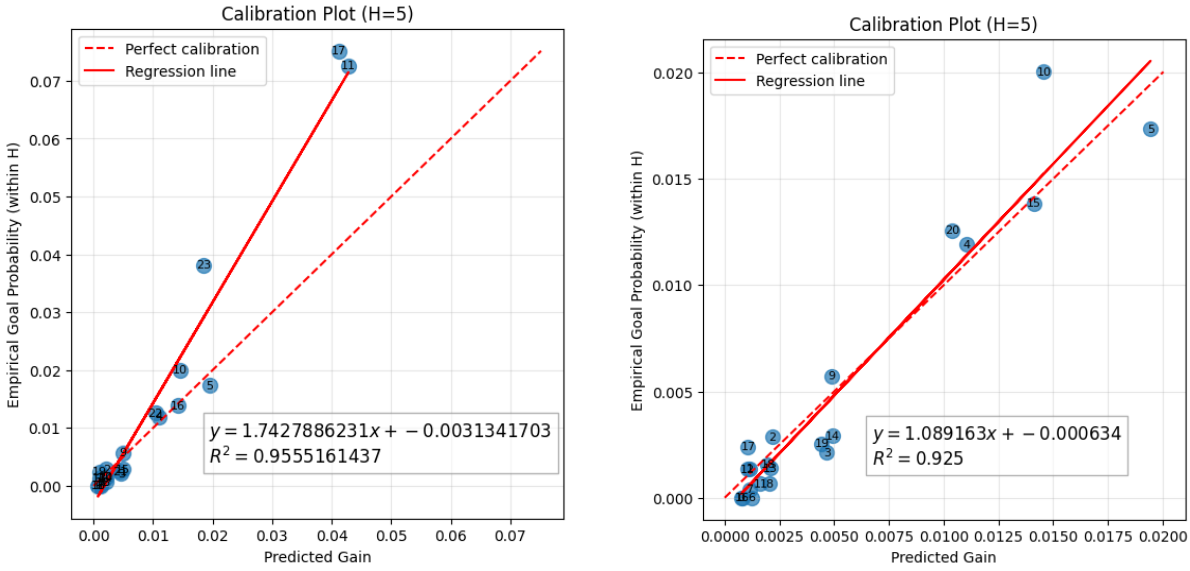


Figure 4.11: Scatter plot with error bars comparing the 95% confidence intervals for the gain of each zone for the bootstrap distribution to the original data.

From Figure 4.11, it is clear that the original gain scores are very similar to the bootstrapped medians. Confidence intervals are minimal from defensive zones and appear to grow towards attacking zones. Having more uncertainty in attacking zones is not surprising since there are fewer samples. However, the attacking zones' confidence intervals are not large, which gives evidence that the bootstrapped gain distribution is valid.

Calibration

The calibration of the gain was assessed by plotting empirical probabilities against predicted probabilities. The initial regression slope of approximately 1.7 indicates calibration issues compared to the perfect calibration slope of 1. Excluding zones 11, 17, and 23 decreases the slope to 1.089, which is very strong. The outliers have low empirical counts since they are attacking zones, suggesting the empirical probability does not reflect the underlying gain distribution. However, for the majority of zones, the Markov chain framework produces well-calibrated estimates.



(a) Calibration for $H = 5$. All zones included. (b) Calibration for $H = 5$ without outliers.

Figure 4.12: Calibration scatter plots for predicted gain vs empirical goal probability.

4.2 Player Analysis

In this section, the focus shifts from component-level evaluation to individual player contributions for second ball wins. Using the second ball framework and xSBV model, it becomes possible to quantify which players are most involved in second-ball wins and how impactful those wins are. Unless stated otherwise, the analysis uses the English Premier League data.

4.2.1 Second Ball Win Quantity

With a framework to quantify second ball wins, a list of players with the most number of second ball wins is created. Figure 4.3 shows the top 15 second ball winners with their minutes taken from [65]. It is clear to see that there is now an interpretable second ball win statistic. Players, coaches, and even fans can use this metric to analyze the number of second ball wins by player from past games.

Player	Team	Amount	p90
Danny Drinkwater	Leicester City	95	2.82
Andrew Surman	AFC Bournemouth	84	2.21
Glenn Whelan	Stoke City	77	2.19
Yann M'Vila	Sunderland	73	2.06
Idrissa Gana Gueye	Everton	71	2.08
Victor Wanyama	Southampton	71	2.55
Darren Fletcher	West Brom	70	1.87
Gareth Barry	Everton	68	2.16
Mark Noble	West Ham	68	1.92
Claudio Yacob	West Brom	67	2.15
N'Golo Kante	Leicester City	67	1.99
Eric Dier	Tottenham	66	1.82
Cesc Fabregas	Chelsea	65	2.02
Ashley Westwood	Aston Villa	65	2.15
Yohan Cabaye	Crystal Palace	62	2.07

Table 4.3: Top 15 second ball winners ranked by number of wins. Average number of wins per 90 minutes (p90) is also included.

4.2.2 Player Second Ball Win Visualizations

Next, to illustrate individual player contributions in second ball situations, an example of Declan Rice during the UEFA Euro 2024 tournament is presented. The visualization highlights all of Rice's second ball wins. In addition, second ball wins that led to a shot within the subsequent possession is marked. This type of spatial analysis provides deeper insight into not just how often a player wins second balls, but where on the field they are doing so. Figure 4.13 presents an interpretable second ball visualization designed to help coaches and players better understand second ball performance. These charts highlight individual player second ball tendencies, showing where second ball wins most likely occur. Such visual tools can assist coaches in identifying strengths and weaknesses in a player's second ball winning ability, guiding player development or scouting decisions.

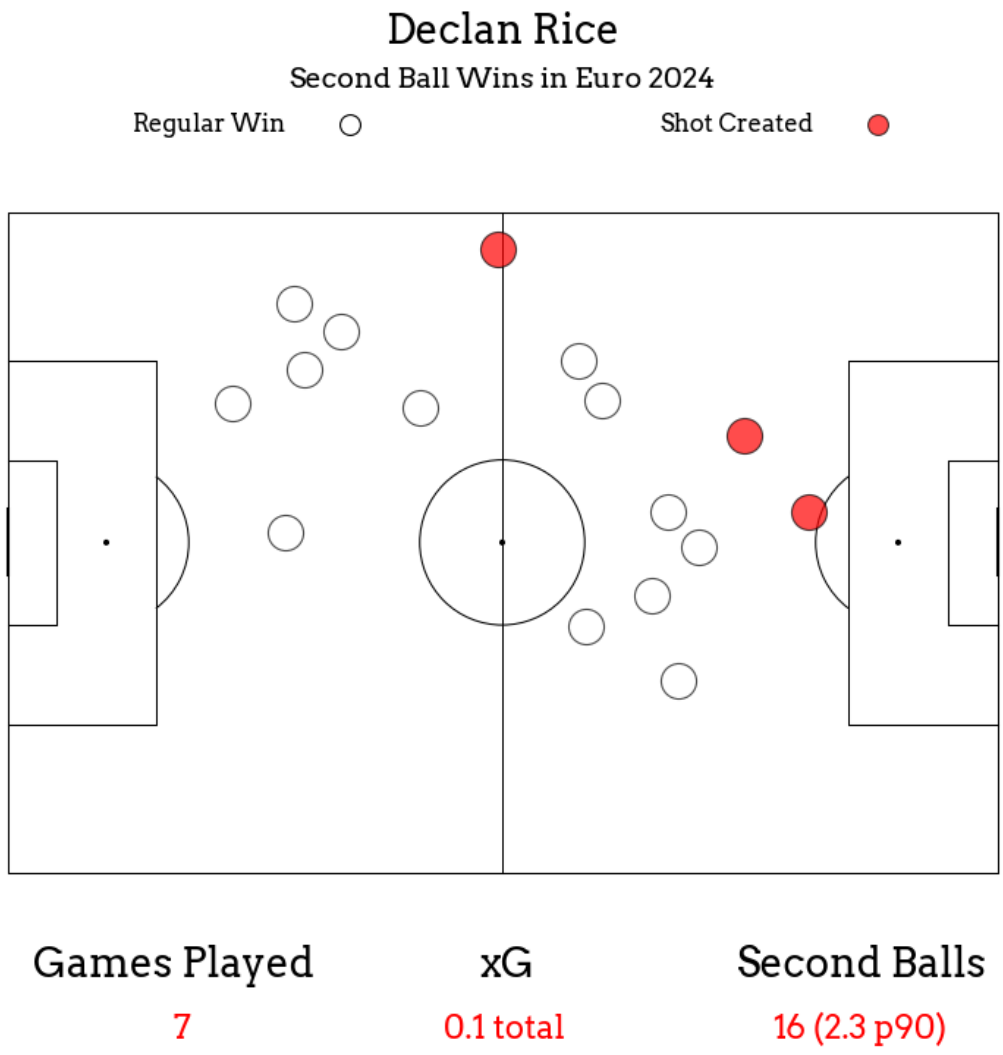


Figure 4.13: Visual of Declan Rice's second ball wins from the 2024 Euros.

4.2.3 Player xSBV

Now for the main result of this thesis, the xSBV for players. Table 4.4 shows the average xSBV for players with more than thirty-eight second ball wins. The cumulative xSBV is also shown along with the average and cumulative gain difference. The total number of second balls won from each player is also listed. It is noted that second ball wins that are won and transitioned into the same zone are not counted since the gain would equal zero. The zones that are not counted are a significant limitation of the xSBV model and are discussed in Section 5.2.

Player	**Avg xSBV**	Total xSBV	Avg Gain Diff	Total Gain Diff	Count
Yann Gérard M'Vila	0.000152	0.006094	0.003940	0.157604	40
James McArthur	0.000081	0.003334	0.001347	0.055214	41
Ben Watson	0.000071	0.002638	0.001220	0.045134	39
Ashley Westwood	0.000071	0.002967	0.002293	0.096327	42
Danny Drinkwater	0.000069	0.004769	0.001901	0.131142	70
Victor Wanyama	0.000010	0.000509	0.000492	0.024101	50
Glenn Whelan	0.000003	0.000128	0.000277	0.013553	49
N'Golo Kanté	-0.000006	-0.000249	0.000219	0.009856	45
Eric Dier	-0.000011	-0.000478	0.000384	0.016890	44
Mark Noble	-0.000046	-0.001893	-0.001255	-0.051451	43

Table 4.4: Top 5 and bottom 5 players ranked by average xSBV.

Yann Gérard M'Vila leads the list with the highest average xSBV (0.000152), supported by a relatively strong total gain differential (0.1576) across 40 events. The others in the top 5 have similar average xSBV scores, but considerably lower than M'vila. Therefore, in terms of creating value from second ball wins, it is concluded that Yann M'vila is the best. As a midfielder from Sunderland who would not widely be regarded as a "top" midfielder, it is exciting that the metric shows him as the best in the Premier League with respect to improving his team's chances of scoring from second ball wins. Looking back at the Premier League table (2.2), notice that Sunderland finished in 17th place, avoiding relegation by one spot.

At the opposite end, Victor Wanyama and Glenn Whelan post near-zero averages, and N'Golo Kanté, Eric Dier, and Mark Noble register negative values. However, it is crucial to note that these metrics only capture the probability of goals and do not reflect defensive value. Many of the lower-ranked players are defensive midfielders whose primary role is to disrupt opposition play and secure possession with safe passes. Without the value of the opponent winning the ball, we cannot fully capture the importance of these players. However, we can use average gain and cumulative gain to get a better idea of their importance.

Player	Avg Gain	**Total Gain**	Count
Danny Drinkwater	0.0190	1.7906	95
Yann Gérard M'Vila	0.0185	1.3508	73
Victor Wanyama	0.0183	1.2798	71
Ashley Westwood	0.0190	1.2352	65
Glenn Whelan	0.0152	1.1687	77
Mark Noble	0.0166	1.0963	68
N'Golo Kanté	0.0159	1.0667	67
Eric Dier	0.0156	1.0295	66
Ben Watson	0.0172	0.9780	59
James McArthur	0.0172	0.9277	54

Table 4.5: Comparison of selected players by average gain, total gain, and number of second ball wins, ordered by total gain.

Even without accounting for defensive contribution, Table 4.5 shows the importance of defensive midfielders. Victor Wanyama and Glenn Whelan both have entered the top 5 comparatively, while the rest have similar scores. So by using part of the full metric, the zonal value of player's second ball wins can be seen.

4.3 Team Analysis

4.3.1 Second Ball Win Quantity

Similarly to quantifying second ball wins for players, the second ball wins for teams can now also be quantified. Second ball losses are also calculated since for every second ball win, the opposing team loses. The binary nature of team second ball wins allows for easy calculation of the losses, which cannot be done as easily for players.

Team	**Wins**	Lost	Total	Win %
Manchester United	585	425	1010	57.92
Liverpool	542	438	980	55.31
Southampton	533	500	1033	51.60
Arsenal	531	363	894	59.40
Crystal Palace	522	514	1036	50.39
Leicester City	510	553	1063	47.98
Norwich City	508	593	1101	46.14
Swansea City	506	452	958	52.82
Tottenham Hotspur	502	419	921	54.51
Everton	480	463	943	50.90
West Ham United	473	495	968	48.86
Chelsea	467	396	863	54.11
West Bromwich Albion	460	597	1057	43.52
Stoke City	458	435	893	51.29
Sunderland	455	601	1056	43.09
Aston Villa	452	520	972	46.50
Watford	451	600	1051	42.91
AFC Bournemouth	451	453	904	49.89
Manchester City	432	413	845	51.12
Newcastle United	413	501	914	45.19

Table 4.6: Team second balls sorted by number of second ball wins (highest to lowest).

An interesting note is that eight of the top 10 teams had a second ball win rate of above 50%. Only three teams outside the top 10 had above a 50% win rate. Notably, Leicester City, the league champions, had a better win rate than only six teams, indicating that second ball win rate alone does not directly relate to success or failure.

4.3.2 Team Second Ball Win Visualization

Similarly to Figure 4.13, an interpretable visualization for team second ball situations is created. Figure 4.14 displays a second ball map for teams, visualizing the location of second ball wins and losses, along with shots created. The visualization provides an overview of a team's spatial second ball patterns, helping to identify strengths and weaknesses.

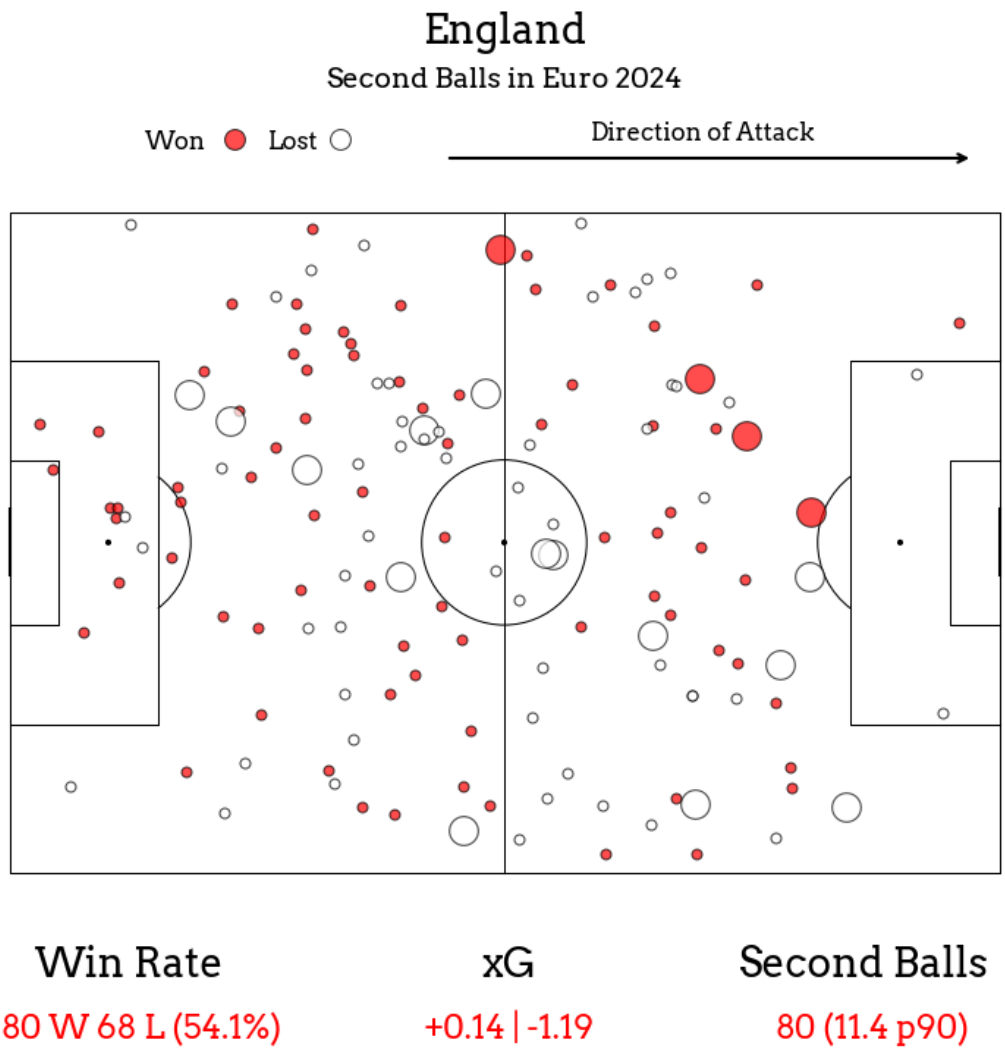


Figure 4.14: Visual of England's second ball wins and losses from the 2024 Euros. Large circles denote a shot created within the following possession.

One insight from Figure 4.14 is the difference in second ball wins and success rate for England. Although England won more second balls than they lost, there were more shots conceded than shots created. The most likely explanation is that since England often dominates possession with a high number of players advanced on the field, they are more prone to counter-attacks. Also, since many weaker teams will defend with many players behind the ball, shots created from second balls may be less likely. Paired with the fact that England were only expected to score 0.14 goals from their second ball wins, whereas they were expected to concede 1.19, it highlights an area England could

improve on. England could benefit from strategies to prevent attacks after losing second balls, while weaker teams could benefit from strategies to create more threatening scoring opportunities from winning second balls.

4.3.3 Team xSBV

Now, the xSBV scores per team can be analyzed. Again, the teams are ranked in order of highest average xSBV to lowest.

Team	**Avg xSBV**	Total xSBV	Avg Gain Diff	Total Gain Diff	Count
West Brom	0.000068	0.020476	0.002118	0.635521	308
Crystal Palace	0.000059	0.020081	0.001532	0.523900	349
Aston Villa	0.000056	0.015796	0.001369	0.383417	286
Sunderland	0.000054	0.015751	0.001949	0.569019	295
Leicester City	0.000054	0.017483	0.002246	0.732334	329
Southampton	0.000046	0.016098	0.001632	0.575958	362
Swansea City	0.000044	0.014084	0.000984	0.316853	326
Norwich City	0.000039	0.012701	0.001358	0.437419	328
Manchester City	0.000039	0.010909	0.001070	0.298503	284
Arsenal	0.000039	0.012995	0.000988	0.332873	345
Liverpool	0.000036	0.012995	0.001347	0.486097	369
Watford	0.000035	0.010062	0.000830	0.236572	292
Manchester United	0.000033	0.012545	0.000766	0.287362	387
Chelsea	0.000031	0.009680	0.000691	0.212803	315
Tottenham Hotspur	0.000028	0.009269	0.001027	0.334641	333
Everton	0.000028	0.008768	0.001060	0.336051	324
AFC Bournemouth	0.000027	0.007748	0.001099	0.318634	298
Newcastle United	0.000026	0.007366	0.000695	0.195879	283
West Ham United	0.000026	0.007690	0.001085	0.322182	307
Stoke City	0.000021	0.005716	0.000675	0.184246	280

Table 4.7: Team-level average and total xSBV, average and total difference gain, and number of second ball wins. Ranked in order of average xSBV.

Figure 4.7 reveals some interesting results. Originally, it was thought that lower table teams would have worse xSBV, but that is not necessarily the case. A more reasonable explanation would be that high xSBV would be associated with teams with a direct, long-ball play style. So it makes sense that worse teams already focus on quick attacks following second ball wins. However, not all lower table teams have high xSBV. For example, Newcastle, the team that was closest to avoiding relegation, had the third-worst xSBV score. The lower score indicates that Newcastle could have benefited from improvements in attacking play following second ball wins. Cases like Newcastle are exactly who this metric is geared toward. There is a possibility that if Newcastle had worked to improve their xSBV, they could have avoided relegation.

4.3.4 Tactics

To provide coaches and players with tactical guidance for winning second balls, heat maps are presented, which show trends of where second balls are likely to land. Although the location prediction component does not predict zones with high accuracy, it does seem to understand the spatial relationships of second ball win locations. As component performance is improved, a similar location distribution is expected, but with higher prediction accuracy. However, from Figure 4.15, it can be shown that a long ball from zone 12 to 16 is most likely to fall into zones 13 and 14. It is also noted that zone 17 has the next highest probability. Therefore, a tactic could be to target zone 16 from goal kicks with many players in zones 13 and 14, with a few players running in behind to zone 17 for flick-on headers. This is one example, but by looking at team-specific visuals of where past success occurred, coaches can use the ideas from the location prediction component to create new tactics for their team.

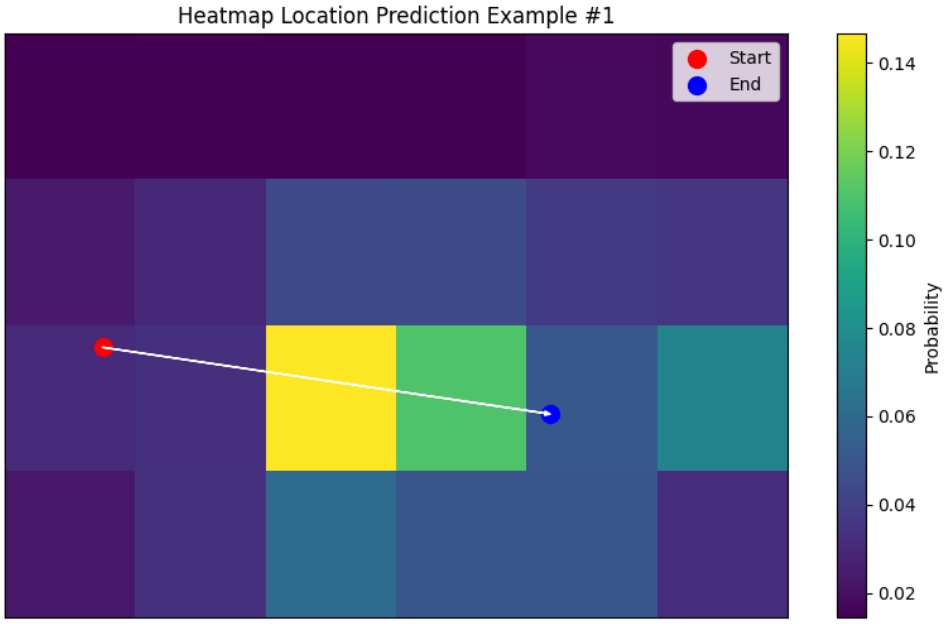


Figure 4.15: Location prediction probability heat map example of a second ball win.

4.4 Summary

This chapter presented the results of the xSBV framework, validating each model component and demonstrating its use in analyzing second ball wins. The location prediction component struggled with top-1 accuracy, still outperforming the naive baseline, but achieved a solid top-3 accuracy given the novel nature of the work. The winning team classifier produced strong AUC-ROC scores, showing it can reliably predict second ball winners. Feature ablation testing and SHAP analysis showed the relative importance of features and offered insight into how the model makes decisions. The gain was analyzed using the bootstrap method to check uncertainty in the distribution of second ball possessions. Beyond raw model performance, visualizations for quantifying second ball wins on a player and team level were created. These outputs allow for easily interpretable results for users to show where teams and players excel/struggle with winning second balls. Finally, xSBV was used to analyze player and team data, demonstrating the model's ability to support second ball performance evaluation. The results highlight xSBV as both a predictive and explanatory tool for second ball wins in soccer.

Chapter 5

Conclusions

5.1 Summary of Findings

To summarize the findings of this thesis, let us take a look back at the original research questions.

Research Questions

1. Can second balls be quantified using available soccer data?
2. Can second ball wins be accurately predicted?
3. What are the key factors influencing second ball wins?
4. Can the potential risk of second balls be predicted?

We have shown (1) through a mathematical framework and definition of second balls, coupled with the code found in Appendix A, second balls were successfully quantified and extracted from the available StatsBomb data. Machine learning models were used to predict second ball wins (2). Location prediction achieved a top-1 and top-3 accuracy of 0.24 and 0.60 respectively. Team winning prediction models achieved an accuracy of 0.72 and AUC-ROC score of 0.799. Although these scores are moderate, there is sufficient

evidence that these scores can be improved and second ball wins can be accurately predicted. Key factors influencing second ball wins (3) were investigated using feature importance metrics and SHAP analysis. To this end, key insights such as contested long balls further into an opponent's half result in a higher chance for team A to win a second ball. Finally, the potential risk of second balls was analyzed using the novel xSBV metric, ranking players performance at winning second balls to create dangerous attacks. Ratings show a quantifiable way to rank the probability of goals occurring from possessions associated with second balls throughout a game. It is stressed that the results simply lay the groundwork for future research on the topic, as they are very early findings and need improvement to be effective at the highest level.

5.2 Limitations

While this thesis offers novel contributions to the quantification and analysis of second balls in soccer, there are limitations to the approach taken, which are discussed below.

5.2.1 Data

One of the primary limitations of this thesis lies in the nature of the available data. Complete player tracking data was not accessible, which restricted the modelling of off-ball player locations. Instead, this thesis relied on publicly available event data from StatsBomb. To incorporate additional spatial context, StatsBomb 360 data was used; however, 360 data is currently limited to select competitions, such as the 2024 UEFA European Championship. As a result, the 360 data was used to supplement small-scale model testing and analysis, but it was not sufficient to independently train robust machine learning models.

Event data is inherently limited because it is most commonly collected by human operators. Liu et al. [66] argues that event data is reliable when collected by well trained

operators. However, the validity across providers is questioned due to differences in definitions for certain events. Biermann et al. [67] discuss how the highly subjective definitions of events can change across data sets leading to results that are not generalizable. Anzer and Bauer [21] find significant differences in the synchronization of time stamps between event and tracking data. Despite these limitations, event data remains the most widely used and common data source for performance analysis, making it a foundation for the research in this thesis.

It should also be mentioned that the main datasets that were used in this thesis were from the 2015/2016 season, which is quite old for soccer-specific data. Physical characteristics of players and team tactics constantly change, and to generalize the results to the present, there could be potential shortcomings since the model was trained on old data. Due to data availability, the use of the 2015/2016 data was a necessary choice; however, whether the age of data affects the results for second balls is unexplored.

5.2.2 Methodology

A key component of this thesis is the extraction and classification of second ball sequences, which required the development of a custom detection framework. The design of these second ball chains was informed by manual annotation of game footage, where second ball events were identified and labelled based on patterns in player behaviour and ball movement. While this approach allowed for domain-informed modelling and generated a reasonable average of approximately 25 second ball events per game, it may undercount certain edge cases or ambiguous situations that fall outside the defined patterns. These potential omissions could introduce bias into the dataset or reduce model robustness in real-world applications.

Another limitation of this thesis is that only second ball *wins* are investigated. The code provided in Appendix A extracts second ball wins from StatsBomb data, not second balls. Analyzing only second ball wins means that we ignored scenarios where no team

establishes possession, i.e, the ball goes out of play from the second ball. In reality, ignoring these scenarios contributes to a lower number of second balls per game, and potential game dynamics that are left unexplored. However, it still can be argued that for the purpose of this thesis that focusing solely on second ball wins helps to narrow the scope and keep the results concise.

Set pieces - free kicks, corners, and throw-ins - were not included in this analysis. Given the difference in nature of open play and set pieces, the different scenarios most likely would have different results, so it is argued they should be modelled and analyzed differently. Because of this, the results are unaffected by set pieces, but also the model and results do and should not be applied to them.

Additionally, the Markov chain model used to estimate the gain of second ball possessions is inherently limited by the Markov property. While simplifying makes the modelling straightforward, it may leave out valuable insight for the xSBV.

5.2.3 Results

The model assigns scores by using the difference in probability of scoring a goal within the following H actions from two different zones. The biggest problem is that if the second ball winner gets the ball and transitions it to the same zone, the resulting difference in gain would equal zero. In reality, the difference should not be zero since winning the second ball itself provides value for your team, even if you transition it into the same zone. Differentiating the gain between loose-ball and secure-ball states would be a natural extension to help improve the model. To address the problem in this work,the single zone gain of where the second ball win is transitioned to is used, which gives an idea of the value of zones that players and teams occupy.

The values from the xSBV model are also hard to interpret. Since the values from the difference in gain and the location prediction component are very small, the xSBV becomes very small too. In retrospect, either the criteria for gain (probability of a goal

within H steps) should have been changed to something more likely (probability of a shot within H steps), or another value needs to be added to the component (value of the other team's attack stopped by the second ball win). If the criteria were changed, then the scores could be larger and easier to understand. However, the model is still valuable since you can compare scores between players and teams. Even if it is hard to understand the value through a small number, knowing that player A is 3x better than player B at providing goal-scoring opportunities from second ball wins is important and interpretable.

Another limitation of the results is the nature of the rankings. Using xSBV, ranking players over the course of a season, distinguishing between the best and worst second ball winners is possible. However, in the nature of sport lies unpredictability, and there is no way of knowing whether the players identified will continue to perform similarly. Many factors can contribute to a decline in player performance, such as mental and physical health issues. So, it is suggested that users of this model never solely pick and choose players using only this metric, but rather utilize the model to help supplement performance evaluation and scouting.

The communication gap between practical use and sports analytics is another limitation. Although coaches are increasingly supportive of sport scientists, most prefer new ideas that come from other coaches or clinics rather than from written reports or academic publications [68]. Therefore, an important role of sports scientists is to transfer data-driven research and ideas into interpretable and actionable performance solutions for coaches, athletes, scouts, and medical staff [69]. The scope of this thesis is only concerned with data-driven analysis, so it is limited in regards to knowledge transfer between analysts and coaches.

5.3 Future Work

While this thesis presents a comprehensive framework for quantifying and evaluating second ball events in soccer, there are several promising directions for future work. These span both technical improvements and potential applications in other sports and domains.

Building on the limitations, all second balls, not just wins, could be included. Given the state of the code, a simple and straightforward extension would be to modify the existing code to include all second balls and analyze how the new ones affect the results. New areas and patterns could be discovered to help teams capitalize on second balls through quick throw-in or corner winning strategies. Also, second balls from set pieces can be included to help gain strategy for a unique but equally important aspect of the game.

At a narrow level, the accuracy of the current models could be significantly improved with access to full tracking data. Although event data provided valuable insight, it is inherently limited in spatial coverage. The availability of continuous tracking data would allow for a more precise understanding of player movement, spacing, and intent — all of which are likely to enhance predictions for both second ball location and winning team outcomes. Furthermore, more advanced modelling techniques such as deep learning architectures (e.g., convolutional or graph neural networks) may offer performance gains over the current tree-based methods.

Adding pitch control to enhance the predictability of second ball wins would be valuable. Using StatsBomb 360 data, the problem is more complex since it does not provide continuous data or player info regarding the players around the ball. However, a simple extension is to use Voronoi diagrams (See Appendix B.7) instead to see if there is evidence that pitch control would even help.

The expected gain should also be incorporated. Recall that it represents the weighted average of the gain from each zone, weighted by the likelihood that the ball is transitioned there. A natural extension is to include the value in the xSBV model. Doing so would

solve the issue of assigning a value of zero when the ball remains in the same zone, while also tracking whether players consistently outperform or underperform the average gain value.

Since the models are trained on men's data alone, it is necessary to extend the research to women-specific models. Davis and Bransen split the groups of training data into just women, just men, and both to find the differences in the accuracy of xG models based on the data used to train them [70]. The models value features between genders differently. For example, men are more likely to score from curling shots towards the far post, whereas women are more likely to score tap-ins from a cross. With women's soccer data becoming more available, it will be important to extend the research to see how results generalize across genders and whether different conclusions can be drawn.

To ensure that ideas from this thesis can be applied in practice, future work should use established theories or practices of knowledge transfer to examine implementation results. Research has shown that providing insights through conversation or discussion with coaches can be particularly effective [68][69]. Building on this, different frameworks such as the Knowledge to Action [71] or SECI Model of Knowledge Creation [72] can be applied to bridge the gap between analyst and coach.

On a broader scale, the conceptual framework developed in this thesis could be adapted to other sports where moments of contested possession or transitional states play a crucial role. In ice hockey, for instance, dump-and-chase situations resemble long balls in soccer, and modelling the recovery of the puck after a dump-in could benefit from a similar chain-based approach. Likewise, in Basketball, rebounds share properties with second ball events — they involve contested space, loose possession, and immediate value transitions. Even outside of sport, this type of modelling could be helpful in domains like robotics [73] (e.g., control state handovers) or finance [74] (e.g., stock price changes and buy times), where expected outcomes are meaningful to estimate.

Ultimately, second ball research has many directions for exploration as well as areas

for improvement. The broader principle of modelling value transitions in contested, dynamic environments — especially where state transitions are observable but outcomes are uncertain — is fertile ground for new research.

5.4 Final Remarks

This thesis has presented a framework for identifying, modelling, and evaluating second balls in soccer. By combining mathematical definitions, machine learning classifiers, and Markov chains, the proposed xSBV (expected second ball value) model captures both the likelihood of winning a second ball and the probability of scoring goals within the following possession. While the approach is rooted in the specific context of second ball analysis, the underlying principles and modelling techniques contribute more broadly to the field of data-driven soccer analytics.

Ultimately, this work serves as a tool for performance evaluation and the groundwork for future extensions. As access to richer datasets such as full tracking data improves and as the soccer analytics community continues to push boundaries, there is significant potential to build on the methods developed here. Whether applied to tactical analysis, player recruitment, or in-game decision making, the ideas introduced in this thesis demonstrate the value of quantitatively understanding transitional moments in the game — and the second ball remains one of the most critical.

Bibliography

[1] Christian Kotitschke. *Soccer analytics data: Beginners guide*. https://www.linkedin.com/pulse/soccer-analytics-data-beginners-guide-christian-kotitschke. June 2020. (Visited on 08/26/2024).

[2] Daniel Link. "Data analytics in professional soccer". In: *Springer Vieweg, Wiesbaden* (2018).

[3] Chun Hang. *Quantifying Second-Ball Wins*. https://lchunhang.medium.com/quantifying-second-ball-wins-d626ac56f108. 2023. (Visited on 08/26/2024).

[4] Paul Wilson. *Guardiola and Koeman recognise difficulties of tackling the full English*. https://www.theguardian.com/football/blog/2016/dec/14/pep-guardiola-ronald-koeman-manchester-city-everton-full-english. Dec. 2016. (Visited on 08/26/2024).

[5] Ivan Sunjic et al. "Win the second balls! The impact of strategic ball recovery on match performance in elite soccer". In: *International Journal of Performance Analysis in Sport* (Feb. 2025).

[6] Jan Van Haaren et al. *Analysing Performance and Playing Style using Ball Event Data*. Nov. 2019.

[7] Mikhail Borodastov. *Tracking data — the most detailed and accurate information about players actions on the pitch*. https://footsci.medium.com/. Mar. 2024. (Visited on 08/26/2024).

[8] Andy Chambers. *StatsBomb Open Data*. https://github.com/statsbomb/open-data. [Data set]. 2023.

[9] FBref.com. *2015-2016 Big 5 European Leagues Stats*. https://fbref.com/en/comps/Big5/2015-2016/2015-2016-Big-5-European-Leagues-Stats.

[10] Jamin D. Speer. "The consequences of promotion and relegation in European soccer leagues: A regression discontinuity approach". In: *Sports Economics Review* 1 (Mar. 2023). ISSN: 2773-1618. DOI: 10.1016/j.serev.2022.100003.

[11] Sam Green. *Assessing The Performance of Premier League Goalscorers*. https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/. Apr. 2012. (Visited on 08/27/2024).

[12] Matthew Mcmullen. *What are Features in Machine Learning and Why it is Important?* https://cogitotech.medium.com/what-are-features-in-machine-learning-and-why-it-is-important-e72f9905b54d. Dec. 2021.

[13] Michael Bertin. *The Third-to-Last Thing I'll Ever Write About Expected Goals*. http://michaelbertin.com/2015/08/28/the-third-to-last-thing-ill-ever-write-about-expected-goals/. Aug. 2015.

[14] Alex Rathke. "An examination of expected goals and shot efficiency in soccer". In: *Journal of Human Sport and Exercise* 12 (2017). DOI: 10.14198/jhse.2017.12.Proc2.05. URL: http://hdl.handle.net/10045/68771.

[15] James Mead, Anthony O'Hare, and Paul McMenemy. "Expected Goals in Football: Improving Model Performance and Demonstrating Value". In: *PloS One* (2023). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0282295.

[16] Pieter Robberechts and Jesse Davis. "How Data Availability Affects the Ability to Learn Good xG Models". In: *Machine Learning and Data Mining for Sports Analytics*. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-64912-8.

[17] Rachel Draelos. *Measuring Performance: AUROC*. https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/. Feb. 2019.

[18] Stephanie Glen. *Brier Score: Definition, Examples*. https://www.statisticshowto.com/brier-score/. (Visited on 11/18/2023).

[19] Michael Caley. *Let's Talk about Expected Goals*. https://cartilagefreecaptain.sbnation.com/2015/4/10/8381071/football-statistics-expected-goals-michael-caley-deadspin. Apr. 2015. (Visited on 11/18/2023).

[20] Patrick Lucey et al. ""Quality vs Quantity": Improved Shot Prediction in Soccer Using Strategic Features from Spatiotemporal Data". In: *MIT SLOAN Sports Analytics Conference*. Boston Convention and Exhibition Center, Feb. 2015. (Visited on 11/18/2023).

[21] Gabriel Anzer and Pascal Bauer. "A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer)". In: *Frontiers in Sports and Active Living* 3 (Mar. 2021). ISSN: 2624-9367. DOI: 10.3389/fspor.2021.624475. (Visited on 11/17/2023).

[22] Danny Pugsley. *A Deeper Look At Shots on Target*. https://bitterandblue.sbnation.com/2013/1/17/3880454/a-look-at-shots-on-target-epl. Bitter and Blue, Jan. 17, 2013. (Visited on 11/18/2023).

[23] FootballCritic. *Premier League Average Passes Per Game Overview for Teams*. https://www.footballcritic.com/premier-league/season-2019-2020/passes-per-game/2/21558. FootballCritic. (Visited on 11/18/2023).

[24] Karun Singh. *Introducing Expected Threat (xT)*. https://karun.in/blog/expected-threat.html. Dec. 2018. (Visited on 08/29/2024).

[25] Sarah Rudd. *A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains*. https://nessis.org/nessis11/rudd.pdf. Harvard University, 2011.

[26] A.J. Peters et al. "Expected Pass Turnovers (xPT) - a model to analyse turnovers from passing events in football". In: *Journal of Sports Sciences* 42 (2024). DOI: 10.1080/02640414.2024.2379697.

[27] Pieter Robberechts, Maaike Van Roy, and Jesse Davis. "Un-xPass: Measuring Soccer Player's Creativity". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Long Beach CA USA: ACM, Aug. 2023. ISBN: 979-8-4007-0103-0. DOI: 10.1145/3580305.3599924. (Visited on 10/19/2023).

[28] Pegah Rahimian et al. "Pass Receiver and Outcome Prediction in Soccer Using Temporal Graph Networks". In: *Machine Learning and Data Mining for Sports Analytics*. Sept. 2023. ISBN: 978-3-031-53832-2. DOI: 10.1007/978-3-031-53833-9_5.

[29] Daniel Link, Steffen Lang, and Philipp Seidenschwarz. "Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data". In: *PLOS ONE* 11 (Dec. 2016). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0168768. (Visited on 07/22/2024).

[30] Tom Decroos et al. "Actions Speak Louder Than Goals: Valuing Player Actions in Soccer". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. July 2019. DOI: 10.1145/3292500.3330758. arXiv: 1802.07127 [stat]. (Visited on 11/17/2023).

[31] Maaike Van Roy et al. "Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP". In: *Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports*. 2020. (Visited on 12/09/2023).

[32] Javier Fernandez, Luke Bornn, and Daniel Cervone. *Decomposing the Immeasurable Sport: A Deep Learning Expected Possession Value Framework for Soccer*. https://www.sloansportsconference.com/research-papers/decomposing-

the-immeasurable-sport-a-deep-learning-expected-possession-value-framework-for-soccer. 2019. (Visited on 07/22/2024).

[33] Javier Fernandez, Luke Bornn, and Daniel Cervone. "A Framework for the Fine-Grained Evaluation of the Instantaneous Expected Value of Soccer Possessions". In: *Machine Learning* (June 2021). ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-021-05989-6. (Visited on 02/22/2024).

[34] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Jan. 2017. DOI: 10.48550/arXiv.1412.6980. (Visited on 09/04/2024).

[35] Borja Burriel and Javier M Buldú. "The Quest for the Right Pass: Quantifying Player's Decision Making". In: *StatsBomb Conference*. 2021.

[36] William Spearman. "Beyond Expected Goals". In: *MIT Sloan Sports Analytics Conference*. Mar. 2018.

[37] William Spearman. "Quantifying Pitch Control". In: *OptaPro Analytics Forum*. Feb. 2016. DOI: 10.13140/RG.2.2.22551.93603.

[38] Laurie Shaw. *Advanced Football: Building and Applying a Pitch Control Model in Python*. Apr. 2020. (Visited on 12/28/2024).

[39] Chaoyi Gu et al. *Player Pressure Map – A Novel Representation of Pressure in Soccer for Evaluating Player Performance in Different Game Contexts*. Jan. 2024. DOI: 10.48550/arXiv.2401.16235. (Visited on 03/06/2024).

[40] Javier Fernández and Luke Bornn. "Wide Open Spaces: A Statistical Technique for Measuring Space Creation in Professional Soccer". In: *MIT Sloan Sports Analytics Conference*. Mar. 2018.

[41] Pieter Robberechts. "Valuing the Art of Pressing". In: *StatsBomb Innovation In Football Conference*. London, UK: StatsBomb, Oct. 2019.

[42] Simon Merckx et al. "Measuring the Effectiveness of Pressing in Soccer". In: *Workshop on Machine Learning and Data Mining for Sports Analytics*. 2021.

[43] Henrik Biermann et al. "Towards Expected Counter - Using Comprehensible Features to Predict Counterattacks". In: *Machine Learning and Data Mining for Sports Analytics*. Cham: Springer Nature Switzerland, 2023. ISBN: 978-3-031-27526-5 978-3-031-27527-2. DOI: 10.1007/978-3-031-27527-2_1. (Visited on 10/19/2023).

[44] Joaquín Gonzalez-Rodenas et al. "Association between Playing Tactics and Creating Scoring Opportunities in Counterattacks from United States Major League Soccer Games". In: *International Journal of Performance Analysis in Sport* 16 (Aug. 1, 2016). ISSN: 2474-8668. DOI: 10.1080/24748668.2016.11868920. URL: https://doi.org/10.1080/24748668.2016.11868920 (visited on 12/28/2024).

[45] Leon Forcher et al. "Is Ball-Possession Style More Physically Demanding than Counter-Attacking? The Influence of Playing Style on Match Performance in Professional Soccer". In: *Frontiers in Psychology* 14 (2023). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2023.1197039.

[46] Mark Carey and Ahmed Walid. "English football is besotted with second balls – but how important are they?" In: *The New York Times* (Mar. 2025). ISSN: 0362-4331. URL: https://www.nytimes.com/athletic/6193329/2025/03/13/measuring-second-balls-premier-league-analysis/.

[47] Stephen Robertson. "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF". In: *ResearchGate* (2004). DOI: 10.1108/00220410410560582. URL: https://www.researchgate.net/publication/238123710_Understanding_Inverse_Document_Frequency_On_Theoretical_Arguments_for_IDF.

[48] Anshuman Singh. *Multiclass Classification in Machine Learning*. Dec. 2024. URL: https://www.appliedaicourse.com/blog/multiclass-classification-in-machine-learning/.

[49] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. URL: https://dl.acm.org/doi/10.1145/2939672.2939785.

[50] Soumyadeep Ghosh et al. "On learning discriminative embeddings for optimized top-k matching". In: *Pattern Recognition* (June 2025). ISSN: 0031-3203. DOI: 10.1016/j.patcog.2025.111341.

[51] Albert Boateng et al. "Modular Analysis of Dataset Balancing Techniques For Binary Classification". In: *2023 Computer Applications & Technological Solutions (CATS)*. Oct. 2023, pp. 1–6. DOI: 10.1109/CATS58046.2023.10424219. URL: https://ieeexplore.ieee.org/abstract/document/10424219.

[52] Fatih Karabiber. *Binary Classification*. https://www.learndatasci.com/glossary/binary-classification/.

[53] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *TEST* 25 (June 2016). ISSN: 1863-8260. DOI: 10.1007/s11749-016-0481-7.

[54] GeeksforGeeks. *Logistic Regression in Machine Learning*. https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/. July 2025.

[55] Rupinder Sekhon and Roberta Bloom. *10.1: Introduction to Markov Chains*. Mar. 2020. URL: https://math.libretexts.org/Bookshelves/Applied_Mathematics/Applied_Finite_Mathematics_(Sekhon_and_Bloom)/10%3A_Markov_Chains/10.01%3A_Introduction_to_Markov_Chains.

[56] D Permana et al. "Convergence of Transition Probability Matrix in CLVMarkov Models". In: *IOP Conference Series: Materials Science and Engineering* (2018). DOI: 10.1088/1757-899X/335/1/012046. URL: https://dx.doi.org/10.1088/1757-899X/335/1/012046.

[57] Vladimir Vovk. "The Fundamental Nature of the Log Loss Function". In: *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of His 75th Birthday*. Cham: Springer International Publishing, 2015. ISBN: 978-3-319-23534-9. DOI: 10.1007/978-3-319-23534-9_20. URL: https://doi.org/10.1007/978-3-319-23534-9_20.

[58] Devopedia. *Confusion Matrix*. https://devopedia.org/confusion-matrix. Aug. 2019.

[59] Željko D Vujovic. "Classification Model Evaluation Metrics". In: *International Journal of Advanced Computer Science and Applications* 12 (2021). ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2021.0120670. URL: http://thesai.org/Publications/ViewPaper?Volume=12&Issue=6&Code=IJACSA&SerialNo=70.

[60] scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html.

[61] XGBoost Developers. *Python API Reference — xgboost 3.1.0-dev documentation*. https://xgboost.readthedocs.io/en/latest/python/python_api.html.

[62] Ziqi Li. "Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost". In: *Computers, Environment and Urban Systems* 96 (Sept. 2022). ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2022.101845.

[63] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. "A Review of Evaluation Metrics in Machine Learning Algorithms". In: *Artificial Intelligence Application in Networks and Systems*. Cham: Springer International Publishing, 2023. ISBN: 978-3-031-35314-7. DOI: 10.1007/978-3-031-35314-7_2.

[64] William Howard Beasley and Joseph Lee Rodgers. "Bootstrapping and Monte Carlo methods". In: *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*. APA hand-

books in psychology®. Washington, DC, US: American Psychological Association, 2012. ISBN: 978-1-4338-1005-3. DOI: 10.1037/13620-022.

[65] FBref.com. *2015-2016 Premier League Player Stats*. https://fbref.com/en/comps/9/2015-2016/stats/2015-2016-Premier-League-Stats. Aug. 2016.

[66] Hongyou Liu et al. "Inter-operator reliability of live football match statistics from OPTA Sportsdata". In: *International Journal of Performance Analysis in Sport* (Dec. 2013). ISSN: 2474-8668. DOI: 10.1080/24748668.2013.11868690.

[67] Henrik Biermann et al. "Synchronization of passes in event and spatiotemporal soccer data". In: *Scientific Reports* (Sept. 2023). ISSN: 2045-2322. DOI: 10.1038/s41598-023-39616-2.

[68] Ian Reade, Wendy Rodgers, and Katie Spriggs. "New Ideas for High Performance Coaches: A Case Study of Knowledge Transfer in Sport Science". In: *International Journal of Sports Science & Coaching* (Sept. 2008). ISSN: 1747-9541. DOI: 10.1260/174795408786238533.

[69] Jonathan D. Bartlett and Barry Drust. "A framework for effective knowledge translation and performance delivery of Sport Scientists in professional sport". In: *European Journal of Sport Science* (2021). ISSN: 1536-7290. DOI: 10.1080/17461391.2020.1842511.

[70] Jesse Davis and Lotte Bransen. *Women's Football Analyzed: Interpretable Expected Goals Models for Women*. https://kuleuven.limo.libis.be/discovery/fulldisplay/lirias3469168/32KUL_KUL:Lirias. 2021. (Visited on 12/07/2023).

[71] Becky Field et al. "Using the Knowledge to Action Framework in practice: a citation analysis and systematic review". In: *Implementation Science* (Nov. 2014). ISSN: 1748-5908. DOI: 10.1186/s13012-014-0172-2.

[72] Siu Loon Hoe. "Tacit knowledge, nonaka and takeuchi seci model and informal knowledge processes". In: *International Journal of Organization Theory & Behavior* (Mar. 2006). ISSN: 1093-4537. DOI: 10.1108/IJOTB-09-04-2006-B002.

[73] Valerio Ortenzi et al. "Object Handovers: A Review for Robotics". In: *IEEE Transactions on Robotics* 37 (Dec. 2021). ISSN: 1941-0468. DOI: 10.1109/TRO.2021.3075365.

[74] Grant Mcqueen and Steven Thorley. "Are Stock Returns Predictable? A Test Using Markov Chains". In: *The Journal of Finance* 46 (1991). ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1991.tb03751.x.

Appendix A

Code

A.1 Helper Functions

```
# Function to get the index of the end of a second-ball possession
# Inputs: df: DataFrame containing event-data
#         index: Index of the second-ball winning event
# Outputs: i: Index of the end of the second-ball possession
def get_possession(df, index):
    i = 0
    # Get the second-ball winning team
    winning_team = df.iloc[index]['team']
    # Loop through events as long as the possession continues
    while df.iloc[index + i]['possession_team'] == winning_team and
       df.iloc[index + i]['out'] != True and df.iloc[index+i]['
       pass_outcome'] not in ['Out', 'Pass Offside'] and df.iloc[
       index + i]['type'] not in ['Half Start', 'Half End', 'Match
       Start', 'Match End', 'Foul Committed', 'Foul Won', 'Injury
       Stoppage', 'Bad Behaviour', 'Substitution']:
        #increment index
```

```
        i += 1
        #below are cases where possession ending does not
           necessarily trigger the above loop to break:
        #goal or off target, this is the end of the possession
        if df.iloc[index + i]['shot_outcome'] in ['Goal', 'Off T']:
            break
        # goal keeper save and held (possession change)(EoP)
        if df.iloc[index + i]['shot_outcome'] == 'Saved' and df.iloc
           [index + i + 1]['goalkeeper_type'] == 'Shot Saved' and (
           df.iloc[index + i + 1]['goalkeeper_outcome'] == 'Success'
            or df.iloc[index + i]['goalkeeper_outcome'] == 'Saved
           Twice'):
            break
        # goal keeper save and then out (EoP)
        if df.iloc[index + i]['shot_outcome'] == 'Saved' and df.iloc
           [index + i + 1]['goalkeeper_type'] == 'Shot Saved' and df
           .iloc[index + i + 1]['goalkeeper_outcome'] == 'Touched
           Out':
            break
    # If the loop ends without a break, it means possession ended
       another way
    return i
```

```
# Function to determine if the first two actions after a second-ball
    winning event are successful
# Inputs: df: DataFrame containing event data
#         index: Index of the event to start from
#         team: Team that won the second ball
# Outputs: True if the next two actions are successful, False
```

```
   otherwise
def next_action_successful(df, index, team):
    #Counter for consecutive passes
    consec_pass = 0
    #starting index of possession, tracks index if needed. Just used
        to check if None in our case.
    start_of_possession = None

    #simple index shift, regular index used in future work.
    index = index - 1

    # Check to see if the first event is a legitimate way to start a
        possession
    if df.iloc[index]['type'] == 'Pass' and pd.isna(df.iloc[index]['
       pass_outcome']):
        consec_pass += 1
        start_of_possession = index
    elif df.iloc[index]['type'] in ['Ball Recovery', 'Carry']:
        start_of_possession = index
    #return false if first event goes out of bounds
    elif df.iloc[index]['out'] == True:
        return False

    i = 1
    # loop through next 5 events, cant be more since from data
       structure
    while (i < 5 and (index + i) < len(df)):
        #get next row
        row = df.iloc[index + i]
```

```
30          # check all cases in which a action fails, aka return false
31          # check for actions where we continue on to next event
32          # check for successful pass and increment counter if so
33          if row['out'] == True:
34              break
35          elif row['type'] == 'Pass' and pd.isna(row['pass_outcome'])
               and row['team'] == team:
36              consec_pass += 1
37          elif row['type'] == 'Pass' and row['team'] != team:
38              break
39          elif row['type'] == 'Pass' and row['pass_outcome'] != '
               Incomplete':
40              break
41          elif row['type'] == 'Duel' and row['team'] != team:
42              pass
43          elif row['type'] == 'Dribble' and row['team'] == team:
44              pass
45          elif row['type'] == 'Dribbled Past' and row['team'] != team:
46              pass
47          elif row['type'] == 'Carry' and row['team'] == team:
48              pass
49          elif row['type'] == 'Foul Committed' and row['team'] != team
               :
50              return True
51          # rare case shot occurs within first 2 actions, so it should
               overide 2 being successful and count
52          elif row['type'] == 'Shot':
53              if start_of_possession == None:
54                  start_of_possession = index + i
```

```
            return True
        else:
            break
        # if we have 2 consecutive passes, we can return true
        if consec_pass == 2:
            return True
        i += 1

    return False
```

```
# Function to determine the outcome of a transition period.
# Inputs: df: DataFrame containing event-data
#         index: Index of the event of the start of the second-ball
#         period: Duration of the transition period in seconds
# Outputs: team: Team that has possession after the transition, or None if no team has possession
#         start_of_possession: Index of the start of the possession, or None if no possession starts
def transition_outcome(df, index, period):
    #setup minute and second
    minute = df.iloc[index]['minute']
    second = (df.iloc[index]['second'] + period) % 60

    #get the team
    team = df.iloc[index]['team']
    consec_pass = 0
    start_of_possession = None

    if df.iloc[index]['type'] == 'Pass' and pd.isna(df.iloc[index]['
```

```
        pass_outcome']):
18          consec_pass += 1
19          start_of_possession = index
20      elif df.iloc[index]['out'] == True:
21          return None, None
22
23      #work around for the fact that we are using a second-based index
24      i = 1
25      if (df.iloc[index]['second'] + period > 54):
26          minute += 1
27      while ((df.iloc[index + i]['minute'] == minute and (df.iloc[
            index + i]['second'] % 60) < second) or (df.iloc[index + i]['
            minute'] == minute - 1)):
28          # Check if the event is out of bounds
29          if df.iloc[index + i]['out'] == True:
30              return None, None
31          # Handle cosecutive passes even if the team changes
32          elif df.iloc[index + i]['type'] == 'Pass' and pd.isna(df.
                iloc[index + i]['pass_outcome']) and df.iloc[index + i]['
                team'] == team:
33              consec_pass += 1
34              if consec_pass == 1:
35                  start_of_possession = index + i
36          elif df.iloc[index + i]['type'] == 'Pass' and pd.isna(df.
                iloc[index + i]['pass_outcome']) and df.iloc[index + i]['
                team'] != team:
37              team = df.iloc[index + i]['team']
38              consec_pass = 1
39              start_of_possession = index + i
```

```
        # Incomplete pass resets counters
        elif df.iloc[index + i]['type'] == 'Pass' and df.iloc[index
           + i]['pass_outcome'] == 'Incomplete':
            #print('test 1')
            consec_pass = 0
            start_of_possession = None
        # Events that terminate the window
        elif df.iloc[index + i]['type'] in ['Half End', 'Match End',
            'Injury Stoppage', 'Bad Behaviour']:
            return None, None
        elif df.iloc[index+i]['type'] == 'Error':
            start_of_possession = None
        # Rare case where shot occurs within the transition period
        elif df.iloc[index + i]['type'] == 'Shot':
            if start_of_possession == None:
                start_of_possession = index + i
            return df.iloc[index + i]['team'], start_of_possession
        # return if we have 2 consecutive passes.
        if consec_pass == 2:
            return team, start_of_possession
        i += 1

    return None, None
```

A.2 Second Ball Extracting Functions

```
# Chain ABA (A) Long Ball -> B Incomplete Pass -> A gets Ball
```

```
# This function identifies ABA Chains, including those with a transition period
# Inputs: eventdf: DataFrame containing event data with columns like 'type', 'team', 'pass_outcome', etc.
#         longballdf: DataFrame containing long ball events with 'pass_outcome'
# Outputs: aba_seq: List of sequences representing ABA chains, where each sequence is a DataFrame slice
def get_aba(eventdf, longballdf):
    aba_seq = []
    count = 0
    aba = longballdf[longballdf['pass_outcome'] == 'Incomplete']
    for i in range(len(aba)):
        index = aba.iloc[i].name
        a = eventdf.iloc[index]['team']
        if ((((eventdf.iloc[index + 1]['type'] == 'Pass' and eventdf.iloc[index + 1]['pass_outcome'] == 'Incomplete') or eventdf.iloc[index + 1]['type'] == 'Clearance' or eventdf.iloc[index + 1]['goalkeeper_type'] == 'Punch') and eventdf.iloc[index + 1]['team'] != a and eventdf.iloc[index+1]['out'] != True)
        and
        ((eventdf.iloc[index + 2]['type'] in ['Pass', 'Ball Recovery', 'Shot'] and eventdf.iloc[index + 2]['team'] == a) and eventdf.iloc[index + 2]['type'] != 'Duel')):
            if next_action_successful(eventdf, index + 3, a):
                eventdf.loc[index, ['2bchain', '2bposs']] = ['aba', index + 2]
```

```
18                longballdf.loc[index, ['2bchain', '2bposs']] = ['aba
                      ', index + 2]
19                end_index = get_possession(eventdf, index + 2)
20                sequence = eventdf.iloc[index:index+3+end_index]
21                aba_seq.append(sequence)
22            else:
23                trans, poss_index = transition_outcome(eventdf,
                      index + 2, 5)
24                if trans == None or poss_index == None:
25                    count += 1
26                elif trans == a:
27                    eventdf.loc[index, ['2bchain', '2bposs']] = ['
                          abata', poss_index]
28                    longballdf.loc[index, ['2bchain', '2bposs']] = [
                          'abata', poss_index]
29                    end_index = get_possession(eventdf, poss_index)
30                    sequence = eventdf.iloc[index:poss_index+
                          end_index]
31                    aba_seq.append(sequence)
32                else:
33                    eventdf.loc[index, ['2bchain', '2bposs']] = ['
                          abatb', poss_index]
34                    longballdf.loc[index, ['2bchain', '2bposs']] = [
                          'abatb', poss_index]
35                    end_index = get_possession(eventdf, poss_index)
36                    sequence = eventdf.iloc[index:poss_index+
                          end_index]
37                    aba_seq.append(sequence)
38    # Uncomment to see the count of ABA chains without a transition
```

```
    #print(count)
    return aba_seq
```

```
# Chain ABB (A) Long Ball -> B Incomplete Pass -> B gets Ball
# This function identifies ABB Chains, including those with a transition period
# Inputs: eventdf: DataFrame containing event data with columns like 'type', 'team', 'pass_outcome', etc.
#         longballdf: DataFrame containing long ball events with 'pass_outcome'
# Outputs: abb_seq: List of sequences representing ABB chains, where each sequence is a DataFrame slice
def get_abb(eventdf, longballdf):
    abb_seq = []
    count = 0
    abb = longballdf[longballdf['pass_outcome'] == 'Incomplete']
    for i in range(len(abb)):
        index = abb.iloc[i].name
        a = eventdf.iloc[index]['team']
        if ((((eventdf.iloc[index + 1]['type'] == 'Pass' and pd.isna(eventdf.iloc[index + 1]['pass_outcome']))\
            or eventdf.iloc[index+1]['goalkeeper_type'] == 'Punch')\
             and eventdf.iloc[index + 1]['team'] != a and eventdf.iloc[index+1]['out'] != True) \
        and
        (eventdf.iloc[index + 2]['type'] in ['Carry', 'Shot', 'Pass'] and eventdf.iloc[index + 2]['team'] != a)):
            if next_action_successful(eventdf, index + 3, eventdf.
```

```
                iloc[index + 2]['team']):
19                  eventdf.loc[index, ['2bchain', '2bposs']] = ['abb',
                        index + 2]
20                  longballdf.loc[index, ['2bchain', '2bposs']] = ['abb
                        ', index + 2]
21                  end_index = get_possession(eventdf, index + 2)
22                  sequence = eventdf.iloc[index:index+3+end_index]
23                  abb_seq.append(sequence)
24              else:
25                  trans, poss_index = transition_outcome(eventdf,
                        index + 2, 5)
26                  if trans == None or poss_index == None:
27                      count += 1
28                  elif trans == a:
29                      eventdf.loc[index, ['2bchain', '2bposs']] = ['
                            abbta', poss_index]
30                      longballdf.loc[index, ['2bchain', '2bposs']] = [
                            'abbta', poss_index]
31                      end_index = get_possession(eventdf, poss_index)
32                      sequence = eventdf.iloc[index:poss_index+
                            end_index]
33                      abb_seq.append(sequence)
34                  else:
35                      eventdf.loc[index, ['2bchain', '2bposs']] = ['
                            abbtb', poss_index]
36                      longballdf.loc[index, ['2bchain', '2bposs']] = [
                            'abbtb', poss_index]
37                      end_index = get_possession(eventdf, poss_index)
38                      sequence = eventdf.iloc[index:poss_index+
```

```
                        end_index]
                    abb_seq.append(sequence)
    # Uncomment to see the count of ABB chains without a transition
    #print(count)
    return abb_seq
```

```
# Chain ADAA (A) Long Ball -> Duel -> A Duel Win -> Team A gets Ball
# This function identifies ADAA Chains, including those with a transition period
# Inputs: eventdf: DataFrame containing event data with columns like 'type', 'team', 'pass_outcome', etc.
#         longballdf: DataFrame containing long ball events with 'pass_outcome'
# Outputs: adaa_seq: List of sequences representing ADAA chains, where each sequence is a DataFrame slice
def get_adaa(eventdf, longballdf):
    adaa_seq = []
    count = 0

    adaa = longballdf[pd.isna(longballdf['pass_outcome'])]
    for i in range(len(adaa)):
        index = adaa.iloc[i].name
        a = eventdf.iloc[index]['team']
        if ((eventdf.iloc[index + 1]['type'] == 'Duel'
        and
        ((eventdf.iloc[index + 2]['type'] == 'Pass' and pd.isna(
            eventdf.iloc[index + 2]['pass_outcome']) and eventdf.iloc
            [index+2]['team'] == a)
        or
```

```
18          (eventdf.iloc[index + 2]['type'] == 'Clearance' and eventdf.
               iloc[index + 2]['out'] != True and eventdf.iloc[index +
               2]['team'] == a))
19          and
20          ((eventdf.iloc[index + 3]['type'] in ['Pass', 'Carry', 'Shot
               ', 'Ball Recovery'] and eventdf.iloc[index + 3]['team']
               == a)))
21          or
22          (eventdf.iloc[index + 1]['type'] == 'Pass' and eventdf.iloc[
               index + 1]['team'] == a and pd.isna(eventdf.iloc[index +
               1]['pass_outcome']) and eventdf.iloc[index + 2]['type']
               == 'Duel' and (eventdf.iloc[index + 3]['type'] == 'Ball
               Recovery' or (eventdf.iloc[index + 3]['type'] == 'Pass'))
                and eventdf.iloc[index + 3]['team'] == a)
23          ):
24              if next_action_successful(eventdf, index + 4, a):
25                  eventdf.loc[index, ['2bchain', '2bposs']] = ['adaa',
                        index + 3]
26                  longballdf.loc[index, ['2bchain', '2bposs']] = ['
                       adaa', index + 3]
27                  end_index = get_possession(eventdf, index + 3)
28                  sequence = eventdf.iloc[index:index+4+end_index]
29                  adaa_seq.append(sequence)
30              else:
31                  trans, poss_index = transition_outcome(eventdf,
                       index + 3, 5)
32                  if trans == None or poss_index == None:
33                      count += 1
34                  elif trans == a:
```

```
                    eventdf.loc[index, ['2bchain', '2bposs']] = ['
                        adaata', poss_index]
                    longballdf.loc[index, ['2bchain', '2bposs']] = [
                        'adaata', poss_index]
                    end_index = get_possession(eventdf, poss_index)
                    sequence = eventdf.iloc[index:poss_index+
                        end_index]
                    adaa_seq.append(sequence)
                else:
                    eventdf.loc[index, ['2bchain', '2bposs']] = ['
                        adaatb', poss_index]
                    longballdf.loc[index, ['2bchain', '2bposs']] = [
                        'adaatb', poss_index]
                    end_index = get_possession(eventdf, poss_index)
                    sequence = eventdf.iloc[index:poss_index+
                        end_index]
                    adaa_seq.append(sequence)
    # Uncomment to see the count of ADAA chains without a transition
    #print(count)
    return adaa_seq
```

```
# Chain ADAB (A) Long Ball -> Duel -> A Duel Win -> Team B gets Ball
# This function identifies ADAB Chains, including those with a
    transition period
# Inputs: eventdf: DataFrame containing event data with columns like
    'type', 'team', 'pass_outcome', etc.
#         longballdf: DataFrame containing long ball events with '
    pass_outcome'
# Outputs: adab_seq: List of sequences representing ADAB chains,
```

```
    where each sequence is a DataFrame slice
def get_adab(eventdf, longballdf):
    adab_seq = []
    count = 0
    adab = longballdf[pd.isna(longballdf['pass_outcome'])]
    for i in range(len(adab)):
        index = adab.iloc[i].name
        a = eventdf.iloc[index]['team']
        if (eventdf.iloc[index + 1]['type'] == 'Duel'
        and
        ((eventdf.iloc[index + 2]['type'] == 'Pass' and eventdf.iloc
           [index + 2]['pass_outcome'] == 'Incomplete' and eventdf.
           iloc[index+2]['team'] == a)
        or
        (eventdf.iloc[index + 2]['type'] == 'Clearance' and eventdf.
           iloc[index + 2]['out'] != True and eventdf.iloc[index +
           2]['team'] == a)
        or
        (eventdf.iloc[index + 2]['type'] == 'Miscontrol' and eventdf
           .iloc[index + 2]['team'] == a))
        and
        ((eventdf.iloc[index + 3]['type'] in ['Pass', 'Interception'
           , 'Shot', 'Ball Recovery', 'Carry', 'Goalkeeper'] and
           eventdf.iloc[index + 3]['team'] != a))):
            if next_action_successful(eventdf, index + 4, eventdf.
               iloc[index + 3]['team']):
                eventdf.loc[index, ['2bchain', '2bposs']] = ['adab',
                    index + 3]
                longballdf.loc[index, ['2bchain', '2bposs']] = ['
```

```
                    adab', index + 3]
                end_index = get_possession(eventdf, index + 3)
                sequence = eventdf.iloc[index:index+4+end_index]
                adab_seq.append(sequence)
            else:
                trans, poss_index = transition_outcome(eventdf,
                    index + 3, 5)
                if trans == None or poss_index == None:
                    count += 1
                elif trans == a:
                    eventdf.loc[index, ['2bchain', '2bposs']] = ['
                        adabta', poss_index]
                    longballdf.loc[index, ['2bchain', '2bposs']] = [
                        'adabta', poss_index]
                    end_index = get_possession(eventdf, poss_index)
                    sequence = eventdf.iloc[index:poss_index+
                        end_index]
                    adab_seq.append(sequence)
                else:
                    eventdf.loc[index, ['2bchain', '2bposs']] = ['
                        adabtb', poss_index]
                    longballdf.loc[index, ['2bchain', '2bposs']] = [
                        'adabtb', poss_index]
                    end_index = get_possession(eventdf, poss_index)
                    sequence = eventdf.iloc[index:poss_index+
                        end_index]
                    adab_seq.append(sequence)
    # Uncomment to see the count of ADAB chains without a transition
    #print(count)
```

```
    return adab_seq
```

```
#Chain ADBA (A) Long Ball -> Duel -> B Duel Win -> Team A gets Ball
# This function identifies ADBA Chains, including those with a transition period
# Inputs: eventdf: DataFrame containing event data with columns like 'type', 'team
#         longballdf: DataFrame containing long ball events with 'pass_outcome'
# Outputs: adba_seq: List of sequences representing ADBA chains, where each sequence

def get_adba(eventdf, longballdf):
    adba_seq = []
    count = 0
    adba = longballdf[longballdf['pass_outcome'] == 'Incomplete']
    for i in range(len(adba)):
        a = adba.iloc[i].team
        index = adba.iloc[i].name
        if ((eventdf.iloc[index + 1]['type'] == 'Duel'
        and
        ((eventdf.iloc[index + 2]['type'] == 'Pass' and eventdf.iloc[index + 2]['pass_outcome'] == 'Incomplete' and eventdf.iloc[index+2]['team'] != a)
        or
        (eventdf.iloc[index + 2]['type'] == 'Clearance' and eventdf.iloc[index + 2]['out'] != True and eventdf.iloc[index + 2]['team'] != a))
        and
```

```
        (eventdf.iloc[index + 3]['type'] in ['Pass', 'Ball Recovery', 'Carry', 'Shot'] and eventdf.iloc[index + 3]['team'] == a))
        or
        (eventdf.iloc[index + 1]['type'] == 'Pass' and eventdf.iloc[index + 1]['team'] != a and eventdf.iloc[index + 1]['pass_outcome'] == 'Incomplete' and eventdf.iloc[index + 2]['type'] == 'Duel' and (eventdf.iloc[index + 3]['type'] in ['Ball Recovery', 'Pass', 'Carry', 'Shot'] and eventdf.iloc[index + 3]['team'] == a))
        ):
            if next_action_successful(eventdf, index + 4, a):
                eventdf.loc[index, ['2bchain', '2bposs']] = ['adba', index + 3]
                longballdf.loc[index, ['2bchain', '2bposs']] = ['adba', index + 3]
                end_index = get_possession(eventdf, index + 3)
                sequence = eventdf.iloc[index:index+4+end_index]
                adba_seq.append(sequence)
            else:
                trans, poss_index = transition_outcome(eventdf, index + 3, 5)
                if trans == None or poss_index == None:
                    count += 1
                elif trans == a:
                    eventdf.loc[index, ['2bchain', '2bposs']] = ['adbata', poss_index]
                    longballdf.loc[index, ['2bchain', '2bposs']] = ['adbata', poss_index]
```

```
                    end_index = get_possession(eventdf, poss_index)
                    sequence = eventdf.iloc[index:poss_index+
                        end_index]
                    adba_seq.append(sequence)
                else:
                    eventdf.loc[index, ['2bchain', '2bposs']] = ['
                        adbatb', poss_index]
                    longballdf.loc[index, ['2bchain', '2bposs']] = [
                        'adbatb', poss_index]
                    end_index = get_possession(eventdf, poss_index)
                    sequence = eventdf.iloc[index:poss_index+
                        end_index]
                    adba_seq.append(sequence)
    # Uncomment to see the count of ADBA chains without a transition
    #print(count)

    return adba_seq
```

```
# Chain ADBB (A) Long Ball -> Duel -> B Duel Win -> Team B gets Ball
# This function identifies ADBB Chains, including those with a
    transition period
# Inputs: eventdf: DataFrame containing event data with columns like
    'type', 'team', 'pass_outcome', etc.
#         longballdf: DataFrame containing long ball events with '
    pass_outcome'
# Outputs: adbb_seq: List of sequences representing ADBB chains,
    where each sequence is a DataFrame slice
def get_adbb(eventdf, longballdf):
    adbb_seq = []
```

```
    count = 0
    adbb = longballdf[longballdf['pass_outcome'] == 'Incomplete']
    for i in range(len(adbb)):
        a = adbb.iloc[i].team
        index = adbb.iloc[i].name
        if (eventdf.iloc[index + 1]['type'] == 'Duel'
        and
        ((eventdf.iloc[index + 2]['type'] == 'Pass' and pd.isna(
           eventdf.iloc[index + 2]['pass_outcome']))
        or
        (eventdf.iloc[index + 2]['type'] == 'Clearance' and \
            eventdf.iloc[index + 2]['out'] != True and eventdf.iloc[
               index + 2]['team'] != a))
        and
        (eventdf.iloc[index + 3]['type'] in ['Pass', 'Ball Recovery'
           , 'Carry', 'Shot'] and \
         eventdf.iloc[index + 3]['team'] != a)):
            if next_action_successful(eventdf, index + 4, eventdf.
               iloc[index + 3]['team']):
                eventdf.loc[index, ['2bchain', '2bposs']] = ['adbb',
                    index + 3]
                longballdf.loc[index, ['2bchain', '2bposs']] = ['
                   adbb', index + 3]
                end_index = get_possession(eventdf, index + 3)
                sequence = eventdf.iloc[index:index+4+end_index]
                adbb_seq.append(sequence)
            else:
                trans, poss_index = transition_outcome(eventdf,
                   index + 3, 5)
```

```
30                  if trans == None or poss_index == None:
31                      count += 1
32                  elif trans == a:
33                      eventdf.loc[index, ['2bchain', '2bposs']] = ['
                            adbbta', poss_index]
34                      longballdf.loc[index, ['2bchain', '2bposs']] = [
                            'adbbta', poss_index]
35                      end_index = get_possession(eventdf, poss_index)
36                      sequence = eventdf.iloc[index:poss_index+
                            end_index]
37                      adbb_seq.append(sequence)
38                  else:
39                      eventdf.loc[index, ['2bchain', '2bposs']] = ['
                            adbbtb', poss_index]
40                      longballdf.loc[index, ['2bchain', '2bposs']] = [
                            'adbbtb', poss_index]
41                      end_index = get_possession(eventdf, poss_index)
42                      sequence = eventdf.iloc[index:poss_index+
                            end_index]
43                      adbb_seq.append(sequence)
44      #print(count) #Uncomment to see the count of ADBB chains without
            a transition
45      return adbb_seq
```

Appendix B

Figures & Tables

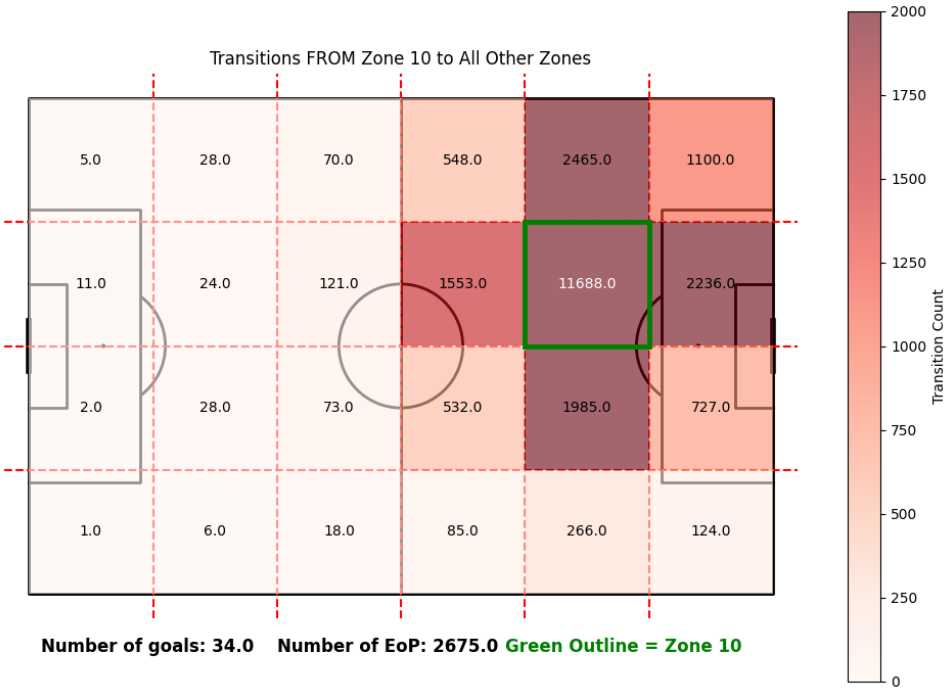


Figure B.1: Transition counts for each zone when ball is in zone 10.

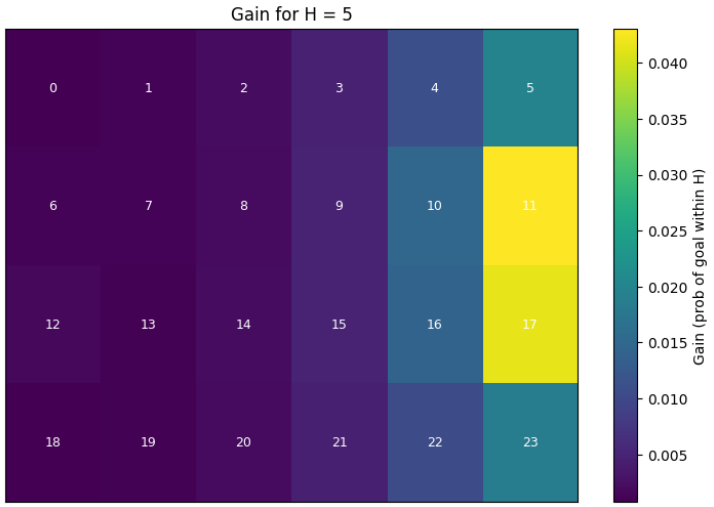


Figure B.2: Heat map of gain values by zone for $H = 5$.

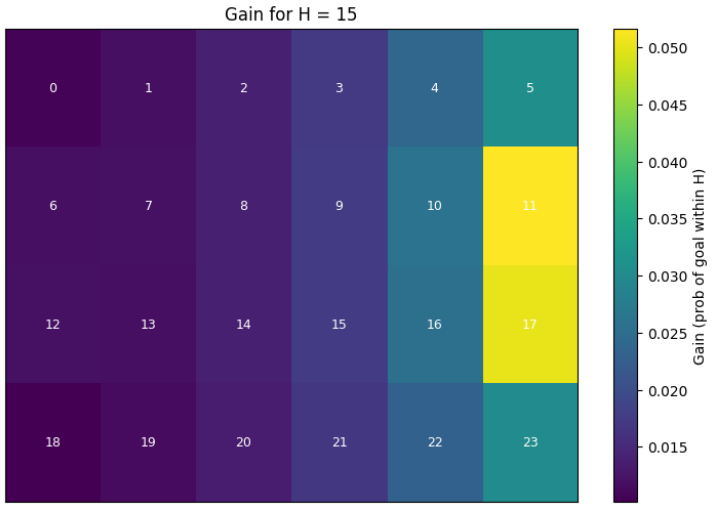


Figure B.3: Heat map of gain values by zone for $H = 15$.

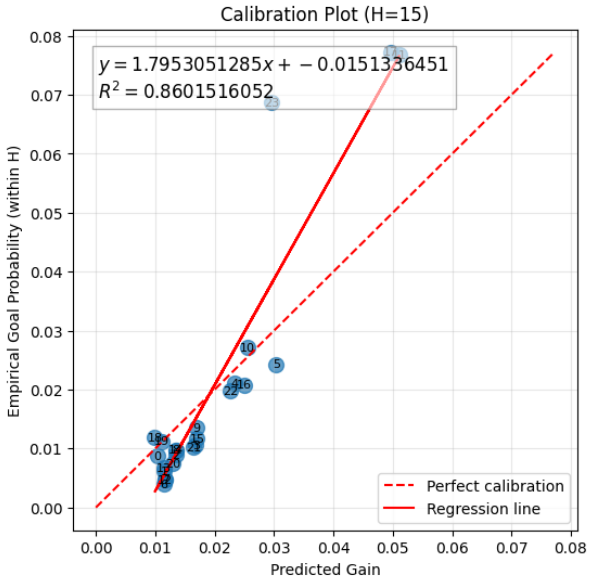


Figure B.4: Calibration plot for predicted gain vs empircal goal probability for $H = 15$ including all zones.

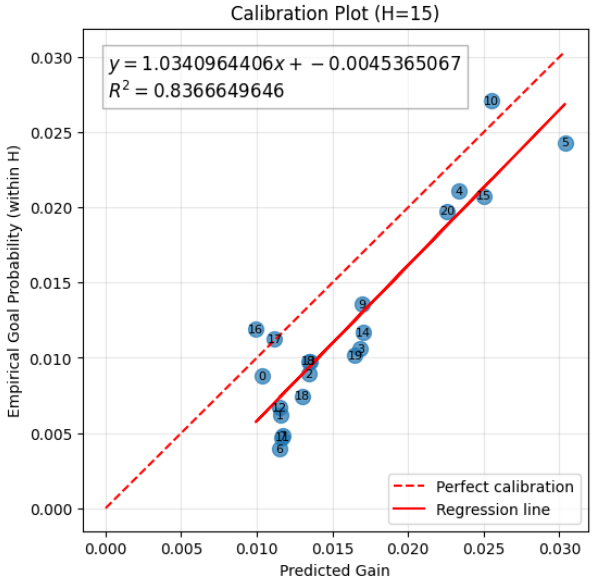


Figure B.5: Calibration plot for predicted gain vs empircal goal probability for $H = 15$ excluding outliers.

Team	Avg Gain	Total Gain	Count
Leicester City	0.0185	9.3671	509
Southampton	0.0184	9.6390	533
West Brom	0.0182	8.2349	460
Crystal Palace	0.0180	9.2755	522
Watford	0.0180	7.9922	451
Norwich City	0.0178	8.9163	508
Aston Villa	0.0176	7.8629	452
West Ham United	0.0175	8.0845	473
Sunderland	0.0173	7.8055	455
Liverpool	0.0171	9.1233	541
Everton	0.0170	8.0102	479
Manchester City	0.0169	7.1949	431
AFC Bournemouth	0.0168	7.4132	450
Arsenal	0.0167	8.7294	531
Stoke City	0.0166	7.5005	458
Manchester United	0.0166	9.5116	585
Tottenham Hotspur	0.0166	8.2034	502
Swansea City	0.0163	8.1407	504
Newcastle United	0.0163	6.7018	413
Chelsea	0.0160	7.3557	467

Table B.1: Team-level average gain, total gain, and count of second ball events. Teams are ranked by average gain.

Zone	Gain	95% CI	Std
0	0.0007623	[0.0006618, 0.0008617]	0.0000519
1	0.0011480	[0.0010424, 0.0012571]	0.0000544
2	0.0021658	[0.0019934, 0.0023455]	0.0000925
3	0.0046661	[0.0043048, 0.0050421]	0.0001914
4	0.0110430	[0.0102119, 0.0119276]	0.0004498
5	0.0194507	[0.0179488, 0.0211324]	0.0008167
6	0.0012362	[0.0010989, 0.0013802]	0.0000714
7	0.0011391	[0.0010266, 0.0012485]	0.0000557
8	0.0020562	[0.0019084, 0.0022288]	0.0000838
9	0.0048862	[0.0045167, 0.0052549]	0.0001946
10	0.0145774	[0.0133351, 0.0158330]	0.0006526
11	0.0428093	[0.0390903, 0.0469527]	0.0020279
12	0.0016369	[0.0009820, 0.0025300]	0.0003937
13	0.0010587	[0.0009457, 0.0011830]	0.0000609
14	0.0020712	[0.0018946, 0.0022810]	0.0000995
15	0.0049489	[0.0045685, 0.0053485]	0.0002045
16	0.0141377	[0.0129694, 0.0154899]	0.0006689
17	0.0412533	[0.0374997, 0.0452974]	0.0020001
18	0.0007887	[0.0006022, 0.0010408]	0.0001148
19	0.0010659	[0.0009335, 0.0012303]	0.0000762
20	0.0019702	[0.0018078, 0.0021378]	0.0000881
21	0.0044324	[0.0040699, 0.0048108]	0.0001885
22	0.0103872	[0.0095578, 0.0112826]	0.0004472
23	0.0185298	[0.0169635, 0.0202002]	0.0008169

Table B.2: Bootstrapped median gain, 95% confidence intervals, and standard deviation by zone. $H = 5$

Zone	Gain	95% CI	Std
0	0.0103227	[0.0095061, 0.0111350]	0.0004182
1	0.0115497	[0.0106450, 0.0124635]	0.0004562
2	0.0134523	[0.0124142, 0.0145195]	0.0005296
3	0.0168522	[0.0155498, 0.0181964]	0.0006675
4	0.0233317	[0.0215104, 0.0252198]	0.0009408
5	0.0303773	[0.0279459, 0.0329114]	0.0012594
6	0.0114976	[0.0106350, 0.0123993]	0.0004486
7	0.0117238	[0.0108507, 0.0126383]	0.0004527
8	0.0134944	[0.0124923, 0.0145504]	0.0005233
9	0.0169788	[0.0157035, 0.0182997]	0.0006590
10	0.0255183	[0.0234010, 0.0275864]	0.0010516
11	0.0511123	[0.0465322, 0.0559680]	0.0023693
12	0.0116745	[0.0105738, 0.0129540]	0.0005953
13	0.0115311	[0.0106945, 0.0124270]	0.0004458
14	0.0134344	[0.0124567, 0.0145032]	0.0005203
15	0.0170523	[0.0157845, 0.0183523]	0.0006551
16	0.0250136	[0.0230176, 0.0271035]	0.0010377
17	0.0496471	[0.0454654, 0.0541503]	0.0022269
18	0.0099276	[0.0090944, 0.0107384]	0.0004225
19	0.0111735	[0.0103485, 0.0120686]	0.0004351
20	0.0129948	[0.0120719, 0.0140201]	0.0005017
21	0.0164906	[0.0152287, 0.0177659]	0.0006430
22	0.0225875	[0.0208606, 0.0244611]	0.0008982
23	0.0296584	[0.0273897, 0.0321136]	0.0012052

Table B.3: Bootstrapped median gain, 95% confidence intervals, and standard deviation by zone. $H = 15$

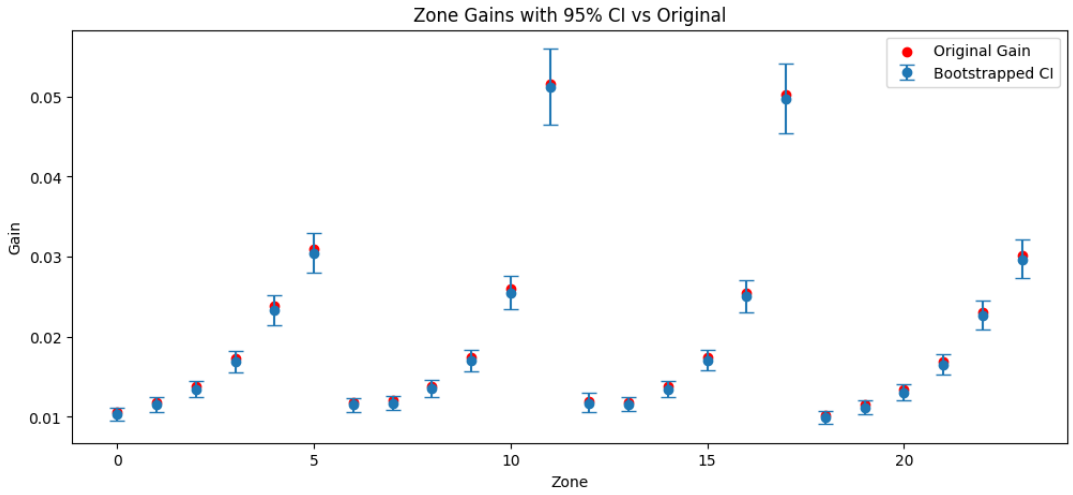


Figure B.6: Scatter plot comparing the original gain to the bootstrapped median gain with error bars for the 95% confidence intervals. $H = 15$.

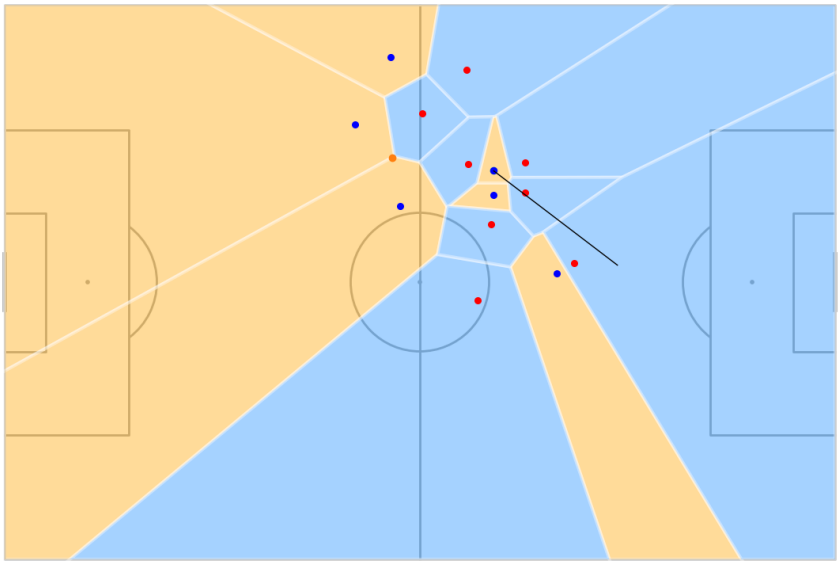


Figure B.7: Voronoi diagram for a pass coupled with StatsBomb 360 data.