

**ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ**



**ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS**

**SCHOOL OF INFORMATION SCIENCES &  
TECHNOLOGY**

**DEPARTMENT OF STATISTICS  
POSTGRADUATE PROGRAM**

**Goal Scoring Performance Prediction of Soccer  
Athletes**

by

**Panagiotis C. Zoris**

**A THESIS**

Submitted to the Department of Statistics of the Athens University  
of Economics and Business in partial fulfilment of the requirements  
for the degree of Master of Science in Statistics

Athens, Greece

September 2023





**ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ**



**ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS**

**ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ**

**ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ**

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ**

**Πρόβλεψη της επίδοσης σκοραρίσματος των  
ποδοσφαιριστών**

**Παναγιώτης Χ. Ζώρης**

**ΔΙΑΤΡΙΒΗ**

Που υποβλήθηκε στο Τμήμα Στατιστικής του Οικονομικού  
Πανεπιστημίου Αθηνών ως μέρος των απαιτήσεων για την απόκτηση  
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα, Ελλάδα

Σεπτέμβριος 2023





## ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Ioannis Ntzoufras for his consistent support and guidance during the running of this thesis.

Furthermore, I would like to thank Bill Mexias and Dennis Tsalikis, the Chief Operations Officer (COO) and Chief Executive Officer (CEO) respectively of Fantasy Sports Interactive (FSI), who offered this thesis and who gave me the opportunity to start my professional career as a data analyst in their company.

I would also like to thank my colleagues, Vasileios Palaskas and Spiros Kolovos, for providing advice regarding analysis.

Finally, many thanks to my parents, Chrysostomos and Konstantina, as well as to my sister, Sofia, for supporting me during the compilation of this thesis.





## VITA

I have always loved making predictions about sports and that was the main reason I studied statistics when I finished high school.

After graduating from university, I worked as a data analyst at Hellenic Competition Commission as I wanted to gain work experience before starting my postgraduate studies.

Choosing the master's program was an easy decision for me, with the Statistics department of the Athens University of Economics and Business (AUEB) being my only choice because of the professors since my undergraduate studies, as well as the important research that is done in the field of sports analytics.

As a postgraduate student, I participated in the 6<sup>th</sup> AUEB Sports Analytics Workshop and after the end of the two semesters of the master, I started working as an analytics intern at FSI. Before the end of my internship, I accepted an offer for the position of junior data analyst from FSI, where I worked until May 2023.





## ABSTRACT

Panagiotis C. Zoris

### **Goal Scoring Performance Prediction of Soccer Athletes**

This thesis is conducted within the framework of the internship provided by FSI. FSI is a software provider for the Sports Betting and Gaming Industry. In this thesis, the purpose is to propose a statistical learning approach for the prediction of the number of goals scored by each soccer athlete in future matches. From the perspective of FSI, those predictions can be used either as future projections of the athletes or as corresponding betting odds markets of those ones.

In this approach, we will use as covariates in our models the available pre-match information which is summarized in two levels: a) team-level statistics and b) soccer athlete-level statistics. Those statistics will be selected based on the Least Absolute Shrinkage and Selection Operator (LASSO) to identify the most important factors affecting the goal scoring performance of soccer athletes.

Using our suggested modeling technique, we calculate the goal scoring probabilities of a list of athletes for various matchdays of the English Premier League in season 2022-2023 and we compare them with those given by the bookies to examine the accuracy of our model in relation to that of bookies.



## ΠΕΡΙΛΗΨΗ

Παναγιώτης Χ. Ζώρης

### **Πρόβλεψη της επίδοσης σκοραρίσματος των ποδοσφαιριστών**

Η παρούσα διπλωματική εργασία διεξάγεται στα πλαίσια της πρακτικής άσκησης που παρέχεται από την FSI. Η FSI είναι πάροχος λογισμικού για την βιομηχανία αθλητικών στοιχημάτων και τυχερών παιχνιδιών. Σε αυτή τη διατριβή, σκοπός είναι να προταθεί μια προσέγγιση στατιστικής μάθησης για την πρόβλεψη του αριθμού των γκολ που θα σημειωθούν από κάθε αθλητή σε μελλοντικούς αγώνες. Από την πλευρά της FSI, αυτές οι προβλέψεις μπορούν να χρησιμοποιηθούν είτε ως μελλοντικές προβλέψεις των αθλητών είτε ως αγορές αποδόσεων στοιχήματος αυτών.

Σε αυτήν την προσέγγιση θα χρησιμοποιήσουμε ως μεταβλητές στα μοντέλα μας την διαθέσιμη πληροφορία πριν από κάθε αγώνα οι οποίες συνοψίζονται σε δύο επίπεδα: α) ομαδικά στατιστικά και β) ατομικά στατιστικά. Αυτά τα στατιστικά θα επιλεγούν με βάση την μέθοδο LASSO για τον προσδιορισμό των σημαντικότερων παραγόντων που επηρεάζουν την απόδοση σκοραρίσματος των αθλητών.

Χρησιμοποιώντας την προτεινόμενη τεχνική μοντελοποίησης μας, υπολογίζουμε τις πιθανότητες σκοραρίσματος μιας συγκεκριμένης λίστας αθλητών για διάφορες αγωνιστικές της αγγλικής Premier League τη σεζόν 2022-2023 και τις συγκρίνουμε με αυτές που δίνουν οι στοιχηματικές εταιρίες προκειμένου να εξετάσουμε την ακρίβεια μοντέλου μας σε σχέση με αυτή των στοιχηματικών εταιριών.



# Contents

<b>1</b>	<b>Introduction .....</b>	<b>1</b>
1.1	Contextualization .....	1
1.1.1	Sports Analytics .....	2
1.1.2	Soccer Analytics .....	3
1.1.3	Sports Betting .....	4
1.1.4	Soccer Betting .....	6
1.2	Motivation .....	7
1.3	Objectives .....	8
1.4	Master Thesis Outline .....	8
<b>2</b>	<b>Related Work .....</b>	<b>9</b>
2.1	Goal Prediction Landscape .....	9
2.2	Elo Ratings .....	12
2.3	Athletes' Goals Scored Prediction .....	13
<b>3</b>	<b>Data .....</b>	<b>14</b>
3.1	Dataset .....	14
3.1.1	Data Origin .....	14
3.1.2	Data Pre-Processing .....	15
3.2	Feature Engineering .....	17
3.2.1	Pre-Match Skill Actions .....	18
3.2.2	Elo Rating System .....	22
3.2.2.1	Mathematical Model .....	22
3.2.2.2	The K-factor .....	23
3.2.2.3	The Expected Wining Probability E ...	23
3.2.2.4	The Home Advantage H .....	25
3.2.2.5	Tuning the Elo parameters H and C .....	25
3.2.2.6	Results .....	26



3.3	Exploratory Data Analysis (EDA) .....	31
3.3.1	The Dependent Variable: Goals .....	32
3.3.2	Relationships Between Variables .....	33
3.3.2.1	Goals and Home Advantage .....	35
3.3.2.2	Goals and Position .....	36
3.3.2.3	Goals and Pre-Match Skill Actions .....	37
<b>4</b>	<b>Models .....</b>	<b>38</b>
4.1	Model Theory .....	38
4.1.1	Linear Mixed Models .....	39
4.1.2	Generalized Linear Mixed Models .....	41
4.2	Model Selection .....	42
4.2.1	Model Selection Criteria .....	42
4.2.1.1	Akaike Information Criterion .....	43
4.2.1.2	Bayesian Information Criterion .....	43
4.2.2	Model Selection by Optimization .....	44
4.2.2.1	Lasso .....	44
4.2.2.2	Lasso for GLMM .....	45
4.2.3	Results .....	46
4.2.4	Final Model .....	51
4.2.4.1	Interpretation .....	52
4.2.4.2	Assumptions .....	55
4.3	Model Evaluation .....	58
4.3.1	Bookmakers' Odds .....	59
4.3.2	Evaluation Metrics .....	60
4.3.2.1	Log Loss .....	61
4.3.2.2	AUC-ROC .....	62
4.3.3	Training and Test Sets .....	64
4.3.4	Comparison with Bookies .....	64
<b>5</b>	<b>Conclusions and Future Work .....</b>	<b>67</b>
	<b>References .....</b>	<b>69</b>
	<b>Appendix: Opta - Skill Actions .....</b>	<b>73</b>



# 1 Introduction

This project is a master's thesis for the Department of Statistics at the Athens University of Economics and Business (AUEB), which offers the Master of Science (M.Sc.) in Statistics.

This project was created in the AUEB with the assistance of Fantasy Sports Interactive (FSI), a recognized and creative supplier of fantasy sports software for the gaming and sports betting industries.

In this project, we'll demonstrate a statistical learning strategy for estimating how many goals each soccer player will score in upcoming games. The introduction's remaining sections are organized as follows: We'll begin by discussing sports analytics in general and outlining some of the most significant developments that have occurred. The same will be done with soccer analytics in particular, and we will mention the many types of market data. The rationale for soccer betting and the most well-known betting markets for this sport will then be discussed, and the chapter's reasons and objectives will be given at the end.

## 1.1 Contextualization

The collaboration of the AUEB with the FSI has been key to the development of this project, especially the support provided by the data scientists of the FSI, Vasileios Palaskas and Spiros Kolovos, as well as the professor of the AUEB, Ioannis Ntzoufras.



## 1.1.1 Sports Analytics

Alamar and Mehrotra (2011) state that a common definition of sports analytics is the process of managing data, putting predictive models into practice, and employing information systems for decision-making in order to gain a competitive advantage on the field. Sports analytics have been used in a variety of contexts. For instance, sports teams analyze players using statistical analysis to decide on the best game plan. Sports organizations rate players and teams, assess the effectiveness of the current set of regulations, and research the viability of enacting new ones. To comprehend the physical and psychological status of players, sports medicine specialists employ statistical techniques.

Baseball was the first sport to have mainstream use of data, as explained by Michael Lewis (2004) in his book *Moneyball: The Art of Winning an Unfair Game*, and its subsequent popular film.

In the book, Lewis provides examples of how Billy Beane, general manager of the Oakland Athletics of Major League Baseball (MLB), put these ideas into practice. Teams scoring runs, in Beane's opinion, was simply the result of some study. In a nutshell, his hypothesis was that teams with higher on-base percentages were more likely to score runs and, as a result, were more likely to win games. Only those players that meet this system were selected by Beane in the draft and through trades. The Athletics became a squad that drew considerably more walks than strikeouts as a direct result. Players did not need to conform to the height, weight, speed, or body composition templates that guided other clubs' decisions under Beane's sabermetrics methodology.

Basketball is another sport where data analytics have an effect. Though it began a few years later than in baseball, the development has been substantial. The well-known 3-Point Revolution is one of sports analytics' most significant contributions to basketball.



Data helped them realize that shooting from mid-range wasn't worth it, given that shooting from a little farther away could get them more points. Shots from midrange appear even less attractive when compared to shots taken merely a few steps farther back; in fact, shots from a distance of 23 feet were valued more than those made from a distance of 2 feet. This realization led to many teams changing their attacking strategies and changed the way the game was being played from that moment onward.

## 1.1.2 Soccer Analytics

As we've already established, sports like baseball and basketball, as well as the realm of soccer, heavily rely on statistics and analytics. Every day, the technical staff, the recruitment department, and other internal organizations at soccer clubs send out a ton of important requests. For example, a manager would be curious to know how many goals a striker should have scored given his opportunities to score, how many saves a goalkeeper should have made given the number of shots on goal he faced, or what methods and tactics the team's upcoming opponent will use.

More teams are making analytics investments in order to provide better answers to these queries. These analytics are made feasible by the vast amount of data that has become more accessible over the past ten years. Elite soccer teams often have positional information about the game and event information about the game. The development of big data technology is influencing the future of performance analysis in elite soccer based on these presumptions.

The two main categories of technical data that are currently gathered during games are matchsheet data and ball event data. Ball event data defines the activities that players take with the ball, such as passes, dribbles, interceptions, tackles, and shoots. Matchsheet data offers extensive information on games, including starting lineups, substitutions, goals, and cards.



The most popular data source for developing football analytics tools is the type of ball event data for the following reasons. First, ball event data is more manageable for processing and analysis compared to tracking data because it has a smaller volume and a simpler structure. Additionally, ball event data proves to be highly valuable for player recruitment, as it offers extensive coverage of players in smaller competitions and crucial youth tournaments. Furthermore, it's accessible for purchase from specialized companies, whereas tracking data is typically restricted to league teams. Also, ball event data is growing in terms of information richness. In addition to detailing ball-related actions, data collection companies have recently begun recording the positions of relevant players during significant events like shots.

To offer a snapshot of the activity at any point throughout a match, *Opta* captures in-depth event data, which is subsequently stored in a database. Experts choose the most intriguing statistics from the database that would fascinate followers everywhere. After that, they mix historical data with live data to produce the metrics that Sky Sports and the BBC utilize.

### 1.1.3 Sports Betting

The playing field has been greatly impacted by sports analytics, but they have also helped fuel the expansion of the sports gambling market. Customers and bookmakers are both particularly interested in sports data. Customers want statistics, particularly sophisticated ones, to help them decide what to bet on, while businesses seek to delight the customers by providing favorable betting odds while still generating a profit.

Soccer attracts the most betting interest among sports. Soccer betting accounts for 70% of the worldwide betting industry, according to *Sportradar*. This places soccer as the top choice for wagers among bettors. While it's challenging to precisely gauge the worldwide soccer betting market due to various illicit platforms, the United Kingdom alone reports an annual soccer betting turnover of more than £1 billion.



The FIFA World Cup, which occurs once every four years, is the soccer event with the highest number of wagers. The FIFA World Cup is the most viewed live event overall on television, averaging 3.2 billion viewers between 2010 and 2024.

210 nations first take part in the competition's qualifying rounds, competing on the field to advance to the final 32. Every step of the route is covered by bets from across the world. Each World Cup competition is thought to attract about \$260 million in wagers. Besides the FIFA World Cup, various local and international leagues provide additional opportunities for gamblers.

American football, as the name suggests, enjoys immense popularity within the United States and abroad. It ranks as the second most bet-on sport, both through legitimate and illegal bookmakers. The Super Bowl, one of the most important NFL competitions, brought in a staggering \$4.76 billion in wagers, both legal and illicit. It was the same story for the 2020 Super Bowl, which attracted an estimated 26 million bettors from around the US and beyond.

The global scale of the soccer betting industry remains challenging to determine accurately, owing to the presence of several illicit platforms. Nonetheless, in the UK alone, soccer attracts over £1 billion in annual wagers. This tournament involves 68 teams, and the ultimate victor emerges after a seven-round contest spanning over a month.

Tennis is another popular sport for wagering, which is not surprising given how many tennis tournaments are hosted every year. The International Tennis Federation hosts 93,000 matches and over 1500 events each year. There is constantly a wager to be made because of the outstanding number of game days. The most well-known tennis competition, Wimbledon, features elite athletes competing for the highly sought after crown.



The event is watched by more than 1 billion people worldwide, and with that enormous audience comes amazing estimates of bets made. The popularity of the sport has also increased as a result of live betting.

Cricket is another popular sport that draws significant bets. During the match between India and Bangladesh in the 2019 Cricket World Cup, wagers on England alone totaled over \$22 million. This demonstrates even more how popular this sport is worldwide, not just in Asia.

### **1.1.4 Soccer Betting**

Every week, there are hundreds of soccer games, and there are even more betting markets accessible thanks to bookies that provide a wide range of odds on every aspect of events. They will provide their rates based on game results and game-related events, such as the number of goals scored, corners won, red and yellow cards displayed, and specific goal scorers. There are plenty of soccer betting markets in the betting space.

The main soccer betting markets are summarized in four categories:

- match results-oriented markets (home win/draw/away win).
- team performance-oriented markets (goals scored, cards, penalties and offsides by each team).
- athlete-level markets (total number of goals scored, cards, penalties and offsides by each athlete).
- long-term markets (which team will be the champion, which player will be the first scorer at the end of the season etc.).



## 1.2 Motivation

The development of an algorithm that predicts how many goals an athlete will score in a game is key to athlete-level markets and specifically to goal-based markets such as Over 0.5 Goals (an athlete with 1 or more goals scored in a match).

Past works on predicting the goal scoring performance of a soccer athlete were focused on the development of models using mainly team-based information, a home advantage index, and some basic statistics about the athlete (his position, the minutes he played and the number of shots that he has). These models separated the creation of shots and their conversion into goals from the overall process of scoring goals. This approach contains two sources of uncertainty, since the correct prediction of goals depends on the correct prediction of the shots that a player will make.

In contrast to this approach, this thesis focuses on the direct predicting the goals scored by soccer athletes using as information not only the above statistics, but also the most effective athlete-level ones which have been selected from a huge database provided by *Opta*. As for team-level information, our approach is based on Elo rating in contrast to most studies that use the categorical variables “team name” and “opponent name”. In this way, we not only keep the information about the quality of the athlete’s team as well as his opponent, but we manage to reduce the dimensions of our algorithm. The use of more detailed athlete-level statistics in combination with the Elo rating as well as with other useful statistics can give more precise insights into the data which will lead us to more accurate predictions.



## 1.3 Objectives

The main objectives of this thesis are the following ones:

- To create pre-match athlete-level statistics in order to use them as predictor variables in our model.
- To implement the Elo Rating System to measure the relative strength of an athlete's team and the strength of his opponent.
- To find significant associations between goals scored and the pre-match information.
- To apply variable screening using LASSO to choose the most significant pre-match information affecting the goal-scoring performance of soccer athletes.
- Finally, to implement a probabilistic comparison of our model-based predictions with the ones from *checkthechance.com*.

## 1.4 Master Thesis Outline

The remainder of the paper will be organized as follows:

- **Chapter 2: Related Work** dives deep into the main studies and developments made in the field of soccer.
- **Chapter 3: Dataset** describes the process followed starting with the data acquisition and finishing with the descriptive analysis.
- **Chapter 4: Models** gives the mathematical theory behind the models, discusses how the best model is selected and compares our predictions with those of the bookies.
- **Chapter 5: Conclusion and Future Work** checks whether the initial goals have been achieved. It also summarizes the findings uncovered during the realization of the project.



## 2 Related Work

In this section, we will go through the previous work done on goals predictions methods in soccer. The research work implemented until now, can provide us valuable insights about the approaches that were explored in the field of athletes' goal scoring ability in soccer and how they performed. We have divided this past research into three categories.

Firstly, we are going to refer to the general landscape of statistical learning techniques related to goal predictions in soccer.

In order to enhance forecasts of the field goals scored by two teams in soccer matches, we will secondly focus on research on how to combine various team ratings as team quality measurements.

Finally, we'll discuss studies that give models for evaluating soccer athletes' individual goal-scoring propensities, which is the primary focus of this thesis.

### 2.1 Goal Prediction Landscape

Since the middle of the 20th century, modeling goals for soccer teams has been a significant academic area. Moroney (1956) and Reep (1971) used both the Poisson distribution and the Negative Binomial distribution to estimate the number of goals scored in a soccer match based on previous team outcomes, and these two authors are credited with developing the first statistical modeling methodologies and insights.

However, it wasn't until Hill approved in 1974 that soccer match outcomes could be analyzed and forecasted using historical data rather than being purely dependent on chance.



Maher (1982) introduced the original invention when he used Poisson distributions to simulate each team's average number of goals scored at home and away as well as their offensive and defensive strengths. He utilized a bivariate Poisson model to account for correlation between the scores and used independent Poisson distributions for the goals scored by each team.

Following that, Dixon, and Coles (1997) were the first to create a model that could provide probabilities for match outcomes and scores while adhering to a Poisson distribution once more. The predicted number of goals for each team is converted into goal probability using the Poisson distribution in the Dixon and Coles model since it is based on a Poisson regression model.

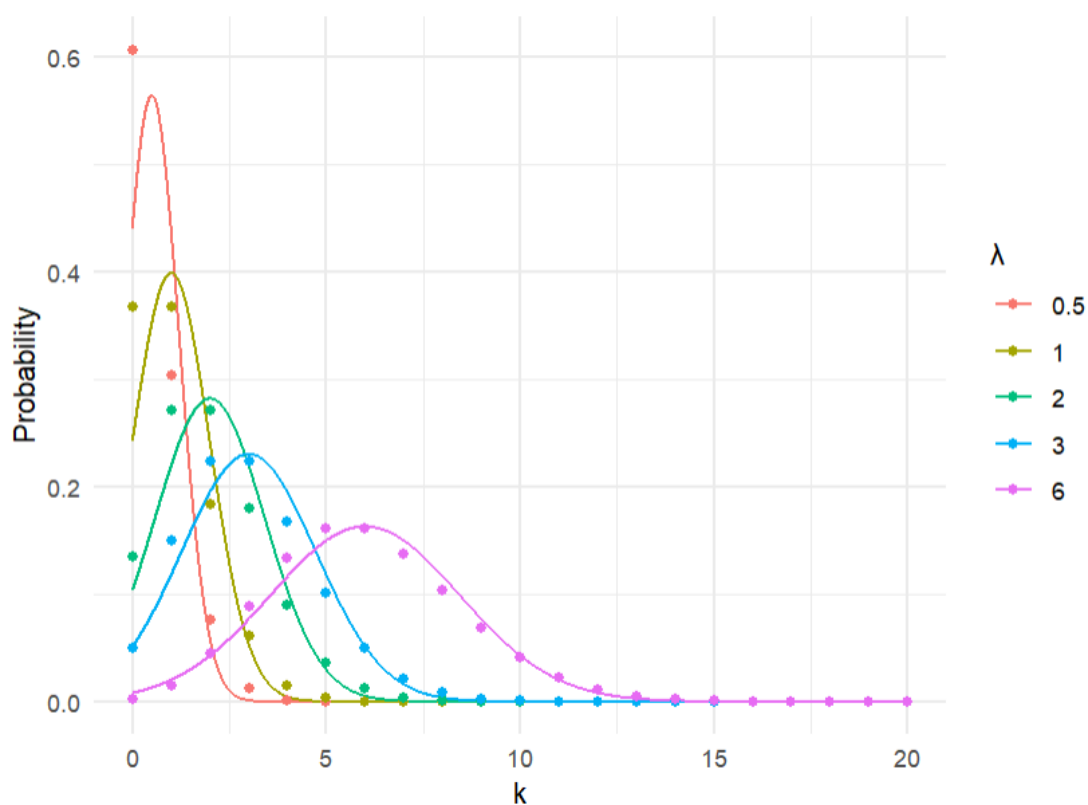


Figure 2.1: Poisson distribution for values of  $\lambda$  and  $k$

$$P(k \text{ goals in a match}) = \frac{e^{-\lambda} \lambda^k}{k!}$$

where  $\lambda$  is the expected number of goals in a game.

By generating the Maximum Likelihood estimates of these ratings based on the outcomes of previous matches, the Dixon and Coles model determines an attacking and defensive rating for each team. In order to develop forecasts, Rue and Salvesen (2002) decided to change the offensive and defensive parameters over time as additional data are obtained. Growder et al. (2002) continued their work to create a more systematic updating algorithm in the same year.

At the beginning of the twenty-first century, academics started modeling match results (win/draw/loss) directly rather than forecasting match scores and using them to give match result probabilities. Forrest and Simmons (2000), for instance, utilized a classifier model to predict the match outcome rather than the goals scored by each player. They were able to circumvent the difficulty of the scores of the two teams being interdependent as a result.

During the same year, Kuypers used variables pulled from a season's match results to generate a model capable of predicting future match results. His research work has been pioneering in the field of creating profitable model-based betting strategies.

By inflating the draw probability, Karlis and Ntzoufras in 2003 expanded the bivariate Poisson model and thoroughly investigated several versions of this model. More recently, Baio and Blangiardo (2010) proposed that a hierarchical model that fits in the Bayesian framework implicitly takes into consideration the correlation between the scores of the two opposing teams. In order to simulate game outcomes in a dynamic framework, Owen (2011) and Koopman and Lit (2014) permitted team parameters to change over time.



## 2.2 Elo Ratings

Elo created the most well-liked rating system in 1978 to evaluate chess athletes. For further applications, such as Buchdahl's 2003 study on utilizing Elo ratings for football to update clubs' relative strength, the name "Elo rating" was retained. Other sports forecasts, including tennis, have also made use of these ratings. In fact, Clarke and Dyte (2000) employed a logistic model with the difference in rankings to estimate match probability, whereas Boulier and Stekler (1999) used computer-generated rankings to enhance forecasts for tennis matches.

In order to forecast the results of football matches, Hvattuma developed an Elo system to create variables used in regression models in 2010. He experimented with two distinct sorts of Elo ratings, one that took into account the outcome of the match (win/loss/draw), and another that solely considered the actual score. He was able to see that Elo ratings produced better outcomes than the other benchmarks he tested his model against by employing a variety of testing benchmarks to assess his predictions. In this project, relative team strength ratings are crucial for encoding historical data and updating each team's relative strength as fresh results are introduced to the model.

Numerous studies have assessed the effect of particular Elo rating formula factors. In order to predict the outcome of a match, Sullivan and Cronin (2016) examined the Elo rating system and applied it to the English Premier League. Incorporating home-field advantage, adapting variable K-factor at different levels in a season, rewarding and penalizing winning and losing streaks, and incorporating game scores (rewarding a win proportionally to the margin of victory) are the four methods they investigate for changing the fundamental Elo scheme. Prediction accuracy is anticipated to rise over the fundamental Elo system with these extra characteristics.



## 2.3 Athlete's Goal Scored Prediction

The majority of the academic research on soccer statistics has focused on the modeling of goals at the team level due to data restrictions. However, because the main focus of this thesis is on modeling player-by-player objectives, it is very helpful to discuss any research that has been done in this particular area.

Using two mixed effects models (Poisson and Logistic Regression, respectively), McHale and Szczepanski (2014) divided the scoring process of individual soccer athletes into the creation of shots and the conversion of shots to goals. They used two different versions of each model: a basic model with only team-specific data and a home field indicator as covariates, which they believed to be the bare minimum to accurately reflect the process, and an expanded model with variables that took player positions and time on the field into account (for the first model) and the number of shots a player has (for the second model). They provided predictions of the potential number of goals that may be expected of a certain player in a game by combining the two components of this process (shot creation and shot conversion to goals).

For this thesis, we will take inspiration from this method as well as those of the above section (especially regarding the distributions used for goals and predictor variables) but we will improve the soccer athlete's expected goals model by adding more detailed team-level and athlete-level features.



## 3 Data

This chapter thoroughly describes the dataset and the data pre-processing techniques before building an algorithm able to determine how many goals a soccer player will score in a game. It is structured as follows:

- **Dataset** introduces the original data, giving details about its format, structure and where it has come from. Then explains how it will be transformed to perform data exploration.
- **Feature Engineering** creates the pre-match skill actions of the soccer athletes and implements the Elo Rating System.
- **Exploratory Data Analysis** investigates the data, summarizes its main characteristics, and provides a better understanding of the variables and the relationships between them.

### 3.1 Dataset

We obtained *Opta's* data for the 2020-21, 2021-22 and 2022-23 (first 16 Gameweeks) seasons of the English Premier League. The Gameweeks of the current season (2022-23) will be used for model evaluation as we will discuss in the next chapter.

#### 3.3.1 Data Origin

The available information in the data is divided into two categories:

- Athlete-level statistics such as goals, assists, passes etc.
- Team-level statistics such as Side (Home/Away), Score, etc.



The dataset provided by *Opta* was originally in XML files. XML file is a text file structured to hold data. It uses things called tags to explain what bits of data it holds and organizes them in a structure that makes it easy to understand and use everywhere.

```
<MatchData detail_id="1" last_modified="2014-08-17T17:00:47+00:00" timestamp_accuracy_id="1" timing_id="1" uID="g755303">
  <MatchInfo MatchDay="1" MatchType="Regular" MatchWinner="t3" Period="FullTime" Venue_id="3250">
    <Date>2014-08-16 17:30:00</Date>
    <TZ>BST</TZ>
  </MatchInfo>
  <Stat Type="Venue">Emirates Stadium</Stat>
  <Stat Type="City">London</Stat>
  <TeamData HalfScore="1" Score="2" Side="Home" TeamRef="t3">
    <Goal Period="FirstHalf" PlayerRef="p51507" Type="Goal" />
    <Goal Period="SecondHalf" PlayerRef="p41792" Type="Goal" />
  </TeamData>
  <TeamData HalfScore="1" Score="1" Side="Away" TeamRef="t31">
    <Goal Period="FirstHalf" PlayerRef="p15284" Type="Goal" />
  </TeamData>
</MatchData>
<MatchData detail_id="1" last_modified="2014-08-17T16:32:51+00:00" timestamp_accuracy_id="1" timing_id="1" uID="g755305">
  <MatchInfo MatchDay="1" MatchType="Regular" Period="FullTime" Venue_id="2516">
    <Date>2014-08-16 15:00:00</Date>
    <TZ>BST</TZ>
  </MatchInfo>
  <Stat Type="Venue">King Power Stadium</Stat>
  <Stat Type="City">Leicester</Stat>
  <TeamData HalfScore="1" Score="2" Side="Home" TeamRef="t13">
    <Goal Period="FirstHalf" PlayerRef="p54316" Type="Goal" />
    <Goal Period="SecondHalf" PlayerRef="p60689" Type="Goal" />
  </TeamData>
  <TeamData HalfScore="2" Score="2" Side="Away" TeamRef="t11">
    <Goal Period="FirstHalf" PlayerRef="p18981" Type="Goal" />
    <Goal Period="FirstHalf" PlayerRef="p18267" Type="Goal" />
  </TeamData>
</MatchData>
<MatchData detail_id="1" last_modified="2014-08-18T10:47:54+00:00" timestamp_accuracy_id="1" timing_id="1" uID="g755307">
  <MatchInfo MatchDay="1" MatchType="Regular" MatchWinner="t80" Period="FullTime" Venue_id="28">
    <Date>2014-08-16 12:45:00</Date>
    <TZ>BST</TZ>
  </MatchInfo>
  <Stat Type="Venue">Old Trafford</Stat>
  <Stat Type="City">Manchester</Stat>
  <TeamData HalfScore="0" Score="1" Side="Home" TeamRef="t1">
    <Goal Period="SecondHalf" PlayerRef="p13017" Type="Goal" />
  </TeamData>
  <TeamData HalfScore="1" Score="2" Side="Away" TeamRef="t80">
    <Goal Period="FirstHalf" PlayerRef="p76542" Type="Goal" />
    <Goal Period="SecondHalf" PlayerRef="p55422" Type="Goal" />
  </TeamData>
</MatchData>
<MatchData detail_id="1" last_modified="2014-08-17T16:44:08+00:00" timestamp_accuracy_id="1" timing_id="1" uID="g755309">
  <MatchInfo MatchDay="1" MatchType="Regular" MatchWinner="t88" Period="FullTime" Venue_id="68">
    <Date>2014-08-16 15:00:00</Date>
    <TZ>BST</TZ>
  </MatchInfo>
```

Figure 3.1: Parsing (Opta) XML File Example. Source: The Information Lab [Brian Prestidge, 2015]

### 3.1.2 Data Pre-Processing

The above data structure is not efficient for exploratory analysis or predictive modelling. We therefore need to clean the dataset in the sense of converting *Opta's* raw data to tidy data. Table 3.1 displays a sample of the cleaned data in a few selected columns.

	<b>PlayerName</b>	<b>matchday</b>	<b>goals</b>
1	Erling Haaland	1	2
2	Erling Haaland	2	0
3	Erling Haaland	3	1
4	Erling Haaland	4	3
5	Erling Haaland	5	3
6	Erling Haaland	6	1
7	Erling Haaland	8	1
8	Erling Haaland	9	3
9	Erling Haaland	10	1
10	Erling Haaland	11	0
11	Erling Haaland	13	2
12	Erling Haaland	16	0

Table 3.1: Tidy (transformed) data.

Each row in the new dataset is an observation that contains information about an athlete in the respective match. On the other hand, the columns refer to the variables for which we have this information.

The total number of variables we have at our disposal is 262 of which 247 are related to the skill actions of athletes (goals, assists, passes, etc.). The remaining 15 ones, which are presented directly below, are related to different information about the athlete and the match he played:

1. season (20/21, 21/22, 22/23)
2. sportevent start time (the start time of the match)
3. uId (the unique id of the match)
4. matchday (the matchday of the league)
5. Status (if a player was started or sub)
6. Score (the final score of the athlete's team)
7. Side (home/away)
8. TeamRef (the unique id of the athlete's team)



9. TeamName (the name of the athlete's team)
10. PlayerRef (the unique id of the athlete)
11. PlayerName (the name of the athlete)
12. Real.Position (the position of the athlete on the pitch)
13. Score\_opponent (the final score of the athlete's opponent team)
14. Opponent\_TeamName (the name of the athlete's opponent team)
15. Opponent\_TeamRef (the unique id of the athlete's opponent team)

Having transformed the data into a useful form, it remains to remove the observations that we do not need. Out of a total of 25,210 observations, 7,090 relate to athletes who are goalkeepers or substitutions. Goalkeepers have an infinitesimal chance of scoring and are therefore superfluous in our sample. Also, we will focus on the athletes who were started because firstly they have comparable data and secondly we want to predict how many goals an athlete will score provided that he will be started. Consequently, we removed the data relating to the above two cases, maintaining the remaining 18,120 observations which correspond to 670 athletes.

## 3.2 Feature Engineering

As we mentioned earlier, there are 247 variables that refer to various actions that each athlete has taken in the match. Since we want to predict the goal scoring ability of an athlete in a match based on information we have for him before this match, we need to transform the above variables into pre-match information. In addition to transforming variables, it is necessary to create new ones that will help us determine more accurately the scoring ability of an athlete such as the skill level of his team as well as that of his opponent. These procedures are described in detail in the corresponding sections below.



### 3.2.1 Pre-Match Skill Actions

The pre-match information so far that we have at our disposal, and we can use to determine the scoring ability of an athlete are the following:

- Side (home/away)
- TeamName (the name of the athlete's team)
- Opponent\_TeamName (the name of the athlete's opponent team)
- Real.Position (the position of the athlete on the pitch)

In order to expand the information about an athlete before the match we want to make the prediction, we will transform all 247 *Opta's* skill actions in the following way:

1. Set the average value of an athlete in each skill action of his past matches into the first observation of that athlete in that transformed skill action.
2. Set the average value of an athlete in each skill action in his last 5 games into the 6<sup>th</sup>, 7<sup>th</sup>, 8<sup>th</sup>, ....,etc. observation of that athlete in that transformed skill action.



3. For the 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup>, and 5<sup>th</sup> observation of an athlete in the transformed skill action set out the following:

- 2<sup>nd</sup> transformed observation = 1<sup>st</sup> original observation /5 + 1<sup>st</sup> transformed observation ×4/5
- 3<sup>rd</sup> transformed observation = (1<sup>st</sup> + 2<sup>nd</sup> original observations) /5 + 1<sup>st</sup> transformed observation ×3/5
- 4<sup>th</sup> transformed observation = (1<sup>st</sup> + 2<sup>nd</sup> + 3<sup>rd</sup> original observations) /5 + 1<sup>st</sup> transformed observation ×2/5
- 5<sup>th</sup> transformed observation = (1<sup>st</sup> + 2<sup>nd</sup> + 3<sup>rd</sup> + 4<sup>th</sup> original observations) /5 + 1<sup>st</sup> transformed observation ×1/5

The optimal lag length, which is 5, was selected using the Akaike Information Criterion (AIC). Specifically, we used as a fixed effect in a generalized linear mixed effects model (detailed mention of these models in the next Chapter) the goals scored by an athlete before the match for which we want to predict how many goals he will score and recorded the model's AIC values for lag 2 to 7 (Figure 3.2).



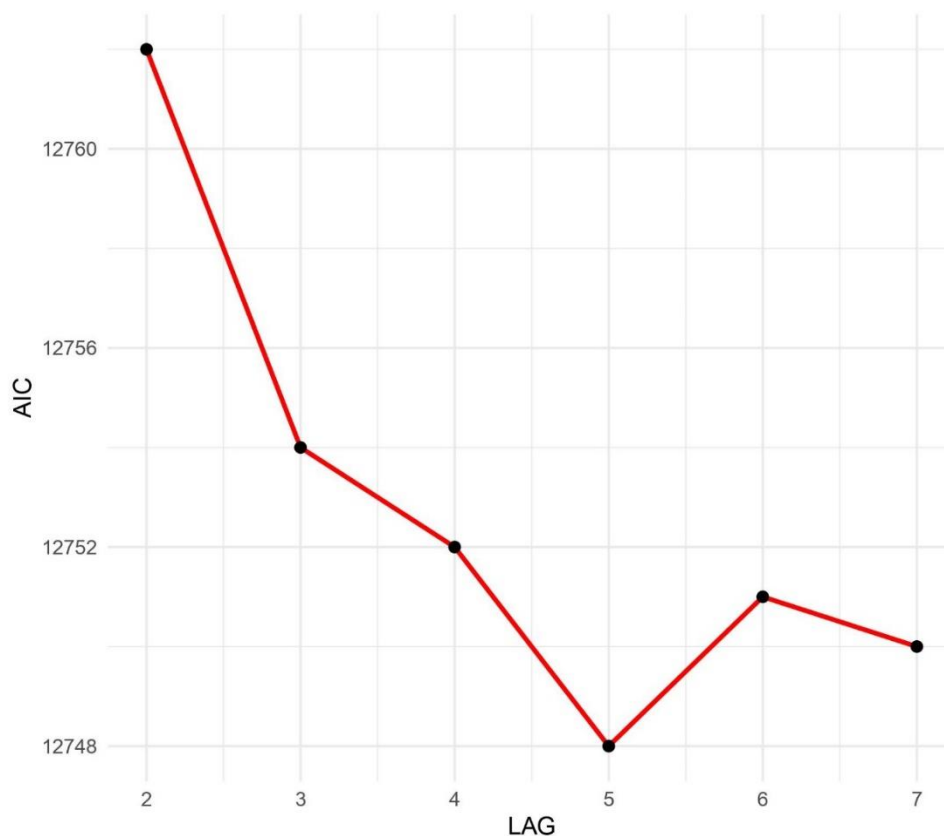


Figure 3.2: Optimal Lag selection.

The lag value that minimizes the model's AIC is 5. Therefore, we use this lag not only in goals but also in all other skill actions of athletes to create pre-match information in the way mentioned above.

In a similar study for fantasy points instead of goals, Bonomo et al. (2014), calculated a tracking average by averaging the points obtained in the 3 previous Gameweek.

In table 3.2, we present as an example of the application of the above method, the goals of Erling Haaland until Gameweek 16 of the 2022-23 season and the transformed pre-match variable (lag 5 goals) which is the lag 5 tracking average of goals.

	PlayerName	matchday	goals	lag 5 goals
1	Erling Haaland	1	2	1.417
2	Erling Haaland	2	0	1.533
3	Erling Haaland	3	1	1.250
4	Erling Haaland	4	3	1.167
5	Erling Haaland	5	3	1.483
6	Erling Haaland	6	1	1.800
7	Erling Haaland	8	1	1.600
8	Erling Haaland	9	3	1.800
9	Erling Haaland	10	1	2.200
10	Erling Haaland	11	0	1.800
11	Erling Haaland	13	2	1.200
12	Erling Haaland	16	0	1.400

Table 3.2: Pre-match skill actions creation: Erling Haaland example.

- 1<sup>st</sup> lag 5 goals (1.417) = Average value of Erling Haaland's goals in his past matches
- 2<sup>nd</sup> lag 5 goals (1.533) = 1<sup>st</sup> goals\_lag\_5 × 4/5 + 1<sup>st</sup> goals/5
- 3<sup>rd</sup> lag 5 goals (1.250) = 1<sup>st</sup> goals\_lag\_5 × 3/5 + (1<sup>st</sup> + 2<sup>nd</sup> goals)/5
- 4<sup>th</sup> lag 5 goals (1.167) = 1<sup>st</sup> goals\_lag\_5 × 2/5 + (1<sup>st</sup> + ... + 3<sup>rd</sup> goals)/5
- 5<sup>th</sup> lag 5 goals (1.483) = 1<sup>st</sup> goals\_lag\_5 × 1/5 + (1<sup>st</sup> + ... + 4<sup>th</sup> goals)/5
- 6<sup>th</sup> lag 5 goals (1.800) = (1<sup>st</sup> + ... + 5<sup>th</sup> goals)/5
- 7<sup>th</sup> lag 5 goals (1.600) = (2<sup>nd</sup> + ... + 6<sup>th</sup> goals)/5
- .....
- 12<sup>th</sup> lag 5 goals (1.400) = (7<sup>th</sup> + ... + 11<sup>th</sup> goals)/5



## 3.2.2 Elo Rating System

The Elo Rating System is a rating system used to assess the relative competitiveness of sports teams and players, and it was briefly discussed in Section 2.2. Elo bases its predictions on the idea that each team's or athlete's performance is symmetrically distributed, with a mean reflecting the underlying strength of the team or the athlete. The Elo system also assumes that the mean value of a team's or athlete's performance (relative strength) very slowly increases over time, even if a team's or athlete's genuine performance may be enhanced with a lot of practice. Also, a weaker team or athlete will receive a higher rating from the Elo system for defeating a stronger opponent than a stronger team or athlete will receive.

### 3.2.2.1 Mathematical Model

When team (or athlete)  $i$  competes with team (or athlete)  $j$ , we have:

$$r_i^{new} = r_i^{old} + K (S_{ij} - E_{ij}) \quad \text{and} \quad r_j^{new} = r_j^{old} + K (S_{ji} - E_{ji}) \quad (1)$$

In the left part of equation (1):

$r_i^{new}$  : the updated (new) rating for  $i$

$r_i^{old}$  : the current (old) rating for  $i$

$K$ : a fixed factor

$S_{ij}$  : the match result (1: win, 0: loss, 0.5: draw) for the team (or athlete)  $i$

$E_{ij}$  : the expected winning probability for the team (or athlete)  $i$  against  $j$

In the right part of the equation (1), the parameters  $r_j^{new}$ ,  $r_j^{old}$ ,  $S_{ji}$ , the fixed parameter  $K$ , and the expected winning probability  $E_{ji}$  are based on  $j$  (against  $i$ ).



### 3.2.2.2 The K factor

$K$  controls the deviation between the existing and new ratings in equation (1). If  $K$  is set too high, the updating rating sensitivity will be strong, which means that if a team performs somewhat better or worse than anticipated, its rating can vary significantly. The Elo updating rating system will not be able to keep up with the teams' actual skill levels if  $K$  is set too low.

The  $K$ -factor frequently changes depending on how significant the competition is. In soccer matches, several  $K$  values are used to represent the match's significance. The importance and level of competitiveness in a match increase as  $K$  increases, and vice versa, when  $K$  decreases.

In the equation (1) we mentioned that  $K$  is a fixed constant, so that every game is weighted equally. However, some Elo models like *FiveThirtyEight's* have a moving  $K$  that is dependent on the margin of victory using a formula of  $k = C + C * \text{margin}$ , where  $C$  is a constant. This means that for every goal the margin increases by, the  $k$  value will increase by  $C$ , starting with a base value  $k$  of  $C$  for draws (margin equals 0). The idea is to help account for and reward / punish team / athletes winning / losing by bigger margins compared to closed matches.

### 3.2.2.3 The Expected Winning Probability E

In the above equation,  $E_{ij}$  represents the expected winning probability of team / athlete  $i$  when playing against team / athlete  $j$  and is usually assumed that is a logistic function of the rating difference of the two teams / athletes.



Due to its s-shaped curve, it is also known as the sigmoid function or sigmoid curve (Figure 3.3).

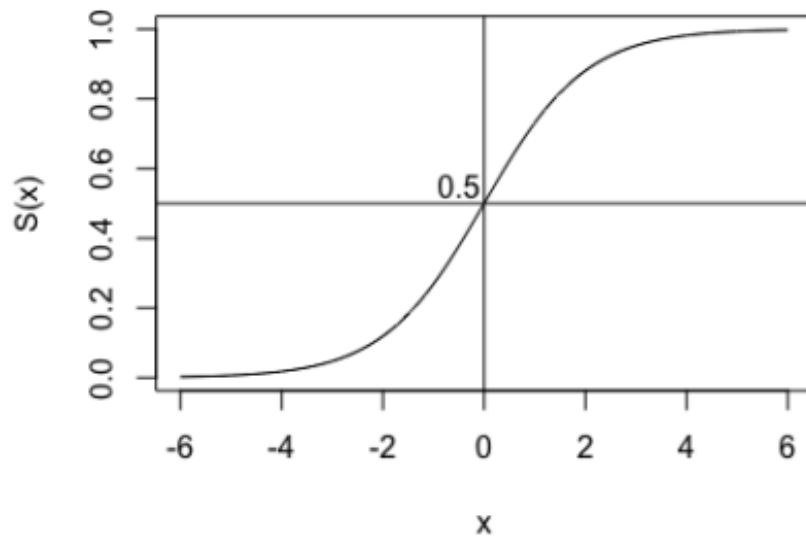


Figure 3.3: The sigmoid curve. Source: Wikipedia

When team  $i$  plays against team  $j$ , and they have rating  $r_i$  and  $r_j$  respectively, the expected winning probability  $E_{ij}$  (for  $i$  against  $j$ ) is:

$$E_{ij} = \frac{1}{1+10^{-(r_i-r_j)/400}} \quad (2)$$

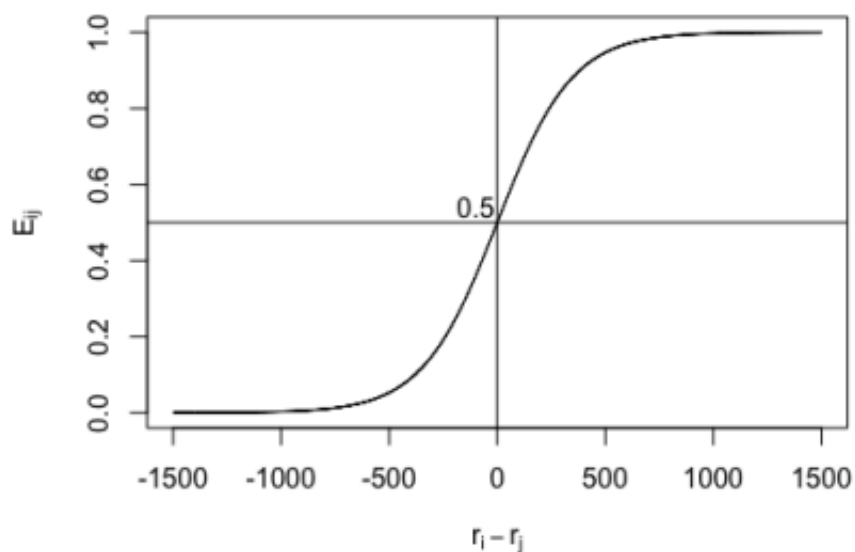


Figure 3.4: The expected winning probability  $E_{ij}$ .

### 3.2.2.4 The Home Advantage H

Home field advantage is occasionally taken into account by Elo algorithms when calculating wins and losses in games. This can result in a considerable improvement in the Elo algorithm's performance. Even though they may not be as good as the competition, home court advantage has historically been a significant factor in favor of the home winning side. This is a result of things like the crowd's ovations and the home team's familiarity with the setting, which gives them an edge throughout the game.

The predicted winning probability function incorporates home field advantage and is expressed as follows:

$$E_h = \frac{1}{1+10^{-(h+r_i-r_j)/400}}, \quad (3)$$

where  $h$  is the gain from home advantage. The chess world's empirical distribution is where the constants 10 and 400 in equations (2) and (3) originate from. They suggest a modification to the x-axis scale in Figure 3.4 when taken as a whole.

### 3.2.2.5 Tuning the Elo parameters H and C

To make the ratings as precise as possible, it is required to adjust the underlying parameters  $H$  and  $C$ , which represent the home advantage and the constant of the  $K$  factor, respectively. For instance, a  $K$  factor that is too little would result in ratings that update too slowly. More recent events won't be well-received by the ratings. Conversely, a high  $K$  factor will overestimate the importance of recent findings. The additional points that were added to the home team's ranking to reflect the home field advantage are equivalent. Poor tuning will result in inaccurate forecasts.



We must gauge the ratings' accuracy in order to tune the system. By comparing the actual outcomes to those projected by the rating gap between the two competing teams, the Elo system adjusts the ratings. This variation can be used to fine-tune the system's settings. We want to adjust the settings such that this disparity is as minimal as feasible since the predictions are more accurate the smaller this difference is.

We apply the following criterion to evaluate the model accuracy in order to formalize this:

$$\Sigma_i [(E_{hi} - S_{hi})^2 - (E_{ai} - S_{ai})^2] \quad (4)$$

where  $E_{hi}$  and  $E_{ai}$  are the expected winning probability of match  $i$  for the home team and the away team, respectively. These expected winning probabilities are a number between 0 and 1 and are calculated based on the ratings of the two teams (see equations 3 and 4).  $S_{hi}$  and  $S_{ai}$  are the actual results of match  $i$ , encoded as 0 for loss, 0.5 for draw and 1 for a win.

### 3.2.2.6 Results

Using the above criterion to all our data we can find the best constant  $C$  using different values between 1 and 30 and setting  $H = 0$ . Then plotting the Mean Squared Error (MSE) of the equation (4) against the constant  $C$  we see that 9 is the best  $C$  value (Figure 3.5).



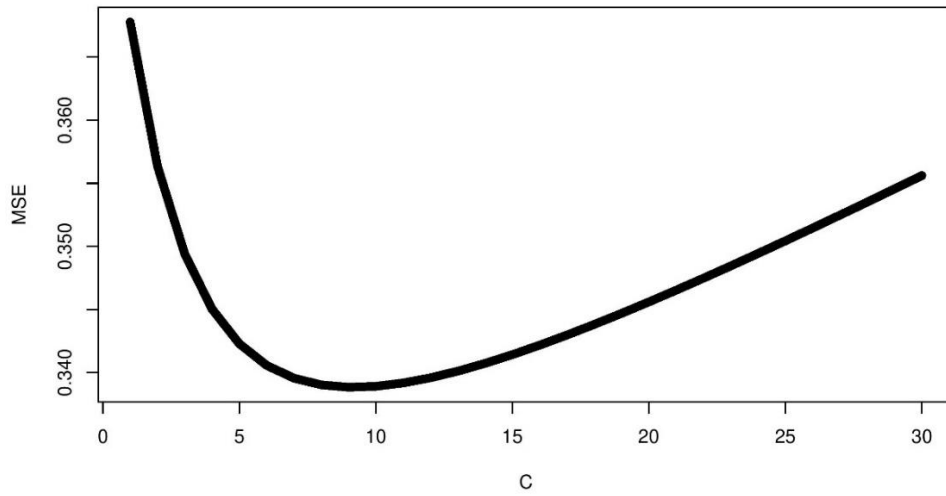


Figure 3.5: Optimal  $C$  selection using the MSE criterion.

We can apply the same strategy to also find the best adjustment for the home field advantage parameter  $H$ . To find the optimal home field advantage we applied the Elo ratings with  $C = 9$ , using different values for the  $H$ .

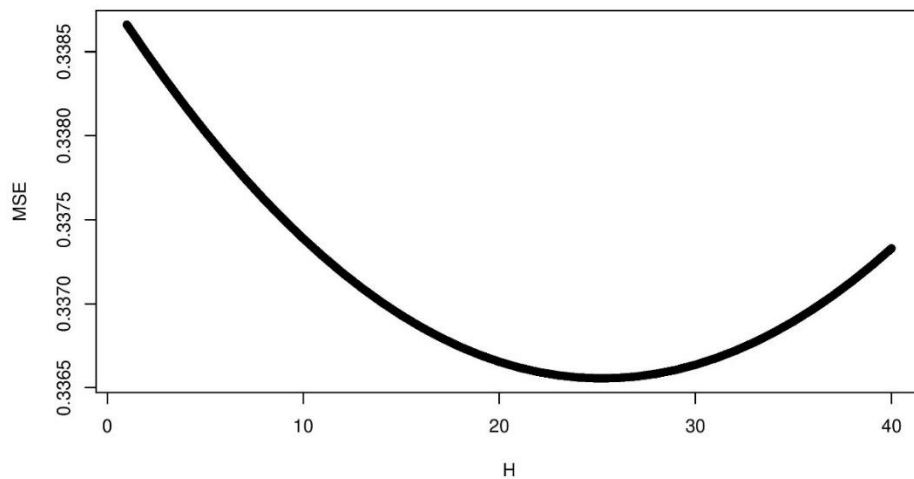


Figure 3.6: Optimal  $H$  selection using the MSE criterion.

From Figure 3.6 we notice that a bonus of 25 points is the best value to add to the elo rating of the home team. We can also find the best values of our parameters independent of each other in the contour below.

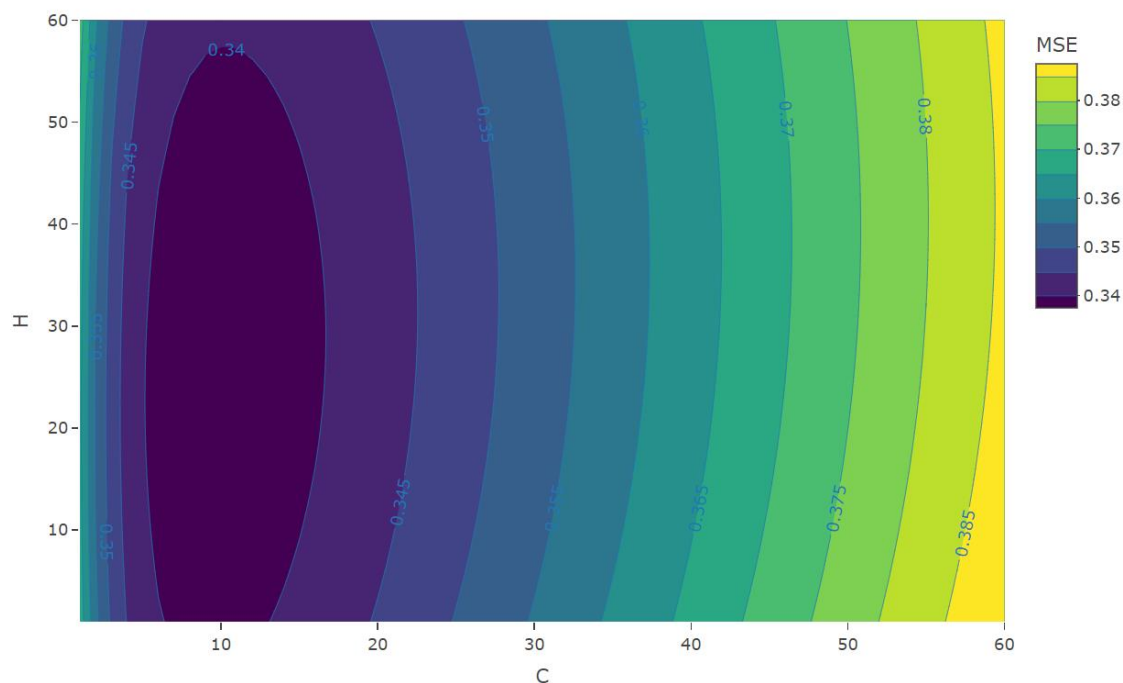


Figure 3.7: Optimal  $C$  and  $H$  selection simultaneously.

Plotting the MSE against both  $C$  and the home field advantage  $H$  we find that the optimal pair of these parameters is 9 and 25 respectively as we concluded in Figures 3.5 and 3.6.

Using these values for our parameters  $C$  and  $H$  respectively, we calculate the final Elo Ratings (until the Gameweek 16) of the last (2022-23) English Premier League teams starting from the 2020-21 season at 1500 elo for all teams.

	Teams 2022–23 Season	Elo Rating Until GW 16	Actual Points Until GW 16
1	Manchester City	1815.793	38
2	Liverpool	1754.908	26
3	Arsenal	1690.434	40
4	Tottenham Hotspur	1650.054	29
5	Chelsea	1610.888	22
6	Newcastle United	1595.932	33
7	Manchester United	1580.797	28
8	Brighton and Hove Albion	1547.298	27
9	Leicester City	1543.305	17
10	West Ham United	1505.217	14
11	Crystal Palace	1493.491	20
12	Brentford	1490.272	22
13	Aston Villa	1488.124	21
14	Bournemouth	1473.631	16
15	Nottingham Forest	1448.182	13
16	Leeds United	1441.233	19
17	Everton	1418.192	14
18	Fulham	1414.497	22
19	Wolverhampton Wanderers	1401.688	10
20	Southampton	1372.718	12

Table 3.3: Elo Rating and Actual Points of 2022-23 English Premier League Teams Until Gameweek 16.

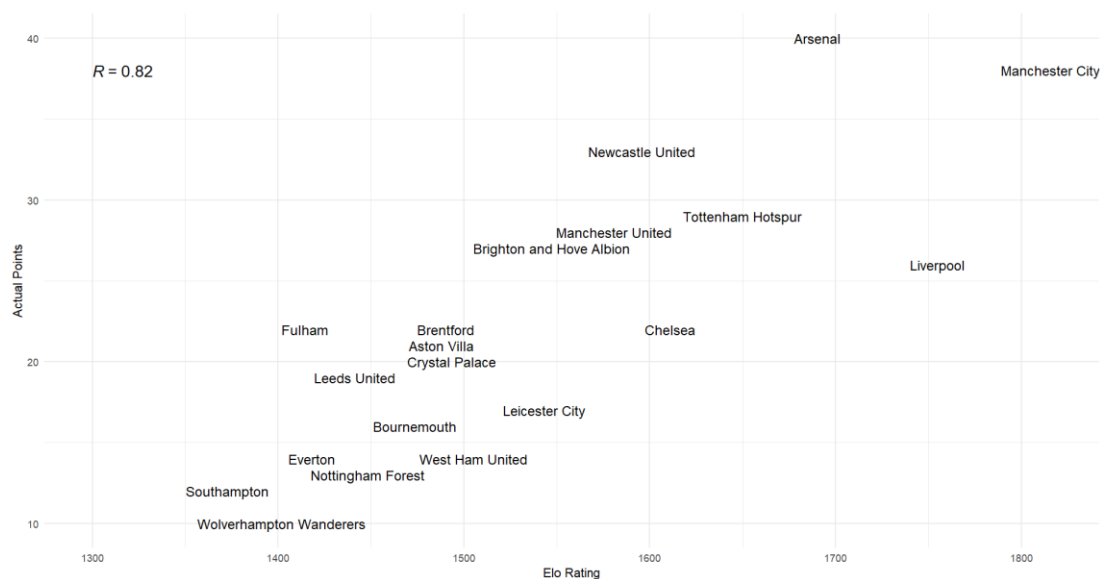


Figure 3.8: Scatter plot of Elo and Actual Points Until Gameweek 16.

On the x axis we see the final elo of the teams after the 16<sup>th</sup> Gameweek of the 2022-23 season (running the algorithm from the first match of the 2020-21 season as mentioned above), while on the y axis we have the actual points for the same period. The strong correlation that these two variables have confirms how well the elo algorithm reflects the dynamics of the teams.

Figure 3.9 shows the distribution and the variability of the Elo Rating grouped by the 2022-23 teams until Gameweek 16. Using boxplots for the distribution of elo rating of each team we got a better picture of the variability of the elo rating in each team but also how much the teams differ between them.

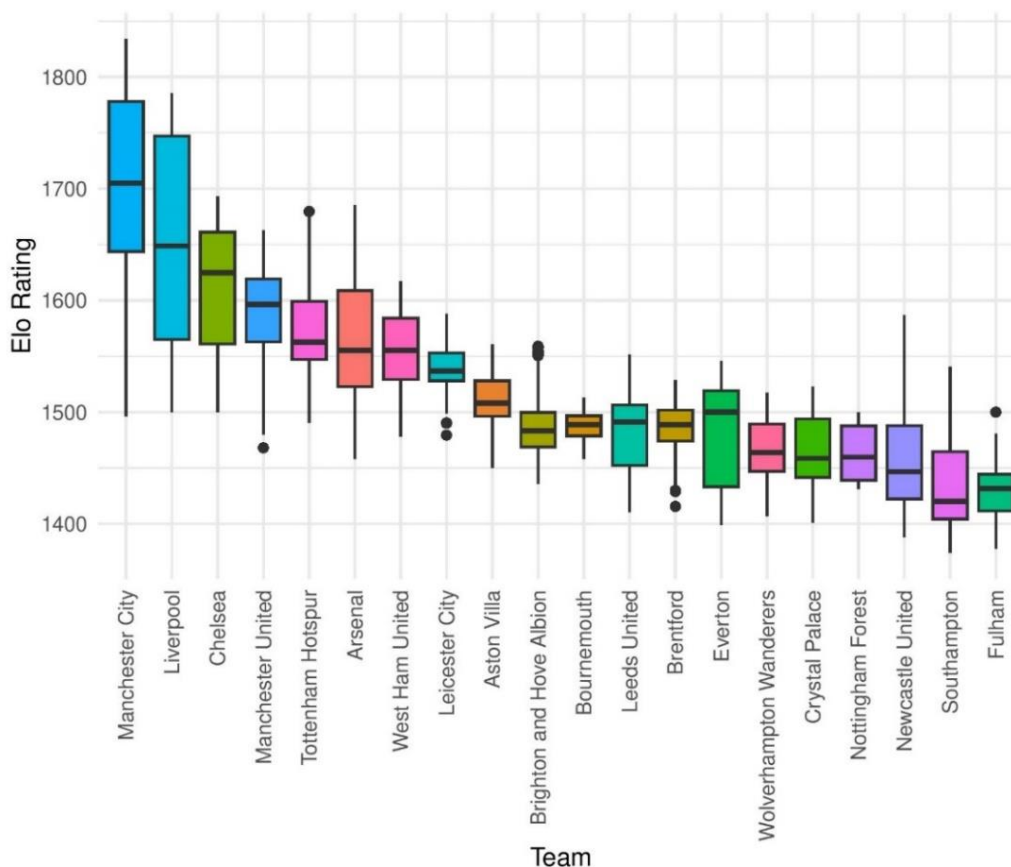


Figure 3.9: Box plot for the Elo Ratings of 2022-23 teams.

Manchester City, 2020-21 and 2021-22 Premier League Champions, is unsurprisingly the team with the highest median Elo points followed by Liverpool. The high variability observed in the distribution of Liverpool is mainly due to this year's (2022-23 season) performance of the team, which is quite different in relation to the previous 2 years (2020-21 and 2021-202) when they were in the first positions of the ranking. On the contrary, Arsenal and Newcastle United, which are this year's pleasant surprises, explain the strong positive skewness that exists around their distribution.

Lastly, Premier League Big 6 (Manchester City, Liverpool, Chelsea, Manchester United, Tottenham and Arsenal) stand out from the rest of the teams, while the teams ranked from the middle of the Elo Rating System and beyond do not seem to differ significantly between them.

### **3.3 Exploratory Data Analysis (EDA)**

Having transformed our data into the appropriate format and creating these variables that will help us predict athletes' scoring ability in a match (independent variables), we can start the exploratory data analysis (EDA).

In order to determine what distribution may be utilized to model our dependent variable, we will specifically study the data of the response (dependent) variable in this part, which is the goals of the athlete in a match. Then, we'll look at how variables relate to one another. To be more precise, we'll look at how home field advantage and position relate to the dependent variable (goals), as well as how the dependent variable relates to the athletes' pre-game skill actions. This will make it possible for us to identify relevant trends in the data and decide which factors are crucial and which have little bearing on the ability of the athletes to score.



### 3.2.1 The Dependent Variable: Goals

As we saw in the previous chapter, most of the research that dealt with the modeling of goals either at the team or individual level was based on the Poisson distribution. The Poisson distribution is a discrete probability distribution used to depict the likelihood of events happening within a given time interval (such as 90 minutes), based on a known average event occurrence rate. The idea that the quantity of occurrences is independent of time is crucial. This suggests that the probability of goals occurring remains the same regardless of how many have been scored so far in the game. Instead, only the average rate of goals ( $\lambda$ ) is used to indicate the number of goals. The figure below shows the relative frequency of goals scored by the (started) athletes of our data in each match.

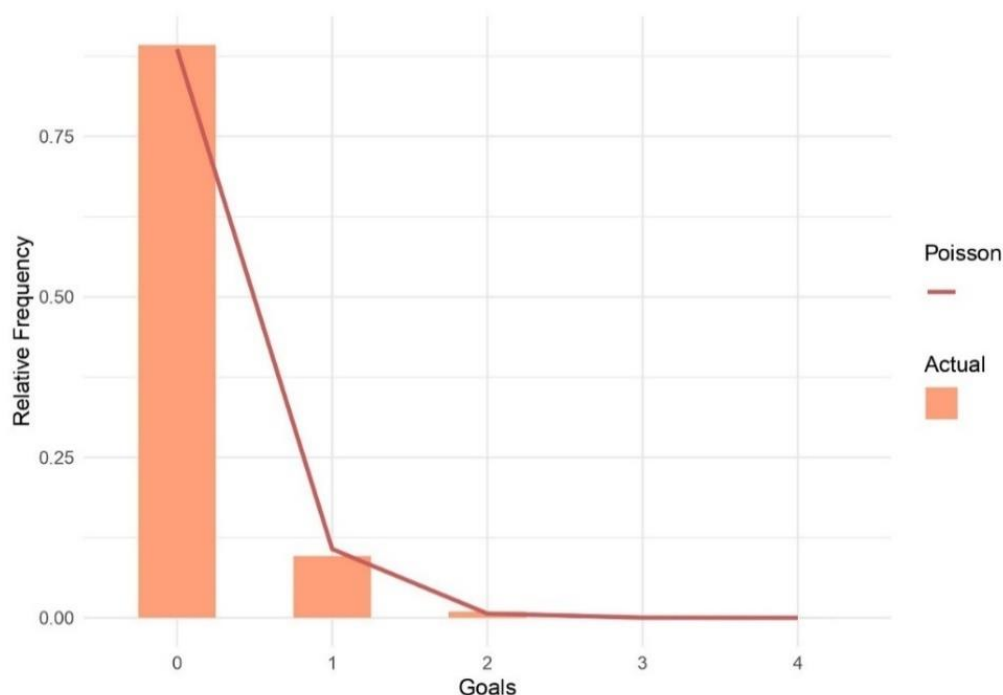


Figure 3.10: Number of Goals per (Started) Athlete per Match.

The average number of goals for a started athlete in a match is 0.12 ( $\lambda$ ) while the percentage of times an athlete will not score in a match reaches 90%. The table below shows the total number of goals scored by all athletes in the 3 seasons we have at our disposal.

Season	Total Goals	Average Goals of 1 Athlete in A Match
1 20/21 (380 matches)	895	0.118
2 21/22 (380 matches)	935	0.123
3 22/23 (146 matches)	359	0.123

Table 3.4: (Started) Athletes' goalscoring statistics grouped by season.

### 3.2.2 Relationships Between Variables

In this section, we will investigate the relationship between goals and the pre-match available information of soccer athletes, which is:

- The home advantage:
  - 0 if the athlete's team is the away team.
  - 1 if the athlete's team is the home team.
  
- The position on the pitch:
  - 1) Central Defender (CD)
  - 2) Full Back (FB)
  - 3) Wing Back (WB)
  - 4) Defensive Midfielder (DM)
  - 5) Central Midfielder (CM)
  - 6) Attacking Midfielder (AM)
  - 7) Winger (W)
  - 8) Second Striker (SS)
  - 9) Striker (S)



- The pre-match skill actions which we created in Sections 3.2.1 and 3.2.2 respectively.

The goal of this analysis is to establish whether there is a statistical connection between goals and the independent variables mentioned above, and if such a connection exists, to assess its strength and direction.

### 3.2.2.1 Goals and Home Advantage

The home team is seen to have a sizable edge over the visiting team in the majority of team sports. Here, we especially want to look at whether an athlete performs much better when his team is the home team than when their team is the away team. In a nutshell, if home field advantage affects an athlete's score output during a game. In Figure 3.11, the error bar for the home team does not cross over with the error bar for the visiting team. This suggests that there may be a big disparity between the mean goals of the athletes on the home teams and those on the away teams. By using the Two-Sample t-Test, this conclusion is supported ( $t(18067) = -2.795$ ,  $p = 0.005$ ).

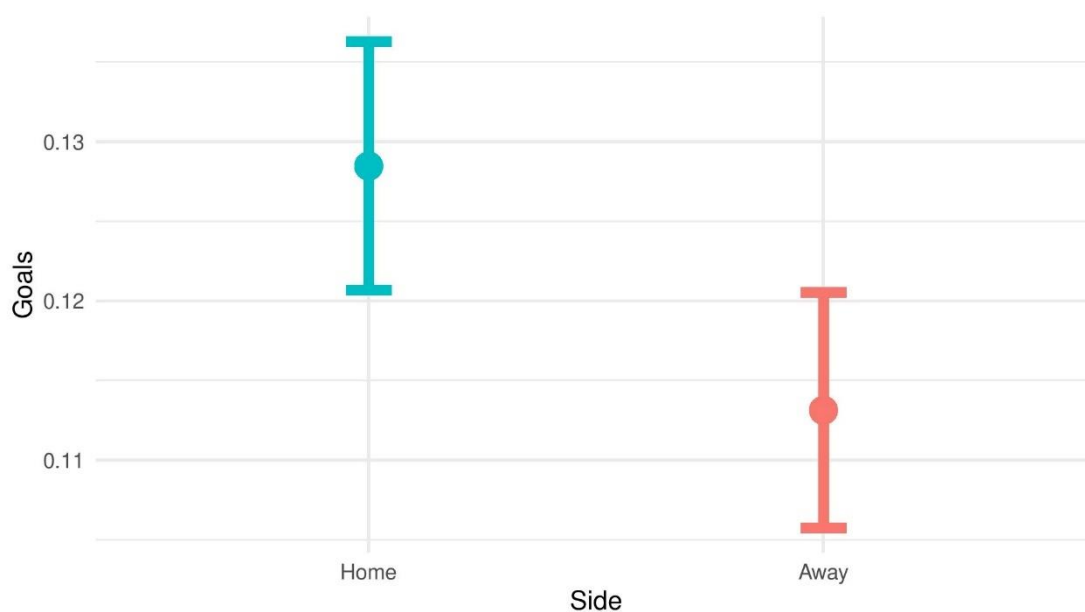


Figure 3.11: Error bars (mean goals +/- 95% CI) grouped by Side.

### 3.2.2.2 Goals and Position

Apart from the home advantage, a variable that makes sense to study in terms of its connection with an athlete's scoring performance is his position on the pitch. For example, offensive athletes naturally score more than defensive athletes due to their position on the field. Figure 3.12 and one-way ANOVA ( $F(2,18117) = 565.6, p < .001$ ) confirm that there is a significant difference in the average scoring performance for each position in the field.

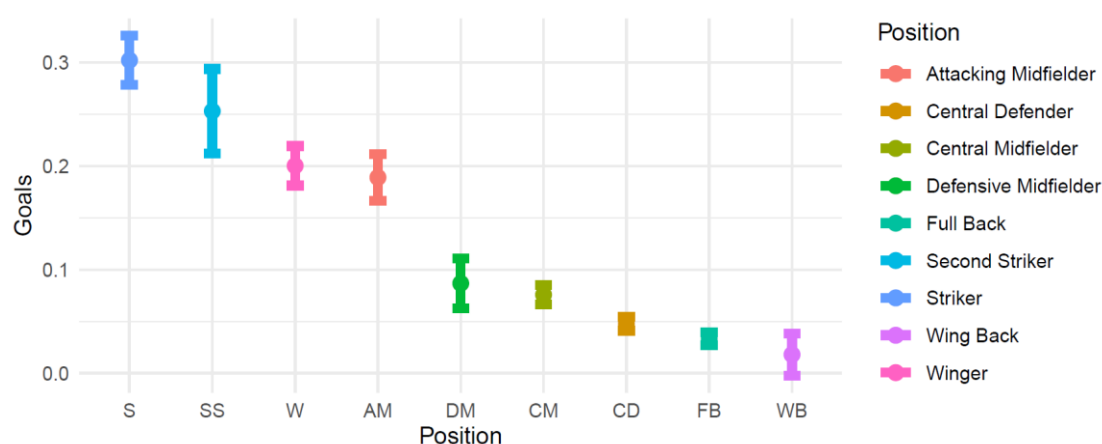


Figure 3.12: Error bars (mean goals +/- 95% CI) grouped by Position.

Strikers score the most goals compared to all positions followed by Second Strikers. Wingers and Attacking Midfielders seem to score the same as well as Defensive Midfielders and Central Midfielders. Next, we find Central Defenders and finally Full Backs and Wing Backs.

To assess whether there is a statistically significant difference in the means among all the pairs of the positions mentioned above we will perform multiple pairwise comparisons. Table 3.5 presents the table of p-values for the pairwise comparisons. Here, the p-values have been adjusted by the Benjamini-Hochberg Method.

	CD	FB	WB	DM	CM	AM	W	SS
FB	0.096	-	-	-	-	-	-	-
WB	0.319	0.592	-	-	-	-	-	-
DM	0.015	0.001	0.033	-	-	-	-	-
CM	0.001	0	0.049	0.495	-	-	-	-
AM	0	0	0	0	0	-	-	-
W	0	0	0	0	0	0.356	-	-
SS	0	0	0	0	0	0	0.001	-
S	0	0	0	0	0	0	0	0.003

Table 3.5: Pairwise comparisons using the Benjamini-Hochberg Method.

Strikers differ significantly from all other positions in terms of average goalscoring performance as well as Second Strikers. On the contrary, Wingers and Attacking Midfielders do not differ significantly between them ( $p = 0.356$ ) and therefore these two positions can be combined into one. Defensive Midfielders and Central Midfielders can be combined into one position too ( $p = 0.496$ ) as well as the 3 remaining positions.

Therefore, the following groupings result from the above analysis:

- Strikers -> (S)
- Second Strikers -> (SS)
- Wingers & Attacking Midfielders -> (AM)
- Defensive Midfielders & Central Midfielders -> (DM)
- Central Defenders, Full Backs & Wing Backs -> Backs (B)

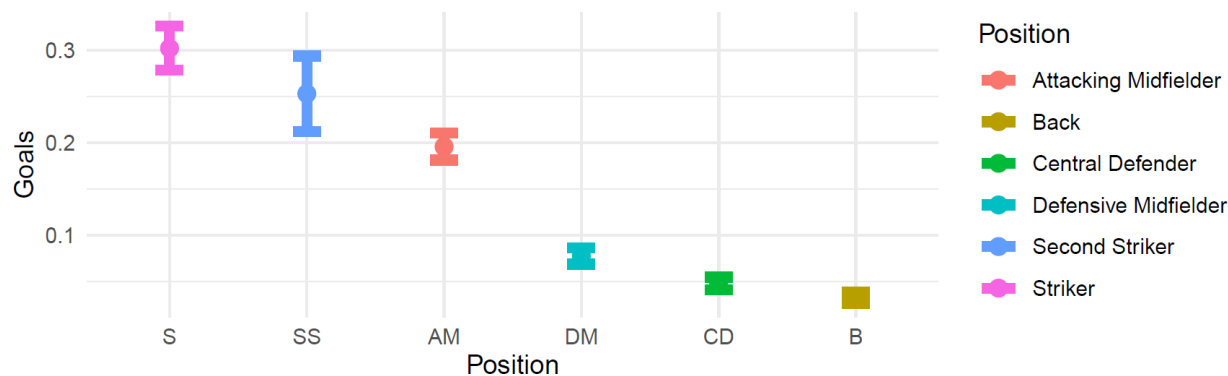


Figure 3.13: Error bars grouped by the new (grouped) Position.

### 3.2.2.3 Goals and Pre-Match Skill Actions

Having studied the relevance of home advantage and position with goals, it remains to examine the correlation between the pre-match skill actions of the athletes and the goals they are expected to score.

Table 3.6 shows the top 10 correlated pre-match skill actions we created in this Chapter.

	<b>Lag 5 skill action</b>	<b>Pearson's r</b>
1	an attempted shot from inside the box	0.28
2	total shots at goal	0.28
3	a shot attempt that came in open play	0.27
4	shot on target	0.27
5	touches inside the opposition's penalty area	0.27
6	goal scored	0.25
7	shot from the centre of the box	0.24
8	goal from a shot inside the box	0.24
9	a goals scored from a clear-cut chance	0.23
10	goals scored from regular play or on a fast break	0.23

Table 3.6: Top 10 correlated pre-match skill actions with goals.

From the table above, we notice that the most important pre-match skill actions related to the athletes' scoring ability are offensive actions. In the next chapter we will examine which of these skill actions will be included in our model for predicting the goals of athletes in future matches.

## 4 Models

This chapter thoroughly dives deep into the modelling process starting with the theory and finishing with the evaluation of the models. It is structured as follows:

- **Model Theory** gives the fundamental theory behind Linear Mixed Models and Generalized Linear Mixed Models.
- **Model Selection** describes the mathematical algorithms used to select the optimal model and applied to this procedure to our data.
- **Model Interpretation** summarizes the most important factors affecting the scoring performance of athletes and interprets the coefficients of the optimal model.
- **Model Evaluation** checks how our model performs comparing, for a list of athletes, the goal scoring probability given by our model with that given by bookmakers using different evaluation metrics.

### 4.1 Model Theory

The goal of this chapter is to present the theory of the Generalized Linear Mixed Models (GLMM) which we will use to predict the scoring ability of athletes. In order to simplify the comprehension of the GLMM, we will provide an introduction to the Linear Mixed Models (LMM).



### 4.1.1 Linear Mixed Models

In our data, we have information for each athlete at multiple points in time (Gameweeks). This type of data is called longitudinal data, meaning that the goals on the same athlete are correlated, since the repeated number of goals is collected on the same athlete. In order to obtain accurate conclusions regarding the regression parameters, it's necessary to incorporate this correlation into the model. The mixed effects model accommodates for this correlation between Gameweeks from the same athlete.

When evaluating longitudinal data, linear mixed models enable the analysis of both between and within-athletes effects. Due to the fact that the effects of the models may be classified as either fixed or random, they are known as mixed effects models (Fitzmaurice et al. 2011, Ch. 8). All the athletes that examine the effects between them share the fixed part, which is the population parameter. The random part, represented by  $b$  in the equation (5), fluctuates arbitrarily between the subjects and mimics the within-athletes effects that are distinctive to a certain athlete.

To illustrate the concept of a Linear Mixed Model (LMM), we present a random intercept model. In this model, the sole random effect is the intercept, and the analysis relies on just a single covariate. The measurement occasions of a subject in the model are represented as  $j = 1, 2, \dots, T_i$  and subjects is denoted by  $i = 1, 2, \dots, n$ . Consequently, the random intercept model may be described as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i} + e_{ij} \quad (5)$$

where  $Y_{ij}$  represents the outcome for the  $i$  subject at  $j$  occasion,  $b_{0i}$  represents the random effect, and  $e_{ij}$  represents the sampling error. The model assumes that both the random effect and residual are independent normal variables with an average value of zero and variance  $V(b_{0i}) = \sigma_b^2$  and  $V(e_{ij}) = \sigma^2$  respectively (Fitzmaurice et al. 2011, 191). Further,  $e_{ij} \perp e_{il}$  is assumed (i.e., residuals are independent across test occasions for  $j \neq l$ ).



The average response can be separated into two components: conditional and marginal means. This allows us to analyze both the effects specific to individual subjects and those pertaining to the overall population. The definitions for these means are as follows:

*Conditional mean*

$$E(Y_{ij}/b_{0i}) = \beta_0 + \beta_1 X_{ij} + b_{0i}$$

*Marginal mean*

$$E(Y_{ij}/b_{0i}) = \beta_0 + \beta_1 X_{ij}$$

A notable characteristic of mixed effects models is their assumption of interdependence among observations within the same subject, while assuming independence between different subjects. The model implies a correlation among observations originating from the same subject (Fitzmaurice et al. 2011, 194). This correlation can be demonstrated by examining the marginal covariance between any pair of responses from the same subject.

$$\begin{aligned} Cov(Y_{ij}, Y_{il}) &= Cov(\beta_0 + \beta_x X_{ij} + b_{0i} + e_{ij}, \beta_0 + \beta_1 X_{il} + b_{0i} + e_{il}) \\ &= Cov(b_{0i} + e_{ij}, b_{0i} + e_{il}) \end{aligned}$$

Since  $e_{ij} \perp e_{il}$  and  $e_i \perp b_0$  is assumed, the expression can be simplified further

$$Cov(Y_{ij}, Y_{il}) = Cov(b_{0i}, b_{0i}) = V(b_{0i}) = \sigma_b^2$$

The LMM is not limited to having just a random intercept; it can also involve random effects applied to other regression parameters. In this scenario, the model would incorporate randomly varying slopes. Using the same example as before, when extending equation (5) to include a random slope, the key distinction is the addition of another coefficient, denoted as  $b_{i1}$ , along with its associated covariate  $x_{ij}$  in the model (equation 6).

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i} + b_{i1} X_{ij} + e_{ij} \quad (6)$$



## 4.1.2 Generalized Linear Mixed Models

An expansion of the LMM is the Generalized Linear Mixed Model (GLMM). The GLMM posits that the conditional distribution of the outcome  $Y_{ij}$  for each subject, given the random effects  $b_i$  and covariates, follows a distribution that falls within the exponential family. (Stroup 2012, Ch. 4.5).

For our study, the number of goals given the athletes, belongs to the Poisson distribution. A GLMM is considered in this case a Poisson mixed effects model or in other words a Poisson regression model with random effects (Fitzmaurice et al. 2011, 400). As a result, it's common to model the response using the logarithmic link function. The linear predictor  $\eta_{ij}$  is identical to the one in LMM

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_i \quad (7)$$

and therefore, the equation (5) becomes:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 X_{ij} + b_{0i}, \quad b_{0i} \sim N(0, \sigma_b^2) \quad (8)$$

where:

- $\log(\lambda_{ij})$  represents the log of the expected count (goals) for the  $j^{th}$  observation (Gameweek) in the  $i^{th}$  group (athlete).
- $\beta_0$  is the fixed intercept, representing the average or overall effect on the log scale across all groups.
- $\beta_1$  is the fixed coefficient associated with the covariate  $x_{ij}$ , representing the effect of  $x_{ij}$  in log scale.
- $X_{ij}$  represents the value of the covariate for the  $j^{th}$  observation in the  $i^{th}$  group.
- $b_{0i}$  represents the random effect or random intercept for the  $i^{th}$  group, capturing the deviation of the intercept for that particular group from the overall intercept.



The random intercept  $b_{0i}$  allows for modelling the variability in the counts (goals) across different groups (athletes). It accounts for the clustering or nesting of observations within groups and captures the group-specific effects. The estimation of parameters in a Poisson GLMM involves maximum likelihood estimation methods, and the resulting estimates can be used for inference and prediction of the count variable  $Y_{ij}$ .

## 4.2 Model Selection

In the last subsection (4.1.2), the theory behind the model we will use (GLMM) to predict the scoring ability of athletes was presented. The next step is to choose the appropriate variables for our model that will help us to predict with the best possible accuracy the scoring performance of the athletes. Specifically, from the 247 pre-match skill actions we created in the previous chapter, we want to select the most important ones for determining the athletes' ability to score in future matches. Using mathematical procedures, which we will outline in the next subsections, this process can be automated.

### 4.2.1 Model Selection Criteria

The most common approach on the comparison and selection of statistical models that fit the same data is the penalized model selection criteria.

In this thesis, the criteria that will be used for the model selection are:

- AIC
- BIC



### 4.2.1.1 Akaike Information Criterion

The model with more explanatory variables is often chosen when a model's quality is merely evaluated in terms of fit. But this might not be a good thing because the model may be overfitted (fits the training data too closely, which leads to poor generalization to new, unseen data). AIC is a statistical measure developed by Akaike (1974) that balances model complexity and goodness of fit identifying simpler models that are less prone to overfitting by penalizing complexity and can be expressed as:

$$AIC = 2k - 2\log(\hat{L}),$$

where  $\hat{L}$  is the value of maximum of the likelihood function and  $k$  expresses the number of model's parameters.

### 4.2.1.2 Bayesian Information Criterion

BIC, like AIC, mitigates the issue of overfitting by imposing a penalty on the log-likelihood estimate for models that exhibit greater complexity. However, what sets BIC apart from AIC is its consideration of the sample size ( $n$ ), whether it's large or small, in the model's evaluation. BIC can be expressed as:

$$BIC = \log(n)k - 2\log(\hat{L}),$$

where  $\hat{L}$  is the maximum likelihood function's value,  $n$  is the number of samples, and  $k$  represents the model's parameter count.



The term  $2k$  in AIC partially penalizes complexity, but the penalty in BIC rises logarithmically with sample size, leading to a more apparent penalty. As a result, BIC favors simpler models than AIC does, encouraging a more frugal method of model selection.

## 4.2.2 Model Selection by Optimization

The earlier model selection methods involved comparing models that were fit by maximizing the Log-Likelihood, and once the parameters and the maximum Log-Likelihood were obtained, these scores were computed to identify the best model.

However, in our dataset, there are more than 247 skill actions, so we require a model selection approach that conducts variable selection as soon as the parameters are estimated. The Least Absolute Shrinkage and Selection Operator (LASSO) is a most-recognized method that accomplishes both parameter estimation and variable selection simultaneously.

### 4.2.2.1 Lasso

Tibshirani's Lasso, introduced in 1996, has gained significant popularity as a regression method employing an  $L_1$ -penalty on the regression coefficients. This results in the shrinking of all coefficients toward zero, with some being precisely set to zero. The fundamental concept is to maximize the log-likelihood  $l(\beta)$  of the model while simultaneously restricting the  $L_1$ -norm of the parameter vector  $\beta$ , yielding the Lasso estimate

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} l(\beta), \quad \text{subject to } \|\beta\|_1 \leq s,$$

with  $s \geq 0$  and with  $\|\cdot\|_1$  denoting the  $L_1$ -norm.



Equivalently the Lasso estimate  $\hat{\beta}$  can be derived by solving the optimization problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} [l(\beta) - \lambda \|\beta\|_1],$$

with  $\lambda \geq 0$

### 4.2.2.2 Lasso for GLMM

As we mentioned in subsection 4.1.1, our data is longitudinal (we have repeated measures for athletes) and therefore our model contains fixed and random effects. As the structure of random effects will influence which fixed effect variables are chosen, Bondell et al. (2010) advocate against separating the fixed and random parts of the model when doing variable selection.

Therefore, we need to simultaneously identify the important skill actions that correspond to both the fixed and random effects components in our GLMM. Our approach for the selection of relevant predictors is a Lasso-type technique which works by combining gradient ascent optimization with the Fisher scoring algorithm and is presented in detail in Groll and Tutz (2011). This algorithm applies  $L_1$  penalty shrinkage to the fixed effects of the GLMM, in the presence of random effects.

It is implemented in the corresponding *glmLasso* function of the corresponding R-package (Groll, 2011a; publicly available via CRAN, see [http:// www.r-project.org](http://www.r-project.org)).



### 4.2.3 Results

Before performing the GLMMLasso to our data we need to determine the penalty parameter  $\lambda$  that controls how much shrinkage we want for our parameters. Generally, the bigger the  $\lambda$ , the more shrinkage it is, thus the fewer the parameters are going to be included in the model. The number of  $\lambda$  is not selected automatically, we therefore need to test and choose the  $\lambda$  that works the best for our data. One way to do this somewhat is by using the information criteria AIC and BIC, which allow for the comparison between models with different number of parameters.

By choosing the model with the lowest AIC and BIC, it is equivalent to choosing the model with highest likelihood but also penalizes models that are too big, because likelihood always increases with the addition of the parameters.

To find the optimal lambda value we build a grid of possible lambda values from which to panelize the model, run the model again and again with a lambda value, and then determine which lambda is the best based on AIC and BIC values. In the following figure we see that the lambda value which minimizes these criteria is  $\lambda = 1800$ .

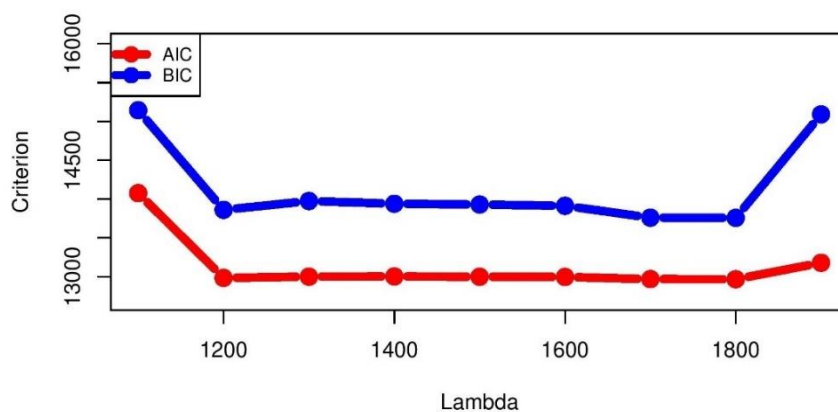


Figure 4.1: Optimal  $\lambda$  selection for the Poisson GLMM model with all the skill actions of the athletes using all data (2020-21, 2021-22 and 2022-23 seasons).

Only three of the 247 pre-match skill actions of the athletes are not set to zero by the lasso regression (Table 4.1), including the average attempted shots from inside the box in the last 5 games, the average total shots at goal in the last 5 games and the average touches in the opposition's penalty area in the last 5 games.

	Lasso beta
<i>(Intercept)</i>	-2.114
<i>an attempted shot from inside the box</i>	0.228
<i>total shots at goal</i>	0.23
<i>touches inside the opposition's penalty area</i>	0.05

Table 4.1: Most significant pre-match skill actions by LASSO.

The above three pre-match skill actions of the athletes selected by LASSO were also the ones with the highest correlation with goals and consequently the result is anything but unexpected. Apart from these pre-match skill actions, we showed in the previous chapter that the home advantage and the position of the player are important for predicting his goalscoring ability in a match and then they can be added as predictor variables along with the above 3 skill actions. Finally, we can add the elo difference of the two teams to the list of our independent variables in order to have a team-based predictor. Summing up, the variables that will make up our model are as follows:

### 1. Fixed part

- Side (Home / Away)
- Position (B, DM, AM, SS, S)
- Elo Difference
- Average attempted shots from inside the box in the last 5 games
- Average total shots at goal in the last 5 games
- Average total touches in the opposition's penalty area in the last 5 games



## 2. Random part

- Athletes

The model fit is shown in the table below:

	<i>Dependent variable:</i>
	Goals
Home	0.118*** (0.043)
Defensive Midfielder (DM)	0.527*** (0.097)
Attacking Midfielder (AM)	1.105*** (0.096)
Second Striker (SS)	1.321*** (0.157)
Striker (S)	1.504*** (0.107)
Elo Difference	0.002*** (0.0002)
An attempted shot from inside the box	0.035 (0.075)
Touches in Opponent Box Aver	0.064*** (0.022)
Total Scoring Attempts Aver	0.137** (0.056)
Constant	-3.533*** (0.072)
Number of Athletes	670
SD (Athletes)	0.354
Number of Observations	18120
AIC	12325
BIC	12411
R Squared Conditional	0.22
R Squared Marginal	0.18

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

Table 4.2: GLMM proposed by glmmLasso using all data.

We notice that the coefficient of the average attempted shots from inside the box in the last 5 games is not statistically significant at any level of significance, even though that variable has the highest correlation with goals. This happens because this variable is highly correlated with the other two pre-match skill actions (Multicollinearity) as we can see in the figure below:



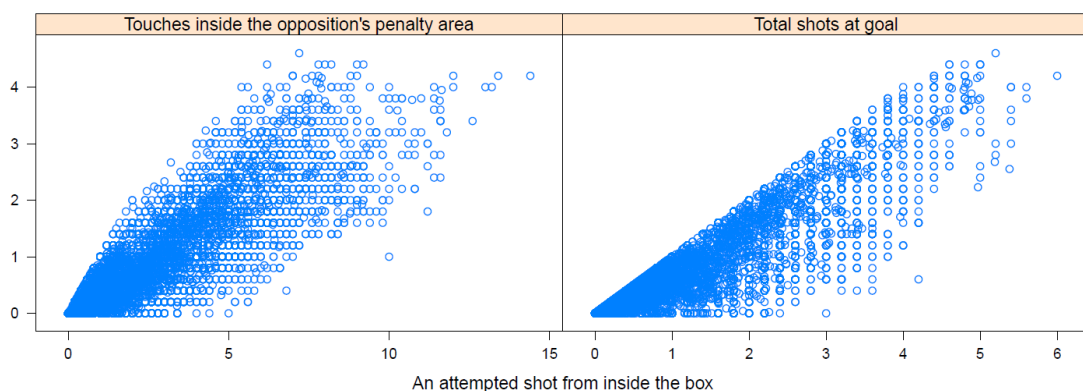


Figure 4.2: Highly correlated predictors (Multicollinearity).

In a multiple regression model, when two or more predictor variables exhibit strong correlation, this statistical phenomenon is referred to as multicollinearity, which results in an increase in the standard errors of the coefficients. In turn, higher standard errors suggest that certain independent variable coefficients may not be determined to be substantially different from 0. In other words, multicollinearity makes some variables statistically insignificant when they should be significant by exaggerating the standard errors. Those coefficients could be important in the absence of multicollinearity (and hence with reduced standard errors).

The variance inflation factor (VIF), which evaluates how much the variance of an estimated regression coefficient rises if our predictors are linked, is one technique to quantify multicollinearity. The VIFs will all be 1 if there are no associated factors. A poor correlation between that predictor and other predictors is indicated by a VIF of less than 5. VIF values greater than 10 are a symptom of excessive, intolerable correlation of model predictors, whereas values between 5 and 10 suggest a moderate correlation. The VIFs values for our predictors are shown in table 4.3:

	<b>Term</b>	<b>VIF</b>
1	Side	1.00
2	Position	1.93
3	Elo Difference	1.04
4	Touches inside the opposition's penalty area	3.31
5	Total shots at goal	4.88
6	An attempted shot from inside the box	6.46

Table 4.3: VIFs values for model's predictors.

As expected, the average attempted shots from inside the box in the last 5 games has the highest VIF value, which is above 5 (moderate correlation). This leads to the removal of this variable from the model. Figure 35 shows the VIFs values for the model's predictors after the removal of this variable. As can be seen in the table 4.4, the new model does not have the problem of multicollinearity.

	<b>Term</b>	<b>VIF</b>
1	Side	1.00
2	Position	1.71
3	Elo Difference	1.04
4	Touches inside the opposition's penalty area	2.49
5	Total shots at goal	2.42

Table 4.4: VIFs values for the final model's predictors.

### 4.3 Final Model

Table 4.5 shows the results from fitting the final model:

	<i>Dependent variable:</i>
	Goals
Home	0.118 <sup>***</sup> (0.043)
Defensive Midfielder (DM)	0.519 <sup>***</sup> (0.096)
Attacking Midfielder (AM)	1.093 <sup>***</sup> (0.094)
Second Striker (SS)	1.312 <sup>***</sup> (0.156)
Striker (S)	1.504 <sup>***</sup> (0.107)
Elo Difference	0.002 <sup>***</sup> (0.0002)
Touches in Opponent Box Aver	0.069 <sup>***</sup> (0.019)
Total Scoring Attempts Aver	0.156 <sup>***</sup> (0.040)
Constant	-3.535 <sup>***</sup> (0.072)
Number of Athletes	670
SD (Athletes)	0.355
Number of Observations	18120
AIC	12324
BIC	12402
R Squared Conditional	0.22
R Squared Marginal	0.18

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

Table 4.5: Estimates for our final GLMM using all data.

All the coefficients of the final model are statistically significant (including the random effect as its estimated standard deviation component is greater than zero), and we can then move on to their interpretation.



### 4.3.1 Interpretation

Our final model in Table 4.5 has the form:

$$\log(\lambda_{ij}) = \beta_0 + (Home)_{ij}\beta_1 + (DM)_{ij}\beta_2 + (AM)_{ij}\beta_3 + (SS)_{ij}\beta_4 + (S)_{ij}\beta_5 + \\ (Elo\ Difference)_{ij}\beta_6 + \\ (Touches\ inside\ the\ opposition's\ player\ area)_{ij}\beta_7 + \\ (Tota\ shots\ at\ goal)_{ij}\beta_8 + b_{0i},$$

where  $\lambda_{ij}$  denotes the expected number of goals score by Athlete  $i$  at game  $j$  and  $b_{0i} \sim N(0, \sigma_b^2)$  represents athlete-specific random intercepts. The interpretation of the effects in the model is as follows:

#### 1. Fixed Effects:

- $\beta_0$  (-3.535): The expected number of goals for an athlete whose team is the away team, his position is Back (B), the elo difference between his team and his opponent is zero, his average touches inside the opposition's penalty area in his last 5 games is zero and his average total scoring shots at goal in his last 5 games is zero, is  $e^{-3.535} = 0.03$ .
- $\beta_1$  (0.118): The expected number of goals for an athlete whose team is the home team, is  $(e^{0.118} - 1)100\% = 12.52\%$  higher than the expected number of goals for the same athlete when his team is the away team, assuming that the remain variables remain constant.
- $\beta_2$  (0.519): The expected number of goals for an athlete who plays as Defensive Midfielder (DM), is  $(e^{0.519} - 1)100\% = 68.03\%$  higher than the expected number of goals for the same athlete who plays as Back (B), assuming that the remain variables remain constant.



- $\beta_3$  (1.093): The expected number of goals for an athlete who plays as Attacking Midfielder (AM), is  $(e^{1.093} - 1)100\% = 198.32\%$  higher than the expected number of goals for the same athlete who plays as Back (B), assuming that the remain variables remain constant.
- $\beta_4$  (1.312): The expected number of goals for an athlete who plays as Second Striker (SS), is  $(e^{1.312} - 1)100\% = 271.36\%$  higher than the expected number of goals for the same athlete who plays as Back (B), assuming that the remain variables remain constant.
- $\beta_5$  (1.504): The expected number of goals for an athlete who plays as Striker (S), is  $(e^{1.504} - 1)100\% = 349.97\%$  higher than the expected number of goals for the same athlete who plays as Back (B), assuming that the remain variables remain constant.
- $\beta_6$  (0.002): Each additional point in elo difference over an athlete's team is associated with  $(e^{0.002}-1)100\% = 0.2\%$  more goals assuming that the remain variables remain constant.
- $\beta_7$  (0.069): Each additional touch in an athlete's average total touches inside the opposition's penalty area in the last 5 games is associated with  $(e^{0.069}-1)100\% = 7.19\%$  more goals assuming that the remain variables remain constant.
- $\beta_8$  (0.156): Each additional shot in an athlete's average total shots at goal in the last 5 games is associated with  $(e^{0.156}-1)100\% = 16.88\%$  more goals assuming that the remain variables remain constant.

## 2. Random Effects

- $\sigma_b$  (0.355): The random intercept standard deviation, or between-subject standard deviation ( $\sigma_b$ ), indicates how much athletes differ from each other.



Table 4.6 presents the 10 athletes with the highest random intercept using the final model of the Table 4.5

	<b>Player</b>	<b>Random Intercept</b>
1	James Ward–Prowse	0.85
2	Ilkay Gündogan	0.75
3	Jorge Luiz Frello Filho	0.75
4	Heung–Min Son	0.74
5	Wilfried Zaha	0.65
6	Bruno Guimarães Rodriguez Moura	0.64
7	Ferran Torres	0.60
8	Gareth Bale	0.59
9	James Maddison	0.57
10	Matheus Felliipe Costa Pereira	0.55

Table 4.6: Top 10 athletes with the highest random intercepts of the final model of Table 4.5.

### 4.3.2 Assumptions

Some presumptions must be true for the parameter estimations of the final model to be fair and, as a result, for the statistical inference to be accurate. The model may not effectively represent the underlying data process if the model assumptions are not followed, which might lead to incorrect estimations and hypothesis testing.

Assessing the assumptions helps identify potential outliers, influential observations, or patterns in the residuals that may require further investigation or model refinement. By confirming that the model assumptions are reasonably met, we gain confidence in the reliability and generalizability of the findings and can make more accurate interpretations and predictions based on the model.



The first assumption is that the response variable (goals) follows a Poisson distribution, meaning that it represents counts (goals), and the mean and variance are approximately equal. This assumption was tested in the previous chapter (subsection 3.2.1), and we can therefore consider our data a Poisson distribution with mean and variance approximately equal.

The second assumption is that the random effects are normally distributed. This assumption allows for the estimation of the variance components associated with the random effects and captures the variability between different groups (in our case, athletes). However, it is important to note that the normality assumption applies only to distribution of the random effects, not the response variable itself which follows a Poisson distribution as we have shown previously. To assess the normality assumption for random effects, we conduct the following quantile-quantile (Q-Q) plot.

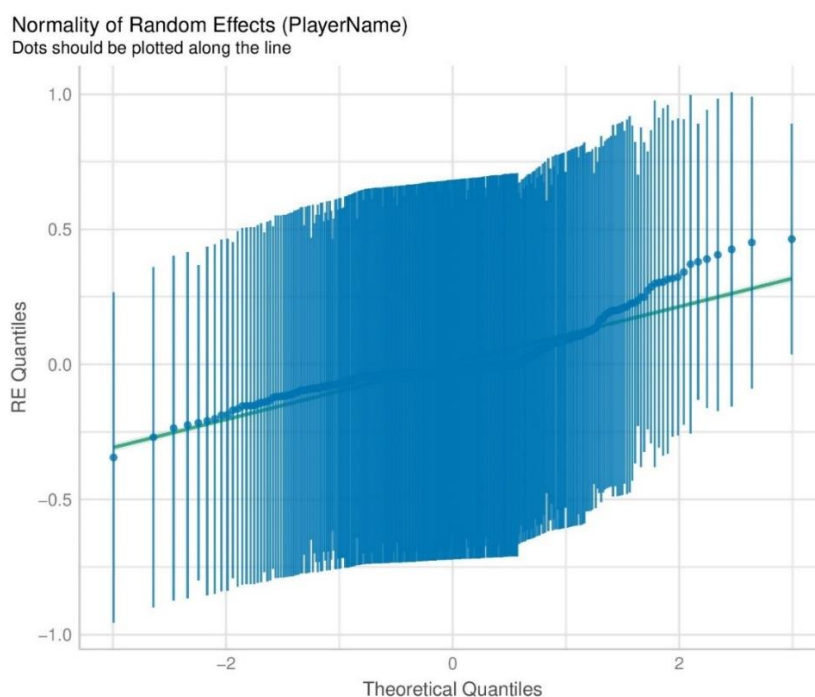


Figure 4.3:Q-Q plot for random effects of the final model of Table 4.5.

The dots in Figure 4.5 represent the random intercepts estimated for each athlete separately, while the line represents the 95% confidence interval of each random intercept estimate. Most points are plotted along the diagonal line and therefore the normal distribution seems a reasonable assumption.

Finally, it is important to check for influential points in our data. Influential points refer to observations that have a significant impact on the model's estimates, standard errors, hypothesis tests, and model fit. In figure 4.10 we assess the influence of individual observations using Cook's distance which is a measure of the influence of each observation on the model estimates and quantifies the change in model estimates when a particular observation is removed. Observations with large Cook's distances are considered influential. Figure 4.6 shows that there are no influential points.

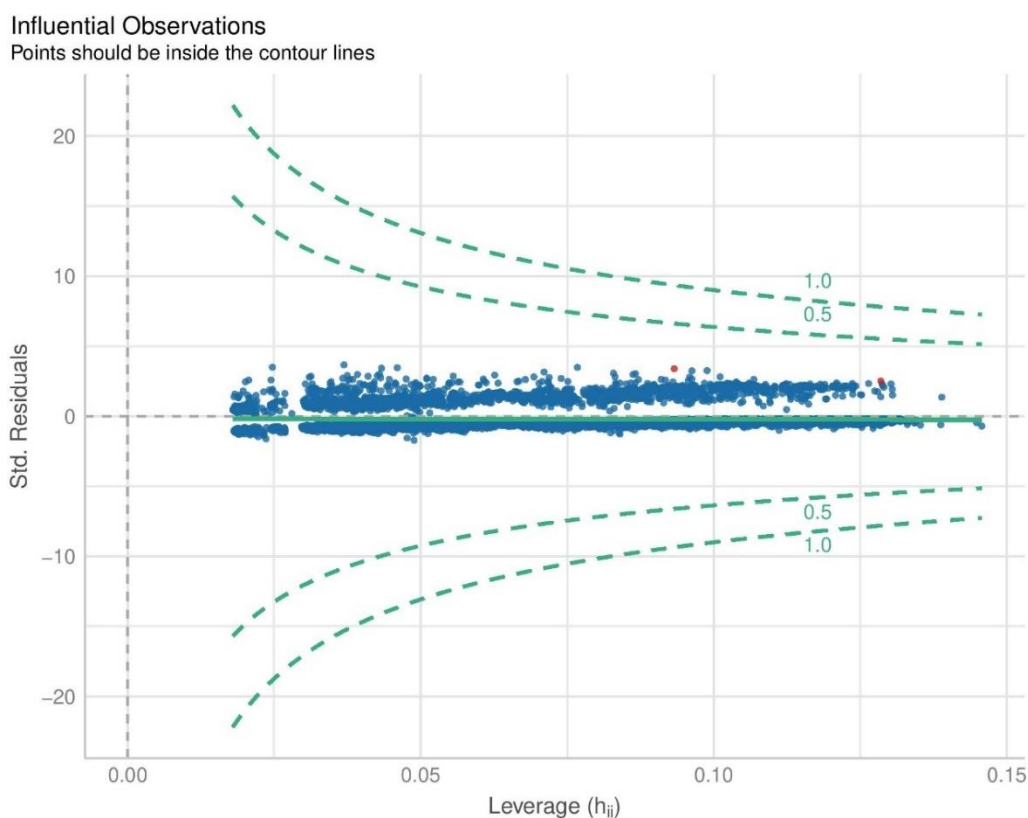


Figure 4.4: Cook's distance for influential points detection of the final model of Table 4.5.

## 4.3 Model Evaluation

In this section, we will evaluate our model comparing its predictive performance against actual betting odds to see if we come close to professional bookmakers' complex models.

It is structured as follows:

- **Bookmakers' Odds** refers to the goalscoring odds given by bookies for various athletes each Gameweek and how they can be compared with the counterparts given by our model.
- **Evaluation Metrics** presents the theory behind the evaluation metrics that we will use to assess the performance of our model compared to the bookies' model.
- **Training and Test Sets** splits the data into training and testing sets. The training sets, which will contain one year data, will be used for training our model and the testing sets, which will contain 1 Gameweek data, will be used for testing.
- **Comparison with Bookies** compares the results of the predictions given by our model with those of bookies for a specific list of athletes in different Gameweeks.

### 4.3.1 Bookmakers' Odds

The odds of the bookmakers that we will use to compare with those from our model will be those from the website *checkthechance.com* which is a website that was created in August 2020 and use the information available on the bookies market to supply probabilities on sports. Their model adjusts for the margins used by the bookies to get the most accurate predictions, and it's all displayed conveniently in percentages. Over time, *checkthechance.com* has been tailored towards Fantasy Premier League (FPL), and it provides Clean Sheet tables, Goalscorer tables as well as individual athlete predictions.



As far as Goalscorer tables, which in this case are the ones we are interested in, *Check the Chance* gives every Gameweek the goalscoring probability for a specific list of athletes. In the figure below, we see an example which concerns the first Gameweek of the 2022-2023 season.

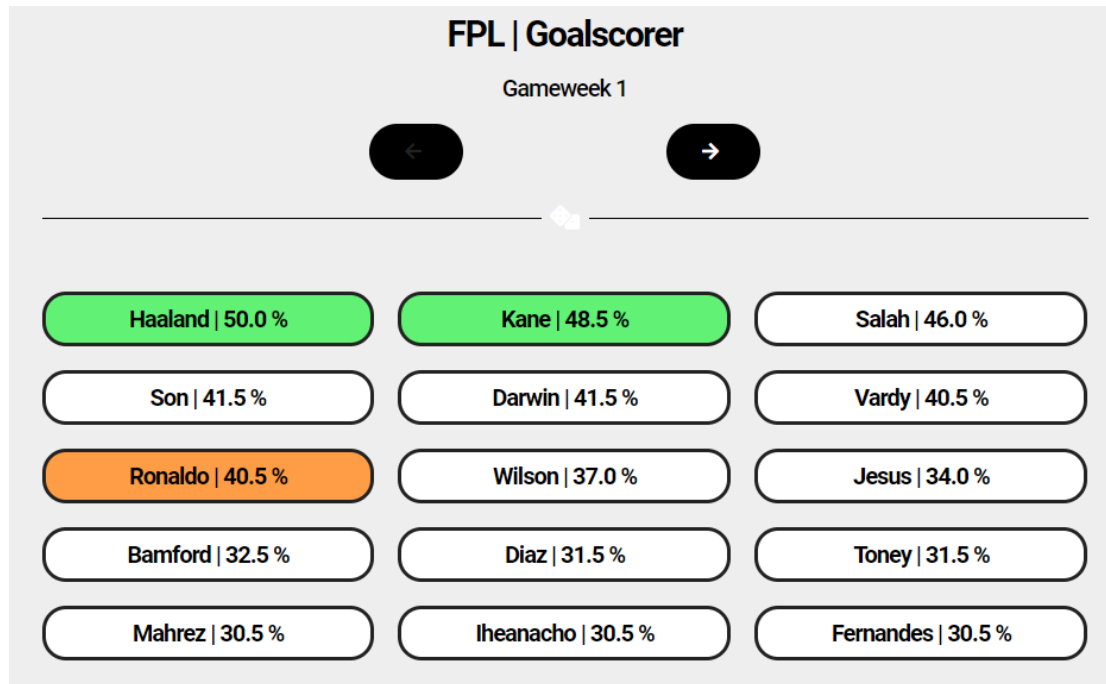


Table 4.5: Goalscoring probabilities for a specific list of athletes by bookies. Source: *checkthechance.com*.

### 4.3.2 Evaluation Metrics

As we saw in the subsection above, bookies give the goalscoring probability each Gameweek for a specific list of athletes. Therefore, in order to compare our model with that of the bookies, we need to 1) convert the expected goals of the athletes estimated by our model into goalscoring probabilities and 2) use the appropriate metrics to compare the predictive ability of the two models.

As for the first part, an athlete's goalscoring probability can be calculated through Poisson distribution as:

$$P(k > 0) = 1 - P(k = 0) = 1 - \frac{e^{-\lambda} \lambda^0}{0!},$$

where  $\lambda$  represents the expected number of athlete's goals in a match.

As regards part two, to determine a model's performance, several metrics are employed in the literature; the metric to choose depends on the study's objectives. For example, it makes no sense to use accuracy as a metric for goalscoring probabilities. A goalscoring probability of 0.49 and 0.01 will be classified as a no-goal, but if it is actually a goal the error in the first case is a lot lower than the second one. The metrics that are used to evaluate the performance of the two models are described below.

### 4.3.2.1 Log Loss

Instead of focusing only on the most likely class, log loss assesses the accuracy of a classifier by the probability for every potential class. The cross-entropy between the distribution of the predictions and the actual labels is known as log loss. Cross entropy calculates the additional unpredictability that results from assuming a different distribution than the underlying model distribution. The equation of the log loss in a binary classification problem is given by:

$$LL = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

where  $N$  is the number of observations in the test set,  $y$  is the actual/true value (0 or 1), and  $p$  is the prediction probability.



In figure 4.7 a binary classification problem's log loss functions are fit. As can be observed, prediction accuracy increases as log loss decreases. Large forecast errors are particularly harshly penalized due to the logarithmic function's shape.

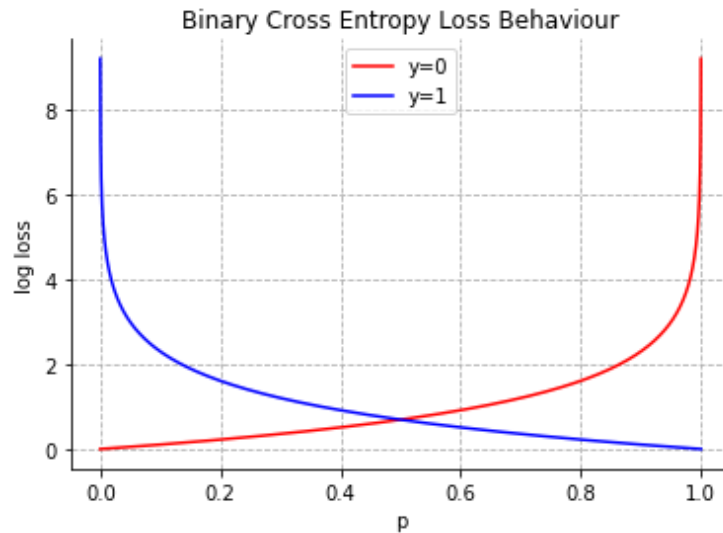


Figure 4.5: Graphical representation of log loss functions of a binary classification problem.

### 4.3.2.2 AUC-ROC

AUC-ROC is a performance indicator that indicates a model's ability to differentiate across classes. AUC-ROC is an amalgamation of the terms "Area Under Curve" (AUC) and "Receiver Operating Characteristics" (ROC). At various categorization thresholds, a ROC curve compares the True Positive Rate with the False Positive Rate. A higher AUC-ROC score denotes greater prediction accuracy since the AUC calculates the area under the ROC curve. AUC has the benefit of being independent of the selected classification criterion. Distinct AUC-ROC curve examples are displayed in Figure 4.11, each with a distinct graph and score. It demonstrates that a higher AUC-ROC value corresponds to improved precision accuracy. The dotted line indicates the expectation for the AUC-ROC for random guessing, so it is even for possible for the AUC-ROC to have a lower score than random guessing.

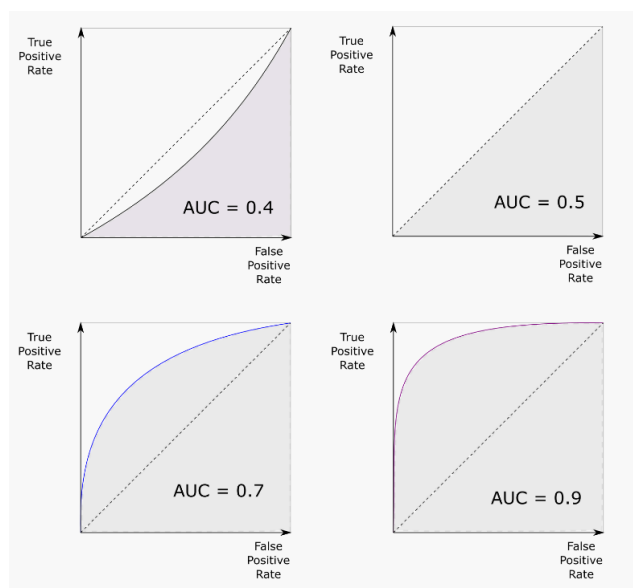


Figure 4.6: Graphical representation of different AUC-ROC curves.

### 4.3.3 Training and Test Sets

The table below gives the goalscoring probabilities of the athletes selected by the bookies in the Gameweek 16. To estimate their goalscoring probabilities, we train our model in 1 year data before the match we want to make the prediction and predict the goalscoring probability this Gameweek.

	Player	Team	Opponent	Goals	Model	Bookies
1	Darwin Núñez	Liverpool	Southampton	1	0.56	0.42
2	Erling Haaland	Manchester City	Brentford	0	0.76	0.61
3	Gabriel Fernando de Jesus	Arsenal	Wolverhampton Wanderers	0	0.39	0.37
4	Harry Kane	Tottenham Hotspur	Leeds United	1	0.59	0.45
5	Kevin De Bruyne	Manchester City	Brentford	0	0.38	0.30
6	Marcus Rashford	Manchester United	Fulham	0	0.24	0.30
7	Mohamed Salah Ghaly	Liverpool	Southampton	0	0.52	0.48
8	Phil Foden	Manchester City	Brentford	1	0.40	0.32
9	Richarlison de Andrade	Tottenham Hotspur	Leeds United	0	0.34	0.31
10	Roberto Firmino Barbosa de Oliveira	Liverpool	Southampton	1	0.40	0.38

Table 4.8: Goalscoring probabilities for athletes selected by the bookies.

“Goals” is 1 if the athlete scored in the corresponding Gameweek and 0 otherwise. “Model” is the goalscoring probability estimated by our model and “Bookies” is the goalscoring probability given by the bookies. The Log-Loss for our model is 0.676 while for bookies’ model is 0.691. On the other hand, AUC for our model is 0.75 while for bookies’ model is 0.625. From both the Log-Loss and the AUC, it seems that our model has a better predictive ability than that of the bookies for this Gameweek.

Because with only one Gameweek for test set we do not get very safe conclusions, we create various test sets for which we train our model on train sets which are one year data before the corresponding test set.

- Train 1: 16/09/2021 – 04/09/2022  
Test 1: 16/09/2022 – 18/09/2022 (Gameweek 8)
- Train 2: 01/10/2021 – 18/09/2022  
Test 2: 01/10/2022 – 03/10/2022 (Gameweek 9)
- Train 3: 08/10/2021 – 03/10/2022  
Test 3: 08/10/2022 – 10/10/2022 (Gameweek 10)
- Train 4: 14/10/2021 – 10/10/2022  
Test 4: 14/10/2022 – 16/10/2022 (Gameweek 11)
- Train 5: 18/10/2021 – 16/10/2022  
Test 5: 18/10/2022 – 20/10/2022 (Gameweek 12)
- Train 6: 22/10/2021 – 20/10/2022  
Test 6: 22/10/2022 – 24/10/2022 (Gameweek 13)
- Train 7: 29/10/2021 – 24/10/2022  
Test 7: 29/10/2022 – 30/10/2022 (Gameweek 14)



- Train 8: 05/11/2021 – 30/10/2022  
Test 8: 05/11/2022 – 06/11/2022 (Gameweek 15)
- Train 9: 12/11/2021 – 06/11/2022  
Test 9: 12/11/2022 – 13/11/2022 (Gameweek 16)

### 4.3.4 Comparison with Bookies

Figures 4.7 and 4.8 show the log-loss and the AUC respectively of the two models.

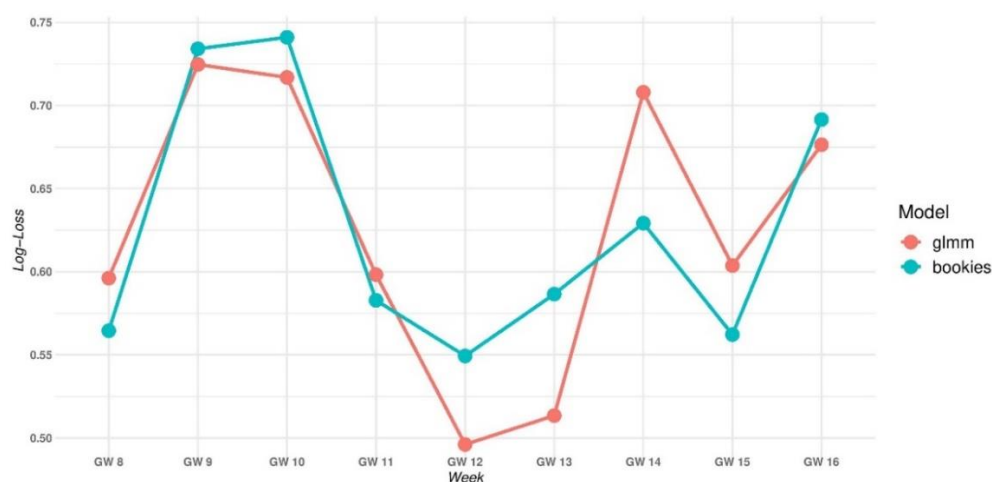


Figure 4.7: Log-Loss of the two models in different test sets.

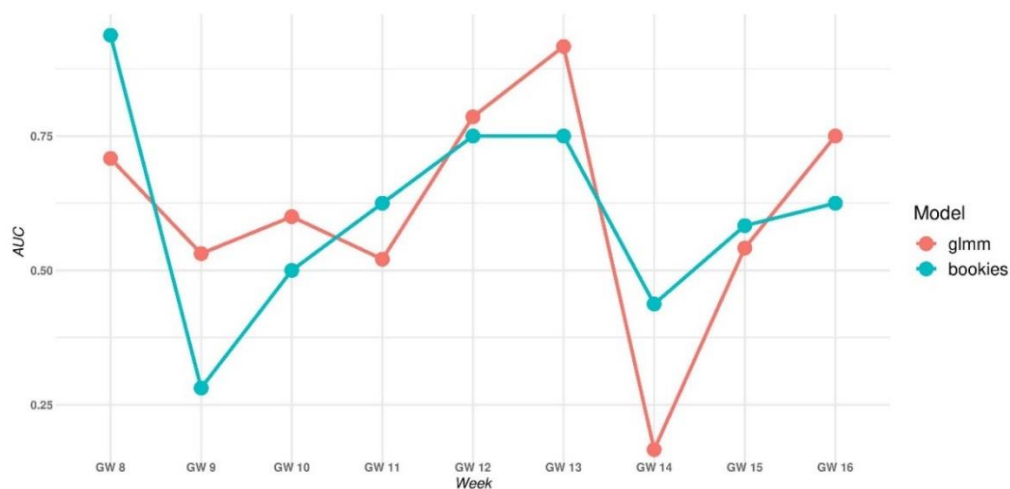


Figure 4.8: AUC of the two models in different test sets

From figure 4.7 we notice that our model performs better in 5 of the 9 test Gameweeks (Gameweeks 9, 10, 12, 13 and 16) in relation to the bookies' model as the log-loss in these Gameweeks is smaller. We would reach the same conclusion if we observed the immediately following figure (Figure 4.8), as in these test Gameweeks our model has a higher AUC value than bookies and as we mentioned above, the greater the AUC of a model, the greater predictive accuracy it has.

To decide which model is finally the best, we calculate the average of these two metrics for the two models in the 9 test Gameweeks. Therefore, the average values of the log-loss and AUC for our model are 0.62 and 0.63 respectively while the corresponding values for the bookies model are 0.61 and 0.67. Therefore, on average, the bookies' model is slightly better.

Beyond the above metrics, it would be interesting to see what a player's profit would be if he bet on the scoring of an athlete using the two models or otherwise from the side of the betting companies, which model brings them the most money (the least profit of the players).

In order to answer the above questions, we will initially use as an example the athletes of Table 4.8 which concerns the 16<sup>th</sup> Gameweek. Suppose a player bets 50 euros for the athletes of Table 4.8 (Gameweek 16) to score and specifically for Darwin Núñez, using our model. Then, as this athlete scored, he would win 50 euros times the inverse of the probability of scoring (bet market) of our model (1.79). Therefore, he would win 89.5 euros and his profit would amount to 39.5 euros. Table 4.9 presents the player's profits depending on which model he would bet on.



	Player	Goals	Model	Bookies	Model Profit	Bookies Profit
1	Darwin Núñez	1	1.79	2.38	39.5	69.0
2	Erling Haaland	0	1.32	1.64	-50.0	-50.0
3	Gabriel Fernando de Jesus	0	2.56	2.70	-50.0	-50.0
4	Harry Kane	1	1.69	2.22	34.5	61.0
5	Kevin De Bruyne	0	2.63	3.33	-50.0	-50.0
6	Marcus Rashford	0	4.17	3.33	-50.0	-50.0
7	Mohamed Salah Ghaly	0	1.92	2.08	-50.0	-50.0
8	Phil Foden	1	2.50	3.12	75.0	106.0
9	Richarlison de Andrade	0	2.94	2.23	-50.0	-50.0
10	Roberto Firmino Barbosa de Oliveira	1	2.50	2.63	75.0	81.5

Table 4.9: Player's profits depending on which model he will bet on for the athletes of Table 4.8 (Gameweek 16).

The average profit (per athlete) for the player who would bet with our model to the above athletes and for that Gameweek would be -7.6 euros (loss) while the corresponding average profit (per athlete) for the player who would bet with bookies' model would be 1.75 euros. Therefore, for a betting company, our model would be more profitable.

As far as the average profit (per athlete) in the last 9 Gameweeks (8-16) is concerned, then betting with our model we would have an average profit of -3.94 euros (loss) while if we bet with the bookies' model we would have an average profit of -4.74 euros (loss). Therefore, the bookies' model would be the one that would bring more money on average to the betting companies.

## 5 Conclusions and Future Work

This thesis aimed to develop a comprehensive framework for predicting the goal-scoring performance of soccer athletes. Through an in-depth analysis of various factors that influence goal scoring, ranging from individual athlete attributes to team dynamics, we have successfully constructed a predictive model that demonstrates promising accuracy and potentially real-world applicability.

The findings of this thesis underscore the importance of a holistic approach when evaluating goal scoring performance. By considering a multitude of variables such as athletes' historical performance (goals, total scoring attempts, touches in opponent box), team dynamics (home advantage, Elo rating) and the position on the pitch, our predictive model showcases its ability to capture the nuanced nature of soccer gameplay. The integration of statistical learning techniques (generalized linear mixed model) enabled us to effectively model the above relationships with the data and generate meaningful predictions.

While this thesis represents a substantial step towards the prediction of goal-scoring performance of soccer athletes, several avenues for future research and improvements remain. Here are some directions that want exploration:

1. Different number of lags in athletes' skill actions: In this thesis, we used lag 5 based on AIC criterion in order to create pre-match skill actions. A different approach of choosing the lag might have given a value that would have improved the predictive ability of the model.
2. Prediction for the 1<sup>st</sup> match of an athlete: In our model, we used the performance of an athlete in his past matches to predict his goal-scoring performance in the current match. Therefore, an improvement that needs to be made in future research is to develop a methodology exclusively for each athlete's first match.



3. Lower Elo Rating for promoted teams: In the Elo rating we developed for this thesis, all teams started with the same Elo rating. However, a team that has just been promoted to the Premier League needs to start with a lower Elo because it is considered to have lower potential than the rest of the Premier League teams.
4. Athlete's Injury and Recovery: Our model does not take into account whether an athlete is coming back from an injury. For future research, we suggest investigating the impact of athlete injuries and recovery periods on goal-scoring performance, as these events can significantly disrupt an athlete's rhythm and form.
5. External Factors: Last but not least, we suggest expanding the model to incorporate external factors such as weather conditions or stadium atmosphere. These variables could contribute significantly to goal-scoring variations.

By addressing these areas, future research endeavors can build upon the foundation laid by this thesis, advancing the state of the art in goal-scoring performance prediction and enriching the soccer industry's analytical capabilities.



## References

- Alamar, B., & Mehrotra, V. (2011). Beyond 'Moneyball': The rapidly evolving world of sports analytics, Part I. *Analytics Magazine*.
- Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton.
- Shea, S. (2018). THE 3-POINT REVOLUTION. *ShotTracker*.
- Moroney, M.J. (1956). *Facts from Figures* (3rd ed.). Penguin.
- Reep, C. (1971). Skill and chance in ball games. *Journal of the Royal Society Series A*, **131**, 581-585.
- Hill, I.D. (1974). Association football and statistical inference. *Applied Statistics*, **23**, 203-208.
- Maher, M.J. (1982). Modelling association football scores. *Statistica Neerlandica*, **36**, 109-118.
- Dixon, M.J., & Coles, S.C. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, **46**, 265-280.
- Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Statistician*, **49**, 399-418.
- Growder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English Football League matches for betting. *Statistician*, **51**, 157-168.



Forrest, D., & Simmons, R. (2000). Forecasting sport: The behavior and performance of football tipsters. *International Journal of Forecasting*, **16**, 317-331.

Kuypers, T. (2000). Information and efficiency: An empirical study of a fixed odds betting market. *Applied Economics*, **32**, 1353-1363.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D*, **52**, 381-393.

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of the Royal Statistical Society*, **37**, 253-264.

Owen, A. (2011). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, **22**, 99-113.

Koopman, S., & Lit, R. (2014). A dynamic bivariate Poisson model for analyzing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A*, **178**, 167-186.

Elo, A.E. (1978). *The Rating of Chess Players, Past and Present*. Arco Publishing.

Buchdahl, J. (2003). *Fixed Odds Sports Betting: Statistical Forecasting and Risk Management*. High Stakes.

Boulier, B.L., & Stekler, H.O. (1999). Are sports seedings good predictors? An evaluation. *International Journal of Forecasting*, **15**, 83-91.



Clarke, S.R., & Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, **7**, 585-594.

Hvattum, L. M., & Arntzen, H. (2010). Using Elo ratings for match result prediction in association football. *International Journal of Forecasting*, **23**, 460-470.

Sullivan, C., & Cronin, C. (2016). Improving Elo rankings for sports: Experimenting on the English Premier League. *Virginia Tech*.

McHale, N., & Szczepanski. (2014). A mixed effects model for identifying goal scoring ability of footballers. *Journal of the Royal Statistical Society*, **177**, 397-417.

Bonomo, F., Duran, G., & Marengo, J. (2014). Mathematical programming as a tool for virtual soccer coaches: A case study of a fantasy sport game. *International Transactions in Operational Research*.

Fitzmaurice, G.M., Laird, J.H., & Laird, N.M. (2011). Applied Longitudinal Analysis (2nd ed.). *Hoboken, New Jersey: Wiley*.

Sammut, C., & Webb, G.I. (2017). Bias Variance Decomposition. *Springer US*.

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

Hurvich, C.M., & Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.



Cavanaugh, J.E. (1997). Unifying the derivations for the Akaike and the corrected Akaike information criteria. *Statistics & Probability Letters*, **33**, 201-208.

Burnham, K.P., & Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, **33**, 261-304.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, **58**, 267-288.

Bondell, H.D., Krishna, A., & Ghosh, S.K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear-Mixed-Effects Models. *Biometrics*, **66**, 1069-1077.

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*, **24**, 137-154.

Bhandari, A. (2023). Multicollinearity: Causes, Effects, and Detection Using VIF. *Analytics Vidhya*.



## Appendix: Opta - Skill Actions

Accurate passes
Accurate corner distributions into the box
Accurate crosses (via corners, open play and setpieces)
Accurate crosses (including freekicks but excluding corners)
An accurate pass that ends in the attacking half of the pitch (including crosses but excluding throw-ins and keeper throws)
A successful pass by a striker who has received the ball with his back to goal and then plays the ball backtowards team-mates (excluding throw-ins, crosses, keeper throws)
All accurate passes (excluding throw-ins, keeperthrows and crosses)
Accurate passes that leave a player one-on-one withthe goalkeeper
All accurate throw-ins
Unsuccessful aerial duels
Successful aerial duels
Shot from the centre of the box
Total shots from outside the box, in the centre (less than 35 yards out)
Shot from the right of the box
Shot from the left of the box
Shot that came after a corner was taken
Shot resulting from a counter-attack

Free kick goals
A direct free kick saved by the goalkeeper
Total shots from direct free kicks
A direct free kick that misses its target (the goal)
A direct free kick that hits the goalpost (no goal)
A goal in the top centre of the goalpost
A goal in the top left of the goalpost
A goal in the top right of the goalpost
A goal in the bottom centre of the goalpost
A goal in the bottom left of the goalpost
A goal in the bottom right of the goalpost
A headed goal
A headed attempt off target
A headed attempt that hits the woodwork
A headed attempt that is saved by the goalkeeper
A headed attempt that is saved by the goalkeeper
Shot attempt from inside the box that is blocked (excludes any shots saved/cleared off the line in a defensive act – this is not the same as a regular goalkeeper save on the line)
Goal from a shot inside the box
Shot from inside the box that goes wide of or over the goal
Shot from inside the box that hits the post or bar

Shot from inside the box that are saved by the goalkeeper
own goal from inside the box
own goal from outside the box
Goal from a left-footed shot
Left-footed shot that was saved by the goalkeeper
Total left -footed shots
Attempt that misses - over the crossbar
Attempt that misses - over the crossbar and left
Attempt that misses - over the crossbar and right
Attempt that misses - to the left
Attempt that misses - to the right
Shot from outside the box that is blocked before reaching the goal (excludes any shots saved/cleared off the line in a defensive act – this is not the same as a regular goalkeeper save on the line)
Goal from a shot outside the box
Shot from outside the box that goes wide of or over the goal
Shot from outside the box that hits the post or bar
Attempt from outside the box to the left side of the goal
Attempt from outside the box to the right side of the goal
Shot from outside the box that was saved by the goalkeeper (i.e. not blocked by an outfield player on the line)
Attempt from deep outside the box to the left side of the goal
Attempt from deep outside the box to the right side of the goal
Goal scored with other body part

Attempt from long range to the centre of the goal
Attempt from long range to the left of the goal
Attempt from long range to the right of the goal
A shot attempt where the attacker was in a one-on-one situation against the goal keeper
An attempt that missed that was close but high
An attempt that missed that was close but high right
An attempt that missed that was close but high left
An attempt that missed that was close but left
An attempt that missed that was close but right
A shot attempt that came in open play or in a fast break
Penalty goal
A penalty shot that goes wide of the goal or over the crossbar
A penalty that hits the woodwork
A penalty that was saved by the goalkeeper
A shot that hits the crossbar
A shot that hits the left post
A shot that hits the right post
Right-footed goals
Right-footed shots that were saved by the goalkeeper (i.e. not blocked by an outfield player on the line)
Total right-footed shots
A shot that came directly after a set-piece (corner, free kick or throw-in)
Shot that is saved in the top centre part of the goal (i.e. not blocked by an outfield player on the line)

A shot that is saved in the top left of the goal (i.e. not blocked by an outfield player on the line)
A shot that is saved in the top right of the goal (i.e. not blocked by an outfield player on the line)
A shot that is saved in the bottom centre part of the goal (i.e. not blocked by an outfield player on the line)
A shot that is saved in the low, left part of the goal (i.e. not blocked by an outfield player on the line)
A shot that is saved in the low, right part of the goal (i.e. not blocked by an outfield player on the line)
An attempt conceded from a shot from inside the box
An attempt conceded from a shot from outside the box
An attempted shot from inside the box
An attempted shot from outside the box
Free kick for passing back to goalkeeper
When a player takes possession of a loose ball
Shot blocked (excludes any shots saved/cleared off the line in a defensive act, which is not the same as a regular goalkeeper save on the line)
Challenge that did not make contact, i.e. a missed challenge where the player was dribbled past (therefore there cannot be a challenge_won metric)
No goals conceded in the game (player must play 90 minutes)
Clearance/blocked shot off the line
Any major talking point or error made by the referee
Player takes a corner
Goalkeeper tries to catch a cross but misses the ball
A cross caught by the goalkeeper when delivered from within 18 yards from the by-line.

A cross caught by the goalkeeper when delivered from further than 18 yards from the by-line.
Foul for play which could cause serious injury to an opponent
A goal scored by a defender
Player is dispossessed on the ball by an opponent – no dribble involved
Goalkeeper dives and catches the shot
Goalkeeper dives and parries/deflects the ball to a safe area
Duel over the possession of the ball where a player loses the ball
Duel over the possession of the ball where a player wins the ball
A successful defensive clearance – where a player under pressure kicks the ball clear of the defensive zone or/and out of play
A successful defensive clearance via a header – where a player under pressure heads the ball clear of the defensive zone or/and out of play
A mistake which leads to the opposition scoring
A mistake which leads to an opposition shot
A pass where the ball moves from outside the final third into the final third
Free kicks won when fouled or as a result of dangerous play by the opposition (does not include handball or penalties)
Free kicks conceded to the opposition via fouls, handballs, dangerous play, 6-second violations, or back passes (includes penalties)
Goals scored by strikers
Fouls committed
A foul that occurs in the final third of the pitch (in relation to the attacking player's final third, i.e. the attacking third)
Kick Off – game starts
Goalkeeper successfully takes possession by diving down to collect a loose ball that a striker is chasing
Player assists a goal by passing the ball to the player who scored the goal

Qualifier for an assist to confirm that the assisting player had a direct intention to set up a goal attempt
Goalkeeper restarts play with a goal kick
Goal scored
Goal conceded
Goal conceded from a shot inside the box
Goal conceded from a shot outside the box
Goalkeeper catches a cross
goals scored from regular play or on a fast break
Free kick given for handball (included in the foul lost statistic but not foul won)
Clearance via a header
Pass from a header (this statistic always implies a successful headed pass)
A defending player intercepts a pass between opposition players
Interception made within the penalty area
Goalkeeper picks the ball up - usually under pressure
Goalkeeper throw-outs
Player successfully beat the last man in a dribble
A player makes a defensive action/tackle and is the last person between the opponent and the goal
Long pass made from a player's own half into the opposition's half
Successful long pass from a player's own half into the opposition's half
Corner conceded
Minutes played by player
Goals scored by midfielders

Awarded to the last defender when an offside decision is given
Player assists another player, who takes a shot but misses the goal
Player assists another player, who takes a shot that is on target (includes goals)
Shot on target
Defender blocks a shot
Own goal conceded
Pass to the left wing (attacking half)
Pass to the right wing (attacking half)
Penalty goal conceded (and scored) against the team
Penalty awarded against the team in question (including handballs)
Penalty save made
Player fouled within the penalty box (excludes any penalties won for handball)
The percentage of overall ball possession the given team had during the game
Shot that hits the post or bar
Goalkeeper punches the ball clear from a high ball
Red card
Shot saved from shot inside the box
Shot saved from shot outside the box
Shot saved from a direct free kick
Total goalkeeper saves
Second yellow card given
A pass to create an opportunity for another player to assist a goal

Shot that goes wide of or over the goal, or hits the post/bar
Foul conceded by goalkeeper holding the ball beyond the six-second rule
A shot blocked from an attempt inside the six-yard box
Goalkeeper saves a shot by standing and catching
Goalkeeper saves a shot by standing and deflecting/parrying
Number of substitutions made
Player assists a shot (including goals). Also known as 'chances created' or key passes.
Total number of passes that end in the player's own half – excluding throw-ins and keeper throws
Total number of clearances
A dribble past a player attempted
Total corners that reached the box
Total number of crosses
Total number of crosses that are not from corners
Total number of fast breaks that occurred
Total number of passes that end in the opposition half (includes crosses).
Total number of high claims by goalkeeper
Number of long balls launched forward into an area of the pitch rather than to a specific team-mate
Total passes by a striker who has received the ball with his back to goal and then plays the ball back towards team-mates
Total passes longer than 35 yards
Total offsides
An aggregate of all attempted (successful or unsuccessful) passes excluding throw-ins, keeper throws and crosses.
Total shots at goal

Number of players substituted off
Number of players substituted on
Total number of tackles
Total number of through-balls.
Total number of throw-ins
Total number of yellow cards awarded to the team overall
Total number of red cards awarded to the team overall
Total sum of a team's on-the-ball events
Previously collected as a standalone stat, this is now added as a qualifier to events where possession is lost
Number of fouls on the player
Total dribbles where a team player beats an opponent – no over-runs
Total corners forced by the team
Total tackles won
Yellow cards awarded to the team
Total number of passes (successful or unsuccessful) which are “flicked” on to a running team-mate - usually a header
Total number of passes which are “flicked” on and successfully find a team-mate – usually a header
Total number of passes (successful or unsuccessful) which are lofted into the air and not along the ground – does not include crosses
Total number of successful passes played in the air and not along the ground – does not include crosses
Total number of opposition crosses blocked by the team
A defender shields the ball with his body from an opponent as the ball rolls out of play
Throw-in taken incorrectly
Total number of opposition crosses blocked by the team resulting in possession being won
Goalkeeper faced a penalty kick
An attacking player reaches the byline and passes the ball in a backwards direction (successful or unsuccessful)



An attacking player reaches the byline and successfully passes the ball in a backwards direction to a team mate
Goalkeeper attempts to come off his line and win possession of the ball
Goalkeeper comes off his line, wins the ball and wins possession for his team
A goal assist from an open play situation
A goal assist from a corner, free kick or throw-in, where the assist itself may be several passes after the set play in question.
A shot assist from an open play situation
A shot assist from a corner, freekick or throw-in, where the assist itself may be several passes after the set play in question
A player attempts a dribble but hits the ball too far ahead and loses it
Interception where the player wins and retains possession of the ball
A pass which led to a clear-cut scoring opportunity e.g. one-on-one situation or a shot from just a few yards out
A clear-cut scoring opportunity which was not converted / scored
A goals scored from a clear-cut chance
A player only touched the ball and lost possession – bad control
A pass forward
A pass backward
A pass to a player on the left hand side of the pitch in the opposition half (excludes throw-ins, keeper throws and crosses)
A pass to a player on the right hand side of the pitch in the opposition half
Total number of successful passes where the ball ends in the final third of the pitch
Total number of passes (successful or unsuccessful) where ball ends in the final third of the pitch
Red card rescinded by the referee
Combination of ALL diving saves and catches, including saves where the ball ends up still in danger in the possession of an attacking player
Total number of times possession was regained in the defensive third of the field

Total number of times possession was regained in the midfield third of the field
Total number of times possession was regained in the attacking third of the field
Total losses of possession
Goals scored from a fast break situation
Shots made following a fast break situation
A pass where the ball moves from outside to inside the penalty area
Total passes which end within the attacking third
Total number of occasions the ball hit the woodwork in any situation
Assist directly from a corner or free kick
A ball played into the box directly from a free kick (successful or unsuccessful)
A ball played into the box directly from a free kick that successfully found a team-mate
Any type of pass in open play (successful or unsuccessful)
Any type of pass to a team-mate in open play
A foul event with "attempted tackles" qualifier assigned
A player blocks a pass at close range
An attacking player who plays a pass/cross/shot through an attempted block
A heavily deflected pass
A shot blocked by a defender and scored on the rebound
A shot saved by the goalkeeper but scored on rebound
A shot that hits the woodwork and is scored on rebound
A foul won that leads to a direct free kick scored
A player forces a handball which results in a free kick scored
A player shoots or passes, forcing a defender to put the ball in his own goal
An assist assigned to a player who wins a penalty which is taken and scored by a team-mate
Points dropped from winning positions
Points gained from losing positions
Total touches inside the opposition's penalty area
The position within the formation

