Developing an English and Spanish Football League Result Predictor using Excel.

INDEX

Contents

Abstract.....3

CHAPTER 1: Introduction.....4

CHAPTER 2: Problem Statement.....9

CHAPTER 3: Literature Review.....11

CHAPTER 4: Objective.....13

CHAPTER 5: Methodology.....15

CHAPTER 6: Implementation and Analytics.....32

CHAPTER 7: Conclusions.....44

CHAPTER 8: Future development of the model.....45

Abstract

The field of sports analytics has gained significant attention in recent years, driven by the growing demand for accurate forecasts and insights into the outcomes of different sporting events. This project is focused on developing a football prediction system where advanced statistical analysis and predictive modelling approaches are used.

The project intends to provide forecasts that can assist football aficionados, professionals, and bettors in making informed selections by leveraging historical match data, team performance statistics, and other pertinent aspects.

To determine match outcomes, the project takes a statistics-driven approach, employing Poisson Distribution and statistics. Goals scored, goals conceded, Expected Goals (xG), Expected Assists (xA), Home Form, Away Form, and historical records are among these criteria. The project tries to find patterns and trends that lead to effective forecasts by accumulating, pre-processing, and analysing massive amounts of football data.

The generated prediction model will be rigorously evaluated and validated using an extensive number of performance metrics, in addition to back and forward testing.

Furthermore, the project recognises the necessity of interpretability and candour for prediction systems, and attempts will be made to provide pertinent explanations behind the predictions generated.

This project is expected to produce a dependable and user-friendly football prediction system that can be used by football fans, analysts, and betting specialists.

By harnessing the power of data analysis and predictive modelling, this project aims to enhance the understanding and prediction of football match outcomes, ultimately contributing to the advancement of football analytics and decision-making processes in the sports industry.

CHAPTER 1: Introduction

1.1 European Football

European football is a rich tapestry of leagues, with every league having its own distinct style and devoted fan base. Along with the English Premier League, four other important leagues dominate the European football landscape: **La Liga** in Spain, **Serie A** in Italy, **Ligue 1** in France, and the **Bundesliga** in Germany. These leagues feature some of the world's most gifted players and provide an enthralling blend of skill, strategies, and emotion.

In terms of global viewership, the **English Premier League** is often regarded as the leading force. It is known for its fierce competition and is the most difficult league in the world, with enormous financial resources. Famous clubs such as Manchester United, Manchester City, Arsenal FC, Chelsea FC, Liverpool FC, and Tottenham Hotspur draw millions of viewers each year, contributing to the league's unequalled popularity.

The league averages 643 million viewers per game and has a global broadcast audience of more than 3.2 billion people.

The Spanish **La Liga** is known for its technical brilliance and attacking flair, and it is home to prestigious clubs such as Barcelona and Real Madrid. La Liga, with its star-studded rosters and strong rivalries, attracts a sizable global audience, with an **estimated viewership of over 2 billion people**. The league's emphasis on exquisite passing, imaginative playmaking, and mesmerising goal scoring feats has cemented its position as one of the world's most watched football championships.

The **Italian Serie A** is noted for its technical brilliance and defensive strength, and it has storied teams such as Juventus, AC Milan, and Inter Milan. The league has a long history and a reputation for producing excellent defenders and tactical masterminds. With an **estimated global audience of over 1 billion people**, Serie A delivers a riveting combination of sophisticated football, strong rivalries, and fanatical supporters, making it a spectacle in its own right.

The **French Ligue 1** league, which includes PSG, Marseille, and Lyon, mixes flair, pace, and young energy. The league's offensive style of play and steady inflow of fresh talent has earned it a large domestic and international following. **Ligue 1 has a projected global audience of about 900 million people**, with fascinating matches and rising players catching the attention of football fans worldwide.

The **German Bundesliga** is renowned for its high-intensity matches and constant pressing, and it is home to powerhouse clubs like as Bayern Munich and Borussia Dortmund. The Bundesliga, known for its colourful atmosphere and passionate supporters, has an **estimated global audience of over 1 billion people**. The league's dedication to developing young players and encouraging

a fast-paced, attacking style of football has earned it a reputation as one of the world's most thrilling football leagues.

The purpose of the foregoing introduction is to showcase the sport's global popularity, which has grown and continues to develop throughout the years.

1.2 English Premier League

English football, also known as 'soccer' (only in the United States), is dear to the hearts of millions of people in the United Kingdom and many more throughout the world. It is strongly ingrained in the nation's culture and heritage, with a history extending back over a century. The sport is played at many different levels, ranging from grassroots amateur leagues to professional divisions. English football is recognised for its fervent supporters, magnificent stadiums, and heated club rivalries. The sport acts as a uniting factor, bringing people from all walks of life together to cheer on their favourite teams and participate in the excitement of the game.

The English Premier League (EPL) is the highest level of English football. The EPL, which began in 1992, marks the highest level of competition in the country. It consists of 20 teams that compete over the course of a season, playing a total of 38 matches apiece. The league draws great talent from all around the world, making it one of the most competitive and exciting football leagues in the world. The English Premier League captivates spectators

both locally and worldwide with its fast-paced action, skilled players, and unpredictable outcomes, producing strong support and intense enthusiasm.

1.3 Betting and Odds in Football

Considering the worldwide appeal, it is no surprise that Betting has seamlessly integrated itself into the football culture, serving as an essential component of the fan experience. By introducing an additional level of thrill and involvement to the sport, it enables enthusiasts to put their expertise and instincts to the test, as they make predictions about match results and various other factors.

Betting on football involves placing wagers on various aspects, such as

i. The final result,
ii. number of goals scored, or
iii. individual player performances.

Bookmakers determine the ***odds*** for each possible outcome**, reflecting the probability or likelihood of its occurrence**.

"The odds provided by bookmakers serve as a guide for bettors, indicating the potential returns they can expect if their predictions prove correct."

Odds

Odds are calculated based on a range of factors, including

i. team form, player availability,
ii. head-to-head records, and
iii. other statistical indicators.

To estimate the probability and establish the odds, bookmakers use complex algorithms and analysis. There are many sorts of bets available, providing fans with a variety of betting options. From simple win-draw-lose bets to

more sophisticated accumulators and handicap betting, the world of football betting offers a diverse range of possibilities to suit a wide range of tastes.

1.4 Fantasy Premier League

The Fantasy Premier League (FPL) is a virtual tournament in which football fans may manage their own clubs. Participants are given a budget and are charged with assembling fantasy squads of real-life players from the English Premier League.

Managers must make strategic decisions during the season, such as selecting the starting lineup, making transfers, and adopting tactical alterations. Points are provided depending on the performance of the chosen players in actual matches, with goals, assists, clean sheets, and other metrics all adding to the overall point total.

The FPL has grown in popularity among football enthusiasts, with over 11,000,000 people registered for the 2022 - 2023 season, providing a new opportunity to participate with the sport and compete against friends, colleagues, and other participants. Individuals can demonstrate their football knowledge, managerial abilities, and ability to recognise prospective breakthrough players. As managers actively monitor the performances of their selected players and make smart decisions to maximise their points, the game encourages interaction and healthy competition. The Fantasy Premier League has grown into a big phenomenon, spawning its own community, conversations, and even prize tournaments, all of which contribute to the whole football experience.

This chapter establishes the groundwork for the succeeding chapters by offering an in-depth explanation of English football, including the significance of the English Premier League, the role of betting and odds, and the birth of the Fantasy Premier League. We will go into the construction of

a football prediction system in the following parts, investigating its fundamental methodologies, data processing approaches, and prospective applications.

CHAPTER 2: Problem Statement

The subject addressed in this study is the unpredictability of football match outcomes and the difficulty of effectively forecasting them, especially when fans are financially and emotionally invested. Given that the majority of these people may make such selections based on personal prejudices and preferences, we might conclude that the decisions are decided on whim rather than organised analytical thinking.

Football is a very dynamic sport that is impacted by a variety of elements such as team form, individual performances, tactical plans, injuries, and external situations. This unpredictability presents a big difficulty for football fans, sports analysts, and bettors who want to make informed predictions and judgements based on match outcomes.

While football outcome predictors abound, a congested market with excessive membership fees limits fans' capacity to engage in betting activities with confidence and make informed decisions about their fantasy football teams.

Furthermore, the growing popularity of football, as well as the advent of many football leagues at both the domestic and international levels, has broadened the scope and complexity of prediction attempts. Each league has its own distinct set of traits, team dynamics, and playing styles, making it even more challenging to construct a complete prediction model that takes these subtleties into consideration.

As a result, there is a need not only to develop an effective and reliable football result predictor that uses advanced analytics techniques, machine learning algorithms, and real-time data to generate accurate predictions for football match outcomes, but also to educate participants and make such a tool available.

A predictor of this type would allow football fans, bettors, and fantasy football participants to make better educated judgements, increase their involvement with the sport, and increase their chances of winning betting and fantasy football tournaments.

Addressing these challenges and developing a reliable football result predictor necessitates a thorough understanding of the factors influencing match outcomes, analysis of historical data, identification of relevant features, selection of appropriate machine learning algorithms, and integration of real-time data sources.

While the model I constructed is fairly basic, it provides the groundwork for a more advanced system in the future that will be more accurate than it is now.

This project seeks to contribute to the area of sports analytics by producing an accurate and trustworthy predictor and providing significant insights and forecasts for football aficionados, bookmakers, bettors, and fantasy football participants.

CHAPTER 3: Literature Review

- Statistical football prediction has been a topic of discussion since the 1990s. *Moroney* produced one of the earliest known models in *1956* [2], finding that both the Poisson distribution and the negative binomial distribution effectively suited the results of football games.
  This laid the foundation for further research in this area.
- *Reep and Benjamin* used the negative binomial distribution to analyse the sequence of ball passing between players during football matches in *1968*[3]. They observed that football had patterns that could be somewhat predicted, disputing the concept that it was entirely random [4]. This realisation paved the way for the development of more accurate prediction models.
- In 1982, *Michael Maher* pioneered a model that predicted the outcomes of football matches between teams of varying ability levels [5]. The goals scored by opponents throughout a game, according to Maher's model, followed a Poisson distribution. The parameters of the model were set

based on the gap between offensive and defensive abilities, with modifications made for home field advantage. This model offered a good foundation for studying the elements that influence match results.

- The importance of home field advantage in football forecasts was extensively examined in a *1992 essay by Cornea and Carron* [6]. They summarised the modelling methodologies for this feature, providing light on its importance in effectively forecasting match results.
- In *1999, Knorr-Held* investigated the time-dependence of team strengths and proposed a more realistic way to evaluating football teams based on recursive Bayesian estimate [7]. When compared to projections based on standard average data, this strategy provided a more detailed knowledge of football clubs' shifting dynamics and capabilities.

These studies demonstrate the history of football prediction models, beginning with Moroney's fundamental work and progressing to the addition of home field advantage and time-dependency in more contemporary systems. This study intends to construct a complete and accurate football prediction model by combining findings from several investigations.

CHAPTER 4: Objective

- **To accurately predict the outcomes of football matches:** The model's primary goal is to create accurate predictions for the outcomes of football matches. The model seeks to produce trustworthy forecasts of match results by utilising statistical approaches and considering pertinent aspects such as team performance, team statistics, and historical data.
- **To analyse and incorporate key influencing factors:** The model aims to discover and include important influencing factors that influence the result of football matches. Team form, home field advantage, and other pertinent considerations will be among these. By taking these characteristics into account, the model hopes to improve the accuracy of its predictions.
- **To assess team strengths and weaknesses:** The approach tries to evaluate football teams' relative strengths and weaknesses. The model seeks to give insights into the strengths and weaknesses of teams by analysing team and player performance measures, historical data, and other pertinent factors, allowing for more educated forecasts.

- **To identify value betting opportunities:** The methodology may also be used to discover value betting opportunities. The algorithm tries to identify matches where the odds undervalue specific teams or outcomes by analysing the odds supplied by bookmakers and comparing them to the estimated probability of match outcomes, offering possible chances for lucrative betting.
- **To provide actionable insights for fantasy premier league:** The model's goal is to give players in fantasy premier leagues with actionable knowledge. The algorithm attempts to give recommendations for team selection, captain selections, and transfer decisions by analysing player statistics, form, and future games, therefore improving the success of fantasy premier league managers.
- **To continuously improve and update the model:** The goal is to create a model that can adapt and improve over time. The goal is to constantly improve the accuracy and reliability of the prediction model by adding new data, optimising algorithms, and analysing model performance, while remaining up to speed with developing trends and dynamics in football.

CHAPTER 5: Methodology

5.1 OVERVIEW

The scope of this model can be categorized into two distinct types:

i) Goals Model

ii) Expected Goals (xG) Model

Before, delving deeper into the models let us take a moment to explore Goals, the xG model and the distinctions between them.

In football, the term "Goals" represents the cumulative number of times a team successfully scores by getting the ball across the goal line and into the opposing team's net. It is the tangible outcome of a scoring effort and directly contributes to the final score of a match, ultimately determining the winner or loser.

On the other hand, "Expected Goals" (often abbreviated as xG) is a statistical metric that evaluates the quality or probability of a scoring chance in football. It serves as a predictive value, assessing the likelihood of a shot resulting in a goal based on various factors. These factors include shot location, angle, distance, assist type, and other contextual variables such as defensive actions and the use of a dominant or weak foot while shooting.

Expected Goals are derived from the analysis of extensive historical data using complex models and algorithms. These models assign a numerical value to each scoring opportunity, indicating the likelihood of it resulting in a goal.

Greater xG values correspond to increased scoring probabilities, whereas lower xG values indicate diminished chances of scoring.

The differentiation between Goals and Expected Goals lies in their interpretation and purpose. Expected Goals provide a statistical estimate of the probability of scoring based on the characteristics of the shots taken. On the other hand, Goals represent the actual outcome of scoring during a match. Expected Goals are valuable for assessing the quality of scoring opportunities and a team's attacking performance, regardless of whether those opportunities were converted into goals.

While Goals reflect the visible results displayed on the scoreboard, Expected Goals offer deeper insights into a team's offensive performance and efficiency. By comparing Goals and Expected Goals, analysts and coaches can gain a better understanding of a team's ability to convert scoring opportunities and identify areas for improvement.

In my model, I utilized both the Goals and Expected Goals techniques to evaluate their performance for the Premier League Season 2022-2023.

5.2 TOOL

The primary tool utilized in constructing this model is Excel, a robust spreadsheet software extensively employed for data analysis, calculation, and presentation. Excel offers a wide range of features and capabilities that enable users to efficiently organize, manipulate, and visualize data. When implementing a football prediction model, Excel proves invaluable due to its numerous tools and functions specifically tailored to enhance the process.

1. **VLOOKUP:** In Excel, the VLOOKUP (Vertical Lookup) function enables users to search for a particular value in a vertical column and retrieve corresponding information from another column. This functionality proves highly advantageous when handling extensive datasets, such as football match data, as it allows for the retrieval of specific details like team names, scores, or pertinent statistics. By leveraging VLOOKUP, you can effectively acquire and integrate the required data into your prediction model, streamlining the process.
2. **Conditional Formatting:** Excel's Conditional Formatting is a powerful tool that empowers users to emphasize cells or ranges based on predefined conditions or criteria. This feature proves immensely useful when visualizing and analyzing football data, as it enables the application of formatting rules to highlight patterns, trends, or specific outcomes. For instance, you can utilize Conditional Formatting to emphasize winning teams, identify matches with a high potential for scoring, or highlight notable deviations from expected results. By harnessing the capabilities of Conditional Formatting, you can extract valuable insights from your data and make informed decisions when developing your football prediction model.

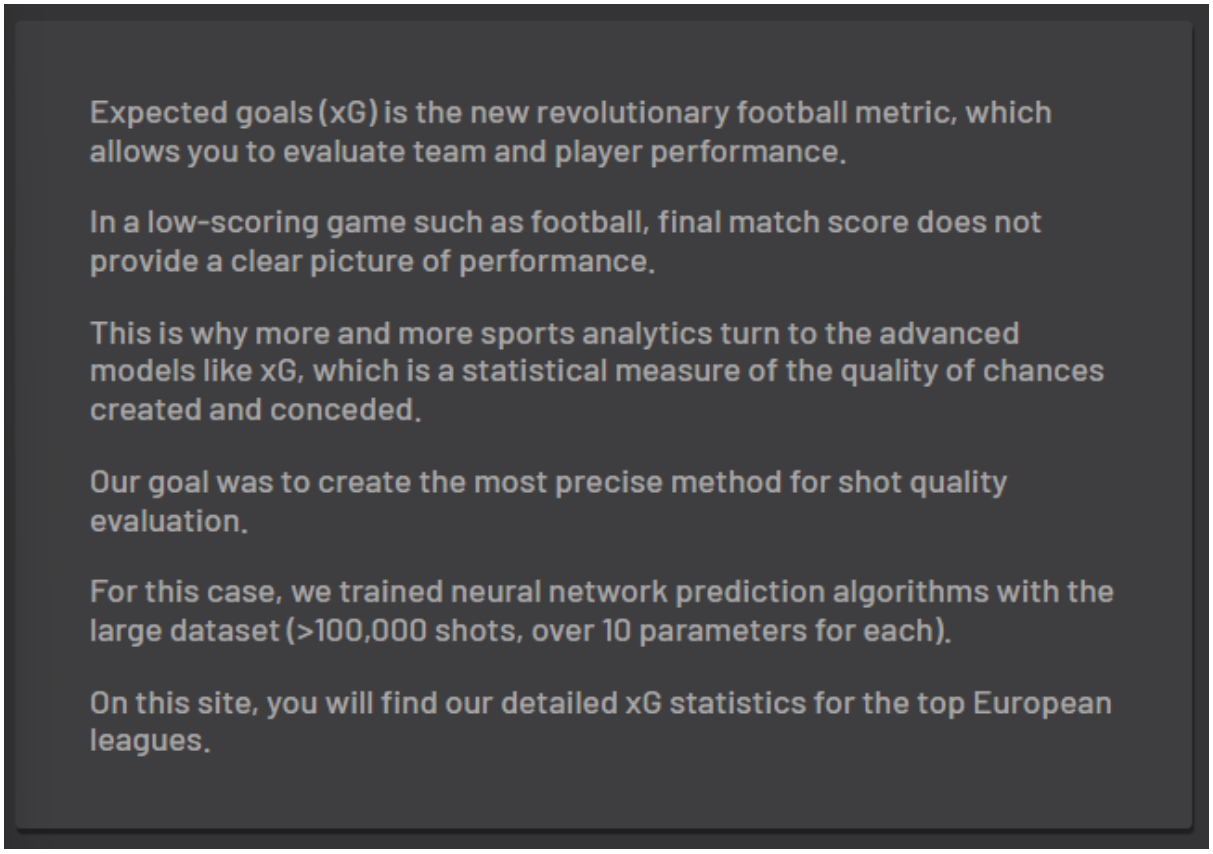
3. **Data Validation:** Excel's Data Validation is a functionality that aids in guaranteeing the precision and consistency of data by imposing validation rules or constraints on input values. When constructing a football prediction model, data validation can be employed to establish specific criteria for entering match data, including team names, scores, or other pertinent variables. By defining validation rules, such as numeric ranges, text lengths, or pre-existing lists, you can reduce errors and ensure the validity and reliability of the data integrated into your model. The implementation of data validation enhances the overall integrity and dependability of your football prediction model.

To summarize, Excel is a flexible software offering a multitude of robust features that facilitate the implementation of a football prediction model. Through the utilization of tools such as VLOOKUP, Conditional Formatting, and Data Validation, you can effectively organize, analyze, and verify the required data, thereby enhancing the precision and efficacy of your model.

5.3 METHOD

The forecast relies on a mathematical concept and utilizes freely accessible data from public websites. Specifically, I obtained the data from understat.com, which I selected due to its convenient accessibility and user-friendly website interface. This platform provided comprehensive football game statistics for the entire

season, allowing me to easily access and gather the necessary information.



(Via.understat.com)

With the exception of xG, all the statistics found on this website can be obtained from various sports websites, platforms, or applications since they represent the tangible events that occur during a football game, including Goals Scored, Saves, Shots Taken, and more. However, it is important to acknowledge that the calculation of xG may vary across platforms due to differences in methodologies. In the case of understat.com, they employed a neural network prediction algorithm trained with a dataset of 100,000 individual shots, utilizing 10 parameters for each shot, as illustrated in the accompanying graphic. It is important to note that my xG model is derived from understat.com's interpretation of xG.

5.4 DATA

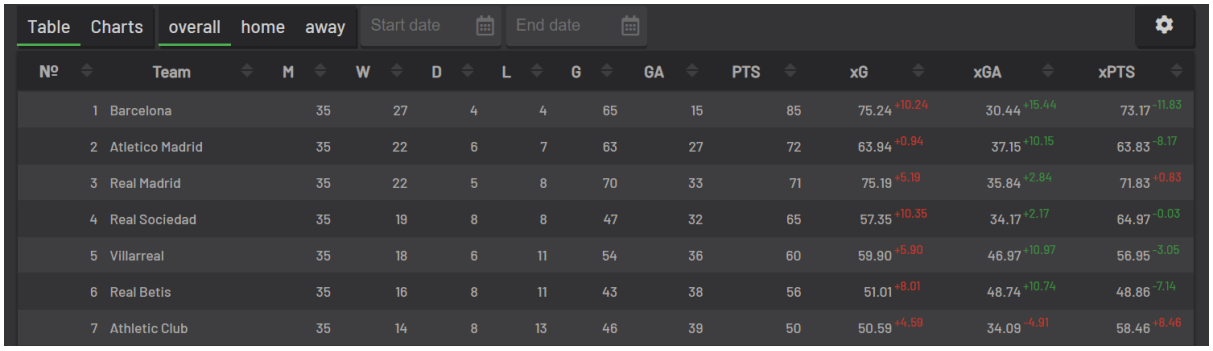


Table Charts overall home away Start date End date

№	Team	M	W	D	L	G	GA	PTS	xG	xGA	xPTS
1	Barcelona	35	27	4	4	65	15	85	75.24 +10.24	30.44 +15.44	73.17 -11.83
2	Atletico Madrid	35	22	6	7	63	27	72	63.94 +0.94	37.15 +10.15	63.83 -8.17
3	Real Madrid	35	22	5	8	70	33	71	75.19 +5.19	35.84 +2.84	71.83 +0.83
4	Real Sociedad	35	19	8	8	47	32	65	57.35 +10.35	34.17 +2.17	64.97 -0.03
5	Villarreal	35	18	6	11	54	36	60	59.90 +5.90	46.97 +10.97	56.95 -3.05
6	Real Betis	35	16	8	11	43	38	56	51.01 +8.01	48.74 +10.74	48.86 -7.14
7	Athletic Club	35	14	8	13	46	39	50	50.59 +4.59	34.09 -4.91	58.46 +8.46

The data is quite simple and in fact we need very few data points to make our predictions. In this table we can see that we can we all the teams and how many games they have played (M), the number of games they have won (W), games they have drawn (D), the games they have lost (L). Further we can see the stats that we are the most concerned with, the number of goals scored by the team (G), the number of goals they have conceded (GA). G and GA will go into helping us predicting using the Goals model. We need not concern ourselves with the points the team has earned (PTS) or the expected points that the team may earn (xPTS). Expected Goals (xG) and expected goals against (xGA) are important for us to consider in terms of our Expected Goals model.

5.5 HOME AND AWAY

Generally, in any league format of football every team in the league must face every other team twice in the entire season, these games are played once on a team's own home stadium (Home) and once on the opponent's home stadium

(Away). In the Spanish La Liga there are 20 teams that means every team must face the other 19 teams twice in the season. 19 games will be played at their own home turf and 19 games on all the opponent's home turfs.

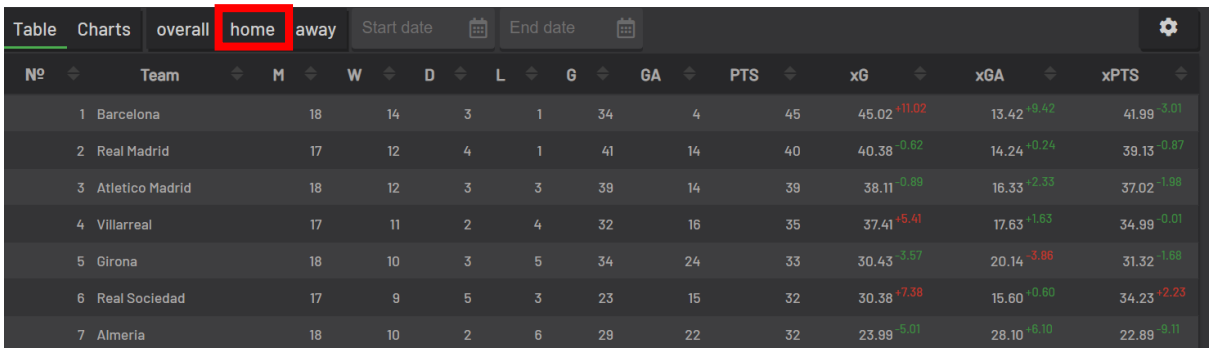


Table | Charts | overall | home | away | Start date | End date

№	Team	M	W	D	L	G	GA	PTS	xG	xGA	xPTS
1	Barcelona	18	14	3	1	34	4	45	45.02 +11.02	13.42 +9.42	41.99 -3.01
2	Real Madrid	17	12	4	1	41	14	40	40.38 -0.62	14.24 +0.24	39.13 -0.87
3	Atletico Madrid	18	12	3	3	39	14	39	38.11 -0.89	16.33 +2.33	37.02 -1.98
4	Villarreal	17	11	2	4	32	16	35	37.41 +5.41	17.63 +1.63	34.99 -0.01
5	Girona	18	10	3	5	34	24	33	30.43 -3.57	20.14 -3.86	31.32 -1.68
6	Real Sociedad	17	9	5	3	23	15	32	30.38 +7.38	15.60 +0.60	34.23 +2.23
7	Almeria	18	10	2	6	29	22	32	23.99 -5.01	28.10 +6.10	22.89 -9.11

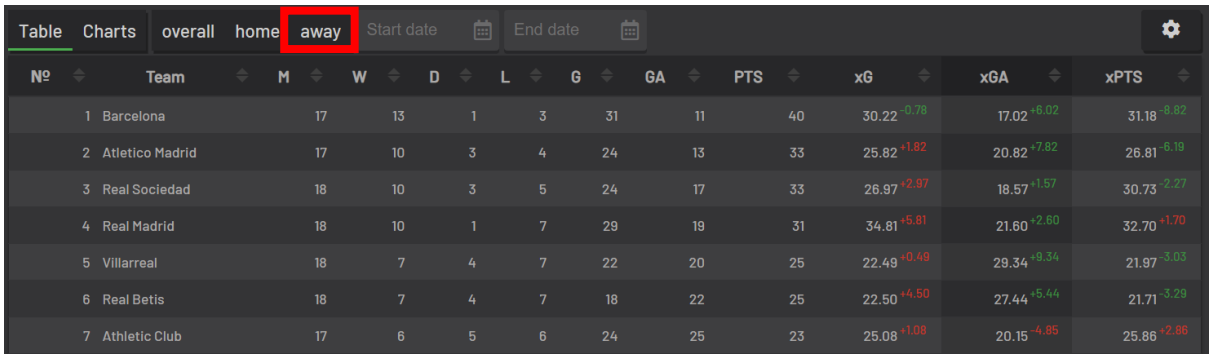


Table | Charts | overall | home | away | Start date | End date

№	Team	M	W	D	L	G	GA	PTS	xG	xGA	xPTS
1	Barcelona	17	13	1	3	31	11	40	30.22 -0.78	17.02 +6.02	31.18 -8.82
2	Atletico Madrid	17	10	3	4	24	13	33	25.82 +1.82	20.82 +7.82	26.81 -6.19
3	Real Sociedad	18	10	3	5	24	17	33	26.97 +2.97	18.57 +1.57	30.73 -2.27
4	Real Madrid	18	10	1	7	29	19	31	34.81 +5.81	21.60 +2.60	32.70 +1.70
5	Villarreal	18	7	4	7	22	20	25	22.49 +0.49	29.34 +9.34	21.97 -3.03
6	Real Betis	18	7	4	7	18	22	25	22.50 +4.50	27.44 +5.44	21.71 -3.29
7	Athletic Club	17	6	5	6	24	25	23	25.08 +1.08	20.15 -4.85	25.86 +2.86

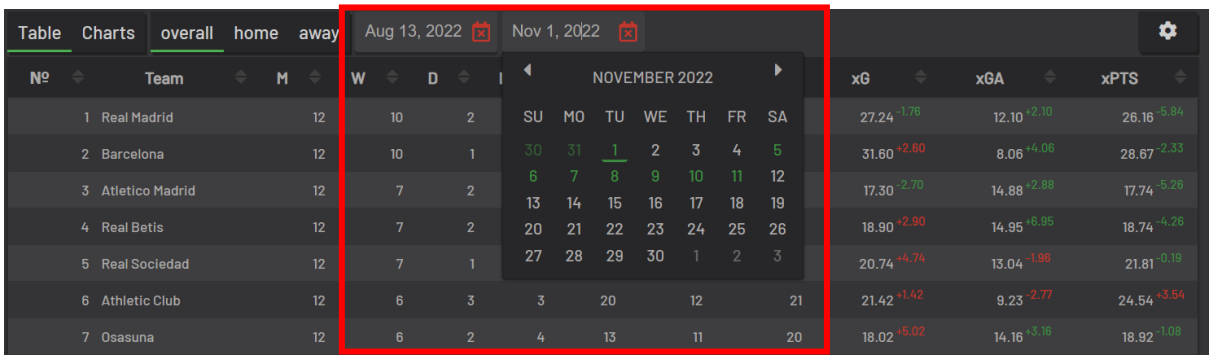
Here we can see clearly see that understat.com allows us to easily split a team's statistics based on how they performed at Home and how they performed Away. Home and away are important features that seem to have an influence on the results of a game as noted in *Michael Maher's paper*. [5]

5.6 GAMEWEEKS

As already discussed, every team must play 38 games in the entirety of a season. The season goes on from August till the following year's May month. We must note that it takes 2 teams to play a game therefore with 20 teams in the

league there are 10 games in one gameweek. The first gameweek is classified as Gameweek 1 and it goes on till Gameweek 38. This makes it 380 matches played in an entire season.

In order to back test, the data I was able to use understat's calendar feature where I was able to take all the data chronologically. For example, if I wanted to predict the results of Gameweek 13, I had the choice to only retrieve the data from GW1 – GW12. This selection criteria can be seen in the image below.



5.7 METRICS USED

Both the Goals model and the xG models have the same methodology the only difference is the goals model considers G and GA (Actual GS) as the input, whereas, the xG model considers the xG and xGA (Expected GS) model as input.

GS/pg. home and GA/pg. home

	HOME									
№	Team	M	W	D	L	G	GA	PTS	GS/pg	GC/pg
1	Barcelona	12	10	2	0	25	1	32	=T3/P3	0.0833333
2	Real Madrid	12	8	4	0	26	9	28	2.1666667	0.75
3	Villarreal	12	7	2	3	17	10	23	1.4166667	0.8333333

	AWAY									
№	Team	M	W	D	L	G	GA	PTS	GS/pg	GC/pg
1	Barcelona	13	11	0	2	22	7	33	=AF3/AB3	0.5384615
2	Real Madrid	13	9	1	3	24	10	28	1.8461538	0.7692308
3	Atletico Madrid	13	8	3	2	17	8	27	1.3076923	0.6153846

The first and foremost metric we need to calculate is the average number of goals scored and goals conceded by a team at their home ground and we will do the same for the team when it is playing away.

Avg. GS/pg. home and Avg. GA/pg. home

	HOME									
№	Team	M	W	D	L	G	GA	PTS	GS/pg	GC/pg
1	Barcelona	12	10	2	0	25	1	32	2.0833333	0.0833333
2	Real Madrid	12	8	4	0	26	9	28	2.1666667	0.75
3	Villarreal	12	7	2	3	17	10	23	1.4166667	0.8333333
4	Almeria	13	7	1	5	20	17	22	1.5384615	1.3076923
5	Atletico Madrid	12	6	3	3	22	11	21	1.8333333	0.9166667
6	Real Betis	12	6	3	3	18	12	21	1.5	1
7	Girona	13	6	3	4	23	18	21	1.7692308	1.3846154
8	Mallorca	13	6	3	4	13	10	21	1	0.7692308
9	Valencia	13	6	2	5	20	12	20	1.5384615	0.9230769
10	Athletic Club	13	6	2	5	18	11	20	1.3846154	0.8461538
11	Real Valladolid	12	6	2	4	11	10	20	0.9166667	0.8333333
12	Celta Vigo	13	5	4	4	19	14	19	1.4615385	1.0769231
13	Rayo Vallecano	12	5	4	3	16	12	19	1.3333333	1
14	Real Sociedad	12	5	4	3	13	12	19	1.0833333	1
15	Osasuna	12	6	1	5	12	11	19	1	0.9166667
16	Sevilla	13	5	3	5	16	16	18	1.2307692	1.2307692
17	Cadiz	13	4	6	3	14	17	18	1.0769231	1.3076923
18	Getafe	13	4	4	5	14	16	16	1.0769231	1.2307692
19	Espanyol	12	3	4	5	16	20	13	1.3333333	1.6666667
20	Elche	13	1	4	8	11	21	7	0.8461538	1.6153846
	Average								=AVERAGE(W3:W22)	

Once we have the initial metrics in place for all teams, the next metric we need to calculate is the average goals scored at home for the entire league and the average goals conceded for all teams at home.

This phase will be repeated for the away statistic, which will provide the average goals scored away from home for the entire league as well as the average goals conceded for all teams away from home.

5.8 VLOOKUP

Total Avg Home gpg	1.38
Total Avg Away gpg	1.03
Avg Home Team gpg scored at Home	1.54
Avg Away Team conceded per game to Home side	0.62
Avg away Team gpg scored against home team	1.31
Avg Home Team gpg conceded to Away side	0.92

After setting up our initial table I used the VLOOKUP feature in excel to filter out 2 statistics for the 2 teams we are trying to match up.

1. Home team

- Avg. goals per game scored at home

- Avg. goals per game conceded at home

2. Away team

- Avg. goals per game scored away from home

- Avg. goals per game conceded away from home

Using the VLOOKUP, we just have to select the home and away team and the statistics will automatically be updated.

5.9 HOME ATTACK and AWAY DEFENCE

$$\text{Home attack} = \frac{\text{Avg. Home Team's goals per game scored at Home}}{\text{Avg. Home goals scored per game}}$$

Home attack is the ratio of average of the home team's goals per game scored at home, against, the average number of goals scored at home by all the teams in the league.

$$\text{Away Defence} = \frac{\text{Avg. Away Team's goals per game conceded at Away}}{\text{Avg. Home goals scored per game}}$$

Away defence is the ratio of average of the away team's goals per game conceded away from home, against, the average number of goals scored at home by all the teams in the league.

5.10 AWAY ATTACK and HOME DEFENCE

$$\text{Away attack} = \frac{\text{Avg. Away Team's goals per game scored at Away}}{\text{Avg. Away goals scored per game}}$$

Away attack is the ratio of average of the away team's goals per game scored at away, against, the average number of goals scored at away by all the teams in the league.

$$\text{Home Defence} = \frac{\text{Avg. Home Team's goals per game conceded at Home}}{\text{Avg. Away goals scored per game}}$$

Home defence is the ratio of average of the away team's goals per game scored at away, against, the average number of goals scored at away by all the teams in the league.

5.12 PREDICTIONS

1) Predicted Home Goals

Predicted Home Goals = Home Attack x Away Defence x Total Avg. of home goals per game

2) Predicted Away Goals

Predicted Away Goals = Away Attack x Home Defence x Total Avg. of away goals per game

3) Predicted Total Goals

Predicted Total Goals = Projected Home Goals + Projected Away Goals

5.13 POISSON DISTRBUTION

The Poisson distribution is a commonly employed probability distribution for simulating the occurrence of events within a specified interval of time or space. This distribution takes its name from Siméon Denis Poisson, a renowned French mathematician who played a significant role in introducing and popularizing the concept during the early 19th century.

The Poisson distribution is a useful tool when analyzing occurrences of rare and independent events.

The probability mass function (PMF) of the Poisson distribution is defined by the following formula:

$P(X = k) = (e^{(-\lambda)} * \lambda^k) / k!$

Where:

- $P(X = k)$ represents the probability of observing k events.
- e is the base of the natural logarithm (approximately equal to 2.71828).
- λ (lambda) is the average rate or intensity of events occurring in the given interval.

- k is the number of events.

In our context, the Poisson distribution is utilized to model the goal-scoring patterns for both the Home Team and the Away Team in a football match. By employing these probabilities, we can construct a matrix that represents the likelihood of each possible scoreline in a game between home and away teams

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probabaility	50.34%	34.55%	11.86%	2.71%	0.47%	0.06%	0.01%	0.00%	0.00%
0	31.14%	15.68%	10.76%	3.69%	0.84%	0.14%	0.02%	0.00%	0.00%	0.00%
1	36.33%	18.29%	12.55%	4.31%	0.99%	0.17%	0.02%	0.00%	0.00%	0.00%
2	21.19%	10.67%	7.32%	2.51%	0.57%	0.10%	0.01%	0.00%	0.00%	0.00%
3	8.24%	4.15%	2.85%	0.98%	0.22%	0.04%	0.01%	0.00%	0.00%	0.00%
4	2.40%	1.21%	0.83%	0.29%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%
5	0.56%	0.28%	0.14%	0.07%	0.04%	0.02%	0.00%	0.00%	0.00%	0.00%
6	0.11%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Initially, we can observe the probability mass function depicting the likelihood of the Home team scoring 0 – 8 goals in a game.

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probabaility	50.34%	34.55%	11.86%	2.71%	0.47%	0.06%	0.01%	0.00%	0.00%
0	31.14%	15.68%	10.76%	3.69%	0.84%	0.14%	0.02%	0.00%	0.00%	0.00%
1	36.33%	18.29%	12.55%	4.31%	0.99%	0.17%	0.02%	0.00%	0.00%	0.00%
2	21.19%	10.67%	7.32%	2.51%	0.57%	0.10%	0.01%	0.00%	0.00%	0.00%
3	8.24%	4.15%	2.85%	0.98%	0.22%	0.04%	0.01%	0.00%	0.00%	0.00%
4	2.40%	1.21%	0.83%	0.29%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%
5	0.56%	0.28%	0.14%	0.07%	0.04%	0.02%	0.00%	0.00%	0.00%	0.00%
6	0.11%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Secondly, we can see the probability of the Mass function of Away team scoring 0 – 8 goals in a game independent of home teams actions.

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probabaility	50.34%	34.55%	11.86%	2.71%	0.47%	0.06%	0.01%	0.00%	0.00%
0	31.14%	15.68%	10.76%	3.69%	0.84%	0.14%	0.02%	0.00%	0.00%	0.00%
1	36.33%	18.29%	12.55%	4.31%	0.99%	0.17%	0.02%	0.00%	0.00%	0.00%
2	21.19%	10.67%	7.32%	2.51%	0.57%	0.10%	0.01%	0.00%	0.00%	0.00%
3	8.24%	4.15%	2.85%	0.98%	0.22%	0.04%	0.01%	0.00%	0.00%	0.00%
4	2.40%	1.21%	0.83%	0.29%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%
5	0.56%	0.28%	0.14%	0.07%	0.04%	0.02%	0.00%	0.00%	0.00%	0.00%
6	0.11%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Thirdly, we turn our attention to the probability of the Away team scoring NO goals and the Home team scoring 0-8 goals.

We can interpret the scores for the Home team as:

Home Goals	Away Goals	Probability
0	0	15.68%
1	0	10.76%

2	0	3.69%
3	0	0.84%
4	0	0.14%

Similarly, we can determine the probabilities from the Point of View of the Away team

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probabaility	50.34%	34.55%	11.86%	2.71%	0.47%	0.06%	0.01%	0.00%	0.00%
0	31.14%	15.68%	10.76%	3.69%	0.84%	0.14%	0.02%	0.00%	0.00%	0.00%
1	36.33%	18.29%	12.55%	4.31%	0.99%	0.17%	0.02%	0.00%	0.00%	0.00%
2	21.19%	10.67%	7.32%	2.51%	0.57%	0.10%	0.01%	0.00%	0.00%	0.00%
3	8.24%	4.15%	2.85%	0.98%	0.22%	0.04%	0.01%	0.00%	0.00%	0.00%
4	2.40%	1.21%	0.83%	0.29%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%
5	0.56%	0.28%	0.14%	0.07%	0.04%	0.02%	0.00%	0.00%	0.00%	0.00%
6	0.11%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

We can interpret the scores for the Away team as:

Home Goals	Away Goals	Probability
0	0	15.68%
0	1	18.29%
0	2	10.67%
0	3	4.15%
0	4	1.21%

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probabaility	50.34%	34.55%	11.86%	2.71%	0.47%	0.06%	0.01%	0.00%	0.00%
0	31.14%	15.68%	10.76%	3.69%	0.84%	0.14%	0.02%	0.00%	0.00%	0.00%
1	36.33%	18.29%	12.55%	4.31%	0.99%	0.17%	0.02%	0.00%	0.00%	0.00%
2	21.19%	10.67%	7.32%	2.51%	0.57%	0.10%	0.01%	0.00%	0.00%	0.00%
3	8.24%	4.15%	2.85%	0.98%	0.22%	0.04%	0.01%	0.00%	0.00%	0.00%
4	2.40%	1.21%	0.83%	0.29%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%
5	0.56%	0.28%	0.14%	0.07%	0.04%	0.02%	0.00%	0.00%	0.00%	0.00%
6	0.11%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

This enables us to determine the probabilities of all the results that are favourable to the home team (In green).

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probabaility	50.34%	34.55%	11.86%	2.71%	0.47%	0.06%	0.01%	0.00%	0.00%
0	31.14%	15.68%	10.76%	3.69%	0.84%	0.14%	0.02%	0.00%	0.00%	0.00%
1	36.33%	18.29%	12.55%	4.31%	0.99%	0.17%	0.02%	0.00%	0.00%	0.00%
2	21.19%	10.67%	7.32%	2.51%	0.57%	0.10%	0.01%	0.00%	0.00%	0.00%
3	8.24%	4.15%	2.85%	0.98%	0.22%	0.04%	0.01%	0.00%	0.00%	0.00%
4	2.40%	1.21%	0.83%	0.29%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%
5	0.56%	0.28%	0.14%	0.07%	0.04%	0.02%	0.00%	0.00%	0.00%	0.00%
6	0.11%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

The probabilities of all the results that are favourable to the away team (In red)

	Home Goals	0	1	2	3	4	5	6	7	8
Away Goals	Probabaility	50.34%	34.55%	11.86%	2.71%	0.47%	0.06%	0.01%	0.00%	0.00%
0	31.14%	15.68%	10.76%	3.69%	0.84%	0.14%	0.02%	0.00%	0.00%	0.00%
1	36.33%	18.29%	12.55%	4.31%	0.99%	0.17%	0.02%	0.00%	0.00%	0.00%
2	21.19%	10.67%	7.32%	2.51%	0.57%	0.10%	0.01%	0.00%	0.00%	0.00%
3	8.24%	4.15%	2.85%	0.98%	0.22%	0.04%	0.01%	0.00%	0.00%	0.00%
4	2.40%	1.21%	0.83%	0.29%	0.07%	0.01%	0.00%	0.00%	0.00%	0.00%
5	0.56%	0.28%	0.14%	0.07%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%
6	0.11%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

The highlighted Yellow Cells indicate the outcomes where both teams score an equal number of goals, resulting in a draw.

By employing the Poisson distribution, it is possible to estimate the probability of observing a specific goal count by considering a team's average scoring rate. For instance, if we know a team's average goals per game, we can utilize the Poisson distribution to calculate the probabilities of scoring 0, 1, 2, or more goals in a given match.

5.14 USER DASHBOARD

Barcelona	Real Madrid
Projected Home Goals 1.16	Projected Away Goals 0.15

	Home Win	Draw	Away Win
% Chance	63.47%	31.84%	4.70%
Implied Odds	1.58	3.14	21.28

Under Goal Markets		
Goals	% Chance	Implied Odds.
0.5	26.97%	3.71
1.5	62.31%	1.60
2.5	85.47%	1.17
3.5	95.59%	1.05
4.5	98.90%	1.01

Over Goal Markets		
Goals	% Chance	Imp Odds.
0.5	73.03%	1.37
1.5	37.69%	2.65
2.5	14.53%	6.88
3.5	4.41%	22.65
4.5	1.10%	90.87

Correct Score	Imp Odds
0-0	3.71
0-1	24.93
0-2	335.37
0-3	6765.99
1-0	3.19
1-1	21.46
1-2	288.68
1-3	5824.17
2-0	5.49
2-1	36.95
2-2	497.00
2-3	10026.88
3-0	14.19
3-1	95.42
3-2	1283.44
3-3	25893.43
Any other home Win	32.75
Any other away Win	21.28
Any other draw	2381766.14

Based on the various metrics and probabilities, I have developed a user-friendly dashboard that allows users to select the Home and Away teams. Once selected, the dashboard provides the following data. The first is the likelihood of different game outcomes, which includes the probability of the Home team winning, the game ending in a draw, or the Away team winning.

In the above case we can see if Barcelona is the Home team and Real Madrid is the Away team there is an 63.47% that Barcelona will win, 31.84% chance that the game will be a draw and only a 4.7% chance that Real Madrid will win.

5.15 IMPLIED ODDS

Implied odds can be defined as the reciprocal of the probability of an event occurring, which is inferred from the odds provided by bookmakers. In the context of football matches, bookmakers assign odds to different outcomes such as a team's victory, a draw, or a loss. These odds reflect the bookmakers' assessment of the likelihood of each respective outcome.

To calculate implied odds, we invert the probability obtained from our Poisson Distribution. In the example mentioned, the probability of the Home team winning is 63.47%. The implied probability can be calculated as 1 divided by 63.47%, resulting in 1.58. Therefore, the implied odds for a Home win are 1.58, for a draw are 3.14, and for an Away win are 21.28. This indicates that the odds are in favor of the Home team winning.

Implied odds are significant for bettors as they help assess the value in a specific betting opportunity. When a bettor believes that the actual probability of an outcome exceeds the calculated implied probability derived from the odds, they may consider it a favorable betting opportunity. This process is commonly known as identifying "value" within the odds.

Bettors often compare the implied odds derived from bookmakers' odds with their own evaluations of the probabilities. If they perceive that the actual probability is higher than the implied probability, they may consider placing a wager on that particular outcome. Conversely, if they believe that the bookmakers' odds undervalue the actual probability, they may view it as a profitable betting opportunity.

Understanding implied odds empowers bettors to make informed decisions and identify potential value bets within football betting markets. This involves evaluating the odds set by bookmakers, calculating the implied probabilities, and then comparing them with one's own predictions or assessments of the actual probabilities for different outcomes in a football match.

5.16 CORRECT SCORE

Now that we have all the metrics in place, we can conveniently view all the possible odds for every possible result in the following table

Correct Score	Imp Odds
0-0	3.71
0-1	24.93
0-2	335.37
0-3	6765.99
1-0	3.19
1-1	21.46
1-2	288.68
1-3	5824.17
2-0	5.49
2-1	36.95
2-2	497.00
2-3	10026.88
3-0	14.19
3-1	95.42
3-2	1283.44
3-3	25893.43
Any other home Win	32.75
Any other away Win	21.28
Any other draw	2381766.14

The green score indicates a home win, the red score represents an away win, and the grey score signifies a draw. Additionally, we can observe the implied odds associated with each outcome. In the example provided, we can conclude that the odds are most favourable for a home team victory with a score of 1-0, which has the lowest odds of 3.19.

It is important to note that any other home win, away win, or draw should only be considered if the "Imp Odds" for any of the three outcomes is lower than the odds associated with the visible scores in the table above.

CHAPTER 6: Implementation and Analytics

6.1 Implementing the model

First let us identify the gameweek for which the results we want to predict in the Spanish La Liga. To give a clear understanding let us pick Gameweek 26 (18th March 2023) which has already occurred at the time of writing this dissertation.

Above are the games and their actual results that took place in Gameweek 26 in the Spanish La Liga.

6.2 Input the Data

The first thing we need to do is input the appropriate data for all the teams before GW26, Saturday 18th March.

We will need to have the data segregated as per the team's results at home and team's results away from home.

6.3 Home data

Aug 13 - March 16 (Home)

№	Team	M	W	D	L	G	GA	PTS	xG	xGA	xPTS
1	Barcelona	12	10	2	0	25	1	32	32.85+7.85	8.30+7.30	29.08-2.92
2	Real Madrid	12	8	4	0	26	9	28	27.42+1.42	9.57+0.57	27.75-0.25
3	Villarreal	12	7	2	3	17	10	23	22.20+5.20	10.61+0.61	23.63+0.63
4	Almeria	13	7	1	5	20	17	22	15.70-4.30	22.67+5.67	13.78-8.22
5	Atletico Madrid	12	6	3	3	22	11	21	21.58-0.42	13.73+2.73	20.77-0.23
6	Real Betis	12	6	3	3	18	12	21	20.92+2.92	15.36+3.36	18.88-2.12
7	Girona	13	6	3	4	23	18	21	21.81-1.19	13.93-4.07	22.67+1.67
8	Mallorca	13	6	3	4	13	10	21	15.67+2.67	11.09+1.09	21.61+0.61
9	Valencia	13	6	2	5	20	12	20	21.27+1.27	13.01+1.01	23.68+3.68
10	Athletic Club	13	6	2	5	18	11	20	21.50+3.50	9.47-1.53	26.46+6.46

The comprehensive data of home football results encompasses various aspects of team performance. It includes the number of wins, draws, and losses, reflecting the outcomes of matches played on their home turf. Additionally, the data captures the goals scored by the team, highlighting their attacking prowess, while assists provide insight into the players' ability to create goal-scoring opportunities for their teammates. Points, an essential metric, reveal the team's overall success in accumulating points throughout the season. Furthermore, expected goals (xG) assess the quality of scoring opportunities created, offering an insight into the team's attacking efficiency. Similarly, expected assists (xA) gauge the probability of a pass leading to a goal, showcasing the creative abilities of the team. Expected points (xPts) provide an analytical estimation of the team's performance based on

underlying metrics. Together, these data points paint a vivid picture of a team's performance, both in terms of results and the underlying statistical indicators.

6.4 Away data

Aug 13 - March 16 (Away)

№	Team	M	W	D	L	G	GA	PTS	xG	xGA	xPTS
1	Barcelona	13	11	0	2	22	7	33	22.90 +0.90	12.30 +5.30	23.73 -9.27
2	Real Madrid	13	9	1	3	24	10	28	25.21 +1.21	13.17 +3.17	26.31 -1.69
3	Atletico Madrid	13	8	3	2	17	8	27	18.54 +1.54	14.37 +6.37	20.72 -6.28
4	Real Sociedad	13	8	2	3	20	12	26	18.84 -1.16	11.21 -0.79	23.02 -2.98
5	Real Betis	13	6	3	4	15	14	21	19.04 +4.04	18.49 +4.49	18.32 -2.68
6	Rayo Vallecano	13	4	4	5	13	16	16	12.69 -0.31	18.74 +2.74	14.31 -1.69
7	Villarreal	13	4	3	6	12	14	15	15.07 +3.07	22.74 +8.74	14.43 -0.57
8	Osasuna	13	3	6	4	10	13	15	9.80 -0.20	19.71 +6.71	10.48 -4.52
9	Espanyol	13	3	5	5	15	19	14	12.99 -2.01	19.72 +0.72	13.11 -0.89
10	Athletic Club	12	3	4	5	15	16	13	14.97 -0.03	12.94 -3.06	16.65 +3.65

The away football results data provides a comprehensive overview of a team's performance in matches played on opponents' grounds. It encompasses key metrics such as wins, draws, and losses, indicating the outcomes of away games. The number of goals scored away from home showcases the team's ability to find the back of the net in challenging environments. Similarly, assists recorded in away matches highlight the players' capability to create goal-scoring opportunities even when playing in unfamiliar surroundings. Points accumulated in away fixtures indicate the team's success in securing positive results on the road. Furthermore, expected goals (xG) for away games provide insights into the team's attacking threat and efficiency when playing away from their home stadium. Expected assists (xA) assess the team's ability to generate goal-scoring opportunities for teammates on opponents' grounds. Expected points (xPts) offer a statistical estimation of the team's performance in away matches based on underlying metrics. Collectively, these data points offer a

comprehensive analysis of a team's performance in away games, shedding light on their effectiveness in unfamiliar environments.

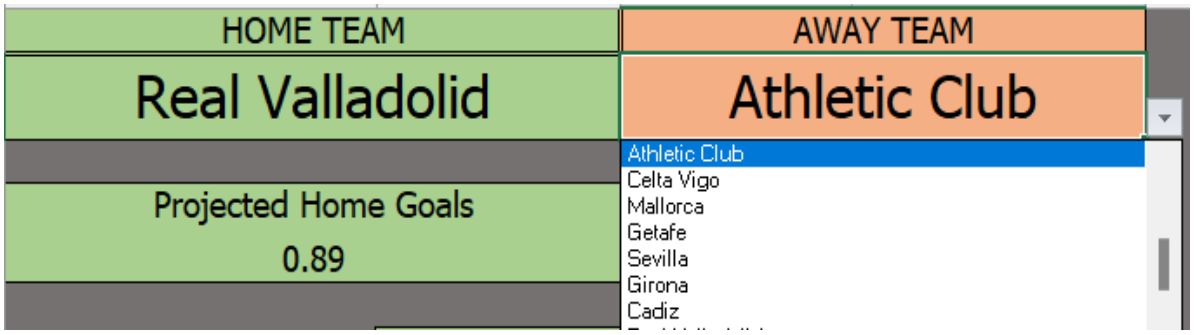
Update the home and away data for the model to be able to predict.

	HOME									
№	Team	M	W	D	L	G	GA	PTS	GS/pg	GC/pg
1	Barcelona	12	10	2	0	25	1	32	2.0833333	0.0833333
2	Real Madrid	12	8	4	0	26	9	28	2.1666667	0.75
3	Villarreal	12	7	2	3	17	10	23	1.4166667	0.8333333
4	Almeria	13	7	1	5	20	17	22	1.5384615	1.3076923
5	Atletico Madrid	12	6	3	3	22	11	21	1.8333333	0.9166667
6	Real Betis	12	6	3	3	18	12	21	1.5	1
7	Girona	13	6	3	4	23	18	21	1.7692308	1.3846154
8	Mallorca	13	6	3	4	13	10	21	1	0.7692308
9	Valencia	13	6	2	5	20	12	20	1.5384615	0.9230769
10	Athletic Club	13	6	2	5	18	11	20	1.3846154	0.8461538
11	Real Valladolid	12	6	2	4	11	10	20	0.9166667	0.8333333
12	Celta Vigo	13	5	4	4	19	14	19	1.4615385	1.0769231
13	Rayo Vallecano	12	5	4	3	16	12	19	1.3333333	1
14	Real Sociedad	12	5	4	3	13	12	19	1.0833333	1
15	Osasuna	12	6	1	5	12	11	19	1	0.9166667
16	Sevilla	13	5	3	5	16	16	18	1.2307692	1.2307692
17	Cadiz	13	4	6	3	14	17	18	1.0769231	1.3076923
18	Getafe	13	4	4	5	14	16	16	1.0769231	1.2307692
19	Espanyol	12	3	4	5	16	20	13	1.3333333	1.6666667
20	Elche	13	1	4	8	11	21	7	0.8461538	1.6153846
	Average								1.3794872	1.0346154

	AWAY									
№	Team	M	W	D	L	G	GA	PTS	GS/pg	GC/pg
1	Barcelona	13	11	0	2	22	7	33	1.6923077	0.5384615
2	Real Madrid	13	9	1	3	24	10	28	1.8461538	0.7692308
3	Atletico Madrid	13	8	3	2	17	8	27	1.3076923	0.6153846
4	Real Sociedad	13	8	2	3	20	12	26	1.5384615	0.9230769
5	Real Betis	13	6	3	4	15	14	21	1.1538462	1.0769231
6	Rayo Vallecano	13	4	4	5	13	16	16	1	1.2307692
7	Villarreal	13	4	3	6	12	14	15	0.9230769	1.0769231
8	Osasuna	13	3	6	4	10	13	15	0.7692308	1
9	Espanyol	13	3	5	5	15	19	14	1.1538462	1.4615385
10	Athletic Club	12	3	4	5	15	16	13	1.25	1.3333333
11	Celta Vigo	12	3	3	6	10	20	12	0.8333333	1.6666667
12	Mallorca	12	3	2	7	9	16	11	0.75	1.3333333
13	Getafe	12	2	4	6	11	18	10	0.9166667	1.5
14	Sevilla	12	2	4	6	13	24	10	1.0833333	2
15	Girona	12	2	3	7	15	21	9	1.25	1.75
16	Cadiz	12	2	3	7	6	20	9	0.5	1.6666667
17	Real Valladolid	13	2	2	9	8	25	8	0.6153846	1.9230769
18	Valencia	12	1	3	8	8	16	6	0.6666667	1.3333333
19	Elche	12	1	3	8	8	28	6	0.6666667	2.3333333
20	Almeria	12	0	3	9	9	27	3	0.75	2.25
	Average								1.0333333	1.3891026

6.5 Select the appropriate teams

The first game is Real Valladolid (Home) and Athletic Club (Away)



Right off the bat we get the projected goals that may be scored by both the teams in this fixture.

Projected Home Goals 0.89		Projected Away Goals 1.01	
	Home Win	Draw	Away Win
% Chance	30.86%	31.81%	37.23%
Implied Odds	3.24	3.14	2.69

Athletic Club is likely to win in terms of goals and away win odds.

Home Attack	0.66
Away Defence	0.97
Away Attack	1.21
Home Defence	0.81

We see the Home Defence is outweighed by the Away Attack and thus gives the Away side a slight edge.

6.6 Score possibilities

Correct Score	Imp Odds
0-0	6.64
0-1	6.59
0-2	13.10
0-3	39.02
1-0	7.49
1-1	7.44
1-2	14.78
1-3	44.05
2-0	16.91
2-1	16.80
2-2	33.37
2-3	99.43
3-0	57.27
3-1	56.88
3-2	112.99
3-3	336.66

6.7 Noise

Apart from this specific observation, I have noticed that the score 1-1 seems to have the lowest probability almost every time. This is maybe because every team tends to score and concede goals over the course of the season.

The caveat is that if the percentage distribution of home win, draw, and away win suggests any outcome other than a draw, we can disregard the noise and proceed to consider the second most probable result.

This case is diffirent as it shows that there is no noise and 0-1 (Implied odds 6.59) is the most probable result to Athletic Club. However, the real result was 1-3 to Athletic Club so we got the winning team right however, the we did not get the score perfectly right which is acceptable.

Therefore, the model does have its limits.

6.8 FIXTURE 2 [Real Betis vs Mallorca]

Real Betis	Mallorca
Projected Home Goals	Projected Away Goals
1.45	0.72

The odds for the game are shown as below:

	Home Win	Draw	Away Win
% Chance	54.47%	26.84%	18.64%
Implied Odds	1.84	3.73	5.37

This game favours the home team much more than the away team and thus the home team has a slightly higher probability of winning than the away team.

Correct Score	Imp Odds
0-0	8.80
0-1	12.14
0-2	33.49
0-3	138.60
1-0	6.07
1-1	8.37
1-2	23.10
1-3	95.60
2-0	8.37
2-1	11.55
2-2	31.87
2-3	131.88
3-0	17.33
3-1	23.90
3-2	65.94
3-3	272.89

Noise

Again, in this instance the probability of a home win is slightly higher so we can ignore the draw and move on to the next possibility which is 1-0 (Implied odds 6.07) to Real Betis.

That is clearly the result that took happened in actuality. Not only did the model predict the winner but also got the scores perfectly right.

6.9 FIXTURE 3 [Espanyol vs Celta Vigo]

I am purposely taking a fixture where the model got the prediction wrong so we can see the limitations of this model

Espanyol	Celta Vigo
Projected Home Goals	Projected Away Goals
1.61	1.34

	Home Win	Draw	Away Win
% Chance	43.79%	24.28%	31.24%
Implied Odds	2.28	4.12	3.20

Here, we can clearly see the model indicates for a Home win

Correct Score	Imp Odds
0-0	19.17
0-1	14.28
0-2	21.27
0-3	47.54
1-0	11.90
1-1	8.86
1-2	13.21
1-3	29.51
2-0	14.77
2-1	11.01
2-2	16.40
2-3	36.64
3-0	27.51
3-1	20.50
3-2	30.54
3-3	68.24

The predicted result is 1-1 (Implied odds 8.86) to Espanyol however the real result was 1-3 to Celta Vigo. It was the 13th most likely prediction according to the model.

Therefore, the model does have its limits.

6.10 BACKTESTING

In order to validate the model, I performed some back testing with all the available data all the till GW10. All 10 previous Gameweeks cannot be predicted accurately because for this model to work there has to be some Home and away data for every team

The results of the back tested data are as follows

Gameweek	Perfect Prediction (Goals model)	Result Accuracy (Goals model)	Perfect Prediction (xG Model)	Result Accuracy (xG Model)
GW10	10%	60%	10%	60%
GW11	0%	50%	10%	60%
GW12	20%	60%	50%	60%
GW13	10%	30%	20%	50%
GW14	10%	80%	20%	80%
GW15	20%	50%	20%	60%
GW16	20%	50%	30%	60%
GW17	50%	75%	13%	38%
GW18	20%	70%	30%	70%
GW19	20%	60%	10%	40%
GW20	17%	50%	25%	50%
GW21	30%	50%	20%	60%
GW22	30%	80%	20%	70%
GW23	10%	30%	10%	30%
GW24	0%	30%	0%	50%
GW25	40%	50%	50%	80%
GW26	10%	80%	40%	80%
GW27	10%	40%	0%	40%
GW28	0%	70%	0%	50%
GW29	20%	50%	20%	60%

GW30	30%	70%	30%	90%
GW31	10%	60%	10%	60%
GW32	10%	60%	10%	70%
GW33	30%	80%	20%	80%
GW34	10%	70%	20%	70%
GW35	--	--	--	--
GW36	--	--	--	--
GW37	--	--	--	--
GW38	--	--	--	--
TOTAL	**0.1748**	**0.582**	**0.1952**	**0.6072**

In the above table the first column marks the gameweek.

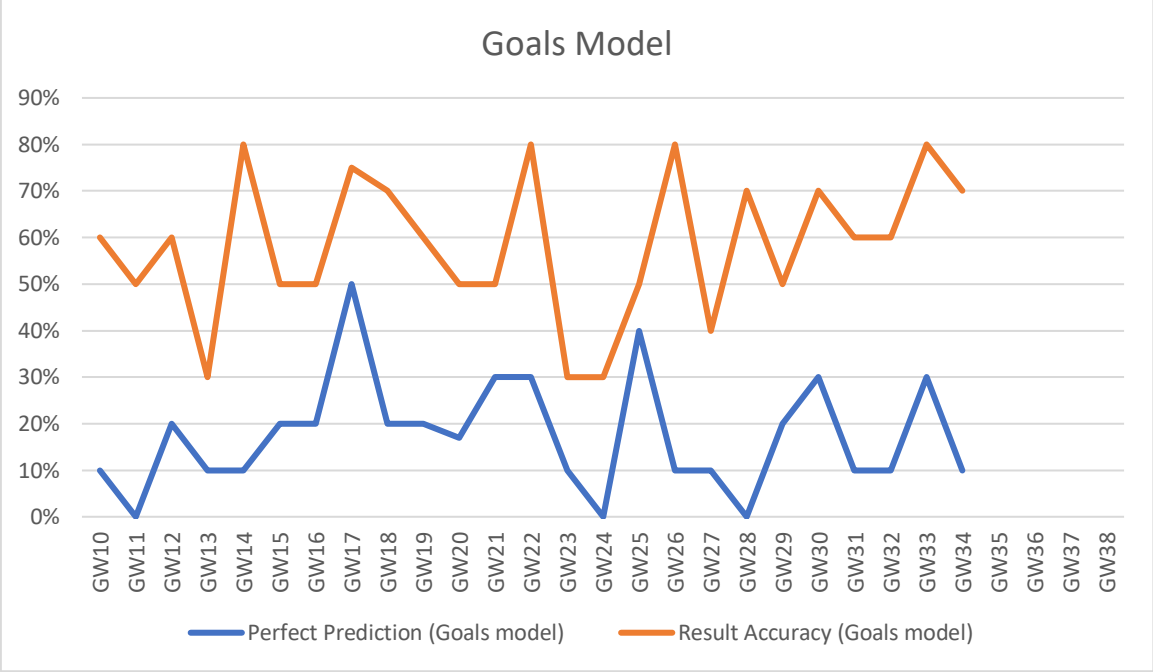
Perfect Prediction (Goals model) – Shows us how many results in that Gameweek were predicted accurately, down to the exact scoreline, using the goals model

Result Accuracy (Goals model) – Shows us how many results did the model get right in the Gameweek irrespective of the scoreline, using the goals model

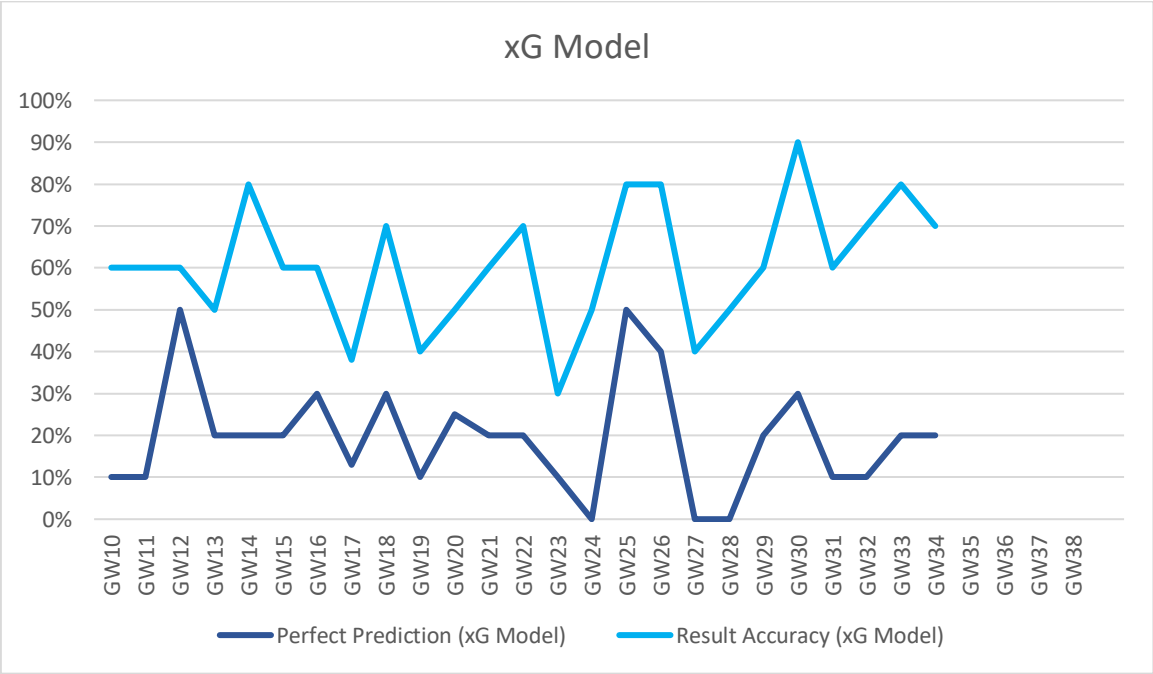
Perfect Prediction (xG model) – Shows us how many results in that Gameweek were predicted accurately, down to the exact scoreline, using the expected goals model

Result Accuracy (xG model) – Shows us how many results did the model get right in the Gameweek irrespective of the scoreline, using the expected goals model

6.11 MODEL SUMMARY



As we can see, over the season the Goals model has provided 17.46% perfect predictions and 58.2% correct predictions. The results seem to improve over time with more and more data.



Here we can see, over the season the Expected Goals model has provided 19.5% perfect predictions and 60.7% correct predictions.

By examining the charts, we can observe that the Expected Goals (xG) model exhibits a higher level of reliability in terms of both perfect predictions and result accuracy. This is evident in the noticeable gap between GW13 (Game Week 13) and GW22 (Game Week 22), indicating the model's superior performance during that period. The larger the gap, the more pronounced the divergence between the xG model's predictions and the actual outcomes, thus demonstrating its enhanced reliability.

6.12 FORWARD TESTING

I had the opportunity to conduct a forward test as the development of my model coincided with the approach of GW 34.

Gameweek 34

Goals Model

Actual			
Team	Home	Away	Team
Mallorca	1	0	Cadiz
Real Sociedad	2	2	Girona
Osasuna	3	1	Almeria
Villarreal	5	1	Athletic Club
Real Madrid	1	0	Getafe
Celta Vigo	1	2	Valencia
Elche	1	0	Atletico Madrid
Real Valladolid	0	3	Sevilla
Espanyol	2	4	Barcelona
Real Betis	3	1	Rayo Vallecano

Predicted			
Team	Home	Away	Team
Mallorca	1	0	Cadiz
Real Sociedad	1	0	Girona
Osasuna	1	0	Almeria
Villarreal	1	0	Athletic Club
Real Madrid	2	0	Getafe
Celta Vigo	1	0	Valencia
Elche	0	2	Atletico Madrid
Real Valladolid	0	1	Sevilla
Espanyol	0	2	Barcelona
Real Betis	1	0	Rayo Vallecano
Perfect	10%		
Accuracy:	70%		

xG Model

Actual			
Team	Home	Away	Team
Mallorca	1	0	Cadiz
Real Sociedad	2	2	Girona
Osasuna	3	1	Almeria
Villarreal	5	1	Athletic Club
Real Madrid	1	0	Getafe
Celta Vigo	1	2	Valencia
Elche	1	0	Atletico Madrid
Real Valladolid	0	3	Sevilla
Espanyol	2	4	Barcelona
Real Betis	3	1	Rayo Vallecano

Predicted			
Team	Home	Away	Team
Mallorca	1	0	Cadiz
Real Sociedad	1	0	Girona
Osasuna	2	1	Almeria
Villarreal	1	0	Athletic Club
Real Madrid	2	0	Getafe
Celta Vigo	1	0	Valencia
Elche	1	2	Atletico Madrid
Real Valladolid	1	0	Sevilla
Espanyol	0	1	Barcelona
Real Betis	1	0	Rayo Vallecano
Perfect	20%		
Accuracy:	70%		

As we can see both the models were able to get 63.64% result accuracy. However, the xG model got 1 more result perfect in comparison to Goals model.

CHAPTER 7: Conclusions

- When it comes to predicting football match outcomes, focusing on a team's Goals Scored and Goals Conceded, both at home and away, can form the basis of a reliable predictive model.
- The Poisson Distribution provides a useful framework for understanding the results of football games from a conceptual and visual perspective.
- While the model has its strengths, it also has limitations that need to be acknowledged.
- One significant limitation of the discussed model is its failure to account for important factors such as injuries, suspensions, and form from other competitions, which can have a significant impact on football matches.
- Despite not considering these factors, the model has still demonstrated the ability to predict approximately 50% of the variance in football results by relying solely on two factors.
- Another limitation is the introduction of noise when making result predictions, but this can be mitigated by analyzing the dispersion of win probabilities for each team. [6.7]
- It is worth noting that the predictive model assigns probabilities to every possible result and ranks them based on their odds. Interestingly, the top three odds (excluding noise) often align with the actual outcomes observed.

CHAPTER 8: Further development of the model

Although this model is still a work in progress, it is important to acknowledge that there are other existing models in the market that serve the same purpose and may offer higher accuracy. However, there is still potential for further exploration and refinement within this particular model by considering different permutations and combinations of factors.

1) In this model, it should be noted that only "Cumulative Data" has been considered when inputting the data. This means that the predictions for upcoming gameweeks are based on the complete data of the entire season.
   - An argument can be put forth that as data becomes more outdated, it may become irrelevant and contribute to the noise in predictions.
   - For instance, if a team replaces its manager during the season, relying on previous xG and xGA data could lead to inaccurate predictions due to the introduction of a new playing style under the new manager.
   - The same can apply if a team experiences significant player injuries or transfers. To address this, I suggest the concept of utilizing segmented data.
   - Segmented data entails considering data for teams starting from a significant point in the season rather than the beginning.

2) Granular Analysis for each team involves conducting a thorough examination of the players comprising the team.
   - When discussing xG and xGA for a team, these metrics are actually aggregates based on the performance of each individual player.
   - This leads to the question of whether we can develop a model to predict which player is most likely to score or assist in a given game.

3) The logical progression for the existing model is to incorporate additional features.
   - The inclusion of weather conditions as a feature is a potential enhancement, considering their impact on player movement and ball dynamics.

- Morale is a non-tangible feature but it plays a big role in the psychology of any player as poor team with high morale can cause an upset against a team which is better.

4) Apart from projections this concept can help with a structured data collection system
   - There is certainly a requirement for a well-organized platform that provides accessible data for individuals associated with the football community, similar to the user-friendly interface offered by understat.com.

5) In my opinion, the most significant and valuable application for this model is to develop a platform that offers comprehensive and detailed analytics for every team and player, allowing for in-depth analysis.
   - A platform of this kind would gain immense popularity within the football community, particularly among fans interested in betting and fantasy sports.
   - Monetization opportunities could be explored by launching social media pages and offering subscription-based memberships to access the platform's advanced features and insights.

In addition to football, it is evident that similar models can be developed for other sports as well.

In non-sport domains, the Poisson distribution can be utilized in several predictive scenarios. For instance, call centre data analysis can benefit from this distribution by modeling the number of incoming calls per unit of time. This enables call centre managers to estimate the probability of encountering different call volumes at specific time intervals, allowing for effective resource allocation and staffing decisions.

Moreover, the Poisson distribution finds application in diverse areas such as insurance risk assessment, earthquake occurrence analysis, and website traffic analysis. These fields often involve events that occur randomly and independently, and the Poisson distribution provides a valuable mathematical framework for assessing the probability of rare events. As a result, it plays a critical role in various research and analytical domains beyond the realm of sports.

Bibliography

[1] Viewer ship stats - https://khelnow.com/football/top-ten-most-watched-football-leagues-in-the-world

[2] Moroney M. J. (1956) *Facts from figures*. 3rd edition, Penguin, London. [Literature Review (1,3)]

[3] C. Reep and B. Benjamin (1968) *Skill and chance in association football*. Journal of the Royal Statistical Society, Series A, 131, 581-585.

[4] Hill I.D. (1974), *Association football and statistical inference*. Applied statistics, 23, 203-208.

[5] Maher M.J. (1982), *Modelling Association Football scores*. Statistica Neerlandica, 36, 109-118

[6] Cornea K.S. and Carron A.V. (1992) *The home advantage in sports competitions: a literature review*. Journal of Sport and Exercise Physiology, 14, 13-27.

[7] Knorr-Held, Leonhard (1997) *Dynamic Rating of Sports Teams*. (REVISED 1999). Collaborative Research Center 386, Discussion Paper 98

Fixture List: https://fantasy.premierleague.com/

xG interpretation: How to Build An Expected Goals Model 1: Data and Model

Kie Millet: How to create football Prediction model