MSc Data Science for Business
Research Paper
Academic Year 2021-2022

Data driven betting strategy for football matchs

FOURNIER Paul

Under the supervision of

Prof. Vincent FRAITOT

June 30, 2022

Contents

1 Executive Summary

The online sports betting industry has been booming since its legalization, and with more revenue generated each year, one is tempted to ask if the sports betting market is exploitable. Betting companies have been quick to adopt data analysis to maximize returns. They purchase the consulting services of mathematical modeling experts to derive their betting odds and identify the profiles of the most profitable betters to block them [1].

This research paper explores data modeling methods applied to predicting outcomes of football games with the goal of generating profits on betting markets. Football is the most popular sport worldwide, yet predicting the outcome of games is not trivial. Firstly, I present the intricacies of the business model of betting companies. I explain the ways they maximize profits and minimize risks by setting odds and adjusting them through time. I also present the point of view of betters and the potential betting opportunities available. Then I describe the data available and used by fans, betting odds providers and football clubs. This is an important part of the paper that focuses on the resources (with respect to data) available for modeling and gives insights into the current state of the data revolution that football experiences.

I have done extensive work towards researching and understanding of the literature available on football predictions. I present hallmarks of the published academic papers to familiarize the reader with the state of the art and build a theoretical understanding of the methods attempted. The research can be split between the predictions made before the game starts and the predictions made while the game is unfolding. The mathematical formulas are explained for relevant models.

Finally, I present the implementations of those models and describe the adaptations and improvements, I made. I detail the way I evaluated the models in a way that is realistic if I were to apply this technique in real life using back-testing. Ultimately, I present the portfolio management techniques used to maximize returns. I show that the performance of the betting odds and show that they are not perfect which opens the door to opportunity. Yet, the methods implemented did not outperform them. I interpret the results with a deep analysis of the predictive models. Overall, this paper concludes that applying the existing literature on football outcome prediction to betting markets will not yield profits.

2 Preface

First and foremost, I would like to thank my tutor, Vincent Fraitot, for his availability and his recommendations.

I would like to thank LiveOddsAPI, FiveThirtyEight and Football.co.uk for making their data public and for enabling me to use it throughout this paper.

I would like to thank HEC for this great opportunity to explore and research a topic of my choice. Their support and resources alongside the lectures throughout the Master were instrumental for the success of this exercise.

3 Introduction

The turnover of the gambling sector in France in 2021 is around 10,7 Billion Euros [2]. Yet, like many industries during the Covid-19 pandemic, the gambling sector suffered initially. Indeed the chaos of the global sanitary crisis in 2020 caused sporting events to be postponed or outright cancelled and the physical points of sale to close (because they are non-essential businesses). Yet since 2021, the gambling sector is thriving; all key performance indicators point towards a promising future for the industry. For instance, the number of active players on online sports betting platforms increased by 16% from 2021 to 2022, reaching a total of 4.5 Million. Additionally, the total sum of wagered amounts skyrocketed to 7 891€M, a 47% increase, which lead to a total revenue of 1 355€M made by bookmakers, which is a 44% increase.

One possible explanation for the meteoric rise of the popularity of online betting platforms is the steady growth of advertisement made by the betting companies to capture new audiences. From sponsorships of major football teams and events (like Winnamax's sponsorship with RC Lens and Lille OSC), to endorsements of influencers and content creators (for instance Betclic's sponsorship of *Wiloo* on YouTube and Winnamax sponsorship of *Villebrequin* on Youtube) and most visibly, the traditional television and media advertisements. A study conducted by Addictions France found that 40% of billboards in the subway in Paris and a fourth of television ads prior to football games (during the Euros 2021) were for gambling services [3]. While the number of gambling advertisements have increased, the marketing manoeuvres employed by the betting companies raised some concerns. Indeed, Addictions France claims that marketing campaigns target the young and the low-income minorities by adopting slang and emotional slogans. This trends was picked up by the Gambling National Authority (ANJ) that acted out and forbade one campaign from Winnamax to circulate. The commission later stated its determination to regulate the market more thoroughly [4].

Historically in France, the Francaise des Jeux (FDJ) and PMU were the only organizations legally authorized to organize gambling. They would respectively operate the lottery and horse race betting through their physical points of sales. Yet since 2010, the French government has legalized online gambling and opened the market to new and foreign competitors [5]. This legalization was a strategy to cut down on the illegal gambling websites operated by foreign companies and conversely to regulate and tax the companies locally. The French government estimated that the legalization of gaming would bring in 100€M per year. To do so, the French government takes 7.5% of players' bets for horse racing and sports betting and 2% of bets for poker.

While gambling is a very profitable industry that generates steady tax revenue, some negative trends have become apparent. Firstly the number of young gamblers is increasing and secondly, the number of addicted gamblers also increases rapidly. Isabelle Falque-Pierrotin, president of the aforementioned ANJ says that "more than a third of 15 to 17 year olds have engaged in online gambling" (despite it being illegal for minors to gamble) due to the constant exposure of ads [6]. Alarmingly, Thomas Gaon, clinical psychologist at Marmottan Hospital has observed an increase in the number of cases of addicted gamblers and noted : "there is a rejuvenation of pathological gamblers, they often started underage. This is due in particular to an extremely powerful marketing, very generational, supported by advertisements in loop, on television and on social networks". A 2012 study by the Institute for Economic Research at the University of Neuchatel, estimated that excessive gambling (problem and pathological) costs the Swiss community 525€M each year in the form of additional health expenses, reduced work performance and loss of health-related quality of life [7].

The gambling sector can be separated into two categories: traditional gambling (like the lottery and casino games) and sports bets which will be the focus of this paper. Sport betting is the process of staking a wager on the outcome of an event. The range of events one can bet on varies from betting company to another but they are defined as the noticeable events that occur during a sports match (like the number of goals scored, the winner, the goalscorer etc.) or over a season (the winner of a league or the top goalscorer). Each outcome is associated with a betting odd, which is used to compute the returns yielded from the event occurring. There are different ways to represent betting odds but the most widely used format in Europe is decimal odds. To compute the returns yielded from a successful bet, one simply multiplies the wager by the decimal odds. For instance, say one bets 10€ on the home team to win with the betting odds of 2.1. If the home team wins, one receives $10 \times 2.1 = 21$€. Note that one does not earn 21€ in addition to the initial stake, one should consider the returns to be 11€ plus the buy-in. It follows that in the case where the game ends with a draw or the away team winning, then one outright loses the 10€. Historically the betting odds were determined by human experts and they were generated by each individual bookmaker, but the sports betting industry has evolved and now most betting platforms purchase the odds from a centralized company. Companies such as Smartodds have grown to use cutting edge statistical research and sports modelling to offer accurate odds. Betting companies thus buy the football outcome odds and offer the service of betting on the events.

Of all the events that can be modelled, we will focus in this paper on the outcome of a football game; we will look at football games where either team wins or the game ends with a draw. In this thesis I will explain the intricacies of the business model of betting companies and also the approach of gamblers. I will review academic literature published on the subject to familiarize myself with the state of the art and build a theoretical understanding of the methods attempted. I will present the results of existing models and then show the adaptations and improvements made on these models. For the sake of simplicity, I kept the scope of the data to the English Premier League and I will train a model on season 2017-18 to 2020-21 and then evaluating it on the 2021-22 season. We will analyse the performance of the betting odds and show that they are not perfect which opens the door to opportunity. And finally I will offer a betting strategy to attempt to exploit this.

4 Setting the problem

4.1 Bookmakers

4.1.1 Margin

We have already explained how betting companies, also referred to as bookmakers, provide a service to clients by offering a payout proportional to the wager conditional on the prediction occurring. Like any business, betting companies are looking to make and maximize profits. To do so, the odds proposed are always tipped in the favor of the betting site. This is called the margin, which refers to the safety cushion applied to the odds that can be interpreted as the cut taken by the house to offer the service of betting and to offset the risk taken by the bookmaker for offering this service. Indeed, if a betting site had no margin, it would only make profits if betters were to bet on the wrong even more than on the right event. Take the example of a football game with the following betting odds: $o_h = 5.25$, $o_d = 3.75$, $o_a = 1.7$ (denoting the home win, draw and away odds respectively). To compute the margin taken by the betting companies, we sum the inverse of the decimal odds:

$$Margin = 1 - \sum_{r\in(h,d,a)} \frac{1}{o_r} = 1 - (\frac{1}{1.7} + \frac{1}{3.75} + \frac{1}{5.25}) = 0.045$$

The margin taken by the betting site in this example is 4.5%. To put this into perspective, if we were to bet on all outcomes of that event, then we would lose 4.5% of the sum of our wagers. Note that the placed wager would be proportional to the odds of the corresponding event and not the same wager for each outcome. However, if the margin is zero, then the house is running a friction-less service and betting on all outcomes would payoff the same as the wager. Now if the house has a negative margin, then the house is running at a deficit and betting on all outcomes of the game would always make a profit. This of course never happens. Thus the phrase "the house always wins" used to talk about the casino games can also be used for betting odds. On average, the margin is between 4 and 5% for football. Interestingly, in the last dozen years, the margin of English bookmakers has halved from an average of 9% in 2005 to an average of 5% in 2018. The competition becoming ruthless between betting companies drives them to lower their margin in order to attract betters with the most profitable odds [8].

4.1.2 Hedging risks

Betting companies like any business try to maximize profits while minimizing risks. One simple way they do this is by enabling early withdrawal: where a user can retrieve a fraction of his wager before the end of a game if his bet currently is fulfilled. This is an attractive marketing topic for betting companies, but from a business perspective, it represents an opportunity to minimize the payout given to the better which minimizes risk.

Another way the betting companies hedge risk is by adjusting the odds through time. There are two reasons a bookmaker might adjust the odds, the first one is to take into account new information that has come to light and the second one is to balance the books. The first reason is straightforward, the bookmakers reevaluate the likelihood of an event occurring when new information is available and ensure the betting odds reflect that. For instance if a player got injured during the previous game or during training this might affect the team's performing in the upcoming game. The second reason stems from the fact that betting companies accept wagers practically unconditionally (the only exception being the minimum and maximum betting amounts). This means that as soon as an event is available for wager, anyone can place a bet on it. Now assume that for a specific event, the crowd unevenly bets on a single outcome of the event. The bookmaker is now in a risky position where in the case of the outcome favored by the crowd occurs, then they have to payout a massive amount of money. For this to be economically viable, the same betting company would need other events (with equally uneven bets) to offset the loss. Which would mean the betting company would offset its risky position by being in multiply risky position. This would be unreasonable because on the off-chance all the favored events occur, the betting site would lose considerable capital. Thus betting companies try to maximize the spread of wagers on the events. In other words, they attempt to have an balanced amount staked on each outcomes of an event (adjusted for the betting odds). To do so, they adjust the odds through time to attract the new betters to wager on the outcome unfavored by previous betters in order to cover their exposure. Therefore, with balanced bets on outcomes, the betting companies payoff the winners with that specific event's looser and make a profit with the aforementioned margins.

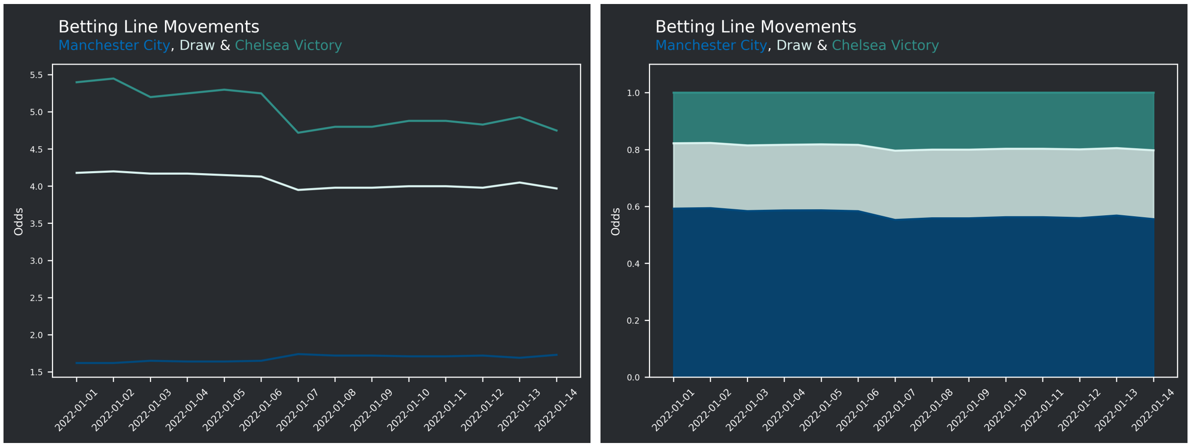


Figure 1: (a) Betting line movement shown with raw odds (b) with implied odds

Take for example the game between Manchester City and Chelsea Football Club that took place January 15th 2021. Figure 1 shows the betting odds from betting site Betclic from the opening odds (from the day they are released) to the closing odds (last available pre-game odds). I scraped the data using The Odds API [1] which gives the live odds of many betting websites. I collected this data daily between January 1st 2021 and January 15th 2021 at midday [2]. The figure also shows the implied odds which is a method of processing betting odds by dividing the inverse of the odds by the sum of the inverse of the odds. This is a way to represent the betting odds without the bookmaker margin, for the outcome $r \in (home, draw, away)$:

$$I_r = \frac{1}{\frac{o_r}{\frac{1}{o_h}+\frac{1}{o_d}+\frac{1}{o_a}}}$$

There is no way to know if the betting companies apply margins uniformly to all betting lines, but this method gives an estimate for the betting companies predicted outcomes.

[1] https://the-odds-api.com/

[2] available here https://github.com/Fournierp/FPL

Manchester City were favorites, sitting comfortably on top of the league and having won its previous 11 games while Chelsea drew 4 of their last 5 games and were going to play away from home. Chelsea, despite being considered the underdogs for this fixture by starting with implied odds of 17%, were favored by the crowd. So much so that the betting site Betclic adjusted the odds considerably from 5.4 to 4.75 (a 12% decrease), even though the implied odds stayed quite stable going from 18% to 20%. The decreasing betting odds for an outcome signifies that the the crowd heavily bets in its favor and the bookmakers adjust their odds accordingly. It follows that since the home team had odds of 1.6 to begin with and the away team had 5.4, the amount wagered on the away team would need to be at least $\frac{5.4}{1.6} = 3.375$ greater than on the home team to create a considerable imbalance in the betting spread. Without the information about the number of wagers placed on each outcome and the exact amount wagered, it is not possible to prove that the betting sites adjusted the odds to hedge risk. Yet since the implied odds were stable, one can assume that the betting companies did not reevaluate the betting odds in favor of the underdogs due to new public information but in fact to hedge risk.

The process of adjusting betting odds through time is referred to as the betting line movements. Some research has been made on analysing the betting line movement as a basis for betting strategies. Strategies involving betting on the outcome whose odds are decreasing (called reversing the line movement) have for instance been developed [9].

4.2 Betters

4.2.1 Bookmaker selection

From the point of view of a better, the optimal strategy is to select a betting site based on the proposed betting odds. Indeed, the betting odds directly indicate the payoff one might get in case of the event occurring. Even though most betting sites purchase the services of experts that produce the odds for them, as we have seen in the previous section, each betting site has its own recipe for the margin applied to an event and for the adjustment of their odds based on the bets placed. This means that sometimes the betting odds across different bookmakers might differ. In this case, it is in the better's interest to select the bookmaker with the highest decimal odds for the selected event.

4.2.2 Arbitrage

Arbitrage is the practice of taking advantage of a difference in prices in two or more markets. With respect to betting it is the process of placing bets on all sides of an event chosen carefully such that regardless of the outcome, a profit will be made. In the section on bookmakers, we introduced the concept of margins which make it impossible for a better to place wagers on all events and expect to make a profit. Yet the better can spread the wagers across multiple bookmakers and if the betting odds are different enough, an arbitrage opportunity might be available. For the sake of simplifying the calculations of this example, let us consider a betting opportunity on a football game where we predict which team wins and no draws can happen (for instance an elimination game where penalty shootouts will always decide a winner). Let the fictional betting odds from two different bookmakers be the following:

	Home win	Away win	Margin
Bookmaker A	1.4	2.9	6%
Bookmaker B	1.6	2.4	4%

Table 1: Betting odds

To evaluate if there is an arbitrage opportunity one has to compute the sum of inverse of the maximum decimal odds. If the sum is below one, then this means that this combination will yield a profit. To express it mathematically given our example betting odds, we denote the odds from bookmaker A for the results r as o_r^A.

$$ArbitrageMargin = \frac{1}{max(o_h^A, o_h^B)} + \frac{1}{max(o_a^A, o_a^B)} = \frac{1}{1.6} + \frac{1}{2.9} = 0.97$$

By betting on the home team at Bookmaker B and on the away team at Bookmaker A, one stands to make a 3% profit. To do so, one has to place stakes carefully to get the payoff :

$$Stake_h = \frac{max(o_a^A, o_a^B)}{max(o_h^A, o_h^B) + max(o_a^A, o_a^B)} \times 100 = \frac{2.9}{2.9 + 1.6} \times 100 = 64.4\%$$

$$Stake_a = \frac{max(o_h^A, o_h^B)}{max(o_h^A, o_h^B) + max(o_a^A, o_a^B)} \times 100 = \frac{1.6}{2.9 + 1.6} \times 100 = 35.6\%$$

Assume we have a total of a 100€ to wager on this event, then we should place 64.4€ on the home team with bookmaker B and 35.6€ on the away team with bookmaker A. Then one will be able to benefit 3.04€ in case of a home win:

$$Profit_h = (64.4 \times 1.6 + 35.6 \times 0) - 100 = 3.04$$

As well as a benefit of 3.24€ in case of an away win:

$$Profit_a = (64.4 \times 0 + 35.6 \times 2.9) - 100 = 3.24$$

So an overall average expected profit of 3.14€ (assuming a equal probability of probability of events occurring):

$$Profit = (64.4 \times 1.6 + 35.6 \times 2.9) \times \frac{1}{2} - 100 = 3.14$$

Betting companies are very much aware of the opportunity that arbitrage betting represents. They adjust the betting odds very carefully making these opportunities very rare. In addition, the bookmaker margins are also a layer of security that bookmakers manipulate to avoid these opportunities.

5 Data

This section is dedicated to describing the range of data types and sources available in the sphere of football. This data is nowadays used by fans, betting odds providers and football clubs, and could be of great use to model football game outcomes.

5.1 Raw Statistics

The first element one may use to make an informed bet are statistics. Statistics are the records of what occurred during a game. The statistics can be either general team related or individual player performance statistics. The most basic team statistics one can look at is the historical goals scored and conceded which can give an idea of the strength of a team's defence and offence. Statistics like number of fouls made and conceded, number of shots, shots on target, percentage of possession, number of corner kicks and many more will give insights into the play style of a team. More specific player statistics can be the number of passes, percentage of successful passes completion, number of shots, goals ... These statistics will give insight into the relative importance of a player in the team as well as the role (either a defensive asset, creative or decisive player) and also the current form of the player. All put together, one can identify the weaknesses and strengths of teams and base their predictions upon those. With the popularity of statistics, many newspapers and websites make these statistics available freely for the public.

5.2 Press conference

Press conferences are set up before and after games so that journalists ask questions to each teams' coaches. Generally, journalists get to ask about tactical matchup, player performances and most importantly, player injuries. During pre-game press conferences, coaches will sometimes give information about a player's injury, recovery period or medical results. During post-game press conferences, coaches will sometimes give initial medical analysis about a player that got subbed off during the game due to injury concerns. It comes to no surprise that a team with a lot of injuries on key players might be destabilized during their next game. A better or a model could retrieve valuable information from press conferences. Some general newspapers sometimes offer the big talking points but usually local newspapers and the specific team's website will relay the entire press conference to the public.

5.3 Refined metrics

The beauty of football or arguably what makes a scoreless draw not entertaining is that goals are rare events. I calculate that on average, the home team scores 1.5 goals and the away team only scores 1.2 goals (the data comprises of all English Premier League games from 2016 to 2022). Rare and random events can be modeled using a Poisson distribution which expresses the probability of a given number of goals scored in a fixed interval of time or space. We need these events to occur with a known constant mean rate and independently of the time since the last event. In Figure 2, we can see that goals closely follow a Poisson distribution. They do not follow perfectly given that we cannot be sure that events have constant mean rate: a substitution or an injury may alter greatly a team's capacity to score during a game. Nor can we ensure that they happen independently of the time since the last event: a team might take advantage of the momentum of scoring a goal to pounce and score many more quickly thereafter. Note that looking solely at the average goals scored by teams during the Covid-19 Pandemic (when games were played without an audience), we see that the home team scores 1.35 goals and the away team scores 1.34 goals. We see that the home field advantage, widely accepted as a considerable winning boost to the home team, might come from the support of the crowd rather than sheer familiarity with the stadium. Figure

2 indeed shows the effect of playing at home has on the number of goals scored. The right figure shows that games played without a crowd (due to Covid-19 regulations) tend to have less difference between goals scored distributions.

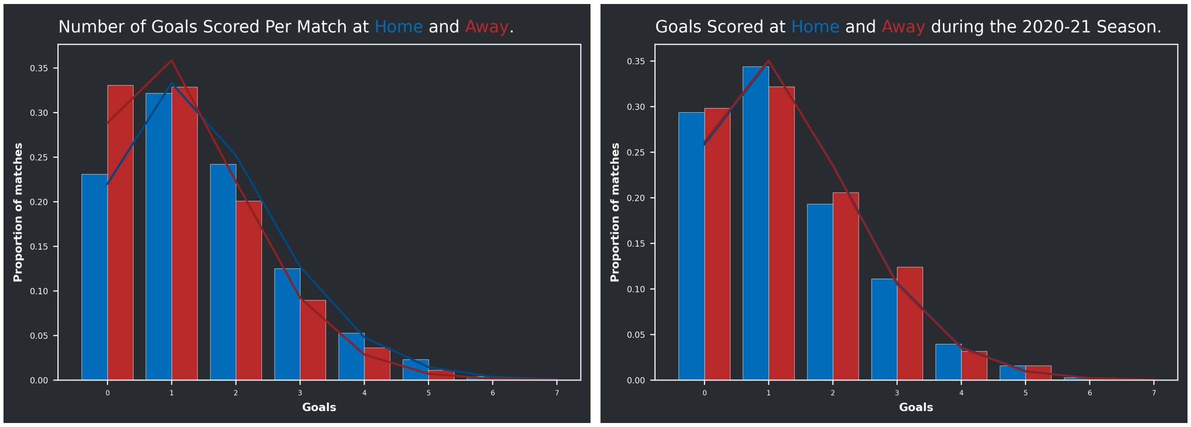
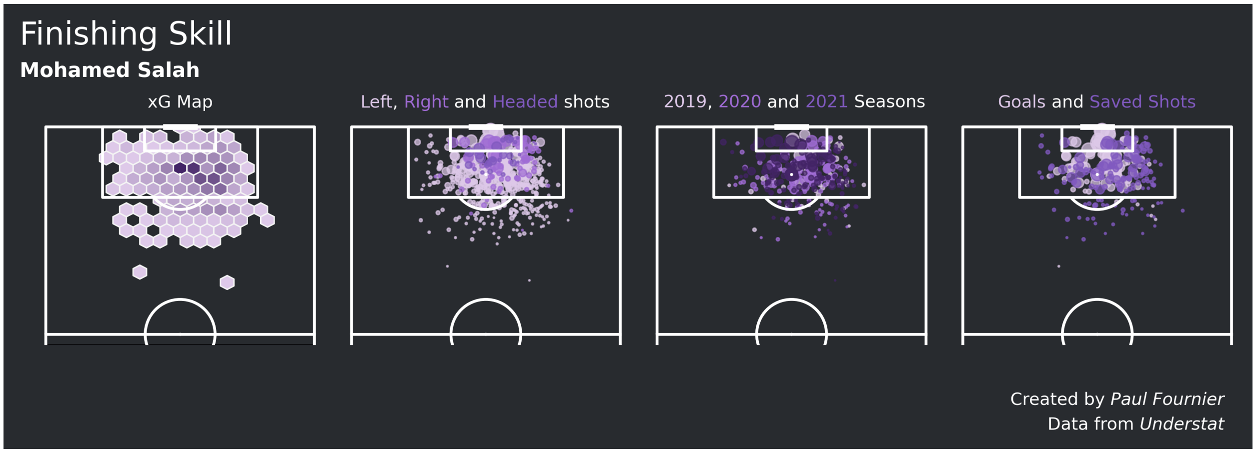


Figure 2: Distribution of goals (histogram) in the English Premier League along side a Poisson distribution (lines) with rate equal to the mean goals scored

Assuming that goals follow the Poisson distribution shown in Figure 2, from a statistical point of view the result of a football match is almost as much noise as it is signal. Take the home goals which follows a Poisson distribution with rate 1.5, the standard deviation would be then be 1.22. Thus the noise (1.22) is only slightly smaller than the signal (1.5).



Created by *Paul Fournier*
Data from *Understat*

Figure 3: Mapping Mohamed Salah's likelihood of scoring from inside the penalty box

Considerable research efforts have been put into developing ways to replace the raw statistics widely used for decades (mainly the goals scored) with more precise and less noisy metrics. In 2012, sports data collection and analysis company Opta introduced a new metric called

Expected Goals (XG) [10]. Expected Goals are an attempt to define the quality of shots. An XG model is a machine learning model that uses data to predict if a shot is likely to lead to a goal. The most important features that one can use is the position of the player shooting, the position of the goalkeeper as well as the positions of nearby opponents. Additional improvements can be made by including tracking data about last few seconds of player position, preferred foot used and the context of the shot (whether its a penalty, a corner or a counter attack) [11]. The value of XG over goals as a metric for a team or a player's quality is great to measure both the volume of chances created as well as the quality of those. Using raw statistics like shots and shots on target may be misleading: a team that shoots a lot from long distance will have a lot of shots and maybe a lot of shots on target but most likely will not have a lot of goals. From a tactical perspective this could be interpreted as a team that fails to get decisive and clear goal scoring opportunities. A team however having a fewer shots but from more dangerous positions might perform better. Looking at Figure 3 which maps the likelihood of goals to the position from where a shot is taken by Liverpool player Mohamed Salah, we see that intuitively the closer we are to goal the better the chance is. XG enables us to profile teams players and determine their capacity to convert goal scoring chances into actual goals.

Another advantage of XG is the fact that they are less noisy than goals. Over the same period as the goals, the sample mean for XG is 1.58 (similar as the mean rate of goals scored of 1.5) while the sample variance is 0.84, which is smaller than the 1.22 variance from the Poisson distribution.

5.4 Event Data

Event data represents the specific ball actions that happen on the pitch recorded at specific time-frames. Those events include the time and positions of passes, shots and interceptions etc. Event data can be used to evaluate more intricate performance metrics like the value of passes [12] or the team's attacking playing style [13]. Indeed ball event data may give us insights into the way a team attacks, maybe by going through the wings more than down the middle or maybe by passing the ball many times around the penalty box or predominantly with via direct through balls. Figure 4 shows an example use of event data, where passes are mapped from passer to receiver. From such mappings, we can gather insight about the general team structure, formation, player roles and contributions.

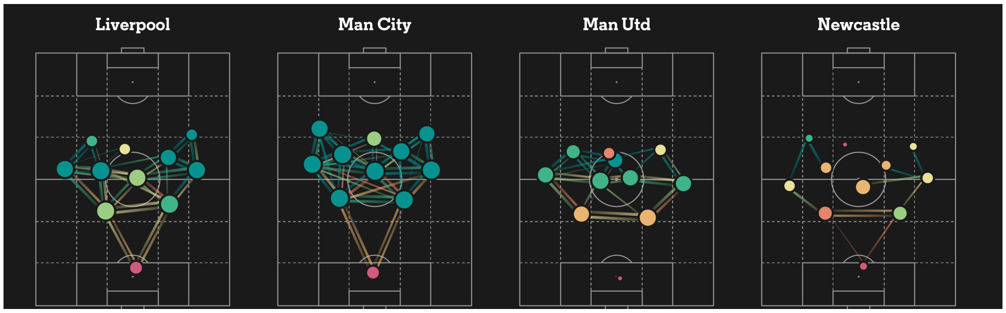


Figure 4: Selection of Premier League teams pass maps [14]

Further research has developed other metrics using event data, like Expected Assists which assigns credit to passers when the receiver shoots. By combining a predictive model's XG and the event data, one can back-propagate the contribution towards a goal-scoring opportunity back to the passes. This enables us to identify a creative player that creates shot opportunities for teammates. Other models like Expected Threat measure the offensive contribution any action on the ball leads to, so it includes ball carries and dribbles as well as passes. And finally off-ball scoring opportunities have been used to measure the likelihood a team scores using only statistics not related to shots. For instance data about the ball possession, number of tackles etc. could be used as input data for a model to predict goals [15]. Event data is more granular than statistics and refined metrics but they are less available (because companies collect and sell them to football clubs).

5.5 Tracking Data

Finally, tracking data is a record of each player and ball position on the field throughout the match. Tracking data, while it is bigger in volume and harder to structure, is much more rich in information. Tracking data can give a lot of insight in both the attack and the defensive style of play of any team. From an attacking perspective, it can give more context of players not holding the ball. A player's run without the ball can draw a defender out of position and open new lines of passes for its teammates without contributing with the ball. Tracking data can help identify players that by merely positioning themselves smartly will improve the team's chance of scoring. From a defensive standpoint, it will enable analysts to observe the positional quality of defenders. Pitch control is one of the potential use of tracking data [15] which shows the areas of the pitch one team has more control over, so an area where the ball will be contested more aggressively and risks being lost. Tracking data is even more granular than event but also more rare (fewer companies collect it since it requires more complex techniques).

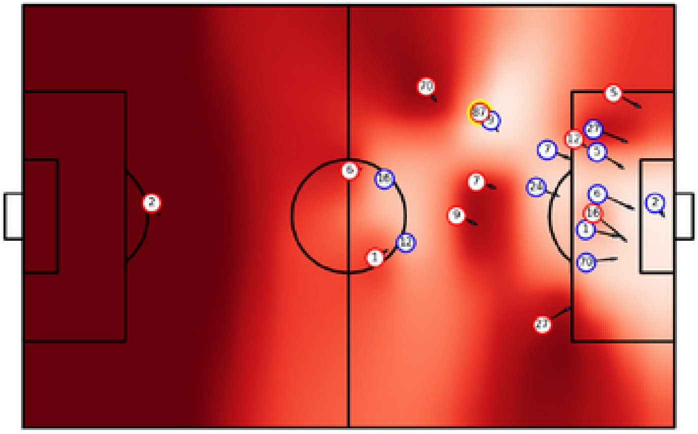


Figure 5: Pitch control [15]

6 Literature Review

In this section, we will explore some of the hallmarks of research applied to predicting football match outcomes. We can differentiate the research into two categories: the pre-game predictions and the live odds. Pre-game predictions are likelihoods computed prior to the beginning of the game such that only information about the past is used. Live odds on the other hand are the likelihoods computed while the game is happening such that information about the game is also used to evaluate the most probable winner.

6.1 Pre-grame Predictions

6.1.1 Statistical Models

Let us explore the statistical models that can be applied to evaluate the winner of a game. A statistical model is a mathematical model that relies on a set of statistical assumptions concerning the sample of data. In other words, we assume that the data follows a statistical distribution and evaluate the parameters of this distribution.

- A **Thurstone-Mosteller model** is designed to evaluate the probabilities of paired comparisons. The paired comparison in question is the opposition between two teams. A Thurstone-Mosteller model evaluates latent variables that stand for the performance of a team. A latent variable is not observed but rather evaluated by a model using other observed variables. The model assumes that the latent variables for describing the team performance are drawn independently from a normal distribution, with constant variance [16]. Let us define a few symbols: let M be the total number of matches where a total of T teams of are confronted. We will denote the latent continuous variables as $Y_{i,m}$ for team i's performance in match m. It follows that a team i wins over team j in match m if $Y_{i,m} - Y_{j,m} > d$ (and conversely) and the match ends with a draw if $|Y_{i,m} - Y_{j,m}| < d$ such that d is a threshold representing the chance of a draw. In football only the final score determines the winner, so the latent variable $Y_{i,m}$ for performance can be evaluated using the final score. It is drawn from a normal distribution with mean r_i and a variance constant across all teams: $Y_{i,m} \sim \mathcal{N}(r_i, \sigma^2)$. The model assumes that the performances of teams are independent so: $Y_{i,m} - Y_{j,m} \sim \mathcal{N}(r_i - r_j, 2\sigma^2)$. As we have seen in the section on refined metrics, the home team in football enjoys a performance boost which we can evaluate using a variable h and refer to as the home-field advantage. We can now compute the probabilities of each outcomes as follows:

$$P(Home_{ijm}) = P(Y_{i,m} - Y_{j,m} > d) = \Phi(\frac{(r_i + h) - r_j - d}{\sigma\sqrt{2}})$$

$$P(Away_{ijm}) = P(Y_{j,m} - Y_{i,m} > d) = \Phi(\frac{r_j - (r_i + h) - d}{\sigma\sqrt{2}})$$

$$P(Draw_{ijm}) = 1 - P(Home_{ijm}) - P(Away_{ijm})$$

where Φ represents the cumulative distribution function of the standard normal distribution. A Thurstone-Mosteller model uses a single observation and two variables per match in order to evaluate a total of $(T + 1)$ variables.

- A **Bradley-Terry model** is also designed to predict the outcome of pairwise comparisons. But it assumes that latent variables for describing the team performance are independently drawn from a logistic distribution, with constant scale factor [17]. Thus $Y_{i,m} \sim logistic(r_i, s)$ and $Y_{i,m} - Y_{j,m} \sim logistic(r_i - r_j, s)$. We can compute the probabilities of each outcomes as follows:

$$P(Home_{ijm}) = P(Y_{i,m} - Y_{j,m} > d) = \frac{1}{1 + exp(-\frac{(r_i+h)-r_j-d}{s})}$$

$$P(Away_{ijm}) = P(Y_{j,m} - Y_{i,m} > d) = \frac{1}{1 + exp(-\frac{r_j-(r_i+h)-d}{s})}$$

$$P(Draw_{ijm}) = 1 - P(Home_{ijm}) - P(Away_{ijm})$$

A Bradley-Terry model uses a single observation and two variables per match in order to evaluate a total of $(T+1)$ variables.

- A **Poisson model** is designed to predict the number of goals a team will score versus another. A Poisson model assumes that the number of goals scored in a game follows a Poisson distribution (as we suggested in the section on refined metrics). Maher [18] proposed that $G_{i,m}$ and $G_{j,m}$ be independent random variables for the goals scored by teams i and j in match m. We can denote $\lambda_{i,m}$ and $\lambda_{j,m}$ as the rate of $G_{i,m}$ and $G_{j,m}$ respectively such that $G_{i,m} \sim Pois(\lambda_{i,m})$. Given this we can compute the probability of a specific scoreline as follows:

$$P(G_{i,m} = x, G_{j,m} = y) = P(G_{i,m} = x)P(G_{j,m} = y) = \frac{\lambda_{i,m}^{x}}{x!} exp(-\lambda_{i,m}) \frac{\lambda_{j,m}^{y}}{y!} exp(-\lambda_{j,m})$$

To evaluate the probability of a team winning we can simply sum the probabilities of the scorelines that would represent that team winning.

$$P(Home_{ijm}) = \sum_{i,j=1}^{T} \sum_{m=1}^{M} P(G_{i,m} > G_{j,m}) = P(G_{i,m} = 1, G_{j,m} = 0) + P(G_{i,m} = 2, G_{j,m} = 1) + \ldots$$

$$P(Away_{ijm}) = \sum_{i,j=1}^{T} \sum_{m=1}^{M} P(G_{i,m} < G_{j,m}) = P(G_{i,m} = 0, G_{j,m} = 1) + P(G_{i,m} = 1, G_{j,m} = 2) + \ldots$$

$$P(Draw_{ijm}) = \sum_{i,j=1}^{T} \sum_{m=1}^{M} P(G_{i,m} = G_{j,m}) = P(G_{i,m} = 0, G_{j,m} = 0) + P(G_{i,m} = 1, G_{j,m} = 1) + \ldots$$

In the Maher paper, the scoring rates were assumed to take the following form: $\lambda_{i,m} = exp(c + (a_i + h) - d_j)$ and $\lambda_{j,m} = exp(c + a_j - (d_i + h))$ where a and d represent the attacking and defensive capabilities, h and c represent the home-field advantage and the constant scoring rate. A Poisson model uses two observations (the home and away goals scored) and six variables per match in order to evaluate a total of $(2T + 2)$ variables.

- **Dixon-Coles** is a variation of a Poisson model such that it models the dependence between the home and away scores. In a paper in 1997 [19], researchers hypothesized that the assumption of independence of the home score and away score does not hold for low-scoring games. Indeed looking back at Figure 2, one can see that the Poisson distribution with rate of the mean number of goals scored is more prone to error for low scores than high scores. As such, the authors proposed to adjust the probabilities score 0-0, 1-0, 0-1 and 1-1 as follows:

$$P(X_{i,m} = x, Y_{j,m} = y) = \tau(x,y)\frac{\lambda_{i,m}^{x}}{x!}exp(-\lambda_{i,m})\frac{\lambda_{j,m}^{y}}{y!}exp(-\lambda_{j,m})$$

where

$$\tau(x,y) = \left\{ \begin{array}{cc} 1-\rho\lambda_{i,m}\lambda_{j,m} & \text{if x = y = 0} \\ 1+\rho\lambda_{i,m} & \text{if x = 0, y = 1} \\ 1+\rho\lambda_{j,m} & \text{if x = 1, y = 0} \\ 1-\rho & \text{if x = y = 1} \\ 1 & \text{otherwise} \end{array} \right\}$$

A Dixon-Coles model uses two observations and six variables per match in order to evaluate a total of $(2T + 4)$.

For these three models, we have the formula for evaluating the likelihood of outcomes, but another step is necessary to evaluate the parameters used in those formulas. The method used in the papers referred to previously is maximum likelihood estimation (MLE) which refers to the estimation of a probability distribution parameters using observed data. Like the name suggest, we will maximize a likelihood function such that under the statistical assumptions, the observed data will be the most probable. In our case the observed data is the results of past football games.

$$L = \prod_{m=1}^{M} \prod_{i\in(1...T)} \prod_{j\in(1...T)} \prod_{R\in(H,D,A)} P(R_{ijm})^{x_{R_{i}jm}}$$

In the formula above, R_{ijm} represents the results between team i and j in game m and $x_{R_{i}jm}$ is equal to 1 if the result between team i and j in game m is R and 0 otherwise. In other words, the likelihood function is the product of all match outcomes predictions (i.e. the odds of the observed outcome). For instance for a sample game that ends with a draw, only $P(D)$ will be used in the likelihood function. The purpose of MLE is to find the parameters that maximize the formula, which in other words would mean that the predicted results of the historical events would as close as possible to the realized results.

With this MLE formula, the same importance is given to games that took place a very long time ago. Yet, teams tend to have very different play-styles and more importantly strengths from game to game and overall from year to year. This is due in part to changes in the squad due to injury, transfers or player performance as well as the coach. To account for this, we can attribute an importance factor to games such that recent games are more important to predict accurately than very old games. In other words, we force the MLE to evaluate parameters that reflect the recent past better [20]. Let us introduce a weight factor for game $m : w_m$, and raise the probability to that power in the likelihood formula:

$$L = \prod_{m=1}^{M} \prod_{i\in(1...T)} \prod_{j\in(1...T)} \prod_{R\in(H,D,A)} P(R_{ijm})^{x_{R_{ijm}} w_m}$$

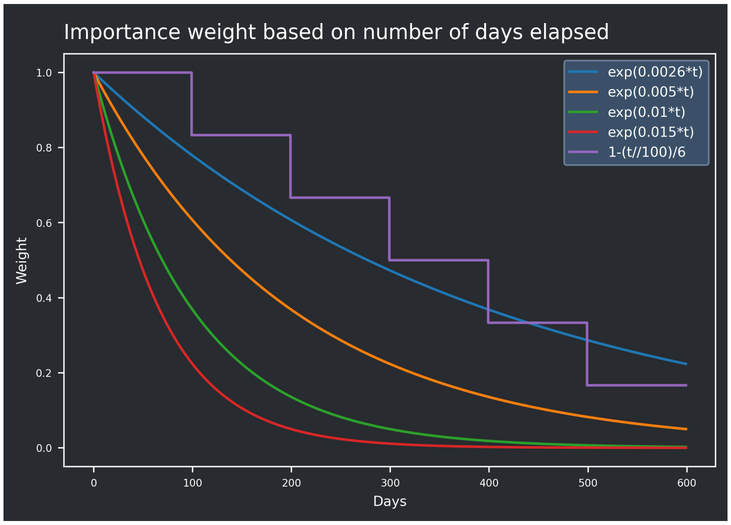


Figure 6: Example weight factor function that can be applied to statistical models

It follows that the smaller the weight factor the less important the game and subsequently the higher the probability will be scaled to, so its impact on the likelihood function will be minimized. For instance a game that occurred a long time ago, such that its weight factor is 0.2, is evaluated to have probability 0.1 under the current parameter estimation (this means the model does not predict it accurately). Thus the likelihood would be $0.1^{0.2} = 0.63$ which is much larger than the un-scaled original prediction. We can compute the weight factor in a number of ways: either using a continuous deprecation or step-wise. We can use the continuous formula : $exp(d * t)$ where d is the decay rate and t is the number of day that elapsed since the game occurred. Or we can use a step-wise function: $1 - (t//100)/6$ which removes a equal amount for at every multiple of 100 days elapsed. Several weight functions are graphed in Figure 6.

- A **Bayesian Model** can also be used to model football outcomes as was proposed in research from Baio-Blangiardo [21]. Maximizing the likelihood of observing the observed data is a frequentist approach to statistical modeling. The estimated parameters are in the form of a point estimation (with a confidence interval). In a Bayesian method however, the estimated parameters are in the form of a probability distribution. Parameters are considered random variables and we represent them as probability distributions to capture the uncertainty of their real value. This means that we can assign a probability value for each potential value of the parameters. This evaluation of parameter probability relies on the Bayes theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

where $P(\theta)$ is the prior distribution of our model parameters, which is an assumption made on the relationship between observed data and model parameters. $P(X|\theta)$ is the likelihood which indicates the fit between observed data and the prior. While this step is also done for frequentist methodology, for the bayesian methodology, this provides a probability for each value in the parameter space. $P(X)$ is the marginal distribution of the predictors. In other words, the probability of having observed the data given the parameter space. While the frequentist method solely relies on observations, a bayesian method also relies on the assumption of parameter distributions and refines that assumption with new data samples.

The model Baio-Blangiardo proposed is hierarchical, which assumes that model parameters are related because they come from a common distribution. This can be interpreted as a common distribution representing the team ratings from which is draw individual team strengths. For the Poisson model, we assumed that the Poisson variables for the scored goals are independent and for Dixon-Coles, we then correlated them. Yet with this hierarchical architecture, this correlation is already taken into account by their relationship (through the common distribution). The authors estimated the number of goals scored followed a similar

Poisson distribution to previous research: $\lambda_{i,m} = exp((a_i + h) - d_j)$. Where they assigned a prior distribution to the estimated variables in the model as follows:

$h \sim N(0, 0.0001)$ for the home-field advantage

$o \sim N(\mu_a, \tau_a)\ \forall|\mu_a| \sim N(0, 0.0001)\ \forall|\tau_a| \sim Gamma(0.1, 0.1)$ for the attacking rating

$d \sim N(\mu_d, \tau_d)\ \forall|\mu_d| \sim N(0, 0.0001)\ \forall|\tau_d| \sim Gamma(0.1, 0.1)$ for the defensive rating

- The **Elo Rating** was introduced in 1960 by a Hungarian-American chess master named Arpad Elo as a rating system to track the relative performance of players in zero-sum games. The rating system, named after him, has been used in Chess ever since. Much like previous models, the Elo rating evaluates team ratings based the outcomes of games (not accounting underlying performance) but it accounts for the opponent's strength. Let R_h and R_a be the ratings of the home and away teams. We can compute the expected probability of each team winning as follows:

$$P(Home) = \frac{1}{1 + 10^{\frac{R_a - R_h}{400}}}$$

$$P(Away) = \frac{1}{1 + 10^{\frac{R_h - R_a}{400}}}$$

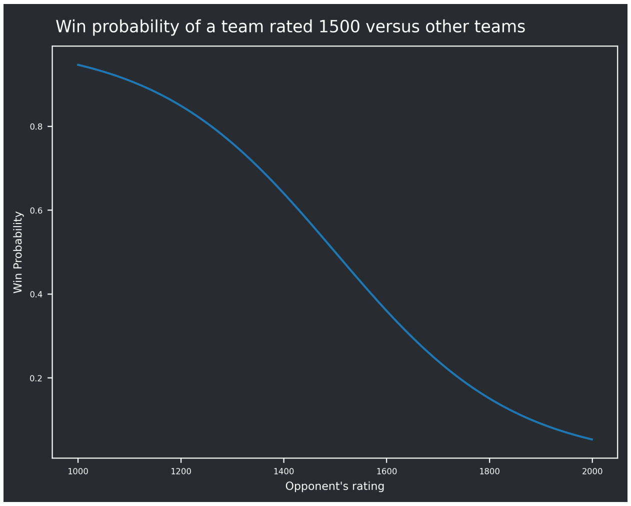


Figure 7: Win probability for different oppositions

Note that the original system cannot predict a draw. In the framework proposed by Elo, ratings are initially set to a default value and as players compete, their ratings are updated with the following formula:

$$R'_h = R_h + K(S_h - P(Home))$$

In the formula, K is a hyper-parameter denoting the update rate of ratings: it scales the update such that K is the maximum amount of rating one can gain or lose at once. In the formula, S_h is one if the Home team won and 0 if they lost. Here we can also account for games that end with a draw. By setting S_h to 0.5 (or any value between 0 and 1) for draws, we can scale ratings such that a draw will be accounted as an under-performance by the favorite. Even though this model cannot predict the probability of a draw, it can still account for them in the rating update. Thus, we could apply this system to football (where draws can happen) and either infer a zero probability of draws for each game (which would hinder the predictive performance) or smooth the odds spread.

6.1.2 Machine Learning Models

The statistical models we have seen up to now estimate a team performance using raw results from games. Yet, using Machine Learning models to evaluate the likelihood of each outcome would enable more flexibility for the inputs of the model. While we could train a ML model solely on historical scores, we can also consider using additional data (mentioned in the Data section) to gain an edge and reduce the noise in previous models. One paper has done considerable research into the data that is valuable to include into different models to achieve accurate results [22]. Let k be a hyper-parameter for the number of historical games to consider, the authors suggested to perform the following feature engineering:

- Compute the average corners, shots on target and goals over the past k games to capture more detail about the team performance.

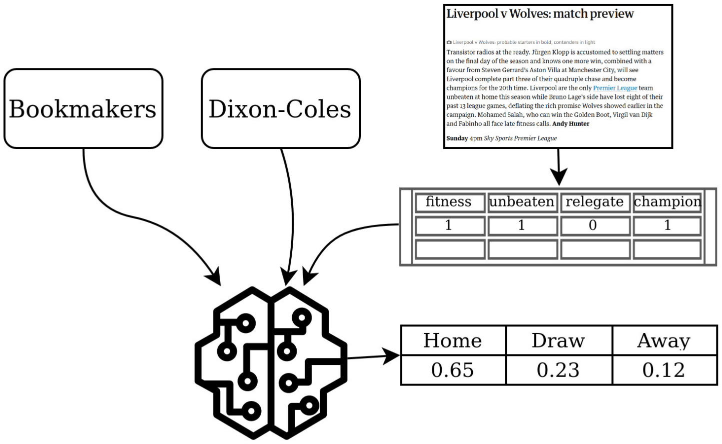
- Generate a streak value with formula $\delta_j = \sum_{p=j-k}^{j-1} \frac{r_p}{3k}$ where r_p is the number of points won from game p. This encapsulates recent team form and ability of a team to concretely win games.

- Compute a weighted streak feature with formula $\omega_j = \sum_{p=j-k}^{j-1} \frac{p-(j-k-1)r_p}{3k(k+1)}$. This encapsulates recent team form with greater emphasis on most recent games.

- Incorporate granular team ratings. The popular football game franchise Fifa releases a new edition of its game each year and attributes to each team a new rating based on the players transferred in and out. Using a public database of historic Fifa ratings, the authors used the attack, midfield, defense and overall rating of teams to account for team specific strengths and weaknesses. Note that this rating is static for a given season and does not takes into account team performance during a season.

- Compute a form feature. By attributing a rating that changes based on the performance during a season, the model can capture the relative strengths that is fluid enough to handle changes in performance throughout the season. The initial form rating given to each team in a season is 1 and is reset at the beginning of each season. Similarly to an Elo rating, this form rating transfers "rating points" between teams based on the results of games. The formula is the following for team α to win is : $\xi_j^\alpha = \xi_{j-1}^\alpha + \gamma\xi_{j-1}^\beta$ and for a draw: : $\xi_j^\alpha = \xi_{j-1}^\alpha - \gamma(\xi_{j-1}^\alpha - \xi_{j-1}^\beta)$ for γ a hyper-parameter.

Let us explore a model using a creative type of input and architecture. Using a match preview from a popular British newspaper, the authors built a Natural Language Processing pipeline to predict the outcome of a game [23]. The authors propose a novel way to ensemble the expertise of bookmaker odds, Dixon-Coles, and a text-vectorization model into a domain expert. Firstly, one should look at the articles to understand what they are made of. The articles present the likely starting team, a short summary of the recent form and the importance of the game for each team, and information about injuries and suspensions. The paper proposes to only use the short summary and processes it using a Count Vectorizer. Count Vectorizer is a numerical representation of text where each word is counted and the order and context is ignored. To represent text using only the count of occurrence of word, one builds an array of size m equal to the total number of unique words in the training set. The size of the corpus can be minimized to contain only important words and thus limit the array dimension by ignoring words that provide no context. Then for each sample, one counts the occurrence of words and sets the value of the array to that number at the index of the word. This text representation is then given to Random Forest algorithm which predicts three numbers representing the home and away win probability and the draw probability. Finally, they train an ensemble model taking as inputs the predicted odds of the bookmakers, a Dixon-Coles model and the random forest trained on text. An ensemble model is an ML method that trains a single expert model from the outputs of multiple noisy models, called weak learners. The author's hypothesis is that bookmaker odds are noisy due to the adjustments made to the odds to minimize the risks, Dixon-Coles is also noisy because it only account for the past result and not the context for the next game. So combining the three would capture information about both the context and the team strengths. The overarching idea is that unique games like derbies (games between rivals), relegation or title race games will have a different vocabulary which will provide context into how the upcoming game will unravel.



Liverpool v Wolves: match preview

Liverpool v Wolves: probable starters in bold, contenders in light

Transistor radios at the ready. Jürgen Klopp is accustomed to settling matters on the final day of the season and knows one more win, combined with a favour from Steven Gerrard's Aston Villa at Manchester City, will see Liverpool complete part three of their quadruple chase and become champions for the 20th time. Liverpool are the only Premier League team unbeaten at home this season while Bruno Lage's side have lost eight of their past 13 league games, deflating the rich promise Wolves showed earlier in the campaign. Mohamed Salah, who can win the Golden Boot, Virgil van Dijk and Fabinho all face late fitness calls. **Andy Hunter**

Sunday 4pm *Sky Sports Premier League*

fitness	unbeaten	relegate	champion
1	1	0	1

Home	Draw	Away
0.65	0.23	0.12

Figure 8: Model architecture

6.2 Live odds

Live odds are used by betting companies to propose betting opportunities for its clients during the game. This offers a better user experience for clients to continue engaging with wagers during the game. But it also presents a risk with respect to the synchronization of the odds with real life. Indeed if the update of betting odds is slightly delayed, a client could theoretically place a bet on an event for which he knows the outcome. This is especially problematic if the data necessary for the model to run takes time to be collected (for instance event or tracking data) or the model itself requires a lot of computing power.

Live odds can be summarized by the prediction of the results, conditioned on the knowledge of the current state space. The state space is chosen by the modelers. For instance, research was conducted to evaluate the live odds with a Bayesian model using a combination of statistics, event and tracking data [24]. In this paper, the authors divide the game into 100 time frames and used a model to predict at time-frame (t), the number of goals scored by the end of time-frame (t+1). Even though a football game lasts for 90 minutes, stoppage time is added at the end of each half to account for minutes that were lost due to fouls, injuries or goal celebrations. The authors account for the variable game lengths by dividing the game into 100 time-frames (50 for each half). They represented the state space S_t with contextual features about the game (within the past t time-frames) as well as more high level historical data. Indeed the model takes as input: relative team strength in the form of Elo ratings which accounts for high level team dynamics. It also uses contextual feature like the percent of game time played and the scoreline, the number of yellow and red cards, the number of goal scoring opportunities. The authors also used very advanced features derived from event and tracking data: number of attacking passes (which is the passes attempted within 30 meters of the opponent goal) and the rolling average Expected Threat over the last 12 time frames. The model assumes that the expected number of goals that the home and away team will score are independent Poisson distributions:

$$P(G_{h,t+1} = x, G_{a,t+1} = y|S_t) = P(G_{h,t+1} = x|S_t)P(G_{a,t+1} = y|S_t)$$

One interesting finding is that the feature importance varies throughout the game. At the beginning of the game, the XT feature is non informative yet it peaks at two instance: slightly before the end of the first half and right at the end of the second half. Which shows the use of very dense data like tracking data.

7 Implementation

Modeling football outcomes is synonymous with a classification problem. The classification in question will be that of categorizing a football match into one of three possible classes: the home team winning, a draw or the away team winning. In our case where we combine a football model with a betting strategy, knowing the relative probabilities of each outcome is valuable. Thus the model will be a multi-class classifier where each outcome is attributed a number that represents the probability or the likelihood of the outcomes to occur.

The task of modeling football outcomes can also be tackled with a regression model. This is the approach of the statistical models we discussed in the literature review. These models estimated a team's strength based on different assumptions. This is a regressions step where a continuous rating is attributed to each individual team. These approaches used a formula to translate these ratings to a probability distribution.

7.1 Evaluation

There is a multitude of metrics that can be used to evaluate a classification model. The most widely used (and most interpretable) is the accuracy. Accuracy is a percentage of the data that was correctly classified. Mathematically, it is the number of correct classifications c divided by the total number of samples M:

$$Accuracy = \frac{c}{M}$$

We can and should use other metric in order to derive a better understanding of the model performance and its weakness, mainly to understand the type of data sample it models accurately and the type it fails for. Being able to understand and explain the decisions a model makes (in other words its outputs) is crucial to improve it and also to rely on it comfortably when making decisions with it. Choosing the metric to maximize and to evaluate a model with is important in a modeling problem. There is not a single metric to solve all modeling problems, so using the correct metric that precisely applies to the problem at hand is an important design decision. Accuracy is a metric that accounts for the model's maximum prediction, thus it should be the metric used in settings where only the maximum class prediction is taken into account. In our case, we want to know the relative probabilities of all the outcomes of a game, thus we should consider a metric that evaluates the fit or the similarity between all the predicted likelihoods and all the actual outcomes. There exists a score function for probabilistic functions called the Brier Score. It computes the average squared error of all outcomes. Let p_i be the predicted probability for outcome i and r_i be the actual value of that outcome, then we have the Brier Score is computed as follows:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$$

In the context of football modeling, the outcomes are correlated and ordered. In a traditional classification task, each class is independent. Yet for some tasks, the classes are correlated, such that one is more similar to others. For football, the outcome of the home team winning is more similar to a draw than to an away win. Looking at the score lines will be more intuitive, the goal difference for the home team to win needs to be greater than zero while for a draw it needs to be equal to zero which is closer to a negative goal difference needed for an away win. This is where the Ranked Probability Score is useful. It accounts for the similarity of outcomes when measuring probabilistic models by measuring the fit of the cumulative distributions of the prediction and the outcomes using the formula:

$$RPS = \frac{1}{n-1}\sum_{i=1}^{n}(\sum_{j=1}^{i}p_j - \sum_{j=1}^{i}r_j)^2$$

In our case, $n = 3$, because there are only three possible outcomes. So to interpret this formula literally, it computes the gradual difference of the cumulative probabilities of event (i.e. first the home win, then the home win and the draw and finally all three events) with the sum of cumulative outcomes. The RPS is a score between 0 and 1 where a low value indicates a close match between the probability distribution and the results while a high values indicates that the two are very different.

7.2 Back-testing

Having explained the evaluation metrics used, I will in this section present the evaluation framework. The models were trained using data ranging from the 2017-18 season to the 2020-21 season. All the models were training using games results data from online news media FiveThirtyEight [3]. Then each model was evaluated using the entire 2021-22 season. The method used is called back-testing where for each game-week, the model is trained on all available historical data and makes a prediction for the upcoming game-week. A game-week is the set of all games to be played within the coming calendar week. Football seasons are scheduled at the beginning of the year such that every team plays every other week. This means that throughout the season, teams always have played the same number of games. Note that a few exceptions make the previous statement partly true: some league games are postponed and replaced with games from other competitions and Covid-19 was more recently a very frequent cause for postponed games since leagues needed to avoid teams infecting each others. Thus on average a game-week is made of 10 games (each of the 20 teams playing once) with a few exceptions where there are fewer or more than 10.

In the English Premier League, there are 38 game-weeks so throughout the back-testing, the model was trained 38 times and made 38 batches of predictions. For each game-week, I saved the batch of predictions such that I can compute season-wide metrics. The advantage of doing back-testing over a method of training the model once and predicting on the entire season at once is that we can measure the model's capacity to account for the season and team trends. This approach is a simulation to a real use of the model for betting during a season.

7.3 Models

In this section I will describe the models I implemented and explain the coding adaptation needed to make them work in practice. I will explore their results and detail the hyper-parameters I used.

[3] https://data.fivethirtyeight.com/soccer-spi

7.3.1 Baseline Models

Prior to diving into the depths of model building, we should build very basic, if not simplistic models and measure their performance. Having a few set of models that behave differently in an interpretable manner will enable us to compare any model built later with these baselines and accept or reject any one of them.

The dataset is imbalanced in favor of home wins. Like we already discussed, the football team playing at home benefits from a home field advantage which makes them more likely to score so more likely to win. Between seasons 2017-18 and seasons 2021-22, a total of 1900 games were played for the EPL. A total of 833 were won by the home team (so 43.8% of the dataset), 433 were draws (22.8% of the dataset) and 634 were won by the away team (33.3%). A typical dummy model to start with is one that predicts the most frequent class, so in our case a home team win. We can also build dummy classifiers that always predict a draw or an away win. Another typical baseline model is one that predicts a random class. Given that the evaluation metric we use accounts for predicted odds fit, we should make a baseline model that outputs random odds (that sum to one). In addition, we can also look at the results of other prediction models. For instance, bookmaker odds can be converted into implied odds using the formula we mentioned earlier.

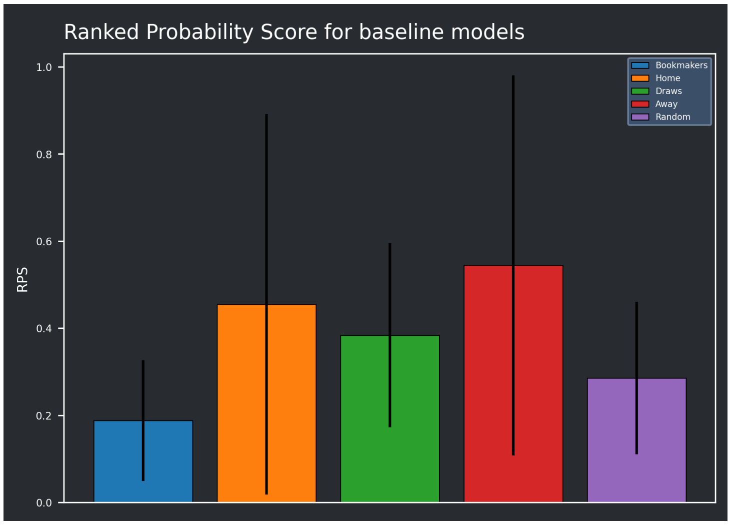


Figure 9: Season wide mean Ranked Probability Score for Baseline models (histogram) along with standard deviation (bars)

These models are starting points and their evaluation metrics are the targets that new models should beat. Indeed, if no new model can beat either a baseline or a benchmark model, then we should use those models to make decision (or no model at all). It comes to no surprise that the mean RPS for Bookmaker odds is the lowest. The bookmaker odds are not perfect, though. A RPS of 0.189 presents an opportunity for improvement. It is more surprising however that random odds outperform the biased models towards home and away wins or draws. This could be explained by the standard deviation for the biased models which is the highest.

7.3.2 Statistical Models

As mentioned in the Literature Review, to fit the statistical models, we need to maximize the likelihood function. I used the python library *scipy.optimize* to perform optimization where the objective function is the likelihood function. I initialized the parameters of the statistical distributions randomly using a uniform values between 0 and 1. Yet I imposed a constraint on the optimization to ensure reproducibility. Indeed, there are many possible solutions that can minimize the likelihood function and we would like to have parameters that are similar if not identical when reproducing numerous trials. Thus, we force the estimated parameters to sum to the total number of parameters T:

$$\sum_{i=1}^{T} a_i = \sum_{i=1}^{T} d_i = T$$

Another constraint I impose is to bound the ratings within a range: $r \in (0, 3)$. I also limit the number of optimization iterations that can be attempted. As I mentioned in the Back-testing section, I am training each model 38 times, so to limit the total running time needed, I set the maximum number of iterations to 100. To ensure that those 100 iterations are enough to find a plausible solution, I decided to reuse the previous game-week's parameters as the initial parameters for the next MLE. In other words, I did not re-initialise the parameters randomly for each new game-week and estimate parameters from scratch (which could have been very time consuming). Instead I am using the information learned from past MLE and am fine-tuning the parameters each game-week using the new available data. Note that this is not especially ideal since I run into the risk of stumbling into a local minima and not being able to escape it after. One other implementation adaptation I made was to minimizing the negative log likelihood function instead of a maximizing the likelihood function. The reason being that the likelihood formula involves the product of many small numbers: we deal with probabilities (so values between zero and one). The risk for a computer implementation is to run into floating point precision (also known as underflow) where the space needed to represent the decimals of a number is too large for the computer causing either imprecision due to rounding/truncating or outright errors. Thus applying a logarithmic function to the likelihood function converts the products into sums which avoids the risk of underflows. Since the Python library only has a minimization functionality, we can then minimize the negative of the log-likelihood, which is equivalent as maximizing the original.

In this implementation, design choices like the maximum number of iterations and the fact that past parameters are refined with future estimations could impact the model performance at the benefit of ease of use. The only hyper-parameter to choose is the weight factor applied to the historical data. To find the optimal weight decay factor, one has to train models with different weight values and select the one that performs best.

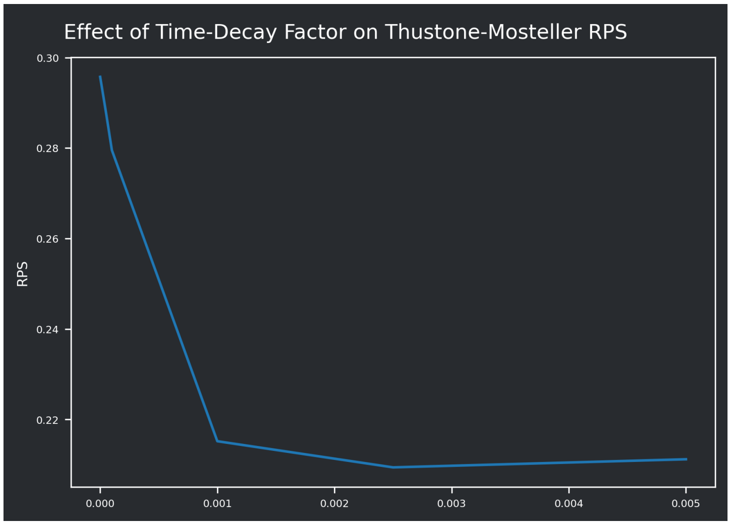


Figure 10: Graph of model performance (RPS) plotted for different hyper-parameters

For a Thurstone-Mosteller model for instance, I find that the time-decay factor $d = 0.0025$ in formula mentioned earlier: $exp(d \times t)$.

Thurstone-Mosteller	Bradley-Terry	Poisson	Dixon Coles
$exp(0.0025 \times t)$	$exp(0.0025 \times t)$	$exp(0.001 \times t)$	$exp(0.001 \times t)$

Table 2: Optimal hyper-parameters found

Interestingly, the Thurstone-Mosteller and Bradley-Terry models compare similarly with RPS score of 0.2094 and 0.2108 respectively. This suggests that the assumption for the

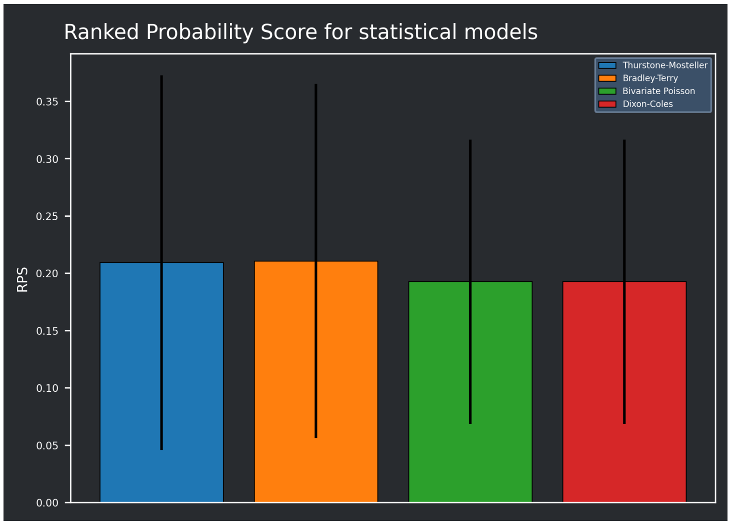


Figure 11: Season wide mean Ranked Probability Score for best Statistical models (histogram) along with standard deviation (bars)

statistical distribution that the team ratings follow is lightly in favor of a Normal distribution. While the Poisson and Dixon-Coles models perform similarly with RPS score of 0.1927 and 0.1927 respectively. The models that estimate two parameters per team (Dixon-Coles and Poisson) do however, perform better than the models estimating one general rating.

7.3.3 Probabilistic Programming

To fit a Bayesian model, I used Python library *PyMC3*, and I estimated goals scored as Poisson variables as follows:

$$\lambda_{i,m} = exp(c + (a_i + h) - d_j)$$

$$\lambda_{j,m} = exp(c + a_j - (d_i + h))$$

With prior distributions:

$c \sim N(0, 1)$ which is the model intercept (or the default scoring rate).
$h \sim N(0, 1)$ which is the home-field advantage.
$o \sim N(0, \sigma_o)\ \forall|\sigma_o| \sim N(0, 2)$ which is the offensive rating.
$d \sim N(0, \sigma_d)\ \forall|\sigma_d| \sim N(0, 2)$ which is the defensive rating.

Notice that I used different priors and that the values for the variance I assign are much larger than those reported in the research paper. This is a way to program the uncertainty of my knowledge of the prior distributions. Much like in the paper and in previous models, I imposed a constraint on the sum of variables o and d for reproducibility:

$$\sum_{i=1}^{T} a_i = \sum_{i=1}^{T} d_i = 0$$

During the training phase of the probabilistic model, the package finds the optimal probability distribution of parameters (under the prior distributions defined). It does so by sampling data points from the training set and gradually updating the prior distributions. An advantage of this Bayesian implementation using *PyMC3* over the frequentist implementation with *scipy.minimze* is that the observed data can be decimal numbers. Indeed, previously when modeling goals scored using Poisson random variables, the observed variables needed to be integers. Yet, with this package, we could model surrogate variables that are decimal numbers instead. We could use Expected Goals, discussed in the Refined Statistics section, for instance.

	Goal Scored	Expected Goals
RPS	0.1943	0.1987

Table 3: Model Results

7.3.4 Elo

For the Elo rating model I built, teams are initially given a rating of 1500. Since I am only considering Premier League games, every season 3 teams are promoted (and 3 are relegated). This means that potentially 3 teams that were never rated in the system before will be playing games weekly. Thus, in my implementation, I attributed an initial rating of 1350 to teams that are promoted. Giving them a rating below the 1500 I initially gave to all teams is a design choice made when considering that the general level of the Championship league (the league below the PL) is below that of the PL. In addition, to include the concept of home teams having an advantage over their opponent, I adapted the formulas from the original rating system as follows:

$$P(Home) = \frac{1}{1 + 10^{\frac{R_a - (R_h + h)}{400}}}$$

$$P(Away) = \frac{1}{1 + 10^{\frac{(R_h+h)-R_a}{400}}}$$

Another design choice I made was to smooth the home and away win probabilities. Like I mentioned earlier, the home win is closer to a draw than to an away win. Thus we could attribute odds for a draw based proportionally on the individual team strengths. The formula I used is the Hubbert Curve which is a symmetric logistic distribution curve:

$$P(Draw) = \frac{e^{\frac{r_h - r_a}{100}}}{(1 + e^{\frac{r_h - r_a}{100}})^2}$$

To put this formula into perspective, it assigns a probability of a draw to 25% when teams are equally rated. The maximum draw probability is 25% and the probability of draw decreases with a greater difference between teams. Note when calibrating the draw formula, I found that not including the home-field advantage lead to optimal results.

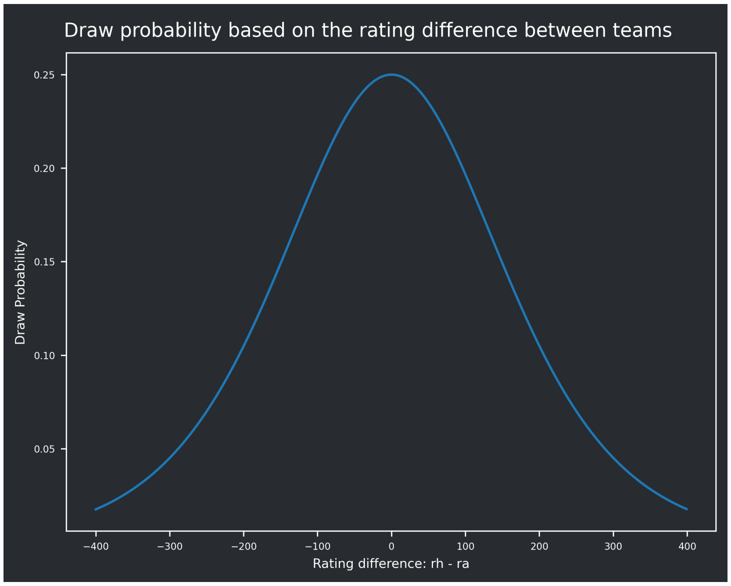


Figure 12: Hubbert curve plotted for different team ratings differences

Then I smoothed the home and away win probabilities to ensure that the sum of predicted probabilities is equal to 1. The formula takes away a proportional amount from the home and away probabilities as follows:

$$P(Home)' = P(Home) - P(Home) \times P(Draw)$$

$$P(Away)' = P(Away) - P(Away) \times P(Draw)$$

Another adaptation I made to the original Elo ranking system is to account for performance. While for the original rating system, only the results matter, in my implementation, I increased the rating update if the goal difference in the game is big. A great goal difference is an indicator of one team being much worse than the other. So if one team scores more than 2 goals than its opponent, then its rating should increase more after the game and reversely if it concedes more than 2 goals than its opponent. Indeed, I needed to penalize the defensive performances as well as reward the attacking performances in order to keep the symmetric nature of the rating update. The new rating update formula is the following:

$$R_h' = R_h + K(S_h - P(Home)) + \left\{ \begin{array}{cc} +\delta & \text{if } \mathrm{G}_h - G_a > 2 \\ -\delta & \text{if } \mathrm{G}_a - G_h > 2 \\ 0 & \text{if } \mathrm{G}_h = G_a \end{array} \right\}$$

$$R_a' = R_a + K(S_a - P(Away)) + \left\{ \begin{array}{cc} +\delta & \text{if } \mathrm{G}_a - G_h > 2 \\ -\delta & \text{if } \mathrm{G}_h - G_a > 2 \\ 0 & \text{if } \mathrm{G}_h = G_a \end{array} \right\}$$

In an Elo rating algorithm, the hyper-parameters to select are the home-field advantage which I found to be optimal for $h = 50$. To put into perspective, for two teams with equal ratings of 1500 confronting, without a home-field advantage, the home team would win with probability 50% compared to 57.1% with the home field advantage (without accounting for draws). The value K for the amount of rating transferred is also a hyper-parameter that I assign to 20. The value of δ for the Elo rating transferred in case of good and poor performance is another hyper-parameter to evaluate. I found that $\delta = 5$ is optimal. To interpret this value, if a team performs really well and scores at least 3 more goals than its opponent, it stands to gain and extra 25% of the maximum amount of rating it would stand to gain (in other words $\frac{5}{20}$), which is a small upgrade that does not overshadow the result of the game. Table 4 shows the gradual metric improvements made due to each incremental model change.

	Original	HFA	HFA+Draw	HFA+Draw+Performance
RPS	0.2165	0.2108	0.1988	0.1962

Table 4: Model Results

7.4 Betting Strategy

The first intuition one might have is to bet on the favored outcome, the one with highest likelihood of occurring. Yet in the context of a betting strategy, we don't want to maximize the number of correct predictions, we want to maximize profits. To do so one has to carefully select the bet according to the betting odds. Indeed the betting companies also are aware of which team is most likely to win and adjust their odds accordingly. Then one has to select the optimal bet amount to maximize profit and minimize the risk of ending with an empty balance.

The betting strategy is evaluated using betting odds data from Football.co.uk [4] which gathers closing betting odds from multiple betting sites. The odds from a single bookmaker were used throughout the season, those from Bet365. The outcome predictions used to evaluate the betting strategy were those of the model that performed best so a Dixon-Coles model.

7.4.1 Kelly criterion

The first betting strategy I implemented is based on the expected value of outcomes and the Kelly criterion. To maximize profit we will compute a simple statistic called the expectation. In statistics, the expectation of a random variable is the weighted average of all the possible outcomes. The expectation of a bet is the average payoff one can expect to receive. We know that the payoff is the multiplication of the wager with the betting odd which we will weight using the inferred likelihood of the outcomes. Let us express it mathematically for event an A (where A $\in \{Home, Draw, Away\}$) such that w_A is the wager placed on that outcome, o_A is the bookmaker odds for that outcome and $\overline{A}$ is the set of all outcomes excluding A and finally $P(A)$ the estimated probability of event A occurring:

$$EV(A) = w_A \times o_A \times P(A) + \sum_{e \in \overline{A}} P(e) \times 0 - w_A$$

To simplify this, we can simply let the payoff be the decimal betting odds (i.e. we would stake a single unit of money). The formula can be further simplified such that the expected value is equal to the odds of the event times the prediction of this event (because when betting on one event the payoff from the other events happening will be nill).

[4] https://www.football-data.co.uk/englandm.php

$$EV(A) = o_A \times P(A) - 1$$

If the expected value for a bet is greater than the total wager, then it means that given the odds and the outcome prediction, the bet would yield a profit on average. However if the expected value for a bet is lower than the overal wager, then the bet would would lose money on average.

Now that we have identified a bet that we expect will yield a positive payoff, we have to determine the amount to stake on it. Betting the entire balance on a single event might seem appealing since if it occurs the payoff will be the maximum possible but if it does not, one does not have new betting opportunities. Managing the account balance for betting is about acknowledging the randomness of the outcomes and allowing oneself to have to margin for error. Some strategies have been elaborated to maximize wealth growth over time for investing and could be used in the context of gambling. The Kelly Criterion is one mathematical formula used to determine the optimal amount of money to put into a single bet. The strength of the Kelly Criterion is the fact that it takes into account the likelihood of the event as well as the payoff. Thus it is a formula that balances the uncertainty and randomness (or the risk) with the payoff (i.e. the reward). The formula is the following:

$$K(A) = P(A) - \frac{1 - P(A)}{o_A - 1}$$

Once we identified an event A that has a positive expected value, we can evaluate the amount to stake. If $K(A)$ is positive, then we should wager $K(A)$ times the account balance on the event A. K is a value between 0 and 1 and represents the percentage of the account balance to stake on bets. For instance when Liverpool faced Wolverhampton at home, the outcome with maximum expected value, given the closing betting odds and a Dixon-Coles prediction model, was the draw.

	Liverpool win	Draw	Wolverhampton win
Odds	1.14	8.5	15.0
Predictions	67.5%	20.7%	11.7%
Expected Value	0.769981	1.758363	1.752104

Table 5: Example event: Liverpool vs. Wolverhampton (22/05/2022)

The Kelly Criterion for the draw event is the maximum of the curve shown in Figure 13. Intuitively, the curve shows that when placing a wager amount of 0 on that event, the expected account balance after the event will be the initial account balance. Yet when placing a large wager, the risk incurred increases and the expected account balance after the event will be 0. The optimal bet derived from this curve is 10.1% of the account balance

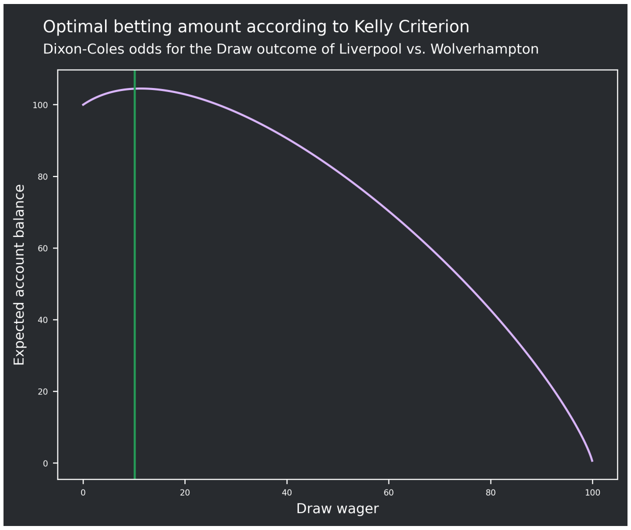


Figure 13: Example of the optimal bet amount to place using Kelly Criterion

To evaluate the combination of the prediction model along with the betting framework, we run inference on a whole batch of football games, in our case season 2021-22, and save the predictions. For each game in the season, we compute the expected value of each outcome and for the maximum one, we determine the bet amount with the Kelly criterion. Then we evaluate the payoff or the loss based on the observed outcome.

7.4.2 Kelly criterion optimization

The formula for the Kelly criterion I wrote before actually only applies for events where only two outcomes are possible. Indeed, it account for the $P(A)$ and $1 - P(A)$ and in the case of football outcomes where there are actually 3 outcomes. This simplified formula does not work especially given that per event, only a single bet is place (on the highest expected value event). Thus I explored the Kelly criterion formula when generalized to a multiple outcome case and will develop the mathematics its relies on [25]. For a two outcome problem, we want to maximize

$$r = (1 + K(A)o_a)^p \times (1 - K(A))^{1-p}$$

which is the growth rate when betting on outcome A with betting odds o_A and probability p. This is equivalent to finding the maximum of the logarithm of this formula:

$$E = log(r) = plog(1 + K(A)o_A) + (1 - p)log(1 - K(A))$$

The maximum of this formula is given by the value for which its derivative equals zero:

$$\frac{dE}{dK(A)} = \frac{po_A}{1 + K(A)^* o_A} + \frac{-(1-p)}{1 - K(A)^*} = 0$$

which when simplified gives the Kelly Criterion formula detailed before:

$$K(A)^* = p - \frac{1-p}{o_A}$$

We can extend this formula to multi-outcome event as follows:

$$r = \prod_{i \in (h,d,a)} (1 + K_i \times o_i - \sum_{j \in (h,d,a)} K_j)^{p_i}$$

Which literally is the product of the account balance in case of a win raised to the power of the estimated probability of a victorious bet. The account balance in case of a victorious bet is equal to the initial account balance subtracted by the sum of all the bets and with the winnings for the bet added.

$$E = log(r) = \sum_{i \in (h,d,a)} p_i log(1 + K_i \times o_i - \sum_{j \in (h,d,a)} K_j)$$

Notice that in this formula, we can place bets on multiple outcomes (represented by the variable K_i). In this formula, we represent the initial account balance to be 1 unit of money. It is out of the scope of this paper to develop the formula for the derivative $\frac{dE}{K_i}$, since I implemented the maximization of the equation using optimization with Python library *scipy.optimize*. The following constraints should be added to the optimization model:

$$\sum_{j \in (h,d,a)} K_j \leq balance$$

This enforces the total wagered amount on a single event to not exceed the amount available.

$$K_j \geq 0 \forall j \in (h, d, a)$$

This enforces that each wagered amount to be positive.

$$K_j \notin]0; 0.1[\forall j \in (h, d, a)$$

This constraint is necessary since in practice betting companies have a minimum wager policy of 10 cents. So a bet is either 0 or greater than or equal to 10 cents.

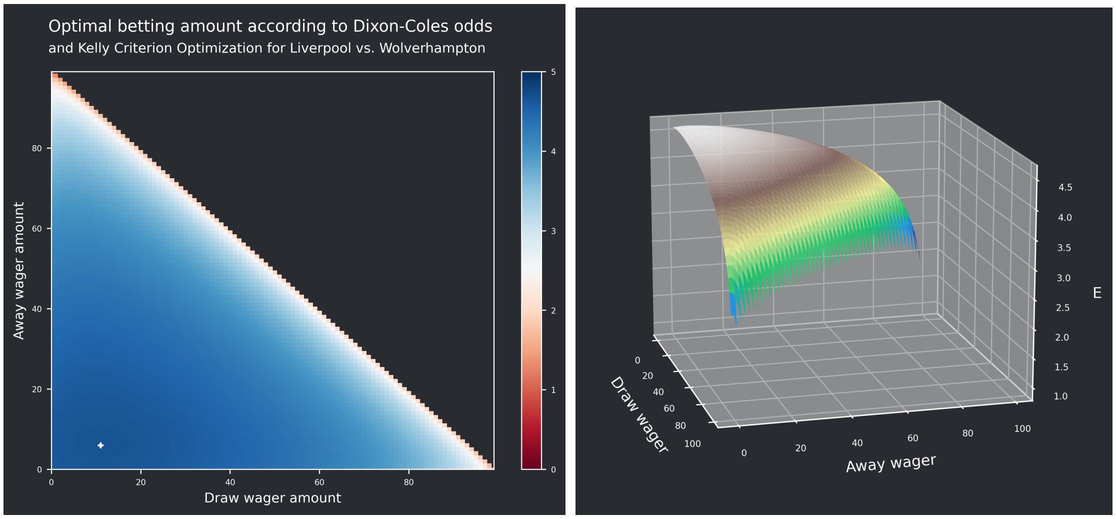


Figure 14: Example of the optimal bet amount to place using Kelly Criterion Optimization

In Figure 13, we saw that optimizing the payoff was equivalent to finding the maximum of a curve. For Kelly Criterion Optimization, we have to maximize function E with 3 parameters: the bets on home, draw and away win. Figure 15 shows the values for this function such that the bet on the home win outcome is fixed at 0 (to facilitate the graphing and interpretation). In fact, the home win outcome is ignored because it is set to zero when optimized which we could have anticipated given it had the lowest expected value from Table 5. The left image shows that when the sum of the wagers is greater than 0, the value of E is below zero. This means that the constraint I used in my optimization model is redundant but I will keep it to accelerate the optimization. The maximum value of the function E is shown with the white cross when $K_d = 10.1\%$ and $K_a = 6\%$. The image on the right shows that the curve E (at least for this event) is quite smooth and we can expect to not be stuck within local minimas during optimization.

To evaluate the combination of the prediction model along with the betting framework, we run inference on a whole batch of football games, in our case season 2021-22, and save the predictions. For each game in the season, we solve the Kelly Criterion optimization model described to determine the combination of bet amount to place. Then we evaluate the payoff or the loss based on the observed outcome.

7.5 Results

This sections will discuss the results obtain and offer insights. Firstly, it is important to note, that among the model from the literature that I reproduced, none outperforms the bookmakers odds. The bookmakers record a Ranked Probability Score of 0.189 over the 2021-22 season. The closest is the Dixon-Coles model with a RPS of 0.1927. Overall, more complex models that estimated more parameters performed better, which indicates that predicting football outcomes is an intricate problem.

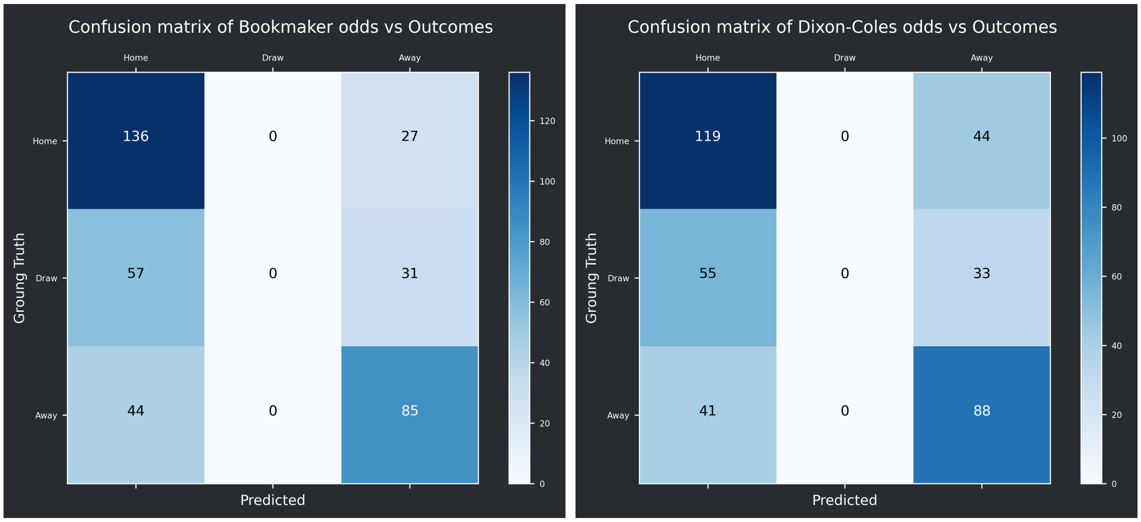


Figure 15: Model results comparison

Interestingly, neither model predicted a single draw throughout the entire season, despite draws occurring 88 times over said season (so 23.1% of games). The main difference between the two models is the distribution of predicted outcomes. The Dixon-Coles model seems to fail to capture the home-field advantage as well as the bookmakers. Indeed the bookmakers predicted a home win for 62% of the season games compared to 56.7% for the Dixon-Coles model. This means that besides the maximum predicted event being misidentified more often, this also means that the predictions of the less likely outcomes are mis-represented.

In other words, the probability distribution will also be incorrectly modelled. Overall the Dixon-Coles model favors away teams more than bookmakers. This is examplified by the fact that Dixon-Coles predicts correctly more away wins than bookmamkers. But due to the imbalance of the outcome distribution, it overall performs worse. This phenomenon of over-estimating the winning chances of the away team is going to affect the results of the betting framework.

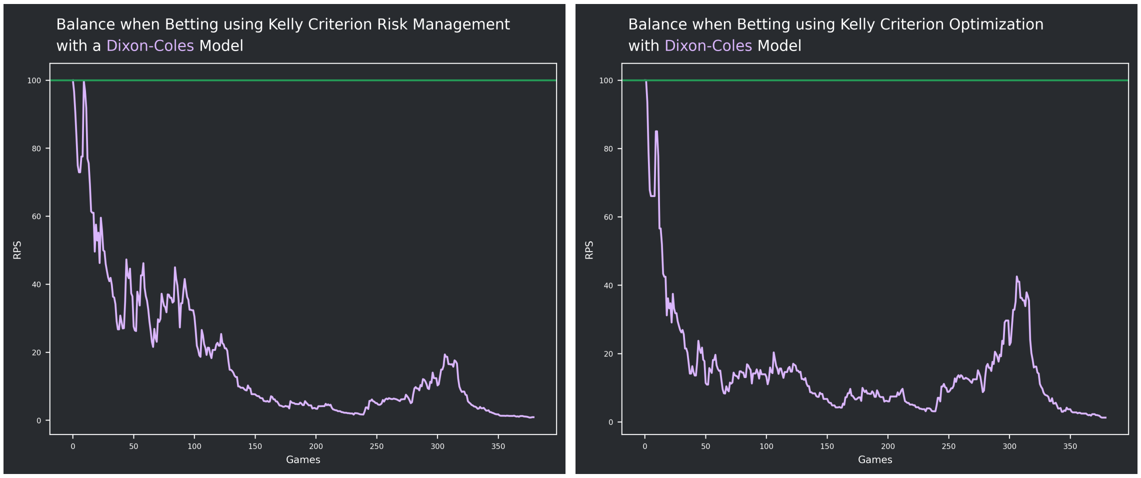


Figure 16: Result of the betting strategy over a season

Both betting frameworks rely on the predictions of the models. As is commonly known in Data Science, ”Garbage in, garbage out”. Even though, this phrase is more often used to refer to the input variables given to a predictive model, this still applies to the input given to the optimization models. It follows that with faulty outcome predictions, the optimal betting wagers will be in favor of the wrong outcomes. The first betting framework relying on the Kelly Criterion, ends the season with 0.83% of the starting balance. This betting framework suggests to not bet on 52 occasions and the other times, it places bets on the home team 91 times, on draws 13 times, and on the away team 224. Interestingly, despite never predicting a draw, the Kelly Criterion still suggests that there is an edge to exploit for those outcomes. The bets were victorious 90 times and losses 238 times. The second betting framework relying on Kelly Criterion Optimization, ends the season with 1.14% of the starting balance. This betting framework suggests to not bet on 107 occasions and the other times, it places bets on the home team 69 times, on draws 126 times, and on the away team 81. Interestingly, like we mentioned the Kelly Criterion Optimization maximizes returns with respect to bets on all possible outcomes. Thus the model suggests to place a bet on more than one outcome 119 times (or 43.6% of betting times). The bets were victorious 90 times and losses 183 times. Even though the Kelly Criterion Optimization framework performs slightly better than the Kelly Criterion, the results clearly show that the overestimation of away teams' winning chances hinders the betting strategy.

8 Conclusion

The main takeaway from the data driven strategy developed in this paper, is that it is not ready to be applied to football betting markets. Although the Kelly Criterion Optimization method is a sound approach to optimizing returns, it relies on faulty data generated by the prediction models implemented. The area where improvements are needed is obviously the predictive models. As I mentioned in the Baseline Models section, if no model can outperform the baselines or the benchmarks, then they should not be used. The models from the literature that I reproduced in this paper were all using result data as input. Yet as we have mentioned earlier, scores in football are random or at least noisy. Goals being rare occasions in a game are soft indicators of a teams strength. Thus models that incorporate more advanced metrics should be able to perform better. Building a set of features of raw statistics and refined performance statistics along with contextual data about team dynamics and rivalries are interesting paths to explore. Using more complex machine learning models should handle the complexity of the task.

The predictive models I described and implemented are fine-tuned for the task of football outcome predictions. Yet some adaptations could be made to apply them to other sports. The Elo rating system, originally made for zero-sum games could easily be applied to sports with only two outcomes (like tennis, rugby, basketball etc.). For the models that predict draws, like Poisson or Thurstone-Mosteller, we could apply them to two outcome sports by ignoring draw predictions and re-scaling home and away wins accordingly. Applying the statistical models to sports with more frequent scoring opportunities should also be straightforward: one could simply modify the constraints on team strength to predict goals more often. Additionally, we could apply these predictive models to other betting opportunities. Indeed, the Poisson or Dixon-Coles model that predict the number of goals scored could also be applied to bet on the number of goals scored in a game. Overall both the Kelly Criterion and Kelly Criterion Optimization betting frameworks are technically sport agnostic. Thus they could be applied to any events (game winners, number of goals scored etc.) given the predicted outcome events.

9 Appendix

9.1 A. Dixon-Code model code sample

```python
import pandas as pd
import numpy as np

from scipy.stats import poisson
from scipy.optimize import minimize
```

Listing 1: Python library dependencies

```python
def _rho_correction(x, y, lambda_x, mu_y, rho):
    return np.select(
        [
            (x == 0) & (y == 0),
            (x == 0) & (y == 1),
            (x == 1) & (y == 0),
            (x == 1) & (y == 1),
        ],
        [
            1 - (lambda_x * mu_y * rho),
            1 + (lambda_x * rho),
            1 + (mu_y * rho),
            1 - rho,
        ],
        default=1,
    )
```

Listing 2: Rho correction method for low scores correlation

```python
def neg_log_likelihood(parameters, games):
    # Gather team specific parameters (attacking and defensive strengths)
    parameter_df = (...) # Redacted since its not relevant

    # Merge team specific parameters with each games
    fixtures_df = (...) # Redacted since its not relevant

    # Predict the number of goals scored
    score1_inferred = (
        np.exp(
            fixtures_df["home_adv"] +
            fixtures_df["intercept"] +
            fixtures_df["attack1"] -
            fixtures_df["defence2"])
            )
    score2_inferred = (
        np.exp(
            fixtures_df["intercept"] +
            fixtures_df["attack2"] -
            fixtures_df["defence1"])
            )
    # Compute the likelihood of the observed goals scored
    score1_loglikelihood = poisson.logpmf(
        fixtures_df["score1"],
        score1_inferred)
    score2_loglikelihood = poisson.logpmf(
        fixtures_df["score2"],
        score2_inferred)
    # Sum the likelihood for all games
    return -(
        (
            score1_loglikelihood +
            score2_loglikelihood +
            np.log(self._rho_correction(
                fixtures_df["score1"],
                fixtures_df["score2"],
                score1_inferred,
                score2_inferred,
                parameters[-3]))
            ) * fixtures_df['weight']
        ).sum()
```

Listing 3: Method to compute the Likelihood of observed data

```
def maximum_likelihood_estimation():
    # Set strength ratings to have unique set of values for reproducibility
    constraints = [{
        "type": "eq",
        "fun": lambda x:
            sum(x[: league_size]) - league_size
        }]
    # Set the maximum and minimum values the parameters can take
    bounds = [(0, 3)] * league_size * 2
    bounds += [(-.2, .2)]
    bounds += [(0, 1)] * 2

    solution = minimize(
        neg_log_likelihood,
        parameters,
        args=games,
        constraints=constraints,
        bounds=bounds,
        options={’disp’: False, ’maxiter’: 100})

    parameters = solution["x"]
```

Listing 4: Method to optimize parameter by minimizing the MLE

9.2 B. Kelly Criterion Optimization Code sample

```
def kco(parameters, odds, predictions, balance):
    ending_bankroll = np.zeros(3) + balance
    winnings = parameters * odds # Compute conditional winnings
    ending_bankroll += winnings
    losses = np.sum(parameters) # Compute total wagers
    ending_bankroll -= losses
    # Compute expected account balance
    return - (np.log(ending_bankroll) * predictions).sum()
```

Listing 5: Method that computes the growth formula

```
def bet(row):
    global account_balance
    parameters = np.concatenate((np.zeros(3),)) # Initialize bets to zero
    odds = np.array([row.home_win, row.draw, row.away_win]) # Betting odds
    predictions = np.array([row.home_win_p, row.draw_p, row.away_win_p])
    # Force total betting amount to be below or equal to account balance
    constraints = (
        {
            "type": "ineq",
            "fun": lambda x:
                - sum(x) + account_balance },
        )
    # Force bets to be positive
    bounds = [(0, None)] * 3
    # Compute optimal parameters
    solution = minimize(
        kco,
        parameters,
        args=(odds, predictions, account_balance),
        constraints=constraints,
        bounds=bounds,
        options={'disp': False, 'maxiter': 1000}
        )
    bet_amount = solution['x']

    # Remove all wagers from the account
    account_balance -= sum(bet_amount)

    # Get result of the bet
    if row['winner'] == 0:
        account_balance += bet_amount[0] * odds[0]

    if row['winner'] == 1:
        account_balance += bet_amount[1] * odds[1]

    if row['winner'] == 2:
        account_balance += bet_amount[2] * odds[2]

    return account_balance

account_balance = 100
df_dc['balance'] = df_dc.apply(bet, axis=1)
```

Listing 6: Method that simulate betting

References

[1] J. Lecot. “Paris sportifs : Comment les operateurs bloquent les joueurs gagnants.” (), [Online]. Available: https://www.liberation.fr/societe/paris-sportifs-comment-les-operateurs-bloquent-les-joueurs-gagnants-20210702_JXZMB4C2NNA4JP72GJHDX (accessed: 07.05.2022).

[2] ANJ. “Marche 2021 : Tres forte dynamique pour les paris sportifs en ligne & croissance de la loterie.” (), [Online]. Available: https://anj.fr/marche-2021-tres-forte-dynamique-pour-les-paris-sportifs-en-ligne-croissance-de-la-loterie. (accessed: 07.05.2022).

[3] M. Debry. “Pubs pour les paris sportifs pendant l’euro : Une reglementation efficace s’impose pour lutter contre les addictions.” (), [Online]. Available: https://www.pourquoidocteur.fr/Articles/Question-d-actu/36820-Pubs-paris-sportifs-l-Euro-une-reglementation-efficace-s-impose-pour-lutter-les-addictions. (accessed: 09.06.2022).

[4] C. Palierse. “Paris sportifs : Vers un recadrage de la publicite.” (), [Online]. Available: https://www.lesechos.fr/industrie-services/services-conseils/paris-sportifs-vers-un-recadrage-de-la-publicite-1333526. (accessed: 09.06.2022).

[5] V. Collet. “Les paris en ligne desormais legaux en france.” (), [Online]. Available: https://www.lefigaro.fr/conjoncture/2010/04/07/04016-20100407ARTFIG00037-les-paris-en-ligne-desormais-legaux-en-france-.php. (accessed: 09.06.2022).

[6] C. Delporte. “Paris sportifs : Pour l’amour du risque.” (), [Online]. Available: https://www.lesechos.fr/weekend/perso/paris-sportifs-pour-lamour-du-risque-1409644. (accessed: 09.06.2022).

[7] C. Imsand. “Quand les jeux d’argent tournent au cauchemar.” (), [Online]. Available: https://www.letemps.ch/suisse/jeux-dargent-tournent-cauchemar. (accessed: 09.06.2022).

[8] P. Marek. “Bookmakers’ efficiency in english football leagues.” (), [Online]. Available: https://www.researchgate.net/publication/327631361_Bookmakers'_Efficiency_in_English_Football_Leagues. (accessed: 09.06.2022).

[9] J. F. E. Moore. “Betting with house money: Reverse line movement based strategies in college football totals markets.” (), [Online]. Available: https://doi.org/10.1007/s12197-019-09479-3. (accessed: 07.05.2022).

[10] S. Green. “Assessing the performance of premier league goalscorers.” (), [Online]. Available: https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/. (accessed: 09.06.2022).

[11] L. et al. “Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data.” (), [Online]. Available: https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5fee09c092fcdb0989d51ecf_1034_rppaper_SoccerPaper5.pdf. (accessed: 09.06.2022).

[12] J. V. H. Lotte Bransen and M. van de Velden. “Measuring soccer playersâ contributions to chance creation by valuing their passes.” (), [Online]. Available: https://janvanhaaren.be/assets/papers/jqas-2019-passes.pdf. (accessed: 09.06.2022).

[13] T. D. L. B. Jan Van Haaren Pieter Robberechts and J. Davis. “Analyzing performance and playing style using ball event data.” (), [Online]. Available: https://janvanhaaren.be/assets/papers/bih-2019-event-data.pdf. (accessed: 09.06.2022).

[14] T. Athletic. “How each premier league team pass.” (), [Online]. Available: https://theathletic.com/3346644/2022/06/07/premier-league-pass-networks/. (accessed: 07.05.2022).

[15] W. Spearman. “Beyond expected goals.” (), [Online]. Available: https://www.researchgate.net/publication/327139841_Beyond_Expected_Goals. (accessed: 09.06.2022).

[16] F. Mosteller. “Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations.” (), [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-44956-2_8. (accessed: 09.06.2022).

[17] R. Bradley and M. Terry. “Rank analysis of incomplete block designs: I. the method of paired comparisons. biometrika.” (), [Online]. Available: https://doi.org/10.1093/biomet/39.3-4.324. (accessed: 09.06.2022).

[18] M. J. Maher. “Modelling association football scores.” (), [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.1982.tb00782.x. (accessed: 09.06.2022).

[19] M. J. Dixon and S. G. Coles. “Modelling association football scores and inefficiencies in the football betting market.” ().

[20] W. T. d. Ley C and E. HV. “Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches.” (), [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1471082X18817650. (accessed: 07.05.2022).

[21] G. Baio1 and M. A. Blangiardo. “Bayesian hierarchical model for the prediction of football results.” (), [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/16040/1/16040.pdf. (accessed: 07.05.2022).

[22] R. Baboota and H. Kaur. “Predictive analysis and modelling football results using machine learning approach for english premier league.” (), [Online]. Available: https://doi.org/10.1016/j.ijforecast.2018.01.003. (accessed: 07.05.2022).

[23] T. J. N. Ryan Beal Stuart E. Middleton and S. D. Ramchurn. “Combining machine learning and human experts to predict match outcomes in football: A baseline model.” (), [Online]. Available: https://arxiv.org/abs/2012.04380. (accessed: 07.05.2022).

[24] J. V. H. a. P. R. Jesse Davis. “A bayesian approach to in-game win probability in soccer.” (), [Online]. Available: https://arxiv.org/pdf/1906.05029.pdf. (accessed: 07.05.2022).

[25] J. H. Kushal Shah and D. Samangy. “A poisson betting model with a kelly criterion element for european soccer.” ().