A Profitable Model For Predicting the Over/Under Market in Football

Edward Wheatcroft

London School of Economics and Political Science, Houghton Street, London, United Kingdom, WC2A 2AE.

Abstract

The over/under 2.5 goals betting market allows gamblers to bet on whether the total number of goals in a football match will exceed 2.5. In this paper, a set of ratings, named 'Generalised Attacking Performance' (GAP) ratings, are defined which measure the attacking and defensive performance of each team in a league. GAP ratings are used to forecast matches in ten European football leagues and their profitability is tested in the over/under market using two value betting strategies. GAP ratings with match statistics such as shots and shots on target as inputs are shown to yield better predictive value than the number of goals. An average profit of around 0.8 percent per bet taken is demonstrated over twelve years when using only shots and corners (and not goals) as inputs. The betting strategy is shown to be robust by comparing it to a random betting strategy.

1 Introduction

Interest in forecasting the outcomes of sporting events has grown a great deal in recent years. This is perhaps partly as a result of greater access to data (The Economist (2017)), greater computational power and an increase in the number of available betting markets. It is now commonplace for media companies to include estimates of the probabilities of different outcomes of sporting events in their coverage and consumers appear to be largely comfortable with this. This practice is perhaps most common in the United States where probabilistic forecasts of National Football League (NFL), National Basketball Association (NBA) and Major League Baseball (MLB) events are routinely disseminated to the public and have been for some years (ESPN (2018)). Probabilistic forecasting for the purpose of gambling has also taken off in recent years. Increased competition in the bookmaking industry has led to smaller profit margins and hence greater opportunities for gamblers. Betting exchanges allow gamblers to bet against each other with one party 'laying' a bet for another and the exchange taking a small cut of the return. Crucially, under this arrangement, there is no incentive for the company to exclude those customers making consistent profits and thus, if high quality forecast models can be built, opportunities for consistent profit making exist.

The purpose of this paper is twofold. First, a set of ratings, named Generalised Attacking Performance (GAP) ratings, are introduced, which measure the attacking and defensive ability of each team within a sports league. GAP ratings can take as inputs any measure of attacking performance which could, for example, be match statistics such as shots or shots on target or might simply be goals. Second, it is demonstrated how a profit can be made by using GAP ratings to form probabilistic forecasts for the over/under 2.5 goals market (which allows gamblers to bet on whether the total number of goals in a match will exceed 2.5 or not) in association football (hereafter, just referred to as 'football'). It is found that, assuming that the maximum odds over all bookmakers on the BetBrain website were available, using forecasts formed using GAP ratings with the number of shots and corners as inputs alongside a level stakes value betting strategy would have provided an average profit margin of around 0.8 percent over the last 12 years over a total of 68,672 bets. It is subsequently demonstrated that the probability of achieving such a return by chance (i.e. by betting randomly with the same frequency) is extremely small (so small that, over one million trials, we were unable to find such a high return by randomly choosing bets at this frequency) and thus that the strategy appears to be robust in finding effective betting opportunities. A second strategy based on the Kelly criterion is tested and shown to be capable of yielding a larger profit by varying the stake according to the degree of 'value' in the odds. The analysis is repeated using average odds and it is found that a loss is made under both betting strategies. The implication of this in terms of the efficiency of the market is then discussed.

Whilst a vast number of betting markets are now available, obtaining odds data from a large number of games for these markets is generally difficult. A large archive of historical odds from the over/under 2.5 goals and match outcome markets are available, however, and this justifies the choice of the former in this case. A future paper will consider the use of GAP ratings in the match outcome market.

A large number of academic papers have focused on producing ratings systems for sets of players or teams in a sporting context. This has been done both with specific sports in mind such as football, and in a more general context with the aim of providing ratings systems that translate between sports. Almost certainly the most well known approach to producing sports ratings is the Elo rating system which has a long history in sport and forms an important basis for a range of ratings systems. Elo ratings were first designed to produce rankings for chess players and the system was implemented by the United States Chess Federation in 1960 (Elo (1978)). The Elo system assigns ratings to each player and, from those ratings, probabilities of the outcome of a game between two players or teams is estimated. The two ratings are then updated to reflect the outcome once the game is finished. The system was initially designed for the case in which outcomes are binary (i.e. there are no ties), but, more recently, has been extended to account for draws, making it applicable to sports such as football, in which draws are common. At the updating stage (once a game has been played), the system computes the difference between the estimated probabilities and the outcome (assigned a one, a zero, or 0.5 for a draw). As such, the system in its original form does not account for the *size* of a win; for example, in football, there would be no additional increase in a team's ranking from winning by several goals over winning by one goal. Elo ratings have been demonstrated in a football context and found to perform favourably with respect to six other rating systems (Hvattum & Arntzen (2010)). In 2018, Fifa switched to an Elo rating system to produce its international football world rankings (Fifa (2018)). Elo ratings have also been applied to other sports such as, for example, Rugby League (Carbone et al. (2016)) and video games (Suznjevic et al. (2015)), whilst `fivethirtyeight.com` produce probabilities for NFL (FiveThirtyEight (2019*a*)) and NBA (FiveThirtyEight (2019*b*)) based on Elo ratings.

The Elo rating system provides a single rating for each participating team relating to the overall ability of that player or team. This approach has been taken for other ratings systems. The pi-rating system assigns a home and an away overall rating to each team in a league, both of which are updated after each match in which a team is involved. The change in rating is dependent on the winning margin of each team but is tapered such that additional goals on top of already large winning margins are assigned less weight (Constantinou & Fenton (2013)). The approach taken in this paper is to assign attacking and defensive ratings to each team and to use those ratings to calculate match probabilities. This approach has been taken by a large number of authors. Maher (1982) used fixed ratings for each team in combination with a Poisson model to estimate the number of goals scored by each team (but not match probabilities). This approach was extended by Dixon & Coles (1997) who used similar attack and defence ratings in a Poisson regression model to estimate match probabilities. They combined these with a value betting strategy (described and used later in this paper) to demonstrate a significant profit for cases in which the model suggested a large discrepancy between the estimated probabilities and the probabilities implied by the odds. Whilst the approach of Dixon and Coles assumed a fixed attack and defensive rating over time for each team, this was extended by Rue & Salvesen (2000) who defined a Bayesian model to allow the ratings to vary over time. Dixon & Pope (2004) used a slightly modified version of the model proposed by Dixon and Coles to demonstrate a profit using a wider range of published bookmaker odds. Other examples of the use of attacking and defensive ratings systems include Karlis & Ntzoufras (2003), Lee (1997) and Baker & McHale (2015).

The use of machine learning in prediction has become more and more popular in recent years. A 2017 special issue in the journal 'Machine Learning' presents the results of a 'Soccer Prediction Challenge' in which participants were provided with results from 216,743 past football matches from around the world and were asked to make predictions regarding 206 future matches (Berrar et al. (2019)). A number of the entrants achieved very encouraging results. Hubáček et al. (2019) won the competition using gradient boosted trees whilst Constantinou (2019) finished second using a model called Dolores which combines dynamic ratings with Hybrid Bayesian Networks. Other past uses of machine learning to predict football results include O'Donoghue et al. (2004), Van Haaren & Van den Broeck (2015), Joseph et al. (2006) and Hucaljuk & Rakipović (2011).

A number of other approaches to football prediction have been taken over the years. These include Bayesian generalised linear models (Rue & Salvesen (2000)), Bayesian Networks (Constantinou et al. (2012)) and neural networks (Huang & Chang (2010)).

The question of how to translate team ratings into probabilistic forecasts was addressed by Goddard (2005). Two approaches are generally in use. The first approach uses ratings as inputs to Poisson regression models, simulating the number of goals scored by each team and calculating the probabilities of each outcome accordingly. The other approach predicts the probability of each result directly using methods such as ordered probit regression. It was found that there is little difference in the performance of the two approaches. In this paper, the latter approach is used with the probability of exceeding 2.5 goals modelled directly using logistic regression.

In recent years, the idea that other match statistics might provide a better indication of the relative performance of the teams than the actual number of goals scored has been investigated. The reasoning behind this is that there is a significant element of chance involved in scoring a goal and that the number of goals actually scored might be indicative of the relative performances of each team. More effort has thus been invested in attempting to use data to understand the relative performance of the two teams in the hope that this will provide a better indication of future performance (Cintia et al. (2015)). The concept of 'expected goals' has also taken off in recent years (Rathke (2017), Opta (2018)). The aim here is simply to attempt to calculate the number of goals a team would be 'expected' to score given the location and nature of all its shots taken during the match (Eggels (2016)). The use of shot data is conceptually very similar to that of expected goals. The difference, however, is that, in the former case, a constant weighting is placed on each shot, since the nature and location of each one is not taken into account.

The use of match statistics such as shots, shots on target and corners has become more feasible in recent years due to the greater availability of these data. Free data on overall match statistics such as shots, corners, fouls, red and yellow cards as well as bookmakers' odds are freely available in an easy-to-use format at the popular and widely used Football-data.co.uk website (Football Data (2018)). It is from this source that the data used in this paper are taken. Other sources collect a vast amount of data for matches in top European leagues including shot locations, shot methods (headed, right footed, left footed etc), the locations of and outcomes of passes, changes of possession and much more (Opta Football (2018)). The cost of obtaining these data is often prohibitively expensive for individuals, however, and these data are not considered in this paper.

This paper is organised as follows. First, in section 2, background information is given regarding bookmakers' odds and how these can be interpreted as implied probabilities. In section 3, the data utilised in this paper is described. In section 4, GAP ratings are defined. In section 5, the approach taken in this paper to producing probabilistic forecasts from GAP ratings is described. Approaches to forecast evaluation and parameter selection are described in section 6. The approach taken to dealing with promotions and relegations is described in section 7. The two betting strategies used in this paper are described in section 8. The experimental design is defined in section 9 and the results are described in section 10. Section 11 is used for discussion.

2 Background

This paper makes extensive use of betting odds, using them both as inputs with which to produce probabilistic forecasts and as a basis with which to demonstrate the performance of the forecasts as a potential money making tool. Before proceeding, however, it is useful to define exactly what is meant when betting odds are referred to. The format of betting odds varies somewhat in different regions of the world. In the context of this paper, it is convenient to consider 'decimal', or 'European style', betting odds (to which other formats can easily be converted). Decimal odds simply indicate by how much the stake is multiplied in the event of that bet being successful. For example, if the odds offered on an event are 3, a unit stake would generate a return of 3. The 'fractional', or 'British style' odds, in this case would be given as '2/1' indicating a profit of 2 for a stake of 1. Hereafter, in this paper, betting odds are given in decimal format.

Another useful concept is that of the 'odds implied' probability. Define the decimal odds for the ith event to be O_i. The 'odds implied' probability is simply defined as the inverse, i.e. $r_i = \frac{1}{O_i}$. For example, if the odds on an event are $O_i = 3$ then $r_i = \frac{1}{3}$. Whilst, conventionally, probabilities over a set of exhaustive events are required to add up to one, this is *not* the case for odds implied probabilities and, in fact, the sum of odds implied probabilities for an event will usually exceed one; the excess representing the bookmaker's profit margin or 'overround'. The overround represents a significant challenge to gamblers because it means that the bookmakers shorten their odds in order to make a profit.

3 Data

This paper utilises the repository of past football match data available at www.football-data.co.uk which supplies free-to-access match-by-match data for a range of European Leagues dating back, in some cases, as far back as the 1993/1994 season. Whilst early years of the data set feature only the final result of each match, in more recent years, the information provided has grown more rich such that, for each football match in any of 10 different European leagues, full and half time results, the number of shots, shots on target and number of corners taken by each team are given along with betting odds of the final result (home win, away win or draw) and of the over/under 2.5 goal market from multiple bookmakers. The maximum and average odds for the match outcome and over/under 2.5 goal markets over many bookmakers, collected from the BetBrain odds comparison website, are also given. From the 2017/2018 season onwards, the number of leagues in which all of this information is available has been increased to 22 (though data from the extra 12 leagues are not utilised in this paper).

In this paper, a rating system is introduced, called the GAP rating system, and those ratings are used as inputs to produce probabilistic forecasts. The available data provide an opportunity for robust evaluation of forecasts and of potential betting strategies. In this paper, the performance of the forecasts is compared with different inputs to the GAP ratings representing different measures of attacking performance, including shots and corners. Only leagues for which this information is available along with the maximum and average BetBrain odds in the over/under 2.5 goal market are thus considered. The leagues considered in this paper are summarised in table 1. The total number of football matches considered is thus 54,437. The data set does not include cup games, playoffs or any other extra matches during the regular season and thus these are not considered in this paper.

4 Generalised Attacking Performance (GAP) Ratings

Generalised Attacking Performance (GAP) ratings are now formally defined. Consider a football league consisting of N different teams who all play each other a number of times over a season. For a given football match, let S_h and S_a be some measure of the attacking performance of the home and away

League	First available season	Number of matches
English Premier League	2005/2006	6460
English Championship	2005/2006	6624
English League One	2005/2006	6624
English League Two	2005/2006	6624
English National League	2005/2006	6488
Scottish Premier League	2005/2006	2940
Spanish Primera Liga	2005/2006	4910
Italian Serie A	2005/2006	4908
French Ligue One	2005/2006	4908
German Bundesliga	2005/2006	3951

Table 1: Football league data used in this paper.

teams respectively, where the definition of 'attacking performance' (hereafter frequently referred to as the 'GAP rating input') is given by the user and is usually derived from match statistics. Most obviously, the input could be defined as the number of goals scored by each team but could also be the number of shots taken, the number of shots on target, the number of corners, a combination of these, or any other indication of attacking performance. Each team is given a separate attacking and defensive GAP rating for home and away games defined as follows:

- H_i^a - The attacking GAP rating of the ith team for home games
- H_i^d - The defensive GAP rating of the ith team for home games
- A_i^a - The attacking GAP rating of the ith team for away games
- A_i^d - The defensive GAP rating of the ith team for away games

The attacking GAP ratings relate roughly to the expected attacking performance that a team should achieve against an average team in the league whilst the defensive ratings relate to the expected attacking performance of an average opposing team. The best teams in the league will thus have a high attacking rating and a low defensive rating. The following rule is used to update the ith team's GAP ratings after a home match against the jth team:

$$
\begin{aligned}
H_i^a &= \max(H_i^a + \lambda\phi_1(S_h - \frac{H_i^a + A_j^d}{2}), 0), \\
A_i^a &= \max(A_i^a + \lambda(1-\phi_1)(S_h - \frac{H_i^a + A_j^d}{2}), 0), \\
H_i^d &= \max(H_i^d + \lambda\phi_1(S_a - \frac{A_j^a + H_i^d}{2}), 0), \\
A_i^d &= \max(A_i^d + \lambda(1-\phi_1)(S_a - \frac{A_j^a + H_i^d}{2}), 0)
\end{aligned}
\tag{1}
$$

The GAP ratings for the jth team (away to the ith team) are updated as follows:

$$
\begin{aligned}
A_j^a &= \max(A_j^a + \lambda\phi_2(S_a - \frac{A_j^a + H_i^d}{2}), 0), \\
H_j^a &= \max(H_j^a + \lambda(1-\phi_2)(S_a - \frac{A_j^a + H_i^d}{2}), 0), \\
A_j^d &= \max(A_j^d + \lambda\phi_2(S_h - \frac{H_i^a + A_j^d}{2}), 0), \\
H_j^d &= \max(H_j^d + \lambda(1-\phi_2)(S_h - \frac{H_i^a + A_j^d}{2}), 0),
\end{aligned}
\tag{2}
$$

where, $\lambda > 0$, $0 < \phi_1 < 1$ and $0 < \phi_2 < 1$ are parameters to be selected.

For a given match, a home team can be said to have outperformed expectations in an attacking sense if its attacking performance is greater than the mean of its attacking rating and the opposition's defensive rating and thus, when this is the case, both its home and away attacking ratings are increased (or decreased, if its attacking performance is lower than expected). The same is true if an away team outperforms expectations. A team's defensive ratings are impacted in a similar way such that, if an opposing team's attacking performance is lower than expected, its defensive rating is decreased (note that lower defensive ratings indicate better defensive performance). The parameter λ determines the influence of the last match on the ratings of each team whilst ϕ_1 and ϕ_2 determine the influence of a home match on a team's away ratings and the influence of an away match on a team's home ratings respectively. If $\phi_1 = 0$, the home team's away ratings will not be affected whilst the same is true of the away team's home ratings if $\phi_2 = 0$. The maximum operator is included to ensure that the GAP ratings for each team cannot become negative. Parameter selection is discussed in section 6.

The GAP rating system described above has some similarities to, and, to some extent, is inspired by, the pi-rating system defined in Constantinou & Fenton (2013). Some of the similarities between the two ratings systems are described below:

1. The parameter λ performs the same function as for both GAP ratings and pi-ratings. The Elo rating system also has a similar learning parameter.

2. Both GAP ratings and pi-ratings consist of separate home and away ratings for each team (in order to account for the impact of home advantage).

3. In GAP ratings, the parameters ϕ_1 and ϕ_2 govern the impact of a home match on a team's away ratings and the impact of an away match on a team's home rating respectively. In the pi-rating system, a parameter γ is defined with a similar function which serves both purposes (it is necessary for the one parameter to serve both purposes to ensure that the sum of the ratings over the entire league remains zero).

4. A pi-rating can roughly be interpreted as a team's expected winning (or losing) margin over an average team in a league. GAP ratings can similarly be interpreted as the expected number of attacking plays (as defined) of a team against an average opponent.

Despite the above similarities, there are a number of key differences:

1. GAP ratings are defined more generally than pi-ratings such that any measure of attacking performance can be used as an input. It would, however, be straightforward to extend pi-ratings in this way also.

2. In the GAP rating system, each team has four ratings corresponding to its attacking and defensive performance and the location of the match (home or away). Pi-ratings do not distinguish attacking and defensive ability and thus each team is assigned two ratings corresponding to the expected difference in the number of goals scored by each team. This makes GAP ratings more suitable for forecasting the total number of goals in a match.

3. GAP ratings define the expected number of attacking plays by a home (away) team to be the average of their home (away) GAP rating and the away (home) defensive GAP rating of the opposition. Pi-ratings define the expected winning (losing) margin of the home team to be the difference between the pi-ratings of the two teams.

4. pi-ratings are constrained such that the sum of the ratings over the league is zero. For GAP ratings, the only constraint is that all ratings must be greater than or equal to zero.

5. As described in the list of similarities above, GAP ratings have separate parameters for the effect of a home team's performance on its away rating and an away team's performance on its home rating. This extra flexibility is possible in the former case because of the difference in the constraint described above.

5 From GAP Ratings to Probabilistic Forecasts

Here, an approach to using GAP ratings to predict the probability that there will be three or more goals in the match is described. Although a variety of approaches could be taken to do this, in this paper, a simple logistic regression approach is taken, allowing for multiple variables to influence the estimated probability. Let $y = \log(\frac{\hat{p}}{1-\hat{p}})$ where $\hat{p}$ is the estimated probability of the total number of goals exceeding 2.5. A regression model is defined as

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1(H_i^a + H_i^d + A_i^a + A_i^d) + \hat{\beta}_2 r_i \quad (3)$$

where $r_i = \frac{1}{o_i}$ is the odds implied probability (which may be derived from a single bookmaker or averaged over several). The parameters are estimated using least squares estimation. From $\hat{y}$, the estimated probability $\hat{p}$ can easily be derived.

6 Parameter Selection and Forecast Evaluation

GAP ratings require the selection of three parameters: λ, a parameter that governs the impact of the most recent match on a team's ratings and ϕ_1 and

ϕ_2, parameters that govern the impact of a home match on a team's away ratings and of an away match on a team's home ratings respectively. In addition, the selection of a number of extra parameters is required depending on the approach taken to forming probabilistic forecasts. Under the approach described in section 5, and which is used to calculate the results shown in this paper, a further three parameters are used in the logistic regression making the total number of parameters six.

Given the probabilistic nature of the forecasts, parameter selection is performed with the aim of maximising forecast performance given a scoring rule, a function of a probabilistic forecast and its outcome that, when averaged over many forecasts and outcomes, measures the skill of the forecasting system (other approaches such as attempting to maximise profit could also be used but are not considered here). An important property of scoring rules is propriety. A score is *proper* if it is optimised in expectation by the true underlying probability or probability distribution from which the outcome was drawn. It is often argued that only scoring rules that are proper should be used in practice since, otherwise, a forecaster may be encouraged to issue forecasts that don't reflect their true beliefs (Bröcker & Smith (2008)).

The scoring rule of choice in this paper is the *ignorance score* (Good (1952), Roulston & Smith (2002)), given by $S(p(Y)) = -\log_2(p(Y))$ where $p(Y)$ is the probability or (in the continuous case) probability density placed on the outcome Y. The ignorance score is proper and is also conveniently linked to the more well known concept of maximum likelihood in that the aim is to maximise the mean of the logarithm of the probability placed on the outcomes. A benefit of the ignorance score is in its interpretation. The difference between two ignorance scores calculated using two different forecast systems of the same outcome can be interpreted as the number of bits of information gained from using one over the other. For example, if forecast system A has a mean ignorance of 0.75 and forecast system B has a mean ignorance of 0.5, forecast system B provides 0.25 extra bits of information, on average, than forecast system A since ignorance is a negatively oriented score. This can be interpreted as placing $2^{0.25} \approx 1.19$ more probability on the outcome, on average.

Parameter selection is performed by minimising the mean ignorance score over all previous matches. This is done using the Nelder-Mead simplex algorithm as implemented in the fminsearch function in Matlab. For computational reasons, this is done with respect to the three parameters that impact the ratings and the logistic regression parameters are optimised using least squares given the ratings that result. The parameters are therefore not all optimised simultaneously. Since it is computationally expensive to optimise the ratings parameters, this is done only between seasons and over all leagues simultaneously. The logistic regression parameters, however, are optimised after every round of games since it is computationally cheap to do so.

A useful concept commonly used in weather forecasting is that of the 'climatology'. A climatology is simply defined as a distribution of past states, over some given time period. An example would be the distribution of maximum temperatures at a specific location on a given day over the last 100 years or, in the binary case, the proportion of days on which it has rained over that time. The climatology gives a useful summary of past states but can also be treated as a forecast itself. For example, weather forecasts cannot generally be expected to have skill beyond around two or three weeks and thus, if one wanted to gain insight into how, say, the temperature might pan out a year from now at a given location, the climatology would likely provide the best information available. The concept of climatology can easily be transferred into sports forecasting where, in this case, observing three or more goals in a football match is conceptually similar to observing a rainy day. The climatology is therefore used as an alternative baseline forecast, the skill of which is compared with that of the forecast system defined in this paper. If the forecast system cannot outperform the climatology, it is of little use in practice. The probability of exceeding 2.5 goals is, in fact, found to be around 50 percent. All mean ignorance scores in this paper are quoted relative to that of the climatology (i.e. with the climatological ignorance subtracted) such that, if the mean ignorance is less than zero, the forecast system outperforms the climatology, on average.

An alternative approach to the parameter selection approach performed in this paper would be to maximise the profit made through the betting strategy. However, this would cause the function to be optimised to be less smooth and would therefore make optimisation more challenging as the risk of falling into a local maximum (minimum) would be increased. This is therefore left as a suggestion for possible future work.

7 Dealing With Promotions and Relegations

It is almost universally the case that football leagues feature promotion and/or relegation, i.e. if a team finishes top or close to the top of a league, they are offered the chance to play in the league above, whilst teams that finish close to the bottom are 'relegated' to the league below. This links into the broader question of how a team's ratings should transfer from one season to another. Whilst a team that remains in the same league from one season to the next already has a set of ratings with which to start the following season, teams that enter the league will not and thus some approach is required to initialise the ratings of such teams. A number of approaches could be taken to this, varying somewhat in complexity. For example, data on summer signings, preseason bookmakers' odds and predictions from pundits could all be utilised (though the latter may be hard to quantify). There is also a question of how realistic it is for those teams that stay in the same league to maintain their ratings from one season to the next given managerial changes, player transfers etc. and thus extra data could be used to adjust the ratings before the start of each season. In this paper, however, for simplicity, if a team remains in the same division, they retain their ratings from the previous season whilst teams that are promoted to the league (i.e. come from the league below) are given the average ratings of the teams relegated from that league in the previous season and teams relegated to that league are given the average rating of the promoted teams. Note, however, that it is common for promoted teams to outperform the relegated teams they replace. For example, it has been found that, in the English Premier League, the promoted teams tend to achieve around 8 more points on average than the relegated teams in the previous season (Constantinou & Fenton (2017)).

8 Betting Strategies

This paper uses two different betting strategies based on the concept of 'value betting'. In value betting, whether a bet is placed or not is determined by whether the probabilistic forecast (if taken as a true probability) implies that the odds have 'value'. The odds on a particular outcome are said to have value if they are longer (i.e. generate a larger return) than they should be, given the true probability[1] of that outcome. Specifically, the odds are said to have value if $p_i > r_i$ where, as defined in section 2, p_i represents the 'true' probability and r_i represents the odds implied probability. When this

[1]Here, we take the frequentist definition of probability, i.e. that the probability is the expected relative frequency of outcomes were the match able to be repeated independently an infinite number of times

is the case, the expected profit for the gambler is positive and, thus, by only betting when there is value, they would profit in the long term. In practice, however, due to inevitable limitations inherent in any real world predictive model, the underlying probability can only be estimated and thus will differ from the true probability. Nonetheless a betting strategy can be defined by replacing the true probability, which is not known, with an estimated probability. This strategy is referred to throughout this paper as the *Level Stakes* strategy. Under this strategy, a unit bet is made on the ith outcome if and only if the forecast probability exceeds the odds implied probability, i.e. if $\hat{p}_i > r_i$, where $\hat{p}_i$ is the forecast probability.

An alternative strategy makes use of the Kelly Criterion (Kelly Jr (1956)). Under the Kelly criterion, the proportion of one's wealth placed on a particular outcome is given by

$$f = \max(\frac{rp - 1}{r - 1}, 0) \tag{4}$$

where p is the gambler's estimated probability of the outcome and r is the decimal odds offered by the bookmaker. If the forecast probability is less than the odds-implied probability, the bet is not taken since the stake is zero. The idea behind the Kelly criterion is that the bankroll of the gambler 'grows' over time by increasing the stake proportionally to their bank. In this paper, however, the approach of Boshnakov et al. (2017) is taken, in which equation 4 is always taken as a proportion of 1 and therefore does not depend on the bankroll. Each stake is multiplied by a constant such that the average stake when a bet is taken is exactly 1. This is done so that the results are directly comparable with the level stakes case above. This is referred to in the results section of this paper as the *Kelly strategy*.

In value betting, considering forecast probabilities rather than actual probabilities removes the guarantee of making a profit in expectation and thus the performance of the forecasts and betting strategy will depend heavily on the skill of the forecasts or, more specifically, the ability of the forecasts to identify profitable betting opportunities. With this in mind, some additional conditions are placed on whether to bet or not on a specific match. Bets are not placed until both teams have played at least six league games during that season in order for the ratings to have sufficiently 'learned' about the attacking and defensive ratings of the teams. Although teams remaining in the same league from one season to the next retain their rating, no account is taken of transfer activity and other close-season changes and so the ratings are prone to being uninformative at this stage of the season. The first six matches are chosen to be ineligible for betting in this case because there is some evidence that this is the point at which the league starts to settle down and reflect the overall strength of the teams (Cronin (2019)). In addition, bets are never placed on the last six games for each team as games at this time of the season can be particularly unpredictable due to differences in motivation between teams. For example, a team fighting relegation at the end of the season will have more to play for than a team safely in mid-table at that stage and so there is more incentive for the former to pick their best team or put more effort into winning the game. Indeed, bookmakers will often make the relegation threatened team favourites to win in those situations. These kinds of factors are hard to factor into a mathematical model and whether to bet or not may require some human judgement beyond the scope of this paper.

9 Experimental Design

The performance of the probabilistic forecasts formed using GAP ratings along with the two betting strategies is demonstrated with the following experimental design. Attacking and defensive GAP ratings are calculated for each team and updated after each match. Promotions and relegations are dealt with using the approach described in section 7 in which newly promoted teams are initialised with the average ratings of the relegated teams from the previous season and relegated teams with the average ratings of the promoted teams. For each match, the current attacking and defensive GAP ratings of each team and the odds-implied probabilities from the BetBrain maximum odds are then used to calculate a forecast probability of the number of goals exceeding 2.5 using the logistic regression model defined by equation 3. The parameter values used to calculate the ratings are optimised over all available data in previous seasons and used for the entirety of that season whilst the logistic regression parameters are optimised after every round of games (see section 6 for details). Data from the 2000/2001 season are used solely for parameter selection purposes since no previous data are available. Both the Level Stakes and the Kelly strategies defined in section 8 are then used and the profit/loss for each match based on these strategies are calculated. This is repeated under the assumptions of both maximum and average Betbrain odds. As explained in section 8, the first and last six matches of the season for each team are not treated as eligible for betting.

The above experimental design is applied six times using different measures of attacking performance as inputs to the GAP ratings. The six different inputs tested are listed in table 2. The results of the experiment using each GAP rating input are described and compared in the results section.

	Measure
1	Goals
2	Shots
3	Shots on Target
4	Corners
5	Shots and Corners
6	Shots on Target and Corners

Table 2: GAP rating inputs tested

10 Results

The stated aim of this paper is to build strategies with which to make a long term gambling profit on European football matches. The obvious first consideration is therefore whether the forecasts and betting strategy are able to achieve this. Here, the profitability of both betting strategies is assessed in two different settings. The majority of the analysis is performed in the case in which the maximum BetBrain odds are assumed to be available. The main results are then repeated but with the profit/loss calculated using the average odds. The effect of betting with odds that are a weighted average of the maximum and average odds is then assessed. Finally, the forecast skill of the forecasts under each GAP rating is compared.

10.1 Maximum Odds

The cumulative profit a gambler *would* have made by applying the Level Stakes strategy over all of the considered seasons using the maximum BetBrain odds is shown as a function of time in figure 1 for each input to the GAP ratings. The profit in each season for each input is shown in table 3. Under four of the considered inputs, a profit would have been made over this period. In the cases in which only shots and both shots and corners are considered, the total profit is substantial, reaching around 500 units. This corresponds to an average profit of around 0.8 percent per bet taken. It is unclear whether the small financial gain in considering corners alongside shots is due to chance or a genuine improvement in the ability to find value bets. There is a substantial difference between the profit made from using only shots on target and from including corners as well, however, suggesting that corners may be a useful addition in terms of the performance of the strategy. It is also interesting to note that using only corners as inputs outperforms using only shots on target. Finally, strikingly, when goals scored is used as the GAP rating input, the gambling return is far lower than for each of the other measures, resulting in a large loss. It is worth noting that, were the gambler to bet on the number of goals exceeding 2.5 in all possible matches (that is including those deemed not to have value), they would make an average loss of 1.81 percent whilst, if they were to bet that the number of goals would be less than 2.5, they would make an average loss of 1.46 percent. There is thus a small bias in which better odds are typically offered on there being fewer than 2.5 goals. This, however, is not enough to be exploitable in terms of making a profit.

Season	G	S	ST	C	S + C	ST + C
2005/06	+35.14	+14.16	+0.95	−6.50	+41.54	−7.38
2006/07	−18.50	−37.09	−88.54	+0.12	+11.81	−3.33
2007/08	−51.35	+80.93	+47.69	+45.70	+65.74	+45.20
2008/09	−115.48	−26.17	+46.45	−12.60	−39, 82	−1.57
2009/10	−70.59	+76.49	−14.41	−39.40	+66.06	−10.01
2010/11	−106.79	+58.15	−9.07	+55.23	+45.22	+39.80
2011/12	−28.46	+173.16	+131.73	−13.41	+127.94	+88.79
2012/13	−31.59	+38.15	+16.50	+9.61	+91.54	+47.25
2013/14	−19.60	+13.44	−16.41	+29.17	+17.38	−4.38
2014/15	−99.30	−49.63	−86.33	+1.88	+16.64	−19.26
2015/16	−82.91	−13.05	−127.51	−94.08	+5.29	−110.63
2016/17	−2.91	−18.06	−102.92	−56.59	−13.65	−102.91
2017/18	+31.50	+89.51	+53.51	+105.94	+99.36	+88.45
Total	−631.12	+414.72	−148.36	+25.07	+535.01	+50.02

Table 3: Total profit in each season from using the Level Stakes strategy with maximum BetBrain odds when using goals (G), shots (S), shots on target (ST), corners (C), shots and corners (S+C) and shots on target and corners (ST+C) as the input to the GAP ratings. The input that results in the highest profit is highlighted in green.

Whilst the above results consider the performance of the forecasts over all leagues considered, it is also of interest to assess the performance over individual leagues. The overall profit that would have been made in each

Figure 1: Cumulative profit over time from using the Level Stakes betting strategy with the maximum BetBrain odds for each input to the GAP ratings.

league under the Level Stakes strategy is shown in table 4 for each GAP rating input. The strategy appears to be fairly robust. For example, using shots and corners as the GAP rating input would have made a profit in six out of the ten leagues considered with the English Premier League and Championship returning the largest profits. In fact, for each league considered, at least one of the inputs would have resulted in a profit. Notably, however, using goals as the GAP rating input would have resulted in an overall loss in all ten leagues considered.

The results above demonstrate that using the forecasts with the Level Stakes betting strategy would have produced a cumulative profit over the seasons considered. However, from these results alone, it is not possible to tell whether this could have occurred entirely by chance, i.e. the forecasts

League	G	S	ST	C	S + C	ST + C
English Premier League	−56.14	+129.70	−46.19	+41.18	+139.63	−10.72
English Championship	−82.19	+187.79	+34.00	+78.91	+204.05	+62.29
English League One	−27.24	−40.74	−26.60	−68.85	−14.14	−31.40
English League Two	−44.11	+3.62	+9.93	−22.93	−2.16	−49.36
English National League	−47.30	+21.39	+90.31	+8.38	+99.58	+97.05
Scottish Premier League	−43.49	+35.53	−33.33	+133.27	+80.82	+22.32
Spanish La Liga	−105.74	+41.74	+6.58	−33.79	+44.38	+23.18
French Ligue One	−98.56	+26.82	−70.57	−53.97	−33.45	−31.15
Italian Serie A	−65.56	−30.92	−69.15	−62.19	−26.65	−8.75
German Bundesliga	−60.79	+39.79	−43.34	+5.06	+42.95	−23.44
Combined	−631.12	+414.72	−148.36	+25.07	+535.01	+50.02

Table 4: Total profit for each league, along with the combined profit, from using the Level Stakes strategy with maximum BetBrain odds when using goals (G), shots (S), shots on target (ST), corners (C), shots and corners (S+C) and shots on target and corners (ST+C) as the input to the GAP ratings. The input that results in the highest profit is highlighted in green.

got lucky in identifying bets that resulted in a profit over time. In order to assess this, actual betting performance is compared with the results that would typically be expected were the gambler to bet randomly with the same probabilities. A simple approach is therefore taken whereby the number of bets actually taken in a season is preserved but those bets are chosen randomly, rather than using the outlined betting strategy. This process is repeated multiple times with different realisations to give a range of possible scenarios of how the profit and loss might turn out when this is done. If the actual results are typical of what happens when bets are made randomly, then little confidence should be had in the performance of the forecasts for identifying value bets. In figure 2, the results of doing this using shots and corners as the input to the GAP ratings are shown in the upper panel. Here, the green line shows the actual profit and loss over time whereas each black line shows the profit and loss achieved by randomly choosing bets under different realisations. In the lower panel, the red line shows the profit/loss obtained from only betting when the forecasts suggest negative value and the black lines show the profit/loss from betting randomly at the same frequency. Since the actual profit made is far higher than what would be typical if the bets were chosen randomly, a great deal of confidence is gained that the forecasts are informative for the purpose of this strategy. Moreover, it is clear that the forecasts are also capable of identifying bets that offer negative value. The latter could potentially be useful when gambling on a betting exchange in which gamblers are able to act like a bookmaker and 'lay' bets for other gamblers. Given the lack of available betting exchange data, this is not tested in this paper, however.

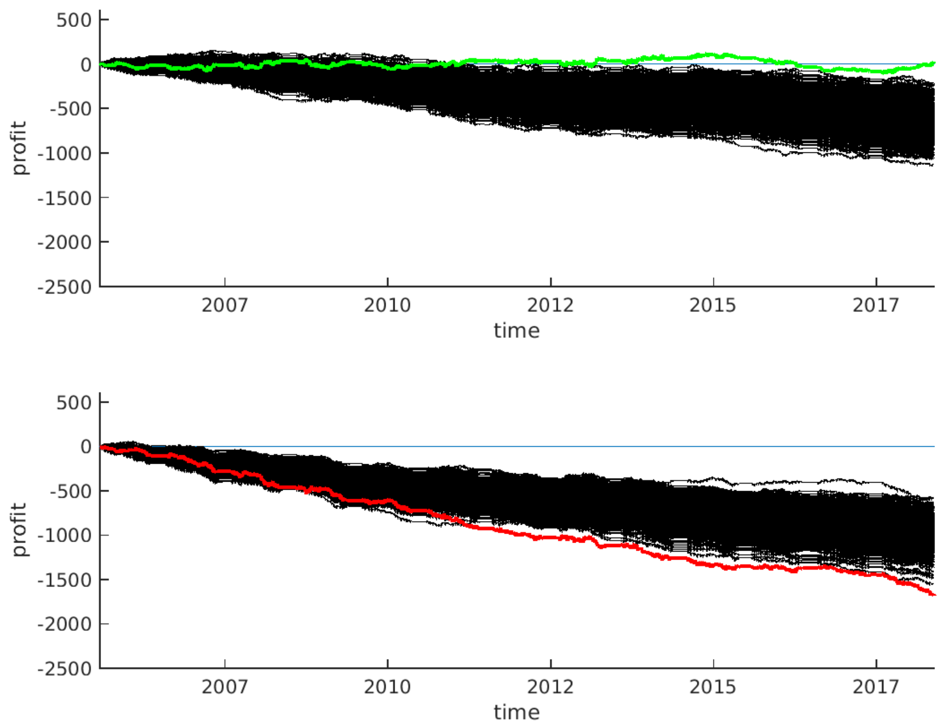

Figure 2: Upper panel: Cumulative profit as a function of time using the Level Stakes betting strategy (green line) and when bets are randomly taken at the same frequency (black lines) using maximum BetBrain odds. Lower panel: Cumulative profit as a function of time betting only when the forecast suggests there is negative value (red line) and when bets are taken randomly at the same frequency (black lines).

In order to illustrate the results of the above process for each GAP rating input, the proportion of random bet scenarios with a higher profit/loss is shown in figure 3 (note the logarithmic scale) as a function of time (calculated at the end of each season). Here, in all cases other than when goals are used as inputs to the GAP ratings, the betting strategy is eventually able to outperform all random bet scenarios tested. When goals are used as the input, on the other hand, there is little evidence that betting performance is better than that which could be achieved by chance.

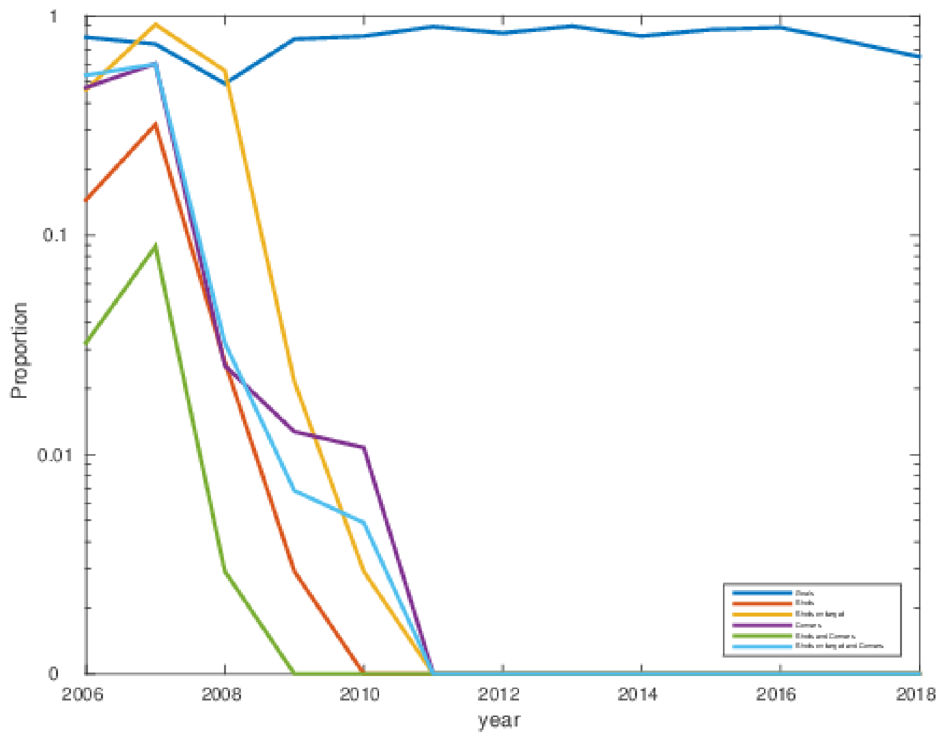

Figure 3: The proportion of random bet scenarios with higher cumulative profit as a function of time for each GAP rating input considered when using maximum BetBrain odds.

Whilst the results given above suggest that the Level Stakes betting strategy is able to identify enough value bets to make a robust long term profit for certain GAP rating inputs, it is important to rule out the possibility that the forecasts don't simply identify arbitrage opportunities. An arbitrage opportunity is a case in which opportunities exist for strategically placing bets with bookmakers that have differing odds and guaranteeing a profit. Such an opportunity is available when the overround is less than zero. The mean overround of bets selected by the forecasts and betting strategy in each league and for each GAP rating input is shown in table 5. Here, in all cases, the mean overround is greater than zero, showing that a profit is made despite the built in bookmakers' profit margin of the chosen bets and hence the favourable results do not simply occur as a result of identifying arbitrage opportunities.

League	G	S	ST	C	S + C	ST + C
English Premier League	0.011	0.011	0.011	0.011	0.012	0.011
English Championship	0.013	0.014	0.014	0.013	0.014	0.014
English League One	0.016	0.016	0.016	0.016	0.016	0.016
English League Two	0.016	0.016	0.016	0.016	0.016	0.016
English National League	0.017	0.017	0.018	0.018	0.018	0.018
Scottish Premier League	0.015	0.016	0.016	0.016	0.016	0.016
Spanish La Liga	0.012	0.012	0.012	0.012	0.012	0.012
French Ligue One	0.015	0.015	0.015	0.015	0.013	0.014
Italian Serie A	0.012	0.016	0.012	0.013	0.012	0.013
German Bundesliga	0.012	0.014	0.013	0.012	0.012	0.012
Combined	0.014	0.014	0.014	0.014	0.014	0.014

Table 5: Mean overround of bets placed in each league, along with that of all the leagues combined, when using goals (G), shots (S), shots on target (ST), corners (C), shots and corners (S+C) and shots on target and corners (ST+C) as inputs to the GAP ratings and using maximum BetBrain odds.

The forecasts considered in this paper are formed using logistic regression with the sum of both the attacking and defensive ratings for the home and away team as a single term (equation 3). However, an alternative is to use the GAP ratings as separate inputs to allow each of the ratings to impact the forecasts to a potentially different extent. Although details of the results are omitted here, this was tested over all eligible matches (i.e. more than six matches into the season and more than six matches before the end of the season) for each GAP rating input to assess whether this improved the forecasts. The resulting forecasts were found to have less support than the forecasts considered in this paper, when tested using Akaike's Information Criterion (AIC). These results imply that there is little evidence to distinguish the importance of each rating in terms of building the forecasts and so this is not considered further.

Another issue worth visiting is to check that the profit/loss achieved using the Level Stakes betting strategy does not simply occur as a result of testing too many inputs to the GAP ratings and finding one that works well in terms of making a profit by chance. A simple way to account for multiple testing is to use the Bonferroni correction (Bonferroni (1936)). The Bonferroni correction accounts for multiple tests by adjusting the significance level α to account for the number of hypotheses tested. Under the Bonferroni correction, if a significance level of α is required for a single test, and m tests are performed, the required significance level of each of the tests is adjusted to $\frac{\alpha}{m}$.

In figure 3, the profit/loss from each GAP rating input was compared with that achieved from randomly betting at the same frequency. This was done by testing 1024 different random sets of bets. The proportion of randomly selected sets of bets achieving a greater profit than that of the Level Stakes betting strategy can then be interpreted as an estimated p-value of the profit/loss. Here, this is done again but with one million random sets of bets in order to estimate the p-value more accurately. This can then be compared to adjusted significance levels calculated using the Bonferroni correction. The estimated p-value for each GAP rating input is shown in table 6, along with the proportion of random sets of bets that would have returned a profit. Each are compared with adjusted significance levels at the 5%, 1% and 0.1% levels (denoted with one, two and three stars respectively). Since the profit/loss for each GAP rating input other than goals is highly significant even when multiple testing is accounted for, the results in this paper do not appear to result from testing too many inputs. In addition, the fact that, in each case, only a handful of randomly chosen sets would have resulted in a profit demonstrates the difficulty of profiting over this number of bets without truly having some ability to identify value.

GAP rating input	Estimated p-value	Prop. trials with profit
Goals	0.541765	0.000017
Shots	0.000000***	0.000009
Shots on Target	0.000124**	0.000000
Corners	0.000003***	0.000004
Shots and Corners	0.000000***	0.000008
Shots on Target and Corners	0.000000***	0.000000

Table 6: Estimated p-values for each GAP rating input along with the proportion of random sets of bets returning a profit when using maximum BetBrain odds. Estimated p-values with one, two and three stars next to them indicate significance at the 5%, 1% and 0.1% levels respectively, adjusted using the Bonferroni correction.

All of the above results make use of the Level Stakes strategy in which a unit bet is taken if a forecast implies the odds offer 'value'. However, in section 8, another strategy, called the Kelly strategy, was defined in which the stake is allowed to vary based on the difference in implied probability between the forecast and the odds on offer. Odds perceived to have greater 'value' are assigned higher stakes. Here, the profit/loss of applying this strategy with the forecasts defined in this paper is demonstrated. In each case, the stakes are adjusted so that the *average* stake is one. This makes the profit/loss directly comparable with the level stakes strategy which has formed the basis of the results so far.

The profit/loss of using each GAP rating input along with the Kelly strategy with maximum BetBrain odds is shown as a function of time in figure 4. Here, the solid lines represent the profit/loss of the Kelly strategy whilst, for comparison, the dotted lines represent the profit/loss of the Level Stakes strategy. The results here are striking. In all cases, there is an improvement in terms of profit, whilst forecasts formed using each of the measures of attacking performance other than goals result in a profit. This represents a substantial improvement and highlights the potential for the use of alternative betting strategies alongside the forecasts produced.

10.2 Average odds

Selected results are now repeated for the case in which average rather than maximum BetBrain odds are assumed to be available. The cumulative profit under each GAP rating input both for the Level Stakes and Kelly betting strategies are shown in figure 5 as a function of time. In contrast to the maximum odds case, a substantial loss is made, demonstrating that the forecasts do not appear to be informative enough to overcome the substantial overround in this case (around 6.8 percent).

Despite the substantial loss, it is worthwhile to consider the performance of each of the betting strategies compared to a case in which bets are randomly taken at the same frequency. Similarly to figure 3, the proportion of random bet scenarios (i.e scenarios in which the same number of bets are selected as under the Level Stakes betting strategy but randomly rather than using the forecast probability) in which the cumulative profit is higher than that achieved by the Level Stakes strategy is shown at the end of each season in figure 6. Here, despite the fact that a loss is made, the strategy performs better than would be the case were bets taken randomly. This means that, whilst the strategy is unable to find enough value bets to make a long term profit, it is able to identify bets that have a low 'negative value' and is there-

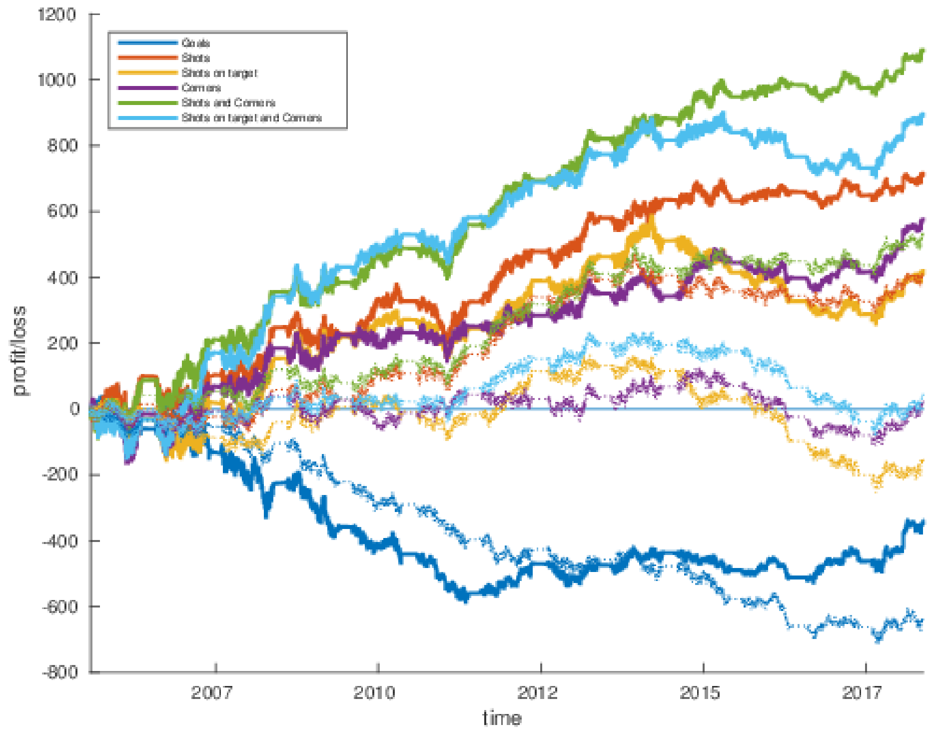

Figure 4: Cumulative profit/loss of using GAP ratings with each input alongside the Kelly strategy (solid lines) and the level stakes strategy (dotted lines) when using maximum odds.

fore successful in reducing losses over time (though in practice it would be better simply not to bet at all).

10.3 Contrasting Betting Performance Under Maximum and Average Odds

There is a stark difference between the results achieved when assuming maximum and average Betbrain odds. In the former case, a substantial and robust profit is made whilst, in the latter, there is a clear loss. It is worth considering the implication of this both in terms of the performance of the forecasts and the efficiency of the market. To make a profit under a value betting strategy

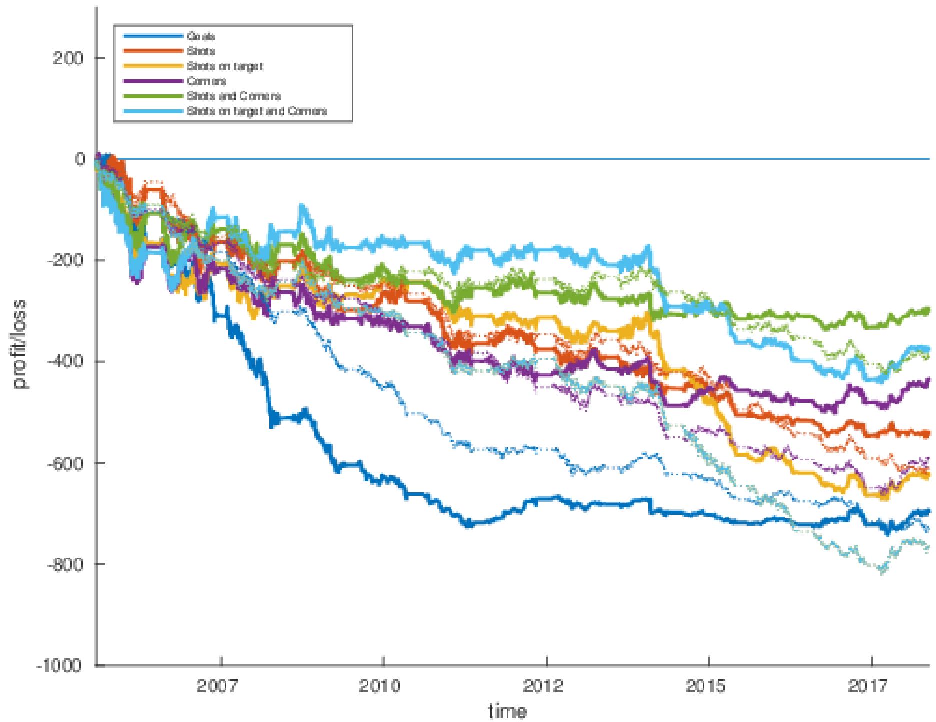

Figure 5: Cumulative profit/loss of using GAP ratings with each input alongside the Kelly strategy (solid lines) and the level stakes strategy (dotted lines) when using average BetBrain odds.

such as the Level Stakes and Kelly strategies, two essential ingredients are required. Firstly, there must be odds available in which the probability of the outcome is higher than the odds-implied probability, that is odds that offer *value*. If such opportunities exist, it can be said that the market is inefficient since a market that takes into account all information would be reflective of the true probability (plus a profit margin). The second requirement is that the forecasts are able to identify such opportunities. Someone in possession of the true probabilities would have the tools to do this and would therefore easily be able to make a profit over time by only betting on odds that offer value. In practice, given the inherent limitations in all mathematical models, the forecasts are not expected to reflect the probabilities of each outcome

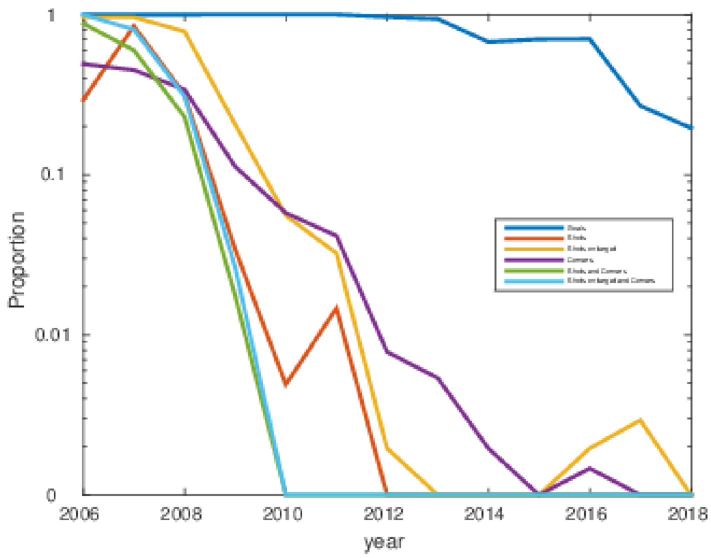

Figure 6: The proportion of random bet scenarios with higher cumulative profit as a function of time for each GAP rating input considered when using average BetBrain odds.

perfectly. Given this, in order to make a long term expected profit, the expected profit from value bets that are successfully identified must not be cancelled out by bets that are incorrectly selected and in fact have negative value.

The fact that a robust profit is made in the maximum odds case implies that both of the requirements to make a profit are satisfied, that is the market defined by the maximum BetBrain odds is inefficient and that the forecasts are informative enough to identify value bets. For the average odds case, the fact that a loss is made has a profound implication in that it confirms that the forecasts cannot represent perfect probabilities. This is because, if the average odds market were efficient, no bets would be taken and the profit/loss would be zero. If the market were inefficient, value bets would be identified and an expected profit made. The substantial loss can therefore only result from limitations in the performance of the forecast probabilities. In addition, since a profit cannot be demonstrated, it is impossible to conclude whether the average odds market is inefficient or not.

The results above clearly show that the overround in the odds impacts the profitability of both betting strategies. Whilst using the maximum odds clearly results in a sustained profit, using the average odds results in a loss. In fact, there is a substantial difference in the overround of the maximum odds and the average odds. In the former case, the mean overround is 1.57 percent whilst, in the latter, it is a substantially higher 6.8 percent. In practice, finding and betting with the maximum odds may be a prohibitively time consuming task. However, by carefully choosing a selection of bookmakers and selecting the best odds among those, a substantially lower effective overround may be possible than if the average odds were taken. Another potential way to reduce the overround is to use a betting exchange. Betting exchanges often have a far lower effective overround than bookmakers, even once commission is taken into account. For example, if even odds (i.e. decimal odds of 2) are offered on both outcomes in the over/under 2.5 goal market, a commission rate of 5 percent such as that charged by Betfair would result in effective odds of 1.95 on both outcomes, corresponding to an effective overround of $\frac{1}{1.95} + \frac{1}{1.95} - 1 = 0.026$ or 2.6 percent.

The performance of the Level Stakes and Kelly betting strategies in a setting in which the odds on offer are higher than the average but lower than the maximum BetBrain odds is now considered. To find odds that are a weighted average of the two, the following relation is used:

$$\tilde{r} = \frac{1}{\omega \frac{1}{r_i^{\max}} + (1-\omega)\frac{1}{r_i^{\text{ave}}}} \tag{5}$$

where $r_i^{\max}$ and r_i^{ave} are the maximum and average odds respectively, ω determines the distance between them, and the resulting odds are $\tilde{r}$. The reason the odds are inverted first is so that they are weighted in terms of their odds-implied probabilities. The overall profit from both strategies is shown as a function of ω in figure 7 under both the Level Stakes strategy (top) and the Kelly strategy (bottom). Here, as previously shown, for both strategies, using the average odds produces a loss for all GAP rating inputs. However, a profit can still be made when the odds are substantially lower than the maximum and this suggests that the forecasts may be effective in achieving a profit even when the best odds are not available or it is not practical to compare the odds of a large number of bookmakers.

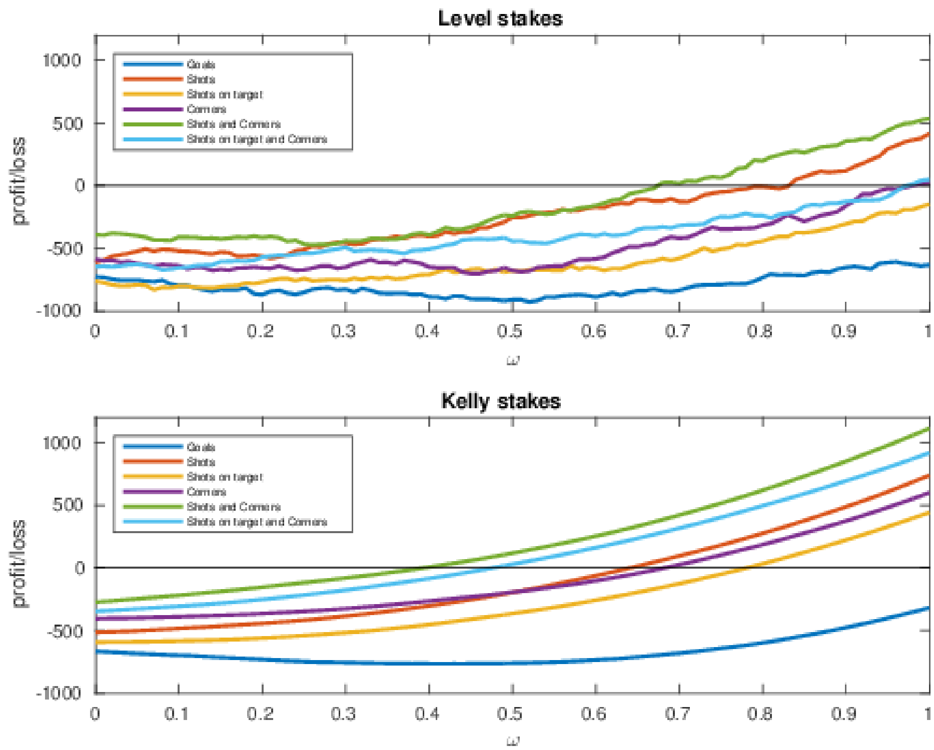

Figure 7: Total profit as a function of ω for the level stakes (top) and Kelly (bottom) betting strategies for each GAP rating input.

10.4 Assessing Forecast Accuracy

Whilst the profit made is a useful indication of the performance of the forecasts under each GAP rating input, it is also of interest to evaluate the forecast skill (accuracy). The advantage of this is twofold. Firstly, someone may wish to use the forecasts for some other purpose, perhaps with some different betting strategy or even for some different purpose altogether such as choosing whether to attend a match or not! Secondly, it is interesting to see whether there is general agreement between the relative forecast skill and betting performance when using each of the GAP rating inputs. The mean ignorance score of the forecasts under each GAP rating input is shown for each league in table 7. Here, there appears to be reasonable agreement between the skill of the forecasts and the betting performance. This is a useful observation since, if increased forecast skill generally results in larger profits, some confidence can be had that optimising the parameters with respect to the ignorance score is a reasonable thing to do. This, in fact, need not be the case, however. In the two betting strategies outlined in section 8, a profit is expected to be made when the 'true' probability is higher than the odds implied probability and thus a higher profit will be expected from forecasts that are able to identify value bets more often. The ignorance score, on the other hand, favours forecasts that place the most probability on the outcome, on average. It is therefore the case that, for the Level Stakes betting strategy, the forecasts need not be the most accurate, they just need to be capable of successfully identifying enough value bets to make a profit.

There is, in fact, a connection between the ignorance score and gambling returns. It can be shown that, if in a series of bets, a gambler sets the proportion of their wealth placed on each outcome to be their estimated probability distribution, their expected return is 2 raised to the power of the mean ignorance of the house (note that the ignorance score of the house can be calculated even if the sum of the odds-implied probabilities do not add up to one; this, however, gives the house a distinct advantage) relative to that of the gambler (Roulston & Smith (2002)). For both of the betting strategies in this paper, however, bets are only taken if the odds offer 'value'. The efficacy of both strategies can thus only be tested empirically.

11 Discussion

In this paper, the GAP rating system has been defined and demonstrated as a basis with which to build probabilistic forecasts. Alongside two value betting strategies, the resulting forecasts have been shown to be capable of producing a robust profit in the over/under 2.5 goal betting market using match data as inputs. At first glance, the results may seem somewhat counterintuitive since one might have expected the number of goals *actually* scored in past matches to be the main factor impacting the number of goals scored in future matches. In football, however, perhaps more than in many other sports, scoring is a

League	G	S	ST	C	S + C	ST + C
English Premier League	−1.36	−1.48	−1.46	−1.52	−1.62	−1.48
English Championship	−0.39	−0.39	−0.39	−0.38	−0.58	−0.39
English League One	+0.05	+0.12	+0.12	+0.08	+0.04	+0.12
English League Two	+0.08	+0.12	+0.11	+0.12	−0.00	+0.12
English National League	−0.33	−0.52	−0.52	−0.54	−0.47	−0.52
Scottish Premier League	−0.38	−0.62	−0.41	−1.02	−0.63	−0.62
Spanish La Liga	−3.08	−3.36	−3.44	−3.23	−3.12	−3.36
French Ligue One	−1.37	−1.79	−1.59	−1.84	−1.99	−1.79
Italian Serie A	−1.05	−1.22	−1.15	−0.98	−0.86	−1.22
German Bundesliga	−2.39	−2.20	−2.17	−2.03	−2.10	−2.22
Combined	−0.87	−0.97	−0.94	−0.96	−0.99	−0.97

Table 7: Mean ignorance score for each league, along with the combined score, when using goals (G), shots (S), shots on target (ST), corners (C), shots and corners (S+C) and shots on target and corners (ST+C) as the input to the GAP ratings. The input that results in the best mean ignorance score is highlighted in green. For clarity, each entry in the table has been multiplied by 100.

relatively difficult task with a substantial element of chance involved. This means that, whilst a team may dominate a match, they may simply be unlucky or have an 'off day' in terms of converting chances. Conversely, a team with relatively few chances may simply get lucky and still manage to score. If one team has a large number of chances and fails to score while another has relatively few but still manages to score at least one goal, this information may not be very informative when predicting the number of goals that might occur in future matches. A team that has more shots and corners is usually far more likely to score than those that achieve fewer. It therefore seems reasonable that shots and corners should provide a better indication of how the game went than the number of goals scored.

The GAP rating system has a number of similarities with another well know rating system in the football forecasting literature called the pi-rating system. These similarities and a number of key differences have been described. One key difference is that, unlike pi-ratings, which estimate the difference in the number of goals scored by two teams, GAP ratings differentiate between the attacking and defensive capabilities of each team. Consequently, whilst there is no obvious way of using pi-ratings to create informative forecasts for the over/under market (since the expected difference in the number of goals scored does not give a strong indication of the total number of goals scored), GAP ratings are ideally suited to this. GAP ratings are also directly applicable to the match outcome market in which the potential profitability of pi-ratings has been demonstrated (Constantinou & Fenton (2013)). Given the demonstrated success of using match statistics rather than goals as inputs for the ratings, an interesting avenue for further research is to test both the pi-rating and GAP rating systems with different match statistics as inputs in this market. In fact, this will form the basis of another paper (in preparation). Preliminary results suggest that both sets of ratings are capable of yielding a robust profit and that, consistently with the results in this paper, match statistics other than goals tend to produce far more accurate forecasts and a larger profit in each case.

It is worth noting that, whilst the forecasts and betting strategies presented in this paper have been found to perform extremely favourably in that they produce a clear long term profit in the over/under market, there appears to be some evidence that the results from the strategy would have worsened in recent years, since the cumulative profit appears to have 'levelled off' (i.e. the cumulative profit does not appear to have been growing as quickly). It is not clear why this should be the case and this forms the basis of potential future work. One possible explanation for this is that, in recent years, the betting odds have started to take better account of additional performance data such as shots and corners. This seems like a plausible explanation given the increase in the availability of in-match data and greater interest in statistical modelling for sporting events. One explanation that can be ruled out is an upward shift in the typical profit margin of the bookmakers, which has in fact fallen in recent years.

The efficacy of GAP ratings for producing forecasts for the over/under 2.5 goal has been demonstrated and the potential to extend their use to the match outcome market discussed. Testing the performance of the ratings in these two markets is straightforward because historical odds are easily available. However, in the modern game, a vast number of other betting markets are often available. For example, many bookmakers give odds on whether the total number of goals will exceed 1.5, 3.5, 4.5 etc and it would be easy to extend the use of GAP ratings to these markets. Other examples of markets in which the ratings could be utilised include the total number of corners, the half time score and the number of goals scored by a single team. There is also significant potential for the GAP rating system to be applied to other sports, particularly those in which there are well defined statistics other than goals or points that signify attacking performance. In basketball or ice hockey, for example, the number of shots in a game could be used as an input for the ratings. In American football, a measure of attacking performance could be the number of yards gained by each team, the number of times each team got close to scoring or some other similar measure.

The results in this paper demonstrate that measures of attacking performance other than goals can be more informative in predicting future goalscoring performance than goals themselves, so much so that the former can result in profitable betting strategies whilst the latter does not. Given the wide array of match statistics available, there is great potential to come up with innovative measures of the attacking performance of football teams. The GAP rating system described provides a solid basis on which to incorporate these data to build informative and profitable probabilistic forecasts.

A Parameter Values

The following tables show the parameter values selected at the beginning of the stated season for a given GAP rating input.

Shots

Season	λ	ϕ_1	ϕ_2
2005/06	0.459	0.481	0.566
2006/07	0.449	0.449	0.700
2007/08	0.446	0.478	0.654
2008/09	0.476	0.480	0.689
2009/10	0.405	0.477	0.852
2010/11	0.426	0.476	0.813
2011/12	0.464	0.491	0.660
2012/13	0.468	0.491	0.637
2013/14	0.451	0.491	0.560
2014/15	0.407	0.492	0.568
2015/16	0.437	0.496	0.566
2016/17	0.445	0.497	0.551
2017/18	0.437	0.496	0.566

Goals

Season	λ	ϕ_1	ϕ_2
2005/06	0.395	0.109	0.909
2006/07	0.103	0.569	0.490
2007/08	0.474	0.144	0.924
2008/09	0.478	0.487	0.391
2009/10	0.210	0.566	0.439
2010/11	0.226	0.495	0.483
2011/12	0.158	0.534	0.500
2012/13	0.121	0.559	0.504
2013/14	0.121	0.519	0.520
2014/15	0.424	0.492	0.542
2015/16	0.086	0.595	0.528
2016/17	0.075	0.534	0.665
2017/18	0.078	0.515	0.847

Shots and Corners

Season	λ	ϕ_1	ϕ_2
2005/06	0.183	0.491	0.739
2006/07	0.455	0.454	0.715
2007/08	0.450	0.459	0.710
2008/09	0.435	0.477	0.671
2009/10	0.475	0.483	0.625
2010/11	0.466	0.484	0.650
2011/12	0.469	0.496	0.603
2012/13	0.481	0.492	0.553
2013/14	0.422	0.492	0.541
2014/15	0.436	0.497	0.543
2015/16	0.452	0.499	0.535
2016/17	0.425	0.497	0.543
2017/18	0.436	0.493	0.546

Shots on Target

Season	λ	ϕ_1	ϕ_2
2005/06	0.176	0.551	0.724
2006/07	0.449	0.464	0.735
2007/08	0.466	0.471	0.694
2008/09	0.450	0.478	0.762
2009/10	0.413	0.482	0.868
2010/11	0.434	0.475	0.800
2011/12	0.500	0.485	0.738
2012/13	0.480	0.491	0.734
2013/14	0.467	0.490	0.786
2014/15	0.478	0.490	0.793
2015/16	0.499	0.504	0.804
2016/17	0.494	0.513	0.816
2017/18	0.500	0.504	0.796

Shots on Target and Corners

Season	λ	ϕ_1	ϕ_2
2005/06	0.163	0.510	0.795
2006/07	0.445	0.463	0.723
2007/08	0.407	0.460	0.779
2008/09	0.434	0.475	0.662
2009/10	0.423	0.484	0.622
2010/11	0.445	0.484	0.653
2011/12	0.432	0.495	0.614
2012/13	0.444	0.503	0.558
2013/14	0.480	0.491	0.567
2014/15	0.422	0.496	0.560
2015/16	0.432	0.509	0.541
2016/17	0.441	0.517	0.557
2017/18	0.490	0.519	0.552

Corners

Season	λ	ϕ_1	ϕ_2
2005/06	0.173	0.468	0.818
2006/07	0.464	0.495	0.763
2007/08	0.405	0.474	0.771
2008/09	0.388	0.482	0.640
2009/10	0.321	0.485	0.572
2010/11	0.366	0.503	0.581
2011/12	0.342	0.498	0.571
2012/13	0.311	0.491	0.548
2013/14	0.348	0.493	0.552
2014/15	0.426	0.488	0.547
2015/16	0.297	0.487	0.525
2016/17	0.297	0.490	0.538
2017/18	0.426	0.488	0.547

GAP rating input	α	β_1	β_2
Goals	-2.0629	-0.0026	$+4.0918$
Shots	-2.3009	$+0.0070$	$+3.8969$
Shots on Target	-2.1576	$+0.0086$	$+3.9135$
Corners	-2.4001	$+0.0176$	$+3.9910$
Shots and Corners	-2.4287	$+0.0070$	$+3.8912$
Shots on Target and Corners	-2.3010	$+0.0081$	$+3.8815$

Table 8: Logistic regression parameter values for each GAP rating input, calculated using all available years of data.

Acknowledgements

I am indebted to two anonymous reviewers and a handling editor who provided invaluable feedback on earlier versions of this paper.

References

References

Baker, R. D. & McHale, I. G. (2015), 'Time varying ratings in association football: the all-time greatest team is..', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178**(2), 481–492.

Berrar, D., Lopes, P., Davis, J. & Dubitzky, W. (2019), 'Guest editorial: special issue on machine learning for soccer', *Machine Learning* **108**(1), 1–7.

Bonferroni, C. (1936), 'Teoria statistica delle classi e calcolo delle probabilita', *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* **8**, 3–62.

Boshnakov, G., Kharrat, T. & McHale, I. G. (2017), 'A bivariate weibull count model for forecasting association football scores', *International Journal of Forecasting* **33**(2), 458–466.

Bröcker, J. & Smith, L. A. (2008), 'From ensemble forecasts to predictive distribution functions', *Tellus A* **60**(4), 663–678.
URL: *http://dx.doi.org/10.1111/j.1600-0870.2008.00333.x*

Carbone, J., Corke, T. & Moisiadis, F. (2016), 'The rugby league prediction model: Using an elo-based approach to predict the outcome of national rugby league (nrl) matches', *International Educational Scientific Research Journal* **2**(5), 26–30.

Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D. & Malvaldi, M. (2015), The harsh rule of the goals: Data-driven performance indicators for football teams, *in* 'Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on', IEEE, pp. 1–10.

Constantinou, A. C. (2019), 'Dolores: a model that predicts football match outcomes from all over the world', *Machine Learning* **108**(1), 49–75.

Constantinou, A. C. & Fenton, N. E. (2013), 'Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries', *Journal of Quantitative Analysis in Sports* **9**(1), 37–50.

Constantinou, A. C., Fenton, N. E. & Neil, M. (2012), 'pi-football: A bayesian network model for forecasting association football match outcomes', *Knowledge-Based Systems* **36**, 322 – 339.

Constantinou, A. & Fenton, N. (2017), 'Towards smart-data: Improving predictive accuracy in long-term football team performance', *Knowledge-Based Systems* **124**, 93–104.

Cronin, B. (2019), 'How important are the first six games of the season?', `https://www.pinnacle.com/en/betting-articles/Soccer/importance-of-first-six-games-in-soccer/J3R2FJ6DY6MBZVW2`. Accessed: 18/04/2019.

Dixon, M. J. & Coles, S. G. (1997), 'Modelling association football scores and inefficiencies in the football betting market', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(2), 265–280.

Dixon, M. J. & Pope, P. F. (2004), 'The value of statistical forecasts in the uk association football betting market', *International journal of forecasting* **20**(4), 697–711.

Eggels, H. (2016), Expected goals in soccer: Explaining match results using predictive analytics, *in* 'The Machine Learning and Data Mining for Sports Analytics workshop', p. 16.

Elo, A. E. (1978), *The rating of chessplayers, past and present*, Arco Pub.

ESPN (2018), 'Espn's nba basketball power index playoff odds', `http://www.espn.co.uk/nba/story/_/page/BPI-Playoff-Odds/espn-nba-basketball-power-index-playoff-odds`. Accessed: 07/07/2018.

Fifa (2018), 'Revision of the fifa / coca-cola world ranking', `https://resources.fifa.com/image/upload/fifa-world-ranking-technical-explanation-revision.pdf?cloudid=edbm045h0udbwkqew35a`. Accessed: 27/04/2019.

FiveThirtyEight (2019*a*), 'The complete history of the nfl', `https://projects.fivethirtyeight.com/complete-history-of-the-nfl/`. Accessed: 29/03/2019.

FiveThirtyEight (2019*b*), 'Nba elo ratings', `https://fivethirtyeight.com/tag/nba-elo-ratings/`. Accessed: 29/03/2019.

Football Data (2018), 'Football data', `http://www.football-data.co.uk/`. Accessed: 09/10/2018.

Goddard, J. (2005), 'Regression models for forecasting goals and match results in association football', *International Journal of forecasting* **21**(2), 331–340.

Good, I. J. (1952), 'Rational decisions', *Journal of the Royal Statistical Society: Series B* **14**, 107–114.

Huang, K.-Y. & Chang, W.-L. (2010), A neural network method for prediction of 2006 world cup football game, *in* 'Neural Networks (IJCNN), The 2010 International Joint Conference on', IEEE, pp. 1–8.

Hubáček, O., Šourek, G. & Železnỳ, F. (2019), 'Learning to predict soccer results from relational data with gradient boosted trees', *Machine Learning* **108**(1), 29–47.

Hucaljuk, J. & Rakipović, A. (2011), Predicting football scores using machine learning techniques, *in* 'MIPRO, 2011 Proceedings of the 34th International Convention', IEEE, pp. 1623–1627.

Hvattum, L. M. & Arntzen, H. (2010), 'Using elo ratings for match result prediction in association football', *International Journal of forecasting* **26**(3), 460–470.

Joseph, A., Fenton, N. E. & Neil, M. (2006), 'Predicting football results using bayesian nets and other machine learning techniques', *Knowledge-Based Systems* **19**(7), 544–553.

Karlis, D. & Ntzoufras, I. (2003), 'Analysis of sports data by using bivariate poisson models', *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**(3), 381–393.

Kelly Jr, J. (1956), 'A new interpretation of the information rate', *Bell System Technical Journal* **35**, 917–926.

Lee, A. J. (1997), 'Modeling scores in the premier league: is manchester united really the best?', *Chance* **10**(1), 15–19.

Maher, M. J. (1982), 'Modelling association football scores', *Statistica Neerlandica* **36**(3), 109–118.

Opta (2018), 'Assessing the performance of premier league goalscorers', `https://www.optasportspro.com/about/optapro-blog/posts/2012/blog-assessing-the-performance-of-premier-league-goalscorers/`. Accessed: 09/10/2018.

Opta Football (2018), 'Opta Football', `https://www.optasports.com/sports/football/`. Accessed: 09/10/2018.

O'Donoghue, P., Dubitzky, W., Lopes, P., Berrar, D., Lagan, K., Hassan, D., Bairner, A. & Darby, P. (2004), 'An evaluation of quantitative and qualitative methods of predicting the 2002 fifa world cup', *Journal of Sports Sciences* **22**(6), 513–514.

Rathke, A. (2017), 'An examination of expected goals and shot efficiency in soccer'.

Roulston, M. S. & Smith, L. A. (2002), 'Evaluating probabilistic forecasts using information theory', *Monthly Weather Review* **130**, 1653–1660.

Rue, H. & Salvesen, O. (2000), 'Prediction and retrospective analysis of soccer matches in a league', *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**(3), 399–418.

Suznjevic, M., Matijasevic, M. & Konfic, J. (2015), Application context based algorithm for player skill evaluation in moba games, *in* '2015 International Workshop on Network and Systems Support for Games (NetGames)', IEEE, pp. 1–6.

The Economist (2017), 'How data changed gambling - the economist explains', `https://www.economist.com/the-economist-explains/2017/07/19/how-data-changed-gambling`. Accessed: 07/10/2018.

Van Haaren, J. & Van den Broeck, G. (2015), Relational learning for football-related predictions, *in* 'Latest Advances In Inductive Logic Programming', World Scientific, pp. 237–244.