ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting football match results using a player rating based model

Benjamin Holmes [a,b], Ian G. McHale [a,*]

[a] Centre for Sports Business, University of Liverpool Management School, UK
[b] Department of Mathematics, University of Liverpool, UK

ARTICLE INFO

Keywords:
Sports forecasting
Football
Betting
Rating
Ranking

ABSTRACT

The paper presents a model for forecasting the results of football matches, which takes into account the abilities of the players on each team. The advantage of this approach is that the dynamic nature of team strengths is incorporated into the model directly. We test our model against the bookmaker's predictions and in a Kelly-type betting strategy applied to the pre-match win/draw/loss market. The new model results in significant positive returns to betting.

1. Introduction

Models for predicting the outcomes of football matches use historical information on the teams competing to obtain team ratings. These estimated team ratings are then used to generate estimated probabilities of the result (win, draw, loss) or scoreline (0-0, 1-0, 0-1, etc.). For example, arguably the most famous of all models for football match scorelines, the Dixon and Coles (1997) model (which is itself based on Maher, 1982), uses historical information on goals scored and conceded by teams to estimate team attack and defence strengths.

Much of the literature on forecasting models in football has focused on allowing the estimated strengths to vary with time. For example, Dixon and Coles (1997) apply a down-weighting in the likelihood function so that matches played further in the past influence a team's estimated strength less than matches played more recently; Baker and McHale (2015) assume team strengths vary deterministically over a long time period; and Crowder et al. (2002), Owen (2011), and Koopman and Lit (2015) adopt models that allow the strengths to vary stochastically from match to match.

Although work has been done on allowing team strengths to vary, little attention has been paid to why team strengths vary and the physical mechanism that drives these variations in strengths from match to match and season to season. It seems likely that the leading cause of the dynamic nature of team strengths is that the identity of the team members varies, and the changing quality of these players means the team strengths themselves change.

In this paper, we focus not on estimating teams' ratings to forecast future results but on using player ratings to forecast team results. In doing so, we deal directly with the mechanism that causes team strengths to vary. Consequently, our modelling framework results in a well-performing model, in terms of forecasting accuracy and when compared with the betting market, despite its simplicity.

The paper is structured as follows. In the following section, we review the recent literature on forecasting models in football. Section 3 describes the data we use. Section 4 then introduces our player ratings system. Models estimating the level of interaction between two opposing players are introduced in Section 5. Section 6 presents several models for forecasting results using the Skellam distribution for modelling goal difference. The results of out-of-sample predictions and a betting simulation are detailed in Section 7. Finally, Section 8 concludes the

* Corresponding author.
E-mail address: ian.mchale@liverpool.ac.uk (I.G. McHale).

https://doi.org/10.1016/j.ijforecast.2023.03.002

Please cite this article as: B. Holmes and I.G. McHale, Forecasting football match results using a player rating based model. International Journal of Forecasting (2023), https://doi.org/10.1016/j.ijforecast.2023.03.002.

paper with a summary of our findings and thoughts for future work.

2. Recent literature

Since Maher (1982) and Dixon and Coles (1997), many published models for forecasting the scorelines (and/or results) have continued to use the same basic specification. Team attack and defence strengths are estimated, and a team's attack strength interacts with the opposing team's defence strength and vice versa. The beauty of this specification cannot be understated. It represents the reality of football: the attackers on one team interact with the opposition's defending players. Boshnakov et al. (2016) follow the lead of Maher (1982) in their model specification and use a bivariate Weibull count distribution as the underlying probability distribution for the counts of goals.

The Elo rating system has been used to model football (see, for example, Hvattum and Arntzen (2010)), estimates team strengths based on previous results, and includes a method for updating the team strengths as new results are recorded. The pi-ratings of Constantinou et al. (2012) and the GAP ratings of Wheatcroft (2020) follow similarly in which team ratings are updated as new information is recorded.

Following the 'Soccer Prediction Challenge' (Dubitzky et al., 2019), a flurry of papers adopting machine learning techniques were published. Berrar et al. (2019) 'won' the competition with an ensemble of gradient-boosted trees. But perhaps most noteworthy is their conclusion that incorporating domain knowledge in forecasting models for football is a key driver of forecasting success. Hubacek et al. (2019) came a close second using a combination of rating models for teams, including pi-ratings, Elo and Google PageRank. Constantinou (2019) performed well using the pi-ratings, as did Tsokos et al. (2019), who used a Poisson model with scoring intensities allowed to vary with time according to an INLA process. In our round-up of machine learning models, we mention da Costa et al. (2021), who estimated the probability of both teams to score using machine learning classifiers but notably used team-level variables.

Despite the efforts of the machine learning community, the marginal gains in terms of predictive accuracy are limited. For example, the best-performing model in the 'Soccer Prediction Challenge' achieved an accuracy of 53.88%, whereas the worst (of the serious entries) had an accuracy of 50.49%. As Berrar et al. (2019) stated, domain knowledge is a key driver to success, and machine learning algorithms have the unattractive property of not representing reality. Like the Maher (1982) and Dixon and Coles (1997), our model has the attractive property of representing the reality of how football is played.

Despite the clear benefits, there have been few attempts to utilise a player-based model for forecasting football match results. Kharrat (2016) and Lasek (2019) use player ratings from the popular FIFA video game franchise in forecasting models. Arntzen and Hvattum (2021) utilised the difference in the simple average of the regularised plus-minus player ratings on the two teams as the basis of a forecasting model. Peeters (2018) do not use player ratings to forecast match results. Instead, they use a simple average of the crowd-sourced player transfer valuations from Transfermarkt.com for the two teams and find that their predictions outperform the team-based rating model they use. A major contribution of the model we propose here is that we do not use simple averages of player ratings on the two teams as the basis for generating the forecasts. Instead, we build a model to mimic how each player on one team interacts with each player on the opposition team.

3. Data

The data requirements of our model are non-trivial–which would be the case for any player-based model. Three required data sources cover the player ratings, match event, and odds data. Each data set was obtained for all seasons from 2013/14 to 2020/21. All processing of data and subsequent modelling was performed using the R programming language (R. Core Team, 2022).

3.1. WhoScored player matchday ratings

Our modelling framework requires individual player ratings of the players on the pitch (the line-ups and identity of players on each team are announced a minimum of 30 minutes before a match and often known well in advance) as inputs into a model for the results of matches. We collected match performance ratings published by WhoScored.com. In total, we used 1,505,177 individual ratings attributed to 24,167 unique players. These ratings spanned 14/07/2013 to 31/05/2021.

3.2. Match event data

Match event data describes all of the actions (shots, passes, tackles, interceptions, etc.) within a match and is becoming more commonplace in football literature. We use such event data as part of a series of multinomial models to estimate the level of interaction between two opposing players (which will be introduced in Section 5). InStat provided the match event data.

For the multinomial models, we required all defensive actions: aerial duels, blocks, ground duels, interceptions, and all shots. In addition, we required the playing positions of players and the formations of both teams. We ensured that position or formation changes during a match were accounted for.

The final forecasting model uses matches within the top five European leagues from seasons 15/16 to 20/21 (this will be discussed further in Section 6.1). The multinomial models are based on actions within the top five leagues from 13/14 to 14/15 to ensure our predictions are fully out-of-sample. The final dataset included 764,712 defensive actions and 108,286 shots.

Match information such as the results, scoreline, identities of teams and players, and formations of the teams were also provided within the InStat data. Further, details of any changes in team formation occurring during a match, and the timing of the change, were recorded.

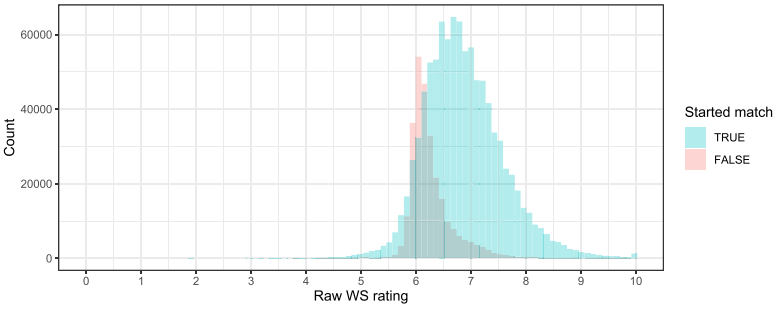

Fig. 1. Histogram of the raw WhoScored ratings achieved by all players. Players are separated by whether they started the match or not.

Table 1
Top ten WhoScored average ratings achieved by players over a two-year rolling window (RWS). Players must have made at least ten appearances and can only appear once in the table.

Player	Date	Team	League	RWS
Neymar	09/04/2019	PSG	France Ligue 1	8.821
Lionel Messi	19/12/2018	Barcelona	Spain LaLiga	8.698
Hakim Ziyech	12/02/2020	Ajax	Netherlands Eredivisie	8.429
Carlos Vela	20/08/2020	Los Angeles FC	USA Major League Soccer	8.394
Cristiano Ronaldo	02/09/2016	Real Madrid	Spain LaLiga	8.240
James Tavernier	07/12/2020	Rangers	Scotland Premiership	8.174
Kylian Mbappé	05/12/2020	PSG	France Ligue 1	8.149
Robert Lewandowski	23/05/2021	Bayern	Germany Bundesliga	8.098
Luuk de Jong	19/09/2019	Sevilla	Spain LaLiga	8.073
Zlatan Ibrahimovic	21/02/2017	Man Utd	England Premier League	8.026

3.3. Odds data

Finally, historical betting odds were used to test the predictive capabilities of our forecasting model and home-win, draw, and away-win odds were obtained for Bet365 (a bookmaker) from football-data.co.uk.

4. Player ratings: League-adjusted WhoScored ratings

Before describing the modelling framework, we present the player ratings system we use as the forecasting model's input.

The basis of our player-ratings model (which will be described in more detail in Section 5 and Section 6) are the matchday ratings published by WhoScored.com. WhoScored publishes performance ratings for every player within a match based on their in-match actions. Although the methodology for calculating the ratings is not fully in the public domain, the general, top-level concept is described on the WhoScored website.[1] To summarise, a player in a match starts with a rating score of 6. As the game progresses, a player receives points for actions deemed to impact team performance positively and is penalised for actions that have been judged to have a negative impact on the team. Players can earn a maximum match performance score of 10. There have been 1084 instances where a player has received a 10. The famed "MSN" trio, Lionel Messi, Luis Suaréz and Neymar, hold the records with 52, 20, and 22 perfect ratings, respectively. The unfortunate recipient of the lowest recorded rating is Oier Olazábal, who was the goalkeeper for Granada when they lost 9-1 to Real Madrid in 2015, resulting in a score of just 1.89. The resultant ratings are popular amongst fans and the media. Fig. 1 displays the histogram of ratings achieved by players. We separate players by whether they started in the match or came on as a substitute.

By taking an average of the WhoScored match performance ratings for a player, one obtains an idea of how good the player is and how they might be expected to perform in future matches. Let a player's raw WhoScored rating (RWS) be the average rating they achieved over two years. Table 1 shows the highest RWS ratings of individual players throughout the data.

Although some high-profile and widely accepted top players offer reassurance that the WhoScored ratings are meaningful, there are some unexpected names in Table 1 (namely: Carlos Vela, James Tavernier, and Luuk de Jong). This highlights three potential problems with the WhoScored ratings and taking a simple average of the match performance ratings. First, it appears that the methodology does not adjust the match performance ratings for the league's quality (the quality of the players within a league) such that performances in different leagues are not directly comparable. As such, an adjustment to the WhoScored player ratings is needed to account for the strength of the league and the players within that league.

Second, some players appear in a small number of games. Taking a simple average of their performance ratings to estimate how they can be expected to perform in the future is likely to result in volatile, unrealistically high or low average ratings. Third, it is likely that more recent performances by a player are more relevant to how that

[1] See https://www.whoscored.com/Explanations

Table 2
Top ten Adjusted WhoScored (AWS) ratings achieved by players. There is no minimum number of games required for a player to appear in the table.

Player	Date	Team	League	AWS
Lionel Messi	19/05/2021	Barcelona	Spain LaLiga	1.799
Neymar	20/02/2018	PSG	France Ligue 1	1.603
Cristiano Ronaldo	07/08/2015	Real Madrid	Spain LaLiga	1.397
Robert Lewandowski	23/05/2021	Bayern	Germany Bundesliga	1.225
Kylian Mbappé	01/11/2020	PSG	France Ligue 1	1.166
Kevin De Bruyne	20/01/2021	Man City	England Premier League	1.106
Eden Hazard	10/05/2019	Chelsea	England Premier League	1.100
Zlatan Ibrahimovic	20/08/2016	Man Utd	England Premier League	1.091
Hakim Ziyech	11/03/2019	Ajax	Netherlands Eredivisie	1.075
Harry Kane	21/01/2018	Tottenham	England Premier League	1.053

player is expected to play than performances further in the past.

These problems can be addressed in a regression model, which we use to generate '*adjusted WhoScored ratings*' (AWS). The dependent variable equals the 'raw' WhoScored match performance rating. The covariates include dummy variables for the player and league and a home indicator to allow for home advantage. To be explicit, suppose we observe $y_1, \ldots, y_N$ WhoScored ratings. For observation i, let $p(i)$ denote the player who achieved that rating, let $l(i)$ denote the league it was achieved in and let $h(i)$ indicate whether the player was competing at their home ground. Then our model can be written as

$$y_i = \alpha_0 + h(i)\alpha_1 + \beta_{p(i)} + \gamma_{l(i)} + e_i, \quad (1)$$

where $e_i \sim \mathrm{N}(0, \sigma^2)$ and β and γ are the estimated ratings of each player and league, respectively, whilst α_0 is the intercept, and α_1 represents a home advantage parameter.

To account for players with small numbers of games, we shrink the ratings towards the average rating by adding 'fake' games. In these fake games, we assume a player competed in a match within their current league. They receive a rating equal to the average within their current league, and we assume home advantage is equal to 0.5 (which means these games are effectively played at a neutral venue). By increasing the weight of these pseudo-observations, we can adjust the level of shrinkage. Letting ω denote the weight, when $\omega = 0$, the model has no shrinkage. As ω increases, the level of shrinkage increases. This itself is a hyper-parameter that must be tuned.

To account for changing player ability and form, we weight the observations to allow for match performance ratings further in the past to have a smaller effect on the coefficient estimate of the player dummies than more recent match performances. We apply an exponential weighting scheme to observations as was used by Dixon and Coles (1997) and others since. We include only games played within ψ years before the rating date, where the weight is $\exp(-\phi \cdot t/3.5)$ and t is the number of days the performance is from the calculation day.

To tune the hyper-parameters, ω, ψ and ϕ, we aimed to minimise the RMSE when predicting a player's future performance throughout the validation data. We found the values that minimised the RMSE were $\psi = 2.00$, $\phi = 0.0062$ and $\omega = 7.00$. The estimated value of ϕ is such that matches one year ago have a weight approximately one-half that of the most recent matches when estimating the player's adjusted WhoScored rating.

Table 2 shows the resulting top ten players according to the adjusted WhoScored ratings, which we denote by AWS. Due to the shrinkage, there is no need to set a threshold of ten games for the minimum number of matches a player must have played to appear in the table. The list of players is a who's who of the top footballers offering strong reassurance that the new adjusted ratings are meaningful. The surprise inclusions in Table 1 have now disappeared from the top 10. Vela and Tavernier were competing in the MLS and Scottish Premiership at the time of their top ratings. De Jong transferred from the Eredivisie on 01/07/2019, shortly before his maximum RWS rating. Consequently, the high RWS ratings achieved by these three players consisted of good performances in relatively easier leagues. The AWS ratings account for the weaker leagues; hence, the adjusted ratings are lower.

A potential problem with the AWS ratings presented in Table 2 is that forward players dominate it. Indeed, the majority of the top 50 players are forwards. Of course, it may be that the best players in the world are forwards; after all, they attract the highest wages and transfer fees. But it is also possible that the AWS ratings are biased towards forward players. One could use different rating systems in our modelling framework, but as demonstrated by the performance of the forecasting model (see later), the AWS ratings perform well.

As a sense check of the newly adjusted WhoScored ratings, we calculate the average rating of the eleven starting players for each team. Table 3 shows the results. As for the individual players, the identity of the teams making up this table raises confidence that the ratings are meaningful.

An interesting aside to the main topic here is the estimated league strengths. This is an important area of research in itself, as when clubs recruit players from leagues other than their own, it is essential to gauge whether a player will be able to play as well in the new league as they have done in the current league. Fig. 2 shows the estimated league adjustments to player match performance ratings (where the second tier of football in England, the English Championship, is the reference league). It is probably no surprise that the English Premier League is the most difficult league throughout the data. Each match performance rating is worth around 0.25 more than the same score in the English Championship. Another interesting finding is the rise of the

Fig. 2. Plot of the league strengths over time. Note that we plot the negative value of the actual estimate, given a more negative value implies the league is harder. The reference league is the English Championship.

Table 3
Highest average AWS and RWS ratings for individual clubs' starting 11 players.

Team	Date	AWS	RWS
Man City	29/01/2019	0.589	7.212
Barcelona	24/02/2018	0.562	7.410
Liverpool	29/02/2020	0.519	7.174
Tottenham	20/08/2017	0.517	7.144
Man Utd	31/01/2018	0.514	7.154
Arsenal	15/10/2016	0.511	7.228
Real Madrid	21/09/2016	0.477	7.380
Chelsea	03/01/2018	0.468	7.129
PSG	19/01/2019	0.468	7.295
Juventus	29/09/2018	0.459	7.276

French Ligue 1. During the 2016/17 and 2017/18 seasons, ratings in the Championship were worth more than in Ligue 1 (the estimated coefficient was negative). Still, as of the 2021/22 season, a rating in Ligue 1 is worth more than the Championship and the German Bundesliga.

5. Including player level ratings in a team results forecasting model

In this section, we present our general methodology for including player-level ratings/metrics in models for forecasting the results of matches. The specification we present has two benefits. First, it replicates the way players on opposing teams interact in matches. Second, the specification can be used in many different models as it produces a single covariate (or, in machine learning terminology, a single 'feature').

In any given match, depending on the teams' formations, an individual player will play most of the game competing with one opposing player, and competing against other opposing players less frequently.

For example, a team's left winger plays in an advanced position on the left. Left-wingers thus compete with the opposition's right-back more frequently than they will compete with, for example, the opposition's left-back (on the other side of the pitch). Indeed, in match strategy decisions, a right-back's primary responsibility is to stop the attacking threat of a left-winger.

A complicating matter is that the level of interaction between players on opposing teams depends on the formations the two teams are playing. A common formation for a team is to play with four defenders, four midfielders, and two strikers (all teams must play with a single goalkeeper). This formation is known as 4-4-2. Another common formation is 4-3-3 (four defenders, three midfielders and three attackers). The level of interaction between the left-winger on the team playing a 4-4-2 formation and the right-back on the opposing team depends on whether the opposing team is playing in a 4-4-2 or 4-3-3 formation.

In a model for match outcomes based on individual player ratings, the level of interaction between players in different positions and on teams with various formations must be correctly represented. To accomplish this, we estimate the proportion of events a player in a given position in a given formation (e.g., left striker in a 4-4-2 formation) interacts with each player in given positions on the opposing team, given their formation. We estimate these proportions using two multinomial models explained in the following two subsections.

5.1. Estimating players' interaction with opposing outfield players

The first multinomial model estimates the level of interaction between opposing *outfield* players using data on events in which two players are involved (and recorded in the data): tackles (won or lost), blocks, interceptions, and aerial duels. Suppose there are M events indexed by $j = 1 \dots M$. Let pos_j^a and pos_j^d denote the positions of the attacking and defending players who were involved in the jth action, respectively. Further, let $form_j^a$ and $form_j^d$ denote the formations the attacking and defending teams were using for the jth event. The dependent variable is the position of the defending player. The independent variables comprise the attacking player's position and the two teams' formations.

There are 25 unique outfield positions; thus, 24 logit models associated with the opponent's outfield position are estimated. Each model is estimated relative to a reference category; in our case, the central attacking midfield (CAM) position and separate coefficients are estimated for each logit model.

For instance, the logit model, which estimates the probability the opponent's right-back (RB) attempts a defensive action, is

$$\log\left(\frac{P(pos_j^d = \mathrm{k})}{P(pos_d = \mathrm{CAM})}\right) = int^{\mathrm{RB}} + \beta^{\mathrm{RB}}_{pos_j^d} + \beta^{\mathrm{RB}}_{form_j^a} + \beta^{\mathrm{RB}}_{form_j^d}, \quad (2)$$

where $k = 1 \ldots 25$ is an index for the 25 playing positions.

When generating predictions from this model, 25 non-zero probabilities are calculated. However, only ten outfield players are possible.[2] Consequently, we normalise these ten values to sum to one. The value $p_{i,j}$ ($j \neq \mathrm{GK}$) thus represents the probability that a defensive action against attacking player i will be attempted by defending position j, which measures the overall level of interaction between the two positions.

5.2. Estimating players' interaction with the opposing goalkeeper

Similarly, we use data on shot events to determine the level of interaction between outfield players and the opposing goalkeepers. This model is needed because goalkeepers do not tend to interact with opponent players in the duel-type events used as the basis for our first model above. This time, the dependent variable is the player's position who has taken a shot. The position of the defending player is always the goalkeeper, so only the two teams' formations are used as independent variables. As in the previous model, we normalise the ten values associated with players actually on the pitch.

The value $p_{i,\mathrm{GK}}$ represents the probability that a given shot will be attempted by position i. This measures the level of interaction between i and the opponent's goalkeeper.

5.3. Examples

Fig. 3 shows the results of these multinomial models for two example cases. The first plot shows how a left striker (LST), playing on a team in a 4-4-2 formation, interacts with each player on the opposition team, also playing a 4-4-2 formation. 22.8% of their interactions are with the right-centreback (RCB), whilst just 2.3% of their interactions are with the left-striker (LST) on the opposing team. We see a 21.1% chance that the left-striker will attempt a given shot, indicating their interaction with the goalkeeper.

The second plot shows that a left-winger (LW) in a 4-3-3 formation has 75.4% of his interactions with the opposing right-back (who is playing in a 4-4-2 formation) and 12.2% of his interactions against the opposition's right midfielder (RM). We note that the LST in a 4-4-2 will interact with the opponent's RST 0.022 of the time. For an LW in a 4-3-3, this increases to 0.031, suggesting a winger will, on average, play more defensively than a striker, which accords with intuition. Further, there is a 20.8% chance that the left-winger will attempt a shot.

We propose the following metric to measure the difference in two teams' strengths in a match.

$$\Delta = \sum_i \sum_j p_{ij}(\mathrm{AWS}_i - \mathrm{AWS}_j) \quad (3)$$

where AWS_i is the AWS rating of the ith player on the home team, and AWS_j is the AWS rating of the jth player on the away team. p_{ij} is the weight estimated from the multinomial models described above. The first summation provides a weighted difference between a player's rating and each of the opposition team's player's ratings. The second summation calculates the first sum for each of the players.

6. Forecasting models

6.1. Data

Having trained our multinomial positional models on the 13/14 and 14/15 data, we use the remaining data (15/16–20/21) for modelling results. This ensures the probabilities generated by the multinomial models are themselves out-of-sample.

As is common practice, we use the first 80% of the 15/16–20/21 data for training and the remaining 20% for testing. The order of the data is maintained so that no leakage occurs.

There are several parameters to tune: the hyperparameters of the player ratings model, time-weightings in the team-based models we use for comparison (see Section 7.1), and optimal thresholds for betting strategies (see Section 7.2). We split the training set again to optimise these parameters and ensure results are fully out-of-sample, keeping the last 20% as a validation set. Fig. 4 displays these splits graphically.

Consequently, we present fully out-of-sample results in all the experiments reported herein. During the Covid pandemic, football matches were played behind closed doors. We removed these matches from our analysis as the home advantage is known to have been distorted during these games where no fans were present (see, for example, McCarrick et al. (2021)). This leaves us with a sample of 6824 matches between 12th August 2016 and 23rd May 2021. Of this sample, we use the final 20% as testing data.

6.2. Skellam model

Throughout the literature on forecasting in football, there has been considerable focus on estimating the scoring rates of teams. The pioneering idea of Maher (1982) was allowing teams to have separate attack and defence abilities. The scoring rate of the home team is estimated using the home team's attack strength and the away team's defence strength, and vice-versa for the away

[2] We note that the predicted probabilities of players not on the pitch are extremely small.

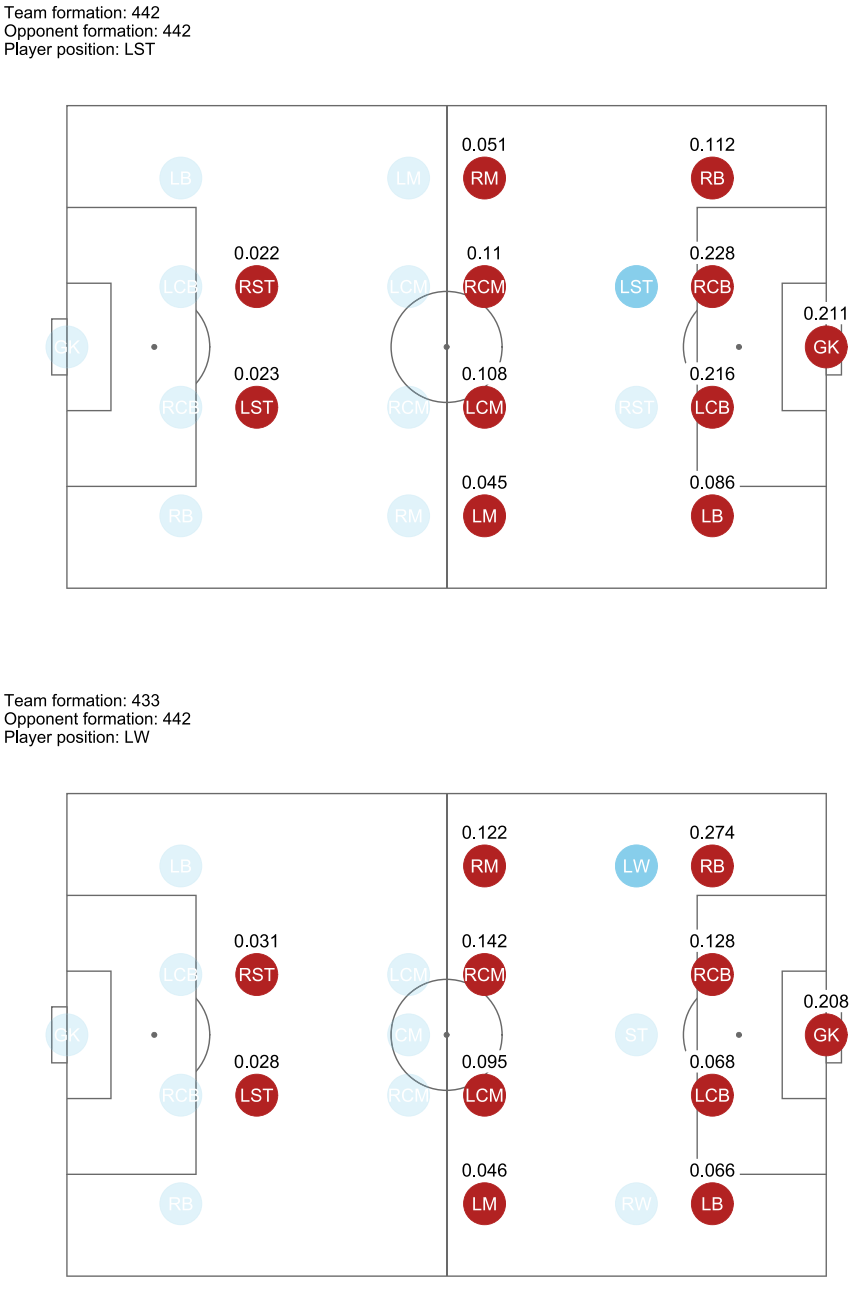

Fig. 3. Examples of the player weights for two different scenarios. The defending team is coloured in red.

Fig. 4. Plot detailing how data was split during the three main stages of this work. Black indicates the portion of data used for training models, whilst grey represents data used for testing.

Table 4
Model results for the Skellam regression model. Variables are standardised to have a mean of 0 and a variance of 1. The associated p-values are given in the right column.

Dependent variable:		
Goal-difference	Coefficient	p-value
β_{0h}	0.3670	0.0000
β_{1h}	0.2493	0.0000
β_{2h}	0.0583	0.0088
β_{0a}	0.0500	0.0686
β_{1a}	−0.3403	0.0000
β_{2a}	−0.0635	0.0269
Observations	5459	
Log-likelihood	−10374.17	
AIC	20760.34	

team's scoring rate. The original model used independent Poisson regressions, and since then, researchers have developed more complex frameworks, for instance, a bivariate-copula Weibull regression (Boshnakov et al., 2016).

Our variables, which represent the difference between teams' overall strengths, do not fit as naturally in this framework. Instead of modelling the goals scored by each team, we can estimate the goal difference using a Skellam regression. Of course, the probabilities of each match outcome are easily obtained by summing the relevant goal difference probabilities.

For game i, suppose the goal-difference is GD_i, Δ_i^{Start} is the weighted difference in the adjusted ratings of the players who start the match defined in Eq. (3), and Δ_i^{Sub} is the unweighted difference in the substitutes adjusted ratings.[3] Since the Skellam distribution represents the difference between two independent Poisson random variables, we estimate two coefficients for each independent variable. These coefficients naturally correspond to the scoring rates of the home and away teams, which we have included in the following notation for the model. Consequently, we estimate

$$\text{GD}_i \sim \text{Skellam}(\lambda_{ih}, \lambda_{ia}) \tag{4}$$

$$\log(\lambda_{ih}) = \beta_{0h} + \beta_{1h}\Delta_i^{\text{Start}} + \beta_{2h}\Delta_i^{\text{Sub}} \tag{5}$$

$$\log(\lambda_{ia}) = \beta_{0a} + \beta_{1a}\Delta_i^{\text{Start}} + \beta_{2a}\Delta_i^{\text{Sub}}. \tag{6}$$

The estimated coefficients of this model are displayed in Table 4. Note that the variables were scaled and centred when fitting the model. We see a significant home advantage, and the effect size aligns with past literature. The weighted sum of differences in player ratings is highly statistically significant. The difference in the ratings of the substitutes is also highly statistically significant, with a smaller estimated coefficient. This is to be expected as it will matter less to the outcome of the match than the strength of the starting players on each team. We also observe a significant home advantage.

[3] We use the unweighted difference in ratings for the substitutes as it isn't known before the game if the substitutes will be used, and for how long they will play. If they are used, it is unknown what position they will play (though they are likely to play in their specialised roles).

7. Out-of-sample testing

7.1. Scoring rules

To compare and benchmark models for forecasting football match results against other models, we advocate using scoring rules (see, for example, Johnstone et al. (2013)) and examining returns based on betting.

Since both the adjusted WhoScored ratings and positional weights are novel features of our model, we fit several variations to investigate whether both additions add predictive power.

Let $\Delta_{full}^{\text{Start}}$ be the sum of differences in the AWS ratings of opposing players, weighted by the level of interaction between them–that is, the variable defined in Eq. (3). Let $\Delta_{adj}^{\text{Start}}$ be the difference in the average AWS ratings of the opposing teams' starting players. Similarly, let $\Delta_{raw}^{\text{Start}}$ be the difference in the average RWS ratings of the opposing teams' starting players. Finally, let $\Delta_{adj}^{\text{Sub}}$ or $\Delta_{raw}^{\text{Sub}}$ be the difference in the average AWS or RWS ratings of the opposing teams' starting players, respectively. We consider four models defined as follows:

- Model skellam_{full} is the Skellam model displayed in Table 4. This uses $\Delta_{full}^{\text{Start}}$ and $\Delta_{adj}^{\text{Sub}}$ as covariates.
- Model skellam_{adj} removes the player interaction weights, thus using $\Delta_{adj}^{\text{Start}}$ and $\Delta_{adj}^{\text{Sub}}$ as covariates.
- Model skellam_{raw} removes the WhoScored rating adjustments, thus using $\Delta_{raw}^{\text{Start}}$ and $\Delta_{raw}^{\text{Sub}}$ as covariates.
- Model skellam_{team} is a team-based model, where we include dummy variables for each team which can take values 1 or −1 if they are the home or away team, respectively. A value of 0 indicates the team is not involved in a match. As with the player ratings, team strengths are updated using all results available before a match.

We can compare the results from these four models to assess whether the adjusted WhoScored ratings improve on the raw average ratings. We also include how the interaction between opposing players improves results and whether having player-based information is better than team-based information.

We include time-weighting in the team ratings-based models and test whether including shrinkage through 'fake' games improves the fit. We optimise these parameters through cross-validation on the validation data to minimise the average Brier score.

We report the accuracy and Brier score achieved by the models. The Brier score is the most commonly used scoring rule in forecasting literature. Whilst accuracy is not a proper scoring rule, it is the most intuitive to a reader and is interpretable in and of itself.

Table 5 shows the Brier scores for the different models and the bookmaker implied probabilities. We use the odds available from Bet365 through www.football-data.co.uk and remove the bookmaker's margin by scaling the implied probabilities to sum to 1.

In terms of accuracy (the proportion of matches in which the outcome with the highest implied probability occurs), the bookmaker performs the best. However, we

Table 5
Scoring rules for several models used to predict the results of football matches. Also shown are the corresponding results derived from the bookmaker's (Bet365) implied probabilities.

Model	n	Accuracy	Brier
Bet365	1350	52.74%	0.5877
skellam$_{full}$	1350	52.00%	0.5955
skellam$_{adj}$	1350	51.85%	0.5957
skellam$_{team}$	1350	52.89%	0.5962
skellam$_{raw}$	1350	51.33%	0.6029

note the tiny differences in these figures. A baseline model of predicting a home win in every match, regardless of the teams, results in an accuracy of 45.11%. The marginal gain from going from the simplest model of all available to the best-performing model (the bookmakers) is perplexingly small. Even the 'best' performing machine learning model from the Soccer Prediction Challenge (Dubitzky et al., 2019) achieved an accuracy only slightly higher at 53.88% (though, of course, it was achieved on a different dataset so direct comparisons are not possible).

The bookmaker also achieves the best Brier score. Of our models, skellam$_{full}$ performs best; the small gap behind the bookmakers provides encouragement. We see that skellam$_{full}$ improves on skellam$_{adj}$, assuring that the opposing player interaction weightings improve performance (albeit only to a small degree). There is a notable improvement when moving from RWS to AWS ratings, justifying the use of our player-rating framework. The most interesting results are that all models utilising our AWS player ratings beat the team-based models.

We note that these results are robust to other scoring rules, for instance: the rank-probability score used throughout the Soccer Prediction Challenge (Dubitzky et al., 2019); and the ignorance score used in Wheatcroft (2021).

7.2. Betting

In the previous section, we showed that our model achieved results similar to the bookmakers according to the Brier Score and the accuracy. However, unlike the bookmakers, a bettor does not have to 'invest' in every match. We now use the models to bet with and investigate what returns on investment are obtained on the 1X2 (home win, draw, away win) market. In betting against the bookmakers, despite the bookmakers having an advantage built into their odds (the margin or vig), the bettor has an advantage in that they can choose not to bet.

Our investment strategy is based on the Kelly Criterion (Kelly, 1956) and is the same as the one used in Boshnakov et al. (2016). The Kelly criterion is borne from a desire to maximise long-run log-utility, and it results in an investment strategy where the bettor invests a fraction f of his overall wealth

$$f = \frac{(b+1)p - 1}{b},$$

where p is the bettor's estimate of the probability of an event (e.g. the home team winning the game), and b is the (fractional) odds offered by the bookmaker (where $1/(b+1)$ can be interpreted loosely as the bookmaker's implied probability of the event occurring).

Whereas the usual Kelly strategy follows a rolling bankroll updated after each bet, we stake the equivalent of one "unit" multiplied by f for each bet. Effectively we reset our bankroll to one after each bet.

An additional 'protection' was also introduced: we restrict ourselves to 'quality bets' when the expected value of any bet is above a threshold. For each game, there are three possible events to bet on: home win, draw, and away win.[4] For event type A, we only bet if

$$EV(A) = P(A) \times Odds(A) - 1 > t,$$

where t is a threshold parameter and effectively protects the investment strategy when the bookmaker knows more than the model. As mentioned in Section 6.1, we split our data into training, validation, and testing sets. To determine the optimal threshold, we fit an initial model to the training data (without the validation data). The optimal betting threshold is then determined by finding the maximal return on investment (ROI) on the validation data (which, at this point, is out-of-sample to the training data). Finally, the full model is fit using the training and validation data, and the out-of-sample betting results are calculated using the pre-determined threshold on the test data.

In addition to looking at the returns to investment, we believe it is important to consider the Sharpe ratio as we see in finance. The Sharpe ratio is a measure for calculating risk-adjusted returns and is defined as the rate of return per unit of volatility. Just as in finance, we calculate the Sharpe ratio as the ROI over all bets divided by the standard deviation of the ROI of each individual bet. The result is then annualised by multiplying by $\sqrt{n}$ where n is the total number of bets. A general rule of thumb is that a Sharpe ratio of 1 or higher is considered good (and the higher the better, as the investment achieves higher returns at lower risk).

We compare the results of the modified Kelly strategy with those of using simpler flat staking strategies. In these schemes, we place one unit on the most likely outcome according to the model. As with the Kelly strategies, we find the optimal expected value threshold and report the corresponding results.

It should be noted that the bet set may differ when using flat or Kelly staking strategies. When applying flat stakes, the user bets on the outcome they believe is the most likely. However, in Kelly scenarios, the user places bets based on the expected value of the bet. Consequently, this choice may not be the most probable outcome according to the user.

The results of betting with skellam$_{full}$ are given in Table 6. Given the literature on forecasting in football, these returns are very promising, especially given the high

[4] To be explicit, this means we may bet on a maximum of three outcomes (in the unlikely situation that all have positive expected value). This also means we may bet on outcomes which are not necessarily the most likely result predicted by the model.

Table 6
Results for several betting strategies using the skellam$_{full}$ model.

Strategy	t	N	Accuracy (%)	Stakes	Profits	ROI (%)	Sharpe
Kelly	0.1866	556	24.10	65.36	7.81	11.96	1.07
Kelly	0.0000	1457	29.44	105.42	6.03	5.72	1.02
Flat	0.1760	199	37.19	199.00	9.04	4.54	0.45
Flat	0.0000	568	45.25	568.00	16.93	2.98	0.59
Flat		1350	52.00	1350.00	−32.56	−2.41	−0.87

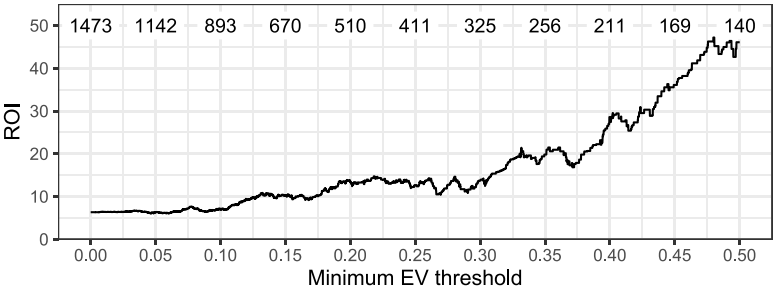

Fig. 5. Plot displaying the ROI that would be achieved using the skellam$_{full}$ model for betting under the modified Kelly strategy for different minimum expected value thresholds.

number of bets being placed. For example, Koopman and Lit (2015) placed just 50 bets over two seasons.

We find that only the most basic strategy- flat staking with no value threshold- results in losses. Both Kelly strategies performed well, achieving very promising ROIs and Sharpe ratios. We highlight that these results have been obtained on a large number of bets. Whilst flat stakes with $t = 0$ and $t = 0.1866$ both achieve positive returns, the Sharpe is less than 1 indicating more risk than reward.

Fig. 5 shows the relationship between the minimum expected value threshold and the ROI achieved by the skellam$_{full}$ model under the modified Kelly staking strategy. Also shown is the number of bets placed along the chart's top. The number of bets decreases as the threshold increases, but the ROI increases to very high levels.

8. Closing remarks

In this paper, we have presented a new model for forecasting the results of football matches. The model is a 'player-based' model as opposed to the previously published 'team-based' models of Maher (1982) and Dixon and Coles (1997). We developed a novel rating framework which adjusts publicly available player matchday ratings to ensure comparability across leagues. Further, we introduced multinomial models to account for the level of interaction between two opposing players, knowing that different formations dictate how often a player will compete against a particular opponent.

Player-based models rely heavily on data but solve the major issue with team-based models. There is no need to worry about time-varying team strengths: the mechanism which causes the dynamics is modelled directly, that is, the changing line-ups of the teams and the changing short-term form of the players. Admittedly, the model is data-hungry, but databases of player ratings now exist, and access to these should become increasingly easy in the future.

We have demonstrated the goodness-of-fit of the model. Scoring rules suggest the model performs very well compared to bookmakers. Even when we perform the sternest test of all forecasting models, examining the returns to betting, the results are positive to the extent that we achieve positive returns to betting on the 1X2 market.

Our results have implications in economics studies of market efficiency and the practice of trading in football. For example, the player-based model may reduce, at least to some extent, the reliance of bookmakers on expert traders to adjust predictions from a team-based model in light of information about the actual line-up of players, say when a star player is injured. Currently, traders are typically required to adjust model probabilities subjectively. Our player-based model does this automatically.

Future work on this type of model is promising. One could, for example, model the interactions of players on the same team. Football fans often believe some players play well together and are greater together than the sum of their abilities. A model including some interaction between players on the same team would be able to identify whether this was true. Another area for potential improvement of the model is to use 'better' player ratings. Here we use WhoScored ratings, but these ratings may be weak. For example, there may be a bias towards forward players in the WhoScored ratings (given that the top ten are exclusively forwards). One could even use this model to rate the player ratings themselves. For example, one rating of players is their pass completion percentage. This could be used as the metric feeding the forecasting model (instead of the WhoScored rating), and the model's performance is used to measure the usefulness of players' pass completion percentage as a predictor of future team performance. Many player-level metrics could be tested, compared, and rated in this framework for their usefulness.

Lastly, we note the model could be used to develop recruitment tools for football clubs and predict the potential impact a new player might have on a club's results.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Arntzen, H., & Hvattum, L. M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, *21*(5), 449–470.

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., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, *108*(1), 97–126.

Boshnakov, G., Kharrat, T., & McHale, I. (2016). A bivariate weibull count model for association football scores. *Journal of International Forecasting*.

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

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of english football league matches for betting. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *51*(2), 157–168.

da Costa, I. B., Marinho, L. B., & Pires, C. E. S. (2021). Forecasting football results and exploiting betting markets: The case of "both teams to score". *International Journal of Forecasting*.

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.

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The Open International Soccer Database for machine learning. *Machine Learning*, *108*(1), 9–28.

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

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

Johnstone, D. J., Jones, S., Jose, V. R. R., & Peat, M. (2013). Measures of the economic value of probabilities of bankruptcy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(3), 635–653.

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, *35*(4), 917–926.

Kharrat, T. (2016). *A journey across football modelling with application to algorithmic trading* (Ph.D. thesis).

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 167–186.

Lasek, J. (2019). *New data-driven rating systems for association football* (Ph.D. thesis).

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

McCarrick, D., Bilalic, M., Neave, N., & Wolfson, S. (2021). Home advantage during the covid-19 pandemic: Analyses of european football leagues. *Psychology of Sport and Exercise*, *56*, Article 102013.

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

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, *34*(1), 17–29.

R. Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G., & Király, F. (2019). Modeling outcomes of soccer matches. *Machine Learning*, *108*(1), 77–95.

Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, *36*(3), 916–932.

Wheatcroft, E. (2021). Evaluating probabilistic forecasts of football matches: the case against the ranked probability score. *Journal of Quantitative Analysis in Sports*, *17*(4), 273–287.