Euro 2016 Predictions Using Team Rating Systems

Jan Lasek

jan.lasek@deepsense.io

Abstract. In this study we employ several rating systems to generate predictions for the outcome of 2016 European Championships in association football. To this end, we first estimate probabilities of match results between all competing nations using the rating systems. Secondly, via Monte Carlo simulations we compute probabilities of advancing past a given stage of the tournament. The approach was developed for the Euro 2016 Prediction Competition organized within Sport Analytics Workshop at ECML/PKDD 2016.

Keywords: team ratings, rating systems, predictions, Euro 2016

1 Introduction

In this paper we demonstrate how to utilize various football team rating systems in order to make predictions for outcome of the Euro 2016 championships in association football. Our approach consists of two steps: firstly, we employ several team rating systems to estimate team strength parameters. Using these models we calculate probabilities of outcome of each possible match-up result in the tournament. The predictions are next used to simulate tournament outcome in a Monte Carlo experiment. The paper is an extended version of our blog post on using team rating systems for generating predictions for the tournament [3].

We proceed with the description of rating systems that we are going to use.

2 Rating systems for football teams

There have been multiple rating systems for various sport developed throughout the years. We discuss here three different models for rating teams in association football: the ordinal logistic regression model, the least squares model and the Poisson model.

Ordinal regression ratings. The first rating system discussed is ordered logistic regression as the match results model [1, 7]. Under this model, each team is associated with a single parameter – a rating – reflecting its strength. Teams' strength parameters are estimated based on the outcomes of games between the teams. Let r_i, r_j be ratings of two teams i and j and with team i playing

at home ground. According to the model, if H and A denote a home and away team win, respectively, and D corresponds to a draw, the probabilities of these events are linked with teams' ratings parameters with the following equations

$$\begin{aligned}\mathbb{P}(H) &= \frac{1}{1+e^{c-(r_j-r_j+h)}},\\ \mathbb{P}(D) &= \frac{1}{1+e^{-c-(r_j-r_j+h)}} - \frac{1}{1+e^{c-(r_j-r_j+h)}},\\ \mathbb{P}(A) &= 1 - \frac{1}{1+e^{-c-(r_j-r_j+h)}},\end{aligned}$$

where $c > 0$ is an intercept and h is a parameter introduced to account for the home team advantage [9]. To estimate the model's parameter weighted maximum likelihood method can be used. The estimation proceeds as follows. Let $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ denote the vector of teams' ratings. Let us denote by $L(M|\mathbf{r}, h, c)$ the weighted log-likelihood function of the results observed in dataset of matches M given model parameters $\mathbf{r}, h, c$. Each match $m \in M$ is described as a tuple $m = (i, j, k, t)$ where i, j are indexes encoding particular teams, k is the type of a match (friendly, qualifier to a major tournament or a major tournament match) and t is time elapsed from the *estimation period*. The estimation period is understood as the time at which we want to estimate ratings for. The log-likelihood function is weighted by both the time the game took place and the importance of a match. It is defined as

$$L(M|\mathbf{r}, h, c) = \frac{1}{|M|} \sum_{m \in M} \phi(m) \cdot \log \mathbb{P}(R_m),$$

where $R_m \in \{H, D, A\}$ is the actual result of match m. We assume that the weighting function has the form of $\phi(m) = \alpha(k) \cdot e^{-\beta t}$, where $\alpha(\cdot)$ is a function that maps match type to a numerical value representing its importance and β is a time decay parameter. The idea here is to give a higher weight for recent results as well as different weights according to match type. To estimate team ratings we minimise

$$-L(M|\mathbf{r}, h, c) + \lambda \cdot \left(\frac{1}{2}(1-\gamma)\|\mathbf{r}\|_2^2 + \gamma\|\mathbf{r}\|_1 \right),$$

where $||\cdot||_1$ and $||\cdot||_2$ are L_1 and L_2 norms, respectively, and $\gamma \in [0, 1]$ and λ are parameters for Elastic Net regularisation component [12].

Least squares ratings. The next rating system is based on a simple observation that the difference $s_i - s_j$ in the scores produced by the teams should correspond to the difference in ratings

$$s_i - s_j = r_i - r_j + h.$$

Again, h is a correction for the home team i advantage. The rating system's name originates from its estimation method: one finds ratings r_i such that the sum

of squared differences (over a set of games) between the two sides of the above equation is minimal [11]. The sum of squares function can be weighted in an analogous manner as discussed in case of the ordinal logistic regression model.

For the least squares model, we still need to generate probabilities for particular outcomes. This is done first by computing a logistic regression model with binary outcomes as described in [6]. Next, the binary outcomes are mapped to a three-way-outcome by a method proposed in [10].

Poisson model. The final rating system that we discuss is based on the assumption that the goals scored by a team can be modelled as a Poisson distributed variable. The mean rate of this variable is dependent on the attacking capabilities of a team and the defensive skills of its opponent. This extends ratings to two parameters – offensive and defensive skills per team as opposed to a single parameter in the methods discussed above.

Given the attacking and defensive skills of teams i and j, a_i, a_j and d_i, d_j, respectively, the rates of Poisson variables for a home team i and visiting team j, λ and μ respectively, are modelled as:

$$\lambda = c + h + a_i - d_j,$$

$$\mu = c + a_j - d_i,$$

where c is an intercept and h accounts for home team advantage. Under this model, the probability of a score x to y is a product of two individual Poisson variables with rates λ and μ respectively and equal to $\frac{\lambda^x \cdot e^{-\lambda}}{x!} \cdot \frac{\mu^y \cdot e^{-\mu}}{y!}$. Given a dataset of matches, one can estimate the team rating parameters using the maximum likelihood method. The likelihood can be weighted in a similar manner as discussed in case of ordinal logistic regression model. Here, for simplicity we employ the basic version of the model that assumes that the Poisson variables corresponding to the goals scored by the teams, given their rating parameters, are independent [8]. There are studies which relax this assumption [8, 4].

3 Tuning the predictive performance

We used the rating systems presented here to estimate win, draw and loss probabilities for every pair of possible match-ups among the 24 teams participating in Euro 2016. Given these probabilities, we simulated the tournament multiple times and computed each team's probability of winning it all. We used the database of international football match results provided at `http://laenderspiel.cmuck.de/`.[1]

First of all, the rating systems involve some adjustable parameters e.g., weights for importance of matches, a weighing function for most recent results and regularization parameters. We tuned these parameters (by exploring a grid

[1] Thanks to the website's maintainer Christian Muck for generously exporting the data.

of values) to maximize the predictive accuracy of the models: using a sample of games, we predicted their results and evaluated them. For tuning the parameters, we chose matches from major international tournaments – World Cup finals, European Championships and Copa America.

The parameters of the ratings systems are chosen for World Cup finals held between 1994 and 2010 (5 tournaments), UEFA European Championships 1996-2008 (4) and Copa America finals 1999-2011 (5). This accounts for a set of 562 matches. In the competition, the prediction accuracy is evaluated using logarithmic loss (logloss). Accordingly, we use this metric to tune the models' parameters. This error metric is calculated as $\frac{1}{m}\sum_{i=1}^{m}\log\mathbb{P}(R_m)$, where $\mathbb{P}(R_m)$ is the probability of the final outcome of i-th game in data attributed by the model, $i = 1, 2, \ldots, m$. A more direct interpretation could be provided by accuracy that is defined as the percentage of matches that were correctly predicted by a given method. To estimate the final efficacy of the methods we present results on the validation sample comprising of 2014 World Cup finals, 2012 UEFA European Championships and 2015 Copa America. To provide some context for the numbers, we present a benchmark solution of random guessing and probabilities derived from an average of bookmakers' odds. A random guess yields a logloss of $-\log(1/3) \approx 1.1$ and accuracy of 33% for a three-way outcome. We also show scores achieved by two benchmark solutions based on the Elo model: EloRatings.net and FIFA Women World Rankings methodology [2, 5].

Table 1. Evaluation of the final test set (112 matches).

Method	**Logloss**	**Accuracy**
Bookmakers	0.9726	52%
Ensemble	0.9950	56%
Least squares	0.9985	55%
Poisson	0.9991	55%
Ordinal regression	1.0002	52%
FIFA Women World Rankings	1.0060	50%
EloRatings.net	1.0189	51%
Random guess	1.0986	33%

The results achieved by bookmakers (in terms of logloss) are better than all the individual rating methods. Of course, the bookmakers can include some additional information on player injuries, suspensions or a teams form during the contest – this provides them with an advantage over the models. Including such external information would be the next step to enhancing the accuracy of the presented models. In any case, the accuracy of predictions is slightly better in case of the rating systems. The bottom row of the table presents results for an ensemble method – which is the average of predictions for the three best performing methods: least squares, Poisson and ordinal regression ratings. It is a simple method for increasing the predictive power of individual models. We observe that this method slightly improves logloss while maintaining accuracy.

4 Simulations of tournament outcome

Given match outcome probabilities for each possible match-up, we simulated 1,000,000 tournaments (that many repetitions appear to provide stable results). We sampled only win, draw and loss results. If - after considering head-to-head results - the teams are still tied in the group stage, we resolved such ties randomly. According to the tournament's official rules, we should use goal differences, however, this information is not available in our simulation.[2] If there is a draw in the play-offs, we sample the result again.

Table 2 presents the predictions generated using the ensemble of the three introduced ratings systems. The consecutive columns indicate the probability of advancing to a given stage of the competition. For example, the number next to Portugal in the first column indicates that there is a 91.37% chance that it will advance past the group stage. The last column indicates a team's chance of winning the whole tournament.

Table 2. Predictions of elimination stage of Euro 2016 tournament.

Team	**Group stage**	**Quarterfinal**	**Semifinal**	**Final**	**Champions**
France	98.01%	82.6%	67.71%	51.21%	37.55%
Spain	92.60%	72.24%	51.11%	33.95%	19.08%
Germany	94.71%	70.41%	45.99%	24.88%	13.21%
England	93.52%	67.5%	40.87%	22.25%	10.40%
Belgium	84.38%	48.2%	26.10%	11.51%	4.55%
Portugal	91.37%	54.70%	26.31%	12.09%	4.42%
Italy	72.43%	33.38%	14.83%	5.26%	1.55%
Ukraine	76.81%	37.05%	15.5%	5.53%	1.52%
Croatia	66.00%	31.92%	14.65%	5.27%	1.50%
Russia	75.34%	37.84%	13.07%	4.29%	1.14%
Turkey	61.90%	27.97%	12.07%	4.00%	1.05%
Switzerland	69.98%	30.49%	11.80%	3.97%	0.88%
Poland	67.40%	26.58%	9.35%	2.77%	0.60%
Sweden	57.89%	20.76%	7.45%	2.11%	0.47%
Romania	62.64%	23.82%	8.07%	2.35%	0.45%
Austria	71.63%	27.01%	7.46%	2.07%	0.43%
Slovakia	63.66%	25.57%	6.96%	1.79%	0.37%
Republic of Ireland	54.68%	18.64%	6.38%	1.72%	0.35%
Czech Republic	46.28%	16.19%	5.60%	1.44%	0.29%
Hungary	56.86%	16.08%	3.37%	0.69%	0.11%
Iceland	47.81%	11.32%	2.02%	0.36%	0.05%
Albania	31.46%	6.62%	1.26%	0.19%	0.02%
Wales	34.29%	7.98%	1.19%	0.16%	0.02%
Northern Ireland	28.32%	5.11%	0.88%	0.13%	0.01%

[2] Notably, coin-tosses were used to resolve ties (if the game was tied after extra-time) before the penalty shoot-out was "invented." For instance, on its way to winning Euro 1968, Italy "won" its semifinal with the USSR through a coin toss.

5 Discussion

We see that France tops the ranking for the championship race in terms of associated probability. The 12th man is behind them – they are playing at home and the methods we used give them some edge due to this fact. On the other hand, the prediction for four-time World Cup winners Italy is somewhat discouraging. In recent years, Italy has seen disappointing results, including draws with Armenia, Haiti and Luxembourg (not to mention their 2010 and 2014 World Cup records). However, what the rating system could not infer is the fact that the Italian team usually rises to the occasion when faced with a major challenge – which usually happens at the big tournaments.

The rating methods presented here have some limitations. There are many factors influencing match results and we only covered simple predictive models based on historical data. Naturally, one could use some external and more sophisticated information e.g., players and their skills, and include it in a model. This could greatly improve the models' accuracy.

References

1. Aitchison, J. and Silvey, S.D.: The Generalization of Probit Analysis to the Case of Multiple Responses. Biometrika Vol. 44, No. 1–2, pp. 131–140 (1957)
2. EloRatings.net, `http://eloratings.net/`, on-line resources, last access date 6 July 2016
3. Euro 2016 Predictions Using Team Rating Systems, `http://deepsense.io/euro-2016-predictions/`, on-line resources, last access date 6 July 2016
4. Dixon, M.J. and Coles, S.G.: Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 46, No. 2, pp. 265–280 (1997)
5. FIFA Women's World Ranking Methodology, `http://www.fifa.com/fifa-world-ranking/procedure/women.html`, on-line resources, last access date 4 July 2016
6. Glickman, M.: Parameter Estimation in Large Dynamic Paired Comparison Experiments, Applied Statistics, Vol. 48, No. 3, pp. 377–394 (1999)
7. Koning R.H.: Balance in Competition in Dutch Soccer. Journal of the Royal Statistical Society: Series C (The Statistician), Vol. 49, No. 3, pp. 419–431 (2000)
8. Maher, M.J.: Modelling Association Football Scores. Statistica Neerlandica, Vol. 36, No. 3, pp. 109–118 (1982)
9. Pollard, R.: Home Advantage in Football: A Current Review of an Unsolved Puzzle, The Open Sports Sciences Journal, Vol. 1, No. 1, pp. 12–14 (2008)
10. Schrader, A.: Developing and Elo Rating for Major League Soccer and Predicting End of Season Finish, `https://drive.google.com/file/d/0Bxr6KEe4KY_OYnJuLUw1WF9GcGs/view?pli=1`, on-line resources, last access date 6 July 2016
11. Stefani, R.: Football and Basketball Predictions Using Least Squares, IEEE Transactions on Systems, Man and Cybernetics, Vol. 7, No. 2, pp. 117–121 (1977)
12. Zou, H. and Hastie, T.: Reguralization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B (Statistical Methodology), Vol. 67, pp. 301–320 (2005)