DE GRUYTER

Requires Authentication | Published by **De Gruyter** | July 10, 2019

A hybrid random forest to predict soccer matches in international tournaments

Andreas Groll, Cristophe Ley, Gunther Schauberger and Hans Van Eetvelde

From the journal Journal of Quantitative Analysis in Sports
https://doi.org/10.1515/jqas-2018-0060

Citations 30

Abstract

In this work, we propose a new hybrid modeling approach for the scores of international soccer matches which combines *random forests* with *Poisson ranking methods*. While the random forest is based on the competing teams' covariate information, the latter method estimates ability parameters on historical match data that adequately reflect the current strength of the teams. We compare the new *hybrid random forest* model to its separate building blocks as well as to conventional Poisson regression models with regard to their predictive performance on all matches from the four FIFA World Cups 2002–2014. It turns out that by combining the random forest with the team ability parameters from the ranking methods as an additional covariate the predictive power can be improved substantially. Finally, the hybrid random forest is used (in advance of the tournament) to predict the FIFA World Cup 2018. To complete our analysis on the previous World Cup data, the corresponding 64 matches serve as an independent validation data set and we are able to confirm the compelling predictive potential of the hybrid random forest which clearly outperforms all other methods including the betting odds.

Keywords: FIFA World Cup 2018; random forests; soccer; sports tournaments; team abilities

Appendix

A Some notations and definitions

Kronecker's delta, which is used in Section 4 in the formula of the multinomial likelihood and the RPS, is defined as follows:

$$\delta_{ij} = \begin{cases} 1, & \text{if} \quad i = j, \\ 0, & \text{otherwise.} \end{cases}$$

The Skellam distribution, which is also used in Section 4, is the discrete probability distribution of the integer random variable that is defined as the difference $K := Y_1 - Y_2$ of two independent Poisson distributed random variables Y_1, Y_2 with respective event rates λ_1, λ_2. The corresponding probability mass function is given by

$$P(K=k)=e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{k/2} I_k\left(2\sqrt{\lambda_1\lambda_2}\right), k\in\mathbb{Z},$$

where $I_k(\cdot)$ is the modified Bessel function of the first kind (for more details, see Skellam 1946). Now let Y_1 and Y_2 denote the (conditionally independent) Poisson-distributed numbers of goals of two soccer teams competing in a match. Then, the three probabilities $P(Y_1 > Y_2)$, $P(Y_1 = Y_2)$ and $P(Y_1 < Y_2)$ can be easily obtained by computing $P(K > 0)$, $P(K = 0)$ and $P(K < 0)$ via the Skellam distribution.

B Lasso regression for soccer data

An alternative, more traditional approach which is often applied for modeling soccer results is based on regression. In the most popular case the scores of the competing teams are treated as (conditionally) independent variables following a Poisson distribution (conditioned on certain covariates), as introduced in the seminal works of Maher (1982) and Dixon and Coles (1997). Similar to the random forests, the methods described here can also be directly applied to data in the format of Table 2 from Section 2.1. Hence, each score is treated as a single observation and one obtains two observations per match. Accordingly, for n teams the respective model has the form

$$\begin{aligned} Y_{ijk}|\boldsymbol{x}_{ik},\boldsymbol{x}_{jk} &\sim Po\left(\lambda_{ijk}\right), \\ \log\left(\lambda_{ijk}\right) &= \beta_0+\left(\boldsymbol{x}_{ik}-\boldsymbol{x}_{jk}\right)^\top\boldsymbol{\beta}+\boldsymbol{z}_{ik}^\top\boldsymbol{\gamma}+\boldsymbol{z}_{jk}^\top\boldsymbol{\delta}, \end{aligned} \qquad (2)$$

where Y_{ijk} denotes the score of team i against team j in tournament k with $i,j\in\{1,\ldots,n\}, i\neq j$. The metric characteristics of both competing teams are captured in the p-dimensional vectors $\boldsymbol{x}_{ik},\boldsymbol{x}_{jk}$, while $\boldsymbol{z}_{ik}$ and $\boldsymbol{z}_{jk}$ capture dummy variables for the categorical covariates *Host*, *Continent*, *Confed* and *Nation.Coach* (built, for example, by reference encoding), separately for the considered teams and their respective opponents. Furthermore, β is a parameter vector which captures the linear effects of all metric covariate differences and γ and δ collect the effects of the dummy variables corresponding to the teams and their opponents, respectively. For notational convenience, we collect all covariate effects in the $\tilde{p}$-dimensional real-valued vector $\boldsymbol{\theta}^\top=\left(\boldsymbol{\beta}^\top,\boldsymbol{\gamma}^\top,\boldsymbol{\delta}^\top\right)$.

Due to a rather large number of potential covariates in our data, we use regularization techniques when estimating the models to allow for variable selection and to avoid overfitting. In the following, we will introduce such a basic regularization approach, namely the conventional Lasso (Tibshirani 1996). For estimation, instead of the regular likelihood $l\left(\beta_0,\boldsymbol{\theta}\right)$ the penalized likelihood

$$l_p\left(\beta_0,\boldsymbol{\theta}\right)=l\left(\beta_0,\boldsymbol{\theta}\right)-\lambda P\left(\boldsymbol{\theta}\right) \qquad (3)$$

is maximized, where $P\left(\boldsymbol{\theta}\right)=\sum_{v=1}^{\tilde{p}}\left|\theta_v\right|$ denotes the ordinary Lasso penalty with tuning parameter λ. The optimal value for the tuning parameter λ will be determined by (standard) 10-fold cross-validation (CV) simply as the parameter that minimizes the CV error. The model will be fitted using the function `cv.glmnet` from the R-package `glmnet` (Friedman, Hastie, and Tibshirani 2010). In contrast to the similar ridge penalty (Hoerl and Kennard 1970), which penalizes squared parameters instead of absolute values, Lasso does not only shrink parameters towards zero, but is able to set them to exactly zero. Therefore, depending on the chosen value of the tuning parameter, Lasso also enforces variable selection.

Possible extensions

While the Lasso method described above was chosen as the reference method to compare the predictive power of the hybrid model, in the literature also several alternatives and extensions are discussed. In the following, we shortly sketch some possible modifications. As a first possible extension of the model (2), the linear predictor can be augmented by team-specific attack and defense effects for all competing teams. This extension was used in Groll et al. (2015) to predict the FIFA World Cup 2014. There, each couple of attack and defense parameters corresponding to a team has been treated as a group and, hence, the Group Lasso penalty proposed by Yuan and Lin (2006) has been applied on those parameter groups.

Alternatively, if the model (2) shall be extended from linear to smooth covariate effects $f(\cdot)$ for metric covariates, boosting techniques designed for generalized additive models could be used, such as the `gamboost` algorithm from the `mboost` package (Hothorn et al. 2017). Instead of the Poisson distribution the negative binomial distribution could be used as the response distribution when considering distributions for count data, which is less restrictive as it overcomes the rather strict assumption of the expectation equating the variance. Schauberger and Groll (2018) investigated two different boosting approaches for this model class. However, no overdispersion compared to the Poisson assumption was detected and the models reduced back to the Poisson case.

Altogether, in Schauberger and Groll (2018) the simple Lasso from (3) with predictor structure (2) turned out to be the best-performing regression approach, though slightly outperformed by the random forests from Section 3.1.

C Comparison of FIFA ranking, Elo rating and estimated abilities

Table 8 compares the ranking of the 32 participating teams in the FIFA World Cup 2018 according to estimated abilities (left column), Elo rating (center column) and FIFA ranking (right column). The ranking according to the estimated abilities and the Elo ratings are very similar (Spearman correlation of 0.94), while both have a smaller correlation with the FIFA ranking (Spearman correlation of 0.86 and 0.90, respectively).

All three methods rank Germany and Brazil as the two top teams. Notable differences between the rankings can be seen, for example, for Spain and Belgium. Both the estimated abilities and the Elo rating rank Spain third while it is ranked ninth by FIFA. Belgium is ranked rather inhomogenously in positions 6, 8 and 3 by the different methods. More details on the comparison of estimated team abilities and the FIFA rank can be found in Ley et al. (2019).

Table 8:

Ranking of the participants of the FIFA World Cup 2018 according to estimated abilities (left), Elo rating (center) and FIFA ranking (right).

Table 9:

Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2018 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup together with winning probabilities based on the ODDSET odds.

D Probabilities for FIFA World Cup 2018 Winner

In this section, the hybrid random forest is applied to (new) data for the World Cup 2018 in Russia (in advance of the tournament) to predict winning probabilities for all teams and to predict the tournament course.

The abilities were estimated by the bivariate Poisson model with a half period of 3 years. All matches of the 228 national teams played since 2010-06-13 up to 2018-06-06 are used for the estimation, what results in a total of more than 7000 matches. All further predictor variables are taken as the latest values shortly before the World Cup (and using the finally announced squads of 23 players for all nations).

For each match in the World Cup 2018, the hybrid random forest can be used to predict an expected number of goals for both teams. Given the expected number of goals, a real result is drawn by assuming two (conditionally) independent Poisson distributions for both scores. Based on these results, all 48 matches from the group stage can be simulated and final group standings can be calculated. Due to the fact that real results are simulated, we can precisely follow the official FIFA rules when determining the final group standings[5]. This enables us to determine the matches in the round-of-sixteen and we can continue by simulating the knockout stage. In the case of draws in the knockout stage, we simulate extra-time by a second simulated result. However, here we multiply the expected number of goals by the factor 0.33 to account for the shorter time to score (30 min instead of 90 min). In the case of a further draw in extra-time we simulate the penalty shootout by a (virtual) coin flip.

Following this strategy, a whole tournament run can be simulated, which we repeat 100,000 times. Based on these simulations, for each of the 32 participating teams probabilities to reach the single knockout stages and, finally, to win the tournament are obtained. These are summarized in Table 9 together with the winning probabilities based on the ODDSET odds for comparison.

We can see that, according to our hybrid random forest model, Spain was the favored team with a predicted winning probability of 13.7% followed by Germany, France, Brazil and Belgium. Overall, this result seems in line with the probabilities from the bookmakers, as we can see in the last column. While Oddset favors Germany and Brazil, the hybrid random forest model predicts a slight advantage for Spain. However, we can see no clear favorite, as several teams seem to have good chances. In retrospect, the early drop-outs of Germany and Spain seem rather surprising. While Spain at least played a successful group stage finishing in first place, Germany performed unexpectedly bad with two defeats during the group stage. The probability for such an early drop-out of Germany was predicted to be only around 22% and, therefore, could be seen as the biggest surprise of the tournament. Spain failed in the round-of-16 against host Russia in a penalty shoot-out and, hence, did not reach the quarter finals (the probability for this event had been predicted to be about 39%). Beside the probabilities of becoming world champion, Table 9 provides some further interesting insights also for the single stages within the tournament. For example, it is interesting to see that the two favored teams Spain and Germany had almost equal chances to at least reach the round-of-sixteen (80.5% and 78.0%, respectively), while the probabilities to at least reach the quarter finals differ significantly. While Spain had a probability of 61.2% to reach at least the quarter finals, Germany only achieved a probability of 49.0%. Obviously, in contrast to Spain, Germany had a rather high chance to meet a strong opponent in the round-of-sixteen. In case they would have reached the round-of-sixteen, Germany would have faced either Brazil, Switzerland, Serbia or Costa Rica, while Spain would have faced Uruguay, Russia, Saudi Arabia or Egypt. In the following rounds, Germany catches up to Spain finally ending up with almost equal winning probabilities.

Most probable tournament course

Finally, based on the 100,000 simulations, we also provide the most probable tournament course. For each of the eight groups we selected the most probable final group standing, while also considering the order of the first two places, but not the irrelevant order of the teams on places three and four. The results together with the corresponding probabilities are presented in Table 10.

Obviously, there are large differences with respect to the groups' balances. While in Group B and Group G the model forecasts Spain followed by Portugal as well as Belgium followed by England with rather high probabilities of 27.7% and 26.7%, respectively, other groups such as Group D, Group E, Group F and Group H seem to be more volatile. Now that we know the true tournament outcome, it is worth a note that indeed in Group B and G the first two places were exactly taken by the two forecasted teams, while in Group F and H there were some surprises.

Table 10:

Most probable final group standings together with the corresponding probabilities for the FIFA World Cup 2018 based on 100,000 simulation runs.

Moreover, we provide the most probable course of the knockout stage in Figure 4. The most likely round-of-sixteen directly results from those teams qualifying for the knockout stage in Table 10. For all following matches we compute the probabilities for the respective two teams (say team A and team B) to go to the next stage. This is done by applying the Skellam distribution to first get the probabilities for *A wins*, *draw* and *B wins* after 90 minutes. Second, the probability for *draw* is distributed between teams A and B again following the principles of extra-time and penalty shootouts we already applied for draws in the knockout stage in the previous section. This way the probabilities for *A wins* and *B wins* add up to 1, as is necessary for the knockout stage. In Figure 4, the probabilities accompanying the edges of the tournament tree represent the probability of the favored team to proceed to the next stage.

Figure 4: Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2018 based on 100,000 simulation runs.

Figure 4:

Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2018 based on 100,000 simulation runs.

In the most probable tournament course Germany wins the World Cup. However, again it becomes obvious that with (in that case) Switzerland the German team would have had to face a much stronger opponent than Spain in the round-of-sixteen. Even though they still were the favorite in this match, they would have succeeded to move on to the quarter finals only with a probability of 58%. While in the most probable course of the knock-out stage, though having tough times in all single stages, Germany would have made its way into the final and defended the title, the previous section showed that generally still Spain was the most likely winner.

We wish to attract the reader's attention to the fact that, despite being the most probable tournament course, due to the myriad of possible constellations this exact tournament course still was extremely unlikely: if we take the product of all single probabilities of Table 10 and Figure 4, its overall probability yields $7.63 \cdot 10^{-9}\%$. Hence, deviations of the true tournament course from the model's most probable one were not only possible, but very likely.

References

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones. 2016. "mlr: Machine Learning in R." *Journal of Machine Learning Research* 17:1–5. http://jmlr.org/papers/v17/15-066.html (http://jmlr.org/papers/v17/15-066.html) .

Boshnakov, G., T. Kharrat, and I. G. McHale. 2017. "A Bivariate Weibull Count Model for Forecasting Association Football Scores." *International Journal of Forecasting* 33:458–466. http://www.sciencedirect.com/science/article/pii/S0169207017300018 (http://www.sciencedirect.com/science/article/pii/S0169207017300018) . 10.1016/j.ijforecast.2016.11.006 (https://doi.org/10.1016/j.ijforecast.2016.11.006)

Breiman, L. 2001. "Random Forests." *Machine Learning* 45:5–32. 10.1023/A:1010933404324 (https://doi.org/10.1023/A:1010933404324)

Breiman, L., J. H. Friedman, R. A. Olshen, and J. C. Stone. 1984. *Classification and Regression Trees.* Monterey, CA: Wadsworth.

Dixon, M. J. and S. G. Coles. 1997. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46:265–280. 10.1111/1467-9876.00065 (https://doi.org/10.1111/1467-9876.00065)

Dyte, D. and S. R. Clarke. 2000. "A Ratings Based Poisson Model for World Cup Soccer Simulation." *Journal of the Operational Research Society* 51(8):993–998. 10.1057/palgrave.jors.2600997 (https://doi.org/10.1057/palgrave.jors.2600997)

Friedman, J., T. Hastie, and R. Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33:1. 10.18637/jss.v033.i01 (https://doi.org/10.18637/jss.v033.i01)

Gneiting, T. and A. E. Raftery. 2007. “Strictly Proper Scoring Rules, Prediction, and Estimation.” *Journal of the American Statistical Association* 102:359–378.
10.1198/016214506000001437 (https://doi.org/10.1198/016214506000001437)

Groll, A. and J. Abedieh. 2013. “Spain Retains its Title and Sets a New Record – Generalized Linear Mixed Models on European Football Championships.” *Journal of Quantitative Analysis in Sports* 9:51–66.
10.1515/jqas-2012-0046 (https://doi.org/10.1515/jqas-2012-0046)

Groll, A., T. Kneib, A. Mayr, and G. Schauberger. 2018. “On the Dependency of Soccer Scores – A Sparse Bivariate Poisson Model for the UEFA European Football Championship 2016.” *Journal of Quantitative Analysis in Sports* 14:65–79.
10.1515/jqas-2017-0067 (https://doi.org/10.1515/jqas-2017-0067)

Groll, A., G. Schauberger, and G. Tutz. 2015. “Prediction of Major International Soccer Tournaments Based on Team-Specific Regularized Poisson Regression: An Application to the FIFA World Cup 2014.” *Journal of Quantitative Analysis in Sports* 11:97–115.
10.1515/jqas-2014-0051 (https://doi.org/10.1515/jqas-2014-0051)

Hoerl, A. E. and R. W. Kennard. 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics* 12:55–67.
10.1080/00401706.1970.10488634 (https://doi.org/10.1080/00401706.1970.10488634)

Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan. 2006. “Survival Ensembles.” *Biostatistics* 7:355–373.
10.1093/biostatistics/kxj011 (https://doi.org/10.1093/biostatistics/kxj011)

PubMed (https://pubmed.ncbi.nlm.nih.gov/16344280/)

Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner. 2017. *mboost: Model-Based Boosting*.
https://CRAN.R-project.org/package=mboost (https://CRAN.R-project.org/package=mboost) , R package version 2.8-1.

Karlis, D. and I. Ntzoufras. 2003. “Analysis of Sports Data by Using Bivariate Poisson Models.” *The Statistician* 52:381–393.
10.1111/1467-9884.00366 (https://doi.org/10.1111/1467-9884.00366)

Kelly, J. L. 1956. “A New Interpretation of Information Rate.” *Bell System Technical Journal* 35:917–926.
http://dx.doi.org/10.1002/j.1538-7305.1956.tb03809.x (http://dx.doi.org/10.1002/j.1538-7305.1956.tb03809.x) .
10.1142/9789814293501_0003 (https://doi.org/10.1142/9789814293501_0003)

Koopman, S. J. and R. Lit. 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)* 178:167–186.
10.1111/rssa.12042 (https://doi.org/10.1111/rssa.12042)

Leitner, C., A. Zeileis, and K. Hornik. 2010. “Forecasting Sports Tournaments by Ratings of (Prob)Abilities: A Comparison for the EURO 2008.” *International Journal of Forecasting* 26(3):471–481.
10.1016/j.ijforecast.2009.10.001 (https://doi.org/10.1016/j.ijforecast.2009.10.001)

Ley, C., T. Van de Wiele, and H. Van Eetvelde. 2019. “Ranking Soccer Teams on the Basis of their Current Strength: A Comparison of Maximum Likelihood Approaches.” *Statistical Modelling* 19:55–77.
https://doi.org/10.1177/1471082X18817650 (https://doi.org/10.1177/1471082X18817650) .
10.1177/1471082X18817650 (https://doi.org/10.1177/1471082X18817650)

Maher, M. J. 1982. “Modelling Association Football Scores.” *Statistica Neerlandica* 36:109–118.
10.1111/j.1467-9574.1982.tb00782.x (https://doi.org/10.1111/j.1467-9574.1982.tb00782.x)

McHale, I. and P. Scarf. 2007. “Modelling Soccer Matches Using Bivariate Discrete Distributions with General Dependence Structure.” *Statistica Neerlandica* 61:432–445. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2007.00368.x (https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2007.00368.x) .
10.1111/j.1467-9574.2007.00368.x (https://doi.org/10.1111/j.1467-9574.2007.00368.x)

McHale, I. G. and P. A. Scarf. 2011. “Modelling the Dependence of Goals Scored by Opposing Teams in International Soccer Matches.” *Statistical Modelling* 41:219–236.
10.1177/1471082X1001100303 (https://doi.org/10.1177/1471082X1001100303)

Probst, P. and A.-L. Boulesteix. 2017. “To Tune or not to Tune the Number of Trees in Random Forest?” *Journal of Machine Learning Research* 18:181:1–181:18.

Quinlan, J. R. 1986. “Induction of Decision Trees.” *Machine Learning* 1:81–106.
10.1007/BF00116251 (https://doi.org/10.1007/BF00116251)

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (https://www.R-project.org/) .

Schauberger, G. and A. Groll. 2018. “Predicting Matches in International Football Tournaments with Random Forests.” *Statistical Modelling* 18:460–482. https://doi.org/10.1177/1471082X18799934 (https://doi.org/10.1177/1471082X18799934) .
10.1177/1471082X18799934 (https://doi.org/10.1177/1471082X18799934)

Skellam, J. G. 1946. “The Frequency Distribution of the Difference between Two Poisson Variates Belonging to Different Populations.” *Journal of the Royal Statistical Society. Series A (General)* 109:296–296.
10.2307/2981372 (https://doi.org/10.2307/2981372)

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8:25.
10.1186/1471-2105-8-25 (https://doi.org/10.1186/1471-2105-8-25)

PubMed (https://pubmed.ncbi.nlm.nih.gov/17254353/)
PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1796903/)

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics* 9:307.
10.1186/1471-2105-9-307 (https://doi.org/10.1186/1471-2105-9-307)

PubMed (https://pubmed.ncbi.nlm.nih.gov/18620558/)
PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491635/)

Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society* B58:267–288.
10.1111/j.2517-6161.1996.tb02080.x (https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)

Wright, M. N. and A. Ziegler. 2017. “Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” *Journal of Statistical Software* 77:1–17.
10.18637/jss.v077.i01 (https://doi.org/10.18637/jss.v077.i01)

Yuan, M. and Y. Lin. 2006. “Model Selection and Estimation in Regression with Grouped Variables.” *Journal of the Royal Statistical Society* B68:49–67.
10.1111/j.1467-9868.2005.00532.x (https://doi.org/10.1111/j.1467-9868.2005.00532.x)

Published Online: 2019-07-10
Published in Print: 2019-10-25

— oder —

*PDF***30,00 €**

From the journal

Journal of Quantitative Analysis in Sports

Volume 15 Issue 4

Articles in the same Issue

Frontmatter

Offensive or defensive play in soccer: a game-theoretical approach

A hybrid random forest to predict soccer matches in international tournaments

Bayesian statistics meets sports: a comprehensive review

A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Using a Markov decision process to model the value of the sacrifice bunt

Combinatorial models of cross-country dual meets: what is a big victory?