Andreas Groll: A hybrid random forest approach for modeling and prediction of international football matches

Andreas Groll, Faculty of Statistics, TU Dortmund University, Vogelpothweg 87, 44227 Dortmund, Germany
(E-mail: groll@statistik.tu-dortmund.de)

Abstract. Many approaches that analyze and predict the results of international matches in football/soccer are based on statistical models incorporating several potentially influential features with respect to a national team's sportive success, such as the bookmakers' ratings or the FIFA ranking. Based on all matches from the four previous FIFA World Cups 2002-2014, we compare the most common *regression models* that are based on the teams' feature information with regard to their predictive performances. Furthermore, an alternative modeling class is investigated, so-called *random forests* (Breimann, 2001)., which can be seen as mixture between machine learning and statistical modeling and are known for their high predictive power.
Within the framework of Generalized Linear Models (GLMs), the most frequently used type of regression models in the literature is the *Poisson model*. It can easily be combined with different regularization methods such as penalization (see, e.g., Groll and Abedieh, 2013; Groll et al., 2015) or boosting (Groll et al., 2018).
Our main focus, however, is on the incorporation of so-called hybrid predictors, i.e. features which were obtained by a separate statistical model. We are particularly interested in how those can improve the predictive performance of the models. (Groll et al., 2019).
For these different modeling techniques, the predictive performance with regard to several goodness-of-fit measures is compared. Based on the estimates of the best performing method all match outcomes of the FIFA World Cup 2018 in Russia are repeatedly simulated (1,000,000 times), resulting in winning probabilities for all participating national teams.
Finally, we shortly sketch how we have progressed in this research line since the FIFA World Cup 2018.

Keywords: Football, FIFA World Cups, Poisson regression, Random forests, Regularization, hybrid modeling.

References

1. L. Breimann (2001). Random Forests. *Machine Learning*, **45(1)**, 5-32.
3. A. Groll 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(1)**, 51-66.
4. A. Groll, 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(2)**, 65-79.
5. A. Groll, 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(2)**, 97-115.
6. A. Groll, A., C. Ley, G. Schauberger, and H. Van Eetvelde (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, **15(4)**, 271-287.