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

A summary of the talk given by Andreas Groll at the research seminar

Robert Bajons

Introduction

The modeling and prediction of soccer matches have raised increased popularity among the statistical community in the past years and have become a major research area within the field of sports statistics (cf. Groll and Schauberger (2019)). In the literature, there are two main directions for modeling match results. The first approach, which will be the focus of this review is to model the score of the competing teams in various forms. The second approach is to model match outcomes directly, i.e. to model win, draw or loss probabilities.

The main model class used for modeling the score of teams is Poisson regression. In general, the idea is to model the number of goals scored by each competing team in a single football match. That is we consider two variables, $X_{ij} \sim Po(\lambda_{ij})$ and $Y_{ij} \sim Po(\mu_{ij})$, where X_{ij} denotes the goals scored from team i playing against team j and Y_{ij} the goals scored from team j when playing against team i. The trick is now to find an accurate estimate for the expected number of goals of λ_{ij} and μ_{ij}.

In the simplest case, conditional on the teams' abilities or covariates such as economic and sportive factors of the country, the two Poisson distributions are treated as independent (Groll, Schauberger, and Tutz (2015)). Many approaches however allow for dependence between the two score variables. A commonly used approach is the bivariate poisson model initially proposed by Karlis and Ntzoufras (2003), which is able to account for (positive) dependencies between the scores. In a closely related approach, such models are used as ranking methods for football teams. In this sense, no covariates are taken into account, but rather (defensive and offensive) ability parameters are estimated from a large set of matches, for an overview see Ley, Wiele, and Eetvelde (2019). The advantage of such an approach is that there is no need for collecting a potentially huge amount of covariates. Furthermore, the ranking approach and the pure covariate approach can easily be combined.

A fundamentally different approach for modeling the score is to use a machine learning algorithm such as random forests. Schauberger and Groll (2018) investigated the predictive potential of random forests in the context of international football matches and compared different types of random forests on data containing all matches of the FIFA World Cups 2002–2014 with conventional regression methods for count data, such as the Poisson models from above.

Finally, it can be shown, that the combination of the random forest with the ability estimates derived from Poisson models yields improved overall results. Groll, Ley, Schauberger, and Eetvelde (2019) show the superior performance of their so-called *hybrid random forest model* over typical regression approaches.

From Poisson Regression to Random Forests

Independent Poisson Models

In this class of models, the single scores are used as response variables and (conditionally on the covariates) a Poisson distribution is assumed. As mentioned in the introduction a crucial assumption is conditional independence of the two scores of one match given covariates. Each score is treated as a single observation so that per match there are two observations. Accordingly, for n teams the respective model has the form

$$\begin{aligned} Y_{ijk} \mid \boldsymbol{x}_{ik}, \boldsymbol{x}_{jk} &\sim \text{Po}\left(\lambda_{ijk}\right), \\ \log\left(\lambda_{ijk}\right) &= \beta_0 + \left(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk}\right)^\top \boldsymbol{\beta} + \boldsymbol{z}_{ik}^\top \boldsymbol{\gamma} + \boldsymbol{z}_{jk}^\top \boldsymbol{\delta}. \end{aligned} \tag{1}$$

Here, Y_{ijk} denotes the score of team i against team j in tournament k, where $i, j \in \{1, \ldots, n\}, i \neq j$. The metric characteristics of both competing teams are captured in the p-dimensional vectors x_{ik}, x_{jk}, while z_{ik} and z_{jk} capture dummy variables for the categorical covariates separately for the considered teams and their respective opponents (cf. Schauberger and Groll (2018)). For these variables, it is not sensible to build differences between the respective values. Furthermore, $\boldsymbol{\beta}$ is a parameter vector which captures the linear effects of all metric covariate differences and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ collect the effects of the dummy variables corresponding to the teams and their opponents, respectively. Groll, Schauberger, and Tutz (2015) extend the formulation of equation (1) to allow for team attacking and defensive ability parameters of the teams such that[1]

$$\begin{aligned} y_{ijk} \mid \boldsymbol{x}_{ik}, \boldsymbol{x}_{jk} &\sim \text{Pois}\left(\lambda_{ijk}\right) \\ \log\left(\lambda_{ijk}\right) &= \beta_0 + \left(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk}\right)^\top \boldsymbol{\beta} + \boldsymbol{z}_{ik}^\top \boldsymbol{\gamma} + \boldsymbol{z}_{jk}^\top \boldsymbol{\delta} + att_i - def_j. \end{aligned} \tag{2}$$

Now in the presence of a high dimensional set of covariates, a usual procedure is to perform regularized estimation, i.e. to apply penalization in the estimation procedure in order to reduce the variance of the parameter estimates and provide better predictive performance than regularized estimators. Thus in order to estimate the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})$ an additional penalization term is added to the likelihood function. In case of ability parameters $\boldsymbol{att} = (att_1, \ldots, att_n)$ $\boldsymbol{def} = (def_1, \ldots, def_n)$ added as in (2), add extra constraint termed *group lasso* penalty is added, such that both effects corresponding to the same team form a group of parameters. In group lasso, groups of parameters can be defined where variable selection is then applied to the group as a whole. Therefore, such a penalty shrinks whole groups of parameters to 0. Thus for estimation, instead of the regular likelihood $l\left(\beta_0, \boldsymbol{\theta}\right)$ the penalized likelihood

$$l_p\left(\beta_0, \boldsymbol{\theta}\right) = l\left(\beta_0, \boldsymbol{\theta}\right) + \lambda P\left(\beta_0, \tilde{\boldsymbol{\theta}}\right) \tag{3}$$

is maximized, where $P\left(\beta_0, \tilde{\boldsymbol{\theta}}\right)$ is either the ordinary lasso penalty (for equation (1)) or the a penalty of the form

$$P\left(\beta_0, \tilde{\boldsymbol{\theta}}\right) = P\left(\beta_0, \boldsymbol{\theta}, \boldsymbol{att}, \boldsymbol{def}\right) = \sum_{v=1}^{\tilde{p}} |\theta_v| + \sqrt{2} \sum_{i=1}^{n} \sqrt{att_i^2 + def_i^2} \tag{4}$$

for equation (2). In both cases λ is a tuning parameter determined by cross-validation.

Poisson Ranking Models

In this section, we describe how Poisson models can be used to obtain rankings that reflect a team's current ability. As mentioned in the introduction and seen from equation (2), there is a close connection to independent poisson models of section . In the most simple case, one could simply ignore the exogenous variables in the model (2). Such ranking models are appealing for their simplicity (i.e. no need for possibly tedious data collection) as well as for their nice interpretability in terms of being able to compare teams. Ley, Wiele, and

[1] Note that actually in their work Groll, Schauberger, and Tutz (2015) did not include the categorical covariates, i.e. they do not estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. However, as shown in Schauberger and Groll (2018) they can be added quite straightforwardly, after having taken care of identifiability issues.

Eetvelde (2019) provide a nice overview of possible models in this category and we will follow Groll, Ley, Schauberger, and Eetvelde (2019) and describe the ranking models in term of the most elaborate model namely the bivariate Poisson model from Karlis and Ntzoufras (2003). While these models on their own possess only moderate predictive abilities, the main idea here is to provide a ranking of teams as well as team strength parameters, which may help in predicting match outcomes when using them in a hybrid model.

The bivariate Poisson model can be formalized in the following way. If we have M matches featuring a total of n teams, we write Y_{ijm} the random variable number of goals scored by team i against team j ($i, j \in \{1, \ldots, n\}$) in match m (where $m \in \{1, \ldots, M\}$). The joint probability function of the home and away score is then given by the bivariate Poisson probability mass function,

$$\begin{aligned} \mathrm{P}\left(Y_{ijm} = z, Y_{jim} = y\right) = & \frac{\lambda_{ijm}^{z} \lambda_{jim}^{y}}{z!y!} \exp\left(-\left(\lambda_{ijm} + \lambda_{jim} + \lambda_C\right)\right) \\ & \cdot \sum_{k=0}^{\min(z,y)} \binom{z}{k} \binom{y}{k} k! \left(\frac{\lambda_C}{\lambda_{ijm}\lambda_{jim}}\right)^{k}, \end{aligned} \tag{5}$$

where λ_C is a covariance parameter assumed to be constant over all matches and λ_{ijm} is the expected number of goals for team i against team j in match m, which we model as

$$\log\left(\lambda_{ijm}\right) = \beta_0 + \left(r_i - r_j\right) + h \cdot \mathbb{I}(\text{ team } i \text{ playing at home}) \,, \tag{6}$$

where β_0 is a common intercept and r_i and r_j are the strength parameters of team i and team j, respectively. Since the ratings are unique up to addition by a constant, we add the constraint that the sum of the ratings has to equal zero. The last term h represents the home effect and is only added if team i plays at home. Note that we have the independent Poisson model if $\lambda_C = 0$.

Estimation of strength parameters is traditionally again done via maximum likelihood. However, in order to account for the fact that team strengths vary in time and we are mostly interested in the actual strength before a tournament, Groll, Ley, Schauberger, and Eetvelde (2019) suggest using a weighted maximum likelihood approach. In this way, they adjust the importance of data points, i.e. matches, for estimation of the strength parameters, such that more recent matches have more influence on the estimation. Furthermore, another weight is placed on the type of match that is played such that important matches (e.g. tournament elimination matches) are given more weight than less relevant matches (e.g. friendly matches). The overall weighted likelihood then reads

$$L = \prod_{m=1}^{M} \left(\mathrm{P}\left(Y_{ijm} = y_{ijm}, Y_{jim} = y_{jim}\right)\right)^{w_{type,m} \cdot w_{time,m}} , \tag{7}$$

where y_{ijm} and y_{jim} stand for the actual number of goals scored by teams i and j in match m. The explicit formulations of the weight functions $w_{type,m}$ and $w_{time,m}$ can be found in Groll, Ley, Schauberger, and Eetvelde (2019) and are based on findings from Ley, Wiele, and Eetvelde (2019). The values of the strength parameters $r_1, \ldots, r_n$, which determine the resulting ranking, are computed numerically as maximum likelihood estimates on the basis of historic match data. These parameters also allow to predict future match outcomes thanks to the formula (6).

Random Forest Models and Hybrid Forms

We finally turn to a completely different class of models, namely random forests. In this brief review, we refrain from revisiting the general ideas of random forest and refer the interested reader to Schauberger and Groll (2018), who provide an overview as well as further references. They also discuss how such models are used for modeling football matches. As random forests are a very flexible class of machine learning models, they can be used in various ways, however in order to concur with the rest of this summary, we only discuss their usage for predicting the number of goals scored.

Since the response variable of interest is the metric variable number of goals, regression trees are used for predicting the expected number of goals. Schauberger and Groll (2018) mention two different variants of random forests, one being the classical algorithm as introduced by Breiman (2001), the other one being presented by Hothorn, Hornik, and Zeileis (2006), which uses the principle of conditional inference to construct the final trees. The advantage of the latter variant of random forests is that they avoid selection bias in cases where the covariates have different scales, e.g. numerical vs. categorical with many categories. Schauberger and Groll (2018) further show that the conditional random forests also result in better predictive performance for the football data. Finally, in order to use the estimates of the expected number of goals from the random forests for prediction of match results or tournaments, the predicted value from these models is used as an estimate for the event rate λ of a Poisson distribution. The idea is thus similar to the previous sections, i.e. two independent Poisson distributions (conditional on the covariates) for both scores are used to model match outcomes.

Groll, Ley, Schauberger, and Eetvelde (2019) extend the idea of using random forest by combining ranking models as described in section and the random forest idea resulting in what they call a *hybrid random forest model*. They propose to use the ranking approach to generate a new (highly informative) covariate that can be incorporated into the random forest model. It turns out that the additional strength variable, although seemingly similar to exogenous variables such as the FIFA ranking of teams, is much more informative. Further, Groll, Ley, Schauberger, and Eetvelde (2019) show that their hybrid approach leads to superior predictive performance when modelling match outcomes.

Application to World Cup Data

Various versions of the hybrid random forest have been applied to major tournaments of men's and women's football in Groll, Ley, Schauberger, and Eetvelde (2019), Groll, Ley, Schauberger, Eetvelde, and Zeileis (2019) and Groll et al. (2021). We briefly describe the main approach as presented in Groll, Ley, Schauberger, and Eetvelde (2019). In order to model and predict tournament outcomes such as the FIFA world cup 2018, the idea is to train a hybrid random forest only on available data from past world cups. Information from other international matches such as friendly matches, qualifier matches, or matches from other tournaments is only indirectly included from the estimated ability parameters that are fed into the random forest prediction model.

Based on the hybrid random forest fitted on the data of 4 past world cups (FIFA world cups 2002–2014), the authors simulated the FIFA world cup 2018 100,000 times. These simulations allowed to compute tournament-winning probabilities for all participating teams as well as the most probable tournament course. A retrospective analysis of the world cup shows that the performance of the hybrid random forest is able to outperform other approaches such as pure ranking methods, different hybrid variants (such as a lasso hybrid glm) and that the model is even able to outperform bookmaker betting odds.

Discussion

In his talk *A hybrid random forest approach for modeling and prediction of international football matches*, Andreas Groll gave an overview of the derivation and evolution of an approach to model international football tournaments. The so-called *hybrid random forest* model combines early works on ranking football teams using Poisson models and machine learning algorithms such as random forests. The author showed that the hybrid approach performs better than simple approaches and that incorporating team abilities via ranking models are very informative covariates for the prediction of matches.

The hybrid approach can be nicely extended in such a way that more covariates from other statistical models can be incorporated. Groll et al. (2021) for example add two hybrid variables derived from a bookmaker consensus model, see e.g. Leitner, Zeileis, and Hornik (2010), and a plus-minus rating model, see e.g. Hvattum and Gelade (2021). Moreover, an interesting extension mentioned is to compare the random forest with other state-of-the-art machine learning algorithm such as gradient boosting.

Finally, an interesting application is to use similar ideas for women's football tournaments. While Groll, Ley, Schauberger, Eetvelde, and Zeileis (2019) intend to adopt the same methodology to the FIFA women's 2019 world cup it is unclear whether it is sensible to imitate the procedure for women's football. Michels, Ötting, and Karlis (2023) for example provide an interesting alternative to classical poisson models for deriving ability parameters for women national teams. Their work is based on the observation that in women's football, scorelines have different characteristics than in men's football. This could be nicely incorporated in order to provide a more accurate *hybrid* covariate for the hybrid random forest model. Furthermore, it is not clear whether in women's sport there are different influential exogenous variables that need to be taken into account. A survey analyzing the covariate effect for women's football could provide more interesting insight into this fact and this could benefit the prediction models described in this summary.

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Groll, Andreas, Lars Magnus Hvattum, Christophe Ley, Franziska Popp, Günther Schauberger, Hans Van Eetvelde, and Achim Zeileis. 2021. "Hybrid Machine Learning Forecasts for the UEFA EURO 2020." *ArXiv* abs/2106.05799.

Groll, Andreas, Christophe Ley, Gunther Schauberger, Hans Van Eetvelde, and Achim Zeileis. 2019. "Hybrid Machine Learning Forecasts for the FIFA Women's World Cup 2019." *ArXiv*. https://arxiv.org/abs/1906.01131.

Groll, Andreas, Cristophe Ley, Gunther Schauberger, and Hans Van Eetvelde. 2019. *Journal of Quantitative Analysis in Sports* 15 (4): 271–87. https://doi.org/doi:10.1515/jqas-2018-0060.

Groll, Andreas, and Gunther Schauberger. 2019. "Prediction of Soccer Matches." In *Wiley StatsRef: Statistics Reference Online*, 1–7. John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781118445112.stat08162.

Groll, Andreas, Gunther Schauberger, and Gerhard Tutz. 2015. *Journal of Quantitative Analysis in Sports* 11 (2): 97–115. https://doi.org/doi:10.1515/jqas-2014-0051.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15 (3): 651–74. https://doi.org/10.1198/106186006X133933.

Hvattum, Lars Magnus, and Garry A. Gelade. 2021. "Comparing Bottom-up and Top-down Ratings for Individual Soccer Players." *International Journal of Computer Science in Sport* 20 (1): 23–42. https://doi.org/doi:10.2478/ijcss-2021-0002.

Karlis, Dimitris, and Ioannis Ntzoufras. 2003. "Analysis of Sports Data by Using Bivariate Poisson Models." *Journal of the Royal Statistical Society: Series D (The Statistician)* 52 (3): 381–93. https://doi.org/https://doi.org/10.1111/1467-9884.00366.

Leitner, Christoph, Achim Zeileis, and Kurt Hornik. 2010. "Forecasting Sports Tournaments by Ratings of (Prob)abilities: A Comparison for the EURO 2008." *International Journal of Forecasting* 26 (3): 471–81. https://doi.org/https://doi.org/10.1016/j.ijforecast.2009.10.001.

Ley, Christophe, Tom Van de Wiele, and Hans Van Eetvelde. 2019. "Ranking Soccer Teams on the Basis of Their Current Strength: A Comparison of Maximum Likelihood Approaches." *Statistical Modelling* 19 (1): 55–73. https://doi.org/10.1177/1471082X18817650.

Michels, Rouven, Marius Ötting, and Dimitris Karlis. 2023. "Extending the Dixon and Coles Model: An Application to Women's Football Data." *ArXiv*. https://arxiv.org/abs/2307.02139.

Schauberger, Gunther, and Andreas Groll. 2018. "Predicting Matches in International Football Tournaments with Random Forests." *Statistical Modelling* 18 (5-6): 460–82. https://doi.org/10.1177/1471082X18799934.