

Hybrid Machine Learning Forecasts for the UEFA EURO 2020

A. Groll ^{*} L. M. Hvattum [†] C. Ley [‡] F. Popp [§]
G. Schaubberger [¶] H. Van Eetvelde ^{||} A. Zeileis ^{**}

June 7, 2021

Abstract Three state-of-the-art statistical ranking methods for forecasting football matches are combined with several other predictors in a hybrid machine learning model. Namely an ability estimate for every team based on historic matches; an ability estimate for every team based on bookmaker consensus; average plus-minus player ratings based on their individual performances in their home clubs and national teams; and further team covariates (e.g., market value, team structure) and country-specific socio-economic factors (population, GDP). The proposed combined approach is used for learning the number of goals scored in the matches from the four previous UEFA EUROs 2004-2016 and then applied to current information to forecast the upcoming UEFA EURO 2020. Based on the resulting estimates, the tournament is simulated repeatedly and winning probabilities are obtained for all teams. A random forest model favors the current World Champion France with a winning probability of 14.8% before England (13.5%) and Spain (12.3%). Additionally, we provide survival probabilities for all teams and at all tournament stages.

Keywords: UEFA EURO 2020, Football, Machine Learning, Team abilities, Sports tournaments.

1 Introduction

The use of statistical and machine learning models to predict the outcome of international football tournaments, such as European championships (EUROs) or FIFA World Cups, has become pretty popular in recent years. One model class that is frequently used is the class of Poisson regression models. These directly model the number of goals scored by both

^{*}Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany, groll@statistik.tu-dortmund.de

[†]Molde University College, Molde, Norway, Lars.M.Hvattum@himolde.no

[‡]Faculty of Sciences, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, 9000 Gent, Belgium, Christophe.Ley@UGent.be

[§]Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany, franziska.popp@tu-dortmund.de

[¶]Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, g.schauberger@tum.de

^{||}Faculty of Sciences, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, 9000 Gent, Belgium, hans.vaneetvelde@ugent.be

^{**}Department of Statistics, Universität Innsbruck, Austria, Achim.Zeileis@R-project.org

competing teams in a single football match. Let $X_{ij} \sim Po(\lambda_{ij})$ and $Y_{ij} \sim Po(\mu_{ij})$ denote the goals of the first and second team, respectively, in a match between teams i and j , where $i, j \in \{1, \dots, n\}$ and n denotes the total number of teams in the regarded set of matches. For the (non-negative) intensity parameters λ_{ij} and μ_{ij} , which reflect the expected numbers of goals, several modeling strategies exist, which incorporate playing abilities or covariates of the competing teams in different ways.

In the simplest case, conditional on the teams' abilities or covariates, the two Poisson distributions are treated as independent. For example, Dyte and Clarke (2000) applied this model to data from FIFA World Cups and let the Poisson intensities of both competing teams depend on their FIFA ranks. Groll and Abedieh (2013) and Groll et al. (2015) considered a large set of potentially influential variables for UEFA EURO and World Cup data, respectively, and used L_1 -penalized approaches to detect a sparse set of relevant predictors. Based on these, forecasts for the UEFA EURO 2012 and FIFA World Cup 2014 tournaments were provided. These approaches showed that, when many covariates are regarded or the predictive power of the individual variables is not clear in advance, regularized estimation approaches can be beneficial.

These approaches can be generalized in different ways to allow for dependent scores. For example, Dixon and Coles (1997) identified a (slightly negative) correlation between the scores and introduced an additional dependence parameter. Karlis and Ntzoufras (2003) and Groll et al. (2018) modeled the scores of both teams by a bivariate Poisson distribution, which is able to account for (positive) dependencies between the scores. If also negative dependencies should be accounted for, copula-based models can be used (see, e.g., McHale and Scarf, 2007, McHale and Scarf, 2011 or Boshnakov et al., 2017). A regularized copula regression technique was proposed by van der Wurp et al. (2020).

Closely related to the covariate-based Poisson regression models are Poisson-based ranking methods for football teams. On the basis of a (typically large) set of matches, ability parameters reflecting the current strength of the teams can be estimated by means of maximum likelihood. An overview of the most frequently used Poisson-based ranking methods was provided by Ley et al. (2019).

An alternative ranking approach that is solely based on bookmakers' odds was proposed by Leitner et al. (2010). They calculate winning probabilities for each team by aggregating winning odds from several online bookmakers. Based on these winning probabilities, by inverse tournament simulation team-specific *bookmaker consensus abilities* can be computed by paired comparison models, automatically stripping the effects of the tournament draw. Next, pairwise probabilities for each possible game at the corresponding tournament can be predicted and, finally, the whole tournament can be simulated.

Yet another ranking approach, the plus-minus player rating, calculates ratings of individual players based on the performance of their teams as a whole using underlying match data, both on a national and international level, containing information on the starting line-ups as well as on certain events such as substitutions, red cards, and goals scored, (Hvattum, 2019; Pantuso and Hvattum, 2021).

A fundamentally different modeling approach is based on a random forest – a popular ensemble learning method for classification and regression (Breiman, 2001), which originates from the machine learning and data mining community. Schauburger 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. The random forests provided very satisfactory results

and generally outperformed the regression approaches. Groll et al. (2019a) and Groll et al. (2019b) showed on both women’s and men’s FIFA World Cup data that the predictive performance of random forests could be further improved by combining it with additional ranking methods, leading to what they call a *hybrid random forest model*. The term *hybrid* shall emphasize that some of the features used are themselves estimates from separate statistical models.

In the present work, we carry this strategy forward and combine the random forest on the one hand, but also a so-called *extreme gradient boosting* (xgboost) approach on the other hand, with the Poisson ranking methods from Ley et al. (2019), the bookmaker consensus abilities from Leitner et al. (2010) and plus-minus player ratings from Pantuso and Hvattum (2021). The xgboost method is a sequential ensemble technique, which is known in the machine learning community for its high predictive power. So in a sense, this results in *hybrid* or *combined ranking-based* machine learning approaches, similar to Groll et al. (2019b). The model is fitted to all matches of the UEFA EUROs 2004-2016 and based on the resulting estimates, the UEFA EURO 2020 is then simulated 100,000 times to determine winning probabilities for all 24 participating teams.

A word of caution needs to be said regarding the still ongoing COVID-19 pandemic. Only a very much reduced number of fans will be allowed to attend the matches and support their teams, the preparation of the national teams is certainly different from their usual habits due to sanitary restrictions, and it is not unlikely that various players will need to quarantine because they are tested positive at some point or have been in contact with a person affected by COVID-19. All these aspects, combined with the general uncertainty (which probably gets accentuated due to the many travels from country to country at this UEFA EURO tournament), are very likely to affect performances. While the impact of crowd home advantage has been studied over the past year for domestic competitions (Wunderlich et al., 2021), it is yet unknown what this implies during a competition like the UEFA European championship. Therefore, we do expect our predictions, as well as those of other machine learning models, to yield less reliable results than in normal times, since they are trained on COVID-19-free competitions.

The remainder of the manuscript is structured as follows. In Section 2 we describe the four underlying data sets. The first data set covers all matches of the four preceding UEFA EUROs 2004-2016 including covariate information, the second consists of the match results of all international matches played by all national teams during certain time periods. The third data set contains the winning odds from several bookmakers for the single UEFA EUROs regarded in this analysis, and the fourth is based on match data specifying the starting line-ups together with information regarding substitutions, red cards, and goals scored. Next, in Section 3 we briefly explain the basic idea of random forests and extreme gradient boosting, as well as of the three different ranking methods and, finally, how they can be combined to yield hybrid machine learning models. In Section 4, we fit the hybrid machine learning models to the data of the four UEFA EUROs 2004-2016 and investigate their predictive power. Based on the model fit of the model with the best predictive performance, the UEFA EURO 2020 is simulated repeatedly and winning probabilities for all teams are presented (Section 5). Finally, we conclude in Section 6.

2 Data

In this section, we briefly describe four fundamentally different types of data that can be used to model and predict international football tournaments such as the UEFA EURO. The

first type of data covers variables that characterize the participating teams of the single tournaments and connects them to the results of the matches that were played during these tournaments. The second type of data is simply based on the match results of all international matches played by all national teams during certain time periods. These data do not only cover the matches from the specific tournaments but also all qualifiers and friendly matches. The third type of data contains the winning odds from different bookmakers separately for single UEFA EUROs. Finally, the fourth type of data is based on match data specifying the starting line-ups together with information regarding in-game events such as substitutions, red cards, and goals scored.

2.1 Covariate data

The first type of data we describe covers all matches of the four UEFA EUROs 2004-2016 together with several potential influence variables. Basically, we use a similar set of covariates as introduced in Groll et al. (2018), but also added a dummy indicating whether a match is a group or a knock-out stage match. For each participating team, the covariates are observed either for the year of the respective UEFA EURO (e.g., GDP per capita) or shortly before the start of the UEFA EURO (e.g., average age), and, therefore, vary from one UEFA EURO to another.

Several of the variables contain information about the recent performance and sportive success of national teams, as the current form of a national team is supposed to have an influence on the team's success in the upcoming tournament. One additional covariate in this regard, which we will introduce later, is reflecting the national teams' current playing abilities and is related to the second type of data introduced in Section 2.2. The estimates of these ability parameters are based on a separate Poisson ranking model, see Section 3.3 for details, and are denoted by *HistAbility*. Another additional covariate, which is also introduced later, reflects the bookmaker consensus abilities on a log scale (denoted by *logability*) from Leitner et al. (2010) and is related to the third type of data introduced in Section 2.3. Details on this ranking method can be found in Section 3.4. The last type of additional covariates are different versions of the plus-minus (PM) rating of each team as proposed by Pantuso and Hvattum (2021) and described in Section 3.5, which are related to the fourth type of data introduced in Section 2.4.

Beside these sportive variables, certain economic factors as well as variables describing the structure of a team's squad are collected, which are now described in more detail.

Economic Factors:

GDP per capita. To account for the general increase of the (logarithmized) gross domestic product (GDP) during 2004–2016, a ratio of the GDP per capita of the respective country and the worldwide average GDP per capita is used (source: <http://data.worldbank.org/indicator/NY.GDP.PCAP.CD>).

Population. The (logarithmized) population size is used in relation to the respective global population to account for the general world population growth¹.

¹In order to collect data for all participating countries at the UEFA EUROs 2004-2020, different sources had to be used. Amongst the most useful ones are <http://www.wko.at>, <http://www.statista.com/> and <http://epp.eurostat.ec.europa.eu>. For some years the populations of Russia and Ukraine had to be searched individually.

Home advantage:

Host. A dummy variable indicating if a national team is a hosting country.

Neighbor. A dummy variable indicating if a national team is from a neighboring country of the host of the UEFA EURO (including the host itself).

Sportive factors:

Market value. Estimates of the teams' average market values can be found on the webpage <http://www.transfermarkt.de>². For each national team participating in a UEFA EURO these market value estimates (on a log-scale) have been collected (retrospectively) right before the start of the respective tournament.

FIFA ranking. The FIFA ranking system ranks all national teams based on their performance over the last four years³.

UEFA points. The associations' club coefficients rankings are based on the results of each association's clubs in the five previous UEFA CL and Europa League (previously UEFA Cup) seasons⁴.

UEFA starting places. The number of starting places of the corresponding national league for both the UEFA CL and Europa League.

Factors describing the team's structure:

The following variables describe the structure of the teams. They were observed with the 23-player-squad⁵ nominated for the respective UEFA EUROs and were obtained manually both from the website of the German football magazine *kicker*, <http://kicker.de>, and from <http://transfermarkt.de>.

(Second) maximum number of teammates. For each squad, both the maximum and second maximum number of teammates playing together in the same domestic team are counted.

Absolute difference from optimal age. First, the average age of each squad is collected. Then, the absolute distance between each team's average age and the overall average age over all teams (serving as a proxy for the optimal age) is calculated.

²Unfortunately, the archive of the webpage was established not until 4th October 2004, so the average market values of the national teams that we used for the UEFA EURO 2004 can only be seen as a rough approximation, as market values certainly changed after the UEFA EURO 2004.

³The exact formula for the calculation of the underlying FIFA points and all rankings since implementation of the FIFA ranking system can be found at the official FIFA website: <http://de.fifa.com/worldranking/index.html>. Since the calculation formula of the FIFA points changed after both the World Cups 2006 and 2018, the rankings according to FIFA points are used instead of the points. The FIFA ranking was introduced in August 1993.

⁴The exact formula for the calculation of the underlying UEFA points and all rankings since implementation of the UEFA ranking system can be found at the official UEFA website: <http://www.uefa.com/memberassociations/uefarankings/country/index.html>. The rankings determine the number of places allocated to an association (country) in the forthcoming UEFA club competitions.

⁵Note that due to some exceptional rules related to the COVID-19 pandemic, at UEFA EURO 2020 all teams are allowed to nominate 26 players. To make those covariates that are based on the player numbers within the squad comparable to the historic tournaments, we multiply them by the factor 23/26.

Number of Champions League (Europa League) players. As a measurement of the success of the players on club level, the number of players in the semi finals (taking place only few weeks before the respective World Cup) of the UEFA Champions League (CL) and Europa League (EL), respectively, is counted.









Number of players abroad/Legionnaires. For each squad, the number of players playing in clubs abroad (in the season preceding the respective UEFA EURO) is counted.

Factors describing the team’s coach.

Also attributes of a national team’s coach may influence the performance of the team. Therefore, the *age* of the coach (again, as absolute difference from the optimal age approximated by the overall mean) is observed together with a dummy variable⁶, indicating whether the coach has the same *nationality* as his team or not.

In addition, we include a dummy variable indicating whether a certain match is a group- or a knockout match. The motivation for this is that football teams might change their playing style and be more cautious in knockout matches. In total, together with the ranking variables from the different ranking methods described in more detail below, this adds up to 17 variables which were collected separately for each UEFA EURO and each participating team. As an illustration, Table 1 shows the results (1a) and (parts of) the covariates (1b) of the respective teams, exemplarily for the first four matches of the UEFA EURO 2004. We use this data excerpt to illustrate how the final data set is constructed.

Table 1: Exemplary table showing the results of four matches and parts of the covariates of the involved teams.

(a) Table of results				(b) Table of covariates						
				EURO	Team	HistAbility	logability	ave.PM	FIFA.rank	...
POR		1-2	 GRE	2004	Portugal	1.26	0.36	0.133	20	...
ESP		1-0	 RUS	2004	Greece	0.91	-0.38	0.057	34	...
GRE		1-1	 ESP	2004	Spain	1.34	0.33	0.157	3	...
RUS		0-2	 POR	2004	Russia	0.95	-0.31	0.076	30	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

For the modeling techniques that we shall introduce in the following sections, all covariates (including the dummies for *Host*, *Continent* and *Nationality of coach*) are incorporated in the form of differences between the two competing teams. For example, the final variable *HistAbility* will be the difference between the current Poisson ability rankings of both teams. As all machine learning model introduced later use the number of goals of each team directly as the response variable, each match corresponds to two different observations, one per team. For the covariates, we consider differences which are computed from the perspective of the first-named team. The dummy variable *groupstage* corresponds to a single column in the design matrix and is either zero or one for both rows corresponding to the same match. For illustration, the resulting final data structure for the exemplary matches from Table 1 is displayed in Table 2.

⁶These two variables are available on several football data providers, see, for example, <http://www.kicker.de/>.

Table 2: Exemplary table illustrating the data structure.

Goals	Team	Opponent	Group stage	Historic match abilities	Bookmaker abilities	average PM player ranking	FIFA rank	...
1	Portugal	Greece	1	0.34	0.74	0.076	-14	...
2	Greece	Portugal	1	-0.34	-0.74	-0.076	14	...
1	Spain	Russia	1	0.39	0.64	0.081	-27	...
0	Russia	Spain	1	-0.39	-0.64	-0.081	27	...
1	Greece	Spain	1	-0.42	-0.71	-0.100	31	...
1	Spain	Greece	1	0.42	0.71	0.100	-31	...
0	Russia	Portugal	1	-0.31	-0.67	-0.057	10	...
2	Portugal	Russia	1	0.31	0.67	0.057	-10	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.2 Historic match results

The data used for estimating the abilities of the teams consist of the results of every international match played in the last 8 years preceding the considered UEFA EURO. Besides the number of goals, we also need the information of the venue of the match in order to correct for the home effect and the moment in time when a match was played. The reason is that, in the ranking method described in Section 3.3, each match is assigned a weight depending on the time elapsed since the match took place. For example, Table 3 shows an excerpt of the historic match data used to obtain ability estimates for the teams at the UEFA EURO 2004.

Table 3: Historical results from matches prior to a tournament used for estimating current Poisson abilities, exemplarily for the UEFA EURO 2004

Date	Home team	Away team	Score	Country	Neutral
2004-06-06	Czech Republic	Estonia	2-0	Czech Republic	no
2004-06-06	France	Ukraine	1-0	France	no
2004-06-06	Germany	Hungary	0-2	Germany	no
2004-06-06	Latvia	Azerbaijan	2-2	Latvia	no
⋮	⋮	⋮	⋮	⋮	⋮

2.3 Bookmaker data

The basis for the bookmaker consensus model from Leitner et al. (2010), which is explained in more detail in Section 3.4, are the “outright” winning odds for the entire tournament. For the upcoming UEFA EURO 2020 these have been obtained on 2021-05-31 for all 24 teams from 19 online bookmakers via <https://www.oddschecker.com/> and <https://www.bwin.com/>, respectively (see Table 11 in Appendix B). For the tournaments in 2008, 2012, and 2016 we have used the data from Leitner et al. (2010) and Zeileis et al. (2012, 2016), respectively. For the UEFA EURO 2004 we could not find any online collections of outright winning odds but were provided with the odds from the German state betting agency ODDSET⁷ (upon request; see Table 10 in Appendix B).

⁷The possibility of betting on the overall cup winner before the start of the tournament is pretty novel. The German state betting agency ODDSET offered the bet for the first time at the UEFA EURO 2020.

2.4 Plus-minus player rating data

Principally, plus-minus player ratings are based on match data, both on a domestic and international level, specifying the starting line-ups together with information regarding certain events such as substitutions, red cards, and goals scored. For substitutions and red cards, additional information is required regarding which players are involved as well as the time when the event takes place, whereas for goals scored it suffices to know the time when they happen. The matches are then split into segments of maximal duration such that the set of players present on the pitch does not change within a segment. An example of the underlying data sources is displayed in Tables 4 and 5, respectively.

Once this data has been used to calculate individual player ratings, it can be combined with information about squads to calculate different covariates: 1) The mean PM rating of the players in a squad, 2) the median PM rating of the players in a squad, 3) the average PM rating of the 11 highest rated players within a squad, and 4) the number of players that were not included in the squad but that both had a rating that would qualify for the top 11 players in the squad and that had appeared in at least one match for their national team in the last two years before the tournament.

The idea of 4) is that if a team qualifies for a EURO while using their best players, and then some players are injured or otherwise missing from the final squad, then the ratings as given by e.g., Poisson-models will overestimate the quality of the team. Knowing that the squad is likely missing some key players can be helpful with respect to the predictive performance.

Table 4: Underlying data set for deriving different plus-minus player ratings and related features, exemplarily for the UEFA EURO 2004 (part I).

Date	MatchID	Home team	Away team	Neutral
2003-08-31	58218	Man. City	Arsenal	no
⋮	⋮	⋮	⋮	⋮
2003-09-10	87611	Czech Republic	Netherlands	no
⋮	⋮	⋮	⋮	⋮

Table 5: Underlying data set for deriving different plus-minus player ratings and related features, exemplarily for the UEFA EURO 2004 (part II). Abbreviations used in the table: Home team (HT), away team (AT), segment ID (SID), red cards (RC), and goals scored (GS).

MatchID	SID	Time	HT players	AT players	RC (at start)		GS (during)	
					HT	AT	HT	AT
58218	1	0–68	Lehmann, Tarnat, ...	Seaman, Cole, ...	0	0	1	1
58218	2	68–76	Lehmann, Tarnat, ...	Seaman, Cole, ...	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
87611	1	0–13	Cech, Grygera, ...	van der Sar, Stam, ...	0	0	0	0
87611	2	13–20	Cech, Grygera, ...	van der Sar, Stam, ...	0	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
87611	8	81–94	Cech, Ujfalusi, ...	van der Sar, Stam, ...	0	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3 Hybrid machine learning models

In this section, we propose to use hybrid machine learning approaches that combine the information from all four types of data bases introduced above. The proposed methods combine a random forest and an extreme gradient boosting approach, respectively, based on conventional covariate data, with the abilities estimated on the historic match results as used by the Poisson ranking methods, with the abilities obtained from the bookmaker consensus approach and with the plus-minus player rankings. Before introducing the proposed hybrid method, we first separately present the basic ideas of the four model components.

3.1 Random forests

Random forests, originally proposed by Breiman (2001), are an aggregation of a (large) number of classification or regression trees (CARTs). CARTs (Breiman et al., 1984) repeatedly partition the predictor space mostly using binary splits. The goal of the partitioning process is to find partitions such that the respective response values are very homogeneous within a partition but very heterogeneous between partitions. CARTs can be used both for metric responses (regression trees) and for nominal or ordinal responses (classification trees). For prediction, all response values within a partition are aggregated either by averaging (in regression trees) or simply by counting and using majority vote (in classification trees). In this work, we use trees (and, accordingly, random forests) for the prediction of the number of goals a team scores in a match of a UEFA EURO tournament.

Random forests are the aggregation of a large number B (e.g., $B = 5000$) of trees, grown on B bootstrap samples from the original data set. Combining many trees has the advantage that the resulting predictions inherit the feature of unbiasedness from the single trees while reducing the variance of the predictions. For a short introduction to random forests and how they can specifically be used for football data, see Groll et al. (2019a).

In R (R Core Team, 2018), two slightly different variants of regression forests are available: the classical random forest algorithm proposed by Breiman (2001) from the R-package *ranger* (Wright and Ziegler, 2017), and a modification implemented in the function `cforest` from the *party* package⁸. In Schauburger and Groll (2018) and Groll et al. (2019a), the latter package turned out to be superior, but both approaches are tested in this manuscript.

3.2 Extreme gradient boosting

As alternative to parallel ensemble methods like the random forest approach from above, we also consider sequential ensembles. A famous approach in this context is *boosting*, a technique which stems from the machine learning community (Freund and Schapire, 1996) and was later adapted to estimate predictors for statistical models (Friedman et al., 2000; Friedman, 2001). Generally, the concept of an iterative boosting algorithm is to additively combine many weak learners to a powerful ensemble that achieves high accuracy (Schapire, 1990). A main advantage of statistical boosting algorithms is their flexibility for high-dimensional data and their ability to incorporate variable selection in the fitting process

⁸Here, the single trees are constructed following the principle of conditional inference trees as proposed by Hothorn et al. (2006). The main advantage of these conditional inference trees is that they avoid selection bias if covariates have different scales, e.g., numerical vs. categorical with many categories (see, for example, Strobl et al., 2007, and Strobl et al., 2008, for details). Conditional forests share the feature of conditional inference trees of avoiding biased variable selection.

(Mayr et al., 2014). An extensive and enlightening, general overview on (gradient) boosting algorithms can be found in Bühlmann and Hothorn (2007).

Friedman (2001) introduced the idea of gradient tree boosting, using decision trees as learners. The decision trees are repeatedly fitted on the residuals of the previous fit and, hence, are combined to a sequential ensemble. This technique was then further improved by Chen and Guestrin (2016) via introducing additional regularization in the objective function. The regularization terms make the single trees weak learners to avoid overfitting. In a certain boosting iteration, the next tree is additively incorporated into the ensemble after multiplication with a rather small learning rate which makes the learners even weaker. The method is called *extreme gradient boosting*, in short *xgboost*, and is known in the machine learning community for its high predictive power. It has been very successful in prestigious machine learning competitions, such as those organized by Kaggle (<https://www.kaggle.com>). The approach is implemented in the `xgb.train` function from the `xgboost` R package (Chen et al., 2021).

One important aspect is that `xgboost` involves several tuning parameters, such as e.g. the learning rate, the optimal number of boosting steps and several penalty parameters. For this purpose, we specified reasonable, discrete parameter grids and used multivariate 10-fold cross validation via the `xgb.cv` function to determine optimal tuning parameters.

3.3 Current ability ranking based on historic matches

In this section we describe how (based on historic match data, see Section 2.2) Poisson models can be used to obtain rankings that reflect a team’s current ability. We will restrict our attention to the best-performing model according to the comparison achieved in Ley et al. (2019), namely the bivariate Poisson model. The main idea consists in assigning a strength parameter to every team and in estimating those parameters over a period of M matches via weighted maximum likelihood based on time depreciation.

The time decay function is defined as follows: a match played x_m days back gets a weight of

$$w_{time,m}(x_m) = \left(\frac{1}{2}\right)^{\frac{x_m}{\text{Half period}}},$$

meaning that, for instance, a match played *Half period* days ago only contributes half as much as a match played today. We stress that the *Half period* refers to calendar days in a year, not match days. In the present case we use a Half period of 3 years (i.e. 1095 days) based on an optimization procedure to determine which Half period led to the best prediction for men’s football matches based on the average Rank Probability Score (RPS; Gneiting and Raftery, 2007)

The bivariate Poisson ranking model is based on a proposal from Karlis and Ntzoufras (2003) and can be described as follows. 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* ($i, j \in \{1, \dots, n\}$) in match m (where $m \in \{1, \dots, M\}$). The joint probability function of the home and away score is then given by the bivariate Poisson probability mass function,

$$P(Y_{ijm} = z, Y_{jim} = y) = \frac{\lambda_{ijm}^z \lambda_{jim}^y}{z!y!} \exp(-(\lambda_{ijm} + \lambda_{jim} + \lambda_C)) \sum_{k=0}^{\min(z,y)} \binom{z}{k} \binom{y}{k} k! \left(\frac{\lambda_C}{\lambda_{ijm}\lambda_{jim}}\right)^k,$$

where $\lambda_C \geq 0$ is a covariance parameter assumed to be constant over all matches and $\lambda_{ijm} \geq 0$ is the expected number of goals for team i against team j in match m , which we model as

$$\log(\lambda_{ijm}) = \beta_0 + (r_i - r_j) + h \cdot \mathbb{1}(\text{team } i \text{ playing at home}), \quad (1)$$

where $\beta_0 \in \mathbb{R}$ is a common intercept and $r_i, r_j \in \mathbb{R}$ 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 $h \in \mathbb{R}$ represents the home effect and is only added if team i plays at home. We get an independent Poisson model if $\lambda_C = 0$. The overall (weighted) likelihood function then reads

$$L = \prod_{m=1}^M (\mathbb{P}(Y_{ijm} = y_{ijm}, Y_{jim} = y_{jim}))^{w_{time,m}},$$

where y_{ijm} and y_{jim} stand for the actual number of goals scored by teams i and j in match m . The values of the strength parameters r_1, \dots, r_n , which allow ranking the different teams, are computed numerically as maximum likelihood estimates on the basis of historic match data as described in Section 2.2. These parameters also allow to predict future match outcomes thanks to the Equation (1).

3.4 Bookmaker consensus model

Prior to the tournament, on 2021-05-31, we obtained long-term winning odds from 19 online bookmakers. However, before these odds can be transformed to winning probabilities, the stake has to be accounted for and the profit margin of the bookmaker (better known as the “overround”) has to be removed (for further details see Henery, 1999; Forrest et al., 2005). Here, it is assumed that the quoted odds are derived from the underlying “true” odds as: *quoted odds* = *odds* · δ + 1, where +1 is the stake (which is to be paid back to the bookmakers’ customers in case they win) and $\delta < 1$ is the proportion of the bets that is actually paid out by the bookmakers. The overround is the remaining proportion $1 - \delta$ and the main basis of the bookmakers’ profits (see also Wikipedia, 2019 and the links therein). Assuming that each bookmaker’s δ is constant across the various teams in the tournament (see Leitner et al., 2010, for all details), we obtain overrounds for all bookmakers with a median value of 17.3%.

To aggregate the overround-adjusted odds across the 19 bookmakers, we transform them to the log-odds (or logit) scale for averaging (as in Leitner et al., 2010). The bookmaker consensus is computed as the mean winning log-odds for each team across bookmakers and then transformed back to the winning probability scale.

In a second step the bookmakers’ odds are employed to infer the contenders’ relative abilities (or strengths). To do so, an “inverse” tournament simulation based on team-specific abilities is used. The idea is the following:

1. If team abilities are available, pairwise winning probabilities can be derived for each possible match using the classical Bradley and Terry (1952) model. This model is similar to the Elo rating (Elo, 2008), popular in sports, and computes the probability that a Team A beats a Team B by their associated abilities (or strengths):

$$\Pr(A \text{ beats } B) = \frac{\textit{ability}_A}{\textit{ability}_A + \textit{ability}_B}.$$

2. Given these pairwise winning probabilities, the whole tournament can be easily simulated to see which team proceeds to which stage in the tournament and which team finally wins⁹.

⁹By adopting the classical Bradley-Terry model, the simulation of each match yields only a winner and

- Such a tournament simulation can then be run sufficiently often (here 100,000 times) to obtain relative frequencies for each team to win the tournament.

Here, we use the iterative approach of Leitner et al. (2010) to find team abilities so that the resulting simulated winning probabilities (from 100,000 runs) closely match the bookmaker consensus probabilities. This allows to strip the effects of the tournament draw (with weaker/easier and stronger/more difficult groups), yielding a log-ability measure (on the log-odds scale) for each team.

3.5 Plus-minus player ratings

This section describes a method for calculating ratings of individual players based on the performance of their teams as a whole, known as *plus-minus (PM) ratings* (Hvattum, 2019). The starting point is match data specifying the starting line-ups together with information regarding substitutions, red cards, and goals scored, see Section 2.4 for more details.

The basic idea of adjusted PM ratings is to formulate a regression model where the dependent variable corresponds to the observed goal difference within a segment and the covariates include indicator variables for the presence of players. The regression coefficients of the player specific indicator variables can then be interpreted as player ratings. In a basic form, let y_i be the observed goal difference for segment i , taken from the perspective of the home team. Furthermore, let $x_{ij} = 1$ if player j appears on the pitch for the home team during segment i , $x_{ij} = -1$ if the player appears for the away team, and $x_{ij} = 0$ otherwise. Then, player ratings β_j can be obtained as the estimated regression coefficients of a simple linear regression:

$$y_i = \sum_j \beta_j x_{ij} + \varepsilon_i,$$

where ε_i is an error term. When applying PM ratings to football players, the above specification is too simplistic, and additional covariates and estimation tricks must be incorporated. We use here the ratings as described by Pantuso and Hvattum (2021), including the following adjustments:

- A country-specific home-field advantage is modelled by additional covariates.
- Red cards are handled by introducing additional covariates for each potential red card and by weighing the appearances of the remaining players.
- Players’ ratings are adjusted by a factor that depends on their age.
- Another adjustment of players’ ratings is made based on the set of league tournaments in which they have participated.
- Each observation is weighted based on three criteria: the time since the match was played, the duration of the corresponding segment, and the goal difference at the beginning and end of the segment.

a loser, without the possibility of a tie or any further information about the number of goals or the goal difference. This is sufficient for the knock-out stage of the tournament, as it reflects the fact that the actual matches always have a winner (if necessary through overtime and penalties). However, for the group phase within the simulation this approach might result in tied teams. If necessary, such ties are resolved through additional “fictitious” matches between the tied teams in order to obtain unique winners and runner-ups of the groups.

- The estimation of ratings is based on regularization in the form of ridge regression (see Hoerl and Kennard, 1970).
- For coefficients corresponding to player ratings, the regularization is adjusted ad hoc, under the assumption that a player is more likely to be of similar playing strength as the most common teammates, rather than an average player.

The resulting PM ratings have been examined in past studies. Gelade and Hvattum (2020) showed how the ratings are related to certain key-performance indicators based on event data, while Hvattum and Gelade (2021) compared them to an alternative rating system based on valuing individual player actions. Arntzen and Hvattum (2021) found that the PM ratings of players in the starting line-ups of teams contained relevant information for predicting match outcomes, in particular in combination with a team rating based on the Elo system. For this project, we calculated the average, median and “best-11-player” PM team rankings as well as the number of “missing important PM players” as introduced in Section 2.4. The first three variables are extremely correlated (> 0.98). During the tuning of our models we found that the *average PM player rating* was slightly outperforming the other two, so we only included it together with the missing player in the hybrid machine learning models, which we introduce in the next section.

3.6 Combine methods to hybrid machine learning models

In order to link the information provided by the covariate data, the historic match data, the bookmakers’ odds and the plus-minus player rating related data, we now combine the random forest approach from Section 3.1 and the extreme gradient boosting approach from Section 3.2 with the ranking methods from Sections 3.3-3.5. We propose to use the ranking approaches to generate new (highly informative) covariates that can be incorporated into the statistical models. For that purpose, we estimate current Poisson ranking team abilities r_i based on historic match data (see Section 3.3) as well as for the “# of missing players” (see Section 3.5).

Moreover, for each UEFA EURO we use the winning odds provided by the bookmakers and calculate the team log-abilities $s_i, i = 1, \dots, N$, of all $N \in \{16, 24\}$ participating teams shortly before the start of the respective tournament (see Section 3.4). This procedure gives us the estimates \hat{s}_i as an additional covariate covering the strength for all teams participating in a certain UEFA EURO. Actually, this variable turns out to be much more informative than e.g. the FIFA ranking, see Section 5.1.

Finally, based on historic match segment data, we estimate the (average) plus-minus team rankings $pm_i, i = 1, \dots, N$, of all participating teams shortly before the start of the respective tournament (see again Section 3.5). The corresponding estimates \widehat{pm}_i again serve as another additional covariate. Also this variable turns out to be rather relevant, see again Section 5.1. The newly generated variables can be added to the covariate data based on previous UEFA EUROs and a random forest, an xgboost model or actually any other statistical or machine learning model, such as e.g. a lasso-regularized regression model¹⁰, can be fitted to these data. Lasso regression was used for example in Groll et al. (2015) to predict the FIFA World Cup 2014. More details on how classical regression approaches can be used for the modeling and prediction of football matches can also be found in Groll et al. (2020). Based

¹⁰The idea of Lasso penalization was first introduced by Tibshirani (1996). Such a Lasso regression model can be easily tuned and fitted using the function `cv.glmnet` from the R-package `glmnet` (Friedman et al., 2010).

on these models, new matches (e.g., matches from an upcoming UEFA EURO tournament) can be predicted. Exemplarily for the random forest, a new observation is predicted by dropping down its covariate values from each of the B regression trees, resulting in B distinct predictions. The average of those is then used as a point estimate of the expected numbers of goals conditioning on the covariate values. Similarly, also for the *xgboost* and the Poisson Lasso regression model the covariate values of the new observation can be plugged into the sequential tree ensemble and the corresponding linear predictor, respectively. In order to be able to use the point estimates from the tree-based models for the prediction of the outcome of single matches or a whole tournament, we follow Groll et al. (2019a) and treat the predicted expected value for the number of goals as an estimate for the intensity λ of a Poisson distribution $Po(\lambda)$. For the Poisson lasso regression model, this is implicitly done anyway. This way we can randomly draw results for single matches and compute probabilities for the match outcomes *win*, *draw* and *loss* by using two independent Poisson distributions (conditional on the covariates) for both scores.

4 Model performance

In the following, we investigate the predictive performance of the proposed hybrid machine learning models and compare them also to a more conventional (regularized) Poisson regression approach. This method is also “hybrid” in the sense that it includes both the covariate data from Section 2.1 and the ranking variables from Sections 3.3-3.5. It links the feature information to the number of goals in a log-linear Poisson model. The model is estimated using a penalized likelihood approach. The details for this method can also be found in Groll et al. (2020).

Altogether, the following four approaches are now compared with regard to their predictive performance: two random forest implementations, *ranger* and *cforest*, the *xgboost* method and a conventional *lasso* Poisson regression model. For this purpose, we apply the following general procedure on the UEFA EURO 2004-2016 data for all methods:

1. Form a training data set containing three out of four UEFA EUROs.
2. Fit each of the methods to the training data.
3. Predict the left-out UEFA EURO using each of the prediction methods.
4. Iterate steps 1-3 such that each UEFA EURO is once the left-out one.
5. Compare predicted and real outcomes for all prediction methods.

This procedure ensures that each match from the total data set is once part of the test data and we obtain out-of-sample predictions for all matches. In step 5, several different performance measures for the quality of the predictions are investigated.

Let $\tilde{y}_i \in \{1, 2, 3\}$ be the true ordinal match outcomes for all $i = 1, \dots, N$ matches from the four considered UEFA EUROs. Additionally, let $\hat{\pi}_{1i}, \hat{\pi}_{2i}, \hat{\pi}_{3i}$, $i = 1, \dots, N$, be the predicted probabilities for the match outcomes obtained by one of the different methods mentioned above. These can be computed by assuming that the numbers of goals follow (conditionally) independent Poisson distributions, where the event rates λ_{1i} and λ_{2i} for the scores of match i are estimated by the respective predicted expected values. Let G_{1i} and G_{2i} denote the random variables representing the number of goals scored by two competing teams in match

i. Then, the probabilities $\hat{\pi}_{1i} = P(G_{1i} > G_{2i})$, $\hat{\pi}_{2i} = P(G_{1i} = G_{2i})$ and $\hat{\pi}_{3i} = P(G_{1i} < G_{2i})$, which are based on the corresponding Poisson distributions $G_{1i} \sim Po(\hat{\lambda}_{1i})$ and $G_{2i} \sim Po(\hat{\lambda}_{2i})$ with estimates $\hat{\lambda}_{1i}$ and $\hat{\lambda}_{2i}$, can be easily calculated via the Skellam distribution. For a short description of the Skellam distribution, see Appendix A. Based on these predicted probabilities, we use three different performance measures to compare the predictive power of the methods:

- the multinomial *likelihood*, which for a single match outcome is defined as $\hat{\pi}_{1i}^{\delta_{1\tilde{y}_i}} \hat{\pi}_{2i}^{\delta_{2\tilde{y}_i}} \hat{\pi}_{3i}^{\delta_{3\tilde{y}_i}}$, with $\delta_{r\tilde{y}_i}$ denoting Kronecker’s delta, which is defined in Appendix A. The multinomial likelihood reflects the probability of a correct prediction. Hence, a large value reflects a good fit.
- the *classification rate*, based on the indicator functions $\mathbb{1}(\tilde{y}_i = \arg \max_{r \in \{1,2,3\}} (\hat{\pi}_{ri}))$, indicating whether match *i* was correctly classified. Again, a large value of the classification rate reflects a good fit. However, note that along the lines of Gneiting and Raftery (2007) the classification rate does not constitute a proper scoring rule.
- the *rank probability score* (RPS), which, in contrast to both measures introduced above, explicitly accounts for the ordinal structure of the responses. For our purpose, it can be defined as $\frac{1}{3-1} \sum_{r=1}^{3-1} \left(\sum_{l=1}^r (\hat{\pi}_{li} - \delta_{l\tilde{y}_i}) \right)^2$. As the RPS is an error measure, here a low value represents a good fit.

Odds provided by bookmakers serve as a natural benchmark for these predictive performance measures. For this purpose, we collected the so-called “three-way” odds for (almost) all matches of the UEFA EUROs 2004-2016¹¹. By taking the three quantities $\tilde{\pi}_{ri} = 1/\text{odds}_{ri}$, $r \in \{1, 2, 3\}$, of a match *i* and by normalizing with $c_i := \sum_{r=1}^3 \tilde{\pi}_{ri}$ in order to adjust for the bookmaker’s margins, the odds can be directly transformed into probabilities using $\hat{\pi}_{ri} = \tilde{\pi}_{ri}/c_i$ ¹².

Table 6 displays the results for these (ordinal) performance measures for the four prediction methods as well as for the bookmakers, averaged over 144 matches from the four UEFA EUROs 2004-2016 (regarding the results in regular time, i.e. after 90 minutes and without possible extra time or penalty shootout). First of all, it turns out that all approaches achieve a slightly worse performance compared to our previous analysis on FIFA World Cups in Groll et al. (2019a).

The hybrid cforest slightly outperforms the other approaches with respect to the multinomial *likelihood*, closely followed by the xgboost approach, and is also (together with xgboost) the best regarding the *classification rate*, getting even close to the bookmakers as the natural benchmark. In terms of the RPS, all four methods are rather comparable with a slight advantage for the lasso Poisson regression model.

¹¹Three-way odds consider only the match tendency with possible results *victory team 1*, *draw* or *defeat team 1* and are usually fixed some days before the corresponding match takes place. This allows the bookmakers to incorporate current information (e.g., injuries of important players) into the odds during the run of a tournament. The three-way odds were obtained from the website <http://www.betexplorer.com/>.

¹²The transformed probabilities implicitly assume that the bookmaker’s margins are equally distributed on the three possible match tendencies.

Table 6: Comparison of the prediction methods for ordinal match outcomes *victory team 1*, *draw* or *defeat team 1* (regarding the results in regular time, i.e. after 90 minutes and without possible extra time or penalty shootout; best performing approach in bold font); additionally, the predictions based on the bookmakers’ odds are shown as a natural benchmark (bottom line).

	Likelihood	Class. Rate	RPS
ranger	0.372	0.458	0.216
cforest	0.382	0.486	0.213
xgboost	0.380	0.486	0.217
lasso	0.379	0.458	0.210
bookmakers	0.400	0.493	0.203

As the proposed methods can also be used to simulate the tournament course of an upcoming tournament (see also Section 5.2 for a prediction of the UEFA EURO 2020), we are also interested in the performance of the regarded methods with respect to the prediction of the exact number of goals. In order to identify the teams that qualify for the knockout stage, the precise final group standings need to be determined. To be able to do so, the precise results of the matches in the group stage play a crucial role¹³.

For this reason, we also evaluate the methods’ performances with regard to the quadratic error between the observed and predicted number of goals for each match and each team, as well as between the observed and predicted goal difference. Now let y_{ijk} , for $i, j = 1, \dots, n$ and $k \in \{2004, 2008, 2012, 2016\}$, denote the observed numbers of goals scored by team i against team j in tournament k and \hat{y}_{ijk} a corresponding predicted value, obtained by one of the compared methods. Then we calculate the two absolute errors $|y_{ijk} - \hat{y}_{ijk}|$ and $|(y_{ijk} - y_{jik}) - (\hat{y}_{ijk} - \hat{y}_{jik})|$ for all N matches of the four UEFA EUROs 2004-2016. Finally, per method we calculate (mean) absolute errors. Note that in this case the odds provided by the bookmakers cannot be used for comparison. Table 7 shows that actually all four methods yield rather similar results, with slight advantages for lasso.

Table 7: Comparison of the prediction methods for the exact number of goals and the goal difference based on mean absolute error (best performing approach in bold font).

	Goals	Goal Difference
ranger	0.862	1.176
cforest	0.862	1.166
xgboost	0.883	1.162
lasso	0.846	1.148

Altogether, based on these results and our positive experiences in earlier research projects (Groll et al., 2019a,b), we assess the hybrid cforest method to be the best and most reliable choice for forecasting the upcoming UEFA EURO 2020 tournament.

But note that also the xgboost approach seems to be principally very promising for modeling and predicting football matches. However, we experienced some higher sensitivity and

¹³The final group standings are determined by (1) the number of points, (2) the goal difference and (3) the number of scored goals. If several teams coincide with respect to all of these three criteria, a separate chart is calculated based on the matches between the coinciding teams only. Here, again the final standing of the teams is determined following criteria (1)–(3). If still no distinct decision can be taken, the decision is induced by lot.

instability during the (more sophisticated) tuning process compared to the other methods, which we believe is mostly due to small sample size of our training data. In particular, as we had to always exclude one full UEFA EURO from the training data in our leave-one-tournament-out strategy in order to assess the models' performances on external validation data, we believe that the full potential of xgboost is not yet fully exploited in this competition.

5 Modeling the UEFA EURO 2020

We now fit the proposed hybrid cforest model to the full UEFA EURO 2004-2016 data. Next, we calculate the Poisson ranking ability parameters based on historic match data over the 8 years preceding the UEFA EURO 2020, as well as the bookmaker consensus abilities based on the winning odds from 19 different bookmakers, and the average PM player ratings as well as the number of important PM players missing in the squad. Based on conventional covariate data and those additional ability and rating variables, the fitted cforest model will be used to simulate the UEFA EURO 2020 tournament 100,000 times to determine winning probabilities for all 24 participating teams.

5.1 Fitting the hybrid cforest to the UEFA EUROs 2004-2016 data

We next fit the hybrid cforest to the complete data set covering the four UEFA EUROs 2004-2016. As suggested in regression settings, the optimal number of input variables randomly sampled as candidates at each node is set to $m_{\text{try}} = \lfloor \sqrt{p} \rfloor = 4$. The best way to understand the role of the single predictor variables in a complicated, blackbox-type machine learning model such as the cforest is the so-called variable importance (Breiman, 2001). Typically, the variable importance of a predictor is measured by permuting each of the predictors separately in the out-of-bag observations of each tree. Out-of-bag observations are observations which are not part of the respective subsample or bootstrap sample that is used to fit a tree. Permuting a variable means that within the variable each value is randomly assigned to a location within the vector. If, for example, *FIFA.rank* is permuted, the *FIFA.rank* of the German team in 2004 could be assigned to the *FIFA.rank* of the Spanish team in 2016. When permuting variables randomly, they lose their information with respect to the response variable (if they have any). Then, one measures the loss of prediction accuracy compared to the case where the variable is not permuted. Permuting variables with a high importance will lead to a higher loss of prediction accuracy than permuting values with low importance. Figure 1 shows bar plots of the variable importance values for all variables in the hybrid cforest applied to the data of the UEFA EUROs 2004-2016. Interestingly, the market value is the most important predictor in the cforest model, followed by the abilities from the bookmaker consensus approach. But also the *number of CL players*, the *average PM player rating* and the *UEFA points* seem to be more informative compared e.g. to the *FIFA rank*. Besides those variables, also the *Poisson ranking abilities* and the *GDP* contain relevant information concerning the current strengths of the teams. Hence, it is definitely worth the effort to estimate such abilities in separate statistical models. For a more detailed comparison of the team abilities and the *FIFA rank*, see Table 8.

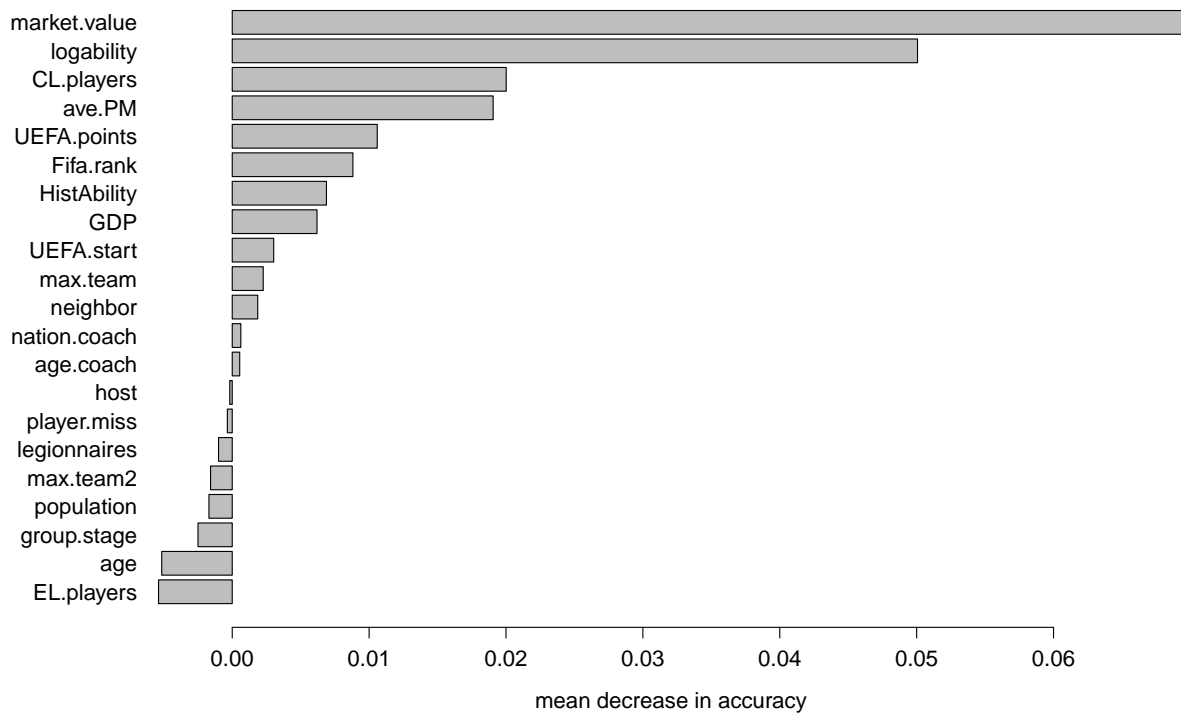


Figure 1: Bar plot displaying the variable importance in the hybrid cforest model applied to UEFA EURO 2004-2016 data.

5.2 Probabilities for UEFA EURO 2020 winner

In this section, the hybrid cforest model is applied to (new) data for the UEFA EURO 2020 (in advance of the tournament) to predict winning probabilities for all teams and to predict the tournament course.

The Poisson abilities were estimated by a bivariate Poisson model with a half period of 3 years (i.e. 1095 days). All matches of the 282 national teams played since 2003-05-27 up to 2021-05-26 are used for the estimation, what results in a total of 6953 matches. All further predictor variables are taken as the latest values shortly before the UEFA EURO (and using the final squads of 26 players for all nations). The bookmaker consensus abilities are based on the average odds of 19 bookmakers.

Note that due to the fact that the upcoming UEFA EURO tournament will not have a single hosting country, but the matches are distributed all over Europe, we believe that a potential home effect will be much less pronounced compared to earlier tournaments, as also the average travel distances for teams and fans should decrease. Moreover, due to the ongoing COVID-19 pandemic, the numbers of spectators allowed are substantially reduced compared to earlier tournaments. For these reasons, we decided to set the *home* and *neighbor* dummy variables for all 24 teams equal to zero.

For each match in the UEFA EURO 2020, the hybrid cforest 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 36 matches from the group stage can be simulated and final group

Table 8: Ranking of the participants of the UEFA EURO 2020 according to estimated historic match abilities, bookmaker consensus abilities, average PM player rating and FIFA ranking.

























	Historic match abilities	Bookmaker consensus abilities	PM player rating	FIFA ranking
1	Belgium	England	Germany	Belgium
2	Spain	France	England	France
3	England	Belgium	France	England
4	France	Germany	Spain	Portugal
5	Portugal	Portugal	Belgium	Spain
6	Germany	Spain	Portugal	Italy
7	Netherlands	Italy	Netherlands	Denmark
8	Italy	Netherlands	Italy	Germany
9	Denmark	Denmark	Switzerland	Switzerland
10	Switzerland	Croatia	Denmark	Croatia
11	Poland	Turkey	Austria	Netherlands
12	Croatia	Switzerland	Croatia	Wales
13	Sweden	Russia	Turkey	Sweden
14	Russia	Sweden	Ukraine	Poland
15	Wales	Poland	Czech Republic	Austria
16	Austria	Ukraine	Poland	Ukraine
17	Turkey	Austria	Sweden	Turkey
18	Ukraine	Czech Republic	Russia	Slovakia
19	Czech Republic	Wales	Wales	Hungary
20	Slovakia	Hungary	Scotland	Russia
21	Scotland	Scotland	Slovakia	Czech Republic
22	Finland	Finland	Hungary	Scotland
23	Hungary	Slovakia	Finland	Finland
24	North Macedonia	North Macedonia	North Macedonia	North Macedonia

standings can be calculated. Due to the fact that the full score-lines are simulated, we can precisely follow the official UEFA rules when determining the final group standings (see again Footnote 13). 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 $1/3$ 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 24 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 (average) winning probabilities based on 19 different bookmakers for comparison.

We can see that, according to our hybrid cforest model, the current FIFA World champion France is the favored team with a predicted winning probability of 14.8% followed by England, Spain, Portugal and Germany. Overall, this result seems mostly in line with the probabilities from the bookmakers, as we can see in the last column. However, e.g. for

Table 9: Estimated probabilities (in %) for reaching the different stages in the UEFA EURO 2020 for all 24 teams based on 100,000 simulation runs of the UEFA EURO 2020 together with (average) winning probabilities based on the odds of 19 bookmakers.

			Round of 16	Quarter finals	Semi finals	Final	European Champion	Bookmakers consensus
1.		FRA	89.7	59.5	39.7	25.0	14.8	15.0
2.		ENG	94.6	56.3	35.3	22.9	13.5	14.8
3.		ESP	94.0	66.8	35.4	21.9	12.3	9.9
4.		POR	85.3	52.3	31.7	18.6	10.1	9.0
5.		GER	85.3	52.3	32.5	18.8	10.1	9.6
6.		BEL	91.5	54.6	32.2	16.2	8.3	12.1
7.		ITA	88.8	56.6	32.2	15.9	7.9	7.5
8.		NED	93.4	52.1	28.3	13.0	6.1	6.5
9.		DEN	84.5	44.4	23.2	10.2	4.6	2.9
10.		CRO	78.0	36.8	16.3	7.4	3.1	2.4
11.		SUI	72.3	34.7	15.3	5.7	2.2	1.1
12.		AUT	80.9	33.2	13.5	4.6	1.5	0.8
13.		POL	66.2	29.8	10.2	3.9	1.2	1.1
14.		SWE	59.8	25.6	8.7	3.2	1.0	0.9
15.		TUR	53.3	20.8	7.8	2.4	0.7	1.6
16.		WAL	53.7	20.6	7.4	2.1	0.6	0.6
17.		SCO	49.8	19.3	6.3	2.0	0.6	0.4
18.		RUS	52.0	16.8	5.3	1.4	0.4	0.9
19.		CZE	40.8	14.4	4.4	1.3	0.3	0.6
20.		UKR	57.4	16.9	5.1	1.3	0.3	1.0
21.		SVK	44.9	16.5	4.6	1.3	0.3	0.3
22.		FIN	37.1	9.3	2.3	0.5	0.1	0.2
23.		NMD	32.9	7.1	1.6	0.3	0.1	0.2
24.		HUN	13.9	3.4	1.0	0.2	0.0	0.2

Belgium, which the bookmakers rate on place three with a winning chance of 12.1%, the cforest model calculates substantially lower chances (8.3%). Beside the probabilities of becoming European champion, Table 9 provides some further interesting insights also for the single stages within the tournament. For example, it is interesting to see that while Belgium has a much higher probability to reach the round of 16 than Germany (91.5% vs. 85.3%), Germany has higher chances to reach the final (16.2% vs. 18.8%) and become European champion (8.3% vs. 10.1%). This is probably related to the fact that German is assigned to a very tough group with France and Portugal. It is also interesting that some teams have a fair chance to reach the round of 16, but then their chances decrease drastically to reach the next round. Ukraine, for example, has a moderate chance of 57.4% to reach the knockout stage, but only a rather low chance of 16.9% to reach the quarter finals.

The lowest chances are given to Hungary, though they do not have the weakest squad (see Table 8). The reason is their very low chance to attain the Round of 16 because of their bad luck with the draw: they are in the strongest group with France, Germany and Portugal

6 Concluding remarks

In this work, we proposed hybrid modeling approaches for the scores of international football matches which combine random forests, an extreme gradient boosting approach and lasso-penalized Poisson regression with several different ranking methods, namely a current ability ranking based on historic matches, abilities based on bookmakers' odds and plus-minus player rankings. While the machine learning models are principally based on the competing teams' covariate information, the latter components provide ability parameters, which serve as adequate estimates of the current team strengths as well as of the information contained in the bookmakers' odds. In order to combine the methods, the current ability and PM player ranking methods need to be repeatedly applied to historical match data preceding each UEFA EURO from the training data. This way, for each UEFA EURO in the training data and each participating team ability estimates are obtained. Similarly, the bookmaker consensus abilities are obtained by inverse tournament simulation based on the aggregated winning odds from several online bookmakers. These rankings and ability estimates can be added as additional covariates to the set of covariates used in the machine learning procedures. We compared the predictive performances of the approaches in a leave-one-tournament-out competition on the data of the four preceding UEFA EUROs 2004-2016 and found the highest potential in the cforest model, closely followed by the xgboost model. Additionally, based on the estimates of the hybrid xgboost model on the training data, we repeatedly simulated the upcoming UEFA EURO 2020 100,000 times. According to the simulations, the current World champion France with a winning probability of 14.8% is the top favorite for winning the title, followed by England (13.5%) and Spain (12.3%). Furthermore, survival probabilities for all teams and at all tournament stages are provided. Even though in our prediction competition in Section 4 the *extreme gradient boosting* technique was not yet able to substantially outperform the other approaches, though being well-known in the machine learning community for its high predictive power, we believe that this is partly due to the small sample size of our training data of the four UEFA EUROs 2004-2016 (and in the competition the training data size even decreased due to the leave-one-tournament-out strategy). Compared to the other methods, xgboost involves a more sophisticated tuning process with several tuning parameters and, hence, to fully exploit its high potential the more training data are available the better. On our small training data sets, however, we experienced a rather high sensitiveness and instability during the tuning process. Nevertheless, we think that our results already indicate that the method is very promising for modeling and predicting football matches. We are looking forward with excitement to comparing the cforest and xgboost approach again in an extensive ex-post analysis, when the UEFA EURO 2020 is finished.

Acknowledgment

We thank Jonas Heiner for his tremendous effort in helping us to collect the covariate data.

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:

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

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

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

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 Additional data material

	FRA	ITA	NED	POR	ESP	GER	ENG	CZE
Oddset	3.25	5	5.5	6	6.5	7	7	7
	SWE	DEN	RUS	GRE	CRO	BUL	SUI	LVA
Oddset	20	20	40	45	45	60	60	100

Table 10: Quoted odds from ODDSET for the 16 teams in the EURO 2004.

	FRA	ENG	BEL	ESP	GER	POR	ITA	NED
bwin	5.5	6.0	7.50	9.0	8.0	9.00	11.00	13.0
bet365	5.5	6.0	7.00	8.5	8.0	9.00	12.00	13.0
Sky Bet	6.0	6.0	7.00	9.0	9.0	9.00	12.00	12.0
Paddy Power	6.0	5.0	7.50	8.0	9.0	11.00	9.00	13.0
William Hill	5.5	6.0	7.00	8.5	9.0	9.00	12.00	13.0
Betfair Sportsbook	6.0	5.0	7.50	8.0	9.0	11.00	9.00	13.0
Bet Victor	5.5	5.5	7.00	9.0	10.0	9.00	11.00	13.0
Unibet	6.0	6.5	7.00	10.0	10.0	9.00	13.00	15.0
Mansion Bet	6.0	6.0	7.00	8.5	8.0	10.00	12.00	12.0
Smarkets Sportsbook	6.0	6.0	7.60	9.8	10.0	10.00	12.50	15.5
Betway	6.0	6.0	7.00	8.5	9.0	10.00	12.00	13.0
Boylesports	5.5	6.0	7.00	8.5	8.5	9.00	12.00	13.0
10Bet	6.0	6.0	7.00	8.5	8.5	9.00	12.00	12.0
Sport Nation	5.5	5.5	7.00	8.5	10.0	8.00	11.00	12.0
Vbet	5.5	5.9	6.80	8.5	8.0	8.80	12.00	13.0
Sporting Index	6.0	6.0	6.50	8.0	8.5	10.00	12.00	13.0
RedZone	6.0	6.0	7.75	8.5	8.5	8.75	10.75	12.0
Spreadex	5.5	5.5	6.50	8.5	9.0	10.00	10.00	13.0
Smarkets	5.8	6.0	7.80	9.6	9.8	10.80	12.20	15.2
	DEN	CRO	TUR	SUI	POL	UKR	SWE	RUS
bwin	29	34	41	81	81	81	101	81
bet365	29	34	51	67	81	51	101	67
Sky Bet	34	29	51	67	81	67	81	51
Paddy Power	26	31	67	91	67	91	91	76
William Hill	29	34	51	81	81	101	81	101
Betfair Sportsbook	26	31	67	91	67	91	91	76
Bet Victor	26	41	51	81	81	126	81	126
Unibet	34	41	61	71	81	101	101	101
Mansion Bet	31	31	51	67	67	81	67	101
Smarkets Sportsbook	34	44	65	90	100	100	120	240
Betway	29	34	51	81	81	67	101	80
Boylesports	29	34	51	67	81	81	81	81
10Bet	31	34	51	67	81	81	81	81
Sport Nation	29	34	51	67	81	81	81	81
Vbet	29	34	51	67	81	51	101	67
Sporting Index	29	34	61	67	67	81	101	101
RedZone	29	34	34	67	81	81	81	81
Spreadex	29	29	51	51	81	81	81	126
Smarkets	33	43	64	88	98	98	118	235
	AUT	CZE	WAL	SCO	SVK	FIN	HUN	MKD
bwin	101	151	101	151	201	201	201	201
bet365	81	151	201	251	251	501	401	501
Sky Bet	101	151	101	151	151	151	251	501
Paddy Power	91	126	126	251	326	326	326	501
William Hill	126	126	126	251	251	251	401	501
Betfair Sportsbook	91	126	126	251	326	326	326	501
Bet Victor	126	126	126	251	501	501	501	501
Unibet	201	151	101	301	301	501	801	501
Mansion Bet	101	101	126	201	301	401	401	501
Smarkets Sportsbook	200	190	260	300	300	500	500	500
Betway	80	101	151	251	501	501	501	501
Boylesports	101	101	101	201	251	251	401	501
10Bet	81	126	151	201	201	401	401	501
Sport Nation	81	126	151	201	201	401	401	501
Vbet	81	151	201	251	251	501	401	501
Sporting Index	101	151	151	251	251	301	301	501
RedZone	81	126	151	201	201	401	401	501
Spreadex	151	81	151	251	501	401	501	401
Smarkets	196	186	255	294	294	490	490	490

Table 11: Quoted odds from 19 online bookmakers for the 24 teams in the EURO 2020 obtained on 2021-05-31 from <https://www.oddschecker.com/> and <https://www.bwin.com/>, respectively.

References

Arntzen, H. and Hvattum, L. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*. In press.

- Boshnakov, G., Kharrat, T., and McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466.
- Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, J. C. (1984). *Classification and Regression Trees*. Wadsworth, Monterey, CA.
- Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22:477–522.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2021). *xgboost: Extreme Gradient Boosting*. R package version 1.3.2.1.
- Dixon, M. J. and 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.
- Dyte, D. and Clarke, S. R. (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society*, 51 (8):993–998.
- Elo, A. E. (2008). *The Rating of Chess Players, Past and Present*. Ishi Press, San Rafael.
- Forrest, D., Goddard, J., and Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21(3):551–564.
- Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, San Francisco, CA.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:337–407.
- Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407.
- Gelade, G. and Hvattum, L. (2020). On the relationship between $+/-$ ratings and event-level performance statistics. *Journal of Sports Analytics*, 6:85–97.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

- Groll, A. and Abedieh, J. (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.
- Groll, A., Kneib, T., Mayr, A., and Schauburger, G. (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.
- Groll, A., Ley, C., Schauburger, G., and Van Eetvelde, H. (2019a). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15:271–287.
- Groll, A., Ley, C., Schauburger, G., Van Eetvelde, H., and Zeileis, A. (2019b). Hybrid machine learning forecasts for the FIFA Women’s World Cup 2019. *arXiv preprint arXiv:1906.01131*.
- Groll, A., Schauburger, G., and Tutz, G. (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.
- Groll, A., Schauburger, G., and Van Eetvelde, H. (2020). Ranking and prediction models. In Ley, C. and Dominicy, Y., editors, *Science Meets Sports: When Statistics Are More Than Numbers*, pages 95–122. Cambridge Scholars Publishing.
- Henery, R. J. (1999). Measures of over-round in performance index betting. *Journal of the Royal Statistical Society D*, 48(3):435–439.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.
- Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and van der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7:355–373.
- Hvattum, L. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, 18:1–23.
- Hvattum, L. and Gelade, G. (2021). Comparing bottom-up and top-down ratings for individual soccer players. *International Journal of Computer Science in Sport*, 20:23–42.
- Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393.
- Leitner, C., Zeileis, A., and Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26 (3):471–481.
- Ley, C., Van de Wiele, T., and Van Eetvelde, H. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, 19(1):55–77.
- Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014). The evolution of boosting algorithms - from machine learning to statistical modelling. *Methods of Information in Medicine*, 53(6):419–427.

- McHale, I. and Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, 61(4):432–445.
- McHale, I. G. and Scarf, P. A. (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, 41(3):219–236.
- Pantuso, G. and Hvattum, L. (2021). Maximizing performance with an eye on the finances: a chance-constrained model for football transfer market decisions. *TOP*. In press.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5:197–227.
- Schauberger, G. and Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5-6):460–482.
- 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(Pt 3):296–296.
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, B* 58:267–288.
- van der Wurp, H., Groll, A., Kneib, T., Marra, G., and Radice, R. (2020). Generalised joint regression for count data: a penalty extension for competitive settings. *Statistics and Computing*, 30(5):1419–1432.
- Wikipedia (2019). Odds — wikipedia, the free encyclopedia. Online, accessed 2019-05-24.
- Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.
- Wunderlich, F., Weigelt, M., Rein, R., and Memmert, D. (2021). How does spectator presence affect football? home advantage remains in european top-class football matches played without spectators during the covid-19 pandemic. *Plos one*, 16(3):e0248590.
- Zeileis, A., Leitner, C., and Hornik, K. (2012). History repeating: Spain beats Germany in the EURO 2012 final. Working Paper 2012-09, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck.
- Zeileis, A., Leitner, C., and Hornik, K. (2016). Predictive bookmaker consensus model for the UEFA Euro 2016. Working Paper 2016-15, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck.