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

Andreas Groll

Department of Statistics,
TU Dortmund University

(with J. Abedieh, L.M. Hvaatum, C. Ley, A. Mayr, T. Kneib,
G. Schauberger, G. Tutz, H. Van Eetvelde & A. Zeileis)

Research Seminar Summer Term 2023
June 16^{th} 2023

Who will celebrate?

Sources: youtube.com,EMAJ Magazine,youfrisky.com,Bailiwick Express

Who will cry?

Sources: youtube.com,pinterest,BBC,Daily Mail

Before the tournament starts:

Before the tournament starts:

Before the tournament starts:

ODDSET
DIE SPORTWETTE
VON LOTTO®

Sources: dfb.de, kicktipp.de

kicktipp
WM

Before the tournament starts:

ODDSET
DIE SPORTWETTE
VON LOTTO®

kicktipp
WM

Sources: dfb.de, kicktipp.de

Sources: duda.news, welt.de

How can the prediction of a major football tournament be done a bit more sophisticated?

Theoretical Background

Theoretical Background

For a general summary, see, for example:

Groll, A. and G. Schauberger (2019). Prediction of Soccer Matches. *Wiley StatsRef: Statistics Reference Online*, 1-7.

Part I: Regression-based Methods

Model for international football tournaments

$$y_{ijk}|\boldsymbol{x}_{ik},\boldsymbol{x}_{jk} \sim Pois(\lambda_{ijk}) \quad i,j \in \{1,\ldots,n\},\ i \neq j$$

$$\lambda_{ijk} = \exp\left(\beta_0 + (\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})^\top \boldsymbol{\beta}\right)$$

n: Number of teams

y_{ijk}: Number of goals scored by team i against opponent j at tournament k

$\boldsymbol{x}_{ik}$, $\boldsymbol{x}_{jk}$: Covariate vectors of team i and opponent j varying over tournaments

$\boldsymbol{\beta}$: Parameter vector of covariate effects

Regularized estimation

Maximize penalized log-likelihood

$$l_p(\beta_0, \boldsymbol{\beta}) = l(\beta_0, \boldsymbol{\beta}) - \xi J(\boldsymbol{\beta})$$

Regularized estimation

Maximize penalized log-likelihood

$$
\begin{aligned}
l_p(\beta_0, \boldsymbol{\beta}) &= l(\beta_0, \boldsymbol{\beta}) - \xi J(\boldsymbol{\beta}) \\
&= l(\beta_0, \boldsymbol{\beta}) - \xi \sum_{i=1}^{p} |\beta_i| ,
\end{aligned}
$$

with lasso penalty term (Tibshirani, 1996):

$$
J(\boldsymbol{\beta}) = \sum_{i=1}^{p} |\beta_i| .
$$

The model can be estimated with the `R`-package `glmnet` (Friedman et al., 2010).

Regularized estimation

Maximize penalized log-likelihood

$$
\begin{aligned}
l_p(\beta_0, \boldsymbol{\beta}) &= l(\beta_0, \boldsymbol{\beta}) - \xi J(\boldsymbol{\beta}) \\
&= l(\beta_0, \boldsymbol{\beta}) - \xi \sum_{i=1}^{p} |\beta_i| ,
\end{aligned}
$$

with lasso penalty term (Tibshirani, 1996):

$$
J(\boldsymbol{\beta}) = \sum_{i=1}^{p} |\beta_i| .
$$

The model can be estimated with the `R`-package `glmnet` (Friedman et al., 2010).

Versions used for: EURO 2012 (Groll and Abedieh, 2013); World Cup 2014 (Groll et al., 2015); EURO 2016 (Groll et al., 2018)

Part II: Ranking Methods

Independent Poisson ranking model

$$
\begin{aligned}
Y_{ijm} &\sim Pois(\lambda_{ijm}), \\
\lambda_{ijm} &= \exp\Big(\beta_0 + (r_i - r_j) + h \cdot \mathbb{1}(\text{team } i \text{ playing at home})\Big)
\end{aligned}
$$

n: Number of teams

M: Number of matches

y_{ijm}: Number of goals scored by team i against opponent j in match m

r_i, r_j: strengths / ability parameters of team i and team j

h: home effect; added if team i plays at home

Independent Poisson ranking model

Likelihood function:

$$L = \prod_{m=1}^{M} \left(\frac{\lambda_{ijm}^{y_{ijm}}}{y_{ijm}!} \exp(-\lambda_{ijm}) \cdot \frac{\lambda_{jim}^{y_{jim}}}{y_{jim}!} \exp(-\lambda_{jim}) \right)^{w_{type,m} \cdot w_{time,m}} ,$$

with weights

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

and

$$w_{type,m} \in \{1, 2.5, 3, 4\} \qquad \text{(depending on type of match)}.$$

Independent Poisson ranking model

Likelihood function:

$$L = \prod_{m=1}^{M} \left(\frac{\lambda_{ijm}^{y_{ijm}}}{y_{ijm}!} \exp(-\lambda_{ijm}) \cdot \frac{\lambda_{jim}^{y_{jim}}}{y_{jim}!} \exp(-\lambda_{jim}) \right)^{w_{type,m} \cdot w_{time,m}} ,$$

with weights

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

and

$$w_{type,m} \in \{1, 2.5, 3, 4\} \qquad \text{(depending on type of match)}\,.$$

Different extensions, for example, **bivariate Poisson models**. Ley et al. (2018) show that bivariate Poisson with Half Period of 3 years is best for prediction.

Part III: Random Forests

Random Forests

- introduced by Breiman (2001)
- **principle**: aggregation of (large) number of **classification / regression trees**

  $\Longrightarrow$ can be used both for classification & regression purposes

Random Forests

- introduced by Breiman (2001)
- **principle**: aggregation of (large) number of **classification / regression trees**

  $\Longrightarrow$ can be used both for classification & regression purposes
- **final predictions**: single tree predictions are aggregated, either by majority vote (classification) or by averaging (regression)

Random Forests

- introduced by Breiman (2001)
- **principle**: aggregation of (large) number of **classification / regression trees**

  $\Longrightarrow$ can be used both for classification & regression purposes
- **final predictions**: single tree predictions are aggregated, either by majority vote (classification) or by averaging (regression)
- feature space is partitioned recursively, each partition has its own prediction

Random Forests

- introduced by Breiman (2001)
- **principle**: aggregation of (large) number of **classification / regression trees**

  $\Longrightarrow$ can be used both for classification & regression purposes
- **final predictions**: single tree predictions are aggregated, either by majority vote (classification) or by averaging (regression)
- feature space is partitioned recursively, each partition has its own prediction
- find split with strongest difference between the two new partitions w.r.t. some criterion

Random Forests

- introduced by Breiman (2001)
- **principle**: aggregation of (large) number of **classification / regression trees**

  $\Longrightarrow$ can be used both for classification & regression purposes
- **final predictions**: single tree predictions are aggregated, either by majority vote (classification) or by averaging (regression)
- feature space is partitioned recursively, each partition has its own prediction
- find split with strongest difference between the two new partitions w.r.t. some criterion
- Observations within the same partition as similar as possible, observations from different partitions very different (w.r.t. response variable)

Random Forests

- introduced by Breiman (2001)
- **principle**: aggregation of (large) number of **classification / regression trees**

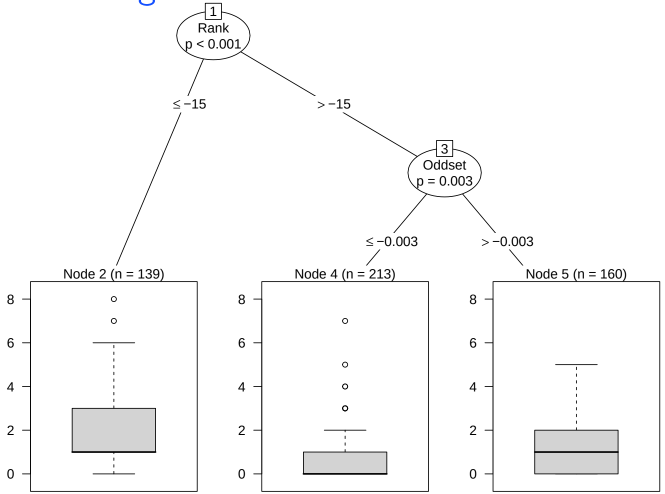
  $\Longrightarrow$ can be used both for classification & regression purposes
- **final predictions**: single tree predictions are aggregated, either by majority vote (classification) or by averaging (regression)
- feature space is partitioned recursively, each partition has its own prediction
- find split with strongest difference between the two new partitions w.r.t. some criterion
- Observations within the same partition as similar as possible, observations from different partitions very different (w.r.t. response variable)
- a single tree is usually pruned (lower variance but increases bias)

Random Forests

- introduced by Breiman (2001)
- **principle**: aggregation of (large) number of **classification / regression trees**

  $\Longrightarrow$ can be used both for classification & regression purposes
- **final predictions**: single tree predictions are aggregated, either by majority vote (classification) or by averaging (regression)
- feature space is partitioned recursively, each partition has its own prediction
- find split with strongest difference between the two new partitions w.r.t. some criterion
- Observations within the same partition as similar as possible, observations from different partitions very different (w.r.t. response variable)
- a single tree is usually pruned (lower variance but increases bias)
- visualized in dendrogram

Dendrogram of regression tree



Exemplary regression tree for FIFA World Cup 2002 – 2014 data using the function `ctree` from the `R`-package `party` (Hothorn et al., 2006). **Response**: *Number of goals*; **predictors**: only *FIFA Rank and Oddset* are used.

Random Forests

- repeatedly grow different regression trees
- main goal: decrease variance

Random Forests

- repeatedly grow different regression trees
- main goal: decrease variance $\Longrightarrow$ decrease correlation between single trees.

Random Forests

- repeatedly grow different regression trees
- main goal: decrease variance $\Longrightarrow$ decrease correlation between single trees.
- $\Longrightarrow$ two different randomisation steps:

  1) trees are not applied to the original sample but to **bootstrap samples** or random subsamples of the data.

  2) at each node a **(random) subset of the predictors** is drawn that are used to find the best split.

Random Forests

- repeatedly grow different regression trees
- main goal: decrease variance $\Longrightarrow$ decrease correlation between single trees.
- $\Longrightarrow$ two different randomisation steps:
  1) trees are not applied to the original sample but to **bootstrap samples** or random subsamples of the data.
  2) at each node a **(random) subset of the predictors** is drawn that are used to find the best split.
- by de-correlating and combining many trees $\Longrightarrow$ predictions with **low bias and reduced variance**

Random Forests for football

- response: metric variable *Number of Goals*
- predefined number of trees B (e.g., $B = 5000$) is fitted based on (bootstrap samples of) the training data
- prediction of new observation: covariate values are dropped down each of the regression trees, resulting in B predictions $\Longrightarrow$ average
- use predicted expected value as event rate $\hat{\lambda}$ of a Poisson distribution $Po(\lambda)$

Random Forests for football

- response: metric variable *Number of Goals*
- predefined number of trees B (e.g., $B = 5000$) is fitted based on (bootstrap samples of) the training data
- prediction of new observation: covariate values are dropped down each of the regression trees, resulting in B predictions $\Longrightarrow$ average
- use predicted expected value as event rate $\hat{\lambda}$ of a Poisson distribution $Po(\lambda)$
- 2 slightly different variants:
  1) classical RF algorithm proposed by Breiman (2001) from the `R`-package `ranger` (Wright and Ziegler, 2017)
  2) RFs based conditional inference trees: `cforest` from the `party` package (Hothorn et al., 2006)

Application to FIFA World Cups

Covariates

Data basis: World Cups 2002–2014

Covariates

Data basis: World Cups 2002–2014

- **Economic Factors:**
  GDP per capita, population

Covariates

Data basis: World Cups 2002–2014

- **Economic Factors:**
  GDP per capita, population

- **Sportive Factors:**
  Winning probs by bookmakers (Oddset), FIFA rank

Covariates

Data basis: World Cups 2002–2014

- **Economic Factors:**
  GDP per capita, population

- **Sportive Factors:**
  Winning probs by bookmakers (Oddset), FIFA rank

- **Home advantage:**
  host of the world cup, same continent as host, continent

Covariates

Data basis: World Cups 2002–2014

- **Economic Factors:**
  GDP per capita, population
- **Sportive Factors:**
  Winning probs by bookmakers (Oddset), FIFA rank
- **Home advantage:**
  host of the world cup, same continent as host, continent
- **Factors describing the team's structure**
  (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad

Covariates

Data basis: World Cups 2002–2014

- **Economic Factors:**
  GDP per capita, population
- **Sportive Factors:**
  Winning probs by bookmakers (Oddset), FIFA rank
- **Home advantage:**
  host of the world cup, same continent as host, continent
- **Factors describing the team's structure**
  (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad
- **Factors describing the team's coach**
  age, nationality, tenure

Covariates

Data basis: World Cups 2002–2014

- **Economic Factors:**
  GDP per capita, population

- **Sportive Factors:**
  Winning probs by bookmakers (Oddset), FIFA rank

- **Home advantage:**
  host of the world cup, same continent as host, continent

- **Factors describing the team's structure**
  (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad

- **Factors describing the team's coach**
  age, nationality, tenure

All variables are incorporated as differences between the team whose goals are considered and its opponent!

Extract of the design matrix

FRA		0:0		URU
URU		1:2		DEN

Team	Age	Rank	Oddset	...
France	28.3	1	0.149	...
Uruguay	25.3	24	0.009	...
Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋱

Extract of the design matrix

FRA 0:0 URU
URU 1:2 DEN

Team	Age	Rank	Oddset	...
France	28.3	1	0.149	...
Uruguay	25.3	24	0.009	...
Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋱

Goals	Team	Opponent	Age	Rank	Oddset	...
0	France	Uruguay	3.00	-23	0.140	...
0	Uruguay	France	-3.00	23	-0.140	...
1	Uruguay	Denmark	-2.10	4	-0.003	...
2	Denmark	Uruguay	2.10	-4	0.003	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Comparison of predictive performance: WC 2002-2014 data

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

Comparison of predictive performance: WC 2002-2014 data

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

We combine both the random forest and the LASSO with the ability estimates from the ranking method, calling those hybrid models!

Prediction of match outcomes

- true ordinal match outcomes: $\tilde{y}_1, \dots, \tilde{y}_N$ with $\tilde{y}_i \in \{1,2,3\}$, for all matches N from the 4 World Cups.
- predicted probabilities $\hat{\pi}_{1i}, \hat{\pi}_{2i}, \hat{\pi}_{3i}, \; i = 1, \dots, N$,
- Let G_{1i} and G_{2i} denote the goals scored by 2 competing teams in match i

  $\Longrightarrow$ compute $\hat{\pi}_{1i} = P(G_{1i} > G_{2i}), \hat{\pi}_{2i} = P(G_{1i} = G_{2i})$ and $\hat{\pi}_{3i} = P(G_{1i} < G_{2i})$ 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 $\hat{\lambda}_{2i}$ (Skellam distribution)
- **benchmark**: **bookmakers** $\Longrightarrow$ compute the 3 quantities $\tilde{\pi}_{ri} = 1/\text{odds}_r$, $r \in \{1,2,3\}$, normalize with $c_i := \sum_{r=1}^{3} \tilde{\pi}_{ri}$ (adjust for bookmakers' margins)

  $\Longrightarrow$ estimated probabilities $\hat{\pi}_{ri} = \tilde{\pi}_{ri}/c_i$

Prediction of match outcomes

3 Performance measures:

(a) **multinomial *likelihood*** (probability of correct prediction): for single match 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 δ_{ri} denoting Kronecker's delta

(b) **classification rate**: is match i correctly classified using the indicator function

$$\mathbb{1}(\tilde{y}_i = \underset{r \in \{1,2,3\}}{\arg\max} \, (\hat{\pi}_{ri}))$$

(c) **rank probability score** (RPS; explicitly accounts for the ordinal structure):

$$\frac{1}{3-1} \sum_{r=1}^{3-1} \left(\sum_{l=1}^{r} \hat{\pi}_{li} - \delta_{l\tilde{y}_i} \right)^2$$

Prediction of match outcomes

	Likelihood	Class. Rate	RPS
Hybrid Random Forest	0.419	0.556	0.187
Random Forest	0.410	0.548	0.192
Ranking	0.415	0.532	0.190
Lasso	0.419	0.524	0.198
Hybrid Lasso	0.429	0.540	0.194
Bookmakers	0.425	0.524	0.188

Comparison of different prediction methods for ordinal outcome based on multinomial likelihood, classification rate and ranked probability score (RPS)

Prediction of exact numbers of goals

- let now y_{ijk}, for $i, j = 1, \ldots, n$ and $k \in \{2002, 2006, 2010, 2014\}$, denote the observed number of goals scored by team i against team j in tournament k
- $\hat{y}_{ijk}$ the corresponding predicted value
- 2 quadratic errors: $(y_{ijk} - \hat{y}_{ijk})^2$ and $\left((y_{ijk} - y_{jik}) - (\hat{y}_{ijk} - \hat{y}_{jik})\right)^2$

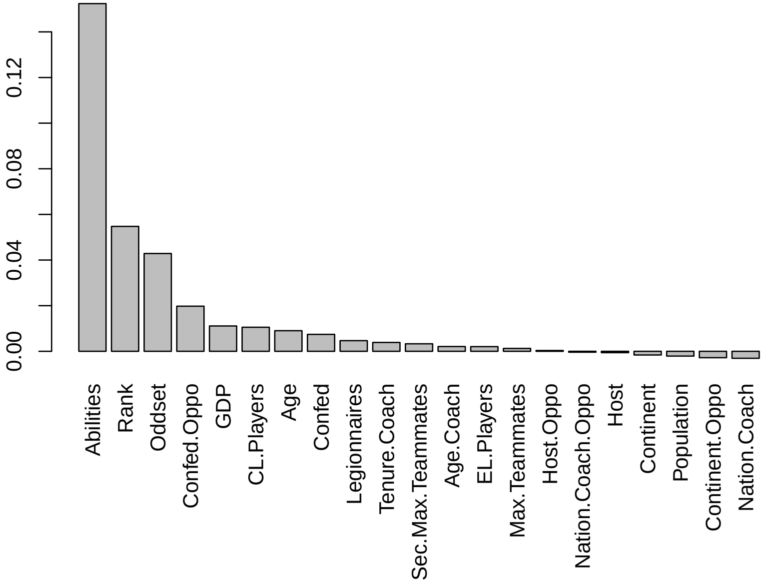
Prediction of exact numbers of goals

	Goal Difference	Goals
Hybrid Random Forest	2.473	1.296
Random Forest	2.543	1.330
Ranking	2.560	1.349
Lasso	2.835	1.421
Hybrid Lasso	2.809	1.427

Comparison of different prediction methods for the exact number of goals and the goal difference based on MSE

Prediction of FIFA World Cup 2018

Variable importance



Winning probabilities

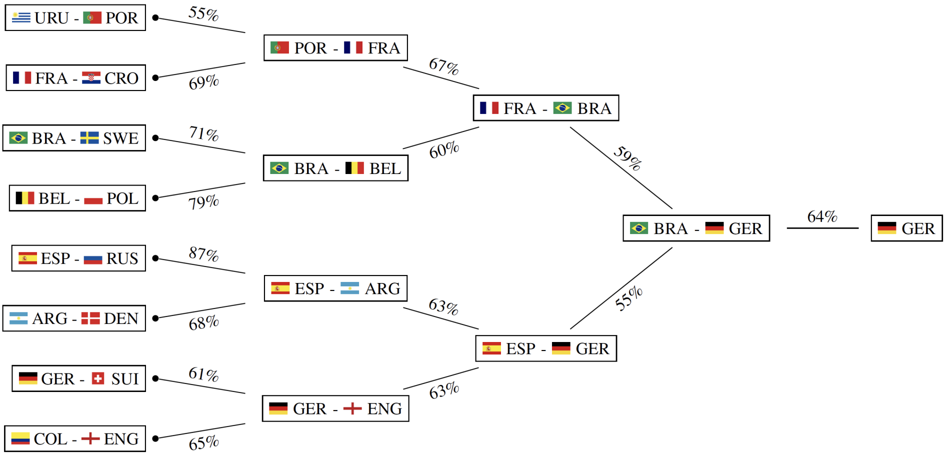
			Round of 16	Quarter finals	Semi finals	Final	World Champion	Oddset
1.		ESP	88.4	73.1	47.9	28.9	17.8	11.8
2.		GER	86.5	58.0	39.8	26.3	17.1	15.0
3.		BRA	83.5	51.6	34.1	21.9	12.3	15.0
4.		FRA	85.5	56.1	36.9	20.8	11.2	11.8
5.		BEL	86.3	64.5	35.7	20.4	10.4	8.3
6.		ARG	81.6	50.5	29.8	15.2	7.3	8.3
7.		ENG	79.8	57.0	29.8	15.6	7.1	4.6
8.		POR	67.5	46.1	19.8	7.3	2.5	3.8
9.		CRO	65.9	30.8	15.6	6.0	2.2	3.0
10.		SUI	58.9	30.6	13.1	5.6	2.2	1.0
11.		COL	79.2	33.1	14.0	5.7	2.1	1.8
12.		DEN	59.0	26.1	12.4	4.8	1.7	1.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Most probable group stage

Group A 28.7%	Group B 38.5%	Group C 31.5%	Group D 30.7%
1. URU	1. ESP	1. FRA	1. ARG
2. RUS	2. POR	2. DEN	2. CRO
KSA	MOR	AUS	ICE
EGY	IRN	PER	NGA

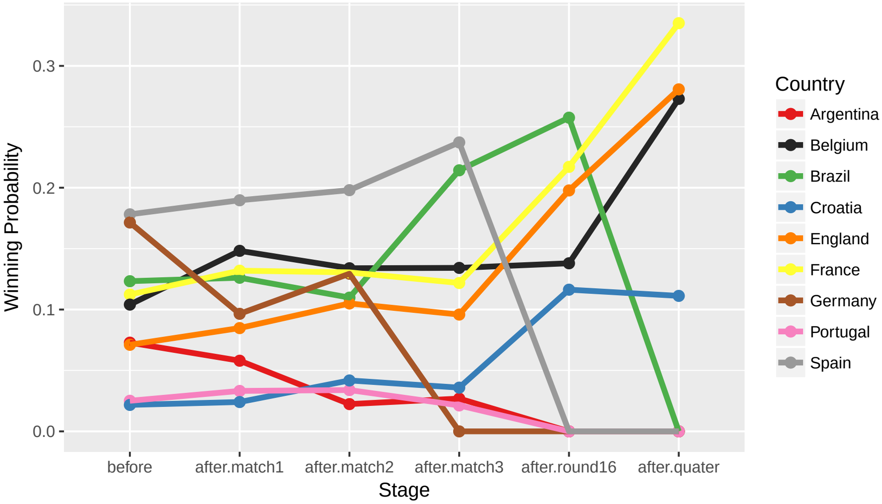
Group E 29.0%	Group F 29.9%	Group G 38.1%	Group H 26.5%
1. BRA	1. GER	1. BEL	1. COL
2. SUI	2. SWE	2. ENG	2. POL
CRC	MEX	PAN	SEN
SRB	KOR	TUN	JPN

Most probable knockout stage



Winning probabilities over time

Time course of the winning probabilities for the nine (originally) favored teams:



Performance I

	Likelihood	Class. Rate	RPS
Hybrid Random Forest	0.440	0.609	0.188
Random Forest	0.433	0.609	0.191
Lasso	0.424	0.547	0.207
Hybrid Lasso	0.434	0.609	0.201
Ranking	0.423	0.578	0.197
Bookmakers	0.438	0.562	0.194

Performance I

	Likelihood	Class. Rate	RPS
Hybrid Random Forest	0.440	0.609	0.188
Random Forest	0.433	0.609	0.191
Lasso	0.424	0.547	0.207
Hybrid Lasso	0.434	0.609	0.201
Ranking	0.423	0.578	0.197
Bookmakers	0.438	0.562	0.194

	Goal Difference	Goals
Hybrid Random Forest	1.181	2.113
Random Forest	1.209	2.177
Lasso	1.216	2.333
Hybrid Lasso	1.187	2.270
Ranking	1.253	2.171

Performance II

Final standing in forecast competition fifaexperts.com (> 500 participants):

Submit your forecasts | Check your results | **Scoreboard** | Your league

1. Esportes em Números: 4650 points
2. Andreas Groll: 4644 points
3. Danilo Lopes: 4634 points
4. Natanael Prata: 4634 points
5. Chance de Gol: 4611 points
6. Wilson Chaves: 4597 points
7. Sigma Benedek: 4589 points
8. Márcio Diniz: 4587 points
9. Francesco Beatrice: 4574 points
10. Alun Owen: 4565 points
11. Tolstói Tói: 4558 points
12. Magne Aldrin: 4557 points

Performance III

Final standing in forecast competition Kicktipp (with colleagues):

Gesamtübersicht

Spieltagspunkte ▼

		Spieltage													
Pos	Name	1	2	3	4	5	6	7	Ac	Vi	Ha	Fi	B	S	G
1	stats_model	14	13	**14**	9	12	**10**	**19**	13	7	4	4	28	2,50	147
2	Hendrik	**20**	**14**	9	9	11	5	8	12	**9**	4	0	28	1,83	129
3	Katharina	12	11	9	10	**15**	**10**	11	16	7	3	2	20	1,50	126
4	Katrin	12	**14**	8	6	12	4	15	**18**	7	4	2	24	0,83	126
5	Lukas	10	12	9	6	9	6	4	15	7	3	6	**32**	1,00	119
6	Jona	10	9	6	10	9	6	11	12	8	6	**7**	24	1,00	118
7	Hilsi	16	8	7	7	10	2	6	14	**9**	**7**	2	24	1,50	112
8	Borussenengel	13	10	10	**11**	14	2	5	14	5	4	2	16	1,00	106

Performance IV

Final standing in WC-forecast competition from Prof. Claus Ekstrøm :

	log.loss
Groll, Ley, Schauberger, VanEetvelde	-11.69
Ekstrom (Skellam)	-11.72
Ekstrom (ELO)	-13.48
Random guessing	-14.56

And the winner is the prediction by **Groll, Ley, Schauberger, VanEetvelde** (although not by much). Well done! Time to prepare the prediction algorithms for the next tournament - and hopefully we can get more people to participate.

Performance V

Betting strategies:

For every match i and each of the possible three outcomes $r \in \{1, 2, 3\}$ calculate expected return:

$$E[return_{ri}] = \hat{\pi}_{ri} * odds_{ri} - 1\,.$$

Performance V

Betting strategies:

For every match i and each of the possible three outcomes $r \in \{1,2,3\}$ calculate expected return:

$$E[return_{ri}] = \hat{\pi}_{ri} * odds_{ri} - 1\,.$$

Choose outcome with highest expected return and only place bet if expected return is positive:

$$\max_{r \in \{1,2,3\}} E[return_{ri}] > \tau = 0\,.$$

Performance V

Betting strategies:

For every match i and each of the possible three outcomes $r \in \{1, 2, 3\}$ calculate expected return:

$$E[return_{ri}] = \hat{\pi}_{ri} * odds_{ri} - 1\,.$$

Choose outcome with highest expected return and only place bet if expected return is positive:

$$\max_{r \in \{1,2,3\}} E[return_{ri}] > \tau = 0\,.$$

Koopman and Lit (2015): use different values of the threshold $\tau > 0 \implies$ overall mean return could be increased.

Performance V

Betting strategies:

For every match i and each of the possible three outcomes $r \in \{1, 2, 3\}$ calculate expected return:

$$E[return_{ri}] = \hat{\pi}_{ri} * odds_{ri} - 1 \,.$$

Choose outcome with highest expected return and only place bet if expected return is positive:

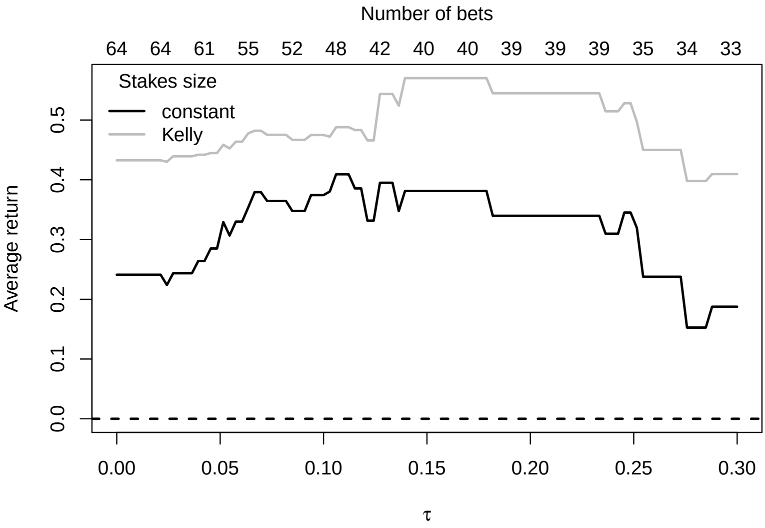
$$\max_{r \in \{1,2,3\}} E[return_{ri}] > \tau = 0 \,.$$

Koopman and Lit (2015): use different values of the threshold $\tau > 0 \Longrightarrow$ overall mean return could be increased.

Boshnakov et al. (2017): use varying stake sizes based on the Kelly criterion (Kelly, 1956). $\Longrightarrow$ determines optimal stake for single bets in order to maximize the return considering size of the odds and the winning probability.

Performance V

Betting strategies:



Recent extensions:

- more "hybrid" features
- XGBoost

Recent extensions

For the prediction of the **UEFA EURO 2020** (Groll et al., 2021), beside the current ability ranking based on historic matches (Ley et al., 2018), we included two additional **hybrid features**:

- bookmaker consensus abilities (Leitner et al., 2010)

Recent extensions

For the prediction of the **UEFA EURO 2020** (Groll et al., 2021), beside the current ability ranking based on historic matches (Ley et al., 2018), we included two additional **hybrid features**:

- bookmaker consensus abilities (Leitner et al., 2010)
- plus-minus player ratings (Hvattum & Gelade, 2021)

Recent extensions

For the prediction of the **UEFA EURO 2020** (Groll et al., 2021), beside the current ability ranking based on historic matches (Ley et al., 2018), we included two additional **hybrid features**:

- bookmaker consensus abilities (Leitner et al., 2010)
- plus-minus player ratings (Hvattum & Gelade, 2021)

Moreover, we compared the random forest with an ***extreme gradient boosting approach*** (XGBoost; Chen and Guestrin, 2016).

Summary

Regarded models & predictive performance:

- (Regularized) regression approaches vs. random forests vs. ranking methods
- random forests & ranking methods perform pretty good (almost as good as bookmakers)

Summary

Regarded models & predictive performance:

- (Regularized) regression approaches vs. random forests vs. ranking methods
- random forests & ranking methods perform pretty good (almost as good as bookmakers)
- $\Longrightarrow$ combine random forests & ranking methods to hybrid random forest

Summary

Regarded models & predictive performance:

- (Regularized) regression approaches vs. random forests vs. ranking methods
- random forests & ranking methods perform pretty good (almost as good as bookmakers)
- $\Longrightarrow$ combine random forests & ranking methods to hybrid random forest
- $\Longrightarrow$ combination outperforms bookmakers (on FIFA WC 2002 – 2014 data)

Summary

Regarded models & predictive performance:

- (Regularized) regression approaches vs. random forests vs. ranking methods
- random forests & ranking methods perform pretty good (almost as good as bookmakers)
- $\Longrightarrow$ combine random forests & ranking methods to hybrid random forest
- $\Longrightarrow$ combination outperforms bookmakers (on FIFA WC 2002 – 2014 data)

FIFA WC 2018 prediction:

- Spain favorite with 17.8%, closely follow by Germany (17.1%); then: Brazil, France, Belgium (**before the tournament start**)

Summary

Regarded models & predictive performance:

- (Regularized) regression approaches vs. random forests vs. ranking methods
- random forests & ranking methods perform pretty good (almost as good as bookmakers)
- $\Longrightarrow$ combine random forests & ranking methods to hybrid random forest
- $\Longrightarrow$ combination outperforms bookmakers (on FIFA WC 2002 – 2014 data)

FIFA WC 2018 prediction:

- Spain favorite with 17.8%, closely follow by Germany (17.1%); then: Brazil, France, Belgium (**before the tournament start**)
- Performance: Germany & Spain already dropped out; **but**: very good performance **on average**

Summary

Regarded models & predictive performance:

- (Regularized) regression approaches vs. random forests vs. ranking methods
- random forests & ranking methods perform pretty good (almost as good as bookmakers)
- $\Longrightarrow$ combine random forests & ranking methods to hybrid random forest
- $\Longrightarrow$ combination outperforms bookmakers (on FIFA WC 2002 – 2014 data)

FIFA WC 2018 prediction:

- Spain favorite with 17.8%, closely follow by Germany (17.1%); then: Brazil, France, Belgium (**before the tournament start**)
- Performance: Germany & Spain already dropped out; **but**: very good performance **on average**
- Conclusion: single match outcome / tournament winner almost impossible to predict, but in general very adequate model

References

Boshnakov, G., T. Kharrat and I. G. McHale (2017): A bivariate Weibull count model for forecasting association football scores, *International Journal of Forecasting 33*, 458–466.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794.

Friedman, J., T. Hastie and R. Tibshirani (2010): Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, 33, 1.

Groll, A. and J. Abedieh (2013). Spain retains its title and sets a new record - generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports 9*(1), 51–66.

Groll, A., G. Schauberger, and G. Tutz (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports 11*(2), 97–115.

Groll, A., T. Kneib, A. Mayr, and G. Schauberger (2018). On the dependency of soccer scores – A sparse bivariate Poisson model for the UEFA European Football Championship 2016. *Journal of Quantitative Analysis in Sports 14(2)*, 65-79.

References II

Groll, A., C. Ley, H. Van Eetvelde and G. Schauberger (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports 15*, 271-287.

Groll, A., C. Ley, G. Schauberger, H. Van Eetvelde, and A. Zeileis, (2019). Hybrid Machine Learning Forecasts for the FIFA Women's World Cup 2019. *arXiv preprint arXiv:1906.01131*

Groll, A., G. Schauberger, and H. van Eetvelde (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.

Groll, A., L.M. Hvattum, C. Ley, F. Popp, G. Schauberger, H. Van Eetvelde, H. and A. Zeileis (2021). Hybrid machine learning forecasts for the UEFA EURO 2020. *arXiv preprint arXiv:2106.05799*

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*, 651–674.

Hvattum, L. and G. Gelade (2021). Comparing bottom-up and top-down ratings for individual soccer players. *International Journal of Computer Science in Sport 20*, 23–42.

References III

Kelly, J. L. (1956): A new interpretation of information rate. *Bell System Technical Journal 35*, 917–926.

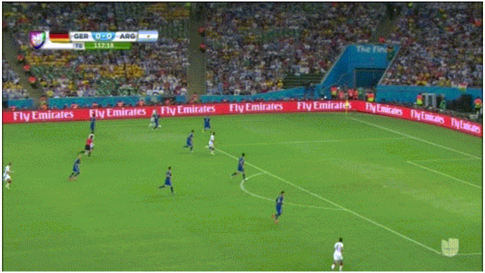
Koopman, S. J. and R. Lit (2015): A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*, 167–186.

Ley, C., T. Van de Wiele and H. Van Eetvelde (2018): Ranking soccer teams on basis of their current strength: a comparison of maximum likelihood approaches, *Statistical Modelling 19*, 55–77.

Wright, M. N. and A. Ziegler (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software 77*(1), 1–17.

Schauberger, G. and A. Groll (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling 18*(5-6), 1–23.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B 58*, 267–288.



European champion 2024?

Sources: gifrific.com, dfb.de

Thank you for your attention!

Link arXiv Working Paper: https://arxiv.org/abs/2106.05799

and

Blog with interactive graphs: https://www.zeileis.org/news/euro2020/

Appendix

Alternative approach

Copula regression:

- van der Wurp, H., A. Groll, T. Kneib, G. Marra, and R. Radice (2020) Generalised joint regression for count data: a penalty extension for competitive settings. *Statistics and Computing 30*, 1419–1432.
- van der Wurp, H. and A. Groll (2023a) Introducing LASSO-type penalisation to generalised joint regression modelling for count data. *Advances in Statistical Analysis 107*, 127–151.
- van der Wurp, H. and A. Groll (2023b). Using (copula) regression and machine learning to model and predict football results in major European leagues. *Statistica Applicata*. To appear.

Sources: For The Win - USATODAY.com, Tadias Magazine

Similar model used for the FIFA Women's World Cup 2019 in France

(Working paper on arXiv: https://arxiv.org/pdf/1906.01131.pdf)

Similar model used for the FIFA Women's World Cup 2019 in France

(Working paper on arXiv: https://arxiv.org/pdf/1906.01131.pdf)

(Blog: http://bit.ly/fifa-women-2019)

Sources: For The Win - USATODAY.com, Tadias Magazine

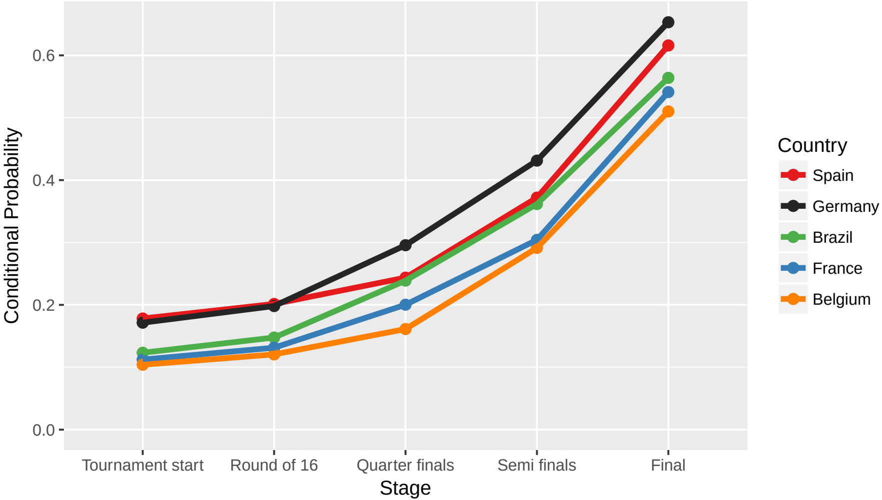
Source: The New Yorker

Winning probabilities

		Round of 16	Quarter finals	Semi finals	Final	World Champion	Bookmakers
1.	USA	98.4	75.5	53.4	39.6	28.1	17.7
2.	FRA	95.9	66.8	40.7	25.4	14.3	18.2
3.	ENG	96.1	69.8	45.3	23.8	13.3	11.0
4.	GER	95.4	66.3	36.9	22.9	12.9	12.4
5.	NED	92.7	47.1	25.9	12.0	5.1	6.0
6.	SWE	91.2	50.7	24.8	12.1	4.4	3.3
7.	BRA	88.7	51.2	25.5	10.5	3.9	3.8
8.	AUS	89.0	50.0	24.2	10.1	3.8	4.7
9.	ESP	81.5	43.8	20.1	9.4	3.6	3.6
10.	JPN	82.5	43.3	21.1	8.0	2.7	5.3
11.	CAN	85.4	33.2	14.7	5.7	2.0	3.1
12.	ITA	81.7	38.8	16.7	5.8	1.9	1.6
13.	NOR	75.0	33.7	13.1	4.6	1.5	2.2
14.	CHN	72.5	29.0	9.5	3.1	0.8	1.5
15.	SCO	66.6	24.5	8.3	2.4	0.7	0.9
16.	KOR	64.8	23.6	7.3	2.0	0.5	1.2
17.	NZL	65.4	16.1	4.9	1.2	0.3	1.1
18.	THA	36.9	7.9	1.8	0.3	0.1	0.2
19.	NGA	30.1	6.5	1.3	0.2	0.0	0.4
20.	ARG	22.6	5.2	1.0	0.2	0.0	0.7
21.	CHI	26.2	5.4	1.1	0.2	0.0	0.7
22.	CMR	26.6	5.1	1.1	0.2	0.0	0.2
23.	RSA	19.6	3.9	0.8	0.1	0.0	0.3
24.	JAM	15.1	2.7	0.4	0.1	0.0	0.1

Conditional winning probabilities

Winning probabilities conditional on reaching the single stages of the tournament for the five favored teams:



Winning probabilities after group stage

		Quarter finals	Semi finals	Final	World Champion
1.	ESP	88.2	61.1	42.2	23.7
2.	BRA	79.9	51.2	35.6	21.4
3.	BEL	85.1	40.9	24.1	13.4
4.	FRA	63.4	43.6	22.1	12.2
5.	ENG	71.6	45.4	20.1	9.6
6.	SUI	60.6	24.1	9.7	3.6
7.	CRO	56.1	20.8	10.2	3.6
8.	ARG	36.6	21.6	7.0	2.7
9.	DEN	43.9	15.2	6.8	2.4
10.	POR	55.1	19.0	5.5	2.1
11.	COL	28.4	15.9	5.2	1.8
12.	SWE	39.4	14.7	5.1	1.5
13.	URU	44.9	15.8	4.0	1.4
14.	MEX	20.1	4.7	1.2	0.3
15.	RUS	11.8	2.8	0.7	0.1
16.	JPN	14.9	3.1	0.6	0.1