

Prediction of the FIFA World Cup 2018

A random forest approach with an emphasis on
estimated team ability parameters

- Data structure
- Basic idea of random forests, (regularized) Poisson regression and ranking methods and compare their predictive performances
- The best-performing model, which is a combination of random forests and ranking methods, is fitted to the data and used to predict the FIFA World Cup 2018

DATA

- **GDP per capita**

Wealthier countries tend to be sportier

Number of grass pitches

To account for the general increase of the gross domestic product (GDP) during 2002 – 2014

- **Population:**

The population size is used in relation to the respective global population to account for the general world population growth

- **ODDSET probability**

Convert bookmaker odds provided by the German state betting agency ODDSET into winning probabilities

- **FIFA rank:**

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

DATA

- **Host**

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

- **Continent**

A dummy variable indicating if a national team is from the same continent as the host of the World Cup (including the host itself)

- **Confederation**

This categorical variable comprises the teams' confederation with six possible values: Africa (CAF); Asia (AFC); Europe (UEFA); North, Central America and Caribbean (CONCACAF); Oceania (OFC); South America (CONMEBOL).

DATA

- **Team Structure**

Maximum number of teammates for each squad, both the maximum and second maximum number of teammates playing together in the same national club is counted.

- **Average age**

The average age of each squad is collected.

- **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 UEFA Europa League (EL) are counted.

- **Number of players abroad/Legionnaires**

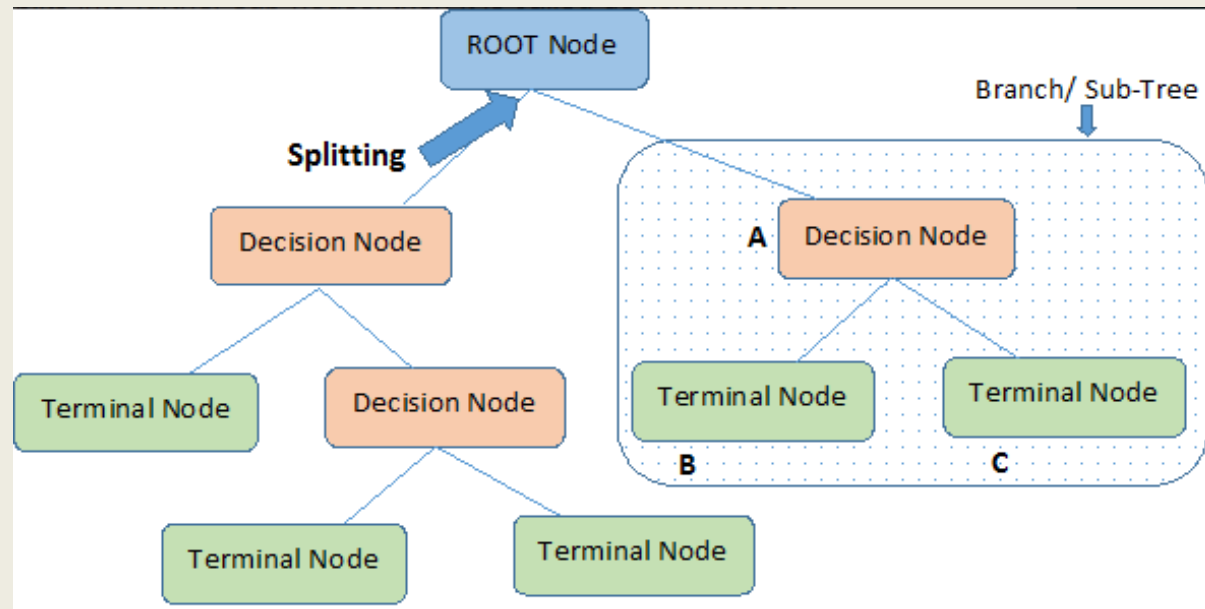
For each squad, the number of players playing in clubs abroad (in the season preceding the respective World Cup) is counted

- **Factors describing the team's coach**

For the coaches of the teams, Age and duration of their Tenure are observed. Furthermore, a dummy variable is included, if a coach has the same Nationality as his team.

Decision Tree

1. Root node
2. Splitting
3. Decision node
4. Leaf/ terminal node
5. Pruning

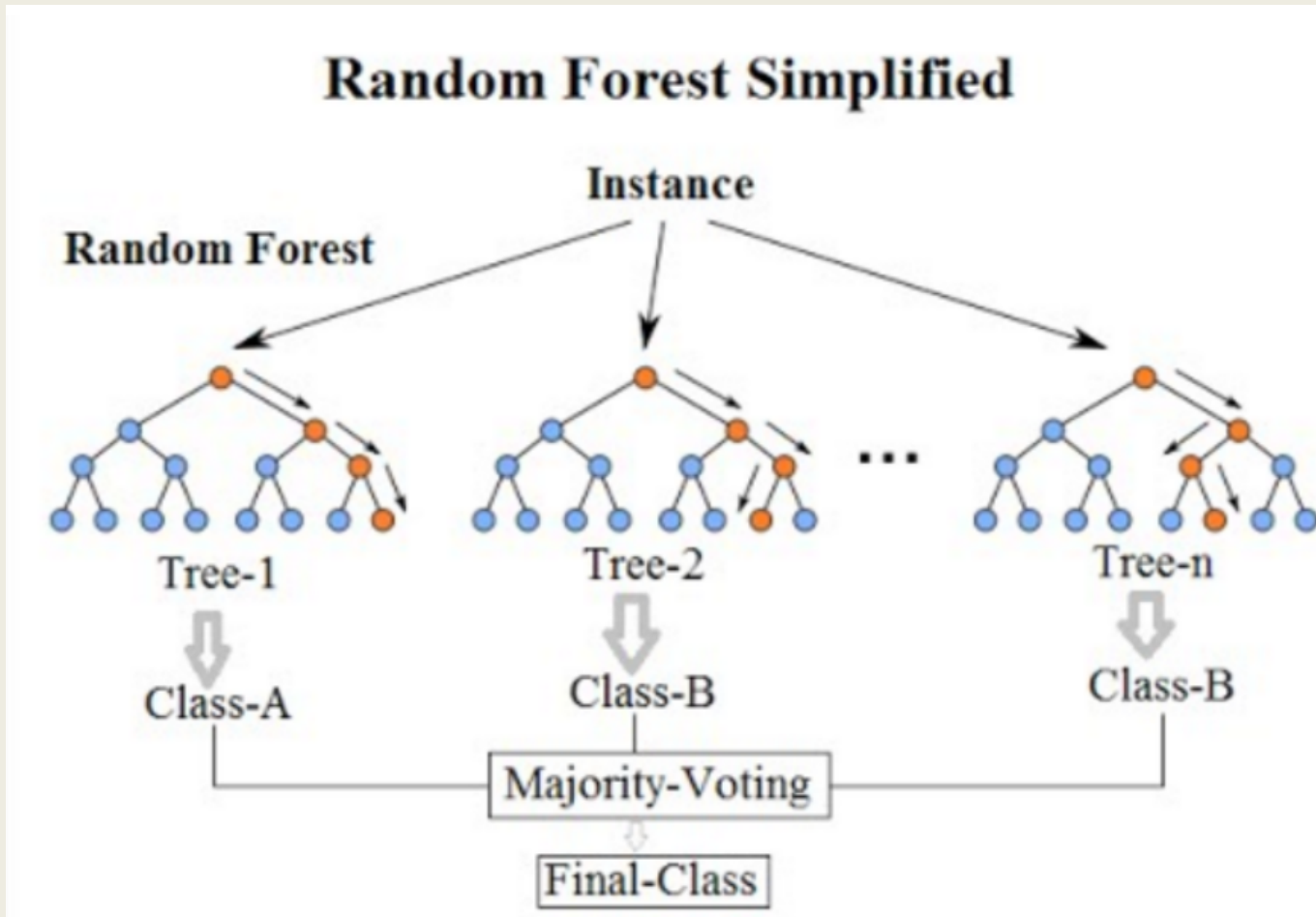


Decision Tree

- **Advantages:**
 - Easy to understand
 - Useful in data exploration
 - Less data cleaning required
 - Data type is not a constraint
 - Non parametric method
- **Disadvantages**
 - Over fitting
 - Not a good fit for continuous variables

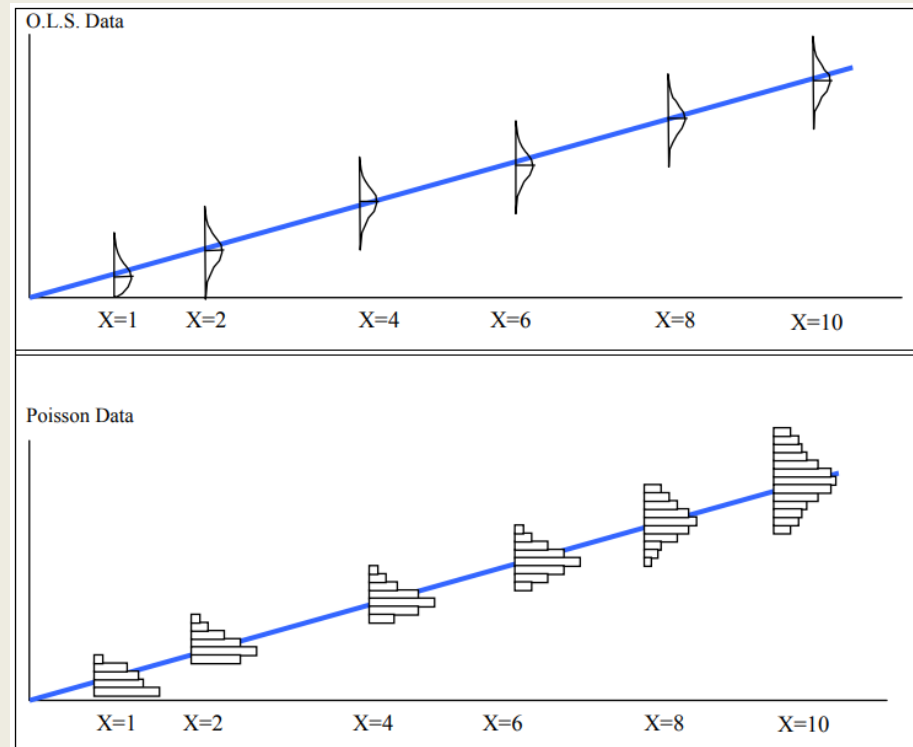
Random Forest

Bagging many decision trees!



Poisson Regression

- A regression model for count data
- Assumptions (besides all the independencies!):
 - Usually rare events
 - Equal mean and variance



Methods used in the paper

































- Models:
 - Random Forest
 - Ranking (bivariate Poisson regression)
 - LASSO (penalized Poisson regression)
- Performance measures:
 - Likelihood (multinomial dist.)
 - Classification rate
 - Rank probability score (RPS)

Model comparison results

	Likelihood	Class. Rate	RPS
Random Forest	0.410	0.548	0.192
Lasso	0.419	0.524	0.198
Ranking	0.415	0.532	0.190
Bookmakers	0.425	0.524	0.188

	Goal Difference	Goals
Random Forest	2.543	1.330
Lasso	2.835	1.421
Ranking	2.560	1.349

Table 8: Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2018 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup together with winning probabilities based on the ODDSET odds.

			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
13.		URU	86.6	37.5	13.5	4.4	1.3	2.8
14.		SWE	54.0	21.7	8.0	3.1	1.0	0.8
15.		POL	60.6	18.9	6.8	2.3	0.7	1.5
16.		PER	39.2	15.4	6.6	2.1	0.6	0.4
17.		ICE	36.6	12.9	5.3	1.7	0.5	0.6
18.		SRB	36.2	13.8	4.7	1.5	0.4	0.6
19.		SEN	39.7	10.9	3.7	1.1	0.3	0.6
20.		MOR	30.3	14.8	4.0	1.0	0.3	0.3
21.		TUN	22.8	8.9	2.8	0.8	0.2	0.2
22.		MEX	41.5	13.9	3.7	1.1	0.2	1.0
23.		CRC	21.4	6.4	1.7	0.4	0.1	0.3
24.		EGY	45.5	10.3	2.1	0.4	0.1	0.6
25.		RUS	50.4	10.5	2.4	0.4	0.1	2.2
26.		NGA	15.8	4.0	1.2	0.3	0.1	0.6
27.		AUS	16.2	4.2	1.2	0.3	0.1	0.3
28.		JPN	20.5	4.1	0.9	0.2	0.0	0.6
29.		KOR	17.9	4.0	0.8	0.2	0.0	0.6
30.		IRN	13.8	5.1	0.9	0.1	0.0	0.3
31.		PAN	11.1	2.5	0.5	0.1	0.0	0.1
32.		KSA	17.5	2.6	0.4	0.0	0.0	0.1

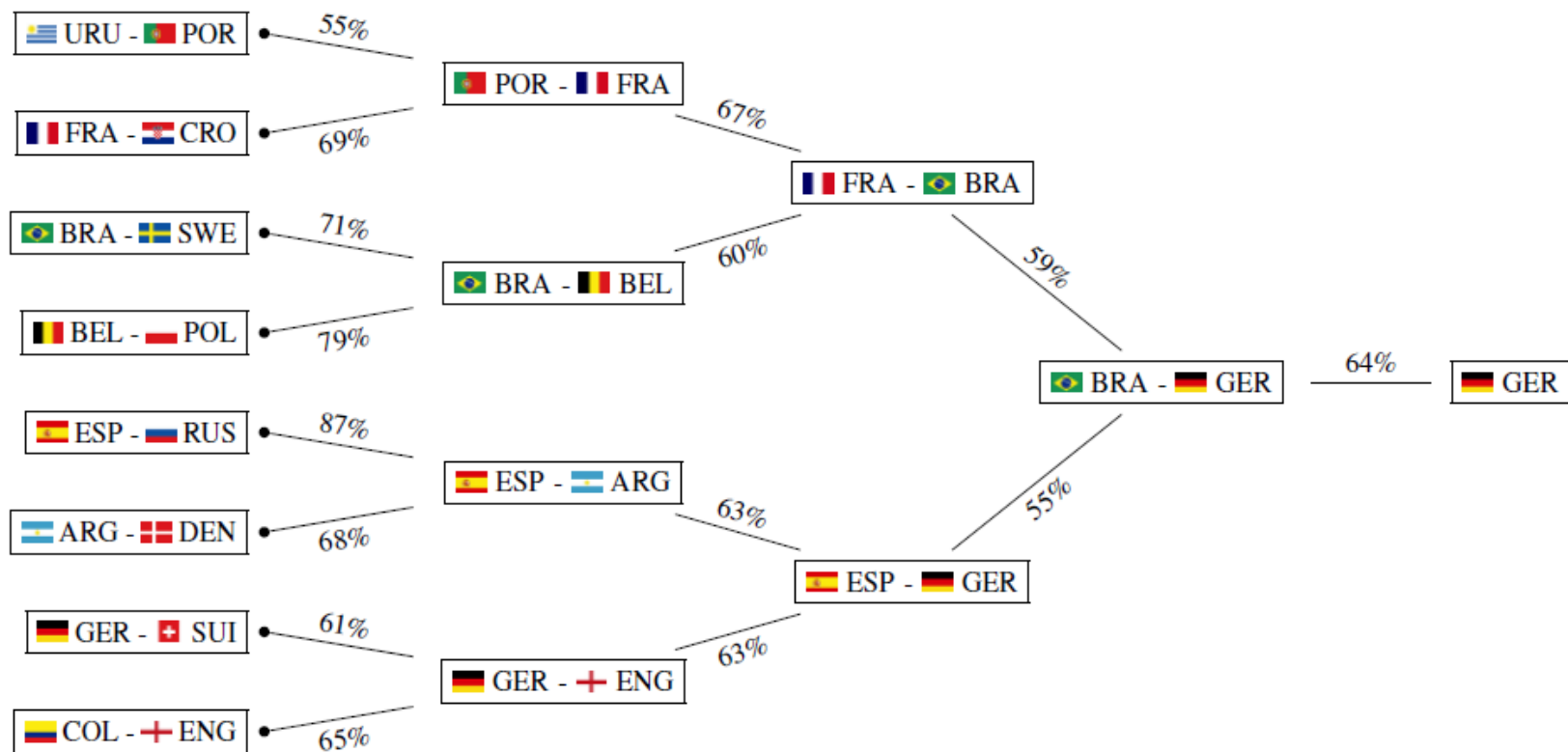


Figure 5: Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2018 based on 100,000 simulation runs.