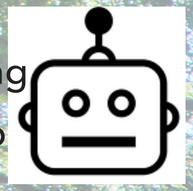
Machine Learning
Reading Group



A SHORT PRESENTATION ON

RANDOM FORESTS

04.07.2018

Alina Miron

DECISION TREES

IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
1	10	200	1	0	1	Yes
1	8	175	0	1	1	No
1	9	175	1	0	1	No
1	9	210	1	0	1	Yes
1	9	205	1	0	0	Yes
0	2	210	0	0	1	No

DECISION TREES

IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
1	10	200	1	0	1	Yes
1	8	175	0	1	1	No
1	9	175	1	0	1	No
1	9	210	1	0	1	Yes
1	9	205	1	0	0	Yes
0	2	210	0	0	1	No

DECISION TREES: HOW DO YOU CONSTRUCT THE TREE?

Decision trees usually work top-down, by choosing a variable at each step that best splits the set of items. This can be done by using:

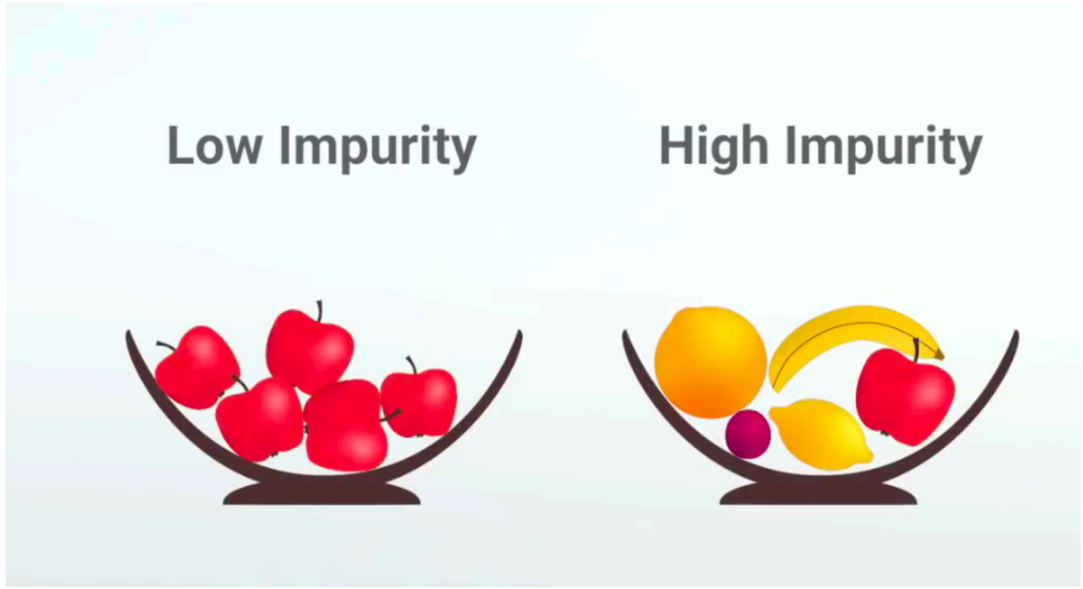
- Gini impurity
- Information gain

THE BASIC IDEA OF

DECISION TREES: GINI IMPURITY

measurement of the likelihood of an **incorrect classification** of a new instance of a random variable, if that new instance were **randomly classified according to the distribution of class labels** from the data set

$$G(k) = 1 - \sum (p_i^2)$$



DECISION TREES

$$G(k) = 1 - \sum (p_i^2)$$

IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
1	10	200	1	0	1	Yes
1	8	175	0	1	1	No
1	9	175	1	0	1	No
1	9	210	1	0	1	Yes
1	9	205	1	0	0	Yes
0	2	210	0	0	1	No

$$G(k) = 1 - (0.5^2 + 0.5^2) = 0.5$$

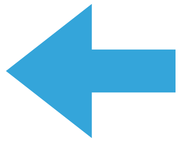
DECISION TREES LEARNING: INFORMATION GAIN

Find the question that reduces our uncertainty the most!

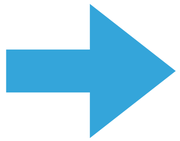
IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
1	10	200	1	0	1	Yes
1	8	175	0	1	1	No
1	9	175	1	0	1	No
1	9	210	1	0	1	Yes
1	9	205	1	0	0	Yes
0	2	210	0	0	1	No

DECISION TREES LEARNING: INFORMATION GAIN

IsRound	IsApple?
1	Yes
1	No
1	No
1	Yes
1	Yes
0	No



IsRound	IsApple?
1	Yes
1	No
1	No
1	Yes
1	Yes



IsRound	IsApple?
0	No

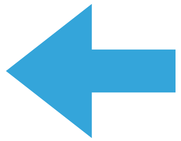
Impurity: 1 - (3/5 ^2 +2/5^2)

= 1 - (0.36+0.16) = 0.48

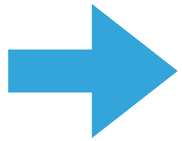
Impurity: 0

DECISION TREES LEARNING: INFORMATION GAIN

IsRound	IsApple?
1	Yes
1	No
1	No
1	Yes
1	Yes
0	No



IsRound	IsApple?
1	Yes
1	No
1	No
1	Yes
1	Yes



IsRound	IsApple?
0	No

Impurity: 1 - (3/5 ^2 +2/5^2)

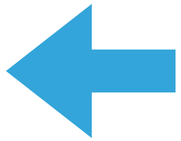
= 1 - (0.36+0.16) = 0.48

Impurity: 0

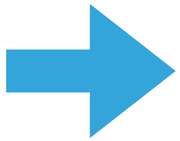
Average Impurity: 0.48 * 5/6 + 0 * 1/6 = 0.4

DECISION TREES LEARNING: INFORMATION GAIN

IsRound	IsApple?
1	Yes
1	No
1	No
1	Yes
1	Yes
0	No



IsRound	IsApple?
1	Yes
1	No
1	No
1	Yes
1	Yes



IsRound	IsApple?
0	No

Impurity: 1 - (3/5 ^2 +2/5^2)

= 1 - (0.36+0.16) = 0.48

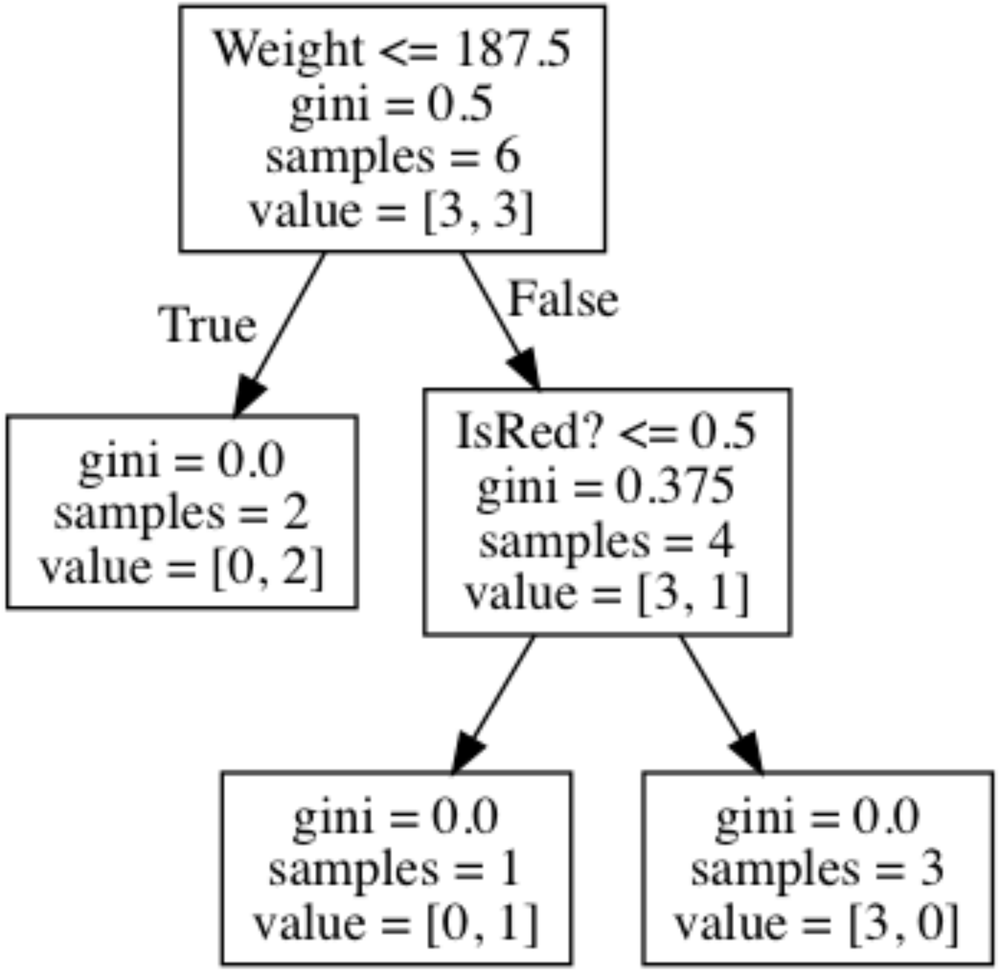
Impurity: 0

Information gain: Initial Impurity - The current node impurity = 0.5 - 0.4 = 0.1

DECISION TREES LEARNING

- You compute the information gain for every single node, and you choose the node that maximises the information gain!

DECISION TREES LEARNING



THE BASIC IDEA OF

TREE BAGGING

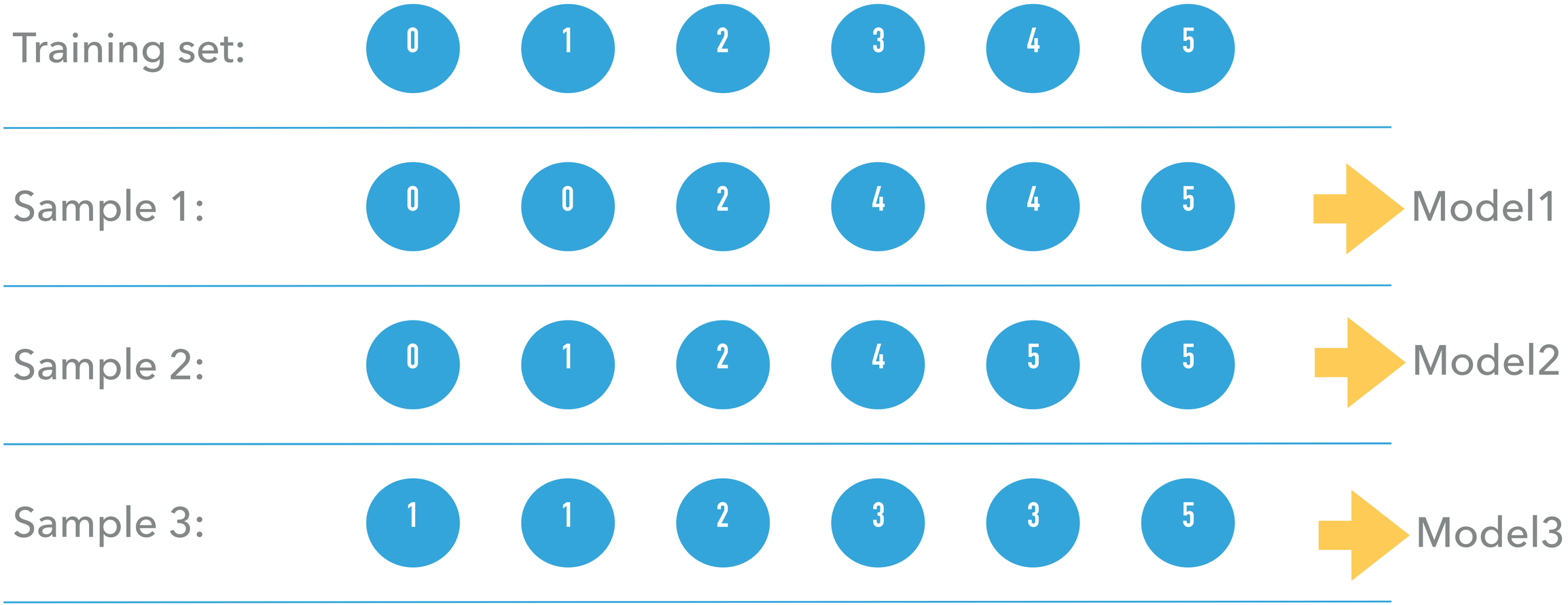
- BOOTSTRAP SAMPLES OF YOUR TRAINING SET!
- ON EVERY SET TRAIN A NEW TREE
- AVERAGE THE RESULTS OVER THE TREES

BAGGING

ID	IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
0	1	10	200	1	0	1	Yes
1	1	8	175	0	1	1	No
2	1	9	175	1	0	1	No
3	1	9	210	1	0	1	Yes
4	1	9	205	1	0	0	Yes
5	0	2	210	0	0	1	No

THE BASIC IDEA OF

BAGGING



THE BASIC IDEA OF

RANDOM FORESTS

- BOOTSTRAP SAMPLES OF YOUR TRAINING SET
- **RANDOM SELECT A SET OF FEATURES**
- ON EVERY SET TRAIN A NEW TREE
- AVERAGE THE RESULTS OVER THE TREES

RANDOM FORESTS

ID	IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
0	1	10	200	1	0	1	Yes
1	1	8	175	0	1	1	No
2	1	9	175	1	0	1	No
3	1	9	210	1	0	1	Yes
4	1	9	205	1	0	0	Yes
5	0	2	210	0	0	1	No

RANDOM FORESTS

ID	IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
0	1	10	200	1	0	1	Yes
1	1	8	175	0	1	1	No
2	1	9	175	1	0	1	No
3	1	9	210	1	0	1	Yes
4	1	9	205	1	0	0	Yes
5	0	2	210	0	0	1	No

RANDOM FORESTS

ID	IsRound	Diameter	Weight	IsRed?	IsOrange	IsSweet	IsApple?
0	1	10	200	1	0	1	Yes
1	1	8	175	0	1	1	No
2	1	9	175	1	0	1	No
3	1	9	210	1	0	1	Yes
4	1	9	205	1	0	0	Yes
5	0	2	210	0	0	1	No

THE BASIC IDEA OF

XGBOOST

- IN MOST KAGGLE COMPETITIONS IT REPLACED RANDOM FORESTS AS A CLASSIFICATION METHOD
- RANDOM FORESTS AND BOOSTED TREES ARE NOT DIFFERENT IN TERMS OF MODEL, THE DIFFERENCE IS HOW WE TRAIN THEM

Prediction of the FIFA World Cup 2018 – A random forest approach with an emphasis on estimated team ability parameters

Andreas Groll * Christophe Ley † Gunther Schauberger ‡
Hans Van Eetvelde §

June 8, 2018

Abstract In this work, we compare three different modeling approaches for the scores of soccer matches with regard to their predictive performances based on all matches from the four previous FIFA World Cups 2002 – 2014: *Poisson regression models*, *random forests* and *ranking methods*. While the former two are based on the teams' covariate information, the latter method estimates adequate ability parameters that reflect the current strength of the teams best. Within this comparison the best-performing prediction methods on the training data turn out to be the ranking methods and the random forests. However, we show that by combining the random forest with the team ability parameters from the ranking methods as an additional covariate we can improve the predictive power substantially. Finally, this combination of methods is chosen as the final model and based on its estimates, the FIFA World Cup 2018 is simulated repeatedly and winning probabilities are obtained for all teams. The model slightly favors Spain before the defending champion Germany. Additionally, we provide survival probabilities for all teams and at all tournament stages as well as the most probable tournament outcome.

1 Introduction

Like the previous FIFA World Cup 2014, also the up-coming tournament in Russia has caught the attention of several modelers who try to predict the tournament winner. One approach that has already produced reasonable results for several of the past European championships (EUROs) and FIFA World Cups is based on the prospective information contained in bookmakers' odds (Leitner, Zeileis, and Hornik, 2010b, Zeileis, Leitner, and Hornik, 2012, 2014, 2016). Nowadays, for such major tournaments bookmakers offer a bet on the winner in advance of the tournament. By aggregating the winning odds from several online bookmakers and transforming those into winning probabilities, inverse tournament simulation can be used to compute team-specific abilities, see Leitner, Zeileis, and Hornik (2010a). With the team-specific abilities all single matches are simulated via paired comparisons and, hence, the complete tournament course is obtained. Using this approach, Zeileis, Leitner, and Hornik (2018) forecast Brazil to win the FIFA World Cup 2018 with a probability of 16.6%, followed by Germany (15.8%) and Spain (12.5%).

A fundamentally different modeling approach is based on random (decision) forests – an ensemble learning method for classification, regression and other tasks proposed by Breiman (2001). The method originates from the machine learning and data mining community and operates by first constructing a multitude of so-called decision trees (see, e.g., Quinlan, 1986; Breiman, Friedman, Olshen, and Stone, 1984) on training data. The predictions from the individual trees are then summarized, either by taking the mode of the predicted classes (in classification) or by averaging the predicted values (in regression). This way, random forests reduce the tendency of overfitting and the variance compared to regular decision trees, and, hence, are a common powerful tool for prediction. In preliminary work from Schauberger and Groll (2018) the predictive performance of different types of random forests has been compared on data containing all matches of the FIFA World Cups 2002 – 2014 with conventional regression methods for count data, such as the Poisson models mentioned above. It turned out that random forests provided very satisfactory results and generally outperformed the regression approaches. Moreover, their predictive performances actually were either close to or even outperforming those of the bookmakers, which serve as natural benchmark. These results motivate us to use random forests in the present work to calculate predictions of the up-coming FIFA World Cup 2018. However, we will show that the already excellent predictive power of the random forests can be further increased if adequate estimates of team ability parameters, reflecting the current strength of the national teams, are incorporated as additional covariates.

The rest of the manuscript is structured as follows: in Section 2 we describe the underlying data set covering all matches of the four preceding FIFA World Cups 2002 – 2014. Next, in Section 3 we briefly explain the basic idea of random forests, (regularized) Poisson regression and ranking methods and compare their predictive performances. Then, the best-performing model, which is a combination of random

2 Data

In this section, we briefly describe the underlying data set covering all matches of the four preceding FIFA World Cups 2002 – 2014 together with several potential influence variables. Basically, we use the same set of covariates that is introduced in Groll et al. (2015). For each participating team, the covariates are observed either for the year of the respective World Cup (e.g., GDP per capita) or shortly before the start of the World Cup (e.g., FIFA ranking), and, therefore, vary from one World Cup to another.

Economic Factors:

GDP per capita. To account for the general increase of the gross domestic product (GDP) during 2002 – 2014, a ratio of the GDP per capita of the respective country and the worldwide average GDP per capita is used (source: `http://unstats.un.org/unsd/snaama/dnllist.asp`).

Population. The population size is used in relation to the respective global population to account for the general world population growth (source: `http://data.worldbank.org/indicator/SP.POP.TOTL`).

Sportive factors:

ODDSET probability. We convert bookmaker odds provided by the German state betting agency ODDSET into winning probabilities. The variable hence reflects the probability for each team to win the respective World Cup[1].

FIFA rank. The FIFA ranking system ranks all national teams based on their performance over the last four years (source: `http://de.fifa.com/worldranking/index.html`).

Home advantage:

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).

Factors describing the team's structure:

The following variables describe the structure of the teams. They were observed with the 23-player-squad nominated for the respective World Cup.

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

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

(a) Table of results

FRA	0:1	SEN
URU	1:2	DEN
FRA	0:0	URU
DEN	1:1	SEN
⋮	⋮	⋮

(b) Table of covariates

World Cup	Team	Age	Rank	Oddset	...
2002	France	28.3	1	0.149	...
2002	Uruguay	25.3	24	0.009	...
2002	Denmark	27.4	20	0.012	...
2002	Senegal	24.3	42	0.006	...
⋮	⋮	⋮	⋮	⋮	⋱

Table 2: Exemplary table illustrating the data structure.

Goals	Team	Opponent	Age	Rank	Oddset	...
0	France	Senegal	4.00	-41	0.14	...
1	Senegal	France	-4.00	41	-0.14	...
1	Uruguay	Denmark	-2.10	4	-0.00	...
2	Denmark	Uruguay	2.10	-4	0.00	...
0	France	Uruguay	3.00	-23	0.14	...
0	Uruguay	France	-3.00	23	-0.14	...
1	Denmark	Senegal	3.10	-22	0.01	...
1	Senegal	Denmark	-3.10	22	-0.01	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

ing general procedure on the World Cup 2002 – 2014 data:

1. *Form a training data set containing three out of four 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.*

Table 4: Comparison of the prediction methods for ordinal match outcomes.

	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

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	ISL
EGY	IRN	PER	NGA

Group A

Main article: 2018 FIFA World Cup Group A

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	Uruguay	3	3	0	0	5	0	+5	**9**	Advance to knockout stage
2	Russia **(H)**	3	2	0	1	8	4	+4	**6**	
3	Saudi Arabia	3	1	0	2	2	7	−5	**3**	
4	Egypt	3	0	0	3	2	6	−4	**0**	

Group B

Main article: 2018 FIFA World Cup Group B

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	Spain	3	1	2	0	6	5	+1	**5**	Advance to knockout stage
2	Portugal	3	1	2	0	5	4	+1	**5**	
3	Iran	3	1	1	1	2	2	0	**4**	
4	Morocco	3	0	1	2	2	4	−2	**1**	

Group C

Main article: 2018 FIFA World Cup Group C

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	France	3	2	1	0	3	1	+2	**7**	Advance to knockout stage
2	Denmark	3	1	2	0	2	1	+1	**5**	
3	Peru	3	1	0	2	2	2	0	**3**	
4	Australia	3	0	1	2	2	5	−3	**1**	

Group D

Main article: 2018 FIFA World Cup Group D

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	Croatia	3	3	0	0	7	1	+6	**9**	Advance to knockout stage
2	Argentina	3	1	1	1	3	5	−2	**4**	
3	Nigeria	3	1	0	2	3	4	−1	**3**	
4	Iceland	3	0	1	2	2	5	−3	**1**	

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

Group E

Main article: 2018 FIFA World Cup Group E

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	Brazil	3	2	1	0	5	1	+4	**7**	Advance to knockout stage
2	Switzerland	3	1	2	0	5	4	+1	**5**	
3	Serbia	3	1	0	2	2	4	−2	**3**	
4	Costa Rica	3	0	1	2	2	5	−3	**1**	

Group F

Main article: 2018 FIFA World Cup Group F

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	Sweden	3	2	0	1	5	2	+3	**6**	Advance to knockout stage
2	Mexico	3	2	0	1	3	4	−1	**6**	
3	South Korea	3	1	0	2	3	3	0	**3**	
4	Germany	3	1	0	2	2	4	−2	**3**	

Group G

Main article: 2018 FIFA World Cup Group G

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	Belgium	3	3	0	0	9	2	+7	**9**	Advance to knockout stage
2	England	3	2	0	1	8	3	+5	**6**	
3	Tunisia	3	1	0	2	5	8	−3	**3**	
4	Panama	3	0	0	3	2	11	−9	**0**	

Group H

Main article: 2018 FIFA World Cup Group H

Pos	Team [v·t·e]	Pld	W	D	L	GF	GA	GD	Pts	Qualification
1	Colombia	3	2	0	1	5	2	+3	**6**	Advance to knockout stage
2	Japan	3	1	1	1	4	4	0	**4**[a]	
3	Senegal	3	1	1	1	4	4	0	**4**[a]	
4	Poland	3	1	0	2	2	5	−3	**3**	

HTTPS://WWW.FIFA.COM/WORLDCUP/GROUPS/

HTTPS://WWW.FIFA.COM/WORLDCUP/MATCHES/#KNOCKOUTPHASE