Analytics Edge Project: Predicting the outcome of the 2014 World Cup

Virgile Galle and Ludovica Rizzo

June 27, 2014

Contents

1 Project Presentation: the Soccer World Cup

The FIFA World Cup is an international association football competition contested by senior's men national teams. The championship has been awarded every four years since the inaugural tournament in 1930 (except in 1942 and 1946 where it was not held because of the Second World War). The current format (since 1998) of the tournament has two parts: a *qualifying phase* where teams compete among their continent and a *final phase* where the "best" 32 teams compete for the title.

For our project we focus on the outcome of the final phase. The final tournament has two stages:

- **Groups** The 32 teams are split into 8 groups of 4 teams. In every group, all the teams play once against each other and the two best teams qualify for the following phase. The outcome of a game can be the victory of a team or a draw.
- **Bracket** 16 teams get out of the group phase and enter a bracket. In every round of the bracket the teams play each other just once, thus there can not be draw games.

The next World Cup will be hosted by Brazil in June 2014. In our project, we use historical data (results from 1994 to 2010 world cups- 5 editions) to make predictions for the next World Cup. We tackle two type of questions:

- What will be the outcome of a single game (in the group and in the bracket phase)?
- In a more aggregate point of view, which teams will qualify for the bracket phase?

We split our data set into a training set (results from 1994, 1998, 2002) and a validation test (2006, 2010). We then use our model to predict for the 2014 World Cup.

2 Data Set

We collected data from the 5 last editions of the World Cup (1994 to 2010). In every world cup there are 64 games, thus our data set has $5 \times 64 = 320$ observations. This number is low compared to others data sets we used in class. We decided not to use more data for several reasons. First of all we believe that soccer has evolved a lot since the 80's and past competitions are not representative of nowadays' soccer. Secondly, before 1994 the FIFA Ranking didn't exist and this is a key feature for our models. For every edition we collected the results of every game and statistical information about every team involved.

2.1 Games

In every World Cup, there are 48 group games and 16 bracket games. For every game we recorded the teams involved, the score, the outcome ("1" if team 1 wins, "2" if team 2 wins and "X" if it is a draw) and the Stage (Group, or round in bracket).

In figure 1, we reported two sample rows of our "Games Results" data set for 1998.

	A	B	C	D	E	F
1	Stage	Team1	Team2	Goals1	Goals2	Winner
2	1st Round, Group A	Brazil	Scotland	2	1	1
3	1st Round, Group A	Morocco	Norway	2	2	X

Figure 1: Example of games results from 1998

2.2 Teams

For every team participating in a World Cup we collected the following data:

- **Name of the team**
- **FIFA Ranking and FIFA points at the beginning of the competition**
  FIFA (the international soccer association) makes a monthly ranking of national teams according to their performances. They give every team a score (computed with a fairly complicated weighted average of the team recent performances) and these scores are used to establish a ranking. We used the last score and ranking previous to the world cup. The way to compute the FIFA score changed over time, we thus scaled our FIFA points to be consistent
- **Participation In a Row**
  This variable counts the number of consecutive times the team participated in the final phase of the World Cup
- **Average Goals Scored and Conceded (Avggf and Avgga) during the qualifying phase**
- **Playoff**
  This is a binary variable that indicates if the team had to go throw a playoff to qualify for the World Cup. The three last variables are indicators of performance during the Qualifying phase.
- **Qualification Zone**
  This variable is a factor that represents the Qualification Zone of the team (Europe, South America, Asia, CONCAF (North and Central America), Africa)
- **Host and Continent**
  Binary variables that indicate if the team is hosting the competition or if it is in his continent.
- **U20**
  Best performance in the last 4 years of the Under 20 youth team

	A	B	C	D	E	F	G	H	I	J	K
1	Team	FIFApts	FIFARanking	PartInaRow	Avggf	Avgga	Playoff	Qualifzone	Host	Continent	U20
2	Argentina	60	6	7	1.44	0.81	0	SA	0	0	1
3	Austria	51	31	1	1.7	0.4	0	SA	0	1	64

Figure 2: Example of team statistics for the 1998 World Cup

In figure 2, we reported two sample rows of our "Teams statistics" data set for 1998.

2.3 Models

Using this data set we tackled two different problems:

- **What is the issue of every game?**
  We will answer this question in section 3. Our problem is split into two different sub problems. For the groups' games there are 3 possible outcomes (one team wins or a draw). We will use a generalization of logistic regression (ordered logistic regression). For bracket's games, there are just two possible outcomes. We will use logistic regression to predict the winner.
- **Which teams will qualify for the bracket phase?**
  We will answer this question in section 4. For every group, we use logistic regression to predict which two teams are more likely to qualify for the bracket phase.

In the two following sections, we will present the models we used and their performances on the train set and the validation test. In the last section we present our models' predictions for the 2014 World Cup.

3 Models to predict the outcome of each game

In this section, we tried to predict the outcome of each game in the group and bracket stage. We will first illustrate the shape of the data set we used and then present the models we used and their performances.

3.1 Data Set

In order to predict the outcome of every game we merged our "Games Results" and our "Teams Statistics" data files. We first built a data frame where every row was a game and with the following columns:

- Name of Team 1
- Name of Team 2
- Outcome of the Game (1,X or 2)
- Statistics of Team 1
- Statistics of Team 2

Our independent variables were the Statistics of the two teams and our dependent variable was the outcome of the game.

When we first ran our models we realized that we needed symmetric data. Intuitively, if we ask our model what will be the outcome of a game where $(Team1 = France, Team2 = Italy)$ or of a game where $(Team2 = Italy, Team1 = France)$ it should give the same result. In the case of a logistic regression, this mathematically translates into the fact that the coefficients for the corresponding features of Team 1 and Team 2 have to have the same absolute value but opposite sign. Concretely, if the coefficient of "FIFARanking.Team1" is 0.5 we want the coefficient of "FIFARanking.Team2" to be -0.5.

To achieve this symmetry we decided to replace the Statistics of the two teams by their difference. For example, if $FIFARanking.Team1 = 3$ and $FIFARanking.Team2 = 10$, we add the feature $FIFARanking = 3 - 10 = -7$. The shape of our data frame becomes:

- Name of Team 1
- Name of Team 2
- Outcome of the Game (1,X or 2)
- Difference of Statistics

We used different models to predict the outcome of every game: CART trees, Random Forests, SVM and logistic regression. We found that the best results were given by models of the "logistic family". We will report in the two following paragraphs the best results we have.

3.2 Predicting the outcome of a game in the group

As we explained above, in group games there are three different outcomes. We use the following notations: "1" means that Team 1 won, "X" represents a draw and "2" means that Team 2 won. In our data set these three classes are more or less equally populated: there are approximately 33% of "1", 33% of draws and 33% of "2".

We use a generalization of the logistic regression called "ordered logit".

Ordered logistic regression The ordered logistic model allows to capture the fact that there are more than 2 possible outcomes and that these outcomes are ordered. For example, in our case, taking the perspective of Team 1 the outcomes are ordered in the following (increasing) way: "2" lose,"X" draw, "1" win. Let's call $j \in$ ["2","X","1"] the level and y the outcome. Let's say that $y \leq j$ if $y \in$ ["2", j"] (where the set is ordered). (For example $t \leq$ "X" $\Leftrightarrow y \in$ {"2","X"}). Let x be the features of the model.

In the classical logistic regression, the odds that $y = 0$ are

$$Odds(y = 0) = e^{\alpha - \beta.x}$$

where α is an intercept and β is the vector of coefficients.

The ordered logistic is built in the same way, the odds that $y \leq j$ are

$$Odd(y \leq j) = e^{\alpha_j - \beta.x}$$

Let's remark that the coefficients β don't depend on the level but the intercept α_j does.

Significant variables
We ran our model using the function clm in the package ordinal. The regression results are in table 3. We found that the significant variables were FIFARanking, PartInaRow and Continent. The signs of the coefficients make sense: a low FIFARanking difference means that Team 1 has higher chances to win (thus the coefficient has to be negative), a high PartInaRow difference means that Team 1 participated more often than Team 2 and statistically a team playing in his continent has more chances to win.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
FIFARanking -0.023388   0.007877  -2.969  0.00299 **
PartInaRow   0.097836   0.030984   3.158  0.00159 **
Continent    0.733048   0.225714   3.248  0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
2|X  -0.8606     0.1847  -4.659
X|1   1.1304     0.1913   5.910
```

Figure 3: Ordered logistic: significant variables

Performances
Here are the confusion matrix and the accuracy of our model. The baseline is a model that predict randomly with probability 33%.

	2	X	1
2	21	19	8
X	12	38	20
1	3	26	33

Figure 4: Confusion matrix on train set

	Accuracy
Train Baseline	33%
Train Model	**37%**
Test Baseline	33%
Test Model	**50%**

Figure 5: Accuracy

We can remark that it is really hard to predict when a game is a draw. This makes sense on a intuitive point of view: there are a lot of games were we are confident that a team will win but we are never really confident that it will be a tie. If we consider only the games where we predict that the game is not a draw (first and last columns of the confusion matrix) the results are really satisfying. Looking just at a subset of games makes sense on a "betting perspective", if our aim is to bet we can select just a subset of games and discard the others.

	2	1
2	**21**	8
X	12	20
1	3	**33**

Figure 6: Confusion matrix on train set (games where we do not predict X)

	Accuracy
Train Baseline	33%
Train Model	**56%**
Test Baseline	33%
Test Model	**48%**

Figure 7: Accuracy (games where we do not predict X)

3.3 Predicting the outcome in a game in the bracket

The main difficulty in predicting the outcome of a group's game is the possibility of a draw, that as we saw is really hard to predict. In this section we try to predict games in the bracket where the outcome is binary, our performances here a more satisfying than in the previous part.

Model and results In this part, we use a classical logistic regression to estimate the outcome of a game. The significant variables are the same as in the previous part.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.17107    0.19373  -0.883 0.377229
FIFARanking -0.02106    0.01038  -2.028 0.042550 *
PartInaRow   0.13476    0.03704   3.638 0.000274 ***
Continent    1.05863    0.31099   3.404 0.000664 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8: Logistic regression for the bracket: significant variables

	Accuracy	AUC
Train Baseline	50%	50%
Train Model	**71%**	**78%**
Validation Baseline	50%	50%
Validation Model	**69%**	**70%**

Table 1: Accuracy and AUC

Here the baseline is random and thus has an of 50%. We can see that the performance of our model is really satisfying, both the AUC and the accuracy strongly outperform the baseline.

4 Logistic Regression for outcome of Group Stage

As we just noticed in Part 3, the ordered logistic has several assets. However it seems that it is very difficult to predict accurately when a draw occurs. Moreover, it seems that we have a very good accuracy in predicting games in the bracket since ties cannot happen. Hence the motivation of predicting the outcome of a group is the following : to predict entirely a World Cup, we do it in two stages : First, we can predict which two teams of each group will qualify for the final bracket and then the result of each game in the bracket. In order to predict accurately the outcome of a group, we are going to consider not each game by itself but only the features of every team that is in the group.

4.1 A New Database

Since we don't want to use the games to predict the outcome of a group, we need a new data base using the features of each team described in Part 2 and the list of each group. For each World Cup, the database has 32 observations (24 for 1994) which corresponds to the number of teams. Every row of our database corresponds to a team in a World Cup. We report its name and its features, followed by the features of the 3 other teams in the same group and a binary variable that indicates whether the team qualifies for the bracket. We denote "Team.1" "Team.2" and "Team.3" the other teams that are in the same group as "Team". Here is the shape of our data set:

- **Team** (str) : Name of the team
- **FIFApts** (num) : The FIFA points normalized like in Part 2
- **PartInaRow** (int) : The consecutive number of participation in the World Cup
- **Team.1** (str) : Name of a team in Team's group
- **FIFApts.1** (num) : The FIFA points also rescaled of Team.1
- **Qualifzone.1** (int) : The Qualifying zone of Team.1 . In Part II, it was introduced as a string. Here we model it by integers ; "SA" (South America) is replaced by 1, "AS" (Asia) by -1 and all the other ones by 0 (this choice will be explained further down)
- Same 3 features for **Team.2** and **Team.3**
- **OutGroup** (int) : Binary variable. 1 if Team qualified for the final bracket, 0 if not.

In this case,the train set is going to be every team participating in the 1994, 1998 and 2002 World Cups (88 observations). The validation set is the 2006 and 2010 World Cups (64 observations) and our test set 2014.

4.2 Evaluation of performance of the model

As in Part 3 , we will evaluate the performance of our model comparing it to benchmarks. Away of measuring our model is the logloss measure. Since the output of our model are probabilities, minimizing the logloss helps us on understanding if our model is 'risky' meaning that it gives high probability even if it's not sure. The formula of the logloss is the following:

$$Logloss(\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(\hat{y_i}) + (1-y_i)log(1-\hat{y_i}))$$

Since our model returns probabilities, then we say that the two teams of each group that have the highest probabilities will pass the group phase.

4.3 Baselines

In order to compare our model to a benchmark, we propose two types of benchmarks. The first one is the 'random' answer. Each team has 0.5 probability of qualifying for the bracket. The other one is a 'smarter' baseline which takes into account the seeds (in terms of FIFA ranking) . The first seed has 80% chance of qualifying, the second 60%, the third 40% and the last 20%. This corresponds to a rational behavior assuming that the FIFA ranking assesses properly the level of each team. The results are available in table 3 to compare with our model.

4.4 Logistic Model

After having tried CART Tree, Random Forests, SVMs and Logistic Regression, it appears that the last one has better performances. One major issue for all models was the fact that our data base is actually asymmetric like we noticed in Part 3. We know that all coefficients corresponding to all features of Team.1,.2 and .3 should be the same. In order to recover this, we permuted the columns and created several logistic models, saved their coefficients and averaged them. With this method, we find exactly the same coefficients. Notice here that we didn't take the difference like in Part 3. Another remark is that we found here that all coefficients for the Qualifzones were the same except for Asia and South America, hence the decision to change the nature of the variable Qualifzone. The coefficients we obtain are in Table 2

Variables	Coefficient
FIFApts	0.578148
PartInaRow	0.171173
FIFApts.1	-0.07393
Qualifzone.1	0.39428
FIFApts.2	-0.07393
Qualifzone.2	0.39428
FIFApts.3	-0.07393
Qualifzone.3	0.39428
Intercept	-0.44935

Table 2: Coefficients of the Logistic Regressions

All the signs of coefficients are easily interpretable. For instance, the more FIFApts Team has, the more likely it has to qualify and this make sense. Another interpretation to make is that teams of South America are more likely to qualify which is also a well known fact in the Soccer world.

4.5 Results

The results concerning the prediction of the outcome of Groups are summarized in Table 3 . Our model works very well compared to the random baseline and better than the smarter baseline in most of the cases and for all performance measures. In addition, it has another asset that is not outlined here : it predicts differently than the baseline. Sometimes it says that seeds lose when it's not the case, but it also know when a non easily predictable team qualifies which can be a real help while betting against bookmaker for instance.

Set	Predictions	Accuracy	Logloss
Train	Random Baseline	0.5	0.693
Train	Smart Baseline	0.729	0.596
Train	Logistic	0.757	0.603
Validation	Random Baseline	0.5	0.693
Validation	Smart Baseline	0.656	0.647
Validation	Logistic	0.687	0.620

Table 3: Results of the outcome of the Groups

5 Predictions for the 2014 World Cup in Brazil

To summarize, in the last two sections we presented three models: a first model that predicts the outcome of a game in the groups stage, a model that predicts the outcome of a game in the bracket phase and a model that

predicts which teams, in every group, will qualify for the bracket phase. We will now apply our three models to make predictions for the 2014 World Cup.

We have two different ways to predict the outcome of the group phase:

1. We can use the model in section 4 that gives the probability that every team is qualified. We will denote this method "OutGroup"
2. We can use the ordered logistic model in section 3 and compute the expected number of points earned by every team in the group phase. In every group, the two teams with the greatest number of points qualify. We will denote this method "Games".

In table 6 in the appendix, we present the results of our two models. We can see that they agree most of time. They disagree on groups that are "hard to predict", where the teams performances are really close to each other.

We then used the predictions in the groups to build a bracket. Since our two models don't always agree we have two different brackets for the "Round of 16 phase". Starting from these two brackets we use the logistic regression of section 3 to predict the outcome of every game. Our results are in in the trees 4 and 7 in the appendix. For every game, we wrote in bold font the name of the expected winner and we reported the probability of this outcome. We can see that our models agree starting from the semifinals.

6 Conclusion

To summarize our work, we used historical data of soccer World Cup to build models to predict the outcome of the 2014 World Cup in Brazil. We tried to answer two different type of questions. First of all, what will be the outcome of every single game (in the groups or in the Bracket phase)? Secondly, in a more aggregate way, which teams will pass the group phase and qualify for the bracket? We found that the best performing models were part of the "logistic regression" family. The significant variables that are able to predict a team's performances are the FIFA Ranking, the Number of consecutive participation to the competition, if the team is playing in its own continent and the qualification zone it comes from.

We realized that predicting games in the group phase is way harder than in the bracket phase because of the possibility of draws, we then used aggregated features to predict which teams will qualify for the bracket instead of trying the predict every single game of the bracket. On a betting perspective, we realized that we are almost never confident that a game will be a draw, but we can be really confident that one team will win. Thus, if we were to bet, we would use just games where we predict that one team is going to win.

Finally, we made predictions for the 2014 World Cup and we can't wait to see if we are right!

Appendix

Table 4: Bracket predicted starting with the method OutGroup

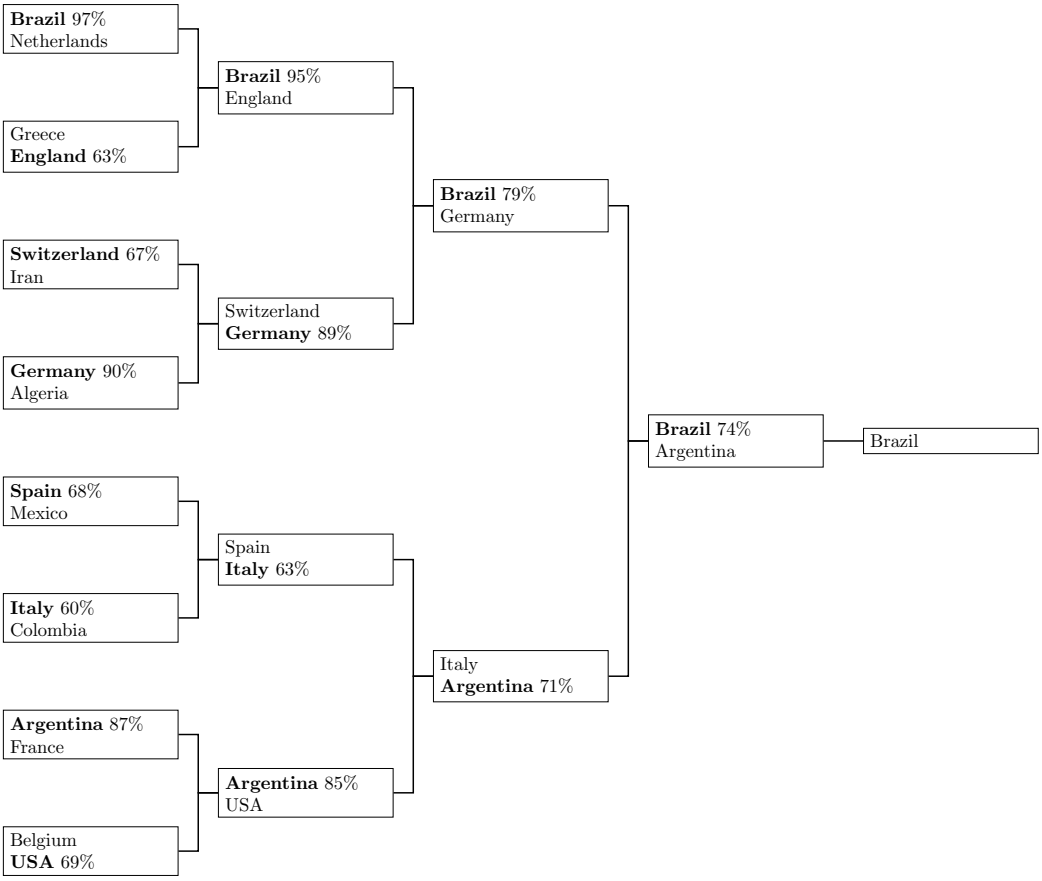

Table 5: Bracket predicted with the method "Games"

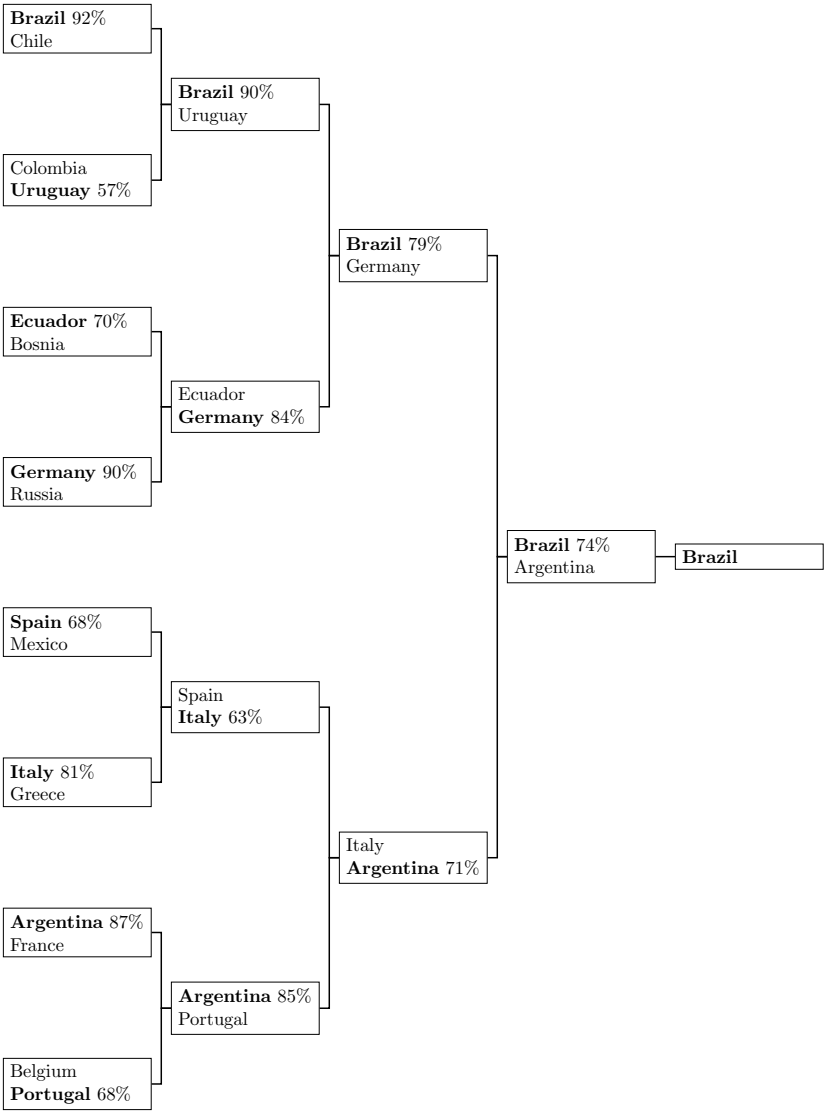

Table 6: Ranking in the groups predict by our two models

A	OutGroups	Prob	Games	Pts
1	Brazil	97.3%	Brazil	8.1
2	Mexico	57.1%	Mexico	3.8
3	Croatia	39.7%	Croatia	2.8
4	Cameroon	26.8%	Cameroon	1.7

B	OutGroups	Prob	Games	Pts
1	Spain	93%	Spain	5.7
2	Netherlands	65.6%	Chile	5.1
3	Chile	64.2%	Netherlands	3.6
4	Australia	31.3%	Australia	1.5

C	OutGroups	Prob	Games	Pts
1	Greece	60.4%	Colombia	5.6
2	Colombia	48.1%	Greece	3.9
3	Ivory Coast	35.4%	Ivory Coast	3.5
4	Japan	31.6%	Japan	2.5

D	OutGroups	Prob	Games	Pts
1	Italy	90.9%	Italy	5.6
2	England	52.4%	Uruguay	4.8
3	Uruguay	45.3%	England	3.7
4	Costa Rica	36.4%	Costa Rica	1.7

E	OutGroups	Prob	Games	Pts
1	Switzerland	53.7%	Ecuador	4.4
2	France	49.8%	France	4.3
3	Honduras	35.7%	Switzerland	4.1
4	Ecuador	27.6%	Honduras	2.6

F	OutGroups	Prob	Games	Pts
1	Argentina	84.5%	Argentina	7.6
2	Iran	40.5%	Bosnia	3.3
3	Bosnia	27.9%	Iran	2.7
4	Nigeria	20.2%	Nigeria	2.5

G	OutGroups	Prob	Games	Pts
1	Germany	96.6%	Germany	6.4
2	USA	68.2%	Portugal	3.8
3	Portugal	63.1%	USA	3.6
4	Ghana	39.7%	Ghana	1.9

H	OutGroups	Prob	Games	Pts
1	Belgium	63%	Belgium	4.1
2	Algeria	61%	Russia	4.0
3	South Korea	47.5%	Algeria	3.8
4	Russia	39.4%	South Korea	3.3

Table 7: Bracket predicted after the Group Phase