FIFA World Cup 2018 Predictor

Pallav Gupta
NUID: 001268703
College of Computer and Information Science
Northeastern University
Boston, MA 02115
gupta.pa@husky.neu.edu

Piyush Goel
NUID: 001236188
College of Computer and Information Science
Northeastern University
Boston, MA 02115
goel.p@husky.neu.edu

Abstract

FIFA world cup is one of the biggest sporting event in the world and it happens once every 4 years. Every 4 years, we see some new teams on the draw and some dominant teams who have been playing since a long time and are regular in the world cup appearances. There are a number of factors that influence a team's preparation or training for the world cup. It can range from availability of good players to the association's budget for a particular team etc. It would be interesting to find out how teams have evolved over time and that can further give us an insight into the next world cup and the predictions for it.

1 Introduction

Being a FIFA enthusiast we wanted to create a model that builds a world cup predictor and predicts the match outcome based on team and player ranking. Also, we want to visualize how team rankings have evolved over years and its correlation with players performance.

Predicting the match outcomes are particularly challenging because of the following reasons:

- Ground/Weather Conditions
- Team morale/bonding
- Player mindset/confidence/injuries/practice/rest
- Coach and strategy against different teams
- Opponents game play understanding
- Sheer LUCK!

2 Related work

Since World Cup 2018 is already over, we can find a lot of World Cup predictions done by multiple scholars around the world.

Zeileis, Leitner, and Hornik (2018) published a paper "Probabilistic forecasts for the 2018 FIFA World Cup based on the bookmaker consensus model" in which they predicted that FIFA world cup 2018 will be conquered by Brazil with an odd of 16.6% followed by Germany and Spain.

Andreas Groll, Christophe Ley, Gunther Schauberger, Hans Van Eetvelde (2018) published a paper "Prediction of the FIFA World Cup 2018 - A random forest approach with an emphasis on estimated team ability parameters" predicted that Spain will swoop the FIFA World Cup 2018 with a probability of 17.8% followed by Germany, Brazil, France and Belgium.

Similarly, Audran, Bolliger, Kolb, Mariscal, and Pilloud (2018) predicted that Germany will win with a probability of 24% followed by Brazil and Spain.

We can note that Germany, Brazil and Spain are the most favored teams in all the above work (not necessarily in that order).

One of the major difference between the above mentioned papers and our methodology is that the above mentioned papers use Bookmakers data as a major feature in order to predict the outcome of a match. Bookmakers data indicates the odds of betting in a match. Bookmaker's odds specify the probability that bookmaker expect for a team to win. While, in our model the major features are rank and point difference of teams and their performance over years.

3 Dataset

All our data was taken from kaggle. We are using 4 datasets.

1) FIFA rankings from 1998 to 2018
This dataset has rankings, which change monthly for a team. The rankings have always been a decent predictor of the team's performance[6].

2) International Soccer Matches from 1948 to 2018
This dataset has information about different matches that happened between different international teams over the years.
We can use this to model on how difference in ranking, points and current rank affects the outcome of a match.

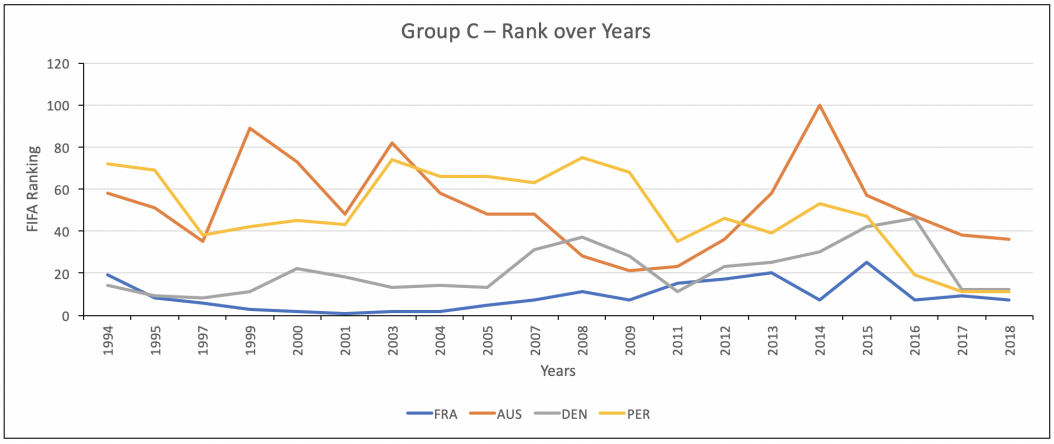
3) FIFA world cup dataset
This dataset has the matches scheduled according to FIFA world cup 2018. We can use this to predict the winner by round by round simulations and similarly measure accuracy.

4) FIFA 18 Complete Player Dataset
This dataset contains varying information about players. It has information about player's personal attributes (Nationality, Club, Photo, Age, Value etc.), player's performance attributes (Overall, Potential, Aggression, Agility etc.) and player's preferred position and ratings at all positions.
We used this dataset to aggregate the potential scores of top 50 players from a national team and store it as team's potential score.

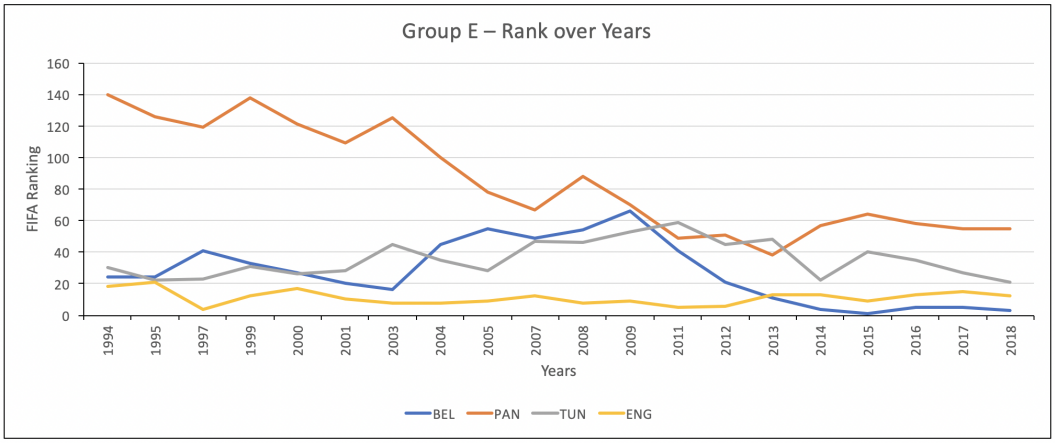
4 Experiments

4.1 Exploratory data analysis results

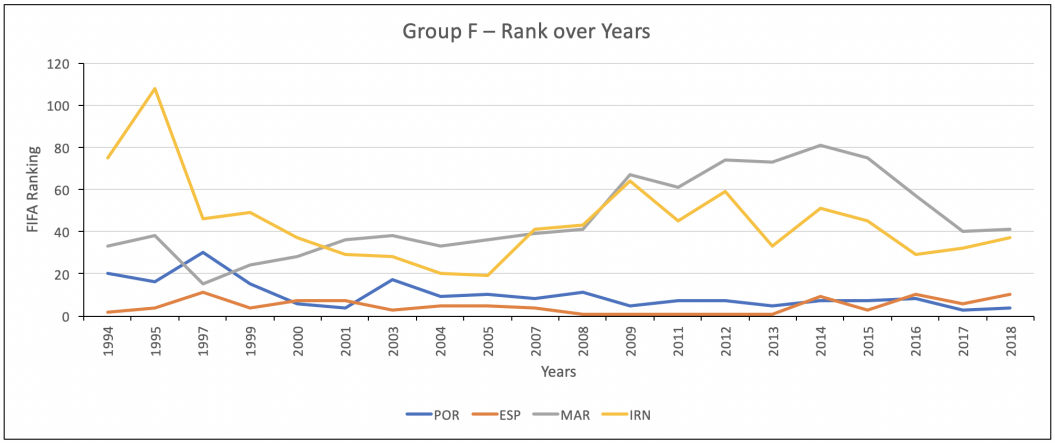
To understand the growth of teams participating in World Cup 2018, we plotted a Rank over Years chart for various teams participating in the World Cup 2018.



If we look at the chart above, France (the world cup winners) have the best rank in their group.



If we look at Belgium, they came 3rd in the competition and have a pretty good ranking.



Spain were the champions in 2010 and we can see how consistent they have been with their rankings even before 2010.

By plotting the Rank over Years chart for all the teams, we learned that each group contains teams with consistently good performance and higher ranking since 1994, while some teams

have variant ranking over years.
By looking at these charts we understood that teams which are consistent over years have more chances or winning a match against teams with inconsistencies. Therefore, we established the team's ranking to be our baseline.
We trained our model using the ranks of teams since 1994 but we found that rank as a feature was not enough.

5 Feature Extraction

1) Point and Rank Difference
We used the rankings from the dataset and merged them with matches to compute the rank and point difference for the two playing nations in a match.

2) Average Rank
For a match, we also computed average rank of the match. This is because the teams which are closer in ranks have a close match and the difference in the goals scored is less as compared to the teams which are far apart in the rankings.

3) Team Potential difference
Another feature that we generated was the potential difference of two teams in a match. The complete player dataset gives the information about player's performance attributes but not for a team. We aggregated the attributes for top 50 players of each team to get a team's potential score. From this, we calculated the difference of team's potential score in a match and included it in our models.

4) Friendly match?
Generally, friendly matches are a way for a team to experiment new things, new players. They don't necessarily play to win a match but to improve their team's coordination, try out different tactics. So this also affected the outcome of a match and hence, we used it in our model as one of the features.

6 Method/Model

In this section, we will discuss multiple approaches used for FIFA World Cup 2018 predictions. We used 3 approaches to model our data. We will see the difference in results for these models in Section 7. We used the sklearn library with Python for implementation purposes.

6.1 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). We used simple Logistic Regression on average rank, rank difference, point difference and team's potential difference to build our model. We used the confusion matrix and AUC score to interpret the models capability as shown in 7.

6.2 Random Forest Classifier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. We use Random Forest Classifier as it is helpful for mixture between statistics and machine learning. We use the same approach as logistic regression and train our model with the similar data to use random forest classifier.

6.3 Gradient Boosting

From Wikipedia, Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. We used

the gradient boosting model with the same data and noticed that it doesn't necessarily outperforms the logistic regression model as seen in section 7

7 Results

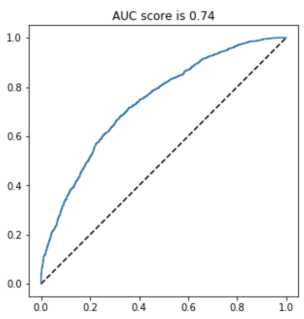
7.1 Quantitative Results

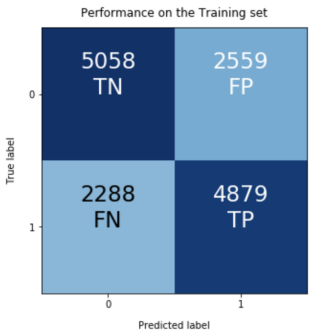
For our evaluation, we used different metrics.
1) We plotted the ROC curve and calculated the AUC score. AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

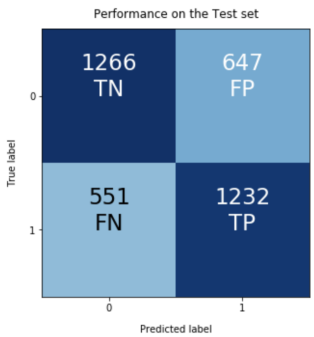
2) Confusion matrix is the go-to evaluation for classification models. The confusion matrix tells us the absolute number of True classifications (True Positives, True Negatives) and False classifications (False Positives, False Negatives)

3) Using the confusion matrix, we calculated the precision and accuracy of our models. Accuracy is the fraction of predictions our model got right. Precision attempts to calculate the proportion of positive identifications that were actually correct.

7.1.1 Logistic Regression

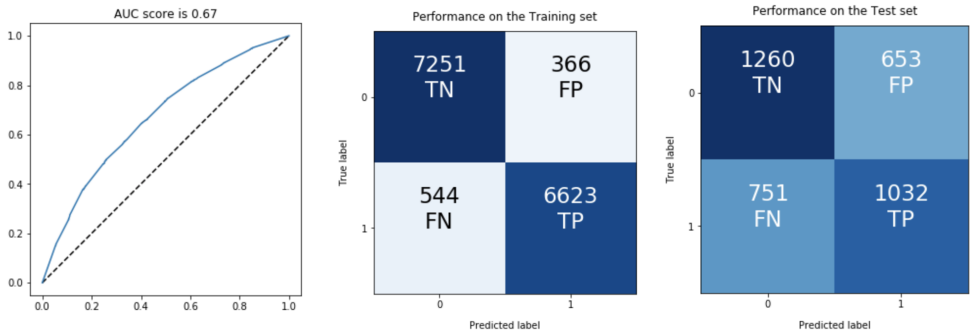






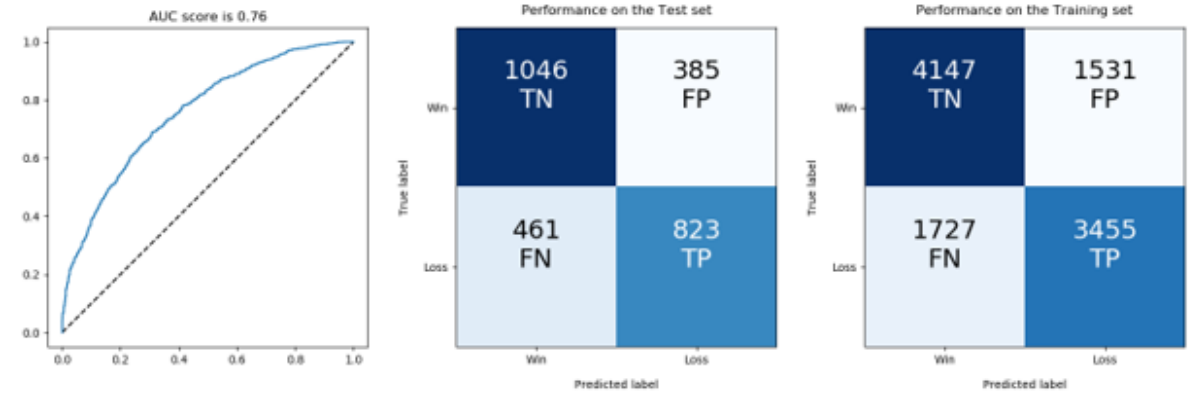
AUC = 0.75
Accuracy = 0.69
Precision = 0.66

7.1.2 Random Forest Classifier



AUC = 0.69
Accuracy = 0.65
Precision = 0.56

7.1.3 Gradient Boosting



AUC = 0.76
Accuracy = 0.69
Precision = 0.64

We see here that Logistic regression and gradient boosting both give a similar performance with Logistic Regression being slightly better.

7.2 Qualitative Results

Our final model predicted the following outcome for FIFA World Cup 2018.

Group Stage:

Team	Matches	Win	Draw	Lose
URU	3	3	0	0
EGY	3	2	0	1
KSA	3	0	1	2
RUS	3	0	1	2

Fig 1: Group A

Team	Matches	Win	Draw	Lose
POR	3	1	2	0
ESP	3	1	2	0
MAR	3	0	3	0
IRN	3	0	3	0

Fig 1: Group B

Team	Matches	Win	Draw	Lose
FRA	3	1	2	0
DEN	3	1	2	0
PER	3	1	2	0
AUS	3	0	0	3

Fig 1: Group C

Team	Matches	Win	Draw	Lose
ARG	3	1	2	0
CRO	3	1	2	0
ISL	3	0	3	0
NGA	3	0	1	2

Fig 1: Group D

Team	Matches	Win	Draw	Lose
BRA	3	1	2	0
SWI	3	1	2	0
CRC	3	0	3	0
SRB	3	0	1	2

Fig 1: Group E

Team	Matches	Win	Draw	Lose
MEX	3	1	2	0
GER	3	0	3	0
SWE	3	0	3	0
KOR	3	0	2	1

Fig 1: Group F

Team	Matches	Win	Draw	Lose
ENG	3	2	1	0
BEL	3	1	2	0
TUN	3	1	1	1
PAN	3	0	0	3

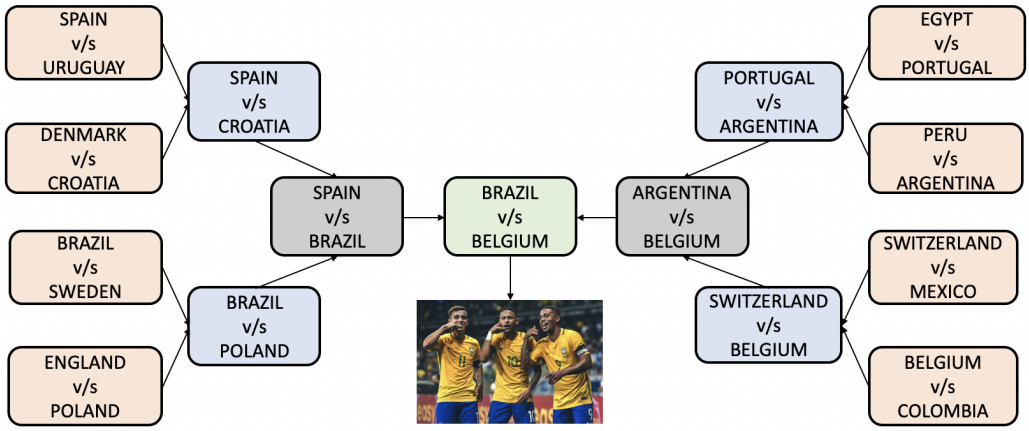
Fig 1: Group G

Team	Matches	Win	Draw	Lose
COL	3	2	1	0
POL	3	1	2	0
SEN	3	1	1	1
JPA	3	0	0	3

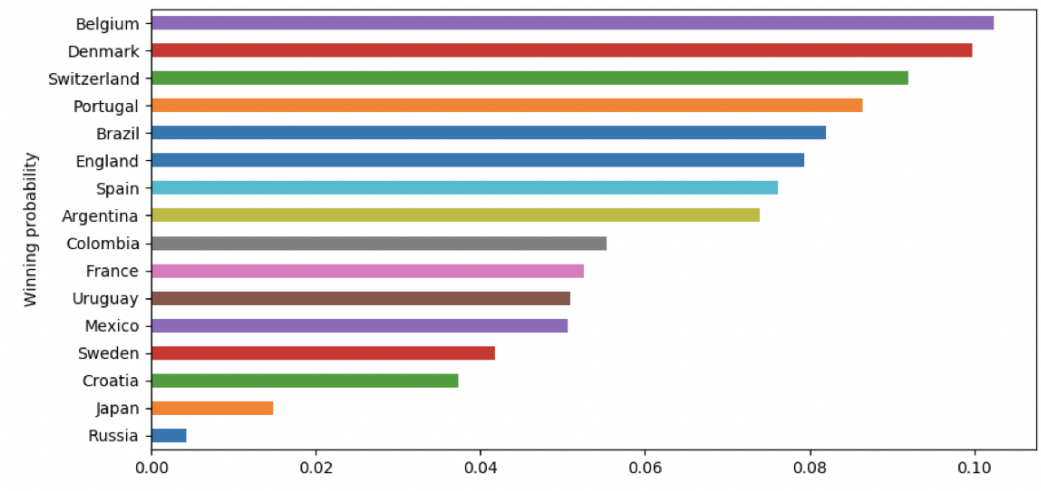
Fig 1: Group H

Our final model predicts 5 group stage results correctly out of 8 group stages.

Knockout Stage:



Our final model predicted that Brazil will win the FIFA World Cup 2018. But when we ran 10,000 simulations, winning probability for Belgium is highest.



References

[1] World FIFA rankings from https://en.wikipedia.org/wiki/FIFA_World_Rankings

[2] World FIFA rankings from https://www.fifa.com/fifa-world-ranking/procedure/men.html

[3] Dataset courtesy of Tadhg Fitzgerald, Mart Jürisoo, and Nuggs

[4] Andreas Groll, Christophe Ley, Gunther Schauberger & Hans Van Eetvelde (2018) Prediction of the FIFA World Cup 2018 - A random forest approach with an emphasis on estimated team ability parameters. *https://arxiv.org/abs/1806.03208*

[5] Zeileis, Leitner, and Hornik (2018) Probabilistic forecasts for the 2018 FIFA World Cup based on the bookmaker consensus model. *https://econpapers.repec.org/paper/innwpaper/2018-09.htm*

[6] Complete player dataset from *https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset*