Machine Learning Prediction of the 2018 football World Cup.

Martin Millon

May 31, 2018

1 Introduction

Football is considered as a rather unpredictable sport. The predictability can be computed either by comparing the odds of the bookmakers and the actual winner of the game, either by defining the favorite as the team having the more win in the current season and comparing with the outcome of the game. The later approach was used in [Ben-Naim et al., 2006] and it was shown that the English Football League is the least predictable league compared to the 4 US League of basketball, baseball, ice hockey and American football. The favorite is winning in only 54% of the case for football whereas it is the case in 64% for basketball.

However, the large amount of data collected at each match is now easily available and make the problem of sport predictions very well suited for machine learning. It corresponds to a typical classification problem between three classes corresponding to the possible outcomes of the games(home win, draw or away win). The problem can also be treated as a regression problem([Delen et al., 2012]). In this case the algorithm aims at predicting the wining margin, which corresponds to the score difference between the winning and loosing team.

A well chosen data set containing the past results of a team or the statistics of the previous games can be used to train Artificial Neural Network (ANN, see [Bunker and Thabtah, 2017] for review). Recent works have reached an accuracy of 67.5% at predicting the result of the NFL (Rugby League), AFL (Australian Rules football), Super Rugby (Rugby Union), and English Premier League Football (EPL) using a (20-10-1) architecture Neural Network([McCabe and Trevathan, 2008]). This is to be compared with the 60-65% of correct predictions of the human experts on the same data set. For football only, where the probability of the game to end on a draw is much higher, the ANN reached 54.6% accuracy at predicting the results of the English Premiere League, whereas human experts stand between 50 and 55% of correct prediction.

In this project, we aim at predicting to outcomes of the group phase of 2018 World Cup. We tested two different classifier, namely a random Forest Classifier and a Artificial Neural Network. Both algorithms were trained on a data set containing all the results of the international football games played since 1872.

2 Data

The data set "International football results from 1872 to 2018" is freely available on kaggle[1]. It consists on a database containing the information of 38930 football games. It includes the competing teams, the score, the date, the type of competition, the city and the country where the game took place and if the match was played in a neutral country.

The odds were taken from the website oddsportal.com[2] which compile the odds prediction of several bookmaker website.

3 Algorithm description

Data Pre-processing : The database contains categorical data such as the name of the competing teams and the type of competition. This needed to be encoded. We use a LabelEncoder

[1] https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017

[2] http://www.oddsportal.com/soccer/world/world-cup-2018

for the Random Forest Classifier and a One Hot Encoder for the Neural Network in order not to introduce any ordering of the data. We notice that both the Random Forest Classifier and the ANN perform better on a reduced data set containing only 4 features, i.e. the home team, the away team, the type of competition and if the game was played on a neutral ground. The remaining features, namely the date, city and country where the match was played are considered irrelevant and thus removed form the data set.

The data were split in a training set and in a test set according to 3 different strategies :

1. Randomly select 80% of the data for the training set and 20% for the testing set
2. Reduce the data set to only the game played after 2000 and randomly split the remaining sample into training and testing set (80/20)
3. Use all the data excpet the game played during the 2014 World Cup in the training set. Use the data of the 2014 World cup as a testing set.

These strategy aim to test if the most recent data are more relevant to predict the outcome of the future game and if it possible to optimize the algorithm for a particular type of competition, namely the World Cup.

Artificial Neural Network : Several networks have been tested and the configuration providing the best results consists in a Input layer of 4 parameters, an hidden layer with 10 nodes and an output layer containing 3 nodes, one for each possible outcome of the game. We used a softmax activation at this latter stage. Three of the four input parameters are categorical parameters and are one hot encoded. This operation results in a 579 parameters input layer. The accuracy is used as the metric to train the neural network.

Random Forest Classifier : A Random Forest Classifier was also implemented to compare the results with the ANN. The algorithm was optimized on a grid of parameters. The combination yielding the higher accuracy is a maximum depth of the tree of 100, the minimum sample per leaves of 10, using the Gini criterion and bootstraping.

4 Results

4.1 Artificial Neural Network :

After training, the ANN has an accuracy of 56% at predicting the winner or if the match will end on a draw. The mixing matrix is shown on figure 1.

The ANN tends to under-predict the number of games ending on a draw. This is possibly because the ANN learns not only about the relative level of team compared to each other but also from the characteristics of football in general, which sees less often the game ends on a draw than on a win. Interestingly, we observe that home team have a larger chance to win the game and this feature is well represented by the ANN.

Other splitting strategy : When reducing the data to games played only after 2000 (splitting strategy 2 in section 3), we reach an even higher accuracy of 58%. This tends to indicate that the most recent data are more relevant and a larger data set does not necessarily improves the ANN performances.

Additional features : We observed that the ANN needs a reduced number of features in input. We tried to train the algorithm on the data set which also contains the date and location of the games and it converged to a trivial solution, giving the same probability of win, draw and lose for all the teams. We conclude that the features should be selected carefully since adding unnecessary information seems to complicate the fitting procedure.

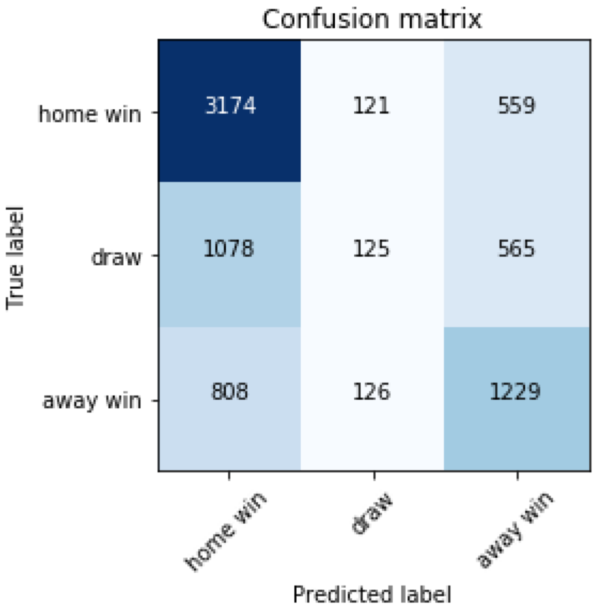


Figure 1: Mixing matrix of the Artificial Neural Network trained with the splitting strategy 1.

World cup 2018 predictions : The predictions of the Neural Network for the 2018 world Cup can be seen in the Appendix A. The algorithm predicts the same result as the bookmaker in 89.6% of the cases. The 5 games that differ from bookmaker predictions are Morocco - Iran, Serbia - Switzerland, Japan - Senegal, Poland - Colombia and England - Belgium. These 5 games do not have a clear favorite and can be considered as rather uncertain. In this sens, the prediction of the ANN seems rather plausible.

The expectation is also computed by multiplying the probability of victory predicted by the algorithm with the odds of the bookmakers. An expectation higher than one means that the bet should be taken.

4.2 Random Forest Classifier :

The Random Forest yielded less good results as the Artificial Neural Network with only 53% accuracy when trained on 80% on the data set and tested on the remaining 20% (splitting strategy 1 defined in section 3). The mixing matrix is shown on Figure 2.

Splitting strategy : We summarized the result of the different splitting strategy in table **??** and compared the accuracy with the ANN.

Splitting strategy	Neural Network	Random Forest Classifier
1	0.56	0.53
2	0.58	0.52
3	-	0.44

Table 1: Accuracy of the two classifiers according to the splitting strategy between training and testing set described in section 3.

World cup 2018 predictions : The Random Forest Classifier showed very poor results at predicting the outcomes of the World Cup 2018 with only 39.58 % of the game in agreement with the bookmakers predictions. We also observed several cases where the obvious favorite is not identified by the algorithm.

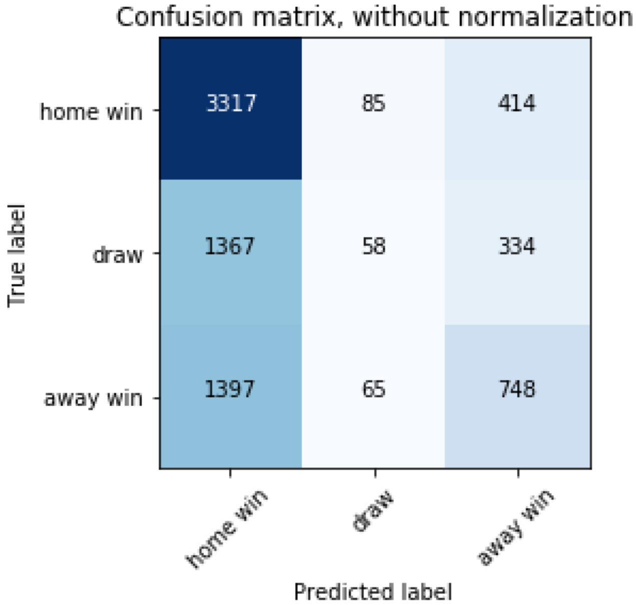


Figure 2: Mixing matrix of the Random Forest Classifier trained with the splitting strategy 1.

5 Conclusion

In this project, we used the records of 38930 football games to train to type of classification algorithm, namely a Random Forest Classifier and Artificial Neural Network with architecture (4-10-3). We tried several splitting strategy between the training set and the test set and reached the conclusion that the data taken after 2000 are more relevant and must be used with a 80/20 split between training and testing set. The Artificial Neural Network showed much better results than the Random Forest Classifier, with 58% of correctly predicted results (win, draw or lose). This is slightly higher than the 55% accuracy that human experts can reach. The predicted results for the group stage of the 2018 World Cup seems plausible with 89.6 % of the predictions in agreement with the bookmakers favorite. But fortunately, the result of a particular football games is still largely unpredictable. What will be the point of watching football if the outcomes can be predicted by a machine ?

References

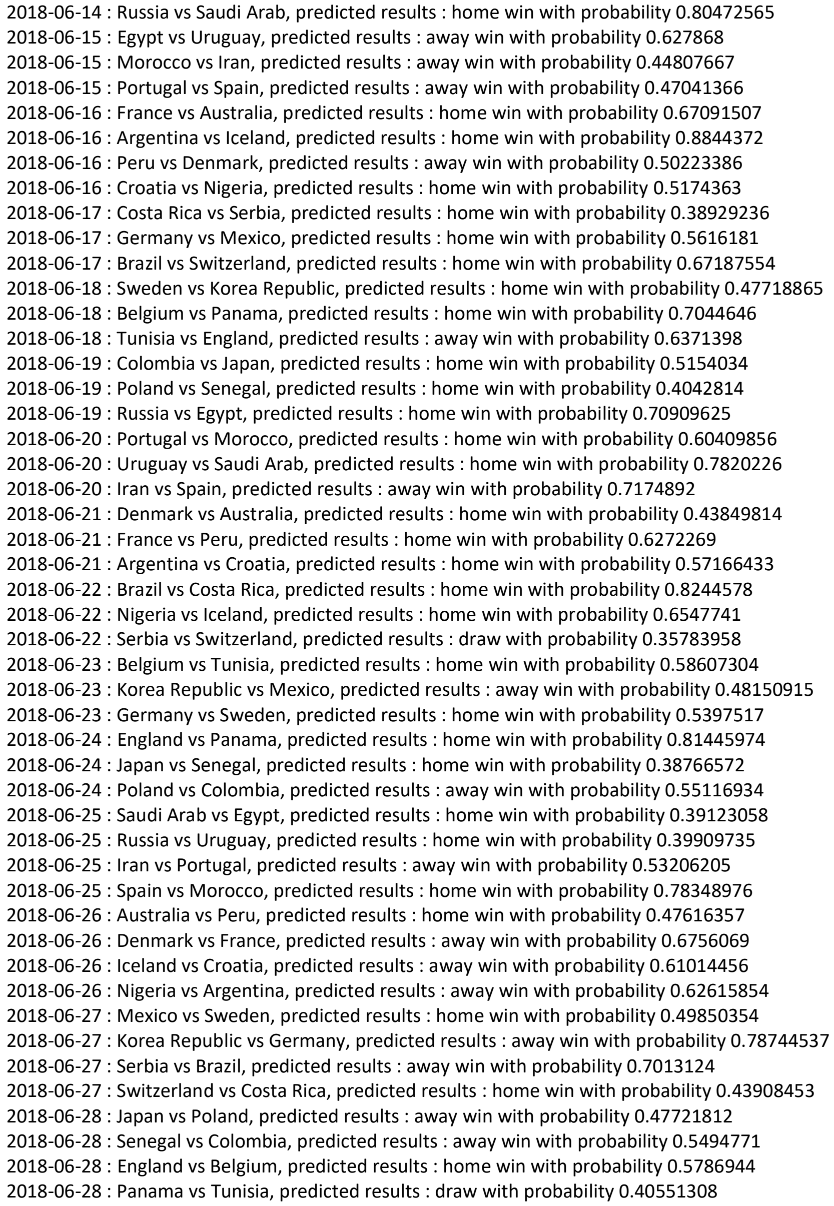
[Ben-Naim et al., 2006] Ben-Naim, E., Vazquez, F., and Redner, S. (2006). Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2(4).

[Bunker and Thabtah, 2017] Bunker, R. P. and Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied Computing and Informatics.*

[Delen et al., 2012] Delen, D., Cogdell, D., and Kasap, N. (2012). A comparative analysis of data mining methods in predicting ncaa bowl outcomes. *International Journal of Forecasting*, 28(2):543–552.

[McCabe and Trevathan, 2008] McCabe, A. and Trevathan, J. (2008). Artificial intelligence in sports prediction. In *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*, pages 1194–1197. IEEE.

A Appendix : Predicted results of the ANN for the 2018 World Cup group stage.



```
2018-06-14 : Russia vs Saudi Arab, predicted results : home win with probability 0.80472565
2018-06-15 : Egypt vs Uruguay, predicted results : away win with probability 0.627868
2018-06-15 : Morocco vs Iran, predicted results : away win with probability 0.44807667
2018-06-15 : Portugal vs Spain, predicted results : away win with probability 0.47041366
2018-06-16 : France vs Australia, predicted results : home win with probability 0.67091507
2018-06-16 : Argentina vs Iceland, predicted results : home win with probability 0.8844372
2018-06-16 : Peru vs Denmark, predicted results : away win with probability 0.50223386
2018-06-16 : Croatia vs Nigeria, predicted results : home win with probability 0.5174363
2018-06-17 : Costa Rica vs Serbia, predicted results : home win with probability 0.38929236
2018-06-17 : Germany vs Mexico, predicted results : home win with probability 0.5616181
2018-06-17 : Brazil vs Switzerland, predicted results : home win with probability 0.67187554
2018-06-18 : Sweden vs Korea Republic, predicted results : home win with probability 0.47718865
2018-06-18 : Belgium vs Panama, predicted results : home win with probability 0.7044646
2018-06-18 : Tunisia vs England, predicted results : away win with probability 0.6371398
2018-06-19 : Colombia vs Japan, predicted results : home win with probability 0.5154034
2018-06-19 : Poland vs Senegal, predicted results : home win with probability 0.4042814
2018-06-19 : Russia vs Egypt, predicted results : home win with probability 0.70909625
2018-06-20 : Portugal vs Morocco, predicted results : home win with probability 0.60409856
2018-06-20 : Uruguay vs Saudi Arab, predicted results : home win with probability 0.7820226
2018-06-20 : Iran vs Spain, predicted results : away win with probability 0.7174892
2018-06-21 : Denmark vs Australia, predicted results : home win with probability 0.43849814
2018-06-21 : France vs Peru, predicted results : home win with probability 0.6272269
2018-06-21 : Argentina vs Croatia, predicted results : home win with probability 0.57166433
2018-06-22 : Brazil vs Costa Rica, predicted results : home win with probability 0.8244578
2018-06-22 : Nigeria vs Iceland, predicted results : home win with probability 0.6547741
2018-06-22 : Serbia vs Switzerland, predicted results : draw with probability 0.35783958
2018-06-23 : Belgium vs Tunisia, predicted results : home win with probability 0.58607304
2018-06-23 : Korea Republic vs Mexico, predicted results : away win with probability 0.48150915
2018-06-23 : Germany vs Sweden, predicted results : home win with probability 0.5397517
2018-06-24 : England vs Panama, predicted results : home win with probability 0.81445974
2018-06-24 : Japan vs Senegal, predicted results : home win with probability 0.38766572
2018-06-24 : Poland vs Colombia, predicted results : away win with probability 0.55116934
2018-06-25 : Saudi Arab vs Egypt, predicted results : home win with probability 0.39123058
2018-06-25 : Russia vs Uruguay, predicted results : home win with probability 0.39909735
2018-06-25 : Iran vs Portugal, predicted results : away win with probability 0.53206205
2018-06-25 : Spain vs Morocco, predicted results : home win with probability 0.78348976
2018-06-26 : Australia vs Peru, predicted results : home win with probability 0.47616357
2018-06-26 : Denmark vs France, predicted results : away win with probability 0.6756069
2018-06-26 : Iceland vs Croatia, predicted results : away win with probability 0.61014456
2018-06-26 : Nigeria vs Argentina, predicted results : away win with probability 0.62615854
2018-06-27 : Mexico vs Sweden, predicted results : home win with probability 0.49850354
2018-06-27 : Korea Republic vs Germany, predicted results : away win with probability 0.78744537
2018-06-27 : Serbia vs Brazil, predicted results : away win with probability 0.7013124
2018-06-27 : Switzerland vs Costa Rica, predicted results : home win with probability 0.43908453
2018-06-28 : Japan vs Poland, predicted results : away win with probability 0.47721812
2018-06-28 : Senegal vs Colombia, predicted results : away win with probability 0.5494771
2018-06-28 : England vs Belgium, predicted results : home win with probability 0.5786944
2018-06-28 : Panama vs Tunisia, predicted results : draw with probability 0.40551308
```