Predicting the Most Probable World Cup Tree with Random Forests - Reproducibility Challenge 2023

Sasha Denouvilliez-Pech
Software Engineering
McGill University, Montreal

Shidan Javaheri
Software Engineering
McGill University, Montreal

Taz Scott-Talib
Software Engineering
McGill University, Montreal

Reproducibility Summary

Scope of Reproducibility

This paper explores a random forest approach to predicting the outcome of the FIFA World Cup in 2018, and the impact of both the complexity of data available and the effect of repeated simulations on model performance. The approach is based on a paper (1) which uses a combination of random forests, Poisson regression and ranking methods to predict the expected number of goals a team will score in a given game, which is then used to determine the winner. However, this approach is beyond our time and complexity constraints. To explore three of the papers main claims, our paper implements a simpler model that uses random forests to predict the most probable winning team. The first claim we explore is that more complex processed features are better data separators and therefore more valuable than less complex and raw features for the a random forest model. Secondly, we explore the implicit claim that repeated simulation of the world cup stages leads to a model with higher predictive power than a single simulation. Finally, we investigate the claim that a combination of random forest, Poisson regression and ranking methods have a higher predictive power than random forests alone. This claim is explored by comparing the results of our simpler model with the results of the more complex one in the paper, since we could time and computational constrains meant that we could not implement the complex model.

Methodology

The code provided by the author of the paper was not very extensive, and despite belonging to the paper, contained data for the cricket world cup instead. This made obtaining the same data that the original paper uses very challenging. To overcome this obstacle, we follow and adapt a Kaggle tutorial (2) to the 2018 FIFA World Cup, to obtain a similar but more simple random forest model, as well as the data to train it. We use its simpler data to investigate the claim that more complex processed features are more valuable than raw ones in prediction. We then perform hyperparameter tuning on a random forest model, and adapt the code of the tutorial to perform 100 simulations of each world cup stage with this model, to explore the impact that repeated simulation had on final results. Finally, obtain the world cup bracket for the 2018 FIFA world cup, which we then compare to the bracket obtained with the more complex model. All of these experiments were performed on Google Colab, with no GPU acceleration, and 12.7 GB of RAM.

Results

The classification score, or accuracy, of our model was around 69% when trained and tested on a dataset of all international matches between the tournament studied (2018) and the one preceding it (2014). This is significantly superior to the score reported in the paper; however, this is due to the reduction to binary labels in our implementation, which when taken into consideration, makes our accuracy closely reflect that of the authors' model (53 vs. 55%). Our model's aggregated tournament bracket is very similar to the one we attempted to replicate, with 53% of projected matches and 73% of advancing teams identical, including all four semi-finalists.

What was easy

The code for our paper came from a Kaggle tutorial (2) for predicting the 2022 FIFA World Cup, that was very easy to follow and adapt to the 2018 world cup. Adjusting this code to make it more modular and to perform the experiments we needed, hyperparameter tuning and repeated simulation, was relatively simple as well.

What was difficult

The author's code was not very helpful (3), which made getting the data and reproducing the model difficult to do precisely, leading us to following a different tutorial. This made identifying the claims challenging as well. In addition, certain elements in the report, such as specifications of what percentages represented, were ambiguous.

Communication with original authors

We did not make contact with the original authors of this paper.

Submitted to ML Reproducibility Challenge 2023. Do not distribute.

1 Introduction

Sports, Football, and in particular the FIFA world cup, captivates billions of viewers all around the world. The results of each game are notoriously unpredictable, as there are an infinite number of factors that influence the outcome of a game, a major factor of which many would term ‘luck’. The challenge of predicting a football game is an enticing one, that has lead to many versions of games, and even entire industries. As a result, our paper investigates three major claims of the paper “Prediction of the FIFA World Cup 2018 – A random forest approach with an emphasis on estimated team ability parameters" (1), while attempting to replicate one of its major results - a prediction of the most probable world cup bracket. Through these claims, we explore the importance of feature engineering for model performance, the impact of repeated simulations, and the performance of a simple random forest model compared to a complex one.

2 Scope of reproducibility

The paper we are attempting to reproduce had some very interesting results regarding using random forests to predict the outcome of football games, but unfortunately did not have a very good code base available. In addition, the models used are beyond our time and complexity limits. Our scope of reproducibility thus aims to replicate the final result of this paper, a predicted world cup bracket for the 2018 FIFA world cup. We will do so with a simpler model and simpler data, to explore three of the claims the paper makes both implicitly and explicitly.

All three of these claims describe the ‘predictive power’ of the model used to determine the most probable world cup bracket, which we will measure both by comparing the world cup bracket to what actually occurred, as well as by calculating the classification rate of the models. The claims we will investigate are listed below, in the order they will appear:

- **Claim 1**: More complex processed features are better data separators and therefore more valuable than less complex and raw features
- **Claim 2**: Repeated simulations of the world cup stages leads to a random forest model with a higher predictive power
- **Claim 3**: A combination of random forest, Poisson regression and ranking methods has a higher predictive power than random forests alone

3 Methodology

The paper we are attempting to reproduce (1) made some very interesting claims we were keen to investigate, but unfortunately did not come with a strong code base - instead, the code given contained data from the cricket world cup. To overcome this obstacle, we follow a Kaggle tutorial (2) to obtain a similar but more simple random forest model, as well as the data to train it. With this model, we then investigate the three claims made by the paper. This meant that we cannot directly reproduce the results of the paper, but get as close as we can with the resources we have available and investigate and describe similarities and differences between our implementations.

To investigate Claim 1, we used the simpler data of the Kaggle model to identify features that would be most important to distinguish between whether a team would win or lose. The tutorial suggests making the outcome of the football game a binary classification problem, by making all draws a loss for the home team. Using this approach, we can identify the features that show the greatest distinction between a when a team achieves victory and defeat, finding features that separate the data well using violin and box plots. In addition, we plot and eliminate highly correlated features.

To investigate Claim 2, we adjusted the code from the Kaggle tutorial to become far more modular, and used this to run the world cup simulation multiple times. We train and perform hyperparameter tuning on a random forest model before using it to simulate the world cup. Due to hardware constraints, we could only do 100 simulations, as opposed to the 100,000 done in the paper. We computed the mean probabilities of victory for each team in each round, passing the most successful teams onto the next round, to avoid discrepancies caused by differing match-ups. The use of probabilistic averages instead of a purely likelihood-based selection method also prevented errors where teams’ success was poorly represented due to high variance. For instance, with our small number of simulations, a tightly contested knockout stage match could see each side progress in close to half of all runs, despite one team having a higher overall

probability. Taking the mean across simulations ensured that so-called "decisive" victories were more important and gave the winning nation a better chance of advancing than a "lucky" win with 50-50 odds.

Finally, to investigate Claim 3, we compared the simpler random forest model from our tutorial to the more complex random forest model from the paper, observing the model's classification scores and reproducing the road to the final tree for the 2018 FIFA World Cup.

3.1 Model descriptions

The model we use to predict the world cup bracket is a random forest model. We perform hyperparameter tuning with sklearn gridsearch, including a range of values for each hyperparameter that also includes the default values. The hyperparameter names and their values provided by the tutorial are shown in a table in Section 3.3. We estimated the number of parameters in our model by summing the number of nodes in all the trees in the forest, and multiplying by the number of trees and the number of weights for each feature + 1, to account for bias. Thus, our random forest model has roughly 44,964 parameters.

3.2 Datasets

Our code used 2 datasets from the Kaggle tutorial(2). These datasets were used because they were simple and easily available, and similar features that were used in the paper we are replicating. Both of these datasets were pre-processed together and combined to give the most meaningful features for our random forest predictive model. We followed the guidelines of the tutorial for data obtaining and pre-processing.

The first dataset used is a Kaggle dataset that contains the results of all international matches from 1822 - 2022 (5). We used the results.csv file to obtain the results of all international games. We combined these results with the second dataset, containing world ranking data of all teams from 1992 to the present (**?**), to give us our set of raw features to use in our analysis. To begin, we filtered the data to contain games between the end of the 2014 world cup, and the start of the 2018 world cup, to use as our training data, and games from the 2018 world cup for our simulation. This data included features like the home team score, the away team score, and attributes to do with their rankings. We also observed and eliminated correlated features, as shown in Figure 2 in the appendices.

We could then compute a few simple processed features for the model, to average values for the team over time in the buildup to the world cup. These features are similar to the sportive and home advantage features used in the paper we are trying to reproduce (1), since the other features were too complex to obtain. The features we used for both teams included the mean goals both scored and conceded, the mean FIFA rank the team faced, the mean FIFA points, the mean games points, and the mean game points by rank. All of these means were calculated both over the 4 years leading up to the world cup, and over the last 5 games, to mirror the importance the original paper gives to games that occur closer to the world next cup (1). 5 games was a tunable hyperparameter that had no effect on the model accuracy when investigated.

These simple processed features were then used in our investigation of Claim 1 and in the importance of more complex features, and served as the basis to obtaining the final features we conclude on in Section 4.1.1 during that exploration. With these final features, we create a dataset with the 2929 games between the end of the 2014 and the beginning of the 2018 world cup, to train and evaluate the model. We then use an 80 - 20 split to obtain the training data, leaving 2343 testing instances and 586 training instances.

3.3 Hyperparameters

The hyperparameter values for our simple random forest model were tuned with sklearn's gridsearch feature, so doing so was very straightforward. The hyperparameter of the number of games before the world cup to make a feature had very little effect on model performance, so 5 was chosen for this value. The most optimal hyperparamters for the random forest model are shown in the table below. These were different to the ones in the demo code associated with the paper, which just used default values.

Table 1: Hyperparameter Values

Hyperparameter Name	Value
Max Depth	10
Min Samples Split	30
Number of Estimators	250
Max Leaf Nodes	None
Min Samples Leaf	1

3.4 Experimental setup and code

Our first experiments use our data to investigate Claim 1: that more complex processed features are better data separators and therefore more valuable than less complex and raw features. These experiments are set up with the same data as described in Section 3.2. We then use violin and box plots to identify features that are good discriminators.

With the data from our previous experiments, we then create a Random Forest Classifier implemented and optimized by the sklearn library. This model, with the hyperparameters listed in Section 3.3, is used to to simulate the 2018 world cup, and then repeat this simulation 100 times to discuss the impact of repetition (Claim 2) on our results. We are then able to compare the world cup bracket we obtain with a simple model to the complex results from the paper to discuss the third and final claim. Our code contains all of these experiments (4).

3.5 Computational requirements

All of our experiments were kept simple to stay within our computational limitations, and thus can all be run on the free version of Google Colab with no GPU acceleration and 12.7Gb of RAM. Our experiments only used one model, a random forest classifier, with the test and training data mentioned in Section 3.2 and the hyperparameters listed in Section 3.3. The modelling of 100 World Cup simulations took on average about eight minutes in our implementation. Feature engineering ran for up to a minute, while all box and violin plots comparing features took mere milliseconds to execute. The only notable exception to this was the plotting of correlated features, which had a mean runtime of 18 seconds.

4 Results

Our results in general support the claims of the main paper, demonstrating that more complex features are more valuable than less complex ones, and that multiple simulations of the world cup bracket leads to results with which we can have a higher confidence in. Our simple model trained on the matches between the 2014 and 2018 world cups performs with a classification rate of 69% when tested on verified outcomes. This is higher than the results of the paper at 55%, due to simplification methods used. Our tournament bracket for the 2018 World Cup averaged over 100 simulations is remarkably similar to the one shown in the work studied, with most winners predicted in the same way, a much more convincing bracket than when a single prediction is made, supporting the second claim.

4.1 Results reproducing original paper

4.1.1 Complex vs. Simple Processed Features

These experiments investigate Claim 1, that more complex processed features are better data separators and therefore more valuable than less complex and raw features. In the paper we are replicating (1), features with the highest importance were ‘Abilities’, ‘Rank’, and ‘Oddset’, all of which are highly complex processed features, too complex for us to replicate. ‘Abilities’ are derived from a Poisson regression model, ‘Rank’ is derived form ranking methods, and ‘Oddset’ are the betting odds in favor of each team, also derived from extensive analysis. Thus, to investigate the importance of complexity in a way that was accessible to us, we changed our dataset to become a binary classification problem - between a home win (0) and an away win (1). We did this by changing all of the draws to a loss for the home team. We then use violin and box plots to identify features that show a greater distinction in their distribution between a 0 and a 1.

With the initial simple features from Section 3.2, Figure 3 in the appendices shows that only the ranking difference between the teams has a significant difference and is a valuable feature. Thus, more processing is required to obtain more valuable features - we need to compute the difference between values for teams, as the values themselves are not enough. The box plot for this attempt is shown in Figure 1 below.

The figure demonstrates that difference of points, differences of points by ranking faced, and difference of ranking faced are good features, for both over the last 4 years before the world cup, and over the last 5 games. Including these features, we also add the ranking difference, the goal difference suffered and conceded, as well as the goals per ranking difference, to our final features, all for both the 4 years before the world cup and the last 5 games+. In addition, we include the boolean isFriendly feature. This final set of features shows that more complex features such as computed differences of averages are more valuable than simple or raw features, and support the claim that features such as ‘Abilities’, ‘Rank’, and ‘Oddset’, all far more complex than these, would greatly improve model performance, as they are features that come from other models themselves.

4.1.2 Repeated Simulation of the World Cup Bracket

By creating abstractions and transforming the tutorial implementation into a more modular approach, we were able to run repeated simulations with separately trained models, and compute the mean probabilities of each tournament match’s outcome. By simulating each round independently, calculating average probabilities and passing the most

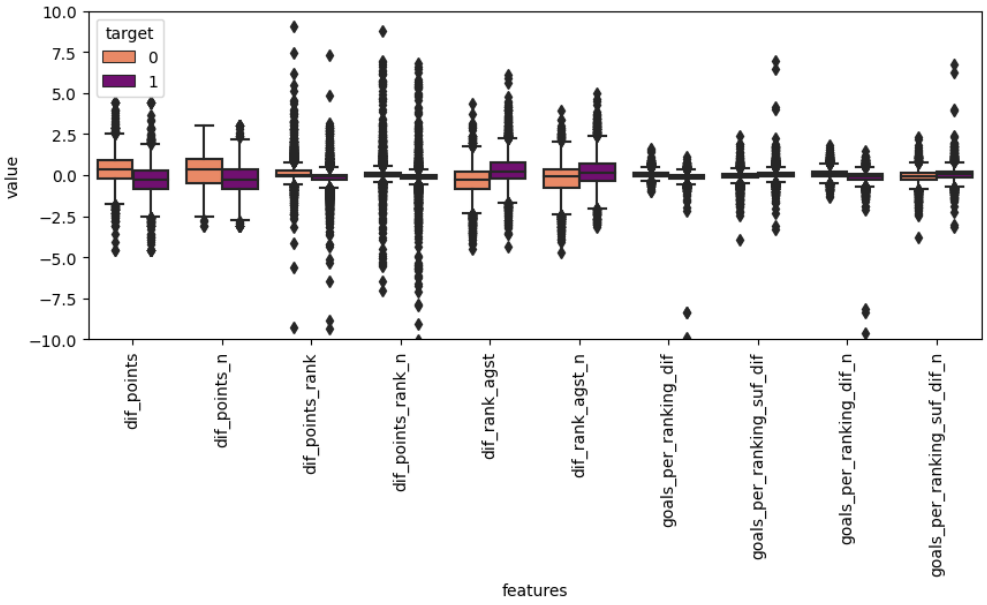

Figure 1: **Box Plot for More Complex Processed Features**

successful teams from each round to the next, we obtained a bracket that avoided several of the unlikely results that occurred in individual runs. The overall bracket from 100 averaged simulations closely resembles that of the paper, both of which can be viewed in the Appendix, Figures 4 and 5. 8/15 (53%) of projected matches are identical in both versions, as well as 11/15 (73%) of advancing teams, which includes all four semi-finalists: France, Brazil, Spain and Germany. Often not the case in individual simulations, where surprise smaller teams advanced in contrast to reality, this validates the second claim stating that repeated runs lead to better predictive power. Mean victory probabilities equally negate variance in match outcomes, which fluctuate above and below the threshold of 50% between individual models. This adds an extra level of regularization on top of what is already provided by the random forest algorithm, further strengthening model accuracy and stability.

4.1.3 Comparing Final Results

When taking into consideration the frequency of draws in football, our overall classification was likely in the 50-53% range (see section 5), less than the authors' model's 55%, demonstrating the strength of combining Poisson-regressed abilities, rankings and random forests (1) over a simple random forest implementation from a standard Python package. This drop in accuracy and reliability reflects the third claim discussed, that a more complex model has a higher predictive power. The paper also demonstrates consideration for special rules and predicts goal totals instead of simple outcomes (further discrepancies are outlined in section 5), which result in a more realistic reflection of the World Cup. Notably, our simulation's placement of Peru in the knockout stages instead of Denmark and favouring of Poland to make the quarter finals, two rather unrealistic results, were both correctly avoided by the version we tried to replicate, and indeed never occurred (9).

5 Discussion

Several simplifying steps were taken in order to make the paper's experiments reproducible with available resources. This lead to experiments which are efficient and easily understandable. The data chosen is easily accessible and does not require a level of preprocessing remotely close to the extent of the Poisson and binomial distributions in the paper (1). Unfortunately, these changes inevitably led to shortcomings in reliability and model accuracy. The most notable modification is that away wins and draws were fused into a single label in order to benefit from a binary classification problem. By grouping these contrasting results into a single category, though we get to work in a smaller dimension, we are reporting accuracy as being much higher than it likely is. This is because our confusion matrix (Figure 6 in Appendix) considers all "draws *and* away wins" predicted to be "draws *or* away wins" as correctly classified instances.

In our training set of 586 matches, 222 fell into this grey zone, and may have falsely been deemed successful. This meant that our accuracy could have been anywhere from 43% to the reported 69%. However, when taking into consideration the frequency of draws in football, which is reported to be around 25% (7), our model's overall classification score was likely in the 50-53% range.

Furthermore, a World Cup does not have home advantage for any team except the host nation. To account for this major difference between qualifying matches and tournament ones, we simulate twice with each team as the home team, and then take the higher probability. If one team is more affected by the home team factor, this gives them a pseudo-advantage in our simulation. In addition, in the group stage, we consider it to be a draw when in this double simulation, one team wins one game and the other team wins the other. We do not take into account how decisively each team may win, which may result in an abnormally high number of draws predicted.

Another major weakness is that specific rules such as tiebreakers for group qualification and extra time or penalty shootouts in the knockout stage were beyond the scope of what we could implement into our simulation. As previously mentioned, our match predictions were also binary, which is a vast reduction in complexity from the discrete goals-per-team estimates computed in the paper (1).

As a counterpoint, one factor playing in our implementation's favour is our training set, which is more recent and more complete than that of the studied work. Our 586 training instances are all from the four years preceding the tournament, whereas the paper's are from four preceding tournaments (going back 16 years from the event), each of which only consists of 64 matches. Despite weighting older matches less in their training, our input is more directly relevant to results and involves players actually participating in current squads. One could also make the argument, however, that training on purely World Cup data like in the studied work captures result patterns specific to tournament play.

A further strength is that when attempting to mirror the original paper giving less importance to older games, we only created features for five preceding games, but upon hyperparameter investigation, it had no effect. This could have been further tuned and observed, however, with different ranges of time periods.

It is also worth noting that classification score using black and white statistical features may not be the best metric for comparing prediction models for sporting events, due to the unpredictably of a game which, in truth, can end in any given way, regardless of presumed team strength (8). Events like Germany's shock elimination (9) are never going to be predicted by a statistical model and elements such as the stadium atmosphere, not quantifiable in machine learning, may play a role in a result.

Ultimately, we conclude that all three claims identified in the original paper are accurate according to our analysis and experiments, but further modifications to both our methods and those of our source could potentially lead to even stronger predictions.

5.1 What was easy

The more or less direct following of a tutorial, paired with a significant simplification of the processes detailed in the paper, meant that our code was reasonably trivial to implement and produced positive results from the beginning. Libraries used also reduced tasks such as visualization with box plots and hyperparameter tuning via cross validation to simply a handful of lines.

5.2 What was difficult

Many of the experiments from the original report were impossible for us to attempt given our hardware resources in Colab and the provided time frame. Hundreds of thousands of simulations would have taken days and exceeded RAM limitations. Assumptions had to be made with regards to interpretations of what certain figures or terms in the report were referencing; for example, percentages in the tournament bracket were not transparent about whether they reflected counts, averages, or a combination of the two. Finally but most significantly, little assistance in terms of code implementation and description of steps followed were provided in accompaniment of the report, which left us to our own devices for a large portion of the data collection and experiments.

5.3 Communication with original authors

We were not able to communicate with the original authors of this paper.

References

[1] A. Groll, C. Ley, G. Schauberger, and H. Van Eetvelde, “Prediction of the fifa world cup 2018 - a random forest approach with an emphasis on estimated team ability parameters," *arXiv preprint arXiv:1806.03208,* 2018.

[2] “Predicting FIFA 2022 World Cup with ML,” kaggle.com. https://www.kaggle.com/code/sslp23/predicting-fifa-2022-world-cup-with-ml/notebook#WC-Simulation (accessed Apr. 26, 2023).

[3] https://github.com/abhinavsagar/ICC-2019-WC-prediction/blob/master/ICC%202019%20WC%20RF.ipynb

[4] https://colab.research.google.com/drive/1C23USKvKZL-Sqnuljl6LiW1fGHcvFDHX#scrollTo=UzYVauXZf9Bb

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

[6] https://www.kaggle.com/datasets/cashncarry/fifaworldranking

[7] “How often do football matches end in a draw?,” Bookie Sign Up Offers, 27-Jan-2023. [Online]. Available: https://www.bookiesignupoffers.com/2020/10/21/how-often-do-football-matches-end-in-a-draw/ [Accessed: 27-Apr-2023].

[8] E. Asikci, “World Cup has seen its share of big upsets in its 92-year history,” Anadolu Ajansı. [Online]. Available: https://www.aa.com.tr/en/sports/world-cup-has-seen-its-share-of-big-upsets-in-its-92-year-history/2745753. [Accessed: 27-Apr-2023].

[9] https://www.fifa.com/tournaments/mens/worldcup/2018russia

6 Appendicies

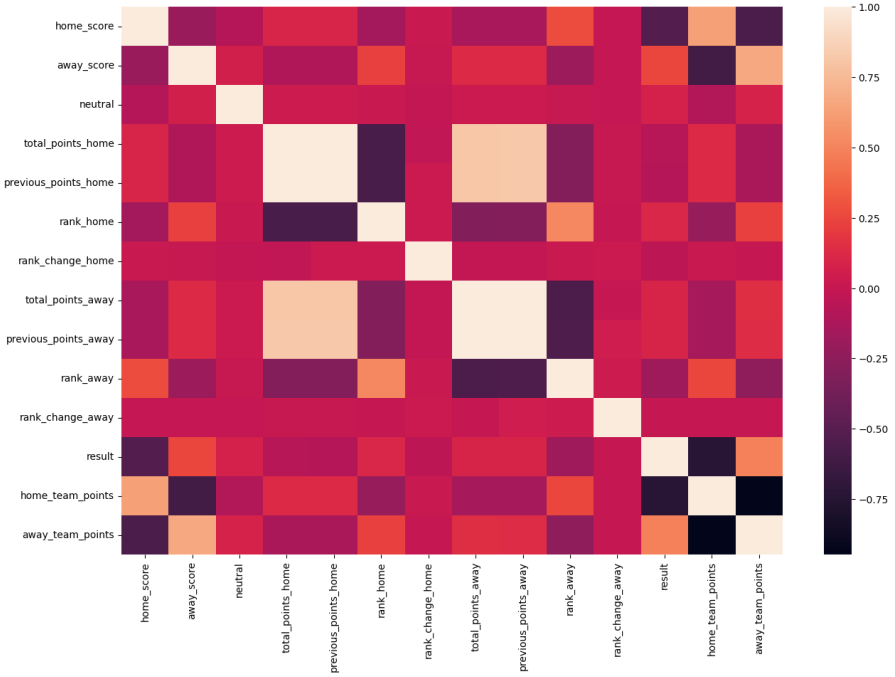

Figure 2: **Correlation Between all Raw Model Features**

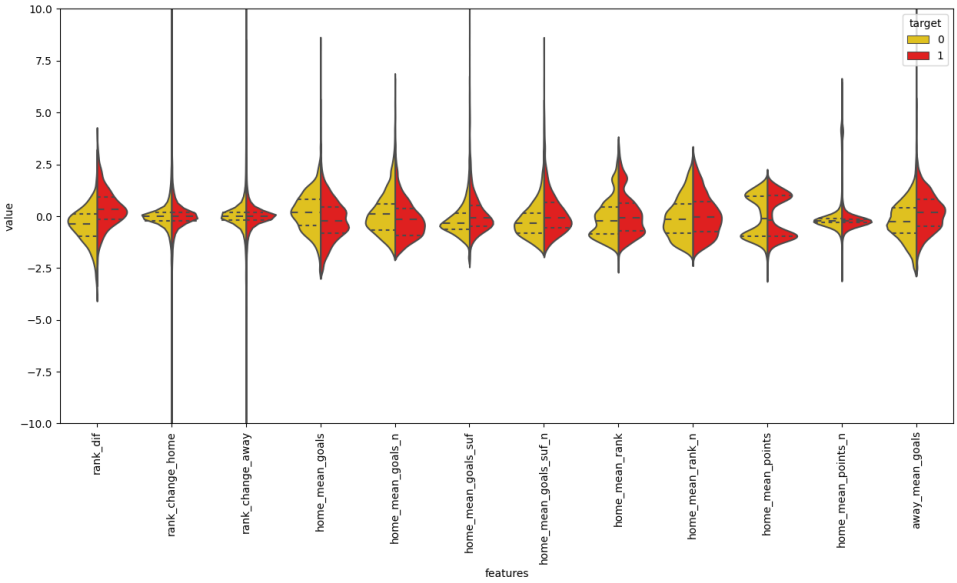

Figure 3: **Violin Plot for Simple Processed Features**

Figure 4: **2018 World Cup Bracket Predicted by our Random Forest Model (100 simulations)**

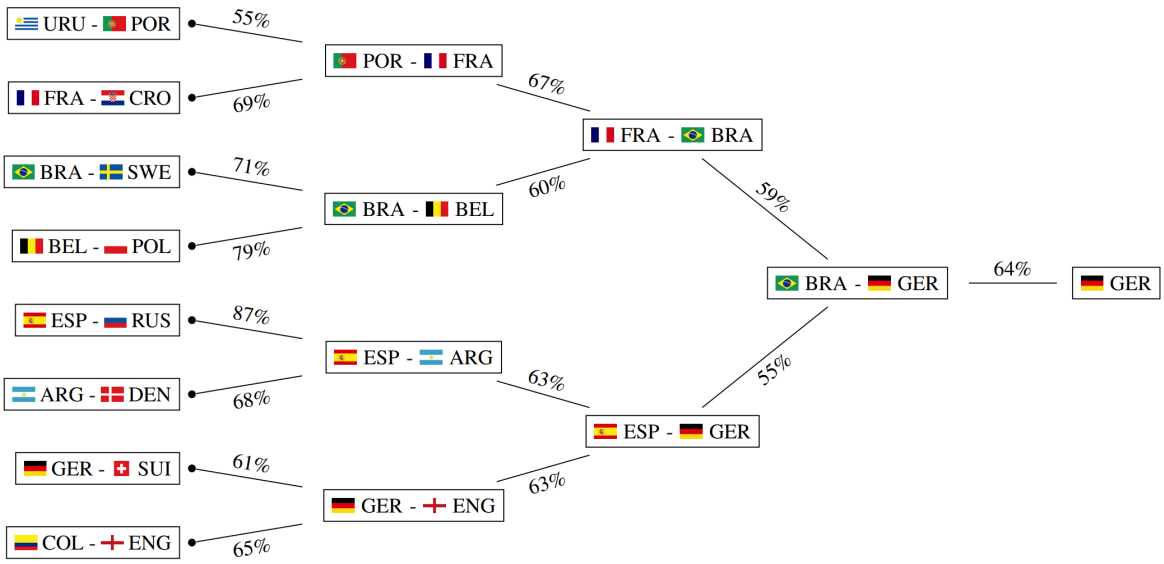

Figure 5: **2018 World Cup Bracket Predicted by the Authors' Model (100,000 simulations)**

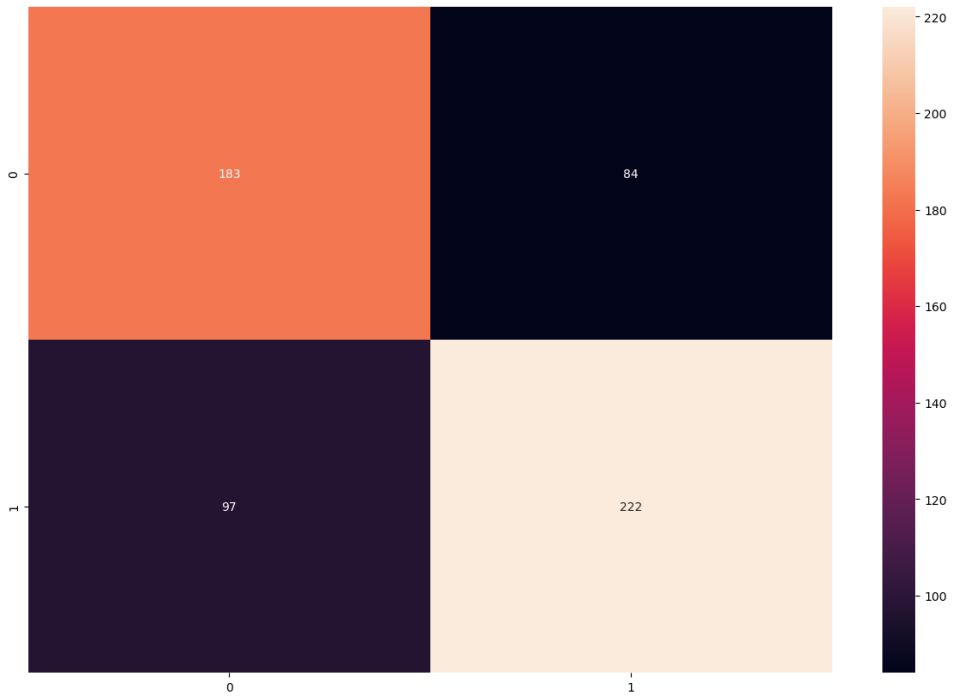

Figure 6: **Confusion Matrix (Test Data Taken from Matches Leading up to 2018)**