Is the FIFA World Cup Draw Truly Random?

Pratik Kulkarni (19400570)
Shrey Parekh (18706941)

Project proposal for 301111 Discovery Project

Supervisor: Dr. Russell Thomson

School of Computer, Data and Mathematical Sciences,
Western Sydney University

Spring, 2021

Table of Contents

1 Background

Is the FIFA World Cup draw truly random?

This section is written by Pratik Kulkarni (19400570)

This is the topic chosen for our Discovery Project. The topic was selected by our supervisor, Dr. Thomson while watching a soccer match and decided to investigate this topic. We intend to investigate whether the FIFA World Cup draw is truly random and if each group has an even chance of being drafted into the groups.

Researchers such as Cea et al. [2020], conducted research on a similar topic. In this paper, they found that one of those main factors that affects the countries chances of being drafted is their home-away status. While this does not directly affect the team ranking by points, a general belief is that teams always play better at home, this factor has not yet been considered in the draw of teams so far. This affects their game due to the home team feeling stronger with a crowd that is supportive of them. This in turn influences their world ranking. Moreover, another factor found in this paper's research was that until the 2018 World Cup, teams would avoid friendly games due to friendlies awarding very few points, teams that would play more friendlies had less chances of going further up the team ranking table as compared to teams who avoided friendly matches and climbed up the ranks.

In the 2018 World Cup, FIFA changed the draw system to a rank-based system, the groups would be split by the team's rankings. While this could be a better system, there are chances that the groups might become unbalanced, especially if a low ranked host country was placed in the top group with the best teams.

The knowledge gap in this topic would be that there is no fairness index created by any researchers. To fill this gap, we would have to make a fairness index for each team that did qualify for our chosen year of the world cup based on their FIFA team ranking in the same groups.

2 Objective

This section is written by Shrey Parekh (18706941)

The Supervisor's goals are centred on providing value and results to the fans. This can be seen through testing the randomness of the FIFA World Cup draw. The goal of this project is to determine the predictions of the simulated World Cup draw. Aim is to discover the fairness index for each team that qualified for the World Cup, based on the FIFA team rankings of the teams in the same grouping. This will allow us to determine that World Cup draw is based on chance or is done through an algorithm.

We will establish whether the fairness index affects the performance through multiple linear regression or non-parametric tests. This will increase the scope of this project as we will be able to produce efficient statistical results to ascertain the influences of chance in the FIFA World Cup draw.

This can be further reinforced by simulating the World Cup draws to test if any countries have a fairness index that is higher or lower than expected by chance alone. This is to proclaim that certain countries will not have a favourable advan-

tage while competing in the World Cup.

3 Hypothesis/question

Is the FIFA World Cup draw truly random? Can we test the randomization of the grouping?

This section is written by Shrey Parekh (18706941)

3.1 Why Ask this Question?

This is a concern fueled by the excitement of the fans of the FIFA world cup. This is important because it raises an interest about who will win this year's trophy. It also generates doubt in the mind of fans as the draw based on chance is fair or is it biased to some teams. They want to know the algorithm that is used to decide the matchups of the team.

The value of answering this question is that it will remove the doubts from people's minds. This will allow the fans to determine that the teams competing are fairly matched up. They will be able to then ascertain that the draws are random, and each team is given a fair chance to progress through the tournament. This will provide confidence for fans as the best performing team will win, not any team based on a biased draw.

3.2 Data Question

Can we accurately predict the winning team of the FIFA World Cup?

3.3 Data

The first part of the data is provided by the supervisor and the second part of the data the team must find the data set or manually scrape it from the websites to get World Cups with groups, rounds, teams participating and the result of the FIFA World Cup. This will then be merged with the existing data set that contains the points and teams from the world.

3.4 Data Science Process

3.4.1 Data Analysis

The Data pipeline that was used to wrangle the raw data is below:

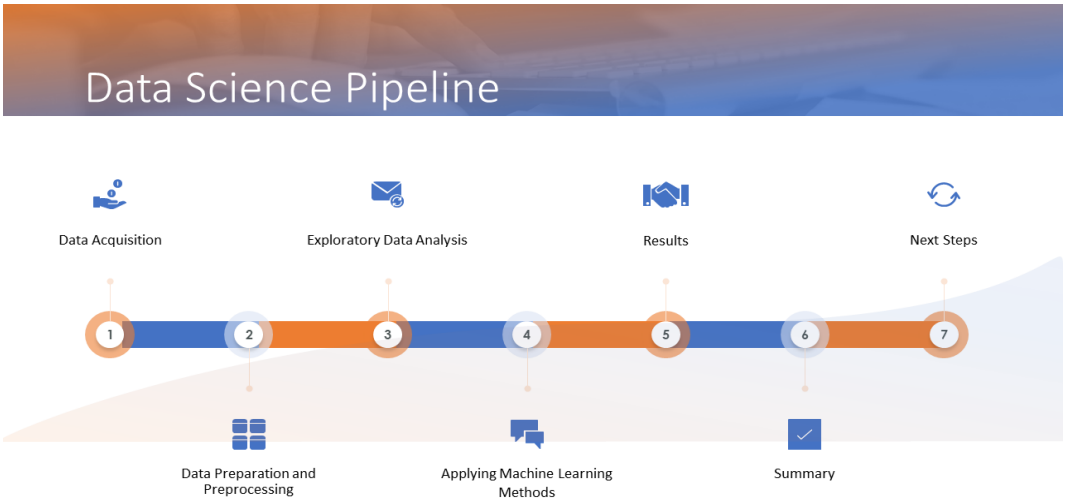


Figure 1: Data Science Pipeline

4 Methodology

This section is written by Shrey Parekh (18706941)

The goal of our project is to create a fairness index for each team that qualified for the world cup based on team rankings in the groups. Test the performance being affected by index through multiple linear regression tests and non-parametric tests. Simulating the world cup draw to observe the draws are fair and not based on chance.

The measures that we need to take to conduct experiments to obtain precise results for this project are that we need to generate a fairness index from the points scored by the teams in their respective groups. This will be calculated using Python libraries such as NumPy and Pandas. Then we will look at the World Cup Result where the most successful team gets the highest amount. A predictive model will be generated based on the fairness index and World Cup Result. This will allow us to demonstrate that group rating and fairness index affects the winning team.

The calculation below is an example to use:

$$lm(result \sim points + fairness + features + residuals)$$

After this step our goal is to test the linear regression model for randomisation of the grouping by simulating the process in which the team members are grouped based on their ranks and points. This is achieved by writing a Python script to

assess the randomisation of groupings.

The detailed outline of the process to complete our project will be illustrated below:

1. Data gathering from the supervisor data was obtained that consisted of FIFA World Cup matches from 1992 to 2021. The team needs to source another data set or will have to scrape data manually to get the groups, result of the World Cup, the rounds in the tournament and the teams that participated. The two data sets will need to be merged to conduct analysis.

2. Performing feature engineering that will allow us to predict the best team of winning the FIFA World Cup. The features that will provide value for this project need to be selected such as groups, World Cup result, points.

3. The feature engineering will allow to create the fairness index based on the points and groups.

4. Exploratory Data Analysis on the data set to obtain insights from the World Cup data and analyse the trends of seeing the patterns in the winning team.

5. Data preparation and pre-processing the data will be cleaned to filter out any missing data, corrupted data from the data set. This will enable the team to get accurate results in predicting the winning team.

6. Applying machine learning models and algorithms will be applied to the data set:

    - Splitting the data set in proportions to 80 to 20 ratio.
    - Then will performing normalization
    - Building the models
    - Then we will predict the winning team
    - Plot the ROC curve

7. Results and evaluation of the models. Evaluating our models to optimise the results of prediction and selecting the best model with the closest accuracy of predicting the winners of the World Cup.

8. Conclusion and Recommendation for FIFA World Cup 2022.

The models and machine learning algorithms that we will use to conduct multiple linear regression and non-parametric tests: logistic regression, Random Forest, Neural net, LightGBM, Naïve Bayes, and Support Vector Machines.

4.1 Block Diagram

The below block diagram displays the methodology we will follow in this project:

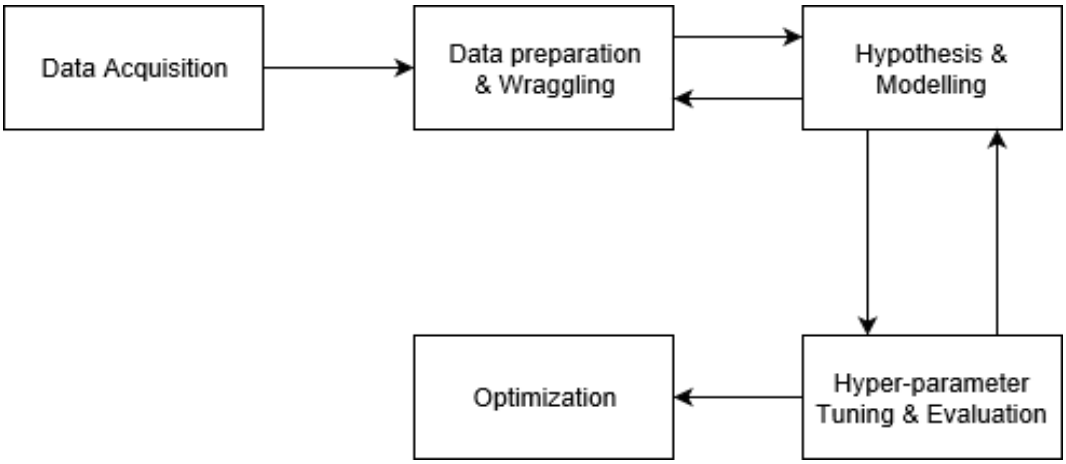


Figure 2: Block Diagram by Mrthlinh [2018]

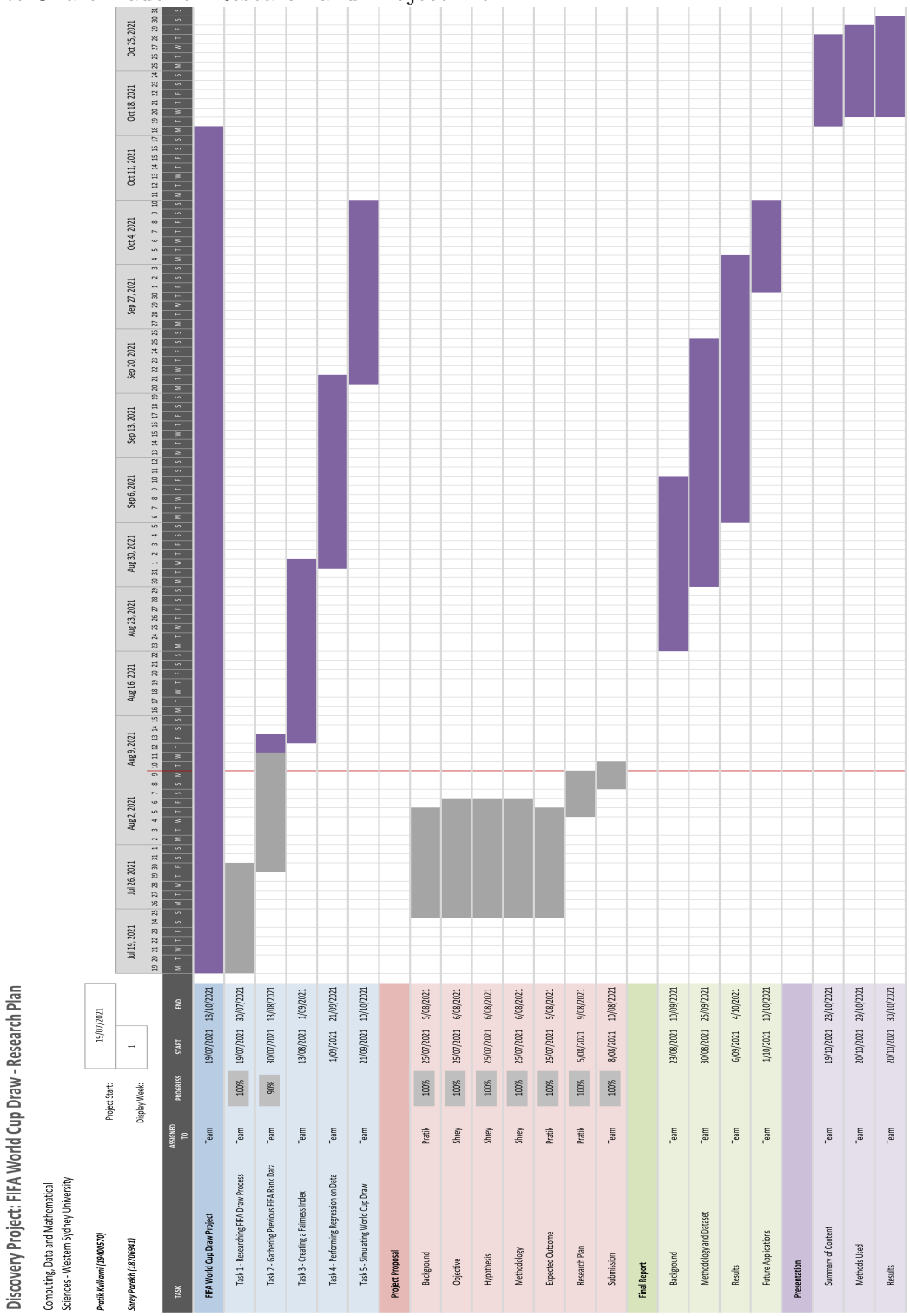
5 Expected outcomes

This section is written by Pratik Kulkarni (19400570)

The expected outcome at the end of this project would be to gain a deeper understanding of the topic. Another outcome would be discovering any trends from gathering data of previous FIFA team rankings, seeing if there are any common reoccurring factors. Assuming the draw is not 'truly random', we also hope to find out if there are any factors that would give some teams a higher chance of being drafted for the World Cup.

6 Program of work

Gantt Chart made for Research and Project Plan:



This section is written by Pratik Kulkarni (19400570)

References

Sebastián Cea, Guillermo Durán, Mario Guajardo, Denis Sauré, Joaquín Siebert, and Gonzalo Zamorano. An analytics approach to the fifa ranking procedure and the world cup final draw. *Annals of Operations Research*, 286(1):119–146, 2020.

Mrthlinh. Fifa-world-cup-prediction/report.md at master · mrthlinh/fifa-world-cup-prediction. *GitHub*, Jul 2018. URL `https://github.com/mrthlinh/FIFA-World-Cup-Prediction/blob/master/report/report.md`.

Rodrigo Nader. Using machine learning to simulate world cup matches. *Towards Data Science*, Jul 2018. URL `https://towardsdatascience.com/using-machine-learning-to-simulate-world-cup-matches-959e24d0731`.

University of Washington. How to write your discovery project proposal. *University of Washington School of Medicine III Requirement*, 2018. URL `https://sites.uw.edu/somcurr2/iii-scholarship-requirement/scholarship-of-discovery/how-to-write-your-project-proposal/`.

Abhinav Raghunathan. Simulating the fifa world cup 2022. *Towards Data Science*, Dec 2020. URL `https://towardsdatascience.com/simulating-the-fifa-world-cup-2022-d363fad7da22`.