World Cup Prediction Report

Trent Meyer

2023-05-08

World Cup Prediction

Introduction

The FIFA World Cup is an event that grips the attention of soccer (or fútbol, football depending on where you are in the world) fans all across the world. This event occurs once every four years, bringing together millions of individuals from six different continents and a wide range of cultures. To have your nation qualify for the World Cup is quite prestigious as only 32 countries (48 in 2026) are able to do so. This rigorous qualifying schedule involves playing the top teams in your region over the course of about two years prior to the next World Cup.

The United States, despite not being completely enthralled like say Brazil, still nonetheless sees a magic atmosphere surrounding the tournament. However, in 2018 the United States lost to Trinidad & Tobago, a team they were supposed to easily beat, in their last qualifying game which effectively knocked them out of that cycle's tournament. This sparked quite a long chain of events in the United States Soccer Federation (USSF, the main governing body for soccer and the national teams) over the next four years. This included completely revamping the youth development system, hiring a new manager, and phasing out all of the aging players in favor for bright, up and coming stars making their presence known in European academies.

This chain of events changed quite a lot for the USSF, as seen since 2018 and today. Our men's national team saw its stars playing at some of the most highly regarded clubs in the world such as Juventus, Chelsea, and Borussia Dortmund. Not only to mention the quality of the clubs, these players were teenagers starting and making an impact for their club. We saw a higher number of teenagers playing in Europe than ever before, meaning our players were developing at the best academies in Europe. Over the next few years, our young team managed to win two regional tournaments and secure our spot in the 2022 World Cup in Qatar.

With such a large event, many want to predict the winner of this prestigious trophy. After 22 tournaments, only eight have managed to win the tournament. Many companies such as FiveThirtyEight and ESPN, to name only a couple, attempt to predict the winner of the upcoming tournament using many different methods.

FiveThirtyEight created a process to predict the winner of the 2018 tournament. This process is explained, along with the data being available in a GitHub repository. Using their data, I am attempting to recreate the probabilities that each team will win each game in the group stage, and use this to determine who is most likely to make it past the group stage. I will also use this same process for the data from the 2022 tournament, which has currently been ongoing in the months of November and December.

Methods

To begin their predictions, each team's World Cup roster was given an SPI rating, which judged the strength of that team. Each team was given an offensive rating, which is "the number of goals that it would

be expected to score against an average team on a neutral field" (Boice, 2018). The defensive rating followed the same situation, except the number of goals they would be expected to concede against an average team. The overall "match" SPI rating is the "percentage of points ... the team would be expected to take if the match were played over and over again" (Boice, 2018). This SPI rating is calculated using a database of international matches dating back to 1905! They also calculated a roster-based SPI rating which looked at each player, and how much they played, in what league, and how elite the competition they are facing each week is.

Once the SPI rating was calculated, they began by predicting the match scores, which are the number of goals that team would need to score to uphold their rating. Using the projected score, they completed a Poisson process where they predicted the probabilities that each team would score zero goals, one goal, two goals, and so on up to ten goals (a little uncommon, but not outlandish).

The next step was to convert these to matrices, and multiply them together to determine the likelihood that one team would score x amount of goals *and* the other team would score y amount of goals. Using this matrix, I was able to calculate the cumulative probability, and generate a random value that would correspond to an outcome based on the matrix. Each match in the group stages was iterated 1000 times, from which we could determine the number of points each team would score in that iteration. Next, I was able to calculate the proportion of times that each team would advance out of the group stages (place in the top two out of four teams in their group).

Results

2018 World Cup

From Figure 1, we are able to see that the teams that made it through the highest proportion of times were Spain (0.916), France (0.873), and Uruguay (0.865). Possibly the most surprising piece of information however is that Germany were the favorites for their groups, however did not actually make it out of the group stage, with Sweden and Mexico advancing instead. The only teams that were expected to advance and did not were Poland and Germany. France were one of the favorites, and end up beating Croatia in the final and winning the whole tournament.

2022 World Cup

From Figure 2, we are able to see that the teams that made it through the highest proportion of times in 2022 were Spain (0.894), Brazil (0.882), and France (0.863) Possibly the most surprising piece of information however is that Germany again were the favorites for their group, however did not actually make it out of the group stage, with Spain and Japan advancing instead. This year had quite a lot more upsets, with four teams who were expected to advance not doing so: Mexico, Denmark, Germany, Uruguay. Australia who were predicted last in the group surprisingly made it out of their group.

Discussion

One of the more difficult aspects of this was that I was not able to completely replicate the probability matrix, as their article only gave limited information on the process. However, with the information given, I learned a lot about predicting World Cup games using skills I learned in linear algebra, probability, and data science. I had to begin by predicting only one game, and once that process was completed, I was able to expand out to one group of four teams. Then once one group was done, I was able to expand to the whole group stage, with eight groups.

The 2022 tournament is at a slightly different time compared to the 2018 version, as the COVID-19 pandemic and the heat in Qatar pushed the tournament into November, rather than June and July. Many players are midway through their seasons, which caused quite a few injuries to star players. We have seen

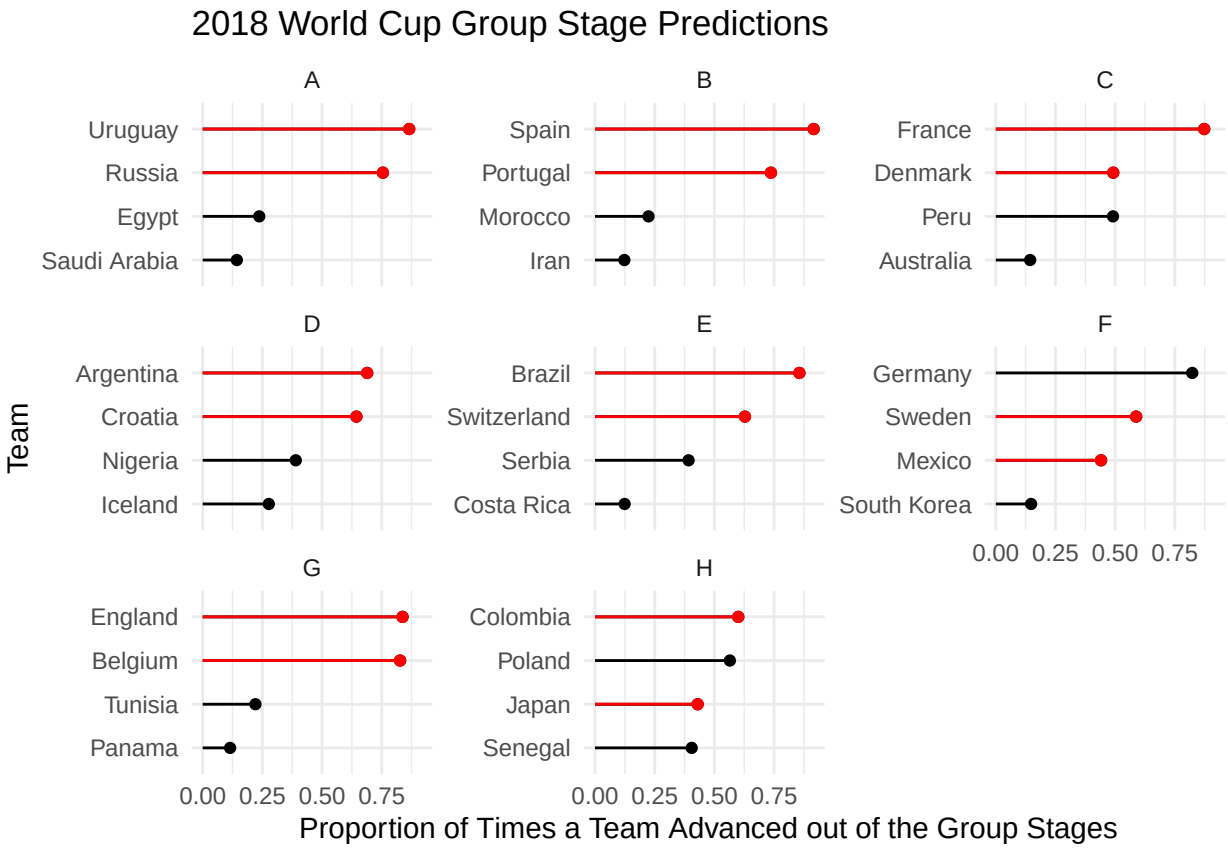

Figure 1: Proportion of times each team advanced out of their respective group

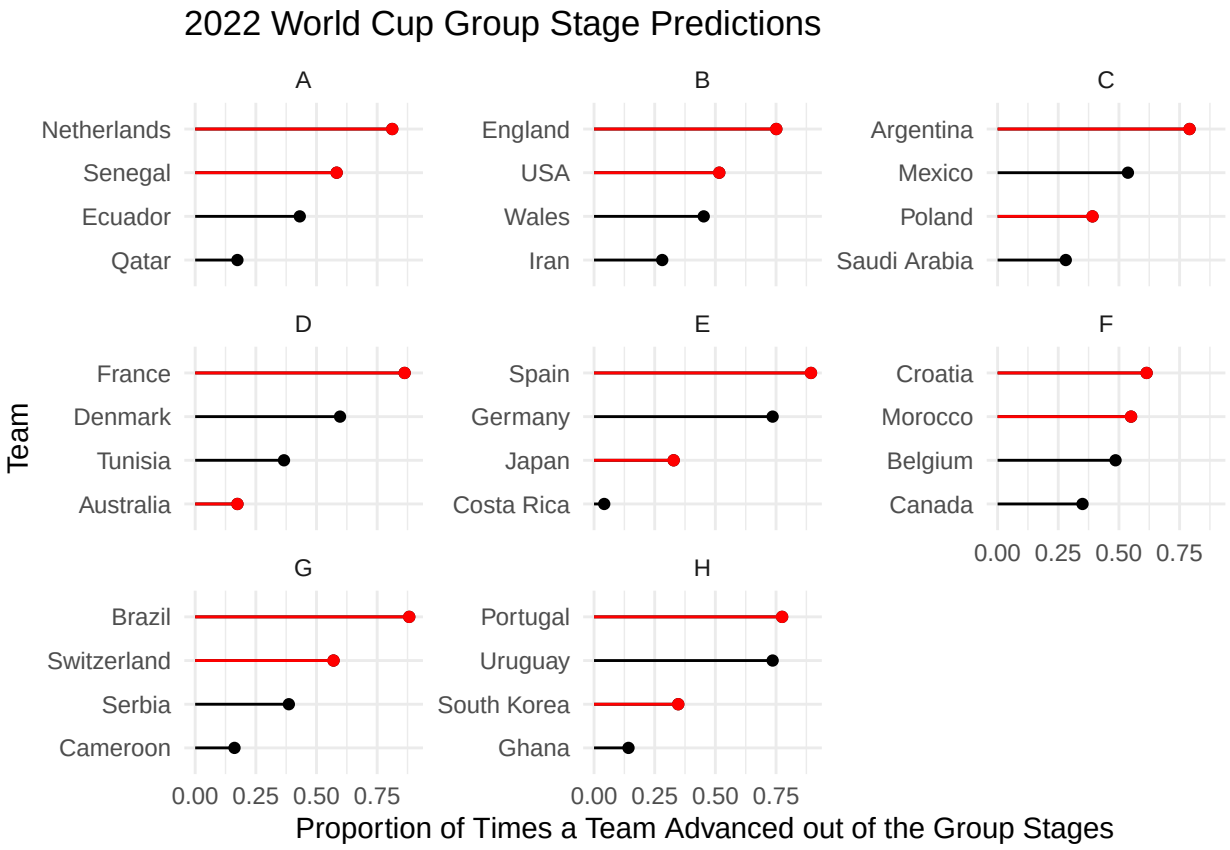

Figure 2: Proportion of times each team advanced out of their respective group

the 2022 tournament be one of the most exciting to watch because of players such as Lionel Messi, Cristiano Ronaldo, and Neymar Jr playing in their last World Cup. Also, this tournament's upsets have shocked many, with some favorites being knocked out in the group stages.

This process took much longer than I expected, and given I only had a semester to work, I was only able to predict the group stages. If given more time, I would be able to figure out a way to set up the knockout stage bracket, and then use the expected scores given to predict who would win that World Cup. This would be quite difficult as you would need to find a way to include the tie breakers, which include point, goal differential, and yellow card accumulation. Once this was completed, you would need to set up a bracket, where the first placed team in one group plays the second placed team in another group.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
library(tidyverse)
library(here)
load(here("data/wc_long_iter_total.rda"))
wc_long_iter_total <- wc_long_iter_total %>% arrange(desc(proptimes))
actually_went_through_18 <- c("France", "Argentina", "Uruguay", "Portugal",
                              "Brazil", "Mexico", "Belgium", "Japan",
                              "Spain", "Russia", "Croatia", "Denmark",
                              "Sweden", "Switzerland", "Colombia", "England")

wc_long_18_through <- wc_long_iter_total %>% filter(team %in% actually_went_through_18)

(ggplot(data = wc_long_iter_total, aes(x = proptimes, y = fct_reorder(team, proptimes))) +
  geom_point() +
  geom_segment(data = wc_long_iter_total, aes(x = 0, xend = proptimes, y = team, yend = team)) +
  geom_segment(data = wc_long_18_through, aes(x = 0, xend = proptimes, y = team, yend = team), colour =
  geom_point(data = wc_long_18_through, aes(x = proptimes, y = team), colour = "red") +
  facet_wrap(~group, scales = "free_y") +
  theme_minimal() +
  labs(x = "Proportion of Times a Team Advanced out of the Group Stages",
       y = "Team",
      title = "2018 World Cup Group Stage Predictions"))
library(tidyverse)
library(here)
load(here("data/wc_long_iter_total_22.rda"))
actually_went_through <- c("Netherlands", "USA", "Argentina", "Australia",
                           "Japan", "Croatia", "Brazil", "South Korea",
                           "England", "Senegal", "France", "Poland",
                           "Morocco", "Spain", "Portugal", "Switzerland")

wc_long_22_through <- wc_long_iter_total_22 %>% filter(team %in% actually_went_through)

wc_long_iter_total_22 <- wc_long_iter_total_22 %>% arrange(desc(proptimes))

(ggplot(data = wc_long_iter_total_22, aes(x = proptimes, y = fct_reorder(team, proptimes))) +
  geom_point() +
  geom_segment(data = wc_long_iter_total_22, aes(x = 0, xend = proptimes, y = team, yend = team)) +
  geom_segment(data = wc_long_22_through, aes(x = 0, xend = proptimes, y = team, yend = team), colour =
  geom_point(data = wc_long_22_through, aes(x = proptimes, y = team), colour = "red") +
  facet_wrap(~group, scales = "free_y") +
```

```
  theme_minimal() +
  labs(x = "Proportion of Times a Team Advanced out of the Group Stages",
       y = "Team",
       title = "2022 World Cup Group Stage Predictions"))
```