CS109 EC PROJECT WRITE UP

Ishan Mehta

Background:
Ever since I was a kid, I was obsessed with sports. I remember I would try and manually put together schedules for the NFL or NBA, trying to comply with all the rules that the actual schedules followed. I would cut out all the names of the players on various teams and do a mock draft and then re-rank all the teams. It was the beginning of a fascination of the intersection between sports and modeling.

As I began to develop more coding skills and a better understanding of probability, I wanted to take the biggest sporting event in the world (coincidentally happening at this exact moment) and try and predict it. I'm talking about the FIFA World Cup, of course.

My Project Goal:
I wanted to create a simulator that would essentially output the probabilities of any team in winning the world cup. Mostly, I just wanted to be able to get a sense of whether or not I could predict probabilities accurately.

Disclaimer:
I will be honest and say, I really did this project for myself. It's something I've wanted to do forever with other sports, and I felt as though soccer could be an easy and of course, relevant sport to start with. That being said, a lot of the methods are not exactly in line with the topics learned in class. However, I was more focused on building a complete and working simulator and I feel as though I have done that!

My Process:
Step 1:

- Gather data from various online sources about relevant team statistics including FIFA ranking, FIFA points, group difficulty, qualifying wins/losses/draws/goal differential/matches played
- I then used formulas within the categories and a broader formula (which I had to tinker with quite a bit) to come up with a single number that represented the strength of each team (Elo Rating)

Step 2:

- I began coding my simulator and read in all the data through a CSV file

Step 3:

- I made the variance assumption of 200^2 based on the basketball variance assumption. I read many papers about that estimate and it seemed to hold with soccer as well. I then began calculating single game probabilities by comparing Elo Ratings with a normal distribution.

Step 4:

- I determined single game outcomes by using a bernoulli on the probability I calculated in the above step

Step 5:

- I built out the entire bracket system (lots of dictionary manipulation) and used these two forms of simulation (ELO and Bernoulli) to help predict a winner

Step 6:

- I used bootstrapping/sampling techniques to run the simulation many times with subtle changes (or no changes) to get a more accurate probability for any team to win the World Cup

Findings:

England, England, England. Ok, but in all seriousness, England has the highest probability of winning the world cup according to my model, sometimes clocking in near 35% or more. I run a full 400 simulations in my demo video in which all of the findings will be demonstrated so I want to spend this section talking a little more about areas that I was particularly interested in.

Firstly, I added in a group difficulty score into my Excel formula to try and account for the weird cases that face the Netherlands and England. They are in objectively weak groups with the Netherlands coming out with a 91% on my group power ranking, meaning they are absolutely better than their other groupmates. I wondered if this would translate into more wins for the teams. It seemed like it did. England's probability, although their ELO is lower than both Belgium and Brazil, is much higher. This comes from their easy group stage (meaning they are less likely to get knocked out early), and their high confidence coming off an easy group stage (meaning they are more likely to play better in the playoffs). I'm glad to see that the group difficulty metric definitely played a role in my findings.

One thing I couldn't fully implement properly in my original Excel calculations was trying to account for the difficulty of qualifying groups. For example, Morocco steamrolled their qualifying matches, winning every single game (the only team to do that.) This made them seem like an extremely strong contender and bolstered their ELO quite a bit. The issue: they had one of the weakest qualifying groups in the whole world, filled with teams who have never been to a World Cup. I realized that teams with easy qualifying groups (Morocco, Denmark, Serbia) would have a higher probability of winning the cup than I actually intended to give them. It’s definitely something to work on in the future.

Taking The Project Further:

As Gordon Ramsay would say, “this is the basic recipe.” I just barely touched the surface of tournament and sports modeling. I want to try implementing this with machine learning algorithms, specifically logistic regression and seeing if I could train a model to predict the cup. I want to have a more clear justification as to how I came up with the initial ELO ratings, about how I am adjusting ELO ratings after wins and losses, and about how I calculate the probability as a whole. I took one of the first (and easiest) methods that popped into my head, but I really want to focus on being more precise and intentional with my formulas and choices. Truthfully, this model is a starting point and I’m really glad I took on this project and allowed myself to try. I’m very excited to see what more I can build with the tools and concepts I am learning here at Stanford.

—---

Works Cited:

https://www.espn.com/soccer/standings/_/league/fifa.worldq.afc

https://statmodeling.stat.columbia.edu/2022/11/19/football-world-cup-2022-predictions-with-stan/

https://www.fifa.com/fifa-world-ranking/men?dateId=id13792