WIZELINE

2018 FIFA WORLD CUP RUSSIA

PREDICTION MODEL

What country is the favorite to win the 2018 FIFA World Cup? Who will make the nets tremble the most and hold up that coveted trophy? Which team will shoot themselves into the football hearts of their countrymen? A lot of people have been asking these questions lately. We didn't just want to ask questions, we wanted to find answers. Responding to our call of distress, our data scientists awoke from their long slumber and built a prediction model. For those of you eager to know how it works, we provide a brief explanation of the model we used.

OUR MODEL

We looked at the goals scored and the match outcomes of the countries participating in the World Cup over a two-year period. Our model only uses data from non-friendly matches, because we believe these are a better representation of the actual prowess of a team.

Goal Intensities and Match Outcomes

The first step to predict the outcome of a match is to estimate the expected number of goals team A scores against team B. Unfortunately, many of the countries facing each other in the World Cup have not played each other in the recent past. This is because there are relatively few matches between national teams in football, and hardly any between countries of different confederations. To circumvent this, we look at the outcome of recent matches of team A, focusing on the number of goals they scored. This data is then triangulated by factoring in the relative defense quality of team B with respect to each of their opponents.

For instance, consider a match between Germany and Switzerland. In the last two years, Switzerland conceded on average 0.60 goals per match. Germany played against Australia in the Confederations Cup, which ended in 3 to 2. Australia conceded on average 1.12 goals per match. Therefore, based on the number of goals scored by Germany against Australia, we expect Germany to score $3 \times \frac{0.60}{1.12} = 1.61$ goals against Switzerland. We expect Germany to score fewer goals against Switzerland, because they have a better defense than Australia. Germany played 20 non-friendly matches in the two-year period, so we average the weighted scores over all opponents. After that, we run a similar analysis for Switzerland because the goal intensities are not symmetric.

We can then model the number of goals team A scores against team B, using a Poisson distribution with the goal intensity $\lambda_{A,B}$ as its parameter. To know the outcome of a match, all we need is the goal difference, $Diff = X - Y$, where X and Y are the number of goals scored by teams A and B, respectively. If $X > Y$, the difference is positive and thus shows that team A wins; If $X = Y$, the difference is equal to zero and thus shows that the match ends in a tie; if $X < Y$, the difference is negative and thus shows that team B wins.

It turns out that, since both X and Y are Poisson distributed, $Diff$ follows a Skellam distribution. It's therefore fairly easy to compute the probability of the aforementioned events. When ties are not allowed, as is the case in the knockout phase of the World Cup, we need to evaluate the probabilities of winning in regular time, overtime (which can be regarded as an independent match of 30 minutes), and in a penalty shoot-out.

Monte Carlo Simulations

The World Cup is composed of a groups stage, in which eight mini round-robin tournaments are held. Thereafter, the winners and runners-up compete in an elimination tournament, which is called the knockout stage. We run hundreds of Monte Carlo simulations to obtain different scenarios, because there is a lot of uncertainty in the results of the groups stage. First, we simulate the final score for each of the 48 matches to know which countries advance to the next stage. Second, we cascade the probability of each country advancing through the knockout tree (quarter-finals, semi-finals, final, winner), using recursion. Finally, we average the probabilities of all the scenarios to obtain unconditional probabilities.

What's cool about this model is that we can update the probabilities with the actual match outcomes as they become available. We also use these to augment the dataset when estimating goal intensities. Furthermore, there's no need to run simulations once the contenders for the round of 16 are known.

That's all folks! We know what you're thinking: "Wait a minute, this is not rocket science". We agree, it is not. This is science in snack form! Oftentimes prediction models are not unintelligible black boxes, but a model understandable to all. As Da Vinci used to say: "Simplicity is the ultimate sophistication." But before you bet a month's worth of salary on the World Cup using our prediction model (we are looking at you, hardcore gamblers), please do remember that the model does not consider player-level data such as alignments, injuries, etc. The World Cup is indeed a championship that is hard to predict. Have fun predicting the match outcomes and the ultimate winner!