Machine Learning & Analysis for Predicting the 2018 FIFA World Cup

Problem statement

The FIFA World Cup is one of the most, if not the most, prominent sporting events in the world. Millions of people gather to watch it in person and billions of people watch the tournament on TV or live streams making it the most widely viewed sporting event in the world. Sharing elements of national pride and relative rare occurrence (it occurs every 4 years as opposed to other events of somewhat similar magnitude like the Superbowl or UEFA Champions League that occur yearly) with the Olympics while still having the commercial appeal and high money stake (a run of good form in the World Cup can be career making and result in contracts and branding opportunities in the 9 figures) of commercial professional sporting events.

Despite being such an important sporting event and the ultimate prize for the most popular sport in the world – the World Cup is relative behind other comparable sporting events in terms of the state of its analytics. In fact football/futbol/soccer analytics in general lags behind the state of the arts for sports such as baseball, basketball, and American football in terms of maturity and development. Because a disproportionate amount of national team matches are part of qualifying for regional tournaments and the World Cup, a large portion a matches a national teams play are against regional opponents. Teams also play friendlies against each other, but those games are taken less seriously than FIFA sanctioned matches. The upshot is that unlike other sports (or even football's professional leagues) where teams play mandated matches against a comprehensive slate of opponents, the amount of structured data is much lower and much of the available data is from segmented sources. As a result, the kind of analytics that would be relatively easy in other sports (compiling a ranking) is much more difficult in world football where a global ranking of national teams might be very useful. That being said, with the rise of companies like Opta, football analytics has increasingly become the lens through which teams, players, fans and media view and interpret the game.

FIFA's answer to the ranking problem in world football, the FIFA Rankings introduced in the 90s, have gained currency as the most common way to predict the relative strength of teams. They have been adopted (especially in the United States) as the standard upon which to base expectations of success in the World Cup. These rankings aren't without controversy, however, as it's not uncommon for them to include counterintuitive results. Moreover there are some suggestions that teams from less prominent federations (e.g. African, North American and Asian teams) are unfairly penalized. Can we create comprehensive and interpretable framework to both accurately represent team strength in the World Cup and appropriately way any factors that may impact team strength and contribute to the result (Do teams from certain continents play well on that continent, do certain fans travel to see their team creating a home climate, do certain teams feel more pressure as a result of fan expectations, are certain FAs more supportive leading to their teams being more prepared for tournaments, etc.)

Project goal: In this project we'll leverage various sources of team and player data to construct FIFA World Cup 2018 prediction models and evaluate those models against the baseline of predictions from simple incorporating FIFA Ranking the measure of team strength.

Data Recources

We've provided you with two datasets to get you started (both courtesy of Kaggle). One is a historical collection of Fifa outcomes between 1872 and 2017. The other is a collection of data Fifa 2018 player data including various ratings, positional data, and some demographics scraped from sofifa.com. We've placed this data in the gitlab repository https://gitlab.com/cs109/worldcup2018. We'll give you access to this repo on request. If you have suggested changes and/or additional data please submit a pull request.

While this data is definitely enough to do some analysis, the general expectation is that a good chunk of this project will involve data collection from various sources, in particular using web scraping techniques to get auxiliary data (to later be converted into features). Some sources include the following:

1. fifa.com - FIFA has a wealth of scrapable data on its site including match outcomes, and historical rankings. This is a good place to start.
2. squawka.com, whoscored.com, sofifa.com, etc, are good places to find statistical information particularly on players, including ratings, demographics, game and league data. You may want to use these to supplement the player statistics data that we provided, especially if you're incorporating historical analysis. For example, we've given you player data for FIFA 18, but if you want to do historical analysis you may need to scrape earlier player ratings and data to get information accurate for earlier time frames.
3. wikipedia.com - may be helpful for venue information, weather, information about countries (GDP/Population), as well as player and national team info.

High-level project goals

1. The first step is to construct your dataset from the given data, scraped data from external sources, and any other sources. Utilize feature engineering and pre-processing techniques to prepare the data for analysis. Feature engineering and pre-processing are especially important. You should put a lot of effort into this part of the project.
2. Next you should accept the FIFA rankings as an accurate representation of team strength, and use the [this] FIFA rankings (with a somewhat limited number additional features) to create as accurate a model as possible of the 2018 World Cup results. While the type of additional features is up to you, some possibilities include features engineered to capture historical trends (European teams tend to do well in tournaments hosted in Europe, South American teams do well in tournaments hosted in the Americas, the host country tends to receive a boost, etc) as well as basic team statistics (average age, number of players playing in top tier professional leagues, etc). In this part of the project the aim is to create a framework for your analysis without the complications involved in creating team strength models from scratch.

3. Using any ata you've collected construct as accurate a model as possible of the 2018 World Cup results. The idea is that you'll create your full model from scratch (in other words without relying on the FIFA rankings as an accurate measure of team strength and minimizing reliance on aggregations from other organizations). Choosing what data to incorporate and how to leverage that data in your analysis will be a major aspect of any approach to a successfully constructing your full model.

4. Compare the results from your full model to the results you obtained from the model that incorporated FIFA rankings. Do they differ significantly? Do any particular features (natural or engineered) contribute to the differences? What features have the greatest influence over your How do you results differ from predicting results by naively leveraging the FIFA rankings? How do your results compare to other predictions and the final results?

References

1. Zeileis A, Leitner C, Hornik K (2018). "Probabilistic Forecasts for the 2018 FIFA World Cup Based on the Bookmaker Consensus Model", Working Paper 2018-09, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, UniversitÃd’t Innsbruck. https://www2.uibk.ac.at/downloads/c4041030/wpaper/2018-09.pdf

2. Lorenz A Gilch and Sebastian MÃijller. On Elo based prediction models for the FIFA Worldcup 2018. https://arxiv.org/abs/1806.01930

3. âĂIJFIFA.com - The Official Website of the FIFA World CupâĎć.âĂİ FIFA.com - FÃľdÃľration Internationale De Football Association (FIFA), www.fifa.com/worldcup/index.html.

4. "Sofifa.com - Players FIFA 18." Sofifa.com - Electronic Arts Sports (EA Sports), https://sofifa.com/