Carlos Flores
Hack Rice 9.0
Identifying Win Conditions for World Cup Matches

Introduction

A championship team can attribute many factors towards their win and match data can be found online in an abundant supply. My main endeavor sought to find whether there are common features amongst the most successful teams. I was not aiming to predict a winner but to find out what attributes proved to be a reliable predictor for a teams win.

The FIFA World Cup was considered given the amount of statistical data and historical data readily available. I found a data set of 100,000 FIFA preliminary matches since 1996. This was the inspiration for the hack and was the starting point to what I have now.

I used the FIFA Rank (from preliminary match records) to infer what predictors would be significant to estimate their placement in the World Cup (championship match records). This was achieved by applying a dummy variable to the World Cup placement in the form of an ELO variable; as seen in various competitive ranking systems. The ELO variable and the FIFA rankings were derived by awarding points to a team for a win/draw/loss. The FIFA organization likes to obfuscate the actual criteria to their calculated FIFA rank. I am not certain on how the FIFA rank is calculated as a result. **The ELO rank is designed so that it would give 9 points for winning, 6 for ending in a tie, and 3 for losing in a match for each World Cup participant.**

The datasets I ended up using are a combination of stats from many sources: Opta, Kaggle, FIFA, IMF, etc.. I included data for all World Cups since 2006. For each country, I have its ELO, PPP, Fouls Committed, Fouls received, Goals Received, Goals Received by Opponent Errors, Goals Scored, Goals Opponents Scored, Penalty Cards Given, FIFA Rankings (before the World Cup), and their Confederation.

Data Preprocessing

The data extraction and preprocessing was time-consuming! It was by far one of the most important factors of this hack because I wanted to develop my Preprocessing skills. The data came from many sources and included

a calculated ELO. While I was able to easily extract and download most of the data, some of the supplementary data proved to be difficult to copy. Therefore I built a web scraper in Selenium and a web crawler in Scrapr to automate the data extraction.

The fifaRankings dataset was initially incomplete because the 2006 World Cup was not found in one contiguous file. The penalties, fouls, and cards had to be scraped individually from FIFAs official website. I used the same scraper with slight alterations to get additional data for the other World Cups (2010, 2014, and 2018).

I was able to find GDP data from the IMF World Bank and created the ELO dataset by calculating the information from the World Cup dataset. The data was not under the same format and the files needed to be uniformly organized. The features, units of measurements, and key values were standardized throughout all of the data sets.

Once the data was obtained it had to be unified. The country names had to be standardized because some references used transliterated formal names or archaic names (like Ireland, Irish Republic, or Republic of Ireland). The tables had to be normalized into BCNF form to avoid overlapping primary keys. Additionally some SQL preprocessing including left joins on the datasets to only include country data for World Cup participants. Whenever data was missing it was imputed by estimates (such as population or GDP data from a year that was close to the targeted estimate). //

Feature Space

The data set includes following features:

Year: The year during which the World Cup occured. It has four possible values: 2006, 2010, 2014, and 2018.

Region: Where the teams come from. For most regions, it is the same as the country they came from. This is a categorical variable that requires one-hot encoding on the data.

ELO: The variable we are trying to predict. It is based on the number of wins, ties, and losses. It is a discrete numerical variable.

Rank: The countries are ranked from 1 to 211 with the highest FIFA Ranking starting at 1 and then going in descending order to where the rank 211 team has the lowest FIFA Rank. The FIFA ranking itself is similar to ELO but it is calculated by the FIFA organization (the formula is not publically available).

PPP: A measure of the financial status of the country. The PPP is typically

considered a continuous numerical variable. It may also be a discrete variable if rounding to the smallest unit of currency is considered.
Fouls committed: The number of fouls the team committed during the world cup. This is an indication of the aggression of the team. This is a discrete numerical variable.
Fouls Received: The number of fouls other teams committed onto the specified team during the world cup. This is an indication of the aggression of the other teams playing in the world cup. This is also a discrete numerical variable.
Goals For: The number of goals a team receives. This includes the number of goals a team scores as well as the number of self-goals a team receives from the other team. This indicates how good a team is. This is a discrete numerical variable.
Goals Against: The number of goals opponents receive. This allows a model to compensate for hard opponents. This is a discrete numerical variable.
Yellow cards: The number of yellow cards a team receive. The number of yellow cards a team receives is one way to measure their aggression. This is a discrete numerical variable.
Double-Yellow Cards: When a team scores two yellow cards, causing a red card. This is one of the indicators showing aggression. This is a numerical discrete variable.
Red Cards: The number of direct red cards a team scores, which is an indicator of aggression. The red cards are a discrete numerical variable.
Rank: The rank is an indication of how good a team is. A good rank is a low number. The rank is a discrete numerical variable.
Confederation: The confederation indicates the region where the teams are from. The confederation is a categorical variable.

Model Selection

Linear Regression,:
Initially I started with a quick scatterplot matrix of a few variables to see if any correlations could be seen or if there were any patterns I could use.