Simon Leahy
Capstone Project 1
Springboard: Data Science Career Track
6/21/2020

Sports are and have for a long time been a big part of global culture. There's something about using physical strength and coordinated finesse that taps into our primal instinct and sparks our desire to compete. Whether watching or playing, the goal is uniform, to win.

Soccer is widely considered the most globally popular sport, yielding the highest viewership and most revenue amongst all others. As a sport heavily reliant on strategy and teamwork, player chemistry is of utmost importance. I plan on using data to determine the degree to which player chemistry plays an impact on team success and what combination of attributes most optimally drives success.

The problem is that in the game of soccer, there is a factor between talent and success, chemistry, that would need to be formally defined so one can measure it. Chemistry is the existence of complementing abilities between players who work together in order to achieve success. Chemistry is also reflected by how strong each team is in specific attributes. Soccer managers are responsible for deciding how to best leverage talent to achieve success, , and identification of the specific attributes that contribute to success more than others can help the manager optimize for success. To address this problem, I will build a machine learning structure that uses the relationship of players skills to extract patterns and combinations that most often achieve success. Attributes such as speed, skill, and vision can be judged by a trained eye after watching a player for an amount of time and a ranking of that player's skill set although often subjective, can be judged and assigned a metric. It is the job of a soccer manager to assemble the best group of players that they can on the basis of attributes given the money that they can provide and use strengths in attributes to their advantage. Managers, who have likely been around the game of soccer their entire lives, can

usually intuit the right combinations of individual attributes to form the best possible team, but can it be quantified into a science? I believe that we can identify the optimal combinations of skill that yield the highest ratio of success to ability.

Client: My clients are soccer managers at the professional level. While the analysis will be performed using data from elite level leagues, the model should be applicable for teams down to the college level. The client will be able to leverage this information for trades and acquisitions or to make adjustments to a current roster.

Data: I will be using data from Footystats, a soccer statistics repository. This site provides data on team statistics, player statistics, and a table with player attributes metrics, all of which will be key to answering the underlying question. The site provides game by game statistics across many leagues all over the world for both players and teams. The site also provides attribute ratings for individual players, which will be the input variables used to assess overall talent of a team.

Approach: Amongst many methods of data preparation and normalization, I will be using several statistical and machine learning methods to both build variables and evaluate the degree to which each attribute contributes to success. The goal will be to first confirm the hypothesis that states that stronger offensive attributes will positively correlate to offensive success. Assuming this is true, I will address the question of which attributes more strongly to success by identifying the strongest attributes of the teams that overachieve based on their expected success vs. actual success metric. There will be several steps to arriving at the success metric, which will involve creating a weighting function to apply to schedule difficulty based on different leagues. Once the predictor and target variables are defined and laid out, I will likely use a variation of multivariate linear regression to identify the team's expected success and isolate combinations of variables that contribute most to the teams that exceed their expectations. I will also extract the standard deviation of attributes of each player in the player sample to identify the attributes from which teams benefit from having a higher variation.

Deliverables: The final product can be code, paper, or slide deck, depending on the capabilities and concern with detail of the end user. The main point to get across are the skill combinations that most contribute to success, but the model itself may change depending on defensive tendencies. In this case, the code can be repurposed to accept inputs from different types of competition.