Simon Leahy
Capstone Project--Final Report
Springboard: Data Science Career Track
7/9/2020

Problem Definition

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 and attribute differential plays an impact on team success and what combination of attributes most optimally drives success. In this project, team success will be defined as offensive points scored throughout a season, as the scope of the analysis focuses on offensive attributes.

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 different complementary abilities that different players possess and use to 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 model that uses the differential between players skills to extract patterns and combinations that successful teams tend to have. 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 through data science? I believe that we can identify the optimal skills that yield the highest ratio of success to ability.

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 jupyter notebook that I developed has been structured such that new teams or updated seasonal data can be added and the analysis will run dynamically based on the new master data files. The client will be able to leverage this information for trades and acquisitions or to make adjustments to a current roster.

I will be using data from Footystats (https://footystats.org/), 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.

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. There will be several steps to arriving at the success metric, which involve creating a weighting function to apply to schedule difficulty based on different leagues. Schedule difficulty is the difference in competition rigor between leagues, and each league is assigned a multiple to adjust offensive success based on how difficult each respective team's league is. This schedule difficulty multiple multiplied by the number of goals scored by each team defines the success metric. Now that the predictor and target variables are defined and laid out, I 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.

- Which attributes most contribute to success
- Which attributes most contribute to overachievement
- Which attributes do teams benefit from having a wide variety of strength levels in (per player)

Data Wrangling

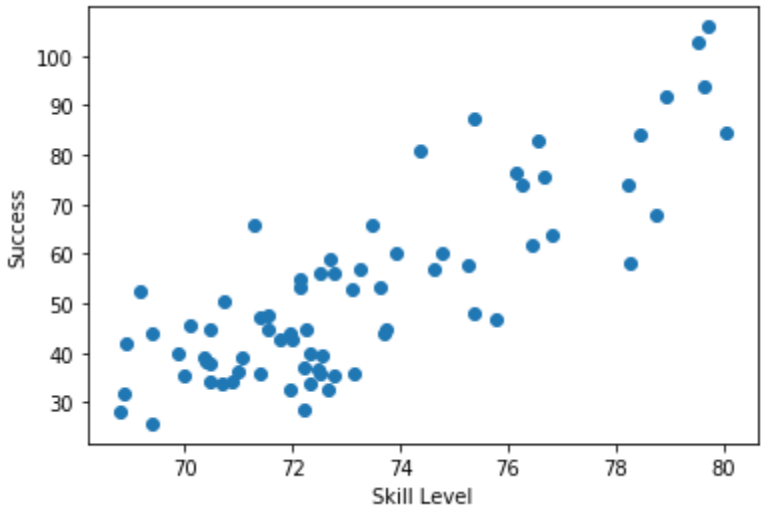
The data cleaning requirements of this project involved both traditional data wrangling functions and ad hoc mathematical applications to manipulate the data into the format necessary to address the question. The first step of importing the csv-formatted data required cleaning, as the dataset included characters native to languages other than English. In order to allow for non-English characters, the encoding argument needed to be set to ‘latin1’. This was applied to the players skill sets csv, players’ teams csv, and teams’ success csv. The next requirement to cleanse the data was to properly weight teams’ success based on the league each team was in. The imported team success data listed all of the goals scored within the 2017/2018 season, but that only includes interleague games. It was necessary to adjust the success metric of goals scored to account for differing difficulties, as otherwise teams with players with lower overall skill sets would yield disproportionately high success metrics. Though the exact weighting is subjective, there are many sources that convey a metric that identify the strength and difficulty of each league. The weighting was in the form of a dictionary,

with league as keys and the key league weight divided by best league weight (i.e. Spain:1662,England:1660.4/1662) rendering each league weight to be multiplied to goals scored as value ranging from 1 to 0.95. The notebook was built in such a way that a different rating system could be easily implemented. After each team was assigned a weighted success metric, the teams were concatenated by row (axis=0) to create a dataframe of all teams and their new offensive success variable.

The next step, to satisfy the scope of the project, was to extract and merge the top five offensively inclined players (as defined by average of offensive skill sets) with their respective teams in order to generate both an average and standard deviation value of those five players skill sets for each team. This was done using sort_values to perform a two-level sort on players by their team and then offensive average, groupby and apply to cut the players data set off at five for each team, pivot_table to aggregate the five players' skill set by average and standard deviation for each team, and merge to join the the new aggregation of players at the team level with each team's weighted success metric. Additionally, column names were wrangled using the rename function and list comprehension to apply "_avg" and "_std" to each offensive attribute where necessary. As mentioned earlier, some teams from the team's data had non-English characters, and thus did not match the english characters in the team field of the players dataset. These rows were dropped using .dropna, as they comprised a small portion of the data. Outliers did not exist within any of the generated data frames and did not need to be addressed.

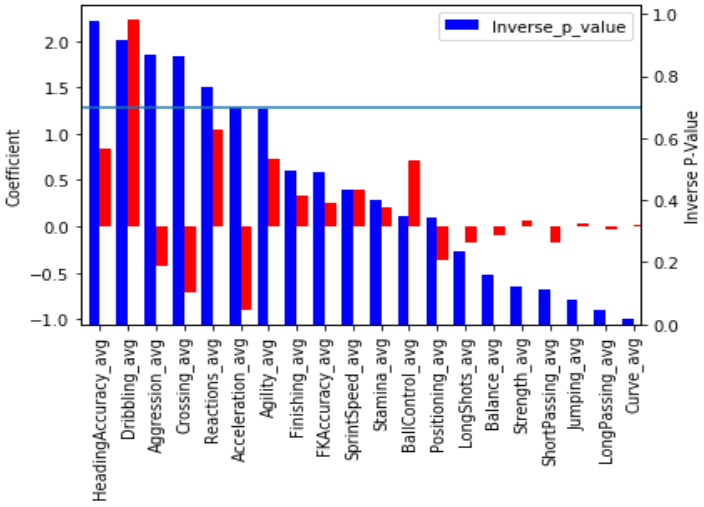
Modeling

The primary assumption that is investigated is the assumption that teams with a higher skill set were more successful. More specifically, teams with a more offensively skilled top-five offensively-inclined unit will yield higher offensive success. This was proven using the np.corrcoef function and scatterplot on the engineered success and skill level variables, that showed both numerically and graphically that there is a strong relationship between the two variables (correlation=.82). This step was necessary to establish a level of confidence in residuals of expected success vs. actual success and would allow me to accurately identify the individual skills that contributed most to the over or underperformance of a team based on expected success.



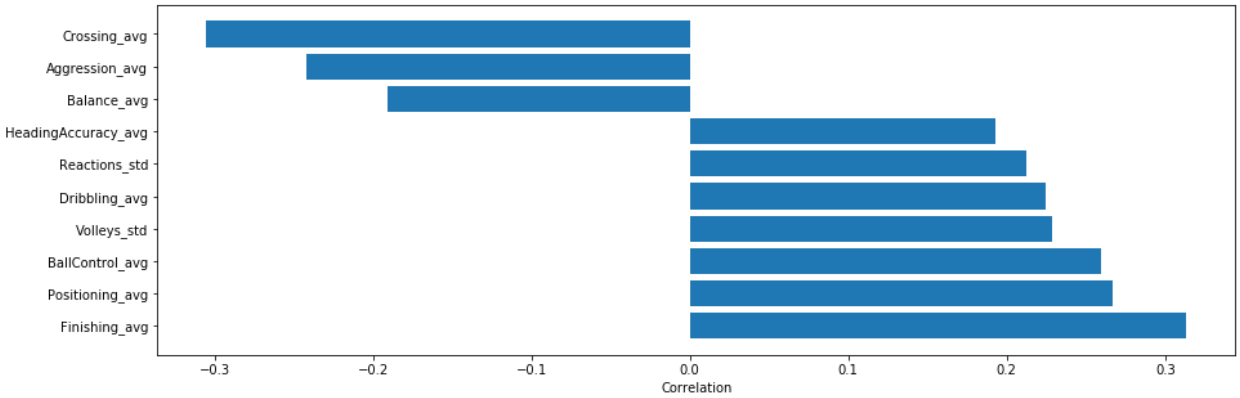
Now that the strength of relationship between skill and success has been established with this dataset, I can look into the individual skills that contribute most to success. This assessment was done using a multivariate linear regression model using each of the team's aggregated attribute metric as independent variables and the

weighted success metric as the dependent variable. The dataset for the multivariate linear regression model consists of 20 different variables that describe 76 different teams' offensive abilities. These variables are preprocessed and formatted to only include inputs of offensive players. The relationship between each attribute and success is determined by coefficient to determine strength of impact and p-value is used to determine degree of statistical significance displayed by each variable within the context of the model. P-value is used to vet the attributes; only variables that have surpassed the selected threshold of statistical significance (<.3) are considered. The grouped bar graph shown below has dual axes with the inverse of the coefficient on the primary y axis and p-value on the secondary y axis. I used inverse so that a higher number would indicate a stronger relationship, which I believe is more easily digestible in a visualization. The visualization indicates that heading accuracy, dribbling, and reactions both meet the pre-set level of statistical significance and have positive coefficients, meaning they can reasonably be considered attributes that contribute to success. Conversely, aggression and crossing both hold a p-value below .3, but have a negative coefficient. This indicates that within the context of the model, teams with higher values in these areas experience less success. The model yielded a .74 adjusted R-Squared and an F-statistic of 1.76e-12. The full scope of p-values and coefficients for each attribute are included in the chart below



The final visualization and one most pertinent to the question at hand is framed similar to the prior visualization. The frame that I am working with includes 1.) the standard deviation of the five players' skill levels in each available attribute, 2.) the average of each skill attribute per team minus the team's overall offensive average, and 3.) the residual between actual success and predicted success based on the prior model. 1 depicts how dispersed each attribute is amongst the top 5 players, 2 shows the extent to which each attribute should be considered a "strength" of that team, and 3 shows how much a team over or underperformed compared to its expected performance based on skill rankings. To illustrate, a team with a dribbling average of 79 but an overall average of 72 would be considered a dribbling-centric team, despite holding a lower score than average in that category amongst its peers. If this team had an expected success metric of 76 but an actual success metric of 85, the fact that dribbling is disproportionately high for this team could be considered a reason that the team overperformed. One instance alone is not enough to draw any conclusions, but if

there is repetition in attribute deltas that contribute to overperformance, it can be reasonably concluded as an important attribute. Variables following the methods of 1 and 2 are correlated against performance metric 3 to determine a detectable pattern. These correlations were generated through an iteration process and those with a correlation of >.2 or <-.2 were considered to be impacting variables. While these threshold values are not typically used to represent significance, the facts that offensive abilities' contribution to success has already been proven and that the target variable is a residual justifies the conclusion that the individual attributes that are higher relative to others can be considered especially impactful on success. Additionally, the notebook is structured such that the user can set the threshold and the visualizations will include variables that surpass them dynamically. Provided is a barchart of all correlations and correlations only of the significant variables. Some of the takeaways are reflective of the multivariate linear regression model of each variable on success, but there are some key takeaways that this specific approach provides. The insights gathered from statistical inference is that crossing, aggression, and balance averages are attributes that negatively correlate with overachievement and heading accuracy avg, reaction std, dribbling avg, volleys std, ball control avg, positioning avg, and finishing avg are most attributes that are significantly positively correlated with success. Additionally, finishing, despite having a minimal coefficient and statistical significance level in the prior analysis, yields the highest correlation of skill delta to success residual. This means that while a finishing might not play the strongest role in success, teams that have a stronger level of finishing relative to other skills do overachieve more often.



Recommendations for the Client

- Heading and dribbling, as confirmed by the multivariate regression model, contribute toward success within the context of that model, but the skill whose excess contributes most to overperformance is finishing. From the stakeholder point of view, this means that a soccer manager looking to improve upon a successful team should ensure his or her team is high in those attributes, but if the goal is to help a struggling team with limited resources outperform, finishing should be the sought after skillset.
- Crossing and aggression are attributes that neither contribute toward success nor overperformance, as those attributes inflate the team overall metric and do not yield any benefit.

Conclusions

I believe that this project has successfully uncovered skill attributes and chemistry combinations that are quantitatively proven have a higher contribution toward

success. More importantly, this project has established a framework for other analytical processes within the soccer or, more broadly, sports community and has the robust nature to field and analyze additional data.

Future Work

Going forward, there are additions that can be made to the notebook that could derive more and deeper insights. Additional team and league data is available and can be added for a more comprehensive and accurate outcome, provided league weightings are included as well. Additional features, such as salary, could be very valuable as well. Salary is an additional constraint, and while intuitively one can assume it follows overall abilities, there are likely attributes for which a player may be over or underpaid, leading to an opportunity to capitalize on attributes that may not yield the highest market value but do significantly contribute to a team's offensive success. Different organizations have different constraints that prevent them from fielding the best talent, so constraint weighting may have to be adjusted at the user level.

- Implement salary as an additional variable with the goal of maximizing team success based on dollars spent per increase in attribute
- Establish purpose for threshold values of statistical significance. Given the limited amount of data, a predictive model could not have been built out. More data would allow the user to prove the significance of a model built on training data on test data and justify or adjust the threshold that is considered statistically significant.

- Include other models (Decision Tree Regression) and increase model complexity
- Extract more data from other leagues or implement bootstrapping concepts to test the model and notebook's robustness