Simon Leahy

Milestone Report

Springboard: Data Science Career Track

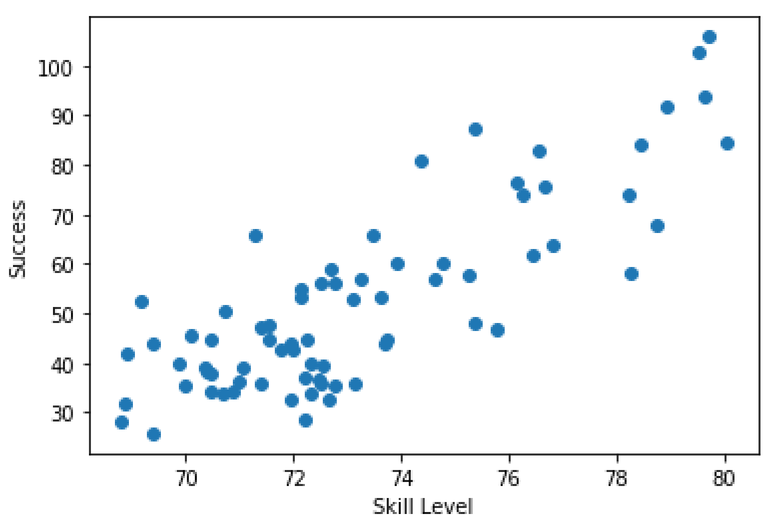
6/30/2020

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.

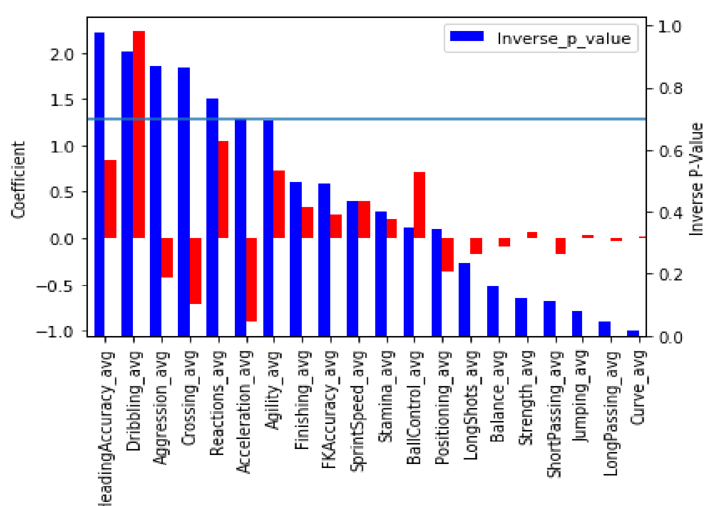
The problem is that in the game of soccer, there is a factor between talent and success, chemistry, that would need to be formally defined so one can measure it. Chemistry is the existence of complementing abilities between players who work together in order to achieve success. Chemistry is also reflected 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. I intend to answer the question of what the attributes are that most contribute to the offensive success of a soccer team, and which attributes and range of attributes amongst teammates increase the likelihood of exceeding expected level of success based on overall offensive skill. 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 and are the target audience for the project, 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.

The data wrangling requirements of this project involved both traditional data preprocessing 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 set 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. As mentioned earlier, some teams from the 'teams' 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.

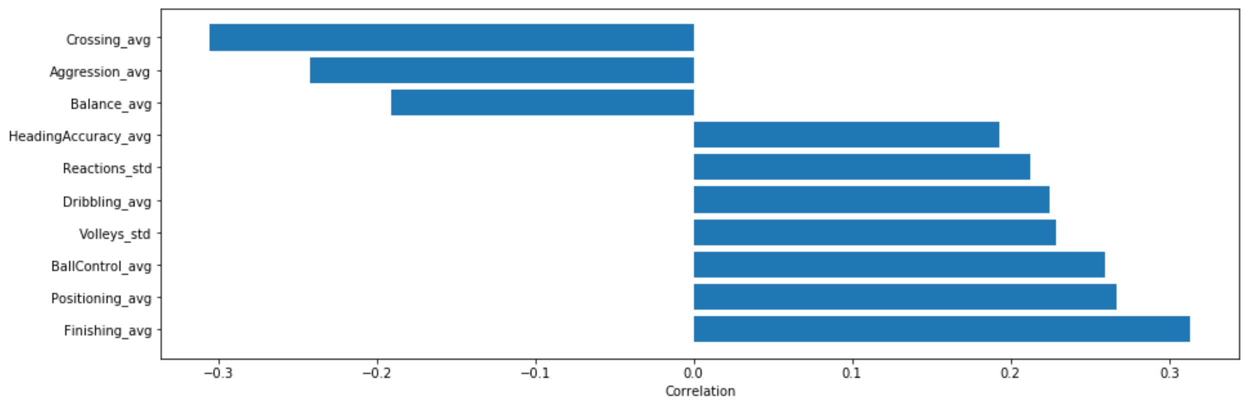
The primary assumption that needs to be proved true is the assumption that teams with a higher skill set were more successful. More specifically, teams with a more offensively skilled top five offensively inclined unit will yield higher offensive success. This was proven using the np.corrcoef function and scatterplot on the engineered success and skill level variables, that showed both numerically and graphically that there was 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 proven, 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 strength of contribution of each was measured by using both the p-value and coefficient of each variable. I displayed each variable on a grouped bar graph with dual axes for the inverse of the p-value and coefficient and after sorting the values by descending p-value. I used inverse so that a higher number would indicate a stronger relationship, which I believe is more easily digestible in a visualization context. Additionally, I will add a horizontal threshold line that denotes variables with an inverse p-value of greater than .7 (<.3 as the significance threshold). The observations that are later corroborated through a separate analysis are that heading accuracy and dribbling both have strong positive coefficient and high level of statistical significance (.02 and 0.08 respectively) and aggression and crossing actually have a negative impact on overall team success as reflected by a negative coefficient.



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 levels in each skill shown and 2.) the average of each skill attribute per team minus the team's overall offensive average, showing how much more or less that attribute is than the team's overall. This number is correlated against the residual of expected minus actual success for each team. Basically, what was each team's relative strength amongst attributes and which of those attributes contributed to out-performing its expected success rate. 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 both 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.


Crossing_avg
Aggression_avg
Balance_avg
HeadingAccuracy_avg
Reactions_std
Dribbling_avg
Volleys_std
BallControl_avg
Positioning_avg
Finishing_avg
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Correlation