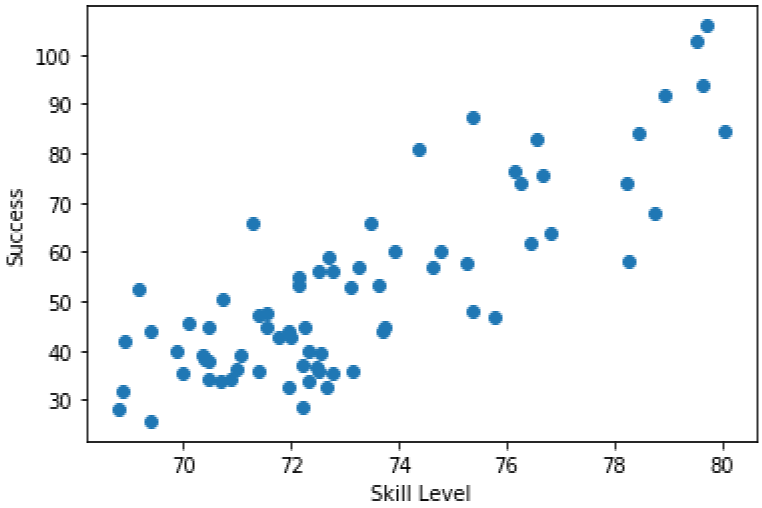
Simon Leahy
Storytelling
Springboard: Data Career Track
6/15/2020

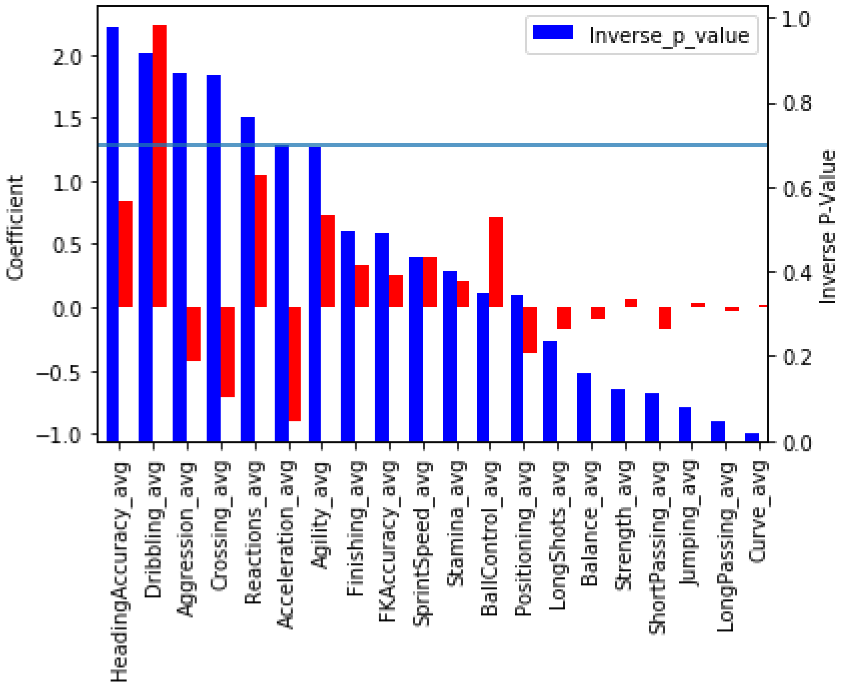
Overarching Question: What are the attributes that most contribute to the offensive success of a soccer team, and what attributes and range of attributes amongst teammates increase the likelihood of exceeding expected level of success based on overall offensive skill?

The storytelling component of the notebook gathers the preprocessed and engineered data and converts it into both insights that can validate or add to the overarching question or help identify invalid data that needs to be further cleaned. The data wrangling portion of the notebook successfully turned extraneous data into a concise dataset engineered to answer a question that requires customized inputs.

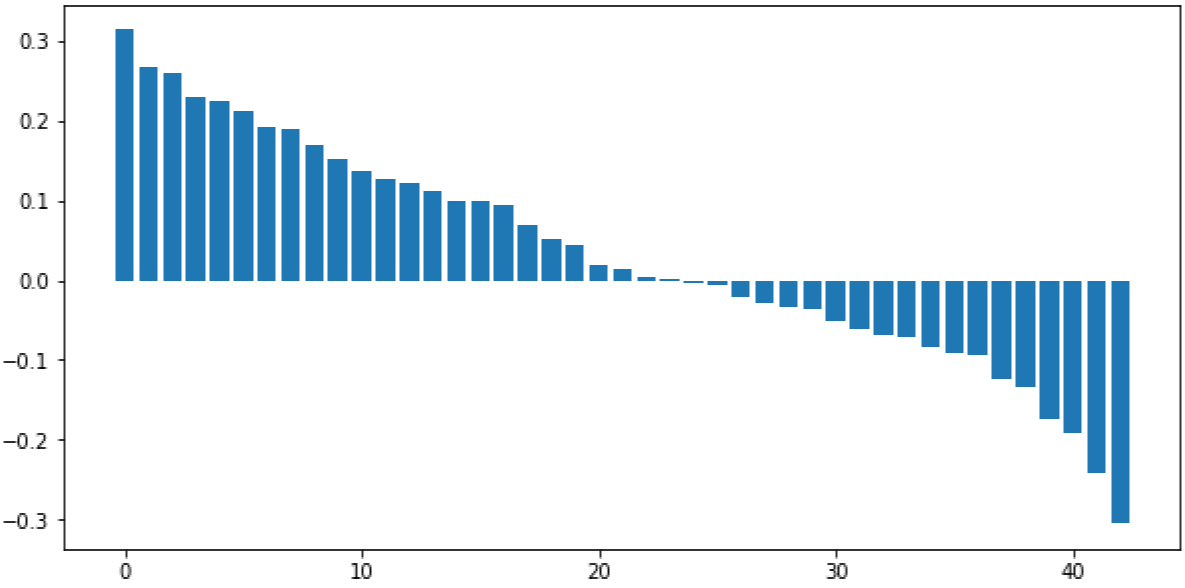
The primary assumption that needs to be proved true is the assumption that teams with a higher skill set were more successful. More specifically, teams with a more offensively skilled top five offensively inclined unit will yield higher offensive success. This was proven using the np.corrcoef function 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.

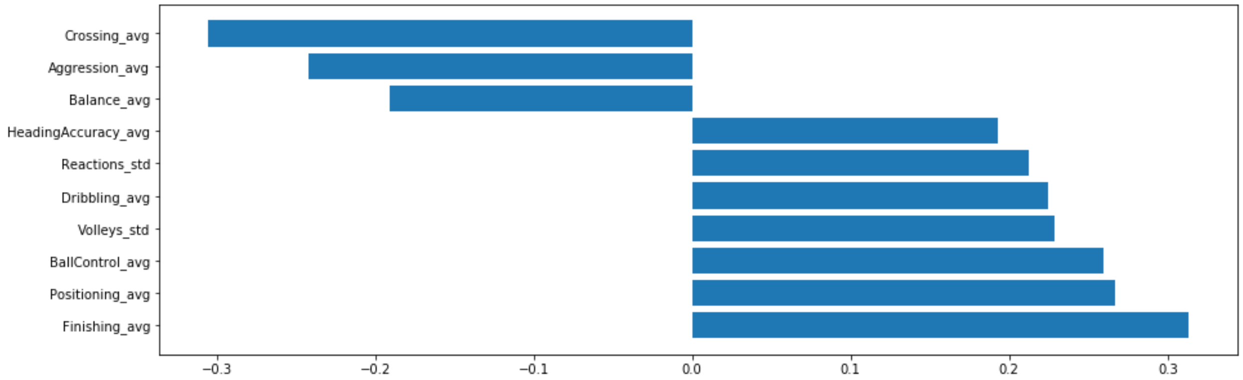


Now that the strength of relationship between skill and success has been proven, I can look into the individual skills that contribute most to success. This assessment was done using a multivariate linear regression model using each of the team's aggregated attribute metric as independent variables and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


0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
10
20
30
40


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