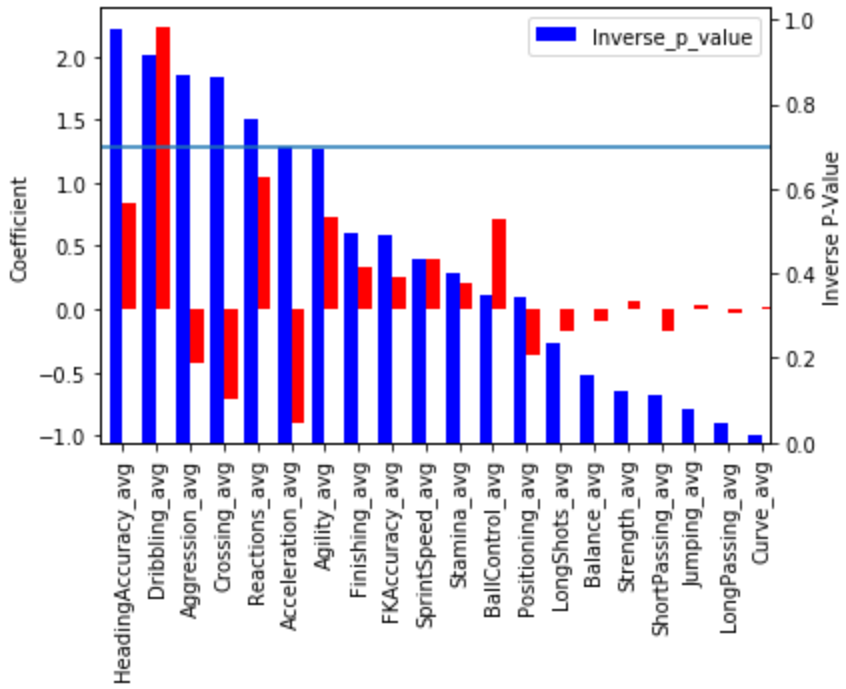
Simon Leahy
In-Depth Analysis
Springboard: Data Science Career Track
7/7/2020

Overarching Question: What are the attributes 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?

The machine learning portion of the notebook uses engineered features and variables from preprocessed data to extract insights into which attributes and combinations of attributes contribute most to offensive success. The raw data, player attributes and team success, is formatted to identify the top five offensively inclined players on each team and use those five players as representatives of the team's skill metrics, and uses those metrics to formulate a way to understand each metric's impact on offensive success. The machine models used are multivariate and univariate linear regression, using average skill rankings and overall skill rankings respectively to predict success outcomes. The scope of the project is not necessarily to predict outcomes, but rather dig into the impacts that each individual skill variable has on success. The univariate regression analysis is used to prove the prerequisite hypothesis, does an increase in skill level contribute to success? This proved to be true with an R^2 of .67 and a positive correlation of .82. Once this hypothesis was proven true, I was able to use individual attributes to assess ability's impact on success at a more detailed level and ideally, render a stronger model.

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 of each variable on success is assessed by coefficient to determine strength of impact and p-value to determine degree of statistical significance. The overall R^2 of the model is .81, indicating the individual variables accurately predict success. The values are

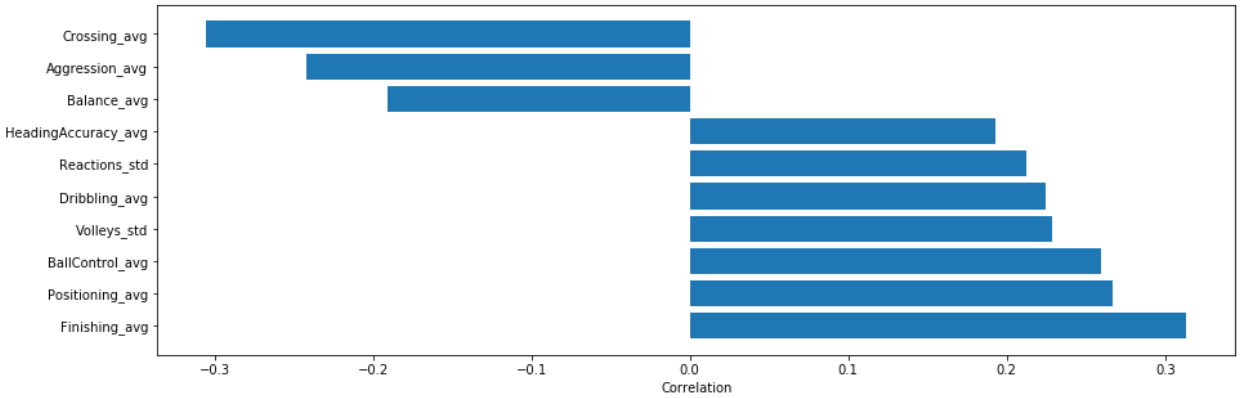
collected and included in the chart below, (summary statistics are visible in attached notebook):



The takeaways are that heading accuracy, dribbling, aggression, crossing, and reactions all have a statistically significant relationship with the success of a soccer team. The coefficients indicate that within the context of the model, only heading accuracy, dribbling, and reactions have a positive impact on the offensive success of each team. While this may seem counterintuitive, this concept is explored in a later analysis.

Next, the model's predicted values are used to compare with each team's actual offensive success to create a list of residuals. The residuals between predicted and actual success are correlated against the delta between each of the teams' individual attribute average and overall average. This output provides clarity to the negative coefficient for aggression and crossing, as it is proven from the correlations run that team's with a higher average of those attributes above their overall average are more likely to have a lower actual success metric than predicted success metric, meaning

they underperform relative to what their overall attributes would otherwise indicate. The correlations of significant residual and skill deltas are shown below:



Finishing, though not proven statistically significant in the linear regression model, has the highest correlation with a high residual. This means that though finishing is not necessarily an attribute to build around, teams outperform their expected success if their finishing average is higher than other attributes.