Simon Leahy
Inferential Statistics
Springboard: Data Science Career Track
6/21/2020

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 inferential statistics portion of the notebook uses raw data, engineered features, and machine learning outcomes as inputs to answer the overarching question. The goal is not necessarily to build a predictive model, but rather to use predictive modeling to assess the strength of impact each variable has on the residual between expected and observed success. To that extent, the variable and outcome pairs are 1) attributes and their impact on success, 2) the difference between each team's attributes and overall average (further referred to as "average delta") and their impact on the residual between success and expected success, and 3) the standard deviation of attribute values of the top five most offensively inclined players from each team (further referred to as "attribute stdev" and its impact on the residual between success and expected success. The assumption that the residual can be looked at as a feature in addition to a margin of error is based on a high R^2 value of .82 for the univariate linear regression model of offensive overall on success and R^2 of .81 multivariate linear regression of individual attributes on success.

The statistical inference required in the preprocessing steps involved feature engineering to convert team success from a variable of simply goals scored to a more comprehensive variable that more accurately portrayed team success. Each team's goal total was multiplied by a scaling system used to weigh total goal count by the difficulty of the team's respective league. These weights are largely subjective, so the notebook is structured such that the user can adjust the weights declared as "league rankings" to reflect a different viewpoint of league difficulty. This process is also critical for the notebooks ability to encompass additional data should it be available, as weaker teams in weaker leagues can be included and their talent to success ratio will be normalized.

The collection of correlations of average deltas and attribute stdev are iteratively produced and reflected via zip function into a tuple of the attribute and it's correlation to success residual. The residual represents the degree to which the team has over or underperformed based on the attribute dependent success variable multivariate regression model. These correlations are compared against one another and visualized

in their entirety to gauge the range and distribution of different values. Next, a default but user-adjustable correlation threshold is set to identify the variables that are most correlated to exceeding expected success or underperforming. 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. Furthermore, ball control, finishing, and positioning are attributes whose overall averages don’t significantly contribute to overall success, but whose deltas do contribute to a high expected success to success residual. This means that players with these attributes can help a team overperform given lower scores in other attributes.