ORIE 5741 Midterm Report

Bjorn Teo, Ignacio Arevalo, Prat Kothiyal

Introduction

Since our initial proposal, our group has decided to shift from studying player's market value obtained from transfermarkt.com to the player's free market value found in the FIFA dataset. This decision is based on our comparison between the two sets of player values and noting their similarity.

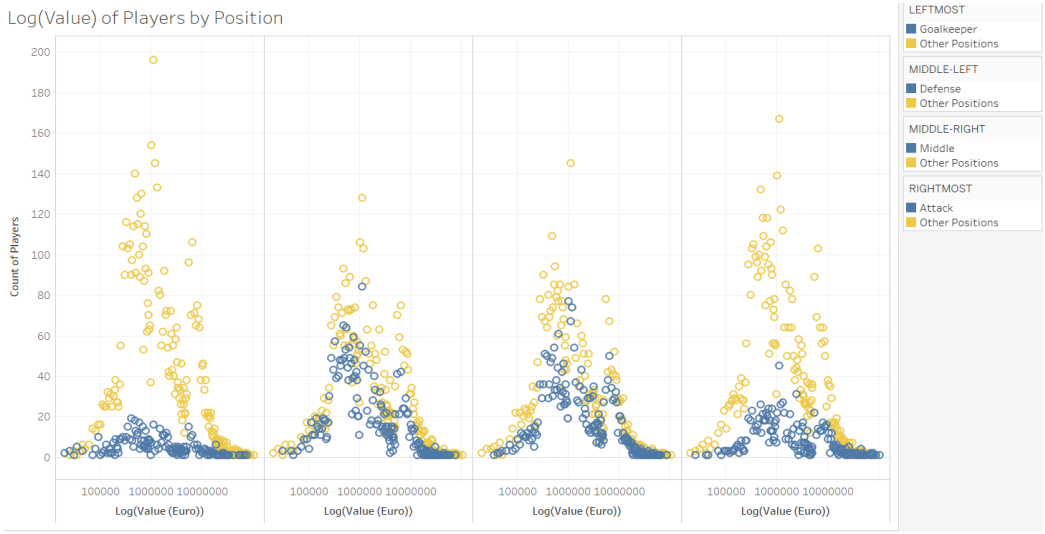
Through our analysis of the data, we found a large difference in the skills present in two categories of players: goalkeepers, and outfield players. In the dataset, goalkeepers have separate attributes that outfield players do not, and they lack the main six physical attributes of outfield players. Thus, we have decided to separate our data based on these 2 categories of position.

Pre-processing

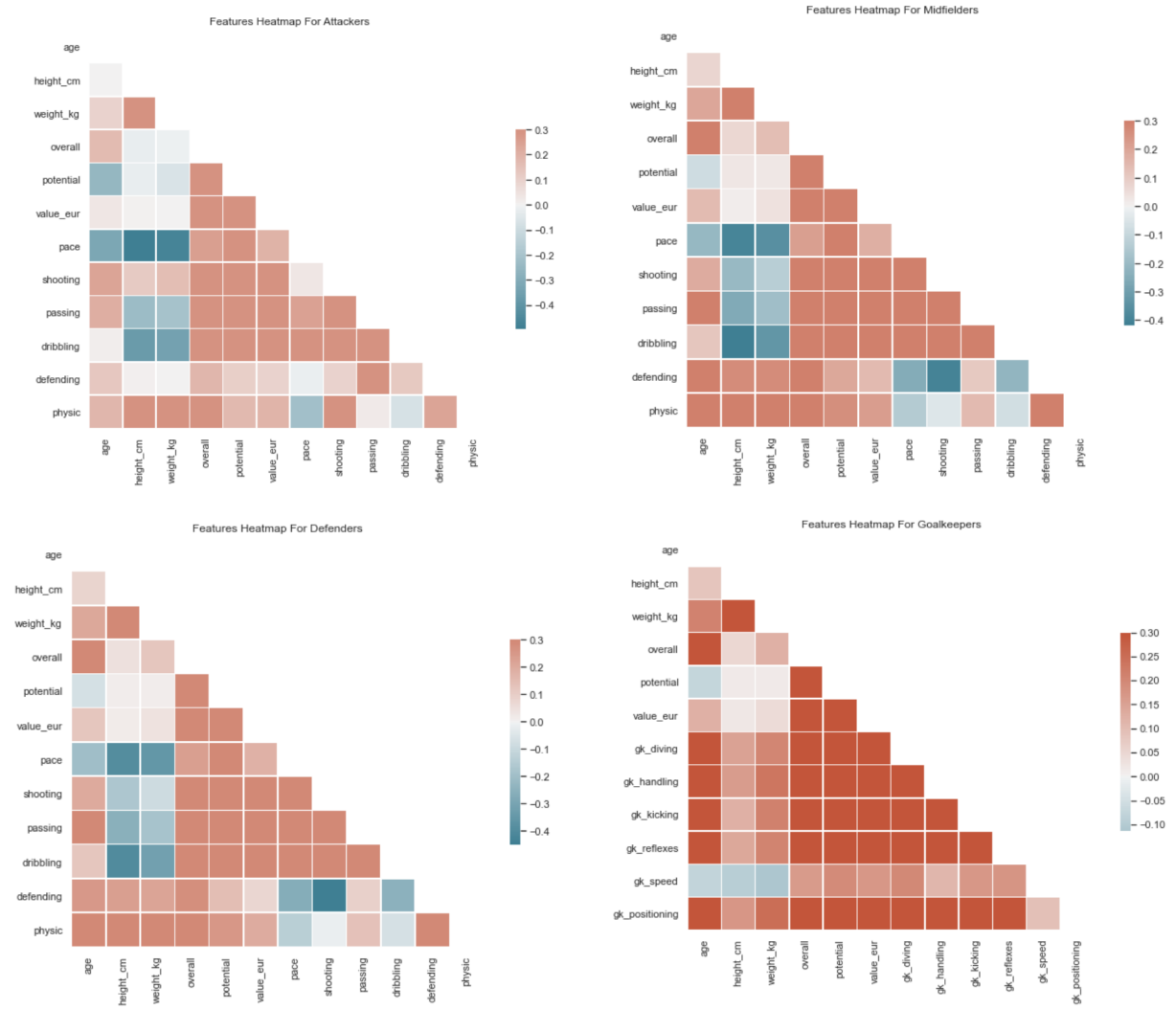
We chose to filter out rows of players to keep only those who have a non-zero market value recorded. Those players with a market value of zero recorded in the data are assumed to have unknown market value, and thus we omit them from our analysis. We found that 31% of our dataset has players with unknown market value, but altogether our non-omitted records are plentiful, being between 10,000 to 15,000 per year over five years of data. For all other null values, we replace these with zero, however of the features we select, most of these are not null as expected. We also choose to transform the market value field to its natural-log, as the net worth varies by several orders of magnitude, and using the natural-log would make a linear trend more apparent. Finally, we choose to transform one of our features, the player_traits feature, into a many-hot vector, as in the original data this column records a subset of traits that each player has been assigned to in the game. Though not included in our initial regression model, we plan to implement this feature in the future.

Based on our proposal, we have selected features that represent ratings of each player's skills, in addition to some biographical information that we believe would be relevant to player performance. These initial features are: age, weight, height, overall rating, potential (i.e, the max rating achievable for a given player), and the six core skill areas (for outfielders: pace, shooting, passing, dribbling, defending, physic; for goalkeepers: diving, handling, kicking, reflexes, speed, positioning). We note that the data also has sub-skills, which are all components of each of the six skills, and whose weighted average equals the above skills. For example, movement_acceleration and movement_sprint_speed are the two sub-skills that make up pace; if we include them along with pace in our input features, it would be redundant as together they represent a linear combination of the pace column.

Exploratory Data Analysis



Above, we used Tableau to visualize the log of market value across all players in our dataset who belong to one of four general positions (in order with the charts below): goalkeepers, defenders, midfielders, and attackers. In general, player value appears normally distributed across all positions. Since there is only one goalkeeper per team, we see a smaller concentration of goalkeeper counts around the mean, but there still is a normal distribution for the goalkeeper market value scatter plot.



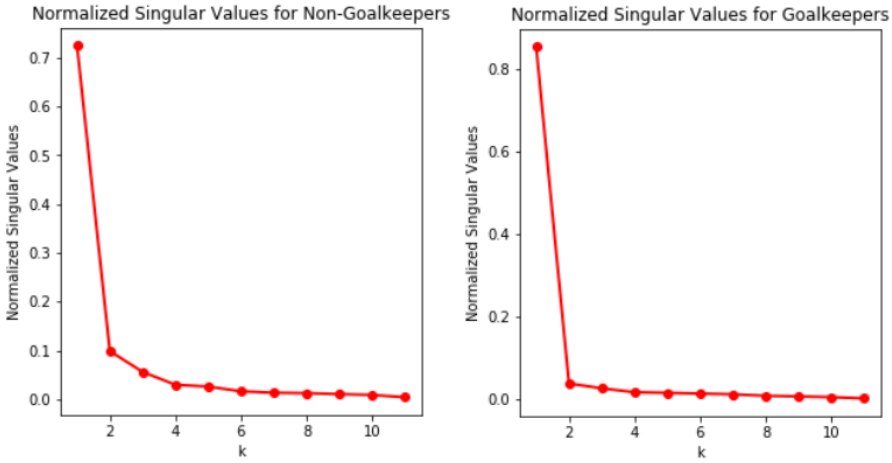
Correlation Heatmaps - Attackers, Midfielders, Defenders, Goalkeepers

The above correlation heatmaps show the relationship between the features we have chosen for the subgroups of players - Attackers, Midfielders, Defenders and Goalkeepers. Many of the features share a positive correlation, which we believe is to be expected as most of these athletic traits are shared with more skilled players. Height and weight are negatively correlated with many offensive stats, which we also believe is to be expected in soccer players. Due to the nature of many positive correlations, we may see that our linear model can be built using a smaller number of our initial proposed features.

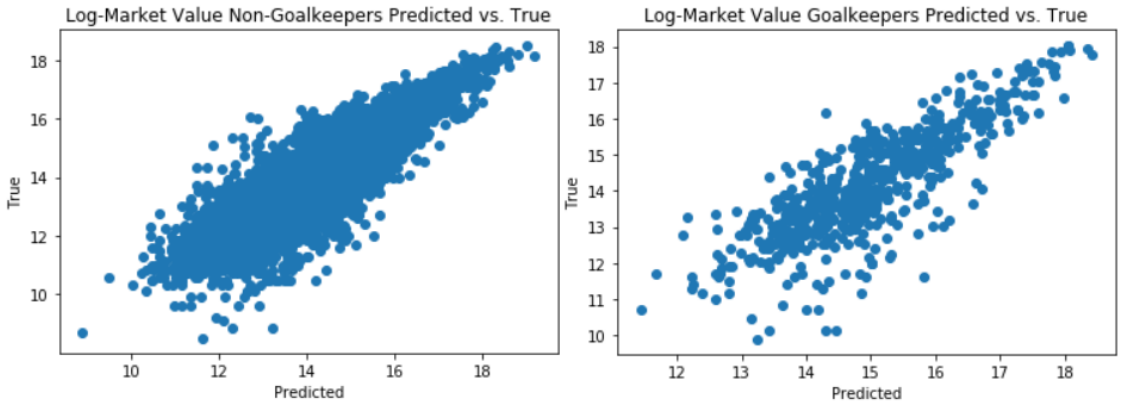
Preliminary Model and Results

To do our PCA we decided to merge our data sets from FIFA 2015 - 2020 to try and predict market value using one year's features to try and predict the respective player's market value next year. To do so we separated the whole data set into random subsets of 65% training, 15% validation, and 20% testing. We did this twice: once for the dataset of all players not including goalkeepers, and then again for the dataset but only keeping goalkeepers since they have six different core skill areas.

After separating the datasets then we found the SVD of the training set and did some analysis on the singular values including finding the respective eigenvalues. Here we see the normalized singular values of which we can see that for non-goalkeepers we should consider keeping 4-6 features as that is when the singular starts flattening out, while for goalkeepers we consider keeping 3-5 features for the reasoning. Choosing from this subset allows us to avoid over and underfitting while choosing the best possible features from our selection to predict future market value.



We then used our validation set to see how well our training set fits a linear regression model using 4-6 features of the non-goalkeepers, and 3-5 features of goalkeepers. Then based on the validation set output of mean squared errors we kept the number of features that managed to get the smallest MSE. For the case of non-goalkeepers the validation set chose 6 features to keep, while for goalkeepers the validation set chose to keep 3 features. We ran this not only on the market value but also the natural-log of it as described above and found that the validation set still chose the same number of features for non-goalkeepers and goalkeepers, respectively. We found that the linear regression fit against the natural-log of market value performed the best for non-goalkeepers and goalkeepers.



Even though these fit considerably better than not taking the natural-log, after taking the exponential we see that our mean squared error is still along the magnitude of 10 trillion, for non-goalkeepers, and 35 billion for goalkeepers (compared to 2 quadrillion, and 4 trillion, respectively). This makes us think that performing a linear regression is not the best way to go about it, giving us some ideas for future plans.

Future Plans

We plan to use Random Forest and k-NN models to predict market value in the future, as these methods can help us categorize skilled players in market value tiers. We will also plan to add Country and Player Traits to our feature set when we begin to build these models.