Nagaprasad Rudrapatna

Predicting the Market Values of Soccer Players

For my final project, I decided to work on a statistical modeling problem. Specifically, I was interested in developing a model that could accurately predict the market values of soccer players based on their attributes. The dataset I used was created by Tan Pengshi Alvin (link to the GitHub repository where the file is located: <https://github.com/tanpengshi/Metis_Project_2_FIFA_Players/tree/master/Scripts%20and%20Data>). Alvin scraped (using the *BeautifulSoup* Python library) information about 19,402 professional soccer players (market values, age, height, weight, FIFA skill ratings, etc.) from the FIFA Index website and then created seven summary attributes for each player. These attributes (ball skills, passing ability, defensive ability, mental qualities, physical qualities, shooting ability, and goalkeeping ability) were calculated by taking the average of the skill ratings corresponding to each category; for instance, a player’s ball skills attribute was the average of his ball control and dribbling skill ratings.

In my analysis, I modified a few aspects of Alvin’s dataset. First, I decided to exclude a player’s weekly salary. It is well-known that a player’s salary is not a good predictor of his market value, especially if he is approaching the end of his contract (since salaries are fixed years before a player’s market value declines/rises). Next, I decided to exclude players whose market value was listed at 1,000 euros. It turns out that 1,000 euros is the minimum FIFA market value and, based on the data, it is usually assigned to any player above the age of 40. However, there were a few players (like Gianluigi Buffon and Hilton) valued at 1,000 euros who had relatively high overall FIFA ratings. These outliers would have significantly influenced the model fitting procedure (I observed this initially), so I decided to remove them from consideration.

Before beginning the model fitting procedure, I randomly split the dataset into two parts: 70% of the players were assigned to the training set and the remaining 30% were assigned to the test set. I had to modify the training-test split to fit locally-weighted least-squares (LOESS) regression models since the *loess* function would only function properly if the best and worst players according to each feature were in the training set (the upper and lower extremes had to be included in the model fitting procedure). For example, since Lionel Messi and Li Xuebo were the highest and lowest-rated players (overall FIFA rating) in the dataset, they were assigned to the training set.

In this analysis, I considered least-squares regression, regularization methods, like ridge and LASSO regression, Gamma regression (appropriate when modeling a positive, continuous response variable), LOESS regression, and generalized additive models (GAMs). I also performed variable selection (backward/forward stepwise selection) on each model and tuned the relevant hyperparameters for the LOESS regression models and the generalized additive models, namely span and degree. The models were fit to minimize training mean-squared error, and their predictive power was assessed according to the magnitude of the test mean-squared and mean absolute errors (if two models had very similar test mean-squared and mean absolute errors, performance on high-value players was the deciding factor). Least-squares regression is typically appropriate when the response variable can adopt any value. In this case, the response variable of interest was a player’s market value, which is strictly positive. So, to apply least-squares regression, the response variable had to be transformed (the residual plots indicated an exponential trend in market value, justifying a logarithmic transformation of the response).

I found that the optimal log-transformed least-squares and Gamma regression models (both given by forward stepwise selection) consistently overestimated the market values of medium to high-value players. Applying the ridge regression penalty to the optimal log-transformed least-squares regression model corrected this issue; however, the ridge regression model’s predictive power was relatively low. The optimal LOESS regression model was a function of a player’s shooting ability, mental qualities, age, and overall FIFA rating, and it achieved decent predictive performance. The optimal GAM was a linear combination of three bivariate LOESS regression fits (that involved a player’s mental qualities, overall FIFA rating, shooting ability, age, and potential FIFA rating) and one univariate LOESS regression fit (that involved a player’s goalkeeping ability), and it outperformed all other types of models. The optimal GAM was particularly successful in predicting the market values of low to medium-value players (but had similar predictive power to the optimal LOESS regression model for high-value players). The optimal GAM overestimated the market values of high-value goalkeepers (this is a common issue in the literature) but not to the degree of previous models, likely due to the addition of the univariate LOESS regression fit. The results of an Analysis of Variance (ANOVA) test confirmed that there was sufficient evidence (i.e. negligible p-value) to conclude that there is a nonlinear relationship between a player’s market value and his attributes. Finally, based on the terms in the optimal GAM, a player’s mental qualities (e.g. composure, aggression) greatly influence his market value. This is a somewhat surprising result! Perhaps this is related to a player’s availability. For instance, if a player is overly aggressive and engages in rash tackles frequently, then he is more likely to receive a red card (which would impact his availability for future matches). It is well-known that, if a player is not available for selection on a consistent basis, then his market value will decline.