Who's Overpaid? Predicting Player Value Using FIFA Data

Rye Julson, Nico Manzonelli, Jeffery Mayolo, Joseph Zuccarelli

December 1, 2021

1 Introduction

FIFA is a popular football simulation video game produced by the company EA Sports. Every year EA Sports releases a new version of the game that includes the real-life players on each team. Each player included in the game comes with a large set of attributes concerning their physical features, athletic ability and skill level. For instance, a few popular attributes are height, weight, pace, shooting, passing, dribbling, defending and physical. In real-life, each FIFA player is payed a specific salary value by team ownership based on their production level on the field. However, is it possible that given the large amount of FIFA data concerning each player, we can build a model that provides better estimates of player salaries than human owners? In the following study, we attempt to predict player value from various player physical, athletic, and skill attributes using machine learning models trained on publicly available data from the video game FIFA.

2 Exploratory Data Analysis

2.1 Collection and Cleaning

The website "sofifa.com" provides relevant and updated FIFA data for each player. After identifying several interesting quantitative and categorical variables for further exploration, we used web scraping via BeautifulSoup to extract 12,000 relevant player variables. We chose to focus on the following variables: player ID, name, origin country, club, primary position, age, height, weight, primary foot, value, attacking, skill, movement, power, mentality, and defending. Player ID is a unique identifier for each player. Our response variable, player value, is the salary amount for each player in euros. Our predictor variables – attacking, skill, movement, power, mentality, defending – represent numerical scores for each player attribute.

After collecting the data, we shifted our focus to cleaning the data in order to enable for further exploration. First, we removed any repeated rows based on player ID. This narrowed our data set to 11,915 unique athletes. Note that we also removed goal keepers from our analysis because they are typically judged based on other variables than the variables of interest in this particular study. This narrowed our data set to 10,792 athletes. Next, we discovered that all observations collected from“sofifa.com” were considered characters; therefore, we properly transformed all quantitative predictors to integers. Finally, we confirmed that our data set does not include any missing values.

2.2 Exploration

First, we examined the summary statistics for all the variables included in our data set (see **Table 1**). Notice that the distributions of all the variables included in our data set appear to be relatively symmetric based off the mean and median values, except for defending and player value.

Table 1: Numerical Summary

	Age	**Height (cm)**	**Weight (kg)**	**Attacking**	**Skill**	**Movement**	**Power**	**Mentality**	**Defending**	**Player Value (euros)**
Mean	26.03	180.58	74.87	289.23	298.66	340.05	325.98	291.30	162.86	4.32×10^6
Median	26.00	180.00	75.00	291.00	302.00	345.00	327.00	291.00	186.00	1.70×10^6
STD	4.17	6.57	6.73	40.79	48.66	43.44	33.37	35.86	54.03	9.30×10^6
Min	16.00	156.00	53.00	165.00	140.00	168.00	184.00	180.00	35.00	0.00×10^0
Max	42.00	203.00	110.00	437.00	470.00	461.00	444.00	414.00	272.00	1.86×10^8

Next, we refined our exploration to the response variable, player value. **Figure 1** included below displays the distribution of player value for all the observations included in our data set. Notice the strong presence of outliers in the normal response boxplot on the left. Therefore, we used a log transformation to transform the response variable, as this will allow us to validate the normality assumption of any linear regression model that we may choose to fit in the future. The right plot included in **Figure 1** shows the distribution of the response variable after a log transformation. Notice that this distribution appears to be much more normal.

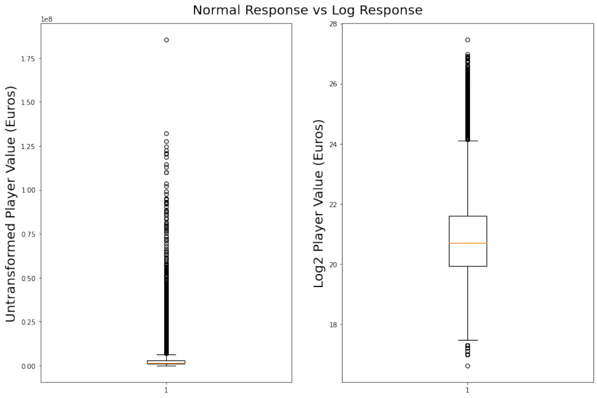

Figure 1: Distribution of Player Value (Regular Scale vs. $\log_2$ Scale)

Finally, we explored the relationship between the response variable, player value, and a few of the predictor variables included in our data set. **Figure 2** included below displays the relationship between player value and the following variables: age, height, weight, attacking, skill, movement, power, mentality, and defending. Notice that there appears to be a positive relationship between player value and the various player attributes, as expected.

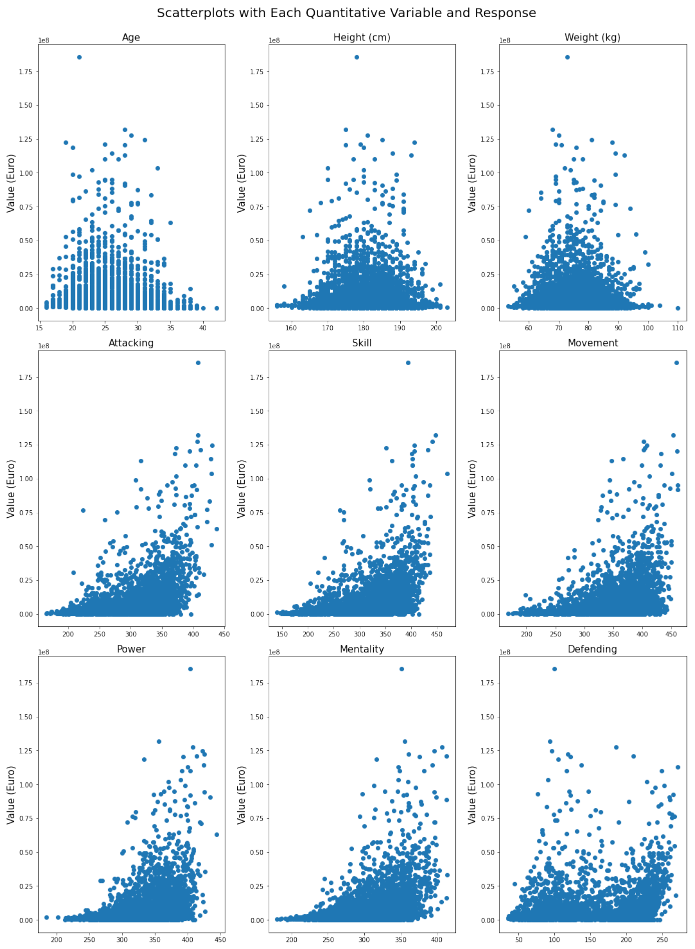

Figure 2: Player Value vs. Predictor Variables

3 Preliminary Model

After exploring our data set, we fit a baseline model to quantify the relationship between our response variable, player value, and the various predictor variables included in the data set. Our baseline model is a multi-linear regression model that is structured as follows:

$$\begin{aligned} \hat{Value} = {} & \beta_0 + \beta_1 Age + \beta_2 Height + \beta_3 Weight + \beta_4 Foot + \beta_5 Attacking \\ & + \beta_6 Skill + \beta_7 Movement + \beta_8 Power + \beta_9 Mentality + \beta_{10} Defending \end{aligned}$$

Using an 80-20 train-test split, this baseline model achieves a training set MSE of $61.27 * 10^{12}$ with an R^2 value of 0.2929 and a test set MSE of $61.35 * 10^{12}$ with an R^2 value of 0.2871. Although this model performs poorly, it serves as a solid baseline model for comparison with the various other types of model that we plan to fit in the future (tree-based methods, regularization methods, etc.).