Supervised Machine Learning - DS5220
FIFA Player Assessment Model & Analytics
Project Milestone 1
03.31.2020

Akshit Jain
Naga Santhosh Kartheek Karnati
Praharsha Singaraju
Thomas Lindstrom-Vautrin

The goal of the project is to develop an assessment model of football players based on their skills, physical attributes and market value to support transfer decisions of football clubs. After preliminary exploratory data analysis we identified a specific set of hypotheses to establish relationships among player attributes in order to develop the player assessment model. The deliverables for the first milestone include findings from the exploratory data analysis and the subsequent results that inspired the learning algorithms.

In the first milestone, we addressed a few low and high risk hypotheses mentioned in the abstract. The results for the low risk hypotheses include the following:

(i) tall, short and strong players are statistically good at heading, dribbling and tackling respectively. **Methodology** - Use scatter plots for heading accuracy vs. height, dribbling vs. height, physique vs. standing tackle and physique vs. sliding tackle. **Findings** - Scatter plot confirms that as the height of the player increases, the player's heading accuracy increases . One interesting finding here is the presence of two clearly distinguishable clusters (Fig-1), where a small cluster contains tall players with low heading accuracy. Upon further investigation, we observed that the small cluster contained only goalkeepers, reserves and substitute players. Filtering out these records, we observe a good linear fit for height against heading accuracy (Fig-2). Scatter plot confirms that as the height of the player increases, the player's dribbling ability decreases. Scatter plot does not conclusively prove that strong players are good at tackling. Moreover, the correlation coefficients between physique and standing tackle, physique and sliding tackle are 0.48, 0.45 respectively. Hence we refute the hypothesis.

(ii) player wage and age are positively correlated upto the age of 31 and negatively correlated after that. **Methodology** - Use scatter plot for wage vs. age. **Findings** - Scatter plot illustrates that the wage increases until the age of 28 rather than 31, before decreasing again. This is confirmed by fitting a linear model for wage against age. Although, we observe an increasing and decreasing trend in wages for players with age less than 28 and more than 28 respectively, it is not a good fit because player wage depends on many other features apart from age.

(iii) there exists a positive correlation between player rating and value. **Methodology** - Use scatter plot for rating vs. value. **Findings -** Scatter plot shows that as rating increases, there is an exponential growth in the value of the player (Fig-3). This is confirmed by observing a linear fit on the log-transformation of value against rating (Fig-4).

The high risk hypotheses include, **(i)** left footed players have a higher overall rating compared to right footed players. **Methodology -** Use one-over-one horizontal box plots of rating for left and right footed players. **Findings -** Upon observation there isn't a discernible difference in rating between the two boxplots. Moreover, the mean rating of left footed and right footed players is 66.7 and 66.1 respectively. Therefore, we refute the hypothesis.

(ii) the top two teams with the highest rated starting eleven for a given year reach the semi finals of Champions League that year. **Methodology** - Find the mean overall rating of a team by picking the best player for each position. **Findings -** Comparing it to the actual tournament, this hypothesis fails for three out of the five years. The idea behind this hypothesis was to model the best playing eleven using player attributes. One can come up with this by finding the player with the highest rating for that position. However, it is not possible to include player relationships as a feature which plays a vital role in team sports.

Keeping our primary goal of building an assessment model and solving various machine learning problems, we further explore ideas not mentioned in the abstract to identify features that influence target variables (i.e. predicting **value**, classifying **work rate**, **nationality** and **player position**). Next, we provide results of how some of the features relate to the target variables.

Based on domain knowledge, we know that the value of a player can depend on several variables. In order to determine the best combination of variables that definitively predict the value of a player, we use forward stepwise selection. The result from forward stepwise selection shows that eight predictors (age, overall, potential, wage_eur, international_reputation, release_clause_eur, power_stamina, defending_sliding_tackle) sufficiently predict value. We also see from the graphs of BIC, RSS, Cp and Adjusted R^2 against the number of predictors that the reduction in error is negligible from five predictors (Fig-5). Hence we choose five (overall, potential, international_reputation, release_clause_eur, power_stamina) as the optimal number of predictors for predicting player value.

For classifying the target variable work rate, our hypothesis is that players with a high work rate earn more than those who work less. Work rate is defined as how much a player works in improving his attack and defense skills. Among the nine categories of work rate, boxplots of wage for High/High, Medium/Medium and Low/Low indicate that there is a significant difference in the wages of players who work more compared to others. Furthermore, the average wage for players with High/High, Medium/Medium and Low/Low work rates is €19631, €6860 and €2971 respectively.

For classifying the player position target variable, our hypothesis is that attackers earn more compared to midfielders and defenders. Based on the position a player plays, they are categorized as attackers, midfielders or defenders. If a player is winger, striker or a forward, he is categorized as an attacker. Similarly if a player is wingback or a centre back he is categorized as a defender. Using one-over-one horizontal box plots of wage for attackers, midfielders and defenders we observed a discernible difference in the wages of attackers compared to players playing in other positions. Furthermore, the mean wage of attackers, midfielders and defenders is €17596, €12802 and €11209 respectively.

Our proposed scope of work for Milestone 2 includes, continuing work on the regression model for predicting player value and building models to **(i)** classify nationality based on skill attributes, **(ii)** classify player position using physical attributes and wage, and **(iii)** classify work rate using defense, attack traits and wage. We hope to identify significant skill attributes, physical and attack attributes and defense attributes that impact player nationality, position and work rate respectively. Since the data is in high dimension, we will use dimensionality reduction techniques like PCA to capture the variance in the data and use principal components to build classification models. Moreover, we will use cross-validation to assess the performance of our models and evaluate the models using a confusion matrix and ROC curve.

Visualizations for Exploratory Data Analysis

Fig-1: Player height vs. heading accuracy

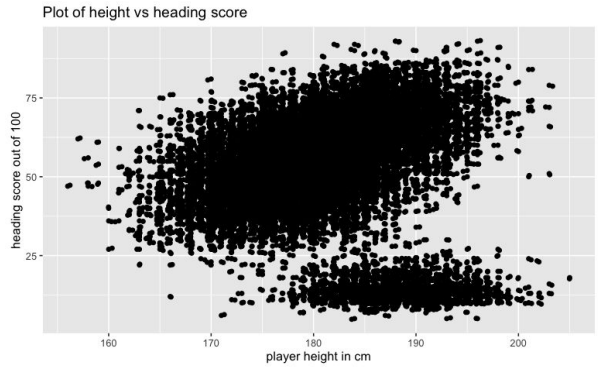
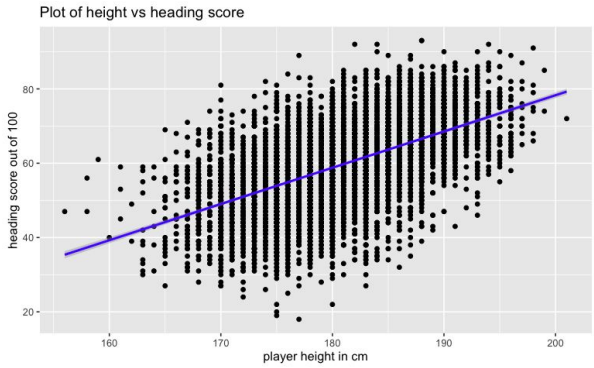


Fig-2: Player height vs. heading accuracy (without reservers, subs and goalkeepers)



* overall is the alias for rating out of 100 in the dataset.

Fig-3: Player rating vs. value_eur

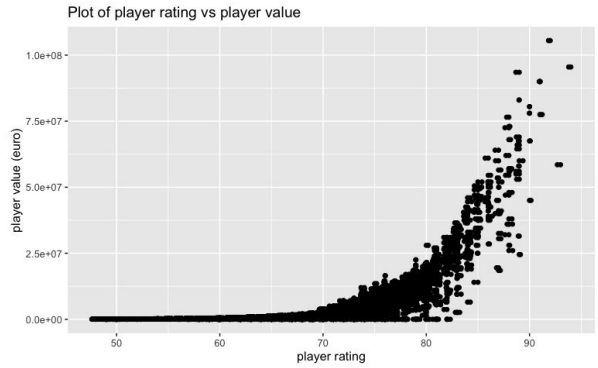
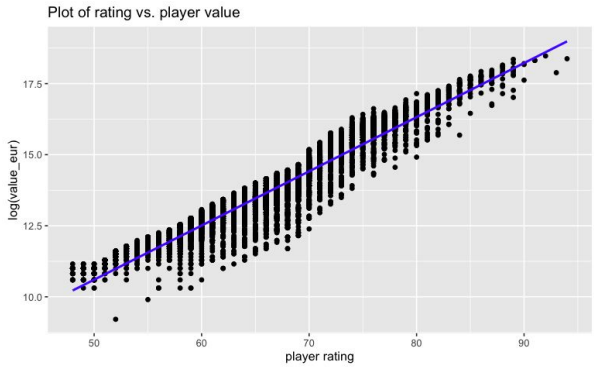


Fig-4: Player rating vs. log(value_eur)



Forward Stepwise Selection - Predicting Player Value

Fig-5: Plots of BIC, RSS, Cp and Adjusted R^2 against the number of predictors, not much change after five predictors.