<u>Supervised Machine Learning - DS5220</u>
<u>Group Project Abstract</u>
<u>03.17.2020</u>

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

Sports analytics is a field that is growing in popularity and application throughout the world. One of the open problems in this field is the assessment of football players based on their skills, physical attributes and market value. The primary aim of this project is to establish football player assessment models using machine learning techniques to support transfer decisions of football clubs. To do so, we will be using publicly available datasets from Kaggle that contain players' data for the Career Mode from FIFA 2015 to FIFA 2020, where each player record is characterised by 104 features. Some of the few important player features include, year, age, body type, work rate, value, skills (e.g. pace, shooting, passing), wage, traits, position, nationality, club, ratings, preferred foot and physical attributes. These features will enable us to analyse the performance of players across seasons and build player assessment models.

Based on domain specific knowledge of the game, we now propose a few low and high risk hypotheses for exploratory analysis. Low risk hypotheses include, (i) tall, short and strong players are statistically good at heading, dribbling and tackling respectively, (ii) player wage and age are positively correlated upto the age of 31 and negatively correlated after that, and (iii) there exists a positive correlation between player rating and value. Some of the high risk hypotheses include, (i) left footed players have a higher overall rating compared to right footed players, and (ii) the starting eleven with the highest overall rating for a given year wins the champions league that year.

Next, we propose a few specific goals to build player assessment models to tackle regression and classification problems. The following goals will be addressed using linear and nonlinear machine learning techniques, (i) to classify nationality based on attributes like shooting, passing, dribbling, defending and pace (e.g. players from Spain are expected to be efficient passers), (ii) to classify player position using physical attributes like height, weight, age, strength, speed and jump (e.g. defenders are expected to be relatively taller and stronger than players in other positions), (iii) to classify work rate using defense and attack traits (e.g. strikers and wingers have high work rates for attack), and (iv) to predict player rating and value based on player attributes.

The goals above will be implemented using advanced supervised learning techniques like ridge and lasso regression, logistic regression, KNN, LDA, QDA, decision trees, random forests and additive models depending on the type of problem. The goals of the assessment models require us to make assumptions about the importance of features. Nevertheless, important predictors will be carefully chosen using feature selection techniques for each proposed question. Cross-validation will be used to determine the model with the least error. Finally, based on the problem we will evaluate the performance of the respective models using metrics like Cp, BIC, Adjusted R^2, confusion matrix, ROC curve.