Capstone Milestone Report: Identifying Top Players and Significant Player Attributes Based on Player Position for European Football

Springboard Intermediate Data Science: Python

Capstone Milestone Report

**INTRODUCTION**

The client, a European Football Club wants a model to predict the aggregated overall rating of a player at each age of the player. They want to identify the current top players for each playing position and the significant attributes that affect the player rating for each position. They will use this list of best players to consider during transfer season. In case they need to make a transfer, they will bring in a replacement, a major part of player selection depends upon which position the player will play at. Depending on the need of the team, the managers will consider the up and coming players who will suit their need. Our goal is to fit a model that predicts the overall rating using only the player attributes available to us in the data source.

**Data Source:** The data we will use for this problem is sourced from several websites such as

http://football-data.mx-api.enetscores.com/: scores, lineup, team formation and events

http://www.football-data.co.uk/: betting odds

http://sofifa.com/: players and teams attributes from EA Sports FIFA games, FIFA series and all FIFA assets property of EA Sports;

It is curated by Hugomathien and made available on Kaggle

Link: https://www.kaggle.com/hugomathien/soccer

We want to identify top players playing at each of the four positions - Forward, Defender, Midfielder and Goalkeeper from the dataset of players playing for different clubs in European Football since 2008 up to 2016. From the player attributes data we get a set of attributes such as overall\_rating (also referred to as player rating), finishing, balance, diving etc. for the players. From the players table, we get the players age, weight and height. We know that \*there is a relationship between the overall rating of a player and the player attributes, the player age, height and weight\* so we use these attributes to build a model to predict the overall rating of a player. Using this model, given the attributes for a player we can project overall rating for the player at a future age

**Data Wrangling**: We will prepare the data for modeling, data is available in separate tables such as players (player id, birthday, weight), player\_attributes (finishing, diving, balance etc.) and matches (player coordinates for a match played). We join these tables to get a dataset with all required player details, attributes and favored playing position. For each player we then aggregate data for each age of the player such that we have average overall rating, average finishing, etc. for each age. This is the final dataset for analysis and model building, to fit a regression model to predict the overall rating we analyze the data to test assumptions and identify patterns.

To get predicted ratings for players at the next age of player, we want data for ratings and attributes of a player at a given age. The data that we have has ratings and attributes at multiple matches for multiple ages of a player. So for one age of a player e.g. 25, there could be more than one match details. To get one data point for each age of a player we aggregate the features at that age using the average function - mean:

Description of the aggregated data:

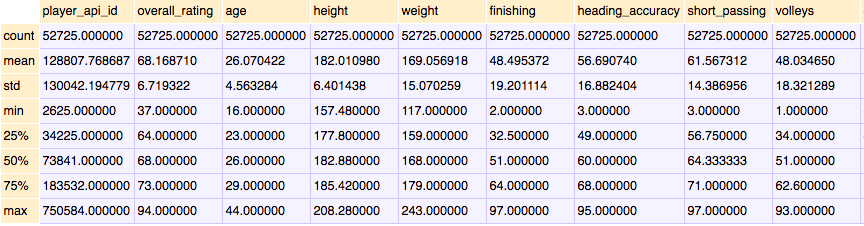


Fig. 1

There are a total of 10582 unique players in this data. These are the fields present in the

Data.

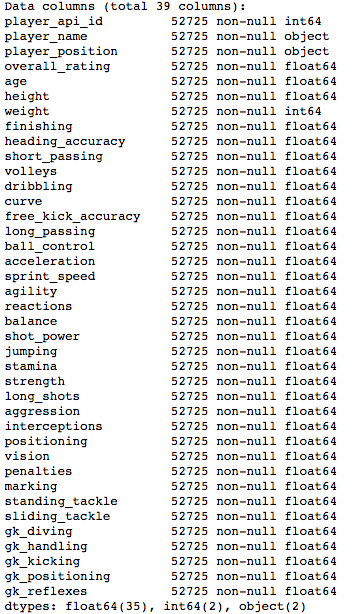


Fig 2

**Limitations:**

There are certain limitations in the data that will affect the robustness of the models that we intend to fit.

1. Not all historical data is available for each player. Knowing that the FIFA Player Ranking is based on historical data of player, we do not have this data for each player. A lot of information will not be learned by the model due to the absence of this data.

2. Data related to ranks of leagues, teams and countries is also not available in this data. This information would have made the model we intend to fit more accurate as these factors play an important role in determining player ranking.

3. There is not enough historical data available for each player, if it were, a time series model could be built to predict the future ratings of players

4. There is country and league data missing for few players and more than 20% players from our player data had not been active in the last two years. These players will be excluded from featuring in the top players lists.

The FIFA 2017 data is now available, we can add that data to enhance and update the player ranking based on current data. With availability of more future data we can identify patterns in attributes of top players to then identify similar patterns in new players. This will help in scoping out up and coming players, this will help the client to make decisions related to player transfers as they could look at signing younger players with higher potentials at a lower cost.

**Data Analysis:**

Player Age Distribution-

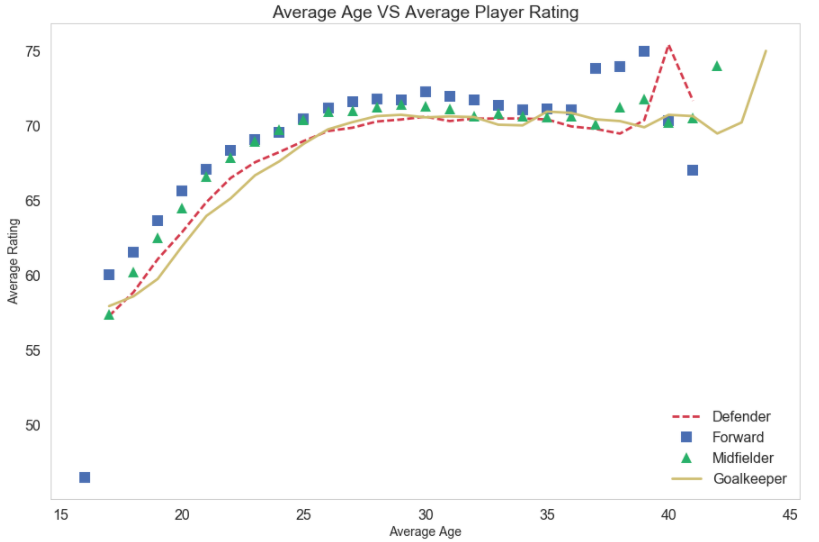
****

Fig 3

Since the relationship of the player age and player rating is of interest to us, as there is a general assumption in football that player rating decreases with age of player, let us explore if this assumption is true for each of the playing positions.

From the above plot, we can see that the player rating increases with age and for Defender, Midfielder and Forwards around the age range of 30-35 the rating starts to drop. While for goalkeepers, the trend is different with ratings remaining high from ages 25 onwards. So the assumption that player rating falls as age increases is true only after age range of 30-33 years.

Difference in Ratings Between Positions –

To predict the player rating we can separate the data into train and test datasets and fit a regression model to the data to find a model with a good fit. However, we know that the data has 36 attributes for each player and according to our assumptions about the game, we think that there is a difference between ratings based on which position the player plays at. If this is true, it would make sense for us to seperate the data by positions and fit seperate models on each position data. For this purpose we test the following hypothesis:

H0: There is no significant difference between the mean ratings of the different position groups

VS

H1: There is significant difference between the mean ratings of the different position groups

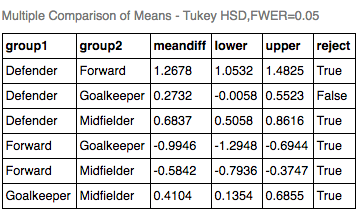
We performed a f-one way anova test to test this hypothesis and the resultant p-value was <0.05 which indicated that we reject the null hypothesis in favor of the alternate.  


Fig 4

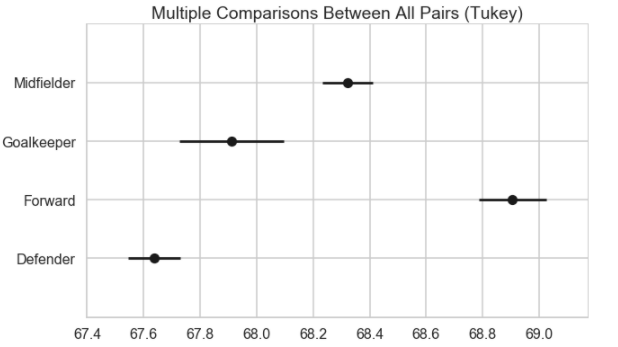


Fig 5

The output of the Tukey test shows the average difference, a confidence interval as well as whether you should reject the null hypothesis for each pair of groups at the given significance level. In this case, the test suggests we reject the null hypothesis for 5 pairs, with an exception of the groups Defender-Goalkeeper. The 95% confidence interval plot reinforces the results visually: only Defender and Goalkeeper groups' confidence intervals overlap.

We have enough evidence to conclude that there is difference in ratings between the positional groups and we go forward with our suggestion to fit separate models for each position.

**Feature Selection and Dimensionality Reduction**: The dataset is very large with many features (36). Feature selection is basis player position where if there is no significant relationship between player rating and feature we discard the feature. Principal Component Analysis will be performed on the scaled feature data to reduce dimensionality and the new components will be selected basis scree plot

**Model Selection**: We split the data into training and test and then find the best fitting model for this transformed (train) data to predict the overall ratings for the next ages of the players (test data). We use the following models for prediction: Linear Regression, Polynomial Regression, Lasso and Ridge Regression and SVR.