# Predicting Team Profiles at the Premier League

Pedro Rodrigues and Sean Pierce

5/9/2022

## Abstract

The purpose of this paper is to help gamblers bet on the soccer transfer market, by predicting the profile of different soccer teams in the English Premier League and test if a model can identify teams by the stats. To do so, we used data from all the players from the 2018-2019 Premier League season and we tested three main models to see which could predict the profile with highest precision, a model with analyst rankings, one with nationality and one without removing any player. In the end, we found a model that was able to predict the player's teams 95% of the time correctly, so it would be able to help someone looking to bet on transfers to have a strong insight of what kind of player each team looks for.

## Introduction

Soccer is the most watched and played sport in the world as it current stands, with approximately 3903 professional level clubs divided into 211 countries who are part of the Fédération Internationale de Football Association (FIFA), an organization with more members than the UN (which currently has 193 nations). To give an perspective, the 2016 Olympic Games is expected to have had a viewership of 3.2 billions if we added all events and all days of the competition, which would lead to 16 days of events, while the 2018 World Cup final had approximately 1.12 billion viewers into a single window of 120 minutes (90 minutes game with 15 min of halftime and extra time added in both halves).

Given the size of soccer in the world, we now focus on Europeans, we need to understand the impact on their lives. It's well known that many countries had problems with the "ultras", which were the most aficionado fans of a team, who would go to create their own dedicated fan club. One of the countries that had the most problem with them was England with the famous hooligans, who supported teams of the English League and would eventually engage in brawls with hooligans from an opposite team. The fervor for soccer is so huge in England that one common practice is the betting on soccer. If you go to websites like Draftkings Sportbook or bet365, you can see that there are plenty of categories to bet on, some are: winner team of the match, first, last or anytime goalscorer per match, categories of goals that a player can score, number of cards, corners, fouls, etc.

This betting market circulates a lot of money every year between bookies and the people who bet. One of the categories for betting is the player transfers, where people can bet on whether a player will transfer from a club to another or will remain in the same, given all speculations. This gain possibility from betting is good to the point of players betting on their own transfers, as it was the case of Kieran Trippier, or even betting in their own matches as it was with Joey Barton. Because of that, we decided to create a predictive model that would be able to find a team profile, by that we mean what kind of player fits each team, if a player who scores many goals and give many assists, but gets a lot of cards is a Liverpool or an Arsenal type of player.

## Methodology

We gathered data from footystats.org for the 2018-2019 season of the English Premier League. The dataset includes 572 players and 47 characteristics for each individual. Each player has qualitative information such as name, age, birthday, what league are they on, what is their current club, what position they play as and

what is their nationality. It also has quantitative variables as number of minutes played, appearances, goals scored, assists, penalty scored and missed, clean sheets (which are the number of games where the teams they are part of have not conceded a single goal), goals conceded and number of red and yellow cards the player received. Most of the quantitative variables are divided into three different variables, they are denoted as "_overall" for the overall stats summing games at home and away, "_home" for the stats only when the player is participating in a home game and "_away" which are the stats for playing away from home. They also have the two different average measures, they include a "_per_90“, which is an average of every stat per what is considered a full match and a”min_per" which is the average minutes taken to get a sense of performance consistency. Additionally, the data also has league ranking for each position, so a player receives a value of “-1” for those it doesn't play in and a value between 1 and the max numbers of players that play in the same one to compared them, in good to inform as well, that there are players who play in multiple positions. So we decided to explore if analyst ranking had more or less impact on transfers than their position in general. To predict the team profiles, we decided to use three different methods:
1. A linear regression with the explanatory variables;
2. A linear regression with the explanatory variables and interaction between them;
3. A random forest model with position as a variable; 4. A random forest model with ranking in positions instead of position.
We divided the data set into a training and a test data, so that with every prediction method, we could test for the percentage of correct predictions and decide the best method by choosing the prediction with the highest correct percentage value. After that we would use our decided method to predict the complete data set.

# Results

We first had to decide which variables were more important when predicting a team profile. As we stated prior, we had different variation of the same variable, as for example the total number of goals who was divided into three categories: the sum of goals, the sum of goals scored at home and the sum of goals scored away. So we decided to create three basic random forest model, using the same variables, however we changed the variables with different variation for each of the forests, so the first forest would have the “overall”, the second would have the “home” values and the third would have the “away” values.

After testing the same random tree with different variations, we found that for majority of the variables, the “overall” variation had a higher effect when predicting the current club of a player. So we decided to use the “overall” variation for all the variables throughout our models. Furthermore, one of the variables that we thought was important “nationality” couldn't be properly used in a random forest model, because there were 67 unique values and the prediction only works with up to 53. To fix that and check the importance of this variable, we decided to look at how many players were born in each country and delete countries with less than three players. After doing that, we removed 48 players, that were spread across 32 countries. In the end we were left with 524 players from 35 countries. Once the data was cleaned enough to use the “nationality” variable as one of our controls, we can see in the following figures 1 and 2, that the “nationality” is the second most relevant variable for the prediction we are interested in, behind “clean sheets”.

**Figure 1 − Variable Importance for prediction using overall**

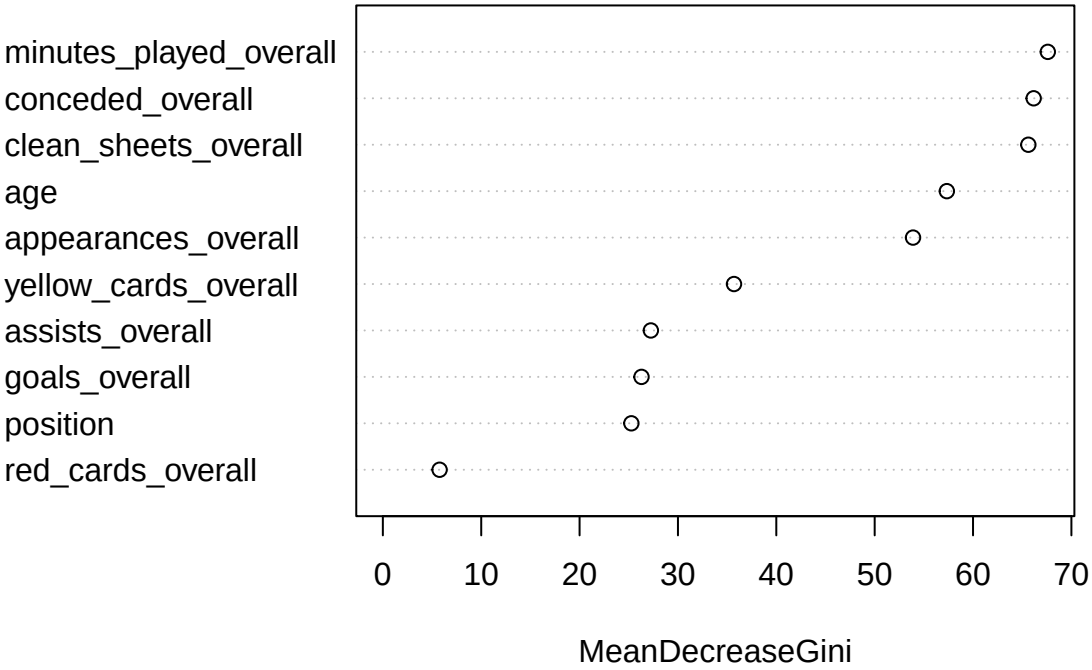
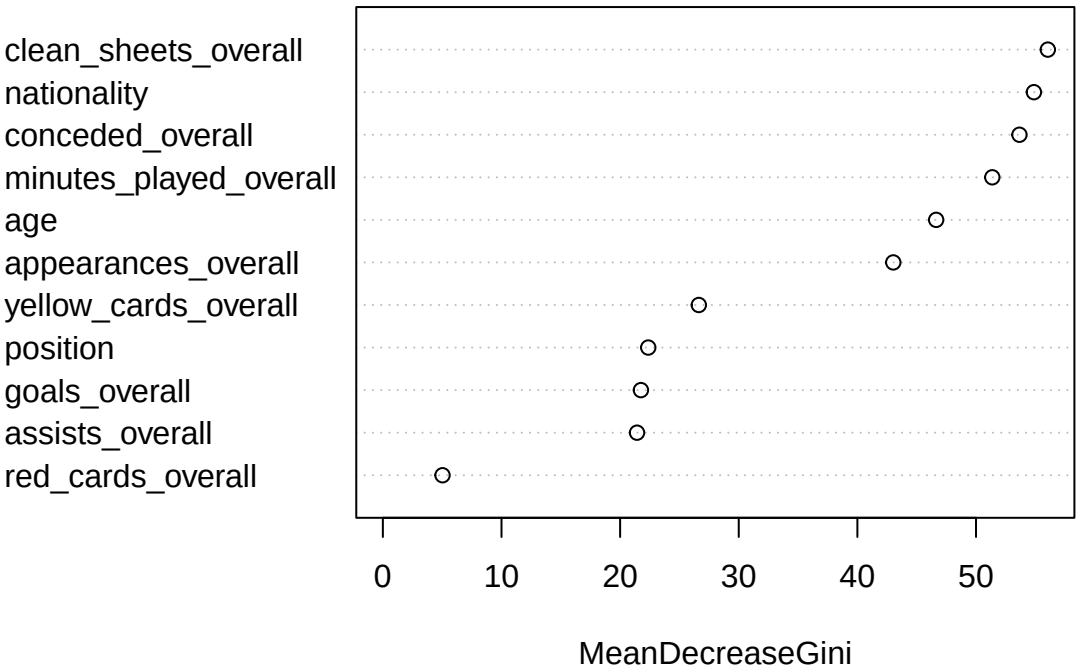



**Figure 2 − Variable Importance with Nationality**




Now that we have decided which variables were more relevant when predicting the "current club", we ran the three different models, the random forest, a medium linear regression and a large linear regression with the important variables. With each prediction model, we saved the predicted results and used them to create an error table, where we have a 21x21 matrix, with the first row and columns being all the possible Premier League clubs and the values on the diagonal are the correct predictions, while the other values are

the wrong predictions. After generating the three matrix of predictions, we calculated the percentage of correct predictions by summing all values on the diagonal and dividing by the number of observations on the testing set. The results are shown in Table 1 below. The best predictive model was the random forest model with the positions, with approximately 19% correct predictions against the 17% correct of the random forest usign the analyst ranking, 4% from the medium model and the 0% from the large model.

Table 1: Correctness of each prediction

| | Percentage of Correct |
|---|---|
| Random Forest | 0.1886792 |
| Random Forest With Rank | 0.1698113 |
| Medium Model | 0.0377358 |
| Large Model | 0.0000000 |

After defining the best model, one of our fears was that we removed important players when removing some countries, so to avoid making this mistake and choosing a wrong model, we decided to run the random forest model, but without the "nationality" variable to check for the percentage of correct predictions it would achieve. After doing the model, we found that it only predicted correctly 93.3% of the players' clubs when accounting for all 572 players, but not using their "nationality".
After that, we once again ran the prediction random forest model, but instead of using the training and the testing sample, we went ahead on using the complete data set of the remaining 524 players, using the "nationality" as a variable When we conducted this prediction, we found that the model was able to predict correctly 95.4% of the players. So the model with less observations, but using the second most important variable was able to predict 2.1% more players. Those values are shown in the Table 2 below, as well as the total value of correct predictions. On the Figure 3, we can see the distribution of predictions made by our model.

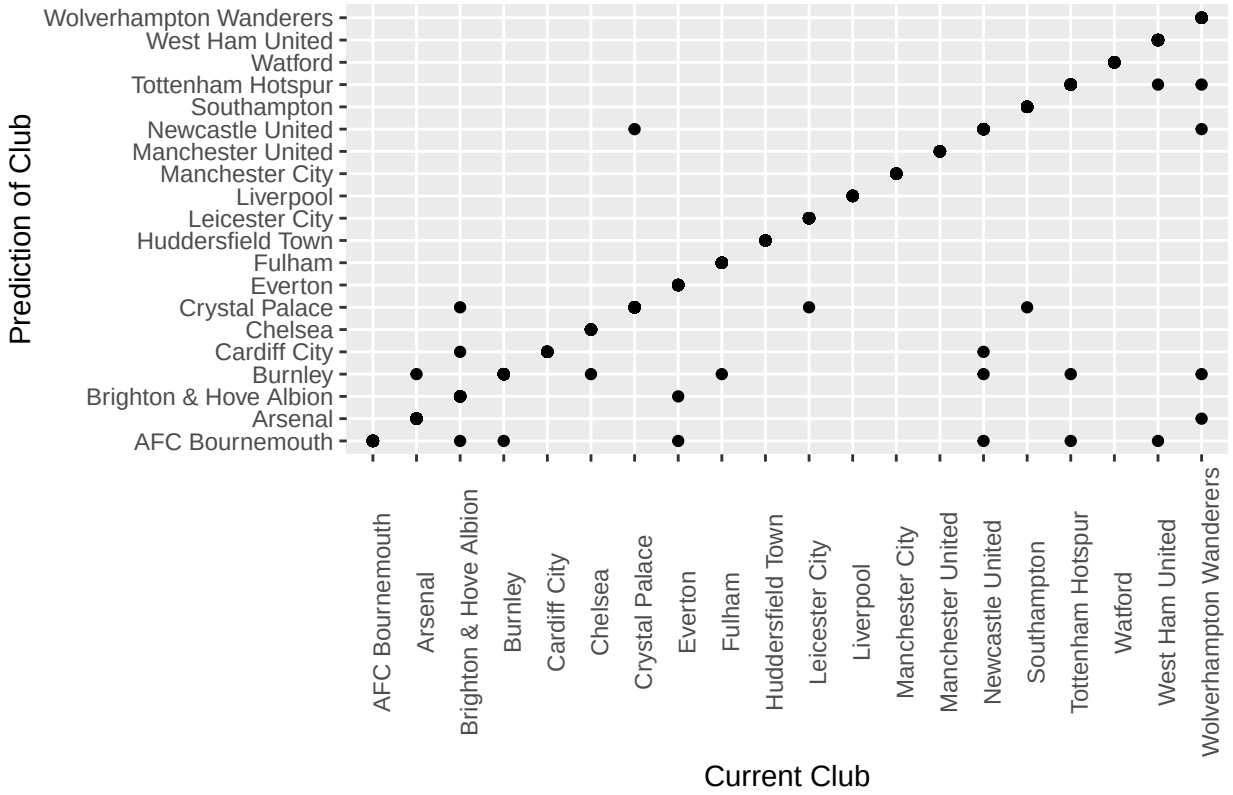
Table 2: Prediction for complete data set

| | Random Forest With Nationality | Random Forest Without Nationality |
|---|---|---|
| Percentage of Correct Predictions | 0.9561069 | 0.9335664 |
| Total Number of Correct Cases | 501.0000000 | 534.0000000 |

# Conclusion

With this research, we found that a random forest model is the best way of predicting a possible "team profile". However if we look at Figure 3, where we see the relation between the predictions and the actual clubs, majority of the miss associations can be explained as teams who tend to play in similar ways and/or have similar characteristics. One example of how clubs are run can be seem in the model classifying at least one player from Brighton & Hove Albion as a Crystal Palace player, because those clubs tend to hire cheaper players, who will have most of the time worse statistics on goals, assists, etc and will usually go for players in their last years of career, to offer small contracts and/or free transfers and will try to get English players. More than that, in the Premier League there is what is known as the "Big 6", that are the biggest 6 teams in the league, composed of: Manchester United, Manchester City, Liverpool, Chelsea, Arsenal and Tottehnam Hotspur. In Figure 3, you can see that those teams have some wrong predictions, but two of them have no wrong prediction, because they have access to considerable more money and better players, so it's hard to wrongly classify them.

Figure 3 − Relation between predictions and actual values




We believe that even though the model was very precise in predicting the players' teams, thus creating a profile for each team and making it easier to identify possible pattern into transfers, it was able to do so due to a clear distinction between teams, as the data comes from a season already over, where the top teams hold the top spots and majority of the players with the highest stats. One way to improve that prediction would be to use data of previous seasons as well, including players that had transfer from or to during previous seasons and adding information on their wages, such that we can easily differentiate the "Big 6" from the rest based on the money availability to spend. Other possible way to improve the model would be to use other league instead of the English and compare the results, however we believe that the English league is the least one-sided and because it is the most competitive, it leads to more similar statistics between blocks of teams. With those similar statistics, it is harder for the model to predict properly, so achieving a high percentage of correct values gives us more confidence to affirm a proper model.

So with this model, a person who decided to bet in transfers for the English Premier League, would have a strong certainty that a selected player that is being rumored will or not fit into the team he is being speculated at, therefore would be able to make a more secure choice when betting.

# Appendix

**Figure A1 − Variable Importance Overall**

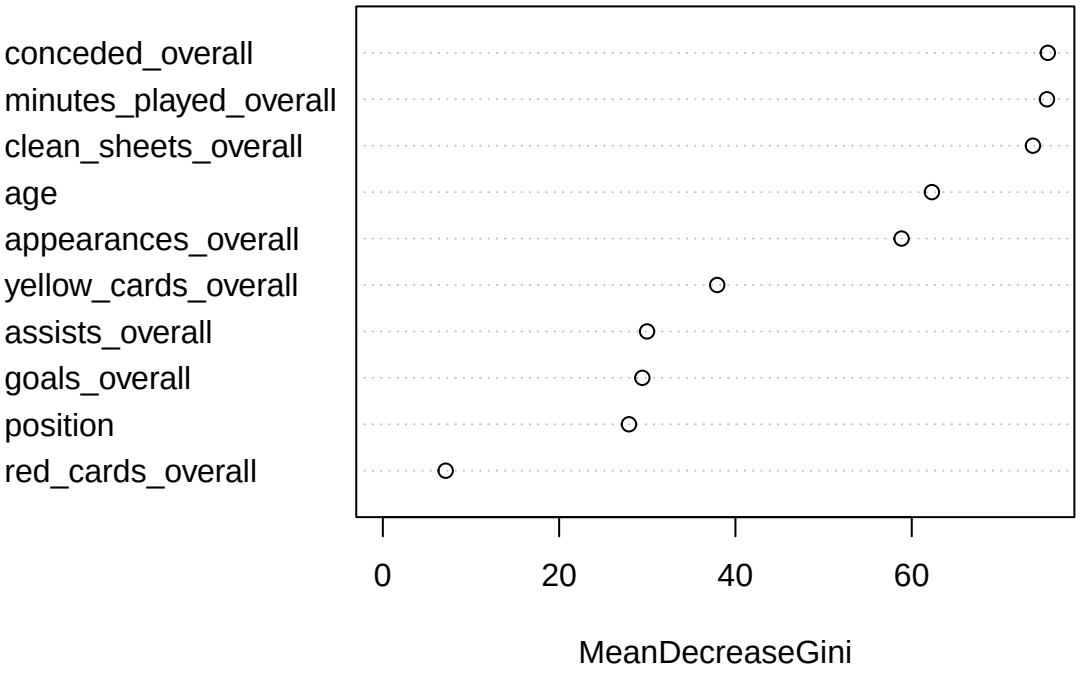



**Figure A2 − Variable Importance Home**

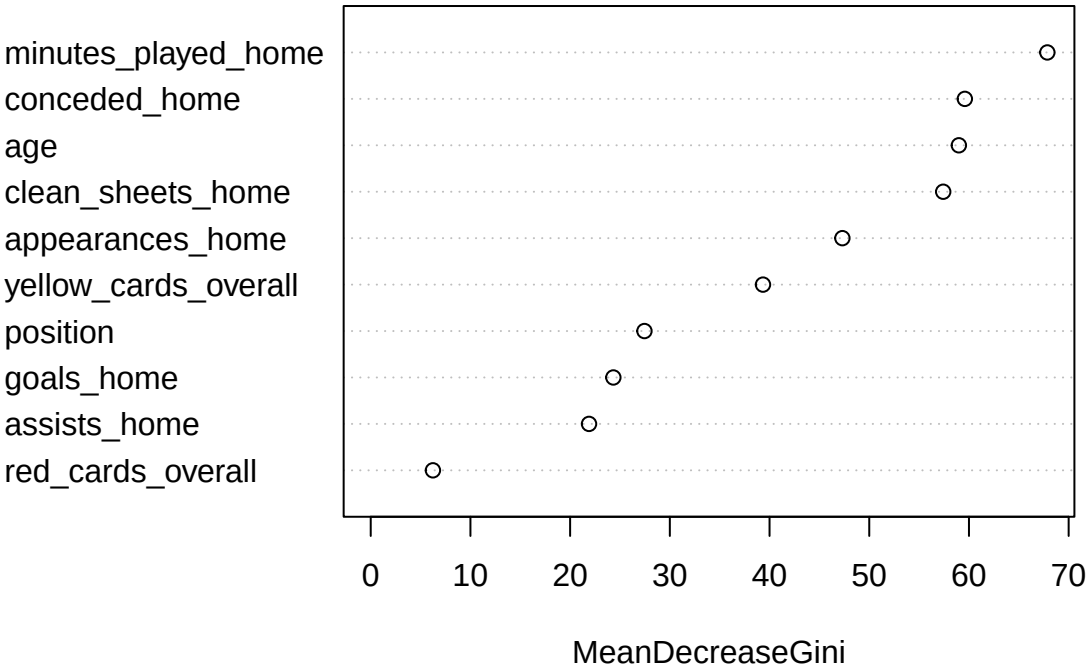


**Figure A3 − Variable Importance Away**

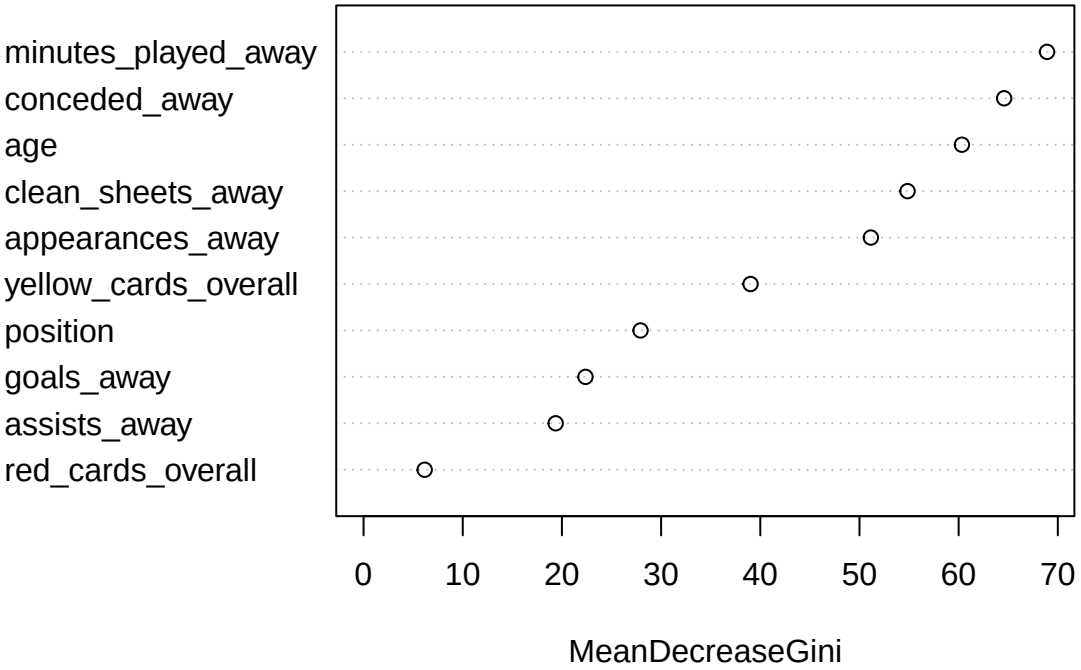



**Figure A4 − Variable Importance Rankings**

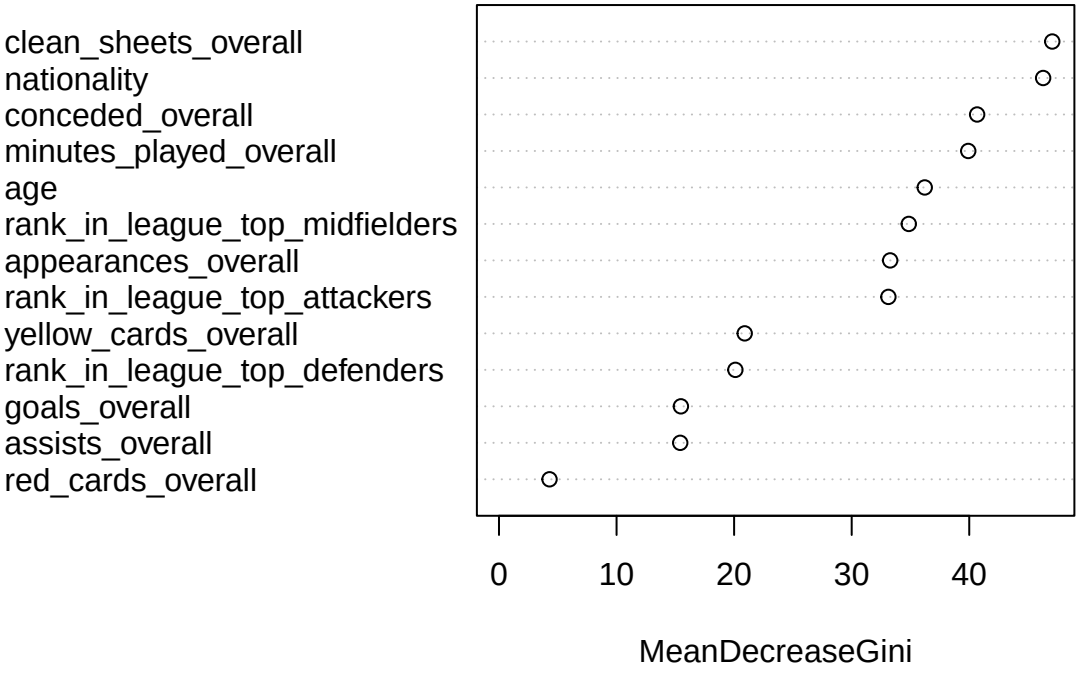


Table 3: Number of players per removed country

| | Number of Players |
|---|---|
| Algeria | 2 |
| Armenia | 1 |
| Australia | 2 |
| Benin | 1 |
| Bosnia_and_Hersegovina | 2 |
| Canada | 1 |
| Chile | 2 |
| Congo_DR | 2 |
| Croatia | 2 |
| Curaçao | 1 |
| Czech_Republic | 2 |
| Ecuador | 1 |
| Equatorial_Guinea | 1 |
| Iran | 1 |
| Israel | 1 |
| Jamaica | 2 |
| Kenya | 1 |
| Mexico | 2 |
| Montenegro | 1 |
| Morocco | 2 |
| New_Zealand | 1 |
| Paraguay | 2 |
| Philippines | 1 |
| Romania | 1 |
| Slovenia | 1 |
| Slovakia | 1 |
| South_Korea | 2 |
| Togo | 1 |
| Turkey | 2 |
| Ukraine | 2 |
| Uruguay | 2 |
| Venezuela | 2 |