FIFA_2021_Player_Ratings_Sedrati

Anass Sedrati

June 20, 2021

Contents

Capstone Project: IDV Learners - Choose your Own - Project Report

1. Introduction

1.1 Aim

The aim of the present project is to apply machine learning techniques beyond standard linear regression in a dataset of our choice, as it is an opportunity to branch out and discover some new data. Given my personal interest for football (called soccer in the USA), and because the European cup is being played at this moment, I chose a dataset related to this sport.

1.2 Method

This project follows the standard data science project methodology suggested by the DataScience Foundation. This methodology contains the following steps: Data collection, examination, cleaning & processing, modelling, prediction, visualization, and continuation.

Given operational constraints in my computer, and the relative big size of data to be treated, I chose from the start to apply three machine learning techniques in this project, as I am confident that they can be applied in my computer. These techniques are: (i) Linear Regression, (ii) Decision Tree, and (iii) Random Forest. The metric used to evaluate these techniques will be the same one used in the first project (MovieLens), and it is the Root Mean Squared Error (abbreviated RMSE), calculated with this formula below:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{u,i}\left(\hat{y}_{u,i} - y_{u,i}\right)^2}$$

I will divide my data into two main sets: Set 1 (70% of the data) and Set 2 (30% of the data). Set 1 will then be divided into two sets: training set (70% of Set 1) and test set (30% of Set 1). Set 2 is our validation set, and will only be used for final hold-out testing to provide the final results for our work.

1.3 Data Collection

I chose a dataset entitled FIFA 21 Complete Player Dataset that I downloaded from Kaggle at this link. The dataset contains various metrics related to football players in the electronic game FIFA 21. Among these metrics, I chose to predict the market value (called Value - The price at which a team buys a player) for a football player. The original dataset contains 107 columns and was disturbing my computer already when loading. I have therefore removed a number of columns that I considered superfluous, especially that their information was available in other columns that I kept. I will explain the details of what I have changed in the section (data cleaning).

The dataset that I have appended to the project and in Github is the one I have cleaned (in Excel), with fewer columns and is easier to handle in my machine. For those wishing to check the full dataset, they can do so by consulting the link I provided earlier.

The dataset has 17.125 rows, each representing a specific football player in FIFA 2021. These players are described by the nine metrics and variables presented in table 1 below (they are the columns of our dataset).

Table 1. Variables within the FIFA dataset

Variable	Class	Example
Name	character	G. Pasquale, Luis García, J. Cole, D. Yorke, Iniesta, D. Odonkor
Age	double	33, 37, 33, 36, 36, 27
Rating	double	69, 71, 71, 68, 81, 66
Nationality	character	Italy, Spain, England, Trinidad & Tobago, Spain, Germany
Club	character	Udinese, KAS Eupen, Coventry City, Sunderland, Vissel Kobe, Alemannia Aachen
Position	character	LWB, CM, CAM, ST, CAM, RW
Potential	double	69, 71, 71, 82, 81, 70
Value	double	625000, 6e+05, 1100000, 0, 5500000, 725000
Total Stats	double	1929, 1906, 1770, 1348, 2014, 1649

2. Analysis

2.1 Data examination

Our dataset has 9 variables (columns) describing player information. The original Kaggle dataset did not provide an explanation of them. Luckily, as a FIFA fan myself, I could recognize them, and provide therefore my explanation here below:

- Name - Name of the football player
- Age - Age of the football player
- Rating - Player rating in FIFA 2021
- Nationality - Country of citizenship of the football player
- Club - Club where the player currently plays
- Position - Position where the player plays (such as goalkeeper or defender)
- Potential - Potential evolution of the player in the future (also interpreted as future rating)
- Value - Market value (price) of the player in €
- Total Stats - A number summing the player skills accross different areas

It is worth to mention that the original dataset contains 107 columns, which cannot all be explained in this report, therefore my decision to cut them before introducing the information in Rstudio.

We can examine our final dataset using the head function in order to have an understanding of the structure and type of information present in it. These can be seen in table 2 below.

Table 2.Head of the FIFA 2021 players dataset

Name	Age	Rating	Nationality	Club	Position	Potential	Value	Total Stats
G. Pasquale	33	69	Italy	Udinese	LWB	69	625000	1929
Luis García	37	71	Spain	KAS Eupen	CM	71	600000	1906
J. Cole	33	71	England	Coventry City	CAM	71	1100000	1770
D. Yorke	36	68	Trinidad & Tobago	Sunderland	ST	82	0	1348
Iniesta	36	81	Spain	Vissel Kobe	CAM	81	5500000	2014
D. Odonkor	27	66	Germany	Alemannia Aachen	RW	70	725000	1649

Table 2 shows that data is rather clean and well prepared. In fact, and as I mentioned, I have prepared it well before putting it in R, as will be explained in the next section.

2.2 Data cleaning & processing

After familiarizing with the dataset, I decided that the variable that I would like to predict for every player will be the market value (in Euro). Therefore, I had to identify the elements that would eventually affect/help predict it.

Being used to play FIFA in Playstation, I know that the total statistics are an important factor to judge a player. They are counted as a combination of several more detailed statistics (such as the ability to defend, dribble, play with the head, etc.). The dataset in Kaggle contains in fact many columns that I have removed (mentality, long shot, stamina, defense, dribbling, and many others), for two reasons. The first reason is that I cannot have all these columns in the dataset (I have tried in the beginning, but the computer was very slow even processing the initial data), and the second reason was that these statistics are part of the column "Total Stats", that is of course in the final dataset. I recognize that removing these variables can affect the quality of the prediction, but given the hardware limitations that I presented, this was the only way to be able to achieve this project.

I have also removed columns that were not relevant for this specific project, such as player picture, club logo, gender (all players are male). Moreover, my predictions do not take into consideration the variables "Nationality" and "Rating", even if they can have an impact on a player value. The reason is to keep the

scope of this project doable in the requested amount of time that I have. The variable "rating" is already included in the total statistics, so removing it can be logical.

To sum up, in this project I will be predicting the value of a FIFA 2021 player (in €) based on the following variables: Total Statistics, Club, Age, Position and Potential.

2.3 Modelling

In order to go into the details of what really affects the value of a player I will implement some hypothesis related to the five variables that I have kept, and test them to see if they do apply. Each hypothesis will be tested towards the 15 most expensive players (in terms of value) to have an idea and see if it makes sense.

Hypothesis 1. Most expensive players have the highest total statistics in FIFA 2021. The verification can be seen in figure 1 below:

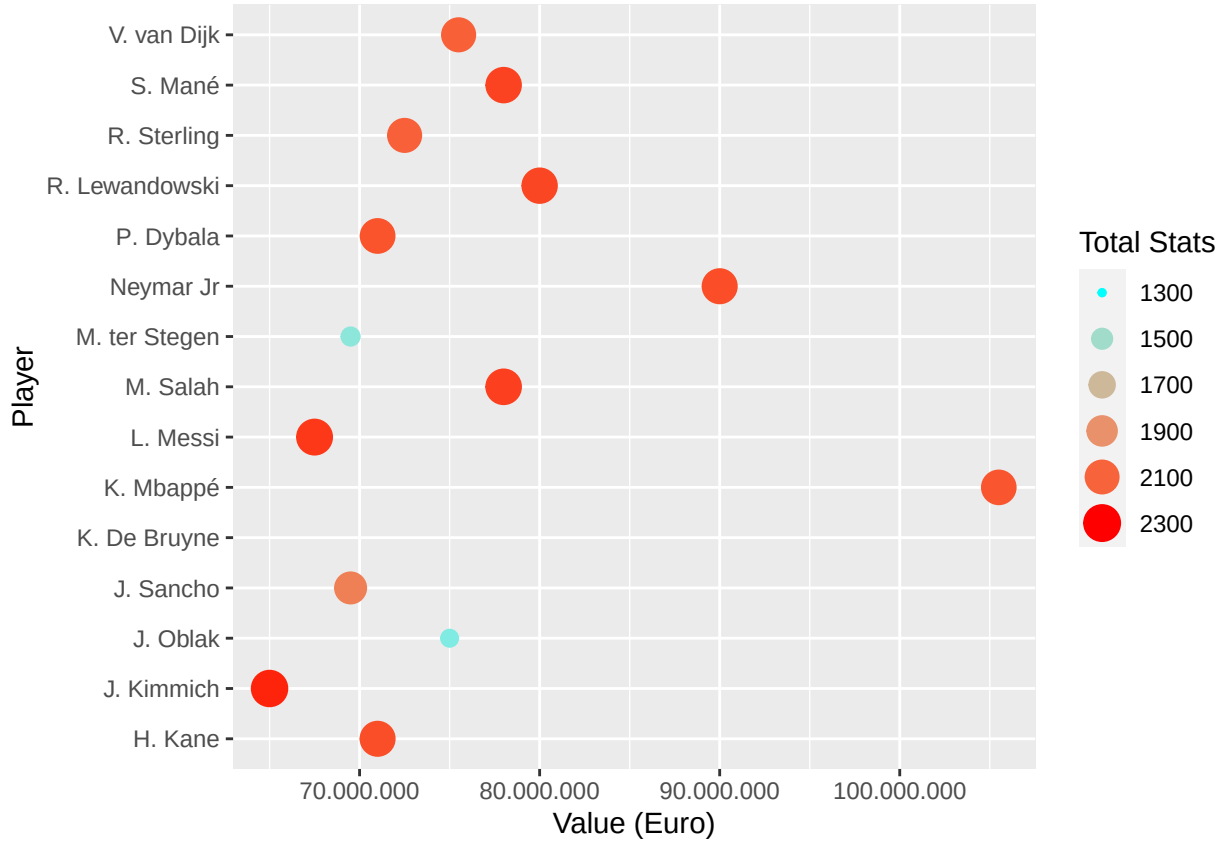

Figure 1. Most expensive players per Total Statistics

Figure 1 shows that the most expensive players are among the highest rated as well. An exception stands however for the goalkeepers (which is normal, given that their statistics are not complete, and are usually low for many skills that are not often used by them). In table 3 below, we can see the list of the 15 players with highest statistics.

Table 3 shows that the players with the highest statistics are not always the most expensive ones. Some players are however present in both table 3 and picture 1 (meaning they are very expensive and have very high statistics). Players with high statistics cover a high range of values (between 20 and 87 million €), which is much higher than the average of the dataset (2.5 million €). Let us now discover more about the data.

Table 3. Players with highest total statistics in the dataset

Name	Age	Club	Position	Potential	Value	Total Stats
L. Suárez	33	Atlético Madrid	ST	87	31500000	2316
K. De Bruyne	29	Manchester City	CAM	91	87000000	2304
Bruno Fernandes	25	Manchester United	CAM	90	63000000	2303
A. Griezmann	29	FC Barcelona	ST	87	50500000	2288
Alex Telles	27	FC Porto	LB	85	31000000	2280
M. Acuña	28	Sevilla FC	LB	83	22000000	2280
Paulinho	31	Guangzhou Evergrande Taobao FC	CM	83	22000000	2279
R. Nainggolan	32	Inter	CDM	83	20000000	2270
J. Kimmich	25	FC Bayern München	CDM	90	65000000	2269
G. Wijnaldum	29	Liverpool	CM	85	37000000	2267
E. Can	26	Borussia Dortmund	CB	84	26500000	2262
Sergio Ramos	34	Real Madrid	CB	89	24500000	2258
L. Goretzka	25	FC Bayern München	CM	88	39500000	2255
L. Modrić	34	Real Madrid	CM	87	24500000	2252
D. Alaba	28	FC Bayern München	CB	84	27500000	2247

Hypothesis 2. Playing at some clubs makes players more valuable.

Below we can see initial information related to this hypothesis:

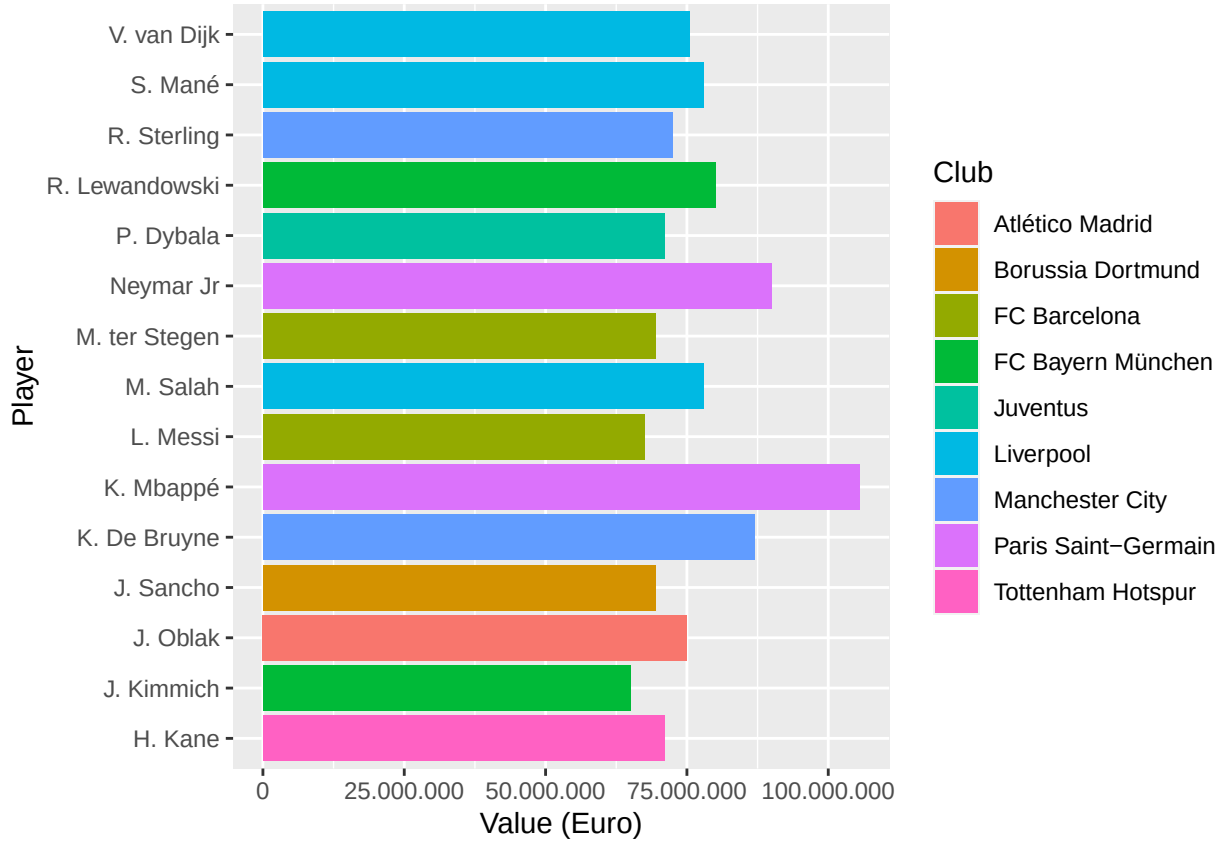

Figure 2. Most expensive players and what club they play for

It is very difficult to draw conclusions from this picture, as no club is specifically distinguished in this list. Moreover, there are 917 clubs present in the dataset, so a deeper investigation is needed to work with clubs. Figure 3 shows the 15 first clubs in terms of most valuable players (by calculating the mean value of their

players).

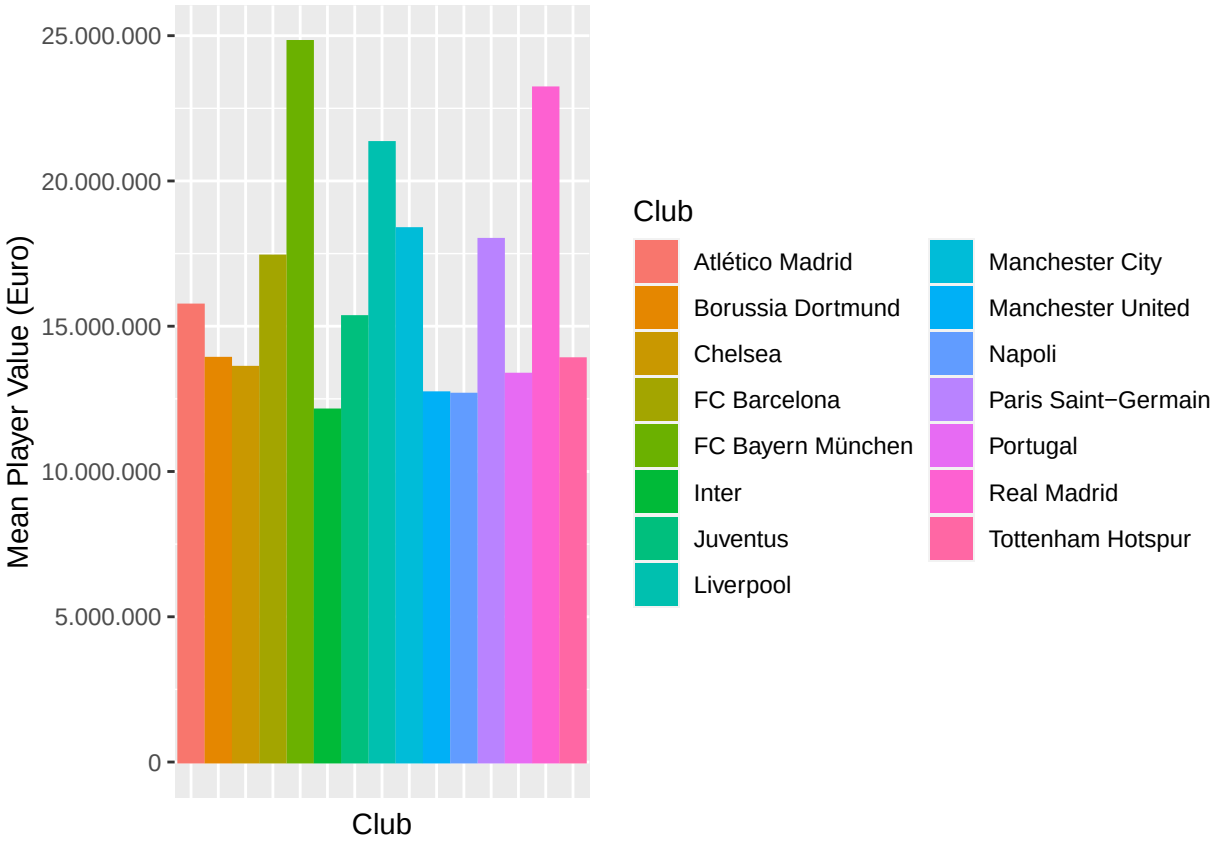

Figure 3. 15 Clubs with most valuable players

From this figure, we can already see that some clubs have a rather higher mean value for their players. If instead of looking at 15 clubs, we look at 100, we can see that the difference becomes bigger as can be observed in figure 4 below (I had to hide the legends as they took a very big part of the plot).

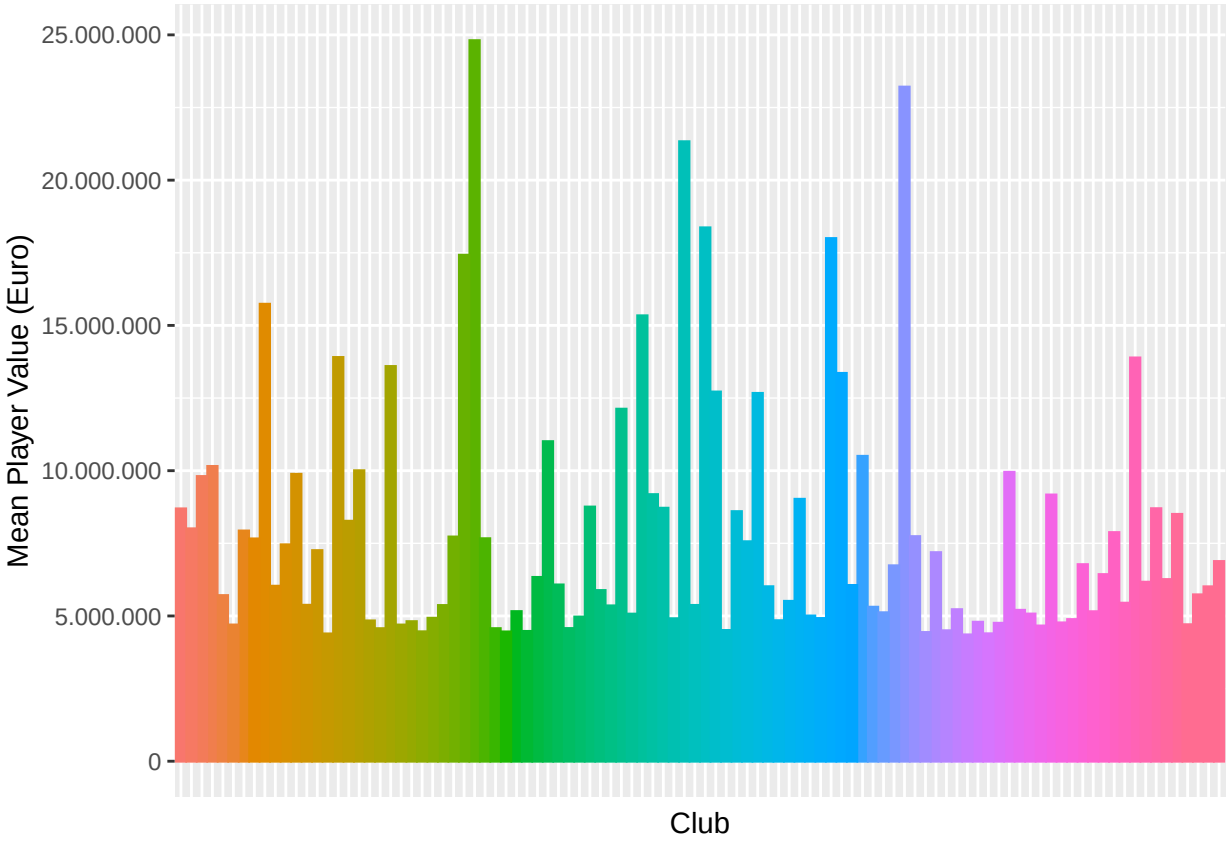

Figure 4. 100 Clubs with most valuable players

From figure 4, we can indeed observe that there are consequent differences between different clubs, and that therefore, it is important to take the club into consideration when predicting the value of a player.

The next variable that we are interested in is the age of a player. Similar to what we did to clubs, we show in figure 5 below the average value of players in this dataset per age.

Figure 5. Players mean value per age

Figure 5 shows that players have their peek value around the age of 28.

In order to remove the outliers and to have more exact predictions, I have removed players aged more than 43, as the ones present in this dataset are (i) few, and (ii) not representative of a regular football player (who usually retires before 40 years old).

Hypothesis 3. Position of football players (in the field) affects their market value.

To explore this hypothesis, we check the 20 more valuable players and their positions. These can be seen in figure 6 below.

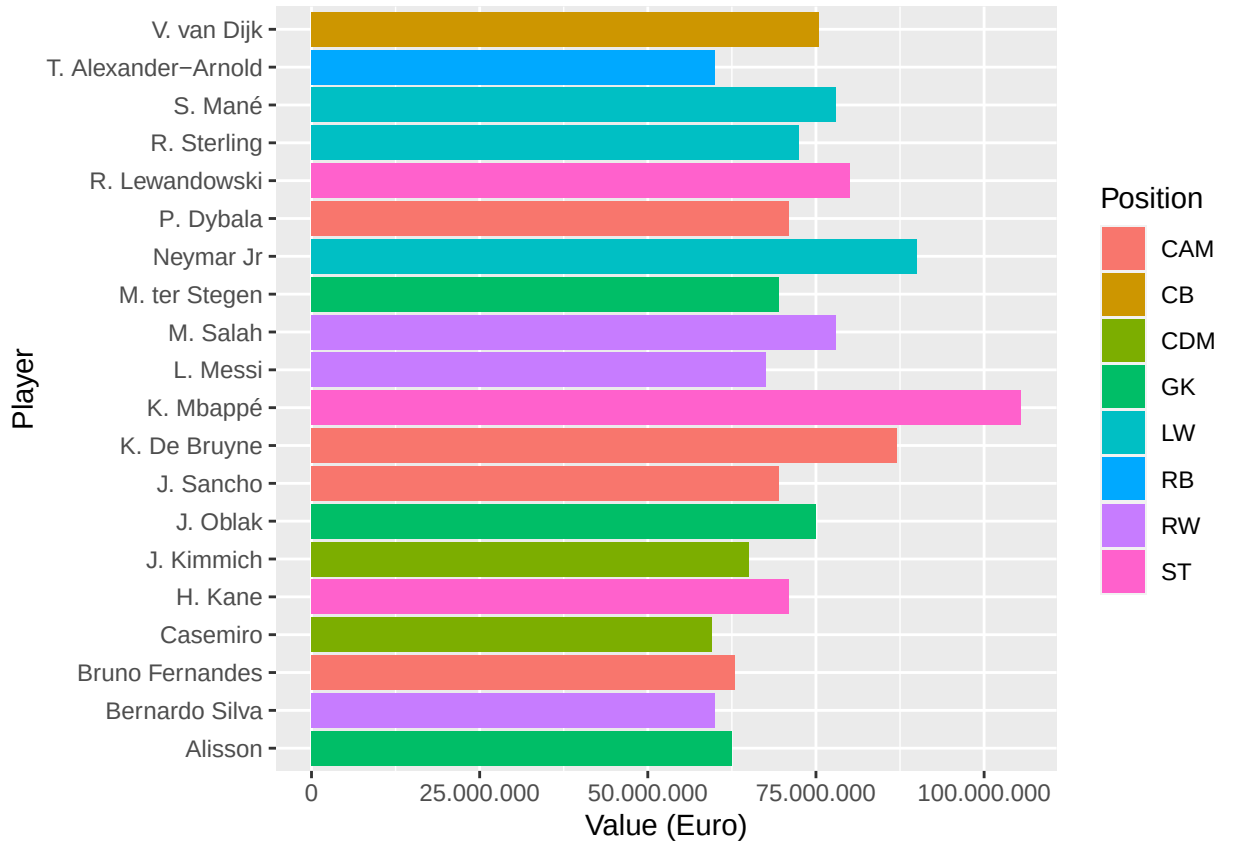

Figure 6. Most expensive players per positions

Out of the 15 available positions, only "half" (7) are present in this list, and most of these expensive players have an offensive profile (9 players are either striker, left wing or right wing). To investigate the player positions further, I looked into the positions per club.

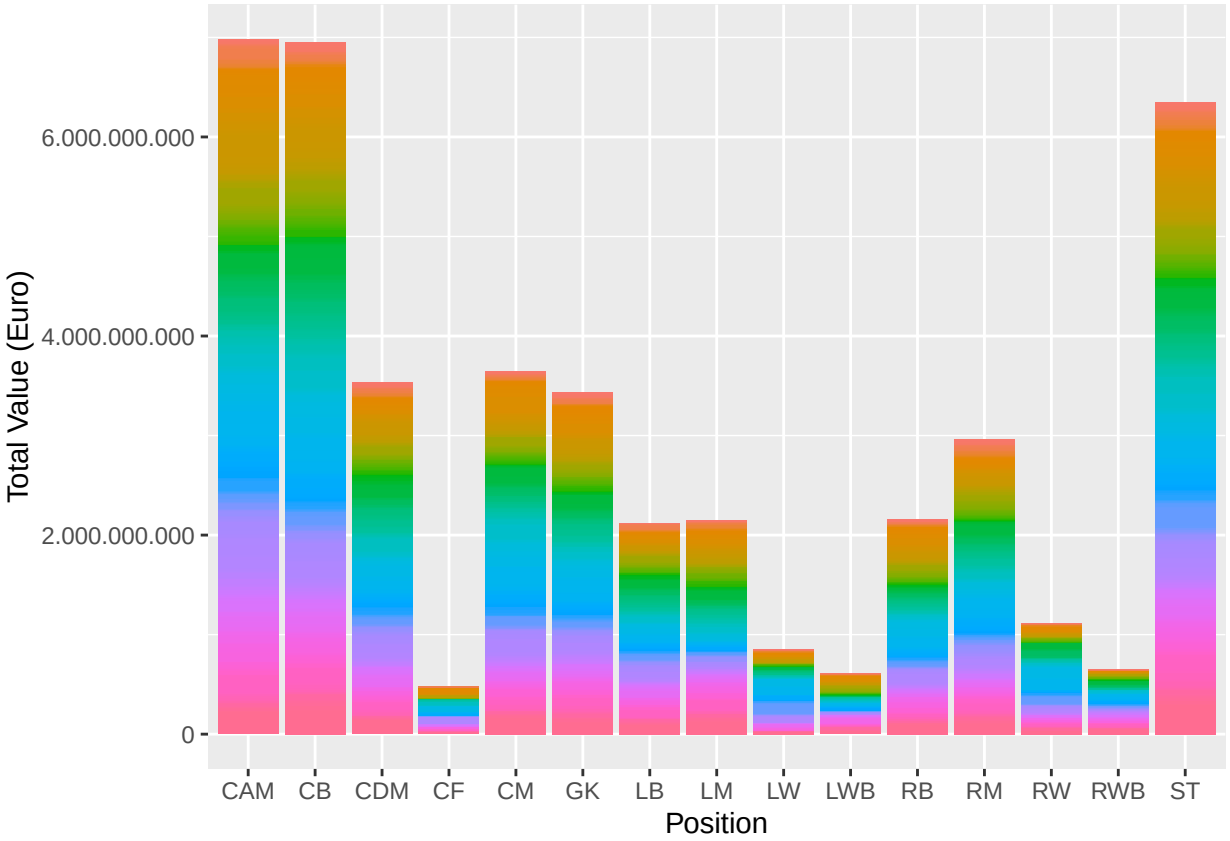

Figure 7. Player value depending on the position per club

I have removed the club names from the legends (as there are 917 clubs), but all positions are present in all clubs at a rather equal percentage. We can clearly see that some positions are more valued than others (even if there is also the possibility that some positions have actually more players than others). The most expensive players are strikers, central defenders, and attacking midfielders. But there is definitely an effect made by the players position on their value.

Hypothesis 4. The potential of a player (a rating number given to a player with prospective of a bright future) affects its value. Figure 8 shows that 15 most expensive players in the dataset and their "potential".

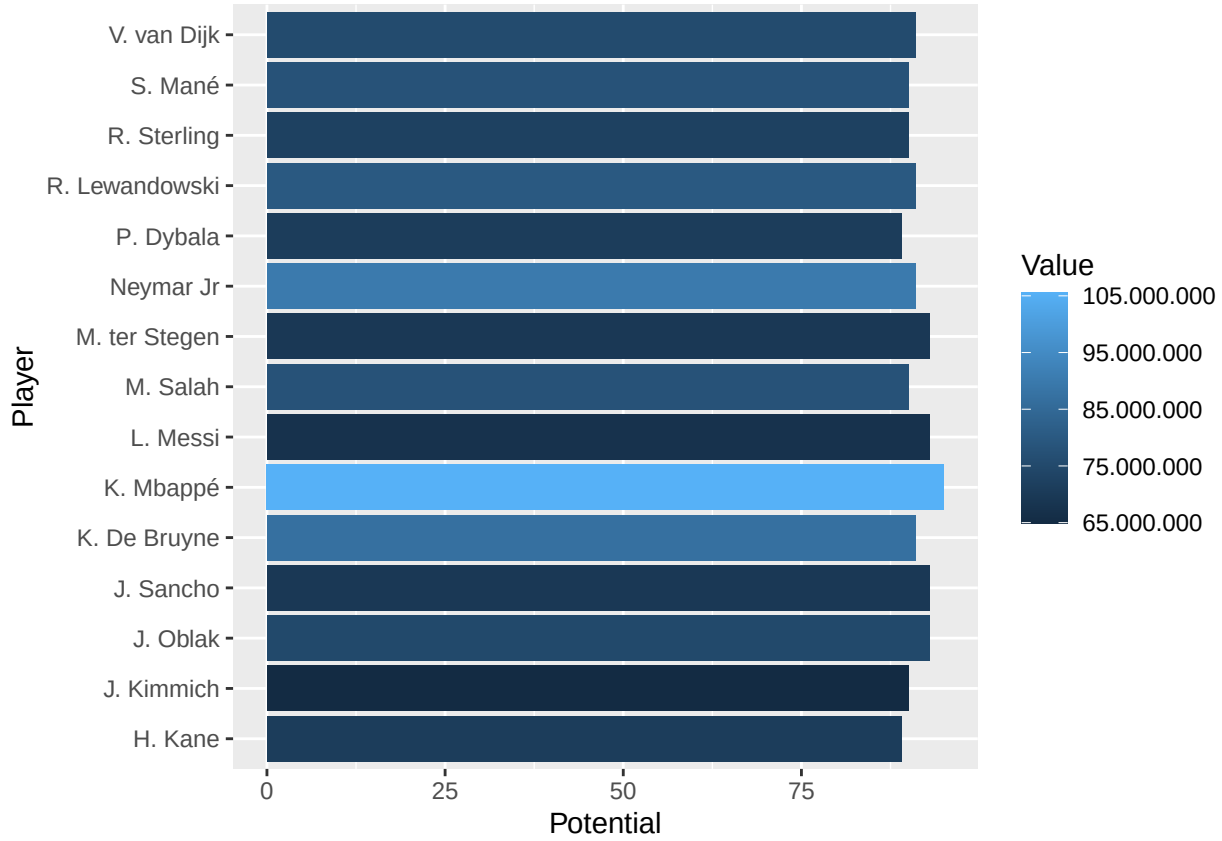

Figure 8. Potential of the most valuable players in FIFA 2021

From figure 8, and given that the mean potential is around 72, we can deduce that the most valuable players have a very high potential (the player with the highest potential in the dataset "95", is in the list). In particular, the lowest potential in the list of most valuable players is 89, which is a high number. It is a strong indication that this hypothesis is likely to be true. To investigate this further, I checked the players value by potential for all the players in the dataset. Since some groups have much fewer players, I did not show the sum of value, but rather the mean value for players having a given potential.

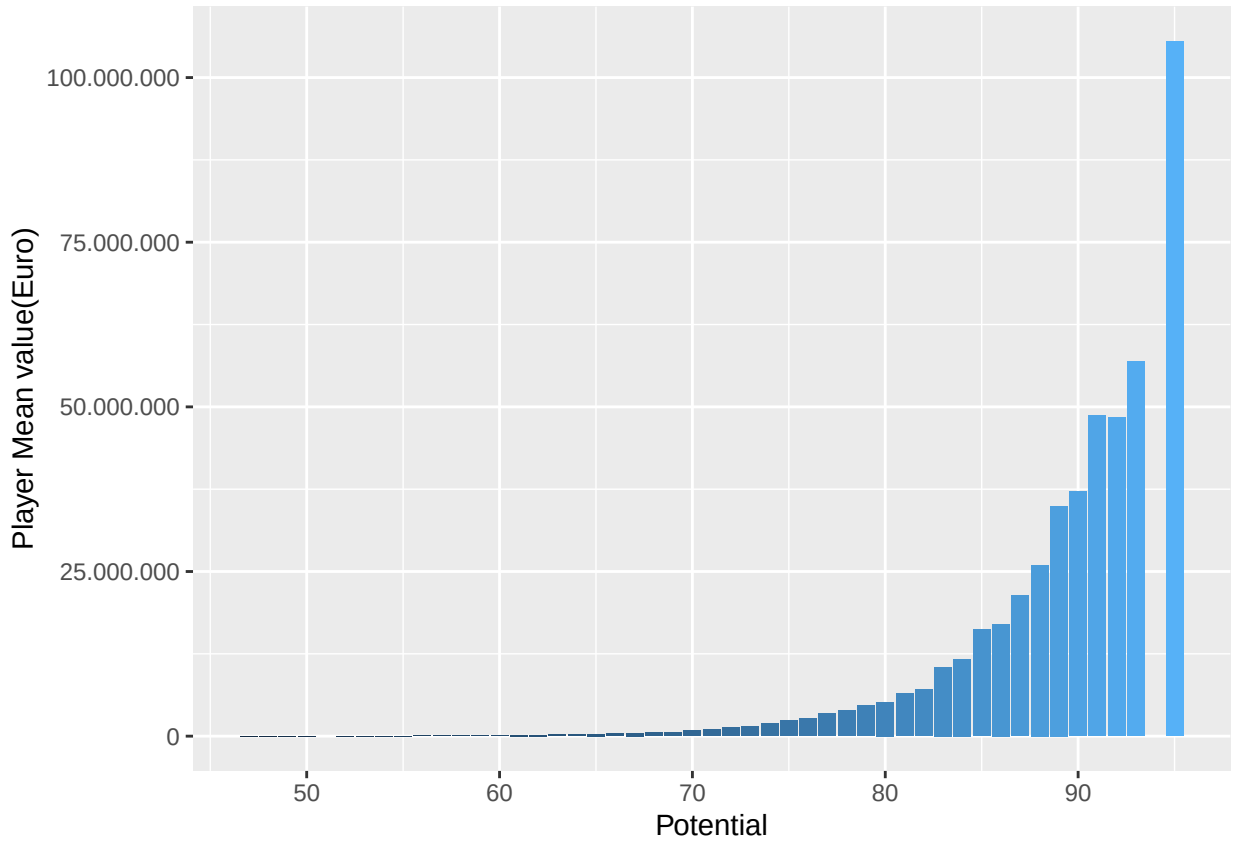

Figure 9. Players mean value per Potential

From figure 9, it is rather obvious that the potential given to a player in FIFA 2021 is directly related with his value, and that the higher the potential is, the more expensive the player is.

To summarize what we have seen until now, all the variables we have studied seem to impact a player's value, with different extends of course. Figure 10 below explains how total statistics, club, position and potential are related to a player's value.

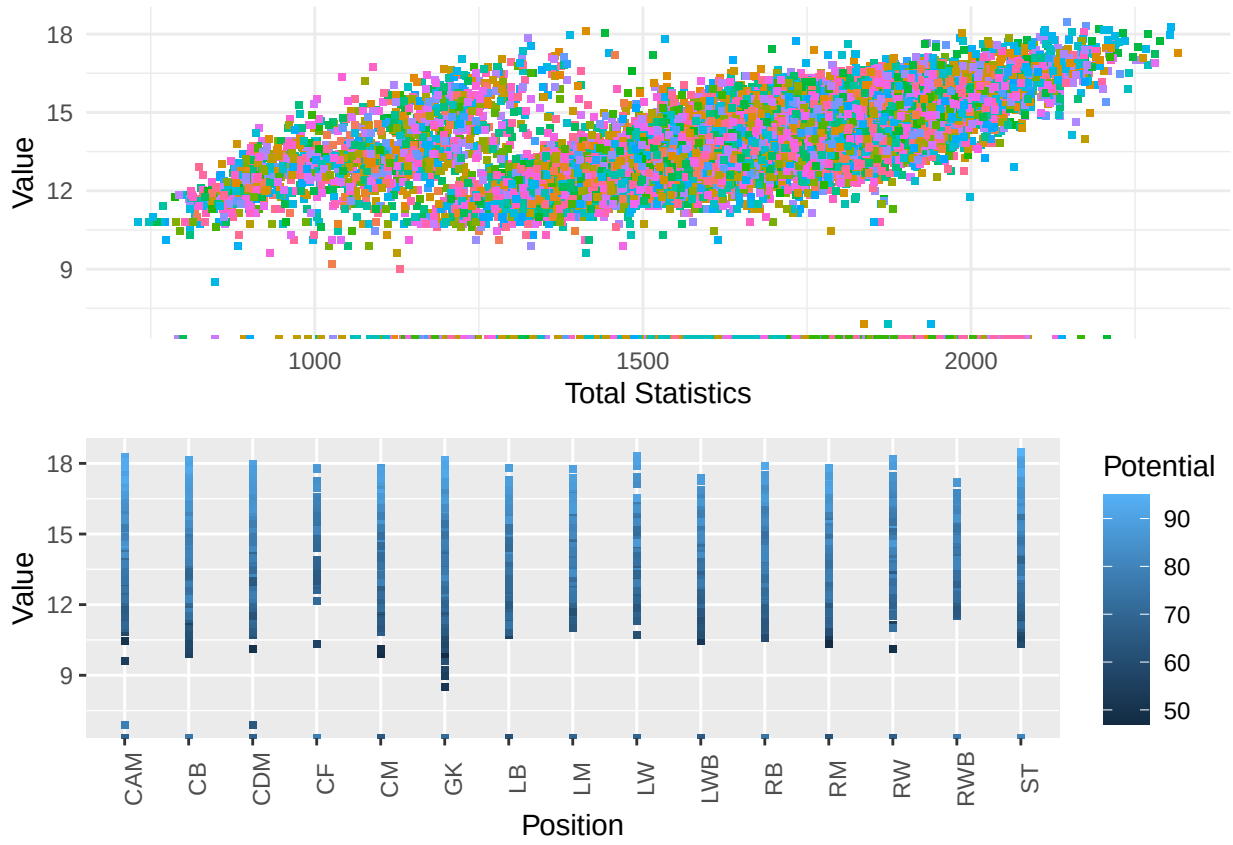

Figure 10. Summary of player related data variables that are going to be used to predict a player's value

It is very complicated to draw any straight forward conclusions from figure 10, especially that there is a wide range of clubs that cannot be very clear in the first graph. In order to study the correlations further, I went on the check how each of the variables we have correlate with the players' value. In particular, I am using total statistics as the main variable, and add other variables to it. For this I used Spearman's rank correlation coefficient, to make sure that the outliers are limited to the value of their ranks. Let us see then in the figures below the different correlations.

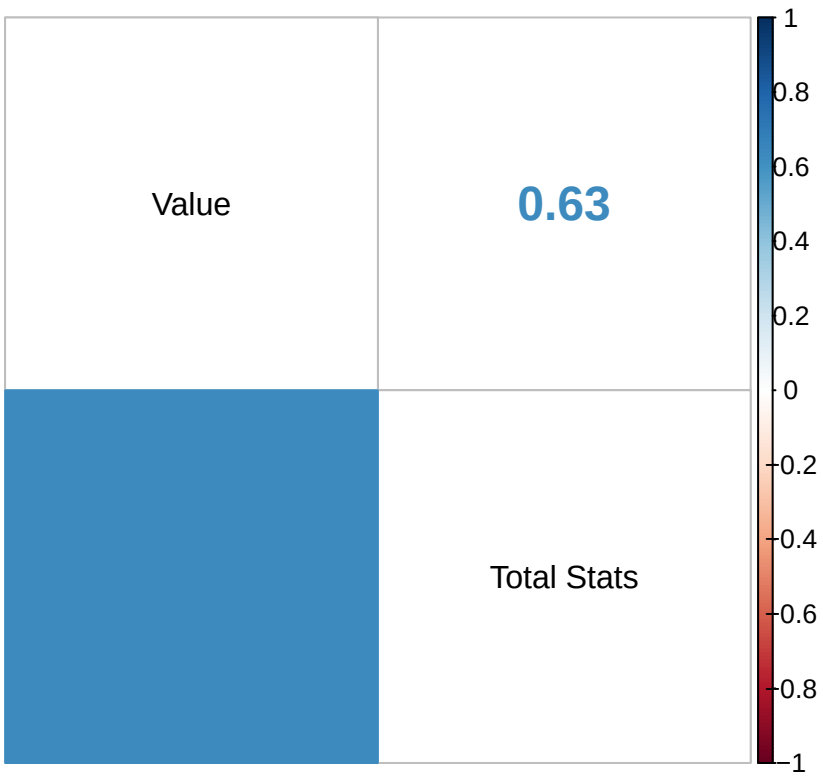

Figure 11. Correlation between players value and total statistics

From this figure, we can observe that there is a good (and positive) correlation between a player's value and his total statistics in FIFA 2021. Let us see how the other variables are doing when added to the statistics.

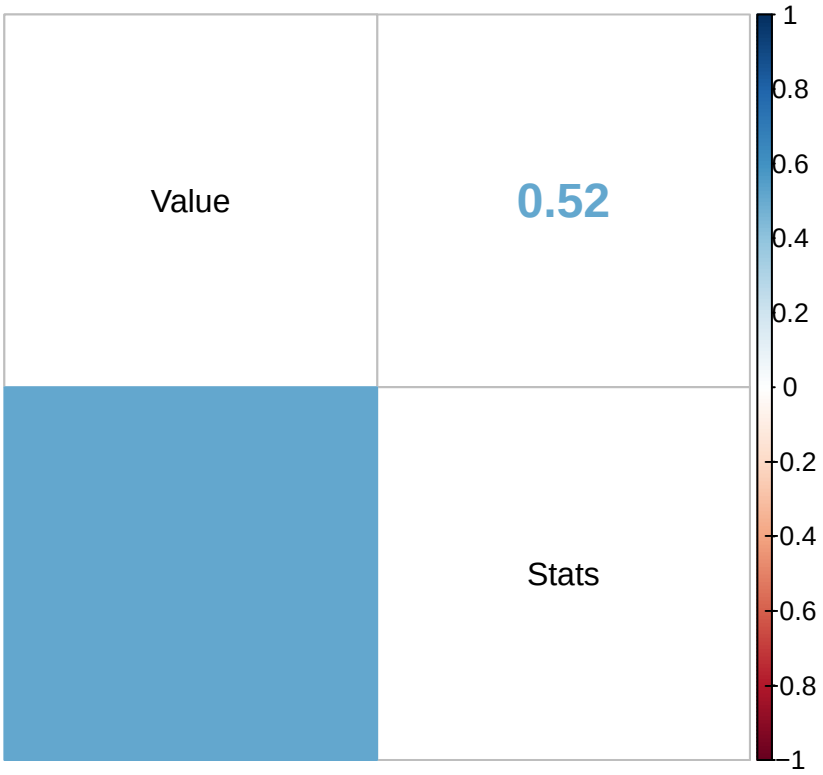

Figure 12.Correlation between value of a specific club and total statistics

From figure 12, we can see the the correlation is still consequent, even that it is less than the previous figure. It is positive, meaning that the bigger the value of a player in a specific club, the higher his statistics will be. The next variable we correlate now to is age as seen in figure 13 below.

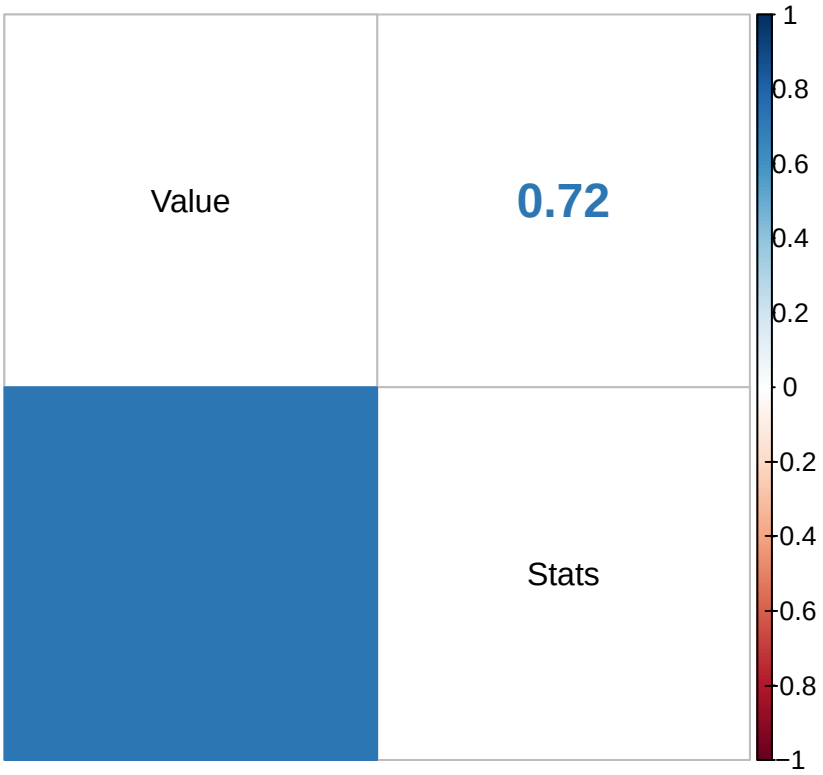

Figure 13.Correlation between value of a specific age and total statistics

Interestingly, we found from the matrix above that the correlation between the value of a player given his age is strongly correlated with his total statistics. The correlation is also positive, meaning that when one is growing, so does the other.

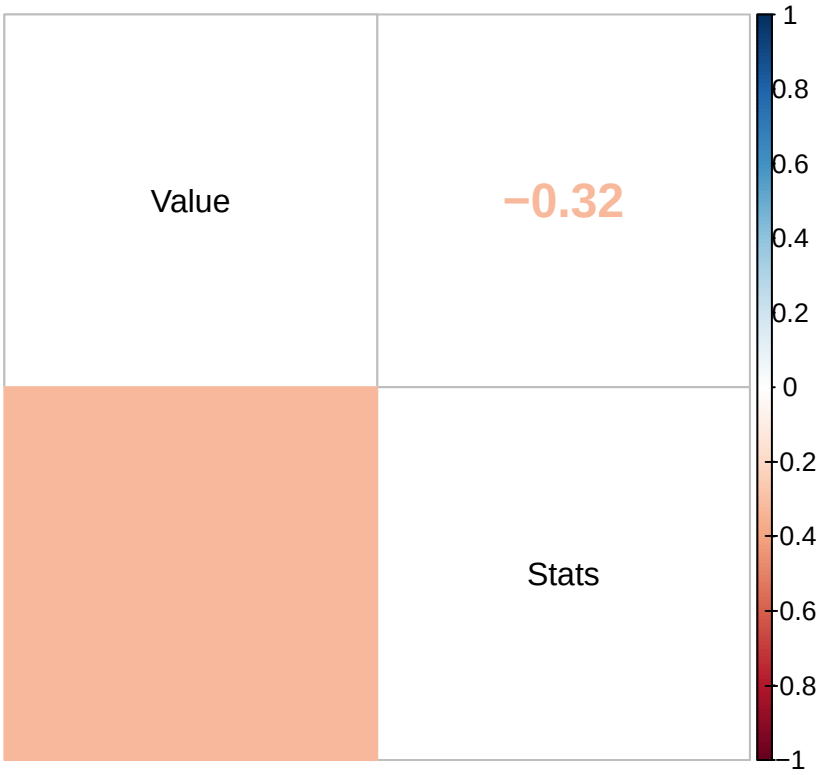

Figure 14. Correlation between value of a specific position and total statistics

From figure 14, we can observe that the correlation between the value of a player given his position in the field is rather weakly correlated with his total statistics. Moreover, the correlation is negative, meaning that when one increases, the other decreases.

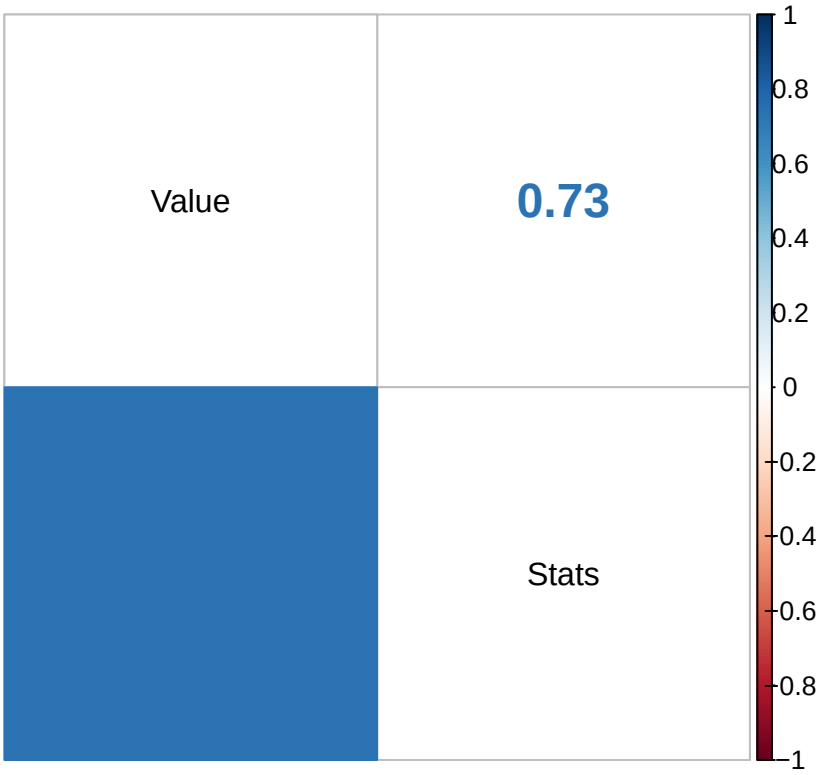

Figure 15. Correlation between value of a specific potential and total statistics

Figure 15 explains to us that when grouped by potential, the player's value is highest correlated with the total statistics, compared to the other variables that we studied.

As a summary of this modeling analysis, we can say indeed that the an important variable that affects the value of a FIFA 2021 football player is his total statistics, but his potential seems to affect the value even more. More importantly, we saw clearly that combining variables with each other bring even bigger correlation, and will for sure bring a better prediction. This is due to the fact the other variables that have a consequent impact as well. In fact, by looking at the figures above, we can see that strong correlations were emerging. From this work and figures above, we can say that the order of importance of the value prediction variables is the following: (1) Potential - (2) Age - (3) Total Statistics - (4) Position - (5) Club.

2.4 Prediction

In this section, I present the three prediction models that I have worked with to train and test my data. I have first split the dataset into test (Set 1 - 70%) and validation (Set 2 - 30%) sets, then split the test set (Set 1) into training (70%) and test set (30%). The validation set will only be used for the final calculations related to the results (part 3 of this work). The three predictions models that I have used are: Linear Regression, Classification and Regression Trees, and Random Forest.

2.4.1 Linear regression Linear regression is a simple supervised machine learning algorithm where the approach is to model the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). It is used to predict values within a continuous range (such as numbers) and is by consequence a good starting technique for us to predict a player´s value.

Linear Regression's main limitation is that it assumes linearity between the dependent variable and the independent variables, while this might not always be the case in concrete scenarios.

In a simple linear regression, the following formula is used for prediction:

$$Y_{u,i} = \mu + \varepsilon_{u,i}$$

In our case, we are using a multivariate linear regression, as we have many variables that we would like to combine. The formula used to predict a players value adding different effects and biases will be:

$$Y_{u,i} = \mu + b_i + b_u + \varepsilon_{u,i}$$

Biases that are tested are the player and potential, that are then combined with age bias, before adding all earlier mentioned to total statistics bias. As observed, we gave the priority to the variables that had the highest impact according to the conclusions we had in our modeling section.

2.4.2 Classification and regression trees (CART) Decision Tree is a supervised learning technique that can be used both for classification and regression problems, through classifying nodes into trees (either classification tree or regression tree). In decision trees, internal nodes represent the features of a dataset, while branches represent the decision rules, and each leaf node represents the outcome.

Algorithms that are used for building decision trees work usually top-down by selecting at each stage the variable that best divides the set of objects. An important challenge with Decision Trees is the identification of the attribute for the root node in each level. There are two popular attribute selection measures: Information Gain and Gini Index.

Information Gain, measures the change in entropy, and is calculated with following formula:

$$\text{entropy}(j) = -\sum_{k=1}^{K} \widehat{p}_{j,k} \log(\widehat{p}_{j,k}), \text{ with } 0 \times \log(0) \text{ defined as } 0$$

and the Gini Index with this one:

$$\text{Gini}(j) = \sum_{k=1}^{K} \widehat{p}_{j,k}(1 - \widehat{p}_{j,k})$$

An inconvenient when using decision tree is that only one variable is tested at a time (when decisions are made), and decision tree can't handle numeric attributes and missing values and last data may be over-fitted or over-classified, if a small sample is tested.

2.4.3 Random Forest Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Random forests undertake dimensional reduction methods, treat missing and outlier values.

An inconvenient of random forests is that their size (consisting of a large number of trees) can make the algorithm very slow, which is a limitation for many people and machines, and is not practical for live predictions.

3. Results

3.1 Visualization

In this project, the aim was to apply machine learning techniques going beyond standard linear regression, and explore a totally new data. Through exploring this FIFA 2021 dataset, and applying decision tree and random forest methods, I believe that I have fulfilled that aim.

I have gathered the results of my findings (RMSE from validation set) from the three machine learning techniques in table 4 below. The table shows how the RMSE (shown in millions) is constantly being improved through the methods (Mean, Linear Regression, Decision Tree, and Random Forest). As noticed, the RMSE value is rather big and I had to check the reason. I found actually four reasons: (1) There is an enormous disparity between the players values, some being very expensive, and some being free, while (2) the dataset is rather small (17000 players in total). The mean player value is around 2.5 Million €, but the variation is big, and many players have values exceeding 80 million €, making the outliers range very big. Also (3), the value of players is in millions, and therefore the variation will also be big, and cannot be shrinked down to tens of Euros, unless having a dataset of billions of players to train with. Finally (4), the total statistics and number of clubs are variable that has a huge range (more than 1000 values for statistics, 917 for the clubs), and therefore requires a much biger dataset than the one we have. This said, I am satisfied with the results I have (there were no calculation mistakes) and believe that the methods were applied in the best way (with of course many improvements that can be implemented in the future).

Table 4. RMSE results

Method	RMSE
Mean	5.035858
Total Stats Effect - Linear regression	4.309631
Potential Effect - Linear regression	3.307320
Decision tree	2.952252
Random Forest	1.645836

The regression tree generated in this project can be observed in figure 16 below.

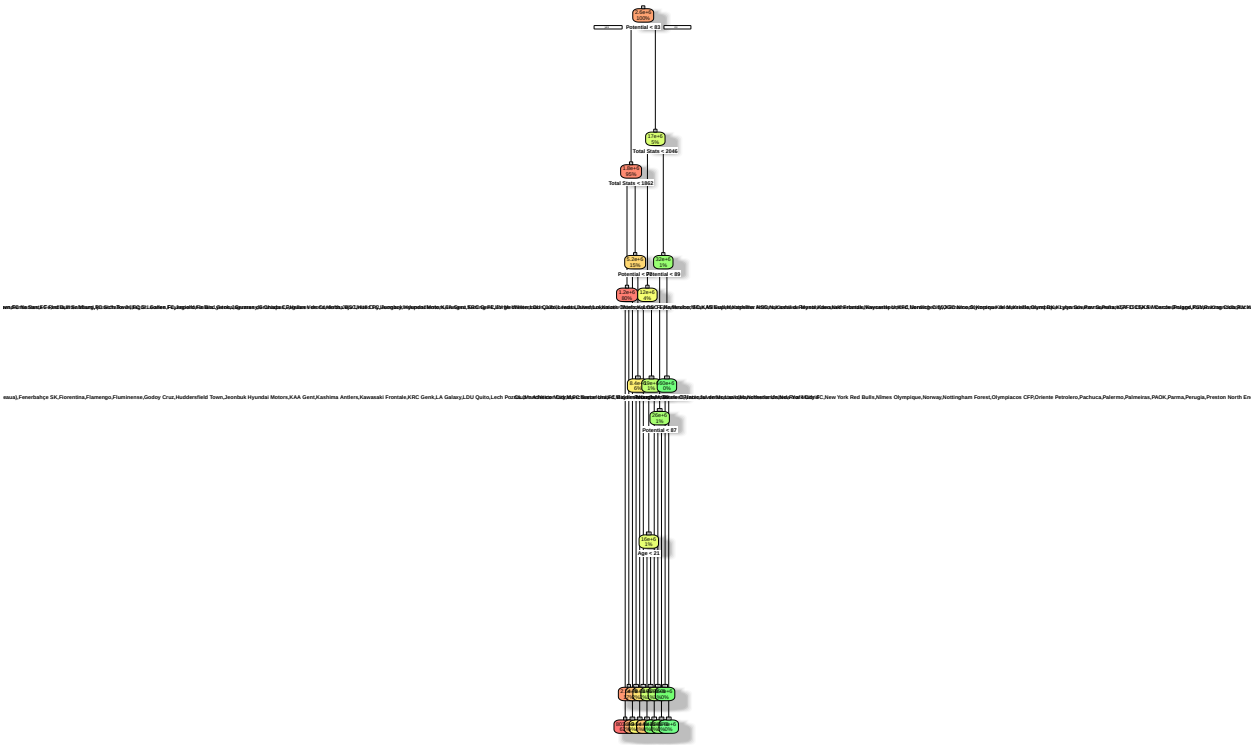

Figure 16. FIFA 2021 players value tree

As noticed, there was a variety of data in the tree, given the complexity of the data that I had, and that some of them were spanning on a large area (such as the number of clubs (917) and the range of total statistics). Moreover, we had a mix of categorical and numerical data, so normalization can be considered in future work.

Figure 17 coming next summarizes the results from our random forest.

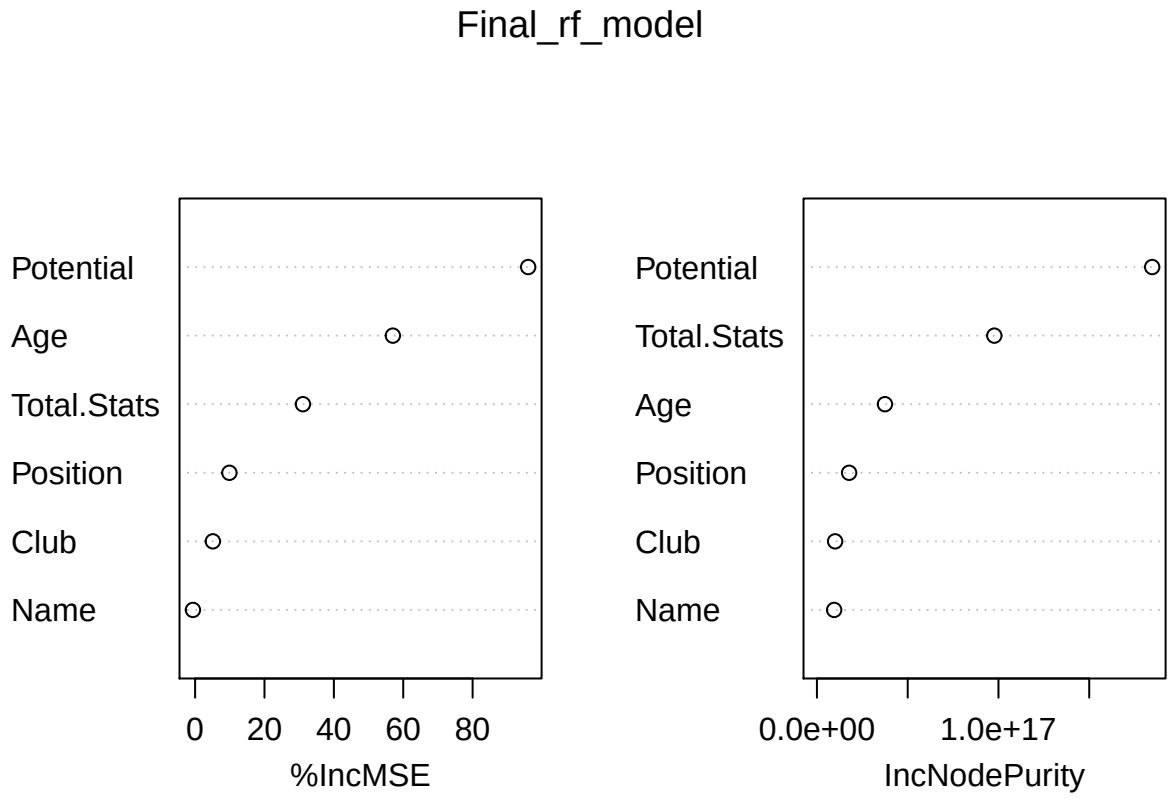

Figure 17. Summary of Random forest results

Increased Mean Square Error (%IncMSE) provides the prediction ability of mean square error with randomly permuted variables, while IncNodePurity calculates the loss function when best splits are selected. From figure 17 we can observe that the variables that are most important in predicting (those having the highest %IncMSE) are potential and age, then total statistics. This goes in line with the work that we did, and with obervations we made earlier, that made us choose these variables in that order for our predictions.

4. Conclusions

4.1 Future Work

In this project, we went through data related to football players in FIFA 2021, by exploring it and trying to predict players value with the help of a number of variables. Although the aim of this project was reached, I believe that there is room for future work, that can improved the results of this work.

First of all, the dataset (of roughly 17.000 rows) was rather small, and the fact that some variables had a huge range (total statistics and clubs) affected the prediction ability and resulted on a high RMSE. In the future, one could concentrate on other variable to test the predictions with, such as rating or nationality, that have a much smaller range (although they cannot be as big in importance compared with the two

others). The fact that some of the data was categorical and not normalized could also have created a risk for over-fitting data.

For the future, it could also be interesting to give more time to the initial dataset (with 107 columns), and investigate the different variables in detail, which would for sure help having a much better accuracy in the predictions. Unfortunately, and as this project has time constraints, this operation was not possible, but this could definitely be reached in future projects.