Football Players Evaluator

Junjie Zhu i Antoni Sanchez Teruel

05/06/2021

Contents

Football Players Evaluator

1 Description of the dataset. Why is it important and what question / problem do you intend to answer?

The dataset is obtained from the statistics of the players of the video game Fifa 2017 (https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global) from the video game developed by the company EA.

`This dataset collects all player information where part of it is fictitious because it quantifies the pl`

The goal of this dataset is to be able to establish models to measure how data relates to reality and how these differences are managed if they exist.

Specifically:

- Given the data, build the probability that the player is part of the national team.
- The consistency of the rating based on objective assessments of the individual characteristics of the player.

---

2 Integration and selection of the data of interest to analyze.

2.1 Reading the file and preparing the data

Let's start by reading the data and selecting the data that are interesting to us for our initial analysis and models.

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('GGally')) install.packages('GGally'); library('GGally')
if (!require('corrplot')) install.packages('corrplot'); library('corrplot')
if (!require('car')) install.packages('car'); library('car')
```

```r
if (!require('caret')) install.packages('caret'); library('caret')

fifa <- read.csv("Fifa.csv", header = TRUE, sep = ",")
head(fifa)
```

```
##                Name Nationality National_Position National_Kit           Club
## 1 Cristiano Ronaldo    Portugal                LS           7    Real Madrid
## 2      Lionel Messi   Argentina                RW          10   FC Barcelona
## 3            Neymar      Brazil                LW          10   FC Barcelona
## 4      Luis SuÃ¡rez     Uruguay                LS           9   FC Barcelona
## 5      Manuel Neuer     Germany                GK           1      FC Bayern
## 6            De Gea       Spain                GK           1 Manchester Utd
##   Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 1            LW        7   07/01/2009            2021     94 185 cm  80 kg
## 2            RW       10   07/01/2004            2018     93 170 cm  72 kg
## 3            LW       11   07/01/2013            2021     92 174 cm  68 kg
## 4            ST        9   07/11/2014            2021     92 182 cm  85 kg
## 5            GK        1   07/01/2011            2021     92 193 cm  92 kg
## 6            GK        1   07/01/2011            2019     90 193 cm  82 kg
##   Preffered_Foot Birth_Date Age Preffered_Position       Work_Rate Weak_foot
## 1          Right 02/05/1985  32              LW/ST      High / Low         4
## 2           Left 06/24/1987  29                 RW Medium / Medium         4
## 3          Right 02/05/1992  25                 LW   High / Medium         5
## 4          Right 01/24/1987  30                 ST   High / Medium         4
## 5          Right 03/27/1986  31                 GK Medium / Medium         4
## 6          Right 11/07/1990  26                 GK Medium / Medium         3
##   Skill_Moves Ball_Control Dribbling Marking Sliding_Tackle Standing_Tackle
## 1           5           93        92      22             23              31
## 2           4           95        97      13             26              28
## 3           5           95        96      21             33              24
## 4           4           91        86      30             38              45
## 5           1           48        30      10             11              10
## 6           1           31        13      13             13              21
##   Aggression Reactions Attacking_Position Interceptions Vision Composure
## 1         63        96                 94            29     85        86
## 2         48        95                 93            22     90        94
## 3         56        88                 90            36     80        80
## 4         78        93                 92            41     84        83
## 5         29        85                 12            30     70        70
## 6         38        88                 12            30     68        60
##   Crossing Short_Pass Long_Pass Acceleration Speed Stamina Strength Balance
## 1       84         83        77           91    92      92       80      63
## 2       77         88        87           92    87      74       59      95
## 3       75         81        75           93    90      79       49      82
## 4       77         83        64           88    77      89       76      60
## 5       15         55        59           58    61      44       83      35
## 6       17         31        32           56    56      25       64      43
##   Agility Jumping Heading Shot_Power Finishing Long_Shots Curve
## 1      90      95      85         92        93         90    81
## 2      90      68      71         85        95         88    89
## 3      96      61      62         78        89         77    79
## 4      86      69      77         87        94         86    86
## 5      52      78      25         25        13         16    14
## 6      57      67      21         31        13         12    21
```

```
##   Freekick_Accuracy Penalties Volleys GK_Positioning GK_Diving GK_Kicking
## 1                76        85      88              14         7         15
## 2                90        74      85              14         6         15
## 3                84        81      83              15         9         15
## 4                84        85      88              33        27         31
## 5                11        47      11              91        89         95
## 6                19        40      13              86        88         87
##   GK_Handling GK_Reflexes
## 1          11          11
## 2          11           8
## 3           9          11
## 4          25          37
## 5          90          89
## 6          85          90
```

The main variables we will use in this activity are:

Variable	Description
Name	Player name
Nationality	Nationality of the player
National_Position	Position in the national team
National_Kit	Team number on the national team
Club	Name of the club
Club_Position	Position of play in the club
Club_Kit	Team number at the club
Club_Joining	Date he started at the club
Contract_Expire	Year end of the contract
Rating	Overall rating of the player, between 0 and 100
Height	Height
Weight	Weight
Preffered_Foot	Little favorite
Age	Age
Preffered_Position	Preferred position
Work_Rate	qualitative assessment in terms of attack-defense
Weak_foot	rating of 1 to 5 control and power of the leg not preferred
Skill_Moves	rating from 1 to 5 of the player's movement ability
The other variables refer to player attributes.	

As explained above, in the first point, the reason that we based on to select these variables from the dataset is because of the information to classify the types of players based on their physique without going into detail in statistics and numerical values of their skills as shown in the game. In this way these results could be extrapolated by possible evaluations of new players and even applied to reality instead of the game.

We see what kind of variables they are and do the first processing if necessary.

Numeric type variables:

Nagional_Kit, Club_Kit, Contract_Expiry (any), Rating, Height, Weight, Age, Weak_foot, Skill_Moves and the rest of the player's stats.

Categorical type variables:

National_Position, Preffered_Foot, Preffered_Position y Work_Rate. It should be noted that the variables Club and Nationality could be considered as a categorical variable, depending on whether players from the same club appear repeatedly.

2.2 Preparació de les dades

```
# Factor of categorical variables
fifa$Nationality <- factor(fifa$Nationality)
fifa$National_Position <- factor(fifa$National_Position)
fifa$Club <- factor(fifa$Club)
fifa$Club_Position <- factor(fifa$Club_Position)
fifa$Preffered_Foot <- factor(fifa$Preffered_Foot)
fifa$Preffered_Position <- factor(fifa$Preffered_Position)
fifa$Work_Rate <- factor(fifa$Work_Rate)


# Height
fifa$Height <- gsub(" [[:alpha:]]*", "", fifa$Height)
fifa$Height <- as.numeric(fifa$Height)

# Weight
fifa$Weight <- gsub(" [[:alpha:]]*", "", fifa$Weight)
fifa$Weight <- as.numeric(fifa$Weight)

head(fifa)
```

```
##                Name Nationality National_Position National_Kit           Club
## 1 Cristiano Ronaldo    Portugal                LS            7    Real Madrid
## 2      Lionel Messi   Argentina                RW           10   FC Barcelona
## 3             Neymar      Brazil                LW           10   FC Barcelona
## 4      Luis SuÃ¡rez     Uruguay                LS            9   FC Barcelona
## 5      Manuel Neuer     Germany                GK            1      FC Bayern
## 6             De Gea       Spain                GK            1 Manchester Utd
##   Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 1            LW        7   07/01/2009            2021     94    185     80
## 2            RW       10   07/01/2004            2018     93    170     72
## 3            LW       11   07/01/2013            2021     92    174     68
## 4            ST        9   07/11/2014            2021     92    182     85
## 5            GK        1   07/01/2011            2021     92    193     92
## 6            GK        1   07/01/2011            2019     90    193     82
##   Preffered_Foot Birth_Date Age Preffered_Position       Work_Rate Weak_foot
## 1          Right 02/05/1985  32              LW/ST      High / Low         4
## 2           Left 06/24/1987  29                 RW Medium / Medium         4
## 3          Right 02/05/1992  25                 LW   High / Medium         5
## 4          Right 01/24/1987  30                 ST   High / Medium         4
## 5          Right 03/27/1986  31                 GK Medium / Medium         4
## 6          Right 11/07/1990  26                 GK Medium / Medium         3
##   Skill_Moves Ball_Control Dribbling Marking Sliding_Tackle Standing_Tackle
## 1           5           93        92      22             23              31
## 2           4           95        97      13             26              28
## 3           5           95        96      21             33              24
## 4           4           91        86      30             38              45
## 5           1           48        30      10             11              10
## 6           1           31        13      13             13              21
##   Aggression Reactions Attacking_Position Interceptions Vision Composure
## 1         63        96                 94            29     85        86
## 2         48        95                 93            22     90        94
## 3         56        88                 90            36     80        80
```

```
## 4         78        93                92           41     84        83
## 5         29        85                12           30     70        70
## 6         38        88                12           30     68        60
##   Crossing Short_Pass Long_Pass Acceleration Speed Stamina Strength Balance
## 1       84         83        77           91    92      92       80      63
## 2       77         88        87           92    87      74       59      95
## 3       75         81        75           93    90      79       49      82
## 4       77         83        64           88    77      89       76      60
## 5       15         55        59           58    61      44       83      35
## 6       17         31        32           56    56      25       64      43
##   Agility Jumping Heading Shot_Power Finishing Long_Shots Curve
## 1      90      95      85         92        93         90    81
## 2      90      68      71         85        95         88    89
## 3      96      61      62         78        89         77    79
## 4      86      69      77         87        94         86    86
## 5      52      78      25         25        13         16    14
## 6      57      67      21         31        13         12    21
##   Freekick_Accuracy Penalties Volleys GK_Positioning GK_Diving GK_Kicking
## 1                76        85      88             14         7         15
## 2                90        74      85             14         6         15
## 3                84        81      83             15         9         15
## 4                84        85      88             33        27         31
## 5                11        47      11             91        89         95
## 6                19        40      13             86        88         87
##   GK_Handling GK_Reflexes
## 1          11          11
## 2          11           8
## 3           9          11
## 4          25          37
## 5          90          89
## 6          85          90
```

We note that we had some variables (Height and Weight) as characters instead of just being numeric. So we proceeded to delete these characters keeping only the numeric values of the field.

For example, Height where we have character type variable "183 cm" becomes numeric 183.

```
fifa2 <- fifa[,1:20]
```

In the end we have a table with the 20 variables that may be of interest, and a total of 17588 observations.

---

3 Data cleaning

3.1 Missing values

Next we will review the missing values of our dataset, it should be noted that in the National_Kit and National_Position ** may ** contain empty values because they refer to the position in the national selection of the player, because not all players can be selected so we will find quite a few empty observations in these columns

```
colSums(is.na(fifa2))
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                  0                  0              16513
##               Club      Club_Position           Club_Kit       Club_Joining
##                  0                  0                  1                  0
```

```
##    Contract_Expiry             Rating             Height             Weight
##                  1                  0                  0                  0
##     Preffered_Foot         Birth_Date                Age Preffered_Position
##                  0                  0                  0                  0
##          Work_Rate          Weak_foot        Skill_Moves       Ball_Control
##                  0                  0                  0                  0
```

```
colSums(fifa2 == "")
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                  0              16513                 NA
##               Club      Club_Position           Club_Kit       Club_Joining
##                  0                  1                 NA                  1
##    Contract_Expiry             Rating             Height             Weight
##                 NA                  0                  0                  0
##     Preffered_Foot         Birth_Date                Age Preffered_Position
##                  0                  0                  0                  0
##          Work_Rate          Weak_foot        Skill_Moves       Ball_Control
##                  0                  0                  0                  0
```

```
colSums(fifa2 <= 0)
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                 NA                 NA                 NA
##               Club      Club_Position           Club_Kit       Club_Joining
##                 NA                 NA                 NA                  1
##    Contract_Expiry             Rating             Height             Weight
##                 NA                  0                  0                  0
##     Preffered_Foot         Birth_Date                Age Preffered_Position
##                 NA                  0                  0                 NA
##          Work_Rate          Weak_foot        Skill_Moves       Ball_Control
##                 NA                  0                  0                  0
```

Note that there is 1 NA in Club_Kit and 1 NA in Contract_Expiry

```
which(is.na(fifa2$Club_Kit), arr.ind=TRUE)
```

```
## [1] 384
```

```
which(is.na(fifa2$Contract_Expiry), arr.ind=TRUE)
```

```
## [1] 384
```

```
which(fifa2$Club_Position == "", arr.ind=TRUE)
```

```
## [1] 384
```

```
# We remove rows with NA except those with NA only in National_Kit
# and National_Position
fifaNet <- fifa2[complete.cases(fifa2[ , -(3:4)]),]
```

As we have seen above, it seems that there is only one player who is missing values outside the excluded columns, so as we observe there is only 1 row difference. We can consider that the change made does not significantly affect the data we have.

```
colSums(is.na(fifaNet))
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                  0                  0              16512
##               Club      Club_Position           Club_Kit       Club_Joining
```

```
##                 0                 0                 0                 0
##    Contract_Expiry            Rating            Height            Weight
##                 0                 0                 0                 0
##     Preffered_Foot        Birth_Date               Age Preffered_Position
##                 0                 0                 0                 0
##          Work_Rate         Weak_foot       Skill_Moves       Ball_Control
##                 0                 0                 0                 0
```

```
colSums(fifaNet == "")
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                  0              16512                 NA
##               Club      Club_Position           Club_Kit       Club_Joining
##                  0                  0                 0                 0
##    Contract_Expiry             Rating            Height            Weight
##                  0                  0                 0                 0
##     Preffered_Foot        Birth_Date               Age Preffered_Position
##                  0                  0                 0                 0
##          Work_Rate         Weak_foot       Skill_Moves       Ball_Control
##                  0                  0                 0                 0
```

```
colSums(fifaNet <= 0)
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                 NA                NA                 NA
##               Club      Club_Position           Club_Kit       Club_Joining
##                 NA                 NA                 0                 0
##    Contract_Expiry             Rating            Height            Weight
##                  0                  0                 0                 0
##     Preffered_Foot        Birth_Date               Age Preffered_Position
##                 NA                  0                 0                NA
##          Work_Rate         Weak_foot       Skill_Moves       Ball_Control
##                 NA                  0                 0                 0
```

The result we obtain is a table with 178587 observations where we do not have absent, empty or zero values. Except for the number in the national position.

3.2 Transformations

In this section we will proceed to discretize some values such as the rating, so that according to the rating of the player is considered:

Type	Rank rating
Excellent	90-100
Very Good	80-89
Bo	70-79
Normal	50-69
Bad	40-49
Very Bad	0-39

```
fifaNet$clasificacion = cut(fifaNet$Rating, c(0,40,50,70,80,90,101),
                            labels = c("Very Bad", "Bad", "Normal",
                                       "Good", "Very Good", "Excellent"),
                            include.lowest = TRUE, right = FALSE)
summary(fifaNet$clasificacion)
```

```
##  Very Bad       Bad    Normal      Good Very Good Excellent
##         0       121     11921      5017       519         9
```

```
fifaNet$clasificacion2 = cut(fifaNet$Rating, c(0,70,101),
                             labels = c("Normal", "Good"), include.lowest = TRUE,
                             right = FALSE)
summary(fifaNet$clasificacion2)
```

```
## Normal   Good
##  12042   5545
```

Additionally Normal players has been separated from good players, where Normal have a rating between 0-69 and good players are between 70-100.

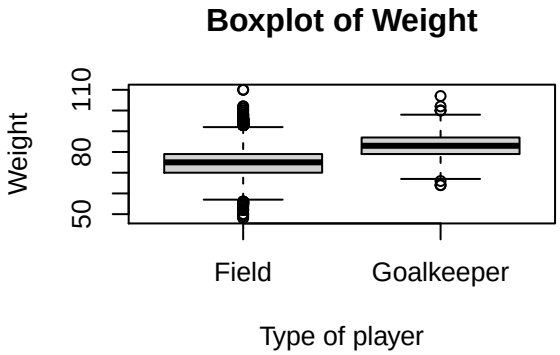
We will differentiate the goalkeeper position from the rest of the players, as this is usually a position with very different characteristics compared to a field player.

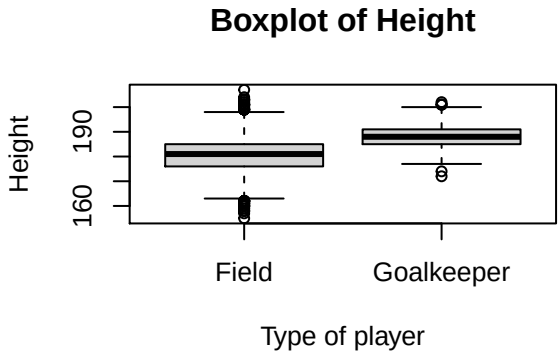
```
# Iterate through the different positions and add TRUE or FALSE depending on
# whether it is a goalkeeper or not, in a new goalkeeper variable
for (i in 1:length(fifaNet$Club_Position)){
  if (fifaNet$Club_Position[i] == "GK"){
    fifaNet$porter[i] = "Goalkeeper"
  } else{
    fifaNet$porter[i] = "Field"
  }
}
fifaNet$porter <- factor(fifaNet$porter)
```

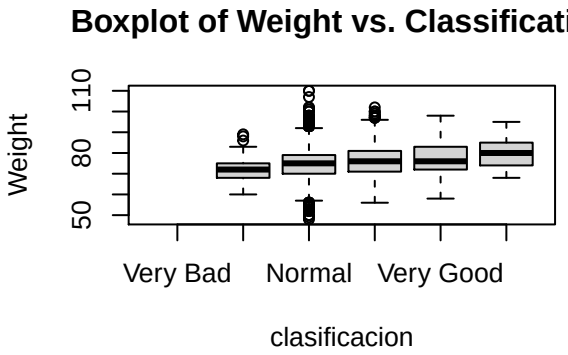
3.3 Extreme values (outliers)

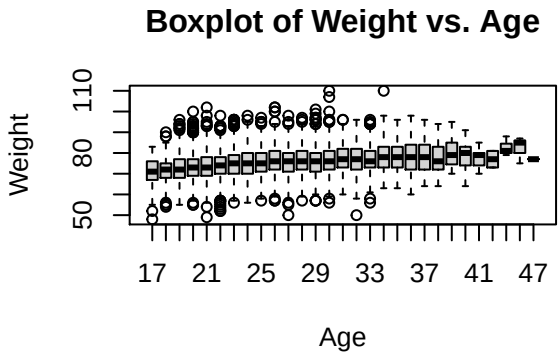
Let's analyze if we have extreme values in ours using the boxplot charts.

```
# Plot boxplot
par(mfrow=c(2,2))
boxplot(fifaNet$Weight ~ fifaNet$porter, main = "Boxplot of Weight", ylab = "Weight", xlab = "Type of pl
boxplot(fifaNet$Height ~ fifaNet$porter, main = "Boxplot of Height", ylab = "Height", xlab = "Type of pl
boxplot(Weight ~ clasificacion , data = fifaNet,
        main = "Boxplot of Weight vs. Classification", ylab = "Weight")
boxplot(Weight ~ Age , data = fifaNet,
        main = "Boxplot of Weight vs. Age", ylab = "Weight")
```









In the Weight chart, we observe how goalkeepers weigh more compared to other players, and the same goes for the case of players 'height.

Weight vs Clasification, we can see how the players considered worse have less weight compared to the best rated players.

Finally in the Wight vs Age chart, we can see that we tend to have more "outliers" in younger players.

But are they really extreme values? Can there be players with more weight than average?

Let's review it:

```
fifaNet %>% arrange_(~ desc(Weight)) %>% head(n = 10)
```

```
##                 Name         Nationality National_Position National_Kit
## 1   Kristof Van Hout             Belgium                 NA           NA
## 2  Adebayo Akinfenwa             England                 NA           NA
## 3        Chris Seitz       United States                 NA           NA
## 4         Bill Hamid       United States                 NA           NA
## 5       Hakeem Araba             England                 NA           NA
## 6         Evan Louro       United States                 NA           NA
## 7     Rene Gilmartin Republic of Ireland                 NA           NA
## 8    Lars Unnerstall             Germany                 NA           NA
## 9         Carl Ikeme             Nigeria                 NA           NA
## 10   Martin PolaÄ\215ek           Slovakia                 NA           NA
##                Club Club_Position Club_Kit Club_Joining Contract_Expiry Rating
## 1      KVC Westerlo           Sub        1   07/01/2015            2017     67
## 2           Wycombe            ST       20   07/10/2016            2017     64
## 3         FC Dallas            GK       18   01/01/2010            2020     68
```

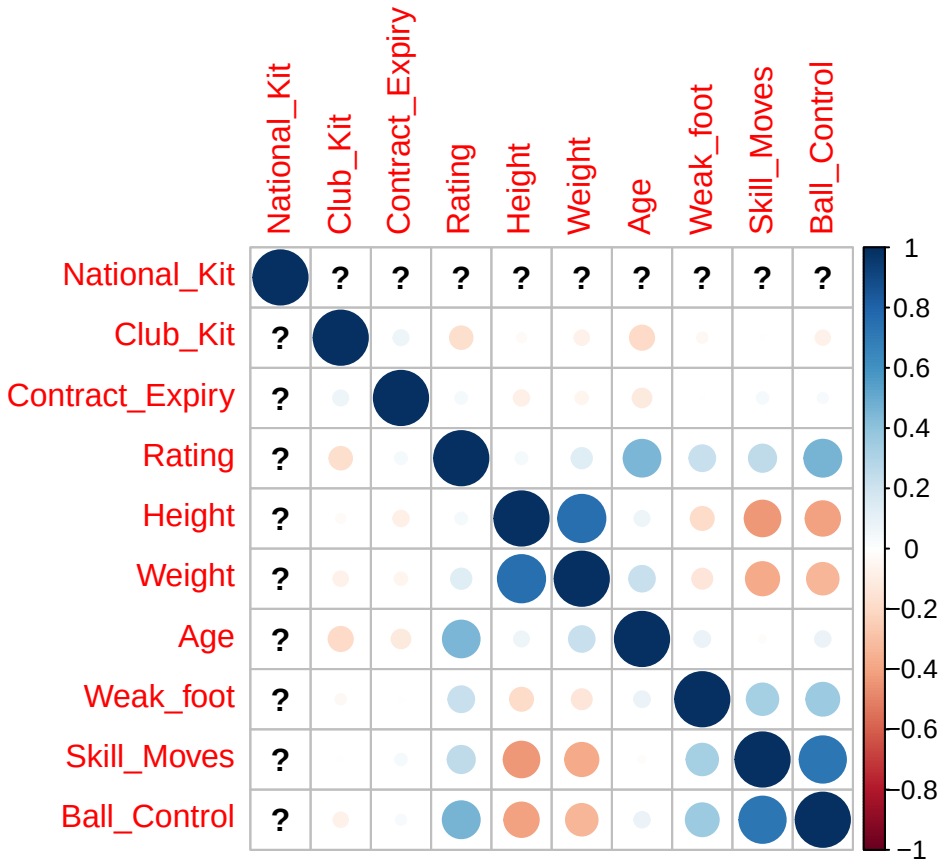
```
## 4       D.C. United            GK       28   01/01/2009            2021     75
## 5    Falkenbergs FF            LS       18   02/27/2015            2021     62
## 6      NY Red Bulls           Res       45   01/23/2017            2020     53
## 7           Watford           Res       13   08/25/2014            2017     59
## 8    F. DÃ¼sseldorf           Sub       19   07/01/2014            2017     72
## 9           Wolves            GK        1   07/01/2003            2019     70
## 10 ZagÅ‚Ä\231bie Lubin            GK        1   07/10/2015            2018     66
##    Height Weight Preffered_Foot Birth_Date Age Preffered_Position
## 1     207    110          Right 02/09/1987  30                 GK
## 2     178    110          Right 05/10/1982  34                 ST
## 3     191    107          Right 03/12/1987  30                 GK
## 4     191    102          Right 11/25/1990  26                 GK
## 5     191    102          Right 02/12/1991  26                 ST
## 6     191    102          Right 01/19/1996  21                 GK
## 7     197    101          Right 05/31/1987  29                 GK
## 8     198    100          Right 07/20/1990  26                 GK
## 9     191    100          Right 06/08/1986  30                 GK
## 10    199    100          Right 04/02/1990  26                 GK
##          Work_Rate Weak_foot Skill_Moves Ball_Control clasificacion
## 1  Medium / Medium         3           1           24        Normal
## 2        Low / Low         3           2           69        Normal
## 3  Medium / Medium         3           1            9        Normal
## 4  Medium / Medium         2           1           20          Good
## 5  Medium / Medium         2           3           58        Normal
## 6  Medium / Medium         2           1           17        Normal
## 7  Medium / Medium         3           1           22        Normal
## 8  Medium / Medium         1           1           20          Good
## 9  Medium / Medium         3           1           24          Good
## 10 Medium / Medium         2           1           19        Normal
##    clasificacion2     porter
## 1          Normal      Field
## 2          Normal      Field
## 3          Normal Goalkeeper
## 4            Good Goalkeeper
## 5          Normal      Field
## 6          Normal      Field
## 7          Normal      Field
## 8            Good      Field
## 9            Good Goalkeeper
## 10         Normal Goalkeeper
```

We note that most of the heaviest players are those of the tallest players, this makes sense on a physical level as a person's height usually affects their weight.

We make a correlation map to confirm the idea:

```
cor1 <- cor(fifaNet[sapply(fifaNet,is.numeric)])

corrplot(cor1, method="circle")
```



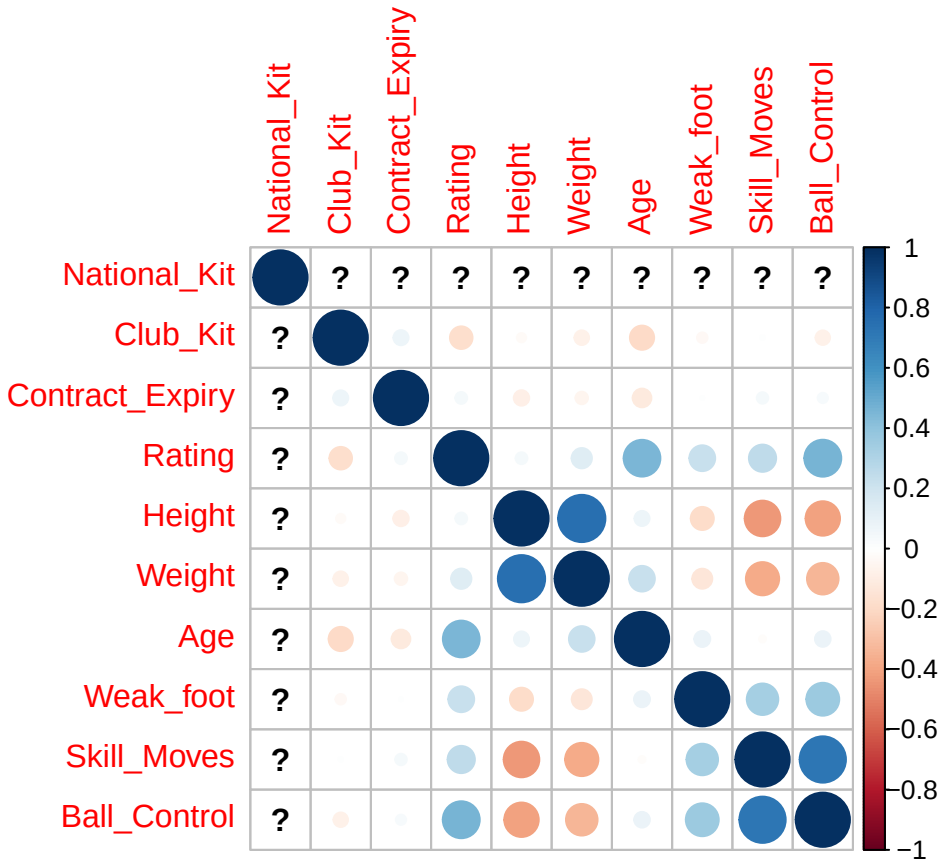
As we observe we have that Weight and Height have a positive correlation and it seems to be close to 0.8 this means that the higher the height the more weight.

---

4 Data analysis

Based on the map of correlations found above:

```
corrplot(cor1, method="circle")
```



We note that apart from Weight-Height we have more interesting correlations between variables.

For example Age and Rating, with a positive correlation which tells us that the older a player tends to have a higher rating. Which makes sense, since an older player means that usually has more experience.

On the other hand, it seems that we have negative correlations Skill_Moves-Height or Ball_Control-Height which is indicating to us that players with more height tend to have less skill with the ball.

4.1 Selection of data to be used

Based on the data selected in point 2 and the correlation map, the variables that interest us most are the following:

Variable	Description
National_Kit	Team number on the national team
Club_Position	Position of play in the club
Rating	Overall rating of the player, between 0 and 100
Height	Height
Weight	Weight
Preffered_Foot	Little favorite
Age	Age
Preffered_Position	Preferred position
Work_Rate	qualitative assessment in terms of attack-defense
Weak_foot	rating of 1 to 5 control and power of the leg not preferred
Skill_Moves	rating from 1 to 5 of the player's movement ability
Ball_Control	1 to 5 rating of the ball control

The reason for selecting this data is because it is physical information of the player and they are the ones who give more information of their ability. Additionally National_kit has been selected as it will be useful for us to know if the player ends up in the national team or not. And club_position allows us to differentiate the different positions of the players, in our case what we will take into account the goalkeeper as he is the only one who has a great variation of requirements and position compared to field players (the other positions).

4.2 Checking the normality and homogeneity of the variance

In our dataset the only values that could study its normality are rating, weight, height and age as they are mainly the only non-categorical variables in our dataset.

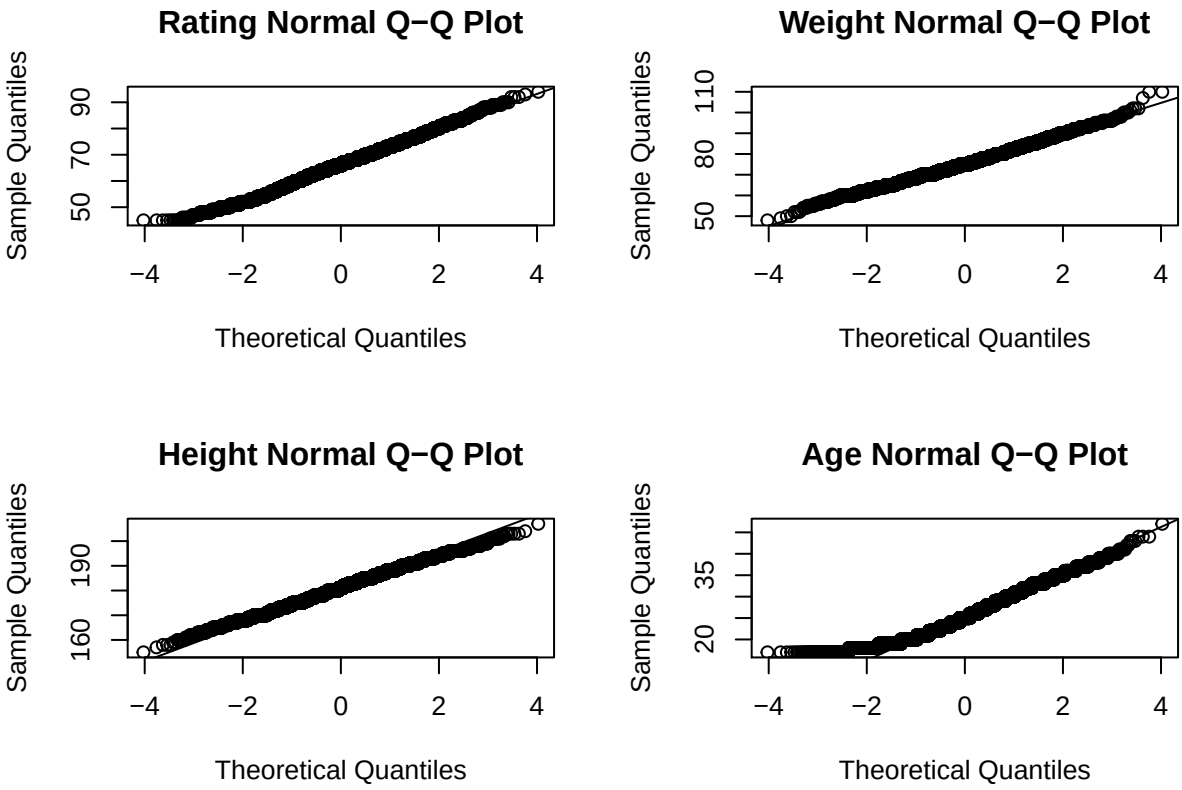
Since we have a large number of observations we will try to visualize their normality through the QQ-plot, generated with the function qqnorm (), which presents the values of the variables on an X axis, this representation of the values is they should align with the diagonal line represented by the qqline () function. This will tell us whether these variables have an approximately normal distribution or not.

```
par(mfrow=c(2,2))
qqnorm(fifaNet$Rating, main='Rating Normal Q-Q Plot')
qqline(fifaNet$Rating)

qqnorm(fifaNet$Weight, main='Weight Normal Q-Q Plot')
qqline(fifaNet$Weight)

qqnorm(fifaNet$Height, main='Height Normal Q-Q Plot')
qqline(fifaNet$Height)

qqnorm(fifaNet$Age, main='Age Normal Q-Q Plot')
qqline(fifaNet$Age)
```



As we can see, it seems that all 4 variables have a distribution that is close to normal, although it should be noted that at lower ages they do not follow this normal much.

To make sure we will perform the Kolmogorov-Smirnov test with these variables, as Shapiro-Wilk works when there are a total of observations between 3 and 5000:

We consider our null hypothesis to be that the population is normally distributed.

```
# Rating
ks.test(fifaNet$Rating, pnorm, mean(fifaNet$Rating), sd(fifaNet$Rating))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Rating
## D = 0.043971, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# Weight
ks.test(fifaNet$Weight, pnorm, mean(fifaNet$Weight), sd(fifaNet$Weight))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Weight
## D = 0.05617, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
# Height
ks.test(fifaNet$Height, pnorm, mean(fifaNet$Height), sd(fifaNet$Height))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Height
## D = 0.048542, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
# Age
ks.test(fifaNet$Age, pnorm, mean(fifaNet$Age), sd(fifaNet$Age))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Age
## D = 0.083494, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

In the Kolmogorov-Smirnov results it gives us that all p-values are less than 0.05, meaning that the null hypothesis must be rejected because p-values is less than the significance level.

But as shown in the qqnorm graphs but taking into account the central limit theorem we can assume that there is normality in the distribution when there is a large number of observations as is our case with 17587 observations.

The theorem tells us that the average of a sample of any data set is becoming more normal as we increase the number of observations.

Next we consider whether there is homogeneity of variance with the players who are goalkeepers, as these are the ones who most differentiate between the other players. We apply the homoscedasticity test with Levene for data with normal distribution:

```r
# Rating
leveneTest(Rating ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     1  3.8863 0.0487 *
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Weight
leveneTest(Weight ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1   14.97 0.0001096 ***
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Height
leveneTest(Height ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
```

```
## group     1  125.39 < 2.2e-16 ***
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Age
leveneTest(Age ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     1  1.3161 0.2513
##       17585
```

As we see almost all have a p-value lower than the significance level (0.05) what is telling us is that the null hypothesis is rejected, meaning that there are statistically different variances for groups of players who are goalkeepers.

On the other hand, we see that in the case of age the p-values is higher than the level of significance, this means that the goalkeepers do not present different variances compared to the field players.

Let's check the homoscedasticity of the variables based on the type of classification (clasificacion2) that we have created where we separate the Normal players from the good ones:

```
# Rating
leveneTest(Rating ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1  519.77 < 2.2e-16 ***
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Weight
leveneTest(Weight ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value   Pr(>F)
## group     1  9.1628 0.002473 **
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Height
leveneTest(Height ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value  Pr(>F)
## group     1  5.3889 0.02028 *
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Age
leveneTest(Age ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1  129.28 < 2.2e-16 ***
```

```
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, the null hypothesis of homogeneity of variances by the Rating variables is rejected, but it is interesting to note that it also happens with age.

In terms of weight and height, it seems that they are homogeneous.

4.3 Data analysis

4.3.1 Hypothesis testing It is often said that younger players have better physique compared to veterans so young players are considered to be better at playing. Is that really so? Starting from goalkeeper and classification2, we ask ourselves the following questions:

- Are players better when they are younger?
- A goalkeeper weighs more than a fielder because they don't have to run around the field?

  Let's make a hypothesis test of the questions asked:

  Are players better when they are younger?

- Null Hypothesis (H0): Good players **AREN'T** the youngest ones
- Alternative Hypothesis (H1): Good players **RE** the youngest ones

  We begin to compare the variances of the two samples:

```
df_1<-fifaNet$Age[fifaNet$clasificacion2=="Good"]
df_2<-fifaNet$Age[fifaNet$clasificacion2=="Normal"]
var.test(df_1,df_2)
```

```
##
##  F test to compare two variances
##
## data:  df_1 and df_2
## F = 0.75159, num df = 5544, denom df = 12041, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7186574 0.7863226
## sample estimates:
## ratio of variances
##          0.7515866
```

The var.test test of R shows us a p-value less than 0.05 so equality of variances cannot be assumed in both populations. Therefore, we apply a test on the average of two independent samples with variance unknown and different. It is a unilateral test from the right.

We proceed to perform the hypothesis test:

```
t.test(fifaNet$Age~relevel(fifaNet$clasificacion2,ref="Good"),m=5,var.equal=F,
       alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  fifaNet$Age by relevel(fifaNet$clasificacion2, ref = "Good")
## t = -28.753, df = 12294, p-value = 1
## alternative hypothesis: true difference in means is greater than 5
## 95 percent confidence interval:
```

```
##  2.910547      Inf
## sample estimates:
##   mean in group Good mean in group Normal
##             27.52985             24.50623
```

For a confidence level of 95%, we have a p-value = 1 greater than the significance level, therefore **we cannot reject the null hypothesis because as the results show, the average age of the good players are 27.53 years and the average age of Normal players are 24.51 years.**

Let's move to the next question:

A goalkeeper weighs more than a fielder because don't they have to run around the field?

- Null hypothesis (H0): Goalkeepers **DO NOT** weigh more than other players
- Alternative Hypothesis (H1): Goalkeepers **DO** weigh more than other players

```
df_1<-fifaNet$Weight[fifaNet$porter=="Goalkeeper"]
df_2<-fifaNet$Weight[fifaNet$porter=="Field"]
var.test(df_1,df_2)
```

```
##
##  F test to compare two variances
##
## data:  df_1 and df_2
## F = 0.78916, num df = 631, denom df = 16954, p-value = 7.335e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7073142 0.8857220
## sample estimates:
## ratio of variances
##          0.7891574
```

The var.test test of R shows us a p-value less than 0.05 so equality of variances cannot be assumed in both populations. Therefore, we apply a test on the average of two independent samples with variance unknown and different.

We proceed to perform the hypothesis test:

```
t.test(fifaNet$Weight~relevel(fifaNet$porter,ref="Goalkeeper"),m=5,var.equal=F,
       alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  fifaNet$Weight by relevel(fifaNet$porter, ref = "Goalkeeper")
## t = 12.446, df = 691.96, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 5
## 95 percent confidence interval:
##  7.639654      Inf
## sample estimates:
## mean in group Goalkeeper      mean in group Field
##                 83.00633                74.96408
```

From the results obtained we have that a p-value lower than the significance level (0.05) tells us that we can reject the null hypothesis in favor of the alternative, therefore, we can say that **the hypothesis that the goalkeepers weigh more than field players, where goalkeepers weigh an average of 83kg and fielders 74kg.** This result makes sense as the goalkeeper is not a position that needs to constantly move around the field but needs good muscle mass to stop the shots having their legs as the only point of support.

4.3.2 Linear regression

4.3.2.1 Model If the rating is a consequence of quality factors we would result in a model that could explain almost all the players in the database. If, on the other hand, this is not the case, we would have to use hidden and external values in numerical valuations to establish the rating.

```
model <- lm(Rating ~ Skill_Moves + Ball_Control + Age + Height + Weight +
              clasificacion, data = fifaNet)
summary(model)
```

```
##
## Call:
## lm(formula = Rating ~ Skill_Moves + Ball_Control + Age + Height +
##     Weight + clasificacion, data = fifaNet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0399  -2.2793   0.1727   2.5029  12.5006
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              22.873042   1.001041  22.849  < 2e-16 ***
## Skill_Moves              -0.377372   0.053679  -7.030 2.14e-12 ***
## Ball_Control              0.126540   0.002482  50.976  < 2e-16 ***
## Age                       0.333803   0.006199  53.847  < 2e-16 ***
## Height                    0.048745   0.006408   7.607 2.95e-14 ***
## Weight                    0.081520   0.006170  13.213  < 2e-16 ***
## clasificacionNormal      10.647914   0.326623  32.600  < 2e-16 ***
## clasificacionGood        18.705150   0.334306  55.952  < 2e-16 ***
## clasificacionVery Good   26.890972   0.367558  73.161  < 2e-16 ***
## clasificacionExcellent   34.445926   1.226067  28.095  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.533 on 17577 degrees of freedom
## Multiple R-squared:  0.7513, Adjusted R-squared:  0.7511
## F-statistic:  5899 on 9 and 17577 DF,  p-value: < 2.2e-16
```

The adjusted R2 value is 0.7511. That is, the model explains 75.11 of the variance in the rating of the players. This indicates that the explanatory ability of the model is good for estimating player scores. The p-value of the model is less than 0.05 and therefore the set of explanatory variables contributes significantly to explaining the rating of the players.

In relation to the separate analysis of the explanatory variables, it is observed that all the variables of the model are significant. The variables Ball_Control, Age, Height, Weight and clasificacion have a positive correlation with the rating, indicating that the player's score increases if the values of these variables increase. In contrast, the Skill_Moves variable has a negative correlation: when this variable takes on more value, and the other variables take on the same value, its score is reduced by about -0.311565

4.3.2.2 Prediction. Let's apply the regression model to predict a player with movement ability 4, ball control 70, age 24, height 179, weight 70 and "Good" rating

```
new<-data.frame(Skill_Moves=4 , Ball_Control =70,  Age=24, Height=179,
    Weight=70, clasificacion= 'Good')
predict(model,new,type="response")
```

```
##        1
## 71.36955
```

The model predicts that a player with these characteristics will have a rating of 72.13. This prediction should be taken with some caution as the R2 is not entirely elevated

Let's look for the player with the highest rating:

```
fifaNet[which.max(fifaNet$Rating),]
```

```
##                Name Nationality National_Position National_Kit        Club
## 1 Cristiano Ronaldo    Portugal                LS            7 Real Madrid
##   Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 1            LW        7   07/01/2009            2021     94    185     80
##   Preffered_Foot Birth_Date Age Preffered_Position  Work_Rate Weak_foot
## 1          Right 02/05/1985  32              LW/ST High / Low         4
##   Skill_Moves Ball_Control clasificacion clasificacion2 porter
## 1           5           93     Excellent           Good  Field
```

Cristiano Ronaldo has the highest Rating of 94.

Now let's make the prediction of its rating we take into account its characteristics:

```
new2<-data.frame(Skill_Moves=5 , Ball_Control =93,  Age=32, Height=185,
    Weight=80, clasificacion= 'Excellent')
predict(model,new2,type="response")
```

```
##        1
## 93.42147
```

We find that its rating predicted is very similar to the rating that we have from the dataset

Let's look for the player with the lowest rating:

```
fifaNet[which.min(fifaNet$Rating),]
```

```
##                Name Nationality National_Position National_Kit          Club
## 17579 Steven Alzate     England                              NA Leyton Orient
##       Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 17579           Res       31   09/21/2016            2017     45    177     65
##       Preffered_Foot Birth_Date Age Preffered_Position       Work_Rate
## 17579          Right 09/08/1998  18               CAM Medium / Medium
##       Weak_foot Skill_Moves Ball_Control clasificacion clasificacion2 porter
## 17579         3           3           46           Bad         Normal  Field
```

The player with the lowest rating is 45

Now let's make the prediction of its rating we take into account its characteristics:

```
new2<-data.frame(Skill_Moves=3 , Ball_Control =46,  Age=32, Height=177,
    Weight=18, clasificacion= 'Bad')
predict(model,new2,type="response")
```

```
##        1
## 48.33872
```

Thus, we find that its rating has a difference of 3 compared to the model.

From the results we see that the model has a fairly high accuracy with extreme values with differences of up to 3 points

4.3.3 Logistic regression

4.3.3.1 Model Finally, in order to know if a player has the potential to go to the national team we will use their national number standard "National_Kit" as a result label. First of all we create a column with international name where 0 = is not selected for selection and 1 = selected for selection.

```
for (i in 1:length(fifaNet$National_Kit)){
  if (is.na(fifaNet$National_Kit[i])){
    fifaNet$internacional[i] = 0
  } else{
    fifaNet$internacional[i] = 1
  }
}
```

Next we create the logistic model from the variables clasificacion2, Rating, Age and Work_Rate. The reason for these variables is because starting from the linear model we see that "clasificacion2" which would be the equivalent of an evaluation if the player is good or not is a weighty estimator with the final result; the rating, obtained from the linear model, is a point that has been considered important when knowing if a player will go to the selection; Age has been selected because as we have seen in the hypothesis test it appears that the best rated players are those who are older so it is estimated that he will be an influential estimator; And finally Work_Rate which similar to the "clasificacion2" is also an evaluation of the player's performance that could be obtained without having to perform a physical examination of the matiex player.

```
# Model logistic
mrl <- glm(internacional ~ clasificacion2 + Rating + Age + Work_Rate,
           family = binomial(link = logit), data = fifaNet)
summary(mrl)
```

```
##
## Call:
## glm(formula = internacional ~ clasificacion2 + Rating + Age +
##     Work_Rate, family = binomial(link = logit), data = fifaNet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5827  -0.3474  -0.2220  -0.1358   3.6109
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -15.338319   0.602103 -25.475  < 2e-16 ***
## clasificacion2Good         0.059286   0.116016   0.511 0.609340
## Rating                     0.196367   0.008738  22.473  < 2e-16 ***
## Age                       -0.025564   0.008558  -2.987 0.002815 **
## Work_RateHigh / Low       -0.434751   0.189956  -2.289 0.022097 *
## Work_RateHigh / Medium    -0.489358   0.132643  -3.689 0.000225 ***
## Work_RateLow / High       -0.536066   0.245249  -2.186 0.028830 *
## Work_RateLow / Low        -1.412971   1.068384  -1.323 0.185991
## Work_RateLow / Medium     -0.554730   0.258379  -2.147 0.031797 *
## Work_RateMedium / High    -0.435535   0.149327  -2.917 0.003538 **
## Work_RateMedium / Low     -0.853741   0.202511  -4.216 2.49e-05 ***
## Work_RateMedium / Medium  -0.553934   0.122836  -4.510 6.50e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8091.8  on 17586  degrees of freedom
## Residual deviance: 6457.3  on 17575  degrees of freedom
```

```
## AIC: 6481.3
##
## Number of Fisher Scoring iterations: 6
```

```
# Devianza
anova(mrl,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: internacional
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           17586     8091.8
## clasificacion2  1  1016.56     17585     7075.2 < 2.2e-16 ***
## Rating          1   581.59     17584     6493.7 < 2.2e-16 ***
## Age             1    10.43     17583     6483.2  0.001237 **
## Work_Rate       8    25.90     17575     6457.3  0.001094 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that Work_Rate is a dichotomous variable with 9 options so an estimator has been obtained for each of the possibilities by taking the reference Work_Rate High / High. It should be noted that Work_Rate Low / Low has a considerably higher negative estimator than other estimators, this is telling us, of course, a player with a low / low performance is much less likely to be selected compared to others.

We observe that the model is valid because the residual deviance is less than the null deviance, where by doing anova we see that all p-values are significant. Therefore, we confirm that the selected variables have weight when obtaining the final model.

4.3.3.2 Confusion matrix Next we check the accuracy of the logistic model to paritr of the dataset used the international column as a result to check the correct prediction:

```
# Exercici 5.2
newFifaNet <- fifaNet[, c("clasificacion2", "Rating", "Age", "Work_Rate")]
pred2 <- predict(mrl, newFifaNet ,type ="response")

matriuConf <- matrix(0, nrow = 3, ncol = 3,
                     dimnames = list(c("Internacional", "No Internacional",
                                       "Total"),c("< 50%", ">= 50%", "Total")))
for (i in 1:length(pred2)){
  if (pred2[i] < 0.50){
    if (fifaNet$internacional[i] == 1){
      matriuConf[1,1] <- matriuConf[1,1] + 1
    }
    else {
      matriuConf[2,1] <- matriuConf[2,1] + 1
    }
  }
  else {
    if (fifaNet$internacional[i] == 1){
      matriuConf[1,2] <- matriuConf[1,2] + 1
```

```
    }
    else {
      matriuConf[2,2] <- matriuConf[2,2] + 1
    }
  }
}
# Total, < 50% i >= 50%
matriuConf[3,1] <- matriuConf[1,1] + matriuConf[2,1]
matriuConf[3,2] <- matriuConf[1,2] + matriuConf[2,2]
# Internacional i No Internacional, Total
matriuConf[1,3] <- matriuConf[1,1] + matriuConf[1,2]
matriuConf[2,3] <- matriuConf[2,1] + matriuConf[2,2]
# Total, Total
matriuConf[3,3] <- matriuConf[3,1] + matriuConf[3,2]

matriuConf
```

```
##                   < 50% >= 50% Total
## Internacional       985     90  1075
## No Internacional  16474     38 16512
## Total             17459    128 17587
```

As we have seen in the matrix there are a total of 16474 players who are correctly ranked this means that the model has an accuracy of 16474/17587 = 93.67%.

On the other hand, there are a high number of players who had a 50% probability of being international and 985 of these have been selected. Similarly there were 38 cases that had a high probability and have not been selected.

4.3.3.3 Prediction. Below is a prediction test of a player who is not expected to be selected by the national team and another with a high probability of being selected:

- **Good Player**: 27-year-old player (average age of good players found in the hypothesis test), a 90-point rating and a Work_Rate rating with High / High
- **Normal Player**: 24-year-old player (average age of Normal players found in hypothesis test), a 69-point rating and a Work_Rate rating with Medium / Medium

```
# Jugador bo
pred3 <- predict(mrl, data.frame(clasificacion2 = "Good", Rating = 90, Age = 27,
                                 Work_Rate = "High / High") ,type ="response")
cat("Probabilitat que el jugador bo sigui seleccionat: ", pred3, "\n")
```

```
## Probabilitat que el jugador bo sigui seleccionat:  0.8460277
```

```
# Jugador Normal
pred4 <- predict(mrl, data.frame(clasificacion2 = "Normal", Rating = 69,
                                 Age = 24, Work_Rate = "Medium / Medium"),
                 type ="response")
cat("Probabilitat que el jugador Normal sigui seleccionat: ", pred4)
```

```
## Probabilitat que el jugador Normal sigui seleccionat:  0.0494321
```

Note that there is a possibility that the first player is selected is 84.60% while the other has a very low probability of 4.94%

We see the possibilities of a young player who has potential:

- **Young player**: 21-year-old player, a 79-point rating and a Work_Rate rating with High / Medium

```r
# Young player
pred5 <- predict(mrl, data.frame(clasificacion2 = "Good", Rating = 79,
                                 Age = 21, Work_Rate = "High / Medium"),
                 type ="response")
cat("Probability that the good player is selected: ", pred5)
```

```
## Probability that the good player is selected:  0.3116895
```

A young player with a rating of 79 (Good) has been placed within the scale specified in point 3.2, we see that the probability of being selected is still relatively low but is much higher compared to a Normal player.

---

5 Conclusions

In this analysis, 3 different types have been performed, a hypothesis test with assumptions of a player's abilities with respect to their physical condition, an analysis with a linear model in order to obtain a numerical assessment of the player from information that could be extracted visually and finally a logistic model to evaluate the possibilities of a player being selected by the national team.

5.1 Hypothesis testing

In the first case we found that a young player does not seem to be better than a more veteran / senior, from here we can draw the conclusion that if you want immediate improvement results will have to sign players aged around 27 years which tend to be better, but if you want to improve a player for the future, either for costs or availability, it is worth signing a player around 24 as they have the potential to improve for the future.

Then we saw that players with the goalkeeper position is the one who tends to have more weight due to the need for strength to stop shots from the strikers of the opposing team.

5.2 Linear model

The variables 'Ball_Control', 'Age', 'Height', 'Weight', 'classification' have a positive and significant correlation with 'Rating', whereas Skill_Moves has a negative correlation. Taken together, the explanatory variables account for 75% of the 'Rating' variance.

5.3 Logistics model

In this section we have made a logistic regression model in order to obtain the probability that a player is selected to play internationally based on certain characteristics. In this section we have seen that the main regressors that affect the probability of being selected is whether it is Good or Normal, its Rating and the Work_Rate. It is interesting to note that depending on which level of Work_Rate we take as a reference the other levels have a greater or lesser effect, in the same way a great effect is the player's rating where naturally a player with a higher rating means he is better.

Contributions Firm ———————— | ————————- Prior research Junjie Zhu, Antoni Sanchez Teruel Writing the answers Junjie Zhu, Antoni Sanchez Teruel Code development Junjie Zhu, Antoni Sanchez Teruel