Guillem Miralles and Miguel Payà

ALL NBA TEAM PREDICITON MODEL

1. SUMMARY: WHAT DO WE WANT?

- ALL NBA TEAM of the Year is an annual NBA award given to the best players of the season. Voting is by a group of sports journalists and announcers from the United States and Canada. The team has been chosen in each NBA season, since its inauguration in 1946. The award consists of three quintets consisting of a total of 15 players, five on each team. It originally consisted of two teams, but in 1988 it was increased to three.

  Players receive five points for each vote on the first team, three points for each vote on the second team, and one point for each vote on the third team. The five players with the highest total number of points enter the first team, with the next five players integrating the second team and the same with the third. There is a position restriction. In each voting of 5 players (of each quintet), 2 players are voted that are guard, in our data.frame "PG" and "SG"; the other 2 being forward, "SF" and "PF" and the last player being center "C". Same with the other two teams.

  They are basically the top 15 players of the season. We will look at the statistics made by all the players in one season (more or less 650), and with this information, we will try to know which player will be in the all-NBA team and which.

2. DATA:

- Our information for making this model is divided into two databases:
  - A database where we have all the information from 1980 to 2017 with all the player statistics for each season. We also introduce the all-nba teams variable called "quintet" that will provide us with information on whether or not the player is in the quintet of the season.
  - And another where we have the information of the data of the current season.

In the first database we have the statistics of each player in a given season. The data have the following variables:

-***Year***: *the season to which the player's statistics belong*
-***Player***: *player name*
-***Pos***: *player position*
-***Age***: *age of the player*
-***Tm***: *player team*
-***G***: *the total number of matches played.*
-***GS***: *the total number of matches played.*
-***MP***: *total minutes played.*
-***PER***: *Rating per minute*
-***TS%:*** *Probability of scoring one point for each shot attempted*
-***3PAr***: *Percentage of three shots attempted per field shot attempted*
-***FTr***: *Number of free throws attempted per attempted field goal.*
-***ORB%***: *Both percent of the offensive rebounds obtained*
-***DRB%:*** *Both percent of the defensive rebounds obtained*
-***TRB***%: *Both percent of the total rebounds obtained*
-***AST%***: *Of all the points scored by the team, both percent of those that have been generated by an assist of the player in question*
-***STL%***: *Both percent of a player's thefts for the team's defensive play while on the court*
-***BLK%***: *Both percent of a player's caps for the team's defensive play while on the court*
-***TOV%:*** *So much percentage of ball losses for each turn a player has the ball*
-***USG***%: *Both percent of the team's offensive plays completed by the player in question*
-***OWS***: *Victories that this player has given to the team only offensively*
-***DWS***: *Victories that this player has given to the team only defensively*
-***WS***: *Victories that this player has given to the team in total.*
-***WS/48***: *Victories that this player has given to the team per game*
-***OBPM***: *Advanced statistics that value a player's contribution offensively for 100 possessions*
-***DBPM***: *Advanced statistics that value a player's contribution defensively for 100 possessions*
-***BPM***: *Advanced statistics that value a player's contribution for 100 possessions*
-***VORP***: *Advanced statistics that deal with the value that a player brings to his team.*
-***FG***: *total shots scored.*
-***FGA***: *total shots fired.*
-***FG***%: *percentage of shots scored.*
-***3P***: *total of 3 shots scored.*
-***3PA***: *total of 3 shots fired.*
-***3P%:*** *percentage of shots from 3 scored.*
-***2P***: *total of 2 shots scored.*
-***2PA***: *total of 2 shots fired*
-***2P***%: *percentage of shots from 2 scored*
-***eFG***%: *Probability of scoring one point for each field goal taken*
-***FT***: *total free throws scored.*

-**FTA**: total free throws taken.
-**FT%**: percentage of free throws scored.
-**ORB**: total offensive rebounds caught.
-**DRB**: total defensive rebounds captured.
-**TRB**: total bounces caught.
-**AST**: total assistance given.
-**STL**: total of robberies committed.
-**BLK**: total plugs placed.
-**TOV**: total ball losses.
-**PF**: total personal misconduct committed.
-**PTS**: total points scored.

+ The variable we enter:

-**quintet**: *If the player is in the all nba team of the season (1 if so and 0 if not).*

- In our second database (current season information) we have fewer variables than in the first. These variables have been adapted so that they are identical to the previous ones. For this issue we have to make a model similar to the one in the previous database, but less accurate due to the lack of variables.

-Player
-Tm
-Pos
-Age
-G
-USG% : *Percentage of offensive plays by the team completed by the player in question*
-TOV%: *Tant per cent de pèrdues de baló per cada volta que un jugador té el baló*
-FTA: *total free throws taken.*
-FT%: *percentage of free throws scored.*
-2PA : *total of 2 shots scored*
-2P%: *total of 2 shots scored*
-PA : *total of 3 shots fired*
-3P%: *percentage of shots from 3 scored eFG.*
-TS%: *Probability of conceding one point for each shot fired*
-PTS: *total points scored*
-AST%: *Of all the points scored by the team, both percent of those that have been generated by an assist of the player in question*
-TRB: *total bounces caught*
-AST: *total assistance given*
-STL: *Both percent of a player's thefts for the team's defensive play while on the court*
-BLK: *total plugs placed*
-TOV: *total ball losses*

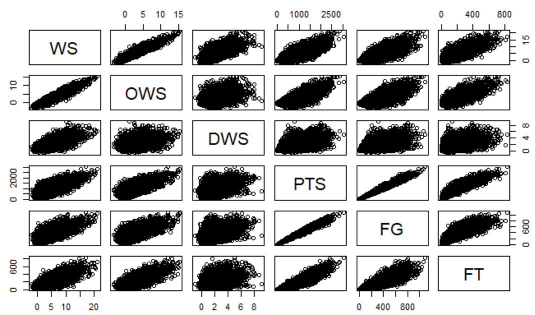
3. PROCESS:

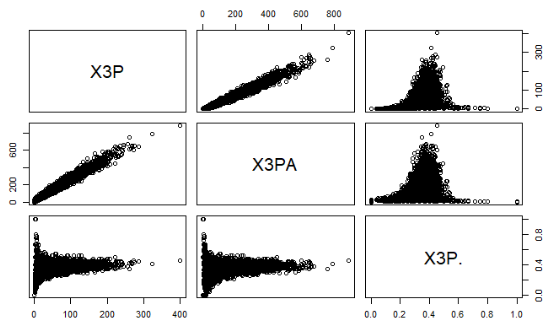
|Complet code at .rmd i HTML|

- We will divide the process into two parts:
    - On the one hand we have the creation of the model with a database that contains all the statistics from 1980 to 2017. In this, we will predict the quintets of the years 2015, 2016 and 2017. In the study we will be able to check which players have predicted correctly, and which ones don't.
    - In the second part we will take the data available this year from the competition to the stop due to the COVID-19. In this one we will not be able to check if the model hits the players or not as this prize has not been awarded yet. We will also work on a data.frame that does not contain the same variables (although it is similar) so we will have to create another model.

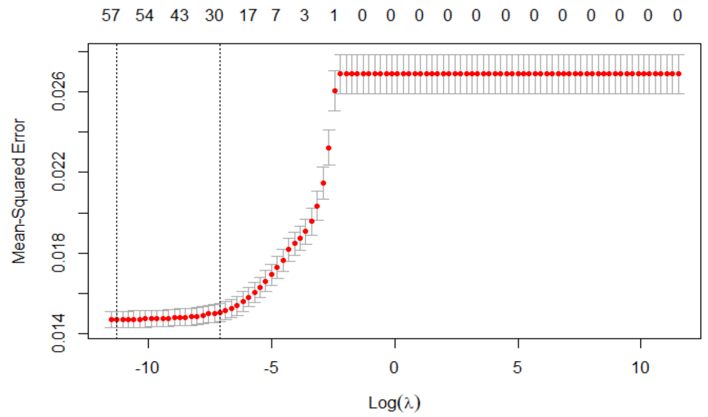
3.1. MODEL CREATION PROCESS WITH KNOWN DATA (1980 - 2017)

- The steps we have followed are as follows:
    - Reading and cleaning the dataframe (reading the csv, setting the variable type, changing null values to 0, deleting rows with duplicate players…)
    - Introduce a variable in dataframe called a quintet, which tells us whether or not the player has been in the best quintet of that season.
    - We set up a training set, and a test set. The training set will have data from 1980 to 2011, and the test set will have data from 2011 to 2017, plus or minus 80% -20%.
    - We visualize the data and observe that many of the variables correlate with each other or do not provide us with relevant information. Therefore, having so many predictive variables, want to regularize them.





- We have observed that there are many correlated variables. Therefore we are going to realize the regularization of variables, with the purpose of to reduce the variance of the same. We use the Lasso method and check if the results obtained are adjusted.




```
 (Intercept)         PosPF         PosSF           Age             G            MP           TS.
2.222875e-02 -1.436547e-03 -3.706560e-03  3.249107e-04  7.727267e-04 -1.716349e-04 -4.249438e-02
         FTr          STL.          TOV.          USG.           DWS            WS           BPM
-1.607393e-02 -7.022702e-04  2.898649e-04 -2.537874e-04  6.191524e-03  1.821565e-02 -4.986440e-04
        VORP            FG           FGA           X2P           FTA            PF           ORB
3.161288e-02  2.369019e-05  1.295914e-04  6.016663e-05  3.683313e-04 -3.212804e-04 -1.507431e-04
         DRB           AST           STL           BLK
1.636141e-04  1.907179e-04 -9.944877e-05  3.400609e-04
```

- We observe that these variables are the ones that the Lasso method indicates to us that they are more explanatory, since they have different coefficients from 0.

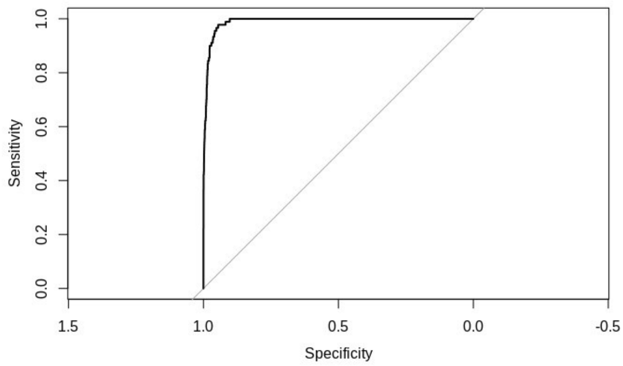
- We perform a Multiple Logistic Regression (GLM) method with these variables in order to do the next step. This command tells us the best glm comparing method using AKAIKE information (AIC).

```
       Df Deviance     AIC
<none>      911.26  941.26
- Age   1   914.56  942.56
- STL.  1   915.38  943.38
- PF    1   917.01  945.01
- ORB   1   918.41  946.41
- TOV.  1   920.89  948.89
- BLK   1   924.19  952.19
- USG.  1   925.62  953.62
- VORP  1   927.48  955.48
- BPM   1   929.51  957.51
- DWS   1   937.73  965.73
- AST   1   968.08  996.08
- G     1   971.50  999.50
- FGA   1   984.52 1012.52
- WS    1  1107.48 1135.48

Call:  glm(formula = quinteto ~ Age + G + STL. + TOV. + USG. + DWS + 
    WS + BPM + VORP + FGA + PF + ORB + AST + BLK, family = "binomial", 
    data = bd)

Coefficients:
(Intercept)          Age            G         STL.         TOV.         USG.          DWS  
 -10.818575     0.045524    -0.133487    -0.220845     0.083679     0.104698     0.447596  
         WS          BPM         VORP          FGA           PF          ORB          AST  
   0.885259     0.166501    -0.498484     0.004465    -0.005797     0.004185     0.004953  
        BLK  
   0.008042  

Degrees of Freedom: 15359 Total (i.e. Null);  15345 Residual
Null Deviance:      3944 
Residual Deviance: 911.3        AIC: 941.3
```

- We see how the variables we are interested in are greatly reduced. As we are performing a logistic regression, in the variables we obtained from the previous point, we perform three models using three different methods which are the ones we will compare. These three methods are: Multiple Logistic Regression (GLM), Quadratic Discriminant Analysis (QDA), and Linear Discriminant Analysis (LDA). We do not take the KNN method because we already know that neighboring values are not interesting for predicting the next value.
- We make comparisons between the models and look at the following results to choose the one that interests us most.

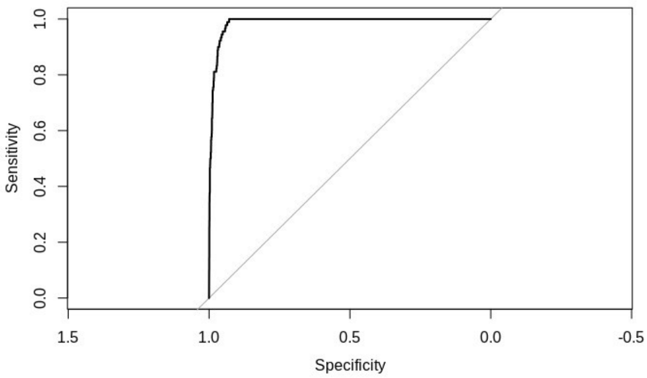
I. LDA

LDA	0	1
0	3422	21
1	47	69
% Hits	98.08935%	
Area under the curve	0.9901	
IC 95%	0.9862-0.994	



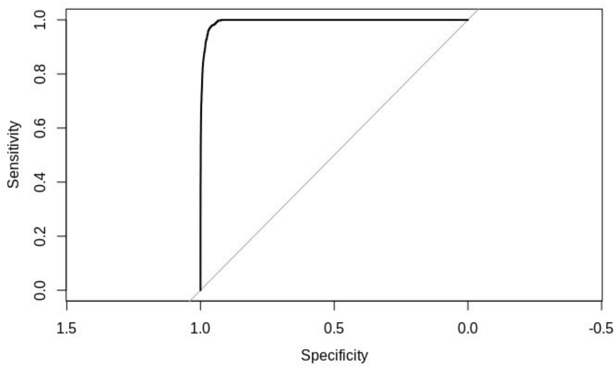
II. QDA

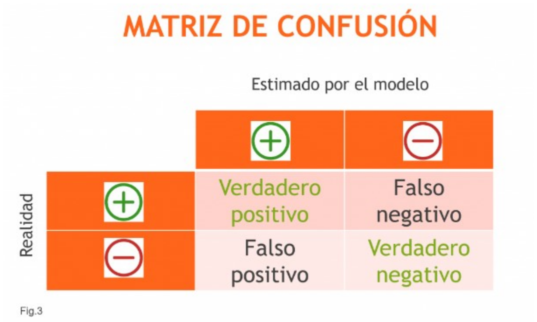
QDA	0	1
0	3296	4
1	173	86
% Hits	95.02669%	
Area under the curve	0.9881	
IC 95%	0.9841-0.992	



III. GLM

GLM	0	1
0	3462	37
1	7	53
% Hits	98.7637 %	
Area under the curve	0.9946	
IC 95%	0.9934-0.9958	





• **We choose the GLM** method as it is the one that best predicts true positives and negatives. On the one hand it is the one that reduces the false positives the most (really what interests us to the mistakes that the model makes), but the false negatives are higher than the other models. We can say that all three models are good, but for the above reasons we will stick with the GLM.

• With the model we have a table with the players who are most likely to be in the quintet. To this table we apply a function in which we take into account the restriction that only certain players can have for each position.

• Finally, we **visualize the results** that we will discuss in point 4.1.RESULTS.

3.2. MODEL CREATION PROCESS WITH THIS YEAR'S DATA

- The process we followed in this case is very similar to the previous one, practically identical. In summary, we do the following steps:

Reading and adapting the data (we change the name of the variables so that they are equal to the other data.frame, change null values to 0…)

We perform a multiple logistic regression with the GLM method. Let's take a step back and choose the variables of interest.

We look again at the different methods of logistic regression to develop the best model. Again we stick with the GLM method. The results obtained are very similar to those of the previous point.

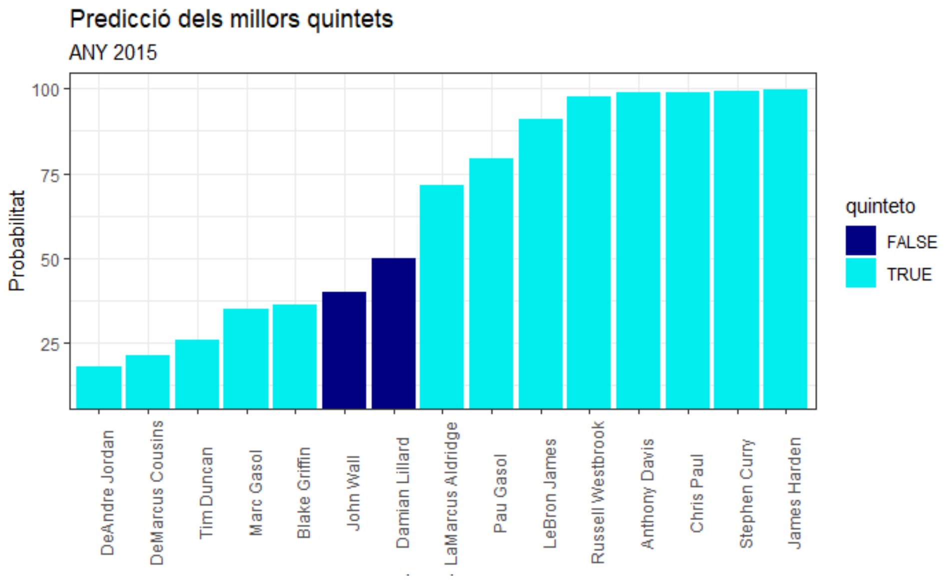
We apply the same function as the one we did before to restrict the positions of the players.

4. RESULTS AND CONCLUSIONS

- At this point, we will analyze the results obtained and draw conclusions from them.

4.1. RESULTATS MODEL ON CONEGUEM LES DADES I CONCLUSIONS

➢ **<u>Prediction 2015:</u>**



Millors jugadors del 2015

Jugador	Edad	Posicio	Probabilitat	Esta en el quintet?
James Harden	25	SG	99.82453	1
Stephen Curry	26	PG	99.46583	1
Chris Paul	29	PG	98.94948	1
Anthony Davis	21	PF	98.68507	1
Russell Westbrook	26	PG	97.67883	1
LeBron James	30	SF	91.07932	1
Pau Gasol	34	PF	79.18571	1
LaMarcus Aldridge	29	PF	71.48565	1
Damian Lillard	24	PG	49.70695	0
John Wall	24	PG	39.83640	0
Blake Griffin	25	PF	36.01446	1
Marc Gasol	30	C	34.85991	1
Tim Duncan	38	C	25.74312	1
DeMarcus Cousins	24	C	21.20210	1
DeAndre Jordan	26	C	17.97916	1

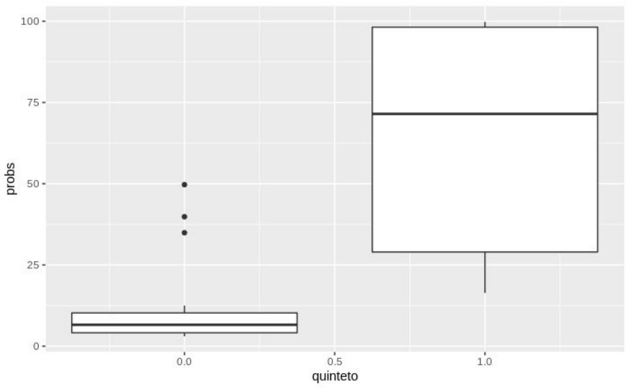
- Checking the model in 2015, we appreciate that the results obtained seem very accurate. Whereas we have a database with many players every season, in this case 650, manages to predict 13 of the 15 players at the ALL NBA TEAM.

Knowing that voting is subjective depending on the player's game, and not on his statistics, we note that our model explains these votes with a very high probability of success.

In the table of substitutions, these are the players who should be in the quintet (Kyrie Irving and Klay Thompson) replacing those who have not been able to correctly predict our model (Jhon Wall and Damian Lillard). We also show the probabilities that our model gives to these players. That way we can learn a little more about our mistakes.

Sustitucions

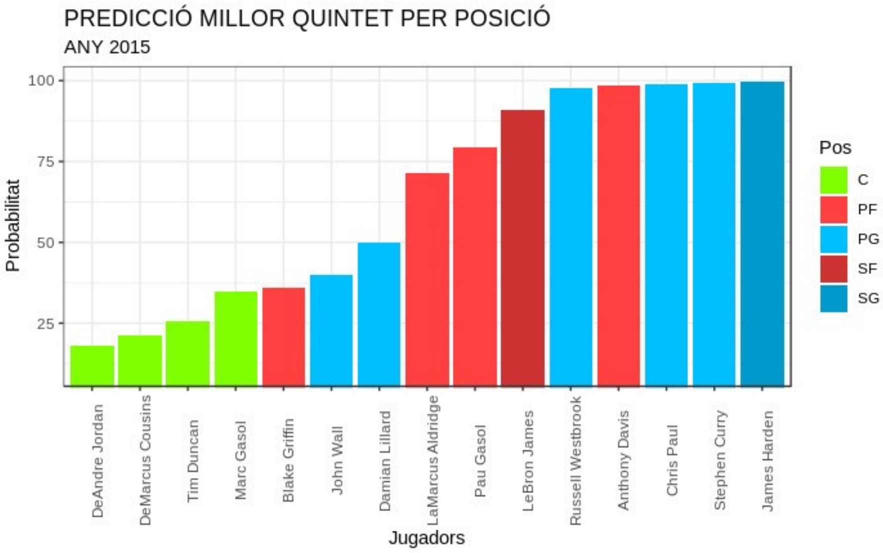
Jugador	Edad	Posicio	Probabilitat	Esta en el quintet?
Damian Lillard	24	PG	49.70695	FALSE
John Wall	24	PG	39.83640	FALSE
Kyrie Irving	22	PG	32.24694	TRUE
Klay Thompson	24	SG	16.39658	TRUE



-To make this boxplot we do it with a database of the 30 players who are most likely to be in the quintet.

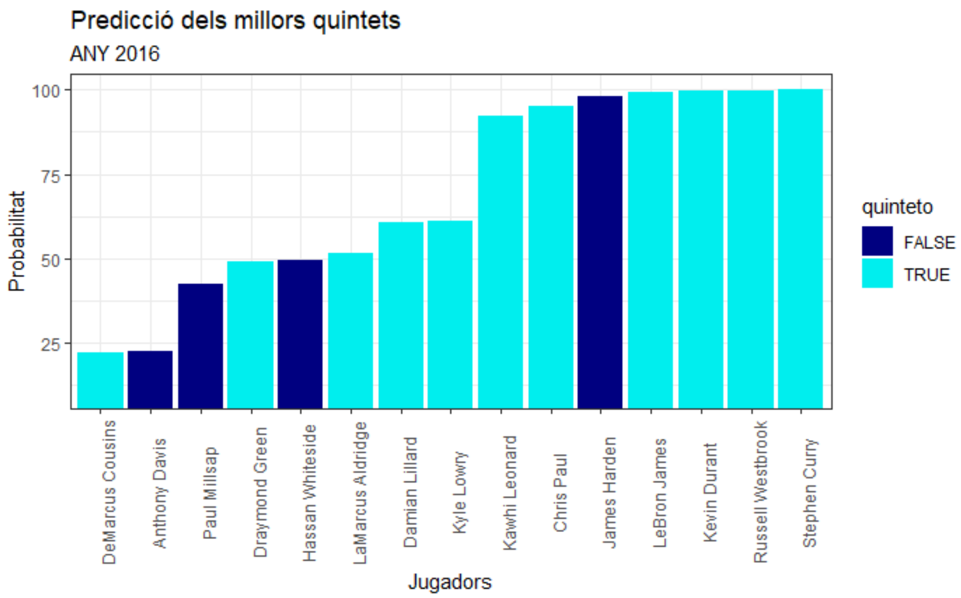
On the right we can see as the median of the players who are going to be in the quintet that year we give a probability of 71.49%, while we give the other players a median of 6.59%.

We note that there are 3 outliers who are the players our model predicts will be. One of them does not enter due to the restriction of positions.



In this graph we can see the division of positions that explains a little more the errors of the model, as John Wall and Damian Lillard (both errors) enter although they have a higher probability than other players in the model, would enter position in the last two places.

➢ Prediction 2016:



Millors jugadors del 2016

Jugador	Edad	Posicio	Probabilitat	Esta en el quintet?
Stephen Curry	27	PG	99.95947	1
Russell Westbrook	27	PG	99.66814	1
Kevin Durant	27	SF	99.59944	1
LeBron James	31	SF	99.27701	1
James Harden	26	SG	97.84001	0
Chris Paul	30	PG	94.96401	1
Kawhi Leonard	24	SF	92.42842	1
Kyle Lowry	29	PG	61.28511	1
Damian Lillard	25	PG	60.52303	1
LaMarcus Aldridge	30	PF	51.54793	1
Hassan Whiteside	26	C	49.35669	0
Draymond Green	25	PF	48.86015	1
Paul Millsap	30	PF	42.52445	0
Anthony Davis	22	C	22.50300	0
DeMarcus Cousins	25	C	22.24952	1

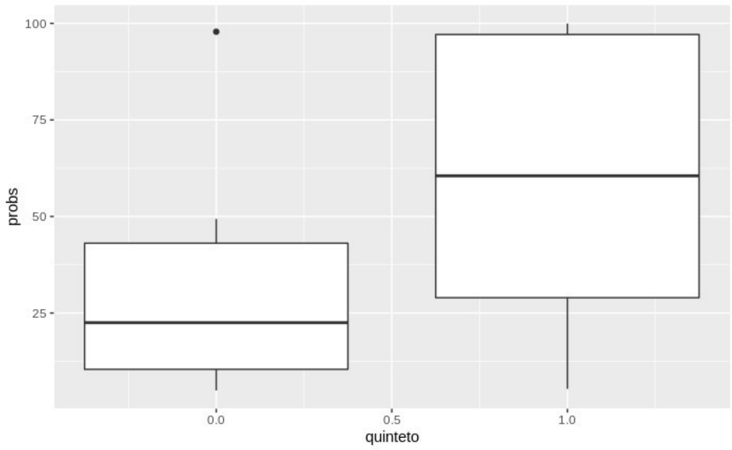
-This year is the year in which we find the most mistakes, especially focusing on the mistake of James Hardem who gives him a 97.8% probability of belonging to the quintets. Researching a bit about the player we realize that he belongs to the quintets from 2013 to 2019 (with the exception of this year) and since 2014 he always appears in the first quintet.

We note that this year is the year in which the player got the fewest wins (a difference of 14 compared to other years), the mistake is because our model does not consider them. We think that this lack of victories influenced the voting. Although his individual statistics were very prominent. This player would have entered the quintet according to the votes in that year in the NBA, but as we know, there is a restriction of positions, which caused him not to enter.

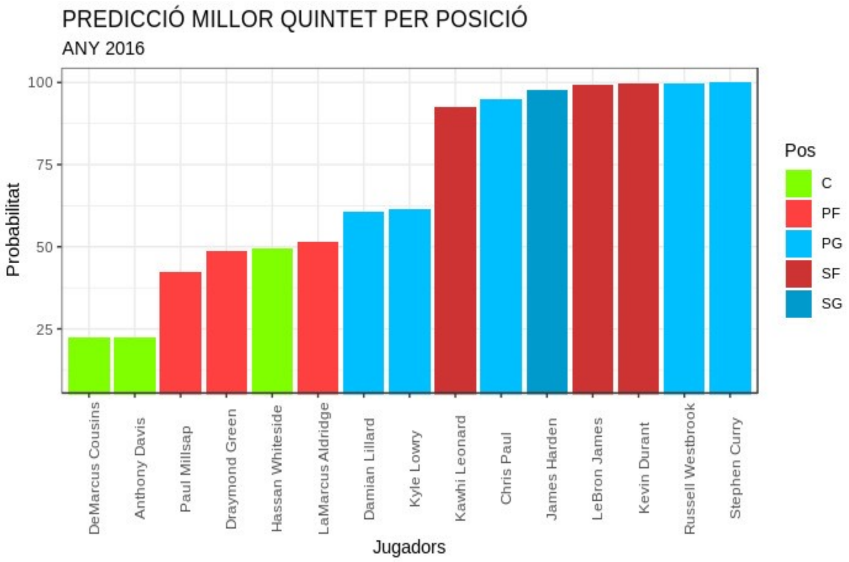
In the table we can find which players have to be in the quintet replacing the model errors.

Sustitucions

Jugador	Edad	Posicio	Probabilitat	Esta en el quintet?
James Harden	26	SG	97.840012	FALSE
Hassan Whiteside	26	C	49.356690	FALSE
Paul Millsap	30	PF	42.524450	FALSE
Anthony Davis	22	C	22.502996	FALSE
Paul George	25	SF	35.686584	TRUE
DeAndre Jordan	27	C	16.846014	TRUE
Andre Drummond	22	C	6.070247	TRUE
Klay Thompson	25	SG	5.361564	TRUE

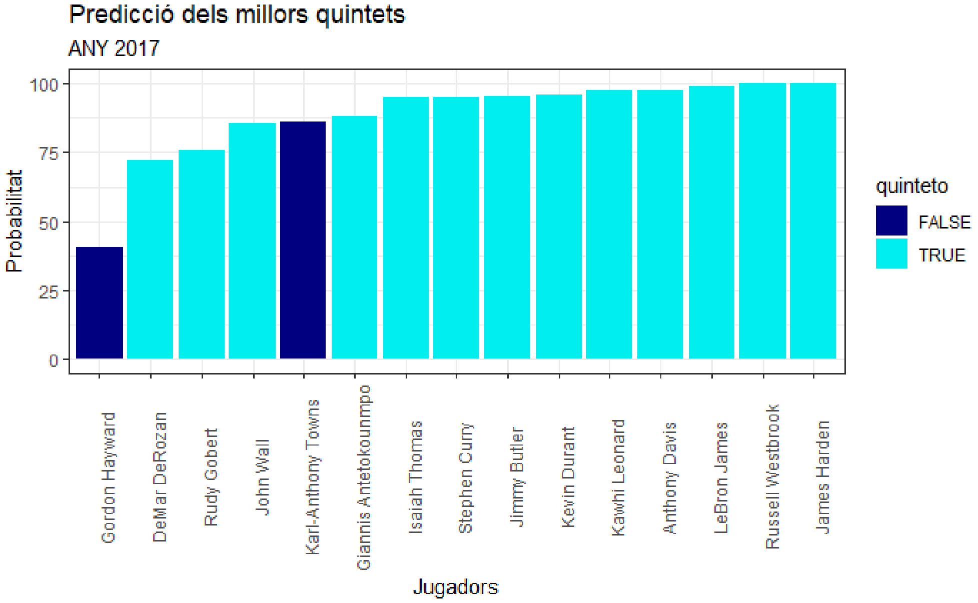


This year we have the highest average of players who are not in the quintet, although we can see that the median probability of players who are in the quintet is twice that of those who are not. Therefore we can consider that we make a good prediction.



As we can see in this model, we already have 6 players in the "guard" position, this fact causes James Harden to not be able to enter this year's quintet.

Prediction 2017:



Millors jugadors del 2017

Jugador	Edad	Posicio	Probabilitat	Esta en el quintet?
James Harden	27	PG	99.90859	1
Russell Westbrook	28	PG	99.90621	1
LeBron James	32	SF	98.82146	1
Anthony Davis	23	C	97.47270	1
Kawhi Leonard	25	SF	97.37003	1
Kevin Durant	28	SF	96.08908	1
Jimmy Butler	27	SF	95.59811	1
Stephen Curry	28	PG	95.06756	1
Isaiah Thomas	27	PG	94.65219	1
Giannis Antetokounmpo	22	SF	88.08650	1
Karl-Anthony Towns	21	C	86.31123	0
John Wall	26	PG	85.78017	1
Rudy Gobert	24	C	75.74142	1
DeMar DeRozan	27	SG	71.95088	1
Gordon Hayward	26	SF	40.50799	0

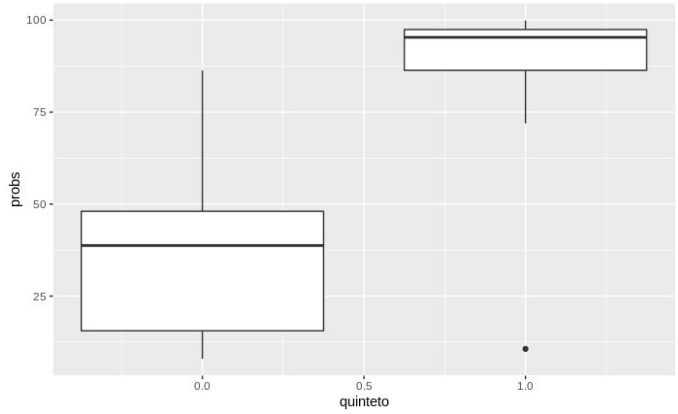
-This year we can consider that there are many players with a very high probability of belonging to the quintet. We note that there are very few errors. There are two mistakes, and they are not in the top 10.

Karl-Anthony Towns of the Minnesota Timberwolves team, has a total of 31 wins and 51 losses. Being these the minimum of victories of all the predicted players. One thing we can also highlight is that this player was 16th in the quintet positions, with 4 points less than Deandre Jordan who came in 15th.

Gordon Hayward that year made the year in best statistics. It was his only year with more than 20 points per game played. It was the only year he was selected for the NBA All Star.

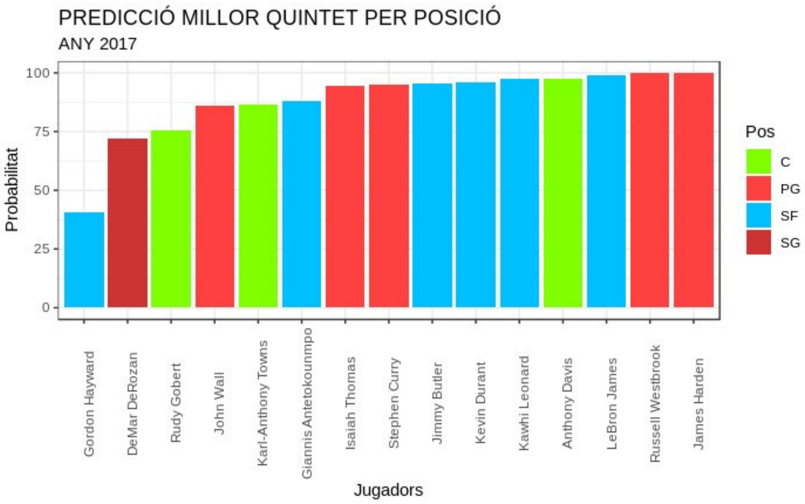
Sustitucions

Jugador	Edad	Posicio	Probabilitat	Esta en el quintet?
Karl-Anthony Towns	21	C	86.311228	FALSE
Gordon Hayward	26	SF	40.507995	FALSE
DeAndre Jordan	28	C	10.635179	TRUE
Draymond Green	26	PF	4.138469	TRUE



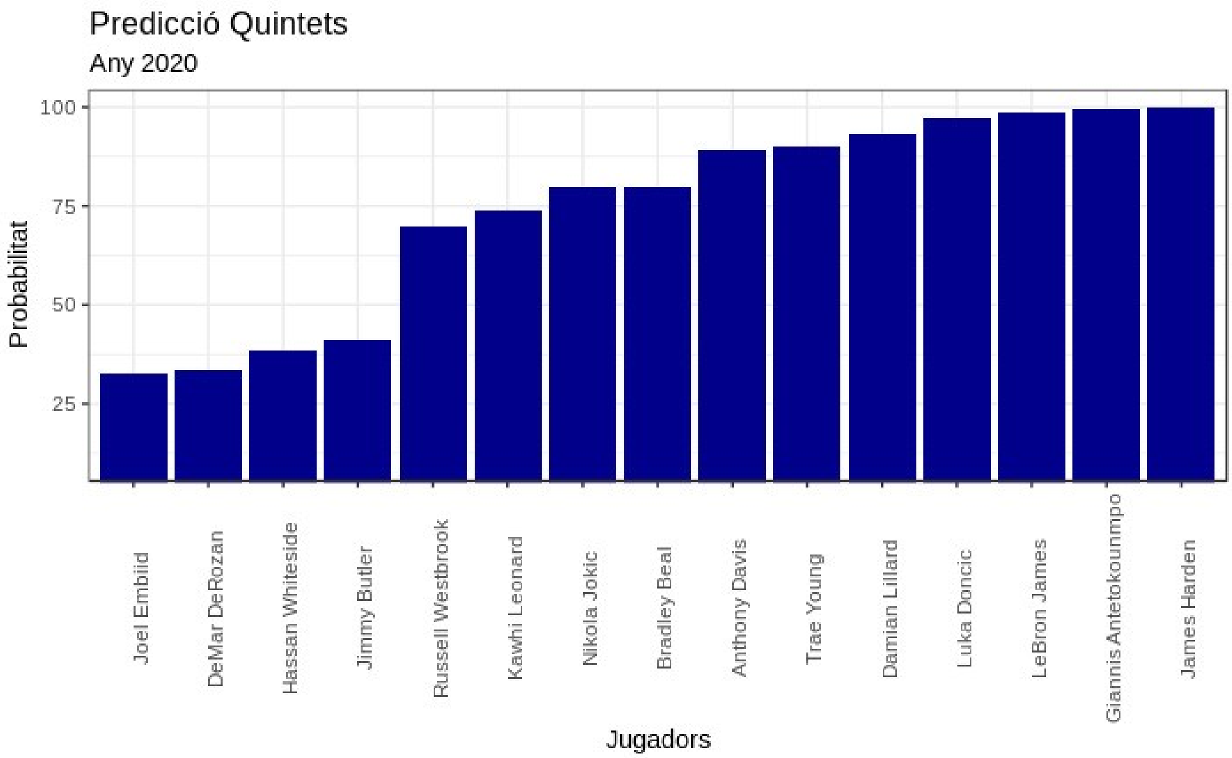
This box diagram shows a big difference between the two groups, a little even more remarkable than the other years. With averages of 8.76% compared to 95.07%

Focusing on the outlier we find, we realize that it is Deandre Jordan with a probability of 10.64% and that he occupies the same position as Karl-Anthony Towns (the mistake of before). Deandre Jordan is a player with a very defensive facet, so he did not have very good statistics, but he has a very good reputation in the league. His team scored 20 more victories this year than the Karl-Anthony Towns team, also entering the playoffs at the top of the table.



In this graph we find it interesting to see how there is no player in the PS position who currently has many changes in the competition. Occupying it to the extent by SF or C players.

4.2. MODEL RESULTS PREDICTING THIS YEAR'S DATA AND CONCLUSIONS



This box diagram shows a big difference between the two groups, a little even more remarkable than the other years. With averages of 8.76% compared to 95.07%

Focusing on the outlier we find, we realize that it is Deandre Jordan with a probability of 10.64% and that he occupies the same position as Karl-Anthony Towns (the mistake of before). Deandre Jordan is a player with a very defensive facet, so he didn't have very good statistics, but he has a very good reputation in the league. His team scored 20 more victories this year than the Karl-Anthony Towns team, also entering the playoffs at the top of the table.

In this graph we find it interesting to see how there is no player in the PS position who currently has many changes in the competition. Occupying it to the extent by SF or C players.

Millors jugadors del 2020

Jugador	Edad	Posicio	Probabilitat
James Harden	30	G	99.82000
Giannis Antetokounmpo	25	F	99.67420
LeBron James	35	F	98.82920
Luka Doncic	21	G-F	97.26112
Damian Lillard	29	G	93.29723
Trae Young	21	G	90.10713
Anthony Davis	27	F-C	89.39499
Bradley Beal	26	G	79.80287
Nikola Jokic	25	C	79.67576
Kawhi Leonard	28	F	74.03270
Russell Westbrook	31	G	69.90656
Jimmy Butler	30	F	41.15219
Hassan Whiteside	30	C	38.69001
DeMar DeRozan	30	F	33.67154
Joel Embiid	25	C-F	32.54220

Puesto	Jugador	Votos
1	Giannis Antetokounmpo	50
2	LeBron James	39
3	James Harden	23
4	Luka Doncic	15
5	Anthony Davis	9

To compare a little more the results of this year, we can see that in the last article published by NBA Spain for the career of the MVP. We note that the 5 players who enter get more than an 89% chance of being in the quintet. The first quintet that can be formed according to the restriction of positions, we would conform it with these 5 positions.

So this year's predictions seem reasonable.

5. FINAL CONCLUISONS

In conclusion, our model obtains a very high reliability when we talk mainly about the 5 players most likely to be in the quintet per year. In players to whom our model gives them lower probabilities, the end results also tend to err a little more. Anyway, we think the reliability that our model gets is very high, as I saw in the results. Making the correct prediction in a total of 37 players out of 45 (82’22%).

As we can see, of these 8 errors, only two fail to give a probability of more than 50%.

In both cases, the team's total victories are significant, as they are not numerous. This is the point where we think our model has found the most mistakes. According to our hypothesis, if we had a database from which to extract this possible variable, our results would significantly improve the study.