Big Data in Sports: Predictive Models for Basketball Player's Performance

Dae-Jin Lee [1]†,and Garritt L. Page [2]

[1] *Basque Center for Applied Mathematics, Bilbao, Spain*
[2] *Department of Statistics, Brigham Young University, Provo UT, USA*

(*Communicated to* MIIR *on 14 April 2021*)

Study Group: ESGI 131. 15–19 May 2017, Bilbao, Spain.

Communicated by: Tim Myers

Industrial Partner: Xpheres Basketball Management/Aryuna

Team Members: Amaia Abanda Elustondo, BCAM, Bilbao; Bruno Flores Barrio, University of La Rioja; Silvia García de Garayo Díaz, UPV/EHU – University of the Basque Country; Manuel Higueras Hernáez, University of La Rioja; Amaia Iparragirre Letamendia, UPV/EHU – University of the Basque Country; Mariam Kamal, UPV/EHU – University of the Basque Country; Gorka Labata Lezaun, University of Zaragoza; Roi Naveiro Flores, ICMAT – Institute of Mathematical Sciences, Madrid; Argyrios Petras, BCAM, Bilbao; Simón Rodríguez Santana, ICMAT - Institute of Mathematical Sciences, Madrid; Quan Wu, UPV/EHU – University of the Basque Country.

Industrial Sector: Sports

Tools: Statistical Modelling, Predictive Models, Clustering, Regression Techniques

Key Words: Basketball, Sports Data Analytics, Big Data

MSC2020 Codes: 62

Summary

Aryuna is a platform that allows to perform advanced data analytics of men's professional basketball statistics of the last 16 seasons in more than 25 professional leagues and 71 FIBA tournaments. The complete database consists of more than 37,000 games and upwards of 20,000 players. Based on a historical database, the report aims to: characterize the performance curve, peak and optimal age in professional men's basketball using performance ratings of players in top basketball leagues; determine a rating correction factor for different basketball leagues, which accounts for intra-league and cross-league variability as well as for player characteristics (position, age, player ratings, etc.); determine which are the most important factors for predicting future outcomes of a basketball player.

† Corresponding Author: dlee@bcamath.org

Challenge 4 – Big Data in Sports: Predictive Models for Basketball Player's Performance

June 2, 2017

Keywords | Basketball; sport data analytics; statistical modelling, predictive models, clustering, regression techniques.

AMS classification | 62-07 (Data Analysis)

1. Introduction and challenge description

Data analytics in professional sports has experienced rapid growth in recent years[1]. Development of predictive tools and techniques began to better measure both player and team performance. Statistics in basketball, for example, evaluate a player's and/or a team's performance[2].

Aryuna© is a platform that allows to perform advanced data analytics of men's professional basketball statistics of the last 16 seasons in more than 25 professional leagues and 71 FIBA tournaments. The complete database consists of more than 37,000 games and upwards of 20,000 players. Based on a historical database, the aim of the challenge is to:

- Characterize the performance curve, peak and optimal age in professional men's basketball using performance ratings of players in top basketball leagues. See [6, 3]
- Determine a rating correction factor for different basketball leagues, which accounts for intra-league and cross-league variability as well as for player characteristics (position, age, player ratings, etc.). See [1]
- Determine which are the most important factors for predicting future outcomes (a successful professional career) of a basketball player. See [2, 6]

[1] `https://en.wikipedia.org/wiki/APBRmetrics#Common_statistics`
[2] `https://en.wikipedia.org/wiki/Basketball_statistics`

- Study statistical models to evaluate the performance of a player based on position, age, skills, league and other characteristics, and their influence in the game. See [5]

2. Data description

The file `Aryuna_sample_DB.csv` contains a sample of the full database. The number of players are 5227, for 16 Seasons (2000-2015), and 6 competitions (Euroliga, Eurocup, ACB, Argentina, ABA, ProA). The total number of variables in the database is 44.

Some important concepts:

- **Variables with suffix 'X100Possessions':** are the statistics produced by the player per 100 Team's possessions, e.g. *'Ptsx100Possessions = 20,5'* means the player scores 20,5 points per 100 Team's possessions.
- **Possession:** In basketball, possessions are defined as the time a team gains offensive possession of the ball until it scores, loses the ball or commits a violation or foul.
- **Usage%:** Usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor. The formula is :

$$100 * ((\text{FieldGoals Attempts} + (0{,}44 * \text{FreeThrows Attempts}) + \text{Turnovers}) * \text{Team minutes}) / (\text{Player minutes} * \text{Team FieldGoals Attempts} + 0{,}44 * \text{Team FreeThrows Attempts} + \text{Team Turnovers})$$

- **Free throw:** are unopposed attempts to score points from a restricted area on the court (the free throw line; informally known as the foul line or the charity stripe), and are generally awarded after a foul on the shooter by the opposing team.
- **Field Goals:** it refers to a basket scored on any shot or tap other than a free throw, worth two or three points depending on the distance of the attempt from the basket.
- **BPM (Box Plus Minus):** is a box score-based metric for evaluating basketball players' quality and contribution to the team. `http://www.basketball-reference.com/about/bpm.html`

There are also some definitions that measure the player performance (called *metrics*).

- **EOPx40M(Efficient Offensive Production per 40 Min):** To calculate EOPx40M we need to get OE (Offensive Efficiency).

- **OE (Offensive Efficiency):** Metric that measures the quality of offensive production. An OE of 1.0 correspond to 100 percent efficiency. OE is the total number of successful offensive possessions the player was directly involved in divided by that player's total number of potential ends of possessions. OE formula:

$$(FieldGoalsMade + Assists) / (FieldGoalsAttempts - OffensiveReb + Assists + Turnovers)$$

In order to compute **EOPx40M**, first we calculate EOP, this metric measures the offensive production with a measure of efficiency (OE), it uses points and assists. To use assists in the formula, it needs to know the value of an assist relative to a point scored, the author consider an assist meaningful if the assist led to a basket at the rim. By average about 38% of assists led to a basket at the rim, if a player had 100 assists, he created 2*0,38*100= 76 points, so 1 assists = 0,76 points. EOP Formula:

$$EOP = (0{,}76 * Assists + Points) * OE$$

$$EOPx40M = (EOP / Player\ seconds\ played) * (40 * 60)$$

Note: OE y EOP are metrics created by Stephen M. Shea in his book Basketball Analytics [8].

3. Objectives

Based on the data collected from **Aryuna©**, the team assigned to this challenge worked in small groups and focused in 4 main goals:

1. Study the performance curve, peak and optimal age of a basketball player in top European leagues.

2. Player's performance and their influence in the game.

3. Rating correction factor for different basketball leagues.

4. Which factors predicts a successful professional career?

Goal 1: Performance of players by age and experience

In order to analyse the performance curves, we chose the variable EOP per 40 minutes (EOPx40M) for each player in one season and one competition. We fitted a mixed-effects model with a quadratic fixed effect for the age of the player interacting with the position. There is also a random effect for each player, to account the variability of each individual. Thus the model is

$$\text{EOPx40M} = \beta_0 \text{Position} + \beta_1 \text{Position:Age} + \beta_2 \text{Position:Age}^2 + u_{\text{Player}} + \epsilon, \tag{1}$$

where $u_{\text{Player}} \sim N(0, \sigma_{\text{Player}})$ is a random effect per player and ϵ is the error term, i.e. $\epsilon \sim N(0, \sigma)$.

Registers of players whose minutes played at the season-competition are lower or equal to 50 minutes are discarded. The data for this model include 10,712 observations for 3,743 different players.

The estimation of the standard deviation of the random effect is $\hat{\sigma}_{\text{Player}} = 2.08$ with p-value < 0.0001, indicating a significant variability for the players.

The fixed effects of this model produce three performance curves, one for each position. The peaks for each position are calculated by calculating their maximums. Let the curve of performance for position p be

$$y = \beta_{0,p} + \beta_{1,p} x + \beta_{2,p} x^2,$$

the derivative is

$$y' = \beta_{1,p} + 2\beta_{2,p} x,$$

and the maximum of the curve (if $\beta_{2,p} < 0$, because $y'' = 2\beta_{2,p}$) is found for the value which solves the derivative equals to 0,

$$\beta_{1,p} + 2\beta_{2,p} x_{\max} = 0 \Rightarrow x_{\max} = -\frac{\beta_{1,p}}{2\beta_{2,p}}.$$

For each position p, $-\hat{\beta}_{1,p}/(2\hat{\beta}_{2,p})$ is the peak performance age and by the δ-method the 95% confidence intervals bound limits are

$$-\frac{\hat{\beta}_{1,p}}{2\hat{\beta}_{2,p}} \pm 1.96 \cdot \left(-\frac{1}{2\hat{\beta}_{2,p}}, \frac{\hat{\beta}_{1,p}}{2\hat{\beta}_{2,p}^2}\right) \cdot \Sigma_{12,p} \cdot \left(-\frac{1}{2\hat{\beta}_{2,p}}, \frac{\hat{\beta}_{1,p}}{2\hat{\beta}_{2,p}^2}\right)^{\mathrm{T}}$$

where $\Sigma_{12,p}$ is the variance covariance matrix of $\beta_{1,p}$ and $\beta_{2,p}$.

Peak performance ages (and their 95% confidence intervals) by position are:

- Center: 27.23 (26.48, 27.98) years;
- Forward: 27.46 (27.00, 27.92) years;

- Guard: 28.08 (27.57, 28.59) years.

These peaks and their statistical uncertainties are between 26 and 29, which are similar to those for the NBA, 27-30, as shown in [6].

Analogously, curves of performance by the experience of the player interacting with the position have been calculated, with a random effect for each player. As there is no information of the debut season of the registered players, it is assumed that new players from the 2001 season younger than 20 are in their first season. The model has the same structure changing the age of the player by the his experience (time since his debut season). Registers of players whose time contributed in the season-competition is lower or equal to 50 minutes are discarded. The data in this model are 986 observations for 283 different players.

Peak performance experiences (and their 95% confidence intervals) by position are:

- Center: 10.63 (2.37, 18.88) years;
- Forward: 7.35 (5.76, 8.93) years;
- Guard: 9.46 (6.38, 12.55) years.

Peak performance experiences are between 7 and 11 years, which are similar to those for the NBA (6-8 years, see [6]). The 95% confidence intervals are huge (mainly for the center players) because of the lack of information (121 centers, 364 forwards and 501 guards). Information about the debut season for each player would give bigger dataset to be analysed in this model.

Figure 1 shows the plot of the fixed effect curves by position for both models.

Extraction of Performance Patterns

In order to deal with individual performance trajectories, we split the database according to different competitions and once a competition was selected, we redimensionalized the database as follows: each row d_i represents a player who played in this competition and each column d_j refers to an age. Each element d_{ij} of the new matrix represents the EOP performance index of the player i by the age j.

First of all, it should be mentioned that there are quite a few gaps or undetermined values in the performance matrix due to the lack of data or the restriction to one competition. For example, if a player has been in the ACB for 5 season, then he moved to the NBA for 4 season and finally he came up for 3 season to the ACB, there would be a 4 years long gap in the ACB-Performance matrix of this player.

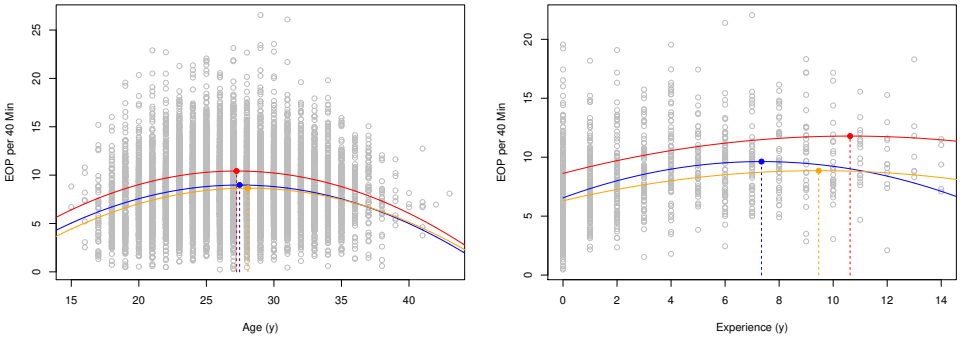

Figure 1: Performance curves by age (left) and experience (right). The grey dots are the observed performances. The solid lines are the performance curves for centers, forwards and guards, and their respective peaks are represented by the solid dots and the dashed lines.

$$
\begin{array}{c}
 \\
P_1 \\
P_2 \\
P_3 \\
\dots \\
P_n
\end{array}
\begin{array}{c}
\begin{array}{cccccc} AGE_{17} & AGE_{18} & \cdots & AGE_{42} & AGE_{43} \end{array} \\
\begin{pmatrix}
\emptyset & \emptyset & \cdots & EOP_{P_1,42} & EOP_{P_1,43} \\
EOP_{P_2,17} & EOP_{P_2,18} & \cdots & \emptyset & \emptyset \\
\emptyset & \emptyset & \cdots & EOP_{P_3,42} & EOP_{P_3,43} \\
 & & \cdots & & \\
 & & & &
\end{pmatrix}
\end{array}
$$

Figure 2: New Performance Matrix

PlayerID	17	18	19	20	21	22	23	24	25	26	27	28	29
118	NA	2.020	NA	8.87	11.80	9.640	9.440	8.160	9.940	11.890	15.510	11.740	NA
121	NA	NA	NA	12.13	10.58	11.470	10.080	13.670	14.420	13.500	NA	13.000	14.130
276	NA	10.590	11.840	12.76	14.03	19.550	NA	NA	NA	16.160	12.580	13.710	10.840
280	NA	NA	NA	15.55	13.98	16.250	14.340	12.050	12.070	13.810	17.150	17.890	12.560
284	NA	8.890	NA	9.52	10.71	13.560	10.710	12.130	15.550	13.660	12.480	13.550	NA
325	NA	11.860	7.640	9.31	8.24	8.950	12.190	9.680	7.580	8.250	7.850	7.780	6.790
326	NA	NA	NA	5.06	10.18	10.520	8.750	11.230	9.000	7.470	9.700	7.740	10.190
329	NA	9.010	3.330	9.38	19.55	9.250	10.230	9.540	21.400	12.630	10.700	11.540	11.370
485	NA	NA	NA	NA	9.96	9.100	8.860	8.800	10.080	9.350	8.980	10.110	8.380
486	NA	NA	NA	NA	7.59	9.980	10.650	12.890	10.320	6.625	9.710	6.850	10.840
489	NA	NA	NA	9.60	8.84	NA	NA	NA	11.010	10.070	11.120	9.560	NA
559	NA	NA	NA	NA	NA	NA	NA	11.140	8.980	11.090	10.060	8.880	12.270
560	NA	NA	NA	NA	NA	NA	NA	NA	10.710	9.040	12.560	9.030	11.110
721	NA	NA	NA	NA	NA	NA	NA	NA	9.430	10.720	7.070	10.600	9.680
830	NA	NA	NA	13.60	NA	8.510	10.570	10.925	12.540	13.540	9.490	16.440	10.690
837	NA	NA	NA	NA	6.36	8.480	5.670	NA	12.110	8.020	11.400	11.390	9.130

Figure 3: Example of Performance Matrix of ACB

Therefore, each row on the matrix corresponds to the performance curve of a player in one competition. In Figure 3 some examples of ACB player's performance curves are shown. We only have the points so we tried to do a regression (find the curve that describes those points) with polynomials approximations of different grades.

Nevertheless, this approximations were not good enough and we wanted to go further. If one looks at A. Hervelle and C. Jimenez player's performance curves (the real data, the points in Figure 3), it can be seen that their pattern is very similar. Consequently, we noticed that it make sense to exists different patterns on the performance curves.

In order to extract meaningful patterns, we took advantage of the fact

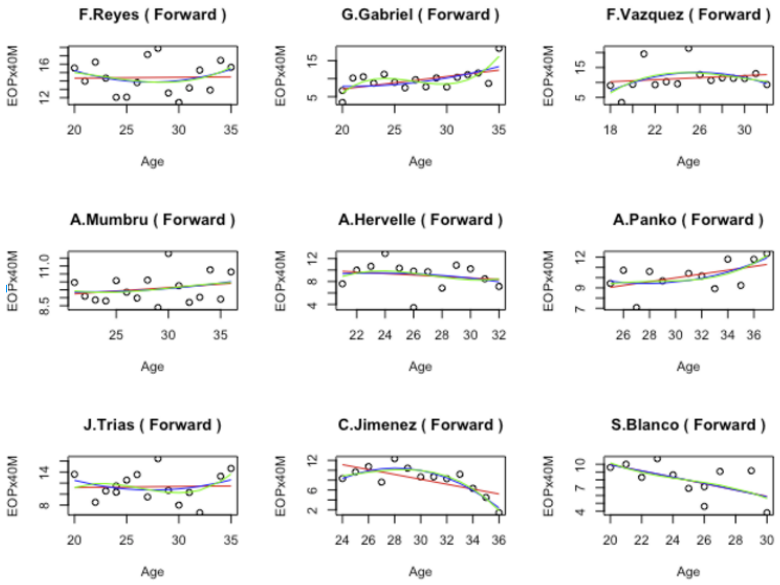

Figure 4: Some performance curves

that this performance curve are, actually, time series. That is, each value has been taken by a specific age and the values are, some kind, sorted. In other words, the values on an individual performance curve can not be considered independently, because they are correlated to each other and this is the main characteristic of the time series. Once this contemplated, the goal is to group the performance curve by their patterns, for which a clustering method was used. Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other clusters, so the aim is to find groups of patterns in the performance curves.

With the intention of clustering the performance curves, a similarity measure between series has to be chosen. Since our interest focuses on the shape of the performance curve, a distance between curve that deals with time warping and shifting was used, namely, Dynamic Time Warping (DTW). DTW is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. It is able to deal with local warping and shifting by searching the optimal alignment between two series that minimizes the distance, so if two series have the same shape but they are out of phase, Dynamic Time Warping will align them and compute the minimal distance between all possible alignments. Figure 5, illustrates the DTW idea.

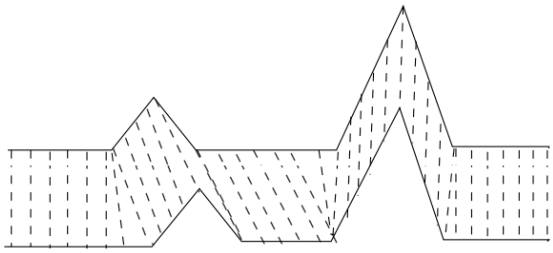

Figure 5: Dynamic Time Warping

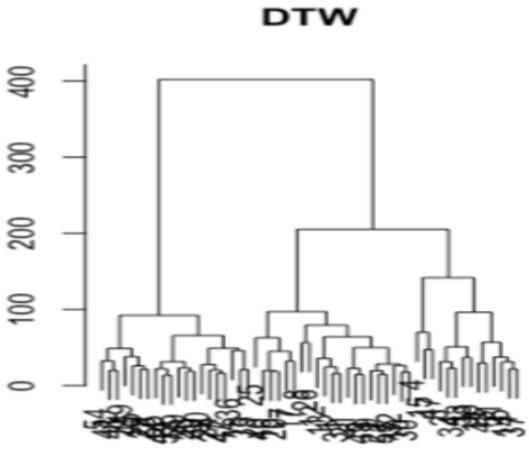

Figure 6: Cluster Dendogram

Therefore, a hierarchical cluster analysis with Dynamic Time Warping was carried out, with the resulting agglomeration dendogram shown in Figure 6. It illustrates the arrangement of the clusters based on the Dynamic Time Warping distance between performance curves, distinguishing 2 or maybe 3 main clusters. We chose to split de database in 3 clusters, but the same analysis could be made for 2 clusters.

The meaning of this partition is that the cluster analysis show that, based on the similarity of the shape of the performance curves, there are 3 main groups. Once the three clusters are divided, we made a regression for each position within each group. The resulting regressions for ACB performance curves are shown in Figure 7. It can be seen the partition on three clusters (each plot shows one cluster) and, in fact, the different patterns for each cluster.

In the first one, the pattern of the performance curve is very similar for Forwards and Guards. Actually, the performance pattern starts with a rela-

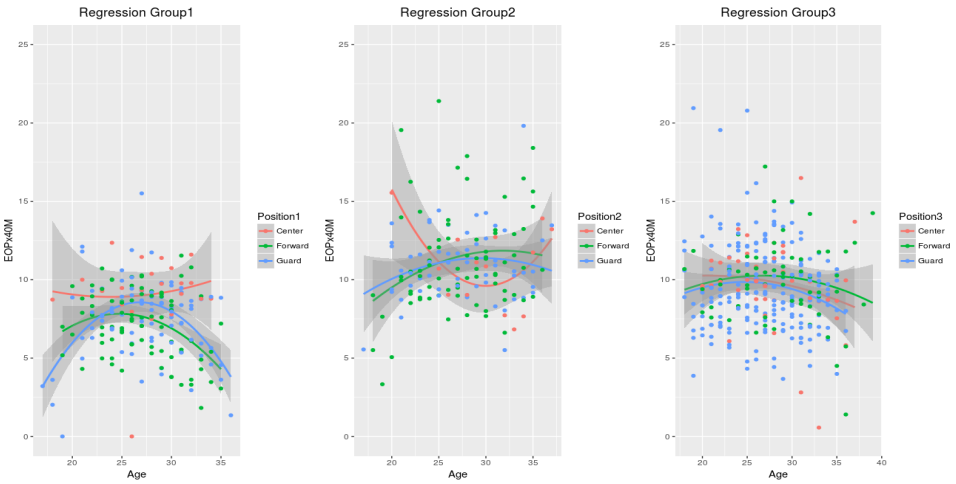

Figure 7: Regression of Clusters by Position (ACB)

tively low EOP in early ages, the players enhance their EOP quickly in the first 5-10 years of competition and reach their performance peak by the age of 25 for Forwards and 27-28 for Guards. After the peak, this kind of players tend to worsen the EOP. It has to be mentioned that for Centers this analysis is not meaningful, due to the lack of data (the error interval coloured by gray is too wide to conclude anything).

The second kind of player (plot in the middle), has a very different pattern of performance. The players in this cluster start with a quite high EOP from a very early ages (as high as the maximum of the player in the first cluster) and they tend to improve, or increase their EOP, over almost their entire career. The performance peak of this kind of player is reached at the age of 31 for Forwards and 29 for Guards. As happened before, there is not enough information to conclude anything for Centers.

The last kind of player starts very similar to the second one, with a quite high EOP in the first years of competition, but instead of keep improving the performance over the years, they reach a peak and start getting worse. For this kind of player, depending on the position, the peak is reached at 27 for Forwards, 24 for Guards. Again, the lack of data does not allow to conclude anything for Centers.

Goal 2: Player's performance and their influence in the game

For this goal, we considered as player's performance measure the shot precision. Figure 8 shows a boxplot per Competition (for all the seasons) with the kernel

density estimator on each side (this is known as *violin plot*). It is shown that the shot precision in all competitions look very similar in 2 points shots (top plots) and very symmetric around their respective averages. The distribution of the shoots for 3 points and free throws are very skewed.

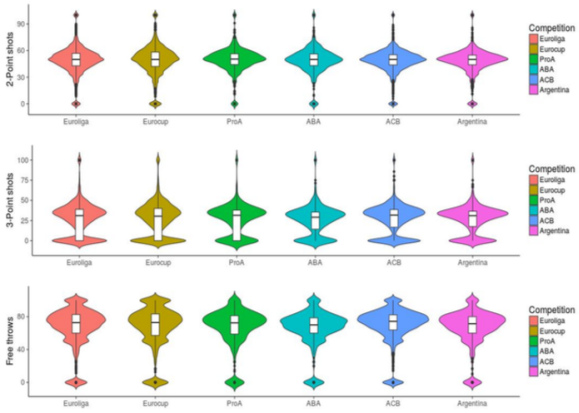

Figure 8: Shot precision by Competition and type of shoot (free, 2 points and points).

Figure 9 shows by player position and age, the increase in % shots in Argentina and Euroleague. There seems to be an increasing percentage in free throws in all the positions. For the rest of the points (2 and 3 points throws) the increase is not so evident due to the high variability of the data.

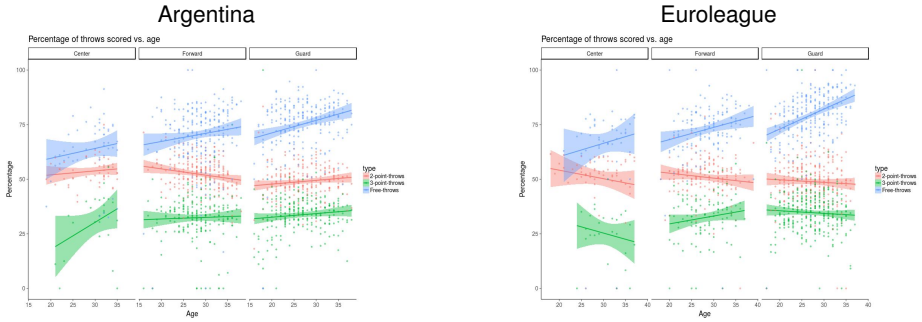

Figure 9: Increase in % shots in Argentina and Euroleague by player position, type of shoot and age.

The conclusions of this goal are not very strong: i) most results are (still) inconclusive due to the need of long-career-players? data; ii) in some rare cases it can be shown that the precision does experience minor changes; and iii) some leagues need more data to obtain better results (if any at all) and finally iv) the nature of each league might be intrinsically different.

Goal 3: Rating correction factor for different basketball leagues

Our second goal was to try to find a conversion factor between the 6 different leagues. As each league has its own characteristics, (some might be more offensive whereas others may use more 3 points shots, for example). Hence, an interesting goal is the need of finding an algorithm that transforms the performance obtained by a player in one league into the performance in another league. This would let them compare different players from different leagues to find the best ones to represent. Interpretation of the results will be left to experts, so we will just focus on the data.

The biggest problem here is to find differences between leagues. Due to the heterogeneity of players, all leagues look similar, as it can be seen in Figure 10. This happens not only for the EOP coefficient but for any other one too. It cannot be seen any difference in terms of mean, variance of distribution that cannot be attributed to noise in the observations.

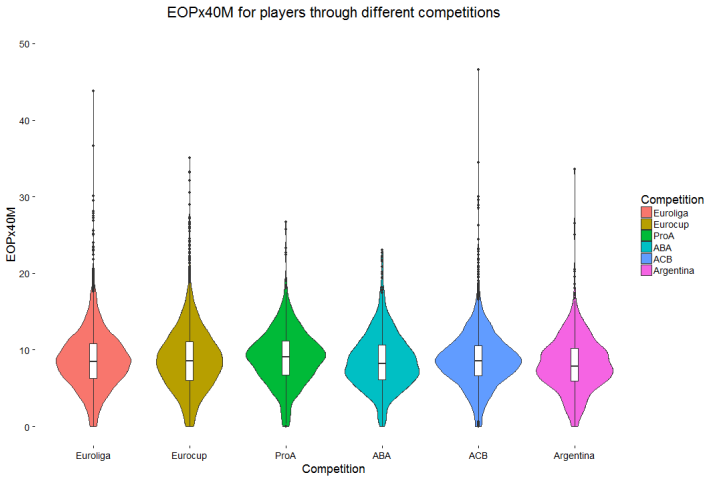

Figure 10: Violin plots obtained for de EOP factor dividing by league.

However, by computing ANOVA (*Analysis Of Variance*) models, we can get some differences for some particular factors and competitions. Using the Euroleague as the base league to compare with, we can see how some means are further than the sum of the standard deviation of both, and all with small p-values, what would imply that there are differences between those means (see Figure 11).

As performance is an abstract concept, many different measures can be built. Indeed, EOP is a measure of offensive performance generated by combination of some basic variables so it might not consider players with high block skills. To consider all different measures, what we propose to do is to predict

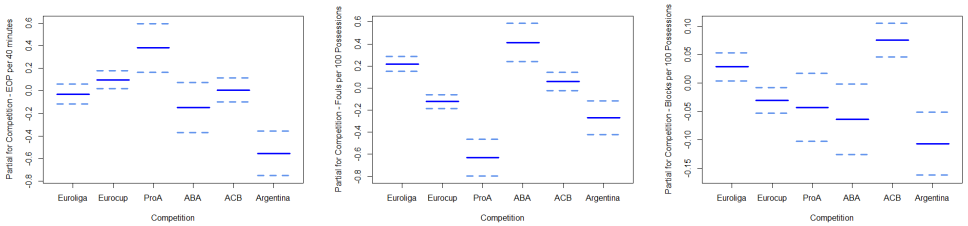

Figure 11: Termplots obtained for EOP (left), Fouls ×100 possessions (center) and Blocks per hundred possessions(right) ANOVA models. The intercept is not taken into account. Continuous lines represent means whereas discontinuous lines represent mean plus or minus standard deviation.

over those basic variables, so that when a different measure is used, the agent just has to apply the formula of the measure to the predictions of the basic variables it uses.

A complete and very visual table with the differences between means for each basic variable and taking into account the p-values of the ANOVA model can be generated so it will help the agent (see Figure 12).

	Points x 100 Possessions			3 Points Scored x 100 Possessions			3 Points Attempted x 100 Possessions			Free Throws Scored x 100 Possessions			Free Throws Attempted x 100 Possessions			Offensive Rebounds x 100 Possessions		
	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value
Intercept	19,97	0,1	2,00E-16	1,79	0,02	2,00E-16	5,32	0,06	2,00E-16	3,95	0,03	2,00E-16	5,52	0,04	2,00E-16	2,61	0,03	2,00E-16
EUROCUP	-0,06	0,14	0,6795	0,08	0,03	0,00844	0,22	0,08	0,0048	-0,12	0,05	0,0107	-0,18	0,06	0,002924	-0,09	0,04	0,0175
ProA	-0,02	0,23	0,9335	0,07	0,05	0,16838	0,33	0,13	0,0132	-0,67	0,08	2,00E-16	-0,85	0,1	2,00E-16	0,11	0,07	0,1165
ABA	-0,48	0,24	0,0449	-0,05	0,06	0,35107	0,32	0,14	0,0206	-0,17	0,08	0,0278	-0,1	0,1	0,34	-0,07	0,07	0,9232
ACB	0,48	0,15	0,0016	0,15	0,035	1,92E-05	0,39	0,09	5,38E-06	0	0,05	0,97	-0,1	0,07	0,1325	0,09	0,04	0,0346
Argentina	-0,38	0,22	0,0798	0,05	0,05	0,33844	0,51	0,12	4,07E-05	-0,34	0,07	2,40E-04	-0,34	0,1	0,000408	-0,26	0,06	3,61E-05
	Defensive Rebounds x 100 Possessions			Steals x 100 Possessions			Turnovers x 100 Possessions			Assists x 100 Possessions			Blocks x 100 Possessions			Fouls x 100 Possessions		
	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value
Intercept	5,95	0,04	2,00E-16	2,03	0,02	2,00E-16	3,79	0,026	2,00E-16	3,47	0,04	2,00E-16	0,74	0,01	2,00E-16	6,99	0,04	2,00E-16
EUROCUP	-0,04	0,05	0,41815	0,01	0,02	0,562	0,05	0,03	0,128	0,14	0,05	0,00509	-0,06	0,02	0,00298	-0,34	0,06	7,65E-10
ProA	0,32	0,09	0,00068	-0,06	0,04	0,116	-0,06	0,06	0,314	0,66	0,08	1,18E-14	-0,07	0,03417	0,0367	-0,85	0,1	2,00E-16
ABA	0,12	0,1	0,20468	-0,17	0,04	5,10E-05	0,08	0,06	0,182	0,45	0,09	2,77E-07	-0,09	0,04	0,00868	0,19	0,1	0,04665
ACB	-0,07	0,06	0,26799	0,11	0,03	3,91E-05	0,02	0,04	0,631	-0,04	0,06	0,42644	0,05	0,02	0,04	-0,16	0,06	0,00995
Argentina	0,5	0,09	1,20E-08	-0,26	0,04	3,36E-12	-0,44	0,06	7,26E-15	-0,64	0,08	1,72E-15	-0,14	0,03	2,58E-05	-0,49	0,09	4,96E-08

Figure 12: Coefficients of the ANOVA model for each of the basic variables. Darker colours imply smaller p-values ranking from smaller than 0.001 to smaller to 0.01, to 0.05 and to 0.1.

To be more accurate, we can divide players into different categories. A good start could be to divide by positions as it is not the same the number of 3 points shots a center can score than this number of scores for a forward or guard. We used an ANOVA model with interactions where now dummies variables for belonging to a league and playing in a certain position are added.

Taking into account all player characteristics doesn't have a big computational impact. Although there are just 3 variables to characterize players (age, height and position), some of them can take too many values. Instead of giv-

ing a conversion factor for each value, that would give an enormous amount of different players, we use a linear model tree. Since ANOVA models are linear models with factors, what we are really doing is to fit an ANOVA model at each leave of the tree. Different values for the minimum number of individuals in each leave allow to adjust the number of types of players.

Following the linear model approach, we performed a linear regression tree, allowing for a recursive partition of the variables. Figure 13 shows the obtained tree.

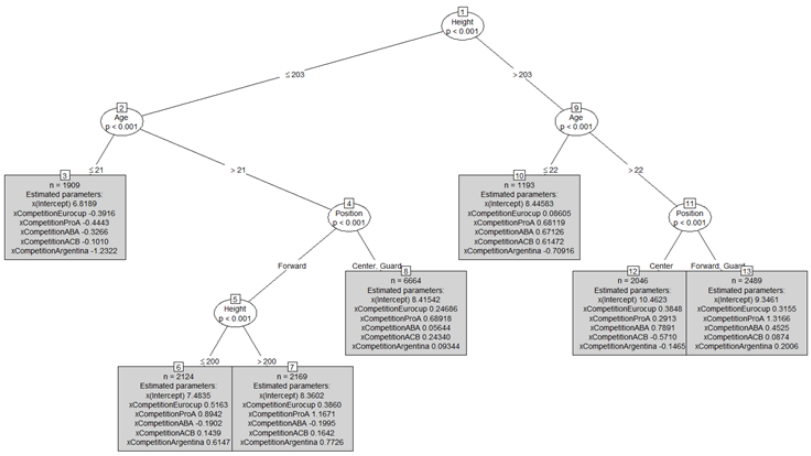

Figure 13: Example of linear model tree to predict the EOP variable adjusted to produce 7 types of players.

The benefits of this last model are that produces good fitting of the data and meantime gives information of the classes of players. All this rules of transformation can be traduced into factors by adding the coefficient in the ANOVA model to the intercept and dividing by the intercept. Thus, a player with a coefficient of 9 in the ACB in a category in which the mean of the Euroleague is 10 and the coefficient of the ANOVA model for the ACB is -0.5 would get a coefficient of:

$$9 \cdot \frac{10}{10 - 0.5} = 9 \cdot 1.05 = 9.4$$

Other ways to obtain this correction factor could be to just consider players that played at two leagues at the same time (one of them should be Euroleague or Eurocup), but we would not have a conversion factor for Argentina league. We could also consider players who were transferred to other leagues. We could observe the level of a player in one league just before leaving and compare it with the level in the other league a few years later, when we think he got used to the new league.

Goal 4: Which factors predicts a successful professional career?

This is the most challenging task. In fact, from our view there is no relevant information in the provided database to quantify a successful professional career. From a basketball point of view, 'success' could be considered as winning a MVP (most valuable player) trophy at least 3 times in a row during a professional career, signing a contract of several millions of $ etc ... However, we tried to answer the following question: "how can we predict success?".

Any statistical model has to be interpretable and have a strong predictive power. Firstly, we considered the variable *EOPx40M* and transformed this variables in a dichotomous variables (0=Fail and 1=Success). Note that this partition is completely subjective, and was performed only to obtain a measure of success. We performed a popular machine learning technique in order to perform a feature (or variable) selection, called *random forest*. Figure 14, shows the number of most important variables for prediction selected by the random forests, i.e. there are 13 out of the 44 variables that are the most relevant variables to predict a successful player based on the transformation we proposed.

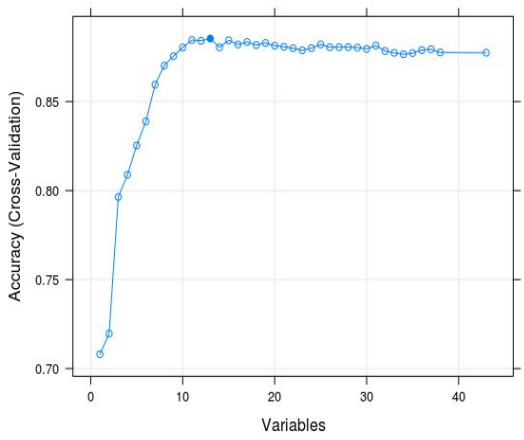

Figure 14: Number of selected features using Random forests and Cross-validation.

Figure 15 presents the selected features. The variables with symbol * are those that are highly correlated as shown in the correlation matrix in Figure 16. The accuracy of the model range between $(0.854, 0.874)$. Table 1 shows the sensitivity and specificity and the positive and negative predictive values.

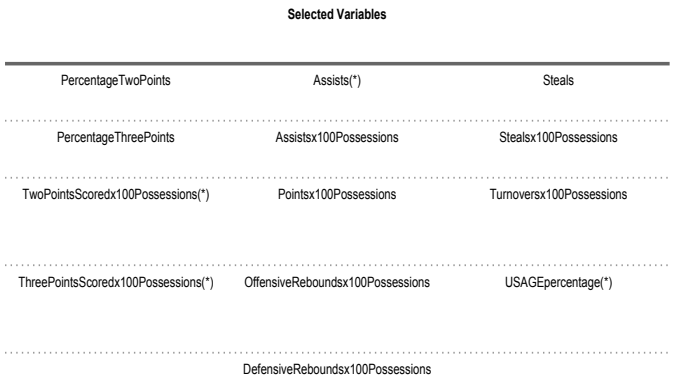

Figure 15: Selected features using Random forests and Cross-validation.

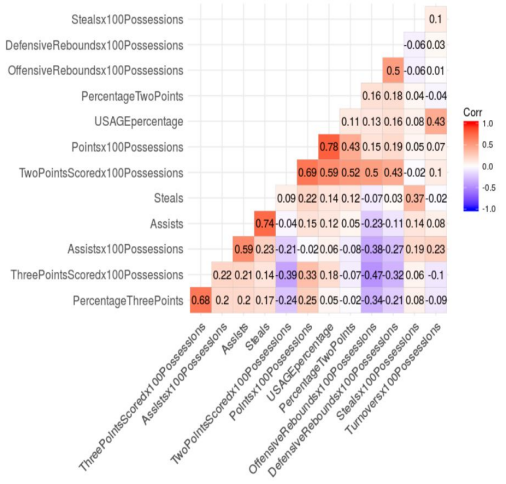

Figure 16: Correlation matrix of selected features.

4. Recommendations and further directions

Most of the patterns found are well-known by specialists in the field of sports analytics in Basketball, but mainly in NBA where the is only one competition and the data has been collected during many years. We provided some statistical support to the data from **Aryuna**©, with special focus on the comparison of several basketball leagues. It is important to mention that metrics are subjective, but they are very useful as it is not easy to find a overall global measure.

We recommend to include in the database some variables such as *"experi-*

Sensitivity	0.85
Specificity	0.88
Pos Pred Val	0.86
Neg Pred Val	0.87

Table 1: Predictive performance measures.

ence" i.e. the number of professional years or other variables to account for the team and coach effects.

The proposed statistical analysis allows for a further validation step in the complete database. However, there are still many open questions.

References

[1] S.M. Berry, C.S. Reese and P.D. Larkey. "Bridging different eras in Sports", *Journal of the American Statistical Association.*Vol. 94, No 447 (1999), pp. 661–676.

[2] S. Bruce. "A Scalable Framework for NBA Player and Team Comparisons Using Player Tracking Data" (2016) `https://arxiv.org/pdf/1511.04351.pdf`

[3] D. Coates and B. Oguntimein. *"The Length and Success of NBA Careers: Does College Production Predict Professional Outcomes"*, Working Papers Series, Paper No. 08–06. International Association of Sports Economists (2008).

[4] P. Fearnhead. and B.M. Taylor. *"On estimating the Ability of NBA players"*. Journal of Quantitative Analysis in Sports. Vol. 7, Issue 3, Article 11, (2011).

[5] J. Kubatko, D. Oliver, K. Pelton and D.T. Rosenbaum. "A starting point for analyzing Basketball statistics", *Journal of Quantitative Analysis in Sports.*Vol. 3, Issue 3, Art 1 (2007).

[6] G.L. Page, B.J. Barney, and A.T. McGuire. "Effect of position, usage rate, and per game minutes played on NBA player production curves". *Journal of Quantitative Analysis in Sports*, Vol. 9, issue 4, Dec (2013).

[7] G.L. Page and F. Quintana. "Predictions based on the Clustering of Heterogeneous Functions via Shape and Subject-specific covariates". *Bayesian Analysis*, 10, number 2, pp. 379–410.

[8] S. Shea. Basketball Analytics: spatial tracking (2014). `http://www.basketballanalyticsbook.com`.

[9] R.P. Schumaker, O.K. Solieman and H. Chen. Sports Data Mining. Integrated Series in Information Systems 26 (2010). Springer.