ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Data science in practice

Predicting career success of NBA player

Authors :

Matthieu Benat
Enguerrand Granoux
Victor Jost
Rym Karime
Reda Zahiri

Professor :

Bruffaerts Christopher

5th April 2018

Abstract

Data analysis is a powerful tool used to extract information from a database. It can be used in multiple fields and is nowadays, a widely recognized tool in the industry. The purpose of this report consists on explaining our analysis of the NBA database. As a matter of fact, we aim at providing a shortlist of young players, between ages 20 to 23 with high potential. We will then recommend this resulting list to NBA managers, which will allow them to invest on potential young players today for reasonable costs.

1 Introduction

The National Basketball Association (NBA) is a men's professional basketball league in North America. It is universally considered to be the premier men's professional basketball league in the world. The NBA is one of the four major professional sports leagues in the United States and Canada and its players are the world's best paid athletes.

The main entry point for the majority of NBA players happens in an annual evening where meet the NBA Commissioner and the leaders of the 30 teams, each team will select in turn a player from the university, high school, or abroad. This is called the NBA draft.

Initially, NBA teams used to select players from universities turn by turn until exhaustion. The order of each round was established in treverse order of the ranking of the past season. Nevertheless, the league decided to introduce a part of chance in obtaining the first choice : the worst team of each conference would play heads or tails their choice. However, accusations that certain teams were wilfully losing in order to gain a chance to participate in the annual coin flip.

In the current lottery system, the league uses a lottery-style ping-pong ball machine with 14 balls numbered from 1 to 14 and 1,000 four-digit combinations assigned to the 14 lottery teams. The worst team receives 250 combinations, the second worst gets 199, the third one 156, and so on. After the first three draft picks are determined, the rest of the teams are ordered in reverse order based on their record in the previous season.

Obviously, to be as competitive as possible for the coming season, NBA teams cannot bet only on the draft of beginner players in NBA. Indeed, in order to improve his team performance, the manager can either buy a top player with high statistics or bet on a potential younger player who started playing in the league just few years in the past.

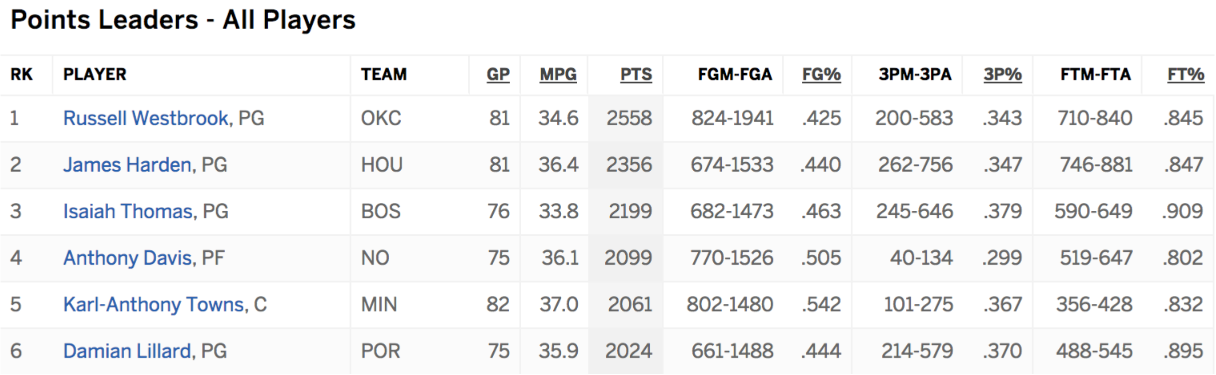
Nonetheless, the issue with top players is notably the high cost, they are tremendously expensive and usually the team can afford to sign only one of them during the transfer period. At this point, our analysis will suggest an alternative to the previous recruitment process based on a panel of selected young players with high potential. Thus, in our analysis we aim to predict the level of players in a few years.

Thereby, a team could invest in multiple young players at lower costs and rely on our predective analysis implying that these players would become top players in a few years. This will have two main advantages for the team. Firstly, they could have better players for reduced costs and secondly they could make more money by selling these players when they won't be needed anymore.

2 Data Acquisition

The NBA website contains plenty of different data on each player and each team. Therefore, the first step is to extract these data toward our analysis software in order to process them. All the statistics have been extracted from ESPN Stat. We chose to collect the data starting fifteen years earlier until the last completed season. This gives us in our case the statistics from 2003 until 2017.

At the first step, we downloaded every individual statistics :

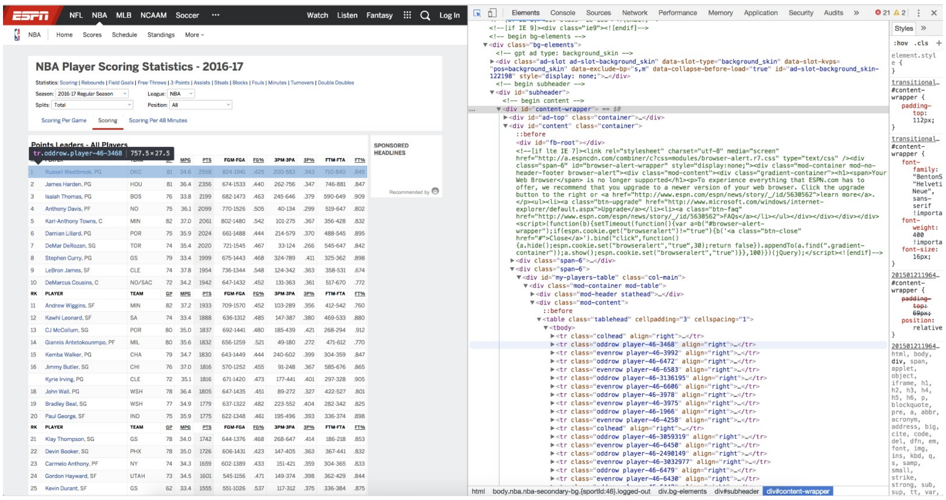


Points Leaders - All Players

RK	PLAYER	TEAM	GP	MPG	PTS	FGM-FGA	FG%	3PM-3PA	3P%	FTM-FTA	FT%
1	Russell Westbrook, PG	OKC	81	34.6	2558	824-1941	.425	200-583	.343	710-840	.845
2	James Harden, PG	HOU	81	36.4	2356	674-1533	.440	262-756	.347	746-881	.847
3	Isaiah Thomas, PG	BOS	76	33.8	2199	682-1473	.463	245-646	.379	590-649	.909
4	Anthony Davis, PF	NO	75	36.1	2099	770-1526	.505	40-134	.299	519-647	.802
5	Karl-Anthony Towns, C	MIN	82	37.0	2061	802-1480	.542	101-275	.367	356-428	.832
6	Damian Lillard, PG	POR	75	35.9	2024	661-1488	.444	214-579	.370	488-545	.895

Figure 1 – Example of the data format

Thanks to Google Chrome's Inspector we were able to understand the architecture of the different pages.



NBA Player Scoring Statistics - 2016-17

Statistics: Scoring | Rebounds | Field Goals | Free-Throws | 3-Points | Assists | Steals | Blocks | Fouls | Minutes | Turnovers | Double Doubles

Season: 2016-17 Regular Season League: NBA
Splits: Total Position: All

Scoring Per Game | Scoring | Scoring Per 48 Minutes

Points Leaders - All Players

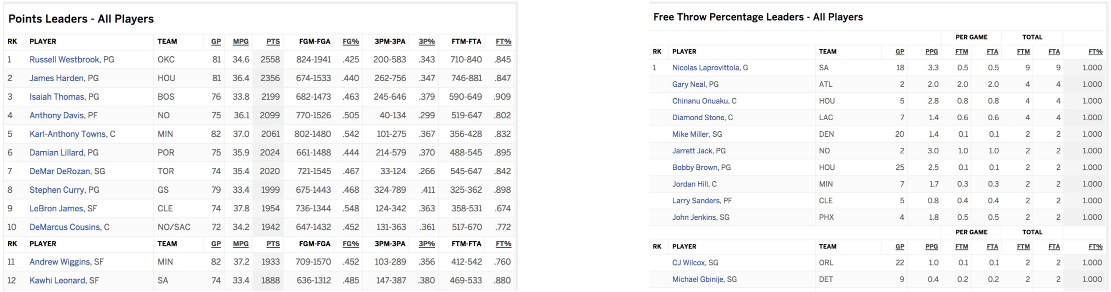
RK	PLAYER	TEAM	GP	MPG	PTS	FGM-FGA	FG%	3PM-3PA	3P%	FTM-FTA	FT%
1	Russell Westbrook, PG	OKC	81	34.6	2558	824-1941	.425	200-583	.343	710-840	.845
2	James Harden, PG	HOU	81	36.4	2356	674-1533	.440	262-756	.347	746-881	.847
3	Isaiah Thomas, PG	BOS	76	33.8	2199	682-1473	.463	245-646	.379	590-649	.909
4	Anthony Davis, PF	NO	75	36.1	2099	770-1526	.505	40-134	.299	519-647	.802
5	Karl-Anthony Towns, C	MIN	82	37.0	2061	802-1480	.542	101-275	.367	356-428	.832
6	Damian Lillard, PG	POR	75	35.9	2024	661-1488	.444	214-579	.370	488-545	.895
7	DeMar DeRozan, SG	TOR	74	35.4	2020	721-1545	.467	33-124	.266	545-647	.842
8	Stephen Curry, PG	GS	79	33.4	1999	675-1443	.468	324-789	.411	325-362	.898
9	LeBron James, SF	CLE	74	37.8	1954	736-1344	.548	124-342	.363	358-531	.674
10	DeMarcus Cousins, C	NO/SAC	72	34.2	1942	647-1432	.452	131-363	.361	517-670	.772
RK	PLAYER	TEAM	GP	MPG	PTS	FGM-FGA	FG%	3PM-3PA	3P%	FTM-FTA	FT%
11	Andrew Wiggins, SF	MIN	82	37.2	1933	709-1570	.452	103-289	.356	412-542	.760
12	Kawhi Leonard, SF	SA	74	33.4	1888	636-1312	.485	147-387	.380	469-533	.880
13	CJ McCollum, SG	POR	80	35.0	1837	692-1441	.480	185-439	.421	268-294	.912
14	Giannis Antetokounmpo, PF	MIL	80	35.6	1832	656-1259	.521	49-180	.272	471-612	.770
15	Kemba Walker, PG	CHA	79	34.7	1830	643-1449	.444	240-602	.399	304-359	.847
16	Jimmy Butler, SG	CHI	76	37.0	1816	570-1252	.455	91-248	.367	585-676	.865
	Kyrie Irving, PG	CLE	72	35.1	1816	671-1420	.473	177-441	.401	297-328	.905
18	John Wall, PG	WSH	78	36.4	1805	647-1435	.451	89-272	.327	422-527	.801
19	Bradley Beal, SG	WSH	77	34.9	1779	637-1322	.482	223-552	.404	282-342	.825
20	Paul George, SF	IND	75	35.9	1775	622-1348	.461	195-496	.393	336-374	.898
RK	PLAYER	TEAM	GP	MPG	PTS	FGM-FGA	FG%	3PM-3PA	3P%	FTM-FTA	FT%
21	Klay Thompson, SG	GS	78	34.0	1742	644-1376	.468	268-647	.414	186-218	.853
22	Devin Booker, SG	PHX	78	35.0	1726	606-1431	.423	147-405	.363	367-441	.832
23	Carmelo Anthony, PF	NY	74	34.3	1659	602-1389	.433	151-421	.359	304-365	.833
24	Gordon Hayward, SF	UTAH	73	34.5	1601	545-1156	.471	149-374	.398	362-429	.844
25	Kevin Durant, SF	GS	62	33.4	1555	551-1026	.537	117-312	.375	336-384	.875

```
<!--[if IE 9]><div class="ie9"><![endif]-->
<!-- begin bg-elements -->
▼<div class="bg-elements">
  <!-- gpt ad type: background_skin -->
 ▶<div class="ad-slot ad-slot-background_skin" data-slot-type="background_skin" data-slot-kvps=
"pos=background_skin" data-exclude-bp="s,m" data-collapse-before-load="true" id="ad-slot-background_skin-
122198" style="display: none;">…</div>
  <!-- begin subheader -->
 ▼<div id="subheader">
    <!-- begin content -->
... ▼<div id="content-wrapper"> == $0
     ▶<div id="ad-top" class="container">…</div>
     ▼<div id="content" class="container">
        ::before
        <div id="fb-root"></div>
        <!--[if lte IE 7]><link rel="stylesheet" charset="utf-8" media="screen"
        href="http://a.espncdn.com/combiner/c?css=modules/browser-alert.r7.css" type="text/css" /><div
        class="span-6" id="browser-alert-wrapper" style="display:none;"><div class="mod-container mod-no-
        header-footer browser-alert"><div class="mod-content"><div class="gradient-container"><h1><span>Your
        Web Browser</span> is no longer supported</h1><p>To experience everything that ESPN.com has to
        offer, we recommend that you upgrade to a newer version of your web browser. Click the upgrade
        button to the right or <a href="http://www.espn.com/espn/news/story/_/id/5630562">learn more</a>.
        </p><ul><li><a class="btn-upgrade" href="http://www.microsoft.com/windows/internet-
        explorer/default.aspx">Upgrade</a></li><li><a class="btn-faq"
        href="http://www.espn.com/espn/news/story/_/id/5630562">FAQs</a></li></ul></div></div></div></div>
        <script>(function(b){setTimeout(function(){var a=b("#browser-alert-
        wrapper");if(espn.cookie.get("browseralert")!="true"){b('<a class="btn-close"
        href="#">Close</a>').bind("click",function()
        {a.hide();espn.cookie.set("browseralert","true",30);return false}).appendTo(a.find(".gradient-
        container"));a.show();espn.cookie.set("browseralert","true")}},100)})(jQuery);</script><![endif]-->
       ▶<div class="span-6">…</div>
       ▼<div class="span-6">
         ▼<div id="my-players-table" class="col-main">
           ▼<div class="mod-container mod-table">
             ▶<div class="mod-header stathead">…</div>
             ▼<div class="mod-content">
                ::before
               ▼<table class="tablehead" cellpadding="3" cellspacing="1">
                 ▼<tbody>
                   ▶<tr class="colhead" align="right">…</tr>
                   ▶<tr class="oddrow player-46-3468" align="right">…</tr>
                   ▶<tr class="evenrow player-46-3992" align="right">…</tr>
                   ▶<tr class="oddrow player-46-6472" align="right">…</tr>
                   ▶<tr class="evenrow player-46-6583" align="right">…</tr>
                   ▶<tr class="oddrow player-46-3136195" align="right">…</tr>
                   ▶<tr class="evenrow player-46-6606" align="right">…</tr>
                   ▶<tr class="oddrow player-46-3978" align="right">…</tr>
                   ▶<tr class="evenrow player-46-3975" align="right">…</tr>
                   ▶<tr class="oddrow player-46-1966" align="right">…</tr>
                   ▶<tr class="evenrow player-46-4258" align="right">…</tr>
                   ▶<tr class="colhead" align="right">…</tr>
                   ▶<tr class="oddrow player-46-3059319" align="right">…</tr>
                   ▶<tr class="evenrow player-46-6450" align="right">…</tr>
                   ▶<tr class="oddrow player-46-2490149" align="right">…</tr>
                   ▶<tr class="evenrow player-46-3032977" align="right">…</tr>
                   ▶<tr class="oddrow player-46-6479" align="right">…</tr>
                   ▶<tr class="evenrow player-46-6430" align="right">…</tr>
```

```
:hov .cls +
element.style {
}
transitional…
#content-wrapper {
  padding-top: 112px;
}
transitional…
#content-wrapper {
  font-family: "BentonS "Helveti Neue", sans-serif !importa
  font-weight: 400 !importa
  font-size: 16px;
}
201501211964…
#content-wrapper {
  ~~padding-top: 69px;~~
  position: relative
}
201501211964…
html, body, div, span, applet, object, iframe, h1, h2, h3, h4, h5, h6, p, blockquote, pre, a, abbr, acronym, address, big, cite, code, del, dfn, em, font, img, ins, kbd, q, s, samp, small, strike, strong, sub, sup, tt, var,
```

Figure 2 – Architecture of the website

Then, in order to recover the data we used the BeautifulSoup Python module. Several problems have arisen for us. Indeed, each page presents many differences : some have 12 statistics, others only 3... Thus, as we wanted our acquisition code to be as general as possible, this represented a challenging issur for our work.



Points Leaders - All Players

RK	PLAYER	TEAM	GP	MPG	PTS	FGM-FGA	FG%	3PM-3PA	3P%	FTM-FTA	FT%
1	Russell Westbrook, PG	OKC	81	34.6	2558	824-1941	.425	200-583	.343	710-840	.845
2	James Harden, PG	HOU	81	36.4	2356	674-1533	.440	262-756	.347	746-881	.847
3	Isaiah Thomas, PG	BOS	76	33.8	2199	682-1473	.463	245-646	.379	590-649	.909
4	Anthony Davis, PF	NO	75	36.1	2099	770-1526	.505	40-134	.299	519-647	.802
5	Karl-Anthony Towns, C	MIN	82	37.0	2061	802-1480	.542	101-275	.367	356-428	.832
6	Damian Lillard, PG	POR	75	35.9	2024	661-1488	.444	214-579	.370	488-545	.895
7	DeMar DeRozan, SG	TOR	74	35.4	2020	721-1545	.467	33-124	.266	545-647	.842
8	Stephen Curry, PG	GS	79	33.4	1999	675-1443	.468	324-789	.411	325-362	.898
9	LeBron James, SF	CLE	74	37.8	1954	736-1344	.548	124-342	.363	358-531	.674
10	DeMarcus Cousins, C	NO/SAC	72	34.2	1942	647-1432	.452	131-363	.361	517-670	.772
RK	PLAYER	TEAM	GP	MPG	PTS	FGM-FGA	FG%	3PM-3PA	3P%	FTM-FTA	FT%
11	Andrew Wiggins, SF	MIN	82	37.2	1933	709-1570	.452	103-289	.356	412-542	.760
12	Kawhi Leonard, SF	SA	74	33.4	1888	636-1312	.485	147-387	.380	469-533	.880

Free Throw Percentage Leaders - All Players

					PER GAME		TOTAL		
RK	PLAYER	TEAM	GP	PPG	FTM	FTA	FTM	FTA	FT%
1	Nicolas Laprovittola, G	SA	18	3.3	0.5	0.5	9	9	1.000
	Gary Neal, PG	ATL	2	2.0	2.0	2.0	4	4	1.000
	Chinanu Onuaku, C	HOU	5	2.8	0.8	0.8	4	4	1.000
	Diamond Stone, C	LAC	7	1.4	0.6	0.6	4	4	1.000
	Mike Miller, SG	DEN	20	1.4	0.1	0.1	2	2	1.000
	Jarrett Jack, PG	NO	2	3.0	1.0	1.0	2	2	1.000
	Bobby Brown, PG	HOU	25	2.5	0.1	0.1	2	2	1.000
	Jordan Hill, C	MIN	7	1.7	0.3	0.3	2	2	1.000
	Larry Sanders, PF	CLE	5	0.8	0.4	0.4	2	2	1.000
	John Jenkins, SG	PHX	4	1.8	0.5	0.5	2	2	1.000
					PER GAME		TOTAL		
RK	PLAYER	TEAM	GP	PPG	FTM	FTA	FTM	FTA	FT%
	CJ Wilcox, SG	ORL	22	1.0	0.1	0.1	2	2	1.000
	Michael Gbinije, SG	DET	9	0.4	0.2	0.2	2	2	1.000

Figure 3 – Structure of the data

Another issue that we faced was the fact that some statistics like « Field Goals »have some caption in multiple sizes in the middle of the data which tend to create a lag when getting the data.

The second step consisted on downloading specifics data from players using their ID acquired in the previous section, such as : Personal Stats for Lebron James. In fact, the idea was to obtain personal informations such as weight, year and location of birth, experience and number in the draft. Those informations are from the current year and are fixed. We decided to stock this information in a ID_Info dictionary in order to avoid getting url call for each players wich would take a very long time. However, we also faced some problem during this step because of missing, wrong statistics of even in reverse order for some players.

2.1 Data selection

After the acquisition, our data frame was organized as follows : Player_ID, Player_Name, PTS, FGM-FGA, FG%, 3PM-3PA, 3P%, FTM-FTA, FT%, OFF, DEF, 2PM-2PA, 2P%, PPS, AST, TO, STL, PF, BLK, FLAG, TECH, EJECT, DBLDBL, TRIDBL, Position.

Here is a short description of the parameters :

GP : Game played
MPG : Minute per game
PTS : Points
FGM-A : Field goals made-attempted
FG% : Field goals percentage
3PM-A : 3-points made-attempted
3P% : 3-Point percentage
FTM-A : Free throws made-attempted
FT% : Free throws percentage
OFF : Offensive rebounds
DEF : Defensive Rebounds
PPS : Points Per Shot
AST : Assists
TO : Turnovers
STL : Steals
PF : Personal Fouls
BLK : Blocks
FLAG : Flagrant Fouls
TECH : Technical Fouls
EJECT : Ejections
DBLDBL : Double Doubles
TRIDBL : Triple Doubles
Position : Position

Before the analysis, we had to choose which information we needed to keep and which one to remove. We know that to optimize the clustering approach, we had to minimize the number of components. That is the reason why we decided to remove every redundant parameters such as in the case of field goal made - field goal attempted - field goal percentage. Indeed, we could keep only two of them and still have the same informations. We deleted then every « attempted »statistics. The same happens with parameters such as game played - minute played.

To go further : Obviously we could have chosen more parameters to increase the accuracy of our analysis. However, some interesting parameters are really difficult to find or to estimate.

For example, a young player can see his progression slowing down due to repetitive injuries. Taking into account this parameter could have been very useful to predict the progression of a player, nevertheless it is very hard to find a complete register with all the injuries and the incapacity time to play since 2003. Moreover, certain injuries must be more critical than others to practice an high level basketball.

Other parameters that could have been very useful to estimate the value of a player are indirect statistics linked to the defensive role. As a matter of fact, NBA experts recently pointed out that these parameters are left behind, distorting the estimate value of a player. If we take the example of the tallest player, Rudy Gobert, who measures 2m16, he is known to intimidate his opponents and pushing them to reduce their field goal attempts. This characteristic is indeed a real added value for the player and should be considered to estimate his level. Although, that kind of information is only available in the statistics for star player, it is almost impossible to find it for all the players.

NBA experts are actively working on a way to integrate these "hidden statistics" which could really improve the estimated value of a player. Thus, at our stage, these kinds of informations where too hard to find or to quantify in order to being implemented in the analysis.

2.2 Data visualization

As mentioned in the precedent section we worked with a lot of different data. Below you will find a table 1 with different characteristics on a sample of data used, that have been really useful during our analysis especially to compute the score of each player.

Data	Game Played	MPG	PTS	AST	Experience
Count	6930	6923	6923	6923	6930
Mean	53.82	20.49	0.44	113.00	8.95
SD	24.60	9.96	0.09	132.83	4.64
Min	1	0	0	0	1
25%	35	12.3	0.399	20	5
50%	61	19.9	0.438	67	9
75%	75	28.7	0.48	154	13
Max	85	43.1	1	925	21

Table 1 – Different characteristics on each data - sample

Obviously for some convenient issues we do not put all the data we had in this table. These kinds of informations of each statistics are essential to estimate the level of a player in the league in comparison to the others. Therefore, they are also essential to compute the score of a player which allow to estimate his level according to our analysis.

3 Data Organization

At this level, we had a data frame of every statistics for each player for all the years. Hence, we merged all the data frames according to the Player Name, Player ID, and the year. We chose to compare players with the same level of experience (i.e compare their first seasons then the second etc...). This allowed us to ignore the year as well as the player age.

We already had the current experience of each player, hence we computed their experience for each season. For players for whom we didn't know the experience, we counted their frequency of appearance in our database (which is equivalent to their number of seasons for the years retained).

Then, in our analysis we focused our interest on young players. Thus, we used the first seasons of a basketball career as a comparison base. Those for who we were not able to find any information regarding their first seasons (for example Shaquille O'Neal who started his career in 1992) were removed of our data frame.

Finally, we set up our data frame to have one line per player and the statistics of the seasons in columns starting from season 1 until season 14. Hence, our data frame has a pyramidal form in the sense that every player has a « first season »but only some of them have the fourteenth one.

Table 2 – Schematic table of our data frame

Name	First season			Second season			Third season		
	Stat 1	Stat 2	Stat 3	Stat 1	Stat 2	Stat 3	Stat 1	Stat 2	Stat 3
New player	1	1	1	-	-	-	-	-	-
Player	1	1	1	2	2	2	-	-	-
Old player	1	1	1	2	2	2	3	3	3

4 ML Models

4.1 Unsupervised Learning : Clustering Approach

At first, we tried a non supervised model (Unsupervised learning). In this model, the computer works without any prediction target. Our goal was to gather the players in different clusters according to their similarities.

We chose to apply K-Means Clustering algorithm for different seasons. The purpose was to find different groups of players with a certain experience (for example after one season of experience). Then, the idea was to analyze data of players with a superior experience and finally try to predict in which group a young player would belong in the future. In order to decide the number of groups, a small analysis has been done. Indeed, the more this number is big, the closest the players will be in it but the less relevant the result is. Finally, to evaluate the trade-off between relevance and accuracy, we ploted it on a graph.

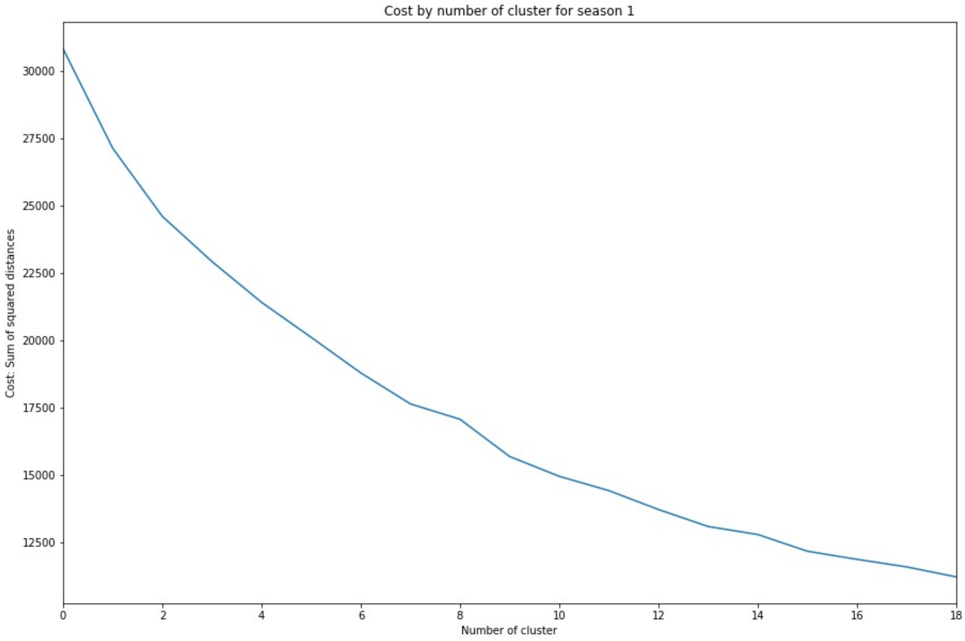


Figure 4 – Variance between each member of a group as a function of the number of groups

K-Means algorithm computes the distance for each feature, the greater the number of features, the less the algorithms will return relevent results. In order to reduce the number of features, we applied PCA algorithm (Principal Component Decomposition).

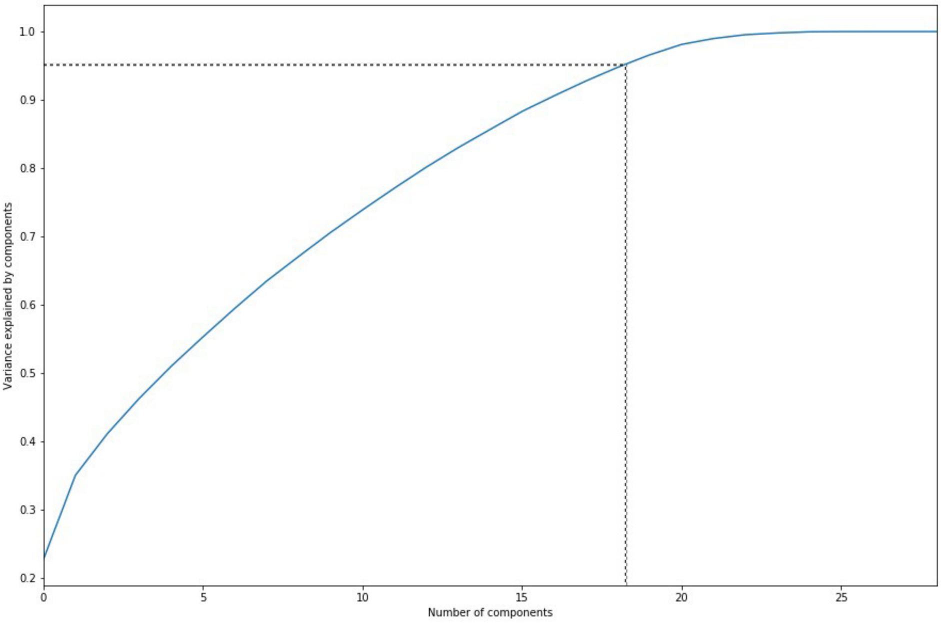


Figure 5 – PCA approach

As this point, we wanted to see how many informations we need to keep - from now, we have 29 features for each seasons. As the number of data decreases along the number of season, the number of informations to keep will also decrease. Hence we computed the variance explained by component as a function of the number of component to see how many features we should analyze, minimizing their number while keeping 95% of informations.
Here we can see the decomposition of the number of points for the season 3 under the principal component 1 and 2.

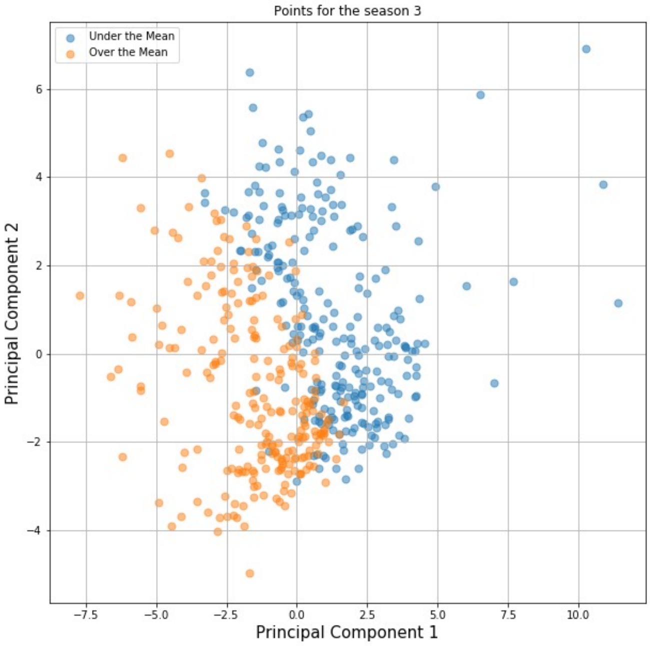


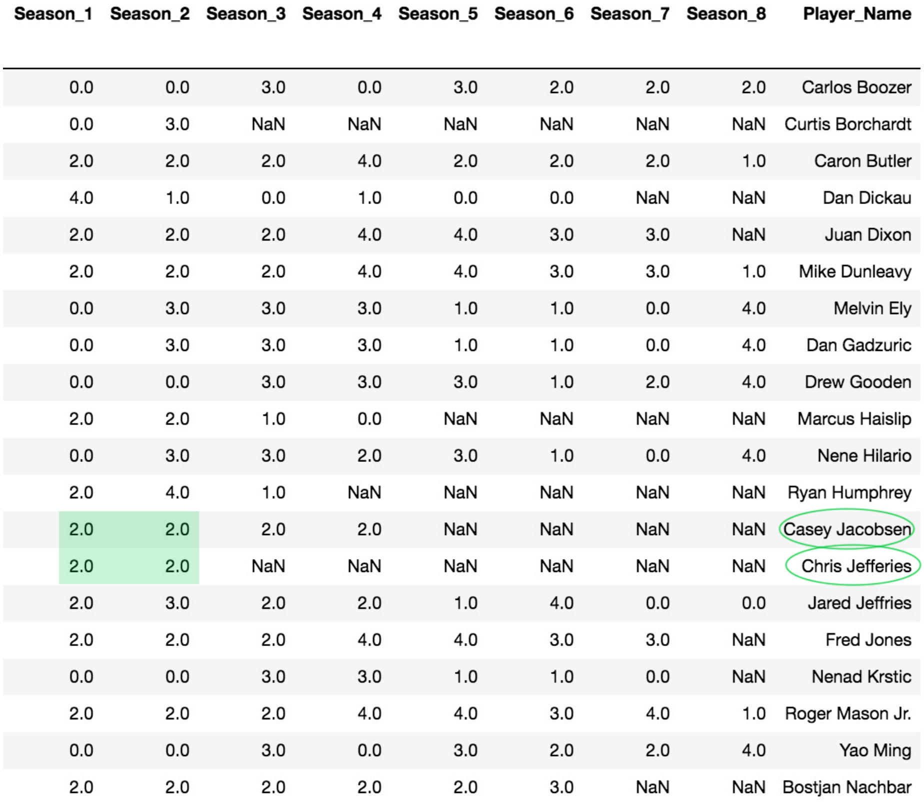
FIGURE 6 – Visualization of PCA Decomposition

Once our data frame were transformed using our PCA, we were able to apply K-Means algorithm leading us to different clusters.

At the end a cluster of players data exists for each players of the same number of years of experience. We did it only between first season to the 8^{th}, since we have just few players for more experimented, and the number of component for the PCA decreases.

Since the K-Means algorithm randomly assigns label to each cluster, we tried to change label name in order to maximize the number of equal label for players. The main issues of this method are that even if we could find some pattern in different cluster, we did not know to which cluster rallied the best player.

Bellow we can see the result of this clustering, we can find some pattern, for example Chris Jefferies seams to start his career like Casey Jacobsen.



Season_1	Season_2	Season_3	Season_4	Season_5	Season_6	Season_7	Season_8	Player_Name
0.0	0.0	3.0	0.0	3.0	2.0	2.0	2.0	Carlos Boozer
0.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	Curtis Borchardt
2.0	2.0	2.0	4.0	2.0	2.0	2.0	1.0	Caron Butler
4.0	1.0	0.0	1.0	0.0	0.0	NaN	NaN	Dan Dickau
2.0	2.0	2.0	4.0	4.0	3.0	3.0	NaN	Juan Dixon
2.0	2.0	2.0	4.0	4.0	3.0	3.0	1.0	Mike Dunleavy
0.0	3.0	3.0	3.0	1.0	1.0	0.0	4.0	Melvin Ely
0.0	3.0	3.0	3.0	1.0	1.0	0.0	4.0	Dan Gadzuric
0.0	0.0	3.0	3.0	3.0	1.0	2.0	4.0	Drew Gooden
2.0	2.0	1.0	0.0	NaN	NaN	NaN	NaN	Marcus Haislip
0.0	3.0	3.0	2.0	3.0	1.0	0.0	4.0	Nene Hilario
2.0	4.0	1.0	NaN	NaN	NaN	NaN	NaN	Ryan Humphrey
2.0	2.0	2.0	2.0	NaN	NaN	NaN	NaN	Casey Jacobsen
2.0	2.0	NaN	NaN	NaN	NaN	NaN	NaN	Chris Jefferies
2.0	3.0	2.0	2.0	1.0	4.0	0.0	0.0	Jared Jeffries
2.0	2.0	2.0	4.0	4.0	3.0	3.0	NaN	Fred Jones
0.0	0.0	3.0	3.0	1.0	1.0	0.0	NaN	Nenad Krstic
2.0	2.0	2.0	4.0	4.0	3.0	4.0	1.0	Roger Mason Jr.
0.0	0.0	3.0	0.0	3.0	2.0	2.0	4.0	Yao Ming
2.0	2.0	2.0	2.0	2.0	3.0	NaN	NaN	Bostjan Nachbar

FIGURE 7 – Cluster method result

However, this analysis was not very relevant because even if we had different clusters, we could not decide with great accuracy which one was the best, and really find a pattern. Hence, we chose to try a supervised approach.
Remark : 1.0, 2.0 ,3.& 4.0 denote the different clusters we found.

4.2 Supervised Learning : Regression

To apply a supervised method one has to define a target. The target can be seen as a grade, an evaluation of the player as a function of his statistics. As this score is very subjective, we tried different combinations, either linear or not of the statistics.

We tried to predict with different algorithms. As we tried a non-linear combination for the metrics, we applied Random forest Regression with grid search in order to optimize Hyperparameters such as the number of estimator and the max depth.

Then, we used a linear combination for the metric, we tried both linear regression and **lasso regression**. Finally, we got a better accuracy with the **lasso regression**. In both case (linear and non linear), we used 3 folds cross validation in order to avoid over-fitting. At the end, we chose to keep a linear combination of the statistics for the metric.

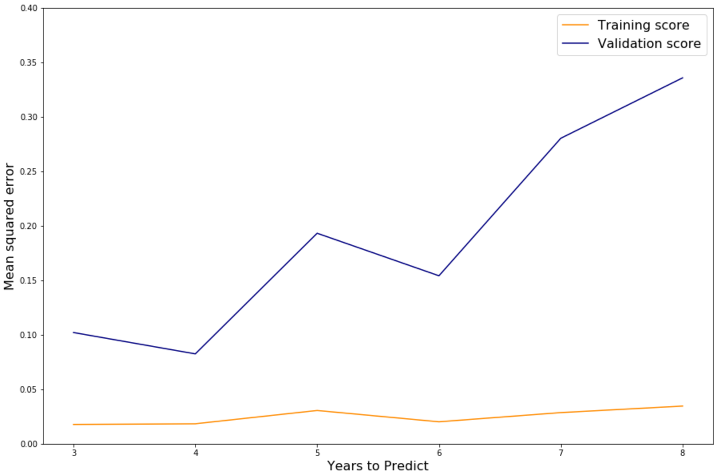


Figure 8 – Mean squared error for Lasso Regression

As we can see on the plot above, the training error is as expected very low. We can also notice that the testing error increases when the season to predict moves away from the initial season. This leads us to believe that our model has some weakness for a long term prediction, which clearly seems normal.

In order to find some metric to measure the skill of a player, we created a rough approximation for player skill. It involved creating a function using as parameters each statistics either positives or negatives one. For the sake of finding the real efficiency of a player, we expressed each statistics by minute played. The function compute the difference between the stats of the player along the season and the mean value of this statistic all over the period studied (2003-2017). Then this value is divided by this mean value so that we have a relative difference. If the score is equal to the mean, the player will not win any points, if he has twice the value, he will win 1 point etc... The idea of normalizing all stats is mandatory in order to weight each statistics in a fair way. Then a weight on each stats was roughly chosen and added to the function.

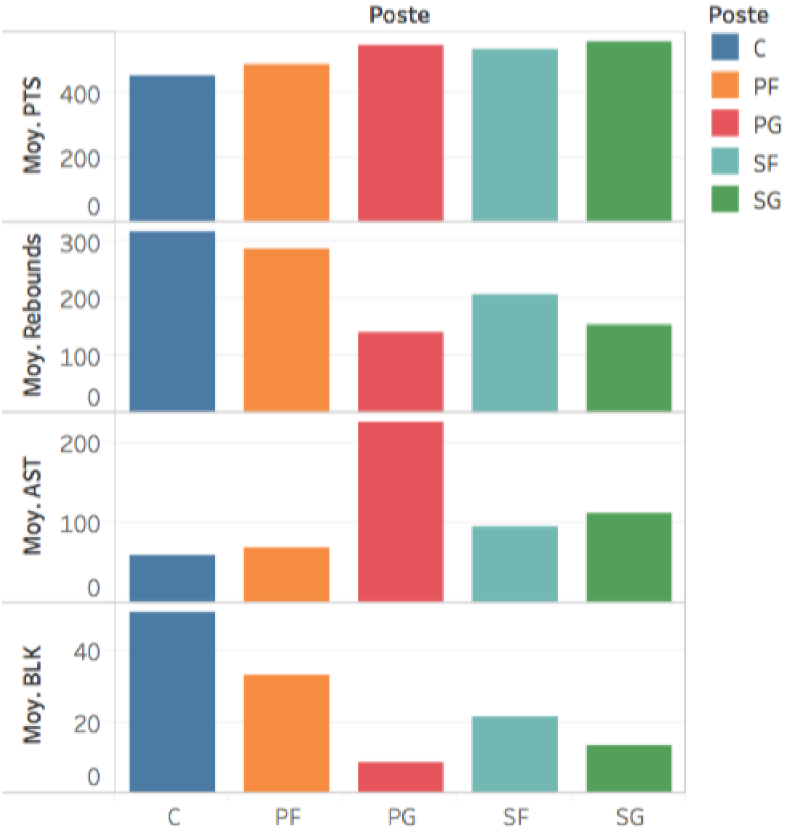


Figure 9 – Average for different statistics for each position

This diagram put forward that since different statistics are not counted in the same way (for example with one shoot you can made 3 points while one assist still stand for one in the table.) we have to implement weight on each statistics to compensate this difference. Indeed, using the previous example of the assist, one assist should not count three times less than a three point made, since this basket probably comes from that assist. This kind of diagram gives us the intuition to weighted all the different statistics.

Finally, for each player, we used the score due to the player's stats for one season. Obviously, our scoring method is not a perfect metric for comparing players against one another ; however, we felt that it would do a good job of showing an individual players change in performance over time.

Once we had a score for each player per season, we chose two experiences N_0 the initial one and N_1 the experience to predict. Hence we can train our algorithm on players with data for all the seasons until N_1 with regard to finally predict the score in N_1 of players with only N_0 seasons.

For this model, we did not have any issue with the running time since our dataset is actually not so big. Indeed, the model run in few minutes. The running time depends on how many years of prediction we want, and on which algorithm of RandomForest takes more time (we lauch the aglorithm for many parameters).

Unfortunately, results are not as good as expected. First, if we want to predict the potential of a player who only played 1 season we just have one year of statistic which is not enough to have a good prediction. We could solve this problem by adding to our model statistics on NCAA, the National Collegiate Athletic Association.

Another issues is that we collect data over 15 years but at the end if we want to predict the evolution of a young player over 6 years, we did not have so many models who have started their career after 2003 and played more than 6 seasons. In order to solve this, we can collect older data and maybe scale then by the difference between NBA season overall level.

Finally, our metric is not perfect. We did not have enough knowledges in NBA in order to find the best metric. In the future we could ask to a consultant and also compute a proper metric for each position.

5 Business Value

In this section, we will discuss the business aspects of our project. First of all, since a long time business took a very important dimension in the sport. Nowadays it is more obvious with more and more powerful investors in the most famous sports. In the USA, basketball is one of them and as we already mentioned it above, the NBA brings together the best players of America and the world. It is the perfect formula for a great show and for big economics fallout. The money comes from everywhere ; it can be the TV rights, the sponsors support, the game tickets, advertising (during the breaks of the game), etc... As competitors, players and staff want to win every single game and win the championship, they want to leave a track in the basketball history. Then, comes from the economic challenge. As a business man, one should notice that a bankable team, is a "visible" team. A bankable team is a team that attracts lot of fans, a team for which the TV are ready to buy the rights to broadcast the game,a team whose the kids ask for the shirt. A "visible" team is a winning one. Thus, the key of all this is to win as mush as possible. During a long time the choices of players were made on human analysis based on their experience through the years. The data analysis has revolutionized the way of thinking. It permits to support the human intuition or, in a much more interesting way to bring out players with high potentials left aside. Therefore, providing predictions of young players is very interesting for every NBA franchise. Indeed it will allow them to buy now young players that have not impressive statistic right now but that probably will make win their team in a few years. Since our model is totally automatized, it can be reused over the time by just updating statistics after each game or each season. We suppose that in the first place we will contact the teams to sell our shortlist of talented players. Therefore, the price should be reasonable as we will be the "rookies" of the game. In a second time, a few years later, at the time we could conclude if our analysis was great or not, we will put forward our results. In the targeted situation (i.e a great analysis) our work will be a guarantee that if the clients follow our instructions they could probably be at the head of the team that will win the championship within 3 or 4 years. Thus we will challenge all the interested teams by our offer and will sell it to the most generous one. We will of course try to improve the method each year to stay attractive and competitive. The possibility to last through time is a real strength for a business model and allow to generate revenue over years and years as long as the model stay efficient. Therefore our investment will be focus exclusively on this goal.

6 Result Presentation

In this section we will present the prediction of the estimate value of a sample of NBA players as expected. To have a readable result we only pick four players here. The purple line is the point where we begin to predict the score of the different players. As we can see in the green curve our model is note perfect at all, indeed since we did not take into account the age of the players the curve should increase during the time.

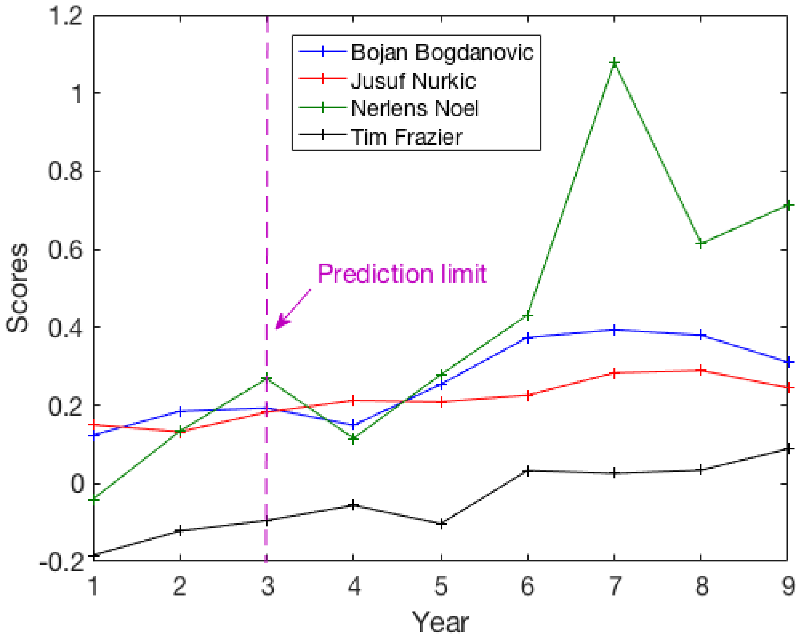


FIGURE 10 – Average for different statistics for each position

In the email you will find a huge pdf with the score of each NBA players for each year between 2003 and 2017. This pdf will allow to make a comparison with the estimate value of the young players. In the reference you will also find the link related to the pdf, where you can pick the player you want using a filter at the top right.

7 Conclusion

In this project we have seen one of the many applications of a data analysis. It is an easy approach to learn to manipulate data and extract some useful informations. However, we figured it could be very difficult to adjust the data in an efficient way to provide the required information. Finally, one has to be aware that these kind of analysis could not ever be perfectly exact. Indeed, human factors such as withstand an high pressure could no be predictable and integrate to the analysis.

References

— http://www.espn.com/nba/statistics

— https://public.tableau.com/profile/benat4706#!/vizhome/scores_3/Feuille1