International Player Translation Project

Name: Kuan-Cheng Fu

Date: 15 March 2021

1. Introduction

An analytics staff needs to provide some recommendations of players currently playing outside of the NBA for the GM to target. In detail, the staff wants to know what metrics predict success for players in the most relevant leagues in Europe and specifically which players he should highlight for this season. Therefore, the aim of this project is to develop a framework which is capable of converting players' European performance to NBA performance.

2. Data Overview

Before starting any analysis work, it's important to completely understand what is in the data and what the samples look like. In the provided SQLite database, there are three tables, "Player Information", "NBA Stats", and "European Stats", housing mock data on players who have played both in the NBA and one of the four main European leagues. The "Player Information" data contains three features, first name, last name, and birth date. Besides, the "NBA Stats" and "European Stats" data both contain first name, last name, season, season type, league, team, and 48 traditional and advanced stats, except that the "NBA Stats" data additionally contains a feature, plus-minus.

48 Traditional and Advanced Stats		
games	steals	estimated possessions
starts	deflections	calculated possessions
minutes	loose balls recovered	plays used
points	blocked shots	team possessions
two points made	personal fouls	usage percentage
two points attempted	personal fouls drawn	true shooting percentage
three points made	offensive fouls	three points attempt rate
three points attempted	charges drawn	free throw rate
free throws made	technical fouls	offensive rebounding percentage
free throws attempted	flagrant fouls	defensive rebounding percentage
blocked shot attempts	ejections	total rebounding percentage
offensive rebounds	points off turnovers	assist percentage
defensive rebounds	points in paint	steal percentage
assists	second chance points	block percentage
screen assists	fast break points	turnover percentage
turnovers	possessions	internal box plus-minus

3. Data Preprocessing

After understanding what is in the data, we then set player ID for the “Player Information” data and set indexes for the both “NBA Stats” and “European Stats” data in order to connect the data more conveniently. After that, we combine the “NBA Stats” and “European Stats” data and sorted the combined data by player ID and then by season in order to divide the players into three groups: players who have only played in the NBA (group 1), players who have only played in Europe (group 2), and players who have played both in the NBA and in Europe (group 3). The group 1 is apparently not our target audience while the players in the group 2 are who we are going to predict success and recommend for highlight. Meanwhile, the traditional and advanced stats of the players in the group 3 are going to be used to train predictive models.

In the beginning, we try to restrict the group 3 to players who have to play in Europe before in the NBA since it seems more reasonable regarding our purpose. However, there are only more than two hundred players left in the group 3, which I believe it's not enough for data modeling and further predicting success for the players in the group2. Therefore, at this stage, we assume that players basically keep their ability and skills so that we can completely focus on capture the relationships between “European Stats” and “NBA Stats”.

	Group 1	**Group 2**	**Group 3**
Number	190	996	477

After defining which players are included in the group 3, we then calculate career stats in Europe and in the NBA, respectively, for each player in the group 3 by using **groupBy** function to sum up or average the “NBA Stats” and “European Stats” data. Regarding the advanced stats, several of them such as true shooting percentage can be recalculated while several of them such as usage percentage can only be averaged instead of recalculating since we don’t have team and opponent team stats. In addition, we also add several new stats including eFG%, two points percentage, three points percentage, free throws percentage, GmSc (Game Score), and FPTS (Fantasy Points). It is worthy to note that now our desirable datasets only contain player ID and its corresponding career stats.

Method	**Traditional and Advanced Stats**
Sum	games, starts, minutes, points, two points made, two points attempted, three points made, free throws made, free throws attempted, blocked shot attempts, offensive rebounds, defensive rebounds, assists, turnovers, steals, blocked shots, personal fouls, personal fouls drawn, plus-minus
Average	possessions, team possessions, usage percentage, offensive rebounding percentage, defensive rebounding percentage, total rebounding percentage, assist percentage, steal percentage, block percentage, internal box plus-minus
Recalculate	true shooting percentage, three points attempt rate, free throw rate, turnover percentage
Add	eFG%, two points percentage, three points percentage, free throws percentage, GmSc (Game Score), FPTS (Fantasy Points).

4. Data modeling

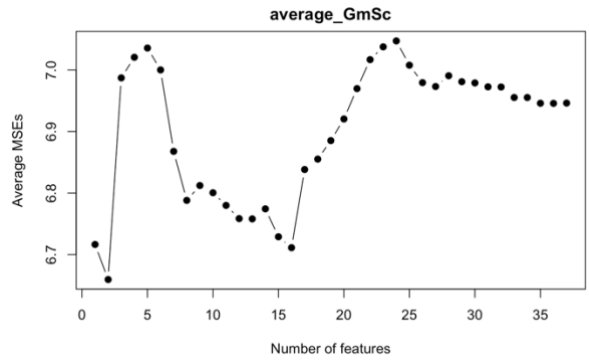
4.1 Multiple Linear Regression

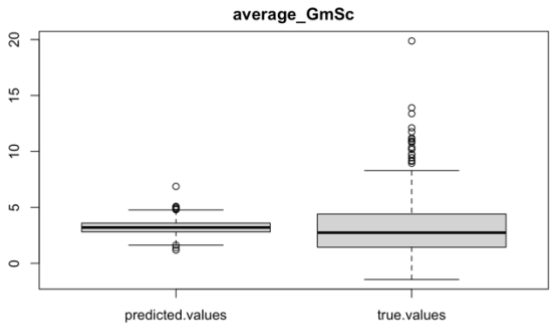
After finishing preparing the datasets, we then have to define our independent variables and dependent variables. Since our goal is to predict success for players, we decide to focus on several advanced stats which are generally used to evaluate players' overall performances, which are GmSc, FPTS, true shooting percentage, turnover percentage, and internal box plus-minus. Intuitively, we regard these five advanced stats from "NBA Stats" data as our dependent variables and regard all the stats from "European Stats" data as independent variables. In order to preliminarily understand the relationships between our independent variables and dependent variables, we first include all independent variables in a multiple linear regression model. The table below shows that no matter which dependent variable is considered, the p-value is way smaller than 0.005 which indicate the corresponding model is statistically significant at a significance level of 0.05 regarding the overall F-test. Meanwhile, we also find out that free throws made and eFG% are collinear variables, which have to be excluded from our independent variables.

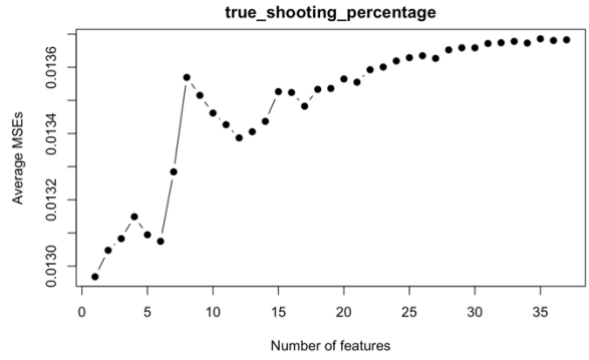
Dependent Variable	Collinear Variables	p-value (F-test)
GmSc	free throws made, eFG%	6.262853e-07
FPTS	free throws made, eFG%	1.005486e-05
true shooting percentage	free throws made, eFG%	4.244815e-02
turnover percentage	free throws made, eFG%	7.016423e-08
internal box plus-minus	free throws made, eFG%	4.726011e-05

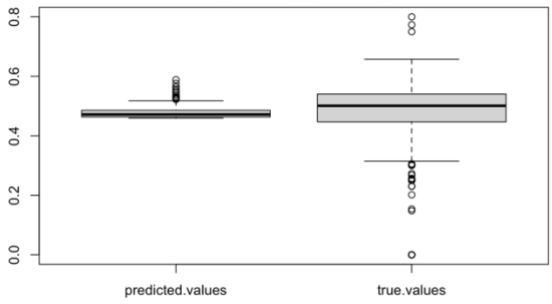
Furthermore, in order to find a better model regarding each dependent variable, we then apply a backward selection based on average MSE using 5-fold cross-validation, a method which aims to find a small subset of independent variables so that the resulting linear model is expected to have the most desirable prediction accuracy. From the figures below, for example, the linear model using GmSc as dependent variable with two independent variables has the smallest average 5-fold cross-validation average MSE when backward selection is performed. Meanwhile, we also notice that the p-values are all way smaller than 0.005 which indicate these five linear models are statistically significant at a significance level of 0.05 regarding the overall F-test. Also, we compare the predicted values and the true values using boxplots.

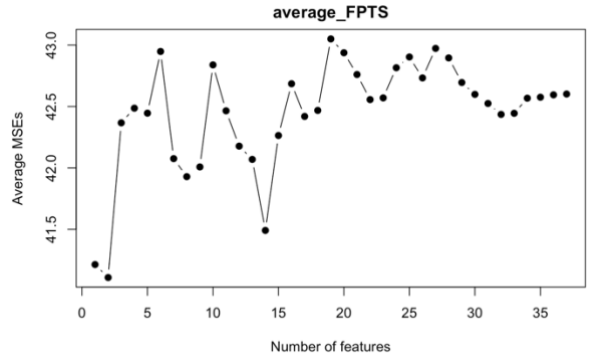
Dependent Variable	Independent Variables	MSE	p-value (F-test)
GmSc	defensive rebounding percentage, GmSc (2)	6.66	2.055232e-06
FPTS	usage percentage, defensive rebounding percentage (2)	41.11	4.191376e-06
true shooting percentage	block percentage mean (1)	0.013	2.195257e-04
turnover percentage	blocked shot attempts, turnover percentage (2)	39.19	2.224509e-15
internal box plus-minus	blocked shot attempts, steals, usage percentage, offensive rebounding percentage, total rebounding percentage, internal box plus-minus (6)	23.36	2.869160e-10

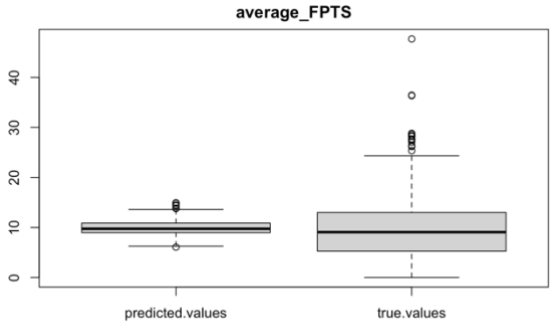

average_GmSc
Average MSEs
Number of features

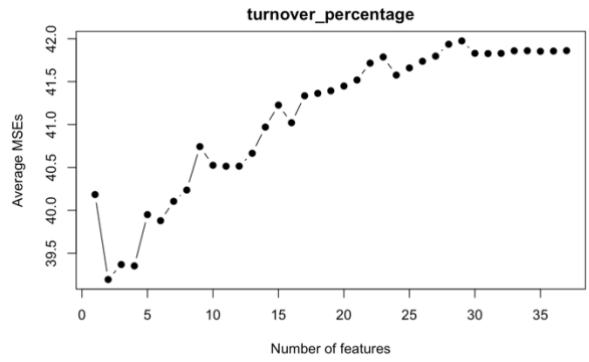

average_GmSc
predicted.values
true.values

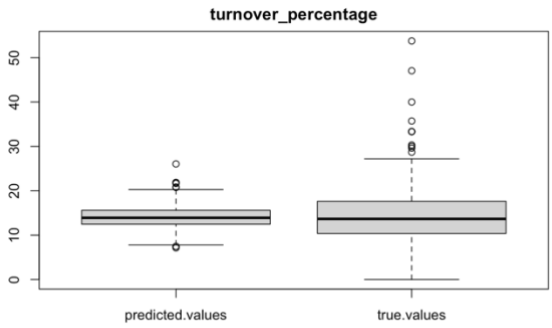

true_shooting_percentage
Average MSEs
Number of features

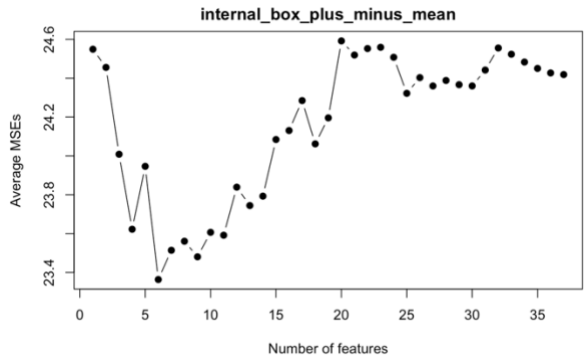

true_shooting_percentage
predicted.values
true.values

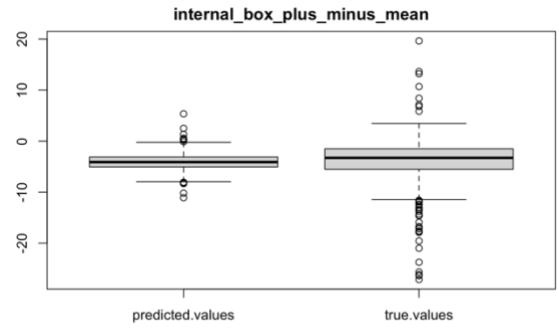

average_FPTS
Average MSEs
Number of features


average_FPTS
predicted.values
true.values


turnover_percentage
Average MSEs
Number of features


turnover_percentage
predicted.values
true.values


internal_box_plus_minus_mean
Average MSEs
Number of features

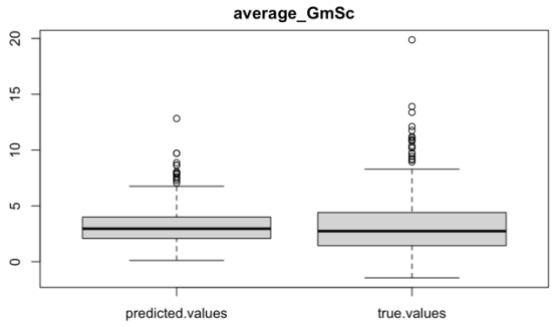

internal_box_plus_minus_mean
predicted.values
true.values

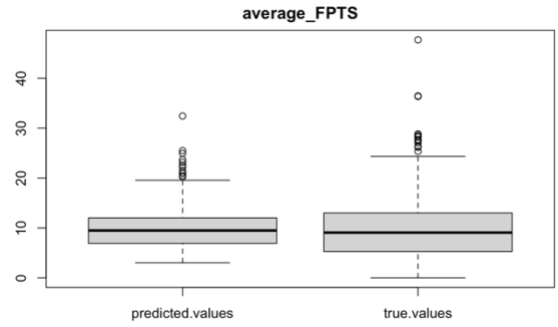
4.2 Random Forest

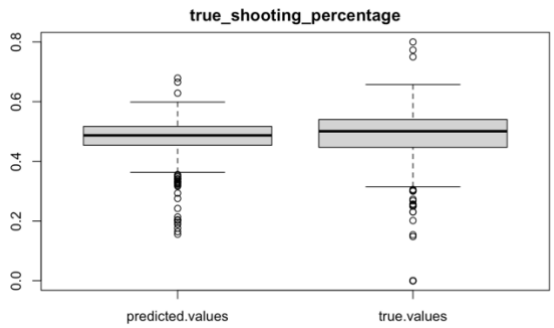
After dealing with the issue with a statistical way, we then try to improve our findings in a machine learning way. Therefore, in order to further reduce average MSE, we decide to apply a random forest model to our datasets. In the same way, we regard GmSc, FPTS, true shooting percentage, turnover percentage, and internal box plus-minus from “NBA Stats” data as our dependent variables and regard all the stats from “European Stats” data as independent variables. Most importantly, we use 5-fold cross-validation to estimate average MSE and further tune the parameters including mtry and ntree.

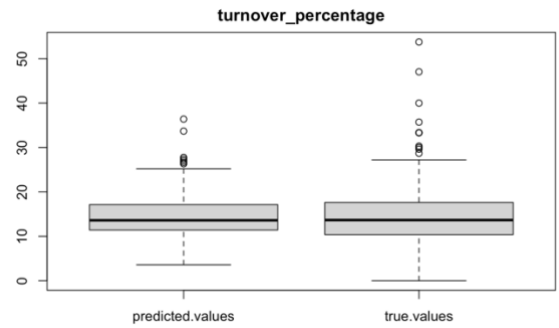
From the table below, since tuning parameters is time-consuming, we only try a small range of values for mtry and ntree. As a result, we can find out that the random forest model doesn’t yield an improvement over the linear model regarding the average MSE. However, from the boxplots below, we can conclude that the random forest model is still better than the linear model since the distributions of the predicted values and the true values are more similar.

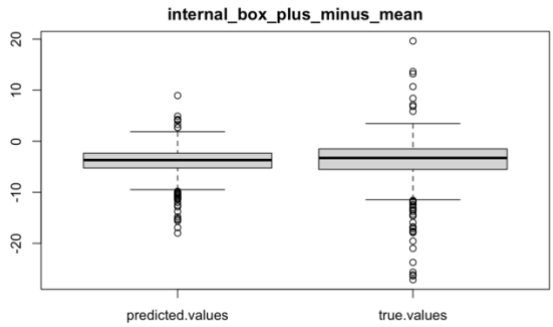
Dependent Variable	mtry	ntree	MSE
GmSc	20	1000	6.66
FPTS	20	500	40.27
true shooting percentage	20	500	0.014
turnover percentage	20	1000	39.34
internal box plus-minus	20	500	24.12









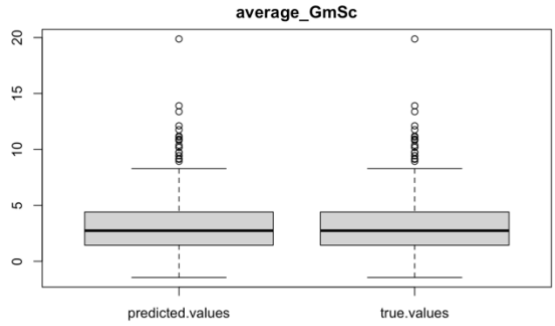


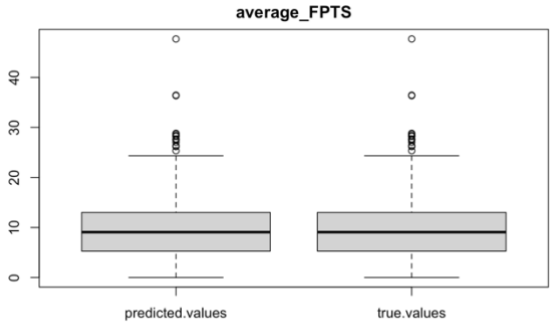
4.3 XGBoost

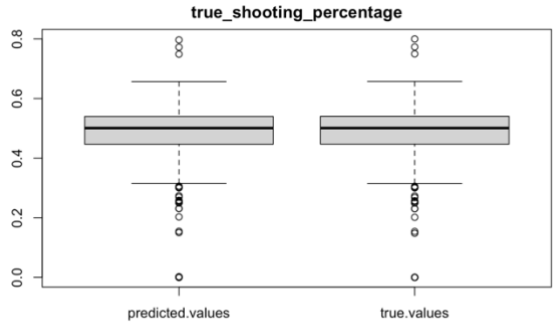
Moreover, we apply a XGBoost model to our datasets. Identically, we regard GmSc, FPTS, true shooting percentage, turnover percentage, and internal box plus-minus from “NBA Stats” data as our dependent variables and regard all the stats from “European Stats” data as independent variables. Also, we use 5-fold cross-validation to estimate average MSE and further tune the parameters including eta and max_depth.

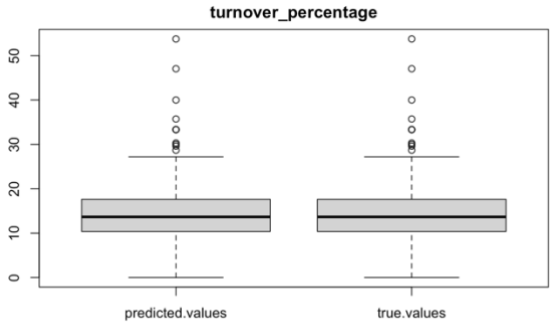
From the boxplots below, we can find out that the XGBoost model seems better than the random forest model since the distributions of the predicted values and the true values are more similar. However, from the table below, the XGBoost model doesn’t yield an improvement over the linear model or the random forest model regarding the average MSE. we believe the reason behind is over-fitting since the XGBoost model is too powerful. Therefore, we have to put lots more efforts on tuning the parameters to avoid over-fitting.

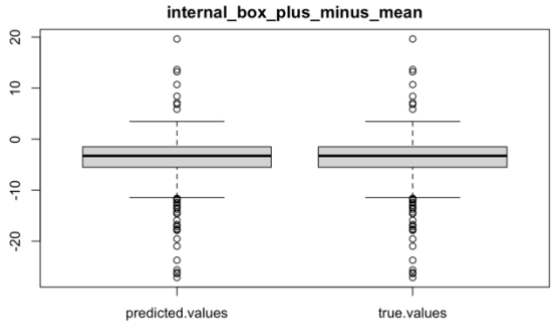
Dependent Variable	max_depth	eta	MSE
GmSc	6	0.3	7.78
FPTS	6	0.3	49.2
true shooting percentage	6	0.3	0.017
turnover percentage	6	0.3	47.01
internal box plus-minus	6	0.3	28.44











5. Result

In summary, we first use multiple linear regression to verify that there are statistical relationships between the “NBA Stats” and “European Stats” data. After that, we use random forest and XGBoost to improve the quality of our predicted values. Since the random forest model and the XGBoost model both perform well, we decide use both models to predict GmSc, FPTS, true shooting percentage, turnover percentage, and internal box plus-minus in the NBA for the players who have only played in Europe (group 2).

However, since we use two models and five evaluation metrics, we finally get ten different prospect rankings. Therefore, we use **RankAggreg** function, which performs aggregation of ordered lists based on the ranks using several different algorithms, to generate one top-50 prospect ranking based on those ten different prospect rankings. Besides, from those top-50 prospects’ stats in Europe, we can notice that they basically have pretty good field goal percentage and play pretty stable minutes on the court. Although the result might not be perfect, I believe that I still provide a reliable framework to deal with this issue.

Full Name	
pablo ostertag	magette watson
lampe borg	jamychal benson
rakocevic calvin	earl mcgee
ridvan labissiere	kljajic knudsen
zaza neal	terry palacio
osetkowskwe looney	eddie siva
mac sanli	khalid madsen
dean collins	shields foote
farley atkins	wolkowyskwe carlisle
la torre mccaskill	o'neale dimsa
arnoldas jacobsen	brandone mccallum
slokar caboclo	printezis akdamar
saben lewis	olek mcconnell
townes ouattara	luke balbi
bargnanwe turiaf	maksim ketner
brodziansky hancock	clyburn hairston
fletcher kravic	xabwe nwaba
deon pavani	alberto williams-goss
pargo buva	karolis nachbar
fontecchio muric	fedor sy
kesteloot cain	brodric powell
eldridge robinson	harrison podkolzin
pelle silva	krunoslav rebraca
johnson huertas	carlo mcguire
derek dickau	akyazilwe doumbouya

	name	G	MP	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	pablo ostertag	45	27.4	4.3	8.4	51.330	1.5	3.6	43.125	1.2	2.0	59.091	2.0	4.0	6.0	1.7	1.2	0.8	0.9	2.0	11.3
2	lampe borg	143	18.4	2.2	4.6	46.988	0.7	1.9	36.940	0.8	1.0	74.830	0.6	3.7	4.3	2.6	0.9	0.5	1.4	2.1	5.8
3	rakocevic calvin	14	23.2	3.9	6.3	62.500	0.0	0.1	0.000	3.1	5.1	61.111	2.6	2.9	5.5	1.0	0.9	0.9	1.6	2.1	11.0
4	ridvan labissiere	18	28.2	6.7	10.2	65.574	0.3	0.4	71.429	3.8	5.4	71.134	3.0	5.6	8.6	2.0	0.6	0.6	2.3	2.4	17.4
5	zaza neal	64	32.4	4.9	9.2	53.650	0.6	1.9	33.333	2.5	3.5	72.321	0.8	5.5	6.3	4.6	1.7	0.5	3.2	2.8	13.0
6	osetkowski looney	25	21.6	6.1	10.0	61.446	0.1	0.2	50.000	2.2	2.6	85.938	2.3	4.0	6.3	0.9	1.2	0.5	1.7	2.8	14.6
7	mac sanli	17	30.4	3.8	7.4	50.794	2.0	4.5	44.737	4.7	5.2	90.909	0.7	4.1	4.8	1.5	1.1	0.1	1.1	2.7	14.2
8	dean collins	22	24.6	6.1	9.5	63.810	0.0	0.0	0.000	5.2	5.9	89.147	2.5	3.0	5.5	2.3	0.9	0.5	2.3	2.2	17.4
9	farley atkins	11	33.9	5.2	10.2	50.893	1.7	4.0	43.182	3.2	3.8	83.333	1.3	4.3	5.5	2.8	1.9	0.4	2.5	3.4	15.3
10	la torre mccaskill	10	26.3	5.6	10.3	54.369	1.3	2.5	52.000	5.4	6.3	85.714	1.6	3.7	5.3	2.8	1.0	0.1	1.9	2.1	17.9
11	arnoldas jacobsen	14	27.7	5.9	11.5	51.553	2.0	4.9	41.176	2.7	3.6	76.000	0.5	2.4	2.9	5.9	1.4	0.1	1.2	3.1	16.6
12	slokar caboclo	12	25.1	4.8	9.0	52.778	0.6	2.3	25.000	2.7	3.8	71.111	1.5	4.1	5.6	1.3	1.9	0.6	2.0	2.7	12.8
13	saben lewis	36	22.5	3.5	6.3	54.825	0.0	0.1	0.000	1.4	1.9	73.134	1.9	3.0	4.8	1.6	0.8	0.6	1.0	2.5	8.3
14	townes ouattara	30	24.5	1.8	4.0	44.538	0.9	2.5	36.842	0.8	1.1	78.125	0.4	3.7	4.1	5.7	1.5	0.2	1.7	1.9	5.3
15	bargnani turiaf	37	25.9	5.0	9.3	53.623	1.4	3.4	41.732	2.1	2.5	82.979	0.7	3.2	3.9	2.0	1.1	0.3	1.1	1.2	13.5
16	brodziansky hancock	15	30.9	5.4	12.8	42.188	2.4	6.6	36.364	5.9	6.5	89.796	0.5	3.9	4.4	4.1	1.3	0.1	2.5	3.1	19.1
17	fletcher kravic	214	22.7	3.8	5.9	63.307	0.0	0.2	24.324	1.1	1.7	65.289	2.3	3.8	6.1	1.2	0.9	0.7	0.9	1.9	8.7
18	deon pavani	13	23.0	4.6	8.0	57.692	1.5	3.5	41.304	1.8	2.0	88.462	1.7	3.8	5.5	0.8	0.7	0.7	1.1	2.2	12.5
19	pargo buva	6	28.8	1.8	3.8	47.826	1.2	2.8	41.176	2.3	2.7	87.500	0.3	5.2	5.5	7.5	0.3	0.0	3.2	2.3	7.2
20	fontecchio muric	40	22.1	4.0	8.0	49.844	1.0	2.6	39.806	2.2	2.6	85.437	2.1	3.5	5.6	2.5	0.9	0.2	2.1	1.6	11.2
21	kesteloot cain	14	26.2	4.2	10.1	41.844	1.4	4.7	30.303	5.7	6.7	85.106	0.3	1.8	2.1	3.6	1.1	0.1	1.9	2.7	15.6
22	eldridge robinson	59	17.4	3.1	6.6	46.907	1.5	3.4	42.857	2.6	3.3	81.250	0.5	2.0	2.5	2.2	0.7	0.1	1.4	2.1	10.3
23	pelle silva	18	23.8	3.8	6.6	57.983	0.1	0.1	50.000	2.1	3.3	62.712	2.2	2.1	4.3	1.2	1.0	1.1	0.8	2.3	9.8
24	johnson huertas	14	26.4	5.6	11.7	48.171	0.7	2.8	25.641	2.2	3.1	72.093	1.6	2.9	4.4	2.7	1.4	0.1	1.4	2.5	14.2
25	derek dickau	33	27.1	4.3	8.3	51.825	1.0	2.5	38.554	2.1	2.8	72.340	1.4	4.2	5.6	2.0	1.3	0.3	1.7	1.8	11.6
26	magette watson	24	24.9	3.6	7.5	48.603	1.3	3.3	40.000	1.7	2.0	85.106	1.7	5.1	6.8	1.5	0.6	0.5	0.8	1.8	10.2
27	jamychal benson	4	21.6	1.8	5.5	31.818	0.2	2.0	12.500	1.2	1.2	100.000	0.0	3.8	3.8	0.8	1.0	0.2	0.8	1.5	5.0
28	earl mcgee	8	22.2	6.1	11.1	55.056	2.4	3.8	63.333	2.6	3.2	80.769	0.5	1.6	2.1	1.4	0.8	0.0	0.8	3.0	17.2
29	kljajic knudsen	8	21.3	3.1	6.1	51.020	0.8	1.1	66.667	2.1	2.8	77.273	1.0	3.1	4.1	0.4	0.9	0.9	1.2	2.0	9.1
30	terry palacio	4	16.6	2.0	6.5	30.769	1.0	2.8	36.364	2.0	2.8	72.727	1.0	3.0	4.0	3.2	0.5	0.0	1.5	1.8	7.0
31	eddie siva	18	30.1	5.2	9.8	53.409	3.1	6.2	50.450	3.6	4.5	79.012	0.7	2.7	3.3	1.4	1.1	0.1	1.7	2.2	17.1
32	khalid madsen	10	13.5	0.9	4.3	20.930	0.2	1.6	12.500	0.4	0.8	50.000	1.2	1.3	2.5	0.5	0.0	0.1	0.2	1.7	2.4
33	shields foote	32	24.7	4.2	8.0	53.333	1.3	3.2	40.196	0.7	0.8	87.500	0.6	3.0	3.6	1.9	1.2	0.5	1.1	2.2	10.4
34	wolkowyski carlisle	215	28.2	5.5	11.8	46.850	1.5	3.9	37.500	3.7	4.7	78.890	0.7	2.4	3.1	2.3	0.7	0.2	2.0	2.0	16.3
35	o'neale dimsa	21	13.1	2.9	4.0	71.429	0.0	0.0	0.000	1.9	2.6	72.222	0.8	1.2	2.0	0.9	0.6	0.1	1.0	1.9	7.6
36	brandone mccallum	6	30.8	5.7	14.2	40.000	2.8	9.2	30.909	5.7	6.5	87.179	0.8	3.5	4.3	2.2	1.2	0.0	2.5	3.2	19.8
37	printezis akdamar	6	5.8	0.5	0.7	75.000	0.3	0.3	100.000	0.3	1.0	33.333	0.2	0.7	0.8	0.7	0.3	0.0	0.3	0.2	1.7
38	olek mcconnell	15	16.1	2.4	5.2	46.154	1.5	3.3	46.000	0.3	0.3	80.000	0.3	1.7	2.1	1.8	0.5	0.0	0.8	1.3	6.6
39	luke balbi	147	26.0	4.5	10.9	41.454	2.6	6.8	37.873	2.4	2.9	83.099	0.4	1.6	2.0	3.5	1.2	0.0	2.0	1.9	14.1
40	maksim ketner	1	26.4	5.0	10.0	50.000	3.0	6.0	50.000	0.0	2.0	0.000	0.0	4.0	4.0	3.0	2.0	0.0	0.0	1.0	13.0
41	clyburn hairston	17	23.6	4.1	7.8	53.030	1.2	2.9	42.000	1.7	2.2	78.378	0.8	2.8	3.6	5.9	1.0	0.3	2.8	1.9	11.2
42	xabi nwaba	29	27.3	6.6	11.5	56.886	0.0	0.0	0.000	2.6	4.0	65.217	2.4	5.0	7.4	1.1	0.7	0.4	2.2	2.5	15.7
43	alberto williams-goss	2	25.8	3.0	11.5	26.087	1.5	5.5	27.273	1.0	1.5	66.667	0.5	2.5	3.0	2.5	0.5	0.0	2.0	4.5	8.5
44	karolis nachbar	2	21.4	5.0	9.0	55.556	1.5	2.5	60.000	3.0	3.5	85.714	1.5	1.5	3.0	1.0	0.5	0.0	0.5	2.0	14.5
45	fedor sy	7	10.5	1.1	3.0	38.095	0.4	2.0	21.429	0.1	0.3	50.000	1.0	0.7	1.7	0.1	0.1	0.0	0.0	0.9	2.9
46	brodric powell	45	16.6	1.9	3.9	47.458	0.8	1.9	43.023	0.7	1.0	68.085	0.5	2.3	2.9	1.9	0.8	0.4	1.4	1.9	5.3
47	harrison podkolzin	20	24.3	4.0	7.9	50.633	1.9	4.6	40.217	0.4	0.7	61.538	0.8	4.0	4.8	1.1	1.2	0.2	1.0	2.0	10.2
48	krunoslav rebraca	6	5.7	0.3	0.7	50.000	0.2	0.2	100.000	0.2	0.3	50.000	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.7	1.0
49	carlo mcguire	24	22.7	2.6	6.1	42.177	1.6	4.0	39.583	0.1	0.3	37.500	0.5	2.2	2.6	0.7	0.7	0.0	0.5	2.1	6.9
50	akyazili doumbouya	24	25.7	2.7	5.3	51.181	0.8	2.0	40.426	2.3	2.7	87.500	1.5	3.5	5.1	1.5	0.7	0.2	0.8	1.6	8.5