Evaluating Euroleague players using RAPM

Task description

The goal of the project is to develop a Machine Learning project to quickly compute RAPM (Regularized Adjusted Plus-Minus) for basketball leagues outside of the NBA, where this kind of advanced metric is uncommon to find. The main goal is to integrate this software analysis with “Sdeng”, a python library I develop which allows to scrape basketball data from leagues websites to csv files or SQL.

If possible, the goal would be to be able to split RAPM in ORAPM (Offensive RAPM) and DRAPM (Defensive RAPM).

Background

Plus Minus (+/-) is a basketball statistic which tries to evaluate the performance of a team when a player is on the court, by summing the score differential across all actions when a player is on court. Nonetheless, Plus Minus (+/-) is known to have several shortcomings.

The main reason is that since basketball is a team sport, a player alone is not entirely accountable for the performance of the team. In fact, if an average player constantly shares the court with very good teammates, he will generally have a high +/-, despite having a reduced impact on the game. On the contrary, a good player in a bad team will likely have a negative +/-, despite bringing positive value to his team.

Adjusted Plus Minus (APM) is an advanced metric that tries to evaluate a player's impact regardless of the teammates and opponents he shares the court with. It was first developed by Dan Rosenbaum in Measuring How NBA Players Help Their Teams Win.. It does so by using a regression considering all the players on court in a given part of the game and the result of that specific part of game. Nonetheless, it has a major problem, which is the tendency to overfit data. To overcome this, Regularized Adjusted Plus Minus (RAPM) was developed, which regularizes data to avoid overfitting and smoothen outliers.

RAPM is a spread and common statistic in NBA, both for the bigger interest that NBA draws, both for the easiness to fetch NBA data. Yet, it is not very common in European basketball, for a combination of factors, such as the absence of a unified stats API, fewer fans and way smaller datasets.

In fact, on average an NBA season has more than 4 times the number of games than a EL season; also, games are longer (48 minutes VS 40 minutes), and are played at a higher pace (in the current season the average possession lasts 16.9 seconds in EL, while 15.2 seconds in NBA).

Methods

A crucial aspect of the project is the data cleaning. Starting from the traditional play-by-play logs, which include the 10 players on-court. Using R package tidyverse, I obtained a dataset containing each stint for each player.

Then, we will use R lm() function and the glm package to perform the regressions.

team	player	+/-	possessions	net rating
H	A	5	25	20
H	B	8	15	53.3
H	C	5	25	20
H	G	-3	10	-30
V	D	-5	25	-20
V	E	-5	25	-20
V	F	-2	5	-40
V	J	-3	20	-15

Adjusted Plus Minus

Adjusted Plus Minus (APM) is a simple linear regression, which uses the net rating as the target variable and possessions as weights.

Nonetheless, the results obtained through this method provide an extremely high standard error and p-values, which does not allow to reject the null hypothesis.

It becomes necessary to apply a kind of regularization, which is why Regularized Adjusted Plus Minus was developed.

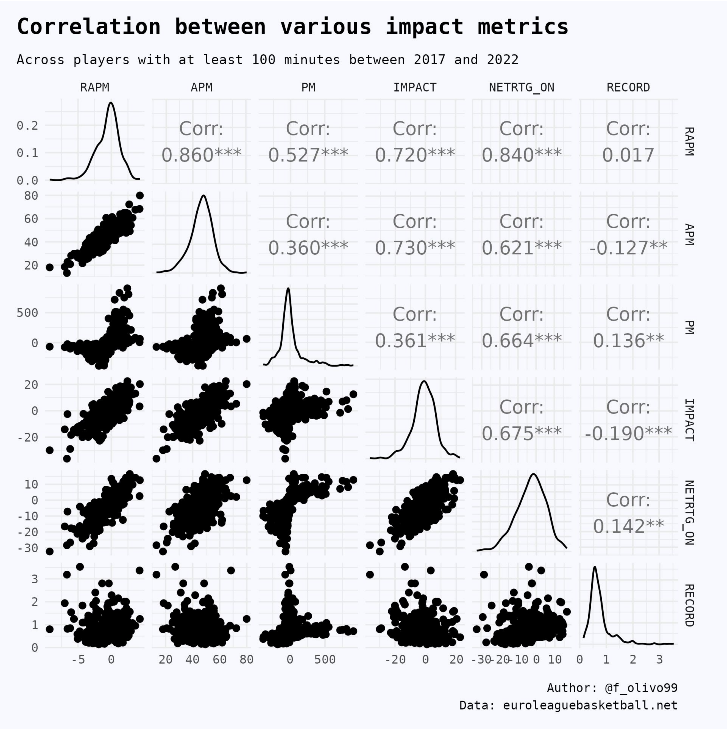
Regularized Adjusted Plus Minus

Regularized Adjusted Plus Minus is an improvement of APM based on a regularized linear regression, in particular on Ridge Regression: it is a technique which overcomes the problem of multicollinearity between variables and prevents overfitting, which means that it is perfect for our goal.

It is an extension of the traditional linear regression method by adding a penalty term, known as the shrinkage parameter, to the least squares objective function. The shrinkage parameter, represented by lambda, controls the magnitude of the coefficients by shrinking the coefficients of less important variables towards zero. This results in a model more robust to noise in the data.

We can find a suitable value for lambda by using cross validation, in particular by trying to minimize deviance. It is fundamental to set `alpha=0` in order to obtain ridge regression.

Results



Correlation between various impact metrics

Across players with at least 100 minutes between 2017 and 2022

Author: @f_olivo99
Data: euroleaguebasketball.net

As we can see, the plot displaying APM vs minutes played shows a very strong funnel shape due to the high variance, which is extremely reduced in the RAPM case.

It is also interesting to note how RAPM is less correlated with record than traditional +/-. As I mentioned in the introduction, it is fair that a good player makes a team good, and similarly a good team has good players, which leads to a certain correlation. Yet, it's quite common to have good players in bad time and vice versa, which justifies a medium correlation.

Improvements

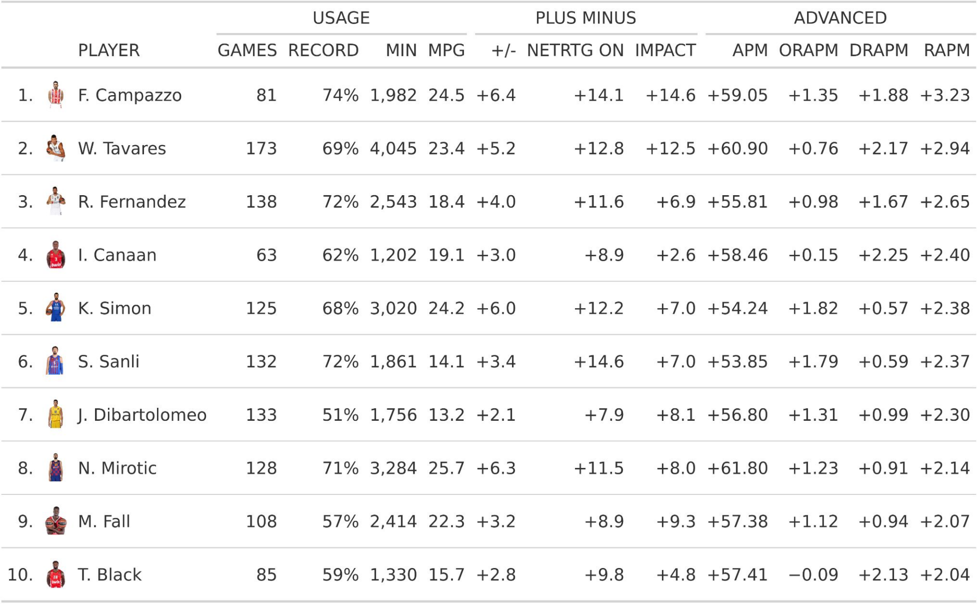
This model works fine, but I think that we could add a few tweaks to make it more insightful: in fact, at the moment each possession has the same weight, which is not the case in basketball. In fact, there are relevant possessions, such as playoffs games or tied games, and the so-called "garbage time", when the difference between the two teams is so big that events during this time are not that important.

At the same time, having play-by-plays allows to segment offensive and defensive RAPM, to evaluate the impact each player has on both sides of the court separately.

We can do this by recreating our dataframe, so that it has 2 new columns, one for the game_type (2x weight if the game is a playoffs game), and one for game phase, namely 0.5 if the action is during garbage time (20 or more points of difference in the 4th quarter), 1.5 if the action is in the clutch (namely, overtime or less than 5 points of difference withing the last 3 minutes of te 4th period), 1 otherwise.

The other difference is that the same stint is repeated twice in the dataset, one time for the home offense and one time for the away offense. We also have twice as many columns, one for the offense and one for the defense, in fact for every player there are the columns `player,offense` and `player,defense`. At the end of the computation we will split offense and defense for each player.

Results after improvements



	PLAYER	USAGE				PLUS MINUS			ADVANCED			
		GAMES	RECORD	MIN	MPG	+/-	NETRTG ON	IMPACT	APM	ORAPM	DRAPM	RAPM
1.	F. Campazzo	81	74%	1,982	24.5	+6.4	+14.1	+14.6	+59.05	+1.35	+1.88	+3.23
2.	W. Tavares	173	69%	4,045	23.4	+5.2	+12.8	+12.5	+60.90	+0.76	+2.17	+2.94
3.	R. Fernandez	138	72%	2,543	18.4	+4.0	+11.6	+6.9	+55.81	+0.98	+1.67	+2.65
4.	I. Canaan	63	62%	1,202	19.1	+3.0	+8.9	+2.6	+58.46	+0.15	+2.25	+2.40
5.	K. Simon	125	68%	3,020	24.2	+6.0	+12.2	+7.0	+54.24	+1.82	+0.57	+2.38
6.	S. Sanli	132	72%	1,861	14.1	+3.4	+14.6	+7.0	+53.85	+1.79	+0.59	+2.37
7.	J. Dibartolomeo	133	51%	1,756	13.2	+2.1	+7.9	+8.1	+56.80	+1.31	+0.99	+2.30
8.	N. Mirotic	128	71%	3,284	25.7	+6.3	+11.5	+8.0	+61.80	+1.23	+0.91	+2.14
9.	M. Fall	108	57%	2,414	22.3	+3.2	+8.9	+9.3	+57.38	+1.12	+0.94	+2.07
10.	T. Black	85	59%	1,330	15.7	+2.8	+9.8	+4.8	+57.41	−0.09	+2.13	+2.04

The results provide some interesting insights: the top 3 includes some of the best Euroleague players ever, but, beside Mirotic, the other 6 players are role players, leaving out of the top ten players like Clyburn (12th), Vesely (13th), Micic (21st) or Vezenkov (53rd)

The new approach allows us to determine on which side of the court the player was more impactful: if it’s not a surprise to find Tavares (3 times Defensive Player of the Year), I would have not expected Canaan and Black to have such defensive values, since they are good defenders but without the same reputation of Tavares. Moreover, Black is the only player of the top 10 having a negative impact on one side of the court, since his ORAPM is negative.

Conclusions and Future developments

Overall, I am satisfied with the results: there are some unexpected values, specially among low usage players, but overall I think that the model effectively managed to measure the impact of players, despite what the boxscore tell. In fact, the insight of this metric is not how much does a player contributes to his team given his points or his defense, but rather his contribution to his team regardless of the teammates he shares the court which.

I think that the results are valid enough to be used as a base for tuning other NBA-oriented metrics to NBA basketball, by finding how are certain metrics correlated with RAPM, and thus the relevance they have in the Euroleague rather than in the NBA.

A next step could be to include prior knowledge about players, such as their boxscore stats, age and role.

Also, I aim to develop an unsupervised clustering method to segment players according to their role on the court, and computing RAPM by cluster rather than by player. This could also lead to evaluating lineup fit according to the available players in a roster.