

Kyle Lew
Dr. Bodwin
DATA 301
Exam 2: Alternate Option for Take Home
Data can be found here: [Colab Notebook](#)

How Data Science can be used in fantasy basketball

Introduction:

Around the world, it's pretty hard to find something with a greater universal love and passion for than sports. In America specifically, major sports leagues have grown into multi-billion dollar industries. The NBA is one of these major sports leagues that has a huge following and therefore is a prime breeding ground for a lot of interesting and diverse data. As such, this is where I decided to focus my research project.

More specifically, I decided to focus on the world of fantasy basketball. To give a quick overview of what fantasy basketball is, it's a fun activity to do with a group of friends in which everyone acts as a manager to a "team" of real NBA players. Each week managers face off against one another in which each team scores fantasy points to see which team comes out on top. Although there are many different variations of formatting fantasy leagues, I focused on the head to head matchup format and looked at two different ways of scoring.

Points Scoring:

This type of scoring is fairly simple and intuitive. Each game statistic is worth some amount of fantasy points, accumulating to a "fantasy score" for each NBA player based off of their stats from any given game. Each week, the manager in each matchup that accumulates the most fantasy points with their team over the week wins. For my research, the point tallies I used were as follows:

$$+ (\text{points} * 1) + (\text{rebounds} * 1.2) + (\text{assist} * 1.5) + (\text{steals} * 3) + (\text{blocks} * 3) - (\text{turnovers} * 1)$$

The drafting strategy for this type of scoring is simple, just draft the players who you think will get the most fantasy points throughout the season (or will have the highest fantasy point average per game).

Categories Scoring:

This type of scoring is slightly more complicated. There are 8 categories of various statistics. Each week in each matchup, teams accumulate total statistics in each of these categories and the manager whose team "wins" the majority of categories wins the matchup. The 8 categories I used were as follows:

Overall Field Goal %, Free Throw %, 3 Points Made, Rebounds, Assists, Steals, Blocks, Points

The drafting strategy for this type of scoring is more involved. You want to draft players based off of which statistics they specialize in. A common strategy for these types of leagues is try to

make a team that collectively performs well in certain categories over others. This type of team tends to perform better than a team that is completely balanced over all 8 categories.

All data taken from the balldontlie API

Pulling/Cleaning player data:

- Pulled player data(height, weight, name, id, etc)
- Pulled 2020 season stat averages
- Transformed the data into dataframes
- Used an inner join to merge the two dataframes (inner join to filter non-current players)
- Created columns for fantasy point per game average, fantasy point season total, per36 stats(averaged statistics by per 36 minutes played)
- Transformed height and minutes played into numerical variables

Clustering Model:

At this step, with the complete dataframe I had, I decided to fit a clustering model with 4 groups. My goal was to get archetypes for NBA players that would help fantasy managers create a plan when trying to draft a balanced team in a categories scoring league. For this model, I created a pipeline with a standard scaler and kmeans clustering model. I fit this pipeline on the 8 stats I mentioned above(except I fit on the per36 stats instead of the pure averages). I chose to use the per36 stats to get a sense of playstyles unbiased by different team schemes. After I fit my model, to get a sense of these groups, I output the means of various statistics along with the minutes played and height for each group. That output can be seen here:

	fg_pct	ft_pct	fg3m	reb_per36	ast_per36	stl_per36	blk_per36	pts_per36	min(sec)	height
type										
0	0.6	0.7	0.4	10.7	2.2	0.9	1.3	16.9	1215.8	6.9
1	0.5	0.8	2.3	5.4	5.4	1.1	0.5	22.9	1949.7	6.5
2	0.2	0.2	0.2	7.6	2.3	2.1	0.6	6.2	432.4	6.6
3	0.4	0.8	1.1	5.3	3.7	1.1	0.5	12.8	1237.7	6.5

As I explored the differences of each group. I decided to give each of them a nickname in my head to better wrap my head around what each group represents. Here are the 4 types:

“The Big Man”: type 0 in the table

This is the tall guy on the team. The guy that gets the rebounds, sets the screens, protects the rim, and all the other dirty work expected of a role playing big man. This can be seen in the table of means as type 0 leads in rebounds and blocks, and bodes the highest average height. Some notable players that fit into this type are Joel Embiid, Anthony Davis, and Domontas Sabonis.

“The GUY”: type 1 in the table.

This is your main guy on the team. The guy who takes the ball down the court, the guy who facilitates the offense, the guy who you want taking the last shot, and probably the guy who scores most of your teams points. This can be seen in the table of means as type 1 leads in minutes played, points scored, assists, and 3 pointers made. Some notable players that fit into this type are Steph Curry, LeBron James, and Jayson Tatum.

“The trash minutes guy”: type 2 in the table

For a lack of a better term, this is the guy on the team who only gets playing time in a “trash minutes”. Maybe a younger guy who the coach is taking a chance on, or maybe an older guy who’s main purpose is to be a locker room presence. This can be seen in the table of means as type 2 shows the least minutes played, points scored, and 3 pointers made.

“The role player”: type 3 in the table

This type of player is crucial to every team. This is the guy who keeps the ball moving, find the open space, takes the catch and shoot threes, and grinds his matchup on defense. In the table of averages, type 3 is right in the middle of most of the stats.

Type Classification:

To better get a sense of these types, let’s say my friend got drafted in 2022. He is a 7 foot center and had an breakout rookie year, averaging 10 points, 15 rebounds, 3 assists, 2 blocks, and 0.5 steals per game. Using a KNeighbors Classifier model on my friend’s height and stats, it is predicted that this hypothetical player is a “Big Man” type player.

Let’s look at some of the top NBA players and what types they are: (sorted by average fantasy points)

	first_name	last_name	team.full_name	height	type
132	Luka	Doncic	Dallas Mavericks	6.583333	1
145	Joel	Embiid	Philadelphia 76ers	7.000000	0
246	Nikola	Jokic	Denver Nuggets	7.000000	1
15	Giannis	Antetokounmpo	Milwaukee Bucks	6.916667	0
117	Anthony	Davis	Los Angeles Lakers	6.833333	0
237	LeBron	James	Los Angeles Lakers	6.666667	1

How to draft your team (in categories scoring) based on these types:

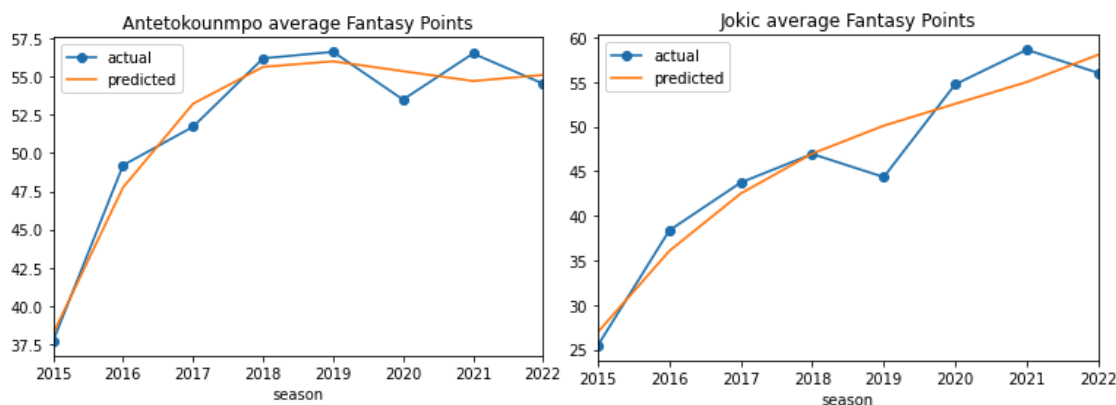
What I found in my data exploration was that the vast majority of top fantasy scoring players are either type 0 or type 1 as can be seen in the table above. As I stated above, one of the best strategies when drafting your team in categories scoring is to focus on a team specialized in certain categories. So next time you are drafting a team in a categories scoring league, firstly I would suggest with your first few picks you take who you think the objectively “best” player is. Then, after the first few picks, based off of what player types you already have, try to draft players in the same player type you already have. For example, if your first 3 picks were all type 0 “Big Men”, I would try to continue this drafting strategy, picking more players that fit into this “Big Men” archetype.

Jokic vs Antetokounmpo:

Nikola Jokic and Giannis Antetokounmpo, two of the most dominant and talented players in recent NBA. The last two NBA MVP awards went to Jokic, and the two before that went to Antetokounmpo. As you can imagine, Jokic and Antetokounmpo have been and still are two of the most valuable players in fantasy basketball. As such, I wanted to explore which player would be better to draft in the next NBA season.

Predicting 2023:

Using a similar techniques as I did for the complete player data, I created dataframes for both Antetokounmpo and Jokic's season averages from the 2015-2022 seasons. I then created a pipeline that created a regression model with polynomial features(with degree=4 to match the shape of the plot) and fit this pipeline onto my Antetokounmpo and Jokic dataframes. Here I plotted my regression model on top of the actual data.



Then, using these regression models, I was able to predict Antetokounmpo and Jokic's average Fantasy Points for the 2023 season.

I predicted Antetokounmpo to average **57.57** fantasy points per game in the 2023 season

I predicted Jokic to average **62.40** fantasy points per game in the 2023 season

So, if you are in a points scoring league for the 2023 season, if both players are still on the board, I would suggest to draft Jokic before Antetokounmpo

Game Predicting:

Let's say you took my advice and drafted Jokic next year. I wanted to create a model that could predict how many fantasy points Jokic would score on any given game night.

Pulling/Cleaning game data:

- Pulled game specific data for Antetokounmpo and Jokic from 2015-2022
- Pulled team data for all 30 NBA teams
- Transformed them into dataframes
- Used the team data to add opponent data into my main dataframe
- Created a column for fantasy points per game
- Created a column for time of year(calendar date)
- Dropped rows with missing values(these rows are games the player didn't play in)

Prediction Model:

For my variables, I decided to have explanatory variables be the calendar date, season, and opponent, and my response variable be fantasy points. (I chose to include calendar date to

account for the time in the season the game happens) For the calendar date and season variables, I applied a standard scaler and for the opponent variable a OneHotEncoder transformer since this variable was categorical. I created a pipeline with these column transformations, and a KNeighborsRegressor model.

Now to put this model to use, let's so you and your buddy are matched up against each other this week. It's the first of January and you guys are tied in fantasy points. You have Jokic and your team and your buddy has Antetokounmpo, and as it turns out, the Denver Nuggets(Jokic's team) and the Milwaukee Bucks(Antetokounmpo's team) are playing each other. Using my model, I predict Jokic will score **41.1** fantasy points and Antetokounmpo will score **39.66** fantasy points. Good job taking my advice and picking Jokic over Antetokounmpo 😊

Conclusion:

Data Science absolutely can be used to help you get an upper hand in your fantasy basketball league. Considering how much data flows through every walk of life, it's impressive how useful data science techniques can be to understand the world on a deeper level.