NBA Position Modeling with College Basketball Player Statistics

Abstract

With more college basketball statistics readily available online, it provides the perfect opportunity to use that data to create models to benefit professional basketball teams in the National Basketball League. Player positions are one of the key components to a well-rounded basketball team, so this project is focused on predicting a player's NBA position, like center, using one of their college player statistics, like blocks per game. Iterating over all possible position and college player statistics combination, models have maximal accuracy of 86.73% and generally explained by the key components of each player position.

Background

In the age of technology, the amount of National Basketball Association (NBA) and college (NCAA) player statistics online grows even more each year. This presents a great opportunity to analyze trends and come up with models. This project focuses on player positions/roles using this data.

The National Basketball Association (NBA) is the men's professional basketball league in North America, composed of 30 teams. The NBA has an annual draft, called the NBA draft, which NBA teams can draft eligible players into the league. These players are typically college players in North America, but international players are also eligible to be drafted. However, my project only considers players who were drafted into the NBA and played college basketball full-time.

The game of basketball is rather complex with five players on each team competing on the court to ultimately make more points, by scoring a ball into a basket, than their opponent. The five players on the court play the following unique positions: point guard, shooting guard, small forward, power forward, and center. For example, the point guard usually handles the ball and is often the shortest on the team. The center is the big man, tallest on the team, in charge of playing defense near the basket and blocking other people's shots.

There are many statistics measured about a player's gameplay, like their field goal percentage, three-point shot percentage, rebounds, assists, and so on. In college basketball, player position is less-defined, with only three general positions: guard, forward, and center. Even though generally guards end up being guards in the NBA, it is not obviously clear what guards end up becoming point guards or shooting guards.

More generally, is there a better way to predict what positions college players end up playing in the NBA? Can a player's statistics in college be used to predict their player position in the NBA? This project tries to tackle this question by creating a regression model that uses a player's college statistics, like points per game in their college basketball career, to predict if that player played center position in the NBA.

Significance

To be a well-rounded team, an NBA team should have a good balance of players who play different positions. So when an NBA team drafts players in the NBA draft, they need to consider the positions that that drafted player can play. Even more so, teams should know which positions a player is most fit to play before they draft that player.

The model created in this project can also be extended to using NBA player statistics to confirm their fit of the current position they play. However, teams' main priority every year is to draft the best up-and-coming players that best fit their team's needs, which is centered around the positions players can play and their capacity to play that position well. Teams also pay their top drafted players millions of dollars and so their decisions are also financially incentivized.[1]

Methods

Even though data from players are readily available online, there is no easily downloadable source of NBA/college basketball data. A web scraper, written with BeautifulSoup in Python, was used to scrape Basketball Reference for NBA player data and Sports Reference College Basketball for college player statistics.[2] [3] Over 5000 different NBA players' information and NBA statistics were scrapped. Specifically, players that were drafted in 1980 or later were used in this project.

Then, player's corresponding college basketball statistics on Sports Reference were scraped, and their NBA and college basketball statistics were matched together using their full name and draft year. This automatic matching process was not perfect and required manual matching of over a hundred different players, who changed their names or were incorrectly inputted in Sports Reference.

For the model, a logistic regression was used to predict if a player played a certain NBA position, like if a player played center in the NBA, using a player's certain college player statistics, like a player's average blocks per game over their entire college season.

To see if the model was good (accurate), performance was calculated in two methods. One was accuracy: what percentage of players did the model predict their NBA positions correctly? And the other was kappa, based on the Kappa coefficient: how much better is our model than a fully randomized model?[4]

Models were created for every pair of possible college player statistics (24 different statistics) and NBA player position (5 different player positions) for a total of 120 different models. Each model's performance was calculated via the two methods above.

Results & Discussion

Overall, the model results made sense and could be explained with knowledge of specific player positions. For example, for point guards, one of the best models to predict if a player's position in the NBA is point guard is to look at their assists per game college statistics. This can be explained because as the player who handles the ball the most, the point guard is usually the one who passes to the open shooter for the shot or to the center for a dunk. This specific model had an accuracy of 86.73% and kappa of 63.13%.

Another example is, to predict if a player's position in the NBA is a center, some of the better explanatory variables are blocks per game (accuracy 79.94% and kappa 43.63%) and 3-point field goal attempts per game (accuracy 75.04% and kappa 41.68%). The former is straightforward as centers are usually the tallest on the court and thus, block a relatively high number of people. The latter is more interesting because the model also takes into account if a player generally does the worst in a category. Centers rarely shoot threes and thus their number of 3-point attempts are low, so it makes sense that a low number three-point attempts can be an identifier of a center.

Are these models good? For a completely random model, which randomly guesses if the player plays that specific NBA position or not, its accuracy will be around 50%, since a player either plays or does not play that position. Thus, accuracies being over 50% is a good sign. Not only does a high accuracy matter, but a high kappa also does. Even if the model was better in accuracy in the completely random model, how much better is it than the completely random model?

The highest accuracy model had an accuracy of 86.73% (the point guard model above) and the highest kappa model had a kappa of 63% (also the point guard model above). Even though the accuracy is higher than 50% and kappa greater than 0%, this shows that the models are far from perfect and can be improved greatly to better predict NBA positions.

For small forward and shooting guard roles, the kappa coefficient is really lacking, where even some of the metrics are negative (worse than a random model). This is because guards and forwards in college basketball tend to have the traits of a point guard and power forward in the NBA respectively, hence the better models for point guards and power forwards. This yields the question, how do NBA teams successfully recruit for these positions?

With all these models, comparisons can be made by comparing which college statistics are the best for predicting if a player played a specific NBA position. Moreover, rankings of which NBA player position is most predictable by a certain college statistics can be made. Full accuracy and kappa value tables can be found in the appendix (A1, A2).

Conclusions

Even though the models were not the most accurate, the logistic regression models provided for a lightweight method to predict NBA positions from a singular college player statistic. The models were easily explained by the way basketball is played by certain player positions, and more can be extrapolated from the results of the models.

Future work can be done to continue to improve the models. Using multiple college statistics could better predict player position, by creating a more specific fit for an NBA position. Segmenting the player data by year and training separate models for different periods of time may be beneficial as many trends are very much based on time, as can be seen in the appendix (A3).

References

[1] https://www.cnbc.com/2018/06/21/how-much-the-2018-nba-drafts-first-pick-will-earn-as-a-rookie.html
[2] https://www.basketball-reference.com/
[3] https://www.sports-reference.com/cbb/
[4] https://en.wikipedia.org/wiki/Cohen%27s_kappa

Appendix

A1. Accuracy tables

	College Statistic	Center	Point Guard	Power Forward	Shooting Guard	Small Forward
1	2−Point Field Goal Attempts Per Game	49.22	54.03	57.92	50.07	54.95
2	2−Point Field Goal Percentage Per Game	67.47	64.00	61.10	57.14	48.94
3	2−Point Field Goals Per Game	55.51	57.27	61.23	52.05	54.87
4	3−Point Field Goal Attempts Per Game	75.04	70.51	67.61	69.09	54.38
5	3−Point Field Goal Percentage Per Game	76.16	53.10	70.43	53.64	58.36
6	3−Point Field Goals Per Game	74.36	70.62	66.60	68.57	54.24
7	Assists Per Game	72.39	86.73	64.34	64.80	47.79
8	Blocks Per Game	79.94	71.69	67.45	59.43	42.98
9	Defensive Rebounds Per Game	65.22	67.80	67.12	56.39	58.83
10	Field Goal Attempts Per Game	64.52	56.39	55.27	59.08	52.35
11	Field Goal Percentage Per Game	71.64	67.60	63.57	60.76	49.61
12	Field Goals Per Game	57.17	49.83	51.12	55.04	53.20
13	Free Throw Attempts Per Game	49.61	53.36	56.67	49.83	48.15
14	Free Throw Percentage Per Game	71.19	62.61	63.73	58.74	47.25
15	Free Throws Per Game	55.49	56.95	47.37	53.36	48.49
16	Games	55.83	50.00	53.53	48.77	52.13
17	Games Started	61.97	50.00	58.18	45.12	45.53
18	Minutes Played Per Game	68.90	59.61	60.53	51.81	46.70
19	Offensive Rebounds Per Game	68.89	75.95	70.52	62.09	55.43
20	Personal Fouls Per Game	58.16	58.16	57.42	61.13	51.94
21	Points Per Game	61.21	53.14	52.97	56.39	52.24
22	Steals Per Game	72.54	74.89	60.81	61.67	53.72
23	Total Rebounds Per Game	65.70	77.35	70.01	60.99	58.58
24	Turnovers Per Game	60.85	68.93	56.49	50.71	49.41

A2. Kappa tables

	College Statistic	Center	Point Guard	Power Forward	Shooting Guard	Small Forward
1	2−Point Field Goal Attempts Per Game	1.37	8.14	9.83	2.62	3.66
2	2−Point Field Goal Percentage Per Game	23.25	21.84	16.30	11.90	−1.02
3	2−Point Field Goals Per Game	3.98	14.34	14.94	6.77	2.76
4	3−Point Field Goal Attempts Per Game	41.68	31.56	32.24	29.49	2.64
5	3−Point Field Goal Percentage Per Game	28.47	13.26	23.94	14.37	−0.92
6	3−Point Field Goals Per Game	40.57	31.31	31.20	28.37	0.31
7	Assists Per Game	36.46	63.13	27.35	16.08	4.72
8	Blocks Per Game	43.63	37.60	20.36	22.28	0.74
9	Defensive Rebounds Per Game	18.73	31.42	24.64	14.46	7.73
10	Field Goal Attempts Per Game	19.71	8.96	7.34	12.91	2.85
11	Field Goal Percentage Per Game	30.93	28.53	20.17	17.84	1.44
12	Field Goals Per Game	8.35	−1.36	0.94	6.24	3.15
13	Free Throw Attempts Per Game	2.37	0.83	7.54	3.44	0.45
14	Free Throw Percentage Per Game	27.24	20.21	18.58	16.20	0.04
15	Free Throws Per Game	8.08	6.30	−1.70	1.95	0.71
16	Games	3.65	3.65	0.58	0.34	0.45
17	Games Started	8.29	7.36	4.70	−0.44	1.55
18	Minutes Played Per Game	22.24	17.24	10.20	7.11	0.86
19	Offensive Rebounds Per Game	25.87	44.42	32.54	22.21	4.90
20	Personal Fouls Per Game	11.03	10.50	12.21	16.71	3.88
21	Points Per Game	13.59	3.67	3.02	8.81	3.03
22	Steals Per Game	35.48	37.51	20.35	13.28	1.46
23	Total Rebounds Per Game	20.32	46.79	31.22	19.07	10.88
24	Turnovers Per Game	15.27	27.10	9.65	−1.03	−0.74

A3. Average Total Season 3-Point Field Goals over Season. This shows that 3-Point shots were not very popular until the early 2000s and thus, models will be different if solely trained on more recent player data.