Investigating the relationship between shot accuracy and distance in comparison to different player positions

Vaida Gulbinskaite

Principles of Data Science - City University of London 2016

ABSTRACT: Big data is getting increasingly popular not only in individual but also in team sports. Since the 2013's big data revolution in NBA, advanced statistical analysis have influenced game strategy in some NBA teams (e.g. Houston Rockets). In this paper I looked at the detailed shot log data from the NBA season 2015 and investigated how distance from the basket impacts shot accuracy. I have found that shooting guards are slightly better at long distance shots and centre players are slightly worse at long distance shots in comparison to all players. I have also concluded that number of shots shooting guards attempted was potentially the same, if not higher from the close proximity to the basket as further away from the basket.

1. Introduction

The recent rise of big data and data science opened the door to new possibilities not only to individual athletes but also team sports. Basketball is not an exception. In 2013, the NBA has implemented a new, state of the art statistical analysis technology – SportsVU (Official NBA release 2013). Since then, basketball data is becoming more in-depth and more accessible to the curious basketball players, team managers and fans alike. Whereas before the big data revolution, team managers had to rely on scout observations and box statistics in order to pick players, now with the adoption of SportsVU, they are able to see every single move a player makes from entering the court to exiting it as a victor or a loser.

I have chosen sports as a domain, because of the increasing granularity and availability of the data. As a Lithuanian, I take pride in being an enthusiastic basketball fan (after all, basketball is often regarded as 'Lithuania's second religion'). For that reason, I have decided to look into basketball data. I have selected a dataset from the 2015 NBA season. This season was not an ordinary season as It was a year after NBA has implemented more advanced statistical data analytics (Official NBA release 2013). Some of the teams, decided to incorporate data insight into the organisation of their team (e.g. Houston Rockets), while others remain sceptical (e.g. LA Lakers, which have been listed at the bottom of The Great Analytics Rankings (Harskamp, 2013). Daryl Morey (general manager of Houston Rockets) is a pioneer in using basketball data analytics in his strategy. He has re-defined the game by introducing 'MoreyBall' - a substitution of long 2 point shots to 3 point shots due to the perceived low value of long 2 point shots (Partnow, 2016).

Basketball observations have encouraged the publication of various scientific articles and journalistic data science pieces. Reich et al (2006) studied the position of the shot and developed hierarchical spatial models for shot-chart data, allowing spatially varying effects of covariates. By looking at shot distances, Goldsberry (2014) has identified that although basketball court 4.700 Square feet, the vast majority of shots (98%) were taken within the 1,300 square feet – "that spans between the baseline and a relatively thin buffer around the 3-point line". Liu and Burton (1999) investigated the effect of distance on shooting accuracy. The researchers have found that when asking people with hardly any basketball experience to shoot a basketball, they get significantly worse as the distance from the basket increases.

1.1 Change in analysis strategy

In my progress report, I was considering to investigate how multiple factors impact shot accuracy. However, by looking at the dataset and considering the length of the report, I have realized that I might have been too ambitious. Therefore, I decided to select one factor and try to investigate it in a more granular level.

1.2 Analysis Goals:

In this data science piece, I will investigate the distribution of shots made in respect to distance from the basket. I will also try to investigate how the success of the shot depends on its distance from the basket. By looking at various articles, I have learned that there is not much research done with regards to comparing types of basketball player positions, therefore I will compare the results between two quite opposite types of basketball player positions: 1. Centre; 2. Shooting Guard.

My hypothesis is that overall basketball players would be more accurate at scoring within close proximity to the basket, similar effect as recorded in (Liu and Burton 1999) but not as significant. However, when controlled for player position, centre players, who are usually tall and play heavily in defence within the close proximity to the basket, will be better at scoring within close proximity to the basket, whereas shooting guards, who are shorter, faster and are expected to play further away from the basket, will be better at scoring further away from the basket. I will also very briefly look into "MoreyBall" (Bonus analysis).

2. Data

2.1 Collection

In order to perform this analysis, I have used 2 datasets. I have selected the first dataset from (Kaggle.com) (ref: Kaggle dataset), which contains data of all the players who took a 2 or 3 point shot during the 2014 - 2015 season (penalty shots excluded). The data also contains where was the shot taken from (distance from the basket), shot result (made or missed), the nearest defender, distance, time on the shot clock and a few more. I have created the second data set by scraping the data from (Basketball-Reference.com) (ref: scraped dataset). I have used a python module called Beautifulsoup (Crummy.com) in order to scrape the data. The scraped dataset contains the average game statistics for each basketball player during 2014 - 2015 season as well as other general information, e.g. age, position and team played.

2.2. Data Preparation

While merging the datasets, I have noticed misspelled player names in the Kaggle dataset. I have decided to create a function that corrected players' names. The defender names were in a format of surname + name rather than name + surname. I have decided to make the names uniform (same format as player names) and therefore split the closest defender column in excel into two columns: 1. Closest defender name, 2. Closest defender surname. I have then joined the columns together in python in a format of name + surname. The closest defender names were also misspelled the same way as player names, for these, I have applied the name correcting function.

I have noticed that some basketball players played for more than one team during the NBA 2015 season. At first I thought to look at the date of the game and assign the team based on where the player played at the time. Investigating this, I have noticed that most of these players played more games in their original team and not their transfer team. For example: Gary Neal was transferred to Minnesota Timberwolves in February 2015 (transfer team), however he only played 11/54 games, therefore I have assigned him to the previous team - Charlotte Bobcats/Hornets (original team), where he played 43/54 games (Basketball-Reference.com).

Lastly, I have decided to normalize string values in order to create binary variables and possibly use them for logistic regression analysis. I have normalized d shot result variable (missed= 0, made = 1).

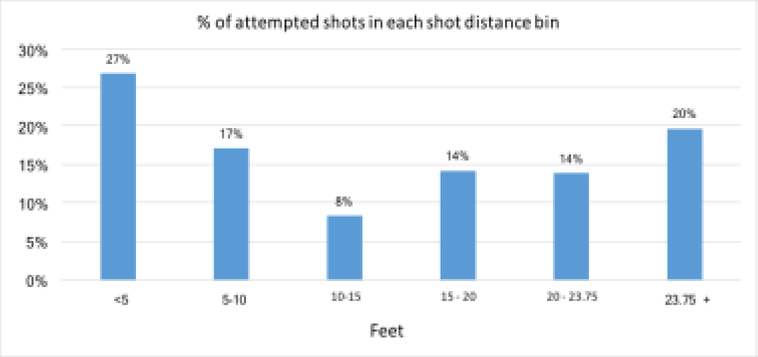
3. Data analysis

3.1 Software used for the analysis

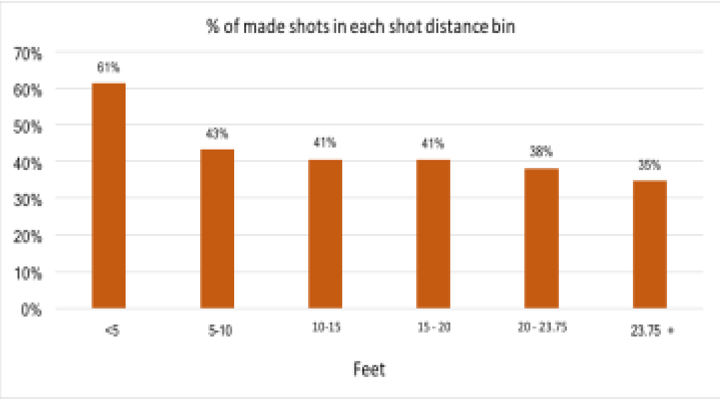
There were a few pieces of software used for the analysis. Python (Jupyter notebook), along with libraries such as pandas, numpy, scipy that were used or data wrangling. Python library statsmodels.api was used in order to perform the correlation analysis. SPSS statistics package (version 23) was used to perform the logistic regression analysis. Descriptive statistics tables were created in python, using pandas' pivot tables. Visuals (charts) were created in excel and SPSS.

3.2 Descriptive statistics

There were 151, 889 shots attempted during the 2015 NBA season, out of which 68,287 (approximately 45%) were successful. 3 point shots accounted for approximately 20% of all the shots attempted and 15% of all shots made. In order to investigate shot percentages further, shot distance variable was binned into 6 bins. Bins 5 and 6 were split at 3-point line - 23.75 feet (Official rules of the National Basketball Association) in order to present a clearer split between 2 point shots and 3 point shots (Graph 1 and Graph 2)



Graph 1: *percentage of attempted shots in each shot distance bin*



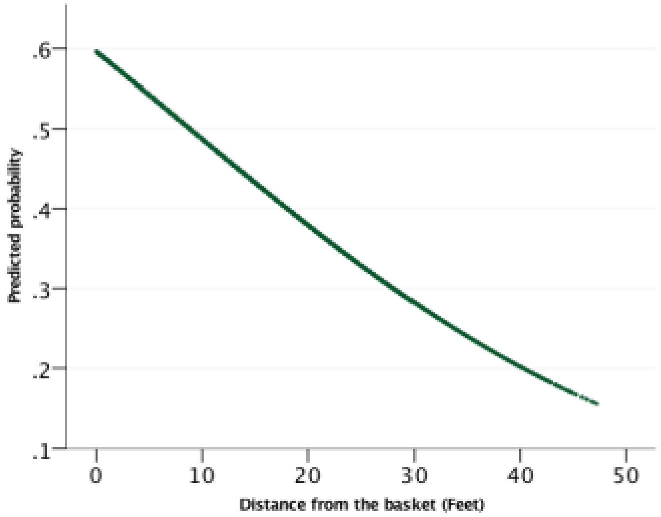
Graph 2: *percentage of made shots in each shot distance bin*

3.3 Inferential statistics

In order to investigate the relationship between the number of shots made and their distance from the basket, Pearsons correlation was performed. A strong, negative correlation r = -.59 (p<.00) was observed between made shot percentage and distance. Pearsons correlation was also performed in order to investigate differences between players who play in centre and shooting guard positions. A strong, negative correlation between distance and number of shots made was observed for centre players r = -.72 (p<.00), on the other side, a weak, negative correlation was observed for shooting guards r = -.18 (p<.00).

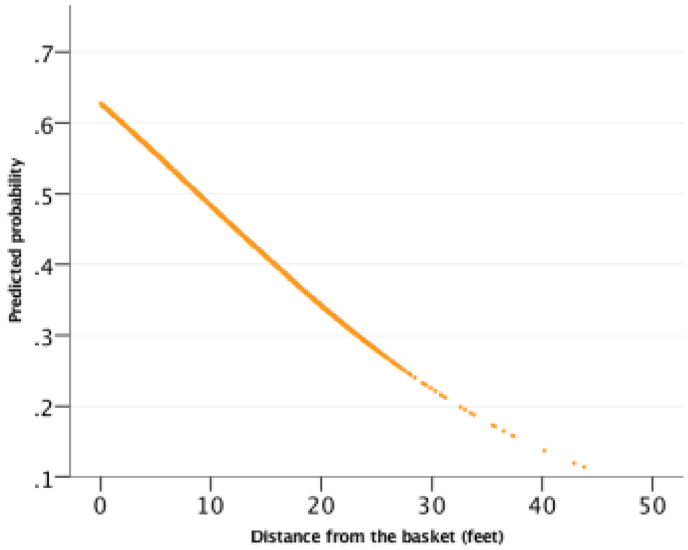
3.4 Logistic regression

Binary Logistic regression was performed in order to determine the probability of shot being made, based on its distance from the basket. Negative regression (B =-.044, p<.00) with odds ratio of .957 was observed. The model predicted 59.4% of the values. In order to identify the variance of the dependent variable (Shot result) can be predicted with the shot distance, Nagelkerke R Square was performed. It showed that less than 5% of the variance could be attributed to the shot distance variable (Chart 3).

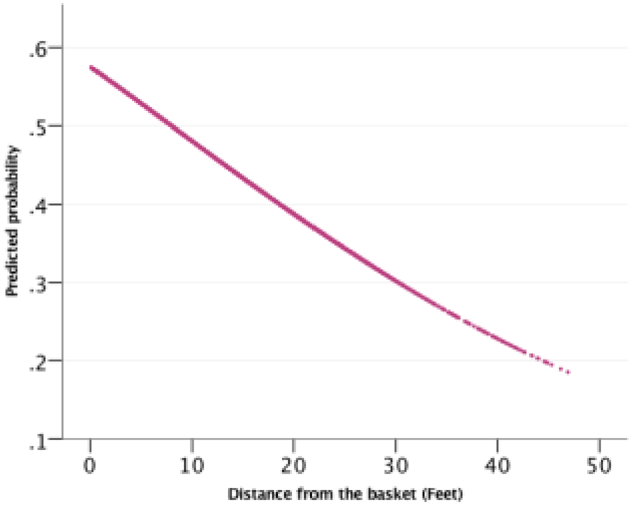


Graph 3: *Estimated probability of shot being made based on its distance from the basket (All players).*

Binary logistic regression was also performed in order to investigate the differences between the two types of player positions (Centre and shooting guard). A negative regression (B =-.059, p<.00) with odds ratio of .943 was observed for the centre players. It was estimated that the regression model for Centre players could predict 58% of the values. Nagelkerke R Square test showed that shot distance accounted for less than 5% of the variance (Graph 4). As for shooting guards, a negative regression (B =-.038, p<.00) with odds ratio of .963 was observed The regression for shooting guards players could predict 60.2% of the values. Nagelkerke R Square test showed that shot distance accounted for less than 3.5% of the variance (Graph 5)



Graph 4: *Estimated probability of shot being made based on its distance from the basket (Centre players).*



Graph 5: *Estimated probability of shot being made based on its distance from the basket (Shooting Guards).*

4. Reflection

4.1 data analysis interpretation

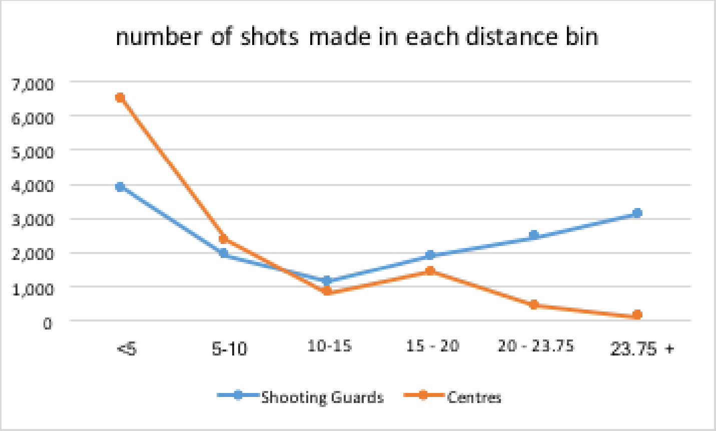
This paper investigated how the distance from the basket impacts the shot as well as how the distance impacts the performance of two types of player positions: 1. Centre - Generally tall, defence – heavy player who generally spend most of their time in a close proximity to the basket; 2. Shooting guard – Possibly the shortest, fastest player who has to be reasonably accurate in making 3 point shots (which suggests that these players mostly play away from the basket).

With descriptive statistics, it was observed that the that majority of the shots (52%) happen between 0 to 15 feet radius away from the basket. It was also observed that the highest percentage of attempted shots happens within the 5 feet radius from the basket, followed by shots, attempted just above the 3 - point line (Graph 1). By looking at the percentage of shots made and observed that shots within the 5 feet radius were made 61% of the time. The percentage of shots made gradually decreased when the distance from the basket decreased, with the largest decline observed between <5 feet radius and 5-10 feet radius (18%) (Graph 2).

Pearsons correlation showed that there was a strong, negative correlation between all shots made and distance away from the basket, which supports my initial hypothesis and partially corresponds with (Liu and Burton 1999): increasing distance has a negative effect on shooting accuracy.

Pearsons correlation was also performed in order to investigate the difference between player positions. Strong negative correlation was observed for Centre players, which corresponds with my hypothesis, that centre players mostly play within close proximity of the basket and hence score significantly more points from the close distances rather than far distances.

A very weak negative correlation was observed for shooting guards I hypothesized that shooting guards will have a positive relationship between number of shots made and distance from the basket, as the nature of their game play dictates that they generally play away from the basket, however weak, negative correlation suggests that Shooting guards make just as many shots within the close proximity to the basket as they do further away from the basket. Graph 6 illustrates a clear difference between the two types of player positions and their shooting style.

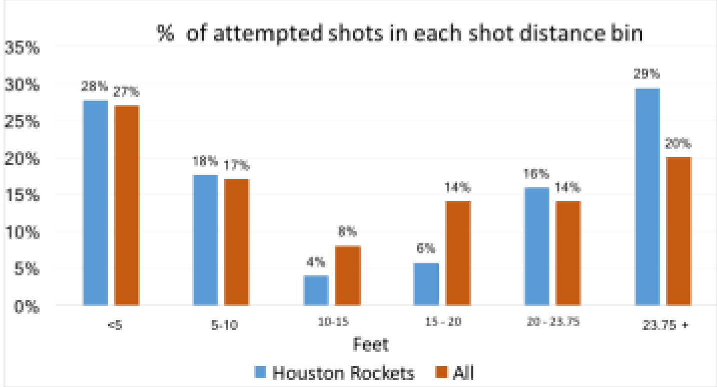


Graph 6: *Number of shots made in each shot distance bins, segmented by type of player position*

In order to look at the effects of distance from the basket on shot accuracy, Binary Logistic regression was performed. Although all models performed in over 50% (Shooting guard - best, Centre - Worst) accuracy shot distance did not account for more than 5% of the variance of the dependent variable (Shot result). The odd ratio suggests that with every extra foot further from the basket, the chances of getting a successful shot decrease by 4.3% for all players, 5.7% for centre players and 3.7% for shooting guards. This somewhat supports my hypothesis as it suggests that shooting guards are slightly better at longer distance shots than all the players and centre players are slightly worse at shooting from long distances than all the players.

In the beginning of the paper, I have briefly talked about “MoreyBall” - a new basketball strategy, adopted by Houston Rockets (Partnow, 2016). It made me curious therefore I have decided to do a bonus analysis and briefly investigate it and compare the percentage of attempted shots in each shot distance bins between all players and Houston Rockets players (Graph 7).

Although the percentage of shots made within 0-10 feet radius is approximately the same and the percentage of attempted shots within the 10 – 20 feet radius decreased in both instances, the percentage of attempted shots dips even lower as the percentage of shots attempted by Houston Rockets In 10 – 15 feet radius decreased by 14% compared to percentage of shots made in 5-10 feet radius. On the opposite side, Houston Rockets makes up by increasing the number of attempted shots above the 3- point line - as the percentage of attempted shots is 9% higher than the percentage of shots, attempted by all the players.



Graph 7: Illustrating MoreyBall: *percentage of shots attempted in each shot distance bins- difference between Houston rockets players and all players*

4.2 Practical use of my research

I believe my tiny data science project could help basketball team managers to alter their strategy in order to utilize centre and shooting guard players better. I have identified that centre players are better at scoring from close distances, which is no surprise to most of the managers, however, shooting guards are potentially just as good at shooting from a very close proximity to the basket as shooting from further away, therefore instead of isolating them to just play far away from the basket, their abilities could also be utilized in a close proximity to the basket.

4.3 Improvements and future research suggestions

I believe that this research could be improved by using coordinates (where exactly shot was attempted in the arena) rather than radius from the basket, as it could give more insight to where shots are easier to be made. Knowing the coordinates, it would be interesting to investigate and assess the difficulty within 2 point shots and within 3 point shots in order to see if for example there are places on a court above the 3-point line that helps basketball players score more accurately.

It could also be interesting to investigate defence. Although the Kaggle dataset had closest defender distance, it was hard to determine defensive strategies (e.g. predict if block happened, look at whether defender made a foul against the offending player). Having this information, it would be interesting to look at points made and defence strategy and compare these between losing and winning team. It would be interesting to see what the losing team tried to change in order to win the game, and how the degree of their rational thinking changed during the game (e.g. Increase in long shots, increased number of fouls made to stop the game every few seconds in the last quarter).

It would also be interesting to look at NBA time series data, in order to see the impact of 2013 NBA big data revolution (Official NBA release 2013) had an impact on the way game is now played and teams are now organised.
Lastly, it would be interesting to investigate “MoreyBall” further. According to (Basketball-Reference.com), during NBA 2015 season Houston Rockets went up from 5-7th place in 2014 to 3-4th place. It would be interesting to see if “MoreyBall” was the reason of their success. I would need to obtain free-throw and defence strategy data (Two very important factors in the game) in order to fully investigate Houston Rockets’ rise in the league.

REFERENCES

[1] Official NBA release (2013) *NBA partners with stats LLC for tracking technology*. Available at: http://www.nba.com/2013/news/09/05/nba-stats-llc-player-tracking-technology/ (Accessed: 10 December 2016).

[2] Harskamp, R. (2013) *The great Analytics rankings*. Available at: http://www.espn.com/espn/feature/story/_/id/12331388/the-great-analytics-rankings (Accessed: 10 December 2016).

[3] Partnow, S. (2016) *MoreyBall, Goodhart’s law, and the limits of Analytics*, VICE sports. Available at: https://sports.vice.com/en_us/article/moreyball-goodharts-law-and-the-limits-of-analytics (Accessed: 10 December 2016).

[4] Reich, B.J., Hodges, J.S., Carlin, B.P. and Reich, A.M. (2006) *‘A spatial analysis of basketball shot chart data’,* The American Statistician, 60(1), pp. 3–12. doi: 10.1198/000313006x90305.

[5] Goldsberry, K. (2014) *Trio Grande Valley: Daryl Morey’s D-League Plan to Do Away With Midrange Shots*, Grantland. Available at: https:// http://grantland.com/the-triangle/trio-grande-valley-daryl-moreys-d-league-plan-to-do-away-with-midrange-shots (Accessed: 10 December 2016).

[6] Liu S, Burton AW . (1999). *Changes in basketball shooting patterns as a function of distance*. Perceptual and Motor Skills, 89(7), p.831.

[7] Kaggle.com. (2016). NBA shot logs | Kaggle. [online] Available at: https://www.kaggle.com/dansbecker/nba-shot-logs [Accessed 14 Dec. 2016].

[8] Basketball-Reference.com. (2016). *Basketball Statistics and History*, Basketball-Reference.com. [online] Available at: http://www.basketball-reference.com/ [Accessed 10 Dec. 2016]

[9] Richardson, L. (2016). *Beautiful Soup: We called him Tortoise because he taught us.*. [online] Crummy.com. Available at: https://www.crummy.com/software/BeautifulSoup/ [Accessed 10 Dec. 2016].

[10] *Official rules of the National Basketball Association,* 2013 -2014. (2013). 1st ed. St. Louis, Mo.: Sporting News Pub., p.9