Analysis of NBA Shot Selection from 2010 to 2017: Understanding the Basketball Meta

David Wong

12/22/2020

Code and data supporting this analysis is available at: https://github.com/waviddong/STA304-Final-NBA-Shot-Analysis

Abstract

In this report, we will analyze how the NBA shot has evolved from 2010 to 2017 and whether or not current discussions regarding the changes in the NBA meta hold ground. We will be looking at a variety of shooting related variables to analyze firstly, whether these changes are correlated to shot success and then perform some exploratory analysis to understand how exactly the NBA shot has changed. From this report, our findings show that there are noteworthy changes in the NBA shot that could possibly indicate changes in the NBA meta, there is strong evidence to show that volume in shooting as well as overall shot accuracy has increased over time. This report will aim to serve as evidence for these claims through implementing a logistical regression analysis and data analysis principles to show and tie together shooting changes over time in the NBA from 2010 to 2017.

Introduction

This report will investigate how both shot selection and shot accuracy has changed from 2010-2017. During this time frame, the basketball meta in the NBA has changed between drastically different meta games. Spectators and fans commonly hear complaints regarding the "good old" NBA days, where posting up in the paint for a 2 PT attempt and a bullish playstyle defined by the big men on teams dictated the pace of the games. Nowadays, analysts claim that basketball is now all about "spacing the floor" and "efficient shot selection" (Jarvis, 2017). These claims will be observed and analyzed in this report. The report will then tie together how the new metagame has contributed to an evolving skill set needed to adapt with the metagame in basketball.

In our report, we will be looking at our independent variables of player id, period of the game, shot type, and action type. These variables will then be implemented in our logistic regression model to understand which shot variable related the best to our explanatory variable, shot made, indicating whether the shot was successful. Furthermore, from our models, we will post stratify each NBA year to predict the overall league shot accuracy over time. Lastly, we will then do some exploratory data analysis by looking at the volumetric changes and shot accuracy changes over time within the NBA league to cross verify the trends that we are seeing in the game. From our analysis of our model and exploratory data analysis, will indicate whether the acclaimed trends of "spacing the floor" to create room for playmaking superstars to select the "efficient shot" is a valid claim. If there is valid evidence that shot accuracy and shot volume have increased over time, we can say that both more shooting opportunities and successful shots have increased over time from 2010 to 2017. Furthermore, from our model diagnostics and analysis, we can also verify each selected independent variable and how it correlates to the success rate of a shot. This report will aim to answer all these questions in regard to the evolution of shooting within the perceived time of metagame transition.

Data

The data is from a user, “ebaek” from GitHub who sourced it from Darren Willman’s NBA Savant data mining API. Darren Willman is the Director for Research & Development for the MLB and has a passion for data analytics, he created this API for open-source tracking NBA data for sport lovers. From the provided data, I cleaned and sorted the data fields by both reclassifying the relevant categorical information into numerical representations. Furthermore, data was cleaned of ‘NA’ values with a total omission of 16,766 data points aggregating up to a total of 1,528,321 total data points in this analysis. The data set that we will post stratify is an aggregate dataset of all NBA shot data from 1997 to 2020 that contains a variety of more data on shot logs. We will use the NBA Savant data mining API shot data to determine sample weights that will be applied to the aggregated from the NBA Stats API dataset found from Data.World. Our NBA Stats API data is 4,729,234 data points in total.

The population would be all NBA players’ shots attempted in the 1997 to 2020 seasons, the frame would be all NBA shot attempts from 2010 to 2017, and the sample would be all NBA shot attempts not omitted from the data due to recording errors from 2010 to 2017.

For the data modelling process, we initially had 72 different action types which were condensed into 4 different classifications of jump shot, layup, dunk or hook shot. These classifications were designated based purely off of the key words in their name of the shot, there were some fringe cases where personal discretion had to be incorporated but it was overall an intuitive classification. Furthermore, it is important to note that the New York Nets used to be the New Jersey Nets until 2012 where ownership changed. Other teams within the NBA Stats API dataset include teams like the Seattle SuperSonics, Charlotte Bobcats, New Orleans Hornets, Oklahoma City Hornets, and Vancouver Grizzlies who existed beyond our desired time frame. During the data modelling process, I included the New Jersey Nets during the years 2010 to 2012 to be its own separate, 31st NBA team. The omission of the team would have resulted in a significant loss in data points for the initial few years and since we are looking at the total shots in the NBA, it would compromise the integrity of our analysis. Other than the New Jersey Nets, older teams have already phased out by then and most of the contemporary 30 teams have held the same name since.

Exploratory Data Visualization

Table 1: Here we are seeing the pairwise correlation coefficients. There are no stand out numbers to take note of as all of these variables seem to have minimal correlation with each other. If there is a real effect of a variable, the probability that variable will be significant is a function of several things, such as the magnitude of the effect, the magnitude of the error variance, the variance of the variable itself etc… Multicollinearity will be further explored in the report.

Looking at Table 1, it is important to note that there does not seem to be much issue with multicollinearity since all variables seem to have low correlation with each other. It is important to assess multicollinearity in a multiple regression model to understand the effect of how each independent variable interacts with each other.

Table 2: As you can see, generally, 3 PT shots are increasingly becoming popular over time as 2 PT shots lose in popularity. This stark increase starts around 2013-2014, which ironically also aligns with the breakout year for Steph Curry and many other superstar shooters in the NBA.

In Table 2, it is important to understand that there is an obvious trend in the volume of 3 PT shooting increasing throughout the years of 2010 to 2017. It is also important to note that although in 2011, there was a significant decrease in overall shooting volume, the NBA was in lock-out which effectively removed 16 games from each team during the season. Proportionally, it remains true that this overall increase in relative shooting volume and an increase in 3 PT volume specifically, indicates a change in shot preferences.

Model

We based our model weights off the individual shooting years in the NBA collected from the NBA Savant API. We then used these weights to post stratify the shooting data collected from the NBA Stats API from 1997 to 2020 to determine the predicted overall play accuracy of all time. We applied a logistic regression model to our training set, William’s API data, since our variable of interest was ‘shot_made_flag’. That is, we are interested in how shot accuracy has changed over time in the NBA. Our model can be described by the following:

$$log(\frac{p}{1-p}) \sim \beta_0 \ + \beta_1 X_{i \ shot \ type \ new} + \beta_2 X_{i \ action \ type \ new}$$

Table 1: Pairwise Correl. Charts 2010-2017

	shot_made_flag	shot_type_new	action_type_new
Pairwise 2010			
shot_made_flag	1.00000	-0.10700	0.04200
shot_type_new	-0.10700	1.00000	-0.18600
action_type_new	0.04200	-0.18600	1.00000
Pairwise 2011			
shot_made_flag.1	1.00000	-0.10700	0.03800
shot_type_new.1	-0.10700	1.00000	-0.19000
action_type_new.1	0.03800	-0.19000	1.00000
Pairwise 2012			
shot_made_flag.2	1.00000	-0.10600	0.02700
shot_type_new.2	-0.10600	1.00000	-0.20300
action_type_new.2	0.02700	-0.20300	1.00000
Pairwise 2013			
shot_made_flag.3	1.00000	-0.11205	0.03669
shot_type_new.3	-0.11205	1.00000	-0.21762
action_type_new.3	0.03669	-0.21762	1.00000
Pairwise 2014			
shot_made_flag.4	1.00000	-0.12000	0.03900
shot_type_new.4	-0.12000	1.00000	-0.22800
action_type_new.4	0.03900	-0.22800	1.00000
Pairwise 2015			
shot_made_flag.5	1.00000	-0.12100	0.04800
shot_type_new.5	-0.12100	1.00000	-0.30400
action_type_new.5	0.04800	-0.30400	1.00000
Pairwise 2016			
shot_made_flag.6	1.00000	-0.13000	0.04400
shot_type_new.6	-0.13000	1.00000	-0.32400
action_type_new.6	0.04400	-0.32400	1.00000
Pairwise 2017			
shot_made_flag.7	1.00000	-0.13700	0.02400
shot_type_new.7	-0.13700	1.00000	-0.35400
action_type_new.7	0.02400	-0.35400	1.00000

where $i = 1, 2, ...n\ and\ p\ =\ probability\ of\ a\ successful\ shot$

Figure 1: As seen above, is the logistical regression model for our test data set, provided by the NBA Stats API.

Figure 2: We see that Shot Type and Action Type are the variables that correlate most strongly with shot accuracy. This means that the success of a player's shot going into the basket is decided between shooting a 2PT or 3PT and the style of shot chosen.

From the total amount of variable provided in the dataset, we decided to only produce a logistic regression model on 'action type' and 'shot type' for a limit on the processing power of my R studio machine as well as the demonstration through a logistic regression on individual years in the training set demonstrating that the only consistent variables that showed strong significance were these two. Thus, it is appropriate to condense our data set as such, but we will still discuss our other variables in our exploratory data analysis discussions. Both logistical models in both the training and test data sets had a fisher score iteration of 4, meaning that it took 4 iterations until our data fit our model accordingly. As for the classification of our variable types, since we were reducing the subcategories of each variable from 72 action types to 4 and from classifying a '2 PT' or '3 PT' attempt to their respective points, it is appropriate to consider them as dummy variables using as.factor() in our model. Properly classifying our variables would allow the ease of calculation for odds ratios and increases the stability and significance of the coefficients

Table 2: Shot Totals for 2PT and 3PT 2010-2017

shot_made_flag	shot_type_new	sum(success)
Shot Sums 2010		
0	2	0
0	3	0
1	2	151114
1	3	47625
Shot Sums 2011		
0	2	0
0	3	0
1	2	118706
1	3	38046
Shot Sums 2012		
0	2	0
0	3	0
1	2	146894
1	3	52788
Shot Sums 2013		
0	2	0
0	3	0
1	2	147084
1	3	57123
Shot Sums 2014		
0	2	0
0	3	0
1	2	145540
1	3	57876
Shot Sums 2015		
0	2	0
0	3	0
1	2	138602
1	3	60540
Shot Sums 2016		
0	2	0
0	3	0
1	2	120532
1	3	60237
Shot Sums 2017		
0	2	0
0	3	0
1	2	120496
1	3	66441

Observations	4729234
Dependent variable	Shot.Made.Flag
Type	Generalized linear model
Family	binomial
Link	logit

in our model (Garavaglia et al., 2016). Thus, it is important for our classification of categorical variables to be classified as dummy variables for our analysis.

$\chi^2(4)$	297752.06
Pseudo-R² (Cragg-Uhler)	0.08
Pseudo-R² (McFadden)	0.05
AIC	6214310.49
BIC	6214377.33

	Est.	S.E.	z val.	p
(Intercept)	1.42	0.00	312.60	0.00
as.factor(shot_type_new)2	-0.13	0.00	-54.42	0.00
as.factor(action_type_new)2	-1.88	0.00	-395.21	0.00
as.factor(action_type_new)3	-1.17	0.00	-238.64	0.00
as.factor(action_type_new)4	-1.42	0.01	-200.80	0.00

Standard errors: MLE

Observations	1528321
Dependent variable	shot_made_flag
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(4)$	99302.35
Pseudo-R² (Cragg-Uhler)	0.08
Pseudo-R² (McFadden)	0.05
AIC	2005518.94
BIC	2005580.14

	Est.	S.E.	z val.	p
(Intercept)	1.44	0.01	171.14	0.00
as.factor(shot_type_new)2	-0.12	0.00	-28.96	0.00
as.factor(action_type_new)2	-1.91	0.01	-216.66	0.00
as.factor(action_type_new)3	-1.13	0.01	-124.75	0.00
as.factor(action_type_new)4	-1.49	0.01	-119.72	0.00

Standard errors: MLE

```
## Warning in kable_styling(., latex_options = "hold_position"): Please specify
## format in kable. kableExtra can customize either HTML or LaTeX outputs. See
## https://haozhu233.github.io/kableExtra/ for details.
```

Table 3: Variance Inflation Factors for NBA Shots 2010-2017

	shot_type_new	team_name_new	action_type_new	period	player_ids
vif_nba10	1.038088	1.001061	1.034593	1.003410	1.001011
vif_nba11	1.037759	1.001472	1.035677	1.002649	1.001621
vif_nba12	1.043411	1.006989	1.041083	1.002513	1.007312
vif_nba13	1.050200	1.004277	1.047812	1.002240	1.005811
vif_nba14	1.054498	1.006773	1.051755	1.001871	1.006610
vif_nba15	1.099611	1.001642	1.097019	1.001533	1.001309
vif_nba16	1.114481	1.007329	1.111326	1.001459	1.005661
vif_nba17	1.140507	1.002020	1.136843	1.000959	1.002052

Table 3: Here, we can see that the Variance Inflation Factor (VIF), is generally close to 1 for all cases. Any VIF greater than 4 is generally a concern that indicates strong multicollinearity (PSU, 2018). Multicollinearity indicates strongly correlated X's meaning that individual X's do not contribute to the prediction of Y over and above other X's. Looking at our pairwise correlation in Table 1, this is further support that our variables are generally well chosen and multicollinearity isn't a big issue.

Furthermore, it may be important to note that although model reduction algorithms like AIC and BIC may be also known as the post stratified model, sees similar trends with the full unweighted model. However, in the post stratified $\hat{Y}$ values, the numbers are a lot more reasonable which also include shot accuracies factored in from 2017 to 2020 as well. Since we are seeing a relatively positive relationship from 2010 to 2017 in shooting accuracy, we may presume that the high overall weighted post stratified model's high accuracy in both 2 PT and 3 PT shooting can be attributed to subsequent shot accuracies increasing.

Table 6: Yhat Values per Dependent Variable PS and Full

	Action PS	mean PS	lower PS	upper PS	Action Full	mean full	lower full	upper full
Dunk	1	0.808	0.808	0.808	1	0.801	0.789	0.808
Jump Shot	2	0.372	0.356	0.385	2	0.370	0.356	0.385
Layup	3	0.577	0.577	0.577	3	0.562	0.547	0.577
Hook Shot	4	0.487	0.487	0.487	4	0.472	0.456	0.487

Table 6: Table 6 displays the PS (post stratified) and full (unweighted) $\hat{Y}$ values for action type.

Table 7: Yhat Values per Dependent Variable 2010-2017

	Action '10	mean '10	lower '10	upper '10	Action '17	mean '17	lower '17	upper '17
Dunk	1	0.462	0.447	0.477	1	0.539	0.533	0.546
Jump Shot	2	0.429	0.343	0.495	2	0.440	0.352	0.523
Layup	3	0.498	0.480	0.514	3	0.492	0.482	0.500
Hook Shot	4	0.518	0.502	0.533	4	0.470	0.463	0.477

Table 7: Dunks and Jump shots seem to increase in popularity over time whereas Hook Shots become less common. Overall, shot accuracy increases more over time as well, despite an increase in shooting volume and 3 PT volume. People have attributed this to an increase in the overall pace of the game and the emphasis on spacing the basketball floor to create room for shooting.

The $\hat{Y}$ values for each years' shot selection of the unweighted full model, and the post stratified model are displayed in Table 6. There is an overall increase of shot accuracy changes to the Jump Shot while an overall decrease to Dunks, Layups and Hook Shots. Similar deductions from Table 7 may be applied here as well, changes in preferences to the Jump Shot indicates a change in shot preferences over time due to a change in their accuracies. The large discrepancies between individual later years and the post stratified models may be because of classification errors when condensing the 72 shot selection responses to 4 categories. Table 7 had to exclude years 2011 to 2016 due to a fitting error when rendering tables; thus, if you were to aggregate the respective years, you would still see a positive and increasing trend in overall shot accuracy in the league.

Discussion

The overall analysis is summarized with these following trends:

- An overall increase in NBA shooting volumes from 2010 to 2017 in both 2 PT and 3 PT attempts.
- Shot selection, defined by selecting between 2 PT or 3 PT and the style of finish a player performs, is correlated with the success of a shot going into the basket.
- An overall increase in NBA shooting accuracies of the Jump Shot, a shot conducted at an unreachable closing range within the hoop, while a decrease in accuracies in the other styles of scoring which are typically used to close into range of the hoop.

With these conclusions made, it is viable to acclaim that there is a definite change in the NBA metagame over time. These changes include all the 3 following changes above, which are also linked to the spectator 'eye-test' when watching our NBA games. When television analysts and avid basketball fans acclaim a drastic change to the metagame has changed the state of shot preferences and shot efficiency, there may be strong evidence of such claims in this report. With an overall aggregate increase in shot volume and shot accuracy, this may be linked to higher point games and shot efficiency.

Weaknesses

There are numerous ways where weaknesses present itself in our analysis of the NBA shot using our post stratification technique. Despite MRP being a great method to adjust for population errors in representing different subgroups, the way we initially classify our subgroups holds a big impact on the output of our model. These data decisions, notably when classifying our 'action type' field, could drastically change the output of our model. In this case, to reasonably condense 72 different ways to shoot a ball at a basketball down to 4 main categories may cause disproportionate representations of certain shot selections. An example of this would be shots that would lie on the grey zone, are 'floaters' considered layups, hook shots or jump shots? Floaters are kind of a hybrid of all 3 shots and should technically be considered its own shot. However, in this report, it is important to consider my computer's capabilities to process a denser model and thus, the decision to classify it as a jump shot may have caused an implicit bias in the logistical regression model. Such classification results in the declaration that a jump shot is equivocally the same as a floater, which to many, it is not. Furthermore, it is important to consider the other factors that we did not include in our model. According to some of our regression outputs in previous years, variables to consider would have been 'period', 'team name' and 'player id'. To incorporate more variables would allow a more robust analysis of how the NBA meta game has truly changed because 'action type' and 'shot type' may not be the only contributing factors. Due to limitations of the machine, we had incorporated a variable selection system to reduce this model down to just 'action type' and 'shot type' to predict our shot success percentage. If the model had incorporated more variables, there could be discussions regarding a more robust analysis about shot selection during periods of the game as well as player/team preference for their style of play. An example of current basketball veterans is Ben Simmons and DeMar DeRozan. During an era where 3 PT shooting volume and overall shot accuracy is at its highest, these players struggle to maintain strong 3 PT shooting averages (Basketball Reference, 2020); however, these players are anomaly cases since they are still considered all stars. To incorporate a more robust regression model would allow us to discuss more about player preferences and how this affects the 'efficient' shot selection meta concurrently in the NBA.

Next Steps

For further analysis, it is beneficial to then try to formulate a way to assess efficiency, perhaps in a different way than contemporary efficiency statistics in the NBA. The "TS %", also known as "True Shooting Percentage", is a common and simple way for analysts to understand the scoring efficiency of a player (Basketball Reference, 2020). However, efficiency as defined by our report is more than just scoring numbers, we clearly identified that the style of shot is highly correlated with its success as well. Thus, perhaps efficiency statistics should also measure stylistic elements of a player's basketball play style. After all, movement within the court during a game is also an indicator of energy expenditure, 'attacking the rim' for a player's finishing game, and awareness of 'spacing the court'. To further investigate claims to changes in the NBA metagame towards efficiency and spacing, it is important to consider a multitude of factors that would enhance our current analysis. Assessing the external validity of our current analysis in such a way would further bring evidence to the ‘eye-test’ that the NBA metagame changes from 2010 to 2017.

References

Jarvis Ryan. (2017). "The NBA Meta". Medium. https://medium.com/@rjarv27/the-nba-meta-3911b7160c28

Penn State University. (2018). "10.7 - Detecting Multicollinearity Using Variance Inflation Factors". Department of Statistics Online Programs. https://online.stat.psu.edu/stat462/node/180/

Shivon Sue-Chee. (2020). "Module 10: Diagnostics in MLR". University of Toronto. pp 13-47.

Ebaek. (2019). "NBA Shot Tracker". NBA Savant. https://github.com/ebaek/NBAShotTracker

Zak Geis. (2020). "2020 - June - NBA Shots (1997-2019)". NBA Stats API. https://data.world/sportsvizsunday/june-2020-nba-shots-1997-2019

Garavaglia Susan, Sharma Asha et al. (2016). "A Smart Guide To Dummy Variables: Four Applications And A Macro". IDRE UCLA. https://stats.idre.ucla.edu/wp-content/uploads/2016/02/p046.pdf

Basketball Reference. (2020). "Glossary". Sports Reference. https://www.basketball-reference.com/about/glossary.html

Basketball Reference. (2020). "NBA & ABA Player Directory". Sports Reference. https://www.basketball-reference.com/players/

R Packages References:

Robinson, D. and Hayes, A. (2020). *broom: Convert Statistical Analysis Objects into Tidy Tibbles.* R package version 0.7.0.

Zhu, H. (2020). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* R package version 1.3.1.

Wickham, H. and Miller, E. (2019). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.* R package version 2.1.1.

Long JA. (2020). *jtools: Analysis and Presentation of Social Scientific Data.* R package version 2.1. 0, https://cran.r-project.org/package=jtools

Fox J, Weisberg S. (2019). *An R Companion to Applied Regression.* Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Lumley T .(2020). *survey: analysis of complex survey samples.* R package version 4.0.

Michael Quinn, Amelia McNamara et al. (2019). *Skimr: Compact and Flexible Summaries of Data.* R package version 1.0.7. https://CRAN.R-project.org/package=skimr

Wickham H, Averick M, Bryan J et al. (2019). *Welcome to the tidyverse.* Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686