Understanding Features of Successful 3 Point Shots in the NBA

Nate Bailey, Karan Bhuwalka, Hin Lee, Tim Zhong
Massachusetts Institute of Technology, MIT
(Dated: December 13, 2018)

I. INTRODUCTION

Basketball is one of the fastest growing sports in the country. Over the past ten years, annual league revenue has increased almost 75% to 7 billion dollars [1]. TV ratings have also soared, increasing by 32% from 2016-2017, with over 20.4 million people tuning in for the NBA finals [2]. As the popularity of the sport has continued to grow, it has seen an influx of data analysts attempting to study the game–essentially following in baseball's footsteps. The extensive use of analytics in baseball has been around for a while, but the true value was revealed with the Moneyball movement where these analytics could be leveraged to build a winning team.

In basketball, the use of data analytics is a useful tool to inform coaches and viewers of trends that may not be visible to the naked eye. Companies such as Second Spectrum are starting to analyze movement data to better understand player movements in the NBA. In this paper, we hope to use similar analysis of movement data to analyze and predict shot quality, as well as evaluate types of players by types of shot attempts to discover similarities between players based on movement data.

A. Data in NBA

In the NBA in particular, player movement data provides information on the location of all 10 players on the court and the ball 25 times per second, and has been available for every NBA game since 2013. This data has been used in the NBA extensively in recent times. Nearly every team in the NBA now has data analysts on their staff while the NBA's best team, the Golden State Warriors, credit much of their success to their analytic acumen. The most prominent use of data analysis in the NBA has been to inform the rise of three point shooting. In 2012, before this data was available, teams took only 18.4 three pointers per game. In 2017 they took 27.

B. Gap in Existing Research

While we found a good amount of research that has tried to quantify probability of shot success, we noticed that most of the inputs contained static information. Typically research focused on distances (from basket, from defender etc.) and angles (of ball trajectory, shooter position etc.). Example plot of number of shots at different location and the relative accuracy are plotted in FIG 1 and FIG. 2. Given basketball involves a lot of movement, we felt that looking at dynamic features could lead to meaningful insights about shot quality.

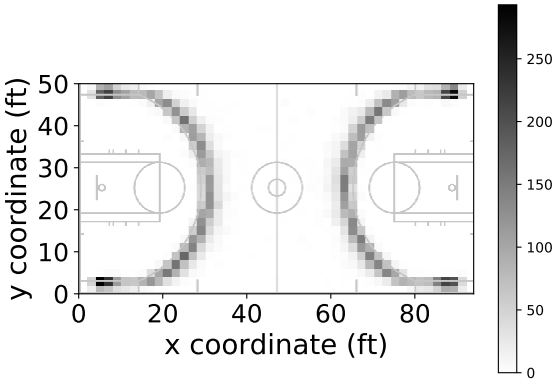

FIG. 1: Number of three point shots taken at different location on the court

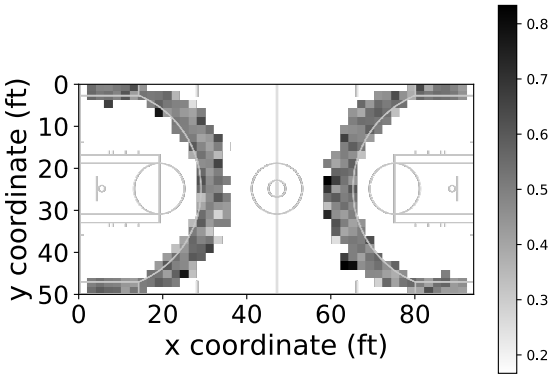

FIG. 2: Accuracy of three point shots taken at different location on the court

This is why our approach involves creating both static and dynamic features that could potentially affect shot success rate and then comparing how the dynamic features perform against previously studied static features. For all the features we define, knowing whether or not they contribute significantly to shot success tells us a lot about what data matters when studying and modelling basketball shots.

C. Questions we address

We limit our analysis to three point shots given that the sport is moving towards taking more of these shots, in part due to recommendations by data analysts. Also restricting our scope in this way makes the processing computationally less difficult, the features clearer to define and allows for using different analytic methods.

We look at player movement data to analyze the following questions:

1. Which dynamic features of three point shots contribute to higher success rate?
2. How do players differ in the type of shots they take?

II. RELATED WORKS

There are many papers regarding shot quality/characteristics in the NBA as sports analytics has become an increasingly popular field. In particular, many papers for the MIT Sloan Sports Analytics Conference center on these topics, and it has also been a popular topic for many predictive analytics studies with the rise of websites such as FiveThirtyEight.

For example, several studies have focused on attempting to characterize the quality of a shot in different self-defined metrics [3] or by augmenting the resolution of the dataset of shots [4]. These specific studies showed to be effective in improving player evaluation–taking features of a shot such as defender distance, angle of shot, and shot distance to better quantify what a "good shot" is. They also show the impact that having a granular dataset has on the ability to quantify "good shots". Our study aims to build off of these studies by leveraging a very granular player movement dataset and by incorporating more features for our own analysis.

Others have tried to predict whether a player will make a shot using predictive models, such as boosting [6] and decision trees [5]. Some of the additional models used in these papers include logistic regression, SVMs, Naive Bayes, Neural Networks, and Random Forests. The downfall of these studies is the lack of accuracy due to having an incomplete/insufficient dataset, and it is a challenge to consider as we move forward using similar datasets. However we hope that wrangling more unique features will help increase the accuracy of our models, or at least reach the level accuracy that is seen in literature.

Finally, some papers have even tried to simulate the implementations of new offensive strategies [7] to see how effective they would be. In this study, they offer a method for testing the impact of team shot policy adjustments on the efficiency of the offense. This particularly has profound effects over the course of a whole season where consistently having a high offensive efficiency is paramount. We aim to take some of this thinking into our own study by examining player tendencies and whether certain players tend to follow their own strategies with shot selection.

III. RAW DATA

We combine two publicly available datasets (found in the same GitHub repository [**?**]) to address this research topic. First, we use SportVU player movement data which provides the location of all 10 players as well as the ball on the court at a resolution of 25 times per second. Second, we use gamelogs (play by play data) that record events of specific types (shot attempts, rebounds, fouls, timeouts, substitutions, and others), the time of the event, and the player(s) involved in the event.

The public subset of the player movement data contains 636 games played between the start of the season on October 27, 2015 and January 23, 2016. For these games, the dataset also includes the corresponding gamelogs. The movement data is broken up by the events listed in the gamelog file, and for each event it contains each 1/25 second long moment between the previous event and the current one. Each moment contains the game clock and shot clock at that time, the players on the court at that time, their locations in (x, y) coordinates, and the ball's location in (x, y, z) coordinates. Details of the data structure are shown in FIG. 3 We have 17,977 three points shots's data with total raw data size of 64.5 GB which are computer intensive to pre-process.

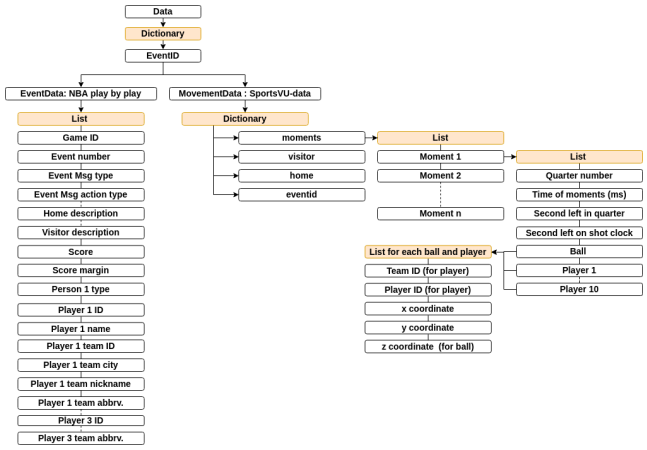

FIG. 3: Data structure of the merged SportsVU and play by play data.

IV. TERMINOLOGY

There are three major ideas and definitions used for pre-processing and generating features and parameters for analysis. They are "Catch-and-Shoot", time of the catch, and time of the shot.

A. Catch-and-Shoot (C&S)

A specific shot of interest is the "Catch-and-Shoot" (C&S) 3 point shot, defined as a shot attempt which is

made without dribbling in between receiving a pass from another player and taking the shot. We believe that the factors that lead to these shots being successful may be different than those for 3 point attempts that come off the dribble.

In our dataset, we identified these shots by observing whether the ball z-position ever fell below 1 foot above the court in between the identified moment in which the shooter gains possession of the ball and the identified moment when the shooter attempts the shot. If the ball stays over 1 foot above the floor of the court in between these attempts, the shot is identified as a catch-and-shoot (C&S) attempt. The definition of catch-and-shoot we use here is the same as in [3], though we had to define our own z-cutoff to identify these shots. We chose 1 foot based on qualitative examination of several example events. In our dataset, almost 80% of the three point shots were C&S.

B. Time(Frame) of the Shot

This is the time(frame) when the shooter take the shot. It is defined as the time of the first local minimum below 10 ft where we backtrack from the top 10 highest ball height (z). We also ran a logic such that the shooter we know took the shot could only have done it if they had possession at the local minimum to avoid data in the same event that there are multiple three point shots. The xy-location of the shooter and the ball, and z-location of the ball at time of the shot are indicated as a red dot in FIG. 4 5 6 respectively.

C. Time(Frame) of the Catch

This is the frame when the shooter catch the ball. It is defined as the time before the time of shot that the ball is within three feet of the shooter. The (x, y)-location of the shooter and the ball, and z-location of the ball at time of the catch are indicated as a black dot in FIG. 4 5 6 respectively.

V. FEATURES, PARAMETERS AND PRE-PROCESSING

A. Shot Clock

Time left on shot clock at the time of the shot.

B. Catch and Shoot

An indicator variable that indicates whether the shot is a catch and shoot. Catch and shoot is true if the ball stays above one foot between time of catch and time of shot since there is no dribbling. As we see in FIG. 6, the shot is a C&S attempt. We can also see the shooter's

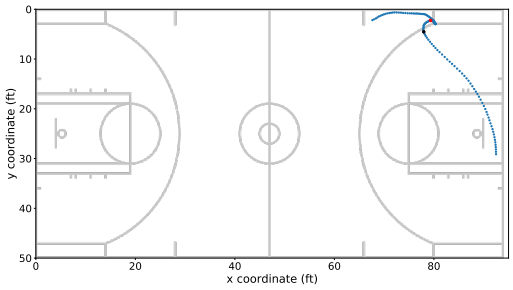

FIG. 4: Position of the shooter from time frame 120 to 260 of a three point catch and shoot with red dot at shot and black dot at catch

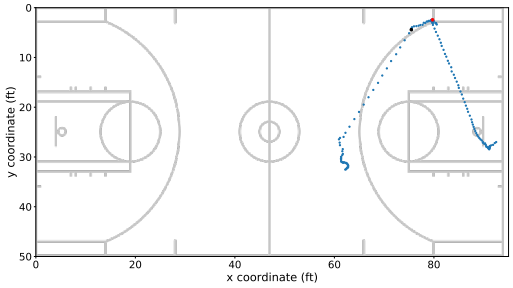

FIG. 5: Plot of the (x, y)-position of the ball position from time frame 120 to 260 of a three point catch and shoot with red dot at shot and black dot at catch

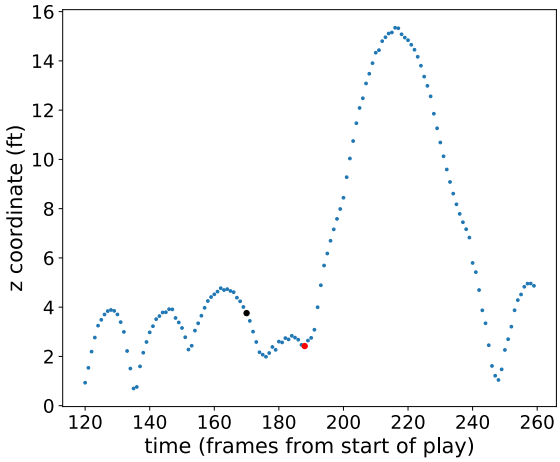

FIG. 6: Plot of the z-position(height) of ball from time frame 120 to 260 of a three point catch and shoot with red dot at shot and black dot at catch

movement and the ball's movement(passing) in FIGs. 4 and 5 also, where it shows a pass leading to the C&S attempt.

C. Time Catch and Shoot

Time between catching and shooting.

D. Velocity Catch and Shoot

The shooter's average velocity between every two subsequent frames between catching and shooting.

E. Distance at Shot

Distance of the shooter from the basket at the time of the shot.

F. Shot Angle

This is the vertical angle of release of the three point shot, which indicates the ball's trajectory.

G. Shooter Travel

Distance travelled by the shooter between the time of catch and time of shot. This distance is indicated by the black arrow in FIG. 7

H. Shooter Move Angle

The angle between shooter travel defined above and the line from the shot location to the basket. This angle is indicated by the red angle in FIG. 7.

I. Defender Distance at Shot

Distance of the closest player of the opposing team at the time of the shot.

J. Defender velocity at Shot

Average speed of the player identified as the closest defender from time of catch to time of shot.

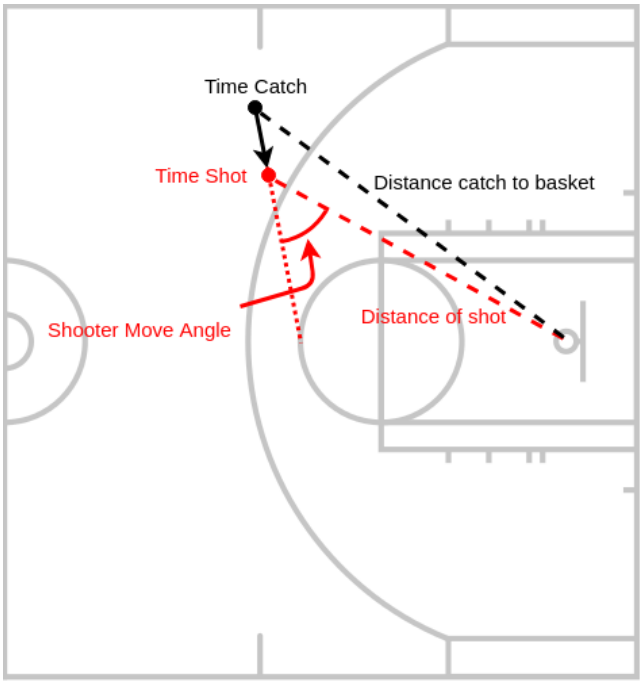

FIG. 7: Visualization of difference features used for a three points shot

K. Pre-Shot Speeds

We suspected before examining the code that a player's speed leading up to a shot could affect that shot's success in several ways. First, a high speed could make the shot harder to defend as the defender must balance keeping up with the shooter with maintaining position to affect the shot. Second, a high speed could make it more difficult for the shooter to have the precision necessary to make a long-range shot, as the distance and angle between the shooter and the basket will be changing quickly.

We identified several ways to define speed that could be relevant. For 3-point shots attempted off the dribble, the speed in a time window during which the player has possession and ending at the time of the shot could be a significant factor in success rate. We identified this form of speed as "Speed Before Shot (n seconds)", where the time parameter of n seconds defines the window before the shot over which the average speed of the shooter is calculated. Note that most catch-and-shoot attempts involve the shooter holding the ball for only fractions of a second, so this measure of speed is used only for non-catch-and-shoot attempts. A timeline for this features are listed in FIG. 8

For catch-and-shoot attempts, we believed that the speed before catching the ball may be an important factor as it can create difficulties for the defender on the play. For these attempts, we defined the average speed of the shooter over an n second window ending with the time of the catch as "Speed Before Catch (n seconds)". For non-catch-and-shoot attempts, we believed this speed

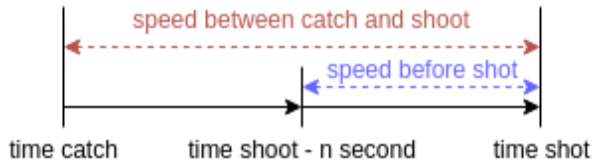

FIG. 8: Timeline of features used for a non catch and shoot three point shot

to be less relevant because the shooter sought to gain an edge over the defender via dribbling rather than based on movement before the catch, so this measure of speed is used only for C&S attempts. A timeline for this features are listed in FIG. 9

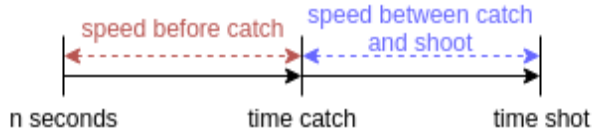

FIG. 9: Timeline of features used for a catch and shoot three point shot

For both types of attempts, we also were interested in the average speed in between the shooter gaining possession of the ball and shooting. For catch-and-shoot attempts, the time period between catching and shooting will be short but the speed may indicate whether the player is shooting on the run or stopping to collect the ball first. For attempts off the dribble, this measure of speed may indicate a player's ability to create separation while in possession of the ball. We defined this quantity as "Speed Between Catch and Shoot".

VI. RESULTS AND ANALYSIS

A. Effect of Speed on Shot Success Probability

Given the many different measures of speed associated with three-point attempts, our goal was to systematically determine which measures of speed had significant impact on shot success. We define making or missing a shot as a random variable M that took values of 1 for a make and 0 for a miss. The shooter's speed on the attempted shot is a random variable S distributed with probability density function $F_S(s)$. We want to know whether the probability of a make conditioned on the shooter's speed is constant across all speeds, which is proportional to the density function of speeds conditioned on whether the shot is a make or a miss:

$$P(M=1|s) = \frac{F_{S|M}(s|M=1)P(M=1)}{F_S(s)} \propto F_{S|M}(s|M=1) \tag{1}$$

To determine this for each different measure of speed we define, we conduct a hypothesis test on the distribution of speed associated with each. The null hypothesis, corresponding to a case where speed does not influence success probability, is:

$$H_0 : F_{S|M}(s|M=1) = F_{S|M}(s|M=0) \tag{2}$$

The alternative hypothesis, corresponding to a case where speed does influence shot probability, is:

$$H_a : F_{S|M}(s|M=1) \neq F_{S|M}(s|M=0) \tag{3}$$

Using the empirical speed distributions for made shots and missed shots as two samples drawn from the distributions of interest, we used the two-sample Kolmogorov-Smirnov test to test these hypotheses. The model for the null hypothesis in this test is that the two samples are drawn from the same distribution. If the maximum difference at any speed is larger than a specific value ($\sqrt{-\frac{n_1+n_2}{2n_1n_2}\log\alpha}$, where α is the significance level), the null hypothesis is rejected. We use $\alpha = 0.1$ for each test, since we are interested in detecting potential features and the penalty for false positives is not very high.

1. *Catch and Shoot Attempts*

For C&S attempts, we tested whether the distributions of speed for made and missed shots were significantly different, where speed was defined as either the shooter's average speed before catching the ball over the previous n seconds ("Speed Before Catch (n seconds)") or the shooter's average speed between catching and shooting the ball ("Speed Between Catch and Shoot").

Table I shows that the shooter's average speed before catching the ball significantly differs between made and missed shots on C&S attempts when computed over the 0.5, 1, 1.5, or 2 seconds immediately preceding the catch. However, for time intervals longer than 2.5 seconds, the distributions of shooter speed do not significantly differ on made or missed shots. The shooter's average speed between catching and shooting the ball on C&S attempts are also significantly different between made and missed shots.

2. *Off-Dribble Attempts*

For three point attempts off-the-dribble (defined as any shot attempt that wasn't a C&S attempt), we tested for significant differences between made and missed shots on speed distributions related to the shooter's average speed in the n seconds immediately before shooting the ball ("Speed Before Shot (n seconds)") or the shooter's average speed between catching and shooting the ball ("Speed Between Catch and Shoot").

Speed Measure	Makes	Misses	D_{max}	Significant?
C&S Attempts				
Speed Before Catch (0.5s)	854	1466	0.080	Yes
Speed Before Catch (1.0s)	841	1426	0.069	Yes
Speed Before Catch (1.5s)	821	1385	0.063	Yes
Speed Before Catch (2.0s)	803	1349	0.052	Yes
Speed Before Catch (2.5s)	778	1307	0.032	No
Speed Before Catch (3.0s)	753	1278	0.032	No
Speed Before Catch (3.5s)	726	1240	0.034	No
Speed Before Catch (4.0s)	708	1208	0.032	No
Speed Before Catch (4.5s)	690	1166	0.030	No
Speed Before Catch (5.0s)	665	1123	0.036	No
Speed Between Catch and Shot	895	1541	0.069	Yes
Off-Dribble Attempts				
Speed Before Shot (0.5s)	244	442	0.042	No
Speed Before Shot (1.0s)	209	378	0.039	No
Speed Before Shot (1.5s)	172	311	0.081	No
Speed Before Shot (2.0s)	144	253	0.085	No
Speed Before Shot (2.5s)	125	215	0.082	No
Speed Before Shot (3.0s)	105	181	0.081	No
Speed Before Shot (3.5s)	91	165	0.114	No
Speed Before Shot (4.0s)	73	144	0.122	No
Speed Before Shot (4.5s)	60	125	0.105	No
Speed Before Shot (5.0s)	49	104	0.130	No
Speed Between Catch and Shot	253	463	0.038	No

TABLE I: Number of makes and misses for which the associated speeds were extracted from the movement data for each speed measure defined and tested, the associated largest gap between the two distributions D_{max}, and whether or not the distribution of speeds for makes and misses were found to be significantly different.

The results shown in Table I indicate that made and missed 3-point shots attempted off-the-dribble do not significantly differ in any measure of speed. This finding indicates that players who move quickly with the ball either before shooting or while in possession with it do not make or miss shots at significantly different rates than those who move less quickly. When compared to C&S attempts, this suggests that very different variables could be involved in determining shot success for off-the-dribble shots, as shooters' speed is not a significant factor.

B. Regression of Features on Shot Result

We run a logistic regression (with L2 regularization) on the features we select to try and understand how they affect shot success.

The following features have a p-value less than 0.05:

Time on Shot Clock, Time between Catching and Shooting, Velocity between Catching and Shooting, Shot Distance, Shooter Travel Toward Basket, Defender Distance

The following features have a p-value greater than 0.05:

Whether the Shot is a Catch and Shoot, Whether the Shooter is from Home or Away Team, Ball Trajectory, Angle of Movement of the Shooter before Shot, Defender Velocity

Logit Regression Results

Dep. Variable:	y	No. Observations:	7368
Model:	Logit	Df Residuals:	7357
Method:	MLE	Df Model:	10
Date:	Sun, 09 Dec 2018	Pseudo R-squ.:	0.006763
Time:	00:54:46	Log-Likelihood:	-4797.3
converged:	True	LL-Null:	-4830.0
		LLR p-value:	3.509e-10

	coef	std err	z	P>|z|	[0.025	0.975]
time_on_shot_clock	0.0193	0.005	4.263	0.000	0.010	0.028
is_catch&shoot	0.0410	0.088	0.466	0.641	-0.131	0.213
is_home	-0.0079	0.049	-0.163	0.871	-0.103	0.087
time_between_c&s	-0.0546	0.026	-2.086	0.037	-0.106	-0.003
velocity_between_c&s	-0.0256	0.011	-2.391	0.017	-0.047	-0.005
shot_distance	-0.0345	0.006	-5.599	0.000	-0.047	-0.022
ball_trajectory	0.0018	0.002	0.752	0.452	-0.003	0.006
shooter_move_angle	-3.742e-05	0.001	-0.055	0.956	-0.001	0.001
shooter_travel	0.0120	0.005	2.448	0.014	0.002	0.022
defender_distance	0.0178	0.007	2.469	0.014	0.004	0.032
defender_velocity	-0.0112	0.006	-1.905	0.057	-0.023	0.000

These results make intuitive sense. Home or Away is less likely to affect shot success than, say, shot distance. Just being a catch and shoot itself, understandably, does not significantly make a shot more likely to go in.

We then remove the insignificant features and re-run the regression model:

Logit Regression Results

Dep. Variable:	y	No. Observations:	7368
Model:	Logit	Df Residuals:	7362
Method:	MLE	Df Model:	5
Date:	Sun, 09 Dec 2018	Pseudo R-squ.:	0.006309
Time:	01:36:03	Log-Likelihood:	-4799.5
converged:	True	LL-Null:	-4830.0
		LLR p-value:	7.767e-12

	coef	std err	z	P>|z|	[0.025	0.975]
time_on_shot_clock	0.0192	0.005	4.276	0.000	0.010	0.028
time_between_c&s	-0.0458	0.022	-2.091	0.036	-0.089	-0.003
velocity_between_c&s	-0.0204	0.010	-2.089	0.037	-0.039	-0.001
shot_distance	-0.0344	0.004	-8.689	0.000	-0.042	-0.027
shooter_travel	0.0096	0.004	2.188	0.029	0.001	0.018
defender_distance	0.0185	0.007	2.632	0.008	0.005	0.032

You can see that the sign of most of the coefficients are what you would expect. Having more time on the shot clock makes a shot more likely to go in, presumably because there is less pressure and more time on aiming. Meanwhile, moving faster between catching and shooting makes a shot less likely to go in. Being closer to the basket makes a shot likelier to go in. If a defender is further away, the success rate is also higher. Moving closer to the basket before shooting also slightly increases probability of success although less movement is preferred. All these results are shown in FIG. 10 and FIG. 11 where they shown the distribution of number of shot attempts on these features and their corresponding accuracy.

Looking more closely at the coefficients, moving closer 1 feet makes a shot 3.4% more likely to go in. According to the results of the Chang et al., moving closer 1 feet makes a shot 3.2% likelier to go in, so we are pretty close. This relationship of accuracy with distance of shot and distance of defender are plotted in FIG. 12. Note that in the figure, the high three point accuracy at defender distance of 12 to 14 feet and 14 to 16 feet with shot distance of 28 to 30 feet are inaccurate because there are only 33 and 12 shots taking yielding extremely noise results.

Interestingly, having more time between catching and shooting makes a shot less likely to go in. One could imagine that this is because dribble shots are harder to put in and also have longer time. However, we know

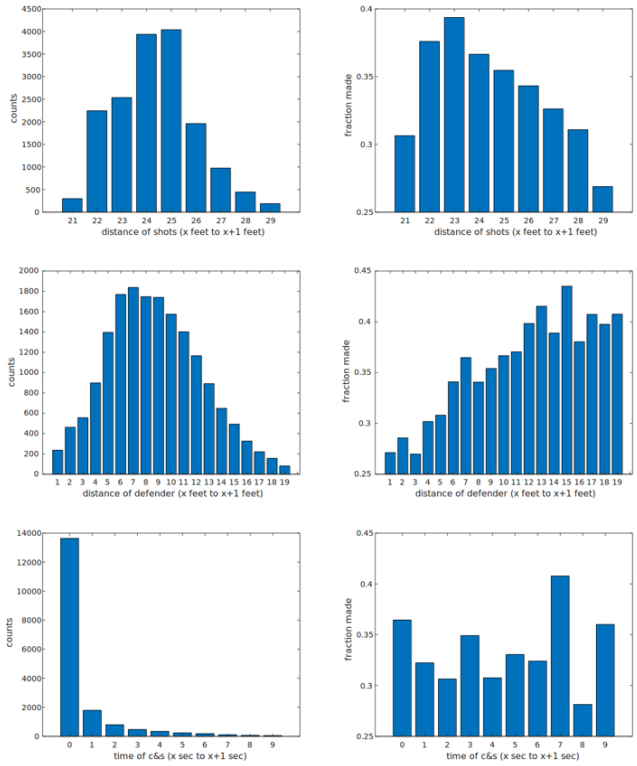

FIG. 10: Left: Number of shot attempts with varying feature, Right: Accuracy of shot attempts with varying feature ; Top: Distance of shot, Middle: Distance of the closest defender, Bottom: Time between catach and shoot

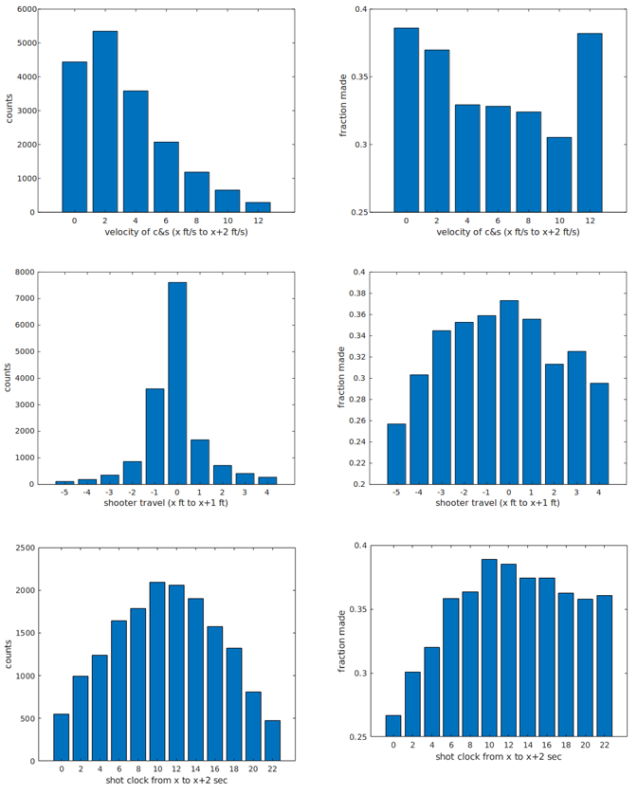

FIG. 11: Left: Number of shot attempts with varying feature, Right: accuracy of shot attempts with varying feature ; Top: Velocity between catch and shoot, Middle: shooter travel, Bottom: Shot clock

that this relationship cannot be just because the shot is a dribble and shoot as opposed to a catch and shoot (that feature had a p-value of .64). Another possible reason for this then, is that the shots that are taken quicker are also those in which players are "freer". However, we found low correlation between lower time between catching and shooting vs defender distance (coefficient= -.308). It may be worth examining if shots that are taken quicker are better because they are more "instinctive". It also possible that better players just happen to shoot faster and that is why we get this negative correlation.

Even when we look at only 'Catch and Shoots', the time between catching and shooting is negatively correlated with shot result.

	coef	std err	z	P>|z|	[0.025	0.975]
time_between_c&s	−0.1873	0.016	−11.787	0.000	−0.218	−0.156

C. Prediction Model

We ran a prediction model based on the significant features. However, we only got an accuracy of 54%.

We identified a few reasons for the bad prediction. While our features significantly affect the likelihood of a shot going in, their effect is numerically quite small. This is understandable as no individual feature has a massive impact given the sheer number of factors that probably affect shot quality. Other features that we don't have, such, as player quality and fatigue, probably overshadow all other effects. Also, to some extent, the likelihood of success is probably random or based on really indi-

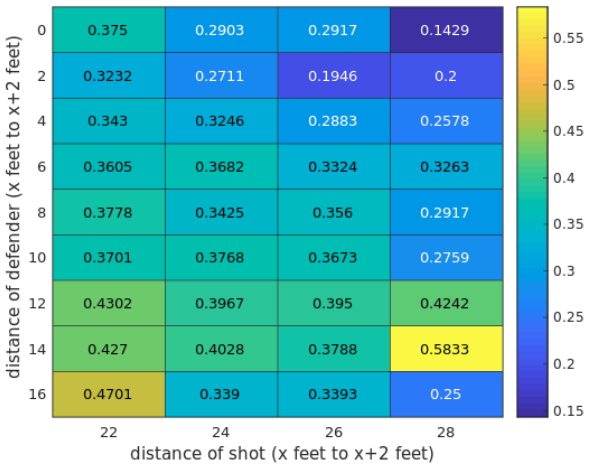

FIG. 12: Accuracy of three point shots at different shot distance and defender distance

vidual micro-characteristics and perhaps cannot be predicted by data. It is probably more valuable in sports to use data to get some inferences about the game rather than make predictions, given the dynamic and inherently unpredictable nature of sports.

While the model may not be able to predict whether exact shots go in or not, it is more useful to see whether the model is able to differentiate between better shots (i.e. higher rates of success) and worse shots. To do this, we find the probability of a shot going in based on the features from our model. We define 6 bins based on these probabilities and look at the actual results of the shots in those bins. From this we can obtain a percentage success rates of the shots in those bins. If our predicted probability of success is close to the actual success rates, our model is able to differentiate between "good" and "bad" shots. Below in FIG 13 you can see how our model performs:

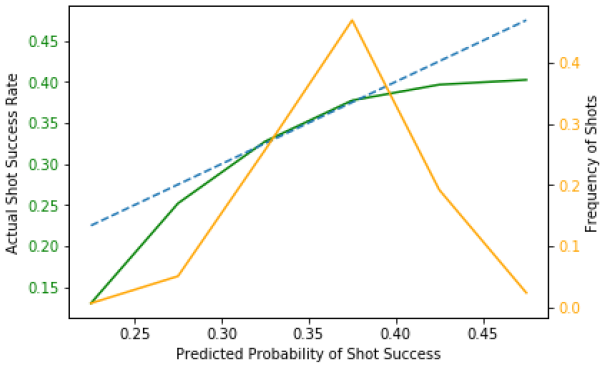

FIG. 13: Actual Rate of Making a Shot vs Predicted Probability of Making

In FIG 13, the x-axis is our predicted probability of shot success, while the green line gives the actual success rate. The blue dotted line is the 45 degree line that shows what the prediction would look like if we predicted 100% accurately. The orange curve is the frequency of shots taken with the respective predicted probabilities.

Our model predicts a higher rate of success for shots that had a very low or very high actual success rate, but as you can see from the directionality of the graph, our predicted probabilities follow the trend of the actual success rate. Moreover, for most of the shots, our models predicted probability is very close to the actual success rate of shots in that bin.

The regression of our prediction on the actual success rate gives an R^2 of **.928** with a standard error of .19 and p-value of .007. This means there is a strong statistically significant relationship between our probability of success and actual rate of success when we consider groups of shots. Although our model can't accurately tell us if any individual shot will go in, it gives us very good numbers for probability of success of the shot based on the features we defined.

Even changing the number of bins (i.e. decreasing bin size) shows a strong relationship. For 12 bins, we get an R^2 of **.841** with p-value of .0006. For 24 bins, we get an R^2 of **.834** with p-value of $3\text{x}10^{-7}$.

In summary, our model performs very well at predicting the probability of a shot going in, however it can't accurately tell us which exact shot will go in and which will not. This is purely because only a small percentage of shots actually go in. so even if we could completely accurately guess the probability of a shot going in, the predictive power of a model would be very low in correctly guessing which exact shots went in.

D. Player Clustering

1. *Methodology*

Using the movement data on 3-point attempts, we used clustering to identify groups of players who exhibit similar movement patterns on 3-point attempts. Because of the differences between catch-and-shoot attempts and off-dribble attempts, we defined the features for clustering by aggregating a player's C&S attempts and off-dribble attempts separately. For each player, we averaged each feature used in the logistic regression over all C&S attempts, with the exception of whether the shot was taken at home or away and ball trajectory and additionally including Speed Before Catch (0.5 seconds). We averaged the same features, with the exception of Speed Before Catch (0.5 seconds) over off-dribble attempts as well. This resulted in 21 features used in the clustering, to which we added the percentage of 3-point attempts which were C&S attempts as a 22nd feature. We selected a cutoff of 60 3-point attempts in the dataset to qualify a player for clustering, which corresponds to roughly 1.5 attempts per game. 116 players met this threshold.

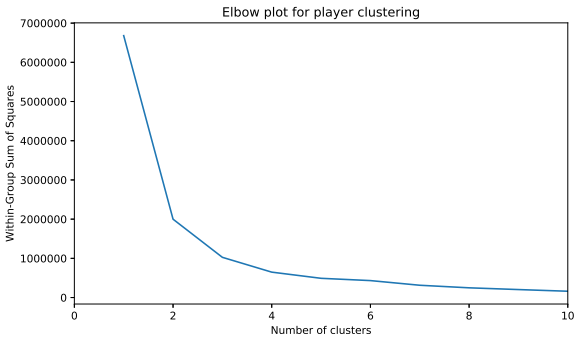

FIG. 14: Within-Group Sum of Squares for Different Values of k Used in k-Means Clustering

k-Means clustering was used to identify the player clusters from these features. FIG. 14 shows the within-group sum of squares as a function of k. From this plot, we identified 2 or 3 as the potential numbers of clusters to

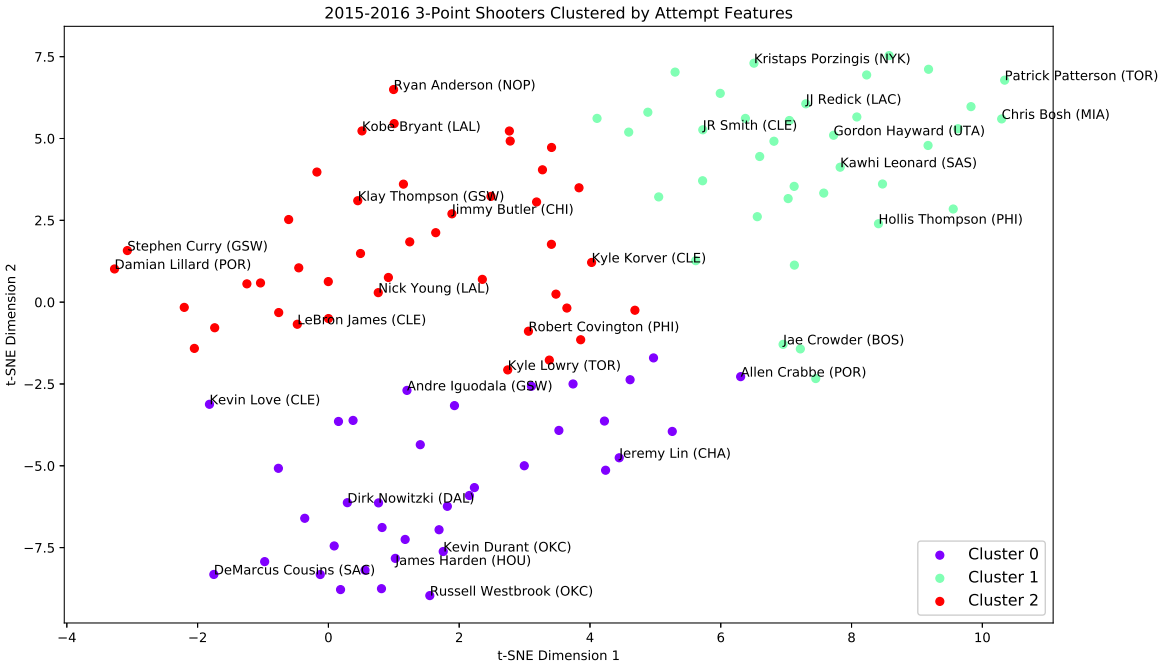

FIG. 15: Player aggregated movement data projected to 2D via t-SNE, with players colored by cluster label

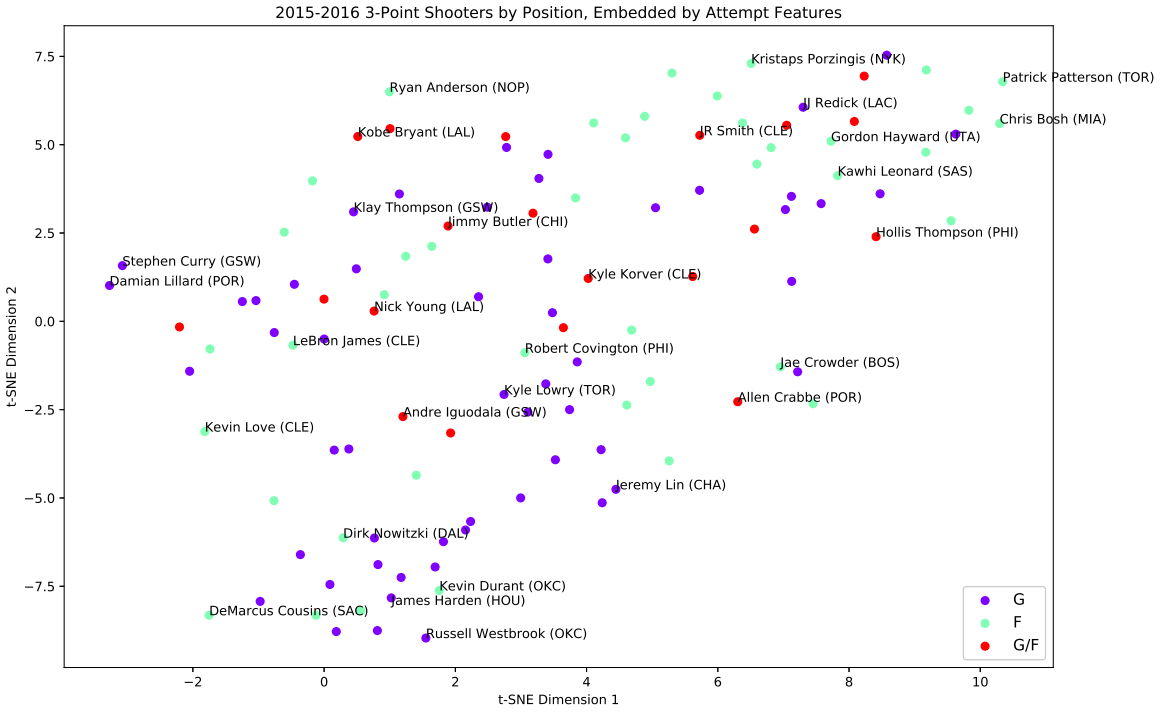

FIG. 16: Player aggregated movement data projected to 2D via t-SNE, with players colored by position

use due to that being the inflection point or "elbow" of the plot. As 3 is more interesting and potentially yields deeper insight into the variety of methods of shooting 3-pointers, we proceeded with that number for the rest of the analysis. k-Medoid clustering was also tried, with the idea that it would select a specific player who best represented the archetype of each cluster as its medoid, but we found that it was not robust and led to different assignments of players to clusters when repeated on the same dataset, so it was scrapped.

FIG. 15 shows the projection of the player movement data via t-SNE into two dimensions, with each individual player colored according to his assigned cluster label and specific players highlighted. This plot demonstrates that the clusters do a good job of grouping together players whose movement data is close to one another, a finding which is reinforced via silhouette analysis. The silhouette plot shown in FIG. 17 shows the silhouette scores for each player grouped by assigned cluster. The silhouette score is a measure of the average distance from a player to the other players in his or her cluster minus the average distance from the same player to all players in the next closest cluster that player could be assigned to. The average silhouette score is 0.23, indicating fairly good clustering, and only a few players in cluster 1 have silhouette scores less than 0, which indicate that they are on average closer to players in a different cluster than to their own.

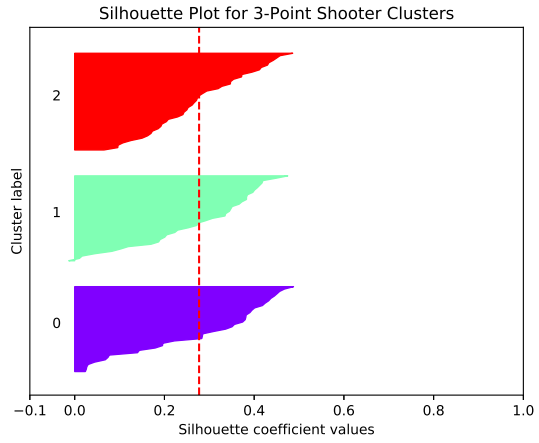

FIG. 17: Silhouette scores for each player grouped by cluster

The choice of 3 clusters also makes comparison to position-based groupings convenient. Using the player positions recorded on stats.nba.com, we assigned each player to one of three positional categories: guards (G), forwards (F), or hybrid guard/forwards (G/F). There were four players classified as hybrid forward/centers in the dataset (Frank Kaminsky, Kristaps Porzingis, Kevin Love, and Meyers Leonard), who we classified as forwards for the purpose of this analysis. No pure centers had enough 3-point attempts in the dataset to qualify for clustering. The t-SNE projection of the data with players colored by position is shown in FIG. 16, and the silhouette analysis is shown in FIG. 18. These plots show much worse results are obtained by grouping players based on their positions rather than by their movement data. In fact, every forward in the dataset is closer on average to players in a different position group than he is to other forwards. This indicates that our clusters identify players that may play different positions but attempt 3-point shots in similar ways, which could provide insight to NBA teams in how to defend against specific opposing 3-point shooters that takes into account more than just their position.

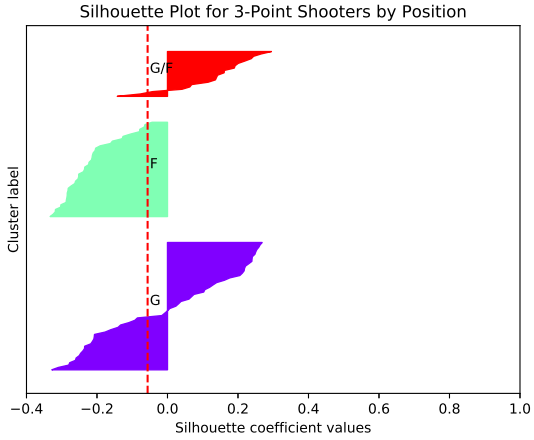

FIG. 18: Silhouette scores for each player grouped by position

2. *Understanding the Clusters*

After identifying these clusters, we were interested to see if we could infer some qualities of the 3-point shooting in each cluster to characterize the styles associated with each. To aid in this, we examined the histograms of different movement features for all shots attempted by players in each cluster, in the hopes that the differences between clusters could be illustrative of their styles.

We identified the shooter's average speed between catching and shooting, the Y-coordinate of the shot on the court, and the movement angle during the shot as movement features which are particularly revealing of the differences between clusters. FIG. 19 shows the distribution of shooters' average speeds between catching and shooting the ball for shooters in each cluster. FIG. 20 shows the distribution of shooters' Y-coordinate on the court, with the low and high values corresponding to the corners of the court while the middle values indicate shots taken somewhere at the top of the 3-point arc closest to the center of the court. Finally, FIG. 21 shows the distribution of movement angles in each cluster. Movement angles between 0 and 90 indicate movement towards the basket, while those between 90 and 180 indicate move-

ment away from the basket.

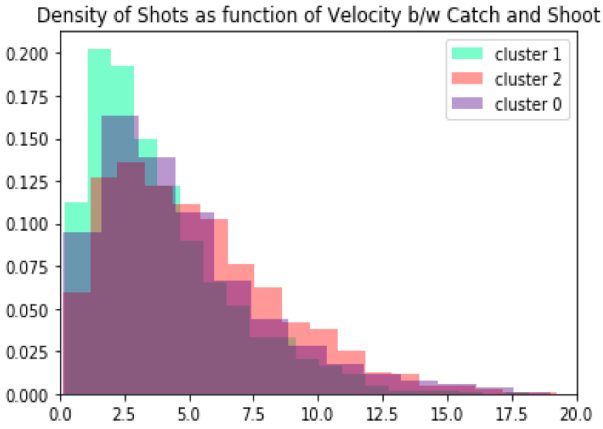

FIG. 19: Player velocity Between catching and shooting, grouped by cluster

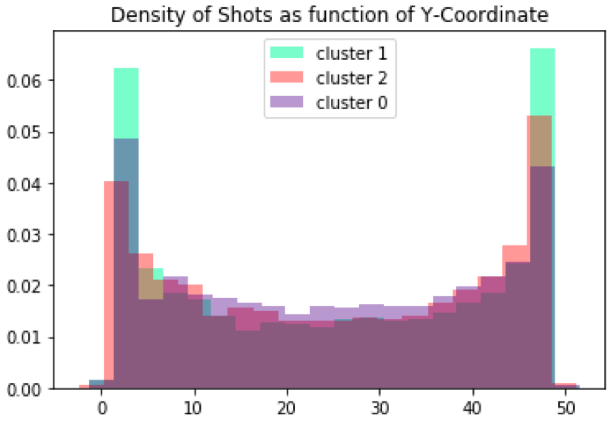

FIG. 20: Player Y coordinate when shot is taken, grouped by cluster

From these trends, we formulate the following cluster descriptions. Cluster 0 has players who move toward the basket while shooting, as seen in FIG.21, where movement angles for players in cluster 0 are generally lower than movement angles for players in the other clusters. Comparatively, players in these clusters also shoot more from the center of the court compared to the other clusters, as seen in FIG. 20. Looking at some of the players in the cluster, we see that many taller players with significant 3-point shooting are in this cluster. It is possible that these players stand on the top of the arc, step forward toward the basket and shoot over the defender. This means they wouldn't have to move perpendicularly to dodge a defender, as their size would mean they can just shoot from a height greater than the defender's block. Some players in this cluster who may fit this description are Dirk Nowitzki and Kevin Durant.

Cluster 1 has players who typically stand free to take

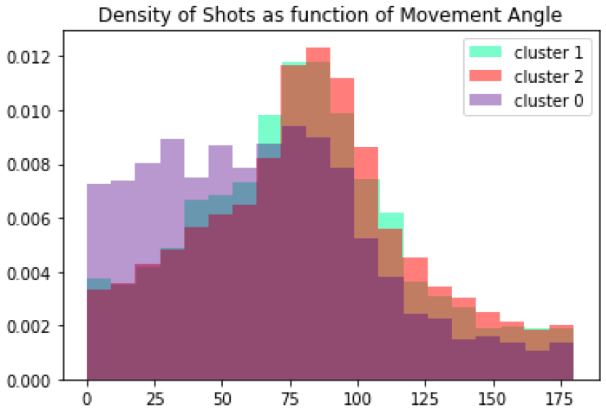

FIG. 21: Player movement angle before shooting, grouped by cluster

the open corner of the court before shooting. FIG. 20 shows that players in cluster 1 take shots from the corner of the court (high or low y values) much more than players in other clusters. Furthermore, their velocities are lower than players in other clusters (see FIG 19). They also take only 15% of their shots off the dribble. Examples of players who fit this description are JR Smith and JJ Redick.

Cluster 2 has players who show much greater perpendicular movement before shooting, as evidenced by their shot angles in FIG 21. They also show a greater velocity before shooting than the other clusters. We can think of them as shot creators who move sideways and backwards to generate space before shooting. It is also possible that they take shots as part of plays that involve much more movement from the players behind the three point line. Examples of players in this cluster include Steph Curry and Kobe Bryant.

It is also interesting to consider players who are classified into a particular cluster but are very close to other cluster. Kyle Korver is classified as being in cluster 2 (players with sideways movement). However, he is quite close to cluster 1 (open corner catch and shoot). This may be useful in identifying hybrid players who share some features of players from different clusters. Another example is LeBron James is also classified in cluster 2 but is close to cluster 0 (taller shooters).

It is important to note that this movement could be driven by both individual characteristics and team strategy. The movement characteristics of a player could depend on the kinds of plays that the team is making in the buildup to the shot. For example a corner catch-and-shooter likely takes up that position because the team plays in a way to create space for him. So these clusters should be interpreted as a combination between players' style and their role in the team three point strategy. We see some semblances of this where teams often seem to have players from different clusters. The Cavaliers, for example, had LeBron (cluster 2, but close to cluster 0)

and JR Smith (cluster 1).

A natural question is whether these styles of 3-point shooting lead to different success rates: is there a method of shooting 3-point attempts which can lead to higher success rate across all players who use it?

We found no significant difference between the shot conversion rate of players in different cluster. Cluster 1 and Cluster 2 players had an average conversion rate of 37.02% and 37.28%, respectively. Cluster 0 had a conversion rate of 34.98%, which is slightly lower but not much. This means that the clustering is not a function of "better" or "worse" shooter but likely only signify strategy and style.

VII. CONCLUSION AND FUTURE WORK

It is important to recognize that the role of data in sports is to help provide marginal improvements to how teams play. We must be careful before overstating how much data can impact an inherently human activity. Given that context, our research provides some inferences that can help understand these improvements. Some of the regression results we obtain match with intuition about what makes a good shot. While some, like the time between catch and shoot are less intuitive and can be explored further.

An area of future work that is interesting would be to see whether players change the clusters they are in after they change teams. This would give us insights on team strategy and how players fit into them. We would also get clarity of whether these clusters signify players' individual styles or team dynamics. People can do work using data-sets of later NBA seasons after players change teams to see this. One example for such analysis could be Kevin Durant who moved from Oklahoma to Golden State in 2017. Teams could potentially use analysis similar this in determining player recruitment to see which kinds of players fit the attributes and style they are looking for.

Other areas of future work would be using the characteristics of these clusters to examine if synergies exist between certain players, or with certain lineups. Perhaps players from certain clusters play well with players from other clusters or maybe even their own clusters. This could provide a data-driven approach for coaches to follow when making substitutions. It would also be interesting to see how the breakdown of a lineup into these clusters would influence offensive/defensive strategies, and vice versa.

[1] Total nba revenue 2001-2017. 2018.

[2] Kareem Abdul-Jabbar. The nba, and not the nfl, is the league of america's future. 2017.

[3] et al. Chang, Yu-Han. Quantifying shot quality in the nba. *MIT Sloan Sports Analytics Conference*, 2014.

[4] Rachel Marty. High-resolution shot capture reveals systematic biases and an improved method for shooter evaluation. *MIT Sloan Sports Analytics Conference*, 2018.

[5] Brett Meehan. Predicting nba shots. 2017.

[6] Megan Robertson. An analysis of nba spatio-temporal data. 2017.

[7] Luke Sandholtz, Nathan Bornn. Replaying the nba. *MIT Sloan Sports Analytics Conference*, 2018.