Ryan Smith

Mapping NBA Shot Quality

Purpose:

Most people who watch basketball can, at least at basic level, understand the value of taking an open, corner three pointer vs. taking a long, contested, two. But what is the exact difference in shot quality? On average, is it advantageous to pass up a semi-open elbow jumper for a not so open, top of the key, three? It's these types of shot decisions that can often win or lose a game and inevitably decide the success of a long season. This project aims to find out which shots are more worthy of an attempt by using techniques in multiple linear regression analysis. In short, I aim to find the correlation between different shot locations and other variables (namely, the affect of the defender) to the efficiency of a shot.

For management it can be extremely helpful to accurately model the makings of a high quality shot. This can eventually affect coaching decisions and influence everything from play designs, defending styles, and to who is encouraged to take what shot. It's also vital to hiring; it can help us fine-tune what players to pursue through free agency (and the draft) by telling us what traits and characteristics we should prioritize. In simple terms if we, as a team, taking higher quality shots than our opponent we should win most nights. But to do so, we have to dissect the makings of a shot and find out, on average, what makes one better than another.

Method:

For this project I used a dataset detailing shot attempts in the NBA (10,000 observations). The data included information like the specific shot coordinates, the dribbles taken before the shot, the velocity of the shooter, the angle of the shooter, of course whether the shot was made or missed, and more. The dataset also contained details pertaining to the closest defender; their distance, angle, and velocity were all included. Noticing that each XY coordinate could be associated to one actually foot on the basketball court, I calculated shot distance. From there I found out how many shots were 3pt field goals and awarded every made three an extra .5pts to its "made" variable. From there I had an easy way of calculating eFG% and went on to use this as the main dependent variable for much of the project. After toying with many variables, as well as different partitions of the court, I landed on mainly using shot distance, defender distance, defender angle, plus 6 basic shot regions (a view and description of these locations are illustrated in the sixLocs.png).

One challenge I worked on was finding an accurate representation of the "defender's affect", a variable I named openness_affect. The variable gets its name from combining a basic calculation for "shot openness" and a variable for "angle affect". "Shot openness" is the defender distance divided by the shot distance and is simple enough. It tells us how far the closest defender is relative to the distance of the shot.

For “angle affect” I played around with different defender angle partitions and landed on 5 angle groupings. The main idea revolving around the fact that a defender behind the shooter should have less of an affect on the shot than a defender directly in front of the shooter. A breakdown of these angle groupings can be found in the somewhat crude illustration, angleChart.png. Next was figuring out how much each angle grouping would affect the likelihood of a shot going in. For this I experimented with a few techniques. I tried logistic regression and different forms of linear regressions to better accurately pinpoint the correlation between the defender’s angle and the likelihood of a make. First I used my own values for each angle grouping and even attempted to use no grouping at all (I experimented with simply using the absolute value of the defender angle –this often included performing some form of data transformation to stabilize the variance in the defender’s angle). I landed on using a fairly basic linear regression, treating my 5 groupings as factor variables (these regression results can be seen in the factorAnglesReg.png). The correlation I found between these groupings and whether the shot was made worked as a multiplier for shot openness.

~ openness_affect = (defender distance / shot distance) * multiplier

Findings:

ThreeCourts.png shows some of the main findings of this project and breaks up all shots by Very Open (approx. 20% of shots), Semi Open (approx. 40% of shots), and Not Very Open (approx. 40% of shots). And while not every “open” shot is created equal, these partitions work as good general groupings to represent the affect the closest defender is having on the shooter. We than can compare the eFG% of each shot region. Right away it’s easy to see how openness affects shot quality (overall shots that are Very Open have 65.5 eFG%, while shots that are Not Very Open have a 33.7 eFG%). While its no revelation that open shots are better than contested ones, we can still make some interesting conclusions about specific shots.

For one, a Semi-Open corner three is better than a “Very Open” mid-range two. Also a “Not Very Open” corner three is a better shot than almost any “Semi Open” shot (that is not a Restricted Shot or Corner Three). These results also hint that a transition three (Very Open) is not advised vs. a layup unless of course that three is in the corner. But even in transition, a “Very Open” corner three could be hard to find (more on this later). Its also note worthy that this doesn’t include the chance of an offense rebound on a transition three, which would be an interesting project in and of itself.

A very simple application to these findings could be basic play calling; a coach could persistently run a play in the beginning of each shot clock designed to find an Open (Very to Semi) Corner Three. If unsuccessful the team could turn to sets that work to find a layup or an open paint attempt. This could also work in tandem with player development. This could involve training players to identify quality shots and then work to improve their ability to make them. One thing we can do is simply rank each

shot type by its eFG% (shot quality) but identifying quality shots is only part of the equation. We also have to consider how difficult it is to find a shot of that quality.

eFGShotBubble.png shows us a ranking of shots by their eFG% but also includes data that attempts to capture how easy those shots are to find. This is done by simply taking all shots attempted in a specific region and than finding how many were of that specific openness. Its important to note that this doesn't exactly show us how easy a shot is to find (i.e. a shot of a certain quality could be found but not taken and thus wont be captured by the dataset, also you could imagine a "Not Very Open" shot should be increasingly easy to find but again is not captured if that shot is not taken). With that said, it does give us valuable information. For instance if a shot of certain region and openness is almost completely absent from our dataset, we can assume it's not very easy to get. With all the info the bubble plot shows us we can see that "Very Open" top threes are of great quality but are few and far between in actual attempts. Overall we can conclude that shots like a "Very Open" Restricted, "Very Open" Non Restricted, and a "Semi Open" Corner Three are both of exceptional quality and also very attainable to find.

One important application of these findings is not only finding players that take and make high quality shots, but also, and perhaps more importantly, finding players that create open (or high quality) shots for others. With All-Star caliber players this is more or less obvious; as they demand attention and double teams, open shots become available to others. But there is, inevitably, diamonds in the rough, so to speak, that are perhaps aggressive and good at demanding attention but under the radar. We can then focus on if a player's teammates actually get open attempts while he is on the floor, somewhat avoiding the more obvious indicators like assists and essentially removing the noise of a specific make or miss (i.e. A player like this could be playing limited minutes and/or playing with below average players -2nd unit, bad team, etc.).

Notes:

The main regression model I used explains about 6% of the variation in the dataset, it can be found in the auxiliary.pdf, and is called olsReg.png. This auxiliary PDF also contains other illustrations I made while working on the project. One of which is specificLocs.png, which maps shots and eFG% to more specific areas of the floor. For this a more broad definition for openness affect was used (in which I created a openness_affect_bool dummy variable and set 30% of shots as open, and the other 70% as not).

Python was used prominently for my data manipulations and visualization building (matplotlib). While I also used it for some statistical analysis, I mostly used STATA for most of my statistical work (different regression techniques, tabulating variables, etc.). While I have done regressions and statistical work in python (and R), I find the quickness and ease of use within STATA to be slightly superior. I also have plenty of ideas of additional variables that could greatly increase the

explanatory power of a model for FG% (and eFG%). And if given the time, I have thought of ways to increase the explanatory power of my openness_affect variable.