Cole Fisher
Intern, Toronto Raptors Basketball Operations

Predicting NBA 3FG% using Multiple Linear Regression

Abstract

A player's ability to shoot the basketball is a very integral aspect of the game of basketball, so being able to predict NBA shooting ability from college statistics cannot be overstated. Unfortunately, a player's ability to make three pointers (3FG%) is not something that translates perfectly well from the NCAA to the NBA. However, this paper will propose a method to predicting NBA 3FG% that performs better than the following two approaches:

a) Assume that everyone will shoot between 33% and 37% (league average is typically 35%)
b) Assume that everyone will shoot a similar percentage to their NCAA 3FG% (plus/minus 2%)

The linear regression model that exceeds the predictability of approaches a) and b) is:

0.129380 + (NCAA 3FG% * 0.2100) + (NCAA FT% * 0.1877)

Please note: there is a glossary at the conclusion of this paper that explains the basketball metrics used throughout.

The Problems

- The 3 point line arc radius in NCAA is 6.32 m whereas in the NBA it is 7.24 m
    - This is a significant difference – a player needs to adjust to the NBA 3 point line. Being able to consistently hit 3FGs in NCAA is physically not the same thing as doing so in the NBA.
- Quality of teammates and differences in offensive systems
    - A player's ability to hit an in-game 3FG is quite dependent on the quality of the shot he gets. It is much easier to hit an uncontested shot than it is to hit a heavily gaurded one. An incredible, play-making point guard that demands the defense's undivided attention might get his teammates easier, more open shots than an average point guard.
    - Different coaches run different offenses that prioritize different aspects of the game. A 3 point shooter playing in a fast-paced, run-in-transition style offense

might get different quality looks than a shooter in a slow, big-man-oriented style offense.

- Small sample sizes
    - Many great NBA prospects play just one or two seasons of NCAA basketball and then declare for the NBA draft. This limits the data used to make predictions – a player might play 25 games in one NCAA season, and he might attempt 4 3FGs per game. By all accounts, a player attempting 4 3FG per game is a comfortable shooter, but that totals only 100 3FGs. This means that *five shots* can be the difference between what we consider an average shooter (34%) and a very good shooter (39%). Therefore, it is important that we consider the volume of shots attempted when we make predictions: surely we do not want to label someone as a 'bad shooter' because he made five less shots in the NCAA than someone considered a 'good shooter.'
        - To somewhat deal with this challenge, the data collected is filtered; only players that have attempted 150 3FGs in both college and the NBA are included.

Data Collection Process

1. I first built a database that aggregates all players' ids from different websites commonly used to collect basketaball data. This database makes automated data scraping effortless because it removes the possibility of gathering incorrect data.
    a. For example, if there are many players named "Joe Smith," for example, this database can be used to ensure that the correct Smith is being scraped. As well, different sites use different spelling conventions: ESPN might use "JR Smith," whereas basketball-reference.com might say "J.R. Smith".

uid	name	birthday	basketball_ref_ID	NBA_com_ID	ESPN_id	sports_ref_ID
1	Derrick Alston	Aug 20 1972	/players/a/alstode01.html	/player/#!/308/	3624	/cbb/players/derrick-alston-1.html
2	Travis Best	Jul 12 1972	/players/b/besttr01.html	/player/#!/696/	58	/cbb/players/travis-best-1.html
3	Melvin Booker	Aug 20 1972	/players/b/bookeme01.html	/player/#!/511/	3645	/cbb/players/melvin-booker-1.html
4	Lazaro Borrell	Sep 20 1972	/players/b/borrela01.html	/player/#!/1953/	78	NULL
5	Shawn Bradley	Mar 22 1972	/players/b/bradlsh01.html	/player/#!/762/	88	/cbb/players/shawn-bradley-1.html
6	Rick Brunson	Jun 14 1972	/players/b/brunsri01.html	/player/#!/1594/	106	/cbb/players/rick-brunson-1.html
7	Randolph Childress	Sep 21 1972	/players/c/childra01.html	/player/#!/719/	1693	/cbb/players/randolph-childress-1.html
8	Robert Churchwell	Feb 20 1972	/players/c/churcro01.html	/player/#!/928/	3663	/cbb/players/robert-churchwell-1.html
9	Bill Curley	May 29 1972	/players/c/curlebi01.html	/player/#!/223/	174	/cbb/players/bill-curley-1.html
10	Yinka Dare	Oct 10 1972	/players/d/dareyi01.html	/player/#!/212/	184	/cbb/players/yinka-dare-1.html
11	Ben Davis	Dec 26 1972	/players/d/davisbe01.html	/player/#!/989/	191	/cbb/players/ben-davis-1.html

Figure 1: First 11 rows of the id database

2. I scraped basketball-reference for NBA stats using the following query:

   For combined seasons; played in the NBA; in the regular season; from 2002-03 to 2016-17; requiring Games >= 50 and Minutes Per Game >= 7 and 3-Pt Field Goal Attempts >= 150; sorted by descending 3-Pt Field Goals

   I scraped sports-reference for NCAA stats using the following query:

   For combined seasons; from 2001-02 to 2015-16; requiring Games >= 20 and 3-Point Field Goal Attempts >= 150; sorted by descending Points

Along with their stats from each website, I made sure to include each player's id in the scrape. I outputted the data to csv files. That allowed me to use the id database to loop through the csv files find all the players in the NBA csv that had matches in the NCAA csv. In the end, I got 225 rows of data to work with.

```
def filter_data(file1, file2, newfile):

    test_data = open(newfile, 'w')
    w = csv.writer(test_data, delimiter=",")

    with open(file1, 'r') as nba_file:
        nba = csv.reader(nba_file, delimiter=",")

        for row_nba in nba:
            if row_nba:
                nba_id = row_nba[0]
                ncaa_id = get_ncaa_id(nba_id)

                with open(file2, "r") as ncaa_file:
                    ncaa = csv.reader(ncaa_file, delimiter=",")
                    for row_ncaa in ncaa:
                        if row_ncaa:
                            id = row_ncaa[0]
                            if ncaa_id == id:
                                print("MATCH: " + row_nba[1])
                                row_to_add = row_ncaa + row_nba
                                w.writerow(row_to_add)

                ncaa_file.close()
    nba_file.close()


def get_ncaa_id(nba):

    cnx = mysql.connector.connect(host='localhost', user='root', database="players_db")
    cursor = cnx.cursor()
    get_ncaa = (
        "SELECT sports_ref_ID from players "
        "WHERE basketball_ref_ID = %s"
    )

    cursor.execute(get_ncaa, (str(nba),))
    ncaa_db = cursor.fetchone()
```

```
    if ncaa_db is None:
        return None
    return ncaa_db[0]
```

Figure 2: python functions that compare the players from the nba and ncaa scrapes, and outputs matches to a new file

Results

Scatter Plots

The following scatter plots show the relationship between certain NCAA shooting metrics and NBA 3FG%.

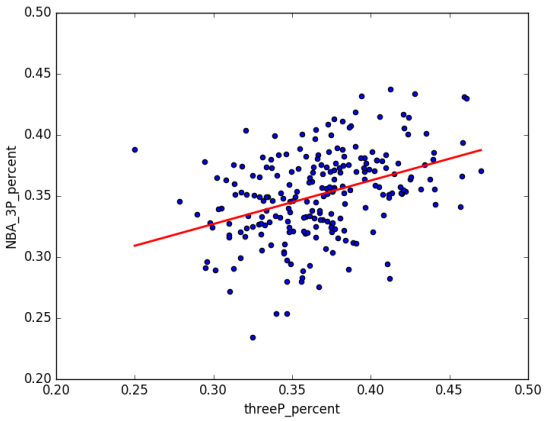

Figure 3: NCAA 3FG% vs NBA 3FG%
R-squared: 0.15197:

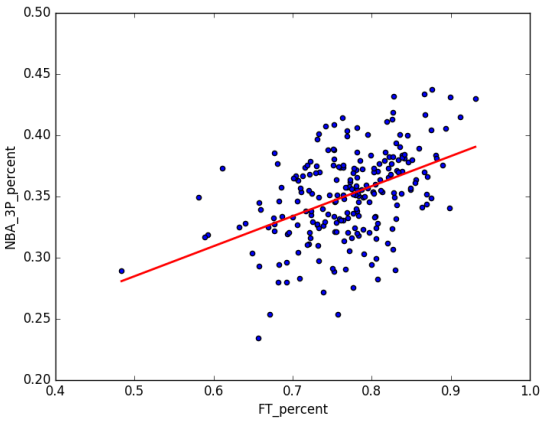

Figure 4: NCAA FT% vs NBA 3FG%
R-squared: 0.20485

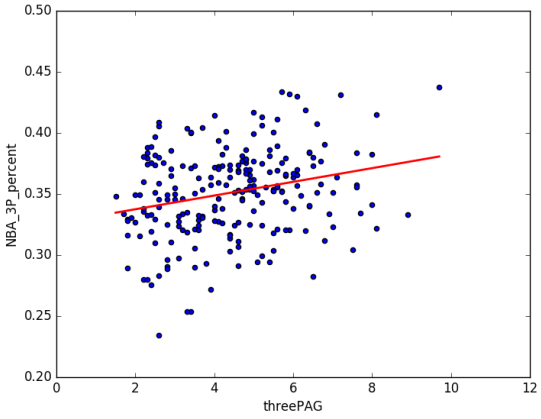

Figure 5: NCAA 3FGA per game vs NBA 3FG%
R-squared: 0.06444

I decided to use NCAA 3FG%, FT% and 3FGA per game as a starting point. Based on the scatter plots, there does seem to be a relationship between NCAA 3FG% and NBA 3FG% as well as

between NCAA FT% and NBA 3FG%. The relationship between NCAA 3FGA per game and NBA 3FG% is less promising.

It makes sense that FT% appears to be related to NBA 3FG% ability. The free throw measures a player's ability to simply shoot the basketball without needing to worry about defense interfering. As a result, it offers the purest, cleanest insight into how a player approaches a simple basketball shot. Good FT% indicates strong shooting form, so NBA front offices believe that a player with a good NCAA FT% can develop into a reliable NBA 3-point shooter.

Unlike FT% which isolates one's ability to shoot, there is lots of "noise" surrounding NCAA 3FG% - it can be effected by teammates, spacing issues, quality of defense, and more. I think this explains why the correlation is weaker than it is with FT%.

Regression Analysis

I used scikit-learn to turn my dataset into training and testing data in order to achieve proper cross validation. I randomly split the data into 180 rows of training data (80%) and 45 rows of testing data (20%). Then I used the training data to build a regression formula which was used to make NBA 3FG% predictions for the test data.

I tried two models: one that incorporated NCAA 3FG%, NCAA FT% and NCAA 3FGA per game (model 1), and one that only used NCAA 3FG% and NCAA FT% (model 2).

After running this experiment 500 times and averaging the results, model 2 emerged more predictive.

What does it mean to be 'predictive'?

There are different ways that people use to predict a player's NBA 3FG%, and I wanted to compare my regression model's prediction to these more traditional approaches.

Option 1 - **Assume the average**: assume that the player will average 35% 3FG% plus/minus 2%

Option 2 - **Assume the same as NCAA:** assume that the player will average his NCAA 3FG% plus/minus 2%

Option 3 - **Regression model:** 0.129380 + (NCAA 3FG% * 0.2100) + (NCAA FT% * 0.1877)
a players' NBA 3FG% will be equal to his predicted 3FG% plus/minus 2%

After 500 iterations, the results are as follows (on average):

Option 1: 18.63/45 players have an NBA 3FG% between 33% and 37%

Option 2: 15.02/45 players have an NBA 3FG% equal to their NCAA 3FG% +/- 2%

Option 3: **22.56/45** players have an NBA 3FG% +/- 2% of the model's 3FG% prediction

What does this mean?

Essentially, this means that the regression model can predict a player's NBA 3FG% +/- 2% with approximately **50.1% accuracy**. This beats out the two other options by a significant margin. Predicting a player will shoot a similar NBA 3FG% to his NCAA 3FG% +/- 2% is only accurate 33.3% of the time, and assuming someone will fall in between the average NBA range (33-37%) only achieves 41% accuracy.

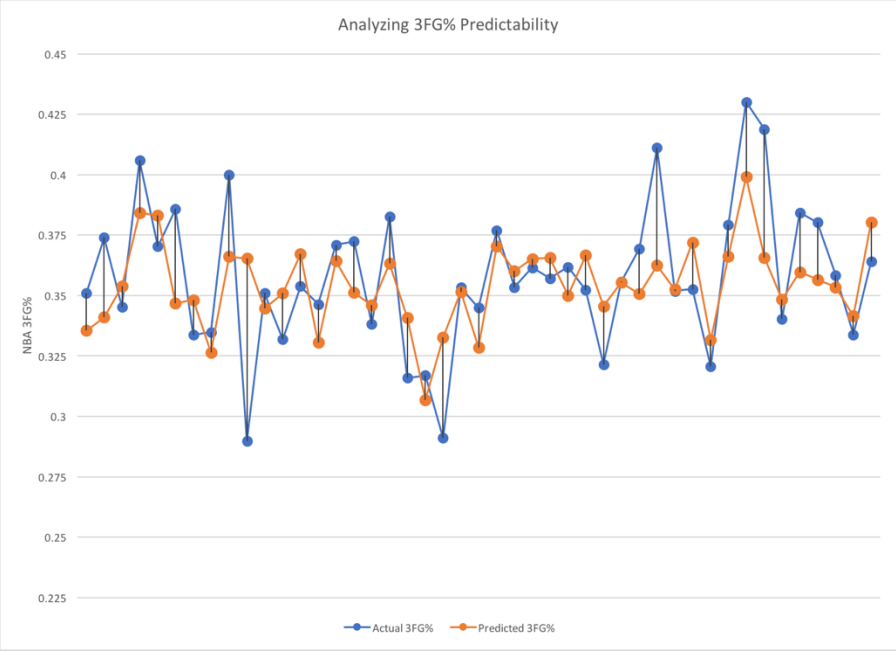

Figure 6: Prediction vs Actual 3FG% for a random 45 player sample

Figure 6 shows the accuracy of the regression model. Typically, the prediction is on the right track – however, it seems to be a little bit too conservative. It does not accurately predict great shooters nor does it accurately predict poor shooters.

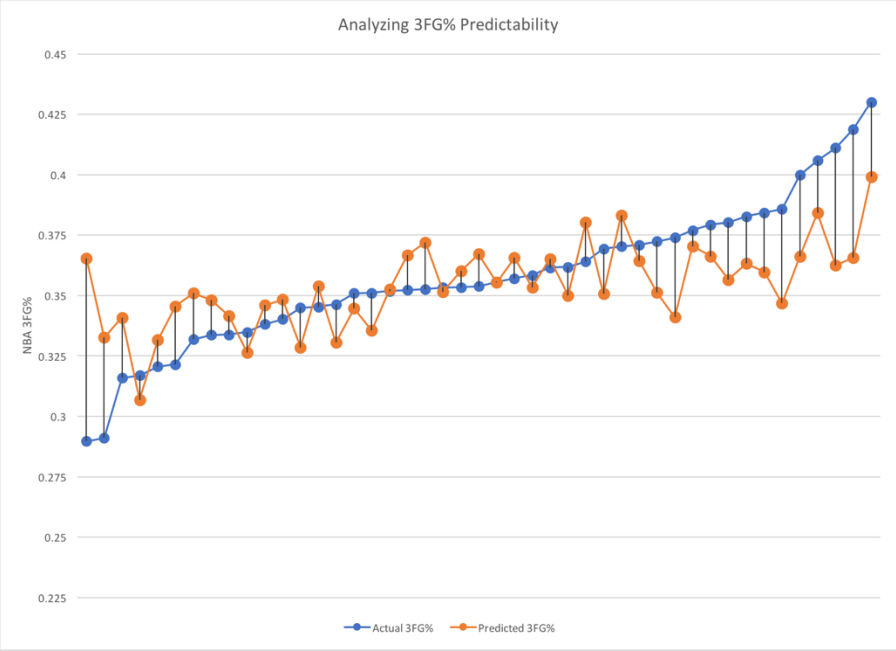

Figure 7: Same data as figure 6, but sorted in ascending order for actual 3FG% (line in blue)

Figure 7 uses the exact same data as figure 6, but this time the data is sorted in ascending order for actual NBA 3FG%. This sorting allows the limitations of the model to become apparent: the model is not good at predicting the shooters that fall outside the 33% - 37% range.
This limitation makes sense: regression tends to *regress*.

I wanted to figure out why the model fails to predict the extreme shooters (good and bad). My first consideration was volume-related. I didn't incorporate NCAA 3FGA per game into my model, so I wondered if perhaps my model struggled to make accurate predictions for players who attempted a relatively low number of three pointers.
For example, if player X happens to hit 60/150 3FGs, his 3FG% is 40%. However, if he hits 50 of those (only a mere 10 shot difference!), his 3FG% drops dramatically to 33.3%. If we keep his FT% constant at 75%, his predicted NBA 3FG% for each scenario is:

- predicted NBA 3FG% = 0.129380 + (0.40 * 0.2100) + (0.75 * 0.1877) **= 0.354%**
- predicted NBA 3FG% = 0.129380 + (0.33 * 0.2100) + (0.75 * 0.1877) **= 0.339%**

That is a 1.5% difference simply because of ten 3 pointers. Surely shot volume must be taken into consideration. However, my model did not improve when I included NCAA 3FGA per game (in fact it became less predictive).

Going forward, the next step should be deeper analysis of shot volume. The following figure helps to illustrate the lack of a relationship between the model's accuracy and NCAA shot volume.

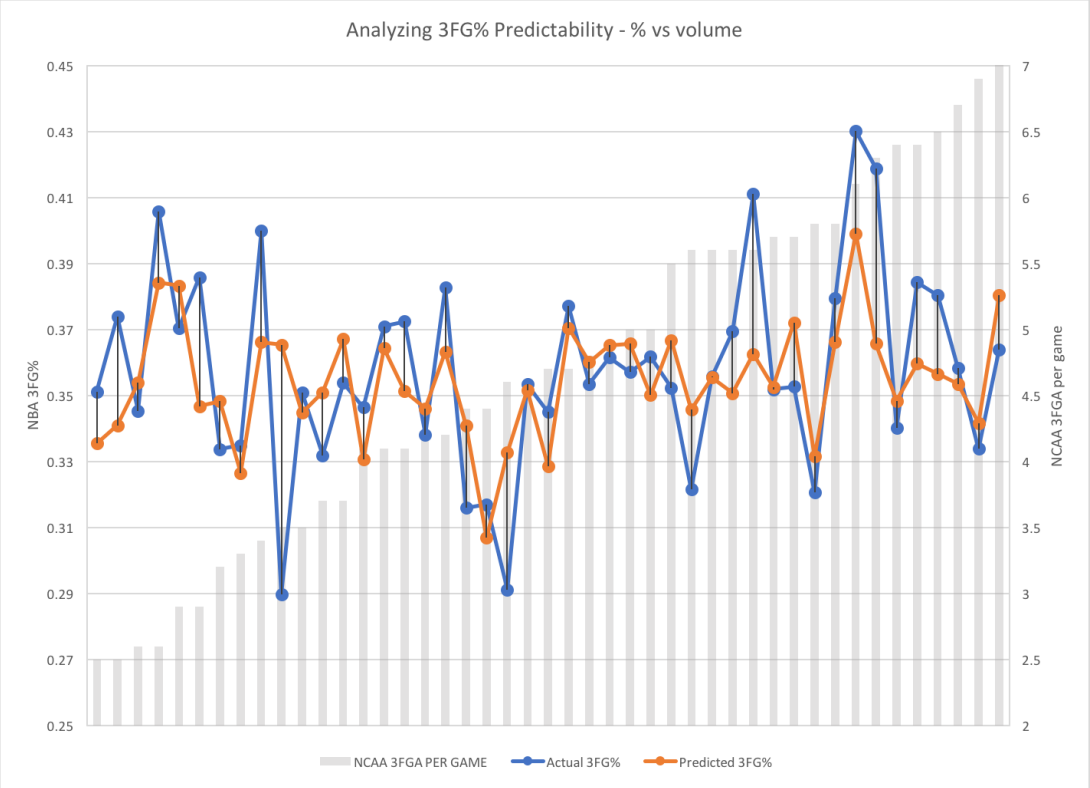

Figure 8: Same data as figure 6/7, sorted for NCAA 3FGA per game

There does not appear to be a correlation between NCAA three point shot attempts per game and overall difference between actual NBA 3FG% and predicted 3FG% (or in other words, the model's error). The model makes poor predictions when shot volume is low, medium and high. The opposite is true as well: the model does quite well for various shot volumes. The following figure further explains the lack of any correlation.

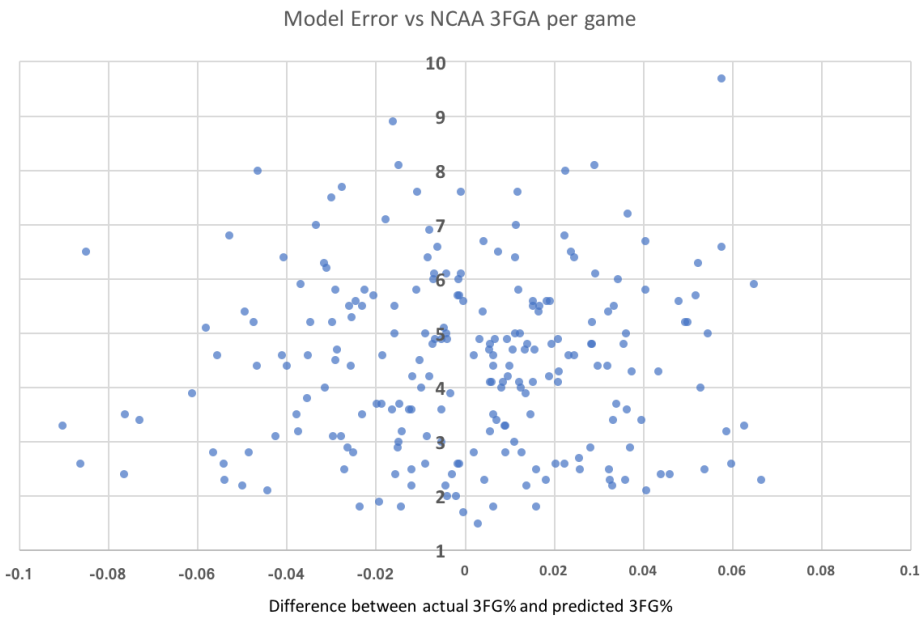

Figure 9: Scatter plot of relationship between the model's error and NCAA 3FGA per game

There looks to be essentially no association between NCAA 3FGA per game and the precision of the model's NBA 3FG% prediction.

Overall, I think my next step in this project would be to figure out how shot volume can be incorporated. I think it certainly deserves consideration in the model but so far the results indicate a lack of a relationship.

Conclusions

NBA 3FG% is mostly unpredictable – the model's 50% accuracy is not precise enough for an NBA team to feel good about using it, however, it does offer a reasonable starting point. The model beats the two traditional, most commonly used approaches to predicting NBA 3FG%. Yet, there is even more room for growth. There must be some sort of volume component that will prove helpful, considering it is currently unaccounted for in the model. Additionally, I think looking at shooting trends over the course of a player's NCAA career might prove to be a worthwhile endeavor. I speculate that a steady year-over-year improvement in free throw shooting implies that a player has improved his shooting mechanics. This bodes well for his future ability to become an NBA three point shooter.

Overall I would conclude the following:

- This regression model is a better predictor than looking only at a player's NCAA 3FG%, which is currently a widely-used approach.
- This model's 50% accuracy deems it a 'work-in-progress.' It should not yet be used to make *conclusions* regarding a prospect's future NBA 3FG% but it does offer a solid foundation for open discussion.
- Shooting is mostly unpredictable – players that were good shooters in college have often been unable to figure out the NBA three pointer and many non-shooters in college have become elite NBA marksmen.
    - This has important implications for NBA organizations: **players can develop shots over time**. Investing heavily in player development and drafting players who are undervalued because of poor shooting metrics might be worthwhile strategies.

Predictions

I have used my model to make predictions for the 2017 draft class (only those who played in the NCAA.) It is sorted by predicted NBA 3FG%.

Some of the players below would not qualify for this study because they did not meet the minimum requirement of 150 NCAA 3FGAs.

Pick#	Name	School	3FG%	FT%	**Predicted NBA 3FG%**
54	Alec Peters	Valparaiso	0.416	0.846	0.375543796
7	Lauri Markkanen	Arizona	0.423	0.835	0.374949207
12	Luke Kennard	Duke	0.382	0.867	0.372344892
10	Zach Collins	Gonzaga	0.476	0.743	0.36881163
46	Sterling Brown	Southern Methodist	0.451	0.77	0.368629109
39	Jawun Evans	Oklahoma State	0.407	0.818	0.368397969
11	Malik Monk	Kentucky	0.397	0.822	0.367048587
29	Derrick White	Colorado	0.396	0.813	0.36514925
24	Tyler Lydon	Syracuse	0.398	0.809	0.36481848
37	Semi Ojeleye	Duke & SMU combined	0.415	0.785	0.363883955
34	Frank Mason	Kansas	0.42	0.763	0.360804606
3	Jayson Tatum	Duke	0.342	0.849	0.360565496
17	D.J. Wilson	Michigan	0.363	0.817	0.358969431

44	Damyean Dotson	Oregon & Houston combined	0.38	0.796	0.358598012
32	Davon Reed	Miami (FL)	0.395	0.779	0.358557363
51	Monte Morris	Iowa State	0.381	0.78	0.355804799
18	TJ Leaf	UCLA	0.466	0.679	0.354698512
45	Dillon Brooks	Oregon	0.362	0.794	0.354442266
41	Tyler Dorsey	Oregon	0.416	0.732	0.354145768
31	Frank Jackson	Duke	0.392	0.755	0.353422458
26	Caleb Swanigan	Purdue	0.376	0.76	0.351000664
55	Nigel Williams-Goss	Washington & Gonzaga combined	0.331	0.806	0.350184101
53	Kadeem Allen	Arizona	0.4	0.725	0.34947155
6	Jonathan Isaac	Florida State	0.348	0.78	0.348874172
56	Jabari Bird	University of California	0.37	0.753	0.348426636
13	Donovan Mitchell	Louisville	0.329	0.788	0.346385427
30	Josh Hart	Villanova	0.389	0.72	0.346222831
48	Sindarius Thornwell	South Carolina	0.34	0.771	0.345504702
42	Thomas Bryant	Indiana	0.373	0.718	0.342487123
2	Lonzo Ball	UCLA	0.412	0.673	0.342231274
35	Ivan Rabb	University of California	0.409	0.666	0.340287303
9	Dennis Smith Jr.	North Carolina State	0.359	0.715	0.338983751
16	Justin Patton	Creighton	0.533	0.517	0.338362061
1	Markelle Fultz	Washington	0.413	0.649	0.337936445
15	Justin Jackson	North Carolina	0.339	0.713	0.334407967
40	Dwayne Bacon	Florida State	0.312	0.733	0.332491494
59	Jaron Blossomgame	Clemson	0.315	0.723	0.331244531
33	Wesley Iwundu	Kansas State	0.338	0.689	0.3296931
52	Edmond Sumner	Xavier	0.285	0.731	0.326445577
5	De'Aaron Fox	Kentucky	0.246	0.739	0.319756452
36	Jonah Bolden	UCLA	0.25	0.733	0.319470316
4	Josh Jackson	Kansas	0.378	0.566	0.315006514
27	Kyle Kuzma	Utah	0.302	0.631	0.3112457
23	OG Anunoby	Indiana	0.365	0.522	0.304017379
38	Jordan Bell	Oregon	0.188	0.627	0.286552726
19	John Collins	Wake Forest	0	0.729	0.266214758
14	Edrice Adebayo	Kentucky	0	0.653	0.251949406
28	Tony Bradley	North Carolina	0	0.619	0.245567538
22	Jarrett Allen	Texas	0	0.564	0.235243928
47	Ike Anigbogu	UCLA	0	0.535	0.22980057
20	Harry Giles	Duke	0	0.5	0.223231

Cole Fisher
Intern, Toronto Raptors Basketball Operations

Glossary

3FG%	Three point percentage
3FGA per game	Number of three pointers attempted per game
FT%	Free throw percentage
NBA	National Basketball Association
NCAA	National Collegiate Athletic Association (college)