The Impact of Biometric Parameters on the Success of NBA Draft Picks

Adarsh Kannan, Brian Kolovich, Brandon Lawrence, Sohail Rafiqi

Southern Methodist University

6425 Boaz Lane

Dallas, TX 75205

**Abstract.** In this paper, we present a machine learning based approach to projecting the success of National Basketball Association (NBA) draft prospects. With the proliferation of data, analytics have increasingly become a critical component in the assessment of professional and collegiate basketball players. The goal of this study is to leverage player biometric data, college statistics, draft selection order, and positional breakdown as modelling features into our prediction algorithms. From this, we hope to gain insight into the key metrics that determine future success. These findings may be useful in aiding basketball personnel in their evaluation and projection of NBA draft prospects.

1 Introduction

THIS NEEDS TO BE REWRITTEN. WHAT IS YOUR PROBLEM STATEMENT?

The National Basketball Association (NBA) has been holding an annual draft for prospects since 1947. The draft is responsible for the highest influx of new players to the league. Recently, analytics in the NBA has become a market opportunity, with a slew of highly educated and experienced professionals in the data science industry taking over higher management positions in the realm of player personnel. This has created a whole new style of basketball and a more efficient manner of selecting players.

One area where data science has not been advanced is in the field of biometrics. THIS IS A FALSE STATEMENT.

Biometrics refers to the branch of statistics that deals with data relating to living organisms.

The biometric parameters collected are determined by the NBA Draft combine, an event that is held before the NBA Draft every year. The parameters include height without shoes, height with shoes, wingspan, standing reach, maximum vertical, maximum vertical reach, vertical with no step, vertical with no step reach, weight, body fat percent, hand length, hand width, number of bench presses, agility, and sprint time.

Basketball statistics have evolved in the last twenty years. Since the advent of analytics in sports, initially in baseball, then trickling into the other major sports, basketball analysis has also improved. Originally statistics that were measured included minutes per game, points per game, rebounds per game, assists per game, blocks per game, steals per game, field goal percentage, three point field goal percentage, and free throw percentage. However, with the field of data science and data analytics making great strides, other statistics have made their way into the daily lexicon of basketball

The advanced statistics that we have used in this study include win shares (WS), win shares per 48 minutes (WS/48), box plus/minus (BPM), and value over replacement (VORP). WS indicates the number of wins a player is responsible for in his career. WS/48 indicates the number of a wins a player is responsible for per 48 minutes – the length of a regulation NBA game. BPM indicates the plus or minus difference with regards to the opposition while the player is on the court playing. VORP indicates the value over the replacement value of the average player.

The National Collegiate Athletic Association (NCAA) consists of 351 teams in 32 conferences in Division 1 play along with many other teams from lesser talented divisions. This dataset will not include players who were drafted either from international competition or from high school. Also, players who were not drafted will not be included. This was decided to determine a type of consistency and the inability to determine a rating of competition for college basketball versus international and high school basketball. Traditionally there are five positions in basketball – point guard, shooting guard, small forward, power forward, and center. However due to the increasingly offensive and higher scoring nature of the NBA over the last twelve years that consists of a much higher volume of three point field goals attempted and scoring, basketball has evolved into being more position-less. Hence, players analyzed in this project will be classified into the following three categories: guards, swing, and big. Guard players consist of players who are viewed as only point guards, or point guards/shooting guards. Swing players consist of players who are viewed as shooting guards, small forwards, shooting guards/small forwards, or small forwards/power forwards. Big players consist of power forwards, centers, or power forwards/centers.

2 NBA Draft Overview

THIS SECTION IS ABOUT THE DRAFT PROCESS AND WHO PARTICIPATES IN IT. NEED A DIFFERENT SECTION ON THE METHOD AND PARAMETERS THAT GO INTO ACTUALLY MAKING A CHOICE OF A PERSON. FUTURE ATHLETIC PERFORMANCE IS ONLY ONE PARAMETER. YOU HAVE MIXED AND CONFUSED THESE TWO VERY DIFFERENT STEPS IN THE PROCESS.

It is important to note that little peer reviewed research exists concerning predicting player performance using publicly known biometric factors such as height or weight. This can be due to publication bias or lack of interest. Either way, this classification problem is very popular in the amateur space. Popular statistics sites like FiveThirtyEight have created statistical models to predict player performance going into the NBA Draft ([Paine](https://fivethirtyeight.com/features/projecting-the-top-50-players-in-the-2015-nba-draft-class/)), but very little of this work has been published.

The NBA draft functions on a combination of lottery and inverse record placement ([NBA.com](http://www.nba.com/news/draft-lottery-history/)WHAT IS THIS? IF IT’S A CITATION, CITE IT. IF IT’S FROM THE NBA SITE, THEN PUT IT AS A FOOTNOTE – WITH WORDS EXPLAINING IT, OF COURSE). Fourteen ping-pong balls are drawn from a lottery machine to determine the draft order for the bottom fourteen teams in the league. The lottery is weighted to give the teams with the worst records the best odds of receiving an early draft pick. The remaining sixteen teams are then given their draft picks based on their performance in the previous season. This process is meant to ensure equity while minimizing perverse incentives such as poor performing teams throwing games to gain a draft advantage in the next season.

There are a few considerations in the draft pick equation. First, a draft pick is a strategic asset for a team. Lower picks yield a larger, more talented selection pool for the team. Picks can be used in trading packages for other players in the league. For example, if a team knows that the players they want are unlikely to be drafted in the first round, they may trade their draft position to another team in exchange for something that the team would value more. With picks, the lower it is, the more valuable it is.

Player salary is another component of the cost equation. Teams seek to maximize performance for their expenses. NBA teams are subject to a salary cap ([NBA.com](http://www.nba.com/article/2017/07/01/nba-salary-cap-set-2017-18-season-99093-million)). Above the expense for performance, this represents an opportunity cost with second and third order consequences for teams close to the cap considering possibilities such as trades and maintaining existing contracts.

The draft can be simplified into a breakeven exercise where the opportunity cost is less than or equal to the value that a given player brings to the team. This is deceptively simple, as each component discussed earlier is weighed against a complete unknown – the player’s future performance. Any insight or edge that a team can have on draft day will pay dividends over the course of the player’s career for the team.

Amateurs that participate in the NBA draft can come from one of three sources – high schools, universities in the National Collegiate Athletic Association (NCAA), or foreign countries (Berri). The data available on each player is different at each level and direct performance comparisons are difficult. Berri discussed this issue when attempting to anticipate draft pick performance based on past collegiate performance statistics. This is why we are limiting our scope to NCAA athletes as well. Berri’s primary finding was that scoring, while valued by the decision makers, was a poor predictor of future success in the NBA. Studies have shown that NBA teams and other competitive basketball clubs recruit on height (Berri, Treme). Another study, by Radzevick in 2016, concluded that an elite player’s contribution to a team was tied with their previous transitional experience during their playing career as much as skill or experience.

Key biometric information is gathered at the NBA draft combine, so all players are on a level playing field from a data perspective.

NEED A SECTION ON THE BIOMETRICS AND HOW THEY ARE COLLECTED…IE A SECTION ON THE DATA AND ITS COLLECTION METHOD

3 Dataset Overview

3.1 Dataset Overview

Our data is predicated on our ability to design a model that could predict the probability of ‘success’ of NBA draft combine participants. Due to sparseness in historical data, we restrained the data to the 2009-2014 NBA draft combines and to the players who participated in *all* biometric measurements. 2014 was chosen as the cut-off year since we believe anything less than that was not sufficient time for a player to realize whether he would be able to remain the NBA. The final dataset in the analysis consists of 194 records across 30 columns.

| **Variable** | **Description** | **Additional Details** | **Variable Type** |
| --- | --- | --- | --- |
| player | Player name | n/a | Categorical |
| college | Player college | n/a | Categorical |
| draft\_yr | Year drafted | n/a | Ordinal |
| fnl\_coll\_rpi | Final Ratings Percentage Index of player's final college season | n/a | Ordinal |
| still\_in\_league | Still playing in NBA, as of start of 2017 NBA season | 1 = yes  0 = no | Numeric |
| age\_first\_yr | Age at the start of 2017 NBA season | n/a | Ordinal |
| draft\_pick | Order of selection in player's respective draft | 1-60 = drafted  61 = undrafted | Numeric |
| hght\_noshoes | Height w/o shoes, (inches) | n/a | Numeric |
| hght\_wtshoes | Height w/ shoes, (inches) | n/a | Numeric |
| wingspan | Wingspan (inches) | n/a | Numeric |
| Ssanding\_reach | Standing reach (inches) | n/a | Numeric |
| vert\_max | Max vertical leap (inches) | n/a | Numeric |
| vert\_maxreach | Max reach from vertical (inches) | n/a | Numeric |
| vert\_nostep | Vertical w/t no steps (inches) | n/a | Numeric |
| vert\_nostep\_reach | Reach from vertical w/ no step (inches) | n/a | Numeric |
| weight | Weight (lbs) | n/a | Numeric |
| body\_fat | Body fat (%) | n/a | Numeric |
| hand\_length | Hand length (inches) | n/a | Numeric |
| hand\_width | Hand width (inches) | n/a | Numeric |
| games | Total number of games played in college | n/a | Numeric |
| pts\_ppg | Average points per game from college career | n/a | Numeric |
| trb | Average rebounds per game from college career | n/a | Numeric |
| ast | Average assists per game from college career | n/a | Numeric |
| fg2\_pct | Average 2 point field goal percentage from college career | n/a | Numeric |
| fg3\_pct | Average 3 point field goal percentage from college career | n/a | Numeric |
| ft\_pct | Average free throw percentage from college career | n/a | Numeric |
| guards | Binary variable indicating guard position | 1=guards  0=not guards | Numeric |
| forwards | Binary variable indicating forward position | 1=forwards  0=not forwards | Numeric |
| centers | Binary variable indicating center position | 1=centers  0=not centers | Numeric |
| drafted | Binary variable indicating drafted or undrafted | 1=drafted  0=undrafted | Numeric |

The biometric data, player name, and draft year were pulled and aggregated directly from a user uploaded dataset via the following Web site: <https://data.world/achou/nba-draft-combine-measurements>.

It consists of a basketball player’s biometric statistics from their respective NBA draft combine, which is held annually prior to the actual draft. The biometric statistics consist of wingspan, hand size, vertical jump, etc.

Additionally, the college statistics and college attended data were scraped from the following Web site and combined with the biometric data: <https://www.sports-reference.com/cbb/players>. For college statistics, we used the averages for all statistics along with total games, which is just the total number of games played.

To derive the final ratings percentage index variable, we downloaded the 2009-2014 ratings percentage index data and cross-referenced it against the player’s college and draft year:

<https://www.ncaa.com/rankings/basketball-men/d1/ncaa-mens-basketball-rpi>

Lastly, we created custom features by referencing existing variables or through manual reconciliation from Web research.

* *Still\_in\_league* - determined by whether the player was actively in the NBA, as of the beginning of the 2017 NBA season; coded as 1 if the player is still in the league; else 0
* *Age\_first\_year* – determined by the age of player at the starting of the NBA season
* *Guards* – coded as 1 if the player is either a point guard or shooting guard; else 0
* *Forwards* – coded as 1 if the player is either a small forward or power forward; else 0
* *Centers* – coded as 1 if the player is a center; else 0

*Drafted* – coded as 1 if the player is drafted in his respective draft; else 0

4 Exploratory Data Analysis

4.1 Pre-Modelling

Due to the multivariate nature of the model comprised of several data on different scales, e.g. average assists per game ranges from 0.1-8.0 while final RPI ranges from 1-206, we standardized the feature values across the dataset. By stabilizing the range and variability of the data, we reduce the risk of certain features exhibiting unequal contributions to the model predictions.

As an initial exploration of the data, we analyzed the distributions from the categorical variables to identify any potential pitfalls or sparseness in the population data. Figure 1 represents the distribution of the players by basketball position. From this, we see that the highest concentration is from the guard position, which comprises ~47% of the population. Forwards comprise ~38% and centers encompass ~15% of the data, respectively. While the position breakdown is skewed towards guards and forwards, it makes sense given that we consolidated the point/shooting guards as guards, small/power forwards as forwards, and centers as a stand-alone position.

*Figure 1*

We then assessed the distribution of players by draft year, as seen in Figure 2. As noted in previous sections, biometric and data in general was inconsistent and sparse in the early to late 2000s. While the percentage of 2009 participants is significantly lower than the other drafts, there are no fundamental differences in how the biometric data was measured, or how college statistics were tabulated that could potentially induce confounding factors in the analysis.

*Figure 2*

An interesting relationship exists between the age of the player upon entering the league and the percentage of those players by age group that remain the league. Based on Figure 3, for ages 19-23 and on average, we see that that the younger players are actually more likely to remain in the league longer than the older players.

*Figure 3*

Lastly, to assess correlation, we calculated the Pearson R Coefficients (Figure 4) and isolated the variables with the highest correlation, i.e. greater than or equal to 0.70 and less than or equal to -0.70.

*Figure 4*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pearson R Coefficients** | | | | | | | | | |
|  | **Height No Shoes** | **Height With Shoes** | **Wingspan** | **Standing Reach** | **Vertical No Step** | **Weight** | **Assists** | **Guards** | **Forwards** | |
| **Hght\_noshoes** | **1** | **0.996** | **0.845** | **0.914** | **0.787** | **0.75** | **-0.708** | **-0.778** | 0.388 | |
| **Hght\_wtshoes** | **0.996** | **1** | **0.844** | **0.915** | **0.782** | **0.756** | **-0.704** | **-0.779** | 0.38 | |
| **Wingspan** | **0.845** | **0.844** | **1** | **0.902** | **0.822** | **0.737** | -0.688 | **-0.711** | 0.366 | |
| **Standing\_reach** | **0.914** | **0.915** | **0.902** | **1** | **0.796** | **0.713** | -0.681 | **-0.724** | 0.359 | |
| **Vert\_nostep\_rch** | **0.787** | **0.782** | **0.822** | **0.796** | **1** | 0.593 | -0.688 | -0.664 | 0.393 | |
| **Weight** | **0.75** | **0.756** | **0.737** | **0.713** | 0.593 | **1** | -0.605 | **-0.72** | 0.374 | |
| **Assists** | **-0.708** | **-0.704** | -0.688 | -0.681 | -0.688 | -0.605 | **1** | 0.647 | -0.397 | |
| **Guards** | **-0.778** | **-0.779** | **-0.711** | **-0.724** | -0.664 | **-0.72** | 0.647 | **1** | **-0.738** | |
| **Forwards** | 0.388 | 0.38 | 0.366 | 0.359 | 0.393 | 0.374 | -0.397 | **-0.738** | **1** | |

5 Model Determination

5.1 Model Selection Criteria

Before we begin the data modelling, we need to determine the best binary classification model for answering our question of interest. How do we predict future success and non-success based off biometric, college statistics, draft order, and position data as described above? Since the answer to our question is a dichotomous non-numeric outcome, it stands to reason that we will focus on supervised classification algorithms as a roadmap for determining the appropriate model type.

With that in mind, we have chosen three distinct supervised classification models as part of our testing plan, which are (1) Logistic Regression, (2) Random Forest, and (3) Support Vector Machine. We will discuss the advantages of each below.

The Logistic Regression (LR) model performs well with features that exhibit linear relationships and with binary responses that can be mathematically separated through linear equations. In instances where the linearity assumption is violated, feature sets may benefit from various types of data transformations, such as a log or higher order transformation. The predictions may also be evaluated as probabilities or as odds ratios, depending on the preferred measurement of the analyst.

Conversely, a Support Vector Machine (SVM) model builds a function (or hyperplane) that aims at separating binary classes of data. The linear SVM algorithm is represented by points in a plot, where it attempts to divide the points into the highest possible separation, also referred to as the margin. While LR models are linear based, a SVM is capable of performing non-linear classification, by using a kernel trick such as a radial basis function.

Lastly, the Random Forest (RF) approach to classification offers considerable differences between both the LR and SVM models. The RF is a decision tree based model that does not require linear relationships in the data. Additionally, since the model is a combination of decision trees, it is suitable for our problem since binary “decisions” are made based on the thresholds computed in each respective branch in the tree.

MY PREVIOUS SPARSE COMMENTARY STANDS. THIS IS POORLY ORGANIZED AND WRITTEN. IT IS NOT MOTIVATED OR OTHERWISE TELEGRAPHED TO THE READER. WHAT’S THE BIG PICTURE? WHY ARE YOU DOING THIS? WHAT ARE YOU DOING? We will be testing various classification methods in an attempt to make predictions about the likelihood of NBA success. The goal is to maximize predictions from two different classification tasks. The first task aims to classify whether an NBA player will have success based on the following factors: 1) biometric data, 2) age prior to first year of NBA experience, and 3) draft order. For our second classification task, we will surmise that by incorporating college statistics and quality of college – the latter which will based on the final Ratings Percentage Index (RPI) ranking from the player’s final college season, that our predictive metrics will improve.

To determine the optimal model for our classification exercises, the quality of each model will be based on the following statistical measures:

* Average Precision score: interpreted as the number of true positives over the number of true positives plus the number of false positives.
* Average Recall score: interpreted as the number of true positives over the number of true positives plus the number of false negatives.
* Average F1 score: defined as the harmonic mean of precision and recall.

By assessing all scores as opposed to just one, we inspect the model’s ability (or lack of ability) to accurately classify success and to reduce the number of false positives and false negatives.

Additionally, one weakness of the accuracy statistic is that it ignores the negatives costs associated with misclassification. To factor this into the assessment of overall model quality, we will build a cost matrix that computes total cost that is predicated on penalties for misclassified predictions (i.e. false positive and false negatives) and credits for correct predictions (i.e. true positives and true negatives). For consistency, we will use the following cost methodology for all of models. A lower score = a better model.

Lastly, to evaluate competing models, we will use the McNemar's test to compare the number of false negatives produced from different algorithms on the same data set. The F1 scores by themselves do not determine if the difference is significant, so the McNemar score serves as additional support for determining the best model.

Prior to fitting our model and making predictions, we will initiate a training/testing data split of 80%/20%. The testing data serves as a truer test of how the model will perform post production, since it does not have the luxury of the ‘expected’ outcome data that is allocated to the training data. To reduce the variance introduced from performing training/testing splits, we will implement a 10-fold cross validation that removes the possibility of the split only having one set. Both measures ensure a properly validated and robust approach prior to model implementation.

3.3 Classification Task 1 - Logistic Regression Model

The logistic regression (LR) model assumes one or more explanatory variables that determine a dichotomous outcome. The LR model predicts the logit transformation of the probability of the question of interest, which may be expressed as:

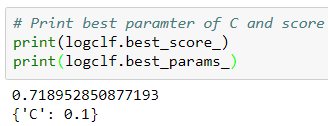
where is the probability of the presence of the particular response. A transformation of the logit may be interpreted as the logged odds, as expressed below:

|  |
| --- |
| Variable |
| Draft\_pick |
| ast |
| Vert\_nostep |
| Age\_first\_yr |
| Hght\_noshoes |
| Body\_fat |
| Vert\_nostep\_reach |
| ft\_pct |
| Wingspan |
| Hand\_length |
| Weight |
| Vert\_max |
| games |
| fnl\_coll\_rpi |
| Hand\_width |
| Hght\_wtshoes |
| Standing\_reach |
| Vert\_maxreach |
| pts\_ppg |
| fg2\_pct |
| fg3\_pct |
| Tms\_played\_for |
| trb |
| Draft\_yr |

For our first classification task, the variables IN TABLE will be used as explanatory measures WHAT IS AN “EXPLANATORY MEASURE”? for predicting the probability of success.

LR Model Results

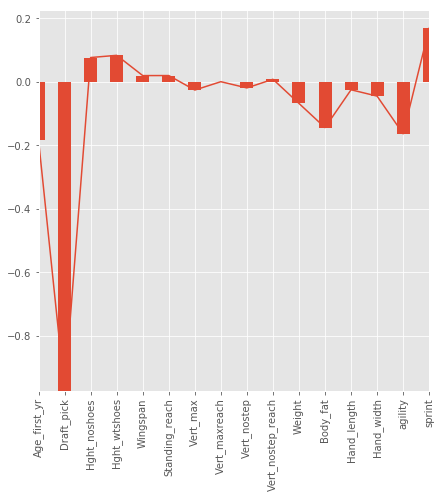
After implementing the custom grid search and cross validation, the optimum model was fit to a C parameter of 0.1 and had an accuracy score of 0.72, as seen below.



To get a better grasp on the variables that had an impact on the model predictions, we generated the coefficient weights from the model. It is clear that the order in which the player was drafted had the most profound effect on whether an NBA player would be successful or not. Other important factors included the age of player upon entering the league, agility and sprint times, and body fat percentage. From the biometric measures, we can infer that player ‘fitness’ is critical for projecting future success. Other secondary measures such as vertical leap, wingspan, hand size, had little to no impact on the model predictions. To determine our coefficients, we also combined college statistics in order to determine the weight of each variable. These were used for both logistic regression models. They are as follows:

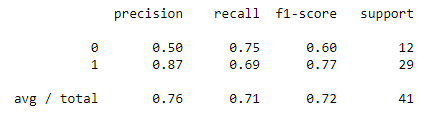
|  |  |
| --- | --- |
|  |  |
| Variable | Weight |
| Draft\_pick | -1.6545 |
| ast | -0.5572 |
| Vert\_nostep | -0.4756 |
| Age\_first\_yr | -0.4498 |
| Hght\_noshoes | -0.2800 |
| Body\_fat | -0.2437 |
| Vert\_nostep\_reach | -0.2360 |
| ft\_pct | -0.1754 |
| Wingspan | -0.1405 |
| Hand\_length | -0.0842 |
| Weight | -0.0489 |
| Vert\_max | -0.0014 |
| games | 0.0001 |
| fnl\_coll\_rpi | 0.0198 |
| Hand\_width | 0.0271 |
| Hght\_wtshoes | 0.0383 |
| Standing\_reach | 0.0729 |
| Vert\_maxreach | 0.0854 |
| pts\_ppg | 0.2492 |
| fg2\_pct | 0.3743 |
| fg3\_pct | 0.5089 |
| Tms\_played\_for | 0.8607 |
| trb | 1.1506 |
| Draft\_yr | 1.3877 |

The graph below plots the weights and shows the relative strengths of the weights:



**Logistic Regression based on Biometrics**

The classification report provides precision, recall, and F1 scores as well as the average scores for the binary responses. The model had a better recall score for unsuccessful players versus successful ones. Conversely, the model had a much better precision score for successful players as opposed to unsuccessful ones. Overall, the model was more adept at predicting success versus predicting no success, as evidenced by the higher average F1-score.



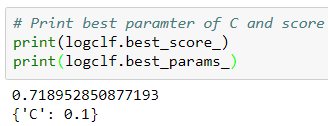
To assess the cost associated with misclassified players, we calculated the cost score associated with the model. This cost sc

Cost Score = (9)(-1) + (3)(10) + (9)(15) + (20)(-1) = 136

Overall, the LR model demonstrated fair predictions from the successful player population, but struggled with predicting non-success, as seen in the respective F1-scores of 0.77 and 0.60. Going forward, we will create Random Forest and Support Vector Machine models to determine which method is most appropriate for our current methodology. Additionally, in classification task 2, we will incorporate the players’ college statistics and quality of college to determine if the model gain additional ‘lift’ from more data.

**Logistic Regression for College Statistics.**

The same process was done for the logistic regression involving college statistics. After implementing the custom grid search and cross validation, the optimum model was fit to a C parameter of 2 and had an accuracy score of 0.75, as seen below.



0.748102894808178

{'C': 2}

The classification report was also done for the college statistics model and is shown below:

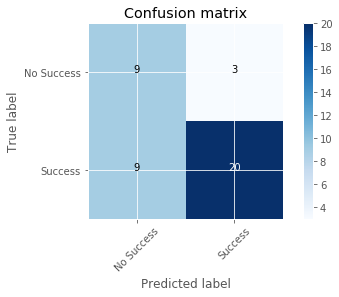
precision recall f1-score support

0 0.58 0.73 0.65 15

1 0.80 0.67 0.73 24

avg / total 0.71 0.69 0.70 39

In addition, we also created a confusion matrix, which summarizes the results of a classification test based on true positives, false negatives, true negatives, and false positives. Starting from the top left and going in a clockwise direction, our confusion matrix shown below indicates our results in the order listed above:



References

1. Berri, David J., et al. “The Short Supply of Tall People: Competitive Imbalance and the National Basketball Association.” *Journal of Economic Issues*, vol. 39, no. 4, 2005, pp. 1029–1041., doi:10.1080/00213624.2005.11506865.
2. Berri, David J., et al. “From college to the pros: predicting the NBA amateur player draft.” *Journal of Productivity Analysis*, vol. 35, no. 1, 2010, pp. 25–35., doi:10.1007/s11123-010-0187-x.
3. “NBA Draft Lottery: Facts, History, Probabilities & All-Time Results.” *NBA.com*, 18 May 2016, www.nba.com/news/draft-lottery-history/.
4. Neil\_Paine. “Projecting The Top 50 Players In The 2015 NBA Draft Class.” *FiveThirtyEight*, FiveThirtyEight, 21 Apr. 2017, fivethirtyeight.com/features/projecting-the-top-50-players-in-the-2015-nba-draft-class/.
5. Radzevick, Joseph R. “Does Transition Experience Improve Newcomer Performance? Evidence From the National Basketball Association.” *Small Group Research*, vol. 47, no. 2, 2016, pp. 207–235., doi:10.1177/1046496416636199.
6. release, Official, and NBA.com. “NBA salary cap set for 2017-18 season at $99.093 million.” *NBA.com*, NBA.com, 1 July 2017, www.nba.com/article/2017/07/01/nba-salary-cap-set-2017-18-season-99093-million.
7. Treme, Julianne, and Robert Burrus. “NCAA basketball: when does recruiting talent translate into wins for power conferences?” *Springer Science Business Media*, Journal of Economics and Finance, 21 May 2015, link.springer.com/article/10.1007%2Fs12197-015-9323-9.