Predicting Rookie Role Players in the NBA

Abhishek Dendukuri - ad5529; Saumya Didwania - sd4469; Neeraj Joshi - nsj244; Arya Tayebi - at3374

1. Business Understanding

Since the start of the Chicago Bulls dominating performances in the mid-90s, the NBA has seen a meteoric rise in popularity not only in the US but around the world. With greater popularity and an influx of cash, greater responsibility follows for teams' general managers to provide both entertaining and successful basketball. One option for general managers is to optimize their team's performance by recruiting well during the annual NBA Draft. Historically, some players who are drafted early don't achieve the future success that teams hope they do, and players drafted later may outperform higher picks and go on to become hall of famers, NBA champions, and gold medalists. A recent example of this is the Lakers two draft picks in 2017. The Lakers selected college superstar Lonzo Ball with the 2nd overall pick and Kyle Kuzma with the 27th overall pick. Three years later, Kuzma has played a significant role in the Lakers NBA championship run while Lonzo has been traded to a non-contending team. Because Lonzo played entertaining college basketball and was on a good team, the Lakers felt obligated to take him with the 2nd pick. Yet, the value and continued success of Kyle Kuzma has been worth much more. Year after year these "anomalies" happen where the best players of the draft are not the highest picks but are distributed throughout the draft, underlying the magnitude of this opportunity. In the Lakers example, Lonzo Ball's first year salary costed the Lakers almost $5 million more than Kyle Kuzma's. In hindsight, the Lakers could have traded their 2nd pick for a seasoned veteran, potentially winning the NBA Championship a year earlier.

NBA teams have a set way of researching players primarily by focusing on stats, watching film, and personal interviews to determine if their skills would translate to the NBA. While there are many areas of analysis on "who the best players of the draft are", very few studies have looked into predicting a player's success 3-4 years after entering the league. Our goal of this paper is to lay out what continued success from NBA players looks like from key drivers in high school, college, and international data. We will look to not

only understand this but also extend it to what separates a star player who falters out of the league in a few years versus a consistent player who adds value to the team year after year.

Previous research in this area, such as that carried out by Krebs and Scheide[1], did not focus on success at the 3-year mark. While they used similar metrics and techniques when evaluating model performance, this paper does not build on their work but considers tackling a similar problem more specific to 3rd year success, the first year a team can exercise an option over the rookie's contract. Other similar papers and teams use value-based team metrics such as *Win Shares* and *Value-Over-Replacement-Player (VORP)* where the main goal is to compare one player to another. Teams will project these metrics onto players based on how they contributed to their teams in high school, college, or internationally and use that to see how well that might translate to the NBA. Additionally, teams will enter the draft knowing what positions they need to strengthen, and may bias their draft pick based on the "best available player" at that position. Our model hopes to use individual player statistics to show if a draft-eligible player will become an important role player for an NBA franchise.

2. Data Understanding

In order to properly assess whether or not a player will achieve future success around the date of the NBA Draft, we're limited to statistics on players in the year prior to the Draft. Between 1995 and 2005, a player could enter the NBA draft directly from high school, after playing in college for a number of years, or after having played professional basketball in a league overseas. However, data on high school basketball games is not robust or consistent enough for us to consider players that enter the NBA directly after graduating 12th grade. In 2005, a collective bargaining agreement (CBA) between the NBA and NBA Players Association required that players be at least 19 years old during the calendar year of the draft, effectively mandating that players spend at least one year in college or in another professional league[2]. By taking this into account, we constrained our data to Drafts from 2006 – 2017, with each Draft including 60 players, totaling 720 candidates for our dataset. Due to a significant portion of players only completing one

[1] Generating Relative Pick Value in the NBA Draft and Predicting Success from College Basketball (wpi.edu)
[2] N.B.A. Draft Will Close Book on High School Stars - The New York Times (nytimes.com)

year of basketball in either college or another professional league, we decided to look at a player's most recent year of statistics prior to the year they entered the Draft. This included adding drafted players that never made it to the NBA, making sure survivorship bias was taken into account as our data did not solely consist of players with NBA statistics in their third year. Certain Draft prospects did not have data in the year prior to their Draft due to injury or academic penalties, bringing our total down to 655 instances.

Due to the timing of a player's first significant contract, we based our outcome variables on a player's 3rd season, prior to any contract decisions. Training data came from two main sources: SportsReference for NCAA athletes and BasketballReference for International athletes. Basketball-Reference was also used to extract our outcome variables. Data was obtained from both sources by web scraping individual player pages for information at a per-game granularity. After initial exploration, it was apparent that the richness of information differed between NCAA Player data and International Player data. NCAA data, coming from a single league, included advanced statistics on each player which could have been indicative of future success in the NBA. Unfortunately, International data didn't have the same granularity of information readily available. With our limited dataset and International players accounting for 13% of our instances, we decided to only include features that were available for both types of prospects. Additionally, with International players coming from multiple leagues, it was difficult to reconcile all of their data in a standardized manner. For example, a player could have statistics from when they played in the annual EuroCup tournament but also for their primary affiliated team. We chose to look at the league or organized event that offered us with the largest sample of games within the same season of play.

Our features are categorized into player demographics and per-game statistics, with a total of 40 variables to predict the success of a draft prospect in their 3rd NBA season. One of the notable omissions from our data model are team statistics. Our model aims to predict the success of a draft entrant in the NBA, regardless of the team that player is drafted to. Due to the high perceived value at the top of the draft, teams engage in "tanking" throughout the season to receive a high pick. These teams generally have severe needs, which provides almost a sure path to the starting lineup for a draftee. In a metric such as Usage Percentage,

a draft player's statistics may be inflated if they are on a bad team simply because they may be the "only option" for a team to score. The metrics we used tend towards individual contributions over the player's value to a team.

It is important to note that our data suffers from a slight selection bias towards players who are regarded as having higher levels of ability, as we only looked at players who entered the NBA through the annual NBA draft. Over 25% of players in the NBA are undrafted and can achieve levels of success comparable to drafted players[3]. However, for our particular business case, it is important to focus on players that provide a team the highest success, given the options available. Given our model's success (see *Modeling and Evaluation)*, we may be able to extend it to undrafted players to understand whether or not they realize success in the NBA three seasons out.

3. Data Preparation & Exploratory Analysis

Success in the NBA can be measured in a variety of ways. We decided upon three initial criteria for success in a player's 3rd NBA season. While we evaluated all of our models on all outcome variables, we've decided to focus our findings on *roleplayer_nba* as it was the most indicative of value for a team. Developing models for each of the outcome variables will prove fruitful for different stages of the draft. See *Deployment* for more on strategies for using these models in different scenarios.

Outcome Variable	Description	Occurrence
still_nba	Prospect plays at least 1 game in 3rd NBA season.	True – **71%** False – **29%**
start_nba	Prospect starts in >50% of games in 3rd NBA season.	True – **21%** False – **79%**
roleplayer_nba	Prospect plays > 20 minutes per game in 3rd NBA season.	True – **36%** False – **64%**

Figure 1.1: The three criteria we selected to measure a player's success in the NBA.

Taking a more focused approach to feature importance, we discovered that certain variables showed greater importance than others. One of those variables, *Years_Experience_Prior*, showed significant

[3] There's Never Been a Better Time to Go Undrafted - The Ringer

differences between players who spent varying amounts of time playing in previous organizations prior to entering the draft (*Figure 1.2*). This relationship seemed counterintuitive in that the less experience prospects have in a more developed organization, the more likely they are to have success in the NBA. This does, however, make sense once we think about the nature of exceptional players entering the NBA as soon as possible to maximize their financial potential. This relationship does not hold true for International players, highlighting the need to create a boolean variable differentiating international and domestic players.

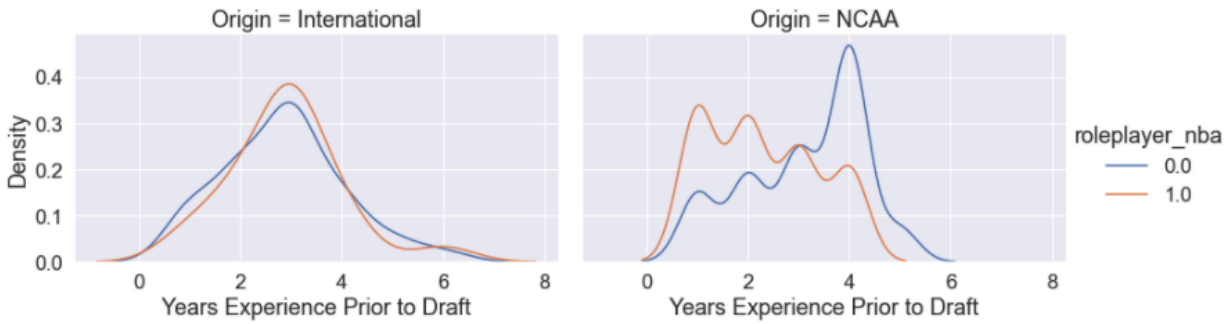


Figure 1.2: Density plots showing the difference in occurrence of the outcome variable, roleplayer_nba, with the number of years that a player spent playing in a league following high school.

Upon further research into recruiting, we found a ranking online, called Recruiting Services Consensus Index (RSCI), which consolidated the rankings from several experts belonging to reputable organizations like ESPN, MaxPreps, Prepstars, and more to obtain an objective top-100 ranking for high school players in the U.S. We found that these rankings were highly correlated with players who end up entering the draft and realizing success in the NBA. Due to the nature of only 100 U.S. high school players being selected, we concluded that the best way to code this information into the model would be to define a boolean variable that indicated whether or not that player was included in the RSCI rankings for the year they graduated high school. In addition to constructing these new features, data was also standardized prior to modeling in order to speed up optimization calculations for our linear based and tree-based models. Finally, to deal with missing values around shooting percentages, we used a simple median imputation within a particular playing position to fill in missing values while also accounting for outliers. Filling in the missing percentages with zeros would incorrectly penalize a player for a type of shot that they never took.

In order to reduce the number of features, we employed a k-means clustering algorithm to group the categorical features that described a players' physical attributes (such as height, weight, position) and playing experience (such as International Boolean, Domestic Conference). In addition to the differences between international players and college players, we found intra-college differences between the leagues a prospective draft pick played in. The data showed between 5-8 clusters would be appropriate, with each cluster having at least 10% of our total dataset. While most players in basketball generally are classified into one of five positions, our data broke most of the players down into only three buckets: Guards, Forwards, and Centers. In order to see if a player's physical attributes would be an important feature in our dataset, we created clusters to help split the players into the natural buckets. Similar to our playing experience cluster, the data showed between 5-7 clusters would be appropriate. We used five clusters for both of the groups of features[4].

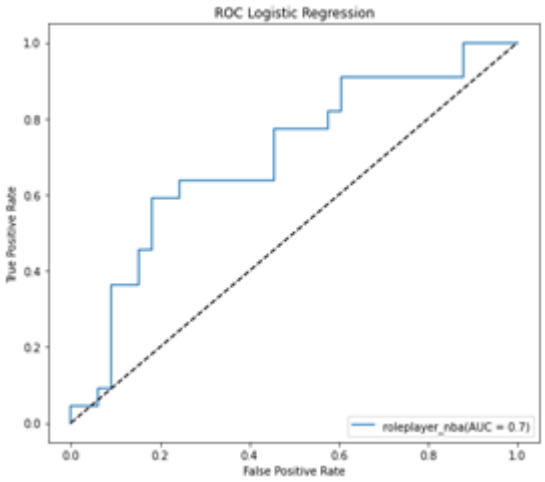
4. Modeling & Evaluation

Our initial hypothesis was to use a Logistic Regression model, as it is extremely well equipped to work with non-categorical data (which basketball stats tend to be) and also uses specified features to make a prediction, limiting bias. Additionally, our data set was small enough to create an effective baseline without tuning hyperparameters. We split our data by draft class – training the model on the draft classes from 2006 to 2016 and testing on the 2017 draft class. As mentioned before, we focused on *roleplayer_nba* as our outcome variable and kept the rest of the features as inputs.

Since a standard accuracy measure would not be able to properly distinguish any inherent skewed sample distribution, we decided to evaluate the Logistic Regression model in the form of an ROC curve and report the corresponding AUC (*Figure 1.3*). While we achieved a decent baseline for our model, we decided to examine the number of false positives and false negatives generated by the algorithm by printing the confusion matrix and classification report (*Figure 1.3*). A false positive represents a player predicted to be a role player when in fact he was not, signifying a contract overpay. A false negative is a player that was

[4] See Appendix A2 and A3 for cluster selection details

predicted to not be a significant contributor to the team, but was, signifying missing on a potentially useful contributor. For some general managers, saving money and not taking risks on players have higher priority and therefore false positives would be more harmful than false negatives. For others, especially those who work for teams in contention to win a championship, finding critical role players is their ethos and they are willing to spend the money for them, potentially treating false negatives as more harmful than false positives. For this reason, we decided to utilize two other evaluation metrics: precision and recall.



roleplayer_nba

	precision	recall	f1-score	support
0.0	0.75	0.82	0.78	33
1.0	0.68	0.59	0.63	22
accuracy			0.73	55
macro avg	0.72	0.70	0.71	55
weighted avg	0.72	0.73	0.72	55

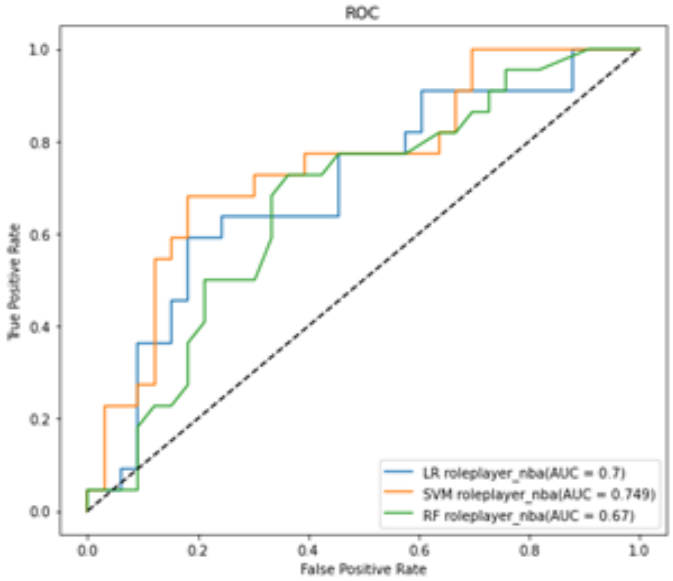
Actual \ Predicted	0	1
0	27	6
1	9	13

Figure 1.3: Evaluation plots showing the ROC Curve, Classification Report, and Confusion Matrix for the initial Logistic Regression Model for roleplayer_nba

Given the differing objectives each team may have, it's important to understand a model's strengths and weaknesses, and also be able to minimize false positives or false negatives depending on the situation at hand. Our precision and recall values of 0.68 and 0.59 for the positive outcome, indicate that our initial model minimizes the number of false negatives more than the number of false positives. Looking at the precision and recall values for our negative outcome, 0.75 and 0.82 respectively, indicates that our model is more capable of predicting players who don't go on to become role players.

As seen in *Figure 1.3*, our data can be prone to false positives and false negatives which complicates our overall feature space. We felt that the Logistic Regression model was an inflexible model since the foundation is based on a logit function. That means, no matter how intricate the features space is, the general trend will be an S curve. As a way to work around this issue, we implemented two more models: Support Vector Machine and Random Forest. SVMs have more flexibility than a Logistic Regression and Random

Forest which, although is more commonly used for multi-class classification, is still useful in tackling high dimensionality problems.



Logistic Regression

	precision	recall	f1-score	support
0.0	0.74	0.88	0.81	33
1.0	0.75	0.55	0.63	22
accuracy			0.75	55
macro avg	0.75	0.71	0.72	55
weighted avg	0.75	0.75	0.74	55

Actual \ Predicted	0	1
0	29	4
1	10	12

SVM

roleplayer_nba

	precision	recall	f1-score	support
0.0	0.75	0.82	0.78	33
1.0	0.68	0.59	0.63	22
accuracy			0.73	55
macro avg	0.72	0.70	0.71	55
weighted avg	0.72	0.73	0.72	55

Actual \ Predicted	0	1
0	27	6
1	9	13

Random Forest

	precision	recall	f1-score	support
0.0	0.65	0.79	0.71	33
1.0	0.53	0.36	0.43	22
accuracy			0.62	55
macro avg	0.59	0.58	0.57	55
weighted avg	0.60	0.62	0.60	55

Actual \ Predicted	0	1
0	30	3
1	15	7

Figure 1.4: Evaluation plots showing the ROC Curve, Classification Report, and Confusion Matrix for the Logistic Regression, SVM, and Random Forest Models for roleplayer_nba

While the AUCs suggested that SVM performed slightly better than Logistic Regression, the total number of false positives and false negatives remained similar (*Figure 1.4*). From here we decided to see if feature selection and hyperparameter tuning could help improve the precision and recall scores for both models.

First, we tried to improve our baseline Random Forest and Logistic Regression models by implementing our own forward stepwise selection. Both of these models ended up with 6 features, fewer than the total number of features we originally had, making our model more interpretable. The most important features identified by forward selection were *'Num_College_Years', '3PA', 'Personal Fouls', '2P%', 'Weight', 'Position_Forward'* for Logistic Regression and *'InternationalLeague_Boolean', 'BLK', 'Num_College_Years', 'Turnovers', 'FTA', 'Position_Guard'* for Random Forest. For forward stepwise selection, SVM only considered *field goal percentage*, classifying all results as the negative class. Since

the plane is one dimensional with the boundary being a point, the prediction error was large. To fix this, we also implemented backward stepwise selection, which reduced the SVM model to 26 total features.

After feature selection, we moved on to hyperparameter tuning. We implemented a GridSearch cross validation algorithm over a normal K-fold cross validation because our dataset is small enough to not encounter long GridSearch runtimes. The parameters we focused on for Logistic Regression were weight, regularization (C), and penalty. For SVM: C, gamma, and kernel type. For Random Forest: max depth, min leaves, min splits, num estimators. To account for the size of the testing set, we implemented 10 folds within the GridSearch.

Logistic Regression

	precision	recall	f1-score	support
0.0	0.67	0.85	0.75	33
1.0	0.62	0.36	0.46	22
accuracy			0.65	55
macro avg	0.64	0.61	0.60	55
weighted avg	0.65	0.65	0.63	55

Actual \ Predicted	0	1
0	28	5
1	14	8

SVM

	precision	recall	f1-score	support
0.0	0.75	0.82	0.78	33
1.0	0.68	0.59	0.63	22
accuracy			0.73	55
macro avg	0.72	0.70	0.71	55
weighted avg	0.72	0.73	0.72	55

Actual \ Predicted	0	1
0	27	6
1	9	13

Random Forest

	precision	recall	f1-score	support
0.0	0.72	0.94	0.82	33
1.0	0.83	0.45	0.59	22
accuracy			0.75	55
macro avg	0.78	0.70	0.70	55
weighted avg	0.77	0.75	0.72	55

Actual \ Predicted	0	1
0	32	1
1	14	8

Figure 1.5: Evaluation plots showing the Classification Report and Confusion Matrix for the Logistic Regression, SVM, and Random Forest Models for roleplayer_nba

From our final models, we can see that while SVM had the same performance in terms of precision and recall as the original Logistic Regression model, we used fewer features – making it a more scalable and interpretable model. To further evaluate our model's performance, we compared a players' draft pick number and our model's probability estimate of them being a role-player in their 3rd NBA season for our test set. We treated a player's draft pick as a form of validation for our model, since it's a measure of how valuable NBA organizations deem a player being. *Figure 1.6* shows the probability our SVM model outputs

for a Draft prospect becoming a role player in their 3rd season against the position that player was drafted. As we'd expect, our model is fairly confident in predicting the positive class for players that are drafted earlier and accordingly, is not as confident about players who are selected in the later parts of the draft. Where we'll see the benefits of our model against traditional draft rankings is at the peripheries of this linear trend. If our model can identify skillful players that are traditionally drafted in the second round or unsuccessful players that are given unwarranted value earlier in the draft, it offers then value beyond the traditional analytics done for Draft prospects.

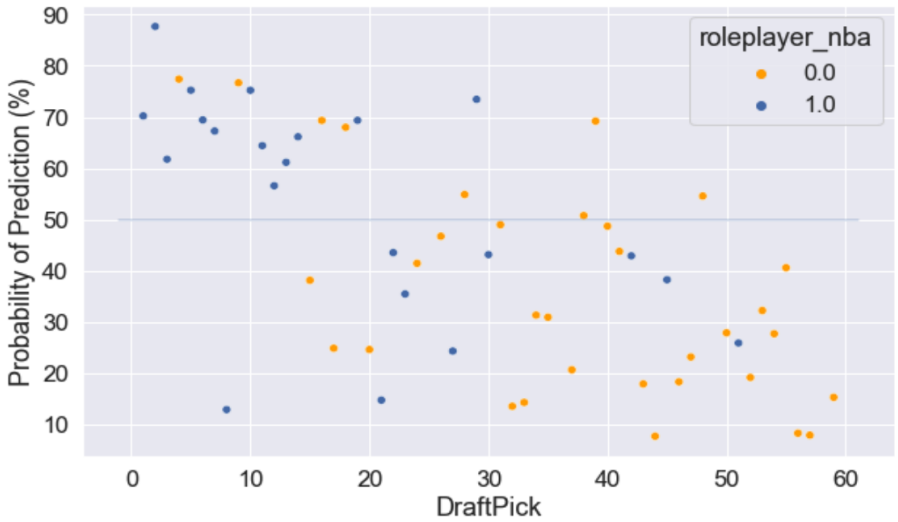


Figure 1.6: The relationship between the probability estimate for our SVM classifier predicting roleplayer_nba and the draft pick number for each of the prospects in the 2017-2018 NBA Draft.

5. Deployment & Future-Proofing

While this model can be applied in multiple ways, the primary use case would be to evaluate potential draft picks. If a team has two draft picks, we'd assume a team would first want to draft a starter that fills a key deficiency on the team. For their other pick, usually in the second round, we'd want a team to select a player that has the highest chance of becoming a starter or a role player. Players drafted in the second round are one of the lowest paid groups of players, and as such, a team can find great value in those picks. As mentioned in *Data Understanding*, there has been an increase in undrafted players over the last few years. While we only looked at players that were drafted in the NBA since 2006, we can expand this model to include all college players that are draft-eligible. Teams can use this model to find undrafted

players that we predict to be significant role players, allowing them to reallocate budgets towards tried-and-tested free agents or super-star players.

Some of the risks associated with this model include changing NBA trends and access to future player data. One of key trends in the recent NBA has been the "Moreyball" revolution, where players replaced longer two-point shots with higher expected value shots.[5] Teams started placing more value on players that can shoot three-pointers[6] and teams such as the Golden State Warriors popularized "small-ball", in which teams used five smaller, quicker players rather than employing a Center. To solve for these potential future impacts, we can penalize certain traits that are less important in more recent seasons. For example, we can add a penalization parameter for players that have a low three-point percentage or players that play the Center position.

Another potential risk in future iterations would be the changing three-point line in college basketball. In the 2019-2020 NCAA season, the three-point line moved back from 20 feet, 9 inches to 22 feet, 1.75 inches – the same dimensions as the international three-point line.[7] The NBA three-point line is almost a foot-and-a-half longer than the international three-point line except for in the corners, where it is approximately the same as the international one. Ideally, we'd like to incorporate shot location in our college dataset to predict the expected NBA points that a college player would score, which might change the importance of three-pointers in our dataset.

As noted earlier, the NBA implemented a rule in 2005 that prevented high school players from entering the NBA directly. If this rule were to be reversed, our model would have to be updated to use high school data instead of just college or international. In our data discovery, high school metrics were much less accessible and based on our current limitations, we would have to impute values for many players. If we impute a median value, as we have been doing, we could find ourselves with a data source that looks like a Cauchy distribution with a tall peak. The players at the tails of the distribution would be the only ones

[5] Nearly Every Team Is Playing Like The Rockets. And That's Hurting The Rockets. | FiveThirtyEight
[6] See Appendix A.4 for Chart showing the same trend in our dataset
[7] NCAA votes to move back men's 3-point line again, among other rule changes | The Seattle Times

our model would be able to provide relevant predictions for. The players in the middle would become a 50-50 guess if there were only a few metrics to set them apart. Additionally, if high school metrics become more important in our model, we might have some ethical concerns with data collection. Since there is currently no standardized data-collection process at the high-school level, our data might be biased towards schools well-funded enough to collect high-school statistics. Schools which might not have the budget for someone to collect stats may be underrepresented in future iterations.

Before this model would be put into production, teams may want to improve on certain features. As noted earlier, NCAA data has more advanced metrics easily accessible and our model could benefit from splitting the two sources of talent. Additionally, our model only attempted to predict if a player would be a significant role-player. We could change our outcome variable to advanced metrics such as *Expected Wins Added* to show the exact impact of a rookie. For this model, we elected a small dataset that only included players that had already been drafted. A further improvement to our model could be to include all draft-eligible players across the NCAA and Internationally to find real "hidden gems." While we don't envision teams using this model to replace their current player evaluation techniques, this model would be a good supplement to help identify overrated and underrated talent.

Appendix

- References

- Beck, Howard. “N.B.A. Draft Will Close Book on High School Stars.” *The New York Times*, The New York Times, 28 June 2005, www.nytimes.com/2005/06/28/sports/basketball/nba-draft-will-close-book-on-high-school-stars.html.
- Dubin, Jared. “Nearly Every Team Is Playing Like The Rockets. And That's Hurting The Rockets.” *FiveThirtyEight*, FiveThirtyEight, 20 Dec. 2018, fivethirtyeight.com/features/nearly-every-team-is-playing-like-the-rockets-and-thats-hurting-the-rockets/.
- Kram, Zach. “There's Never Been a Better Time to Go Undrafted.” *The Ringer*, The Ringer, 16 Nov. 2020, www.theringer.com/2020/11/16/21566647/nba-undrafted-alex-caruso-duncan-robinson.
- Scheide, Jake Connot, and Michael James Krebs. 2019, *Generating Relative Pick Value in the NBA Draft and Predicting Success from College Basketball*.
- The Associated Press. “NCAA Votes to Move Back Men's 3-Point Line Again, among Other Rule Changes.” *The Seattle Times*, The Seattle Times Company, 5 June 2019, www.seattletimes.com/sports/college/ncaa-moving-3-point-line-back-for-first-time-in-a-decade/.

- Contributions

- All: Initial data collection, exploratory data analysis, model tuning, feature selection, write-up
- Abhishek Dendukuri: Modeling, Hyperparameter Tuning
- Saumya Didwania: Initial Modeling, Clustering
- Neeraj Joshi: Data Scraping and Cleaning, Feature Engineering for Boolean Variables
- Arya Tayebi: Initial Data Scraping, Trends, Forward/Backward Selection Implementation

- Additional Figures and Tables

Figure A1: Violin plot showing distributions of Years Experience Prior to Draft colored by our outcome variable, roleplayer_nba, across various years of the NBA Draft.

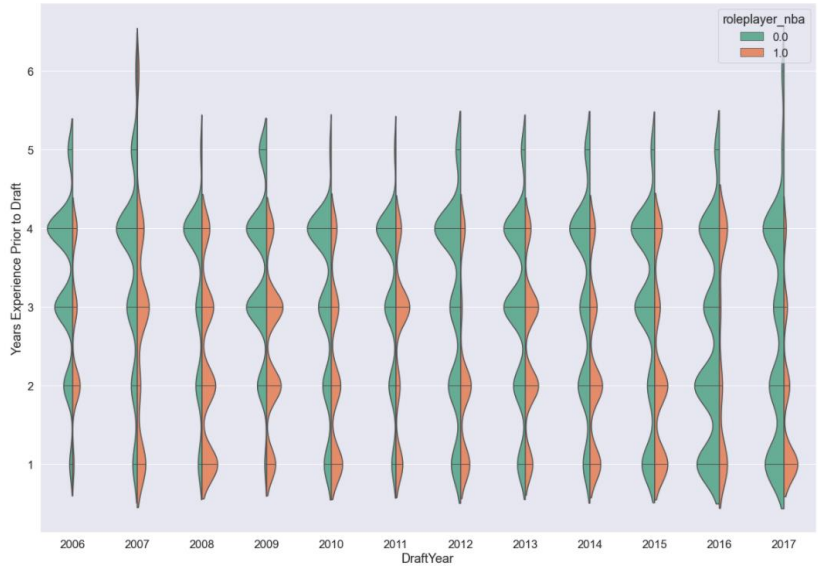


Figure A2: K-means clustering plot to show optimal cluster numbers

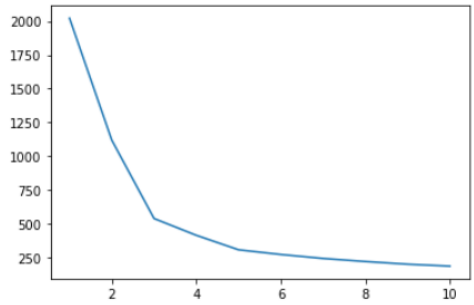


Figure A3: Dendrogram showing distance metrics for cluster size selection

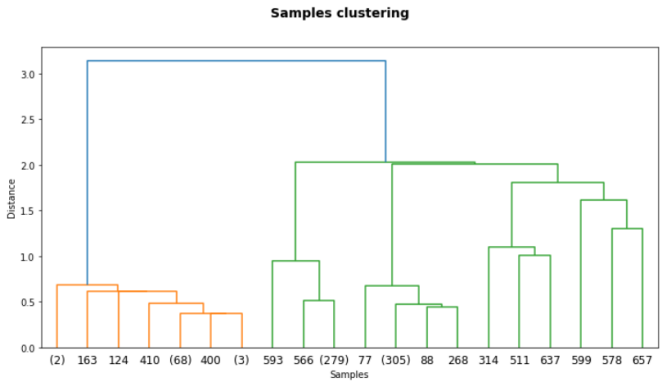


Figure A4: Trends in 3-point shooting in college