Generating Relative Pick Value in the NBA Draft and Predicting Success from College Basketball

A Major Qualifying Project Report:
submitted to the faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the
degree of Bachelor of Science

by

____
Michael Krebs

____
Jake Scheide

Date:

Approved: ________________________
Professor Craig Wills, Major Advisor
MQP-CEW-1901

Abstract

This project analyzes existing basketball player performance metrics, and generates new metrics providing context behind player statistics. Using these metrics, we create a chart quantifying the value of each pick in the NBA Draft. Finally, we create machine learning models that predict the likelihood of NBA success for NCAA student-athletes.

Acknowledgements

We would like to thank Professor Wills for his encouragement of us to apply our technical backgrounds to our shared love for basketball.

We would also like to thank the staff at basketball-reference.com, as their extensive database of college and NBA players made this project possible.

We would finally like to thank Jimmy Johnson, Kevin Pelton, Dr. Aaron Barzilai, and the numerous other analysts who's work informed our own, providing guidance and comparisons.

We hope our work will make a meaningful contribution to the growing body of literature involving sports data mining.

Table of Contents

Table of Figures

Executive Summary

This project's goals are threefold. First, we analyze existing basketball player performance metrics, and use these insights to create new metrics that provide a better comparison for players in the same season. Secondly, we generate a chart that quantifies the value of each pick in the NBA Draft. Finally, we create machine learning models which predict if NCAA Division I student-athletes will accomplish various levels of success in the NBA.

We used Player Efficiency Rating, Value Over Replacement Player, Win Shares and Fantasy Points as our four established metrics. These metrics represent a spectrum of mechanisms that front-offices, coaches, and fans use to evaluate and compare players. Often, these metrics tell different stories about the talent of a player, and can be skewed by injury, players who take a bench role later in their careers, or purely by nature of playing on a bad team. By examining the factors that normalized these metrics, we constructed three additional player performance metrics, with the goal of providing better insight into a comparison between two players in the same season. These metrics were Cumulative Individual Accolades, Basic Percentile and Advanced Percentile.

Using these metrics, we grouped players based on their selection in the NBA Draft, and created visualizations showing the different 'talent curves'. By clustering groups of picks together, we created equations which smoothly estimated the value of each pick. We then collated draft pick only trades made in the NBA since 2005 and settled on a best curve which accurately mapped them. From this, we compared our talent curve for the NBA to both NBA and NFL models, where these charts are actively used by teams for guidance in draft-pick trades.

Finally, we used machine learning to construct linear regression models that classify NCAA DI players based on various success criteria for the NBA. The success criteria we were particularly interested in were being drafted by an NBA team, drafted in the lottery / first round, and playing in an NBA game. These models considered not only the basic and advanced statistics of the players, but also the school they went to, height and weight. These models were extremely good at identifying talented prospects, and many misclassified players were found to have extenuating circumstances.

Overall, this project provides significant value to the front offices of NBA teams who are attempting to maneuver around the uncertainty associated with the NBA Draft. Selecting the right player is extremely important for a team's long-term success, even with lower picks in the draft. By understanding the true value of the team's draft position, and utilizing models such as our own, teams can make more informed draft decisions and extract the maximum value from their picks.

1. Introduction

Basketball is exploding both domestically and abroad, with the most recent National Basketball Association (NBA) season posting record attendance, TV and online viewership numbers (Adgate, 2018). Players now come from 42 countries, with all 30 franchises having at least one non-American player. The league is expanding their outreach into emerging markets such as China, India and Africa, with 300 million people in China playing basketball (Saiidi, 2018). This explosive growth has skyrocketed median team valuations, from $555 million in 2014 to over $1.5bn in 2018 (Routley, 2019).

As the NBA has grown, so has the potential lucrativeness of constructing a championship-winning roster. The Golden State Warriors, winners of three of the last four NBA Championships, find themselves paying $90 million in ‘luxury tax’, an economic penalty on teams which exceed the salary cap (Ramey, 2018). If they maintain their current roster, they will pay $221 million in luxury taxes during the 2020-21 season, more than the actual payroll of $178 million. The Warriors show just how valuable winning in the NBA is, even when paying such high taxes.

With this increased pressure to succeed (and therefore profit), teams must utilize every resource at their disposal to ensure they are accurately evaluating players both at the professional and collegiate level, the latter of which is the primary supplier of young NBA talent. The NBA Draft is held at the end of every season, where each team is awarded two selections in the sixty-pick event. Picks 15-60 are assigned in reverse order of record (where the best record team gets the 30th and 60th picks), and a lottery decides the recipients of the first fourteen picks, with probabilities proportional to standings. Teams are free to trade their rights to a draft pick prior, during, and after the draft lottery, as they try to maneuver up the draft board to obtain the best young talents.

Some teams looking to contend for championships may trade all their draft picks away for veteran contributors, as the Brooklyn Nets did in 2014. They traded three first round picks, as well as the right to swap first round picks (in four consecutive years), to the Boston Celtics for Kevin Garnett, Paul Pierce, and Jason Terry – three championship winning players who declined rapidly following Brooklyn’s acquisition (Greenberg, 2017). The Celtics benefitted even more from the players’ declines, as the struggling Brooklyn ended up receiving the third, first, and eighth selections in the draft- only the rights to the picks belonged to the Celtics.

This project’s goals are threefold. First, we analyze existing basketball player performance metrics, and use these insights to create new metrics that provide a better comparison for players in the same season. Secondly, we generate a chart that quantifies the value of each pick in the NBA Draft. Finally, we create machine learning models that predict if NCAA Division I student-athletes will be drafted or play in the NBA.

This project is timely, relevant, and important to NBA teams which seek to improve their teams through the draft, or trades. By analyzing player performance metrics, teams can contextualize the numbers they often are presented with by their analytics departments when debating a prospective trade. Additionally, analytics professionals can supplement the metrics they currently use with the ones we created, to generate more informed insights. When proposing or deliberating on trades involving draft picks, teams can use our draft pick value chart to ensure they are fairly compensated for the outgoing picks. Finally, front offices can verify their scouts' opinions on a collegiate player using the machine learning models we created to ensure they are selecting players who will be successful in the NBA.

In the remainder of this report, we first break down existing player performance metrics to better understand the mechanisms used by NBA teams when performing trades and contract negotiations. Using this understanding, we design three new player performance metrics that provide a new approach to evaluating talent. By summating the metric values for a set of NBA players, we then generate charts which approximate the relative value of each selection in the NBA Draft. Using draft-pick only trades, we calculate the error of each relative value curve to finally settle on one equation which explains the value of NBA draft picks. From our literature review, we compare our value chart to other NBA charts, as well as numerous NFL value charts, to compare the talent drop-off. Finally, using machine learning, we create models that predict if NCAA DI basketball players will be drafted and/or play in the NBA. The models use statistical data scraped from online sources, as well as the college the player attended, their height, and their weight.

2. Background

2.1 Existing Metrics in the NBA

Although many casual sports fans attribute the numbers revolution in sport to *Moneyball*, statistics and data were driving decision making in sport from as early as the 1920's, with baseball initially pioneering the movement (Schwarz, 2004). Baseball is largely viewed as the easiest game to quantify, as models can describe progress to scoring a run objectively with players moving along the bases. Additionally, each pitch is an independent event, further allowing itself to be analyzed using basic mathematics.

Basketball, on the other hand, is a free flowing, five on five game where missing an open layup after a well-run play counts for the same on the score sheet as a highly contested long-range shot. The complexity of basketball makes it a lot tougher to generate numbers which accurately reflect the talent level of a player or team. Additionally, Dean Oliver posits, the lack of statistics readily counted about defense makes basketball analytics largely skewed towards offensively-minded players (Oliver, 2004). Oliver invented the 'Four Factors' most critical to team success in basketball, namely shooting, rebounding, turnovers, and free throws. Each of the Four Factors are weighted differently and measured using advanced metrics. His book introducing these metrics, *Basketball on Paper*, is widely regarded as the bible of basketball analytics.

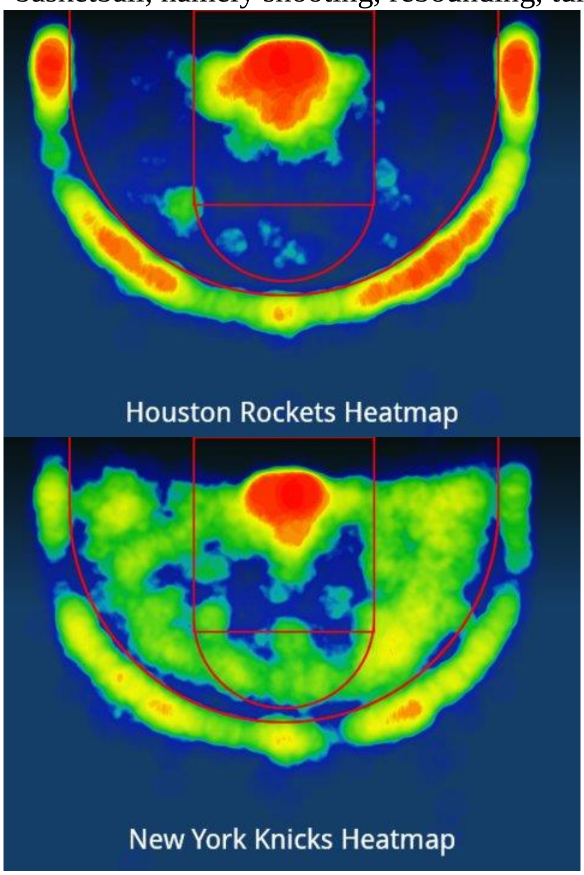


Figure 1: Houston Rockets & New York Knicks Heatmaps

Fast forward 15 years from the book's publication date, and data has truly revolutionized basketball. Teams have discovered the value of the three-point shot, and offenses and teams are now constructed to find threes and layups (Shot Search, n.d.). It's no coincidence that the teams investing the most in analytics, such as the Houston Rockets, are finding the most success. Figure 1 shows the large disparity in shot selection between the Rockets and the New York Knicks – a team languishing at the bottom of the NBA standings.

An analysis of basketball metrics is not something novel, but past papers arbitrarily pick statistics to incorporate into their analysis (Mertz, et al., 2016). For example, including points, rebounds and assists in addition to Win Shares per 48 minutes double counts the basic points, rebounds and assists statistic. Any ranking of players will require careful consideration of the basic statistics that go into the metrics used, as well as any possible normalizing factors used, such as minutes played, team wins, or pace of play.

2.2 Assessing Draft Value in Sports

2.2.1 NFL

One of the project's goals is identifying the value of draft picks in the NBA. In the NFL, there exists a widely known draft value table constructed by former Dallas Cowboys head coach Jimmy Johnson (Johnson, 2019). This draft table was designed to assess what a fair trade would be when trading draft picks. The work done by Barney et. al showcased that draft pick trades did in fact follow closely the values assigned in this draft value table. Indicating either the teams used the draft table to decide if the trade was fair or the table accurately showcases relative value for draft picks. In either case the most important aspect in determining if a draft table is effective is if trades that are made reflect relative values given in the table. Figure 2 displays the first 60 picks and their value from the Jimmy Johnson draft table.

Overall	Value	Normalized	Overall	Value	Normalized
1	3000	100.00	**31**	600	20.00
2	2600	86.67	**32**	590	19.67
3	2200	73.33	**33**	580	19.33
4	1800	60.00	**34**	560	18.67
5	1400	46.67	**35**	550	18.33
6	1600	53.33	**36**	540	18.00
7	1500	50.00	**37**	530	17.67
8	1400	46.67	**38**	520	17.33
9	1350	45.00	**39**	510	17.00
10	1300	43.33	**40**	500	16.67
11	1250	41.67	**41**	490	16.33
12	1200	40.00	**42**	480	16.00
13	1150	38.33	**43**	470	15.67
14	1100	36.67	**44**	460	15.33
15	1050	35.00	**45**	450	15.00
16	1000	33.33	**46**	440	14.67
17	950	31.67	**47**	430	14.33
18	900	30.00	**48**	420	14.00
19	875	29.17	**49**	410	13.67
20	850	28.33	**50**	400	13.33
21	800	26.67	**51**	390	13.00
22	780	26.00	**52**	380	12.67
23	760	25.33	**53**	370	12.33
24	740	24.67	**54**	360	12.00
25	720	24.00	**55**	350	11.67
26	700	23.33	**56**	340	11.33
27	680	22.67	**57**	330	11.00
28	660	22.00	**58**	320	10.67
29	640	21.33	**59**	310	10.33
30	620	20.67	**60**	300	10.00

Figure 2: Jimmy Johnson Draft Table

2.2.2 NBA

However, unlike the NFL, the NBA does not have a publicly known draft value table. NBA draft value tables do exist, one of which was created by ESPN staff writer Kevin Pelton. In Pelton's first draft value table, he confines the value of a pick to only the years played on the rookie contract since unless that player is traded they will be providing value to the team they were selected on (Pelton, Making smart, valuable trades to move up in the draft is harder than it looks, 2015). Pelton acknowledged that in doing so he decreases the value of a top pick because the value they provide after the rookie contract is also likely more than lower picks. He remade his draft value chart with the addition of looking at how players drafted between 2003-07 performed in years 5-9 of their careers (Pelton, Trade down or keep No. 1 pick: Which is more valuable?, 2017). This time frame was considered because this would be the amount of time covered by a maximum rookie extension. Figure 3 displays Kevin Pelton's 2017 draft value table.

Overall Pick	Value	Normalized	Overall Pick	Value	Normalized
1	4000	100.00	**31**	360	9
2	3100	77.50	**32**	350	8.75
3	2670	66.75	**33**	330	8.25
4	2410	60.25	**34**	320	8
5	2240	56.00	**35**	300	7.5
6	2110	52.75	**36**	290	7.25
7	2000	50.00	**37**	280	7
8	1910	47.75	**38**	270	6.75
9	1830	45.75	**39**	250	6.25
10	1720	43.00	**40**	240	6
11	1600	40.00	**41**	230	5.75
12	1500	37.50	**42**	220	5.5
13	1400	35.00	**43**	210	5.25
14	1320	33.00	**44**	200	5
15	1240	31.00	**45**	190	4.75
16	1180	29.5	**46**	180	4.5
17	1130	28.25	**47**	170	4.25
18	1080	27	**48**	160	4
19	1030	25.75	**49**	150	3.75
20	980	24.5	**50**	140	3.5
21	920	23	**51**	130	3.25
22	860	21.5	**52**	120	3
23	800	20	**53**	110	2.75
24	750	18.75	**54**	100	2.5
25	700	17.5	**55**	90	2.25
26	660	16.5	**56**	90	2.25
27	620	15.5	**57**	80	2
28	570	14.25	**58**	70	1.75
29	520	13	**59**	60	1.5
30	470	11.75	**60**	50	1.25

Figure 3: Kevin Pelton Draft Table

2.2.3 Discussion

When comparing the two draft tables side by side, the values are similar. With the 7th pick having the same relative value in each and the 15th pick having a percent difference of 12% in relative value. The major difference between the two tables comes after these first 15 selections as we transition into the latter half of the first round and into the second round for the NBA. NBA players relative value drops below 10% of the first pick's value at the 31st pick which is the first pick of the second round. Contrast that with the first pick in the second round of the NFL (33) which has a relative value of 19.33% of the first pick. These two picks have a percent difference of 73% which is quite substantial. This large difference suggests that the drop off for relative value in a pick decreases faster and steeper in basketball than they do in football.

A main reason for the large difference is there being less people on a team and playing at one time in basketball than in football. In basketball there are more opportunities for a player to make an impact when playing, since they are playing a large portion of the game. On the other hand, a less skilled player has less opportunities to make an impact since only 5 players on the team are on the court at one time. Considering only the regular season a basketball player can play for an entire game for all 82 games (35 minutes for 75 games is more reasonable but the former is still possible), whereas a football player is on the field for roughly half the game, if their offense is on the field the same amount as the defense, if they play every snap for 16 games. Although a single play or performance has a greater impact on the season outcome in football than in basketball; a higher skilled basketball player will be able to provide more consistent value to their team over a less skilled player to a greater effect than in football.

Furthermore, due to the shorter season and limited time on the field, lucky plays or breakout performances are more likely to occur in football than in basketball. This narrows the gap between how much value a great player vs. a good player can contribute over the course of a season because a good player who gets lucky can provide more value to a team in a game than a great player who is more consistent. Consider a highly skilled wide receiver who has 100 yards receiving on 11 catches, but on all of those drives they failed to score any points. On the other hand, a less skilled wide receiver who had one touchdown catch for 88 yards that was due to a free safety tripping. In the context of the game, the great player provided more consistent value, but the good player added 6 points to the team and would provide more value to their team. Football is a lower scoring game than basketball so lucky plays in football like a "pick six" have a huge effect on the outcome of the game and a lucky play in basketball can result in at most 4 points which likely will not affect the game. This sample size issue can also be reflected in baseball, where the 162-game season gives more context to the low likelihood of getting a hit.

2.3 Assessing Draft Value in the NBA

In 2007, Dr. Aaron Barzilai explored the often-overlooked topic of draft value in the National Basketball Association (Barzilai, 2007). Dr. Barzilai assessed the value of each draft pick using 4 metrics (Player Efficiency Rating, Player Wins, Win Shares, and Estimated Salary) over 3 different time periods (career, first 4 years, and years with rookie team) for a total of 12 total metrics. But Barzilai decided that estimated salary was only meaningful for the career time period, so he considered only 10 metrics. Below are the regression lines for the metrics excluding the years with rookie team due the large amount of variability caused by the differing lengths players spend with their rookie team.

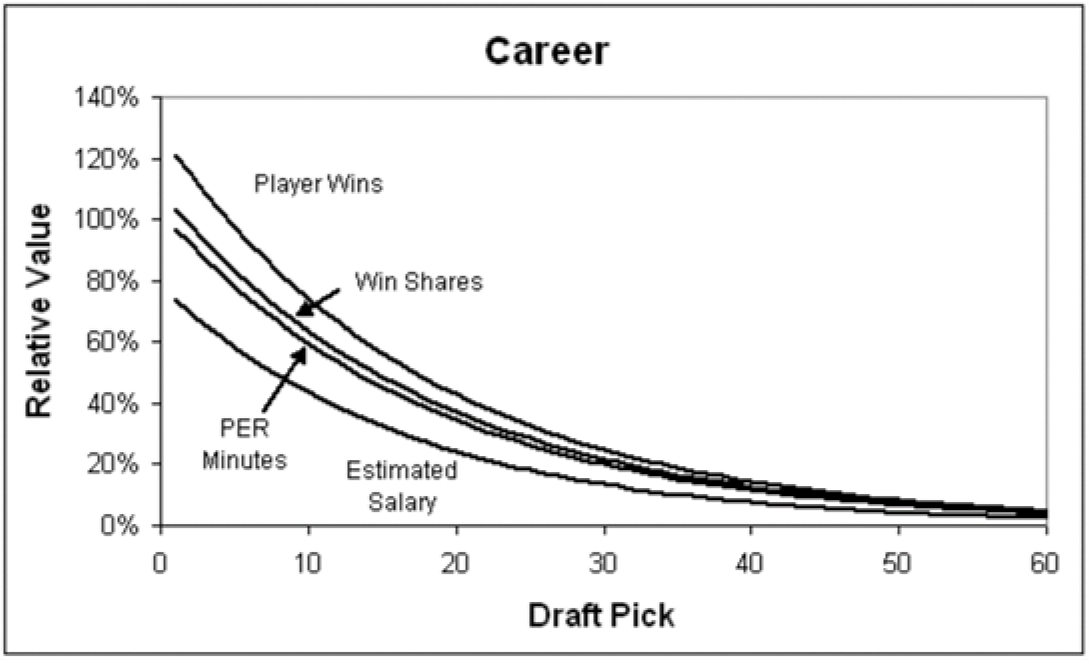


Figure 4: Aaron Barzilai Career Relative Draft Value

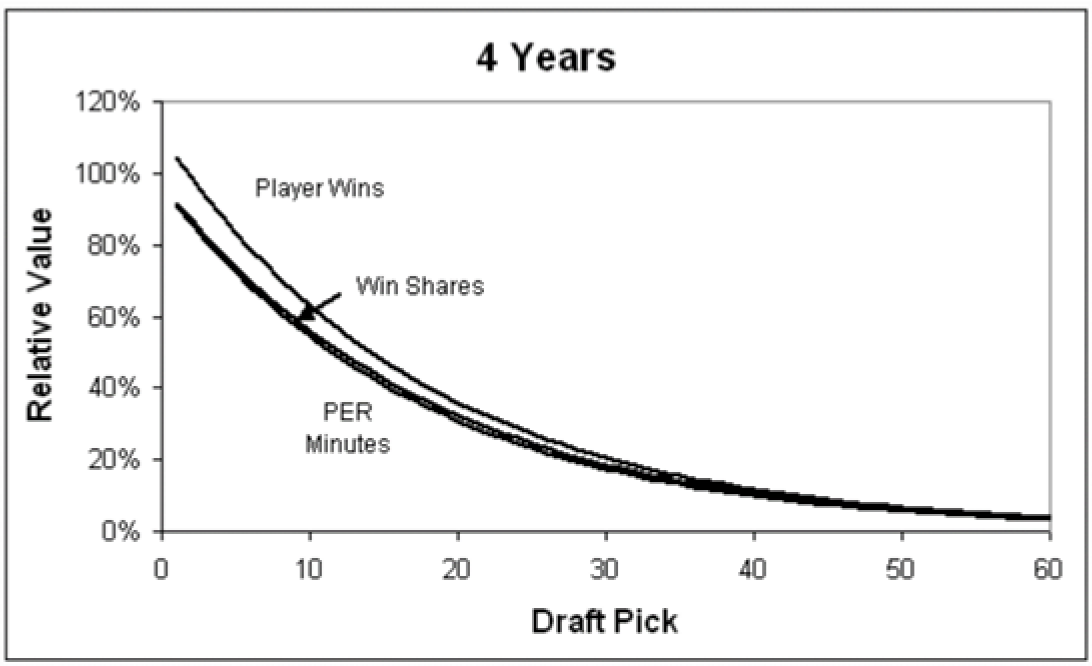


Figure 5: Aaron Barzilai First 4 Years Relative Draft Value

The work done by Dr. Barzilai shows that wins are correlated more to where a player was drafted than PER. A player who was drafted highly, especially a lottery pick, will almost always see the court for a long time. This can be attributed to the fact that higher picks go to lower performing teams. These lower performing teams can take longer to develop these young players and the talent on the team is lower, so the newly drafted player plays far more minutes than a later draft pick who is playing on a perennial playoff team. Although, for most cases a higher pick (earlier selection) is a better player than a lower pick (later selection), there are instances where a later pick will produce more value simply because they are given more chances and could be equally as talented as a lower pick. These late round picks are referred to as steals in the draft and the non-producing early picks are called busts. But in order to figure out if a player is a steal or a bust, they need to have time on the court to showcase their talents. Due to a larger proportion of higher picks getting playing time it makes sense that most people can think of examples of draft busts but not many examples of draft steals. Looking forward, our project will attempt to better quantify what value a player contributes to their team which may shine light on more draft steals.

2.4 Predicting NBA success based on college performance

In American professional sports leagues, drafts are conducted to introduce young talent fairly to all teams. Generally, pick order is decided by inverse order of record, so that worse teams have the first selections and the best chance at picking a superstar. While this system sounds airtight in theory, equality has been increasingly vapid in the NBA. The teams with poor scouting departments-whether it be from personnel or budget limitations-find themselves anchored to the bottom of the standings and making early draft selections each year. Thus, the challenge is to accurately identify successful players from leagues all over the world, using limited data.

Purely numerical statistics are not enough to evaluate a player, however. Players who struggle to make NBA rosters have experienced incredible success in international leagues, with Jimmer Fredette and Stephon Marbury two prime examples. The NCAA is the closest thing to a level playing field NBA teams have to evaluate talent against, as amateur student-athletes play for their college teams. Analysts have used numerical statistics and size in conjunction with subjective scouting to try to predict professional success for collegiate players, to reasonable success. Others try to directly find a relationship between college statistical production and NBA production. What all past models have not done, however, is taking each unique school into consideration when evaluating the likelihood of them making the NBA. Even within the same conference, certain teams are far more likely to send players to professional leagues than others.

2.5 Summary

Overall, we understand that many schools of thought have produced many different numbers to evaluate basketball players. Due to the game’s free-flowing nature, quantifying every effort a player contributes to a team is extremely challenging. Only with recent advancements in player-tracking data are teams beginning to find ways to measure defensive capability, and other factors previously considered intangible.

For the specific application of the draft, relative value is a crucial component of the NFL landscape, where the lengthy draft process leads to many draft pick-only trades. In the NBA, there is a sizable gap in the analysis of draft value, and a lack of discussion regarding the most important statistics to consider when evaluating a prospective talent. We designed our experiments to address these gaps, providing draft value charts for the NBA with statistical rigor, and discovering the most important factors for predicting NCAA DI athlete success in the NBA using machine learning.

3. Design and Methodology

3.1 Determining Scope of the Project

The NBA has had extensive changes to its rules, restrictions on eligibility and size as an association since its creation. In order to best evaluate a modern-day player and produce metrics for their value, it was imperative to consider the time period of the NBA we would include in our dataset. We opted to use data from 1990-2018 in our project. The majority of NBA rules have remained consistent during this timeframe, with one exception being the three-point line's move from 23 feet 9 inches uniformly to 22 feet in 1995 and subsequent extension at the top of the key (corner remained at 22 feet) to 23 feet 9 inches. In the 1990's, more rule changes altered the way on-ball defense was played, removing the ability for the defender to 'hand check' the offensive player. This change was implemented to aid offensive players, making the games higher scoring and thus more entertaining. An important period captured in this dataset is the Jordan years of the NBA. Although not a definitive time period, the NBA in the 90's was focused on physical, defensive play (as demonstrated by the Detroit Pistons' "Bad Boys") to a more offensive and point producing league in the 2000's, with the 3-point explosion revolutionizing the game in the 2010's.

3.2 Collection and Manipulation of the Data

In order to collect the data for our project, we utilized web scraping techniques through the Python package Beautifulsoup. We obtained our information from Basketball-Reference.com which had all of the player data required for the analysis. To produce our dataset, we first iterated through each season and then for each season pulled the player information from three tables. Thee three tables were "per-game", "total" and "advanced." Each of these tables has every player who played a game in that season within the table. Once all of these tables were saved to local spreadsheets, we created functions that cumulatively combined the seasons of data which outputted a single spreadsheet with per-game statistics, total statistics, and advanced statistics for every player in every season they played in the NBA since 1990. To produce the cumulative metric, we also needed to pull data on all-star selections and seasonal awards. We again utilized basketball-reference as for each year they had tables of award summaries that included all award-winning players. These awards were transformed into their own respective column where a 1 indicated they achieved that award and a 0 meant they did not.

3.3 Analyze existing basketball player performance metrics

In professional sports, 'value' can be quantified in many ways. Some measures look purely at statistical output, whereas others take factors such as contract cost, minutes played, and team wins into account. To contextualize our entire project, which involves measuring the performance of basketball players, we analyzed the common metrics used to evaluate players. These four metrics were Player Efficiency Rating (PER), Win Shares (WS), Value over Replacement Player (VORP) and Fantasy Points (FP).

3.4 Feature engineer new player performance metrics addressing shortcomings with existing metrics

After analyzing the existing player performance metrics, we identified potential areas for improvement with different metrics that allowed for a more accurate comparison of players in the same season. These metrics were called Basic Percentile (BP) and Advanced Percentile (AP). Additionally, we created a metric which rewarded recognition rather than statistical output, called Cumulative Individual Accolades (CIA).

3.5 Find the highest value picks based on various measures of cost

One of the most important applications of talent evaluation is the NBA Draft. Each of the thirty teams are assigned two picks, generally in inverse order of team wins. A lottery is conducted for the first fourteen picks, to disincentivize intentional losing of games (commonly referred to as 'tanking') to obtain a highly talented player with the first pick. The NBA rookie salary scale provides an approximation of the talent level available at each pick, which we use with the performance metrics to find the draft picks which provide the highest output per dollar.

3.6 Calculate the approximate value of every pick in the NBA Draft

Another possibility in the NBA Draft is pick trading. Both before and during the draft, teams can swap picks for players or even high picks for multiple lower picks. As such, knowing the value of each position in the draft is critical to teams trying to improve their talent. We use the performance metrics to analyze the drop-off in talent at each pick in the draft.

3.7 Create a Jimmy Johnson-style NBA Draft value chart

Pick trading is far more common in the National Football League (NFL) where there are 224 picks between 32 teams. NFL Analyst Jimmy Johnson created a draft chart in the early 1990's which seeks to quantitatively evaluate the talent available at each pick. We apply this to the NBA and create a value chart which accurately matches past draft pick trades in the NBA.

3.8 Summary

Overall, the key goals of this project section are to identify new avenues for basketball player performance evaluation and using that knowledge to generate useful information regarding the value of draft picks. We verify our approach through comparing the results to existing research done in the NFL and NBA.

4. Results

4.1 Analyze existing basketball player performance metrics

As discussed in the previous section, a crucial decision in evaluating player value is how 'performance' is quantified. Figure 6 lists the top 20 players ranked using the four existing metrics, averaged out over the course of each player's career.

Player	WS	PER	VORP	FP	AVG
LeBron James	1	2	1	1	1.3
Karl Malone	2	4	2	2	2.5
David Robinson	4	1	4	5	3.5
Tim Duncan	8	6	9	4	6.8
Chris Paul	5	7	5	11	7.0
Kevin Durant	7	8	15	7	9.3
Shaquille O'Neal	14	3	18	3	9.5
Michael Jordan	3	11	3	21	9.5
Charles Barkley	10	5	6	21	10.5
Russell Westbrook	21	16	8	6	12.8
Kevin Garnett	17	17	12	8	13.5
John Stockton	6	13	19	16	13.5
Hakeem Olajuwon	21	10	17	10	14.5
James Harden	12	21	11	17	15.3
Clyde Drexler	16	19	7	21	15.8
Stephen Curry	15	21	10	19	16.3
Kobe Bryant	20	15	21	13	17.3
Dirk Nowitzki	13	18	21	18	17.5
Magic Johnson	9	21	21	21	18.0
Dwight Howard	18	21	21	12	18.0
Yao Ming	21	9	21	21	18.0
Allen Iverson	21	21	21	9	18.0
Jason Kidd	21	21	16	15	18.3
Dwyane Wade	21	12	20	21	18.5
Reggie Miller	11	21	21	21	18.5
Scottie Pippen	21	21	13	21	19.0
Larry Bird	21	21	14	21	19.3
Anthony Davis	21	14	21	21	19.3
Gary Payton	21	21	21	14	19.3
Jeff Hornacek	19	21	21	21	20.5
Amare Stoudemire	21	20	21	21	20.8
Patrick Ewing	21	21	21	20	20.8

Figure 6: Top 20 Players by existing metrics

Starting at the top, we can see that there's a reasonable consensus among the top three players. Beyond that, the metrics begin disagreeing quite significantly. For example, Michael Jordan earns third place in Win Shares and VORP, but doesn't feature in the top 20 for Fantasy Points. Because Win Shares distributes production by the number of wins the teams accrues, players on successful teams (such as the 90's Bulls, arguably the greatest team ever) will feature strongly in the WS rankings. Similarly, Magic Johnson's extremely strong Lakers teams boosts his WS rank to 9, which is the only time he features in these standings.

Extrapolating from this chart, if these metrics disagree so significantly for the absolute best players, it's likely that mediocre players will also have large disparities in their statistical rankings by each metric.

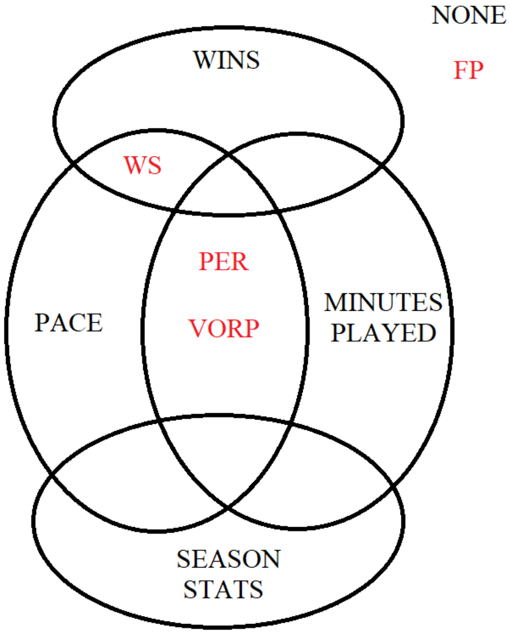


Figure 7: Existing metric Venn diagram

To investigate just what these statistical disparities might be, we broke down each metric to its mathematical formula, to see their components. We were particularly interested in the components which normalized each metric, displayed in Figure 7.

Fantasy Points is the most basic metric – it multiplies each basic 'counting stat' by a coefficient and outputs a number representing the volume of statistical output by a player. The coefficients seek to equalize the value of assists, rebounds, and points. FP does not consider the player's efficiency, or pace of play. Obviously, 20 points in a game ending 74-68 is more valuable than 25 points in a 135-123 game, but FP would rank the latter performance as stronger. By normalizing to pace, the metric would consider the amount of points the player scored per 100 possessions, allowing for a more accurate comparison.

In that case, let's now move to PER, a stat which is normalized to pace, as well as minutes played. It multiplies counting stats by coefficients and analyzes the proportion of team field goals the player's assists contribute towards. Additionally, PER subtracts what its creator, John Hollinger, calls "negative accomplishments" such as turnovers, personal fouls, and missed defensive rebounds. PER's largest flaw is its greatest strength- minutes normalization. Because of limited sample size, the player with the all-time highest PER has only played a few minutes. Adding minimum games or minutes played removes these outliers, but on the other end, players who make significant contributions during their prime, only to decrease in efficiency in their career's twilight are prone to having a low career average PER.

As such, there is no true 'best metric' for evaluating talent. Undoubtedly, every player on this list is a great player in their own right, but such significant difference in the ranking suggests there might be a better way to evaluate talent.

4.2 Feature engineer new player performance metrics addressing shortcomings with existing metrics

4.2.1 Cumulative Individual Accolades

When fans compare players, they often point to the number of individual awards a player accrues over their career. With that in mind, we sought to quantify these awards by examining the mathematical chance that a player accomplishes a certain milestone if all players were randomly selected.

The baseline accomplishment is being named in the 12 active players for each game, which we assign one point to each player. From there, five players are named to the starting lineup (5/12), which is equivalent to 2.4 points. We follow the same methodology for playing a minute on the court, all the way to winning the MVP, which is a 1/450 chance (given 15 players on 30 teams' rosters), thus awarding 450 points.

$$CIA = games\ active * \frac{12}{12}^{-1} + games\ started * \frac{5}{12}^{-1} + minutes\ played * \frac{10}{24}^{-1} + made\ all\ NBA * \frac{5}{450}^{-1} + made\ all\ Defense * \frac{5}{450}^{-1} + made\ all\ Rookie * \frac{5}{60}^{-1} + was\ RotY * \frac{1}{60}^{-1} + was\ MIP * \frac{1}{450}^{-1} + was\ DPOY * \frac{1}{450}^{-1} + was\ MVP * \frac{1}{450}^{-1} + was\ 6MOTY \frac{1}{210}^{-1}$$

Figure 8: CIA Equation

Player	**2018 CIA**
Victor Oladipo	1242
James Harden	1152
Rudy Gobert	976
Anthony Davis	835
Lou Williams	832
LeBron James	816
Jrue Holiday	792
Karl-Anthony Towns	791
Russell Westbrook	789

Figure 9: CIA Top 10

Figure 9 shows the top 10 players as ranked by CIA for 2018. Victor Oladipo won Most Improved Player, made the All-NBA Defensive Team and the All-NBA Third Team. James Harden won MVP and was named to the All-NBA First Team. Because the statistical likelihood of making the Third Team is equivalent to making the First Team, it slightly muddies the data. Similarly, Most Improved Player awards the same points as MVP. While this metric was an interesting twist on the typical in-game analysis of player performance, we found it to be inappropriate to further analyze players using this metric.

4.2.2 Basic Percentile

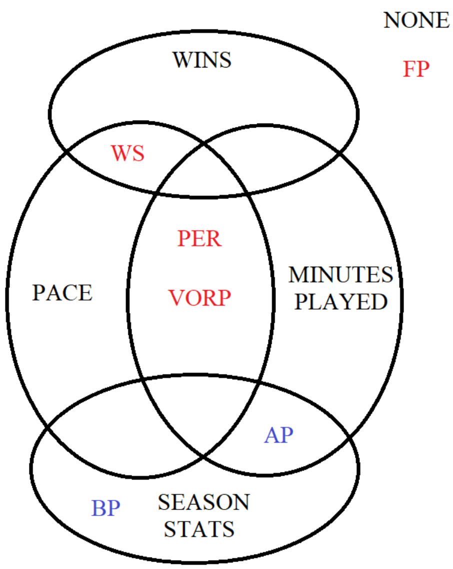


Figure 10: All metrics Venn diagram

There are five major ‘counting stats’ in basketball and are the basis for almost all stats as they tally a player’s basic contributions to their team. The five stats are points, assists, rebounds, blocks and steals. We felt that there was a need to develop a stat that was basic, yet still provided the normalization in the other metrics. When viewing Figure 10, we realized we were not considering stats that looked at the volume produced by a player that only be adjusted to the season they played in. This last clarification is an important one because the speed of the game has increased since the first seasons we were comparing. It would be unfair and improper to treat every season equally and just take the raw outputs of players for these five categories. The pace of the game is higher so there are more points being scored which means more assists and rebounds to be had.

With this in mind, we decided to rank every player by the average of their five major stat categories. The equation is similar to Efficiency but removes the negative parts of the equation and instead ranks the players based on their relative performance compared to the rest of the league.

Efficiency = (Points + Rebounds + Assists + Steals + Blocks) - ((Field Goals Att. - Field Goals Made) + (Free Throws Att. - Free Throws Made) + Turnovers)

Basic Percentile

Each of the five major stat categories turn into ranking where a players rank is determined with the following:
Let X be the stat in question:

X_Rank = sort(All players by X in non increasing order)

E.g. The player who scores the fewest points will be given the PPG_Rank = 1 and the league leader in points will have a PPG_Rank of N, where N is the number of players in that season.

$$\textbf{Basic Percentile} = \frac{(PPG\_Rank + APG\_Rank + TRB\_Rank + BLK\_Rank + STL\_Rank)}{5*N} * 100$$

The reason that we divide by 5 is to get the average rank for all of the 5 major stat categories and we also divide by N to get the percentile of where that player stands off of the total possible score that is achievable. The multiplication by 100 is simply move the metric two decimal places to the right so that the results is easier to read.

The stat gives a raw number that can range from 0-100 and is adjusted to a per season output. A player who leads the league in year X but averages 20 points will get the same PPG_Rank as a player who leads the league in year Y and averages 45 points.

Player	Age	BPercentile
Giannis Antetokounmpo	22	94.53
DeMarcus Cousins	27	94.37
DeMarcus Cousins	26	93.37
Giannis Antetokounmpo	23	93.15
Hakeem Olajuwon	32	92.43
Kevin Garnett	27	92.22
Hakeem Olajuwon	33	92.03
David Robinson	28	92.01
DeMarcus Cousins	24	91.99
DeMarcus Cousins	25	91.97
LeBron James	33	91.96
Hakeem Olajuwon	30	91.95
LeBron James	23	91.93
LeBron James	24	91.87
LeBron James	25	91.86
LeBron James	28	91.86
Kevin Garnett	28	91.64
Chris Webber	26	91.53
Chris Webber	23	91.52
Chris Webber	29	91.50

Figure 11: Top 20 Basic Percentile

The table to the left is the top 20 basic percentile scores since 1990. The reason we believe this metric adds value is it highlights the "stat stuffers" of the NBA, it recognizes the players who have a propensity to add value in all of the major aspects of the game. The idea of adding weights to each of the 5 stat categories was considered. A valid argument for doing so would be since assisting is a vital role to a point guard, we should weigh assists higher than rebounds, a stat usually tied to forwards and centers. For example, a point guard who leads the league in assists but is 200^{th} in rebounds can get the same basic percentile score as another point guard who is say 50^{th} in the league for assists and 150^{th} in rebounds. Some would argue that the league leading assist point guard is providing more value. And in a future iteration perhaps weighting will be added. But the purpose of this stat was to eliminate raw numbers and fancy equations so equally rating all stat categories the same made the most sense.

4.2.3 Advanced Percentile

When evaluating the results and rankings generated by basic percentile it became obvious that there was an aspect missing to the metric. Since basic percentile only looks at per game metrics those players who play more minutes per game were more likely to get higher basic percentile scores. Although minutes played is a good indicator of their perceived value on the team, one of the goals of this project was to try and find undervalued players. For this reason, there was a natural progression which led to the creation of a new metric we call advanced percentile. Instead of looking at raw per game stats, we were now going to calculate the 5 core stats not by their per game output but their accompanying percent metrics.
Therefore the 5 stats we used were TS%*, AST%, TRB%, BLK%, STL%.
They are calculated by the following equations:

*TS% = PTS / (2 * FGA + 0.44 * FTA)*

**The reason we used true shooting percentage is because our data source did not have a metric that fit the same style as the stats below for points. It could have been possible to calculate a similar metric but we reasoned that although true shooting percentage does not take into account how many points a player scored highlighting the efficiency with which they do score we saw as fairly similar in value. In future work it might be best to reevaluate this stat to the Points% which could be calculated with the following equation: Points% = 100* Points/ (((MP / (Tm MP/ 5)) * Tm Points). But since our data source had the below stats but not a metric like the above we decided to use TS%.*

*AST% = 100 * AST / (((MP / (Tm MP/ 5)) * Tm FG) - FG)*
*TRB% = 100 * (TRB * (TmMP / 5)) / (MP * (Tm TRB + Opp TRB)).*
*BLK% = 100 * (BLK * (Tm MP / 5)) / (MP * (Opp FGA - Opp 3PA))*
*STL% = 100 * (STL * (Tm MP / 5)) / (MP * Opp Poss)*

Advanced Percentile

Each of the five major stat categories turn into ranking where a players rank is determined with the following:
Let X be the stat in question:

X_Rank = sort(All players by X in non increasing order)

E.g. The player who scores with the lowest TS% will be given the TS%_Rank = 1 and the league leader in TS% will have a TS%_Rank of N, where N is the number of players in that season.

$$\textbf{\textit{Advanced Percentile}} = \frac{(TS\%\_Rank + AST\%\_Rank + TRB\%\_Rank + BLK\%\_Rank + STL\%\_Rank)}{5*N} * 100$$

For the same reasons as described in basic percentile we divide by 5*N.

Player	Age	A Percentile
Giannis Antetokounmpo	22	87.53
Cole Aldrich	27	86.97
David Robinson	26	86.89
Hakeem Olajuwon	30	86.51
Oliver Miller	23	86.25
Shawn Kemp	24	86.20
Andrei Kirilenko	23	85.99
Kevin Garnett	31	85.90
Kevin Garnett	28	85.82
Kevin Garnett	29	85.76
David Robinson	28	85.61
DeMarcus Cousins	27	85.59
Kevin Garnett	27	85.34
Jordan Bell	23	85.33
David West	37	85.30
Arvydas Sabonis	38	84.91
LeBron James	28	84.78
Andrei Kirilenko	22	84.75
Draymond Green	25	84.29
DeMarcus Cousins	24	84.23

Figure 12: Top 20 Advanced Percentile

The table to the left shows the top 20 advanced percentile scores since 1990. This table is far more interesting to look at as there are players who are not considered all time players like previous metrics we have seen. For example Cole Aldrich when he was 27 (2015-2016 season with the Clippers). In that season he had a TS% of 62.6, TRB% of 19.6, AST% of 10%, BLK % of 6.7, and STL% of 2.9 while playing 13.3 minutes per game. In that season there was 475 players (N=475) and his rankings were the following.

TS%_Rank 452/475 = 95.2

AST%_Rank 238/475 = 50.1

TRB%_Rank 459/475 = 96.6

BLK%_Rank 468 /475 = 98.5

STL%_Rank 453/475 = 96.4

There are two arguments that can be made from his relatively low minutes per game, either this stat overvalues performance for players who play few minutes or Cole Aldrich should have played more minutes that season. Both are rationale and could be explained but it is worth noting that Cole Aldrich had a WS/48 of 0.209 which is reasonably high (243rd all-time best single season WS/48) and is behind only the career WS/48 averages of Michael Jordan, George Mikan, LeBron James and Kawhi Leonard. But regardless of whether Cole Aldrich is being over valued from this metric is not a concern. The purpose of this metric was to highlight seasons like this which are too often overlooked. Of course, there are players who are overvalued from this metric. The leader in TS% from the 2015-2016 year was Rakeem Christmas who had a TS% of 1.00 because he took two shots in 6 minutes and made both and then never played again that year. But this metric also highlights the seasons like Cole Aldrich's and Oliver Miller's which are overlooked and forgotten but show promise in terms of providing value.

4.3 Calculate the approximate value of every pick in the NBA Draft

Following our analysis of existing metrics, and construction of BP and AP, we then group players based on their draft position. First, we summed up the total value of each metric of each draft pick. We included non-drafted players as 'Pick 61', which is displayed on the below graph.

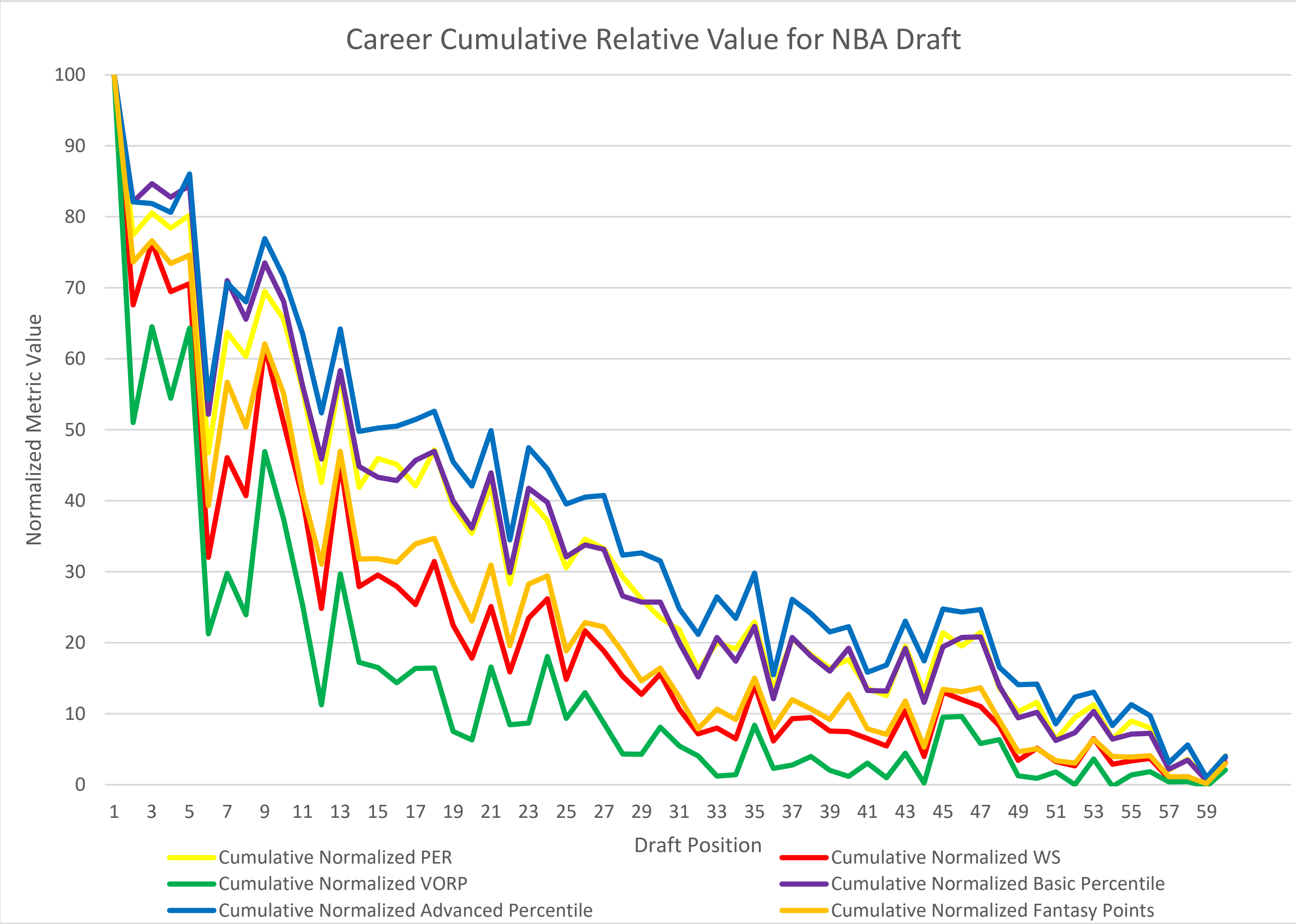


Figure 13: Career Cumulative Relative Value for NBA Draft

This graph is oversensitive to extremely good players, which makes the graph jagged. Additionally, it is notable that undrafted free agents are typically more productive than the final few picks. A potential reason for this is that they're generally older and are more prepared for the rigors of the NBA. In order to provide a more accurate curve, we cluster the draft picks into groups. These groups are 1-3, 4-7, 8-14, 15-30, 31-45, and 46-60. We felt these clusters fall in line with how picks are generally compared to one another.

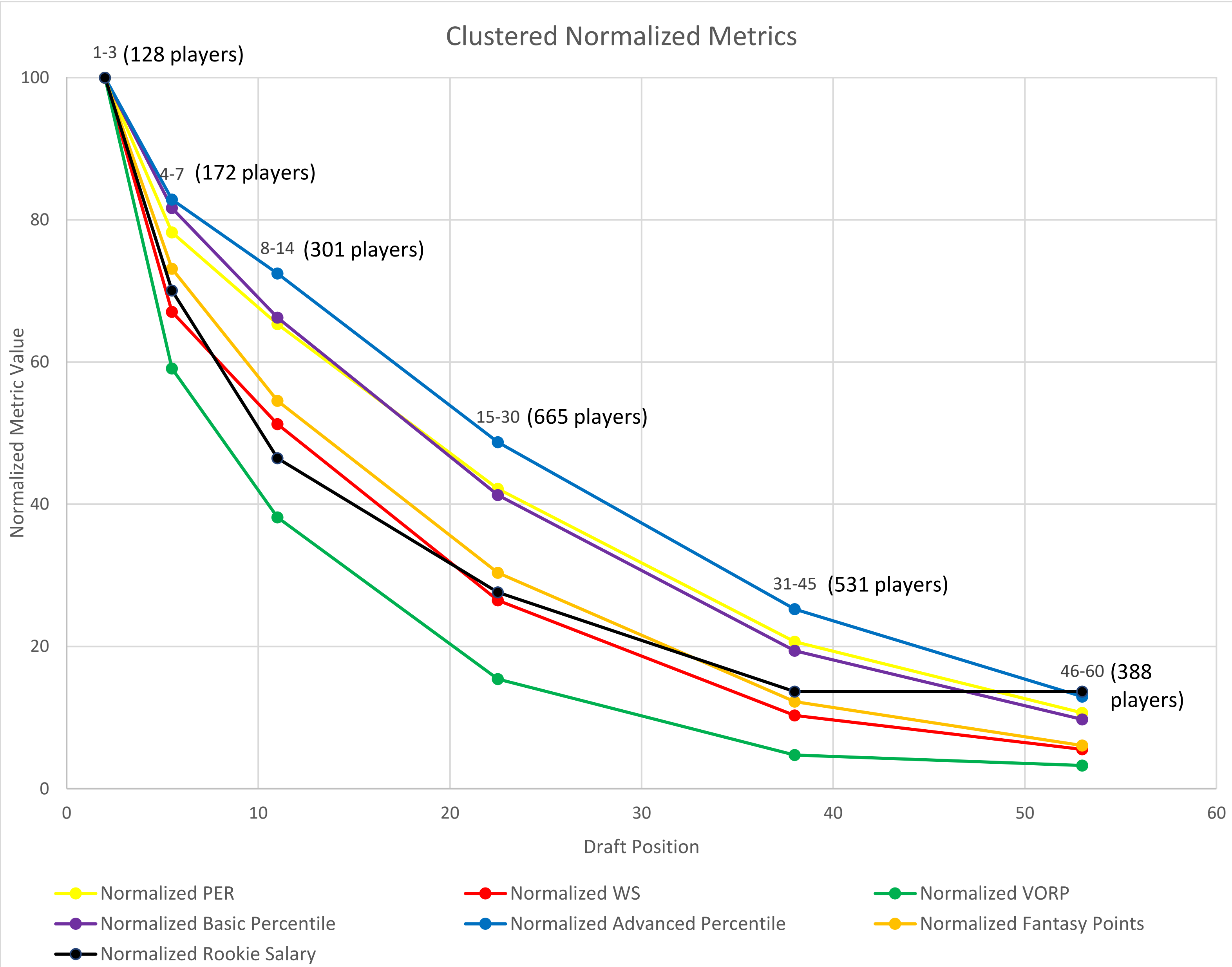


Figure 14: Clustered Career Relative NBA Draft Value

This graph provides a much clearer picture of the values of each metric. Also featured in this graph is the NBA Rookie Salary scale. As there is no mandatory salary for second round picks, we use the league minimum salary. We also display the number of players calculated in each cluster, for context.

Using trendlines, we were able to construct mathematical equations for each metric's value.

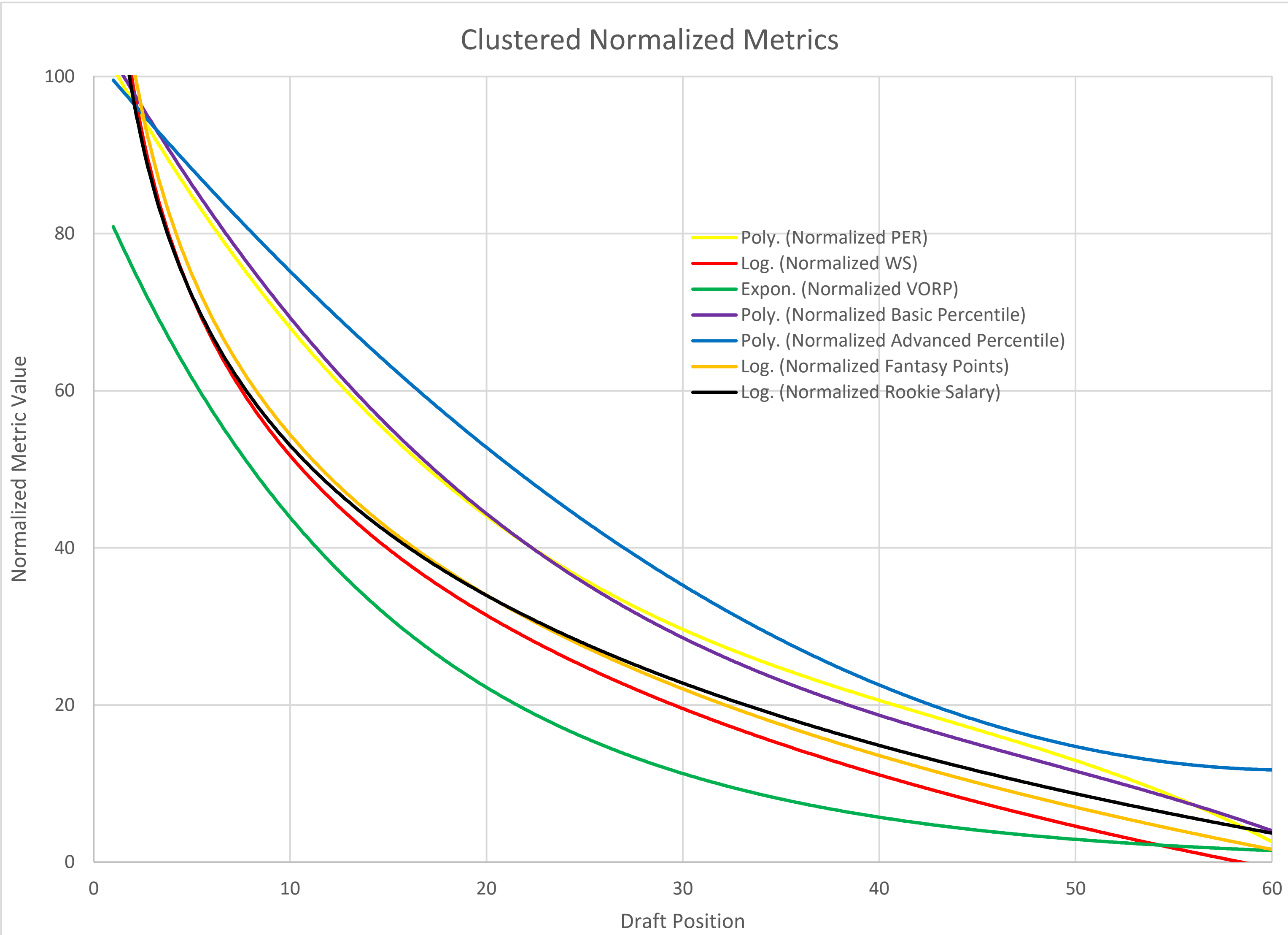


Figure 15: Trendline Clustered Cumulative Relative NBA Draft Value

4.4 Find the highest value picks based on various measure of cost

First, we use the obvious measure of cost, salary, to divide the pick values by. This shows where the best 'bang-for-the-buck' can be found in the NBA draft. We again use the clustering technique to clearly visualize the curves.

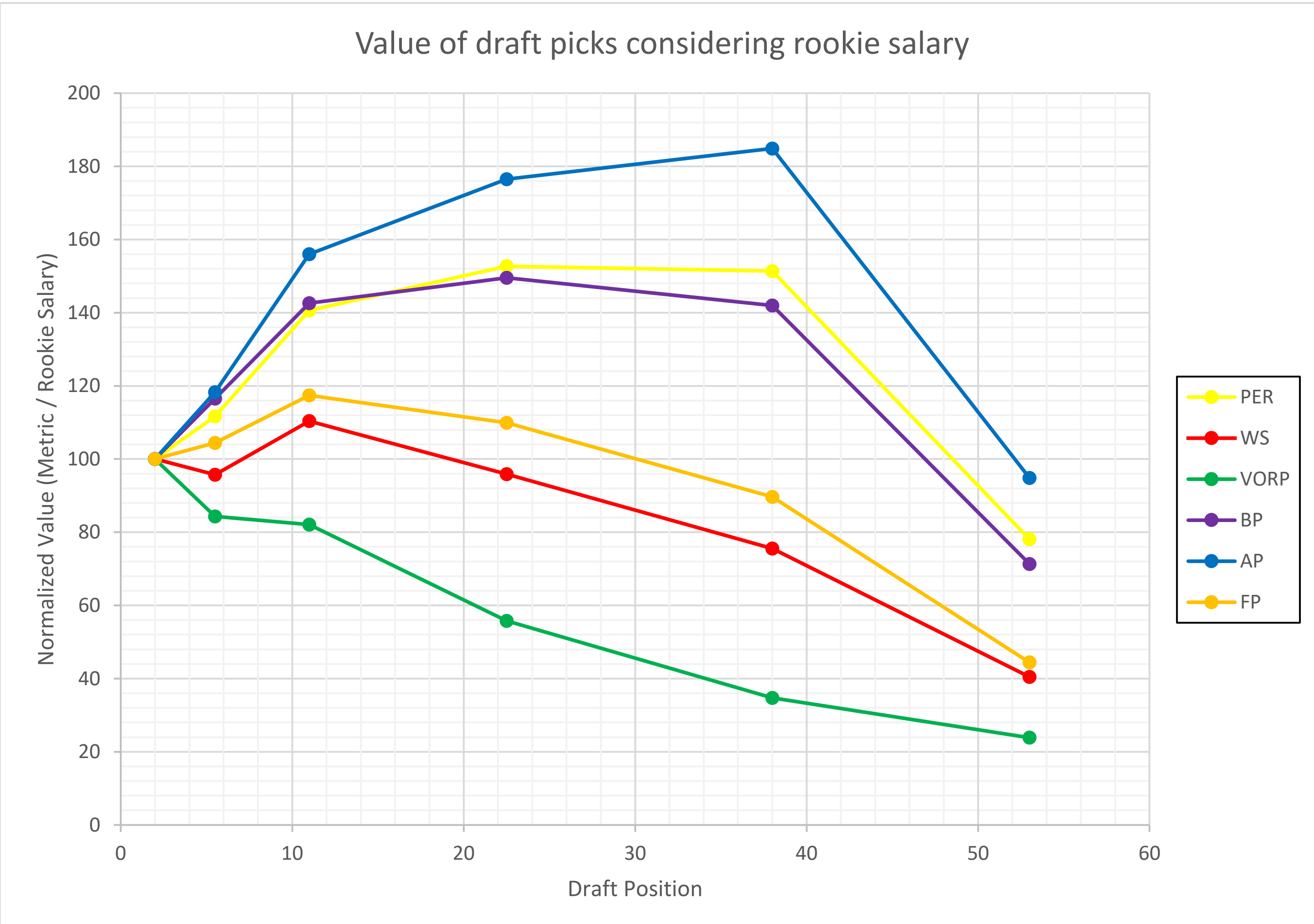


Figure 16: Value per dollar for NBA Rookies

As shown, the metrics disagree greatly in where the highest value can be found. Advanced Percentile suggests the early second round has the best value players, but VORP values the top three picks as the superior selections.

4.5 Create a Jimmy Johnson-style NBA Draft pick value chart

Metric	MAE
VORP	0.0454
WS	0.0707
FP	0.0814
RS	0.0969
AVG	0.1121
PER	0.1494
BP	0.1673
AP	0.1987

Figure 17: Mean Absolute Error of Draft Day Trades based on relative values

We created draft value charts for each pick. NFL Analyst Rich Hill used Jimmy Johnson's chart as a baseline to evaluate draft-pick only trades to create a new draft value chart. With this in mind, we found an assortment of draft-pick only trades in the NBA to evaluate each of the draft charts and select a 'best' chart.

Clearly, VORP is the most accurate chart. In addition to the numerical output of the trade evaluator, selecting VORP is intuitive because it does not cloud the statistical value of a player by normalizing output to wins. On the flip side, because VORP is normalized to pace and minutes played, this provides a more objective value of a hypothetical player who would have equal opportunity on each team.

Position	VORP	Position	VORP	Position	VORP	Position	VORP
1	3000	16	1082	31	390	46	141
2	2803	17	1011	32	364	47	131
3	2619	18	944	33	340	48	123
4	2446	19	882	34	318	49	115
5	2286	20	824	35	297	50	107
6	2135	21	770	36	278	51	100
7	1995	22	719	37	259	52	94
8	1864	23	672	38	242	53	87
9	1741	24	628	39	226	54	82
10	1627	25	587	40	212	55	76
11	1520	26	548	41	198	56	71
12	1420	27	512	42	185	57	67
13	1327	28	478	43	172	58	62
14	1239	29	447	44	161	59	58
15	1158	30	418	45	151	60	54

Figure 18: NBA Draft Relative Numeric Value

Compared to the NFL, the NBA follows a different level of apparent talent drop-off.

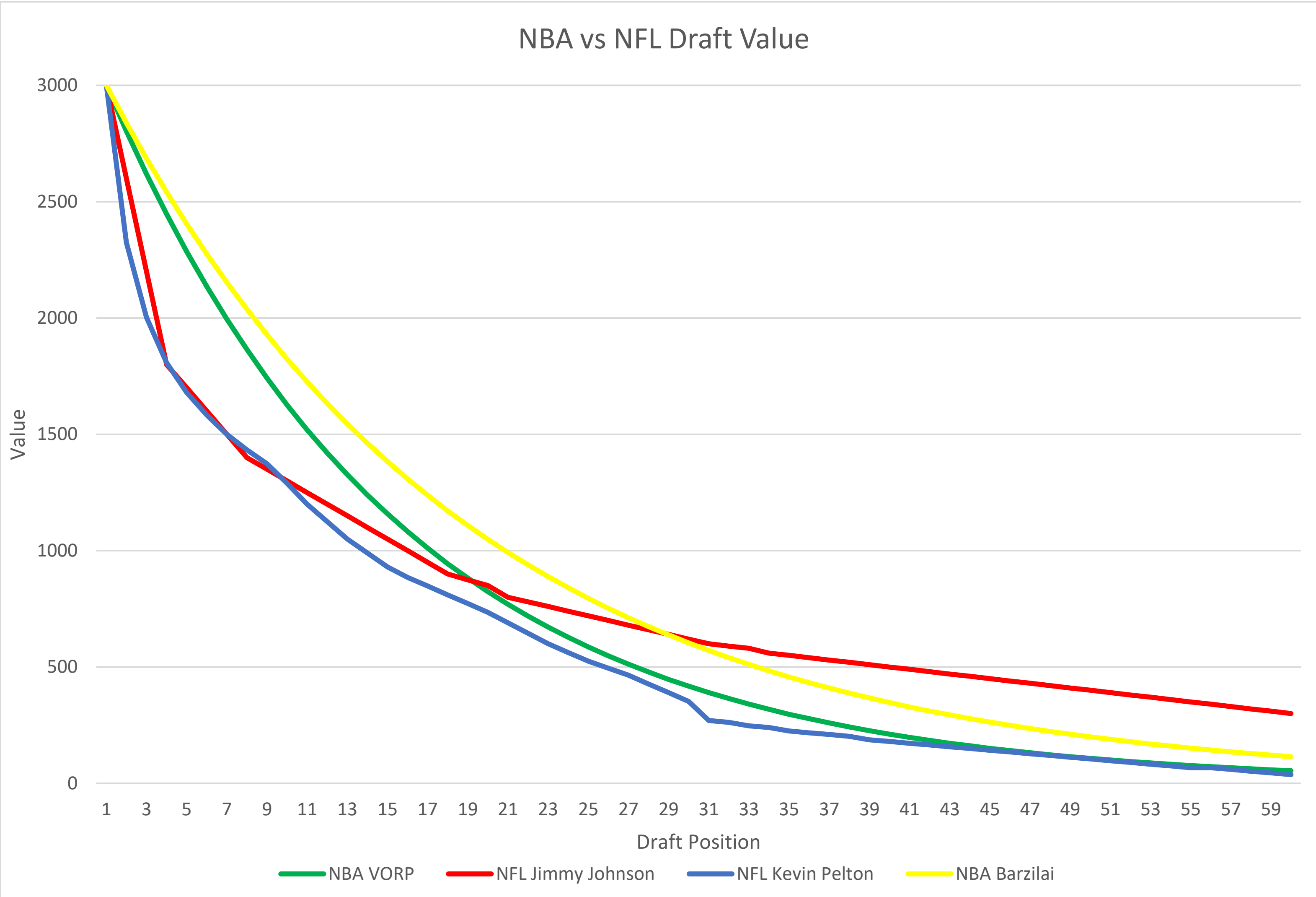


Figure 19: NBA vs NFL Draft Value

For the first 20 picks, NBA talent is relatively better than the same draft pick in the NFL, according to both Kevin Pelton and Jimmy Johnson's charts. After that point, Jimmy Johnson's chart suggest NFL talent is more valuable from picks 20-60. On the other hand, Kevin Pelton's chart closely mirrors the NBA value chart through to the end of the NBA Draft.

5. Design and Methodology for NCAA

5.1 Create a model which predicts various measures of NBA success based on NCAA DI statistics

Following our draft analysis, we pivoted to a more predictive analytics problem. We sought to create models which predicted the best players each year in the NCAA. The NCAA is the primary feeder league for the NBA, so creating a system which models this is critical.

The data we had collected during the first phase of the project lent itself to using primarily draft-based criteria. From the draft position column, we were able to create 'wasDrafted', 'wasFirstRoundPick', and 'wasLotteryPick' fields. By looking up the college player on the pro basketball-reference site, we were able to infer if they ever saw the court for an NBA game.

With four classification targets, we experimented with different machine learning models to find the best fit for each problem. The four models we investigated were:
- Logistic Regression
- Decision Tree
- Random Forest
- Multilayer Perceptron (Neural Networks)

We first used the GridSearchCV function to explore a range of parameters for each model, and then used the best of each individual model in competition with each other. We printed classification reports for all the models and found that Logistic Regression was the most successful model for all four target classifiers.

Once we had a grasp on the value of a player and the expected value from a given draft pick, we set out to predict NBA performance for NCAA Division I players. To do this, we first needed to gather statistics about all NCAA Division I players. Using the same methodology to pull data from Basketball-Refernce.com, we were able to obtain college data from Sports-Reference.com. We were able to pull data from all NCAA division I teams from 2000 – 2018. But due to the lack of consistent IDs for an NCAA player (the ids used in sports reference are not the same as the ones used in basketball reference), we needed to manually enter when a player was drafted and so we focused on college players from 2010 to 2018. When we were evaluating NBA player performance, only in game performance was accounted for, but physical attributes are an important component of evaluating NBA readiness. Thus, we also used height and weight measurements for all NCAA players. To further investigate how physical attributes play a role in NBA success, we also collected data from the NBA Draft Combine from 2010-2018.

After collecting all the above data, we used Python with sklearn, a machine learning package, to predict whether or not a player would make the NBA. We defined making the NBA as playing in an official game during the NBA season. This excludes players who were drafted and never played a game, as well as those who signed contracts and were on NBA rosters but failed to play in a game. These distinctions echo the distinctions that are enforced on the sports reference page in order for a college player to be considered having gone on to play in the NBA. We created and ran a logistic regression, decision tree classifier, random forest classifier, MLP classifier, and

Zero R model to see which model would be best at predicting whether a player would make the NBA. The Zero R model, predicting every player as never making the NBA, was going to be our baseline. Since the vast majority of NCAA DI players never make the NBA, a model that predicts no one will make the NBA is still correct over 99% of the time. But in order to tell a story worth listening to we needed to predict the players who did end up making the NBA.

Once we had a clean dataset, we used stratified sampling to split the data proportionally based on class value. We also normalized the non-target attributes, to make sure no attribute was being artificially weighed more than another. We tinkered with the parameters for each of the models, until we found the best performing set of parameters for each model. At that point, we ran our experiments on each of the target classes, which were: madeNBA, wasDrafted, firstRound and lotteryPick. We then used sk-learn's classification_report to print the resulting precision, recall, accuracy, and f1 score for each of the classes.

To improve the prediction ability of our model, while also using realistic sub sections of NCAA DI players, we broke up our dataset into the following categories.

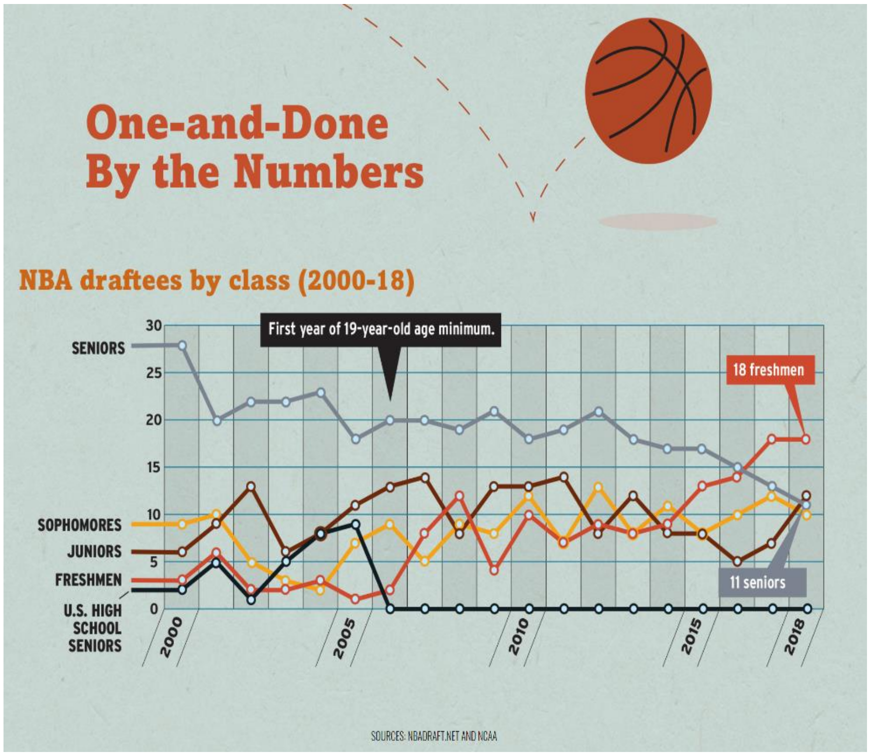


Figure 20: Increasing numbers of freshmen in the NBA (Reynon, 2018)

Freshmen only: We decided that it would be appropriate to only look at players who were in their freshmen year because the trend of freshmen being drafted, especially in lottery selections, has been increasing (seen in Figure 20). From our previous work on NBA performance and the expected value of a pick it was appropriate to put an extra consideration on lottery picks. In the 2018 draft 11 of the 15 lottery picks were freshmen, the other four being international player at 3, junior at 10, sophomore at 12, and junior at 13. In the 2017 draft 11 of the 15 lottery picks were also freshmen. The other four being international at 8, sophomore at 12, sophomore at 13, and junior 15.

Last Year of College: We decided that including the last year a player played would be a good sub section of players to consider as well. This is because this subsection inherently captures a player's last season which could be argued is most likely their best season.

Note: due to an overwhelming amount of empty data points in the combine anthropology and agility datasets we didn't include these metrics. This is due to three main issues, the first is that players who attend the combine rarely perform all the tests, the second is the most notable college players rarely perform any of the tests if they attend at all (the vast majority of NCAA players also do not attend) and lastly the combine usually occurs only a month before the NBA draft and by then most scouts/ fans have already decided who they feel are most draft worthy. For these reasons, we decided that adding the combine metrics to our machine learning models would negatively affect the model's ability to predict NBA readiness.

5.2 Summary

The prediction of whether an NCAA DI player would achieve our target classes, madeNBA, wasDrafted, firstRound, and lotteryPick, are based on machine learning models that trained on thousands of recent NCAA DI player's seasons. We conducted these experiments with multiple machine learning models, including logistic regression, decision trees, random forests, neural nets, and Zero R classifier (as our baseline). We trained and tested these models on all NCAA DI seasons from the freshmen class of 2012 until 2018 and on two subsections of NCAA DI players, freshmen year only and last year of college, to maximize our machine leaning model's predictive ability. The results of this aspect of our project can help NBA teams verify their scouting reports or reveal overlooked collegiate players.

6. Results for NCAA

To generate the most meaningful conclusions and create the best predicting model for NCAA DI players we tested multiple machine learning models on our data. First, we ran each model individually, tweaking the hyperparameters to find the best individual model performance. Then, we evaluated how good a model was by its ability to predict the target attribute, in this case Made NBA, with the entire dataset. We considered the best model to be the model with the highest f1 score for the Made NBA class. Below are the statistics for each model.

Metrics for: madeNBA
Decision Tree

	precision	recall	f1-score	support
No NBA	0.99	0.99	0.99	5653
Made NBA	0.27	0.29	0.28	59
avg / total	0.99	0.98	0.98	5712

Metrics for: madeNBA
Random Forest

	precision	recall	f1-score	support
No NBA	0.99	1.00	1.00	5653
Made NBA	0.80	0.14	0.23	59
avg / total	0.99	0.99	0.99	5712

Metrics for: madeNBA
Logistic Regression

	precision	recall	f1-score	support
No NBA	0.99	1.00	0.99	5653
Made NBA	0.49	0.32	0.39	59
avg / total	0.99	0.99	0.99	5712

Metrics for: madeNBA
Multilayer Perceptron

	precision	recall	f1-score	support
No NBA	0.99	0.99	0.99	5653
Made NBA	0.33	0.27	0.30	59
avg / total	0.99	0.99	0.99	5712

Figure 21: Model experimentation results

We ran the above test with multiple seeds and each time the logistic regression proved to be our best model. In particular, the multilayer perceptron (neural networks) model was particularly volatile based on the random seed. As a result, we use only logistic regression when analyzing different scopes and target we were trying to predict.

As discussed in the previous chapter, we ran twelve experiments featuring four different targets and three distinct datasets. In the interest of brevity, and in order to extract the most meaningful conclusions possible, we have selected the target attribute from each dataset which resulted in the best predictive model. The results of all twelve experiments can be found in the appendices.

Finally, we ran all four experiments once again on the test set of 2018-19 NCAA players to observe our model's predictions for the upcoming 2019 NBA Draft.

6.1 Using all seasons of NCAA DI players

The first group of NCAA DI players that we considered was every season played by every player since the freshmen class of 2012. As mentioned in the design, we excluded players who were not freshmen in 2012 because their previous years were outside of our dataset. Because we initially read in all 2012 players onwards, we confused 2012 seniors with freshmen, thus prompting the shift to 2012 freshmen and beyond only.

The following subsections are the logistic regression's precision, recall and f1 scores for the best target prediction, starting with the entire dataset. This dataset features every player's season as an individual row. Because the data labels a season as being enough to achieve a target, as opposed to a player, the data has multiple rows for the same player with different target values.

Using the entire dataset described above, we found that the best model predicted NCAA DI players being drafted into the NBA.

Logistic Regression

	precision	recall	f1-score	support
Not Drafted	0.99	1.00	1.00	5660
Drafted	0.68	0.44	0.53	52
avg / total	0.99	0.99	0.99	5712

Figure 22: All NCAA season wasDrafted metrics

While initially, the 0.53 f1-score doesn't sound promising, we created a function to map the predictions back to the players' names and then researched their backgrounds. Below is a graph displaying each season as a circle, with the color corresponding the prediction the model made and its correctness. On the x-axis is the numerical value between 0 and 1 that the model predicted for that season. For a logistic regression, a number above 0.5 is determined to be a 1, and vice versa. While the perfect model would have no false negatives or false positives, we would be more confident in a model which has the incorrect predictions concentrated around the 0.5 dividing line.

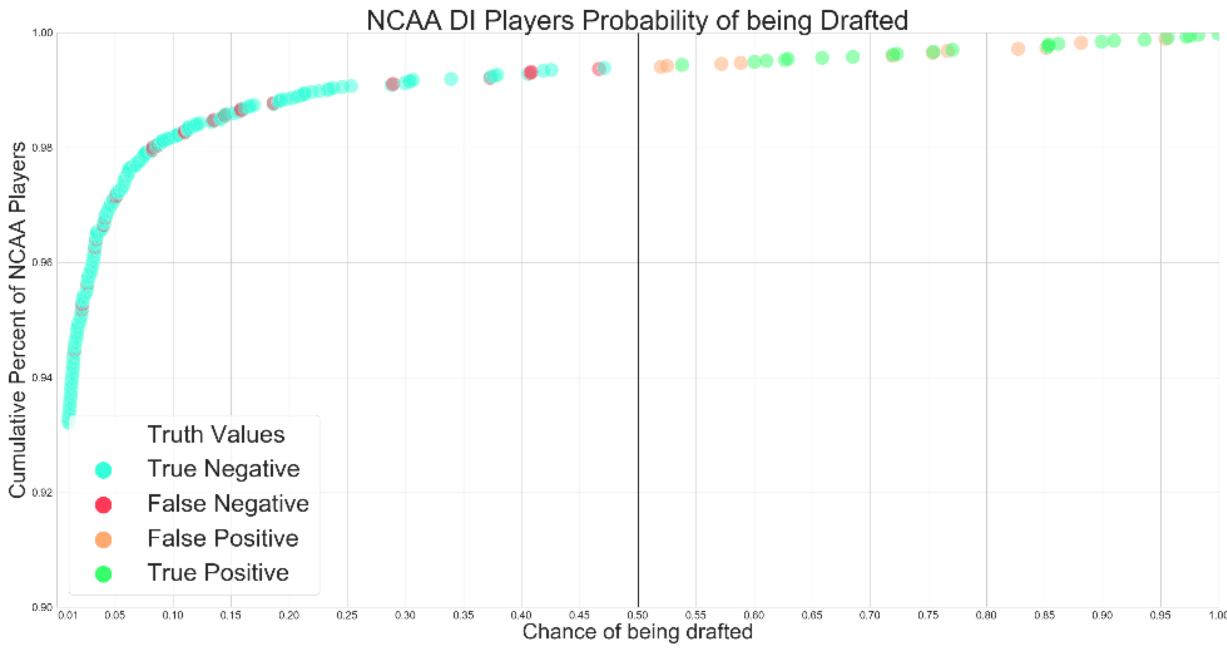


Figure 23: All NCAA seasons wasDrafted breakdown

Most of the players analyzed have an extremely low chance of being drafted, as expected. We found more misses than originally anticipated. From our manual research of the wrongly predicted players, we made an interesting discovery- the model predicted all but one player correctly, with the caveat of them returning to college for another year.

Yellow – Returned to college and was drafted into the NBA
Blue – Plays internationally

Name	Year	Predicted	Actual	Draft Probability	Miss Type
Delon Wright	2013-14	1	0	95.4%	not
Marcus Smart	2012-13	1	0	88.1%	not
Kyle Anderson	2012-13	1	0	85.2%	not
Alec Peters	2015-16	1	0	82.7%	not
Johnathan Motley	2016-17	1	0	76.7%	not
Kris Dunn	2014-15	1	0	75.4%	not
Montrezl Harrell	2013-14	1	0	71.9%	not
Alec Peters	2014-15	1	0	58.9%	not
Daniel Hamilton	2014-15	1	0	57.2%	not
Josh Scott	2015-16	1	0	52.5%	not
Denzel Valentine	2014-15	1	0	52.0%	not
Shane Larkin	2012-13	0	1	46.7%	made
Chris McCullough	2014-15	0	1	40.8%	made
Sviatoslav Mykhailiuk	2017-18	0	1	40.8%	made
Rondae Hollis-Jefferson	2014-15	0	1	37.3%	made
Robbie Hummel	2011-12	0	1	28.9%	made

Bruce Brown	2017-18	0	1	18.7%	made
Khyri Thomas	2017-18	0	1	15.9%	made
Branden Dawson	2014-15	0	1	15.9%	made
Otto Porter	2012-13	0	1	14.5%	made
Devon Hall	2017-18	0	1	13.5%	made
Vince Edwards	2017-18	0	1	11.0%	made
Jevon Carter	2017-18	0	1	11.0%	made
Richaun Holmes	2014-15	0	1	8.7%	made
Hamidou Diallo	2017-18	0	1	8.2%	made
Zach LaVine	2013-14	0	1	5.2%	made
Edmond Sumner	2016-17	0	1	5.1%	made
OG Anunoby	2016-17	0	1	4.2%	made
Deyonta Davis	2015-16	0	1	4.0%	made
Grant Jerrett	2012-13	0	1	3.4%	made
Harry Giles	2016-17	0	1	3.3%	made
Ike Anigbogu	2016-17	0	1	2.6%	made
J.P. Tokoto	2014-15	0	1	2.2%	made
DeAndre Bembry	2015-16	0	1	2.2%	made
Chimezie Metu	2017-18	0	1	2.2%	made
Kevin Hervey	2017-18	0	1	1.5%	made
Sam Dekker	2014-15	0	1	0.6%	made
Jordan Clarkson	2013-14	0	1	0.4%	made
Tyler Harvey	2014-15	0	1	0.2%	made

Figure 24: All NCAA seasons wasDrafted misses

Considering all the players the model predicted to be drafted, only one player was actually not drafted. Because a player must forego their college career when declaring for the NBA Draft, this means that the model was in fact extremely good at detecting draftable players, before they even finished their careers. While the model did miss on a number of solid players, such as Otto Porter and Zach LaVine, we can be confident that when our model does identify a player as draftable, it is likely correct.

One interesting insight from the model is what it views as the most important predictors for either success or failure in the NBA. Below are the top 10 positive and negative coefficients for the statistics used.

Metric	Weight	Metric	Weight
MP	-1.034	Height	1.123
G	-0.8395	WS	0.8628
FT	-0.5135	TOV	0.5359
2P	-0.5004	AST	0.4644
AST%	-0.4954	USG%	0.4493
Grade	-0.3910	2PA	0.4407

TS%	-0.3733	FGA	0.4378
PF	-0.3644	BPM	0.3601
PER	-0.2657	PProd	0.3110
GS	-0.2333	OBPM	0.3055

Figure 25: All NCAA seasons wasDrafted coefficients

For the remainder of the experiments involving this dataset, see Appendix A.

6.2 Using only freshmen year seasons

The second scope of NCAA DI players that we considered was only looking at a player's freshmen year. The rationale behind this decision was that 'one-and-done' players who go to the NBA are likely to show anomalous statistical output, and thus be easily detected by the model. The best target attribute for this model was making the NBA, defined as entering a game and playing at least one second.

```
Logistic Regression
             precision    recall  f1-score   support

     No NBA       1.00      1.00      1.00      2524
   Made NBA       0.71      0.53      0.61        19

avg / total       0.99      0.99      0.99      2543
```

Figure 26: NCAA Freshmen madeNBA metrics

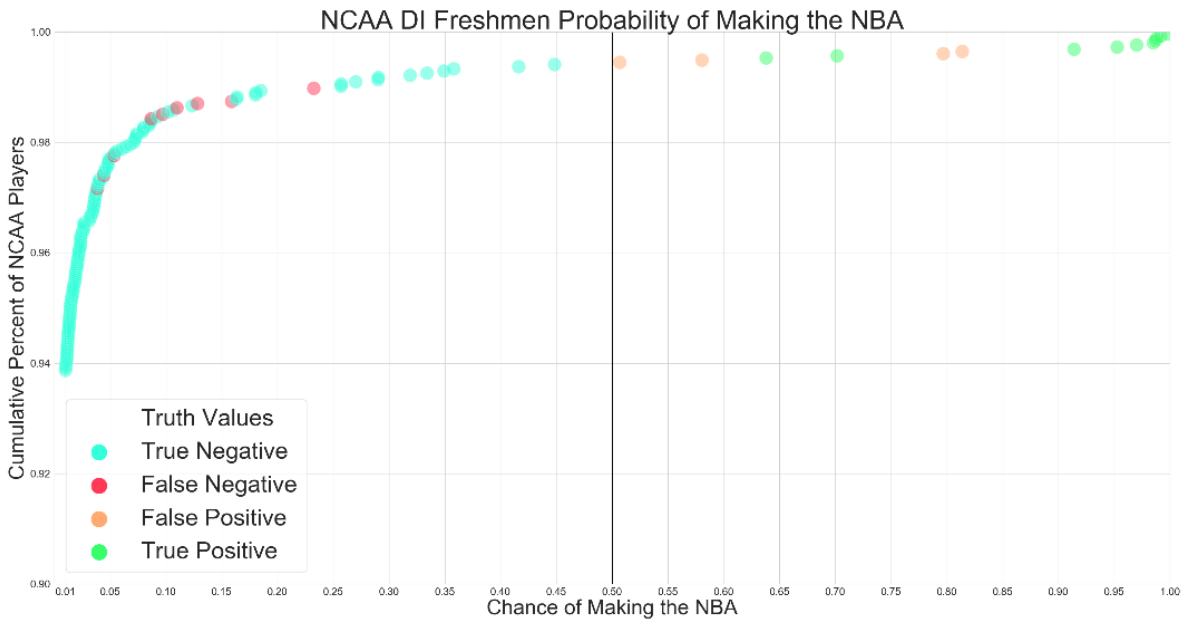


Figure 27: NCAA Freshmen madeNBA breakdown

There are only four false positives, which is promising. All of them are relatively close to the 0.5 region. What we found, however, is three of the four did end up playing in the NBA in later years. Of particular interest is Malcolm Miller, from Holy Cross. The model predicts him to just sneak into the NBA (50.7%) and although he did not declare for the draft until his senior year, the model did in fact identify a player at Holy Cross, not typically a basketball powerhouse, who made the NBA.

Yellow – Did make NBA
Green – Made G League

Name	Year	Predicted	Actual	NBA Probability	Miss Type
Willie Cauley-Stein	2012-13	1	0	81.4%	not
Aaron Harrison	2013-14	1	0	79.7%	not
Melo Trimble	2014-15	1	0	58.1%	not
Malcolm Miller	2013-14	1	0	50.7%	not
Marquis Teague	2011-12	0	1	23.3%	made
Henry Ellenson	2015-16	0	1	15.9%	made
Tyler Ennis	2013-14	0	1	12.8%	made
Chris McCullough	2014-15	0	1	11.0%	made
Justin Patton	2016-17	0	1	9.7%	made
Omari Spellman	2017-18	0	1	8.7%	made
Grant Jerrett	2012-13	0	1	5.3%	made
Deyonta Davis	2015-16	0	1	4.4%	made
Collin Sexton	2017-18	0	1	3.9%	made

Figure 28: NCAA Freshmen madeNBA misses

Collin Sexton is a significant miss, as he played at Alabama and was drafted with the eighth overall pick in the 2018 draft. His lack of size and weight at his position could have contributed to this result.

Metric	Weight	Metric	Weight
MP	-1.152	TOV	0.6605
FT	-0.3357	FTA	0.4857
TRB	-0.2379	WS	0.4516
AST	-0.2200	Height	0.4005
incarnate-word	-0.1936	kentucky	0.2968
FG%	-0.1901	PER	0.2955
mississippi	-0.1836	OWS	0.2905
new-mexico-state	-0.1821	duke	0.2824
north-texas	-0.1783	kansas	0.2775
morgan-state	-0.1708	BLK	0.2722

Figure 29: NCAA Freshmen madeNBA coefficients

6.3 Using only a player's last season

The last scope of NCAA DI players that we considered was only looking at a player's last year they played in college. This scope should eliminate the issue of returning players for all players except players who have not yet left in the 2018-19 season. The best target attribute for this dataset was first round draft picks.

```
Metrics for: firstRound
Logistic Regression
                  precision    recall  f1-score   support

Not First Round        0.99      1.00      1.00      2565
    First Round        0.72      0.45      0.55        29

    avg / total        0.99      0.99      0.99      2594
```

Figure 30: NCAA last season firstRound metrics

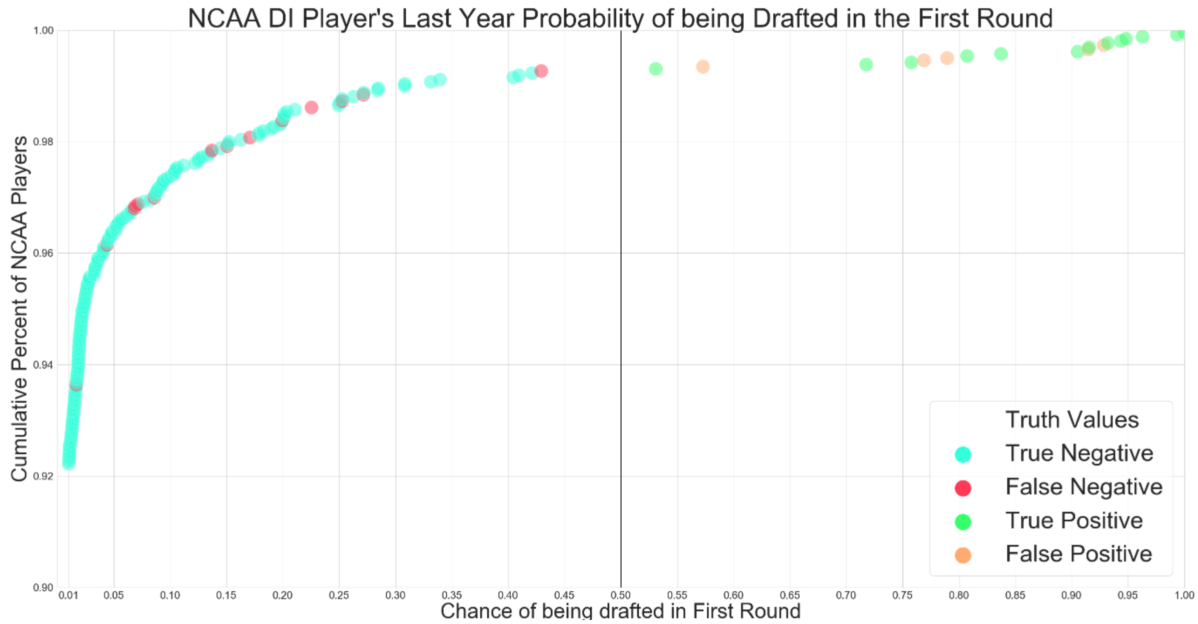


Figure 31: NCAA last season firstRound breakdown

Yellow – Second Round Pick
Green – Undrafted played in G League/ Internationally

Name	Year	Predicted	Actual	1st Round Probability	Miss Type
Davon Usher	2013-14	1	0	92.8%	not
Kenny Kadji	2012-13	1	0	78.9%	not
Cleanthony Early	2013-14	1	0	76.9%	not
Romero Osby	2012-13	1	0	57.3%	not
Josh Hart	2016-17	0	1	42.9%	made
Brice Johnson	2015-16	0	1	27.2%	made
D.J. Wilson	2016-17	0	1	25.3%	made
Donte DiVincenzo	2017-18	0	1	22.5%	made
Terry Rozier	2014-15	0	1	20.0%	made
Marquis Teague	2011-12	0	1	17.1%	made
Justin Patton	2016-17	0	1	15.1%	made
Tony Bradley	2016-17	0	1	13.7%	made
Damian Jones	2015-16	0	1	8.6%	made
Henry Ellenson	2015-16	0	1	7.2%	made
Caris LeVert	2015-16	0	1	6.9%	made
R.J. Hunter	2014-15	0	1	6.8%	made
Josh Okogie	2017-18	0	1	4.4%	made
Chandler Hutchison	2017-18	0	1	1.7%	made
Justin Anderson	2014-15	0	1	0.2%	made

Figure 32: NCAA last season firstRound misses

The model produced four false positives once again, which is not too bad. All four of the false positives are playing professionally at some level, however that is not what the target was predicting.

Metric	Weight	Metric	Weight
MP	-0.5499	OWS	0.6620
BLK%	-0.5127	WS	0.6372
3P	-0.3811	TOV	0.5496
high-point	-0.3048	BLK	0.4379
G	-0.2956	DWS	0.3983
FT%	-0.2919	PProd	0.3709
north-carolina-central	-0.2687	ORB	0.3552
DRB	-0.2686	STL	0.3499
Albany-ny	-0.2644	FTA	0.3425
middle-tennessee	-0.2569	Height	0.3243

Figure 33: NCAA last season firstRound coefficients

6.4 Predicting on the 2019 NCAA DI Players

The below table is our projected ordering of how players will be drafted based on the probability that our model gave them for chances of making the NBA.

Yellow – Was not on the ESPN top 100 best available players

Pick	Team	Year	Player	Prob	Pick	Team	Year	Player	Prob
1	duke	2018-19	Zion Williamson	99.44%	**31**	southern-california	2018-19	Nick Rakocevic	42.77%
2	duke	2018-19	R.J. Barrett	99.09%	**32**	arkansas	2018-19	Daniel Gafford	40.89%
3	murray-state	2018-19	Ja Morant	97.64%	**33**	villanova	2018-19	Eric Paschall	39.61%
4	kentucky	2018-19	PJ Washington	96.08%	**34**	virginia	2018-19	De'Andre Hunter	39.50%
5	furman	2018-19	Matt Rafferty	95.78%	**35**	louisville	2018-19	Jordan Nwora	39.21%
6	oregon	2018-19	Bol Bol	93.29%	**36**	cincinnati	2018-19	Jarron Cumberland	39.10%
7	kansas	2018-19	Dedric Lawson	92.82%	**37**	washington	2018-19	Matisse Thybulle	38.71%
8	tennessee	2018-19	Grant Williams	81.11%	**38**	purdue	2018-19	Carsen Edwards	37.63%
9	michigan-state	2018-19	Cassius Winston	80.80%	**39**	virginia	2018-19	Ty Jerome	37.40%

10	wisconsin	2018-19	Ethan Happ	77.26%	**40**	indiana	2018-19	Romeo Langford	36.63%
11	duke	2018-19	Cam Reddish	74.67%	**41**	marquette	2018-19	Sam Hauser	35.56%
12	kentucky	2018-19	Tyler Herro	73.53%	**42**	louisiana-state	2018-19	Naz Reid	34.06%
13	michigan	2018-19	Jon Teske	69.88%	**43**	florida-state	2018-19	Mfiondu Kabengele	31.53%
14	gonzaga	2018-19	Brandon Clarke	68.86%	**44**	michigan	2018-19	Ignas Brazdeikis	30.48%
15	southern-california	2018-19	Bennie Boatwright	65.88%	**45**	north-carolina	2018-19	Luke Maye	29.06%
16	marquette	2018-19	Markus Howard	65.78%	**46**	north-carolina	2018-19	Coby White	26.61%
17	kentucky	2018-19	Keldon Johnson	65.11%	**47**	syracuse	2018-19	Tyus Battle	26.32%
18	maryland	2018-19	Bruno Fernando	63.45%	**48**	iowa-state	2018-19	Marial Shayok	26.19%
19	indiana	2018-19	Juwan Morgan	63.07%	**49**	michigan-state	2018-19	Xavier Tillman	23.57%
20	north-carolina	2018-19	Cameron Johnson	58.19%	**50**	louisiana-state	2018-19	Skylar Mays	23.27%
21	ucla	2018-19	Moses Brown	51.18%	**51**	bowling-green-state	2018-19	Justin Turner	23.24%
22	ucla	2018-19	Kris Wilkes	51.08%	**52**	syracuse	2018-19	Oshae Brissett	23.18%
23	ucla	2018-19	Jaylen Hands	50.83%	**53**	holy-cross	2018-19	Jehyve Floyd	22.09%
24	kentucky	2018-19	Reid Travis	49.72%	**54**	villanova	2018-19	Phil Booth	21.93%
25	oregon-state	2018-19	Tres Tinkle	49.45%	**55**	georgetown	2018-19	Jessie Govan	21.48%
26	louisiana-state	2018-19	Tremont Waters	47.94%	**56**	louisiana-lafayette	2018-19	Jakeenan Gant	21.33%
27	kansas	2018-19	Udoka Azubuike	46.79%	**57**	kansas	2018-19	Lagerald Vick	19.77%
28	gonzaga	2018-19	Rui Hachimura	45.25%	**58**	syracuse	2018-19	Elijah Hughes	19.01%
29	michigan-state	2018-19	Nick Ward	43.40%	**59**	kentucky	2018-19	EJ Montgomery	18.60%
30	st-johns-ny	2018-19	Shamorie Ponds	43.36%	**60**	vermont	2018-19	Anthony Lamb	18.40%

Figure 34: 2019 NCAA player madeNBA predictions

This order is not what we expect to see in the upcoming NBA draft. The only way this would be the order is if each team picked the most NBA ready player according to our model. Since teams also must pick for positions and needs for their respective teams this approach is flawed in accurately predicting how each pick will go. However, often the top picks transcend need and fit onto rosters as they are star players. In 2019 Zion Williamson and R.J. Barrett are examples of such players as both have been projected top picks since the beginning of the season. Our model does value them as the top two picks with over 99% certainty that they will make the NBA. Of course, machine learning is not necessary to figure out they will be drafted highly but it is good that the model passes the eye test at first glance.

With more investigation and comparing it to the ESPN top 100 players available as of March 14th, 2019 looking at solely if the player we predicted is in this list (not concerned with order) we see that our model had predicted 34 players who appeared on this top 100 best available players. Some of these players were past the 60th best available player but it is worth noting the

ESPN list also includes 20 players who are international and did not attend a college, so our model could not have predicted them. Which means our model predicted 34 out of the top 80 collegiate players according to ESPM. Our model also did better at the earlier and higher skill players as of the top 45 college players on the ESPN site our model predicted 29 of these players correctly, including 17 out of the top 25 players.

The most notable misses are the following players Darius Garland, Jaxson Hayes, Nassir Little, Nickeil Alexander-Walker, KZ Okpala, Kevin Porter, and Talen Horton-Tucker.

The below table is our projected ordering of how players will be drafted based on the probability that our model gave them for chances of being drafted.

Pick	Player	Team	Prob	Pick	Player	Team	Prob
1	R.J. Barrett	duke	99.41%	31	De'Andre Hunter	virginia	38.64%
2	Zion Williamson	duke	98.83%	32	Rui Hachimura	gonzaga	37.44%
3	Ja Morant	murray-state	98.18%	33	Kris Wilkes	ucla	36.96%
4	PJ Washington	kentucky	97.11%	34	Matisse Thybulle	washington	36.28%
5	Bol Bol	oregon	95.25%	35	Marcquise Reed	clemson	31.19%
6	Dedric Lawson	kansas	90.80%	36	Udoka Azubuike	kansas	31.01%
7	Bruno Fernando	maryland	82.49%	37	Eric Carter	delaware	30.80%
8	Tyler Herro	kentucky	79.91%	38	Jalen McDaniels	san-diego-state	30.19%
9	Jarrett Culver	texas-tech	78.92%	39	Justin Turner	bowling-green-state	27.10%
10	Grant Williams	tennessee	75.22%	40	Mfiondu Kabengele	florida-state	26.64%
11	Jason Burnell	jacksonville-state	72.21%	41	Bennie Boatwright	southern-california	26.48%
12	Cam Reddish	duke	70.84%	42	Jaylen Nowell	washington	24.69%
13	Brandon Clarke	gonzaga	69.62%	43	Ignas Brazdeikis	michigan	24.21%
14	Cassius Winston	michigan-state	66.66%	44	Lagerald Vick	kansas	23.96%
15	Keldon Johnson	kentucky	65.84%	45	EJ Montgomery	kentucky	23.78%
16	Markus Howard	marquette	64.17%	46	Ashton Hagans	kentucky	23.70%
17	Cameron Johnson	north-carolina	62.74%	47	Marial Shayok	iowa-state	23.65%
18	Ethan Happ	wisconsin	62.71%	48	Naz Reid	louisiana-state	23.32%
19	Nathan Knight	william-mary	58.66%	49	Coby White	north-carolina	23.16%
20	Matt Rafferty	furman	52.90%	50	Nick Richards	kentucky	22.78%
21	Carsen Edwards	purdue	50.24%	51	Nick Ward	michigan-state	22.63%
22	Reid Travis	kentucky	49.72%	52	Admiral Schofield	tennessee	22.39%
23	Shamorie Ponds	st-johns-ny	47.81%	53	Ky Bowman	boston-college	22.37%
24	Jordan Nwora	louisville	47.58%	54	Luke Maye	north-carolina	22.35%
25	Jon Teske	michigan	45.48%	55	Jarron Cumberland	cincinnati	22.21%
26	Ty Jerome	virginia	44.34%	56	Kyle Guy	virginia	21.37%
27	Moses Brown	ucla	44.08%	57	Oshae Brissett	syracuse	20.83%
28	Tremont Waters	louisiana-state	44.07%	58	Lamine Diane	cal-state-northridge	20.19%
29	Jaylen Hands	ucla	40.36%	59	Sam Hauser	marquette	19.94%
30	Juwan Morgan	indiana	39.95%	60	Nick Rakocevic	southern-california	17.86%

Figure 35: 2019 NCAA players madeNBA probabilities

Similar to the above previous table this ordering is also not indicative of how we think the draft will actually be ordered. But we wanted to test how our model with the target being drafted would fair in predicting the current NCAA players. This model also had 26 players of the top 60 players not be on the ESPN list. And just like the previous table the model fared better at higher skill level players with the same notable misses.

7. Discussion

7.1 Dataset

7.1.1 Levels of Achievement

In order to create our dataset, we had to establish certain criteria for determining if a player made the NBA. The other targets were far more black and white so we did not have to define them. Either they were a first round pick or they were not, lottery pick or not, etc. But for making the NBA we had to define what it meant to make the NBA. Our data was collected from Sports-Reference.com which defined making the NBA as playing an NBA game. It is worth mentioning, however, that every false positive we had the player either returned to college and later made the target or was a later pick than the target (e.g. target was first round and they were a second round pick), played in the G League or Internationally, or returned to college and is playing this year. From this and the fact that we always had more false negatives than false positives we can infer that our model was too tight. But there is no simple way to resolve this issue as the graphs displaying the successes and misses showcase simply lowering the threshold value would not increase the precision and recall of our model. The resolution to this is far more complex and beyond the scope of this project. One potential solution could be creating a more accurate dataset with non-binary labels: for instance, 0 for not playing professionally, 1 for playing abroad, and 2 for playing in the NBA.

A considerable amount also signed NBA contracts they just failed to make the cut when the regular season came around or never saw the court. It is debatable whether these such players who made an NBA roster should be considered having made the NBA. But due to the criteria established by our data source going back and manually editing the data would have been unreasonable.

7.1.2 Returning to College

A further challenge we had to address within our dataset was the players who returned to play college even when they would have made the NBA that year. Players who returned to play in college were unnecessary noise in our dataset and these players did show up as false positive in our predictions. A player like Willie Cauley-Stein in his 2012-13 season decided to go back and play another year at Kentucky. Our model predicted he had a 72% chance of making the NBA, far above the threshold of 50%. And although he later ended up playing in the NBA, his 2012-13 is considered a miss and this adds more complexity to an already complex task.

These players who return to college but were deemed ready for the NBA often are seen across our predictions as misses because they also were likely to be predicted to be drafted, in the first round or a lottery pick. For example, Cody Zeller was “missed” 3 times because our model predicted he would be drafted, be a first round and lottery pick in 2011-12 but since he returned to play another year at college all of these predictions were seen as false positives even though he eventually was a lottery pick. A case could be made that had he declared for the draft in 2011-12 he would have been drafted highly. Overall there are more factors than just if a player would be drafted or play in the NBA as some players choose to stay. These reasons are impossible to account for with the dataset we had access to and will always result in variability for these kinds of predictions.

7.2 Needle in a Haystack

When it comes to predicting how NCAA DI performance will translate into NBA related achievements one is truly trying to find a needle in a haystack. The vast majority of players will never come close to being drafted or playing an NBA game. A model that predicts no one would make the NBA would be correct 99% of the time. But such a model is useless as the only portion people care about is that 1%. Typically, Machine Learning models (especially simple ones) struggle with such a skewed classification problem. It is encouraging, however, that our models were extremely effective, despite their simplicity.

7.3 Coefficients

Below are the three tables with the strong coefficients located above.

All NCAA / Was drafted

Metric	Weight	Metric	Weight
Minutes Played	-1.034	Height	1.123
Games Active	-0.8395	Win Shares	0.8628
Free Throws Made	-0.5135	Turnovers	0.5359
2-pointers Made	-0.5004	Assists	0.4644
Assist Percentage	-0.4954	Usage Rate	0.4493
Year (1 = FR, 4 = SR)	-0.3910	2-point attempts	0.4407
True Shooting %	-0.3733	Field Goal Attempts	0.4378
Personal Fouls	-0.3644	Blocks per minute	0.3601
Player Efficiency Rating	-0.2657	Points Produced	0.3110
Games Started	-0.2333	Offensive Box +/-	0.3055

At a first glance the positive factors make sense, since the NBA is such a large jump from NCAA DI a player’s height is a crucial factor in determining if they could even make the NBA. However, the large factor placed on height may be the reason that a lot of our misses occur. Our models tended to produce false positives on front court players and false negatives on back court players. But this is all part of the imperfection of trying to predict NBA readiness from collegiate data. Some players skill will overcome their physical limitations while some players physical stature is not enough to overcome their lack of skill or seen potential.

Similarly, Minutes Played seems like a strange negative indicator for success. Intuitively we would expect a player to be more successful if they played more minutes, but apparently the opposite is true.

Freshmen / Made NBA

Metric	Weight	Metric	Weight
Minutes Played	-1.152	Turnovers	0.6605
Free Throws Made	-0.3357	Free Throw Attempts	0.4857
Total Rebounds	-0.2379	Win Shares	0.4516
Assists	-0.2200	Height	0.4005
incarnate-word	-0.1936	kentucky	0.2968
Field Goal %	-0.1901	Player Efficiency Rating	0.2955
mississippi	-0.1836	Offensive Win Shares	0.2905
new-mexico-state	-0.1821	duke	0.2824
north-texas	-0.1783	kansas	0.2775
morgan-state	-0.1708	Blocks	0.2722

Interestingly, the strongest positive indicator only has a weight of 0.6605, compared to the previous experiment which weighted Height at 1.123. Free Throws Made and Attempts are both the second strongest predictors, but on opposite sides of the spectrum. Our belief is that the model is simply correcting itself to negate the impact of the features overall.

It also jumps out that the colleges begin to play a stronger role in predictions for freshmen only. The prestige of some of the top basketball schools in the country comes through, with Kentucky, Duke and Kansas all making the top 10 positive predictors.

Last Seasons / First Round Pick

Metric	Weight	Metric	Weight
Minutes Played	-0.5499	Offensive Win Shares	0.6620
Block Percentage	-0.5127	Win Shares	0.6372
3-pointers Made	-0.3811	Turnovers	0.5496
high-point	-0.3048	Blocks	0.4379
Games Active	-0.2956	Defensive Win Shares	0.3983
Free Throw %	-0.2919	Points Produced	0.3709
north-carolina-central	-0.2687	Offensive Rebounds	0.3552
Defensive Rebounds	-0.2686	Steals	0.3499

albany-ny	-0.2644	Free Throw Attempts	0.3425
middle-tennessee	-0.2569	Height	0.3243

For the last seasons dataset, Win Shares is massively important, with the two components of the metric (Offensive WS and Defensive WS) showing up in the strongest positive predictors. Also interesting to note, is that the top basketball schools fall out of the top 10 rankings. Most NBA-quality players from these schools leave after one year, whereas smaller school players typically play all four years, lending credence to this model in particular.

Overall, machine learning models tend to be black-box like. It can be hard to extract meaning from the individual coefficients, even in a simple model like a logistic regression. What is undisputable, however, is that the models overall do an excellent job of predicting NBA success, especially considering the incredibly skewed nature of the dataset. There are certainly many more predicting variables that enter the decision-making process for an actual NBA team when choosing players, such as the player’s mentality, health concerns, and performance in private workouts held just before the draft. While some of these factors are unquantifiable, there are nevertheless improvements that can be made to the data to produce more accurate and usable models.

8. Future Work

The major challenges were presented by the access we had to data sources for all phases of the project. In the predictive component, we faced difficulties with obtaining enough years of data to extract meaningful results, dealing with inaccuracies in the way year of college was modeled, and missing values for physical measurements. Additionally, the need to manually check for players who returned to school or played internationally presented further problems.

Ideally, the dataset would also include international players, and weight their league accordingly. Especially as NBA superstars increasingly hail from outside the US, more importance is being placed on the quality and accuracy of international scouting. With our model able to consistently extract which schools generate more NBA talent, we feel that it would provide similar results if adding European teams into the mix.

One actionable solution to this problem that NBA teams probably have access to is a list of players whom have declared for the NBA Draft for a given year. That way, the model would not concern itself with players who will be returning to school, and likely give better results.

References

Adgate, B. (2018, April 25). *Why the 2017-18 Season Was Great For The NBA*. Retrieved from Forbes: https://www.forbes.com/sites/bradadgate/2018/04/25/the-2017-18-season-was-great-for-the-nba/#5c791e32ecbb

Barzilai, D. A. (2007). *Assessing the Relative Value of Draft Position in the NBA Draft*. Retrieved from 82games.

Greenberg, N. (2017, May 18). *Worst NBA trade ever? 2014 Nets-Celtics trade would have to outdo these four duds.* Retrieved from Washington Post: https://www.washingtonpost.com/news/fancy-stats/wp/2017/05/18/worst-nba-trade-ever-2014-nets-celtics-trade-would-have-to-outdo-these-four-duds/?utm_term=.929a7afb788b

Mertz, J., Hoover, L. D., Burke, J. M., Bellar, D., Jones, M. L., Leitzelar, B., & Judge, W. L. (2016). Ranking the Greatest NBA Players: A Sport Metrics Analysis. *International Journal of Performance Analysis in Sport*, 737-759.

Oliver, D. (2004). *Basketball on Paper.*

Pelton, K. (2015, June 25). *Making smart, valuable trades to move up in the draft is harder than it looks*. Retrieved from ESPN: http://www.espn.com/nba/draft2015/insider/story/_/id/13143349

Pelton, K. (2015, June 25). *Making smart, valuable trades to move up in the draft is harder than it looks*. Retrieved from ESPN: http://www.espn.com/nba/draft2015/insider/story/_/id/13143349

Pelton, K. (2017, June 18). *Trade down or keep No. 1 pick: Which is more valuable?* Retrieved from ESPN: http://www.espn.com/nba/insider/story/_/id/19658707

Ramey, Z. (2018, September 24). *Golden State Warriors: Will the luxury tax bill end the Warriors' dynasty?* Retrieved from Blue Man Hoop: https://bluemanhoop.com/2018/09/24/golden-state-warriors-luxury-bill-end-dynasty/

Reynon, A. (2018, Fall). *NCAA*. Retrieved from The One-and-Done Dilemma: http://www.ncaa.org/static/champion/the-one-and-done-dilemma/

Routley, N. (2019, February 9). *The Data Behind Surging NBA Team Valuations*. Retrieved from Visual Capitalist: https://www.visualcapitalist.com/surging-nba-team-valuations/

Saiidi, U. (2018, November 19). *The NBA is China's most popular sports league. Here's how it happened*. Retrieved from CNBC: https://www.cnbc.com/2018/11/20/the-nba-is-chinas-most-popular-sports-league-heres-how-it-happened.html

Schwarz, A. (2004, July 8). *A numbers revolution*. Retrieved from ESPN: http://www.espn.com/mlb/columns/story?columnist=schwarz_alan&id=1835745

Shot Search. (n.d.). Retrieved from NBA Savant: http://nbasavant.com/shot_search.php

Appendix A: Experiment Results

This appendix contains the remaining results of the experiments we conducted.

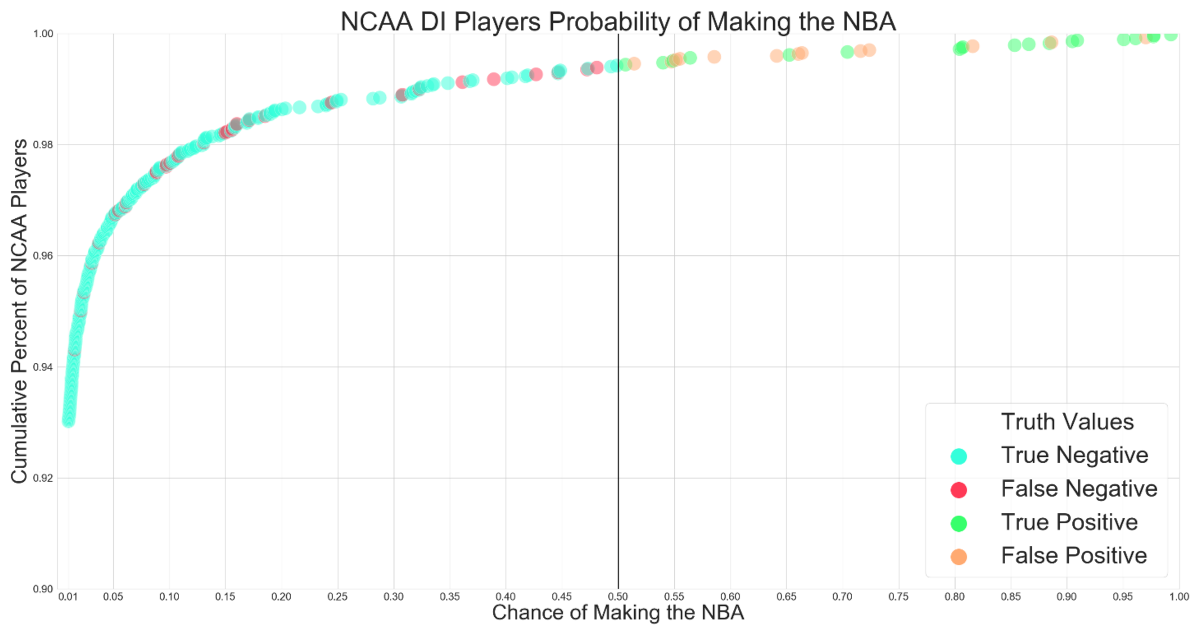
Predicting whether an NCAA DI player will play an NBA game

Our model had a total of 53 Misses where 13 were false positives and 40 were false negatives. The total size of the players was 5712. This means that for players that did not make the NBA we were correct 5630 out of the 5643 times and for players that did make the NBA we were correct 19 out of the 59 times. Below are the corresponding precision, recall, f1-score and support metrics.

The following table shows details about every miss our model had along with what certainty (prob make target) our model predicted this player to achieve the target.
E.g. our model predicted that Grayson Allen had a 97% chance of making the NBA and so it predicted he would make the NBA. Since that year Allen returned to college it was considered a “not” miss type which is a false positive. On the other hand, our model predicted that Dion Waiters had a 44.65% chance of not making the NBA since that is below the threshold of 50% the model predicted he would not make the NBA, but he did end up making the NBA so it is considered a “made” miss type which is a false negative.

Logistic Regression

	precision	recall	f1-score	support
No NBA	0.99	1.00	1.00	5653
Made NBA	0.59	0.32	0.42	59
avg / total	0.99	0.99	0.99	5712



Yellow = Returned to college and went on to play in the NBA
Green = Returned to college and playing this collegiate season
Blue = Returned to college and played in the G League after college

Name	Year	Predicted	Actual	NBA Probability	Miss Type
Grayson Allen	2015-16	1	0	97.0%	not
Kyle Anderson	2012-13	1	0	88.6%	not
Miles Bridges	2016-17	1	0	81.6%	not
Willie Cauley-Stein	2012-13	1	0	72.4%	not
Jontay Porter	2017-18	1	0	71.6%	not
Tyrone Wallace	2014-15	1	0	66.4%	not
Jameel Warney	2014-15	1	0	66.1%	not
Bryce Alford	2014-15	1	0	64.1%	not
Juwan Morgan	2017-18	1	0	58.6%	not
Ivan Rabb	2015-16	1	0	55.5%	not
Kyle Wiltjer	2014-15	1	0	55.1%	not
Bryce Alford	2015-16	1	0	54.8%	not
Antonio Campbell	2015-16	1	0	51.4%	not
Caris LeVert	2015-16	0	1	48.1%	made
Joseph Young	2014-15	0	1	47.2%	made
Dion Waiters	2011-12	0	1	44.6%	made
Dakari Johnson	2014-15	0	1	42.7%	made
Branden Dawson	2014-15	0	1	38.9%	made
Tyler Cavanaugh	2016-17	0	1	36.1%	made
Matt Costello	2015-16	0	1	32.3%	made
Tyler Ennis	2013-14	0	1	30.8%	made

Jerrelle Benimon	2013-14	0	1	18.5%	made
Zach Collins	2016-17	0	1	17.1%	made
Jarrett Allen	2016-17	0	1	16.0%	made
Sterling Brown	2016-17	0	1	15.6%	made
James Michael McAdoo	2013-14	0	1	15.6%	made
Shake Milton	2017-18	0	1	15.1%	made
Damyean Dotson	2016-17	0	1	15.0%	made
Khyri Thomas	2017-18	0	1	13.1%	made
Duncan Robinson	2017-18	0	1	10.9%	made
Zach LaVine	2013-14	0	1	10.8%	made
Travis Wear	2013-14	0	1	10.1%	made
K.J. McDaniels	2013-14	0	1	9.8%	made
Marquese Chriss	2015-16	0	1	9.7%	made
Troy Brown	2017-18	0	1	8.9%	made
Kay Felder	2015-16	0	1	8.8%	made
Jake Layman	2015-16	0	1	7.7%	made
Isaiah Whitehead	2015-16	0	1	6.2%	made
Johnathan Williams	2017-18	0	1	6.1%	made
Diamond Stone	2015-16	0	1	5.5%	made
Fred VanVleet	2015-16	0	1	5.2%	made
Marcus Paige	2015-16	0	1	3.7%	made
Alize Johnson	2017-18	0	1	3.1%	made
Marcus Derrickson	2017-18	0	1	2.4%	made
Shawn Long	2015-16	0	1	2.1%	made
Elfrid Payton	2013-14	0	1	1.6%	made
Ben Bentil	2015-16	0	1	1.0%	made
Kris Dunn	2015-16	0	1	0.9%	made
Wesley Iwundu	2016-17	0	1	0.8%	made
Alan Williams	2014-15	0	1	0.6%	made
Shayne Whittington	2013-14	0	1	0.5%	made

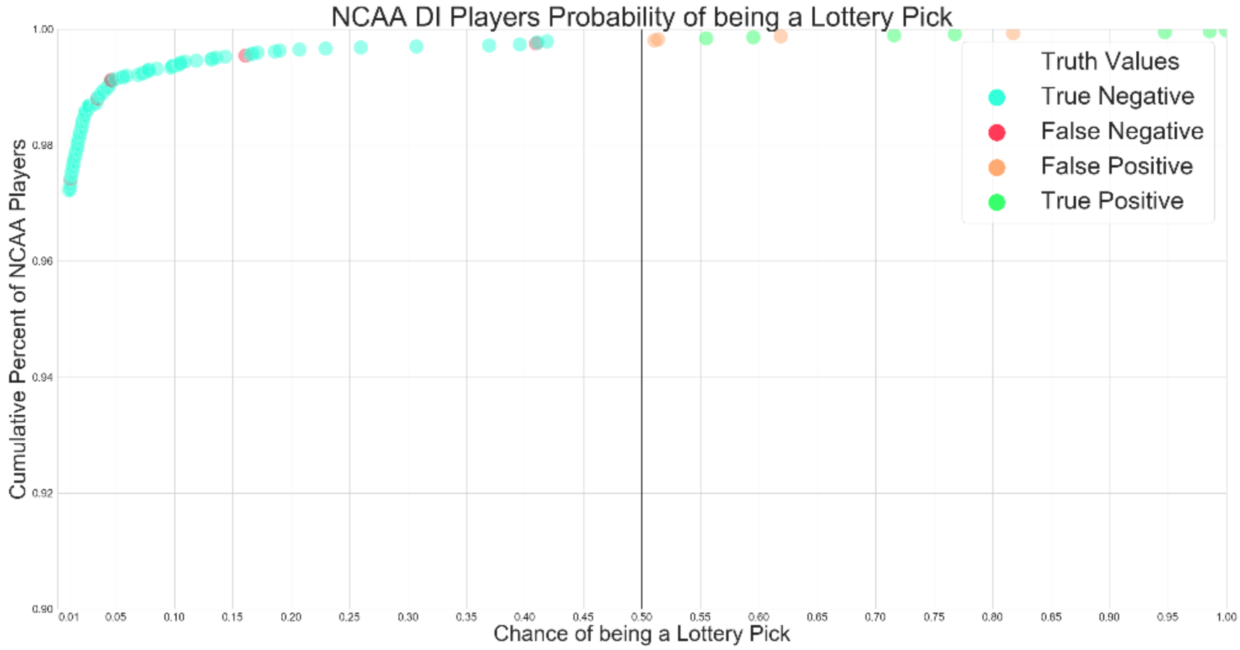
7 ended up playing in the NBA after dataset was collected, 4 are in the G League and the last 2 returned to college expected to be drafted this year (Juwan Morgan, Jontay Porter)

Predicting whether an NCAA DI player will be a lottery pick

```
Logistic Regression
                precision    recall  f1-score   support

Not Lottery          1.00      1.00      1.00      5696
    Lottery          0.64      0.44      0.52        16

avg / total          1.00      1.00      1.00      5712
```



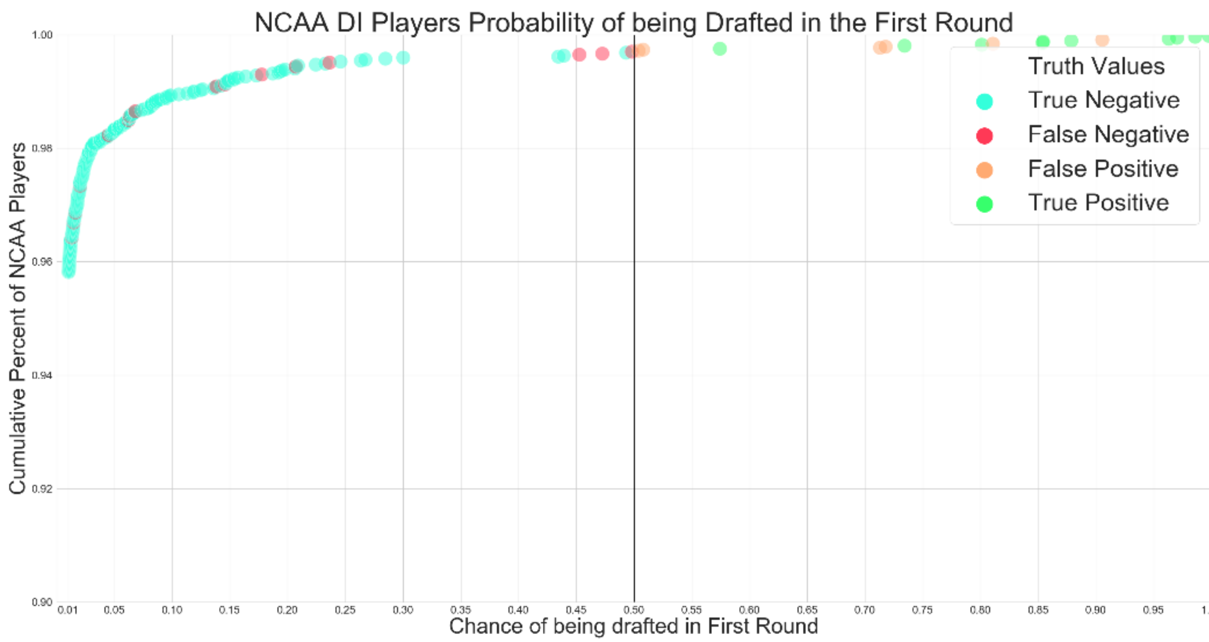
Yellow – Returned and was a lottery pick player
Green – First Round Pick
Blue - Undrafted but played in the NBA

Name	Year	Predicted	Actual	Probability Made	Miss Type
Grayson Allen	2015-16	1	0	81.7%	not
Delon Wright	2013-14	1	0	61.9%	not
Willie Cauley-Stein	2012-13	1	0	51.4%	not
Christian Wood	2014-15	1	0	51.1%	not
Stanley Johnson	2014-15	0	1	40.9%	made
T.J. Warren	2013-14	0	1	16.1%	made
Mikal Bridges	2017-18	0	1	3.4%	made
Bradley Beal	2011-12	0	1	1.1%	made
Donovan Mitchell	2016-17	0	1	0.3%	made
Shabazz Muhammad	2012-13	0	1	0.2%	made
Zach LaVine	2013-14	0	1	0.1%	made
Andre Drummond	2011-12	0	1	0.1%	made

Predicting whether an NCAA DI player will be a first round pick

Logistic Regression

	precision	recall	f1-score	support
Not First Round	1.00	1.00	1.00	5682
First Round	0.62	0.33	0.43	30
avg / total	0.99	1.00	0.99	5712



Yellow – Returned / Was First Round Pick
Green – Undrafted played NBA
Blue – Undrafted played in G League

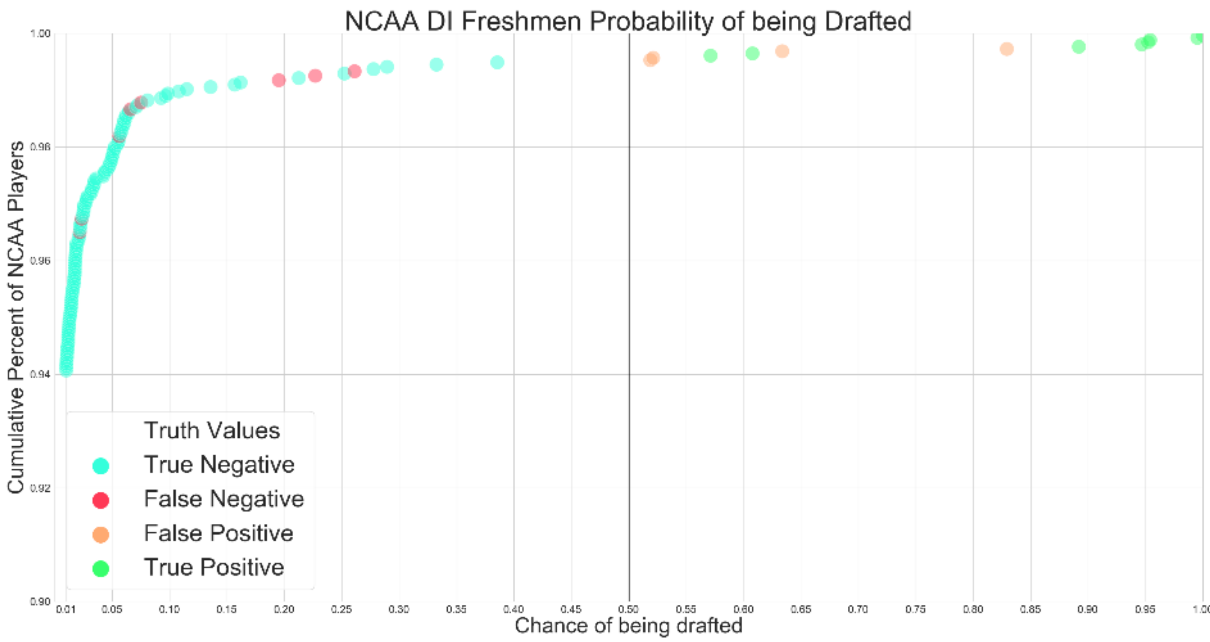
Name	Year	Predicted	Actual	Probability Made	Miss Type
Cody Zeller	2011-12	1	0	90.5%	not
Davon Usher	2013-14	1	0	81.1%	not
Willie Cauley-Stein	2012-13	1	0	71.8%	not
Miles Bridges	2016-17	1	0	71.3%	not
Kris Dunn	2014-15	1	0	50.8%	not
Aaron Harrison	2013-14	1	0	50.4%	not
Dejounte Murray	2015-16	0	1	49.8%	made
Moritz Wagner	2017-18	0	1	47.3%	made
Devin Booker	2014-15	0	1	45.3%	made
Zach Collins	2016-17	0	1	23.6%	made
Jarrett Allen	2016-17	0	1	20.7%	made
Otto Porter	2012-13	0	1	17.8%	made
Caleb Swanigan	2016-17	0	1	14.6%	made
Myles Turner	2014-15	0	1	13.8%	made
OG Anunoby	2016-17	0	1	6.8%	made
Arnett Moultrie	2011-12	0	1	6.4%	made
Henry Ellenson	2015-16	0	1	6.2%	made
Skal Labissiere	2015-16	0	1	4.5%	made
Bradley Beal	2011-12	0	1	2.0%	made
Chandler Hutchison	2017-18	0	1	1.6%	made
Anthony Bennett	2012-13	0	1	1.5%	made
Sam Dekker	2014-15	0	1	1.3%	made
Jerome Robinson	2017-18	0	1	0.8%	made
DeAndre Bembry	2015-16	0	1	0.7%	made
Landry Shamet	2017-18	0	1	0.1%	made

Predicting whether an NCAA DI freshmen will be drafted

```
Logistic Regression
                 precision    recall  f1-score   support

Not Drafted           1.00      1.00      1.00      2523
    Drafted           0.67      0.40      0.50        20

avg / total           0.99      0.99      0.99      2543
```



Yellow – Returned to College was Drafted
Green – Undrafted Played in NBA
Blue – Returned to school, playing this year

Name	Year	Predicted	Actual	Draft Probability	Miss Type
Aaron Harrison	2013-14	1	0	82.9%	not
Luke Kennard	2015-16	1	0	63.4%	not
Trevon Duval	2017-18	1	0	52.1%	not
Jarrett Culver	2017-18	1	0	51.9%	not
Malik Beasley	2015-16	0	1	26.1%	made
Chris McCullough	2014-15	0	1	22.7%	made
Jarrett Allen	2016-17	0	1	19.5%	made
Zach LaVine	2013-14	0	1	7.5%	made
Justin Jackson	2016-17	0	1	5.6%	made
Arnett Moultrie	2011-12	0	1	2.3%	made

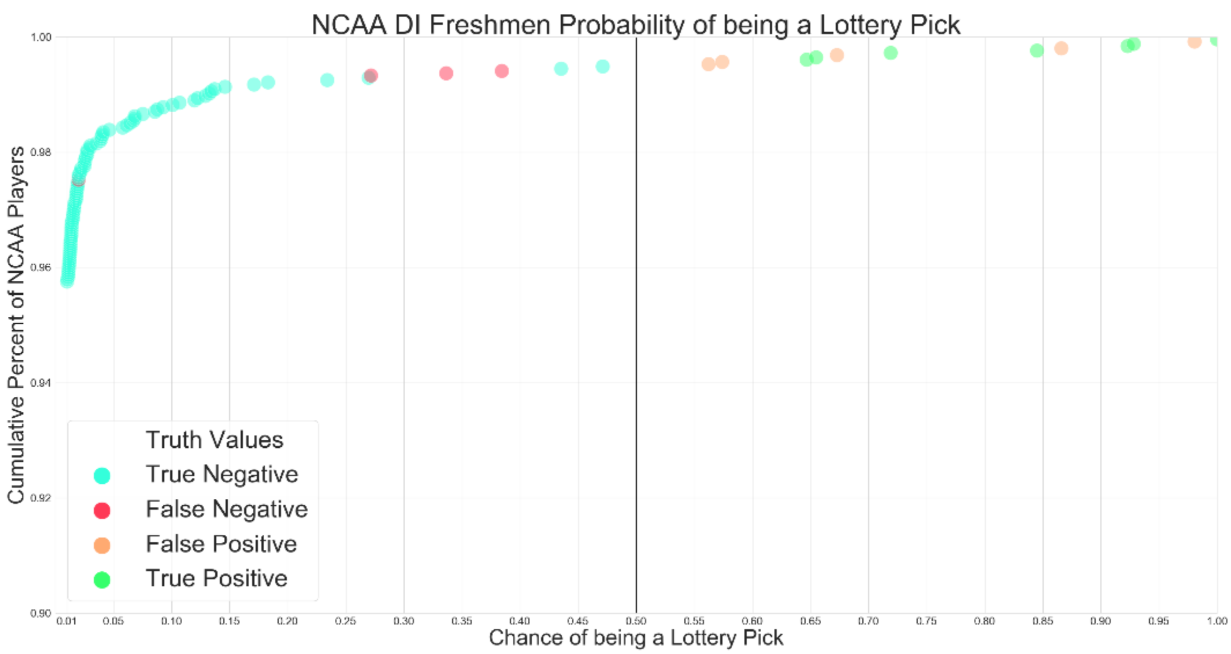
Royce White	2011-12	0	1	2.2%	made
Troy Brown	2017-18	0	1	0.3%	made
Deyonta Davis	2015-16	0	1	0.2%	made
Lonnie Walker	2017-18	0	1	0.2%	made

Predicting whether an NCAA DI freshmen will be a lottery pick

```
Logistic Regression
             precision    recall  f1-score   support

Not Lottery       1.00      1.00      1.00      2532
    Lottery       0.58      0.64      0.61        11

avg / total       1.00      1.00      1.00      2543
```



Yellow- Returned to college was Lottery Pick
Green – First Round Pick
Purple – Second Round Pick
Blue – Undrafted

Name	Year	Predicted	Actual	Lottery Probability	Type
Cody Zeller	2011-12	1	0	98.1%	not
Jordan Adams	2012-13	1	0	86.6%	not
Aaron Harrison	2013-14	1	0	67.3%	not
Willie Cauley-Stein	2012-13	1	0	57.4%	not
Thomas Bryant	2015-16	1	0	56.2%	not
Lauri Markkanen	2016-17	0	1	38.4%	made

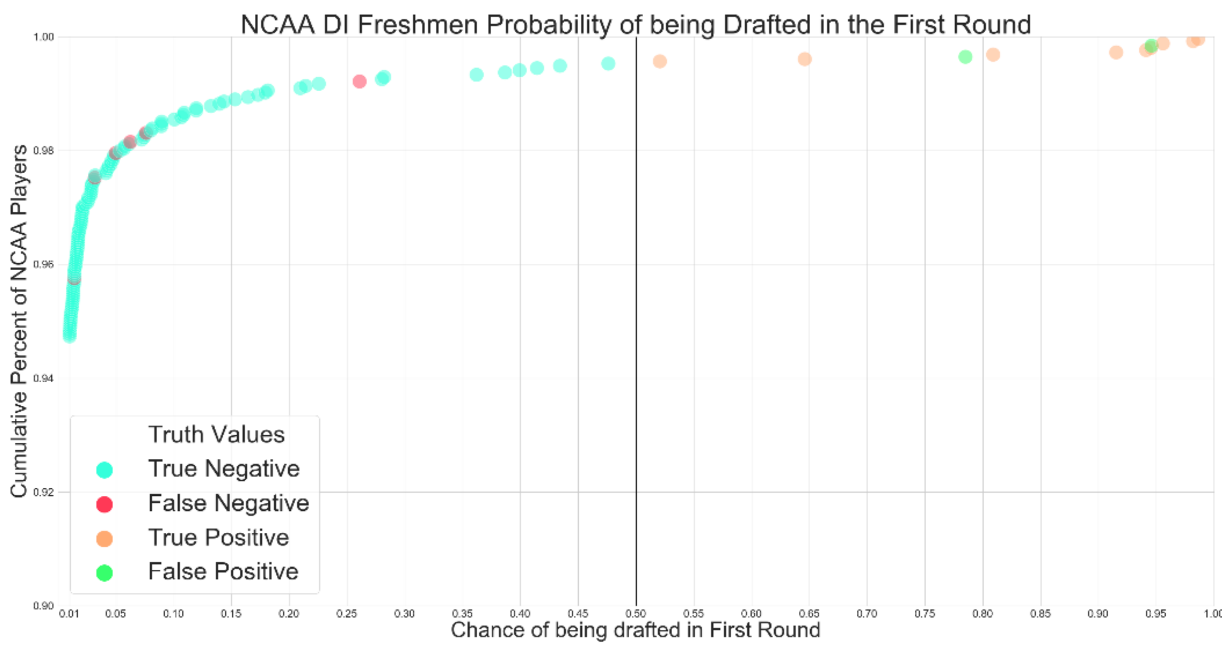
Aaron Gordon	2013-14	0	1	27.2%	made
Ben Simmons	2015-16	0	1	2.0%	made

Predicting whether an NCAA DI freshmen will be a first round pick

```
Logistic Regression
                  precision    recall  f1-score   support

Not First Round       1.00      1.00      1.00      2526
    First Round       0.82      0.53      0.64        17

    avg / total       1.00      1.00      1.00      2543
```



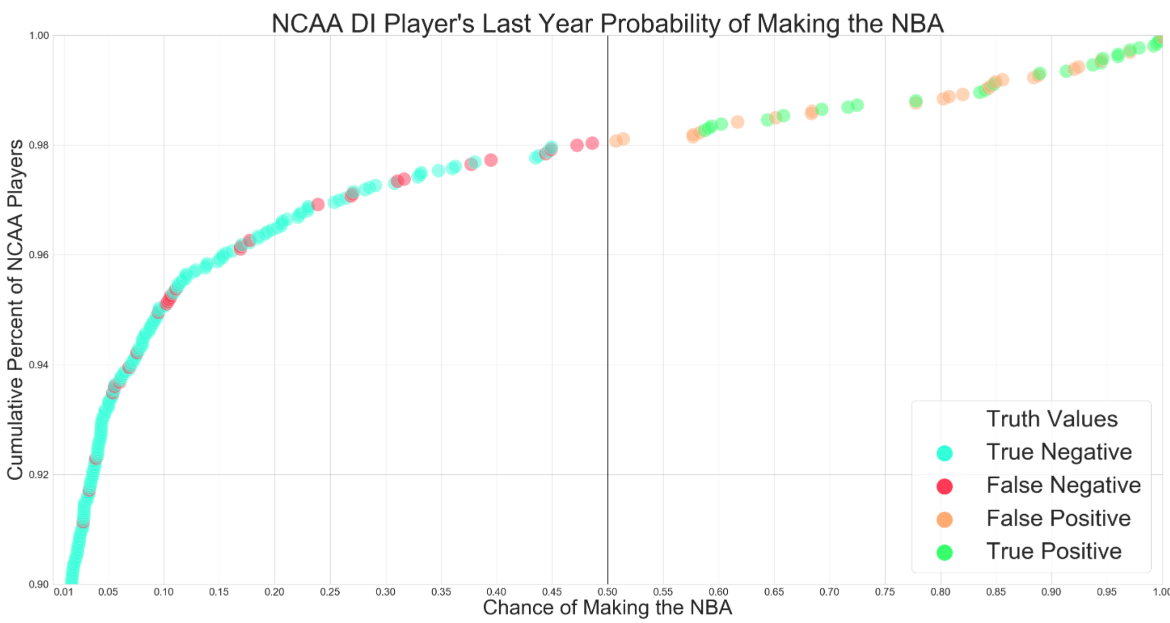
Yellow – Returned to College was First Round
Green - Undrafted

Name	Year	Predicted	Actual	1st Round Prob.	Miss Type
Cody Zeller	2011-12	1	0	94.6%	not
Aaron Harrison	2013-14	1	0	78.5%	not
Austin Rivers	2011-12	0	1	26.1%	made
Omari Spellman	2017-18	0	1	7.6%	made
Zach Collins	2016-17	0	1	5.0%	made
Collin Sexton	2017-18	0	1	3.2%	made
Harry Giles	2016-17	0	1	1.4%	made
Jaylen Brown	2015-16	0	1	0.4%	made
Troy Brown	2017-18	0	1	0.2%	made

Predicting whether an NCAA DI player will play an NBA game

Logistic Regression

	precision	recall	f1-score	support
No NBA	0.99	0.99	0.99	2535
Made NBA	0.52	0.44	0.48	59
avg / total	0.98	0.98	0.98	2594



Yellow – Returned to college made NBA
Green – G League/ International
Blue – Returned to college, playing this year
Purple - Injury

Name	Year	Predicted	Actual	NBA Probability	Miss Type
Brandon McCoy	2017-18	1	0	97.1%	not
Ethan Happ	2017-18	1	0	94.5%	not
Tyus Battle	2017-18	1	0	92.4%	not
Luke Kornet	2016-17	1	0	92.0%	not
Johnathan Motley	2016-17	1	0	88.8%	not
Josh Scott	2015-16	1	0	88.4%	not
Kennedy Meeks	2016-17	1	0	85.6%	not
Kris Wilkes	2017-18	1	0	85.0%	not
Isaiah Austin	2013-14	1	0	84.5%	not
Will Clyburn	2012-13	1	0	84.3%	not
Michael Young	2016-17	1	0	82.0%	not
Perry Ellis	2015-16	1	0	80.8%	not
Eric Mika	2016-17	1	0	80.2%	not
Luke Maye	2017-18	1	0	77.8%	not
Justin Jackson	2013-14	1	0	68.4%	not
Jalen Jones	2015-16	1	0	68.4%	not
Trevon Bluiett	2017-18	1	0	65.1%	not
Ryan Anderson	2015-16	1	0	61.7%	not
Chris Jones	2011-12	1	0	58.3%	not
Jalen Jones	2014-15	1	0	57.7%	not
Brandon Ashley	2014-15	1	0	57.7%	not
Shamorie Ponds	2017-18	1	0	51.4%	not
Cameron Lard	2017-18	1	0	50.8%	not
Troy Williams	2015-16	0	1	48.6%	made
Zach LaVine	2013-14	0	1	47.2%	made
Ike Anigbogu	2016-17	0	1	44.9%	made
Ryan Arcidiacono	2015-16	0	1	44.4%	made
Terry Rozier	2014-15	0	1	39.5%	made
Montrezl Harrell	2014-15	0	1	37.7%	made
Andre Dawkins	2013-14	0	1	31.7%	made
OG Anunoby	2016-17	0	1	31.1%	made
Bryce Dejean-Jones	2014-15	0	1	27.0%	made
Buddy Hield	2015-16	0	1	26.9%	made
Domantas Sabonis	2015-16	0	1	23.9%	made
Raymond Spalding	2017-18	0	1	17.7%	made
Dorian Finney-Smith	2015-16	0	1	16.9%	made
Arnett Moultrie	2011-12	0	1	16.9%	made
Melvin Frazier	2017-18	0	1	11.1%	made

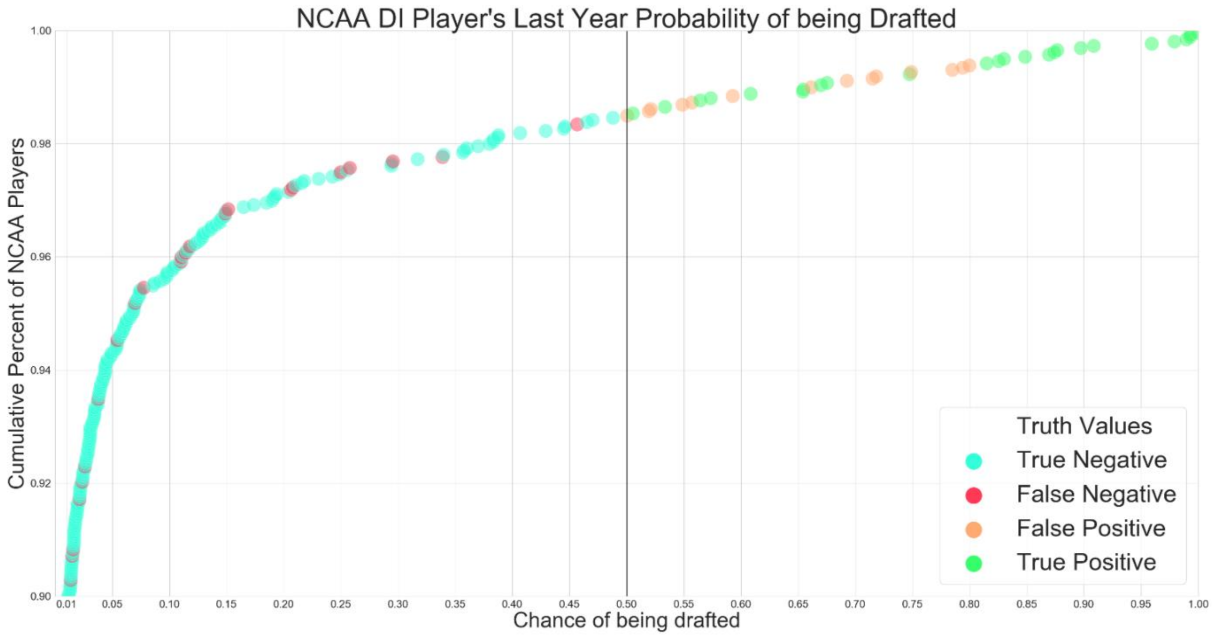
Maurice Ndour	2014-15	0	1	10.7%	made
Marquese Chriss	2015-16	0	1	10.6%	made
Landry Shamet	2017-18	0	1	10.5%	made
Zach Collins	2016-17	0	1	10.4%	made
Pat Connaughton	2014-15	0	1	10.2%	made
Jabari Bird	2016-17	0	1	9.5%	made
John Collins	2016-17	0	1	7.6%	made
Fred VanVleet	2015-16	0	1	6.9%	made
Alize Johnson	2017-18	0	1	6.0%	made
Jarnell Stokes	2013-14	0	1	5.6%	made
DeAndre Bembry	2015-16	0	1	5.4%	made
Malcolm Miller	2014-15	0	1	3.9%	made
Justin Patton	2016-17	0	1	3.3%	made
Larry Nance	2014-15	0	1	2.7%	made
Khyri Thomas	2017-18	0	1	1.3%	made
Johnathan Williams	2017-18	0	1	0.5%	made
Xavier Munford	2013-14	0	1	0.2%	made

Predicting whether an NCAA DI player will be drafted

```
Metrics for: wasDrafted
Logistic Regression
                precision    recall  f1-score   support

Not Drafted       0.99      0.99      0.99      2544
    Drafted       0.64      0.50      0.56        50

avg / total       0.98      0.98      0.98      2594
```



Yellow – Returned to college later drafted
Green – Undrafted played in NBA
Blue – Undrafted played in G League/ Internationally
Purple – Returned to college playing this year or injured

Name	Year	Predicted	Actual	Draft Probability	Miss Type
Kenny Kadji	2012-13	1	0	80.0%	not
Bonzie Colson	2017-18	1	0	79.4%	not
Nathan Knight	2017-18	1	0	78.5%	not
P.J. Hairston	2012-13	1	0	74.9%	not
Derrick Walton	2016-17	1	0	71.5%	not
Isaiah Austin	2013-14	1	0	69.3%	not
Daniel Ochefu	2015-16	1	0	66.1%	not
Kyle Wiltjer	2015-16	1	0	59.3%	not
Justin Pierce	2017-18	1	0	55.7%	not
Shawn Long	2015-16	1	0	54.9%	not
Gary Clark	2017-18	1	0	52.1%	not
Justin Simon	2017-18	1	0	51.9%	not
Zach Auguste	2015-16	1	0	50.0%	not
Marcus Smart	2013-14	0	1	45.7%	made
Monte Morris	2016-17	0	1	33.9%	made
Malik Beasley	2015-16	0	1	29.5%	made
Khyri Thomas	2017-18	0	1	25.8%	made
Jerami Grant	2013-14	0	1	25.0%	made
Lonnie Walker	2017-18	0	1	20.8%	made
Keita Bates-Diop	2017-18	0	1	20.6%	made

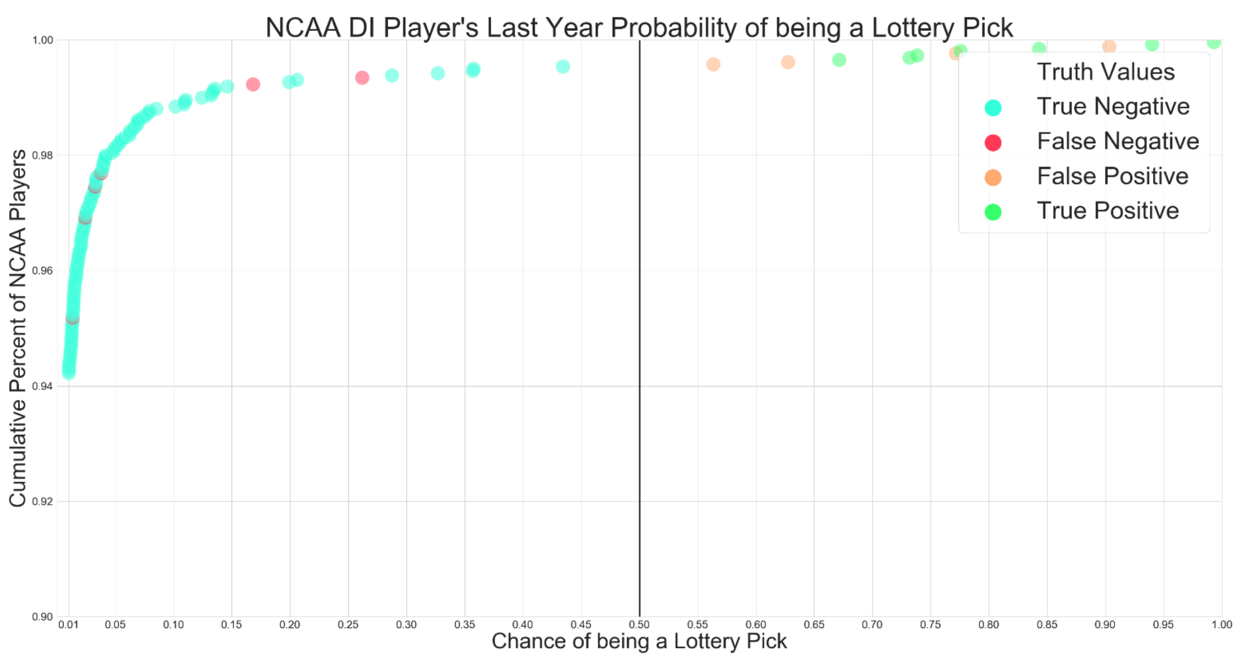
Grant Jerrett	2012-13	0	1	15.1%	made
Tony Bradley	2016-17	0	1	14.9%	made
Tony Carr	2017-18	0	1	11.8%	made
Josh Okogie	2017-18	0	1	11.4%	made
Melvin Frazier	2017-18	0	1	11.0%	made
Chandler Hutchison	2017-18	0	1	11.0%	made
Landry Shamet	2017-18	0	1	7.7%	made
Harry Giles	2016-17	0	1	7.0%	made
Sam Dekker	2014-15	0	1	5.4%	made
Jawun Evans	2016-17	0	1	3.8%	made
Jerome Robinson	2017-18	0	1	2.6%	made
Chimezie Metu	2017-18	0	1	2.4%	made
Jaron Blossomgame	2016-17	0	1	2.1%	made
Robert Williams	2017-18	0	1	1.6%	made
Olivier Hanlan	2014-15	0	1	1.5%	made
Colton Iverson	2012-13	0	1	1.4%	made

Predicting whether an NCAA DI player will be a lottery pick

```
Metrics for: lotteryPick
Logistic Regression
                precision    recall  f1-score   support

Not Lottery          1.00      1.00      1.00      2579
    Lottery          0.64      0.47      0.54        15

avg / total          0.99      1.00      1.00      2594
```



Yellow – Returned was later Lottery Pick
Green – First Round pick
Blue – Second Round Pick
Purple - Returned to college, playing this year

Name	Year	Predicted	Actual	Lottery Probability	Type
Delon Wright	2014-15	1	0	90.3%	not
L.J. Thorpe	2017-18	1	0	77.2%	not
Jordan Adams	2013-14	1	0	62.8%	not
Ivan Rabb	2016-17	1	0	56.3%	not
Mohamed Bamba	2017-18	0	1	26.2%	made
Myles Turner	2014-15	0	1	16.8%	made
Mikal Bridges	2017-18	0	1	3.8%	made
Elfrid Payton	2013-14	0	1	3.3%	made
Anthony Bennett	2012-13	0	1	2.4%	made
Andre Drummond	2011-12	0	1	1.3%	made
Alex Len	2012-13	0	1	0.7%	made