Generating Relative Draft 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

# Abstract

# Acknowledgements

# Table of Contents

[Abstract 2](#_Toc950996)

[Acknowledgements 2](#_Toc950997)

[Table of Contents 2](#_Toc950998)

[Table of Figures 4](#_Toc950999)

[Executive Summary (*Jake, Later)* 4](#_Toc951000)

[1. Introduction *(Later)* 4](#_Toc951001)

[2. Background (*Mike & Jake)* 4](#_Toc951002)

[3. Methodology (*Jake)* 4](#_Toc951003)

[3.1 Analyze existing basketball player performance metrics 4](#_Toc951004)

[3.2 Feature engineer new player performance metrics addressing shortcomings with existing metrics 4](#_Toc951005)

[3.3 Find the highest value picks based on various measures of cost 4](#_Toc951006)

[3.4 Calculate the approximate value of every pick in the NBA Draft 5](#_Toc951007)

[3.5 Create a Jimmy Johnson-style NBA Draft value chart 5](#_Toc951008)

[4. Design (*Mike)* 5](#_Toc951009)

[4.1 Determining Scope of the Project 5](#_Toc951010)

[4.2 Collection and Manipulation of the Data 5](#_Toc951011)

[5. Results (*Jake)* 6](#_Toc951012)

[5.1 Analyze existing basketball player performance metrics 6](#_Toc951013)

[5.2 Feature engineer new player performance metrics addressing shortcomings with existing metrics 7](#_Toc951014)

[5.3 Calculate the approximate value of every pick in the NBA Draft 7](#_Toc951015)

[5.4 Find the highest value picks based on various measure of cost 10](#_Toc951016)

[5.5 Create a Jimmy Johnson-style NBA Draft pick value chart 11](#_Toc951017)

[6. Methodology for NCAA (*Jake)* 13](#_Toc951018)

[6.1 Create a model which predicts various measures of NBA success based on NCAA DI statistics 13](#_Toc951019)

[7. Design for NCAA (*Jake)* 13](#_Toc951020)

[7.1 Create a model which predicts various measures of NBA success based on NCAA DI statistics 13](#_Toc951021)

[8. Results for NCAA (*Mike)* 15](#_Toc951022)

[8.1 Using all years of NCAA DI players 15](#_Toc951023)

[8.1.1 Predicting whether an NCAA DI player will play an NBA game 15](#_Toc951024)

[8.1.2 Predicting which 2018 NCAA DI players would play an NBA game 16](#_Toc951025)

[8.1.3 Predicting whether an NCAA DI player will be drafted 16](#_Toc951026)

[8.1.4 Predicting whether an NCAA DI player will be a lottery pick 16](#_Toc951027)

[8.1.5 Predicting whether an NCAA DI player will be a first round pick 17](#_Toc951028)

[8.1.6 Predicting whether an NCAA DI player will be a second round pick 17](#_Toc951029)

[8.2 Using only freshmen year seasons 17](#_Toc951030)

[8.2.1 Predicting whether an NCAA DI freshmen will play an NBA game 18](#_Toc951031)

[8.2.2 Predicting which 2018 NCAA DI freshmen would play an NBA game 18](#_Toc951032)

[8.2.3 Predicting whether an NCAA DI freshmen will be drafted 18](#_Toc951033)

[8.2.4 Predicting whether an NCAA DI freshmen will be a lottery pick 19](#_Toc951034)

[8.2.5 Predicting whether an NCAA DI freshmen will be a first round pick 19](#_Toc951035)

[8.2.6 Predicting whether an NCAA DI freshmen will be a second round pick 20](#_Toc951036)

[8.3 Using only a player’s last season 20](#_Toc951037)

[8.3.1 Predicting whether an NCAA DI player will play an NBA game 20](#_Toc951038)

[8.3.2 Predicting which 2018 NCAA DI players would play an NBA game 20](#_Toc951039)

[8.3.3 Predicting whether an NCAA DI player will be drafted 21](#_Toc951040)

[8.3.4 Predicting whether an NCAA DI player will be a lottery pick 21](#_Toc951041)

[8.3.5 Predicting whether an NCAA DI player will be a first round pick 22](#_Toc951042)

[8.3.6 Predicting whether an NCAA DI player will be a second round pick 22](#_Toc951043)

[9. Discussion *(Mike)* 22](#_Toc951044)

[9.1 All Division I Players 23](#_Toc951045)

[9.2 Freshmen Division I Players 23](#_Toc951046)

[9.3 Last Year Division I Players 23](#_Toc951047)

[9.4 Dataset Imperfect 23](#_Toc951048)

[9.4.1 NBA / International / G League 23](#_Toc951049)

[9.4.2 Returning to College 23](#_Toc951050)

[9.5 Needle in a Haystack 23](#_Toc951051)

[10. Future Work *(Mike)* 23](#_Toc951052)

[10.1 Predicting which current (2019) NCAA DI players will play an NBA game 23](#_Toc951053)

[10.2 Comparing Draft Value Across Professional Sports 23](#_Toc951054)

[10.3 Considerations for How to Improve Project? 23](#_Toc951055)

# Table of Figures

# Executive Summary (*Jake, Later)*

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 which 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 be drafted or play 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 very 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 which affected 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.

# 1. Introduction *(Later)*

# 2. Background (*Mike & Jake, Later)*

<http://www.nba.com/analysis/rules_history.html>

# 3. Methodology (*Jake)*

## 3.1 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).

## 3.2 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).

## 3.3 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.

## 3.4 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.

## 3.5 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.

# 4. Design (*Mike)*

## 4.1 Determining Scope of the Project

The NBA has had extensive changes to its in game 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 take careful consideration with which time period of the NBA we would include in our dataset. The question being “how far back do you look?” The time period that was decided on was 1990-present day for the following reasons. The first reason is that all of the major modern rules were implemented in a similar way to how they are now. One exception being the three point line was shortened from 23 feet 9 inches uniformly to 22 feet in 1995 then extended only at the top of the key (corner remained at 22 feet) to 23 feet 9 inches. And the other exception being that the shot clock reset on a hit to the backboard up until 1992. The second reason is because we wanted to capture the Jordan years of the NBA. Although not a definitive time period, the NBA in the 90’s was changing from physical play (as demonstrated by the Detroit Piston “Bad Boys”) to a more offensive and point producing league. Within the 90’s there were many rule changes designed to aid offensive players. The last reason is because by starting at 1990 we would have a dataset containing the three “decades” of basketball, the 90’s (1990-00), the 2000’s (00-10), and this decade (10-present) and this provides an easy timeline to understand for the reader.

## 4.2 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. The website that we obtained the information from is Basketball-Reference.com which had all of the data that we thought we would need to collect. To produce our dataset we first iterated through each season and then for each season pulled the 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 programmed algorithms that would cumulatively combine the seasons of data so that in the end we had 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 of the players who were selected to the all star game and those were given an award at the end of the year. These awards were transformed into their own respective column where a 1 indicated they achieved that award and a 0 meant they did not.

# 5. Results (*Jake)*

## 5.1 Analyze existing basketball player performance metrics

As discussed in the methods section, a crucial decision in evaluating player value is how ‘performance’ is quantified. The below table 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 |

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.



To investigate just what these statistical disparities might be, we broke down each metric to its mathematical formula, to see their components.

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.

## 5.2 Feature engineer new player performance metrics addressing shortcomings with existing metrics

CIA BP AP

## 5.3 Calculate the approximate value of every pick in the NBA Draft

First, we summed up the total value of each metric of each draft pick.

This graph is oversensitive to extremely good players who come up at particular positions, which makes the graph jagged. 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

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.

## 5.4 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.

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.

## 5.5 Create a Jimmy Johnson-style NBA Draft pick value chart

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.

|  |  |
| --- | --- |
| Metric | Mean Abs Error |
| VORP | 0.045443858 |
| WS | 0.070661068 |
| FP | 0.081395181 |
| RS | 0.096852287 |
| AVG | 0.112119185 |
| PER | 0.149346209 |
| BP | 0.167310009 |
| AP | 0.198652996 |

Clearly, VORP is the most accurate chart.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DraftPos | VORP | DraftPos | VORP | DraftPos | VORP | DraftPos | 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 |

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

For the first 20 picks, NBA talent is relatively better than the same draft pick in the NFL. However, after that, the NBA talent continues to decline quickly while the NFL flatlines.

# 6. Methodology for NCAA (*Jake)*

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

We designed a series of experiments to examine the ability of machine learning models to predict various success criteria in the NBA. These criteria are as follows:

- Was drafted

- Made NBA (played a regular-season game)

- First round pick

- Lottery pick

All these target values are binary, so this is a classification problem. The models we experimented with were:

- Logistic Regression

- Decision Tree

- Random Forest

- Multi-layer Perceptron (Neural Networks)

# 7. Design for NCAA (*Jake)*

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

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 al NCAA Division I players. Using the same methodology to pull data from Basketball-Refernce.com we were able to pull the 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 need to lack of identifiers 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 where 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 since predicting NBA readiness and expected performance it is also necessary to consider physical attributes. Thus, we also made sure to collect height and weight measurements for all NCAA players. To further investigate how physical attributes play a role into the probability a college player will reach the NBA we also collected data from the NBA combines from 2010-2018.

After collecting all the data that we needed to we integrated 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.

A close up of a map

Description generated with high confidenceOnce 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.

In an attempt 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 to test our model with.

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. 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, an 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 junior15

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 players best season or their

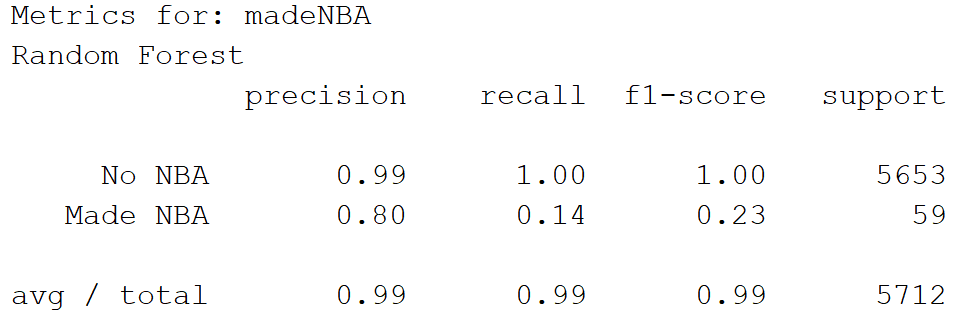
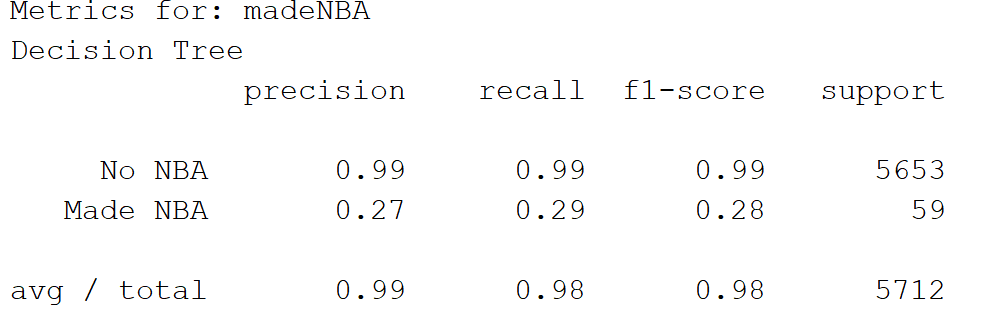
We also ran the model on all of the players in our dataset.

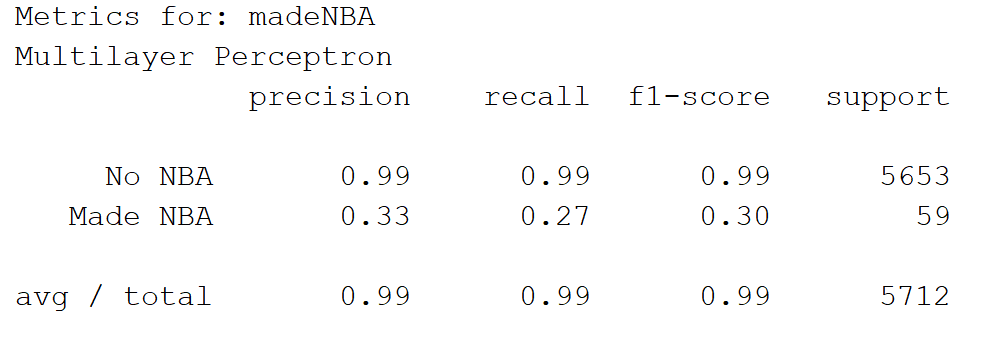
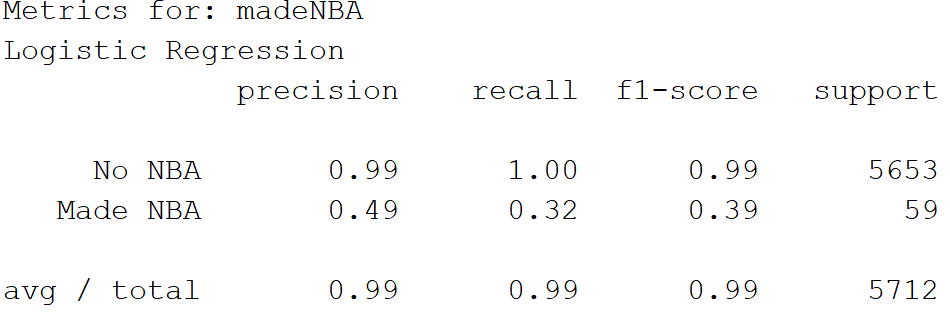
Note: due to an overwhelming amount of empty data points in the combine anthropology and agility datasets we didn’t these metrics. This is due to three main issues, the first is that players who attend the combine rarely perform all the tests, the second is often times the most notable college players 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.

# 8. Results for NCAA (*Mike)*

## 8.1 Machine Learning Models

To generate the most meaningful conclusions and create the best predicting model for NCAA DI players we tested out multiple machine learning models on our data. We evaluated how good a metric was by its ability to predict the made NBA with all of our dataset. We considered the best model to be the model with the highest f1 score. Below are the statistics for each model.





We ran the above test with multiple seeds and each time the logistic regression proved to be our best model. With this in mind we only used the logistic regression when analyzing different scopes and target we were trying to predict.

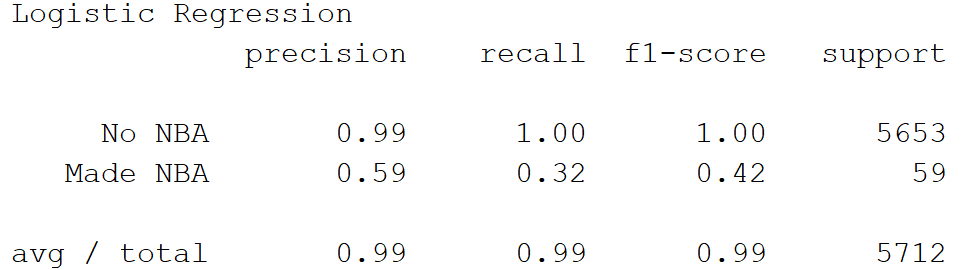
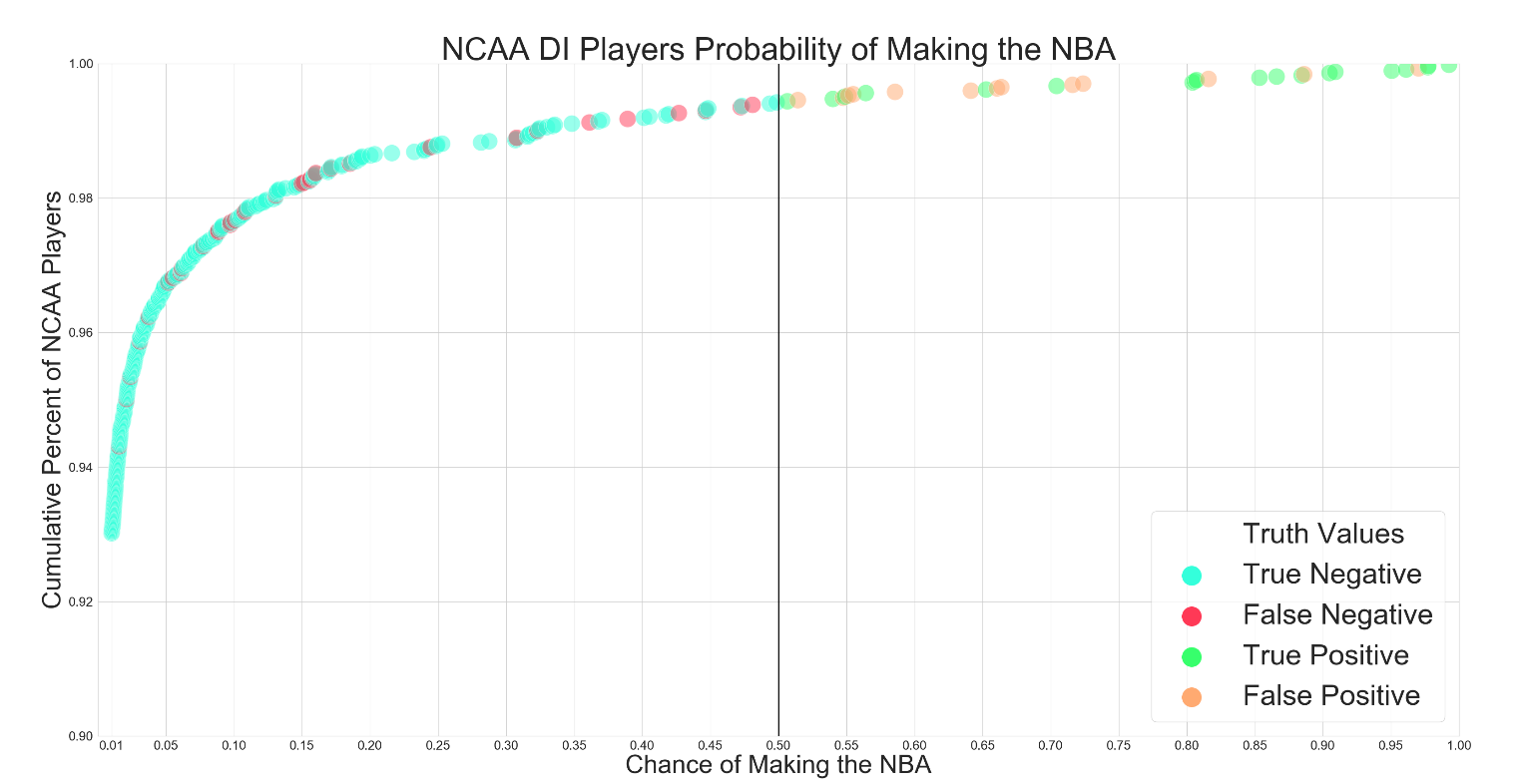
**Group Decision: I know you said in the last feedback that we had too many graphs. The reason they are all in here is because I know you wanted to see everything we tried since this is project. The target of predicting second round was useless for all of the scopes so if we were to publish this I would drop them. But I can write up these sections with more detail. Like I have for 8.2.1. But before I did that I wanted to know which sections we feel tell the best story and which ones I can put less detail into. I think we should do madeNBA and wasDrafted for all of them. And then briefly talk about firstround and lottery for all and fresh but not for last.**

## 8.2 Using all years of NCAA DI players

The first scope 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 and in order to keep the same dataset for the different scopes we considered we needed to look at player’s freshmen year. The following subsections are the logistic regression’s precision, recall and f1 scores for each of our target predictions along with an accompanying scatter plot displaying predicted probabilities that a player within the selected scope has at achieving the target.

### 8.2.1 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.

Yellow = Returned to college and went on to play in the NBA

Green = Returned to college and playing this collegiate season

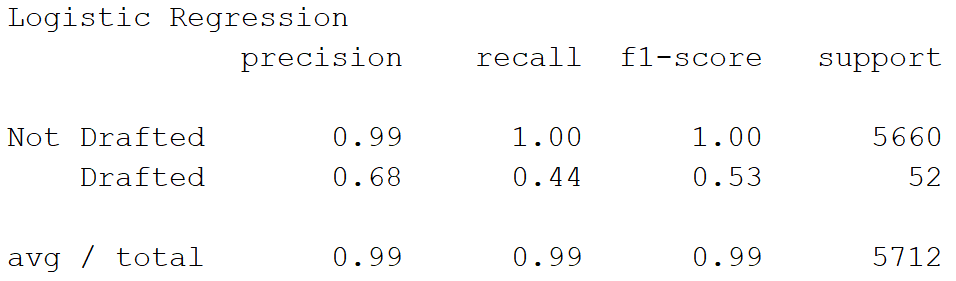
Blue = Returned to college and played in the G League

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Year** | **Predicted** | **Actual** | **MissType** | **Prob. Made Target** |
| Grayson Allen | 2015-16 | 1 | 0 | not | 97.02% |
| Kyle Anderson | 2012-13 | 1 | 0 | not | 88.62% |
| Miles Bridges | 2016-17 | 1 | 0 | not | 81.59% |
| Willie Cauley-Stein | 2012-13 | 1 | 0 | not | 72.38% |
| Jontay Porter | 2017-18 | 1 | 0 | not | 71.60% |
| Tyrone Wallace | 2014-15 | 1 | 0 | not | 66.37% |
| Jameel Warney | 2014-15 | 1 | 0 | not | 66.05% |
| Bryce Alford | 2014-15 | 1 | 0 | not | 64.13% |
| Juwan Morgan | 2017-18 | 1 | 0 | not | 58.56% |
| Ivan Rabb | 2015-16 | 1 | 0 | not | 55.49% |
| Kyle Wiltjer | 2014-15 | 1 | 0 | not | 55.13% |
| Bryce Alford | 2015-16 | 1 | 0 | not | 54.75% |
| Antonio Campbell | 2015-16 | 1 | 0 | not | 51.42% |
| Caris LeVert | 2015-16 | 0 | 1 | made | 48.10% |
| Joseph Young | 2014-15 | 0 | 1 | made | 47.22% |
| Dion Waiters | 2011-12 | 0 | 1 | made | 44.65% |
| Dakari Johnson | 2014-15 | 0 | 1 | made | 42.68% |
| Branden Dawson | 2014-15 | 0 | 1 | made | 38.91% |
| Tyler Cavanaugh | 2016-17 | 0 | 1 | made | 36.12% |
| Matt Costello | 2015-16 | 0 | 1 | made | 32.30% |
| Tyler Ennis | 2013-14 | 0 | 1 | made | 30.77% |
| Jerrelle Benimon | 2013-14 | 0 | 1 | made | 18.53% |
| Zach Collins | 2016-17 | 0 | 1 | made | 17.13% |
| Jarrett Allen | 2016-17 | 0 | 1 | made | 16.01% |
| Sterling Brown | 2016-17 | 0 | 1 | made | 15.61% |
| James Michael McAdoo | 2013-14 | 0 | 1 | made | 15.59% |
| Shake Milton | 2017-18 | 0 | 1 | made | 15.13% |
| Damyean Dotson | 2016-17 | 0 | 1 | made | 14.99% |
| Khyri Thomas | 2017-18 | 0 | 1 | made | 13.07% |
| Duncan Robinson | 2017-18 | 0 | 1 | made | 10.87% |
| Zach LaVine | 2013-14 | 0 | 1 | made | 10.79% |
| Travis Wear | 2013-14 | 0 | 1 | made | 10.07% |
| K.J. McDaniels | 2013-14 | 0 | 1 | made | 9.75% |
| Marquese Chriss | 2015-16 | 0 | 1 | made | 9.72% |
| Troy Brown | 2017-18 | 0 | 1 | made | 8.85% |
| Kay Felder | 2015-16 | 0 | 1 | made | 8.84% |
| Jake Layman | 2015-16 | 0 | 1 | made | 7.73% |
| Isaiah Whitehead | 2015-16 | 0 | 1 | made | 6.18% |
| Johnathan Williams | 2017-18 | 0 | 1 | made | 6.09% |
| Diamond Stone | 2015-16 | 0 | 1 | made | 5.51% |
| Fred VanVleet | 2015-16 | 0 | 1 | made | 5.17% |
| Marcus Paige | 2015-16 | 0 | 1 | made | 3.74% |
| Alize Johnson | 2017-18 | 0 | 1 | made | 3.09% |
| Marcus Derrickson | 2017-18 | 0 | 1 | made | 2.39% |
| Shawn Long | 2015-16 | 0 | 1 | made | 2.10% |
| Elfrid Payton | 2013-14 | 0 | 1 | made | 1.56% |
| Ben Bentil | 2015-16 | 0 | 1 | made | 0.99% |
| Kris Dunn | 2015-16 | 0 | 1 | made | 0.89% |
| Wesley Iwundu | 2016-17 | 0 | 1 | made | 0.78% |
| Alan Williams | 2014-15 | 0 | 1 | made | 0.64% |
| Shayne Whittington | 2013-14 | 0 | 1 | made | 0.51% |

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)

### 8.2.2 Predicting which 2018 NCAA DI players would play an NBA game

### 8.2.3 Predicting whether an NCAA DI player will be drafted



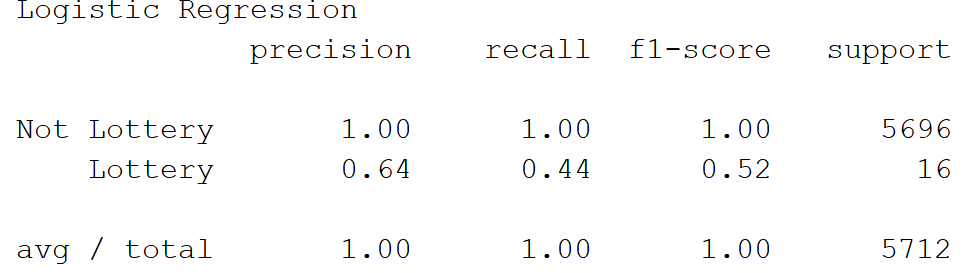
A close up of a map

Description generated with very high confidence40 misses

11 False Positive

29 False Negatives

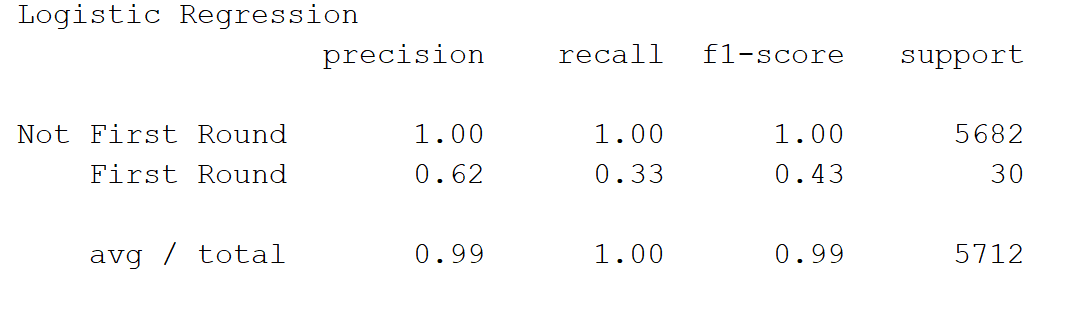
### 8.2.4 Predicting whether an NCAA DI player will be a lottery pick



A close up of a map

Description generated with very high confidence

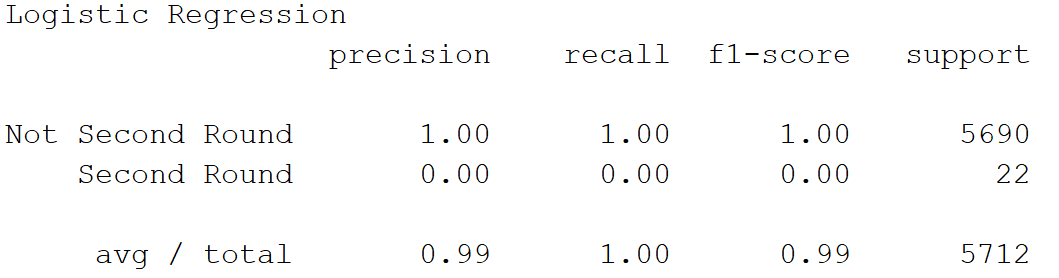
### 8.2.5 Predicting whether an NCAA DI player will be a first round pick



A screenshot of a map

Description generated with very high confidence

### 8.2.6 Predicting whether an NCAA DI player will be a second round pick



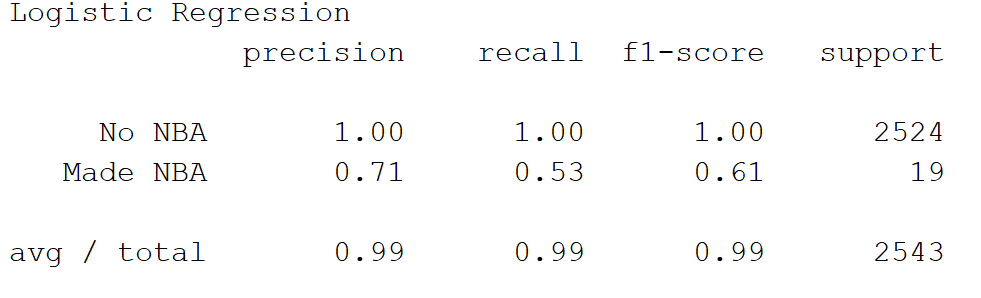
A screenshot of a cell phone

Description generated with very high confidence

## 8.3 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 following subsections are the logistic regression’s precision, recall and f1 scores for each of our targets along with an accompanying scatter plot displaying predicted probabilities that a player within the selected scope has at achieving the target.

### 8.3.1 Predicting whether an NCAA DI freshmen will play an NBA game

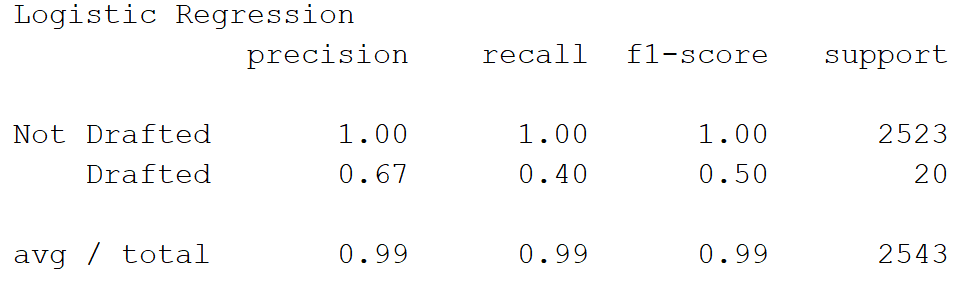


A close up of a map

Description generated with very high confidence

### 8.3.2 Predicting which 2018 NCAA DI freshmen would play an NBA game

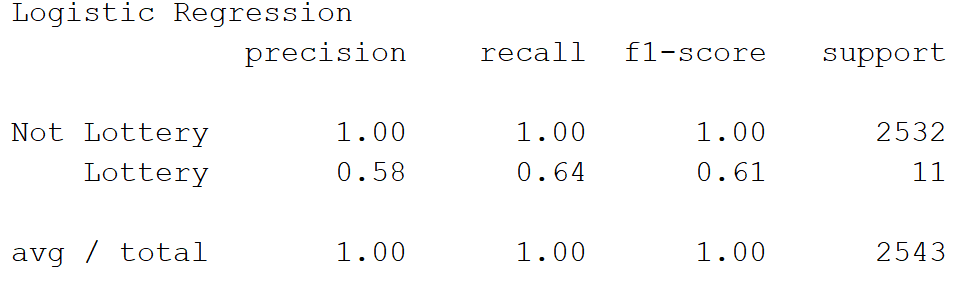
### 8.3.3 Predicting whether an NCAA DI freshmen will be drafted



A close up of text on a white background

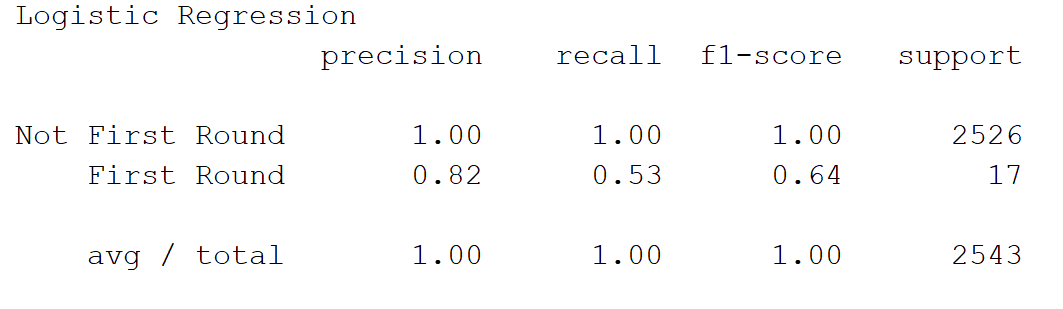
Description generated with very high confidence

### 8.3.4 Predicting whether an NCAA DI freshmen will be a lottery pick

A screenshot of text

Description generated with very high confidence

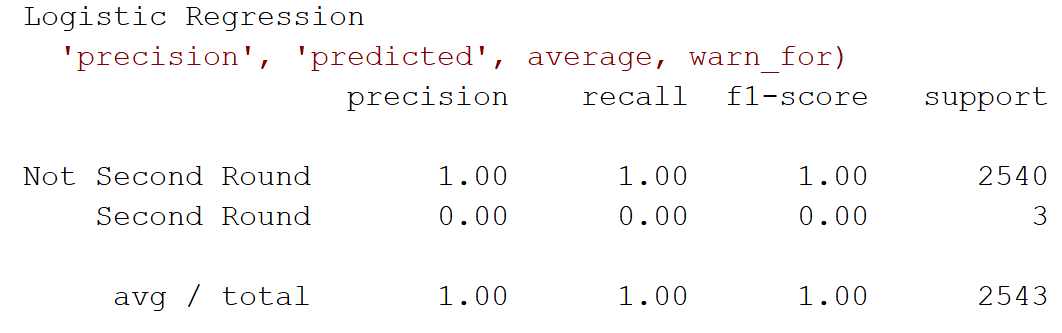
### 8.3.5 Predicting whether an NCAA DI freshmen will be a first round pick



A close up of text on a white background

Description generated with very high confidence

### 8.3.6 Predicting whether an NCAA DI freshmen will be a second round pick



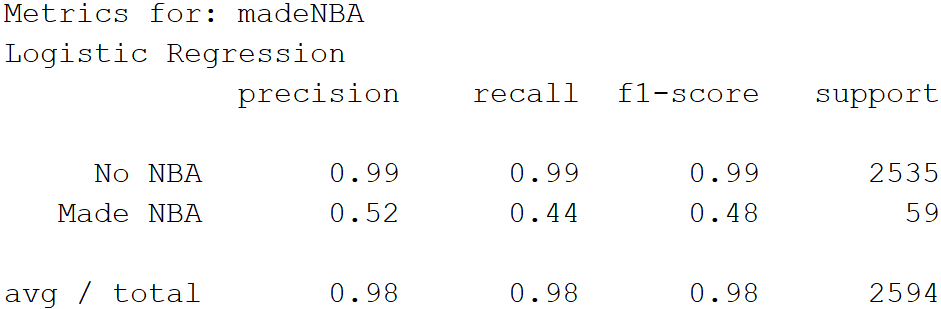
A close up of a map

Description generated with very high confidence

## 8.4 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. The following subsections are the logistic regression’s precision, recall and f1 scores for each of our targets along with an accompanying scatter plot displaying predicted probabilities that a player within the selected scope has at achieving the target.

### 8.4.1 Predicting whether an NCAA DI player will play an NBA game

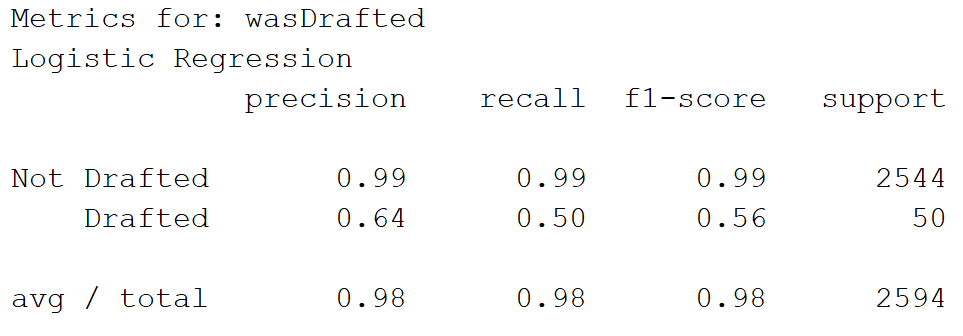


A close up of a map

Description generated with high confidence

### 8.4.2 Predicting which 2018 NCAA DI players would play an NBA game

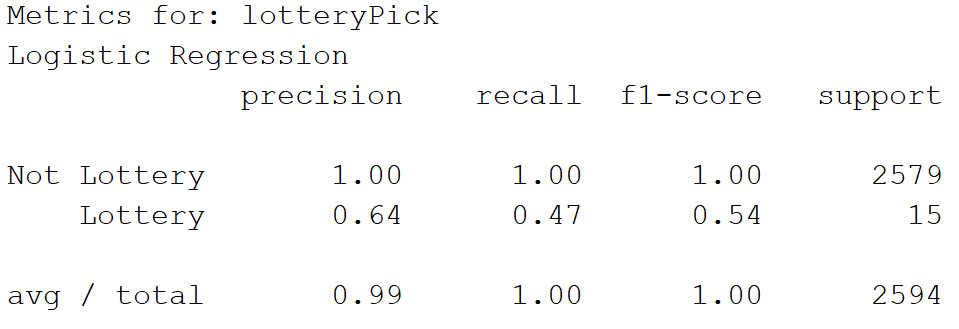
### 8.4.3 Predicting whether an NCAA DI player will be drafted



A close up of a map

Description generated with very high confidence

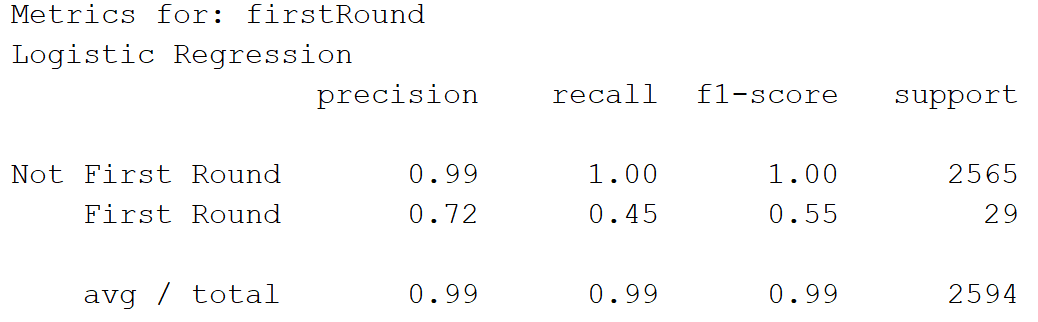
### 8.4.4 Predicting whether an NCAA DI player will be a lottery pick



A close up of text on a white background

Description generated with very high confidence

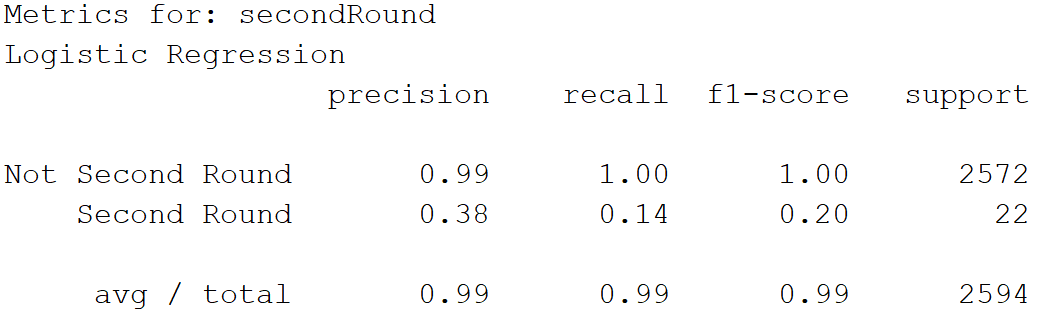
### 8.4.5 Predicting whether an NCAA DI player will be a first round pick



A close up of a map

Description generated with high confidence

### 8.4.6 Predicting whether an NCAA DI player will be a second round pick



A screenshot of a cell phone

Description generated with very high confidence

# 9. Discussion *(Mike)*

## 9.1 All Division I Players

Order of f1-score

Drafted -> Lottery -> NBA -> First -> Second

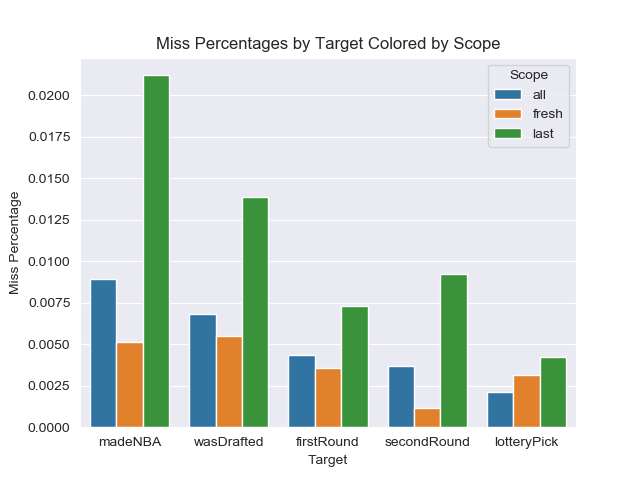
madeNBA: When it came to predicting if a player would make the NBA from looking at all of the seasons played by every player we were correct 19/59 times. However, when you look into the players who we predicted would make the NBA and didn’t 7 of them went on to play in the NBA after that collegiate season most notably because they returned for another year of college, 4 of the players signed NBA contracts for the summer league but were cut and currently play in the NBA G league, and the remaining 2 players played are currently playing in college but both are expected to be drafted Juwan Morgan and Jontay Porter. Of the 40 false negatives 8 of had greater than a 30% predicted chance of making the NBA.

## 9.2 Freshmen Division I Players

First -> Lottery -> NBA -> Drafted -> Second

## 9.3 Last Year Division I Players

Lottery -> First –> Drafted -> NBA -> Second



## 9.4 Dataset Imperfect

### 9.4.1 NBA / International / G League

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 a good portion of our false positives went on to play professional basketball league whether it be the NBA G League or Internationally. 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.

### 9.4.2 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 season adds more complexity to an already complex task. These players who return to college but were ready for the NBA often times are seen across our predictions as misses because they also were likely to be predicted to be drafted and in the first round. For example Aaron Harrison was “missed” 5 times because our model predicted he would play in the NBA, be drafted, be a first round and lottery pick in 2013-14 but since he returned to play another year at college all of these predictions were seen as false positives. A case could be made that had he declared for the draft he would have been drafted highly and played in the NBA. 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 and will always result in variability for these kinds of predictions.

## 9.5 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%.

## 9.5 Coefficients

We attempted to gain insight into that one percent and came away with a few takeaways. Our logistic regression found that the 10 most important metrics in determining if a player would make the NBA are the following with their positive weighting.

|  |  |
| --- | --- |
| Height | 1.38 |
| OWS | 0.92 |
| TOV | 0.87 |
| 3PA | 0.59 |
| AST | 0.56 |
| BLK | 0.55 |
| USG% | 0.54 |
| BPM | 0.52 |
| DRB | 0.42 |
| DBPM | 0.42 |

And the following are the 10 most detrimental metrics to your chances of making the NBA

|  |  |
| --- | --- |
| MP | -1.34 |
| AST% | -0.57 |
| PER | -0.51 |
| BLK% | -0.48 |
| PF | -0.47 |
| 3P | -0.41 |
| ORB% | -0.31 |
| FT | -0.31 |
| C | -0.29 |
| STL% | -0.29 |

# 10. Future Work *(Mike)*

## 10.1 Predicting which current (2019) NCAA DI players will play an NBA game

## 10.2 Comparing Draft Value Across Professional Sports

## 10.3 Considerations for How to Improve Project?