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

# Table of Figures

# Executive Summary

# *Jake*

# 1. Introduction

# 2. Background

# *Mike & Jake*

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

# 3. Methodology

# *Jake*

## 3.1 Analyze existing basketball player performance metrics

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

## 3.3 Find the highest value picks based on various measures of cost

### 3.3.1 Create a Jimmy Johnson-style NBA Draft value chart

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

# 4. Design

## 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. In order 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 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.

# *Mike*

# 5. Results

# 5.1 Analyze existing basketball player performance metrics

# WS VORP PER WS FP

# 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

# 6. Methodology for NCAA

# *Jake*

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

# 7. Design for NCAA

# *Jake*

8.5 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 confidenceIn 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.

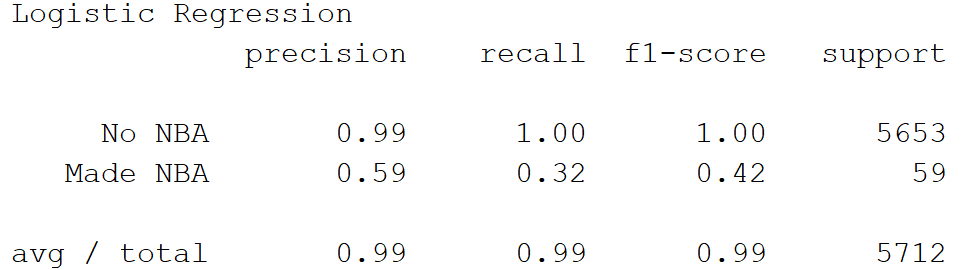
Note: due to an overwhelming amount of empty data points in the combine anthropology and agility datasets we didn’t these metrics. This is due to three main issues, the first is that players who attend the combine rarely perform all the tests, the second is often times the most notable college players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

# Form:

## 8.1 Using all years of NCAA DI players

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



53 Misses

13 False Positives

40 False Negatives

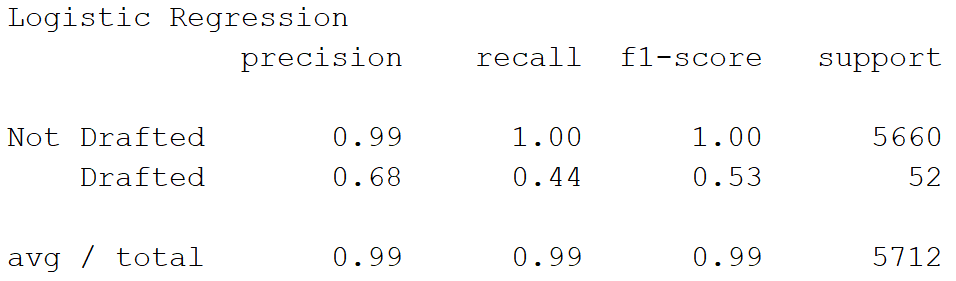
7 ended up playing in the NBA after dataset was collected, 4 are in the G League and the last 2 returned to college expected to be drafted this year ( Juwan Morgan, Jontay Porter)

A close up of a map

Description generated with very high confidence

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

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



40 misses

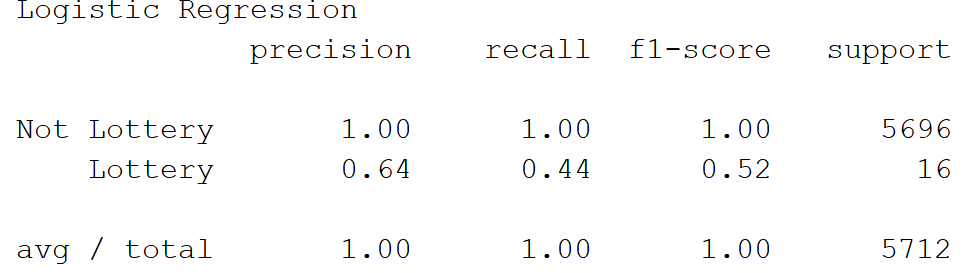
11 False Positive

29 False Negatives

A close up of a map

Description generated with very high confidence

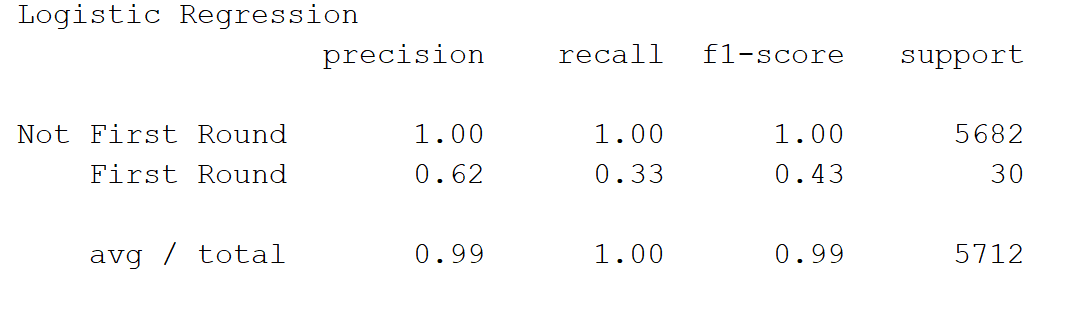
## 8.1.4 Predicting whether an NCAA DI player will be a lottery pick



A close up of a map

Description generated with very high confidence

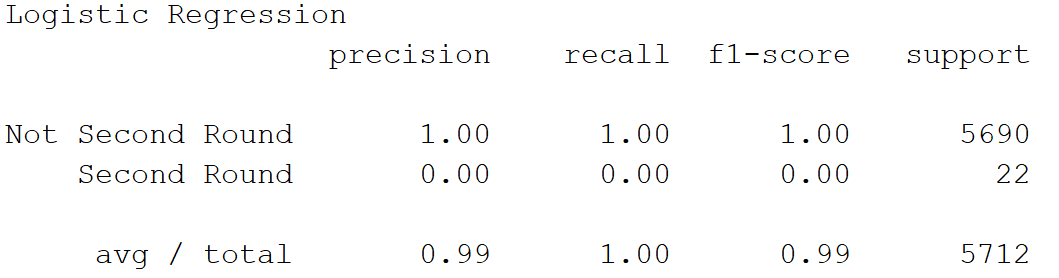
## 8.1.5 Predicting whether an NCAA DI player will be a first round pick



A screenshot of a map

Description generated with very high confidence

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

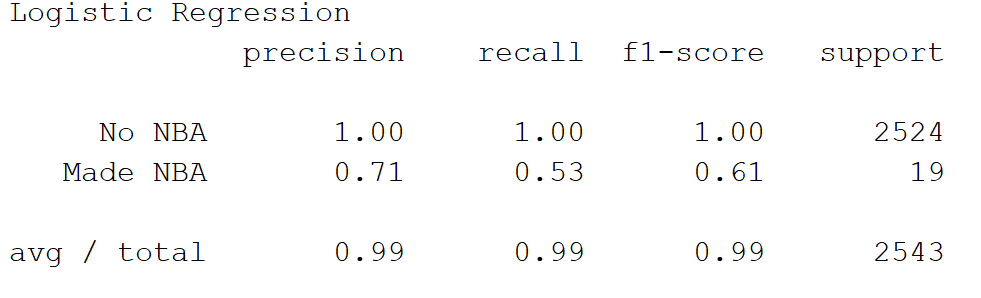


A screenshot of a cell phone

Description generated with very high confidence

## 8.2 Using only freshmen year seasons

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

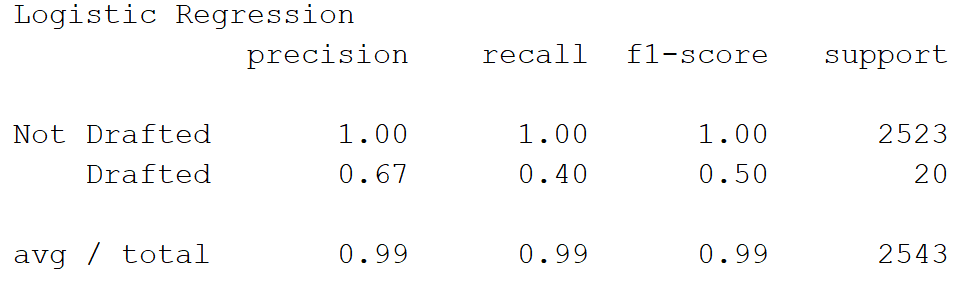


A close up of a map

Description generated with very high confidence

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

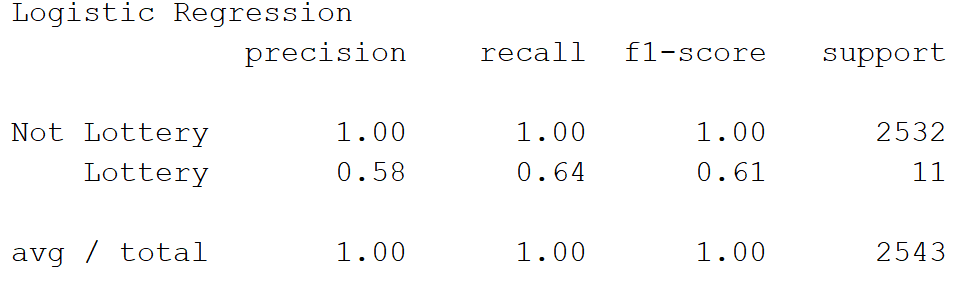
## 8.2.3 Predicting whether an NCAA DI player will be drafted



A close up of text on a white background

Description generated with very high confidence

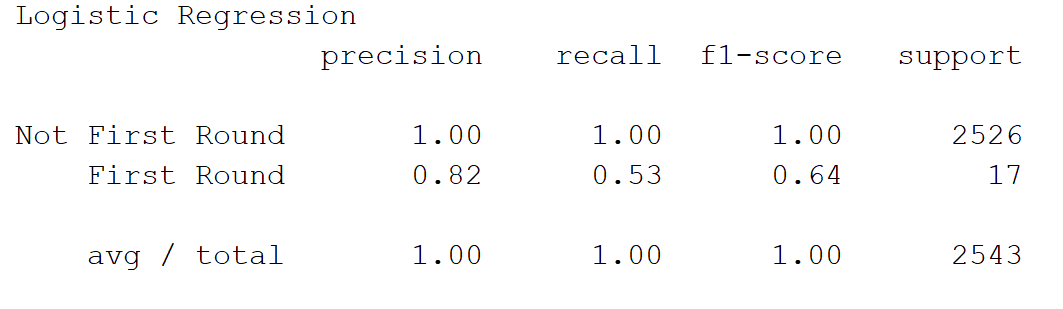
## 8.2.4 Predicting whether an NCAA DI player will be a lottery pick



A screenshot of text

Description generated with very high confidence

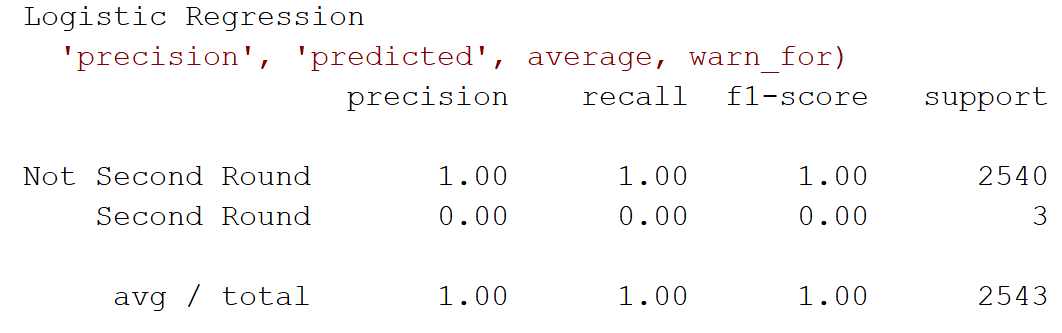
## 8.2.5 Predicting whether an NCAA DI player will be a first round pick



A close up of text on a white background

Description generated with very high confidence

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

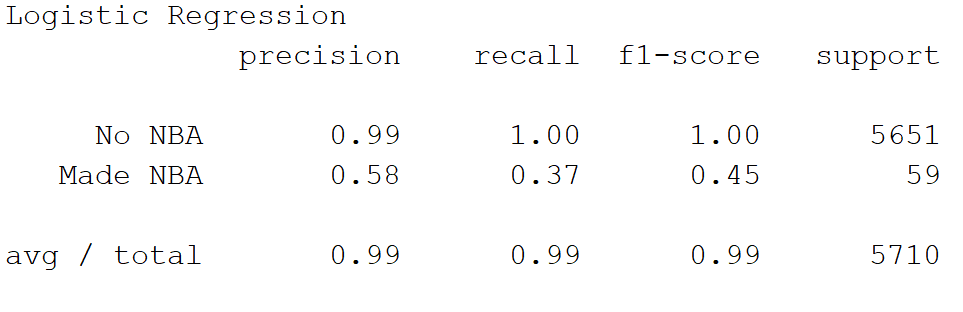


A close up of a map

Description generated with very high confidence

## 8.3 Using only a player’s last season

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

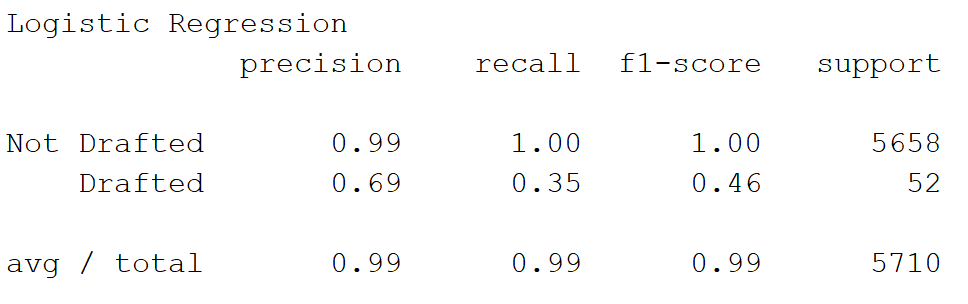


A close up of a map

Description generated with very high confidence

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

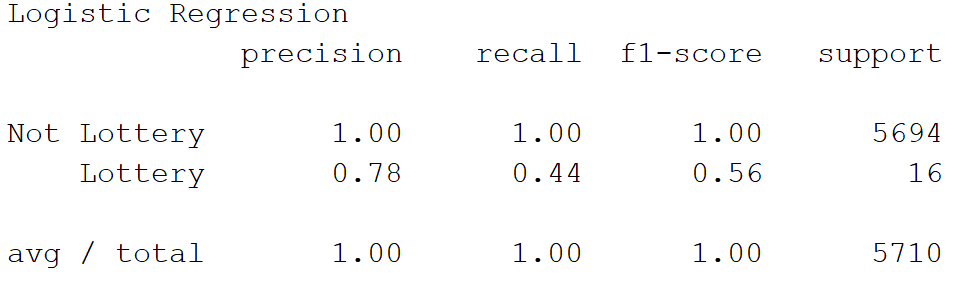
## 8.3.3 Predicting whether an NCAA DI player will be drafted



A close up of a map

Description generated with very high confidence

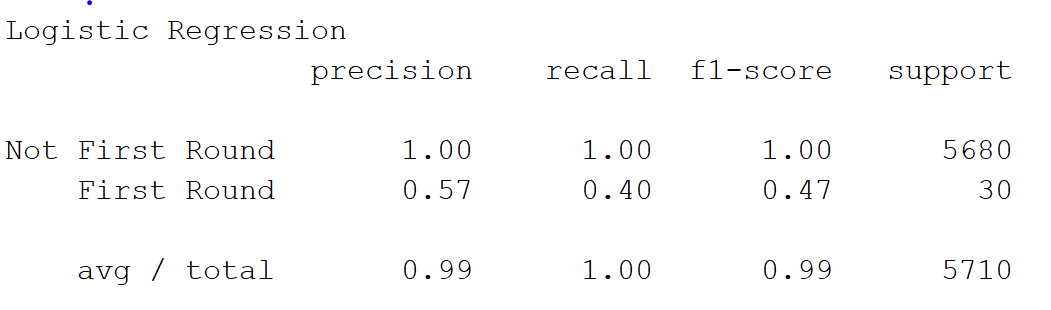
## 8.3.4 Predicting whether an NCAA DI player will be a lottery pick



A screenshot of text

Description generated with very high confidence

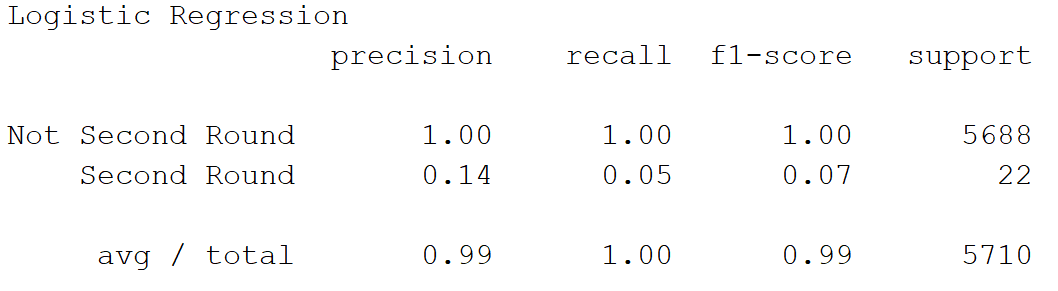
## 8.3.5 Predicting whether an NCAA DI player will be a first round pick



A close up of a map

Description generated with very high confidence

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



A close up of text on a white background

Description generated with very high confidence

# 9. Discussion

# 10. Future Work

## 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?

# 4.1 Sourcing draft data

# 4.2 Experiment design