Data 101 Class Project: Analyzing the Importance of Defense in College Basketball

Anthony Rahner

Arr234

**Introduction:**

Basketball has always been a massive part of my life, and it always will be. Even though I no longer play competitively, I still find the game fascinating to watch and analyze every day in my free time. While the NBA and professional basketball is very entertaining and filled with star players, I find college basketball just as exciting due to its extremely competitive nature and the possibility of upset wins at any time. I have also experienced the excitement of a college basketball game in person by going to some Rutgers games in the past, watching them get strong wins over teams like Seton Hall and Indiana.

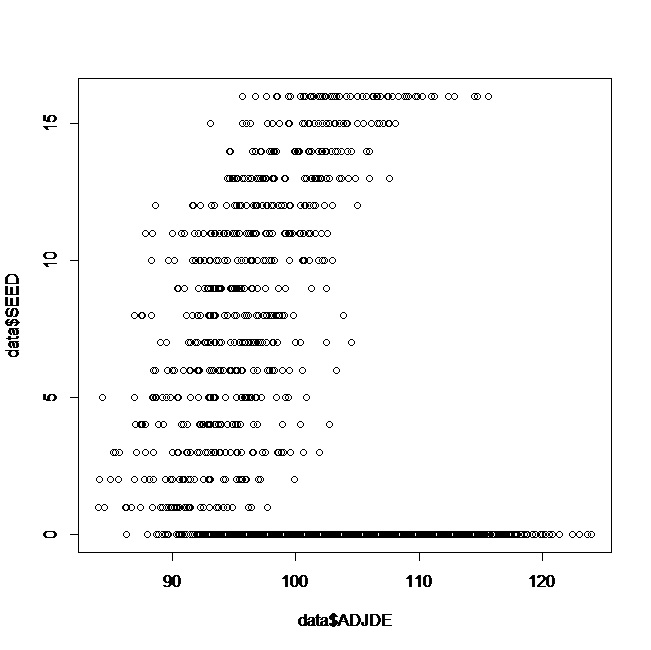
Like many other college teams, Rutgers relies on having a good defense to win games. Because of the lack of maturity and experience of college basketball players, good defenses can rattle opponents’ offenses and prevent them from playing how they want. This makes defense an extremely important part of the sport, and almost every elite program has a consistently great defense. This made me wonder, how do a team’s defensive statistics correlate with their chances of making the NCAA tournament? How many teams can make March Madness without their defense being up to par?

**Materials and Methods:**

To solve this problem, I found a college basketball dataset to study (found here: <https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset/>). In it, there are over three thousand Division 1 men’s basketball teams from the past ten years. It includes important team defensive statisticss about each one. For this research I used:

* Adjusted Defensive Efficiency (ADJDE): An estimate of the points allowed per 100 possessions a team would have against the average Division 1 offense
* Effective Field Goal % Defense (EFG\_D): The average effective field goal percentage allowed by a team (how well teams shoot against their defense)
* Two-Point Shooting % Allowed (X2P\_D)
* Defensive Rebound Rate (DRB)

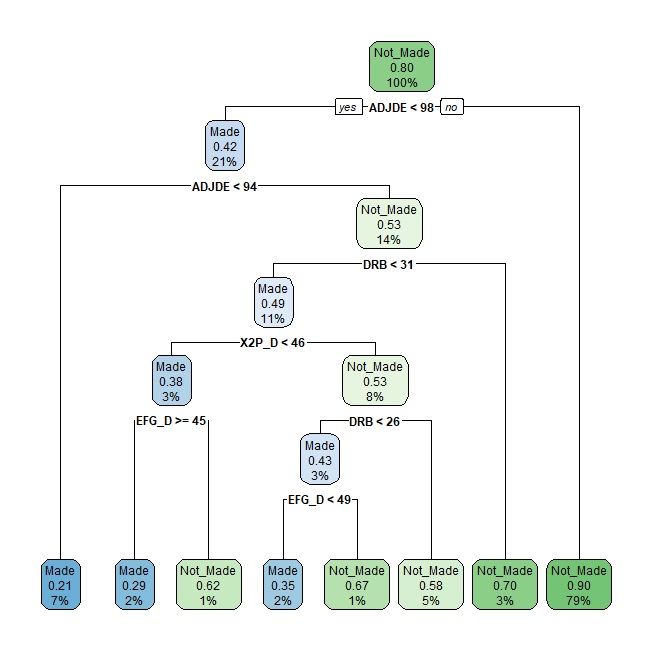
These will be used to find out how important defense is when it comes to making March Madness. The dataset also includes teams’ seeds in the tournament, with an N/A if they did not make it. I went in and changed all of the N/A’s to 0’s so I was able to say that if a team’s seed > 0, they got in. An example of the connection between defense and making the tournament is this scatter plot comparing adjusted defensive efficiency and tournament seeding:



Keep in mind that the lower ADJDE a team has, the less points they allow and the better defense they have. As you can see, once a team’s rating gets worse than 100, their seeding gets much lower. As you get further to the left, the teams with the better ratings are all higher seeds. It is also extremely rare for a team with a very good rating (less than 90) to miss the tournament.

Another adjustment I made to the dataset, which was done in R, was adding a column called made\_tourney, which would be “Made” if a seed was > 0, and “Not\_Made” if a seed was 0. Once that was added, I took a sample of 80% of the dataset to be the training data, with the rest being the testing data. I then put the training data into both an rpart() with a decision tree and a Naïve Bayes classification, and created predictions and confusion matrixes for both afterwards using the testing data.

**Results:**



This is the output of the training data put in an rpart() decision tree. As you can see, the prediction of defense being a key factor of making the tournament is proven in the very first node. In the entire training data, 79% of teams did not make the tournament if their adjusted defensive efficiency was higher (worse) than 98 points. Going further down the tree, each node proves that defense is very important by showing that much less teams make the tournament if they are worse than the statistic listed in the node. The only node that contradicts this is the one that says “EFG\_D >= 45”, but that is only working with 3% of the training data and still shows that it is very rare to make the tournament if you have a bad defense.

I then used the caret package in R to generate confusion matrixes of both the decision tree and a Naïve Bayes classification model. The accuracy of the decision tree predictions was about 0.86, while the accuracy of the Naïve Bayes predictions was 0.81, showing that the decision tree was more accurate than Naïve Bayes.

**Discussion:**

Based on my research, I can come to the conclusion that having a good defense plays a major role in making the tournament. The scatter plot and the decision tree that I generated both show significant information that proves it becomes less and less likely to be selected for March Madness if your defensive statistics are not on the higher end of the spectrum. Not only that, but the predictions for both the decision tree and the Naïve Bayes classification were able to predict at a rate of above 80% accuracy. This shows that there are not many outliers of teams with bad defenses that make the tournament, or teams with great defenses that do not make it. College basketball is known for valuing defensive principles much more than the NBA, and these charts and statistics back up that idea.

**Acknowledgements:**

I mainly got most of my coding from the active textbook and some of the files in our Canvas page. I also used ChatGPT for a few lines of code, such as adding the made\_tourney variable into my dataset, as well as to break down some of my coding sections that I had questions about.

**Literature Cited:**

The only outside sources that I used in this project was my dataset, which was created by Andrew Sundberg and was posted to Kaggle.com.

Full Code and Output: (also included in a .txt file in the assignment)

> library(rpart)

Warning message:

package ‘rpart’ was built under R version 4.3.2

> data <- read.csv("C:/Users/antho/OneDrive/Documents/Data101/classproject/cbb1.csv")

> data$made\_tourney <- as.factor(ifelse(data$SEED > 0, "Made", "Not\_Made"))

> head(data)

TEAM CONF G W ADJOE ADJDE BARTHAG EFG\_O EFG\_D TOR TORD ORB DRB FTR FTRD

1 North Caroli0 ACC 40 33 123.3 94.9 0.9531 52.6 48.1 15.4 18.2 40.7 30.0 32.3 30.4

2 Wisconsin B10 40 36 129.1 93.6 0.9758 54.8 47.7 12.4 15.8 32.1 23.7 36.2 22.4

3 Michigan B10 40 33 114.4 90.4 0.9375 53.9 47.7 14.0 19.5 25.5 24.9 30.7 30.0

4 Texas Tech B12 38 31 115.2 85.2 0.9696 53.5 43.0 17.7 22.8 27.4 28.7 32.9 36.6

5 Gonzaga WCC 39 37 117.8 86.3 0.9728 56.6 41.1 16.2 17.1 30.0 26.2 39.0 26.9

6 Kentucky SEC 40 29 117.2 96.2 0.9062 49.9 46.0 18.1 16.1 42.0 29.7 51.8 36.8

X2P\_O X2P\_D X3P\_O X3P\_D ADJ\_T WAB POSTSEASON SEED YEAR made\_tourney

1 53.9 44.6 32.7 36.2 71.7 8.6 2ND 1 2016 Made

2 54.8 44.7 36.5 37.5 59.3 11.3 2ND 1 2015 Made

3 54.7 46.8 35.2 33.2 65.9 6.9 2ND 3 2018 Made

4 52.8 41.9 36.5 29.7 67.5 7.0 2ND 3 2019 Made

5 56.3 40.0 38.2 29.0 71.5 7.7 2ND 1 2017 Made

6 50.0 44.9 33.2 32.2 65.9 3.9 2ND 8 2014 Made

> plot(data$ADJDE, data$made\_tourney)

> plot(data$ADJDE, data$SEED)

> set.seed(456)

> train\_indices <- sample(1:nrow(data), 0.8 \* nrow(data))

> train\_data <- data[train\_indices, ]

> test\_data <- data[-train\_indices, ]

> tree\_model <- rpart(made\_tourney ~ ADJDE + EFG\_D + DRB + X2P\_D, data = train\_data, method = "class")

> library(rpart.plot)

Warning message:

package ‘rpart.plot’ was built under R version 4.3.2

> rpart.plot(tree\_model)

> test\_predictions <- predict(tree\_model, newdata = test\_data, type = "class")

> library(caret)

Loading required package: ggplot2

Loading required package: lattice

Warning messages:

1: package ‘caret’ was built under R version 4.3.2

2: package ‘ggplot2’ was built under R version 4.3.2

> confusionMatrix(test\_predictions, reference = test\_data$made\_tourney, positive = "Made")

Confusion Matrix and Statistics

Reference

Prediction Made Not\_Made

Made 48 28

Not\_Made 71 558

Accuracy : 0.8596

95% CI : (0.8317, 0.8844)

No Information Rate : 0.8312

P-Value [Acc > NIR] : 0.02301

Kappa : 0.4154

Mcnemar's Test P-Value : 2.43e-05

Sensitivity : 0.40336

Specificity : 0.95222

Pos Pred Value : 0.63158

Neg Pred Value : 0.88712

Prevalence : 0.16879

Detection Rate : 0.06809

Detection Prevalence : 0.10780

Balanced Accuracy : 0.67779

'Positive' Class : Made

> library(e1071)

Warning message:

package ‘e1071’ was built under R version 4.3.2

> nb\_model <- naiveBayes(made\_tourney ~ ADJDE + EFG\_D + DRB + X2P\_D, data = train\_data)

> nb\_predictions <- predict(nb\_model, newdata = test\_data)

> confusionMatrix(nb\_predictions, reference = test\_data$made\_tourney, positive = "Made")

Confusion Matrix and Statistics

Reference

Prediction Made Not\_Made

Made 80 93

Not\_Made 39 493

Accuracy : 0.8128

95% CI : (0.782, 0.8409)

No Information Rate : 0.8312

P-Value [Acc > NIR] : 0.9113

Kappa : 0.4349

Mcnemar's Test P-Value : 3.968e-06

Sensitivity : 0.6723

Specificity : 0.8413

Pos Pred Value : 0.4624

Neg Pred Value : 0.9267

Prevalence : 0.1688

Detection Rate : 0.1135

Detection Prevalence : 0.2454

Balanced Accuracy : 0.7568

'Positive' Class : Made