Predicting March Madness
A Player Based Approach

Darren Lund

Brigham Young University
darrenlund@byu.net

February 21, 2018

Abstract

The NCAA March Madness Tournament provides a unique opportunity for sports analytics. Every year, millions of fans try to predict who will win each and every game, with no one managing to achieve a perfect bracket. Due to the vast amount of data collected in sporting events, the task for achieving the perfect bracket is a perfect for prediction algorithms. While many different methods have already been implemented, this project outlines a player based approach to predict some of the more nuanced factors of basketball, and thus more accurately predict upsets and appropriately named cinderella teams that bust brackets year after year. Data was scraped from .

Problem Statement

The art of predicting the outcome of a competition is at least as old as the ancient Roman Empire. Just as contests and competitions of skill have intrigued humankind, people have wanted to determine beforehand who the victor would be. One of the biggest instances of this in our day and age is the NCAA March Madness (Men's College Basketball) Tournament. A total of 68 teams coming together in a single elimination tournament to determine who will come out on top...and millions of fans predicting who will win each game.

The goal of the NCAA bracket competition, like so many others, is to score more points than others. Points are scored by correctly guessing which of two teams, paired against each other, will win in a single game at a neutral location. Games further along in the tournament are worth more than the initial games, with each successive round worth double the points of the previous round. For example, in the round of 64 (where there are 64 teams competing), games are worth 10 points. In the round of 32, games are worth 20 points, the sweet 16 games are worth 40, elite 8 are worth 80, final four worth 160, and the championship game worth a total of 320 points. The goal then, is to choose teams to maximize the bracket's total score by maximizing the weighted number of correctly chosen games. However, the primary goal of this project isn't just to win in the bracket competition, but to achieve what has never been done before: the perfect bracket. One where every single game was predicted correctly. Thus, instead of optimizing the bracket's score, this project seeks to predict each game independent of that game's position in the tournament. As such, the project will focus on the process for predicting a single game, as opposed to the tournament as a whole. It will do this by predicting the final score of each game individually, round by round, through estimating the number of points scored by each player individually, and thus determining the winner.

Data

All data was scraped from kenpom.com, a website managed by Ken Pomeroy. Ken has collected college basketball data for years, is independent of all teams in the NCAA, and readily gave me permission to use his data for this project when emailed. As such, there is no reason for any of the data to be biased, one way or another. Since the purpose of this project is to predict the results of the NCAA tournament (and thus get a perfect bracket) as opposed to predict who will end up participating in the tour-

nament, data was gathered on the 68 teams per year that have participated in the tournament since 2013. This data includes individual player statistics for each player of these teams on a game-by-game basis, as well as a yearly team scouting report. Thus, each match up can be considered by the individual players who should be participating, as well as how they work as a team.

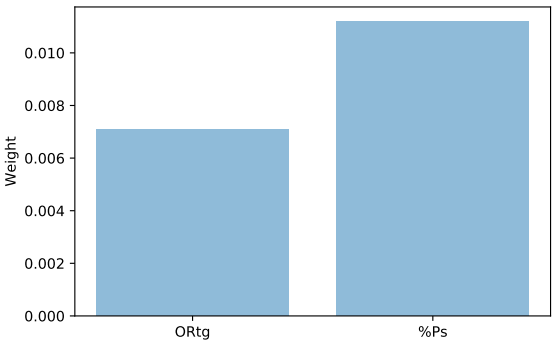
Methods

Player Selection

The first task is to classify players into one of two categories: significant contributors to the tournament, and non-significant contributors. This is because each team has more players than they usually will have play in the tournament, so we need only worry about players who actually will contribute to the tournament games. Whether or not someone was a significant contributor in any given game was determined by the percentage of possessions used during that game. If you were part of less than or equal to 10% of total possessions, then you were not considered a significant contributor. Anything greater was considered a significant contribution to the game. The reason for 10% is because there are multiple possible reasons a player could have some participation, but not actually be a significant contributor. For instance, if a game is going so well that the coach has no fear of losing, that coach may choose to put in players who, up to that point, had contributed nothing to the game so that they can have the experience of playing in the tournament. Alternatively, it's possible that a player started playing in the game, but before they managed to do very much became injured, and had to leave the game. Since there are 5 players on the court at a time, the 10% threshold requires that, on average, you made some contribution to the game every other possession.

Obviously, some of the statistics gathered have little to no bearing on whether or not a player will play in the tournament. For instance, no coach is going to keep you out of the game because you had 0 blocks, especially if you score 20 points on average per game. To determine which statistics were important and which were not, SKLearn's Lasso method was used on the averaged season statistics. The Lasso method utilizes L1 logistical regression for feature selection. Essentially, this amounts to weighting each statistic by how important it is in determining whether or not they will be significant contributors in the tournament while trying to keep the total weights as small as possible. Nonzero weights correspond to features that were important for classification, and zero coefficients those that could be ignored. The only variables with nonzero coefficients were Offensive Rating (ORtg) and Percent of Possessions Used (%Ps), as shown in figure 1. Utilizing these two variables alone achieved 87.89% accuracy [1] in determining which players would contribute to the tournament and which would not.

Figure 1: *Variable weights from Lasso Method (weights of 0 excluded)*



Scoring

Once the players have been determined, the next step is to predict how much each team will score. This is done by 1) approximating the number of scoring possessions each team will get; 2) removing the predicted number of scoring possessions lost due to turnovers; 3) using that number of scoring possessions and rebound probabilities to determine the number of scoring opportunities each team will receive. Those scoring opportunities are then divided among two point and three

[1] Data from 2013-2016 was used for training with 2017 data used for testing.

point attempts, with free-throw attempts calculated separately (due to the nature of the game). Those shots are then distributed to the players most likely to shoot them, and the player's data is used to determine how many they will make. This process will involve multiple machine learning algorithms, and Gibbs Sampling for determining player shot distributions. Each shot will be randomly drawn from said distributions, with the game run many times to get an average winning percentage for each team. This will give the probability that each team will win the game. For now, whichever team has the highest probability of winning will be predicted as the winner. During this process, all number of possessions, attempts, and made shots are rounded to integers to prevent illogical calculations (such as Nigel-Williams Goss of Gonzaga making 6.024 of his 7.593 three point shots).

Results

Analysis

Conclusion