Forecasting the 2023 NCAA Basketball Tournaments

Aleesa Mann and Cyruss Tsurgeon | MSDS 565, Spring 2023

We propose to perform predictive modeling to provide the best winning percentage for the 63 games of the both the Men's and Women's 2023 NCAA Basketball Tournament.

Predicting the winner of individual games and the ultimate champion of the NCAA Basketball Tournament is a fun, yet challenging endeavor many individuals around the world attempt each March-April when the tournament is held. According to the American Gaming Association, as many as 70 million brackets in 2017 (and an estimated 45 million people last year) were completed through some verifiable online sports website (such as CBSSports, ESPN, NCAA, etc) [1].

Although many people complete their bracket in private (and never share them), the longest (verifiable) streak of correct picks in an NCAA tournament bracket is at 49 games (by Greg Nigl of Columbus, Ohio in 2019) [2].

To date, no one has ever completed a 'perfect bracket' - correctly predicting the winner of each of the 63 total games in the NCAA Tournament. In fact, the odds of completing a 'perfect bracket' are astronomical. If you were to just guess or flip a coin the odds of all of your picks being correct are 1 in 9,223,372,036,854,775,808 - that's 1 in 9.2 quintillion. Or if you know a little something about basketball, the odds are estimated to be around 1 in 120.2 billion [3].

Each year, there is a team with 'no chance' who manages to beat a higher seeded team; or a team with a great regular season winning percentage who manages to be upset, unexpectedly, by another team. Hence, the popular reference to the tournament as 'March Madness'. Out of the popularity for the tournament itself and the rising popularity of predictive machine learning algorithms, many data scientists have also entered the fray to try to build a predictive model for the NCAA tournament.

While it would be unrealistic to expect to provide a model which creates the perfect bracket, we believe we can build a model which will outperform the national average which is around 50 percent of total games. In the past eight years of the NCAA.com Bracket Challenge Game, winners have averaged just 49.8 correct games in their brackets [3].

The data we will use will be the 31 csv files found within the Kaggle "March Machine Learning Mania 2023" competition [4]. These data contain files to make adequate predictions about the tournaments including cites (where games were held), teams and team records, past performances for regular season and tournament, and other files.

References:

[1] https://www.americangaming.org/new/espn-70-million-brackets-10-4-billion-in-bets-expected-for-march-madness/ and https://www.americangaming.org/resources/march-madness-2022/

[2] https://www.wavy.com/sports/ncaa-bastketball/who-picked-the-best-march-madness-bracket-of-all-time/

[3] https://www.ncaa.com/news/basketball-men/bracketiq/2022-03-10/perfect-ncaa-bracket-absurd-odds-march-madness-dream

[4] https://www.kaggle.com/competitions/march-machine-learning-mania-2023/data