Method to the Madness: An Analytical Approach to Bracketology

2022 Carnegie Mellon Sports Analytics Conference

September 2, 2022

Abstract

Every year in March, college basketball teams from across the country tune into the NCAA Tournament Selection Show. For better or for worse, some teams already know their fate. Other teams are left to stress about whether they'll hear their name called. The phrase "bracketology" was coined to describe the process of selecting and seeding the field of 68 teams. Some media outlets post daily bracketology projections during the regular season, giving fans an idea of where their team stands. In this paper, I set out to calculate day-by-day probabilities of making the tournament and expected seed for each team in Division 1 men's and women's basketball. These projections are based on the variables that the NCAA Tournament Selection Committee references when creating the bracket. These variables include the team's record, overall strength of schedule, quality wins, and more. The projections are visible via a web app that I have created to make this project more interactive. I also look to close the existing gap between men's and women's basketball in bracketology coverage. While there is no shortage of bracketology projections for the men's tournament, women's projections are much harder to find, especially through an analytical lens.

1 Introduction

The NCAA Tournament is known for all the excitement it jams into a two and a half week span, but there are four months of games prior to that. Many of those games directly or indirectly impact the tournament. One game between two blue blood programs in February can determine which team gets a top seed. The winner of a smaller conference championship game will almost certainly dictate which team goes to the big dance. In some ways, the events leading up to the tournament are just as exciting as the tournament itself, prompting media outlets to publish their "bracketology": a science for projecting the NCAA tournament field. The original bracketology is traced back to 1995 (Dietz, 2006), when Joe Lunardi projected the committee's bracket for the Blue Ribbon College Basketball Yearbook. Lunardi now projects the bracket for ESPN. College

basketball fans from around the country consume bracket projections at any point in the season, even before it starts. In this paper, I introduce a process for creating similar projections purely using stats, and using identical processes for both men's and women's basketball. The NCAA women's basketball tournament has frequently been an afterthought to the NCAA (Bachman & Higgins, 2021), despite its growing popularity (Brooks, 2022). The women's basketball community would greatly benefit from a tool that tracks each team's standing in NCAA tournament selection.

2 Process

First, it's important to have an understanding of the NCAA Tournament Selection Committee's process of creating the bracket. More goes into this than just taking the best 68 teams in the country and placing them where the committee feels they should be. There are 32 conference tournament champions that receive automatic bids and 36 teams that received at-large bids. The NCAA outlines the following procedure for creating the tournament bracket:

1. Select the 36 best at-large teams;
2. Seed the field of 68 teams; and
3. Place the teams into the championship bracket.

This process is identical for both the women's (NCAA, 2022b) and men's (NCAA, 2022a) tournaments. The third step is beyond the scope of this paper. So we will follow the first two steps. Since we will be making projections before the conclusion of conference tournaments, we have to include an extra step before steps 1 and 2. We will follow a process of:

1. Simulate the champions of all conferences;
2. Select the 36 best at-large teams; and
3. Seed the field of 68 teams.

3 Gathering Data

All data in this project was from the hoopR (Gilani, 2021) and wehoop (Gilani & Hutchinson, 2021) R packages. When trying to project a future event by training on cumulative past data, it's important to keep in mind that we cannot look forward in calculating stats. That is, if that goal is to project what the tournament field would look like on December 10th, we can only use results and stats from games on or before December 10th. It was also important to filter out data from games in the NCAA tournament since that cannot be considered when seeding the teams. Conference tournaments, however, can be included since they all are completed before the NCAA tournament field is set. The

training data for the men's side goes back to 2011 since that was the last time the tournament was expanded. The women's training data only goes back to 2014 due to shortage of box score data in 2011-2013. The women's tournament also just recently expanded to 68 teams for the 2022 tournament, so that may play a small impact on the modeling in the future.

4 Modeling

There are three models used in this project. The first is a logistic regression model used for predicting the outcome of a hypothetical game between two teams. It is used for simulations of each conference's tournament, determining which team is most likely to receive the automatic bid. Second is a random forest model to predict which teams are most likely to receive an at-large bid to the NCAA Tournament. Finally, once the 32 projected conference tournament champions and top 36 most likely at-large teams are determined, they are passed to a random forest model that determines which seed each of the 68 participating teams will be assigned, similar to the NCAA's procedure.

4.1 Head-to-Head Model

For the head to head model, I used a logistic regression model using the following variables:

- Whether the game was played at a neutral site
- Home/Away offensive and defensive rebounding rates
- Home/Away offensive and defensive strength of schedule
- Home/Away conferences
- Home/Away offensive and defensive turnover rates
- Home/Away offensive and defensive free throw factors
- Home/Away offensive and defensive steal rates

These variables predict a binary response of whether the home team won the game. Neutral site games still have a listed "home" team which is the reason for the neutral site indicator variable.

4.2 Conference Tournament Simulations

Using the head-to-head model, we simulate a conference tournament 100 times using the Monte Carlo method The team that wins the most simulations is considered the conference champion. However these simulations have assumptions that are violated in the real world of college basketball. It is assumed that each conference has all of its teams participate in the conference tournament and the

bracket is a standard single elimination tournament with a minimal number of first round byes. This assumption was made for simplicity but is simply not the case in reality. The Ivy League only has its top four teams from the regular season participate in their postseason tournament, and many conferences have multiple round byes for higher seeds. In fact, the top two seeds in the West Coast Conference received a three round bye in 2022. Once all conferences are predicted, the projected champion is placed into the field of 68.

4.3 At-Large Model

The seeding model consists of another random forest model that is trained on the final regular season stats for each team in the NCAA that did not receive an automatic bid to the tournament. The binary response variable was a 1 if the team qualified for the tournament, and 0 otherwise. This model returns a probability that the team will make the tournament. This model uses the following variables:

- Overall wins and losses
- Conference wins and losses
- Wins and losses against opponents in major conferences (Big 12, Big Ten, Big East, ACC, Pac-12, SEC)
- Wins and losses against opponents in mid-major conferences (Mountain West, American, Atlantic 10, West Coast Conference, Mid-American Conference, Missouri Valley Conference)
- Offensive and defensive rebounding rates
- Offensive and defensive rating
- Net Rating
- Average opponent offensive and defensive rating
- Average opponent net rating
- Turnover rate
- Opponent turnover rate
- Steal rate
- Opponent steal rate
- Pace
- Free throw factor
- Opponent free throw factor

- Conference

The 36 teams with the highest probability of receiving an at-large bid that were not projected to win their conference are added to the tournament field.

4.4 Seeding Model

The seeding model's training data is the 68 qualifying teams for each NCAA Tournament played since 2014 for women's and 2011 for men's. Using the exact same predictors as the at large model above, our model's response is now the seed that the team received. I used a classification model so this model returns a probability that the team will receive each number seed. A team is first given a label as the seed they are most likely to receive and all 68 teams are ranked 1 to 68 by their predicted seeds. Ties are settled first by the team with the higher probability of being a 1 seed, then probability of being a 2 seed, and the tie breaker criteria continues for all seeds to the 16th seed. We now have a ranked field of 68.

5 Results

An NCAA bracket projection is meant to be more descriptive than it is meant predictive. Bracketology predictions are often meant to be interpreted as "if the season were to end today, the bracket would probably look like this". However, there is still a lot of value in being able to predict the tournament bracket a few weeks out. I will measure the models by how many teams were correctly predicted to be in the field of 68 before conference tournaments start. Conference tournaments usually start in early March, so the projections referenced below were produced on March 1st of the corresponding year.

5.1 Women's Bracket Results

The year and number of teams correctly projected into the bracket are listed below. Note that the 2022 tournament was when the tournament expanded to 68 teams.

Year	Correct Teams
2014	47
2015	50
2016	46
2017	51
2018	53
2019	49
2021	50
2022	51

On average the models correctly predict 49.6 of the teams that make the women's NCAA tournament on March 1st, before conference tournaments start.

5.2 Men's Bracket Results

Below are the results for the men's side.

Year	Correct Teams
2011	49
2012	55
2013	52
2014	52
2015	50
2016	48
2017	55
2018	45
2019	54
2021	51
2022	47

On average, the models correctly predict 50.7 of the teams that make the men's NCAA tournament a couple of weeks before the selection show.

5.3 Shiny App

The Shiny App to view the results from March 1st of each season listed above is available here. This app will be updated into future seasons with new features added.

6 Next Steps

One of the most notable omissions from the models is that of "quality wins". It's difficult to determine which wins would be considered "quality" without looking forward. I plan to continually update and improve this project, and fully intend on adding that at some point in the future. While it would be tedious, it may be worthwhile to account for each conference's unique tournament procedure. At the very least, it would provide a more accurate and higher estimate for top teams' probabilities of winning their tournament. Similarly, as of right now this process does not track the results of conference tournaments as they happen. In the future I plan to have daily updates of conference tournament results. As the collection of play-by-play and event-level data continues, I would like to incorporate some more advanced metrics such as the quality of shots that teams take based on a shot's expected points. The more predictive metrics that can be found, the better our projections will be. Lastly, I would like to include the final step of the NCAA's tournament bracket creation procedures where the actual bracket is made. It would be a challenge but an interesting sort of constraint satisfaction problem.

7 Conclusion

In this paper, I introduced a very early stage method for projecting the NCAA Tournament bracket for both men's and women's college basketball. I created a shiny app that allows users to look through past retroactive projections of the tournament field prior to conference tournaments. All code used in this project can be found in this GitHub repository. The results were very promising. Before conference tournaments even started, the model can correctly place an average of 50 and 49 teams in the men's and women's tournament brackets respectively.

8 Acknowledgements

First and foremost I would like to thank the Carnegie Mellon Sports Analytics Conference for all of their work to put on the conference and this research competition. This project would not have been possible without the work of Saiem Gilani, Geoff Hutchinson, Jason Lee, and Billy Fryer for their efforts in creating the hoopR and wehoop packages. Lastly, I would like to commemorate the late Geoff Hutchinson, who passed away earlier this year. Geoff was a great friend to so many in the sports analytics community and is dearly missed.

References

Bachman, R., & Higgins, L. (2021). Ncaa undervalued women's basketball tournament by millions while prioritizing men's tourney, report finds. Retrieved from https://www.wsj.com/articles/ncaa-undervalued-womens-basketball-tournament-11628018560

Brooks, A. (2022). 2022 ncaa division i women's basketball championship is most-watched season finale in nearly two decades – 4.85 million viewers on espn networks. Retrieved from https://espnpressroom.com/us/press-releases/2022/04/2022-ncaa-division-i-womens-basketball-championship-is-most-watched-season-finale-in-nearly-two-decades-4-85-million-viewers-on-espn-networks/

Dietz, B. (2006). Ten questions for joe lunardi. Retrieved from https://www.news-gazette.com/sports/illini-sports/ten-questions-for-joe-lunardi/article_30ece418-4b7f-57cd-9fad-9188d1dee758.html

Gilani, S. (2021). *hoopr: The sportsdataverse's r package for men's basketball data.* Retrieved from https://hoopR.sportsdataverse.org

Gilani, S., & Hutchinson, G. (2021). *wehoop: The sportsdataverse's r package for women's basketball data.* Retrieved from https://wehoop.sportsdataverse.org

NCAA. (2022a). *2021-22 ncaa division i men's basketball championship principles and procedures for establishing the bracket* (Tech. Rep.). Retrieved from https://www.ncaa.com/_flysystem/public-s3/images/2022/02/13/2021-22%20Principles%20and%20Procedures.pdf

NCAA. (2022b). *2021-22 ncaa division i women's basketball championship principles and procedures for establishing the bracket* (Tech. Rep.). Retrieved from https://ncaaorg.s3.amazonaws.com/championships/sports/basketball/d1/women/D1WBB_BracketPrinciplesandProcedures.pdf