Analyzing the Creation of the March Madness Bracket with a Machine Learning Approach

CS4991 Capstone Report, 2023

Andrew Cornfeld
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
cpm6gh@virginia.edu

ABSTRACT
Bracketology is the process of predicting and creating the March Madness tournament bracket, which is very complex and not well understood. To predict tournament bids and their respective seeds, I utilized Machine Learning techniques to create a model which analyzes a team's resume to determine its seed placement. I imported game data from several Kaggle datasets and cleaned the data using Python packages in a Jupyter Notebook. I then used machine learning techniques and linear regression to predict the impact of all types of wins and losses on a team's seeding. I found that for teams to maximize their opportunity to get a tournament bid with the highest seed, they need to schedule several opportunities against high-caliber competition and avoid scheduling, and especially losing to, low-caliber competition at home. Future work on the model would involve further understanding of seeding teams with few high caliber opportunities.

1. INTRODUCTION

Bracketology is the process of predicting and creating the bracket as well as determining which teams have formed a strong enough resumé to get invited to college basketball's March Madness tournament. Sixty-eight teams get invited each year, 32 of which get auto-bids from winning their conference championship. The other 36 teams are at-large bids, determined by their resume throughout the regular season. Teams are then placed into one of four regions and seeded 1-16. This is all determined by a selection committee—a group of ten members from different conferences and schools nationwide. There are four play-in games for the four lowest at-large bids and the four lowest conference champions.

I created a model to predict a team's seeding based on their resume through a 30-game regular season and conference tournament. This model utilizes the NET (NCAA Evaluation Tool), which is made up of four factors: net efficiency, winning percentage, adjusted win percentage, and team value index. The model also uses the quadrant system and the location of the game: home, away, or neutral. I propose that this model could be used for teams to more effectively schedule their season to earn an at-large bid.

2. RELATED WORKS

Although many have used machine learning techniques to predict outcomes of tournament games, there is less existing research into the creation of the bracket. Lunardi (2022) is the most well-known of bracketologists, and his book on bracketology suggests there is no explicit formula for what a committee any given year favors. Some members may look heavily into box scores and predictive metrics, while others simply refer to the

Associated Press rankings to make their decisions. Lunardi is often consulted by programs to evaluate their schedule. He realizes smaller programs often cannot schedule the winningest competition (Duke, Kansas, etc.) and recommends they schedule other small programs with recent success.

Strack (2023) proposed another change to bracket construction, which involves minimizing the distances teams need to travel to their game sites while maintaining a fair bracket. He analyzed the NCAA tournament rules for seeding and created a penalty value analysis which could lead to fewer required flights. This would work to prevent situations similar to one in the 2023 NCAA Tournament where West Virginia was assigned to play in Birmingham, AL, while same-seeded Florida Atlantic was assigned to Columbus, OH. Geography can play a major role in bracket creation, with the 1st overall seed getting preference on where their games will be played.

3. PROJECT DESIGN

The Kaggle dataset the model is trained on is the 2023 March Madness data, which includes all games from the 2023 season and their outcomes. It also includes the seed each invited team received and their overall seed (EX: one team could be a 1-seed in the south region, but the 2nd overall seed). The model uses linear regression techniques to determine the seeding impact of each type of win or loss. In this way, it creates a formula which allows the number of wins and losses in each quadrant to calculate a seeding value, which can be stack ranked against other teams in contention.

3.1. Review of Game Importance

The two most important factors in the importance of a game are the NET ranking of the opponent, and the location of the game. The committee's quadrant system takes this into account, as explained in Table 1.

	Home	Away	Neutral
Quad. 1	1-30	1-75	1-50
Quad. 2	31-75	76-135	51-100
Quad. 3	76-160	136-240	101-200
Quad. 4	161-363	241-363	201-363

Table 1: Quadrant and Location Breakdown

Some bracketologists also consider Q1A, which is the top half of Q1. The NET rankings also change daily based on new results, so a Q2 win early in the season could become a Q1 win if the defeated opponent continues to produce high results. For this project, I used the NET rankings produced on the day of Selection Sunday (the day the bracket is revealed).

3.2. Data Processing

After obtaining all necessary datasets, I needed to process the data to remove unnecessary columns. Each college had a unique ID number in the "teams" dataset, but across each dataset, the naming was inconsistent (EX: "USC" in one dataset might be "Southern California" in another). I wrote quadrant. I also generated a list of all helper functions to detail team records in each conference tournament winners (which do not need to be considered in the at-large pool)

3.3. Model Creation

I used linear regression to create my model based on the 2022-23 data. My first strategy was to simply stack rank all invited teams with the features of Q1-4 wins/losses and the label being the overall seed (1-68) that they received, and let the linear regression determine the coefficients from there. I then took the sums that were generated by the model (by multiplying their Q1-4 wins/losses with their respective coefficients) to generate my own seed list. To evaluate error on my model, I calculated the residuals (differences

between predicted seed and actual seed) and calculated the root mean squared error. The coefficients and error are shown in Table 2.1 and 2.2. Note that the win coefficients are negative because the best label to receive is 1 and the lowest tournament bid is labeled 68.

	Win	Loss
Q1	-3.308	0.335
Q2	-2.579	1.696
Q3	-0.438	1.724
Q4	-0.325	1.178

Table 2.1: Quadrant Win/Loss Coefficients

Intercept	56.846
RMSE	5.485

Table 2.2: Intercept and RMSE values.

I also classified the error by determining the number of teams the model would have in the tournament that were left out, and vice versa. There were four such teams.

I considered other features of the data to use in the algorithm, such as road win percentage, Q1A wins/losses, Q win percentage, but no combination of features resulted in a lower RMSE than the model detailed above.

4. RESULTS

By looking at the coefficients on the model, we can clearly see that the most impactful win is a Q1 win. Q2 wins are also good to have; a Q1 win that becomes a Q2 win will not negatively impact a team heavily. The intercept implies that every team starts with a sum of 56.846 and can move up or down with each result. The model predicts the lowest at-large team to end with a sum of 43.281. However, only the last four at-large teams have a score over 37, which would put a team safely in the field. In a 30-game season it is impossible to achieve such a score winning only Q3/4 games (61 Q4 or 45 Q3 wins would be necessary), but much more feasible winning Q1/2 games (6 Q1 or 8 Q2 wins).

It is also curious to note that Q4 losses seem to impact teams less heavily than Q2/3 losses. One possible explanation for this is lack of data involving Q4 losses. Only nine of the 36 at large bids had any Q4 losses, and none had more than two.

I also ran the model with the parameters from the 2022-23 dataset on the 2021-22 dataset to evaluate the model's scalability, which was slightly further off, with an RMSE of 8.828. The model appears to scale well to previous data and does not show signs of overfitting.

5. CONCLUSION

Based on the model, any team looking to receive an at-large bid must evaluate the strength of their conference schedule (the last 16-20 games of their 30-game schedule) to make decisions about how difficult their non-conference schedule must be. Teams in high-major conferences (ACC, SEC, Big 10, Big 12, Big East, PAC-12) can rely on most of their conference games being Q1 or Q2 opportunities, whereas teams in mid-major leagues (not high-major) might get only a few Q2 opportunities in conference.

A common solution to this problem for mid-major teams is the buy game. High-major teams will provide mid-majors with a five-figure payout to come to their arena and play. This gives the high-major team another home game to add to their season ticket packages, and the mid-major team gets money for their program and a quality opportunity. High majors need to be careful not to schedule too many of these in their non-conference schedule, because as our model suggests, these Q3 or Q4 wins do very little for the team's resume. On the mid-major side, this can lead to teams going weeks in a row without playing in their home venue, which can reduce fan engagement.

Other opportunities for high majors include multi-team events, which involve teams travelling to a neutral site to play 2-3 games against other high major teams. Teams can also schedule home-and-home series with other teams, where they typically play at one team's home venue one year, and the other team's home venue the next year.

6. FUTURE WORK

One limitation of this model is that it considers each team equally far apart from the team above and below it in the stack ranking. A more accurate but memory intensive way to compute this could include comparing each resume to each other resume room with a one vs. one classifier, and a stack ranking could be determined based on which teams were deemed better in the most one vs. one classifiers.

This model could also be adapted to other college sports, such as determining the college football playoff teams, baseball playoffs and seeding, and others. The model could be expanded if March Madness was extended to include more teams and could also attempt to predict the field of the NIT (tournament for teams that were just outside of the March Madness field).

REFERENCES

Lunardi, J., Smale, D., & Few, M. (2022). *Bracketology: March madness, college basketball, and the creation of a national obsession*. Triumph Books.

Strack, M. (2023, March 21). *Team Assignment and Location Determination for the NCAA March Madness Tournament*. NC State University Libraries. https://repository.lib.ncsu.edu/bitstream/handle/1840.20/40863/etd.pdf?sequence=1