Predicting March Madness Results Using a Quantile Regression Approach

Kimberly Mays, advised by Dr. Eliana Cristou

University of North Carolina at Charlotte

Introduction

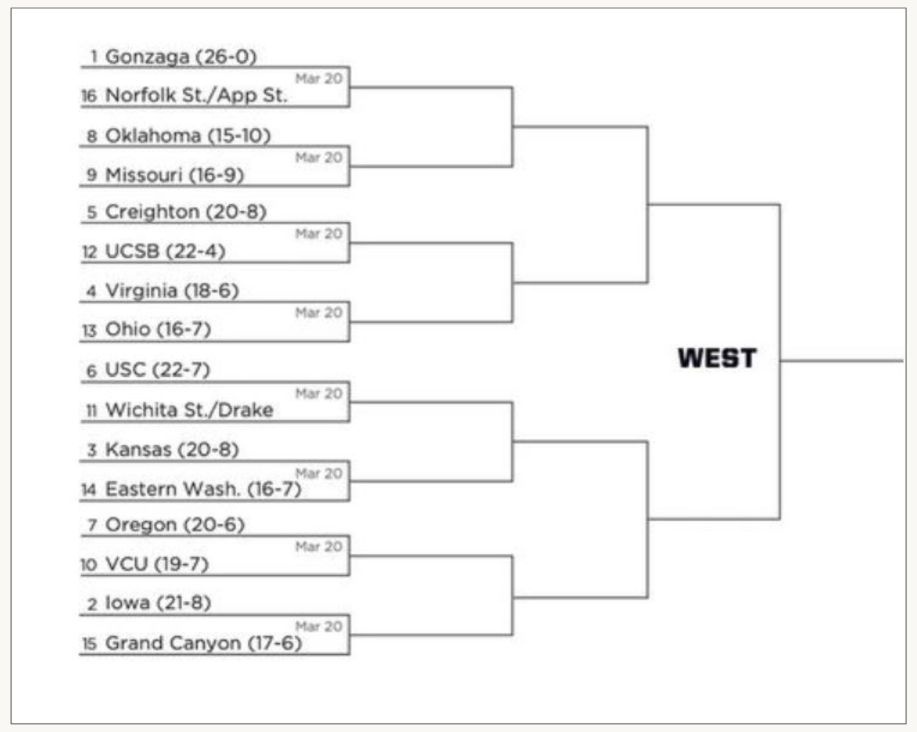
The NCAA Division I Men's Basketball Tournament provides a challenging opportunity to test predictive models:

- Tournament format mostly unchanged since 1985
- Inherent variability of amateur sports
- No perfect bracket to date
- 47 million U.S. bets on the tournament in 2021
- Bracket competitions such as Kaggle's Machine Learning Mania increase interest

Tournament Bracket Design

64 teams divided into 4 regions of 16 teams with 6 single-elimination rounds.

Round 1 pairings based on team seeding with the game seed sum equal to 17 (1 vs. 16, 2 vs. 15, etc.).



Sample region bracket (West region, 2021)

Winning teams advance along the region bracket, with the region winners advancing to Round 5.

Model Comparisons

- Seed: Region seed
- Pomeroy: Pomeroy College Basketball Rankings (kenpom.com)
- Sagarin: Jeff Sagarin's College Basketball Ratings (sagarin.com)
- LRMC: Logistic Regression/Markov Chain (gatech.edu/ jsokol/lrmc)
- Massey: Massey composite rank (masseyratings.com)
- RPI: Rating Percentage Index (collegerpi.com)

Methodology

Let $Y^{(i)}$ be a binary response, denoting a team's win or loss in the ith round, $i = 1, \ldots, 6$, where 0 represents a loss and 1 a win, and $\mathbf{X}$ a p-dimensional vector of predictors.

- **Goal:** Estimate the probability of winning, $P(Y^{(i)} = 1|\mathbf{X} = \mathbf{x})$, for a specific round $i = 1, \ldots, 6$.
- **Approach:** Estimate $P(Y^{(i)} = 1|\mathbf{X} = \mathbf{x})$ by averaging over multiple conditional quantiles $Q_\tau(Y^{(i)}|\mathbf{X} = \mathbf{x})$, $\tau \in (0, 1)$.
- **Model:** Assume $Q_\tau(Y^{(i)}|\mathbf{X} = \mathbf{x}) = g_\tau(\mathbf{B}_\tau^\top \mathbf{x})$, where $\mathbf{B}_\tau$ is a $d_\tau \times p$ matrix, $d_\tau \leq p$, resulting in new sufficient predictors.

Sample Level Algorithm

For each tournament round do the following:

1. Create a grid of quantile levels. For this work, we use equally spaced quantile levels $\frac{k}{10}$, $k = 1, \ldots, 9$.
2. Estimate $\mathbf{B}_\tau^\top \mathbf{x}$ using dimension reduction and form the new sufficient predictors $\widehat{\mathbf{B}}_\tau^\top \mathbf{x}$ following the approach of Christou (2020) [1].
3. Use a nonparametric technique to estimate the conditional quantile. In this work, we use the local linear conditional quantile regression. This gives $\widehat{g}_\tau(\widehat{\mathbf{B}}_\tau \mathbf{x})$.
4. Repeat steps 2 & 3 for the various quantile levels. Estimate $P(Y^{(i)} = 1|\mathbf{X} = \mathbf{x})$ by averaging over quantile levels using the approach of Hashem et al. (2016) [2].

Once the probabilities are calculated, game pairings are considered. The team with the highest probability is selected as winner and advanced to the next round.

Results

Single Scoring							
Year	AQR	Seed	Pomeroy	Sagarin	LRMC	Massey	RPI
2015	42	44	42	**45**	41	41	43
2016	**41**	37	39	40	40	39	38
2017	45	44	44	**46**	44	43	39
2018	36	36	38	39	**40**	39	38
2019	42	41	**44**	43	42	41	N/A
Total	206	202	207	**213**	207	203	158*
% of Points	65.4%	64.1%	65.7%	**67.6%**	65.7%	64.4%	62.7%

Double Scoring							
Year	AQR	Seed	Pomeroy	Sagarin	LRMC	Massey	RPI
2015	**94**	89	81	92	73	79	88
2016	**101**	87	79	82	93	88	73
2017	**140**	82	110	113	88	90	61
2018	63	81	78	**111**	110	79	84
2019	81	92	**127**	94	93	88	N/A
Total	479	431	475	**492**	457	424	306*
% of Points	49.9%	44.9%	49.5%	**51.2%**	47.6%	44.2%	39.8%

Predictions were scored against actual tournament results in both single and double methods to follow standard March Madness bracket scoring:

1. Single scoring: 1 point for every correct game prediction (max = 63)
2. Double scoring: the value of correct predictions double for each round, giving greater weight to end-of-tournament predictions (max = 192)

*Note: RPI (Rating Percentage Index) was discontinued after the 2017-18 season but was a common benchmark metric for seasons prior to 2018-19.

Data & Predictors

Except tournament seed, all predictors represent the season-wide averages:

1. region seed
2. 3 pointers per game
3. field goals per game
4. free throw attempts per game
5. free throws per 100 possessions
6. offensive rebound percentage
7. offensive rebounds per game
8. defensive rebound percentage
9. defensive rebounds per game
10. assists per game
11. fouls per game
12. scoring margin
13. assist to turnover ratio
14. offensive efficiency
15. defensive efficiency

Discussion

- Sagarin had the best performance overall for both single and double scoring; however, the ratings use proprietary metrics.
- Our method uses only freely-available game data and was the best 3 of the 5 years and 1 of the 5 in double and single scoring respectively.
- Of the remaining methods, Pomeroy was the top method 1 year in both scoring methods; Massey, RPI, and seed rankings never earned the highest score.
- 2018 was the only tournament to date where a #1 seed team lost in the first round.
- Algorithm easily adapts to other "successes": covering the spread, upsets by seed, etc.

References

[1] Christou, E. (2020). Central quantile subspace. *Statistics and Computing*, 30, 677-695.

[2] Hashem, H., Vinciotti, V., Alhamzawi, R., & Yu, K. (2016). Quantile regression with group lasso for classification. *Advanced in Data Analysis and Classification*, 10, 375-390.

Acknowledgements

Special thanks to the authors of Hashem et al. (2016) for use of their code.