MARCH MADNESS SEEDING ANALYSIS

By Graham Dynis, Isaiah Kuehl, Grace Burns, & Kaleb Jordan

Introduction to March Madness

- The March Madness tournament that takes place every year features 68 Division 1 basketball teams, some of which are automatic bids and some are at-large bids
- Teams are given a seed of 1-16 based on their performance and given to them by the selection committee
- Our project seeks to answer 2 questions:
  - How does the selection committee select the teams in the field?
  - How does the selection committee seed the teams once they are selected?

Dataset Overview

Two distinct datasets with same predictors

Bubble Selection: 73 observations (2021-2025)

Handpicked teams on the "bubble"

Response Variable: Binary **IN OR OUT**

Seeding: 250 observations (2021-2025)

Every team seeded 1-12 in last 5 tournaments

Response Variable: **Seed**

Predictors:

Resume Metrics: (ESPN Strength of Record, KPI, WAB [when available])

Predictive Metrics: (ESPN BPI, KenPom, Sagarin and Bart Torvik [when available])

NCAA NET Rating

Quad 1 Wins, Quad 1 Winning Percentage, etc.

Power Conference Dummy Variable

Predicting NCAA Tournament Bubble Selection with Binomial GLMs

Goal: Predict March Madness at-large selections near the "bubble" using historical data

Data: 73 teams from 2021–2025, with metrics like SOR, KPI, BPI, KenPom, and more. 2021-2024 observations were used as the training set while 2025 was the test set.

Model: Final model is a binomial GLM using:

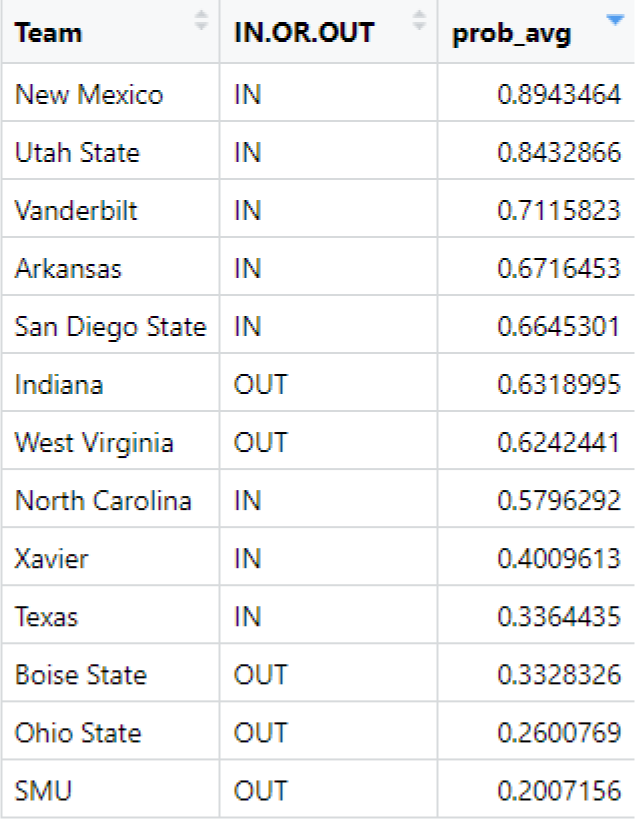
- Resume Metric Average (SOR, KPI, and WAB [2025 only])
- Predictive Metric Average (BPI, KenPom, Sagarin [2021-2023], and Torvik [2025 only])

Final Model: Accuracy and Interpretability

Final Model: *Probability of making tournament =* 7.25013 + (-0.13366 • Resume Average) + (-0.01085 • Predictive Average)

- **2021-2024 training set accuracy: 71.9% (23/32)** on historical data

- **2025 test set accuracy: 75.0% (6/8)**



Team	IN.OR.OUT	prob_avg
New Mexico	IN	0.8943464
Utah State	IN	0.8432866
Vanderbilt	IN	0.7115823
Arkansas	IN	0.6716453
San Diego State	IN	0.6645301
Indiana	OUT	0.6318995
West Virginia	OUT	0.6242441
North Carolina	IN	0.5796292
Xavier	IN	0.4009613
Texas	IN	0.3364435
Boise State	OUT	0.3328326
Ohio State	OUT	0.2600769
SMU	OUT	0.2007156

Forecasting NCAA Tournament Seeds with Ordinal Regression

Goal: Predict NCAA March Madness tournament seedings using team performance metrics

Data: 250 teams from 2021–2025, with metrics like NET, KPI, SOR, BPI, Wins, and conference info. 2021-2024 observations were used as the training set while 2025 was the test set.

Model: Final model is an ordinal regression using:

- Performance metrics (NET, KPI, SOR, Quad 1 Wins)
- Conference-adjusted stats (BPI_cwg, SOR_cwg)
- Interaction terms (e.g., Quad 1 Wins × Quad 1 Opportunities)

Final Model: Accuracy and Interpretability

Baseline Model: *Seed =* (0.082 • BPI) + (0.103 • KPI) + (0.155 • SOR)

Final Model: *Seed =* (0.067 • NET) + (0.068 • KPI) + (0.231 • SOR) + (-1.129 • W) + (0.061 • BPI_cwg) + (0.035 • W:Opp) + (0.050 • KP:Conference Champ) + (-0.003 • KP:SOR_cwg)

- **2021-2024 training set accuracy:**
- **Exact Seed match: 53.5%** on historical data (↑ from 44% in baseline)
- **Within ±1 Seed: 89%** accuracy (captures true seed range very well)

- **2025 test set accuracy:**
- **Exact Seed: 64%**
- **Within ±1: 92%**

Better Predictions = Smarter Brackets

- **Improved Prediction Accuracy**: More accurate seed forecasts help analysts, fans, and selection committees anticipate and understand bracket seeding.

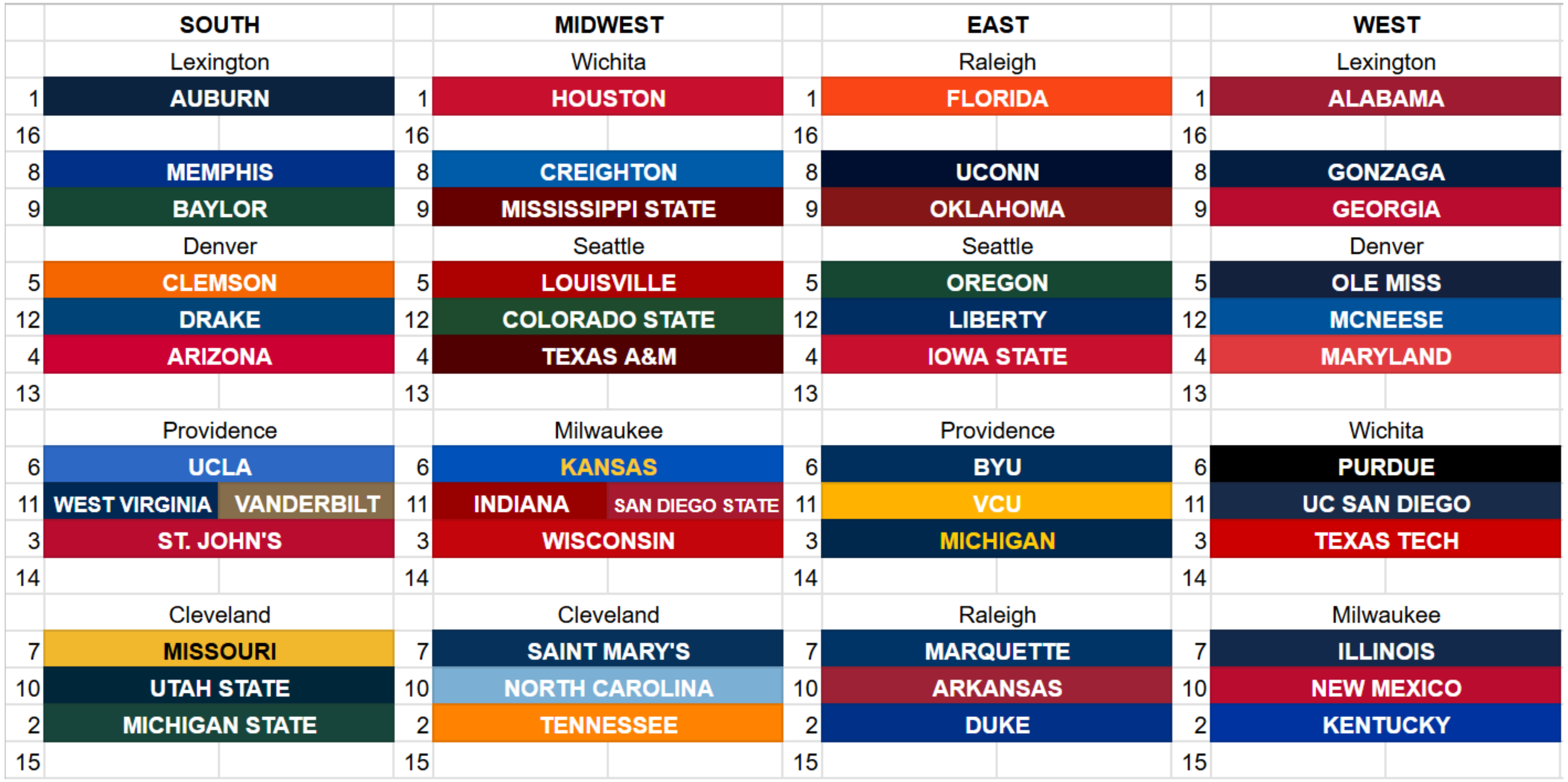
- **True Seeding**: Ordinal logistic regression better models the ranked nature of tournament seeds, avoiding the errors of linear regression (e.g., predicting impossible values like Seed 0 or 13).

- **Reduces Bias**: By adjusting for Power vs. Non-Power Conference effects, the model reduces bias and offers fairer comparisons between teams with different resources and schedules.

- **Key Interactions**:The model’s interaction terms reveal hidden patterns such as how Quad 1 Wins impact seeding more when opportunities are fewer

- **Insights**:The model can be used to accurately identify top teams (e.g., Auburn, Houston, Florida, Alabama) before official bracket release, showing strong real-world utility.

Final Bracket Results



	SOUTH		MIDWEST		EAST		WEST
	Lexington		Wichita		Raleigh		Lexington
1	AUBURN	1	HOUSTON	1	FLORIDA	1	ALABAMA
16		16		16		16	
8	MEMPHIS	8	CREIGHTON	8	UCONN	8	GONZAGA
9	BAYLOR	9	MISSISSIPPI STATE	9	OKLAHOMA	9	GEORGIA
	Denver		Seattle		Seattle		Denver
5	CLEMSON	5	LOUISVILLE	5	OREGON	5	OLE MISS
12	DRAKE	12	COLORADO STATE	12	LIBERTY	12	MCNEESE
4	ARIZONA	4	TEXAS A&M	4	IOWA STATE	4	MARYLAND
13		13		13		13	
	Providence		Milwaukee		Providence		Wichita
6	UCLA	6	KANSAS	6	BYU	6	PURDUE
11	WEST VIRGINIA VANDERBILT	11	INDIANA SAN DIEGO STATE	11	VCU	11	UC SAN DIEGO
3	ST. JOHN'S	3	WISCONSIN	3	MICHIGAN	3	TEXAS TECH
14		14		14		14	
	Cleveland		Cleveland		Raleigh		Milwaukee
7	MISSOURI	7	SAINT MARY'S	7	MARQUETTE	7	ILLINOIS
10	UTAH STATE	10	NORTH CAROLINA	10	ARKANSAS	10	NEW MEXICO
2	MICHIGAN STATE	2	TENNESSEE	2	DUKE	2	KENTUCKY
15		15		15		15	

Risers:

Alabama gets last 1 seed

Kentucky gets last 2 seed

Louisville earns 5 seed

UNC moves to 10 seed

WVU and **Indiana** make field

Fallers:

Duke gets 2

St. John's on 3 line

Purdue falls to a 6

Memphis falls to 8 line

Xavier and Texas fall out of tourney

Takeaways & Next Steps

- Ideally, we would be able to correctly seeds 13-16 in the same way that we predict seeds 1-12, which is more difficult due to the nature of these teams being Mid-Major Conference Champs
- Additionally, we would like to test different models for prediction, such as random forest models, gradient boosting machines, and neural networks.
- In the seeding models we used, only very recent seasons were used, and we would like to increase the size of the dataset.
- Due to the human element of the seeding process by the committee, it is very difficult to predict due to the metrics we used and the metrics they used being slightly different.
- We also would like to hopefully include more variables

The Human Element

- Because what we are attempting to predict in this model is human decisions, it is nearly impossible to fully predict.
- We have come very close, with some slight room for improvement, to being able to predict the first 12 seeds of the NCAA tournament.
- The human element is especially relevant when considering the top of the bracket and the bubble teams.
- The committee also watches games and uses the eye test in their selection process.
- Question to consider: Should the committee be replaced by an unbiased model or continue to have some subjectivity?