March Madness Seeding Analysis

By Kaleb Jordan

TABLE OF CONTENTS

INTRODUCTION

The NCAA March Madness tournament is one of the most anticipated sporting events of the year, with 364 Division I college basketball teams competing for favorable seeding positions that can significantly influence tournament success. By analyzing historical data collected from Bart Torvik's website over the past 5 years, we used statistical methods to assess how accurately we can classify a team's Seed using various resume-based and quality-based metrics. The objective of this study is to build and evaluate high-performance predictive models that classify whether a team will receive a tournament bid and estimate their eventual seed using both resume and quality-based metrics.

METHODS

Our data consists of 1794 observations spanning the last five NCAA Men's March Madness tournaments (68 per year), from Seeds 1-17 (Seed 17 denotes no tournament appearance). Using *Seed* as the target variable, we have the variables: *BPI* (Basketball Power Index), *KP* (KenPom Rankings), *KPI* (Key Performance Indicator), *NET* (NCAA Evaluation Tool), *KPI* (Key Performance Indicator), and *SOR* (Strength of Record), along with *Q1 Wins and Losses* and *Q4 L*. Many variables have also been transformed for higher performance.

An overview of a few of the potential predicting variables:

- **BPI:** BPI is a statistic used to determine a team's performance, rather than the result. Instead, BPI considers factors such as a team's pace (number of possessions per game) and offensive/defensive efficiency.
- **KP:** KP evaluates a team's performances by examining both offensive and defensive ratings, identifying the strengths and weaknesses of each team.
- **KPI:** KPI is a result-based statistic that evaluates team performances per game on a plus-minus 1 evaluation. It determines the margin a team wins or loses by, with 0.0 representing a close game.
- **NET:** NET is an in-depth measurement that involves numerous metrics to determine a team's strength and worthiness of their tournament chances.
- **SOR:** SOR is a way to determine a team's accomplishments throughout the season by investigating the difficulty of their schedule and the actual results.

The first stage was to accurately estimate the 37 of the 333 non-conference championship winning teams that will be seeded in the NCAA Men's Division I tournament. The first of two distinct models used was a logistic regression that calculated the predicted probabilities and assigned a binary classifier, 0 or 1. 0 represents they did not make it, inversely, 1 represents a school did.

We used numerous algorithms and statistical techniques to determine a model that best represents the data. After careful analysis, we identified algorithms that could be useful for a project like this: a **simple linear regression** and an **ordinal logistic regression** model. Both models have their own pros and cons. A simple linear regression model could be applied to ordinal outcomes like Seed, and it would be straightforward to understand, but it treats the response as continuous and assumes equal spacing between Seed thresholds. This can lead to biased or uninterpretable results (e.g., Seed 0). A more appropriate approach is to use an ordinal regression model, which accounts for the ranked nature of the outcome and provides more meaningful probability-based

predictions. Ordinal regression models are easy to interpret (e.g., one-unit increase in BPI affects Seed) and capture the ordinal nature of the data (Seed 1 is better than Seed 2, etc.).

Ordinal regression has some disadvantages over the others, such as its flexibility, as it does not capture non-linearities or interactions. Since it captures the proportional odds, the relationship between the response variable and the predictors remains constant for all thresholds. This may be unrealistic to our data. The proportional odds model assumes that the relationship between predictors and the log-odds of being a lower seed is constant across thresholds. This assumption simplifies interpretation, but may not hold if certain metrics disproportionately affect top or bottom seeds. Below is a diagram of the formula for an ordinal regression model:

An ordinal model gives you cumulative probabilities of the form:

$$\mathbf{P(Y \leq j) = 1 / (1 + \exp[-(\tau j - \eta)])}$$

To get the probability of an exact category, you take:

$$\mathbf{P(Y = j) = P(Y \leq j) - P(Y \leq j - 1)}$$

RESULTS

A baseline linear regression model was fitted to predict NCAA tournament seedings using three predictors: BPI, KPI, and SOR. All three variables were statistically significant predictors of Seed placement, with positive coefficients indicating that higher BPI, KPI, and SOR values (e.g., worse relative rankings) were associated with higher seed numbers (worse seeds). Specifically, SOR had the largest coefficient ($\beta = 0.091$, $p < 0.05$), followed by KPI ($\beta = 0.062$, $p < 0.05$) and BPI ($\beta = 0.036$, $p < 0.05$). The model explained approximately **86% of the variance** in Seed outcomes ($R^2 = 0.8634$), suggesting a strong linear relationship between the predictor metrics and Seed assignment. However, the model's predictive accuracy was relatively limited when Seed values were rounded to the nearest whole number: only **32% (64 / 200)** of Seeds in the training data were correctly predicted.

Figure 1: Linear Regression Table of Coefficients

Variable	Coefficient	P-Value
Intercept	1.354489	< 0.05 (Significant)
BPI	0.036234	< 0.05 (Significant)
KPI	0.061527	< 0.05 (Significant)
SOR	0.090935	< 0.05 (Significant)

Figure 2: Linear Regression Formula

$$\hat{Y} = 1.3545 + (0.0362 \cdot \text{BPI}) + (0.0615 \cdot \text{KPI}) + (0.0909 \cdot \text{SOR}) + \epsilon$$

When the baseline model was applied to teams from the 2025 season, prediction accuracy remained similar, correctly matching **30% (15 / 50)** of actual Seeds. This consistency suggests the baseline model's generalization performance was stable across both training and test sets.

Improving on the baseline linear model, an ordinal logistic regression model (cumulative link model) was fit using the same predictors. Again, BPI, KPI, and SOR were all statistically significant ($p < 0.05$), with SOR ($\beta = 0.155$) demonstrating the most substantial relationship to Seed placement. The ordinal model achieved better classification performance than the baseline linear regression. The confusion matrix conveyed a stronger agreement between predicted and actual Seed values, particularly among the top Seeds (e.g., 15 out of 16 teams were correctly classified as Seed 1) in the training data and a true accuracy of **44% (88 / 200)**. Although some misclassifications occurred, most predictions fell within ±1 of the actual Seed **(86.5%** or **173 / 200)**, highlighting the ordinal model's strength in capturing the ordered nature of the response variable, Seed. The accuracy of the ordinal model for the 2025 dataset was **54% (27 / 50)**, but when allowing for a ±1 Seed margin, accuracy increased substantially to **94% (47 / 50)**. The teams that failed to fall within a ±1 Seed margin were: Creighton (Predicted 7 Seed), Kansas (Predicted 5 Seed), and Louisville (Predicted 4 Seed).

Figure 3: Confusion Matrix Heatmaps

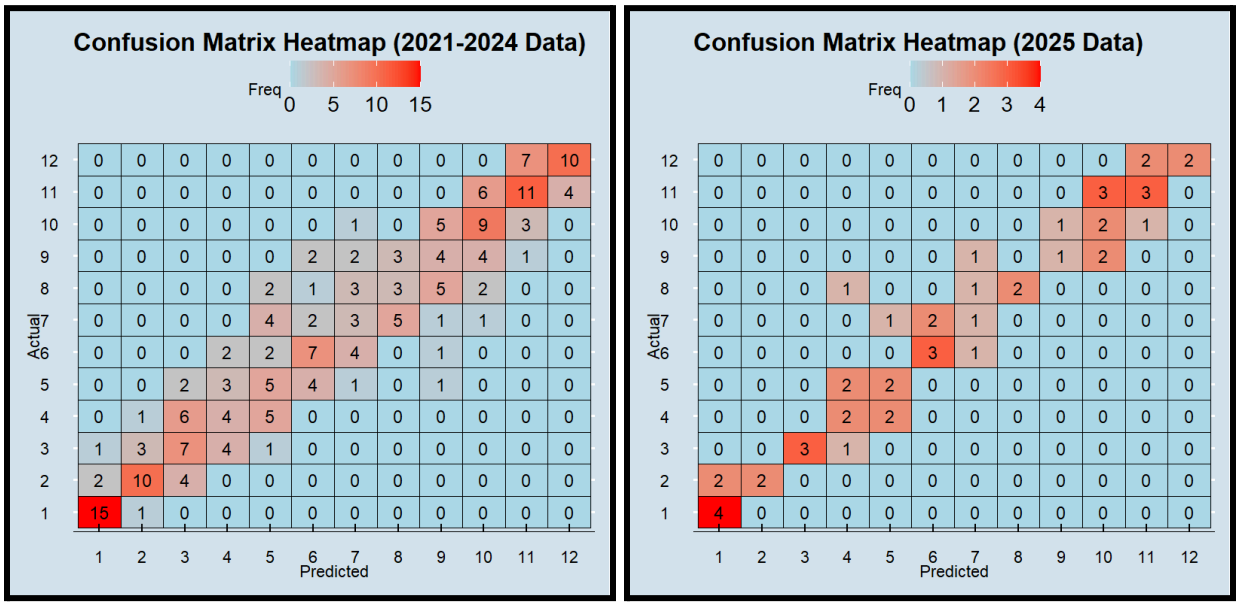

Figure 4: Ordinal Logistic Regression Table of Coefficients

Variable	Coefficient	P-Value
BPI	0.08248	< 0.05 (Significant)
KPI	0.10309	< 0.05 (Significant)
SOR	0.15493	< 0.05 (Significant)

Figure 5: Thresholds Coefficient Table

Threshold	Estimate
1 | 2	1.4075
2 | 3	3.2569
3 | 4	4.7973
4 | 5	6.2177
5 | 6	7.6620
6 | 7	8.9997
7 | 8	10.3115
8 | 9	11.6091
9 | 10	13.0552
10 | 11	15.1358
11 | 12	18.7181

Figure 6: Ordinal Logistic Regression Formula

$$\eta = (0.08248 \cdot \text{BPI}) + (0.10309 \cdot \text{KPI}) + (0.15493 \cdot \text{SOR})$$

For example, Louisville was an 8 Seed for the NCAA tournament in 2025, but predicted to be a 4 Seed. Calculating the linear predictor (η) by using the formula: $(0.08248 \cdot 28) + (0.10309 \cdot 16) + (0.15493 \cdot 11) = 5.66311$. Using the table in Figure 5, we see that the η value falls within 4.7973 and 6.2177, resulting in Louisville being predicted as a 4 Seed.

DISCUSSION

IMPLICATIONS

LIMITATIONS

FUTURE DIRECTIONS

SOURCES

- Torvik, Bart. "NCAA Teamsheet Ranks." barttorvik.com, https://www.barttorvik.com/teamsheets.php?year=2025, Accessed April 10, 2025.
- Coleman, B. & DuMond, J.. (2016). An easily implemented and accurate model for predicting NCAA tournament at-large bids. Journal of Sports Analytics. 2. 1-12. https://www.researchgate.net/publication/307615009_An_easily_implemented_and_accurate_model_for_predicting_NCAA_tournament_at-large_bids.
- Paul, R. J., & Wilson, M. (2012). Political Correctness, Selection Bias, and the NCAA Basketball Tournament. Journal of Sports Economics, 16(2), 201-213. https://doi.org/10.1177/1527002512465413 (Original work published 2015)