Latent Team Strengths using the Bradley-Terry Model on NCAA Men's Basketball Regular Season and Tournament Rankings

Gabriel Alwan, Neo Kok, Kevin Lu, Liz Orraca

Abstract:

This study applies a Bradley-Terry modeling framework to NCAA Division I Men's Basketball regular-season data from 2000-2016 to estimate latent team strengths and compare these model-derived rankings to the official NCAA Tournament seeds. Drawing on a publicly available Kaggle dataset, we filter teams by minimum game thresholds and fit logistic regression models for each season with head-to-head outcomes to derive strength rankings for each team. We then identify the top 64 teams each season and contrast their Bradley-Terry rankings with the NCAA-reported rankings, specifically focusing on conferences of teams, prediction accuracy, and seeding differences. From our results, across all seasons the models accurately seeded 5.7% of teams, with an additional 49.1% of the teams being correctly included in the tournament but seeded incorrectly. We also have an average of 45.2% of predicted teams who did not make the tournament. Across all years, 90% of the teams selected by the models originated from the Power 6 conferences. These conferences were significant at the 5% level in all models while point difference and home games proved to be not statistically significant covariates. By examining the differences between the model fits across all seasons, this approach reveals how closely a data-driven model aligns with the official seeding process and the reliability of statistical ranking methods for analyzing regular season NCAA Men's Basketball data. The incorporation of automatic conference bids is a difference between the model and the NCAA Selection Committee's criteria, which means weaker teams from smaller conferences regardless of their overall performance enter the tournament based on the current criteria. This suggests the March Madness tournament seeding selection process is not designed to be the most competitive. The reliability and accuracy of the models may provide insight into frameworks that analyze data for other sports and the validity of the selection process of teams for postseason tournaments.

Introduction:

The NCAA Division I Men's Basketball March Madness tournament is a yearly momentous sporting event. Each season, the NCAA Men's Basketball Selection Committee selects the top teams to enter the March Madness postseason tournament based on the teams' performances throughout the regular season1. The NCAA Men's Basketball Tournament teams are selected through automatic bids and at-large bids. Automatic bids are given to the 32 Division I conference champions who win their postseason tournaments, regardless of their regular season performance[1]. The Selection Committee awards 32 at-large bids to other deserving teams based on various stats and rankings. Seeding ranks teams 1-16 in each region, with higher seeds rewarded by facing lower seeds in the first round (e.g., No. 1 vs. No. 16), where the No. 1 seed is the highest seed for the best overall ranked team in the region.

To analyze the overall accuracy of the NCAA Selection Committee's selection of the top 64 teams[2], we implement the framework of the Bradley-Terry Model to assess latent strengths between pairwise comparisons of teams. The Bradley-Terry Model is a model widely used in sports, ranking systems, and preference modeling to predict the outcome of pairwise comparisons like competitions or rankings where two entities compete and one emerges as the winner. Our analysis focuses on determining if a data-driven approach (objective criteria) is consistent with the committee's criteria.

Methods:

To conduct this analysis, we utilize the NCAA Basketball database from Kaggle, which consists of three primary datasets[3]. The first dataset provides pairwise win-loss and points information for every NCAA Men's Basketball game during the 2000-2016 season. The second dataset contains detailed information on the teams like their conference. The last dataset that we use includes information on the post-season NCAA March Madness tournament like the tournament seeding.

To pre-process the data, we filter the data to ensure that the data only contains information for Division 1 teams. Then, we merge the first two datasets to combine all of the team, conference, points, and pairwise win-loss information. We then split the data to contain information for each season. We further process data to only include games of teams with more than five home games to reduce the bias of teams with very few games. This leaves us with approximately 20,000 data points per season across over 300 teams.

To determine strength rankings for each team, we implement a generalized linear model under the Bradley-Terry framework. This aims to model the relative probability of any given team beating a reference team, creating a strength ranking. We do this by creating a generalized linear model using a logit link function with a binomial family. We include covariates in the models to account for the role in point difference, home team advantage, and conference. We created a model for each season between 2000 and 2016.

The models use a maximum likelihood estimator to maximize the probability that each team i beats team j which is defined as:

$$P(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

where β_i and β_j represent the strength coefficients for teams i and j, respectively. If $\beta_i > \beta_j$, then $P(i > j)$ will be greater than 0.5, indicating that team i has a higher probability of winning the match up.

For each season, we rank each team based on the strength coefficients of the Bradley-Terry models and arrange them in order of strength. The top 64 teams are assigned "seeds" based on their strength, with the four strongest seeded as a 1, the next four seeded as a 2, and so on. This is done to allow for comparison to the true March Madness seeding, which has four of each seed. We use the third dataset to compare our data-driven rankings with the true March Madness seeds for that season. We measure the deviance of the fitted models and null models similar to the sum of squares used in ordinary linear regression. We then calculate the reduction in deviance when using the fitted models to find the reduced deviance. This measure can be thought of as the R^2 equivalent for generalized linear models and is an accurate method of evaluating model fit[4].

Results:

Figure 1: Number of Predicted Teams by Conference by Season

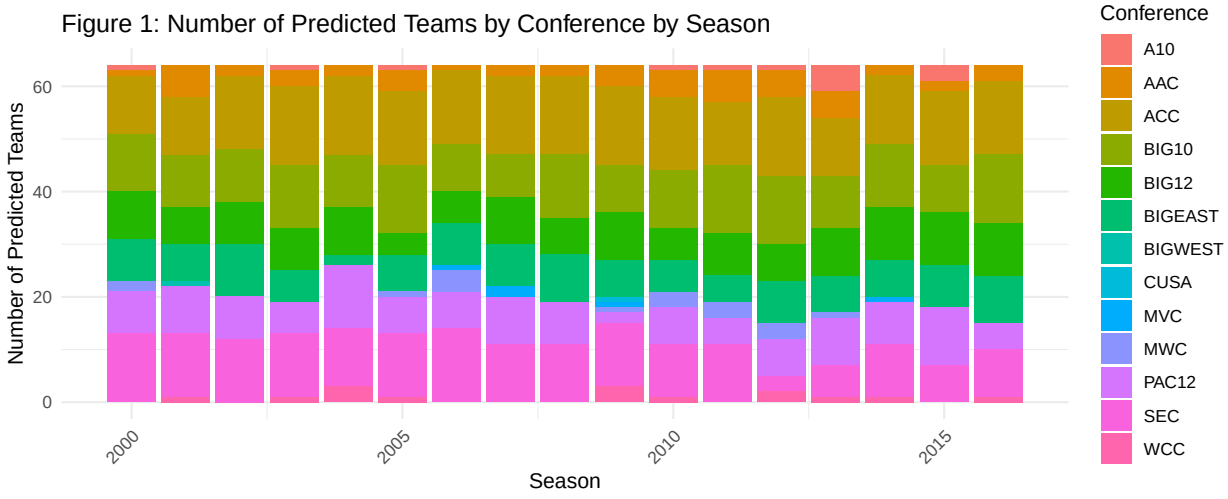

Figure 1 shows the yearly distribution of which conferences the models predicted to make the tournament. The models favor the Power 6 conferences (ACC, BIG 10, BIG 12, BIG EAST, PAC 12, SEC). Specifically, 90% of the teams the models selected came from these conferences across all years. In all of the models, these 6 top conferences were significant at a 5% threshold. Interestingly, point difference and home games were not significant.

Figure 2: Deviance Reduction by Season

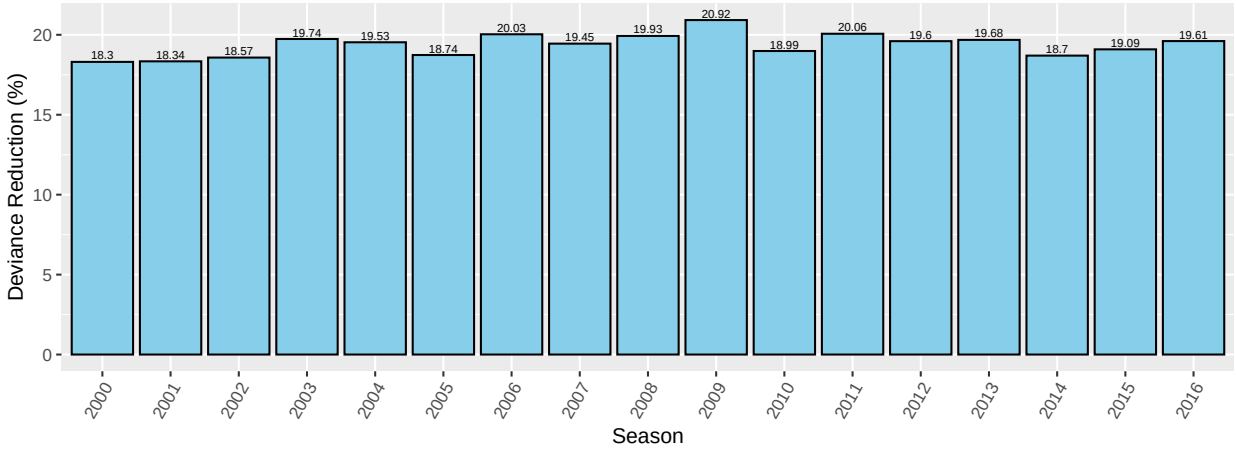

Figure 2 depicts the deviance explained by the models for each respective season. The percentages for each season hover around 20%. This indicates a moderately low fit of the models to the data.

Figure 3: Prediction Accuracy for March Madness (2000–2016)

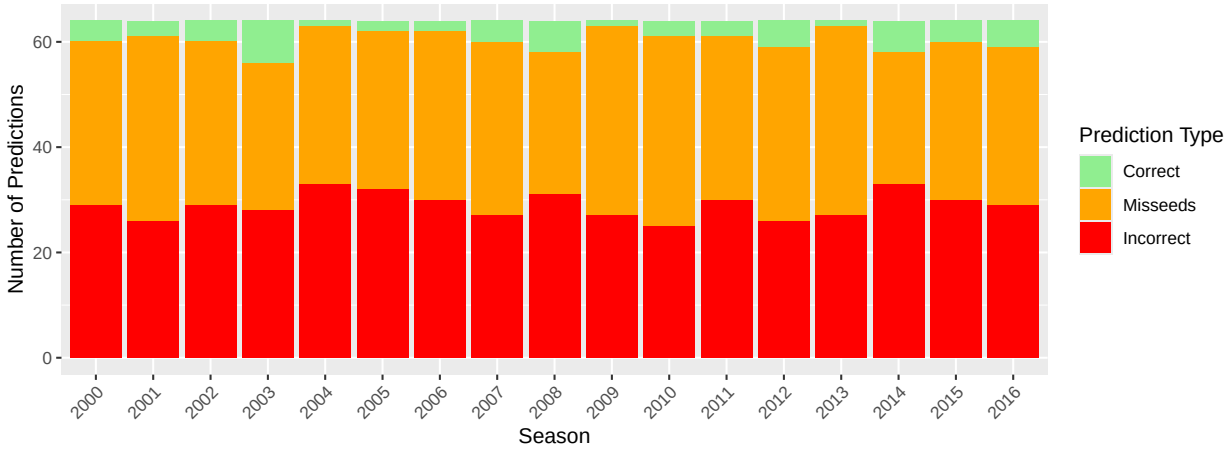

Figure 3 compares the Bradley Terry models' ranking to the true Selection Committee ranking for the top 64 teams per season. On average, approximately 5.7% are seeded exactly correctly, 49.1% are correctly included but not properly seeded, and 45.2% of teams the model ranked in the top 64 are not actually in the tournament.

Figure 4: Seeding Difference as a Percentage by Season

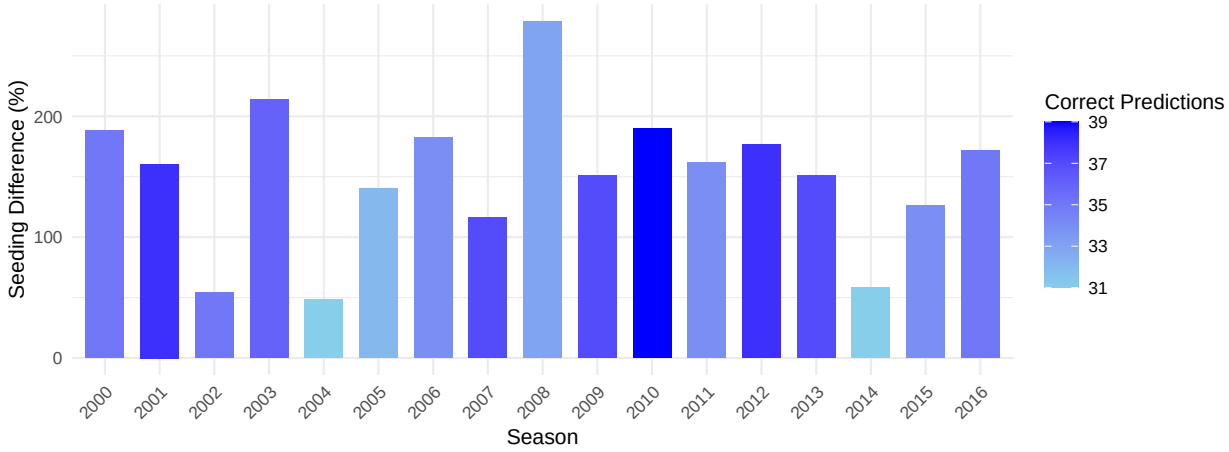

Figure 4 shows the seeding difference as a percent of correct predictions. Interestingly, in seasons where the model had fewer correct predictions in the tournament, the rankings were on average more accurate across the teams (e.g. 2004, 2014), indicating an inverse relationship. Conversely, seasons where the model correctly identified more teams for the tournament also on average had a higher number of incorrect seedings (e.g. 2001, 2010).

Discussion:

In summary, we assess the performance of Bradley-Terry models in predicting the top 64 teams and their seeding for the NCAA March Madness tournament from the seasons 2000-2016. Due to the deviance reduction being at about 20% throughout the seasons, the models have a moderately low fit. A lower deviance reduction value indicates an under-fit and perhaps a need for more covariates in the models. These factors could include injuries, a winning streak, or other variables that are harder to quantify (such as a team's playbook/style of play).

Of the 32 conferences, teams from only 13 conferences (8.9 conferences per season, on average) are ranked in the top 64 strongest teams from the models across all of the seasons combined. Of those 13 conferences, an average of 90% are part of the "Power 6 Conferences", across seasons, which are the six strongest, largest conferences in the competition. Since the official tournament seeding considers both team strengths and auto-qualified conference champions, approximately 45% of teams that the models do not include in the top 64 rankings likely come from this ignorance of the auto-qualification system.

Due to the automatic conference bid selection, many of the smaller, weaker conferences have an included team within the tournament regardless of their actual season performance compared to other teams in more difficult conferences. For example, Gonzaga is a team in a weaker conference (WCC) but makes the March Madness tournament every season across this 16-season period, but only shows up 9 times in the top 64 based on the models in the same period.

Surprisingly, the effect of point differential and home-field advantage did not significantly influence the strength ranking of the models. This could be due to possible collinearity or simply a lack of influence on the models. The effect of a conference on the models was significant at the 0.05 level for all six of the Power 6 conference teams. This is as expected because these conferences are considerably better than non-Power 6 conferences and would therefore give higher strength coefficients to Power 6 conference teams to account for losses against other Power 6 conference teams.

Conclusion:

While the models have relatively low explanatory power for estimating NCAA March Madness seeding, they successfully identify strong teams, especially in stronger conferences. The absence of automatic qualification criteria in the Bradley-Terry model framework makes it inherently difficult to accurately assess the relationship between the model rankings and the true seeding. In the future, we propose implementing the automatic qualification criteria to better evaluate the comparison between the model rankings and the true rankings. We also propose incorporating other covariates to more accurately evaluate the team's strengths. More recent data would also allow for more confidence in generalizing the model's ability.

The disconnect between the models' strength rankings and true seeding highlights a concern about the competitiveness of the NCAA March Madness tournament. With the current system, the Selection Committee is forced to include subpar teams to cater to smaller schools. With only approximately half of the tournament contenders ranking in the top 64 of the model rankings indicates that the tournament is not as competitive as possible. Most of the teams that were in the tournament but not in the top 64 in the models' rankings are teams outside of the Power 6 conferences, further supporting the perspective of a less competitive tournament. These results serve as a recommendation to the Selection Committee to remove or reduce the automatic qualification system to increase the competitiveness of the tournament.

Contributions

Gabe Alwan SQL queried the data along with working on the results section which included coding to get the numbers and the graphs. Elizabeth Orraca created the team final seeding dataset for comparison with our model predictions and summarized the results in the discussion and conclusion section. Neo Kok processed the data, created the models, conducted analyses, and wrote the methods section. Kevin Lu wrote the abstract, introduction, and compiled references for background information.

GitHub link: https://github.com/neokok/NCAABradleyTerry

References

[1] Huang, T.-K., Weng , R. C., & Lin, C.-J. (2006, October 5). Generalized Bradley-Terry Models and Multi-Class Probability Estimates. https://dl.acm.org/doi/pdf/10.5555/1248547.1248551

[2] Wilco, D. (2024, March 7). What is March Madness: The NCAA tournament explained. NCAA. https://www.ncaa.com/news/basketball-men/bracketiq/2023-03-15/what-march-madness-ncaa-tournament-explained

[3] NCAA. NCAA basketball dataset. Kaggle. https://www.kaggle.com/datasets/ncaa/ncaa-basketball/data?select=mbb_historical_teams_games

[4] Stefánsson, G. (1996). Analysis of groundfish survey abundance data: Combining the GLM and delta approaches. ICES Journal of Marine Science, 53(3), 577–588. https://doi.org/10.1006/jmsc.1996.0079