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. By examining the differences between the model fits across sixteen 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 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 season (Wilco 2024). 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 (Wilco 2024). The Selection Committee awards 36 at-large bids to other deserving teams based on various stats and rankings. After the field of 68 teams is finalized, teams are ranked 1-68 and assigned to one of four regions. The top 64 teams proceed after the "First Four" elimination round. 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 after the "First Four", we implemented 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 (Huang et al 2006). Our analysis uses publicly available datasets from Kaggle with regular season data of the NCAA Men's Basketball results. We are ultimately curious if a data-driven approach (objective criteria) is consistent with the committee's criteria.

Methods:

To conduct this analysis, we utilized the NCAA Basketball database from Kaggle, which consisted of three primary datasets. The first dataset provided pairwise win-loss and points information for every NCAA Men's Basketball game during the 2000-2016 season. The second dataset contained detailed information on the teams like their conference. The last dataset that we used included information on the post-season NCAA

March Madness tournament like the tournament seeding. To pre-process the data, we filtered the data to ensure that the data only contained information for Division 1 teams. Then, we combined 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. This data was further processed to only include games of teams with more than five home games to reduce the bias of teams with very few games. This left us with approximately 20,000 data points per season across over 300 teams. To determine strength rankings for each team, we implemented 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 included covariates in the model 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 model uses 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 matchup. For each season, each team was ranked based on the strength coefficients of the Bradley-Terry model and arranged in order of strength. The top 64 teams were 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 measured the deviance of the fitted model and a null model similar to the sum of squares used in ordinary linear regression. We then calculated the reduction in deviance when using the fitted model 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.

Results:

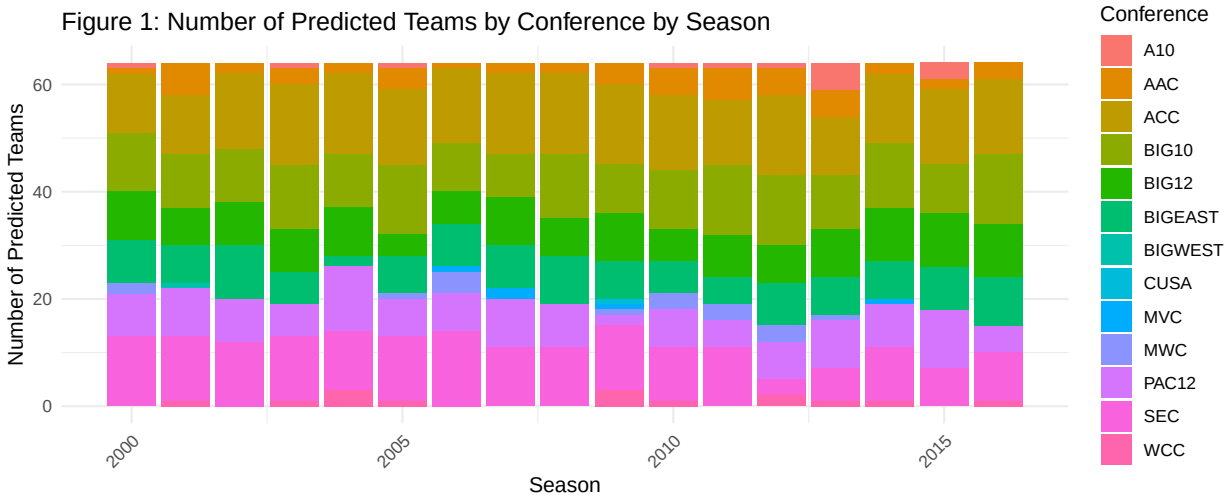

Figure 1 shows the yearly distribution of which conferences our model predicted to make the tournament. As we can see, our model favors the Power 6 conferences (ACC, BIG 10, BIG 12, BIG EAST, PAC 12, SEC). Specifically, 90% of the teams the model selected came from these conferences across all years.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

Figure 2: Deviance Reduction by Season

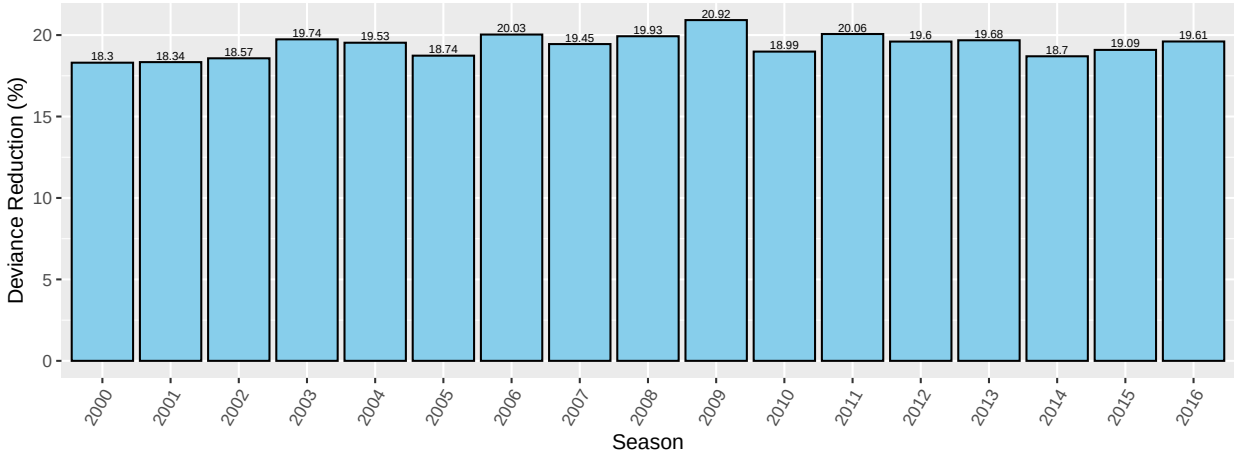

Figure 2 depicts the deviance explained by the model for each respective season. The percentages for each season hover around 20%. This indicates a moderately low fit of the model to the data.

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

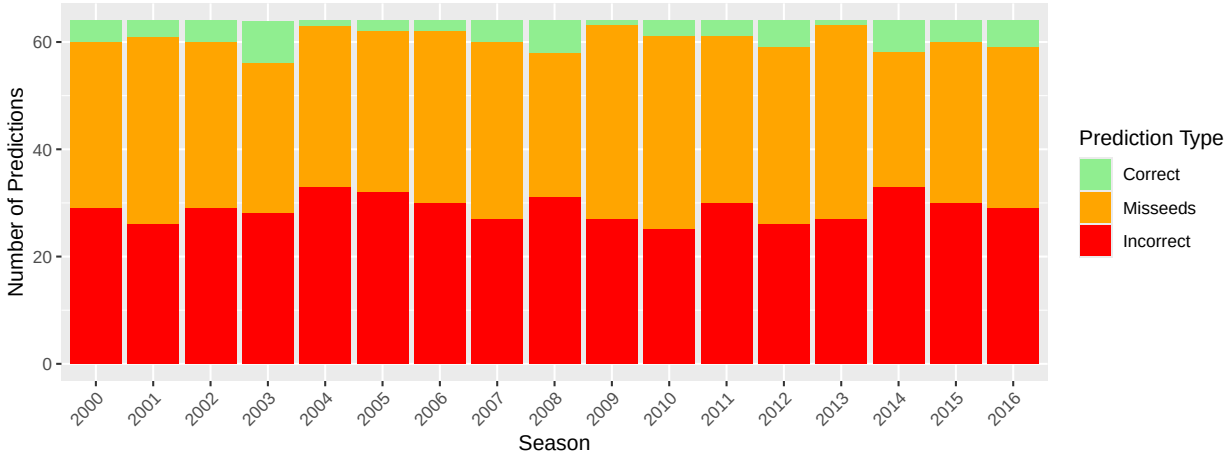

Figure 3 compares the Bradley Terry model ranking to the true Selection Committee ranking for the top 64 teams per season. On average, approximately 5.7% were seeded exactly correctly, 49.1% were correctly included but not properly seeded, and 45.2% of teams our model ranked in the top 64 weren't actually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

Figure 4: Seeding Difference as a Percentage by Season

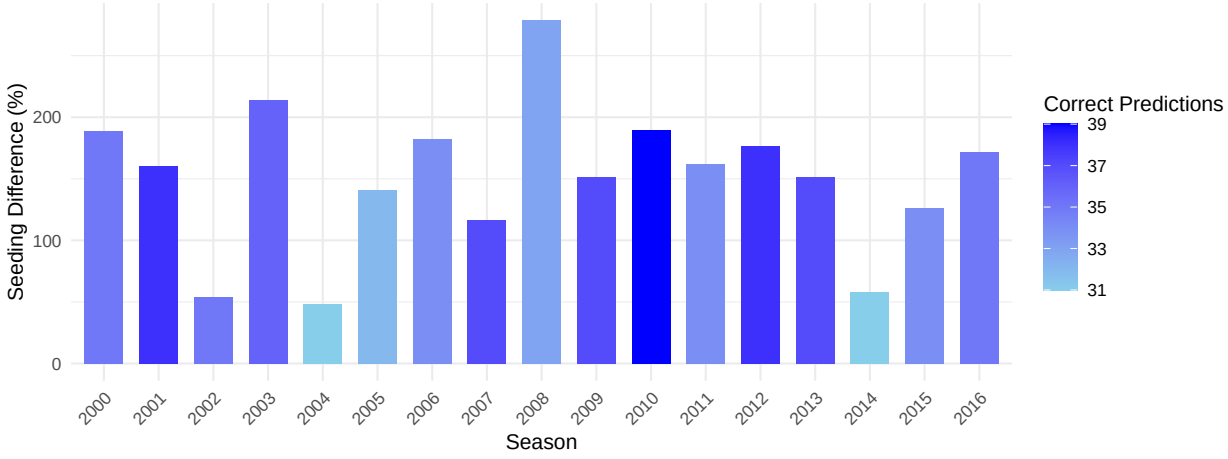

Figure 4 shows the seeding difference as a percent of correct predictions. Interestingly, in seasons where the model had fewer correct predictions in the tournament, our seedings were on average more accurate across the teams (e.g. 2004, 2014), indicating an inverse relationship. Conversely, seasons 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 assessed the performance of a Bradley-Terry model in predicting the top 64 teams and their seeding for the NCAA March Madness tournament from the seasons 2000-2016. Due to our deviance reduction being at about 20% throughout the seasons, which highlights how much better the model performs under the null, the model has a moderately low fit. A lower deviance explained value indicates an under-fit and perhaps a need for more covariates in the model. 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 our model 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 conferences in the competition. Big differences between the model's top 64 teams' strength rankings and the true NCAA March Madness tournament seeding are shown here. Since the official tournament seeding considers both team strengths and auto-qualified conference champions, approximately 45% of teams that the model does not include in the top 64 rankings 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. Despite this limitation, the model consistently identifies the stronger teams within the major conferences.

Surprisingly, the effect of point differential and home-field advantage did not significantly influence the strength ranking of the model. This could be due to possible collinearity or simply a lack of influence on the model. The effect of a conference on the model 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 model has relatively low explanatory power for estimating NCAA March Madness seeding, it successfully identifies strong teams, especially in stronger conferences. The absence of automatic qualification criteria in the Bradley-Terry model 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 competition.