Bidding War: Insight on Making the NCAA Division 1 Basketball Tournament

Michael Dixon, Ashley Hansen, Karan Pandya, Lauren Zengeler

5/8/2022

Abstract

Introduction

March Madness–the annual Division 1 college basketball tournament that consists of 68 national teams–is the National Collegiate Athletic Association's (NCAA) largest revenue maker. On average each year, the men's tournament brings in $800 million dollars, mostly through the television rights to the tournament, constituting 72% of the association's total annual revenue (Dosh, 2021; Blinder, 2022). Of this, on average $600 million dollars goes into the NCAA Basketball Fund, which is how each school that makes the tournament makes money from it (Dosh, 2021). This Fund comprises the Equal Conference Fund and the Basketball Performance Fund. The Equal Conference Fund determines how much money a school's conference–of which there are 32 total–will generate in the tournament based on how many teams make the tournament itself (Steinberg, 2022). The Basketball Performance Fund extends these earnings into the tournament, where as teams advance in the tournament they will make more money for their conference (Steinberg, 2022). There is a unit is awarded to each automatic qualifier (Equal Conference Fund) and one to each at-large berth (Basketball Performance Fund) and then an additional unit is awarded for each game a team wins, but not the championship game (Dosh, 2021). Thus, the total number of units–each one estimated to be worth just over $\$300,000$–a singular team can earn is six and the number of units a conference can earn is based on how many individual teams a conference is able to send to the tournament (Steinberg, 2022). The revenue from this fund is distributed based on a six-year rolling period–one unit will be paid out to a conference once per year for six years starting the year following the tournament (Dosh, 2021). The conference is then able to distribute the money how it pleases to each college, but it is suggested that they distribute it equally among all of the schools in the conference (Blinder, 2022).

The revenue a conference brings in is crucial to the funding and success of the schools in that conference's athletic departments in the future year, as this is a major way in which they get funding. Additionally, making the tournament can increase the public perception of a school, especially if they go far in the tournament (Resch, 2012). The attention gained from performing well in the tournament cannot be taken for granted when the school is smaller and lesser known than the big names (Resch, 2012; Lopresti, 2022). For example, in the 2022 NCAA tournament, Saint Peter's University was the underdog and pulled off a Cinderella run, being the first 15 seed to make the Elite Eight and reached unmatched popularity throughout the tournament, becoming a well-known school that most viewers had only vaguely heard of (Hutchinson, 2022). Further, revenue from the NCAA tournament can lead to increased overall donations and influence the chance of recruiting future players (Resch, 2012; Humphreys and Mondello, 2007). The athletic program of a college depends on how successful they are at recruiting new players, but there is bias in this process when players are unconsciously learning in favor of the school that has recently and historically played better (Resch, 2012). This, combined with future funds to be used for the department, makes it crucial to be able to understand which teams may have the chance to compete in the NCAA tournament.

Our question of interest in this report is to investigate if we can predict a team's chance of making the NCAA men's March Madness Tournament based on past performance (free throw rate, adjusted defensive efficiency, and adjusted offensive efficiency) and their conference. This is a strong way in which to investigate March Madness, as a team making the tournament can be crucial in the future success of not only their

athletic department funding, but in the reputation of the school overall. This is especially important in considering teams that do not come from the dominating conferences. While we are not predicting how well they do in the tournament, predicting their chances of making the tournament can provide insight into if they have the opportunity to pull off a Cinderella run and the chance at the monetary and social success that follows this.

Methods

We found our data set entitled "College Basketball Dataset" on Kaggle. This dataset was published by Andrew Sundberg in 2021 and includes data from the 2013 to 2019 Division I seasons. Data for the 2020 and 2021 seasons were not included in our data file due to the COVID pandemic. To produce this data set, data was extracted from a set on Barttorvik; then postseason, seed, and year were added as columns (Sundberg, 2021). The data includes information on games won, games played, conference, how successful the team was in the playoffs, and a multitude of statistics that reflect the strength of the teams. The variables we chose for our model were free throw rate, conference ranking, adjusted offensive efficiency, and the adjusted defensive efficiency. The adjusted offensive efficiency (ADJOE) was the estimated points scored per 100 possessions against an average Division I defense. For this variable, the higher this statistic, the better the team. The adjusted defensive efficiency (ADJDE) was the estimated points allowed per 100 possessions against average Division I offense. Therefore, for this variable, the lower the statistic, the better the team. These four variables were selected for our model, since they are variables that do not double count certain statistics in basketball. There was no binary outcome variable in the data set; thus, we created a new variable, outcome, by assigning every team that made it to the tournament a 1 and every team that did not make it to the tournament a 0. The data set had a column labeled conference where each team's conference was listed, but it had no numerical value to it. Therefore, we created a new variable, conference ranking, by assigning each conference in the data set a rank (based on data collected from Team Rankings in the 2021-2022 seasons). Instead of using a continuous predictor from 1 to 32 for each conference, we instead created a categorical variable of 40 for the top 8 conferences, 30 for the next 8 strongest, 20 for the following 8, and 10 for the bottom quartile of conferences. This is to better reflect the larger gaps between the dominant conferences and the weaker ones, while nullifying the differences between conferences that are of a similar level of strength. A continuous predictor would assume that the differences between the ranks of the conferences are the same, which is not the case when comparing the higher ranked conferences to the lower ranked ones.

We used a logistic model, $log(\frac{p_x}{1-p_x}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ where p_x is the probability of a team making it to the tournament given the covariates, X_1 is the free throw rate, X_2 is the conference ranking, X_3 is the adjusted offensive efficiency, and X_4 is the adjusted defensive efficiency. We also created a prediction model, $logit(\hat{p_x}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$, that we used to predict the probability of a team making it to the tournament based on certain values for each covariate.

In order to use this model, we assessed the following assumptions: linearity in the log odds, independence, distribution assumptions, and a large sample size. Linearity in the log odds implies that each predictor is linearly associated with the log odds of the outcome event. For continuous predictors, this assumption indicates a one-unit change in the odds ratio is constant for every value in that predictor, as well as constant across values of any other predictor. The linearity of log odds was assessed by looking at the empirical logit plots for the variables. Since all of our plots were roughly linear, this assumption was met. The independence assumption means the outcomes for each observation are independent of other outcomes when the predictors are held constant. In other words, if one team makes it to the tournament, it does not make it more or less likely that another team makes it. Sources in our data that would cause non-independence include cluster sampling, missing covariates, network effects, and measurement error. Since the performance of a team is not affected by other teams, our independence assumption was met. The distribution assumption assumes that we have a binary outcome with a Bernoulli distribution. Since we labeled our outcome variable as 1 if the team makes it to the tournament and 0 if the team does not make it to the tournament, then we do indeed have a binary outcome; thus, our Bernoulli distribution condition is met. Lastly, we met the large sample size assumption, since we have a total of 2455 observations, which is much greater than the minimum of 10 observations.

In this model, we were testing if the probability of making it to the tournament is associated with four

different variables: the free throw rate, the conference ranking, the adjusted offensive efficiency, and the adjusted defensive efficiency. We ran a hypothesis test to test whether there is a significant association between the log odds of making it to the tournament and our covariates at a significance level of $\alpha = 0.05$. Therefore, our null hypothesis assumes that none of these predictors are associated with the probability of making it to the tournament, while the alternative hypothesis assumes that at least one of these variables are significantly associated with making it to the tournament. $H_O : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ H_A : at least one of $\beta_1, \beta_2, \beta_3, \beta_4 \neq 0$

Results

Predictions

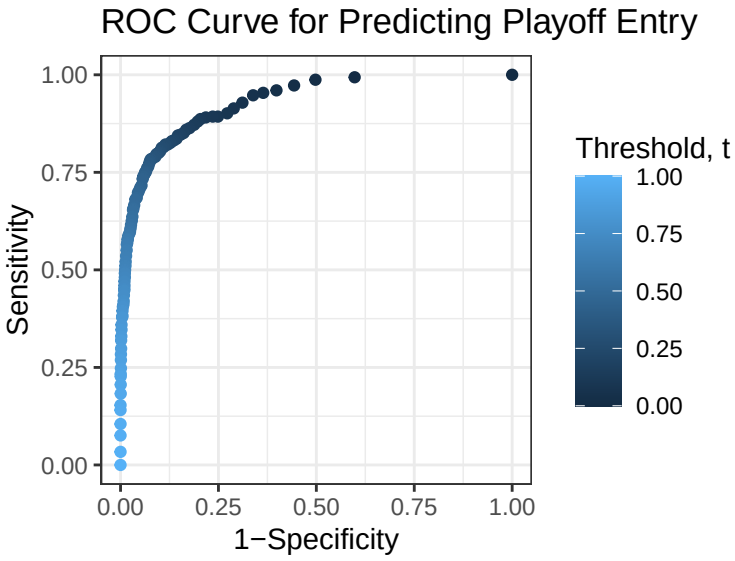
Table 1: Predicted Risk Measures from Logistic Regression Model

FTR	CONFRank	ADJOE	ADJDE	Pred.LogOdds	Pred.Odds	Pred.Probs
35	40	120	98.2	2.147	8.557	0.895

Table 2: Predicted Risk Measures from Logistic Regression Model

FTR	CONFRank	ADJOE	ADJDE	Pred.LogOdds	Pred.Odds	Pred.Probs
30	30	102	90	-0.46	0.631	0.387

Here we created two teams, one who made the playoffs (on top) and one who didn't (on bottom). For the team that made the playoffs, our model predcits that they had a 89.5% chance of making the playoffs. This is in line with what we expect since a team that makes the playoffs should have a higher predicted probability. For the team that didn't make the playoffs, our model predicted that they had a 38.7% chance at making the playoffs. This is also in line with what we would expect since a team that didn't make the playoffs should have a lower predicted probability. In both cases the model is accurately predicting the chances of a team making the playoffs. To further assess the predictive accuracy of our model we plot an ROC curve.



The ROC curve shows that our model is very accurately predicting whether or not a team makes the playoffs based off the covariates in our models. The closer the area underneath the ROC Curve is to 1 the

better the model is at accurately classifying whether a team makes it into the playoffs or not. This ROC curve has an area of 0.931. This indicates, once again, that our model is correctly classifying each team, for most teams.

Discussion

The model created has a strong correlation between all the statistics chosen and predicting receiving a bid in the NCAA tournament for Division 1 basketball. Furthermore, our ROC curve predicts a high degree of accuracy given that it has a high degree of sensitivity for any specificity. We get the desired logistic curve of the ROC, meaning we think it is more accurate than a random classifier. Furthermore, all our predictors were significant at the $\alpha = .05$. Adjusting with our new conference ranking to account for a bigger disparity between the best and the worst, we still see a negative association between being in a "weaker" conference and making the tournament. We can see through our data that traditional statistics such as free throw rate and conference rankings have their place with newer, more advanced metrics such as offensive and defensive efficiency. Finally, we are confident all our LINE assumptions are met and therefore MLE inference is valid on the data.

However there are some key limitations to our data. We treat all bids to the tournament as equivalent, but that is not really true since 32 teams get automatic bids for winning their conference tournament, then the rest of the 36 receive "at large" bids, meaning a selection committee chooses them. This tends to stronger, bigger conference teams getting the at large bids more so than teams from smaller conferences. Therefore future research should be done on the likelihood of a smaller team receiving an at large bid, meaning they were recognized for their skill and talent instead of just automatically qualifying due to the rules of the tournament. Another limitation is we only used two statistics that are based on relative strength, adjusted offensive and adjusted defensive ratings. There is an equation referred to as "BARTHAG" that attempts to take the difference in these two values, adjust it for league average and call it a team's "net rating" (meaning you take the expected amount of points scored, expected amount of points allowed, adjust it for playing an average team and that is the expected margin of the game, i.e. how much they win or lose by). Using something more comprehensive like that could have allowed us to also add in more statistics that show relative strength of a team. Free throw rate is just an absolute value, it does not show how much better a team is at drawing free throws than another one. In a study like this where relative strength seems paramount to getting a bid, relative strength statistics seem more important.

For future studies, we would also like to see how past success of a smaller conference school can influence its future success. Every year, there is a school from a smaller conference that goes further in the tournament than expected, often referred to as a "Cinderella" team. This year, that team was St. Peters from Jersey City, a team with an incredibly small budget compared to other Division 1 teams yet was able to defeat perennial powerhouses Kentucky and Purdue. However, after their successful run their coach took a job at Seton Hall, a much bigger school and several players have now transferred out of the school in pursuit of better opportunities. It seems unlikely for St. Peters to repeat their run again. In contrast, another small school, Loyola Chicago, made a similar run in 2018, and recently made the tournament again. The reverse of the idea is that a school like Duke, probably the most dominant basketball school in the country, has only missed the tournament once since 2009. To miss the tournament for Duke is a major disappointment. A study on how past success in the tournament influences future bids and success within the tournament could also be an area of deep interest.

In conclusion, our study seems to show a high degree of correlation between both performance of a team and the conference within which it plays. Therefore the intuitive idea that "blue blood" schools and conferences have a major advantage over the smaller schools not just in the quality of player they can recruit but from the selection committee seems vindicated. Furthermore, future studies can be conducted to test what seems to be a cyclical nature of success, that is a team is successful, there it becomes even more successful due to past success and so on.

References

Dosh, K. (2021, March 27). *NCAA has confirmed its basketball revenue distribution plan for 2021, but impact of 2020 will linger.* Forbes. Retrieved April 26, 2022, from https://www.forbes.com/sites/kristidosh/2021/02/28/ncaa-has-confirmed-its-basketball-revenue-distribution-plan-for-2021-but-impact-of-2020-will-last/?sh=5c5bc147391a

Humphreys, B. R., & Mondello, M. (2007). *Intercollegiate athletic success and donations at NCAA Division I institutions.* Journal of Sport Management, 21(2), 265-280.

Hutchinson, B. (2022, March 27). *Underdog Saint Peter's ends unprecedented March Madness run with Elite Eight loss.* ABC News. Retrieved May 8, 2022, from https://abcnews.go.com/Sports/underdog-saint-peters-seeks-beat-unc-continue-unprecedented/story?id=83700409

Lopresti, M. (2022, March 26). *There's a strong case Saint Peter's is already the greatest Cinderella ever.* NCAA. Retrieved May 8, 2022, from https://www.ncaa.com/news/basketball-men/article/2022-03-26/strong-case-saint-peters-already-greatest-cinderella-ever

Resch, R. (2017, September 27). *March madness: The importance of the tournaments to the schools.* Bleacher Report. Retrieved April 26, 2022, from https://bleacherreport.com/articles/1093572-march-madness-the-importance-of-the-tournaments-to-the-schools#:~:text=To%20all%20of%20the%20colleges,shine%20and%20ga

Sundberg, A. (2021, March 16). *College basketball dataset.* Kaggle. Retrieved April 26, 2022, from https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset?resource=download

Steinberg, R. (2022, March 14). *How much is a men's NCAA Tournament win worth?* Boardroom. Retrieved April 26, 2022, from https://boardroom.tv/mens-ncaa-tournament-distribution-unit-worth/