Modeling Selection Sunday: An Investigation into the "March Madness" Bracket

Jonathan Olds

Summer 2020

Abstract

Every season, 68 college basketball teams are invited to the NCAA Men's Basketball Championship Tournament, known as "March Madness." Thirty-two teams are given automatic bids, while 36 teams are chosen by a selection committee. Though a broad overview of the process and factors used by the committee to make these selections are known, the exact details of this process and the factors that the committee has determined to be most important are largely unknown. This paper serves as the exploration of an attempt to accurately replicate the selection and seeding process of these teams by the committee through the fitting of a lasso penalized regression model. Across five seasons of data from 2014 to 2019, this model accurately awarded 130 out of 144 potential at-large bids to tournament teams and placed 152 out of a possible 340 teams at their correct seed, while placing 271 within one seed of their correct place.

1 Introduction

The NCAA Men's Basketball Championship Tournament, more commonly known as "March Madness," is one of the biggest spectacles in the sports calendar year. In 2019, the NCAA earned nearly a billion dollars in revenue from media rights, ticket sales, corporate sponsorship and television advertisements throughout the three-week-long tournament. This value represented approximately 75% of their annual revenue [Parker, 2019]. The tournament has massive effects on the gambling industry as well, as the American Gambling Association estimates that Americans filled out 149 million brackets attempting to guess the winner of the tournament and gambled $8.5 billion overall in 2019, which is 25% more than was wagered on the Super Bowl [Kiernan, 2020]. While approximately 175 million fans engage with the tournament across various platforms throughout its duration, the championship game is consistently the biggest attraction, drawing 19.6 million viewers in 2019 [Adgate, 2019, Kiernan, 2020].

March Madness is much more than a single championship game. Sixty-eight teams are chosen each year to compete in the single-elimination tournament. These teams, each of which represents a Division I collegiate institution, are selected, seeded and placed into the tournament bracket by a selection committee. This selection and seeding process occurs on the second or third Sunday of March, a day dubbed “Selection Sunday.” The tournament bracket consists of four regions: Midwest, East, South, and West. Each region is comprised of 16 teams seeded from one (highest) to 16 (lowest). Sixty teams are given a place directly in this bracket, while the other four teams earn their place by winning a game in the “First Four” round, a play-in tournament. At the conclusion of the tournament, the four teams that have won their regions compete in the “Final Four,” consisting of two national semifinals. The two remaining teams compete in a championship game; the winner of that game is crowned the champion of college basketball.

The selection of these 68 teams is a process unique to most sports. In any of the four major sports (NFL, NBA, MLB, NHL), teams make their leagues' postseason tournament based solely on performance. If a team has more wins (or points in the NHL) than enough other teams, they will earn a playoff berth. In contrast, the process of earning a spot in March Madness is much more convoluted. Thirty-two of the 68 open slots are given to the teams that win their conference tournaments in early March. These teams receive automatic bids. The Selection Committee then uses a variety of factors to choose the teams that will earn the remaining thirty-six at-large bids.

The NCAA Selection Committee is known to include several factors in their analysis of the 321 other teams (319 until 2019) to determine the 36 who will earn an at-large bid in the tournament. Among these factors are overall record, NET rating and strength of schedule, though this is far from a comprehensive list [nca, 2020]. This understanding reveals two important facts: (1) the selection process for earning a postseason berth in college basketball relies heavily on human decision-making, and (2) this decision-making process involves the consideration of at least some factors not known to the public. Thus, this project aims to replicate the selection process with statistical modeling.

Many professional and amateur college basketball analysts attempt to predict the tournament field each year. Joe Lunardi is perhaps the most well-known such analyst, who has dubbed this prediction process “bracketology” Negron [2021]. However, from an academic standpoint, most published studies have been centered around predicting the tournament outcome, rather than the tournament field. For example, Wright [2012] utilized probit and ordinary least squares (OLS) regression models to determine significant variables on the outcome of tournament games and test their predictive ability on accurately selecting winners. Ji et al. [2015] looked for a more extensive approach to the same question, as they created a matrix completion approach to estimate the potential performance accomplishments by tournament teams, and then converted these performance accomplishments into game scores of each possible tournament matchup using neural network methods. From that point, win probabilities were derived through various probability adjustments.

To our knowledge, there is only one published study that has academically ventured to predict the tournament field. In 2018, Dutta and Jacobson [2018] created a decision tree that utilized pairwise comparisons to mimic the selection of teams given at-large bids at a seed of 10 or higher. The candidates for potential selection were teams that earned an at-large bid and were seeded at 10 or above, as well as the 1 and 2 seeds in the National Invitation Tournament (NIT). The NIT is another tournament operated by the NCAA, but it is seen as a sort of consolation competition because the teams typically invited to the NIT are the ones that narrowly missed qualifying for the NCAA Tournament. In other words, the NIT gives teams who were among those considered for at-large March Madness berths, but ultimately denied selection the opportunity to compete in a smaller, less-lucrative tournament. coa [2020]. Thus, teams given high seeds in the NIT Tournament are of comparable quality to the lower at-large seeds in the March Madness bracket. This decision tree compared each possible combination of two teams on factors such as the Ratings Percentage Index (RPI), Pomeroy rankings, and other basic statistics. RPI and Pomeroy are different metrics designed to summarize team quality and rank teams accordingly. RPI primarily utilizes winning percentage and strength of schedule, while the backbone of Pomeroy ratings is adjusted efficiency Sukup, Flaherty [2020]. Overall, up to 11 different branches were used until a significant difference was found to determine the better team. After all pairwise comparisons were completed, the teams with the most pairwise victories were selected for the tournament. This method was found to be successful, as in each season from 2012-2016, all but one team selected by the decision tree was invited to the tournament [Dutta and Jacobson, 2018].

This study serves as an introduction into the replication of the Selection Committee's process. However, this study did not apply the decision tree to teams given at-large bids that were seeded below 10, and notably did not attempt to seed the teams selected. Thus, the purpose of this study is to find a replication method that can be used to accurately select and seed 68 teams from the field of current D-I teams (currently 353).

1.1 Data

Since the goal of this study was to replicate the selection of teams for the March Madness tournament, we needed to gather data for each team from prior to Selection Sunday each season. This posed a significant challenge in data collection because most college basketball statistics websites are updated weekly, but the required information was retroactive three or four weeks from the most current data. Even with this challenge, we were able to collect complete information on the variables of interest for each Division-I team, divided by season from 2014-15 to 2019-2020.

The majority of the data were collected through communication with an amateur college basketball statistician, Bart Torvik, who compiles college basketball data on his website [Torvik, 2020], while other data were collected from various pages on the college basketball data website Sports Reference [SR 1, 2020, SR 2, 2020, SR 3, 2020, SR 4, 2020]. The Team Rankings and Warren Nolan sites were utilized to collect the remaining data [Team Rankings, 2020, Warren Nolan, 2020]. Table 1 below describes each covariate and their source.

Abbr.	**Covariate**	**Description**	**Source**
Conf	Conference	Which of the 32 conferences each team belongs to	Torvik [2020]
Record	Record	Each team's record as of Selection Sunday	Torvik [2020]
Adj.OffEff	Adjusted Offensive Efficiency	An estimate of the number of points scored per 100 possessions a team would have against the average D-I defense [Pomeroy, 2002]	Torvik [2020]
Adj.DefEff	Adjusted Defensive Efficiency	An estimate of the number of points given up per 100 possessions a team would have against the average D-I offense [Pomeroy, 2002]	Torvik [2020]
Barthag	Barthag Metric	Bart Torvik's use of adjusted offensive and defensive efficiency and the pythagorean expectation method to create a ranking system similar to the Pomeroy rankings [Torvik]	Torvik [2020]
SOS	Strength of Schedule	A metric of schedule difficulty based on the winning percentage of each team's opponents, calculated for conference games, non-conference games and overall schedule	Torvik [2020]
ConfRecord	Conference Record	Each team's record in games against teams inside their own conference	Torvik [2020]
WAB	Wins Above Bubble Metric	A metric created by Bart Torvik that compares each team's number of wins to the number of wins that an average 'bubble' team would be expected to have with that team's schedule [TheSabre.com, 2020]	Torvik [2020]
EffRankAvg	Efficiency Rank Average	The average of each team's adjusted offensive efficiency rank and adjusted defensive efficiency rank	Torvik [2020]
Power5	Power 5 Indicator	An indicator variable denoting 1 if that team is a member of a 'Power 5' conference - Atlantic Coast Conference (ACC), Big 10 Conference (B10), Big 12 Conference (B12), Southeastern Conference (SEC) and Pacific 12 (Pac-12) - and 0 otherwise	Torvik [2020]
WinP	Win Percentage	Each team's number of wins divided by games played	Torvik [2020]

Abbr.	Covariate	Description	Source
NET	NET Ranking	Each team's ranking according to the NCAA Evaluation Tool, the primary metric used by the Selection Committee since 2019	Warren Nolan [2020]
MakeTourn*	Make Tournament Indicator	An indicator variable denoting 1 if that team made the tournament, and a 0 otherwise	SR 1 [2020]
ConfChamp	Conference Champion Indicator	An indicator variable denoting 1 if that team won their conference tournament and a 0 otherwise	SR 2 [2020]
L12W	Last 12 Wins	The number of wins that each team attained in their final twelve regular season games	SR 3 [2020]
ConfFin	Conference Finish	Each team's ending rank in their respective conferences at the end the regular season	SR 4 [2020]
RPI	RPI Ranking	Each team's ranking according to the Ratings Percentage Index, the primary metric used by the Selection Committee until 2018	Team Rankings [2020]

Table 1: All relevant variables obtained in data collection. The rows highlighted in yellow were deemed both potentially significant through exploratory data analysis and appropriate to include in model fitting based on factors such as what the variable measured and how long the variable was in use.
*Note that MakeTourn represents the variable of interest, not a covariate

2 Methods

Given that 32 teams each year are given automatic bids to the tournament for winning their respective conference tournament, it was decided that including these teams would be irrelevant in the goal of predicting which 36 teams the selection committee selects for at-large bids. Thus, each season's dataset was reduced to include only the teams eligible for at-large bids prior to conducting any analysis.

It was initially of interest to determine which variables were related to a team earning or not earning a tournament bid, given the assumption that each variable that appeared to be associated with a team earning a bid were relevant to the Selection Committee's selection process and should thus be included in a model. Based on an in-depth exploration of the covariates and their graphical relationships with the Make Tournament indicator, the variables highlighted in Table 1 were selected in some form for model fitting.

The belief that each of the highlighted variables were significantly associated with earning a tournament bid was tested by running univariate logistic regression models in R of the form seen in Model 1 below. [R Core Team, 2020]:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \mathrm{X}_{i1} \tag{1}$$

where p represents the probability of a specific team making the tournament. Each variable was determined to be statistically significant in their individual models with $p < 0.001$. However, when a full model including all 15 predictors was built, poor model fit indicated a separation problem within the data.

Thus, it was decided to conduct a lasso penalized logistic regression analysis, using the `glmnet` package [Friedman et al., 2010]. The model structure can be seen in Model 2 below. Just as in the full model, each of the 14 potentially significant variables were included in the model against the make tournament indicator. Wins Above Bubble was removed because its exact origin and derivation procedure could not be found. We fit the model with data from the 2015-2018 seasons, while reserving the 2019 data as the test dataset. This was so we could ultimately conduct a final unbiased evaluation of the model's predictive abilities.

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) =& \beta_0 + \beta_1 \cdot \mathrm{L12W}_i + \\ & \beta_2 \cdot \mathrm{ConfFin}_i + \\ & \beta_3 \cdot \mathrm{RPI}_i + \\ & \beta_4 \cdot \mathrm{Wins}_i + \\ & \beta_5 \cdot \mathrm{Adj.OffEff}_i + \\ & \beta_6 \cdot \mathrm{Adj.DefEff}_i + \\ & \beta_7 \cdot \mathrm{Barthag}_i + \\ & \beta_8 \cdot \mathrm{SOS}_i + \\ & \beta_9 \cdot \mathrm{NonConfSOS}_i + \\ & \beta_{10} \cdot \mathrm{ConfSOS}_i + \\ & \beta_{11} \cdot \mathrm{ConfWinPer}_i + \\ & \beta_{12} \cdot \mathrm{EffRankAvg}_i + \\ & \beta_{13} \cdot \mathrm{Power5}_i + \\ & \beta_{14} \cdot \mathrm{WinP}_i\end{aligned} \tag{2}$$

3 Results

3.1 Model Output

After fitting Model 2, cross validation was used to determine the value of λ — the parameter that determines the strength of the penalization — that would give the most appropriately simple model such that the error is within one standard error of the minimal cross-validated error [Hastie and Qian, 2014]. That value of λ was found to be 0.00307. The coefficients for each β were then calculated at that value of λ:

Variable	**Coefficient Value**
Intercept	-20.464
Last 12 Wins	0.004
Conference Finish	-0.2
RPI Rank	-0.036
Wins	0.183
Adj. Offensive Efficiency	0.154
Adj. Defensive Efficiency	-0.09
Barthag	0
SOS	13.952
Non-Conf SOS	0
Conference SOS	2.715
Conference Win %	0
Efficiency Rank Avg	0
Power 5	0
Win Percentage	0

Figure 1: LASSO regression forces the coefficients of less-contributive variables to be exactly zero, and may also only select one variable among several that are highly-correlated [Kassambara, 2018, Wang et al., 2011]. This can be seen as the coefficients for Barthag, Non-Conf SOS, Conference Win Percentage, Efficiency Rank Avg, Power 5 and Win Percentage were all brought to zero. It is intuitive that Barthag would be highly correlated with RPI, Non-Conf SOS would be somewhat correlated with SOS and Conference SOS, and that Conference Win Percentage and Win Percentage would be correlated with wins. Thus it is unsurprising that the coefficients for variables such as Barthag, Non-Conf SOS, Conference Win Percentage and Win Percentage were brought to zero.

3.2 Baseline Prediction Results & Model Adequacy

As a baseline check of the model's ability to predict the teams that made the tournament, each team's data from the 2019 season were used, in conjunction with the model coefficients to return a vector of predictions, containing each team's probability that they would make the tournament. These probabilities were used to assign predictions as to whether each team would in fact make the tournament. Specifically, if $\mathrm{p}_{\mathrm{team}} > 0.5$, that team was assigned a 1, *i.e.* making the tournament, and if $\mathrm{p}_{\mathrm{team}} < 0.5$, that team was assigned a 0, i.e not making the tournament. These assignments were then juxtaposed with the real assignments, determined by the Selection Committee, in a 2x2 table to assess baseline prediction capability.

		Actual Tournament Selection		
		Miss Tournament	Make Tournament	Total
Model Selection	Miss Tournament	282	4	286
	Make Tournament	3	32	35
	Total	285	36	321

Figure 2: The `glmnet` model accurately selected 32 teams to make the tournament in 2019, while simultaneously erroneously selecting 3 teams that were not selected by the Selection Committee. Thus, the prediction accuracy of the model with p = 0.5 as the cutoff point was 97.8%.

The confusion matrix indicated that the model was strong in its prediction capabilities. To confirm this, the model's ROC curve is given in Figure 3, and the AUC value is calculated.

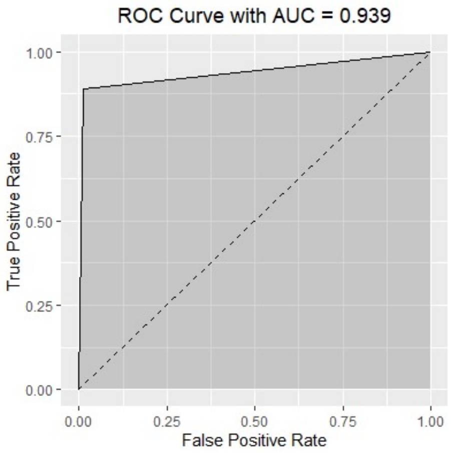


Figure 3: An ROC curve measures the tradeoff between the true positive rate (TPR), or the proportion of teams that made the tournament that were predicted to do so, with the false positive rate (FPR), or the proportion of teams that did not make the tournament that were predicted to make the tournament. We would expect a random classifier to give points along the dotted diagonal line, where TPR would equal FPR and the area under the curve (AUC) would be 0.5. Our AUC value was 0.939, which indicates our model was very efficient in classifying true tournament teams from non-tournament teams.

3.3 Intensive Prediction Results

Based on the results of the confusion matrix for the 2019 season, it was realized that there would be no way to create a probability cutoff that would guarantee the selection of exactly 36 teams for each season. The probability cutoff point would inevitably have variability year-to-year based on a number of factors such as disparity among teams, performance of individual conferences as a whole, etc. Thus, rather than maintaining an arbitrary classification cutoff of p = 0.5, the eligible teams were sorted by their fitted probabilities, and the top 36 were selected.

This new prediction scheme eliminated unnecessary error by ensuring that 36 teams would be chosen for the tournament in each given year. Each season's selections were then checked for accuracy by comparing the teams selected by this method to those selected by the Selection Committee. From 2015-2019, 130 out of a possible 144 teams were selected correctly, with no more than 4 out of 36 being chosen incorrectly for a given year. These results are summarized in the table below and graphically displayed for the 2019 season in the plots that follow. The plots for the remaining four seasons can be found in the 5.

Season	# of Correct Selections	Correct Selections Percentage
2015	34	94.4
2016	32	88.8
2017	35	97.2
2018	32	88.8
2019	33	91.7

Figure 4: For each season, the more intensive prediction scheme proved to be effective, accurately selecting at least 32 tournament teams correctly, and achieving a near-perfect 35/36 for 2017.

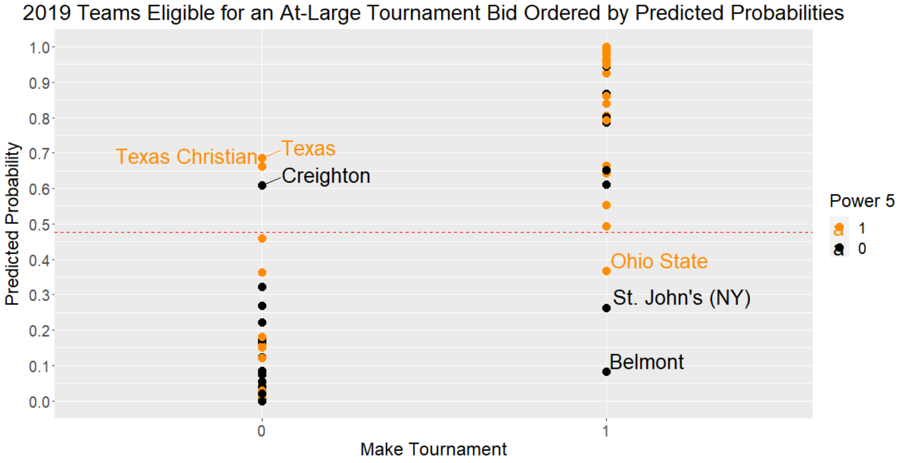


Figure 5: For the 2019 season, we have a proposed probability cutoff of approximately 0.50, represented by the red dashed line. Ohio State, St. John's and Belmont were schools selected by the Selection Committee, but not by the model, while Texas, Texas Christian and Creighton were schools selected by the model, but not by the Selection Committee.

3.4 Variable Trends Over Time

Given the variability in correct predictions year-to-year, the same lasso penalized logistic regression model was run on each individual season's data to check for any trends in model coefficients. It was hypothesized that any changes in a variable's coefficient from year to year represented the Selection Committee putting more or less emphasis on that factor in that season.

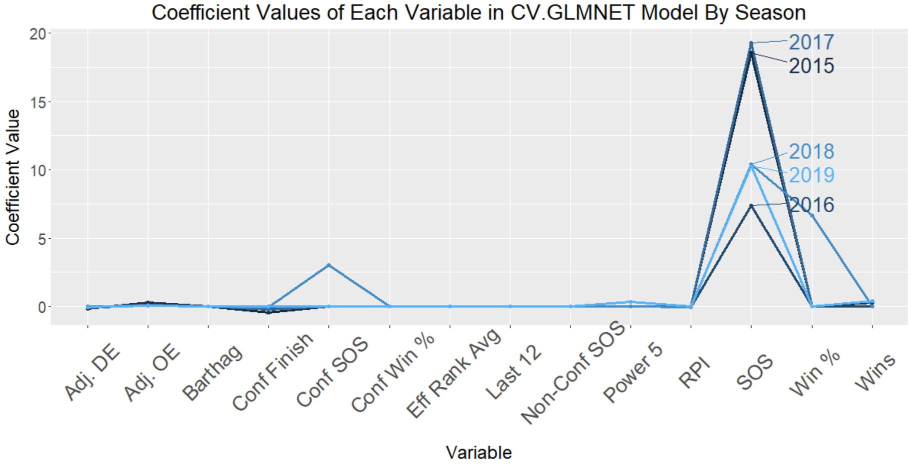


Figure 6: The coefficient values for most variables remain consistent from year to year, with the exception of Strength of Schedule (SOS), whose 2015 and 2017 values appear much larger than those of the other three years.

It is of interest that the two seasons which the model predicted most accurately, 2017 and 2015, are the seasons in which Strength of Schedule (SOS) had the two highest coefficients. Given that the original model returned a coefficient for SOS that was weighted more towards the values from 2017 and 2015 (13.94), the Selection Committee's decision to put less emphasis on SOS in 2016, 2018 and 2019 may help explain why the model did not predict the tournament teams in these seasons quite as well. There is no obvious explanation for the drop in SOS importance in 2016, but the lower values in 2018 and 2019 may be explained by the Selection Committee's inclusion of BPI, Pomeroy and Sagarin ratings in their process prior to the 2018 season [Muma, 2017]. Each of the three ratings systems are known to include a measure of strength of schedule. The introduction of new factors in 2018 may also explain why the Selection Committee appeared to put additional emphasis on Conference SOS and Win Percentage in that season as they attempted to include other metrics seamlessly into the selection process.

3.5 Seeding Results

Part two of the goal in replicating the process of the Selection Committee was testing the model's ability to seed the teams from 1-16. Upon choosing the 68 teams that will earn bids to the tournament, the Selection Committee ranks these teams from 1-68. For the first 10 seeds, the teams ranked in positions 1-4 each earn a 1 seed, the teams ranked in positions 5-8 each earn a 2 seed, and so on. The committee then gives the next six teams, those ranked from 41-46, #11 seeds. Seeds 12-15 are assigned in the same way as the first 10 seeds, and the final six teams, ranked in spots 63-68, each earn # 16 seeds. The last four teams that earn #11 seeds (ranks 43-46) and the last four teams that earn #16 seeds (ranks 65-68) are the eight teams that compete in the four 'First Four' games. To simulate this process, the 36 teams chosen by the model each year were combined with the 32 conference champions from that year and placed in their own data set. To seed these 68 teams, the data were used in combination with the coefficients from the original lasso model to find new predicted probabilities of each team making the tournament. The teams were ordered from 1-68 based on their predicted probabilities, and then assigned a seed based on the process described above. These seed assignments were then compared to those of the Selection Committee for accuracy.

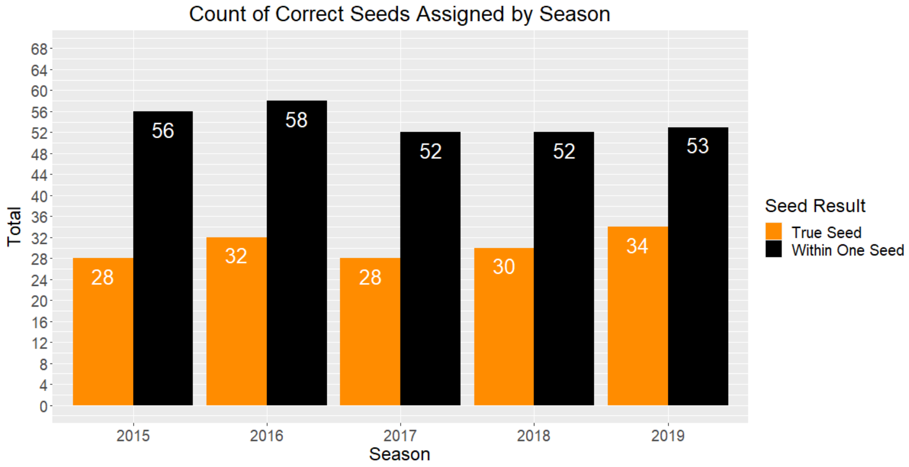
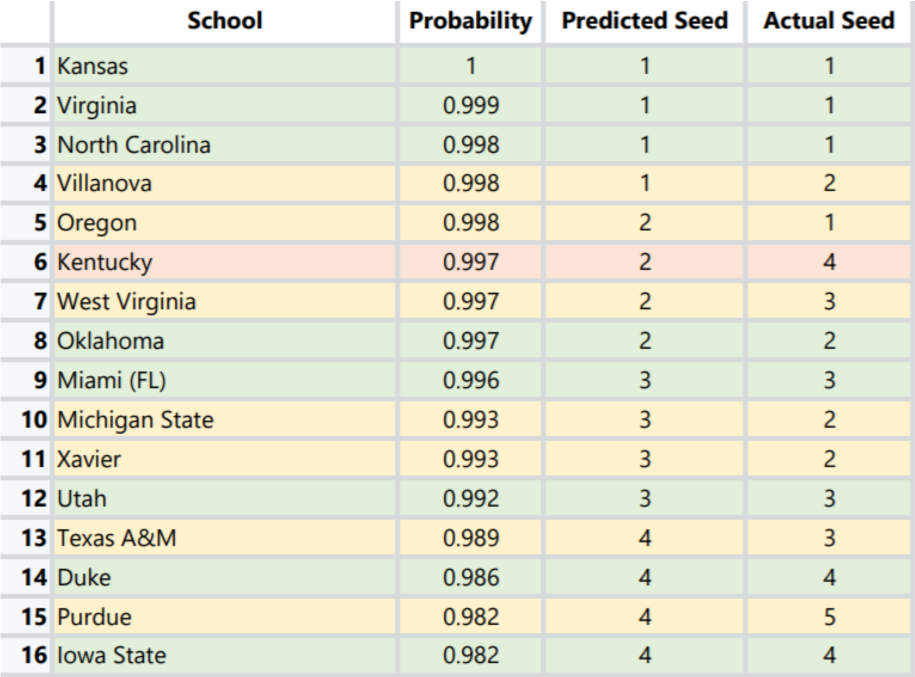


Figure 7: For each season, the model correctly assigned true seeds to between 28 and 34 teams, while correctly assigning seeds within one of the true seed for between 52 and 58 teams.

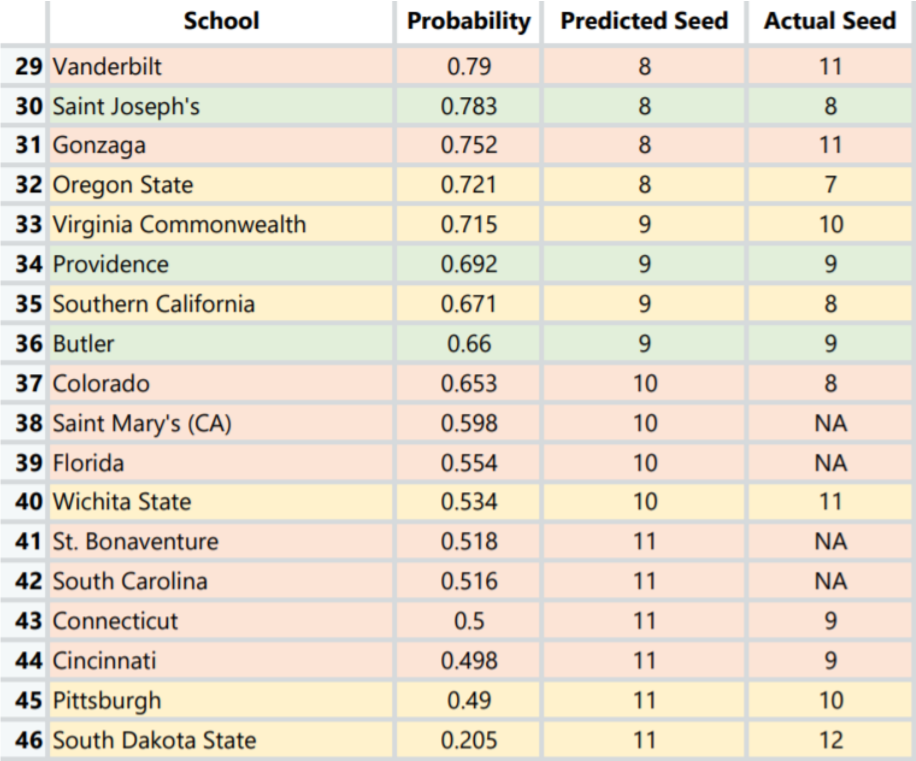
To better demonstrate the comparison between assigned seeds by the model and assigned seeds by the Selection Committee, two charts from the 2016 season are found below.



	School	Probability	Predicted Seed	Actual Seed
1	Kansas	1	1	1
2	Virginia	0.999	1	1
3	North Carolina	0.998	1	1
4	Villanova	0.998	1	2
5	Oregon	0.998	2	1
6	Kentucky	0.997	2	4
7	West Virginia	0.997	2	3
8	Oklahoma	0.997	2	2
9	Miami (FL)	0.996	3	3
10	Michigan State	0.993	3	2
11	Xavier	0.993	3	2
12	Utah	0.992	3	3
13	Texas A&M	0.989	4	3
14	Duke	0.986	4	4
15	Purdue	0.982	4	5
16	Iowa State	0.982	4	4

Figure 8: This chart includes the 16 teams assigned seeds 1-4 by the lasso regression model. Rows outlined in green illustrate that the model assigned the same seed as the Selection Committee. Rows outlined in yellow illustrate that the model assigned a seed within one of the seed assigned by the Selection Committee. Finally, the row outlined in red illustrates that the model assigned a seed that was different than the seed assigned by the Selection Committee by more than one. The probability column is included to show how teams were ordered by descending probability of making the tournament.

The high accuracy of correct seeding across the first 1-7 seeds, or the teams ranked 1-28, as well as the 12-16 seeds, or the teams ranked 47-68, was consistent throughout all five seasons; there were few instances of the model missing a seed by more than one.



	School	Probability	Predicted Seed	Actual Seed
29	Vanderbilt	0.79	8	11
30	Saint Joseph's	0.783	8	8
31	Gonzaga	0.752	8	11
32	Oregon State	0.721	8	7
33	Virginia Commonwealth	0.715	9	10
34	Providence	0.692	9	9
35	Southern California	0.671	9	8
36	Butler	0.66	9	9
37	Colorado	0.653	10	8
38	Saint Mary's (CA)	0.598	10	NA
39	Florida	0.554	10	NA
40	Wichita State	0.534	10	11
41	St. Bonaventure	0.518	11	NA
42	South Carolina	0.516	11	NA
43	Connecticut	0.5	11	9
44	Cincinnati	0.498	11	9
45	Pittsburgh	0.49	11	10
46	South Dakota State	0.205	11	12

Figure 9: This chart includes the 18 teams assigned seeds 8-11 by the lasso regression model. The colored rows have the same significance as in the table above. The values with NA in the ”Actual Seed” column represent schools that were selected for the tournament by the model, but not by the Selection Committee, hence they were not assigned a seed by the committee

Similarly, the low seeding accuracy was also consistent throughout the five seasons. In each season, the model had the most trouble accurately ranking teams in the ’bubble,’ that is, teams outside of the top contenders whose tournament fate was less certain. The frequency of red rows also can be explained by the teams that were incorrectly selected to make the tournament. It was expected that we would see lower seeding accuracy among the teams seeded in the higher positions. Specifically, the seeds 8-11 hold the most uncertainty for the selection committee because the committee must select a limited number of teams from a large pool that could all make similarly legitimate, but not obvious, cases for qualification in the tournament.

4 Discussion

The lasso penalized regression model used in this project was largely successful in correctly selecting the tournament field, achieving a 90.3% overall accurate selection rate, and though it only seeded 44.7% of teams at their correct seed, 79.7% were seeded within one seed of their true seed across the five seasons. This result is fascinating because it seems to align with selection committee rules. Specifically, the NCAA website states that teams may be moved up or down one seed if necessary to meet various seeding principles. Specific examples of these seeding principles are explained further below. NCAA.com [2019].

This project was still limited in both its selection and its seeding procedures. From a selection standpoint, there are other variables that the Selection Committee is known to use to make their selections, such as head-to-head results, record against common opponents, and the number of quality wins/bad losses. Data on each of these variables would have been possible to obtain, but the time constraints of a summer project hindered the ability to access these data. It would be of interest to see if adding these variables to the current penalized regression model would improve prediction accuracy.

However, it should be noted that the selection committee follows an intensive selection process that involves creating small groups of very similar quality, and selecting individual teams from those groups through the use of a ballot. Specifically, the committee creates a list of all schools 'under consideration' for an at-large bid, then votes to select the best eight of those schools. The eight with the most votes are put onto a ballot where the top four are selected by another vote. These four teams earn an at-large bid. From that point, four more teams are added to the ballot, and another vote takes place to select another four teams. This process continues until all 36 bids are awarded [NCAA.com, 2019].

This is a contrast to the model in this project because teams were compared to all other teams on a larger scale, rather than just those in their group of eight, which means the model was hindered in its ability to detect smaller differences between similar teams. Thus, if the project was to be expanded, the primary goal would be to create a method that more closely replicates the vote-select-vote system that the committee uses. The model was also limited in its ability to assign seeds. Besides the fact that the Selection Committee seeds the teams in the same way as they select the at-large teams, the committee also has a number of rules and restrictions on the way that teams can be seeded, as mentioned above. For example, teams from the same conference may not face each other prior to the conference finals if they have played three or more times prior to the tournament NCAA.com [2019]. Furthermore, teams will remain in or as close to their campus as possible. It is possible that a model could account for these restrictions, if data were to be gathered on the schedules and locations of each team and the sites of each first round regional game. Although it would require a large amount of cross-referencing and additional analysis, it would be intriguing to see how much the inclusion of these factors would improve a model's ability to seed teams.

This project was an attempt to use a computer to replicate a human process. The lasso penalized logistic regression model generated plausible results for the selection and seeding of tournament teams, though it did not achieve 100% accuracy due to the exclusion of several important factors and restrictions used by the committee because of time constraints. This model was a successful first attempt to reproduce the selection process of the NCAA Selection Committee and can certainly be modified and improved upon in the future.

5 Appendix

The following four figures depict the lasso penalized regression model's performance in accurately selecting tournament teams in the seasons 2015-2018.

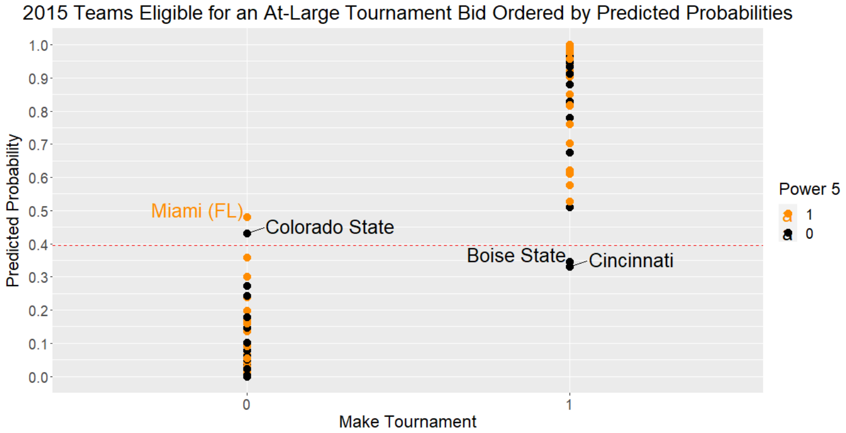


Figure 10: For the 2015 season, we have a proposed probability cutoff of approximately 0.40, represented by the red dashed line. Boise State and Cincinnati were selected by the Selection Committee, but not by the model, while Miami (FL) and Colorado State were selected by the model, but not by the Selection Committee.

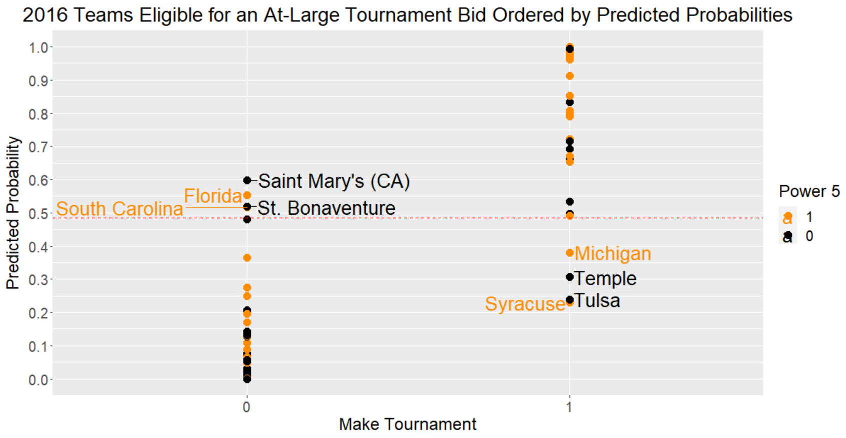


Figure 11: For the 2016 season, we have a proposed probability cutoff of approximately 0.49, represented by the red dashed line. Michigan, Temple, Tulsa and Syracuse were selected by the Selection Committee, but not by the model, while Saint Mary's, Florida, South Carolina and St. Bonaventure were selected by the model, but not by the Selection Committee.

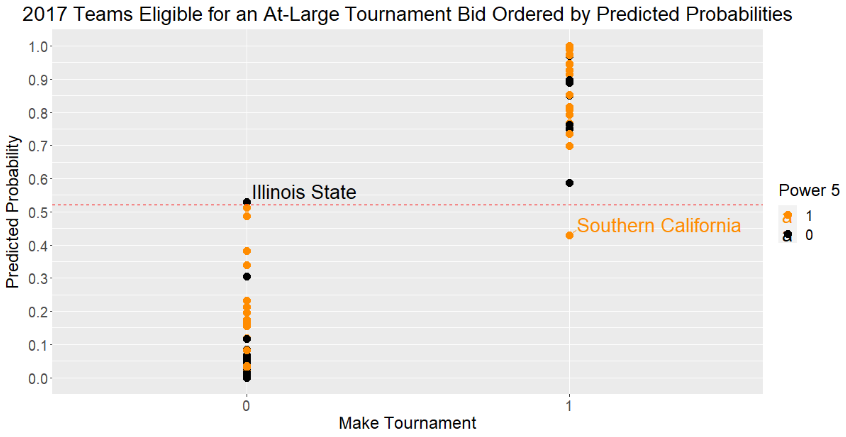


Figure 12: For the 2017 season, we have a proposed probability cutoff of approximately 0.52, represented by the red dashed line. Southern California was selected by the Selection Committee, but not by the model, while Illinois State was selected by the model, but not by the Selection Committee.

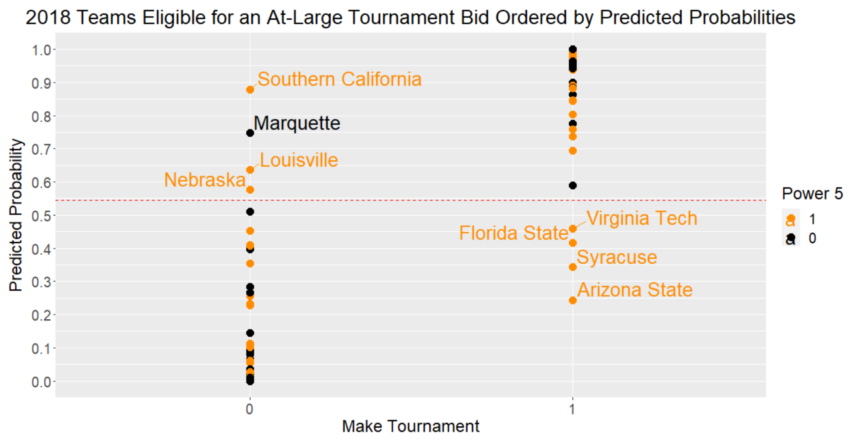


Figure 13: For the 2018 season, we have a proposed probability cutoff of approximately 0.55, represented by the red dashed line. Virginia Tech, Florida State, Syracuse and Arizona State were schools selected by the Selection Committee, but not by the model, while Southern California, Marquette, Louisville and Nebraska were schools selected by the model, but not by the Selection Committee.

References

List of nit champions, by year, May 2020. URL https://www.coachesdatabase.com/list-nit-champions/.

Men's basketball selections 101 - selections, Mar 2020. URL http://www.ncaa.org/about/resources/media-center/mens-basketball-selections-101-selections.

Brad Adgate. 50-plus fun facts about march madness, Mar 2019. URL https://www.forbes.com/sites/bradadgate/2019/03/18/50-fun-facts-about-march-madness/.

Shouvik Dutta and Sheldon Jacobson. Modeling the ncaa basketball tournament selection process using a decision tree. *Journal of Sports Analytics*, 4:65–71, 2018.

Kevin Flaherty. Statistics website kenpom releases 2020-21 rankings, Nov 2020. URL https://247sports.com/LongFormArticle/KenPom-college-basketball-rankings-2021-Baylor-Bears-Gonzaga-Bulldogs-Duke-Blue-Devils-154441771.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

Trevor Hastie and Junyang Qian. Glmnet vignette, Jun 2014. URL https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html.

Hao Ji, Erich O'Saben, Adam Boudion, and Yaohang Li. March madness prediction: A matrix completion approach, 2015. URL https://pdfs.semanticscholar.org/28d0/c0074ee6c7c56740f130402c13b6d0cfa0e4.pdf.

Kassambara. Penalized logistic regression essentials in r: Ridge, lasso and elastic net, Mar 2018. URL http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net.

John S Kiernan. 2020 march madness stats amp; facts, Mar 2020. URL https://wallethub.com/blog/march-madness-statistics/11016/.

Steven Muma. Ncaa men's basketball tournament selection committee finally moving toward inclusion of advancednbsp;metrics, Jan 2017. URL https://www.backingthepack.com/2017/1/13/14269514/ncaa-tournament-pomeroy-ratings-analytics-selection-committee.

NCAA.com. How the field of 68 teams is picked for march madness, Oct 2019. URL https://www.ncaa.com/news/basketball-men/article/2018-10-19/how-field-68-teams-picked-march-madness.

Anna Negron. Espn 'bracketologist' joe lunardi inks new multi-year deal, Feb 2021. URL https://espnpressroom.com/us/press-releases/2021/02/espn-bracketologist-joe-lunardi-inks-new-multi-year-deal/.

Tim Parker. How much does the ncaa make off march madness?, Oct 2019. URL https://www.investopedia.com/articles/investing/031516/how-much-does-ncaa-make-march-madness.asp.

Ken Pomeroy. Ratings glossary: The kenpom.com blog, 2002. URL https://kenpom.com/blog/ratings-glossary/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

SR 1. 2018-19 school stats: College basketball at sports, 2020. URL https://www.sports-reference.com/cbb/seasons/2019-school-stats.html.

SR 2. 2018-19 season summary: College basketball at sports, 2020. URL https://www.sports-reference.com/cbb/seasons/2019.html.

SR 3. 2018-19 abilene christian wildcats schedule and results: College basketball at sports, 2020. URL https://www.sports-reference.com/cbb/schools/abilene-christian/2019-schedule.html.

SR 4. 2018-19 american athletic conference season summary: College basketball at sports, 2020. URL https://www.sports-reference.com/cbb/conferences/aac/2019.html.

Jim Sukup. URL http://rpiratings.com/WhatisRPI.php.

Team Rankings. Ncaa college basketball rpi rankings amp; ratings 2020, 2020. URL https://www.teamrankings.com/ncaa-basketball/rpi-ranking/rpi-rating-by-team.

TheSabre.com. Another stat to look at "wins above bubble" (bart torvik), 2020. URL https://virginia.sportswar.com/mid/13282021/board/basketball/.

Bart Torvik. Rank faq. URL http://adamcwisports.blogspot.com/p/every-possession-counts.html.

Bart Torvik. 2020-21 projections - - customizable college basketball tempo free stats - t-rank, 2020. URL http://www.barttorvik.com/.

Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *The Annals of Applied Statistics*, 5(1):468–485, 2011. doi: 10.1214/10-aoas377.

Warren Nolan. Net rankings, 2020. URL http://warrennolan.com/basketball/2020/net.

Chris Wright. Statistical Predictors of March Madness: An Examination of the NCAA Men's' Basketball Championship. 2012.