Dan Trunzo and Libby Scholz
MCS 100
June 4, 2016

Our Shining Moment: Hierarchical Clustering to Determine NCAA Tournament Seeding

This project tries to correctly predict the NCAA Tournament Selection Committee's bracket arrangement. We look both to correctly predict the 36 at large teams included in the tournament, and to seed teams in a similar way to the selection committee. The NCAA Tournament Selection Committee currently operates in secrecy, and the selection process is not transparent, either for teams or fans. If we successfully replicate the committee's selections, our variable choices shed light on the "black box" of selection. As basketball fans ourselves, we hope to better understand which teams are in, which are out, and why

Existing Selection Process

32 teams automatically qualify for the March NCAA tournament by winning their conference championship tournament. The other 36 "at large" teams are selected by the ten person committee through multiple voting rounds, where a team needs progressively fewer votes from the committee to make the tournament. Once the teams are selected, a similar process follows for seeding, where the selection committee ranks the best eight teams remaining, then the best six, and then the best four as more teams are placed in the bracket. Once all 68 teams are ranked, they are then placed in a bracket. At this point, the actual rank order may be violated to avoid "teams from the same conference… meet[ing] prior to the regional final if they played each other three or more times during the regular season and conference tournament," or prior to the semifinals if they played at least twice. Teams in the first four seeds may not be placed at a "home court disadvantage," and all teams are placed as close to their "natural

region” as possible. In all, a team may move as far as two seeds above or below its true position to accommodate these rules.

Without incorporating the number of meetings between two teams or the distance between possible tournament sites into our model, we cannot accurately seed each of the four regions. However, we are able to predict the true seeds of the teams in a ranked list from first to sixty-eighth.

Data

We sourced our data from College Basketball Reference. Since we wanted to simulate an rating for teams that would be similar to the Ratings Percentage Index (RPI) with more components, we focused on win-loss ratio, deviation from the mean point differential, strength of schedule, offensive rating, defensive rating, and net rating. The win-loss ratio is one of the most important components of the RPI, which is as follows:

$$\textbf{RPI} = (\textbf{WP} * 0.25) + (\textbf{OWP} * 0.50) + (\textbf{OOWP} * 0.25)$$

where WP is Winning Percentage, OWP is Opponents' Winning Percentage and OOWP is Opponents' Opponents' Winning Percentage. Thus, the RPI often favors teams with high win-loss ratios in general. However, RPI rating alone does not translate directly to tournament bids.

We included the deviation from the mean point differential because it gives critical insight to the success of a team. Teams with a positive deviation from the mean point differential suggests that they are skilled teams that are able to win by a greater margin against worse opponents. This is critical when it comes to tournament bids and AP ranking. For example, if Villanova went to two overtimes with St. John’s and only won by 1, they would likely drop in the rankings of the eyes of the selection committee due to the bad win.

The strength of schedule also plays a vital role in tournament seeding. Many were outraged when the 2006-2007 Boise State football team made it to a BCS bowl after going undefeated. Pundits claimed that Boise State had no real competition in the Western Athletic Conference. They ended up grinding out a win in overtime against Oklahoma. A similar example in college basketball is the 2012-2013 Wichita State team. They were 34-0 going into the NCAA Tournament and were given the 1 seed in the Midwest regional. Wichita State played no real contenders, and this is reflected by the fact they were ranked second in the end of season AP poll, but seventh by the coaches poll. Whereas the AP poll is formulaic, the coaches are able to understand that Wichita State played no real opponents. We assume that the selection committee was obligated to give them the 1 seed because they were undefeated, although not 1 seed material. This is reflected in the fact the seven-seeded University of Kentucky knocked them out in the second round.

The offensive, defensive, and net ratings used in our calculations and analysis were provided by College Basketball Reference. Their offensive rating provides the estimated points produced per 100 possessions of a team. Similarly, the defensive rating is the estimated points allowed for every 100 possessions by an opponent. Finally, the net rating is the point spread between these two values. Although the spread might indicate which teams are especially talented, we wanted to include both offense and defensive ratings since teams with exceptionally high offensive ratings and low defensive ratings are typically some of the best teams in the country known for their offense or defense.

Methodology and Results

The first step in answering our research question is determining the appropriate statistical method. We initially built a Bradley-Terry model. However, because of the secrecy of the selection committee, the actual breakdown of weight put on RPI and other factors is unknown. Thus, any model we build would be highly subjective with the possibility of overfitting on our training data. We did not feel comfortable using the Bradley-Terry model as an indicator of the teams that would be in the tournament.

We next tried a logistic regression. Since the RPI is on a scale of 0 to 1, we thought that a binomial logistic regression across all our data would provide accurate results. Although we have all our data, we did not have a value to regress. When we used each team's end of season rank given by College Basketball Reference as the object we would regress upon, the regression simply gave us back the original ranking. We then tried the logistic regression against win-loss ratio, which is the best indicator of RPI in our data. However, after the top twenty teams, the regression had difficulty picking up minute differences in the data. Whereas the RPI can differ between teams by 0.0002, our regression was not sensitive enough.

Finally, we decided on k-means clustering. Not only were we looking for the teams to receive a bid, but we also wanted to predict seeding as well. K-means, to an extent, could provide a clustering of teams by their qualities. We built a data frame to store all of our data, and we only selected the top 200 teams. We decided that any teams in roughly the bottom forty percent have virtually no chance of making the tournament. Then, we ran k-means with 50 clusters to try and break up the teams into sizeable chunks. Unfortunately, it is extremely difficult to add a parameter to restrict clusters such that they all contain the same number of teams.

Once we executed k-means, we wanted a more visual representation of the data, and the opportunity to rank each cluster by how good the teams within it are. In order to

accomplish this task, we used divisive hierarchical clustering. Divisive means that all teams begin in one cluster and are continually broken down. Our matrix of data not only included all our original data, but also the k-means clustering fits. From here, the hclust function in R calculated the Euclidean distance of our matrix. We were able to graph the hierarchical clustering in a dendrogram for the 2013 NCAA tournament [see appendix]. The most important component of the hierarchical clustering function in R is order object it produces. From the order object, hclust ranks every team from one to 200. Now, we are able to select the 32 conference winners and selected the next top 36 teams that would potentially receive at-large bids according to our calculations.

2013 Projected Top Seeds	**2013 Actual Seeding**
Indiana	1
Syracuse	4
The Ohio State University	2
Florida	3
Duke	2
Michigan	4
Louisville	1
Kansas	1

2014 Projected Top Seeds	**2014 Actual Seeding**
Wisconsin	2
Kansas	1
Florida	1
Kentucky	7
Michigan	2
Virginia	1
Michigan State	4
Arizona	1

2015 Projected Top Seeds	2015 Actual Seeds
Virginia	2
Arizona	2
Villanova	1
Duke	1
Wisconsin	1
Gonzaga	2
Kentucky	1
Xavier	6

Analysis

Our model was able to predict on average 85% of the teams that would be in the NCAA tournament. Our most successful year was the one which we trained on, the 2013 NCAA tournament, for which we accurately predicted 62 of the 68 teams for an accuracy of 91.2%. We were unable to predict every team that would make it into the tournament. Furthermore, our clustering was unable to provide accurate seeding for the tournament.

The method put the most weight on strength of schedule and end of season rank. The emphasis on rank might be consistent with a behavioral bias by the selection committee, that they care more about finishing strong than performance throughout the season. Our strength of schedule variable mimics the RPI ratings, which we know the selection committee takes seriously in their deliberations.

However, our model was extremely accurate at tournament success. In the 2013, 2014, and 2015 NCAA tournaments, we accurately predict 5 or more teams in our top 8 results to make it to the Elite Eight. This leads us to believe that our model was better at

predicting tournament viability than tournament entrance. We expect this is the case for several reasons: our model is built only on in-season performance, whereas a selection committee's judgement may be biased towards selecting teams with strong historical records. Our model does its own way of valuing teams, with hereas the selection committee relies heavily on RPI. Related to RPI, our model excludes opponents' opponents' strength of schedule, which RPI considers. Additionally, our model does not differentiate "marquee wins," another factor which likely biases the selection committee.

Areas for Future Research

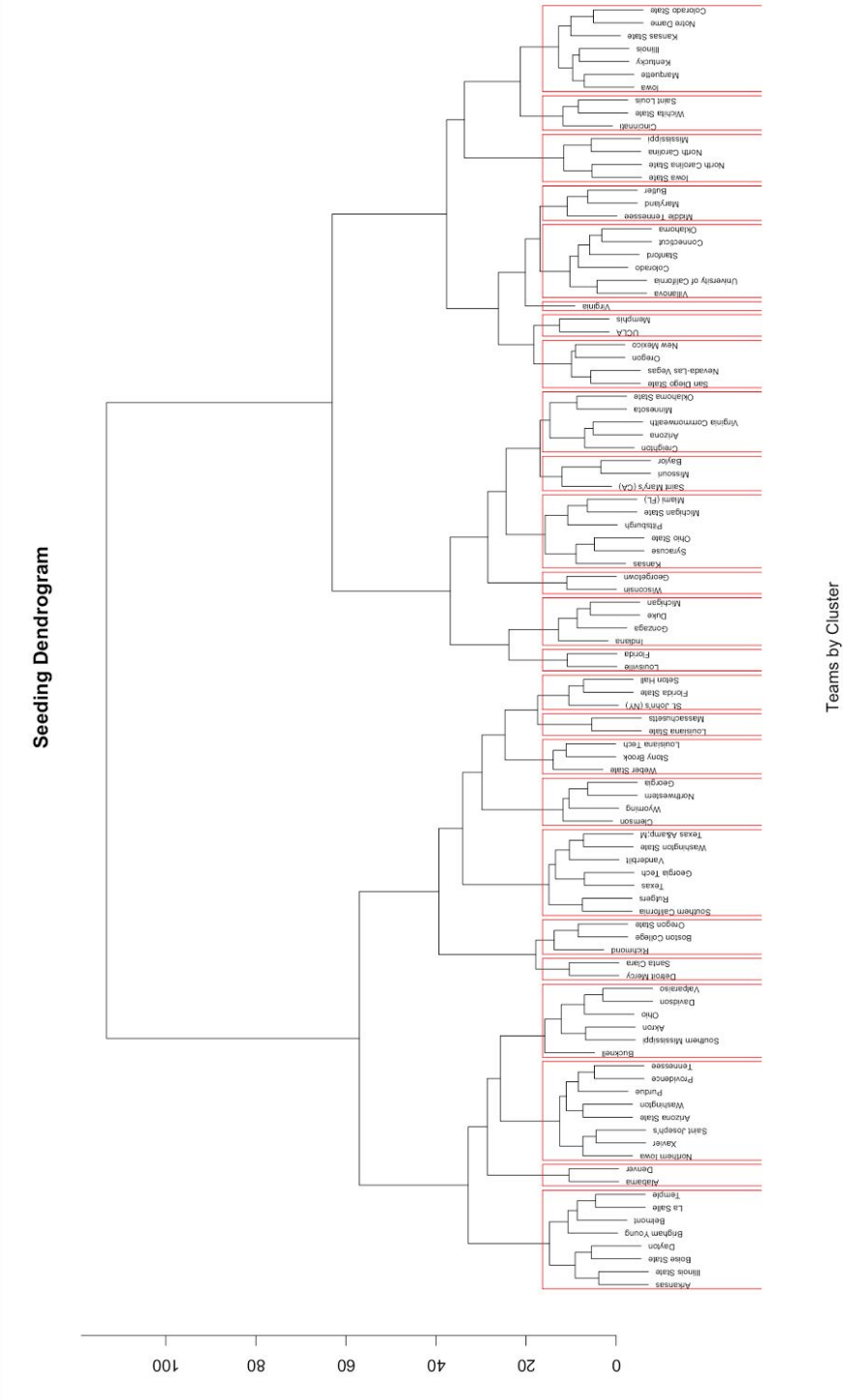
Improving the accuracy of our model will require including more variables that mimic the iterative seeding process. A function that chooses a region for a team based on minimum distance to the tournament site would improve regional placement. A model that can accommodate the small differences in RPI would improve on our existing results because it would include one of the known variables the selection committee considers. We could also introduce a variable for signature wins that would be the rank of the best team beat during a season, in addition to a variable for overall strength of schedule.

Conclusion

Ultimately, our challenge is to automate an entirely manual, subjective process. In some sense, our results shed insight onto the predictability of the selection committee: basing our model entirely on season performance predicted with as much as 85% accuracy what teams made the field. This implies the committee does not consider many additional holistic factors; adding historical program strength or conference revenue might not improve the model significantly. But the committee changes membership and does not discuss their selections year to year; improving our model

beyond the recommendations we provide above might require insight into individual decision making, not more data.

Appendix


Seeding Dendrogram
Teams by Cluster
0
20
40
60
80
100
Arkansas
Illinois State
Boise State
Dayton
Brigham Young
Belmont
La Salle
Temple
Alabama
Denver
Northern Iowa
Xavier
Saint Joseph's
Arizona State
Washington
Purdue
Providence
Tennessee
Bucknell
Southern Mississippi
Akron
Ohio
Davidson
Valparaiso
Detroit Mercy
Santa Clara
Richmond
Boston College
Oregon State
Southern California
Rutgers
Texas
Georgia Tech
Vanderbilt
Wasington State
Texas A&M
Clemson
Wyoming
Northwestern
Georgia
Weber State
Stony Brook
Louisiana Tech
Louisiana State
Massachusetts
St. John's (NY)
Florida State
Seton Hall
Louisville
Florida
Indiana
Gonzaga
Duke
Michigan
Wisconsin
Georgetown
Kansas
Syracuse
Ohio State
Pittsburgh
Michigan State
Miami (FL)
Saint Mary's (CA)
Missouri
Baylor
Creighton
Arizona
Virginia Commonwealth
Minnesota
Oklahoma State
San Diego State
Nevada-Las Vegas
Oregon
New Mexico
UCLA
Memphis
Virginia
Villanova
University of California
Colorado
Stanford
Connecticut
Oklahoma
Middle Tennessee
Maryland
Butler
Iowa State
North Carolina State
North Carolina
Mississippi
Cincinnati
Wichita State
Saint Louis
Iowa
Marquette
Kentucky
Illinois
Kansas State
Notre Dame
Colorado State