Bringing Sanity to March Madness

Alex Marti, Austin Tice, Max Layer, Kendall Land

Background

Every year come mid-March, the NCAA College Basketball tournaments start, and mania promptly ensues. Games that should be blowouts become nail-biters, upsets happen, and a few underdog teams become what are known as “cinderellas”. All of which shows how it has earned the name March Madness®. Our team plans to address the problem of being able to predict the outcome of a game in the tournament.

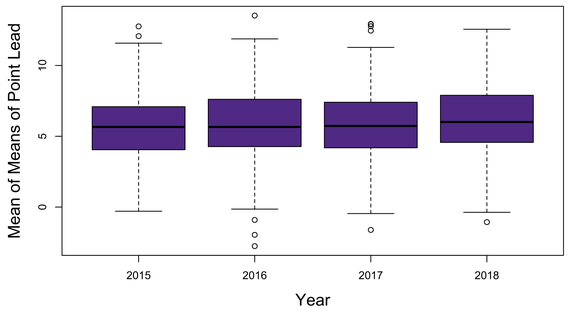
Our goal is to be able to quantify and/or explain a team’s ability to “stay in a game”, their competitiveness, and their “cinderella-ness” in the NCAA College Basketball Tournament, based on how they performed in the regular season. Then, using these attributes we determine for each team, can we predict who will win each game in the tournament.

Stay-In-The-Game

An interesting aspect of a team we wanted to quantify was how a team performed when there was a large point discrepancy. In essence, how could we quantify if particular teams perform better when under pressure or worse when they had a sizable lead on their opponents, and how would such a statistic compare to other measurements of the teams.

To calculate this, we grouped the play by play data based on each game, then took the mean of the difference between the scores of the teams over the course of the game. By grouping that dataset by the winning teams, we could then take the mean of the means of their leads in each winning game.

Plotting this reveals that the average lead of a winning team is consistent, with some exceptional teams either having a consistently large lead over their opponent, or a low and even negative average leads over their opponents, indicating teams that consistently made remarkable comebacks from a losing position.



Competitiveness

Approach

First, we must define what competitiveness is to us, because there is no standard definition of it. To us, competitiveness is the ability for a team to win games and play with the best of the best.

We want to be able to quantify this value for each team so we can compare them against other teams in the tournament to make predictions. We will give them their values based upon the numbers they put up in the regular season. We will accomplish this by training a neural network on previous seasons to create weights that determine the importance of each statistic towards a win.

However, there are few things we need to verify first to see if this method has any chance of holding water, or if we will need to make any adjustments.

1. Do stats for a win in the tournament match up with stats for a win in the regular season.
2. Do teams put up similar stats in the tournament to what they put up in the regular season.

Checking Win Statistics for the Regular Season vs. the Tournament

We used K Nearest Neighbors (k-NN) to see if the stats for wins and losses in the regular season fall into similar groupings with stats for wins and losses in the tournament. We chose k-NN, because we can train a k-NN model on the stats of each team in all the regular season games, and then give it the stats of each team in the tournament, and it will try and give it a label of won or lost based on statistics from other games that have stats close to it.

After trying many different combinations of variables, our misclassification rate was about 22% for most of them. While this number isn’t as low as we were hoping for, it still shows over 75% of the stats for wins and losses were similar enough to correctly predict their labels. But we should still verify that this doesn’t happen because the tournaments are that much different than regular season games. So we did the same thing to test that, except this time we randomly split the regular season games into our training and testing data. And as a result, we got a 22% misclassification rate (exactly the same). So, based on everything we gathered, we decided that the best combination to use for our feature set was:

- Score
- Assists (Ast)
- Steals (Stl)
- Blocks (Blk)
- Turnovers (TO)
- Total Rebounds (TR)
- Field Goal Percentage (FGP)
- Free Throw Percentage (FTP)
- 3-PT Percentage (FGP3)

Checking How Teams Perform in the Regular Season vs. the Tournament

To verify this, we simply took each team's average stats from the regular season and subtracted that value from their stats in each game of the tournament they played in that respective season. This gives how much better or worse each team does in each game. We then took the average of each column so we could have the average difference of every team's performance in every tournament game (see table).

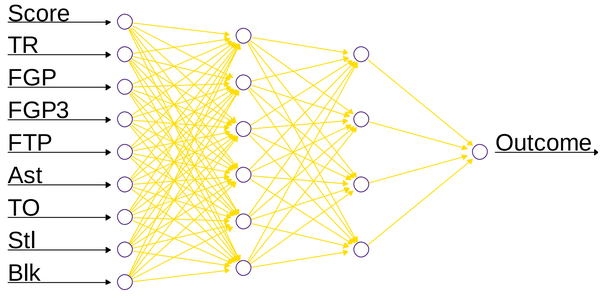
Score	TR	FGP	FGP3
-4.9960	-1.8899	-0.0236	-0.0189

FTP	Ast	TO	Stl	Blk
0.0033	-1.8386	-0.7150	-0.8896	-0.6465

From this, we can see that teams don’t really perform that much different in the tournament than they do in the regular season. So there is no need to adjust our competitiveness scores we get from the regular season.

Creating Competitiveness Scores with a Neural Network

The reason we chose a neural network to determine our scores is because it can determine the weights that correspond best to predicting the likelihood a team will win a game with the given features. So all we have to do is give it the training data and the features we want it to train on, and it will compute the rest. A layout of our model (minus the biases) can be seen in the image below. (Outcome is our competitiveness score)



We trained our model on all seasons before 2019 we had data for and calculated each team’s competitiveness score in 2019 based on their average stats that year. We then compared the scores of the two teams in each game of the tournament and predicted the higher of the two to win.

Given all the matchups, it was able to get 48 out of 67 games (~72%) right. But given only the first round, it was only able to get 27 out of 67 (~40%) right. we also made one that took the opponent’s competitiveness into account, but it did worse, only getting 42 right.

Cinderella-ness

Cinderella, "Cinderella story", and Cinderella-ness are terms used to refer to situations in which competitors achieve far greater success than would reasonably have been expected. In order to properly assess the cinderella-ness of a team in March Madness, we had to define the criteria teams have to meet to become a cinderella upon further success. We decided they need to meet at least one of the following:

1) The team is a 14, 15, or 16 seed.
   - This conclusion was brought about by analyzing the years 2015-2019 and noting that there was only one team from each of these seedings that won.
2) It is the team's first time participating in March Madness, and they are given an unfavorable chance of winning (i.e. a seed of 9 or higher).
   - We want to consider factors such as nervousness and unpreparedness. These are common things that could occur for a team who are having their first appearance in the tournament

Every year there are only a handful of teams that become cinderellas; however, it's these same teams that cause the most complications when it comes to achieving the perfect bracket.

Every year, an average of two new teams enter the tournament meeting the criteria to be a cinderella. This means that, including the 14, 15, and 16 seeds uniquely, approximately 21.88% of teams would fall under the "cinderella" label.

Lessons Learned

We learned how hard it is to implement everything together, as well as just how many factors go into trying to quantify and/or explain each attribute. Our takeaway from this would be to focus more on combining all the attributes into a single model.

Future Work

- Look at what year in college each team’s player are in (Freshman, Sophomore, etc).
- Look at a team’s momentum coming into a game (e.g. are they on a big win streak, did they just beat a really good team, etc).
- Look at the injured players for each team in every game.
- Look more into how the skill of each opponent a team plays in the regular season should factor into their competitiveness score.