Sophie Moore

Professor Plantinga

STAT 201

May 9, 2019

Predicting March Madness: How to Hack Your Bracket

We all know how March Madness works- NCAA division I basketball teams face off tournament style in single elimination rounds until the winner is crowned. People everywhere make brackets predicting who they think will go the distance. Lots of competitions have BIG prize money for winning your group, so the motivation is there! But you have a 1 in 9.2 quintillion chance on getting your bracket right off sheer probability. So if we can find a way to increase our chances of guessing the winners correctly, we want to!

I found a data set on all the NCAA teams and March Madness results from the past 60 years that another user web scraped and uploaded in 2016. His data set pulls data from the AP poll, which has top sports broadcasters and writers rank their top 25 teams and compile the results. It also pulls identifying data about teams and their records going into the tournament each year. Lastly, it compiles teams' final rank in the tournament itself. As is the nature of a data set this huge and web scraped, there are some missing values and discrepancies which I attempted to correct for when possible. The most major thing was that a lot of teams were included in years they had no chance of making the tournament at all, skewing the results, so I removed the teams/years that either did not make the tournament or had an AP Score of 30, indicating they were not expected to make the tournament. The conferences were also often too specific to be meaningful, so I regrouped them into larger regions. It's also important to note that this data isn't independent- since one team winning means another team losing, it has "built in dependence". This is a limitation of the data set, but it is nearly impossible to fully correct for and we can still obtain interesting results from it, since there are so many events.

Because my goal was to try to make the best March Madness bracket I could, I wanted to see if any of the data I had could help me predict how far teams would make it in the tournament. Because my outcome was quantitative (number of rounds), I performed a multiple regression analysis on my data, using NCAA result as my response variable and WL, AP Score, and region as my predictor variables. Although we have already discussed that the independence condition for regression was not quite met, I made a plot of residuals and a QQ-Plot to check the other conditions and found that my data had about equal variance (although it looks funny because one of the variables is discrete), and was nearly normal.

After I had checked my conditions, I ran the test and found the following fitted model for my data. WL and AP Score had p-values less than 0.05, so they were statistically significant and I rejected my null hypothesis, while Region had a p-value greater than 0.05 and so I failed to reject that null hypothesis.

In conclusion, our model is better than guessing, but it still only accounts for 24% of the variation in data, so there are definitely other variables affecting outcome that I did not account for. This analysis is limited by the data and the previously mentioned lack of independence, but the results are still valuable. Moving forward I would be interested to collect data on things like school endowment/funding and coach expertise to see if schools which appear to prioritize their basketball program also tend to do better in the tournament, or whether something like shooting averages is a better predictor.