PlatteVillains Methodology

The methodology for selecting the 2021 NCAA March Madness bracket was determined using previous data from Ken Pomeroy for 2013 to 2019 brackets to determine the probability of different seeds to upset during a match. The twelve metrics presented for each team were then used to feed two different model streams with one being a linear discriminant analysis model in R that was benchmarked against several other models. The other model was a voting classifier model using KNN, random forest classification, and logistic regression in Python to create an average probability from the 3 models. The R model found the probability that the higher seed team would win, and the Python model found the probability that the team would advance to the next round.

Each model was trained on the historical data and used to predict the game outcomes. The two model’s classifications were compared as plots in R for us to determine where the models disagreed. The models disagreed on only 7 picks in the original 32 matchups which were the common upset matches. In this case, the raw model probability for each team in Python was compared to determine how large of a gap was present. If this gap was less than 5%, Vegas betting odds and various sports reports (i.e. Injury, suspensions, covid pauses) were used to determine the winning team with most picks going to the upset team in order to net more total points. The initial 32 advancing teams were then picked.

From here on, a Python model was created to predict who would get into the round of 16 once they made it into the round of 32 using historical data from teams that made it into the round of 32. The probability that a team would make it into the sweet 16 was then used to also predict all future rounds taking into consideration injured team members and seed differences if the probabilities were closer than 0.1. This was due to a lack of historical data that we aggregated for the elite 8 and above. Betting odds could not be used in the sweet 16 due to some of the unique matchups we created. We understand that this methodology is not completely rooted in an analytical method, but we believe that it was the best way to determine close games.

We decided on the tiebreaker by starting with the average sum of score in a championship, which is 145. We added 5 points due to the strong offense for both teams, and then added a random component by adding or subtracting 2 based on a coin flip. The coin flip resulted in an addition of two points leading to a tiebreaker score of 152 points.

### Sources

<https://kenpom.com/>

Ken Pomeroy data was used, and the team cleaned it for our uses for 2013 to 2019.

<https://www.washingtonpost.com/sports/2021/03/18/tiebreaker-national-championship-game-march-madness/>

The Washington Post was used to find the average number of points for the tie breaker.

<https://www.betfirm.com/seeds-national-championship-odds/>

BetFirm compiled several probabilities by seed for the NCAA tournament, these were used early to organize thoughts and decision making.

<https://fivethirtyeight.com/features/the-favorites-and-cinderellas-of-the-mens-ncaa-tournament/>

Five Thirty Eight is a source that compiles “traditional” basketball information that was helpful in decision making.

