March Madness 2020: What Might Have Happened?

Alex Rojas and Michael Lee

In March 2020, as public spaces were being closed nationwide due to the the increasing spread of COVID-19, sporting events became largely infeasible to hold in-person without significant risks to athletes, as well as team and venue staff. As such, the NCAA canceled the pending national championships of all remaining winter sports, as well as the entire seasons of all spring sports. Included in this spate of cancellations was the men's basketball national championship, which marks a period of basketball viewership and fandom commonly known as "March Madness," even among members of the public who do not regularly follow college sports. March Madness participants often fill out brackets predicting the winners of all 63 tournament games, sometimes comparing brackets with friends or entering betting pools. The practice is common enough that the previous President of the United States, Barack Obama, had the tradition of publicly releasing his bracket during all eight years of his presidency (as well as in the years since). In consideration of the nation's passing yearly interest in college basketball, we attempted to simulate the remainder of the canceled men's basketball games, leading up to and including the national championship. To do this, we employed a two-part model: a multilayer perceptron (MLP) to predict the outcome of each game, and a random forest regressor to model each team's changing offensive and defensive statistics with each game. The choice of an MLP for predicting outcomes of sporting events was based on scientific precedent, which showed that such a model could often outperform informed human predictions.[1]

Each year, the NCAA men's basketball tournament includes 68 teams: 32 are "autobids," or the winners of their respective conference tournament, and 36 are "at-large" teams chosen by

[1] Bunker, Rory P., and Fadi Thabtah. "A Machine Learning Framework for Sport Result Prediction." *Applied Computing and Informatics* 15, no. 1 (2019): 27–33. https://doi.org/10.1016/j.aci.2017.09.005.

committee vote and informed by the NCAA Evaluation Tool (NET), among other statistics. The "First Four" games are played between the bottom four autobids and the bottom four at-large teams for a chance at two #11 seeds and two #16 seeds, respectively. Then, the other teams, seeded in committee vote order, are paired up (#1 with #16, #2 with #15, etc.) in each of four regional tournament brackets in such a way as to minimize games between teams from the same conference.[2] The winners of the regional tournaments then meet up to play the national semifinal games (commonly known as the "Final Four," in a series of nicknames for the tournament rounds which also includes the "Sweet Sixteen" and the "Elite Eight"), and then the Final Four winners play the national championship game.

The NET was founded several years ago as an algorithm designed to rank basketball teams based on 5 specific metrics: Team Value Index (TVI), Net Efficiency, Winning Percentage, Adjusted Win Percentage, and Scoring Margin. For a given team, the TVI is a measure of how strong that team's opponents are, as well as whether the team wins and where games are played (as teams tend to play better on their own home court). Also, Net Efficiency measures how efficiently a team plays offensively and defensively by taking a difference of offensive and defensive efficiency ratings, which factor in team points per possession (calculated using a standard formula from box score statistics such as shot attempts and turnovers). Additionally Winning Percentage and Adjusted Win Percentage factor in a team's record, meaning how many games they won and lost, but Winning Percentage is an unprocessed raw feature whereas Adjusted Win Percentage takes into account where games were played, placing the most emphasis on road wins and home losses (which team obviously hope to maximize and minimize, respectively). Then, this May, the NCAA made the decision to adjust its model by only taking into account TVI and adjusted net rating, which factors in offensive and defensive efficiency with strength of schedule.

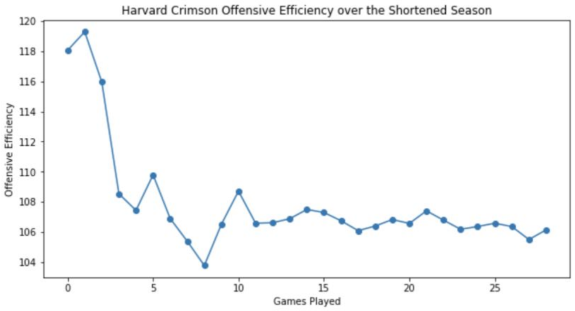
[2] "How the Field of 68 Teams Is Picked for March Madness." NCAA.com, August 17, 2020. https://www.ncaa.com/news/basketball-men/article/2020-08-17/how-field-68-teams-picked-march-madness.

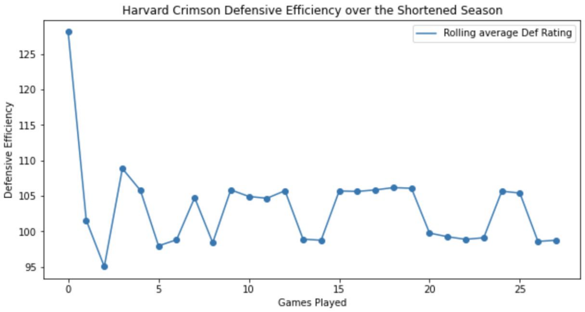
Clearly, when seeding teams for its annual tournament based on the strength of each team, it is not just key for the NCAA to take into account the face value of if a team won games or not, but also how teams perform against tougher opponents in addition to the location of a match. The NCAA tournament is played exclusively at neutral sites, where it is desired that no team has a definite upper hand leading into the game; it is objectively harder to win when a team is away from home, thus considering this in a ranking system for the tournament increases robustness. A seeding system also rewards better teams by matching them up with worse qualifying teams once the tournament begins, thus it is crucial to not just determine who can qualify for the tournament but also how powerful each team is. Because of this change in the seeding algorithm in May 2020, we infer that some of the most important statistics to consider when ranking teams are offensive and defensive with strength of schedule.

In our method, we seek to simulate the 2019-2020 NCAA season by predicting the winner of each game based on a similar heuristic: for a given team in a given game, we take into account where a game was played, the strength of the opponent and their opponents, and offensive and defensive statistics including efficiency. These statistics are readily calculated from standard box score information, which we were able to scrape from ESPN.com for 5,329 D1 games in the 2019-20 season.[3] Yet, as a season rolls on, team statistics are subject to change. This can be due to a multitude of factors, including a “hot” or “cold” stretch, where a team improves or declines significantly in performance by scoring more points or playing better defense, as well as losing players to injury and playing easier or tougher opponents. This can be illustrated in the following images, which demonstrate how a team’s offensive efficiency changes over the course of the season.

Take the Harvard Crimson men’s team. We see that there is some variability in the offensive rating throughout the season. Note that after 12 games played, we notice an improvement in their Offensive Efficiency. In their 27 game season, between games 10 and 15,

[3] e.g. https://www.espn.com/mens-college-basketball/boxscore?gameId=401173213





the Crimson were on a 5 game winning streak, and their offensive and defensive efficiency illustrates the trend well, as the offensive efficiency increases to the 10th game and the defensive efficiency falls from game 10 to game 15. Thus, there is potential to predict how a season will play out based on efficiency, and it is important to take into account when ranking teams as a season progresses.

Because our knowledge of the season ceases at the beginning of the nationwide lockdowns due to the COVID-19 Pandemic, to accurately project a winner for the season, we must add a layer of sophistication to our simulation. After each game, we must be able to project how offensive and defensive efficiency ratings were subject to change, because we have determined these to be so deterministic of a winning team. Realistically, these numbers will change based on how the season progresses; due to the Monte Carlo simulation we are undergoing, each simulation can have vastly different outcomes because they are sampled from a probability distribution. Thus, we need to be able to project the change in offensive and defensive efficiencies of a team to truly have a robust model that changes with every simulation, or we would expect many similar simulations and low variance.

The question remains: how do we best mirror the trajectory of the tournament by including change in efficiencies over time to robustly model many Monte Carlo-simulated postseasons? Calculating change in efficiencies is clearly a regression problem, and because of the ease in determining feature importance, Random Forest Regression is a particularly attractive modeling technique. However, it is important to note that there is a drawback in this technique, which is the fact that because there are thousands of trees, prediction can be slow, particularly in real-time. However, we account for this by using the Hummingbird Python package to offload our random forest regressor to tensor computations, such that the GPU can speed this up by 50 times.

Here, the model that we implement for the real time prediction of change in efficiency accepts the following inputs for every team in every game that has occurred so far, and will project based on these in the future. First, whether a team is playing at home, whether they won, how many points they scored, how many points their opponents scored, previous number of games, prior offensive rating, prior defensive rating, KRACH rating (described below), and strength of schedule. Clearly, we take into account some of the very important ideas that we outlined earlier in this model.

KRACH, which stands for "Ken's Ratings for American College Hockey," is a specific case of a more general paradigm for revealed preference ratings called a Bradley–Terry model.[4] KRACH ratings are defined implicitly for teams $i = 1, ..., n$ with number of wins V_i and numbers N_{ij} of games played between teams i and j by the relation $V_i = \sum_{j=1}^{n} \frac{K_i}{K_i + K_j} N_{ij}$. As such, KRACH's modeling assumptions are defined by the property that the ratio K_i / K_j should convey the expected win odds for team i in a matchup between teams i and j.[5] KRACH additionally has advantages over RPI for the purpose of measuring strength-of-schedule. Although the formula for ratings [6]K_i is defined implicitly, we can calculate the ratings exponentially quickly by rearranging the equation for V_i and iterating. For our purposes, we make a small adjustment to this model, whereby we assign a percentage of each game as a "victory" for each team depending on the margin of victory between the two teams, according to a sigmoid function

[4] "Bradley–Terry Model." Wikipedia, April 1, 2020. https://en.wikipedia.org/wiki/Bradley%E2%80%93Terry_model.

[5] KRACH Ratings. Accessed December 14, 2020. http://www.vaporia.com/sports/krach.html.

[6] "KRACH Ratings for D1 College Hockey." ELynah.com: Your Home for Big Red College Hockey. Accessed December 14, 2020. http://elynah.com/tbrw/tbrw.cgi?krach.

$\frac{1}{1+\exp(-\delta/\alpha S)}$ of the margin δ, the point total S, and a parameter α which controls what we consider to be a "close" game (we chose $\alpha = 5$). This also solves issues with the fact that KRACH is technically undefined for the entire league when any one team has zero wins or zero losses.[7]

In training the efficiency model, we first split up the dataset of already played games into a train and test set. Then, we loop through a list of potential max depths for each tree in the random forest without tuning any other hyperparameters. After noticing that the baseline model indicates that we can accurately predict changes in offensive and defensive ratings (which we do with separate models), we then set up a thorough grid of hyperparameters to search through for the optimal combination that maximizes the performance of the model through cross-validation. We use 3-fold cross validation and a randomized search to save computational time, as a grid search would have been overly exhaustive and with diminishing returns because the random search hyperparameter tuned model did not overly outperform the baseline models to justify such computation. After the search, saving the best model we achieve the following scores for the offensive models: a mean squared error of 3.62, and an R^2 score of .950, and for the defensive models: a mean squared error of 4.04 and an R^2 score of .938.

In order to predict game scores, we trained a multilayer perceptron (MLP) with inputs of difference in log-KRACH rating (i.e. log of expected win odds), offensive and defensive efficiencies for both teams (minus 100 for regularization), average points scored by and against both teams in previous games, and an indicator for whether the game was hosted at a neutral site. In order to incorporate sufficient randomness into our model to account for variable game factors, we observe that the differences between the score predictions from the MLP and the true game scores in our dataset are roughly normally distributed. We therefore calculated the

[7] KRACH is also undefined where there are multiple groups of teams with no crossover matchups. Practically speaking, this is never an issue for D1 basketball, with enough cross-conference games generally played within the first few weeks to generate preliminary KRACH ratings. On the other hand, there is precedent for teams finishing the regular season without a single loss (or teams without a single win).

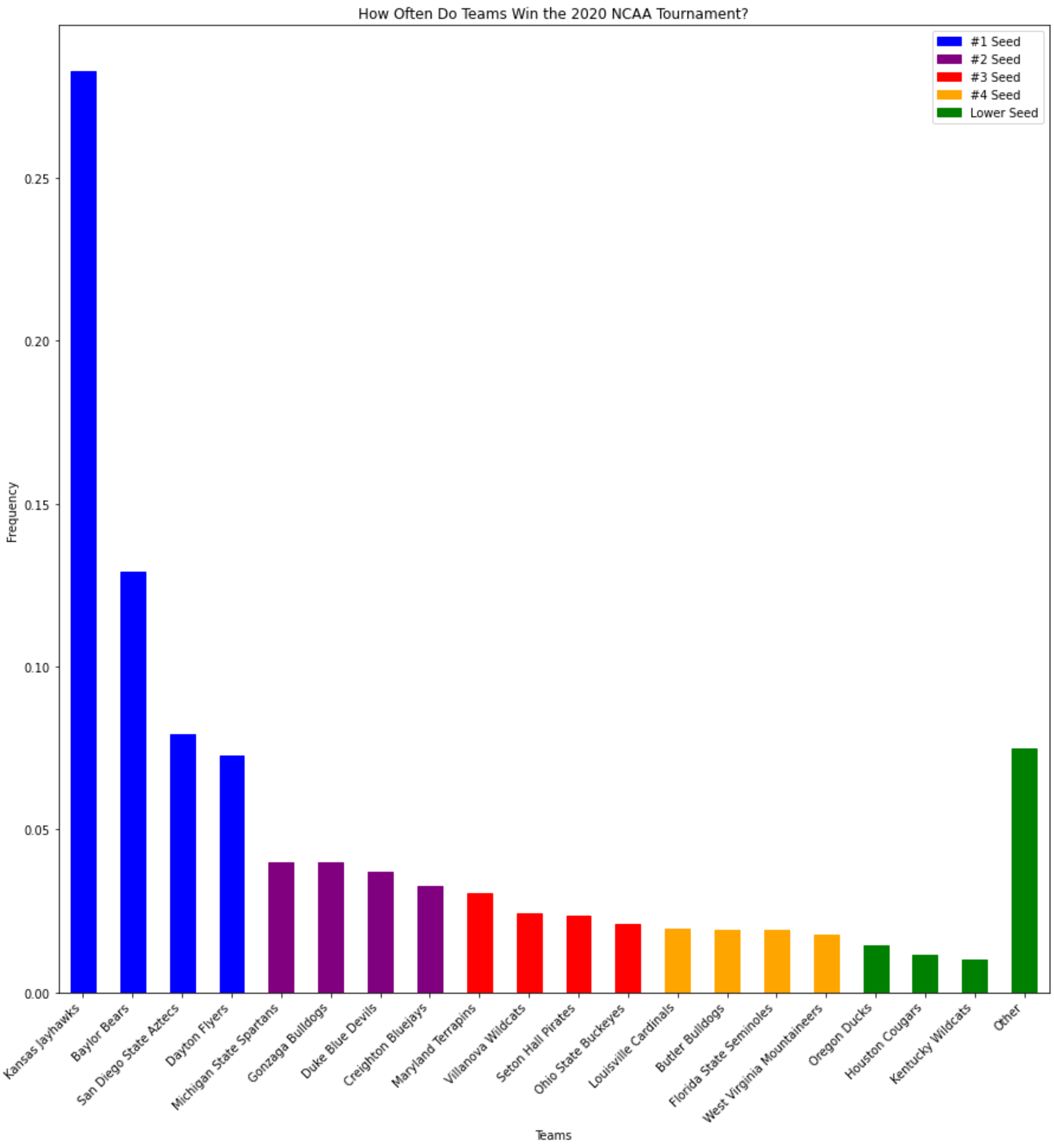
covariance matrix and mean of this distribution and sampled from it to add a random “delta” onto our score predictions. After each game, we then feed these scores back into the random forest regressor to predict the new average offensive and defensive efficiencies for both teams.

Our method for a Monte Carlo simulation of the NCAA tournament therefore follows this outline: (1) Simulate the remaining conference tournament games for the 20 out of 32 NCAA D1 basketball conferences which did not finish playing their tournaments. (2) Recalculate KRACH ratings for the 353 D1 men’s basketball teams that competed in the 2019-2020 season. (3) Automatically seed the NCAA national tournament with 68 teams (32 conference tournament winner autobids, 36 at-large bids determined by KRACH rating). (4) Simulate the resulting 63 NCAA tournament games.

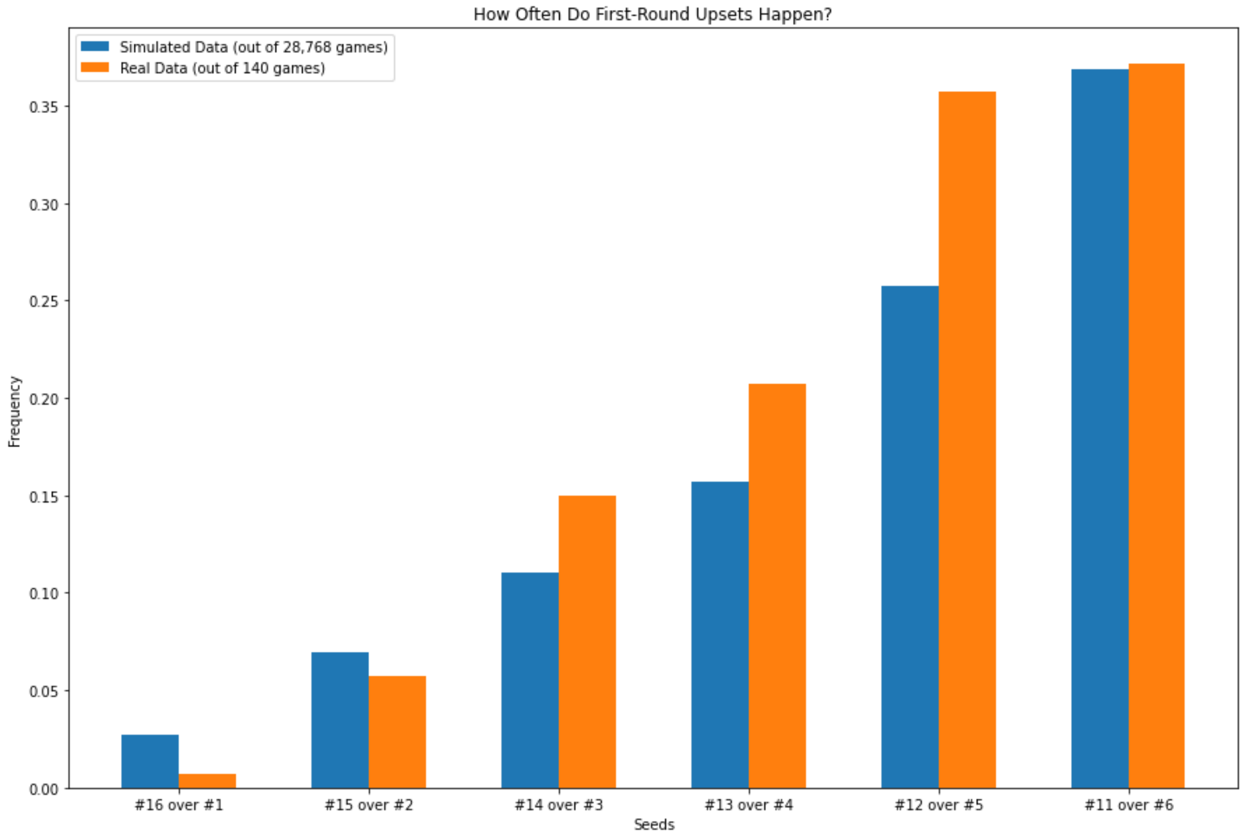
In order to generate a regional assignment for the national tournament that minimized regional matchups between teams in the same conference, we generated 10,000 regional assignments at random and automatically selected the one that adhered most closely to NCAA guidelines. For this project, we performed the above simulations ~7,200 times. Some salient statistics from our Monte Carlo experiment are as follows:

By far the most likely team to win the tournament in our simulations was the Kansas Jayhawks, which was the winner in 28.3% of our simulations. Kansas was ranked #1 by both the Associated Press (which polls sports journalists) and USA Today (which polls basketball coaches)[8] in the final published rankings of the 2019-2020 season. This is borne out in the KRACH ratings including all actual game results, which give Kansas roughly 3:2 direct odds over the next-best team and Kansas’s Big 12 conference-mate, the Baylor Bears. In all, #1 seeds won the NCAA tournament in 56.4% of simulations. In reality, 20 of the 34 NCAA tournaments in the 64-team-tournament era (1985-present) have been won by #1 seeds.

[8] https://www.espn.com/mens-college-basketball/rankings/_/year/2020



Another interesting point of study is the rate of first-round upsets. Fans of March Madness will recall that in 2018, the #1-overall-seed Virginia Cavaliers were upset in the first round by the bottom-seeded UMBC Retrievers, marking the first and only time in 132 (now 140) #1-#16 matchups that a #16 seed has prevailed. Our simulations predict a slightly higher rate of upset for #16 seeds over #1 seeds, at 2.7%, but the fact that there is slight divergence here is unsurprising considering the small sample size for real games relative to the Monte Carlo simulation. The rates of other upsets in our simulations were roughly on-par with real numbers.



In conclusion, we have developed a paradigm for simulation of NCAA basketball tournament results that we feel is indicative of real statistical trends and therefore is a valid means of estimating tournament outcome likelihoods with relatively high accuracy compared to informed human prediction. The broad accessibility of these methods to data scientists and programmers has potentially meaningful implications for industries such as sports betting and sports journalism (where outlets like FiveThirtyEight are mainly using human-tuned models for predicting the NCAA tournament). Furthermore, considering the wealth of sports statistics available in the present day, this dualistic method of neural network score prediction and regression to determine updated performance statistics could be applied easily to other NCAA team sports whose championships were canceled, such as hockey, and potentially individual sports such as wrestling.