**Testing the Limits of College Basketball Playoff Predictions**

**Ryan Presnell**

Introduction

The NCAA men’s basketball tournament notoriously difficult to predict. No one has ever created a perfect bracket— in fact, the longest streak of verifiably correct predictions is 49 games into the tournament without an error[[1]](#endnote-1) (for comparison, there are 67 games in a tournament). Part of the reason for this is that the “best” team doesn’t always win. College basketball fans will remember 16 seed UMBC’s win against 1 seed Virginia back in the 2018 tournament. On average, there are a little above 12 upsets per tournament.[[2]](#endnote-2) In light of the prevalence of unexpected game outcomes, how effective can statistical models be at predicting tournament outcomes? This is the question that will be explored in this paper using Andrew Sundberg’s college basketball dataset from Kaggle[[3]](#endnote-3) and machine learning algorithms such as K-nearest neighbors and decision trees.

Methods

Feature Selection & Concerns

Considering that team statistics from tournament games are included in the calculations for this dataset, the applicability of the results of any models using this data for future predictions must be called into question due to the increased intensity of playoff games. Ideally, I would want to use team statistics that only include “regular season” (including conference tournament, but not NCAA tournament) games. However, the playoff intensity likely does not impact the data enough, especially given the small number of tournament games played on average for each team, to drastically alter prediction efficacy. To ensure that the modeling results retain as much applicability as possible for future predictions, certain features must be altered or removed entirely. Each team’s respective wins and games played either need to be adjusted to their pre-NCAA tournament levels or taken out of consideration for prediction. A team winning the tournament will add 6 wins to their win total (or, although it has never happened,[[4]](#endnote-4) 7 games if they are a 16 seed). This would bias predictions, causing teams who have not yet played in the tournament to be predicted to be knocked out earlier because they have less win totals. For the sake of having measurements that were all collected at the same time (as opposed to having wins measured *before* and other stats measured *after* the tournament, resulting in less interpretability), I elected to remove the games played, wins, and win percentage columns. Since K-nearest neighbors uses Euclidean distance to “vote” on a class, categorical features such as conference and seed (many teams miss the tournament and do not even have seeds), especially being non-binary, would not make sense to convert to a number scale—thus, these features were also removed to ensure a fair playing field for all modeling methods for comparison. A complete list of features is available for view on [Kaggle](https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset).

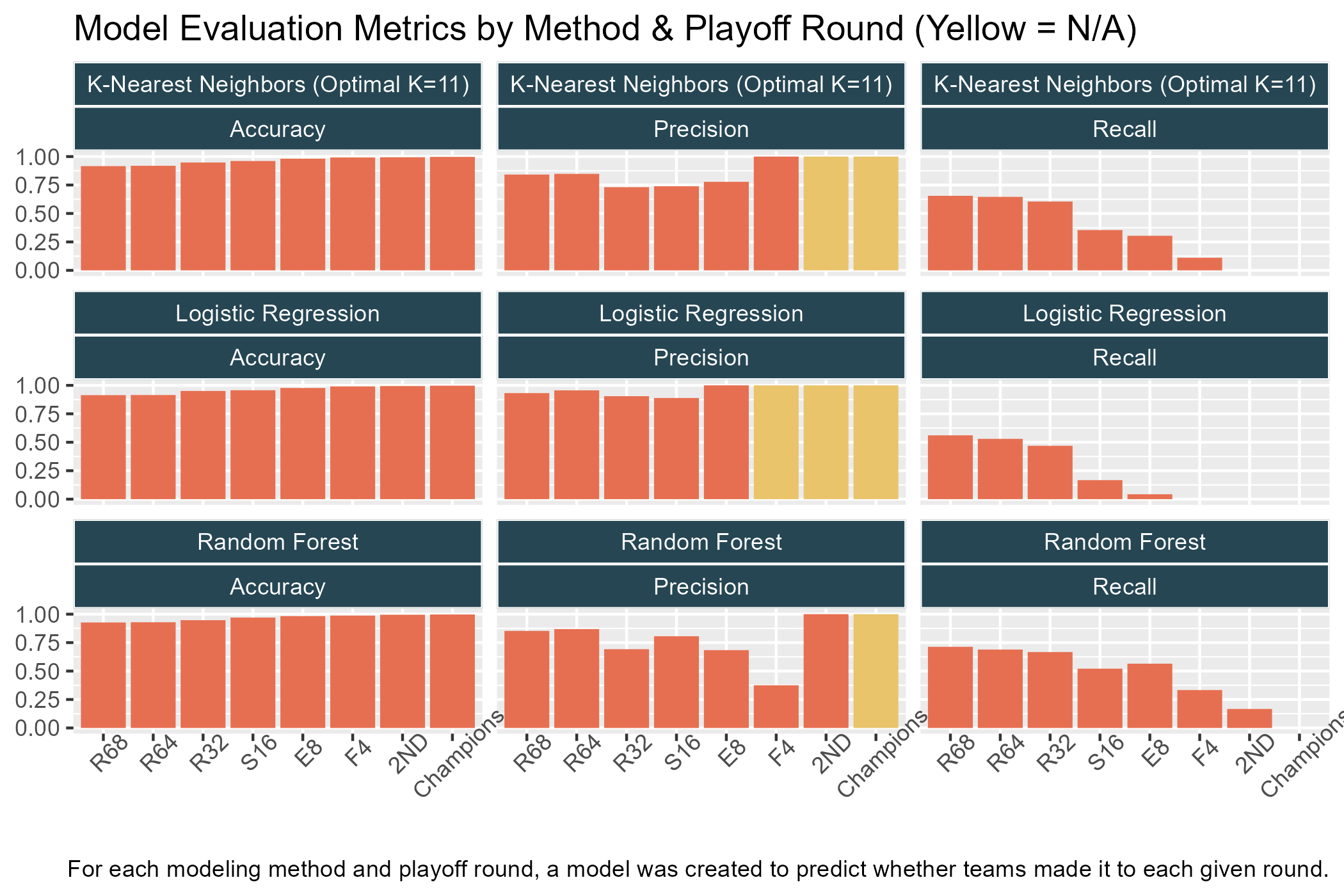
Predicting Team Playoff Exits

First, I decided to attempt to predict the round in which each team exited the tournament (or won), or whether they missed the playoffs entirely. I used four modeling methods for this section: K-nearest neighbors, Naïve Bayes, a classification tree, and random forest. An optimal K-value (11) for the K-nearest neighbors model was found using leave-one-out cross validation. While this K-value resulted in the highest accuracy (82.8%) among KNN models, it also made it more difficult for the model to correctly predict teams to have exited (or won) the tournament in less populous rounds—the model predicted that 0 teams would exit in the round of 68 and the championship, although it did predict that two teams would win the tournament. Naïve Bayes did correctly predict a second-place finish; however, it had the least accuracy (71.6%) with all other methods being at least ten percentage points higher. The classification tree rivals the Naïve Bayes model in terms of lack of usability. While it has much better accuracy on the testing data (83.4%), it only classified teams into two classes: missing the playoffs and exiting in the round of 32. This is not helpful at all for future prediction in terms of filling out a bracket. The random forest model performed the best in terms of accuracy (83.9%) and even correctly predicted a team to have won the championship. Its confusion matrix is pictured below:

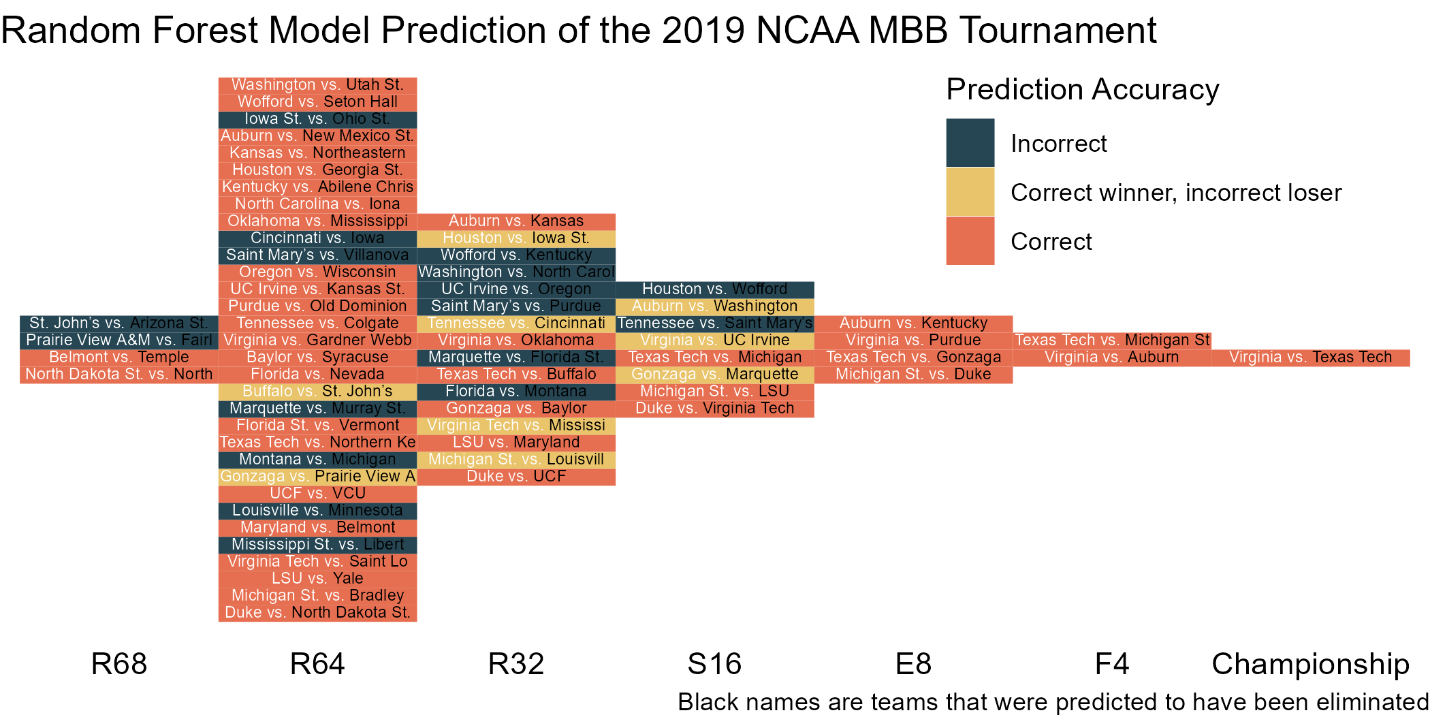


Predicting Whether Teams Survived to a Given Playoff Round

Intuitively, it seems that predicting whether a team would survive at least to a certain round would be easier than predicting the exact round that a team would exit or win the tournament in. Since this is a binary classification problem, I fit a logistic regression model along with a random forest and K-nearest neighbors model (the two best performing modeling methods from the previous section) for each of the 8 rounds of the playoffs. There is no model for whether a team *at least* missed the playoffs because every team either 1) missed the playoffs or 2) did better, i.e., made the playoffs (probability=1). The models seem to confirm the intuition that a binary classification problem would be “easier” in this scenario—the minimum accuracy on the testing data is 91.4%; however, the classification problem becomes increasingly unbalanced as you increase the tournament round. Thus, more criteria (I used precision and recall) are needed for model evaluation. Although the random forest models did not have the best precision values for the elite 8 and final 4, I think it was the method that performed the best given that it was able to identify more teams that survived to the higher rounds.



Re-Predicting the 2019 NCAA MBB Tournament

Since the random forest models have performed the best in each of the previous two sections, I decided to use it to re-predict the NCAA men’s basketball tournament from 2019. I arbitrarily selected the year 2019 because it is the most recent year in the dataset. I re-fit the random forest model on all the data minus that from the year 2019 to have a larger training set as well as to avoid “cheating” by training the model using data that includes the 2019 data values I would be predicting on. Predictions were done by comparing the predicted classes (playoff rounds) for each team. Whichever team had the higher round prediction was the predicted winner. In the event of a tie, an automatic misclassification was substituted for a random guess to prevent the model from having a higher - misleading - accuracy level by randomly guessing. The random forest model predicted 41 games correctly (61.2%), 17 games incorrectly (25.4%), meaning that it predicted the wrong team to win, and 9 games with the correct winner but incorrect loser (13.4%). All the games from the elite 8 onward were predicted correctly. A visualization of the model’s performance can be seen below:

Conclusion & Discussion

Predicting exactly in which round a team exited or won the NCAA tournament was difficult to do accurately. Predicting instead whether a team survived to a given round was more likely to be accurate, but even this suffered from poor results in the highest rounds of the tournament. Random forest models performed the best among those utilized and, in re-predicting the 2019 NCAA tournament, predicted 61.2% of games correctly and all games from the elite 8 onward. Further analysis should refit these models on data that does not include teams that missed the playoffs and compare results. I would also like to see other NCAA tournament re-predictions since the success of the random forest model in the higher rounds of the 2019 tournament is most likely due in part to the fact that I happened to pick the 2019 tournament in particular.

1. <https://www.ncaa.com/news/basketball-men/bracketiq/2022-03-14/longest-ncaa-bracket-has-ever-stayed-perfect> [↑](#endnote-ref-1)
2. <https://www.ncaa.com/news/basketball-men/bracketiq/2018-03-13/heres-how-pick-march-madness-upsets-according-data> [↑](#endnote-ref-2)
3. <https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset> [↑](#endnote-ref-3)
4. <https://www.ncaa.com/news/basketball-men/article/2022-03-27/lowest-seeds-make-mens-final-four-elite-8-and-sweet-16> [↑](#endnote-ref-4)