David Nemirovsky

BISTP8106 – Professor Sun

Midterm Report

3/29/21

Predicting the Outcome of the 2020-2021 NCAA DI Men's Basketball Tournament

Introduction

Every year in March, NCAA Men's DI College Basketball brings excitement to the sports world in the form of a 64-team (68 if you include the first four play-in games) single-elimination tournament in what is known as March Madness. The tournament is broken up into four regions where teams are seeded 1-16 in each, and play each other in a bracket to advance to the Final Four (where the winners of each region play against each other). This time in sports is referred to as March Madness because often times, lower seeded teams and teams that are unlikely to win match-ups, end up going on runs to advance deep in the tournament despite professional rankings and predictions. The aim of this project is to build a model that predicts the winner of this year's tournament, given regular season and tournament stats from years prior.

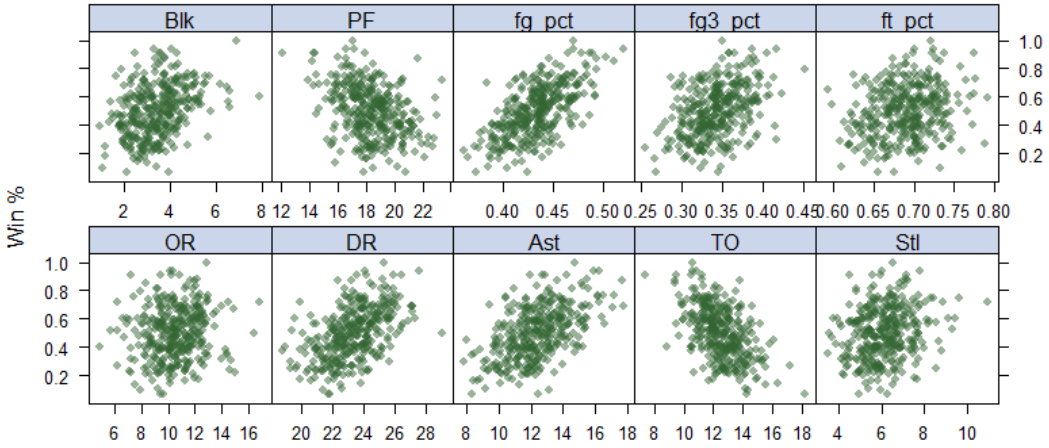
Data was acquired from Kaggle.com's March Machine Learning Mania 2021 – NCAAM competition.[1] For this model, regular season and tournament game data was used from 2015 until this season. For the regular season, data was collected from every game played, and simple stats were recorded for each team in the match-up (points scored, made field goals, attempted field goals, made 3-point field goals, attempted 3-point field goals, made free throws, attempted free throws, offensive and defensive rebounds, assists, blocks, steals, turnovers, and personal fouls). The data was tidied and organized in a way to have the result (win or loss) and stats for a team grouped with its opposing team's stats within a specific match-up. Then, differences in per-game stats were taken across all games in each regular season. That way, models can be fit using game result as the outcome to predict, and stat differences as the predictors. Additional modification included transforming field goal and free throw counting stats to percentage stats, just based on basketball intuition (shooting percentages give better information than shooting count statistics). Data tidying was done using the `tidyverse` package in R and the code can be viewed in this report's additional documents.

Methods

I. Exploratory Data Analysis

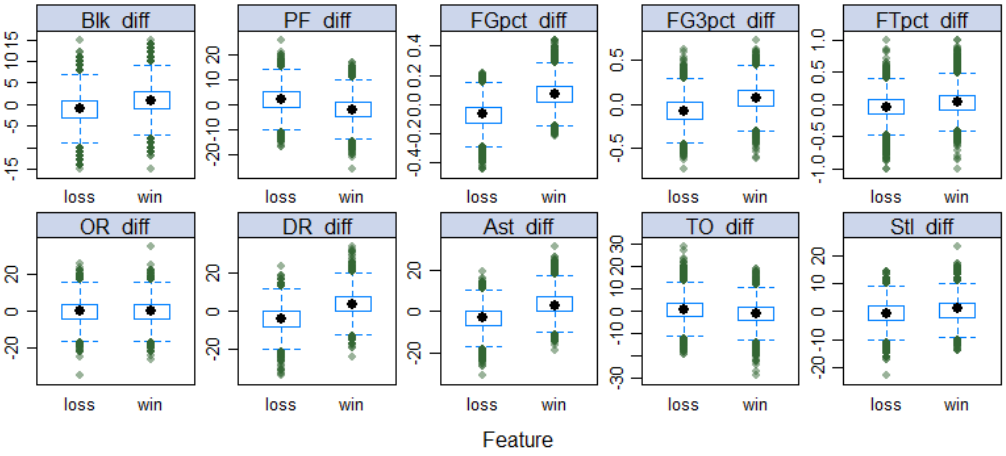
During EDA, it was found that certain predictors may have more of a correlation with game outcomes than other predictors. In this context, the predictors are the stats field goal percentage, 3-point field goal percentage, free throw percentage, offensive and defensive rebounds, assists, blocks, steals, turnovers, and personal fouls committed. The outcome of interest is game result which could be either win or loss (R automatically coded 0 as a loss and 1 as a win). To assess correlation between game result and the various predictors, a new data frame was constructed, obtaining win probabilities for every team, grouped by season, using their average stats per game. According to Figure 1, depicting the scatter plots of these predictors versus win probability during the regular season, it appears as though field goal percentage has a strong positive relationship with win percentage and turnovers have a moderate to strong negative correlation with win percentage. It also appears as though all of the other predictors have a moderate to weak correlation with win probability in the positive direction, with the exception of personal fouls committed.

Figure 1 – Scatter plots of win probability across different seasonal average in-game stats.



Further EDA revealed that when grouping for wins/losses across the six regular seasons (Figure 2), field goal percentage differences were much higher for winning teams than losing teams. In addition, the personal foul difference seemed to be higher for losing teams than winning teams.

Figure 2 – Distributions of game-stat differences, grouped by outcome.



II. Cross-Validation

Regular season data was used as the training data in model building. Three different models were fit using 10-fold cross-validation: logistic regression, multivariate adaptive regression splines (MARS), and k-nearest neighbor (KNN) for each regular season (2015-2020). The best tuning parameters for the MARS model came from 1 degree and 15-18 model terms, as shown in Figure 3a. The optimal tuning parameter, k, for the KNN models were for values of k in between 40 and 70, shown in Figure 3b. For model selection, resampling was done for each season's models and the area under their receiver operating characteristic (ROC) curves was used as the metric for comparing models. The higher area under the curve (AUC) indicated a better fitting model. Figure 4 shows how well the three models do for every season. It can be seen that across every season, the GLM using logistic regression generates the best model, yielding in the highest AUC values. CV and resampling was done using R's `caret` package.

Figure 3a – MARS optimal tuning parameters

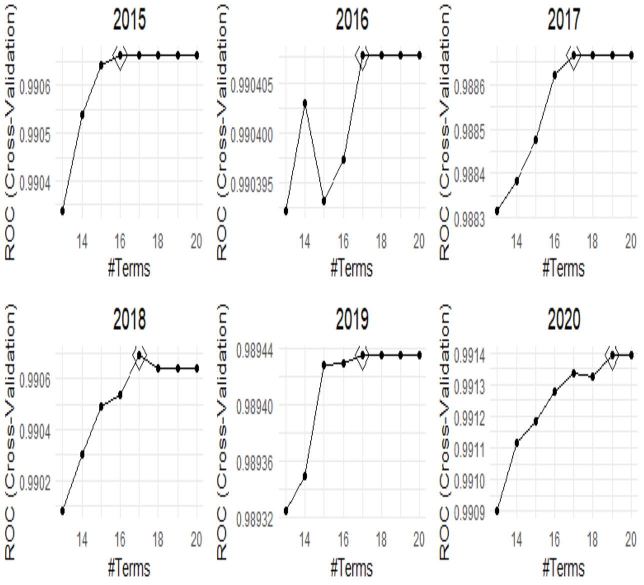


Figure 3b – KNN optimal tuning parameters

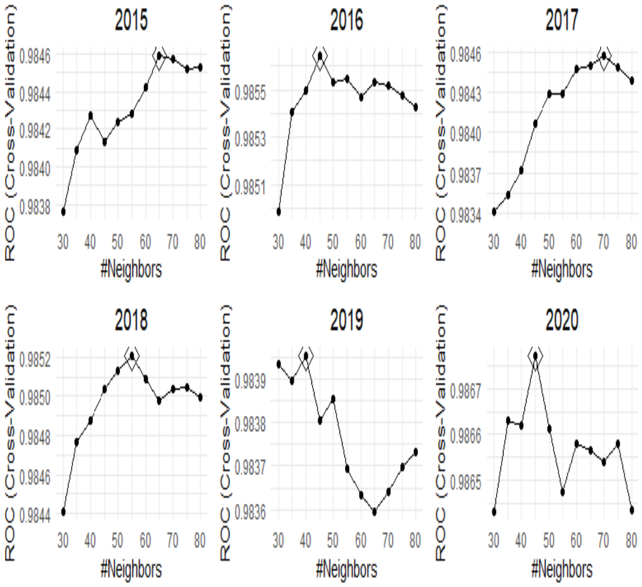
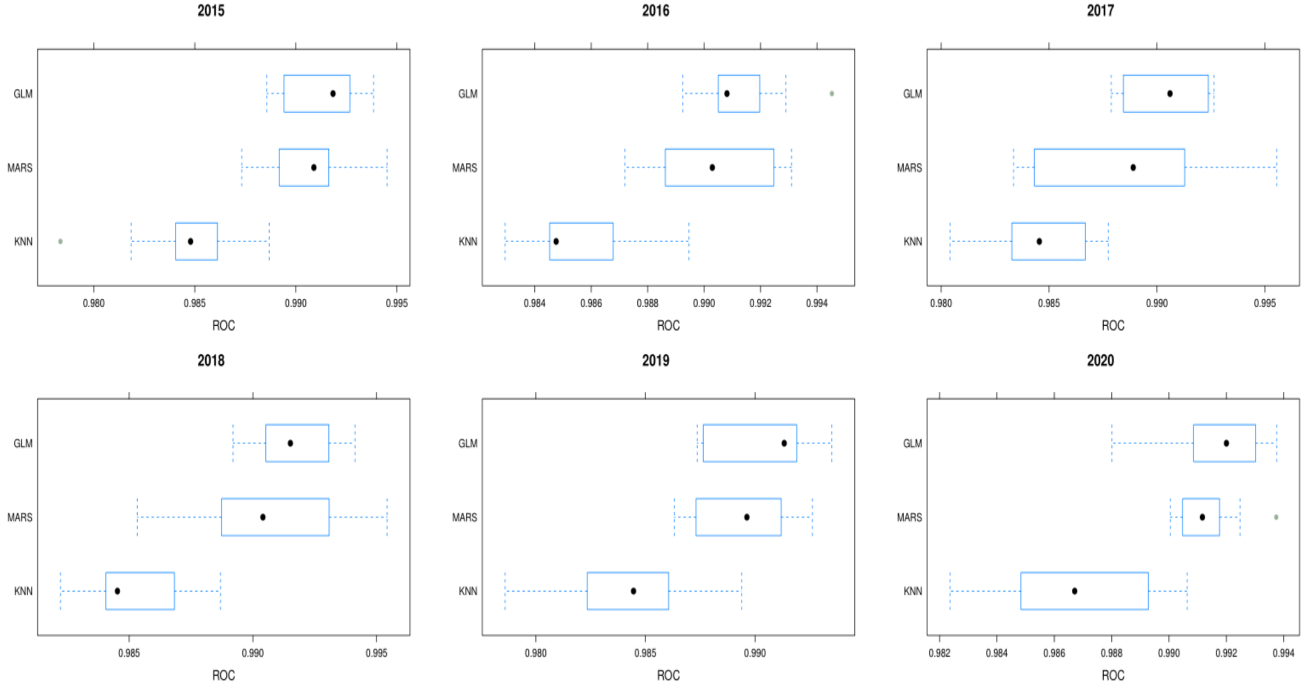


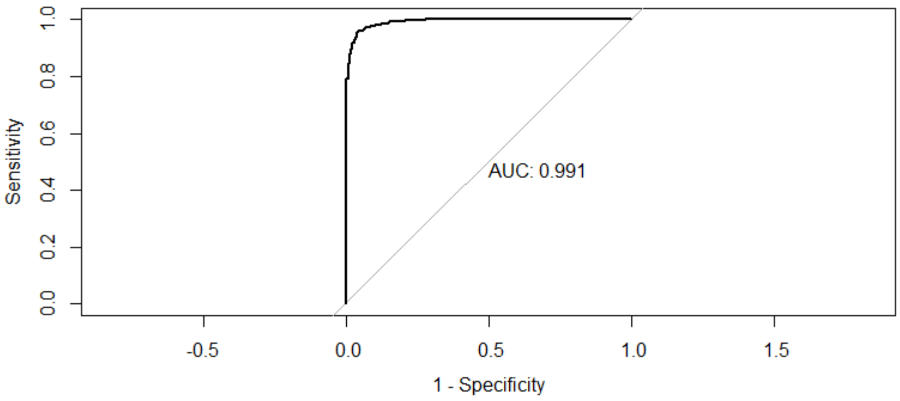
Figure 4 – Box plots of 10-fold CV resampling for the three different models



III. Predicting Tournament Games in Past

Using tournament data from the past for the 2015-2019 seasons as test data, the logistic regression model was fit to determine how well this model performs on tournament data for each season. R's `predict( )` function was used to get predictions across seasons, and then an ROC was generated for all seasons combined. Figure 5 shows the ROC of the logistic regression model on the 2015-2019 tournaments. The model gave an AUC value of 0.991, indicating a good fit.

Figure 5 – ROC of logistic regression model on tournament data (2015-2019)



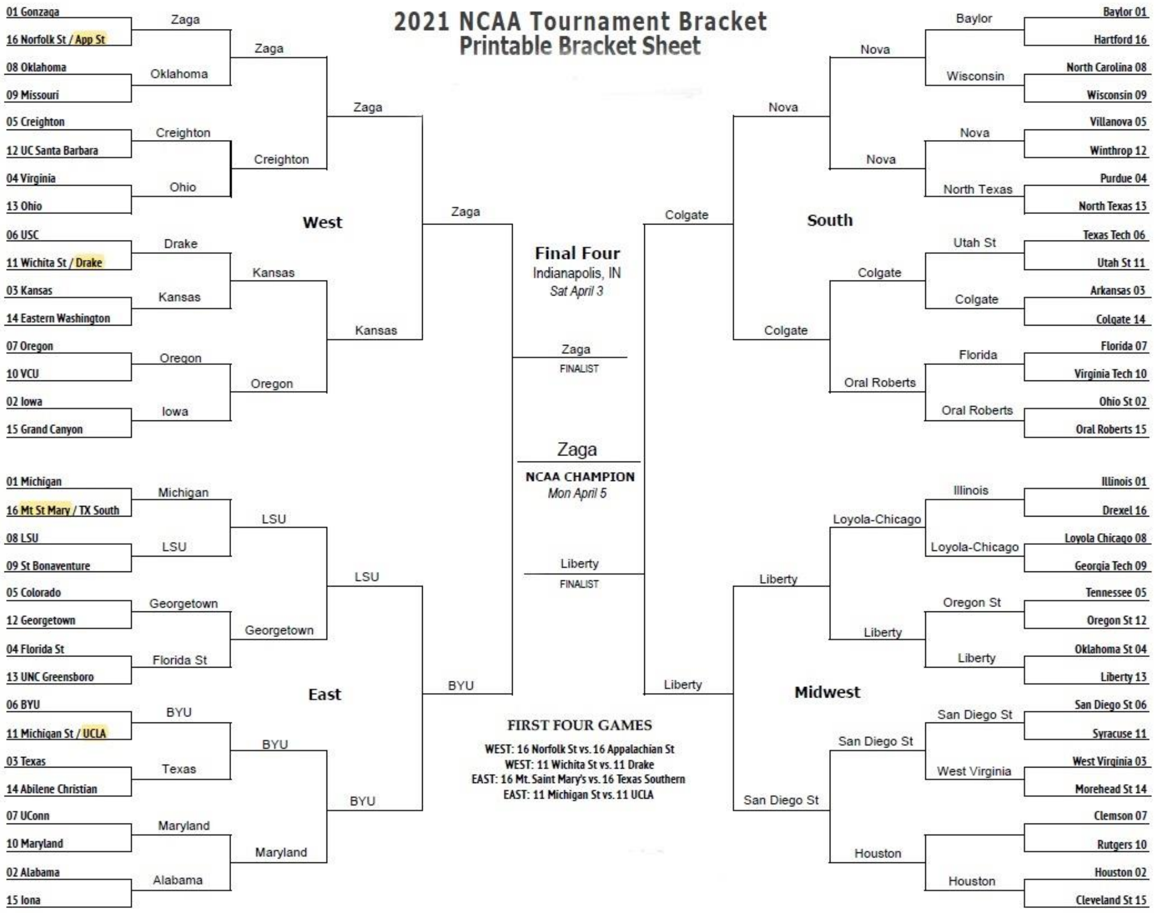
IV. Simulating the 2020-2021 Tournament

To simulate the tournament this year, a data frame that included team seeds and regions was built, combining 2020 regular season data with seed and region information obtained from Kaggle.com. A function, `matchup( )`, was then created, taking two TeamID inputs, running the logistic regression model on their respective stat differences for the 2020 season, and then outputting the TeamID of the team that the model predicts is going to win that match-up. The function was run on the first four play-in games in order to obtain a final, set bracket of team match-ups. Once all the teams were determined, a second function was built, `region_sim( )`, which takes a region (East, West, Midwest, or South) input and outputs the results of that region, giving winners of every match-up in that 16-team region. This function was made by pre-setting match-ups based on initial seeds, and then each subsequent match-up was determined by the previous outcome. The first user-created `matchup( )` function was incorporated into the `region_sim( )` function, ergo the logistic regression model was used to predict the winners of each region. A final function, `f4_sim( )` was then created, similarly to the `region_sim( )` function, which takes each region as an input and outputs the winner of the final match-ups (two semi-finals and one final game).

Results

Each game in the bracket was simulated using the logistic regression model with win probability as the outcome and average in-game stat differences for the 2020 regular season as various predictors. The simulation was only conducted once, as win probabilities were fixed, and resulted in Gonzaga winning the 2020-2021 Men's DI National Basketball Championship tournament. The simulation also had surprising results with 6th-seeded BYU, 13th-seeded Liberty, and 14th-seeded Colgate reaching the Final Four (national semifinals) along with 1st-seeded Gonzaga. The full bracket predictions based on the model are shown below in Figure 6.

Figure 6 – Predicted bracket results of the 2020-2021 NCAA DI Men's Basketball Tournament.[2]



Conclusion

Although the overall tournament #1 seed, Gonzaga, is predicted to win the tournament based on this model, the other results seem highly unlikely because double-digit seeds almost never make it to the Final Four. Since expanding to 64 teams in 1985, only 5 teams seeded #10 or higher have made it to the Final Four (5/140 = 3.6%), and none of them were higher than an 11-seed.[3] Therefore, this model probably won't do the best job in predicting the Final Four or the other games in the tournament. The model failed to account for strength of schedule differences among teams, which could have inflated the stats of some higher-seeded teams, leading them to victory in the model predictions. A way to account for this in the future would be to standardize a ranking based on strength of schedule for each team, which would adjust the predictors by giving higher weight to the stat differences of teams who played tougher match-ups during the regular season than teams who did not have as tough of competition. Another way to make this prediction better would be to incorporate win probability variance into the simulation, and then run multiple simulations to see which team wins the tournament the most number of times.

References

[1] Kaggle.com. "March Machine Learning Mania 2021 – NCAAM competition". Visited 18 March 2021. https://www.kaggle.com/c/ncaam-march-mania-2021/data

[2] TeamRankings. 2021 NCAA Tournament Bracket Printable Bracket Sheet. Visited 21 March 2021. https://www.teamrankings.com/ncaa-tournament/NCAA-Tournament-Printable-Bracket-2021.pdf?ver=1615768848

[3] Wittry, Andy. "Why you should probably pick 2 No. 1 seeds in your NCAA bracket this season". 14 March 2021. Visited 25 March 2021. https://www.ncaa.com/news/basketball-men/bracketiq/2021-02-17/heres-how-many-no-1-seeds-you-should-pick-your-ncaa-tournament-bracket#:~:text=1%20seeds%20have%20made%20the%20Final%20Four%20since%20the%20NCAA,to%2064%20teams%20in%201985.&text=If%20you%20pick%20two%20No,of%20picking%20the%20right%20team.