Using Logistic Regression to Predict March Madness 2021

March Madness brings in over a billion dollars to the NCAA annually by itself, not including the regular season. It has the most widely watched collegiate sporting events along with the College Football Bowl Season, but what the tournament brings that no other sport offers in high level D1 is a massive 68 team bracket. The bracket of teams is an instrument for workplace and friendly bracket challenges to high stakes betting. Being able to guess game winners and losers better than others is met with great respect and bragging rights for the next year. In this analysis, 2003 through 2019 March Madness team data will be used to construct a logistic regression model to predict the results of the 66 2021 March Madness games. The model will be graded using a Cross-Entropy formula.

Section I. Variable Creation and Model Building

Variable Creation

The data was organized as each line being a game where the two teams were randomly assigned team A and team B. The regular season statistics, along with their seed and region, for each team were included and whichever team won the contest. First, a factor variable was created representing whichever team won the game; 1 being that team A won the game and 0 being that team B won.

After creating a binary response variable, it was sought after to create more predictors stemming from the original data. Although only one of these new variables was used in the final model this was a crucial step in building an understanding of the data as well as the impact of Cross-Entropy. Two new types of variables were made: rate data (such as Field Goal Percentage) and combined difference data (such as the difference between seed).

The rate data was created by taking the shots made over the shots attempted for Field Goals, Free Throws and Three-pointers. The difference data created by taking every stat for team A and subtracting it by Team B's stat. This was done for the rate data as well. B subtracted from A was always the order for consistency and simplicity's sake. The thought process was to shorten the model and eliminate an unnecessary large number of predictors, as well as both teams being compared with a single coefficient. It would later come to light that this dampens the predictive quality of the predictor, but this may have been due to a worse model fit with the test data.

After the new variables were created, a simple loop was constructed to run the following Cross-Entropy formula:

$$\frac{1}{T}\sum_{i=1}^{T} -(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

Where y_i is the true outcome of the game (0 or 1) and p_i is the projected outcome of the game by the model (A number between 0 and 1). The goal is to minimize the Cross-Entropy score. Data was split into a training set of 2003 to 2018 games and a test set of 2019. Each variable was fitted into an individual model of like variables; as in team A and B's offensive rebounds were in a model by themselves, same goes for the Field Goals made and attempted, etc. The purpose of this format was to determine which version of each variable is best suited to satisfy a minimum Cross-Entropy (as that is the goal).

This process was repeated for every variable; for example, three models were considered for Three-point Field Goals:

1) The original total makes and attempts data for team A and B.

2) Team A and B three-point percentage.
3) The difference between team A and B three-point percentage.

All three models were run through the Cross-Entropy formula and the model with the lowest score's variables would move on to be considered in model selection. (A March Madness of variables if you will). For most of the variables, the original data received a better score than the new and when the new variable did receive the lower score, in all cases but one, the variable ended up being an insignificant predictor on multiple measurements during variable selection.

Model Building and Variable Selection

A model was fit with all predictors that moved on from the first level of Cross-Entropy comparison, aside from the Region of the team. Region was eliminated early due to similar predictive ability to guessing; whether a team wins and where the team is from are not correlated. Variables in the large model were compared in terms of deviance and AIC in the drop1 R procedure, but mainly the bestglm function was used. The bestglm function is the best subsets procedure specifically designed for glm and logistic regression models. Best Models were produced and compared with AIC and BIC. Using this along with experimentation with the Cross-Entropy formula the following model was found.

The following is the final model built on the March Madness data from 2003-2019, it had the lowest AIC as well as the best Cross-Entropy score. Let P(Y) be the probability that team A won the game, X_1 be the difference in field goal percentage between team A and B, X_2 be the offensive rebounds per game by team A, X_3 be the offensive rebounds per game by team B, X_4 be turnovers per game by team A, X_5 be turnovers per game by team B, X_6 be steals per game by team B, X_7 be the seed of team A, and X_8 be the seed of team B. The logistic regression model is

$$\ln\frac{P(Y=1}{P(Y=0)} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8$$

Where β_0 is the intercept, β_1 is coefficient for field goal percentage difference, β_2 is the coefficient for offensive rebounds by team A, β_3 is the coefficient for offensive rebounds by team B, β_4 is the coefficient for turnovers for team A, β_5 is the coefficient for turnovers by team B, β_6 is the coefficient for steals by team B, and β_7 and β_8 are the coefficients for the seed of team A and B respectively.

The model is interesting in that the steals was only significant for team B. Whenever the combined stat was used the prediction and Cross-Entropy Score was always worse, so the AIC and deviance was respected and steals for team B alone was considered. The most influential predictor for any March Madness game is the seeding. Figure 1 shows the number of games won by each team at each seed from 2003 to 2019.

In figure 1, the pink bars represent team A's wins and the blue represent team B's. The teams were separated in order to get a feel for the data and how it was split. What is interesting about this data is that the teams follow the assumed distribution of worse teams winning less but the data seems to reverse course at seed 11 and again at seed 16. This is due to how the bracket itself is designed; in some ways a team may be better off being a worse seed due to the matchup after round 1. Such us middle seeded teams have the highest likely chance of facing the number 1 seed in round 2. The seeds reflect the chaotic quirks of the structure of the tournament which adds all to the fun and excitement.

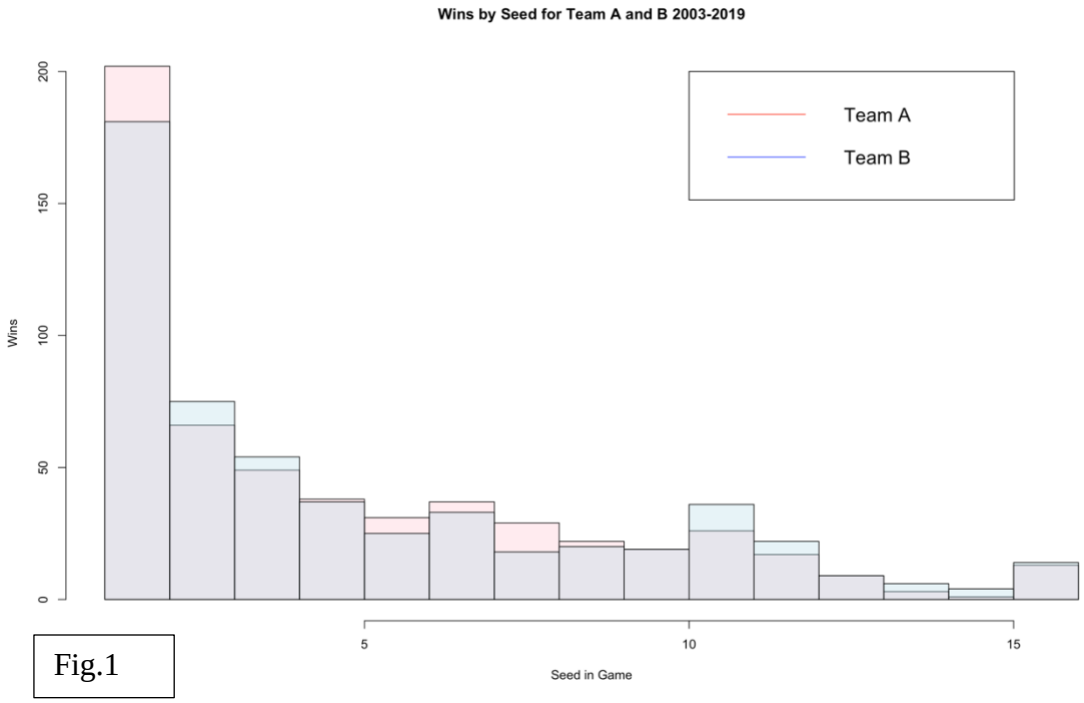

Fig.1

Hopefully going the extra mile and comparing each variable with Cross-Entropy will benefit the model's predicted ability and score to others in the class. It was decided that using Cross-Entropy during variable selection would be the strategy employed as a score is the tool used for comparison between models. Field Percentage Difference is probably the most important factor in the model as hopefully others in the class do not have it. If few or no one has that variable, and it is as influential as is believed then that will bode well for its rank of prediction and score.

Section II. Model Predictive Performance

To test the predictive performance of the model was measured by the data from the 2021 March Madness tournament. This data was not included in the fitting of the model, only the 2003-2019 data was used. The following is the confusion matrix of predicted outcomes versus the true outcomes of the tournament games.

		True Outcome	
		A Wins	B Wins
Predicted Outcome	B Wins	7	**29**
	A Wins	**14**	16

The above confusion matrix results in a 31.8% misclassification rate with 43 out of 66 correct predictions. The model was also evaluated using Cross-Entropy, same as the models before during the building phase. This model received a score of 40.57 or 0.615 when divided by the number of observations which is higher than the levels found when predicting the test data,

but that is to be expected due to introduction of new data. During the building phase, the best Cross-Entropy score given was a 33 or 0.55, but that was with the model with every variable, which would be unrealistic to be a quality predictor to new data. The final model received a 35 or 0.58 when predicting the 2019 test data.

These results bode well for predicting the 2022 tournament. Perhaps incorporating a few interactions or squared values could have been beneficial, but creating the new variables allowed for more creativity. Even though it still felt like even more could have been done variable creation wise. Time constraints prevented further experimenting as running comparisons for every variable with Cross-Entropy was very time consuming. It is to be determined how the score of this model compares to others in the class, but a similar score can be expected when the 2022 data is released.