NCAA March Madness Insights & Predictions

**Team Members**: Zhengchao Ni, Zhefu Peng, Eddie Wu

**Overview & Objectives**

Our group will be analyzing the past data of Division I college men's basketball teams, including both the regular season and tournament statistics, from which we plan to discover aspects of the basketball game that have significant impacts on the outcomes and can likely lead to upset in the tournament. All of these past data will be used as the training set, and the 2017-2018 season games and/or tournament games will be used as the testing set for evaluation of our chosen classifiers. Our preliminary choices of classifiers include Logistic Regression, Random Forest / Gradient Boosted Trees and Neural Network. For the prior two choices, we plan to build the feature vectors consisting of the basic basketball statistics for every pairs of college team (for each year in the season and in the tournament) that played against each other, and the associated target variable would just be 0 (higher seeded team, or the team with better records, wins) or 1 (lower seeded, or the team with worse records, team wins). For Neural Network, the feature vectors would be similar as above, but the target variable would be win-lose pair ((0, 1) or (1, 0), where 1 stands for wining, and 0 for losing) where each entry/game is recorded twice (team1+team2 with (1,0) and team2+team1 with (0,1)). Zhefu will focus on data cleaning, data preparation and some parts of exploratory data analysis; Zhengchao and I will also help with parts of exploratory data analysis to identify significant aspects of the game on outcome, in addition, I will focus more on applying Logistic Regression and Random Forest modeling to the dataset while Zhengchao will focus more on using Neural Network.

**Data Description**

We have two sources of data. First source is from kaggle website (https://www.kaggle.com/c/march-machine-learning-mania-2017/data#), which include the outcomes of games from year 2003 to 2017. Besides the final outcome, the dataset contains the specific value such as the number of rebounds, assists, turnovers, etc, per game. Second source is from "https://www.sports-reference.com/cbb/seasons/2018-school-stats.html", which contains the average basketball statistics information for every team in the 2017-2018 regular season. We plan to merge the tables we need to obtain a comprehensive table through some common columns among different tables (such as team id or team name), from which we could clearly specify the relationship between certain aspects of the games and the outcomes. From the merged table there may be some missing values and abnormal values, so we can just delete them (if there are not a lot) or replace them with mean/median values to keep the data complete. For redundant data, we need to do correlation analysis to lower the data redundancy or reduce the dimension of feature space if applicable. For some features with values that may be too large or too small making the analysis more difficult, we need to do normalization to convert the data to be on comparable scale. Then we plan on adding labels to each entries in the table specifying the outcome of the games, which will be useful later for modeling/prediction.