Non-Technical Summary of NCAA March Madness Prediction

NCAA March Madness Prediction is one of the most popular predictive analysis case each year. The most exciting part of March Madness is the rule of single game elimination, in that, every team has to play their best game in order to be promoted to the next round. The team that can win 6 straight game in March Madness is the champion. As predictive analytical driven students, as well as sports fan, our team will not miss the chance to predict each team’s performance based on data analytical skills. It is also a perfect opportunity to practice what we learned in class to real world.

NCAA March Madness is the single-game elimination tournament which contains the best 64 college basketball teams in United States. Our analysis is based on the dataset which includes every team’s regular season statistics for last 14 years. We will apply multivariate models and analysis on this dataset to predict their post season, i.e., March Madness, performances. It is extremely hard to predict the specific winning percentage of each teams because the standards every team hold is extremely different, and amateur athletes are placed in random environments usually against teams they've never played before in a "win or go home" style of play.

The first analysis we ran is Principal Component Analysis(PCA), it is an analysis that is able to help us determine which variables or set of variables are significant for team to win games in March Madness. As the results stated, a team that has better regular season record (wins/winning percentages) tends to perform better in post season than other teams. As an old saying, “offense wins games, defense wins championships”, which is also approved correctly in this project. According to PCA result, teams that have better defense records, especially in blocks, are more intend to win games in post season. Additionally, the result also shows that teams have more Free Throw Attempt (FTA) are also tend to perform better at post season.

Then we applied PCA results to Linear Discriminant Analysis (LDA), which is able to characterize or separates two or more classes or objects. In this case, wins and losses in the post season are defined as two classes. …

The prediction is hard to achieve not only because we are not capable to apply some techniques that we have not learned in this class, but we are also lack of datasets. The further we predict, the less example we have, which has critical influences in accuracy. Elite 8 teams statistics occupied 12% of the whole datasets and Final 4 teams occupied 6% of the whole datasets. In order to improve the accuracy of prediction, I suggest more data from past year should be collected and applied to analysis. Additionally, more categories should be expanded and collected to the dataset such as average height, average weight, etc.