

Introduction

The topic of our team final project is predicting the results of the 2015 NCAA tournament. We chose this topic because all the members of our team are interested in sports and two of them are very familiar with basketball. Everyone in Duke feels proud of our team's achievements, so do we. Therefore, when we found the data set of NCAA basketball that would allow us to work on this topic, we decided on it immediately. For our final project we want to apply what we have learned to our field of interest.

Since we know what happened in the 2015 tournament then we have a complete set of results with which to compare the prediction our model produces and hopefully we can also take a stab at predicting next year's edition as well.

The data set we will be using is provided by a past competition on Kaggle.com, which ensures that the data is both reliable and rich enough for doing our predictions. Specifically, there are more than 5000 pieces of game data for each year and 34 columns of many other kinds of information for each game including 26 technical skill indexes, which will be useful for us to compare the performance of each team. Except for the detailed data of regular season for 2015, there is also detailed data of both regular season and tournaments from 2003-2014, which can be helpful if we want to consider or compare the performance of past years. With this data set, we can use many methods that we have learned from the class, such as regression and LASSO. Therefore, it is an appropriate data set for the final project.

Some motivating questions we had: Teams in the NCAA change rosters very quickly which means that historical data of past seasons/tournaments has diminishing returns. For example, how useful will be the results of a season that happened 12 years ago in predicting this year's results when all the teams players (possibly coaches) may be completely different. Therefore, we want to know how useful and to what extent should we use historical data in our analysis.

Another question is how useful are each individual game metric to our model? Are no. of overtimes in a game really significant for our predictions? One way we might want to tune our model is by using backwards-step methods to identify the best model to use.