The Data

March Machine Learning Mania 2017 from kaggle

<https://www.kaggle.com/c/march-machine-learning-mania-2017>

Datasets I used (removed all 2017 data):

* Tourney seeds (from 2003 to 2016)
* Tourney Detailed Results
  + I removed the results from the 2016 tournament except the matchups for the first round.
  + I removed all data pre-2003 (from 2003 to 2016), since there were no regular season matchups before 2003 in the other dataset.
* Regular Season Detailed Results (from 2003 to 2016)
* Teams (id number for each team next to their actual name)

Getting the Predictors

For a particular matchup, the team on the left is the team that won that matchup. I implemented a Python script with the Numpy library that averages all the regular season statistics for a team for each year, kept track of how many games each team played and how many times they won those games, and calculated the average points they allowed for the season. I returned these results in a new csv file that has the stats for each team for each year.

Getting the Tournament Matchups

I looked at all the teams that have played in a tournament from 2003 to 2015 and those that are going to play in the 2016 tournament. I created a new csv file that has every tournament matchup and added the regular season averages next to each team for each matchup. Implemented another Python script to get the seeds from a csv file and added a seed attribute next to each team. The team on the left still reflected the team that won. I created a binary response variable that was 1 every time the higher seeded team (closer to one) won and 0 every time the lower seeded team won. If the response is 0, this means that the matchup could be an upset. If the teams had equal seeds, then I checked which team had the most wins. If the team with the most wins won, then the response would be changed to a 1 from a 0.

Getting and Modifying the Data in R

I imported this dataset into R and made a data frame from the data. I changed each team’s ID number to a string with their name. I then calculated new predictors for my model, such as points per possession (ppp), defensive points per possession (dppp), etc. Moreover, I did this for each team in a tournament matchup. I added each new predictor in my data frame next to each team in a matchup and arranged my data frame so that the higher seeded team would be on the left and the lower seeded team would be on the right. Additionally, I separated my training data (tournament matchups from 2003 to 2015) and my testing data (first round of the 2016 tournament with no response). I split my training data into rounds by the daynum attribute. Lastly, I adjusted each numeric predictor for my training and testing data by the league average for that year for that round.

Formulas

Rebounds per game = (OR + DR) / GAMES

Field goal percent = 100 \* FGM / FGA

Total possessions = GAMES \* (FGA - OR + TO – 0.44 \* FTA)

Total points = GAMES \* PPG

Points per possession = Total points / total possessions

Defensive points per possession = total points allowed / total possessions

Effective field goal percentage = 100 \* (FGM + 0.5 \* FGM3) / FGA

True shooting percentage = 100 \* (0.5 \* PPG) / (FGA + 0.44 \* FTA)