Predicting the Madness: Optimizing Your Bracket

We have compiled for you data (described in detail below) from the men's NCAA D1 March Madness basketball tournament from 2003 – 2010 and 2014-2017 (the other years will be used to see how well your model does). Using this data, your objective will be to become a master of ESPN bracketology – i.e. to train a model on this data that scores well on years outside this range, according to ESPN's scoring function:

- Round 2: 10 points per pick
- Round 3: 20 points per pick
- Round 4: 40 points per pick
- Round 5: 80 points per pick
- Round 6: 160 points per pick
- Championship: 320 points per pick

You may only train on the data provided. We have given you regular season and tournament results for 2003 – 2010 and 2014-2017, but only regular season results for 2011-2013. Your job will be to predict the tournament results for 2011, 2012, and 2013 using the model you learned from the other years.[1]

In particular, you would like to learn a function f(x,y) that predicts the winner, where each team is a vector of its statistical features (you select/create which features are important).

You will use this function to produce an output file of predictions for the tournaments in 2011, 2012, and 2013. The structure of that output file is shown in SamplePrediction.csv. The game ID is given by <year>_<team one ID>_<team two ID> and the prediction is 1 or 0, depending on whether you think team one will win over team two. (You can list team one and team two in any order). You need to generate a predict for *every possible matchup of all the teams listed in the tournament, should they meet.* Thus, if there are 64 teams in the tournament (or a few more if there are play-in games), you should produce an output file with 64-choose-2 predictions (that's 2016 predictions).

[1] Note that RPI (rating percentage index) and SOS (strength of schedule) are only available for 2011-2016. So if you decide to consider them as features, know that although they will be available in your testing set, they will only be present for 2014, 2015, and 2016 in your training set.

Your results need to be reproducible and your code readable and well organized. If your code is not clear and your thoughts well organized, we will have no way of knowing what you did.

You may use any tools or algorithms that you like. If you use an algorithm or approach from outside the class, but sure to explain it. Feel free to take this in any direction.

Have fun!

DATA

teams.csv

This file specified the 4-digit ID assigned to each team.

RegularSeasonDetailedResults.csv

This file contains the regular season results from 2003 – 2013.

wfgm - field goals made
wfga - field goals attempted
wfgm3 - three pointers made
wfga3 - three pointers attempted
wftm - free throws made
wfta - free throws attempted
wor - offensive rebounds
wdr - defensive rebounds
wast - assists
wto - turnovers
wstl - steals
wblk - blocks
wpf - personal fouls

TourneyDetailedResults.csv

This file contains the tournament results from 2003-2013.

TourneySeeds.csv

This file identifies the seeds for all teams in each NCAA tournament, from 2003 - 2013.
"season" - the year
"seed" - this is a 3/4-character identifier of the seed, where the first character is either W, X, Y, or Z (identifying the region the team was in) and the next two digits (either 01, 02, ..., 15, or 16) tells you the seed within the region. For play-in teams, there is a fourth character (a or b) to further distinguish the seeds, since teams that face each other in the play-in games will have the same first three characters. For example, the first record in the file is seed W01, which means we are looking at the #1 seed in the W region (which we can see from the "seasons.csv" file was the East region). This seed is also referenced in the "tourney_slots.csv" file that tells us which bracket slots face which other bracket slots in which rounds.

"team" - this identifies the id number of the team, as specified in the teams.csv file

TourneySlots.csv

This file identifies the mechanism by which teams are paired against each other, depending upon their seeds. Because of the existence of play-in games for particular seed numbers, the pairings have small differences from year to year. If there were N teams in the tournament during a particular year, there were N-1 teams eliminated (leaving one champion) and therefore N-1 games played, as well as N-1 slots in the tournament bracket, and thus there will be N-1 records in this file for that season.

"season" - the year

"slot" - this uniquely identifies one of the tournament games. For play-in games, it is a three-character string identifying the seed fulfilled by the winning team, such as W16 or Z13. For regular tournament games, it is a four-character string, where the first two characters tell you which round the game is (R1, R2, R3, R4, R5, or R6) and the second two characters tell you the expected seed of the favored team. Thus the first row is R1W1, identifying the Round 1 game played in the W bracket, where the favored team is the 1 seed. As a further example, the R2W1 slot indicates the Round 2 game that would have the 1 seed from the W bracket, assuming that all favored teams have won up to that point. The slot names are different for the final two rounds, where R5WX identifies the national semifinal game between the winners of regions W and X, and R5YZ identifies the national semifinal game between the winners of regions Y and Z, and R6CH identifies the championship game. The "slot" value is used in other columns in order to represent the advancement and pairings of winners of previous games.

"strongseed" - this indicates the expected stronger-seeded team that plays in this game. For Round 1 games, a team seed is identified in this column (as listed in the "seed" column in the tourney_seeds.csv file), whereas for subsequent games, a slot is identified in this column. In the first record of this file (slot R1W1), we see that seed W01 is the "strongseed", which during the 1985 tournament would have been Georgetown. Whereas for games from Round 2 or later, rather than a team seed, we will see a "slot" referenced in this column. So in the 33rd record of this file (slot R2W1), it tells us that the winners of slots R1W1 and R1W8 will face each other in Round 2. Of course, in the last few games of the tournament - the national semifinals and finals - it's not really meaningful to talk about a "strong seed" or "weak seed", but those games are represented in the same format for the sake of uniformity.

"weakseed" - this indicates the expected weaker-seeded team that plays in this game, assuming all favored teams have won so far. For Round 1 games, a team seed is identified in this column (as listed in the "seed" column in the tourney_seeds.csv file), whereas for subsequent games, a slot is identified in this column.