March Madness Predictions

James D. Wilson

March 2nd, 2017

It is now March, and this can only mean one thing: it's time for some college basketball madness!! Starting on March 16, the best 68 men's college basketball teams will play in a single-ellimination tournament, culminating in one last team as victor - the 2017 NCAA men's college baskeball champion! To keep up with what's going on (I will be), go here:

http://www.ncaa.com/march-madness

Overview

Each year, fans do their best to predict who will win each matchup in the 68 team tournament. The 68 teams are ranked according to their performance throughout the year and placed in a matchup slot in the tournament on *Selection Sunday*, March 12th this year. As an aspiring data scientst, your goal is to predict who will win each team matchup in the tournament.

We don't have the new schedule for this year, and alas it will be released over Spring break. Instead we will use past tournaments and a collection of historical team and player variables to analyze and build a model that predicts each matchup in the 2016 tournament.

A Little Motivation

Before we dive too deeply into the data, check out the Wikipedia description of the NCAA men's basketball tournament:

https://en.wikipedia.org/wiki/NCAA_Division_I_Men%27s_Basketball_Tournament

And, let's watch the last few seconds of the championship game between UNC Chapel Hill (go HEELS!) and the Villanova Wildcats:

https://www.youtube.com/watch?v=EMHoGRp1QrE

. . . . . yes I did watch this game out with friends from UNC last year

Data Description

The *March_Madness.RData* dataset on GitHub contains several data frames of potentially valuable information about the wins, tournament status, regular season statistics, etc. about the Men's basketball teams since as early as 1985. This data was provided in the Kaggle data challenge here:

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

In particular, the dataset contains the following data frames:

- **RegularSeasonCompactResults**
- **sample_submission**
- **Seasons**
- **Teams**
- **TourneyCompactResults**
- **TourneySeeds**
- **TourneySlots**

You'll have to read the above link to get a full description of each data frame.

Note: To shrink the data a little for this case study, I kept only the **RegularSeasonCompactResults** from 2005 onward (so I got rid of results from 1985 to 2004).

Your Objective

Major Goal: Predict which team will win each tournament matchup in 2013, 2014, 2015, and 2016. The matchups that you need to predict are given in the dataset **sample_submission**. Importantly here, we know the truth of who wins! This is provided in the **TourneyCompactResults** data frame in order according to the 67 or more games that are played.

Use exploratory data analysis (summaries, plots, etc.) to determine how you might go about predicting the results of the Tournament games. Consider the following questions to help you along the way –

1) What features of each team are strongly related to regular season wins? Discuss how these are related.
2) What features of each team are strongly related to Tournament wins? Are these the same features that help determine regular season wins?
3) How related are regular season wins and Tournament wins?
4) Combining historical data for each team, how would you go about making a model for determining whether or not a team will win its matchup in the NCAA tournament?

Hints To Get Started

Notice that the data isn't in it's "easiest to use" form. You will need to use subsetting techniques, coercion, and re-labeling to organize data into nice, easy to use data frames. As an example, the Tournament wins are provided in a different format than the regular season wins and losses. You'll have to be careful about how you get these in a useable form. Remember that google.com is your friend!