Name: Vidhi Katkoria (rwa4se) & Gabriel Simmons (gjs3qd)
Date: November 12, 2021

SLGM Project Proposal – Prediction of NCAA Men's Basketball Tournament Teams

Goals:

The goal of this project is to use data available only in the preseason to determine which teams will make the NCAA Tournament. The NCAAM basketball tournament (from now on we shorten to "the tournament") is currently a 68-team basketball tournament comprised of 32 teams that won their respective conferences and 36 "at-large" teams. A conference can be thought of as a "sub-league" of teams. Each conference is made up of a collection of between 8 (Ivy League and WAC) and 15 teams (ACC) in close geographic proximity. The winner of each of these conferences gets an automatic bid to the tournament. The 36 "at-large" teams are comprised of the "best" teams that did not win their conference, as judged by a selection committee at the conclusion of the season. After the 68 teams are selected, they are each assigned a seed by the same committee corresponding to their perceived strength. Lower-numbered seeds are given to better teams. The tournament itself is a simple single-elimination tournament with no reseeding. After an initial play in round that trims the field from 68 teams to 64 teams, six rounds are played the eventually ends with the crowning of a champion.

As previously mentioned, our goal is to determine who will make the tournament based on data available before the season begins. Some examples of this include what percentage of minutes from the previous year is that team returning, the strength of incoming transfer players, that year's recruiting class, performance from the previous season(s), etc. These are just a few examples of data at which we can look.

Finally, it should be noted that our system will *not* be predicting the "best" 68 teams. It will be predicting who makes the tournament, and nothing more. This distinction is important because of various imbalances within NCAA basketball. Front and center of these issues is the power imbalance among conferences, in particular. For example, take the ACC – UVA's conference – and the Colonial Athletic Association (CAA), which contains Virginia schools William & Mary and JMU. The ACC is made up of some of the largest and most well-funded schools in the country, which has led to some of the most prestigious basketball institutions. The ACC has posted the most NCAA tournament wins since 2000 (198), has had the most champions (8), and has submitted 98 "at-large" teams into the tournament. On the other hand, the CAA has submitted just 3 "at-large" bids in that same time frame, and of the teams currently in the conference, they have a combined 2 tournament wins. Remember that every conference winner *automatically* gets into the tournament. Therefore, even though a mediocre ACC team who did not get into the tournament would be likely to beat the CAA conference champion, the CAA will still get to send a team. We hope to create a model that accounts for these and similar disparities.

Data Set(s):

As mentioned in the previous section, we will use various data sets corresponding to different components of a team's makeup going into some season. This will include team stats from the previous season (win-loss record, conference win-loss record, various overall efficiency ratings) as well as returning player and transfer player data. In our initial searching, we found a few interesting data sets on kaggle.com. This one compiled a list of all college basketball players who played at least one minute from 2009-2021. This one has team data since 2013. However, it is quite possible that we will need additional data beyond this. However, this should not be too hard to garner. Barttorvik and Kenpom are both outstanding resources that can be easily scraped using a tool like selenium to grab whatever features we want.

Proposed Method:

This problem screams *classification*; it would have two very distinct classes – those who make the tournament and those who do not. Therefore, a Gaussian class-conditional generative probabilistic model (with some elements of one-hot encoding for categorical data) seems to make sense. However, upon further inspection, an unsupervised learning approach might make even more sense in this case since we are not 100% of the Gaussian nature of some of the data. This led us to the only such method we have learned in this class – EM. We will explore these options and determine which will work best.