Question

How accurately can I predict the 2016 NCAA tournament using only regular season statistics and past tournament results?

Method

I am going to create a logistic regression model for each round of the tournament. The purpose of each model is for prediction and not interpretation. They will predict the probability of the higher seeded team winning a matchup. The training data will be comprised of tournament matchups for each round from 2003 to 2015. For each round, I will only look at matchups from that particular round in order to train my model and have independence. The testing data will be the current round in the 2016 tournament. I will have to build a new matchup schedule for every round.

Limitations

I only have regular season statistics for predictors. My model does not know if a team is in the AP top 25 poll, how much experience the coach has, or how each team did in their conference tournaments. I would assume that these are important predictors that would let me predict the tournament more accurately. The training data for later rounds in the tournament will not have many observations. Therefore, I will not be able to put many predictors in my model with the fear of over fitting. I could also run across perfect separation errors if I put too many variables into my model.