NBN Methodology

John Kearns

11/16/2020

No Bid Nation Model Methodology

One of the most disappointing aspects of following a mid-major team, especially teams in conferences like the CAA, is the lack of in-depth statistical coverage and analysis. There has been a stats explosion for the top teams in the NCAA; player-tracking technology and play-by-play data is changing the way athletics departments and fans are engaging with players. Meanwhile, the open-source score-keeping log at William & Mary's own Kaplan Arena failed to work for half of the Tribe's home games. The CAA may get mentioned in a few tweets by the more famous college basketball statisticians, with most coming only at the end of the season as the conference tournament begins.

Despite the lack of focus at the national level on our Tribe, there is no reason that we cannot give the CAA the level of analysis it deserves ourselves. Some time ago, I came across the excellent work by Luke Benz, a recent Yale graduate who was president of his university's sports analytics club. In his position, he developed his "NCAA Hoops" model (https://rpubs.com/lbenz730/ncaa_hoops_methodology), which ranks the offensive and defensive strength of every team in Division I basketball. His model has performed surprisingly well over the past few years, and he has very kindly released his code and datasets. His work in tailoring a model to focus on the Ivy League inspired me to do the same for the CAA.

For our No Bid Nation model, I sought to build upon Benz's code to create a model capable of both providing statistical rigor not yet consistently produced for the CAA and more generally continue Benz's commitment to providing open source rankings for all of college basketball. Without his initial work on his model, it would have been impossible for me to build it out for the CAA. With that being said, I want to express my respect and thanks to him.

In this methodological overview, I will review the basics of the model, visualize its predictive power, explain its intuition, and outline the CAA-specific features I have built out.

Methodological Basics

Sources of data

I have scraped all game and schedule data from ESPN. Recruiting data is scraped from 247Sports. Much of the historical game and prediction data has also been graciously provided by Luke Benz. Player performance and roster turnover stats come from Bart Torvik.

Score differential

The core of this model is rooted in a regression predicting the score differential of each game. The model is designed as follows:

$$Y_i = \beta_{team} X_{team,i} - \beta_{opp} X_{opp,i} + \beta_{loc} X_{loc,i} + u_i$$

where Y_i is the score differential for the ith game, $X_{team,i}$,$X_{opp,i}$, and $X_{loc,i}$ are indicators for the team, opponent, and location (home, neutral, or away). The key identifying assumption is that game outcomes are independent of each other (i.e., Team A losing to Team B does not change the likelihood they will lose to Team C) and that the error term is distributed normally with mean zero and constant variance. Intuitively, it seems likely that these assumptions hold because each game is generally played under new and changing circumstances.

β_{team}, defined as the "YUSAG coefficient", represents the points better or worse than the average college basketball squad on a neutral court. β_{opp} is interpreted the same and is fit so that $\beta_{team} = \beta_{opp}$ for each team in the dataset. β_{loc} represents home-court advantage, which is estimated to be roughly 3.3 points. This result is in line with other research.

A game's predicted score differential can therefore be predicted as the sum of coefficients corresponding to the specific teams and the location. At this point, it is important to note that the predicted differential will change based on who is defined as the home and away team. For example, imagine William & Mary is hosting Delaware at Kaplan Arena. Assuming the Tribe have a YUSAG coefficient of 1 and Delaware has a YUSAG coefficient of -2.3, meaning that William & Mary is expected to score one more point than the average DI team on a neutral court while Delaware would score 2.3 points less, the score differential from the Tribe's perspective is 1 - (-2.3) + 3.3 = 6.6. The predicted score differential from Delaware's perspective would be -2.3 - 1 - 3.3 = -6.6. In any case, William & Mary would still be 6.6 point favorites.

Offense and Defense

The model can also look at team performance at a more granular level by looking at points score and points allowed:

$$T_i = \beta_0 + \alpha_{team,i} X_{team} - \delta_{opp,i} X_{opp} + \beta_{loc,off,i} X_{loc} + u_i$$

$$O_i = \beta_0 - \delta_{team,i} X_{team} + \alpha_{opp,i} X_{opp} + \beta_{loc,def,i} X_{loc} + u_i$$

β_0 represents the average points scored/allowed across DI basketball, while $\beta_{loc,off/def,i}$ represents the offensive and defensive contributions to home court advantage.

α_{team} consequently represents the number of points greater than baseline a certain team would score on a neutral floor against an average team. δ_{team}, similarly, is the number of points fewer than baseline a certain team would allow on a neutral floor against an average team. The predicted points scored and allowed for a certain game would be calculated in much the same way as score differential.

However, it is important to note that these coefficients are not adjusted for tempo and should not be used to compare the offensive and defensive prowess of teams. For instance, in the year that Virginia won the NCAA Championship, the Cavaliers had a negative offensive rating and a large, positive defensive rating. Thus, despite having a top-25 tempo-adjusted offense, Virginia would score fewer points than average because their defensive solidity meant there were fewer possessions in the game overall to score on. Comparing offensive and defensive ratings for the same team can show what phase of the game, if any, dominates team play. Only the YUSAG coefficient should be used to make claims such as "Team A is better than Team B."

Preseason priors

One difficulty with YUSAG coefficients being solely based on in-season performance is that early predictions may be inaccurate. This can be solved by introducing preseason estimates of team strength. These estimates, based on variables including number of returning players, recruiting, and previous performance, will not be flawless. They will, however, be successful in augmenting early predictions. The stability in team strength coefficients (visualized in Figure 1) indicates that previous performance, and therefore estimated priors, contain useful information for the model.

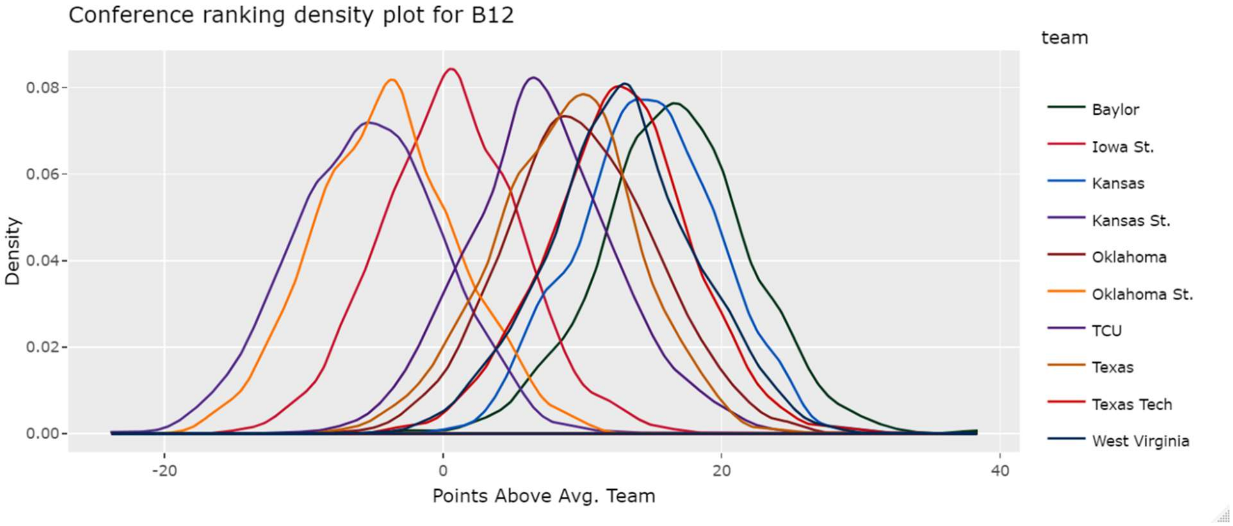
The priors are estimated by regressing historical performance from 2018-20 on a set of relevant variables using 10-fold cross validation (https://www.openml.org/a/estimation-procedures/7#:~:text=Cross%2Dvalidation%20is%20a%20technique,test%20set%20to%20evaluate%20it.), a technique that optimizes coefficients for predictive purposes.

Offensive and defensive priors are estimated using separate regressions. The offensive model is: $\alpha_{i,t+1}$ regressed on $\alpha_{i,t}$, 247 Sports Recruiting Composite Score (team i, season t), returning PPG! (team i, season t+1), transfer recruit PPG! (team i, season t+1), percentage of minutes played by returning players (team i, season t), and indicators for having at least one new 5-star recruit or more than two new 3+-star recruits (team i, season t+1). PPG! refers to Points Over Replacement Player Per Adjusted Game At That Usage (https://www.bigtengeeks.com/new-stat-porpagatu/), a wonderful statistic designed by Bart Torvik that captures the efficiency-adjusted production over replacement. This statistic gives a better sense of a player's contribution than looking at scoring figures alone.

Thus, priors are established as a combination of the previous year's strength, roster turnover, and recruitment. The R-squared of this model (the percentage of variation in next season's offensive performance captured by the model) is 51%. With only these few variables, this model is able to generally ballpark a team's strength (most likely better than the coefficients from the first few games alone would tell you).

The defensive model is similar: $\delta_{i,t+1}$ regressed on $\delta_{i,t}$, 247 Sports Recruiting Composite Score (team i, season t), returning PPG! (team i, season t+1), transfer recruit PPG! (team i, season t+1), percentage of minutes played by returning players (team i, season t), and indicators for having at least three new 5-star recruits. The R-squared for this model is 57%.

One advantage of using regression to predict offensive and defensive rankings is that you can use the resulting standard errors to establish a prior distribution for each team ranking. Since the YUSAG coefficient is just the sum of the offensive and defensive coefficients, the standard deviation for the YUSAG coefficient is $\sqrt{\sigma_{\alpha}^2 + \sigma_{\delta}^2 + 2cov(\alpha, \delta)}$. These standard errors can be used to create fun plots showing the potential rankings for conferences (like this one for the Big 12).



As the season begins, team rankings are established as a weighted sum of the preseason priors and the regression results from the models explained in the first section. The contribution of each new game to this weighted sum depends on the relative position of the game on each team's schedule; more recent games are weighted more heavily than earlier games. This model is therefore attuned to the form of teams, discounting a slow or hot start. Similarly, as a team plays more and more games, its preseason estimate is weighted less and less. Once a team has played 1/2 of its schedule, preseason estimates drop out and the rankings are purely based on the aforementioned regressions.

It is important to reiterate that the rankings at the beginning of the season are likely to be biased in certain, and perhaps unexpected, directions due to quirks in the data. As the season progresses and I accumulate more data, the rankings and projections will improve mightily.

Win Probability

The last major part of the model centers on predicting win probabilities for games that have not played yet. This is crucial for forecasting conference winners, conference tournaments, and NCAA Tourney outcomes.

Using historical data on predicted score differentials using YUSAG coefficients and win outcomes, one can fit a regression model to predict win probabilities. The model simply regresses a binary indicator for winning on predicted score differential. The coefficients of this logistic regression allow you to predict the chance of a team has of winning any game on their schedule. The figure below shows how accurate the model is at predicting outcomes; the closer the dots are to the diagonal line, the more accurate the model is.

So what can this model do?

- Calculate robust strength ratings for every team in NCAA Division I basketball
- Develop pre-season rankings based on prior distributions
- Maintain win prediction model and estimate predicted win totals
- Provide power rankings by Top 25 and by conference
- Simulate regular season conference winners and conference tournaments
- Simulate the NCAA Tournament
- Rate games by tournament probability swing factor
- Varied visualizations for rankings, games, and playoffs

Acknowledgements

I have to thank Luke Benz again for publishing the work he did during his time in Yale. It has served as such a helpful guide, and my management of these files would be impossible without his initial work.

I hope to upload all of my code and data to GitHub soon so people can evaluate the model for themselves.