
ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: “Big” Data Analytics
Proposal (or Final Report)
Basketball Team Statistics as Predictors for Postseason Success

Background

Every year at the end of college basketball season, the most impressive teams are selected to compete in a tournament to determine who will be the national champion. For many teams, even being selected to compete in this tournament is a huge honor and indicates a successful season. Meanwhile, many of the most traditionally powerful teams are expected to make a deep run in this tournament. One of the tournament’s biggest characteristics is its unpredictability: every year fans fill out brackets to try to get as close to correct as they can, though a “perfect bracket” is nearly impossible. Warren Buffett has promised \$1 billion to anyone who can create a perfect bracket, but the odds of getting a perfect bracket are 1 in 9,223,372,036,854,775,808 (9.2 quintillion). For my project, I wanted to analyze some factors and determine what helps a team go farther in the NCAA Tournament.

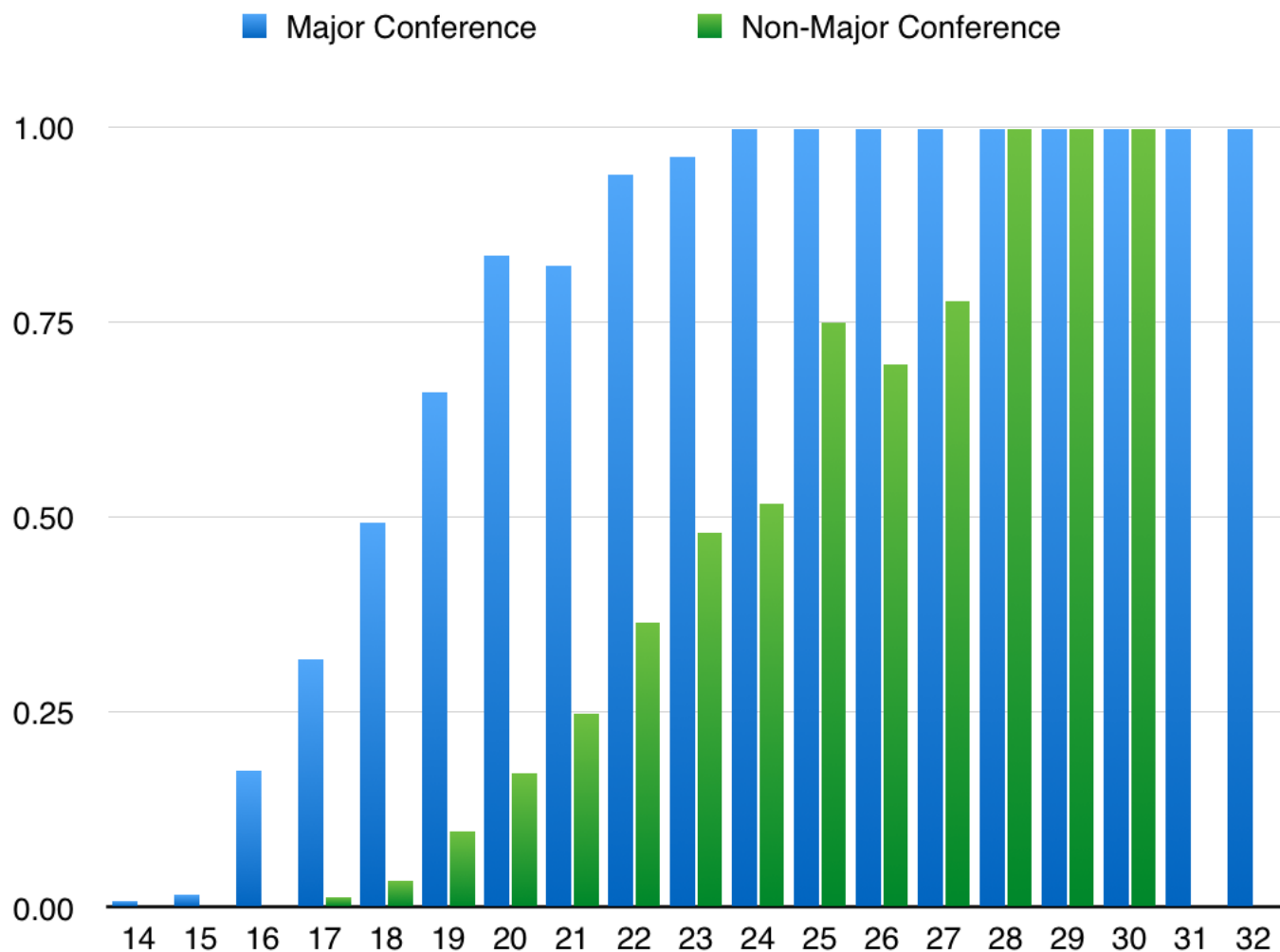
Methodology

First, I selected my factors to analyze. I chose:

- Offensive efficiency: This is a measure of how many points a team scores per 100 possessions. It makes intuitive sense that teams with a better offense will have more success in the tournament.
- Defensive efficiency: This is how many points a team allows per 100 possessions. It makes intuitive sense that a team with a better defense will have more success.
- Tempo: This measures how many possessions a team averages per game. Different teams have different strategies, so there isn’t really an intuitive solution as to whether high-tempo or low-tempo teams will be successful. I wanted to investigate and see if statistically one is better.
- Experience: This is the average years of college basketball experience for each player, weighted by how much they play. It’s “common knowledge” in basketball that more experienced teams play better, so I wanted to test that premise.
- Effective Height: This statistic measures how many inches taller or shorter a team is than the average, weighted towards the power forward and center position.

Next, to quantify “success in the NCAA Tournament,” I decided to represent that with an integer value. That value is 1 if a team simply makes it to the tournament, and 1 is added to it for each game they win (one exception: no points were added for play-in games. This is because a 16-seed defeating another 16-seed in a play-in game is not nearly as important as beating another team. I could’ve weighted it such that winning a play-in is worth 0.1 or so; but I decided that instead of trying to estimate how much a play-in game is worth, I will just not evaluate those). One last thing I

wanted to do was stratify the data. I did this because different teams enter the tournament with wildly differing strengths of schedules. What this means is that maybe it's possible to be a "gimmicky" team and make it to the NCAA Tournament as a lower seed because you beat weaker teams. For evidence of that, consider the difference in standards of play to make it to the tournament for teams from a Major Conference versus teams from a non-Major Conference. I collected data for how many games a team won, and whether they made the tournament, dating back to 1985 when the tournament expanded to 64 teams. This represents roughly 10,000 teams. This graph shows the proportion of teams that make the tournament with each amount of wins:



Clearly, teams from non-Major Conferences must win more games to make it to the tournament. It then follows that their other stats would be higher as well. Rather than trying to come up with a way myself to sort teams into tiers, I relied on what the NCAA has already done: the seeding system. I evaluated the R2 value for each category for teams from each seed category. The results are as follows (on the next page).

Data

	1s	2s	3s	4s	5s	6s	7s	8s	9s	10s	11s	12s	13s	14s	15s	16s	Total
Offense	0.141	0.0222	0.0102	0.0245	0.0343	0.0368	0.0784	0.0265	0.0195	0.0263	0.026	0.0967	0.0004	0.0056	0.0004	Und*	0.2617
Defense	0.0334	0.0267	0.0599	0.1334	0.0002	0.0019	0.064	0.0099	0.0724	0.0648	0.0023	0.0003	0.0485	0.0017	0.052	Und*	0.2046
Tempo	0.0091	0.0024	0.0164	0.046	0.0113	0.001	0.0224	0.0757	0.0336	0.0708	0.0063	0.0105	0.0132	0.0207	0.0248	Und*	0.0011
Experience	0.0017	0.004	0.0029	0.0213	0.033	0	0.0462	0.0181	0.0456	0.0257	0.0001	0.0047	0.0392	0.0906	0.0047	Und*	0.0235
Effective Height	0.1067	0.0249	0.0047	0.0056	0.0733	0	0.0001	0.0022	0.0261	0.058	0.0228	0.0295	0.0974	0.0012	0.002	Und*	0.0682

Red values indicate a negative slope of the linear regression, and green values indicate a positive one. None of these values are particularly high. Judging from this, it appeared that none of these factors had much of an effect. Each category contains around 30 entries. Next I eliminated the stratification by seed, and just evaluated all teams in each category. Here are the results from that:

	1	2	3	4	5	6	7	R ²
Off.	106.9	110.9	112.5	115.3	114.5	116.1	119.0	0.9164
Def.	98.3	95.0	93.7	92.9	91.3	91.9	90.5	0.8740
Tempo	66.5	66.1	66.2	66.9	65.6	65.1	67.0	0.0152
Exp.	1.82	1.75	1.76	1.61	1.65	1.51	1.63	0.6948
E.H.	0.6	1.0	1.6	2.1	1.9	2.0	3.2	0.8817

This shows a much higher correlation in general. The 1s category here has 512 values, and each category after that contains half the entries of the previous category.

Results and Summary

Clearly, the stratification plan did not work. There is a very strong correlation between teams with a good offense, defense, and height making deep runs in the tournament. Interestingly, a team's experience does not seem to have much of an affect at all; in fact, there is a negative correlation there. Experienced teams do not play as well. This is likely because the very best players go to the NBA before finishing their college careers, so having a lot of seniors is essentially indicative of having subpar players. Tempo appears to have basically no effect at all. One of the more interesting results of this experiment is that depending on how one stratifies the data, one can take the same data set and get vastly different results.