NCAA Championship Statistics

Can we determine March Madness contestants based on historical data?

Overview

The purpose of this analysis was to obtain greater understanding of NCAA Division 1 Statistics using PCA analysis and logistic regression.

Datasets Analyzed

The main dataset used was several data pulls of team stats going back 10 to 20 years, and included such stats as Assist/Turnover Ratio, Three-Point Attempts, and Rebounds, etc. These were all pulled from the stats page of the NCAA website and combined into one main csv file using Spark. The combined data then includes Team Name, Season of the Stats in format of 1011 for example, meaning years 2010 through 2011. Overall, about 3700 records combined to analyze.

Data Cleaning

	Team	Years	AST	TO	Ratio	TPG	Avg	3FG%	PTS	PPG	ORebs	DRebs	REB	RPG	FGM	FGA	FG%	FT	FTA	FT%
0	Michigan (Big Ten)	12-13	565	368	1.54	9.435897	NaN	38.5	2,932	75.2	0	0	1,366	0.00	1,093	2,260	48.4	450	642	70.1
1	UC Irvine (Big West)	12-13	580	461	1.26	12.459459	NaN	33.4	2,558	69.1	0	0	1,428	0.00	932	2,151	43.3	445	669	66.5
2	Iowa St. (Big 12)	12-13	560	460	1.22	13.142857	NaN	37.4	2,779	79.4	0	0	1,341	0.00	962	2,111	45.6	509	696	73.1
3	Santa Clara (WCC)	12-13	560	468	1.20	12.315789	NaN	36.1	2,827	74.4	0	0	1,315	0.00	970	2,195	44.2	584	803	72.7
4	Wichita St. (MVC)	12-13	530	492	1.08	12.615385	NaN	33.9	2,720	69.7	0	0	1,497	0.00	954	2,166	44.0	553	791	69.9
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3711	Siena (MAAC)	23-24	342	505	0.68	15.781250	18.8	28.0	1,934	60.4	375	757	1,132	35.38	702	1,758	39.9	361	541	66.7
3712	VMI (SoCon)	23-24	335	507	0.66	15.843750	21.0	31.2	2,204	68.9	299	871	1,170	36.56	818	1,935	42.3	358	517	69.2
3713	Alabama A&M (SWAC)	23-24	369	565	0.65	16.142857	14.7	28.7	2,414	69.0	394	854	1,248	35.66	806	1,918	42.0	654	906	72.2
3714	Mississippi Val. (SWAC)	23-24	271	495	0.55	15.967742	15.8	29.2	1,719	55.5	298	613	911	29.39	613	1,611	38.1	350	532	65.8
3715	Coppin St. (MEAC)	23-24	240	455	0.53	15.689655	18.6	28.3	1,678	57.9	283	590	873	30.10	585	1,571	37.2	355	489	72.6

Problems in training

At first, the goal was to train the model based on individual matchups in an effort to predict a teams performance throughout the NCAA March Madness tournament.

We experienced some issues here, as our data did not include individual matchups so we had to pivot.

We landed on predicting contestants who make it to the tournament, based on season stats.

Contestants who made it into the tournament that year would be given a "Madness" label of 1, and those who did not make it to the tournament would be given a 0.

Resolving Imbalanced Data

Another obstacle we ran into when determining March Madness contestants, is that of 351 Division I NCAA basketball teams, only 68 are considered for the tournament. This presented us with a very imbalanced data set, and in order to fix this imbalance we reduced the number of teams' stats we considered by about half. We kept the top 150 teams by points scored for each season, and then joined the 68 Madness Contestants with the remaining 82 Division I teams.

PCA

Using our combined dataset and cleaning came out with 3 Principal Components

Explained Variance Ratio: [0.6971, 0.1717, 0.0657] suggests that:

The first principal component explains approximately 69.72% of the variance in the data.

The second principal component explains approximately 17.17% of the variance.

The third principal component explains approximately 6.57% of the variance.

Conclusion

- Had we understood the nature of the problem we wanted to solve a little better at the beginning, we might have spent less time experimenting with different models
- With a more complete data set that included individual season matchups, or more time to put that data together manually, we could have predicted individual matchups