NCAA March Madness: Exploring the Mathematics Behind Cinderella Teams

Andrew Huang, Vinai Rachakonda
University of Pennsylvania

Abstract

In March of every year, millions of basketball fans around the globe fill out predictions on the NCAA basketball tournament in hopes of being the first person in history to complete a perfect bracket, one where the outcomes of all 67 games are guessed correctly. The best bracket in 2017 made it to 39 games before making its first mistake. The biggest reason that this feat is practically impossible is due to the nature of Cinderella teams - low-ranked teams that unexpectedly win far more games than projected. This paper seeks to explore the correlation between low-ranked basketball teams' regular season performance and their March Madness outcomes. Through mathematical modeling and machine learning, this paper hopes to discover the mathematics behind Cinderella teams - if such a thing exists.

Introduction

Task Definition: Our team aims to use machine learning to predict the performance of underdog teams in NCAA March Madness. In specific, we want to use regression and classification algorithms to predict the number of games a team will win if they are seeded between #9-16 regionally.

Input: We took data from Sports Reference, which included basic stats (i.e. total points scored, total wins, etc.) along with advanced stats (i.e. TS%, TRB%, ORtg, etc.). In addition, we manually added each team's tournament seed and total tournament games won.

Algorithms: We ran three main types of algorithms – regression, classification, and ordinal regression. The exact algorithms we used were: Linear, Ridge, and Lasso Regression; Linear SVM, Linear SGD, Naïve Bayes; and OrdinalRidge and LAD.

Expectations: We hope to use machine learning and feature analysis to create a bracket that outperforms those made by professional sports analysts for the 2018 NCAA tournament.

Figure 1. No. 16 UMBC upset No. 1 UVA in 2018

Figure 2. No. 11 Loyola-Chicago won 4 games in 2018

Figure 3. Ryan Betley, Penn's leading scorer

Methodology

Results: Using the inputs from Sports Reference and the tournament seeds, we use the algorithms to predict the number of games a team will win. For regression, we record R^2 and RMSE, while we report accuracy scores for classification.

Analysis: To compare both types of algorithms, we developed a Mock Bracket Score (MBS), based on the point totals from the ESPN Bracket Challenge. This score rewards correctly predicted games, but penalizes under or over predictions.

Feature Analysis: We will continue to experiment with feature analysis, including how to treat missing data, or how to reduce dimensionality of our data.

1. Pre-process data and input into algorithms
2. Use predictions to calculate accuracy scores
3. Calculate Mock Bracket Score for each algorithm

Table 1. MBS Score Function (true games won vs. predicted games won)

True games won \ Predicted games won	0	1	2	3	4
0	10	-10	-30	-70	-150
1	-10	30	-20	-60	-140
2	-30	-20	70	-40	-120
3	-70	-60	-40	150	-80
4	-150	-140	-120	-80	310

Results

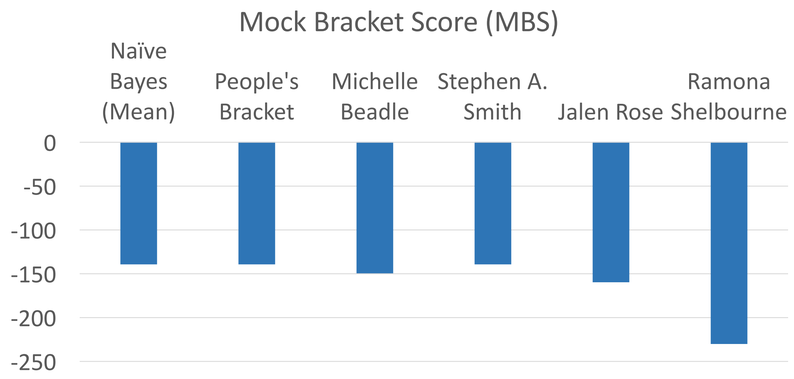
We calculated various accuracy scores of the ML algorithms, listed below in Tables 2 and 3. However, to compare the best algorithms with each other, and to compare them with expert analysts' brackets, we developed the MBS to emulate true bracket scoring.

Table 2. Top 5 Classification Algorithms, by Mock Bracket Score

Classification	Accuracy	Bracket Score
Naïve Bayes, mean	0.71875	-140
Naïve Bayes, drop	0.625	-190
Linear SVM, mean	0.15625	-360
Linear SGD, mean	0.15625	-360
Linear SVM, drop	0.15625	-360

Table 3. Top 5 Regression Algorithms, by Mock Bracket Score

Classification	RMSE	Bracket Score
Linear, mean	40227.70	-140
Lasso, drop	2.40	-1360
OrdinalRidge, drop	3.62	-4120
OrdinalRidge, mean	3.62	-4120
LAD, drop	3.62	-4120



Discussion

Regression: All of the regression algorithms performed extremely poorly. Regardless of algorithm, method of dealing with missing data, or input features, almost every algorithm would either have an extremely high RMSE or have poor bracket scores. The algorithms would either predict all 0's or all 4's.

Classification: These algorithms performed slightly better, but Naïve Bayes was the clear-cut winner. Note that low accuracy scores does not mean a poor bracket score.

We selected Naïve Bayes as our optimal algorithm, and used the mean to fill in missing data points. Our resulting MBS on the 2018 NCAA tournament outperformed the people's bracket (majority picks) and the 4 ESPN analysts. Below, in Table 4, we have the Naïve Bayes MBS for the last 5 NCAA tournaments, where some scores were actually positive. Upon close inspection of the data, we saw that Naïve Bayes was very conservative in guesses.

Table 4. Naïve Bayes, drop, MBS for past 5 years

	2013	2014	2015	2016	2017
Naïve Bayes MBS	60	200	-100	140	-140

Conclusions

We were able to achieve our hypothesis of outperforming experts in the most recent NCAA tournament by using classification algorithms, particularly Naïve Bayes. Closer inspection yielded the conclusion that a more conservative approach (ignoring underdogs) was the most consistent method of success.