Using Statistical Methods to Predict the Outcome of Basketball Games

Sabrina Ball
Joseph Despres

Quick Summary

1. Data

Our data comes from three different data sources Kaggle, NCAA, and a March Madness bracket. First, a Kaggle[1] machine learning challenge that has a comprehensive records of every US division I basketball game played from 2003 to 2020. Second, the NCAA[2] team statistics. Third, the tournament bracket in a nice .xls format we obtained from a blog[3].

Variables

Our response is the outcome win/loss as a function of team statistics. win

Variable name	Description	Type
win		
fg_percent		
opposingfg_percent		
ft_percent		
opposingft_percent		
rpg		
opposingrpg		
st		
opposingst		
to		
opposingto		
opposingbkpg		
bkpg		

[1] https://www.kaggle.com/c/ncaam-march-mania-2021/data
[2] http://stats.ncaa.org/rankings/change_sport_year_div
[3] https://plexkits.com/march-madness-bracket/

2. Questions

Can we use statistical methods to predict outcomes of the March Madness basketball games better than chance? Better than seed? Better than betting markets?

3. Learning

We would like to learn how to make predictions with logistic regression. We would like to come up with a good way to evaluate our predictions. This will be dificult since the outcome is binary so we would need to weight our predictions somehow and index our socres.

4. Methods

Logistic regression with a response prediction function.

5. Issues

The seeds are likely derrived from the predictors we are using. Multicolinearity, will be something to watchout for because better teams probally have better statistics.