Predicting the Outcome of Basketball Games with GLMs

Sabrina Ball and Joseph Despres

Every year top university basketball teams compete in the March Madness tournament. Casual fans and Enthuasists submit predictions gambling small sums of money on the outcomes. This study will determine if which GLM performs best in predicting the results.

Data

Our data comes from three different sources Kaggle, NCAA, and the tournament results. Kagggle provides a comprehensive dataset[1] for all NCAA in season (non-turniment) basektball games from 2001 to 2020. The NCAA provides team-level statistics[2] for each team. We will use those two datasets to train a model that uses team-level statistics to predict basketball game outcomes. We will test our model on the actual tournament results which we obtained with a scraping script[3].
The objective is to predict the individual game outcomes as accurately as possible. Our response is the outcome *win/loss* as a function of team statistics. Win or loss coded 0 or 1. Then we will look at both team's statistics and determine which is more important. In our dataset we have: *field goal percentage* which is the rate attempt to score vs actually scoring. *Free-throw percentage*, is the rate that the team has the opportunity to take an unguarded shot. Cumulative *Three-point goals made*. *Rebounds per game*, counts the times the team recovered the ball after a missed shot. *Steals*, the times a team was able to remove possession of the ball from the opposing team. *Turnover*, the number of times the team lost possession. *Blocks*, The number of times the team was able to block a shot made by the opposing team.

Learning Objectives

We want to learn how to make and validate predictions with GLMs. Additionally, we want an appropriate scoring metric that can compare the performance of various methods. We want to know which GLM performs the best for this particular problem.

Methods and Difficulties

We will use three regression methods, logistic, Poisson, and multinominal, to compare the results to determine which is most accurate. Logistic regression most naturally fits this problem because games do not tie so there is a 50/50 split of wins and losses. Also, we will try Poisson regression to predict the number of points scored by each team. Since there are many close basketball games we will use multinomial regression to predict if the game score is going to be a large difference one way, small difference, or a large difference the other way.

For the most part, assembling the data was not an issue yet I was worried about that. An issue we will face is scoring this model. We switched from making predictions depending on the previous round to being independent of the previous round. We come up with a simple scoring metric then, treat games played as a random sample from all possible games, and use basic statistical methods to infer the prediction accuracy.

[1] https://plexkits.com/march-madness-bracket/
[2] https://stats.ncaa.org/rankings/change_sport_year_div
[3] https://github.com/despresj/March-Madness/blob/main/R/scrape_finals.R