Using GLMs to Predict Basketball Games

Joe Despres & Sabrina Ball

Michigan State University

Using GLMs to Predict Basketball Games

Introduction

Every year top university basketball teams compete in the March Madness tournament. Casual fans and enthusiasts submit predictions gambling small sums of money on the outcomes. This study will determine if which GLM performs best in predicting the results. 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 after it is concluded on April 5th. 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.

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

[1] CITE THIS: https://plexkits.com/march-madness-bracket/

[2] CITE THIS: https://stats.ncaa.org/rankings/change_sport_year_div

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.

We Found. . . .

Exploratory Data Analysis

taking a look at our variables we see that they all have a reasionably normal distribution.

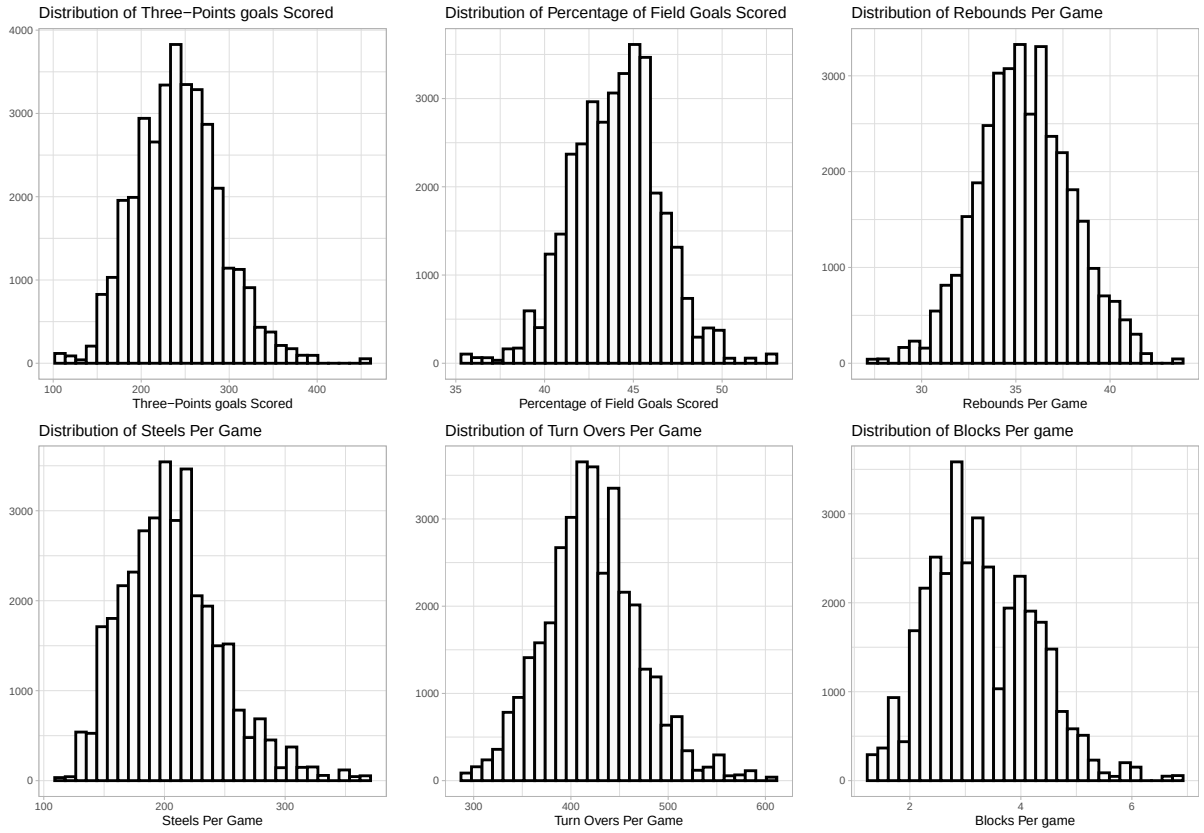

Table 1

Descriptive Statistics

Variable	Mean	Median	Std	Min	Max	Range
Three-Points goals Scored	242.08	241.00	48.18	107.0	454.00	347.00
Field Goals Scored Percentage	44.03	44.10	2.46	35.4	52.60	17.20
Rebounds Per Game	35.48	35.35	2.49	27.6	43.81	16.21

Steels Per Game	206.34	202.00	40.59	116.0	369.00	253.00
Turn Overs Per Game	421.84	419.00	49.01	290.0	604.00	314.00
Blocks Per game	3.30	3.20	0.94	1.3	6.80	5.50

Description

To gather these data, we used the tidyr, (Wickham, 2021) dplyr, (Wickham, François, Henry, & Müller, 2021) and purrr (Henry & Wickham, 2020) package to read-in clean, and prepare our data for analysis. Since this is a predictive model we used all the data we had at our disposal that was not colinear.

Results

Diagnostic Methods

Conculsion

References

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Wickham, H. (2021). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr