Ian Bogley
Jonas Bowman
Professor Rubin
Econ 524
10 March 2021

NBA All-Star Prediction

The aim of our project is to create a model that can predict the National Basketball Association (NBA) all-stars with a high degree of accuracy. To pursue this goal, we gathered data from various online sources and used created predictions using boosting, decision trees and other machine learning techniques

As stated above the goal of this project was to see if we could predict which players would be all-stars in any given NBA season. Understanding this question is important because many basketball historians and analysts believe that all-star voting is skewed because it is voted on by fans. Fan voting is believed to be less accurate as many fans may be biased towards players who have a more exciting style of play or who play for bigger market teams. Our goal is to test, this theory and determine if there are players in the data whose play merited a spot on the team but were ultimately snubbed. This is a prediction problem because we are not asking what causes someone to be an all-star, but rather who are the all-stars.

The data we gathered came from a Kaggle notebook called NBA player stats which we combined with data scraped from the web listing the all-stars and MVPs of each NBA season going back to the league's inception. In order to clean the data, we had to remove a few variables which were causing errors in our predictions and impute the mean and mode for numeric and nominal variables respectively. During the course of our work, we ran into challenges with the time scale of the data. The NBA has changed so much over the course of its existence that we decided to leave out every year before 2000 in order to get more accurate predictions. One major shortcoming of this data is the lack of all-star observations. The pool of all-star players is much smaller than the pool of non-all-star players.

The models used include elastic net logistic regression, linear regression, decision tree, and boosting frameworks. The optimal parameters for the first method were a mixture of 1 (lasso model) and a penalty of 1*10^(-10).The linear regression model also used the same mixture and penalty. The decision tree split variables include steals, free throw percentage, minutes played, and field goal percentage. The best boosting model had 8 splits, a shrinkage parameter of 0.1, with 15 trees. For each of the models, we used accuracy as our metric for tuning. The only exception is the regular regression, which we tuned using root mean squared errors.

There are several ways to measure the success of our model. However, the focus of our metrics was on predicting true positives, as predicting true negatives is relatively easy given that most players in NBA history are not all-stars. Therefore, if true positives are weighted higher, the two options for determining model value are sensitivity and precision. For both sensitivity and precision, the logistic regression model scored the best (sensitivity: linreg= 66%, logistic = 72%, tree=59%,boost = 61%) (Precision: linreg=72%, logistic =81% ,tree = 73%, boost = 79%). One major aspect that limited our performance was the low variability of our outcome. Having so few positive outcome variables made it difficult to predict all-star, and even worse for MVPs. This was a big point of growth in our understanding, as the possible analysis that can be done on an outcome variable depends in large part on the variability of the outcome.