MLB Draft Strategy

Jackson Kennedy and Megan Lieb

Due to the MLB Draft's unique slot value system and exceedingly diverse player pool, it seems reasonable to assume that draft strategy may vary significantly across the league. This assumption is even more reasonable when we consider organizations' drastically different approaches to free agency, which are largely determined by the amount of money their owners are willing to spend, market size, and/or location desirability. The aim of this project was to first identify these different draft strategies, and then to determine whether or not some strategies are more conducive to team success than others. We began with *K*-means clustering before transitioning into a few different forms of regression.

For our cluster analysis, we decided to focus solely on how teams draft in the *first round*. (This is mostly due to the fact that the only data we could find consisted of either teams' first-round picks or *all* of their picks; since the MLB Draft is 20 rounds and the majority of players picked late don't ever make it to the majors, we thought it made the most sense to focus on teams' tendencies at the top of the draft. If given more time, we'd like to extend this to the first three, or maybe even five, rounds.) Our analysis revealed four well-defined clusters, each of which are characterized as having slightly different preferences in regard to the types of player they like to target in the first round. Cluster 1, for example, invests notably less in pitching prospects than Cluster 2; but when teams in Cluster 1 *do* draft pitchers, they're more likely to be straight out of high school. Additionally, we found no meaningful relationships between the clusters and league (AL vs. NL), division, or market size; this opposes both our expectations going in and some of the literature, which describes teams in larger markets as "the most extreme in their pursuit of high school players."

In addition to identifying MLB draft strategies, we were interested in determining which draft statistics best predict the win-loss percentage of MLB teams. We analyzed MLB draft data to identify the strongest predictors of win-loss percentage. Using nonparametric bootstrapping, we found the following variables to be the strongest predictors of win-loss percentage: HS%, Pitcher%, and Signed%. The combination of these predictors yielded extremely low MSE and bias values, suggesting the model was closely predicting the win-loss percentage for all 30 MLB teams. However, when this model was applied to draft data specific to the Arizona Diamondbacks from 1999-2023, it performed poorly. To improve the predictive performance, we ran a random forest model for the win-loss percentage of the Arizona Diamondbacks. The random forest model did a significantly better job at predicting the win-loss percentage of the Arizona Diamondbacks for these particular draft statistics. This suggests that there might be more complex relationships between variables for the Arizona Diamondback draft data that the bootstrapping model was unable to accurately describe.