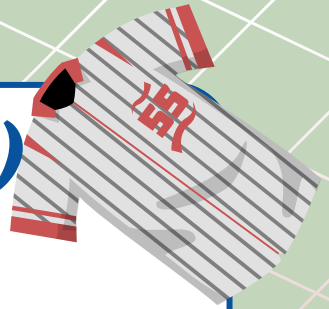
Which Baseball Players Hit it out of Their Ballpark? (Estimate)

Clay Winder and Tommy Skodje

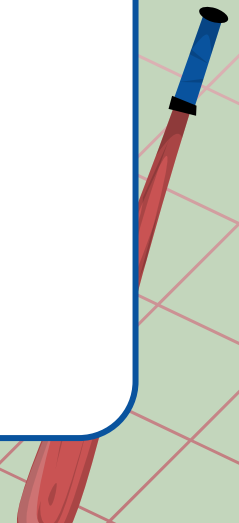
Motivation

- **The big question: How do we identify undervalued major league baseball hitters for the upcoming 2024 season?**
- As a general manager (GM) of a baseball team, there is the timeless question of “how do I get the most value for my money?”
    - The answer to this question starts with defining what “value” means as a hitter.
        - We chose weighted on-base average (wOBA), which is the calculation based on the value of each kind of hit (for example, a home run is more valuable than a single).
    - We can use wOBA as an approximation of “value” as a hitter. Valuable hitters will almost always have a very high wOBA.
- Major League Baseball uses Statcast, a system of cameras installed at every stadium, to calculate special statistics about the players and baseballs themselves, such as exit velocity, launch angle, and player sprint speed.
    - The league use these Statcast statistics to predict a player’s future wOBA (called xwOBA, or expected wOBA). These expected statistics are an important tool general managers use to evaluate a player’s potential performance for the next season.
- We will be trying to outdo xwOBA by coming up with our own prediction of what a player’s wOBA will be for the 2024 season.
- **By comparing our predicted wOBA to Statcast’s xwOBA, we can find hitters who are undervalued!**

Methodology (Preprocessing)



1. Data was gathered from baseballsavant.com. It has 3920 samples and 70 features. Features include (among others) Launch Angle, Home Runs, Sprint Speed, etc.
2. We calculated the percent change in wOBA from year to year for each player, and used this as the label. For example, if player A had a wOBA of .300 in 2015 and a wOBA of .330 in 2016, their percent change would be a 10% increase.
3. After splitting the data into attributes and labels, Standardization was used. This is needed because different statistics have very different looking distributions. For example, sprint speed for most players is very closely “bunched together,” while home runs have much more variance.
4. Pearson correlation was run to perform feature selection. Attributes that were not significantly correlated with the label (< 0.25) were removed. Additionally, attributes that were excessively correlated with one another (> 0.9) were also removed.
5. After preprocessing, our data contained attributes such as sweetspot percentage, walk rate, line drive percent, and in-zone swing percent.



Methodology (Models and Preliminary Results)

- We started with K-Nearest Neighbors.
  - Reasoning: The dataset does not have too many rows, so the downside of being memory-intensive is not as relevant. Also, we are focused on making an interpretable model and determining which attributes are most critical to predicting wOBA.
  - Additionally, prospective major league baseball players are often compared to current players in the league (draft comparisons). We are doing a sort of computer-based player comparison with K-Nearest Neighbors.
  - Using a train-test split of 78-12 and neighbors = 3, we obtained a mean absolute error of 0.04.
  - Existing research obtained a mean absolute error of 0.02.
- Next, we tried gradient boosted trees.
  - Reasoning: we wanted to try another model that can model nonlinear decision boundaries. Additionally, the paper we're comparing our results to used gradient boosting, so it's a good way for us to better compare the impact of the additional statistics we used.
  - Using a train-test split of 78-12, we obtained an MAE of 0.04