Assignment: Signal v. Noise in Batter Outcomes

Your task is to estimate a batted ball outcome model and create a batting leaderboard.

Why are you being asked to do this?

We are simulating a task you might do if you were working as a data scientist for a baseball team.

What (exactly) are you being asked to do?

There are three components to this assignment:

1. Using the data provided (mlb_event_2023.csv), estimate a batted ball outcome model that is better than the very simple one we estimated in class. Use your model to generate predictions for the linear weights of the outcomes of the batted balls in the held-out test set (mlb_batted_ball_test.csv).

2. Split the data into two halves, and perform an empirical validation of the model-based sample size thresholds we estimated in class. Specifically:

   - Is it true that the correlation between first-half xLW3 and second-half LW is higher than the correlation between first-half LW and second-half LW?
   - Is it true that the correlation between first-half xLW2 and second-half xLW3 is higher than the correlation between first-half xLW3 and second-half xLW3?
   - For players with fewer than 150 first-half batted balls, is it true that the correlation between first-half xLW1 and second-half xLW2 is higher than the correlation between first-half xLW2 and second-half xLW2?
   - For players with more than 150 first-half batted balls, is it true that the correlation between first-half xLW2 and second-half xLW2 is higher than the correlation between first-half xLW1 and second-half xLW2?

   For the metrics xLW*, use the model you estimated in component 1, not the naive model from class. When calculating these correlations, take care with how you handle tiny-sample ($n < 30$) players.

3. Create a batting leaderboard that includes every player with at least one plate appearance as a batter in 2023. Your leaderboard must include player name, number of plate appearances and **at most three** additional columns describing the player's batting performance. You choose from traditional statistics (e.g. BB%, K%, ISO, BABIP) as well as the metrics we have covered in class. Your goal is to paint as clear a picture as possible of the player's batting performance subject to the three-number constraint. Which three numbers will you choose?

Submission Requirements

- A PDF report (max 4 pages) summarizing your findings, including at minimum the following:
  - A description of how you estimated your batted ball outcome model
  - A summary of your findings regarding empirical validation of model-based sample size thresholds
  - An explanation of why you choose the three numbers you included in your batting leaderboard
- A CSV of predictions for the held-out batted ball test set, with columns `play_id` and `pred_lw`
- A CSV leaderboard of batting performance in 2023
- A R script (.R file extension) with all of the code you used to generate your report and the CSVs

Reminders

- Prepare your report as if your audience is a baseball executive who has not seen the assignment prompt. Write clearly and concisely, and format your report in a way that makes it easy to read.
- In this class we value **critical thinking**! Don't just parrot what you've been taught—bring your own experiences to bear on the assignment. If you disagree with something we've covered in class, that's strongly encouraged! Be sure to explain your reasoning.
- Please anonymize your submission by removing any personally identifiable information.

How will your grade be determined?

You will get feedback on your work product based on several criteria. Within each of those criteria, the feedback will be: Missing (0%), Needs Improvement (70%), Good (85%) or Exceeds Expectations (100%). Your grade on the assignment will be the average of the grades across criteria. The criteria are:

1. **Description of batted ball outcome model.** Did you explain your model for the batted ball outcome model with sufficient detail that another baseball data scientist could replicate it?
2. **Performance of batted ball outcome model.** Did your test-set predictions out-perform those made by the naive model from class? (a statistically significant difference → Exceeds Expectations)
3. **Validation of sample size thresholds.** Did you correctly calculate the correlations requested in component 2? Did you correctly interpret the results?
4. **Critical thinking.** Does your analysis exhibit a depth of thinking about the problem, or are you just applying the methods we've covered in class?
5. **Written communication.** Did you write clearly and concisely? Did you organize your key ideas with the evidence supporting them? Did you format your report in a way that makes it easy to read?