Assignment #3: Pitch Outcome Model

Your task is to build a swing probability model.

Why are you being asked to do this?

If you ever interview with a baseball club, there is a decent chance you will be asked to estimate one component of a pitch outcome model. Predictive modeling is a broadly useful skill in baseball analytics.

What (exactly) are you being asked to do?

Pitch outcome modeling is more challenging than hit outcome modeling because it involves higher-dimensional observations, and the observations are greater in number (presenting computational challenges). The most computationally challenging component is the swing probability model. Using the provided data (mlb_pitch_2023.csv), you are to train a model for the probability that the batter swings at a pitch, based on pitch characteristics and contextual features, such as batter handedness and count.

Using your swing probability model, predict the swing probability for the test set (mlb_pitch_test.csv). Your predictions will be evaluated using negative log-loss: If you predict p on a pitch that does result in a swing, your loss is $-\log(p)$; if you predict p on a pitch that does NOT result in a swing, your loss is $-\log(1-p)$. Success means beating the simple model from the class R tutorial on the test set. In addition to the probability predictions, you must predict the test negative log-loss for both your model and the simple model from the class R tutorial.

When developing your swing outcome model, you will need to rely on both statistical modeling expertise and baseball domain knowledge for deciding (a) what type of model to use, (b) which features to include, and (c) how to include them. In your writeup, describe your reasoning for your design choices, and describe the tradeoff between your choices and alternative choices you could have made.

Submission Requirements

- A PDF report (max 4 pages) summarizing your findings, including at minimum the following:
  - A description of (and reasoning for) your methodology for predicting swing probability
  - A prediction of the test negative log-loss for your model and the simple model from class
  - At least one data visualization that tells a story about your results
- A CSV of predictions, with columns `pitch_id` and `prob_swing`
- A R script (.R file extension) with all of the code you used to generate your report and the CSV

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 (including file paths in your R script that contain things like a username!).

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 methodology.** Did you explain your methodology with sufficient detail? Did you make sound design choices, from both the statistical modeling perspective and the baseball perspective?

2. **Model performance**. Did your model outperform the simple model on the test set? Did you accurately predict the test-set performances of both models?

3. **Data visualization.** Did you include a data visualization that tells a compelling story?

4. **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?