# Assignment: Pitch Outcome Model

This assignment is designed to simulate the type of open-ended data assessment you might face in an interview with a professional baseball club's analytics department. Your task is to train and validate a contact probability model (conditioned on swing), and then interpret the results to evaluate player performance.

## Part 1: Predictive Modeling

In the pitch outcome model R tutorial, we fit a simple logistic regression model for the probability of the batter making contact with a pitch, assuming that they swing and given the trajectory of the pitch. This simple model was based only on count, pitch location and pitch speed. Your task is to improve upon this simple model by thinking about additional predictor variables you could derive from the pitch tracking data (and how to use them in the logistic regression model). Use the data from pitch.csv (available on Canvas) to fit this improved model. Create a data visualization to illustrate the relationship between the most important pitch tracking features and the contact outcome. Explain the biggest weakness of your model.

Bonus: Additionally fit a gradient boosting model (recommended package: xgboost) and visualize the resulting model fit. Report the tuning parameters you used and explain how you chose them.

## Part 2: Model Validation

Estimate and report the out-of-sample test error for the simple logistic regression model and the improved model you estimated in Part 1, using negative log loss.[1] Describe how you estimated the out-of-sample test error. Generate contact probability predictions for the pitches in pitch_test.csv. Upload your predictions as a CSV file with two columns: pitch_id and prob_contact.

Bonus: Include a gradient boosting model when reporting the out-of-sample test errors. If you do this, generate predictions in the CSV only for the model which you think will perform best on the test data.

## Part 3: Player Evaluation

For each pitcher, calculate observed contact rate per swing (a); expected contact rate per swing (b); and residual (observed – expected) contact rate per swing (c). Perform regression to the mean to estimate each pitcher's true talent for (b) and (c); and sum these two regressed metrics to create predicted contact rate (d). To report your results, create a leaderboard of the best pitchers according to (d), including columns for pitcher name and number of pitches thrown, as well as (a), (b) and (c). Visualize all pitchers by creating a plot of (a) on the x-axis vs. (d) on the y-axis. Describe what you find interesting about these results.

Bonus: Identify two pitchers who exhibit strong evidence of over- and under-performing the contact rate expected from their pitch trajectories (one pitcher from each extreme). Using your understanding of baseball, explain why you think these two pitchers could truly have better or worse contact rates than pitch tracking data would suggest. Rather than describing in general terms why a pitcher might be better/worse than their pitch tracking data, write specifically about what you think could be happening with these pitchers.

## Deliverables

- A PDF report (max 4 pages) with labelled sections corresponding to Part 1, Part 2, and Part 3. This report should contain absolutely no R code. It should be well written and easy to follow. Summarize your findings and leverage data visualization as appropriate.
- A CSV of predictions, with columns `pitch_id` and `prob_contact`
- A R script (.R file extension) with all of the code you used to generate the results for your report

[1] For example, if you predicted contact probability $p$ and the pitch were to result in contact, your error would be $-\log(p)$. If that same pitch were to result in a swing-and-miss, your error would be $-\log(1-p)$.

## Rubric

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. **Predictive Modeling.** Is the implementation correct? Are your interpretations justified?
2. **Model Validation.** Is the implementation correct? Are your interpretations justified?
3. **Player Evaluation.** Is the implementation correct? Are your interpretations justified?
4. **Data Visualization.** Have you used data visualization effectively to convey your findings?
5. **Written Communication.** Is the report well written and easy to follow?

## Reminders

- Please **anonymize** your submission by removing any personally identifiable information (including file paths in your R script that contain things like usernames!).
- In this class we value **creativity** and **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.