

Senior Quantitative Analyst - Take-Home Project

Project Overview

Time Allocation: We expect this project to take 6-10 hours of focused work. Please submit within 7 days of receiving the data.

Deliverables:

1. An HTML report (generated from RMarkdown or Jupyter notebook)
 2. All code used in your analysis (should be reproducible)
 3. Brief README with setup instructions (if necessary)
-

Research Question

"Projecting Hitter Contact Rate: Given a hitter's performance through June, how accurately can we project their contact rate for the remainder of the season?"

Background

Contact rate is a fundamental component of hitting ability. However, with limited sample sizes and varying competition levels, it can be challenging to distinguish genuine skill from random variation.

Your task is to build a model that projects rest-of-season contact rate using data through June, and to quantify the uncertainty in these projections.

Data Provided

You will receive pitch-by-pitch Statcast data from the 2024 season for approximately 400 hitters who saw at least 500 pitches (pitch_data.csv), as well as hitter metadata (hitter_data.csv).

Key Fields:

- `pitch_date`: Date
- `batter_id`: Unique identifier for each hitter
- `pitcher_id`: Unique identifier for each pitcher
- `batter_stand`: Hitter's batting side (L/R)
- `pitcher_throws`: Pitcher's throwing hand (L/R)
- `description`: Pitch outcome (e.g., "swinging_strike", "foul", "in play", "ball", "called_strike")
- `pitch_type`: Type of pitch thrown
- `ball, strike, out`: Count, out state
- `vertical_release`: Release height
- `horizontal_release`: Release side
- `release_speed`: Pitch velocity
- `induced_vertical_break, horizontal_break`: Magnus movement
- `plate_x, plate_z`: Pitch location at home plate

- `sz_top`, `sz_bot`: Strike zone top and bottom
-

Specific Tasks

1. Data Processing & Feature Engineering (20%)

- Define contact rate
- Split the data appropriately
- Create any features you believe are relevant for modeling contact rate
- Handle missing data appropriately

2. Model Development (40%)

Build a model to project rest-of-season contact rate using only data through June.

Your model should:

- Produce a projected contact rate for each hitter for the remainder of the season
- Provide uncertainty estimates (e.g., confidence/credible intervals)
- Account for the varying sample sizes across hitters

Note: You may use any statistical or machine learning approach you deem appropriate. We're interested in your modeling choices and how you justify them.

3. Model Evaluation (20%)

Evaluate your model's performance:

- How accurate were your projections compared to actual rest-of-season performance?
- How well-calibrated were your uncertainty estimates?
- Which types of hitters were easiest/hardest to project accurately?
- What are the key limitations of your approach?

4. Insights & Recommendations (20%)

Translate your findings into actionable insights:

- Which hitters' true contact ability likely exceeds their first-half performance?
 - Which hitters' early-season results likely overstate their underlying skill?
 - How should decision-makers use these projections given the uncertainty?
 - What additional data would most improve projection accuracy, and why? (Be specific about what information would be valuable and how you would incorporate it)
-

Evaluation Criteria

We will evaluate your submission on:

Statistical Rigor:

- Appropriateness of modeling approach

- Proper treatment of uncertainty
- Valid model evaluation methods
- Clear statement of assumptions and limitations

Technical Implementation:

- Clean, well-documented code
- Reproducible analysis
- Efficient computation
- Appropriate use of tools and libraries

Communication:

- Clear problem formulation
- Effective visualizations
- Insights accessible to non-technical decision-makers
- Honest discussion of what the model can and cannot tell us

Baseball Context:

- Sensible feature engineering choices
 - Understanding of the practical problem
 - Realistic assessment of model utility
-

Submission Guidelines

Format:

- Submit as a compressed folder (.zip) containing:
 - HTML report (main deliverable)
 - All code files (R/Python scripts or notebooks)
 - README.md with setup instructions (if necessary)

Code Requirements:

- Should be runnable by us (include package/library requirements)
- Use clear variable names and comments
- Organize logically (data processing → modeling → evaluation → insights)

Report Structure (suggested):

1. Introduction & Research Question
 2. Data & Methodology
 3. Model Development & Justification
 4. Results & Evaluation
 5. Insights & Recommendations
 6. Limitations & Future Directions
-

Notes

- Given the limited time to work on this project, we are less interested in maximizing accuracy than we are in your process and ability to discuss your approach. If you are unable to answer all questions due to time constraints, please explain why and how you would approach with additional time.
 - There is no single "correct" approach to this problem. We're interested in your thought process and decision-making.
 - Clearly communicate any assumptions you make.
 - If you run into issues with the data or have questions about the task, document your assumptions and proceed.
 - We value clear thinking and honest assessment over complex models.
-