Batter Pitch Mixes

Pitch decisions are motivated by a variety of factors, including the pitch types that the pitcher typically throws, that they throw best, and that batters hit worst. Pitching analysis often evaluates a pitcher's "arsenal", which is the set of pitch types that a pitcher throws to batters. Less research has been devoted to the inverse, the set of pitch types that a batter *receives* from pitchers – or a batter's pitch mix.

If a batter's strengths and weaknesses are a motivating factor behind the pitch types that they oppose, then we could expect different batters to face different pitch mixes from one another. Correspondingly, if we could accurately predict the pitch mix that each batter will receive, that information could be incredibly valuable to teams in a variety of areas, ranging from game preparation to player evaluation.

Submission

Given pitch-by-pitch data from the 2021-2023 MLB regular seasons, we would like you to predict the pitch mixes that batters faced in 2024 across three groups: fastballs, breaking balls, and off-speed pitches.

To predict these pitch mixes, we are providing you with a training dataset that contains pitch-level data for all players that faced at least 1,000 MLB pitches from 2021-2023, and a blank prediction dataset containing the same players that *also* faced at least 1,000 MLB pitches in 2024. For each batter in the blank prediction dataset – and only using the information provided – please report the proportions across the three pitch groups that you estimate each batter will have faced in 2024. Remember, you will need to map the pitch types to fastballs (FB), breaking balls (BB), and off-speed pitches (OS).

For your submission, we would like you to include three components:

(1) The code used to build your model and a .csv file of your predictions titled predictions.csv following the sample_submission.csv example;
(2) A concise (~ 1-2 page) technical report, intended for the Director of Analytics. This report should briefly summarize your data, your model, and any limitations to your modeling process for an audience with a high technical expertise, but not enough time to read your code.
(3) A more in-depth piece of communication, intended for the coaching staff. This product should make your model predictions actionable for an audience with less technical expertise, but a strong interest in implementing your findings on the field. This output should at least cover a subset of the players in the submission file. For each player you cover, you should, at a minimum, report the pitch mix you expect that player to see; above and beyond submissions could provide additional context or nuance to the predictions. In particular, prioritize engagement with the data and the output from your model so that the stakeholders can learn more about your predictions (e.g. build a simple dashboard or create a report with graphs and/or tables that effectively visualize and communicate your model predictions).

When evaluating your submission, we will focus on three key areas:

(1) Data Science (50%): This area includes your approach, your modeling decisions, and your use of statistics and baseball knowledge to justify your process;
(2) Communication (49%): This area includes your ability to document your code, communicate technical concepts and limitations, make results and data accessible to a non-technical audience, visualize data, and provide valuable baseball insights;
(3) Accuracy (1%): This area covers how accurately your model predicts the results.

Ideally, we would like you to spend around 8 hours on this take home assessment. Of those 8 hours, we would recommend splitting the time so that 2-3 hours go towards modeling, 1 hour goes towards the technical report, and the remainder goes towards the non-technical communication.

When you submit, upload a zip file containing your submission or submit a public share link to a hosted repository (e.g. GitHub or Google Drive) that contains your submission. If any part of your submission is hosted on another service (e.g. shinyapps.io), please make sure both the URL and the code are included.

Data

We are providing three files for this problem: data.csv, predictions.csv, and sample_submission.csv. Please do not use any public data as supplement; all analysis should use only the data we have provided.

data.csv

This file contains all the training data available for your analysis.

PITCH_TYPE – The type of pitch derived from Statcast (abbreviated)

- CH – Changeup
- CS – Slow Curve
- CU – Curveball
- EP – Eephus
- FA – Other
- FC – Cutter
- FF – 4-Seam Fastball
- FO – Forkball
- FS – Split-Finger
- KC – Knuckle Curve
- KN – Knuckleball
- PO – Pitch Out
- SC – Screwball
- SI – Sinker
- SL – Slider
- ST – Sweeper
- SV – Slurve

PITCH_NAME – The type of pitch derived from Statcast (written out)

PLAYER_NAME - Player's name tied to the event

BATTER_ID – MLB player ID tied to the play event

PITCHER_ID – MLB player ID tied to the play event

BAT_SIDE – Side of the plate batter is standing

THROW_SIDE – Hand pitcher throws with

GAME_PK - Unique Id for Game

GAME_YEAR – Year game took place

GAME_DATE – Date of the game

HOME_TEAM – Abbreviation of the home team

AWAY_TEAM – Abbreviation of the away team

INNING – Pre-pitch inning number

INNING_TOPBOT - Pre-pitch top or bottom of inning

AT_BAT_NUMBER - Plate appearance number of the game

PITCH_NUMBER - Total pitch number of the plate appearance

OUTS_WHEN_UP - Pre-pitch number of outs

BALLS – Pre-pitch number of balls in count

STRIKES – Pre-pitch number of strikes in count

ON_1B - Pre-pitch MLB Player Id of Runner on 1B

ON_2B - Pre-pitch MLB Player Id of Runner on 2B

ON_3B - Pre-pitch MLB Player Id of Runner on 3B

IF_FIELDING_ALIGNMENT - Infield fielding alignment at the time of the pitch

OF_FIELDING_ALIGNMENT - Outfield fielding alignment at the time of the pitch

EVENTS – Event of the resulting plate appearance

DESCRIPTION – Description of the resulting pitch

TYPE – Short hand of pitch result; B = ball, S = strike, X = in play

ZONE – Zone location of the ball when it crosses the plate from the catcher's perspective

PLATE_X – Horizontal position of the ball when it crosses home plate from the catcher's perspective

PLATE_Z – Vertical position of the ball when it crosses home plate from the catcher's perspective

SZ_TOP - Top of the batter's strike zone set by the operator when the ball is halfway to the plate

SZ_BOT - Bottom of the batter's strike zone set by the operator when the ball is halfway to the plate

BB_TYPE – Batted ball type; ground_ball, line_drive, fly_ball, popup

HIT_LOCATION – Position of the first fielder to touch the ball

HC_X - Hit coordinate X of batted ball

HC_Y - Hit coordinate Y of batted ball

HIT_DISTANCE_SC - Projected hit distance of the batted ball

LAUNCH_SPEED - Exit velocity of the batted ball as tracked by Statcast

LAUNCH_ANGLE - Launch angle of the batted ball as tracked by Statcast

ESTIMATED_BA_USING_SPEEDANGLE - Estimated Batting Avg based on launch angle and exit velocity

ESTIMATED_WOBA_USING_SPEEDANGLE - Estimated wOBA based on launch angle and exit velocity

WOBA_VALUE - wOBA value based on result of play

WOBA_DENOM - wOBA denominator based on result of play

BABIP_VALUE - BABIP - based on result of play

ISO_VALUE - ISO value based on result of play

LAUNCH_SPEED_ANGLE - Launch speed/angle zone based on launch angle and exit velocity

1. Weak
2. Topped

3. Under
4. Flare/Burner
5. Solid contact
6. Barrel

HOME_SCORE - Pre-pitch home score

AWAY_SCORE - Pre-pitch away score

BAT_SCORE - Pre-pitch bat team score

FLD_SCORE - Pre-pitch field team score

POST_AWAY_SCORE - Post-pitch home score

POST_HOME_SCORE - Post-pitch away score

POST_BAT_SCORE - Post-pitch bat team score

POST_FLD_SCORE - Post-pitch field team score

DELTA_HOME_WIN_EXP - The change in Win Expectancy before the Plate Appearance and after the Plate Appearance

DELTA_RUN_EXP - The change in Run Expectancy before the Pitch and after the Pitch

predictions.csv

This file is to be used to generate your final submission. This file only contains a column with the BATTER_ID values for which we would like to see your predictions. You will need to save a copy of this file with the prediction columns

sample_submission.csv

This file is what the output from your analysis should look like. You should have one row for each batter and one column for each pitch type for which you make a prediction.