Programming Assignment: Effectiveness of the Infield Shift

Infield shifts have become increasingly prevalent in baseball. Our goal in this assignment is to evaluate the effectiveness of infield shifts by computing each batter's expected BABIP against an infield shift and against a non-shift alignment. For which batters is the shift most effective?

We have provided a dataset for you to perform this analysis. Please prepare a report which addresses the questions below. Your report should include your written responses and any tables or graphs that you think are useful for sharing your findings. In addition to a report, please submit your commented code. Code should be written in Python or R.

Please do not post your code to GitHub or otherwise share the code or the dataset.

Assignment

Part 1. Train a statistical learning model to predict the hit probability of each batted ball in play.

Keep in mind that our goal is to use this model to predict expected BABIP for balls against both shift and non-shift alignments (see 2 below), so your model should account for both the batted ball and the defensive alignment.

We define an infield alignment to be a shift if there are three infielders on the pull side of the infield, and a non-shift otherwise.

- Describe any steps you took to clean the dataset or add additional informative features.
- Discuss how you selected and evaluated the performance of your model.
- Discuss which features are the most important for predicting hit probability.

Part 2. For each batter with at least 100 balls in play in the dataset, compute their expected BABIP if the batter hypothetically faced only shift alignments and their expected BABIP if they hypothetically faced only non-shift alignments.

- For which batters is the shift most effective and most ineffective?
- What are the differences in expected BABIP for these batters against the shift and non-shift?
- How does the shift effectiveness compare between left-handed hitters and right-handed hitters?

The Dataset

infield_bip.csv: This CSV file has StatCast data for every batted ball with landing distance 200 feet or less from the 2018 and 2019 season.

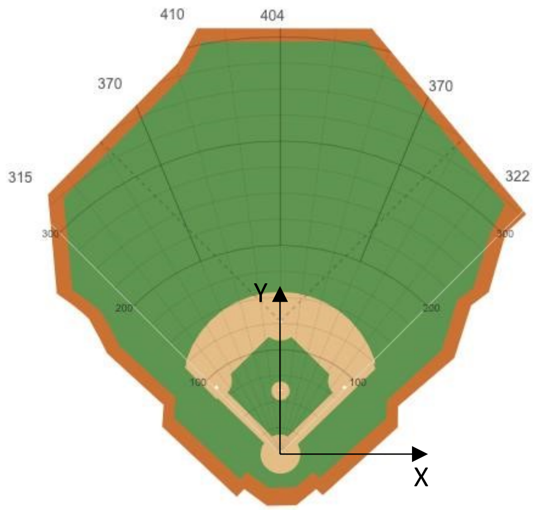
player_lkup.csv: This CSV file provides a lookup for batters and pitchers.

Some additional information:

- **exit_velocity**: the hit speed of the batted ball, in MPH.
- **launch_angle**: the vertical angle of the batted ball, in degrees. -90 degrees is straight down, 0 degrees is on a horizontal plane, +90 degrees is straight up.
- **launch_direction**: the horizontal angle of the batted ball, in degrees. -45 degrees is towards 3B, 0 degrees is up the middle, +45 degrees is towards 1B.

- **hang_time**: time that the ball is in the air, from bat on ball contact to landing (seconds)
- **landing_distance**: landing distance of the batted ball (in feet)
- **x3, y3**: the x & y position of the first baseman at pitch release (in feet). See the coordinate system below.
- **x4, y4**: the x & y position of of the second baseman at pitch release (in feet) See the coordinate system below.
- **x6, y6**: the x & y position of of the shortstop at pitch release (in feet). See the coordinate system below.
- **x5, y5**: the x & y position of of the third baseman at pitch release (in feet). See the coordinate system below.
- **runner_1b**: indicator if there is a runner on 1B.
- **runner_2b**: indicator if there is a runner on 2B.
- **runner_3b**: indicator if there is a runner on 3B.
- **stands**: batter handedness (L or R)
- **throws**: pitcher handedness (L or R)
- **outs:** The number of outs in the inning before the play.
- **strikes:** The number of strikes in the AB before the play.
- **balls**: The number of balls in the AB before the play.
- **game_event_1**: The first game event associated with the batted ball.
- **game_event_2**: The second game event associated with the batted ball.

The origin (0, 0) is at the back tip of home plate. The y-axis is oriented towards the pitcher's mound, the x-axis is oriented perpendicular to


410
404
370
370
315
322
300
300
200
200
100
100
Y
X