**Introduction**

Here is a project proposal to follow up on analysis based on the question below

“One of our infielders, Player X, seems to be struggling in the field. He’s got a great arm, but he’s made a few errors this season and is failing to get to some balls. Could you look into this and identify any problem areas that we can target with drills?”

**Potential Approaches**

1. Exploratory data analysis of Statcast fielding metrics
2. Regression-based model of Outs Above Average
3. Study efficacy of team’s infield shift tactics

**Data**

The dataset used for this analysis will be sourced from MLB Statcast system.

Table I shows the potential structure of the relevant fields of the dataset.

**Table I: MLB Fielding Data**

|  |  |
| --- | --- |
| Field Name | Definition |
| Y Coordinate (Ball) | In/Back (from fielder’s initial starting position) |
| X Coordinate (Ball) | Left/Right (from fielder’s initial starting position |
| Y Coordinate (Fielder) | Fielder’s initial position (In/Back from home) |
| X Coordinate (Fielder) | Fielder’s initial position (Left/Right from home) |
| Batter Speed (ft/sec) | Speed that batter is moving towards first base |
| Throwing Speed | Fielder throwing speed (mph) |
| First Step | Time elapsed from hitter contact to fielder’s first movement |
| First Step efficiency | Angle of deviation from straight line to the ending point of a batted ball vs the actual initial path taken towards the ball |
| Exchange | Time from point fielder receives the ball to releasing |
| Exit Velocity | Speed of ball off bat (mph) |
| Launch Angle | Vertical angle of ball off bat (deg) |
| Shift | 1=Infield Shifted, 2= Infield Not Shifted |

**Proposed Methodology**

The first step would be an exploratory data analysis of the player’s rankings in the fielding metrics. A leaderboard with first step, first step efficiency, exchange times, foot speed (created from X/Y coordinates) and throwing speed for all players can provide a glance at Player X’s strengths and weaknesses in the field.

The next step is to build upon the Outs Above Average model on Baseball Savant and outlined [here](https://technology.mlblogs.com/introducing-infield-outs-above-average-6467e61a98dc). A regression-based machine learning model can estimate a player’s fielding success rate relative to what is expected given the hit trajectory and batter speed. The results from the model can be filtered by hit direction to see the player’s success given where the ball is hit. A variable importance analysis of the features in the model can more specifically provide which fielding metrics are most important.

To visualize this work, an image of the infield with the player’s position at what MLBAM’s Tom Tango calls the “Intercept Point”, the point where the fielder and the ball cross, may be used. This should be overlaid with where the average MLB fielder would be at the intercept point so a comparison of the player to league average may be shown.

Finally, team shift success should be studied. wOBA against on ground balls in shift and non-shift conditions for the Phillies may be compare that to league average. This work is to guard against the possibility that the coach is attributing poor defense to poor positioning.

**Shortcomings**

The success rate model will have some error in it. It’s a model and by design, it will have error in the results. X/Y Coordinate data is generally noisy and may differ from stadium to stadium. The accuracy of the raw data must be fully understood before presenting actionable recommendations. There may be some sample size issues as well. If a coach is observing that a player is slow in a particular direction, there may only be a handful of potential events to analyze.

Sources

mlb.com - <https://technology.mlblogs.com/introducing-infield-outs-above-average-6467e61a98dc>

Baseball Savant - <https://baseballsavant.mlb.com/leaderboard/outs_above_average?type=Fielder&startYear=2021&endYear=2021&split=no&team=&range=year&min=q&pos=&roles=&viz=show&sort=6&sortDir=desc>