Situational Pitch Efficacy

February 8th, 2025

—

Nathan Wright
Colin Montie
Nicolas Thomas

Overview

With the intricate strategies of baseball, decision-making on the mound is critical for a team's success. For this research, our group aims to analyze situational pitch effectiveness by leveraging detailed pitch-level data, including pitch type, location, and batter behavior, to estimate the success probability of various pitches based on specific game contexts.

Research Objectives

- **Factors and Data Sources for Model Development:**
  This study aims to develop a predictive model for assessing pitch effectiveness in baseball games. Key factors and data sources integral to this model include:
    - **Statcast Data:** Comprehensive pitch-level data from Statcast that includes metrics like pitch type, speed, spin rate, and exact pitch location. This data is fundamental for analyzing the dynamics of each pitch and its impact on outcomes.
    - **Brooks Baseball:** Detailed sequencing data on pitch performance, including batter-pitcher matchups and swing outcomes. This source provides insights into how different pitches perform against specific batters under various circumstances.
    - **FanGraphs Metrics:** Advanced metrics such as wOBA and leverage index to assess how pitch types perform against different hitters, which aid in understanding situational effectiveness.

Methodological Approach

The research utilizes a structured approach to model development and analysis:

- **Data Preparation:** Comprehensive selection and cleaning of independent variables (e.g., pitch type, location, swing outcomes) to ensure data quality and consistency.
- **Model Training and Evaluation:** Implementation of machine learning techniques, including Logistic Regression and Random Forest models, to predict pitch effectiveness. Model performance is evaluated using metrics such as accuracy and ROC AUC score.
- **Feature Importance and Selection:** Utilization of feature selection methods to identify the most influential factors affecting pitch success.
- **Predictive Function Development:** Development of a custom predictive function that integrates metrics from selected independent variables to provide tailored predictions for specific pitching scenarios.

Specifications

The R programming language will be leveraged to build the effectiveness model. Data will be collected from Statcast, Brooks Baseball, and FanGraphs, which will ultimately be merged for analysis. Every possible combination of baserunners, outs, and pitch count (0-0,0-1,...) will be

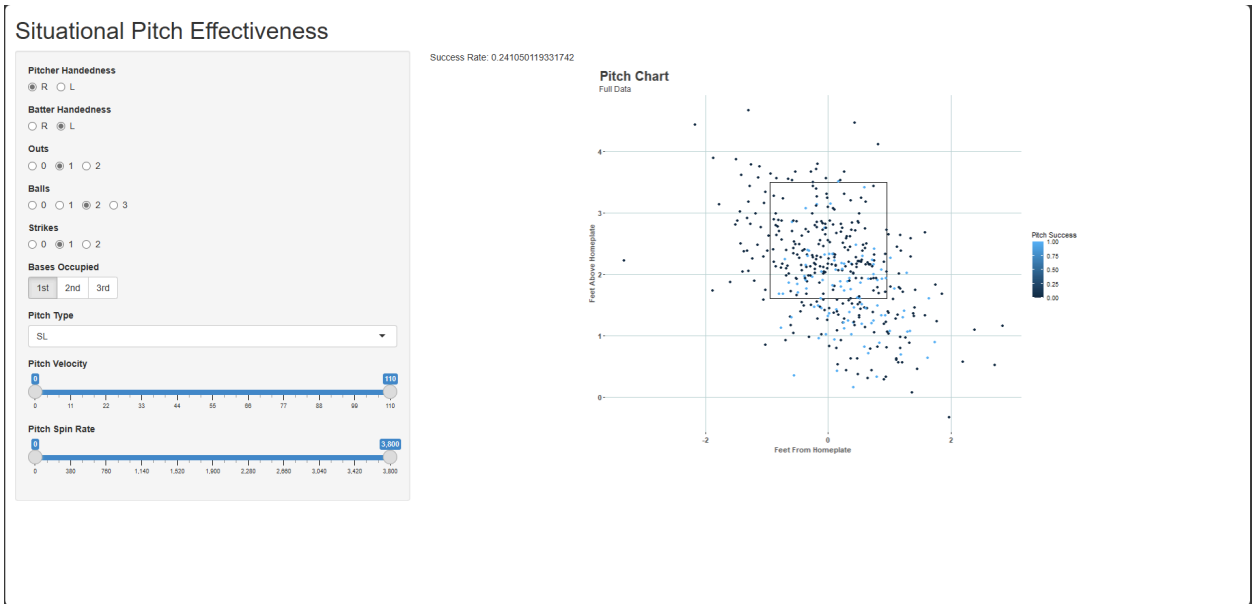
analyzed and success rate will be measured for each possible pitch type and location possibility. Classification algorithms such as Random Forest and traditional regression will be used to fit the gathered data and find the most accurate prediction method. From there, 5-fold cross-validation will be employed to test variable combinations against varying data samples to determine the most predictive factors. Analysis of pitch type and location effectiveness based on situational factors will be conducted, providing numerous visuals to illustrate the variance in success probabilities based on different variables tested in the model.

Step-by-Step Process

1. Imported baseballR package into RStudio
2. Refined dataset to only include 2020 to 2024 seasons (five year sample)
3. Condensed variables to only those relevant for this study: batter_name, pitcher_name, balls, strikes, outs_when_up, pitch_type, pitch_name, plate_x, plate_z, zone, type, events, description, bb_type, inning, inning_topbot, stand, p_throws, pitch_number, bat_score, fld_score, on_3b, on_2b, on_1b, launch_speed_angle, babip_value
4. Generated a listing of every combination of balls, strikes, outs, baserunners on for ease of segmenting master dataset into possible situations
5. Converted certain variables to binary: on_3b, on_2b, on_1b, p_throws, stand
6. Creating whiffs column
7. Defined successful pitch as: whiff, looking strike, or 1:3 launch_speed_angle value (independent of fielding result)
8. Segmented master dataset into separate dataframes for each possible situation created in step 4 listing
9. Stored resulting success rates in a dataframe to be used in analysis finding visualization
10. Used the shinyapp program to generate an interactive application to view success rates

Analysis Results

Results of the study were stored in an interactive visualizer interface app for viewers to edit and test the multitude of combinations of outs, pitch count, runners on, pitcher velocity, spin rate, and handedness, and batter handedness. Listed below is a screenshot of the interface:



Applications and Sport Impact

This research aims to provide actionable insights for pitching coaches and players by identifying optimal pitch strategies under various game situations. By understanding which pitches are most effective against specific batters and circumstances, teams can enhance their decision-making process, leading to improved on-field performance and potentially higher win rates. This work contributes to the evolving landscape of data-driven baseball strategies, emphasizing the importance of situational awareness and success probabilities in pitching decisions.

Sources

baseballR:

@misc{petti_gilani_2021,

  author = {Bill Petti and Saiem Gilani},

  title = {baseballr: The SportsDataverse's R Package for Baseball Data.},

  url = {https://billpetti.github.io/baseballr/},

  year = {2021}

}

baseballsavant: