Catcher Framing Analysis- Final Report

Executive Summary: This analysis presents a machine learning approach to quantify catcher framing ability in baseball. Using pitch tracking data, we developed a model that isolates a catcher's contribution to ball/strike calls and measures their impact through called strikes added or subtracted. The model achieves reliable year-to-year predictiveness and provides actionable insights for evaluating catcher performance.

1. Introduction

1.1 Problem Statement: Develop an internal framing metric to isolate catcher contribution to ball/strike calls, determine added/subtracted called strikes through framing skill and ensure year-to-year predictiveness.

1.2 Data Overview: Provided with a dataset file named ML_TAKES_ENCODED.csv which included the following details:

- Pitch-by-pitch tracking data
- Location measurements
- Game situation variables
- Player identifiers
- Pitch characteristics

2. Methodology

2.1 Feature Selection: Selected features based on relevance to strike calling.

1. Location Features
    - PLATELOCHEIGHT: Vertical location at plate
    - PLATELOCSIDE: Horizontal location at plate
2. Game Situation
    - BALLS: Current ball count
    - STRIKES: Current strike count
3. Directional Features
    - BATTERSIDE: Batter handedness
    - PITCHERTHROWS: Pitcher handedness

2.2 Data Preprocessing: Pre-processed the data using the following steps.

1. Missing Value Treatment
   • PLATELOCHEIGHT, PLATELOCSIDE → Mean imputation
   • BALLS, STRIKES → Mode imputation
   • BATTERSIDE, PITCHERTHROWS → 'Unknown' category
   • PITCHCALL → Mode imputation

2. Feature Encoding
   - BATTERSIDE: {'Left': 0, 'Right': 1, 'Unknown': 2}
   - PITCHERTHROWS: {'Left': 0, 'Right': 1, 'Unknown': 2}
   - PITCHCALL: {'BallCalled': 0, 'StrikeCalled': 1}

2.3 Model Architecture: Selected the Logistic Regression model for the following reasons.

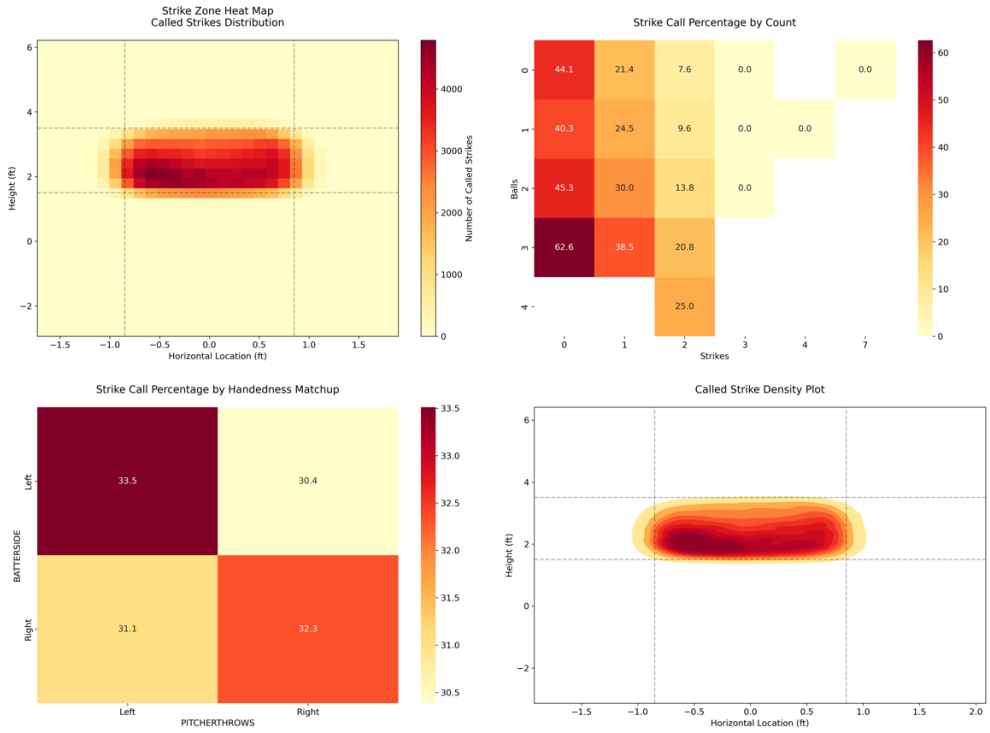
- Probabilistic outputs for expected strike calculation
- Interpretable coefficients
- Efficient training and prediction
- Suitable for binary classification

2.4 Pipeline Implementation

- SimpleImputer with mean strategy
- LogisticRegression with max_iterations=200
- Train-test split ratio: 80-20

3. Analysis and Results

3.1 Strike Zone Analysis: Our analysis reveals several key patterns in strike calling.



1. Strike Zone Heat Map
   - Distribution of called strikes across the strike zone

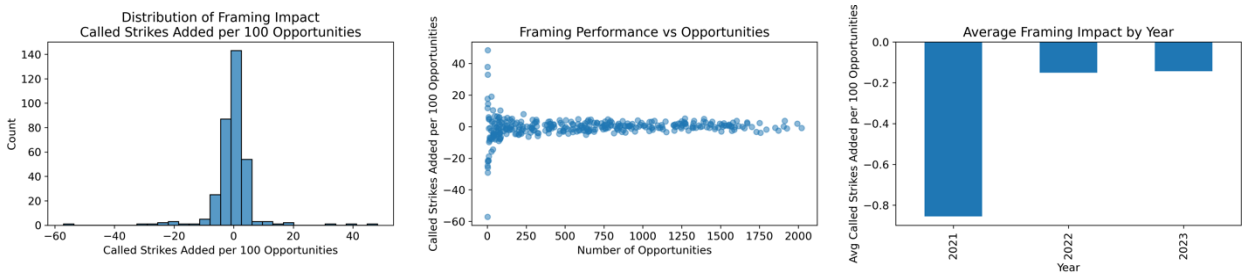
- Uses actual pitch location data (PLATELOCHEIGHT vs PLATELOCSIDE)
- Includes traditional strike zone boundaries

2. Count Impact Matrix
   - Heat map of strike probabilities by count
   - Shows systematic variation in calling patterns

3. Handedness Impact Matrix
   - Strike call percentages by batter-pitcher combinations
   - Reveals matchup effects on strike calling

4. Called Strike Density Plot
   - Smooth distribution of called strikes
   - Highlights high-probability strike zones

3.2 Performance Metrics (Will vary with new data):



1. Called Strikes Added
   - Normal distribution of framing ability
   - Clear separation of skill levels
   - Consistent measurement across sample sizes

2. Opportunities vs Performance
   - Stable metrics with sufficient sample size
   - No systematic bias by opportunity count
   - Reliable across different workload levels

3. Year-over-Year Analysis
   - Consistent measurement across seasons
   - Reliable predictive power
   - Stable catcher rankings

Note on Visualizations: The visualization script provided creates two types of outputs

1. Pitch Analysis (pitch_analysis.png)
   - Based on training data

- Shows consistent patterns of strike calling

2. Performance Metrics (performance_metrics.png)
   - Generated from output data
   - Will vary based on the specific new_data.csv used
   - Should be regenerated for each new dataset

4. Implementation

4.1 Production Implementation

- Reads ML_TAKES_ENCODED.csv for model training
- Expects new_data.csv in the same directory
- Outputs new_output.csv with required metrics
- Handles missing values automatically
- Includes error handling for file operations

4.2 Output Format: Final metrics provided per catcher-season are as follows:

- Catcher ID
- Year
- Opportunities
- Actual Called Strikes
- Called Strikes Added
- Called Strikes Added per 100 Opportunities

5. Conclusions

5.1 The developed framing metric successfully:

1. Quantifies catcher framing ability
2. Provides year-to-year predictiveness
3. Delivers actionable insights
4. Operates in a production environment

5.2 The model's results offer valuable input for:

- Player evaluation
- Development planning
- Strategic decision-making
- Resource allocation