Project Report: MLB Swing Probability Prediction

1. Introduction

The goal of this project is to build a well-calibrated machine learning model that predicts the probability that a batter will swing at a pitch. The dataset is simulated based on MLB pitch-tracking data from pitcher Zac Gallen, including pitch characteristics, release information, and in-game context. Accurately predicting swing probability can help coaches and pitchers optimize pitching strategy, especially in high-leverage situations.

2. Data Description

2.1 Data Source

• Simulated dataset generated using distributions derived from Zac Gallen's pitch profile.
• Approximately 3,567 pitch observations.

2.2 Features

Categorical:
- pitch_name, stand (batter stance)

Numerical:
- release_extension, release_pos_x, release_pos_y, release_pos_z,
  release_speed, release_spin_rate, spin_axis,
  plate_x, plate_z, pfx_x, pfx_z,
  balls, strikes, outs_when_up, sz_top, sz_bot

Target:
- swing (1 = swing, 0 = no swing)

3. Exploratory Data Analysis (EDA)

Key findings from EDA:
• Pitch Type: Changeups have the highest swing rate (~53%), while fastballs and cutters are slightly lower.
• Release Speed: Bimodal distribution at ~82 mph and ~93 mph, reflecting realistic differences between off-speed and fastball pitch types.
• Release Position: Clustered around (-2.8 ft, 6.0 ft), showing consistent mechanics.

Visualizations:

• Proportion of swings by pitch type

- Distribution of release speed
- Hexbin heatmap of release position

4. Modeling Approach

4.1 Preprocessing

- Imputation: mean (numerical), most frequent (categorical)
- Scaling: StandardScaler for numerical variables
- Encoding: OneHotEncoder for categorical variables

4.2 Model

- Algorithm: Random Forest Classifier
- Tuning: GridSearchCV with 5-fold cross-validation
- Hyperparameters searched:
  - n_estimators: [100, 200, 400]
  - max_depth: [20, 40, None]
  - min_samples_split: [4, 10, 20]

4.3 Evaluation Metrics

- Expected Calibration Error (ECE)
- Maximum Calibration Error (MCE)
- Brier Score
- Log Loss
- ROC AUC

5. Results

The model achieved the following performance metrics on the test set:

Metric	Value
Expected Calibration Error (ECE)	0.134
Maximum Calibration Error (MCE)	0.210
Brier Score	0.111
Log Loss	0.377
ROC AUC	0.954

The ROC curve indicates strong discriminative ability, while ECE and MCE suggest the model is close to but not yet fully within acceptable calibration thresholds.

6. Feature Importance Insights

Top predictive features:

- plate_x (0.208)
- plate_z (0.149)
- sz_top / sz_bot (~0.10)
- pfx_z, release_speed, pfx_x (moderate impact)

Pitch location is the primary driver of swing decisions, aligning with baseball intuition.

7. Discussion & Conclusion

If this model were to be applied in a real MLB coaching context, the focus should be on variables that pitchers can directly control or significantly influence. Key features such as release extension, release speed, and pitch type are pitcher-driven and can be strategically adjusted to increase swing probability. Release position also plays a critical role in how a pitch is perceived.

The current model provides strong predictive power but falls slightly short of the calibration standards required for deployment. Further improvements — such as isotonic regression, Platt scaling, or enhanced feature engineering — could turn this into a reliable decision-support tool for pitching strategies.