Baseball ML Analytics

Technical Project Report

Predicting Contact and Hit Outcomes

Using MLB Statcast Data

Generated: February 06, 2026

Stitch Project ID: 9081103112165199837

1. Introduction

Project Background

This project develops ML models to predict baseball outcomes using MLB Statcast data.

The system predicts: (1) whether a batter makes contact, (2) the hit outcome.

Problem Statement

Contact Prediction: Given pitch characteristics, predict contact probability.

Hit Outcome: Given batted ball data, classify as out/single/double/triple/HR.

Scope

Models use 2023-2024 MLB data. Limitations: no park effects, weather, or defensive positioning.

2. Objectives

Contact Model

Goal: Predict contact probability. Target: ROC-AUC > 0.80

Outcome Model

Goal: Classify batted balls. Target: Accuracy > 80%

3. Data Sources

Source: MLB Statcast via pybaseball

Pitch Data: 1.5M pitches (2023-2024)

Batted Balls: 256K events (2023-2024)

Split: Train (2023+H1 2024), Val (Jul-Aug 2024), Test (Sep-Oct 2024)

4. Data Cleaning

Missing Values: Numeric=median, Categorical=mode

Outliers: Exit velocity 20-125 mph, Launch angle -90 to 90 deg

Validation: Required columns verified, date ranges confirmed

5. Feature Engineering

Contact Model (16 features)

Location: plate_x, plate_z, distance_from_center, in_zone

Movement: pfx_x, pfx_z, total_movement

Velocity: release_speed, effective_velocity

Count: balls, strikes, count_advantage, two_strikes, hitters_count

Matchup: same_side, platoon_advantage

Outcome Model (16 features)

Primary: launch_speed, launch_angle

Quality: barrel_indicator, sweet_spot, hard_hit

Spray: spray_angle, distance_from_foul_line, depth_of_hit

Type: is_ground_ball, is_fly_ball, is_line_drive, is_popup

6. Model Development

Contact: LightGBM

Params: num_leaves=31, max_depth=12, lr=0.05, n_estimators=499

Outcome: XGBoost

Params: max_depth=12, lr=0.01, n_estimators=1000

SMOTE: 195K -> 658K samples for class balance

7. Evaluation Results

Contact Model

Test Samples: 125,679

Accuracy: 83.01%

ROC-AUC: 0.8048

Precision: 0.5324 | Recall: 0.0687 | F1: 0.1216

Outcome Model

Test Samples: 21,486

Accuracy: 85.40%

F1 Macro: 0.6427 | F1 Weighted: 0.8713

Per-Class

out: P=0.946 R=0.906 F1=0.925 (n=14628)

single: P=0.855 R=0.761 F1=0.805 (n=4579)

double: P=0.517 R=0.621 F1=0.564 (n=1257)

triple: P=0.036 R=0.317 F1=0.064 (n=104)

home_run: P=0.832 R=0.879 F1=0.855 (n=918)

8. Visualization UI

Platform: Google Stitch (Project ID: 9081103112165199837)

Dashboards: BIP Performance, Model Comparison, Documentation

Design: Teal (#1AA3B0), Space Grotesk, Dark theme

9. Local Deployment

Backend: FastAPI (Python)

Frontend: React + TypeScript + Vite + Tailwind

Endpoints: /health, /api/predict/contact, /api/predict/outcome

10. Conclusion

Achievements

Contact Model ROC-AUC: 80.5%

Outcome Model Accuracy: 85.4%

Full-stack deployment with interactive dashboards

Future Work

Park factors, SHAP explainability, real-time streaming, cloud deployment