MLB Statcast データを用いた 投球タイプ分類と誤分類の構造解析

～カットボール (FC) の識別精度向上に向けて～

花井 龍悟 (B3)

浜田研究室

December 12, 2025

背景と目的

背景: 現代野球におけるデータ活用

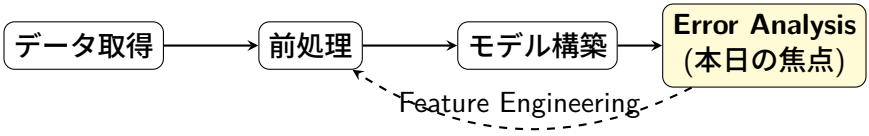
- MLB では全投球の物理データ（Statcast）が計測されている。
- 速度、回転数、変化量などから球種を自動判定する需要が高い。

本研究の目的

1. 機械学習（アンサンブル学習）を用いた高精度な分類器の構築
2. 誤分類（**Error Analysis**）を通じた「人間に近い識別」の実現
3. 特に判別が困難な「カットボール **(FC)**」の精度向上

実装パイプライン

一般的なフローに加え、分析フェーズを重視したサイクルを構築。



- **ライブラリ:** `pybaseball`, `scikit-learn`, `LightGBM` 等
- **モデル:** Random Forest, XGBoost, LightGBM のアンサンブル

モデル構築とベースライン評価

現状のスコア:

- Accuracy: **83.0%**
- Macro F1: 78.5%

課題:

- **全体的には良好だが、特定の球種で迷いが見られる。**
- **特に FC（カットボール）の** Recall **が低い。**

Classification Report
(画像プレースホルダー)

課題の発見：カットボール (FC) の壁

混同行列（Confusion Matrix）による詳細分析。

Confusion Matrix
(FC の誤分類を示す図)

- FC が **SL**（スライダー）および **FF**（ストレート）と混同されている。
- 単純な「変化量」や「球速」だけでは分離できていない可能性。

仮説と検証：なぜ間違えるのか？

仮説

「FC と SL は、変化量 (pfx_x, pfx_z) の分布において重なりが大きく、モデルが境界線を引けていないのではないか？」

検証（物理的特徴量の可視化）：

- 以下の散布図において、赤点（誤分類）が緑点（正解）と完全に重複している。
- 結論: 既存の特徴量（絶対値）だけでは限界がある。

改善へのアプローチ：Feature Engineering

ドメイン知識に基づき、モデルに「コンテキスト」を与える。

1. **相対速度 (Velocity Diff):**
   - その投手の「平均ストレート球速」との差分をとる。
   - 投手ごとの球速差をキャンセルし、球種の定義を明確化。
2. **回転効率 (Spin Efficiency) の推定:**
   - 「回転数が多いのに変化しない（ジャイロ成分）」を FC の特徴として捉える。
   - $\frac{\text{実際の変化量}}{\text{回転数から予測される理論最大変化量}}$

まとめと今後の展望

- **まとめ:**
  - MLB 投球データの分類モデルを構築。
  - Error Analysis により、FC の誤分類原因を物理的に特定。
- **今後の展望:**
  - 提案した新特徴量（相対速度、回転効率）の実装と再学習。
  - デモ用分析ツールの UI 改良。

この後、実際のモデルの動作と分析の様子をデモでお見せします。