

# 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～

浅尾 友志, 川野辺 旭, 花井 龍悟

浜田研究室

2025/12/19

# 背景と目的

## 背景: 現代野球におけるデータ活用

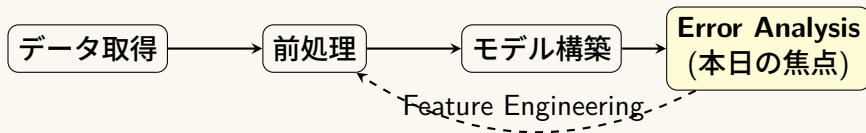
- MLB では全投球の物理データ (Statcast) が計測されている。
- 速度、回転数、変化量などから球種を自動判定する需要が高い。

## 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%

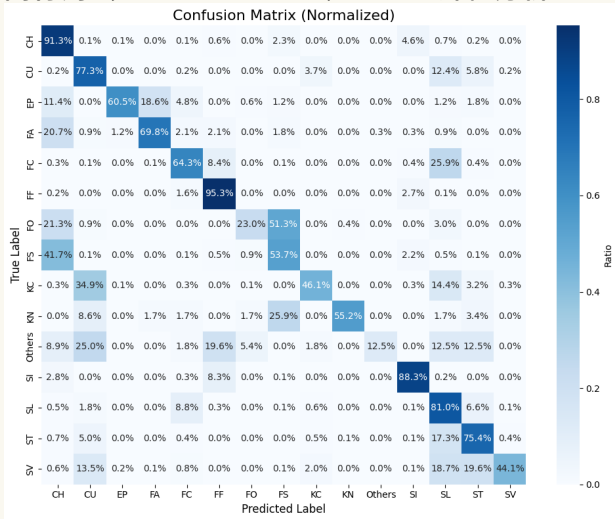
課題:

- 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。

	precision	recall	f1-score	support
CH	0.86	0.91	0.88	23065
CU	0.79	0.77	0.78	14074
EP	0.65	0.60	0.63	167
FA	0.71	0.70	0.70	334
FC	0.71	0.64	0.68	16796
FF	0.94	0.95	0.94	68198
FO	0.42	0.23	0.30	230
FS	0.77	0.54	0.63	4759
KC	0.71	0.46	0.56	4410
KN	0.49	0.55	0.52	58
Others	0.54	0.12	0.20	56
SI	0.90	0.88	0.89	32821
SL	0.75	0.81	0.78	34122
ST	0.73	0.75	0.74	12405
SV	0.71	0.44	0.54	843
accuracy			0.84	212338
macro avg	0.71	0.63	0.65	212338
weighted avg	0.84	0.84	0.84	212338

# 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。



- 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。

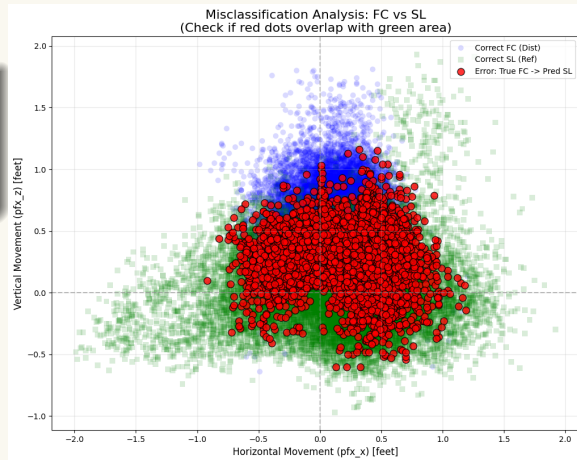
# 結果の分析の一例：カットボール(FC)の壁

## 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）：

- 右図の散布図において、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。



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。