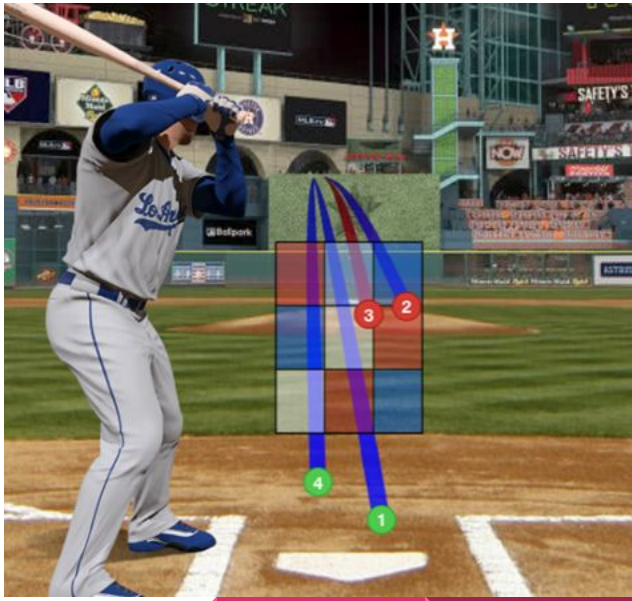
Baseball Pitching Predictor Sprint 3

Yu Song

Overall of the Project

- Predicting pitch type and location.
- Use machine learning models to capture patterns
- Improve batting performance
- Improve franchise record
- “5 million project”



Introduction to the dataset

Game data from 2015 to 2019 season, pitch by pitch in sequence.

Target Variable

- Zone : 1 for strike, 0 for ball
- Pitch type (four-seam, fast ,break, off-speed)

Feature Engineering:

- Past game action (type, zone, result)
- Weather
- Pitch count
- Pitching assignment (Starting or Reliefing)

Cleaning

- Removing pitcher’s entire game script if data is missing.
- Remove unnecessary columns

Key Findings

- Accuracy capped with 55% for binary and 45% for 4-class.
- Ball-Strike count is a key indicator.
  - More four-seam and strikes when ball count is higher than strike count.
  - More breaking ball and balls when strike count is higher than ball count.
- First pitch tends to be a fastball.
- Higher accuracy scrifaces prediction of minority class, especially recalls.

Model Comparison

Accuracy	Ball/Strike Prediction	Pitch Type Prediction
RNN Model	0.54	0.43
Logistic Regression	0.54	0.42
XGboost	0.60	-
Random Forest	0.54	0.45
Ensemble Learning	0.54	-

Criteria:

1. Accuracy
2. Good recall score on minority class
3. Balance between precision and recall

Model Comparison- Recall on Minority Class

Zone/Strike Prediction

	accuracy	recall_ball	recall_strike	precision_ball	precision_strike
Logistic Regression	0.54	0.50	0.60	0.65	0.45
Xgboost	0.60	0.82	0.26	0.62	0.49
Ensemble Learning	0.54	0.46	0.66	0.67	0.45
RNN model	0.54	0.47	0.66	0.69	0.44

Pitch type Prediction

	accuracy	f1_score_Four-seam	f1_score_Other Fastball	f1_score_Breaking	f1_score_Off-Speed	recall_Off-Speed
Logistic Regression	0.42	0.48	0.37	0.47	0.25	0.29
Random Forest Classifier	0.45	0.56	0.31	0.48	0.11	0.07
RNN model	0.43	0.55	0.15	0.51	0.20	0.17

Clustering

Cluster pitchers based on:

- Pitch Mix
- Habit
- Pitch Speed

	Number of Pitcher	Ball/Strike Model	Accuracy	Pitch Type Model	Accuracy
Cluster 1	1007	Ensemble Learning	0.55	Logistic Regression	0.43
Cluster 2	383	Ensemble learning	0.53	Random Forest	0.40
Cluster 3	121	Ensemble learning	0.54	Logistic Regression	0.43

Limit & Future Opportunities

- An overall success rate of 57.5% for making a right prediction.
- Strategy vs. Reality in game-situation.
- More data from both pitching and batting
  - Improve accuracy of the model
  - Improve performance of clustering
- Advanced model to improve accuracy without missing minority class.

Product Demo

Home Team

Away Team

Weather Rainy

Baseball Pitching
Prediction Tool

Inning 4

Top/Bot Top

Home Score 2

Away Score 3

Pitcher Justin Valender

Batter Jalen Durran

First Pitch?

Substitution

2B

Next Pitch will be Four Seam Fastball

3B

1B

Next Pitch will be zone 6, middle inside

Pitch Type Four Seam

Pitch Zone 2

Pitch Result Called Strike

Ball 3

Strike 1

Next Pitch

Next Batter

At Bat result Double

Out 2.00

Switch Sides

Thank you