2024 /11 /05

20:00 - 21:30

11310CS460200 Group 23 Meeting Minutes

Topic	Task Assigned and Model structure confirmation
Place	Discord Voice Chat
Agenda	Discuss about our dataset collection and input format
In attendance	All present
Task Assigned	游松澤: collecting starting pitcher datasets 曾柏勳: collecting closer pitcher datasets 蕭以勝, 楊立慈, 賴允中: model design
Next meeting	Date: 11/12 Time: 9 p.m Objective: Keep working on phase3 (naive implementation of our designed model) Location: Discord Voice Chat

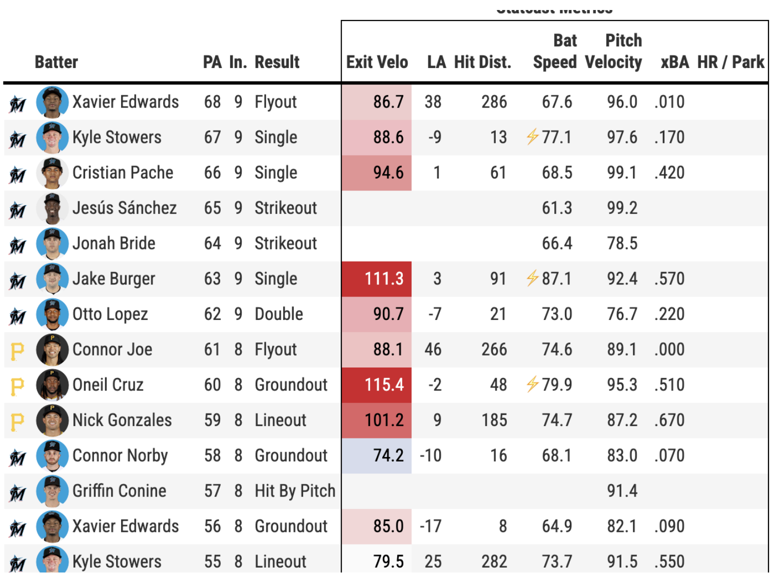
Meeting Summary:

1. We discuss what actually the model input and output is, and also think about the first (easier) version of the dataset csv columns.

This is our concept thoughts of csv:

Date	Pitcher Name	Batter Name	此打席第幾顆球	Pitch-type	isStrike

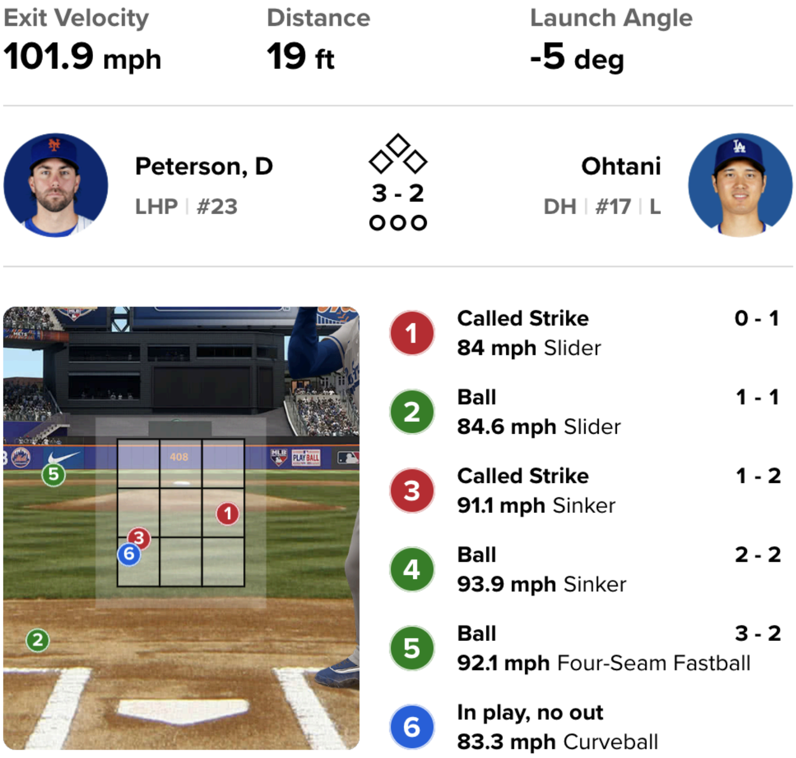
We may add feature like whether this ball isHit (since it may be a foul ball but it will count in isStrike) or weather and so on, but this is for advanced part if we have time to do it. (Since we find out that actually the collection of pitch by pitch dataset is not as easy as we think since we can’t find the raw data as detailed as this in statcast. The game log only provides the ball that has result (out or been a hit) as below)



Statcast Metrics

	Batter	PA	In.	Result	Exit Velo	LA	Hit Dist.	Bat Speed	Pitch Velocity	xBA	HR / Park
M	Xavier Edwards	68	9	Flyout	86.7	38	286	67.6	96.0	.010	
M	Kyle Stowers	67	9	Single	88.6	-9	13	⚡77.1	97.6	.170	
M	Cristian Pache	66	9	Single	94.6	1	61	68.5	99.1	.420	
M	Jesús Sánchez	65	9	Strikeout				61.3	99.2		
M	Jonah Bride	64	9	Strikeout				66.4	78.5		
M	Jake Burger	63	9	Single	111.3	3	91	⚡87.1	92.4	.570	
M	Otto Lopez	62	9	Double	90.7	-7	21	73.0	76.7	.220	
P	Connor Joe	61	8	Flyout	88.1	46	266	74.6	89.1	.000	
P	Oneil Cruz	60	8	Groundout	115.4	-2	48	⚡79.9	95.3	.510	
P	Nick Gonzales	59	8	Lineout	101.2	9	185	74.7	87.2	.670	
M	Connor Norby	58	8	Groundout	74.2	-10	16	68.1	83.0	.070	
M	Griffin Conine	57	8	Hit By Pitch					91.4		
M	Xavier Edwards	56	8	Groundout	85.0	-17	8	64.9	82.1	.090	
M	Kyle Stowers	55	8	Lineout	79.5	25	282	73.7	91.5	.550	

But we can find the detail data in MLB each game like this:



Exit Velocity	Distance	Launch Angle
101.9 mph	19 ft	-5 deg

Peterson, D
LHP | #23

3 - 2

Ohtani
DH | #17 | L

1 Called Strike 0 - 1
84 mph Slider

2 Ball 1 - 1
84.6 mph Slider

3 Called Strike 1 - 2
91.1 mph Sinker

4 Ball 2 - 2
93.9 mph Sinker

5 Ball 3 - 2
92.1 mph Four-Seam Fastball

6 In play, no out
83.3 mph Curveball

However, we need to first find what date this pitcher are pitching. Then, we need to click this game and click every batter in every inning to get this page. This is kinda not systemic and kinda hard to do web crawler. Hence, we aim to find api or other website that have better dataset to collect.

2. We also discussed how we should train our model. Our original thought was that each person should train a model on their own for a pitcher of their choice (possibly of different pitcher types). However, we realized that this is pretty ineffective since everyone will have to do similar work; Also, it fails to apply to a more general setting. We decide that we can instead simply add the pitcher's ID as a feature and let the model figure it out in the training process. This enables us to train our model on a lot more data.

   Additionally, we decided to write the same RNN code and train it on a starting pitcher and a closer. We hope to be able to compare their results and gain insights to what factors might affect their accuracy difference.

3. We decided to divide our work into 3 groups: 2 people on processing the collected data to fit our model's usage, 1 person on sketching the model's rough backbone, and 2 person on implementing the details of the sketch code. This is because we've been busy preparing for midterms of other subjects, so we plan to execute phase 2 & 3 of our timeline in parallel.

A group photo of the discussion session: