

MLB 경기 및 날씨 데이터를 활용한 타구 위치 예측

: Batted Ball Location Prediction
Using MLB Game Statistics and Weather Data

신가연 gayeon0518@gmail.com



Contents

- 1 Introduction
- 2 Dataset Construction
- 3 Modeling & Evaluation
- 4 Conclusion

Introduction

Task Definition

야구



전략적 사고와 예측이 중요한 스포츠

Statcast 시스템으로 타구의 속도, 각도, 비행 경로 등 다양한 메트릭 측정

» 데이터 기반 경기전략 수립 가능

기존 연구의 경우

대부분 이진 혹은 다중 클래스 분류 문제로,
타자의 성과(출루 여부)^[2]
또는 타구 결과(안타/홈런)^[3] 분류에 집중



본 연구는 타구 및 선수 관련 데이터에

날씨 데이터까지 함께 고려하여
타구의 정밀한 낙하 위치(x, y 좌표)를 예측하는
회귀 기반 연구 제안

[2] 김명준, 정준호, "머신러닝(XGBoost)기반 미국프로야구(MLB)의 투구별 안타 및 홈런 예측 모델 개발", 한국정보통신학회논문지, 26, 6, 1325-1333, 2022년 6월

[3] 이승훈, 김준형, "머신러닝을 활용한 미국 프로야구의 투수 및 타자의 유형별 출루 및 아웃 예측 모델", 한국융합학회 학술대회논문집, 제주, 292, 2022년 1월

Introduction

Task Definition

» 예측하고자 하는 것?



X

mlb 경기, 선수 스탯, 날씨 데이터

y

타구 위치

대부분
또는

예측하는

Introduction

Research Purpose

타자의 타구 위치에 적절하게 수비수를 위치시키는 것은 수비에 중요한 전략으로 경기의 승패에 영향

아웃

수비시프트 시행 및 성공



안타

수비시프트 미시행

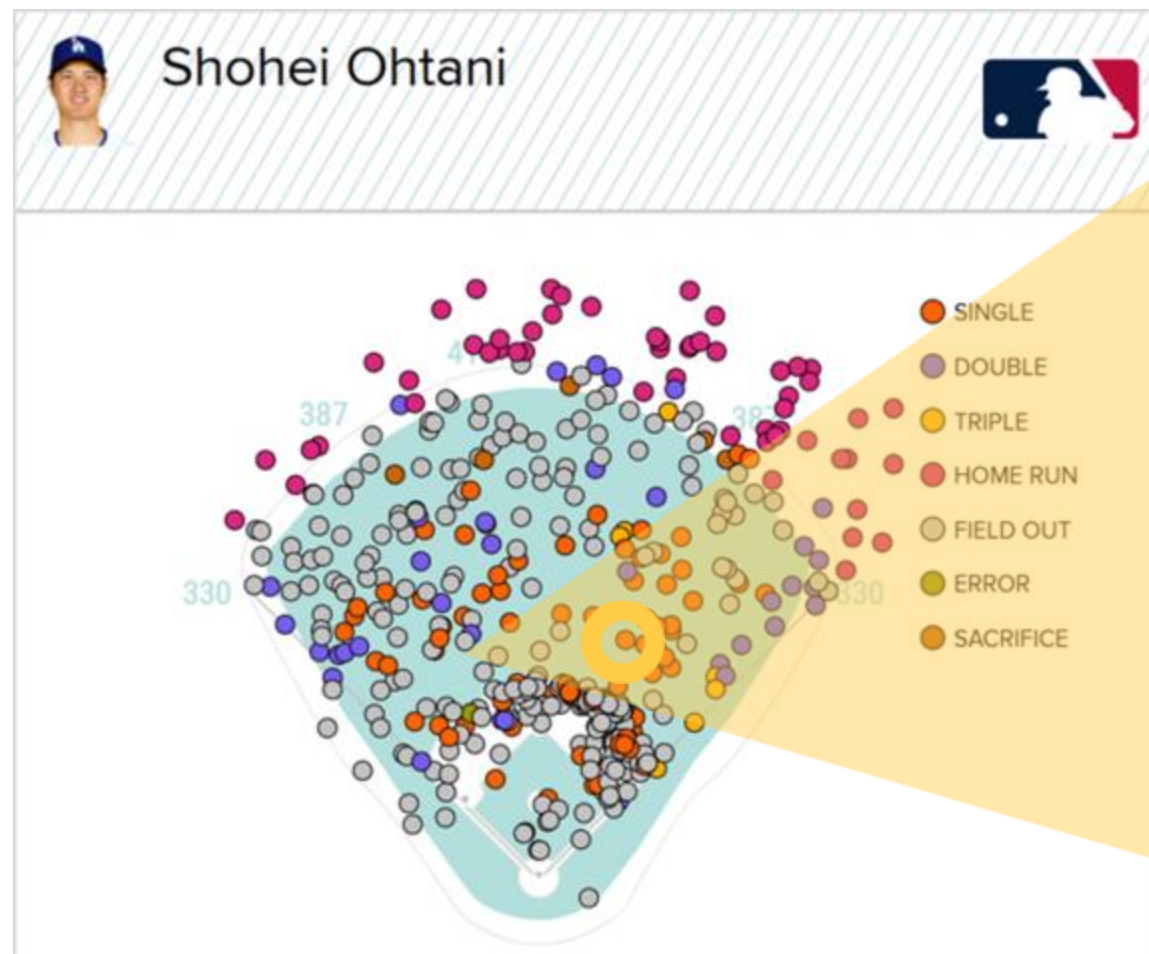


기존 연구와 달리 본 연구는 결과를 예측하여 대응할 수 있게하는 정량적 **인사이트**를 제공
타구 위치 예측을 기반으로 한 수비위치 최적화로 **실점**을 **방지**하는 데 기여

Dataset Construction Data Collection

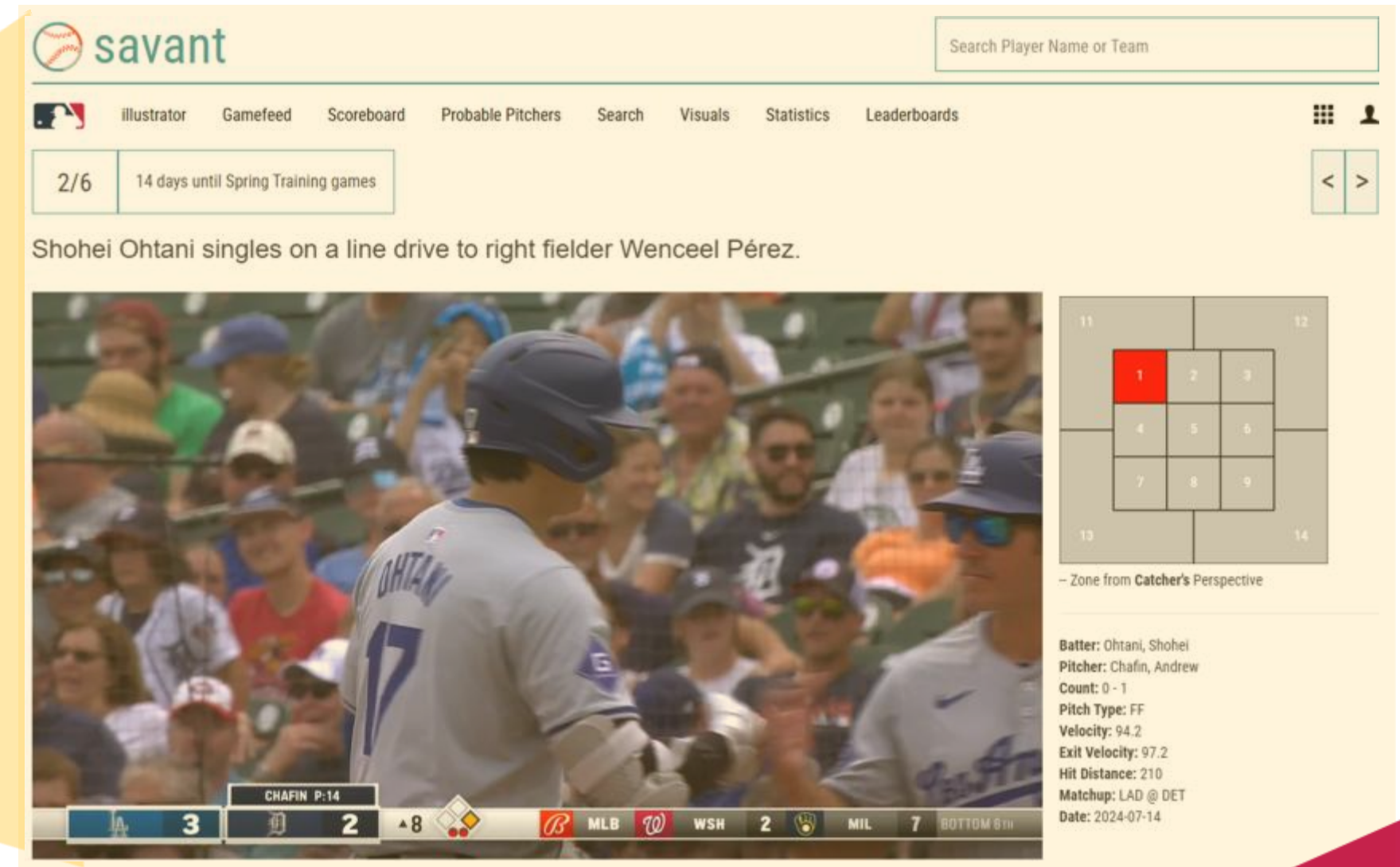
① 타구 정보

“ 2024년 타자 494명의 타구 데이터



<https://baseballsavant.mlb.com/>

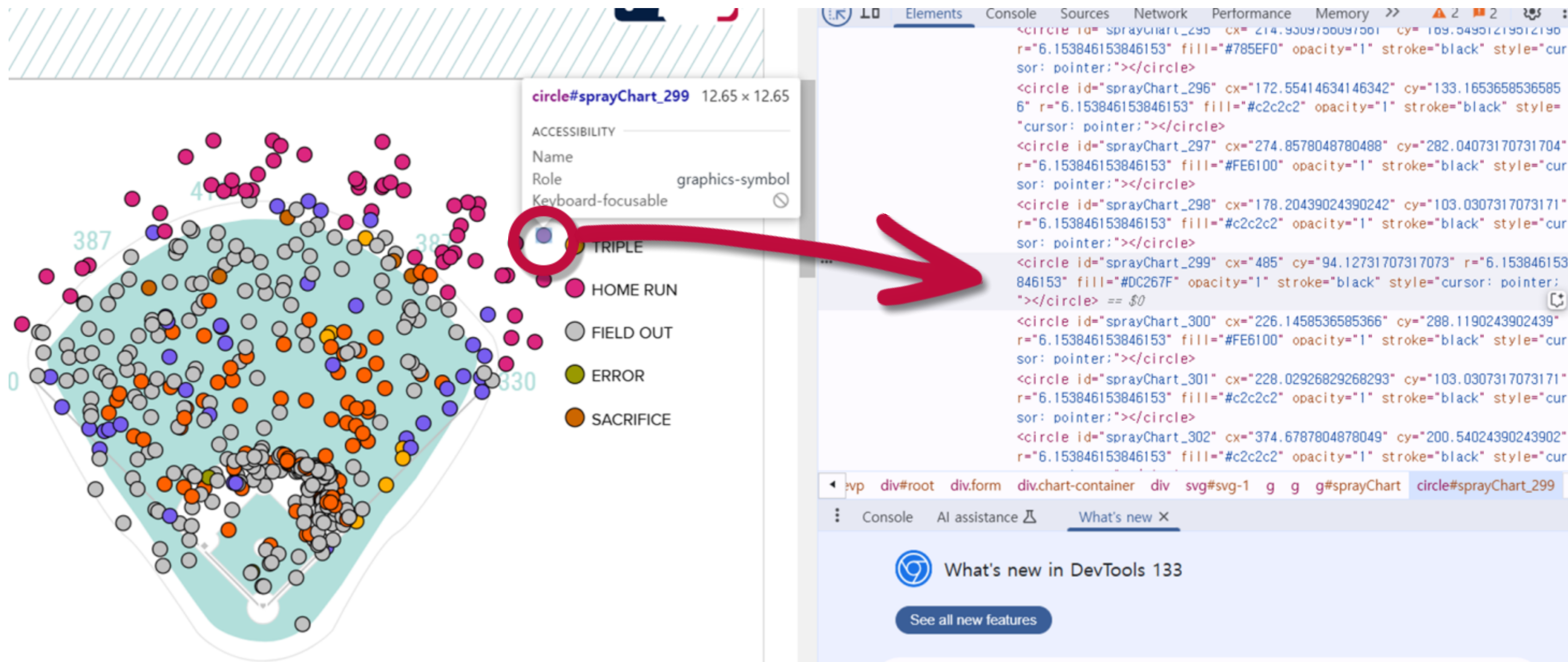
셀레니움을 통한 웹크롤링



Dataset Construction

Data Collection

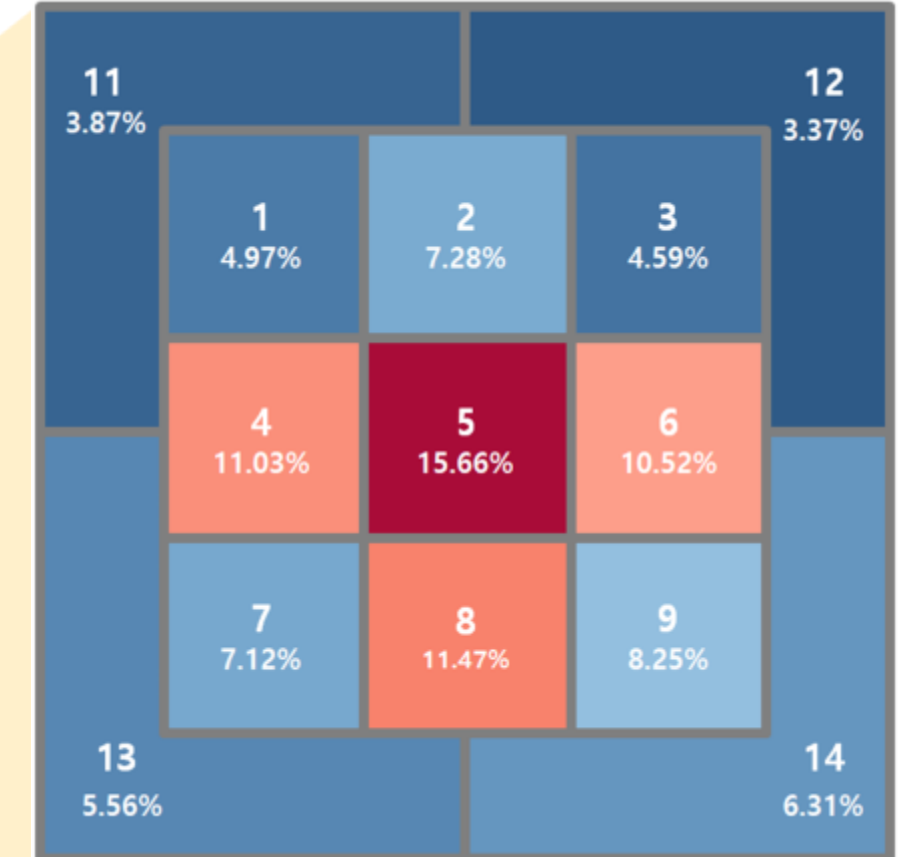
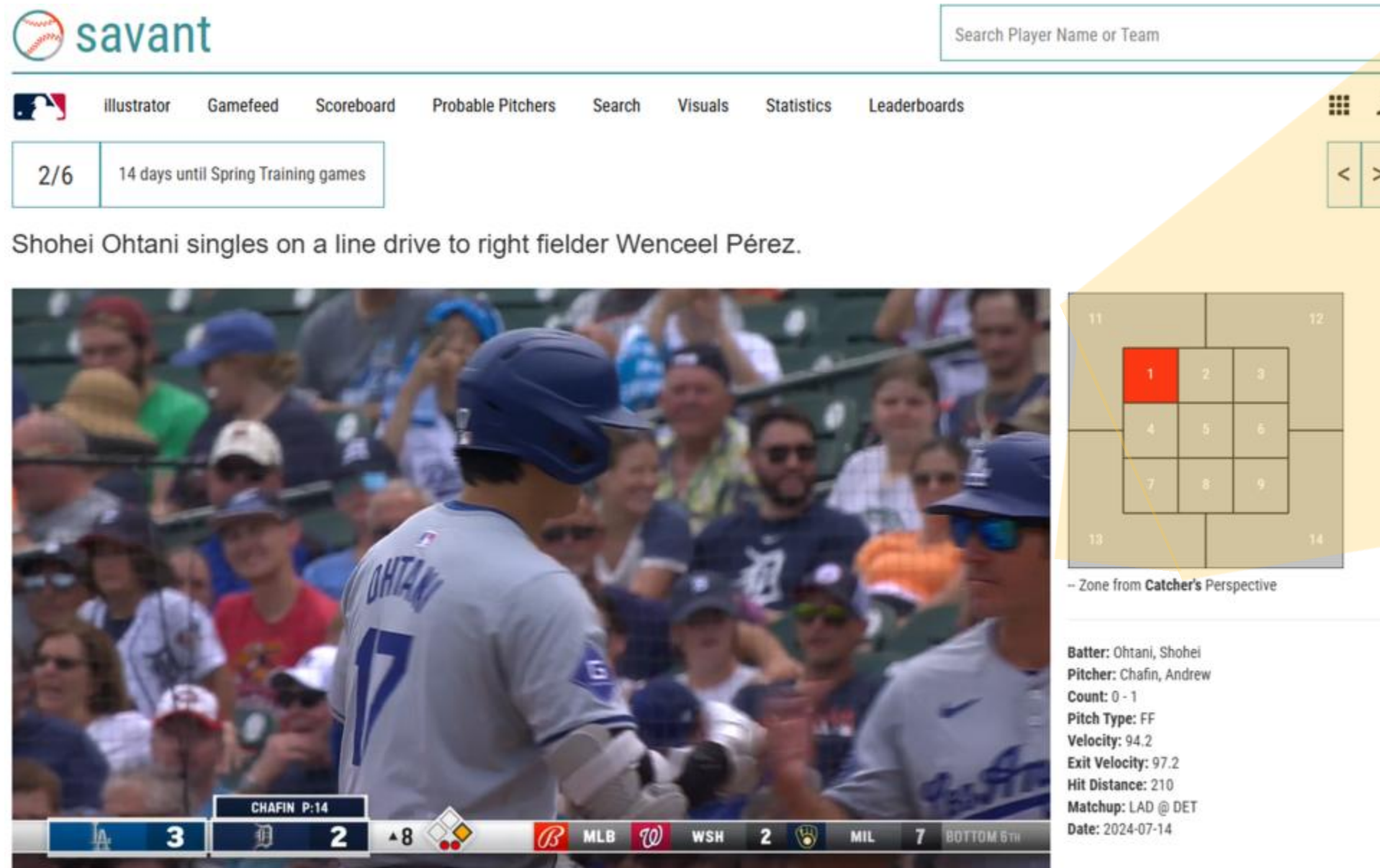
① 타구 정보



Dataset Construction

Data Collection


① 타구 정보




- zone_num
- ball_count
- pitch_type
- velocity
- exit_velocity
- hit_dis
- match_place
- date

Dataset Construction Data Collection

② 선수 정보 savant












Chris Sale

P Bats/Throws: L/L 6' 6" 180LBS | Age: 35

Draft: 2010 | Rd: 1, #13, Chicago White Sox | Florida Gulf Coast

			Age	SLG	ISO	BABIP
			26	.347	.143	.243
			23	.386	.186	.277
			30	.312	.102	.221
			25	.419	.139	.312
			28	.403	.172	.284
			24	.287	.066	.253
			33	.413	.177	.292
8	 Flores, Wilmer	2023	31	.509	.225	.286
9	 Martinez, J.D.	2023	35	.572	.301	.324
10	 Bradley Jr., Jackie	2023	33	.210	.077	.173
11	 Serven, Brian	2023	28	.174	.044	.231
12	 Thomas, Lane	2023	27	.468	.200	.325
13	 Brosseau, Mike	2023	29	.397	.192	.224
14	 Carpenter, Kerry	2023	25	.471	.193	.338

<https://baseballsavant.mlb.com/savant-player/Chris-Sale-694973?stats=statcast-r-pitching-mlb>

623명의 선수 스탯 데이터 수집
: 타자 타격 손, 투수 투구 손, 타자 장타율 등

③ 날씨 정보 Meteostat

Daily Data

This endpoint provides historical weather data. This endpoint is aggregated from multiple days or even months later updating their datasets. Additional model data.

Daily data can be queried for a range of dates.

Endpoint

Daily data is provided through the following endpoint:

GET <https://meteostat.p.rapidapi.com/daily>

Parameters

In order to query data for a specific location, you must provide the latitude and longitude. Do not set the alt parameter.

Parameter	Description
lat	The latitude
lon	The longitude
alt	The elevation
start	The start date
end	The end date
model	Substitution statistics
freq	The time used for
units	The unit parameter

Response

The response body includes the following properties. Please note that all units mentioned below refer to the default units setting.

Parameter	Description	Type
date	The date string (YYYY-MM-DD)	String
tavg	The average air temperature in °C	Float
tmin	The minimum air temperature in °C	Float
tmax	The maximum air temperature in °C	Float
prcp	The daily precipitation total in mm	Float
snow	The maximum snow depth in mm	Integer
wdir	The average wind direction in degrees (°)	Integer
wspd	The average wind speed in km/h	Float
wpgt	The peak wind gust in km/h	Float
pres	The average sea-level air pressure in hPa	Float
tsun	The daily sunshine total in minutes (m)	Integer

```
time tavg tmin tmax prcp snow wdir wspd wpgt pres tsun
2024-08-24 23.4 18.9 28.3 0.0 0.0 306.0 4.8 NaN 1024.0 NaN
```

구장 위치 (위도, 경도) + 날짜 로 온도, 풍속, 강수 등 날씨 데이터 수집

Dataset Construction Preprocessing

범주형 변수 - 라벨인코딩

연속형 변수 - min-max scaling (정규화)

날씨 데이터 - 돐 구장(실내 구장)의 경우

강수량 및 풍속은 0, 기온은 22도로 처리

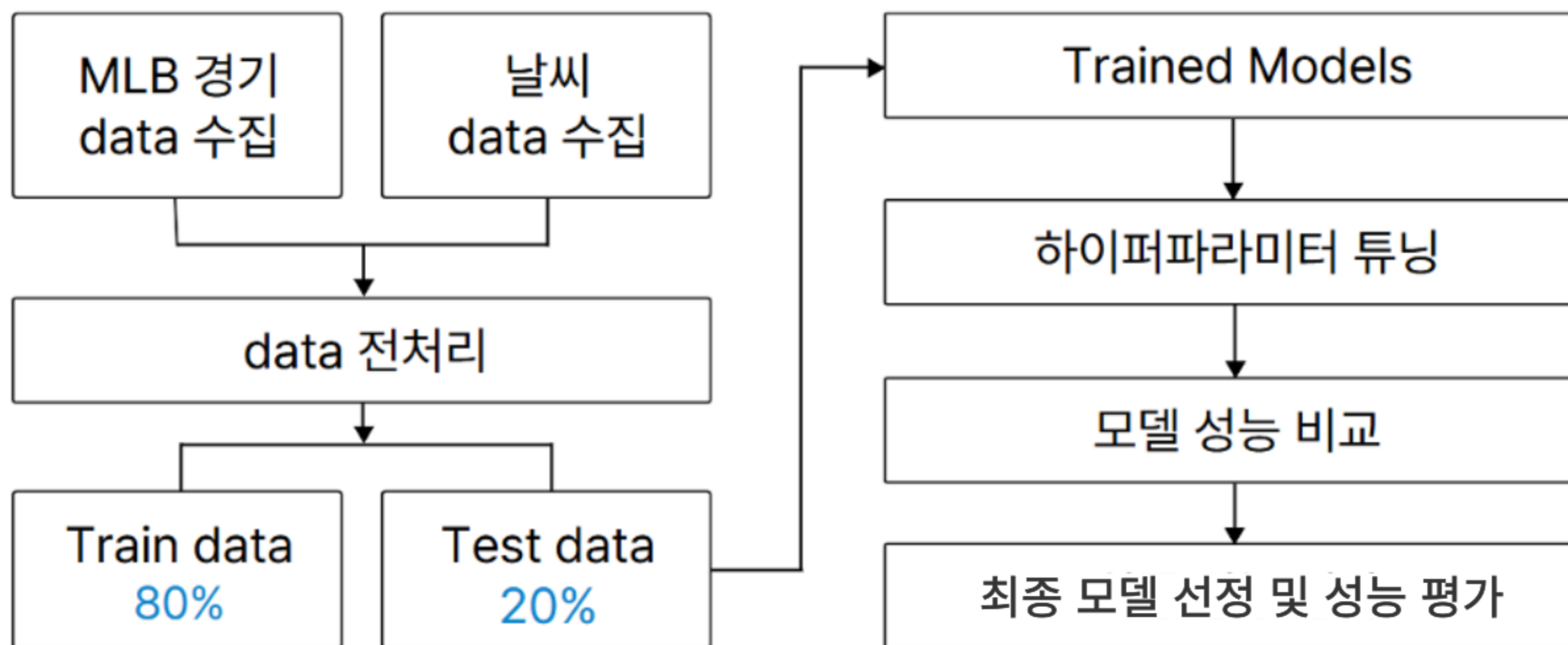
타구 데이터를 기준으로 타자와 날씨 정보를 merge 하여

» HitTrackMLB Dataset 구축

표 1. HitTrackMLB

역할	종류	변수명	의미
종속변수	타구 정보	cx	타구 x좌표
		cy	타구 y좌표
독립변수	선수 정보	player_age	타자 나이
		slg_percent	장타율
		isolated_power	순수 장타력
		babip	인플레이 타구 타율
		batter_hand	타자 타격 손
		pitcher_hand	투수 투구 손
	타구 정보	hit_dis	타구 비거리
		ball_type	타구 결과
		pitch_type	구종
		velocity	구속
		zone_num	투구 위치
		ball	볼
		strike	스트라이크
		exit_velocity	타구 속도
	기상 정보	temp	기온
		wind	풍속
		rain	강수량

Modeling & Evaluation Proposed Scheme



Modeling & Evaluation

Modeling

1

RandomForest

여러 결정트리를 앙상블해 예측하는 모델

2

XGBoost

오차 보정이 뛰어난 부스팅 모델

3

lightGBM

대용량 데이터에 최적화된 부스팅 모델

4

DNN

비선형 관계를 학습하는 딥러닝 기반 모델

Modeling Performance

모델	RMSE	MAPE(%)
RandomForest	46.3	16.2
XGBoost	42.4	13.4
lightGBM	44.2	14.2
DNN	49.8	18.3

>>> 최종 모델로 XGBoost 선정

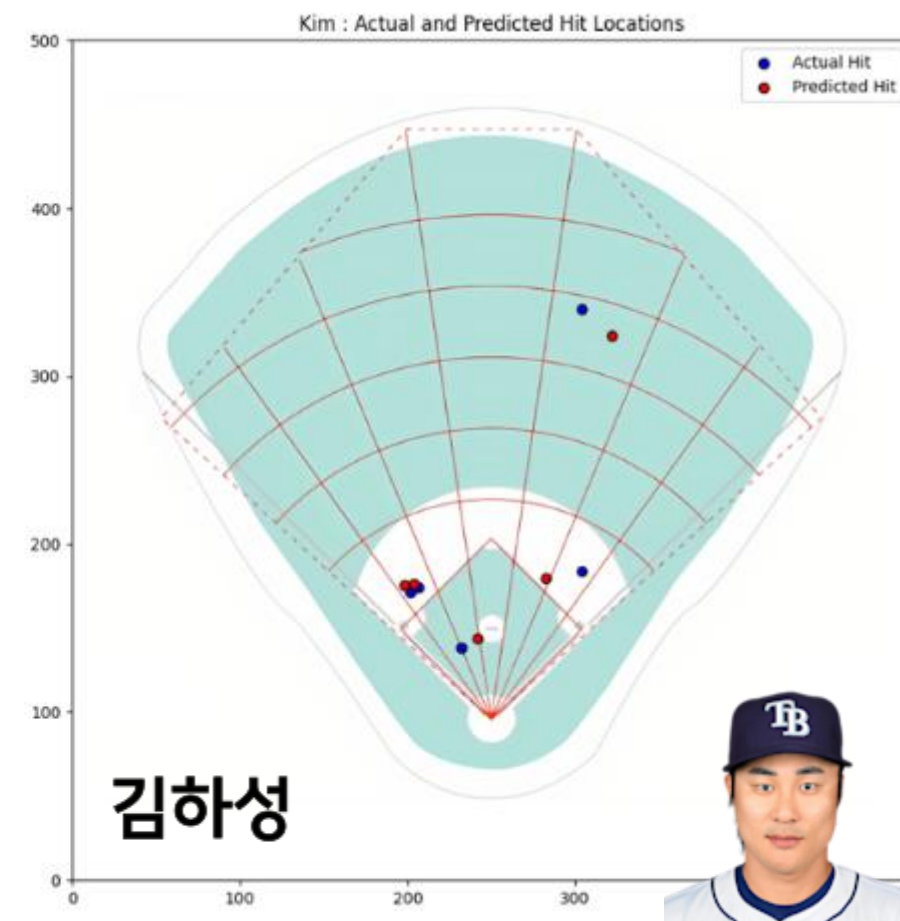
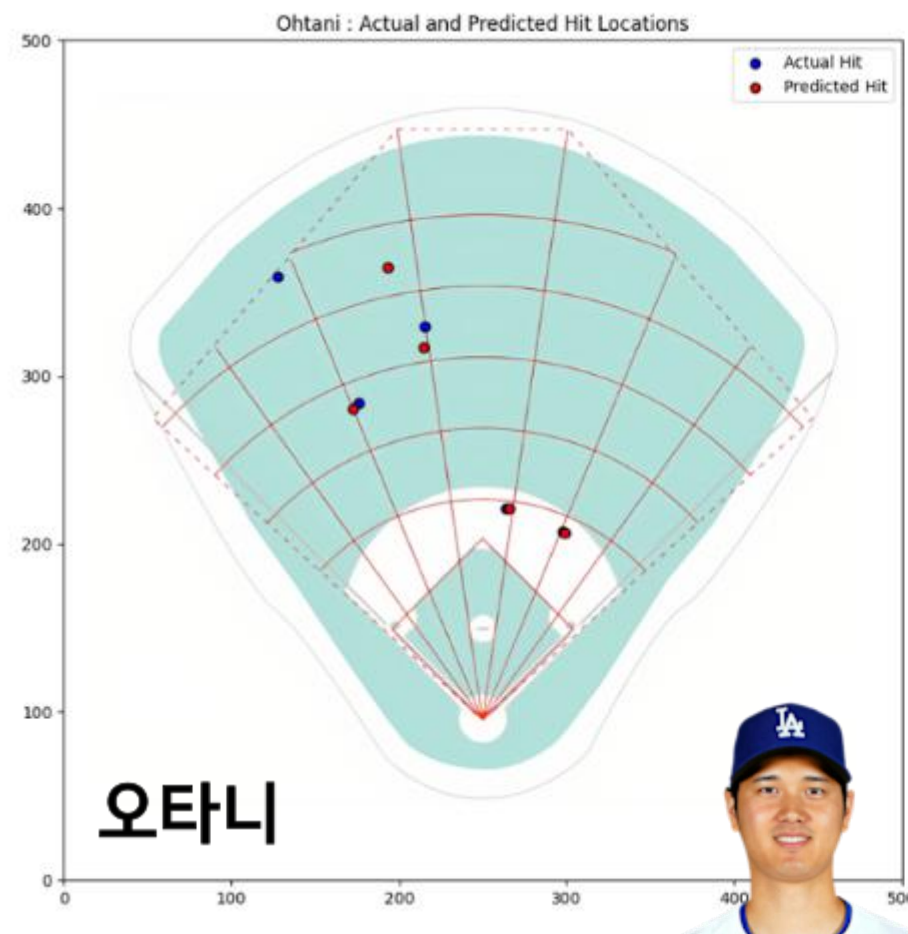
Modeling & Evaluation

Test Evaluation

예측 좌표와 실제 낙하지점 간의 평균 오차를 실제 거리로 환산 : **11.3 m**

mlb 선수들의 평균 sprint speed : **약 8 m/sec** [4]

➡ 2초 이내에 커버할 수 있는 거리(15~17m)와 비슷한 수준의 예측 오차



[4] MLB Advanced Media. Aaron Judge Statcast Hitting Data Internet].Available
: https://baseballsavant.mlb.com/leaderboard/sprint_speedmin_season=2024&max_season=2024&position=&team=&min=10

／ Conclusion

➤ 연구 의의

기존 연구들은 타구 결과 분석에 집중되어 있었으며, 실제 수비에 활용 가능한 ‘타구 위치 예측’은 거의 시도되지 않았다. 본 연구는 타구 정보뿐만 아니라 선수 특성, 경기 상황, 날씨 데이터까지 통합적으로 반영하여, 정확한 타구 낙하 지점을 회귀 방식으로 예측하는 모델을 새롭게 제안하였다.



➤ 연구 활용

선수의 반응 속도와 커버 범위를 고려할 때 실전 적용 가능성이 충분한 수준의 예측 정확도를 확보한 것으로 판단할 수 있다. 적절한 수비위치 선정으로 **실점을 방지하는데 기여**할 수 있다.



➤ 개선 방안

날씨 정보(특히 타구 시 풍량)에 대한 정보를 세부적으로 고려하지 못한 한계점은 남아 있어, 향후 연구에서 타구별 날씨 데이터를 반영하는 모델로 개선하고자 한다.

／

감사합니다

참고 - 구역 설정

“ UZR(Ultimate Zone Rating)

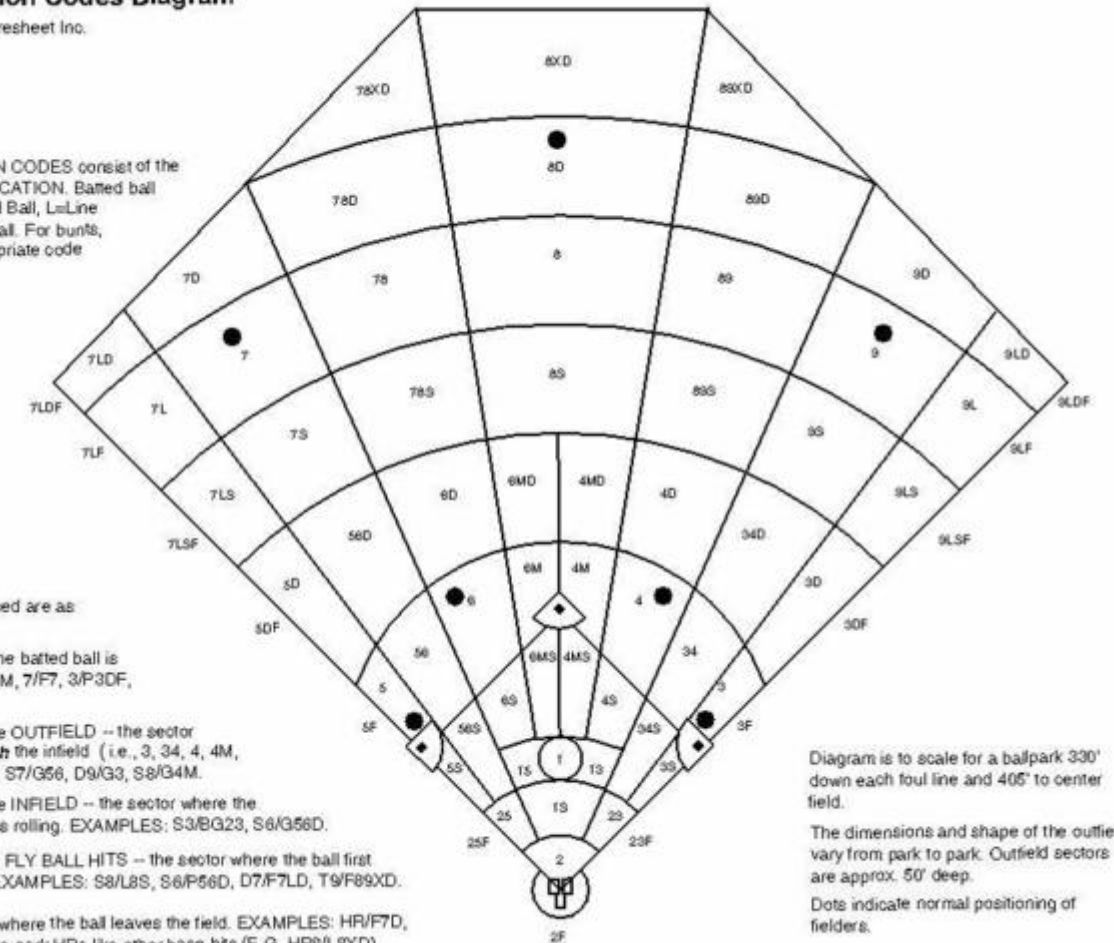
야구장을 여러 개의 작은 사각형 구역으로 나눠서 분석

내야수(INF)는 수비 포지션별 주요 플레이 범위를 기준으로 구역이 정해짐

외야수(OF)는 각 외야 구역을 작은 영역으로 세분화하여 분석

Project Scoresheet Scoring System
Batted Ball Location Codes Diagram
Copyright 1989 Project Scoresheet Inc.

BATTED BALL LOCATION CODES consist of the batted ball TYPE *plus* LOCATION. Batted ball type codes are: G=Ground Ball, L=Line Drive, P=Pop Fly, F=Fly Ball. For bunts, put a "B" before the appropriate code (e.g., "BG" or "BP").



The LOCATION CODES used are as follows:

OUTS -- the sector where the batted ball is fielded. EXAMPLES: 63/G6M, 7/F7, 3/P3DF, 8/L8S.

GROUND BALL HITS to the OUTFIELD -- the sector where the ball goes *through* the infield (i.e., 3, 34, 4, 4M, 6M, 6, 66 or 5). EXAMPLES: S7/G56, D9/G3, S8/G4M.

GROUND BALL HITS to the INFIEL -- the sector where the batted ball is fielded *or* stops rolling. EXAMPLES: S3/BG23, S6/G56D.

LINE DRIVE, POP FLY OR FLY BALL HITS -- the sector where the ball first drops (*not* where it rolls). EXAMPLES: S8/L8S, S6/P56D, D7/F7LD, T9/F89XD.

HOME RUNS -- the sector where the ball leaves the field. EXAMPLES: HR/F7D, HR/F89XD. Score inside-the-park HRs like other base hits (E.G. HR8/L8XD).

Diagram is to scale for a ballpark 330' down each foul line and 405' to center field.

The dimensions and shape of the outfield vary from park to park. Outfield sectors are approx. 50' deep.

Dots indicate normal positioning of fielders.

https://frhyme.github.io/baseball/baseball-eval_defence/

<https://baseballwithr.wordpress.com/2021/12/06/downloading-2021-retrosheet-data-and-batted-ball-locations/>

／ 참고

“구장별로 크기가 다른데 어떻게 했는지?

statcast에 나와있는 타구위치가 구장크기별로 상대적인 비율을 반영한 것이 아닌 절대적인 좌표로 찍히기 때문에 구장크기가 다른것은 큰 제한사항이 되지 않는다.
(구장이 크던 작던 이미지에서 찍히는 점의 위치가 동일)