2021 빅콘테스트

2021 챔피언리그 스포츠테크

KBO

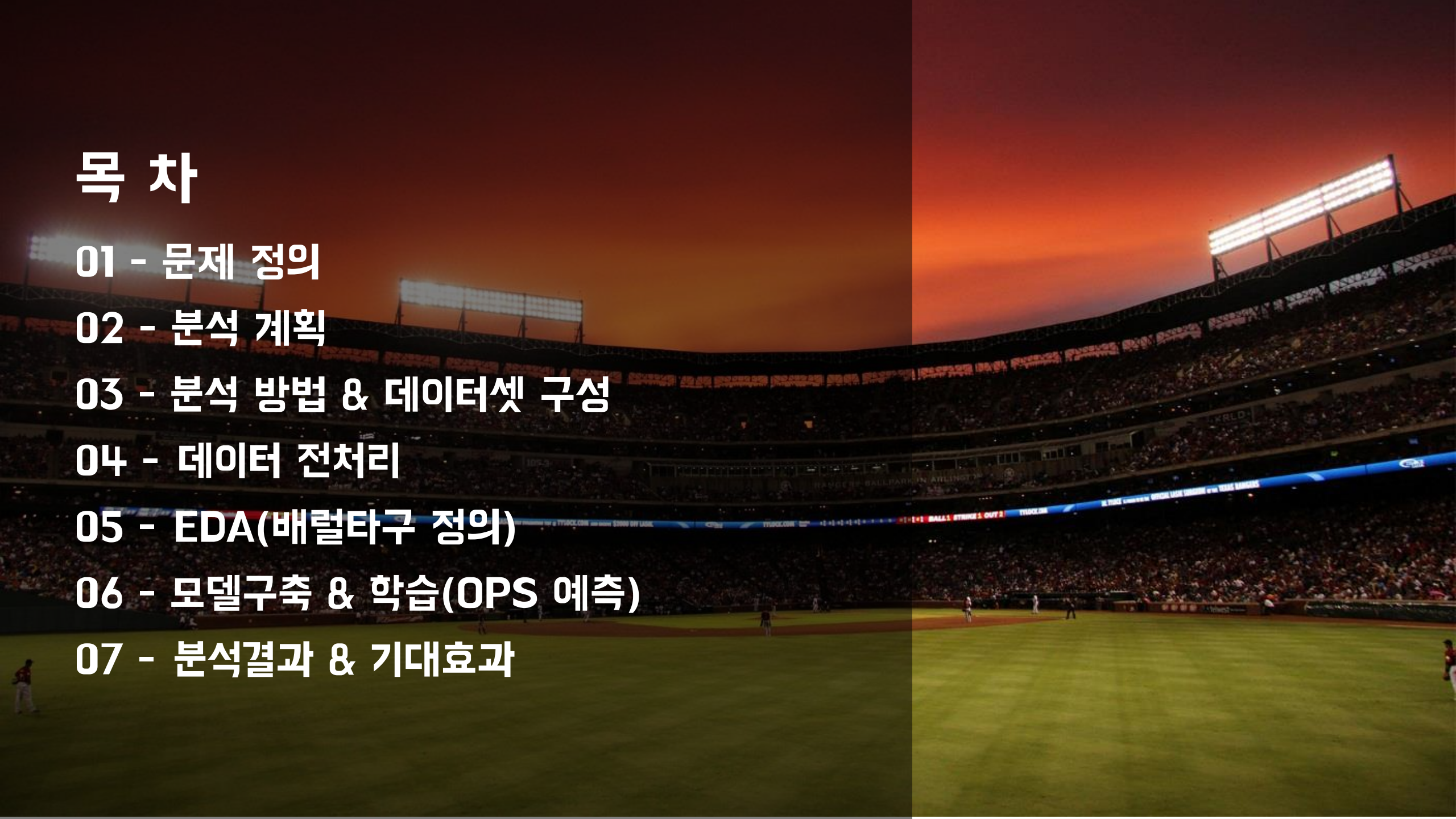
팀명: 패트와 매트

팀장: 최민석(0426minseok@naver.com)

팀원: 박승주(qkrtmdwn124@naver.com)

이성규(sknkdu@daum.net)

조문주(cmj0017@naver.com)



목 차

01 문제정의

01 문제정의

야구 = “기록의 스포츠”

MLB에서 중요한 타구의 질을 판단하는 지표인 배럴

But, KBO에서는 명확한 배럴기준 X.

-> **KBO만의 배럴기준** 정의 필요

OPS(장타율+출루율) 예측에 KBO만의 배럴을 활용한다면?

-> 더 정확한 예측 모델을 만들 수 있을 것으로 기대

02 분석계획

<배럴타구 정의>

활용 데이터	**스포츠투아이** (기본 제공데이터)
분석 기법	**Visualization & EDA** (plotly-scatterplot, boxplot)

02 분석계획

<OPS 예측>

sports2i Information & Interest Provider in Sports + STATIZ + 기상청

활용 데이터	스포츠투아이 (기본 제공데이터) + 스탯티즈 (크롤링 수집데이터) + 기상청 (다운로드 수집데이터)
분석 기법	- Selenium을 통한 데이터수집 - Scikit-learn 패키지를 이용한 ML(Machine Learning) - Tensorflow 패키지를 이용한 DL(Deep Learning)

03 분석방법 & 데이터셋 구성

<배럴타구 정의>

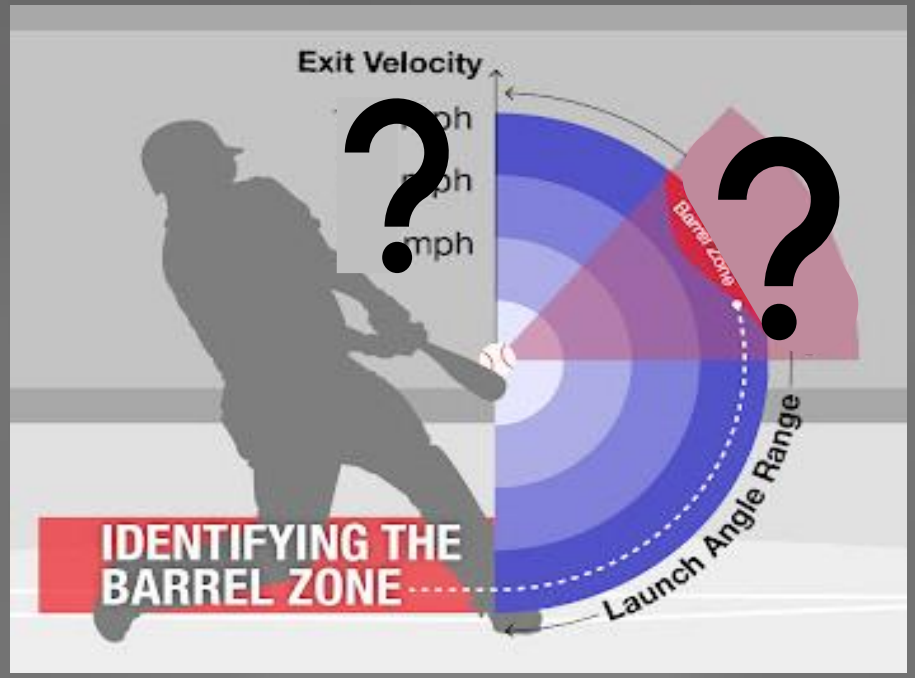
MLB와 KBO의 투수·타자의 능력, 구장 구격 차이 등을 고려

-> 기존 MLB 배럴 타구 기준인

0.5 이상의 타율과 1.5이상의 장타율은 유지

-> 발사각도와 타구속도 분석을 통해 (EDA)

KBO 만의 배럴타구 기준을 찾는 것을 목표



<OPS 예측> - 데이터셋 구성

제공데이터의 선수코드 예시

<선수코드: 67341>

BUT!

오마이뉴스 | 2021.06.09. | 네이버뉴스

'1할 타자' KIA **최형우**... 에이징 커브 직격탄?

타선 중심 **최형우** **부진** 탈출 간절 ▲ 타율 0.191-OPS 0.655로 부진한 KIA 최형우 © KIA 타이거즈 KIA 타이거즈가 3연패에 몰리며 최하위 추락 위기에 직면했다. ...

뉴스 기사와 댓글로 인한 문제 발생시 24시간 센터로 접수해주세요.

전북일보 | 2020.11.30.

전북출신 KIA **최형우**, 4년 만에 **타격왕**

최형우 선수 **최형우**는 지난 30일 서울 임피리얼 팰리스 호텔 그랜드볼룸에서 열린 '2020 신한은행 SOL KBO 시상식'에서 **타격왕**에 올랐다. **최형우**는 2020 시즌 140...

'37세 **타격왕**' **최형우**의 시계는 거꾸로 흐른다 무등일보 | 2020.11.30.
KIA **최형우**, 4년 만에 다시 **타격왕** "데뷔 늦... 뉴스1 | 2020.11.30. | 네이버뉴스
최형우 "**타격왕** 경쟁, 팬들께도 재미를... 마이데일리 | 2020.11.30. | 네이버뉴스
KIA **최형우**, 생애 두 번째 **타격왕** 광남일보 | 2020.11.30.

관련뉴스 8건 전체보기 >

<OPS 예측> - 데이터셋 구성

제공데이터 선수코드

선수코드: 67341

기사에서 보듯이, 2020년의 타격왕

최형우가 2021년 6월 9일기준 1할 타자?!?

<OPS 예측> - 데이터셋 구성

선수코드: 67341

년도 + 선수코드

Ex) 이정후를 1명의 타자로 보는 것이 아니라 각 년도별로 **다른 선수처럼 취급**

선수코드: 2018_67341

선수코드: 2019_67341

선수코드: 2020_67341

선수코드: 2021_67341

<OPS 예측> - 데이터셋 구성

	Train_X		Train_y	
2018년	제공데이터 (스포츠투아이)	수집데이터 (스탯티즈)	제공데이터 (스포츠투아이)	수집데이터 (스탯티즈)
2019년	각 년도별 약 383경기	각 년도별 약 465경기	각 년도별 마지막 105경기	각 년도별 마지막 105경기
2020년				
	Test_X		Test_y	
2021년	제공데이터 (스포츠투아이) 약 383경기	수집데이터 (스탯티즈) 약 465경기	2021년 9월 15일~10월 8일까지 105경기	

-Train set 만족기준-

규정타석 166인

Or

타석수 median

(109타석, 31타석)

- Test set 지정선수 10인-

<OPS 예측> - 분석방법(피쳐셋)

<스탯티즈>

주기별
타자의 성적 ...
(타율, 장타율, 출루율)

<세이버매트릭스>

BB% K%
PSN RC
BB/K XR
Woba IsoP
IsoD IC27
GPA BABIP
RawEqA WRAA
WRC ...

<날씨>

기온 관련
강수량 관련
.agg_dict ...

<스포츠투아이>

배럴타구 관련
안타 1루타
2루타 3루타
홈런 볼넷
도루 도실
희타 희비
사구 고의사구
타순 ...

외부 수집데이터 + 제공데이터

제공데이터

<OPS 예측> - 분석방법(피쳐셋)

출루율 + 장타율

↓

OPS (출루율 + 장타율)

"OPS 예측 모델 하나를 구축하는 것이 아니라
출루율 예측 모델과 장타율 예측 모델을
각각 구축하여 최종적으로 OPS를 산출하는 방법"

04 데이터 전처리

1. 컬럼명 변경

- hit_data.columns

["년도", "경기코드", "투구코드", "선수코드", "팀코드", "이닝", "타구속도","발사각도", "타구종류", "구속", "경기장"]

- player_data.columns

["년도", "선수코드", "이름", "팀코드", "포지션", "나이", "연봉"]

- batter_data.columns

["년도", "선수코드", "출장경기수", "타석", "타수", "타율", "안타", "홈런", "루타", "장타율", "희생플라이", "볼넷", "삼진", "고의사구", "사구", "병살타"]

2. 구장 병합

2019년 NC다이노스 홈구장 이전(마산야구장->창원NC파크)으로 '구장'컬럼의 '마산'과 '창원'의 일치화

3. 이상치 처리

- 1안

규정타석(경기수 *3.1)을 만족한 선수들에 한하여 분석을 진행

- 2안

Train_X (statiz data 기준 465 경기) 타자들의 타석수 median 값인 109타석,
Train_y (마지막 105경기) 타자들의 타석수 median 값인 31타석을 동시에
만족하는 선수들에 한하여 분석을 진행

4. 타구속도와 발사각도 구간 생성 - 배럴타구 정의를 위한 전처리

- 타구속도와 발사각도 값은 소수 둘째자리까지 나타나 있어 무수히 많은 고유값들이 존재하기 때문에, 타구 속도와 발사각도를 **2단위로 구간을 나눠** 분석에 용이하게 변환한다.

  ex) 타구속도가 148~149.9km이면서 타구각도가 24~26° 인 타구: 148

  타구속도가 150~151.9km이면서 타구각도가 24~32° 인 타구: 150

- 구간화한 타구속도와 발사각도를 문자열로 변환 후 결합 해 새로운 **'KEY'**열 생성한다.

  ex) 148/24, 148/26, 150/24, 150/26 ...

05 EDA

<배럴타구 정의>

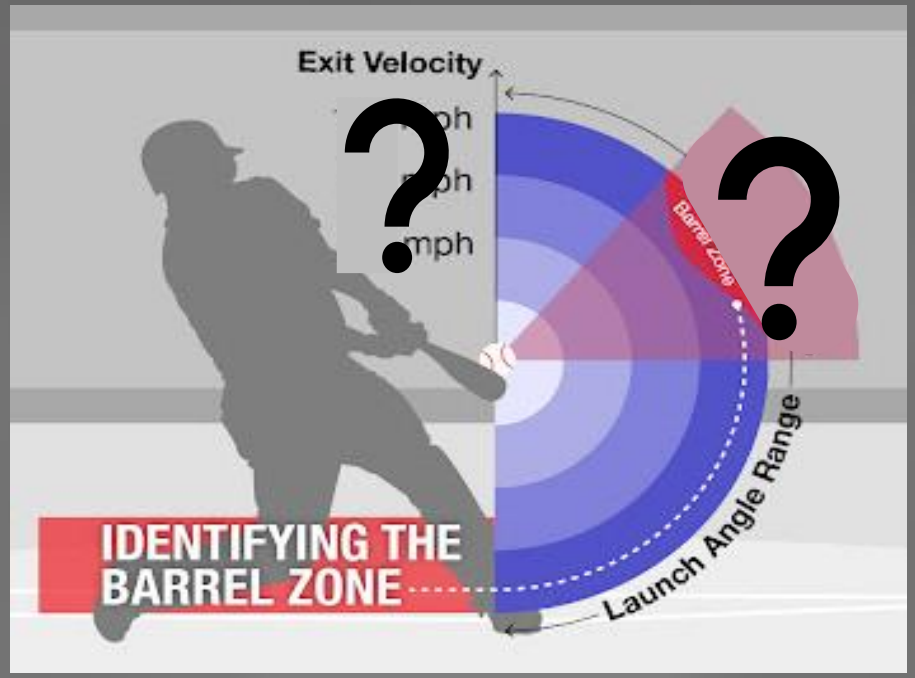
MLB와 KBO의 투수·타자의 능력, 구장 구격 차이 등을 고려

-> 기존 MLB 배럴 타구 기준인

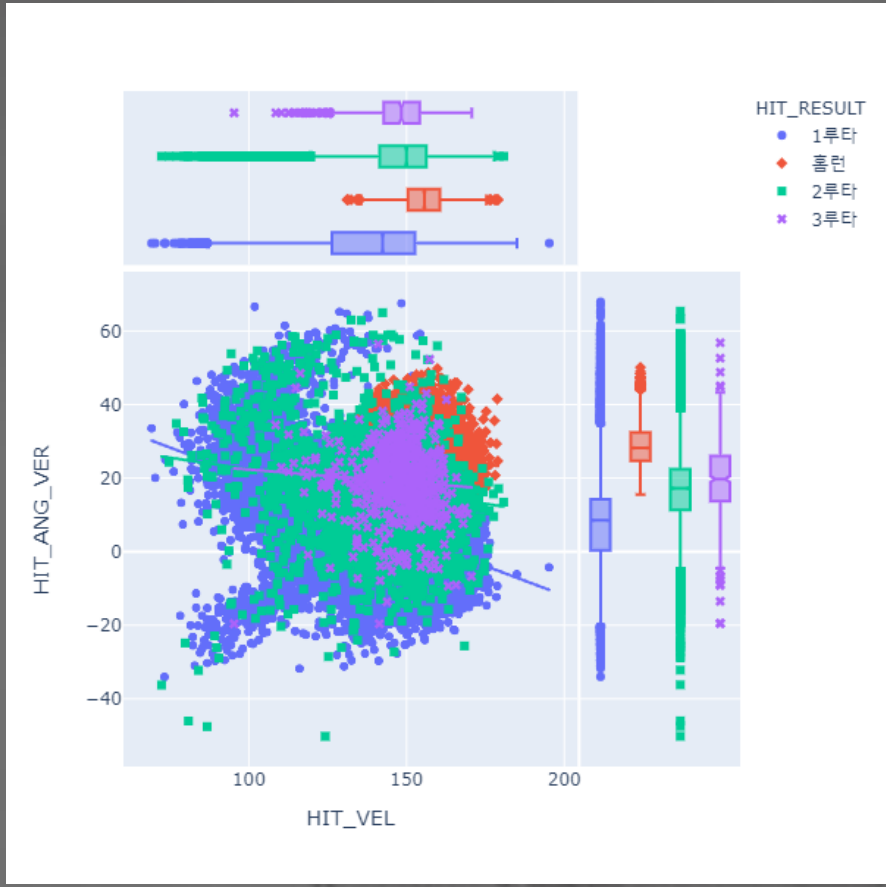
0.5 이상의 타율과 1.5이상의 장타율은 유지

-> 발사각도와 타구속도 분석을 통해 (EDA)

KBO 만의 배럴타구 기준을 찾는 것을 목표



타구결과가 [1루타, 2루타, 3루타 ,홈런]인 데이터로만 그린 산점도+박스플롯



[2루타]의 타구속도 중위값이 149.94, 발사각도 중위값이 17.2

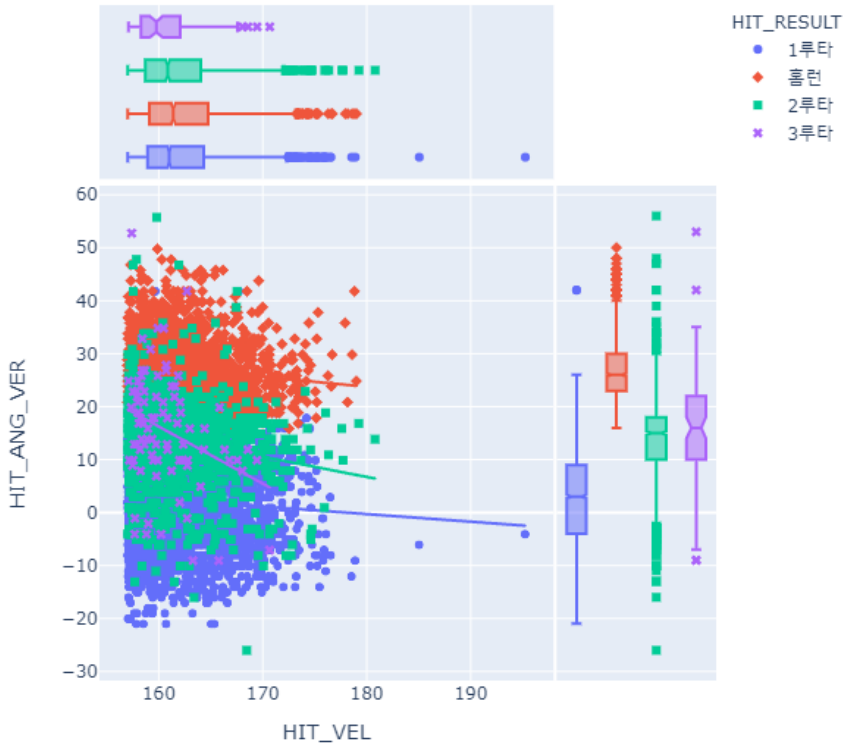
[3루타]의 타구속도 중위값이 148.385, 발사각도 중위값이 19.75

[홈런]의 타구속도 중위값이 155.62 발사각도 중위값이 28.2

- 홈런의 경우 MLB 배럴타구의 기준과 매우 유사한것을 확인.
- 2루타, 3루타의 경우 타구속도와 평균 발사각도가 MLB 배럴타구의 기준보다 낮은 수치임을 확인.

  -> KBO만의 배럴타구를 정의할 필요가 있음.

타구속도가 148 이상 (KBO 배럴타구속도의 시작점) 데이터의 산점도 + 박스 플롯



[2루타]의 발사각도 중위값이 15

[3루타]의 발사각도 중위값이 16

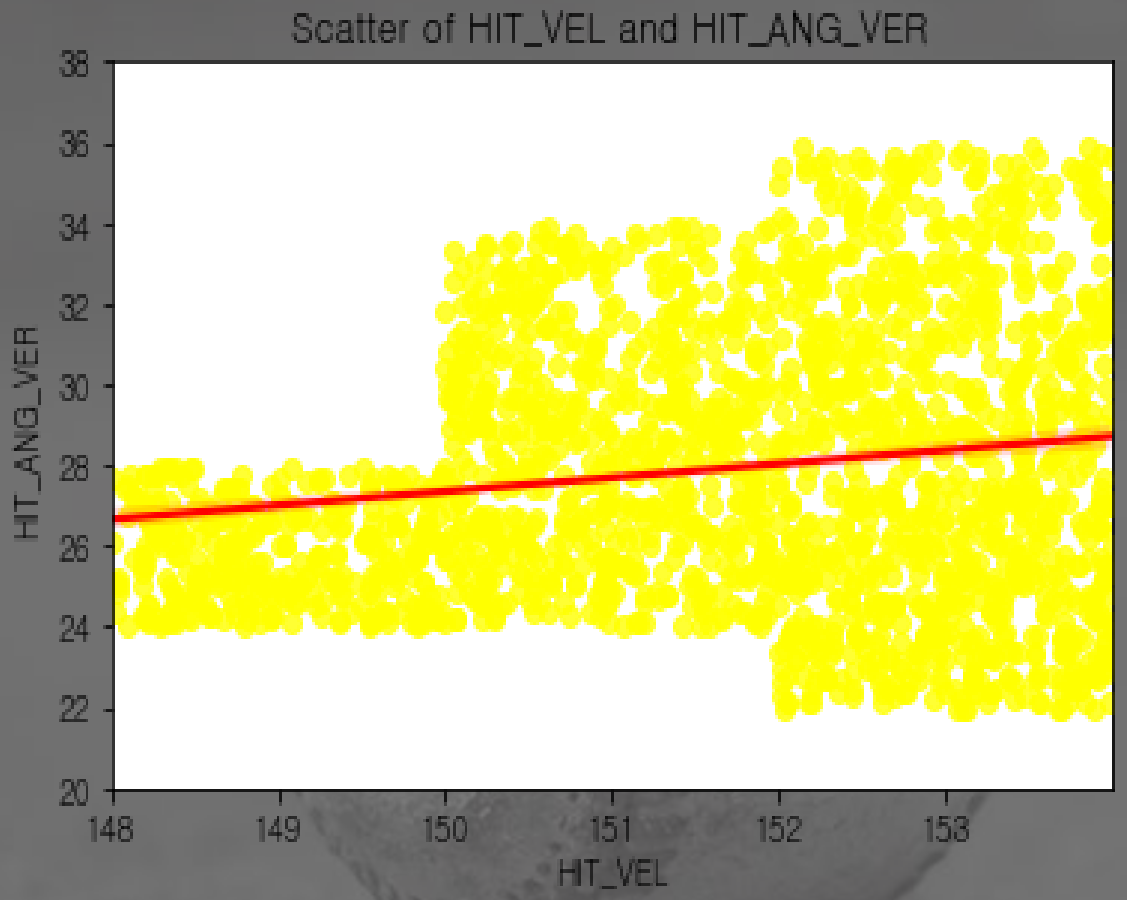
[홈런]의 발사각도 중위값이 26

- 타구속도가 148이상인 타구속도 구간에서,

  타구속도가 증가하면 발사각도가 넓어지는 것을 확인.

  -> 타구속도가 148이상인 데이터 위주로 탐색범위 지정.

전체 타구데이터에 'KEY'를 설정해 배럴 구간 탐색 - 1



1루타, 2루타, 3루타, 홈런이라는 결과를 나타낸 타구들을 안타로 정하고 0과 1로 라벨링을 한 **'안타라벨링'**이라는 새로운 열을 추가했다.

앞서 전처리 단계에서 만든 **'KEY'열**과 위의 **'안타 라벨링'열**로 안타가 가장 많이 나오는 구간의 타율과 장타율을 계산하고 **[타율 0.5이상, 장타율 1.500이상]인 구간**들을 확인했다.

배럴 타구 기준을 충족하는 구간의 범위를 줄여가면서 구간들을 세분화해 배럴타구 기준([타율 0.5이상, 장타율 1.500이상]인 구간)을 충족시키는 타구속도, 발사각도 범위를 찾아냈다.

전체 타구데이터에 'KEY'를 설정해 배럴 구간 탐색 - 2

	key	안타수	타구의수	타구결과의 합	타율	장타율
324	148/24	84	144	246	0.583333	1.708333
325	148/26	70	135	223	0.518519	1.651852
351	150/24	72	124	205	0.580645	1.653226
352	150/26	76	120	255	0.633333	2.125000
353	150/28	88	137	309	0.642336	2.255474
354	150/30	51	96	189	0.531250	1.968750
375	152/22	101	154	267	0.655844	1.733766
376	152/24	102	139	326	0.733813	2.345324
377	152/26	87	121	288	0.719008	2.380165
378	152/28	83	109	292	0.761468	2.678899
379	152/30	59	98	212	0.602041	2.163265
380	152/32	56	95	212	0.589474	2.231579
381	152/34	44	85	170	0.517647	2.000000
398	154/20	107	142	264	0.753521	1.859155
399	154/22	108	141	325	0.765957	2.304965
400	154/24	101	119	368	0.848739	3.092437
401	154/26	93	106	343	0.877358	3.235849
402	154/28	72	90	274	0.800000	3.044444
403	154/30	53	71	202	0.746479	2.845070
404	154/32	53	81	191	0.654321	2.358025
420	156/20	104	137	276	0.759124	2.014599
421	156/22	96	116	316	0.827586	2.724138
422	156/24	86	94	304	0.914894	3.234043
423	156/26	71	86	271	0.825581	3.151163
424	156/28	79	89	304	0.887640	3.415730
439	158/18	79	99	173	0.797980	1.747475
441	158/20	80	91	208	0.879121	2.285714
442	158/22	93	103	326	0.902913	3.165049
443	158/24	79	82	289	0.963415	3.524390
456	160/18	77	100	175	0.770000	1.750000
458	160/20	70	80	221	0.875000	2.762500

[표]에서 볼 수 있듯이, 타구속도가 148km일때 24~26°의 발사각도가 배럴타구이며, 이는 앞서 정의한 [타율 0.5, 장타율 1.5]를 만족하는 **최소 타구속도와 발사각도**이다.

타구속도가 148km이상일때 속도가 2km/h 증가하면 발사각도는 위로 +4° , 아래로 -2° 넓어진다.

전체 타구데이터에 'KEY'를 설정해 배럴 구간 탐색 - 3

key	안타수	타구의수	타구결과의 합	타율	장타율
160/26	42	43	167	0.976744	3.883721
160/30	24	25	94	0.960000	3.760000
162/24	38	38	148	1.000000	3.894737
162/26	44	44	173	1.000000	3.931818
162/30	23	24	92	0.958333	3.833333
162/38	7	7	28	1.000000	4.000000
164/22	28	28	110	1.000000	3.928571
164/26	25	25	100	1.000000	4.000000
164/28	21	21	84	1.000000	4.000000
164/30	14	14	56	1.000000	4.000000
164/32	9	9	36	1.000000	4.000000
164/34	6	6	24	1.000000	4.000000
166/24	26	26	104	1.000000	4.000000
166/26	22	22	88	1.000000	4.000000
166/28	11	11	44	1.000000	4.000000
166/32	7	7	28	1.000000	4.000000
166/34	3	3	12	1.000000	4.000000
166/46	1	1	4	1.000000	4.000000
168/20	16	16	60	1.000000	3.750000
168/22	12	12	46	1.000000	3.833333
168/24	10	10	40	1.000000	4.000000
168/26	13	13	52	1.000000	4.000000
168/28	3	3	12	1.000000	4.000000
168/30	8	8	32	1.000000	4.000000
168/32	8	8	32	1.000000	4.000000
168/44	1	1	4	1.000000	4.000000
170/22	8	8	32	1.000000	4.000000
170/24	9	9	36	1.000000	4.000000
170/26	4	4	16	1.000000	4.000000
170/28	5	5	20	1.000000	4.000000
170/30	3	3	12	1.000000	4.000000
170/34	2	2	8	1.000000	4.000000
170/36	1	1	4	1.000000	4.000000
172/18	2	2	8	1.000000	4.000000
172/20	3	3	12	1.000000	4.000000
172/22	4	4	16	1.000000	4.000000
172/24	6	6	24	1.000000	4.000000
172/26	2	2	8	1.000000	4.000000
172/28	5	5	20	1.000000	4.000000

특히, [표]와 같이 **타구속도가 160km이 넘어가고 발사각도가 8~48° 라면, 99% 이상 홈런**이 나오는 것을 확인하였다.

최종적으로 **배럴타구는 타구속도가 148km 이상일때 타구속도가 X(km/h)만큼 증가하면, 발사각도는 26°를 기준으로 위로 2*X° 아래로 -X°만큼 커지며, 이 발사각도 구간은 8~48 ° 안에서 정의된다.**

06 모델구축 & 학습 (OPS 예측)

- 장타율 예측 모델1 - ML

모델학습

1. 데이터분할: 학습데이터 80%, 평가데이터 20%

2. 사용 모델
'KNeighborsRegressor', 'LinearRegression',
'RandomForestRegressor', 'GradientBoostingRegressor',
'XGBRegressor', 'LGBMRegressor',
'AdaBoostRegressor', 'BaggingRegressor',
'CatBoostRegressor', 'SGDRegressor',
'ExtraTreesRegressor','DecisionTreeRegressor',

3. Random Search
- 평가지표 : RMSE
- cross validation : ShuffleSplit

앙상블

- Averaging
최고성능 LGBM+Linear Regressor

- Stacking 모델이 가장 좋은 성능을 보임
Linear Regression 기반 3개의 모델

- 장타율 예측 모델2 - DNN

모델학습

Stacking에 활용한 피쳐와 다른 피쳐셋으로 딥러닝 INPUT으로 활용.
Dropout을 활용해 과적합 방지.
최고성능의 모델을 찾기 위해 hyper_optimizer+ keras_tuner를 이용한 모델 학습
-> 모델에 parameter tuning 진행

- 출루율 예측 모델

모델학습

1. 데이터분할: 학습데이터 80%, 평가데이터 20%

2. 사용 모델

'KNeighborsRegressor', 'LinearRegression',
'RandomForestRegressor', 'GradientBoostingRegressor',
'XGBRegressor', 'LGBMRegressor',
'AdaBoostRegressor', 'BaggingRegressor',
'CatBoostRegressor', 'SGDRegressor',
'ExtraTreesRegressor','DecisionTreeRegressor',

3. Random Search

- 평가지표 : RMSE
- cross validation : ShuffleSplit

앙상블

- Stacking 모델이 가장 좋은 성능을 보임
  Linear Regression 기반 3개의 모델

07 분석결과 & 기대효과

07 분석결과

- Target 선수 10人의 OPS 예측

선수이름(선수코드)	장타율	출루율	OPS
이정후 (67341)	0.546772	0.399899	0.946671
로맥 (67872)	0.426737	0.336168	0.762905
강백호 (68050)	0.627434	0.391407	1.018841
최정 (75847)	0.472440	0.372368	0.844808
양의지 (76232)	0.563350	0.401204	0.964554
김현수 (76290)	0.451548	0.376581	0.828129
김재환 (78224)	0.505879	0.380917	0.886796
전준우 (78513)	0.468605	0.382125	0.850730
채은성 (79192)	0.506992	0.385475	0.892467
박건우 (79215)	0.492165	0.373432	0.865597

출루율 + 장타율
= OPS

07 기대효과

"KBO만의 배럴타구 기준과 OPS 예측 모델이 불러오는 기대효과"

1. 새로운 훈련 & 전략 도출

- 타자의 경우, 배럴타구를 만들기 위한 훈련 진행 가능.
- 투수의 경우, 타자별로 어떤 투구가 배럴타구로 연결되는지 상관관계 분석을 통한 훈련 진행 가능.
- 팀 차원의 경우, 상대선수에 대한 분석적인 전략 도출 가능.

  -> 이런 훈련들과 도출된 전략으로 경기 결과에 순영향.

2. 배럴 지표화

- 기존 출루율, 장타율, 홈런 등의 성적으로 편성해왔던 **타순 편성**에 배럴 지표와 OPS 예측 모델을 추가로 활용 가능.
- 경기 외적으로 배럴타구율이 높은 타자가 타석에 들어서면, 긴장감과 기대감을 높여 **관객 흥미도**에 기여 가능.
- **선수 계약**에 있어서 배럴 지표도 고려가 가능해 더 합리적인 계약 도출 가능.


Coca-Cola
VOLVO
Bank of America
10 SHEFFIELD
2009 WITH RISP
AVG AB H HR
.250 28 7 0
COVIDIEN
STANLEY
MAKE SOMETHING GREAT
GRANITE CITY
THANK YOU!