

프로야구 배럴을 통한 타자 성적 예측

2021 빅콘테스트 챔피언리그 스포츠테크

팀명: 운수영원

팀장: 박재운 (ui8289@naver.com)

팀원: 곽수연 (sykwak1110@gmail.com)

팀원: 최선영 (cpa1563@naver.com)

팀원: 곽희원 (hewo1217@gmail.com)

INDEX

01 문제 정의 분석 배경

우리의 과제와 방향

02 데이터 소개

외부 데이터 설명

03 배럴 정의

KBO 배럴 정의와 도출 과정

04 데이터 분석

데이터 전처리

05 모델 구축

N개의 모델 비교 및 분석

06 OPS 예측

10명의 선수 9월~10월 OPS

07 활용방안

실질적인 적용방안



문제 정의 분석 배경

01 문제 정의 분석 배경

우리의 과제



KBO 시장에 맞는 배럴 정의하기



10명의 선수 OPS 예측하기

01 문제 정의 분석 배경

우리의 과제



KBO 시장에 맞는 배럴 정의하기

2015년 MLB에 스탯캐스트 시스템 도입
'좋은 타구'로 분류할 수 있는 객관적인 기준 = 배럴
MLB에서는 선수 평가시 의미 있는 지표로 활용

10명의 선수 OPS 예측하기

01 문제 정의 분석 배경

우리의 과제

다양한 요인으로 인해 MLB 배럴 기준과 부적합

우리나라에도 객관적인 지표의 필요성 요구

KBO 타자들의 OPS 예측 모델 구축

KBO 시장에 맞는 배럴 정의하기



10명의 선수 OPS 예측하기



데이터 소개

02 데이터 소개

제공데이터 외부데이터 분석목표

1. 팀 데이터

SEASON_ID	T_ID	T_NM
2016	HH	한화
2016	HT	KIA
2016	KT	KT
2016	LG	LG

2. 선수 데이터(2018~2021)

GYEAR	PCODE	NAME	T_ID	POSITION	AGE_VA	MONEY
2021	50030	소형준	KT	투	19	14000만원
2021	50036	이강준	KT	투	19	3000만원
2021	50040	데스파이네	KT	투	34	500000달러
2021	50054	천성호	KT	내	23	4000만원

3. 타자 기본 데이터(2018~2021)

GYEAR	PCODE	GAMENUM	PA	AB	BA	HIT	HR	TOTB	SLG	SF	BB	KK	IB	HP	GD
2021	50054	25	28	25	0.2	5	0	5	0.2	0	1	5	0	2	0
2021	50150	7	12	10	0.1	1	0	1	0.1	0	1	2	0	1	0
2021	50165	51	205	185	0.243	45	8	78	0.422	0	18	43	0	2	2
2021	50167	14	19	16	0.125	2	0	2	0.125	0	1	7	0	1	0
		타석	타수	타율	안타	홈런	루타	장타율	희생 플라이	볼넷	삼진	고의4구	사구	병살타	

02 데이터 소개

제공데이터 외부데이터 분석목표

4. 경기 데이터

G_ID	GDAY_DS	VISIT_KEY	HOME_KEY	HEADER_NO	GWEEK	STADIUM	ACG
20210403SSWO0	20210403	SS	WO	0	토	고척	1
20210404HHKT0	20210404	HH	KT	0	일	수원	1
20210404HTOB0	20210404	HT	OB	0	일	잠실	1
20210404LGNC0	20210404	LG	NC	0	일	창원	1

5. HIT 데이터

GYEAR	G_ID	PIT_ID	PCODE	T_ID	INN	HIT_VEL	HIT_ANG_VER	HIT_RESULT	PIT_VEL	STADIUM
2021	20210403SSWO0	210403_140101	62415	SS	1	131.7	-5.8	땅볼아웃	144.35	고척
2021	20210403SSWO0	210403_140857	74163	WO	1	116.87	18.4	1루타	132.34	고척
2021	20210403SSWO0	210403_141459	75125	WO	1	160	16.8	2루타	120.78	고척
2021	20210403SSWO0	210403_142105	51463	SS	2	160.37	35.2	플라이	142.66	고척
					이닝	타구속도 (km/h)	발사각도	타격결과	상대 투수 투구속도 (km/h)	

02 데이터 소개

제공데이터 외부데이터 분석목표

1. MLB 선수 데이터

player_id	year	player_age	b_total_pa	b_single	b_double	b_triple	b_home_run	batting_avg	slg_percent	on_base_percent	on_base_plus_slg	exit_velocity_avg	launch_angle_avg	barrel_batted_rate
405395	2020	40	163	20	8	0	6	0.224	0.395	0.27	0.665	88.6	16.4	5.5
408234	2020	37	231	37	4	0	10	0.25	0.417	0.329	0.746	93.2	12.1	9.7
425772	2020	37	68	5	1	1	3	0.161	0.355	0.221	0.575	89.6	20.6	7.7
425783	2020	38	127	18	3	0	5	0.236	0.4	0.323	0.723	90	11.4	10.1
			타석	1루타	2루타	3루타	홈런	평균타구속도	장타율	출루율	OPS	평균타구속도	평균발사각도	배럴률

2. 날씨 데이터

지역명	날짜	평균기온	최저기온	최고기온	평균풍속	최다풍향	평균상대습도	평균증기압	평균현지기압	평균지면온도	경기장
광주	2018-01-01	2.3	-1	7.1	1.6	20	52.6	3.7	1019.1	1.3	광주
광주	2018-01-02	2.2	-2.9	8.4	0.9	20	67	4.8	1020	-0.3	광주
광주	2018-01-03	0.2	-2.5	3.8	1.7	20	52.6	3.2	1021.5	-0.3	광주
광주	2018-01-04	-0.6	-2.3	2.1	1.3	50	54.5	3.2	1017.5	-0.6	광주

02 데이터 소개

제공데이터 외부데이터 분석목표

3. 시계열 데이터

월별	선수명	타수	안타	1루타	2루타	3루타	홈런	사사구	희생플라이
201803	강백호	27	10	4	2	0	4	3	0
201803	김재환	25	5	3	0	0	2	3	1
201803	김현수	29	7	4	2	0	1	3	0
201803	로맥	26	10	7	0	0	3	5	0

02 데이터 소개

제공데이터 외부데이터 분석목표

KBO에 적합한 배럴 정의

- 회귀분석을 통한 OPS와 배럴의 통계적 유의성 확인
- MLB 경기 데이터와 KBO 경기 데이터 비교
- 제공 데이터로부터 파생 변수를 생성하여 배럴 정의

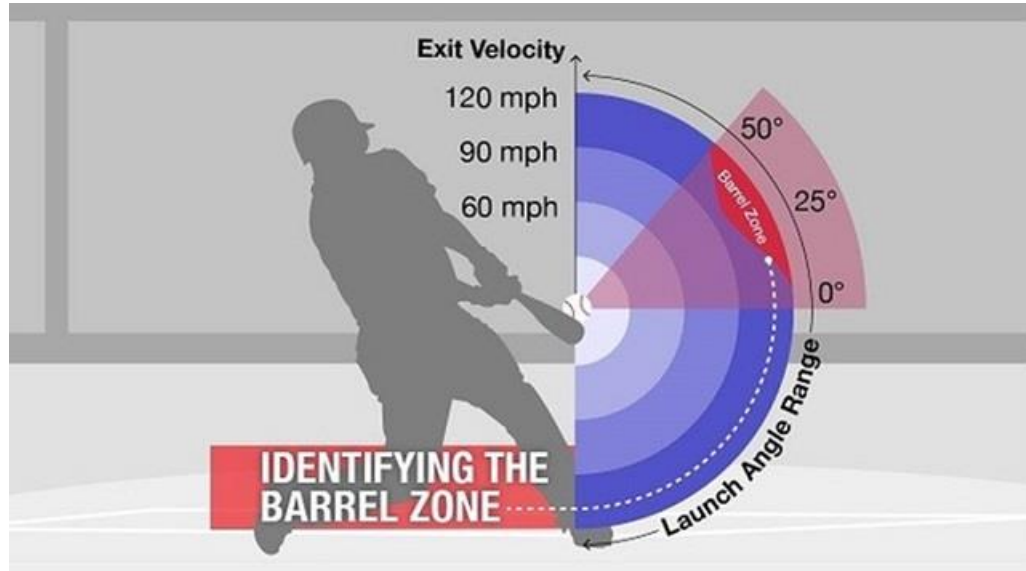
KBO 선수들의 OPS 예측

- 시계열 데이터로 가공한 파생변수를 활용한 OPS예측
- 벡터 자기 회귀(VAR)를 이용한 다변량 시계열 예측 수행
- 우리가 정의한 배럴을 활용한 OPS예측



배럴 정의

03 배럴 정의



“ 배럴(Barrel)이란 타율 0.5, 장타율 1.5이상을 생산하는 타구로 최소 시속 98마일을 기록해야 한다. ”

03 배럴 정의



MLB 데이터를 통해
배럴과 OPS의 연관성을 알아보자

OLS Regression

ordinary least square

선형회귀모델에서 알려지지 않은 파라미터를 추정하기 위한 선형 최소 자승법의 한 종류

1. 사용 목적

회귀분석을 통해 독립변수와 종속변수 사이의 연관성을 검증하기 위함

2. 해석 방법

Coef (회귀계수)

데이터로부터 얻은 계수의 추정치

P-value

귀무 가설이 맞다는 전제 하에, 표본에서 실제로 관측된 통계치와 같거나 더 극단적인 통계치가 관측될 확률

R-squared (결정계수)

추정한 선형 모형이 주어진 자료에 적합한 정도를 재는 척도로, 종속 변수의 변동량 중에서 적용한 모형으로 설명 가능한 부분의 비율을 가리킴

OLS 회귀분석을 사용한 이유

03 배럴 정의

OLS Regression를 통해 배럴의 당위성 확인

OLS Regression Results						
Dep. Variable:	OPS	R-squared:	0.335			
Model:	OLS	Adj. R-squared:	0.334			
Method:	Least Squares	F-statistic:	1032.			
Date:	Mon, 06 Sep 2021	Prob (F-statistic):	8.74e-184			
Time:	22:31:41	Log-Likelihood:	1316.0			
No. Observations:	2053	AIC:	-2628.			
Df Residuals:	2051	BIC:	-2617.			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
intercept	0.5559	0.005	107.172	0.000	0.546	0.566
배럴률	0.0202	0.001	32.127	0.000	0.019	0.021
Omnibus:	263.511	Durbin-Watson:	2.020			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	433.661			
Skew:	-0.868	Prob(JB):	6.79e-95			
Kurtosis:	4.434	Cond. No.	15.4			

모형적합도

'OPS'에 대하여 '배럴률'로 예측하는 회귀분석을 실시한 결과,
이 회귀모형은 통계적으로 유의미하였다.
($F(1,2051) = 1032, p < 0.05$)

독립변수

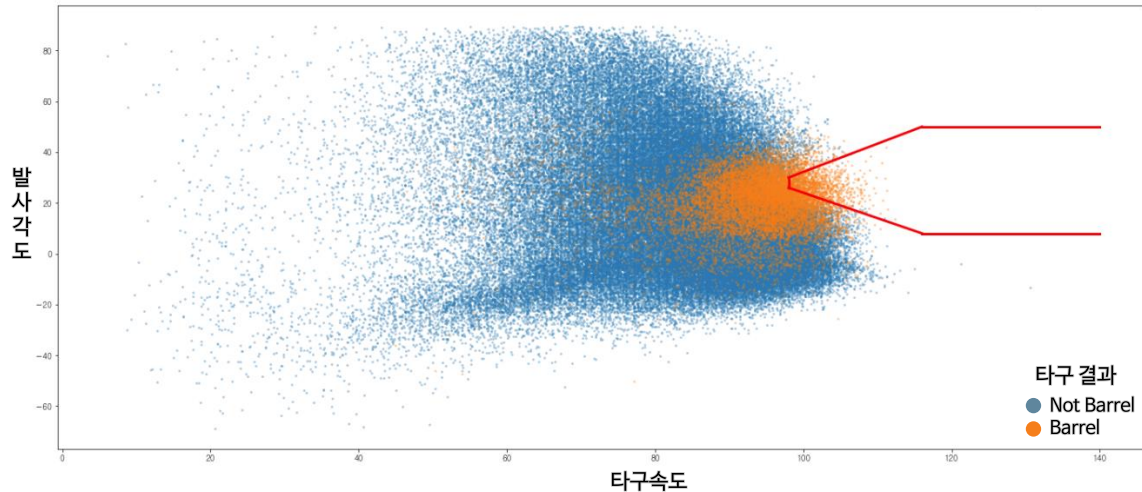
'배럴률'의 회귀계수는 0.0202로 'OPS'에 대하여
유의미한 예측변인인 것으로 나타났다.
($t(2051) = 32.127, p < 0.05$)

현재 MLB의 배럴 기준으로 구한 배럴이
OPS를 구하는데에 미치는 영향도를 확인할 수 있다.

03 배럴 정의

MLB의 배럴 기준과 KBO

KBO 배럴 타구 분포도 및 MLB 배럴 기준 적용



현재 MLB의 배럴 기준을 KBO 타구 분포도에 적용 해보았을 때
배럴에 부합하는 타구는 매우 적은 범위에 포함된다.

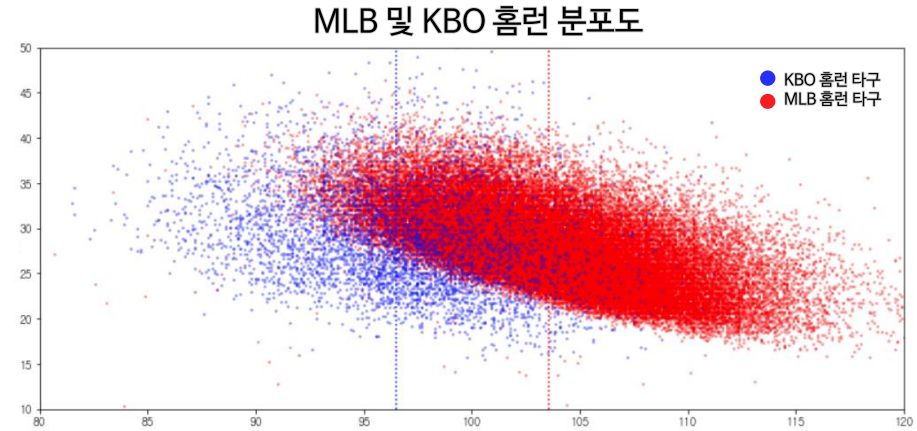
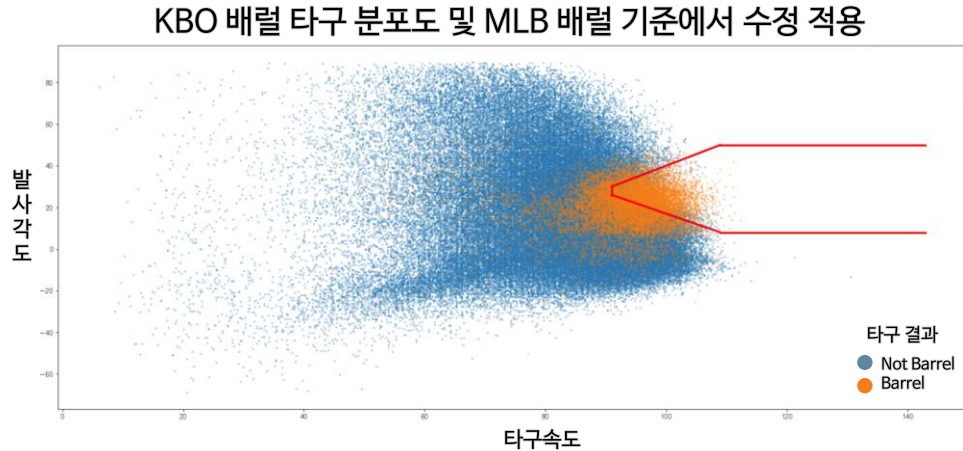


KBO에 적합한 배럴 기준을 확립해야 할 필요성이 있다.

배럴 분포도를 통해 MLB의 배럴 기준이
KBO에 적합하지 않다는 것을 확인할 수 있다.

03 배럴 정의

MLB의 배럴 기준과 KBO



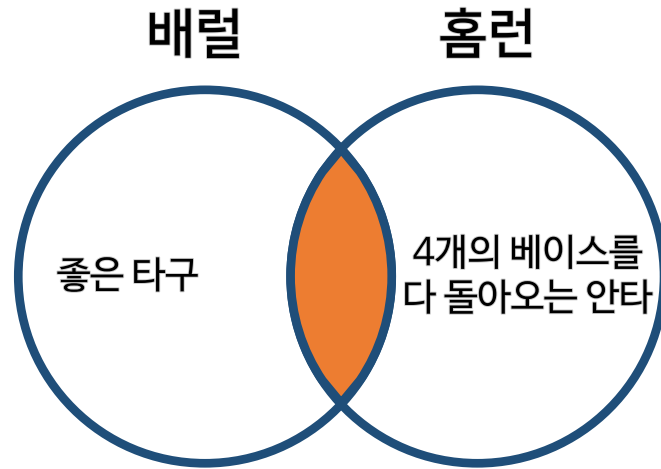
MLB와 KBO의 홈런 평균 타구속도의 차이가 **7차이**가 나는 것을 확인 할 수 있다.

따라서 비교적 많은 배럴 타구를 분류하는 선을 그을 수 있다.

정확도도 91%로 높은 수치를 가진다.

배럴 분포도를 통해 MLB의 배럴 기준이
KBO에 적합하지 않다는 것을 확인할 수 있다.

03 배럴 정의

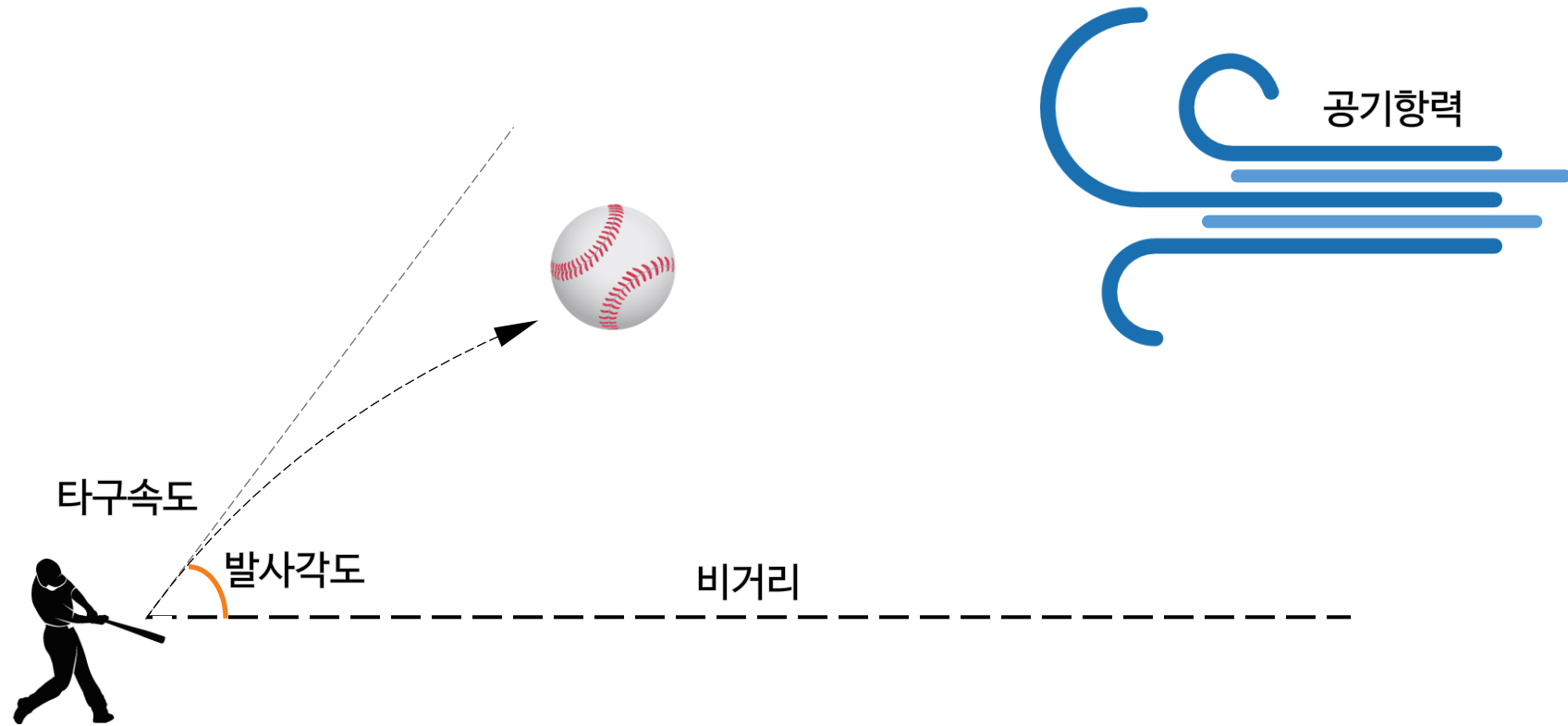


홈런의 긴 **비거리**를 착안해서
배럴을 정의할 때도 **비거리**를 고려해볼 것이다!

우리가 배럴을 정의할 때
비거리를 고려한 이유

03 배럴 정의

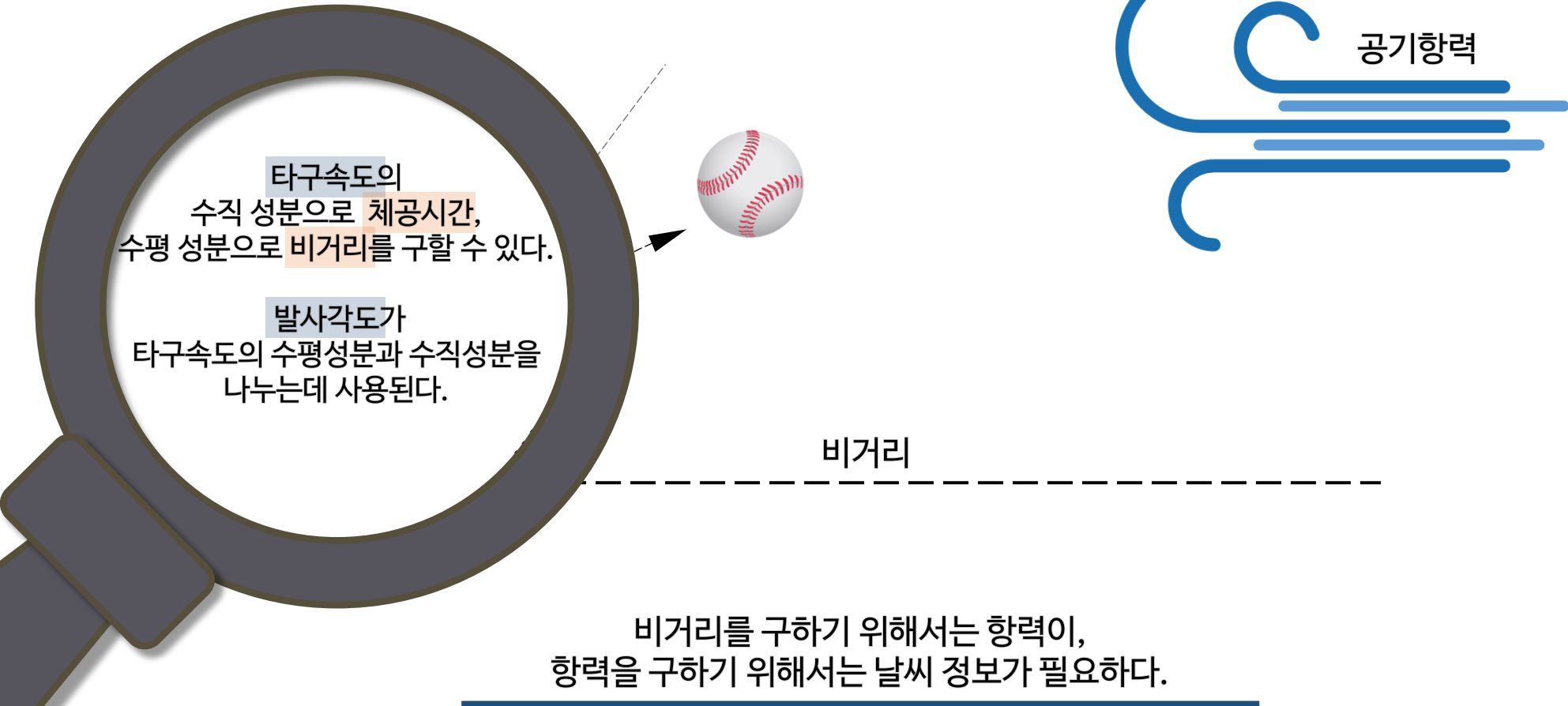
날씨 데이터로 알아보는 비거리



비거리를 구하기 위해서는 항력이,
항력을 구하기 위해서는 날씨 정보가 필요하다.

03 배럴 정의

날씨 데이터로 알아보는 비거리



타구속도의
수직 성분으로 체공시간,
수평 성분으로 비거리를 구할 수 있다.

발사각도가
타구속도의 수평성분과 수직성분을
나누는데 사용된다.

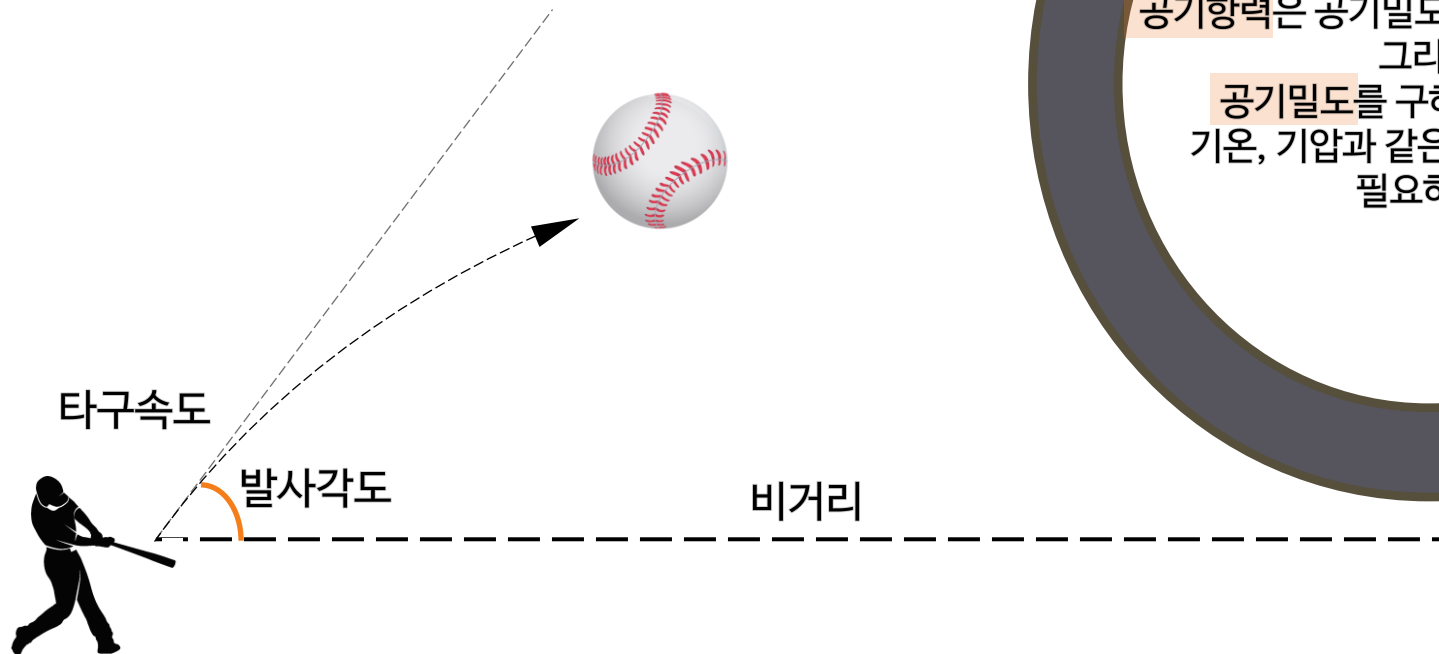
공기항력

비거리

비거리를 구하기 위해서는 항력이,
항력을 구하기 위해서는 날씨 정보가 필요하다.

03 배럴 정의

날씨 데이터로 알아보는 비거리



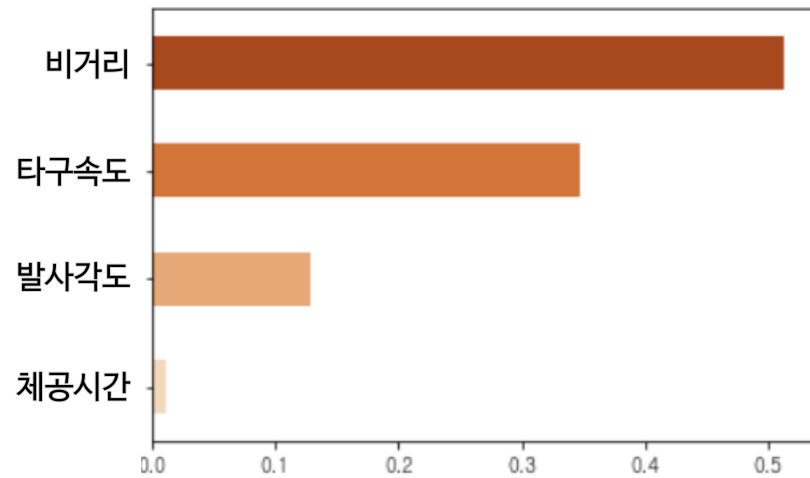
공기항력은 공기밀도의 영향을 받는다.
그리고
공기밀도를 구하기 위해서는
기온, 기압과 같은 날씨 데이터가
필요하다.

비거리를 구하기 위해서는 항력이,
항력을 구하기 위해서는 날씨 정보가 필요하다.

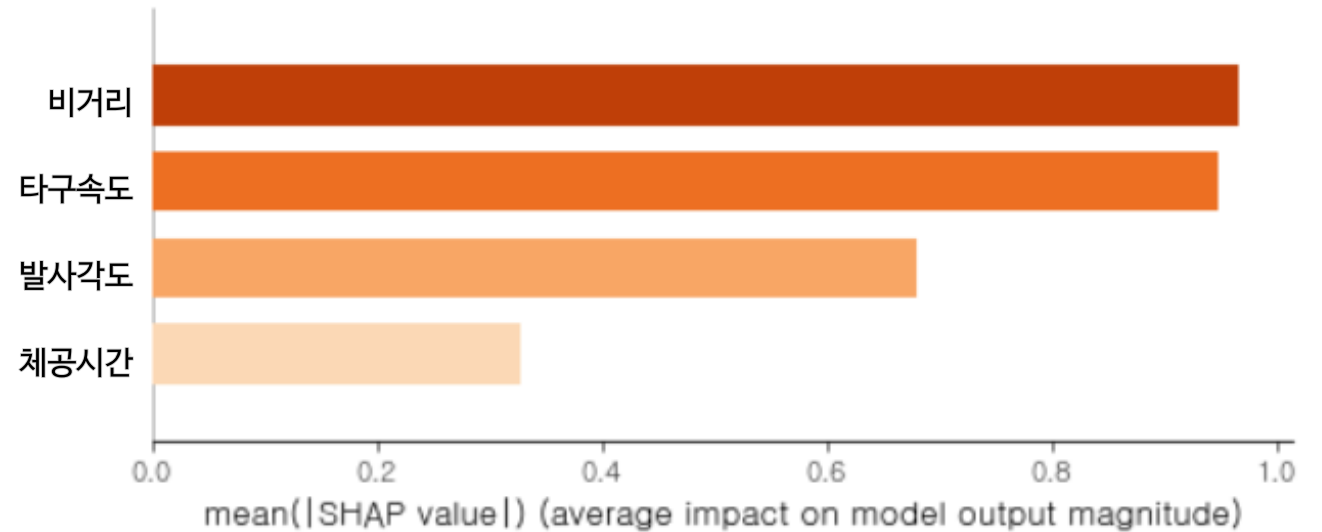
03 배럴 정의

장타(배럴 타구)와 비거리의 상관관계

1. Decision tree



2. SHAP value



특성중요도와 SHAP 두 개의 지표를 통해
비거리가 가장 중요한 특성임을 알 수 있다.

03 배럴 정의

배럴을 정의하는 과정

1. 모델링

SGDC 선형 분류에 비거리와 체공시간을 독립변수로 사용

2. 모델 결과 해석

비거리-체공시간 vs 타구속도-발사각도

3. 배럴 정의

모델 결과를 수식으로 변환하여 우리의 배럴 정의

SGDClassifier

stochastic gradient descent
확률적 기울기 하강법 학습으로 정규화된 선형 모델

1. 사용 목적

배럴타구와 배럴 타구 아닌 것을 이진 분류 하기 위해 사용

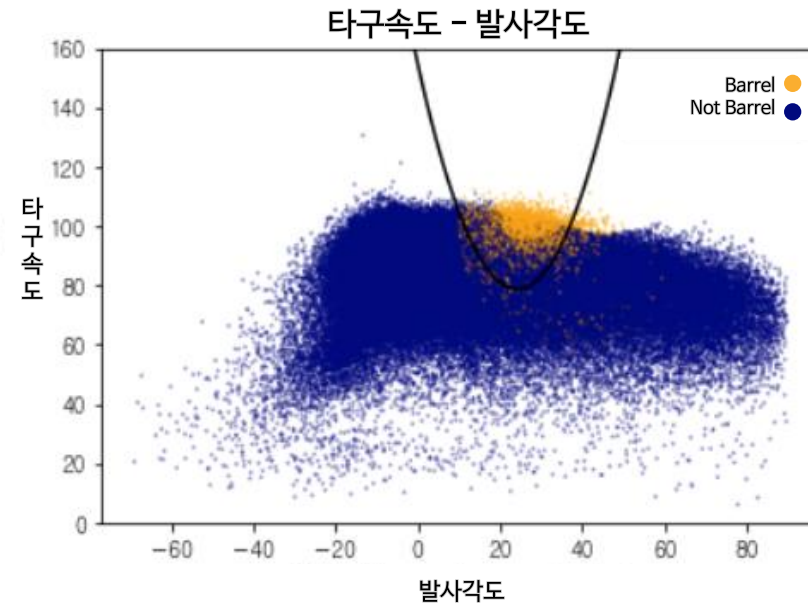
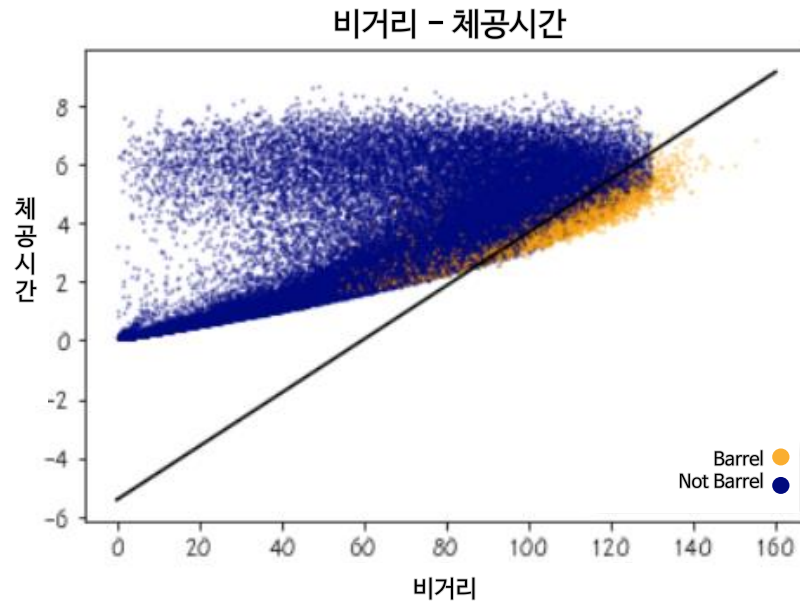
2. 특징

훈련 데이터를 하나씩 독립적으로 처리하기 때문에 큰 데이터 셋을 처리하기에 좋음
분류 기준을 하나의 수식으로 표현할 수 있음

배럴 정의할 때 Decision Tree를 사용한 이유

03 배럴 정의

종속변수 비교

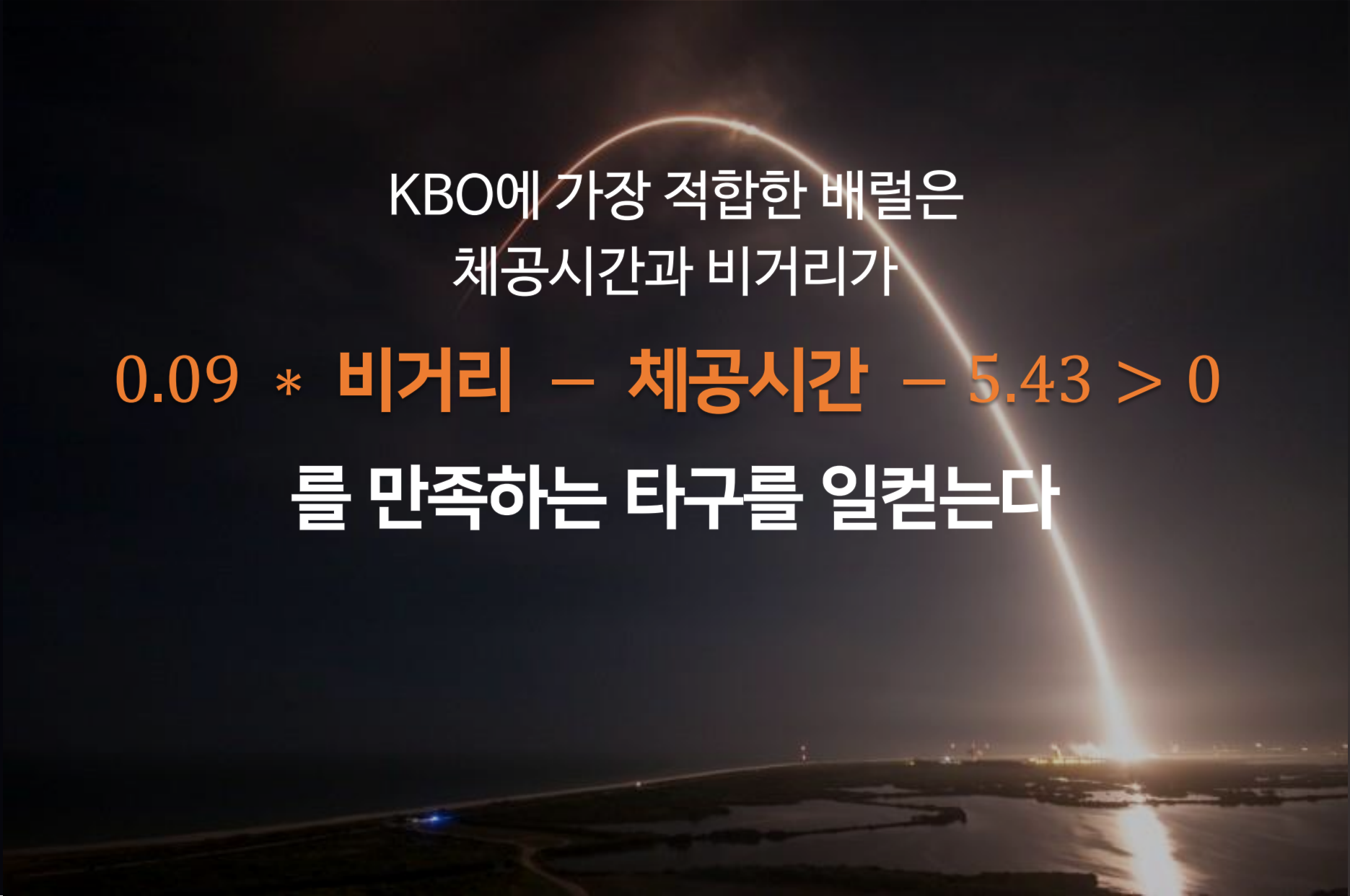


두 가지 종속변수를 비교해보았을 때
비거리-체공시간 종속변수가 높은 정확도(92%)를 보인다.

KBO에 가장 적합한 배럴은
체공시간과 비거리가

$$0.09 * \text{비거리} - \text{체공시간} - 5.43 > 0$$

를 만족하는 타구를 일컫는다



03 배럴 정의

OLS Regression를 통해 배럴의 당위성 확인

OLS Regression Results						
Dep. Variable:	장타율		R-squared:	0.220		
Model:	OLS		Adj. R-squared:	0.220		
Method:	Least Squares		F-statistic:	1304.		
Date:	Wed, 15 Sep 2021		Prob (F-statistic):	8.54e-252		
Time:	00:04:47		Log-Likelihood:	-886.65		
No. Observations:	4623		AIC:	1777.		
Df Residuals:	4621		BIC:	1790.		
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
intercept	0.4112	0.005	80.470	0.000	0.401	0.421
배럴률	1.6012	0.044	36.114	0.000	1.514	1.688
Omnibus:	1241.793	Durbin-Watson:	1.991			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25601.910			
Skew:	0.771	Prob(JB):	0.00			
Kurtosis:	14.425	Cond. No.	10.3			

모형적합도

'장타율'에 대하여 '배럴률'로 예측하는 회귀분석을 실시한 결과,
이 회귀모형은 통계적으로 유의미하였다.
(F1, 4621) = 1304, $p < 0.05$)

독립변수

'배럴률'의 회귀계수는 1.6012로, '장타율'에 대하여
유의미한 예측변인인 것으로 나타났다.
(t(4621) = 36.114, $p < 0.05$)

배럴이 OPS를 예측하는데에 영향력 있는 변수인 것을 알 수 있다.



데이터 분석

04 데이터 분석

데이터 전처리

연봉 데이터 결측치 제거

- | 선수코드 '50802'의 2020년 연봉 데이터
- | 선수코드 '50802'의 홈런 1개

강우 콜드게임 데이터 제거

- | '이닝' 특성의 최대값이 6이하면 강우 콜드게임으로 판단
- | 경기코드 20180519NCKT0, 20180524LTSS0는 데이터셋 오류로 인해 포함된 경기

불필요한 데이터(투수) 제거

- | 투수의 데이터는 불필요하다고 판단
- | 포지션 = '투'인 데이터 제거

정규 타석 미달 데이터 제거

- | 정규시즌 규정타석 기준 = 팀경기수 * 3.1
- | 기준 미달의 데이터 제거

04 데이터 분석

데이터 전처리

연봉 단위 통일

- 만원, 달러가 혼재되어 있는 연봉 단위
- '원' 단위로 통일해줌

개명 선수

- 활동 중 개명을 한 선수의 중복된 선수코드
- 현재 이름 기준으로 전처리

선수코드	선수명	선수코드	선수명
15509	62895 한동민	15509	62895 한유성
114402	62895 한유성	32284	63559 백동훈
32284	63559 백민기	31890	63905 윤철준
63038	63559 백동훈	3858	64717 지시완
31890	63905 윤대영	34259	67207 이유찬
118639	63905 윤철준		
3858	64717 지성준		
116160	64717 지시완		
34259	67207 이병희		
65169	67207 이유찬		

희생플라이 컬럼 수정

- 트래킹 데이터와 실제 데이터 사이에 차이
- KBO 공식 홈페이지를 통해 차이나는 데이터 수정



모델 구축

05 모델 구축

모델 구축을 위한 데이터프레임

우리가 예측해야 하는 10명의 선수들의
2018년~2021년 8월
OPS 관련 데이터를 활용한다.

	월별	선수명	타수	안타	1루타	2루타	3루타	홈런	사사구	득점	희생플라이	선수코드	배럴률
0	201803	강백호	27	10	4	2	0	4	3	7	0	68050	0.250000
1	201803	김재환	25	5	3	0	0	2	3	5	1	78224	0.285714
2	201803	김현수	29	7	4	2	0	1	3	5	0	76290	0.105263
3	201803	로맥	26	10	7	0	0	3	5	9	0	67872	0.214286
4	201803	박건우	30	9	7	1	1	0	1	5	0	79215	0.043478
...
248	202107	양의지	17	5	3	1	0	1	1	1	2	76232	0.133333
249	202107	이정후	23	9	8	1	0	0	7	5	0	67341	0.050000
250	202107	전준우	27	9	5	4	0	0	3	6	0	78513	0.000000
251	202107	채은성	13	4	2	2	0	0	3	3	1	79192	0.153846
252	202107	최정	14	3	2	1	0	0	5	4	1	75847	0.000000

253 rows × 13 columns

OPS 예측에 사용할 데이터프레임 생성

VAR

벡터 자기회귀(Vector Auto Regression)

1. 우리가 하고 싶은 모델링

OPS 관련 변수들을 활용한 시계열 예측

2. VAR

다변량 시계열 분석이 가능한 모형

3. 우리의 모델링 방향성

각각의 변수에 대해 하나의 모델을 구성하고 동시에 추정함으로써 상호작용 고려

모델 선택 이유 및 모델 구축 방향성 제시



OPS 예측

우리가 예측해야 하는 10명의 선수들의
최종 OPS



2021년 9월 15일~30일: 14경기
2021년 10월 1일~8일: 7경기

조화 평균으로 OPS 예측
9월 OPS * 14/21 + 10월 OPS * 7/21

예측 기간을 고려하여 조화 평균을 이용한 OPS 예측

06 OPS 예측

우리가 예측해야 하는 10명의 선수들의
최종 OPS



2021년 9월 15일~30일: 14경기
2021년 10월 1일~8일: 7경기

조화 평균으로 OPS 예측
9월 OPS * 14/21 + 10월 OPS * 7/21

PCODE	OPS	장타율	출루율
76232	0.995774	0.572739	0.423035
68050	0.84115	0.46983	0.37132
75847	1.008615	0.593741	0.414873
67341	0.873131	0.471251	0.40188
79192	0.847945	0.488704	0.359242
78224	0.976023	0.580061	0.395962
78513	0.818339	0.449584	0.368755
76290	0.902504	0.522329	0.380175
79215	0.930486	0.517072	0.413414
67872	0.950169	0.555126	0.395043

최종적으로 우리가 예측한 10명의 선수들의 OPS

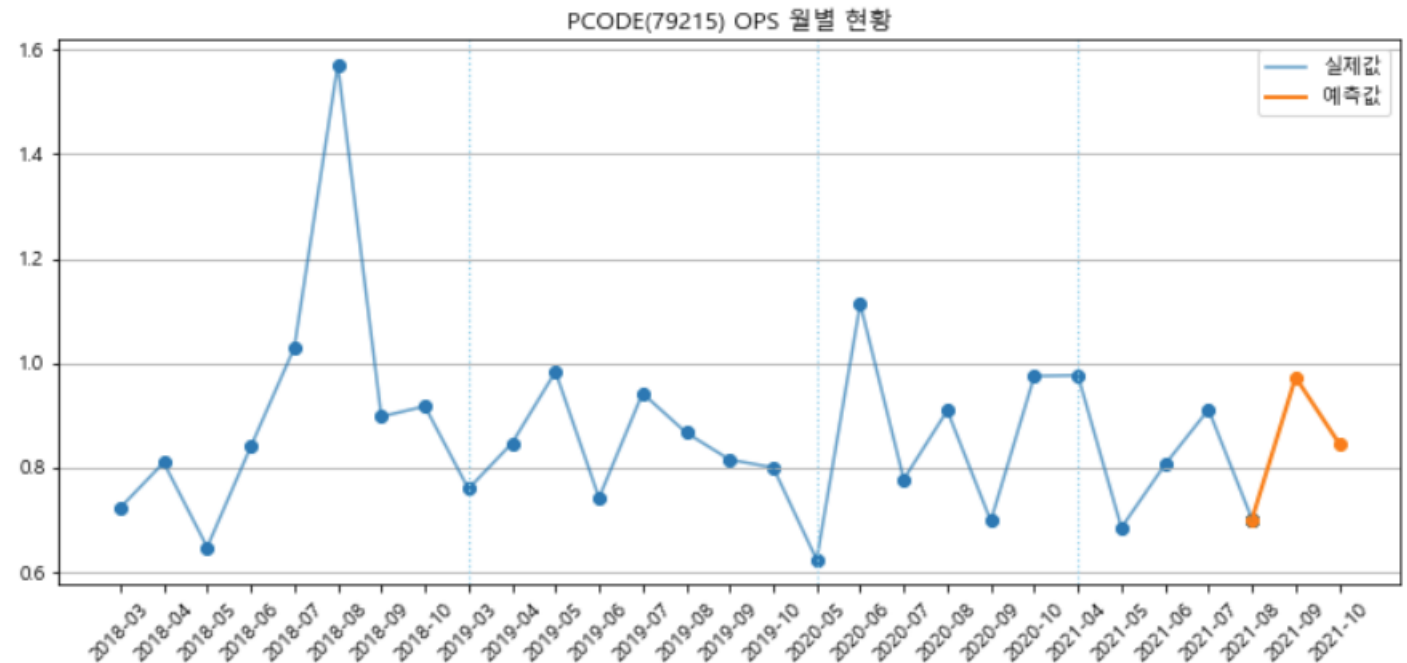
06 OPS 예측

최종 예측 그래프

우리가 예측해야하는
2021.09~2021.10의 OPS가

현재(2018~2021.07) 까지
OPS의 범위 내에서

안정적으로 예측된 것을 확인할 수 있다.

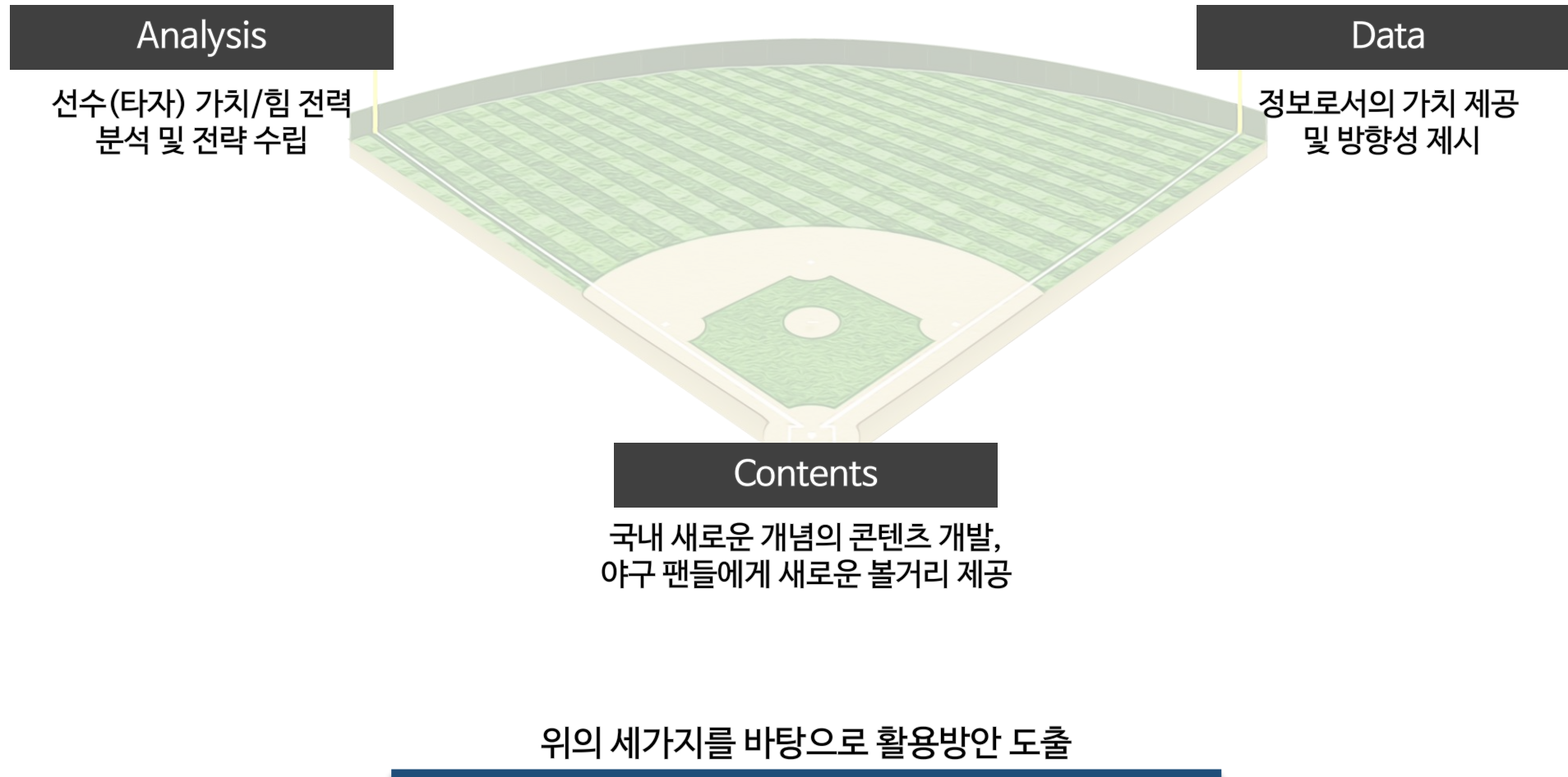


예측한 결과에 대한 해석



활용 방안

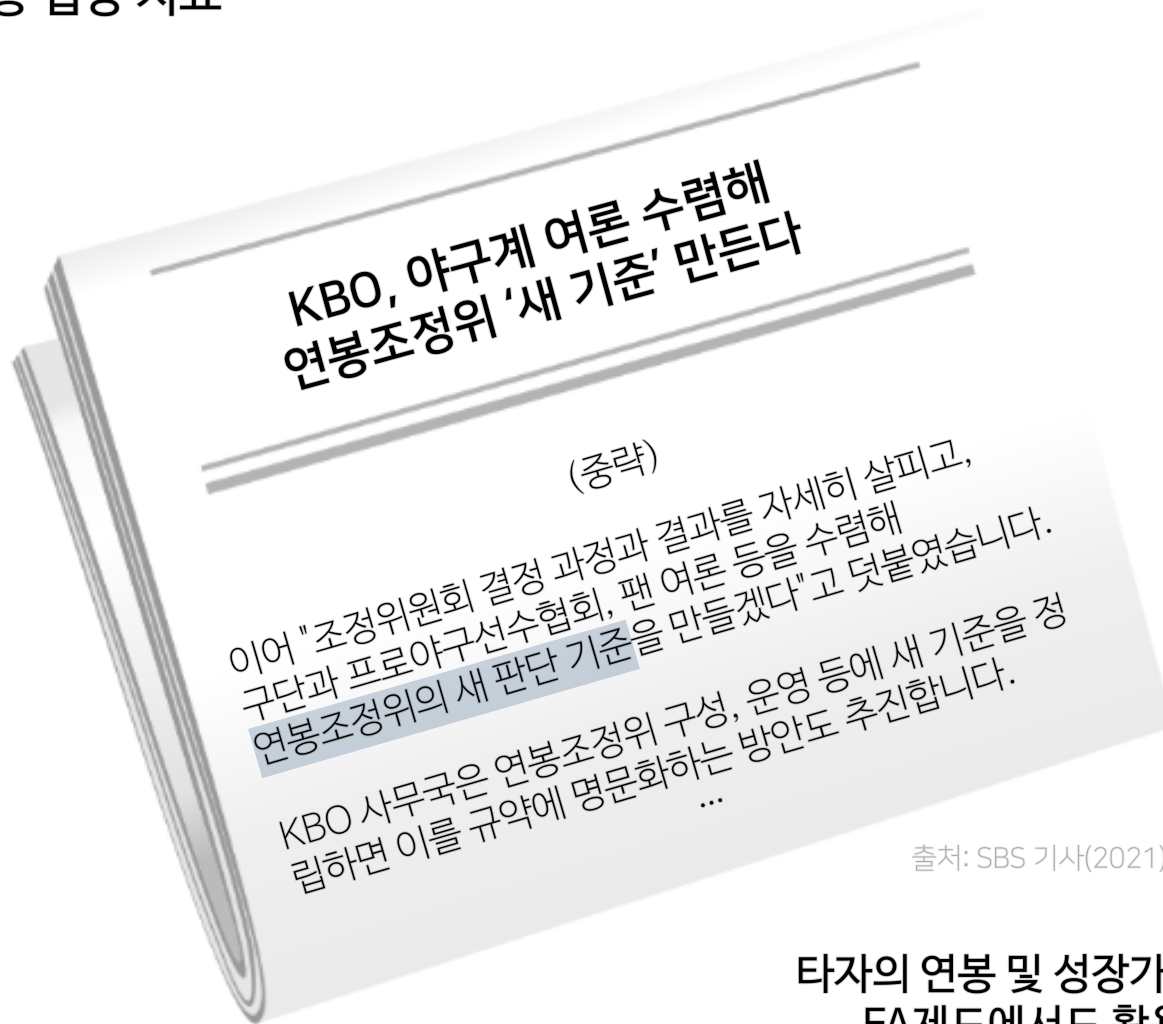
07 활용 방안





위의 세가지를 바탕으로 활용방안 도출

연봉 협상 지표



출처: SBS 기사(2021)

새 기준으로 OPS 예측에 배럴을 활용할 수 있다.

현재 타자의 OPS와 삼진 비율이 타자의 능력을 평가하는 지표
높은 정확도로 OPS를 예측할 수 있다면
연봉협상 테이블에서 유의미하게 사용할 수 있다.

타자의 연봉 및 성장가능성 판단 지표:
FA제도에서도 활용 가능하다.

KBO 선수 훈련 체계 구축



출처: 일본 덴츠 기사(2020)

일본의 Deep Nine

동영상으로 신체정보 정량화
AI 자세 추정 어플리케이션
일본 프로구단에 시범 도입



출처: 네이버 포스트 AI가 바꾸는 세상(2020)

스포츠에서 포즈 인식

배럴타구 자세를 학습한 딥러닝 모델 생성
자세와 일치하는지 여부를 판단
배럴타구를 생산하는 자세 훈련 가능

딥러닝 포즈 인식을 활용한 방안:
배럴 타구를 치기 위한 훈련 방향 제시 가능하다.

참고 링크

MLB 데이터

<https://baseballsavant.mlb.com>

배럴 정의 참고

<https://www.mlb.com>

야구 용어

https://ko.wikipedia.org/wiki/%EC%95%BC%EA%B5%AC_%EA%B8%B0%EB%A1%9D

KBO 데이터

<http://www.statiz.co.kr/main.php>

KBO 선수 데이터

<https://www.koreabaseball.com/Default.aspx>

야구공의 항력 계수 논문

<https://physics.csuchico.edu/baseball/Pubs/drag.pdf>

공기의 밀도 구하는 공식

https://en.wikipedia.org/wiki/Density_of_air

기상청 날씨 데이터

<https://data.kma.go.kr/cmmn/main.do>

var 관련 논문

<https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=DIK00015514988&dbt=DIK0>

ops와 득점 관한 연구 논문

<https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE08963452>

shap

<https://github.com/slundberg/shap>



감사합니다

궁금한 점이 있다면 자유롭게 말씀해주세요

팀명: 운수영원

팀원: 박재^운, 곽^수연, 최선^영, 곽희^원