



2021 BIG CONTEST

프로야구 배럴(Barrel)을 통한 타자 성적(OPS) 예측

효도원

팀장 김효림(gyfla810@naver.com)

팀원 변성도(dmdsns@naver.com)

팀원 장원석(ws9623@naver.com)



목차

1. 문제 정의

2. 데이터 개요 및 분석 목표

3. 데이터 전처리

4. EDA 및 배럴타구 정의

5. 모델 구축 및 훈련

6. 분석 결과

7. 결론 및 느낀점

8. 향후 계획

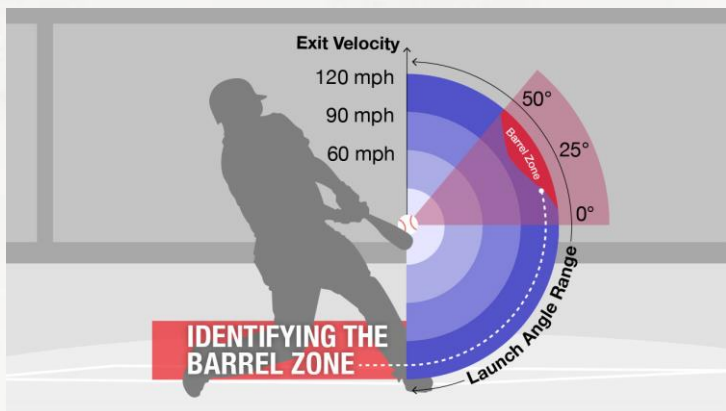


1. 문제 정의

제공데이터 및 분석의 필요성

배럴타구란?

타구 속도와 발사각도의 조합상 평균적으로
타율 .500, 장타율 1.500이상을 생산하는 타구

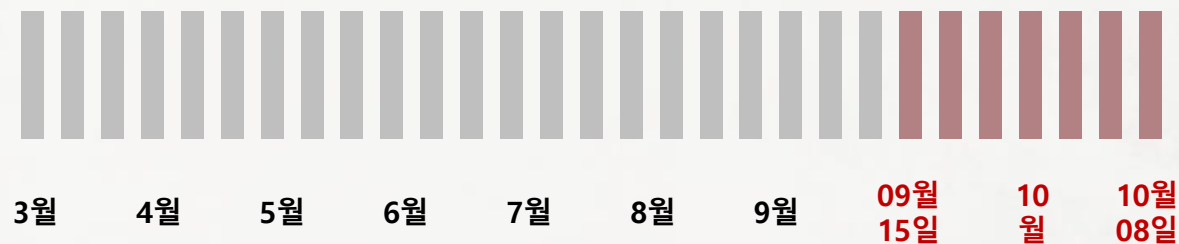


메이저리그 기준이 아닌
한국 프로야구 사정에 맞는 배럴타구 정의

OPS란?

선수의 출루 능력과 장타 능력을
하나의 숫자로 간편하게 나타낼 수 있는 지표

[KBO리그 2021시즌]



2021시즌 9월 15일 ~ 10월 8일
특정 타자들의 OPS 성적 예측



2. 데이터 개요 및 분석목표

기본 제공 데이터



타구 트래킹
2018~2021



타자 연도별 성적
2018~2021



선수 정보
2018~2021



경기일정
2021



팀 이름 정보
2018~2021



월별 타자 성적
2018.03 ~ 2021.09

■ 추가 파일(크롤링)

2. 데이터 개요 및 분석목표

분석 방법 및 목표

1. 전처리 및 EDA



데이터 전처리
및
탐색적 분석

2. 배럴타구 정의



타격 트래킹 데이터 기반
한국형 배럴타구 정의

3. 모델링



분석 모델 선정
및
타자 OPS 예측

4. 결과 해석



모델링 결과에 따른 해석
및
향후 활용 방안 모색

3. 데이터 전처리

이상치 제거 – 번트타구, 타구속도, 발사각도

번트 타구 삭제

*번트 : 야구방망이를 스윙하지 않고 내야에 공이 천천히 구르도록 일부러 볼에 배트를 살짝 갖다 대는 것

```
hts = hts.loc[(hts['타격결과'] != '번트안타') &
              (hts['타격결과'] != '번트아웃') &
              (hts['타격결과'] != '희생번트')]
hts.reset_index(drop = True, inplace = True)
```

번트는 방망이를 휘둘러 타격하는 것이 아닌,
갖다 대는 것이므로
배럴타구의 정의에 맞지 않아 삭제

타구속도, 발사각도 이상치 제거

```
# 이상치 제거
# 함수 정의
def outlier(df, column):

    Q1 = np.quantile(df[column], 0.25)
    Q3 = np.quantile(df[column], 0.75)
    IQR = Q3 - Q1
    minimum = Q1 - (IQR * 1.5)
    maximum = Q3 + (IQR * 1.5)
    print('-----', column, '의 이상치-----')
    print('IQR : ', IQR)
    print('minimum : ', minimum)
    print('maximum : ', maximum)

    NumOfOutlier = df[(df[column] > maximum) | (df[column] < minimum)].shape[0]
    print('upper bound 이상치 개수 : ', df[df[column] > maximum].shape[0])
    print('lower bound 이상치 개수 : ', df[df[column] < minimum].shape[0])
    print('총 이상치 개수 : ', NumOfOutlier)
```

```
1 outlier(hts, '타구속도')
```

```
----- 타구속도 의 이상치-----
IQR : 27.670000000000016
minimum : 80.34499999999997
maximum : 191.02500000000003
upper bound 이상치 개수 : 2
lower bound 이상치 개수 : 2137
총 이상치 개수 : 2139
```

```
1 outlier(hts, '발사각도')
```

```
----- 발사각도 의 이상치-----
IQR : 33.6
minimum : -51.800000000000004
maximum : 82.60000000000001
upper bound 이상치 개수 : 610
lower bound 이상치 개수 : 25
총 이상치 개수 : 635
```

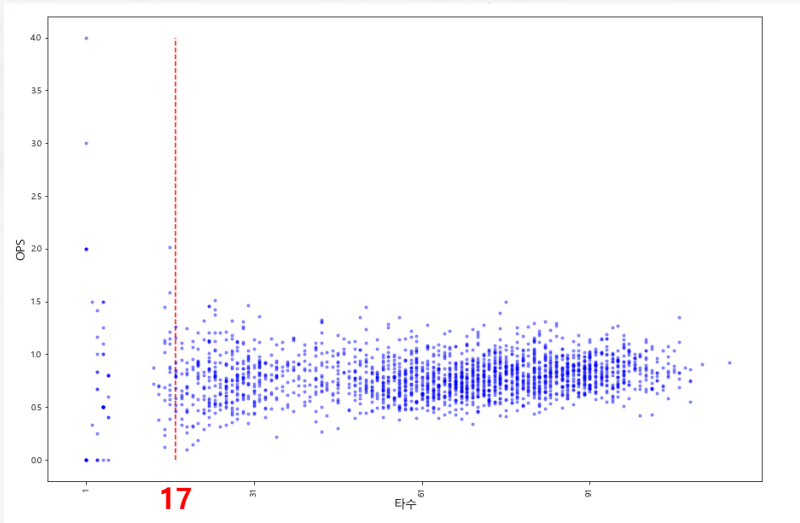
타구속도와 발사각도에 대한 이상치 제거



3. 데이터 전처리

이상치 제거 – 규정타수, OPS

일정 타수 이하 선수 데이터 제거



```
outlier(hitter, 'OPS')
```

```
----- OPS 의 이상치 -----  
IQR : 0.245  
minimum : 0.2934999999999999  
maximum : 1.2734999999999999  
upper bound 이상치 개수 : 38  
lower bound 이상치 개수 : 23  
총 이상치 개수 : 61
```

타수가 16미만에서는 OPS가 너무 크거나 작아서
분석에 영향을 끼칠것으로 예상되어 이상치로 간주하고 제거

3. 데이터 전처리

데이터 분할

우선 연도별로 장타인 집단과 장타가 아닌 집단을 비교하기 위해 hts파일에 타격 결과를 이용하여 데이터 분할

```
good_2018 = hts_2018[(hts_2018['HIT_RESULT'] == '2루타') | (hts_2018['HIT_RESULT'] == '3루타') | (hts_2018['HIT_RESULT'] == '홈런')]  
good_2019 = hts_2019[(hts_2019['HIT_RESULT'] == '2루타') | (hts_2019['HIT_RESULT'] == '3루타') | (hts_2019['HIT_RESULT'] == '홈런')]  
good_2020 = hts_2020[(hts_2020['HIT_RESULT'] == '2루타') | (hts_2020['HIT_RESULT'] == '3루타') | (hts_2020['HIT_RESULT'] == '홈런')]  
good_2021 = hts_2021[(hts_2021['HIT_RESULT'] == '2루타') | (hts_2021['HIT_RESULT'] == '3루타') | (hts_2021['HIT_RESULT'] == '홈런')]
```

또한 연도별로 안타인 집단과 안타가 아닌 집단을 비교하기 위해 hts파일에 타격 결과를 이용하여 데이터 분할

```
hit_2018 = hts_2018[(hts_2018['HIT_RESULT'] == '내야안타(1루타)') | (hts_2018['HIT_RESULT'] == '1루타') |  
hit_2019 = hts_2019[(hts_2019['HIT_RESULT'] == '내야안타(1루타)') | (hts_2019['HIT_RESULT'] == '1루타') |  
hit_2020 = hts_2020[(hts_2020['HIT_RESULT'] == '내야안타(1루타)') | (hts_2020['HIT_RESULT'] == '1루타') |  
hit_2021 = hts_2021[(hts_2021['HIT_RESULT'] == '내야안타(1루타)') | (hts_2021['HIT_RESULT'] == '1루타') |  
  
(hts_2018['HIT_RESULT'] == '2루타') | (hts_2018['HIT_RESULT'] == '3루타') | (hts_2018['HIT_RESULT'] == '홈런')]  
(hts_2019['HIT_RESULT'] == '2루타') | (hts_2019['HIT_RESULT'] == '3루타') | (hts_2019['HIT_RESULT'] == '홈런')]  
(hts_2020['HIT_RESULT'] == '2루타') | (hts_2020['HIT_RESULT'] == '3루타') | (hts_2020['HIT_RESULT'] == '홈런')]  
(hts_2021['HIT_RESULT'] == '2루타') | (hts_2021['HIT_RESULT'] == '3루타') | (hts_2021['HIT_RESULT'] == '홈런')]
```


3. 데이터 전처리

데이터 추가 – 2018.03 ~ 2021.09 월별 타자 성적

정확한 성적 예측을 위해 기존 연도별 타자 성적을 세분화 하기 위해 월별 타자 성적 데이터 추가

STATIZ

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

2021 2018 시작 끝 팀:전체 포지션: 정규 규정 상황: 승선

상황별기록

기간: 3월 03-01 ~ 11-30 요일:전체 시간:전체 홈/원정:전체 vs:전체 구장:전체 As:전체 타순:전체

투수:전체 이닝:전체 상대:전체 심판:전체 아웃:전체 주자:전체 점수:전체 카운트:전체 After:전체

니:전체

검색

세부/상황별 기록은 주포지션, 홈선고, 홈선대, 환산, As 등이 적용되지 않습니다. [자료 : 18 타석이상]

순위	이름	팀	타율	타석	타수	득점	안타	2타	3타	홈런	루타	타점	도루	도실	볼넷	사구	고4	삼진	병살	희타	희비	타율	출루	장타	OPS
1	양의지	18.5	.500	29	24	5	12	4	0	1	19	4	0	0	3	2	0	1	1	0	0	.500	.586	.792	1.378
2	호잉	18.5	.500	22	22	5	11	2	1	1	18	4	3	1	0	0	0	1	0	0	0	.500	.500	.818	1.318
3	김주찬	18.5	.467	21	15	4	7	5	0	1	15	7	0	0	2	0	0	3	0	1	3	.467	.450	1.000	1.450
4	박용택	18.5	.444	31	27	3	12	6	0	0	18	2	0	0	4	0	0	6	1	0	0	.444	.516	.667	1.183
5	정현	18.5	.444	18	18	3	8	3	0	0	11	0	0	0	0	0	0	4	1	0	0	.444	.444	.611	1.056
6	송광민	18.5	.440	27	25	4	11	1	0	2	18	5	0	0	2	0	0	6	0	0	0	.440	.482	.720	1.202
7	박병호	18.5	.435	35	23	6	10	0	0	3	19	8	0	1	9	3	0	5	1	0	0	.435	.629	.826	1.455
8	김동원	18.5	.417	26	24	6	10	1	0	4	23	13	2	0	2	0	0	4	0	0	0	.417	.462	.958	1.420
9	김민성	18.5	.414	31	29	4	12	4	0	3	25	10	0	0	2	0	0	7	0	0	0	.414	.452	.862	1.314
10	이정후	18.5	.407	34	27	7	11	3	0	0	14	2	0	0	4	1	1	3	0	2	0	.407	.500	.519	1.019

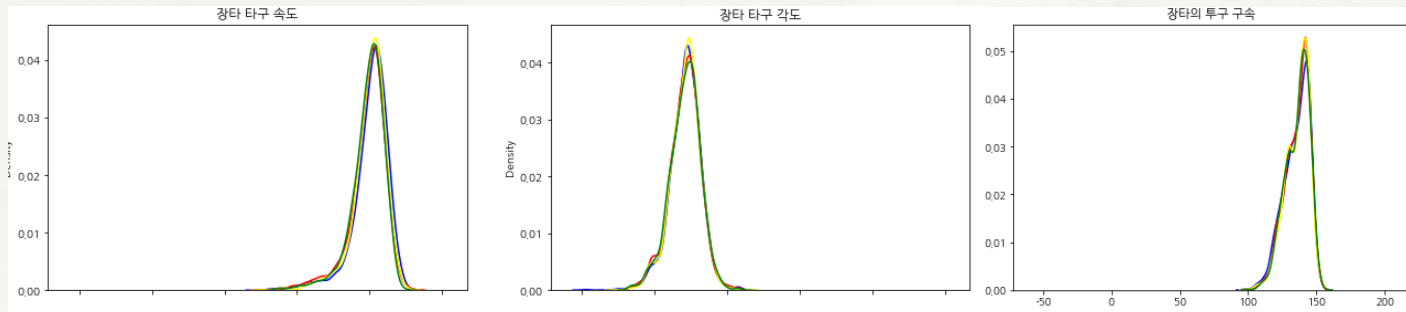
	이름	시즌	month	도루	도실	타석	타수	득점	안타	2타	...	사구	고4	삼진	병살	희타	희비	타율	출루	장타	OPS
0	양의지	18	3	0	0	29	24	5	12	4	...	2	0	1	1	0	0	0.500	0.586	0.792	1.378
1	호잉	18	3	3	1	22	22	5	11	2	...	0	0	1	0	0	0	0.500	0.500	0.818	1.318
2	김주찬	18	3	0	0	21	15	4	7	5	...	0	0	3	0	1	3	0.467	0.450	1.000	1.450
3	박용택	18	3	0	0	31	27	3	12	6	...	0	0	6	1	0	0	0.444	0.516	0.667	1.183
4	정현	18	3	0	0	18	18	3	8	3	...	0	0	4	1	0	0	0.444	0.444	0.611	1.056
...
454	추신수	21	9	0	0	36	33	3	5	1	...	0	0	11	0	0	0	0.152	0.222	0.182	0.404
455	고종욱	21	9	1	0	29	27	2	4	0	...	0	0	6	0	0	0	0.148	0.207	0.148	0.355
456	박병호	21	9	0	0	44	36	4	5	2	...	0	1	13	0	0	0	0.139	0.296	0.361	0.657
457	박준영	21	9	0	0	33	28	2	3	1	...	2	0	12	1	0	0	0.107	0.242	0.143	0.385
458	최인호	21	9	0	0	29	22	1	1	0	...	0	0	7	0	0	1	0.045	0.241	0.045	0.287

2143 rows x 25 columns

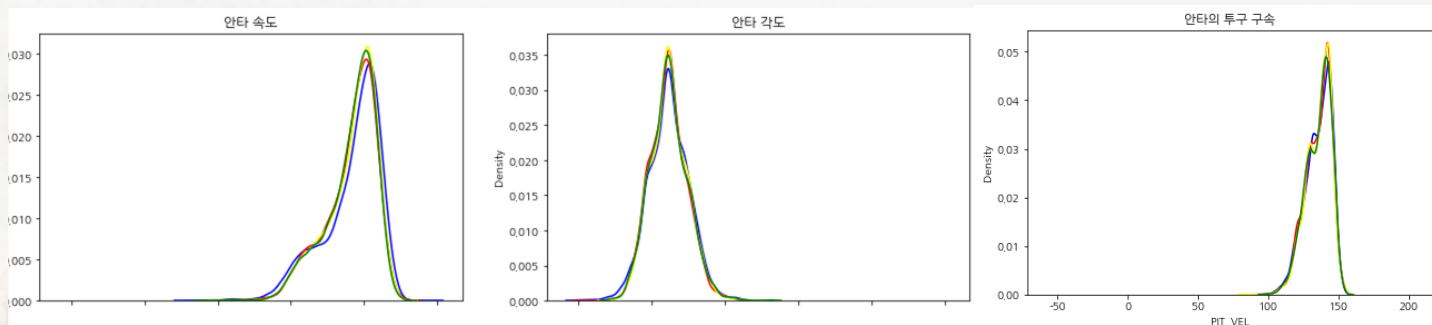
4. EDA 및 배럴타구 정의

장타, 안타 별 각종 분포 확인

연도별로 장타 그룹과 장타가 아닌 그룹의 타구속도, 발사 각도, 투구 구속 분포 확인



연도별로 안타 그룹과 안타가 아닌 그룹의 타구속도, 발사 각도, 투구 구속 분포 확인



연도 별 그룹간의 차이가 없어 보이므로 데이터를 합쳐 분석 진행

4. EDA 및 배럴타구 정의

안타와 아웃 그룹 간 타구속도, 발사각도 평균 검정

```
1 hit_vel = hts_hit['타구속도']
2
3 out_vel = hts_out['타구속도']
4 print(shapiro(hit_vel))
5 print(shapiro(out_vel))
6 print()
7 print(stats.levene(hit_vel,out_vel))
8
9 t_result = stats.ttest_ind(hit_vel,out_vel, equal_var = False)
10 t, p = t_result.statistic.round(3), t_result.pvalue.round(3)
11 print()
12 print('2-Sample welchs t-test')
13 print('t통계량 : {}'.format(t))
14 print('p-value : {}'.format(p))
```

```
ShapiroResult(statistic=0.9345917105674744, pvalue=0.0)
ShapiroResult(statistic=0.9756127595901489, pvalue=0.0)
```

```
LeveneResult(statistic=264.4300401814186, pvalue=2.1567685866934197e-59)
```

```
2-Sample welchs t-test
t통계량 : 95.203
p-value : 0.0
```

```
1 hit_ang = hts_hit['발사각도']
2
3 out_ang = hts_out['발사각도']
4 print(shapiro(hit_ang))
5 print(shapiro(out_ang))
6 print()
7 print(stats.levene(hit_ang,out_ang))
8
9 t_result = stats.ttest_ind(hit_ang,out_ang, equal_var = False)
10 t, p = t_result.statistic.round(3), t_result.pvalue.round(3)
11 print()
12 print('2-Sample welchs t-test')
13 print('t통계량 : {}'.format(t))
14 print('p-value : {}'.format(p))
```

```
ShapiroResult(statistic=0.9973626136779785, pvalue=3.903095480122203e-26)
ShapiroResult(statistic=0.9650865197181702, pvalue=0.0)
```

```
LeveneResult(statistic=31769.370176951594, pvalue=0.0)
```

```
2-Sample welchs t-test
t통계량 : -77.643
p-value : 0.0
```

정규성, 등분산성 검정 결과 정규성은 만족, 등분산성은 만족하지 않음

두 검정 모두 $p\text{-value} < 0.05$ 이므로

타구속도와 발사각도에 대한 두 집단 간의 유의미한 차이가 있다

4. EDA 및 배럴타구 정의

안타와 아웃 집단간의 투구구속에 대한 두 집단 평균 검정

```
1 hit_pit = hts_hit['투구구속']
2
3 out_pit = hts_out['투구구속']
4 print(shapiro(hit_pit))
5 print(shapiro(out_pit))
6 print()
7 print(stats.levene(hit_pit,out_pit))
8
9 t_result = stats.ttest_ind(hit_pit,out_pit, equal_var = False)
10 t, p = t_result.statistic.round(3), t_result.pvalue.round(3)
11 print()
12 print('2-Sample welchs t-test')
13 print('t통계량 : {}'.format(t))
14 print('p-value : {}'.format(p))

ShapiroResult(statistic=0.9665027260780334, pvalue=0.0)
ShapiroResult(statistic=0.966221034526825, pvalue=0.0)

LeveneResult(statistic=11.025712944859189, pvalue=0.0008988402248235634)

2-Sample welchs t-test
t통계량 : 1.911
p-value : 0.056
```

정규성, 등분산성 검정 결과
정규성은 만족하나
등분산성은 만족하지 않음
p-value>0.05 이므로
두 집단간의 유의미한 차이가 없다.

안타와 아웃집단의 유의미한 변수는 타구 속도, 발사각도이다.

4. EDA 및 배럴타구 정의

장타와 장타가아닌 집단간의 타구속도와 발사각도에 대한 두집단 평균 검정

```
1 good_vel = hts_good['타구속도']
2
3 not_good_vel = hts_not_good['타구속도']
4 print(shapiro(good_vel))
5 print(shapiro(not_good_vel))
6 print()
7 print(stats.levene(good_vel, not_good_vel))
8
9 t_result = stats.ttest_ind(good_vel, not_good_vel, equal_var = False)
10 t, p = t_result.statistic.round(3), t_result.pvalue.round(3)
11 print()
12 print('2-Sample welchs t-test')
13 print('t통계량 : {}'.format(t))
14 print('p-value : {}'.format(p))
```

```
ShapiroResult(statistic=0.9008217453956604, pvalue=0.0)
ShapiroResult(statistic=0.9737948775291443, pvalue=0.0)
```

```
LeveneResult(statistic=3940.835310987613, pvalue=0.0)
```

```
2-Sample welchs t-test
t통계량 : 136.373
p-value : 0.0
```

```
1 good_ang = hts_good['발사각도']
2
3 not_good_ang = hts_not_good['발사각도']
4 print(shapiro(good_ang))
5 print(shapiro(not_good_ang))
6 print()
7 print(stats.levene(good_ang, not_good_ang))
8
9 t_result = stats.ttest_ind(good_ang, not_good_ang, equal_var = False)
10 t, p = t_result.statistic.round(3), t_result.pvalue.round(3)
11 print()
12 print('2-Sample welchs t-test')
13 print('t통계량 : {}'.format(t))
14 print('p-value : {}'.format(p))
```

```
ShapiroResult(statistic=0.9835506081581116, pvalue=1.2352279600869578e-36)
ShapiroResult(statistic=0.9613795280456543, pvalue=0.0)
```

```
LeveneResult(statistic=8034.237117605266, pvalue=0.0)
```

```
2-Sample welchs t-test
t통계량 : 136.373
p-value : 0.0
```

정규성, 등분산성 검정 결과 정규성은 만족하나 등분산성은 만족하지 않음

두 검정 모두 $p\text{-value} < 0.05$ 이므로

타구속도와 발사각도에 대한 두 집단 간의 유의미한 차이가 있다.

4. EDA 및 배럴타구 정의

장타와 장타가 아닌 집단간의 투구구속에 대한 두집단 평균 검정

```
1 good_pit = hts_good['투구구속']
2
3 not_good_pit = hts_not_good['투구구속']
4 print(shapiro(good_pit))
5 print(shapiro(not_good_pit))
6 print()
7 print(stats.levene(good_pit, not_good_pit))
8
9 t_result = stats.ttest_ind(good_pit, not_good_pit, equal_var = False)
10 t, p = t_result.statistic.round(3), t_result.pvalue.round(3)
11 print()
12 print('2-Sample welchs t-test')
13 print('t통계량 : {}'.format(t))
14 print('p-value : {}'.format(p))
```

```
ShapiroResult(statistic=0.9616132974624634, pvalue=0.0)
ShapiroResult(statistic=0.9669203758239746, pvalue=0.0)
```

```
LeveneResult(statistic=15.223847134834338, pvalue=9.554301742678714e-05)
```

```
2-Sample welchs t-test
t통계량 : 1.46
p-value : 0.144
```

정규성, 등분산성 검정 결과 정규성은 만족
하나 등분산성은 만족하지 않음

p-value > 0.05 이므로
두 집단간의 유의미한 차이가 없다.

따라서 장타와 장타가 아닌 집단의 유의미한 변수는 타구속도, 발사각도이다.

4. EDA 및 배럴타구 정의

배럴타구 정의 – 특정 데이터 추출

배럴타구의 특징을 찾기 위해 2루타, 3루타, 홈런 타구를 추출

```
hts_barrel = hts[(hts['타격결과'] == '2루타') |(hts['타격결과'] == '3루타') |(hts['타격결과'] == '홈런') ]
```

추출된 타구들에 대한 이상치 제거

```
1 outlier(hts_barrel, '타구속도')
```

```
----- 타구속도 의 이상치-----  
IQR : 12.95750000000001  
minimum : 125.35374999999998  
maximum : 177.18375000000003  
upper bound 이상치 개수 : 9  
lower bound 이상치 개수 : 641  
총 이상치 개수 : 650
```

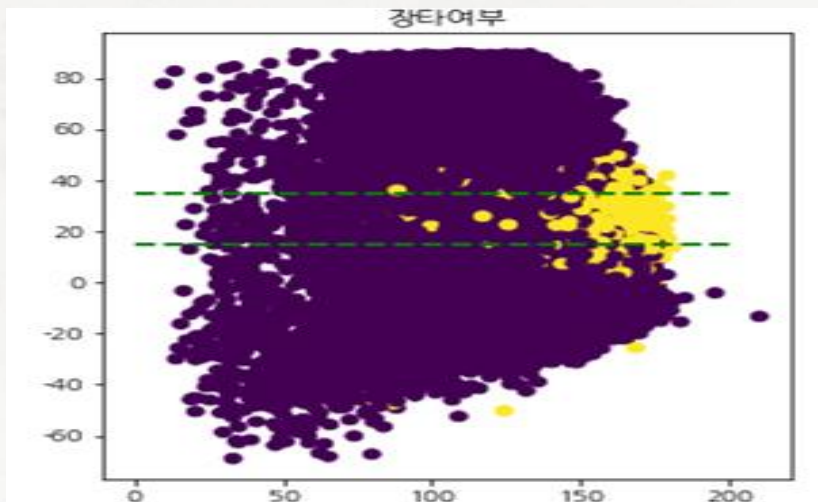
```
1 outlier(hts_outlier, '발사각도')
```

```
----- 발사각도 의 이상치-----  
IQR : 12.799999999999999  
minimum : -4.299999999999999  
maximum : 46.9  
upper bound 이상치 개수 : 47  
lower bound 이상치 개수 : 233  
총 이상치 개수 : 280
```

4. EDA 및 배럴타구 정의

배럴타구 정의 - 특정 각도, 속도 추출

산점도(발사각도, 타구속도)



안타와 장타를 결정하는데 있어
타구속도보다 **발사각도가 더 중요**

각도 : 15~35, 타구속도에 대한 타율

각도 15~35 고정

- 타구속도 146 ~ 180일때 -> $7270 / 11603 = 62.66\%$
- 타구속도 146 ~ 170일때 -> $7162 / 11492 = 62.32\%$
- 타구속도 146 ~ 165일때 -> $6726 / 11003 = 61.13\%$
- 타구속도 147 ~ 180일때 -> $6968 / 10768 = 64.71\%$
- 타구속도 148 ~ 180일때 -> $6589 / 9872 = 66.74\%$
- 타구속도 149 ~ 180일때 -> $6219 / 9051 = 68.71\%$
- 타구속도 150 ~ 180일때 -> $5819 / 8257 = 70.47\%$
- 타구속도 151 ~ 180일때 -> $5420 / 7453 = 72.72\%$
- 타구속도 155 ~ 180일때 -> $3608 / 4535 = 79.56\%$
- 타구속도 140 ~ 180일때 -> $8751 / 16510 = 53.00\%$
- 타구속도 143 ~ 180일때 -> $8117 / 14161 = 57.32\%$
- 타구속도 145 ~ 180일때 -> $7590 / 12460 = 60.91\%$


배럴 타구의 기준인
5할을 넘기는 것을 알 수 있음



4. EDA 및 배럴타구 정의

배럴타구 정의 - 발사각도 재정의

- 타구 속도 : 높으면 높을 수록 좋음
 - 발사 각도 : 높으면 높을 수록 좋은게 아님 - 각도는 잘 맞은 타구의 각도의 일정 범위의 각도가 좋은 각도임
 - 위의 잘 맞은 타구(2루타, 3루타, 홈런)의 발사 각도 : **21.01도**가 좋은 각도라고 가정
 - **공식-**
 - **100 - |잘 맞은 각도 - x|**
 - 예시
 - 1. 발사각도 23도의 홈런 : $\$100 - |21 - 23| = 98\$$
 - 2. 발사각도 50의 플라이 아웃 : $\$100 - |21 - 50| = 71\$$
 - 3. 발사각도 -23도의 땅볼 아웃 : $\$100 - |21 - (-23)| = 56\$$
- => 각도가 **21.01(최적의 각도)**도에서 음의 방향이거나 양의 방향으로 멀어질 수록 점수가 낮아짐



배럴 지수 정의 : 재정의된 발사각도 + 타구속도

4. EDA 및 배럴타구 정의

배럴타구 정의 - 배럴지수 정의

표준화 - Minmax

	타구속도	파생_발사각도
0	0.488539	0.695866
1	0.523687	0.844101
2	0.688252	0.734551
3	0.318720	0.894663
4	0.778892	0.754213
...
117682	0.581471	0.636236
117683	0.198090	0.848956
117684	0.635053	0.826484
117685	0.628176	0.959911
117686	0.574021	0.899518

타구 속도와 재정의 된 발사각도(파생_발사각도)에 대한
Minmax 정규화 실시

배럴지수 생성

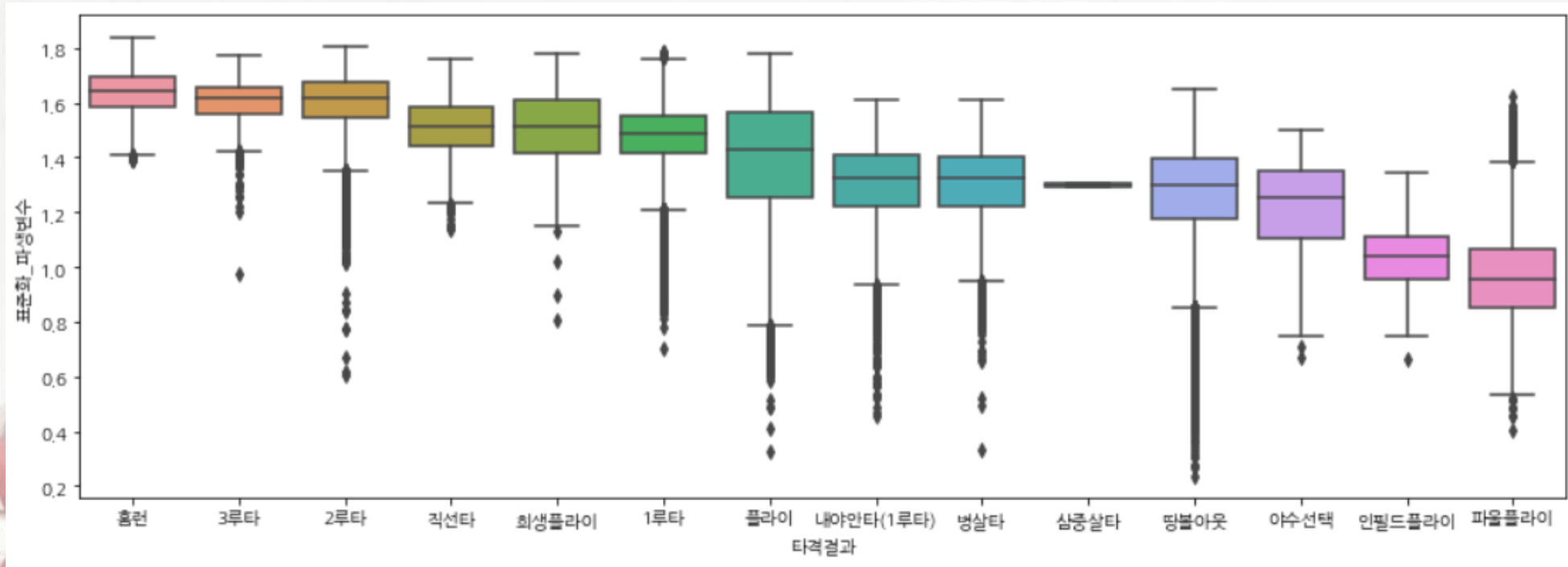
```
타격결과
파울플라이      0.703743
인필드플라이    0.721521
야수선택        1.010030
땅볼아웃        1.082918
병살타          1.120203
내야안타(1루타) 1.124069
삼중살타        1.156991
플라이          1.231936
1루타           1.347629
희생플라이      1.382895
직선타          1.394471
2루타           1.520521
3루타           1.527353
홈런            1.604240
Name: 표준화_파생변수, dtype: float64
```

배럴지수 : 재정의된 발사각도(파생_발사각도) + 타구속도
배럴지수가 높을 수록 안타, 장타가 될 확률이 높음

4. EDA 및 배럴타구 정의

배럴타구 정의 – 배럴지수 정의

타격 결과 별 배럴 지수 Box plot



4. EDA 및 배럴타구 정의

배럴타구 정의 - 특정 각도, 속도 추출

배럴 지수(barrel_point)를 수정하면서 타율 0.5, 장타율 1.5 이상의 배럴 지수 파악

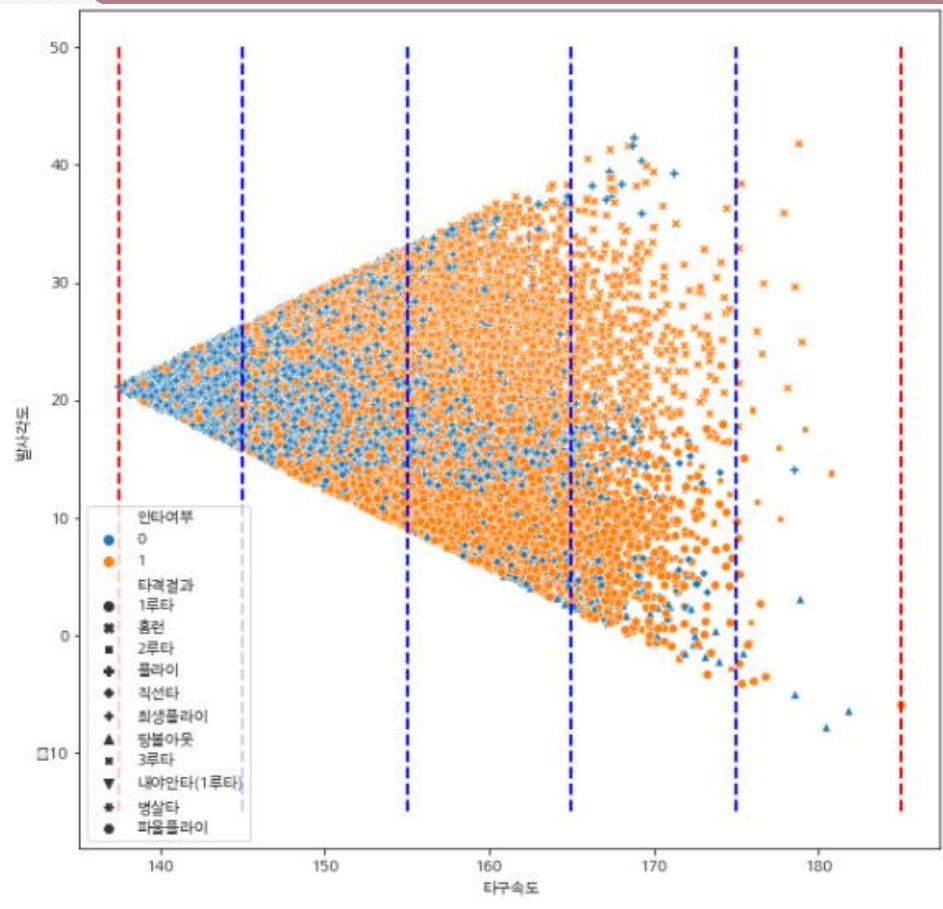
```
# barrel_point를 조절하면서 안타 5할, 장타율 1.5의 범위를 찾아야 함  
barrel_point = 1.54469  
  
# 위의 barrel_point의 이상의 타구들을 barrel1으로 설정  
barrel1 = hts_pasaeng[(hts_pasaeng['표준화_파생변수'] > barrel_point)]  
  
# 그리고 위의 barrel_point의 이상을 넘는 타구는 배럴여부 col에 1로 표시  
hts_pasaeng['배럴여부'] = 0  
hts_pasaeng['배럴여부'][(hts_pasaeng['표준화_파생변수'] > barrel_point)] = 1
```

배럴지수 1.54469이상일 때 타율 0.5, 장타율 1.5이상을 만들어낼 수 있는 타구로 파악

	연도	선수코드	이닝	타구속도	발사각도	투구구속	안타여부	장타여부
count	17635.000000	17635.000000	17635.000000	17635.000000	17635.000000	17635.000000	17635.000000	17635.000000
mean	2019.243493	70308.235498	4.965523	154.485846	18.981276	136.736811	0.646612	1.500482
std	1.047809	7288.576648	2.578154	7.224759	6.721918	8.876599	0.478035	1.455840
min	2018.000000	50054.000000	1.000000	137.490000	-7.900000	88.290000	0.000000	0.000000
25%	2018.000000	64300.000000	3.000000	149.160000	14.300000	130.775000	0.000000	0.000000
50%	2019.000000	70410.000000	5.000000	154.310000	19.200000	138.680000	1.000000	1.000000
75%	2020.000000	76313.000000	7.000000	159.495000	23.700000	143.510000	1.000000	2.000000
max	2021.000000	99810.000000	12.000000	185.050000	42.300000	156.480000	1.000000	4.000000

4. EDA 및 배럴타구 정의

배럴타구 정의 – 배럴지수 기반 산점도



배럴지수 1.54469이상일 때
발사각도와 타구 속도에 대한 산점도

타구 속도 : 137.49 ~ 144.99 / 발사 각도 : 15.9 ~ 25.8
타구 속도 : 145.0 ~ 154.99 / 발사 각도 : 9.3 ~ 32.7
타구 속도 : 155.0 ~ 164.99 / 발사 각도 : 2.6 ~ 38.3
타구 속도 : 165.0 ~ 174.79 / 발사 각도 : -3.4 ~ 42.3
타구 속도 : 175.07 ~ 181.87 / 발사 각도 : -7.9 ~ 41.8

각 구간 별 배럴에 해당하는 타구 속도 및 발사 각도

발사각도가
최적의 발사각도(21.01도)일 때는
타구속도가 낮아도 배럴 타구지만,
발사각도의 범위가 넓어질 수록
타구속도가 빨라지는 것을 볼 수 있음`

4. EDA 및 배럴타구 정의

배럴타구 정의 - 각종 지표 종합

각도 고정 시 배럴 기준

각도 15~35 고정

- 타구속도 146 ~ 180일때 -> $7270 / 11603 = 62.66\%$
- 타구속도 146 ~ 170일때 -> $7162 / 11492 = 62.32\%$
- 타구속도 146 ~ 165일때 -> $6726 / 11003 = 61.13\%$
- 타구속도 147 ~ 180일때 -> $6968 / 10768 = 64.71\%$
- 타구속도 148 ~ 180일때 -> $6589 / 9872 = 66.74\%$
- 타구속도 149 ~ 180일때 -> $6219 / 9051 = 68.71\%$
- 타구속도 150 ~ 180일때 -> $5819 / 8257 = 70.47\%$
- 타구속도 151 ~ 180일때 -> $5420 / 7453 = 72.72\%$
- 타구속도 155 ~ 180일때 -> $3608 / 4535 = 79.56\%$
- 타구속도 140 ~ 180일때 -> $8751 / 16510 = 53.00\%$
- 타구속도 143 ~ 180일때 -> $8117 / 14161 = 57.32\%$
- 타구속도 145 ~ 180일때 -> $7590 / 12460 = 60.91\%$

배럴 지수 생성 시 배럴 기준

타구 속도 : 137.49 ~ 144.99 / 발사 각도 : 15.9 ~ 25.8
타구 속도 : 145.0 ~ 154.99 / 발사 각도 : 9.3 ~ 32.7
타구 속도 : 155.0 ~ 164.99 / 발사 각도 : 2.6 ~ 38.3
타구 속도 : 165.0 ~ 174.79 / 발사 각도 : -3.4 ~ 42.3
타구 속도 : 175.07 ~ 181.87 / 발사 각도 : -7.9 ~ 41.8

< 배럴타구 >

- 발사 각도 : 15도 ~ 37도

- 타구 속도 : 140km/h ~ 180km/h



5. 모델 구축 및 훈련

모델링 과정 도식화

전처리

- 결측치 처리
- 이상치 조정

데이터 크롤링
: 월 별 타자 성적

- 파생변수 추가
- 배럴 비율
 - 세이버 메트릭스 지표

상관관계 분석

데이터 취합

모델링

Regression

Lasso
Ridge

Ensemble

Random Forest
XGBoost
CatBoost

최종 모델 선정

RMSE, MAE
성능 평가

최종 모델 선정

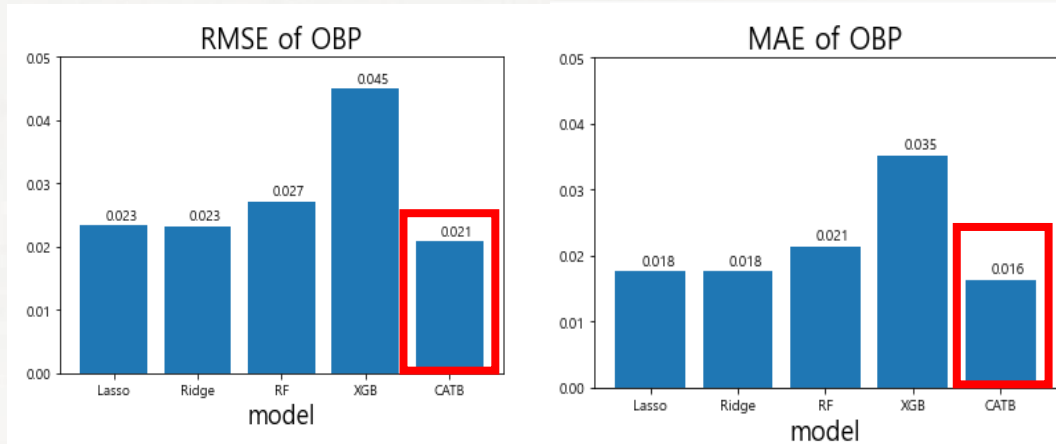
Test



6. 분석 결과

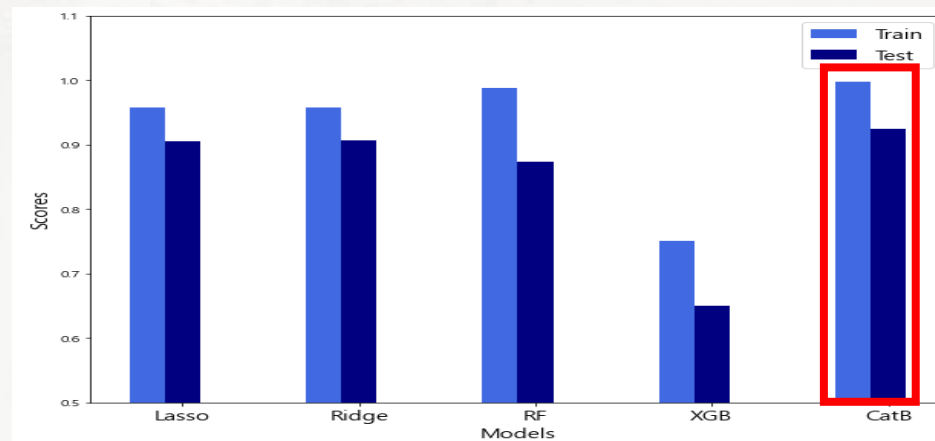
분석결과 - 출루율

모델 별 성능 비교



모델	RMSE	MAE
Lasso	0.023	0.018
Ridge	0.023	0.018
RandomForest	0.027	0.021
XGBoost	0.045	0.035
CatBoost	0.021	0.016

모델 별 정확도 비교



모델	Train Accuracy	Test Accuracy
Lasso	0.957	0.905
Ridge	0.958	0.906
RandomForest	0.988	0.872
XGBoost	0.750	0.650
CatBoost	0.998	0.925

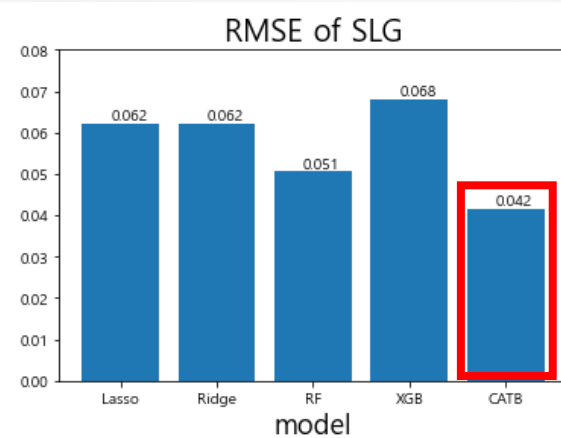
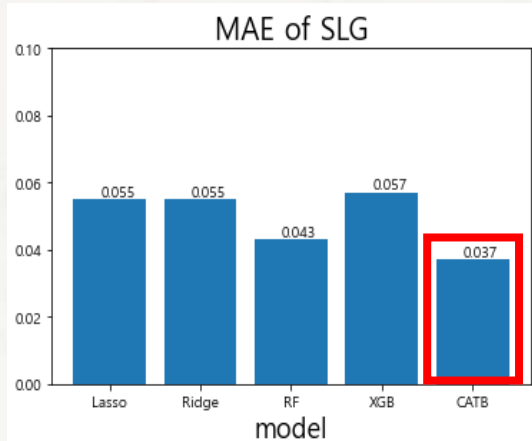


출루율(OBP) 모델로 CatBoost 선정

6. 분석 결과

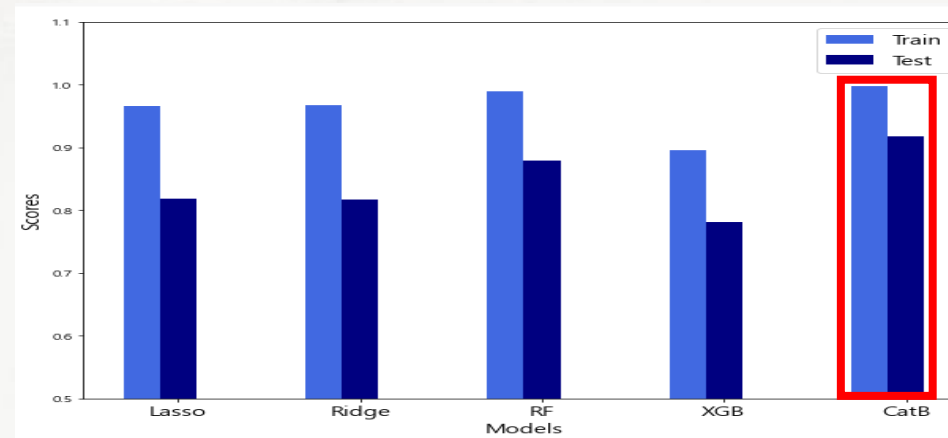
분석결과 - 장타율

모델 별 성능 비교



모델	RMSE	MAE
Lasso	0.032	0.022
Ridge	0.032	0.022
RandomForest	0.034	0.022
XGBoost	0.049	0.035
CatBoost	0.035	0.015

모델 별 정확도 비교



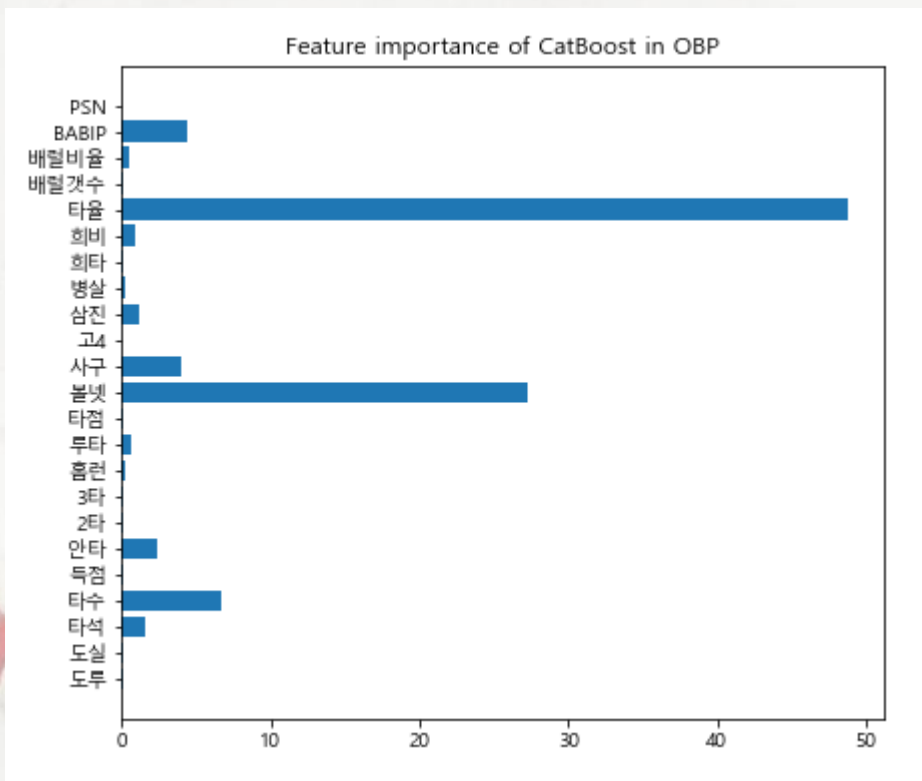
모델	Train Accuracy	Test Accuracy
Lasso	0.966	0.817
Ridge	0.966	0.817
RandomForest	0.989	0.879
XGBoost	0.895	0.781
CatBoost	0.997	0.917

장타율(SLG) 모델로 CatBoost 선정

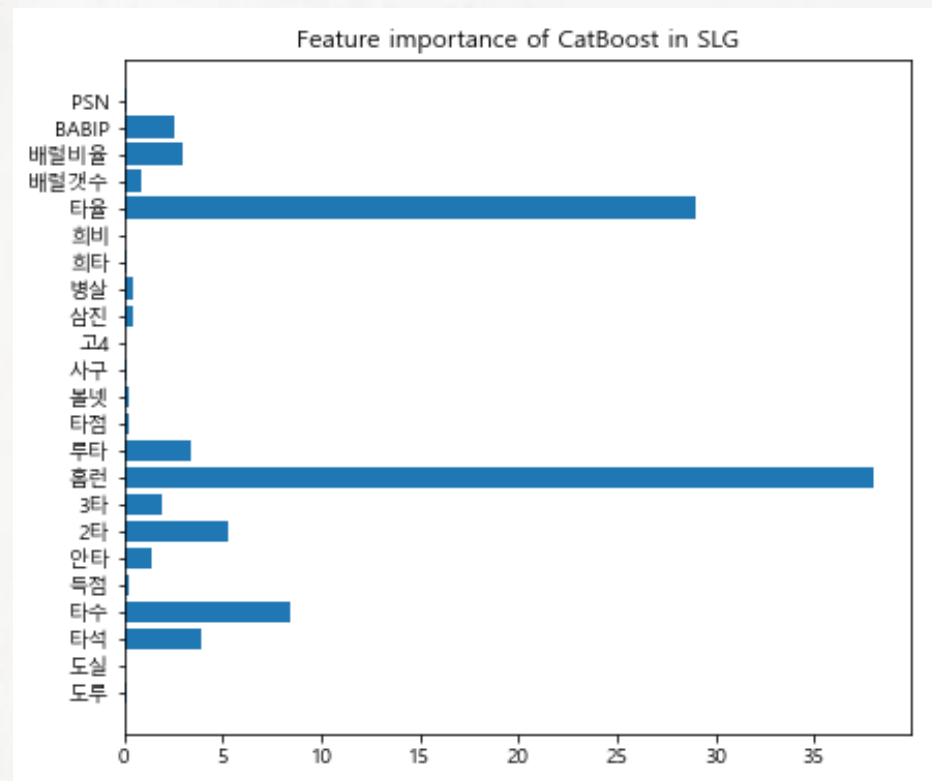
6. 분석 결과

출루율, 장타율 변수 중요도

출루율(OBP) – CatBoost 모델



장타율(SLG) – CatBoost 모델



6. 분석 결과

최종 예측 결과

예측 데이터 입력

```
pcode = [76232, 68050, 75847, 67341, 79192, 78224, 78513, 76290, 79215, 67872]
final_test = hitter[hitter['선수코드'].isin(pcode)].groupby('선수코드').mean()
final_test = final_test.reindex(index = [76232, 68050, 75847, 67341, 79192, 78224, 78513, 76290, 79215, 67872])
final_test
```

	Unnamed: 0	시즌	month	도루	도실	타석	타수	득점	안타	2타	...	홈타	희비	타율	출루율	장타율	OPS	배달갯수	배달비율	BABIP	PSN
선수코드																					
76232	1062.370370	19.370370	6.703704	0.629630	0.185185	69.592593	60.037037	10.814815	20.629630	3.851852	...	0.037037	0.814815	0.347704	0.420741	0.586704	1.007667	8.111111	12.002222	0.340078	0.784829
68050	1062.538462	19.384615	6.576923	1.000000	0.576923	81.230769	70.538462	12.769231	23.192308	4.923077	...	0.000000	0.692308	0.331000	0.408115	0.551769	0.959923	7.346154	9.640038	0.373985	0.995604
75847	1104.461538	19.423077	6.538462	1.076923	0.423077	78.846154	64.500000	13.153846	17.653846	3.038462	...	0.038462	0.961538	0.266808	0.391077	0.531808	0.922923	8.000000	9.417308	0.278117	1.543258
67341	1066.521739	19.391304	6.478261	1.739130	0.695652	87.086957	76.869565	12.956522	26.043478	5.826087	...	0.347826	0.956522	0.348652	0.413304	0.503826	0.917174	5.608696	5.961696	0.365002	1.589216
79192	943.809524	19.190476	6.238095	0.380952	0.333333	80.238095	72.476190	10.714286	22.809524	3.952381	...	0.000000	1.095238	0.308000	0.362286	0.478857	0.841429	8.428571	10.490190	0.336284	0.620301
78224	1045.629630	19.296296	6.703704	0.444444	0.111111	80.666667	68.962963	11.814815	20.074074	3.481481	...	0.000000	1.037037	0.277000	0.364037	0.492852	0.856889	9.185185	10.285963	0.327846	0.668921
78513	1071.142857	19.357143	6.714286	0.892857	0.571429	82.285714	73.535714	12.964286	22.571429	4.607143	...	0.214286	0.678571	0.291143	0.348893	0.478107	0.827071	7.642857	8.513000	0.303074	1.052165
76290	1104.480000	19.480000	6.360000	0.200000	0.240000	85.520000	75.120000	12.800000	24.240000	5.160000	...	0.000000	1.280000	0.301240	0.373000	0.468200	0.841080	9.080000	9.939440	0.308119	0.148485
79215	1057.600000	19.360000	6.600000	1.320000	0.320000	75.160000	66.360000	12.320000	21.000000	4.360000	...	0.200000	0.920000	0.324280	0.385280	0.462280	0.847560	7.640000	9.878920	0.354326	1.280159
67872	1039.538462	19.307692	6.538462	0.807692	0.384615	83.038462	70.153846	12.384615	19.807692	3.423077	...	0.000000	0.730769	0.282615	0.382885	0.535846	0.918692	8.461538	10.017769	0.303703	1.208408

10 rows × 29 columns

최종 예측 결과

	NO.	PCODE	OPS	장타율	출루율
1	1	76232	1.259256	0.780595	0.478662
2	2	68050	0.921170	0.501902	0.419268
3	3	75847	0.910208	0.512006	0.398202
4	4	67341	0.899348	0.480799	0.418549
5	5	79192	0.652467	0.355960	0.296507
6	6	78224	0.757806	0.429407	0.328399
7	7	78513	0.664444	0.374292	0.290152
8	8	76290	0.699995	0.378212	0.321784
9	9	79215	0.809932	0.437871	0.372061
10	10	67872	0.831693	0.474749	0.356944

선수 코드 별 4년치 데이터

평균 값을 입력



7. 결론 및 느낀점

기대 효과

1. 한국형 배럴타구

한국 실정에 맞는
배럴타구 정의로
다양한 야구 지표 생성

2. 합리적 의사결정

연봉협상, 선수 트레이드,
구단운영 등
다양한 합리적 의사결정
가능

3. 다양할 즐길거리

선수를 평가할 수 있는
지표가 늘어남에 따라
즐길거리 증가

KBO 시장 성장

7. 결론 및 느낀점

느낀점

빅콘테스트 데이터분석을 통해
데이터 분석의 A-Z를 직접 경험해
볼 수 있어서 유익한 기회였던것 같다.
팀원들과 함께 고민하고 다양한 시도를
해볼 수 있어서 앞으로 데이터분석을
하는데 많은 도움이 될 것 같다.



김효림

평소 관심있던 분야의 데이터 분석을
하면서 도메인 지식의 중요성을
깨달았고, 기존의 도메인 지식과
데이터 결과가 상충하는 경우에
의사결정하는 방법을 배울 수
있었습니다. 무엇보다 팀원들과 함께
고민하며 성장한 좋은 경험이었습니다.



변성도

평소 관심있던 분야의 데이터 분석을
하면서 도메인 지식의 중요성을
깨달았고, 기존의 도메인 지식과
데이터 결과가 상충하는 경우에
의사결정하는 방법을 배울 수
있었습니다. 무엇보다 팀원들과 함께
고민하며 성장한 좋은 경험이었습니다.



장원석



8. 향후 계획

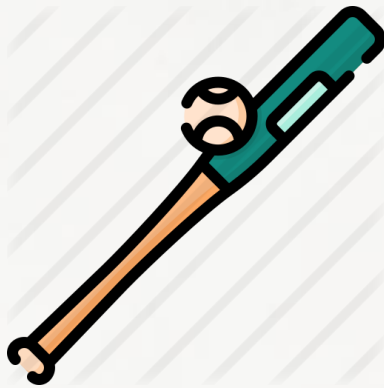
향후 보완점

1. 비거리 추가



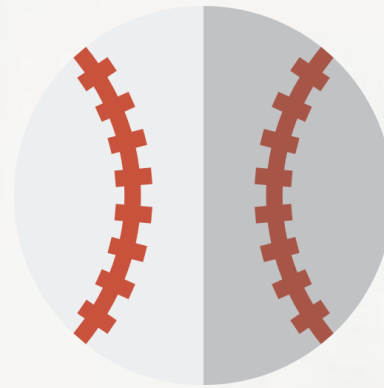
비거리 트래킹 데이터로
더욱 정교한 배럴 타구 정의

2. 다양한 정보 활용



타구 트래킹 데이터에
발사각도, 투구속도, 타격속도 외
날씨, 투구 구질 등
다양한 지표 추가

3. 객관적 지표 추가



사람의 차이로 생기는 변수가 아닌
야구공의 반발계수 등
고정할 수 있는 지표 추가




2021 BIG CONTEST

감사합니다

팀장 김효림

팀원 변성도

팀원 장원석



참고문헌

- 참고 논문 -

- 한국프로야구에서 타자능력의 측정(2014, 이장택)
- 한국 프로야구 타자의 경기력요인 분석(2015, 양도업 외 3인)
- 한국프로야구에서 타자력 지수 제안(2016, 홍종선 외 2인)
- 한국프로야구 타자력 예측모형 개발(2017, 홍종선, 신동식)

- 참고 사이트 -

- 데이터 크롤링 : <http://www.statiz.co.kr/>
- KBO : <https://www.koreabaseball.com/>
- 팬그래프 닷컴 : <https://www.fangraphs.com/>
- 베이스볼 아메리카 : <https://www.baseballamerica.com/>