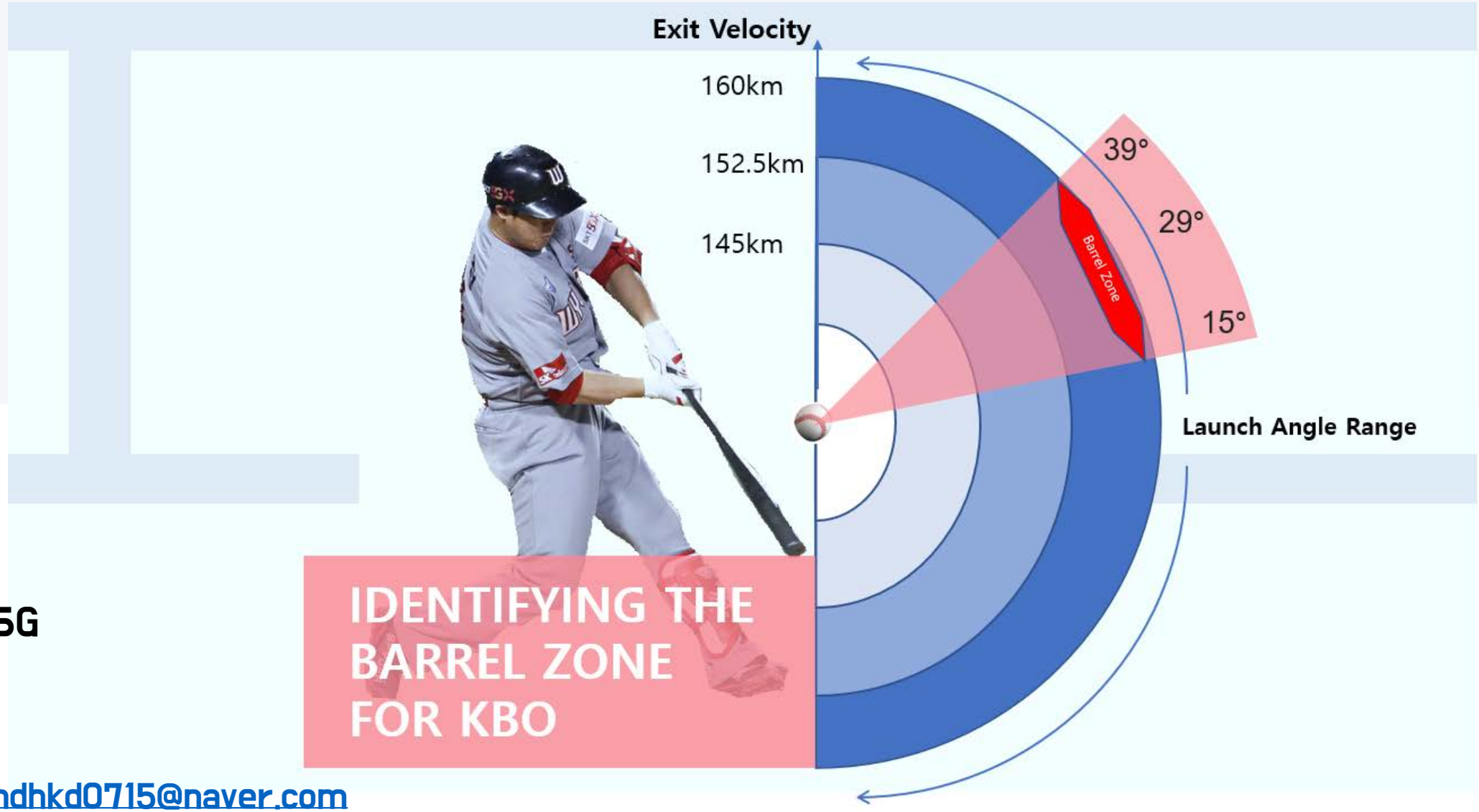
「프로야구 배럴(Barrel)」을 통한 타자 성적 예측



팀명 : BASSG

팀장

- 송민규 dirndhkd0715@naver.com

팀원 - 지윤혁 yhji1127@naver.com, **조영진** zammanbou0825@gmail.com, **백찬진** anback14@gmail.com

INDEX

1-1 문제정의

sports2i
Information & Interest Provider in Sports

스포츠투아이에서 제공하는 야구데이터를 활용하여

① 좋은 타구(배럴)에 대하여 정의

② 타자 성적 예측 모형 개발을 통한 타자의 OPS 예측

1-2 데이터 소개

출처	제공데이터 목록
빅콘테스트 내부데이터	1. 2021 빅콘테스트_데이터분석분야_챔피언리그_스포츠테크_HTS_2018, 2019, 2020, 2021
	2. 2021 빅콘테스트_데이터분석분야_챔피언리그_스포츠테크_경기일정_2021
	3. 2021 빅콘테스트_데이터분석분야_챔피언리그_스포츠테크_선수_2018, 2019, 2020, 2021
	4. 2021 빅콘테스트_데이터분석분야_챔피언리그_스포츠테크_타자 기본_2018, 2019, 2020, 2021
	5. 2021 빅콘테스트_데이터분석분야_챔피언리그_스포츠테크_팀
외부데이터	KBO 기록실 데이터 (https://www.koreabaseball.com/Record/Player/HitterBasic/Basic1.aspx)

1-3 사용 환경

```
Pandas : 1.3.2
Numpy : 1.19.5
Scikit-Learn : 0.23.2
seaborn : 0.11.2
matplotlib : 3.3.4
pycaret : 2.3.3
Python 3.8.8
```

```
fbprophet : 0.7.1
```

(1) 데이터 중복 처리

PIT_ID가 중복되는 Record가 224개 존재.

-> PIT_ID가 시간-분-초로 이루어져 있어 같은 날에 시분초가 다 똑같으면 PIT_ID가 중복.

	GYEAR	G_ID	PIT_ID	PCODE	T_ID	INN	HIT_VEL	HIT_ANG_VER	HIT_RESULT	PIT_VEL	STADIUM
10	2018	20180324HHWO0	180324_145431	62797	HH	3	63.08	-34.8	땅볼아웃	145.41	고척
107	2018	20180324LGNC0	180324_145431	62931	NC	3	108.48	22.9	1루타	145.78	마산

	GYEAR	G_ID	PIT_ID	PCODE	T_ID	INN	HIT_VEL	HIT_ANG_VER	HIT_RESULT	PIT_VEL	STADIUM
758	2018	20180328KTSK0	180328_193206	74215	KT	4	171.67	-3.8	1루타	129.06	문학
813	2018	20180328LGWO0	180328_193206	67341	WO	3	92.80	52.0	플라이	121.20	고척

>> 해결책 : PIT_ID의 각 열을 year, month, day, hour, minute, second로 쪼개어 독립 열로 분리 후 PIT_ID 열 삭제.

(2) 결측치 처리

MONEY 열에서 총 5개의 결측치 확인.

	GYEAR	PCODE	NAME	T_ID	POSITION	AGE_VA	MONEY
64	2020	50802	화이트	SK	내	29	NaN

	GYEAR	PCODE	NAME	T_ID	POSITION	AGE_VA	MONEY
442	2021	67610	김석환	HT	내	22	NaN
457	2021	68069	고명성	KT	내	22	NaN
529	2021	69645	장지수	HT	투	21	NaN
548	2021	71752	김태균	HH	내	39	NaN

》 해결책 : 결측치 있는 행 삭제.

(3) 데이터 전처리

1. 날짜 열처리

- HTS 데이터의 날짜 열을 '년', '월', '일', '시간', '분', '초 ' 로 변환.

2. 데이터 병합

- HTS, Player, Hitter, Team, Calendar 데이터 병합.

	GYEAR	G_ID	PIT_ID	PCODE	T_ID	INN	HIT_VEL	HIT_ANG_VER	HIT_RESULT	PIT_VEL	...	SLG	SF	BB	KK	IB	HP	GD	OBP	OPS	BABIP
0	2018	20181013LTHT0	2018-10-13 18:22:29	76368	HT	5	106.43	18.6	1루타	117.36	...	0.667	0	7	10	0	0	0	0.525	1.192	0.590909
1	2018	20180502HTLT0	2018-05-02 22:09:32	76368	HT	9	154.67	-4.1	2루타	135.06	...	0.667	0	7	10	0	0	0	0.525	1.192	0.590909
2	2018	20180503HTLT0	2018-05-03 21:39:17	76368	HT	9	151.62	-17.8	1루타	145.63	...	0.667	0	7	10	0	0	0	0.525	1.192	0.590909

(1) Kde Plot을 이용한 좋은 타구들의 밀집도 비교

- Kde Plot 사용한 이유

  각 변수별 밀집 분포를 확인하여 비교적 밀집도가 높은 타구들의 범위를 파악 하고자 함.

  이를 통해, 보다 일반화된 '배럴'의 범위를 구하고자 한다.

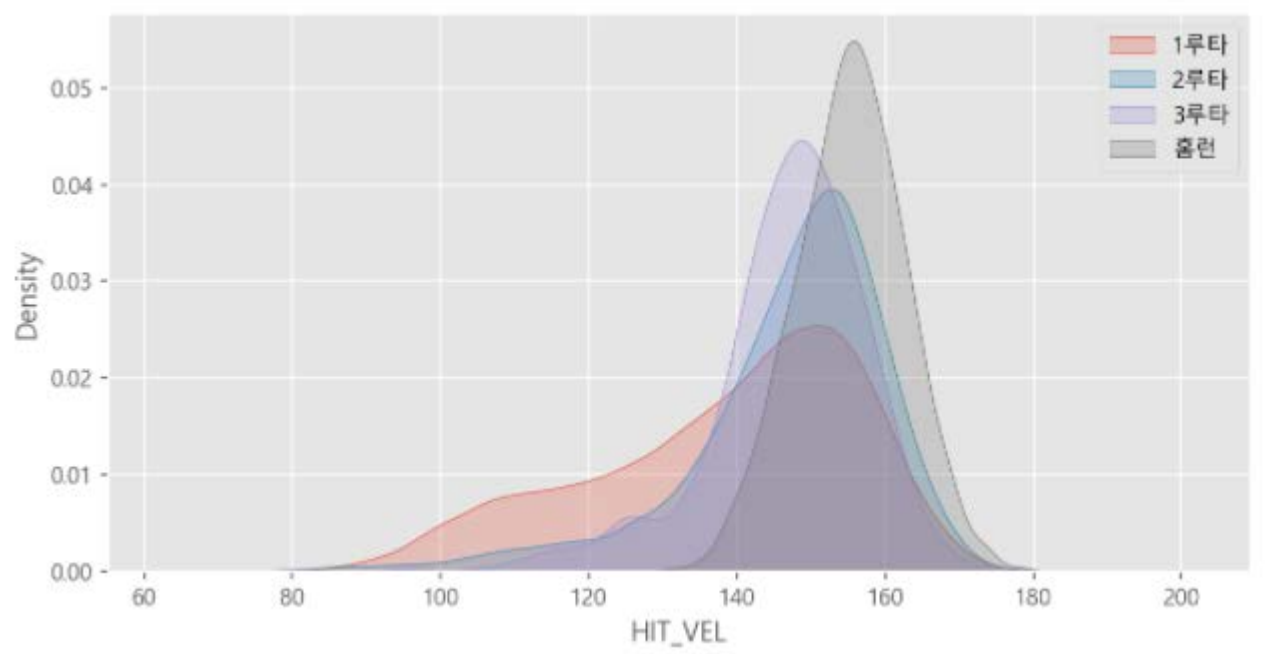
- Kde Plot 사용한 방법

  각 변수별 Kde plot을 시각화 (2차원으로 시각화)

2. 'KBO' 배럴 정의

2-2 좋은 타구들의 특징 분석

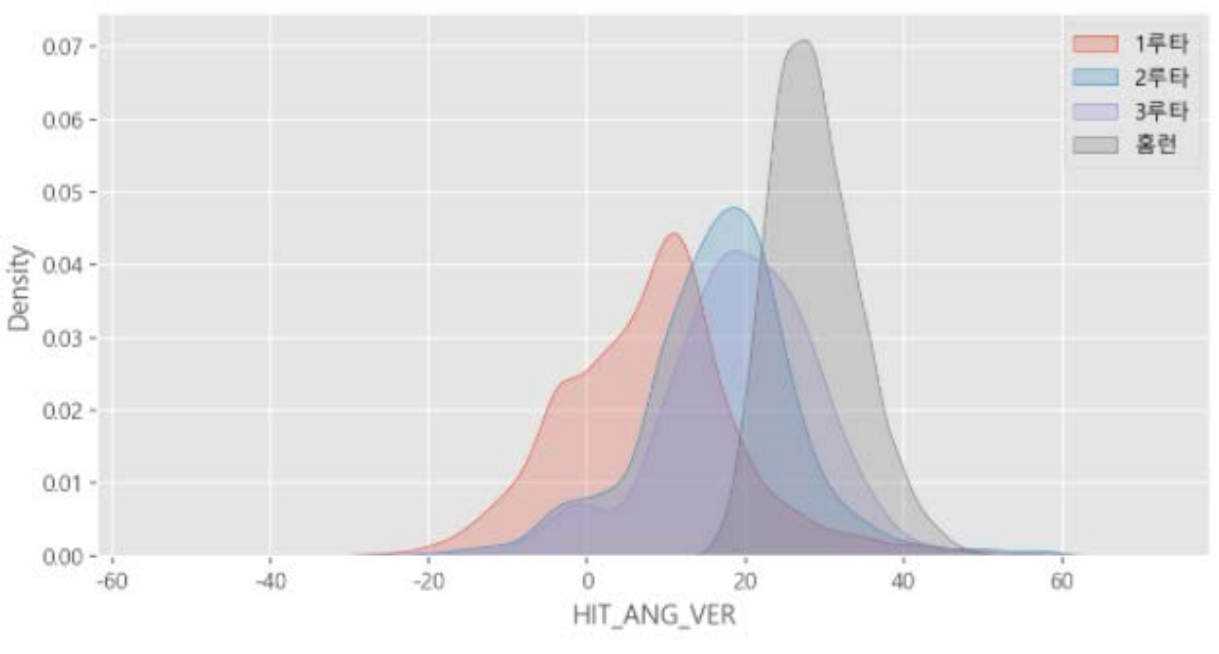
(1) Kde Plot을 이용한 좋은 타구들의 밀집도 비교 - 타구 속도



- 홈런은 타구 속도 150 km/h ~ 160 km/h
- 3루타는 타구 속도 140 km/h ~ 160 km/h
- 2루타는 타구 속도 135 km/h ~ 160 km/h
- 1루타는 타구 속도 130 km/h ~ 160 km/h

2-2 좋은 타구들의 특징 분석

(1) Kde Plot을 이용한 좋은 타구들의 밀집도 비교 - 발사 각도

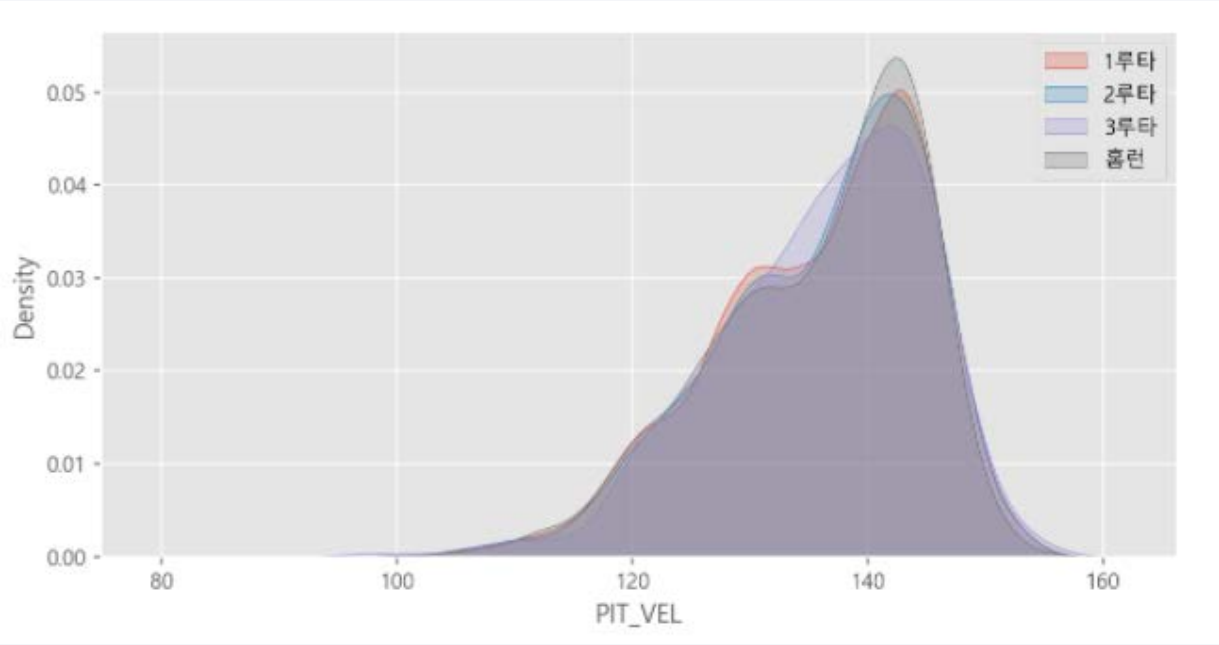


- 발사각 23° ~ 33° 에 밀집
- 발사각 10° ~ 30° 에 밀집
- 발사각 10° ~ 30° 에 밀집
- 발사각 -5° ~ 20° 에 밀집

2. 'KBO' 배럴 정의

2-2 좋은 타구들의 특징 분석

(1) Kde Plot을 이용한 좋은 타구들의 밀집도 비교 - 투구 속도



- 좋은 타구들의 밀집 분포는 투구 속도에서 거의 비슷한 분포를 보이기 때문에 큰 관련이 없다는 것을 알 수 있음.

2-2 좋은 타구들의 특징 분석

(2) 좋은 타구와 투구 속도 간에 관계가 없는 것인지에 대한 검증

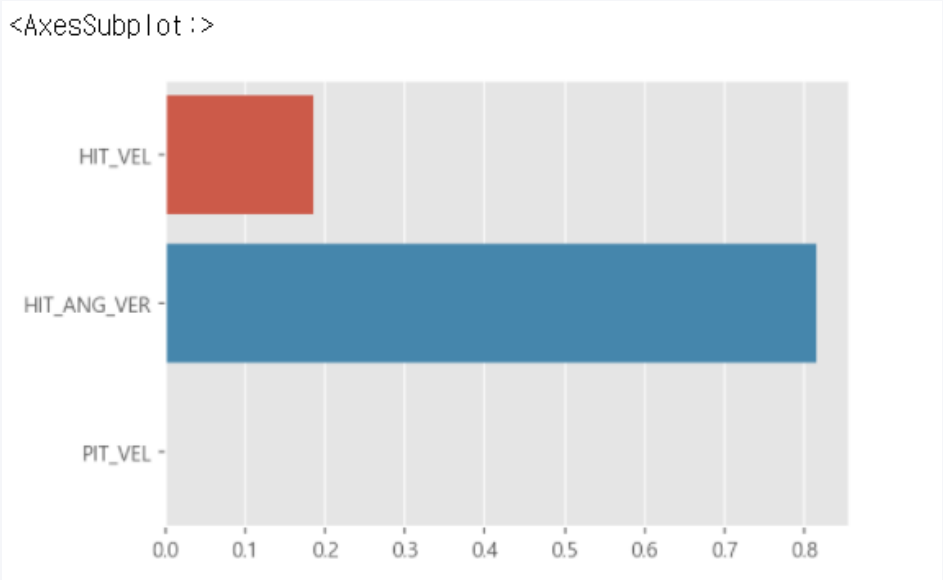
결정트리기반 분류

- 어떤 변수가 중요한지 검증 해 보는 방법으로 결정 트리 기반 분류를 선택

- 좋은 타구를 1, 좋지 않은 타구를 0으로 분류하는 모델을 만들어 모델의 성능을 검증

- 성능과 일반화 능력이 좋은 모델을 선정해 feature importance를 확인

결론 : 투구 속도는 좋은 타구를 분류 하는 데에 있어서 영향이 거의 없다고 판단



2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(1) Kde Plot을 이용한 각 타구별 밀집도 비교 및 배럴 정의

- 투구속도를 제외하고 발사각과 타구속도만으로 3차원 Kde Plot 시각화.

좋은 타구들의 중복 범위를 배럴로 설정.

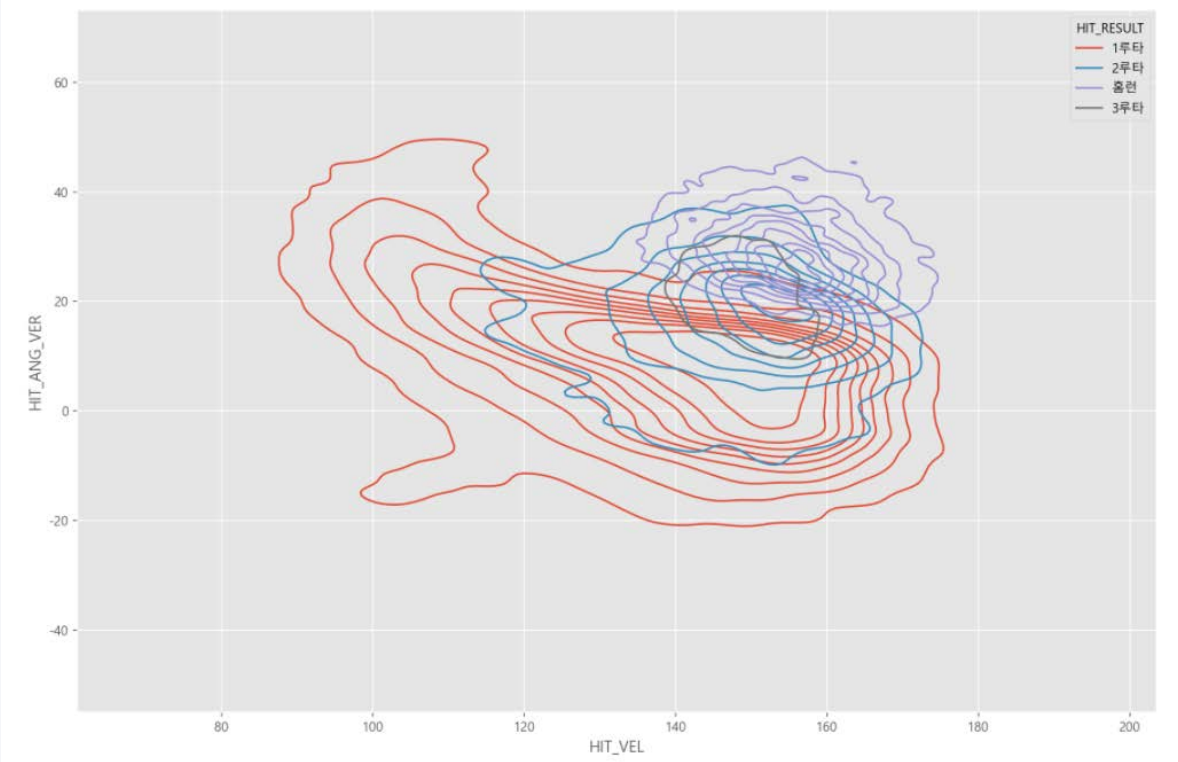
안좋은 타구들의 밀집 분포와 배럴 범위를 비교하여 설정한 배럴이 타당한지 검증.

* 중복 범위는 좋은 타구들이 가지는 공통 범위이기 때문에 이들이 가지는 공통적인 특징이 될 수 있다고 판단

2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(1) Kde Plot을 이용한 각 타구별 밀집도 비교 및 배럴 정의



우선적으로 배럴을 가장 간단하게 정의해보는 방법은

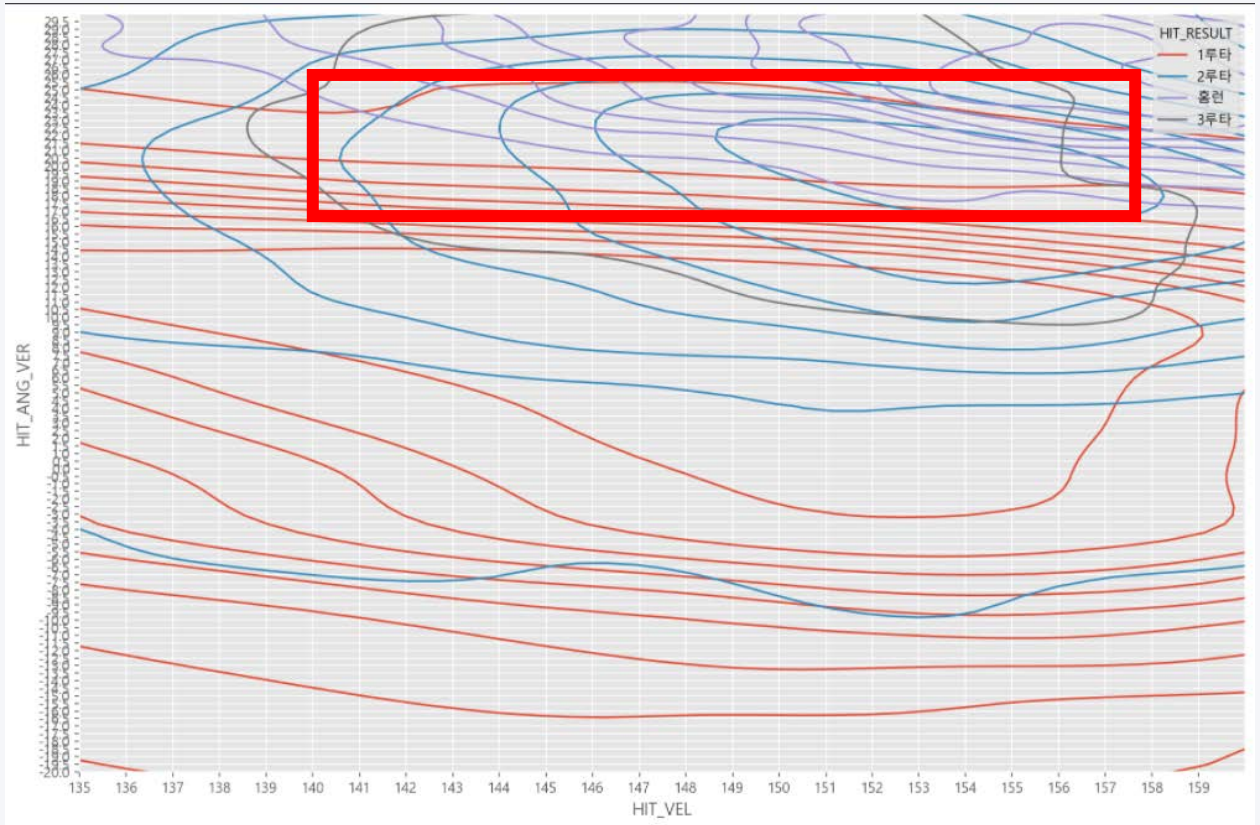
안타로 분류되는 1루타, 2루타, 3루타, 홈런의 밀집된

분포가 모두 겹치는 부분을 배럴로 정의하는 것.

2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(1) Kde Plot을 이용한 각 타구별 밀집도 비교 및 배럴 정의



* 배럴 정의

타구속도 : 140km/h ~ 159km/h.

발사각 : 17.4° ~ 26°를 배럴로 정의.

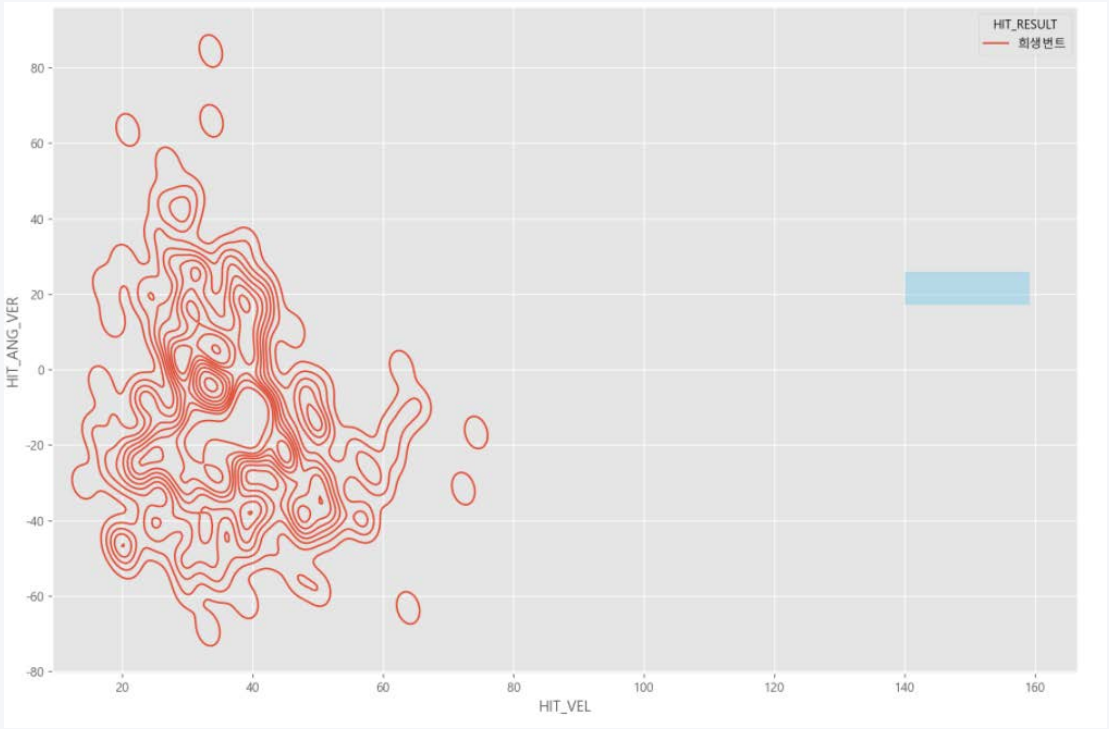
* 검증

정의한 '배럴'의 범위에서 비안타 타구들이 높은 확률로 등장 할 수 있기 때문에 이 범위를 '배럴'로 정의하는 것은 무리가 있음.

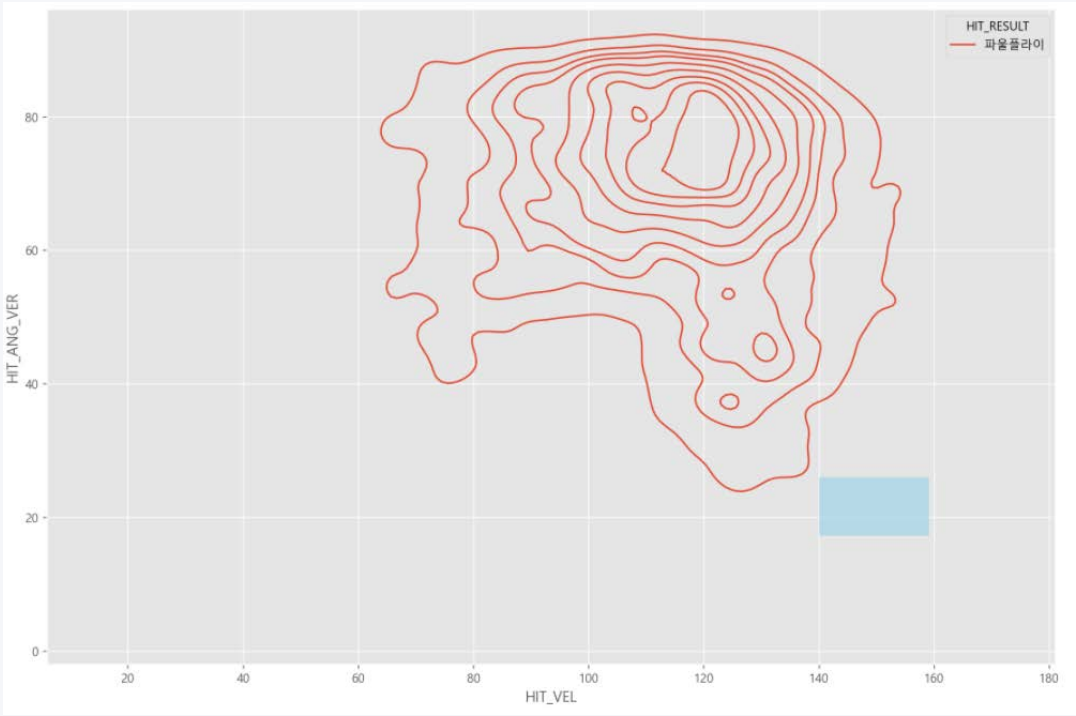
2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(2) 검증 - 범위와 겹치지 않는 비안타 타구들



〈 희생번트 〉

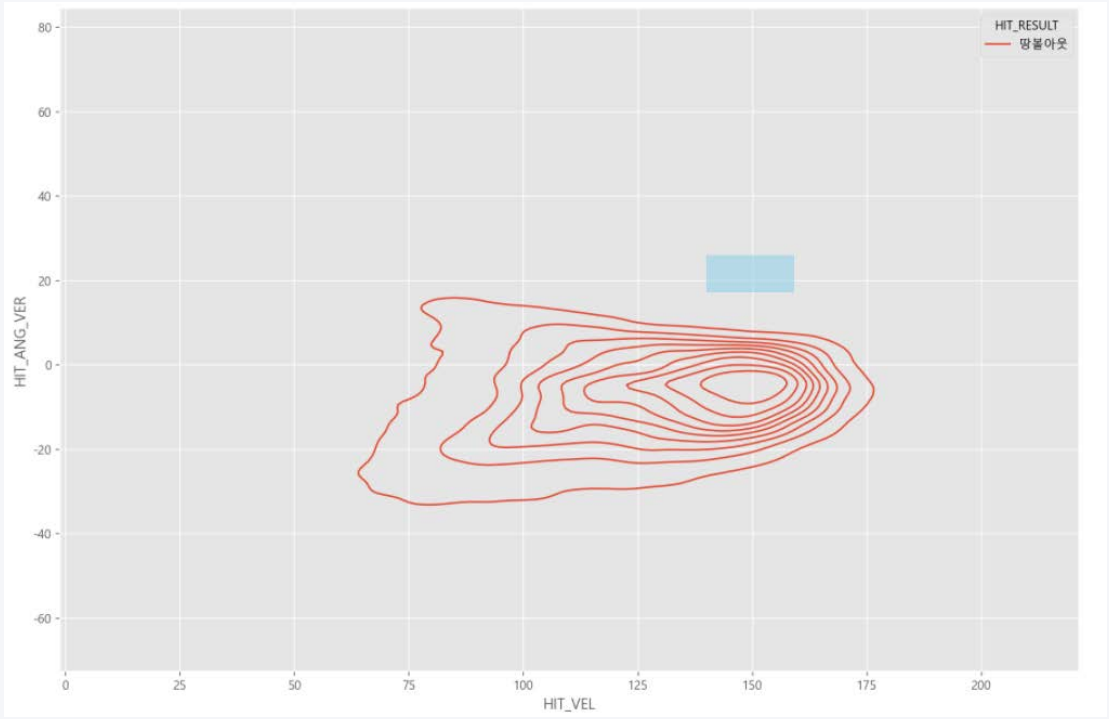


〈 파울플라이 〉

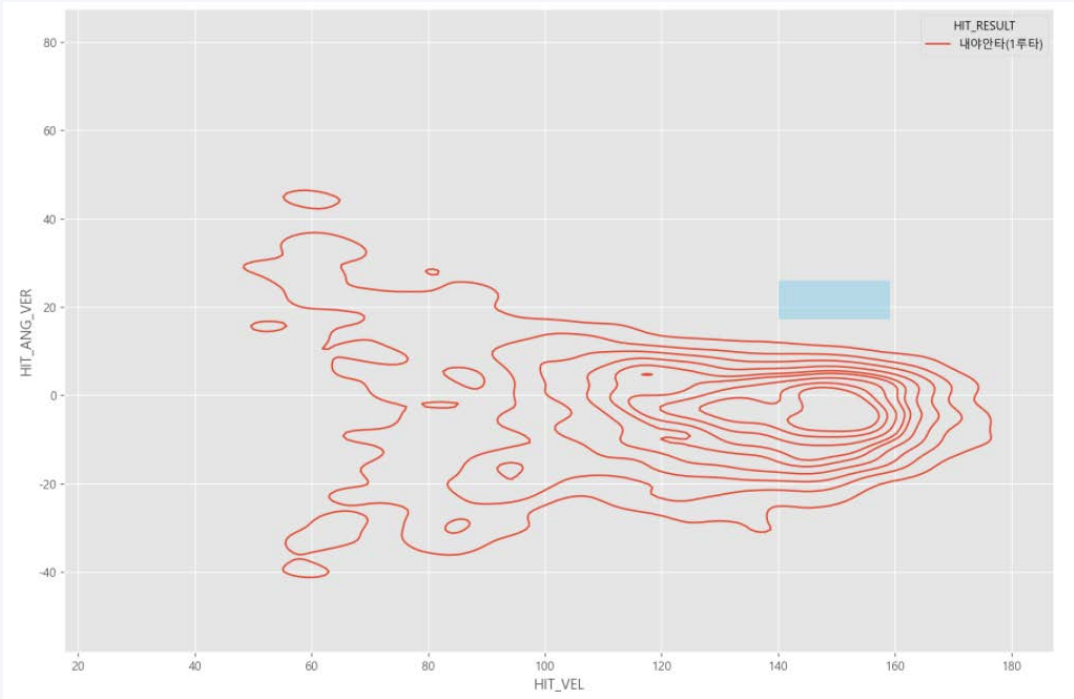
2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(2) 검증 - 범위와 겹치지 않는 비안타 타구들



〈 땅볼아웃 〉

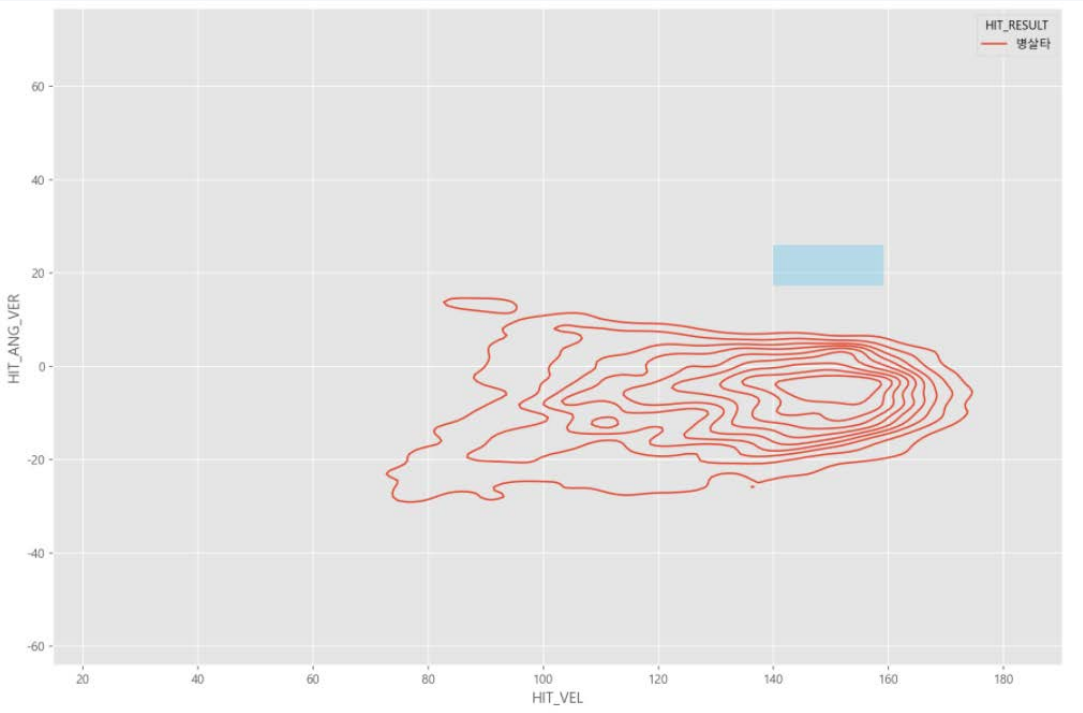


〈 내야안타 (1루타) 〉

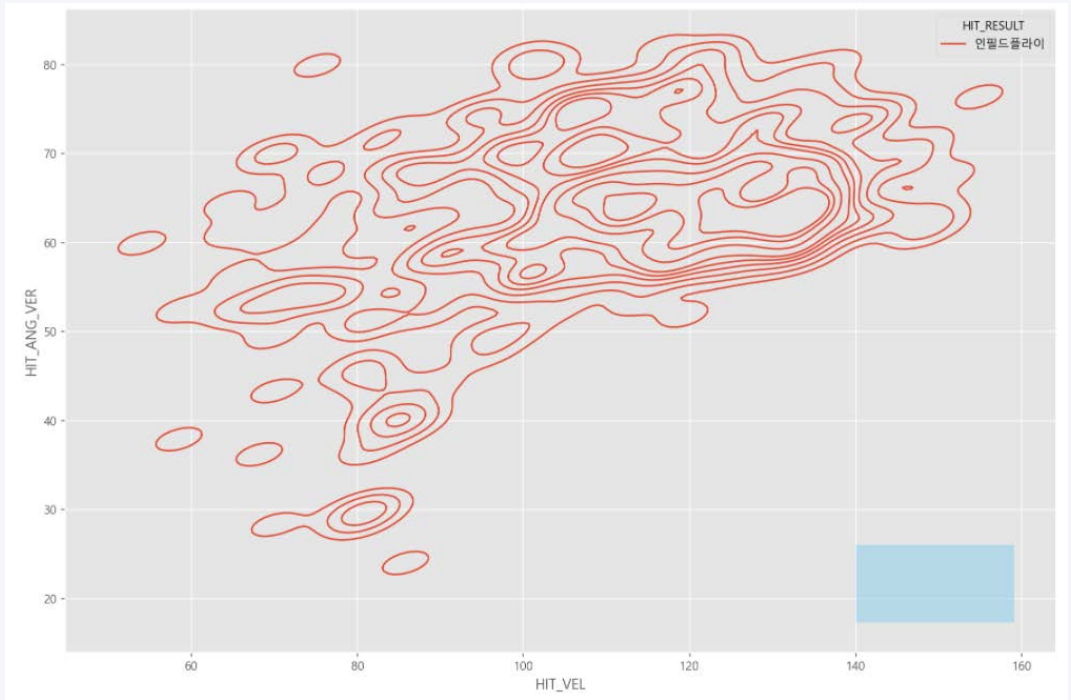
2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(2) 검증 - 범위와 겹치지 않는 비안타 타구들



〈 병살타 〉

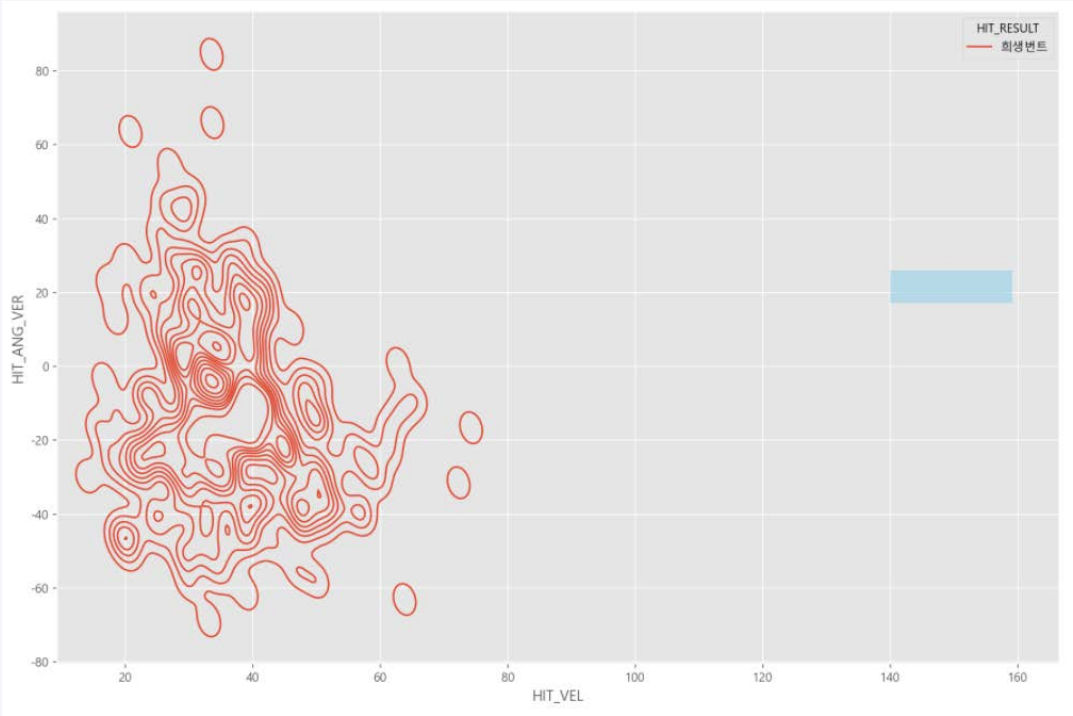


〈 인필드 플라이〉

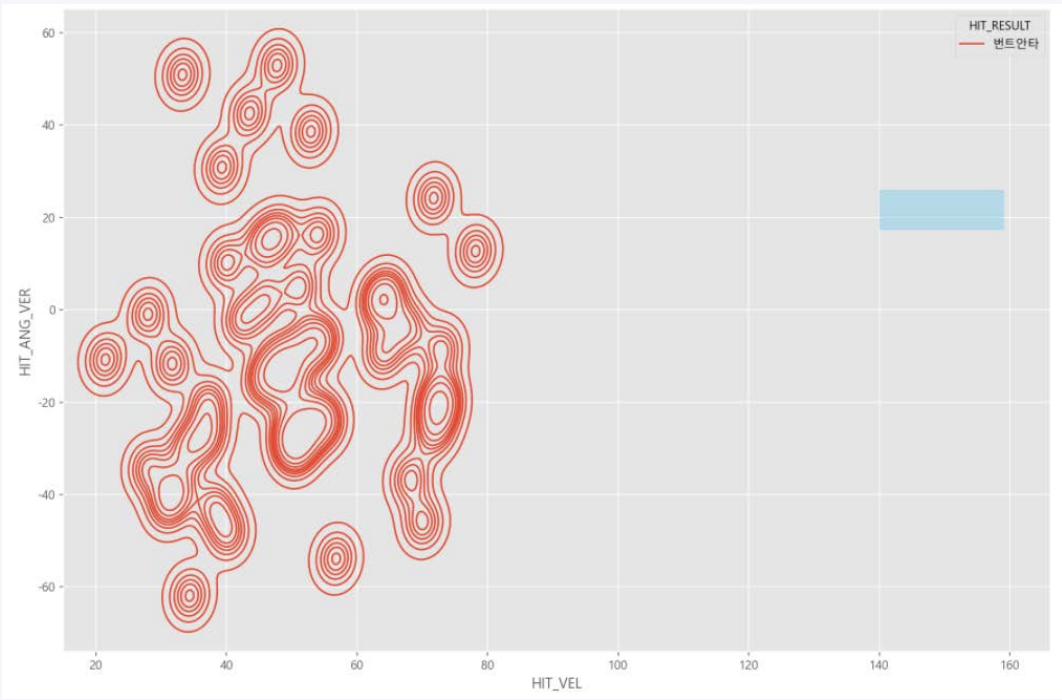
2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(2) 검증 - 범위와 겹치지 않는 비안타 타구들



〈 희생번트 〉

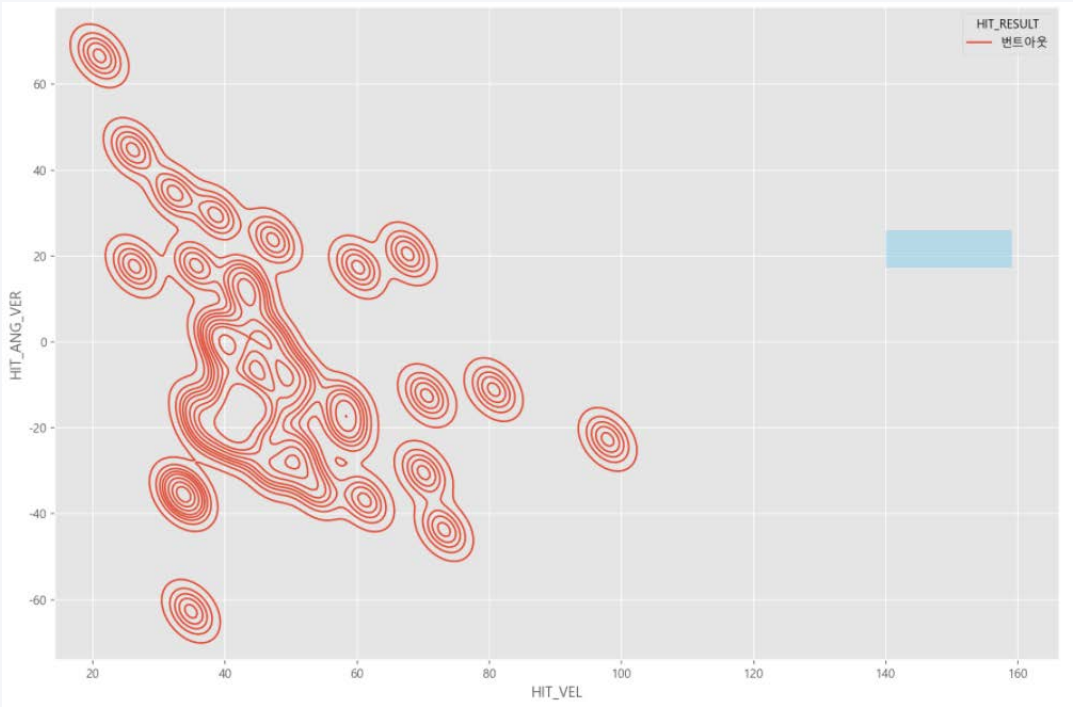


〈 번트안타 〉

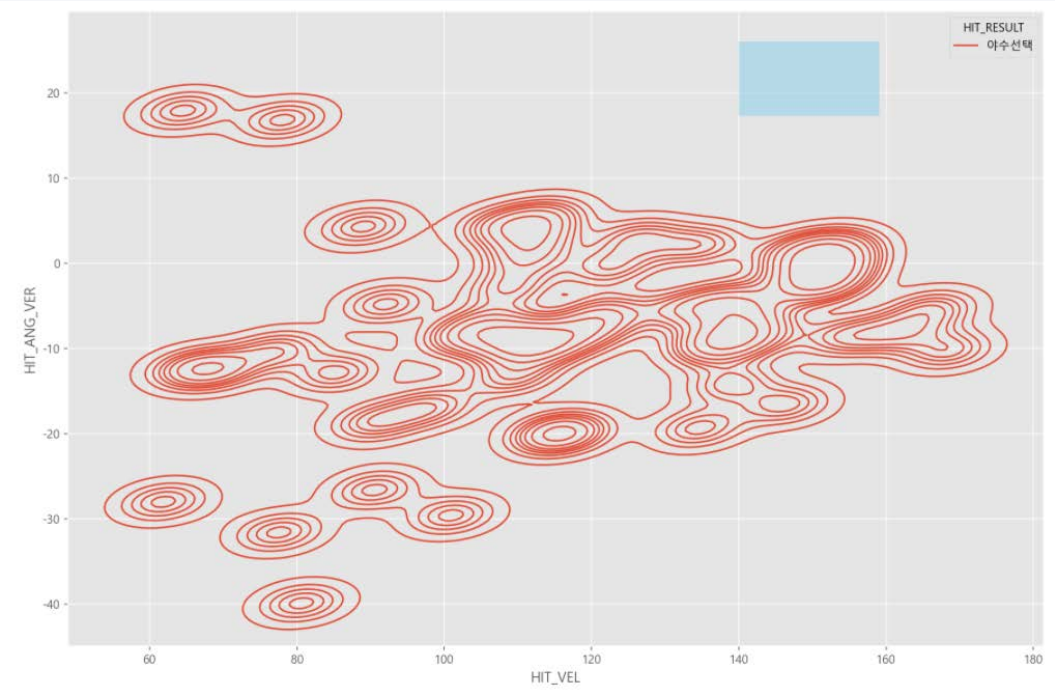
2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(2) 검증 - 범위와 겹치지 않는 비안타 타구들



〈 번트아웃 〉

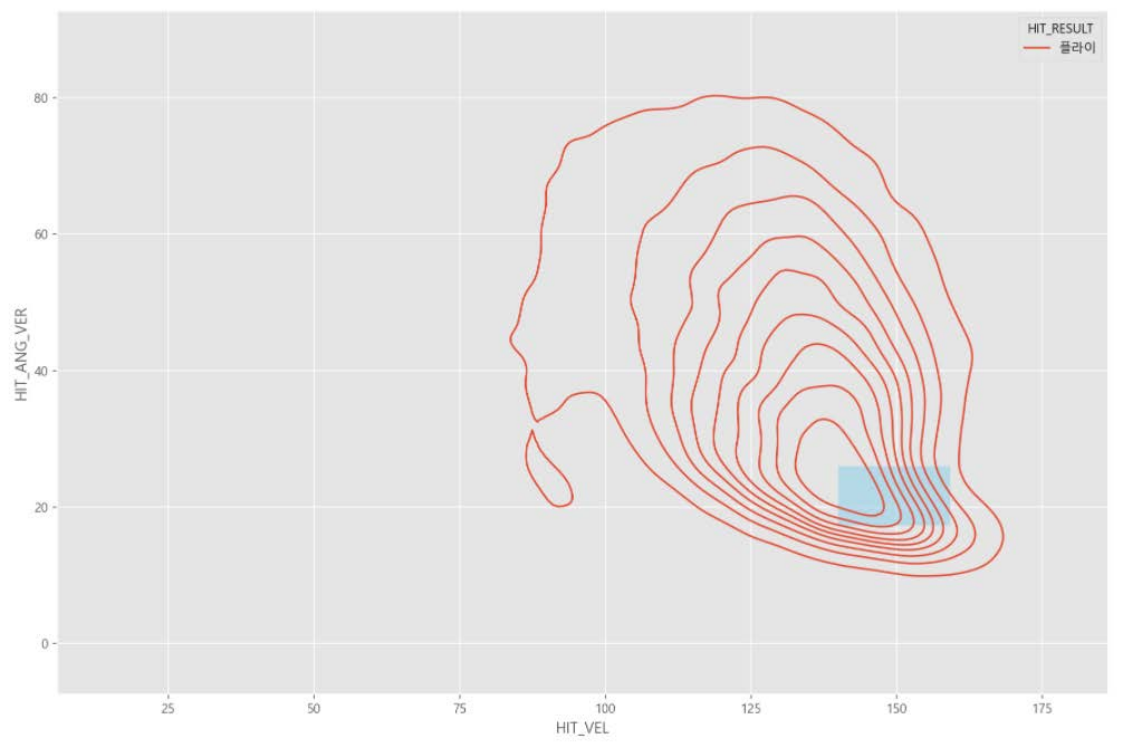


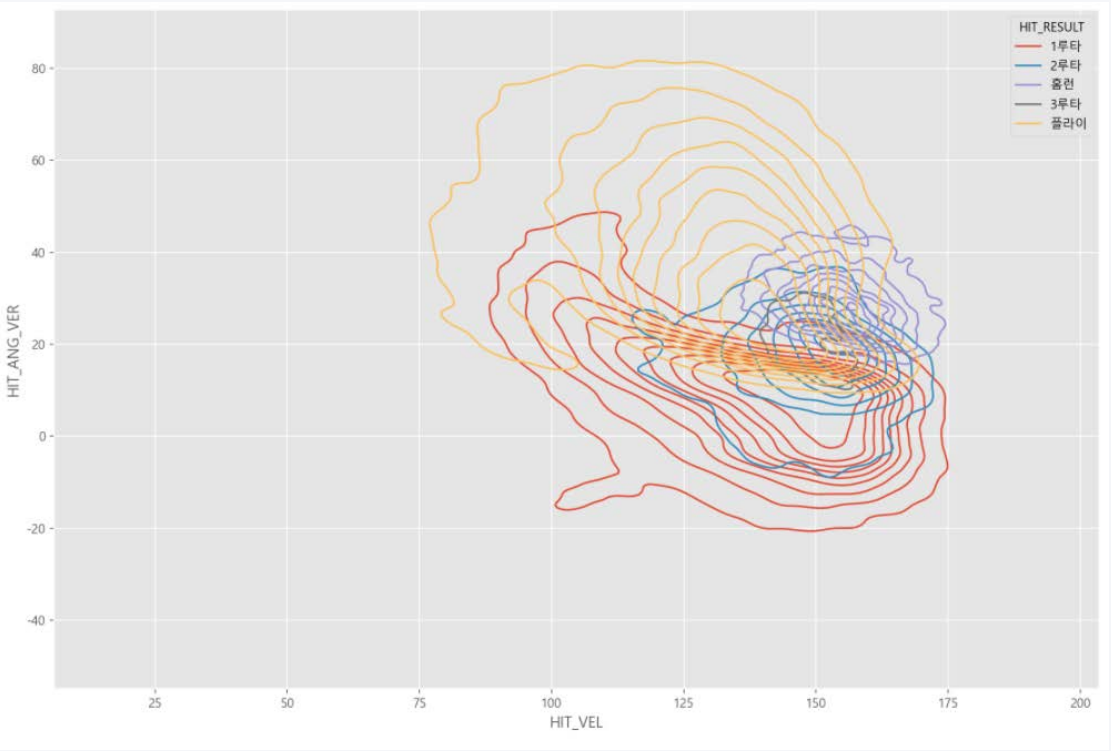
〈 야수선택 〉

2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(2) 검증 - 범위와 겹치는 비안타 타구들





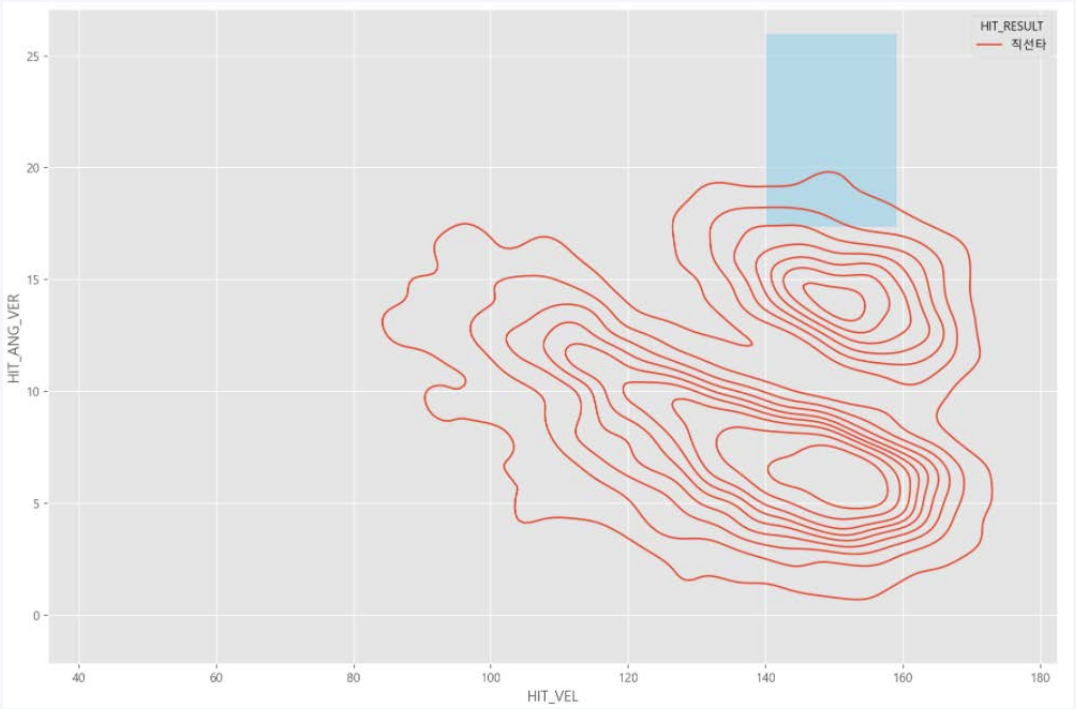
〈 플라이 〉

밀집 분포를 안타 타구 전체의 밀집 분포와 비교해봤을 때 상당 부분 겹치는 것을 알 수 있음.

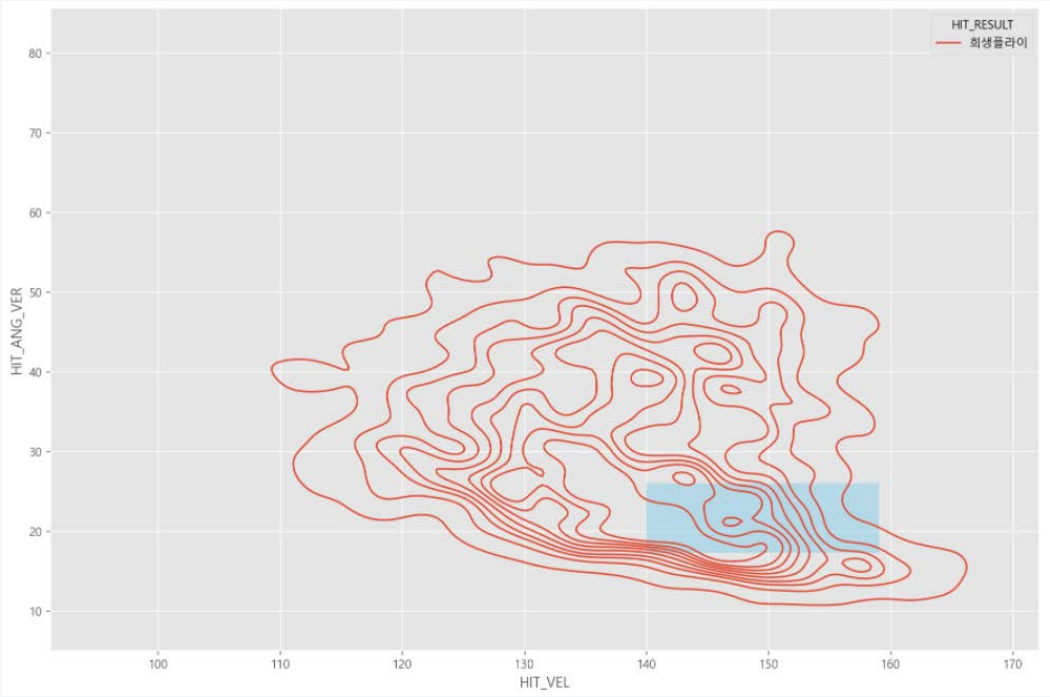
2. 'KBO' 배럴 정의

2-3 KDE plot을 통한 '배럴' 정의

(2) 검증 - 범위와 겹치지 않는 비안타 타구들



〈 직선타 〉



〈 희생플라이 〉

(3) 결론 및 보완과 해결책

- 결론

앞서 정의한 '배럴'은 [플라이], [희생플라이], [직선타]와 상당히 겹치는 모습을 보임.
따라서 앞서 정의했던 범위를 '배럴'로 정의하기에는 근거가 부족.
타구별로 밀집도가 높은 변수들의 범위를 확인해보니, 배럴을 설정할 수 있는 범위를 발견.
-> 이상치 제거의 필요성

- 보완과 해결책

타구들이 가지는 결과의 구체적인 확률값을 구하는 것이 가장 합리적이라고 판단.
이렇게 나온 확률값을 배럴 타구 정의에 사용.

2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(1) 개요

- 확률 추정이란 발사각과 타구 속도와의 조합으로 어떤 타구가 될 것인지에 대한 확률을 추정하는 것.

- 즉, 확률값 = 신뢰도라고 볼 수 있음.

- 구체적으로, 데이터의 발사각, 타구속도를 데이터로 타구 결과를 예측하는 분류 문제로 정의하여 타구 결과에 대한 확률값을 추정.

- 추정된 확률을 바탕으로 어떤 타구가 안타가 될 확률 0.6 이상 그리고 장타가 될 확률이 1.4이상 되는 타구들을 '배럴'로 정의.

- '배럴'들을 scatter plot으로 시각화하여 KBO에 적합한 배럴의 범위를 정의.

(2) 모델링 - IsolationForest를 이용한 이상치 제거

- IsolationForest 사용 이유 : 이상치를 시각적으로 보고 제거하는 방법을 시도해보았으나. 마땅한 경계선을 확정하기 어려움.

- 따라서. 결정트리 모델 기반으로 이상치를 합리적으로 판단하는 알고리즘을 가진 IsolationForest를 사용.

```
1루타 : 이상치 제거 완료(1462 : 5.00%)
2루타 : 이상치 제거 완료(408 : 5.01%)
홈런 : 이상치 제거 완료(237 : 5.01%)
파울플라이 : 이상치 제거 완료(205 : 5.01%)
3루타 : 이상치 제거 완료(34 : 5.06%)
땅볼아웃 : 이상치 제거 완료(1326 : 5.00%)
플라이 : 이상치 제거 완료(1827 : 5.00%)
직선타 : 이상치 제거 완료(180 : 5.00%)
내야안타(1루타) : 이상치 제거 완료(91 : 5.03%)
병살타 : 이상치 제거 완료(155 : 5.02%)
희생플라이 : 이상치 제거 완료(81 : 5.05%)
인필드플라이 : 이상치 제거 완료(17 : 5.14%)
희생번트 : 이상치 제거 완료(12 : 5.02%)
번트안타 : 이상치 제거 완료(3 : 6.25%)
번트아웃 : 이상치 제거 완료(3 : 6.82%)
야수선택 : 이상치 제거 완료(3 : 5.77%)
삼중살타 : 이상치 제거 완료(0 : 0.00%)
```

(2) 모델링 - Pycaret을 이용한 최적의 모델 탐색

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.6765	0.9217	0.2781	0.6106	0.6292	0.5748	0.5801	22.0020
qda	Quadratic Discriminant Analysis	0.6463	0.9061	0.2839	0.5911	0.6069	0.5374	0.5456	0.5420
lda	Linear Discriminant Analysis	0.6055	0.8969	0.2473	0.5539	0.5448	0.4768	0.4895	0.4360
nb	Naive Bayes	0.6224	0.8950	0.2762	0.5369	0.5735	0.5050	0.5127	0.3960

- 발사각과 타구 속도를 통해 어떤 타구 결과를 가질지에 대한 확률을 예측.

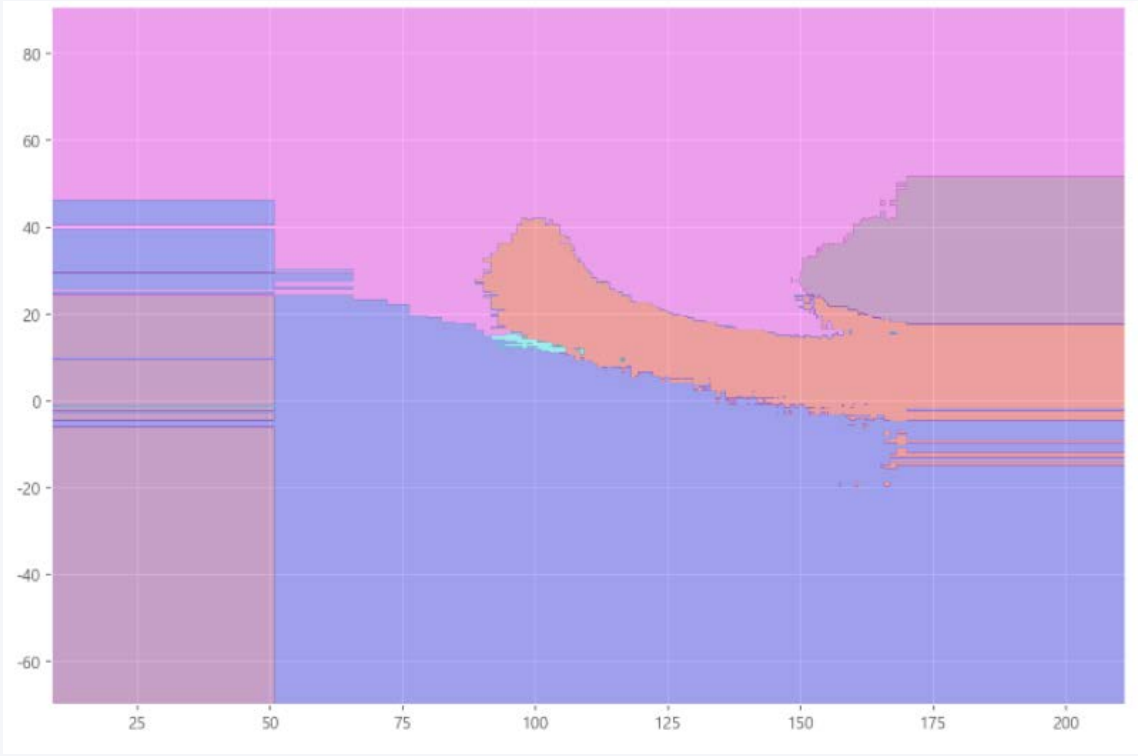
- 가장 성능이 좋은 모델을 뽑기 위하여, AutoML 방식의 Pycaret을 사용.

- Pycaret 결과를 바탕으로
  1. 성능이 좋은 모델과 2. 일반화가 잘 될 것으로 예상되는 모델 4가지를 선정하여 분석.

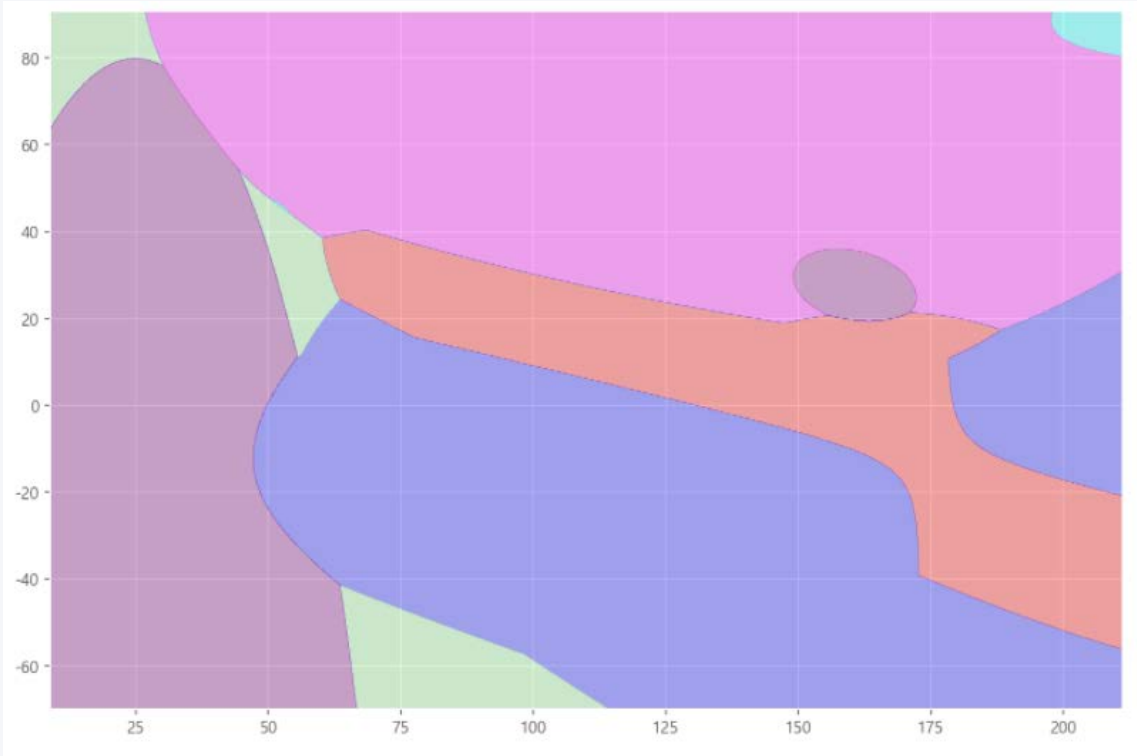
2. 'KBO' 배럴 정의

2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(2) 모델링 - 결정경계 그리기



〈 CatBoost Classifier 〉

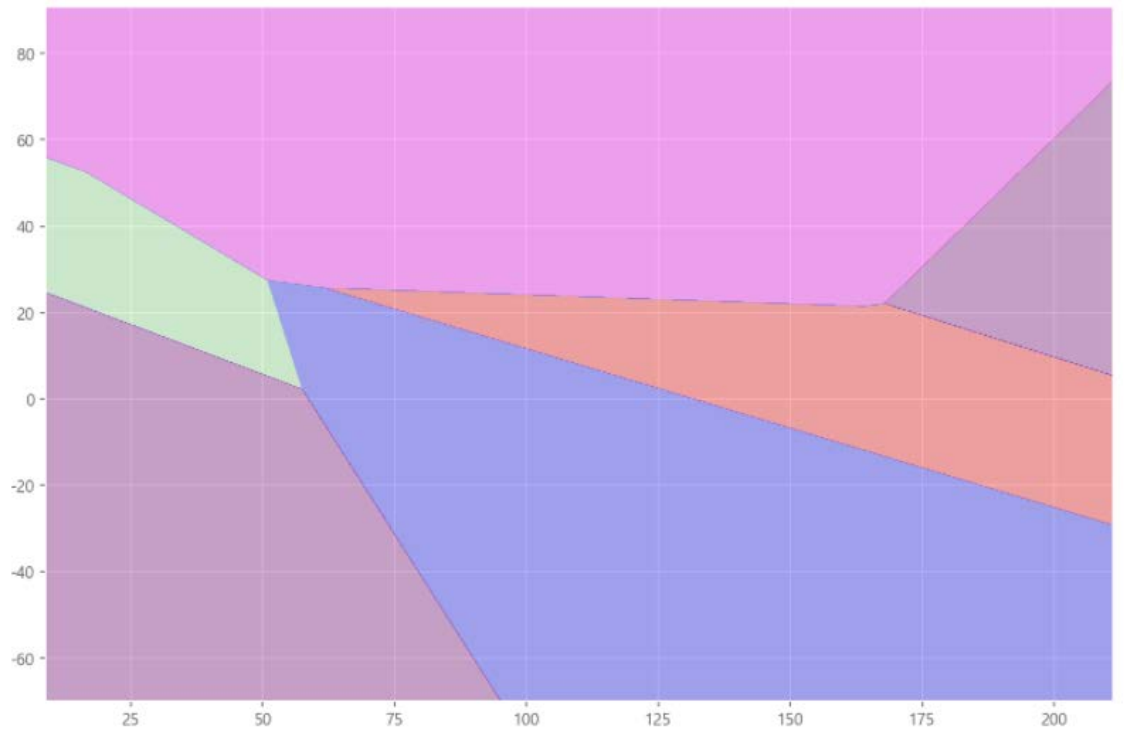


〈 Quadratic Discriminant Analysis 〉

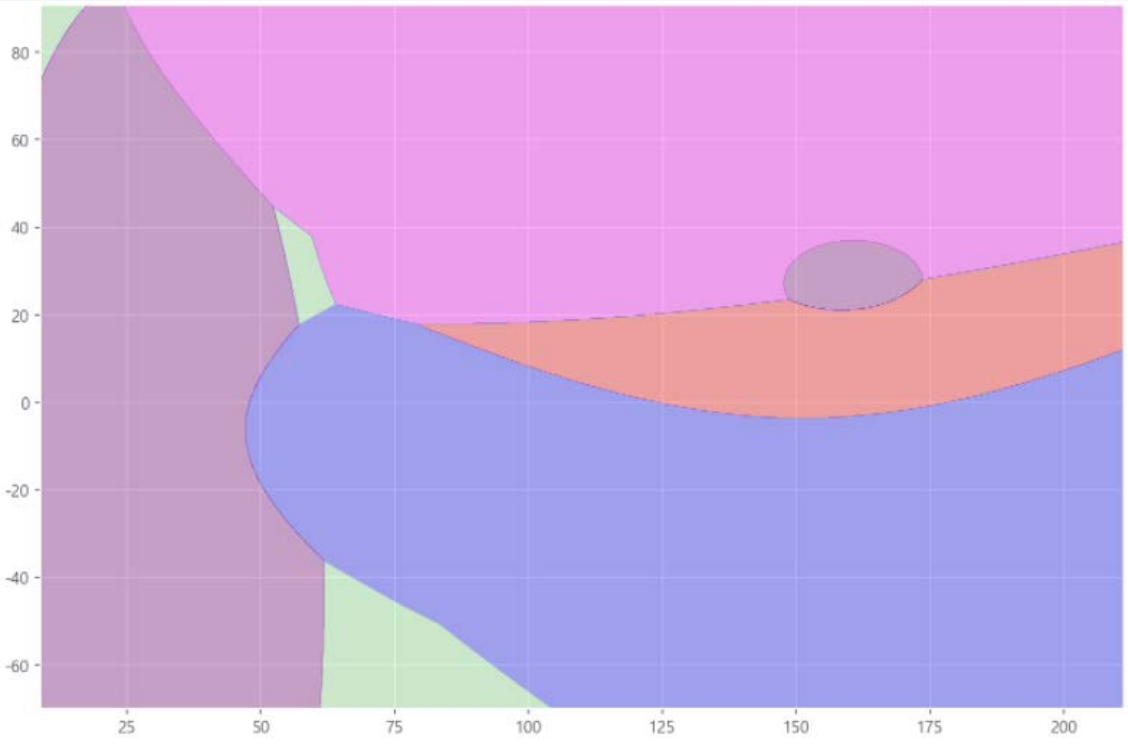
2. 'KBO' 배럴 정의

2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(2) 모델링 - 결정경계 그리기



〈 Linear Discriminant Analysis〉



〈Gaussian Naive Bayes 〉

(2) 모델링 - Gaussian Naive Bayes Model 선정이유

- 과적합 여부를 판단 해 보았을 때, QDA와 GaussianNB가 비교적 smooth한 결정 경계를 보이고 있음.

- 두 가지 모두로 분석을 진행 해 보았으나, 더 적합한 결과를 보이는 GaussianNB를 최종 모델로 선정.

검증

- GaussianNB 모델은 모든 변수가 독립이어야 한다는 전제가 있기 때문에 VIF 다중공선성 여부를 검토.

- 그 결과, 모든 변수의 VIF가 1에 가깝기 때문에 다중공선성이 거의 없다고 판단하였고, 변수 발사각(X1)과 타구속도(X2)가 독립이라는 것을 확인.

OLS Regression Results

Dep. Variable:	label	**R-squared:**	0.379
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	3.677e+04
Date:	Sun, 12 Sep 2021	**Prob (F-statistic):**	0.00
Time:	18:19:29	**Log-Likelihood:**	-3.4116e+05
No. Observations:	120726	**AIC:**	6.823e+05
Df Residuals:	120723	**BIC:**	6.824e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.8428	0.077	63.085	0.000	4.692	4.993
HIT_VEL	-0.0066	0.001	-11.796	0.000	-0.008	-0.005
HIT_ANG_VER	0.1314	0.000	267.159	0.000	0.130	0.132

Omnibus:	19895.297	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5159.149
Skew:	-0.203	**Prob(JB):**	0.00
Kurtosis:	2.072	**Cond. No.**	891.

변수 x1 VIF : 1.0168968610690836
변수 x2 VIF : 1.0168968610690834

2. 'KBO' 배럴 정의

2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(3) 확률기반 '배럴' 정의

- 표

- 2018년에는 KBO 공인구의 반발계수가 높아 OBP와 SLG가 비정상적으로 높은 모습을 보이기 때문에 2018년도 KBO기록은 제외

KBO	OBP	SLG
2018	0.286	0.45
2019	0.267	0.385
2020	0.273	0.409
2021	0.260	0.389
평균	0.272	0.408
2018 제외 평균	0.266	0.394

- 타율을 0.6, 장타율을 1.4로 조정한 이유

- 2018년도의 기록을 제외한 KBO의 OBP와 SLG의 평균은 각각 0.266, 0.394

- 이는 MLB와 비교 해 보았을 때, OBP는 약 8% 높고, SLG는 약 6% 낮음.

- 따라서, 오차 범위와 계산의 편의성을 고려하여 조정량을 10%로 결정.

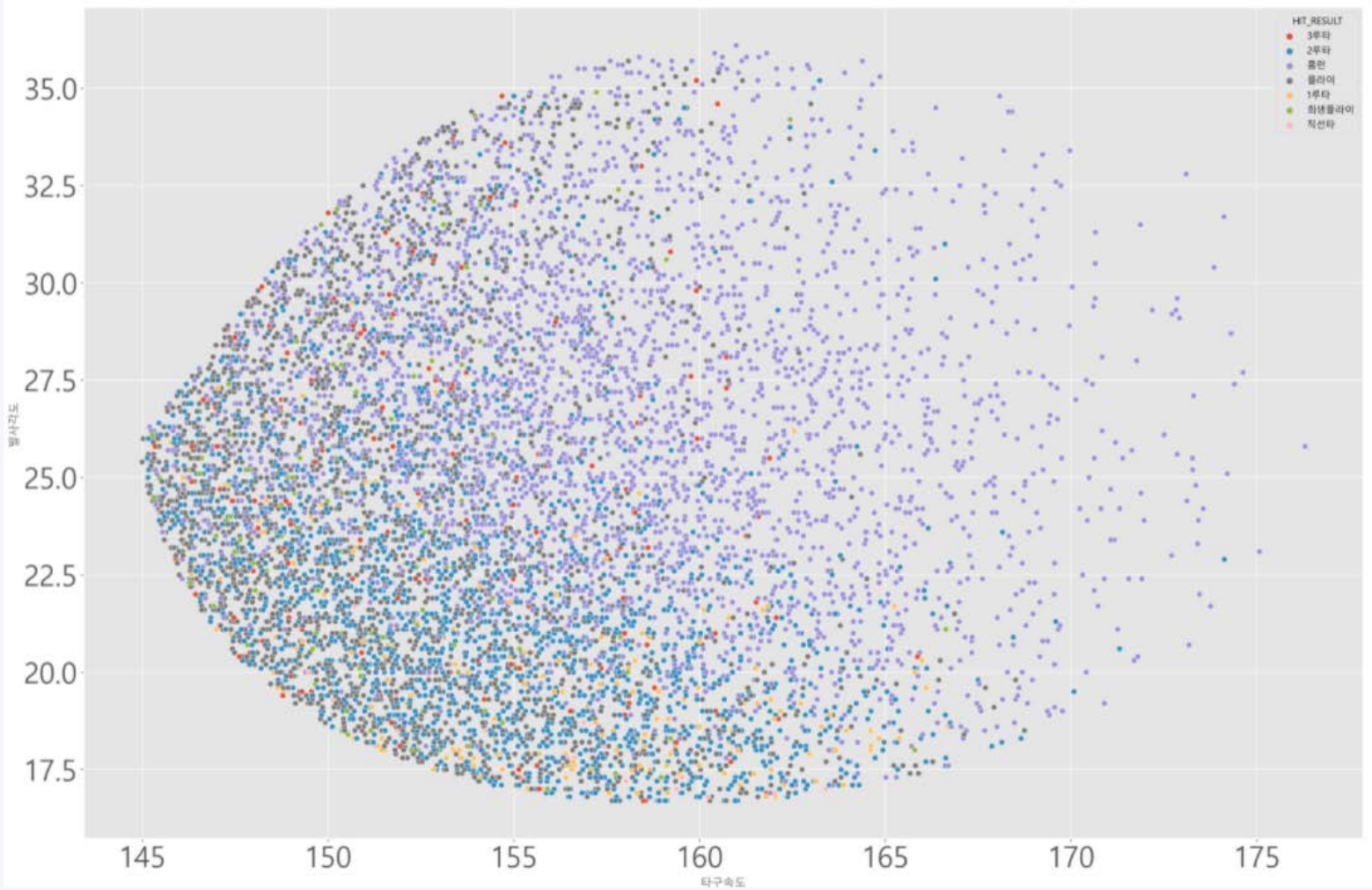
MLB	OBP	SLG
2018	0.248	0.409
2019	0.252	0.435
2020	0.245	0.418
2021	0.243	0.409
평균	0.247	0.418

출처
KBO 기록실 (https://www.koreabaseball.com/Record/Player/HitterBasic/Basic1.aspx)
BASEBALL REFERENCE (https://www.baseball-reference.com/leagues/majors/bat.shtml)

2. 'KBO' 배럴 정의

2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(3) 확률기반 '배럴' 정의 - 배럴타구의 scatterplot



- 타율과 장타율을 조정한 결과에 해당하는 모든 타구의 산점도를 그림.

- 이 범위에 해당하는 모든 타구를 '배럴'로 정의.

- 배럴임에도 불구하고 호수비로 인하여 아웃으로 기록된 타구들도 존재하기 때문에 희생플라이, 직선타들 또한 ' 배럴 ' 이라고 판단.

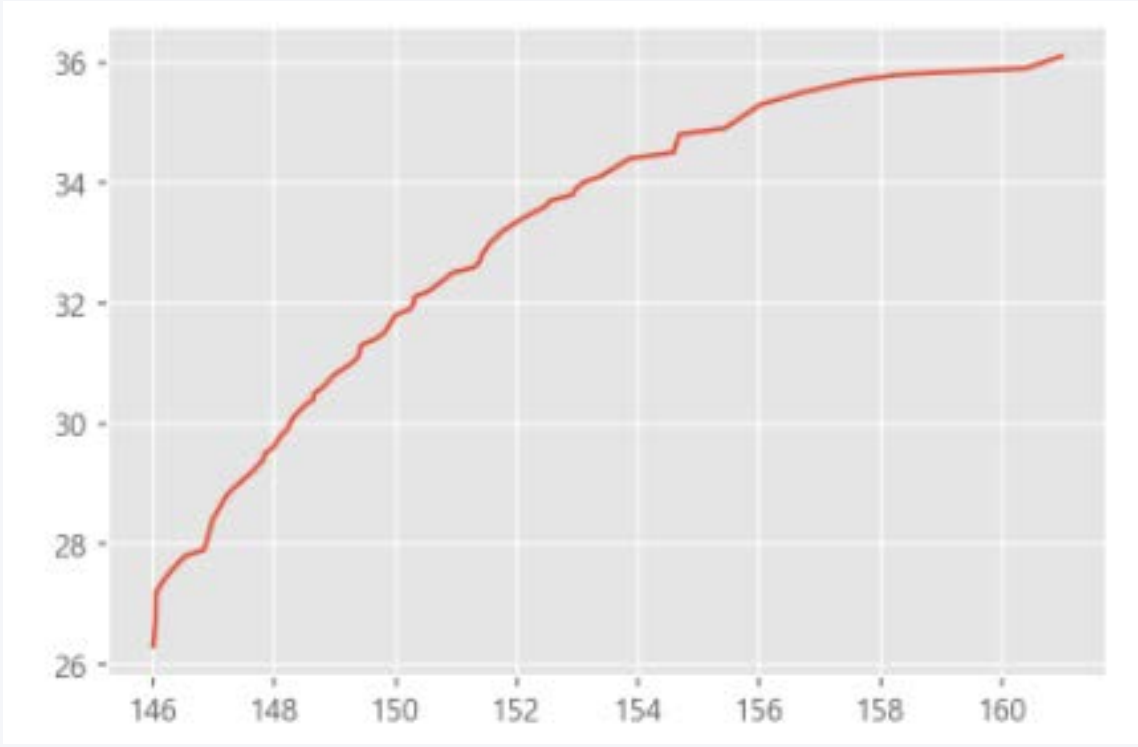
(3) 확률기반 '배럴' 정의 - 경계 범위 설정

- 위 그래프의 값들을 전부 포함할 수 있는 배럴의 범위를 설정하기 위하여 경계선을 설정할 필요가 있음.

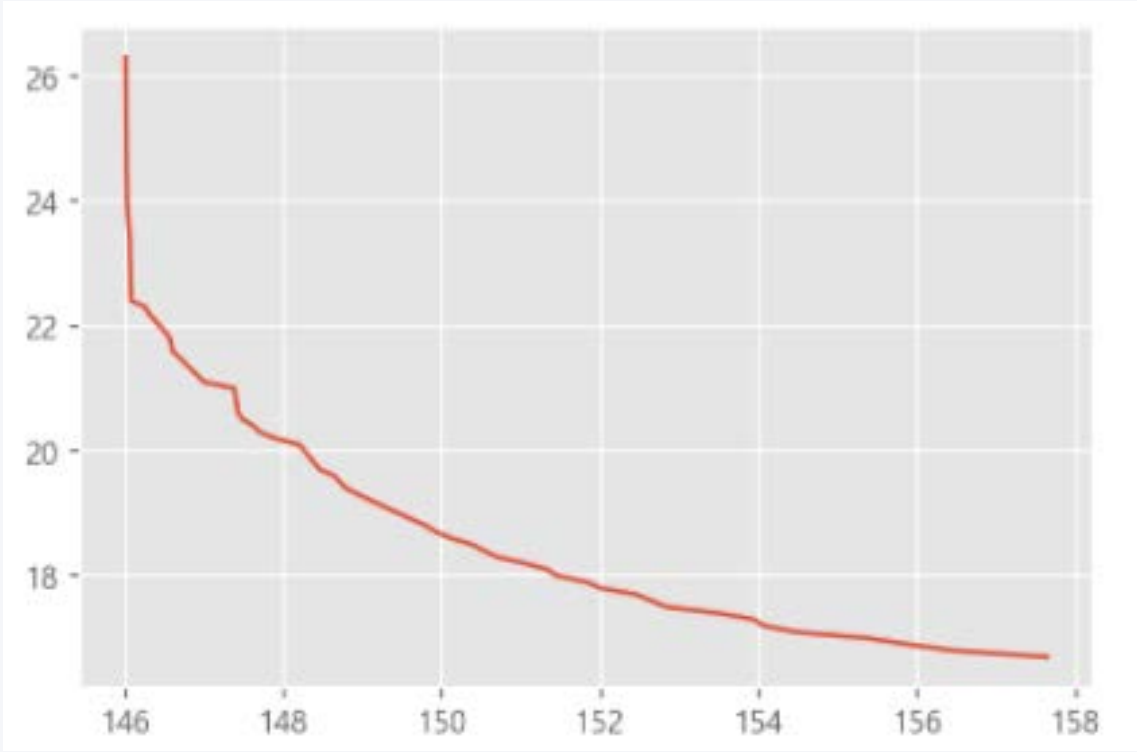
- 위쪽 그래프의 경우 속도가 늘어날수록 발사각이 높아지는 모습을 보이고 아래쪽 그래프의 경우 속도가 늘어날수록 발사각이 작아지는 모습을 보이기 때문에 위와 아래, 두 개로 나누어서 경계선을 그림.

(3) 확률기반 '배럴' 정의 - 경계 범위 설정 (경계선 구하기)

- 위쪽 경계선



- 아래쪽 경계선



2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(3) 확률기반 '배럴' 정의 - 경계 범위 설정 (경계선의 이차식 구하기)

- 위쪽 경계선

Dep. Variable:	HIT_ANG_VER	**R-squared:**	0.995
Model:	OLS	**Adj. R-squared:**	0.995
Method:	Least Squares	**F-statistic:**	5629.
Date:	Sun, 12 Sep 2021	**Prob (F-statistic):**	1.88e-62
Time:	18:19:35	**Log-Likelihood:**	16.509
No. Observations:	56	**AIC:**	-27.02
Df Residuals:	53	**BIC:**	-20.94
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1256.0353	35.451	-35.430	0.000	-1327.141	-1184.929
HIT_VEL	16.2370	0.465	34.888	0.000	15.304	17.171
np.square(HIT_VEL)	-0.0510	0.002	-33.415	0.000	-0.054	-0.048

Omnibus:	35.408	**Durbin-Watson:**	0.593
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	101.220
Skew:	-1.795	**Prob(JB):**	1.05e-22
Kurtosis:	8.523	**Cond. No.**	3.26e+07

- 아래쪽 경계선

Dep. Variable:	HIT_ANG_VER	**R-squared:**	0.911
Model:	OLS	**Adj. R-squared:**	0.907
Method:	Least Squares	**F-statistic:**	209.8
Date:	Sun, 12 Sep 2021	**Prob (F-statistic):**	2.90e-22
Time:	18:19:40	**Log-Likelihood:**	-45.587
No. Observations:	44	**AIC:**	97.17
Df Residuals:	41	**BIC:**	102.5
Df Model:	2		
Covariance Type:	nonrobust		

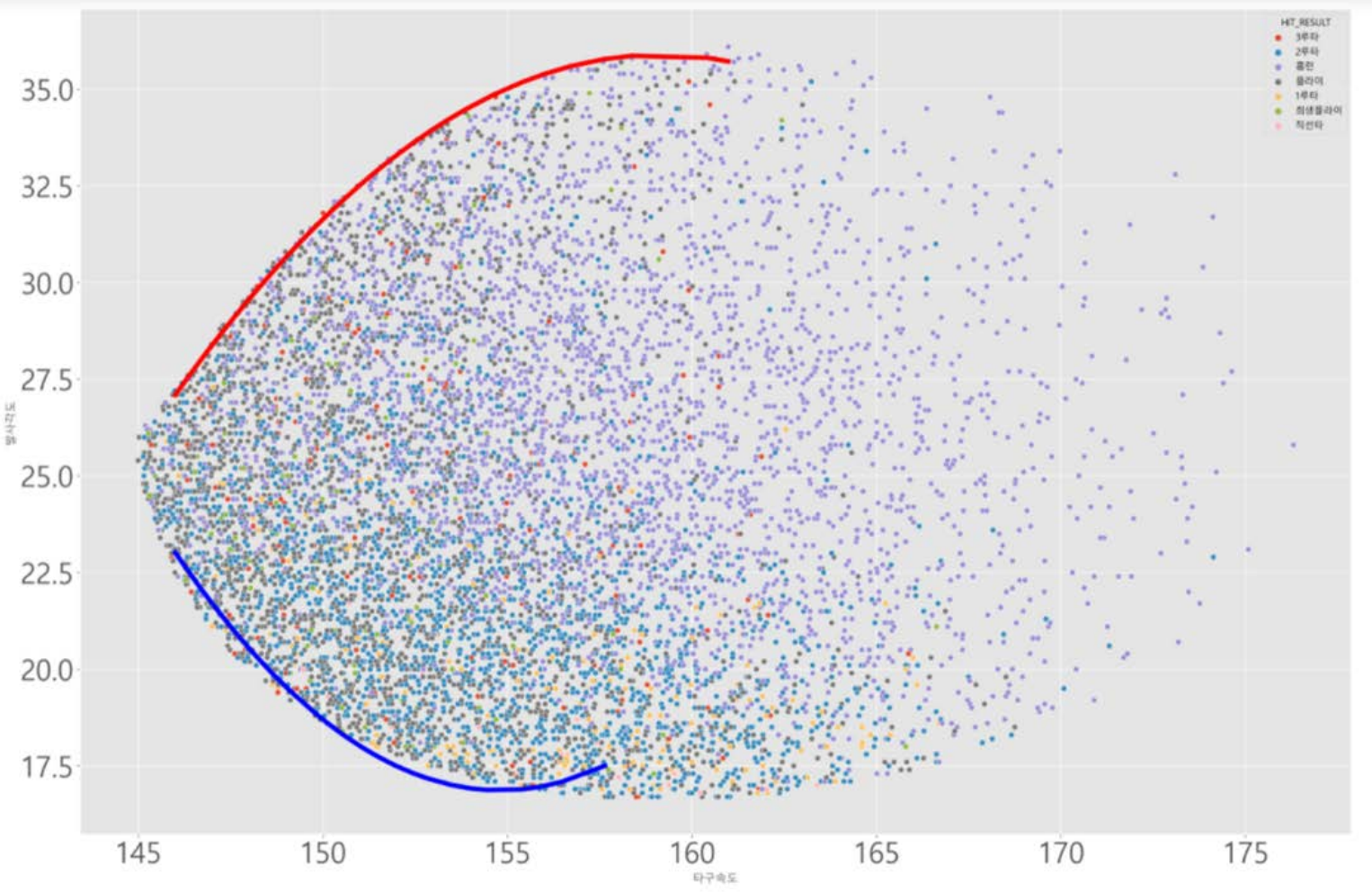
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1917.2246	235.803	8.131	0.000	1441.010	2393.439
HIT_VEL	-24.5514	3.127	-7.851	0.000	-30.867	-18.236
np.square(HIT_VEL)	0.0793	0.010	7.651	0.000	0.058	0.100

Omnibus:	48.854	**Durbin-Watson:**	0.261
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	243.411
Skew:	2.715	**Prob(JB):**	1.39e-53
Kurtosis:	13.163	**Cond. No.**	4.97e+07

2. 'KBO' 배럴 정의

2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(4) 접선의 방정식을 그래프로 그리기 - 이차식의 그래프 그리기

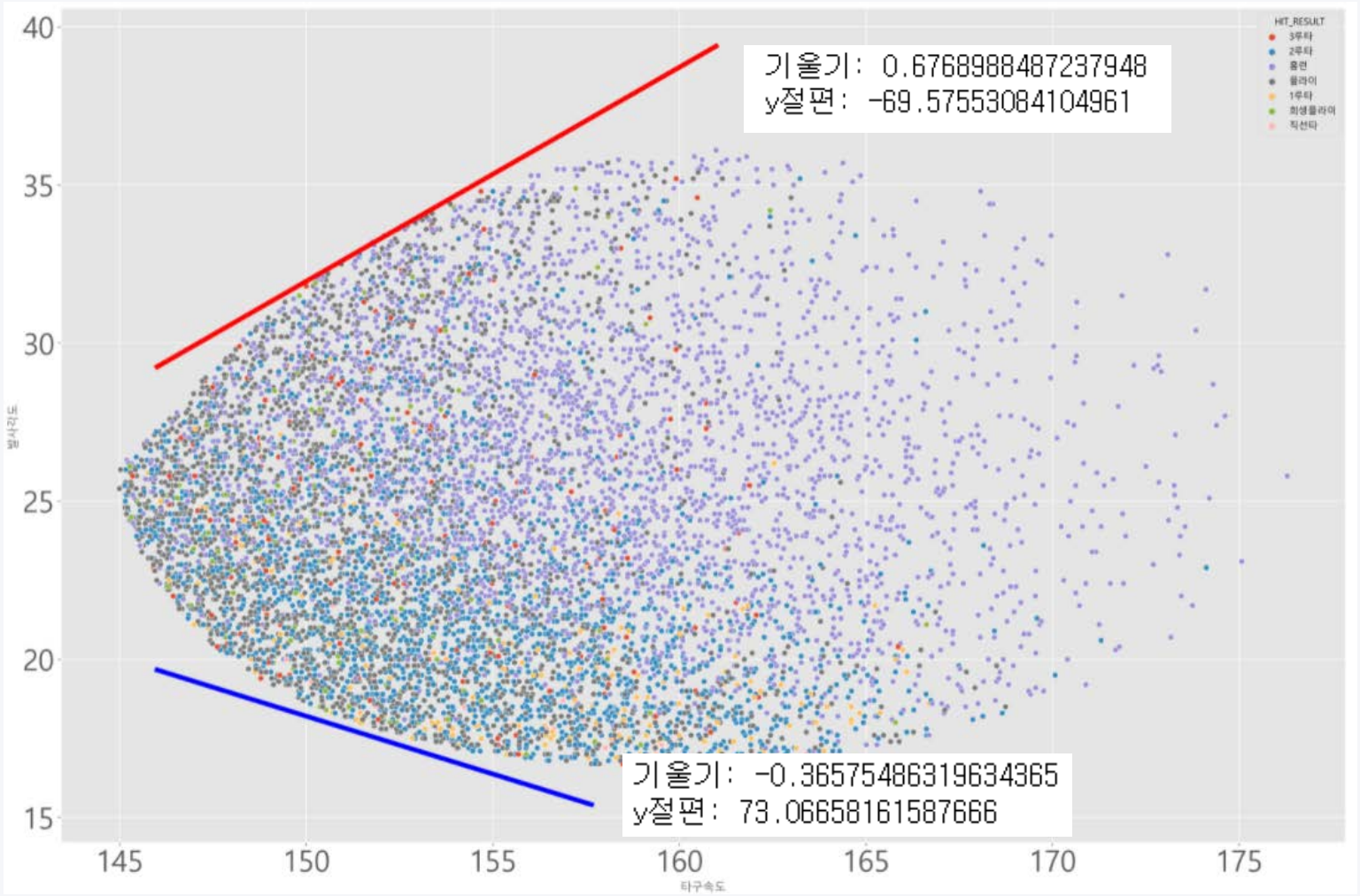


- 조금 더 보편화된 '배럴'의 정의를 구하기 위해서 위에서 그린 이차식을 대표 할 수 있는 일차식이 필요하다.

- 따라서. 타구속도 145km/h ~ 160km/h사이의 중앙값인 152.5km/h에서의 접선의 방정식을 가장 보편화된 식으로 판단하였다.

2-4 모델 기반 확률 추정을 통한 '배럴' 정의

(4) 접선의 방정식을 그래프로 그리기 - 이차식의 접선의 방정식



- 배럴의 범위는 속도가 145km/h 일 때, 20°~29°이고 1km/h 늘어날 때마다 위로 0.68° 밑으로 0.37° 늘어남.

- 속도가 160km/h 이상인 경우 발사각의 범주가 더 이상 커지지 않기 때문에 발사각이 14°~ 39° 사이에 있다면 속도에 상관없이 배럴로 정의.

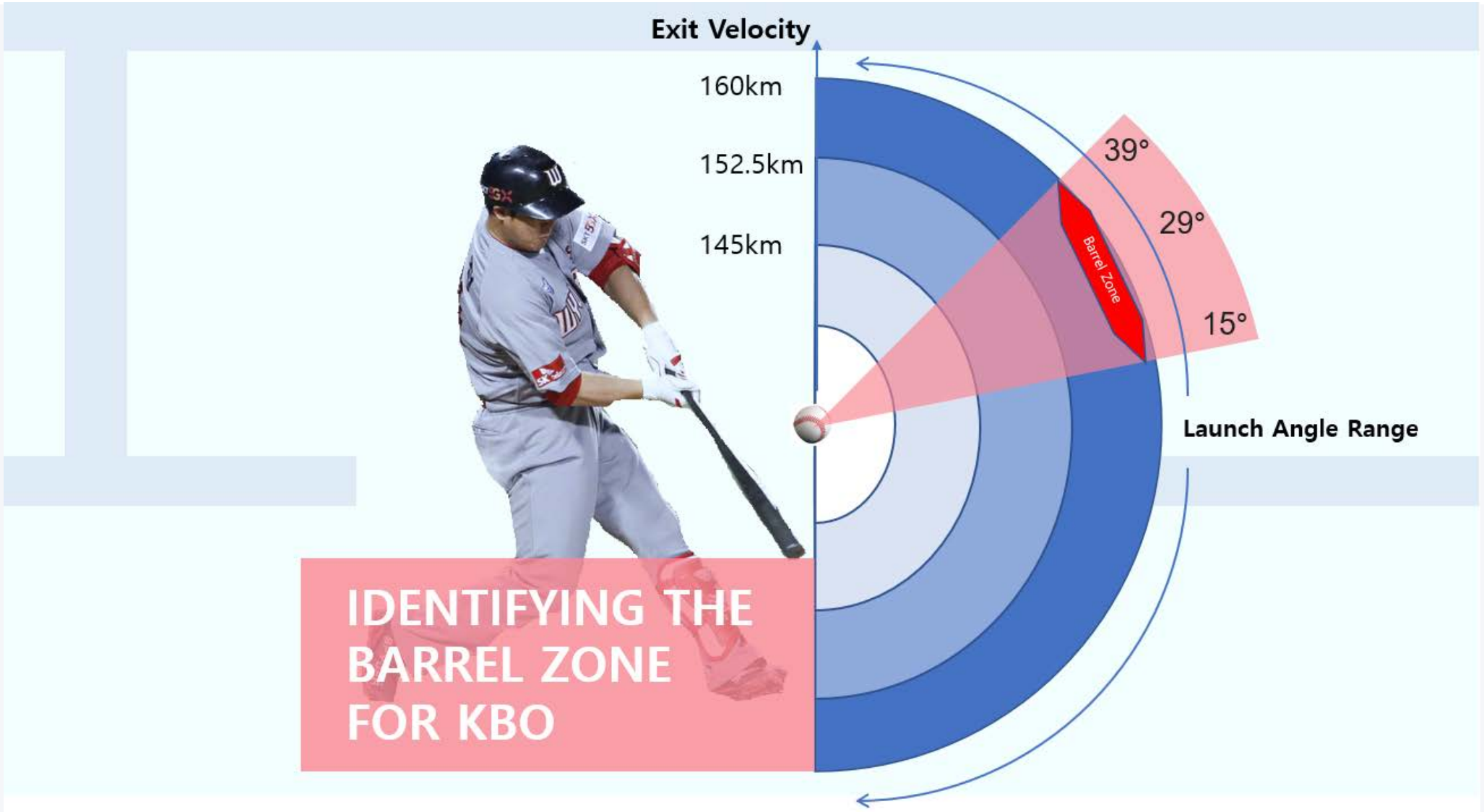
- KBO식 배럴을 타구 속도와 발사각의 조합상 평균적으로 타율 0.600, 장타율 1.400 이상을 생산하는 잘 맞은 타구로 정의.

- KBO식 배럴은 타구 속도 최소 145km/h을 기록해야 하며 발사각이 20°~ 29°가 되어야 하고 145km/h 가 넘는 타구는 발사각의 범주가 조금씩 커짐.

- 속도가 160km/h 이상인 경우 발사각이 14°~ 39°에서 더이상 커지지 않기 때문에 발사각이 14°~ 39° 사이에 있다면 속도에 상관없이 배럴로 정의.

2. 'KBO' 배럴 정의

2-5 결론



3. 타자 성적 예측

3-1 개요

(1) 개요

1. 외부 데이터인 KBO기록실로부터 월별 예측해야 할 선수 10명의 기록을 가져옴

2. 가져온 기록을 월별 OBP(출루율)와 SLG(장타율)로 구분하여 분석 진행함

3. 일별 기록은 데이터가 상당히 불규칙하기 때문에 예측에 어려움이 있어 월별로 집계하여 분석함

4. 예측 모델로는 facebook Prophet을 사용.

(2) facebook prophet 구성요소

$$y(t) = g(t) + s(t) + h(t) + \epsilon_i$$

- Facebook Prophet 모델의 주요 구성 요소는 Trend, Seasonality, Holiday가 있음.

g(t)는 데이터의 전체 추세(Trend)를 잡아줄 수 있고

s(t)는 반복적인 패턴(Seasonality)을 잡아줄 수 있고

h(t)는 불규칙적인 패턴(Holiday)을 잡아줄 수 있음.

(3) facebook prophet 사용이유

- Facebook Prophet은 전체 추세도 잘 잡는 동시에 규칙적, 비규칙적 패턴 모두를 잡을 수 있기 때문에 시계열 분석 및 예측에서 상당히 유연한 모델이라고 생각.
- 또한, 데이터가 선수 마다 기록이 끝나는 월이 다른 경우도 있고, 중간 월에 기록이 없는 선수들도 있음.
- 일반 시계열 모델로는 이 빈 구간을 채워 넣어야 하는데 데이터의 절대적인 양이 부족해 이 방법은 어려움.
- 하지만 Facebook Prophet은 빈 구간을 채워 넣지 않아도 각 월의 패턴을 잘 잡아 줌. 따라서 이와 같은 이유로 Facebook Prophet 모델을 차용함.

3-2 데이터 전처리

- 기본 데이터에는 OBP 및 SLG에 대한 정보가 없으므로 OBP 및 SLG열을 추가.
- fbprophet으로 예측을 진행할 때, 0인 값은 좋은 예측을 하는데 방해를 주기 때문에 이를 이상치로 판단하여 분석 데이터에서 제외함.
- OBP, SLG 열 추가 및 날짜 칼럼 ds 생성.
- 데이터를 prophet에 넣을 수 있도록 ds와 y 열로 조정.

3-3 선수별 OBP 및 SLG 예측

- 21년 7월 전까지의 데이터를 학습데이터로, 21년 8월 실제값을 검증 데이터로 사용하여 예측값을 비교하여 가장 적합한 하이퍼파라미터로 튜닝

- 보라색 점은 실제 8월 Target 값이고, 빨간 그래프가 예측값, 파란 그래프가 실제값임.

- 특히, 하이퍼파라미터 튜닝 시 무조건 8월의 검증 실제값과 가까운 예측값을 모델로 뽑아내면 과적합이 되는 경우가 발생하므로 Manual Search를 통해 적절한 하이퍼파라미터로 튜닝.

- 예측 기간이 9월 15일~10월 8일로 경기의 수가 대략 2:1이기 때문에 9월, 10월 예측결과를 2:1로 가중 평균하여 최종 예측값으로 사용.

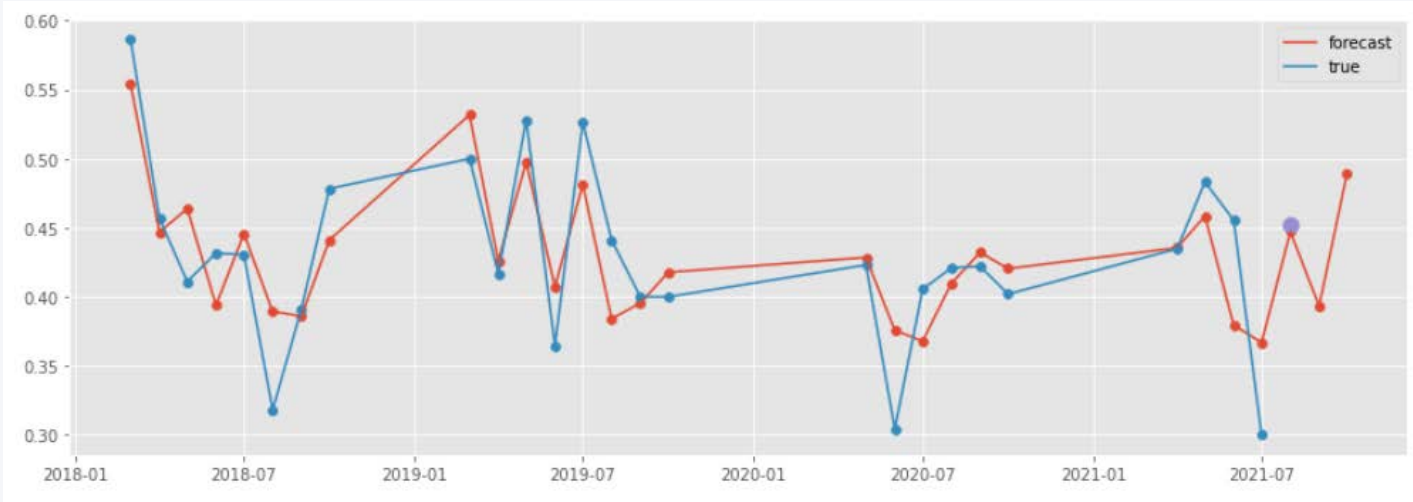
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(1) 양의지 (76232)

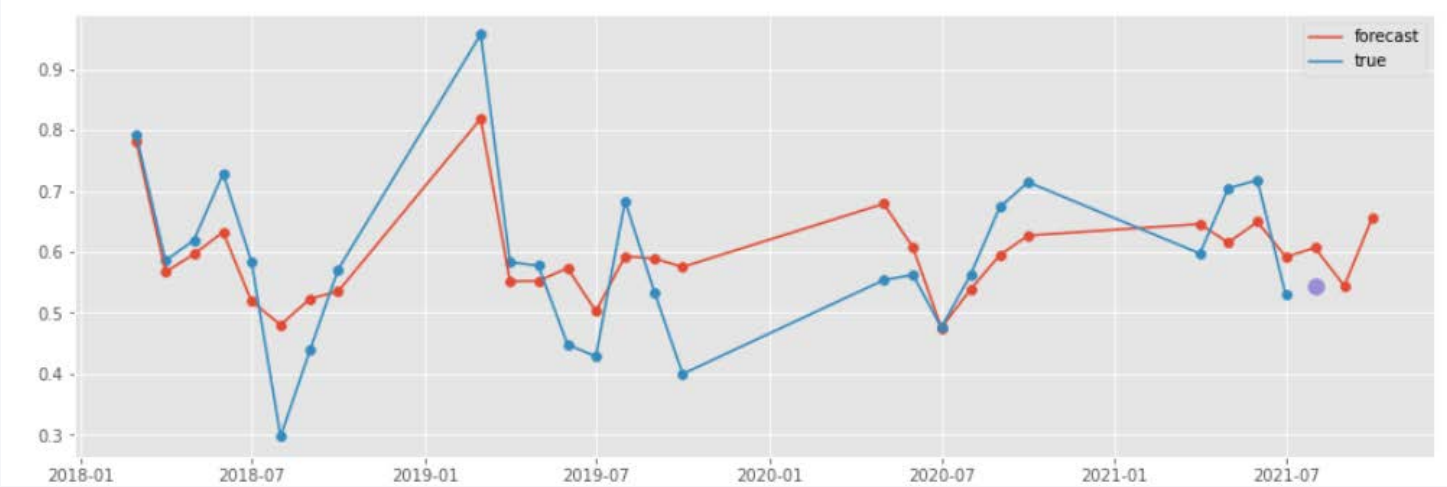
OBP

9월 예측 : 0.3933253895000263
10월 예측 : 0.4891708179389940 3
총 예측 : 0.4252738656463488 4



SLG

9월 예측 : 0.5444690440894934
10월 예측 : 0.657253705720927
총 예측 : 0.5820639312999712



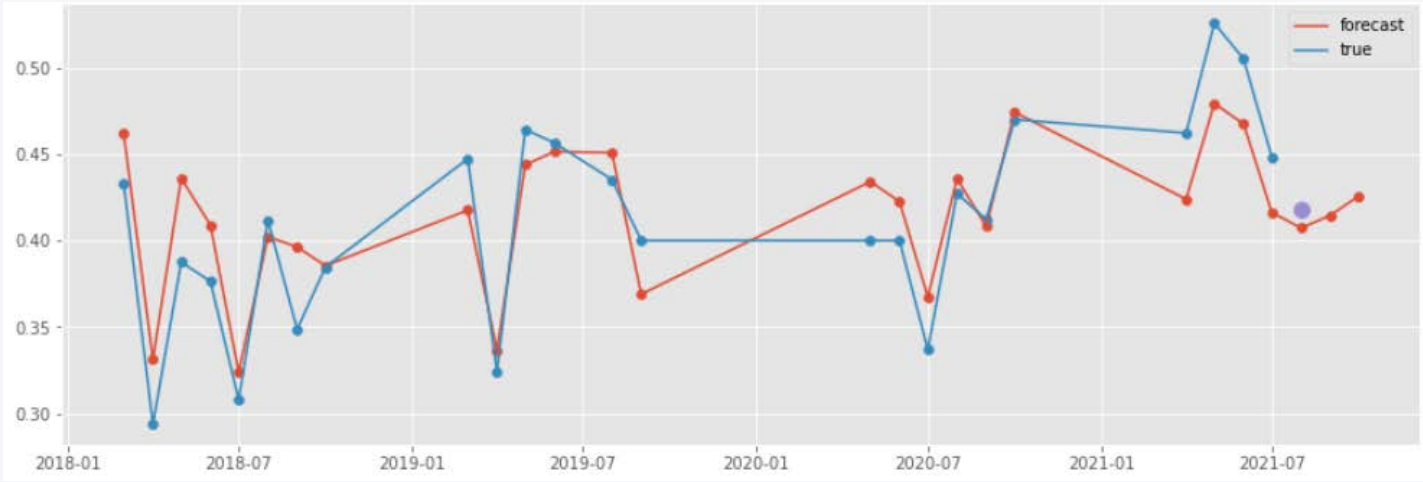
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(2) 강백호 (68050)

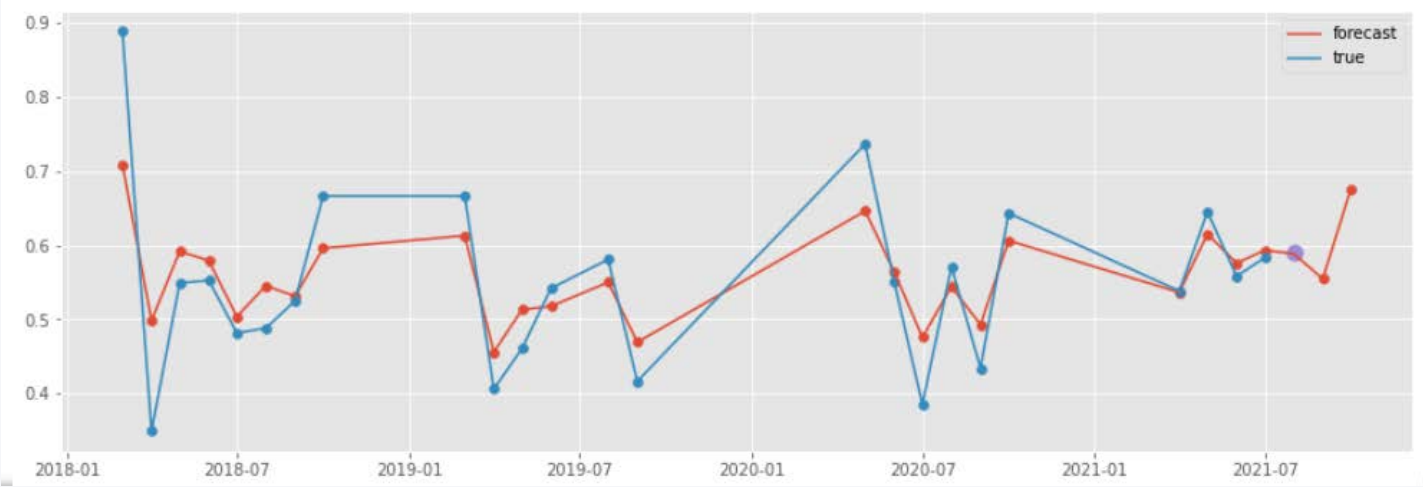
OBP

9월 예측 : 0.41455154991876536
10월 예측 : 0.4257654199069039
총 예측 : 0.4182895065814782



SLG

9월 예측 : 0.5549102334117465
10월 예측 : 0.6767278420416215
총 예측 : 0.5955161029550382



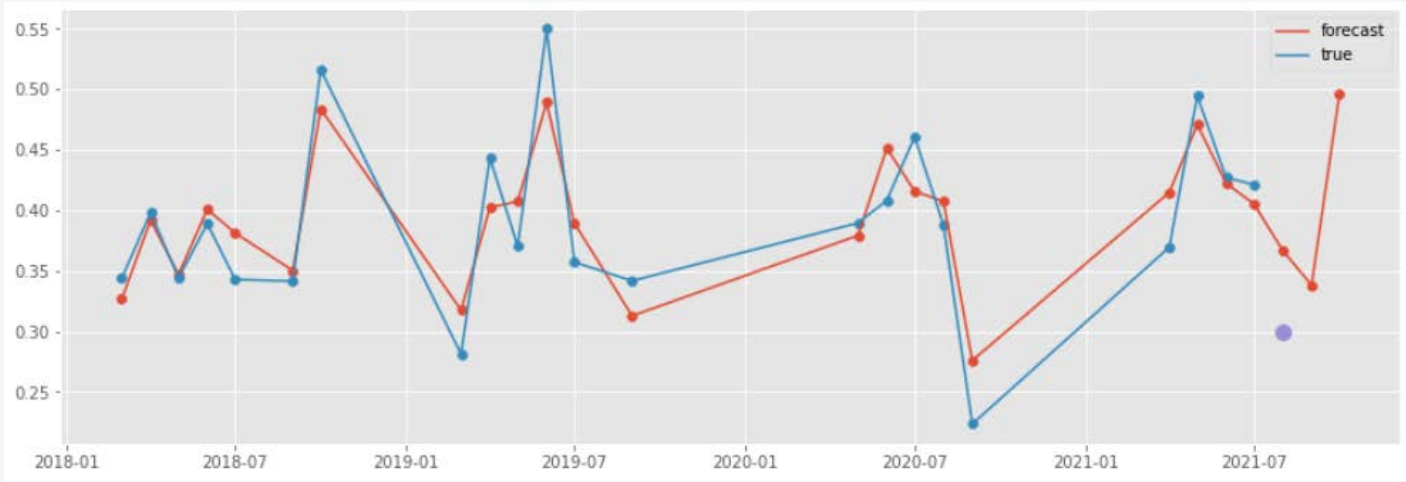
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(3) 최정 (75847)

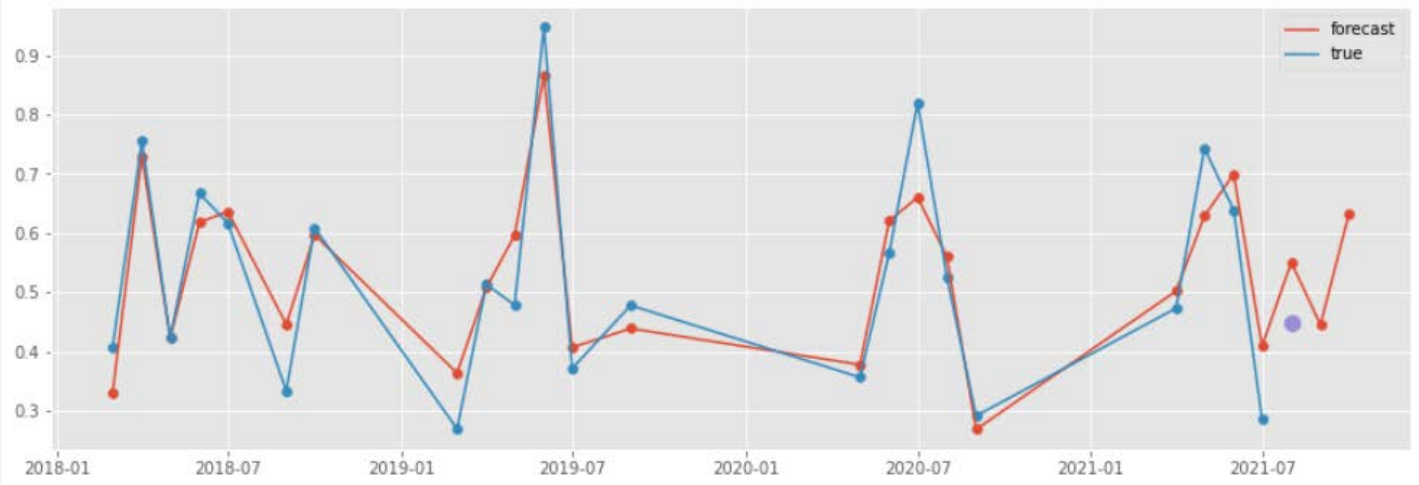
OBP

9월 예측 : 0.3381714384517708
10월 예측 : 0.4960020062152482
총 예측 : 0.39078162770626323



SLG

9월 예측 : 0.44496225780876203
10월 예측 : 0.6313837787318031
총 예측 : 0.507102764783109



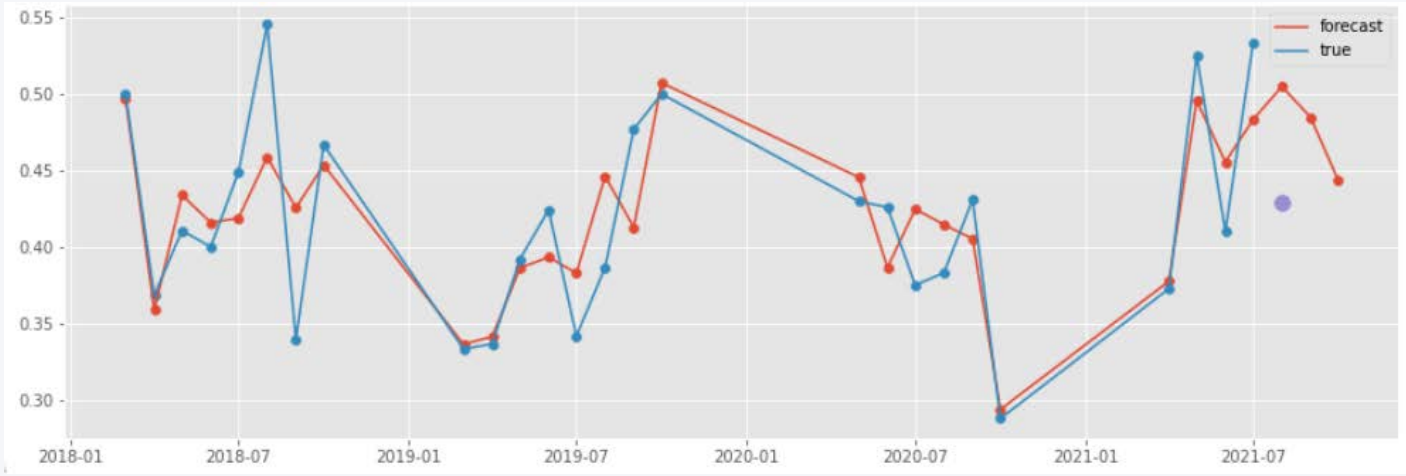
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(4) 이정후 (67341)

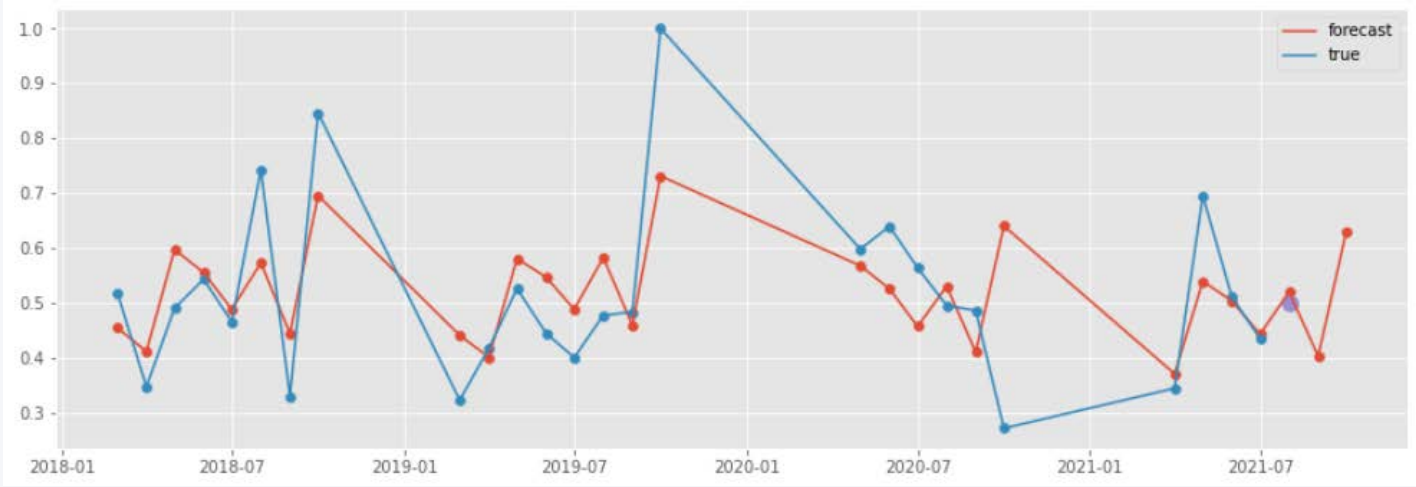
OBP

9월 예측 : 0.4847532489329454
10월 예측 : 0.443442880575841 25
총 예측 : 0.470983126147244



SLG

9월 예측 : 0.4026727435890564
10월 예측 : 0.6299830622221234
총 예측 : 0.4784428498000787



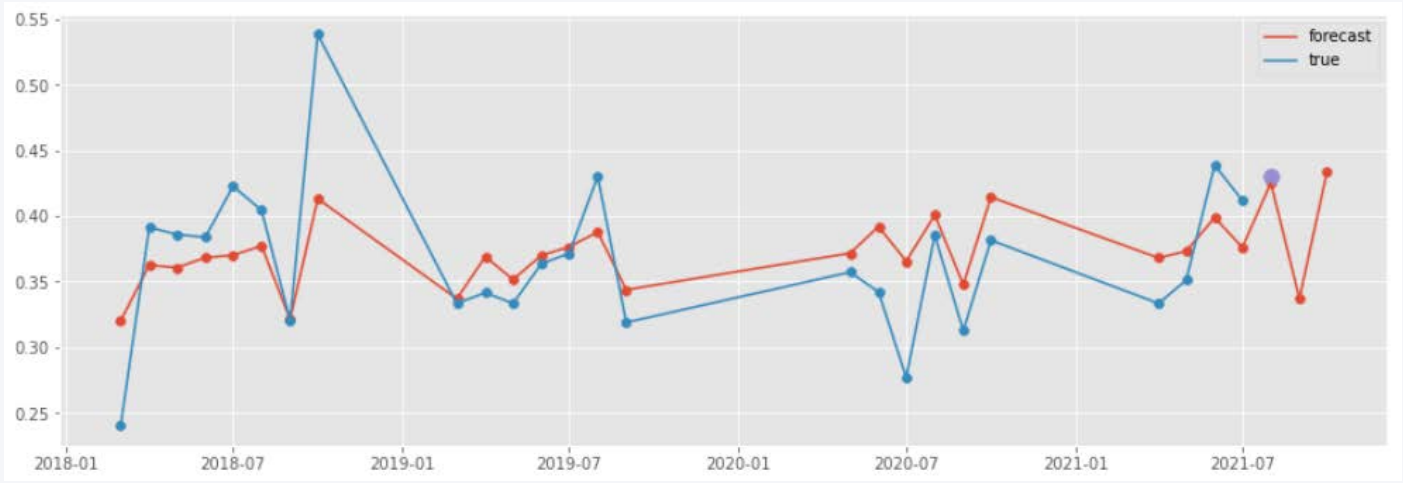
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(5) 채은성 (79192)

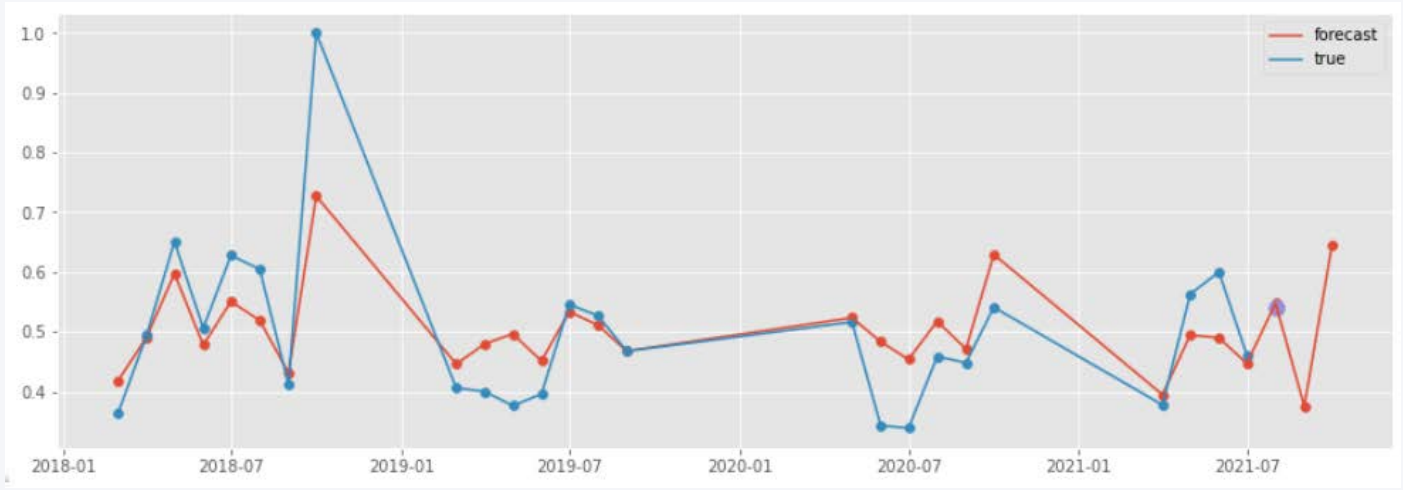
OBP

9월 예측 : 0.3370630674112327
10월 예측 : 0.43389087810497295
총 예측 : 0.3693390043091461



SLG

9월 예측 : 0.3757280187797599
10월 예측 : 0.64564655865864
총 예측 : 0.4657008654060532 7



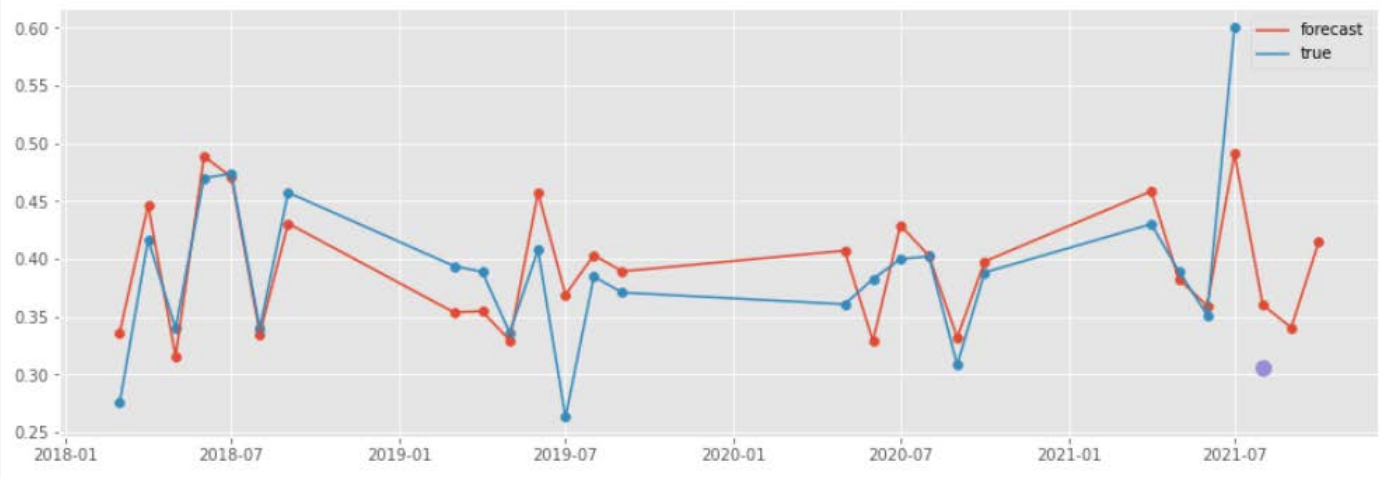
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(6) 김재환 (78224)

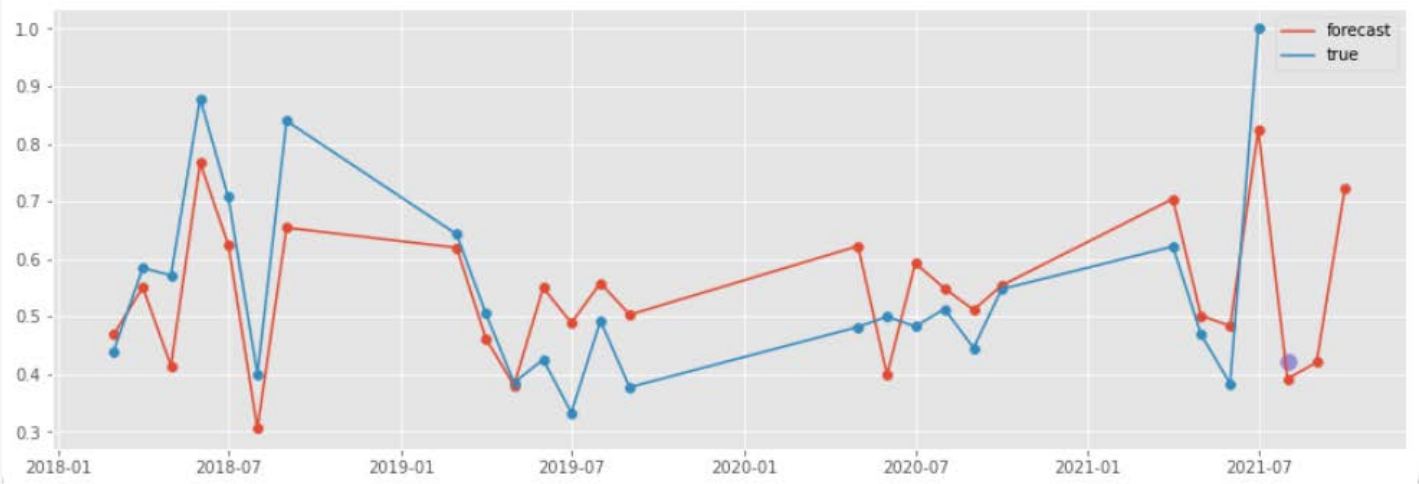
OBP

9월 예측 : 0.3406269553448055
10월 예측 : 0.4146973694569787
총 예측 : 0.3653170933821966



SLG

9월 예측 : 0.42053627576109 96
10월 예측 : 0.7214841641480546
총 예측 : 0.5208522385567512



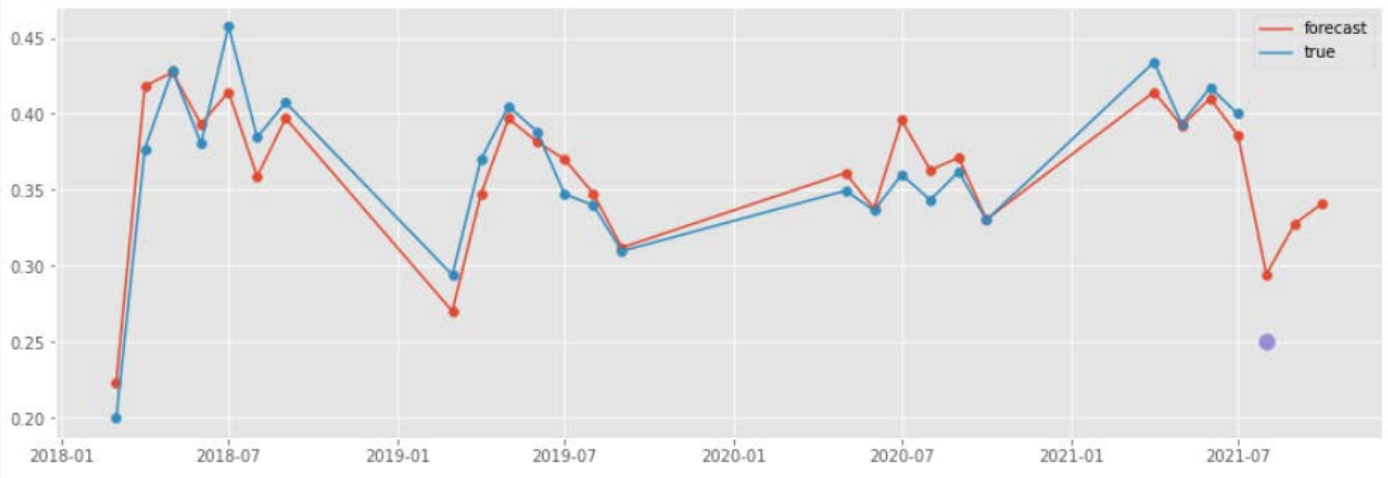
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(7) 전준우 (78513)

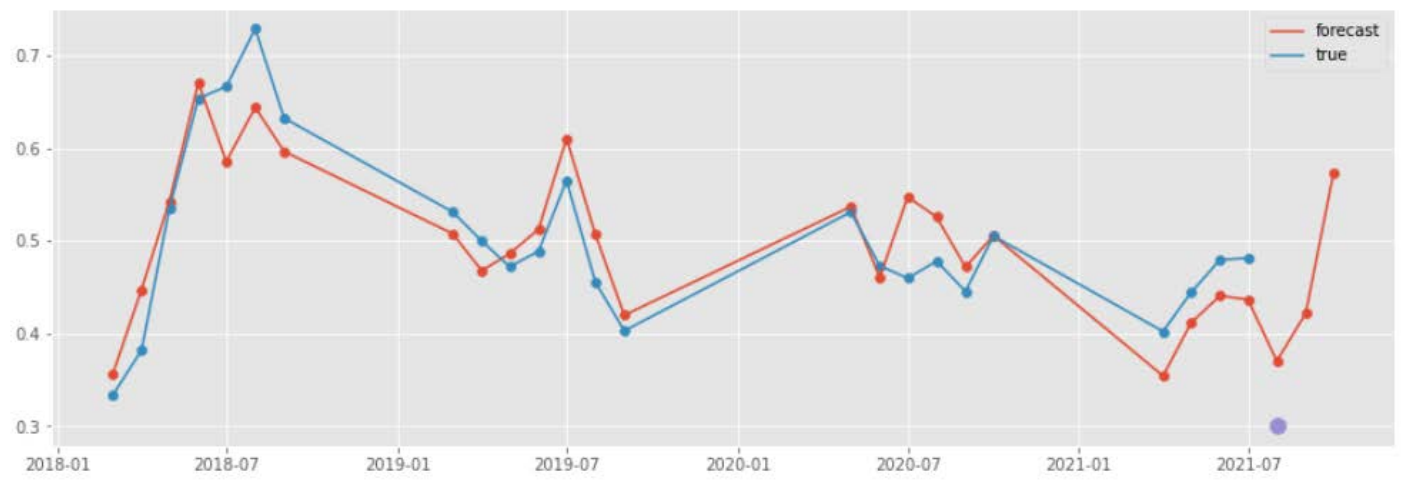
OBP

9월 예측 : 0.3279424383717292
10월 예측 : 0.34127037374528457
총 예측 : 0.3323850834962476



SLG

9월 예측 : 0.42167491675670 61
10월 예측 : 0.5742216846185075
총 예측 : 0.4725238393773066



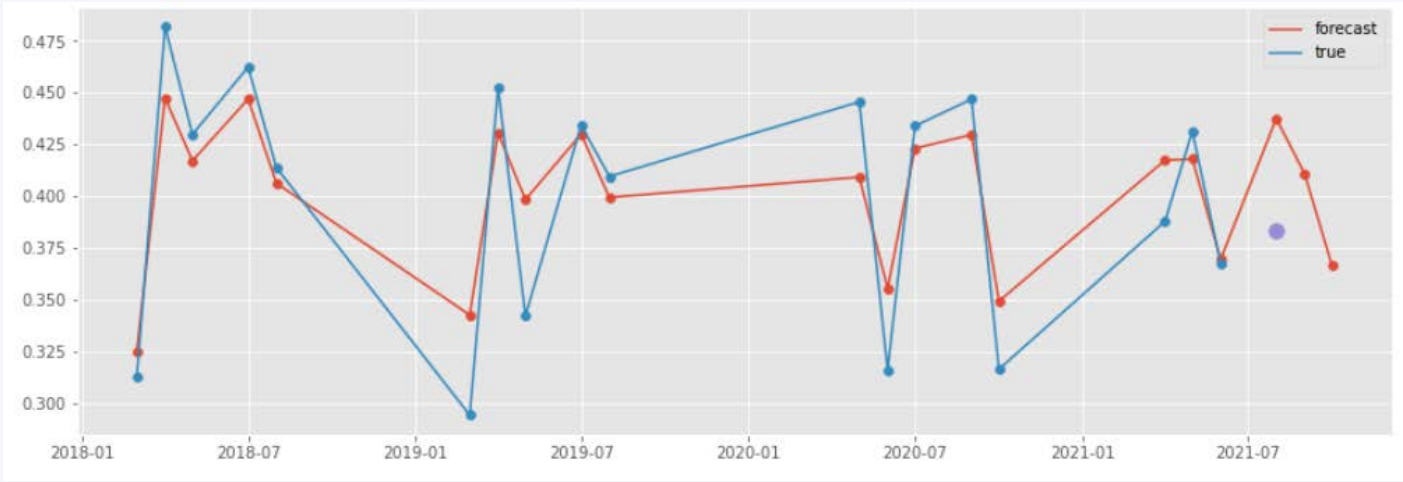
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(8) 김현수 (76290)

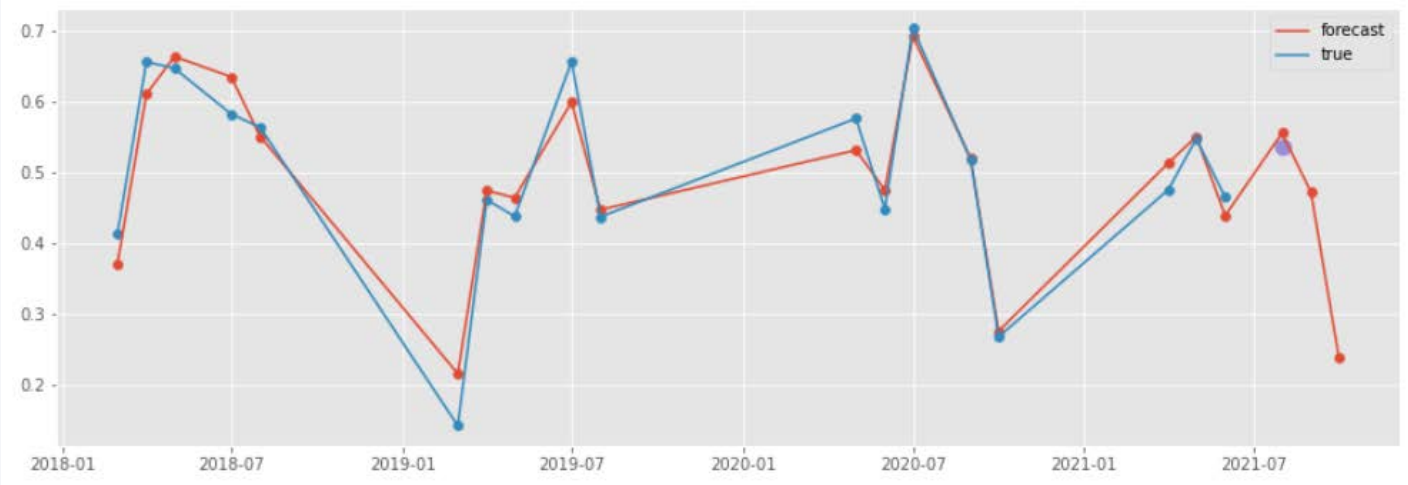
OBP

9월 예측 : 0.41087431895353765
10월 예측 : 0.36660456855653073
총 예측 : 0.39611773548786866



SLG

9월 예측 : 0.47226163289384354
10월 예측 : 0.23903963553206764
총 예측 : 0.39452096710658485



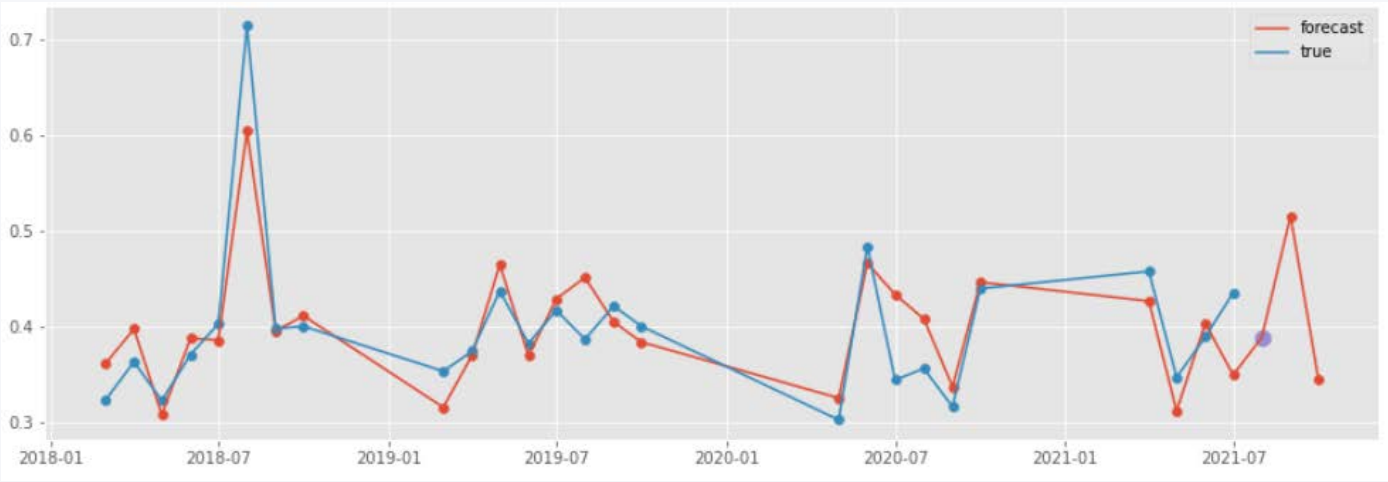
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(9) 박건우 (79215)

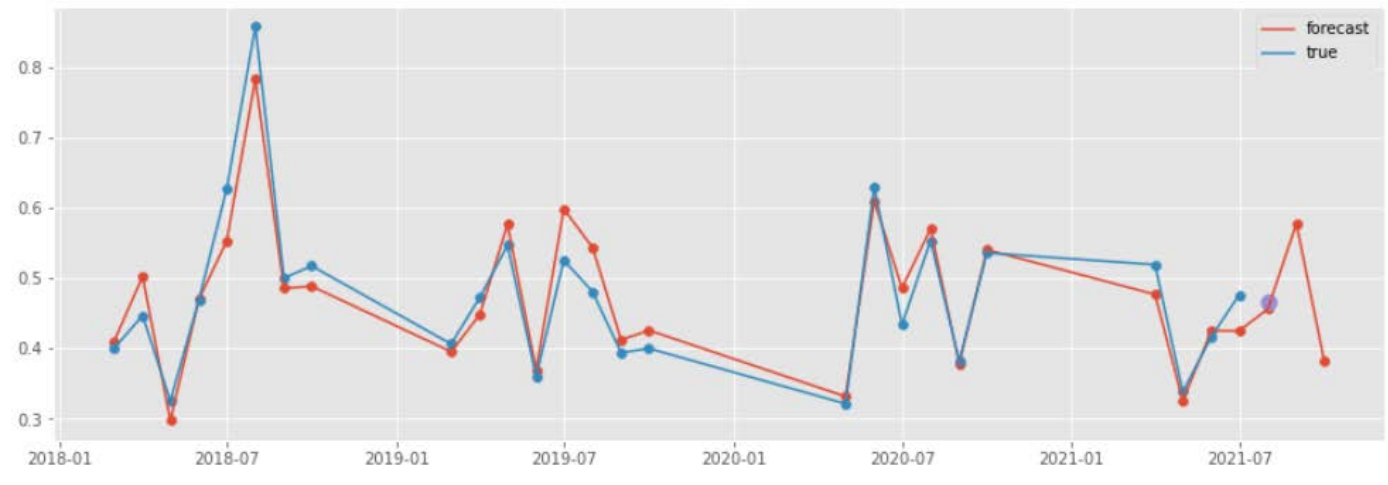
OBP

9월 예측 : 0.5154239476634909
10월 예측 : 0.3452997351295874
총 예측 : 0.458715876818856 34



SLG

9월 예측 : 0.5772351949343718
10월 예측 : 0.38309947060752486
총 예측 : 0.5125232868254228



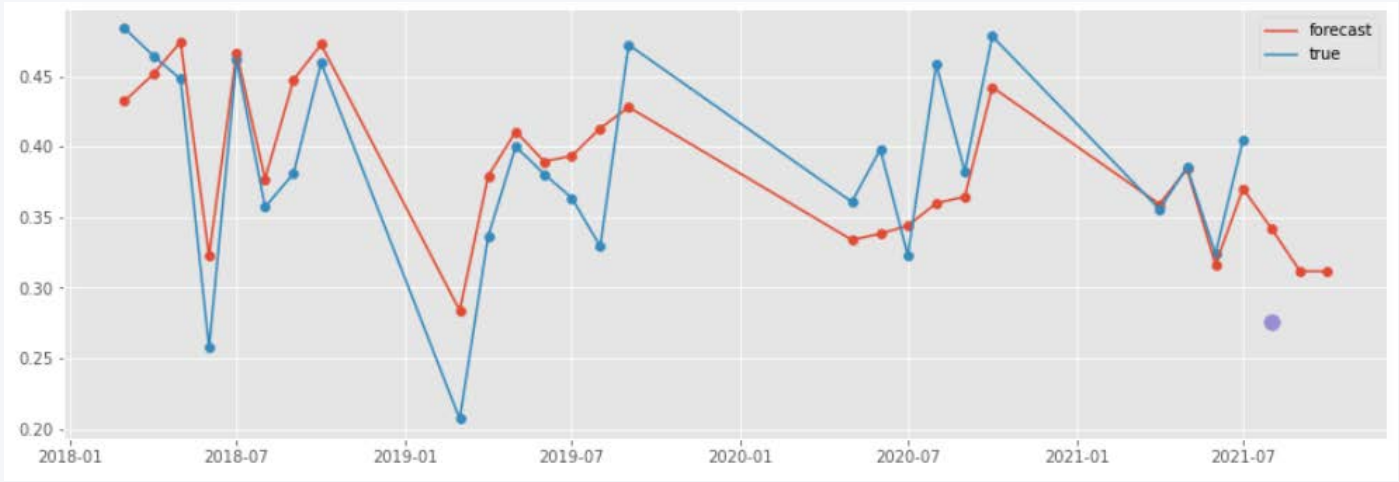
3. 타자 성적 예측

3-3 선수별 OBP 및 SLG 예측

(10) 로맥 (67872)

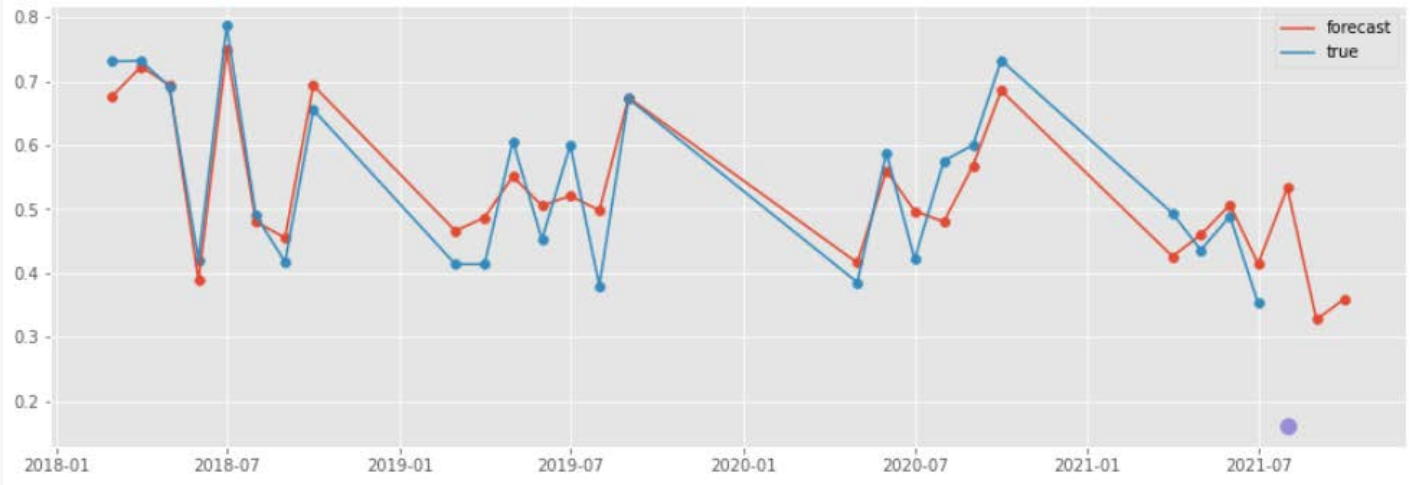
OBP

9월 예측 : 0.31185925378870166
10월 예측 : 0.3115536332571723
총 예측 : 0.3117573802781919



SLG

9월 예측 : 0.3273265245925721
10월 예측 : 0.35993120904448334
총 예측 : 0.3381947527432092



3. 타자 성적 예측

3-4 선수별 최종 OPS 예측

NO.	PCODE	OPS	장타율	출루율
NaN	(예시)	1.090000	0.680000	0.417000
1	76232	1.007338	0.582064	0.425274
2	68050	1.013806	0.595516	0.418290
3	75847	0.897884	0.507103	0.390782
4	67341	0.949426	0.478443	0.470983
5	79192	0.835040	0.465701	0.369339
6	78224	0.886169	0.520852	0.365317
7	78513	0.804909	0.472524	0.332385
8	76290	0.790639	0.394521	0.396118
9	79215	0.971239	0.512523	0.458716
10	67872	0.649952	0.338195	0.311757
*설명회 자료 참고	NaN	NaN	NaN	NaN

감사합니다.