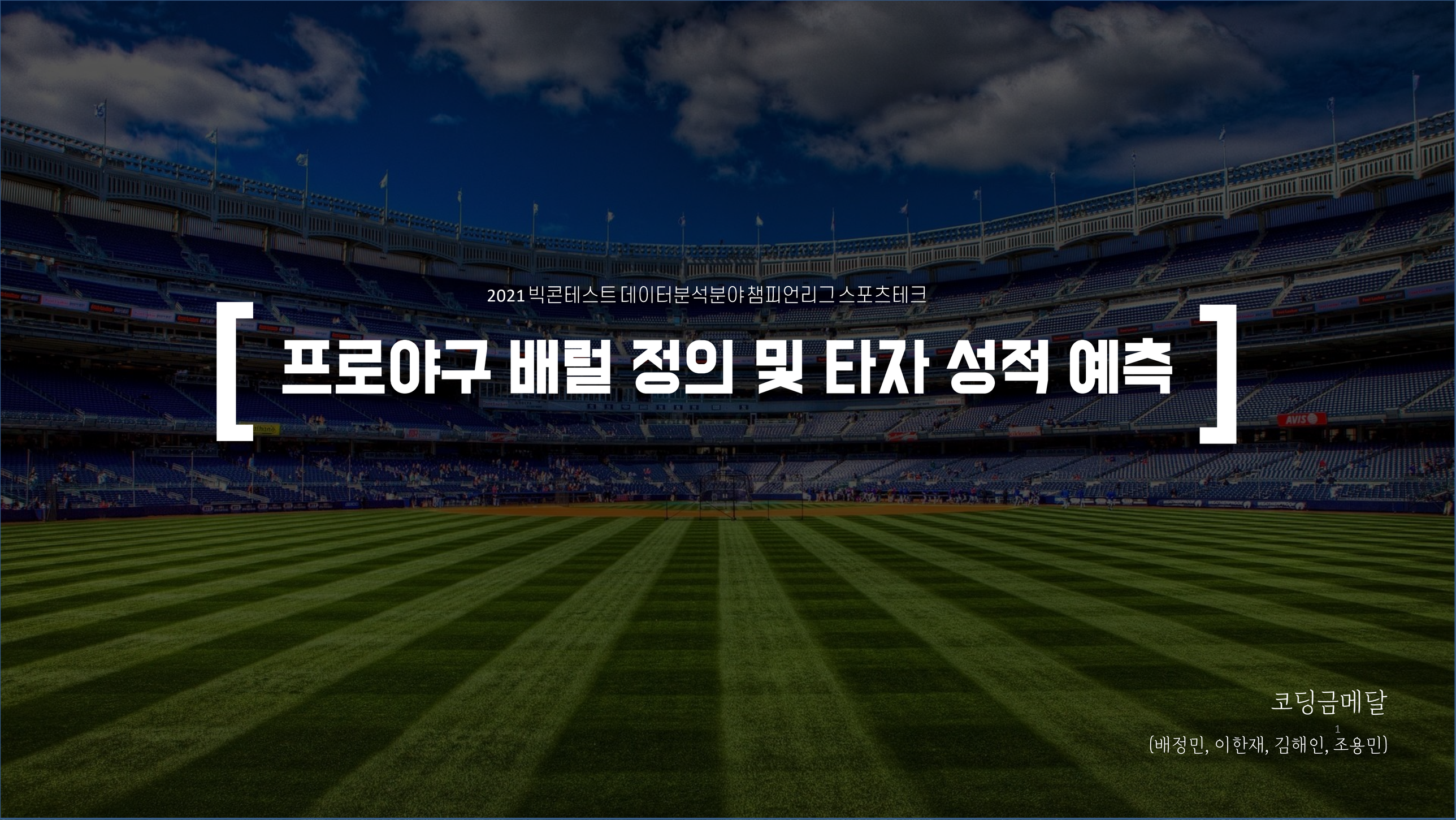


2021 빅콘테스트 데이터분석분야 챔피언리그 스포츠테크

[프로야구 배럴 정의 및 타자 성적 예측]

코딩금메달

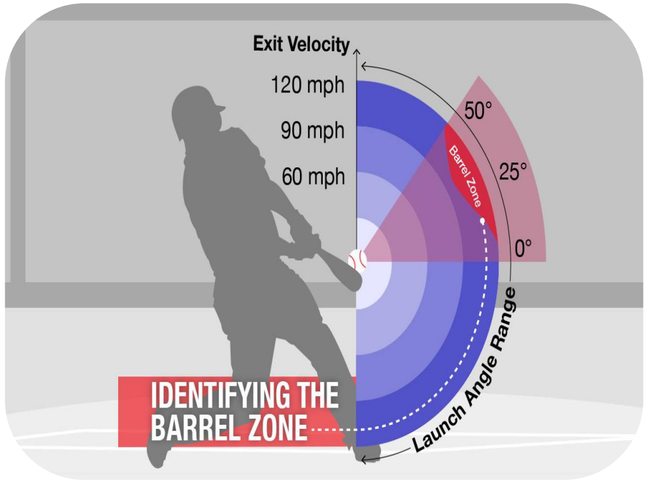
(배정민, 이한재, 김해인, 조용민)

CONTENTS

1 분석개요

분석 개요

배경



MLB의 “ *barrels* ”

배럴(barrels)은 2015년 MLB 트래킹 데이터의 사용과 함께 등장한 개념이다.
MLB 타자로부터 형성된 기준을 사용해 좋은 타구로 배럴이 정의되었으며,
정의된 배럴의 타구속도와 발사각도 경계 기준점으로 타구들이 구분된다.

KBO에 MLB에서 정의된 배럴을 활용하는 것은 적절하지 않다.
따라서 KBO의 성격이 반영된 새로운 기준으로 배럴을 재정의 해야 한다.

분석 개요

분석목표

기존의 타격 변수

타자 기본 변수

세이버메트릭스 변수

새로 정의한 배럴

KBO의 배럴

타석에서 생산되는 단순 결과물(홈런, 안타 등)만으로 설명하기 어려운 타자 고유의 능력이 반영되도록 변수를 정의한다.

배럴을 활용한다면, 기존의 타격 변수들의 값과 상호보완적으로 작용하여 타자의 능력을 보다 객관적으로 평가할 수 있다.

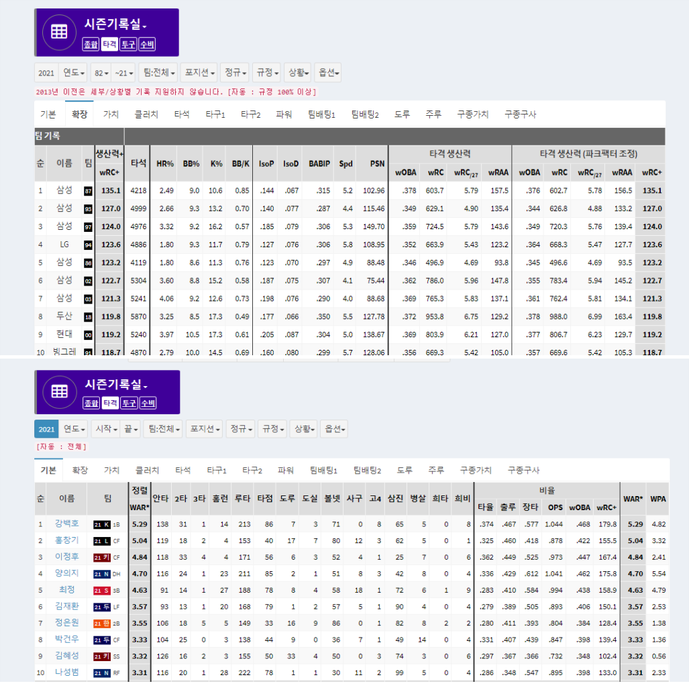
이를 통해 타자들의 보편적 평가 변수들인 OPS , 출루율 , 장타율을 예측하는 모델을 구축한다.

2 배럴 정의 

배럴 정의

데이터 크롤링

1 | STATIZ

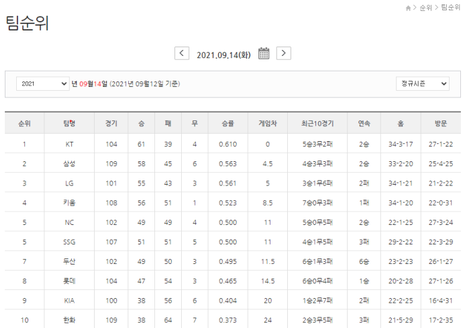


1982 ~ 2021

연도별 선수 기록 크롤링

연도별 팀 기록 크롤링

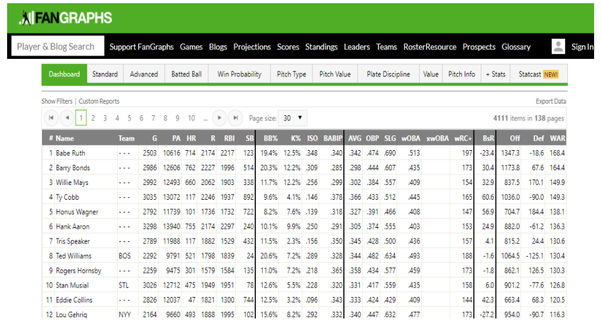
2 | KBO 기록실



1982 ~ 2021

연도별 팀 성적 크롤링

3 | Fangraphs



1982 ~ 2021

연도별 선수 기록 크롤링

배럴 정의

진행 과정

1 변수 선택

- 배럴의 기준을 팀 승리에 영향을 미치는 타구로 판단한다.
- 팀 승리에 영향을 끼치는 변수를 선택하고 KBO와 MLB의 차이를 확인한다.

2 두 리그의 차이 확인

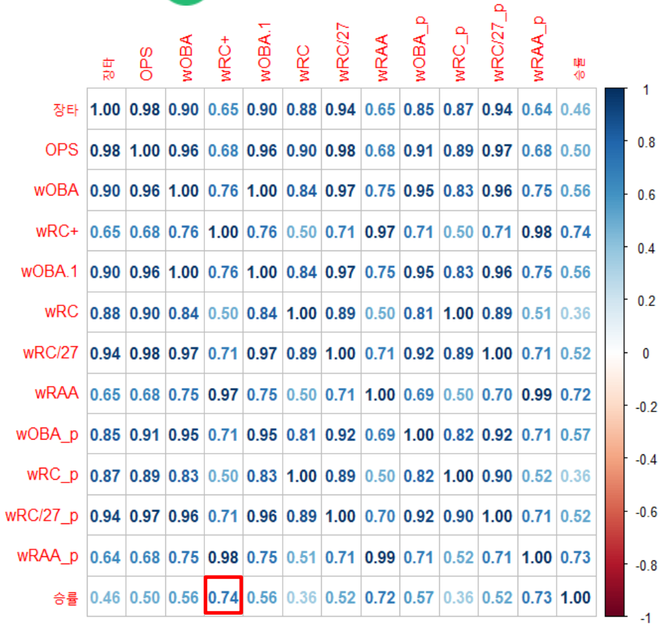
- 선정된 변수에 대해 두 리그에서 타율/장타율의 차이가 있는지 확인한다.

3 배럴 구간 선정

- 두 리그의 차이로 인해 재정의된 타구속도와 발사각도로 KBO의 배럴 구간을 정의한다.
- 관측되지 않은 범위에 대해서 물리적인 가능성을 고려해 배럴의 범위를 확장한다.

배럴 정의
변수 선택

승률과의 상관계수



wRC+와의 상관성이 가장 높아
wRC+를 기준으로 배럴을 재정의한다.

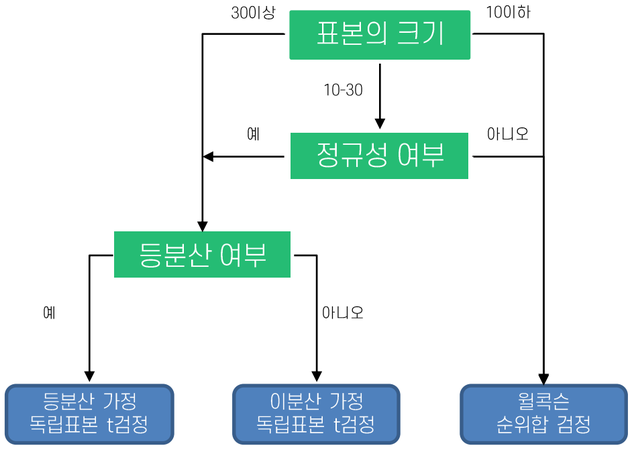
wRC+ 그룹화

wRC+	설명	그룹명
160~	Excellent	1
140~160	Great	2
115~140	Above Average	3
100~115	Average	4
80~110	Below Average	5
75~80	Poor	6
~75	Awful	7

동일 wRC+로 계산하기에는 관측치의
수가 적어, 그룹화를 진행한다.

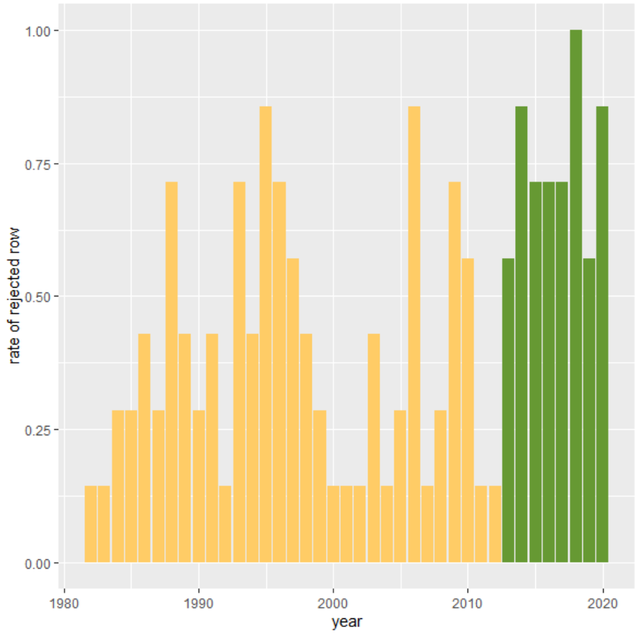
배럴 정의

두 리그의 차이 확인



두 리그의 타율이 차이가 있는지 확인하기 위한 검정 기준

검정결과



2013년 이후 부터 KBO와 MLB의
타율 차이가 있는 데이터가 절반 이상 존재

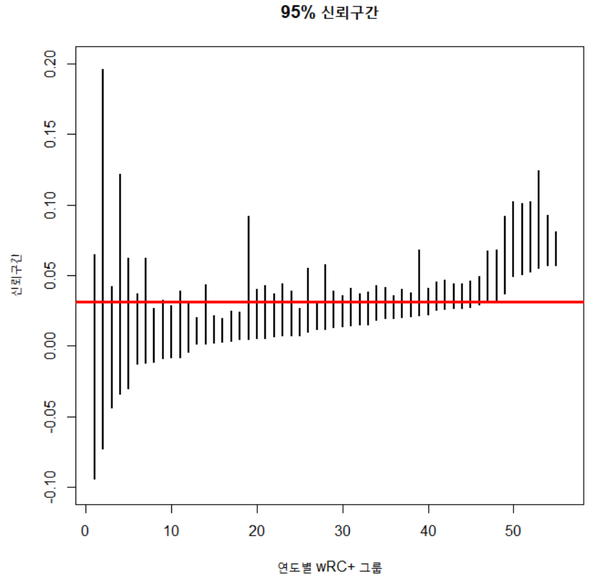
2013년 이후부터 연도 간 큰 차이를 보이지 않음

2013년 이후의 자료를 이용해
새로운 배럴을 도출

배럴 정의

두 리그의 차이 확인

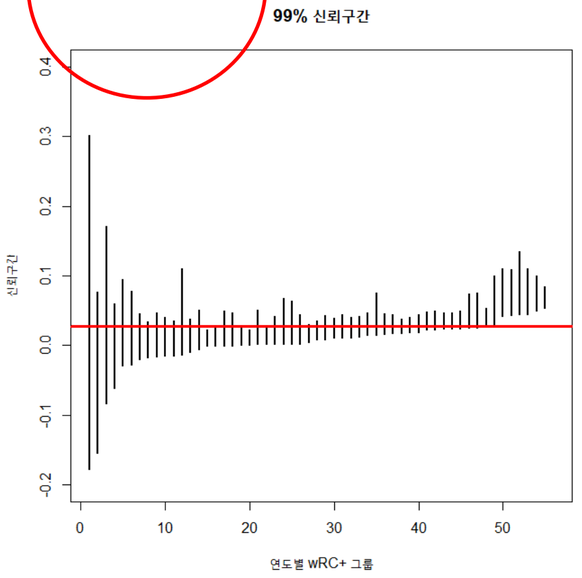
1| 95% 신뢰구간



56개의 신뢰구간 중 40개의 신뢰구간에
구간(0.03083, 0.03108)이 겹친다.

99% 채택

2| 99% 신뢰구간

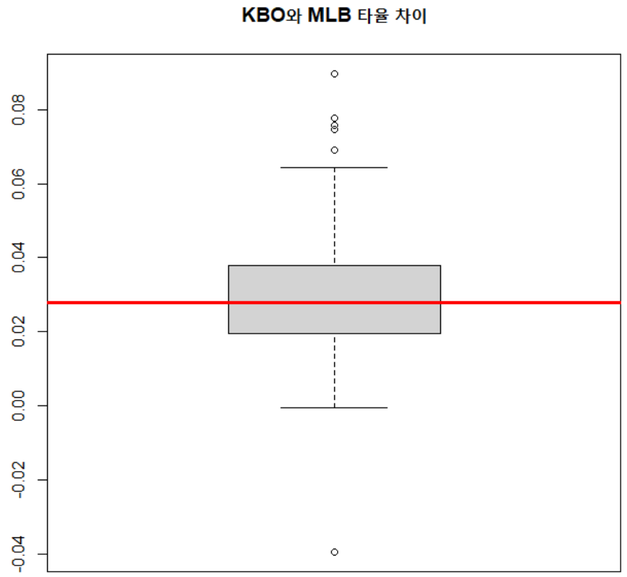


56개의 신뢰구간 중 45개의 신뢰구간에 구간(0.02444 , 0.02476)
또는 구간(0.02566 , 0.02687)이 겹친다.

배럴 정의

두 리그의 차이 확인

점 추정치 활용



Min	1Q	Median	Mean	3Q	Max
-0.01450	0.01656	0.02763	0.03198	0.03789	0.08973

- 점 추정치의 중앙값인 0.02763에 가까운 구간(0.02566, 0.02687) 내에 실제 차이가 존재한다고 생각한다.
- 구간의 중앙값인 0.026265를 KBO와 MLB의 타율 차이로 활용한다.

배럴 정의

배럴 구간

1| 5단위로 정의

	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3	1	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	14	13	14	10	4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
12	36	41	32	42	24	10	15	4	1	0	1	0	0	0	0	0	0	0	0	0
13	86	102	128	104	121	89	66	42	19	5	5	0	0	0	0	0	0	0	0	0
14	184	201	220	243	301	323	317	246	146	70	17	2	1	1	0	0	0	0	0	0
15	301	308	360	444	541	619	714	631	479	289	122	26	5	1	0	0	0	0	0	1
16	422	436	471	589	734	929	1065	1163	984	553	262	83	25	2	1	0	0	0	0	0
17	387	490	566	754	902	1173	1473	1568	1290	732	280	66	7	0	0	0	1	0	0	0
18	362	439	507	669	817	1051	1205	1325	1011	572	204	38	3	0	0	0	0	0	0	0
19	387	435	558	730	871	1070	1280	1324	1021	560	225	58	4	0	0	0	0	0	0	0
20	377	454	604	714	890	1093	1248	1330	1000	609	219	44	2	1	0	0	0	0	0	0
21	354	467	576	789	956	1102	1218	1173	895	527	195	31	5	0	0	0	0	0	0	0
22	377	470	637	730	971	1107	1216	1054	801	371	143	32	5	0	0	0	0	0	0	0
23	423	466	561	761	871	1042	1011	880	592	283	100	28	3	0	0	0	0	0	0	0
24	362	429	537	705	792	825	773	658	363	137	53	7	1	0	0	0	0	0	0	0
25	298	426	473	593	633	655	557	422	222	94	24	5	2	0	0	0	0	0	0	0
26	286	351	433	548	574	480	387	260	120	71	12	0	1	0	0	0	0	0	0	0
27	253	340	402	420	433	368	284	164	69	32	3	0	0	0	0	0	0	0	0	0
28	273	327	357	373	368	296	180	99	39	6	5	0	0	0	0	0	0	0	0	0
29	208	264	316	309	283	212	133	68	29	4	0	0	0	0	0	0	0	0	0	0
30	196	243	263	289	226	158	74	39	9	1	0	0	0	0	0	0	0	0	0	0
31	205	212	227	197	145	83	37	16	8	1	0	0	0	0	0	0	0	0	0	0
32	175	183	188	142	97	46	29	5	1	0	0	0	0	0	0	0	0	0	0	0
33	145	147	133	92	60	31	13	3	0	0	0	0	0	0	0	0	0	0	0	0
34	129	129	81	62	33	10	1	2	0	0	0	0	0	0	0	0	0	0	0	0

HTS 자료에서 타구속도와 타구 발사각도를 각각
5 단위로 나누어 해당 구간의 타율과 장타율을 확인

발사 각도(°)	
그룹 번호	구간
1	-80도 미만
2	-80도 이상 -75도 미만
3	-75도 이상 -70도 미만
…	…
20	10도 이상 15도 미만
21	15도 이상 20도 미만
22	20도 이상 25도 미만
23	25도 이상 30도 미만
24	30도 이상 35도 미만
25	35도 이상 40도 미만
26	40도 이상 45도 미만
27	45도 이상 50도 미만
…	…
32	70도 이상 75도 미만
33	75도 이상 80도 미만
34	80도 이상

타구 속도(km/h)	
그룹 번호	구간
1	10미만
2	100이상 150미만
3	150이상 200미만
…	…
30	1500이상 1550미만
31	1550이상 1600미만
32	1600이상 1650미만
33	1650이상 1700미만
34	1700이상 1750미만
35	1750이상 1800미만
36	1800이상 1850미만
37	1850이상 1900미만
…	…
40	2000이상 2050미만
41	2050이상 2100미만
42	2100이상

배럴 정의

배럴 구간 선정

2| 세분화 1

	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3	1	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	14	13	14	10	4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	36	41	32	42	24	10	15	2	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
13	86	102	128	104	121	89	66	26	16	14	5	4	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0
14	184	201	220	243	301	323	317	135	111	96	50	36	34	10	7	2	0	1	0	0	1	0	0	0	0	0	0
15	301	308	360	444	541	619	714	327	304	258	221	165	124	86	36	17	9	3	2	1	0	0	0	0	0	0	1
16	422	436	471	589	734	929	1065	615	548	509	475	328	225	177	85	57	26	19	6	2	0	1	0	0	0	0	0
17	387	490	566	754	902	1173	1473	794	774	723	567	438	294	183	97	52	14	7	0	0	0	0	0	1	0	0	0
18	362	439	507	669	817	1051	1205	680	645	573	438	364	208	144	60	28	10	2	1	0	0	0	0	0	0	0	0
19	387	435	558	730	871	1070	1280	634	690	566	455	327	233	144	81	48	10	3	1	0	0	0	0	0	0	0	0
20	204	249	305	334	437	524	606	301	333	288	231	191	122	74	45	14	10	1	0	0	0	0	0	0	0	0	0
21	173	205	299	380	453	569	642	349	347	260	221	169	127	75	25	15	5	0	1	1	0	0	0	0	0	0	0
22	156	235	281	388	442	556	628	339	297	255	227	179	106	72	39	13	6	1	2	0	0	0	0	0	0	0	0
23	198	232	295	401	514	546	590	280	257	222	191	151	91	51	33	8	4	2	0	0	0	0	0	0	0	0	0
24	184	246	299	354	472	547	584	283	268	233	178	122	83	56	23	10	5	1	1	0	0	0	0	0	0	0	0
25	193	224	338	376	499	560	632	239	264	205	185	105	61	43	21	10	7	2	1	0	0	0	0	0	0	0	0
26	423	466	561	761	871	1042	1011	456	424	348	244	167	116	68	32	15	13	2	1	0	0	0	0	0	0	0	0
27	180	216	258	375	432	403	403	181	161	115	72	49	32	18	16	3	2	0	0	0	0	0	0	0	0	0	0
28	182	213	279	330	360	422	370	181	135	104	72	35	21	9	10	1	1	1	0	0	0	0	0	0	0	0	0
29	137	214	267	306	340	363	292	136	97	70	59	40	20	10	2	2	1	0	1	0	0	0	0	0	0	0	0
30	161	212	206	287	293	292	265	118	71	53	40	25	9	7	5	2	0	1	0	0	0	0	0	0	0	0	0
31	144	164	231	255	308	258	219	86	65	41	24	24	16	4	6	0	0	0	1	0	0	0	0	0	0	0	0
32	142	187	202	293	266	222	168	57	52	31	24	23	8	1	1	0	0	0	0	0	0	0	0	0	0	0	0
33	137	186	227	211	241	200	155	59	32	23	17	10	7	3	0	0	0	0	0	0	0	0	0	0	0	0	0
34	116	154	175	209	192	168	129	42	31	20	9	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- 배럴로 판단된 타구속도 30~43그룹을 2.5단위로 세분화
- 일정 타구속도(43 그룹)를 넘어가는 타구는 HTS 데이터에 거의 존재하지 않아 세분화 불필요

발사 각도(°)	
그룹 번호	구간
1	-80도 미만
2	-80도 이상 -75도 미만
3	-75도 이상 -70도 미만
…	…
20	10도 이상 15도 미만
21	15도 이상 20도 미만
22	20도 이상 25도 미만
23	25도 이상 30도 미만
24	30도 이상 35도 미만
25	35도 이상 40도 미만
26	40도 이상 45도 미만
27	45도 이상 50도 미만
…	…
32	70도 이상 75도 미만
33	75도 이상 80도 미만
34	80도 이상

타구 속도(km/h)	
그룹 번호	구간
1	10미만
2	10이상 15미만
3	15이상 20미만
…	…
30	150이상 152.5미만
31	152.5이상 155미만
32	155이상 157.5미만
33	157.5이상 160미만
…	…
40	175이상 177.5미만
41	177.5이상 180미만
42	180이상 182.5미만
43	182.5이상 185미만
…	…
48	205이상 210미만
49	210이상

배럴 정의

배럴 구간 선정

2| 세분화 2

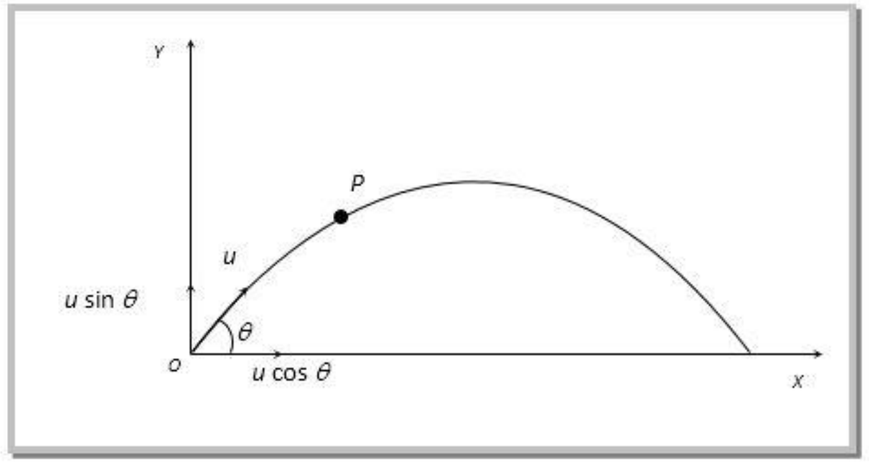
	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3	1	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	14	13	14	10	4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
12	36	41	32	42	24	10	15	2	2	1	0	0	0	0	1	0	0	0	0	0
13	86	102	128	104	121	89	66	26	16	14	5	4	1	5	0	0	0	0	0	0
14	184	201	220	243	301	323	317	135	111	96	50	36	34	10	7	2	0	1	1	0
15	301	308	360	445	541	619	714	327	304	258	221	165	124	86	36	17	9	5	1	0
16	422	436	471	590	734	929	1065	615	548	509	475	328	225	177	85	57	26	25	2	0
17	387	490	566	755	902	1173	1473	794	774	723	567	438	294	183	97	52	14	7	0	0
18	362	439	507	669	817	1051	1205	680	645	573	438	364	208	144	60	28	10	3	0	0
19	387	435	558	730	871	1070	1280	634	690	566	455	327	233	144	81	48	10	4	0	0
20	204	249	305	334	437	524	606	301	333	288	231	191	122	74	45	14	10	1	0	0
21	173	205	299	380	453	569	642	349	347	260	221	169	127	75	25	15	5	1	1	0
22	156	235	281	388	442	556	628	339	297	255	227	179	106	72	39	13	6	3	0	0
23	198	232	295	401	514	546	590	280	257	222	191	151	91	51	33	8	4	2	0	0
24	184	246	299	354	472	547	584	283	268	233	178	122	83	56	23	10	5	2	0	0
25	193	224	338	376	499	560	632	239	264	205	185	105	61	43	21	10	7	3	0	0
26	423	466	561	761	871	1042	1011	456	424	348	244	167	116	68	32	15	13	3	0	0
27	180	216	258	375	432	403	403	181	161	115	72	49	32	18	16	3	2	0	0	0
28	182	213	279	330	360	422	370	181	135	104	72	35	21	9	10	1	1	1	0	0
29	137	214	267	306	340	363	292	136	97	70	59	40	20	10	2	2	1	1	0	0
30	161	212	206	287	293	292	265	118	71	53	40	25	9	7	5	2	0	1	0	0
31	144	164	231	255	308	258	219	86	65	41	24	24	16	4	6	0	0	1	0	0
32	142	187	202	293	266	222	168	57	52	31	24	23	8	1	1	0	0	0	0	0
33	137	186	227	211	241	200	155	59	32	23	17	10	7	3	0	0	0	0	0	0
34	116	154	175	209	192	168	129	42	31	20	9	14	1	0	0	0	0	0	0	0

- [7,24], [12,37]의 경우 관측된 값이 1개이고 다른 범위와 떨어져 있기 때문에 배럴 타구로 여기지 않는다.

발사 각도(°)	
그룹 번호	구간
1	-80도 미만
2	-80도 이상 -75도 미만
3	-75도 이상 -70도 미만
…	…
20	10도 이상 15도 미만
21	15도 이상 20도 미만
22	20도 이상 25도 미만
23	25도 이상 30도 미만
24	30도 이상 35도 미만
25	35도 이상 40도 미만
26	40도 이상 45도 미만
27	45도 이상 50도 미만
…	…
32	70도 이상 75도 미만
33	75도 이상 80도 미만
34	80도 이상

타구 속도(km/h)	
그룹 번호	구간
1	10미만
2	100이상 150미만
3	150이상 200미만
…	…
30	1500이상 1525미만
31	1525이상 1550미만
32	1550이상 1575미만
33	1575이상 1600미만
…	…
39	1725이상 1750미만
40	1750이상 1800미만
41	1800이상 1850미만
42	1850이상

- 타구속도가 150km/h ~ 152.5km/h일 때, 발사각도가 30° ~ 55°이면 배럴 타구이고, 타구 속도가 2.5km/h늘어날 때마다 발사각도는 약 5°씩 범위가 증가한다.



배럴 정의

배럴 구간 선정

물리 개념을 통한 배럴 구간 확장

✓ 확장된 구간

43	44	45	46	47	48	49
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	1
0	1	0	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

실제 관측되지 않은 값에서 물리적으로 홈런이 가능한 타구들의 구간을 찾는다.

HTS 데이터의 타구속도와 발사각도를 활용하여 비거리를 계산하고, 모두 홈런이 나오는 특정 비거리를 구한다.

배럴 정의

배럴 구간 선정

물리 개념을 통한 배럴 구간 확장

m = 질량(0.145kg)
g = 중력가속도(9.80665)
v_0 = 타구속도(m/s)
θ = 발사각도($radian$)
A = 단면적(m^2)(0.00435)
ρ = 공기의 밀도(1.205)
c_d = 항력계수(0.372)

$$v_x = \frac{mv_0cos(\theta)}{\frac{1}{2}v_0cos(\theta)t + m}$$

X축 속도

$$v_y = \sqrt{\frac{2mg}{\rho Ac_d}}\tan(-\frac{t}{2m}\rho Ac_d\sqrt{\frac{2mg}{\rho Ac_d}} + tan^{-1}(\frac{v_0sin(\theta)}{\sqrt{\frac{2mg}{\rho Ac_d}}}))$$

Y축 속도

$$s_x = s_{x0} + \frac{2m}{\rho Ac_d}\left[\log\left(\frac{\rho Ac_d v_0\cos(\theta)\,t}{2} + m\right) - \log(m)\right]$$

X축 이동거리

$$s_y = s_{y0} - \frac{\rho Ac_d}{2m}\left[\ln\left|\sec(--\frac{t}{2m}\rho Ac_d\sqrt{\frac{2mg}{\rho Ac_d}} + tan^{-1}(\frac{v_0\,sin(\theta)}{\sqrt{\frac{2mg}{\rho Ac_d}}})\right| - \ln\left|\sec(tan^{-1}(\frac{v_0\sin(\theta)}{\sqrt{\frac{2mg}{\rho Ac_d}}})\right|\right]$$

Y축 이동거리

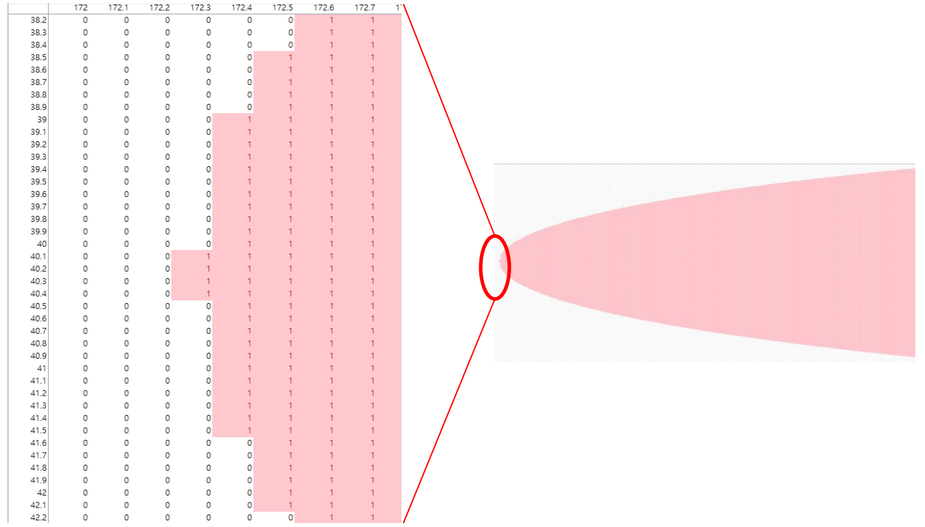
위 계산식을 통해 Y축 거리가 00이 되는 순간을 t로 놓고 X축 거리를 계산한다.

배럴 정의

배럴 구간 선정

물리 개념을 통한 배럴 구간 확장

```
> hts1[hts1$dx >= 124, "HIT_RESULT"]
[1] "홈런" "홈런" "홈런" "홈런"
```



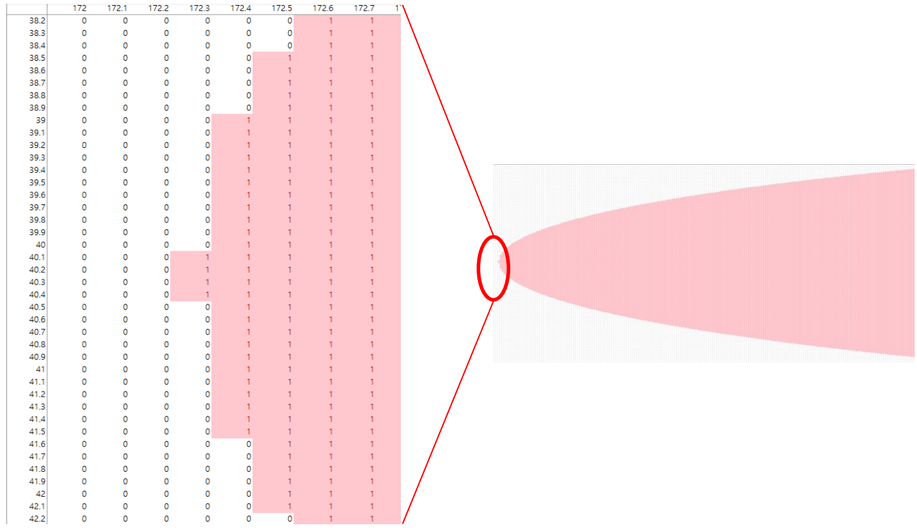
- HTS 자료에서 계산한 비거리가 124m 이상인 타구들이 모두 홈런
- 비거리가 124m 이상인 타구들은 타구속도가 172.3km/h 이상부터 발생하며, 발사각도가 40.1도~ 40.4 도부터 점차 증가하는 경향을 보인다.

배럴 정의

배럴 구간 선정

최종 구간 정의

	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3	1	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	14	13	14	10	4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
12	36	41	32	42	24	10	15	2	2	1	0	0	0	0	1	0	0	0	0	0
13	86	102	128	104	121	89	66	26	16	14	5	4	1	5	0	0	0	0	0	0
14	184	201	220	243	301	323	317	135	111	96	50	36	34	10	7	2	0	1	1	0
15	301	308	360	445	541	619	714	327	304	258	221	165	124	86	36	17	9	5	1	0
16	422	436	471	590	734	929	1065	615	548	509	475	328	225	177	85	57	26	25	2	0
17	387	490	566	755	902	1173	1473	794	774	723	567	438	294	183	97	52	14	7	0	0
18	362	439	507	669	817	1051	1205	680	645	573	438	364	208	144	60	28	10	3	0	0
19	387	435	558	730	871	1070	1280	634	690	566	455	327	233	144	81	48	10	4	0	0
20	204	249	305	334	437	524	606	301	333	288	231	191	122	74	45	14	10	1	0	0
21	173	205	299	380	453	569	642	349	347	260	221	169	127	75	25	15	5	1	1	0
22	156	235	281	388	442	556	628	339	297	255	227	179	106	72	39	13	6	3	0	0
23	198	232	295	401	514	546	590	280	257	222	191	151	91	51	33	8	4	2	0	0
24	184	246	299	354	472	547	584	283	268	233	178	122	83	56	23	10	5	2	0	0
25	193	224	338	376	499	560	632	239	264	205	185	105	61	43	21	10	7	3	0	0
26	423	466	561	761	871	1042	1011	456	424	348	244	167	116	68	32	15	13	3	0	0
27	180	216	258	375	432	403	403	181	161	115	72	49	32	18	16	3	2	0	0	0
28	182	213	279	330	360	422	370	181	135	104	72	35	21	9	10	1	1	1	0	0
29	137	214	267	306	340	363	292	136	97	70	59	40	20	10	2	2	1	1	0	0
30	161	212	206	287	293	292	265	118	71	53	40	25	9	7	5	2	0	1	0	0
31	144	164	231	255	308	258	219	86	65	41	24	24	16	4	6	0	0	1	0	0
32	142	187	202	293	266	222	168	57	52	31	24	23	8	1	1	0	0	0	0	0
33	137	186	227	211	241	200	155	59	32	23	17	10	7	3	0	0	0	0	0	0
34	116	154	175	209	192	168	129	42	31	20	9	14	1	0	0	0	0	0	0	0



	172	172.1	172.2	172.3	172.4	172.5	172.6	172.7	1
38.2	0	0	0	0	0	0	1	1	
38.3	0	0	0	0	0	0	1	1	
38.4	0	0	0	0	0	0	1	1	
38.5	0	0	0	0	0	1	1	1	
38.6	0	0	0	0	0	1	1	1	
38.7	0	0	0	0	0	1	1	1	
38.8	0	0	0	0	0	1	1	1	
38.9	0	0	0	0	0	1	1	1	
39	0	0	0	0	1	1	1	1	
39.1	0	0	0	0	1	1	1	1	
39.2	0	0	0	0	1	1	1	1	
39.3	0	0	0	0	1	1	1	1	
39.4	0	0	0	0	1	1	1	1	
39.5	0	0	0	0	1	1	1	1	
39.6	0	0	0	0	1	1	1	1	
39.7	0	0	0	0	1	1	1	1	
39.8	0	0	0	0	1	1	1	1	
39.9	0	0	0	0	1	1	1	1	
40	0	0	0	0	1	1	1	1	
40.1	0	0	0	1	1	1	1	1	
40.2	0	0	0	1	1	1	1	1	
40.3	0	0	0	1	1	1	1	1	
40.4	0	0	0	1	1	1	1	1	
40.5	0	0	0	0	1	1	1	1	
40.6	0	0	0	0	1	1	1	1	
40.7	0	0	0	0	1	1	1	1	
40.8	0	0	0	0	1	1	1	1	
40.9	0	0	0	0	1	1	1	1	
41	0	0	0	0	1	1	1	1	
41.1	0	0	0	0	1	1	1	1	
41.2	0	0	0	0	1	1	1	1	
41.3	0	0	0	0	1	1	1	1	
41.4	0	0	0	0	1	1	1	1	
41.5	0	0	0	0	1	1	1	1	
41.6	0	0	0	0	0	1	1	1	
41.7	0	0	0	0	0	1	1	1	
41.8	0	0	0	0	0	1	1	1	
41.9	0	0	0	0	0	1	1	1	
42	0	0	0	0	0	1	1	1	
42.1	0	0	0	0	0	1	1	1	
42.2	0	0	0	0	0	0	1	1	

- 타구속도가 150km/h ~ 152.5km/h일 때, 발사각도가 30도 ~ 55도이면 배럴 타구이고,
  타구 속도가 2.5km/h늘어날 때마다 발사각도의 범위가 약 5도 씩 늘어나는 경향을 보인다.

- 추가적으로 타구속도가 172.3km/h 이상이 되면
  발사각도가 40.1도 ~ 40.4도부터 증가하는 경향을 보인다.

3 데이터 소개

데이터 소개

외부데이터

Sabermetrics

야구에 게임 이론과 통계학적 방법론을 적극 도입하여 기존 야구 기록의 부실한 부분을 보완하고, 선수의 가치를 비롯한 '야구의 본질'에 대해 좀더 학문적이고 깊이 있는 접근을 시도하는 방법론

Why?

더 다양한 변수를 포함하기 위해	7월 11일 이후 데이터를 반영하기 위해	타자의 경기 능력을 직접적으로 나타내는 지표를 사용하기 위해

2018-2020

*STATIZ에서 제공하는 데이터

장타율, OPS, wRC+, WAR, WPA, BABIP, wOBA, wRC/27, wRAA

파크팩터 조정 변수
wOBA_p, wRC/27_p, wRC/27_p, wRAA_p

2021 (2021.08.16 기준)

*STATIZ에서 제공하는 데이터

출루율, 장타율, OPS, wRC+, WAR, WPA, BABIP, wOBA, wRC/27, wRAA

파크팩터 조정 변수
wOBA_p, wRC/27_p, wRC/27_p, wRAA_p

데이터 소개

사용데이터

2018 - 2020

타자 data

2018 ~ 2020

PCODE	장타율
68050	0.524
67872	0.508
67341	0.524
…	…

선수 data

2018 ~ 2021

GYEAR	T_ID	NAME
2018	KT	강백호
2019	SK	로맥
2020	WO	이정후
…	…	…

hts data

2018 ~ 2021.07.11

barrels_p
0.104816
0.122905
0.045652
…

STATIZ

2018 ~ 2021.08.16

wRC+	WAR	…	출루율	OPS
115.3	2.1	…	0.356	0.879
137.1	4.06	…	0.37	0.878
143.6	5.64	…	0.397	0.921
…	…	…	…	…

2021

PCODE	GYEAR	T_ID	NAME
76232	2021	NC	양의지
…	…	…	…

barrels_p
0.091346
…

wRC+	WAR	…	출루율	OPS
188.7	4.3	…		1.11
…	…	…	…	…

barrels_p : 배럴 수/타석

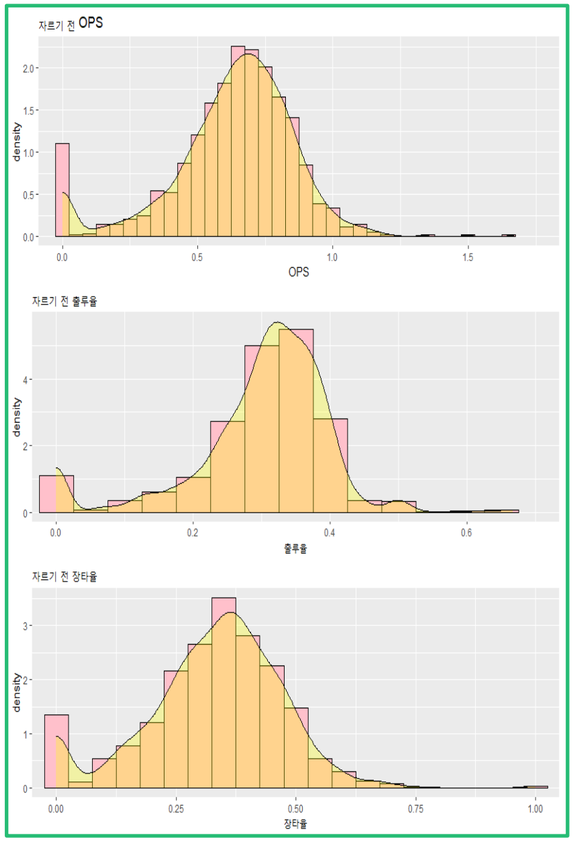
hts데이터와 STATIZ 데이터의 기간 차이를
고려하기 위해 비율로 사용

4 전처리 및 EDA

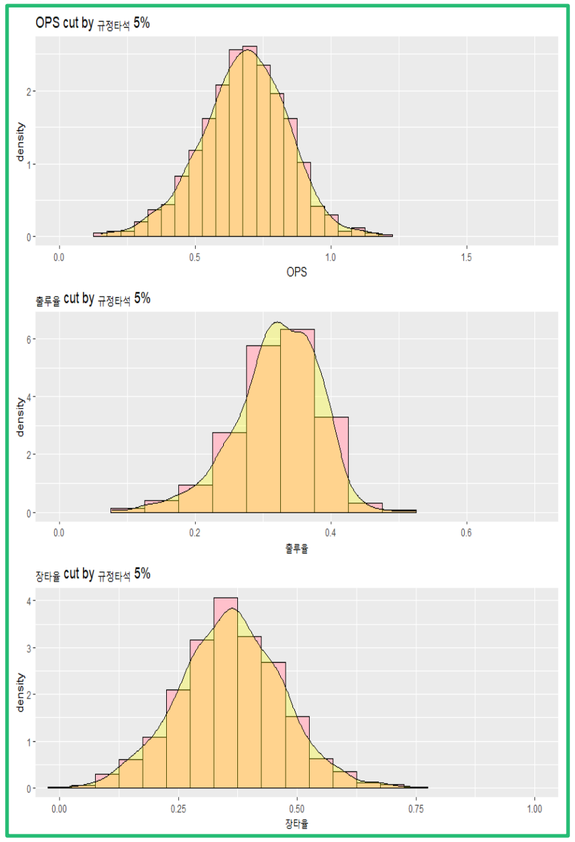
전처리 및 EDA

이상치 제거

전



후



손실률

규정타석 = 경기수 * 3.1

2018~2020년의 규정타석
144 * 3.1 = 446.4
2021년의 규정타석
79 * 3.1 = 244.9

기준점	손실률
3%	12.98%
5%	18.53%
7%	24.46%
10%	29.16%
15%	37.35%

선택 기준

정규분포의 모양을 크게 해치지 않으며,
데이터의 손실률이 너무 높지 않은 5% 선택

규정타석의 5%미만 행 제거

전처리 및 EDA

변수 선택

선택 조건

예측할 변수의 계산식에 직접적으로 포함되는 변수들은 제외

계산식

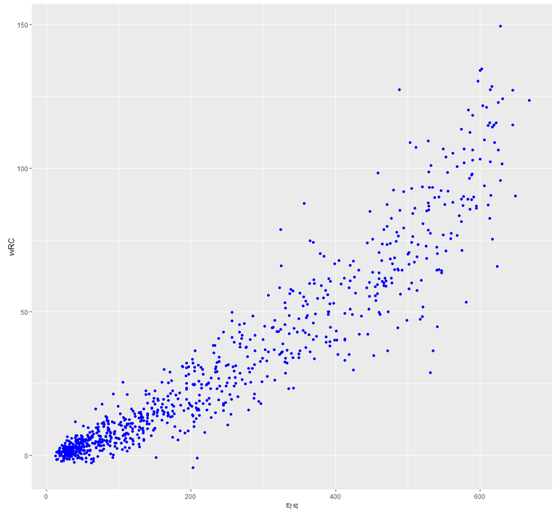
출루율 = (안타 + 볼넷 + 몸에맞는볼) / (타수 + 볼넷 +몸에맞는볼 + 희생타)

장타율 = (1루타 + 2 * 2루타 +3 * 3루타 +4 * 홈런) / 타수

OPS = 출루율 + 장타율

비율 형태의 변수들을 선택

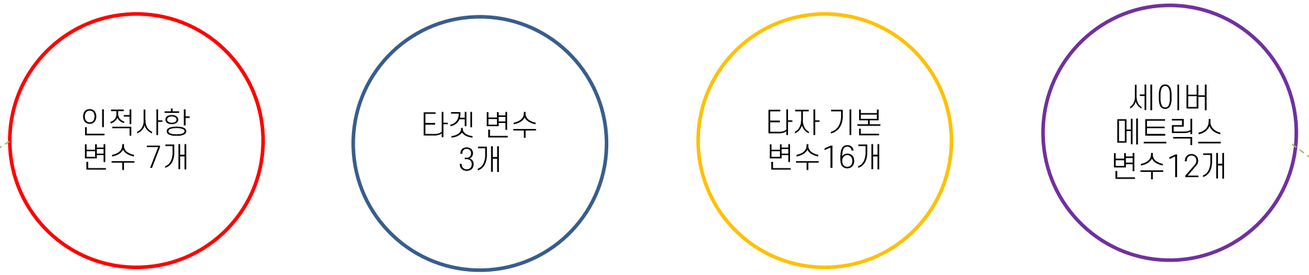
주어진 데이터를 일반화하여 해당 기간(9/15~10/8)을 예측하므로 단순히 타석 증가에 따라 값의 스케일이 영향을 받는 변수들은 제외하고 비율 형태의 변수들을 선택한다.



ex) wRC 는 타석이 많아 질수록 값이 커진다.

전처리 및 EDA

변수 선택



PCODE
50054
50054
50066
...
99810

OPS	장타율	출루율
0.532	0.246	0.286
0.486	0.200	0.286
0.617	0.367	0.250
...	...	...
0.787	0.417	0.370

타율
0.104816
0.122905
0.045652
...
0.318

WAR	BABIP	wOBA	wRC/27	wOBA_p	wRC/27_p	barrels_p
-0.25	0.259	0.261	0.24	0.259	2.15	0
-0.05	0.250	0.253	1.79	0.251	1.74	0
0.04	0.278	0.276	2.53	0.274	2.15	0
...	...	...	...	...	...	...
0.48	0.351	0.351	5.98	0.346	5.79	0.0183

전처리 및 EDA
변수 선택

타겟 변수와의 상관계수가 0.5보다 높은 변수를 선택

독립 변수 내에서 같은 유형의 변수일 경우,
상관계수가 더 높은 변수만 사용

출루율

WAR	타율	BABIP	wRC/27_p	wRC/27	wRC+	wOBA_p	wOBA	barrels_p
0.367	0.853	0.706	0.932	0.933	0.932	0.938	0.939	0.270

장타율

WAR	타율	BABIP	wRC/27_p	wRC/27	wRC+	wOBA_p	wOBA	barrels_p
0.343	0.820	0.586	0.903	0.905	0.902	0.909	0.911	0.593

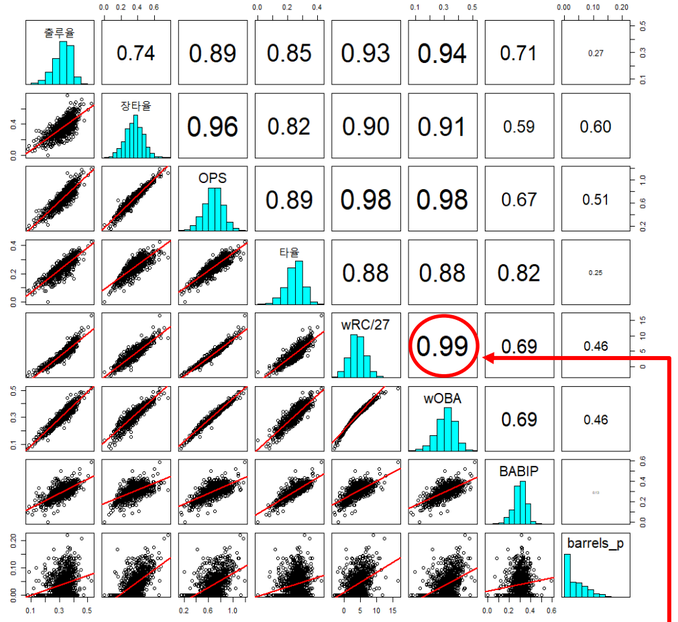
OPS

WAR	타율	BABIP	wRC/27_p	wRC/27	wRC+	wOBA_p	wOBA	barrels_p
0.404	0.887	0.672	0.974	0.976	0.973	0.981	0.982	0.508

출루율은 타율, BABIP, wRC/27, wOBA 사용
장타율은 타율, BABIP, wRC/27, wOBA, barrels_p 사용
OPS는 타율, BABIP, wRC/27, wOBA , barrels_p 사용

전처리 및 EDA
변수 선택

종속변수에 대해 상관성이 높은 변수들



이대로 모델링을 진행하면,
다중공선성 의 위험이 있음

다중공선성 확인 및 고려

종속변수

출루율

X	타율	wRC/27	wOBA	BABIP
VIF(X=x)	7.73	34.10	36.52	3.21

X	타율		wOBA	BABIP
VIF(X=x)	7.73		4.70	3.20

X	타율	wRC/27		BABIP
VIF(X=x)	7.24	4.39		3.18

모든 독립변수의
VIF ≤ 10,
다중공선성 제거

장타율

또는

OPS

X	타율	wRC/27	wOBA	BABIP	barrels_p
VIF(X=x)	8.35	34.40	37.15	3.22	1.50

X	타율		wOBA	BABIP	barrels_p
VIF(X=x)	8.33		6.44	3.22	1.47

X	타율	wRC/27		BABIP	barrels_p
VIF(X=x)	7.71	5.96		3.20	1.47

모든 독립변수의
VIF ≤ 10,
다중공선성 제거

5 모델링

Modeling

가중평균을 이용한 test data

예측기간(9/15~10/8)의 데이터는 존재하지 않기때문에 4년간 타자들의 데이터를 이용해 test dataset 생성

지수평활법 차용

$$
\begin{aligned}
F_{n+1} &= \alpha Z_n + (1-\alpha)F_n \\
&= \alpha Z_n + (1-\alpha)[\alpha Z_{n-1} + (1-\alpha)F_{n-1}] \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2 F_{n-1} \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2[\alpha Z_{n-2} + (1-\alpha)F_{n-2}] \\
&\vdots \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + \alpha(1-\alpha)^2 Z_{n-2} + \alpha(1-\alpha)^3 Z_{n-3} + \cdots
\end{aligned}
$$

시간의 흐름에 따라 최근 시계열에
더 많은 가중치를 부여하여 미래를 예측하는 방법

방법

2021년 데이터 가중치 α
2020년 데이터 가중치 $\alpha(1-\alpha)$
2019년 데이터 가중치 $\alpha(1-\alpha)^2$
2018년 데이터 가중치 $\alpha(1-\alpha)^3$

"가중평균"

*평활상수 α

평활상수 결정

지수평활법은 평활상수에 따라 각 시점의 가중치가
정해지기 때문에 평활상수의 설정이 중요

1| 4년치의 데이터가 모두 존재하는 타자들의 데이터만 선택

2| 지수평활법을 차용한 가중치를 설정하여 변수별로
RMSE가 가장 작은 평활상수 설정

평활상수는 0.1부터 0.9까지 지정하며

이때, 2020년 데이터 가중치 α
2019년 데이터 가중치 $\alpha(1-\alpha)$
2018년 데이터 가중치 $\alpha(1-\alpha)^2$
로 설정한다.

Modeling

가중평균을 이용한 test data

평활상수 결정

α가 0.3일때 rmse가 최소인 변수	타율	0.05324
	BABIP	0.06225
α가 0.4일때 rmse가 최소인 변수	wOBA	0.05078
	wRC/27	2.03473
	wOBA_p	0.05091
α가 0.5일때 rmse가 최소인 변수	wRC+	34.47565
	wRC/27	1.49046
	wOBA_p	0.03021
α가 0.6일때 rmse가 최소인 변수	barrels_p	0.03017

Test Dataset

	PCODE	이름	타율_test	wRC._test	WAR_test	BABIP_test	wOBA_test	wRC.27_test	wOBA_p_test	wRC_p_test	barrels_p_test
41	67341	이정후	0.343271	149.640000	4.687333	0.358685	0.415449	8.333934	0.415202	89.993333	0.026285
42	67872	로맥	0.269257	126.953333	2.732000	0.287474	0.391559	7.139338	0.386673	77.473333	0.112476
43	68050	강백호	0.352985	170.860000	4.541333	0.400642	0.447151	10.362757	0.444298	94.713333	0.073721
61	75847	최정	0.276262	149.793333	4.397333	0.283749	0.420735	8.480735	0.415673	81.226667	0.109189
62	76232	양의지	0.348146	175.980000	5.189333	0.334920	0.460485	10.350588	0.455404	89.760000	0.089261
65	76290	김현수	0.310235	137.240000	3.216667	0.310916	0.389956	7.154816	0.400603	80.693333	0.104083
77	78224	김재환	0.282050	146.346667	3.742667	0.322358	0.399783	7.579154	0.412202	81.553333	0.144711
79	78513	전준우	0.305999	124.113333	3.047333	0.320700	0.384732	6.740037	0.380886	76.826667	0.061376
86	79192	채은성	0.311482	140.033333	2.676667	0.335627	0.389162	6.896949	0.400901	63.306667	0.085554
87	79215	박건우	0.320540	130.740000	2.992000	0.353663	0.384180	6.713493	0.389393	65.873333	0.061179

Modeling
모델링 방법

1 | 모델 선정이유

행의 수(1063)가 많지 않기 때문에
과적합의 위험이 적은 모델 사용

Random Forest **Xgboost**

변수 중요도를 통해 사용 변수와
OPS, 장타율, 출루율의 관계를 파악

2 | 모델링 순서

1

wRC/27과 wOBA 사이
다중공선성 존재 우려

wRC+를 제외한 모델과
wOBA를 제외한 모델을 생성

2

OPS, 장타율, 출루율을
각각 예측하는 모델 생성

예측 정확도가 높은
2개의 종속변수만 선택 후
계산식을 통해 예측값 도출

*장타율 + 출루율 = OPS

3

위 과정을 Xgboost와
RF 모델로 반복해서 수행

데이터셋을 8:2로 나누고
성능 평가를 통해 최종 모델 선택

Modeling

모델링 평가

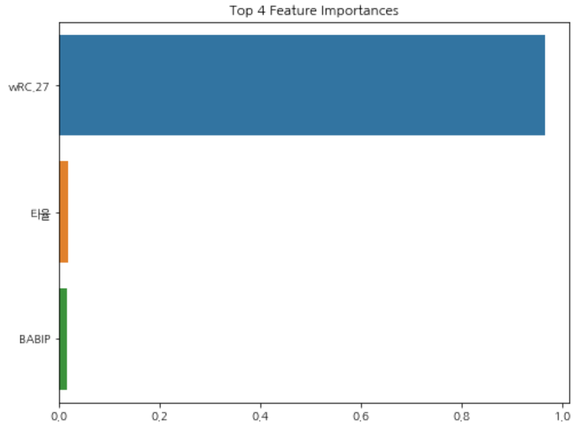
종속변수	독립변수	Random Forest	Xgboost	
출루율	wRC/27, 타율, BABIP	R-squared : 0.8892 MAE : 0.0146	R-squared : 0.8963 MAE : 0.0141	✓
	wOBA, 타율, BABIP	R-squared : 0.8702 MAE : 0.0158	R-squared : 0.8809 MAE : 0.0150	
장타율	wRC/27, 타율, BABIP, barrels_p	R-squared : 0.8648 MAE : 0.0297	R-squared : 0.8691 MAE : 0.0293	장타율의 성능이 가장 낮은 것을 확인
	wOBA, 타율, BABIP, barrels_p	R-squared : 0.8776 MAE : 0.0285	R-squared : 0.8789 MAE : 0.0284	
OPS	wRC/27, 타율, BABIP, barrels_p	R-squared : 0.9565 MAE : 0.0236	R-squared : 0.9632 MAE : 0.0213	
	wOBA, 타율, BABIP, barrels_p	R-squared : 0.9637 MAE : 0.0231	R-squared : 0.9707 MAE : 0.0211	✓

Modeling

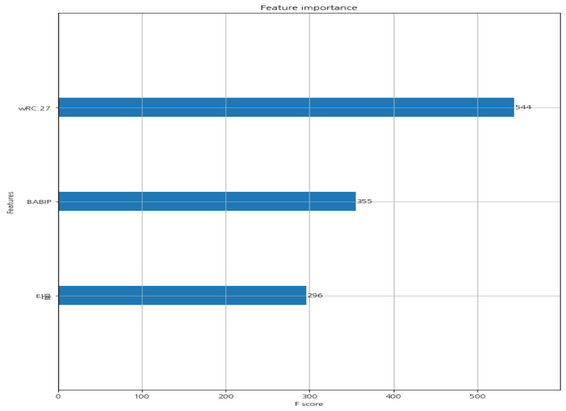
변수중요도

1 | 출루율 ~ wRC/27 + 타율 + BABIP

Random Forest

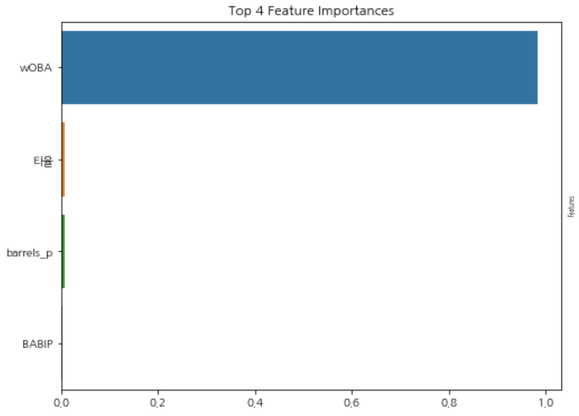


Xgboost

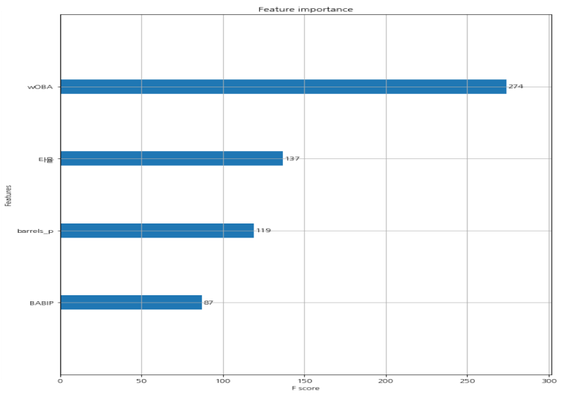


2 | OPS ~ wOBA + 타율 + BABIP + barrels_p

Random Forest



Xgboost



> “ Xgboost의 성능이 Random Forest보다 뛰어나며,
> 예측에 3가지 변수가 적절하게 반영되고 있음을 확인 ”

Modeling

장타율 모델비교

08.17부터 09.13까지의 장타율 데이터를 활용해 모델의 오차와 계산 오차를 비교

사용 모델

Xgboost : 장타율 ~ wOBA + 타율 + BABIP + barrels_p

```
xgboost = xgb.XGBRegressor(colsample_bytree = 0.7,
                           learning_rate = 0.01, max_depth = 3, n_estimators = 500,
                           subsample = 0.5, seed = 1234, objective = 'reg:squarederror')
```

<모델 성능>
R-squared : 0.8789
MAE : 0.0284

오차 비교

	PCODE	이름	예측장타율	실제장타율	계산장타율	model_error	calculate_error
1	67872	로맥	0.504521	0.297	0.498230	0.207521	0.201230
2	68050	강백호	0.576962	0.551	0.558732	0.025962	0.007732
3	75847	최정	0.550158	0.662	0.567963	0.111842	0.094037
4	76232	양의지	0.631248	0.440	0.617700	0.191248	0.177700
5	76290	김현수	0.524798	0.452	0.500728	0.072798	0.048728
6	78224	김재환	0.530622	0.488	0.539046	0.042622	0.051046
7	78513	전준우	0.487915	0.405	0.490815	0.082915	0.085815
8	79192	채은성	0.510634	0.417	0.496845	0.093634	0.079845
9	79215	박건우	0.480640	0.474	0.477320	0.006640	0.003320

<평균 오차>

model error	0.0928
calculate error	0.0833

계산식을통해 도출한 error가더 작으므로
계산식을 활용해 장타율 예측

Modeling

최종 모델

출루율

출루율 ~ wRC/27 + 타율 + BABIP

```
#하이퍼 파라미터 튜닝
params = { 'max_depth': [3, 5, 10],
           'learning_rate': [0.01, 0.05, 0.1],
           'n_estimators': [100, 200, 300, 500],
           'colsample_bytree': [0.5, 0.7, 1],
           'subsample' : [0.5, 0.7, 1]}

xgboost = xgb.XGBRegressor(objective='reg:squarederror', seed = 1234)

clf = GridSearchCV(estimator = xgboost,
                   param_grid = params,
                   scoring = 'neg_mean_squared_error',
                   cv = 4,
                   verbose = 1)

clf.fit(출루율_X, 출루율_y)

print("Best parameters:", clf.best_params_)
print("Lowest RMSE: ", (-clf.best_score_)**(1/2.0))`
```

```
Fitting 4 folds for each of 324 candidates, totalling 1296 fits
Best parameters: {'colsample_bytree': 1, 'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 200,
'subsample': 0.5}
Lowest RMSE: 0.019715239711205316
```

R-squared	0.8963
MSE	0.00036
MAE	0.0141

OPS

OPS ~ wOBA + 타율 + BABIP + barrels_p

```
#하이퍼 파라미터 튜닝
params = { 'max_depth': [3, 5, 10],
           'learning_rate': [0.01, 0.05, 0.1],
           'n_estimators': [100, 200, 300, 500],
           'colsample_bytree': [0.5, 0.7, 1],
           'subsample' : [0.5, 0.7, 1]}

xgboost = xgb.XGBRegressor(objective='reg:squarederror', seed = 1234)

clf = GridSearchCV(estimator = xgboost,
                   param_grid = params,
                   scoring = 'neg_mean_squared_error',
                   cv = 4,
                   verbose = 1)

clf.fit(OPS_X, OPS_y)

print("Best parameters:", clf.best_params_)
print("Lowest RMSE: ", (-clf.best_score_)**(1/2.0))`
```

```
Fitting 4 folds for each of 324 candidates, totalling 1296 fits
Best parameters: {'colsample_bytree': 1, 'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 200,
'subsample': 0.5}
Lowest RMSE: 0.02751851442960648
```

R-squared	0.9707
MSE	0.00072
MAE	0.0211

6 결론 및 한계점

결론 및 한계점

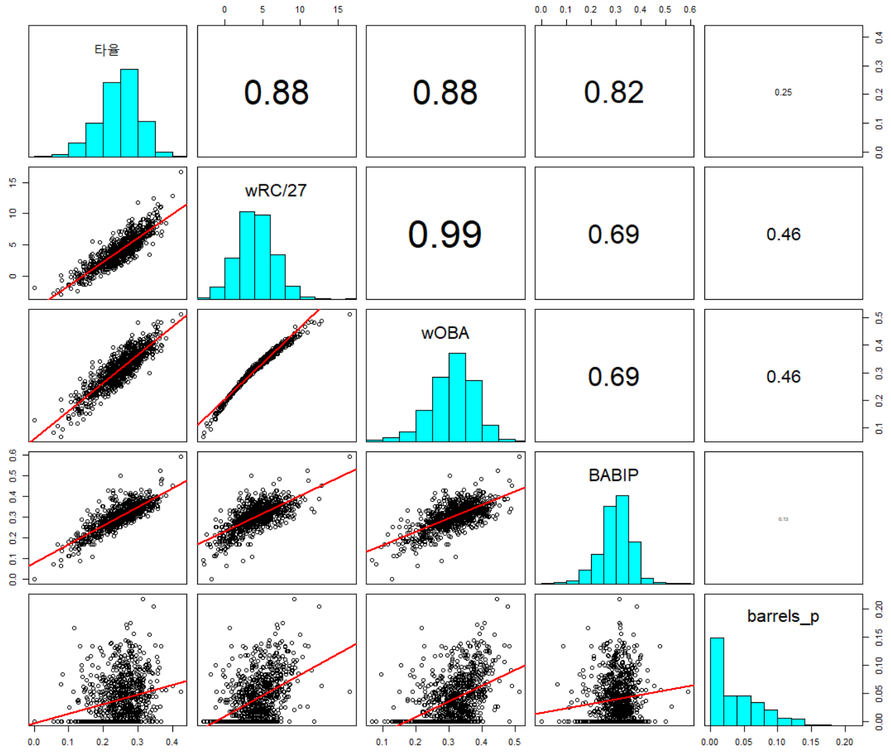
예측 결과

PCODE	이름	출루율	OPS	장타율
67341	이정후	0.409944	0.927237	0.517292
67872	로맥	0.381083	0.879313	0.498230
68050	강백호	0.458383	1.017115	0.558732
75847	최정	0.396734	0.964698	0.567963
76232	양의지	0.439308	1.057008	0.617700
76290	김현수	0.383300	0.884029	0.500729
78224	김재환	0.382095	0.921141	0.539046
78513	전준우	0.375059	0.865875	0.490816
79192	채은성	0.384273	0.881118	0.496846
79215	박건우	0.383252	0.860573	0.477320

결론 및 한계점

결론

1 | 객관성이 뚜렷한 지표 barrel_p

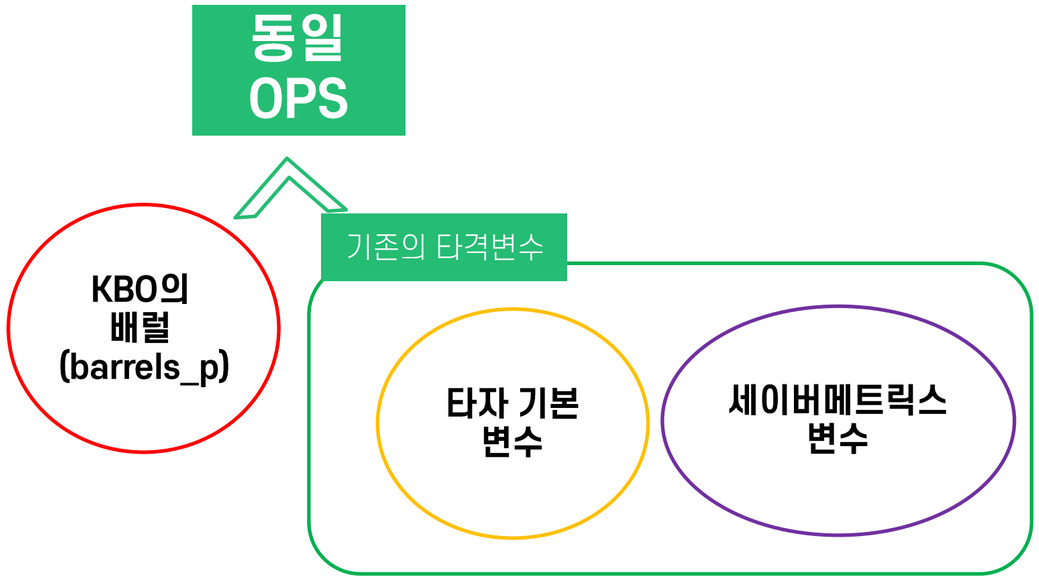


BABIP 와 타변수들 간의 상관성이 대체로 높지만, BABIP 와 barrels_p의 상관성은 매우 낮다. BABIP 의 뜻이 인플레이에서의 타구의 타율 임을 고려할 때, BABIP가 운의 척도로 인식되어지므로 배럴은 운의 요소가 배제된 타자의 순수한 능력의 척도로 판단할 수 있다.

결론 및 한계점

결론

2 | 타자의 순수실력에 대한 평가



Case 1)

동일 OPS를 갖고 있을 때, barrles_p 가 높고 기존의 타격변수가 낮다면 타자의 순수실력이 더 높다고 판단할 수 있다.

Case 2)

반대로 동일 OPS를 갖고 있을 때, barrles_p 가 낮고 기존의 타격변수가 높다면 타자의 순수실력이 더 낮다고 판단할 수 있다.

“ 동일한 OPS를 갖고 있을 때, 배럴 비율과 기존의 타격변수의 대소비교를 통해 타자의 순수 실력을 평가할 수 있다. ”

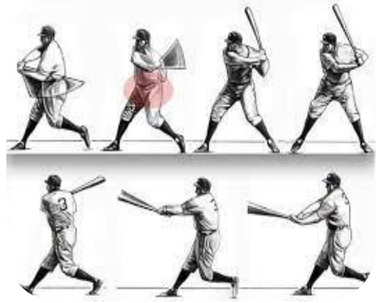
결론 및 한계점

결론

3 | 활용방안

경기 멤버 재구성 및 새로운 교체 전략 수립

유망주의 발전가능성 확인



선수들의 효과적인 타격방법 연구

결론 및 한계점

결론

4 | 한계점

1

트래킹 데이터의 부족

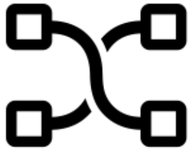
배럴 변수를 만들 때,
hts 트래킹 데이터를
활용해야만
했기 때문에 더 많은 기간으로
확장할 수 없었다.

2

시계열성 판단

시계열성을 충분히 반영하지
못하여 타자의 급격한 부진을
예측하기 어려움

3



모델링

모델이 학습을 잘 하여도,
가중평균 데이터셋을 이용하므로
더 많은 에러가 발생할 수 있음.

감사합니다