2020 빅콘테스트 데이터분석 분야
퓨처스 리그

건너건너

[팀장]

안희재 (heejae2563@naver.com)

숙명여자대학교 통계학과 4학년

[팀원]

김민정 (kmmnjng528@gmail.com)

숙명여자대학교 통계학과 3학년

김채현 (7chaney25@gmail.com)

숙명여자대학교 통계학과 3학년

이혜원 (hyewon903@gmail.com)

숙명여자대학교 통계학과 3학년

Contents

EDA

승률

사용할 데이터 불러오기

1) 제공 데이터

스포츠투아이_제공데이터(.CSV)_시즌별, 시트별 구분.csv

- 2016 ~2020년 선수 csv
- 2016 ~ 2020년 팀타자 csv
- 2016 ~ 2020년 팀투수 csv

2) 추가로 사용한 데이터

- 2015 ~ 2019년 각 팀별 승률
- 2015 ~ 2019년 각 팀의 상대팀에 대한 상대승률

데이터 크기 및 결측치 확인

1) 크기

(6400, 22)

2) 결측치

선수 데이터 셋에서 결측치 발견

- 연봉, 계약금 없이 옵션으로 계약을 체결한 선수.

(WO팀 밴헤켄선수)

- 특수한 경우이므로 제거하고 진행

타율 & 방어율

사용할 데이터 불러오기

1) 제공 데이터

스포츠투아이_제공데이터(.CSV)_시즌별, 시트별 구분.csv

- 2016 ~ 2020년 팀타자 csv
- 2016 ~ 2020년 팀투수 csv

데이터 크기 및 결측치 확인

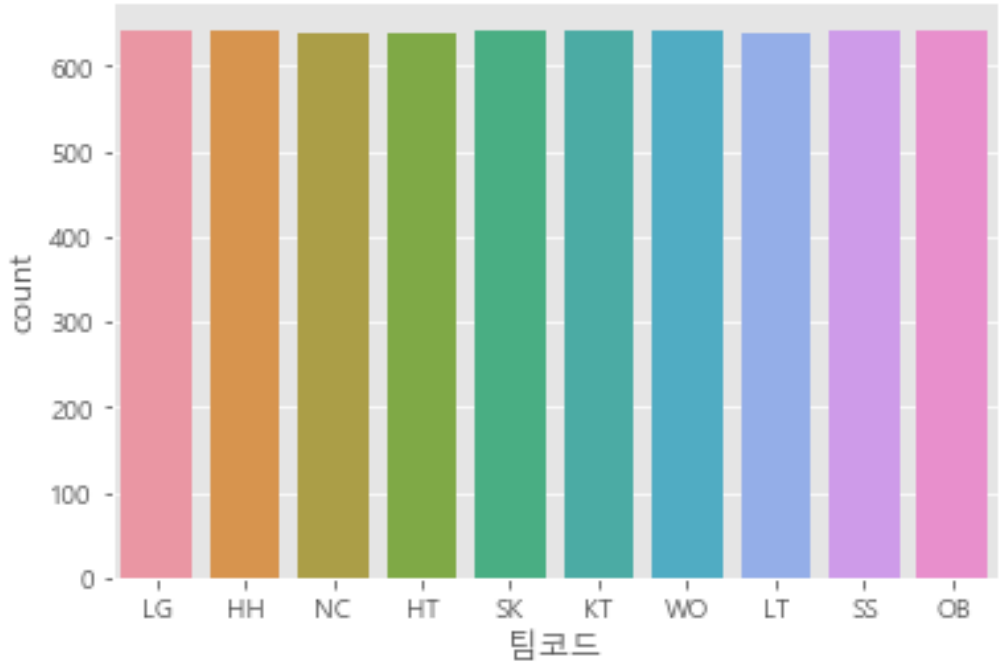
1) 크기

(6400, 28)

2) 결측치

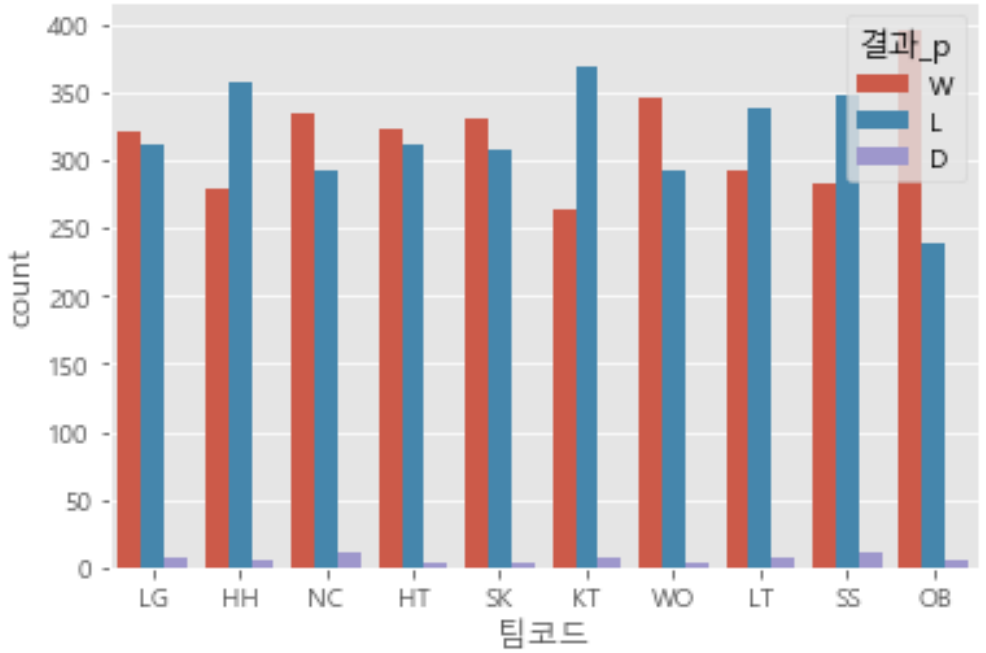
결측치 없음

팀별 경기수



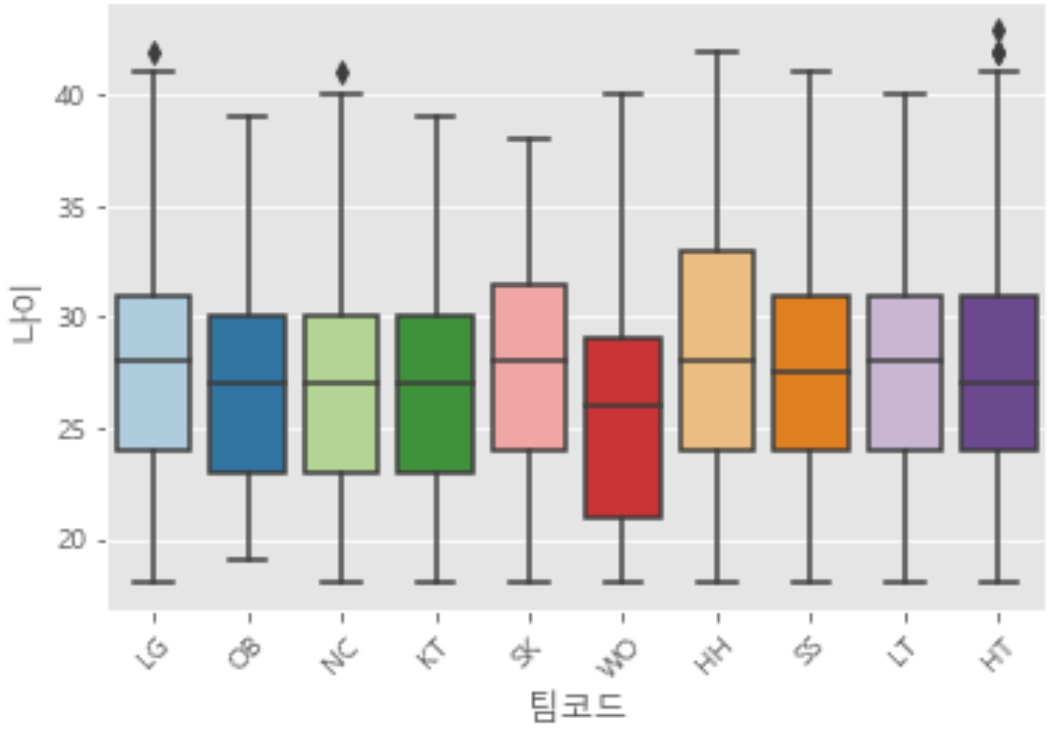
야구 경기에 참여하는 팀은
총 10팀으로 경기 수는 모두 비슷함

팀별 경기 결과



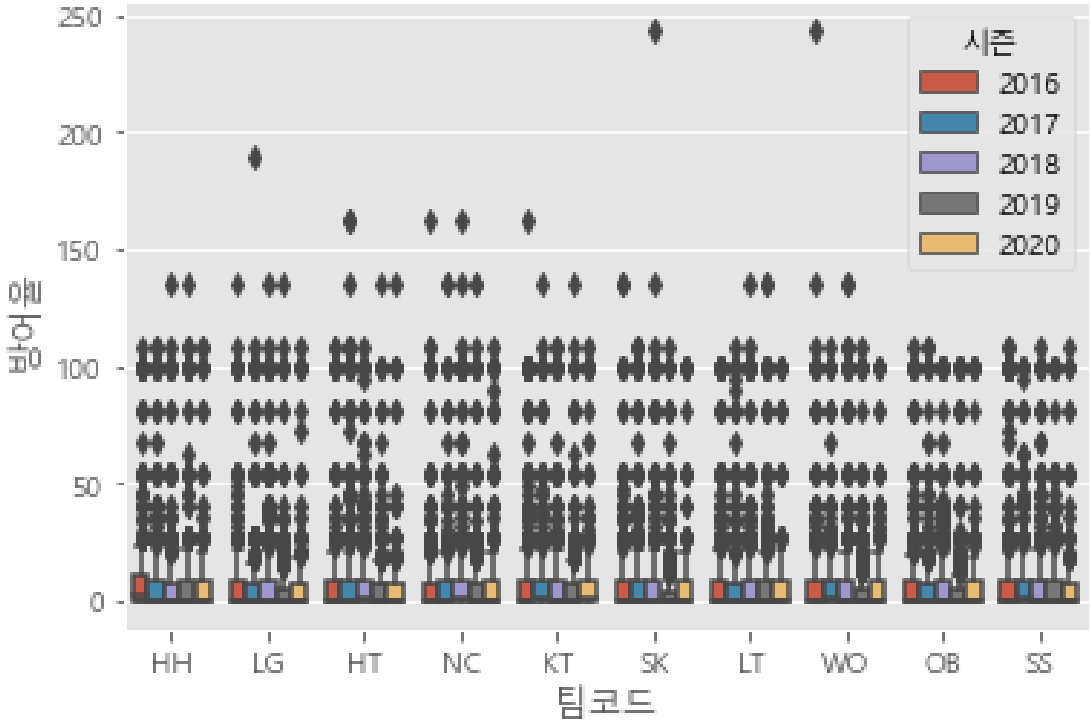
팀 LG, NV, HT, SK, WO, OB는
승리의 횟수가 패배의 횟수보다 많음

팀별 선수들의 나이 상자 그림



주로 20대 중반부터 20대 후반의 나이대로 분포함

방어율 이상치 판단 boxplot



Feature Engineering

팀투수

변수명	계산 방법	비고
피안타율_p	안타수 / 타수	타수가 0인경우 99.9
평균자책점_p	자책점 / (이닝*3 / 3)*9	이닝*3이 0인 경우 99.9
실점/이닝_p	실점 / (이닝*3 / 3)*9	이닝*3이 0인 경우 99.9
삼진/이닝_p	삼진 / (이닝*3 / 3) *9	이닝*3이 0인 경우 99.9
볼넷/이닝_p	4구 / (이닝*3 / 3) *9	이닝*3이 0인 경우 99.9
삼진/볼넷_p	삼진 / 4구	4구의 수가 0인 경우 99.9
피홈런/이닝_p	홈런 / (이닝*3 / 3) *9	이닝*3이 0인 경우 99.9
타자/이닝_p	타자 / (이닝*3 / 3) *9	이닝*3이 0인 경우 99.9
WHIP_p	안타+4수 / 이닝수	이닝*3이 0인 경우 99.9 고의사구 포함, 사구 제외

팀타자

변수명	계산	비고
출루율_h	(안타 + 사구 + 4구) / (타수 + 사구 + 4구)	(타수 + 사구 + 4구) 가 0 인 경우 99.9
장타율_h	(안타 + 2루타 + 3루타 + 4루타 + 홈런) / 타수	타수가 0인 경우 99.9
타율_h	안타 / 타수	타수가 0인 경우 99.9
사구/타수_h	(사구 + 4구) / 타수	타수가 0인 경우 99.9
삼진/타수_h	삼진 / 타수	타수가 0인 경우 99.9

선수

변수명	설명	비고
나이	시즌별, 각 팀별, 선수들의 나이의 평균	
연봉	시즌별, 각 팀별, 선수들의 연봉의 평균	단위 : 만원

추가 사용 피처

변수명	설명	비고
전년도승률	팀별 전년도 경기 승률	KBO 홈페이지 참고
전년도상대승률	팀별 전년도 경기 상대승률	KBO 홈페이지 참고

사용하지 않는 피처 제거 - 팀타자

변수명	정의	비고
더블헤더코드	두 팀이 같은 날 계속해서 두 경기를 치르는 것	0,1,2의 값을 가짐
초말	Bottom 말 Top 초	
2루타_h		
3루타_h		
홈런_h		
4구_h	타자가 볼카운트에서 4개의 볼을 얻는 상황	
사구_h	타자의 몸에 공이 맞은 경우	
삼진_h	스트라이크를 세번 당하는 것	

사용하지 않는 피처 제거 – 팀투수

변수명	계산 방법	비고
더블헤더코드	두 팀이 같은 날 계속해서 두 경기를 치르는 것	0,1,2의 값을 가짐
초말	Bottom 말 Top 초	
타자_p	타자가 타석에 선 모든 타석수의 합	
타수_p	총 타석수에서 볼넷, 사구, 희생번트, 희생플라이, 타격방해, 주루방해 6가지를 제외한 횟수	
안타_p	2루타, 3루타, 홈런을 포함	
2루타_p, 3루타_p, 홈런_p		
4구_p	투수가 스트라이크존을 벗어나는 투구를 4번하게 됨	
사구_p	투수가 던진 볼에 몸을 맞아 타자가 1루로 진루	
삼진_p	스트라이크를 세번 당하는 것	

범주형 자료 변환

범주형 변수들을 숫자로 변환

변환 전	변환 후	변환 전	변환 후
HH	0	NC	5
HT	1	OB	6
KT	2	SK	7
LG	3	SS	8
LT	4	WO	9

시즌 변수 생성

올스타전 날짜 기준으로 상반기와 하반기로 분할

기준 날짜		
2016년 상반기	2016년 7월 13일 이후	2016년 하반기
2017년 상반기	2017년 7월 13일 이후	2017년 하반기
2018년 상반기	2018년 7월 12일 이후	2018년 하반기
2019년 상반기	2019년 7월 18일 이후	2019년 하반기
2020년 상반기	2020년 7월 20일 이후	2020년 하반기

시행착오

IDEA 1. 팀코드/상대팀코드 기준으로 데이터 나누기

한 팀이 어떤 팀을 맡게 되는지가 중요하다고 생각됨
' 팀코드/상대팀코드' 기준으로 (LG vs HH, LG vs HT, ...) 데이터를 나눈 후 분석

HH vs HT

```
mth_HH_HT = mth[(mth['팀코드'] == 'HH') & (mth['상대팀코드'] == 'HT')]
mth_HH_HT['구장'] = 0
mth_HH_HT['구장'].loc[(mth['홈팀코드'] == 'HH')] = 1
mth_HH_HT.head()
```

	게임키	일자	원정팀코드	홈팀코드	더블헤더코드	요일	구장	팀코드	상대팀코드	초말	타자	타수	타점	득점	안타	2루타	3루타	홈런	도루	도루실패	희타	희비	4구	고4
128	20180410HTHH0	20180410	HT	HH	0	화	1	HH	HT	B	36	30	4	4	7	0	0	2	1	0	0	0	4	0
138	20180411HTHH0	20180411	HT	HH	0	수	1	HH	HT	B	37	31	6	6	11	1	0	0	2	1	0	1	3	0
148	20180412HTHH0	20180412	HT	HH	0	목	1	HH	HT	B	44	40	15	15	17	6	0	2	0	0	0	0	3	0
253	20180425HHHT0	20180425	HH	HT	0	수	0	HH	HT	T	36	34	3	3	6	1	0	1	0	0	0	0	2	0
263	20180426HHHT0	20180426	HH	HT	0	목	0	HH	HT	T	37	34	2	3	7	1	0	0	0	0	0	0	3	0

IDEA 2. 경기 구장이 홈인지 원정인지를 기준으로 데이터 나누기

홈팀에서 경기하는것과 원정에서 경기를 하는것에 차이가 있다고 판단됨
홈인 경우 1, 원정인 경우 0으로 '구장' 피처 생성

HH vs LG ¶

```
mth_HH_LG = mth[(mth['팀코드'] == 'HH') & (mth['상대팀코드'] == 'LG')]
mth_HH_LG['구장'] = 0
mth_HH_LG['구장'].loc[(mth['홈팀코드'] == 'HH')] = 1
mth_HH_LG.head()
```

	게임키	일자	원정팀코드	홈팀코드	더블헤더코드	요일	구장	팀코드	상대팀코드	초말	타자	타수	타점	득점	안타	2루타	3루타	홈런	도루	도루실패	희타	희비	4구	고4	
306	20180501LGHH0	20180501	LG	HH	0	화	1	HH	LG	B	39	34	6	6	11	0	1	3	0	1	1	0	3	0	
314	20180502LGHH0	20180502	LG	HH	0	수	1	HH	LG	B	36	34	4	4	9	3	0	1	1	0	0	0	2	0	
324	20180503LGHH0	20180503	LG	HH	0	목	1	HH	LG	B	36	31	6	7	11	2	0	1	0	0	1	1	3	1	
431	20180518HHLG0	20180518	HH	LG	0	금	0	HH	LG	T	35	32	3	4	10	0	0	1	0	1	0	0	2	0	
441	20180519HHLG0	20180519	HH	LG	0	토	0	HH	LG	T	34	31	2	2	6	2	0	0	0	0	1	0	1	0	

IDEA 3. 타자, 투수 데이터 관련

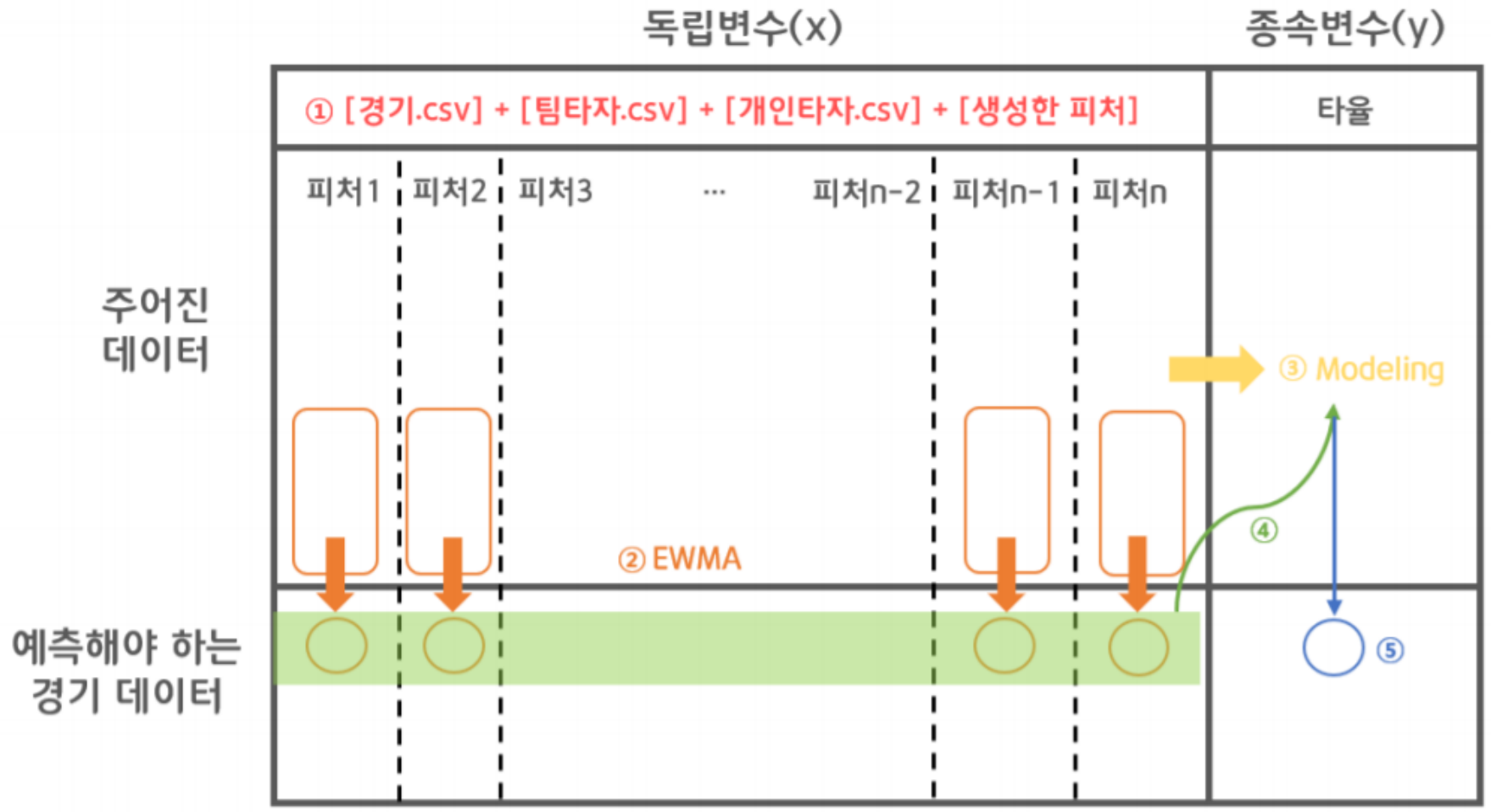
타자

- 타자의 경우 보통 나오는 선수들이 비슷하게 나오기 때문에 팀 데이터 사용

투수

- 투수의 경우 보통 5명이 돌아가면서 경기를 나오기 때문에 그 선수들의 개인 데이터 이용
- 구원투수의 경우에는 어느 선수가 나올지 예측하기 힘드므로 돌아가면서 나오는 투수를 제외한 나머지 투수들의 데이터를 평균내서 이용

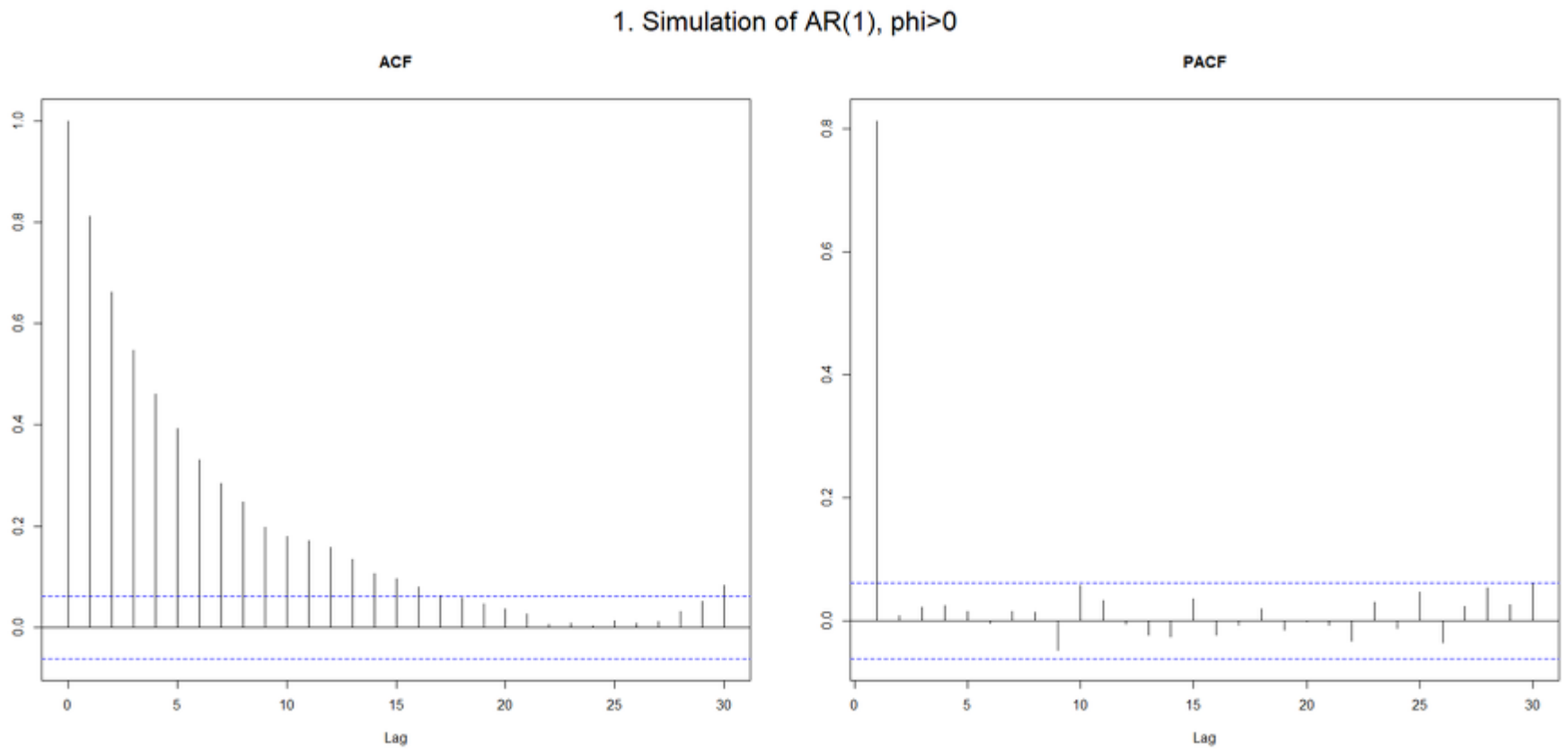
IDEA 4. 전년도 경기 결과에 가중주기



[승률, 방어율도 동일한 흐름으로 진행]

IDEA 4. 전년도 경기 결과에 가중주기

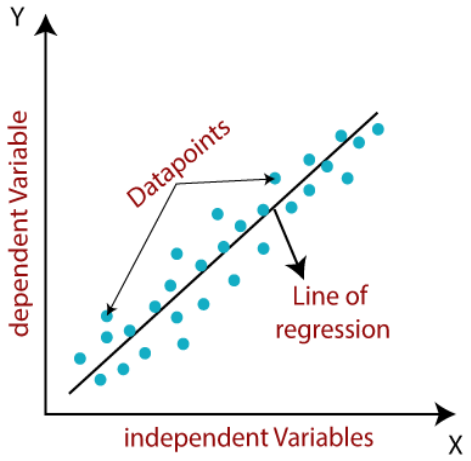
시계열모형 AR(1)



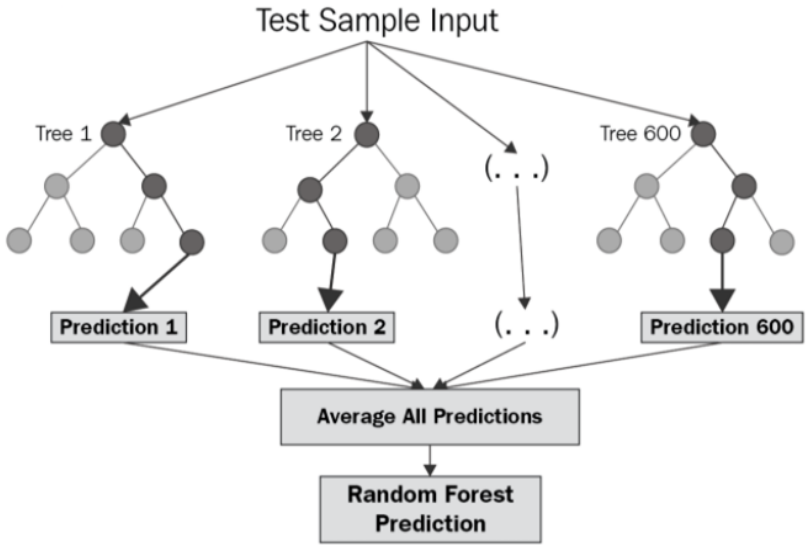
IDEA 5. 선발투수를 고려하여 가중주기

Algorithm

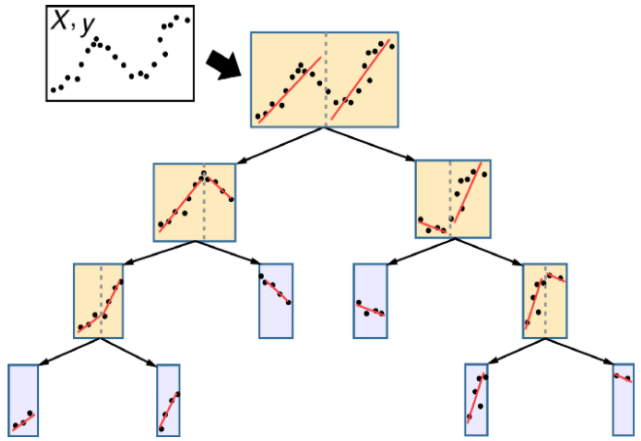
Linear Regression



Random Forest Regressor

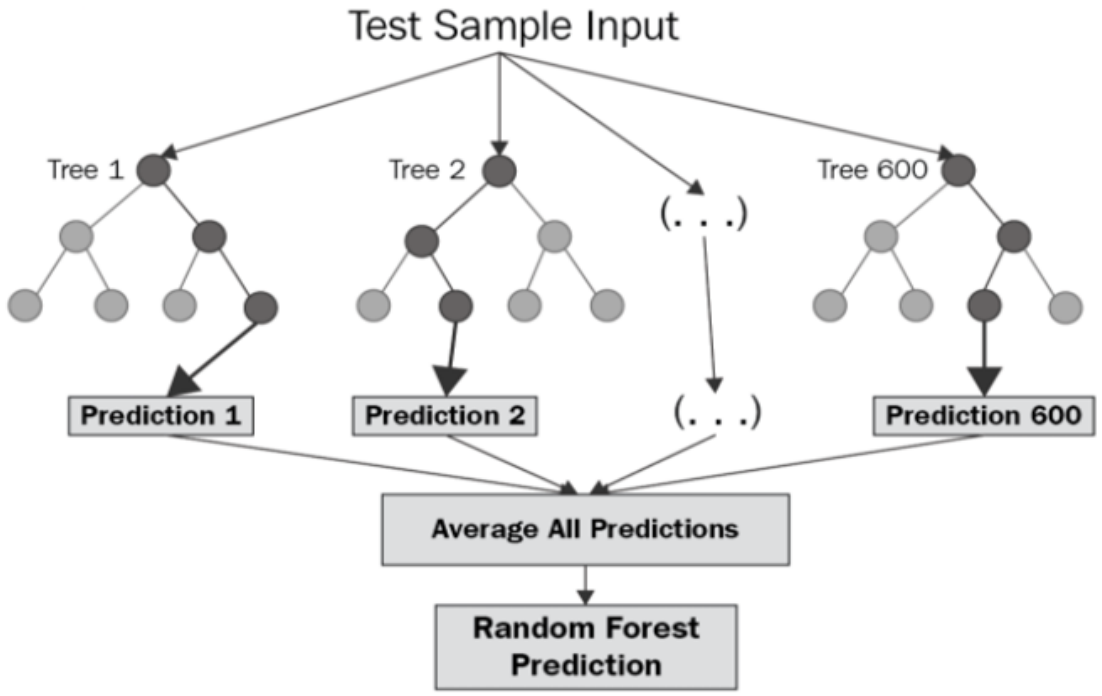


Decision Tree Regressor



Random Forest Regressor

같은 알고리즘으로 여러 개의 분류기를 만들어서 보팅으로 최종 결정하는 배깅(bagging)의 대표적인 알고리즘

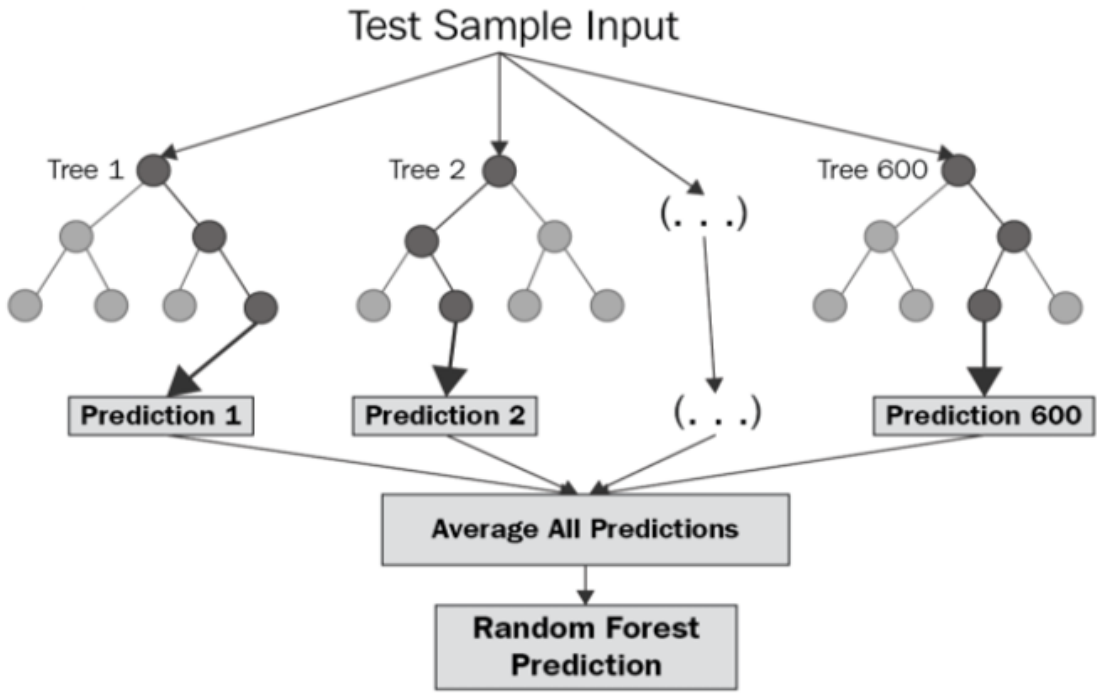


- B개의 동일하게 분포된 데이터셋을 만들고, 거기서 트리를 깊게 만든다.
  : Bias가 작고 분산이 큰 트리

- B개의 트리의 평균을 취한다.
  : 평균을 취함으로써, bias는 개개의 트리에 대해서 같은 상태로 남게 되고, 분산은 줄어든다.

Random Forest Regressor

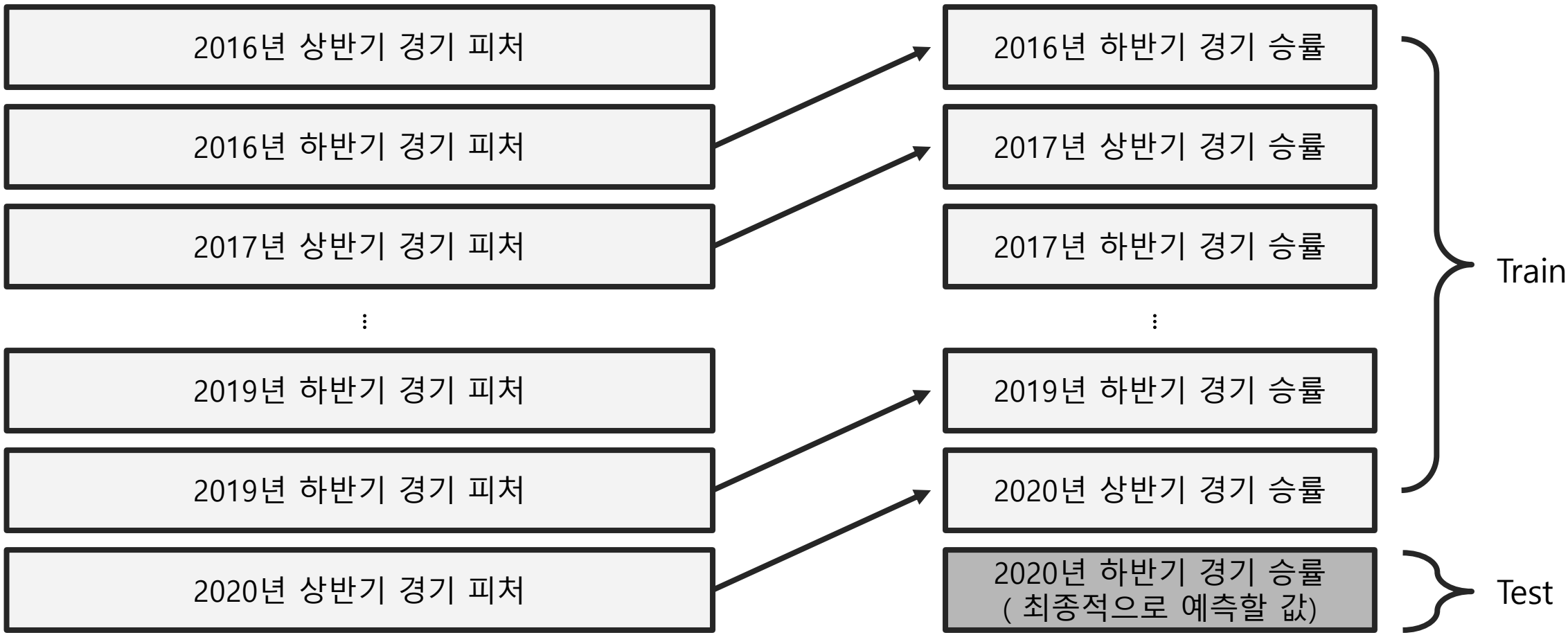
같은 알고리즘으로 여러 개의 분류기를 만들어서 보팅으로 최종 결정하는 배깅(bagging)의 대표적인 알고리즘



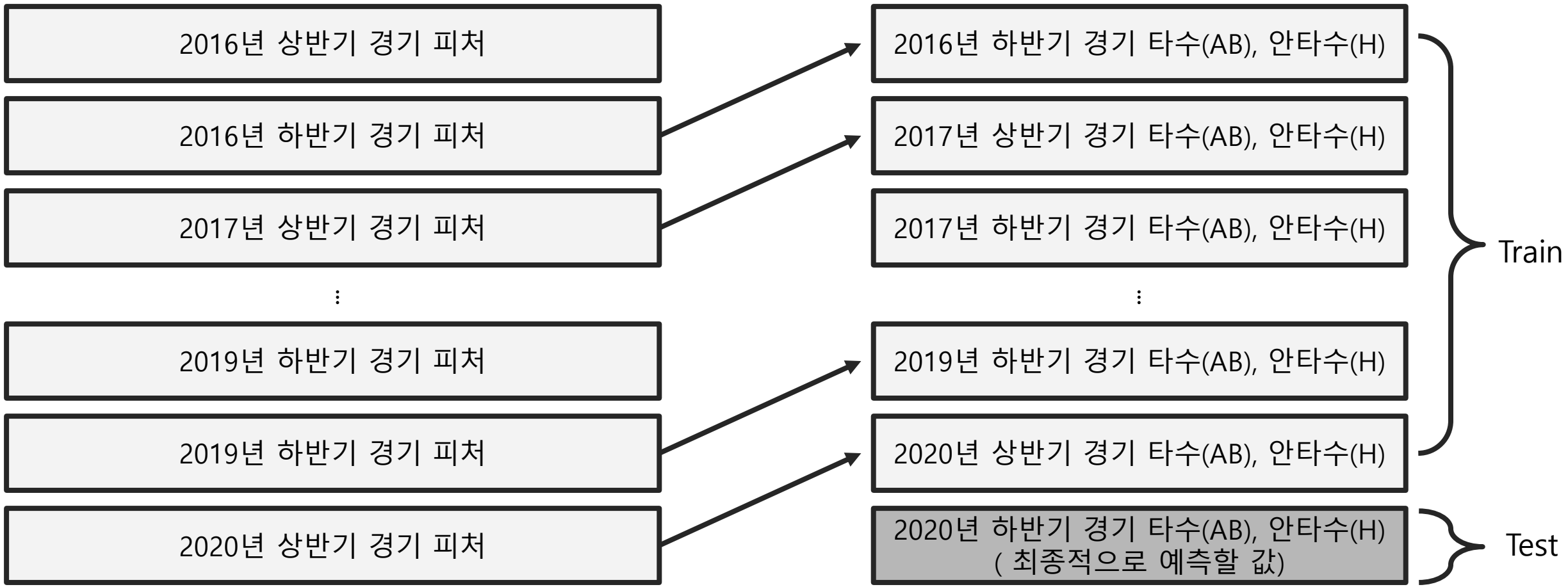
- **앙상블 알고리즘** 중 비교적 빠른 수행 속도를 가지고 있으며, 다양한 영역에서 높은 예측 성능을 보임

- 여러 개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링하여 개별적으로 학습을 수행한 뒤 최종적으로 모든 분류기가 보팅을 통해 예측 결정

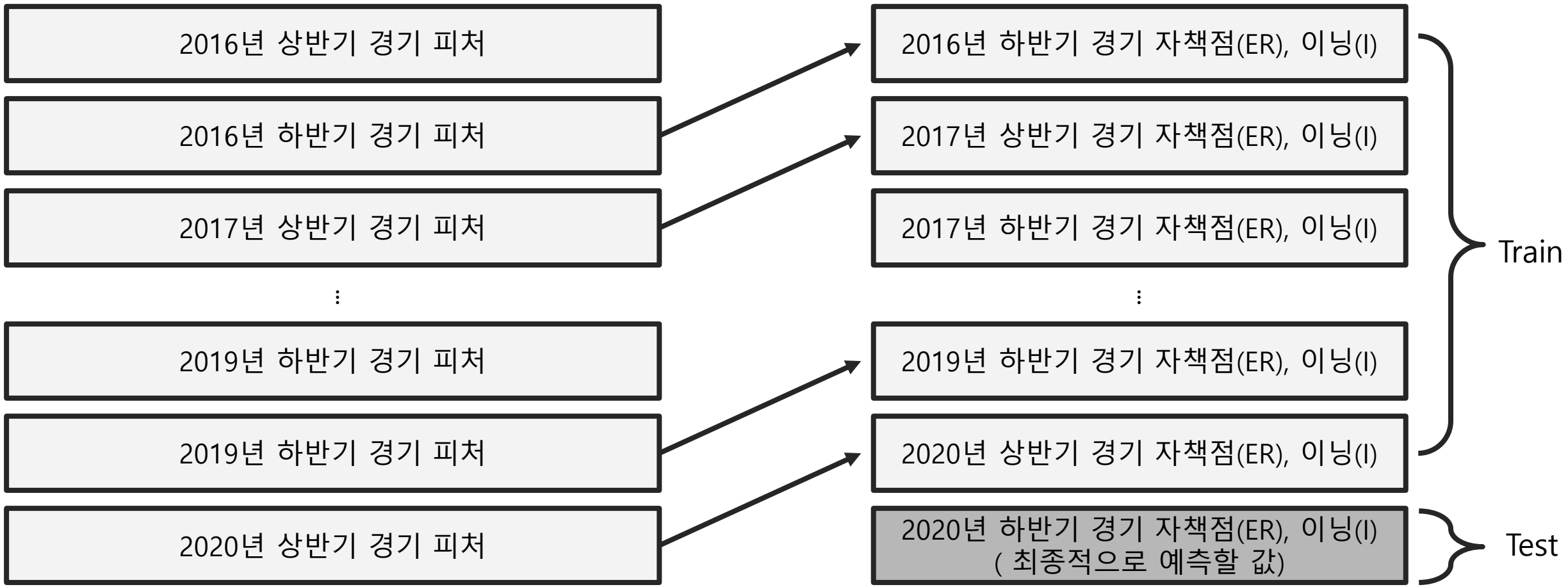
승률


2016년 상반기 경기 피처
2016년 하반기 경기 피처
2017년 상반기 경기 피처
2019년 하반기 경기 피처
2019년 하반기 경기 피처
2020년 상반기 경기 피처
2016년 하반기 경기 승률
2017년 상반기 경기 승률
2017년 하반기 경기 승률
2019년 하반기 경기 승률
2020년 상반기 경기 승률
2020년 하반기 경기 승률
(최종적으로 예측할 값)
Train
Test

타율


2016년 상반기 경기 피처
2016년 하반기 경기 피처
2017년 상반기 경기 피처
2019년 하반기 경기 피처
2019년 하반기 경기 피처
2020년 상반기 경기 피처
2016년 하반기 경기 타수(AB), 안타수(H)
2017년 상반기 경기 타수(AB), 안타수(H)
2017년 하반기 경기 타수(AB), 안타수(H)
2019년 하반기 경기 타수(AB), 안타수(H)
2020년 상반기 경기 타수(AB), 안타수(H)
2020년 하반기 경기 타수(AB), 안타수(H)
(최종적으로 예측할 값)
Train
Test

방어율


2016년 상반기 경기 피처
2016년 하반기 경기 피처
2017년 상반기 경기 피처
2019년 하반기 경기 피처
2019년 하반기 경기 피처
2020년 상반기 경기 피처
2016년 하반기 경기 자책점(ER), 이닝(I)
2017년 상반기 경기 자책점(ER), 이닝(I)
2017년 하반기 경기 자책점(ER), 이닝(I)
2019년 하반기 경기 자책점(ER), 이닝(I)
2020년 상반기 경기 자책점(ER), 이닝(I)
2020년 하반기 경기 자책점(ER), 이닝(I)
(최종적으로 예측할 값)
Train
Test

정규시즌 잔여 경기에 대한 각 팀별 승률 예측

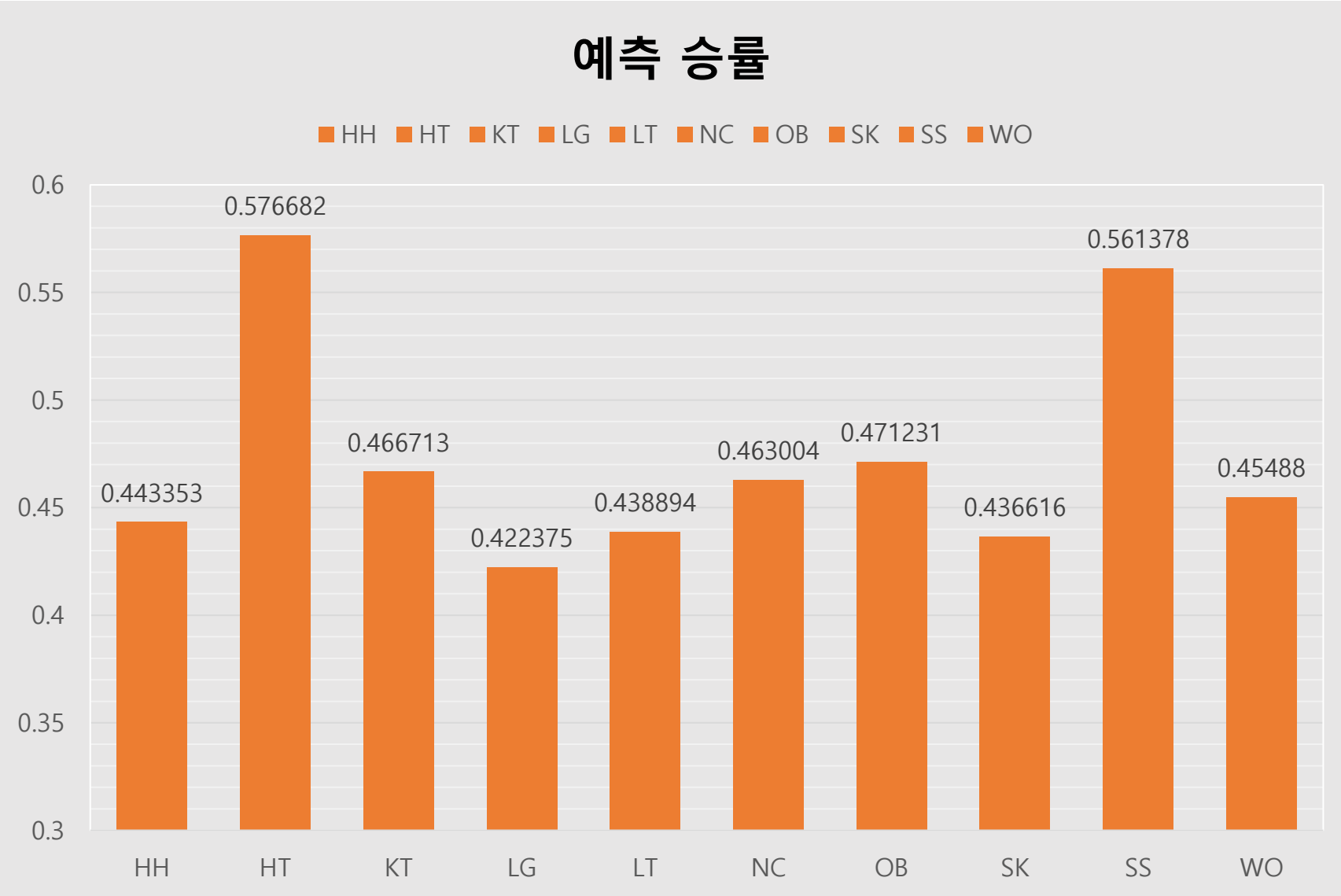
승률 반응 변수 생성

- 데이터셋의 '결과'변수를 이용하여 승률 계산

- 2016년 상반기 변수의 반응변수 → 2016년 하반기 경기 승률
- 2016년 하반기 변수의 반응변수 → 2017년 상반기 경기 승률
- 2017년 상반기 변수의 반응변수 → 2017년 하반기 경기 승률

⋮

- 2019년 상반기 변수의 반응변수 → 2019년 하반기 경기 승률
- 2019년 하반기 변수의 반응변수 → 2020년 상반기 경기 승률
- 2020년 상반기 변수의 반응변수 → NULL


예측 승률
HH
HT
KT
LG
LT
NC
OB
SK
SS
WO
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.443353
0.576682
0.466713
0.422375
0.438894
0.463004
0.471231
0.436616
0.561378
0.45488
HH
HT
KT
LG
LT
NC
OB
SK
SS
WO

정규시즌 잔여 경기에 대한 각 팀별 타율 예측

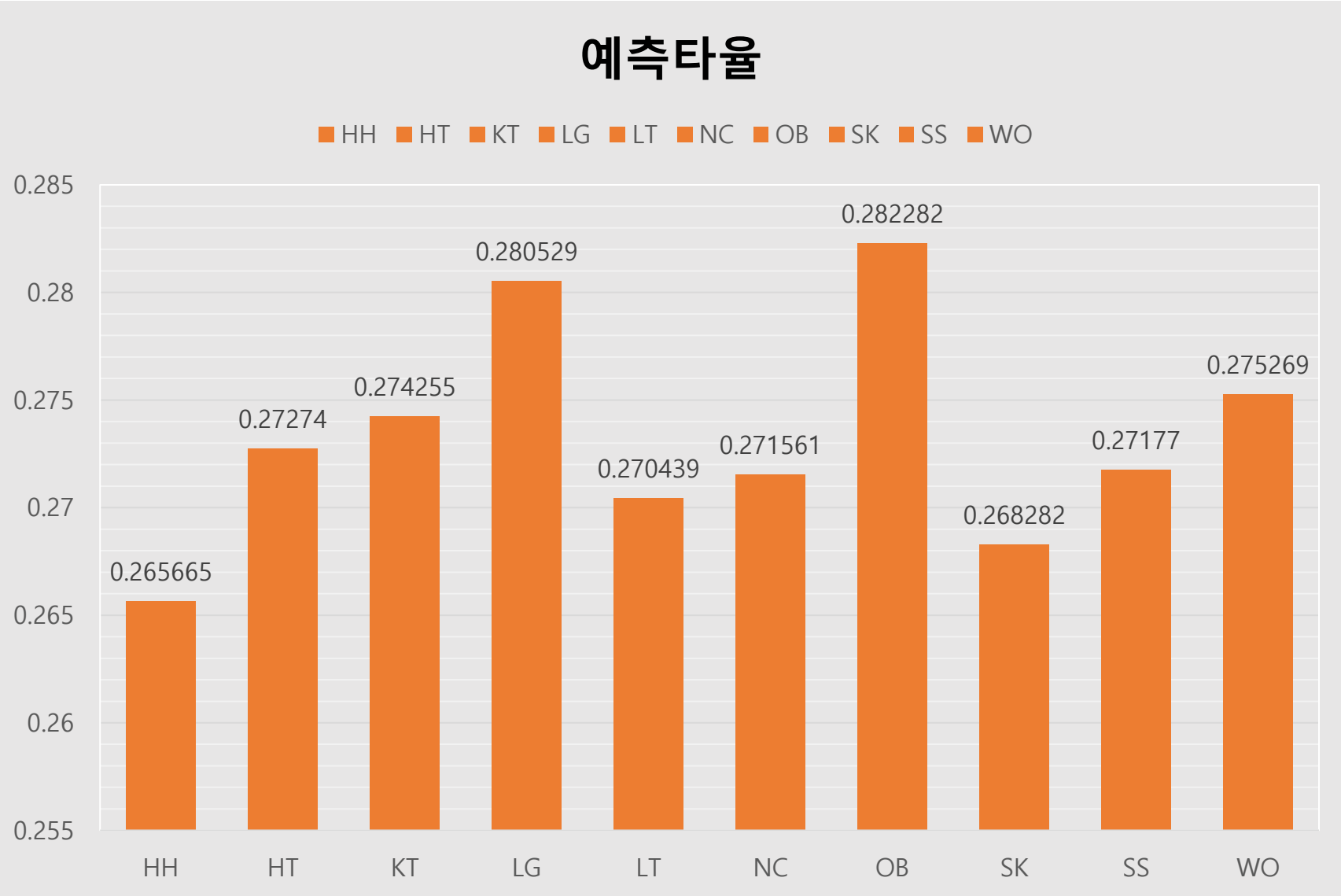
타율 반응 변수 생성

- 데이터셋의 '결과'변수를 이용하여 승률 계산

- 2016년 상반기 변수의 반응변수 → 2016년 하반기 경기 타수(AB), 안타수(H)
- 2016년 하반기 변수의 반응변수 → 2017년 상반기 경기 타수(AB), 안타수(H)
- 2017년 상반기 변수의 반응변수 → 2017년 하반기 경기 타수(AB), 안타수(H)

⋮

- 2019년 상반기 변수의 반응변수 → 2019년 하반기 경기 타수(AB), 안타수(H)
- 2019년 하반기 변수의 반응변수 → 2020년 상반기 경기 타수(AB), 안타수(H)
- 2020년 상반기 변수의 반응변수 → NULL


예측타율
HH
HT
KT
LG
LT
NC
OB
SK
SS
WO
0.285
0.28
0.275
0.27
0.265
0.26
0.255
0.265665
0.27274
0.274255
0.280529
0.270439
0.271561
0.282282
0.268282
0.27177
0.275269

정규시즌 잔여 경기에 대한 각 팀별 방어율 예측

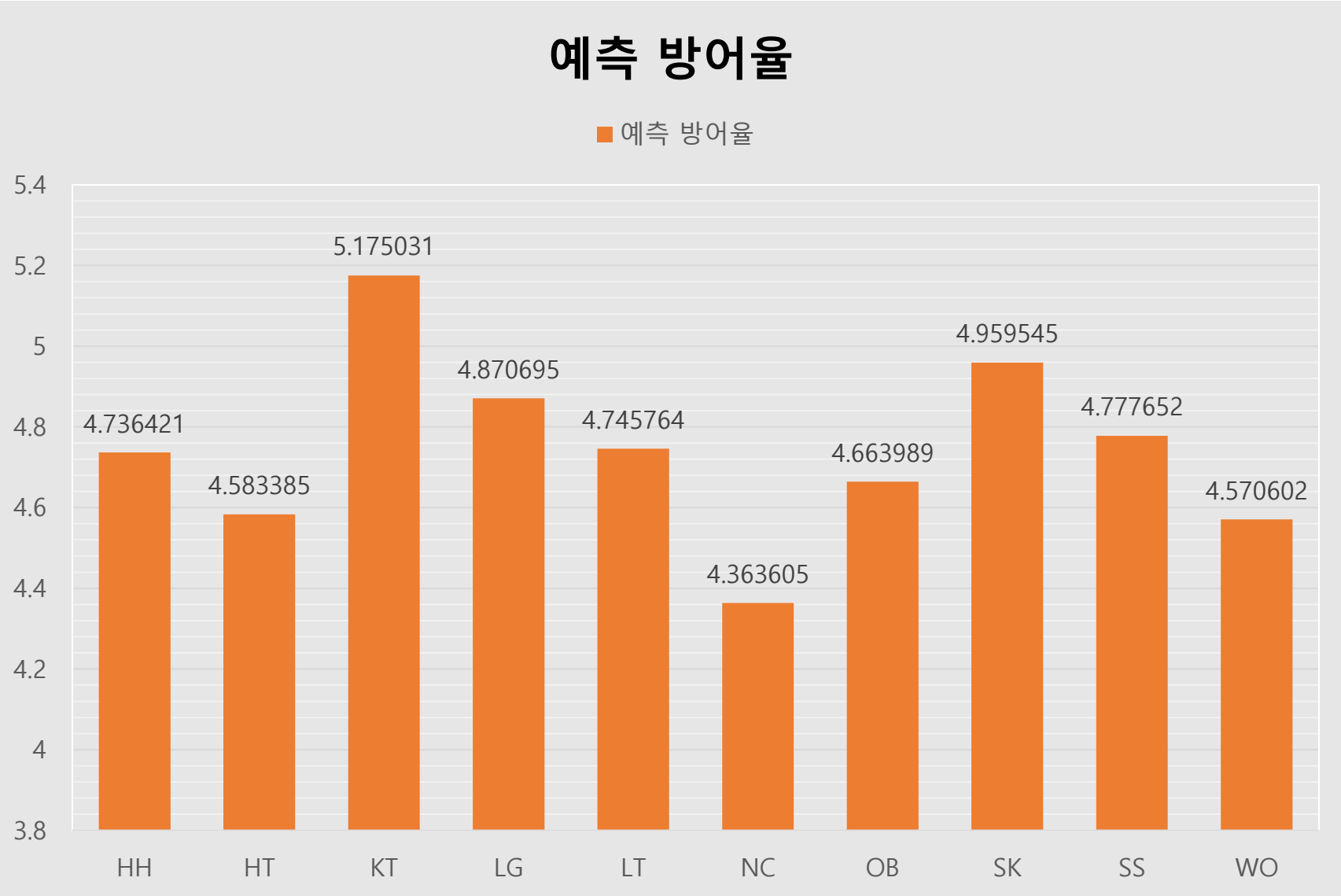
방어율 반응 변수 생성

- 데이터셋의 '결과'변수를 이용하여 승률 계산

- 2016년 상반기 변수의 반응변수 → 2016년 하반기 경기 자책점(ER), 이닝(I)
- 2016년 하반기 변수의 반응변수 → 2017년 상반기 경기 자책점(ER), 이닝(I)
- 2017년 상반기 변수의 반응변수 → 2017년 하반기 경기 자책점(ER), 이닝(I)

⋮

- 2019년 상반기 변수의 반응변수 → 2019년 하반기 경기 자책점(ER), 이닝(I)
- 2019년 하반기 변수의 반응변수 → 2020년 상반기 경기 자책점(ER), 이닝(I)
- 2020년 상반기 변수의 반응변수 → NULL


예측 방어율
예측 방어율
5.4
5.2
5
4.8
4.6
4.4
4.2
4
3.8
4.736421
4.583385
5.175031
4.870695
4.745764
4.363605
4.663989
4.959545
4.777652
4.570602
HH
HT
KT
LG
LT
NC
OB
SK
SS
WO


Thank you :)