KBO 승률/타율/평균자책점 예측

팀 통마늘

팀장: 신보현 (sbh0613@yonsei.ac.kr)
팀원: 류재원 (jwryu0812@gmail.com)

INDEX

Chapter 1

주제 및 데이터 이해

주제 이해

1. 대회 주제

9월 28일 이후 진행되는 KBO 정규시즌의 **잔여 경기**에 대한 각 팀의 승률, 타율, 평균자책점 예측

- 타율: 팀의 공격력을 나타내는 지표
- 평균자책점(방어율): 팀의 방어력을 나타내는 지표

주제 이해

2. KBO 리그

대한민국 프로 야구 경기. 매년 3월 말 ~ 4월 초에 개막하여 10월 초에 시즌이 종료되며,
시즌 종료 이후에 포스트시즌이 진행됨.

- 올해는 코로나19의 여파로 5월 5일 개막 / 10월 말 시즌 종료 예상

주제 이해

3. 세이버메트릭스

1971년 밥 데이비스가 창시한 SABR(The Society for American Baseball Research)라는 모임에서 만들어진, 야구를 통계학적/수학적으로 분석하는 방법론.

기존 주먹구구식 선수 평가론을 보완하고, 객관적인 평가를 위해 창안된 이론이다.

4. 피타고리안 승률

세이버메트릭스에서 고안된 **승률을 추정**하기 위한 공식. 야구가 득점을 많이 하고 실점을 적게 하면 이기는 스포츠라는 사실에서 시작되었다. 공식은 아래와 같다.

$$\mathrm{P} = \frac{W^n}{W^n + L^n}$$

W = 득점, L = 실점, n = 상수(KBO에서는 n = 1.86 사용)

바로 승률을 예측하기 보다, 득점과 실점을 바탕으로 계산한 피타고리안 승률을 사용하기로 결정.

Chapter 2

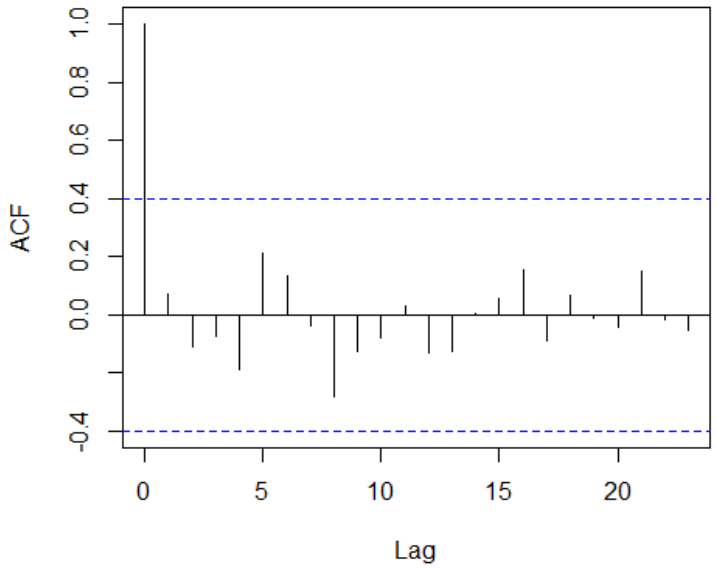
EDA

EDA

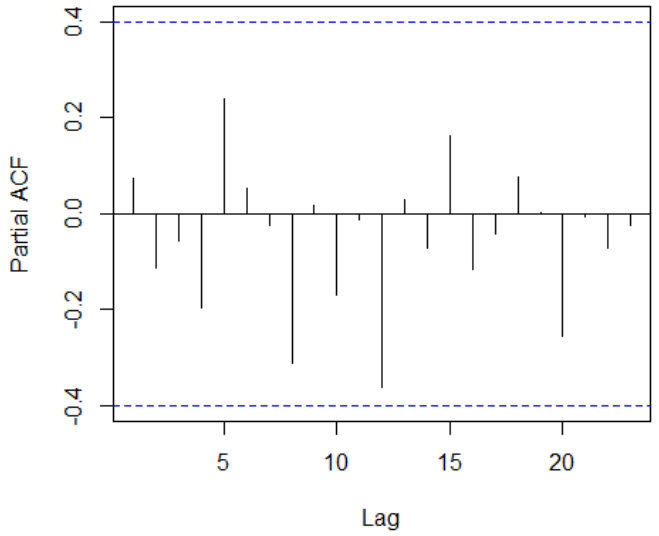
1. 시간 의존성

Chatfield's Analysis of Time Series (1980), The Advanced Theory of Statistics (1966)

NC 월간 타율의 ACF Plot



NC 월간 타율의 PACF Plot



- 표본으로부터 계산된 ACF와 PACF 95% 신뢰구간 안에 들어와 있다.
- ACF, PACF 그림으로부터, ARIMA 모형을 식별할 수 없었다.
- 모든 팀의 타율, 자책점, 승률에 대해서 해당 과정을 반복한 결과 한화 팀 제외, 동일한 결론을 얻었다.
- 분산 안정화를 통해 ARCH GARCH 모형이나 외생변수를 고려할 수 있는 VAR 모형을 시도해보았지만 아래의 이유로 최종 모형에서는 선택하지 않았다.
    - 고려해야할 가정 사항(Stationary, Granger Causality 등)이 많음
    - 이를 모든 팀(10개)의 승률, 자책점, 타율에 대해서 한다면 30개의 모형을 만들어야하므로 비효율적이라고 판단

데이터 이해

2. 시계열 자료

시간	201604	201605	201606	...	202009	202010
구단	HH	HH	HH		KT	KT
타율	0.267	0.273	0.281	...	0.267	0.283
자책점	3.4	2.7	2.85		4.67	5.76
승률	0.172	0.823	0.659	...	0.713	0.652

대부분의 기록이 시간에 따라 관측된 데이터이다.

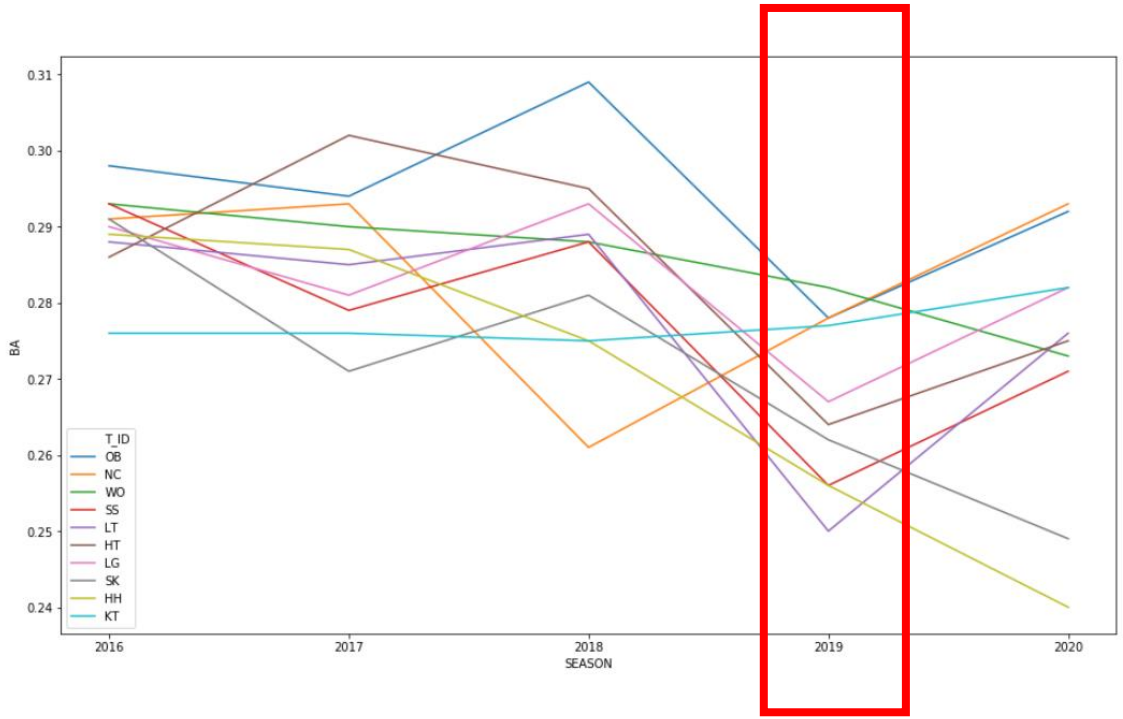
시즌 별로 순환되고, 시즌 내에서도 순차적으로 데이터가 관측된다.

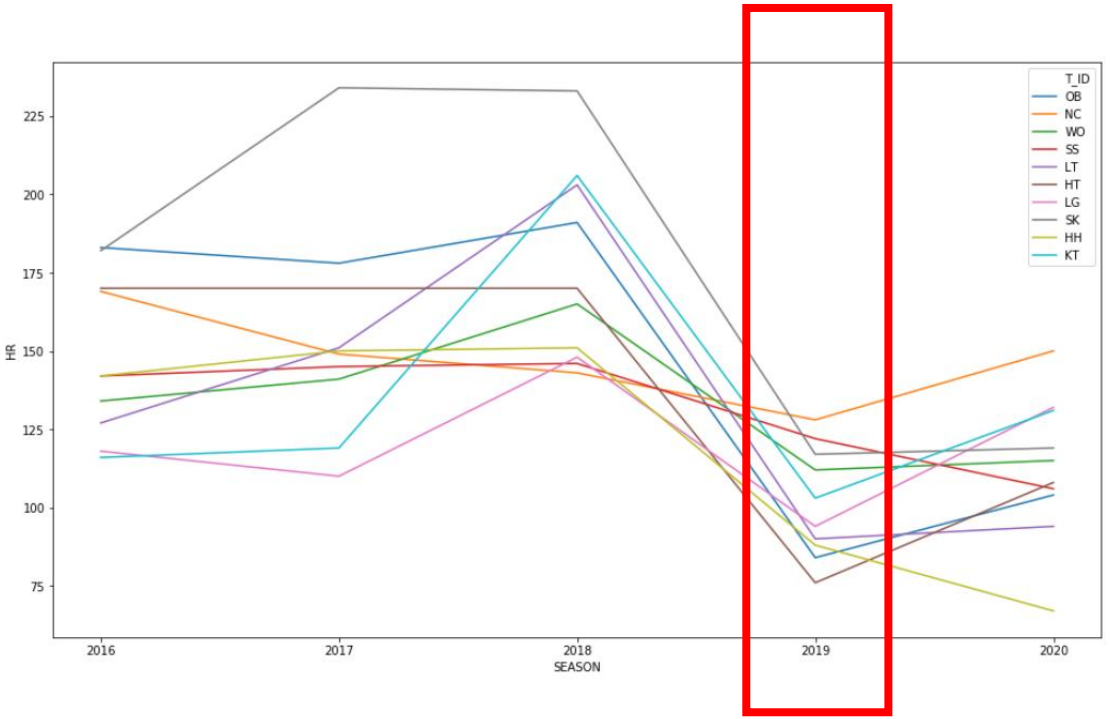
Seasonality와, 시즌 내에서 Time Dependency가 존재할 수 있다.

EDA

3. KBO 공인구 변경

연도별 각 팀의 타율/홈런 변화





관련기사

연합뉴스

끝모를 타고투저에 제동건다...KBO 공인구 반발계수 하향

기사입력 2018.12.21. 오전 11:54 | 최종수정 2018.12.21. 오전 11:56 기사원문

주간경향

KBO 공인구 달라져 홈런이 줄었다?

기사입력 2019.04.29. 오전 11:04 기사원문 스크랩 본문듣기 · 설정

2018시즌 이후 공인구가 변경되어, 기존의 타고투저에서 투고타저로 흐름이 변함 → 공인구 관련 더미변수 추가

EDA 결론

시계열 모형을 고려하지 않는다.

승률 계산시, 피타고리안 승률을 고려한다.

KBO 공인구가 바뀐 시점을 고려한다.

Chapter 3

Feature Engineering

Feature Engineering

0. 문제 제기

01

데이터 수 부족

시즌 단위로 학습 시 총 데이터
단 40개(4시즌 10개 팀) 뿐

월 단위로 split 하더라도
여전히 데이터 부족
(4시즌 x 10개 팀 x 6개월 = 약 240개)

데이터 간 중요도가 일정하지 않다
개막 이후 3개월치 데이터가
1개월치 데이터보다 더 설명력 높음

▷ 월 단위로 split
(잔여경기 특성 고려)

▷ 계산 기간을 다르게 하면
많은 데이터 활용 가능

▷ 데이터 별 가중치 설정
(ex) 90경기 3, 30경기 1

02

클래식 스탯 위주의 구성

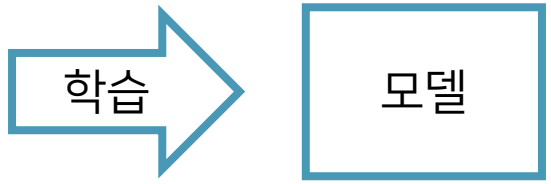
▷ 야구 데이터 분석 웹사이트 '스탯티즈'의
세이버메트릭스 지표 크롤링

Feature Engineering

1. 데이터 부족 문제 해결

- 앞서, 시간 의존성에 대한 EDA를 토대로 변수 계산 시에 바로 이전 시점이 아니라 시즌 초부터의 경기를 고려하였다.
- 이때, 한 시즌 내에서 시점 별로 진행된 경기가 다르기 때문에, 추후 모델링 진행을 진행할 때 **데이터 별로 다른 가중치**를 주었다.

팀	시즌	타율	홈런	…	반응변수	가중치
HH	2016	개막 이후 1개월 계산값 (4월)	개막 이후 1개월 계산값 (4월)	개막 이후 1개월 계산값 (4월)	201605 한화 타율	1
HH	2016	개막 이후 2개월 계산값 (4월~5월)	개막 이후 2개월 계산값 (4월~5월)	개막 이후 2개월 계산값 (4월~5월)	201606 한화 타율	2
HH	2016	개막 이후 3개월 계산값 (4월~6월)	개막 이후 3개월 계산값 (4월~6월)	개막 이후 3개월 계산값 (4월~6월)	201607 한화 타율	3
HH	2016	개막 이후 4개월 계산값 (4월~7월)	개막 이후 4개월 계산값 (4월~7월)	개막 이후 4개월 계산값 (4월~7월)	201608 한화 타율	4
…	…	…	…	…	…	…
NC	2019	개막 이후 5개월 계산값 (4월~8월)	개막 이후 5개월 계산값 (4월~8월)	개막 이후 5개월 계산값 (4월~8월)	201909 NC 타율	5



Feature Engineering

2. 타자 관련 추가 변수 생성

타격 관련 변수

타율 BA	안타 수 / 타수 HIT / AB
장타율 SLG	루타 / 타수 (H1 + 2*H2 + 3*H3 + 4*HR) / AB
순수 장타율 IsoP	장타율 – 타율 SLG - BA
출루율 OBP	타자가 출루에 성공한 비율 (HIT+BB+HP)/(AB+BB+HP+SF)
On Base Plus Slusgging OPS	출루율 + 장타율 SLG + OBP
Gross Product Average GPA	OPS 보완 지표 (1.8*OBP + SLG) / 4
Contact Quality CQ	인플레이 타구에 대한 퀄리티 (0.5HIT+0.3TB) / (AB-KK)

선구안 관련 변수

볼삼비 BBK	삼진 대비 볼넷 비율 BB/KK
볼삼비 점수 BBKScore	볼넷, 삼진이 득점에 기여하는 정도 0.4BB – 0.3KK
삼진비 K_PERCENT	타수 대비 삼진 비율 KK / (AB – IB)
볼넷비 BB_PERCENT	타수 대비 볼넷 비율 (BB – IB) / (AB – IB)

Feature Engineering

2. 타자 관련 추가 변수 생성

득점 관련 변수

Runs Created RC	타자의 득점 기여도 (위키피디아 참조)
Runs Created 27 RC27	출전 이닝을 보정한 타자의 득점 기여도 (27RC / (AB – HIT + CS + GD + SH + SF)

기타 변수

Batted Average on Balls In Play BABIP	인플레이 타구가 안타가 된 비율 (HIT – HR) / (AB – KK –HR + SF)
BABIP_minus_BA	타율과 BABIP의 차이 BABIP - BA
Adjusted Weighted On Base Average Adj_wOBA	타자의 생산력을 측정하는 지표로 선형회귀분석을 통해 생성됨. WAR을 계산하기 위해 사용되는 지표이다. (위키피디아 참조)

Feature Engineering

3. 투수 관련 추가 변수 생성

투수 관련 변수

평균자책점 ERA	투수가 9이닝당 허용한 자책점 ER * 9 / INN
피안타율 OAVG	투수가 상대한 타석수에 비해 안타를 허용할 확률 HIT / AB
피장타율 OSLG	투수가 허용할 누타의 기댓값 (H1 + 2*H2 + 3*H3 + 4*HR) / AB
피출루율 OOBP	투수가 타자에게 출루를 허용할 확률 (HIT+BB+HP)/(AB+BB+HP+SF)
피OPS OOPS	피안타율과 피장타율의 합 SLG + OBP

Feature Engineering

3. 투수 관련 추가 변수 생성

투수 관련 변수

볼삼비 KBB	볼넷과 삼진의 비 BB/KK
OBABIP	인플레이 타구가 안타가 된 비율 (HIT – HR) / (AB – KK –HR + SF)
Walks Plus Hits Divided by Innings Pitched WHIP	이닝 당 출루 허용률
Fielding Independent Pitching FIP	조정방어율 ERA의 단점을 보완하기 위해 만들어진 지표
Win Probability Added WPA	경기당 투수의 승리 기여도

Master Table

T_ID	monthkey	타자		투수		weight	반응변수
		클래식 지표	추가 비율 지표	클래식 지표	추가 비율 지표		
HH	201604	HIT/경기	OPS	HIT/경기	OOPS	1~5	타율 평균자책점 득/실점
…	…	H2/경기	SLG	K/경기	WHIP		
NC	202008	…	…	…	…		

Chapter 4

Modeling

Modeling: 모형 선정

PLS, Lasso

- 차원 축소 및 변수 선택을 통해 상관성이 있는 변수를 줄임
- 관측치마다 가중치를 다르게 줄 수 없음
- 각 변수마다 계수를 추정하지만,
  이를 통해 변수의 중요도를 판단할 수는 없음

Tree 계열 모형(Ensemble, Boosting)

- node를 split할 때마다 impurity을 낮추는 변수를 선택함으로써 상관성 있는 변수를 다룸
- 관측치마다 가중치를 다르게 줄 수 있음
- 변수 중요도를 제공함
- **최종 선택 모형**

Modeling: 파라미터 튜닝

GridSearch Optimization

- 파라미터 공간을 Discrete하게 탐색
- 파라미터 공간을 늘리면 탐색 시간이 매우 오래 걸릴 수 있음
- 파라미터 공간을 Discrete하게 제한해야 하기 때문에 Optimal Point을 포함하지 못할 수도 있음

Bayesian Optimization (hyperopt)

- 파라미터 공간을 Continuous하게 탐색
- Continous하게 탐색하기 때문에 파라미터 공간을 늘리더라도 탐색 시간이 선형적으로 증가하지 않음
- GridSearch에 비해서 파라미터 공간의 제약이 상대적으로 적음
- **최종 선택 방법**

Algorithms for Hyper-Parameter Optimization, James Bergstra et al.,

Modeling: 튜닝한 파라미터 리스트

RandomForest / AdaBoost

- 과적합 방지 파라미터
  - max_depth
  - min_samples_split
  - min_samples_leaf
- 기타 파라미터
  - n_estimators
  - max_features

XGB / LightGBM / GBM

- 과적합 방지 파라미터
  - max_depth
  - min_child_weight
  - reg_alpha
  - reg_lambda
- 기타 파라미터
  - n_estimators
  - learning_rate
  - colsample_bytree

- 특히, 과적합이 우려되어서 관련 파라미터를 세심하게 튜닝하였다.
- Boosting 모형은 훈련시에 early_stopping을 통해서 과적합을 피하고자 하였다.

Modeling: 팀 타율 예측 결과

주요 선택 변수

- Feature Engineering을 통해 만들어진 변수
- 연도 구분 더미 변수 (공인구 구분)
- 경기 수로 보정 및 가공하지 않은 클래식 지표
  (예: 안타 / 볼넷 개수 등)

모델 학습 결과

모델명	RMSE
RF	0.016
Ada Boosting	0.0182
Gradient Boosting	0.0184
LGBoosting	0.0180
XGBoosting	0.0110

Modeling: 팀 평균자책점 예측 결과

주요 선택 변수

- Feature Engineering을 통해 만들어진 변수
- 연도 구분 더미 변수 (공인구 구분)
- 경기 수로 보정 및 가공하지 않은 클래식 지표
  (예: 안타 / 볼넷 개수 등)

모델 학습 결과

모델명	RMSE
RF	0.918
Ada Boosting	0.926
Gradient Boosting	0.938
LGBoosting	0.949
XGBoosting	0.989

Modeling: 팀 승률 예측 결과

모델 학습 결과

모델명	득점 RMSE	실점 RMSE	승률 RMSE
RF	0.784	0.878	0.098
Ada Boosting	0.783	0.89	0.104
Gradient Boosting	0.78	0.82	0.104
LGBoosting	0.761	0.84	0.098
XGBoosting	0.8	0.899	0.107

- Feature Engineering을 통해 만들어진 변수
- 연도 구분 더미 변수 (공인구 구분)
- 경기 수로 보정 및 가공하지 않은 클래식 지표
  (예: 안타 / 볼넷 개수 등)
- 남은 기간의 득점, 실점을 예측하고 이를 피타고리안 승률로 엮어서 최종 승률을 계산한다.

Chapter 5

Interpreting Models

변수 중요도

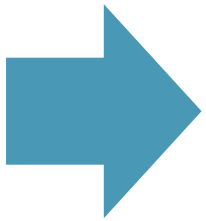
- RegressorTree 계열 모형은 Mean Decrease Impurity을 통해 변수 중요도를 제공한다.
- MSE을 통해서 Impurity을 계산했기 때문에, MSE 감소를 통해서 변수 중요도를 계산하였다.
- 여러 tree 모형을 사용한만큼, 각 모형이 변수를 바라보는 관점을 종합한다.
- 각 모형의 변수 중요도 순위를 합산하여 최종 변수 중요도를 산출한다.

Interpreting Modeling

최종 Global 변수 중요도 산출 과정

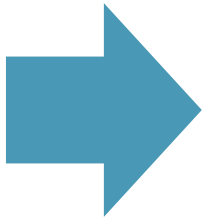
- 각 모형에서 변수 중요도를 내림차순
- 각 변수 별로 순서를 부여

Ex) A변수는 RandomForest 에서 1순위



- 각 모형에 대해서 반복해서 시행

Ex) A변수는 XGBoost, AdaBoost에서 각각 2, 3순위



- 각 모형에서 순위를 합산

Ex) A, B 변수의 6개 모델에서 산출한 순위를 합한 결과 각 3, 10.

-> A 변수가 전체적으로 더 중요함을 알 수 있음

Interpreting Modeling: 타율 모델링 변수 중요도 결과

변수명	Random Forest	AdaBoost	GradientBoost	LightGBM	XGBoost	순위 합
공인구 여부	1	0	1	1	0	3
WOBA	3	3	3	0	3	12
OBP	4	4	0	2	7	17
삼진 비율	10	1	9	4	15	39
장타율	9	9	14	10	4	46
볼삼비	22	8	7	5	6	48
안타 개수	6	15	2	7	18	48
삼진 개수	13	2	18	15	2	50
볼넷 개수	21	13	4	3	9	51

- 타자의 경우 분석 결과 '일단 출루를 해야 득점을 한다'는 세이버매트릭스의 가장 유명한 명언을 떠올리게 한다.
- 출루율에 관한 많은 지표(wOBA, OBP 등)가 안타보다 중요한 가치를 보이는 것에서 확인할 수 있다.
- 반면 요즘 인기있는 지표인 RC(득점 생산력)의 경우 생각보다 중요한 것은 아니라는 것을 알 수 있었다

※ 안타, 삼진, 볼넷 개수는 경기 수를 보정하였다.

Interpreting Modeling: 평균자책점 모델링 변수 중요도 결과

변수명	Random Forest	AdaBoost	GradientBoost	LightGBM	XGBoost	순위 합
홈런 개수	0	0	0	1	2	3
공인구 여부	2	1	2	8	0	13
피장타율	1	2	7	3	1	14
FIP	3	3	1	11	3	21
WPA	5	6	8	0	5	24
3루타 개수	9	5	3	5	8	30
삼진 개수	11	4	9	2	9	35
피OPS	4	11	15	6	4	40
볼넷 개수	12	7	13	4	11	47

- 투수의 퍼포먼스를 향상시키는 가장 큰 키워드는 '장타억제'로 볼 수 있다. 경기당 피홈런, 피장타율 등이 평균자책점에 큰 영향을 주는 것으로 분석되기 때문이다.
- 보통 야구에서는 투수에게 타자와의 승부에서 도망가지 말고 정면승부할 것을 최고의 덕목으로 여기지만, 분석 결과 볼넷이나 삼진 등이 큰 영향을 미치지 않는 것으로 보아 강타자에게는 볼넷을 내주더라도 유인구 위주의 까다로운 승부를 해야 한다는 것을 알 수 있다.

※ 안타, 삼진, 볼넷 개수는 경기 수를 보정하였다.

Interpreting Modeling: 득점 모델링 변수 중요도 결과

변수명	Random Forest	AdaBoost	GradientBoost	LightGBM	XGBoost	순위 합
CQ	2	2	6	5	1	16
공인구 여부	11	0	4	0	2	17
장타율	0	1	13	3	3	20
RC27	5	11	0	14	0	30
홈런 개수	7	9	3	2	11	32
WOBA	8	5	8	10	6	37
BA	6	8	2	4	17	37
OBP	12	3	10	9	18	52
삼진 비율	14	4	22	11	5	56

- 인플레이 타구의 타율과 관련한 지표의 중요성이 높다(CQ 등)
- 최근 인플레이 타구가 안타로 만드는 능력을 더 이상 운으로 여기지 않고, 강한 타구를 만들어 안타 확률을 높이고자 하는 타격 트렌드와 일치
- 한 번의 타격으로 더 많은 베이스를 진루하는 것에 더 높은 가치를 두는 경향과도 일치(장타율, 홈런 등)

※ 안타, 삼진, 볼넷 개수는 경기 수를 보정하였다.

Interpreting Modeling: 실점 모델링 변수 중요도 결과

변수명	Random Forest	AdaBoost	GradientBoost	LightGBM	XGBoost	순위 합
피홈런 개수	0	1	0	5	1	7
공인구 여부	4	2	1	2	0	9
FIP	1	0	7	0	3	11
피장타율	2	3	5	1	2	13
WPA	5	9	6	4	9	33
ERA	6	4	10	6	8	34
피OPS	3	14	4	12	5	38
HP 개수	11	13	2	3	14	43
3루타 개수	10	10	8	9	6	43

- 인플레이 타구의 타율과 관련한 지표의 중요성이 높다(FIP, BABIP 등)
- 최근 인플레이 타구가 안타로 만드는 능력을 더 이상 운으로 여기지 않고, 강한 타구를 만들어 안타 확률을 높이고자 하는 타격 트렌드와 일치
- 한 번의 타격으로 더 많은 베이스를 진루하는 것에 더 높은 가치를 두는 경향과도 일치(피장타, 피3루타 등)

※ 안타, 삼진, 볼넷 개수는 경기 수를 보정하였다.

Appendix

실제 경기 결과

해당 부분은 대회 종료 후 추가하였습니다.

Final Result

예측 결과

팀	타율	방어율	승률
한화	0.241	5.20	0.277
KIA	0.279	4.84	0.560
KT	0.278	4.61	0.580
LG	0.276	5.11	0.556
롯데	0.271	4.96	0.576
NC	0.287	4.58	0.499
두산	0.272	4.38	0.574
SK	0.260	5.93	0.357
삼성	0.271	5.04	0.402
키움	0.286	4.58	0.608

실제 경기 결과

팀	타율	방어율	승률
한화	0.263	5.64	0.400
KIA	0.273	6.28	0.414
KT	0.290	4.32	0.556
LG	0.251	3.75	0.583
롯데	0.274	4.56	0.464
NC	0.283	4.62	0.480
두산	0.297	3.46	0.680
SK	0.256	5.15	0.458
삼성	0.252	3.86	0.522
키움	0.246	4.19	0.450

해당 부분은 대회 종료 후 추가하였습니다.

Final Result

RMSE

팀	타율	방어율	승률
한화	0.000455867	0.198941685	0.015035480
KIA	0.000037965	2.071619844	0.021521729
KT	0.000145024	0.088725962	0.000609229
LG	0.000612096	1.860815583	0.000770468
롯데	0.000003528	0.161091726	0.012481274
NC	0.000014614	0.001756014	0.000378417
두산	0.000650424	0.853749460	0.011241042
SK	0.000020334	0.593021277	0.010326663
삼성	0.000354651	1.403623383	0.014323188
키움	0.001599339	0.156562364	0.025073686

RMSE 종합

타율	방어율	승률
0.019733	0.859646	0.1057172

- KIA의 경우 외국인 1선발이 예기치 못한 일(가정사)로 시즌을 일찍 마무리하였고, 투수진의 연쇄 붕괴로 이어져 예상을 크게 벗어났다.
- 키움의 경우 중심타선이 부상에서 회복되는 속도가 늦었고, 구단 내 정치 싸움으로 인해 예상보다 크게 낮은 타율/승률을 기록하게 되었다.
- 1, 2위를 기록한 KT와 NC의 경우 안정적인 전력과 뎁스를 바탕으로 예상 기록에 가장 가까운 결과를 기록했다.

해당 부분은 대회 종료 후 추가하였습니다.

감사합니다