세이버메트릭스에 따른 KBO 연봉분석을 통해 FA등급 개선 제안

김도연, 김태현, 심기열, 이관형

목차

배경 설명

FA란?

Free Agent의 약어로 KBO에서는 신인 계약기간 8년이 지나면 자유롭게 다른 팀과 협상할 수 있는 자격을 취득 가능
또한 FA계약 이후 4년 뒤 재취득 가능
FA 자격으로 다른 팀으로 이적할 경우 규정에 따라 원 소속팀에 보상이 발생

FA 등급이란?

FA자격을 취득한 선수가 다른 팀으로 이적하게 될 경우, 원래 연봉을 기준으로 A/B/C 세 등급을 분류하여 보상의 정도를 책정한다.

FA 등급제 전략

선수입장에서 FA계약은 연봉외에도 계약금을 통해 큰 금액을 한번에 얻을 수 있게 되므로 FA 자격을 취득했을때 큰 이득을 얻을 수 있도록 자격 취득을 연기하거나 직전연봉을 낮춰서 계약하는 전략을 보인다.
이에 구단은 해당 선수를 FA계약 전 트레이드 하거나 아직 FA자격을 취득하지 않은 선수와 미리 다년계약을 하는 형식으로 계약하게 된다.

개선 방향

현재 상황 (As-is)

- 기존 연구[1]에 따르면 현재 FA등급제는 **경쟁균형**에 악영향을 주기 때문에 개선의 필요성의 논의되어야 한다고 주장한다.
- 선수협은 FA등급제가 B와 C등급의 선수들이 쉽게 팀을 찾을 수 있는 **개선안이라고 보기 어렵다**고 주장한다.[2]
- FA자격 취득시 연봉을 기준으로 등급이 산정되므로 직전 계약시 일부러 연봉을 낮게 계약하는 선수들도 존재한다.

조건

- 경쟁균형에 도움되어야 함
- 구단과 선수 입장이 모두 개선되어야함

개선 방향 (To-be)

- 보상규정이 더 완화된 **D등급을 신설**하여 자유계약선수처럼 이적할 수 있도록 함
- 연봉기준이 아니라 **퍼포먼스 기준**으로 등급을 산정하여 등급 기준을 더 명확히 함
- 핵심선수는 지키고, 스쿼드 멤버는 자유롭게 이적할 수 있는 **시장을 활성화**하여 FA미아가 나오지 않도록 유도함

1. 이인엽, 한진욱, 「프로야구 경쟁균형을 위한 FA제도의 문제점 및 개선 방안 연구」 2023, vol.28, no.3, pp. 95-107
2. 배영은, "KBO·선수협, 오랜 줄다리기 끝 'FA 제도' 변화 첫걸음 막후", 일요신문, 2019.12.06

자료 수집

1. 경기 기록

자료 출처: statiz.com, KBO

KBO 10개 구단의 기록이 있는 전체 선수들의 데이터를 수집

예: OPS, RBI, ERA, WHIP, F%, WAR등 세이버메트릭스 지표와

출전 경기수, 득점, 타점, 세이브, 홀드, 보살 등 기록 지표 수집

2. 연봉 데이터

자료 출처: 구단 공식 홈페이지, 언론사 뉴스, KBO공식 홈페이지

계약금 총액을 계약 연수로 나누어 연평균 환산, 연봉과 합산

대상 선수: 신인 계약 선수 (드래프트 지명자), FA 계약 체결 선수, 외국인 용병 선수

예: 계약금 40억(4년) + 연봉 10억/년 → 연평균 20억으로 반영

그리고 전체 타자, 투수 분리해서 비교 분석

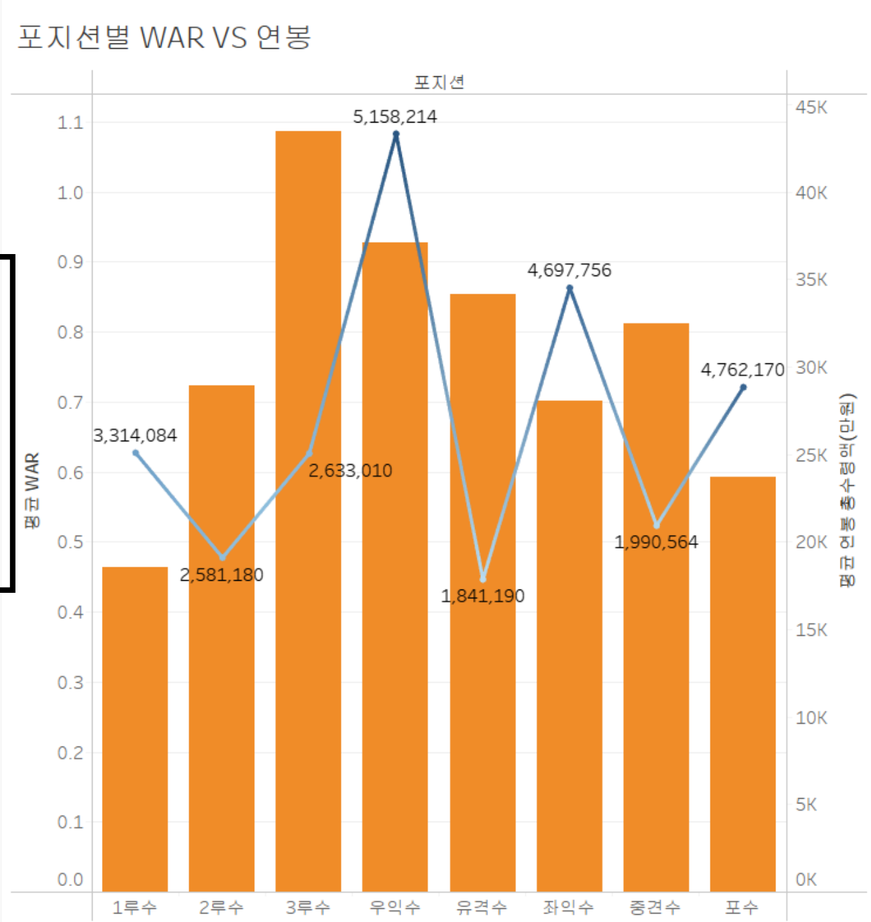
3. 기간 분석 범위: 2021 ~ 2024 시즌 (최근 4년)

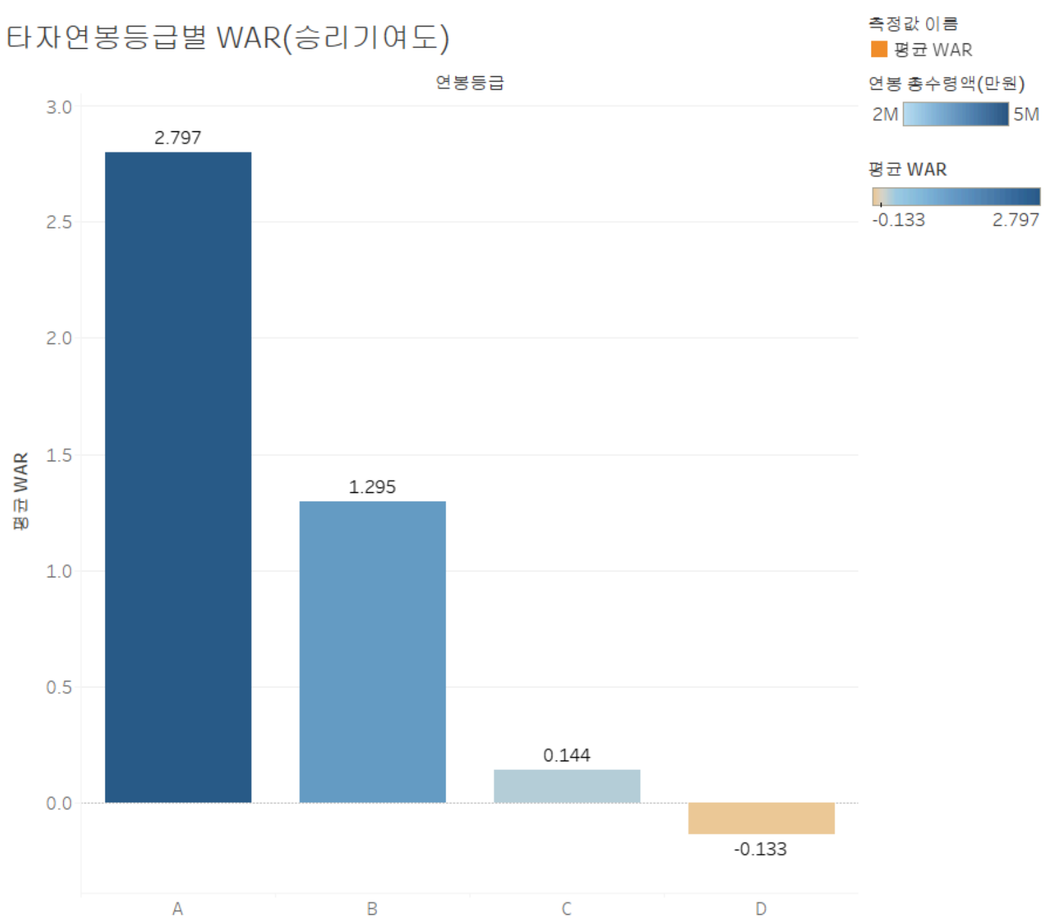
조사 불가능 기록: 비공개 계약 세부 내역, 부상 경과, 내부 스카우팅 리포트, 훈련 데이터 등 비공개 정보

데이터셋 살펴보기

포지션 별 지표

- 개별 포지션에서는 WAR과 연봉 간 직접적인 연결고리가 약해 보이지만, 전체 선수 집단으로 확장하면 **WAR이 연봉과 상관관계가 있음**을 확인

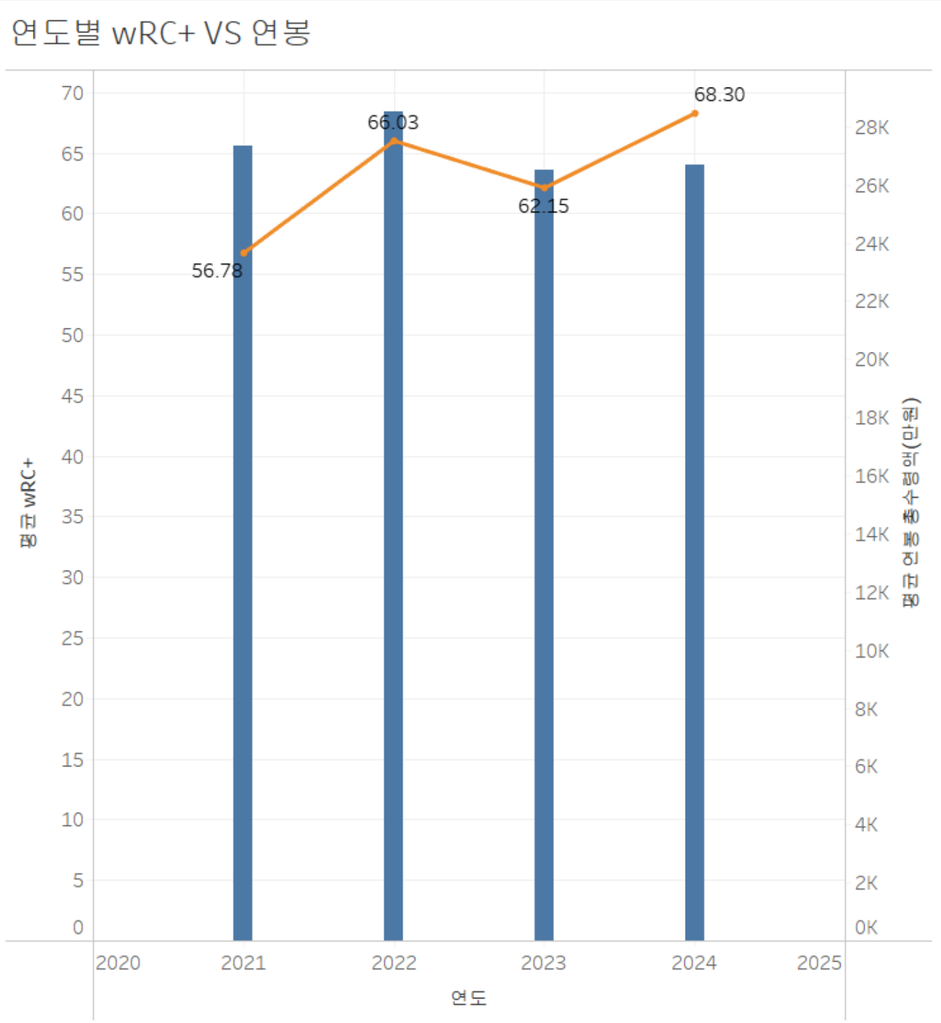


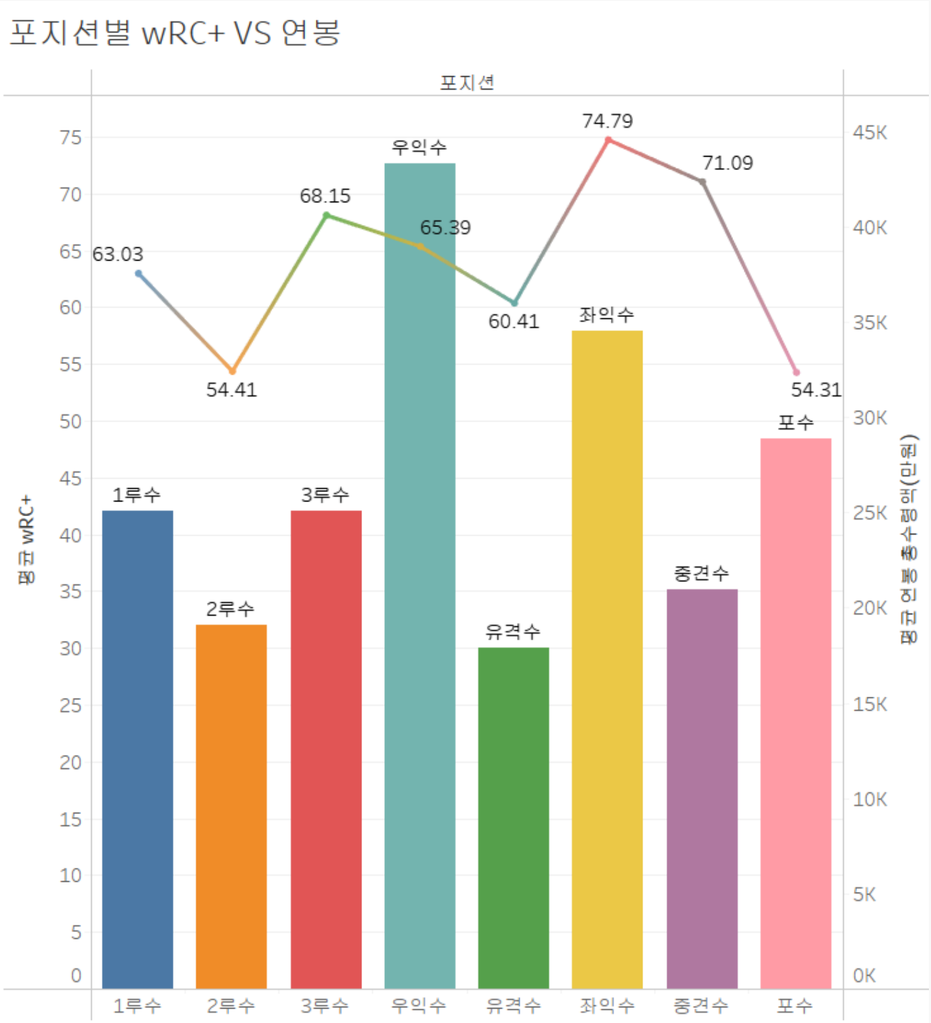


데이터셋 살펴보기

타자 타격 지표

- wRC+는 공격 기여도의 포지션 차이를 드러냄
  → 외야수 공격 기여도가 높게 나타남
- 연봉은 공격 지표만으로 설명되지 않음
  → 수비 기여, 포지션 희소성, 시장 수요가 함께 반영
- 포지션별 연봉 예측에는 **공격+수비+희소성 종합 고려** 필요

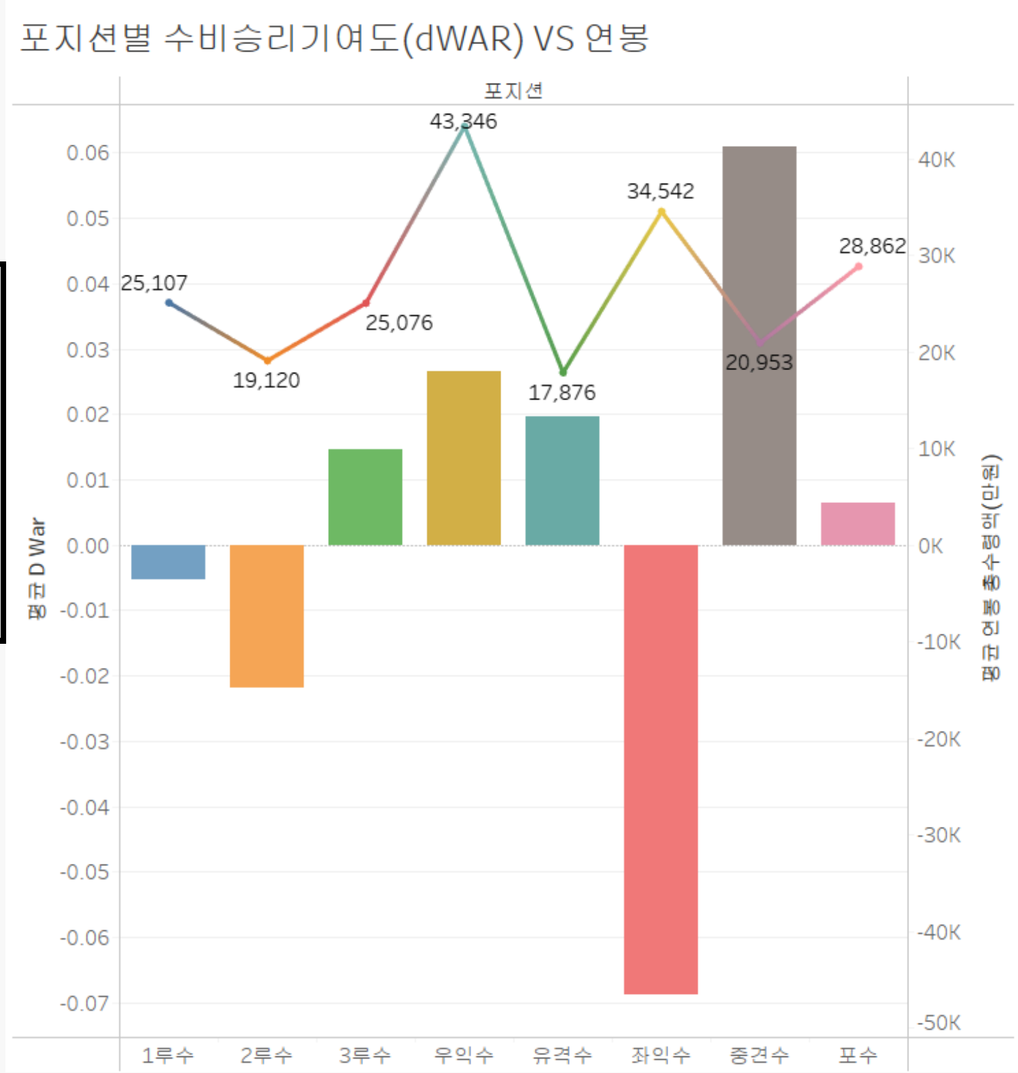


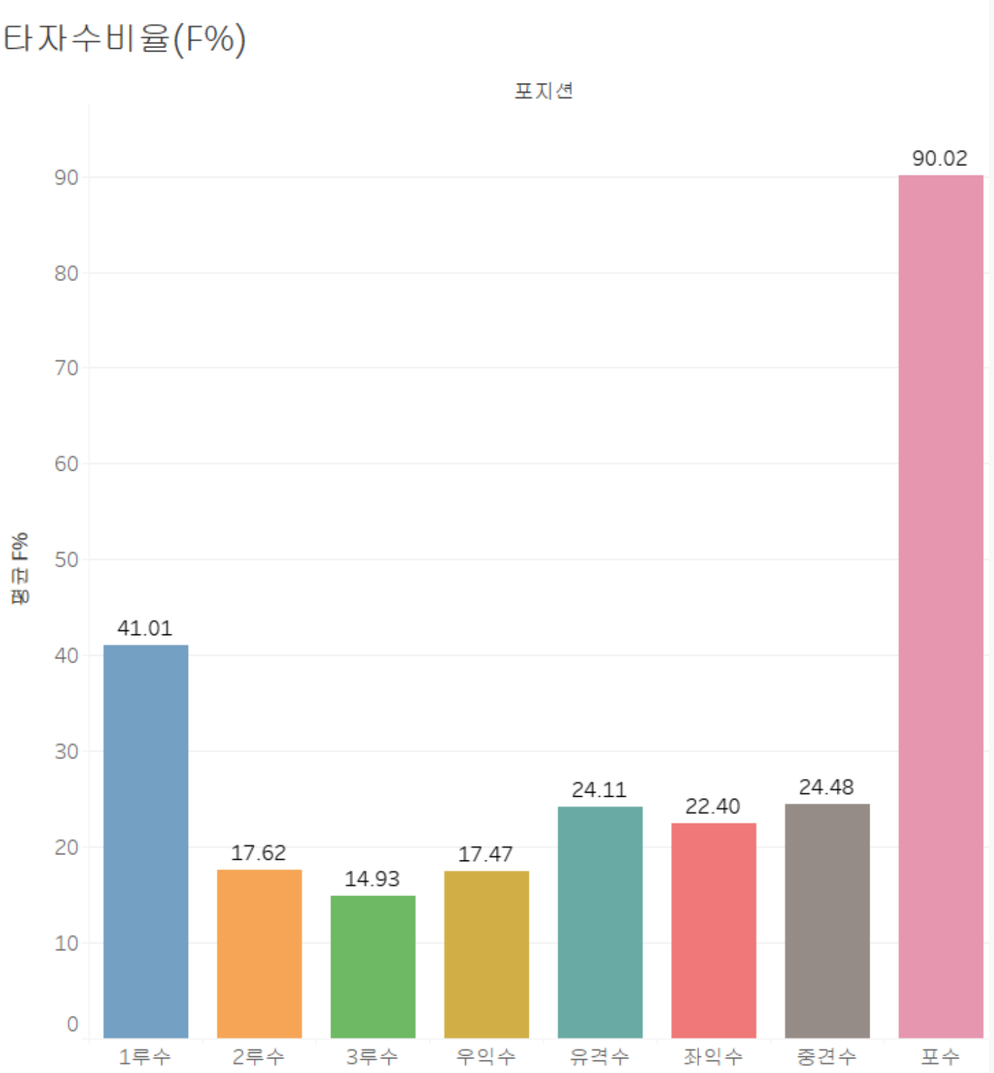


데이터셋 살펴보기

타자 수비 지표

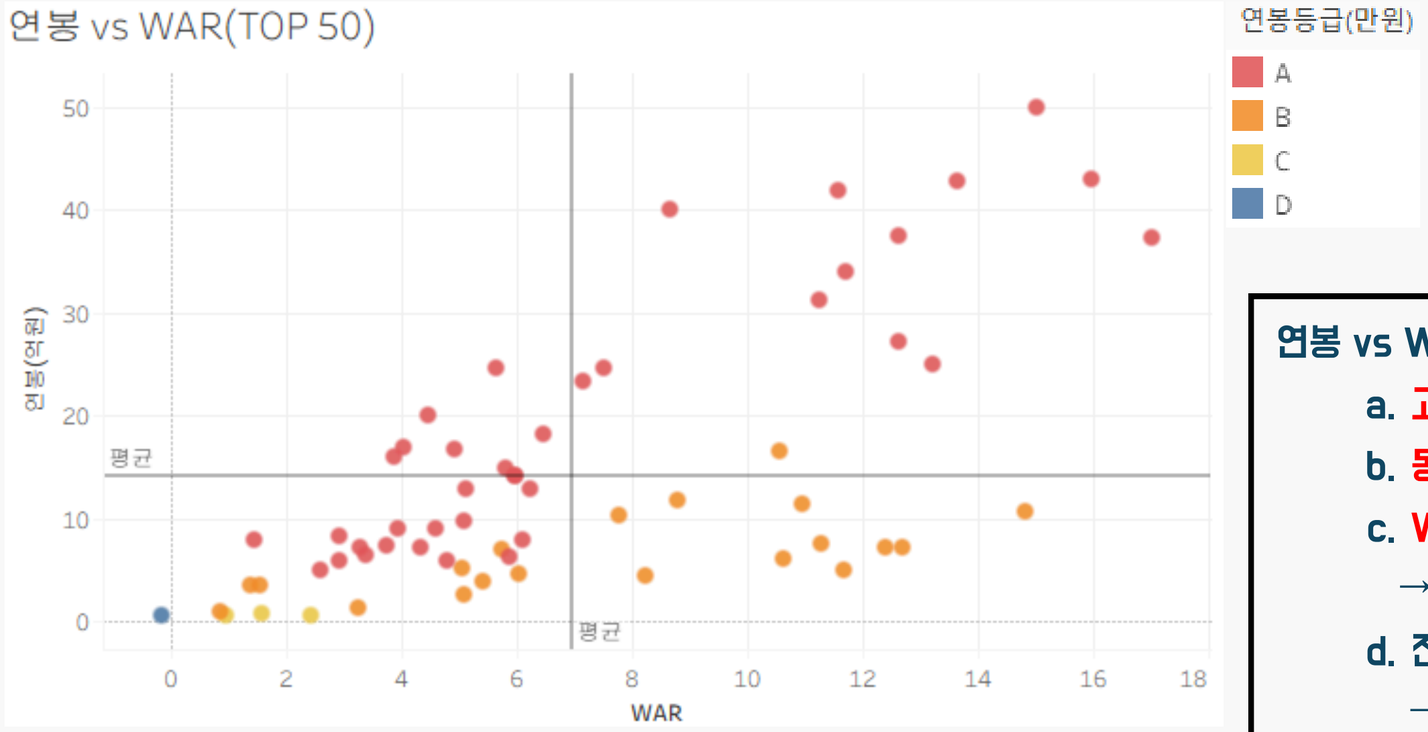
- 수비에서 중견수, 포수의 지표가 두드러짐.
- 수비 지표만으로는 연봉 반여 여부가 잘 보이지 않음





데이터셋 살펴보기

투수의 투구 지표 & 연봉 데이터 분석 (WAR은 높을수록 좋은 성과)



연봉 vs WAR

a. 고연봉·저WAR 선수 다수 존재 → 성과 대비 과대평가

b. 동일 WAR에서도 연봉 차이가 큼

c. WAR 상위권 선수는 주로 A/B 등급에 집중

→ 성과에 비례해 상위 연봉대에 포진

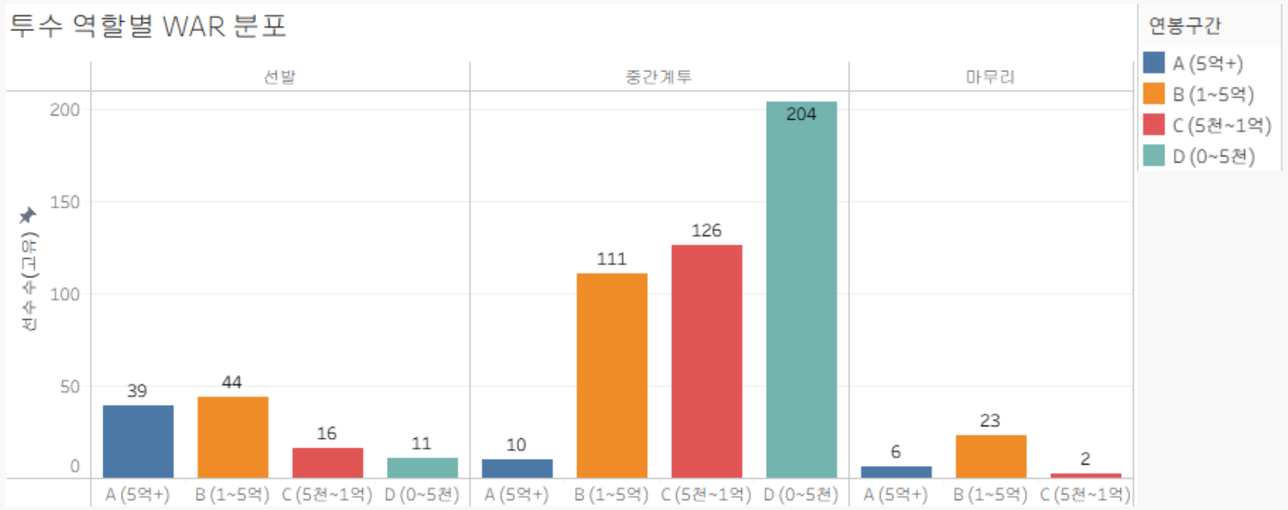
d. 전체적으로 연봉과 WAR 간 완벽한 선형 상관은 없음

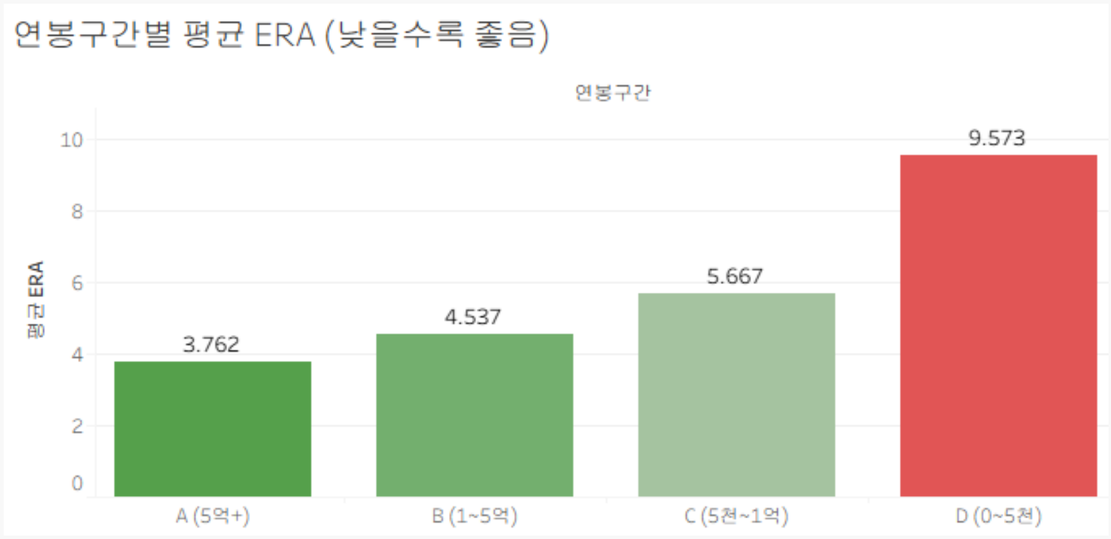
→ 성적 외 요인(나이, 외국인 여부, 계약 구조 등)이 연봉 결정에 영향.

연봉은 성과(WAR)와 부분적으로만 연관되며, 외부 요인으로 인해 성과 대비 불균형이 존재

데이터셋 살펴보기

투수의 투구 지표 & 연봉 데이터 분석





투수 역할별 WAR 분포

a. 선발투수는 고연봉·고WAR 선수들이 집중

b. 중간계투/마무리는 성과 편차가 크고 연봉 구간이 섞여 있음

c. 투수 역할별 연봉·성과 간 격차 구조가 뚜렷하게 드러남

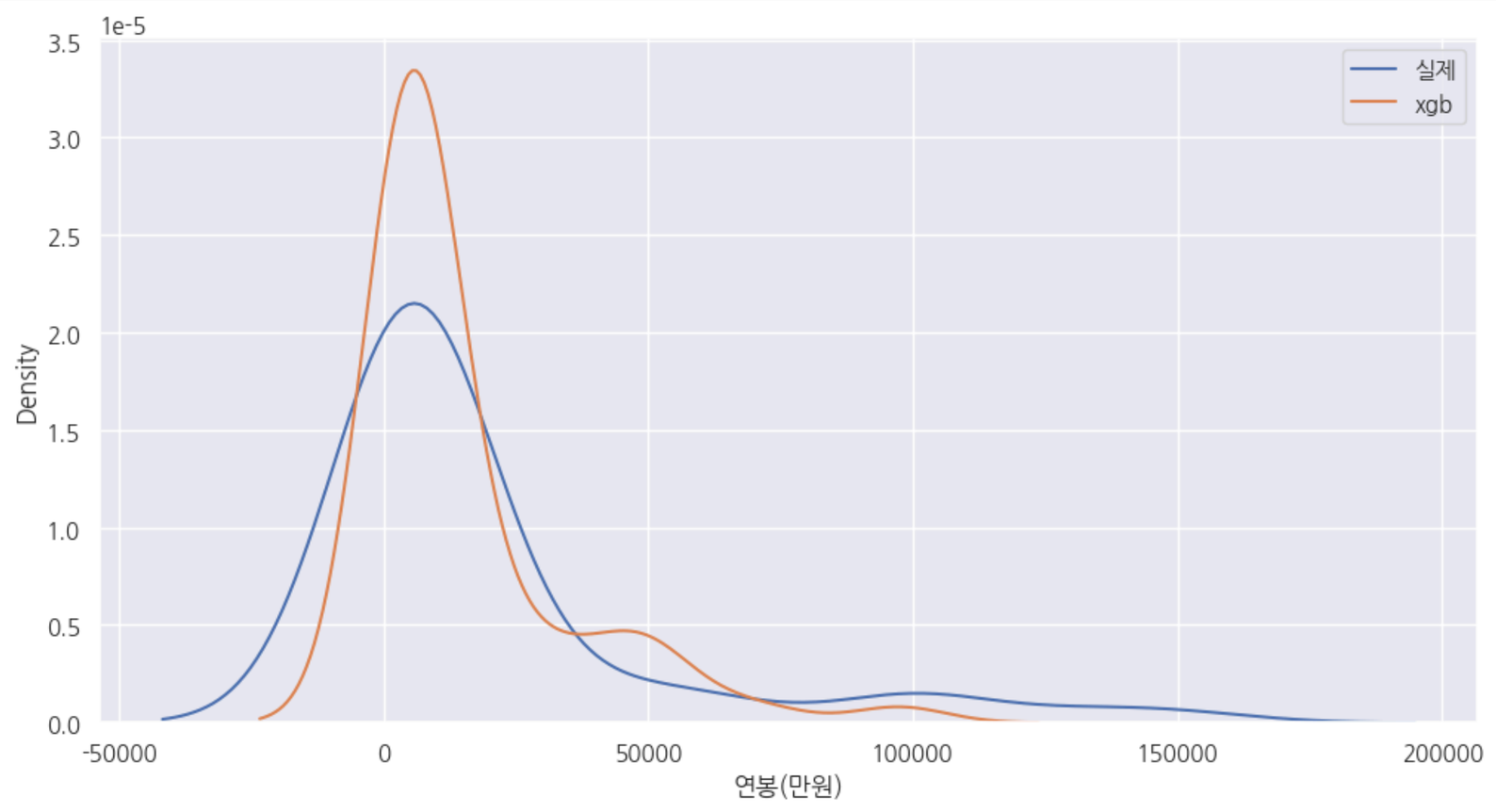
연봉구간별 평균 ERA

a. ERA는 낮을수록 우수 → A/B구간은 평균 4점대 이하로 안정적

b. C구간은 5점대, D구간은 9점대 이상으로 압도적 열세

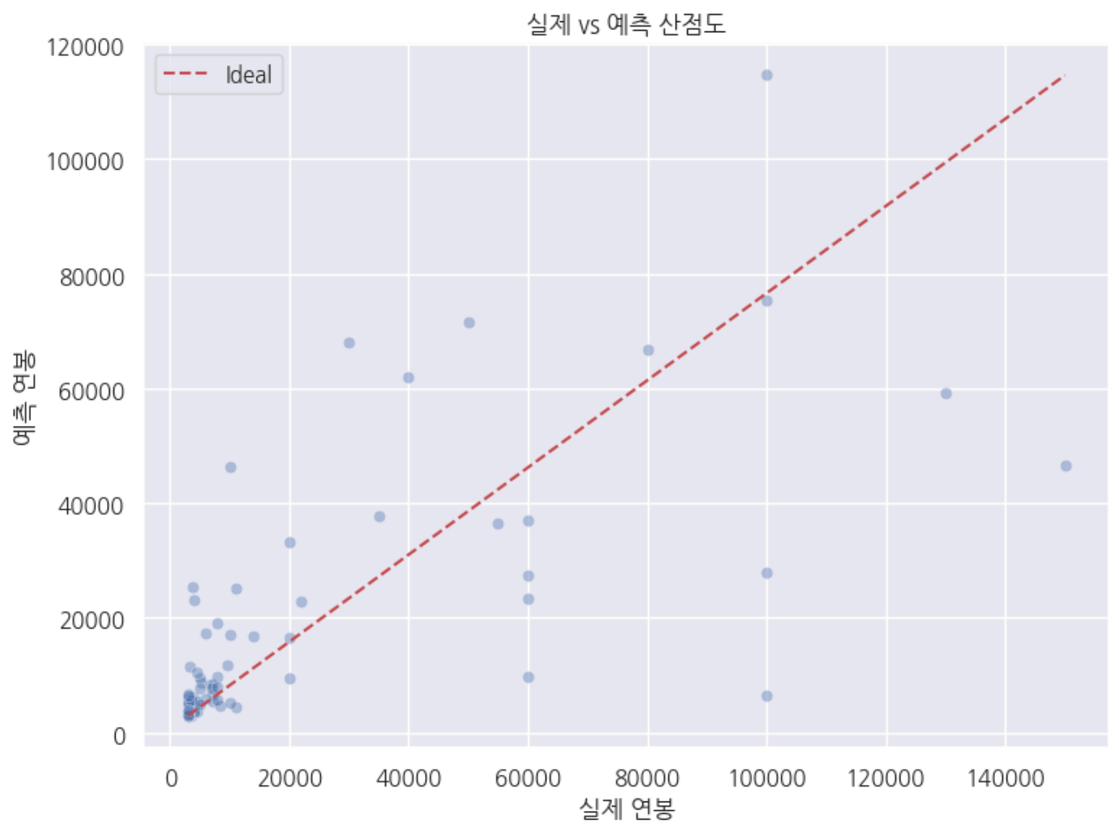
c. 저연봉일수록 실점 억제력이 떨어지는 패턴이 분명하게 나타남

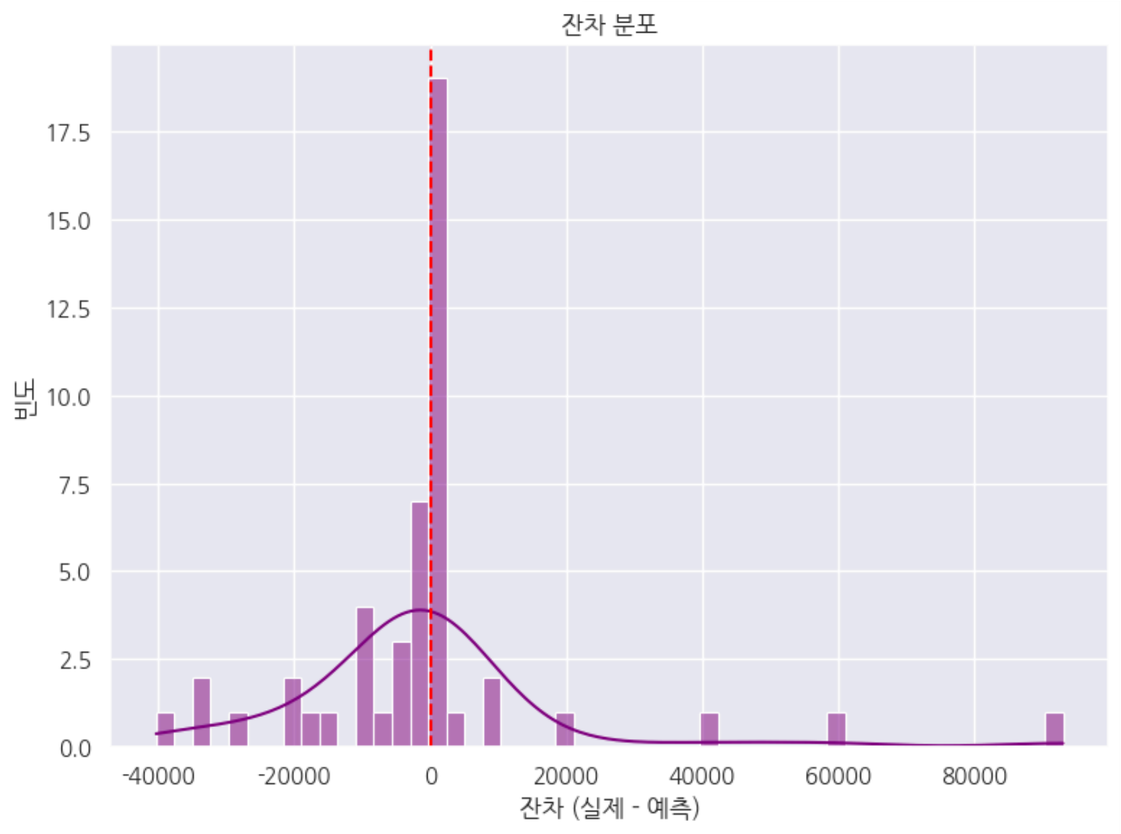
머신러닝



- 연봉 분포가 정규분포를 하지않고 스탯들과 선형관계가 없기 때문에 선형 모델이 적합하지 않음
- 트리 기반의 RandomForest나 LGBM, XgBoost등의 부스팅 모델 사용
- RMSE, RAE, R2score 지표를 사용해 종합적으로 판단

머신러닝 - 분포 확인





- 높은 연봉을 받는 선수들에 MSE등의 지표가 크게 영향을 받음
- 해당 선수를 목표로 추가 자료수집이 필요
- 또한 전체적인 지표에서 문제확인.

수상과 연봉 상관관계

상관계수 히트맵

총수상횟수 0.313.

골든글러브 0.252.

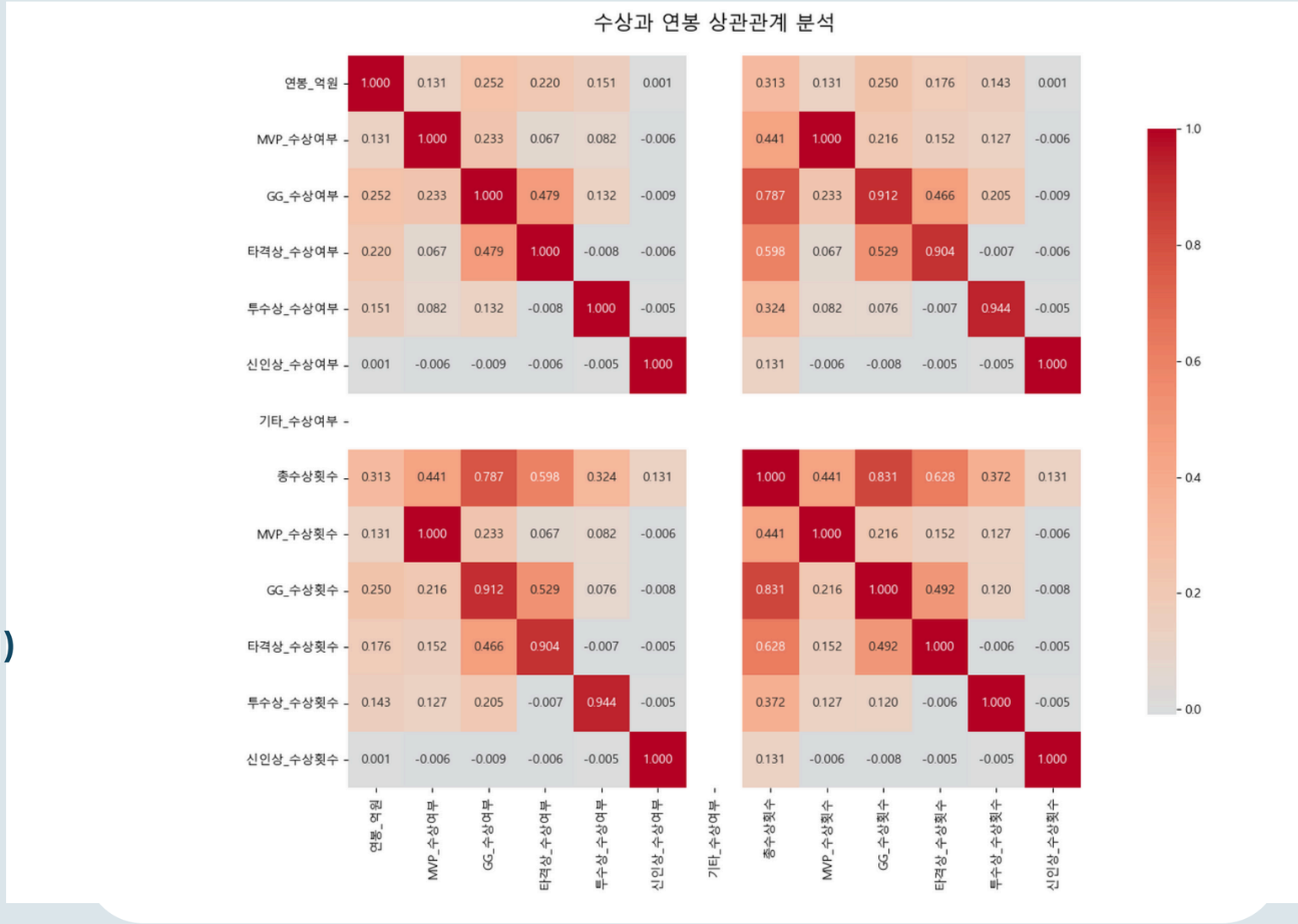
타격상 0.2197 등의 연관 관계

신인상은 거의 0 (연봉 영향 없음)

수상 경력은 연봉과 통계적으로 유의한 상관관계 (p<0.001)

MVP·타격상·골든글러브·투수상은 강력한 연봉 프리미엄

신인상은 연봉 인상과 무관



수상과 연봉 상관관계

고연봉 선수 대부분이 수상자이지만

→ 수상이 곧 고연봉을 보장하지는 않음

수상 횟수 ↑ → 일정 수준 이상에서는 추가 효과 미미

수상자 평균 연봉이 높지만

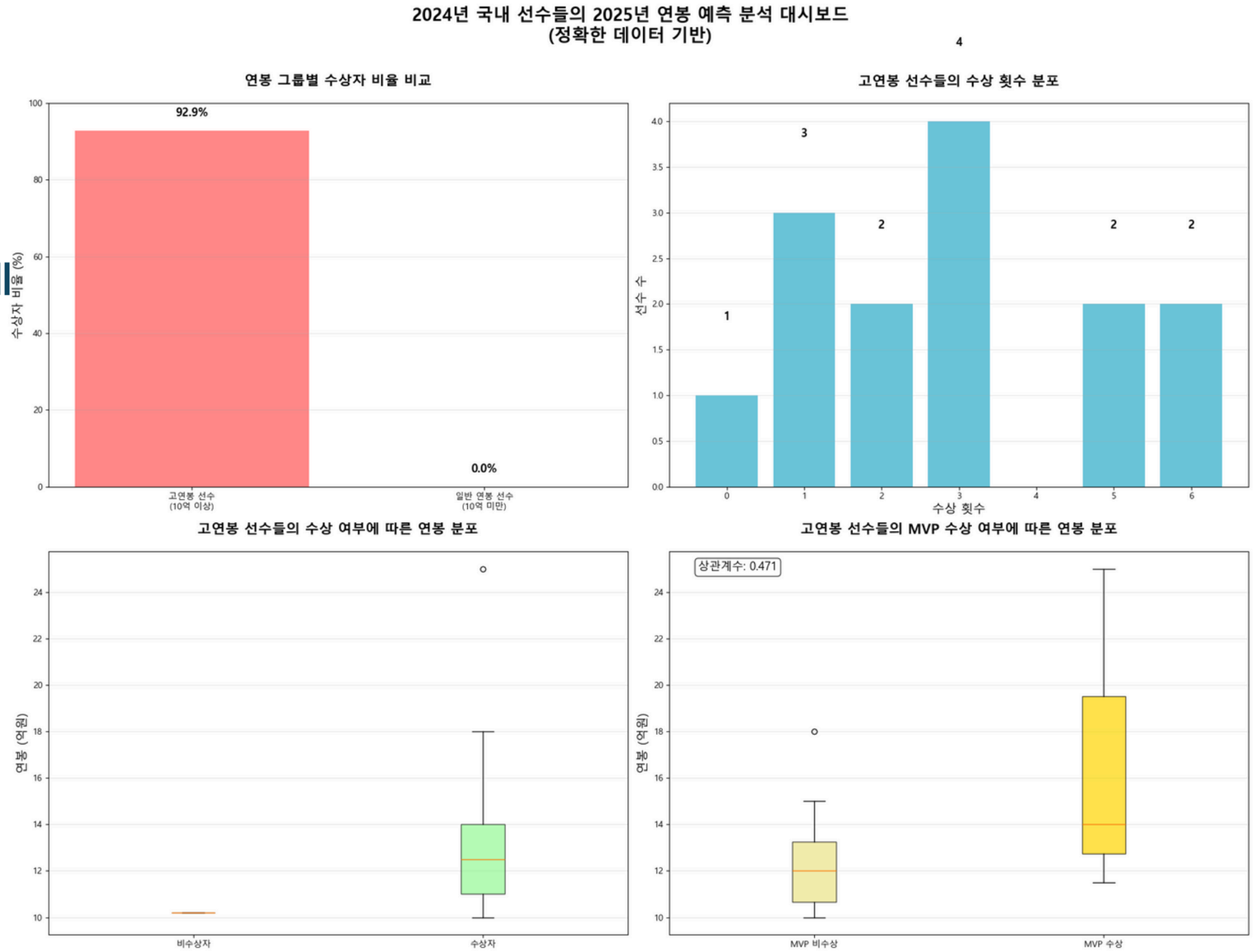
→ 분포 겹침 큼, 설명력 낮음

MVP는 프리미엄 요인이지만

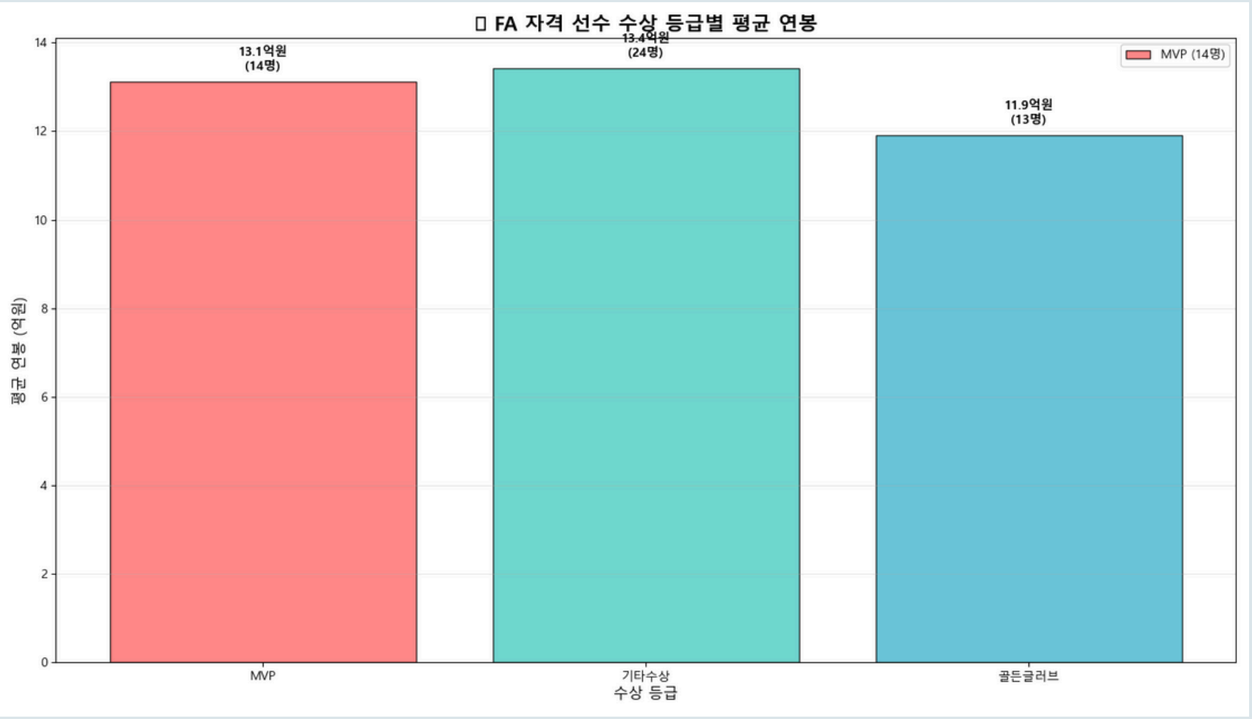
→ 절대적 결정 요인 아님

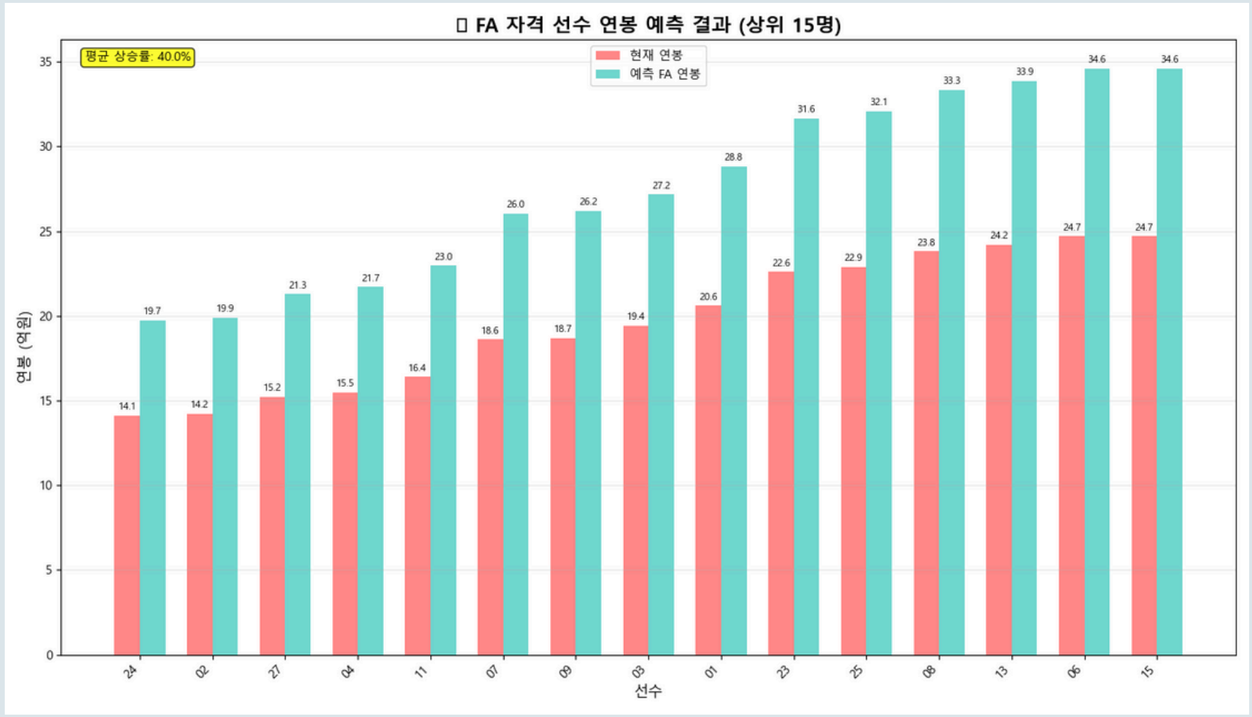
결론: 연봉은 수상 경력보다 FA 시점·

포지션 가치·시장 상황에 더 크게 좌우됨



수상과 연봉 상관관계 예측 (10억 이상)





3. 수상 등급별 평균 연봉

MVP: 13.1억 / 기타: 13.4억 / 골든글러브: 11.9억

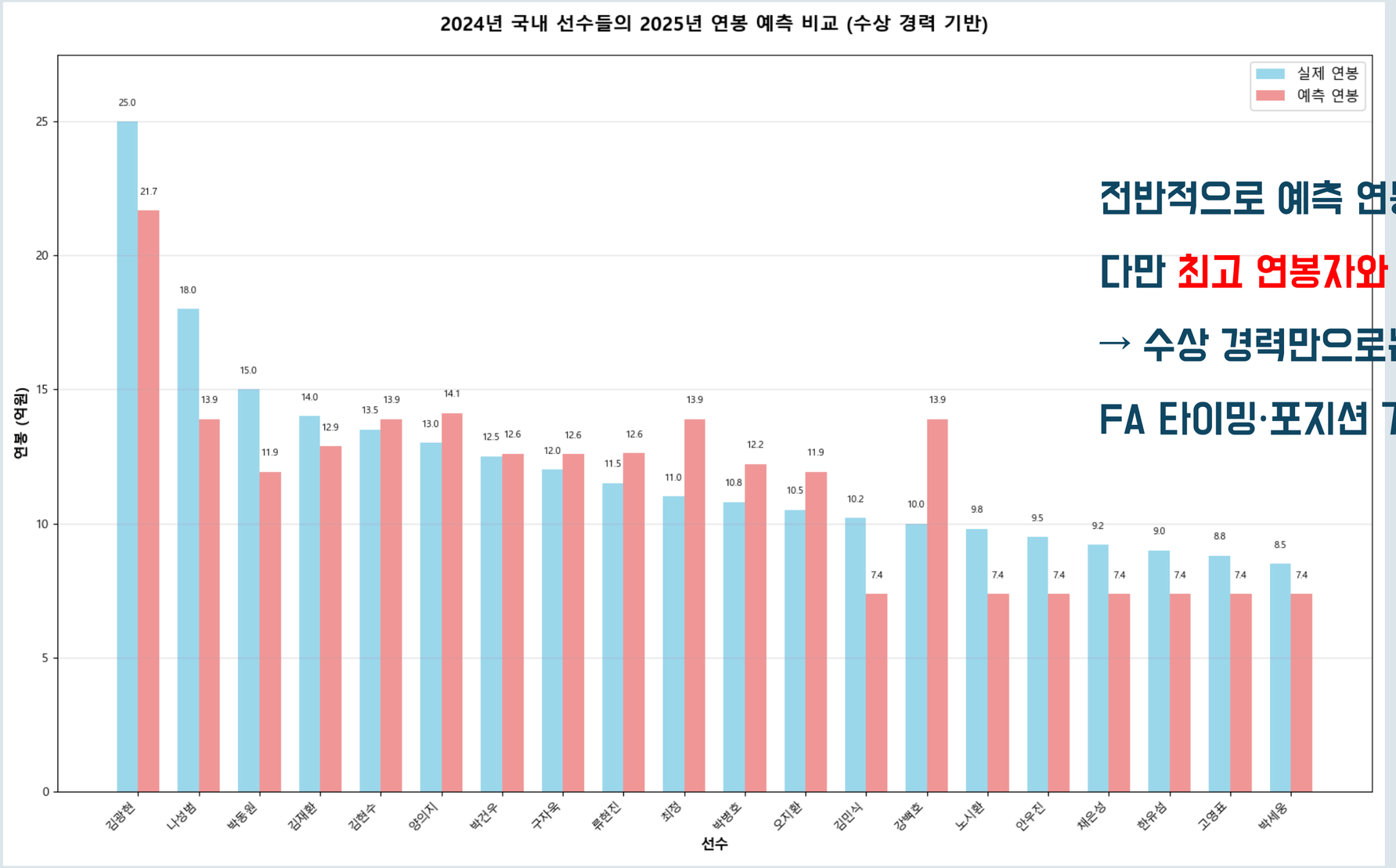
MVP·기타 차이 적음, 골든글러브는 평균 낮음

4. 수상 여부별 평균 연봉

수상 有: 13.4억 / 수상 無: 19.0억

무수상자도 FA 시점 조건 좋으면 고연봉

수상과 연봉 상관관계 예측 (10억 이상)



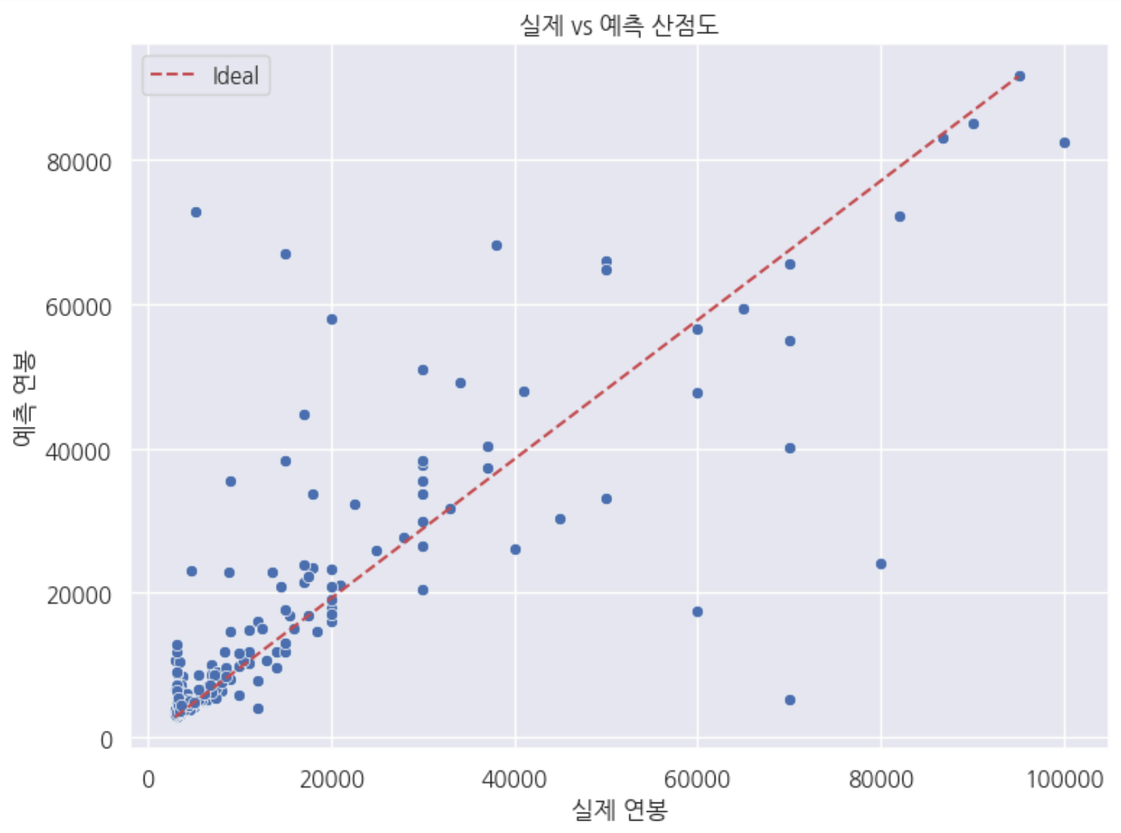
전반적으로 예측 연봉과 실제 연봉 간 큰 차이 없음.

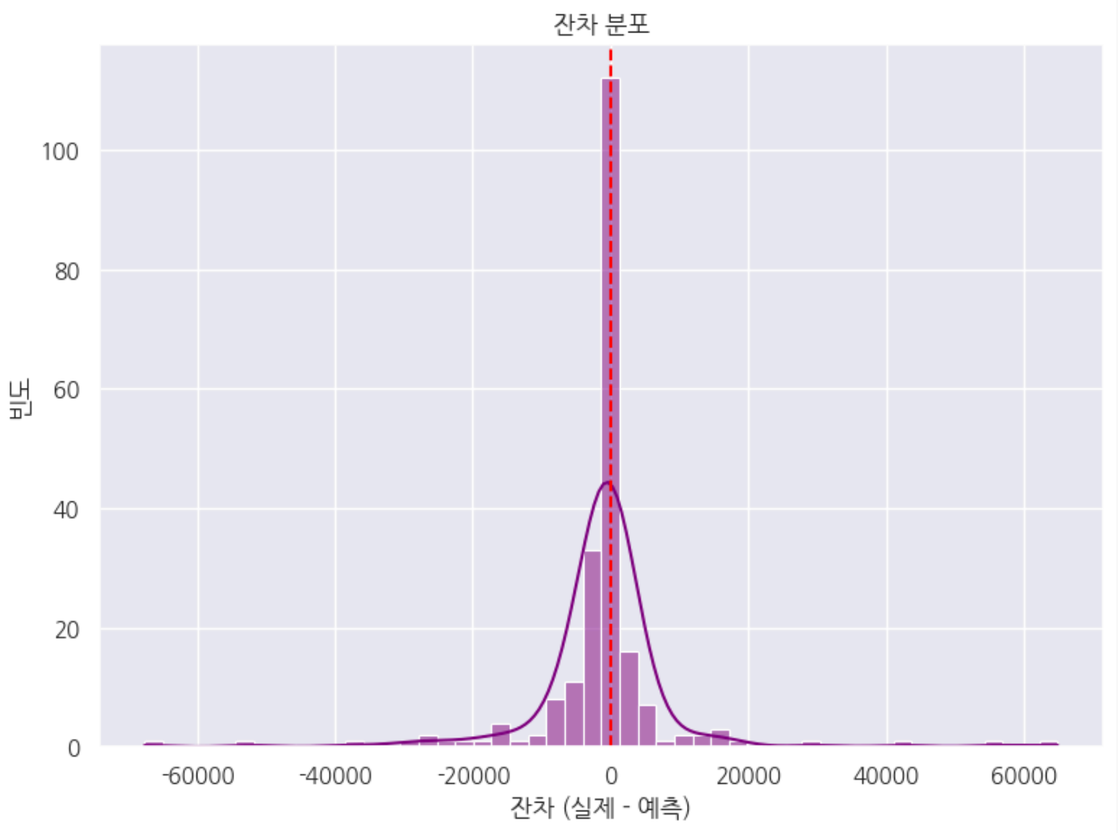
다만 **최고 연봉자와 일부 중·하위권 선수에서 오차 발생.**

→ 수상 경력만으로는 연봉을 안정적으로 설명하기 어려움.

FA 타이밍·포지션 가치·시장 상황 등 다른 요인 고려 필요.

수상경력을 고려한 개선





- 수상횟수를 Feature로 추가
- 10억 이하의 선수들로 필터링해 지표에 혼동을 줄이고자함
- 시각적으로도 개선정도를 확인 할 수 있음

머신러닝 - 분류 (2021년 ~ 2024년 데이터)

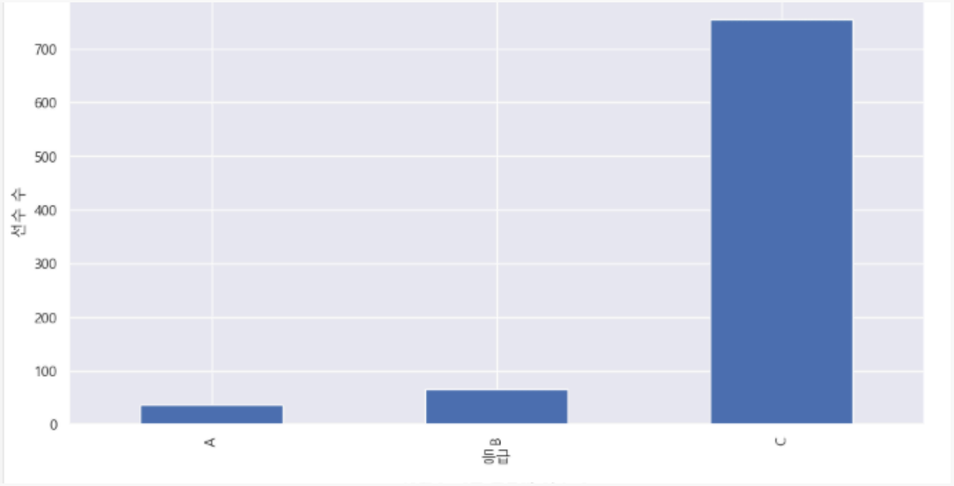
분석·예측은 '공식 FA등급'이 아니라, **편향을 줄인** '연봉 4구간' 라벨로 진행

연봉 + 신인 계약금 + 외국인 선수 계약금 + FA 계약금 ÷ 계약기간 → 최종 연봉

신인·외국인 일시 반영, FA 계약기간 분할 → '**최종 연봉**' 산출, 연봉 데이터 A-D **4구간 라벨링**

누수 방지: 팀/전체 연봉순위·**미래 연봉/계약** 제외, **결측/불일치** 제거

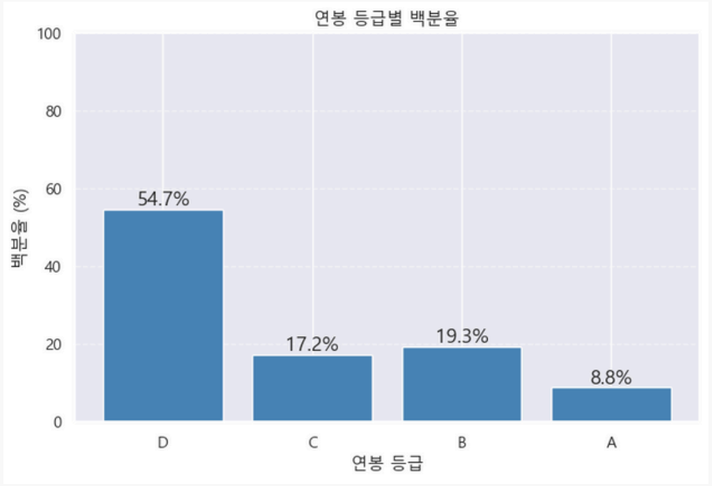
공식 FA등급 분포



- KBO 공식 FA등급 기준으로 선수 수를 집계한 분포. (2021년 ~ 2024년)
- 특정 등급에 선수들이 **과도하게 몰림**(제도/협상/신인 규정 등 **비정량 요인** 영향).

  ⇒ 학습 라벨로 쓰기엔 **편향·노이즈가 큼.**

학습용 연봉 등급 분포



- 학습을 위해 연봉 금액 4구간으로 단순·균형화
- 구간: 0-5천 / 5천-1억 / 1-5억 / 5억+.

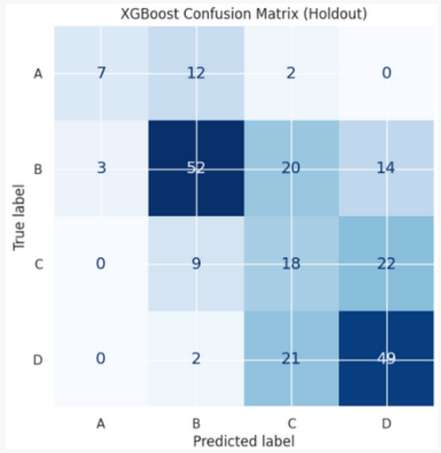
  ⇒ **클래스 불균형 완화 + 라벨 일관성↑**

머신러닝 - 모델 성능(F1 스코어) & 혼동행렬

2021년 ~ 2024년 투수 데이터

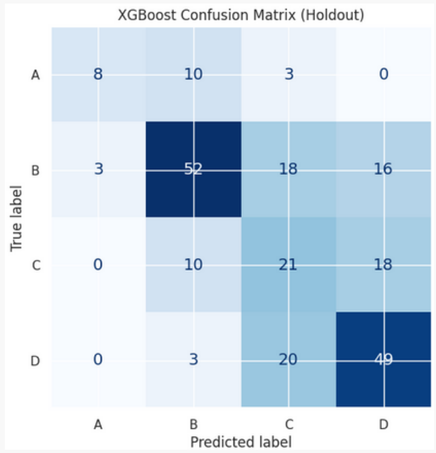
1차(기본 스탯)

Model	F1 Macro	F1 Micro
XGBoost Tuning	0.6031	0.6280
LGBM Tuning	0.5995	0.6301
SoftVoting Tuning	0.5953	0.6260
HardVoting Tuning	0.5923	0.6156
GradientBoost Basic	0.5882	0.6073



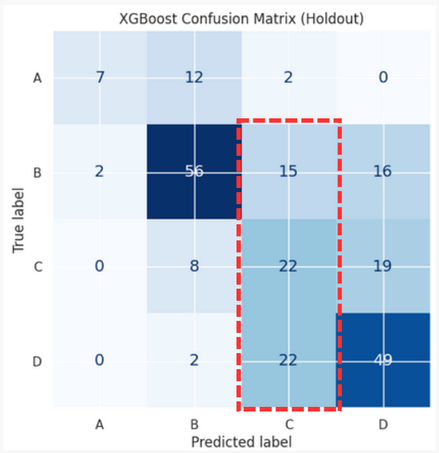
2차(+선발/중간/마무리)

Model	F1 Macro	F1 Micro
XGBoost Tuning	0.6274	0.6598
LGBM Tuning	0.6049	0.6392
SoftVoting Tuning	0.5885	0.6247
HardVoting Tuning	0.5864	0.6144
XGBoost Basic	0.5786	0.6124



3차(+신인)

Model	F1 Macro	F1 Micro
XGBoost Tuning	0.6368	0.6528
HardVoting Basic	0.6228	0.6384
SoftVoting Basic	0.6223	0.6405
LGBM Basic	0.6221	0.6405
LGBM Tuning	0.6128	0.6280



F1 스코어

- 최종적으로 XGBoost가 가장 높은 성능을 보임

혼동행렬

- 1차: 오류가 C↔B/D 에 집중
- 2차: 역할 피처로 C행 오프대각 감소
- 3차: 가장 완만 — 대각선 진해짐, C 오류 최소

가로=실제, 세로=예측, 대각선=정답,

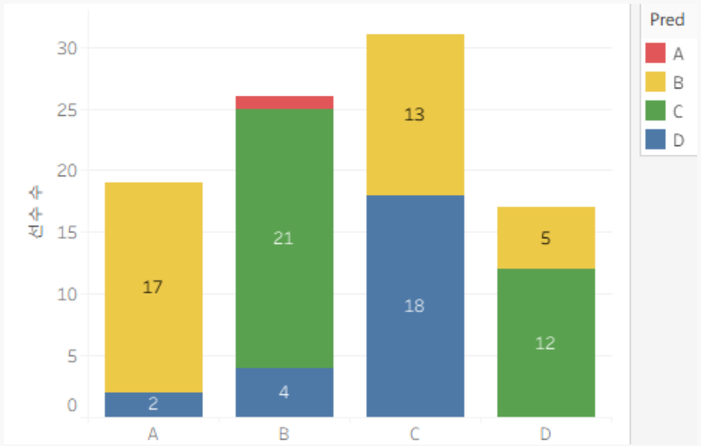
대각선 밖=오류(진할수록 비중 큼)

데이터: 2021–2024 투수 / 학습≈950, 예측≈200 / 지표: F1-macro(클래스 동일가중), F1-micro(전체 가중)

머신러닝 - 예측 결과(오분류 분석)

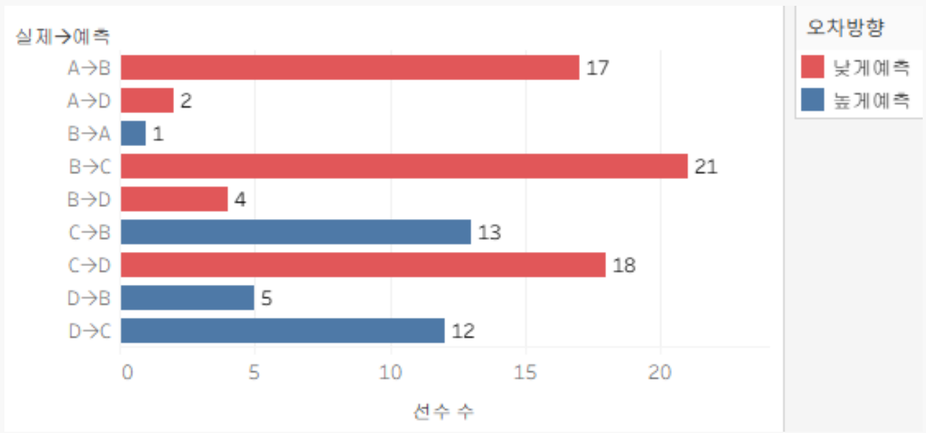
오류의 대부분은 C를 B·D로 헷갈림 / 팀별 편차가 있고, 2단계 이상 큰 오차는 소수의 예외

실제등급별 오분류



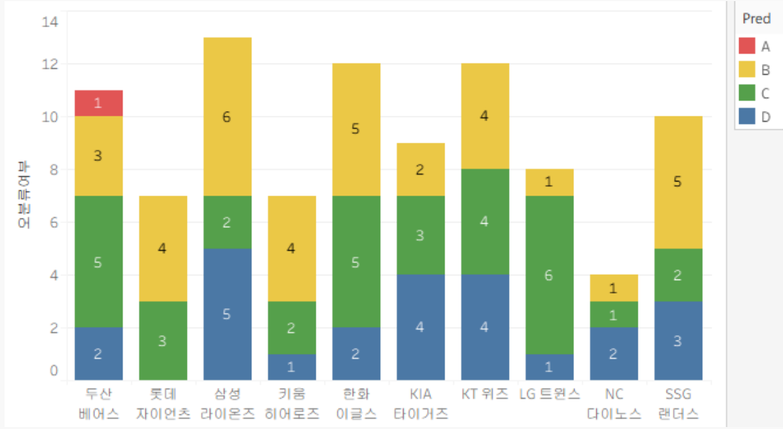
C↔B/D 경계 불안정

오분류 흐름(실제→예측)



대부분 한 단계 차이, 큰 오차는 소수

팀별 예측 등급



큰 오차는 일부 팀에 몰림

1. 전체 오류의 1/3이 C 경계에 집중 (합 31건 : 전체 93건 중 33%) .
2. 팀 별 차이가 존재 : 큰 오차가 특정 팀에 상대적으로 집중.
3. 두 단계 이상 큰 오차는 드물어 전반적으로 방향성은 유지

머신러닝 - F1 스코어

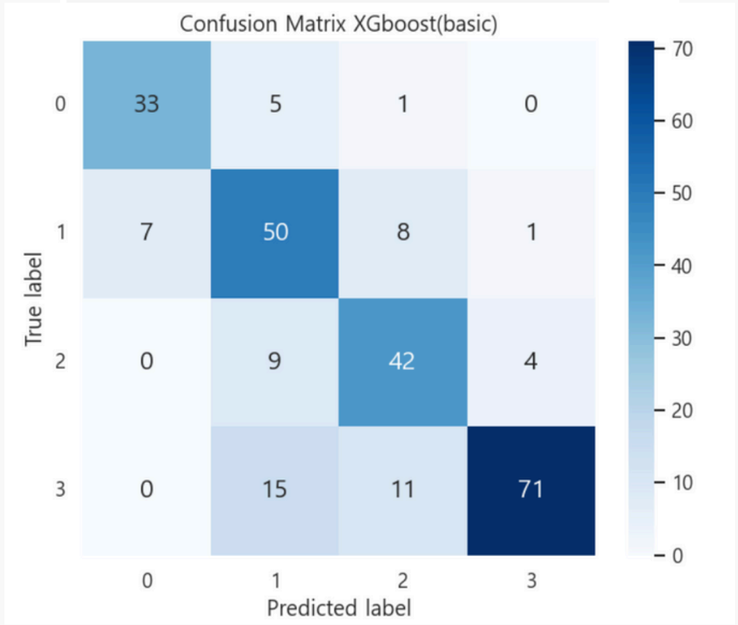
2021년 ~ 2024년 타자 데이터

기본 모델과
하이퍼 파라미터 학습

- 기본 모델보다 튜닝 모델이 소폭 개선된 성능을 보임
- LGBM은 안정적인 성능과 낮은 오분류율을 확인

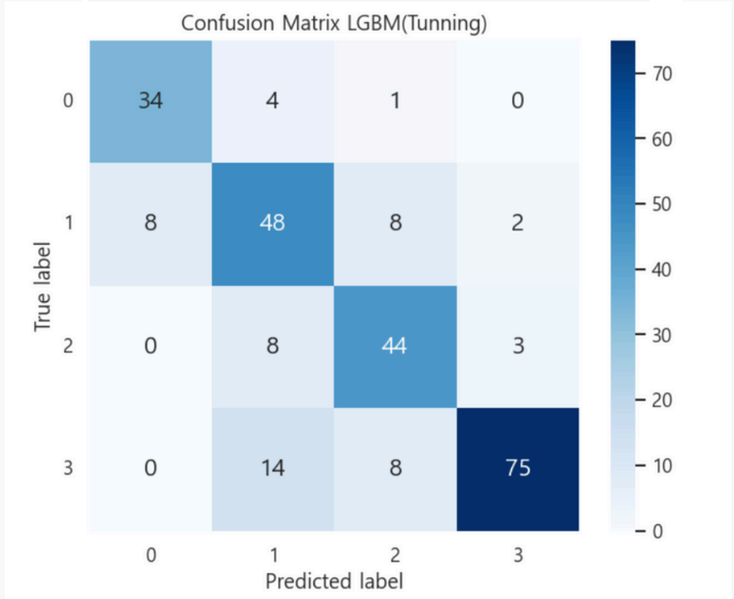
기본 모델 XGBoost

	f1 score
XGBoost Basic micro	0.831007
GradientBoost Basic micro	0.826860
SoftVoting Basic micro	0.826842
LGBM Basic micro	0.822788
XGBoost Basic macro	0.815996
HardVoting Basic micro	0.815957
RandomForest Basic micro	0.815920



하이퍼파라미터 튜닝 LGBM

	f1 score
HardVoting Tuning micro	0.855461
LGBM Tuning micro	0.850000
GradientBoosting Tuning micro	0.840485
HardVoting Tuning macro	0.840316
SoftVoting Tuning micro	0.837745
LGBM Tuning macro	0.837151
XGBoost Tuning micro	0.832303



데이터: 2021-2024 타자 / 지표: F1-macro(클래스 동일가중), F1-micro(전체 가중)

머신러닝 - 오분류 분석

연차 효과(계약 제도)

- 1-3년차 구간은 **연봉 하한/협상 구조/FA 미자격** 등으로 실연봉이 성적을 늦게 따라감

- 모델은 성적이 좋으면 등급을 올려 찍는 경향
  → **D(실제) vs B/C(예측) 오분류**

오분류 (실제→예측):

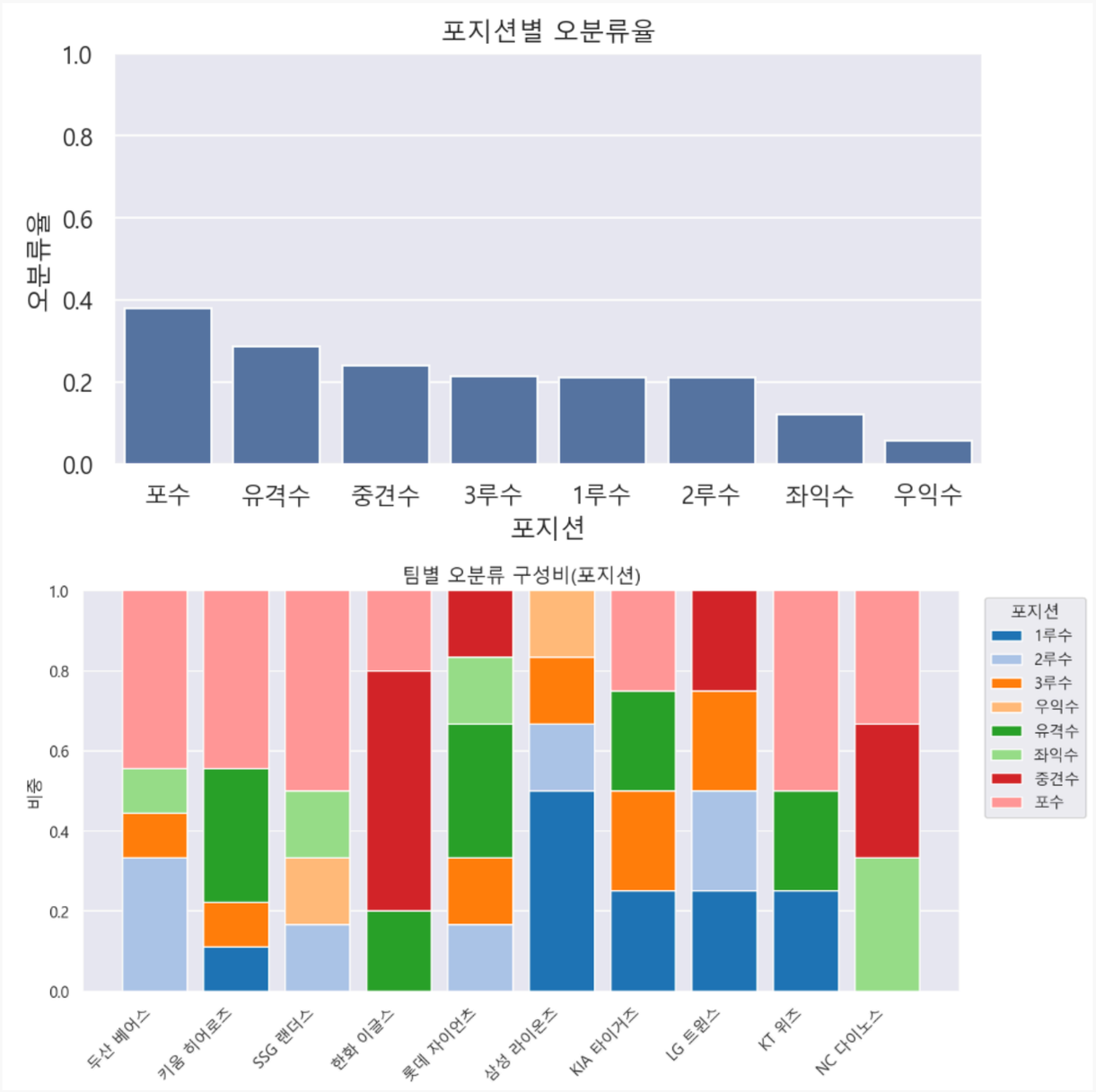
	true	pred	건수
0	D	B	14
1	B	A	8
2	B	C	8
3	C	B	8
4	D	C	8
5	A	B	4
6	C	D	3
7	B	D	2
8	A	C	1

선수	팀	포지션	연도	TRUE	Pred	연차
김두현	KIA 타이거즈	유격수	2024	D	B	1년차
김범석	LG 트윈스	1루수	2024	D	B	2년차
김현종	LG 트윈스	중견수	2024	D	B	1년차
박한결	NC 다이노스	좌익수	2024	D	B	2년차
신용석	NC 다이노스	포수	2024	D	B	2년차
정현승	SSG 랜더스	우익수	2024	D	B	1년차
여동건	두산 베어스	2루수	2024	D	B	1년차
전다민	두산 베어스	좌익수	2024	D	B	1년차
류현준	두산 베어스	포수	2024	D	B	1년차
강성우	롯데 자이언츠	유격수	2024	D	B	1년차
이호준	롯데 자이언츠	유격수	2024	D	B	1년차
이재상	키움 히어로즈	유격수	2024	D	B	1년차
김동헌	키움 히어로즈	포수	2024	D	B	2년차
한경빈	한화 이글스	유격수	2024	D	B	3년차

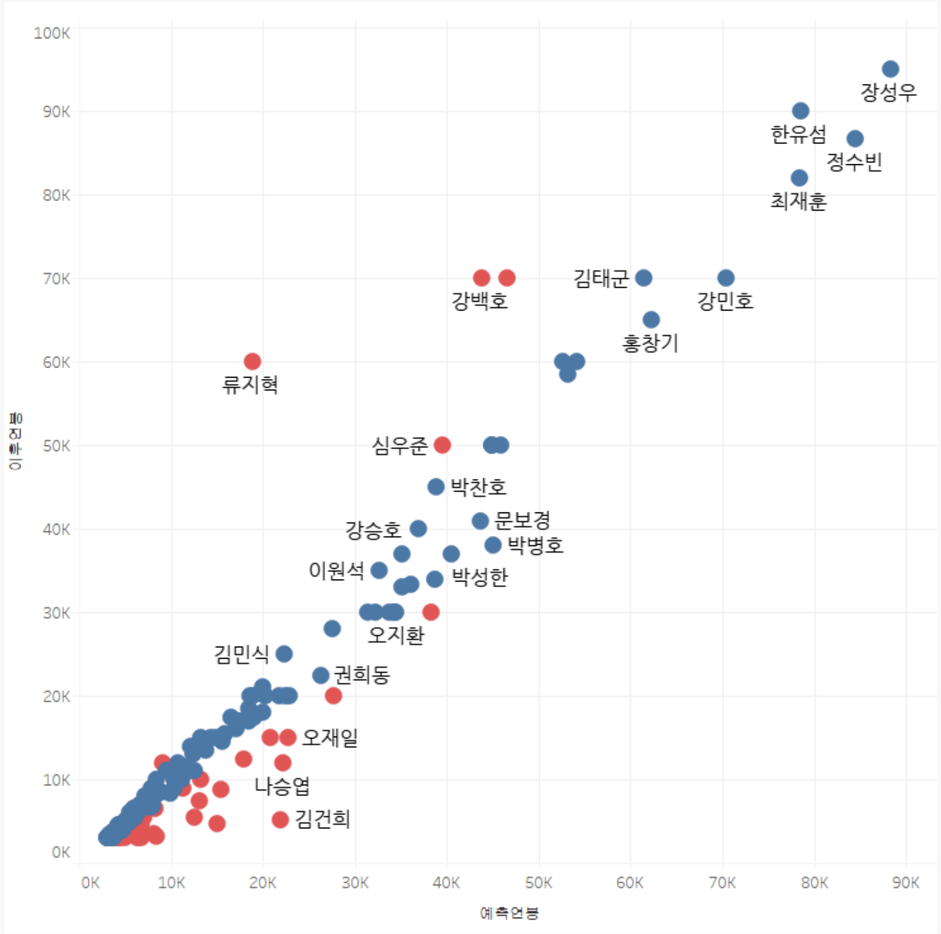
머신러닝 - 오분류 분석

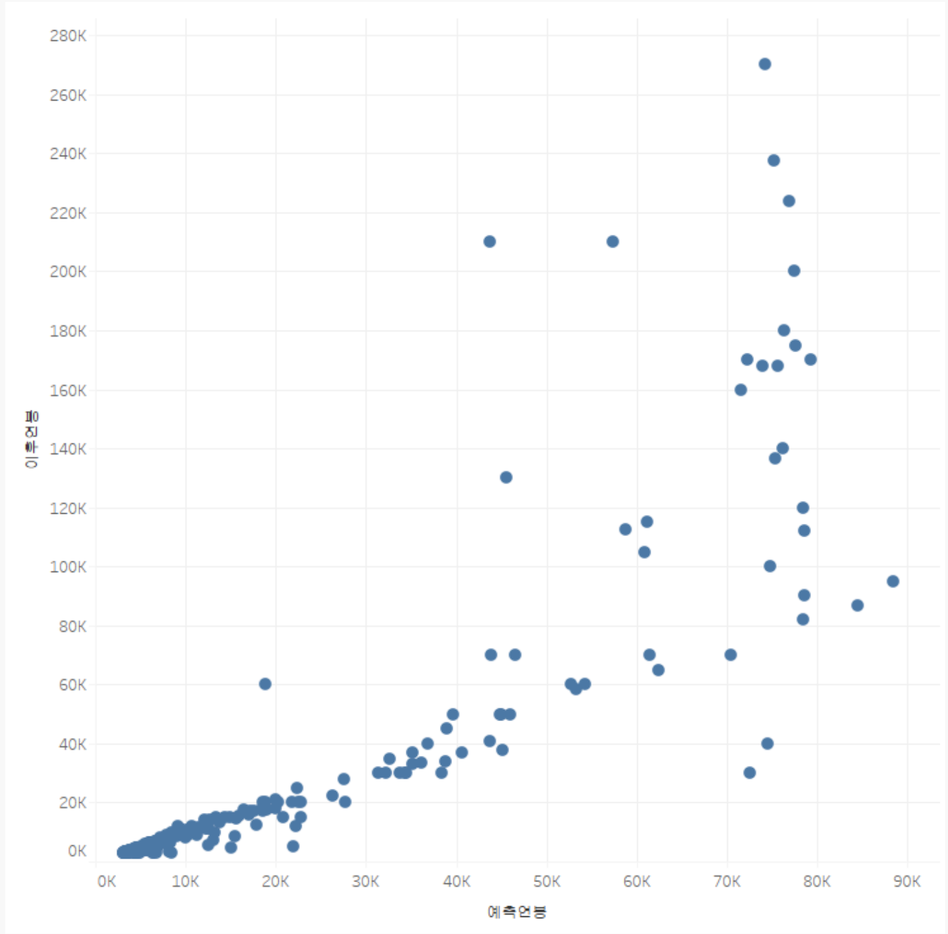
역할·포지션·팀 정책

- 포수/유격수/중견수처럼 **수비가치/희소성 포지션**은 지표 대비 연봉이 다르게 움직일 수 있음
- 팀별 인사/예산/재계약 관성도 반영 필요(같은 성적이라도 팀 따라 급여 등급이 다름).



추가된 칼럼을 반영한 개선





- 팀의 연봉규모(상위40인 샐러리캡), 포지션 정보, 계약금을 연봉에 반영한 데이터셋으로 회귀분석 진행
- 포지션 변경이나 장기부상등의 수집하지 못한 데이터가 오류의 주 원인
- 10억 이상의 선수에 대한 예측은 실패

결론 - 24년 FA계약

- 회귀분석을 토대로 2024년 FA자격이 있는 선수들을 대상으로 변경되는 선수들을 조사.
- 수집한 데이터는 연차를 고려하지않아 등급이 오른 베테랑 선수가 눈에 띔
- D등급을 신설하고자 하였으나 해당 등급의 선수는 FA신청요건을 충족시키지 못함

등급 변경선수	원FA등급	변경FA등급
최정	C	A
엄상백	B	A
허경민	B	A
구승민	A	B
최원태	A	B
우규민	C	B
임기영	B	C
장현식	B	C
류지혁	B	C
오재일	B	C
장현식	B	C
진해수	B	C
이재원	B	C
하주석	B	C

회고 및 개선 가능성

데이터 수집

연봉에 영향을 주는 요소들이 수집하기 어려운 형태여서 반영하기 어려웠다.
KBO의 구단과 선수숫자가 적어 수집에 한계가 있었다.

머신러닝

머신러닝 하이퍼 파라미터 튜닝이나 모델 선정등에서 기술적으로 부족하여서 오랜 시간이 걸린것이 아쉬웠다.

추후 연구 제안

시계열 연구를 통해 한해의 성적이 아니라 누적된 성적의 변화를 연봉에 반영할 수 있다면 충분한 개선이 될 것으로 보인다.
FA등급 뿐만이 아니라 퍼포먼스 지표가 대안이 될 수 있는 카스포인트 등의 지표 개선에도 사용될 수 있을 것으로 생각된다.

Thank you