2023

KBO 프로야구 선수 연봉 예측

김서현

Contents

KBO 타자 연봉 예측

1. 분석 배경

- 분석 배경 소개
- 현황 분석

2. 데이터 분석

- 연봉 예측을 위한 모델 적용 및 분석

3. 연봉 예측

- 분석 결과
- 연봉 예측

현재 프로야구 연봉 산정 시스템

구단별로 100개 이상의 평가 항목을 게임마다 적용하고 점수로 환산해 합산하는 방식

+

프랜차이즈 스타

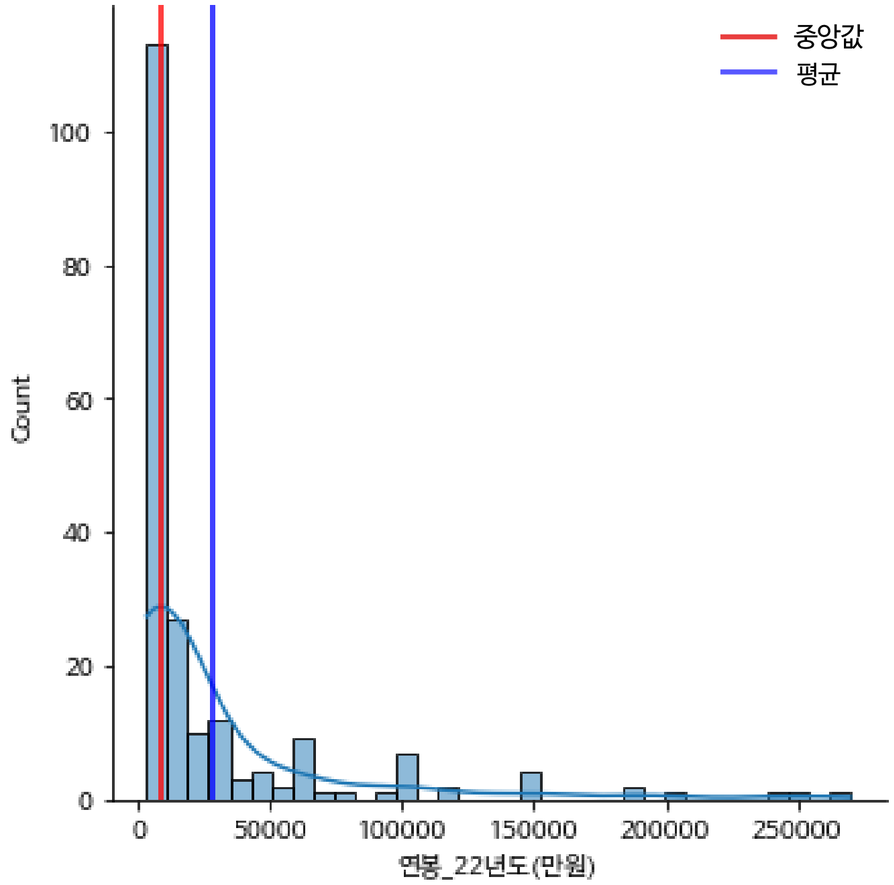
FA
(자유계약선수)

선수들의
그라운드 밖
행동

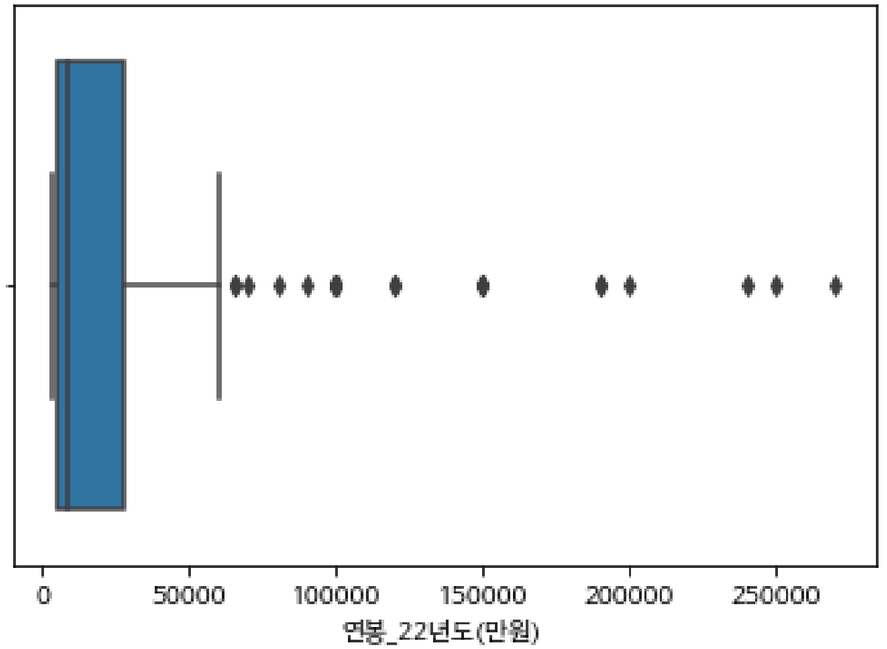
구단별 세부 기준

현황 분석

2022년도 KBO 타자 연봉 히스토그램



2022년도 KBO 타자 연봉 상자그림



분석 데이터

21_KBO_batter.csv

이름
팀/포지션
G
타석
타수
득점
안타
2타
3타
…

255 X 29

21_KBO_salary.csv

팀
선수
연봉_21년도(만원)

716 X 3

22_KBO_salary.csv

팀
선수
연봉_22년도(만원)

676 X 3

분석 데이터

22_KBO_salary.csv

+

21_KBO_salary.csv

22년도 연봉 데이터를 기준으로
두 연봉 데이터 병합

(이적한 선수들을 최신 팀으로
지정해야 하기 때문)

+

21_KBO_batter.csv

선수 이름과 팀 명을 기준으로
데이터 병합

세이버 메트릭스 지표 변수 생성

Feature Engineering

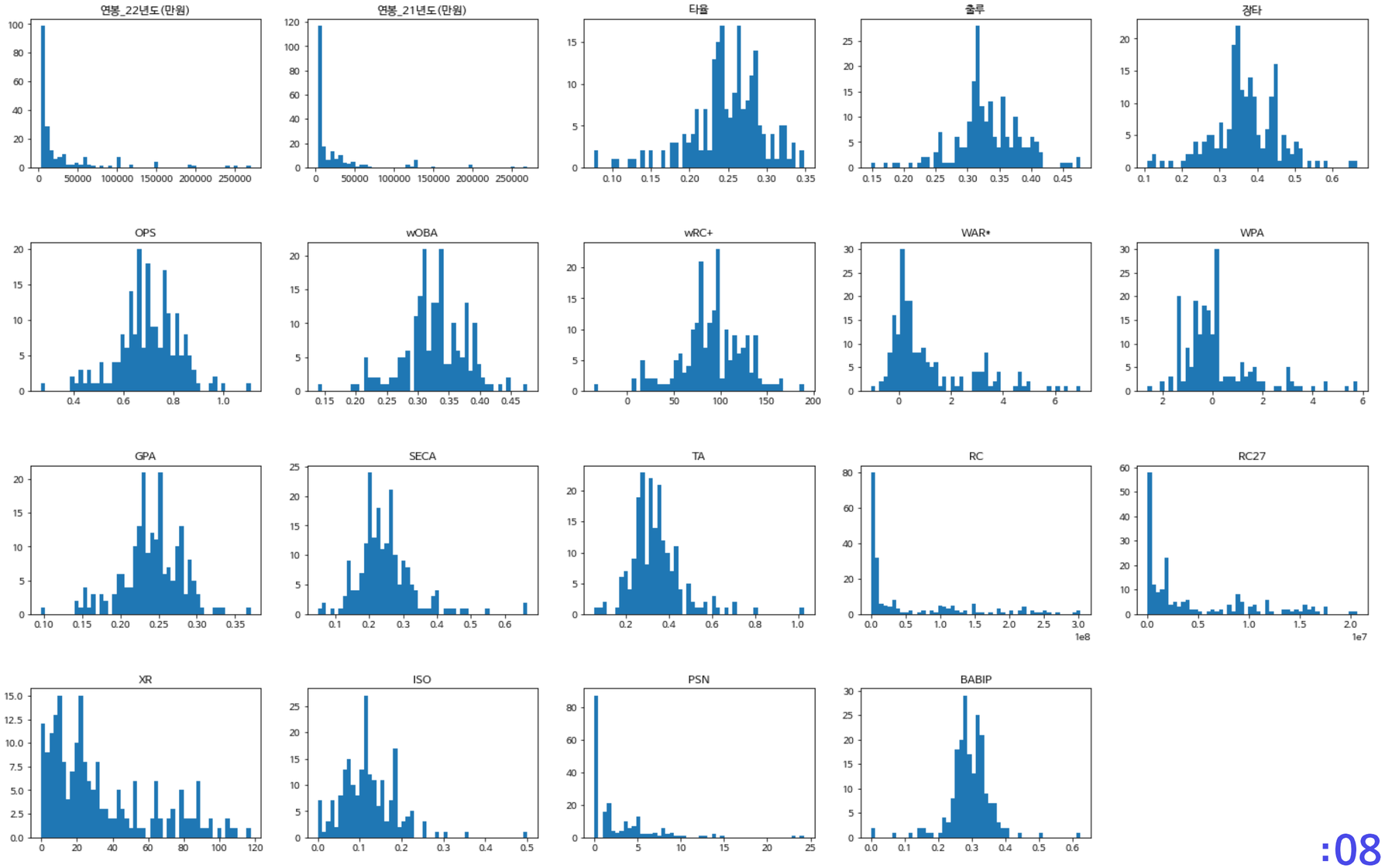
기존의 데이터를 이용하여 세이버 메트릭스 지표 feature 생성

- GPA : 타자의 공격 능력 (OPS 개선)
- SECA : 타율에서 장타의 요소를 추출한 지표
- TA : 타자가 1아웃당 얼마나 많은 루타를 얻을지를 나타내는 지표
- RC : 선수의 활약을 득점에 공헌한 값으로 평가하는 지표
- RC/27 : 27 아웃 (9 이닝 × 3 아웃 = 1 경기)에서 평균 몇점을 취할지 산출한 지표
- XR : 중회귀 분석을 이용하여 각 플레이에 가중치를 두어 점수에 미치는 영향을 수치화한 지표
- ISO : 순수장타율
- PSN : 파워와 스피드 두 능력이 모두 뛰어난 선수 확인하는 지표 (타자의 호타준족 정도)
- BABIP : 홈런 이외에 필드로 날아간 타구의 안타가 될 비율

세이버 메트릭스 지표 변수 생성

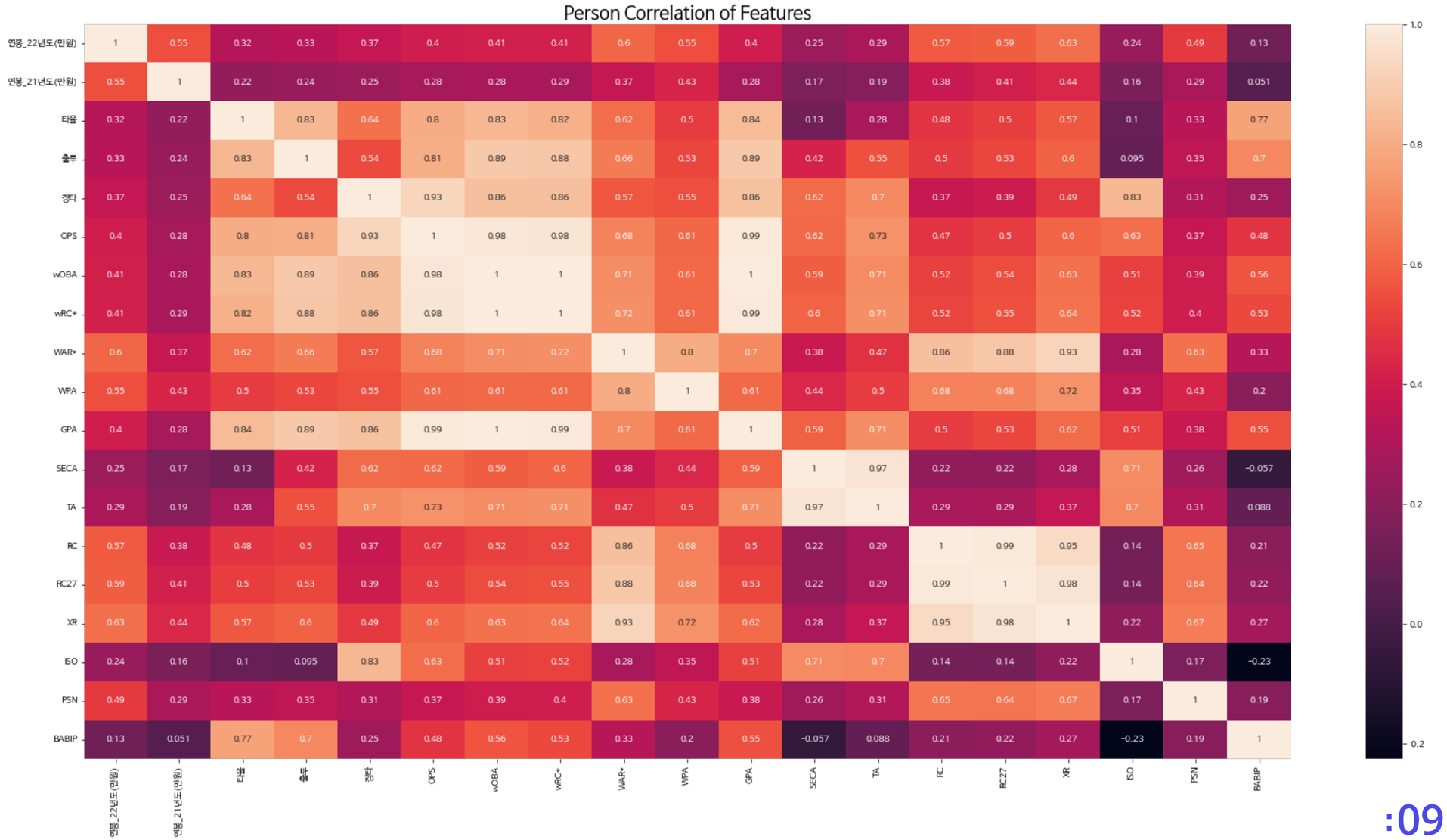
각 Feature 별 Histogram

각 feature의 분포 확인



세이버 메트릭스 지표 변수 생성

상관계수 히트맵



상관계수 〉 0.8 인 Feature

- OPS – wOBA, wRC+, GPA
- wOBA – GPA
- wRC+ – GPA
- WAR+ – XR, RC, RC27
- SECA – TA
- RC – XR, RC27
- RC27 – XR

선형회귀모델 적용 및 학습

기준모델

평균을 이용하여 기준모델 생성

MSE	3198260156.89
RMSE	56553.16
MAE	34043.35
R2	-0.002

다중선형회귀모델

검증 데이터셋 모델 평가

MSE	1949816392.70
RMSE	44156.73
MAE	26073.60
R2	0.389

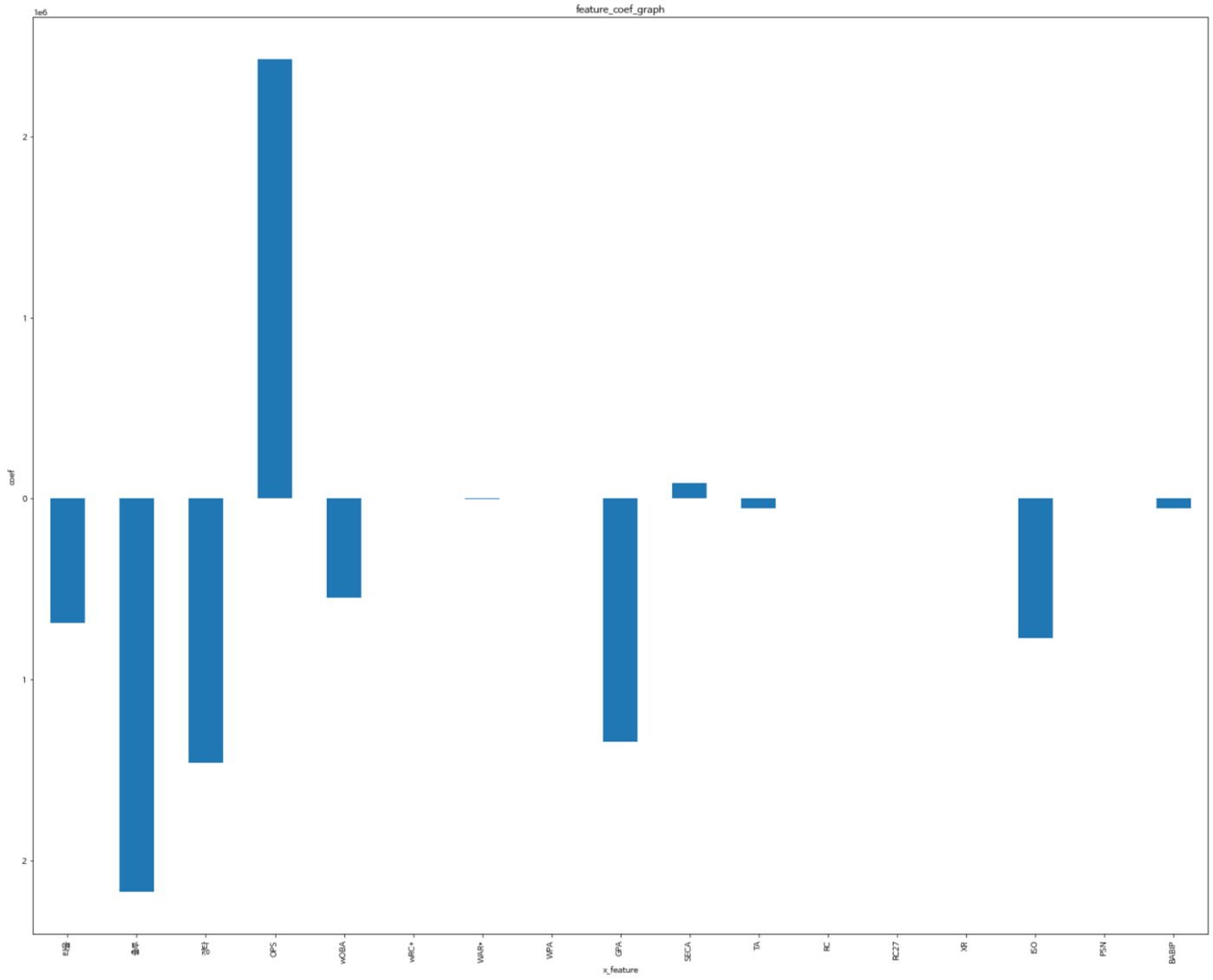
선형회귀모델 적용 및 학습

다중선형회귀모델

회귀계수 그래프

회귀계수가 양수인 Feature

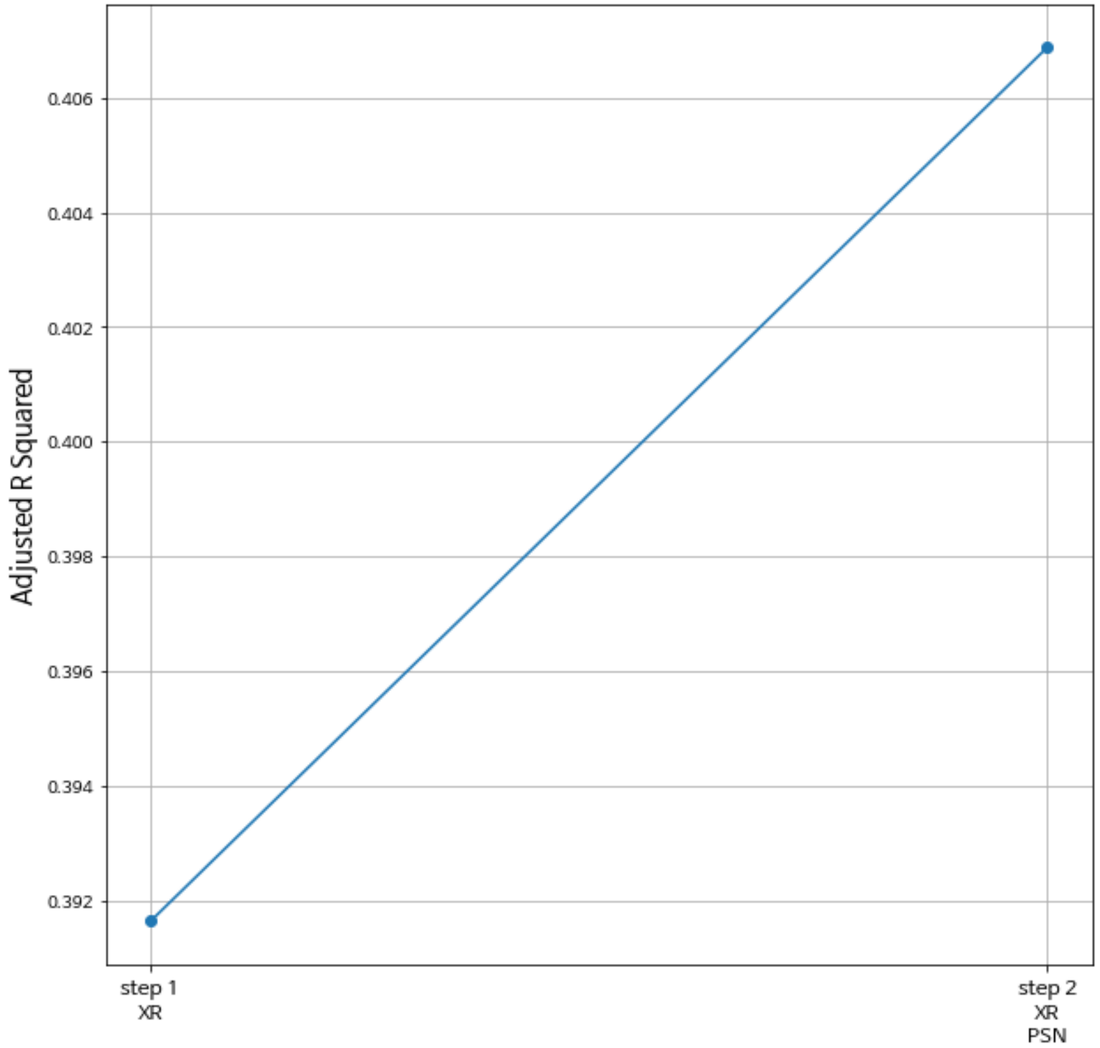
- OPS
- WAR*
- SECA



변수 선택

단계별 선택법

XR, PSN 변수 선택됨



선형회귀모델 적용 및 학습

선형회귀모델 재학습

선택된 XR, PSN을 독립변수로 하여 다시 선형회귀모델 학습

검증 데이터셋 모델 평가

MSE	1834722844.60
RMSE	42833.66
MAE	26081.98
R2	0.425

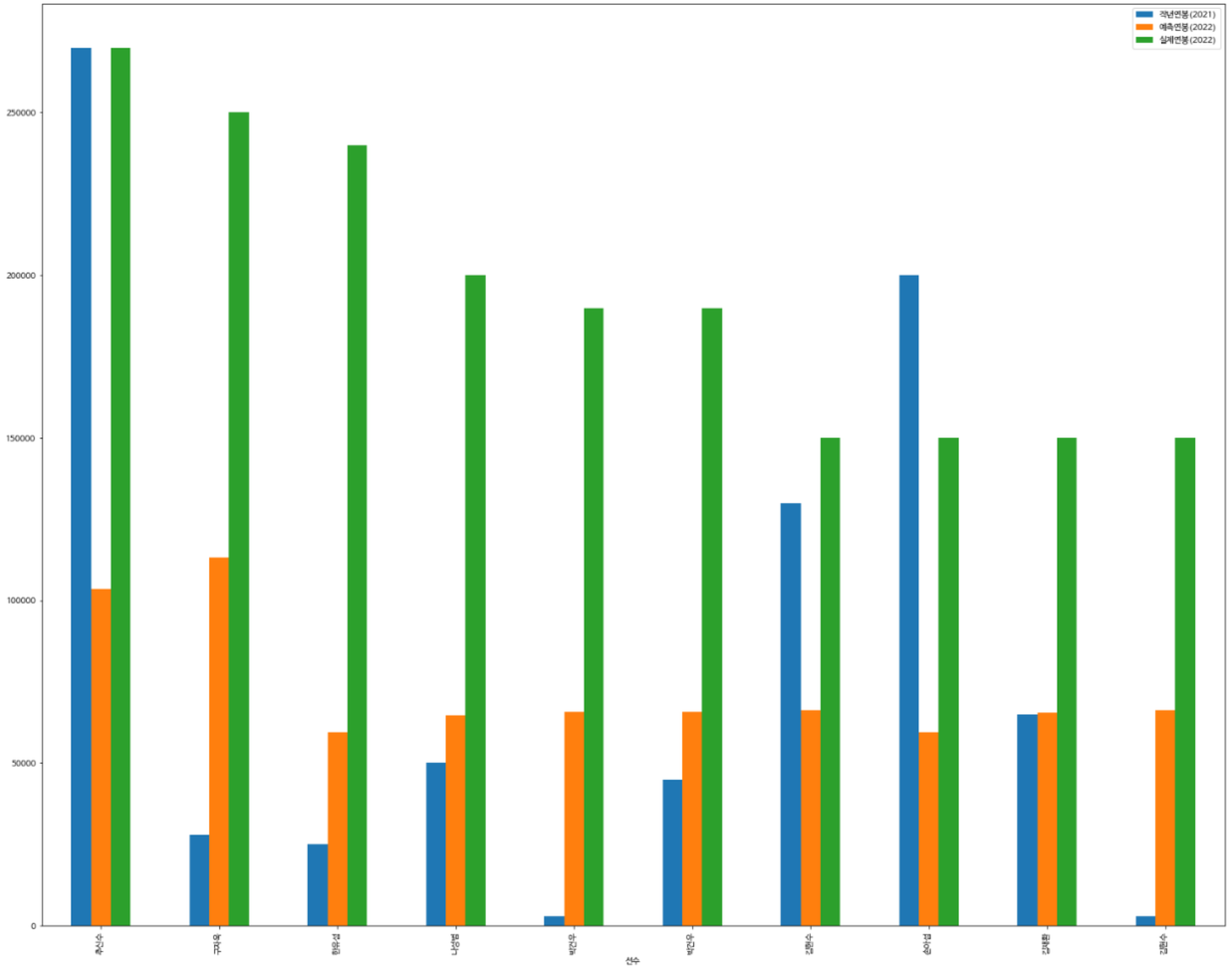
테스트 데이터셋 모델 평가

MSE	2336908880.57
RMSE	48341.59
MAE	26772.92
R2	0.292

처음 선형회귀모델
테스트 데이터셋 평가

MSE	2278913330.83
RMSE	47737.97
MAE	28756.06
R2	0.309

분석 결과



작년연봉(2021)
예측연봉(2022)
실제연봉(2022)

모델을 이용하여 예측한 연봉이
실제 연봉과 차이가 큼

> **모델 하이퍼파라미터 조절이나**
> **다른 독립변수 고려 필요**


THANK
YOU