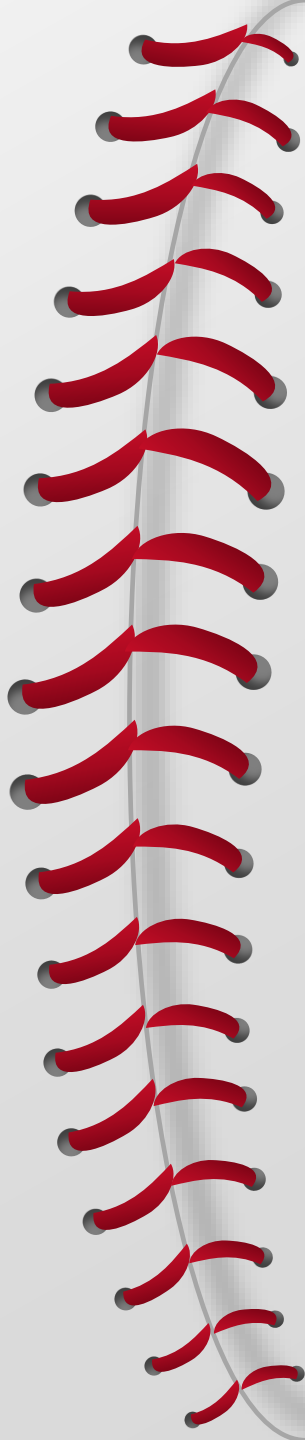
KOREAN SERIES

야구선수 연봉 예측

부제 : 전설의 투수, **故**최동원 선수가 현재 현역이라면 연봉이 얼마일까?

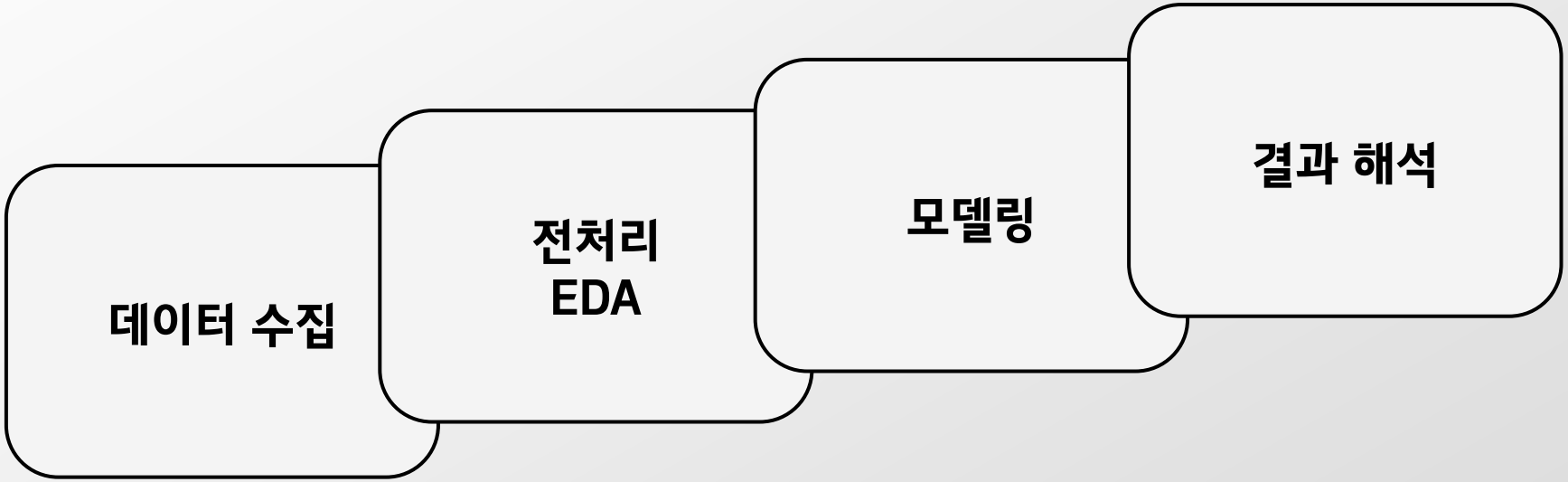


최동원, 그는 어떤 선수?

롯데의 영웅, 최동원

- 통산 평균자책점 2위
- 단일 시즌 최다 승: 2위(1984)
- 단일 시즌 최다 이닝: 2위(1984), 4위(1986)
- 단일 시즌 최다 탈삼진: 1위(1984)
- 역대 유일 5년연속 (83-87) 200이닝-14완투-2점대이하 FIP
- **역대 최초 통산 100승-1000탈삼진(1990)**
- 단일 한국시리즈 최다 승: 4승(1984)
- 단일 한국시리즈 최다 선발승: 3선발승(1984)
- 단일 한국시리즈 최다 완투승: 3완투승(1984)
- 단일 한국시리즈 최다 탈삼진: 35탈삼진(1984)
- **한국시리즈 최초 완봉승(1984)**
- 한국시리즈 최초 선발 전원 탈삼진(1984)
- **6년 연속 선발승 전경기 완투승(1983-1988)**

INDEX



데이터 수집

- BeautifulSoup & Selenium을 통한 동적 크롤링

*** 투구 기록을 기준으로 LEFT OUTER JOIN**
(N년도의 연봉을 N-1년도의 기록과 조인)

데이터 수집

- 최종 데이터프레임 형태

* 연봉(만원)

선수명	팀명	연도	ERA	G	CG	…	WHIP	연봉
최동원	롯데	1983	2.89	38	16.0	…	NaN	NaN
배경환	롯데	1983	3.18	37	4.0	…	NaN	NaN
				…				
오승환	삼성	2020	2.64	45	NaN	…	1.24	120,000

* 결측치

1. 1980년대 데이터 : 연봉, WHIP
2. 2010년대 데이터 : CG, SHO, TBF, 연봉, WAR

데이터 수집

- 최종 데이터프레임 형태

열 제거 →

행 제거 →

칼럼명	Data_Type	Unique 수	결측치 수	비고
선수명	object	725	0	-
팀명	object	18	0	-
CG	float64	21	1534	결측치 존재
SHO	float64	8	1534	결측치 존재
TBF	float64	347	1534	결측치 존재
ERA	object	655	0	'-' 존재
WPCT	object	132	0	'-' 존재
WHIP	object	219	452	'-' 및 결측치 존재
WAR	float64	846	948	결측치 존재
연봉	object	118	1169	결측치 존재
IP	object	525	0	1/3과 같은 분수 존재

전처리

Type 변환

* ERA, WPCT, WHIP : '-' 값을 가지는 행을 제거함으로써 수치형으로 변환
* 연봉 : comma(,)를 제외함으로써 수치형으로 변환 $(2{,}000 \rightarrow 2000)$
* IP : 분수를 값으로 표현 $(1/3 \rightarrow 0.3333\ldots)$

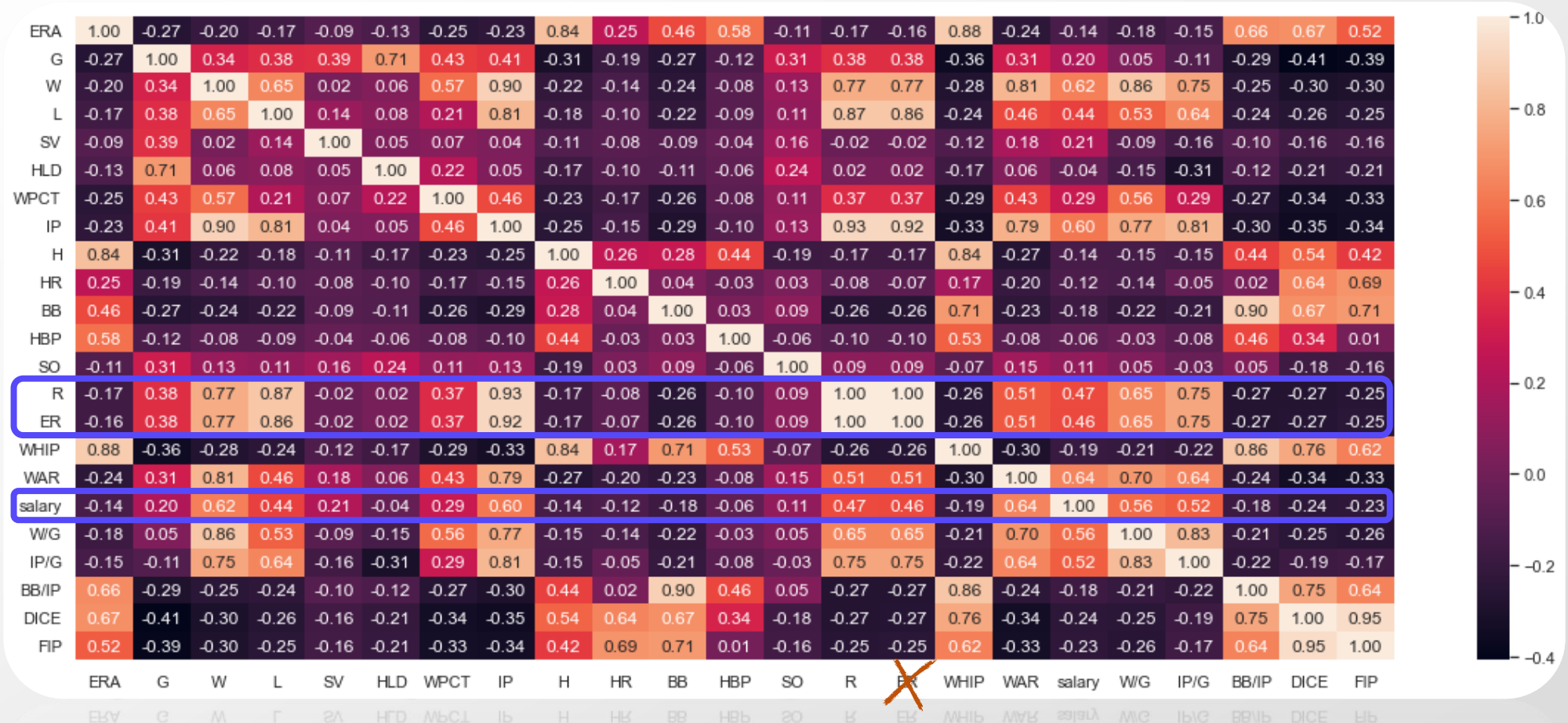
변수 가공

* WHIP : 사구 기록을 포함하여 계수를 조정 $\cdots \frac{(BB+H)}{IP} \rightarrow \frac{(BB+H+HBP)}{IP}$
* DICE : 투수가 통제하는 영역을 자책점의 형태로 표현 $\cdots 3.0 + \frac{13*(HR)+3*(BB+HBP)-2*K}{IP}$
* FIP : DICE의 개량된 형태 $\cdots 3.20 + \frac{13*(HR)+3*(BB)-2*K}{IP}$
* 경기 당 승수 : 총 경기 수로 나누어 $Normalizing \cdots \frac{W}{G}$
* 경기 당 이닝 : 총 경기 수로 나누어 $Normalizing \cdots \frac{IP}{G}$
* 이닝 당 볼넷 : 볼넷 뿐만 아니라 사구를 포함하여 $Normalizing \cdots 9 * \frac{(BB+HBP)}{IP}$
* SO, BB, HBP, HR, H : 9이닝 동안의 기록으로 환산하여 $Normalizing \cdots 9 * \frac{(Variable)}{IP}$

투수 평가지표 논문

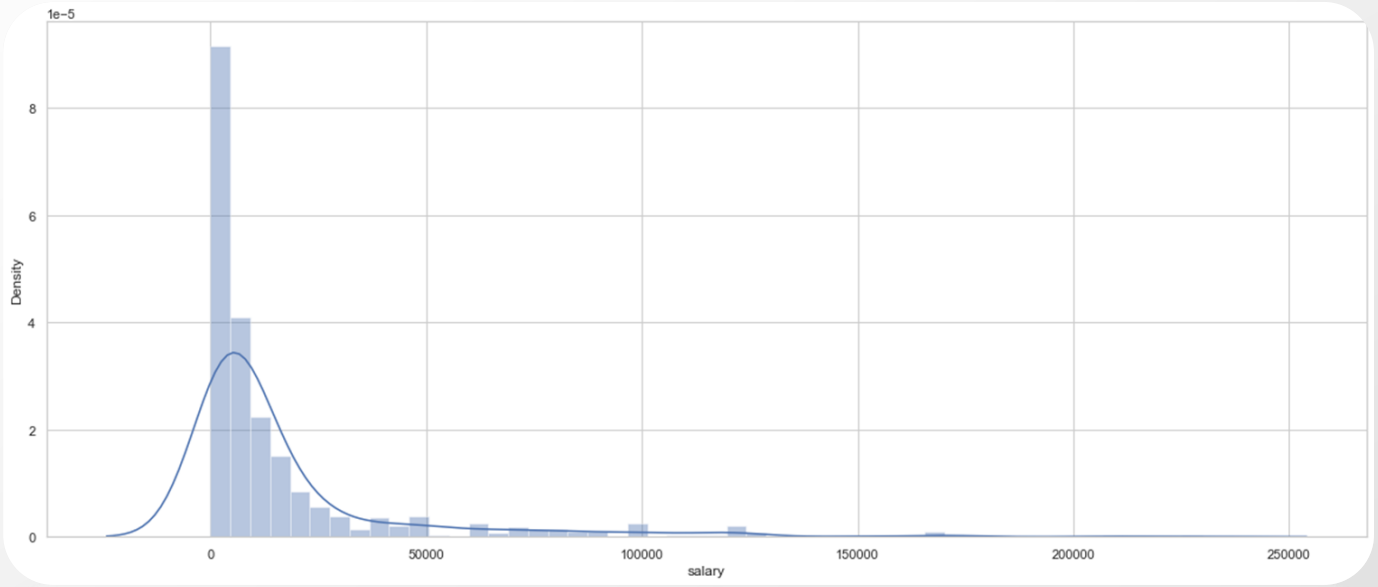
EDA

- 상관관계 히트맵


ERA 1.00 -0.27 -0.20 -0.17 -0.09 -0.13 -0.25 -0.23 0.84 0.25 0.46 0.58 -0.11 -0.17 -0.16 0.88 -0.24 -0.14 -0.18 -0.15 0.66 0.67 0.52
G -0.27 1.00 0.34 0.38 0.39 0.71 0.43 0.41 -0.31 -0.19 -0.27 -0.12 0.31 0.38 0.38 -0.36 0.31 0.20 0.05 -0.11 -0.29 -0.41 -0.39
W -0.20 0.34 1.00 0.65 0.02 0.06 0.57 0.90 -0.22 -0.14 -0.24 -0.08 0.13 0.77 0.77 -0.28 0.81 0.62 0.86 0.75 -0.25 -0.30 -0.30
L -0.17 0.38 0.65 1.00 0.14 0.08 0.21 0.81 -0.18 -0.10 -0.22 -0.09 0.11 0.87 0.86 -0.24 0.46 0.44 0.53 0.64 -0.24 -0.26 -0.25
SV -0.09 0.39 0.02 0.14 1.00 0.05 0.07 0.04 -0.11 -0.08 -0.09 -0.04 0.16 -0.02 -0.02 -0.12 0.18 0.21 -0.09 -0.16 -0.10 -0.16 -0.16
HLD -0.13 0.71 0.06 0.08 0.05 1.00 0.22 0.05 -0.17 -0.10 -0.11 -0.06 0.24 0.02 0.02 -0.17 0.06 -0.04 -0.15 -0.31 -0.12 -0.21 -0.21
WPCT -0.25 0.43 0.57 0.21 0.07 0.22 1.00 0.46 -0.23 -0.17 -0.26 -0.08 0.11 0.37 0.37 -0.29 0.43 0.29 0.56 0.29 -0.27 -0.34 -0.33
IP -0.23 0.41 0.90 0.81 0.04 0.05 0.46 1.00 -0.25 -0.15 -0.29 -0.10 0.13 0.93 0.92 -0.33 0.79 0.60 0.77 0.81 -0.30 -0.35 -0.34
H 0.84 -0.31 -0.22 -0.18 -0.11 -0.17 -0.23 -0.25 1.00 0.26 0.28 0.44 -0.19 -0.17 -0.17 0.84 -0.27 -0.14 -0.15 -0.15 0.44 0.54 0.42
HR 0.25 -0.19 -0.14 -0.10 -0.08 -0.10 -0.17 -0.15 0.26 1.00 0.04 -0.03 0.03 -0.08 -0.07 0.17 -0.20 -0.12 -0.14 -0.05 0.02 0.64 0.69
BB 0.46 -0.27 -0.24 -0.22 -0.09 -0.11 -0.26 -0.29 0.28 0.04 1.00 0.03 0.09 -0.26 -0.26 0.71 -0.23 -0.18 -0.22 -0.21 0.90 0.67 0.71
HBP 0.58 -0.12 -0.08 -0.09 -0.04 -0.06 -0.08 -0.10 0.44 -0.03 0.03 1.00 -0.06 -0.10 -0.10 0.53 -0.08 -0.06 -0.03 -0.08 0.46 0.34 0.01
SO -0.11 0.31 0.13 0.11 0.16 0.24 0.11 0.13 -0.19 0.03 0.09 -0.06 1.00 0.09 0.09 -0.07 0.15 0.11 0.05 -0.03 0.05 -0.18 -0.16
R -0.17 0.38 0.77 0.87 -0.02 0.02 0.37 0.93 -0.17 -0.08 -0.26 -0.10 0.09 1.00 1.00 -0.26 0.51 0.47 0.65 0.75 -0.27 -0.27 -0.25
ER -0.16 0.38 0.77 0.86 -0.02 0.02 0.37 0.92 -0.17 -0.07 -0.26 -0.10 0.09 1.00 1.00 -0.26 0.51 0.46 0.65 0.75 -0.27 -0.27 -0.25
WHIP 0.88 -0.36 -0.28 -0.24 -0.12 -0.17 -0.29 -0.33 0.84 0.17 0.71 0.53 -0.07 -0.26 -0.26 1.00 -0.30 -0.19 -0.21 -0.22 0.86 0.76 0.62
WAR -0.24 0.31 0.81 0.46 0.18 0.06 0.43 0.79 -0.27 -0.20 -0.23 -0.08 0.15 0.51 0.51 -0.30 1.00 0.64 0.70 0.64 -0.24 -0.34 -0.33
salary -0.14 0.20 0.62 0.44 0.21 -0.04 0.29 0.60 -0.14 -0.12 -0.18 -0.06 0.11 0.47 0.46 -0.19 0.64 1.00 0.56 0.52 -0.18 -0.24 -0.23
W/G -0.18 0.05 0.86 0.53 -0.09 -0.15 0.56 0.77 -0.15 -0.14 -0.22 -0.03 0.05 0.65 0.65 -0.21 0.70 0.56 1.00 0.83 -0.21 -0.25 -0.26
IP/G -0.15 -0.11 0.75 0.64 -0.16 -0.31 0.29 0.81 -0.15 -0.05 -0.21 -0.08 -0.03 0.75 0.75 -0.22 0.64 0.52 0.83 1.00 -0.22 -0.19 -0.17
BB/IP 0.66 -0.29 -0.25 -0.24 -0.10 -0.12 -0.27 -0.30 0.44 0.02 0.90 0.46 0.05 -0.27 -0.27 0.86 -0.24 -0.18 -0.21 -0.22 1.00 0.75 0.64
DICE 0.67 -0.41 -0.30 -0.26 -0.16 -0.21 -0.34 -0.35 0.54 0.64 0.67 0.34 -0.18 -0.27 -0.27 0.76 -0.34 -0.24 -0.25 -0.19 0.75 1.00 0.95
FIP 0.52 -0.39 -0.30 -0.25 -0.16 -0.21 -0.33 -0.34 0.42 0.69 0.71 0.01 -0.16 -0.25 -0.25 0.62 -0.33 -0.23 -0.26 -0.17 0.64 0.95 1.00
ERA G W L SV HLD WPCT IP H HR BB HBP SO R ER WHIP WAR salary W/G IP/G BB/IP DICE FIP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

EDA

- 연봉의 분포

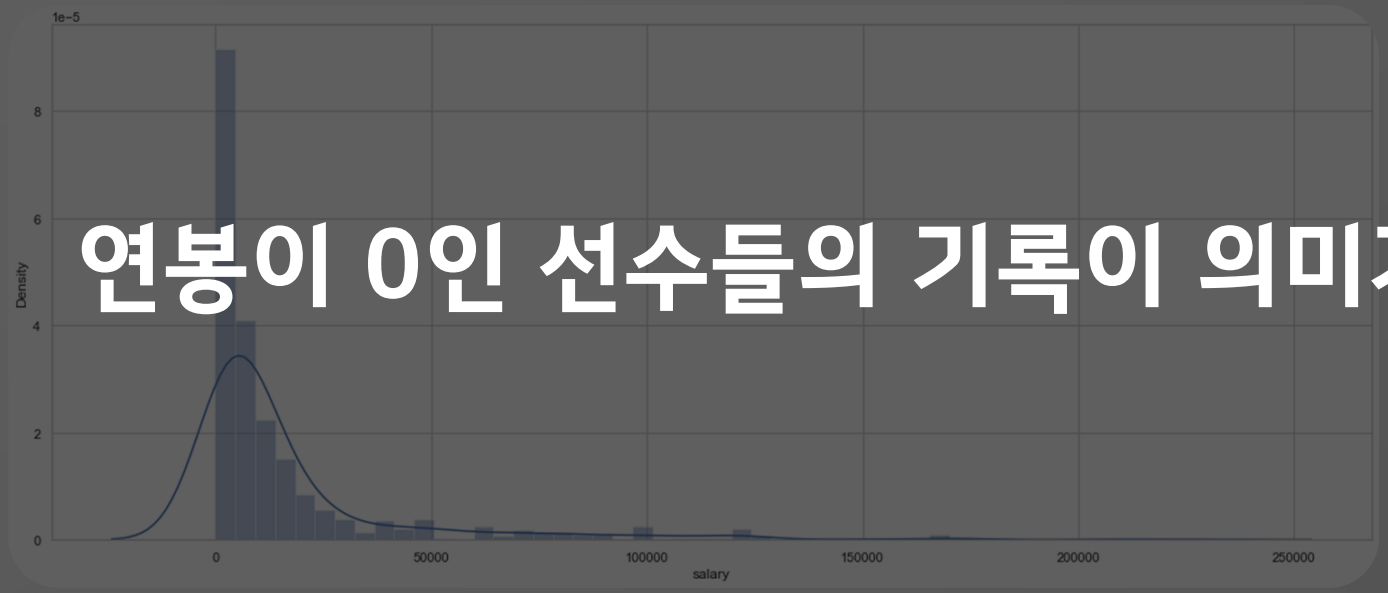


연봉	Count
23억	1
21억	1
17억	3
...	
0.27억	29
0억	67

상당히 Right-Skewed (Positive Skewness)

EDA

- 연봉의 분포



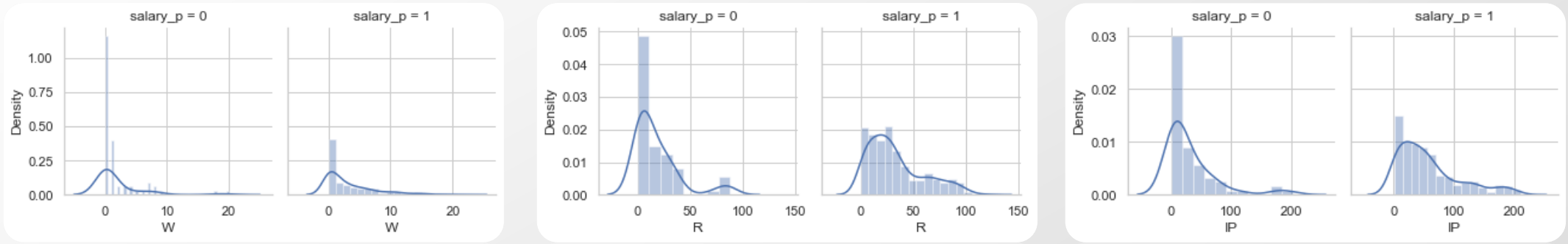
연봉이 0인 선수들의 기록이 의미가 있을까?

연봉	Count
23억	1
21억	1
1억	3
...	
0.27억	29
0억	67

상당히 Right-Skewed (Positive Skewness)

EDA

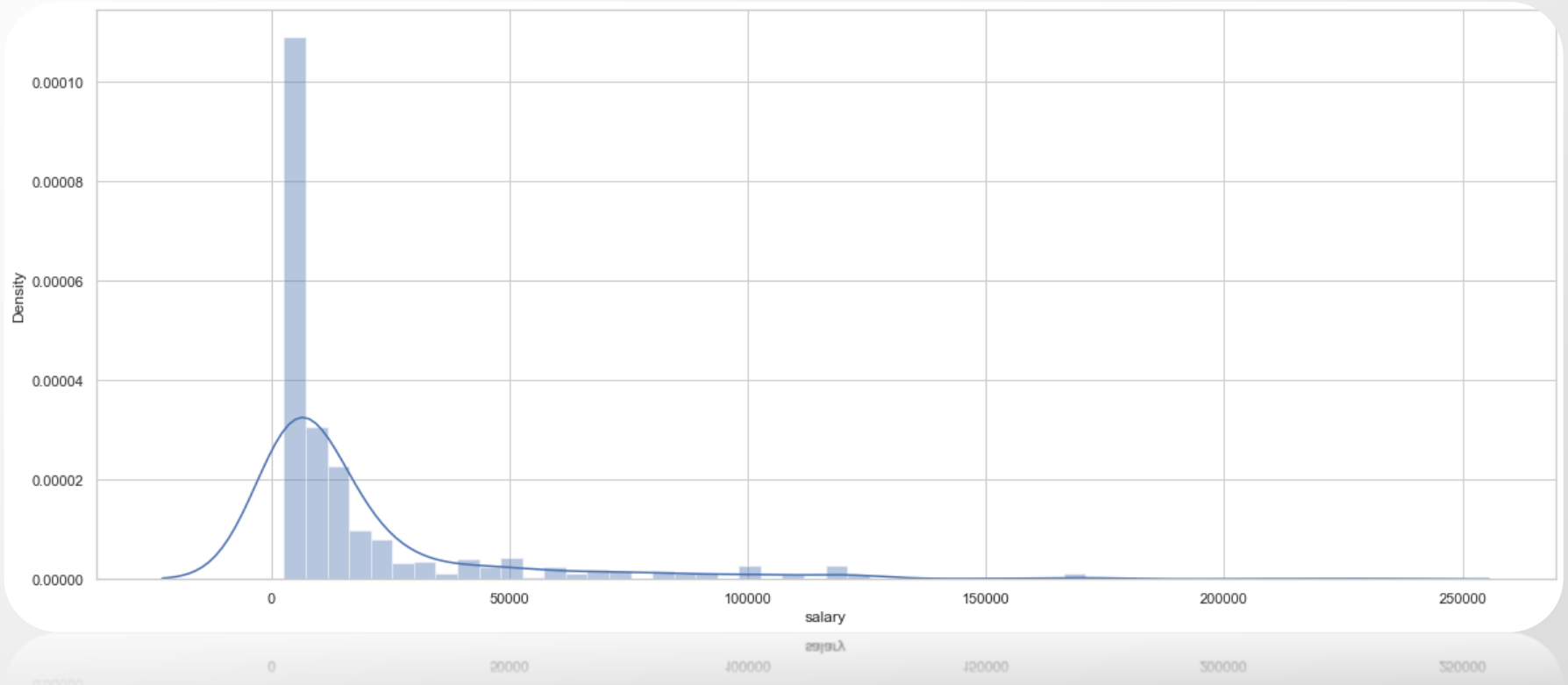
- 연봉이 0인 집단 vs 연봉이 있는 집단



연봉과 상관관계가 높은 변수들 기준, 집단 간 분포의 차이가 존재

EDA

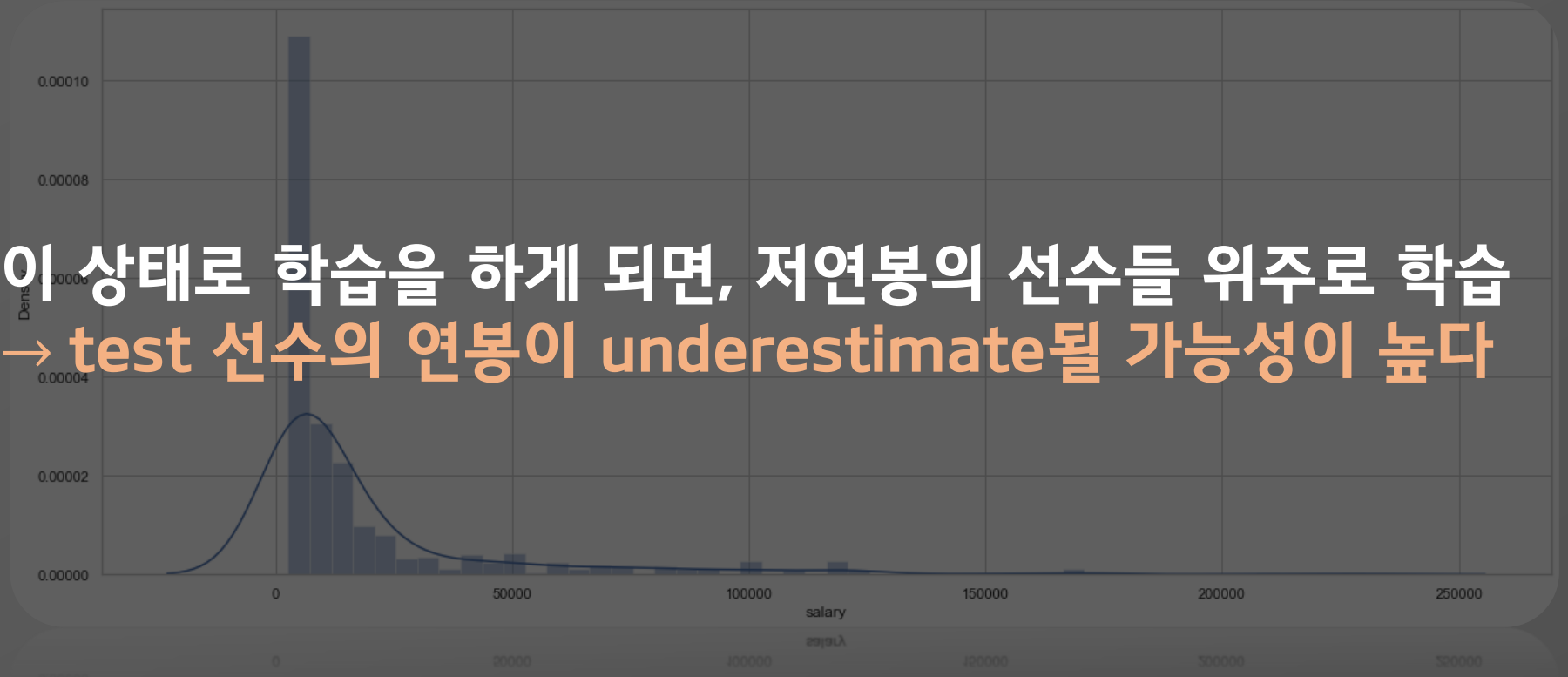
- 연봉이 0인 집단 vs 연봉이 있는 집단



연봉이 0인 선수들을 제외해도, 여전히 Right-Skewed

EDA

- 연봉이 0인 집단 vs 연봉이 있는 집단



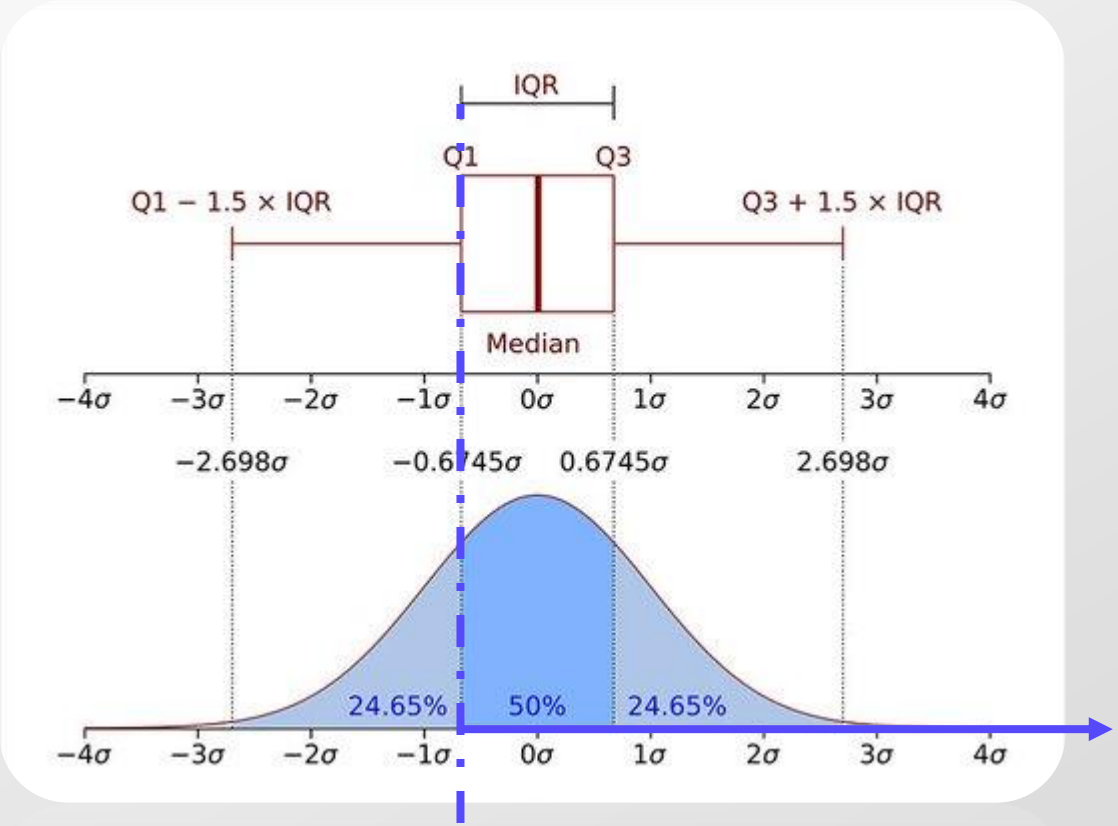
이 상태로 학습을 하게 되면, 저연봉의 선수들 위주로 학습
→ test 선수의 연봉이 underestimate될 가능성이 높다

연봉이 0인 선수들을 제외해도, 여전히 Right-Skewed

EDA

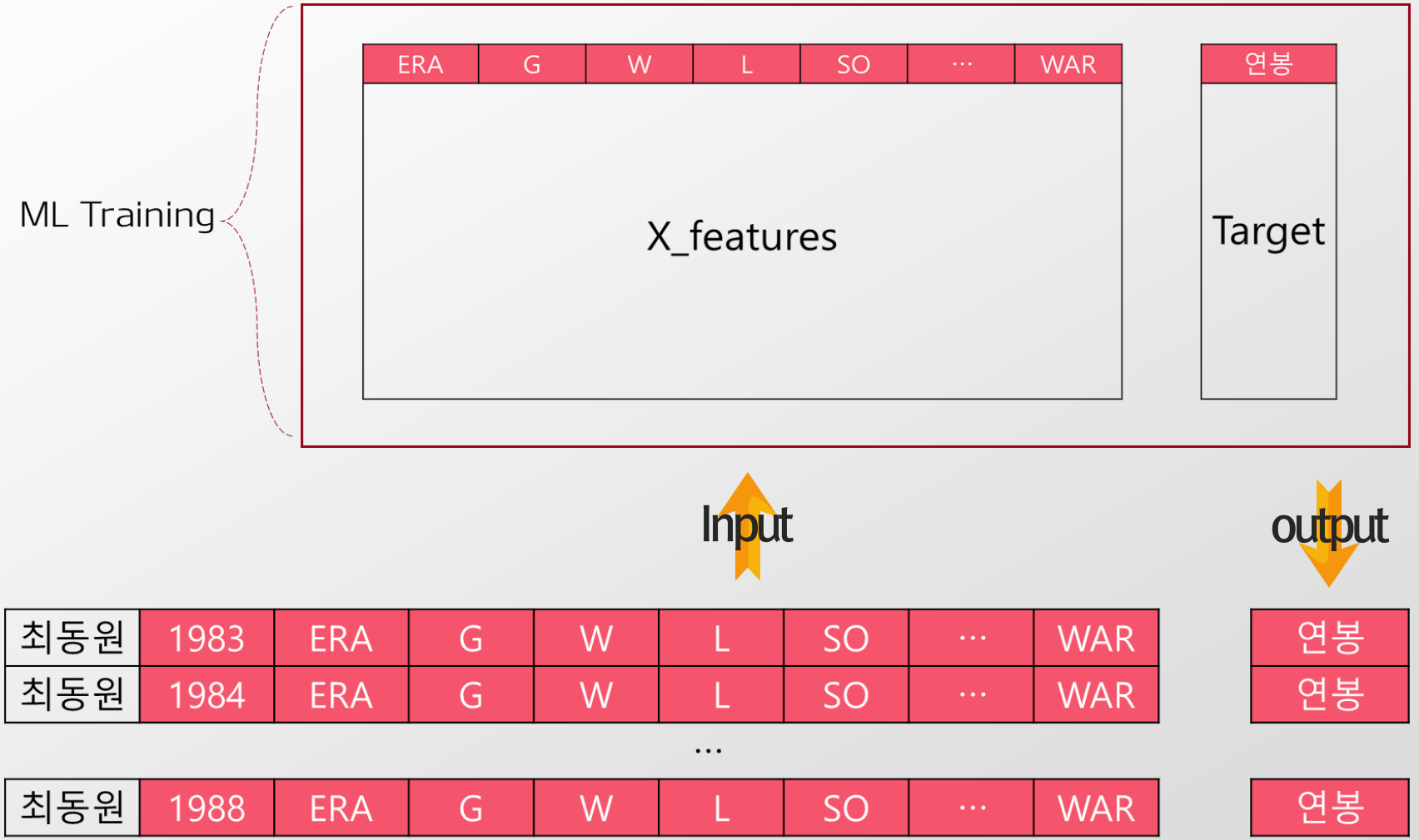
- 이상치 확인 후 제거

Q1,Q3로부터 IQR의 1.5배 거리 이상의 데이터를 이상치로 판단 → **Q1 미만의 데이터를 이상치로 판단**



모델링

- 목표 : train 된 모델에 최동원 선수의 기록을 넣어 **연봉을 예측**하자



모델링

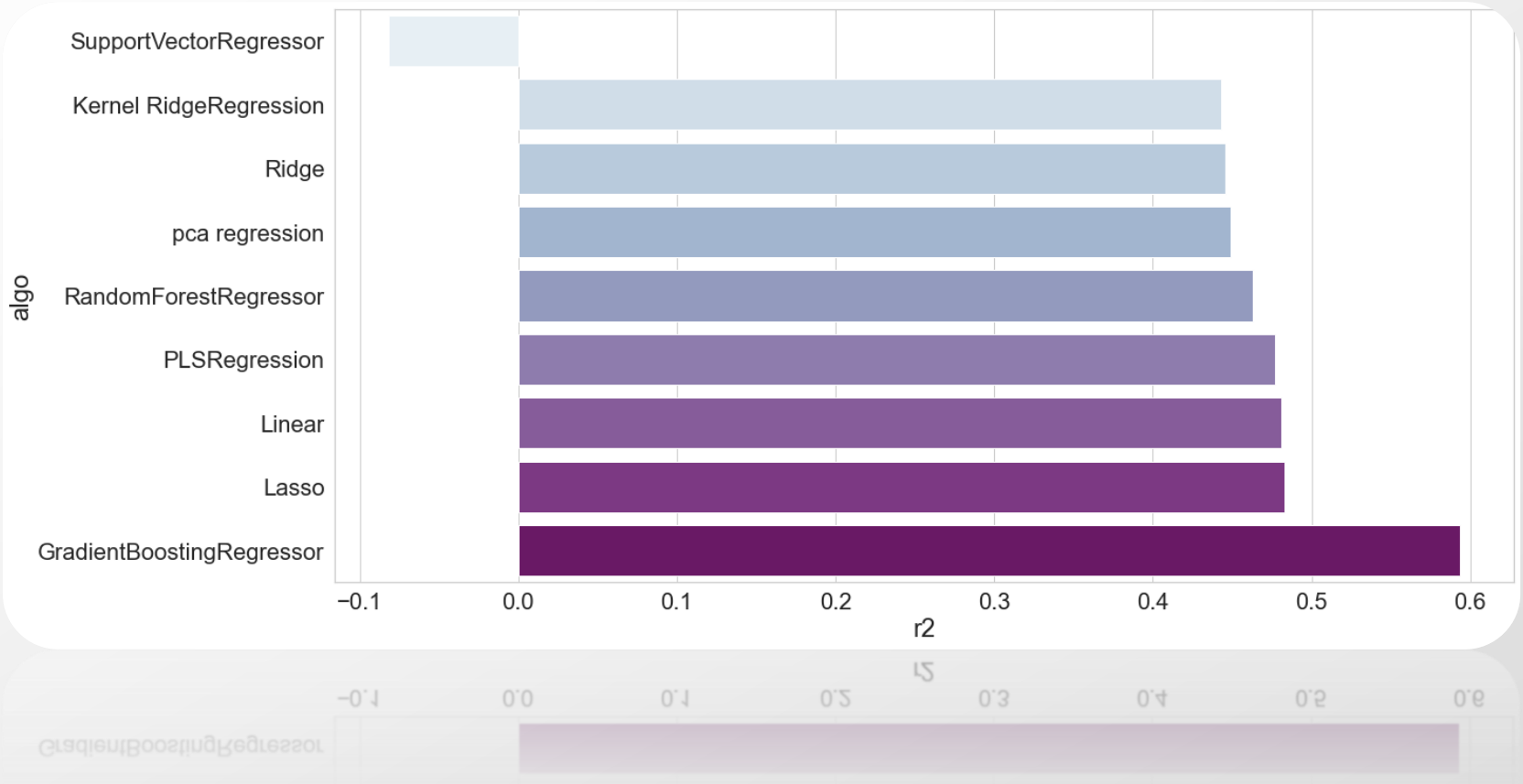
- Scaling

모델	size	x_scaler	y_scaler	rmse	mae
LinearRegression()	0.2	RobustScaler()	RobustScaler()	18770.709623	12808.984662
Lasso()	0.2	RobustScaler()	RobustScaler()	23459.535081	16848.045382
Ridge()	0.2	RobustScaler()	RobustScaler()	18107.139033	12705.545068
ElasticNet()	0.2	RobustScaler()	RobustScaler()	20076.511790	14579.143076
LinearRegression()	0.2	RobustScaler()	MinMaxScaler()	18770.709623	12808.984662
...	...	...	...	...	...
ElasticNet()	0.4	MaxAbsScaler()	StandardScaler()	32858.197387	21165.896657
LinearRegression()	0.4	MaxAbsScaler()	MaxAbsScaler()	23415.133888	14609.216050
Lasso()	0.4	MaxAbsScaler()	MaxAbsScaler()	32858.197387	21165.896657
Ridge()	0.4	MaxAbsScaler()	MaxAbsScaler()	23479.910309	14479.220468
ElasticNet()	0.4	MaxAbsScaler()	MaxAbsScaler()	32858.197387	21165.896657

	Scaling 전	Scaling 후
Linear Regression RMSE	21906	18770
Ridge Regression RMSE	21879	18107
Lasso Regression RMSE	21905	17254
ElasticNet Regression RMSE	22626	20076

모델링

- Pipeline 모델



Gradient Boosting 회귀 : $\{RMSE : 20940,\ R^2\ score : 0.5938\}$

결과 해석

- 한계

	1983	1984	1985	1986	1987	1988
salary	112154.084529	162023.601067	115814.127657	113983.735599	115191.172906	92392.572614

Params : $\{\max\_depth = 8, \max\_features = log2, \min\_samples\_split = 0.9\}$

1. 1980년대 선수들의 연봉이 존재하지 않아, 과거의 연봉 분포를 파악할 수 없는 점
2. 고연봉 선수들의 데이터가 부족하여 높은 스탯을 보유한 선수의 연봉을 예측하는 것에 한계
3. 최종 데이터셋의 행 개수가 622개로, 매우 적은 데이터 양 … underfitting 가능성 농후

* min_samples_split : 0.9 * sample 개수

감사합니다