Regression project

KBO 프로야구 연봉 예측 모델

pronounce12@gmail.com
조현상

github notion

01 프로야구 연봉 예측 데이터셋

소개 및 분석 목적

1. 목적 :

KBO 연봉예측 모델은 프로야구 선수들의 각종 기록 데이터와 연봉 데이터를 활용하여 다음 시즌 선수 개인의 연봉을 예측하는 모델입니다.

2. 프로젝트 전제 조건:

- 데이터셋은 2020년의 수집된 약 **285**명의 데이터로 구성됩니다.
- 데이터는 선수의 기록 정보와 연봉에 관한 정보를 각각의 데이터셋에 포함하고 있습니다.
- 프로젝트의 주요 관심사 및 중점 분석 대상은 "**WAR**" 컬럼으로, 이 컬럼은 선수의 지표를 종합적으로 나타내는 중요한 변수입니다. 이를 통해 선수의 기록과 연봉의 연관성 패턴 및 예측 모델을 개발하고자 합니다.
- 투수와 타자의 평가지표가 다르기에 이번 프로젝트에서는 투수를 다루었습니다.

3. 데이터 크롤링:

스탯티즈(http://www.statiz.co.kr/main.php)의 **2020선수정보**와
(http://www.statiz.co.kr/stat.phpopt=0&sopt=0&re=1&ys=2020&ye=2020&se=0&te=&tm=&ty=0&qu=auto&po=0&as=&ae=&hi=&un=&pl=&da=1&o1=WAR&o2=OutCount&de=1&lr=0&tr=&cv=&ml=1&sn=30&si=&cn=)
연봉정보 탭(http://www.statiz.co.kr/salary.php)을 크롤링 하였습니다.

4. 데이터 병합:

player_pay2020과 player_pat2021, 2020KBO data의 선수 이름(이름, player)으로 병합하였습니다.

5. 컬럼 소개

pay_2021, pay_2020

- player (선수 이름)
- year (년도)
- team (팀)
- **pay (연봉)**

2020pitcher

- 순 (순위)
- 이름 (선수 이름)
- 연도 (데이터의 연도)
- **WAR (Wins Above Replacement, 대체 선수 대비 승리 기여도)**
- 출장 (출장 횟수)
- 완투 (완투한 경기 수)
- 완봉 (완봉한 경기 수)
- 선발 (선발로 나선 경기 수)
- 승 (승리한 경기 수)
- 패 (패배한 경기 수)
- 세이브 (세이브한 경기 수)
- 홀드 (홀드한 경기 수)
- 이닝 (투구 이닝 수)
- 실점 (실점 수)
- 자책 (자책점 수)
- 타자 (상대한 타자 수)
- 안타 (허용한 안타 수)
- 2타 (허용한 2루타 수)
- 3타 (허용한 3루타 수)
- 홈런 (허용한 홈런 수)
- 볼넷 (허용한 볼넷 수)
- 고4 (고의 사구 수)
- 사구 (허용한 사구 수)
- 삼진 (기록한 삼진 수)
- 보크 (보크 수)
- 폭투 (폭투 수)
- ERA (Earned Run Average, 평균 자책점)
- FIP (Fielding Independent Pitching, 수비 무관 투수 평가지표)
- WHIP (Walks and Hits per Inning Pitched, 이닝당 출루 허용률)
- ERA+ (ERA 보정지표, 리그 평균 대비 ERA)
- FIP+ (FIP 보정지표, 리그 평균 대비 FIP)
- WAR2 (Wins Above Replacement, 대체 선수 대비 승리 기여도) - [WAR 지표와 중복]
- WPA (Win Probability Added, 경기 중 승률 변동 기여도)

02 Data Crawling

스탯티즈
스탯티즈는 대한민국 프로야구의 다양한 통계와 선수 정보를 제공하는 웹사이트입니다. 이 사이트는 선수들의 경기별, 시즌별 성적과 연봉 등 그 외에도 여러 정보를 상세하게 제공하고 있어 연봉 예측 모델의 데이터 분석을 위해 활용하였습니다.

BeautifulSoup
BeautifulSoup은 웹 크롤링을 위한 파이썬 라이브러리로, HTML과 XML 문서에서 데이터를 추출하는 데 유용합니다. 스탯티즈의 웹 페이지 구조를 분석한 후, BeautifulSoup를 활용하여 필요한 선수 데이터와 연봉 정보를 추출하였습니다.

CSV파일로 반환
추출된 데이터는 분석을 용이하게 하기 위해 CSV(Comma-Separated Values) 형식으로 저장하였습니다.

03 Data 전처리

연봉 데이터 병합 및 전처리 및 전처리

	선수	연도	팀	연봉(만원)
0	양현종	2020	KIA	230000
1	최형우	2020	KIA	150000
2	가뇽	2020	KIA	65000
3	나지완	2020	KIA	60000
4	터커	2020	KIA	55000

+

merge

	선수	연도	팀	연봉(만원)
0	최형우	2021	KIA	150000
1	브룩스	2021	KIA	100000
2	터커	2021	KIA	70000
3	나지완	2021	KIA	60000
4	다카하시	2021	KIA	60000

2020년 연봉 데이터와 2021년 연봉데이터를 선수 이름을 Key로 병합하였습니다. 은퇴선수나 군입대 선수들, 연봉 데이터가 0인 선수들은 기사와 공식 사이트를 참조하여 업데이트 하였습니다.

결측치 및 중복행 제거

결측치

isna().sum() →

```
순       4      안타      2
이름      2      2타      2
연도      2      3타      2
WAR     2      홈런      2
출장      2      볼넷      2
완투      2      고4       2
완봉      2      사구      2
선발      2      삼진      2
승       2      보크      2
패       2      폭투      2
세이브     2      ERA     2
홀드      2      FIP     2
이닝      2      WHIP    2
실점      2      ERA+    2
자책      2      FIP+    2
타자      2      WAR2    2
WPA     2
```

중복행

duplicated().sum() →

```
In [87]: pitcher_2020.duplicated().sum()
Out[87]: 1
```

연봉 데이터와 선수 데이터 병합

	순	이름	연도	WAR	출장	완투	완봉	선발	승	패	...
0	1.0	알칸타라	20두	8.31	31.0	1.0	0.0	31.0	20.0	2.0	...
1	2.0	스트레일리	20롯	7.53	31.0	0.0	0.0	31.0	15.0	4.0	...
2	3.0	브룩스	20K	7.16	23.0	1.0	1.0	23.0	11.0	4.0	...
3	4.0	루친스키	20N	5.59	30.0	0.0	0.0	30.0	19.0	5.0	...
4	5.0	요키시	20키	5.53	27.0	0.0	0.0	27.0	12.0	7.0	...
...	...	...	...	...	...	...	...	...	...	...	...
280	281.0	윤정현	20키	-0.86	15.0	0.0	0.0	4.0	0.0	1.0	...
281	282.0	장지훈	20삼	-0.98	29.0	0.0	0.0	0.0	0.0	1.0	...
282	283.0	장현식	20NK	-1.67	37.0	0.0	0.0	3.0	4.0	4.0	...
283	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
284	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...

285 rows × 33 columns

+

merge

	player	year_2020	team_2020	2020	2021
0	최형우	2020	KIA	150000	150000
1	나지완	2020	KIA	60000	60000
2	터커	2020	KIA	55000	70000
3	브룩스	2020	KIA	47900	100000
4	김선빈	2020	KIA	45000	45000
...	...	...	...	...	...
491	이강준	2020	kt	2700	2700
492	천성호	2020	kt	2700	2700
493	윤종휘	2020	kt	2700	2700
494	소형준	2020	kt	2700	2700
495	윤준혁	2020	kt	2700	2700

496 rows × 5 columns

2020년 KBO 프로야구 투수 스탯 데이터와
이전에 병합한 2020년, 2021년 연봉데이터를 선수 이름을 Key로 병합하였습니다. 은퇴선수나 군입대 선수들, 연봉 데이터가 0인 선수들은
기사와 공식 사이트를 참조하여 업데이트 하였습니다.

컬럼명 수정 및 삭제

순 이름 연도 WAR
출장 완투 완봉 선발
승 패 세이브 홀드
이닝 실점 자책 타자
안타 2타 3타
홈런 볼넷 고4 사구
삼진 보크 폭투 ERA
FIP WHIP
ERA+ FIP+ WAR2 WPA
year_2020 team_2020
2020 2021

→

선수명 팀명
승 패 세 홀드
경기 선발 이닝
삼진/9 볼넷/9 홈런/9
ERA FIP
KFIP WAR
연봉(2020)
연봉(2021)

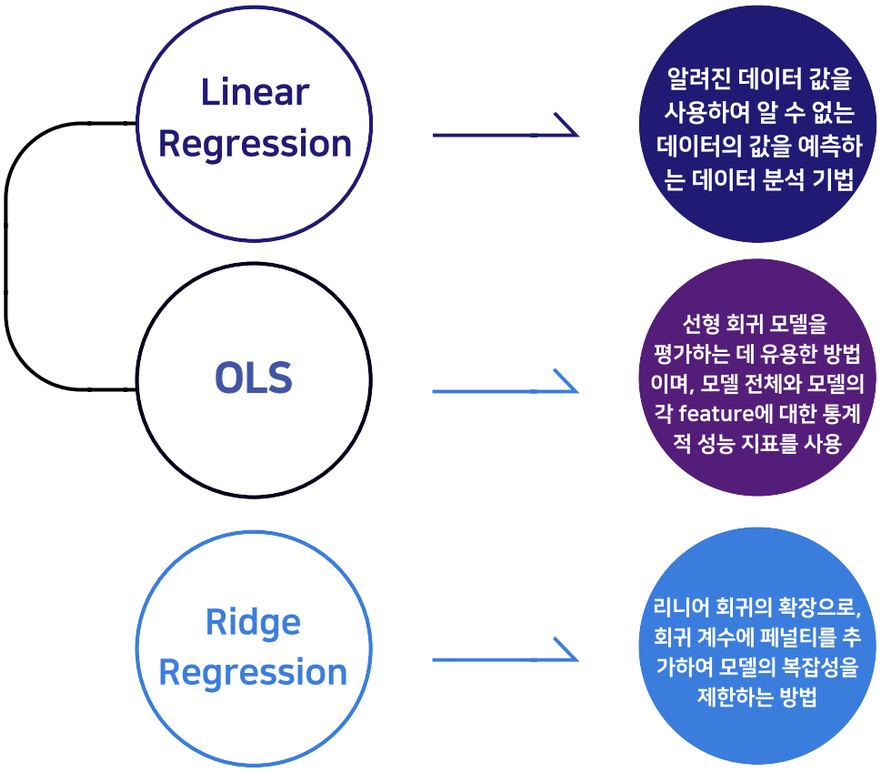
연봉을 결정하는데 비교적 중요도가 떨어지는 컬럼들을 삭제하고, 각 컬럼들의 이름을 직관적으로 파악할 수 있게 컬럼명을 수정하였습니다.

경기 당 지표로 변환

```
columns_to_divide = ['삼진/9', '볼넷/9', '홈런/9']
for col in columns_to_divide:
result_data[col] = (result_data[col] / 9).round(2)
```

비교적 이닝이 많은 선발투수와, 계투, 마무리 투수간의 삼진, 볼넷, 홈런 갯수에 대한 불균형을 최소화하기 위해 삼진, 볼넷, 피홈런에 대한 지표를 경기(9이닝)당 갯수로 변경하였습니다.

04 사용한 회귀 기법



Linear Regression

선형 회귀(Linear Regression)는 데이터의 분포를 최선으로 대표하는 **직선**(또는 고차원에서는 초평면)을 찾는 것을 목표로 합니다. 이 직선은 독립 변수와 종속 변수 간의 관계를 선형적으로 모델링하며, 주어진 데이터 포인트에 대해 **오차를 최소화**하는 방향으로 학습됩니다.

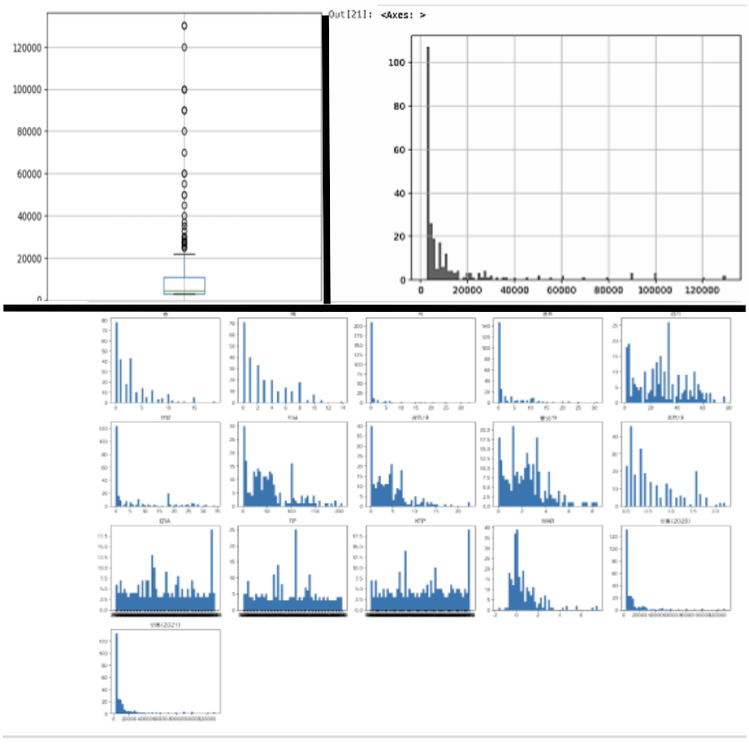
OLS(Ordinary-Least-Square)

최소제곱법(Ordinary Least Squares, OLS)은 선형 회귀 분석에서 사용되는 방법으로, 실제 값과 예측 값의 차이(잔차)의 제곱을 최소화하는 파라미터를 찾는 방법입니다. OLS는 모델의 파라미터 추정에 있어서 편향되지 않으며, **큰 표본에서는 효율적**입니다. 하지만, 데이터에 이상치가 있거나 독립 변수간의 고도의 상관성(다중공선성)이 있을 경우, 민감하게 반응할 수 있습니다.

Ridge Regression

릿지회귀는 선형 회귀의 확장 형태로, 회귀 계수에 **L2 규제**를 추가하여 모델의 복잡성을 제한합니다. 이 규제는 모델의 계수를 작게 만들어, 데이터의 노이즈나 다중공선성으로 인한 과적합을 방지하는 데 도움을 줍니다. 릿지회귀의 핵심 파라미터는 규제의 강도를 조절하는 **알파(α)** 값입니다. 알파가 0이면 릿지회귀는 일반 선형 회귀와 동일해지며, 알파가 증가할수록 모든 회귀 계수를 0에 가깝게 만들어 복잡성이 줄어듭니다

05 Histogram



이상치를 제거하지 않은 이유

01 고액 연봉자의 특수성

고액 연봉자 데이터는 프로야구 선수의 시장 가치와 그들의 특별한 스킬셋 또는 기여도를 반영합니다. 이러한 데이터는 특정 선수가 게임에 끼치는 영향력이나 팀 내에서의 중요성을 나타내며, 따라서 연봉 구조 분석에서 중요한 역할을 합니다.

02 통계적 다양성

프로야구는 선수 개개인의 다양한 스탯과 기록을 포함하는 풍부한 데이터를 제공합니다. 연봉은 이러한 스탯에 기반하여 결정되므로, 스탯과 연봉 간의 상관관계를 이해하기 위해 모든 데이터 포인트가 중요합니다.

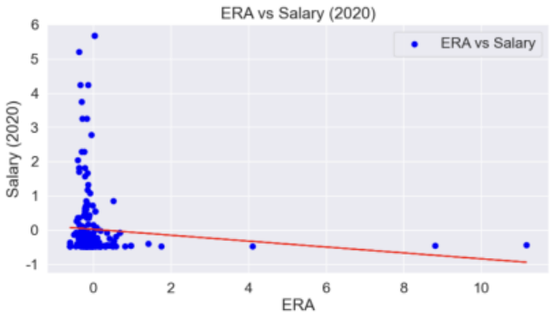
03 이상치의 중요성

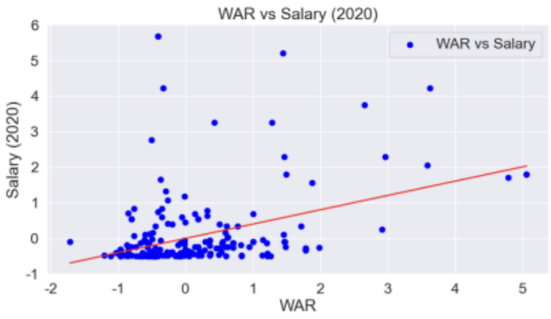
이상치는 실제 선수의 기록과 연봉 사이의 관계를 파악하는 데 중요한 역할을 합니다. 이러한 이상치는 연봉 결정에 영향을 미치는 숨겨진 변수나 예외적인 성과를 반영할 수 있으며, 분석에서 제외되어서는 안 됩니다.

04 신인과 스타 선수의 연봉 차이

프로야구 연봉 데이터는 신인 선수와 스타 선수 간의 연봉 차이를 명확히 드러내며, 이는 스포츠 산업에서의 경력과 성과에 따른 보상 구조를 이해하는 데 필수적입니다. 이상치는 이러한 차이를 더욱 명확히 하며, 분석에서 이를 제거하는 것은 전체적인 연봉 분포의 이해를 저해할 수 있습니다.

06 Scatter plot



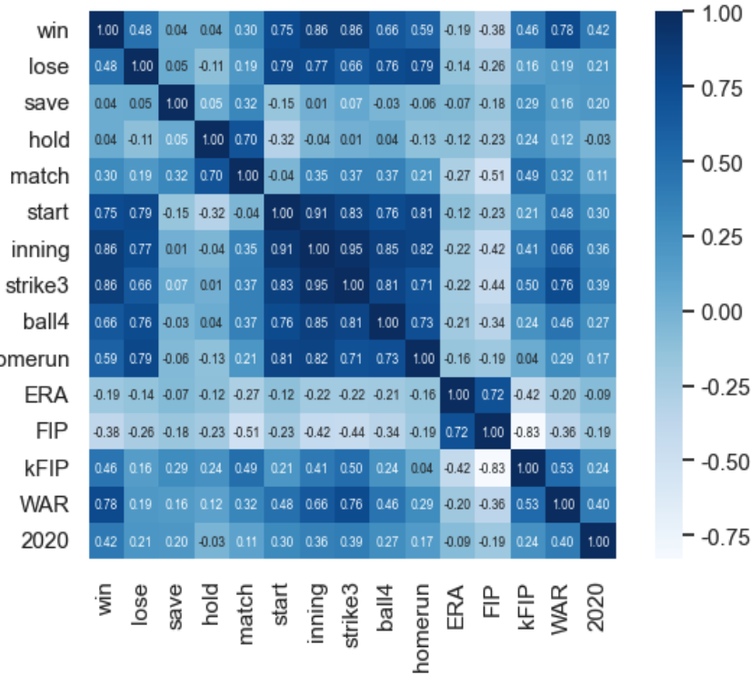


연봉과 ERA, WAR 산점도

ERA(Earned Run Average, 평균자책점)와 연봉 정보, 그리고 WAR(Wins Above Replacement, 대체 선수 대비 승리 기여도)와 연봉 정보 각각에 대한 산점도를 표현한 scatter plot 그래프입니다.

scatter plot를 통해 ERA와 연봉, 그리고 WAR와 연봉 간의 상관 관계를 파악할 수 있습니다. 일반적으로, ERA가 낮을수록(더 좋은 투수 성적) 연봉이 높아지고, WAR가 높을수록(더 많은 승리 기여) 연봉이 높아지는것을 알수 있습니다.

07 Heatmap



HeatMap 연관성 지표

win (승리) : 0.42
WAR (대체 선수 대비 승리 기여도) : 0.40
strike3 (9이닝 당 삼진) 0.39
ining (이닝) : 0.36
start (선발) : 0.30

해석

HeatMap에 따르면 거의 모든 스탯이 연봉정보와 연관이 되어있습니다. 특히 win과 WAR, strike3, ining, start는 2020(연봉)과 강한 양의 상관 관계를 가집니다. 이것은 특정 선수들이 시즌을 진행하는 동안 누적한 스탯이 투수들의 연봉과 연관될 수 있음을 뜻합니다.

08 OLS

Omnibus:	159.357	**Durbin-Watson:**	1.897
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	25830.728
Skew:	2.120	**Prob(JB):**	0.00
Kurtosis:	59.813	**Cond. No.**	2.55e+16

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.722
Model:	OLS	**Adj. R-squared:**	0.682
Method:	Least Squares	**F-statistic:**	17.96
Date:	Tue, 07 Nov 2023	**Prob (F-statistic):**	1.47e-34
Time:	13:11:24	**Log-Likelihood:**	-2074.5
No. Observations:	191	**AIC:**	4199.
Df Residuals:	166	**BIC:**	4280.
Df Model:	24		
Covariance Type:	nonrobust		

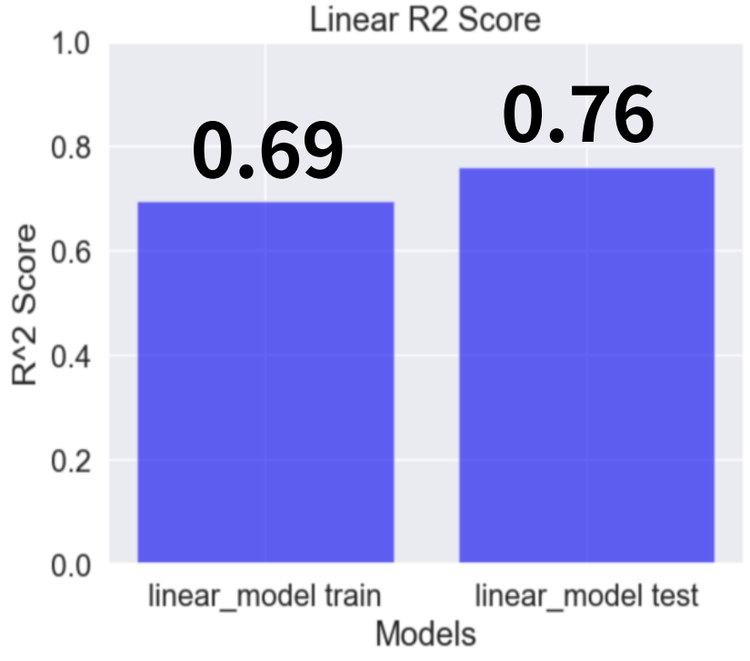
해석

R-squared (0.722): 독립 변수가 종속 변수의 변동성을 얼마나 잘 설명하는지 나타내는 값입니다. 프로야구 연봉예측 모델의 경우, 종속 변수의 약 72.2%가 모델에 의해 설명이 가능합니다.

조정된 R-squared (0.682): R-squared보다 약간 낮은데, 이는 예측 변수의 수를 고려하여 조정된 것이므로 정상이라고 판단했습니다.

F-통계량 (17.96): F-통계량은 모델의 전체적인 유의성을 검사하는 데 사용됩니다. 관련된 Prob(F-statistic) 값이 매우 낮다는 것은 모델이 유의하다는 것을 의미합니다.

09 Linear - R2 Score



R2 스코어

Train R2 Score = 0.6952114931039778

Test R2 Score= 0.7618121847729287

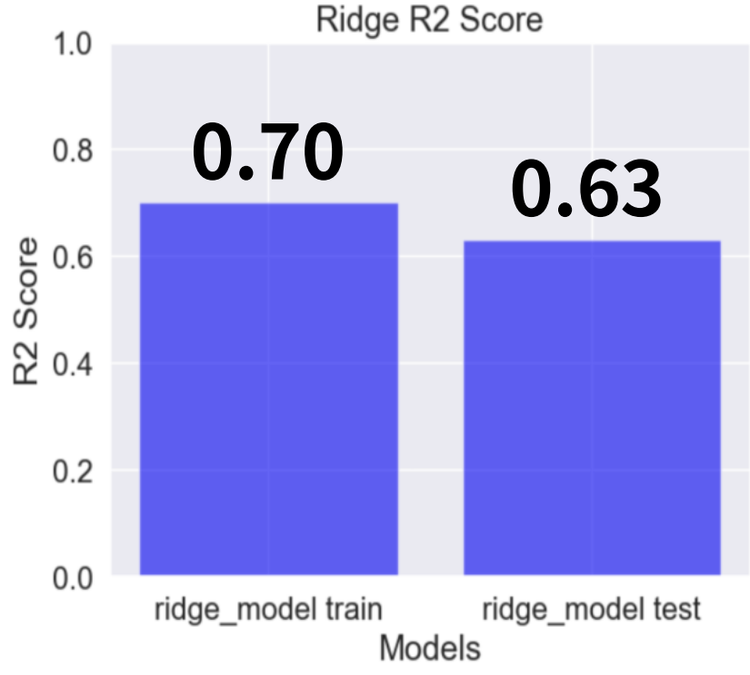
해석

프로야구 선수들의 연봉을 예측하기 위해 구축된 회귀 모델은 훈련 데이터에 대해 약 0.695의 스코어를 달성했습니다. 이는 모델이 훈련 데이터에 대해서 어느정도의 예측을 수행해내지만 모든 변동성을 잡아내지는 못한다는것을 뜻합니다.

한편, 테스트 데이터에 대해서는 0.762의 스코어를 기록했습니다. 테스트 데이터에 대한 스코어가 더 높다는 것은 모델이 과적합(overfitting)되지 않고, 새로운 데이터에 대해서도 일관되게 잘 일반화되고 있음을 뜻합니다.

이러한 결과는 회귀 모델이 선수들의 직전 해 스탯을 기반으로 그들의 연봉을 예측하는 데 있어서 비교적 정확하고 신뢰할 수 있는 성능을 보여준다고 해석할 수 있습니다. 하지만 모델의 설명력이 완벽하지 않기 때문에, 연봉 예측에 있어서 선수들의 나이나 fa상태 등의 변수를 고려해봐야합니다

10 Ridge - R2 Score



R2 스코어

Train R2 Score = 0.7019951926599164

Test R2 Score= 0.6315797677317416

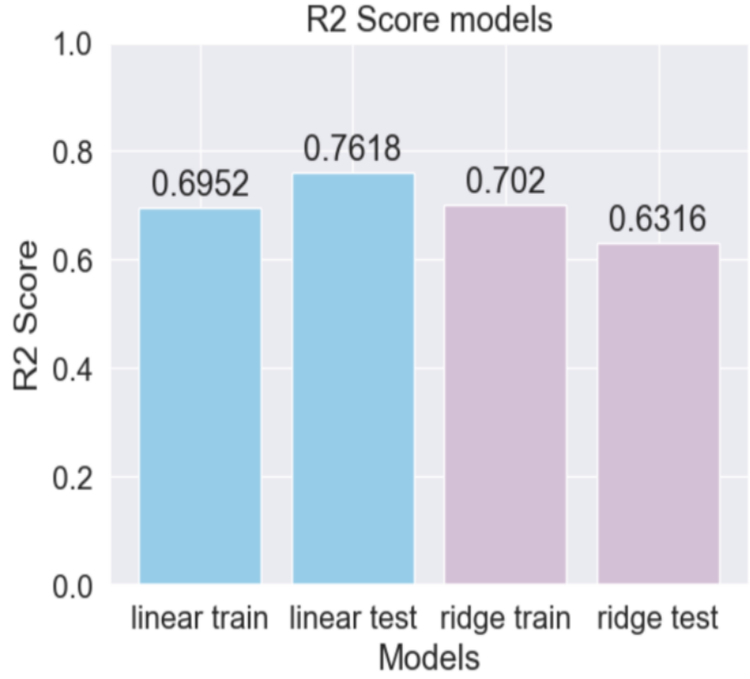
해석

릿지회귀의 R2 Score가 선형회귀의 R2스코어보다 낮은 이유에 대해 확인해봐야했습니다. 관련 내용을 확인해 본 결과 두가지 이유로 추려볼 수 있었습니다.

1. 정규화의 영향: 릿지 회귀는 계수에 대한 제약을 추가하여 모델의 복잡성을 줄입니다. 데이터가 작고 모델이 과적합되지 않는 경우, 릿지 회귀의 정규화는 모델 성능을 오히려 저하시킬 수 있습니다.

2. 데이터 특성: 데이터의 특성이 선형 회귀 모델에 더 적합할 수 있습니다. 예를 들어, 변수들 간의 관계가 실제로 선형이고 복잡한 모델이 필요하지 않은 경우, 정규화는 불필요하게 모델의 성능을 저하시킬 수 있습니다.

11 Total - R2 Score



결론

linear train = 0.6952

linear test = 0.7618

ridge train = 0.702

ridge test = 0.6316

선형 회귀 모델이 릿지 회귀 모델보다 우수한 예측 성능을 보이고 있습니다.

릿지 회귀는 규제를 통해 모델의 복잡도를 줄이고 과적합을 방지하는 장점이 있지만, 연봉예측 모델에서는 규제가 모델의 성능을 저하시킨 것으로 판단했습니다. 데이터가 과적합의 위험이 적거나, 모델의 복잡성이 문제가 되지 않는 상황이라면, 규제가 없는 선형 회귀가 더 적절할 수 있다는 것을 배웠습니다.

따라서, 테스트 데이터에 대한 성능을 중시하고, 데이터가 과적합의 위험이 적은 상황에서는 선형 회귀 모델을 선택하는 것이 바람직하다는 것을 알수있었습니다.

12 느낀점

선형회귀와 OLS

프로젝트를 시작하기 전, 선형회귀와 OLS의 관계에 대해 오해하고 있었습니다. 제가 원래 이해하고 있었던 OLS는 선형회귀, 릿지회귀, 라쏘회귀와 같이 OLS라는 회귀가 있다고 이해를 하고 있었습니다.
새롭게 이해한 OLS는 선형회귀는 데이터 간의 선형 관계를 모델링하는데, OLS는 이 때 사용되는 주요 방법입니다. OLS는 데이터 포인트와 회귀선 사이의 거리(잔차)를 최소화하여 최적의 선형 관계를 찾습니다. 이번 프로젝트를 통해 선형회귀 모델이 어떻게 작동하는지 잘 이해할 수 있게 되었습니다.

릿지회귀와 선형회귀의 성능

프로야구 선수 연봉 데이터 분석을 통해 릿지회귀와 선형회귀의 성능 차이를 경험했습니다. 데이터가 복잡하지 않은 경우, 릿지회귀의 규제가 오히려 모델 성능을 저하시킬 수 있음을 발견했습니다. 이를 통해 모든 상황에 일률적으로 적용할 수 있는 회귀 방법은 없으며, 데이터의 특성에 따라 적절한 모델을 선택하는 것이 중요하다는 것을 깨달았습니다.

감사합니다