Hitter's "Contract Year Boost Effect" & Predicted Contract size

이주원, 김이주, 정재훈, 이환욱

YSAL
YONSEI SPORTS ANALYTICS LAB

Table Of Contents

Introduction – Research Background

YSAL
YONSEI SPORTS ANALYTICS LAB

→ 타자들의 FA 직전 기록은 전보다 좋아질 것인가

→FA 계약 직전 시즌에 이전보다 월등히 좋은 성적을 기록하는 선수들이 있음

- FA 직전에 좋은 성적을 기록할 경우, 계약 규모에 긍정적 영향

→ 자신의 가치 향상을 위해 더 좋은 기록을 낼 수도, 부담으로 인해 안 좋아질 수도, 연관 없을 수도

Introduction – Research Background

YSAL
YONSEI SPORTS ANALYTICS LAB

→ FA 직전 시즌과 타자의 기록이 유의미한 연관성을 보이는지

- 타자의 통산 성적과 fa 직전 성적의 차이

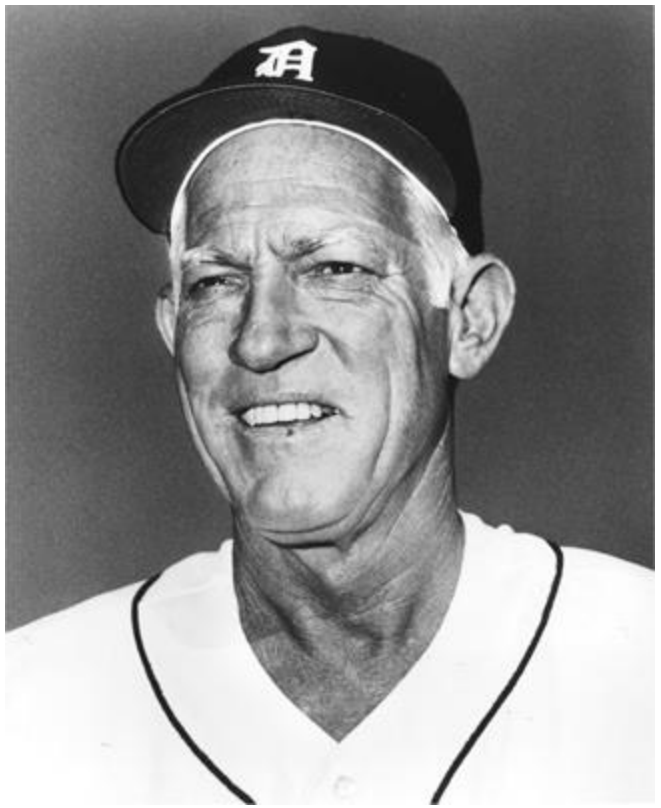
→ 타자의 계약 직전 기록 및 통산 기록에 따른 FA 계약 규모 예측

- 적정 연봉 제시 모델

Introduction – Work Cited

Do Hitters Boost Their Performance During Their Contract Years?

written by Heather M. O'Neill



Sparkey Anderson :

FA를 앞둔 선수의 기록은 분명 좋아진다

Ex. Raul Ibanez

2005 OPS .792

2006 OPS .869 (Contract Year)

2 Years contract

2007 OPS .831

2008 OPS .837 (Contract Year)

3 Years contract

Introduction – Work Cited

YSAL
YONSEI SPORTS ANALYTICS LAB

연구 시 고려 사항

1. 투수와 타자의 경우 분리 – 타자만 대상으로 함

- 투수의 경우, 개인 기록과 팀의 상황을 분리하기 어려움

Ex. 구종선택, 로케이션 등은 팀의 수비 및 상황과 관련 있으며 불펜의 상황도 고려

2. 6년 이상 활약한 선수의 4년간 성적을 기반으로 한 FA 계약을 다룸

- 어느 정도 데이터가 쌓인 선수의 기록을 바탕으로 연구해야 예측율이 높아짐

Introduction – Work Cited

OPS+	등급	수준
175	매우훌륭	MVP
150	훌륭	베스트나인 이상
125	좋음	올스타
100	평균	주전
75	나쁨	벤치멤버 이하

선행 연구

- OLS 회귀 분석을 통해 나이, 경력, 팀의 성적 등으로 선수의 특징 제한

1. 계약 직전 연도에 타자의 ops가 유의미하게 증가 (+ 4.2% ~ 5.5%)
2. 은퇴 직전 선수의 경우 ops 감소 (- 11.2% ~ 13.2%)
3. 구단주의 입장에서 선수의 경기력 일관성보다 최근 경기력 중시

종속변수로 ops100 선정

타자의 종합적인 타격 성적 반영 + 리그, 홈구장 영향 설명에 OPS보다 유리

Process – Work Cited

YSAL
YONSEI SPORTS ANALYTICS LAB

$$OPS100_{i,t} = \beta_0 + \beta_1^{(+)} * GAMES_{i,t} + \beta_2^{(+)} * PLAYOFF_{i,t} + \beta_3^{(+)} * PROBRET_{i,t} + \beta_4^{(-)} * CONTRACTYR_{i,t}$$

$$+ a_i + u_{i,t},$$

시즌 t의 선수 i의 ops100 회귀모형

Probret : 은퇴확률 추정

Contractyr: : 시즌 t가 계약연도인지(=1) 여부(=0)

각 B 계수 위의 부호 : 독립 변수의 증가가 ops100에 미치는 예상 영향

오차 1. ai : 관찰 불가능한 모든 요인 (타고난 능력, 직업 윤리, 추진력 등)

오차 2. μi,t : 사고, 날씨 등으로 인한 오차

Process – Work Cited

YSAL
YONSEI SPORTS ANALYTICS LAB

Table 2. Descriptive Statistics for Contract Year versus Non-Contract Year

		CONTRACT YEAR					NON-CONTRACT YEAR			
	N	MEAN	ST. DEV.	MINIMUM	MAXIMUM	N	MEAN	ST. DEV.	MINIMUM	MAXIMUM
OPS100	546	85.9	30.41	-21	182	470	97.2	29.87	-39	192
NOPLAY	546	0.231	0.42	0	1	470	0.032	0.18	0	1
AGE	546	33.59	3.28	26	48	470	32.25	3.02	24	47
DL	546	19.39	33.74	0	163	470	17.53	31.99	0	193
EXP	546	11.6	3.29	7	26	470	10.6	3	6	25
PLAYOFF	546	0.333	0.47	0	1	470	0.309	0.46	0	1
GAMES	546	95.23	40.57	7	162	470	115.49	36.88	10	162

계약 연도 상태별로 기술 통계량 분류

플레이오프와 DL을 제외한 모든 변수의 평균 차이 $p < .001$

계약 연도의 평균 ops100은 비계약 연도의 97.2 보다 낮은 85.9

(FA 직전 시즌의 기록이 그렇지 않을 때보다 상회할 것이라는 이론과는 반대)

Process – Work cited

YSAL
YONSEI SPORTS ANALYTICS LAB

OLS 추정의 한계

오차항 a의 측정 불가능한 선수 특성이 일부 독립변수와 상관관계가 있을 때 발생하는,
생략된 변수 편의 존재

FE 추정 (‘고정 효과’)

실제 관측치 - 선수의 시간에 따른 각 변수의 평균
편향이 해결되지만, 추정된 계수에 대한 통계적 유의성 손상 가능성

Process – Work cited

$$PROBRET_{i,t} = \alpha_0 + \alpha_1 * EXP_{i,t} + \alpha_2 * EXP2_{i,t} + \alpha_3 * DL_{i,t} + \alpha_4 * OPS100_{i,t} + a_i + v_t + u_{i,t,}$$

시즌 t에서 선수 i의 은퇴 확률에 대한 회귀 방정식

선수의 은퇴에 영향을 주는 변수 : DL에 등재된 기간, 타격 성적, 경력 등

OPS 100이 1 증가하면 은퇴 가능성 0.4% 감소

활약이 클수록 은퇴하지 않고 계속 출전할 것임을 의미

Process – Work cited

YSAL
YONSEI SPORTS ANALYTICS LAB

Conclusion

FE 추정 사용으로 선수의 행동 변화 설명 가능

OLS 추정 시 발생하는 편향 줄일 수 있음

도출된 결론

1. FA 직전 연도의 타자의 조정 OPS는 비계약연도 기간보다 6.7% 증가할 것으로 예상됨
2. 은퇴 선수는 계약 종료 직전, 성적이 떨어짐.

Process – Data Preprocessing

YSAL
YONSEI SPORTS ANALYTICS LAB

```
mlb_final_probret = mlb_final.copy()
mlb_final_probret['played_year_squared'] = mlb_final_probret['played_year'] ** 2
mlb_final_probret
```

	player	cont_year	cont_length	dollars	season	teamID	POS	G	OPS+	injury	played_year	final_year	ps	cont_year_d	played_year_squared
0	A.J. Ellis	2016	1	2500000	2016	PHI	C	64	63.0	0	8	0	0	1	64
1	A.J. Ellis	2016	1	2500000	2017	MIA	C	51	82.0	0	9	0	0	0	81
2	A.J. Ellis	2016	1	2500000	2018	SDN	C	66	104.0	0	10	1	0	0	100
3	AJ Pollock	2018	5	60000000	2016	ARI	OF	12	85.0	146	4	0	0	0	16
4	AJ Pollock	2018	5	60000000	2017	ARI	OF	112	100.0	51	5	0	1	0	25
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2148	Yoshi Tsutsugo	2021	1	4000000	2021	PIT	1B	81	88.0	31	1	0	0	1	1
1786	Zack Cozart	2017	3	38000000	2016	CIN	SS	121	92.0	7	5	0	0	0	25
1787	Zack Cozart	2017	3	38000000	2017	CIN	SS	122	140.0	21	6	0	0	1	36
1788	Zack Cozart	2017	3	38000000	2018	LAA	3B	58	81.0	109	7	0	0	0	49
1789	Zack Cozart	2017	3	38000000	2019	LAA	3B	38	-12.0	140	8	1	0	0	64

1790 rows × 15 columns

칼럼에서 제시한 은퇴확률 회귀식 활용

EXP, EXP2 : 선수의 경력, EXP2는 EXP의 제곱

EXP2 를 독립변수로 활용하기 위해 'played_year_squared' 열 추가

Process – Data Preprocessing

```
[ ]  # 특성 스케일링
     scaler = StandardScaler()
     X_scaled = scaler.fit_transform(X)

     # 주성분 분석 (PCA)을 이용한 차원 축소
     pca = PCA(n_components=2)
     X_pca = pca.fit_transform(X_scaled)

     # 다중 로지스틱 회귀 모델 훈련
     model_skl = LogisticRegression()
     result_skl = model_skl.fit(X_pca, y)
```

```
[ ]  # 결정 경계 시각화를 위한 함수 정의
     def plot_decision_boundary(X, y, model):
         # 산점도 그리기
         plt.scatter(X[:, 0], X[:, 1], c=y, cmap='coolwarm', marker='o', edgecolors='k')

         # 결정 경계 그리기
         x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
         y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
         xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.02), np.arange(y_min, y_max, 0.02))
         Z = model.predict(np.c_[xx.ravel(), yy.ravel()])
         Z = Z.reshape(xx.shape)
         plt.contourf(xx, yy, Z, alpha=0.4, cmap='coolwarm')
         plt.xlim(xx.min(), xx.max())
         plt.ylim(yy.min(), yy.max())
         plt.xlabel('Principal Component 1')
         plt.ylabel('Principal Component 2')
         plt.title('Decision Boundary')

     # 결정 경계 시각화
     plot_decision_boundary(X_pca, y, model_skl)
     plt.show()
```

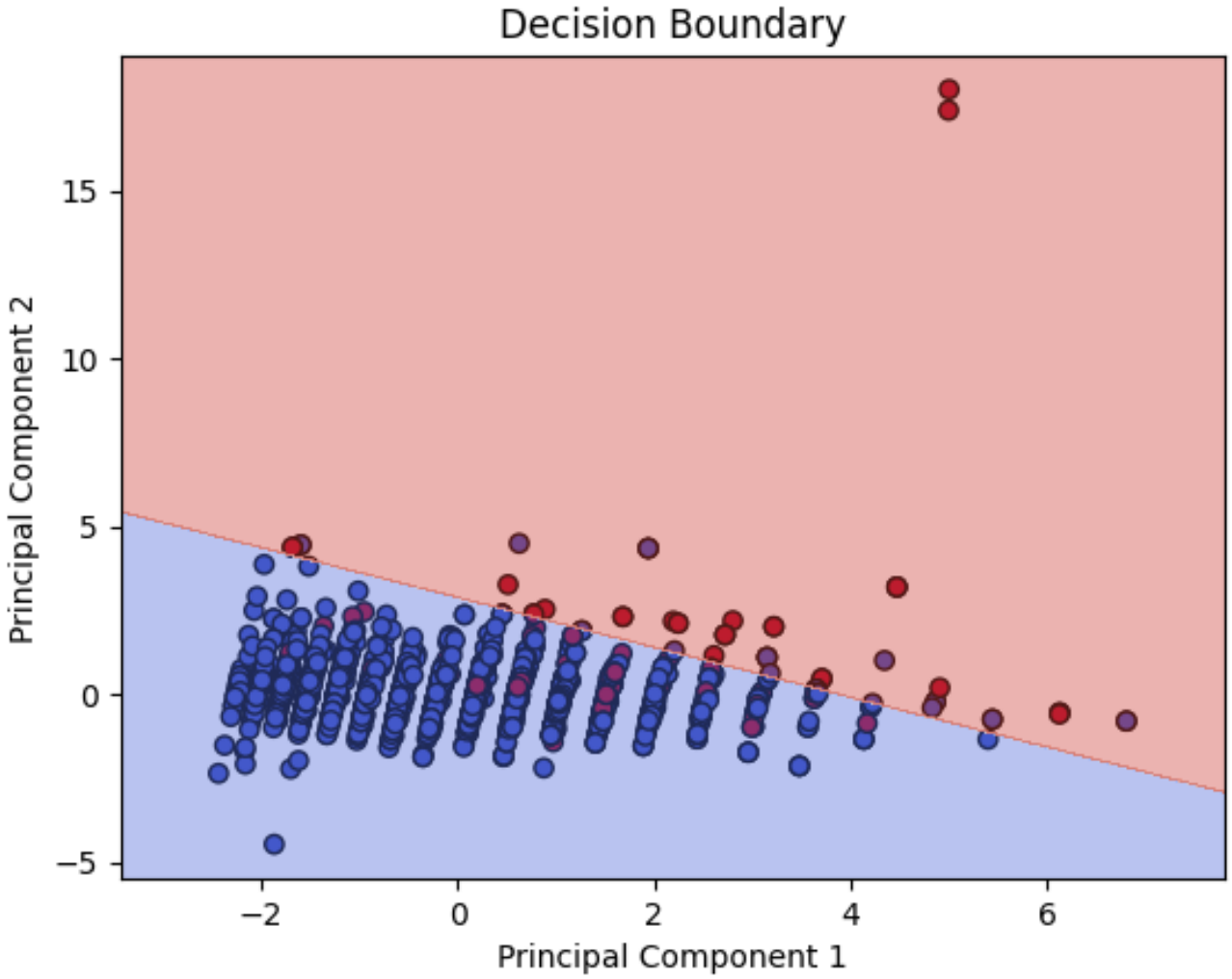
결측값 존재 시 회귀모델이 산출되지 않기 때문에
결측치 여부를 확인하고 0으로 대체

- 독립변수가 4개이므로 PCA를 통한 차원축소

- 범주형 변수가 포함되어 있기 때문에
- “로지스틱 회귀모델” 선택

독립 변수의 선형 결합을 통한 사건 발생 가능성 예측

Process



모델의 정확도가 90.8%정도로 적절하나
PCA를 통한 차원 축소 모델로는
각 변수의 영향력 확인 불가능

Process

차원축소 진행하지 않고 모든 변수의 특성을 반영한 로지스틱 회귀 모델 추정

$$PROBRET = -3.38 + .157*EXP + .006*EXP^2 + .0004*DL - .004*OPS100$$
$$(.0475)\ (.0001)\ (.135)\ (.001)$$
$$\text{Correctly Predicted} = 94\%^{23}$$

```
Optimization terminated successfully.
         Current function value: 0.215027
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:             final_year   No. Observations:                 1790
Model:                          Logit   Df Residuals:                     1785
Method:                           MLE   Df Model:                            4
Date:                Sat, 06 Apr 2024   Pseudo R-squ.:                 0.2859
Time:                        11:14:01   Log-Likelihood:                -384.90
converged:                       True   LL-Null:                       -538.99
Covariance Type:            nonrobust   LLR p-value:                1.852e-65
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.1141      0.566     -7.268      0.000      -5.224      -3.005
played_year    0.6172      0.118      5.228      0.000       0.386       0.849
played_year_squared    -0.0157      0.006     -2.726      0.006      -0.027      -0.004
injury         0.0042      0.002      1.775      0.076      -0.000       0.009
OPS+          -0.0286      0.003    -10.976      0.000      -0.034      -0.024
==============================================================================
```

논문에서 제시한 회귀모델 식과 비교해
상수항과 EXP를 제외한 독립변수들의 회귀계수가 유사함

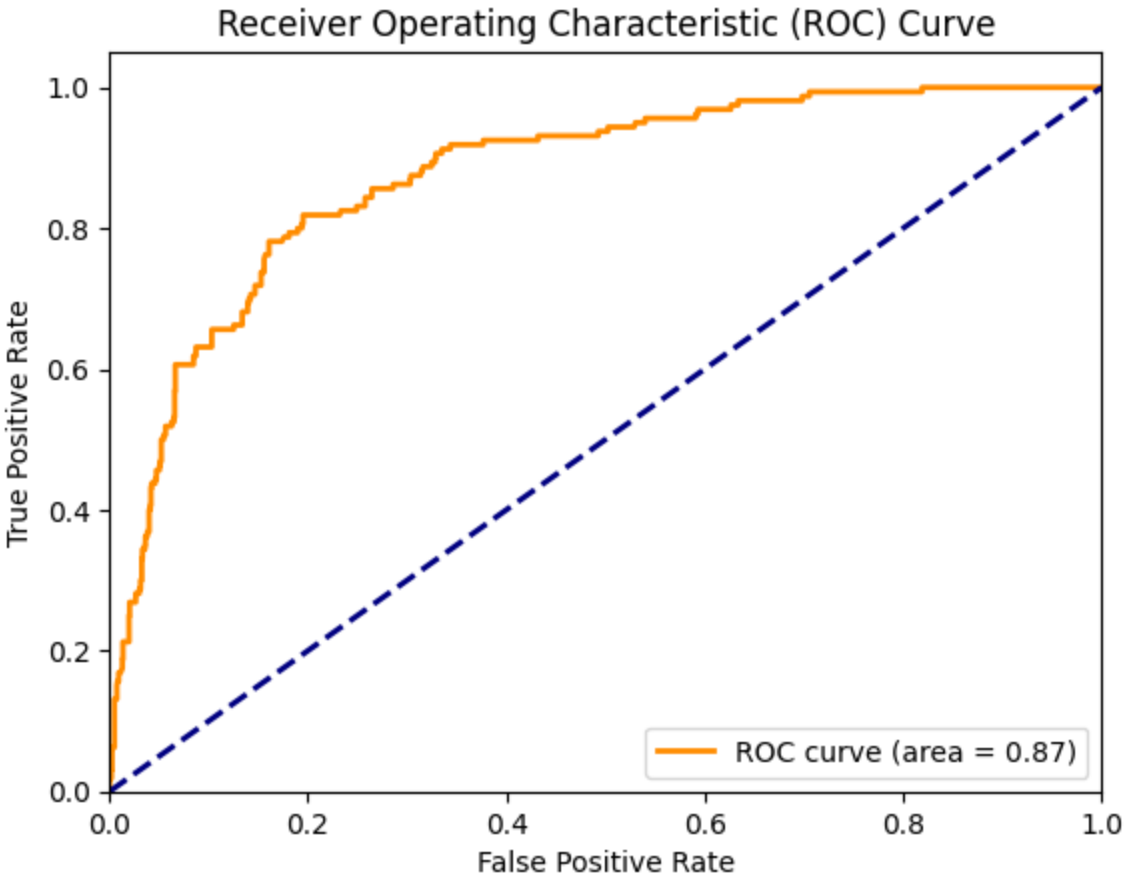
Process

선수의 경력이 은퇴확률에
가장 큰 영향을 미친다는 것을 알 수 있음

Injury도 약하지만 은퇴확률과 양의 관계를 가짐

OPS+는 은퇴확률과 음의 관계를 가짐

Process

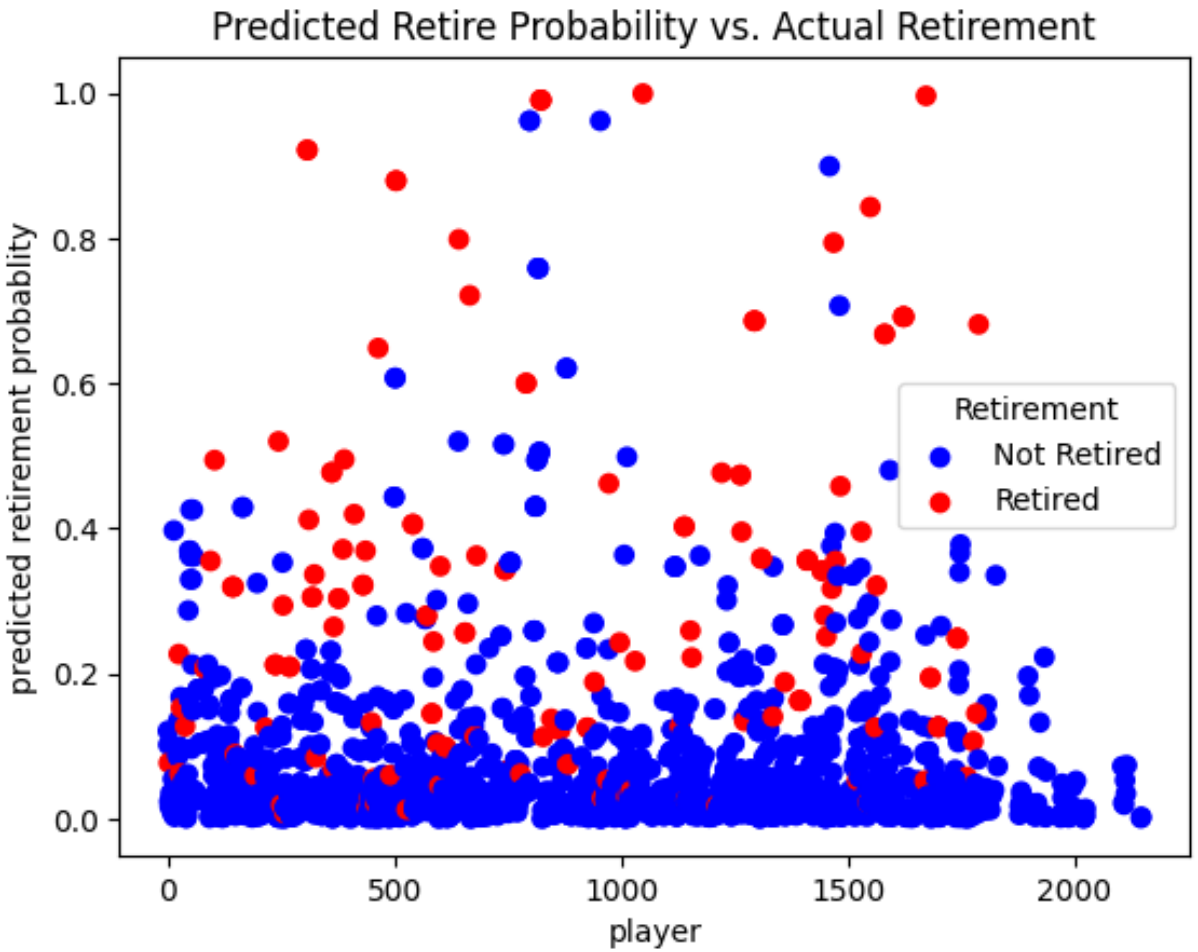


산출한 모델을 바탕으로 그려낸 ROC Curve

Area Under the Curve(AUC)의 값이 0.87

(일반적으로 AUC값이 0.7이상일 때, 모델이 수용가능함을 나타냄)

Process



Y축 : 예측한 모델에서 산출된 은퇴확률
X축 : 은퇴 여부 (red : o, blue : x)

은퇴하지 않은 선수들의 분포는 예측 은퇴 확률이 낮은 쪽에 포진

은퇴한 선수들의 분포는 예측 은퇴 확률이 높은 쪽에 포진

Accuracy: 91.5%

Process

player	season	teamID	POS	G	OPS+	injury	played_year	final_year	ps	cont_year_d	probret
Alex Gordon	2016	KCA	OF	128	85.0	34	9	0	0	0	0.10924975058220700
Alex Gordon	2017	KCA	OF	148	63.0	0	10	0	0	0	0.20976041783687100
Alex Gordon	2018	KCA	OF	141	90.0	15	11	0	0	0	0.15139272089194400
Alex Gordon	2019	KCA	OF	150	95.0	0	12	0	0	1	0.15676211823889300
Alex Gordon	2020	KCA	OF	50	65.0	0	13	1	0	0	0.3544879557098950

알렉스 고든(Alex Gordon)

모델로 예측한 은퇴확률 16년 10% -> 17년 20% -> 18년 15% -> 19년 15% -> 20년 35%
은퇴 직전 해에 은퇴확률이 20%p 상승했고 그대로 은퇴함

Process

YSAL
YONSEI SPORTS ANALYTICS LAB

player	season	teamID	POS	G	OPS+	injury	played_year	final_year	ps	cont_year_d	probret
Brian Dozier	2016	MIN	2B	155	134.0	0	4	0	0	0	0.0031739448954143 70
Brian Dozier	2017	MIN	2B	152	126.0	0	5	0	1	0	0.0064074403838550 20
Brian Dozier	2018	LAN	2B	151	89.0	0	6	0	1	1	0.0280668307519492 00
Brian Dozier	2019	WAS	2B	135	98.0	0	7	0	1	0	0.0327909994518957 0
Brian Dozier	2020	NYN	2B	7	-9.0	0	8	1	0	0	0.5107295992974520

브라이언 도져(Brian Dozier)

모델로 예측한 은퇴확률 16년 0.3% -> 17년 0.6% -> 2.8% -> 3.2% -> 51%
은퇴 직전 해에 은퇴확률이 47.8%p 상승했고 그대로 은퇴함

Process – Estimating OPS+

$$OPS100_{i,t} = \underset{(+)}{\beta_0} + \underset{(+)}{\beta_1 * GAMES_{i,t}} + \underset{(+)}{\beta_2 * PLAYOFF_{i,t}} + \underset{(-)}{\beta_3 * PROBRET_{i,t}} + \beta_4 * CONTRACTYR_{i,t}$$

$$+ a_i + u_{i,t},$$

논문에서 제시한 OPS+ 회귀 추정식

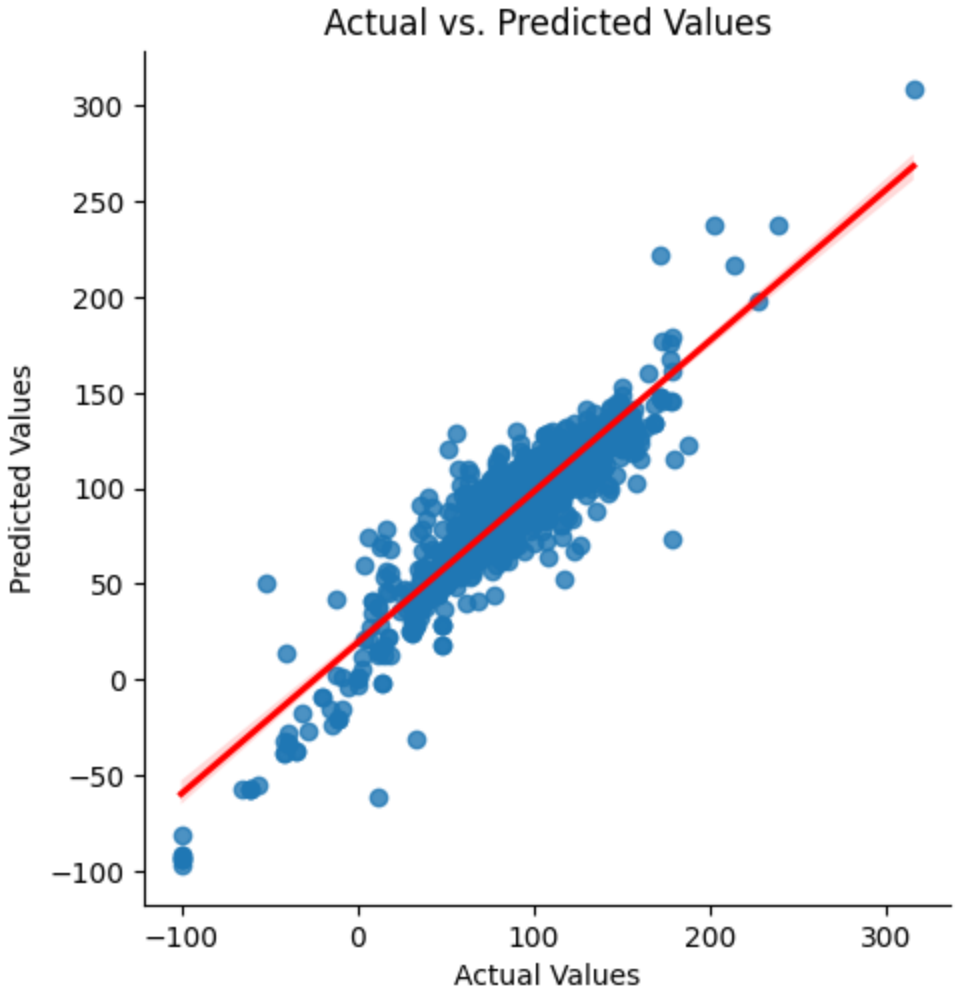
PLAYOFF, CONTRACTYR 모두 범주형 변수이므로 이진변수로 나타냄.
XGBoost를 통한 회귀 분석을 진행함.
(학습과 분류가 빠르고, 과적합이 잘 일어나지 않음)

앞서 예측한 은퇴확률과 회귀 추정식으로 OPS+ 예측

Process – Actual & predicted values

```
import xgboost as xgb
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split

# X와 y를 정의합니다.
X = mlb_final_probret[['G', 'pa', 'probret', 'cont_year_d']]
y = mlb_final_probret['OPS+']

# 데이터 분할
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# XGBoost 모델 설정
xg_reg = xgb.XGBRegressor(objective ='reg:squarederror')

# RandomizedSearch를 위한 하이퍼파라미터 분포 설정
param_dist = {
    'learning_rate': [0.05, 0.1, 0.2],
    'max_depth': [3, 5, 7],
    'n_estimators': [50, 100, 200]
}

# RandomizedSearchCV 객체 생성
random_search = RandomizedSearchCV(estimator=xg_reg, param_distributions=param_dist, n_iter=10, cv=3, scoring='r2', random_state=42)

# 랜덤 탐색 수행
random_search.fit(X_train, y_train)

# 최적의 하이퍼파라미터 조합
best_params = random_search.best_params_
print("Best Parameters:", best_params)

# 최적의 모델
best_model = random_search.best_estimator_

# 테스트 데이터로 예측
y_pred = best_model.predict(X_test)

# 평가 - R-squared
r_squared = r2_score(y_test, y_pred)
print("R-squared:", r_squared)
```

```
Best Parameters: {'n_estimators': 100, 'max_depth': 5, 'learning_rate': 0.1}
R-squared: 0.6046998611290142
```



R-Squared 값: 0.6

OPS+ 예측치의 회귀직선 & OPS+ 실제값

Process

player	cont_year	cont_length	dollars	season	teamID	POS	G	OPS+	predicted_OPS+	injury	played_year	final_year	ps	cont_year_d
Corey Seager	2021	10	325000000	2016	LAN	SS	157	134.0	137.99428	0	1	0	1	0
Corey Seager	2021	10	325000000	2017	LAN	SS	145	126.0	106.80458	0	2	0	1	0
Corey Seager	2021	10	325000000	2018	LAN	SS	26	103.0	109.010155	154	3	0	1	0
Corey Seager	2021	10	325000000	2019	LAN	SS	134	112.0	110.816414	29	4	0	1	0
Corey Seager	2021	10	325000000	2020	LAN	SS	52	150.0	146.40904	0	5	0	1	0
Corey Seager	2021	10	325000000	2021	LAN	SS	95	142.0	112.63557	76	6	0	1	1

Year	Age	Tm	Lg	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	Pos	Awards
2022 ★	28	TEX	AL	151	663	593	91	145	24	1	33	83	3	0	58	103	.245	.317	.455	.772	117	270	14	7	0	5	7	*6/D	AS
2023 ★	29	TEX	AL	119	536	477	88	156	**42**	0	33	96	2	1	49	88	.327	.390	.623	1.013	169	297	9	4	0	6	9	*6/D	AS,MVP-2,SS
2024	30	TEX	AL	7	34	29	6	11	1	0	1	5	0	0	5	4	.379	.471	.517	.988	184	15	0	0	0	0	1	*/6D	

코리 시거(Corey Seager)

FA 직전 해가 아닐 때는 OPS+ 예측 모델로 예측한 수치만큼의 OPS+를 나타냄

FA 직전 해 모델의 예측 OPS+ 약 113을 크게 상회하는 142만큼의 성적을 냄

10년 3억 2500만 달러의 대박 계약을 이끌어냄

Process

YSAL
YONSEI SPORTS ANALYTICS LAB

player	cont_year	cont_length	dollars	season	teamID	POS	G	OPS+	predicted_OPS+	injury	played_year	final_year	ps	cont_year_d
Anthony Rendon	2019	7	245000000	2016	WAS	3B	156	108.0	124.810585	0	3	0	1	0
Anthony Rendon	2019	7	245000000	2017	WAS	3B	147	139.0	116.06383	0	4	0	1	0
Anthony Rendon	2019	7	245000000	2018	WAS	3B	136	137.0	109.08457	27	5	0	0	0
Anthony Rendon	2019	7	245000000	2019	WAS	3B	146	157.0	129.63396	19	6	0	1	1
Anthony Rendon	2019	7	245000000	2020	LAA	3B	52	150.0	129.13528	0	7	0	0	0
Anthony Rendon	2019	7	245000000	2021	LAA	3B	58	94.0	86.429436	140	8	0	0	0

앤서니 렌던(Anthony Rendon)

FA 직전 해에 예측 OPS+ 129보다 크게 상회하는 OPS+ 157의 성적을 냄

워싱턴 시절에도 꾸준히 예측치보다 높은 성적을 기록.

Process – Actual & predicted values

YSAL
YONSEI SPORTS ANALYTICS LAB

	player	cont_year	cont_length	dollars	season	teamID	POS	G	OPS+	injury	played_year	final_year	ps	cont_year_d	played_year_squared	probret	predicted_OPS+
0	A.J. Ellis	2016	1	2500000	2016	PHI	C	64	63.0	0.0	8	0	0	1	64	0.121142	67.760887
1	A.J. Ellis	2016	1	2500000	2017	MIA	C	51	82.0	0.0	9	0	0	0	81	0.102063	77.813988
2	A.J. Ellis	2016	1	2500000	2018	SDN	C	66	104.0	0.0	10	1	0	0	100	0.076948	89.066940
3	AJ Pollock	2018	5	60000000	2016	ARI	OF	12	85.0	146.0	4	0	0	0	16	0.023888	62.481506
4	AJ Pollock	2018	5	60000000	2017	ARI	OF	112	100.0	51.0	5	0	1	0	25	0.016855	108.651848
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2148	Yoshi Tsutsugo	2021	1	4000000	2021	PIT	1B	81	88.0	31.0	1	0	0	1	1	0.002730	98.257301
1786	Zack Cozart	2017	3	38000000	2016	CIN	SS	121	92.0	7.0	5	0	0	0	25	0.017574	93.610413
1787	Zack Cozart	2017	3	38000000	2017	CIN	SS	122	140.0	21.0	6	0	0	1	36	0.007439	122.035622
1788	Zack Cozart	2017	3	38000000	2018	LAA	3B	58	81.0	109.0	7	0	0	0	49	0.081861	84.200096
1789	Zack Cozart	2017	3	38000000	2019	LAA	3B	38	-12.0	140.0	8	1	0	0	64	0.681219	2.445260

Df에 probret, 예상 OPS+ 열 추가

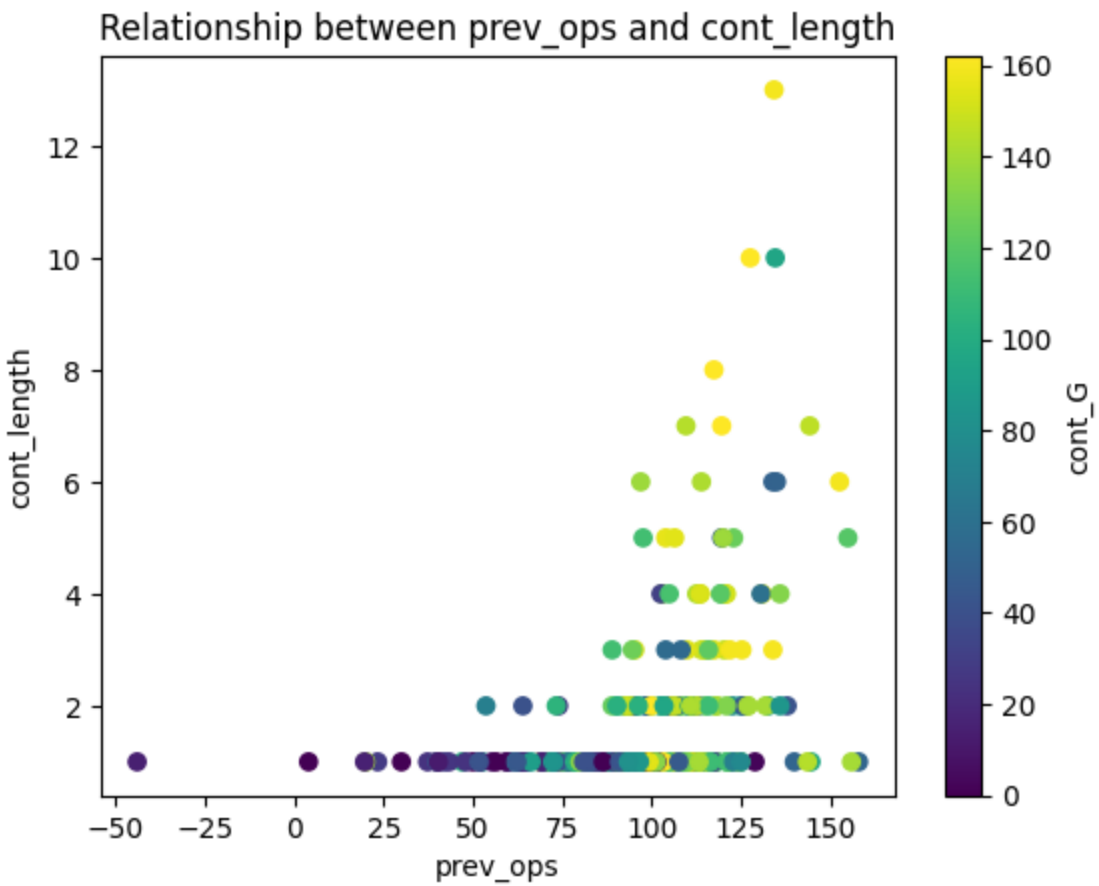
Process

	player	prev_ops	after_ops	dollars	cont_year	cont_length	played_year	prev_year	after_year	POS	predicted_OPS+	cont_year_OPS+	cont_G	cont_next_OPS+
0	A.J. Ellis	63.000000	82.000000	2.500000e+06	2016	1	8	1	1	C	74.513760	63.0	64	82.0
1	AJ Pollock	97.666667	124.333333	1.200000e+07	2018	5	6	3	3	OF	107.741135	108.0	113	107.0
2	Abraham Almonte	64.666667	93.000000	9.900000e+05	2020	1	7	3	1	OF	-0.489301	-5.0	7	93.0
3	Adam Duvall	100.333333	102.000000	5.000000e+06	2020	1	6	3	1	OF	115.222250	114.0	57	102.0
4	Adam Jones	103.000000	87.000000	3.000000e+06	2018	1	12	3	1	OF	92.991120	101.0	145	87.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
335	Yasmani Grandal	113.666667	133.500000	1.825000e+07	2019	4	7	3	2	C	119.848236	119.0	153	112.0
336	Yoenis Cespedes	136.000000	109.666667	2.750000e+07	2016	4	4	1	3	OF	130.464170	136.0	132	135.0
337	Yonder Alonso	111.000000	83.000000	8.000000e+06	2017	2	7	2	2	1B	114.871980	134.0	142	98.0
338	Yoshi Tsutsugo	93.500000	0.000000	4.000000e+06	2021	1	1	2	0	OF	102.868010	88.0	81	0.0
339	Zack Cozart	116.000000	34.500000	1.266667e+07	2017	3	6	2	2	SS	119.260090	140.0	122	81.0

340 rows × 14 columns

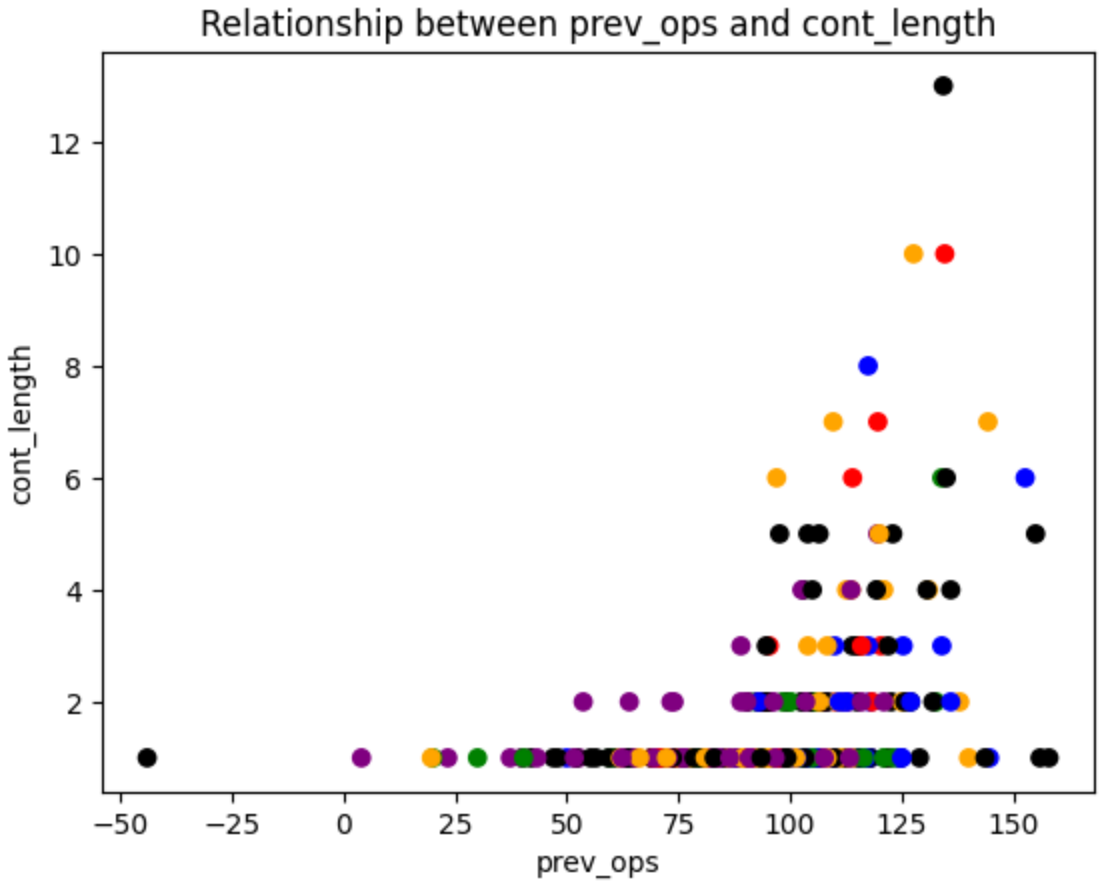
연봉과 연관 있을 것 같은 변수만 추출

Process



경기 수와 계약 연도 수 비교

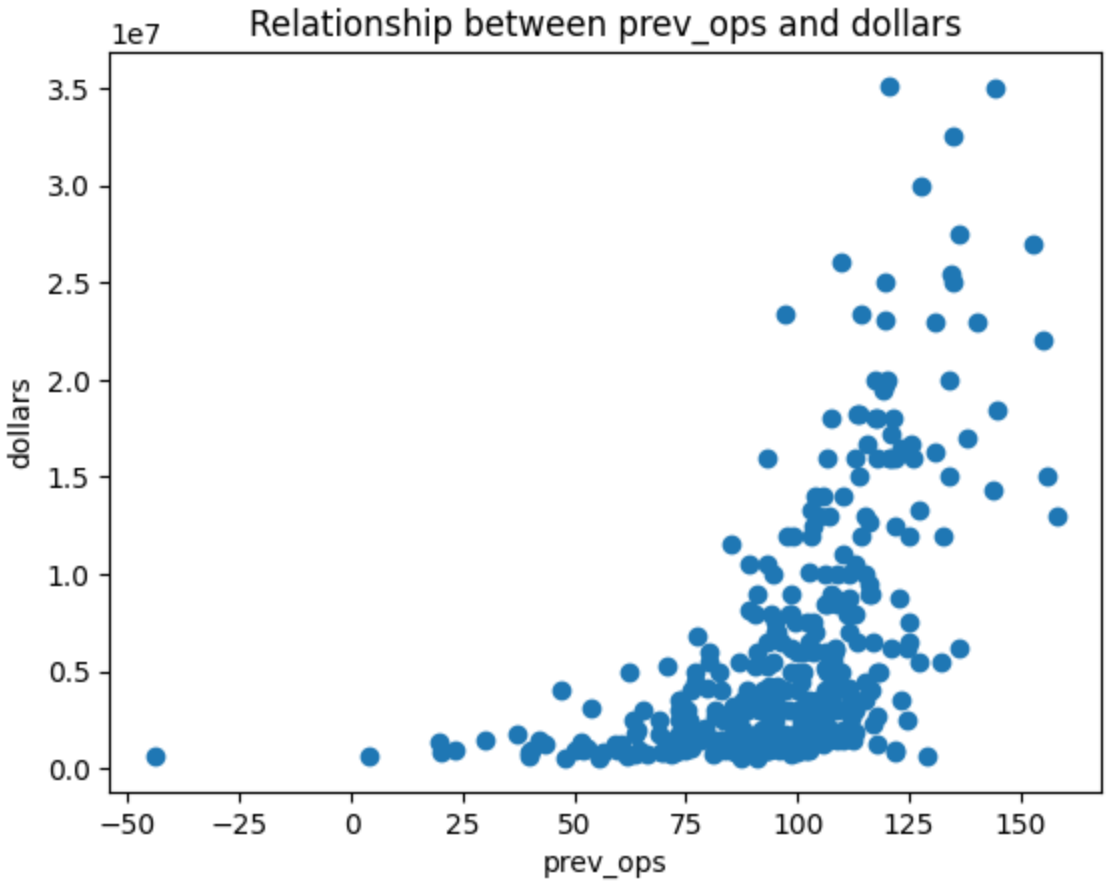
- 경기 수가 증가할 수록 다년 계약할 확률 상승



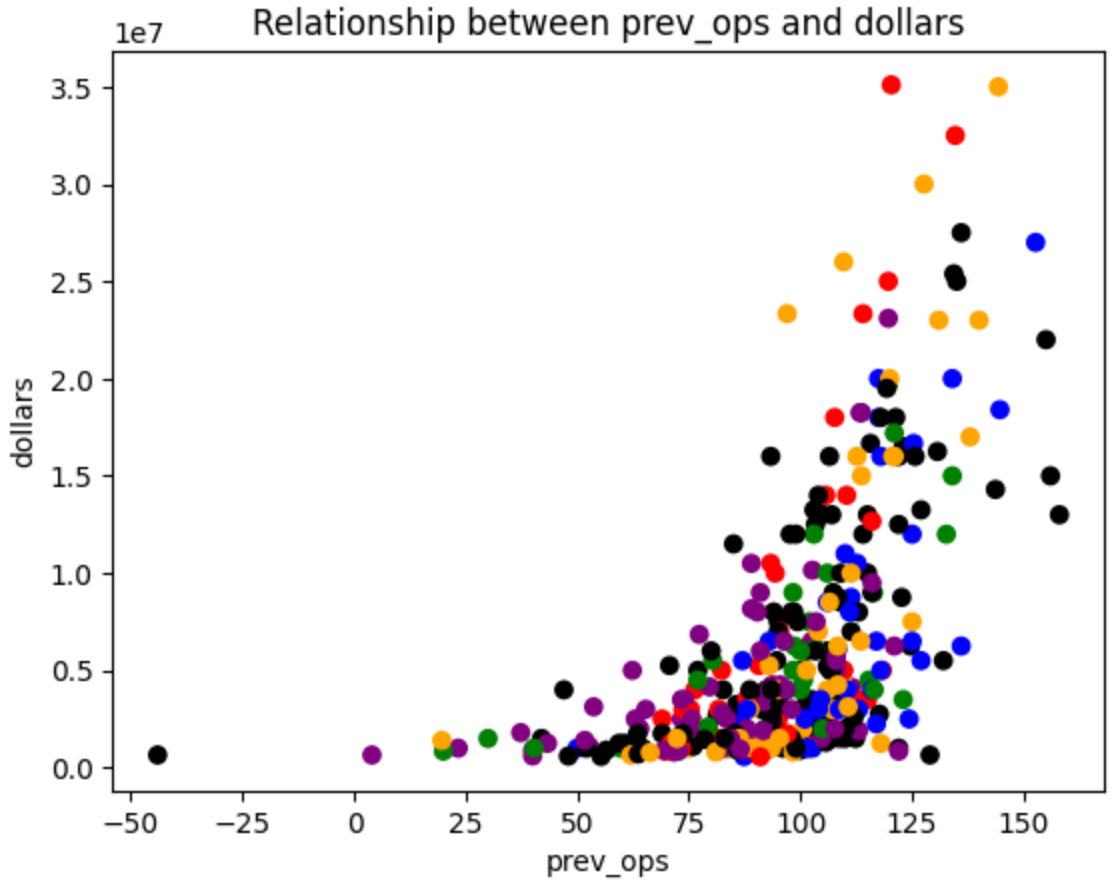
포지션과 계약 연도 수 비교

- 포지션은 계약 기간에 큰 연관 없어보임

Process



계약 전까지의 OPS+ 평균과 계약 연봉과의 관계
- OPS+가 증가할수록 연봉 역시 선형적으로 증가

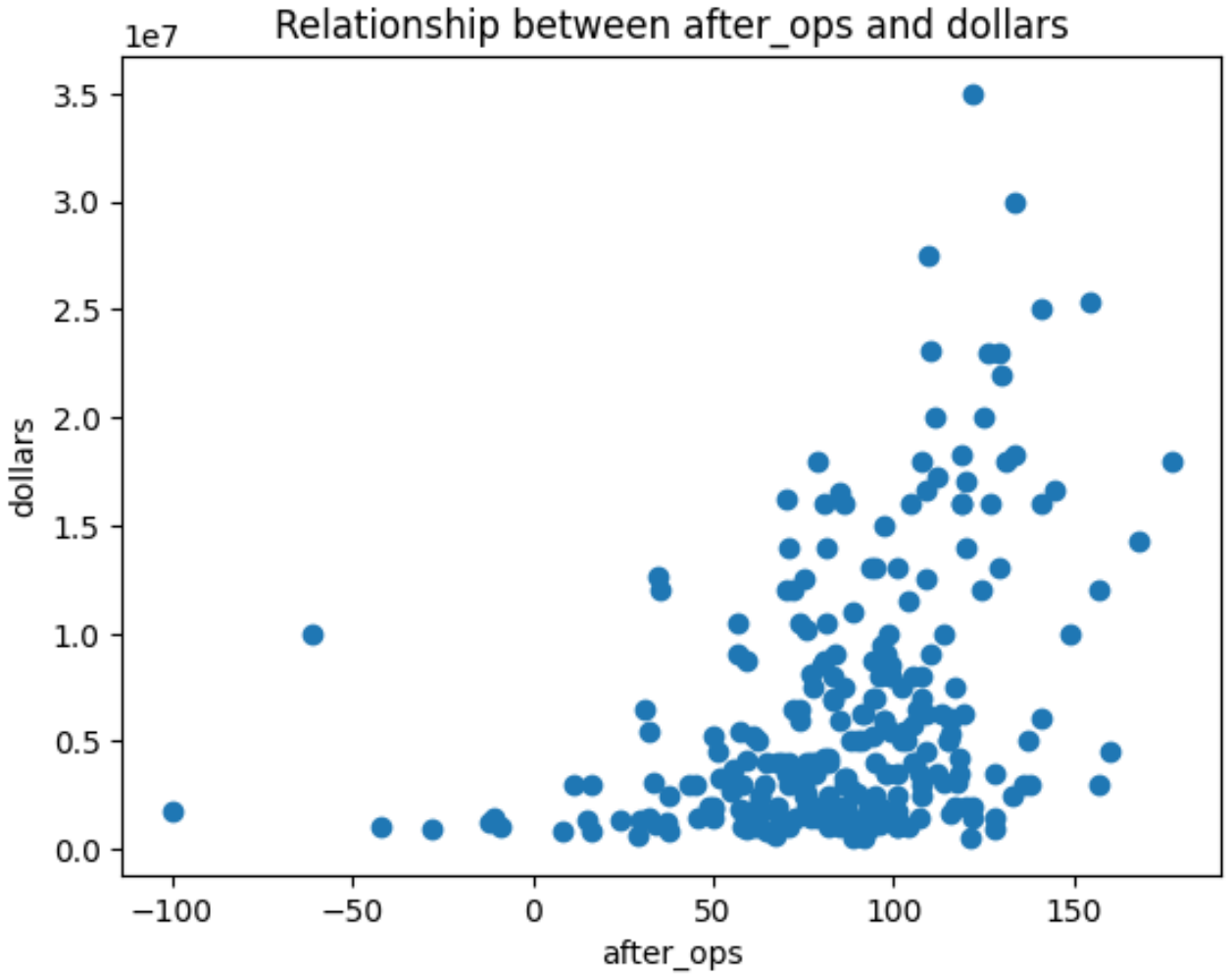


포지션과 연봉 비교
- 역시 포지션과 연봉에 큰 연관 없어보임

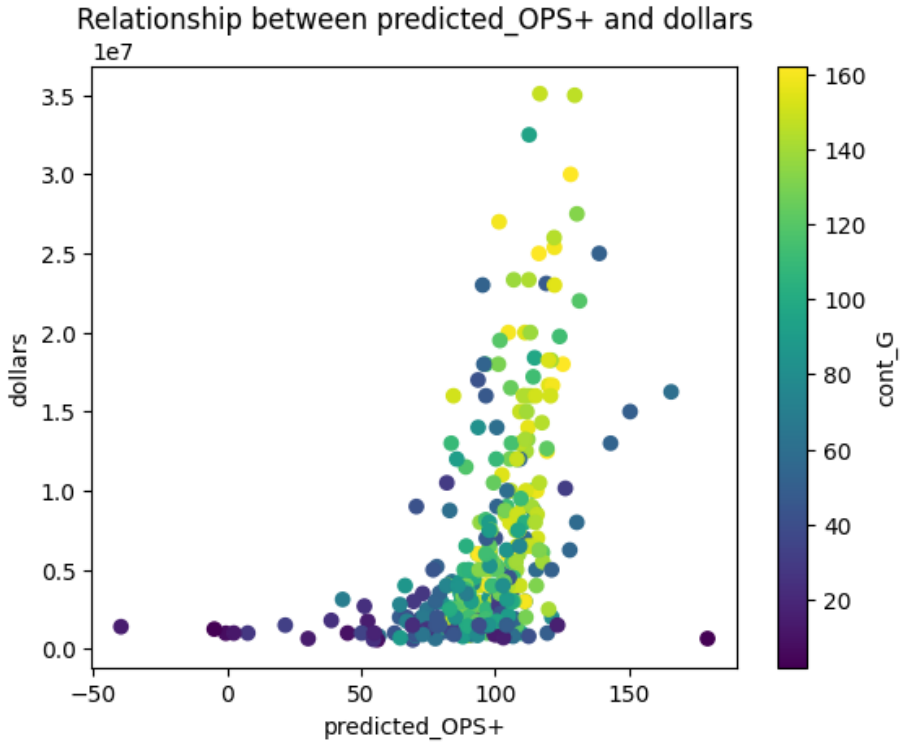
Process

계약 이후 OPS와 연봉의 관계

- 큰 연관없어보임
- 연봉은 높으나 낮은 OPS 보이는 선수 존재



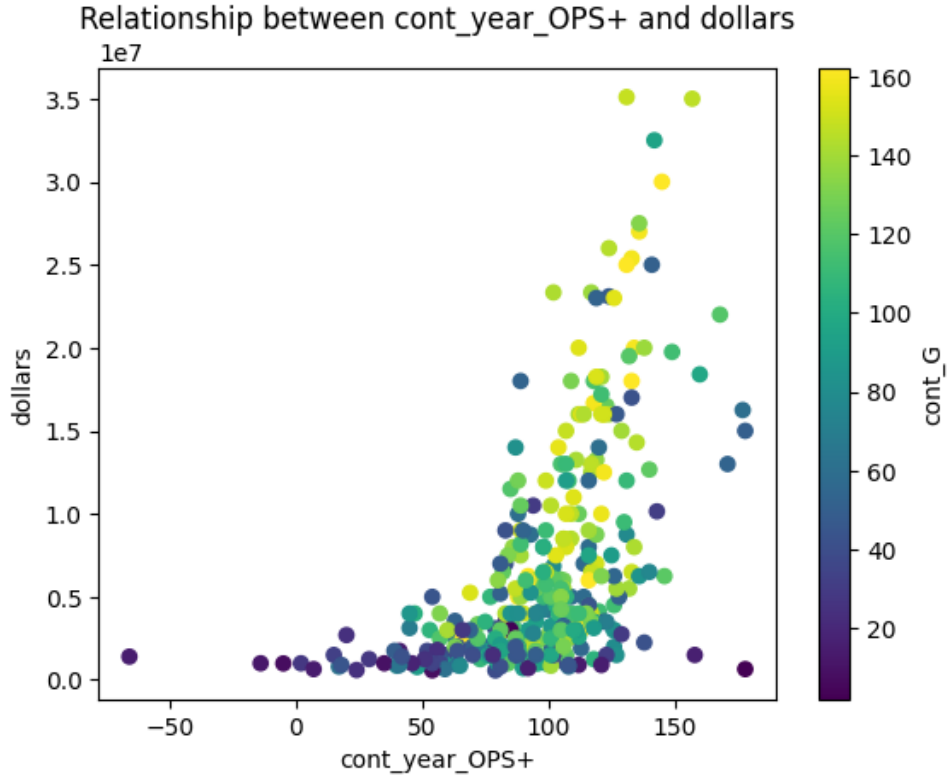
Process



OPS+ 예측치만으로는 연봉을 예측하기 어려움.

100을 넘어갔을 때부터 경기 수가 중요함.

OPS+와 경기수를 합쳐서 연봉을 예측



FA 해당 연도의 OPS+ 와 연봉의 관계
최근 성적(FA 직전 성적)이 계약 규모 결정에 유의미한 영향 미침.
(OPS+가 100을 넘어갈 시 경기 수도 중요함)

Process - modeling

```
[ ]  # X 값에는 계약 연도 OPS+, 계약 연도 경기수, 계약 연도, 이전 OPS 평균치, 계약 기간, 예측 OPS+를 넣고 y 값에는 연봉을 넣는다.
     X = result_df[['cont_year_OPS+', 'cont_G', 'prev_ops', 'cont_length', 'predicted_OPS+']]
     y = result_df['dollars']
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.35, random_state=42)
```

```
[ ]  clf = xgb.XGBRegressor()
```

```
[ ]  # 적절한 파라미터 값을 찾는다.
     param_dist = {
         'enable_categorical': [True, False],
         'objective': ['reg:squarederror', 'reg:absoluteerror'],
         'learning_rate': [0.05, 0.1, 0.2],
         'max_depth': [3, 5, 7],
         'n_estimators': [10, 15, 30]
     }

     random_search = RandomizedSearchCV(estimator=clf, param_distributions=param_dist, n_iter=10, cv=3, scoring='r2', random_state=42)

     random_search.fit(X_train, y_train)

     best_params = random_search.best_params_
     print("Best Parameters:", best_params)
```

```
Best Parameters: {'objective': 'reg:absoluteerror', 'n_estimators': 30, 'max_depth': 3, 'learning_rate': 0.2, 'enable_categorical': False}
```

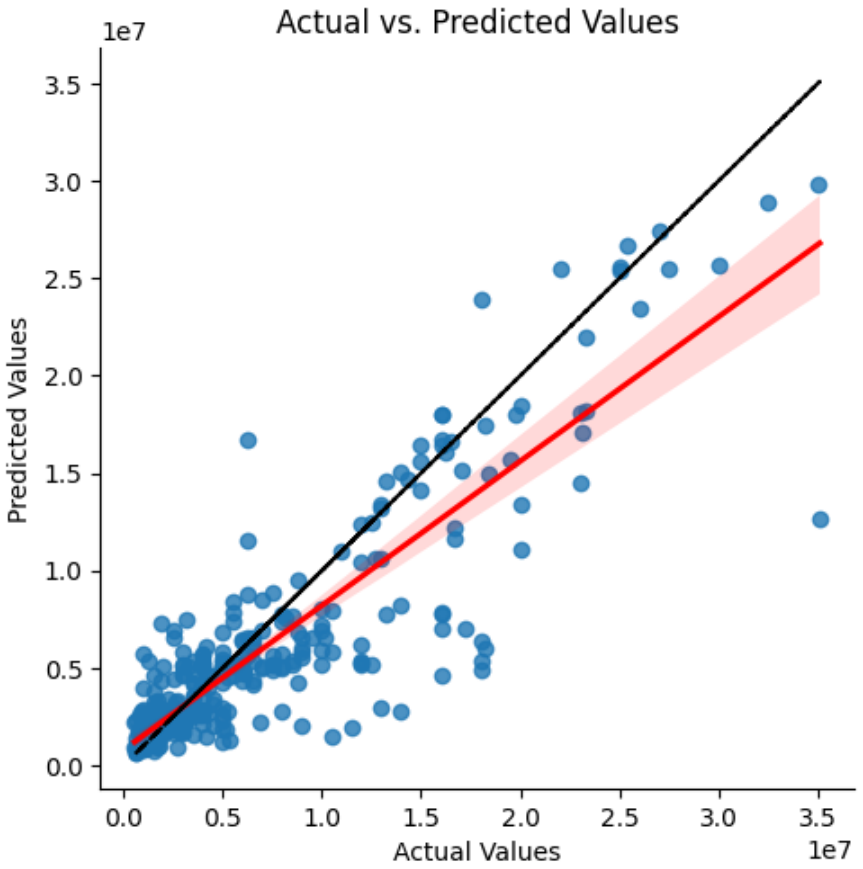
시각화를 통해 중요해 보이는 변수 선정 후 회귀분석 진행
Gridsearch보다는 Randomsearch를 선택
Grid의 최적 간격을 구하여 진행하기에는 비효율적임.

Process - modeling

```
# 결과
best_model = random_search.best_estimator_

y_pred = best_model.predict(X_test)

r_squared = r2_score(y_test, y_pred)
print("R-squared:", r_squared)
```

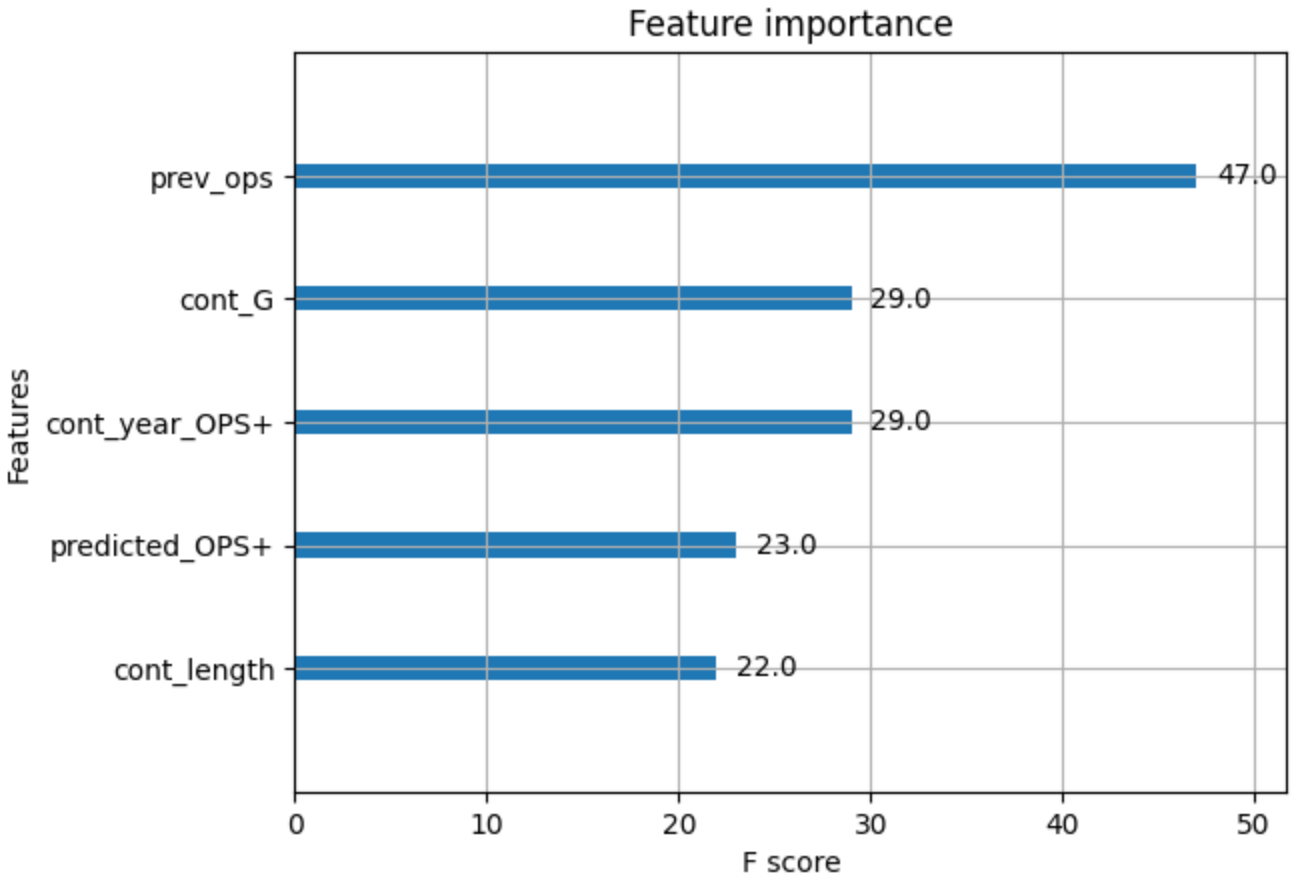
```
R-squared: 0.6620454525120703
```

결과
R-squared 값이 약 0.66



시각화해봤을 때,
예측이 잘 됐다 판단할 수 있음

Process - modeling



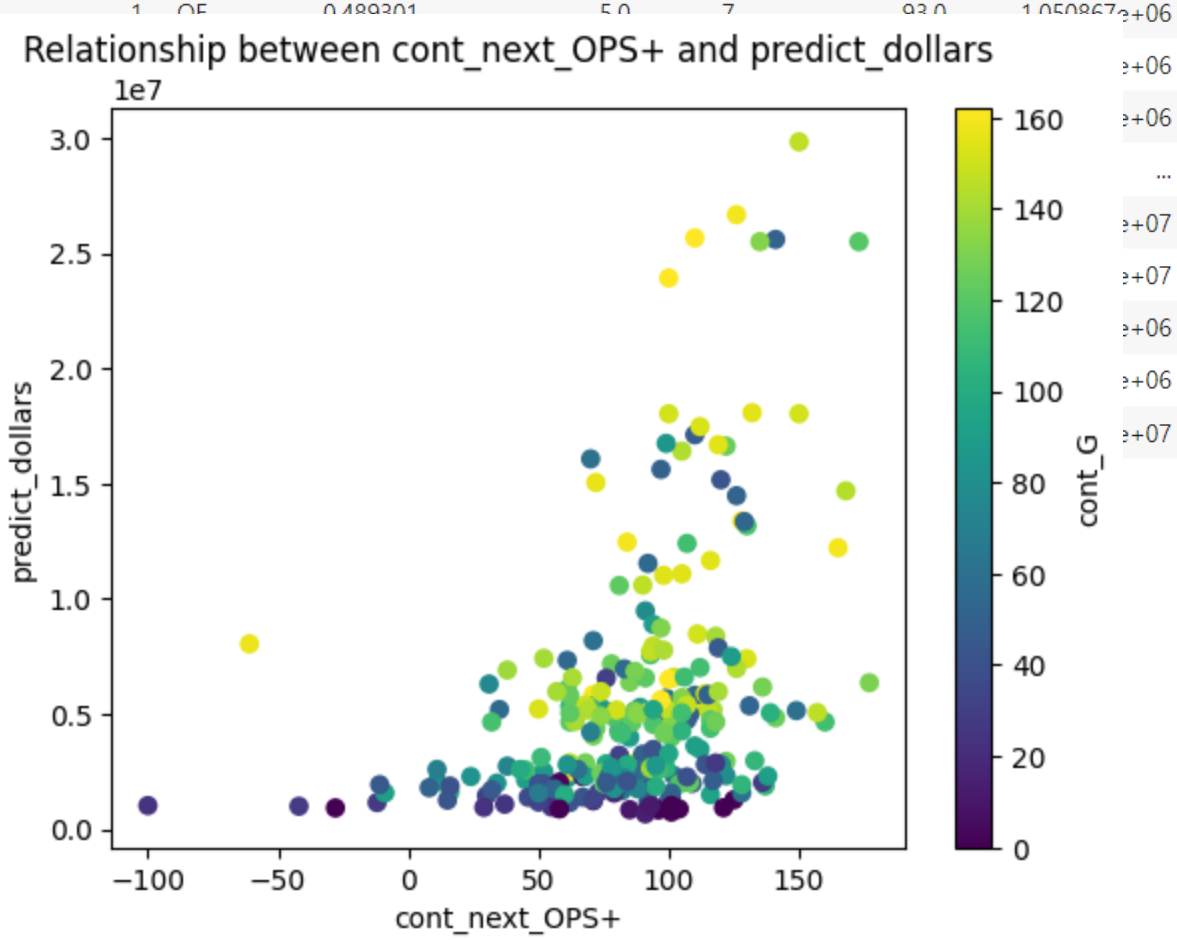
각 변수의 FA 계약 규모에 대한 영향력

1. prev_ops
2. Cont_G
3. Con_year_OPS+
4. Predicted_OPS+

기존의 활약, 경기 수, 계약 직전 활약이 중요함을 알 수 있음

Process

	player	prev_ops	after_ops	dollars	cont_year	cont_length	played_year	prev_year	after_year	POS	predicted_OPS+	cont_year_OPS+	cont_G	cont_next_OPS+	predict_dollars
0	A.J. Ellis	63.000000	82.000000	2.500000e+06	2016	1	8	1	1	C	74.513760	63.0	64	82.0	1.670403e+06
1	AJ Pollock	97.666667	124.333333	1.200000e+07	2018	5	6	3	3	OF	107.741135	108.0	113	107.0	1.240323e+07
2	Abraham Almonte	64.666667	93.000000	9.900000e+05	2020	1	7	3	1	OF	0.489301	5.0	7	93.0	1.050867e+06
3	Adam Duvall	100.333333	102.000000	5.000000e+06	2020	1	6	3							e+06
4	Adam Jones	103.000000	87.000000	3.000000e+06	2018	1	12	3							e+06
...	...	...	...	...	...	...	...	...							...
335	Yasmani Grandal	113.666667	133.500000	1.825000e+07	2019	4	7	3							e+07
336	Yoenis Cespedes	136.000000	109.666667	2.750000e+07	2016	4	4	1							e+07
337	Yonder Alonso	111.000000	83.000000	8.000000e+06	2017	2	7	2							e+06
338	Yoshi Tsutsugo	93.500000	0.000000	4.000000e+06	2021	1	1	2							e+06
339	Zack Cozart	116.000000	34.500000	1.266667e+07	2017	3	6	2							e+07



Predict_dollars 열 추가

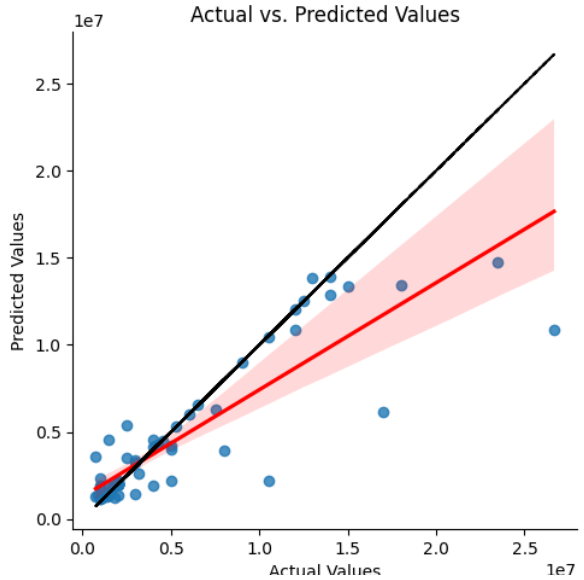
계약 직후 연봉과 OPS+ 그래프를 봤을 때
성적이 크게 하락한 고액연봉자를 발견할 수 있음

Process - modeling

모델을 통해 예측한 2024년 FA 대상자들의 계약 규모

R_squared : 0.58

Ex) 야스마니 그랜달(Yasmani Grandal)
예측 연봉: 393만$
실제:250만$
예측보다 적게 받음.
연봉 대비 효율 높을 것으로 예상



	player	prev_ops	after_ops	dollars	cont_year	cont_length	played_year	prev_year	after_year	predicted_OPS+	cont_year_OPS+	cont_G	cont_next_OPS+	predict_dollars
0	Aaron Hicks	88.000000	0	740000.0	2023	1	11	3	0	84.872559	106.0	93	0	3614110.750
1	Adam Duvall	102.333333	0	3000000.0	2023	1	10	3	0	87.922501	119.0	92	0	3357212.250
2	Adam Frazier	96.000000	0	4500000.0	2023	1	8	3	0	114.850449	94.0	141	0	4508165.000
3	Amed Rosario	98.333333	0	1500000.0	2023	1	7	3	0	107.075127	89.0	142	0	4564544.000
4	Andrew Knizner	83.000000	0	1825000.0	2023	1	5	2	0	90.754013	92.0	70	0	1197113.125
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
60	Travis Jankowski	61.500000	0	1700000.0	2023	1	9	2	0	90.945900	90.0	107	0	1810757.375
61	Tyler Wade	75.000000	0	850000.0	2023	1	7	3	0	61.396919	80.0	26	0	1345450.375
62	Victor Caratini	85.000000	0	6000000.0	2023	2	7	3	0	75.287384	95.0	62	0	6004695.500
63	Whit Merrifield	92.333333	0	8000000.0	2023	1	8	3	0	111.303688	94.0	145	0	3935465.000
64	Yasmani Grandal	98.666667	0	2500000.0	2023	1	12	3	0	83.631149	77.0	118	0	3534710.250

65 rows × 14 columns

Result

YSAL
YONSEI SPORTS ANALYTICS LAB

결론

FA 계약에 영향을 주는 것에는 prev_ops, cont G, Cont_year_ops+, predicted_ops+
활약이 뛰어날수록, 경기 출전 수가 많을수록 더 많은 연봉과 장기계약을 보장받음

실제로 FA 직전 활약이 중요하며, 경기 출전 수 역시 중요하다는 것을 알 수 있음.

계약 후 FA 직전 성적에 미치지 못하는 성적을 기록하는 선수가 있으며 FA Boost effect 존재한다는 결론 내릴 수 있음

Result

YSAL
YONSEI SPORTS ANALYTICS LAB

한계

1. OPS+와 같은 공격지표만을 고려함
 - 선수의 몸값에는 스타성과 같은 기타 요소들도 포함됨

2. 예측 모델의 결과, 예측값보다 실제값이 높게 나왔는데 원인을 명확히 밝혀내지 못함.

Result

추후 연구방향

Ex1) 선수들이 새로운 장기 계약을 맺은 후 성적 하락 여부

- 계약을 맺은 후 성적 하락으로 이어진 타자를 발견할 수 있었음.

Ex2) 투수의 몸값 추정
- 투수의 성적을 평가하는 절대적 지표가 없어서 투수를 프로젝트에서 제외함
- 투수의 성적을 평가하는 지표를 통한 몸값 추정

Recommendations – References

- Heather M.O'Neill (2014). Do Hitters Boost Their Performance During Their Contract Years

- 'On-base Plus Slugging Plus (OPS+)' , MLB.com

- 'Logistic Regression' , Habituaion of Memos, 2022.3.8

Feedback & QnA