예측모델 수정

Makeprediction1에서 index, a (반복횟수), b (유사한 공연 추출 개수)를 i,15,5로 지정

[진행과정]

: df2(올림픽공원 데이터)에서 **EXCCLC_EVENT_NMPR_CO** (정산인원)을 설명할 수 있는 예측변수가 무엇이 있을까 알아보기 위해 회귀분석 진행

: 유의한 변수를 찾고 vector3에서 추출되는 정산인원 데이터랑 같이 예측 진행

[전처리]

```
# 공연장소
df2.loc[df2["FCLTY_NM"] == "올림픽홀","FCLTY_NM"] = 1
df2.loc[df2["FCLTY_NM"] == "KSPO DOME(체조경기장)","FCLTY_NM"] = 2
df2.loc[df2["FCLTY_NM"] == "핸드볼경기장","FCLTY_NM"] = 3
df2.loc[df2["FCLTY_NM"] == "우리금융아트홀","FCLTY_NM"] = 4

# 문화예술행사 (1), 대중공연 (2), 체육행사 (3), 공공행사 (4), 순수예술공연 (5)
(dummy variable)
df2.loc[df2["EVENT_SDIV_NM"] == "문화예술행사","EVENT_SDIV_NM"] = 1
df2.loc[df2["EVENT_SDIV_NM"] == "대중공연","EVENT_SDIV_NM"] = 2
df2.loc[df2["EVENT_SDIV_NM"] == "체육행사","EVENT_SDIV_NM"] = 3
df2.loc[df2["EVENT_SDIV_NM"] == "공공행사","EVENT_SDIV_NM"] = 4
df2.loc[df2["EVENT_SDIV_NM"] == "순수예술공연","EVENT_SDIV_NM"] = 5

# 내부 (0), 외부 (1)로 코딩 (dummy variable)
df2.loc[df2["ISE_ELSE_FLAG_NM"] == "내부","ISE_ELSE_FLAG_NM"] = 0
df2.loc[df2["ISE_ELSE_FLAG_NM"] == "외부","ISE_ELSE_FLAG_NM"] = 1

#train & test set 분할
X_train, X_test, y_train, y_test = train_test_split(X,y, train_size =
0.7, test_size = 0.3, random_state = 1)
```

[회귀분석 진행]

y: 정산인원

X: 공연장소, 공연유형, 신청인원, 내외 여부, 사용일수, 년도, 월

- X (설명변수)간의 상관관계

	FCLTY_NM	EVENT_SDIV_NM	REQST_EVENT_NMPR_CO	ISE_ELSE_FLAG_NM	USE_DAY_CO	년도	월
FCLTY_NM	1.000000	0.305766	0.035101	0.623287	0.088973	-0.002849	0.052721
EVENT_SDIV_NM	0.305766	1.000000	-0.035917	0.288861	-0.103908	-0.112817	0.015665
REQST_EVENT_NMPR_CO	0.035101	-0.035917	1.000000	0.091299	0.171814	-0.028120	0.011050
ISE_ELSE_FLAG_NM	0.623287	0.288861	0.091299	1.000000	-0.053116	-0.022870	0.076262
USE_DAY_CO	0.088973	-0.103908	0.171814	-0.053116	1.000000	-0.019921	-0.066020
년도	-0.002849	-0.112817	-0.028120	-0.022870	-0.019921	1.000000	-0.047312
월	0.052721	0.015665	0.011050	0.076262	-0.066020	-0.047312	1.000000

- X의 VIF

	VIF Factor	features
0	3.617375	FCLTY_NM
1	4.602291	EVENT_SDIV_NM
2	1.166826	REQST_EVENT_NMPR_CO
3	1.904663	ISE_ELSE_FLAG_NM
4	1.279254	USE_DAY_CO
5	9.688729	년도
6	5.618623	월

해석)

상관관계- (공연장소, 내외부): 당연히 높을 수밖에 없음. 둘 중 하나만 쓰던지 interaction term 만들던지 조치

VIF- 모두 10 이상이기에 다중공선성 문제 배제 가능

Model 1

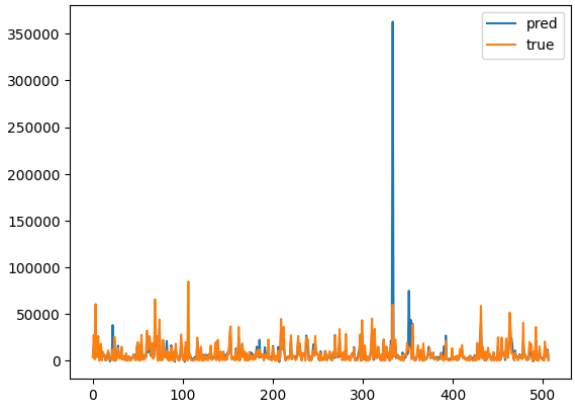
OLS Regression Results

Dep. Variable:	EXCCLC_EVENT_NMPR_CO	**R-squared:**	0.661
Model:	OLS	**Adj. R-squared:**	0.659
Method:	Least Squares	**F-statistic:**	326.6
Date:	Tue, 25 Jul 2023	**Prob (F-statistic):**	3.98e-270
Time:	06:17:03	**Log-Likelihood:**	-12707.
No. Observations:	1180	**AIC:**	2.543e+04
Df Residuals:	1172	**BIC:**	2.547e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.436e+05	2.06e+05	0.697	0.486	-2.6e+05	5.47e+05
FCLTY_NM	-20.2050	117.397	-0.172	0.863	-250.538	210.127
EVENT_SDIV_NM	-743.2034	350.498	-2.120	0.034	-1430.877	-55.530
REQST_EVENT_NMPR_CO	0.6354	0.014	46.541	0.000	0.609	0.662
ISE_ELSE_FLAG_NM	1051.2282	1300.161	0.809	0.419	-1499.675	3602.132
USE_DAY_CO	9.6843	23.370	0.414	0.679	-36.167	55.536
년도	-69.9451	102.049	-0.685	0.493	-270.165	130.275
월	23.0491	100.310	0.230	0.818	-173.759	219.857

Omnibus:	609.174	**Durbin-Watson:**	2.016
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	7001669.046
Skew:	-0.174	**Prob(JB):**	0.00
Kurtosis:	380.368	**Cond. No.**	1.63e+07

해석) R^2로 꽤 선형 상관관계가 있어보이기는 하나 p-value를 보면 공연유형, 신청인원 외의 변수는 유의하지 않은 것으로 나타났다.

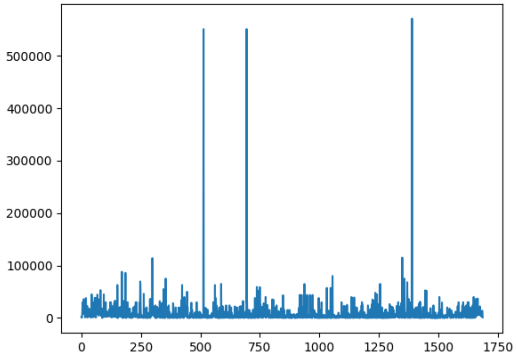


해당 model로 예측했을 때, 왜 300~400번째 공연은 실제값보다 훨씬 크게 예측했는지 보기

⇨ 확인해보니

공공행사	2019한성백제문화제	2019-09-19	2019-10-02	570000	평화의광장

그래서 신청인원 확인해보니



공공행사	2019한성백제문화제	2019-09-19	2019-10-02	570000	60000
	태권도퍼포먼스 “탈”			114000	14226
	제 14회 한성백제문화제			550000	600000
	제 15회 한성백제문화제			550000	100000
	2019 박효신 20주년 콘서트			115000	101251

⇨ 신청인원과 정산인원이 차이 많이 나는 데이터들이 예측력을 흐리는 듯 보임.

⇨ 신청인원이 정산인원보다 매우 큰 경우 데이터에서 제외하고자함.

2배 차이

```
df2[df2['REQST_EVENT_NMPR_CO']>=2*df2['EXCCLC_EVENT_NMPR_CO']] #189 개
```
3배 차이

```
df2[df2['REQST_EVENT_NMPR_CO']>=3*df2['EXCCLC_EVENT_NMPR_CO']] #88 개
```
4배 차이

```
df2[df2['REQST_EVENT_NMPR_CO']>=4*df2['EXCCLC_EVENT_NMPR_CO']] #54 개
```

⇨ 추가적으로 신청인원보다 정산인원이 매우 큰 경우도 제외함.

```
df2[df2['REQST_EVENT_NMPR_CO']*4<df2['EXCCLC_EVENT_NMPR_CO']] #33 개
```

이상치 제외해서 (1598, 13)의 df2_new 생성

Model 2

OLS Regression Results

Dep. Variable:	EXCCLC_EVENT_NMPR_CO	**R-squared:**	0.954
Model:	OLS	**Adj. R-squared:**	0.954
Method:	Least Squares	**F-statistic:**	3289.
Date:	Thu, 27 Jul 2023	**Prob (F-statistic):**	0.00
Time:	05:53:32	**Log-Likelihood:**	-10943.
No. Observations:	1118	**AIC:**	2.190e+04
Df Residuals:	1110	**BIC:**	2.194e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.751e+04	7.78e+04	-0.225	0.822	-1.7e+05	1.35e+05
FCLTY_NM	108.4686	51.149	2.121	0.034	8.108	208.829
EVENT_SDIV_NM	370.6878	143.011	2.592	0.010	90.085	651.290
REQST_EVENT_NMPR_CO	1.0118	0.007	148.508	0.000	0.998	1.025
ISE_ELSE_FLAG_NM	-683.4970	538.736	-1.269	0.205	-1740.554	373.560
USE_DAY_CO	-38.6910	9.262	-4.177	0.000	-56.864	-20.518
년도	8.0514	38.584	0.209	0.835	-67.653	83.756
월	-59.1172	38.552	-1.533	0.125	-134.760	16.526

Omnibus:	605.903	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39910.404
Skew:	-1.673	**Prob(JB):**	0.00
Kurtosis:	32.079	**Cond. No.**	1.28e+07

: 기존 train_set에서 극 이상치 87개 제거하니 R^2가 0.954로 (overfitting)/ train_set_new에서도 내외여부랑 연도, 월은 p-value는 매우 크기에 해당 변수를 제거하고자함. (신청인원과 같이 사용일수의 유의성이 매우 커짐.)

⇨ 공연 장소, 공연 유형, 신청인원, 사용일수만 남기자

Model3

⇨ 공연 장소, 공연 유형, 신청인원, 사용일수만 사용한 모델

역시 model 2와 R^2 똑같다, 결과 차이가 거의 없음

OLS Regression Results

Dep. Variable:	EXCCLC_EVENT_NMPR_CO	**R-squared:**	0.954
Model:	OLS	**Adj. R-squared:**	0.954
Method:	Least Squares	**F-statistic:**	5749.
Date:	Thu, 27 Jul 2023	**Prob (F-statistic):**	0.00
Time:	06:00:40	**Log-Likelihood:**	-10945.
No. Observations:	1118	**AIC:**	2.190e+04
Df Residuals:	1113	**BIC:**	2.193e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1582.9024	284.845	-5.557	0.000	-2141.796	-1024.009
FCLTY_NM	62.3521	39.386	1.583	0.114	-14.928	139.632
EVENT_SDIV_NM	345.3942	140.525	2.458	0.014	69.670	621.118
REQST_EVENT_NMPR_CO	1.0109	0.007	148.949	0.000	0.998	1.024
USE_DAY_CO	-36.4933	9.194	-3.969	0.000	-54.533	-18.454

Omnibus:	599.656	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39967.237
Skew:	-1.641	**Prob(JB):**	0.00
Kurtosis:	32.107	**Cond. No.**	4.99e+04

Model4

: 신청인원 변수만 제외

```
lr4 = smf.ols(formula= 'EXCCLC_EVENT_NMPR_CO~ FCLTY_NM+EVENT_SDIV_NM+USE_DAY_CO', data= train_set_new)
model4 = lr4.fit()
model4.summary()
```

OLS Regression Results

Dep. Variable:	EXCCLC_EVENT_NMPR_CO	**R-squared:**	0.034
Model:	OLS	**Adj. R-squared:**	0.031
Method:	Least Squares	**F-statistic:**	12.90
Date:	Thu, 27 Jul 2023	**Prob (F-statistic):**	2.76e-08
Time:	06:08:04	**Log-Likelihood:**	-12645.
No. Observations:	1118	**AIC:**	2.530e+04
Df Residuals:	1114	**BIC:**	2.532e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7747.2555	1270.774	6.096	0.000	5253.874	1.02e+04
FCLTY_NM	441.2411	179.747	2.455	0.014	88.560	793.922
EVENT_SDIV_NM	-1799.4133	639.272	-2.815	0.005	-3053.726	-545.100
USE_DAY_CO	206.1673	41.382	4.982	0.000	124.973	287.362

Omnibus:	2735.053	**Durbin-Watson:**	2.013
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23565155.578
Skew:	24.016	**Prob(JB):**	0.00
Kurtosis:	712.622	**Cond. No.**	36.2

Model5

: 표준화해도 변화가 없었다.

[변수 선정]

⇨ 공연 장소, 공연 유형, 신청인원, 사용일수만 사용한 모델

[Makeprediction2 함수 생성]

공연 장소, 공연 유형, 신청인원, 사용일수와 vector3에서 추출된 값 활용

신청인원, 정산인원 4배 이상 차이나는 df2_new 사용한 vector3_new 생성해서 구하려고 했는 데, vector3_new(6,1)의 경우 유사한 공연 자체를 못 찾는 문제가 생겨서 vector3와 df2 사용해서 makeprediction2 생성

```
df2_pred = df2_pred.iloc[:,[0,1,5,6,8]]
df2_pred
```

	FCLTY_NM	EVENT_SDIV_NM	REQST_EVENT_NMPR_CO	USE_DAY_CO
0	8	4	2000	1
1	17	4	300	1
2	18	3	1600	2
3	9	1	3500	11
4	14	1	20000	62
...	...	...	...	...
1682	5	1	6000	7
1683	3	1	8000	5
1684	2	1	14000	8
1685	7	4	10000	2
1686	4	2	1100	2

=〉 사용할 4개의 변수만을 가지는 df2_pred의 우측에 vector3열 여러 개 추가하려고 했는 데, vector3(1,5)의 경우는 무조건 vector3 한 개만 추출하기에 vector3열 한 개만 병합

```
df2_pred.join(vector3_col,how='inner') #1687번째 행은 도출 X
```

	FCLTY_NM	EVENT_SDIV_NM	REQST_EVENT_NMPR_CO	EXCCLC_EVENT_NMPR_CO	USE_DAY_CO	vector3_col
0	8	4	2000	2000	1	2000
1	17	4	300	300	1	2000
2	18	3	1600	1600	2	2000
3	9	1	3500	2395	11	2000
4	14	1	20000	14153	62	2000
...	...	...	...	...	...	...
1681	3	1	2600	3219	3	7128
1682	5	1	6000	4504	7	7128
1683	3	1	8000	9493	5	6925
1684	2	1	14000	15100	8	9493
1685	7	4	10000	10000	2	3219

1686 rows × 6 columns

⇨ 모형

```
model_pred = smf.ols(formula= 'EXCCLC_EVENT_NMPR_CO~ FCLTY_NM+EVENT_SDIV_NM+REQST_EVENT_NMPR_CO+USE_DAY_CO+vector3_col',data= df2_pred).fit()
model_pred.summary()
```

OLS Regression Results

Dep. Variable:	EXCCLC_EVENT_NMPR_CO	**R-squared:**	0.532
Model:	OLS	**Adj. R-squared:**	0.530
Method:	Least Squares	**F-statistic:**	381.3
Date:	Fri, 28 Jul 2023	**Prob (F-statistic):**	1.59e-273
Time:	03:15:09	**Log-Likelihood:**	-18228.
No. Observations:	1686	**AIC:**	3.647e+04
Df Residuals:	1680	**BIC:**	3.650e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4662.8678	673.838	6.920	0.000	3341.217	5984.519
FCLTY_NM	34.3906	84.279	0.408	0.683	-130.912	199.694
EVENT_SDIV_NM	-1215.1607	304.711	-3.988	0.000	-1812.813	-617.508
REQST_EVENT_NMPR_CO	0.4831	0.011	42.149	0.000	0.461	0.506
USE_DAY_CO	50.4868	22.304	2.264	0.024	6.740	94.233
vector3_col	-0.0161	0.029	-0.548	0.584	-0.074	0.041

Omnibus:	2540.099	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	13462629.966
Skew:	7.922	**Prob(JB):**	0.00
Kurtosis:	440.479	**Cond. No.**	6.68e+04

⇨ Prediction 함수

```
def makeprediction2(index,df):
  input_data = list(df.loc[index])[[0,1,2,4,5]]
  result = model_pred.predict(input_data)
  return result
```