KCSE 2023 튜토리얼 - 자료 분석

백규승 선임 연구원

AI LAB

2023.02.08

obzen
Action on Knowledge

개요

- 회귀분석과 인과추론
- 인과추론을 이용한 마케팅 분석 예
- 질의응답

회귀분석과 인과추론

Renew dataset

- 구독 서비스를 제공하는 어느 인터넷 사이트의 방문자에 대한 정보 dataset
- 각 고객에 대해 다음과 같은 정보를 가짐

sales_calls	광고 전화 수	**discount**	제공받은 할인률
interaction	수신된 광고 전화 수	**monthly_usage**	당월 기준 이용률
economy	고객의 (상대적인)경제 상황	**ad_spend**	고객에게 사용한 광고료
last_upgrade	최근 상품 업그레이드 시기	**bug_reported**	버그 보고 수
product_need	상품에 대한 욕구	**did_renew**	재구독 여부 (0 or 1)

- 목적: 고객에게 사용한 광고료(ad_spend)가 실제 재구독 여부(did_renew)에 어떤 영향을 끼쳤는가?
- 참고: 전체 고객의 재구독률: 30.4%

Renew dataset

- description

```
data.describe()
```

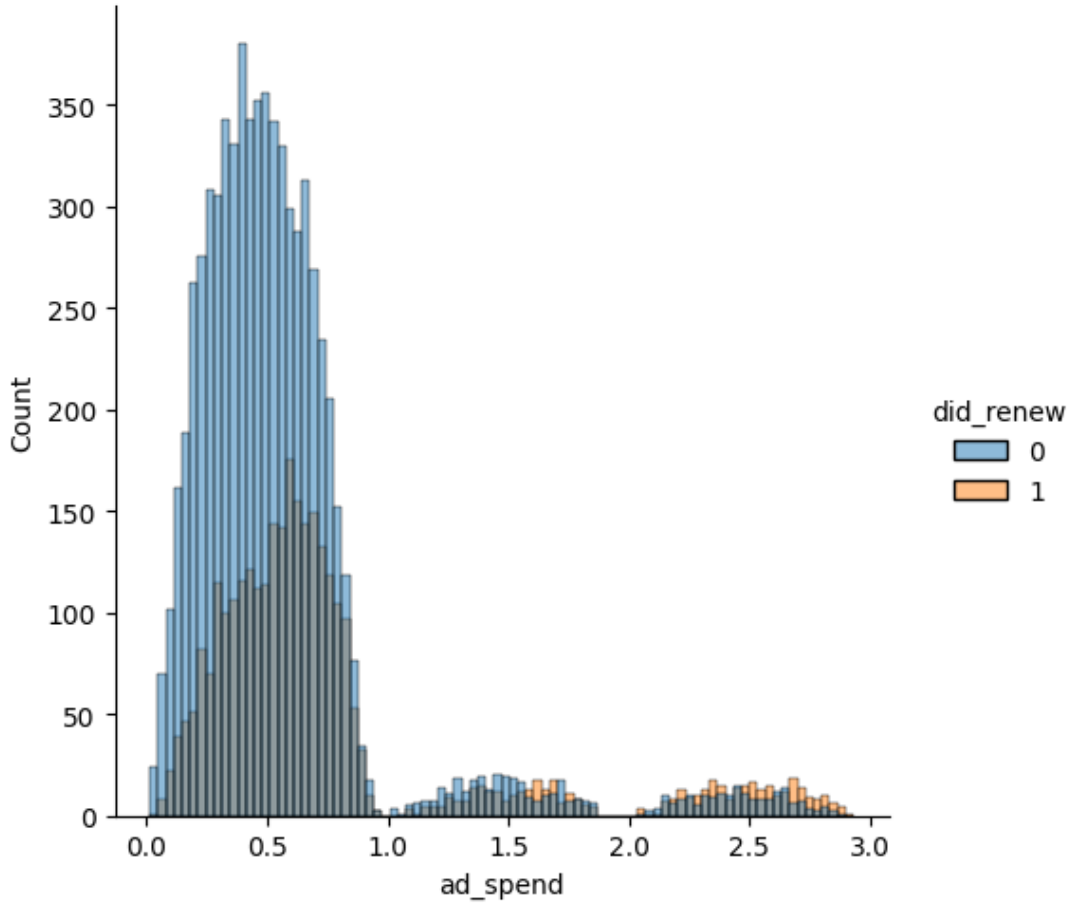
	sales_calls	interactions	economy	last_upgrade	product_need	discount	monthly_usage	ad_spend	bugs_reported	did_renew
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000
mean	1.987900	2.194800	0.496817	10.027655	0.205747	0.238824	0.510413	0.623815	0.60620	0.304500
std	1.229107	1.304923	0.288887	5.736958	1.005001	0.090296	0.213940	0.503217	0.79359	0.460218
min	0.000000	0.000000	0.000345	0.005662	-3.756813	0.011001	0.015036	0.014785	0.00000	0.000000
25%	1.000000	1.000000	0.243905	5.051535	-0.467320	0.171214	0.342557	0.340137	0.00000	0.000000
50%	2.000000	2.000000	0.499632	10.032572	0.206263	0.238871	0.512767	0.517237	0.00000	0.000000
75%	3.000000	3.000000	0.746817	14.993790	0.872302	0.305277	0.679745	0.697351	1.00000	1.000000
max	4.000000	7.000000	0.999919	19.996676	4.124378	0.483810	0.981948	2.929800	6.00000	1.000000

EDA

- ad_spend와 did_renew 간의 관계를 알아보고자 histplot을 그림

```
sns.displot(data, x = "ad_spend", hue = "did_renew")
```

```
<seaborn.axisgrid.FacetGrid at 0x1d1e98a3640>
```



EDA

- ad_spend가 1, 2를 지날 때 마다 분포가 많이 바뀌는 것을 확인 가능.
- ad_spend가 0~1, 1~2, 2 초과인 구간에서 재구독률을 계산

```
ad_spend_bin = pd.cut(data["ad_spend"], bins = [0,1,2, np.inf], labels = [0,1,2])
for i in range(3):
    print(data["did_renew"].loc[ad_spend_bin==i].mean())
```

```
0.2831545183054972
0.4274353876739563
0.5921052631578947
```

- Ad_spend가 클수록 재구독률이 상당히 큰 것을 확인할 수 있다.
- ➔ ad_spend를 늘리면 재구독률이 증가하는 것이 아닌가?

logistic regression

- 모든 변수를 이용해서 재구독률을 예상하는 logistic regression 모형 생성
- ad_spend 변수의 계수가 매우 크며, p-value 값도 매우 작은 것을 확인 가능
- ad_spend를 늘리면 확실히 구독자수를 늘릴 수 있는 것인가?

```
import statsmodels.api as sm

sm_model = sm.Logit(Y, X).fit(disp=0)
sm_model.summary()
```

Logit Regression Results

Dep. Variable:	did_renew	**No. Observations:**	10000
Model:	Logit	**Df Residuals:**	9991
Method:	MLE	**Df Model:**	8
Date:	Wed, 08 Feb 2023	**Pseudo R-squ.:**	0.1804
Time:	09:48:11	**Log-Likelihood:**	-5037.5
converged:	True	**LL-Null:**	-6146.3
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
sales_calls	-0.0730	0.057	-1.284	0.199	-0.184	0.038
interactions	0.0842	0.053	1.576	0.115	-0.021	0.189
economy	0.3228	0.080	4.014	0.000	0.165	0.480
last_upgrade	-0.0563	0.005	-12.211	0.000	-0.065	-0.047
product_need	1.0288	0.036	28.913	0.000	0.959	1.099
discount	-3.1323	0.267	-11.726	0.000	-3.656	-2.609
monthly_usage	-0.6999	0.132	-5.306	0.000	-0.958	-0.441
ad_spend	0.4442	0.056	7.936	0.000	0.334	0.554
bugs_reported	-0.0477	0.034	-1.394	0.163	-0.115	0.019

treatment effect 추정

- econml을 이용해서 ad_spend를 treatment로 보고 단순한 선형 treatment effect 모형을 추정

- 모든 계수가 전체적으로 작은 것을 확인할 수 있음.
  - p-value가 매우 큼. 대부분 0으로 간주 가능
  - 유일하게 p-value가 작은 last_upgrade의 경우에도 계수가 매우 작음

```
from econml.dml import LinearDML

est = LinearDML()
est.fit(data["did_renew"], data["ad_spend"],
        X=data.drop(["did_renew", "ad_spend"], axis = 1) )
est.summary()
```

Coefficient Results

	point_estimate	stderr	zstat	pvalue	ci_lower	ci_upper
sales_calls	-0.034	0.027	-1.291	0.197	-0.086	0.018
interactions	0.035	0.025	1.378	0.168	-0.015	0.084
economy	-0.004	0.038	-0.102	0.919	-0.079	0.071
last_upgrade	-0.009	0.002	-3.681	0.0	-0.014	-0.004
product_need	0.006	0.014	0.45	0.653	-0.021	0.033
discount	0.088	0.151	0.586	0.558	-0.207	0.384
monthly_usage	-0.065	0.057	-1.154	0.249	-0.177	0.046
bugs_reported	0.015	0.016	0.919	0.358	-0.017	0.047

CATE Intercept Results

	point_estimate	stderr	zstat	pvalue	ci_lower	ci_upper
cate_intercept	0.09	0.054	1.674	0.094	-0.015	0.195

A linear parametric conditional average treatment effect (CATE) model was fitted:

$$Y = \Theta(X) \cdot T + g(X, W) + c$$

where for every outcome i and treatment j the CATE $\Theta_{ij}(X)$ has the form:

$$\Theta_{ij}(X) = X' coef_{ij} + cate\_intercept_{ij}$$

Coefficient Results table portrays the $coef_{ij}$ parameter vector for each outcome i and treatment j. Intercept Results table portrays the $cate\_intercept_{ij}$ parameter.

treatment effect 추정

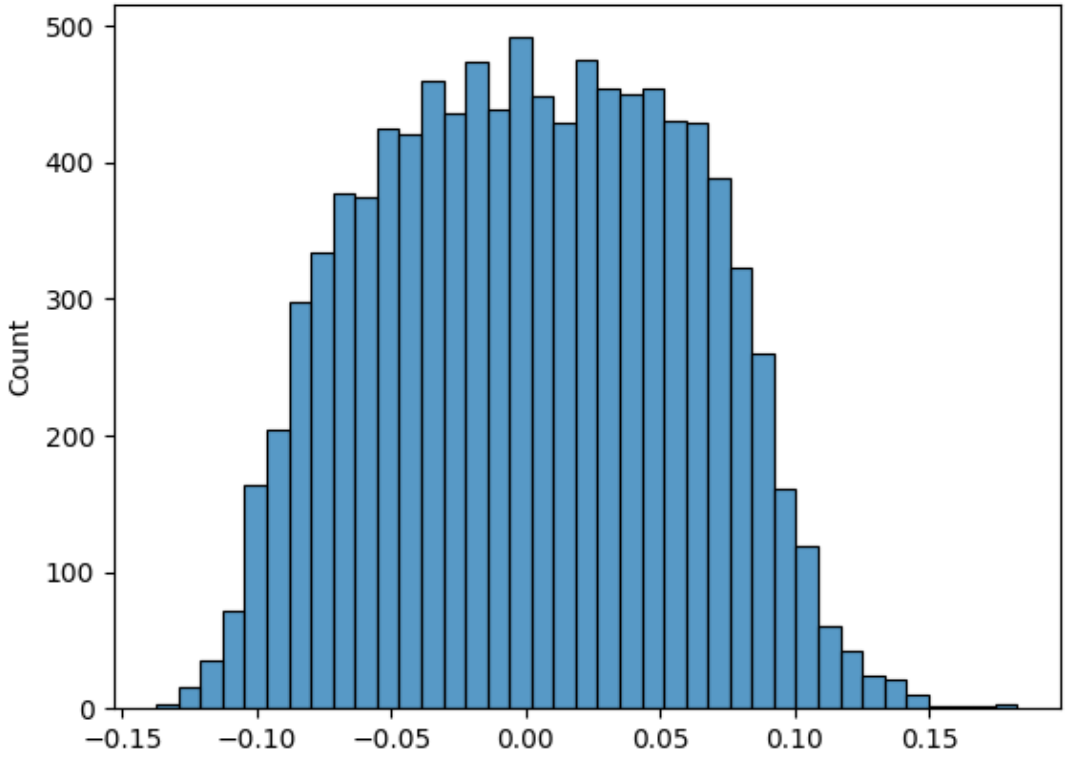
- econml을 이용해서 ad_spend를 treatment로 보고 단순한 선형 treatment effect 모형을 추정

- 실제 각 고객에 대한 treatment effect 추정값 또한 매우 작음
  - 전체 고객의 treatment effect 값의 평균은 0.002
  - 평균적으로 ad_spend 1단위 증가가 재구독률을 0.2% 상승시켰음을 나타냄

```
causal_effect = est.effect(data.drop(["did_renew", "ad_spend"], axis = 1))
print(causal_effect.mean())
sns.histplot(causal_effect)
```

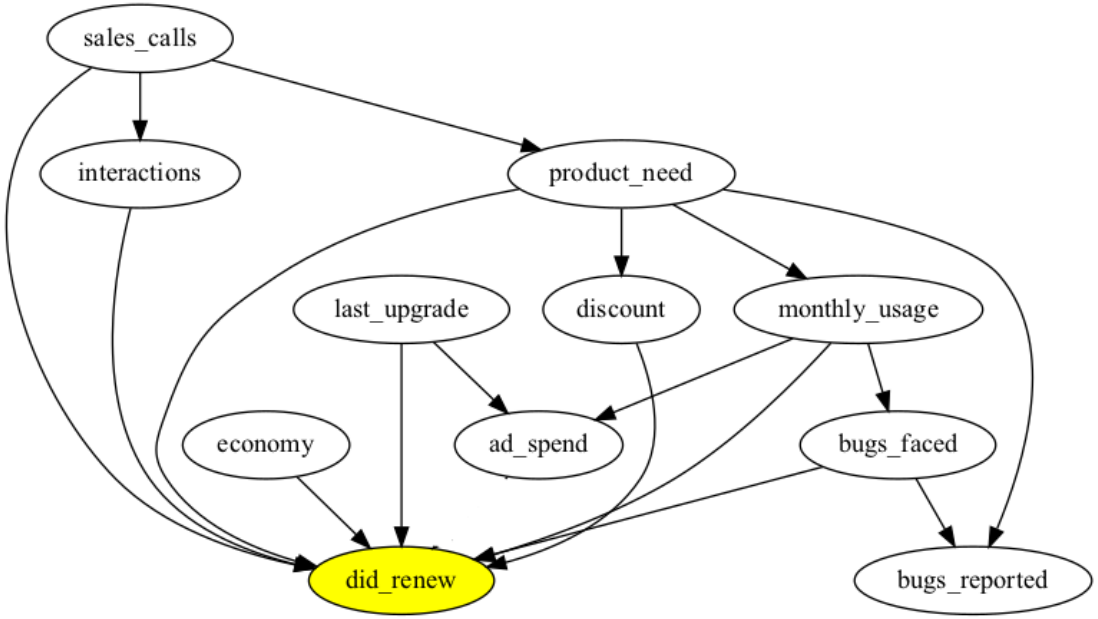
```
0.0020878841342368334
```

```
<AxesSubplot: ylabel='Count'>
```



사실은...

- 실제 특정 사이트에서 수집된 데이터가 아닌, 가상의 생성된 데이터 셋
- 재구독 여부는 ad_spend 값과는 무관
- ad_spend에 영향을 끼치는 last_upgrade, monthly_usage가 did_renew에 영향을 끼치기 때문에 앞에서의 호도된 결과가 나옴
- 단순 회귀 분석으로는 이러한 결과를 분리해 낼 수 없다.



인과 추론을 활용한 마케팅 설계

obzen
Action on Knowledge

캠페인 정보

■ 캠페인 개요

캠페인 대상		어플리케이션 미설치자
캠페인 비용	채널	LMS 발송 비용 (인당 25원)
	오퍼	-
캠페인 성과 지표		LMS 발송 당일 어플리케이션 설치 후 로그인 여부

■ 기존 캠페인

- point + 이용 금액으로 고객을 선정
- 기존 고객들에 대한 정보 분석 ➔
  point가 높은 고객들이 앱을 활발히 이용한다 + 이용 금액이 많은 고객들이 앱을 활발히 이용한다.
- 마케팅 내용과 무관한 고객 선정 방식

캠페인 설계

- 설계를 위한 캠페인: A/B test를 신규로 설계
  - 전체 앱 미설치 고객 중 일부를 선별해서 마케팅을 진행
  - 결과를 바탕으로 고객별 캠페인 효과를 측정

- 다음 캠페인 설계: 앱 미설치 고객 120만명을 50만 / 50만 / 20만으로 임의로 분할 (집단 A/B/C)
- 집단 A는 기존 마케터들의 기준을 바탕으로 고객을 선정해서 마케팅
- 집단 B는 위의 설계 캠페인의 결과를 바탕으로 캠페인 효과가 우수한 고객을 선정해서 마케팅
  - 선정 고객 수는 집단 A에 맞춤
- 집단 C는 대조군으로 마케팅을 실시하지 않음

캠페인 결과

■ 타겟팅 방식別 캠페인 효과 비교

타겟팅 방식	전체 고객 수 (명)			앱 로그인 고객 수 (명)			증분 고객 수 (명)
	전체	캠페인 수행	미수행	전체	캠페인 수행	미수행	
기존 마케팅 방식	500,000	80,000	420,000	1,481 (100%)	654	827	266 (100%)
캠페인 효과 기반 타겟 마케팅		80,000	420,000	1,661 (112%)	998	663	446 (168%)

*집단 C에서의 앱 로그인 고객수: 486명
이를 기반으로 추정된 자연반응 고객수: 1215명

캠페인 결과

■ 타겟팅 방식別 캠페인 효과 비교

타겟팅 방식	전체 고객 수 (명)			앱 로그인 고객 수 (명)			증분 고객 수 (명)
	전체	캠페인 수행	미수행	전체	캠페인 수행	미수행	
기존 마케팅 방식	500,000	80,000	420,000	1,481 (100%)	654	827	266 (100%)
캠페인 효과 기반 타겟 마케팅		80,000	420,000	1,661 (112%)	998	663	446 (168%)

*집단 C에서의 앱 로그인 고객수: 486명
이를 기반으로 추정된 자연반응 고객수: 1215명

- 기존 방식과 동일한 고객수에 대해 마케팅을 진행했음에도 불구하고, 전체 반응 고객 수를 더 많이 얻을 수 있었음
- 대조군을 따로 설정함으로써 각 캠페인으로 인한 증분 고객 수를 정확하게 얻을 수 있음

- 활용 방안:
  - 기존 마케팅 방식을 그대로 대체
  - 기존 마케팅 방식보다 대상 고객 수를 더 축소하면서 효과는 유지

질의 응답

obzen
Action on Knowledge