HW13 Mini Project

2021321148
Insik Cho

<초록>

주어진 데이터는 Starbucks 데이터로, 14280명의 고객에 대한 event offer의 결과에 대한 데이터입니다. 이를 바탕으로, 1. 고객의 offer 확인에 영향을 주는 매체 , 2. 고객의 offer 이행에 기여하는 요소 파악을 목적으로 했습니다. 이를 바탕으로, 고객이 offer를 확인하는 시간을 줄이기 위해서는 social 매체를 이용하는게 필요하며, 고객이 offer를 이행하는데 결정적인 변수는 고객의 멤버십 가입기간임을 확인할 수 있었습니다. 이를 바탕으로, 회사 측에서 광고 전략을 세워야 함을 제시합니다.

<목차>

1. 서론 및 데이터 전처리

□ 주어진 데이터는 Starbucks 데이터로, 14280명의 고객에 대한 event offer의 결과에 대한 데이터입니다. 이를 바탕으로, 여기서 뽑아낼 수 있는 함의로 다음 세가지를 선정했습니다.

1. 고객의 offer 확인에 영향을 주는 매체

주어진 데이터에서 파악해야 하는 것 중 하나로 해당 부분을 고려했습니다. 고객이 offer를 확인을 해야지, 고객이 offer를 이행하여 회사의 수익 증진에 기여할 수 있는 만큼, offer를 확인이 큰 매체를 파악하는 게 필요하다고 생각했습니다.

2. 고객의 offer 이행에 기여하는 요소

고객의 offer 이행에 기여하는 요소를 파악할 시, 회사 측에서 고객이 매장에 방문할 유인을 높일 수 있을 것이라 생각했습니다. 이를 고려하여, 주어진 변수 중 difficulty, reward, duration 이 offer 이행에 기여할 것이라 예상해 이를 바탕으로 데이터 분석을 했습니다.

□ 데이터 전처리
o email은 모든 고객에게 전달되었으므로, 결과 도출에 영향을 미칠 수 없으므로 제거했습니다.

o membership 가입 관련으로 가입날짜(member_year,month,day), 가입이후 지난 날,년(members_since_in_days, years)의 5가지 변수가 있습니다. 이들의 상관계수가 높게 나타나고, 여러 변수가 투입되는 것이 결과해석에 유의미한 차이를 주지 못 할거라 생각하여 members_since_in_days를 제외한 변수를 모두 제거했습니다.

o 그 외 데이터마다 분포, 이상치 등을 확인했으나 크게 문제가 없어서 그대로 진행했습니다.

o 모든 분석에서 decision tree의 경우 scaling을 적용하지 않았고, 그 외 방법론에서는 StandardScaler를 적용하였습니다.

2. 고객의 offer 확인에 영향을 주는 매체

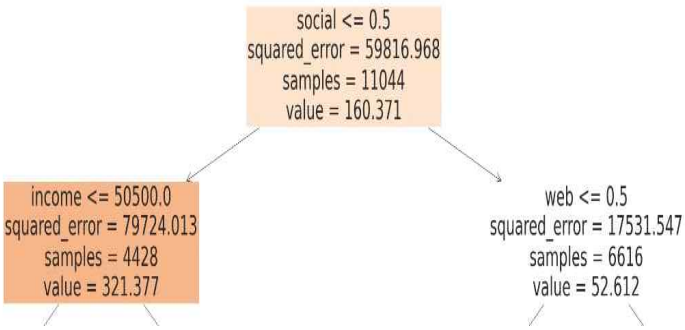
□ 고객이 offer를 받고, offer를 확인할 때 까지의 시간(offer received - offer viewed)를 종속변수로, 나머지 변수를 독립변수로 사용하여 decision tree로 이에 영향을 미치는 변수를 파악했습니다. 이때, offer를 보지 않은 경우, test에 소요된 최대 시간인 600시간을 소모한 것으로 처리했습니다. 따라서 연속형 종속변수를 사용하므로, 회귀분석을 진행했습니다.

매체 이외에도 모든 변수를 독립변수로 포함하였지만, view의 특성상 소득 정도를 제외하고는 매체가 가장 큰 영향을 끼칠 것이라 예상하였습니다.

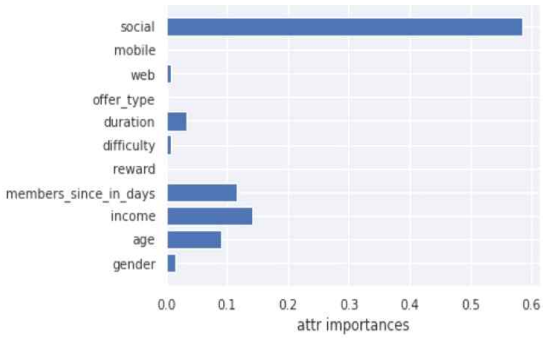
	event	time	offer_id	customer_ids
37	offer received	0	3	4
38	offer received	168	1	4
39	offer viewed	168	1	4
40	offer received	336	9	4
41	offer viewed	348	9	4
42	offer completed	456	9	4

	taken_times
37	600
38	0
40	12

□ 분석을 위해서는 DecisionTreeRegressor를 이용하였고, Bayesian Optimization을 통해 hyperparameter tuning을 진행했습니다. 이를 통해 얻을 수 있는 tree shape과 feature importance는 다음과 같습니다.
(*tree의 경우, 가장 위의 3개 부분만 스크린샷을 찍었습니다.)



<Decision tree shape>



<Decision tree feature importance>

트리에서 첫 번째 분기점이 social이라는 점(social=1이면 평균 view 시간이 감소)과 feature importance 상에서도 social이 가장 큰 영향력을 주는 것으로 나타나는 것을 볼 때, 고객의 offer 확인률을 높이려면 social 매체를 이용해야 하는 것으로 생각됩니다.

□ MAPE비교

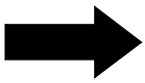
```
Pruned Tree      : 6.6513395664324216e+16
Default Tree     : 7.056087961606516e+16
Random Forest    : 6.767632259601913e+16
```

Pruned tree가 default tree 또는 random forest 보다 더 낮은 mape를 나타낸다는 측면에서, pruning이 잘 이루어짐을 확인할 수 있습니다.

3. 고객의 offer 이행에 기여하는 요소

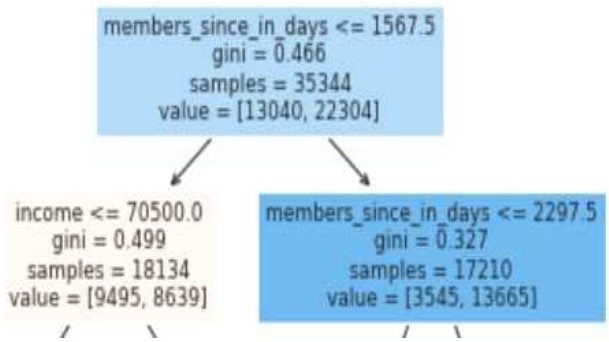
□ 데이터에서 offer_type=2 (informational)인 경우를 제외하고, BOGO 또는 discount인 경우 completed이면 값 1을, completed이 없으면 값 0을 부여한 변수를 생성하였습니다. 이원종속변수를 사용했으므로, 분류분석을 진행했습니다.

	event	offer_id	offer_type	customer_ids
37	offer received	3	2	4
38	offer received	1	1	4
39	offer viewed	1	1	4
40	offer received	9	1	4
41	offer viewed	9	1	4
42	offer completed	9	1	4

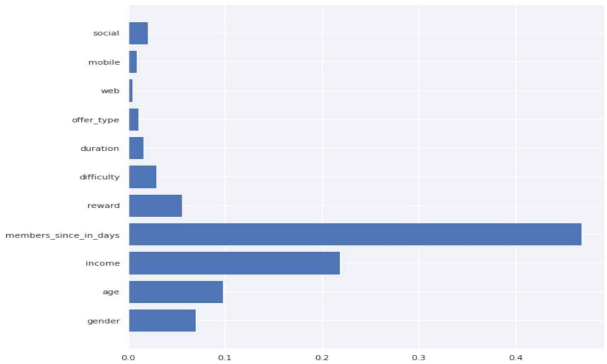


	offer_id	offer_type	customer_ids	completions
38	1	1	4	0
40	9	1	4	1

□ 분석을 위해서는 DesionTreeClassifier를 이용하였고, Bayesian Optimization을 통해 hyperparameter tuning을 진행하였습니다. 이 외에도, 변수의 영향력 체크를 위해 Logistic Rgression도 추가로 진행해보았습니다.



<decision tree-prunes>



<decision tree- feature importance>

앞서 이행여부에 영향을 주는 요소로 difficulty, reward, duration이 있을 것으로 예상했지만, offer를 받는 개인의 가입기간, 소득에 더 큰 영향을 받는 것으로 나타났습니다.

단, difficulty, reward, duration 역시 영향을 주기는 하며, logistic regression을 통해 이를 검정 시,

	coef	std err	z	P>|z|
const	-2.9468	0.217	-13.603	0.000
gender	-0.5096	0.024	-21.305	0.000
age	0.0040	0.001	5.564	0.000
income	2.291e-05	6.22e-07	36.834	0.000
members_since_in_days	0.0012	3.11e-05	39.561	0.000
reward	0.0099	0.017	0.595	0.552
difficulty	-0.1380	0.013	-10.774	0.000
duration	0.1597	0.014	11.543	0.000
offer_type	0.5412	0.093	5.789	0.000
web	0.1124	0.053	2.102	0.036
mobile	-0.4367	0.107	-4.094	0.000
social	0.3253	0.035	9.310	0.000

좌측의 검정결과로부터 reward를 제외한 다른 변수들은 유의미한 영향을 주나, reward의 영향에 대해서는 귀무가설을 기각할 수 없다는 것을 알 수 있습니다.

Feature extraction으로 확인하기로는 reward가 영향력을 지니나, Logistic regression의 결과로 인해 그 의미에 대해서는 의심을 가질 수 있습니다.

<Logistic regression>

□AUC비교

```
default tree AUC:  0.64755
pruned tree AUC: 0.6812089
randomforest AUC: 0.689991
logistic  AUC:  0.5
```

AUC 값을 통해 pruned tree가 randomforest 못지 않은 예측력을 보임을 확인할 수 있습니다. 추가적으로, Logistic 회귀가 해당 모델에서 적합하지 않음을 확인할 수 있습니다. 이 점을 고려하여, 변수 해석에 있어서 decision tree의 해석결과에 조금 더 집중해도 괜찮을 것 같습니다.

4. 결론

1. 고객의 offer 확인에 영향을 주는 매체로는 Social 매체가 더 크다는 것을 확인할 수 있었습니다. View가 이루어져야지 Completion이 발생하는 만큼, 기업 입장에서는 social 매체를 통한 offer의 홍보가 이루어져야 한다고 생각합니다.

2. 고객이 offer를 이행하는데 있어서 기업이 부과하는 reward, difficulty, duration 역시 영향을 미치긴 하지만 이보다는 멤버십 가입기간, 개인의 소득이 더 큰 영향을 주는 것으로 나타났습니다.

따라서, reward나 difficulty의 조절을 통한 offer의 공급 역시 중요하지만, 이용 고객의 특성(멤버십 가입기간이 길고, 소득이 낮고, 성별이 여자일 시)에 따라서 offer 제시를 더 많이 할 지, 적게 할 지를 결정하는 것이 더 중요하다고 생각합니다.