

2021 D&A

구매내역과 클릭스트림 데이터를
이용한 고객의 성별, 연령대
예측 경진대회

교수님저희싫어하시조
최종 2위



목차

01

Feature 설명

02

Scaling

03

One-hot
encoding

04

W2V

05

Feature
Seletion

06

모델링

07

하이퍼파라미
터 튜닝

08

Predict

Feature

❖ 구매피처

- [총구매액] [구매건수] [평균구매액] [최대구매액][최소구매액]
- [고가상품구매율]

❖ 방문피처

- [주말방문비율][내점일수][구매빈도][세션접속일수]

❖ 검색피처

- [총페이지조회건수][상위키워드검색합]

❖ 접속시간피처

- [총접속시간대비 최소][총접속시간대비 최대][총접속시간대비 평균]

Scaling

- Min-Max, Standard scaler, Robust scaler

- 시도 해보았으나 그닥,,

- ❖ 결론 : power transform -> 거듭제곱 변환

One-hot Encoding

❖ 범주형 피쳐

- 다른 범주형 피쳐들은 과적합의 요소가 있었음
- 아래의 두 조합이 가장 좋은 결과를 보였다

◦ [주구매경로]

◦ [구매지역]

W2V

❖ 하나씩 넣어보면서 성능 비교를 해보았다.

- PD_NM : 상품명
- CLAC2_NM : 상품중분류명
- CLAC3_NM : 상품소분류명
- PD_ADD_NM : 구매한 상품의 추가 정보
 - 사이즈 정보가 들어있어서 w2v했음
- PD_BRA_NM : 구매한 상품의 브랜드
- KWD_NM : 검색창에 입력한 검색 키워드
- PD_C : 구매한 상품 코드

W2V

- Oversample : 10
- Vector_size (size) : 60
 - 가장 좋은 성능을 보임
- 해당 피쳐 w2v결과 결측치가 발생하는 항목이 있었음
 - Drop 보다 fillna(0)이 더 좋은 성능을 보임

Feature Selection

- LGBM 기반 selection
 - 458 -> 105 개의 feature로 select 됨
 - 성능은 그닥,,
 - Select percentile
 - $p = 95$, feature는 140개로 select 됨
 - 성능은 그닥,,
- ❖ 결론 : 다 넣자 -> 최고성능 (?)

모델링

- LGBM, CATBoost, Ensemble, Stacking 등 여러 모델들을 시도해봄
- 여러 모델들 중 LGBM과 CATBoost가 가장 좋은 성능을 보임
 - 위의 두 모델로 앙상블을 진행하였으나 성능은 그닥,,
 - 이중 스택킹을 진행하였으나 성능은 그닥,,
- ❖ 결론 : 단일모델 (LGBM, CATBoost)
 - 기본에 충실해보자!

Hyperparameter tuning

- LGBM

- BayesianOptimization

- Skf : `n_split = 5`

- Crossvalidation score : `neg_log_loss, cv = skf`

- BayesianOptimization : `init_points = 5, n_iter = 10`

- > 총 15번 탐색

Hyperparameter tuning

- CATBoost

- iterations = 10000, learning_rate = 0.01, eval_metric = 'MultiClass'

- Skf : n_splits : 10

- > 지나친 소요 대비 낮은 성능

- > 오히려 튜닝을 전혀 하지 않은 기본 모델이 좋았음.....

Predict

❖ public score

➤ LGBM (튜닝 ㅇ) : 0.32060

➤ CATBoost (튜닝 x) : 0.3217

느낀점

- EDA의 중요성
- 많은 시도를 해봤던 것이 좋은 성능을 낼 수 있게 함
- 의외로 기대했던 것들은 좋은 성과를 내지 못함
-> 기본적인것들이 더 좋았었음
- 다른 경진대회 나가서도 많은 시도를 해봐야겠다고 느낌
- 분류에선 boosting 계열의 모델들의 성능이 좋다고 생각함