수상작리뷰
[Private 1st, 0.70019] CatBoost + 2-way optimization - DACON

[신용카드 고객 세그먼트 분류 AI 경진대회]

◈ 주제

👉 금융 데이터를 바탕으로 신용카드 고객 세그먼트 분류하는 AI 알고리즘 개발

◈ 데이터

- train.csv :고객별 금융활동 데이터. 각 정보 내 parquet파일로 구성 / 회원정보, 신용정보, 승인매출정보, 청구입금정보, 잔액정보, 채널정보, 마케팅정보, 성과정보, 고객 세그먼트(A~E)
- test.csv :고객별 금융활동 데이터. 각 정보 내 parquet파일로 구성 / 회원정보, 신용정보, 승인매출정보, 청구입금정보, 잔액정보, 채널정보, 마케팅정보, 성과정보
- sample_submission.csv : 샘플별 고유 ID , 예측한 세그먼트
- 데이터명세.xlsx : 고객 금융활동 데이터에 대한 명세

◈ 코드 리뷰

(1) 라이브러리 설치, 불러오기, 패키지 임포트
- sys, catboost, optuna, pandas, numpy, sklearn…

(2) 데이터 정리
- train, test 데이터세트가 폴더별로 나뉘어져있으므로 train, test 최종 데이터세트로 병합 (base_clean_train, base_clean_test)

(3) BASE 모델링 # A100 GPU 사용.
- 🔷 Catboost + Optuna (단일 valid split) 예측 및 저장
- 🔷 Catboost + 10-Fold CV 예측 및 저장

(4) VIP 분류 모델 # L3 GPU 사용.
- **A, B 예측을 정확히 하는것이 중요하다고 판단 → AB클래스만 학습하는 모델 별도 생성)**
  - 🔷 AB 별도 Catboost 단일 모델 + 10-Fold CV 예측 및 저장

(5) 세그먼트 수정 및 최종 예측 저장
- VIP 분류 모델을 활용해 세그먼트를 AB 로 수정 수행
- 최종 결과 예측 및 저장

◈ **배울점**

- 피처 개수가 많고 타켓값이 5개인만큼, 최대한 다양한 알고리즘 방법을 수행함.
- VIP 모델 분류가 중요하다는 사실을 판단하는 능력.
- 하이퍼 파라미터를 튜닝 할 때, 무작정 수행하지 않고 과적합을 방지하는 선에서 optuna를 수행한 점.