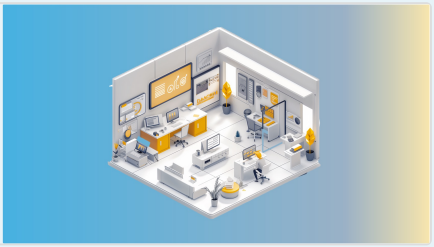
(예시) 웹 로그 기반 조회수 예측 해커톤

💡 **[DACON] 웹 로그 기반 조회수 예측 해커톤**

웹 로그 기반 조회수 예측 해커톤

출처 : DACON - Data Science Competition

https://dacon.io/competitions/official/236226/overview/description



데이터

- **train**: 한 세션에서 발생한 다양한 정보를 기록한 `웹 로그`
  - `sessionID` : 세션ID
  - `TARGET` : 예측 목표 = 세션에서 발생한 총 조회수
- **test**: 한 세션에서 발생한 다양한 정보를 기록한 웹 로그

코드 흐름

(1) EDA

- data 파악
- target 분포 확인
- unique한 값의 개수가 20개 이하인 컬럼 선택 ⏩ target과의 관계 파악
- `bounced=1` 일 때 항상 `TARGET=1`인 것을 관찰 ⏩ `bounced=0` 인 데이터만 training

(2) Feature Engineering & Processing

- 결측값 존재하는 컬럼, 불필요한 칼럼, 중복 데이터 → drop
- `ChatGPT`를 이용한 파생변수 생성
  1. device + quality → `기술적 환경`

2. continent + subcontinent → `지역`
3. traffic source + traffic medium → `트래픽 소스`

(3) Modeling: Catboost Regressor

- `optuna`로 hyperparameter 튜닝 후 학습 진행
- Stratified KFold를 사용한 cross-validation
- target의 `nunique=117`이라서(아주 많지는 않아서) StratifiedKFold 사용할 수 있겠다고 판단
  - `userID`를 groups 변수에 할당 → dataset 분할 후 해당 컬럼 삭제

  ⭐ 이 때 rmse크게 감소
- Soft voting: fold마다 학습 진행 → 결과를 평균내서 적용

(4) 후처리

- EDA에서 관찰한 TARGET 범위: 1~386

  ⏩ 모델 예측 결과 중 1 이하인 값 → 1로 변환

차별점, 배울점

- chatGPT를 활용해 새로운 파생변수 생성한 점이 새롭다.
- 회귀 문제이긴 하지만, TARGET의 unique 값 개수를 살펴보고 StratifiedKFold를 적용한 점이 차별화된 포인트이다.
- EDA를 기반으로 한 학습 데이터 처리나 예측값의 후처리도 중요하다.