# CTR Prediction

카카오 서비스들의 여러 게재 지면에서 다양한 광고가 노출되고 있으며, 광고의 CTR을 예측하는 것은 데이터 밸류팀이 해결해야하는 중요한 비즈니스 문제입니다. 주어진 광고 노출-클릭 데이터를 이용해 테스트 데이터의 CTR을 예측하세요.

첨부파일들 (training_00, test_00) 에  여러 게재 지면에 노출된 디스플레이 광고(DA)의 노출-클릭 데이터가 담겨있습니다. 데이터는 학습 데이터 1개, 테스트 데이터 1개의 csv 파일로 구성되어 있습니다. 테스트 데이터는 타겟 필드가 없으며, 테스트 데이터의 row별 CTR을 0~1 사이의 실수로 예측하시면 됩니다. 테스트 데이터의 평가는 log-loss 메트릭을 사용합니다.

분석 방법 및 모델링에 사용하는 프로그래밍 언어, 모델, 라이브러리 등은 모두 자유이며, 분석 과정부터 예측 결과까지의 내용을 노트북 등의 편하신 양식으로 정리하여 결과 파일과 함께 제출하시면 됩니다.

과제는 아래의 항목들을 참고해 작성하여 주시기 바랍니다.
- 과제 해결과정에서 참고하신 논문 혹은 문헌이 있다면 파일 내 명시
- 과제 해결과정에서 의사결정을 위한 가정, 혹은 선택의 이유를 파일 내 명시
- 과정에서 성공적이지 않은 시도도 의미가 있었다면 파일내 기록

제출해야 할 항목은 두 가지입니다.
1. 테스트 데이터의 예측값: index, CTR(소수점 여섯째 자리) 두 개의 컬럼을 가진 csv 파일
2. 분석 및 모델링 과정을 설명할 수 있는 자료: ipython/R 노트북 파일 등의 자유 양식

해당 데이터의 필드는 아래와 같은 순서로 되어 있으며, 다음은 각 필드의 설명입니다.
-index: row id
-userid: 유저 id
-gender: 추정된 성별
-age: 추정된 나이
-slotid: 광고 게재 지면
-device: 디바이스 타입
-connection_type: 네트워크 연결 타입
-activity1: 유저의 활동 관련 정보1
-activity2: 유저의 활동 관련 정보2
-activity3: 유저의 활동 관련 정보3
-interest_list: 유저의 n개 관심사 ('|'로 구분되며 최대 5개)
-interest_count: 유저의 총 관심사 개수
-ad_feature1: 광고이미지 id
-ad_feature2: 광고소재 id
-ad_feature3: 광고주 id
-ad_feature4: 광고그룹 id
-timestamp: 한국 기준 광고 노출 시각
-click: 타겟 필드, 노출된 광고 클릭 여부

데이터 사이즈는 아래와 같으며 첨부파일들의 크기가 같은지 최종 확인바랍니다.
-학습 데이터 총 row 개수: 1,527,731
-테스트 데이터 총 row 개수: 209,106