



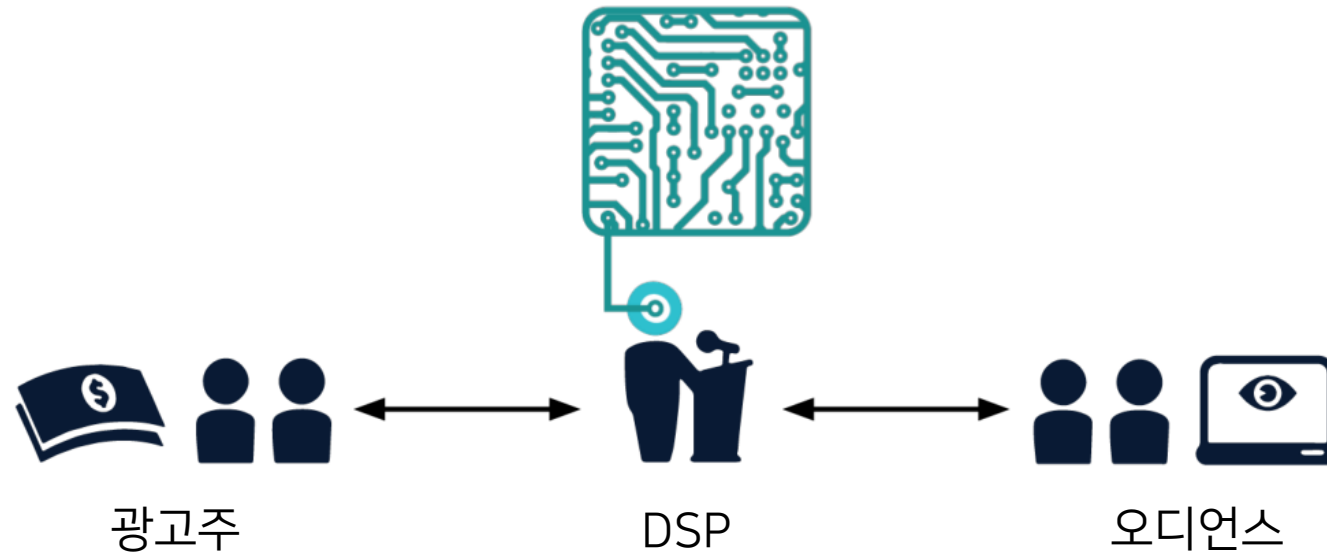
2019 IGAWorks BIG DATA Competition 대회 설명

2019.12.16.

개요

대회주제

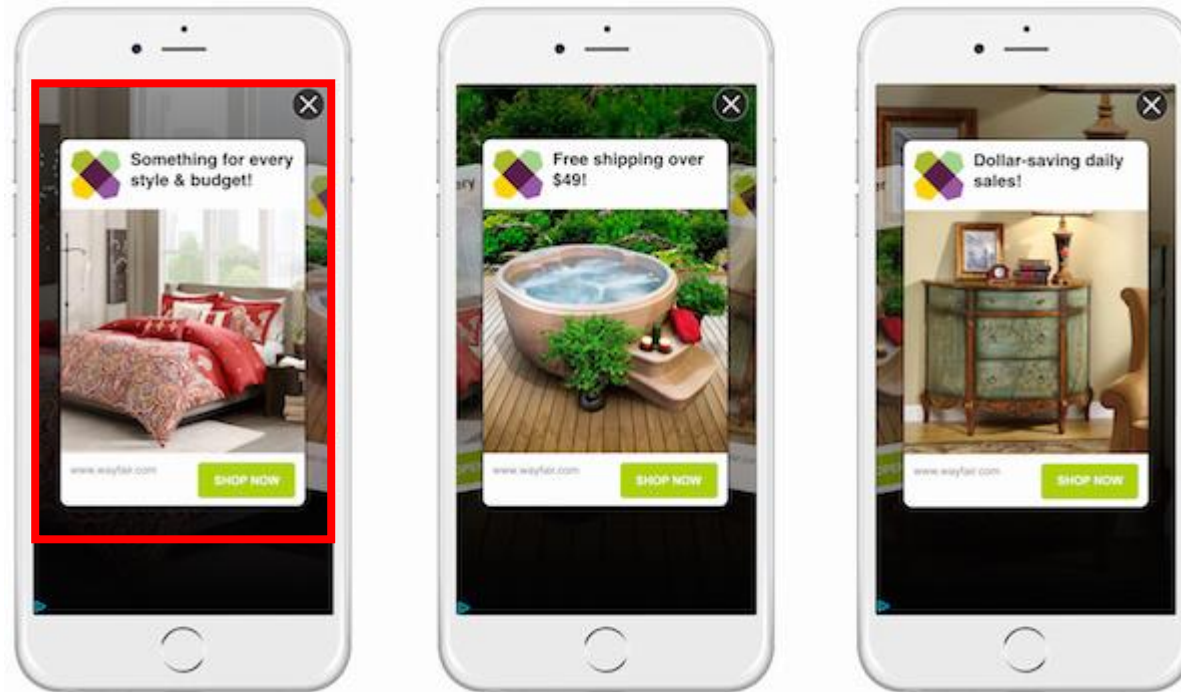
- 사용자와 방문한 페이지가 주어지면 주어진 광고를 클릭할 확률 추정 모형 개발
 - 온라인 광고 시장은 비식별 데이터를 사용하여 오디언스 프로파일링
 - 개인화 추천을 통해 광고를 게재함



- 광고주와 인터넷 오디언스 사이의 최적 연결을 만드는 알고리즘 설계

분석대상


- TradingWorks(트레이딩웍스)
 - 월 500억 건 이상의 모바일 데이터
 - 3천 개 이상의 매체
 - Unique User 4천만명 규모



분석대상


- TradingWorks(트레이딩웍스)의 특징
 - 모바일 App에 특화
 - A.I 예측 마케팅은 예측된 오디언스별로 개인화된 크리에이티브를 자동 생성
 - 글로벌 주요 매체에 Auto Bidding 시스템을 통해 최적의 순간, 최적의 인벤토리에 노출

1. Only Mobile App User



- ✓ 100% Mobile App Inventory

2. 글로벌 TOP Publisher 연동



- ✓ 4만여 App으로부터 월 300억건 광고 요청 처리
- ✓ 국내 / 글로벌 100여곳의 미디어 파트너 연동

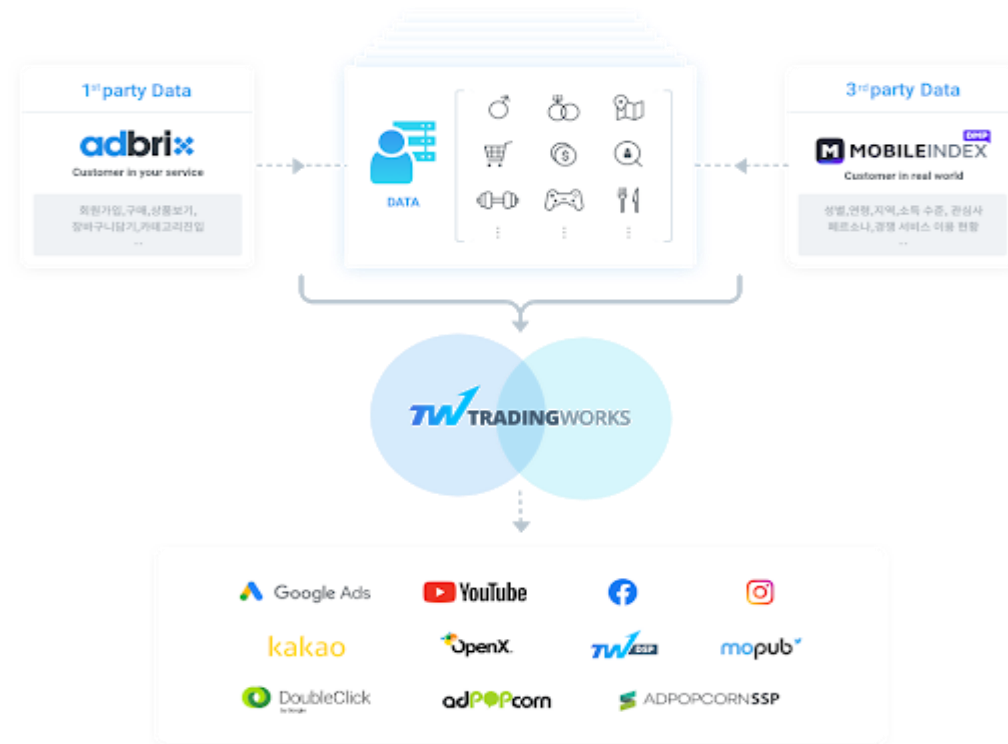
3. 국내외 모든 유저 Coverage



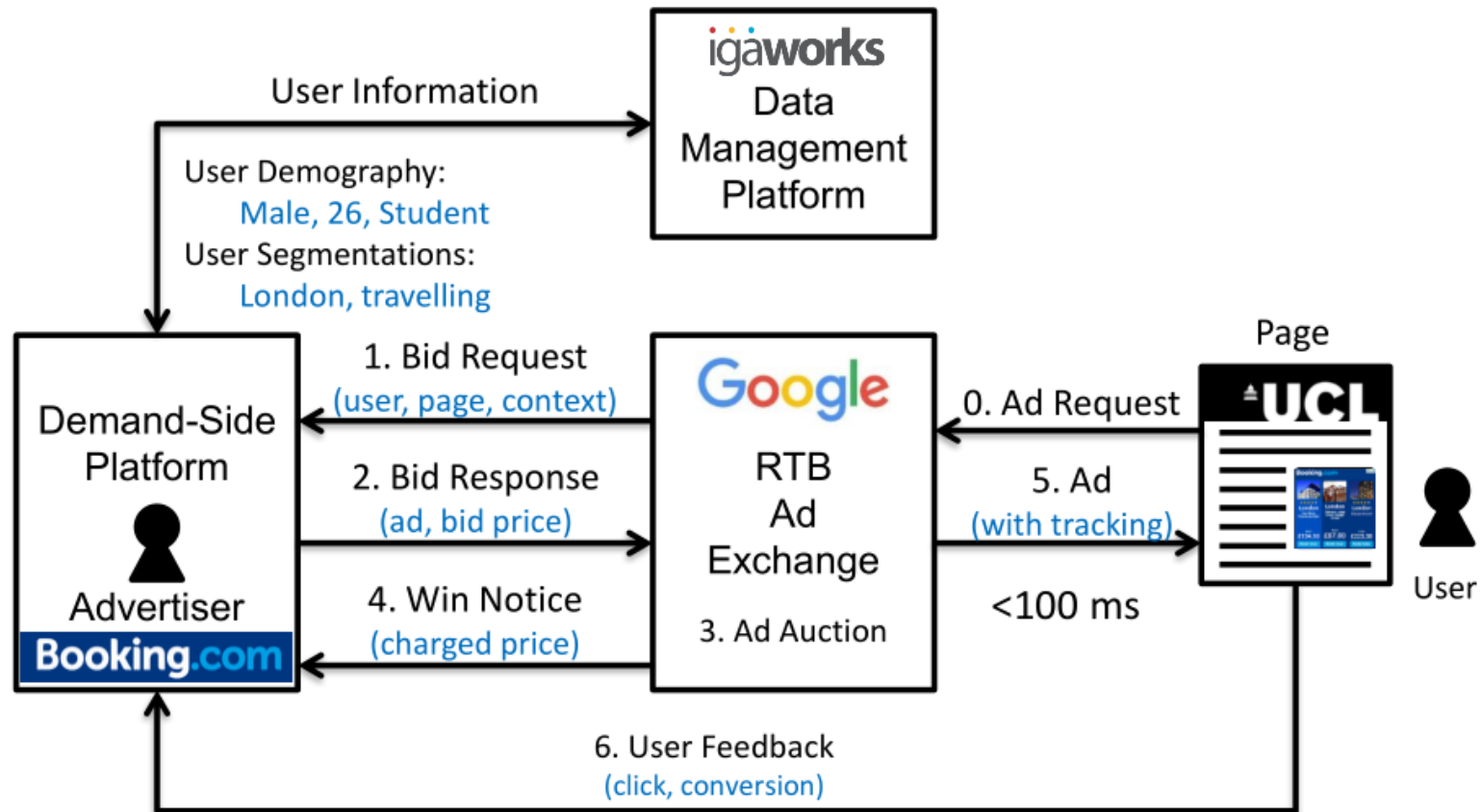
- ✓ 국내 4,500만명의 Mobile User Coverage
- ✓ 글로벌 200여개 국가 Coverage

광고 데이터 분석

- 오디언스의 모바일 행동을 관찰 가능
- 노이즈를 포함한 많은 정보들 속에서 연령, 성별, 관심사, 지역정보, 라이프스타일, 구매력 등을 분석하고 모델링
- 모델링의 결과가 건 당 100ms 의 실시간으로 반응하기 때문에 매출과 직결되어 있음



RTB 경매 방식



CTR Prediction

- CTR Prediction은 아래와 같은 정보를 이용해 클릭 여부를 예측함
- 여기서 나온 CTR을 바탕으로 입찰 전략을 통해 RTB 경매에 입찰할 입찰가를 결정함

- Date: 20160320
- Hour: 14
- Weekday: 7
- IP: 119.163.222.*
- Region: England
- City: London
- Country: UK
- Ad Exchange: Google
- Domain: yahoo.co.uk
- URL: <http://www.yahoo.co.uk/abc/xyz.html>
- OS: Windows
- Browser: Chrome
- Ad size: 300*250
- Ad ID: a1890
- User tags: Sports, Electronics



Click (1) or not (0)?

Predicted CTR (0.15)

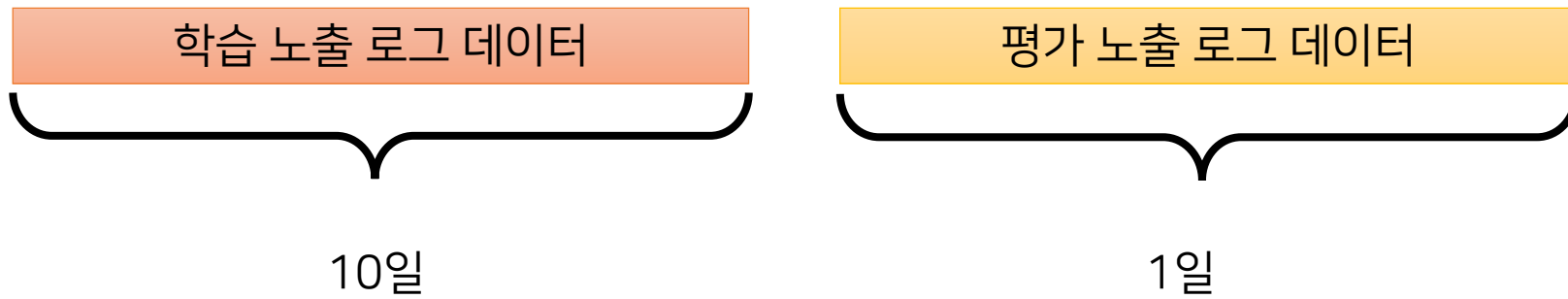
문제의 의도와 목표

- 정확한 클릭 확률 추정은 입찰가 결정에 도움을 줌

문제 설명

데이터 구성

- 10일치 노출 로그 데이터를 이용하여 모델 학습
- 그 다음날 1일치 노출 로그 데이터에 대하여 실제 클릭이 난 로그를 구별



데이터 구성

- 학습 및 평가 데이터 구성 방식 및 규모
 - Train 데이터 : 5500000 row
 - Test 데이터 : 550000 row
- 데이터 종류별 CSV 파일 제공 (총 3종)
 - 예측 대상은 개별 row 기준

데이터세트	데이터 내용
train.csv	학습 기간 노출 로그 데이터
test.csv	평가 기간 노출 로그 데이터
audience_profile.csv	오디언스 관련 정보 모음

데이터 구성

- train.csv, test.csv
 - 노출 로그 데이터

변수	설명
click	클릭 여부, test에는 없음
event_datetime	로그 발생 시간
ssp_id	SSP 아이디
campaign_id	캠페인 아이디
adset_id	광고 아이디
placement_type	광고 타입
media_id	미디어 아이디
media_name	미디어 한글이름
media_bundle	미디어 앱명
media_domain	미디어 도메인
publisher_id	매체사 아이디
publisher_name	매체사 이름
device_ifa	기기 구별 아이디
device_os	기기 OS
device_os_version	기기 OS 버전
device_model	기기 모델명
device_carrier	기기 통신사
device_make	기기 제조사
device_connection_type	기기 연결방식
device_language	기기 언어
device_country	기기 국가
device_region	기기 지역
device_city	기기 도시
advertisement_id	광고주 아이디

데이터 구성

- audience_profile.csv
 - 오디언스 관련 정보 모음

변수	설명
device_ifa	기기 구별 아이디
age	연령 (추정)
gender	성별 (추정)
marry	기혼여부 (추정)
install_pack	설치된 앱 정보
cate_code	IGAW 카테고리별 등급
predicted_house_price	자산 가격(추정)
asset_index	자산 지수(추정)

평가 방법

- 예측 성능+재현성 테스트+서류 심사
- 예측 성능
 - 참가팀이 제공한 클릭 확률에 따른 logloss를 계산함 $-(y \log(p) + (1 - y) \log(1 - p))$
- 재현성 테스트
 - 모델링 단계별 소스 코드 및 관련 자료 제출
 - 제출한 코드를 이용해 최종 예측 결과가 얼마나 쉽고 정확하게 재현되는지 측정
 - 모델 학습과 평가 시에 걸린 시간도 평가 (1건 당 예측 시간으로 측정)
- 서류 심사
 - 탐색적 자료 분석, 전처리, 모델링 및 튜닝 등 전체 분석 과정에 대한 설명 문서
 - 체계적이고 논리적인 접근, 적절한 시각화

평가 방법 상세

재현성 테스트

- 모델링 단계별로 모듈(전부 소문자)을 구분하여 소스 코드 및 관련 자료 저장
 - 아래 모듈 파일을 모아 "팀이름.zip"로 압축하여 제출

모듈명	모듈 내용
preprocess	원본 데이터 전처리 코드
model	최종 모델 학습용 코드
predict	테스트 데이터와 모델을 이용하여 최종 답안지를 생성하는 코드

평가 방법 상세

- preprocess 모듈 구성
 - 전처리 코드 : 원본 데이터 파일들을 불러들여, 최종 모델의 input이 되는 데이터 파일을 저장하는 코드
 - input : 원본 데이터 파일
 - output : 최종 모델의 input이 되는 데이터 파일
 - 데이터 파일 명명 규칙 : 팀이름/데이터세트_preprocess_숫자.확장자

원본 파일

- train.csv
- test.csv
- audience_profile.csv



preprocess.py



preprocess 결과

- train_preprocess_1.csv
- test_preprocess_1.csv

평가 방법 상세

- model 모듈 구성
 - 모델링 코드 : preprocess 모듈로 나온 데이터를 이용하여, 최종 답안지 생성에 사용되는 모델 객체를 저장하는 코드
 - input : preprocess 코드에서 나온 최종 모델의 input이 되는 데이터
 - output : model 폴더에 최종 답안지 생성을 위해 사용되는 모델 객체
 - 모델 객체 : 모델링 코드를 통해 생성되는 최종 예측 모델

preprocess 결과

- train_preprocess_1.csv
- test_preprocess_1.csv



create_model.py



model 결과

final_model.sav

평가 방법 상세

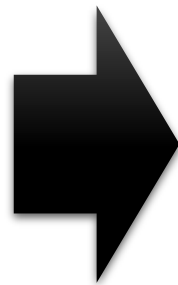
- predict 모듈 구성
 - 예측 코드 : preprocess의 결과 데이터와 model 모듈의 결과 최종 모델 객체를 불러들여 최종 답안지를 생성하는 코드
 - input :
 - 최종 모델의 input이 되는 데이터
 - model 모듈의 결과 최종 모델 객체
 - output :
 - predict/test_predict.csv

preprocess 결과

- train_preprocess_1.csv
- test_preprocess_1.csv

model 결과

- final_model.sav



predict.py



predict 결과

test_predict.csv

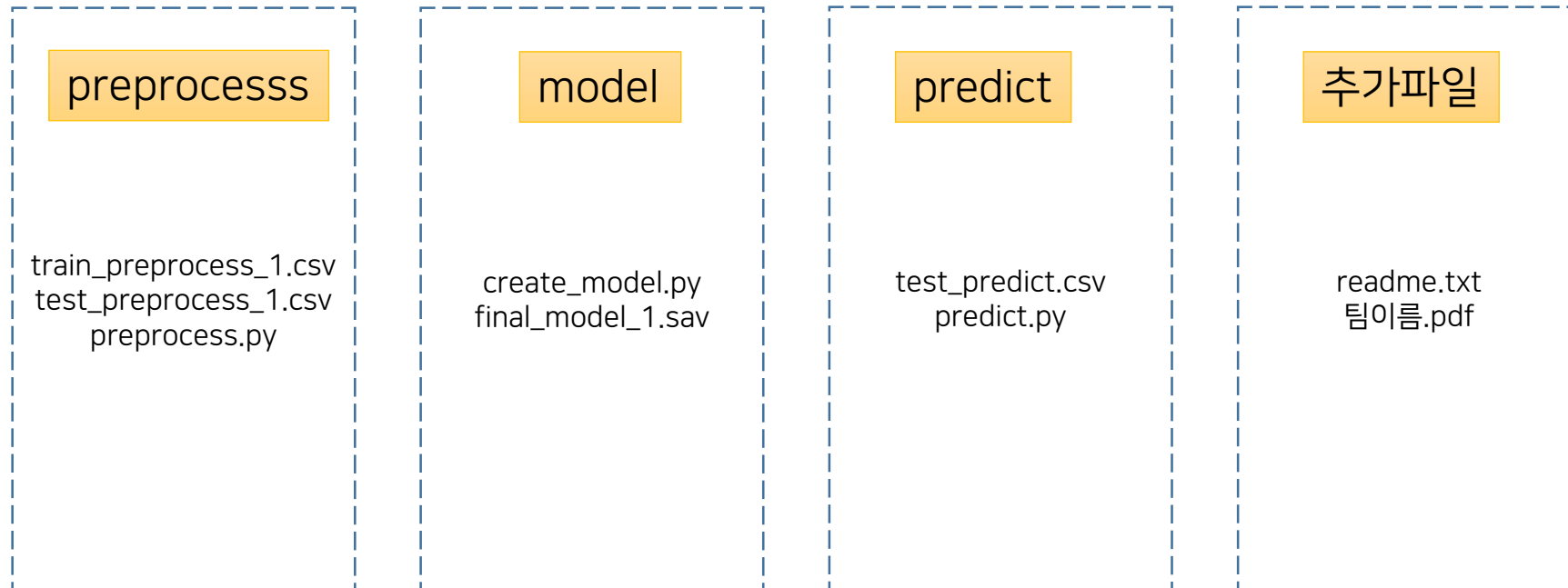
평가 방법 상세

- 추가파일
 - readme.txt
 - 코드 실행에 필요한 패키지/라이브러리/모듈 리스트 및 실행 환경
 - 코드 실행 순서 및 방법
 - 팀이름.pdf
 - 서류 심사용 자료

평가 방법 상세

- 최종 자료 구성 예시
 - python 기준으로 작성된 예시이며 사용 가능한 언어 및 도구 제한은 없음 (단, 상용툴 이용시 평가 불가)

팀이름



평가 방법 상세

- 서류 심사
- 본선 심사 발표 용도
 - 탐색적 자료 분석, 전처리, 모델링 및 튜닝 등 전체 분석 과정을 설명하는 문서
 - 적용 된 학습 알고리즘에 대한 설명 및 모델 해석
 - 논리적이고 체계적인 접근, 이해를 돕기 위한 시각화

자율 평가

- 최종 결과 제출에 앞서 평가 데이터에 대한 예측 성능 확인 및 벤치마킹을 위한 리더보드 제공 (1월 중)
- 어뷰징 방지를 위해 중간 평가는 **전체 평가 데이터의 20%만을 측정**한 결과 제공
- 점수 해킹 및 과도한 트래픽 부하를 막기 위해 지원자 별 **1일 5회로 횟수 제한**
- 자율 평가는 성능 확인 및 벤치마킹을 위해서만 제공되며, 최종 평가에 영향을 끼치지 않음

최종 평가

- 최종 예측 성능은 마지막에 제출한 예측 결과를 기준으로 전체 성능 측정하여 평가
- 예측 성능 및 재현성 테스트 결과로 이후 서류 심사 진행
- 최종 후보 5개 팀 선별하여 본선 심사 진행
 - 참가 인원 및 순위에 따라 서류 심사 대상자 및 본선 심사 대상자 수 변경 가능

수상자 혜택

- IGAWorks 2020년 데이터 부문 채용 시 서류 전형 면제

Q&A

Q1. 데이터셋에 ‘앱 정보’ 변수의 내용이 무엇인지 궁금합니다.

- ‘앱 정보’ 변수는 안드로이드 기기라면, 해당 시점에 기기에 설치 된 앱입니다.

Q2. 변수들이 암호화가 되어있는데 변수들에 대한 추가적인 설명은 얻을 수 없나요?

- 변수들은 전부 명목변수로써, 제공된 데이터셋의 변수명을 보시면 모델링을 하는데 있어 필요한 정보는 충분히 파악하실 수 있습니다.

Q3. 광고 위치가 어디인지 알 수 있나요?

- 하단인지 전면인지 등의 위치는 알 수 없습니다.

Q4. 광고주가 다양하고 매체도 다양한데 클릭할 1건의 데이터에 대한 정보는 지면과 광고주 모두 유니크할 것으로 예상됩니다. 지면, 광고주 정보들을 모두 고려해야 하나요?

- 맞습니다. 로그 자료를 이용해서 진행됩니다. 로그 데이터는 ‘누구’인지는 알 수 없으나 공통적인 값들이 찍혀 있을 테니 그 값들에 유의해서 진행합니다.
예를 들어, 특정 광고주의 클릭 확률을 얼마였다는 것을 학습하여 테스트 셋에 넣어도 문제를 풀 수 있습니다.

Q5. 정리를 하자면, 배너를 클릭할 것인지 아닌지 이 1건의 데이터에 대한 정보를 맞추는 것인가요?

- 10일간의 노출에 클릭 하였는지 아닌지에 대한 여부가 담겨져 있습니다. 이 학습 자료를 통해 노출과 클릭 사이에 확률적 모형을 개발한 뒤에 테스트 셋에는 그 클릭 여부가 없기 때문에 이 값을 맞추시면 됩니다.

Q6. 모형을 학습할 때 앙상블이 효과적이지 않다는 말씀이신가요?

- logloss를 줄일 수 있다면 어떠한 모형을 사용하셔도 무방하지만 모형의 정확도와 학습 시간을 종합적으로 평가하기에 이에 적합한 모형을 만드시면 됩니다.
- 한 건의 request에 대해 광고를 3개, 100개 등 다양하게 보여줄 수 있으나 이 결정이 100ms에 결정되어야 합니다.
- 로그 로스를 낮추는 것 뿐만 아니라 시간을 줄일 수 있는 것을 고려하는 것이 필요합니다.

Q7. Train set 1개, Test set 1개가 제공되는데 내부적으로 따로 평가를 하는 데이터 세트가 있나요?

- 내부적으로 평가할 수 있는 private 데이터를 따로 두지 않습니다.
test set이 전체 평가 자료이고, 리더보드로 평가받을 수 있는건 20% 뿐입니다. 추후 재현성 테스트때는 test set 100%를 이용해 평가하게 됩니다.

Q8. 각자 개인의 컴퓨터로 분석을 진행해야하는지 클라우드로 분석할 수 있는 환경을 제공해주는지 궁금합니다.

- 개인 컴퓨터로 분석을 진행하고 재현할 수 있는 코드를 제출해주시면 됩니다.

Q9. 재현성 테스트는 리더보드에서 확인할 수 있나요?

- 모델 평가만 확인할 수 있습니다. 클라우드를 통해 동시 학습 및 평가를 진행합니다.

Q10. 분석 하려면 도메인 지식이 필요할것 같은데, igaworks 데이터만 가지고 해야하나요? 외부의 카테고리 코드나 새로운 자료를 크롤링 해도 되나요?

- 외부 데이터는 사용 불가입니다.

Q11. 학습 시간이 평가 기준 중 하나인데 컴퓨터의 성능에 따라 측정되는 시간이 다른데 정확한 학습 시간을 확인하는 방법이 있을까요?

- 제출코드를 검증할 때 똑같은 성능의 컴퓨터로 실행하여 시간을 측정하기 때문에 개인 컴퓨터에서 최대한 학습 시간을 줄여보는것이 좋습니다

Q12. 모형의 정확도와 학습 속도를 종합적으로 평가하는 기준이 무엇인지 궁금합니다.

- 재현성 테스트를 거쳐 logloss의 등수와 inference time의 등수를 매긴 뒤, $\text{logloss 등수} * 0.7 + \text{inference time 등수} * 0.3$ 으로 새로 등수를 정하고, 동점일 경우엔 logloss의 낮은 값이 상위가 됩니다.

Q13. 결과, predict 는 0,1,0,1,로 제출하나요?

- logloss를 통해 계산을 하기 때문에 0,1,0,1로 제출하시지 말고 확률값으로 제출하면 됩니다.

Q14. 데이터의 불균형은 일반적인 수준인가요? 10일이라는 데이터가 이벤트 없이 일반적인 경우의 데이터인지 궁금합니다.

- 클릭 , 비 클릭에 대해 서로 다른 전략으로 샘플링하였습니다.

Q15. 코드에 주석을 달지 말고 따로 설명을 텍스트 파일로 만들면 되나요?

- 코드는 재현성 테스트를 위한 용도이며 코드를 각각 'Preprocessing', 'Modeling', 'Predict' 부분으로 모듈화하여 올려주시고 주석을 정리한 파일을 따로 제출해주시면 됩니다.

Q16. 추후에 발표를 하게되면 자료를 만드는 양식이 있나요?

- 자유로운 양식으로 준비해주시면 됩니다.