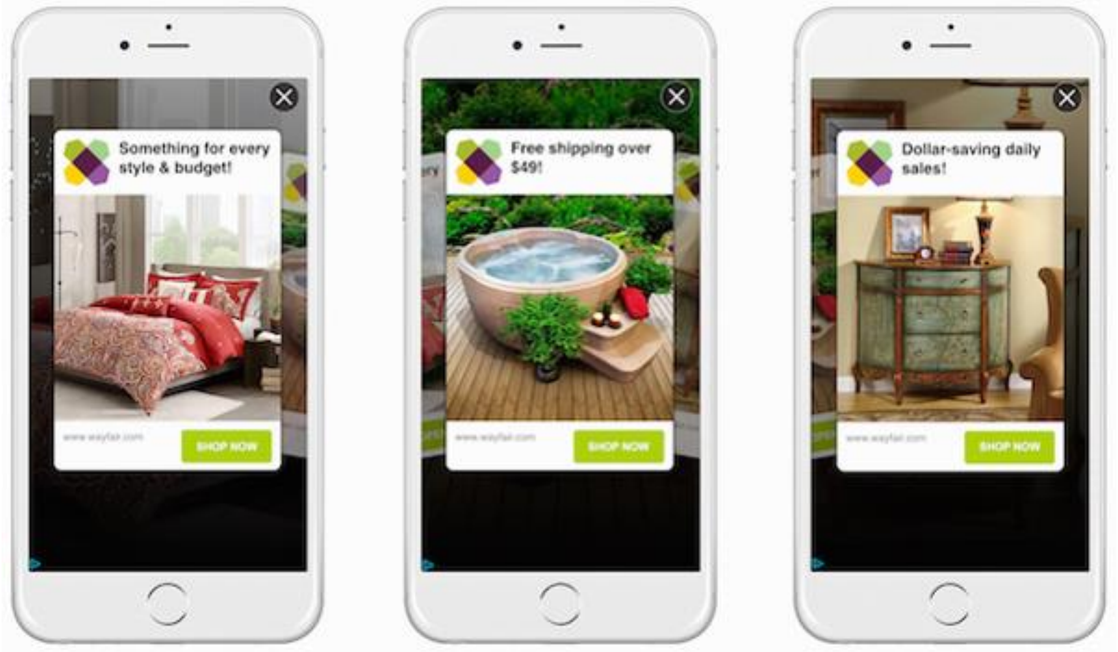
연구 프로젝트 최종 발표

DA 2019312260
김미소

01. 프로젝트 주제

“ 모바일 광고 클릭률 예측 모형 개발 (CTR Prediction) “

- 사용자와 방문한 페이지가 주어지면 광고를 클릭할 확률 추정 모형 개발
- 9일 분량의 모바일 광고 데이터를 활용하여, 다음날 1일치의 개별 노출의 클릭 확률 예측



02. 활용 데이터

- 데이터 소스: IGAWorks
  (2019 IGAWorks big data competition 공모전에서 제공받은 데이터로, 대회측에 대회 이외의 용도로 사용 허가 받음)
- 데이터 크기 및 기간:
  - 광고 노출 로그 데이터: 5,500,000 row (1.43 GB) / 10일치
  - audience 데이터:1,000,001 row (9.04 GB)

[광고 노출 로그 데이터]

변수	설명
click	클릭 여부
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

[audience 데이터]

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)

03. 데이터 전처리 및 EDA – 데이터프레임 병합

- 분석을 위해 두가지 데이터프레임을 병합하는 과정 필요
- 광고 노출 로그 데이터를 메인 데이터프레임으로, audience 데이터를 추가 활용 데이터프레임으로 사용하기 위해 left join 진행 시, audience 데이터 변수에 많은 결측치가 발생하는 문제점 존재
- 데이터의 손실을 감안하더라도 audience 데이터를 보다 정확하고 적절하게 활용하고자 inner join 실시

[광고 노출 로그 데이터]

변수	설명
click	클릭 여부
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

[audience 데이터]

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)

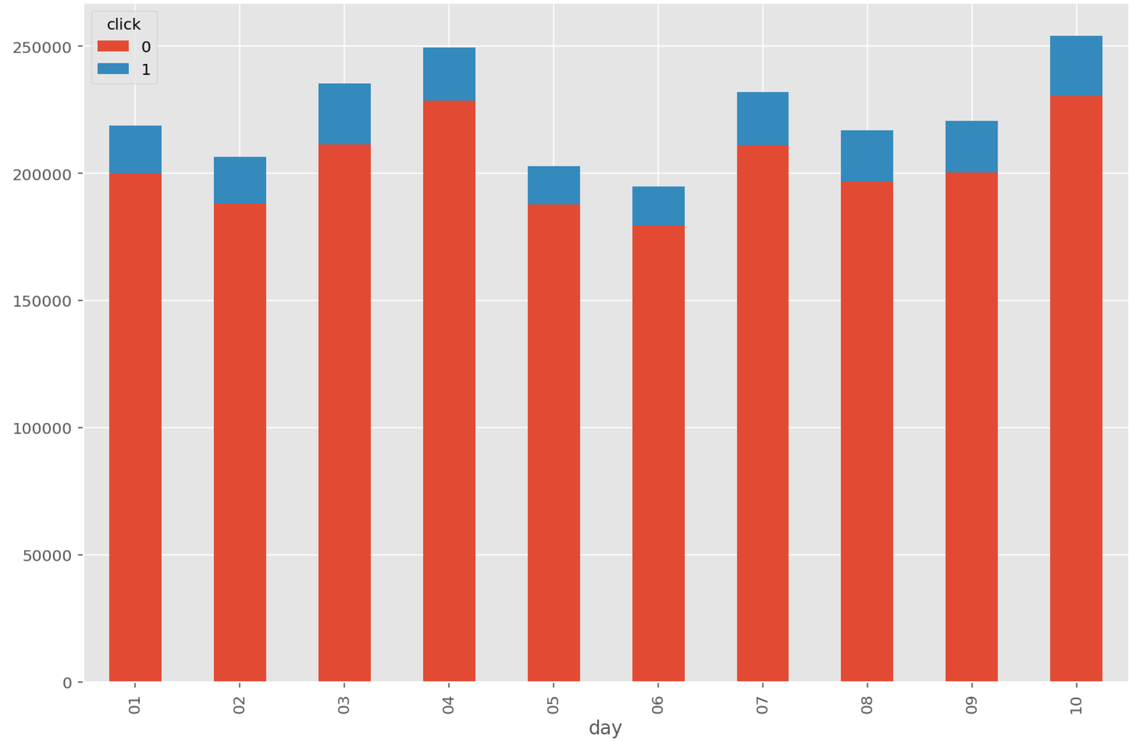
device_ifa 변수를 key로 inner join 실시

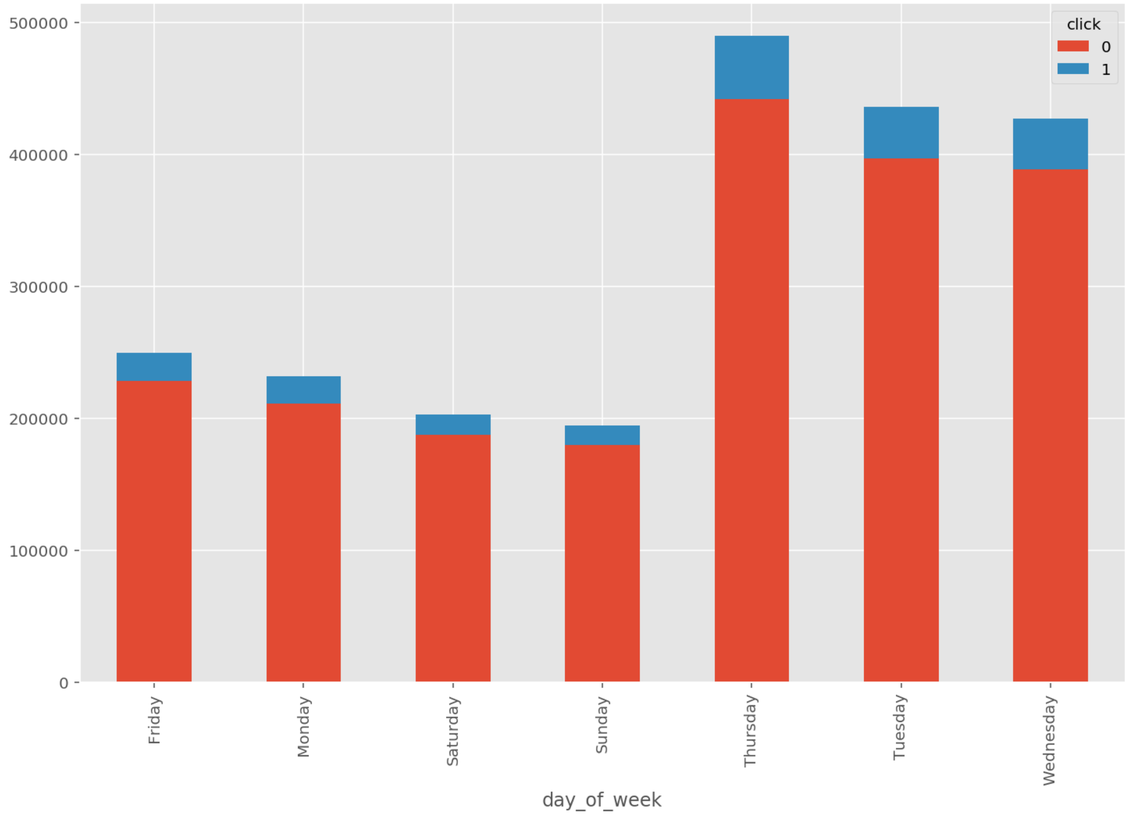
- join 이전 광고 노출 로그 데이터: 5,500,000 rows
- join 이후 병합된 데이터: 2,232,520 rows

join 이후 상당한 데이터 손실이 발생하였지만, 병합 후의 데이터도 분석하기에 여전히 큰 데이터이기 때문에 이를 활용하는 방향으로 진행함

03. 데이터 전처리 및 EDA – 추가 변수 생성 및 데이터 분포 확인

- 로그 발생 시간인 'event_datetime' 변수로 부터,
  일자(day), 시간(hour), 요일(day_of_week), 주말 및 공휴일 유무(holiday)를 나타내는 추가 변수 생성
- holiday 변수는 전체 데이터 기간이 2019년 10월 1일부터 2019년 10월 10일까지 이므로,
  3일(개천절), 5일(토), 6일(일), 9일(한글날)에 해당되는 날짜는 1, 나머지는 0인 binary 변수

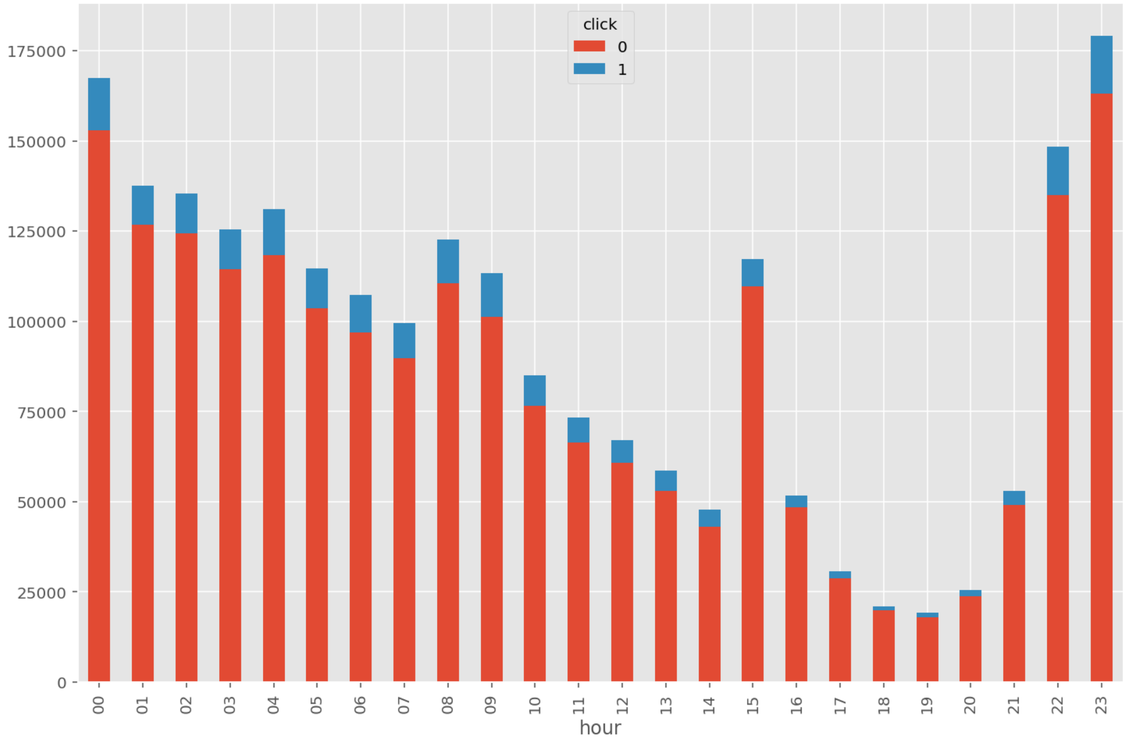


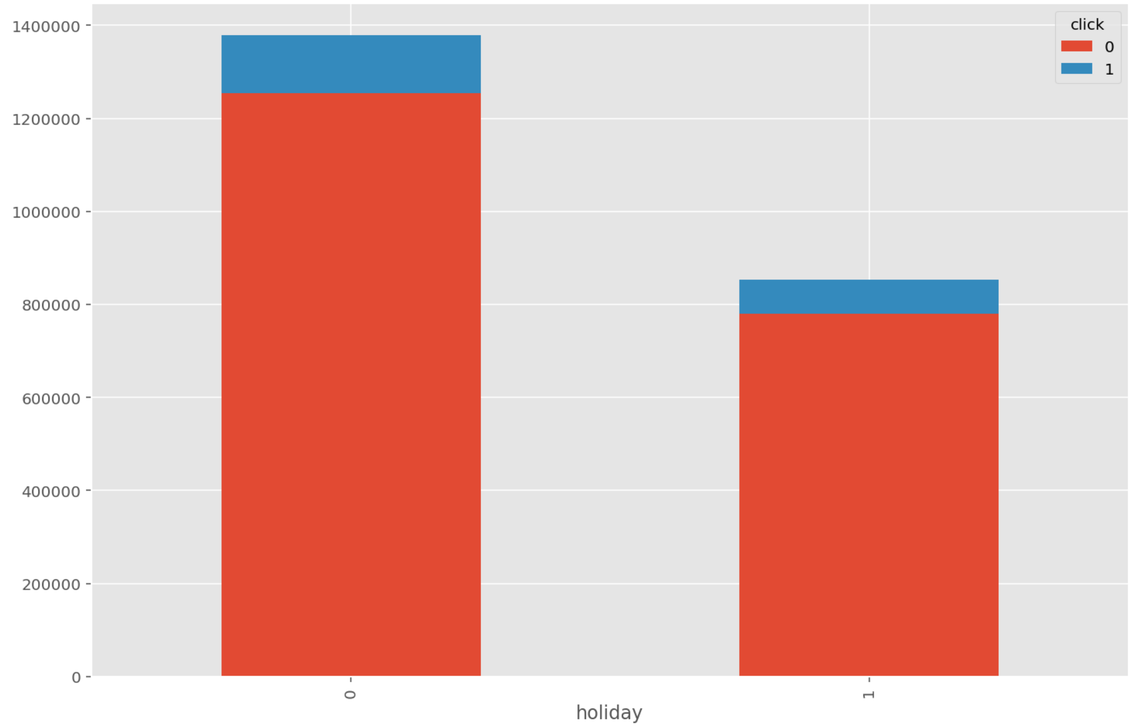


- 전체 데이터의와 label의 분포는 전체 날짜에 골고루 분포함을 확인할 수 있음
- 화, 수, 목에 다른 요일에 비해 데이터의 양이 많은 이유는 전체 데이터 기간에 해당 요일이 2번 반복되기 때문 (이를 감안하면 요일별로도 비슷한 분포)

03. 데이터 전처리 및 EDA – 추가 변수 생성 및 데이터 분포 확인

- 로그 발생 시간인 'event_datetime' 변수로 부터,
  일자(day), 시간(hour), 요일(day_of_week), 주말 및 공휴일 유무(holiday)를 나타내는 추가 변수 생성
- holiday 변수는 전체 데이터 기간이 2019년 10월 1일부터 2019년 10월 10일까지 이므로,
  3일(개천절), 5일(토), 6일(일), 9일(한글날)에 해당되는 날짜는 1, 나머지는 0인 binary 변수





- 로그 발생 시간대는 주로 밤~새벽 시간대에 많이 발생하고, 저녁시간대에 가장 적게 발생함을 확인
- 주말 및 공휴일 유무에 따른 광고 클릭 비율은 크게 차이가 없는 것을 알 수 있음

결측치 처리

- 다른 변수들에는 결측치가 존재하지 않으나, 자산 추정 가격(predicted_house_price) 변수에 상당한 결측치가 존재함을 확인
- audience의 자산, 즉 경제적 능력은 광고 클릭 확률과 높은 관련이 있을 것이라 생각되어 predicted_house_price에 존재하는 결측치는 연령, 성별, 기혼여부에 따라 그룹화된 그룹별 평균으로 대체

변수 제거

- 레벨값이 하나뿐인 범주형 변수 제거
  : 기기 OS(device_os), 기기 국가(device_country)
- 데이터 제공시 오류가 있는 변수 제거
  : 자산 추정 지수(asset_index)

03. 데이터 전처리 및 EDA – 범주형 변수 인코딩

- 자산 추정 가격(predicted_house_price) 변수를 제외하고 모든 변수가 범주형 변수
  → 적절한 인코딩 필요

변수	설명	레벨값
click	클릭 여부	2
ssp_id	SSP 아이디	17
campaign_id	캠페인 아이디	178
adset_id	광고 아이디	844
placement_type	광고 타입	4
media_id	미디어 아이디	4041
media_name	미디어 한글이름	4520
media_bundle	미디어 앱명	3881
media_domain	미디어 도메인	135
publisher_id	매체사 아이디	2808
publisher_name	매체사 이름	981
device_ifa	기기 구별 아이디	767197
device_os_version	기기 OS 버전	54
device_model	기기 모델명	778
device_carrier	기기 통신사	347
device_make	기기 제조사	181

변수	설명	레벨값
device_connection_type	기기 연결방식	8
device_language	기기 언어	23
device_region	기기 지역	130
device_city	기기 도시	983
advertisement_id	광고주 아이디	28
age	연령 (추정)	12
gender	성별 (추정)	2
marry	기혼여부 (추정)	2
Install_pack	설치된 앱 정보	767197
cate_code	IGAW 카테고리 등급	767197
predicted_house_price	자산 가격 (추정)	numeric 변수
day	날짜	10
hour	시간대	24
day_of_week	요일	7
holiday	주말 및 공휴일 유무	2

03. 데이터 전처리 및 EDA – 범주형 변수 인코딩

Machine Learning 모델 적용시

- One-hot-encoding
  : 레벨값이 적고 범주 레벨에 따라 target 변수에 유의미한 차이를 보이는 변수의 경우
  ex) placement_type (광고 타입)
- Frequency encoding
  : 레벨값이 많은 경우
  ex) publisher_name (매체사 이름)
- Target Mean encoding
  : 레벨값이 많고 target 변수와 밀접하게 관련된 변수의 경우 범주별로 target 비율을 구해 해당 값으로 대체
  (단, target 변수의 정보를 사용하므로 Data Leakage 와 Overfitting 문제 발생 위험 → Smoothing, CV loop 방법을 통해 해당 문제 최소화)

Deep Learning 모델 적용시

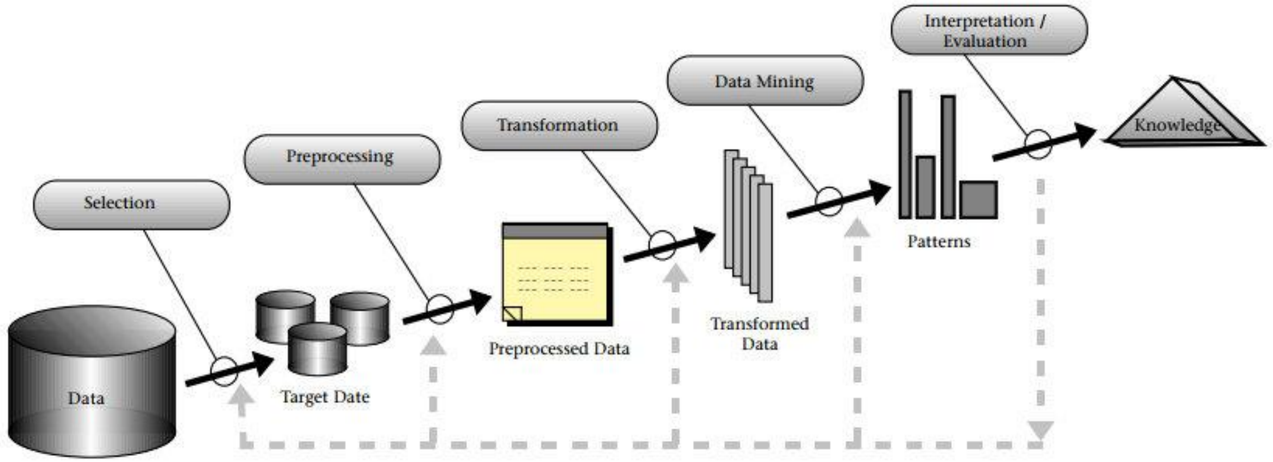
- 범주형 변수들을 Embedding 시켜 NN의 input으로 사용

  참고 논문) Entity Embeddings of Categorical Variables / Cheng Guo, Felix Berkhahn

Future Plan: 모델 별로 적절한 encoding 방식을 적용해보고 최적 encoding 방식과 모델 선정

04. 분석 과정 및 결과

- 분석 기법 방법 및 프로세스:
  KDD 분석 프로세스에 따라 머신러닝 및 딥러닝 기법을 활용하여 광고주와 인터넷 오디언스 사이의 최적 연결을 만드는 알고리즘 설계



- 모델 정확도 평가 방법:
  개별 클릭 확률에 따른 log loss를 평가 지표로 사용

$$\text{Log Loss} = \sum_{(x,y)\in D} -y\log(y') - (1-y)\log(1-y')$$

04. 분석 과정 및 결과 - 최종 전처리 방법 선택

- LightGBM model을 base model로, Frequency encoding을 적용한 데이터 vs 적용하지 않은 데이터 비교[1]
- Target mean encoding의 경우 종속변수의 분포를 알고 있어야 하므로 unseen data 에는 적용할 수 없어 실무적으로 활용할 수 없음 → 전처리 방법에서 제외

변수	설명	인코딩 방식
click	클릭 여부	-
ssp_id	SSP 아이디	frequency
campaign_id	캠페인 아이디	frequency
adset_id	광고 아이디	frequency
placement_type	광고 타입	one-hot
media_id	미디어 아이디	frequency
media_name	미디어 한글이름	frequency
media_bundle	미디어 앱명	frequency
media_domain	미디어 도메인	frequency
publisher_id	매체사 아이디	frequency
publisher_name	매체사 이름	frequency
device_ifa	기기 구별 아이디	frequency
device_os_version	기기 OS 버전	frequency
device_model	기기 모델명	frequency
device_carrier	기기 통신사	frequency
device_make	기기 제조사	frequency

변수	설명	인코딩 방식
device_connection_type	기기 연결방식	one-hot
device_language	기기 언어	frequency
device_region	기기 지역	frequency
device_city	기기 도시	frequency
advertisement_id	광고주 아이디	frequency
age	연령 (추정)	-
gender	성별 (추정)	one-hot
marry	기혼여부 (추정)	binary
Install_pack	설치된 앱 정보	frequency
cate_code	IGAW 카테고리 등급	frequency
predicted_house_price	자산 가격 (추정)	-
day	날짜	-
hour	시간대	-
day_of_week	요일	one-hot
holiday	주말 및 공휴일	-

전처리 방법	log loss
encoding 적용 X	0.2472
frequency encoding	0.2466

Frequency encoding을 적용한 데이터를 최종 활용 데이터로 선정

참고: [1] LightGBM의 경우 범주형 변수에 대해 encoding을 하지 않아도 모델 자체적으로 학습 가능

04. 분석 과정 및 결과 – 변수 선택 및 Over sampling

변수 선택

- SelectFromModel을 통한 변수 선택 진행
- 총 39개의 변수 중 19개 변수가 선택 됨

'age', 'predicted_house_price', 'hour', 'campaign_id_encode', 'adset_id_encode', 'media_id_encode', 'media_name_encode', 'media_bundle_encode', 'publisher_id_encode', 'publisher_name_encode', 'device_ifa_encode', 'device_os_version_encode', 'device_model_encode', 'device_carrier_encode', 'device_region_encode', 'device_city_encode', 'advertisement_id_encode', 'install_pack_encode', 'cate_code_encode'

적용 방법	log loss
변수 선택 적용 X	0.2466
변수 선택 적용 0	0.2475

변수 선택 적용 X

Over sampling

- 다음과 같이 target data의 편향 확인
- SMOTE를 통해 Over sampling 진행

적용 방법	log loss
Over sampling 적용 X	0.2466
Over samping 적용 0	0.2509

Over sampling 적용 X

04. 분석 과정 및 결과 – 파라미터 최적화 및 최종 모델

- 모바일 광고의 경우 광고 건당 100ms로 실시간으로 반응하기 때문에 모델의 train 및 predict에 소요되는 시간 최소화 필요
- 따라서 여러 모델 간 앙상블 모델 보다는 **'단일 모델'**을 사용하고 train 및 predict 속도가 가장 빠른 **'LGBM'**을 최종 모델로 사용

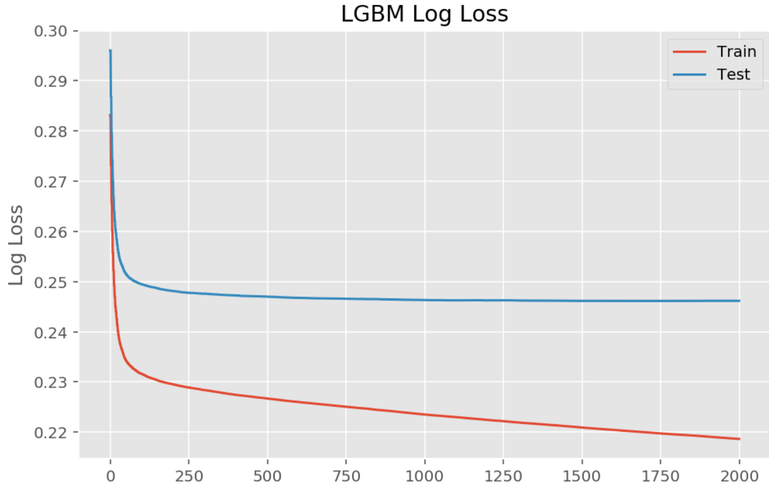
파라미터 최적화

- Bayesian Optimization을 통해 파라미터 튜닝 진행
- 최종 파라미터

```
params = {
    'boosting_type': 'gbdt',
    'objective': 'binary',
    'num_leaves': 49,
    'learning_rate': 0.1128527143464075,
    'feature_fraction':  0.24089912543315553,
    'bagging_freq': 15,
    'verbose': 0,
    'max_bin': 495,
    'num_iterations': 1000,
    'min_data_in_leaf': 10,
    'min_sum_hessian_in_leaf': 10,
    'random_state': 42
}
```

최종 모델

- 최종 LGBM 모델 log loss: 0.2461

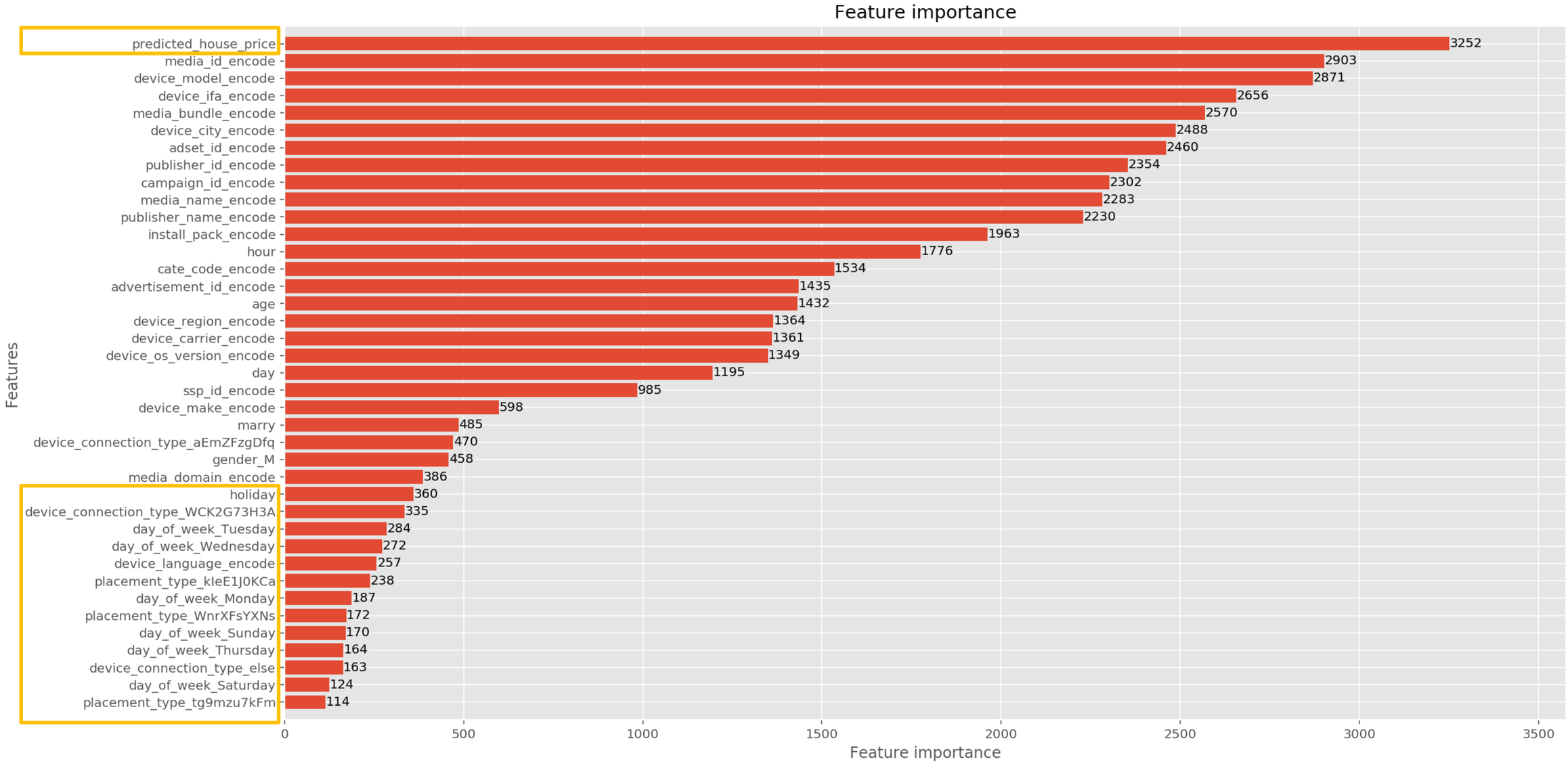


- 다른 모델과 성능 및 학습 시간 비교

모델	log loss	학습 시간
LGBM	0.2461	2m 28s
Random Forest	0.2517	23m 6s
CatBoost	0.2478	20m 36s
XGBoost	0.2508	24m 15s

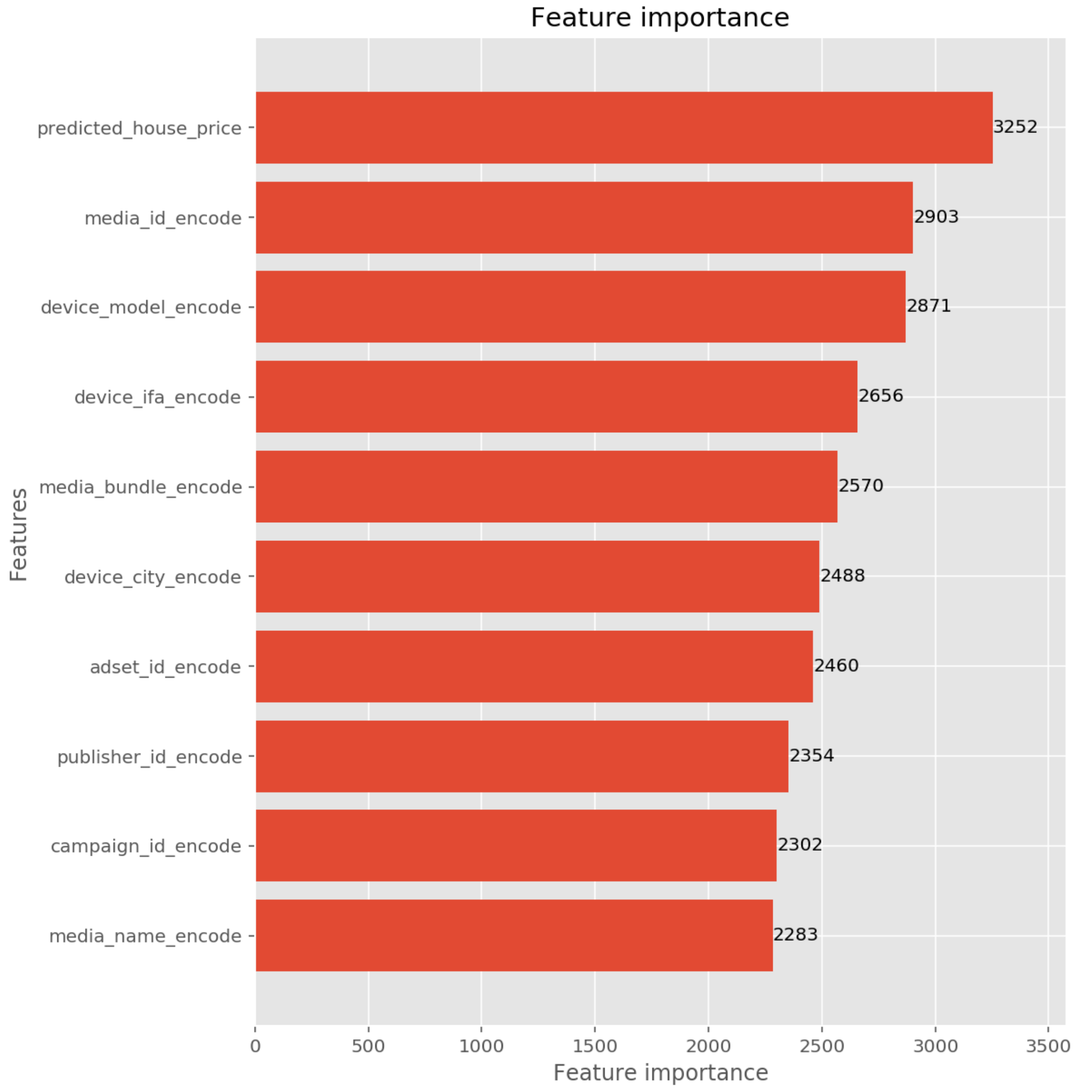
04. 분석 과정 및 결과 – 모델 결과 해석

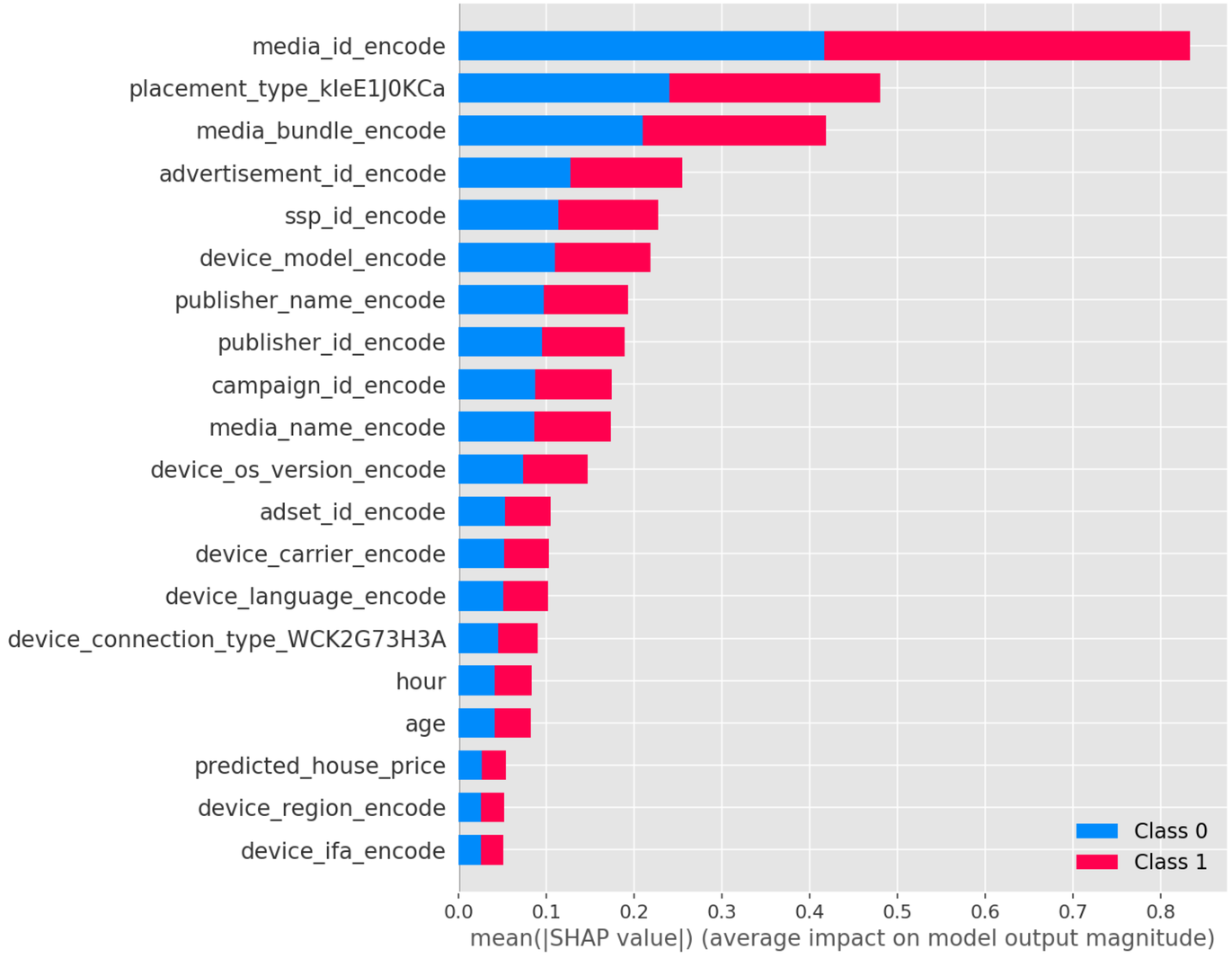
- 변수 중요도 시각화 결과, 자산 추정 가격(predicted_house_price)가 가장 중요한 변수로 나타남
- 주말 및 공휴일 유무와 요일 변수의 경우 중요도가 낮음



04. 분석 과정 및 결과 – 모델 결과 해석

- 변수 중요도 상위 10개 변수와 SHAP value 비교

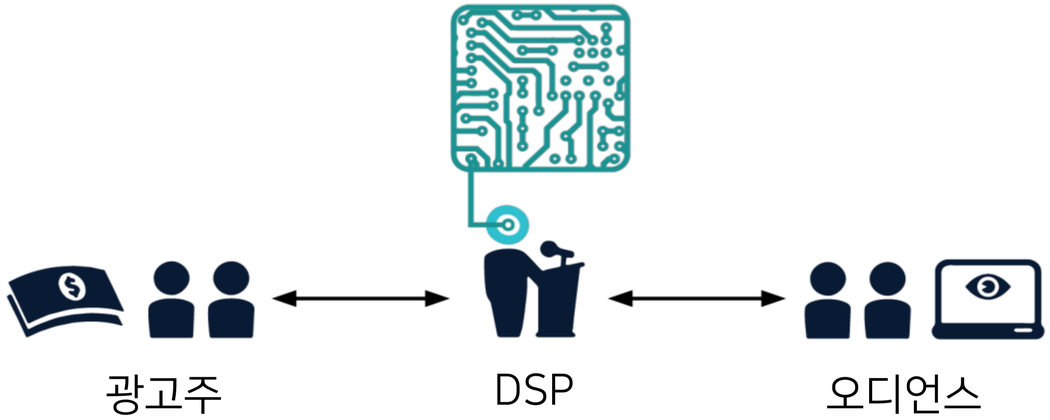




05. 결과 활용 방안 및 향후 계획

- 결과 활용 방안:

기존 광고 로그 데이터 분석 모델링은 결과가 건 당 100ms의 실시간으로 반응하기 때문에 매출과 매우 밀접하게 직결되어 있어 광고 산업에서 유용하게 활용될 수 있다. 따라서 해당 프로젝트를 통해 오디언스의 연령, 성별, 관심사, 지역정보, 라이프스타일, 구매력 등을 분석하고 이들의 모바일 행동을 관찰하여 정확도가 높은 CTR 예측 모델을 구축하고자 한다.
예측된 CTR을 바탕으로 광고 입찰 전략을 통해 RTB 경매에 입찰할 입찰가를 결정하는데 기여할 수 있다.



- 향후 계획:

- 범주형 변수들을 Embedding 시켜 NN의 input으로 사용한 Deep Learning 모델 생성
- 기존 LGBM 모델과 성능 및 속도 비교

THANK YOU