스태킹 앙상블을 활용한 웹 광고 클릭률 예측

2024.06.06

김태형, 함유진

목차

1. 문제 정의 및 데이터 설명

웹 광고 클릭률 예측

- 7일간의 웹 로그를 기반으로 하루 동안의 광고 클릭률을 예측하는 AI 알고리즘을 개발함.
- 데이터 설명은 다음과 같음.
- ID: train 데이터 샘플 고유 ID
- Click: 예측 목표인 클릭 여부 (0: 클릭하지 않음, 1: 클릭)
- F01 ~ F 39 : 각 클릭 로그와 연관된 Feature
- 개인정보 보호를 위해 상세 정보는 비식별 처리됨.
- 학습 및 평가 데이터는 다음과 같이 결측치가 많음.

	ID	Click	F01	F02	F03	F04	F05
1	TRAIN_00000000	1	NSLHFNS	AVKQTCL	DTZFPRW	114.0	ISVX
2	TRAIN_00000001	0	VGIVWZQ	LSUSMVO	PQGWFJZ	26.0	NFRV
3	TRAIN_00000002	0	JCDXFYU	PILDDJU	IAGJDOH	119.0	LFPUE
4	TRAIN_00000003	1	PSMFWTP	ZYAVJHP		15.0	ATQP
5	TRAIN_00000004	0	SLCRICD	QPQWGXA		13.0	CHZG
6	TRAIN_00000005	0	ZRTPTHN	WOAIOXV		10.0	KGRH
7	TRAIN_00000006	1	JCDXFYU	PILDDJU	IAGJDOH		LFPUE
8	TRAIN_00000007	0	KTNHUXF	YSUQLFP			FNSJIG
9	TRAIN_00000008	0	AOKIKZU	PKLWPDW		2.0	VRSGS
10	TRAIN_00000009	0	JCDXFYU	PILDDJU	IAGJDOH	15.0	LFPUE

2. 데이터 전처리

탐색적 데이터 분석

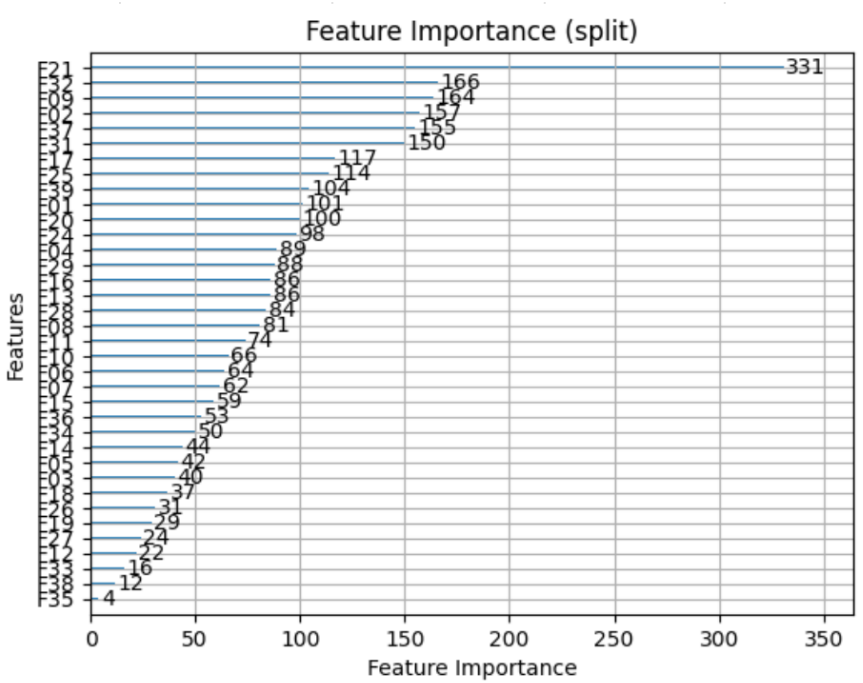
- 웹 로그 데이터는 대용량, 클래스 불균형, 고차원(High Cardinality) 등을 특징으로 가짐.
- 학습 데이터는 28,605,391 개수, 평가 데이터는 4,538,541 개수로 대용량 데이터임.
- 종속 변수 Click은 0이 23,035,531(ratio=0.805286) 개수, 1이 5,569,860(ratio=0.194714) 개수로 클래스 불균형 데이터임.
- 독립 변수 F01~F39 중 수치형 변수는 14개 , 범주형 변수는 26개임.
- 범주형 변수는 고차원 데이터로 Count Encoder를 활용한 Frequency Encoding을 진행함.
- 수치형 변수는 최빈값인 0으로 결측치를 대체함.

범주형 변수에 대한 고유한 값의 개수

F01 nunique: 4760931
F02 nunique: 304405
F03 nunique: 64
F05 nunique: 5343557
F07 nunique: 151200
F08 nunique: 79
F09 nunique: 27551
F10 nunique: 1404255
F12 nunique: 4174064
F13 nunique: 1307
F15 nunique: 4
F16 nunique: 15467
F17 nunique: 10
F20 nunique: 178603
F21 nunique: 33
F22 nunique: 7187
F23 nunique: 950
F25 nunique: 10700
F26 nunique: 2205
F28 nunique: 55
F30 nunique: 19444
F31 nunique: 14
F34 nunique: 3165581
F35 nunique: 3
F37 nunique: 9423
F39 nunique: 6800

3. 방법론

Feature Engineering



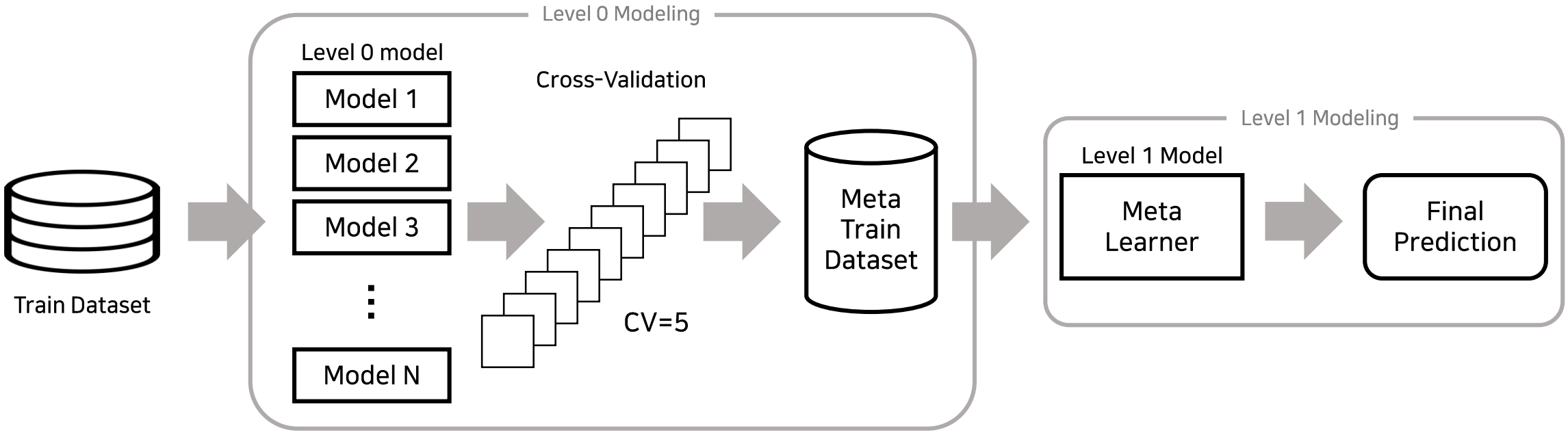
'F21' 에 대한 'F32'의 평균값 파생 변수

	F21	F32
0	892952	380.0
1	9100506	466.0
2	5573998	197.0
3	778969	8640.0
4	4652049	41774.0

	F21	F32	**mean_by_F21_F32**
0	892952	380.0	20483.091906
1	9100506	466.0	20413.441983
2	5573998	197.0	11706.724265
3	778969	8640.0	19655.257965
4	4652049	41774.0	20174.624671

- LightGBM을 활용하여 Feature Importance를 계산함.
- 변수 중요도가 높은 변수를 기준으로 범주형 변수에 대한 수치형 변수의 평균, 최소값, 최대값 등 기초 통계량 파생 변수를 생성함.
- 예를 들어, 변수 중요도가 가장 높은 범주형 변수 'F21'에 대한 수치형 변수 'F32'의 평균값 파생 변수를 만들어 사용함.

Stacked Generalization (Stacking Ensemble)



- Stacking은 다양한 예측 모형의 출력 값을 최종 예측 모형의 입력 값으로 사용하는 앙상블 기법임.
- Level 0 Modeling을 통해 선택된 모형의 예측 값으로 Meta Train Dataset을 구축함.
- 그 후, Level 1 Modeling을 통해 Meta Learner을 학습하여 최종 예측을 진행함.
- XGBoost, LightGBM, Catboost 등 3가지 모델을 Level 0 Model로 사용함.
- Logistic regression을 Level 1 Model로 사용함.

4. 실험 결과 및 결론

Experiments & Results

- 최종 제출 모델은 stratified 5-fold를 활용한 XGBoost, LightGBM, CatBoost를 Stacked Generalization한 결과임.
- Optuna를 활용하여 하이퍼 파라미터를 탐색함.
- XGBoost의 Validation AUC는 0.78488임.
- LightGBM의 Validation AUC는 0.77376임.
- CatBoost의 Validation AUC는 0.76746임.
- Stacking 모델의 Public AUC는 약 0.7865임.
- Feature Engineering을 활용한 Stacking 모델의 Public or Private AUC는 약 0.788~0.789 임.
- 데이터는 전처리 후 Parquet 파일로 저장 후 사용함.
- 모델 최적화 및 파생변수가 성능 향상에 유의미한 결과를 보임.

Thank you ☺

taehyeong93@korea.ac.kr