LG U+

2022 유플러스 AI Ground

싹쓰리

김성연(리더)
구진범
이환주

1.EDA

- Top-10 (성별, 나이, 시간)
- 분류별 유사도
- 데이터 편차
- 희소 행렬
- Train/Valid

2.모델접근법

- CatBoost
- Rule-based
- Rule-based + Baseline
- Rule-based + LightGCN

3.모델실용성

- Batch serving이 가능한 이유
- Batch serving 특징

4.모델개선방안

- 모델 개선 방안

팀원을 소개합니다

김성연(리더)

EDA

LightGCN

발표

구진범

EDA

CatBoost

발표 자료 제작

이환주

EDA

Rule-based

발표 자료 제작

EDA

Top 10 – 성별

남자아이의 Top 10

여자아이의 Top 10

EDA

Top 10 – 나이

3세 이하 아이의 Top 10

4-5세 아이의 Top 10

EDA

Top 10 – 나이

6-7세 아이의 Top 10

8-9세 아이의 Top 10

10-13세 아이의 Top 10

EDA

Top 10 – 시간

3월 상반기의 Top 10

3월 하반기의 Top 10

EDA

Top 10 – 시간

4월 상반기의 Top 10

4월 하반기의 Top 10

EDA

분류별 유사도

성별 간 유사도

	M - F
recall@100	0.7755
ndcg@100	0.72

나이 간 유사도

	age1 - age2	age1 - age3	age1 - age4	age1 - age5	age2 - age3	
recall@100	0.5043	0.2165	0.2155	0.1852	0.5568	
ndcg@100	0.43	0.16	0.15	0.15	0.46	
	age2 - age4	age2 - age5	age3 - age4	age3 - age5	age4 - age5	avg
recall@100	0.5311	0.3120	0.8146	0.5477	0.5340	0.4418
ndcg@100	0.44	0.26	0.76	0.51	0.49	0.381

3세 이하, 4-5세, 6-7세, 8-9세, 10-13세로 분류

날짜 간 유사도

	date1 - date2	date1 - date3	date1 - date4	date2 - date3	date2 - date4	date3 - date4	avg
recall@100	0.7966	0.6880	0.6427	0.7883	0.7089	0.8067	0.7385
ndcg@100	0.75	0.62	0.58	0.73	0.65	0.77	0.683

3월 상반기, 3월 하반기, 4월 상반기, 4월 하반기로 분류

EDA

분류별 유사도

부모 관심 키워드

	과학기술	정서/사회성	자연탐구	바른생활/안전	
recall@100	0.5438	0.8783	0.8489	0.8286	
ndcg@100	0.46	0.84	0.81	0.78	
	활동/운동	음악예술	언어논리	수리논리	avg
recall@100	0.7750	0.7799	0.7837	0.5364	0.7468
ndcg@100	0.72	0.72	0.72	0.44	0.686

아이 관심 키워드

	노래/율동	동물/식물	동화	만들기	숫자/계산
recall@100	0.4450	0.8310	0.6622	0.5321	0.5350
ndcg@100	0.35	0.78	0.58	0.44	0.45
	외국어	친구/사람	탈 것/기계	활동/운동	avg
recall@100	0.5930	0.8235	0.7948	0.8392	0.6729
ndcg@100	0.51	0.77	0.74	0.79	0.601

EDA

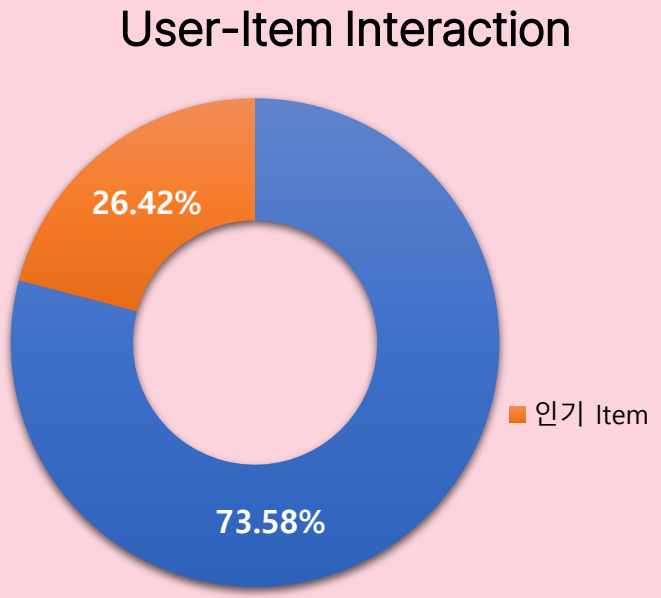
데이터 편차

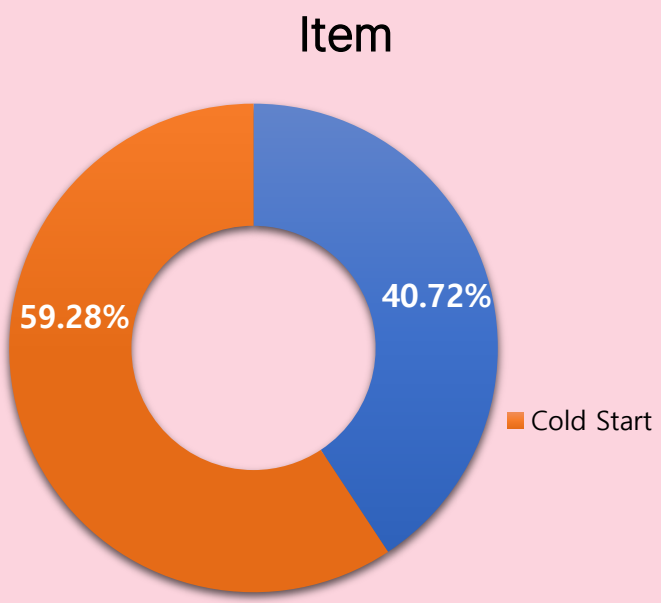
인기 아이템 - 상위 1% 아이템

→ 26.42%

Cold start - 5명 이하가 시청한 아이템

→ 59.28%





EDA

희소행렬

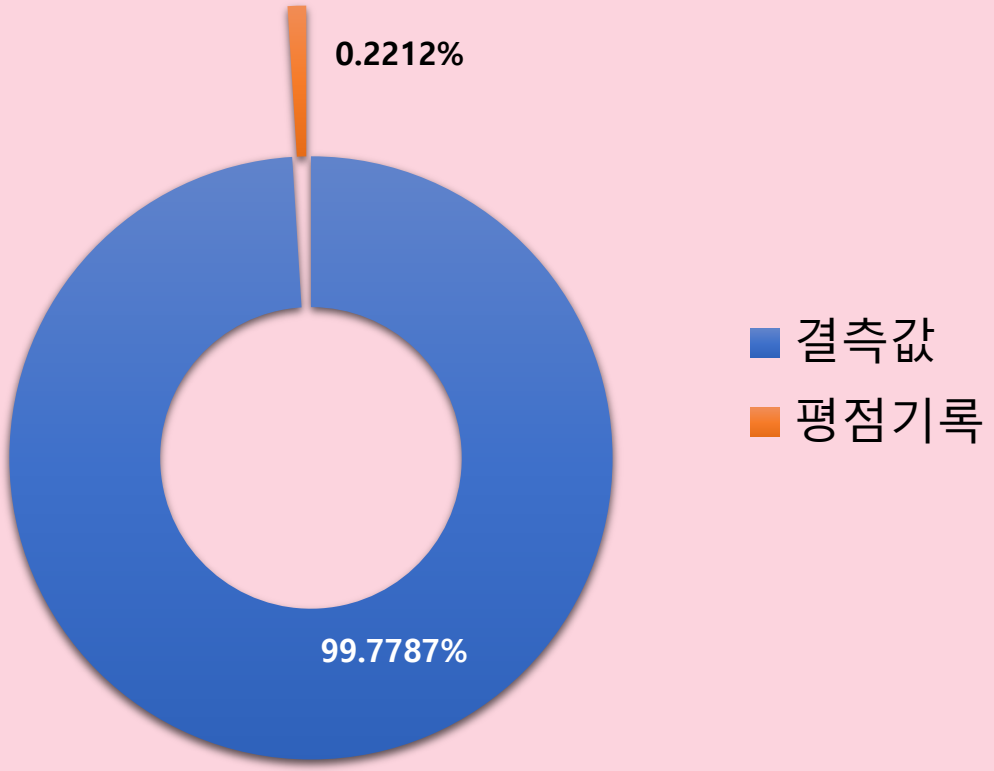
유저 수: 8311, 아이템 수: 20695

최대 시청 기록 개수: 유저 수 * 아이템 수

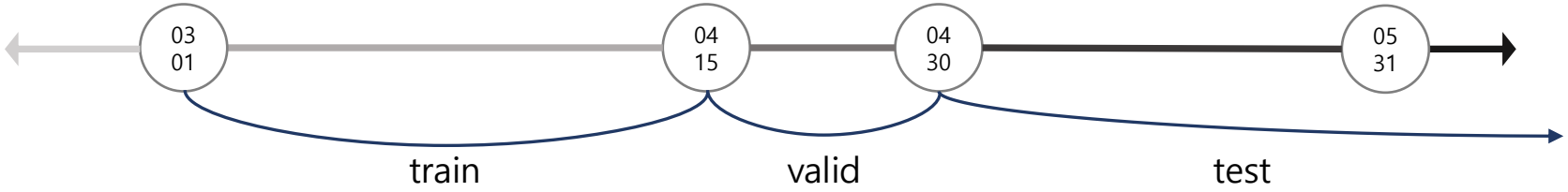
= 8311 * 20695 = 171996145

시청 기록: 380547

희소 행렬(결측 값 비율) = 99.7787%



Train/Valid



3월과 4월의 데이터 → 5, 6, 7월의 시청 기록

랜덤으로 train/valid를 나누는 것이 아닌 시계열 성질을 고려

→ 4월 15일을 기준으로 하여 train/valid를 설정

모델 접근법: CatBoost

데이터 특성

범주형 변수

- Baseline에서 사용하지 않은 user의 정보와 item의 정보를 사용

Benchmarks

Quality | Learning speed

☑ Tuned ☑ Default

	CatBoost		LightGBM		XGBoost		H2O	
	Tuned	Default	Tuned	Default	Tuned	Default	Tuned	Default
Adult	**0.26974**	0.27298 +1.21%	0.27602 +2.33%	0.28716 +6.46%	0.27542 +2.11%	0.28009 +3.84%	0.27510 +1.99%	0.27607 +2.35%
Amazon	**0.13772**	0.13811 +0.29%	0.16360 +18.80%	0.16716 +21.38%	0.16327 +18.56%	0.16536 +20.07%	0.16264 +18.10%	0.16950 +23.08%
Click prediction	**0.39090**	0.39112 +0.06%	0.39633 +1.39%	0.39749 +1.69%	0.39624 +1.37%	0.39764 +1.73%	0.39759 +1.72%	0.39785 +1.78%
KDD appetency	0.07151	**0.07138** −0.19%	0.07179 +0.40%	0.07482 +4.63%	0.07176 +0.35%	0.07466 +4.41%	0.07246 +1.33%	0.07355 +2.86%

Features

1  **Great quality without parameter tuning**

Reduce time spent on parameter tuning, because CatBoost provides great results with default parameters

2 

Categorical features support

Improve your training results with CatBoost that allows you to use non-numeric factors, instead of having to pre-process your data or spend time and effort turning it to numbers.

CatBoost 모델 선택

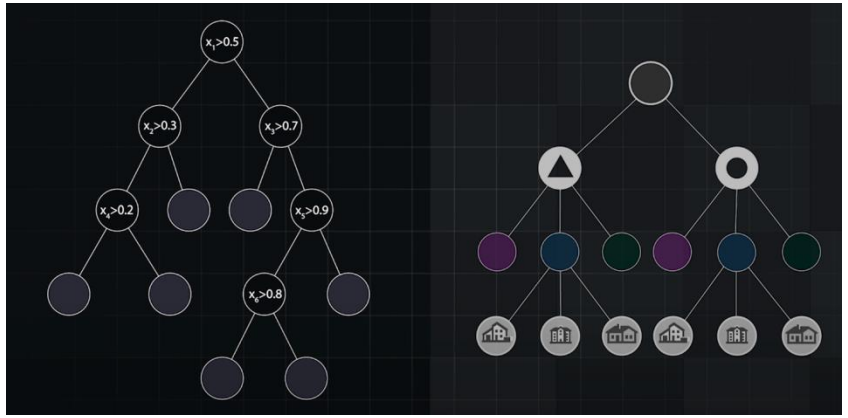
- 정형 데이터를 다루는 분류 및 회귀 문제에서 트리 기반 부스팅 모델이 가장 좋은 성능을 보여주고 있음 (in Kaggle…)
- 트리 기반 부스팅 모델 중 범주형 변수 처리에 특화된 CatBoost 선택

모델 접근법: CatBoost

데이터 전처리

Negative sampling & 결측치 제거

- 시청 시간 비율이 5% 이하인 경우 제거
- 각 유저의 각 앨범 당 평균 시청 비율을 사용
- Train에서 일정 비율만큼 negative sample을 랜덤하게 추가



CatBoost 모델에 적용

- CatBoost 모델이 범주형 변수에 강하기 때문에, 수치형 변수인 run_time, age 변수를 범주형으로 군집화

모델 접근법: CatBoost

사용한 Feature

Users – Age

- 나이 그룹 별 평점 평균에 유의미한 차이 존재
- 3세 이하, 4-5세, 6-7세, 8-9세, 10-13세로 분류

3세 이하 인기 앨범

10세 이상 인기 앨범

관심사에 노래/율동이 있는 아이

관심사에 노래/율동이 없는 아이

Users – 아이 관심 키워드

- 아이 관심 키워드 별 유의미한 차이 존재
- 유의미한 차이를 보이는 노래/율동, 만들기, 숫자/계산 키워드를 사용

Rule-Based

- 데이터를 분석한 결과 신규 컨텐츠 보다 기존 시청 컨텐츠 소모하는 경우가 다수

1. 기존 시청 컨텐츠 + Baseline
2. 기존 시청 컨텐츠 + CatBoost
3. 기존 시청 컨텐츠 + 나이 별 인기 컨텐츠
4. 기존 시청 컨텐츠 + Baseline 임베딩 기반 클러스터링
5. 기존 시청 컨텐츠 + LightGCN 임베딩 기반 클러스터링

Rule-Based

- Baseline/CatBoost/Age
- 기존 시청 컨텐츠에 추가로 추천하는 방식으로 실험 진행

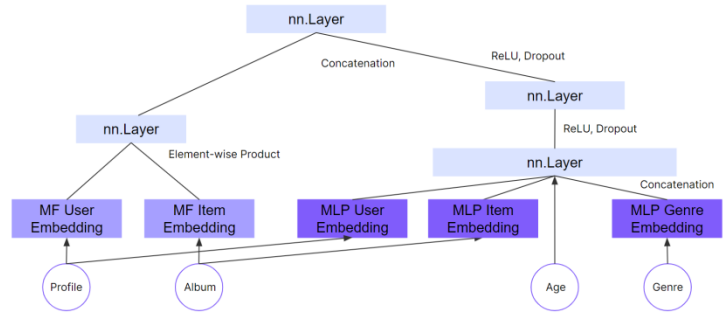
	Baseline	CatBoost	Age
score	0.2301 / 0.1762	0.2278 / 0.1737	0.2295 / 0.1761
recall@25	0.2335 / 0.1810	0.2310 / 0.1783	0.2326 / 0.1807
ndcg@25	0.2199 / 0.1619	0.2180 / 0.1600	0.2203 / 0.1624

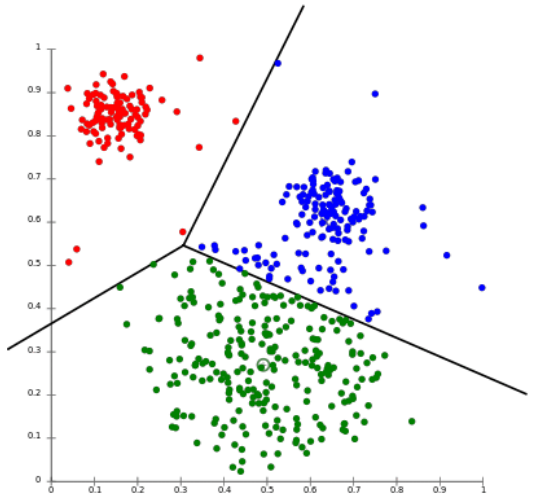
Rule-Based +Baseline

데이터 전처리

NeuMF 모델에서 임베딩 추출

- MF User Embedding / MLP User Embedding을 사용





추출된 임베딩으로 군집 만들기

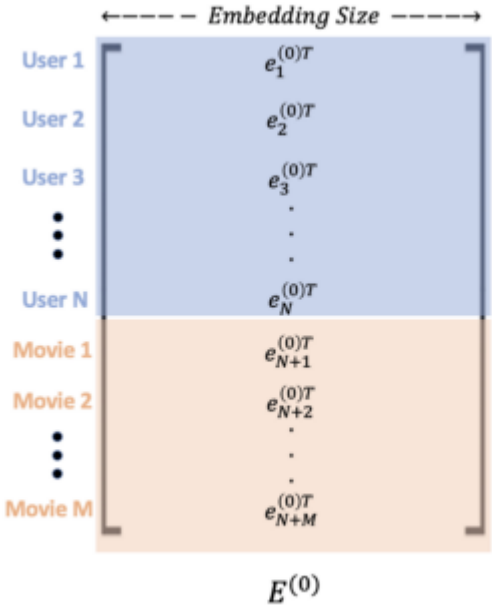
- 추출된 임베딩을 K-means 클러스터링을 통해 군집 제작
- 군집별 추천 서비스 제공

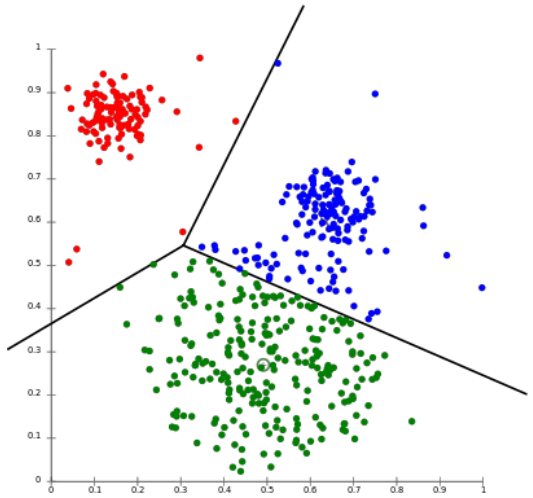
Rule-Based +LightGCN

데이터 전처리

LightGCN 모델에서 임베딩 추출

- 유저-아이템간 상호작용을 잘 고려하는 그래프 기반 LightGCN 모델을 이용해 유저별 임베딩 값 추출



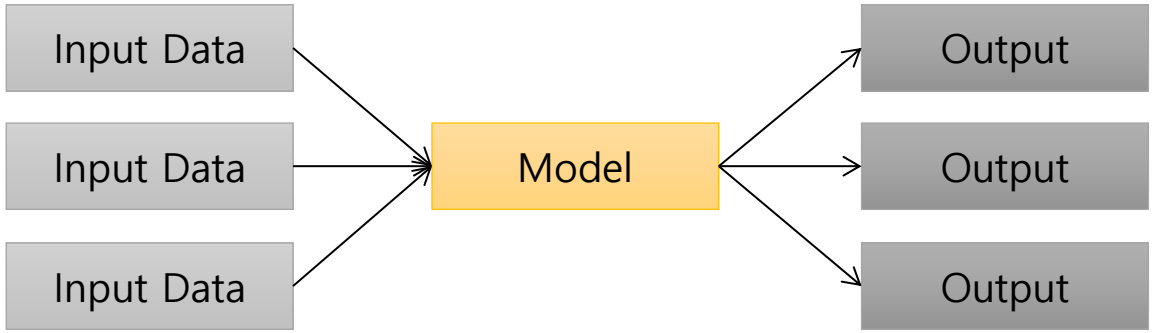


추출된 임베딩으로 군집 만들기

- 추출된 임베딩을 K-means 클러스터링을 통해 군집 제작
- 군집별 추천 서비스 제공

모델 실용성

Batch serving이 가능한 이유?



1. Sequence를 고려하지 않음
2. 최신 반응을 보고 실시간으로 추천하는 모델이 아님

→ 일정 주기를 가지는 batch serving으로 충분함

모델 실용성

Batch serving 특징

- Batch serving 가능
- 1일 단위로 batch serving 해도 될 듯
- 코드를 함수화하여 주기적으로 실행 가능
- 실시간 처리가 아니기 때문에 비용 측면에서도 부담이 없음
- online serving보다 구현이 쉽고 간단하며, 한번에 많은 데이터를 처리하므로 latency 문제가 없음
- airflow, cron job 등으로 스케줄링 작업을 진행할 수 있음

모델
개선 방안

모델 개선 방안

1. 연관성뿐만 아니라 "참신함"도 고려

→ 인기 있는 아이템만 계속 추천하는 대신 인기 없지만 질 좋은 아이템을 가지고 광고처럼 잘 보이는 곳에 띄어줌

→ 수익 창출 가능

2. 성능 지표에 popularity뿐만 아니라 age도 추가

→ 인기도에 기반하여 아이템을 추천하다 보면 오래된 영상만 추천할 가능성이 높아짐

→ 최신성도 반영하기 위해 age라는 변수도 추가

3. 추천 시스템 작동 시 클릭 여부 나타내는 feedback log 생성

→ Negative sampling을 하지 않아도 되므로 모델의 완성도가 높아짐

모델
개선 방안

모델 개선 방안

4. 아이템의 내용 정보를 더 많이 수집

→ Contents-based 추천을 잘 할 수 있음

→ 최신 영상도 잘 추천할 수 있음

5. 모델 최신화 주기적으로 진행

→ 영상 데이터는 유행에 민감하기 때문

6. Recall이나 NDCG 외에 다른 요소도 복합적으로 고려

→ 예를 들어, 참신함(Serendipity) 또는 새로움(Novelty)