2022 유플러스 AI Ground

아이들나라 고객 콘텐츠 추천

새싹팀 심승현, 박선홍, 박현주

새싹 팀원소개

심승현 (팀장)
모델링 및 테스트

박선홍 (팀원)
모델링 및 테스트

박현주 (팀원)
데이터 분석, 시각화

목차

Chapter 1.

데이터 인사이트

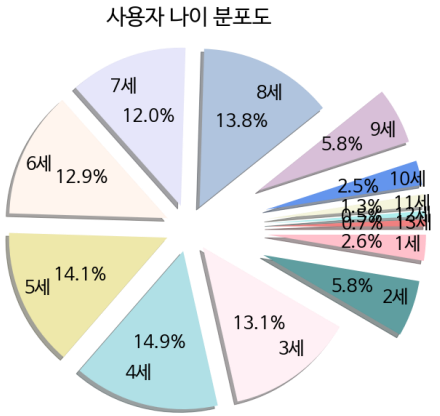
01. 문제 정의

유아-아동 전용 미디어 서비스 '아이들나라'의 실제 사용자 데이터를 바탕으로,
프로필별 맞춤형 콘텐츠 추천 AI 모델 개발.

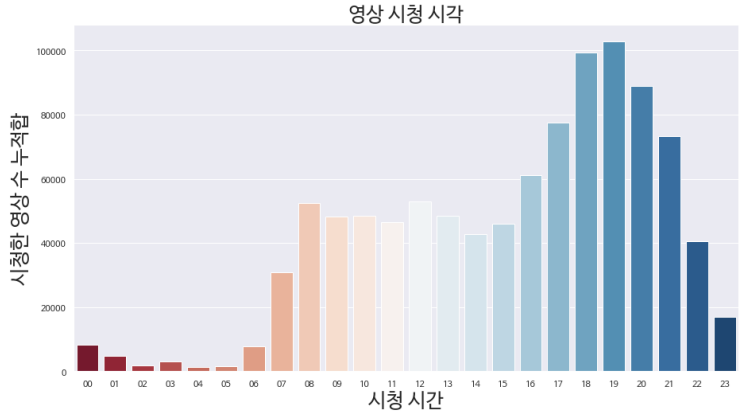
프로필 정보와 부모의 관심자 정보 그리고 시청 이력과 메타 데이터를 활용하여,
유저별 25개의 추천 콘텐츠 예측 모델 개발
ndcgk@25와 recall@25의 가중치 평균을 통해 모델 성능 평가

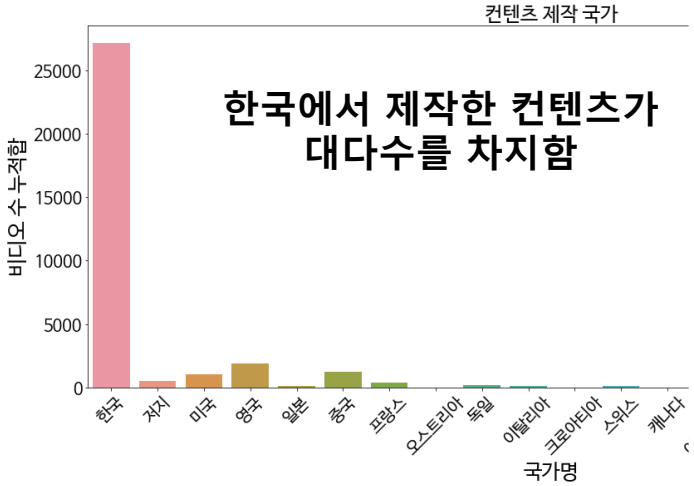
02. EDA

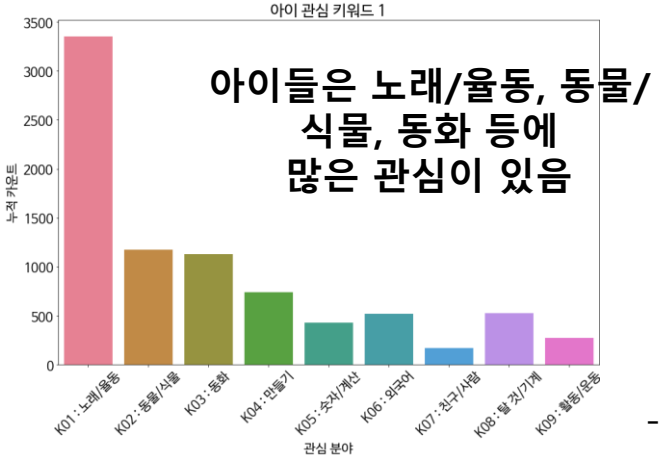
총 80% 이상의 사용자
=
3세 ~ 8세 어린이
(서로 비슷한 비율)



오후 4시~ 9시 :
시청률이 가장 높은 시간대



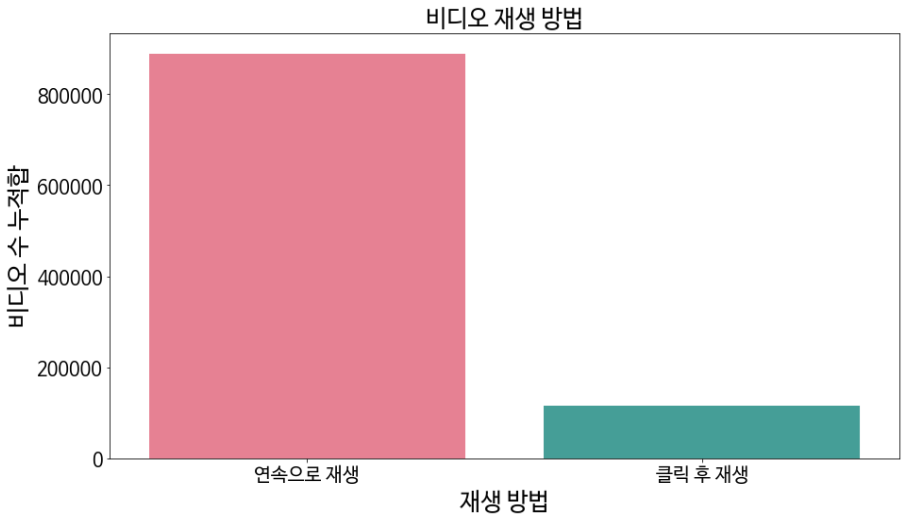




02. EDA

연속되어 재생 : 88.45%
클릭 후 재생 : 11.55%

대부분의 유저가
컨텐츠를 끝까지 시청함



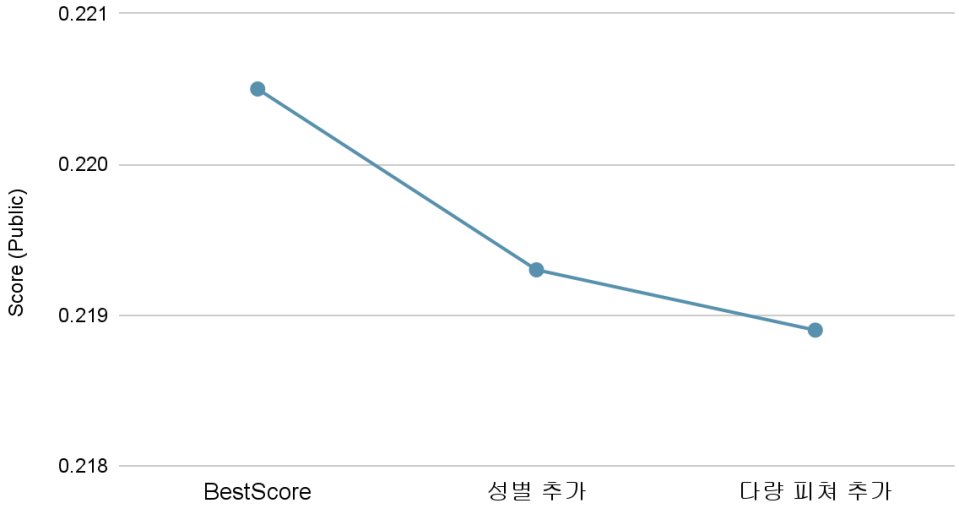
추천 알고리즘에 session_id 가 중요한 피쳐가 될 것 같다!

02. EDA

모델 성능 개선을 위한 Feature Selection에서의 어려움

- 개념적으로 모델에 유의미 할 것으로 여겨지는 Feature가 모델 성능에 큰 영향을 주지 않음
- 많은 데이터(많은 Feature)가 곧 모델의 성능을 개선시키는 것은 아님

Feature에 따른 성능 변화



history

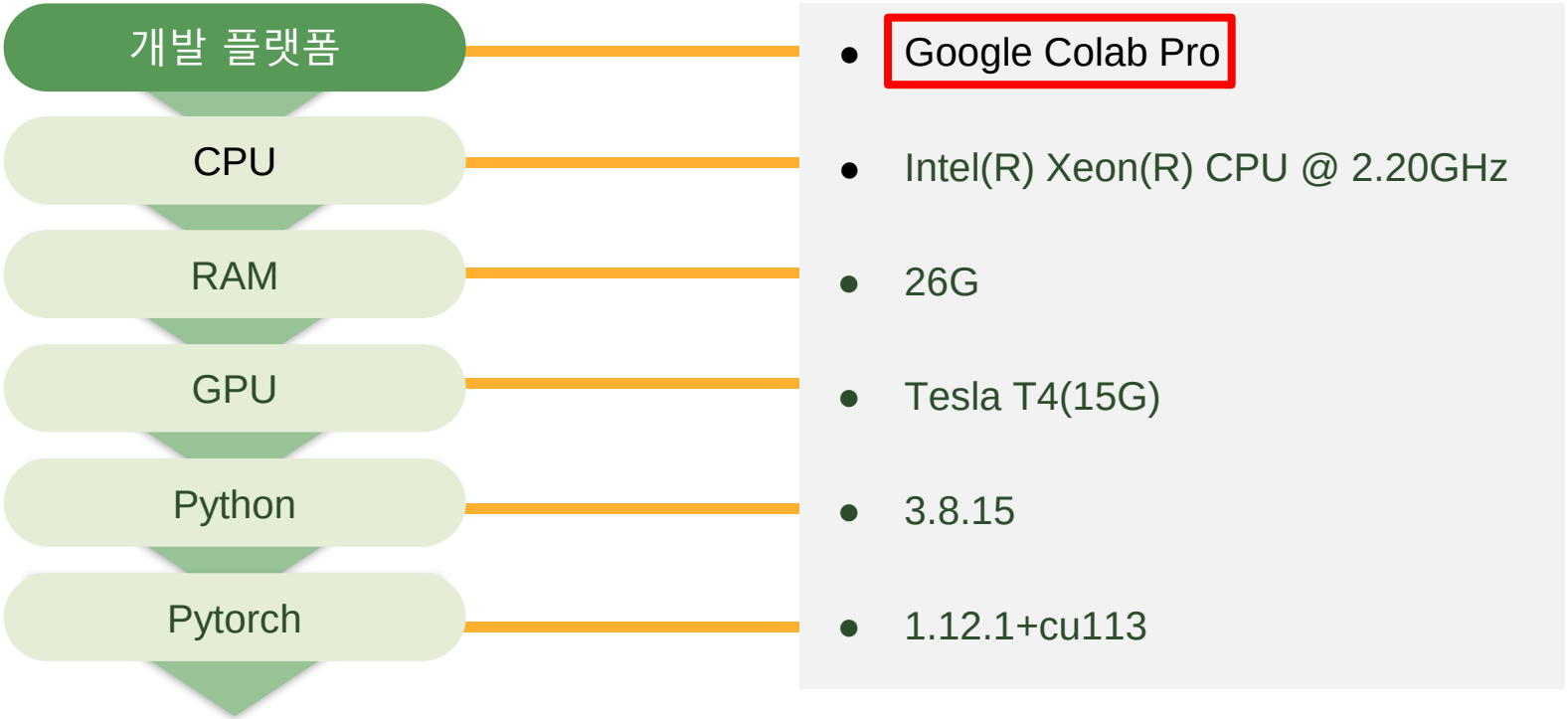
profile

meta

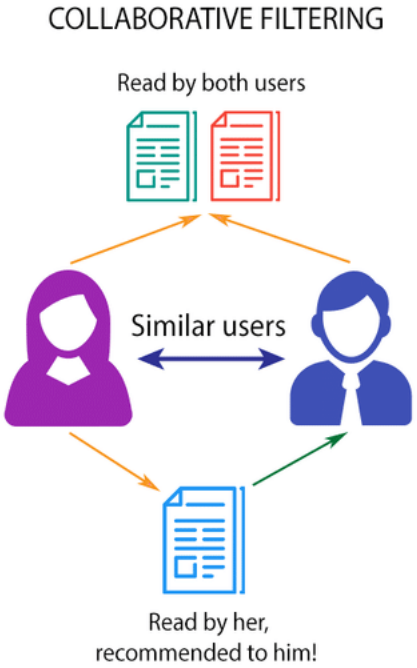
Chapter 2.

모델 학습 및 추론

01. 개발환경


개발 플랫폼
CPU
RAM
GPU
Python
Pytorch
Google Colab Pro
Intel(R) Xeon(R) CPU @ 2.20GHz
26G
Tesla T4(15G)
3.8.15
1.12.1+cu113

02. 모델링



Neural Collaborative Filtering (NCF)

1) Implicit Feedback 을 토대로 Item을 User에게 추천
2) DL을 Matrix Factorization 에서 User-Item Interaction 에 적용
3) 손실함수로 MSE가 아닌, Binary Cross-Entropy 사용
4) Point-wise Loss + Negative Sampling 사용
5) Dot-Product(GMF)와 MLP의 장점을 모두 살린 네트워크 구조

간단한 모델 구조로 준수한 성능 보장

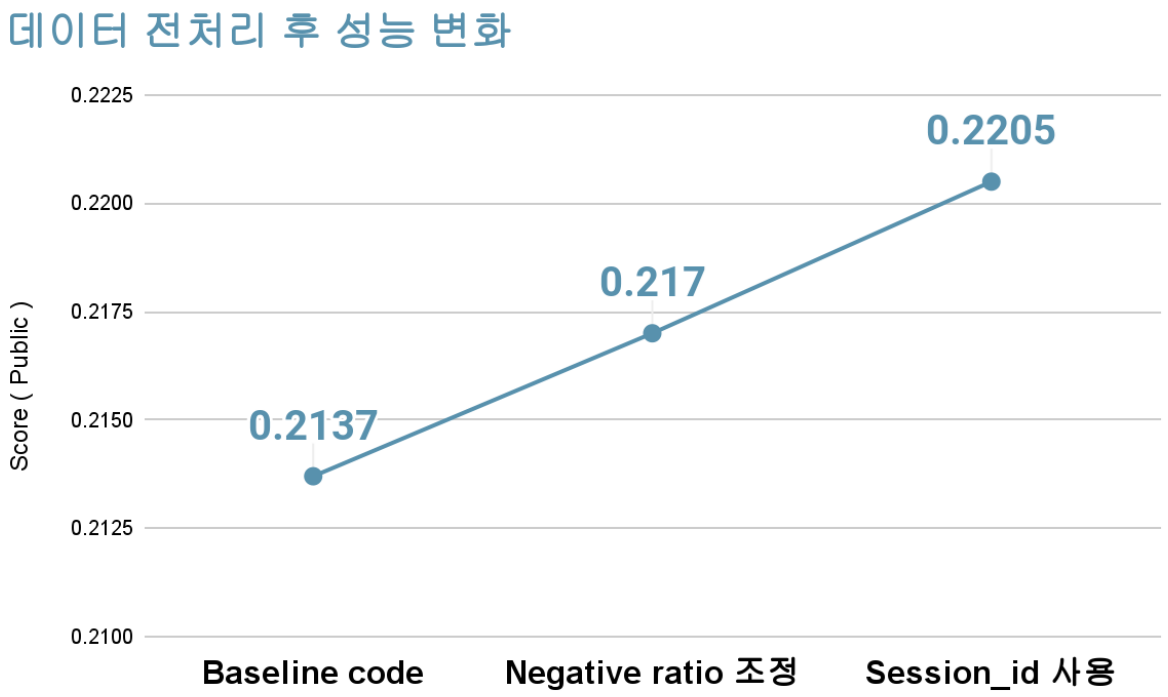
02. 모델링

모델 소요 시간	
Train 학습	1 Hour 50 Min
Model Load	1.7 Sec
Pred. 예측	2 Min
하이퍼 파라미터	
batch_size	256
emb_dim	128
layer_dim	128
epochs	20
learning_rate	0.001

```
==========================================================
Layer (type:depth-idx)                   Param #
==========================================================
NeuMF                                    --
├─Embedding: 1-1                         8,456,448
├─Embedding: 1-2                         6,634,752
├─Embedding: 1-3                         8,456,448
├─Embedding: 1-4                         6,634,752
├─Embedding: 1-5                         406
├─Sequential: 1-6                        --
│    └─Linear: 2-1                       135,168
│    └─ReLU: 2-2                         --
│    └─Dropout: 2-3                      --
│    └─Linear: 2-4                       32,896
│    └─ReLU: 2-5                         --
│    └─Dropout: 2-6                      --
├─Linear: 1-7                            385
==========================================================
Total params: 30,351,255
Trainable params: 30,351,255
Non-trainable params: 0
==========================================================
```

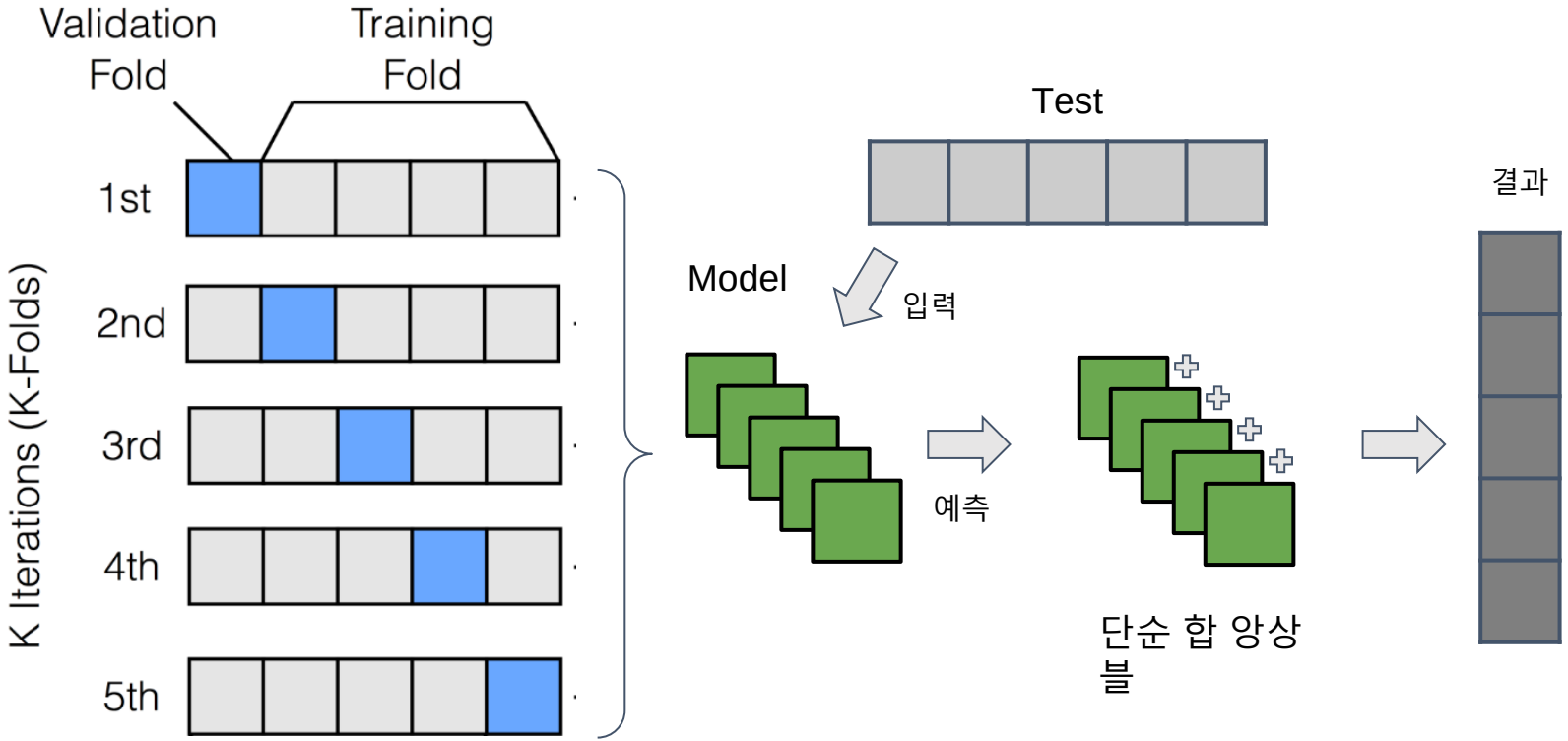
모델 구조

03. 데이터 전처리 후 성능 변화



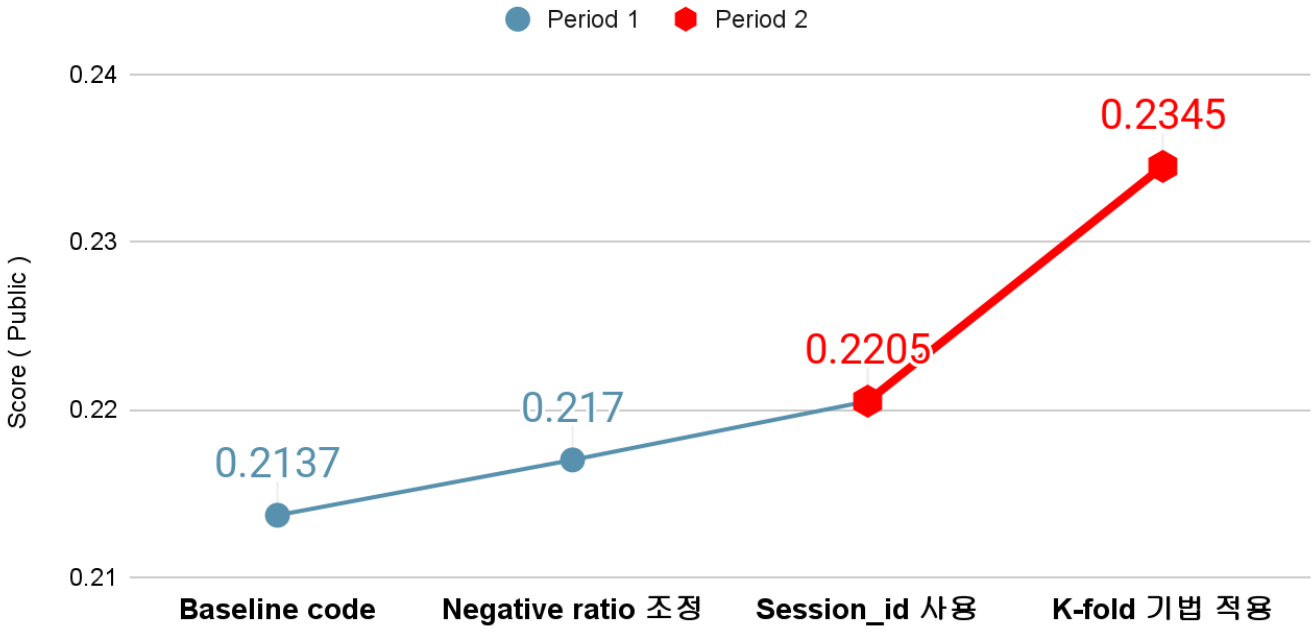
Negative ratio 조정 및 Session_id 를 기준으로 중복되는 데이터 제거 후, 성능이 향상

04. 사용한 기법 : K-Fold Cross Validation



05. K-Fold 교차 검증 후 성능 향상

K-fold 기법 적용 후 성능 향상



K-fold 교차 검증 기법을 사용하고, 결과를 앙상블한 후 성능 향상

Chapter 3.

결론

01. 결론

간단한 방법으로 모델의 성능을 약 **10%** 향상

간단한 전처리	K-Fold	앙상블

02. 개선점_Optuna

Optuna: 하이퍼파라미터 최적화를 위한 프레임워크. 파라미터의 범위를 지정해주거나, 파라미터가 될 수 있는 목록을 설정하면 매 Trial 마다 파라미터를 변경하면서, 최적의 파라미터 튜닝

Step 1.

neg_ratio 튜닝을 통해 모델에 적용할 최적의 데이터 셋 생성

Step 2.

Optuna를 이용한 최적의 데이터 셋을 바탕으로 모델의 HyperParameter튜닝

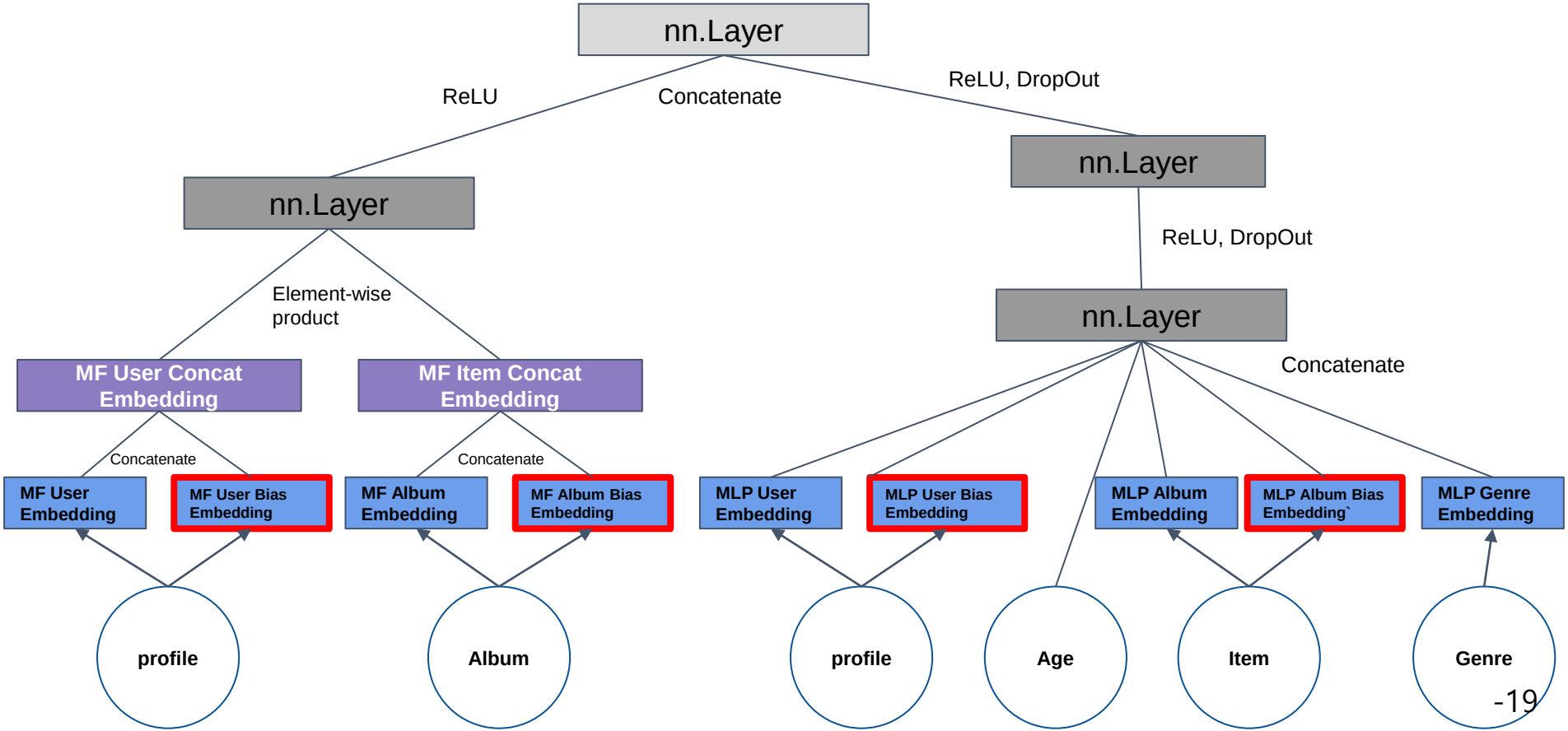
-emb_dim

-layer_dim

-learning_rate

-dropout

02. 개선점_모델 구조 개선



Thank You!