세계 추천시스템 대회 RecSys2023 Challenge에서 엔비디아를 꺾고 세계 7위, 국내 1위를 위한 여정

Corca ML Engineer 전태현

Profile

전태현

- Machine Learning Engineer @ Corca (2022.04.18 ~ Current)
- 광고 추천 ML 모델 개발
- LLM 기반 대화형 검색/추천 시스템 개발 중

RecSys Challenge 2023

ACM이 주관하는 학회 중 RecSys는 추천시스템 분야 학회에서 주관하는 대회

twitter, spotify 등 매년 다양한 대기업이 주최

올해는 인도의 SNS 회사인 ShareChat이 주관
온라인 광고가 노출이 된 이후 사용자가 해당 광고 상품을 다운로드할 확률을 예측하는 대회

Timeline

- 4월에 열려서 6월 말 마감.
- 대회 끝나고 3주 뒤 논문 제출
- 8월에 논문 Accpeted!
- 9월에 논문 발표를 위해 학회로 해외 출장!

RecSys Challenge 2023

- 1위는 Layer 6 AI로, 수 년간 꾸준히 순위권에 드는 멋진 회사
- 2위는 Intel, 3등은 화웨이였다!
- 4, 5, 6위는 어떤 기업인지 모르나, 최종 논문 발표에는 없었다
- 8위는 엔비디아로 2020년, 2021년 우승자. 대회를 위한 팀이 존재

Position	Team Name	Score	Captain	Submission Date	Last Updated Time	Number of Submissions
1	Layer 6 AI	5.744062	hocayi	23/06/2023 12:31	23/06/2023 12:32	56
2	LearningFE	5.892977	XC	23/06/2023 13:56	23/06/2023 14:02	74
3	hahaha	5.904369	doubleQ	22/06/2023 19:17	22/06/2023 19:18	223
4	Ainvest	5.949816	AIME	23/06/2023 12:18	23/06/2023 12:23	8
5	Shield	5.958641	ShawnSong	22/06/2023 15:47	22/06/2023 15:48	402
6	AmazMe	6.010449	AmazMe	23/06/2023 21:01	23/06/2023 21:01	72
7	Corca	6.015522	Taehee	23/06/2023 20:21	23/06/2023 20:22	702
8	NVIDIA RAPIDS	6.022057	Gilberto Titericz Junior	23/06/2023 21:46	23/06/2023 21:51	270
9		6.059065	aporia	18/06/2023 23:01	18/06/2023 23:02	867
10	PPRec	6.113701	Poovaiah	22/06/2023 05:36	22/06/2023 05:41	120
11	Sam	6.115381	Sejoon	23/06/2023 18:31	23/06/2023 18:32	1435

어려웠던 점

대회 운영 방식

- 리더보드 Score 계산 방식이 이상해서 재채점해서 순위 변동이 일어남
- 주최측에서 Metric 계산식을 제공했지만 중요한 상수를 제공하지 않아서 모델링 조금만 바꿔어도 점수 편차가 너무 심했음
- 데이터 익명화로 인해 피쳐가 어떤 의미를 갖는지 모르기 때문에 데이터 분포나 상관관계 등 통계적인 정보들에만 의존을 해서 모델링을 진행함
- 대회 운영을 매끄럽지 못해서 미안하다고 막판에 대회를 4일을 연장함 (?!?!?!?!)

Dataset

ShareChat Dataset Information

- 21일의 train data를 가지고 22일의 유저의 install 확률을 예측하는 대회
- 출제자의 의도 중 하나가 Data Privacy였기 때문에 모든 데이터는 익명화 되어 있음
- 심지어 전처리되어서 예측하기가 더 어려웠음
- 데이터에 제공된 Feature는 Categorical, Numerical, Embedding .. 등등
- 그럼 뭐하나.. 무슨 의민지 모르는데..
- 보통 Click->Install의 funnel을 가지는데, 이 데이터 셋의 일부는 Click 없이도 Install이 발생했다

(is_clicked, is_installed)	Distribution
(0, 0)	67.73%
(1, 0)	14.87%
(0, 1)	10.29%
(1, 1)	7.11%

Our Solution

Our Key Points: Feature Interaction & Ensemble

- Data Analysis & EDA
    - Find out Domain Knowledge Features
    - Find out Correlated Features
- Feature Engineering
    - Frequency Encoding
    - Target Encoding
    - Elapsed & Recent Days
- Modeling
    - Tree based-Model
    - Multi-Task Model
    - Deep Learning Model
    - DCAF (Our Proposed Model)
- Ensemble

Our Solution

Our Key Points: Feature Interaction & Ensemble

- Data Analysis & EDA
  - Find out Domain Knowledge Features
  - Find out Correlated Features
- Feature Engineering
  - Frequency Encoding
  - Target Encoding
  - Elapsed & Recent Days
- Modeling
  - Tree based-Model
  - Multi-Task Model
  - Deep Learning Model
  - DCAF (Our Proposed Model)
- Ensemble

Our Solution – DCAF (Deep Cross Attention Factorizational Machine)

어떻게 하면 피쳐들간의 관계를 극대화해서 잡아낼 수 있을까?

- Feature Interaction을 극대화해서 잡아보자!
  - Attention Network
  - Cross Network
  - Factorization Machine

=> 3개를 결합해보자!

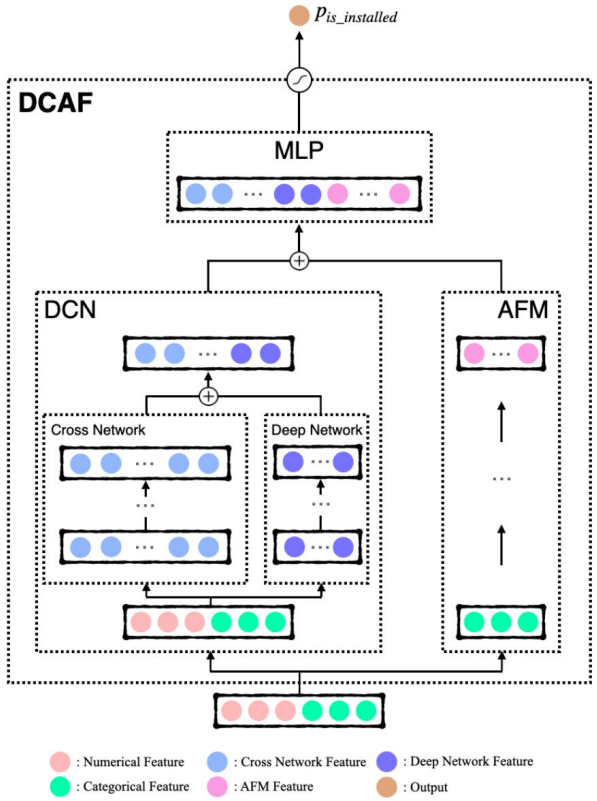
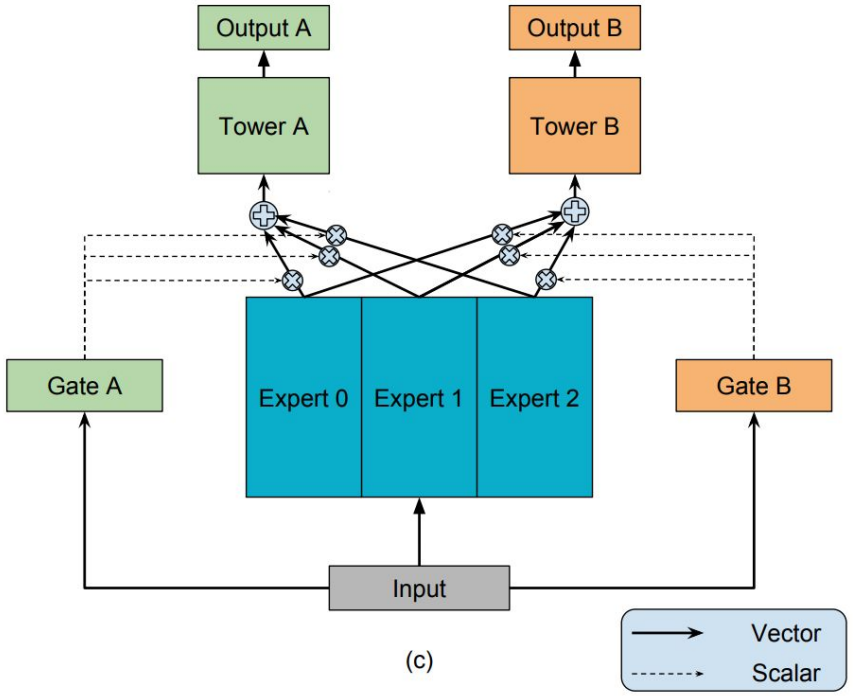


Fig. 3. Deep Cross Attentional Factorization Machine architecture

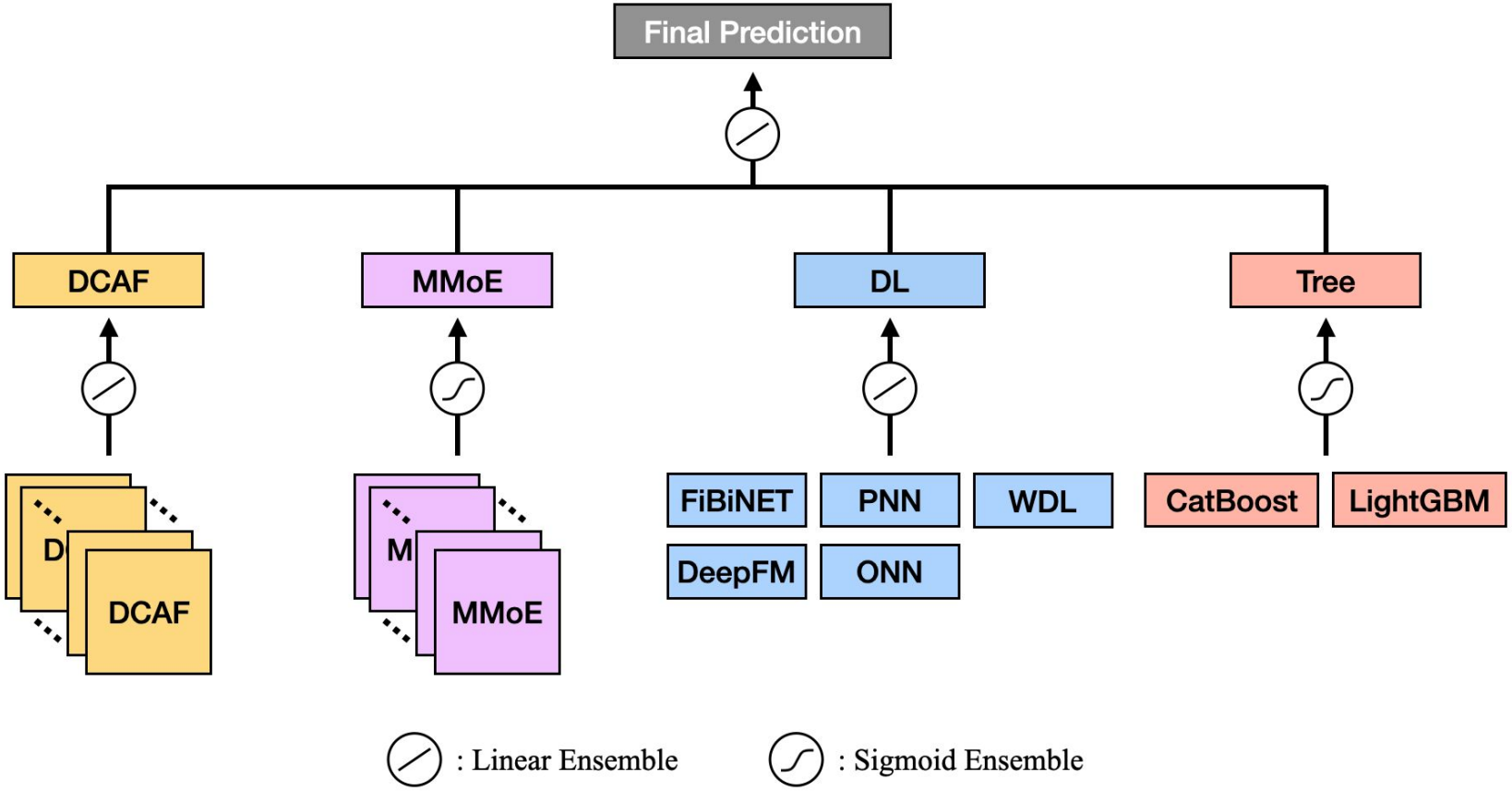
Our Solution – MultiTask Learning

Click이 일어나야 install이 발생한다. 이 말은 click과 install의 상관관계가 높다는 뜻

- click 정보를 어떻게 잘 사용할 수 있을까?
- click과 install이 동시에 label로 주기 때문에 학습이 어려움 (테스트 때 feature로 사용할 수 없음)
- click과 Install를 동시에 학습하는 Multi-Task Model 사용
- 그 중에서 MMoE라는 모델 사용



Our Solution – Ensemble



LeaderBoard

Position	Team Name	Score	Captain	Submission Date	Last Updated Time	Number of Submissions
1	Layer 6 AI	5.744062	hocayi	23/06/2023 12:31	23/06/2023 12:32	56
2	LearningFE	5.892977	XC	23/06/2023 13:56	23/06/2023 14:02	74
3	hahaha	5.904369	doubleQ	22/06/2023 19:17	22/06/2023 19:18	223
4	Ainvest	5.949816	AIME	23/06/2023 12:18	23/06/2023 12:23	8
5	Shield	5.958641	ShawnSong	22/06/2023 15:47	22/06/2023 15:48	402
6	AmazMe	6.010449	AmazMe	23/06/2023 21:01	23/06/2023 21:01	72
7	Corca	6.015522	Taehee	23/06/2023 20:21	23/06/2023 20:22	702
8	NVIDIA RAPIDS	6.022057	Gilberto Titericz Junior	23/06/2023 21:46	23/06/2023 21:51	270
9		6.059065	aporia	18/06/2023 23:01	18/06/2023 23:02	867
10	PPRec	6.113701	Poovaiah	22/06/2023 05:36	22/06/2023 05:41	120
11	Sam	6.115381	Sejoon	23/06/2023 18:31	23/06/2023 18:32	1435

우승할 수 있었던 요인

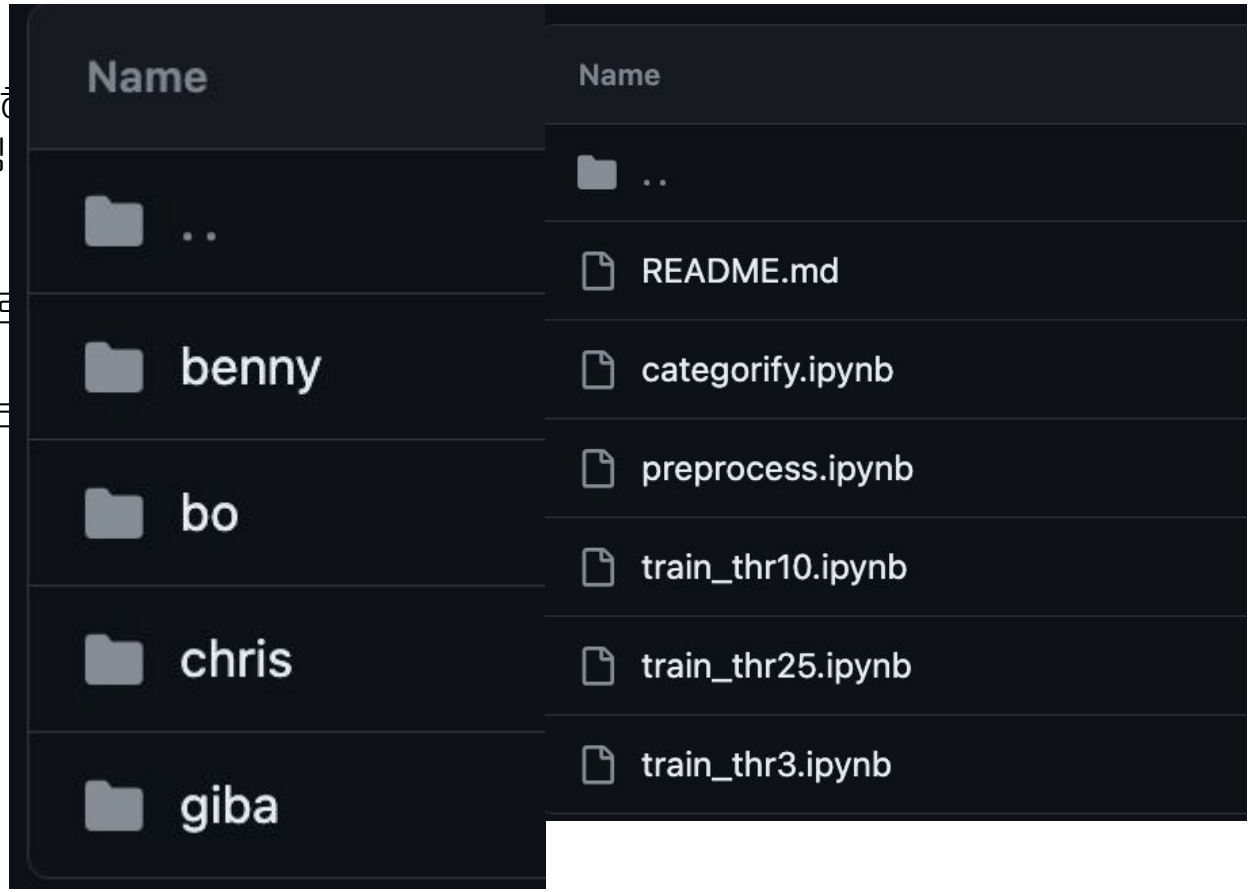
- 하고 있는 프로젝트인 DSP와 유사한 데이터와 도메인 특성을 가짐
- 중요한 것은 꺾이지 않는 마음.. 3개월동안 주말이나 평일 밤에 모여서 업무 지장되지 않는 선에서 노력했다.
    - 탈주자도 없었다!
    - 그래서 서로 으쌰으쌰 할 수 있었음
- 모두가 올라운더였음. 전처리-EDA-모델링 전부 가능해서 협업이 원할했음
    - 한명이 SOTA를 찍고 솔루션을 올리면 다른 팀원들이 그것에 기반한 아이디어로 또 SOTA를 찍고 ..
    - 양성 피드백 무한 반복
- 모델이 학습 속도가 굉장히 빨랐기 때문에 실험을 빠르고 많이 할 수 있는 시스템을 구축한 것
    - 모든 코드를 Slack으로 공유함
    - 처음에는 모든 실험 환경을 파이프라인화 했는데 이게 오히려 효율성을 저하시킴

=> 우리가 잘하고 있다는 것을 알 수 있었음

우승할 수 있었던 요인

- 하고 있는 프로젝트인 DSP와 유사한 데이터와 도메인 특성을 가짐
- 중요한 것은 꺾이지 않는 마음.. 3개월동안 주말이나 평일 밤에 모여서 업무 지장되지 않는 선에서 노력했다.
    - 탈주자도 없었다!
    - 그래서 서로 으쌰으쌰 할 수 있었음
- 모두가 올라운더였음. 전처리-EDA-모델링 전부 가능ㅎ
    - 한명이 SOTA를 찍고 솔루션을 올리면 다른 팀 찍고 ..
    - 양성 피드백 무한 반복
- 모델이 학습 속도가 굉장히 빨랐기 때문에 실험을 빠ㄹ
    - 모든 코드를 Slack으로 공유함
    - 처음에는 모든 실험 환경을 파이프라인화 했는

=> 우리가 잘하고 있다는 것을 알 수 있었음



우승할 수 있었던 요인

- 하고 있는 프로젝트인 DSP와 유사한 데이터와 도메인
- 중요한 것은 꺾이지 않는 마음.. 3개월동안 주말이나
  노력했다.
  - 탈주자도 없었다!
  - 그래서 서로 으쌰으쌰 할 수 있었음
- 모두가 올라운더였음. 전처리-EDA-모델링 전부 가능
  - 한명이 SOTA를 찍고 솔루션을 올리면 다른 팀
    찍고 ..
  - 양성 피드백 무한 반복
- 모델이 학습 속도가 굉장히 빨랐기 때문에 실험을 빠

=> 우리가 잘하고 있다는 것을 알 수 있었음

```
for i, quant in enumerate(quantiles):
    ddf['quantile'] = (ddf['quantile']+(ddf['a_follower_count']>quant).astype('int8')).astype('int8')

ddf['date'] = cudf.to_datetime(ddf['timestamp'], unit='s')

ddf['a_ff_rate'] = (ddf['a_following_count'] / ddf['a_follower_count']).astype('float32')
ddf['b_ff_rate'] = (ddf['b_follower_count']  / ddf['b_following_count']).astype('float32')
ddf['ab_fing_rate'] = (ddf['a_following_count'] / ddf['b_following_count']).astype('float32')
ddf['ab_fer_rate'] = (ddf['a_follower_count'] / (1+ddf['b_following_count'])).astype('float32')
ddf['a_age'] = ddf['a_account_creation'].astype('int16') + 128
ddf['b_age'] = ddf['b_account_creation'].astype('int16') + 128
ddf['ab_age_dff'] = ddf['b_age'] - ddf['a_age']
ddf['ab_age_rate'] = ddf['a_age']/(1+ddf['b_age'])

## Normalize
for col in NUMERIC_COLUMNS:
    if col == 'tw_len_quest':
        ddf[col] = np.clip(ddf[col].values.get(),0,None)
    if ddf[col].dtype == 'uint16':
        ddf[col].astype('int32')

    if col == 'ab_age_dff':
        ddf[col] = ddf[col] / 256.
    elif 'int' in str(ddf[col].dtype) or 'float' in str(ddf[col].dtype):
        ddf[col] = np.log1p(ddf[col])

    if ddf[col].dtype == 'float64':
        ddf[col] = ddf[col].astype('float32')

## get categorical embedding id
for col in CAT_COLUMNS:
    ddf[col] = ddf[col].astype('float')
    if col in ['a_user_id','b_user_id']:
        mapping_col = 'a_user_id_b_user_id'
    else:
        mapping_col = col
    mapping = cudf.read_parquet(f'/raid/recsys_pre_TE_w_tok/workflow_232parts_joint_thr25/categories/unique.{
    mapping.columns = ['index',col]
    ddf = ddf.merge(mapping, how='left', on=col).drop(columns=[col]).rename(columns={'index':col})
    ddf[col] = ddf[col].fillna(0).astype('int')

label_names = ['reply', 'retweet', 'retweet_comment', 'like']
DONT_USE = ['timestamp','a_account_creation','b_account_creation','engage_time',
            'fold', 'dt_dow', 'a_account_creation',
            'b_account_creation', 'elapsed_time', 'links','domains','hashtags','id', 'date', 'is_train',
            'tw_hash0', 'tw_hash1', 'tw_hash2', 'tw_http0', 'tw_uhash', 'tw_hash', 'tw_word0',
            'tw_word1', 'tw_word2', 'tw_word3', 'tw_word4', 'dt_minute', 'dt_second',
           'dt_day', 'group', 'text', 'tweet_id', 'tw_original_user0', 'tw_original_user1', 'tw_original_user
            'tw_rt_user0', 'tw_original_http0', 'tw_tweet',]
DONT_USE = [c for c in ddf.columns if c in DONT_USE]
gc.collect(); gc.collect()

return ddf.drop(columns=DONT_USE)
```

Papers

Capturing Performance and Privacy by Assembling Avengers of Online Advertising

TAEHEE KIM, Corca, Inc., Republic of Korea
SEUNGYUN BAEK, Corca, Inc., Republic of Korea
TAEHYEON JEON, Corca, Inc., Republic of Korea
HOJIN JUNG, Corca, Inc., Republic of Korea
JOONHONG KIM, Corca, Inc., Republic of Korea
TAEHO LEE, Corca, Inc., Republic of Korea

RecSys 학회 후기

- Let's go Singapore~
- 확실히 추천 시스템이라는 주제 특성상 학회임에도 대부분 기업이 참가
    - 어딜 둘러봐도 대기업 출신 사람들..
- 유튜브, 넷플릭스, 다양한 대기업들도 우리랑 같은 고민을 한다.
- 가장 차이점은 **문제 정의**
- 최근 트렌드는 추천을 잘하기 위한 방법론보다
    - 우리가 잘하고 있는지, online-offline gap을 줄이는 방법과
    - 실험을 더 빠르고 효율적으로 하는 방법이 있는지에 대한 추천**"시스템"** 에 집중

Links

Github: https://github.com/corca-ai/recsys-challenge-2023

후기 Blog:
https://medium.com/corca/recsys-challenge-2023-%EA%B5%AD%EB%82%B4-1%EC%9C%84-%EC%BD%94%EB%A5%B4%EC%B9%B4-%ED%8C%80%EC%9D%98-%EB%8C%80%ED%9A%8C%EB%B6%80%ED%84%B0-%ED%95%99%ED%9A%8C%EA%B9%8C%EC%A7%80%EC%9D%98-%EC%97%AC%EC%A0%95-1-2-e384ad1aa238

Paper:

We are Hiring!

- Ad Tech 회사
- 회사 설립 2년 반, Pre-A 70억 유치
- 추천 시스템과 LLM 기반 Application 개발하실 분을 찾습니다~
- Solution
  - DSP
  - 개인화 배너 광고 추천 시스템 (ADCIO)
  - 대화형 추천 시스템 (ADCIO Agent)
  - LLM 기반 Solver (Agent Village)

Corca Opening Position

🚀 인재풀

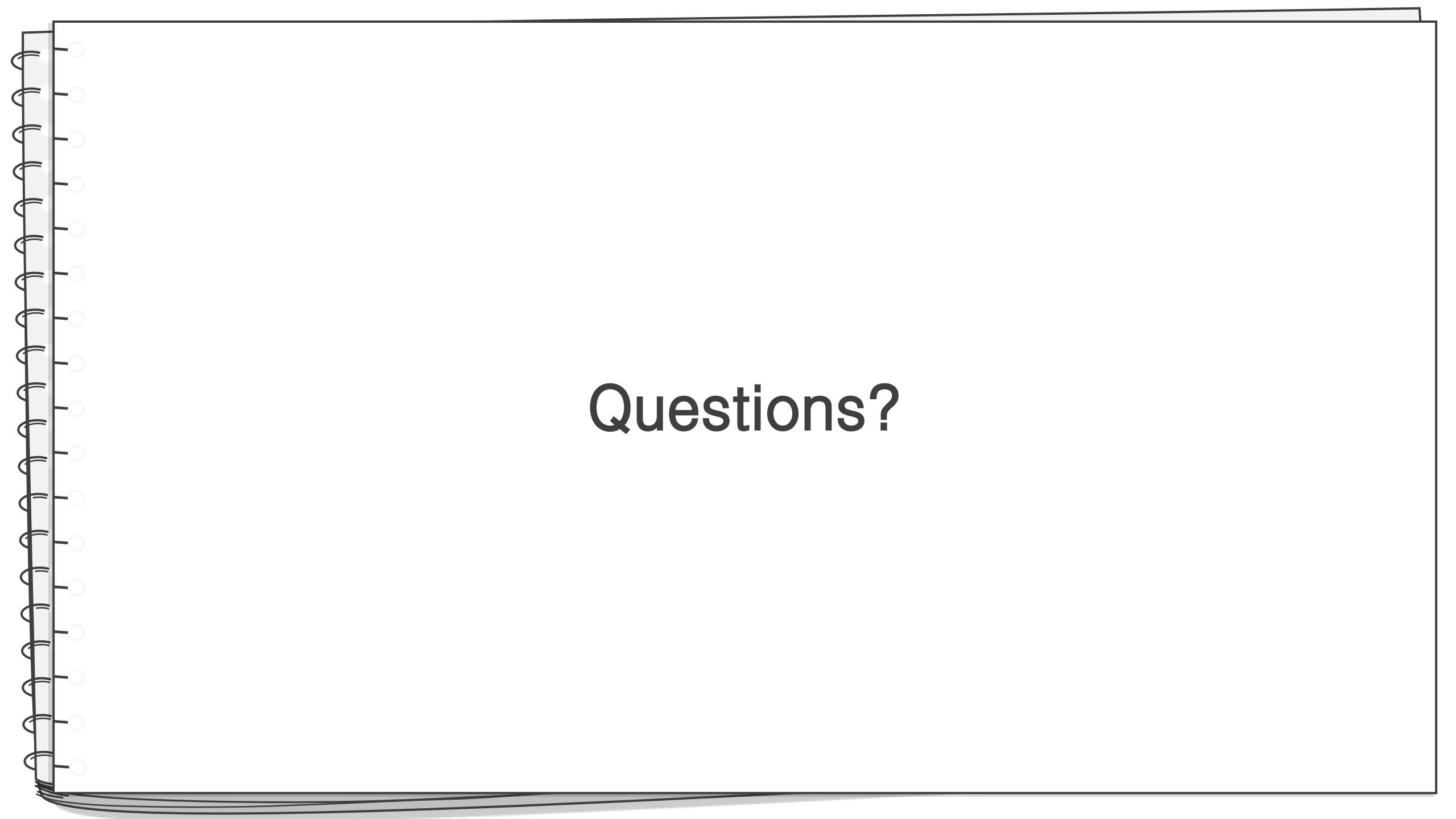
Position	채용 형태
인재풀 등록	신입 경력

🤖 Tech

Position	Part	채용 형태
ML Engineer	ML	신입 경력
Frontend Engineer (3년 이상)	Dev	경력
Backend Engineer	Dev	신입 경력
DevOps Engineer	Dev	신입 경력

🎁 Product

Position	채용 형태
Product Manager (3년 이상)	경력


Questions?


Thank You!