

Level2 DKT 프로젝트

캠퍼 솔루션 발표 및 프로젝트 랩업

Recsys 03조 (Recdol)

강찬미 T5009

박동연 T5080

서민석 T5102

이준영 T5158

주혜인 T5208

Contents

프로젝트 개요 및 목표

- 프로젝트 개요
- 프로젝트 목표

프로젝트 진행방법

- 협업 과정
- 타임라인

EDA

각 Approach별 설명

- Tabular
- Sequential
- Graph

학습 방법 및 전략

- 학습 방법 및 평가 지표
- 제출 전략

프로젝트 회고

The background is a solid light blue color. It features several overlapping circles of various shades of blue, some with a slight gradient. On the left side, there is a large, stylized number '1' in a light blue color, which is partially obscured by the circles.

프로젝트 개요 및 목표

프로젝트 개요

프로젝트 목표

프로젝트 개요



Level 2 프로젝트의 주제인 DKT에 대해서 설명합니다.

D_{ee}p K_{now}ledge T_{racing}

"지식 상태"를 추적하는 딥러닝 방법론
학생 개인의 이해도에 맞춰 개인화된 교육을 제공

해당 대회에서는 학생이 풀 문제 목록과 정답 여부를 통해 최종 문제에 대한 정답 확률 예측

프로젝트 목표

 지난 level 1 프로젝트의 아쉬웠던 점을 토대로 저희 팀만의 프로젝트 목표를 수립하였습니다.

새로운 베이스라인 작성

제공된 베이스라인을 참고하여
우리 팀만의 자체적인 베이스라인 구축

스프린트 방식 도입

전반적인 계획 수립을 통한
체계적인 진행을 위해
스프린트 방식 도입

적극적인 Github 활용

이전보다 적극적인 issue 사용
PR을 통한 code review 활성화

다양한 Tool 사용

pytorch lightning, hydra, github
action 등과 같은 다양한 tool 경험해보고자 함

프로젝트 진행방법

협업과정

타임라인

협업 과정

노션에 Jira를 토대로 애자일 방법론을 적용한 칸반보드를 제작하여 협업을 수행하였습니다.

프로젝트 개요

Introduce

아쉬웠던 점 모음집

해결 과정 모음집

실험실

고민 해결 일지

성능 기록장

아카이브

Daily Report

Tips

회의록

Conventions

Milestones

Wrapup Report

Jira-ver-Recdol

백로그

스프린트

Jira-ver-Recdol do...

나는 지금

GRU 활용 모델 추가

t-fix up 도전

saint+ 파라미터 튜닝

gcn 모듈화!

feature 추가하는 중

스프린트 보드

스프린트 3 - 기간 : 5월 15일 (월) - 5월 22일 (월)

보드 | 타임라인 | 표

Not started 2

Sequential: GPT2 Exp

박동연

IMPL

Graph: Ultra GCN

주혜인

IMPL

1

In progress 6

Feature 추가

서민석

IMPL

Sequential : GRUATTN, GPT2

박동연

IMPL

Done 15

Graph: SrGNN 추가

주혜인

IMPL

Sequential : SAINT+

준영 이

IMPL

노션을 활용한 효율적인 협업

협업 과정

노션에 Jira를 토대로 애자일 방법론을 적용한 칸반보드를 제작하여 협업을 수행하였습니다.

프로젝트 개요

- Introduce
- 아쉬웠던 점 모음집
- 해결 과정 모음집

실험실

- 고민 해결 일지
- 성능 기록장

나는 지금

- GRU 활용 모델 추가
- t-fix up 도전
- saint+ 파라미터 튜닝
- gcn 모듈화!
- feature 추가하는 중

스프린트 보드

스프린트 3 - 기간 : 5

Not started 2

Sequential: GPT2 Exp

Graph: Ultra GCN

IMPL

고민 해결 일지

누군가 프로젝트 과정에서 필요하다고 생각되는 것을 적어두면, 피어세션 시간마다 확인하고 해결합니다.

표

유형	Aa 문제	해결사	해결
협업과정	나의 pre-commit 적용기	서민석, 주혜인	<input checked="" type="checkbox"/>
협업과정	ouput dir 학기		<input type="checkbox"/>
문서화	EDA한 내용도 노트북 형태로 깃허브에 푸시?		<input type="checkbox"/>
전처리	sequence 길이를 가변적으로 쓸 수 있도록 해주세요!	박동연	<input checked="" type="checkbox"/>
전처리	긴 길이의 sequence를 처음으로써 데이터 중장 효과를 볼 수 있도록 해주세요!		<input type="checkbox"/>
기타	양상을 soft voting을 위해 확률값을 submit 파일에 나오게 하기		<input checked="" type="checkbox"/>
기타	새로운 베이스라인이 필요합니다	박동연, 서민석, 주혜인	<input checked="" type="checkbox"/>
모델링	validation set 구축	서민석	<input checked="" type="checkbox"/>
전처리	User와 문제에 대하여 Moving Average를 저장하는 전처리 단계가 있으면 좋겠습니다!		<input type="checkbox"/>
전처리	lgbm에 label을 피쳐로 넣으면.. 생기는 오류	서민석, 주혜인, 강장미, 박동연	<input checked="" type="checkbox"/>

고민 해결 일지를 통한 문제 해결과정 문서화

협업 과정

노션에 Jira를 토대로 애자일 방법론을 적용한 칸반보드를 제작하여 협업을 수행하였습니다.

프로젝트 개요

Introduce

아쉬웠던 점 모음집

해결 과정 모음집

실험실

고민 해결 일지

성능 기록장

아카이브

Daily Report

Tips

회의록

Conventions

Milestones

Wrapup Report

Jira-ver-Recdol

백로그

스프린트

Jira-ver-Recdol do...

나는 지금

GRU 활용 모델 추가

t-fix up 도전 🤔

saint+ 파라미터 튜닝

gcn 모듈화!

feature 추가하는 중 🤔

스프린트 보드

스프린트 3 - 기간 : 5월 15일 (월) - 5월 22일 (월)

보드 타임라인 표

Not started 2

Sequential: GPT2 Exp

박동연

IMPL

Graph: Ultra GCN

주혜인

IMPL

In progress 6

Feature 추가

서민석

IMPL

Sequential : GRUATTN, GPT2

박동연

IMPL

Done 15

Graph: SrGNN 추가

주혜인

IMPL

Sequential : SAINT+

준영 이

IMPL

칸반 보드를 사용한 스프린트 단위 업무 관리

 노션에 Jira를 토대로 애자일 방법론을 적용한 칸반보드를 제작하여 협업을 수행하였습니다.

스프린트 목표

- ✓ 새로운 툴 도입 → PyTorch Lightning, Hydra

완료된 작업 목록

- Baseline 구축
 - Baseline - Tabular
- EDA - 결측치, 이상치 탐색
- Github 세팅
 - Github template

완료되지 않은 작업 목록

- Baseline - Sequential 구축
이유: 코드가 돌아가는 상태이기는 하나, 학습이 잘 되는지 성능 점검, 모델 추가, 모듈화, 로깅, wandb 작업이 이행되지 않았음
- Baseline - Graph
이유: lightning wrapper로 코드를 구현했지만 데이터 로드 과정에서의 오류를 아직 해결하지 못함. 이 외에도 부가기능(hydra, wandb, 각종 결과 파일 저장) 미완
- Github action
이유: 우선순위가 낮다고 생각해서 이번 스프린트에서는 진행 안 함. 담당자 할당 필요
- Validation set 구축
이유: Baseline - Tabular 작업이 계획했던 것보다 오래 걸려서 못했음

스프린트 단위 팀 자체 회고 진행

다양한 협업툴 사용 - Git

 pre-commit, github issue, github action을 사용해 생산성을 향상시켰습니다.

Git 컨벤션

: 커밋 메시지, github flow 전략 도입

Pre-commit 활용








: black 포맷터를 이용해 코드 스타일 통일

Github Issue 활용

: 작업할 목록을 Issue에 정리

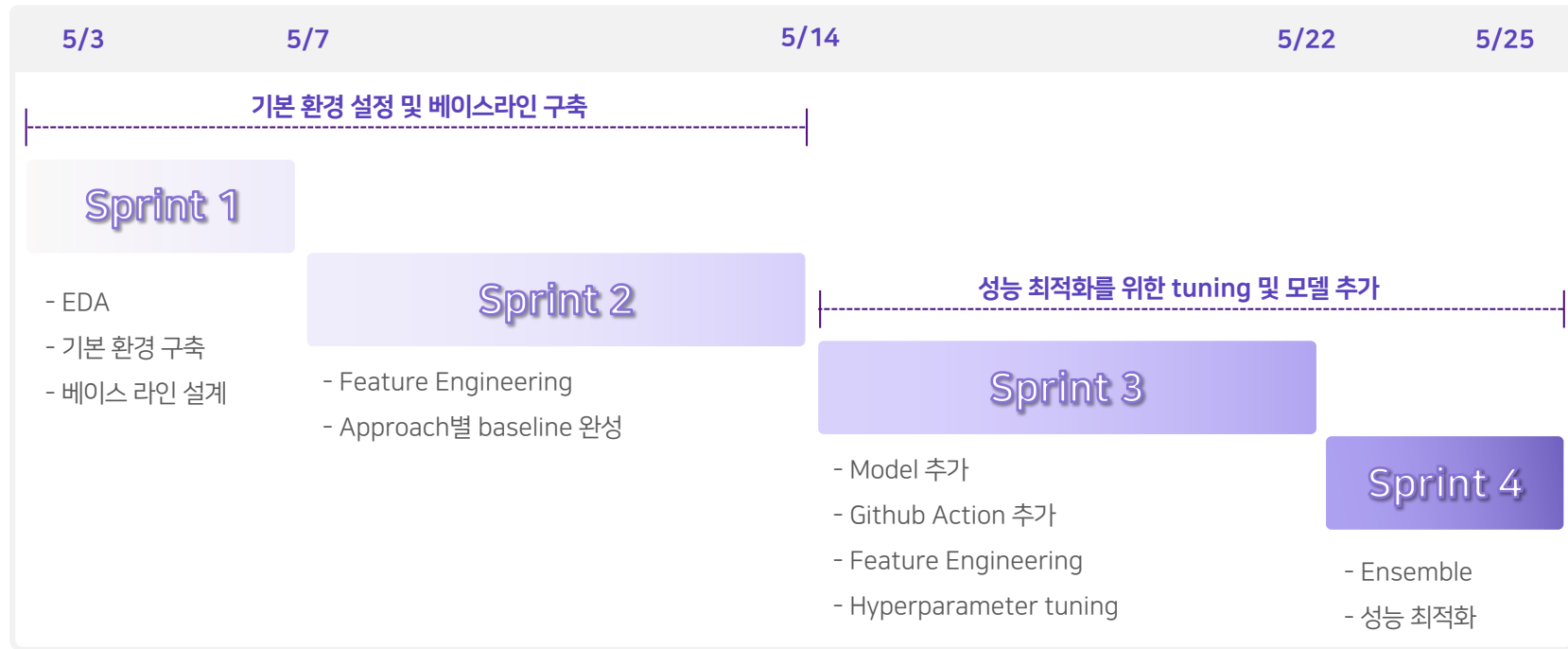
Github Action을 이용한 자동 테스트

: pytest를 이용한 모델 러닝 테스트

<input type="checkbox"/>	🔗 0 Open ✓ 90 Closed	Author ▾	Label ▾	Projects ▾	Milestones ▾	Reviews ▾	Assignee ▾	Sort ▾
<input type="checkbox"/>	🔗 fix: be longer session len in dummy test data ✓	bug	chore			👁 1		💬 1
	#150 by 2jun0 was merged 3 days ago • Approved							
<input type="checkbox"/>	🔗 refactor: tabular approach ✓	bug	enhancement	refactor		👁 4		💬 3
	#144 by alstjrdlzz was merged 3 days ago • Approved							
<input type="checkbox"/>	🔗 feat: add checkpoint to graph baseline ✓	enhancement				👁 1		💬 1
	#139 by juhyein was merged 4 days ago • Approved							
<input type="checkbox"/>	🔗 fix: add wandb logging & dropout parameter in config ✓	bug				👁 1		💬 3
	#138 by DyeonPark was merged 4 days ago • Approved							
<input type="checkbox"/>	🔗 feat: add T-fix up in SAINT+ ✓	enhancement				👁 1		💬 7
	#136 by kCMI113 was merged 4 days ago • Approved							
<input type="checkbox"/>	🔗 feat: implement XGboost in tabular approach ✓	enhancement				👁 1		💬 2
	#135 by 2jun0 was merged 5 days ago • Approved							
<input type="checkbox"/>	🔗 feat: add linear_warmup scheduler ✓	enhancement				👁 1		💬 4
	#134 by kCMI113 was merged 4 days ago • Approved							



4주간의 스프린트 단위 프로젝트 타임라인입니다.





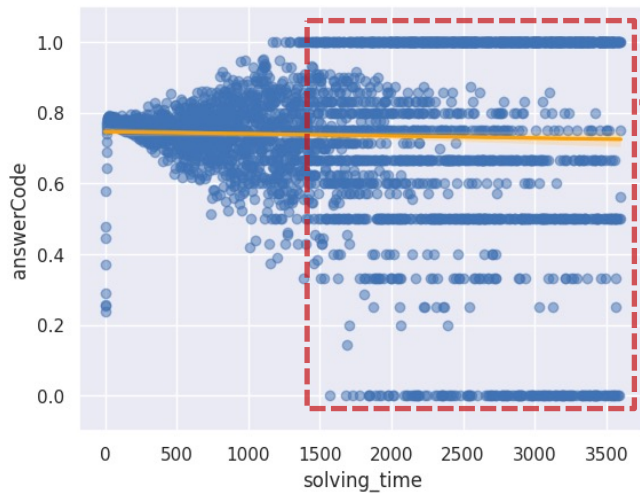
EDA

EDA : 문제 풀이시간



풀이시간은 (다음 행 Timestamp) - (현재 행 Timestamp) 입니다.

· 문제 풀이시간별 정답률



풀이시간이 길어질수록 낮은 정답률의 비중이 높아짐

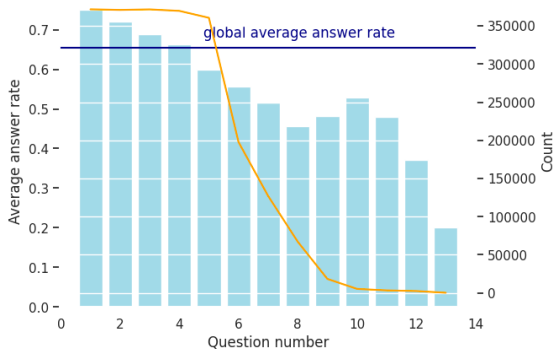


현재 문제의 풀이시간은 미래정보이기 때문에,
직전 문제의 풀이시간을 모델에 활용

EDA : Question Num

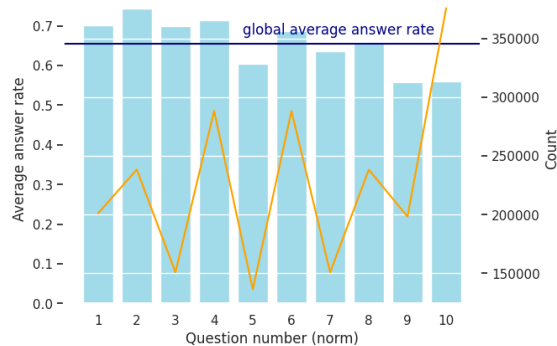
🐱 Question Num은 assessmentItemID의 뒷번호 3개를 의미합니다. ex) A000000XXX

• Question Num별 정답률



문제 번호가 커질수록 정답률이 낮아짐

• Normalized Question Num별 정답률



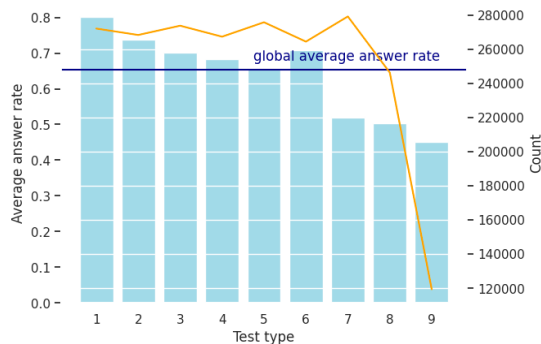
Normalized Question Num은 정규화된 Question Num(1~10)
시험지(testId) 마다 문제 수가 달라 이 값을 모델에 활용

EDA : Test Type



Test Type은 testId의 3번째 숫자를 의미합니다. ex) AOX0000000

· Test Type별 정답률



Test Type이 커질수록 정답률이 낮아짐

· Test Type과 KnowledgeTag의 관계

KnowledgeTag	
testId	
1	[5485, 7581, 5834, 7593, 6308, 7597, 7600, 759...
2	[8132, 8134, 8135, 8137, 8136, 8091, 8092, 809...
3	[407, 307, 7309, 7310, 308, 332, 331, 7321, 41...
4	[2048, 2047, 2049, 2050, 2053, 11214, 2054, 20...
5	[2617, 2619, 3691, 3682, 202, 3751, 3752, 3753...
6	[7224, 7225, 7226, 7227, 7228, 7229, 7230, 586...
7	[4746, 3804, 3806, 164, 162, 5656, 5782, 5620,...
8	[4605, 1394, 1396, 1397, 1395, 1356, 4657, 465...
9	[4697, 10174, 78, 4699, 4709, 4723, 4724, 4725...
중복된 tag {7863}	

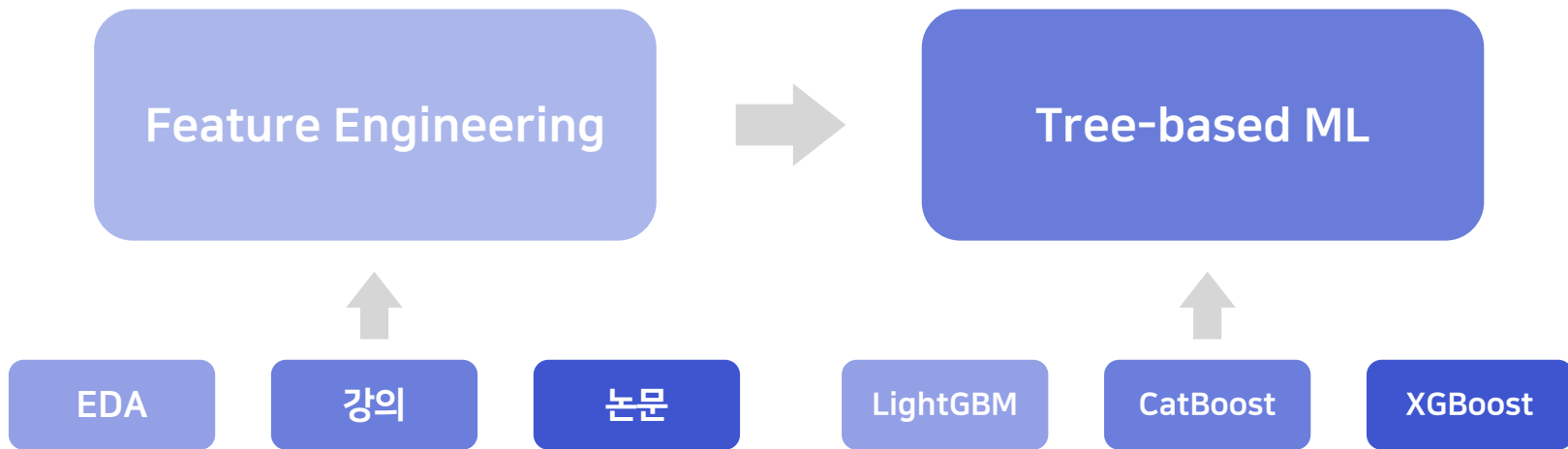
TestType과 Tag는 대분류, 소분류로 생각할 수 있음
Tag는 하나의 TestType에 속해있음 (Tag#7863제외)

4 각 approach별 설명

Tabular
Sequential
Graph



Tabular Approach 진행과정을 설명합니다.



Tabular : 파생 변수 생성

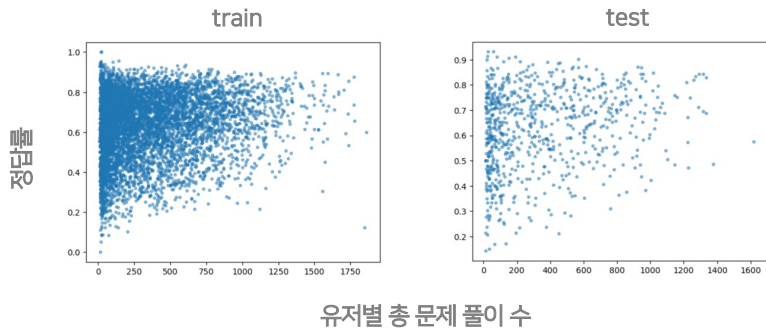
Tabular Sequential Graph



대표적인 파생변수 생성 전략에 대해 설명합니다.

EDA

train 데이터와 test 데이터에서 공통적으로 관찰되는 데이터의 특성을 모델에 학습시킴



총 문제 풀이 수가 증가함에 따라
정답률이 0.5 ~ 0.9 사이로 수렴하는 특성 관찰



'유저별 총 문제 풀이 수' 파생변수 생성

논문

여러 DKT 논문에서 '문제'와 '정답여부'에 대한 임베딩을 feature로 채택하는 것에 착안



 데이터 버저닝을 도입하여 실험에 소요되는 시간을 단축시킨 과정을 설명합니다.

문제점

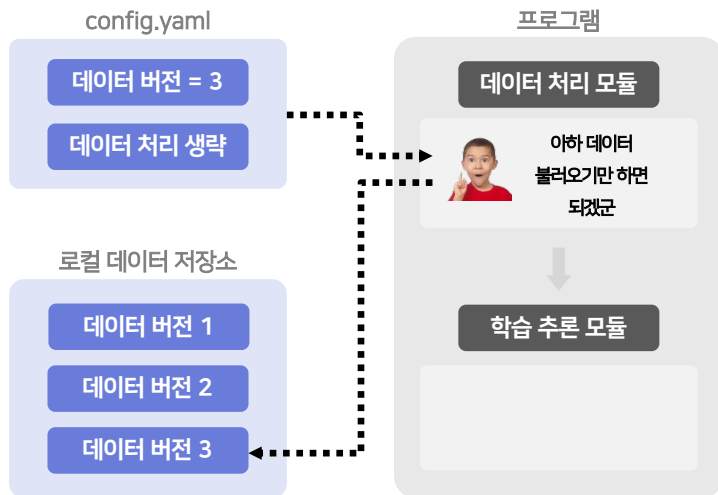
프로그램을 실행하면 데이터 처리부터 추론까지 end to end로 진행되는 로직



???: 데이터 처리는 처음 한번만 진행하고 처리된 데이터를 저장해서 사용하면 되지 않을까?

해결

데이터 버저닝 도입을 통한 데이터 처리 반복 최소화





Sequential Approach 기반으로 다양한 모델들을 사용해보았습니다.

LSTM

Baseline에서 제공된 모델로, 자체적인 베이스라인을 구축하고 정상적으로 동작하는지 확인하는 것에 주로 사용하였습니다

LSTMATTN

Baseline에서 제공된 모델로, LSTM에 Attention을 더한 모델입니다

GRUATTN

유저 아이디별로 묶었을 때, 데이터의 수가 적은 것을 감안하여 LSTM보다 일반적으로 더 적은 양의 데이터셋에서 잘 동작하는 GRU에 Attention을 더한 모델입니다

BERT

Baseline에서 제공된 모델로, Transformer의 인코더를 사용하는 모델입니다

GPT2

데이터를 증강한 후 커진 규모의 데이터셋에 대해서 잘 수행될 것으로 추측되는 GPT2 모델을 Transformer 라이브러리에서 호출함으로써 사용하였습니다

LQTR

강의의 실습 코드를 토대로 구현한 모델로, 트랜스포머 인코더, LSTM, DNN을 더했으며 인코더에서 마지막 쿼리만 사용하여 낮은 복잡도를 갖습니다

SAINT+

강의와 논문을 토대로 직접 구현한 모델로, Transformer에 시간 정보를 활용한 모델입니다



Sequential Approach 기반으로 다양한 모델들을 사용해보았습니다.

LSTM

Baseline에서 제공된 모델로, 자체적인 베이스라인을 구축하고 정상적으로 동작하는지 확인하는 것에 주로 사용하였습니다

LSTMATTN

Baseline에서 제공된 모델로, LSTM에 Attention을 더한 모델입니다

✓
GRUATTN

유저 아이디별로 묶었을 때, 데이터의 수가 적은 것을 감안하여 LSTM보다 일반적으로 더 적은 양의 데이터셋에서 잘 동작하는 GRU에 Attention을 더한 모델입니다

BERT

Baseline에서 제공된 모델로, Transformer의 인코더를 사용하는 모델입니다

✓
GPT2

데이터를 증강한 후 커진 규모의 데이터셋에 대해서 잘 수행될 것으로 추측되는 GPT2 모델을 Transformer 라이브러리에서 호출함으로써 사용하였습니다

✓
LQTR

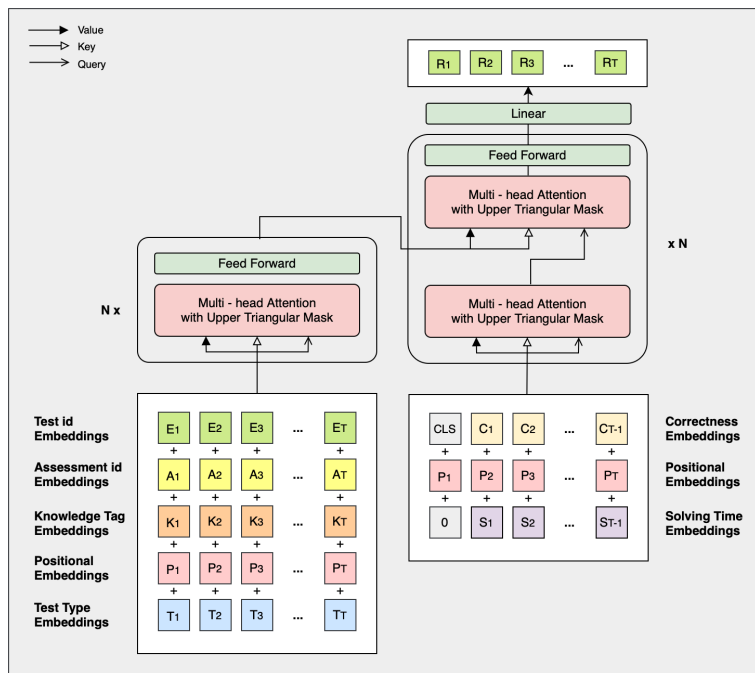
강의의 실습 코드를 토대로 구현한 모델로, 트랜스포머 인코더, LSTM, DNN을 더했으며 인코더에서 마지막 쿼리만 사용하여 낮은 복잡도를 갖습니다

✓
SAINT+

강의와 논문을 토대로 직접 구현한 모델로, Transformer에 시간 정보를 활용한 모델입니다



I-Scream dataset에 맞게 입력 임베딩을 수정하였습니다.



입력 임베딩

Elapsed Time과 Lag Time을 Solving Time(0~300s)로 구현

Test Id, Test Type 추가

T-fixup

모델의 깊이에 따라 transformer parameter를 scaling하는 가중치 초기화

Data Augmentation

sliding window 방식으로 하나의 sequence에서 여러 개로 증강

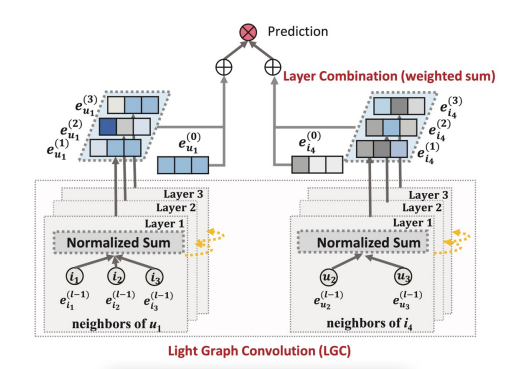
Parameter Customization

논문에 제시된 모델 구조
 encoder layer 4, decoder layer 4, head 8
 ↓
 encoder layer 2, decoder layer 2, head 2



LightGCN의 구조와 실험과정에 대해 소개합니다.

모델 구조



출처: LightGCN Simplifying and Powering Graph Convolution Network for Recommendation

이웃 노드의 임베딩의 가중합으로 GCN을 적용한 모델

실험 과정

LightGCN이 소개된 논문의 Experiment 참고

layer의 범위 2~4로 제한

배치 단위 학습 & epoch 증가



layer의 범위는 작을수록 auc가 높았으며
epoch는 900회가 넘었을 때 수렴하는 양상

학습 방법 및 전략

학습 방법 및 평가 지표

제출 전략

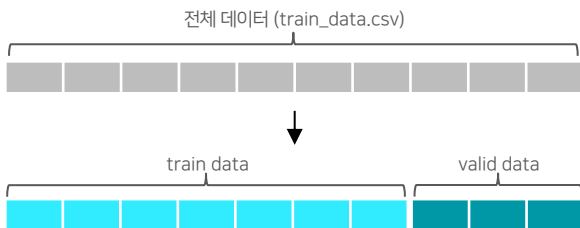
🌱 학습 방법 및 평가 지표



우리 팀이 사용한 학습 방법과 공통 평가 지표를 소개합니다.

Holdout

train_data.csv 전체를 7:3 비율로 train 데이터와 valid 데이터로 나누어서 학습 및 검증



대회에서 사용하는 지표 사용

train_auc

train_acc

train_loss

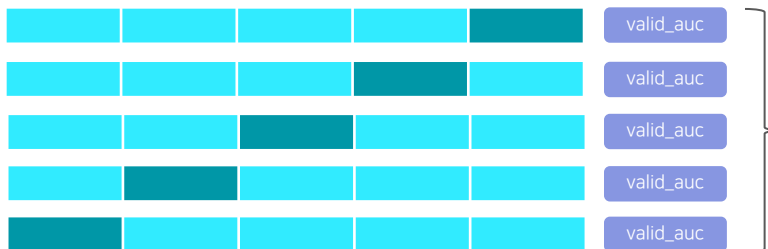
valid_auc

valid_acc

valid_loss

k-fold CV

train_data.csv 전체를 5개의 동일한 크기의 폴드로 나누어서 cross validation 방법 적용



대회 지표 + 자체 성능평가 지표 (cv_score)

cv_score

&

train_auc

train_acc

train_loss

valid_auc

valid_acc

valid_loss



앙상블 전략

 성능 향상 및 최종 제출을 위해 사용한 제출 전략을 토대로 앙상블을 사용하였습니다.

Step 1

각 모델 따라서 정답을 예측하는 방법 및 추이, 잘 예측하는 경우가 다를 것이라고 가정



앙상블 전략



성능 향상 및 최종 제출을 위해 사용한 제출 전략을 토대로 앙상블을 사용하였습니다.

Step 1

각 모델 따라서 정답을 예측하는 방법 및 추이, 잘 예측하는 경우가 다를 것이라고 가정

Step 2

각 모델이 예측한 값과 정답값의 분포 추이를 비교

앙상블 전략

 성능 향상 및 최종 제출을 위해 사용한 제출 전략을 토대로 앙상블을 사용하였습니다.

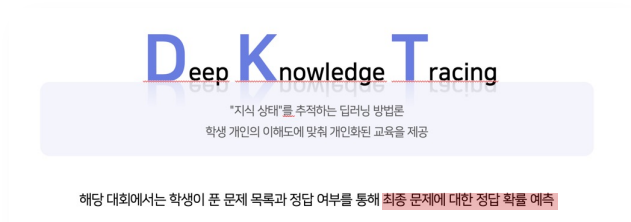
Step 1

각 모델 따라서 정답을 예측하는 방법 및 추이, 잘 예측하는 경우가 다들 것이라고 가정

Step 2

Validation Set 구성 방식

1page 프로젝트 개요



DKT의 목표와 부합하는지 객관적으로 평가하기 위해

test_data의 $n_u - 1$ 번째 문제를 validation set으로 구성

* n_u : user u 가 푼 문제 수

앙상블 전략

 성능 향상 및 최종 제출을 위해 사용한 제출 전략을 토대로 앙상블을 사용하였습니다.

Step 1

각 모델 따라서 정답을 예측하는 방법 및 추이, 잘 예측하는 경우가 다를 것이라고 가정

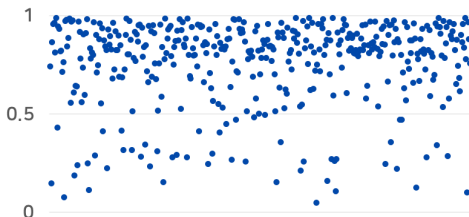
Step 2

각 모델이 예측한 값과 정답값의 분포 추이를 비교

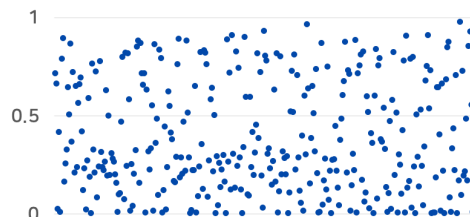
각 경우에 따라 잘 예측하지만,
오답의 경우 중간값으로 예측

SAINT+

정답값이 1인 경우



정답값이 0인 경우



앙상블 전략

 성능 향상 및 최종 제출을 위해 사용한 제출 전략을 토대로 앙상블을 사용하였습니다.

Step 1

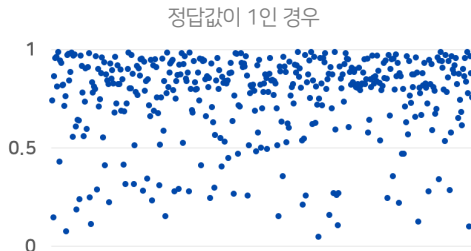
각 모델 따라서 정답을 예측하는 방법 및 추이, 잘 예측하는 경우가 다를 것이라고 가정

Step 2

각 모델이 예측한 값과 정답값의 분포 추이를 비교

각 경우에 따라 잘 예측하지만,
오답의 경우 중간값으로 예측

SAINT+

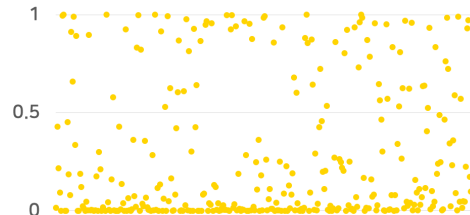
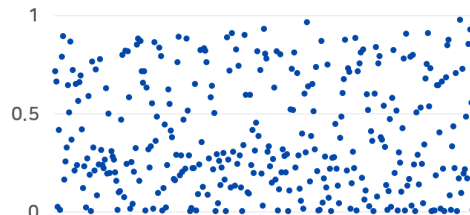


정답과 오답을 중간값보다는
극단값으로 예측하는 경향이 있음

LightGCN



정답값이 0인 경우



앙상블 전략

 성능 향상 및 최종 제출을 위해 사용한 제출 전략을 토대로 앙상블을 사용하였습니다.

Step 1

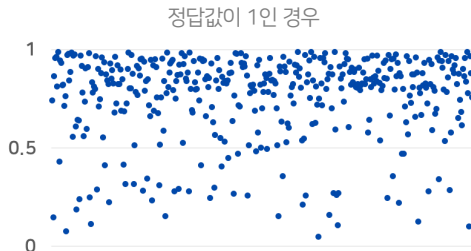
각 모델 따라서 정답을 예측하는 방법 및 추이, 잘 예측하는 경우가 다를 것이라고 가정

Step 2

각 모델이 예측한 값과 정답값의 분포 추이를 비교

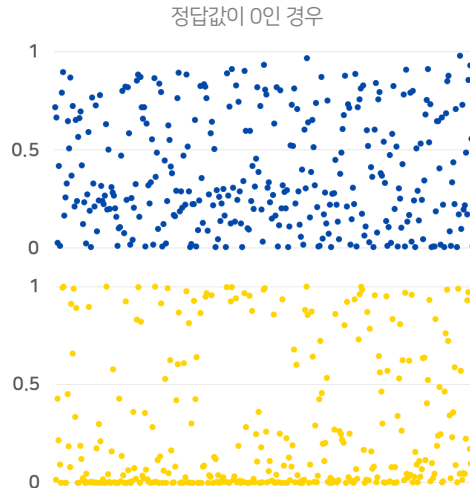
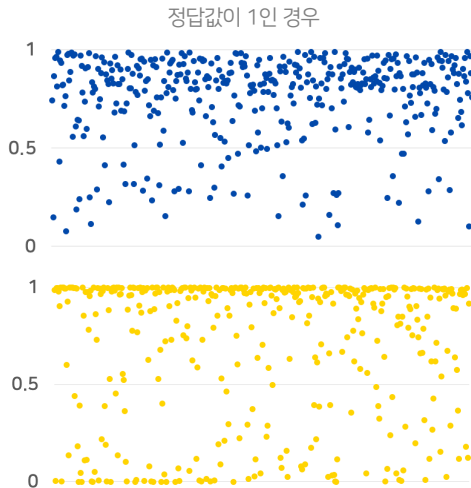
각 경우에 따라 잘 예측하지만,
오답의 경우 중간값으로 예측

SAINT+



정답과 오답을 중간값보다는
극단값으로 예측하는 경향이 있음

LightGCN



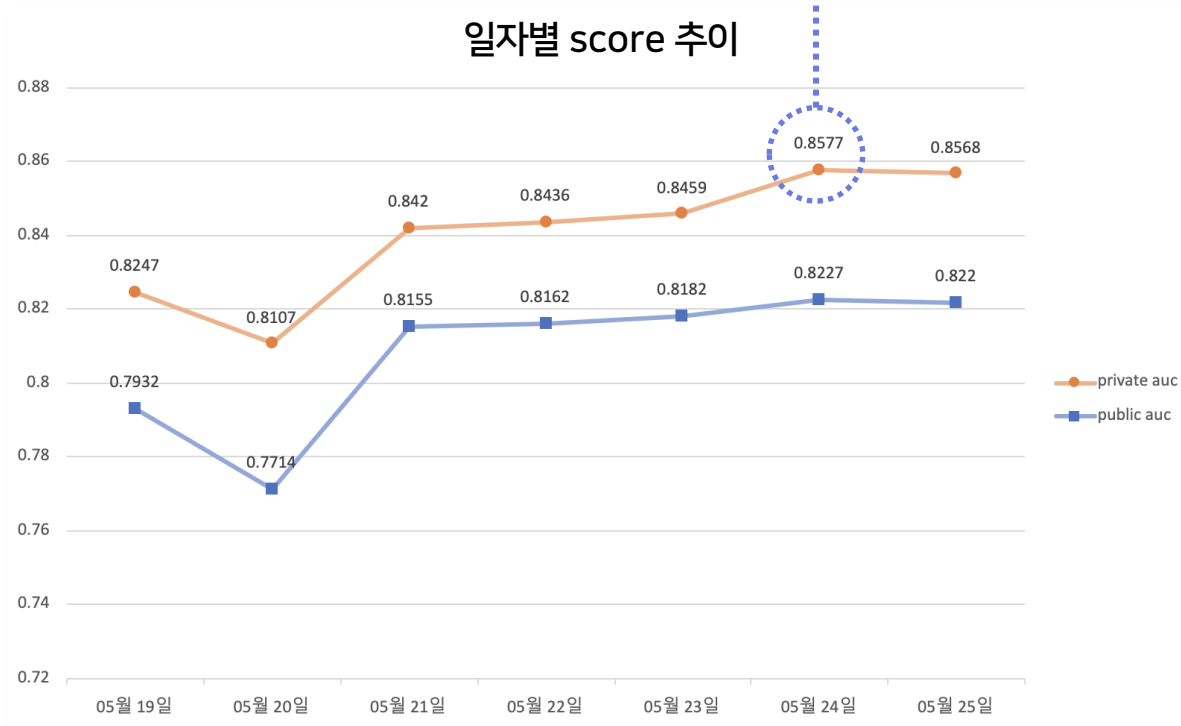
Step 3

서로 다른 예측 분포를 가진 모델들을 같이 다양한 기법으로 앙상블하여, 상호 보완하는 효과를 내는 것을 기대

🌱 성능 차이 및 솔루션 모델



프로젝트 스프린트4 기간 동안의 public/private score 추이입니다.



Simple weighted ensemble

SAINT+ (0.8)

LigthGCN (0.2)

Stacking ensemble

1st SAINT+

2nd SAINT+

LigthGCN

LightGBM

LSTMATTN

LSTM

솔루션 모델 조합 목록

The background is a solid light blue color. It features several overlapping circles of various shades of blue, some with a slight gradient. A prominent, thick white curved line starts from the left side, curves upwards and then downwards, forming a large, open 'C' shape. The text '프로젝트 회고' is written in white, bold Korean characters on the right side of the image.

프로젝트 회고

프로젝트 회고 : 배운 점



프로젝트를 하며 배운 점을 소개합니다.

1

코드 리뷰의 중요성 코드 리뷰를 통해 더 좋은 코드에 대해 고민하고 의논하기!



2jun0 2 weeks ago

반환 타입을 dict[str, Tensor]로 할 수 있을 것 같아요

2

함께 자라기를 통한 빠른 문제 해결 궁금하거나 풀리지 않는 점은 바로 게더에서 질문하기!

3

베이스라인 구축을 통한 모델 구조 및 프로세스 파악

4

새로운 툴(Pytorch lightning & Hydra)을 통한 효율적 실험 환경 구성

5

충분한 데이터의 중요성 augmentation으로 인한 성능 향상

프로젝트 회고 : 아쉬운 점



프로젝트를 하며 아쉬운 점을 소개합니다.

- 1 **시퀀스 구성 방식** 경우에 따라 한 시퀀스 내에서 시간의 범위가 좁은 경우와 긴 경우가 있는데, 시퀀스 구성을 시간 관점에서 해보면 어땠을까?
- 2 **ML 프로젝트에서의 Unit Test** 저희 팀은 아직 적절한 방법을 찾지 못했는데, 혹시 이번에 하신 팀이 계신다면 댓글이나 디엠으로 알려주세요!
- 3 **피처 관리 및 데이터 버저닝의 어려움**
- 4 **불필요한 로그 파일 관리** 새로운 툴은 편리했지만 각 프레임워크가 자동으로 생성하는 불필요한 로그 파일이 너무 많아졌다는 점!
- 5 **베이스 라인 구축 과정에서의 시간 소모**

감사합니다

Thank you

Recsys 03조 (Recdol)

강찬미 T5009

박동연 T5080

서민석 T5102

이준영 T5158

주혜인 T5208