



Networking Day

online

2021/06/22
10:00-17:20

boostcamp^{ai tech}

SGD

Share, **G**row, **D**ay by day

[DKT-10] No_Caffeine_No_Gain

T1194 정희석
T1168 이창우
T1155 이애나
T1119 안유진
T1098 선재우

SGD(Stochastic Gradient Descent)가 뭐라구?

- SGD는 느리지만, **다양한 방향**을 겪으며 Adam 보다 **Global Minimum**에 더 가깝게 나아가는 알고리즘입니다.

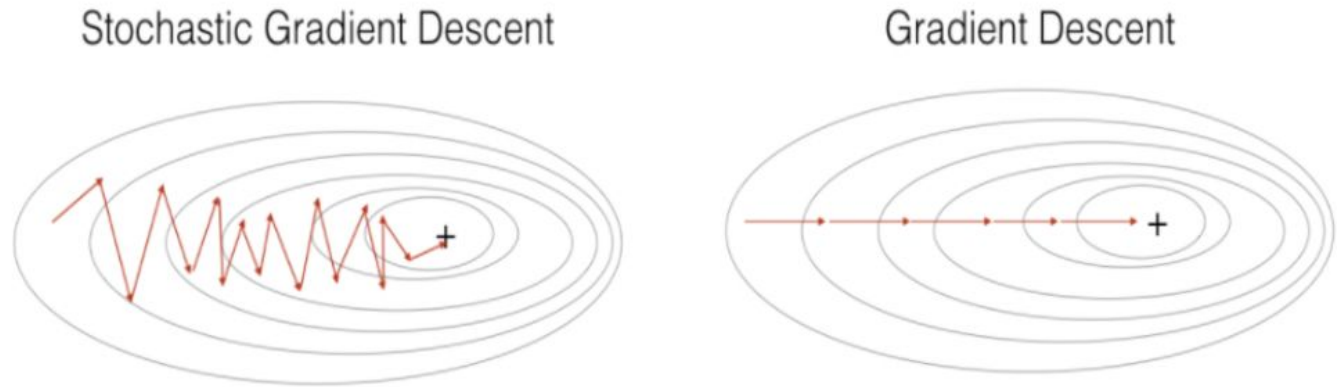


그림 출처: <https://engmrk.com/mini-batch-gd/>

우리는 왜 SGD인가?

- 우리 팀은 **Global Minimum**으로 가기 위해 **다양한 방향**을 지향합니다.
- 단, 나아간 길들을 팀과 **반드시 공유**하고 **적용**하여 더 좋은 **방향**으로 나아가기 때문입니다.

SGD

1. Share

1.1 Set up

1.2 목적이 뭐야?

2. Grow

2.1 Valid Set 에 관하여

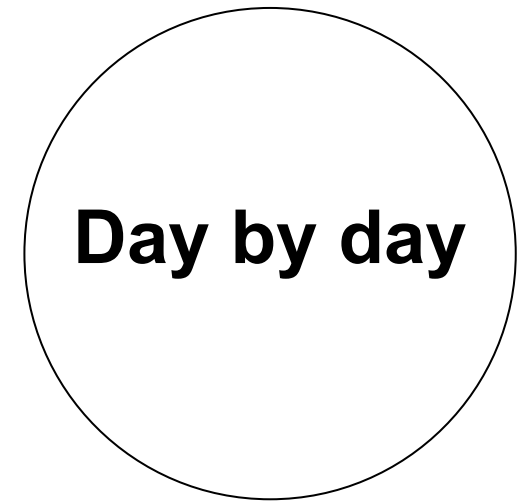
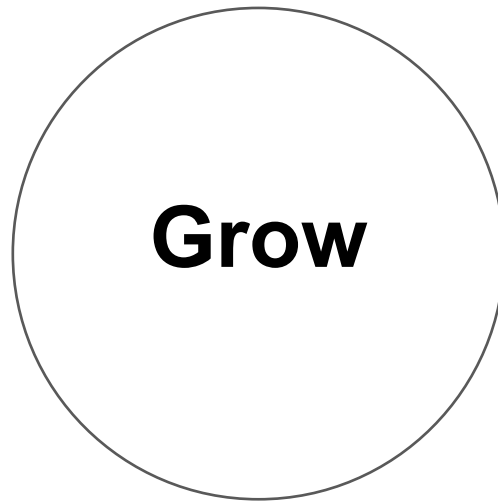
2.2 Transformer 성장기

2.3 Kaggle 대회 솔루션과 TABNET

3. Day by Day

3.1 발자취

Chapter 1



1.1 Set up

실험 노트

문제 정의

해결 아이디어

진행 상황

결과

평가

split_data 데이터 증강 재실험

Description	BERT 에서 효과가 있다!
시작일	2021년 5월 31일
제안자	Ⓜ HS J
실험자	Ⓜ HS J
종료일	2021년 6월 01일
진행상황	완료
카테고리	Data Processing
속성 추가	

📌 댓글 추가

문제 정의

ModuleList 가 적용된 코드로 split_data 증강에 대한 실험을 다시 수행

LSTM 과 BERT 에 실험

<https://www.notion.so/Feature-Engineering-63f6f1b386d0473c90e24e735a8b3d92>

1.1 Set up

제출 기록

boostcamp.stages.ai의 description == 제출 결과의 이름

이름은 실험내용_index

실험 노트

하이퍼파라미터

결과

실험 wandb 포함

BERT_decoder T-Fixup

≡ 변경 사항	Bert Decoder에 T-Fixup을 적용
≡ AUROC	0.7485
≡ ACCURACY	0.6720
👤 제출한 사람	 anna lee
≡ Created At	21.06.06.(Sun) 23:41
≡ 실험 노트	<u>T-Fixup 실험</u>
+ 속성 추가	

 댓글 추가

실험 노트

T-Fixup 실험

하이퍼파라미터

ANSWER_COLUMN=['answerCode'], EXCLUDE_COLUMN=['Timestamp', 'testId'], Tfixup=True, USERID_COLUMN=['userID'], USE_COLUMN=['KnowledgeTag', 'assessmentItemID',

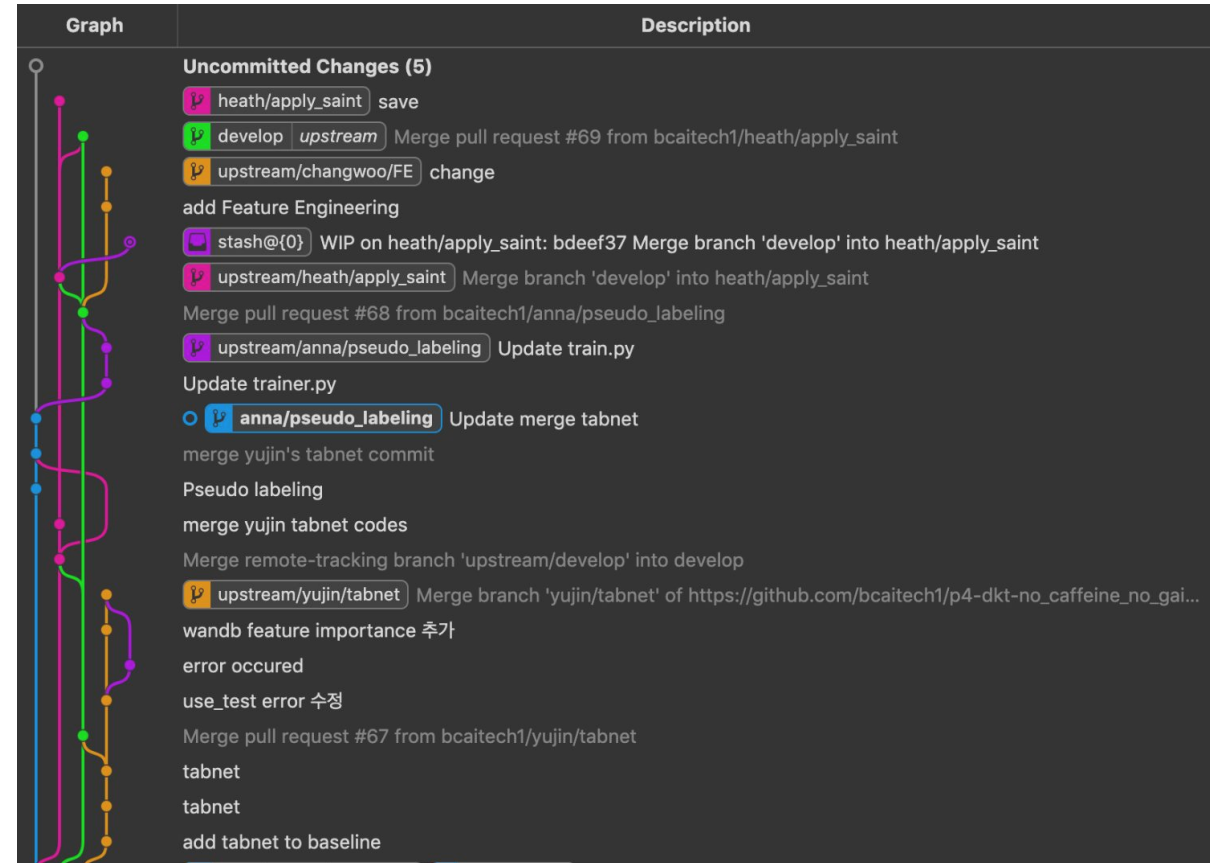
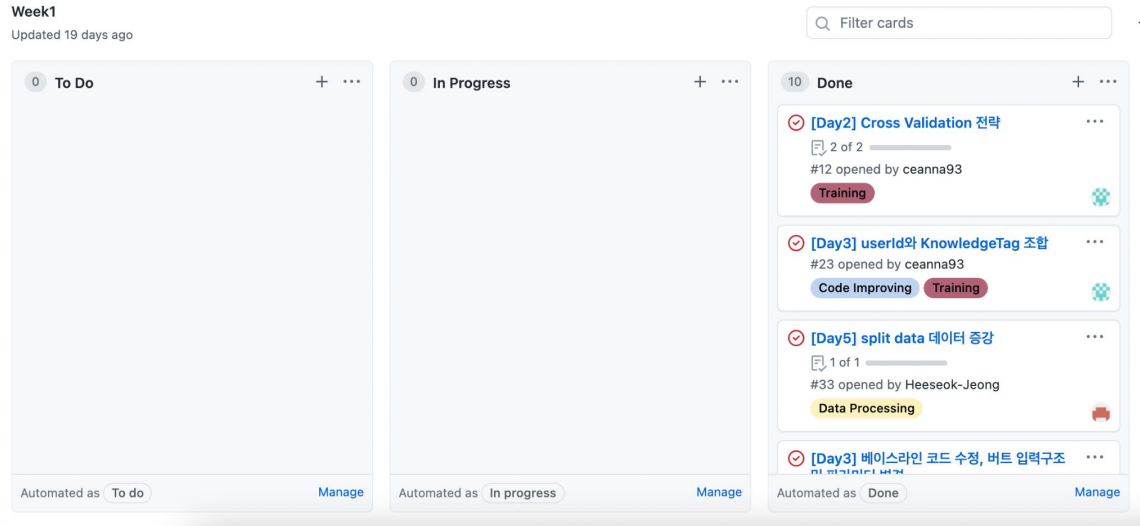
1.1 Set up

Gitflow

기능별로 branch를 만들고 develop branch로 merge

코드 모듈화

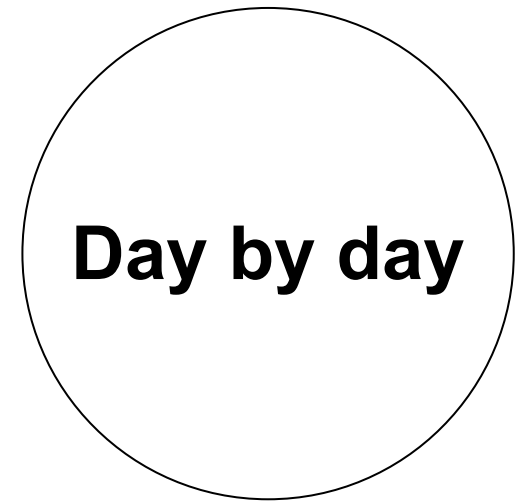
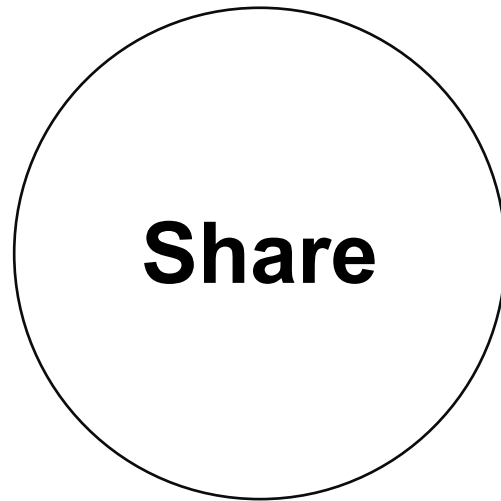
할 일과 PR을 프로젝트 탭으로 관리



1.2 목적이 뭐야?

개인이 아닌 **모두**의 성장!

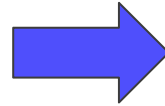
Chapter 2



2.1 Valid Set 에 관하여

Valid Score

0.6816
0.7216
0.7777



LB Score

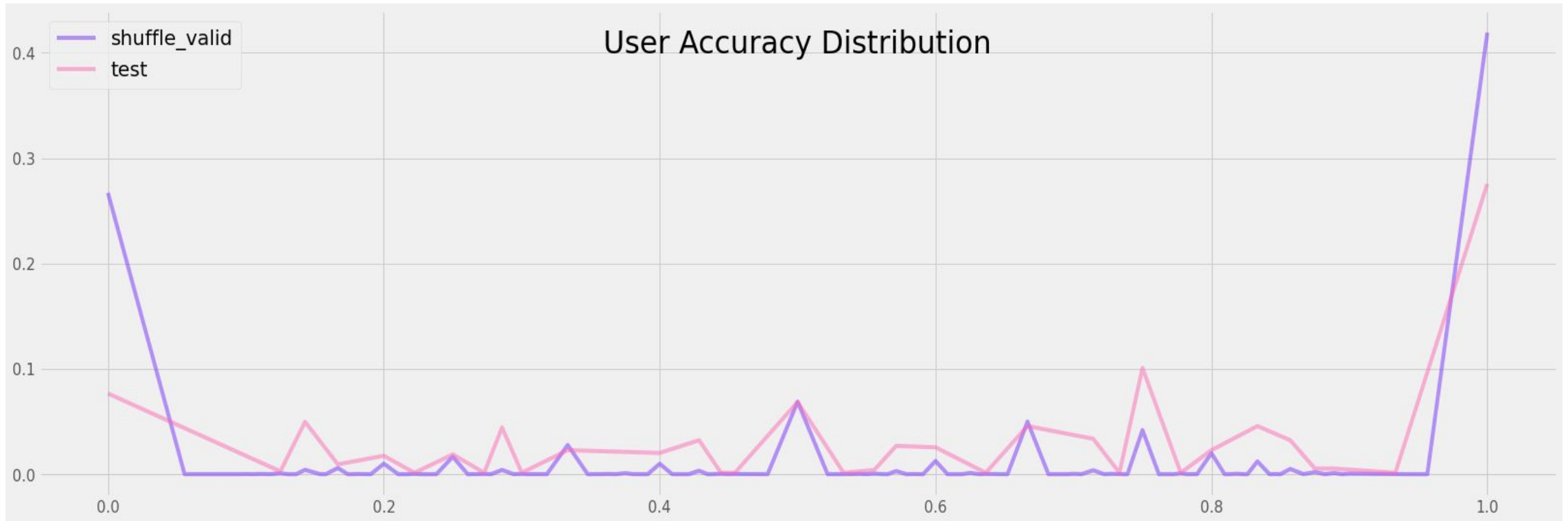
0.7232
0.6892
0.7435

± 0.048

2.1 Valid Set 에 관하여

user 정답률 분포 시각화 [shuffle_valid, test]

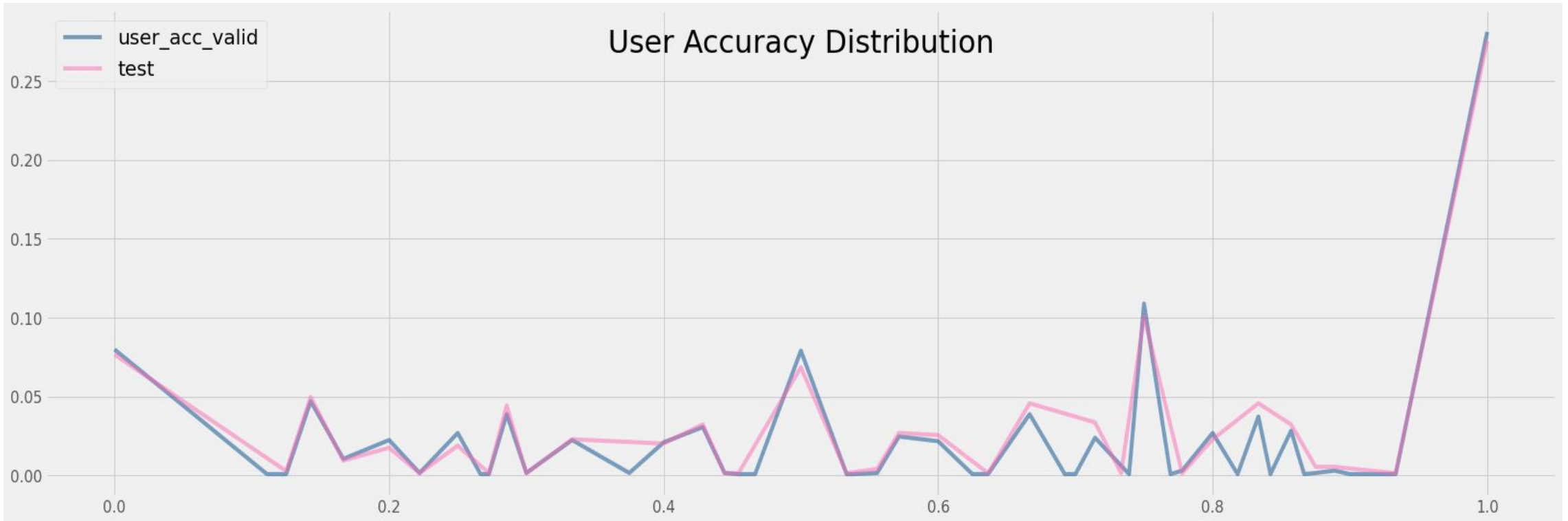
train 데이터에서 임의로 뽑은 **valid** 데이터는 **test** 와 분포가 많이 다르다!



2.1 Valid Set 에 관하여

user 정답률 분포 시각화 [user_acc_valid, test]

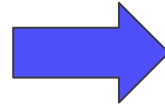
train 데이터에서 유저 정답률 순으로 정렬하고, index 가 0, 5 번째인 user 를 valid 로 넣음 (8:2 비율)
user_acc_valid 와 **test** 의 분포가 비슷해졌다!



2.1 Valid Set 에 관하여

Valid Score

0.7494
0.7474
0.7516



LB Score

0.7419
0.7440
0.7340

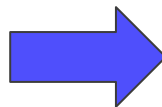
± 0.017

2.2 Transformer 성장기



AUROC 0.724

Transformer-decoder



과거 성적만 보는게 어때?



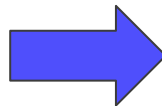
AUROC 0.744

2.2 Transformer 성장기



AUROC 0.744

T-Fixup



빠르고 안정적인 학습의 초기화 기법



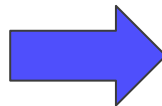
AUROC 0.768

2.2 Transformer 성장기



AUROC 0.744

**Data
Augmentation
(Sliding Window)**



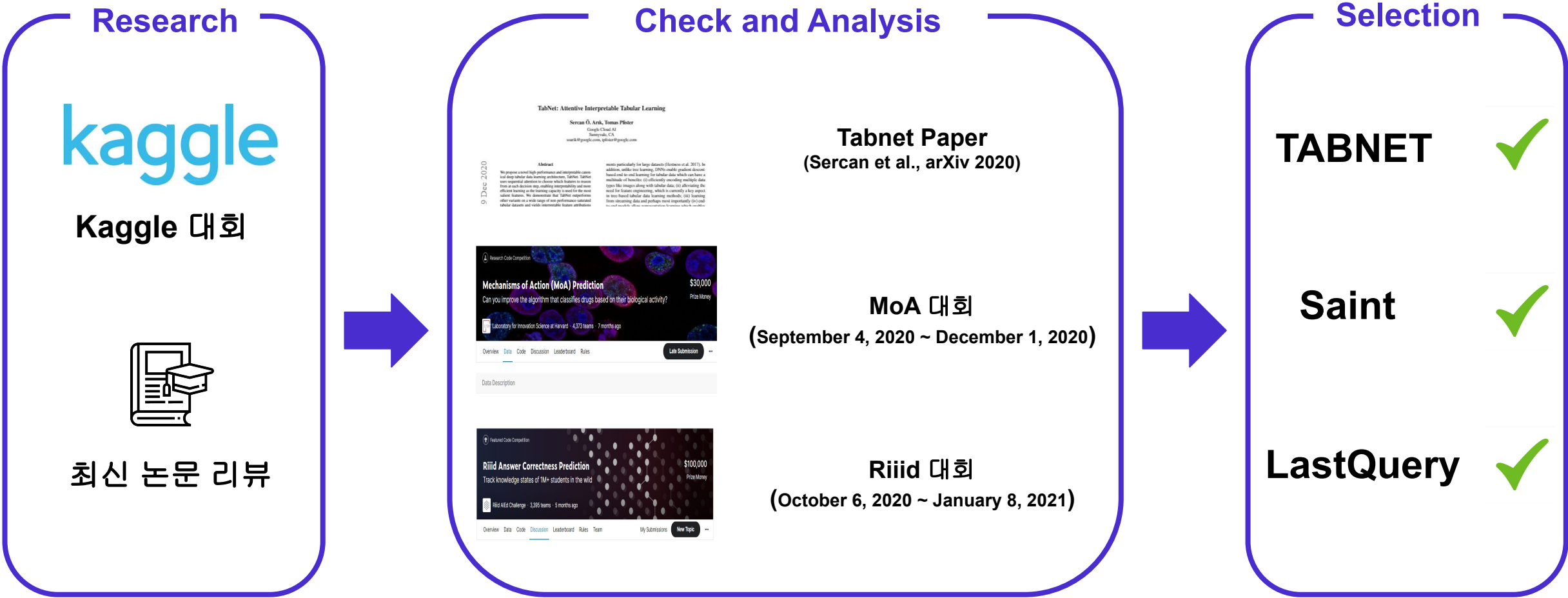
max_seq_len : 40
max_seq_len : 80
max_seq_len : 200 (median)
max_seq_len : 400 (average)



AUROC 0.775

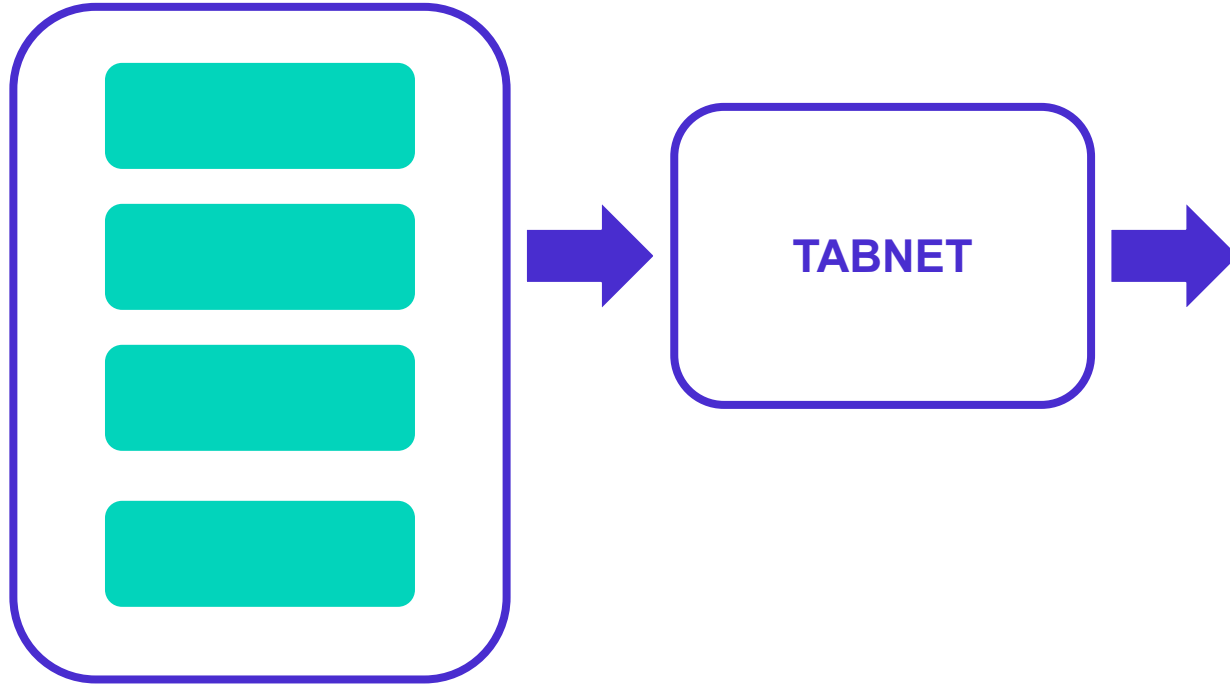
length 40 or 80 is best!

2.3 Kaggle 솔루션과 TABNET

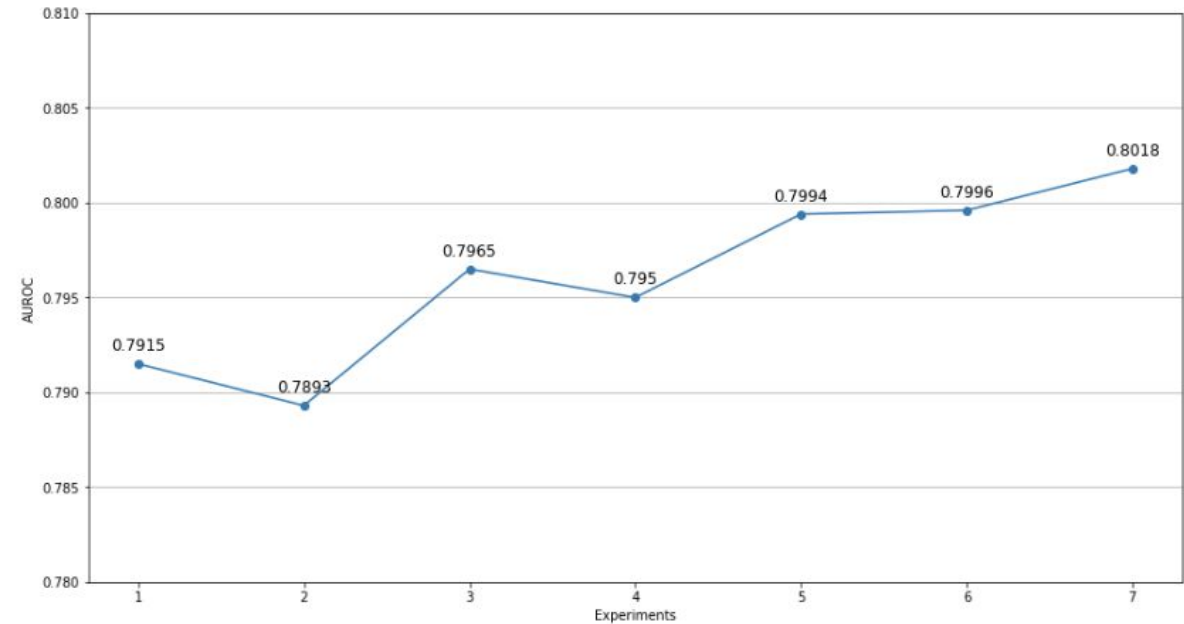


2.3 Kaggle 솔루션과 TABNET

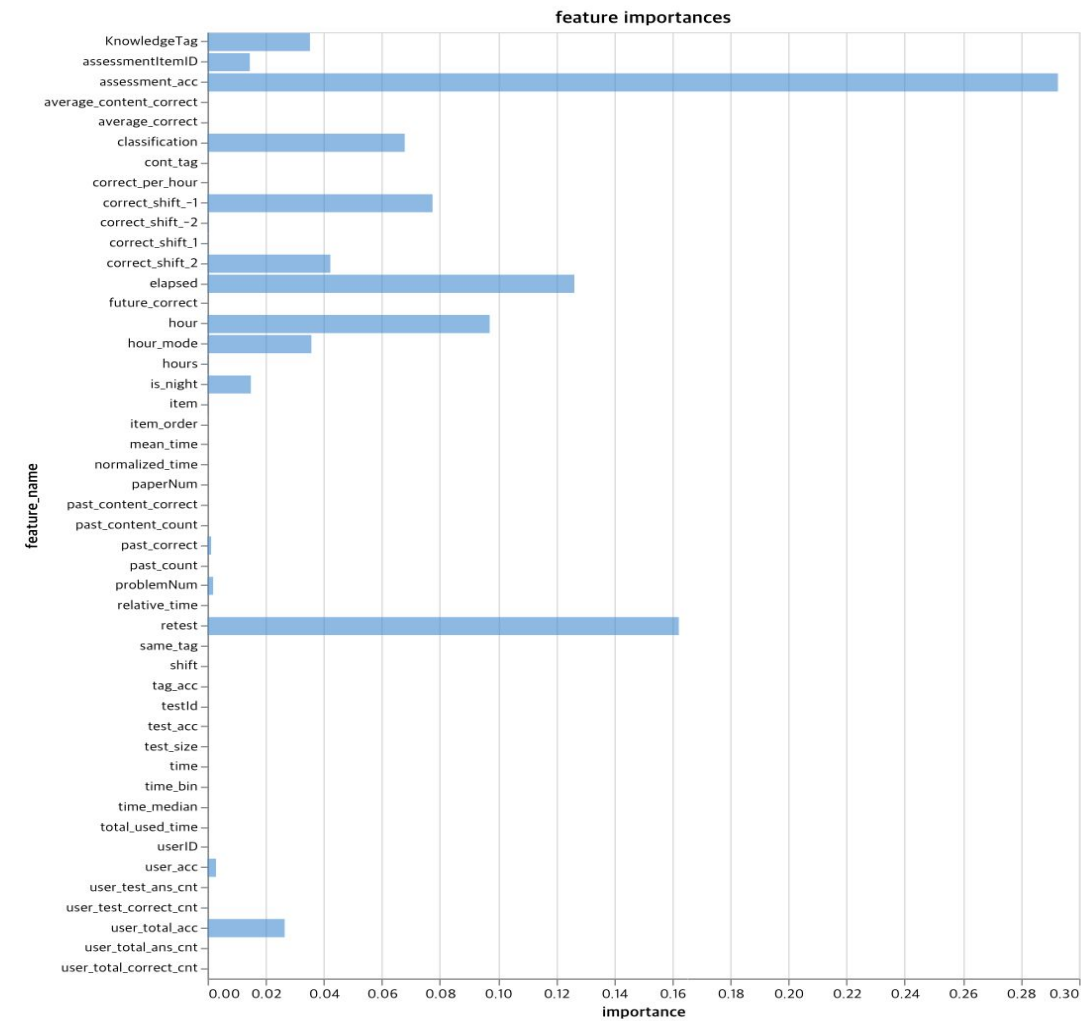
22 of 57 features
from feature
engineering



AUROC 0.8018



2.3 Kaggle 솔루션과 TABNET



Features

- Knowledge Tag
- assessmentItemID
- classification
- correct_shift_-1
- correct_shift_2
- elapsed
- hour
- hour_mode
- is_night
- retest
- user_acc
- user_total_acc

Chapter 3



Share

Grow

**Day by
day**

3.1 발자취

실험 노트

5/25 - 6/11

개수: 39

📅 시작일	📅 종료일	Aa 제목	Description	≡ 카테고리	👤 제안자	👤 실험자	≡ 진행상황
2021년 6월 06일	2021년 6월 06일	📄 feature순서에 따라 모델 학습이 달라진다?!		Data Processing Training	(유) 유진 안	(유) 유진 안	완료
2021년 6월 07일	2021년 6월 07일	📄 Transformer Encoder에 T-Fixup 적용	LayerNorm 없이 Transformer Encoder에 T-Fixup 적용	Code Review Modeling	a anna lee	a anna lee	완료
2021년 6월 08일	2021년 6월 10일	📄 Tabnet 논문 리뷰	Tabnet 논문 리뷰	Paper Review	(유) 유진 안 재우 선	재우 선 (유) 유진 안	완료
2021년 6월 07일	2021년 6월 09일	📄 데이터 증강에 Valid 넣지 않기	수렴이 안한다	Data Processing Modeling	(H) HS J	(H) HS J	완료
2021년 6월 04일	2021년 6월 08일	📄 합리적인 valid data 만들기	과연 믿을 수 있을까?	Data Processing	창우 이	창우 이	완료
2021년 6월 09일	2021년 6월 12일	📄 fixed_train 에 증강 적용하기	수렴을 위해 에포크를 늘리고 SGD 사용	Data Processing Modeling	(H) HS J	(H) HS J	완료
2021년 6월 10일	2021년 6월 10일	📄 Pseudo Labeling		Training	a anna lee	a anna lee	완료
2021년 6월 09일		📄 Transformer 성능 내기	천하제일 Transformer 대회	Modeling Data Processing	(H) HS J	(H) HS J	완료
2021년 6월 10일	2021년 6월 11일	📄 Feature Engineering		Training EDA	a anna lee	a anna lee	완료
2021년 6월 10일	2021년 6월 11일	📄 max_seq_len 와 성능	트랜스포머에서 길이를 늘리자	Data Processing	(H) HS J	(H) HS J	완료
2021년 6월 10일	2021년 6월 11일	📄 Feature importance	Feature들의 중요도 뽑아내기	EDA	재우 선	재우 선	완료
2021년 6월 10일	2021년 6월 11일	📄 데이터 분리 shuffle vs 유저 정답률	original_fixed vs fixed 데이터	Data Processing	(H) HS J	(H) HS J	완료












개수 39

3.1 발자취

Github Commits

5/19 - 6/18

개수: 98










	Heeseok-Jeong Merge pull request #73 from bcaitech1/health/final_code_improvi...	78f44e9 8 hours ago	 98 commits
	dkt Remove comments	8 hours ago	
	make_custom_data put making files into make_custom_data folder and add ensemble	10 hours ago	
	.gitignore Merge remote-tracking branch 'origin/yujin/baseline' into JAEWOOSU...	24 days ago	
	README.md Initial settings	25 days ago	
	args.py Code Improving	8 hours ago	
	ensemble.py put making files into make_custom_data folder and add ensemble	10 hours ago	
	inference.py put making files into make_custom_data folder and add ensemble	10 hours ago	
	requirements.txt put making files into make_custom_data folder and add ensemble	10 hours ago	
	train.py Code Improving	8 hours ago	

3.1 발자취

Github Issue

5/26 - 6/16

개수: 28

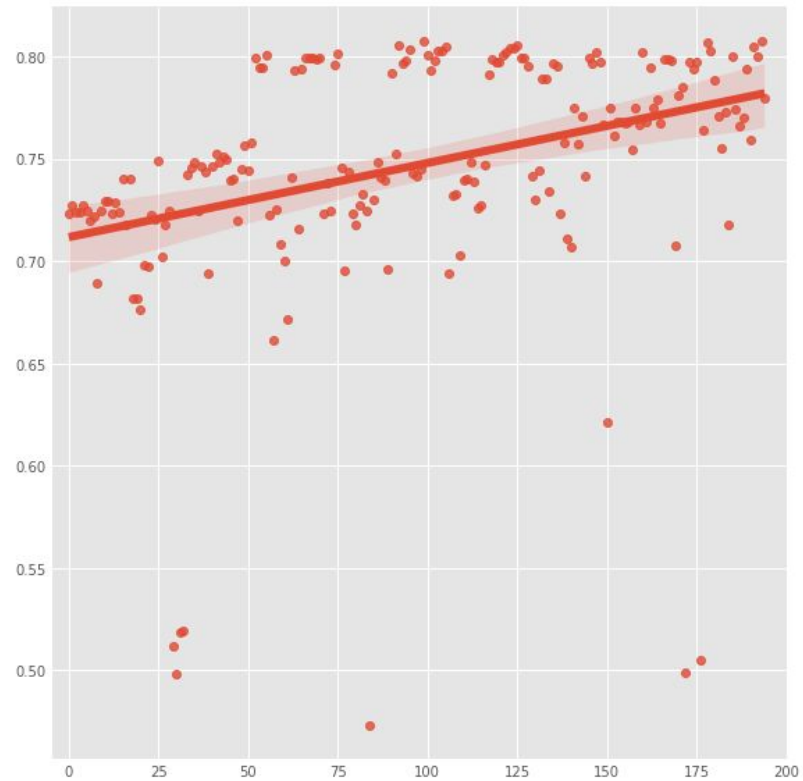
<input type="checkbox"/>	<input checked="" type="radio"/> 0 Open	<input checked="" type="radio"/> 28 Closed	Author ▾	Label ▾	Projects ▾	Milestones ▾	Assignee ▾	Sort ▾
<input type="checkbox"/>	<input checked="" type="radio"/>	[Day 13] Feature Engineering	Data Processing	EDA				
		#64 by ceanna93 was closed 3 days ago	📄 15 of 15					
<input type="checkbox"/>	<input checked="" type="radio"/>	[Day12] 데이터 증강에 valid 넣지 않기	Data Processing	Modeling				
		#62 by Heeseek-Jeong was closed 5 days ago	📄 6 of 6					
<input type="checkbox"/>	<input checked="" type="radio"/>	[Day 12] Transformer Encoder에 T-Fixup 적용	Modeling	Training				 1
		#60 by ceanna93 was closed 9 days ago	📄 2 of 3					
<input type="checkbox"/>	<input checked="" type="radio"/>	LastQuery 후속 실험	Modeling					
		#59 by dkswndms4782 was closed 2 days ago	📄 0 of 4					
<input type="checkbox"/>	<input checked="" type="radio"/>	[DAY 10] Saint 분석	Modeling	Paper Review	Training			
		#57 by ceanna93 was closed 9 days ago	📄 3 of 3					
<input type="checkbox"/>	<input checked="" type="radio"/>	[DAY 10] tabnet 사용하기	Modeling					
		#56 by JAEWOOSUN was closed 3 days ago	📄 6 of 6					
<input type="checkbox"/>	<input checked="" type="radio"/>	[Day10] LastQuery	Modeling					
		#55 by dkswndms4782 was closed 10 days ago	📄 5 of 5					
<input type="checkbox"/>	<input checked="" type="radio"/>	[Day7] LGBM	Training					
		#43 by changwoomon was closed 9 days ago	📄 6 of 6					
<input type="checkbox"/>	<input checked="" type="radio"/>	[Day7] Sliding Window 적용	Data Processing					
		#42 by ceanna93 was closed 13 days ago	📄 3 of 3					

3.1 발자취

제출 기록

5/24 - 6/15

개수: 195



Aa 이름	≡ ACCURA...	≡ AUROC	≡ 변경 사항	👤 제출한 사람	≡ Created At	≡ 실험 노트	+
Baseline_8	0.6613	0.7239	시간 features 추가	창우 이	21.05.26.(Wed) 07:45	시차, 시간대 변수 추가	
Baseline	0.6667	0.7233	베이스라인 돌림	창우 이	21.05.24.(Mon) 11:12		
Baseline_7	0.6667	0.7233	시간 features 추가 (오류)	창우 이	21.05.26.(Wed) 02:52	시차, 시간대 변수 추가	
LSTM_augmentation_2	0.6613	0.7232		HS J	21.05.31.(Mon) 17:35		
bert_6	0.6344	0.7232	continuous_bert	재우 선	21.06.02.(Wed) 00:48	Continuous embedding	
LSTM_pseudo	0.6532	0.7232	oof stacking, pseudo labeling 코드 실행 결과	anna lee	21.06.10.(Thu) 01:31		
Baseline_11	0.6613	0.7227	.ipynb → .py 변환	창우 이	21.05.27.(Thu) 04:48	시차, 시간대 변수 추가	
LSTM_augmentation	0.6452	0.7226	유저 데이터 쪼개서 모두 사용	HS J	21.05.30.(Sun) 10:08	데이터 증강	
Istmattn	0.6667	0.7220	model : LSTM + Attention	유진 안	21.05.25.(Tue) 17:10		
Baseline_12	0.6559	0.7203	.ipynb → .py 변환 Baseline_11 동일 inference 잘못함	창우 이	21.05.27.(Thu) 05:09	시차, 시간대 변수 추가	
LGBM	0.6371	0.7201	model : LGBM	유진 안	21.05.25.(Tue) 17:10		
Baseline_28	0.6720	0.7199	epoch - 19 test만 학습	창우 이	21.05.29.(Sat) 19:51		
Istmattn_3	0.6613	0.7179	3 features 추가	재우 선	21.05.26.(Wed) 15:33	대분류, 시험지 번호, 문제번호	
개수 195							

End of Document

Thank You.