

0. Team 소개

Naver Boostcamp AI Tech 5기 RecSys 07 조 낮가리지 않는 법 추천해조

팀 구성원 : 강은비_T5006, 김철현_T5066, 이한정_T5166, 최민수_T5216

1. Book Rating Prediction

1- 1. 프로젝트 개요

프로젝트 주제

사용자의 책 평점 데이터를 바탕으로 사용자가 어떤 책을 더 선호할지 예측하는 태스크

프로젝트 기간

2022.04.10 ~ 2022.04.20

프로젝트 실행환경

개발 환경

- V100 GPU 서버

협업 및 개발 Tools

- Notion
- Zoom
- Slack
- Python
- JupyterLab
- Optuna

1- 2. 프로젝트 팀 구성 및 역할

- 최민수(팀장)

EDA, 데이터 전처리, CatBoost 모델 설계 및 최적화, 팀 목표 설정 및 스케줄 관리

- 강은비

EDA, 데이터 전처리, LGBM 최적화

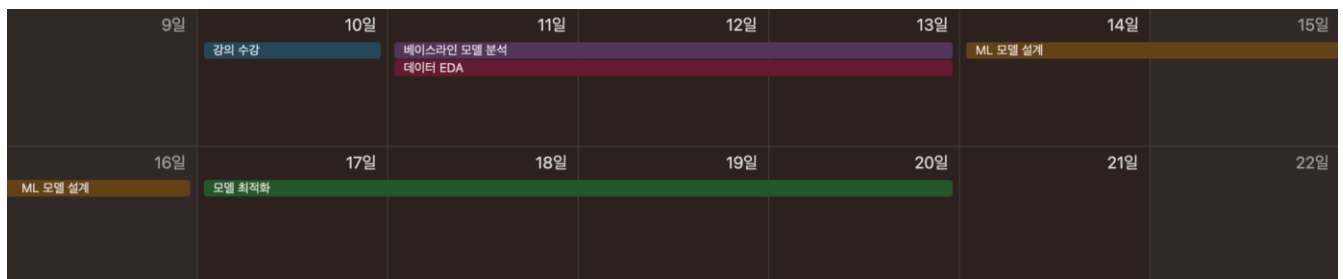
- 김철현

EDA, 데이터 전처리, CatBoost 모델 최적화

- 이한정

EDA, 데이터 전처리

1- 3. 프로젝트 수행 절차



1. 프로젝트 진행 전 사전 강의 수강 (4/10)
2. 데이터 EDA 와 병렬적으로 베이스라인 코드 분석 및 최적화 (4/11 ~ 4/13)
3. 베이스라인 코드가 아닌 새로운 ML 모델(CatBoost, LGBM) 설계 및 분석 (4/14 ~ 4/16)
4. 앞서 설계한 ML 모델 최적화 (4/17 ~ 4/20)

1- 4. 프로젝트 수행 결과

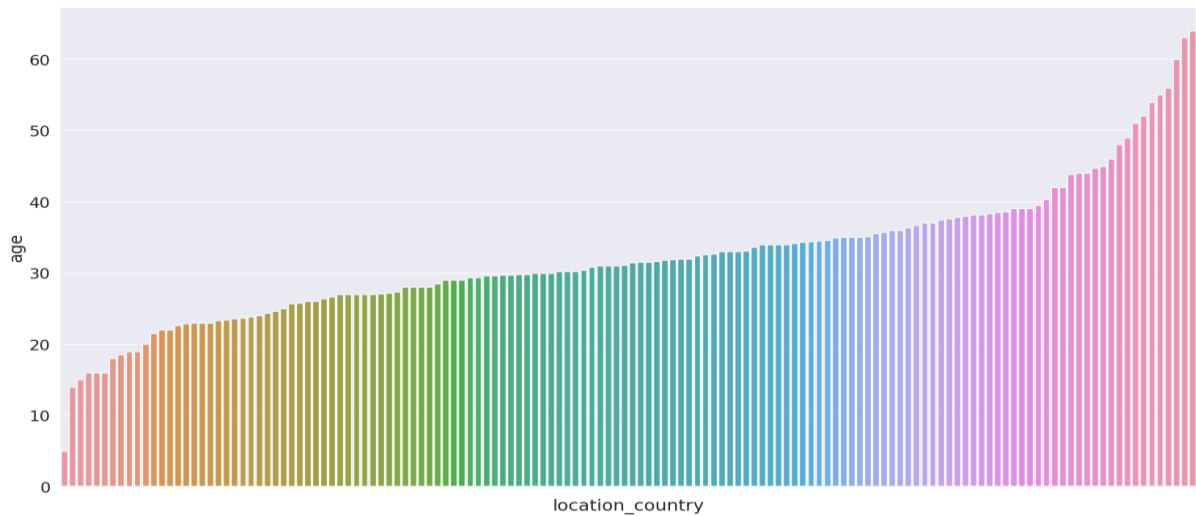
EDA

0. Users - Books Matrix

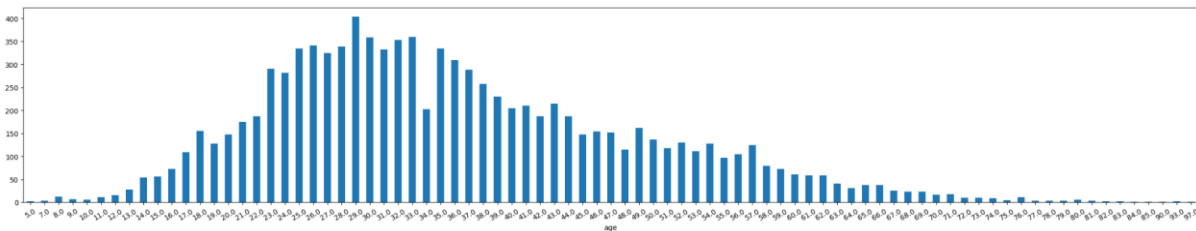
- Sparse Matrix
 - user 가 책에 매긴 평점 matrix 는 sparse matrix 이다.
 - $306795(\text{평점 개수}) / (68069(\text{유저 수}) * 149570(\text{책 수})) = 0.0000301$

1. Users

- user_id, location, age 로 이루어져 있다.
- location, age 에 결측값이 존재한다.
- Age
 - 약 40%의 값이 결측값이다.

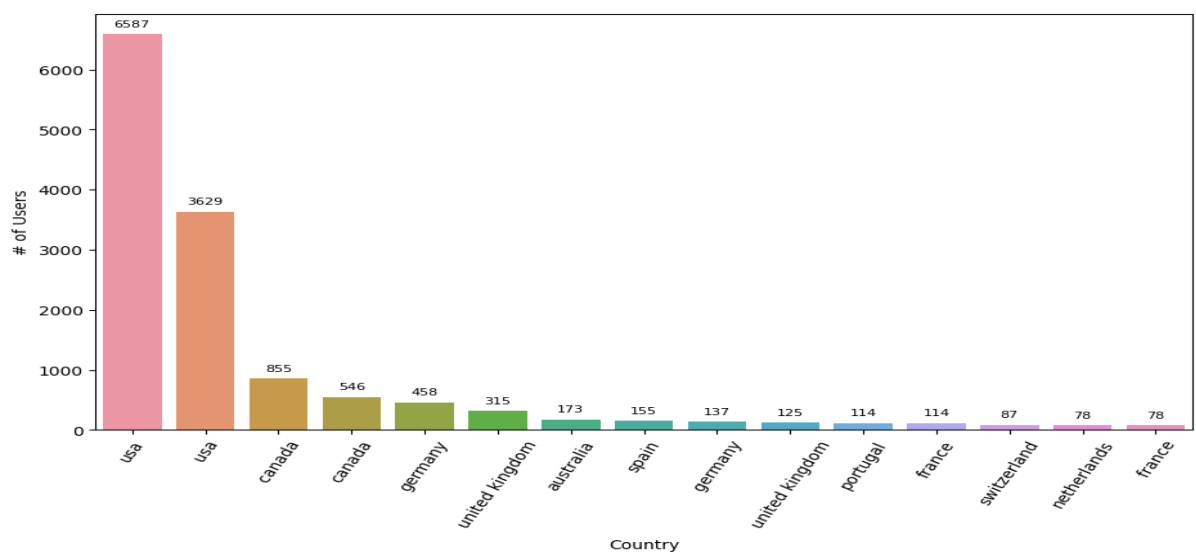


<국가별 나이 평균에 유의미한 차이가 존재>



<20 대 초반 ~ 30 대 중반 까지의 사용자가 대다수>

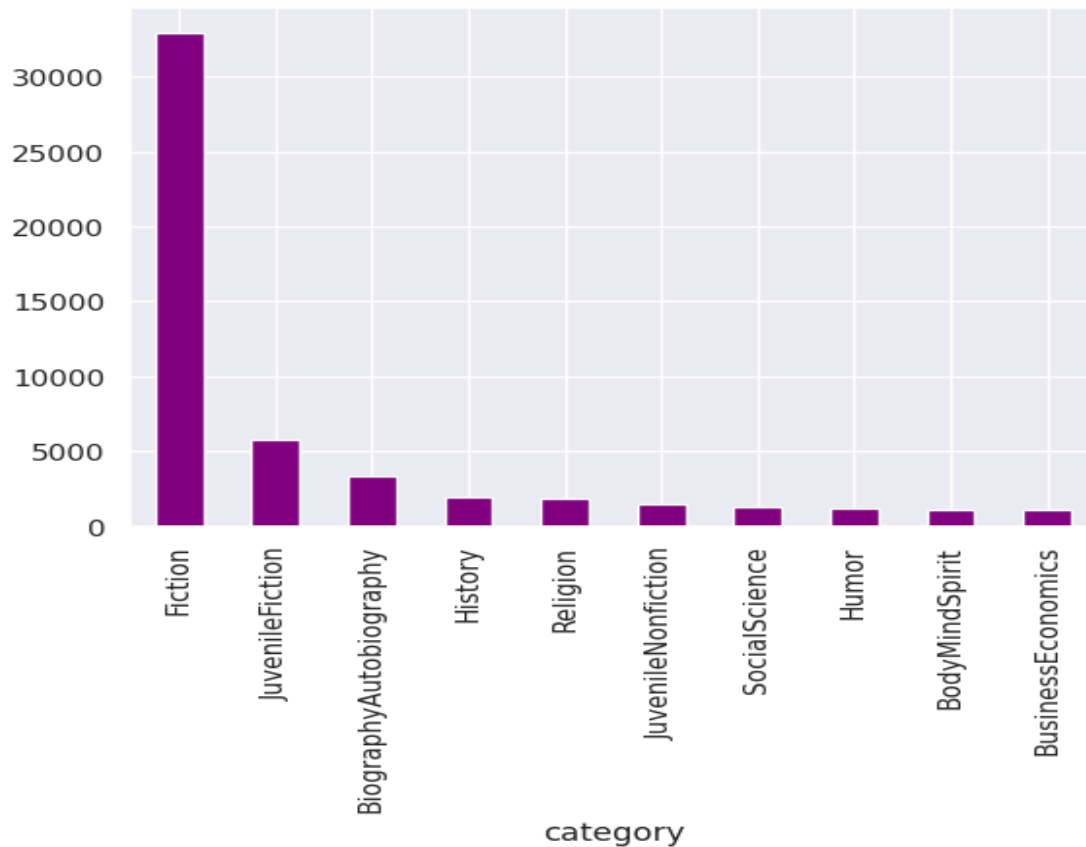
- location
 - 지역, 주, 국가의 정보가 들어있다.



<국가의 경우 대다수가 미국에 거주>

2. Books

- isbn, book_title, book_author, year_of_publication, publisher, img_url, language, category, img_path, summary 의 column 으로 이루어져 있다.
- language, category, summary 에 결측값이 존재한다.
- isbn
 - isbn 은 모두 10 자리이다.
 - isbn 은 4 개의 그룹으로 나누어 볼 수 있다. (출판 국가, 출판사 번호, 항목 번호, 확인 숫자)
- language
 - 'en', nan, 'es', 'fr', 'de', 'da', 'it', 'ru', 'nl', 'th'의 값으로 이루어져 있으며 대부분이 'en'이다.
- category
 - 3000 개가 넘는 값으로 이루어져 있지만, 그 중 10 번 이상 등장하는 값은 48 개 뿐이다.



<Fiction 이 차지하는 부분이 대다수>

데이터 전처리

Catboost v1

1. Users

- location
 - location 을 country, state, city 로 나누어서 구성
- age
 - country 별 평균 age 를 구하여 이를 통해 결측치를 처리함
- N_ratings, avg_rating
 - 각각의 유저가 평가한 횟수인 N_ratings 라는 새로운 column 추가
 - 각각의 유저가 매긴 평균 평점인 avg_rating 이라는 새로운 column 추가
 - 평가를 한번도 하지 않아 avg_rating 이 존재하지 않는다면 중간값인 5 로 대체
- 완성된 Users 테이블

	user_id	location	age	N_ratings	avg_rating	city	state	country
0	8	timmins, ontario, canada	35.582181	7	4.428571	timmins	ontario	canada
1	11400	ottawa, ontario, canada	49.000000	12	6.750000	ottawa	ontario	canada
2	11676	n/a, n/a, n/a	36.399133	5520	6.779891	cassina rizzardi	fife	sweden
3	67544	toronto, ontario, canada	30.000000	7	7.285714	toronto	ontario	canada
4	85526	victoria, british columbia, canada	36.000000	120	7.666667	victoria	british columbia	canada
...
68087	278376	danville, pennsylvania, usa	54.000000	1	7.000000	danville	pennsylvania	usa
68088	278621	victoria, delaware, canada	74.000000	1	8.000000	victoria	delaware	canada
68089	278636	irvington, alabama, usa	37.859796	1	2.000000	irvington	alabama	usa
68090	278659	vancouver, washington, usa	33.000000	1	10.000000	vancouver	washington	usa
68091	278713	albuquerque, new mexico, usa	63.000000	1	7.000000	albuquerque	new mexico	usa

2. Books

- Category
 - 결측치를 처리하기 위해 50 개 이상을 갖는 카테고리 선정
 - 50 개 이상을 갖는 카테고리는 대표성을 지닌다고 생각하여 랜덤 샘플링을 통해 결측치 처리

- **Book_author**

- 결측치를 처리하기 위해 5 개 이상을 갖는 작가를 선정

- 5 개 이상을 갖는 작가는 유의미한 정보라고 판단하여 랜덤 샘플링을 통해 결측치 처리

- **Summary, Language**

- 모델을 학습시켜 **Feature Importance** 를 확인해 본 결과 중요하지 않다고 판단하여 삭제

- **완성된 Books Table**

	isbn	book_title	book_author	year_of_publication	publisher	category
0	0002005018	Clara Callan	Richard Bruce Wright	2001.0	HarperFlamingo Canada	['actresses']
1	0060973129	Decision in Normandy	Carlo D'Este	1991.0	HarperPerennial	['1940-1949']
2	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999.0	Farrar Straus Giroux	['medical']
3	0399135782	The Kitchen God's Wife	Amy Tan	1991.0	Putnam Pub Group	['fiction']
4	0425176428	What If?: The World's Foremost Military Histor...	Robert Cowley	2000.0	Berkley Publishing Group	['history']
...
149565	067161746X	The Bachelor Home Companion: A Practical Guide...	P.J. O'Rourke	1987.0	Pocket Books	['humor']
149566	0767907566	All Elevations Unknown: An Adventure in the He...	Sam Lightner	2001.0	Broadway Books	['nature']
149567	0884159221	Why stop?: A guide to Texas historical roadsid...	Claude Dooley	1985.0	Lone Star Books	['pets']
149568	0912333022	The Are You Being Served? Stories: 'Camping In...	Jeremy Lloyd	1997.0	Kqed Books	['fiction']
149569	1569661057	Dallas Street Map Guide and Directory, 2000 Ed...	Mapsc	1999.0	American Map Corporation	['pets']

Catboost v2

1. Users

- **age**

- Users 데이터에서는 age 부분에 약 40% 정도 결측치가 존재

- 처음에는 평균으로 결측치를 채운 다음에 10 단위로 구간화 한 뒤에 성능을 체크했음

- 두번째 방법으로는 랜덤 샘플링을 사용

- 평균으로 결측치를 채우고 10 단위로 구간화 했을 경우 20 ~30 대 구간에 너무 많은 데이터가 집중되어 있음을 발견 → 나이 분포를 최대한 그대로 유지시켜주고 구간화도 시행하지 않음

- **완성된 Users Table**

	user_id	age	location_city	location_state	location_country
0	8	34.0	timmins	ontario	canada
1	11400	49.0	ottawa	ontario	canada
2	11676	33.0	christchurch	canterbury	newzealand
3	67544	30.0	toronto	ontario	canada
4	85526	36.0	victoria	britishcolumbia	canada
...
68087	278376	54.0	danville	pennsylvania	usa
68088	278621	74.0	victoria	delaware	canada
68089	278636	30.0	irvington	alabama	usa
68090	278659	33.0	vancouver	washington	usa
68091	278713	63.0	albuquerque	newmexico	usa

68092 rows x 5 columns

2. Books

- language

- Books 데이터에서는 language 부분에 약 45% 정도 결측치가 존재
- ISBN의 첫번째 숫자는 책이 출판된 국가나 지역을 뜻하기 때문에 첫번째 숫자를 이용해 language 결측치를 보완
- 이후 남은 language의 경우 random sampling으로 보완

- publisher

- 출판한 책이 다수인 출판사를 기준으로 출판사 이름을 통합(기준: 10개 이상 출판)
- 이후 통합되지 못한 출판사의 경우 others로 처리

- 완성된 Books Table

	isbn	book_title	book_author	year_of_publication	publisher	language	category_high
0	0002005018	clara callan	richard bruce wright	2001.0	flamingo	en	arts
1	0060973129	decision in normandy	carlo d este	1991.0	perennial	en	biographies
2	0374157065	flu the story of the great influenza pandemic ...	gina bari kolata	1999.0	farrar straus giroux	en	medical
3	0399135782	the kitchen god s wife	amy tan	1991.0	putnam	en	literature
4	0425176428	what if the world s foremost military historia...	robert cowley	2000.0	berkley publishing group	en	history
...
149565	067161746X	the bachelor home companion a practical guide ...	p j o rourke	1987.0	pocket	en	entertainment
149566	0767907566	all elevations unknown an adventure in the hea...	sam lightner	2001.0	broadway	en	science
149567	0884159221	why stop a guide to texas historical roadside ...	claudie dooley	1985.0	others	en	science
149568	0912333022	the are you being served stories camping in an...	jeremy lloyd	1997.0	others	en	literature
149569	1569661057	dallas street map guide and directory 2000 edi...	mapsco	1999.0	others	en	literature

149570 rows × 7 columns

LightGBM

1.Users

- age

- 결측의 경우 국가별 평균 나이에 차이가 상당히 존재하므로 국가별 평균나이로 처리함

- user 별 평균 평점

- rating의 전체 평균은 7 점대로 높은 편이나 user 별 평균 평점 분포를 확인한 결과 평가 성향이 다양한 것으로 확인됨
- 기록이 없는 user의 결측치를 처리하지 않고 학습했을 때 심한 과적합 문제 발생
- => user 정보를 활용하여 lgb으로 예측한 평균 평점으로 결측 처리하여 성능 개선

- location

- city, state, country 중 한 카테고리라도 존재할 경우 다른 카테고리를 통해 결측 처리
- state 를 사용하지 않았을 때 성능이 향상되어 city, country 만 사용함

- 완성된 Users Table

	user_id	age	location_city	location_country	avg
0	8	35.733836	timmins	canada	4.428571
1	11400	49.000000	ottawa	canada	6.750000
2	11676	35.733836	toronto	canada	6.779891
3	67544	30.000000	toronto	canada	7.285714
4	85526	36.000000	vancouver	canada	7.666667
...
68087	278376	54.000000	danville	usa	7.000000
68088	278621	74.000000	vancouver	canada	8.000000
68089	278636	37.845259	irvington	usa	2.000000
68090	278659	33.000000	vancouver	canada	10.000000
68091	278713	63.000000	albuquerque	usa	7.000000

2.Books

- category

- 카테고리의 경우 매우 많은 범주를 가져 5 개이상 나타난 카테고리는 통합
- 결측의 경우 작가 별 최다 카테고리를 이용하여 처리

- language

- 기존 language 분포에 따라 랜덤 샘플링하여 처리함.

- publisher

- publisher 역시 매우 많은 범주가 존재하여 isbn 을 이용해 대표 publisher 로 통합함

- 완성된 Books Table

	isbn	book_author	year_of_publication	publisher	language	category	new_cate
0	0002005018	Richard Bruce Wright	2001.0	harpercollins	en	act	arts
1	0060973129	Carlo D'Este	1991.0	harpercollins	en	19401949	others
2	0374157065	Gina Bari Kolata	1999.0	farrar straus giroux	en	medical	medical
3	0399135782	Amy Tan	1991.0	putnam pub group	en	fiction	literature
4	0425176428	Robert Cowley	2000.0	berkley publishing group	en	history	history
...
149565	067161746X	P.J. O'Rourke	1987.0	pocket	en	humor	entertainment
149566	0767907566	Sam Lightner	2001.0	broadway books	en	nature	nature
149567	0884159221	Claude Dooley	1985.0	bridge publications	de	others	others
149568	0912333022	Jeremy Lloyd	1997.0	pub group west	en	fiction	literature
149569	1569661057	Mapsco	1999.0	soho press	en	others	others

모델 분석

LightGBM

- 범주형 변수가 많은 데이터에서 특히 높은 성능을 보임
- gradient가 가장 큰 노드부터 분할하는 **leaf-wise** 방식을 사용해 빠른 속도로 학습 가능
- Category 형 피처의 자동 변환 및 최적 분할 가능

CatBoost

- 범주형 변수에 강력한 성능을 보이는 모델
- Ordered boosting
- Random permutation
- Categorical feature combination

모델 고도화

Stratified K-Fold CV

5-Fold 또는 10-Fold의 평균 RMSE를 계산해 줌으로써 조금 더 뛰어난 일반화 성능 기대

Optuna - LGBM (주요 Hyperparameter)

- N_estimators
생성할 Tree 개수로 10~500까지의 범위로 설정
- Num_leaves
Tree의 최대 leaf 개수로 2~100까지의 범위로 설정
- Max_depth
트리의 깊이로 3~10까지의 범위로 설정
- Subsample
학습할 데이터 샘플링 비율로 0.5~1까지의 범위로 설정

- Colbysample_bytree

Tree 학습 시 선택할 feature 의 비율로 0.5~1 까지의 범위로 설정

- Reg_alpha

L1 정규화 강도로 0~1 까지의 범위로 설정

Optuna - CatBoost (주요 Hyperparameter)

- cat_features

카테고리형 feature 로 country, book_author, publisher ... 등을 활용

- Bagging_temperature

Bagging 효과의 정도로 약 0.01 ~ 100 까지의 범위로 설정

- N_estimators

앙상블할때 활용할 트리의 개수로 약 1000 ~ 10000 까지의 범위로 설정

- Max_depth

트리의 최대 깊이로 약 4 ~ 12 까지의 범위로 설정

- Random_strength

무작위성의 강도로 약 0 ~ 100 까지의 범위로 설정

- Min_child_samples

Leaf 노드에 필요한 최소 샘플 수로 약 40 ~ 100 까지의 범위로 설정

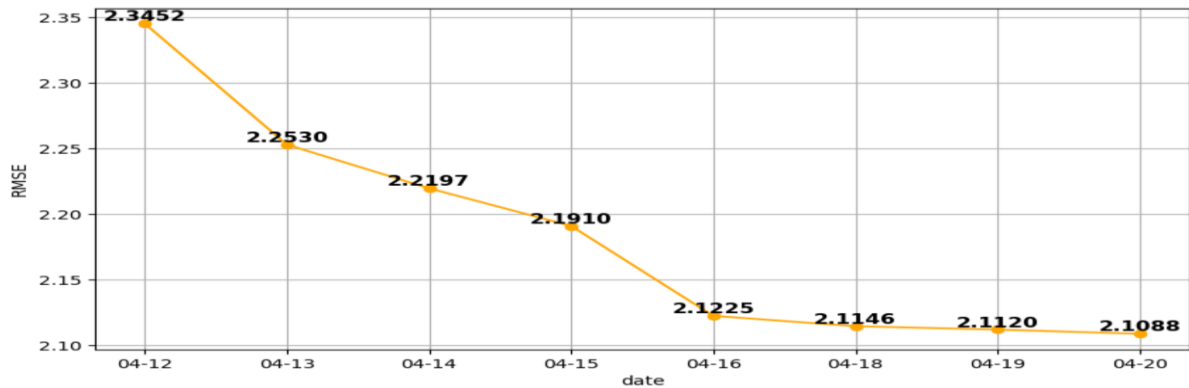
Ensemble

이처럼 모델 고도화 과정을 거친 후 제출해 본 결과 단일 모델의 성능은 그렇게 뛰어나지 않았음



따라서 앙상블을 통해 다양한 모델의 예측력을 결합하고 과적합을 방지하며 더 뛰어난 일반화 성능 기대

다양한 앙상블 조합을 탐색해 보았고 그 결과 최종 output 을 약 8 : 1 : 1 로 결합하여 최적의 성능 도출

최종 결과



Public 1 위

1	RecSys_07조	 	2.1088	63	3d
---	------------	---	--------	----	----

Private 1 위

1	RecSys_07조	 	2.1061	63	3d
---	------------	---	--------	----	----

1- 5. 자체 평가 의견

잘했던 점

- 성능이 좋은 모델에 집중하여 최고의 성능을 얻음
- 2 주 동안 꾸준히 제출하여 성능 테스트 함
- 다양한 방식의 **Feature Engineering** 진행
- 베이스라인 코드에서 벗어나 새로운 ML 모델 설계

아쉬웠던 점

- 딥러닝 모델로 좋은 성능을 내지 못함
- GPU 서버를 100% 활용하지 못함

배운점

- **Data Leakage** 의 위험성
- 다양한 시도의 중요성
- 협업 능력이 아직 부족함

2. 개인 회고

강은비

프로젝트 초반에는 부스트캠프 참가 목적이었던 “이해에 기반한 딥러닝 활용 능력 향상”에 맞게 DL 모델들을 더 공부하고 여러 시도들을 많이 했었다. 특히 그나마 더 잘 이해했다고 생각했던 FM 계열 모델들에 꽂혔었다. 하지만 생각과는 다르게 **valid loss**가 잘 개선되지 않았고, 예러도 많이 났었는데, 이를 해결하기 위해 또 많은 시간을 소요했었다.

들인 시간에 비해 결과가 좋지 않아서 초조한 마음이 들었다. 특히 팀원분들이 높은 성능을 냈던 것을 보고, “나도 성적에 도움이 되야할텐데...”라는 걱정이 많이 들었다.

이런저런 걱정으로 안되는 것은 포기하고 성능이 좋다는 부스팅계열 모델들을 시도했고, 놀랍게도 성능이 정말 많이 개선되었다. 늦게 시도해 본게 아쉬웠고, 좀 더 시간이 주어졌다면 많은 실험을 통해 더 일반화된 모델로 개선할 수 있었을 것 같았다. 프로젝트가 끝날때쯤 되니까 아이디어들이 많이 떠올랐는데, 다 시도해보지 못해서 아쉬움이 많이 남았다.

1등을 하고 나서 팀원분들과 우리가 한 방식이 과연 적합한가에 대해 이런저런 이야기를 나누었으나, 마스터 클래스를 통해 우리 조가 했던 방식도 충분히 의미가 있음을 느꼈다. 특히 급하게 발표를 준비하느라 힘들었는데, 발표 덕분에 현직자 멘토님의 피드백을 들을 수 있어서 정말 의미 있는 시간이 되었다.

팀으로 하지 않았다면 아마 나는 계속 CNN_FM에 매달렸을 것 같다... 팀원들 덕분에 여기서 벗어나서 다른 시도를 해보고 성능 개선에 좀 더 집중할 수 있었고, 높은 성적을 내서 좋은 피드백을 받을 수 있는 기회도 얻게 되었다.

김철현

2 주동안 재밌게 했다!! 좋은 GPU 도 써보고 첫 대회인데 1 등도 해보고 좋은 경험이었다.

대회 초반에는 **argparse** 모듈을 처음 써보았고 모듈화가 되어있는 코드를 돌려본 것도 오랜만이었어서 많이 해맸었다. 첫 대회인만큼 모델에 익숙해지고 **EDA** 를 제대로 해보자! 라는 마음가짐으로 시작했었는데 리더보드가 열리고 팀원들이 점수를 올리는 것을 보고 있자니 나도 점수에 욕심이 생겼다!

급하게 구글링을 해서 성능을 끌어올리려 했고 **Catboost** 모델과 **Optuna** 최적화를 사용했을 때 결과가 엄청 좋게 나와서 신기하기도 했지만 한편으로는 내가 직접 짠 코드가 아니어서 이게 의미가 있나 싶기도 했다. 그래도 덕분에 **Optuna** 에 대한 사용법과 여러가지 전처리 기법들을 익혀서 도움은 많이 된 것 같다. 5 시간동안 모델이 돌아가는 걸 보았을 때는 정말 신기했다!

팀원들과는 대회 초반에 어떻게 소통을 해야할지 너무 어려웠다. **github** 이나 **wandb** 등을 같이 쓰면서 기록해보고 싶었는데 쉽지 않았다! **EDA** 나 전처리와 관련된 부분도 이야기를 많이 나누고 시작했으면 좋았는데 내가 부족한 부분이 많아서 얘기를 꺼내기 어려웠던 부분도 있었었다.

대회 초반에 시간을 잘 못 쓴것도 너무 후회된다. 대회 후반으로 갈 수록 시간이 너무 부족하다는 것을 느꼈다.

마지막으로 팀원들에게 너무 고맙고 미안하다는 말을 하고싶다. 프로젝트에 적극 참여하도록 자극시켜주었고 내가 많이 부족했지만 한번도 부담을 주지 않았었다. 마지막 날에 자료를 늦게 준비한 것은 정말 너무 미안하다. 그날 정말 많이 배운 것 같다. 너무 개인적으로 행동했었고 팀 프로젝트에 대해 다시한번 생각해보는 시간을 가졌다.

많이 아쉽고 많이 배운 프로젝트였다! 다들 수고많으셨습니다!!

이한정

지금까지 이와 비슷한 프로젝트를 진행했을 때, 모델링보다는 데이터 전처리와 후처리에서 많은 효과를 봤었습니다. 따라서 이번 프로젝트에서도 데이터 전처리 및 결측값 채우기에 집중하였습니다. 따라서 다양한 모델을 사용해 보지 못한 점이 아쉬움으로 남습니다.

저희 조가 이번 프로젝트에서 가장 많은 성능 개선을 볼 수 있었던 부분은 다양한 모델을 최적화해보고 그 모델들을 앙상블 했을 때였습니다. 이에 반해 저는 모델 최적화보다는 결측값 채우기에 집중하느라 숲을 보지 못하고 나무를 본 채로 프로젝트를 진행했던 것 같습니다. 무작정 이전에 좋은 성능을 내었다고 해서 이번에도 같은 방식이 통할 거라는 생각은 하면 안 된다는 것을 알게 되었습니다.

마스터 클래스에서도 많은 영감을 받게 되었는데 우선 성능 높이기에 집중하느라 **dl** 모델을 다양하게 사용해 보지 못한 점이 있습니다. 또한 **attention layer** 를 추가해 볼 것을 말씀해 주셨는데 이 부분은 아직 완벽히 이해가 되지 않아 조금 더 공부해봐야 할 것 같습니다. 또한, 처음 알았던 부분이 있었는데 **GPU** 로 학습시켰을 경우와 **CPU** 로 학습시켰을 경우의 모델의 성능이 다르다는 점입니다. 이 부분에 대해서도 아직 정확한 이유를 알 지 못해 이유에 대해 학습하고, 이후에 적절하게 **CPU** 와 **GPU** 를 사용할 수 있도록 하고 싶습니다. 더불어 데이터 전처리 시 **outlier** 를 고려하지 못 한 점도 있는데, 데이터 양이 많지 않은 상황에서 **outlier** 로 인해 많은 영향을 받을 수 있다는 점을 간과하였습니다. 이 외에도 **target encoding**, **optuna**, **feature engineering**, 딥러닝의 성능이 좋지 않은 이유 등에 대해서도 설명해 주셔서 아주 알찬 마스터 클래스를 보낼 수 있었던 것 같습니다.

이번 프로젝트 기간 동안 개인적인 이유로 프로젝트에 집중을 하지 못했었는데, 안 그래도 팀원이 적은 상황에서 저마저도 도움이 되지 못해 죄송스러웠습니다. 그럼에도 불구하고 1 등한 팀원들이 너무 멋있다고 생각합니다. **level 1** 동안 착하고 대단한 팀원들과 같이 해서 즐거웠습니다! 전공자인 저보다도 다들 잘 하시는 것 같습니다. **level 2** 에서도 다들 행복하게 지내셨으면 좋겠습니다..!

최민수

우선, 이 책 평점 예측 프로젝트를 진행할 때 제 목표는 크게 2 가지였습니다. 첫 번째로, 지금까지 배운 지식을 기반으로 직접 다양한 추천 모델을 코드로 구현해보고자 했습니다. 두 번째로, 이러한 모델을 다양한 최적화 기법과 모델에 대한 연구를 통해 더 나은 성능을 내보고 싶었습니다.

이러한 목표를 달성하기 위해 저는 베이스라인 코드를 분석하였습니다. DCN, CNN_FM 등 다양한 모델에 대해 분석을 진행하고 데이터 전처리 과정도 주의깊게 살펴보았습니다. 코드가 어떤 식으로 구현되어 있는지, 데이터가 어떻게 정제되고 모델의 입력으로 사용되는지를 분석하였습니다. 그 결과, 베이스라인 코드로는 원하는 수준의 성능을 얻기 힘들 것이라고 판단하고 새로운 모델에 대한 탐색을 시작하였습니다.

이 대회에 데이터인 범주형 데이터에 좋은 성능을 보여주는 CatBoost 라는 모델에 대해 학습을 진행해 보았습니다. 모든 범주형 데이터를 one-hot encoding 등의 방식으로 처리하기에는 어려움이 있었고, 성능도 좋지 않았기 때문에 CatBoost 모델을 선택하고 더 발전시켜보기로 결정하였습니다. CatBoost 모델을 설계하면서 다양한 feature engineering 을 시도할 수 있었습니다. 학습 데이터에는 많은 feature 들이 존재하였는데, 데이터를 더 잘 이해하기 위해 EDA 를 통해 통계치를 확인하고 결측치를 처리하는 등의 논리적인 방법으로 feature engineering 을 수행하였습니다. 이후에는 모델을 학습시켜 feature importance 를 확인하고, 필요에 따라 새로운 feature 를 추가하거나 삭제하여 모델을 더 고도화하였습니다. 이러한 다양한 시도를 통해 원하는 성능을 달성할 수 있었고, 최종적으로 1 위라는 좋은 결과도 얻을 수 있었습니다.

이를 통해 딥러닝이 항상 뛰어나다고 생각했었지만, 머신러닝도 적절히 활용한다면 정말 좋은 성능을 보여줄 수 있다는 것을 확인할 수 있었습니다. 또한, feature engineering 의 중요성에 대해 깨닫게 되었습니다. 많은 팀들이 CatBoost 모델을 활용했지만, feature engineering 에 따라 전혀 다른 결과가 나온다는 것을 알 수 있었습니다. 모델 설계뿐만 아니라 데이터에 대한 이해와 다양한 전처리가 얼마나 중요한지를 직접 경험해 볼 수 있었습니다.

이 프로젝트를 진행하면서 많은 한계와 아쉬운 점도 물론 존재했습니다. 첫 번째로, 타겟 인코딩 과정에서 data leakage 를 고려하지 못한 것이었습니다. Feature engineering 과정에서 "avg_rating"이라는 새로운 컬럼을 생성하여 학습을 진행했는데, 이는 책 평점을 예측하는 태스크에서 target leakage 를 유발할 수 있는 부분이었습니다. 이 과정에서 과적합이 발생할 수 있고, 여러 가지의 리스크가 존재한다는 것을 뒤늦게 깨달았습니다. 따라서 binning 이나 사용자를 군집으로 나누어 타겟 인코딩을 진행했다면 더 나은 일반화 성능을 얻을 수 있었을 것으로 생각합니다.

두 번째로, Optuna 를 통해 하이퍼파라미터를 최적화하는 과정에서 다양한 조합을 시도해보지 못한 것 같습니다. Optuna 와 같은 라이브러리를 사용한다면 global optima 로 수렴하기보다는 각 파라미터 간의 상호작용 때문에 주로 local optima 로 수렴하게 되고, 항상 비슷한 값의 결과가 나왔던 것 같습니다. 다양한 조합을 시도했다면 더 일반화 성능이 높은 모델을 얻을 수 있었을 것으로 생각합니다. 또한 CatBoost 모델을 GPU 가 아닌 CPU 로 학습했다면 더 좋은 예측 성능을 얻을 수 있었을 것으로 예상됩니다.

마지막으로, 다양한 딥러닝 모델을 구현해보지 못한 것이 아쉬웠습니다. 기본 베이스라인 코드와 강의를 통해 배운 코드에 대해서는 모델을 구현하고 최적화를 시도해봤지만, 너무 성능에만 집중하여 다양한 시도를 하지 못했습니다. **Attention mechanism**을 활용하여 모델을 구현해보거나, 머신러닝 모델과 딥러닝 모델을 다양하게 앙상블하여 사용했다면 더 다양한 **representation**을 학습하고 활용하여 더 좋은 결과를 얻을 수 있었을 것으로 생각합니다.

이렇게 여러 가지 한계를 마주했지만, 하나씩 분석하고 보완하여 다음 프로젝트를 진행할 때에는 다양한 시도를 해볼 수 있을 것 같습니다. **Feature engineering**을 할 때 **data leakage**는 큰 문제가 될 수 있기 때문에 더욱 신중하게 시도하고, 다양한 모델을 구성하여 최적의 모델을 찾아내기 위해 딥러닝과 머신러닝 모델을 적절히 활용할 것입니다. 또한, 모델을 설계한 후 **hyperparameter** 최적화를 진행할 때에는 모델에 대한 이해를 토대로 다양한 조합에 대한 탐색을 진행할 것입니다.