Wrap-Up Report - 책

1. Team Wrap Up Report

1) 프로젝트 개요

프로젝트 주제

책과 관련된 정보와 소비자의 정보, 그리고 소비자가 실제로 부여한 평점을 활용하여 **사용자가 주어진 책에 대해 얼마나 평점을 부여**할지에 대해 예측합니다.

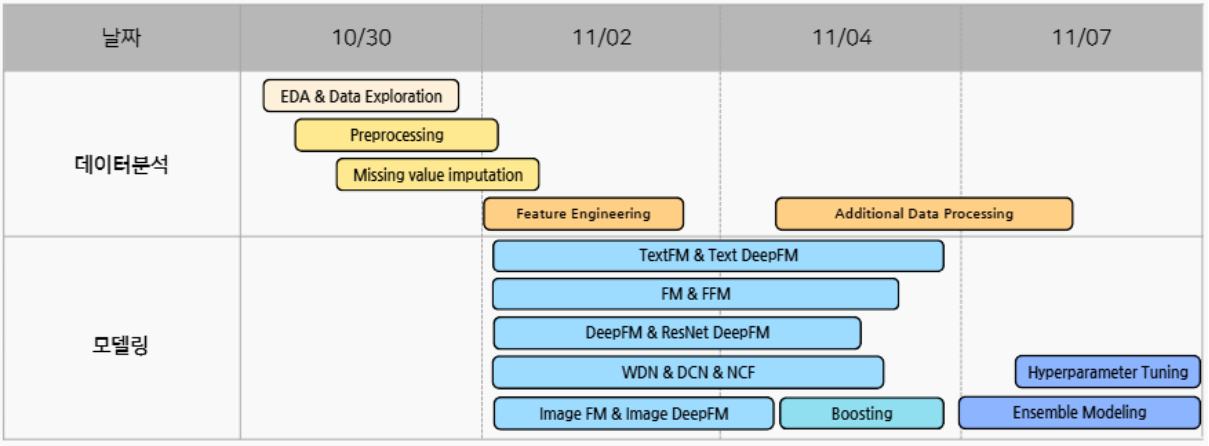
해당 경진대회는 소비자들의 책 구매 결정에 대한 도움을 주기 위한 **개인화된 상품 추천 대회**입니다.

2) 프로젝트 팀 구성 및 역할

이름	역할
강현구	EDA, DeepFM, ResNet_DeepFM
서동준	TextFM, TextDeepFM, LGBM
이도걸	WDN, DCN, NCF, Boosting
이수미	Add features, FM, FFM
최윤혜	Data Preprocessing, ImageFM, Image DeepFM, CatBoost

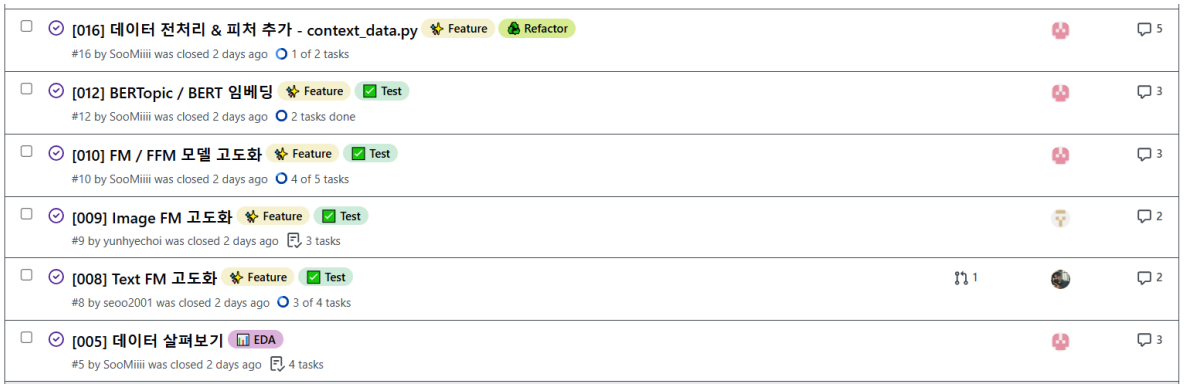
3) 프로젝트 수행 절차 및 방법

수행 과정



협업 문화

1. GitHub 활용



- 체계적이고 효율적인 워크플로우 유지
- 발생한 이슈나 아이디어들을 Issue로 관리
- 회의록 작성 및 공유

2. Wandb & notion 활용

Name	Project	State	Created	User
DeepFM_baseline	book-rating-prediction	Finished	1 day ago	ardkyer
DCN_baseline	book-rating-prediction	Finished	1 day ago	smlee-
WDN_baselinecc	book-rating-prediction	Finished	1 day ago	smlee-
FFM_baseline	book-rating-prediction	Finished	1 day ago	smlee-
Text_DeepFM_baseline	book-rating-prediction	Killed	2 days ago	seoo20001
Text_DeepFM_baseline	book-rating-prediction	Finished	2 days ago	seoo20001

실험 날짜	담당자	Aa 파일명	제출 여부	모델	# Valid RM...	# Valid MAE	# 리더보드...	설명	원인 분석
4년 11월 6일	Dogeol Lee	20241106_150003_WDN	O	WDN	1.7817	1.3062	2.2847	전처리 + 복잡도 낮춰서 + split 0.2 + dropout 0.5 + lr: 1e-3	
4년 11월 6일	최윤혜	catboost_context_data	O	CatBoost	1.548336345		2.3896	수미 코드로 바꾸고 context_data에 catboost 대강	
4년 11월 6일	Dogeol Lee	20241106_161055_DCN	O	DCN	1.97337	1.38389	2.3724	lr: 1e-3, embed_dim: 32, cross_layer_num: 3, mlp_dims: [32, 64, 32], dropout: 0.5	
4년 11월 6일	seodongjoon	lgbm with text_vector_pca	O	LightGBM	2.1959		2.1944		
	이수미(자연과	20241106_08570 열기	O	FM	1.83442	1.27675	2.3986		
4년 11월 6일	seodongjoon	LGBM, Text_DeepFM 앙상블	O	LightGBM, Text_DeepFM			2.1845	average ensemble	
4년 11월 6일	이수미(자연과	20241106_140456_FM	O	FM	1.87816	1.31604	2.3922	윤혜언니 전처리 적용	큰 차이가 없다.....
	이수미(자연과		X	FFM	1.93	1.381		모델만 바꿈 / 과적합됨(train 점수 0.8대)	
	이수미(자연과		X	FM	1.89753	1.34843		SentenceTransformer embedding 를 추가	
4년 11월 7일	최윤혜	20241107_catboost_feature_drop	O	CatBoost	2.1351		2.1495	books데이터 전처리 + 수치형 변수 제거	
	이수미(자연과	20241107_052414_FM	O	FM	1.80733	1.27032	2.3994	epoch 줄임 / title_1 임베딩 추가	FM의 한계인가....

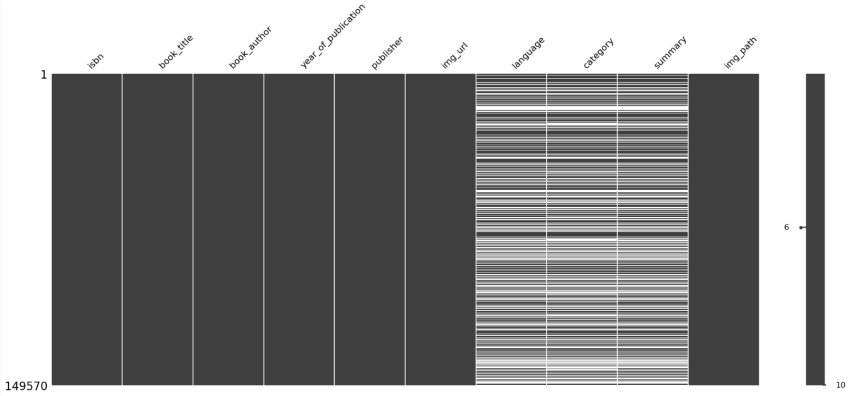
- Wandb와 Notion을 활용해 실험 결과 기록
- 주요 발견 사항 상세 기록 및 저장, 확인 가능

3. Zoom을 활용한 커뮤니케이션
   - 정기적인 아이디어 회의 진행
   - 팀원 간 원활한 소통 및 의견 교환 촉진

4) 프로젝트 수행 결과

EDA & 데이터 전처리

- language, category, summary에 결측치가 아주 많음.



  - language의 44%가 결측치

- language 변수
  - 결측치 처리
    - ISBN 값을 통해 language 결측치 대체

        - 기존 결측치 67227개 → **105개**
- category 변수
    - 통합
        - Sentence Bert 임베딩 활용 카테고리 통합
        - AgglomerativeClustering (계층적 클러스터링)
        - 기존 카테고리 4292개 → **654개**
    - 결측치 처리
        - 동일 저자, 제목의 category 결측치는 최빈값으로 대체
        - 기존 결측치 68,851개 → **27,203개**
- location 변수
    - 이상치 대체
        - 동일 country, city 이면서 state만 다른 경우, 최빈 조합으로 대체
        - 동일 state, city 이면서 country만 다른 경우, 최빈 조합으로 대체
- 추가 변수
    - 데이터의 특성을 더 잘 반영하고자 user_id와 author를 기준으로 총 7개의 변수를 추가
        - user_id: 평균 평점, 평점 개수, 평점 표준편차
        - author: 책 개수, 평균 평점, 독자 수, 평점표준편차
    - 부스팅 모델을 통한 feature importance 확인을 통한 변수 제거
        - `user_avg_rating`, `author_avg_rating`이 유독 높음
        - rating을 평균 낸 변수라서 validation에서 data leakage 발생한 것으로 보임
        - 모델이 해당 변수에 너무 의존하도록 학습되므로 public score에서 성능 하락 → 변수 제거

추가 시도

- Adversarial Validation 실험

1. train, test 예측

	adv_score	feature
0	0.590183	rating_count
1	0.584625	user_avg_rating
2	0.578559	author_unique_readers
3	0.578528	author_book_count
4	0.576913	author_avg_rating
5	0.499834	language
6	0.497827	location_country
7	0.497684	age_range
8	0.497627	location_state
9	0.497359	publication_range
10	0.495376	user_id
11	0.495371	category
12	0.495369	location_city
13	0.491749	book_title
14	0.491441	book_author
15	0.491188	publisher
16	0.488370	isbn

2. train, valid 예측

	adv_score	feature
0	0.587997	rating_count
1	0.582553	user_avg_rating
2	0.578643	author_unique_readers
3	0.578248	author_book_count
4	0.575700	author_avg_rating
5	0.500522	language
6	0.499649	age_range
7	0.497840	location_country
8	0.497676	publication_range
9	0.497412	location_city
10	0.496612	category
11	0.496554	location_state
12	0.494765	user_id
13	0.492156	book_title
14	0.491323	publisher
15	0.490719	book_author
16	0.489356	isbn

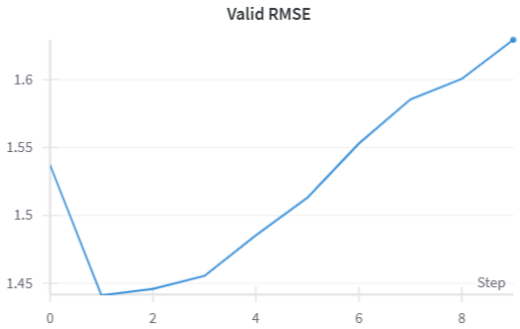
    - train 데이터와 valid/test 간의 피쳐 분포 차이가 존재하는지 확인
        - train, test셋을 각각 0,1로 target
        - 각 변수별로 lgbm 모델 적합
        - target 정확도 계산
        - 0.75 이상인 변수 제외
    - 0.5 이상의 정확도를 보이는 변수가 있긴 하지만 모두 0.6 이하로 엄청 높진 않음
- 카테고리 통일 노력
    - **BERTopic**: BERT 기반의 토픽 모델링 구현체
        - 제목 + 카테고리 + 요약을 하나의 list로 만들어서 토픽으로 임베딩
        - 결과: -1(이상치) 가 너무 많이 나와서 사용하지 않음
    - SentenceTransformer embedding
        - **Sentence-BERT**의 all-MiniLM-L6-v2모델을 활용해서 title 혹은 title+category+summary 를 수치벡터로 변환
        - 384 차원의 임베딩 벡터를 NN 모듈을 통해 **n차원으로 축소**
        - book 데이터에 추가 → 유의미한 성능 향상을 보이지는 않음.

모델 개요

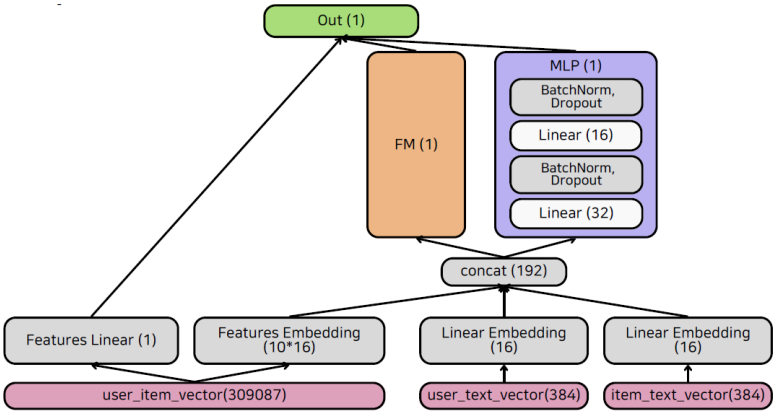
- 기본 모델

모델명	특징
FM	피처 간 상호작용을 저차원 벡터의 내적으로 모델링
FFM	각 피처가 속한 필드에 따라 다른 임베딩을 학습
NCF	딥러닝을 활용한 협업 필터링 모델로 사용자-아이템 상호작용을 비선형적으로 모델링
DCN, WDN	피처 간 교차항을 자동으로 학습하는 심층 신경망 구조로 광범위한 피처 조합 학습
DeepFM	FM과 딥러닝을 결합한 모델로 저차원과 고차원 피처 상호작용을 동시에 학습
Image FM, Image DeepFM, Resnet DeepFM	이미지 피처를 활용하기 위해 CNN 구조로 이미지 특징을 추출하여 FM/DeepFM과 결합
Text FM, Text DeepFM	텍스트 정보를 활용하기 위해 텍스트 임베딩을 FM/DeepFM과 결합
LGBM, XGB, CatBoost	결정 트리 기반의 그래디언트 부스팅 앙상블 모델
Catboost_cat	rating 재범주화를 통해 자주 등장하는 rating 예측에 집중시킨 모델

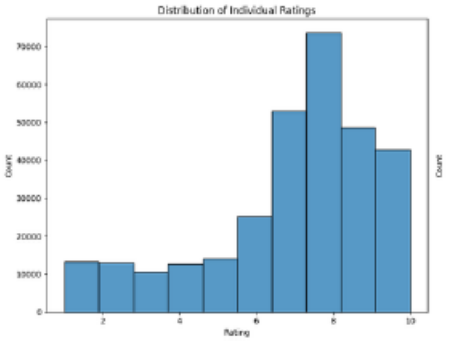
    - 모델 개선 시도
        - FM, FFM, NCF, DCN, WDN, DeepFM, Image FM, Image DeepFM, ResNet DeepFM
            - valid RMSE가 지속적으로 증가하는 불안정한 학습 곡선으로 과적합 경향이 두드러짐
            - 과적합 방지를 위하여 모델 단순화를 시도하였으나, 크게 개선되지 않음



- Text DeepFM
  - Auto Tokenizer의 pretrained 모델 변경
    - `bert-base-uncased'` → `sentence-transformers/all-MiniLM-L6-v2`
  - 모델 구조 변경
    - `user_text_vector`, `item_text_vector` 의 임베딩 파라미터를 분리하여 개별적으로 학습
    - text vector에 대해서 MLP를 적용하여 저차원으로 추가 임베딩
  - 데이터셋 추가
    - context 데이터 추가



- Catboost_cat
  - 기본적으로 rating 값은 7~10에 몰려 있음
  - 따라서 드물게 등장하는 값들을 재범주화함으로써 자주 등장하는 rating 값 예측 정확도를 높이고자 함
  - 재범주화 후, 동일하게 regression 문제로 학습
    - 범주화 기준

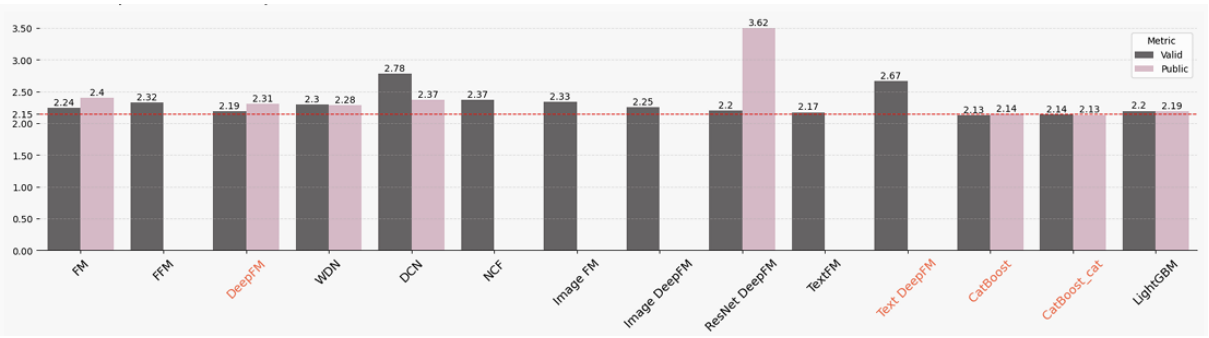


구간	값
1~3	2
4~6	5
7~7	7
8~8	8
9~9	9
10~10	10

모델 성능 평가 및 앙상블

- 모델별 성능 평가

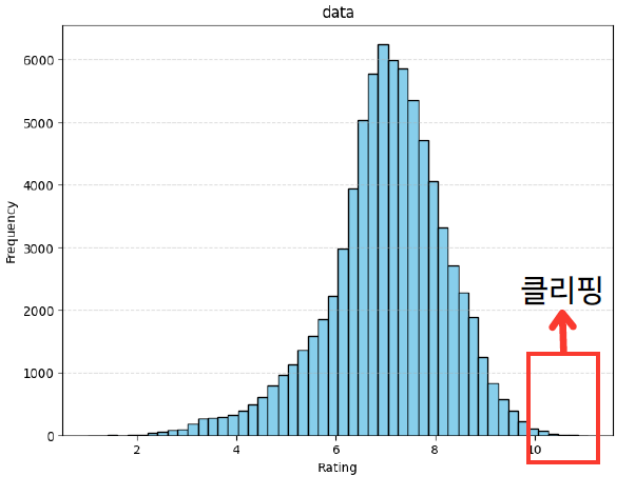
모델명	Validation RMSE	public Leaderboard RMSE
FM	2.24	2.4
FFM	2.32	
DeepFM	2.19	2.31
WDN	2.3	2.28
DCN	2.78	2.37
NCF	2.38	
Image FM	2.33	
Image DeepFM	2.25	
ResNet DeepFM	2.2	3.62
TextFM	2.17	
Text DeepFM	2.67	
Catboost	2.13	2.14
Catboost_cat(재범주화)	2.13	2.13
LGBM	2.2	2.19



- 모델 선정 기준
  - 아래 두 가지 기준에 따라 모델을 각각 두개씩 선정

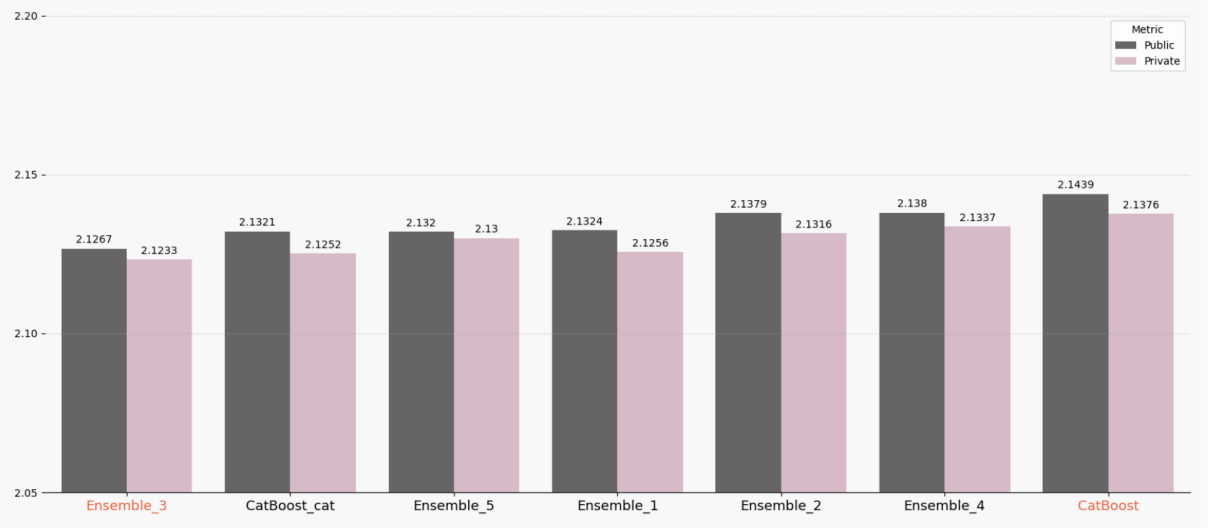
1. **Boosting Model**: Validation Score, Public Score < **2.15**
   →
   CatBoost, CatBoost_범주화
2. **DL Model**:Validation Score < **2.2**
   →
   DeepFM, Text_DeepFM

- 앙상블 모델
  - **average ensemble** 후에 **클리핑**
    - 1 미만인 경우 1로 대체
    - 10 초과인 경우 10으로 대체



  - 앙상블 목록 & 최종 결과

No.	Ensemble Configuration	Validation RMSE	Leaderboard
1	Catboost + Catboost_cat	2.1324	2.1256
2	Catboost + Catboost_cat + DeepFM	2.1379	2.1316
3	Catboost + Catboost_cat + Text DeepFM	**2.1267**	**2.1233**
4	Catboost + Catboost_cat + FM + Text DeepFM	2.138	2.1337
5	Catboost_cat + Text DeepFM	2.132	2.13



- 최종 선정 모델: **Ensemble_3**, **Catboost**
  - 리더보드 결과: **Public Score 2위**(2.1267), **Private Score 2위**(2.1233)

    → 앙상블 모델의 성능이 가장 좋았다.

자체 평가 의견

- 새롭게 배운 점
  - 체계적으로 짜여있는 베이스라인 코드를 통해 코드 모듈화를 배울 수 있었다.
  - shell script로 실행하는 코드의 편리함을 느꼈다.
  - wandb를 활용하여 실험을 관리하고 협업하는 방법을 배웠다.
- **잘했던 점**
  - 대회에서 좋은 순위를 얻을 수 있었다.
  - 하나 더 추가하기~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
- **아쉬운 점**
  - 시간도 부족했고 summary에 null 값이 많아서 LDA나 Top2Vec와 같은 토픽 모델링을 적용시켜보지 못했다.
  - 여러 베이스라인 모델이 학습되면서 Valid Loss가 수렴하지 않는 문제를 해결하지 못하였다.
  - Text 모델의 vector 생성 프롬프트에 Category를 추가하는 방법을 시간이 부족해서 시도해보지 못했다.

2. 개인 회고

1) 강현구_T7502

초기 목표와 달성한 정도

추천시스템에 대해 처음 배워보는 만큼 솔직히 여러 가지로 시도해보고 싶었다. 근데 시간이 상대적으로 부족해서 초기 목표를 모든 모델을 다 사용해 보자. 이렇게까지는 아니었던것 같다.

주어진 시간에 비해 목표를 약간 낮게 설정해서 그런가 아 이번 프로젝트에 진짜 아무것도 안했다 이런느낌은 아닌 정도인것 같다.

내가 가장 신경 썼던 점

원래 바람직한 학습은 기초부터 행해져야 하는게 맞긴한데 이번엔 시간이 너무 없어서 결과를 우선 내고 그 다음 학습하는 방식으로 진행한 느낌이 없지않아 있다. gpt의 도움도 이전보다 조금 많이 받았고. 바람직하지 않다는 것을 알긴하지만 빠른 학습을 위해 그런 방향으로 진행한 것 같다.

구현하지 못한 아이디어

많은 것을 구현하고 싶었지만 하지 못했다. 뭘 알아야 구현을 하고싶어지기 마련인데 아는게 부족하여 아쉽다.

한계와 아쉬웠던 점

문제를 정확히 인식하고 해결하는 걸 가장 중요시하지만 내가 베이스라인과 여러모델에 대해 정확히 이해하지 못했다는 것을 느끼고 있기에 더욱 아쉽다.

다음에 시도해볼 것

우선 기초적인 공부를 다시 하고

다음번엔 시간이 조금 기니까 streamlit을 사용해보고 싶다. 시간이 만약 된다면.

마지막 한마디

벌써 프로젝트 3개를 마쳤다. 1개만 더 끝내고 좀 돌아보는 시간을 가지자.

2) 서동준_T7527

초기 목표와 달성한 정도

추천 시스템 관련 프로젝트 경험이 없었기 때문에, 추천 시스템 프로젝트의 데이터와 모델을 잘 이해하는 것을 목표로 임했다. 짧은 기간에 최대 효율로 많은 모델을 공부하고 적용해보면서 초기 목표에 만족스러운 지식들을 얻은 것 같다.

내가 가장 신경 썼던 점

맡았던 모델을 잘 이해하고 성능을 개선해보는 경험을 가장 중요하게 생각했다. FM, DeepFM 모델의 기본적인 구조를 이해하고, Text 임베딩 모델의 학습 과정을 이해하려고 노력했다. 다양한 실험을 해보면서 Text DeepFM 모델을 최종 앙상블에 활용할 수 있는 수준으로 성능을 개선했다.

구현하지 못한 아이디어

프로젝트 후반부에 팀원의 아이디어로 Text Vector 생성 프롬프트에 category와 같은 다른 Text 데이터도 활용하는 방법을 시간적인 한계로 시도해보지 못했다.

한계와 아쉬웠던 점

모든 모델을 다 살펴보기에는 시간이 부족해서 팀원끼리 모델을 나눠서 프로젝트를 진행했다. 너무 짧은 기간동안 진행된 프로젝트라 모든 모델을 살펴볼 시간이 없었던게 아쉬웠다.

CF 기반 모델을 활용하기에는 User-Item 상호작용 데이터가 너무 부족했다. 그래서 콘텐츠 기반 모델이나 부스팅 모델에서만 좋은 성능을 낼 수 있었고, 다른 모델들을 잘 활용하기 어려웠다.

다음에 시도해볼 것

그래프 기반 추천 모델들도 시도해보고 싶다.

마지막 한마디

추천 모델 성능 올리기 어렵다.. 대회에서 높은 점수를 받는 모델이 실제 서비스 환경에서도 좋은 성능을 보일지가 궁금하다.

3) 이도걸_T7540

초기 목표와 달성한 정도

1. 등수보다는 학습
    - 처음 해보는 추천 프로젝트이기에 프로젝트 등수보다는 학습에 초점을 두려고 했었다.
    - 하지만 프로젝트 결과를 의식하고 이것저것 하다보니 학습도 애매하게 했고 프로젝트고 애매하게 해서 어떤게 남았는지 잘 모르겠다.
2. Baseline Code 구조 분석
    - 베이스라인 코드가 잘 제공되었기에 구조와 방식들을 잘 파악해서 다음 프로젝트에 적용하고자 했다.
    - 어느정도 구조와 방법들을 익혔고 다음 프로젝트에서 적용해봐야겠다.

내가 가장 신경 썼던 점

1. 이전 프로젝트의 코드를 잘 활용하기
    - 짧은 시간안에 새로운 모델을 추가하기는 어려울 것으로 생각해서 이전에 진행했던 프로젝트에서 바로바로 실험할 수 있는 모델이나 함수들을 최대한 활용하려고 했다.

구현하지 못한 아이디어

1. Classifier로 구현한다면 어떤 결과가 나왔을까?
    - Classifier를 학습하기 위한 과정 중간에 다른 길로 새어버려서 실제로 학습해보질 못했다.
    - 흥미로운 결과가 나왔을 것 같기도 하다.
2. 다른 Tree 모델이나 RF로 학습 했다면 결과가 어땠을까?
    - Catboost의 성능을 이기기는 어려웠겠지만 앙상블용 모델로 활용할 수 있었을 것 같다.

한계와 아쉬웠던 점

1. 시간 부족으로 인한 모델 이해도 부족
    - 모델의 구조나 특성을 이해하지 못한채로 프로젝트를 진행해서 파라미터를 바꾸는 것 이외의 수정을 하기에 어려웠다.
    - 모델의 Validation Loss가 발산하는 문제에 대해 해결하지 못했다.
2. 실제로 프로그래밍 한 것이 없다는 점
    - Baseline code에서 많은 모델과 전처리등 프로젝트에 필요한 거의 모든 것을 제공해주었기에 실제로 내가 작성한 코드는 많이 없다는 점이 아쉬웠다.

- 프로그래머보다는 러너가 된 것 같아 흥미가 조금 떨어졌던 것 같다.

다음에 시도해볼 것

1. Baseline code와 우리팀의 base 코드의 결합
    - Baseline code를 우리팀에서 사용하던 구조에 맞게 바꿔서 조금 더 코드이해도를 높이고 사용하기 익숙한 방향으로 작업해보고 싶다.
2. 다양한 데이터 EDA와 Feature embedding
    - 내가 만든 Feature가 학습에 좋은 영향을 끼치는 경험을 할 수 있으면 좋겠다.
3. 평소 사용해보지 못 한 모델들을 사용
    - 항상 사용하는 모델말고 다양한 모델들을 사용해보고 싶다.

마지막 한마디

진짜 정신없고 바쁜 한 주였따..

4) 이수미_T7541

초기 목표와 달성한 정도

이전까지 다뤄보지 않았던 추천 시스템의 다양한 모델과 모듈화된 baseline 코드 작동 방식 이해를 목표로 하였으나 생각보다 주어진 시간이 짧아서 내가 맡은 부분을 제외하고는 건드리지 못하였다. 그래도 FM 모델은 이해할 수 있었다.

내가 가장 신경 썼던 점

내가 맡았던 FM 모델의 성능 향상과 통합되지 않은 category 변수의 통합 부분에 가장 많은 시간을 쏟았다.

구현하지 못한 아이디어

제목, 카테고리, 요약 데이터를 이용하여 토픽 모델링을 해서 보다 정확한 카테고리로 분류하고자 시도는 하였지만 적절한 결과가 나오지 못했다.

한계와 아쉬웠던 점

- EDA도 모델링도 둘 다 애매하게 건들기만 해서 유의미한 결과를 도출하지는 못한 것 같다. 하나를 하더라도 제대로 해내는 것이 더 중요하다고 느꼈다.
- 모델을 학습시킬 때 valid RMSE 점수가 지속적으로 오르는 현상이 있었는데 원인을 파악하지 못했다.
- FM 모델을 주로 맡아서 그런건지, 평점 예측이라 그런건지 지금까지 공부해왔던 머신러닝 모델들과 추천 모델의 차이점이 느껴지지 않았다.
- 기한 내에 해내기 위해 분량을 나누다보니, 이미지와 텍스트 모델에는 손도 대보지 않았다는 점이 아쉽다.

다음에 시도해볼 것

- 하나를 집중적으로 파기
- baseline code 처럼 실행 파일도 모듈화 해보기

마지막 한마디

아쉬움이 많이 남는 프로젝트지만 그럼에도 불구하고 얻어가는 것은 있는 것 같다. 다시 열심히 해봐야겠다.

5) 최윤혜_T7549

초기 목표와 달성한 정도

- 추천 모델 학습하기
    - 추천 모델을 써보는 것은 처음이었기 때문에 모델 구조나 학습 방식 같은 것들을 이번 기회에 제대로 배워보고자 했다. 하지만 이미 베이스 코드가 만들어져 있는 상태에서 주어진 시간이 짧다보니 모델에 대한 공부보다는 성능 올리기에 집중하게 된 것 같다. 담당했던 Image FM 쪽도 대략적인 모델 구조만 이해하고 넘어간 것 같아서 아쉽다.

내가 가장 신경 썼던 점

- 데이터 전처리
    - 결측치가 많은 데이터였다. 그래서 books 데이터의 location, category, language 변수의 결측치 처리에 시간을 많이 썼다. 최대한 정확한 값으로 결측치를 대체하기 위하여 변수 특징들을 꼼꼼히 살펴보았고, 결측치 처리 이후에 모델 성능이 개선될 수 있었다.

구현하지 못한 아이디어

- context, image, text 정보 모두 활용하기
    - image FM, text FM 처럼 image나 text만을 활용하는 모델에서 발전시켜, image vector와 text vector를 연결하여 한 모델 내에서 image와 text를 모두 함께 처리했다면 어땠을지 궁금하다.

한계와 아쉬웠던 점

- 시간이 너무 짧았다.
    - 너무 짧은 시간 내에 프로젝트를 진행하다 보니 기대했던 것 만큼 배우지 못한 것 같다. 결국에는 추천 모델은 포기하고 데이터와 부스팅 모델에 시간을 쓰게 됐다. 좀 더 새로운 추천 모델을 공부하고 싶었는데 이전 프로젝트에서 해온 분석을 또다시 하게 된 것 같아서 아쉽다.

다음에 시도해볼 것

- 추천 모델 제대로 활용해보기. 모델 구조 이해도 높이고, 데이터에 적합한 모델이 무엇인지 잘 고민해봐야겠다.
- 베이스라인 코드처럼 코드 모듈화와 shell script 활용도를 높여봐야겠다.

마지막 한마디

추천 모델 어렵다..