

책 평점 예측 대회

RECSYS-08조 LEVEL2 프로젝트 랩업 리포트

프로젝트 개요

프로젝트 개요

목표: 사용자의 책 선호도를 예측해 개인화된 책 추천 모델 개발
상세 목표: 유저와 책 메타 데이터를 결합해 각 특징을 학습하고, 새로운
 유저-아이템 쌍에 대한 평점을 예측하는 모델을 구축
기간: 2024.10.28 ~ 2024.11.08 (2주)
예측 타겟: 평점('rating' 변수)
평가지표 RMSE

팀 구성 및 역할

팀원 수: 6명
팀 구조: • 조유솔: Text Embedding, Multimodality, WDN, DIN
 • 박세연: Feature Engineering, Multimodality(MLP), LGBM
 • 배현우: EDA, LGBM + HPO
 • 김영찬: EDA, NGCF modeling
 • 박광진: EDA, Multimodal learning(ViT + BERT)
 • 박재현: EDA, data preprocessing, Multimodality,
 LGBM, CatBoost

협업/개발 툴

코드 공유

 GitHub

실험 관리

 Notion TeamSpace

개발 환경




- IDE: VSCode
- 작업 환경: 로컬, 원격 GPU서버(Linux V100 서버)
- 가상 환경: Conda

타임라인

기간: 2024.10.28 ~ 2024.11.08 (2주)

- (1) 베이스라인 + EDA: 2024.10.28 ~ 2024.11.03 (1w)
- (2) 전처리 및 데이터 가공: 2024.11.02 ~ 2024.11.4(3d)
- (3) 모델링: 2024.11.05 ~ 2024.11.08(4d)

기술 스택

- 프로그래밍 언어:  Python, Bash Script, YAML
- 프레임워크 :  PyTorch  Scikit-learn
- 모델: DeepFM, WDN, FFM, ViT, LightGBM, XGBoost, Catboost 등
- 시각화: Seaborn, Matplotlib, Plotly
- 기타: Argparse, OmegaConf 등

데이터셋 소개

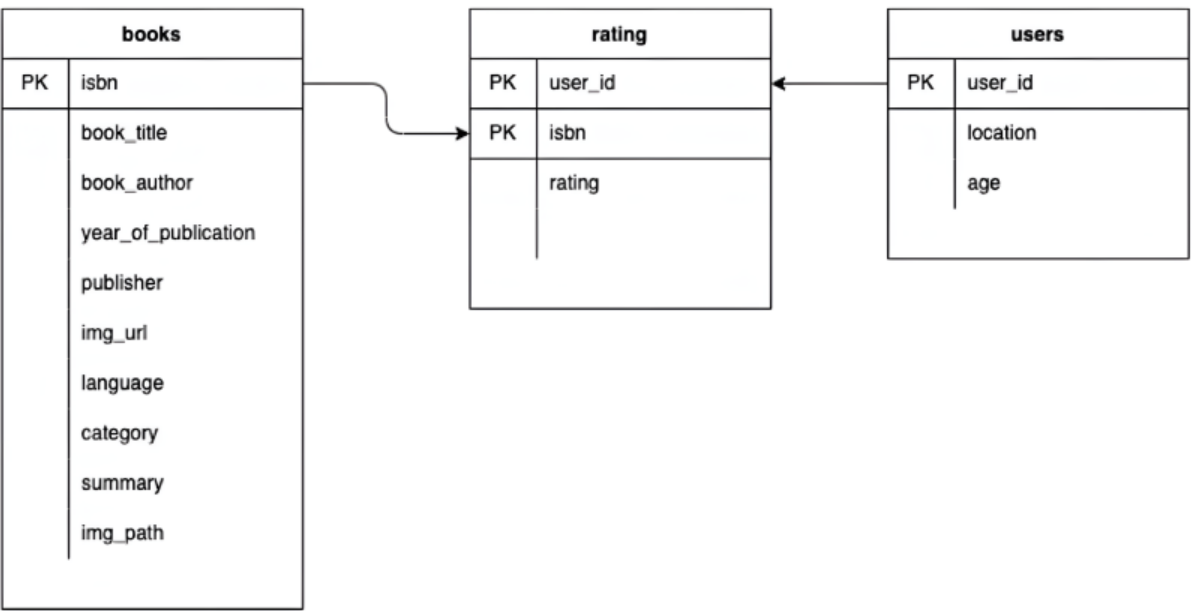
파일

유저, 책, 유저-책 상호작용에 대한 정형/비정형 데이터

구성

- books.csv: 149,570권의 책 정보 (제목, 저자, 출판연도, 커버 이미지 등) [149,570행]
- users.csv: 68,092명의 사용자 정보 (지역, 나이)[68,092행]
- train_ratings.csv: 59,803명의 사용자가 129,777권의 책에 남긴 평점 [306,795행]
- test_ratings.csv: 평점 예측을 위한 사용자-책 목록 (평점 0) [76,699행]

Data Schema



users.csv

| user_id | location | age |
|---------|--------------------------|------|
| 8 | timmins, ontario, canada | NaN |
| 11400 | ottawa, ontario, canada | 49.0 |

{}_ratings.csv

| user_id | isbn | rating |
|---------|------------|--------|
| 8 | 0002005018 | 4 |
| 67544 | 0002005018 | 7 |

books.csv

| isbn | book_title | book_author | year_of_publication | publisher | img_url | language | category | summary | img_path |
|------------|----------------------|------------------|---------------------|-----------------|------------------|----------|---------------|------------------------|-----------------|
| 0002005018 | Clara Callan | Richard Bruce.. | 2001.0 | HarperFlam.. | http://images... | en | ['Actresses'] | In a small town in .. | images/000201.. |
| 0374157065 | Flu: The Story of .. | Gina Bari Kolata | 1999.0 | Farrar Straus.. | http://images.. | en | ['Medical'] | Describes the great .. | images/037401.. |

툴

EDA 과정에서 Python의 C++ 기반 라이브러리인 graph-tool을 사용하여 데이터의 구조적 특성을 분석

분석 결과

- [훈련 및 테스트 데이터 분석] 각 데이터에서 평점이 가장 많은 유저 상위 30명을 분석, train과 test 데이터가 비슷한 분포를 보임
→ 두 데이터가 원래 하나의 데이터였을 가능성이 있다고 추측
- [비정상 계정 탐색] 평점이 지나치게 높고, 표준편차가 낮은 유저를 대상으로 평점이 5개 이상인 경우를 추려내어 스팸 계정 여부를 확인
→ 각 유저가 평가한 책들이 서로 겹치지 않는 점을 발견하여, 실제 스팸 계정은 아니고 다만 평가 기준이 후한 사용자로 판단

전처리

- isbn 컬럼에 isbn13, isbn10 형식 혼재 → isbn13으로 통일
- location 컬럼을 국가, 주, 도시로 분할해서 하나씩 실험
- summary, category 컬럼은 결측치가 2/3이상 존재해 제거
- 수치형 변수로 age, year_of_publication 사용
- age 결측치 중앙값으로 채움
- book_title, book_author, publisher는 깨진 문자가 있어 변환
- year_of_publication은 이상치 처리 → 1920년~

Feature Engineering

- 사용한 데이터
이미지 데이터 및 텍스트 데이터는 제공된 데이터셋 그대로 사용.
정형 데이터 모두를 카테고리 형태로 변환 후 모델에 적용.
- Feature Engineering
유저 평균 평점 및 책의 평균 평점을 새로운 컬럼으로 추가하여 실험.
user_id 컬럼과의 다중공선성 때문에 정확도가 더 낮아지는 결과를 얻음.
- 최종적으로 사용한 데이터셋
multimodal을 이용한 모델링 시에는, 이미지(책 표지), 텍스트(책 제목, 작가, 요약), 범주형(나이, 출간연도, 사용자 아이디, 책 넘버)를 사용
최종적으로 LGBM 모델링 시에는, 정형 데이터만을 베이스라인과 같은 전처리 방식을 사용하여 구성

MLP

- 배경

이번 데이터셋은 이미지 데이터, 텍스트 데이터와 같은 비정형 데이터가 존재함. 따라서 각각의 비정형 데이터를 같은 벡터 공간에 표현하여 최종 아웃풋을 내볼 수 있음.

- 실험 내용

아래와 같은 pre-trained 모델들을 이용하여 제공된 데이터셋에 맞게 추가로 학습 시키는 fine-tuning 과정을 거침.

1. 이미지 데이터 - resnet18, resnet50, ViT
2. 텍스트 데이터 - BERT, roberta, electra

비정형 데이터를 임베딩하여 벡터 형태로 변환한 후, 정형 데이터와 함께 mlp에 입력하여 최종 예측 수행. 연속형 데이터를 범주형 데이터로 변환하거나, 표준화하여 mlp에 맞는 형태로 변환.

다만, 엄청 많은 카테고리 데이터를 어떻게 모델에 맞게 변형해야할 지 고민하는 과정을 겪음.

결국 제외하고 실험.

- 실험 결과

이미지, 텍스트, 정형 데이터를 각각 단독으로 사용하는 경우(베이스라인 모델)보다는 좋은 결과를 내는 것처럼 보임.

각 데이터 형태에 잘 작동하는 즉, 좋은 결과를 내는 데이터가 존재하여 시너지를 내는 것으로 추측가능

WDM

- 배경

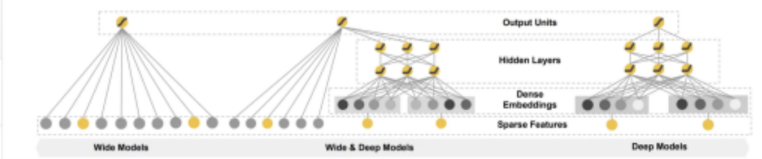
기존 multimodal 모델은 2-layer MLP, 단순 context 정보와 image, text 임베딩 정보를 2개의 FC레이어에 통과시키는 식으로 구성됨. 단순한 특성과 복잡한 관계를 동시에 학습하는 구조가 필요하다고 판단하여 Wide Layer, Deep Layer 를 모두 가진 WDM 모델을 적용함.

- 적용사항:

- Wide 파트 추가 = 간단한 특성(ex. user_mean_rating, book_mean_rating 등)을 단순히 더하는 FeaturesLinear Layer 구현
- Deep 파트 추가 = 기존 MLP 구조 활용해 이미지, 텍스트, 수치형 데이터 통합
- Wide와 Deep의 출력 결합, 최종 rating 예측값 생성
- 이후 모델 경량화를 위해 Tokenizer, Text Embedding 모델을 ELECTRA 로 변경

- 결과표

| | 2-Layer MLP | WDM | WDM with ELECTRA |
|-------|-------------|-------------|------------------|
| RMSE | 2.4803 | 2.4786 | 2.3798 |
| 소요 시간 | 55m / epoch | 60m / epoch | 49m / epoch |



- 결과 해석

- 기존의 2-Layer MLP와 비교했을 때, WDM의 정확도 향상은 0.0685%에 불과한 반면 개별 에포크당 소요시간은 9% 증가. → 성능향상보다 비용 증가율이 더 큼
- BERT 기반의 텍스트 임베딩이 주요 요인이라 판단함. BERT는 큰 모델 크기로 인해 학습은 물론 추론 비용이 높은 편. 따라서 BERT의 light 버전이라, 속도가 빠르고 효율적인 ELECTRA로 임베딩 알고리즘 변경
- BERT기반 WDM 대비: 정확도 +3.986%, 에포크당 소요시간 -18.33%
- 2-Layer MLP 대비: 정확도 +4.05%, 에포크당 소요시간 -10.91%

Multimodal models + LGBM, CatBoost

- 실험 배경:

본 실험은 제공된 다양한 데이터를 활용하여 사용자 책 평점을 예측하는 것을 목표로 함. 사전학습된 모델 (RoBERTa, ResNet50, EfficientNet, CLIP)은 텍스트와 이미지 데이터를 임베딩하는 데 활용되었으며, 이를 기반으로 MLP 모델을 학습함. 또한, 정형 데이터와 카테고리형 데이터의 특성을 활용해 각각 LGBM 과 CatBoost 모델을 사용하여 별도 실험을 수행함.

- 실험 요약

사전학습된 모델인 RoBERTa, ResNet50을 활용하여 제공된 데이터셋 임베딩 후 MLP 모델을 학습함. 이 모델을 기준으로 피처를 수정해가며 실험을 진행함. 이후, EfficientNet, CLIP을 활용해 추가 실험을 진행함. 이 중 CLIP 모델이 가장 우수한 성능을 보였지만 epoch 당 가장 오랜 학습시간이 걸림

LGBM으로 텍스트와 정형 데이터만을, CatBoost 모델로는 카테고리형 데이터만을 학습함.

- 결과 요약

모든 텍스트와 정형 데이터가 주로 카테고리형이라 CatBoost 모델이 가장 우수한 성능을 보였음. Optuna를 적용해 성능을 최대화함.

이미지 데이터는 책 표지로 구성되어, 유저의 책 평점 예측에 큰 영향을 미치지 않았던 것으로 판단됨.

GCN

| 모델 | 장점 | 단점 |
|----------|------------------------------------------------------------------------------------|-----------------------------------------------------------|
| GCN | - 그래프 구조를 반영한 정보 집약 가능 - 노드의 지역적 정보와 인접 노드 간의 관계를 효과적으로 학습 | - 계산 비용이 높아 대규모 그래프에는 비효율적 - 깊이가 깊어질수록 오버스무딩 문제 발생 가능 |
| LightGCN | - 파라미터 효율성 개선 - 연산 간소화로 인한 빠른 학습 시간 - 경량화로 메모리 사용 감소 | - 피처 다양성이 낮아 복잡한 관계 학습에는 한계 - 단순화로 인한 성능 저하가 개인적으로 우려됨 |
| NGCF | - 사용 가능한 피처가 더 다양해 각 노드 간 상호작용을 보다 정교하게 학습 - 네트워크의 구조적 특성과 사용자-아이템 관계를 함께 학습 가능 | - 학습 시간이 길고 메모리 사용량이 높음 - 연산 복잡도가 증가하여 구현 및 최적화가 어려움 |

1. 모델 선택 근거

경진대회 환경에서는 시간이 제한된 경우도 많으나, 이번 프로젝트에서는 가능한 주어진 데이터를 모두 활용하는 것이 더 좋은 결과를 얻을 가능성이 있다고 판단. 이러한 이유로, 계산을 간소화한 LightGCN 대신, 딥러닝 모델을 사용하기로 결정.

2. NGCF 모델의 장점

GCN과 NGCF 논문을 참고하여 NGCF는 GCN 대비 다양한 피처를 반영할 수 있다는 점에서 더 적합하다고 판단. 예를 들어, NGCF는 단순한 인접 노드 정보뿐만 아니라 노드 간의 상호작용과 사용자-아이템 관계를 세밀하게 표현할 수 있어 데이터의 복잡한 패턴을 학습하기에 유리하므로, 본 프로젝트에 더 알맞다고 판단.

3. 모델 학습 과정

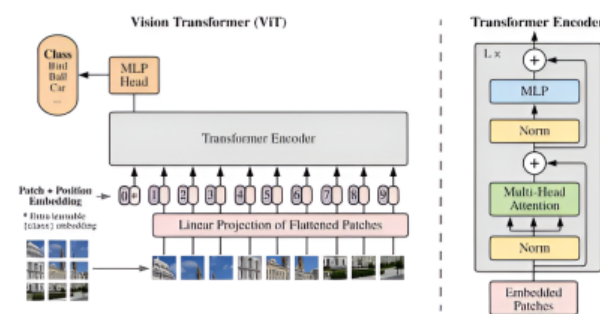
인터넷에 공개된 NGCF 코드를 사용. 프로젝트 특성에 맞게 AI의 도움 없이, 코드의 각 모듈을 직접 읽어보고 가진 데이터에 맞게 수정하여 학습을 진행. NGCF 모델은 복잡한 연산으로 인해 학습시간에 6시간 이상 소요되어, 시간 부족으로 리더보드에 제출하지 못함. 중간평가에서 validation RMSE는 7~8 정도로 확인.

ViT

- 배경

금번 대회는 정형데이터와 함께 비정형 데이터가 존재하여, Multimodal Learning이 필요하다고 판단되어 이미지 인코더 및 텍스트 인코더를 각각 지정하여 처리해 보았음. 그 중, 이미지 인코더로 우수한 성능을 낸다고 알려진 Transformer 계열 모델인 ViT를 사용하여 보았음.

- ViT?



2021년 Google Research에서 발표한 내용으로, Image Classification 문제에 Transformer 구조를 적용시킨 모델. CNN 구조 대부분을 Transformer로 대체한 모델로서, CNN에 비해 훨씬 적은 계산 리소스로도 우수한 성능을 낸다는 사실이 입증된 모델임.

- 선택 이유

- ViT는 Transformer 기반 모델로서, NLP를 처리하는 아키텍처를 동일한 아키텍처를 갖고 있음. 이에, 텍스트와 이미지 데이터를 동일한 구조로 처리할 수 있어 Multimodal Learning에 적합한 구조를 제공할 것으로 예상되어 수행해 보았음.

- 결과

사전학습 모델 이용(google/vit-base-patch16-224)

| | |
|--------------------|----------------------------------|
| Text Encoder with, | BERT |
| Epoch | 5회 |
| RMSE | 2.4611(private) / 2.4626(public) |
| Training time | 110 min/epoch |

- 결과 해석

예상보다 좋은 성능을 보지 못하였고, 이를 아래 두가지 이유로 해석해봄

- 이번 대회의 데이터셋의 크기에 비해, 과한 모델을 사용하여서 그럴 것이라 예상
→ ViT는 대규모 데이터셋에 적합하도록 pre-trained된 모델, 학습 데이터가 적을 경우 Overfitting
- 이미지가 제공하는 정보량이 텍스트에 비해 현저히 낮음
→ 이미지는 표지만 제공할 뿐, 책에 대한 어떠한 정보도 제공하지 않음. 반면, 텍스트 정보는 Summary 부터 시작하여 여러가지 정보를 제공함.

GBDT

- 실험 배경:

이미지 데이터와 텍스트 데이터를 제외한 컨텍스트만 포함 정형 데이터만을 사용하게 된다면 트리 기반 모델이 잘 작동하지 않을까 하는 가설에서 출발. 빠르게 검증해보기 위해 LightGBM 모델을 사용하여 학습 진행.

- 실험 내용:

이미지와 텍스트를 제외하고 유저-아이템 상호작용 데이터와 컨텍스트 데이터를 카테고리화 변환하여 LGBM을 학습하여 성능을 확인함. Optuna를 사용해 하이퍼 파라미터를 튜닝 한 뒤에 KFold를 통해 Out-of-fold 소프트 보팅 앙상블을 진행.

- 실험 결과:

| | Valid | Test(public, private) |
|-------|--------|-----------------------|
| RMSE | 2.20 | 2.1940 / 2.1933 |
| 소요 시간 | 2m 11s | 2m 11s |

테스트 데이터에 대해 단일 모델로 가장 좋은 성능을 보임. 단순 유저 평균 평점과 책 평균 평점 피쳐는 성능의 하락을 불러와 제외.

정형 데이터의 강자 답게 트리 기반인 LGBM이 다른 모델들을 쉽게 제치는 것을 볼 수 있음. 사용자의 평점을 예측하는 회귀 문제에서는 정형 데이터에 대해 트리 모델이 유리한 것을 확인.

최종 모델

1. LightGBM KFold Out-of-fold (RMSE: 2.1940)
2. LightGBM KFold Out-of-fold + WDN Weighted Ensemble (RMSE: 2.1841)