BOOK RATING PREDICTION
PROJECT REPORT

RecSys Team 04

사용자의 기존 도서 평점 데이터를 기반으로 미래 평점을 예측하는 모델 개발 프로젝트

프로젝트 개요

핵심 목표

RMSE (Root Mean Square Error) 최소화

오차를 제곱하므로 정답과 예측값의 차이가 큰 값이 포함되면 점수가 크게 상승합니다. 전체적으로 크게 벗어나는 값이 적도록 예측하는 것이 핵심입니다.

User Data

+

Book Data

⬇️

Predicted Rating

데이터셋 개요

- **Users:** 68,092명
- **Books:** 149,570권
- **Ratings:** Train 306,795건 / Test 76,699건
- **Images:** 149,523개 (Cover Images)

팀 구성 및 역할 분담

"명확한 역할 분담과 구조화된 협업 프로세스를 통해 프로젝트의 안정성을 확보했습니다."

Project Manager

백승
개발 총괄, 마일스톤 관리

Technical Lead

김지민
기술 검증, 코드 리뷰, VAE

Data Engineering Lead

민유경
스키마 정의, 버전 관리, 피쳐엔지니어링

Data Processing Lead

이소진
전처리, 클러스터링

Model Architecture Lead

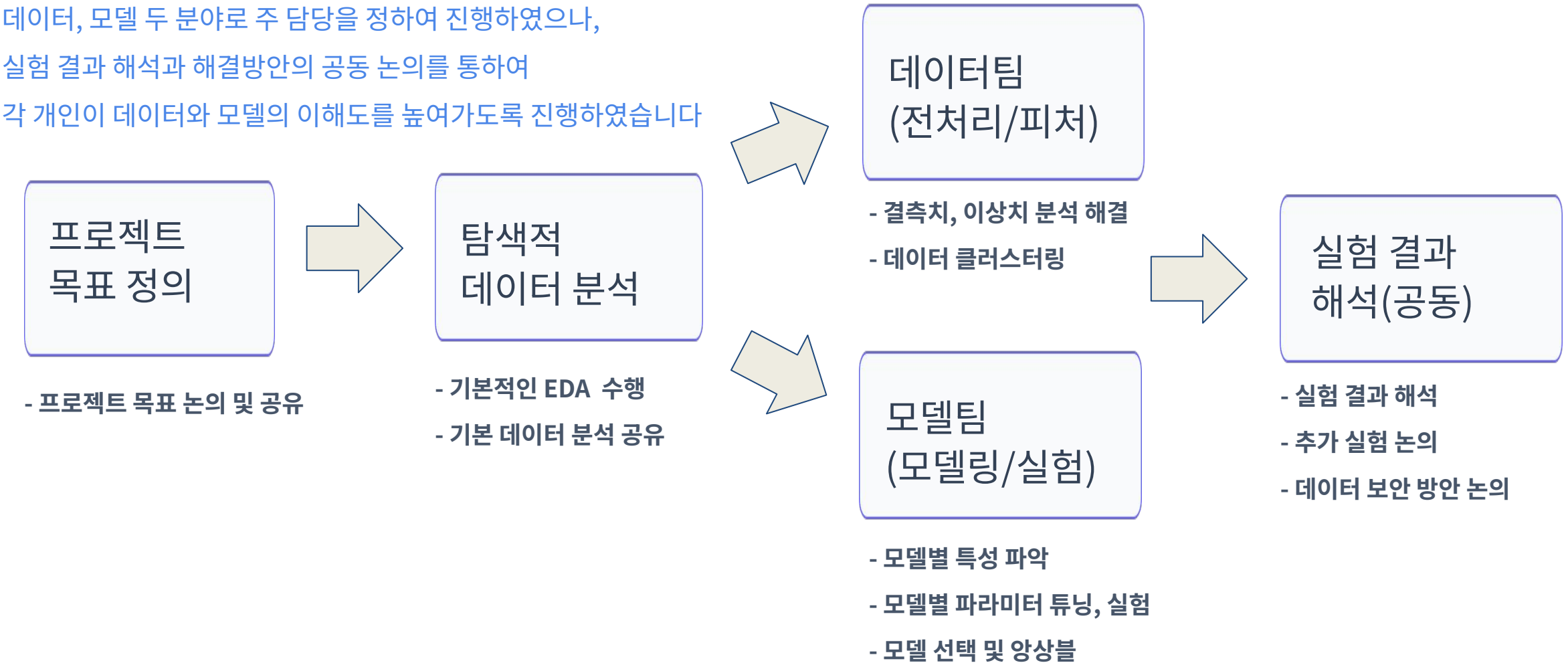
여지호
모델 및 파라미터 정의

Model Analysis Lead

조주익
성능 검증, 신규 실험 제안

프로젝트 수행 절차

데이터, 모델 두 분야로 주 담당을 정하여 진행하였으나,
실험 결과 해석과 해결방안의 공동 논의를 통하여
각 개인이 데이터와 모델의 이해도를 높여가도록 진행하였습니다



협업 방식

실험

Slack 캔버스를 활용한 실험 공유

실험 재현 & 실험 중복 방지를 위해

6	CatBoost	2.1827517	2.1408500	O	2.1547	catboost 패키지 이용, context data 이용, catboost 패키지의 CatBoostRegressor를 이용하기 때문에 wandb 로그 저장에 어려움이 있어서 wandb 는 이용하지 않도록 작성됨(wandb 적용하다가 실패 ㅠㅠ)	@조주익_T8196
7	VAE	5562.058	8663.683	O	4.056	4e-4, dropout 0.4, 29 epoch 돌렸고 갑자기 loss가 꽤 잘 나와서 제출해봤습니다 → VAE는 아쉽지만 폐기하기로 결정	@김지민_T8049
8	CatBoost	2.1534119	2.1356998	O	2.1487	1000 iter으로는 충분히 수렴하지 않아서 5000으로 두고 early stopping 지점까지 돌려보았습니다 → 3563회까지 학습됨 early_stopping_rounds=50, # 50 iteration 동안 개선 없으면 조기 종료	@김지민_T8049
9	FM	1.797	2.193	O	2.2046	전에 데이터 처리쪽에서 문제가 있어서 rmse가 이상했는지 지금 돌리니까 2.2 정도로 잘 작동하네요 FM도 써보면 좋을 것 같습니다 lr: 4e-4 epoch: 5 batch_size:512	@김지민_T8049
10	Text_DeepFM	2.152	2.188	O	2.3202	lr: 4e-4 best epoch: 2 batch_size: 1024	@김지민_T8049

개발

Github를 통한 코드 관리

각자 Branch에서 main으로 병합

코드 & 커밋 컨벤션

코드의 가시성과 일관성 확보로

효율적인 개발

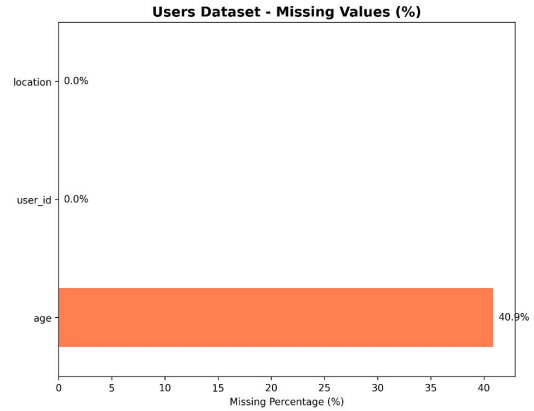
태그	제목 (영어로 요약해 적기)
Feat:	새로운 기능 추가
Fix:	버그 수정
Docs:	문서 수정
Style:	코드 변경이 없는 수정 - 코드 포맷팅, 세미콜론 추가 등
Refactor:	리팩토링
Test:	테스트 코드 추가
Chore:	위 태그에 포함되지 않는 수정사항 (빌드 스크립트 수정, asset 추가/삭제 등)
Design:	CSS 등 UI 변경
Comment:	주석 변경
Init:	초기 설정
Rename:	파일명/폴더명/경로 수정
Remove:	파일/폴더 삭제

본문 작성

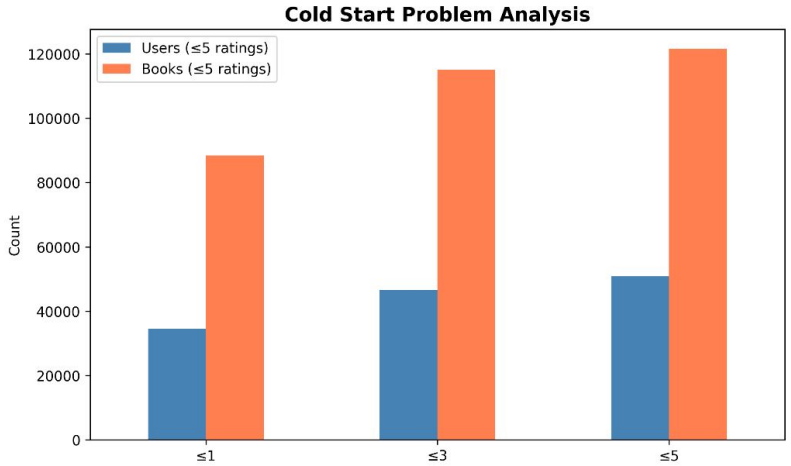
- 최대한 상세히 작성
- '무엇을', '왜' 변경했는지 설명

데이터 분석

결측치가 높은 데이터

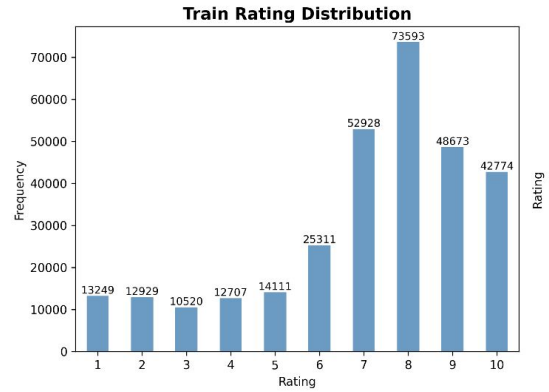


Cold Start 위험이 큰 데이터



rank	Feature	Importance (%)
1	user_id	31.54
2	location_city	6.18
3	title_cluster	5.82
4	location_state	5.09
5	author_norm	4.15

낮은 평점이 극히 드문 Sparse 데이터



➡ 사용자에게 점수를 매기는 성향이 있다는 뜻.

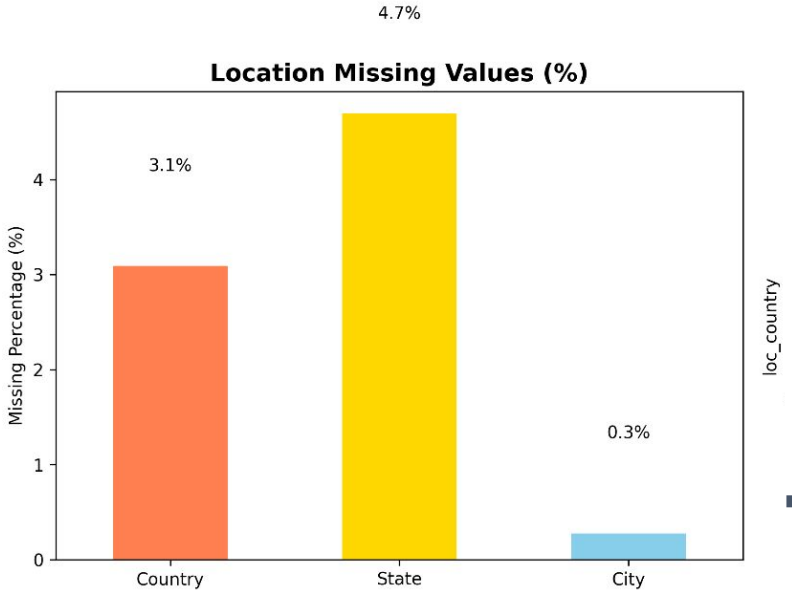
따라서 정보가 적거나 없는 사용자가 나오면

제대로 예측할 수 없음.

데이터 전처리 (1)

User 데이터 처리

Feature	Problem	Solution
User Age	결측치 40.9% (27,833건)	중간값(34세) 대체
Location	비정형 데이터 혼재 (2,658건)	추론 기반 복원



➡ Location 3요소.
결측 비율이 높다.

추론 기반 Location 복원

1. Location을 Country, State, City 로 split
2. 결측치를 추론 기반으로 복원
   a. State 있고 Country 없으면 → 같은 State의 최빈 Country로 채움
   b. City 있고 State 없으면 → 같은 City의 최빈 State로 채움

데이터 전처리 (2)

Book 데이터 처리

Feature	Problem	Solution
Year of Publication	이상치 3건	중간값(1996) 대체
Language	결측치 44.9 % (67,227건)	최빈값(en) 대체
Author	결측치 1건	unknown 대체
Category	결측치 46.3% (68,851건)	단계적 추론

⚠ Insight: Language를 ISBN 기반으로 추론?

가설: ISBN 기반 추론이 최빈값 대체보다 정확할 것이다.

검증: 실제 언어 분포는 정확히 반영했으나,
모델 성능은 오히려 미미하게 하락함.

➡ 데이터의 '정확성'과 모델의 '성능'은 별개의 검증이 필요함
최종적으로 최빈값 대체를 채택.

예상 원인

CatBoost 모델이 Language를 선택하지 않았다

ISBN 안에 이미 Language 정보가 내포되어 있기 때문.

추론 자체는 100% 성공했고,

Language가 평점에 기여하는 정도는

Category의 ⅓ 수준으로 낮지 않다.

	항목	v1	v2
0	처리 방법	최빈값 대체	ISBN 기반 추론
1	총 처리 건수	67,227건	67,227건
2	고유 언어 수	1개	19개
3	최다 언어	en	en
4	최다 언어 비율	100.0%	83.8%

[표] ISBN을 활용한 Language 추론 자체는 유의미했음을 보여줌

데이터 전처리 (3)

Book 데이터 처리 - Category 단계적 추론

Category 필드 결측치 46.03%(68,851건)를 해결하기 위한 **5단계 프로세스**

1

저자/년도 기반

동일 저자/출판 시기 유사성 활용
(32,314건 해결)

2

출판사 기반

출판사별 전문 장르 활용
(30,914건 해결)

3

제목 키워드

제목 내 장르 키워드dict 작성 후 매칭
(1,523건 해결)

4

임베딩 클러스터링

Sentence-BERT 활용
(4,100건 해결)

5

최종 처리

잔여 결측치 'unknown'
(100% 해결)

클러스터링: Noise to Signal

문제 정의

High Cardinality: 텍스트 기반 피처(Category, Title, Summary)는 고유값이 너무 많아 모델이 ID처럼 인식하는 노이즈(Noise)로 작용할 위험이 있음.

해결 전략

의미 기반 클러스터링: 각 피처를 임베딩하여 의미적으로 유사한 그룹으로 묶는 파생 피처 생성.

- ✓ 표기 변형 통합 (sci-fi vs science fiction)
- ✓ 통계적 안정성 및 일반화 성능 향상

Before (Noise)

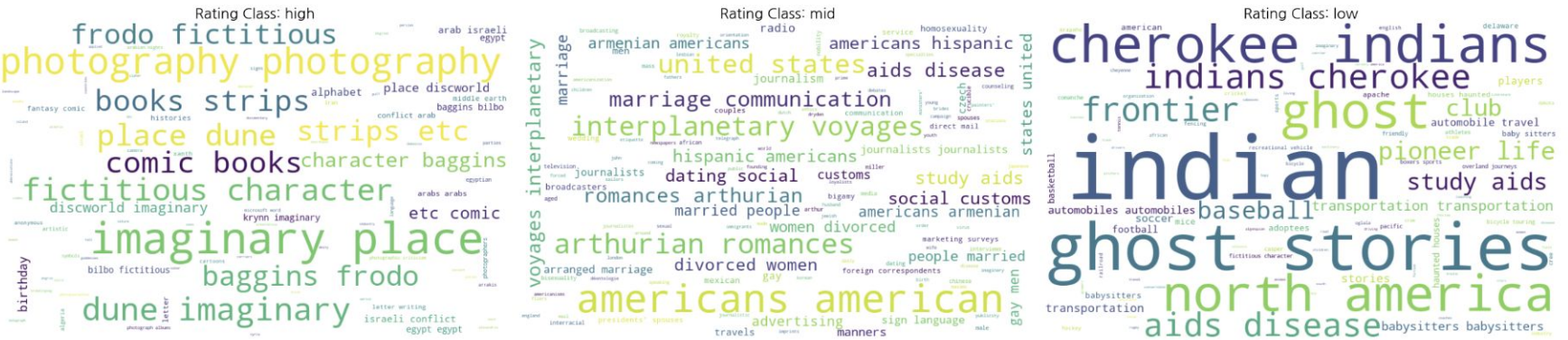
sci-fi | Science Fiction | Space Opera | Future War | Cyberpunk

↓

After (Signal)

Cluster: Science Fiction & Fantasy

↓



문제 정의

주요 Feature Importance 결과 (상위 일부)
(지금까지의 처리,엔지니어링 기반으로 catboost 돌린결과)

- user_id
- location_city, location_state
- publisher, book_author
- summary_cluster, category_cluster
- age_of_book, age_group

특히, author 와 publisher 관련 피쳐가 상위권을 차지하여, 이 두 요소에 대한 추가적인 파생 피쳐 설계가 필요하다고 판단.

피쳐 추가

가설

1.많은 책을 출간한 작가는 독자층이 확실하고 →평균평점도 안정적 일 것

2.데이터가 많은 (리뷰가 많은) or 책이 많은 대형 출판사일수록 → 출판하는 책 퀄리티가 높고 →rating 관련 패턴 있지 않을까?

3. 책이 최신일수록 rating 패턴이 옛날 책과 다를 것 -> 따라서, 모델입장에서 언제 나왔는가 보다 얼마나 오래된 책인가 에 대한 직관적 신호를 줄 수 있게 함.

- author_review_count
- author_book_count
- publisher_review_count
- publisher_book_count
  +) long_tail 분포이므로, 이 count 피쳐에 대해,
  log 와 bin 한 버전의 피쳐도 추가함
- age_of_book.

성능 개선 (catboost test loss 기준)

2.1442 -> 2.1394

feature_importance 결과

“의미·빈도 기반 파생 피쳐가 원본 식별자보다 높은 예측력을 보이며, Feature Engineering의 효과를 입증함”

모델 개요

협업 필터링 기반 잠재 요인 모델(Latent Factor Model)

활용 데이터 : User ID, Item ID, Ratings(Interaction)

활용 모델 : NCF(Neural CF), MF(Matrix Factorization), VAE

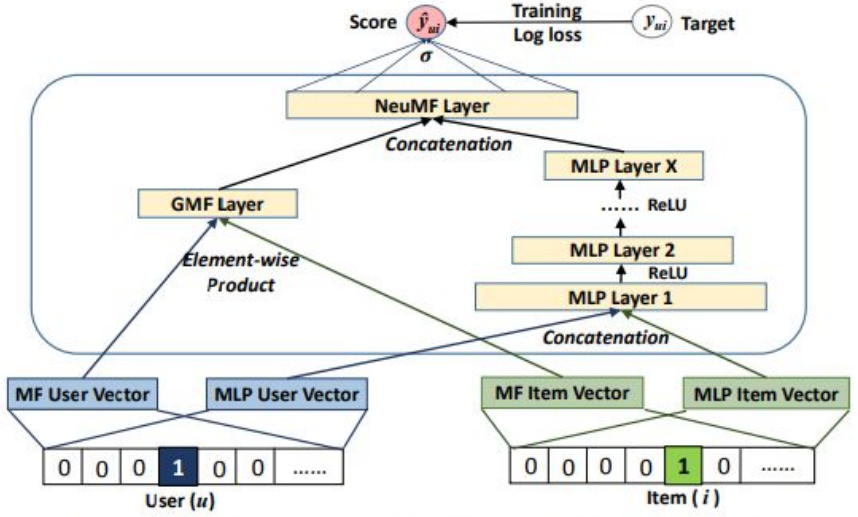


Figure 3: Neural matrix factorization model

MF

NCF

User/Item **고유성** 학습

User/Item ID를 같은 크기의 고차원 벡터 공간으로 **임베딩**

각각의 User와 Item 마다의 **고유성** 부여

행렬 분해(MF) 및 비선형 결합(MLP)

MF: 임베딩 벡터간의 **내적**으로 **선형**적인 상호작용 계산

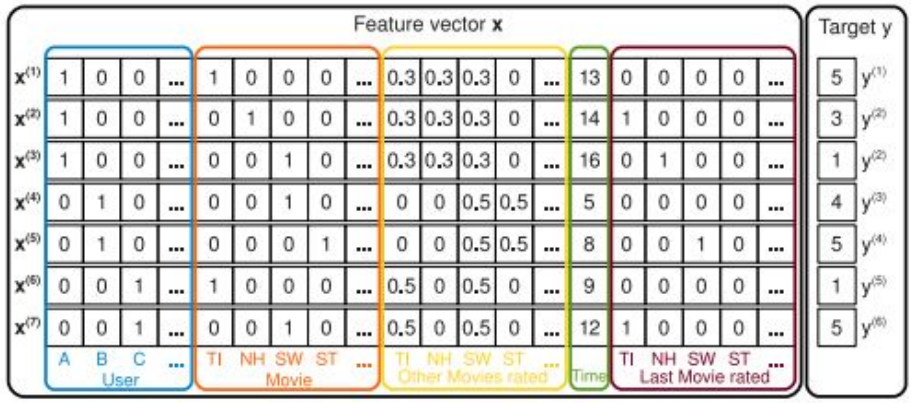
MLP: 임베딩 벡터간의 **비선형**적인 상호작용 계산

모델 개요

컨텍스트 기반 특성 상호작용 모델(Factorization Machine)

활용 데이터 : Interaction Data, User/Item Meta Data(Context Data)

활용 모델 : FM, DeepFM, Image_DeepFM



$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

Glob_Bias Feat_Bias Feature Interaction

다양한 요인들에 대한 학습

Context 정보를 Feature, Rating을 Label로 활용

Context를 **임베딩**하여 **MLP** 혹은 **내적**으로 결합가능

딥러닝으로 Feature간 상호작용 정도 계산가능

확장성

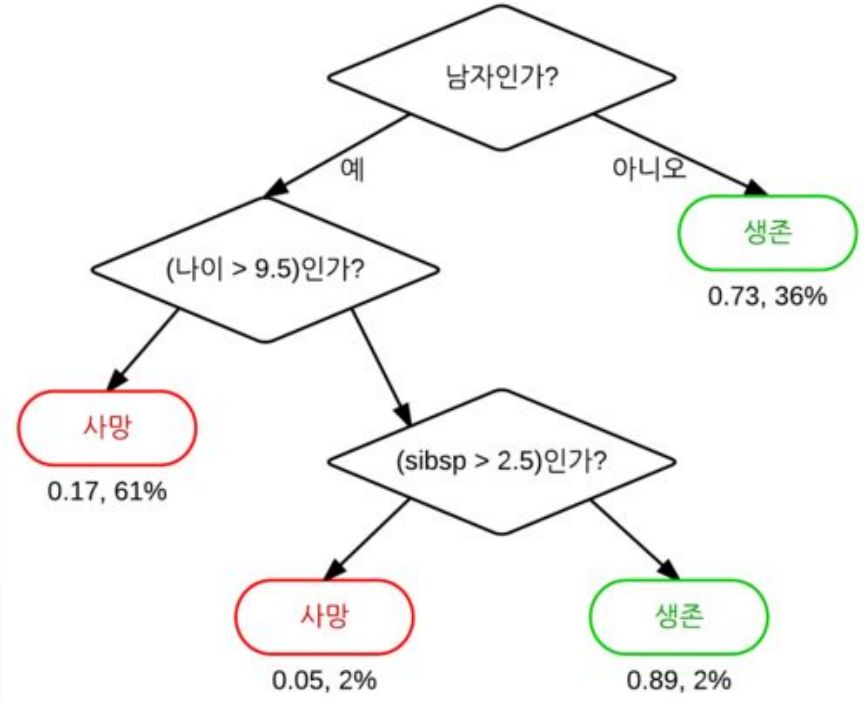
Context 정보를 추가하여 Feature를 늘릴 수 있음

모델 개요

의사 결정 트리 모델(Decision Tree Model)

활용 데이터 : User ID, Item ID, Context Data

활용 모델 : CatBoost, XGBoost



출처 : Wikipedia

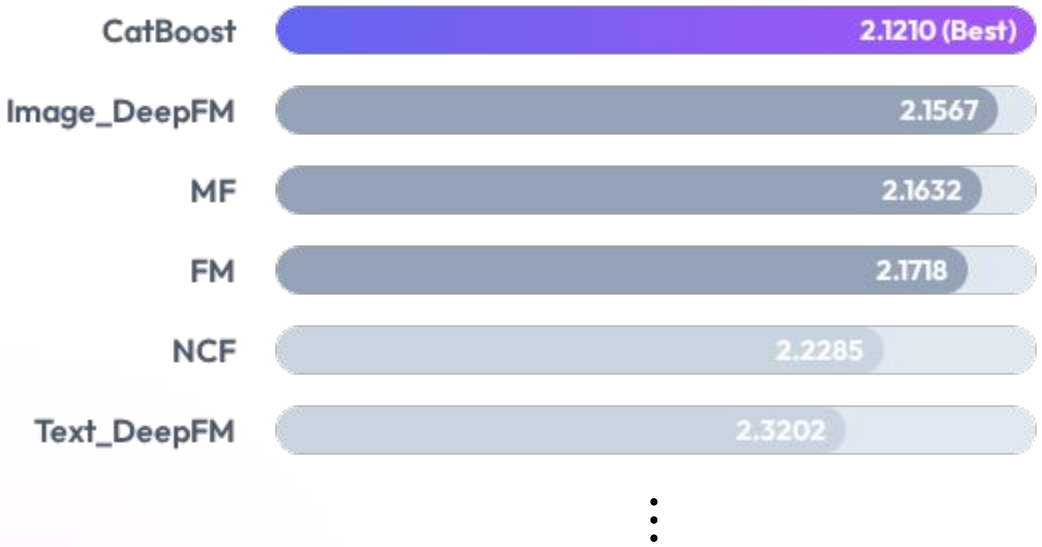
단순하지만 좋은 성능

딥러닝 모델이 아니지만 여전히 정형 데이터 분석의 정점!

Categorical Data 처리

다른 모델들은 카테고리 별로 차원을 늘려야 하는 문제가 있지만 결정 트리 모델은 모델 내부에서 처리가 가능

모델 성능 검증: 단일 모델 경쟁



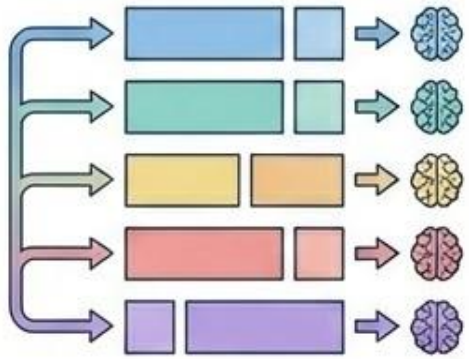
모델 선정 결과

다양한 모델 아키텍처(협업 필터링, 컨텍스트 기반, 트리 기반)를 실험한 결과, **CatBoost**가 Valid Loss와 Public Score 모두에서 가장 우수한 성능을 기록했습니다.

Note: CatBoost는 특히 범주형 변수 처리에 특화되어 있어, 정제된 메타데이터의 효과를 가장 잘 반영했습니다.

모델 평가 및 개선: K-fold Averaging

K-Fold Cross Validation

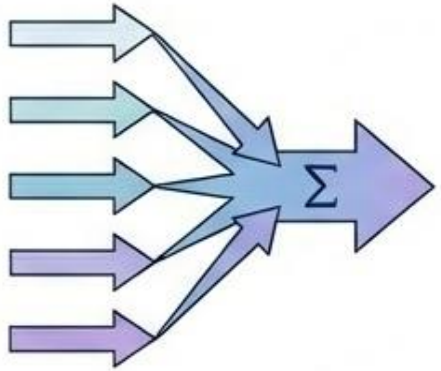


전체 학습 데이터를 5개의 Fold로
분할하여, 각기 다른 데이터셋으로
K번의 독립적인 학습 및 검증을 수행

과적합 방지 (Robustness)

단일 Validataion Set에 모델이
과적합되는 것을 방지하고,
데이터의 편향(Bias)을 최소화하여
일반화 성능을 확보

Ensemble Averaging



최종적으로 5개 모델이 예측한 평점
(Rating)을 산술 평균하여
분산을 줄이고 예측 안정성을 높임

최종 결과: 최적화와 직관의 시너지

SLSQP 최적화 방식

수학적 최적화를 통해 가중치 산출 시도.

- CatBoost (0.7096) + MF (0.2904)
- Image_DeepFM에 0%의 가중치 부여
  ㄴ 데이터상으로는 합치지 않는게 낫다고 판단

Public Score: 2.1172
Private Score: 2.1108

경험적 Soft Voting (Final)

도메인 지식과 직관을 바탕으로 이질적 모델 결합.

- CatBoost (0.8) + Image_DeepFM (0.1) + MF (0.1)
- 서로 다른 특성(이미지, 잠재요인)의 시너지 유도

Public Score: 2.1164 (Winner)
Private Score: 2.1097 (Winner)

"수학적 최적화는 강력하지만, 서로 다른 특성을 포착하는 모델들이 결합될 때 최고의 성능을 냈습니다."

팀 자체 평가

프로젝트를 통해 배운 점 또는 시사점

- **도메인 지식:** 추천 시스템에서 기술적 깊이만큼이나 중요한 것은 데이터의 특성에 대한 이해였습니다.
- **목표 공유:** 초기에 배경과 목표를 명확히 한 것이 흔들리지 않는 기준점이 되었습니다.
- **계획과 유연성의 조화 :** 완전한 통제하의 실험은 어려웠지만, 실험 후 결과를 공동으로 해석하는 과정에 집중하여 인사이트를 얻을 수 있었습니다.

시도 했으나 잘 되지 않았던 것들

- VAE (Variational Auto Encoder)
- Rounding 후처리
- Stacking Ensemble
- Cold Start 대응 전략