머신러닝 예측 모델을 통한

애니메이션 추천 시스템 개발

Codestates
[AIB] 9기 안나

Contents

애니메이션 추천 Database

데이터 선정 이유

- 콘텐츠 사업에서 작품 퀄리티 만큼 중요한 '**추천 서비스**'
- 모든 커머셜 사이트에서도 고객 정보를 활용한 맞춤형 추천 서비스는 매출에 직결됨

- 최신 트렌드에 발맞춰 머신 러닝을 활용하여 **유저의 추천여부를 예측하고,** 간단한 **작품 추천 시스템** 개발

데이터 구성 설명

myanimelist.net 추출 데이터

애니메이션과 **사용 유저 정보** 구분

1. Anime : 12,294 records (7 특성)

- Anime_id, name, genre, type, episodes, rating, members

2. Rating : 7,813,737 records (3 특성)

- user_id, anime_id, rating

→ 데이터 **변환** 및 **맵핑**하여 학습에 사용 할 최종 데이터세트를 생성

애니메이션 추천 Database

문제 정의

- 어떤 애니메이션 평가가 높을까?
- 어떤 애니메이션을 추천할까?

→ 매우 복잡한 문제

→ **특정 애니메이션을 유저가 추천할지, 말지 (binary question)** 검증

검증 방법

- **지도학습**(supervised machine learning)
- **분류** (추천, 비추천) 문제로 접근
- 작품간 유사도를 이용하여 작품 추천

가설 설정

1. 추천 여부 (특성 추가)
   - **target: 유저의 추천 여부**
   - [평점 > 기존 평점 평균] 일 경우

2. 추천에 영향을 미치는 특성들 검증
   - 애니메이션 **평점**이 높을 수록 추천
   - **커뮤니티** 인원 많을 수록 추천
   - 인기 있는 **장르, 타입**일수록 추천
   - **평소 기대치**가 낮은 유저일수록 추천
   - **평가를 많이 남기는** 유저일수록 추천

데이터 전처리

Anime
(12,294개)

anime_id	name	genre	type	episodes	rating	members
32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1	9.37	200630

- 결측치, 중복데이터 제거
- 데이터 타입 정리
- 동일 항목은 하나만 사용

Rating
(7,813,737개)

user_id	anime_id	rating
42653	16498	8.0

- 장르 구분 (1,2,3구분)
  → 종류 갯수: 40, 43, 42
  → 구분1만 사용

사이트 작품평점

유저 정보

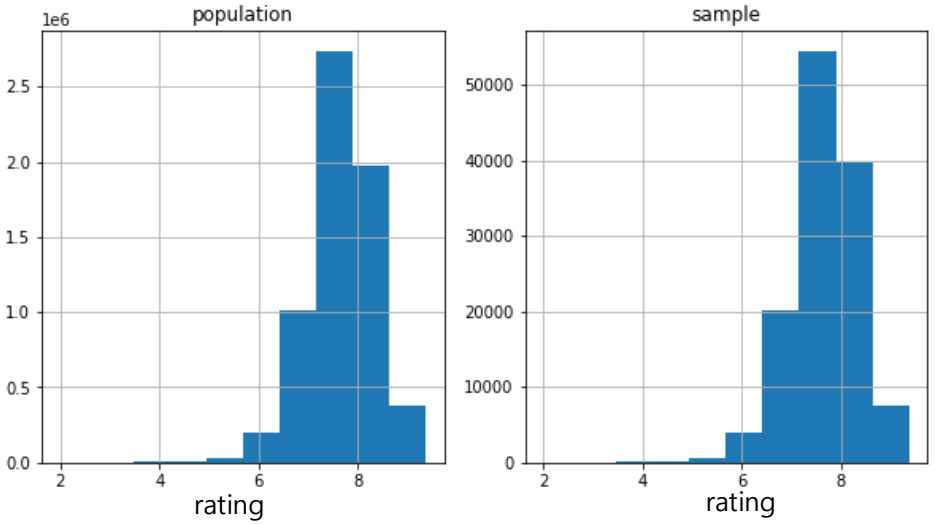
최종
(6,337,145개)

anime_id	genre	type	episodes	rating	members	user_id	Rating_user	Rating_mean	user_exp	user_review	recommendation
22145	Comedy	TV	10	8.37	122895	1031	8.0	7.696581	4	234	True

타겟 분포
추천비율: 53%

- anime 데이터에 rating 데이터를 합치고, 유저별 평가 정보를 통해 추천여부를 추가한다.
- **recommendation** : 개별 평점 > 기존 평점평균 (target)
- **user_exp** : user_mean 을 등급화, 유저의 기대치가 어떤지 (기대가 낮으면 평가가 높다)
- **user_review** : user가 남긴 평가 갯수 추가 (평가를 많이 남겼는지, 아닌지)

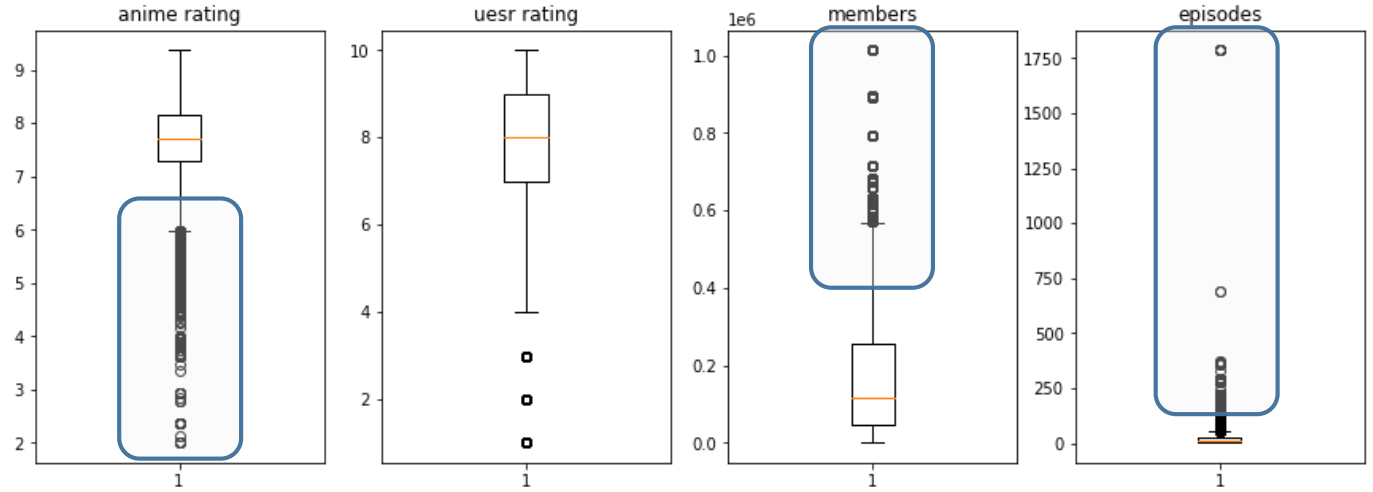
샘플링(Sampling)



- 최종 데이터 셋의 크기가 매우 큼 (6,337,145개)
- 다양한 방식으로 샘플링이 가능하지만 모집단 특성 반영하는 **랜덤샘플링**으로 추출

- 평점 및 타겟 분포에서 모집단을 대표하는 샘플을 추출하여 사용 (2%, **12만개**)

극단치(Outlier) 확인

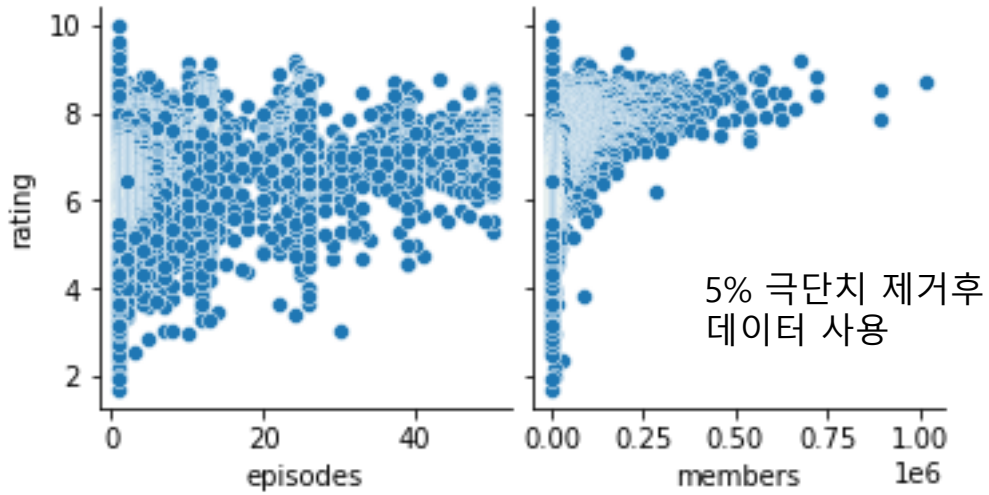


- 데이터 분포가 평균에서 많이 벗어나는 극단치가 각 항목별로 5% 미만에 포함됨

- 머신러닝 학습을 위해 해당 데이터 제거 후 진행해도 성능개선에 큰 영향이 없어 **이후 데이터 해석의 편의를 위해 처음 데이터를 그대로 사용**

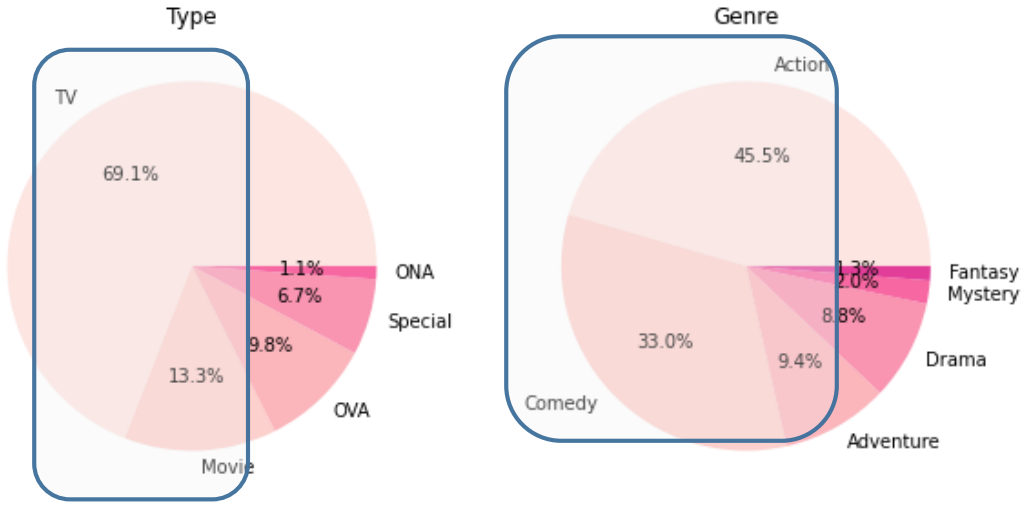
탐색적 데이터 분석 (EDA)

1) 평점과 데이터 상관관계



- 제작편수가 많으면, 해당 애니메이션 그룹의 인원이 많은 편이고, 평가도 높은 경향성을 보인다.
- 수치형 데이터와 (episodes, members) 강한 **선형적 상관관계는 성립하지 않는다**

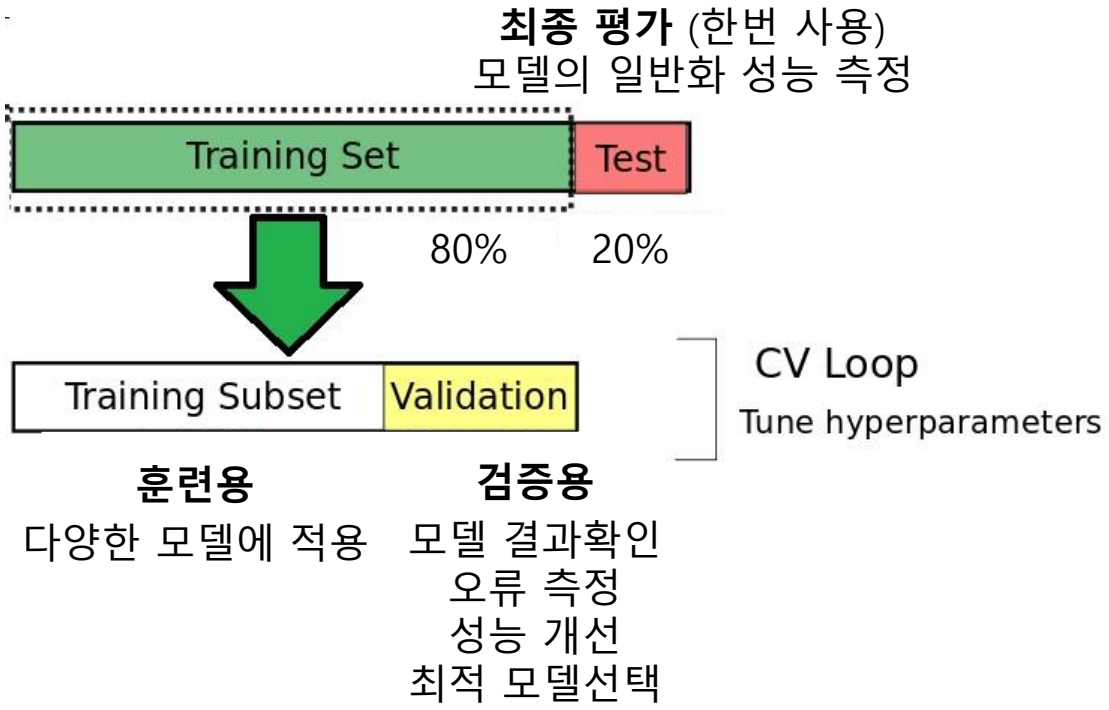
2) 가장 인기가 많은 타입 및 장르



- **TV**시리즈가 전체의 69%를 차지한다.
- **액션, 코미디**가 전체의 78%를 넘게 차지한다.
- 구분인자가 많지만 특정 1~2개 항목에 **편중**되어 있어 작품의 특성을 구분하는 주요 인자가 되기 어려워 보임

머신러닝

1) 훈련/ 검증 SET 나누기



2) 기준 모델

- 가장 간단하면서도 직관적
- 최소한의 성능을 나타내는 기준
- **분류**: 타겟의 **최빈 클래스** 사용

target	최빈값	정확도
True	True	O
True	True	O
False	True	X
False	True	X
True	True	O

- training accuracy: **0.533** (기준)

머신러닝 모델간 성능 비교

앙상블 모델 (여러 트리 사용)

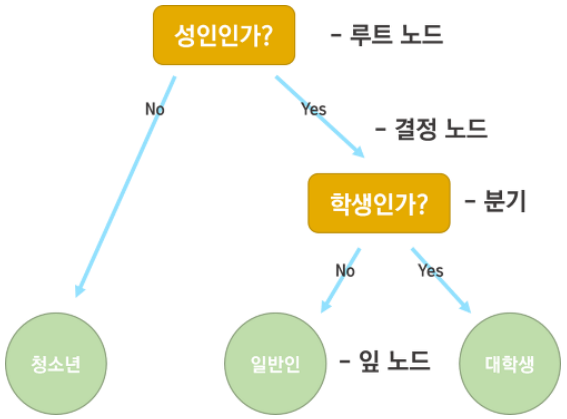
1) 기준 모델

target	최빈값	정확도
True	True	O
True	True	O
False	True	X
False	True	X
True	True	O

- 훈련 정확도 : 0. 53
- 검증 정확도 :

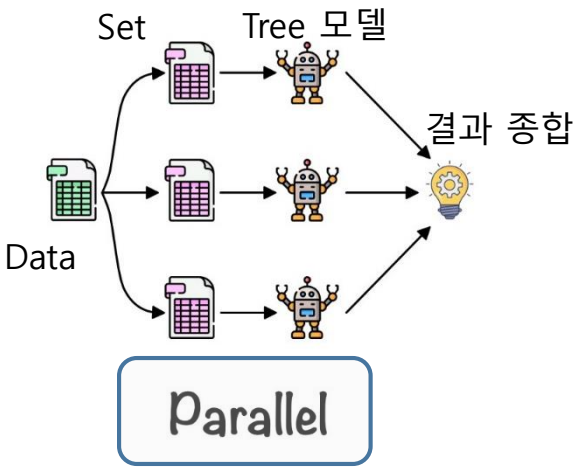
- F1 Score :
- 모델 튜닝 후 F1 :

2) 결정 트리



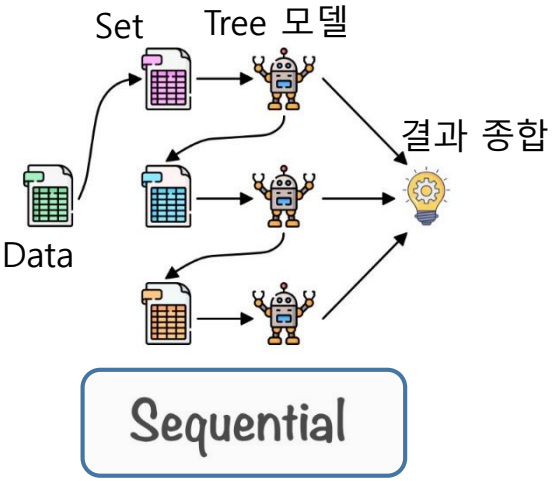
1.0
0.99

3) 랜덤포레스트



1.0
0.99

4) 부스팅



Data leakage 발견

1.0
0.99

머신러닝 모델간 성능 비교

앙상블 모델 (여러 트리 사용)

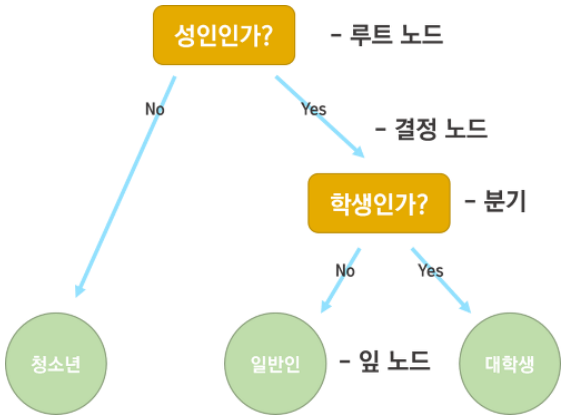
1) 기준 모델

target	최빈값	정확도
True	True	O
True	True	O
False	True	X
False	True	X
True	True	O

- 훈련 정확도 : 0. 53
- 검증 정확도 :

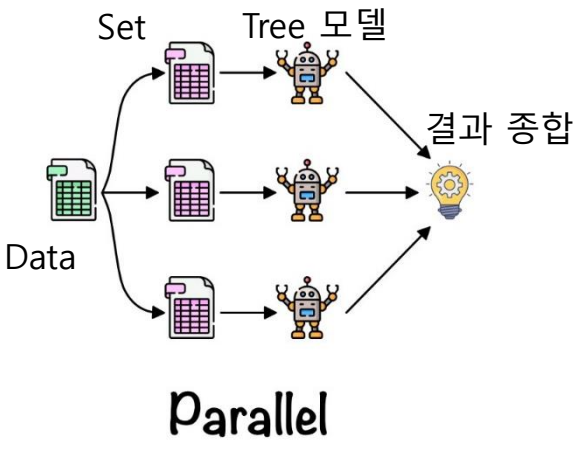
- F1 Score :
- 모델 튜닝 후 F1 :

2) 결정 트리



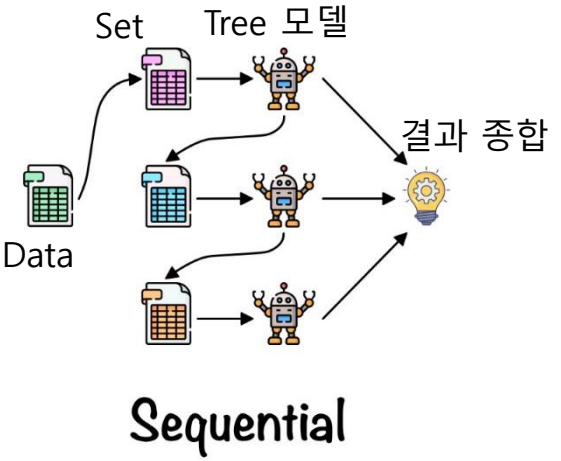
1.0
0.58
0.61

3) 랜덤포레스트



1.0
0.64
0.67
0.72

4) 부스팅



0.67
0.67
0.68
0.69

랜덤 포레스트 모델 평가지표 해석

1) 사용 데이터 및 성능측정

- **랜덤 포레스트** 모델 선정
- 최적인 **하이퍼파라미터** 적용

- 훈련용+검증용 학습에 사용 (10만개)
- 테스트용 데이터로 최종검증 (2만개)

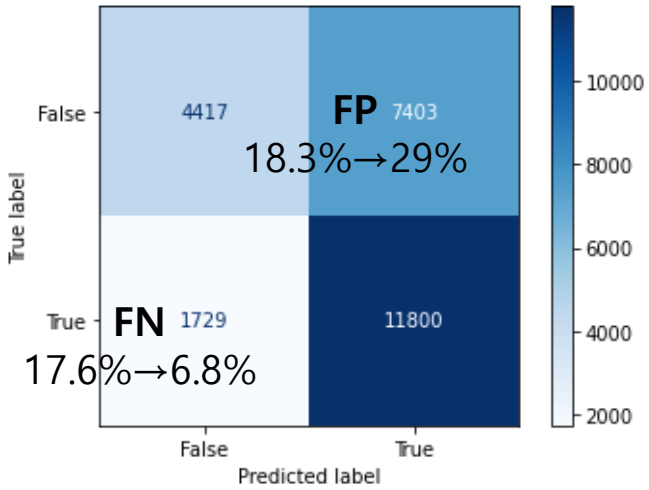
- 훈련 정확도: 0.65
- F1 score**: 0.72**

2) 평가지표

[Confusion matrix]

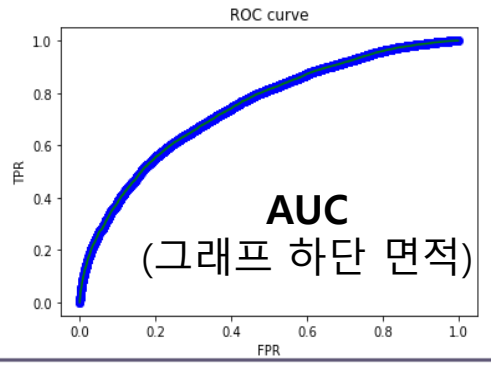
- 추천에 가중치를 주어 모델 성능을 개선
- 오류의 종류가 변함
- 추가적으로 FP를 낮추는 노력이 필요하다

예측 값들을 분류 (정답/오답)



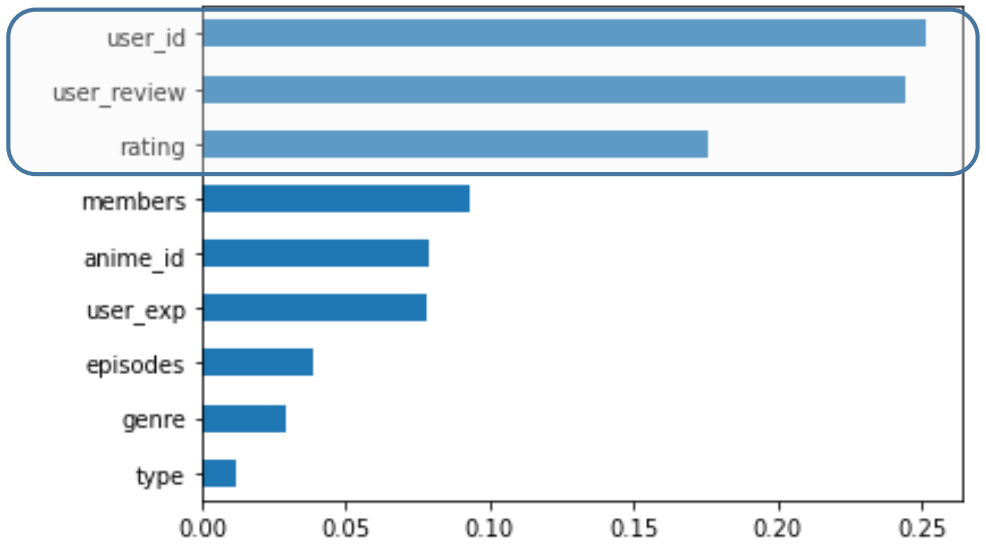
[ROC curve, AUC Score]

- 1에 가까울 수록 모델 성능이 좋음
- AUC: 0.68 → 0.75로 개선



랜덤 포레스트 모델 평가지표 해석

3) 특성 중요도 (Feature importance)



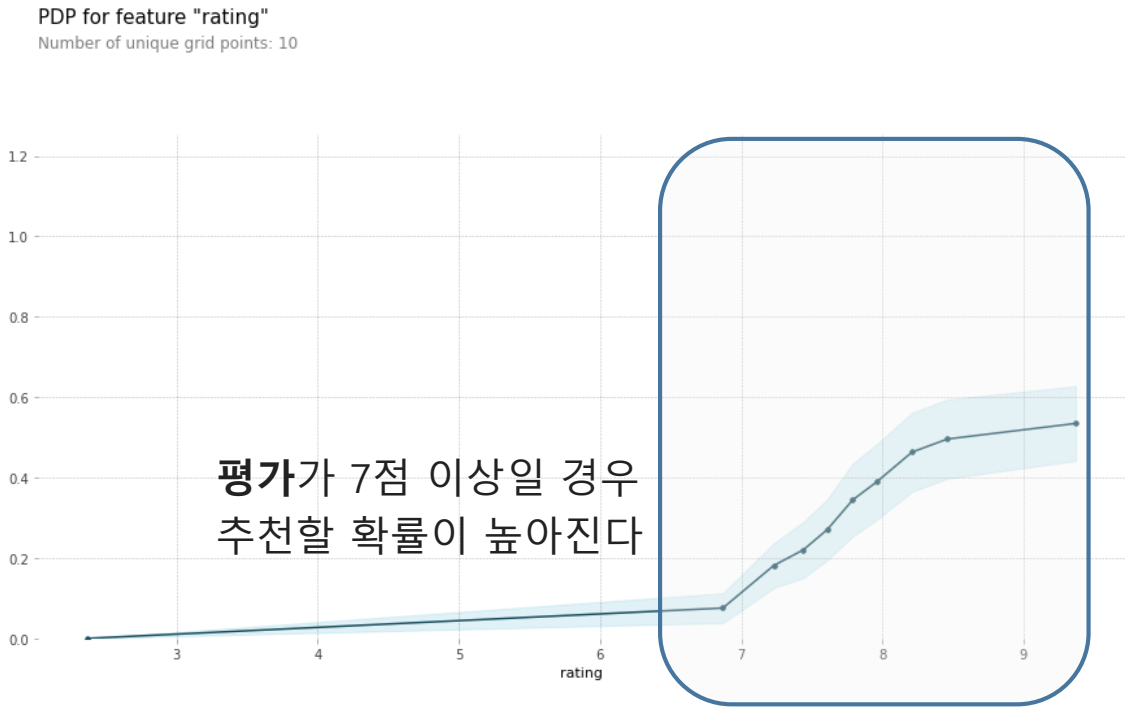
4) 순열 중요도 (Permutation importance)

Weight	Feature
0.0834 ± 0.0026	rating
0.0080 ± 0.0012	user_exp
0.0013 ± 0.0011	user_review
-0.0000 ± 0.0002	type
-0.0001 ± 0.0008	anime_id
-0.0002 ± 0.0007	user_id
-0.0006 ± 0.0007	genre
-0.0012 ± 0.0005	episodes
-0.0018 ± 0.0010	members

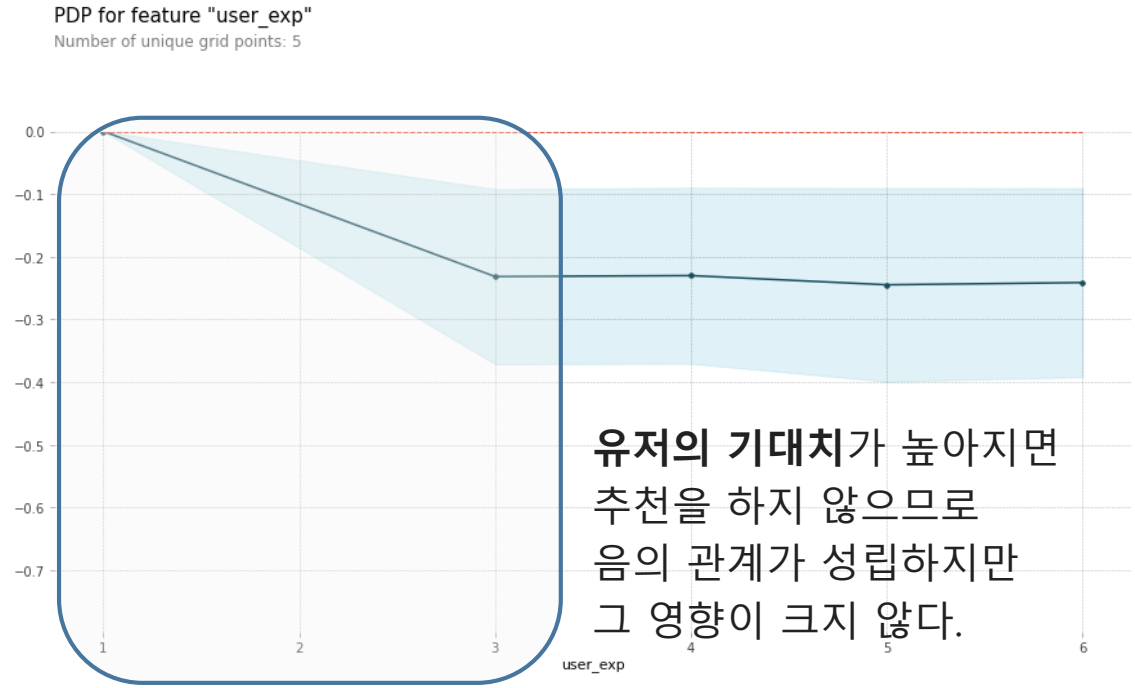
- **특성 중요도**: 해당 특성이 트리가 나뉠 때 얼마나 자주, 일찍 사용되는가에 따라 중요한 정도를 나타낸다
- **순열 중요도**: 관심 있는 특성에만 무작위로 노이즈를 주고 전/후 성능이 감소한 수치를 측정한다
- 특성의 **복잡도**(구분이 매우 다양하다)가 높으면 분기에 많이 사용되어 특성중요도가 높게 측정될 수 있다

PDP (Partial Dependence Plot)

- 트리 모델이 여러 개 합쳐져 만들어진 랜덤포레스트 모델은 각 트리마다 사용된 특성이 달라 이해가 어렵다
- PDP를 통해 **개별 특성**이 **타겟**에 어떻게 작용하는지 알아볼 수 있다



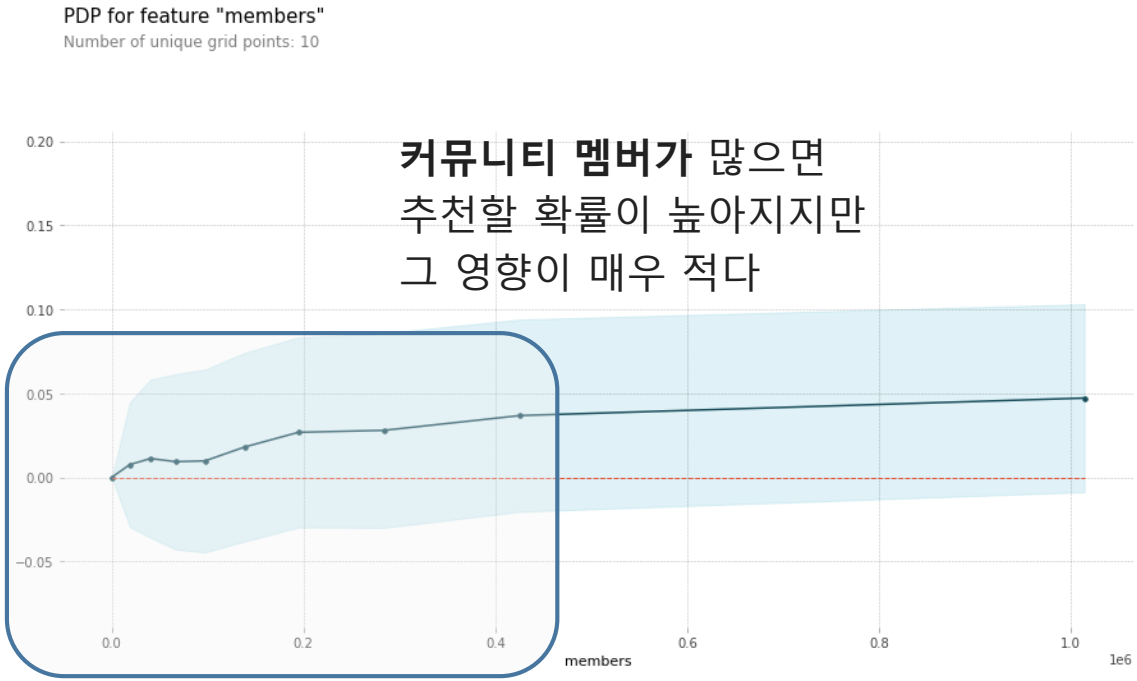
평가가 7점 이상일 경우
추천할 확률이 높아진다



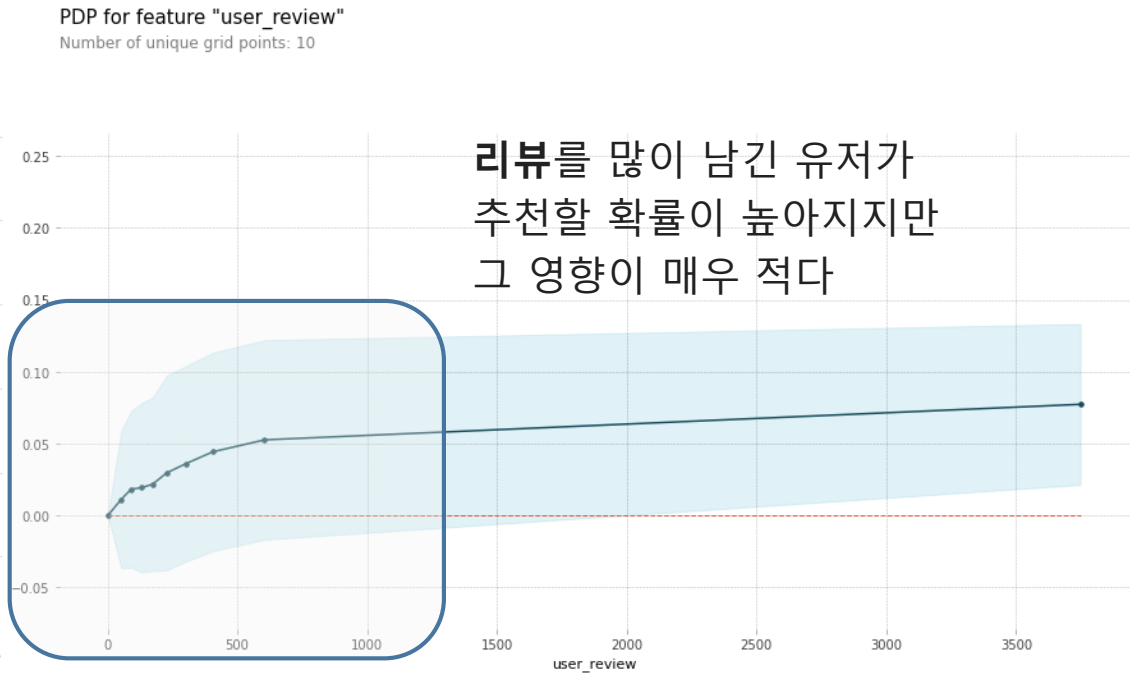
유저의 기대치가 높아지면
추천을 하지 않으므로
음의 관계가 성립하지만
그 영향이 크지 않다.

PDP (Partial Dependence Plot)

- 트리 모델이 여러 개 합쳐져 만들어진 랜덤포레스트 모델은 각 트리마다 사용된 특성이 달라 이해가 어렵다
- PDP를 통해 **개별 특성**이 **타겟**에 어떻게 작용하는지 알아볼 수 있다



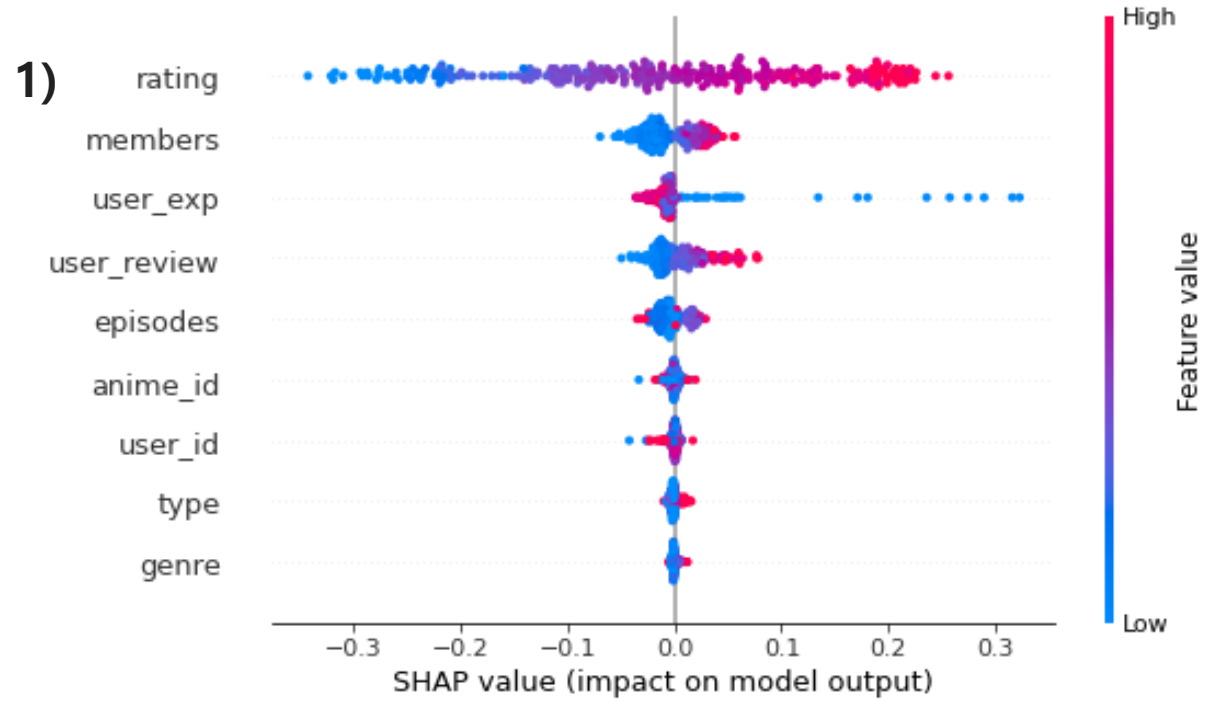
커뮤니티 멤버가 많으면
추천할 확률이 높아지지만
그 영향이 매우 적다



리뷰를 많이 남긴 유저가
추천할 확률이 높아지지만
그 영향이 매우 적다

SHAP (Shapley Additive explanation)

1)



2)

- **개별 관측치**에 대한 설명을 도와줌
- 게임이론을 바탕으로, Game 에서 각 Player 의 기여분을 계산하는 방법
- 계산된 SHAP Value가 높으면 해당모델에 특성의 영향도가 높다고 판단

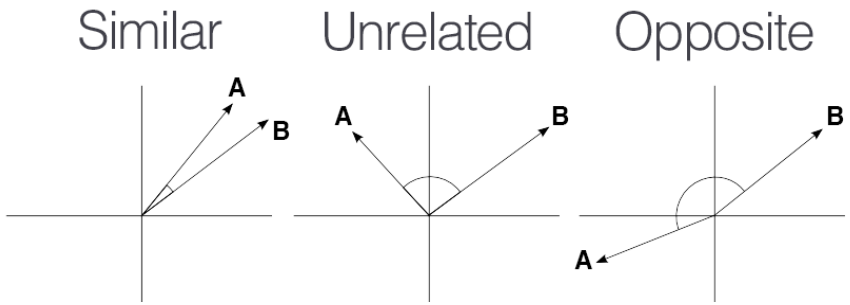
1) **특정 수의 샘플**을 뽑아 해당 특성의 영향을 구할 수 있다 (300개 샘플분석 결과)

2) **개별 샘플**에 어떤 영향이 있는지도 구체적으로 예측할 수 있다. 예를 들어 anime_id와 user_id를 특정하여 추천여부를 확인하고, 그 결과에 영향을 준 특성들을 확인 할 수 있다.

애니메이션 추천 시스템 개발

- 가장 주요한 특성인 **평점**과 **유저**의 관계를 이용하여, **유사도**를 판단하여 작품추천 시스템을 만든다
- 특정 애니메이션을 봤을 때, 해당 작품과 유사도가 높은 다른 작품을 추천한다

유사도 측정 (Cosine Similarity)



- 두 정보(평점, 유저)를 벡터로 표현
- 두 벡터 간 코사인 각도를 이용하여 유사도 구함
- 코사인 유사도: -1 이상 1 이하의 값을 가지며 값이 1에 가까울수록 유사도가 높다고 판단

애니메이션 추천 시스템

- 평가 수가 적은 것은 제외 (상위 10% 사용)
- Sklearn (cosine_similarity) 사용
- 시청한 애니메이션 이름을 넣으면, 유사도가 높은 작품 5개를 추천한다

```
[150] anime_recommendation('Death Note')

0초

Recommended because you watched Death Note:

#1: Code Geass: Hangyaku no Lelouch R2, 30.62% match
#2: Code Geass: Hangyaku no Lelouch, 29.7% match
#3: Fullmetal Alchemist: Brotherhood, 27.16% match
#4: Steins;Gate, 23.84% match
#5: Shingeki no Kyojin, 22.17% match
```

한계점 및 추후 발전방향

한계점

- 추천은 다양한 방식으로 가능하고, 절대적인 기준이 없기 때문에 개인에 맞춰 본인이 했던 평균 평점보다 새로운 작품 평점이 높으면 추천하는 모델링을 구상했다.
- 결과를 보면 결국 작품의 **전체 평점**의 영향이 가장 크다. 평점은 많은 정보를 함축하지만, 그저 공공의 인기를 반영하기 보다는 **개인화된 추천**을 위해 다른 특성들을 더 활용할 수 있는 방안에 대한 고민이 필요하다.

추후 발전 방향

- 추천을 예측하는 모델이 **개인화**가 잘 되었는지 더 검증해보면 유용할 것 같다.
  - ✓ 유저별 평점 분포를 살펴보고, 평균이 아니라 **상위 특정% 위에 속할 경우 추천**하는 것으로 **비교 기준을 더 강화**할 수 있다.
  - ✓ 유저가 기존에 평가한 작품수가 많지 않을 경우, 강화된 기준을 넘어 추천하는 작품 수가 적을 수 있다. 분석결과를 신뢰할 수 있는 데이터 양인지도 검증해야 한다.

Thank you for your attention!