도서 추천 알고리즘

AI 17 정미정

목차

프로젝트 목표

어디서?

"밀리의 서재" 기술개발팀

무엇을?

평점 도입과 함께 취향에 맞는
도서 추천

왜?

기존 사용자들의 이탈을 막고,
신규 사용자의 유입을 기대

데이터 셋 소개

캐글 데이터셋:Book Recommendation Dataset

User.csv

사용자의 정보를 담고 있는 데이터셋

- User-ID
- Location
- Age

Books.csv

책 정보를 담고 있는 데이터셋

- ISBN: 도서번호
- Book-Title
- Book-Author
- Year-of-Publication
- Publisher
- Image-URL-S
- Image-URL-M
- Image-URL-L

Ratings.csv

사용자들이 책에 매긴 점수를 담고 있는 데이터셋

- User-ID
- ISBN
- Book-Rating

머신러닝 문제 정의

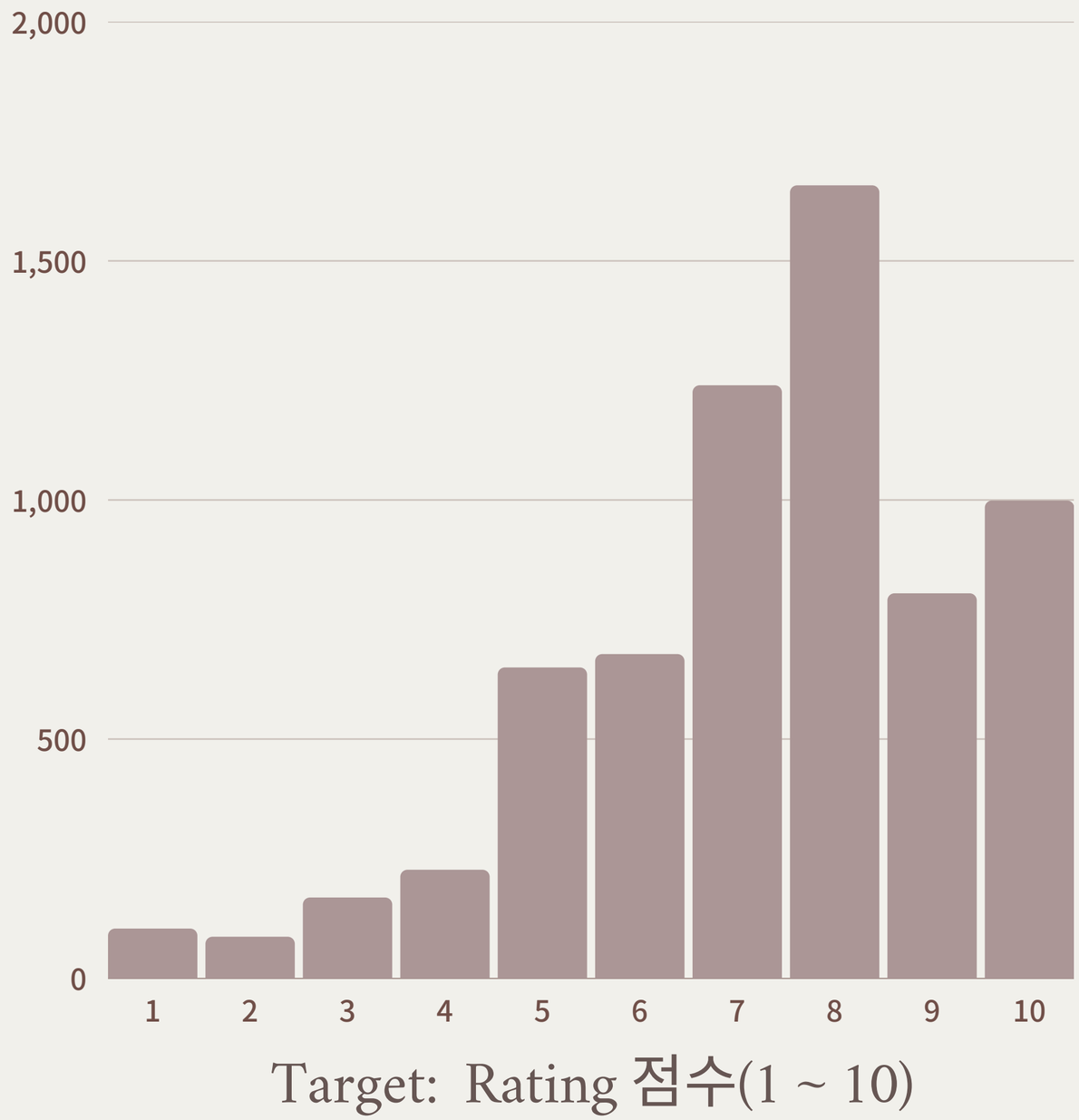
한 명의 사용자에 대한 도서 평점을 예측할 수 있다.

데이터 셋

- 가장 많은 평점을 남긴 한 개의 User-ID에 관한 데이터

목표

- 도서 별 특징에 따른 평점을 예측할 수 있다.
- 다중 분류 알고리즘을 활용할 수 있다.



데이터 EDA 과정

데이터 병합 과정

- Users.csv + Ratings.csv -> User-ID 기준

```
origin = pd.merge(users, ratings, on='User-ID', how='right')
```

- 위의 데이터셋 + Books.csv -> ISBN 기준

```
origin = pd.merge(origin, books, on='ISBN', how='left')
```

데이터 EDA 과정

feature engineering

- Location 컬럼 = 나라 이름 + 주 이름 + 도시 이름

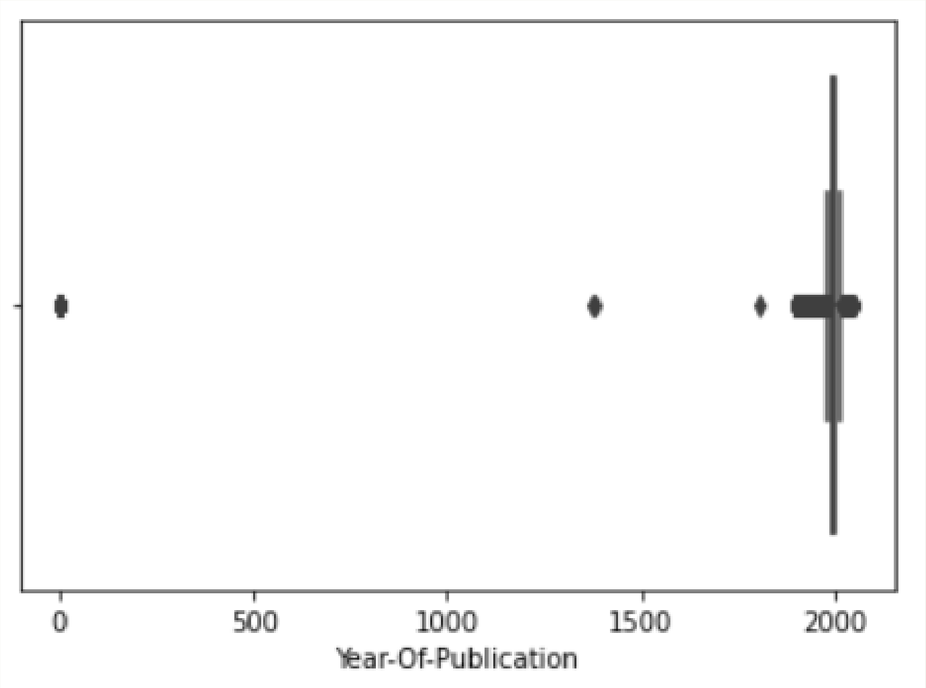
=> 각각 country, state, city로 분리

```
# feature engineering
location = df['Location'].str.split(', ', n=2, expand=True)
location.columns=['city', 'state', 'country']

df.loc[:, 'city'] = location['city']
df.loc[:, 'state'] = location['state']
df.loc[:, 'country'] = location['country']
```

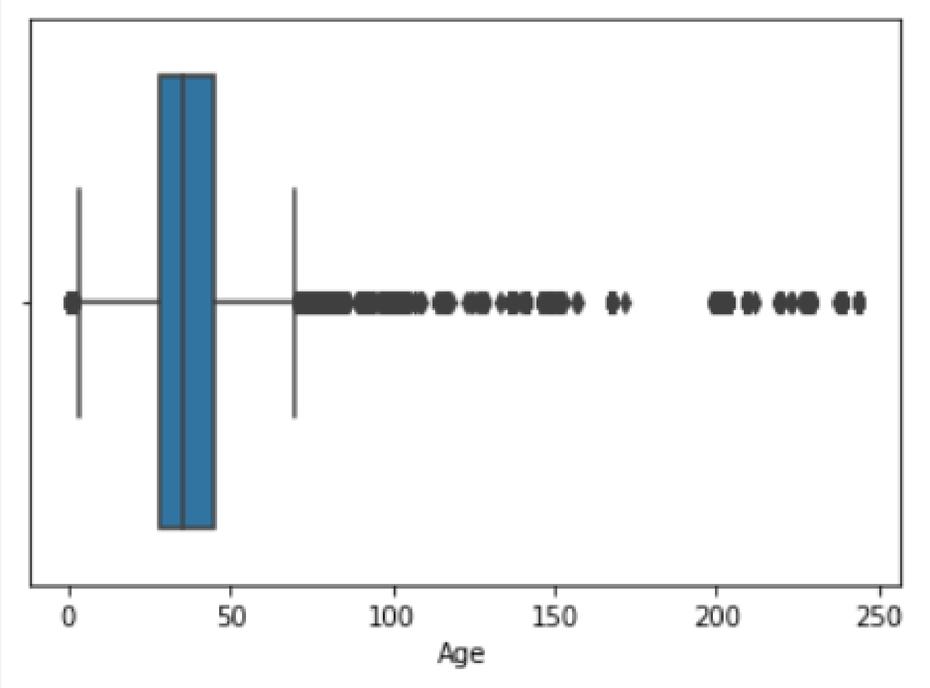
데이터 EDA 과정

데이터 전처리 - 이상치 처리



Year

0, 2050년 같은 이상치가 존재,
year의 범위를 1975년부터 2012년까지로 제한



Age

0, 244살과 같은 이상치가 존재,
범위를 14세 이상부터 100세 이하로 제한

데이터 EDA 과정

데이터 전처리 - 타입 변경

- Year: int -> object

```
df['Year-Of-Publication'] = df['Year-Of-Publication'].astype("O")
```

- User-ID: int -> object

```
df['User-ID'] = df['User-ID'].astype('O')
```

데이터 EDA 과정

데이터 전처리 - 불필요한 컬럼 제거

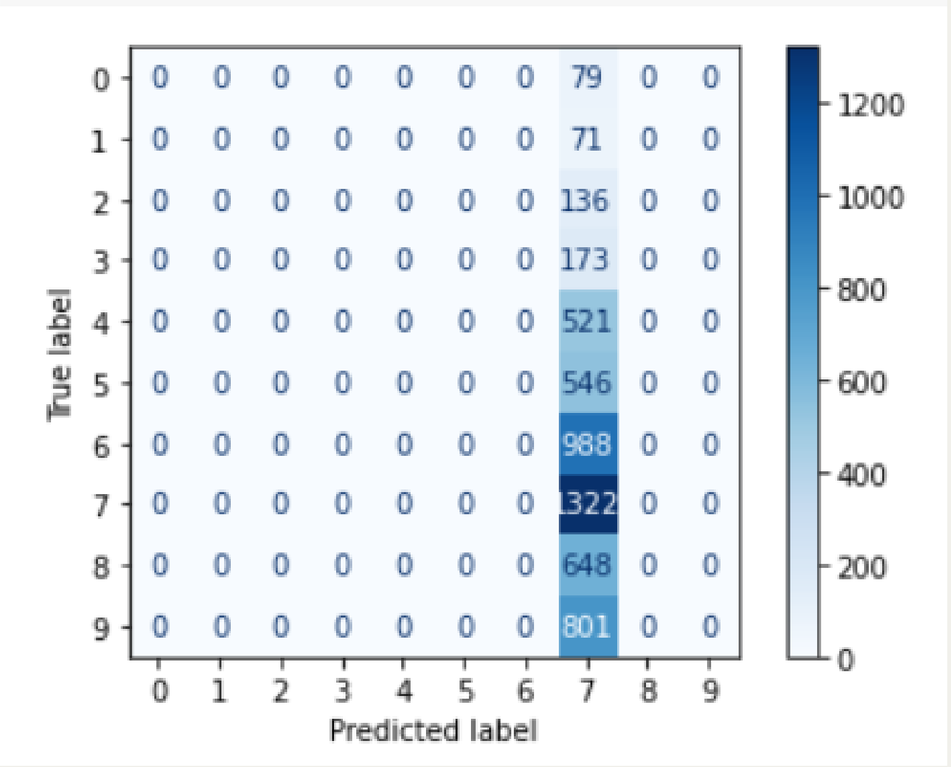
- **Image 관련 컬럼과 Location 컬럼 삭제**

```
# 컬럼 삭제
drop_col = ['Image-URL-S', 'Image-URL-M', 'Image-URL-L', 'Location']
df.drop(drop_col, axis=1, inplace=True)
```

모델 학습 및 검증

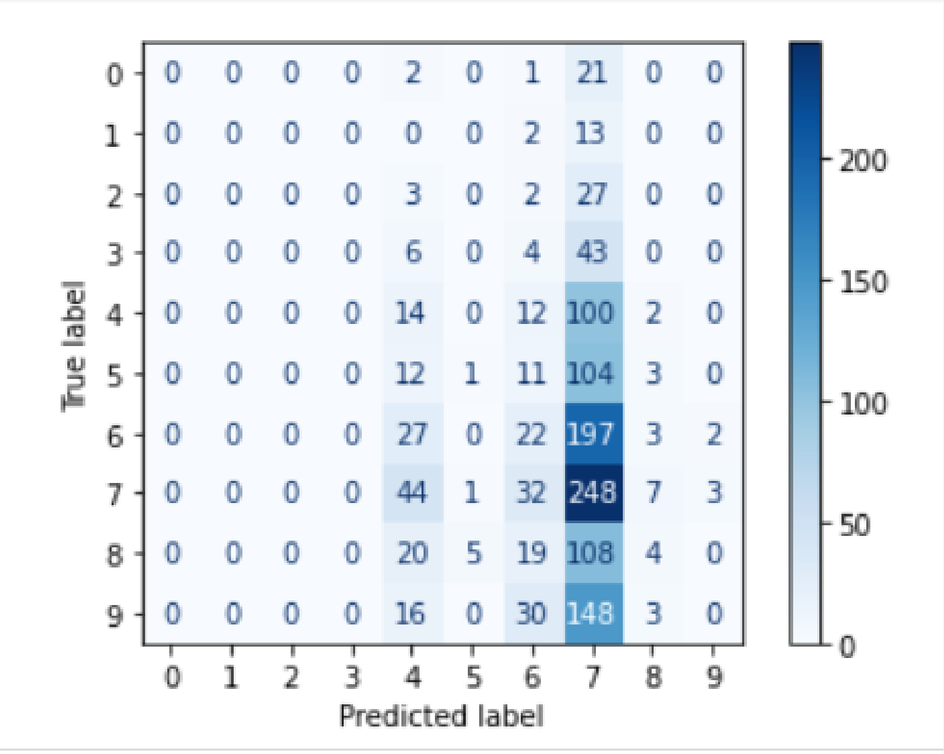
baseline: 최빈값으로 설정

	precision	recall	f1-score	support
1	0.00	0.00	0.00	79
2	0.00	0.00	0.00	71
3	0.00	0.00	0.00	136
4	0.00	0.00	0.00	173
5	0.00	0.00	0.00	521
6	0.00	0.00	0.00	546
7	0.00	0.00	0.00	988
8	0.25	1.00	0.40	1322
9	0.00	0.00	0.00	648
10	0.00	0.00	0.00	801
accuracy			0.25	5285
macro avg	0.03	0.10	0.04	5285
weighted avg	0.06	0.25	0.10	5285



모델 학습 및 검증

DecisionTreeClassifier



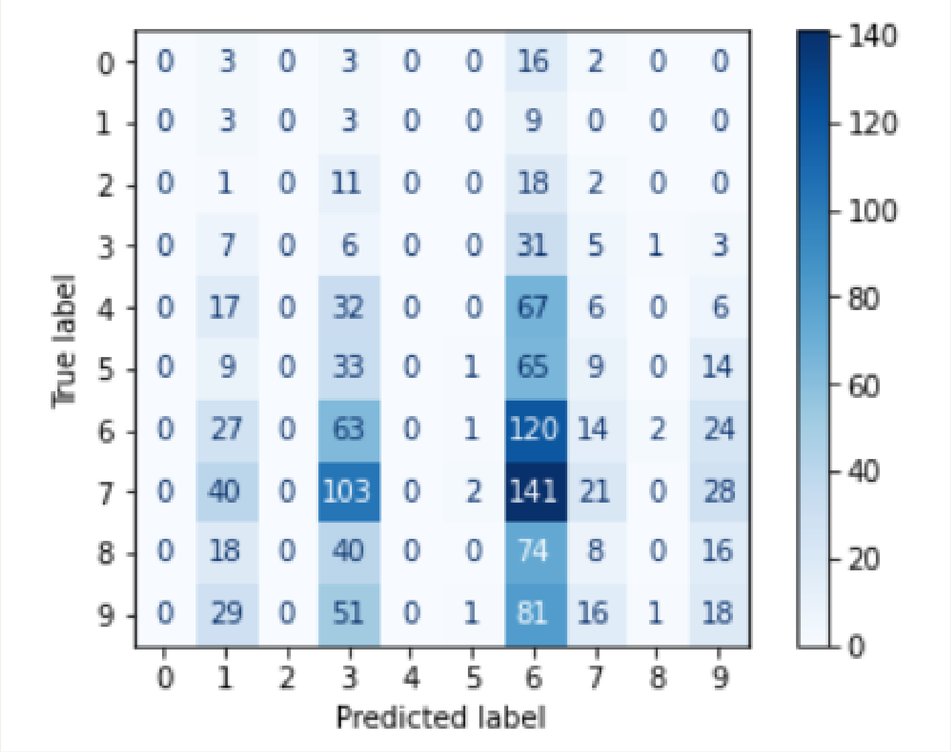

```
학습데이터 AUC:
0.50801166241158

검증데이터 평가지표:
정확도: 0.2186, 정밀도: 0.1382, 재현율: 0.2186, F1: 0.1319, AUC: 0.4980
```

- 정확도: 전체 샘플 중 맞게 예측한 샘플의 수
- 정밀도: 양성 클래스에 속한다고 출력한 샘플 중
  실제로 양성 클래스에 속하는 샘플 수의 비율
- 재현율: 실제 양성 클래스에 속한 표본 중
  양성 클래스에 속한다고 출력한 표본 수의 비율
- f1 score: 정밀도와 재현율의 가중조화평균
- auc: roc curve 면적

모델 학습 및 검증

RandomForestClassifier + UnderSampling

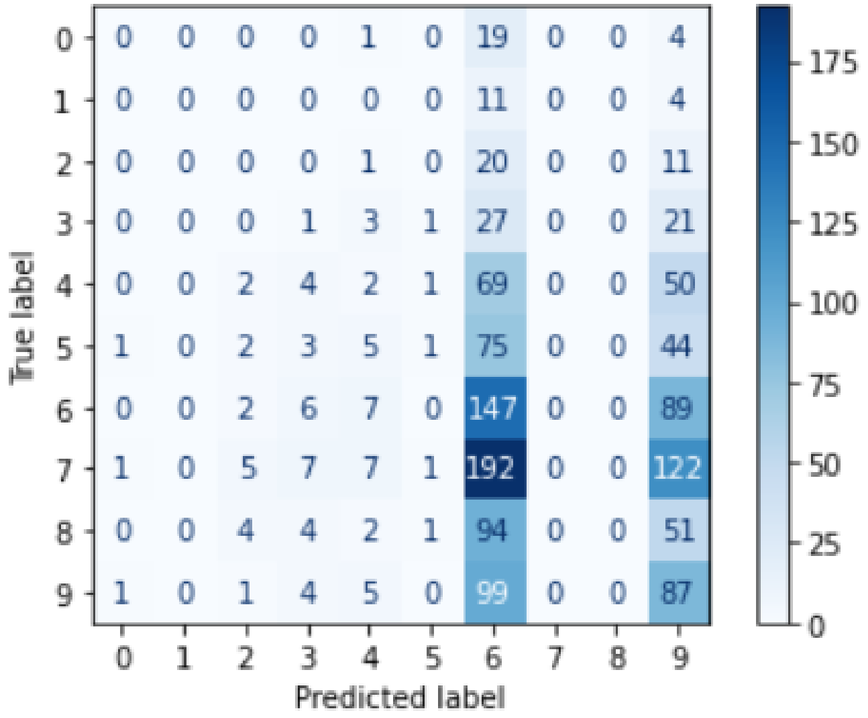


학습데이터 AUC:
0.50801166241158

검증데이터 평가지표:
정확도: 0.1278, 정밀도: 0.1461, 재현율: 0.1278, F1: 0.0983, AUC: 0.5156

모델 학습 및 검증

ExtraTreeClassifier + Undersampling

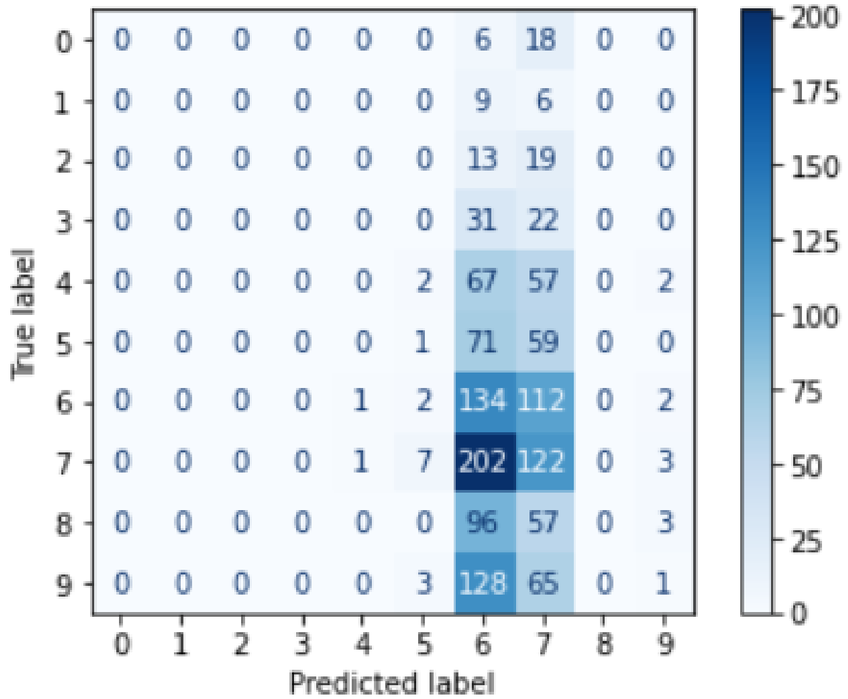


학습데이터 AUC:
0.50801166241158

검증데이터 평가지표:
정확도: 0.1800, 정밀도: 0.0910, 재현율: 0.1800, F1: 0.0986, AUC: 0.5136

모델 학습 및 검증

XGBClassifier + 가중치

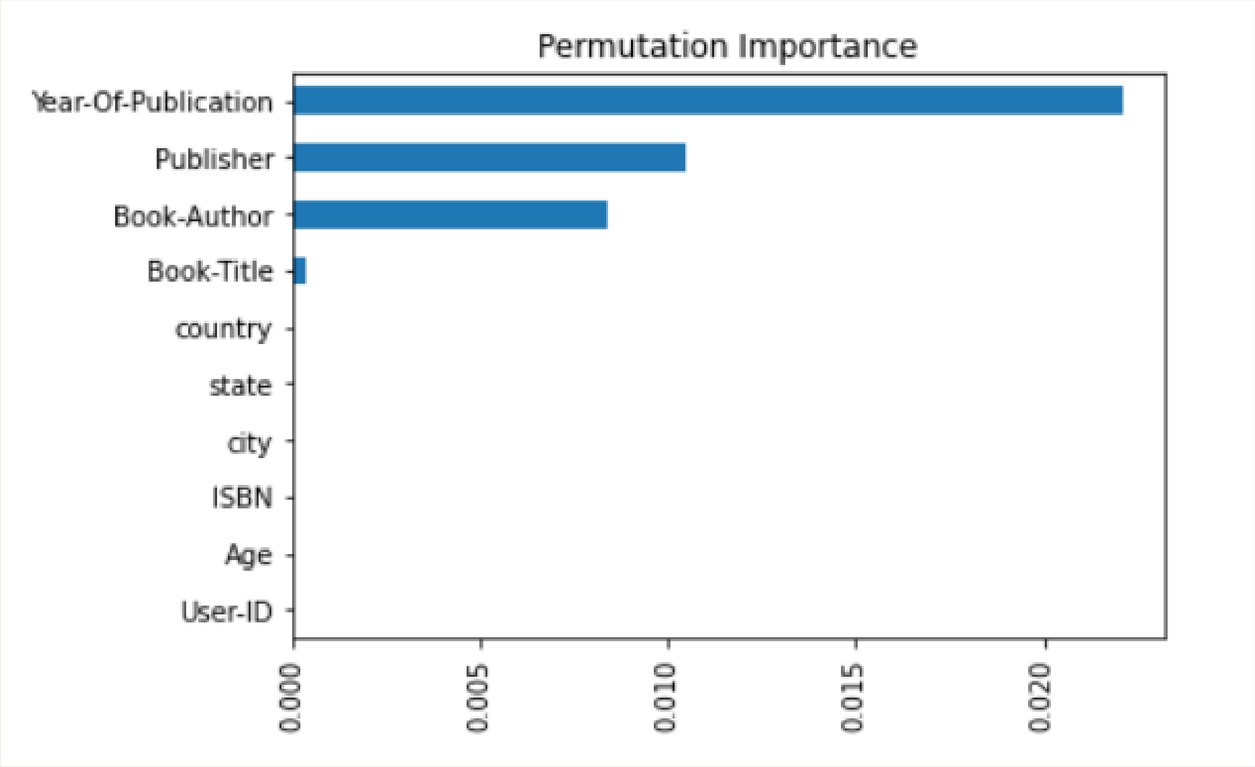


학습데이터 AUC:
0.5094540357093563

검증데이터 평가지표:
정확도: 0.1952, 정밀도: 0.1113, 재현율: 0.1952, F1: 0.1242, AUC: 0.5136

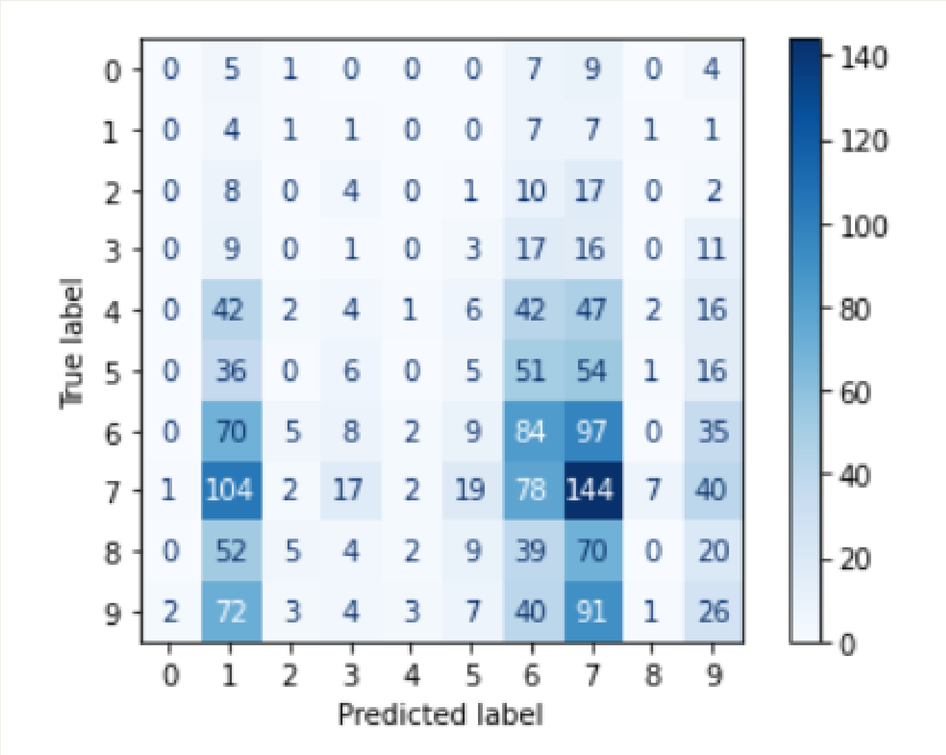
모델 해석

Permutation Importance: RandomForestClassifier 모델 활용



모델 해석

테스트 데이터 셋 결과

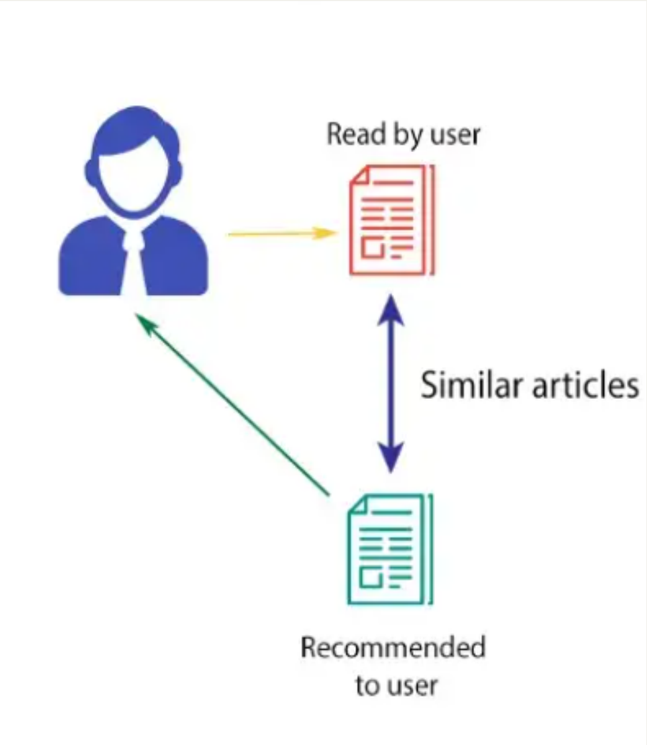


모델: RandomForestClassifier /특성: 출판연도, 출판사, 작가, 책 제목 사용

테스트데이터 평가지표:
정확도: 0.1604, 정밀도: 0.1496, 재현율: 0.1604, F1: 0.1459, AUC: 0.4978

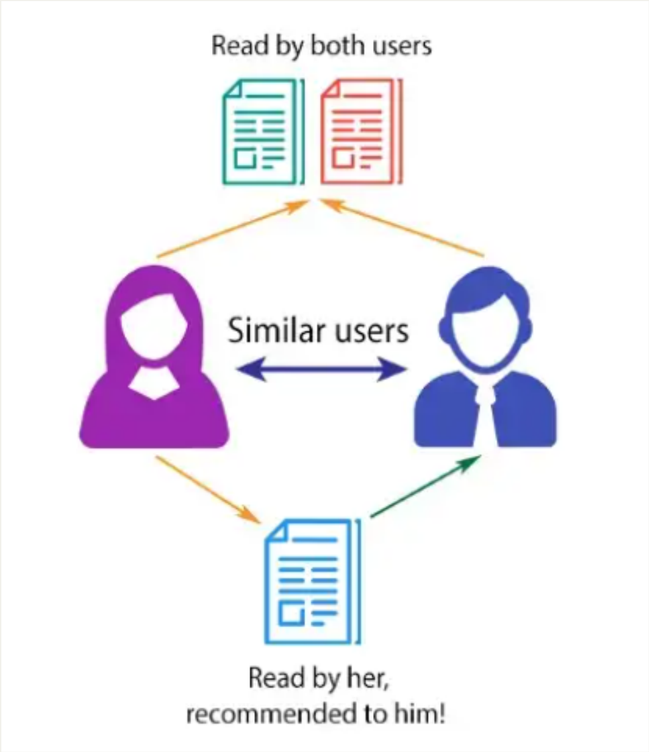
향후 진행할 사항

협업 필터링을 활용해 여러 사용자들의 점수를 예측하고 특정 점수를 넘어가면 추천하는 알고리즘을 구현



콘텐츠 기반 필터링

아이템의 특성과 사용자의 선호도를 비교해 추천하는 방식



협업 필터링

유사한 사용 행동을 파악하여, 비슷한 성향의
사람들에게 아이템을 추천

참고 자료

- https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/
- https://romg2.github.io/mlguide/03_%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-%EC%99%84%EB%B2%BD%EA%B0%80%EC%9D%B4%EB%93%9C-04.-%EB%B6%84%EB%A5%98-XGBoost/
- https://dining-developer.tistory.com/27
- https://aimb.tistory.com/152
- https://towardsdatascience.com/the-right-way-of-using-smote-with-cross-validation-92a8d09d00c7
- https://tensorflow.blog/2017/11/30/%EB%8D%94%EC%9A%B1-%EB%9E%9C%EB%8D%A4%ED%95%9C-%ED%8F%AC%EB%A0%88%EC%8A%A4%ED%8A%B8-%EC%9D%B5%EC%8A%A4%ED%8A%B8%EB%A6%BC-%EB%9E%9C%EB%8D%A4-%ED%8A%B8%EB%A6%ACextratreesclassifier/
- https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset?select=Ratings.csv
- https://datascienceschool.net/03%20machine%20learning/09.04%20%EB%B6%84%EB%A5%98%20%EC%84%B1%EB%8A%A5%ED%8F%89%EA%B0%80.html

THANK YOU

감사합니다