6조

[IC-PBL]

어플리케이션 평점 예측

데이터마이닝

윤성호 소프트웨어 2017012242

윤재승 산업경영 2016006371

이수아 소프트웨어 2017012333

이민재 소프트웨어 2017012306

이다현 로봇 2019018577

Contents

주제 및 목표

App

시나리오

한 IT 회사에서 IOS 기반 어플리케이션을 안드로이드 기반에서도 출시하는 것이 좋을 지 판단하고자 한다. 따라서 한 부서에서 Google PlayStore에서 어플리케이션을 출시하였을 때 사용자의 평점이 어떨 지 예측하는 과제를 맡게 되었다. **App Store 기반에서 해당 어플리케이션에 대한 정보들을 기반으로 사용하여 Google PlayStore 출시 시에 어느 정도의 평점을 받을 수 있을 지 예측하기 위해 Google PlayStore 내 어플리케이션 정보를 통한 평점 예측** 프로젝트를 진행하고자 한다. 이를 통해 안드로이드 기반 어플리케이션을 출시하였을 때 회사에게 이득이 되는 지 생각해 본다.

목표

Google Playstore의 기존 앱 데이터를 기반으로 사용자의 평점을 분석하고, 학습 알고리즘을 통해 기대 평점을 확인한다.

문제

평점 구간을 나누어 어플의 평점을 예측하는 분류(Classification) 문제로 접근

Raw 데이터 소개

Feature

> 데이터 정보

: 구글에서 제공하는 Cloud VM Instance API를 이용해 수집된 **Google Playstore 안드로이드 앱 데이터**로 2021년 6월까지 Google Playstore에 출시된 어플들의 이름, 평점, 설치 수, 가격 등의 정보가 담겨있다.

> 데이터 출처

: **Kaggle**
(https://www.kaggle.com/gauthamp10/google-playstore-apps)

> 데이터 수

: **약 230만 개**

> 속성 수

: **24개**

#	Column	Dtype	Description
0	App Name	object	앱 이름
1	App Id	object	앱 고유 ID
2	Category	object	앱 카테고리
3	Rating	float	평가 점수 (0~5)
4	Rating Count	float	점수 평가 부여 수
5	Installs	object	다운로드 수
6	Minimum Installs	float	최소 다운로드 수
7	Maximum Installs	int	최대 다운로드 수
8	Free	bool	앱 무료 여부
9	Price	float	앱 가격
10	Currency	object	앱 판매 재화
11	Size	object	앱 크기
12	Minimum Android	object	앱 구동을 위한 최소 안드로이드 버전
13	Developer Id	object	개발자 ID
14	Developer Website	object	개발자 웹 사이트 주소
15	Developer Email	object	개발자 E-mail 주소
16	Released	object	배포 시점
17	Last Updated	object	최근 업데이트 날짜
18	Content Rating	object	이용 가능 나이
19	Privacy Policy	object	개인 보호 정책 설명 홈페이지 링크
20	Ad Supported	bool	광고 지원 여부
21	In App Purchases	bool	앱 내 결제 여부
22	Editiors Choice	bool	GooglePlayStore 편집팀 추천 여부
23	Scraped Time	object	데이터 수집 시간

데이터 전처리 - 속성 선택

속성 선택

1) 학습을 위해 필요 없다고 확연하게 판단된 속성들과 고유한 값들이 있는 속성 제거

: App Name, App Id, Developer Id, Developer Website, Developer Email, Privacy Policy, Scraped Time, Minimum Android

2) 데이터 불균형이 심한 속성 제거

: Editors Choice, Currency

Editors Choice: False 값이 매우 적고 True 값이 압도적으로 많이 존재
Currency: USD에 대한 데이터가 압도적으로 많이 존재

3) 독립 속성 간 중복성을 갖는 속성 제거

: Installs, Free

1의 자리 수 까지 세분하게 표시된 Maximum Installs 속성을 남겨두고 비슷한 의미를 갖는 Installs와 Minimum Installs 속성 제거
Free는 Price가 0인 경우와 중복되는 내용이므로 Price 속성을 남겨두고 Free 속성 제거

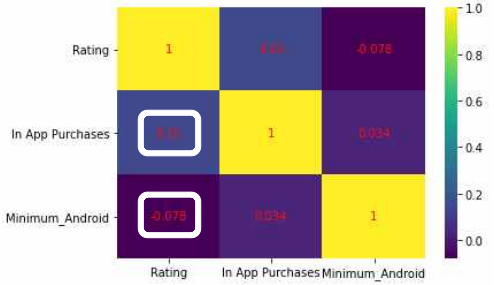
데이터 전처리 - 속성 선택

속성 선택

4) 종속 변수와의 관련성이 없다고 판단된 속성 제거

: Minimum Android, In App Purchases

Rating과의 상관관계가 낮은 수치를 보임



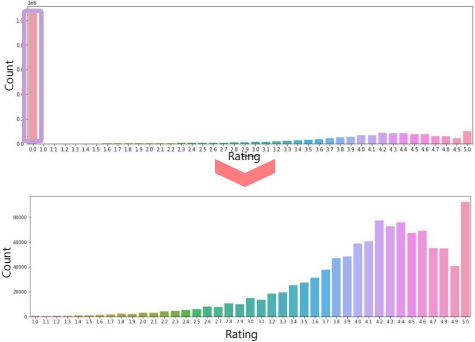
데이터 전처리 - 결측치 제거

Drop

1) Rating

: Rating이 null 혹은 0이면 제거

평점을 주는 경우 1점 이상부터 부여가 가능하다는 점과 0 이외의 평점의 그래프 분포를 분석하기 용이하게 하기 위한 점을 고려하여 0점인 경우와 null인 경우를 제외시킴



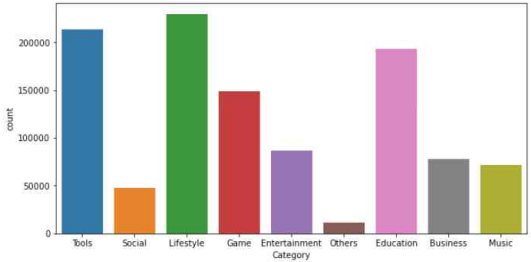
데이터 전처리 - 범위 간소화

Category

1) Category

: 비슷한 의미를 갖는 카테고리를 합쳐 48개의 카테고리를 9개의 카테고리로 정리

: Business, Education, Entertainment, Game, Lifestyle, Music, Social, Tools, Others로 간소화



데이터 전처리 - 범위 간소화

Category

2) Content Rating

: 비슷한 연령대로 나뉜 카테고리를 합쳐 6개의 카테고리를 3개의 카테고리로 정리

: Adults, Everyone, Teen로 간소화

Everyone	1081795
Teen	111570
Mature 17+	37054
Everyone 10+	22551
Unrated	125
Adults only 18+	87

Everyone	1084017
Teen	109253
Adults	36573

Everyone : Everyone, Unrated
Teen : Teen , Everyone 10+
Adults : 17+, 18+

3) Last Update, Released

: 년, 월, 일의 정보를 년 정보만 남겨 종속 변수와 관계를 파악하기 쉽게 정리

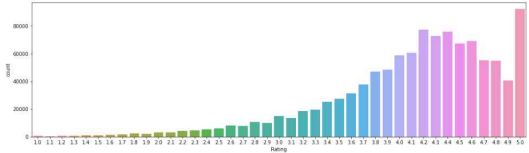
데이터 전처리 - 범위 간소화

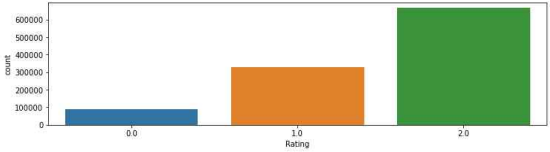
Category

3) Rating

: 분류 문제를 위해 수치형의 데이터 타입을 3개의 범위로 나누어 명목형 타입으로 변경

그룹 별 데이터의 수의 불균형을 최소화하고자 범위를 임의로 정하여 나눔





Group 0: 1 ~ 3
Group 1: 3 ~ 4
Group 2: 4 ~ 5

데이터 전처리 - 데이터 단위 정리

unit

1) Size

: 어플의 크기 단위를 KB(킬로바이트)로 통일

2) Maximum Installs

: 단위 K(1,000)로 변경

'K', 'k' => (수치)
'M', 'm' => (수치)*10^3
'G', 'g' => (수치)*10^6
'단위X' => (수치)/10^3

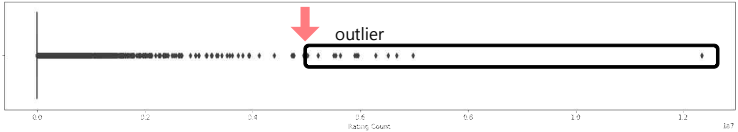
데이터 전처리 - 이상치 제거

Outlier

1) Rating Count

: 데이터가 **정규분포를 따르지 않기 때문에 IQR로 데이터를 제거 시 361 이상 데이터가 제거**되어 적합하지 않다고 판단

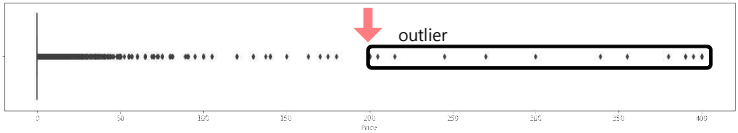
따라서 다양한 Rating Count값을 이용한 학습 결과를 얻기 위해서 데이터 분포를 보고 임의적으로 기준을 **0.5*1e7로 정하여 기준값 이상 제거**



2) Price

: **대부분의 Price가 0$이므로 IQR로 데이터 제거 시 0$가 아니면 모두 이상치로 처리**되어 적합하지 않다고 판단

따라서 Price에 따른 학습 효과를 보기 위해 데이터 분포를 보고 임의적으로 기준을 **200$로 정하여 기준값 이상 제거**



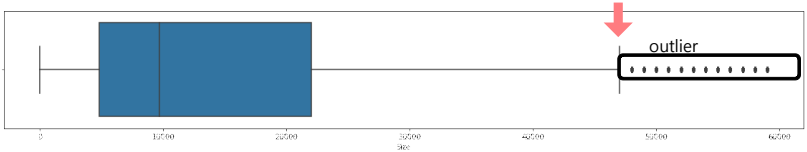
데이터 전처리 - 이상치 제거

Outlier

3) Size

: **IQR을 통해 평균적인 앱 사이즈에 비해 특수하게 큰 사이즈의 데이터 제거**

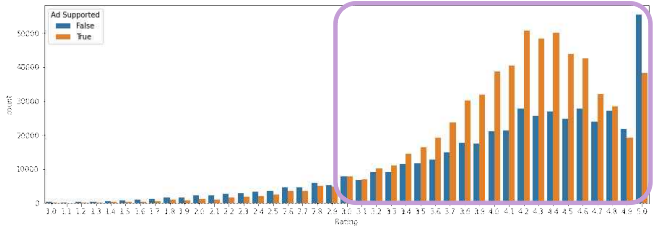
따라서 IQR upper bound인 **49,850이상 제거**



데이터 분석

1) Ad Supported

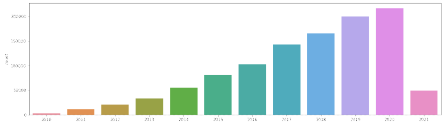
: Rating이 3.0 ~4 후반 사이일 때 광고가 있는 어플의 비율이 더 높음



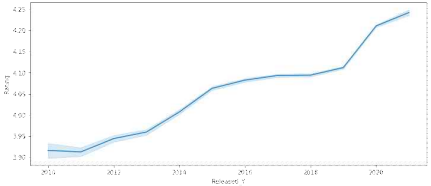
데이터 분석

2) Released

: 최근일수록 출시된 앱의 개수가 많아지는 것으로 보아 앱에 대한 전반적인 관심과 사용이 많다는 것을 볼 수 있음



2021년은 데이터가 완전히 수집되지 않아 적음



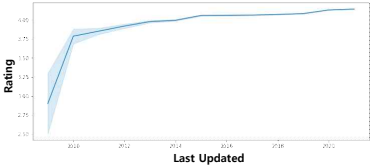
출시한지 오래된 앱 일수록 평균 Rating이 낮은 경향을 보임

데이터 분석

3) Last Updated

: 최근 업데이트된 앱일 수록 앱의 크기와 평점이 높게 나타나는 경향을 볼 수 있음

Size가 큰 앱들은 비교적 최근일수록 많음



업데이트 년도가 최근일 수록 Rating이 높게 나타남

최종 데이터셋

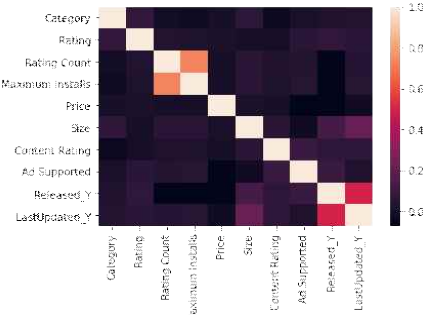
Dataset

> 데이터 수: 약 108만 개

> 속성 수: 10개 (독립 변수: 9개, 종속 변수: Rating)

> 상관관계
: Rating과의 상관관계가 높은 속성은 뚜렷하게 나타나지 않음
: Rating Count와 Maximum Installs의 상관관계가 높은 것을 통해 다운로드 수가 많기 때문에 평가도 많이 받게 되는 선형적 관계를 파악할 수 있음

[상관관계]



학습을 위한 데이터 처리

Check

1) 인코딩

명목형 속성들은 수치로 변환을 시켜주기 위해 인코딩을 진행

: One-Hot Encoder, Oridinal Encoder, Label Encoding 등 사용

2) 스케일링

수치형 속성들에 대해 큰 수의 값이 종속변수에 큰 영향을 줄 수 있으므로 정규화 진행

: MinMax, Robust, Standard Scaler 등 사용

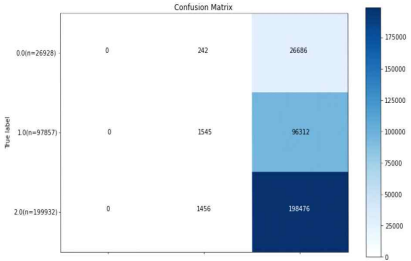
```
Categorical = ['Category','Content Rating','Ad Supported']
Numerical = ['Rating Count','Maximum Installs','Price','Size','Released_Y','LastUpdated_Y']
label = ['Rating']
```

학습을 위한 데이터 처리

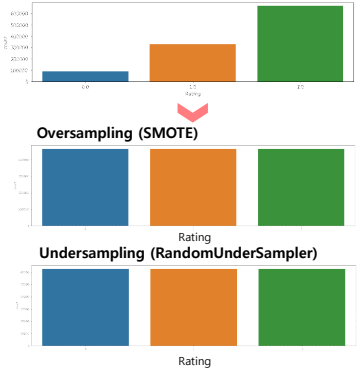
Check

3) Resampling

주어진 데이터로 학습시키면 Rating이 2인 데이터의 양이 상대적으로 많기 때문에
Accuracy 는 높지만 상대적으로 데이터 개수가 많은 2로만 예측하는 경향



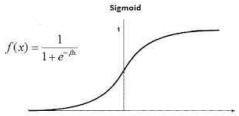
: 따라서 UnderSampling & OverSampling이 필요할 것이라 판단



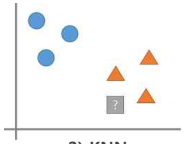
학습

Check

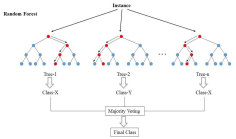
> 학습 모델



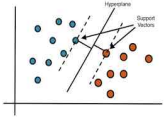
1) Logistic Regression



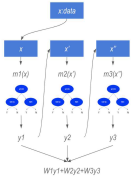
2) KNN



3) RandomForest



4) SVM



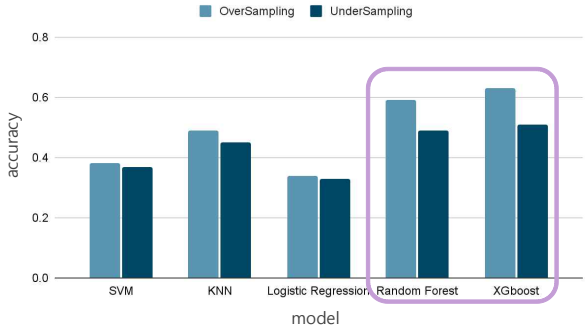
5) XGBoost

학습 결과

Resampling

> 학습 모델 별 정확도

전체적으로 OverSampling의 결과가 UnderSampling의 결과보다 높아 Oversampling 데이터 사용



학습 결과 - XGboost

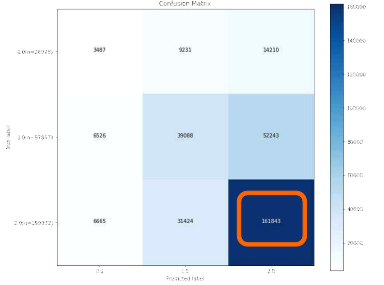
XGboost

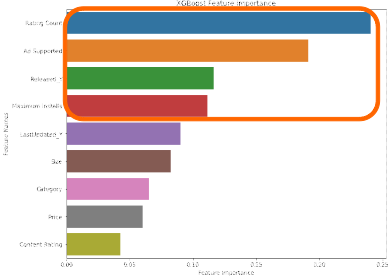
> XGboost

: 학습 결과 63%의 정확도를 나타나는 것을 확인할 수 있음

: Rating 그룹 2에 대한 예측 결과가 다른 그룹에 비해 정밀도, 재현율의 결과가 확연히 높게 나타나는 것을 볼 수 있음

: Rating Count, Ad Supported , Released, Maximum Installs가 학습 결과의 약 50%의 중요성을 차지하는 것을 확인할 수 있음





학습 결과 - RandomForest

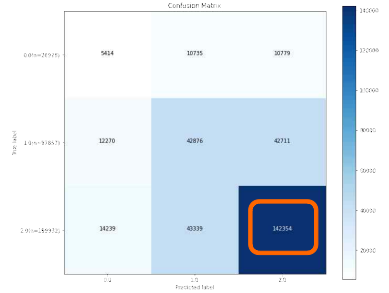
Random Forest

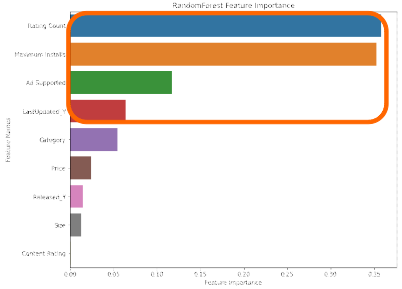
> RandomForest

: 학습 결과 59%의 정확도를 나타나는 것을 확인할 수 있음

: Rating 그룹 2에 대한 예측 결과가 다른 그룹에 비해 정밀도, 재현율의 결과가 확연히 높게 나타나는 것을 볼 수 있음

: Rating Count, Maximum Installs, Ad Supported, LastUpdated가 학습 결과의 약 90%의 중요성을 차지하는 것을 확인할 수 있음





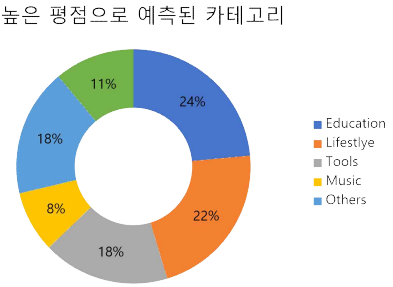
결과 분석

Category

> Category와의 관계 분석

4 ~ 5점의 평점으로 분류된 Rating Group 2로 예측한 Category 종류 별 비율을 통해
Education, Lifestyle, Tools, Game, Music 순서로 높은 평점을 받는 Category임을 확인할 수 있음

: Education, Lifestyle, Tools, Game, Music과 관련된 어플리케이션일 경우 높은 평점을 기대할 수 있을 것으로 판단됨



결과 분석

Content Rating

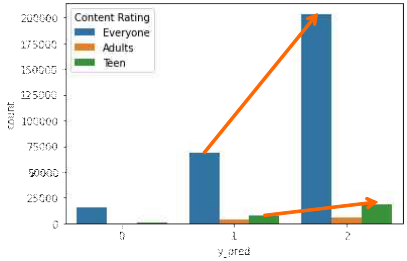
> Content Rating와의 관계 분석

Content Rating별 Rating 예측 비율을 통해

전체 연령 사용이 가능한 Everyone 그룹이 좋은 평가 점수를 받았음을 확인할 수 있음

Content Rating 항목 별 데이터 수를 고려하여 보았을 때 Everyone, Teen의 경우 Group 2의 비율이 다른 Group에 비해 확연히 높고, Adults의 경우 Group 1과 Group 2의 비율이 눈에 띄는 차이를 보여주진 않음

: **전체 연령 및 청소년 가능** 앱일 경우 **높은 평점을 기대**할 수 있을 것으로 판단됨



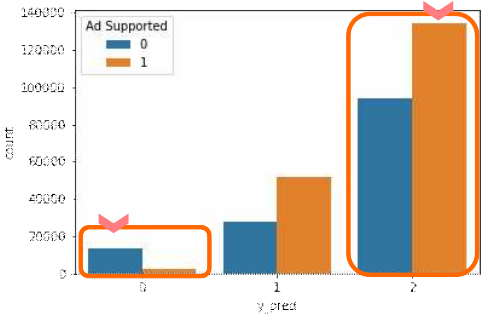
결과 분석

Ad Supported

> Ad Supported와의 관계 분석

Ad Supported에 따른 Rating 예측 비율을 통해
평점이 낮은 Group 0인 경우 광고 지원을 받는 경우가 지원을 받지 않는 경우보다 적고,
평점이 높은 Group 1, 2인 경우 광고 지원을 받는 경우가 지원을 받지 않는 경우보다 많은 것을 확인할 수 있음

: **광고를 지원받는 앱**이 높은 평점을 기대할 수 있을 것으로 판단됨



결과 적용

Ad Supported

> 문제사항 적용

IT 회사의 어플리케이션이 'Eduction', 'LifeStyle', 'Tools', 'Game'과 관련되고,
사용 가능 연령의 범위가 넓고,
광고를 지원받고,
최대 다운로드 수가 높은 경우
Google PlayStore에서의 높은 평점을 기대할 수 있을 것으로 예상됨.

위의 경우를 만족하여 높은 평점을 기대할 수 있다고 판단된다면 Google PlayStore에 출시하여 안드로이드 사용자를 늘리는 것이 회사에 이득이 될 수 있다고 판단해 볼 수 있음

소감

Check

> **실생활에서 접할 수 있는 문제상황을 가정하여 분류 문제를 진행해 봄으로써 주변에서 접할 수 있는 문제에서 예측을 통한 분석이 필요한 경우 데이터 분석 과정과 학습을 통해 접근할 수 있는 역량을 기를 수 있었음**

> **수집된 데이터를 사용하기 위한 데이터 전처리 과정이 큰 비중을 차지함을 깨달음**

데이터셋에서 학습을 위한 속성의 선택과 이상치를 가지는 데이터를 처리하는 일들을 하면서 데이터 전처리에 대한 이해를 할 수 있었음.

> **다양한 학습 모델을 통해 도출되는 결과의 차이를 비교하며 적합한 모델을 찾아보는 과정의 의미**

여러가지 학습 모델로 학습을 시키면서 각 모델의 특징을 파악할 수 있었고, 해당 주제에 적합한 학습 모델이 무엇인지도 알 수 있었음.

The END