**목표**

우리는 다양한 음악 정보와 사용자 정보, 사용자의 음악 접근 정보를 이용해서 한 달 이내에 사용자가 해당 음악을 다시 들을 지 여부를 판별해야 한다.

**활용 가능한 데이터** (\*: 연속값)

음악 정보: 음악 길이\*, 장르, 제목, isrc 번호, 가수, 작사가, 작곡가, 언어

사용자 정보: 도시, 연령\*, 성별, 등록일\*, 만료일\*, 등록 방법

음악 접근 정보(train, test): 사용자, 음악, 접근 메뉴, 접근 방법, 레이아웃

음악 간접 데이터: Music Fingerprint\*, 유튜브 사용자 반응, 발매일\*, …

Outlying Data에 대해서도 적절한 선처리를 해 주어야 한다. 특히 간접 데이터는 외부 데이터나 웹 크롤링을 통해 수집할 수 있으나, 모든 음악에 대해 간접 데이터를 수집할 수 있는 것은 아니므로 수집되지 않은 노래에 대해 적절한 처리를 해 주어야 한다.

**데이터 경향성 파악**

모든 데이터를 한 가지 모델에 끼워 넣어도 되지만, 그렇게 하면 좋은 예측이 나올 수 없다. 따라서 모델을 세우기 전에 데이터 자체를 분석하고 내재되어 있는 경향성을 파악하는 것이 중요하다.

데이터 경향성 가설 예시:

* 연령 분포는 20대 사용자를 Peak로 하여 종 모양 곡선을 그릴 것이다.
* 작사가와 작곡가는 사용자의 음악 선택에 별 영향을 주지 않을 것이다.
* 연령과 음악의 발매일은 음의 상관관계를 가질 것이다.
* 다시 듣기 빈도는 장르에 따라 차이를 보일 것이다.
* 검색을 통해 들은 사람들은 공개된 재생목록에서 들은 사람보다 다시 들을 확률이 높다.

다만, 음악 접근정보에서 음악을 들은 일시가 주어지지 않음에 유의해야 한다\*\*. 이 때문에 “밤에 들은 음악은 다시 들을 확률이 높다”, “계절에 따라 유행하는 장르가 다르다”와 같은 가설은 검증할 수 없다. 물론 KKBOX 측에서 나중에 이 정보를 공개할 수도 있다.

\*\* 참고: https://www.kaggle.com/c/kkbox-music-recommendation-challenge/discussion/40854

가설을 검증하려면 실제 데이터를 통계 처리하여 상관 계수와 같은 항목을 확인해 본다. 3변수 이상에서 상관관계가 나타나는 경우도 있으므로, 다양한 가설을 세우고 검증해 보아야 한다.

**모델 세우기**

파악한 데이터 경향성을 바탕으로 모델을 작성할 수 있다. GBM, Decision Tree와 같은 다양한 모델이 있는데, 우리 데이터를 분석하기에 어떤 모델이 적절할 지 검토해 보아야 한다. 다양한 모델들이 파이썬 라이브러리로 나와 있으므로 확인해 볼 필요가 있다. 필요하다면 여러 모델을 테스트하여 어떤 모델이 좋을 지 판단할 수도 있다.

특정 모델의 경우, 모델이 우리의 예측을 얼마나 잘 반영하고 있는지 확인할 수도 있다. 예를 들어 Decision Tree의 경우 위쪽 노드의 분류 기준과 우리가 세운 가설을 비교할 수 있으며, 거꾸로 생성된 Decision Tree로부터 적절한 가설을 찾아낼 수도 있다.