Melon Diversity 정량적 해석(7주)

프로젝트 개요

- 추천시스템에는 accuracy 뿐만 diversity, novelty, serendipity 등의 고려요소가 있습니다. 이러한 요소들은 장기적으로 유저의 retention에 영향을 미칠 것이라는 가설이 있습니다.
- 이러한 가설을 기반으로 분석해낸 것이 아래의 논문입니다.
- Algorithmic Effects on the Diversity of Consumption on Spotify https://www.cs.toronto.edu/~ashton/pubs/alg-effects-spotify-www2020.pdf

결과

- 위의 링크 논문을 Melon 데이터로 재현했습니다.
- 다양성이 무엇인지, 유저의 다양성이 retention에 미치는 긍정적인 영향을 분석하였습니다. 높은 다양성의 유저는 retention할 확률이 높았습니다. 다양성은 증가시켜 마땅한 긍정적인 지표였습니다.
- 다양성이 증가하는 유저가 어떤 유저인지 분석했습니다.
- 개인의 다양성 수준과 추천의 다양성 수준에 따른 유저 반응 차이를 분석했습니다.
- GS-score라는 diversity에 대한 지표는 계산 cost가 매우 높은 지표입니다. 따라서 이를 계산이 좀 더 용이한 지표로 대체하고자 대체 지표를 찾아보았습니다.

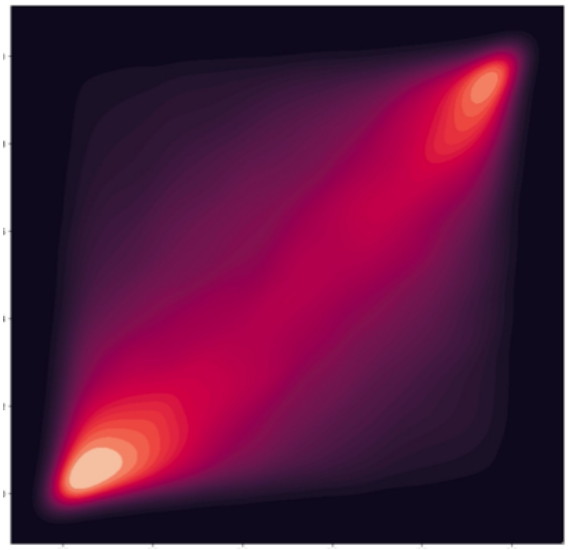
배운 점

- 쿼리를 날리면서 효율적인 쿼리가 무엇인지 배웠습니다. 효율적인 데이터 처리에 관심이 깊어졌습니다.
  - 1달에 약 30GB의 로그가 있었습니다. 이를 한번에 가져오려면 반드시 쿼리가 죽었습니다.
  - user_id%MOD==0, 1, ..., MOD-1의 방식을 통해서 쿼리를 나눠보냈습니다. 이를 통해서 데이터를 효율적으로 가져올 수 있었습니다.
  - 가져와서 전처리 등을 수행하는 것보다 쿼리를 통해서 전처리를 끝내서 가져오는 것이 더 빠르며 효율적이었습니다.
- 분석은 문제에 대한 이해가 중요하며, 데이터의 특징과 한계를 이해해야하고, 도출한 인사이트가 타당한지 끊임없이 검증해야함을 배웠습니다.
- 분석을 전달할 때는 자연스러운 스토리텔링을 통해서 전달해야 훨씬 더 이해하기 쉬움을 배웠습니다.

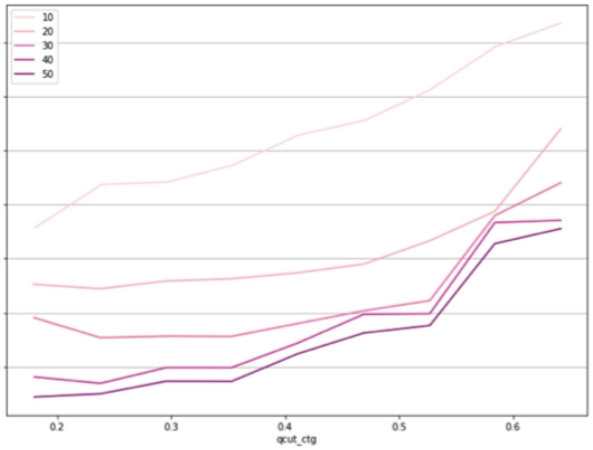
기술 keyword :

diversity, k-means clustering, entorpy, odds ratio

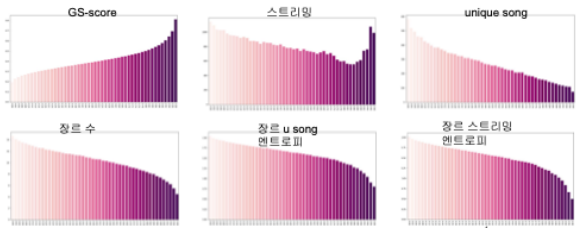
pandas, seaborn등 각종 시각화 툴



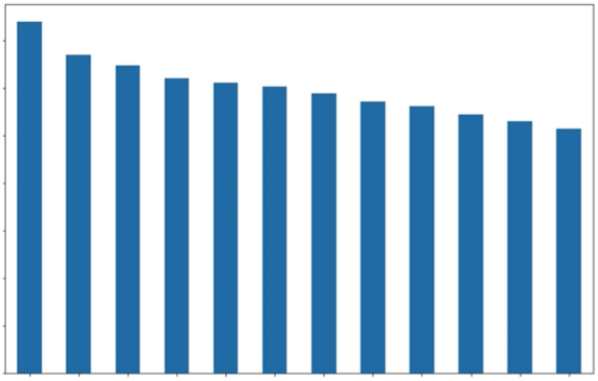
1년 뒤 diversity 변화 추이



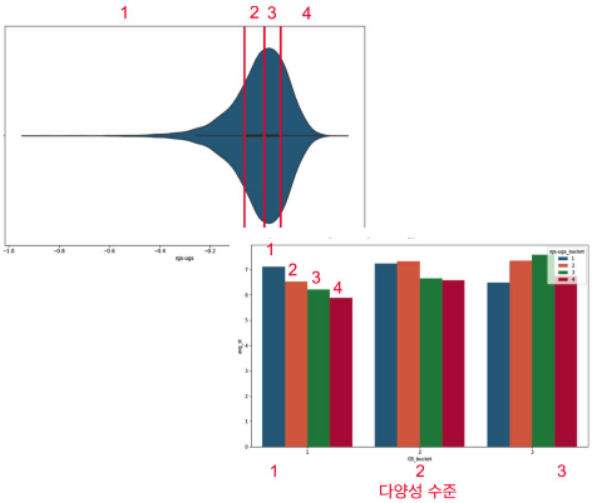
diversity와 activity에 따른 이탈률



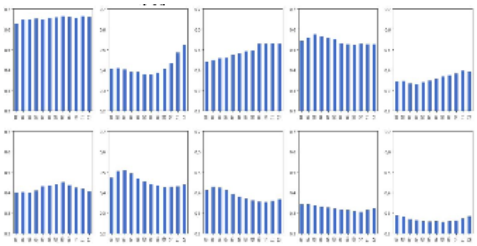
대체지표 탐색



diversity의 월별 상관계수



추천의 다양성 정도에 따른 유저 반응 분석



diversity의 경향성에 따른 군집 분석(이탈률 분석)