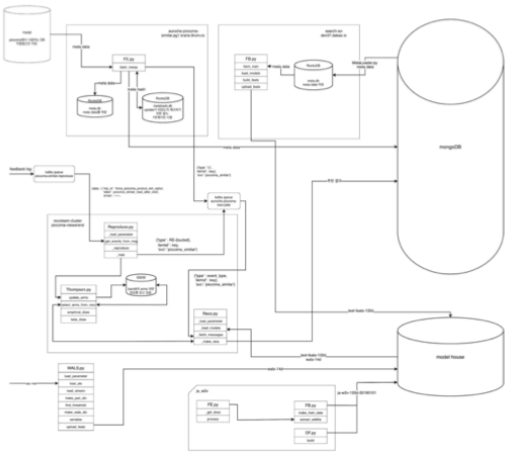
포트폴리오

In Kakao

Piccoma 연관추천 개선 프로젝트(3주)



대외비일지도 몰라서 해상도를 많이 낮췄습니다.

프로젝트 개요

- Piccoma의 추천 구좌 중 연관추천 구좌의 모델을 개선하는 프로젝트
- 일차적으로 위의 데이터 플로우를 이해하였습니다.

프로젝트 아이디어

- read after click이 1회만 발생하는 작품은 표지만 매력적인 소위 표지 사기 작품일 확률이 높을 것이라는 아이디어에서 시작했습니다.(read after click : 추천 구좌를 통해서 발생하는 모든 클릭 log, 연속적인 작품의 열람도 포함)

eda 결과

- 전체 유료(1화부터 유료) 작품의 경우, 1 개의 작품만 보고 끝내는 비율이 2화 이상을 보는 경우의 2배가 넘었습니다.
- 1개 이상 무료로 공개된 작품들에서는 1개의 작품만 보는 경우와 2개 이상을 보는 작품의 비율이 비슷했습니다.
- 이는 전체 유료 작품에서 표지사기로 유입되는 케이스가 상당히 많다는 것을 유추할 수 있습니다.

결과 예측

- 이러한 결과를 바탕으로 생각해보았을 때, 표지만 매력적인 작품은 해당 로직을 통해 없어질 것이며, 2화 이상 지속할만한 작품들을 노출 시킬 것으로 예상했습니다.
- 이는 한 번의 클릭 이 후에 연속적인 작품의 열람 가능성이 높아질 것으로 예상합니다.
- 하지만 표지라도 매력적인 작품들이 없어질 것이므로 이로 인한 클릭이 줄어들 수 있습니다.

결과

- CVR과 CTR둘 다 떨어졌지만, CTR이 더 많이 떨어졌습니다.
- 이는 작품을 들어가서 연속적으로 보는 비율이 더 높아졌음을 시사합니다.
- 하지만 표지사기 작품이 없어지고, 없어진 자리를 더 매력적인 작품이 채워주지는 못했습니다.

해결한 이슈

- 과정
  - 카프카 관련 이슈를 겪었습니다.
  - 카프카의 리스너는 정해진 시간(max poll time) 안에 다시 메세지를 fetch 해와야 합니다.
  - 이를 제대로 인지하지 못했고, 여러 개의 실험이 돌아가면서 max poll time을 넘기며, process가 주기적으로 꺼지는 이슈였습니다.
- 해결
  - fetch messages를 5000 → 3000으로 조정하였습니다.
  - 줄어든 처리량 때문에 렉이 계속 쌓였습니다.
  - 프로세스를 1개 더 띄우면서 이를 해결할 수 있었습니다.

배운 점

- 데이터 파이프라인 간의 의존성을 낮추는 것이 효율적이라는 것을 배웠습니다.
- 이슈를 해결하며, 협업의 중요성을 깨달았습니다.
  - 혼자서 원인 분석을 하는 것보다 훨씬 빠르게 분석할 수 있었고, 성공적으로 이슈를 해결할 수 있었습니다.
- 현재 데이터엔지니어를 꿈꾸는 계기가 되었습니다.

기술 keyword :

MAB(multi arm bandit), Tomson Sampling, CF(Collaborative Filtering), CB(Contents based filtering), reciprocal rank fusion, tf-idf, ALS(Alternating Least Square)

RocksDB, SSDB, mysql, kafka

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.

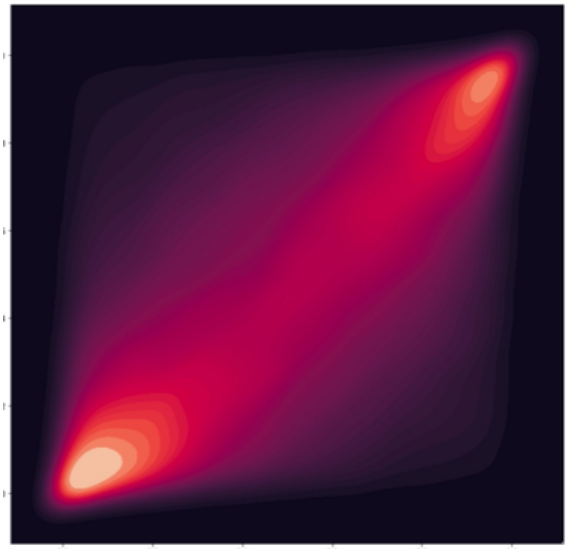
배운 점

- 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.

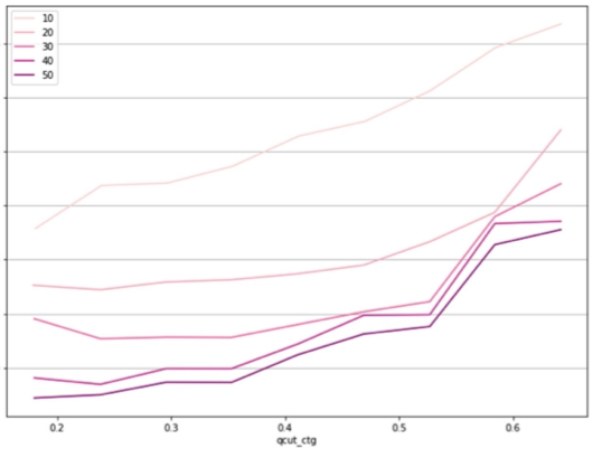
기술 keyword :

diversity, k-means clustering, entorpy, odds ratio

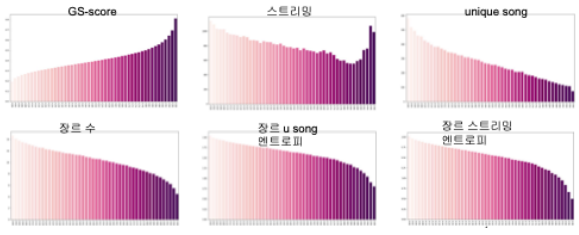
pandas, seaborn등 각종 시각화 툴



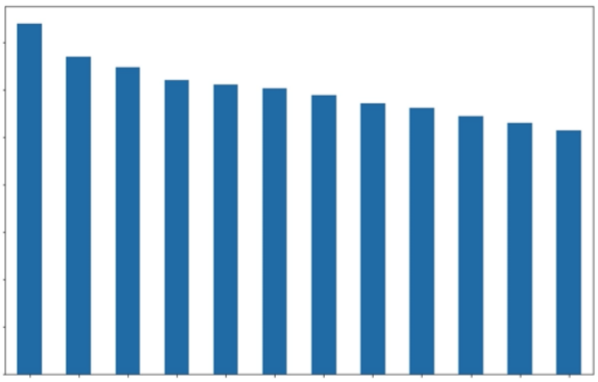
1년 뒤 diversity 변화 추이



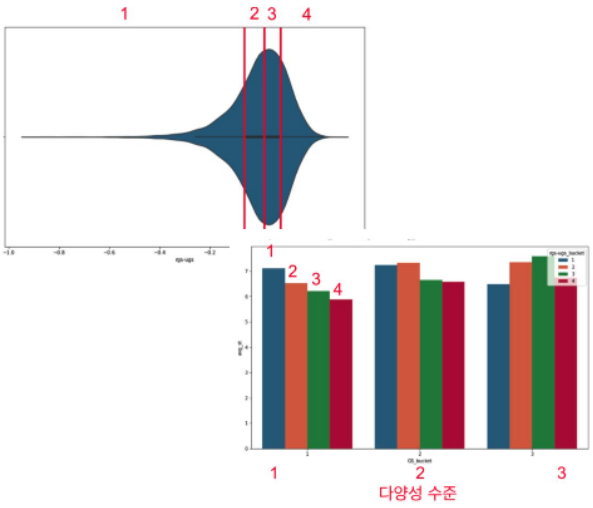
diversity와 activity에 따른 이탈률



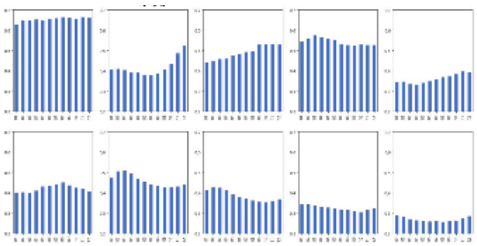
대체지표 탐색



diversity의 월별 상관계수



추천의 다양성 정도에 따른 유저 반응 분석



diversity의 경향성에 따른 군집 분석(이탈률 분석)

In Naver boostcamp AI Tech

마스크 분류모델 대회(10일)

대회 개요

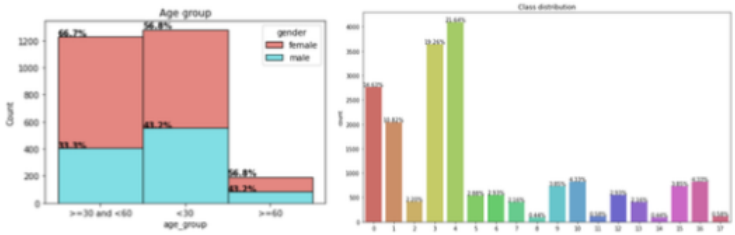
- 마스크 쓴 인물 사진의 나이와 마스크 착용 여부, 성별을 맞추는 대회입니다.

무엇을 했는가?

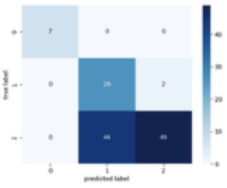
- baseline code를 작성했습니다.
- 데이터에 대한 오류 검수 및 eda를 진행했습니다.
- 몇 가지 모델에 대한 실험을 하였습니다.
- confusion matrix를 학습 코드에 도입했습니다. 이를 통해서 모델이 나이 예측에 어려움을 겪는 것을 파악하고 이를 개선하고자 실험을 하였습니다.
- ansemble 코드를 작성했습니다.

배운 점

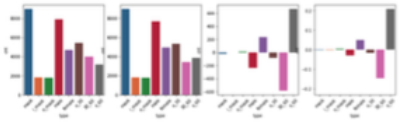
- 협업
  - git hub의 칸반보드를 활용하여 협업이 어떻게 이루어져야 효율적인지 배울 수 있었습니다.
  - 기록이 중요하다는 것을 배웠습니다. 기록은 서로의 중복된 실험을 막아줬으며, 실패를 통해서 서로 배울 수 있었습니다.
- 시각화
  - 모델의 문제점이 명확히 드러났습니다.
  - 모델의 성능을 대략적으로 예측해볼 수 있어 제출횟수를 유용하게 사용할 수 있었습니다.



나이대별 성별 분포 및 class별 분포



confusion matrix를 통한 validation set에 대한 시각화



퍼블릭 테스트 데이터에 대한 모델 시각화