(1페이지)

안녕하십니까 빅데이터 처리 및 응용 1조 팀프로젝트 발표를 맡게 된 정보융합학부 정호빈입니다.

저희 주제는 전자상거래 데이터를 이용하여 추천시스템과 ANOVA분석을 구현해 보는 것입니다.

(2페이지)

목차는 다음과 같습니다. 데이터 소개를 시작으로 데이터 전처리 과정, ANOVA 분석 결과, 추천 시스템 구축 결과에 대해 설명드리겠습니다.

(3페이지)

조원 및 조원들의 역할은 다음과 같습니다.

(4페이지)

데이터를 소개해 드리자면 데이터는 실제 전자상거래에서의 사용자 행동데이터를 사용하였습니다. 데이터 크기는 14기가 정도이며 변수는 9개로 적지만 관측치는 1억개 이상

입니다.

데이터의 특징은 다음과 같습니다. 우선 전자상거래에서 쌓이는 데이터들이 변수이며, 그 중에는 numeric 변수 뿐만 아니라 categorical 변수도 존재합니다. 또한 전처리가 되어있지 않아 전처리가 필요한 데이터입니다.

저희 프로젝트의 목표는 다음과 같습니다. 첫째로는 ANOVA분석으로 사용자의 이벤트 수에 따라 평균구매가격의 차이가 존재하는지를 파악하는 것이고 두 번째로는 추천시스템 모델로 사용자가 구매하지 않은 물건 중 사용자가 좋아할만한 상품들을 추천하는 것입니다.

(5페이지)

데이터 전처리 과정에 대해서 살펴보겠습니다. 저희가 사용하는 데이터가 워낙 대용량이라 전처리를 진행하는것이 매우 어려웠고 시간도 많이 소요 되었습니다. 특히 메모리를 심하게 잡아먹어 해결하는데 많은 시간이 걸렸습니다.

먼저 결측치를 행 기준으로 삭제하였고 NAN값은 N으로 변경하여 프로그램이 인식하게 처리하였습니다. 그 다음 Category\_code에서 “.”을 기준으로 스플릿을 하여 각 특성으로 새로운 범주를 생성하였습니다.

(6페이지)

이후로 유효한 열을 선정하여 label encoding을 하였고 이로 인해 각 변수 간의 관계가 긴밀해져 학습이 용이하게 되었습니다. label encoding한 정보는 dict 형식으로 저장하여 두개의 csv에서 동일한 값을 동일한 label로 치환하거나 신규값이 있다면 해당 dict를 추가하게 하였습니다. 이를 바탕으로 마지막에 label encoding으로 숫자로 바뀐 값들이 다시 원래의 변수명으로 표시되게 하였습니다. 마지막으로 두 개의 전처리한 csv파일을 join하여 합쳤습니다. 결과적으로 위와 같은 전처리를 통해 14GB에 가까운 데이터를 8GB로 줄여 40% 이상 효율적으로 만들어 뒤에서 진행하는 ANOVA분석과 ALS추천모델을 전처리 전보다 빠르고 정확하게 진행 할 수 있었습니다.

(7페이지)

다음은 앞서말한 category를 스플릿하여 나온 새로운 범주이고,

(8페이지)

다음은 레이블 인코딩을 한 데이터셋 사진입니다.

(9페이지)

이제 ANOVA분석 한 것에 대해 설명드리겠습니다.

ANOVA 혹은 분산분석이란 세 개 이상 다수의 집단의 모평균을 서로 비교하고자 할 때 집단 내 분산과 집단 간 분산의 비가 따르는 F분포를 이용하여 가설검정 하는 방법입니다.

ANOVA의 한가지 예시를 들자면 한국 일본 중국 각 국가 간의 학습기술에 따른 성적비교를 하는 것입니다. 이 경우 세 개 이상의 모평균을 비교해야 하므로 ANOVA분석을 사용할 수 있습니다.

(10페이지)

저희들은 이번 프로젝트에서 One-way ANOVA를 파이스파크로 구현하였습니다. 다음 사진은 ANOVA 테이블과 아노바 테이블에 나온 값들을 구하는 공식들입니다.

(11페이지)

각 사용자에게서 발생한 이벤트 수의 수준을 사분위 수 기준으로 나누어 얻은 4개의 사용자 그룹간 구매가격의 모평균이 유의미한 차이를 보이는지 확인해 보았습니다.

레벤의 등분산 검정, 사후 분석은 통계패키지를 이용하면 가능하긴 하지만 구체적인 동작방식이 학부생 수준을 넘어서는 것이라 pyspark로는 구현하지 못하였습니다.

(12페이지)

우선 귀무가설과 대립가설을 세우고 검정을 시작하도록 하겠습니다. 귀무가설은 event수에 따라 각 그룹간 구매가격에 차이가 없다이고 대립가설은 적어도 한 쌍의 그룹에서 구매가격에 차이가 있다라고 설정하였습니다. 검정통계량은 집단내 분산과 집단간 분산의 비율로 계산할 수 있고 이를 기반으로 p-value를 계산하여 유의수준과 비교하였습니다.

(13페이지)

이제 ANOVA를 위해 어떤 작업들을 했는지 살펴보겠습니다. 우선 전처리된 데이터로부터 쿼리문을 적절하게 사용하여 event수와 평균구매가격을 담은 데이터프레임을 추출하였으며 이후 중복되는 열들은 삭제하였습니다.

(14페이지)

이후 event수에 대해 udf를 새로 정의하여 사분위수의 관계에 따라 4가지로 나누었습니다.

이후 직접구현한 one way anova함수에 데이터 프레임을 넣었습니다.

(15페이지)

다음은 앞서말한 과정을 더 와닿게 사진을 첨부한 것입니다.

(16페이지)

one-way anova함수는 데이터 프레임과 그룹을 나타내는 열 이름, 구매가격을 나타내는 열이름을 입력받아 집단내 제곱합, 집단간 제곱합, 검정통계랑, 각 자유도를 반환하는 함수로 이루어져 있습니다. 분석을 한 결과 p-value값이 상당히 작아 그룹간 평균의 차이가 유의미 하다는 것을 알 수 있었으며 따라서 귀무가설을 기각할 수 있었습니다.

(17페이지)

직접 구현하지는 않았지만 사후분석을 실시하여 어떤 그룹에서 유의미한 차이가

있나 확인할 수 있습니다. 이 분석결과는 회사의 입장에서 기존 유저와 신규 유저 중

평균가격이 낮은 쪽을 위한 프로모션 및 홍보 진행을 위한 근거로 유용하게 사용될

수 있을거라고 생각됩니다.

(18페이지)

다음은 추천시스템에 대해 설명드리겠습니다. 저희 팀은 ALS 추천 시스템 모델을 적용하여 특정 사용자에게 선호할 만한 제품을 추천하는 함수를 새로 구현하였습니다. 추천시스템을 이용하기 위해 이벤트 타입별 가중치를 설정하였습니다. 이벤트 타입은 view, cart, purchase로 각각 가중치를 0.1 0.2 0.3를 주었습니다.

(19페이지)

10월 11월 데이터를 합친후 결측치를 제거하였으며 불필요한 열을 제거하여 데이터 프레임을 생성하였습니다.

(20페이지)

다음으로 유저별 각 제품에 대한 가중치 총점을 구하여 새로운 데이터 프레임을 생성하였고 기존의 데이터프레임에 존재하는 유저 아이디, 프로덕트 아이디가 중복된 것들을 제거하였습니다. 이후 유저아이디와 프로덕트 아이디가 같은것끼리 조인하였습니다.

(21페이지)

다음은 조인한 데이터 프레임의 사진입니다.

(22페이지)

모델을 학습시킬때는 유저아이디와 프로덕트 아이디 토탈 스코어값만 포함된 데이터 프레임을 새로 생성하여 진행하였습니다. ALS추천모델을 만들고 결과를 RMSE 지표를 사용하여 평가하였습니다. RMSE는 0.445 정도로 준수한 결과를 보였습니다.

(23페이지)

이후 제품 아이디 데이터 프레임과 특정유저가 본 제품관련 데이터 프레임을 생성하여 이 둘을 프로덕트 아이디를 기준으로 조인하여 이 사용자가 보지 못한 제품들을 추출하였습니다.

이후 만들어진 ALS모델에 적용시켜 총점을 예측하고 내림차순으로 나타내었습니다.

(24페이지)

다음사진은 실제로 유저에게 추천할 제품을 나타낸 것입니다.

(25페이지)

ALS모델을 구축하면서 아쉬웠던 점으로는 제품과 유저간 상호작용이 있었다면 더 좋은 결과를 보였을 것이라고 생각하고 이 ALS모델을 통해 각 유저의 선호도를 파악해 이익창출과 전자상거래 이용자 들에게 도움을 줄 수 있을것이라고 생각합니다.

(26페이지)

저희 프로젝트를 간단히 요약해보자면 ANOVA테스트 한 결과 이벤트 수에 따라 평균 구매가격의 모평균 차이가 유의미하다는 것을 알 수 있었습니다. 이 분석결과는 전자상 거래 사이트를 운영하는 회사 입장에서 기존 유저와 신규유저 중 평균구매가격이 낮을 쪽을 위한 프로모션이나 홍보를 진행하기 위한 근거로 유용하게 사용할 수 있을것이라 생각됩니다.

(27페이지)

또한 ALS추천모델을 적용하여 특정 유저가 선호할 만한 제품을 추천하는 함수를 직접 구현하였으며 이 추천시스템을 통해 각 유저들에 대한 선호도를 파악하여 전자상거래 사이트 운영자들의 이익창출과 이를 이용하는 유저들에게도 도움이 될 것이라고 생각합니다.

(28페이지)

이상으로 1조 발표 마치겠습니다. 감사합니다.