**기계학습기초 프로젝트 발표**

**[1 page 표지]**

중간발표 이후 주제를 다시 생각해보고 기존의 머신러닝 모델과 직접 구현한 모델을 활용할 수 있는 방안으로 변경하였습니다.

**[2 page 목차]**

주제, 데이터, 모델, 보완점, 결론 순서로 발표를 진행하겠습니다.

**[4 page 주제]**

다양한 e-커머스 플랫폼이 존재하는 가운데 소비자는 자신의 구매 성향에 맞는 플랫폼을 선택한다. 기업은 이러한 소비자의 행동 분석을 통해서 이익의 증대를 목표로 한다.

기업은 어떤 소비자들이 충성 고객이 될 것인가 예측을 할 수 있다면 이익의 증대를 기대할 수 있다. e-커머스 소비자의 행동 데이터를 통해 재 방문 빈도에 따른 고객의 충성도를 예측하는 것을 주제로 삼았습니다.

**[6 page 데이터수집1]**

캐글의 e-커머스 사용자의 행동 데이터를 사용했으며 2019년 10월부터 2020년 4월까지 7개월간의 데이터를 수집하였습니다.

**[7 page 데이터수집2]**

소비자의 행동을 하나의 이벤트로 분류한 데이터로

이벤트 타임은 이벤트가 발생한 시간, 이벤트 타입은 view, cart, remove\_from\_cart, purchase 4가지로 구성된 발생한 이벤트의 종류를 나타냅니다.

품목 id를 나타내는 product\_id의 경우 각 id에 해당하는 품목정보가 없어 제거하였습니다.

category\_id 이벤트가 발생한 품목의 카테고리에 대한 고유 id, category\_code 품목에 대한 카테고리로 메인 카테고리와 최대3개의 서브 카테고리로 구성되어 있습니다.

category\_id는 category\_code를 통해 구분이 가능하기 때문에 제거

brand는 판매 기업이며 확인결과 상위 몇몇 기업에 대한 편향이 심해 제거하였습니다.

user\_id는 각 소비자에 따른 고유한 id, user\_session은 특정 소비자가 접속한 세션에 따른 고유한 값으로 소비자가 접속이후 시간이 지나고 다시 접속하게 되면 기존의 세션은 다른 고유값으로 초기화됩니다.

**[8 page 데이터수집3]**

e-커머스를 사용하는 소비자의 view, cart, remove\_from\_cart, purchase 4가지 행동에 대한 데이터를 저희가 수립한 주제인 한 번이라도 구매한 고객에 대해 재방문을 예측하는 것, 즉 소비자의 충성도 예측에 맞춰 purchase에 해당하는 데이터를 추출, 사용하였고 이는 전체 데이터의 약10%에 해당합니다.

결측치는 제거하여 사용하였으며 총 6백8만7백6십1개의 행으로 구성되어있습니다.

**[9 page 데이터분석-카테고리]**

카테고리는 초기 데이터에서 메인 카테고리와 최대3개의 서브 카테고리로 분류되어 있습니다. 이를 구분하여 분포를 확인해본 결과 분포가 매우 지엽적으로 메인 카테고리에 따라 총 13개의 카테고리로 구분하였습니다.

**[10 page 데이터분석-이벤트타임]**

이벤트가 발생한 시간을 나타내는 이벤트 타임을 이용하여 시간단위, 요일단위, 주중 또는 주말 단위로 구분하였습니다.

시간의 경우 4개의 구간으로 나누어 사용했습니다.

**[11 page 데이터분석-원핫코딩]**

카테고리와 시간, 요일, 주중주말에 해당하는 범주형 데이터에 대해 원 핫 코딩을 수행했습니다.

**[13 page 데이터분석-유저세션1]**

저희 주제는 한 번이라고 구매한 소비자의 재방문에 따라 소비자의 충성도를 예측하는 것으로 각 소비자의 재방문 빈도와 구매 이력에 대한 데이터가 필요합니다. 따라서 각 소비자의 고유한 id와 동일 접속을 나타내는 user\_session을 이용하여 각 소비자의 session별 구매이력으로 데이터를 조작했습니다.

**[14 page 데이터분석-유저세션2]**

event\_time을 이용하여 소비자의 session순서를 알 수 있고 user\_id를 통해 소비자의 총 접속횟수 total\_session을 구할 수 있었습니다.

**[15 page 데이터분석-구매횟수]**

각 소비자의 session별 구매이력으로 조작한 데이터를 통해 session별 구매 횟수를 구하고 분포확인을 통해 이상치를 제거하였습니다.

**[16 page 데이터분석-구매금액1]**

session별 총 구매 금액을 구하고 분석하여 역시 이상치를 제거하였습니다.

**[17 page 데이터분석-구매금액2]**

price의 경우 다른 특성과 스케일의 차이가 커 정규화를 진행하였습니다.

[참고]

< Z-점수 정규화, 이상치(outlier) 문제를 피하는 데이터 정규화 전략>

(값 - 평균)/(표준편차) 평균이면 0

**[18 page 데이터분석-로열티]**

앞서 구한 소비자의 총 접속 횟수 total\_session에 따라 소비자의 충성도를 나타내는customer\_royalty로 변경하였습니다. 총 5개의 구간으로 설정하였고 각 구간의 경계는 구간 분포를 고려하였습니다.

**[19 page 데이터분석-요약]**

분석과 전처리과정을 통해 소비자 행동 데이터에서 구매행동에 해당하는 데이터를 추출 후 결측치를 제거, 범주형 데이터를 구간구분이후 원 핫 코딩하였습니다.

소비자의 행동 별 데이터를 각 소비자의 session별 구매이력에 관한 데이터로 조작하여 총 구매횟수에 따라 저희가 예측하고자 하는 목적변수 customer\_royalty를 결정하였습니다.

이러한 과정을 거쳐 총 4백4십3만5백3십5개의 행을 가진 데이터를 학습에 사용했습니다.

[참고]

초기 데이터에는 우리가 구하고자 하는 목적변수가 없었다. 따라서 초기 데이터를 분석하고 과정을 거쳐 고객의 충성도라는 목적변수를 정의하였다.