안녕하세요. 저희는 MOMA팀의 발표를 맡은 김수빈, 박기정입니다.

발표를 시작하겠습니다. 목차는 다음과 같습니다.

1.개요

1)분석 목적 및 필요성 (p3)  
왼쪽 그림은 ai 고객데이터 통합플랫폼 다이티 내 칼럼에서 추출한 내용입니다. 이는 전체 고객 수가 같더라도 반복 구매고객과 VIP 고객이 많아질 경우 전체 매출 볼륨이 커짐을 의미합니다.

오른쪽 그림은 온라인 쇼핑 멤버십 트렌드 리포트의 내용을 발췌한 것입니다.

이에 따르면 온라인 쇼핑 멤버십 이용자는 비가입자대비 해당 쇼핑몰 방문 빈도 및 지출 비용이 더 높은 것을 확인할 수 있습니다.

따라서 CJ 더 마켓의 ‘the 프라임’ 가입 고객 수를 증가시켜 반복 구매자 및 vip 고객을 확보하고 궁극적으로 매출의 증대를 꾀하고자 본 분석을 진행했습니다.

2)데이터 파악 (p4)

데이터 파악 결과 다음과 같습니다.

3)가설설정 (p5)

본 분석의 궁극적인 목적은 the 프라임 가입 고객 수를 증가시키는 것입니다. 이를 위해서는 prime 고객만이 가지는 특성을 파악하여야 하므로 “일반 회원과 임직원의 프라임/비프라임 회원 간 구매 특성에 **뚜렷한** 차이가 있을 것이다”라고”(다음과 같이) 가설을 설정했습니다.

1:08

2. 데이터 전처리

1)eda - employee\_yn, prime\_yn (p6)

employee비율과 prime비율을 확인하기 위해서 다음과 같이 시각화를 통해 확인했습니다.

1) eda -이상치 확인 (p7)

양적변수인 net\_order\_qty 와 net\_order\_amt의 boxplot을 그려서 분포를 확인했습니다.

1) eda - 이상치 파악(p8)

구체적으로 잠재적 이상치 후보개수를 파악하기 위해 각 데이터에서 z-score 3이상인 값과 1.5IQR 및 3IQR 기준으로 확인해본 결과 다음과 같습니다. 이상치후보 값을 직접 확인해 본 결과 대부분 설 선물세트였고, 이와 같은 현상이 많이 나타나는 이유는 설이라는 명절 특성 상 선물 세트를 여러 개 구매하는 고객층으로 인한 것으로 볼 수 있습니다. 본 분석에서 TRAIN, TEST데이터 모두 설이 포함된 데이터이므로 이를 이상치로 판단하지 않고 분석에 사용했습니다.

여기서부터는 프라임회원 여부 예측을 중심으로 각 컬럼을 분석한 결과를 말씀드리겠습니다.

다음 나오는 두개의 컬럼은 분석결과 프라임 회원 예측에 유의미하지 않아 모델에 제외되었으므로 시간관계상 제출한 ppt와 코드를 참고해주시면 감사하겠습니다.

2)order\_date -> holiday(p9)

order\_date를 이용해서 날짜가 비영업일이면 1 영업일이면 0 값을 갖는 파생변수 holiday를 생성했습니다. 그리고 다음과 같이 임직원과 고객 데이터 각각에서 프라임회원과 비프라임회원과의 차이를 확인해 본 결과 같은 경향을 보였습니다. 본 분석은 프라임 회원예측에 목적을 두고 있으므로 유의미한 차이가 보이지 않아 holiday변수는 모델에 포함시키지 않았습니다.

3-1) product\_name -> set\_yn(p10)

다음으로 product\_name을 사용하여 set\_yn이라는 파생변수를 생성했습니다.

unique한 product\_name에 다음과 같이 x2개처럼 묶음 상품인 경우면 1 개별상품이면 0으로 값을 할당했습니다.

set\_yn 시각화 (p11)

임직원과 고객 데이터 각각에서 묶음상품 여부의 비율에 따른 프라임과 비프라임회원을 비교한 결과, 차이가 뚜렷하지 않아 유의미하다고 판단하기엔 무리가 있으므로 set\_yn 변수를 모델에 포함시키지 않았습니다.

3-2) product\_name -> category (p12)

product\_name 컬럼의 unique한 상품은 3104개 였습니다. 하지만 프라임과 비프라임 회원의 더욱 뚜렷한 구매성향 차이를 관측하기 위해서는 상품군을 기준으로 분석하는 것이 합리적이라는 생각하에 cj 더 마켓 홈페이지 카테고리란을 기준으로 product\_name 컬럼을 이용해 다음과 같이 category 파생변수를 생성했습니다.

category - Apriori 분석 (임직원) (p13)

카테고리 변수가 프라임 회원 예측하는데 도움이 되는 지를 확인해보기 위해 먼저 임직원데이터에서 Apriori분석을 진행했습니다.

왼쪽사진은 prime회원에 대한 분석 결과, 오른쪽은 비프라임회원에 대한 결과입니다.

빨간색 박스를 보시면 prime회원의 경우 카테고리 3,4,6,7 에 속하는 상품을 구매할 때 카테고리 2에 속하는 상품을 구매하는 경향이 있음을 알 수 있는 반면 비프라임회원의 경우 카테고리 1번에 속하는 상품을 구매하는 경향을 보입니다.

같은 방식으로 파란색 박스와 노란색 박스 각각 비교해보면 같은 결과로 귀결됨을 확인할 수 있습니다.3:46

category - Apriori 분석 (고객) (p14)

마찬가지로 고객데이터에서 프라임회원과 비프라임회원 간 뚜렷한 구매성향 차이가 존재함을 확인할 수 있습니다.

따라서 임직원데이터와 고객데이터에서 category가 프라임회원 여부 예측하는 데 도움된다는 판단 하에 모델에 포함시켰습니다.

4) scd -> scd\_count

프라임 회원은 cj 더마켓을 이용해서 주로 장을 볼 것이기 때문에 여러 상품을 구매할 것이고, 비프라임 회원은 특정 상품이 타 쇼핑몰 대비 가격 경쟁력이 존재할 때 구매할 것입니다. 따라서 프라임/비프라임 간 구매 폼목의 개수 차이가 존재한다는 판단 하에 scd 컬럼을 이용하여 하나의 주문번호에 따른 구매 품목의 개수를 의미하는 scd\_count를 만들었습니다. 확인 결과 employee와 customer 모두 product의 수가 적을 때 비프라임 회원의 비율이 높고, 반대의 경우 프라임 회원의 비율이 높게 나타나 프라임 예측에 도움이 될 것으로 보여 모델에 사용하였습니다. 4:34

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

5) net\_order\_amt (기정 시작)

프라임 회원이 비프라임 회원에 비해 혜택이 많으므로 주문 금액이 낮을 것이라는 생각 하에 상품의 한 단위 당 가격 데이터를 얻고자 net\_order\_amt에 어떠한 스케일링이 적용되었는 지 예측을 진행하였습니다.

이를 위해 net\_order\_amt의 분포 확인 결과 모두 정규분포와 유사한 형태가 나타났고, 범위가 0과1 사이가 아니므로 min-max와 max-abs는 제외하였습니다.

그리고 net\_order\_amt는 최종 주문금액 데이터이므로 net\_order\_qty에 영향을 받을 수밖에 없고 qty가 왜도가 매우 크게 나타났으므로 amt 또한 그럴 것이라 생각했습니다.

Standard와 Robust의 경우 스케일러 특성 상 분포의 형태 자체를 변화시킬수는 없으므로 이 또한 제외했습니다.

반면, exponential 적용 결과 qty와 유사하게 왜도가 매우 큰 분포가 나타난 것으로 보아 로그 변환이 유력하다고 판단했습니다.

추가적인 근거로 amt에 exponential을 적용하고, 이를 qty로 나누어서 price\_per\_unit라는 파생변수를 얻은 다음 해당 상품의 실제 가격과 비교한 결과 가격이 거의 일치하는 데이터가 상당 부분 존재했습니다. 따라서 로그 변환 가정에 대한 근거가 충분하므로 price\_per\_unit을 상품 한 단위 당 가격으로 간주했습니다.

실제로 프라임 회원이 비프라임 회원에 비해 구매 금액 평균이 낮은 것으로 나타났고, 따라서 price\_per\_unit을 모델에 사용하였습니다.5:49

6) gender & age\_group

gender와 prime\_yn과의 관계를 시각화 해 본 결과 employee와 customer 모두 남성의 프라임 회원 비율이 높은 것으로 나타났고, 프라임 예측에 도움이 될 것이라는 판단 하에 모델에 사용하였습니다.

age\_group은 (prime\_yn, age\_grp, category) 그룹화를 통해 연령대별로 프라임에 따라 구매 품목에 차이가 있는지를 확인하였습니다.

그 결과 employee와 customer 모두 동일 age 그룹 내에서 프라임 여부에 따라 구매 품목별 차이가 있는 것으로 나타나 유의미하다고 판단하여 모델에 사용하였습니다.

3. 모델링

1) 모델 선택

위 내용을 종합한 결과 모델에 사용할 변수는 위와 같습니다.

그리고 다양한 분류 모델들을 디폴트 상태로 적합한결과, employee와 customer 에서 모두 f1 스코어가 가장 높은 XGBoost를 최종 모델로 선택하였습니다.

2) 하이퍼 파라미터 튜닝

XGBoost의 하이퍼 파라미터 튜닝 과정입니다.

학습의 효율성을 위해 0.1의 높은 learning\_rate와 1000번의 반복 수행횟수로 시작하여 다음과 같은 과정을 따랐고, 최종적으로 저희의 목표 f1\_score인 0.9에 맞추어 모델을 완성하였습니다.

다음은 employee의 최종모델입니다.

다음은 customer의 최종모델입니다. 7:15 ->6:49

4. 결과

1) f1 score 기반 모델 결과 설명

각 모델 적합 결과 평가 지표 값은 다음과 같습니다.

그 중 f1 스코어는 employee에서 0.8912, customer에서 0.9027이 나왔고, 이 모델들을 이용해서 test 데이터에 대해 전달받은 f1 스코어는 0.8158이었습니다

2) 예상 기대효과

분석 결과 예상되는 기대효과는 각 컬럼 및 파생변수 별로 화면 내용과 같고, category를 이용해서 프라임 회원 확보를 위한 마케팅적 방안을 수립할 수 있습니다.

먼저 임직원 전용 상품이 있듯이, 프라임 회원이 주로 함께 구매하는 카테고리를 묶어서 프라임 회원 전용으로 판매한다면, 기존 프라임 회원의 이용 빈도를 늘릴 수 있을 것입니다.

또, 비프라임 회원이 주로 같이 구매하는 카테고리를 묶어서 프라임 회원 전용으로 판매한다면, 비프라임 회원을 프라임 회원으로 끌어들이는 효과를 얻을 수 있을 것입니다.

추가로 카테고리 연관성 분석 결과를 활용하여 광고에 이용하는 방안도 있습니다. 예를 들어 만약 프라임 회원이 1,4,7 카테고리를 같이 구매한다면, 1번 카테고리 상품을 구매하려 할 때 4,7번 카테고리 상품을 옆 배너에 띄우는 전략입니다.

앞서 제시한 전략적 마케팅을 활용한다면 기존 프라임 고객을 유지하면서 새로운 프라임 회원 또한 확보할 수 있을 것이고, 이는 본 분석의 목적인 반복구매 고객의 수의 증가와 매출 증가로도 이어질 수 있을 것으로 기대됩니다. 8:07

—------------------------------------------------------------------------------------------------------------------------

태클 예상 질문

1. 왜 employee와 customer에 같은 변수들을 추가했나
2. apriori 해석
3. category 처리를 왜 python 코드로 짜서 하지 않았느냐
4. 각각의 컬럼을 생각해낸 이유?

(learning rate 0.1시작에 대한 태클 대비)

XGBoost에서 하이퍼 파라미터 튜닝 시에는 0.1~0.3 정도의 값을 사용하고, 최종 모형 학습 시에는 0.05이하의 값을 사용하는 것이 일반적입니다.

그리고 만약 learning\_rate를 지나치게 크게 잡았을 때 모형 적합이 잘 되지 않는다면, 이 값을 기준으로 튜닝한 다른 하이퍼 파라미터의 조합이 learning\_rate를 낮추었을때는 최적 조합이 아니게 될 가능성이 높습니다.

따라서 0.1~0.3 중 가장 낮은 값인 0.1로 하이퍼 파라미터 튜닝을 진행하였고, 이후 0.05 이하의 값으로 낮추는 방법을 사용하였습니다.

(최종 learning rate를 0.01과 0.015로 설정한 이유에 대한 태클 대비)

구글링 결과 train데이터와 test데이터의 분포가 유사하고, 암 발생여부 예측과 같이 타겟의 분포 및 확률이 치우쳐져 있는 경우가 아니라면, train에서의 f1스코어가 0.85~0.9 정도라면 test에서 또한 0.85정도 나온다는 정보를 얻었습니다. 그리고 f1스코어가 0.95정도로 높아진다면 일반적으로 overfitting으로 인해 모델의 예측력이 감소한다는 정보 또한 얻었습니다. 따라서 이를 토대로 learning rate를 0.01로 내려본 결과 customer에서 f1스코어가 약 0.9가 나왔고, 이에 맞추기 위해 employee를 0.015로 다시 올려서 약 0.89라는 결과를 얻었습니다.