[2023 BDA 데이터 분석·활용 공모전]

팀명: 데마 (김수빈, 백지연, 추은서)

참여 트랙	**Track 2: 모델링 고도화**
분석 목적 및 필요성	분석의 최종 목적에 부합하고, 꼭 필요하다고 생각되는 변수, 파생 변수들을 채택하여 모델링에 사용할 예정이다. <기존 변수> 1. scd (주문번호) 2. net_order_amt (주문 금액) 3. gender(성별) 4. age_grp(나이) 5. employee_yn(임직원 유무) 6. order_date(주문 날짜) <파생 변수> 1. non_main_yn(비주력 상품 여부) -> product_name(상품명)에서 파생된 변수. 비주력 상품의 구매 여부는 프라임 회원의 구매 행태를 반영한다고 판단하여 생성한 변수 2. net_order_type(동시 주문 상품 개수) -> 동일한 주문번호 기준, 구매한 상품의 가짓수를 나타내는 변수. 프라임/일반 회원 간 동시 주문 상품 개수가 다르게 나타나겠다고 판단하여 생성한 변수 3.event_yn(설날/이벤트 기간 내 주문) -> 상품 구매 시기가 이벤트 기간 내 포함되어 있는지를 나타내는 변수. 프라임/ 일반 회원 간 이벤트의 영향이 다르게 나타나겠다고 판단하여 생성한 변수 <선택되지 않은 변수> 1.product_name(상품명) -> 상품명을 그대로 반영하는 것은 의미 없다고 판단하여 non_main(비주력 상품 여부) 변수를 생성한 뒤, 최종 모델링에서는 제외 2.net_order_qty(주문 수량) -> net_order_qty 는 대소 비교에 사용하기 적절하지 않아 최종 모델링에서는 제외

가설 수립	중요하다고 생각되는 변수를 토대로 가설을 수립하였다. 1. 연령대는 프라임 회원 여부에 영향을 줄 것이다 2. 성별은 프라임 회원 여부에 영향을 줄 것이다 3. 프라임 회원들은 일반 회원들보다 CJ 더 마켓에서 비주력 상품들을 더 적극적으로 구매할 것이다. 4. 프라임 회원은 일반 회원보다 동시 주문 상품 종류의 개수가 많을 것이다. 5. 일반 회원은 프라임 회원보다 할인 및 이벤트 기간 내에 주문하는 비중이 클 것이다. 6. 프라임 회원은 일반 회원보다 주문 금액이 더 클 것이다.
예상 기대효과	성공적인 유료 멤버십을 통해 수익을 극대화하기 위해서는, 기존 멤버십 가입자들의 유지뿐만 아니라 새로운 멤버십 가입자를 확보해야 한다. 멤버십 가입자가 될 잠재력을 지닌 고객을 예측하기 위해서는 일반 회원들과 차별화되는 멤버십 회원들의 특징에 대한 심도 있는 분석이 필요하다. 본 팀에서는 멤버십 회원과 일반 회원을 구분 짓는 데 가용한 변수를 최대한으로 고려하고자 노력하였다. 또한 본 팀에서 제안하는 파생 변수들은 새로운 프라임 멤버십 회원 확보에 효과적으로 사용될 수 있다. 타당한 근거를 바탕으로 변수를 생성하고자 노력하였기 때문에, 기존 멤버십 회원을 예측하는 것뿐만 아니라 새로운 멤버십 회원을 유치하는 데 큰 도움이 될 것이다.

1. 분석 목적 및 필요성

1) 데이터 예시

하단의 표는 최종 모델링에 사용하게 될 Data Frame의 예시이다.

(1) group by 이전

scd(주문번호)	non_main_yn	net_order_amt	gender	age_grp	employee_yn	order_date	event_yn
20230124153976	1	0.523	F	30	N	20230102	0
20230124153976	0	0.523	F	30	N	20230102	0

(2) group by 이후

scd(주문번호)	non_main	net_order_amt	net_order_type	gender	age_grp	employee_yn	order_date	event_yn
20230124153976	1	1.046	2	F	30	N	20230102	0

2) 기존 변수

(1) scd (주문번호)

유니크한 번호로서, 동일한 주문번호를 가진 데이터는 같은 고객이 동시에 주문한 데이터로 간주한다.
- 한 명의 고객이 동시에 주문한 상품들을 하나의 행으로 나타내기 위한 기준 변수로 사용한다.

(2) net_order_amt (주문 금액)

배송비나 쿠폰 할인 등이 모두 적용된 최종 주문 금액 데이터이며, 스케일링 처리가 된 상태이다.
- 주문번호를 기준으로 개별 고객별 group by를 적용하여 총주문 금액을 계산하기 위해 사용한다.

(3) gender (성별)

남성일 경우 M, 여성일 경우 F로 표기되는 고객의 성별 데이터이다.

(4) age_grp (나이)

10대부터 50대까지 숫자로 표기되는 고객의 나이 데이터이다.

(5) employee_yn (임직원 유무)

임직원일 경우 Y, 임직원이 아닐 경우 N으로 표기되는 임직원 유무 표기 데이터이다.
- 임직원 모델과 비 임직원 모델을 각각 만들기 위해 데이터셋을 분리할 때 사용한다.

(6) order_date (주문 날짜)

고객이 해당 제품을 주문한 날짜로, 연도와 달, 일 정보가 모두 포함된 데이터이다.
- CJ 더마켓에서 진행하는 이벤트 기간 내에 주문한 상품인지 파악하기 위해 사용한다.

3) 파생 변수

(1) non_main_yn (비주력 상품 여부)

product_name (상품명)이 CJ 더마켓의 비주력 상품인지 여부에 대한 정보를 담은 데이터이다.
- CJ 더마켓 홈페이지에 기재된 카테고리를 기준으로 밥/죽/면, 국/김치/김/반찬/두부, 만두/피자/치킨, 핫도그/떡

볶이/간식, 돈까스/함박/구이, 스팸/닭가슴살/소시지, 밀키트는 CJ 더마켓의 주력 상품인 가정간편식이므로 '주력 상품'으로 구분하며, 그 외의 양념/소스/가루/오일, 건강식품, 신선식품 그리고 음료/생수/시럽은 '비주력 상품'으로 구분한다.
- 비주력 상품의 경우 1로, 주력 상품의 경우 0으로 표기하며, 주문 번호를 기준으로 group by 후 하나의 데이터(행)으로 통합한 이후에는 non_main (비주력 상품 개수 or 비율) 변수로 대체될 것이다.
- non_main은 추후 원본 데이터를 살펴본 후 전체 데이터 중 비주력 상품이 차지하는 비중에 따라 변수 사용 목적에 맞는 전처리를 추가로 적용할 예정이다.

(2) net_order_type (동시 주문 상품의 개수)
동일한 주문 번호를 기준으로 같은 고객이 동시에 주문한 상품의 개수에 대한 데이터이다.
- 동일한 주문 번호를 가진 데이터는 모두 같은 net_order_type 데이터를 가지게 된다.

(3) event_yn (설날/이벤트 기간 내 주문)
CJ 더마켓에서 진행하는 설맞이 설 선물 대전이나 더마켓 세일 페스타 등의 이벤트 기간 내에 상품을 주문했는지 여부를 표기하는 데이터이다.
- 설날/이벤트 기간 내에 주문한 데이터의 경우 1, 주문하지 않은 데이터의 경우 0을 부여한다.
- order_date (주문 날짜)를 기준으로 하며, product_name (상품명)에 이벤트 관련 상품임이 표기된 경우도 포함한다.

4) 선택되지 않은 변수
(1) product_name (상품명)
non_main_yn (주력상품 여부) 변수 생성에 사용한 후 최종 모델링에서는 제외한다.
- 추후 데이터를 동일한 주문 번호를 기준으로 group by 후 하나의 데이터(행)으로 통합하여 사용할 예정이기에 다양한 상품명이 담겨 있는 product_name 변수를 그대로 모델링에 적용하는 것은 적절하지 않다고 판단하였다. product_name은 파생 변수인 non_main_yn을 통해 비주력 상품의 주문 개수 및 비중에 대한 정보를 담은 적절한 변수로 치환하여 더 효과적으로 사용할 것이다.
- product_name에 포함된 단위별 용량(중량) 정보의 활용 가능성에 대해서도 논의해 보았으나, 모델링에 사용하기에 부족하다고 판단하였다. 최근 경기 불황과 고물가가 지속되면서 대용량 제품을 싸게 구매하는 소비 행태가 확산되고 있다. G마켓이 2023년 1월부터 2월까지의 상품 거래액을 작년 동기간과 비교한 결과, 대용량 제품의 전체 신장률은 12%를 기록했으며 그중에서도 장기간 보관이 용이한 가공식품이 전년 동기 63% 증가하였다.[1] CJ더마켓은 CJ제일제당의 식품 전문몰로서 가정간편식 등의 가공식품을 주력으로 하고 있기 때문에 이러한 소비 트렌드에 더욱 큰 영향을 받을 것으로 보인다. 대용량 제품 소비 트렌드는 공통으로 적용되기 때문에 멤버십 회원인지 판별하는 데에 영향 변수로 사용하기엔 어려울 것이다.

(2) net_order_qty (주문 수량)
net_order_qty은 대소 비교에 사용하기 적절하지 않아 최종 모델링에서 제외한다.
- net_order_qty을 기준으로 많고 적음을 고려하게 된다면 대용량 제품이나 건강식품 등 고가의 제품을 1개 주문한 경우보다 저가의 제품을 5개 주문한 경우를 많음으로 간주하게 되는데, 이는 일반적인 모델링에서 기대하는 대량 구매로 단정하기엔 부족하다.
- net_order_qty에서 제공하는 대소에 대한 정보는 net_order_amt (주문 금액)에 포함되어 있으며, 오히려 net_order_amt이 제품의 용량 및 금액 등이 반영된 총체적인 많고 적음에 대한 정보를 제공하고 있으므로 net_order_qty을 net_order_amt으로 대체하는 것에 큰 무리가 없을 것으로 보인다.

[1] 'CJ더마켓' 1주년, 식품 전문몰 도약 박차…매출·가입자수 급증, 전자신문 etnews, 2020년 7월

2. 가설 설정

1) 일반 회원 (비 임직원) 모델링

(1) 연령대는 프라임 회원 여부에 영향을 줄 것이다.

오픈서베이의 온라인 식료품 구매 트렌드 리포트에 따르면, 20대의 18.2%, 30대의 24.6%, 40대의 21% 그리고 50대의 12%가 식료품 구매 시 오프라인보다 온라인 채널을 선호한다.[2] 또한 오픈서베이의 온라인 쇼핑 멤버십 트렌드 리포트에 따르면, 온라인 쇼핑 멤버십의 이용자는 3~40대의 비중이 상대적으로 높으며, 특히 20대의 경우 멤버십 가입률이 가장 낮게 나타난다.[3] 이를 종합해 봤을 때, 온라인으로 식료품을 판매하는 채널인 CJ 더마켓의 멤버십 서비스 더프라임 또한 연령대에 따라 가입률에 유의미한 차이를 보일 것으로 판단된다.

(2) 성별은 프라임 회원 여부에 영향을 줄 것이다.

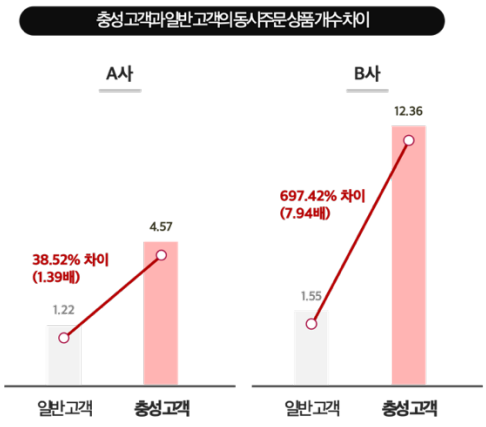
통계청의 월평균 간편식 구입 지출액 조사에 따르면, 여성은 월평균 97,920원, 남성은 월평균 89,020원으로 여성이 남성보다 간편식 소비에 더 적극적임을 확인할 수 있다.[4] 여성과 남성의 식품 소비 행태에 존재하는 유의미한 차이는, 간편식을 주력 상품으로 취급하는 CJ 더마켓의 프라임 멤버십에도 동일하게 적용될 것이다. 따라서 성별은 프라임 회원 여부에 영향을 미칠 것으로 예상할 수 있다.

(3) 고객이 구매한 상품이 CJ 더마켓의 주력 상품인 가정 간편식인지, 그 외의 비주력 상품인지가 프라임 회원 여부에 영향을 줄 것이다.

프라임 회원들은 일반 회원들보다 CJ 더마켓에서 신선 식품이나 건강식품 등 타 온라인 식료품 채널에서도 쉽게 만나볼 수 있는 비주력 상품을 더 적극적으로 구매할 것이다. 멤버십 서비스는 할인과 같은 금전적인 혜택을 제공함으로써 취급 상품을 가격적으로 부담이 가지 않는 '저관여 상품'으로 만들어 고객들이 이탈하지 않고 채널을 장기적으로 이용할 수 있도록 설계된 전략이다. 멤버십 회원들은 이러한 전략을 바탕으로 구매 채널에 높은 신뢰를 가지고 제품을 반복 구매 및 이용하는 충성 고객으로 전환된다.

CJ 더마켓의 프라임 회원들 또한 멤버십을 통해 형성된 높은 충성도로 인해 CJ 더마켓을 주 온라인 구매 채널로 이용할 것이다. 이는 프라임 회원들로 하여금 CJ 더마켓의 주력 상품인 가정 간편식품 뿐만 아니라 비주력 상품을 구매할 때에도 CJ 더마켓을 이용하도록 이끌 것이다. 따라서, 고객이 구매한 비주력 상품의 개수 및 비중은 프라임 회원 여부를 판별하는 데에 유의미한 변수로 사용할 수 있다.

(4) 동시주문 상품 종류의 개수가 프라임 회원 여부에 영향을 줄 것이다.



출처 | 빅인 이커머스 트렌드 리포트 (2021)

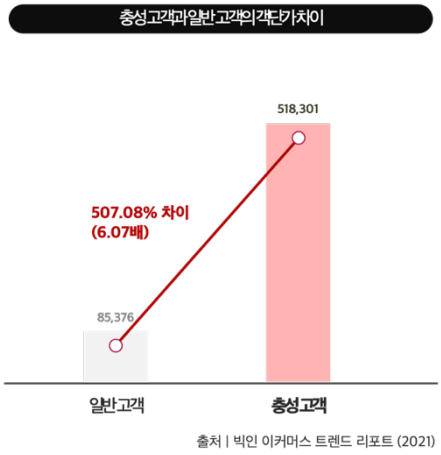
[2] '온라인 식료품 구매 트렌드 리포트 2023' , opensurvey, 2023년 1월
[3] '온라인 쇼핑 멤버십 트렌드 리포트 2022' , opensurvey, 2022년 6월
[4] 통계청, '월평균 간편식(HMR) 구입 지출액' , 가공식품소비자태도조사, 2018 ~ 2022

프라임 회원들은 일반 회원보다 더 다양한 상품 가짓수를 구매할 것이다. 빅인의 이커머스 트렌드 리포트에 따르면 충성 고객은 4.57개, 일반 고객은 1.67개로 충성 고객은 일반 고객 대비 평균 173.74% 더 다양한 상품 가짓수를 구매한다. 두 고객층의 구매 상품 수는 가장 적게 차이 나는 곳에서는 38.52%, 많게는 697.42%로 유의한 차이를 보인다.[5] CJ 더마켓의 더프라임 멤버십은 차별적인 혜택을 통해 충성 고객을 확보하기 위한 제도이므로, 충성 고객인 프라임 회원과 일반 회원을 구분 지을 때 주문한 상품의 가짓수가 효과적으로 이용될 수 있을 것이다.

(5) 할인 및 이벤트 기간 내에 주문했는지가 프라임 회원 여부에 영향을 줄 것이다.

일반 고객들은 충성 고객보다 세일 상품을 구매하는 경우가 많아 기업 입장에서 이익률이 낮다는 것은 고객관계관리(CRM) 관점에서 저명한 사실이다. 충성 고객은 기업에 대한 신뢰를 우선으로 하는 관계 지향적 구매를 하며, 일반 고객들은 가격에 대한 관심도를 기반으로 하는 거래 지향적 구매를 한다. 할인이나 이벤트 기간에 구매하는 고객을 분석한다면 관계 지향적 구매를 하는 프라임 회원과 거래 지향적 구매를 하는 일반 회원을 판별하는 데에 큰 도움이 될 것이다.

(6) 주문 금액의 많고 적음이 프라임 회원 여부에 영향을 줄 것이다.



출처 | 빅인 이커머스 트렌드 리포트 (2021)

앞선 가설에서 언급했듯이, 일반 회원들은 거래 지향적 구매를 하므로 할인 행사 상품이나 이벤트 상품 위주로 구매한다. 따라서 프라임 회원보다 상대적으로 주문 금액이 적을 것으로 예상된다. 또한, 빅인의 이커머스 트렌드 리포트에 따르면 충성 고객은 일반 고객 대비 한 번 구매할 때의 객단가가 평균적으로 507.08% 높게 나타난다. (빅인, 2021) 따라서 주문 금액은 프라임 회원 여부 판단에 필요한 변수로 선택되어야 한다.

2) 임직원 모델링

유의사항

1. 본 쿠폰 적용 상품은 CJ더마켓 홈페이지, 앱 내 '마이페이지 > 할인쿠폰 > 적용상품보기' 에서 확인 가능합니다.
2. 본 쿠폰은 CJ더마켓의 운영 방침 및 재고에 따라 쿠폰 적용 상품이 사전 안내 없이 상시 변경될 수 있습니다.

⋮

7. CJ 임직원은 참여하실 수 없습니다.

[5] 빅인사이트 2021 이커머스 트렌드 리포트 Vol.03, 2021년 7월

임직원 모델링의 경우 일반 회원 모델링과 다르게 최종 변수 선택에서 event_yn (이벤트 유무)을 제외한다. CJ 임직원에게는 CJ 더마켓에서 상품을 구매할 때 40% 할인 혜택이 제공된다. 임직원이 프라임 회원일 경우, 무료 배송 등의 추가적인 혜택을 받는다. 그러나 CJ 임직원은 CJ 더마켓에서 진행하는 쿠폰 발행 이벤트, 첫 구매 이벤트 등의 이벤트에 참여할 수 없다. 또한, CJ 임직원이 갖는 할인 혜택은 할인 이벤트 시 중복되어 적용되지 않으므로 CJ 임직원의 구매 결정에 이벤트 여부는 영향을 미치지 않을 것이다.

3. 전처리

1) 변수 처리 과정

(1) scd (주문번호)

추가적인 전처리 과정 없이 원본 데이터 그대로 사용한다.

(2) net_order_amt (주문 금액)

개별 제품의 구매 가격보다 고객별 총주문 금액이 프라임 여부 예측 모델링에 더 적합한 데이터라 판단하였다. scd (주문번호)를 기준으로 group by를 진행한 뒤, sum 메소드를 이용해 도출된 고객별 총주문 금액을 최종적으로 사용한다.

(3) gender (성별)

범주형(M/F)으로 제시된 데이터이므로 원핫 인코딩을 통해 숫자형으로 변환한다.

(4) age_grp (나이)

범주형(10/20/30/40/50)으로 제시된 데이터이므로 원핫 인코딩을 통해 숫자형으로 변환한다. 나이는 순서가 없는 nominal 변수이기 때문에, 데이터에 순서를 부여하는 라벨 인코딩 방식은 적합하지 않아 선택하지 않았다.

(5) order_date (주문 날짜)

추후 event_yn (설날/이벤트 기간 내 주문) 변수 생성에 활용되므로 datetime 형식으로 변환한다.

(6) product_name (상품명)

- product_name (상품명)에서 이벤트, 용량 등을 제외하고 상품 이름만을 추출한다. 추출된 상품명을 바탕으로 주력 및 비주력으로 상품을 구분한 뒤 새롭게 생성한 non_main_yn (비주력 상품 여부) 변수에 할당한다. 이때, 비주력 상품의 구매 여부가 프라임 회원 예측에 중요한 요인이므로 비주력 상품을 1로, 주력 상품을 0으로 표시한다.
- 이후 scd (주문번호)를 기준으로 group by하여 고객별 장바구니 형태로 데이터를 가공한다. 고객이 주문한 상품별로 non_main_yn 변숫값을 확인한 뒤, 이를 바탕으로 non_main (비주력 상품 개수 or 비율) 변숫값을 결정한다. 최종 모델링에는 product_name, non_main_yn은 제외하고 non_main만을 선택해 사용한다.
- 현재 원본 데이터가 주어지지 않아 주력 및 비주력 상품의 비율을 확인할 수 없다. 따라서, non_main 변수는 전체 데이터 분포를 통해 불균형 여부를 판단한 뒤, 결과에 따라 다른 전처리 과정을 적용할 예정이다.

< Case 1: 비주력 상품을 구매한 고객의 비율이 유의미하게 높지 않을 때 >
비주력 상품을 하나라도 구매한 경우, non_main (비주력 상품 개수) 변수에 1로 표시한다.

scd(주문번호)	product_name	non_main_yn	non_main
20230124153976	잔칫집 식혜 240ml 30입	1	1
	고메 오리지널 핫도그 400g	0	
	비비고 왕교자 1.05kg	0	

< Case 2: 비주력 상품을 구매한 고객의 비율이 유의미하게 높을 때 >
전체 구매 상품 중 비주력 상품의 비율을 계산한 후, non_main (비주력 상품 비율) 변수에 계산값을 표시한다.

scd(주문번호)	product_name	non_main_yn	non_main
20230124153976	잔칫집 식혜 240ml 30입	1	1/3 = 0.33
	고메 오리지널 핫도그 400g	0	
	비비고 왕교자 1.05kg	0	

(7) net_order_type (동시 주문 상품의 개수)
scd (주문번호)로 group by 후 고객이 주문한 총 상품 종류의 개수를 count한다. 이후 주문 번호별로 새로운 변수인 net_order_type을 생성하고, 계산한 총 상품 종류의 개수를 할당한다.
- 제시된 데이터가 2023년 1월에 한정된 데이터이고, 동일 고객의 재구매는 일어나지 않았다는 가정이 존재하므로 단순 식별자의 역할을 하는 scd는 group by를 통한 데이터 가공 작업 이후 삭제한다.

(8) event_yn (설날/이벤트 기간 내 주문)
event_yn은 order_date (주문 날짜)와 product_name (상품명), 두 가지 변수를 이용해 처리한다.
- CJ 더마켓에서 2023년 1월 진행한 이벤트의 기간을 파악한 후, order_date 변수를 기준으로 이벤트 기간에 포함될 경우 event_yn에 1 값을, 포함되지 않을 경우 0 값을 넣어준다. CJ 더마켓에서 2023년 1월 진행한 이벤트는 다음과 같다.
    a. 설날 상품 특가 판매 기간, 설날 쿠폰 증정: 1/4 ~ 1/18
    b. 더마켓 세일페스타 등 이벤트 기간: 1/1 ~ 1/10
- product_name 변수를 이용하여 [이벤트], [설] 등의 특가 상품 표식이 상품명에 기재되어 있는지 파악한 후, 존재할 경우 event_yn에 1 값을, 존재하지 않을 경우 0 값을 넣어준다.

2) 최종 변수 선택
scd (주문번호), product_name (상품명), non_main_yn (비주력 상품 여부)는 파생 변수 생성 후 삭제한다.
- 이 외의 변수는 모두 모델링에 사용한다.

3) 임직원 데이터셋 처리
우선 범주형(Y/N) 변수인 employee_yn(임직원 유무) 변수에 원핫 인코딩을 적용하여 숫자형으로 변환한다. CJ 더마켓의 임직원 회원과 비임직원 회원에 대해 서로 다른 모델링을 적용해야 하므로, employee_yn을 변수를 기준으로 임직원 및 비임직원 데이터셋을 분리한다.

- 앞서 임직원 모델링에 대한 가설 설정에서 언급했듯이, CJ 임직원의 구매 결정에 이벤트의 유무는 큰 영향을 미치지 않기 때문에 임직원 모델링에서 event_yn (이벤트 유무) 변수는 사용하지 않는다. 따라서 임직원 데이터셋에서는 employee_yn과 event_yn 변수를, 비 임직원 데이터셋에서는 employee_yn를 삭제한다.

4. 모델링

1) 데이터 특성 파악

모델링의 성능을 향상시키기 위해서는 데이터 특성을 기반으로 적절한 모델을 선택해야 한다. CJ 더마켓 프라임 회원 예측에 사용될 데이터는 다음과 같은 특성을 가질 것으로 예상된다.

	비임직원	임직원
타겟변수 불균형	크다	크다
독립변수 개수	많다	많다
변수들 간의 상호작용	있다	있다
고객 데이터 양	비교적 많다	비교적 적다

(1) 타겟변수 불균형

CJ더마켓은 2020년 누적 가입자 수 300만 명을 달성하였으나, 그중 2만 명 정도만이 더프라임 유료 멤버십 회원으로 가입돼 있었다.[6] 2023년 CJ더마켓의 누적 가입자 수가 이전과 크게 달라지지 않았고, 더프라임 멤버십 혜택이 축소되고 있기 때문에 여전히 프라임 회원이 일반 회원에 비해 현저하게 작을 것으로 추정된다. 따라서 타겟변수인 prime_yn (프라임 회원 유무)는 매우 불균형한 변수일 것이다.

- 불균형 데이터를 바탕으로 모델이 학습을 진행한다면 예측 결과가 왜곡될 수 있다. 데이터가 불균형하다면 예측 모델은 분포도가 높은 클래스에 가중치를 많이 두게 되는데, 이는 모델이 다양한 유형에 대해 적절한 학습을 진행하는 것을 방해한다. 또한 이러한 데이터는 과적합 문제가 발생할 가능성이 높아 일반화 정도를 낮추므로 모델의 예측 성능 및 설명력을 저하시키는 원인이 된다. 이러한 문제점을 해결하기 위해, 프라임 회원 데이터에 대해 오버 샘플링 방식을 적용하여 데이터를 증식한 후, 각 클래스의 비율을 맞추어 학습을 진행할 예정이다.

(2) 변수들 간의 상호작용

본격적인 모델링 과정에 앞서, 상관분석을 통해 다중공선성 여부를 확인해야 한다. 다중공선성이란 독립변수 간에 강한 상관관계가 존재할 때 나타나는 성질로, target 예측에 부정적인 영향을 주기에 경계해야 하는 문제이다. 변수 간 상관관계를 분석한 뒤, 강한 상관관계를 보이는 변수에 대해서는 삭제하거나 상관 정도를 가중치로 활용하여 모델링에 반영함으로써 다중공선성으로 인한 문제를 최소화하고자 한다.

- 다중공선성 문제는 tree 기반의 모델에서는 큰 문제가 되지 않으므로 선형 모델을 사용할 경우에만 처리한다.

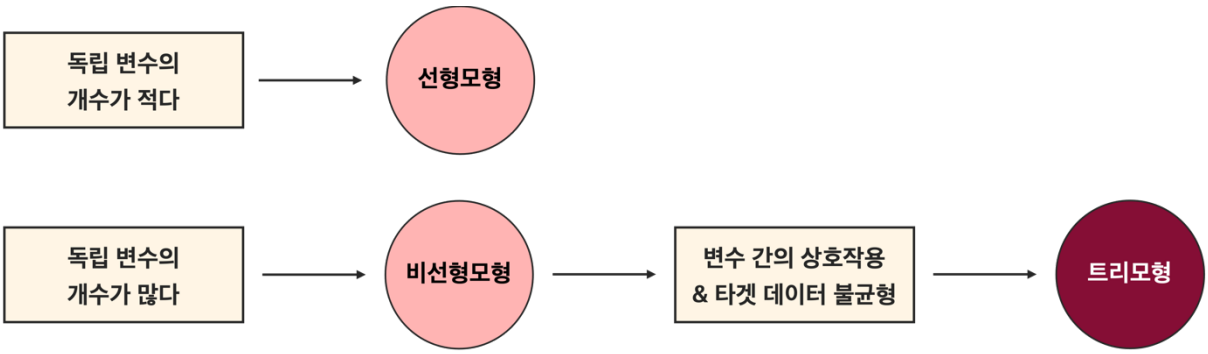
(3) 고객 데이터양

임직원 고객 데이터가 비 임직원 고객 데이터보다 비교적 적을 것으로 예상된다. CJ제일제당 관계자의 인터뷰에 따르면, CJ더마켓은 자사몰이라는 특성상 대외적으로 소비자들에게 임직원 전용 구매 플랫폼으로 인식되어 임직원들의 매출이 더 높았으나 리뉴얼 이후 임직원보다 일반 회원들의 매출이 더 높게 나타나는 추세임을 알 수 있

[6] "무조건 7% 할인" CJ더마켓 유료회원 'the 프라임' 확 바꾼다, 뉴스1, 2020년 11월

다.[7] CJ더마켓은 최근 20% 이상의 매출 성장세를 보이며 온라인 자사몰 강화에 집중하고 있기 때문에 이러한 추세가 현재까지도 지속되었을 것이다.
- 실제 데이터를 살펴보고 임직원 및 비임직원 고객 데이터양을 직접적으로 비교한 뒤 모델을 선택할 것이다.

2) 모델 선택 과정



독립 변수의 개수가 많을 때는 선형보다는 비선형일 가능성이 높다. 비선형모형 중에서도 변수 간의 상호작용을 방지하고, 타겟 데이터의 불균형성에 가중치를 부여해 해결할 수 있는 트리 모형이 데이터 특성에 알맞을 것으로 판단된다.
- 같은 트리 모형이라도 종류에 따라 장단점이 상이하기 때문에 데이터를 살펴본 후 사용할 모델을 확정한다.

3) 최종 모델 선택

	비임직원	임직원 (대량의 데이터)	임직원 (소량의 데이터)
모델	XGBoost, Random Forest, Gradient Boosting	XGBoost, Random Forest, Gradient Boosting	Decision Tree

- 변수들 간의 상호작용과 타겟 데이터의 불균형성을 손쉽게 해결할 수 있고 대규모 데이터에서도 좋은 성능을 보이는 XGBoost와 RandomForest를 최종 모델로 선택하고자 한다.
- 임직원 모델링의 경우, 임직원 데이터 확인 후 대규모 데이터가 아니라면 소규모 데이터에서 좋은 성능을 보이는 Decision Tree 모델을 사용한다.

5. 예상 기대효과

(1) 멤버십 회원 분석 과정의 기대 효과

충성 고객이란, 반복적으로 구매할 뿐만 아니라 주변 사람들에게 적극적으로 추천할 의향을 가진 고객을 뜻한다. 충성 고객은 평균적으로 전체 고객의 20%밖에 해당하지 않지만, 전체 매출의 80% 이상에 기여하고 있다. 이들은 신규 고객보다 구매 전환율이 높아 매출에 크게 기여하며, 경쟁 브랜드가 더 높은 접근성이나 낮은 가격으로 상품을 제공하더라도 구매를 유지하기 때문에 유입 비용을 절감하는 효과를 가져온다.
충성 고객의 중요성이 더욱 커짐에 따라, 최근 많은 기업이 안정적으로 수익을 창출하고 충성도 높은 고객을 락인하기 위한 전략으로 유료 멤버십 서비스를 활용하고 있다. CJ더마켓의 유료 멤버십 서비스 더프라임 또한 CJ더마켓의 충성 고객 유치에 핵심적인 역할을 할 것으로 보인다. 특히 CJ더마켓은 CJ제일제당의 자사몰로서 여러 경쟁 브랜드가 함께 입점해 있는 오픈마켓에 비해 충성 고객을 확보하기에 용이하고 안정적인 매출을 낼 수 있

[7] "자사몰 힘주는 식품업계 (1) CJ제일제당, '더마켓'으로 유통플랫폼 넘본다", 한국금융, 2021년 9월

다는 장점이 존재한다. 따라서 더프라임 멤버십의 성행을 통해 충성 고객 육성이 효과적으로 이루어질 수 있다면, 자사몰이라는 특성에 힘입어 CJ더마켓의 매출 증가에 많은 도움이 될 것으로 예상된다.
성공적인 유료 멤버십을 통해 수익을 극대화하기 위해서는, 기존 멤버십 가입자들의 유지뿐만 아니라 새로운 멤버십 가입자를 확보해야 한다. 기존 멤버십 고객들의 구매 행동을 토대로 공통의 특성을 식별하고 이들과 유사한 행태를 보이는 고객을 분류한다면 멤버십 가입자가 될 잠재력을 지닌 고객을 일정 수준 예측할 수 있을 것이다. 따라서 일반 회원들과 차별화되는 멤버십 회원들의 특징에 대한 심도 있는 분석이 필요하다.

(2) 본 팀에서 진행하는 예측 모델링의 기대 효과

본 팀에서는 멤버십 회원과 일반 회원을 구분 짓는 데 가용한 변수를 최대한으로 고려하고자 노력하였다. 최종 모델링에 사용할 변수를 선택할 때는 명확한 근거로 뒷받침하는 것을 가장 중요시하였기에 추측이나 추정이 아닌 논문이나 기사, 통계 자료를 바탕으로 꼼꼼하게 자료조사를 진행했다. 실제로 프로젝트 초기 단계에서 프라임 회원이 일반 회원보다 대용량 상품을 더 자주 구입한다는 정보를 바탕으로 조사를 진행했으나, 최근 경기 불황과 고물가가 지속되면서 대용량 소비 트렌드가 이어지고 있어 멤버십 회원만의 특징으로 보기엔 어렵다고 판단하였다. 충분히 설득력 있어 보이는 근거 자료가 존재하더라도, 이와 상충되는 연구 결과의 유무를 반드시 확인하였다.
본 팀에서 제안하는 파생 변수들은 새로운 프라임 멤버십 회원 확보에 효과적으로 사용될 수 있다. 일반 회원이 상품을 주로 이벤트 기간에 구매한다는 점을 이용한다면, 이벤트 기간 동안 프라임 회원만을 위한 독점 행사를 동시에 진행해 프라임 회원의 혜택에 대해 홍보하고, 가입을 유도할 수 있다. 특히 식료품 시장은 전환 비용이 낮고 가격이 미치는 영향이 크기 때문에 프라임 회원들에게만 독점적으로 제공되는 할인 혜택이 존재한다는 사실은 고객들에게 매력적으로 다가올 것이다. 또한, 프라임 회원들이 일반 회원들보다 비주류 상품을 주로 구매한다는 점을 이용해, 프라임 회원들을 위한 비주류 품목 위주의 할인 이벤트를 진행할 수도 있다. 본 팀에서는 타당한 근거를 바탕으로 변수를 생성하고자 노력하였기 때문에, 기존 멤버십 회원을 예측하는 것뿐만 아니라 새로운 멤버십 회원을 유치하는 데 큰 도움이 될 것이다.