2 0 2 3

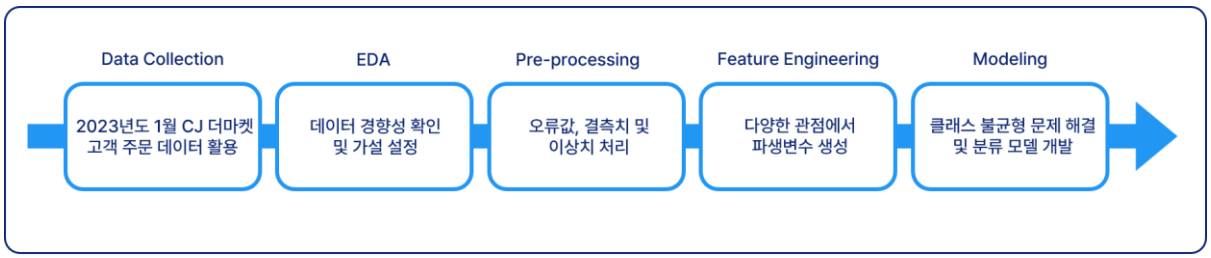
모델링 고도화

분 석 보 고 서

2023.05.09

팀 명	쓰리라찹
팀 원 명	김지은, 김채원, 이현준, 천예은

I. 기획서 요약



< 분석 스토리 보드 >

본 분석보고서는 **EDA – Pre-processing – Feature Engineering – Modeling**의 순서로 진행된다. **EDA** 과정에서는 target 값에 대한 분석을 진행하고, 각 컬럼 별 데이터 경향성을 확인한다. 기존 컬럼을 특징에 따라 3가지 관점(상품 정보, 고객 정보, 주문 정보)으로 분류하고, 이를 바탕으로 예측 가설을 수립하여 분석을 진행한다. **Pre-processing** 단계에서는 KNN Imputer, DBScan, Studentized Residual 방법론을 활용하여 데이터에 대한 전처리를 진행한다. 변수 별 특성에 따라 데이터(행)을 삭제하는 방법, 다른 값으로 대체하는 방법을 통해 결측치를 처리한다. 또한, 변수 별 특성에 따라 범주를 벗어난 값이나 자료의 중심으로부터 떨어진 거리, 잔차를 이용해 이상치를 확인하고, 데이터(행)을 삭제하는 방법, 다른 값으로 대체하는 방법을 사용하여 이상치를 처리한다. **Feature Engineering**의 경우, categorical 변수와 numeric 변수를 기준으로 나누어 피처 생성을 진행한다. 데이터셋을 나누는 기준인 임직원 여부 컬럼과 target 인 프라임회원 여부 컬럼을 중점적으로 활용하고, 특히 unique값이 가장 많은 product_name을 많이 활용하여 피처를 생성하는데 차별점을 둔다. **Modeling** 단계에서는 Imbalanced Class와 다양한 예측 모델 실험에 초점을 맞춘다. 머신러닝에서 타겟의 클래스가 불균형한 경우, 모델 학습에 문제가 될 수 있다. 따라서 모델링에서는 Class Imbalance 문제 해소에 집중한다. 크게 데이터 자체를 가공하는 방법과 모델 자체에서 해결하는 방법을 통해 문제를 해결하고자 한다. 마지막으로는 머신러닝/딥러닝 예측 모델을 개발하여 얻을 수 있는 기대효과에 대해서 알아본다.

II. 분석 목적 및 필요성

본 분석 보고서의 목적은 CJ 더마켓의 임직원 여부와 프라임 회원의 여부를 예측함으로써 다양한 비즈니스 인사이트를 도출하는 것이다. 이를 통해 CJ 더마켓은 효과적인 마케팅 전략 수립, 고객 맞춤형 서비스 제공, 고객 이탈 감소, 매출 증대와 같은 다양한 이점을 얻을 수 있다.

이를 위해 크게 3가지 측면에 주목하고자 한다. 첫째, 이상치 및 결측치 처리에 주목한다. 주어진 데이터에서 결측치나 이상치가 발생할 가능성이 있으므로 각 변수의 특성에 따라 알맞은 기법을 사용하여 처리한다. 또한 통계적 방법을 적용하여 보다 더 효과적인 결과를 도출한다. 둘째, 변수 특성을 고려하여 유용한 feature를 생성한다. 이를 통해 데이터의 정보를 더욱 효과적으로 활용하는 동시에, 모델의 성능을 향상한다. 셋째, Class Imbalance 문제에 집중한다. 임직원 데이터와 비임직원 데이터 모두 타겟에 대해 클래스 불균형 문제가 발생할 것으로 예상된다. 이를 해결하기 위해 oversampling, outlier Detection Model, Cost-sensitive Modeling과 같은 기법을 활용하여 데이터의 균형을 맞추고자 한다. 위와 같은 방법을 통해 목적을 달성하고, CJ 더마켓의 비즈니스에 대한 인사이트를 제공하며, 지속적인 모델 개선을 통해 더욱 나은 결과를 도출하고자 한다.

III. 세부 내용

i. EDA(탐색적 데이터 분석)

scd(주문번호)	product_name	net_order_qty	net_order_amt	gender	age_grp	employee_yn	order_date	prime_yn
20230124153976	잔칫집 식혜 240ml 30입	0.521	0.523	M/F	10/20/30/40/50	Y/N	20230102	Y/N

원본 데이터는 2개의 numeric columns와 7개의 categorical columns로 구성되어 있다. 이 중 employee_yn 컬럼을 기준으로 전체 데이터셋을 임직원 데이터셋과 비임직원 데이터셋으로 구분하고, target으로는 prime_yn을 사용한다.

(1) 컬럼 분류 및 클러스터링

컬럼 별 특성에 따라 target을 제외한 7가지 컬럼들을 아래 표와 같이 **3가지 관점(상품 정보, 주문 정보, 고객 정보)**으로 분류한다. 각 관점은 프라임회원 여부에 영향을 미치는 요소들로, **각 요소 내 컬럼들의 구매 패턴에 대해 클러스터링을 진행함**으로써 컬럼 별 데이터 양상을 파악하고자 한다. 즉 프라임회원 여부 결정에 영향을 미치는 요인들을 기준으로 이후 EDA 과정을 진행한다.

상품 정보	product_name, net_order_qty, net_order_amt
고객 정보	gender, age_grp
주문 정보	scd, order_date

< 컬럼 분류 표 >

(2) target 값 분석

먼저 두개의 데이터셋(임직원/비임직원)에 대해 target 값 분석을 진행한다. 각 데이터셋의 target 값에 대해 pie chart 를 그림으로써 **프라임회원의 비율을 파악하고, 데이터의 불균형성을 확인**한다. 임직원 데이터셋과 비임직원 데이터셋에 대해 프라임회원 비율의 차이가 두드러지는지 확인한다.

프라임회원의 고객 정보(gender, age_grp) 특징을 분석한다. bar plot 을 통해 프라임회원과 일반회원의 성별과 나이를 확인한다. 이를 통해 프라임회원의 주요 연령대와 성별을 파악한다.

(3) 컬럼 별 데이터 양상 확인

이후 다양한 방법의 시각화를 통해 각 컬럼에 대한 데이터의 양상을 확인한다. 특히 주문 정보에 해당하는 **order_date 컬럼의 경우, lineplot 을 그림으로써 데이터의 주기성을 파악**한다. 나머지 컬럼의 경우 bar plot 을 주로 사용한다. 각 컬럼에 대해 임직원 데이터셋과 비임직원 데이터셋의 차이가 있는지, 그 차이가 두드러지는지를 확인한다.

(4) 가설 설정

다음은 본격적인 분석 프로세스 이전, 일반 상식과 통계 자료를 바탕으로 세운 예측 가설이다. 아래 가설들은 원본 데이터에 대한 사전 지식이 없는 상황에서 이후 분석 과정을 원활하게 진행하기 위해 수립하였다. 본선 대회 데이터 공개 이후에는 해당 데이터에 대한 EDA 를 바탕으로 가설 수정 및 추가적인 가설 수립을 진행하고자 한다.

· 임직원의 경우, 프라임회원의 비중이 높을 것이다.

임직원의 경우, 1 년에 2 만원의 금액을 내고 더 큰 이익을 볼 수 있을 것이라고 예측된다. 특히 임직원 카드는 CJ 브랜드 구입 및 이용 시 40% 할인 혜택을 제공하므로, 임직원은 프라임회원의 혜택을 누리기 위해 프라임회원에 가입했을 것이라 추측한다.

· 임직원의 경우, 프라임회원은 40~50 대 남성이 많을 것이다.

CJ 제일제당 정보 기준, 기업 내 임직원 비율은 2019 년도 남성 74.05%, 2020 년도 남성 72.65%, 2021 년도 72.21%로 남성 임직원이 여성 임직원에 비해 더 많다. 지난 3 년간 남성 임직원의 비율이 70% 이상이므로, 비임직원에 비해 프라임회원 중 남성의 비율이 더 높을 것으로 예상된다.

또한 타사(이마트몰)의 모바일 쇼핑 연령대별 비중 변화를 확인했을 때 40 대의 비중이 가장 높았으며, 50 대의 모바일 쇼핑 비중 상승률이 가파르므로 40~50 대 남성으로 대상을 축소하였다.

· 임직원의 경우, 프라임회원은 구매 수량에 비해 구매 금액이 적을 것이다.

임직원은 40% 할인을 추가적으로 제공받으므로, 구매 수량에 비해 구매 금액이 적을 것이다.

· 비임직원의 경우, 프라임회원은 30~40 대가 많을 것이다.

온라인 쇼핑몰을 이용하는 비율에 대한 통계 자료를 확인했을 때, 30~40 대의 비율이 가장 높았으므로 위와 같은 가설을 수립하였다.

· 프라임회원의 개별 상품 결제 금액은 일반회원의 개별 상품 결제 금액에 비해 낮을 것이다.

프라임회원의 혜택(할인, 기프트 카드)을 적용했을 경우, 프라임회원의 최종 결제 금액은 일반 회원보다 낮을 것이다.

· ID 별 총 평균구매액과 각 ID 의 구매액을 비교했을 때, 후자가 크면 일반 회원일 확률이 높을 것이다.

<소비자의 온라인 식품 구매 경험과 구매 및 이용 만족의 영향요인에 관한 질적 연구>에 따르면, 온라인 식품 구매 경험자는 주로 무료 배송 이상 금액으로 구매하며, 필요 시에 추가 상품을 구매하는 특성을 보인다고 한다. 일반 회원의 경우 무료배송을 이용하기 위해 일정 금액(3 만원) 이상 실결제가 이루어졌을 것으로 예측되며, 프라임 회원의 경우 월 1 회 무료배송 쿠폰이 있기 때문에 실결제 금액은 그보다 작을 것으로 예측된다.

· 명절 선물세트를 구매한 회원들은 일반 회원일 확률이 높을 것이다.

2023년 1월 당시, 명절 선물세트 할인 행사가 진행됐다. 10만 원 이상 구매 시 40% 할인 쿠폰을 적용할 수 있어 일반 회원의 유입이 많을 것이라 예측된다. 명절 세트를 구매하는 사람들은 일시적으로 혜택을 보려는 사람들이기에, 굳이 프라임 회원으로 전환하지 않고 할인된 행사 가격으로 구매할 것이라 추측된다. **따라서 명절 요소는 분석 과정에서의 예외 사항으로 두고, 이와 같은 요소를 반영하여 더욱 일반화된 모델을 구현하고자 한다.**

ii. Pre-processing(전처리)

(1) 결측치

target 변수에 대한 결측치는 없다고 가정하고, target 변수를 제외한 7개의 변수에 대한 결측치를 처리한다.

7개 변수의 각 변수 별 결측치 데이터의 비율이 5% 이하일 경우, 결측치 데이터(행)를 삭제하는 방법을 사용한다. 반면, 결측치 데이터의 개수가 많아 기존 데이터 결과와 전혀 다른 결과를 도출할 것으로 예상되는 경우에는 아래와 같이 각 변수에 따라 결측치 처리를 다르게 진행한다.

・상품 정보 변수 : product_name, net_order_qty, net_order_amt

product_name에 대한 결측치가 존재하는 경우, 다른 변수를 통해 예측 및 추정이 불가능한 변수이므로 **해당 데이터(행)을 삭제하는 방법을 사용**한다.

net_order_qty, net_order_amt에 대한 결측치가 존재하는 경우, **수치형 변수**라는 특징을 반영하여 다른 값으로 대체하는 방법을 사용한다. net_order_amt는 각 scd(주문번호) 별 할인 쿠폰 적용, 적립금 사용 등의 개별 정보에 따라, 같은 product_name이라도 반드시 같은 값을 갖는다고 확신할 수 없다. 또한, 해당 변수들은 스케일링을 포함한 전처리가 이미 된 변수이므로 가장 흔하게 사용하는 방법인 평균값, 최빈값, 중앙값 등의 통계값으로 대체하는 방법에 대해 확신을 가질 수 없다. 따라서 **KNN Imputer, MICE(Multivariate Imputation by Chained Equation)의 알고리즘 방법**을 사용한다.

・고객 정보 변수 : gender, age_grp

gender, age_grp에 대한 결측치가 존재하는 경우, 범주형 변수의 특성에 따라 **최빈값으로 대체**하는 방법을 사용한다.

・주문 정보 변수: scd(주문번호), order_date

scd(주문번호)에 대한 결측치가 존재하는 경우, 다른 변수를 통해 예측 및 추정이 불가능한 변수이므로 **해당 데이터(행)을 삭제**하는 방법을 사용한다.

order_date에 대한 결측치가 존재하는 경우, 범주형 변수의 특성에 따라 **최빈값으로 대체**하는 방법을 사용한다.

(2) 이상치

target 변수에 대한 결측치는 없다고 가정하고, target 변수를 제외한 7개의 변수에 대한 결측치를 처리한다.

· 범주형 변수

먼저 범주형 변수의 이상치를 확인한다. 범주형 변수인 product_name, gender, age_grp 중 gender와 age_grp에 대해서는 각 범주 값의 unique값을 확인한 후 범주 이외의 값에 대해서는 해당 데이터(행)를 이상치로 분류한다. 또한, order_date는 주어진 데이터 기간에서 벗어난 값에 대해서는 해당 데이터(행)을 이상치로 분류한다.

· 수치형 변수

수치형 변수에 대해서는 데이터 형태에 따라 3가지 방법을 적용할 수 있다. 수치형 변수인 net_order_qty, net_order_amt의 이상치를 확인하기 위해서 자료의 중심으로부터 상대적 거리를 사용하는 방법과 잔차를 이용한 방법이 있다.

먼저 **IQR** 방법론이 있다. 데이터의 분포가 정규 분포를 이루지 않거나 한 쪽으로 skewed한 경우에 활용하며, $(Q1 - 1.5 * IQR)$보다 작거나 $(Q3 + 1.5 * IQR)$보다 큰 데이터를 이상치로 분류한다.

또한, **DBScan(Density Based Spatial Clustering of Applications with Noise)**을 적용할 수 있다. 밀도 기반의 클러스터링 알고리즘으로 어떠한 클러스터에도 포함되지 않는 데이터를 이상치로 분류한다.

마지막으로, **회귀 모형을 활용한 스튜던트화 잔차(studentized residual)**를 사용하고자 한다. 스튜던트화 잔차의 절대값이 3보다 크면 이상치로 분류한다. 여기서 이상치 결정 기준으로 잔차를 추정표준오차로 사용하는 이유는 변수(데이터)의 단위에 따라 이상치로 잘못 판단되는 경우를 방지하기 위함이다.

(3) 이상치 처리

수치형 변수에 대해 정성적인 방법만으로도 이상치 탐지를 진행할 수 있지만, 정량적인 다양한 방법을 활용하고자 하는 이유는 데이터에 대해 **일반화 가능한 기준을 확립**하기 위함이다. 각 기준에 따라 이상치로 분류된 기준이 존재하므로, 다른 데이터가 들어왔을 경우에도 일반화 가능한 기준을 확립하고자 한다. 따라서 해당 방법들을 적절하게 사용하여, 공통적으로 이상치라고 분류되는 데이터에 대해서는 이상치 처리를 추가로 진행한다.

이상치 처리에 대해서도 다양한 방법론이 존재하므로, 이상치가 발생한 이유에 따라 다른 방법론을 적용한다. 우선 해당 이상치가 수집이나 기록에 의한 오류에 의해 발생한 것이라면, 그 데이터는 제외하는 방법을 사용한다.

그렇지 않은 경우, 범주형 데이터에 대해서는 이상치 데이터(행)를 삭제하는 방법을 사용한다. 임의의 판단으로 범주 값을 지정하는 경우, 기존 데이터가 설명하는 특성에 영향을 미칠 수 있기 때문이다.

수치형 데이터에 대해서 각 변수 별 이상치 데이터의 비율이 5% 이하라면, 범주형 데이터와 같이 해당 이상치 데이터(행)를 삭제하는 방법을 사용한다. 반면 이상치 데이터의 비율이 커 삭제 시 기존 데이터 결과와 다른 결과를 도출할 것으로 예상되는 경우, 데이터 변경 방법을 사용한다. 임의나 평균의 표준편차를 이용하여 상한 값과 하한 값을 설정하고, 상한 값보다 크면 상한 값으로 하한 값보다 작으면 하한 값으로 대체하여 이상치를 처리한다.

(4) 회귀 진단

회귀 모형에 적용하기 전에 회귀 진단을 진행해야 한다. 먼저, 데이터의 비선형성(등분산성)을 확인하기 위해 각 변수 별 잔차 대 적합치 그래프를 통해 잔차 분석을 진행한다. 등분산성을 만족한다면 0을 중심으로 $\hat{y}$ 값에 관계없이 일정 범위 내에서 특정한 패턴을 가지지 않게 분포하게 된다. 그렇지 않은 경우는 다음의 두 방법과 같은 polynomial regression을 활용한다. $\hat{y}$ 값이 커지면서 잔차의 폭이 커지는 경우는 log 변환 항을 추가하고, $\hat{y}$ 값이 커지면서 잔차가 하강(상승)하다가 상승(하강)하는 경우는 제곱항을 추가하여 등분산성을 충족하도록 한다.

iii. Feature Engineering(파생변수 생성)

아래는 **범주형 파생변수** 목록이다.

No.	파생변수명	변수 설명
1	scd_cnt	한 번에 주문한 product 종류 및 수량
2	scd_net_order_amt_var	한 고객 당 구매한 제품 가격의 분산
3	scd_net_order_qty_max	각 고객이 한 번 주문할 때 가장 많이 구매한 단일 제품의 수량
4	scd_net_order_qty_min	각 고객이 한 번 주문할 때 가장 적게 구매한 단일 제품의 수량
5	scd_product_name_unique_cnt	각 고객이 구매한 고유한 제품 수량 개
6	order_date_week	order_date를 week별로 범주화
7	order_date_day	order_date를 day별로 split
8	product_release_order_date_differ	최신 제품을 좋아하는 고객 파악
9	week_vs_weekend	주말에 구매한 제품의 비율과 평일에 구매한 제품의 비율
10	order_date_daily	요일 별 구매 컬럼 원핫 인코딩
11	order_dat_daily_net_order_amt_sum	요일 별 주문 총액
12	gender_age_grp_mean	성별 별 나이 평균
13	gender_net_order_qty	성별 별 주문 수량
14	gender_net_order_amt	성별 별 주문 금액
15	age_grp_net_order_qty_cnt_sum_mean	나이대별 총주문수량의 평균
16	age_grp_net_order_qty_max	최대구매수량 연령층

No.	파생변수명	변수 설명
17	age_grp_net_order_amt	최대구매금액 연령층
18	age_grp_product_name_cnt	나이대별 구매 품목 빈도수 순위
19	product_name_cnt	상품 구매 빈도수
20	top_n	상위 n개의 상품
21	bottom_n	하위 n개의 상품
22	product_category	CJ더마켓 웹페이지의 category를 활용한 범주화
23	product_category_age	category x age
24	product_category_max	각 고객이 가장 많이 구매한 제품 카테고리
25	product_category_min	각 고객이 가장 적게 구매한 제품 카테고리
26	product_category_cnt	각 제품 카테고리별로 구매한 빈도수 파악
27	product_category_net_order_qty_mean	각 카테고리에서 평균적으로 구매하는 제품 수량
28	product_category_net_order_amt_mean	각 카테고리에서 평균적으로 구매하는 제품 가격
29	product_name_net_order_amt_max	상품 별 최대 금액
30	product_name_net_order_amt_min	상품 별 최소 금액
31	product_name_net_order_amt_mean	상품 별 평균 금액
32	popular_product_purchase_o_x	매출이 높은 제품을 인기 제품으로 정의한 후, 구매 여부 0,1 column 생성
33	product_category_pur_o_x	특정 제품 카테고리의 구매 여부

아래는 **수치형 파생변수** 목록이다.

No.	파생변수명	변수 설명
1	net_order_qty*net_order_amt	주문금액과 주문수량의 곱
2	net_order_amt_max	가장 높은 주문 금액
3	net_order_amt_min	가장 낮은 주문 금액
4	net_order_qty_max	가장 높은 주문 수량
5	net_order_qty_min	가장 낮은 주문 수량
6	scd_net_order_amt_max-min	고객이 한 번에 주문한 제품들 중 제품 가격의 최고값과 최저값의 차이
7	net_order_amt_group	제품 가격에 따라 그룹으로 범주화
8	net_order_amt_group_cnt	제품에 가격에 따라 범주화한 후, 각 그룹별로 주문한 제품의 총 수량
9	net_order_amt_qty_group	제품의 주문량에 따른 그룹 범주화

위와 같은 파생변수들을 생성함으로써 모델의 예측 성능을 향상하고, 더욱 체계화된 분석을 진행하고자 한다. 추후 공개되는 본선 데이터의 양상에 따라 추가적인 Feature Engineering 과정을 거칠 수 있다.

iv. Modeling

모델링에 들어가기 앞서 모델링에 사용할 데이터의 특성을 알아보고자 한다. 앞서 설정한 대전제에 따라 임직원 데이터와 비임직원 데이터 모두 타겟에 대해 **클래스 불균형 문제**가 있을 것이라 예측된다. 타겟의 클래스가 불균형한 경우 성능 지표 왜곡, 과적합, 소수 클래스의 정보 손실, 분류 임곗값 문제 등 여러 문제가 발생할 수 있으며, 이와 같은 문제들은 예측 모델의 성능을 좌우할 수 있다. 따라서 클래스 불균형 문제를 해결하는 것은 필수적이다. 클래스 불균형 문제를 해소한 이후에는 머신러닝 및 딥러닝 모델을 통해 실제 예측 모델을 구현하고자 한다. 이에 본 Modeling 파트에서는 클래스 불균형을 해소할 수 있는 방법론 및 모델링 과정에 대해 다루고자 한다.

(1) Class Imbalance 문제 해결 방안

클래스 불균형 문제의 해결 방안은 다양하나, 본 분석 보고서에서는 Over-Sampling, Outlier Detection Model, Cost-Sensitive Modeling을 활용하는 방법론 중 일부를 택하고자 한다.

· Over-Sampling

Over-sampling은 원본 데이터 자체를 가공해서 클래스 불균형 문제를 해결하는 기법으로 비율이 적은 클래스에 해당하는 데이터를 추가로 만든다. 이러한 Over-Sampling 기법에는 Resampling, SMOTE, Borderline-SMOTE, ADASYN 등이 존재한다. 따라서 원본 데이터에 각각의 샘플링 기법을 적용해 데이터셋을 생성한 후 모델링을 진행한다.

· Outlier Detection Model

앞서 진행한 Over-Sampling은 데이터 자체를 가공해서 클래스 불균형 문제를 해소하는 방법론이다. 하지만 Outlier Detection Model은 모델이 스스로 비율이 적은 클래스를 탐지한다. 따라서 데이터 자체를 가공하지 않고 원본 데이터를 모델에 학습시켜 클래스 불균형 문제를 해소할 수 있다. 이러한 Outlier Detection Model로는 Isolation Forest, Local Outlier Forest, DBSCAN, One – Class SVM 등이 존재한다. 따라서 Outlier Detection Model에 대해서는 원본 데이터를 활용해 모델링을 진행한다.

· Cost-sensitive Modeling

이외에도 몇몇 머신러닝 모델에는 클래스 불균형 문제를 해결하기 위한 가중치 파라미터가 존재한다. 예를 들면, 사이킷런의 로지스틱 회귀 모델의 'class_weight' 파라미터는 딕셔너리 입력값을 통해 클래스에 가중치를 부여한다. 해당 방법을 통해 원본 데이터셋과 일반적인 머신러닝 모델들을 추가로 활용하여 클래스 불균형 문제를 해소함과 동시에 모델링을 진행하고자 한다.

이와 같은 방법론을 통해모델링 전 혹은 모델링 중에 클래스 불균형 문제를 해소할 수 있다. 따라서 클래스 불균형 문제를 해소하기 위해, 오버샘플링을 통해 클래스 불균형을 해소한 데이터셋과 원본 데이터셋 두 가지를 사용해 학습을 진행하고자 한다.

(2) Feature Scaling

머신러닝의 모델은 모델에 사용되는 데이터의 범위가 너무 크거나 작으면 학습이 잘 진행되지 않을 수 있다. 따라서 각 Feature별로 데이터의 범위를 일정하게 조정해주는 스케일링이 필요하다. 스케일링 기법에는 StandardScaler, MinMaxScaler, MaxAbsScaler, RobustScaler 등이 있으며 데이터의 특징에 따라서 적절한 기법을 사용할 수 있다. 스케일링을 진행하게 되면 모든 Feature는 범위가 일정해지기 때문에 머신러닝 모델의 학습에 도움이 된다.

(3) Modeling Process

· Train Test Split

각각의 데이터셋에 대해서 학습에 사용할 Train 데이터와 평가에 사용할 Test 데이터를 나눈다. 생성한 두 데이터셋에 대해서 오버샘플링 데이터셋은 OS 데이터, 원본 데이터셋은 OR 데이터로 임의로 지칭하고자 한다. 이때 학습 데이터와 평가 데이터의 클래스 비율을 맞춰주기 위해서 사이킷런의 train_test_split() 함수의 'stratify' 파라미터를 사용한다.

· Training & Tuning

다음으로는 생성한 데이터셋을 활용해 모델을 직접 학습시킨다. 먼저 OS 데이터에 대해서 여러 가지 머신러닝 모델을 적용해 볼 수 있다.

Logistic Regression : 2진 분류의 Task에 적합한 모델로 간단하게 사용해 볼 수 있다

Decision Tree/Random Forest : 기본적인 트리 모델로 간단하면서 좋은 성능과 모델에 대한 설명이 가능하다.

Boosting 기반 모델 : GBM, XGB, LGBM, Catboost 등의 부스팅 기반 모델을 통해 높은 성능의 모델을 만들 수 있다.

또한 머신러닝 모델에 비해 더욱 높은 예측 정확도를 기대할 수 있는 **딥러닝 모델의 사용도 고려**한다. 그 중, 우리는 **정형데이터에서 강력한 예측 성능과 해석 가능성을 제공하는 TabNet**을 활용하고자 한다. 입력 데이터의 feature와 임베딩을 활용해 feature의 중요도를 자동으로 학습하여 분류 결정을 내리는 모델로, 뛰어난 일반화된 성능을 얻을 수 있다는 기대효과를 지닌다.

두 번째로, OR 데이터에 대해서는 **이상치 탐지 모델들과 코스트-민감 모델링**을 활용해 볼 수 있다. 앞서 설명한 이상치 탐지 모델들을 통해서 OR 데이터를 학습시키고 성능을 평가할 수 있다. 이후 일반적인 머신러닝 모델들을 사용할 때 비중이 적은 클래스에 대한 가중치를 주는 파라미터를 활용해 모델을 학습하고자 한다.

위와 같이 여러 모델을 통해서 학습을 진행하면서 해당 모델의 성능을 평가해보는 것도 중요하다. 따라서 **K-폴드** 방식을 통해서 학습한 모델의 성능을 확인해 볼 수 있다. 물론 test 데이터를 통해서 일반화 성능도 확인해 볼 필요가 있지만, test 데이터를 통해서 계속해서 성능을 평가할 경우 test 데이터에 과적합이 일어날 수 있다. 따라서 test 데이터를 활용한 성능 평가는 가장 마지막에만 진행하는 것으로 한다. 추가로 **하이퍼파라미**

터 튜닝을 통해서 각 모델들의 성능 고도화를 진행한다.

이와 같이 모델의 성능을 고도화한 뒤에는 각 모델의 최적 파라미터를 토대로 전체 데이터에 재학습을 진행한다. train 데이터와 test 데이터를 합한 원본 데이터로 모델을 학습시킨 뒤 예측을 진행한다.

(4) Ensemble

학습과 튜닝으로 통해 성능이 고도화된 모델들은 앙상블을 통해서 보다 더 좋은 성능으로 개선될 수 있다. 앙상블은 서로 다르게 학습된 모델일수록 그 효과가 극대화되기 때문에 OS 데이터로 학습한 모델과 OR 데이터로 학습한 모델들을 서로 앙상블 했을 때 효과가 좋을 것으로 예상된다.

Ⅳ. 기대효과

최근 이커머스 시장의 경쟁이 심화되면서, 기업들은 더욱 체계적인 고객 관리와 마케팅 전략이 필요하다. 해당 파트에서는 이커머스 플랫폼에서 유료 회원을 예측하는 머신러닝/딥러닝 모델을 개발하여 얻을 수 있는 기대효과를 분석하고자 한다.

1. 효과적인 마케팅 전략 수립

예측 모델을 통해 유료 회원이 될 가능성이 높은 고객들을 미리 파악할 수 있다. 이를 통해, 마케팅 자원과 예산을 효과적으로 배분하고, 타겟팅 전략을 개선할 수 있다.

2. 고객 맞춤형 서비스 제공

각 고객의 유료 회원 전환 가능성을 알게 되면, 개인별 맞춤형 서비스나 프로모션을 제공할 수 있다. 이를 통해 고객 만족도와 브랜드 충성도를 높일 수 있다.

3. 고객 이탈 감소 및 매출 증대

유료회원이 될 가능성이 낮은 고객들을 파악하여, 이들을 위한 특별한 서비스나 혜택을 제공함으로써 고객 이탈을 방지할 수 있다. 이를 통해 매출 증대 효과를 가져올 수 있다.

4. 비즈니스 인사이트 도출

예측 모델은 유료 회원 전환에 영향을 미치는 주요 요인을 파악할 수 있게 도와준다. 이를 통해 비즈니스 인사이트를 얻어 전략 개선에 활용할 수 있다.

5. 지속적인 모델 개선

예측 모델은 데이터가 축적될수록 성능이 개선된다. 따라서 시간이 지남에 따라 모델의 정확도가 높아져 더욱 효과적인 예측이 가능해진다.