# 프라임 회원 식별을 위한 예측 분류 모델링: 마케팅 전략 추진을 위한 고객 데이터 활용

---

## 1. 기획서 요약

- 분석 목적
    - CJ 더마켓의 고객 데이터를 기반으로 일반 회원과 임직원의 프라임 회원을 예측하는 분류 모델을 개발한다.
    - 단일 상품 분석을 통해 각 상품마다 프라임 회원이 구매하는 시기를 특정함으로써 프라임 회원의 행동 패턴에 대한 이해를 높인다.
    - 상품 조합 분석을 통해 프라임 회원이 일반적으로 구매하는 상품의 조합을 특정함으로써 프라임 회원의 행동 패턴에 대한 이해를 높인다.
- 세부 내용
    - 데이터 전처리부터 가설 수립, feature 활용 방향, feature engineering, EDA 과정, 모델링, 모델 평가, 모델 최적화, 모델 해석의 전 과정을 담았다.
    - 본선 진출 후 실제 데이터를 받게 되면, 해당 내용을 전부 적용해볼 예정이다.
- 예상 기대 효과
    - 일반 회원과 프라임 회원을 정확하게 구분할 수 있는 예측 분류 모델 개발의 기대 효과는 고객의 행동과 선호도에 대한 인사이트를 제공할 수 있다.
    - 두 유형의 고객을 차별화하는 요인이나 패턴을 이해함으로써 CJ더마켓은 마케팅 전략을 맞춤화하여 특정 고객 세그먼트를 타겟팅하고 고객 유지율을 높일 수 있다.
    - 단일 상품 분석 후, 할인율 1%마다 총 판매액의 편차를 계산해 가격민감도를 계산할 수 있고 이를 통해 기업의 이윤을 극대화할 수 있는 적정가격을 찾을 수 있다.
    - 군집화를 통해 CJ 더마켓이 고객데이터에서 잠재적인 프라임 회원을 식별하고 프로모션 또는 개인화 제안을 타겟팅하는 데 도움이 될 수 있다. 이를 통해 마케팅 노력의 효율성을 높이고 고객 확보 비용을 줄일 수 있다.

## 2. 분석 목적 및 필요성

## **2-1.** 분석 목적

- CJ 더마켓의 고객 데이터를 기반으로 일반 회원과 프라임 회원을 구분할 수 있는 예측 분류 모델을 개발하는 것이다. 두 유형의 고객을 차별화하는 요인 또는 패턴을 이해하고 임의의 고객을 적절한 멤버십 범주로 정확하게 분류할 수 있는 모델을 구축하는 것을 목표로 한다.
- 단일 상품 분석을 통해 각 상품마다 프라임 회원이 구매하는 시기를 특정함으로써 프라임 회원의 행동 패턴을 이해할 수 있고, 상품 조합 분석을 통해 프라임 회원이 일반적으로 구매하는 상품의 조합을 특정함으로써 프라임 회원의 행동 패턴을 이해할 수 있다.

## **2-2.** 필요성

전반적으로 일반 회원과 프라임 회원을 구분할 수 있는 예측 분류 모델을 개발하면 고객 행동에 대한 귀중한 통찰력을 제공하고 CJ The Market이 마케팅 및 비즈니스 전략을 최적화하여 고객 만족도와 매출 성장을 높일 수 있다.

- 타겟 마케팅: CJ 더마켓은 일반 회원과 프라임 회원을 차별화하는 요인이나 패턴을 파악하여 각 그룹에 맞는 마케팅 전략과 프로모션을 제공하여 개인화된 쇼핑 경험을 제공하고 고객 유지 및 충성도를 높일 수 있다.
- 비용절감 : CJ더마켓은 신규고객을 적절한 멤버십 카테고리로 정확하게 분류하여 자원을 효율적으로 배분하고 회원에게 혜택과 서비스를 제공하는 비용을 절감할 수 있다.
- 수익 증대: 프라임 회원은 충성도가 높고 플랫폼에서 더 많은 돈을 쓸 가능성이 높으므로 잠재적인 프라임 회원을 정확하게 식별하고 프라임 회원 가입을 유도하면 플랫폼 수익을 높이는 데 도움이 될 수 있다.

## **2-3.** 분석 및 예측

- 단일 상품 분석 : 각 상품당 프라임 회원의 구매 시기와 할인율을 통해 일반적으로 프라임 회원이 해당 상품을 어떤 시기, 할인율에 구매하는지 분석할 수 있다. 또한 이를 통해 프라임 회원 예측시 상품의 상태 정보(시기, 할인율, 일반/프라임회원의 구매 유사도 등)에 따라 상품에 대한 가중치를 부여할 수 있다.
- 상품 조합 분석 : 같은 주문 번호를 가진 상품들의 조합과 총 구매 금액을 통해 프라임 또는 일반 회원의 구매 일반적인 상품 조합, 평균 구매 금액 등을 분석할 수 있다. 또한 이를 통해 프라임 회원 예측시 상품의 조합 및 총 구매 금액에 따라 가중치를 부여할 수 있다.

- 단일 상품 예측 : 단일 상품 분석을 통해 얻은 해당 상품의 상태 정보와 회원의 구매 금액을 통해 프라임 회원을 예측할 수 있다.
- 상품 조합 예측 : 2가지 분석을 통해 얻은 해당 상품의 상태 정보와 회원이 구매한 상품의 조합 정보, 회원의 구매 금액을 통해 프라임 회원을 예측할 수 있다.
- 프라임 회원 군집 특정 : 프라임 회원의 군집내의 일반 회원은 잠재적 프라임 회원이라고 생각할 수 있고 이는 마케팅적 요소로 활용할 수 있으므로 프라임 회원의 군집을 찾아 분석한다.

# 3. 세부 내용

## 3-1. 데이터 전처리

수집된 데이터를 정리하고 전처리하여 누락된 값을 처리한다. 이 단계에는 데이터 정리, 기능 엔지니어링, 범주형 변수 처리와 같은 작업이 포함된다.

### 결측치 처리

우선 결측치를 처리해야 분석을 진행할 수 있기 때문에 결측치를 먼저 처리한다.

프라임 회원의 경우 일반 회원의 데이터보다 중요하다고 볼 수 있다. 따라서 일반 회원의 데이터에서 결측치가 발생하였을 경우 제거를 하고 그렇지 않다면 아래와 같은 방법들을 활용해 결측치를 처리한다.

1. 대치 - 추정한 값으로 누락된 값을 채운다.
    1. 평균, 중앙값, 최빈값(범주형일 때 용이), 회귀(Stochastic regression imputation)
    2. Hot-deck imputation
2. 추정
    1. Extraplolation, Interplolation
3. 알고리즘에 의한 처리
    1. xgboost, lightgbm에서 use_missing=False인자
    2. KNN에서 feature similarity 적용해 대체한다. (이상치에 민감함)
4. MICE (Multivariate Imputation by Chained Equation)
    1. 결측치를 한번만 대체하는 게 아닌 여러 번 대체해보면서 결측치의 불확실성을 체크해 결측치를 대체한다.
    2. 이산, 연속 변수에도 flexible하게 사용되는 장점이다.

5. 딥러닝 이용
    1. 학습을 통해 최적의 결측치를 찾는다.
    2. 범주형 변수에 유용하다.
    3. 대용량 데이터에는 속도가 느리다.
    4. 결측치 존재하는 feature가 여러 개면 하나의 feature만 impute할 수 있다.
    5. 하이퍼 파라미터 튜닝이 필요하다.

## 가설

- 할인율에 따른 구매 가격 시나리오
    - 0~7% 할인된 가격으로 구매한 회원은 모두 일반회원이고, 7% 이상 할인된 가격으로 구매한 회원중 프라임 회원은 7% 할인을 받은 회원이므로 시나리오는 다음과 같다.
    - 적립금 사용 (일반)
    - 7% 할인 (프라임)
    - 상품 할인 (일반)
    - 상품 할인 + 적립금 (일반)
    - 상품 할인 + 7% 할인 (프라임)
    - 적립금 + 7% 할인 (프라임)
    - 상품 할인 + 적립금 + 7% 할인 (프라임)
- 배송 유형에 따른 상품 구매
    - 같은 배송 유형인 상품의 총 주문 금액이 3만원 이상이어야 배송비가 무료이므로 배송유형에 따라서 특정 상품의 구매가 유도될 수 있다.
- 누적 금액
    - 프라임 회원은 멤버십과 관련된 추가 혜택 및 인센티브로 인해 일반 회원에 비해 평균적으로 더 많은 비용을 지출하는 경향이 있다.
- 최대 금액
    - 일반 회원에 비해 프라임 회원이 상품 한개의 최대 금액이 높을 수 있다.
- 평균 금액
    - 프라임 회원일 경우 제품을 다양하게 구매하기 때문에 평균 금액이 낮고, 일반 회원일 경우 특정 제품만 구매할 확률이 높기 때문에 평균 금액이 높을 수 있다.
- 특정 주차 및 요일에 주문 수가 많을 수 있다.
    - 특정 주차에 프로모션을 진행해 주문할 확률이 높다.
    - 월 초에 주문할 확률이 높다.
    - 주말에 주문할 확률이 높다.

    - 임직원 데이터의 경우 특정 일자에 월급이 들어와 특정 주에 주문할 확률이 높다.
- 데이터 불균형
    - 일반 회원과 프라임 회원의 차이가 극단적으로 발생하지 않을 것이라고 판단하여, 실제 경계값에서의 데이터를 oversampling 하는 기법 중 BorderlineSMOTE를 사용하여 불균형을 해소할 수 있다.
- net_order_qty & net_order_amt
    - QnA에 의하면 같은 주문 번호를 가진 상품들의 값을 모두 더한 것이 총 주문 수량, 총 주문 금액이므로 두 컬럼은 단순히 특정 값으로 나눠서 스케일링 되었다고 볼 수 있다.
    - 고로 스케일링 되었어도 net_order_qty의 값은 가장 작은 값의 정수배로 이루어져 있다고 볼 수 있다.
- 정가
    - 주문 금액은 항상 정가보다 작거나 같으므로 동일한 net_order_qty라면 net_order_amt가 가장 큰 주문 금액이 정가로 구매한 금액이라고 볼 수 있다.
- 판매가
    - 날마다 상품의 할인율이 다르므로 정가를 추정한 방식과 동일하게 날짜별로 판매가를 추정할 수 있다.
- 상품 할인을 제외한 적립금 및 프라임 할인으로 인한 할인율
    - 날짜별로 상품의 판매가를 추정할 수 있으므로 주문 금액과 해당 날짜의 판매가를 비교해 회원이 적립금 및 프라임 할인, 쿠폰으로 얼마나 할인을 받았는지 추정할 수 있다.
- 할인율에 따른 구매 빈도 분포의 유사도
    - 어떤 상품에 대해 일반/프라임 회원이 똑같은 반응을 보인다면 할인율에 따른 구매 빈도 분포가 같다고 할 수 있다. 두 분포의 유사도를 계산해 유사도가 크다면 해당 상품은 프라임 회원 예측에 도움이 되지 않는 상품이라고 할 수 있다.
    - 두 분포의 유사도는 할인율이 연속적이지 않다고 가정하고 각 할인율에 따른 빈도수의 차의 제곱을 모두 더한것을 유사도로 정의한다.

## **Feature** 활용 방향

**scd(**주문번호**)**

- 주문번호를 기준으로 상품들을 그룹화할 수 있다.

**product_name**

- 상품명을 기준으로 대/중분류 카테고리를 생성할 수 있다.
- 상품명 또는 대/중분류 카테고리를 이용해 라벨인코딩, 원핫인코딩, 엔티티 임베딩, 빈도수 인코딩을 할 수 있다.

**net_order_qty**

- 가장 낮은 값이 1개임을 가정하고 주문 수량을 정수로 변환할 수 있다.

**net_order_amt**

- 같은 주문번호인 상품들을 통해 해당 주문의 총 주문 금액, 최대 금액, 평균 금액을 계산할 수 있다.
- 가장 높은 값을 정가로 추정할 수 있고, 추정된 정가와 주문 금액에 따라 할인율을 추정할 수 있다.

**gender, age_grp**

- 성별과 나이를 합친 카테고리로 변경할 수 있다. ex) 10M, 10F, 20M, 20F

**order_date**

- 평일/휴일 카테고리를 생성할 수 있다.
- 해당 날짜가 1월의 몇번째 주차인지 계산할 수 있다.

## Feature Engineering

공통

- product_category 추가 : 상품의 카테고리 정보 추가 (예: 음료, 고기, 조미료 등)
    - 더마켓 홈페이지 내의 상품 카테고리에 해당하는 대/중분류로 카테고리 정보를 추가한다.
- shipping_type 추가 : 크롤링을 통해 상품의 배송 유형(상온&실온, 냉장&냉동, 공장직배송, 쿡킷 등) 정보를 추가한다.
- discount_rate 추가 : 결제된 금액을 정가로 나눠서 할인율 값을 구한다.
    - 상품의 정보 중 qty가 가장 작은 것을 1개, amt가 가장 큰 것을 정가라고 가정하고 각 주문 상품 당 amt를 추정된 정가를 통해 할인율을 구할 수 있다.
- gender_age 추가 : gender, age_grp 컬럼을 하나의 컬럼으로 변경한다.
    - (남자, 10대), (남자, 20대), … , (여자,40대), (여자,50대)로 총 10개 값으로 결합한다.
- weekday 추가 : order_date를 통해 평일, 휴일을 구분하는 컬럼을 추가한다.

- n_week 추가 : 해당 날짜가 몇번째 주차인지 값을 추가한다.
- order_date 삭제 : 평일/휴일 및 n주차 컬럼이 있으므로 order_date 컬럼을 삭제한다.

단일 상품 분석

- fixed_price 추가 : 전체 기간 동안의 데이터 중 qty가 가장 작고 amt가 가장 큰 가격을 정가로 설정한다.
- 판매가 데이터 생성 : 전체 상품에 대해 1~30일 동안의 판매 가격을 추정해 새로운 데이터 셋을 생성한다.
- product_discount_rate : 정가와 해당 날짜의 판매가를 통해 해당 날짜의 상품 할인율을 계산한다.
- client_discount_rate : 주문가격과 해당 날짜의 판매가를 통해 고객의 적립금 및 쿠폰, 프라임 할인으로 인한 할인율을 계산한다.
- normal_prime_rate_total : 전체 기간 동안의 일반/프라임 구매 중 프라임 회원의 구매 비율
- normal_prime_rate_today: 오늘 일반/프라임 구매 중 프라임 회원의 구매 비율
- prime_rate_today : 전체 기간 동안의 프라임 구매 중 오늘 프라임 회원의 구매 비율
- dist_similarity : 할인율에 따른 일반/프라임 회원의 구매 빈도 분포의 유사도, 이를 통해 해당 상품이 얼마나 프라임 회원 예측에 도움이 되는지 알 수 있다.

상품 조합 분석

- product_name 수정 : 상품명 중 숫자 및 불용어 제거
    - 숫자의 경우에는 정규표현식을 통하여 텍스트만 남기고 용량 및 수량(ml, 개입)의 경우 불용어로 추가하여 제거한다.
- total_price 추가 : 같은 주문 번호를 가진 모든 상품들의 금액을 합친 값
- 최대 금액을 범주형 데이터로 변환한 컬럼 추가 : 연속형 데이터를 특정 구간으로 나누어 범주형으로 바꾼다. (예: 적다, 중간, 많다)

# 3-2. EDA (Exploratory Data Analysis)

탐색적 데이터 분석을 수행하여 변수 분포에 대한 통찰력을 얻으면서 이상치를 처리하고, 패턴 또는 추세를 식별하고, 다양한 기능과 대상 변수(프라임 멤버쉽) 간의 관계를 이해한다.

1. 변수 분포
    1. 변수의 범위, 중심 경향 및 산포를 확인한다.

    2. 데이터 불균형이 존재하는지 확인하고 샘플링과 같은 추가적인 전처리를 통해 불균형을 해소한다.
    3. 모든 컬럼의 분포를 확인해 인사이트를 얻고 일반회원과 프라임회원의 특징을 파악한다.
2. 변수 간의 관계
    1. 상관관계 분석을 통해 데이터 종속성을 발견하고 특이사항(ex. 다중공선성)이 있다면 적절한 처리를 한다.

### 이상치 처리

- 시각화(box plot, scatter plot), 통계 방법(z-score, 사분위수 범위) 등을 통해 데이터의 분포 및 이상치를 확인하고 극심한 이상치는 제거한다.
- 이상치를 처리하면서 다음과 같은 사항들을 확인한다.
    - 데이터 품질
        - 데이터 수집 또는 데이터 입력의 오류가 있는 가능성을 확인 가능
        - 부정확한 예측을 할 위험을 줄이는 데 도움이 됨
    - 모델 성능
        - 평균, 분산 및 상관관계와 같은 통계 측정을 왜곡할 수 있음
        - 이상치에 민감한 모델은 신뢰할 수 없는 모델로 이어짐
    - 데이터 해석
        - 프라임 혹은 일반 회원만의 특이한 현상에 대한 통찰력을 제공할 수 있음
            - 특별한 주의와 분석이 필요한 독특하거나 극단적인 경우를 나타냄
        - 데이터를 더 깊게 이해하고 잠재적인 기회나 테스크를 식별하는 데도 도움이 될 수 있음

## **3-3.** 모델 선택 및 훈련

### 단일 상품 분석 및 예측

상품 할인율, 고객 할인율, 해당 상품의 프라임 회원 구매 비율, 할인율에 따른 일반/프라임 회원의 구매 유사도 등을 통해 프라임 회원이 해당 상품을 어떤 시기, 할인율에 구매하는지 분석하고 상품의 상태에 따른 가중치를 부여해 프라임 회원 예측을 진행한다.

### 상품 조합 분석 및 예측

같은 주문 번호를 가진 상품들의 조합과 총 구매 금액을 통해 프라임 또는 일반 회원의 구매 일반적인 상품 조합, 평균 구매 금액 등을 분석할 수 있다. 또한 이를 통해 프라임 회원 예측시 상품의 조합 및 총 구매 금액에 따라 가중치를 부여할 수 있다.

## 사용가능한 모델

주어진 데이터 세트와 분석 목적에 적합한 적절한 분류 알고리즘을 선택한다.

1. XGBoost
    1. 트리 기반 앙상블 모델이며, 다른 트리 기반들의 모델과 비슷하게 level-wised 방식으로 성장을 한다.
    2. 또한 One-hot Encoding을 적용하였을 때의 데이터에도 좋은 일반화 성능을 보인다.
    3. 추가적으로 XGBoost의 알고리즘의 하나 중 default direction을 통하여 모델을 최적화 하는데 데이터에서의 default direction을 찾아서 모델을 적합해 나아가기 때문에 최적의 성능을 낼 수 있다.
2. LightGBM
    1. 트리 기반 앙상블 모델이며, 다른 트리 기반들의 모델과는 다르게 leaf-wised 방식으로 성장을 한다.
    2. 따라서 XGBoost 모델보다 속도적인 측면에서 더 효율적이다.
    3. 또한 lightgbm을 사용했을 경우 범주형 데이터를 category 형태로 처리한 후 피처가 자동 변환되며 최적의 분할법을 통하여 분할 된다.
    4. 트리 기반의 모델에서의 범주형 변수를 분할 할 때 이진 변수(범주의 수가 적은 경우)보다 범주형 변수의 수가 많을 경우(범주의 수 q) 일 때 2^(q-1) -1 개의 분할법이 존재하여 더 모델의 분리의 성능에 좋다.
3. Penalized logistic Regression
    1. 기존의 로지스틱 회귀분석에 L1 penalty를 적용하여 Lasso회귀분석을 적용하여 중요도가 덜 한 feature에 기울기가 0이 되기에 종속변수에 중요한 영향을 주는 feature들에 대하여 알 수 있어 추가적으로 마케팅을 할 시에 어떠한 요인이 중요한지 파악할 수 있다.
4. Tabular Neural Network (TabNet)
    1. 딥러닝과 decision tree의 아이디어를 결합해 테이블 형식 데이터 분석을 위한 해석가능하고 효과적인 모델이다.
    2. 고차원 및 희소 기능 공간을 처리하는 능력으로 알려져 있다.
    3. 고차원 정형 데이터 문제를 해결한다.

## 모델 선택 및 훈련

- 범주형 데이터를 사용하므로 stratified-kFold를 통하여 데이터를 분할한 후 모델링을 진행한다.
- 종속변수와 독립변수 사이의 선형 및 비선형 관계를 확인하고 그에 맞는 알고리즘을 선택한다.
- 적절한 교육 및 검증 기술을 사용하여 전처리된 데이터에서 선택한 모델을 학습시킨다.
- TabNet과 xgboost 등과의 앙상블 모델이 대부분의 경우에서 나은 성능을 보여주기 때문에 f1-score를 높이는 관점에서 바라볼 때 stacking 알고리즘을 사용하는 것을 고려한다.

프라임 회원 군집 특정

K-means, 계층적 클러스터링, DBSCAN 등과 같은 클러스터링 알고리즘을 사용해 군집을 특정한다.

## **3-4.** 모델 평가

주최 측에서 제시한 f1-score을 성능 평가 지표로 활용해서 고객을 프라임 및 일반 회원에 대한 예측 분류 모델의 성능을 평가한다.

f1-score이외에도 ROC-AUC 커브를 통한 일반화, 또한 데이터가 불균형하기에 Precision-Recall Curve등 여러 복합적인 지표를 통하여 일반화 시킨다.

## **3-5.** 모델 최적화

성능을 더욱 향상시키기 위해 하이퍼 파라미터를 조정하거나 더 나은 앙상블 방법을 탐색하여 모델을 미세 조정한다. 다양한 구성을 반복적으로 실험하고 모델 성능에 미치는 요인들에 대해 분석한다.

- 과적합으로 인한 성능 저하가 일어난 경우, 특성 중요도(feature importance)를 파악하여 중요도가 낮은 변수를 제거한다.
- GridSearch, RandomSearch, HyperOpt, Optuna 등의 방법을 이용해 최적화된 하이퍼 파라미터를 탐색한다.
    - Xgboost, Lightgbm 모델의 경우 베이지안 서치 기반의 HyperOpt를 사용하여 최적화한다.

## **3-6.** 모델 해석

모델링을 통해 나온 결과를 단순히 분류에만 사용하는 것이 아닌 모델을 통하여 구할 수 있는 추가적인 부분을 XAI (Explainable Artificial Intelligence)기법을 활용하여 마케팅적 인사이트 혹은 비즈니스에 활용한다.

- 모델을 해석할 때에 특성 중요도 파악을 통하여 어떠한 피처가 분류에 영향을 미치는 지에 대하여 알 수 있다. 예를 들어서 총 가격이 영향을 많이 미친다면 이러한 점을 고려하여 추가적인 마케팅에 활용할 수 있다.
- 그러나 특성 중요도만으로는 피처의 영향이 긍정적인지 부정적인지를 판단하기 어렵다. 이때, 부분 의존성 플롯(Partial Dependence Plot)을 활용하여 각 피처가 모델에 어떤 방향으로 영향을 미치는지 파악할 수 있다.
    - 부분 의존성 플롯은 특정 피처의 영향력과 함께 그 피처의 마진 효과까지 보여주기 때문에, 모델의 오차 범위도 동시에 확인할 수 있다.
    - 이를 통해 모델의 예측 결과에 대한 이해를 높일 수 있으며, 이를 바탕으로 더 나은 결정을 내릴 수 있다.
    - 특성 중요도와 함께 이를 사용한다면, 피처의 영향력과 그 방향성을 종합적으로 이해하고 모델 개선에 활용할 수 있다.
- XAI 기법을 활용하여 마케팅 및 비즈니스에 활용
    - 고객 세분화: 피처들의 영향력과 방향성을 고려하여 고객군을 세분화하고, 각 세그먼트별로 맞춤형 마케팅 전략을 수립할 수 있다.
    - 제품 개선 : 모델이 중요하게 여기는 피처들을 바탕으로 제품 개선에 집중할 수 있다. 예를들어 프라임 회원인지에 대해 요일이 매우 큰 영향이 미치는 것으로 판단이 되면, 각 요일별로 추가적인 혜택을 주는등의 전략을 수정하여 개선을 할 수 있다.
    - 리소스 최적화 : 피처의 영향력을 분석해 효과적인 리소스 배분을 결정할 수 있다. 예를 들어, 특정 피처가 프라임회원인지에 큰 영향을 미친다면, 해당 피처와 관련된 활동에 더 많은 리소스를 투자할 수 있다
    - 경쟁 분석 : 모델 해석 결과를 바탕으로 시장의 트렌드와 경쟁사의 전략을 분석하여 시장에서 독보적인 위치를 확보하기 위한 전략을 수립할 수 있다.
- 하지만 XAI 기법을 활용할 때, 과대해석 및 실버불렛에 대한 주의가 필요하다.
    - 과대해석 : 모델 해석 결과를 분석할 때, 너무 세부적인 정보에 집중하거나 복잡한 패턴을 간과하는 등의 과대해석이 발생할 수 있다. 해설 결과를 신중하게 검토하고 여러 기법을 사용하여 검증해야 한다. 또한, 도메인 전문가의 의견을 참고하여 모델 해석 결과와 현실과의 괴리를 최소화 할 수 있다.
    - 실버불렛 : XAI 기법은 모델의 해석을 돕는 도구이지, 모든 문제를 해결해주는 만능 도구가 아니기 때문에 모델의 정확성 로버스트성, 공정성

등의 측면에서도 개선이 필요할 수 있으며, 다양한 기법과 접근법을 조합하여 최적의 솔루션을 찾아야 한다.

- 이러한 과대해석과 실버 불렛에 대한 주의를 기억하며, 비즈니스 전략과 모델 개선을 통하여 추가적으로 비즈니스에 긍정적인 영향을 미칠 수 있다.

변수 선택 **(**고려**)**

컬럼이 많지 않기 때문에 변수를 선택하여 모델에 적용하는 것은 과적합 문제를 보고 판단한다. 과적합이 될 경우 여러가지 변수 선택법을 활용할 수 있는데 전처리 방법에서 사용하는 Filter Method 기법 혹은 Wrapper method 기법 혹은 Embedded Method 기법을 활용하여 할 수 있다.

# **4.** 예상 기대 효과

전반적으로 예측 분류 모델 개발의 기대 효과는 고객 세분화를 개선하고 고객 참여 및 유지를 늘리며 CJ더마켓의 매출 성장을 주도하는 것이다.

- 일반 회원과 프라임 회원을 정확하게 구분할 수 있는 예측 분류 모델 개발의 기대 효과는 고객의 행동과 선호도에 대한 인사이트를 제공할 수 있다.
- 두 유형의 고객을 차별화하는 요인이나 패턴을 이해함으로써 CJ더마켓은 마케팅 전략을 맞춤화하여 특정 고객 세그먼트를 타겟팅하고 고객 유지율을 높일 수 있다.
- 단일 상품 분석 후, 할인율 1%마다 총 판매액의 편차를 계산해 가격민감도를 계산할 수 있고 이를 통해 기업의 이윤을 극대화할 수 있는 적정가격을 찾을 수 있다.
- 군집화를 통해 CJ 더마켓이 고객데이터에서 잠재적인 프라임 회원을 식별하고 프로모션 또는 개인화 제안을 타겟팅하는 데 도움이 될 수 있다. 이를 통해 마케팅 노력의 효율성을 높이고 고객 확보 비용을 줄일 수 있다.