모델링 고도화 제출 자료

生卽必死 死卽必生 **(생즉필사 필사즉생)**

팀원 : 김해우(조장), 류병하, 신기성, 이은지

목차

요약 데이터

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

I. 기획서 요약

본 연구는 제시된 데이터를 학습하여 일반회원, 임직원의 프라임 회원 여부를 예측하는 모델을 만드는 것을 목적으로 한다. 이를 위해 요약 데이터를 살펴보았고 한정적인 정보에서 feature 생성과 모델링에 다양한 접근을 시도했다. 각종 논문을 참고하고 실제 CJ 더마켓 홈페이지에서 판매되는 상품명을 scraping 하여 일반화 성능 향상을 위한 모델링 방안을 고려했다. 결과적으로 성능이 좋은 모델을 통해 인사이트를 도출해 내어 프라임 회원 수의 증가와 CJ 더마켓의 매출에 도움이 될 수 있는 마케팅의 방향성을 제시하며 기획서를 마무리하고자 한다.

II. 분석 목적 및 필요성

2023 년도 1 월 CJ 더마켓 고객 주문 데이터를 활용하여 CJ 더마켓의 1)일반회원에 대한 프라임 회원 예측과 2)임직원의 프라임 회원을 예측하는 것이 우리의 목표이다.

일반회원과 임직원의 프라임 회원을 예측하는 분석 목적은 다음과 같다.

1. 고객 분류 및 마케팅 전략 수립

   가) 일반회원과 프라임 회원의 분류 정확도를 높여 각 그룹의 특성을 파악하고, 해당 그룹에 맞는 마케팅 전략을 수립할 수 있다. 예를 들어, 일반회원과 임직원의 프라임 회원은 각각 소비 패턴이 다를 것이므로 각각 다른 할인 혜택이나 서비스를 제공함으로써 더 많은 고객을 유치할 수 있고 이는 곧 매출 증대와 비용 절감으로 이어진다.
   나) 두 모델의 분류방식과 feature 중요도의 차이를 해석한다면, 일반회원과 프라임 회원이 아닌 임직원을 프라임 회원으로 전환하기 위한 각각의 전략 수립을 세부적으로 세울 수 있다.
   다) 또한 프라임 회원의 소비양상과 비슷한 소비양상을 가지는 일반회원에 대해 집중적인 마케팅 전략을 펼칠 수 있다.

2. 고객 이탈 예방

   가) 고객의 구매 패턴이나 행동 양식을 분석하여, 프라임 회원의 이탈을 예방할 수 있다. 예를 들어, 고객의 구매 빈도가 급격히 감소하면, 해당 고객이 이탈할 가능성이 높아진다는 것을 예측할 수 있다. 이에 따라, 이탈 가능성이 높은 고객에게 더 많은 혜택을 제공해 이탈을 예방할 수 있다.

위와 같은 분석 목적을 달성하기 위해선, 적절한 feature engineering, 모델링 및 최적화 작업이 필요하다. 따라서 아래의 방법으로 분석을 진행해 정확한 예측 모델을 만들고, 이를 토대로 비즈니스 가치를 창출할 수 있도록 한다.

III. 세부내용

모델링 시작에 앞서,

이번 공모전은 처음부터 전체 데이터가 아닌 요약 데이터만 주어졌기 때문에, 모델링 계획을 시작하기 전에 모델링할 데이터의 양을 고려해야 했다. 우리는 이번 공모전에서 데이터가 충분히 많을 것이라고 추측한다. 이유는 2020 년 연말 기사에서 CJ 더마켓의 매출이 작년 대비 500 억에서 200 억이 증가하여 700 억에

이르렀다는 보도[1]가 있었고, 2023 년 기사에서는 CJ 제일제당 관계자가 CJ 더마켓 매출 신장률이 전년 대비 20% 이상으로 지속적으로 성장 중이라고 말한 것에 있다.[2]

따라서 현재 매출액이 약 1000 억으로 추정되며, 1 월 매출액은 약 80 억이라고 가정할 수 있다. 한 번 주문 시 평균 금액을 1 만원에서 10 만원으로 모두 고려하더라도, 데이터의 개수는 80 만개에서 8 만개 정도로 충분하다고 판단한다.

또한 이번 공모전에는 일반회원 데이터와 임직원 데이터 두 가지 종류의 데이터로 모델링을 해야하는 특수한 경우이다. 일반 회원의 데이터는 충분히 많을 것으로 예상되지만, 임직원 데이터의 양은 일반회원 데이터에 비해 현저히 적을 것으로 예상된다. 따라서, 데이터가 적은 경우와 데이터가 많은 경우 두 가지 상황을 모두 고려하여 모델링 계획을 수립하고자 한다.

일반적인 Machine learning 프로젝트에는 일련의 과정들이 있다. 데이터 사이에서 인사이트를 얻기 위한 EDA, 많은 데이터 중에서 중요한 변수를 찾기 위한 Feature Engineering, 데이터 결측치 처리 등으로 데이터의 퀄리티를 높이기 위한 Data Cleansing 작업이 필요하다. 그리고 이후에는 모델 선택을 위한 과정을 순서대로 기획을 진행하게 된다. 이번 프로젝트에서는 일반회원 데이터와 임직원 데이터 두 가지 경우에 대해 차별을 두어 모델링 계획을 수립할 예정이다.

EDA

Q&A 를 통해 주문 수량과 주문 금액이 스케일링 되어있다. 답변을 받았기 때문에, 주어진 요약 데이터는 기본적인 EDA 가 수행되었다고 가정한다. 따라서, Feature Engineering 단계에서는 결측치 처리, 데이터 필터링, 이상치 처리 등을 제외하고 기획했다.

i. 컬럼 정의서

1 . Categorical Features

- **[product_name]**

  가) 이번 공모전에서 가장 중요한 feature 라고 생각한다. net_order_qty 과 net_order_amt 은 스케일링까지 마친 feature 이고, 나머지 gender, age_grp, order_date 은 비교적 간단하게 표기되어 있는 feature 들 이므로 product_name 을

[1] https://www.pinpointnews.co.kr/news/articleView.html?idxno=23399
[2] https://www.etnews.com/20230110000042

어떻게 세분화하고 범주화 하고 또 이것으로 어떻게 Feature Extrantion 을 진행하냐에 따라 모델 성능에 지대한 영향을 미칠 것이라 예상했다.

나) 가진 데이터가 하나의 행만 가지고 있어 product_name 에 대한 분석이 비효율적이라 생각해 우리는 CJ 더마켓을 crawling 해(페이지[3]의 모든 제품명을 crawling 함) 현재 CJ 더마켓 에서 판매되는 257 개의 제품명을 가지고 product_name 에 대한 Feature Engineering 을 진행했다.

다) 다음 내용은 제시된 [product_name] 문자열의 데이터를 효율적으로 사용하기 위한 방법론이다. [product_name]은 상품의 이름이 담겨 있으며, 상품의 이름에는 브랜드·상품명·용량 및 개수가 담겨있다. 또한 행사상품으로 여러 상품이 묶음으로 판매가 되는 경우도 있다. 따라서 [product_name]은 한 열에 의미가 있는 단어의 집합(corpus)이 여러 개로 묶여 있는 형태라고 접근하여 분석하고자 한다.

- ['promotion'] - product_name 에서 파생
    - 상품명 중 [2+1], [가정의달 추천], [행사] 등 가장 앞에 등장하는 대괄호[ ]를 추출하여 하나의 컬럼을 만들었다.
    - 이 feature 를 통해 CJ 더마켓에서 진행하는 행사가 프라임 회원의 여부에 주는 영향을 확인할 수 있다.
- ['product_set_1, product_set_2, product_set_3...'] - product_name 에서 파생
    - 묶음상품의 경우 각 상품이 +단위로 연결되어 있는 것을 확인했다. 따라서 +로 분리하여 묶음상품에 포함된 상품을 각각의 열로 나타낸다.
    - 이를 통해 고객이 어떤 상품들을 구매하고자 했는지 모델이 학습할 수 있도록 한다.
- ['set_yn'] - product_name 에서 파생
    - 주문한 상품이 묶음상품인지 아닌지를 y/n 으로 나타내는 feature 이다.

---

[3] https://www.cjthemarket.com/pc/event/new/cjsalefesta?evntId=604375&code=TNABSP&utm_source=naver&utm_medium=sa_bsa&utm_campaign=salefesta_may&utm_content=main_text_pc&utm_term=더마켓&n_media=27758&n_query=더마켓&n_rank=1&n_ad_group=grp-a001-04-000000028955678&n_ad=nad-a001-04-000000241178642&n_keyword_id=nkw-a001-04-000004762738506&n_keyword=더마켓&n_campaign_type=4&n_contract=tct-a001-04-000000000703673&n_ad_group_type=5&NaPm=ct%3Dlh4cgno8|ci%3D0yy0003le4Xym6YTVf2W|tr%3Dbrnd|hk%3D67bbdb0252e7b17760670ac211bc4e7a8deb07ca&plnInfo=0,0,0,0,0,0,0,0,0,1,0&plnPage=1,1,1,1,1,1,1,1,1,6,1&scrollTop=36585

    - 이를 통해 묶음상품을 선호하는 구매자가 프라임 회원에 어떤 영향을 미치는지 알아보고자 한다.

- ['product_item'] - product_name 에서 파생
    - 상품명에서 행사·브랜드·용량·개수를 제외한 순수 제품의 품목을 나타내는 feature 를생성한다.
    - 이 feature 를 통해 어떤 식품을 구매하는 지에 따라 프라임 회원에 미치는 영향을 알아보고자 한다.

- [gender]
    - 구매자의 성별을 나타낸다.
    - 아래 인코딩 란에서 처리방법 제시

2 . Numeric Features

- [net_order_qty]
    - 구매 수량을 정의한 컬럼이다.
    - 스케일링을 마친 경우라 따로 손대지 않는 것이 적절하다고 판단했다.

- [net_order_amt]
    - 구매 금액을 정의한 컬럼이다.
    - 스케일링을 마친 경우라 따로 손대지 않는 것이 적절하다고 판단했다.

- [age_grp]
    - 구매자의 나이대를 정의한 컬럼이다.

- ['working_age']
    : 멤버십 가입의 경우 정기소득이 존재하는 고객이 가입할 확률이 높다. 이번 연구에서는 경제활동가능인구의 연령을 20~60 대로 보고 [age_grp]에서 조건에 부합하면 1, 부합하지 않는다면 0 을 부여한다.

- ['age_amt_mean'] : 1 월 구매자들의 나이 대 별 구매금액을 나타내는 feature 이다.
    - 나이대로 해당 그룹을 나누고, 나이대별 구매금액의 양상을 파악할 수 있다.
    - (Pandas dataframe groupby 를 사용하여 만들며 기준열은 ['age_group'] 이 되고, agg func 은 mean 을 사용한다.)

- 예를 들어, 20 대 프라임 회원과 50 대 프라임 회원의 주된 구매 상품은 확연한 차이가 있을 것이다.
- 대한민국 20 대 초중반의 61.8 % 는 원룸에 살고 있다.[4] 그에 반해 대한민국 50 대 이상의 92.6 % 는 기혼인구이다.[5] 이러한 나이대 별 생활환경의 차이는 유의미한 구매양상의 차이를 빚을 것이며 어떠한 데이터가 프라임 회원일지 아닐지에 대한 분류에 도움을 줄 수 있을 것이다.

- ['brand_1, brand_2, brand_3...'] - product_name 에서 파생

    - 행사명이 포함된 대괄호를 제외하면 상품명의 맨 앞에 상품의 브랜드가 등장하는 것을 확인할 수 있다.
      CJ 더마켓의 브랜드관에서 CJ 더마켓에 판매되고 있는 20 개의 브랜드들을 list 로 정리해 각 브랜드명을 새로운 feature 로 생성한 뒤, 브랜드가 상품명에 등장한다면 그 브랜드에 해당하는 열에 1 을 더한다.
      이를 통해 묶음상품의 경우 한 주문에 등장하는 브랜드의 빈도 또한 확인할 수 있다.

- ['product_amount'] - product_name 에서 파생

    - 상품명의 뒷부분에서 등장하는 mg, g, kg, ml, l 등 단위를 기준으로 용량을 추출한 후 단위를 절대적인 양으로 통일시킨다(g→mg, l→ml).

    - 이를 통해 프라임 회원이 어느정도 용량의 제품을 선호하는지 알아보고자 한다.

- ['important_words'] - product_name 에서 파생

    - 프라임 회원들로 구성된 데이터 프레임을 새로이 형성하여, 해당 데이터 프레임의 ['product_name'] feature 를 분석한다.

    - string 으로 이루어진 feature 이기에, str.split 과 count method 를 사용하여 해당 열에서 비교적 많이 나온 단어들을 뽑아낼 수 있다.

    - 해당 단어들을 '중요도가 높은 단어'로 정의하고 전체 데이터 프레임에 대하여 모든 행의 ['product_name'] 에 '중요도가 높은 단어' 가 몇 개 들어가 있는지를 count 하여

[4] https://www.dailypop.kr/news/articleView.html?idxno=46389
[5] https://www.joongang.co.kr/article/25012605#home

['important_words'] feature 에 새롭게 저장한다.

- 프라임 회원들이 보이는 구매양상을 프라임 회원들이 구매한 상품의 이름들의 패턴으로부터 분석하여 데이터들이 ' 얼마나 프라임 회원에 가까운 구매였는가' 를 알려줄 수 있는 feature 이다.

- ['product_count']

  - 상품명의 뒷부분에서 등장하는 개수, 개, 입, 캡슐, 번들 등 단위를 기준으로 개수를 추출하여 feature 를 생성한다.

  - 이를 통해 프라임 회원이 몇 개로 구성된 제품을 선호하는지 알아보고자 한다.

※ 추가 feature 생성 방법론 : synthetic feature

원본 데이터에는 없지만 기존 특징을 결합하거나 변환하여 feature 를 생성한다. 이러한 합성 feature 를 만들어 알고리즘에 더 관련성이 높고 유용한 데이터를 제공해 알고리즘의 성능을 향상시킬 수 있다.[6]

- Feature importance features

: 여러 모델들을 사용해 Feature importance 가 공통적으로 높은 피처들을 선별해 그 피처들 끼리의 연산을 진행한 피처를 추가해 중요한 피처에 대한 데이터의 가중을 높이거나, 중요도가 높은 피처들을 범주화를 진행해 새로운 피처들을 생성한다.

  - 이때, 모델에서 중요도가 높은 컬럼들 간의 결합이 numeric 일지, categorical 일지 기획단계에서는 정확히 알 수 없다.

이렇게 feature 들을 수정 및 생성을 진행한 뒤, 각 feature 들을 categorical feature 와 numeric feature 로 구분한 뒤 categorical feature 에는 Encoding 을 numeric feature 에는 Scaling 을 진행한다.

ii. 전처리 방식

1. Encoding – categorical features

[6] [System and Method for Prediction Using Synthetic Features and Gradient Boosted Decision Tree]

- product_item - Binary Encoding :

    - 위에서 언급했듯, product_item feature 는 제품의 행사·브랜드·용량·개수를 제외한 순수 제품의 품목을 나타내는 feature 이다.
    - 해당 feature 를 통해 얻고자 하는 정보는 "프라임 회원이 어떤 순수 제품을 선호하는가" 이므로, encoding 이후에도 각각의 value 가 자신의 고유한 값을 표현할 수 있어야 한다.
    - CJ 더마켓 에서는 다양한 종류의 제품을 다루기 때문에, 해당 column 은 high - cardinality categorical feature 일 것으로 예상된다. 그러한 경우, Binary encoding 을 사용하여 높은 cardinality 의 value 들에 각각을 나타낼 수 있는 번호를 1~n 까지 매기고 해당 수를 이진변환하여 할당된 이진수의 자리수 별 column 을 새롭게 생성한다. (해당 과정에서 특정 value 에 대한 가중은 발생하지 않는다.)

- promotion - Target Encoding :
    - promotion 은 product_name 에서 대괄호 안에 위치한 CJ 더마켓에서 진행하는 행사를 뽑아 프라임회원의 여부에 주는 영향을 확인하기 위한 feature 이다.
    - 257 개의 제품명을 scraping 하는 과정에서 임의로 생성한 promotion feature 에서의 cardinality 는 4 였다.
    - promotion 은 각각 값의 의미를 담고있어야 한다는 encoding 의 주요 목적과 cardinality 가 낮을 것이라는 예측을 기반으로 target encoding 을 실시한다. 데이터들의 타겟값을 사용하여 각 value 의 의미를 유지하고, low - cardinality feature 이기 때문에 target 값을 사용하므로써 나오는 data leakage 문제는 최소화된다.

- gender - One-hot Encoding :
    - gender 는 M/F 로 두개의 값을 가지는 Categorical feature 이므로 데이터를 완전히 보전하는 one - hot encoding 을 사용한다.

2. Scaling – numeric features

- 모든 feature 들의 데이터 분포나 범위를 동일하게 조정 함으로써 모델링에 도움을 준다.
- 스케일링이 필요한 수치형 feature 인 net_order_qty 와 net_order_amt 는 스케일링이 진행되었다고 사전에 고지되어 그 두개의 수치형 변수는 제외하고, 새로 생성된 feature 들 중에 수치형 변수들이 많으니 스케일링을 진행한다.

3. Sampling

- employee_yn : Y 인 경우

- 임직원들의 데이터이기 때문에 prime 회원과 아닌 회원의 차이는 총 데이터를 가지고 EDA 를 진행한 뒤 Sampling 을 고려해야 하지만 임직원의 데이터가 많이 부족할것으로 판단되어 적은량의 데이터를 handling 할때에도 성능 상승을 기대할 수 있는 sampling 을 사용할 것이다.[7]
- Sampling 기법으론 UnderSampling 보다 정보 손실이 적고, 분류 정확도가 높은 OverSampling 방법을 채택하여 데이터 셋의 균형을 맞춘다.
- OverSampling 기법 중에서도 참고문헌[8]의 내용을 바탕으로 다른 기법보다 성능이 우수하다고 알려진 SMOTE 기법을 사용해 EDA 진행 후 클래스 불균형이 일어난 방향으로 데이터를 늘린다.

- employee_yn : N 인 경우
  - 임직원들이 아닌 고객들의 데이터이기 때문에 앞서 언급했듯 충분히 많을 것이라 판단되지만, 일반 고객들의 데이터는 클래스 불균형이 확실하다고 판단되어 Sampling 사용할 것이다.
  - 위와 같은 이유로 OverSampling 기법을 사용할 것이며 그 중 SMOTE 기법을 사용해 일반회원중 prime 회원의 데이터를 늘리는 Sampling 을 진행할 것이다.

4. Clustering

- K-means 와 같은 함수를 사용해 Clustering 을 진행해서 데이터를 군집화한 feature 를 추가해 분류에 도움을 주는 방향을 도모해본다.

iii. 모델링 방식

1. employee_yn : Y 인 경우
   - 데이터가 상대적으로 매우 적을 것으로 예상되므로, 파라미터 튜닝으로 성능을 높이는 것보다는 여러 분류기의 도움을 받아 앙상블 모델링을 진행하는 것이 더 효과적일 것으로 판단된다.
   - 최근 데이콘에서 작은 크기의 데이터로 진행된 competition 에서 가장 높은 성적을 달성한 팀을 참조[9]해, sklearn 에 내장된 모든 Classifier 및 Boosting 계열 모델, 딥러닝 모델 등의 모델들을 all_algorithm 을 사용해 성능을 평가하고, 가장 성능이 좋은 모델을 선정한 후 VotingClassifier 모델을 최종으로 사용하여 앙상블 모델링을 진행한다.

2. employee_yn : N 인 경우
   - 데이터가 충분히 많은 경우의 모델링으로, product_name 이 가장 cardinality 가 높은 feature 이므로 이를 활용한 Feature Extraction 을 진행한다.

---

[7] https://www.kaggle.com/code/rafjaa/dealing-with-very-small-datasets

[8] ["Deterministic oversampling methods based on SMOTE", 2019]

[9] https://dacon.io/competitions/official/236035/overview/description

- product_name 은 변수도 많고 범주형 변수의 비중이 높기 때문에, 범주형 변수의 비중이 높은 데이터셋에서 예측 성능이 우수하다고 알려진 CatBoost 모델을 우선적으로 사용하여 다양한 파라미터 튜닝을 진행해본다.

두 경우를 모두 고려하여, Machine Learning 진행 시에는 어떤 모델링이 가장 좋다는 판단은 어렵다. 따라서, 다양한 모델들을 사용해보고 성능을 비교해보며 진행하는 것이 가장 효과적일 것이다. 또한, black box 문제로 인해 수많은 모델들의 내부를 완벽하게 확인할 수 없고, 모든 데이터는 동일하지 않기 때문에 최선의 결과를 도출해내기 위해선 계속해서 시도하고 실험해보는 것이 불가피하다.

IV. 예상 기대효과 및 방향성 제시

CJ 더마켓의 고객 정보를 분석해 정회원과 임직원 프라임회원을 예측함으로써, 다음과 같은 예상 기대효과가 있을 수 있다.

1. 정확한 고객 타겟마케팅
- 일반 회원과 prime 회원 분류를 통해 해당 회원들의 성향을 파악하고, 해당 정보를 이용해 적합한 제품 및 서비스를 제공할 수 있으며, 타겟 마케팅 효과를 극대화할 수 있다.

2. 고객 만족도 향상
- 고객 분류 및 맞춤형 서비스 제공으로 인해 고객의 만족도가 높아지고, 이에 따라 재구매율 및 매출이 증대될 수 있다.

3. 비즈니스 경쟁력 강화
- 정확도가 높은 예측 모델을 활용하여 CJ 더마켓은 경쟁 업체들과 차별화된 서비스를 제공할 수 있다.

4. 비용 절감
- 정확한 예측을 통해 CJ 더마켓은 불필요한 마케팅 비용을 절감할 수 있다.

CJ 더마켓은 일반회원에서 프라임회원 가입을 촉진시키기 위해 다음과 같은 마케팅 방향성을 제시할 수 있다.

- 프라임회원을 예측하는 모델의 일반화성능을 높인다면 프라임회원으로 분류 했지만 실제로는 프라임회원이 아닌 회원이 있을 때(Confusion Matrix 에서 FP 에 해당한다.) 그 회원을 ‘ 프라임회원으로 전환 가능성이 높은 회원’으로 간주할 수 있다. FP 에 해당하는 회원에게 더욱 적극적인 광고와 프로모션을 통해 프라임 회원의 수를 늘릴 수 있다고 기대한다.