2 0 2 3

CJ THE MARKET
고객 예측 모델링

쓰 리 라 찹 | 김지은, 김채원, 이현준, 천예은

CONTENTS

01

분석 배경 및 개요

- 분석 목적 및 필요성
- 스토리보드

01. 분석 배경 및 개요

분석 목적 및 필요성

효과적 마케팅 전략 수립

고객 맞춤형 서비스 제공

고객 이탈 감소

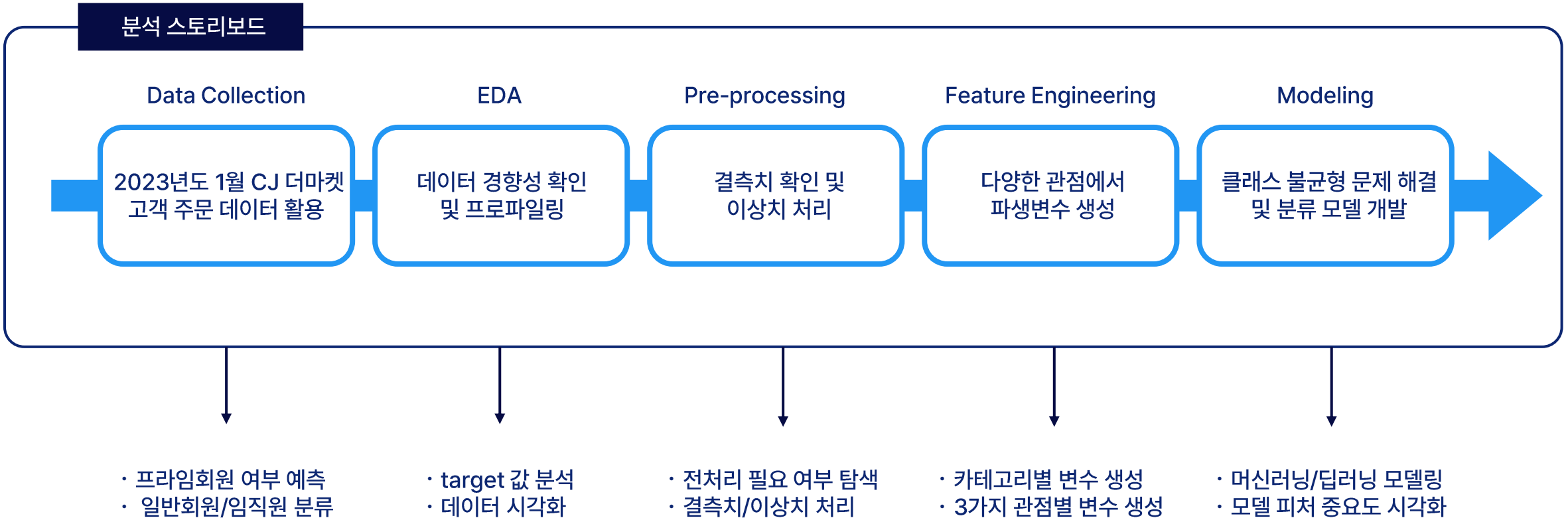
매출 증대

→ **CJ 더마켓 프라임 회원 여부를 예측함**으로써 다양한 비즈니스 인사이트 도출

→ **고객의 구매 성향 파악**을 위한 EDA 진행

01. 분석 배경 및 개요

스토리보드



02

탐색적 데이터 분석

- 공통 특성
- target 값 분석
- 임직원/비임직원 분석
- 공통 특성 분석

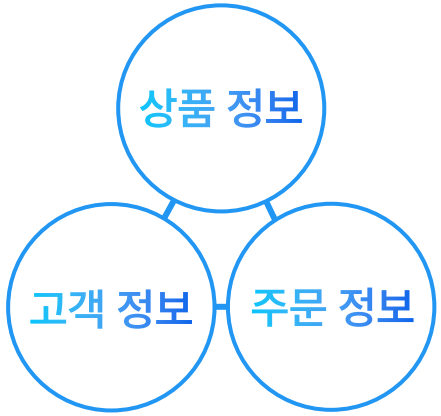
02. 탐색적 데이터 분석

공통 특성

공통 특성에 따라 컬럼 분류

- 3가지 특성(상품 정보, 주문 정보, 고객 정보) 추출
- 컬럼 별 특성에 따라 target을 제외한 7가지 컬럼을 3가지 관점으로 분류

- 프라임회원 여부 결정에 영향을 미치는 요인들을 기준으로 이후 EDA 과정을 진행



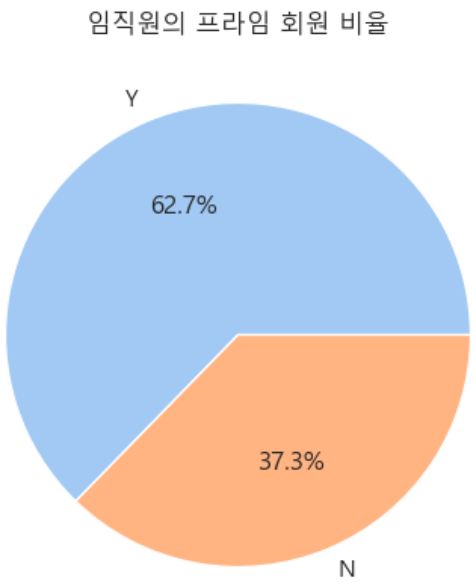
컬럼 분류 표

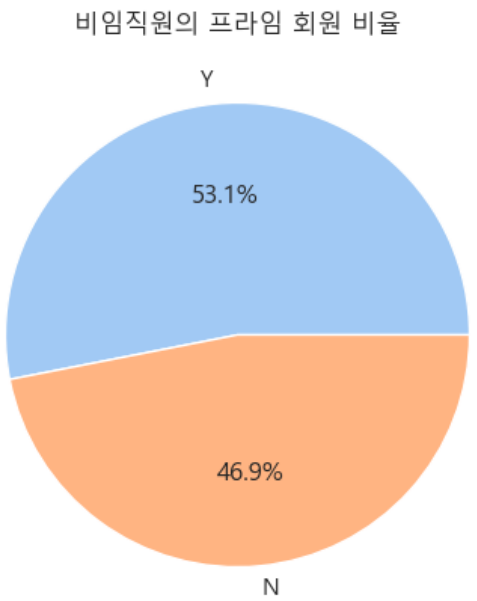
상품 정보	product_name, net_order_qty, net_order_amt
고객 정보	gender, age_grp
주문 정보	scd, order_date

02. 탐색적 데이터 분석

target 값 분석

임직원 / 비임직원 target 값 분석





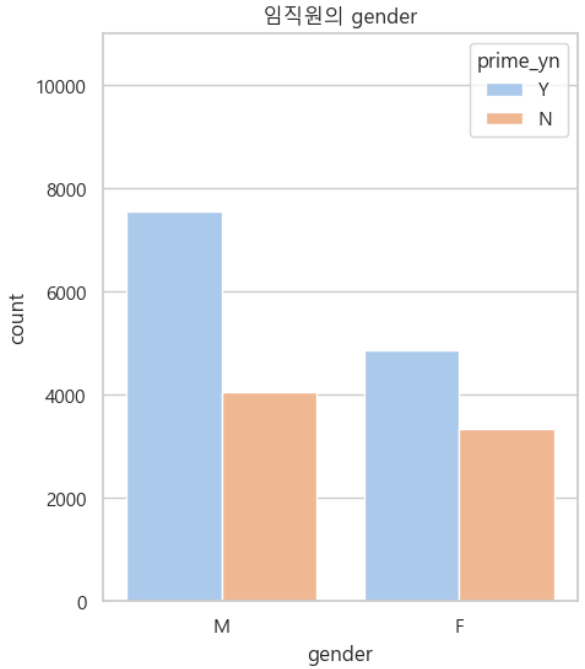
- 임직원의 경우, 프라임 회원의 비율이 더 높음
- 비임직원의 경우, 프라임 회원의 비율이 소폭 더 높음

→ 프라임 회원 비율이 비교적 균형을 이루므로 임직원, 비임직원의 데이터셋이 **불균형 하지 않다**고 판단

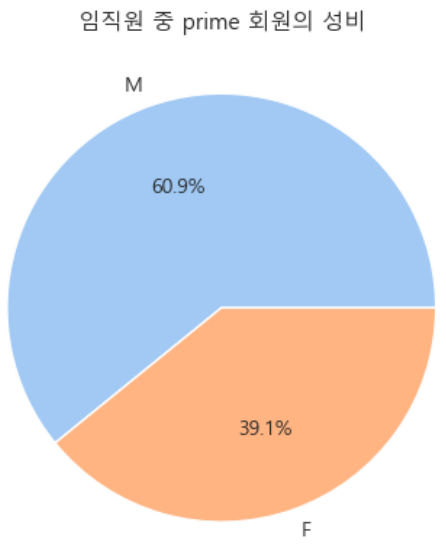
02. 탐색적 데이터 분석

임직원 데이터셋 분석

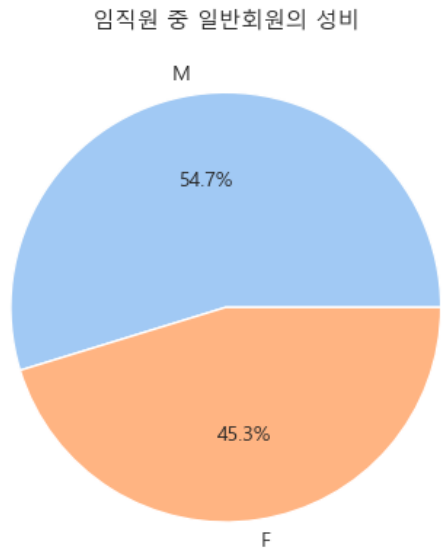
임직원 데이터셋 분석 – gender



・임직원의 성비는 약 6:4로 남성이 여성보다 많음



・임직원 중 프라임 회원의 성비는
약 6:4로 남성이 여성보다 많음



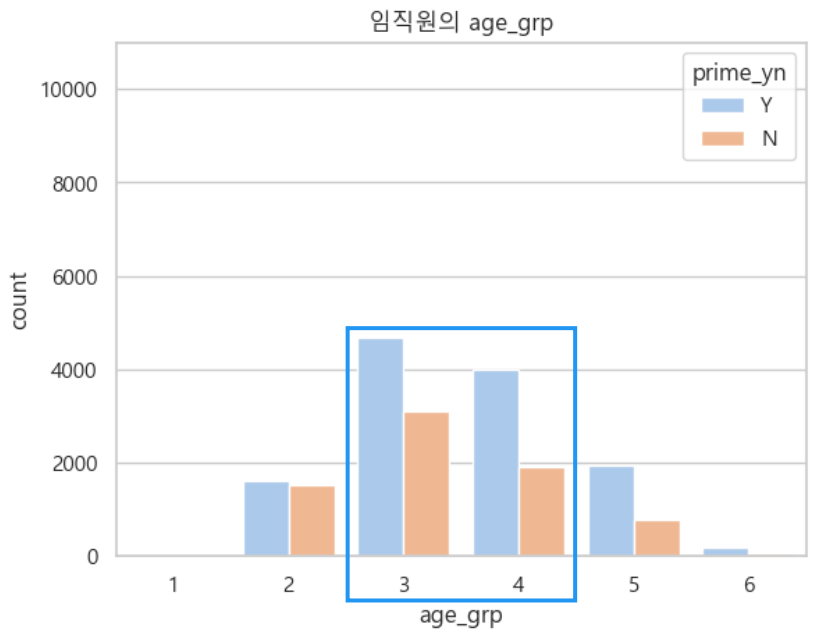
・임직원 중 일반회원의 성비는
약 5.5:4.5로 큰 차이가 없음

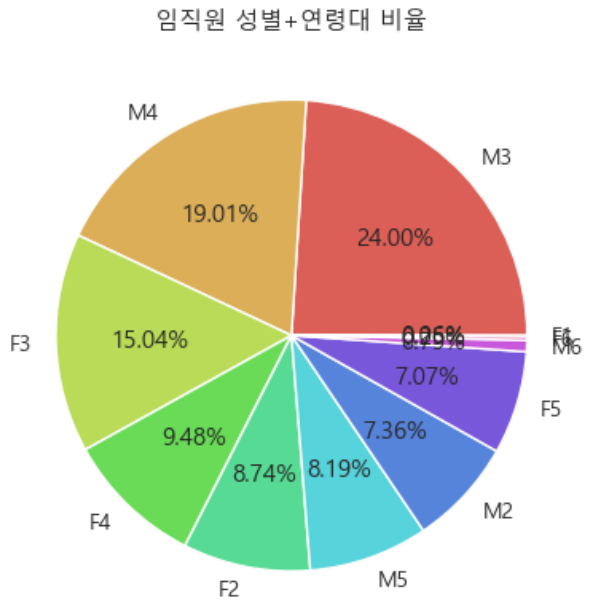
임직원 중 프라임 회원의 **주요 성별은 남성**임을 확인

02. 탐색적 데이터 분석

임직원 데이터셋 분석

임직원 데이터셋 분석 – age_grp





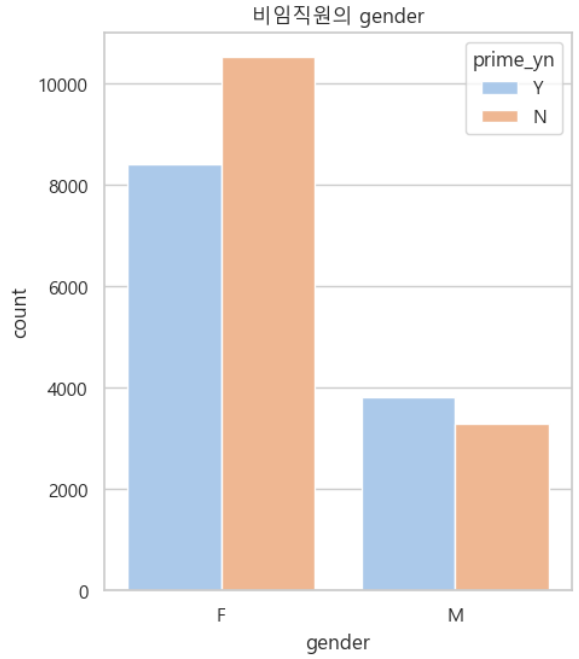
- 30대 남성
- 40대 남성
- 30대 여성
- 40대 여성

임직원 중 프라임 회원의 **연령대는 30, 40대가 70%**를 이루었으며, **30대 남성**이 가장 많았음

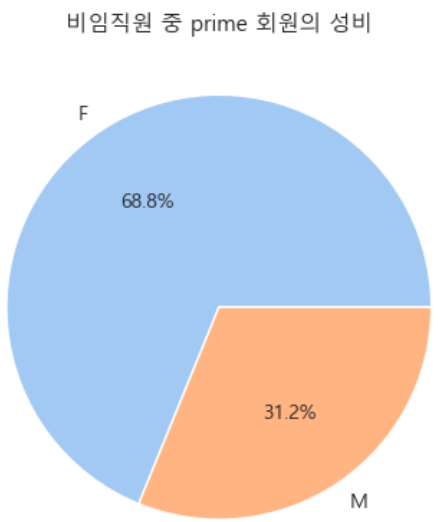
02. 탐색적 데이터 분석

비임직원 데이터셋 분석

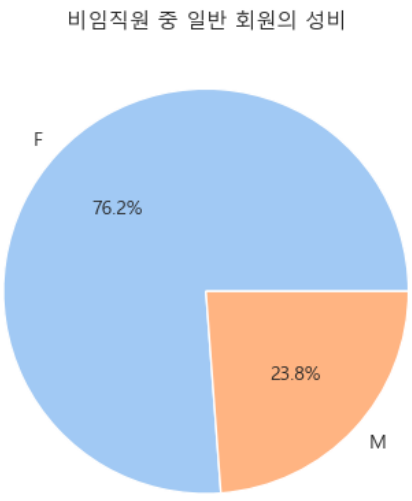
비임직원 데이터셋 분석 – gender



・비임직원의 성비는 약 7:3으로 여성이 남성보다 많음



・비임직원 중 프라임 회원의 성비는 약 7:3으로 여성이 남성보다 많음



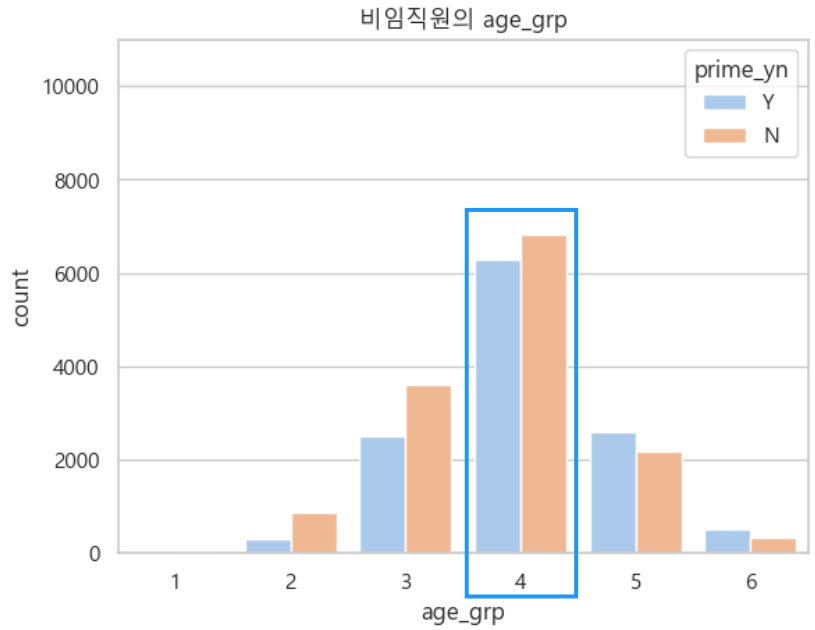
・비임직원 중 일반회원의 성비는 약 7.5:2.5로 여성이 남성보다 많음

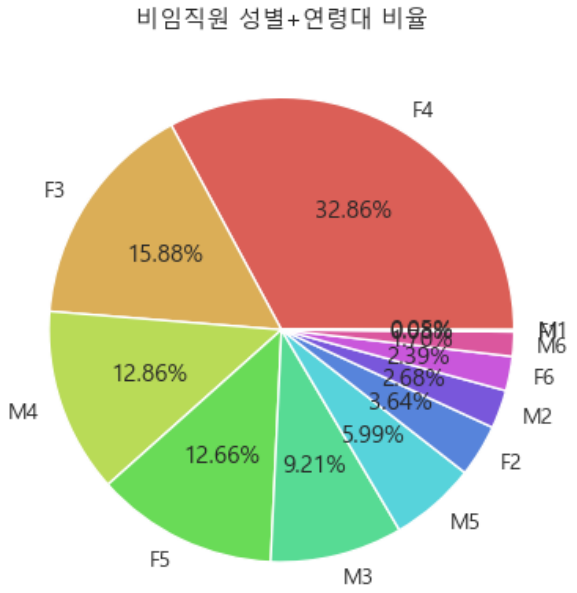
비임직원 중 프라임 회원의 **주요 성별은 여성**임을 확인

02. 탐색적 데이터 분석

비임직원 데이터셋 분석

비임직원 데이터셋 분석 – age_grp





- 40대 여성
- 30대 여성
- 40대 남성
- 50대 여성

비임직원 중 프라임 회원의 **연령대는 40대가 51%** 를 이루었으며, **40대 여성**이 가장 많았음

02. 탐색적 데이터 분석

임직원/비임직원 분석

임직원 데이터셋 EDA 결과

임직원 중 프라임 회원의 **연령대는 30, 40대가 70%**를 이루었으며, **30대 남성**이 가장 많았음

30대 남성 > 40대 남성 > 30대 여성 > 40대 여성 순

- CJ 제일제당 인적 자원 현황 기준, 임직원 남성의 비율이 7:3 정도로 높음
- 이마트몰 모바일 쇼핑 이용 통계에 따르면, 30대~40대의 비율이 가장 높음

비임직원 데이터셋 EDA 결과

비임직원 중 프라임 회원의 **연령대는 40대가 51%** 를 이루었으며, **40대 여성**이 가장 많았음

40대 여성 > 30대 여성 > 40대 남성 > 50대 여성 순

- 50대의 모바일 쇼핑 비중 상승률이 가파름
- 이마트몰 모바일 쇼핑 이용 통계에 따르면, 30대~40대의 비율이 가장 높음

02. 탐색적 데이터 분석

공통 특성 분석

1. 상품 정보 종합 관점

product_name

- 전체 : 3113
- 임직원 : 2553
- 비임직원 : 1878

net_order_qty

- 임직원 : 88, 138 등의 이상치 존재
- 비임직원 : 180, 198 등의 이상치 존재

net_order_amt

- 임직원 / 비임직원간 큰 차이가 존재하지 않음

임직원의 경우, **[임직원] 상품이 존재**하기 때문에 상품 수 차이가 존재하는 것으로 추정

net_order_qty 컬럼의 경우, **이상치 처리** 필요

02. 탐색적 데이터 분석

공통 특성 분석

구매 정보

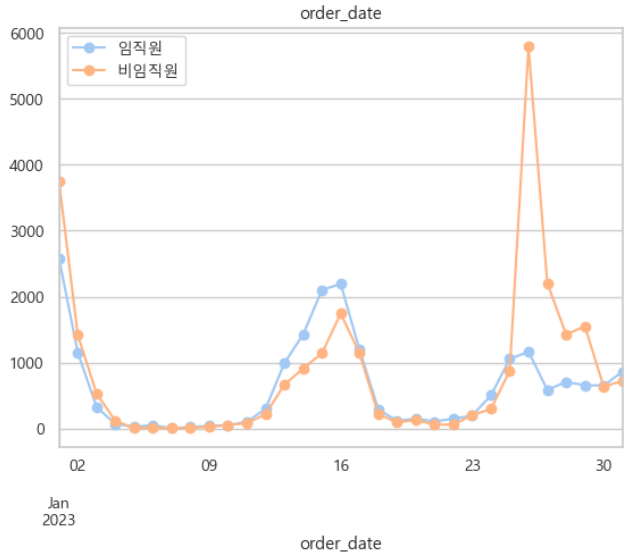
2. 구매 정보 종합 관점

scd

- 전체 : 10653
- 임직원 : 4713
- 비임직원 : 5940

order_date

- 전체 : 1일 주문이 가장 많고, 13~17일에는 주문이 증가 추세를 보임
- 임직원 : 1, 16, 15, 14, 17일 순으로 주문이 많음을 확인
- 비임직원 : 26, 1, 27, 16, 29일 순으로 주문이 많음을 확인



임직원과 비교하여, **비임직원의 주문건수가 더 많음**을 확인

임직원의 경우, **1일과 설 연휴 일주일 전**의 주문량이 많고, **비임직원**은 **1일과 월말**에 주문이 많음

03

데이터 전처리

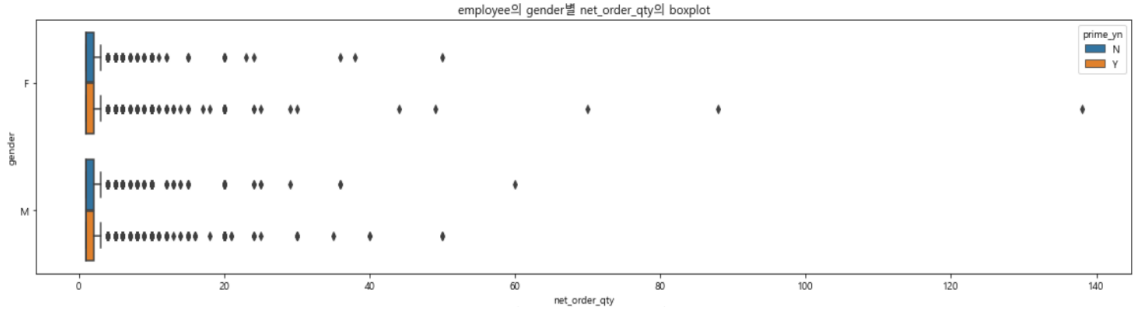
- 이상치 처리
- 피처 엔지니어링

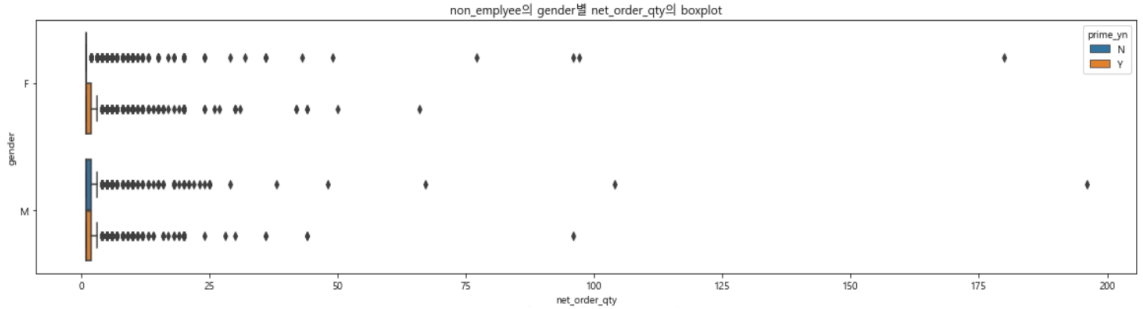
03. 데이터 전처리

이상치 처리

이상치 처리

- EDA를 통해 이상치가 의심되는 net_order_qty의 피처를 확인함
- Boxplot EDA를 통해 이상치 자세히 확인






```
employee['net_order_qty'] = employee['net_order_qty'].apply(lambda x : 50 if x >= 60 else x)
non_employee['net_order_qty'] = non_employee['net_order_qty'].apply(lambda x : 50 if x >= 55 else x)
```

- 구매수량이 1~3개로 치우쳐진 데이터
- 정성적인 방법을 사용할 경우, 데이터 손실 문제가 발생할 위험이 존재
- 따라서 데이터의 분포를 극단적으로 벗어난 값들에 대해 일정 범위 내 최댓값으로 변환함

03. 데이터 전처리

피처 엔지니어링

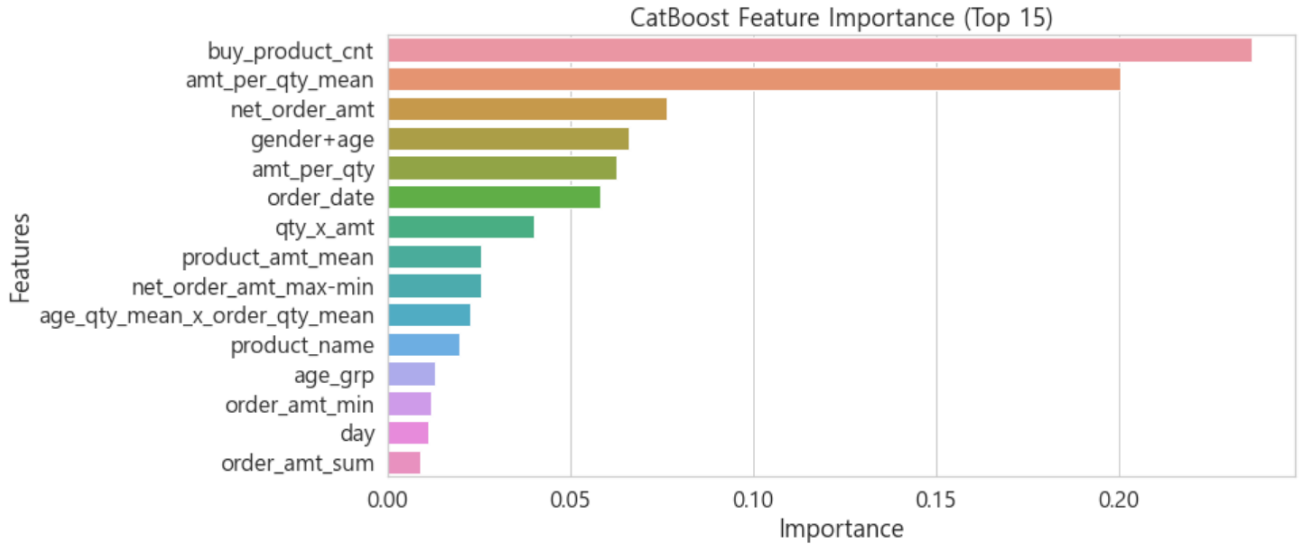
공통 특성에 따른 피처 생성

상품 정보	고객 정보	주문 정보
product_name, net_order_qty, net_order_amt	gender, age_grp	scd, order_date
・amt_per_qty: 구매 상품의 개당 가격 ・top5: 매출 높은 5가지 상품 구매 여부 ・bottom5: 매출 낮은 5가지 상품 구매 여부 ・product_newyear: 설날 상품 구매 여부 ・product_qty_max: 상품별 구매 개수의 최댓값 ・net_order_amt_max_min: 상품별 구매 금액의 max-min ・qty_x_amt: 상품 개수 x 상품 가격 ・cnt_order_mean: 구매품목수 x 구매 상품 개당 가격의 평균	・age_qty_mean: 나이별 상품 수량의 평균 ・age_qty_max: 나이별 상품 수량의 최댓값 ・age_amt_mean: 나이별 상품 가격의 평균 ・age_amt_max: 나이별 상품 가격의 최댓값 ・gender+age: 성별 + 나이	・day: 날짜 ・day_label: 주차 ・order_weekend: 주말 구매 여부 ・holiday: 공휴일 구매 여부 ・holiday_prev7d: 공휴일 일주일 전 구매 여부 ・week for 10: 월 초중말(10일 단위)

→ 컬럼별 특성에 따라 **3가지 관점(상품 정보, 고객 정보, 주문 정보)**으로 분류한 후, 피처 생성

03. 데이터 전처리
피처 엔지니어링

Feature importance top 5



< Top 3 >

1. buy_product_cnt : 1인당 구매수량(scd별)
2. amt_per_qty_mean : 구매상품의 개당 가격의 평균
3. net_order_amt : 상품 가격

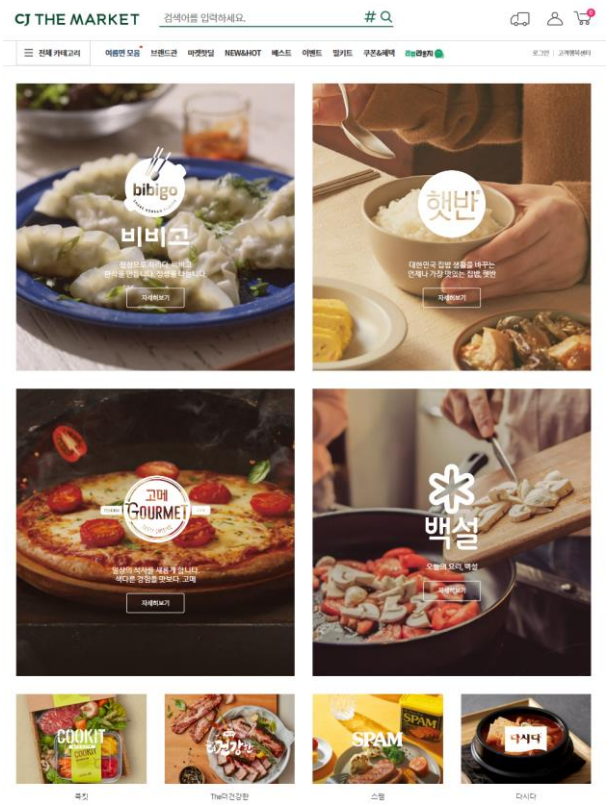
→ 공통점 : Top3 모두 **상품 정보**와 관련된 피처

→ 3가지 정보 중 **상품 정보에 가장 민감**하며, 특히 **수치형 변수**가 중요도에 많은 영향을 끼침

03. 데이터 전처리

피처 엔지니어링

구매자의 특성에 따른 product_name의 세분화



구매자의 특성에 따른 product		
1	brand	브랜드 표기 상품 구매 여부
2	비비고	비비고 상품 구매 여부
3	햇반	햇반 상품 구매 여부
4	스팸	스팸 상품 구매 여부
5	고메	고메 상품 구매 여부
6	삼호	삼호 상품 구매 여부
7	백설	백설 상품 구매 여부
8	head_employee	[임직원] 상품 구매 여부
9	1box	1box 상품 구매 여부
10	Water	(아이시스, 삼다수, 에비앙) 구매 여부
11	닭가슴살	닭가슴살 구매 여부
12	Oil	Oil 상품 구매 여부
13	mandu	만두 상품 구매 여부
14	hotdog	핫도그 상품 구매 여부
15	soup	국, 탕, 찌개 상품 구매 여부
16	bab	볶음, 주먹, 비빔밥 상품 구매 여부
17	fried	튀김 상품 구매 여부
18	seaweed	김 상품 구매 여부
19	auto_delivery	자동배송 top10

03. 데이터 전처리

피처 엔지니어링

	원본 피처	생성한 피처			
1	scd	age_qty_sum	order_amt_max	product_newyear	order_qty_sum
2	product_name	age_qty_mean	order_amt_min	whether_rice	order_qty_mean
3	net_order_qty	age_qty_max	order_amt_std	비비고	order_qty_max
4	net_order_amt	age_qty_std	order_amt_lam	햇반	order_qty_std
5	gender	age_qty_lam	order_amt_max_x_order_qty_max	스팸	order_qty_lam
6	age_grp	age_amt_sum	amt_x_qty	고메	top5
7	order_date	age_amt_mean	net_order_amt_max-min	삼호	order_weekend
8		age_amt_max	amt_per_qty	백설	holiday
9		age_amt_min	amt_per_qty_mean	brand	holiday_prev7d
10		age_amt_std	product_amt_mean	1box	week for 10
11		age_amt_lam	product_qty_max	water	soup
12		age_order_mean x age_qty_mean	buy_product_cnt	닭가슴살	bab
13		age+gender	day	oil	fried
14		order_amt_sum	day_label	mandu	kim
15		order_amt_mean	bottom5	hotdog	auto_delivery

04

모델링

- 데이터 검증 방법
- 머신러닝 모델링
- 딥러닝 모델링
- 모델 시각화

04. 모델링

데이터 검증 방법

Data Split

- Model Tuning으로 인한 과적합 방지 및 정확한 검증을 위해 Validation Dataset 구성
- Validation dataset의 size는 0.3으로 설정

Stratified-Kfold를 통한 교차검증

- 분류 문제(0과 1로 예측)에 적합
- 데이터 레이블 분포의 균형성 회복
- 임직원 dataset의 프라임회원 여부 비율은 약 6:4

04. 모델링

머신러닝 모델링

Random Forest

Dataset(임직원)	F1_Score
train	0.845
validation	0.866

Dataset(비임직원)	F1_Score
train	0.759
validation	0.791

Why Forest Model?
- overfitting 예방 및 정확도 향상
- classification 모델에서 유용

LightGBM

Dataset(임직원)	F1_Score
train	0.873
validation	0.874

Dataset(비임직원)	F1_Score
train	0.788
validation	0.794

CatBoost

Dataset(임직원)	F1_Score
train	0.878
validation	0.885

Dataset(비임직원)	F1_Score
train	0.799
validation	0.819

Why Boosting Model?
- weight 조정 가능
- 불균형 데이터에서 좋은 성능

04. 모델링

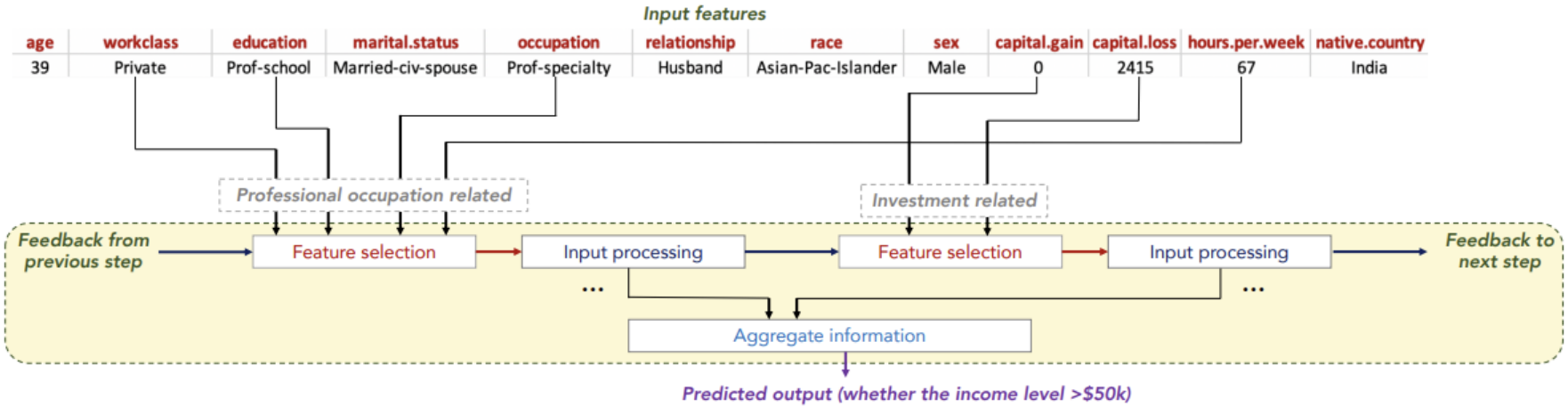
딥러닝 모델링

TabNet

Dataset(임직원)	F1_Score
train	0.743
validation	0.709

Dataset(비임직원)	F1_Score
train	0.669
validation	0.649

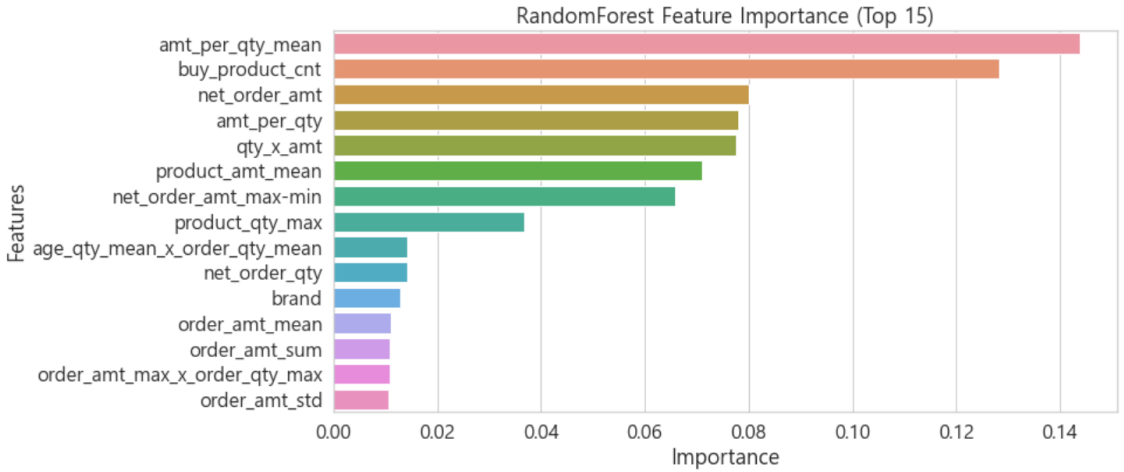
Why TabNet?
- 학습 수행 시, 어떠한 feature에 집중할지를 모델이 스스로 결정
- 편향이 개입될 여지가 줄어듦

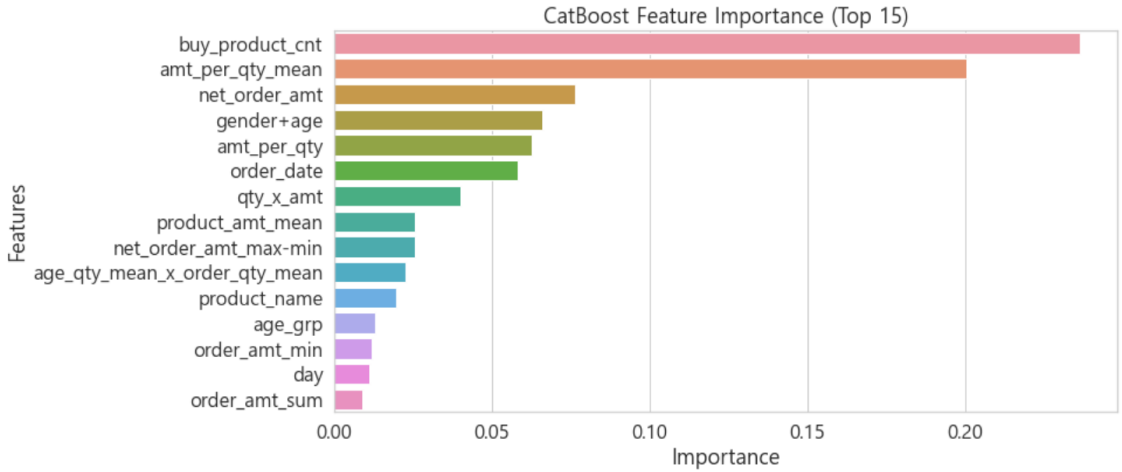


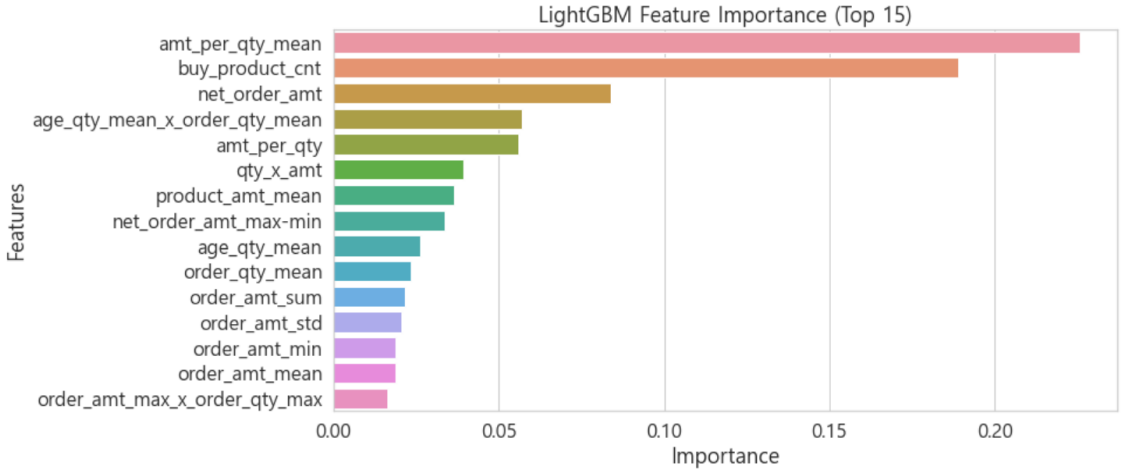
- Transformer와 Tree 구조를 결합한 딥러닝 모델
- 정형 데이터를 처리하는 데 사용될 수 있는 최초의 딥러닝 모델

04. 모델링

모델 시각화







- 3가지 모델에 대한 Feature Importance를 확인한 결과, 상품정보와 관련된 피처들의 중요도가 높게 다루어지는 것을 확인함
- 모든 모델들의 Top3 피처는 'amt_per_qty_mean', 'buy_product_cnt', 'net_order_amt'로 동일함

05

결론

결론

변수 측면

- 모델들의 Feature Importance를 통해 알 수 있듯이, 상품 정보는 프라임 회원 예측에 중대한 영향을 미친다.
- 구매 수량이 증가하고 구매 가격이 높아질수록 프라임 회원일 가능성이 높아짐을 알 수 있다.
- 따라서, 상품 정보를 기반으로 의미 있는 피처를 적절히 생성한다면, 프라임 회원 여부를 예측하는데 효과적일 것이다.

- 분석 결과에 따르면, 평균 개당 구매 가격이 낮은 소비자가 프라임 회원인 경향성을 보인다.
- 즉 프라임 회원들은 다양한 할인 혜택을 받기 때문에, 평균 개당 구매 가격이 일반 회원에 비해 낮음을 확인할 수 있다.

- 비임직원의 경우, 프라임 회원 중 40대 여성의 비율이 가장 높았으며, 일반회원 역시 40대 여성의 비율이 높았다.
- 즉 프라임 회원으로 전환 가능한 주요 타겟은 40대 여성이 될 것이다.

결론

모델 측면

- 보다 나은 성능을 위해서는 **경량화 된 모델**을 선택하는 것이 좋다.
- **피처의 개수를 줄이거나 모델 구성을 단순화**하여 **과적합을 줄이는 방향**으로 진행하는 것이 좋을 것으로 판단된다.

- 이를 통해 지금보다 가벼운 모델을 구축한다면, 예측 성능을 향상시키는데 효과적일 것으로 보인다.

THANK YOU

김지은 - kimje1101@naver.com
김채원 - clkimcw@gmail.com
이현준 - dlguswns522@naver.com
천예은 - tpdk9556@naver.com