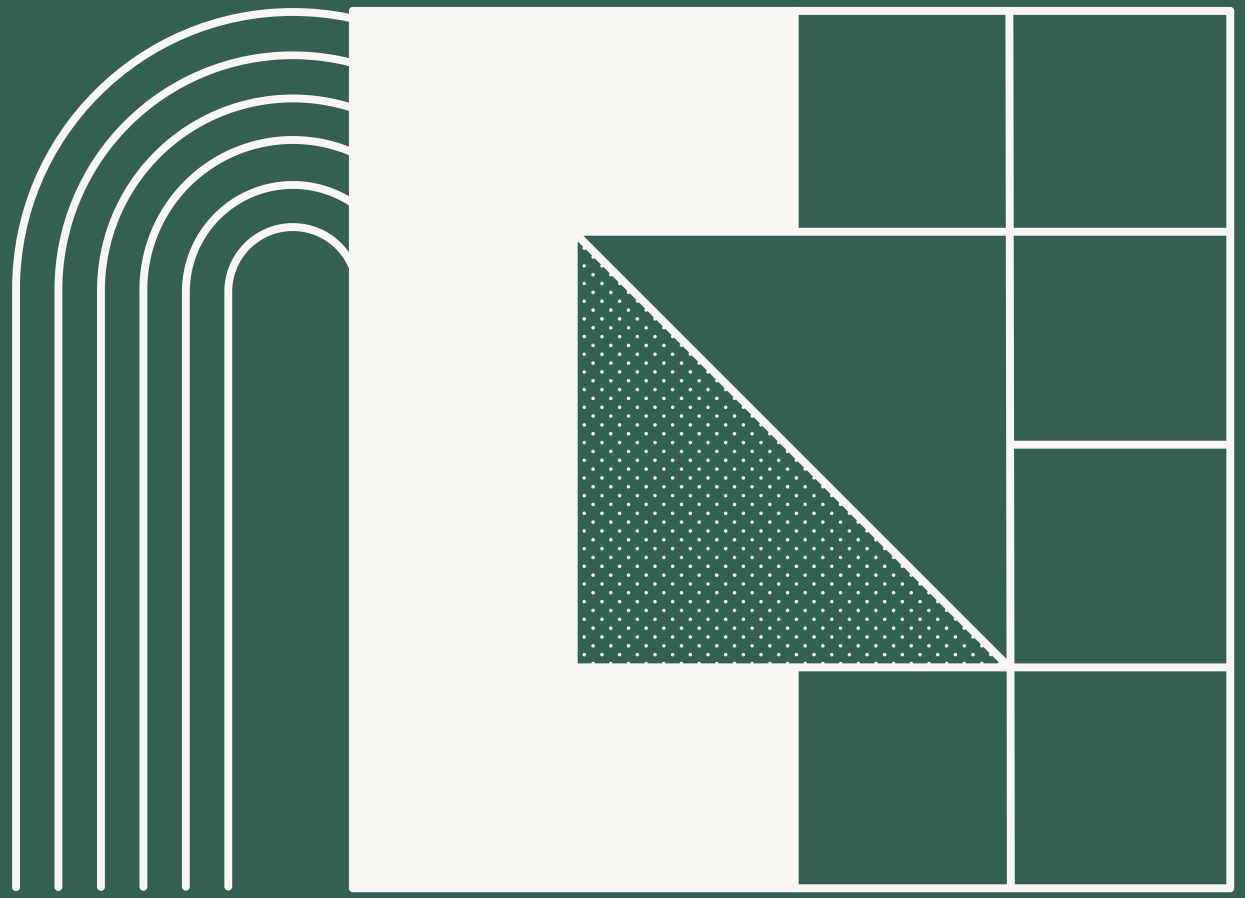
2023 BDA 데이터 분석·활용 공모전

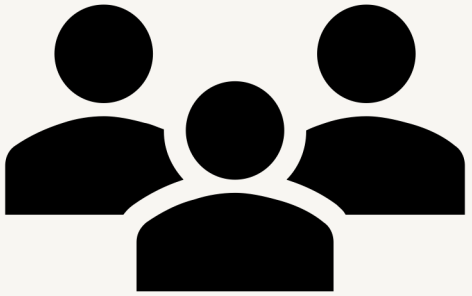
제일세당
김준, 강지아, 김동하, 송윤주

목차

분석 개요



프라임 회원이란?

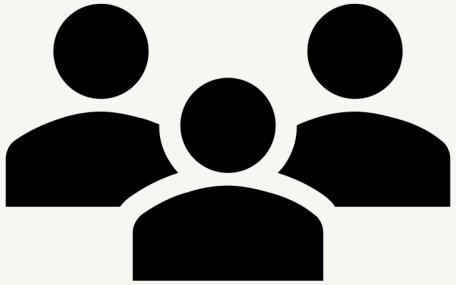


THE프라임 멤버십

- 가입 / 갱신 시 기프트 카드 지급
(월간 이용권 2천원, 연간 이용권 2만 4천원)
- 7% 무제한 할인

(23년 1월 기준)

프라임 회원 예측



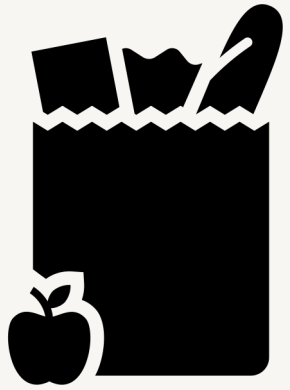
THE프라임 멤버십

WHY?

프라임 회원 예측을 통해 고정적인 고객(단골 고객)의 양상을 파악해,
이를 비즈니스 전략 수립에 이용

분석 방향

상품 조합 분석



프라임 회원이
주로 구매하는
상품 조합 파악

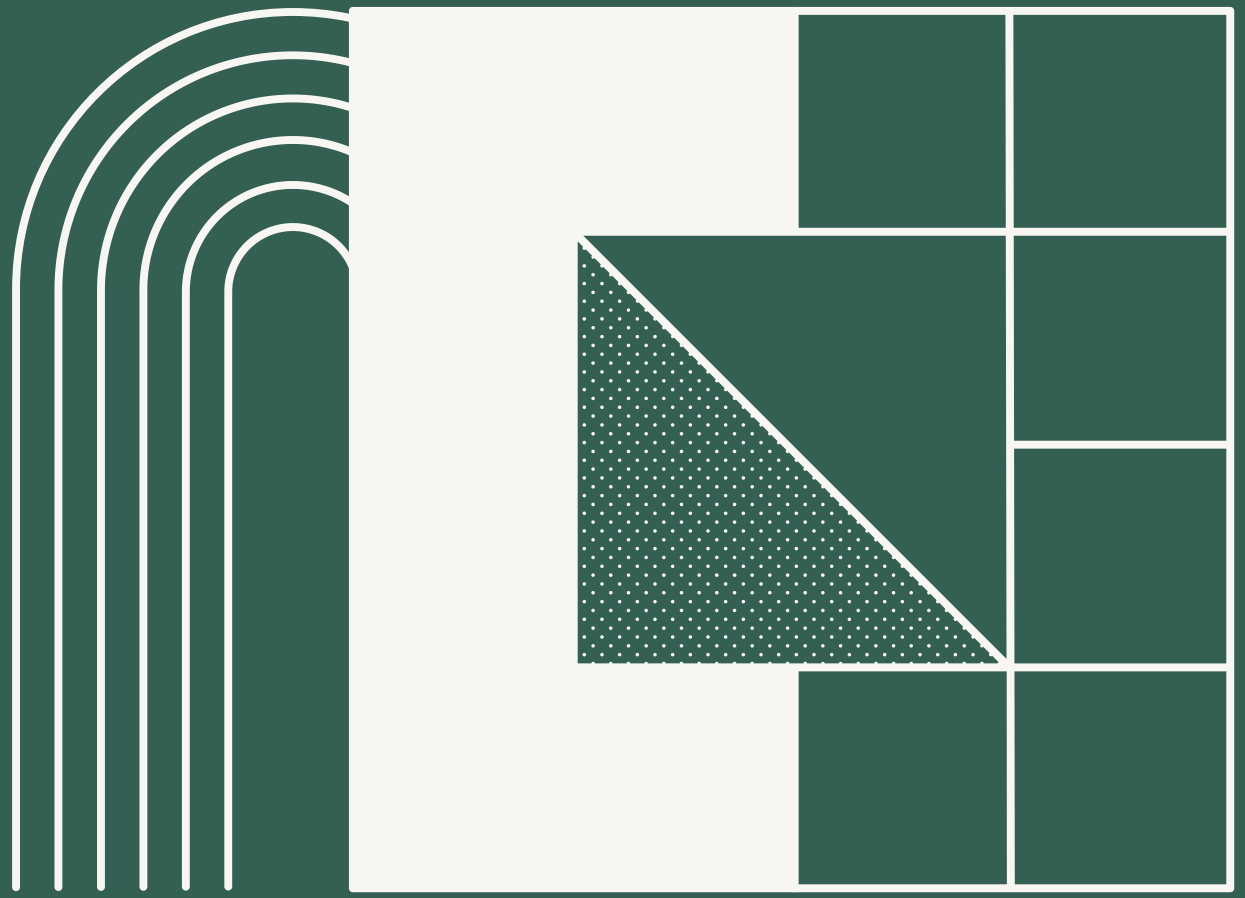
단일 상품 분석

프라임 회원이
주로 구매하는 상품의
특징(시기, 할인율 등) 파악

가설 소개

가설 1	프라임 혜택이 7% 할인이므로 할인율이 프라임 회원 예측에 중요한 요소일 것이다.
가설 2	프라임 회원은 [냉동], [셰프컬렉션]과 같은 괄호 요소가 있는 제품을 선호할 것이다.
가설 3	주차 및 요일에 따라 주문 수에 차이가 있다.

데이터 전처리



파생변수 생성

PRICE

주문금액(net_order_amt)에 Exponential(e^x)을 취한 값

net_order_amt
8.667336
9.933483
8.667336
10.771512
9.266815

Exponential

price
5810.0
20609.0
5810.0
47644.0
10581.0

FIXED_PRICE

주문가격(price)을
주문수량(net_order_qty)으로 나눠 단가를
구한 다음 상품당 최대 단가를 정가로 설정

product_name	price	net_order_qty	fixed_price
비비고 한마리고등어구이 140g	125641.0	20	6282.5
비비고 한마리고등어구이 140g	62866.0	15	6282.5
비비고 한마리고등어구이 140g	20984.0	5	6282.5
비비고 한마리고등어구이 140g	17965.0	3	6282.5
비비고 한마리고등어구이 140g	4197.0	1	6282.5

DISCOUNT_RATE

$$1 - \frac{\text{주문가격(price)}}{\text{정가(fixed\_price)} \times \text{주문수량(net\_order\_qty)}}$$

price	fixed_price	net_order_qty	discount_rate
17761.0	3582.25	8	0.380243
25597.0	8533.00	3	0.000078
6485.0	5040.50	2	0.356711
13121.0	1378.00	10	0.047823
27628.0	28981.00	1	0.046686

가격 스케일링

PRICE

문제점

기존 net_order_amt를 그대로 사용할 경우 상품 개수의 변화에 따른 가격 변화가 비례하지 않아 대소비교를 할 수 없다.

product_name	net_order_qty	net_order_amt	price
햇반 소프트밀 단호박죽 420g	1	7.952967	2844.0
햇반 소프트밀 단호박죽 420g	1	8.254269	3844.0
햇반 소프트밀 단호박죽 420g	4	9.506065	13441.0

과정 및 결과

1. 값이 커질수록 값 사이의 차이가 줄어 로그 스케일로 변환된 것이라고 판단했고, 이에 따라 총 주문금액을 단순 합으로 계산할 수 없다고 판단
2. 실제 값과 로그값을 비교해 자연로그(ln)으로 스케일링 됐음을 확인
3. 값에 Exponential를 취함으로써 대소비교가 가능한 가격정보 도출

정가 추정

FIXED_PRICE

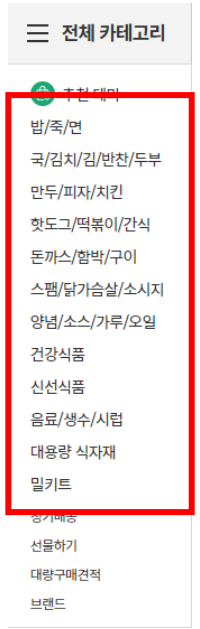
과정 및 결과

1. Exponential을 취한 주문가격(price)을 주문수량(net_order_qty)으로 나눠 단가(unit_price)를 구한다.
2. 같은 상품의 단가 중 가장 높은 금액을 정가라고 추정한다.

product_name	price	net_order_qty	unit_price	fixed_price
비비고 한마리고등어구이 140g	12565.0	2	6282.500000	6282.5
비비고 한마리고등어구이 140g	18847.0	3	6282.333333	6282.5
비비고 한마리고등어구이 140g	125641.0	20	6282.050000	6282.5
비비고 한마리고등어구이 140g	5989.0	1	5989.000000	6282.5
비비고 한마리고등어구이 140g	5989.0	1	5989.000000	6282.5

파생변수 생성

CATEGORY



product_name	category
비비고 새우계란죽 420g	음료/생수/시럽
[UPCYCLE]햇반 흑미밥 210gx24개(1box)	밥/죽/면
햇반 소프트밀 단호박죽 420g	밥/죽/면
[UPCYCLE]햇반 210gx24개(1box)	밥/죽/면
[2023설사원선물신청] CJ명가 토종김 5호	국/김치/김/반찬/두부

1. CJ더마켓에서 12개 카테고리와 상품명을 크롤링한 후, 데이터에 매핑하여 카테고리 추출
2. 라벨 인코딩(label encoding) 적용

BRACKET_FREQUENCY

product_name	bracket_in
[임직원]가쓰오우동 4인분933.2gx2개	[임직원]
[UPCYCLE]햇반 흑미밥 210gx24개(1box)	[UPCYCLE]
[2023설사원선물신청] 스팸12호	[2023설사원선물신청]
[혈행건강] 리턴업 혈행+기억력케어 60캡슐(1개월)	[혈행건강]
[60%선할인적용]BYO 면역플러스 60캡슐X2개	[60%선할인적용]

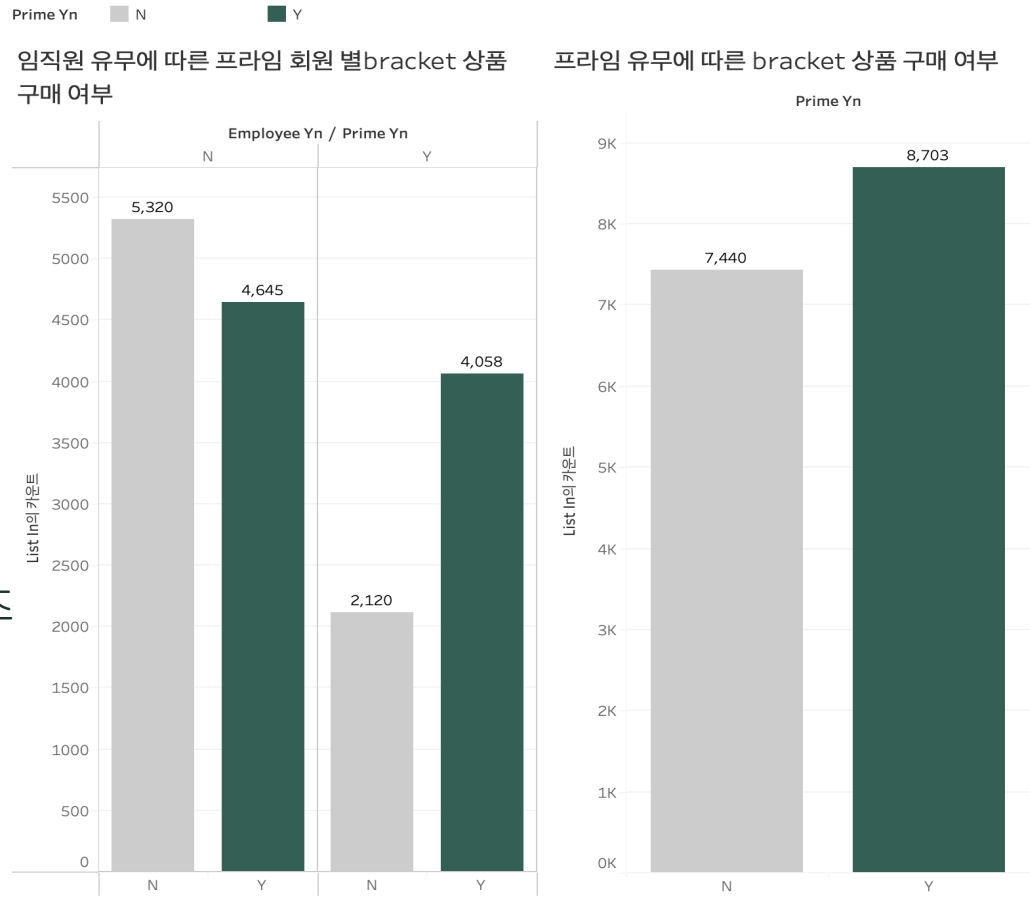
bracket_frequency
0.143929
0.020345
0.091624
0.000144
0.000072

1. 상품명(product_name)에서 괄호 안 항목 추출
2. 빈도수 인코딩(frequency encoding) 적용

괄호 요소 빈도수 측정

bracket_frequency

1. 전체 데이터에서는 괄호 요소가 프라임 예측에 중요한 요소가 아님
2. 하지만, 임직원 데이터의 경우 괄호 요소가 있는 데이터가 프라임일 확률이 2배
3. 프라임 회원 데이터 중 특정 괄호요소가 얼마나 자주 등장하는지를 빈도수 인코딩을 통해 도출



데이터 전처리

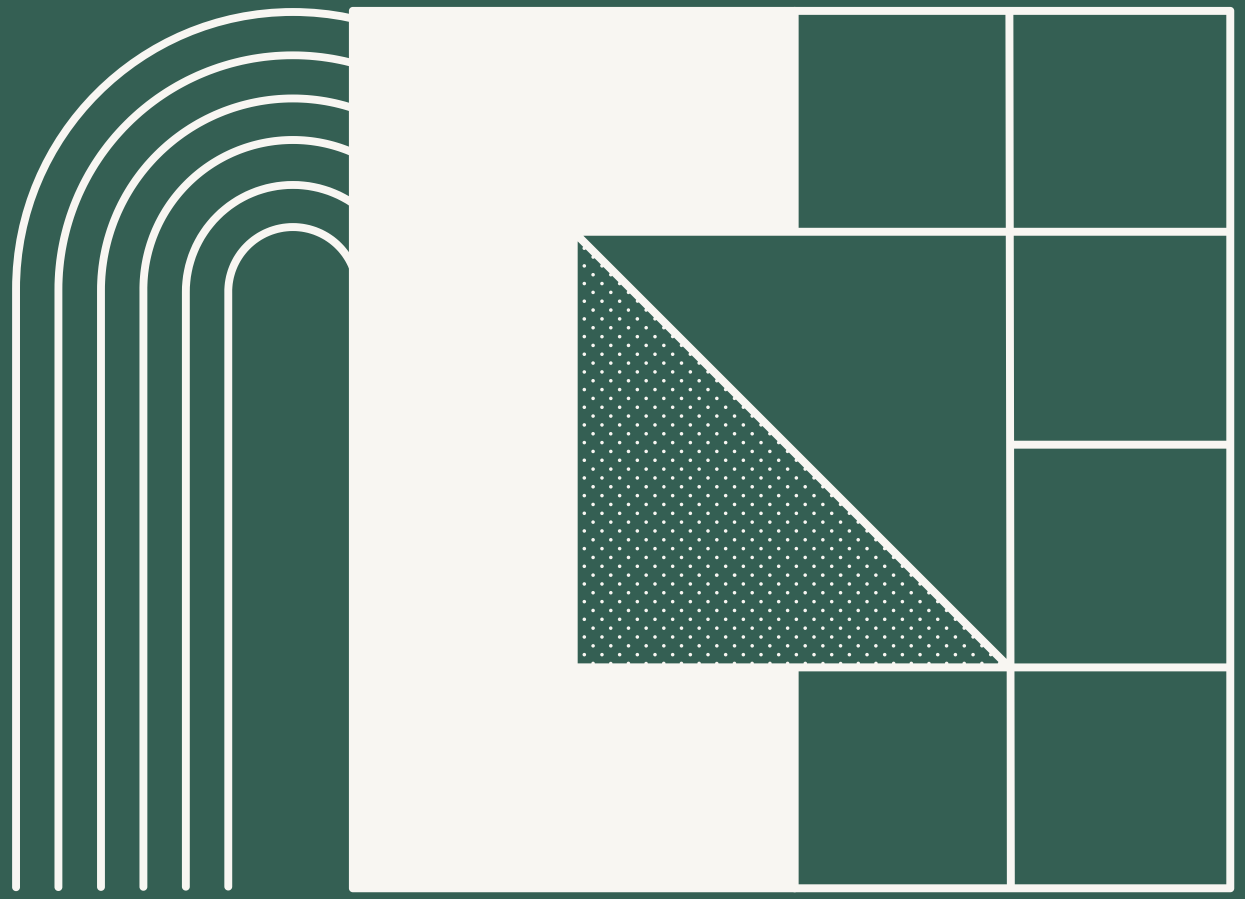
❶ train 데이터 (전처리 전)

scd	product_name	net_order_qty	net_order_amt	gender	age_grp	employee_yn	order_date	prime_yn
20230124153976	잔칫집 식혜 240ml 30입	1	9.803170	F	2	Y	20230124	N
20230124155563	백설 한입쏙 비엔나 120g*2	1	8.256607	M	3	Y	20230124	N
20230125158386	비비고 왕교자 1.05kg	1	9.348449	F	4	N	20230125	N
20230126164638	고메 바삭쫄깃한 탕수육 900g	1	9.667259	F	4	N	20230126	Y
20230125159705	햇반 매일잡곡밥210g	20	9.994653	M	4	N	20230125	Y

❷ train 데이터 (전처리 후)

gender	age_grp	category	week	holiday	price	fixed_price	discount_rate	bracket_frequency
0	1	1	5	1	18091.0	18091.00	0.000000	0.689705
1	2	7	5	1	3853.0	3853.00	0.000000	0.689705
0	1	11	5	1	7903.0	7903.00	0.000000	0.689705
0	2	6	5	1	8083.0	8083.00	0.000000	0.689705
1	2	7	5	1	17761.0	3582.25	0.380243	0.689705

모델 선택 및 훈련



모델 선택

LightGBM

XGBoost

tabnet

모델 훈련

데이터 세트 분할

1. 임직원 유무
2. Train과 Valid : 주문번호 그룹화

오브젝트 타입 인코딩

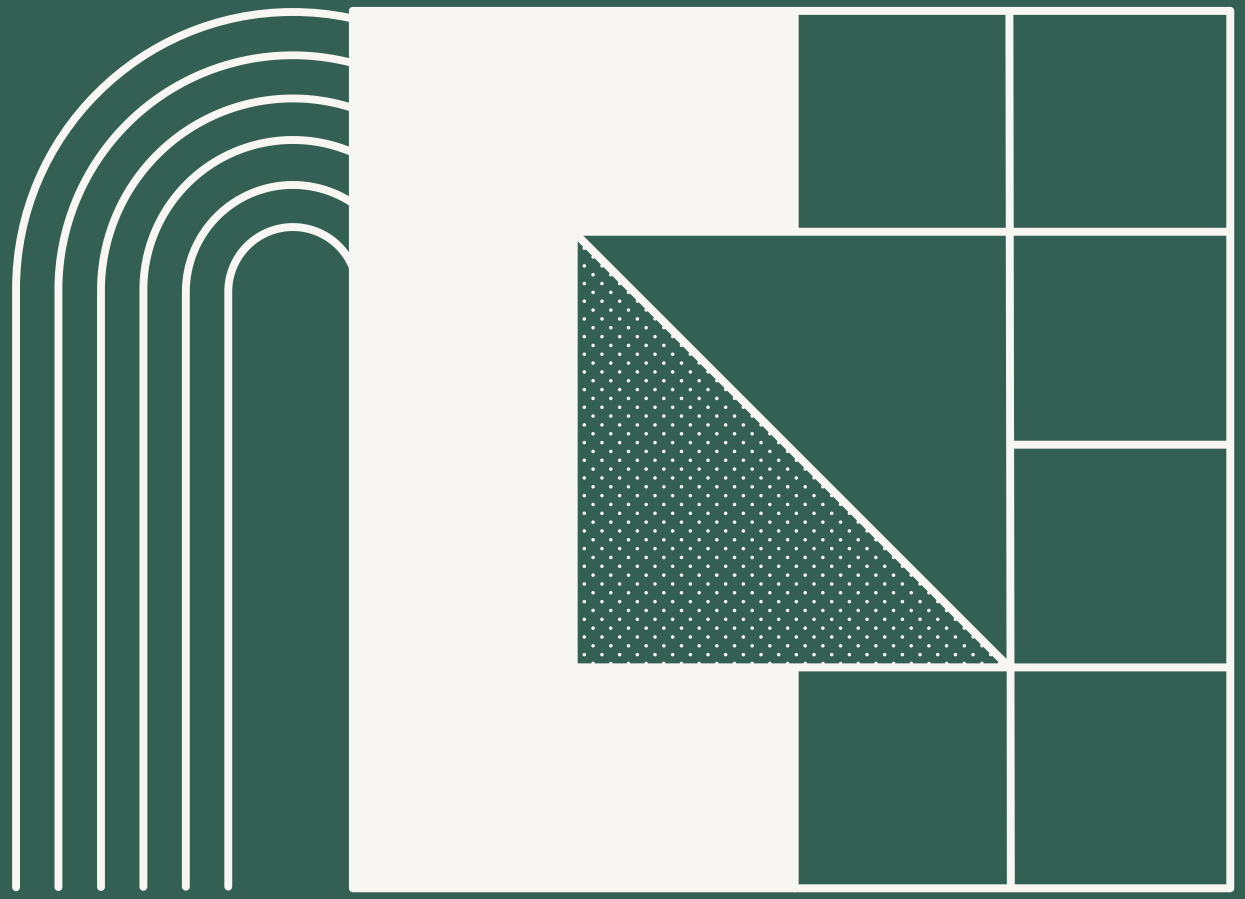
- One-hot
- Label
- Frequency

상관계수 확인

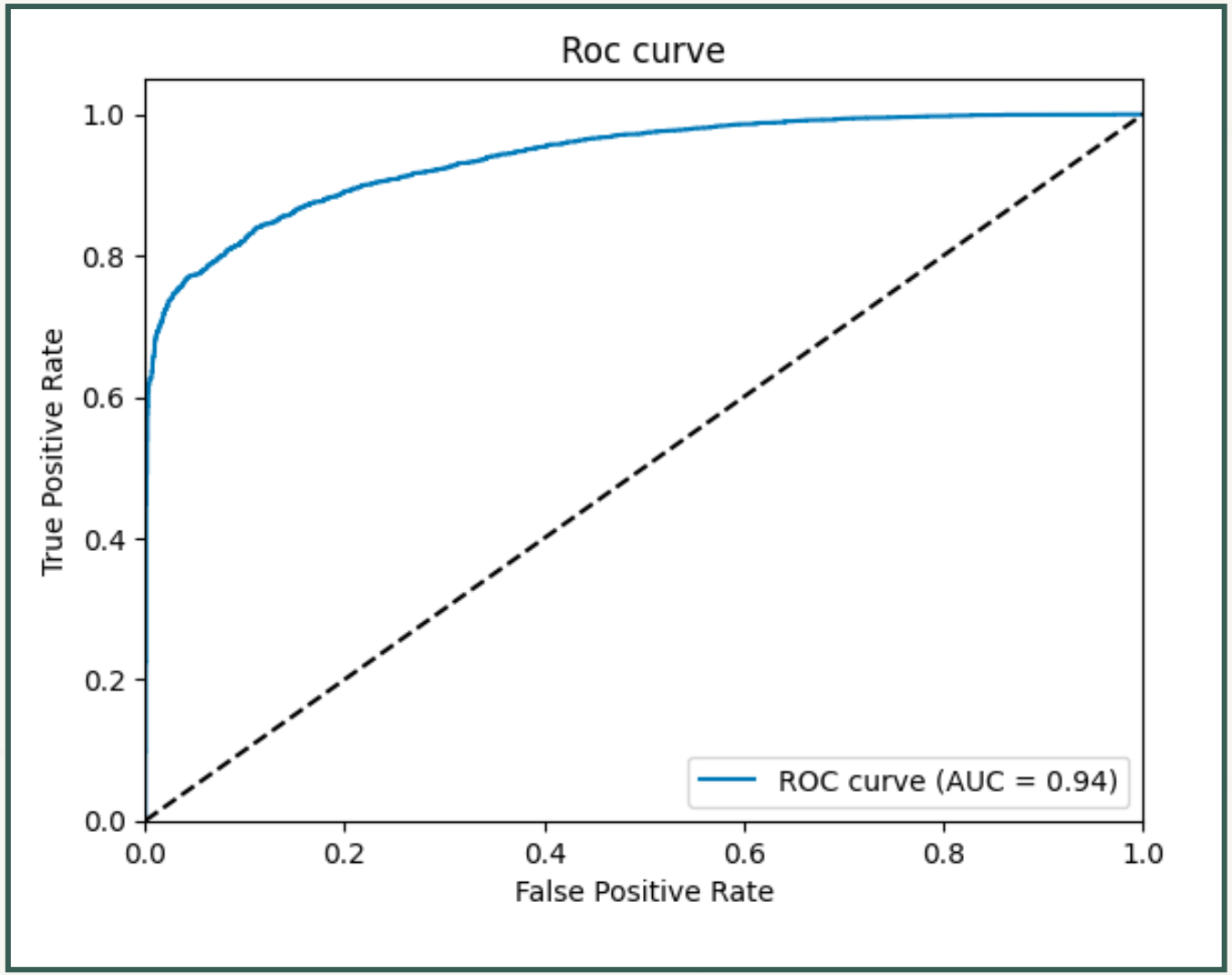
- [net_order_qty]
  price <-> P: +0.68
- [net_order_amt]
  price <-> P: +0.58
  unit_price <-> P: 0.67

모델 평가 및 최적화

(임직원 데이터 기준)



모델 평가 Baseline



- 정확도: 0.855
- 정밀도: 0.8796
- 재현율: 0.8840
- F1: 0.8818
- AUC: 0.9392

모델 최적화

'learning_rate': 0.30568863319033185

'max_depth': 11

'min_child_weight': 0

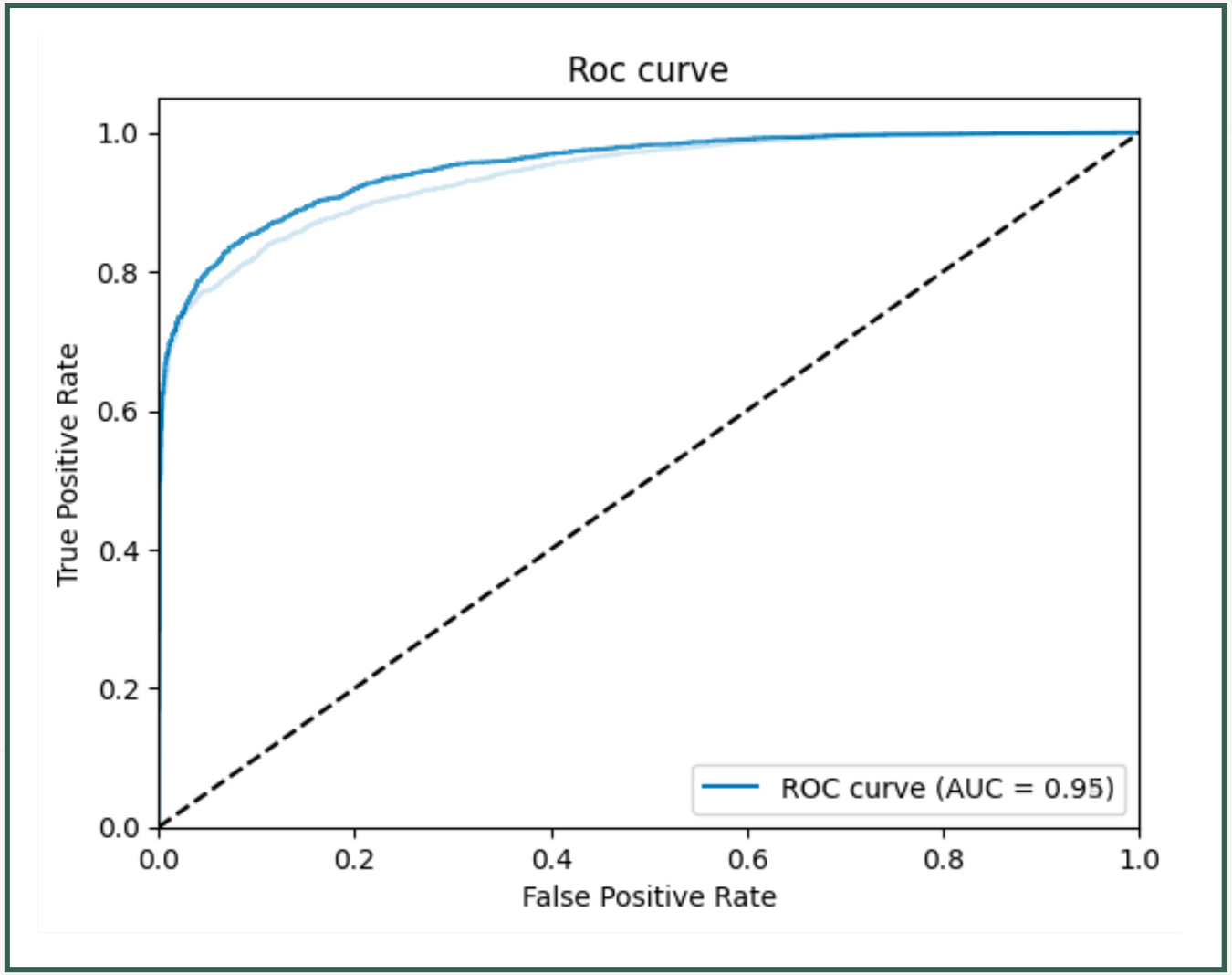
'n_estimators': 84

'reg_alpha': 0.056176064172264 4

'reg_lambda': 0.012132700049299814

scale_pos_weight': 5.552834155959141

모델 평가 최적화



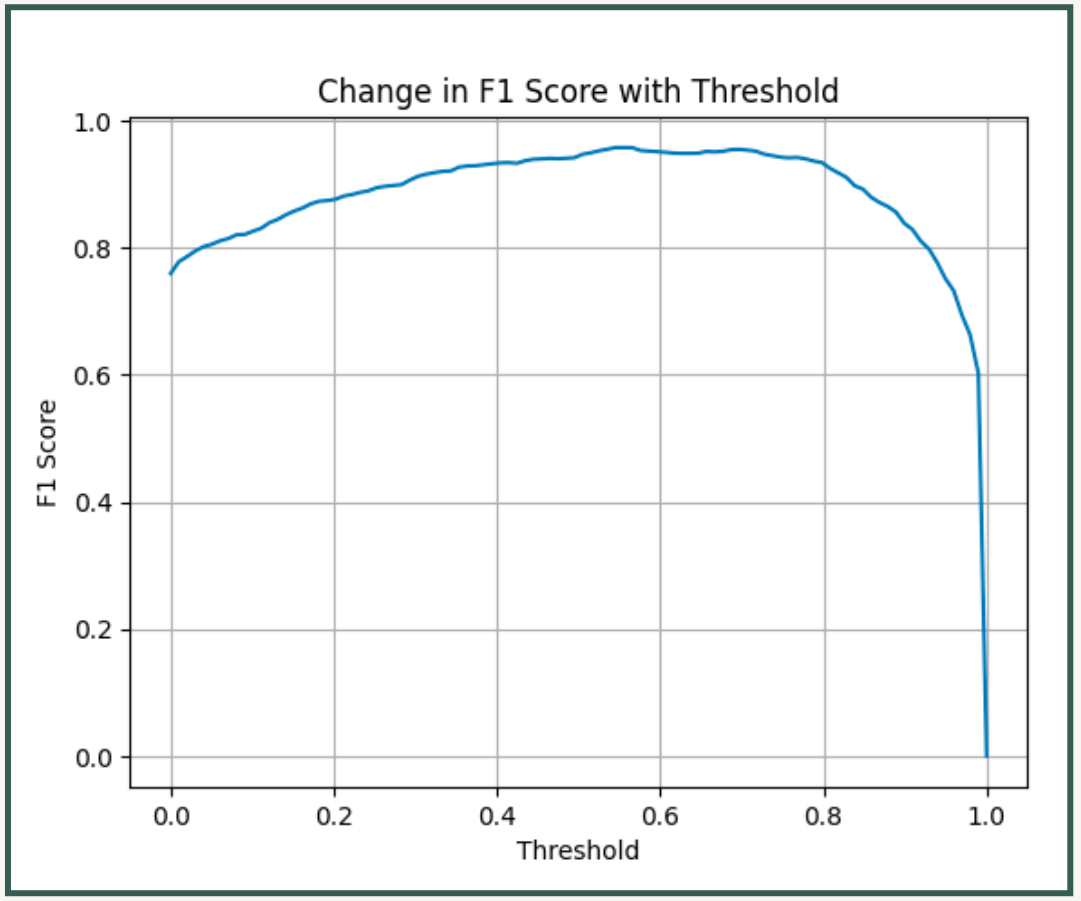
- 정확도: 0.8725
- 정밀도: 0.8770
- 재현율: 0.9208
- F1: 0.8984
- AUC: 0.9517

예측값 후처리

scd		ori_preds
20230101963226	0	0.312020
	1	0.117598
	2	0.803821
	3	0.010159
20230101963247	0	0.998577
	1	0.998342
	2	0.851543
	3	0.954682
	4	0.997341

Mean

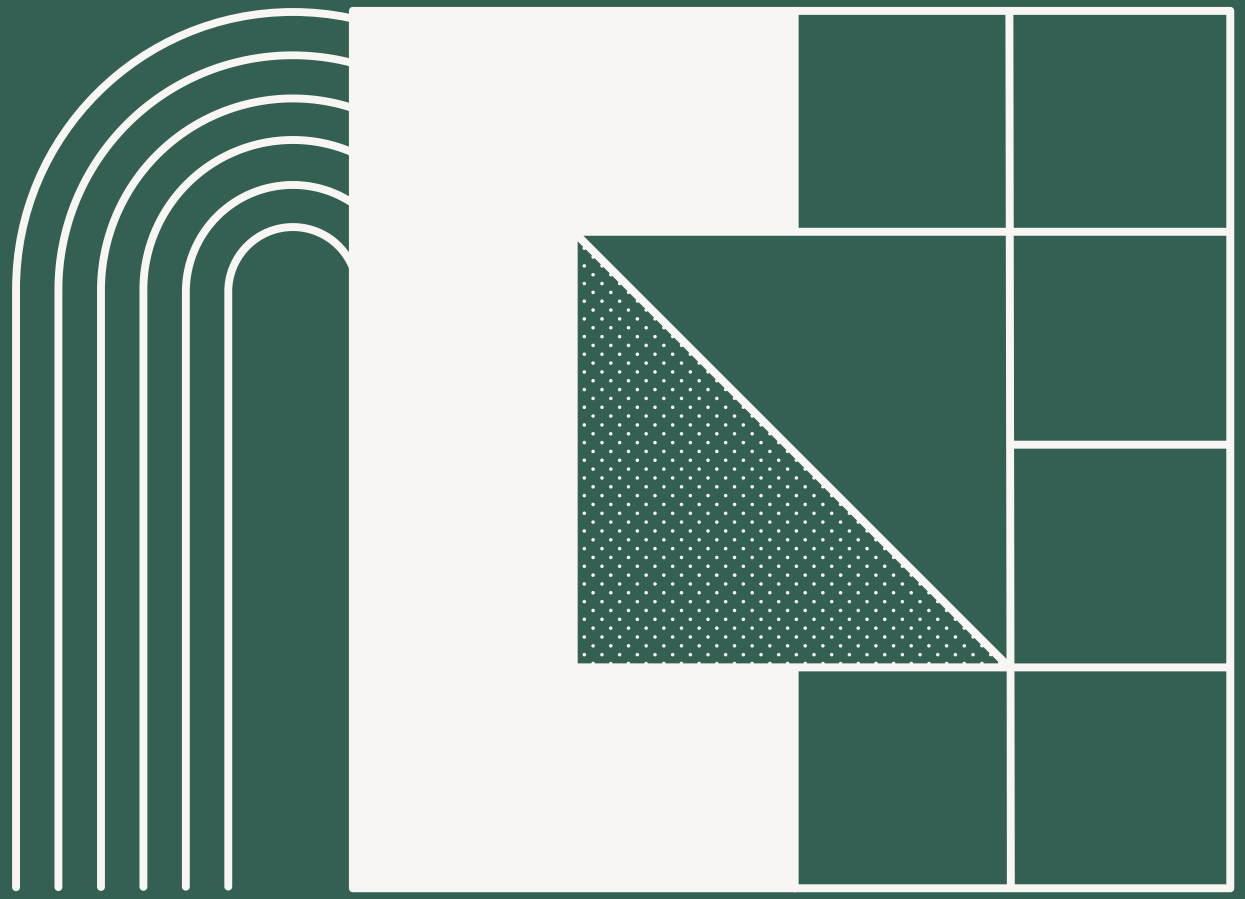
preds
0.310900
0.310900
0.310900
0.310900
0.960097
0.960097
0.960097
0.960097
0.960097
0.984513



Threshold : 0.545454

F1-score : 0.9574

모델 적용 및 해석

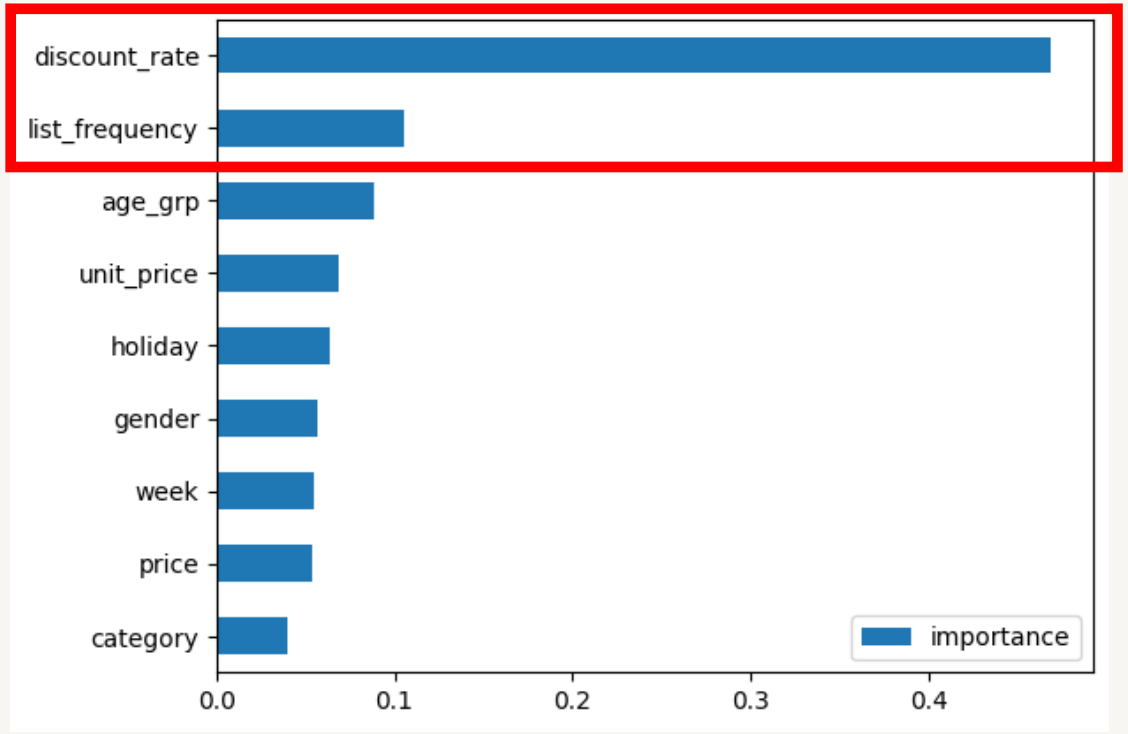


모델 적용

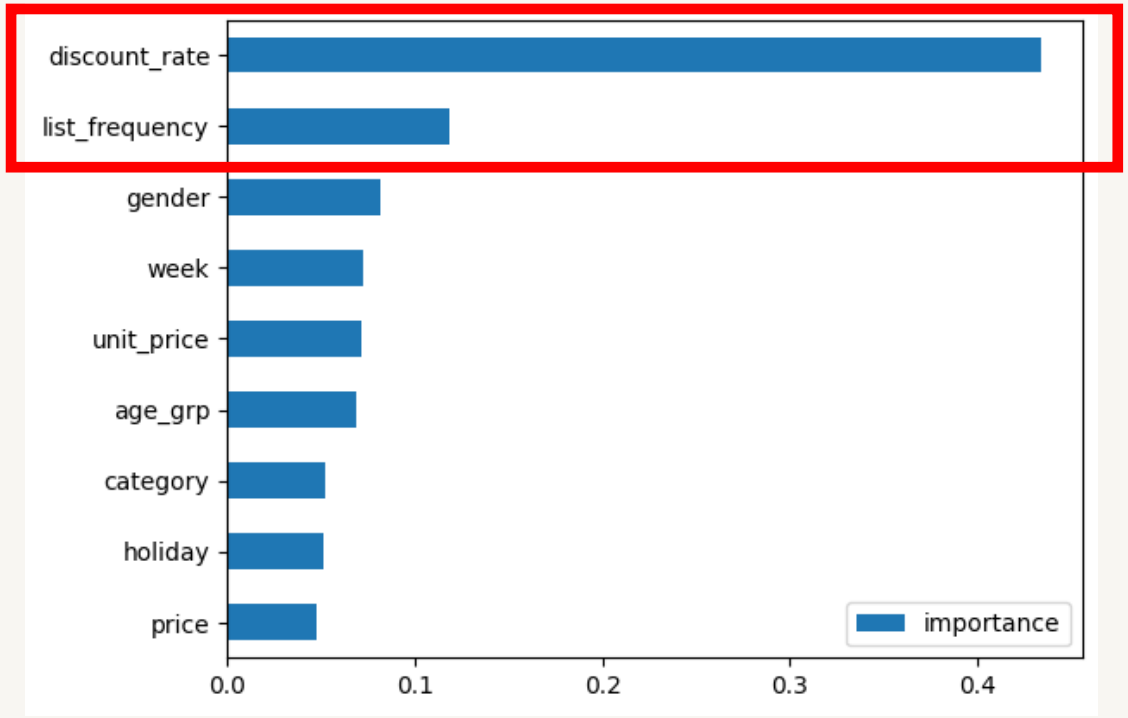
전처리 → 특성 공학 적용 → 인코딩 → 데이터 세트 분할 → 훈련된 모델 적용 → 후처리 threshold 적용

Test 데이터 예측

모델 해석 특성 중요도



임직원

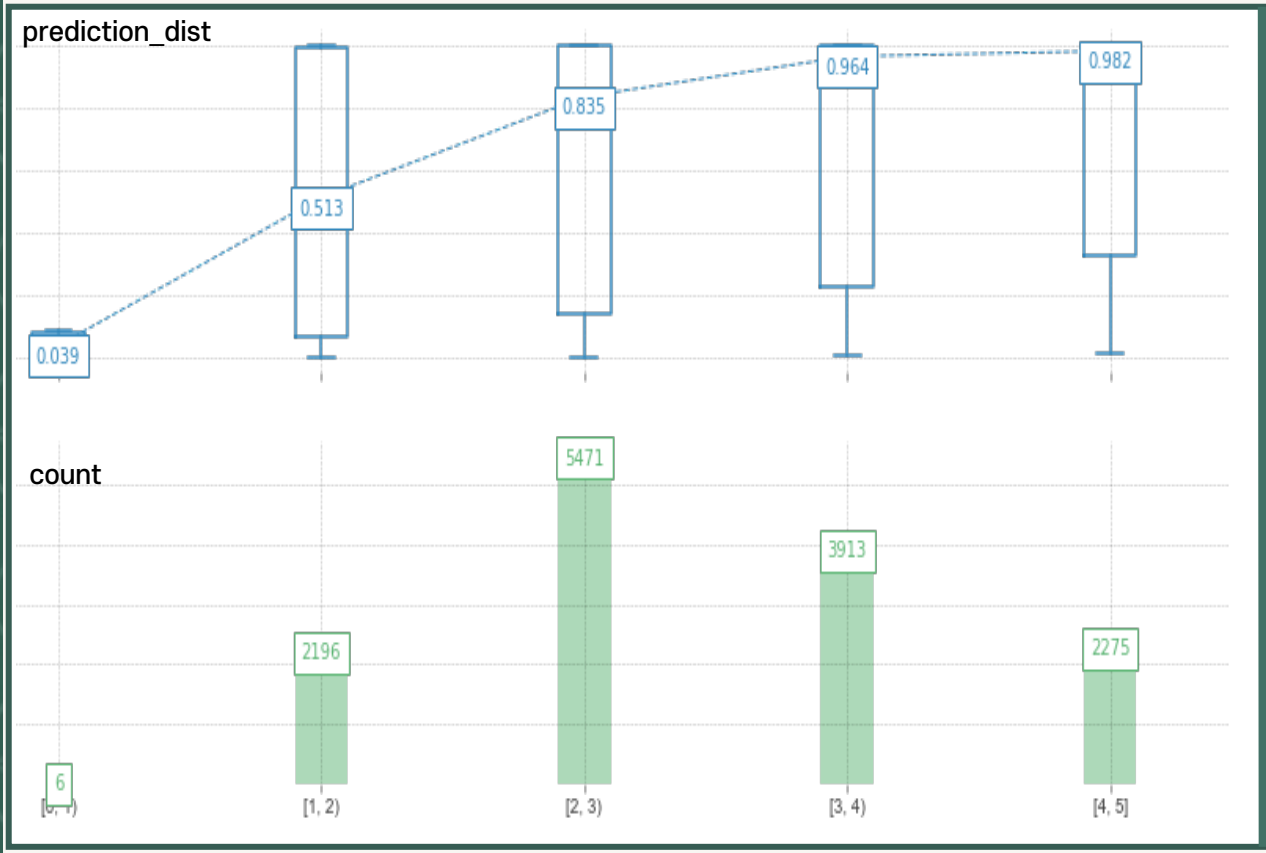


일반회원

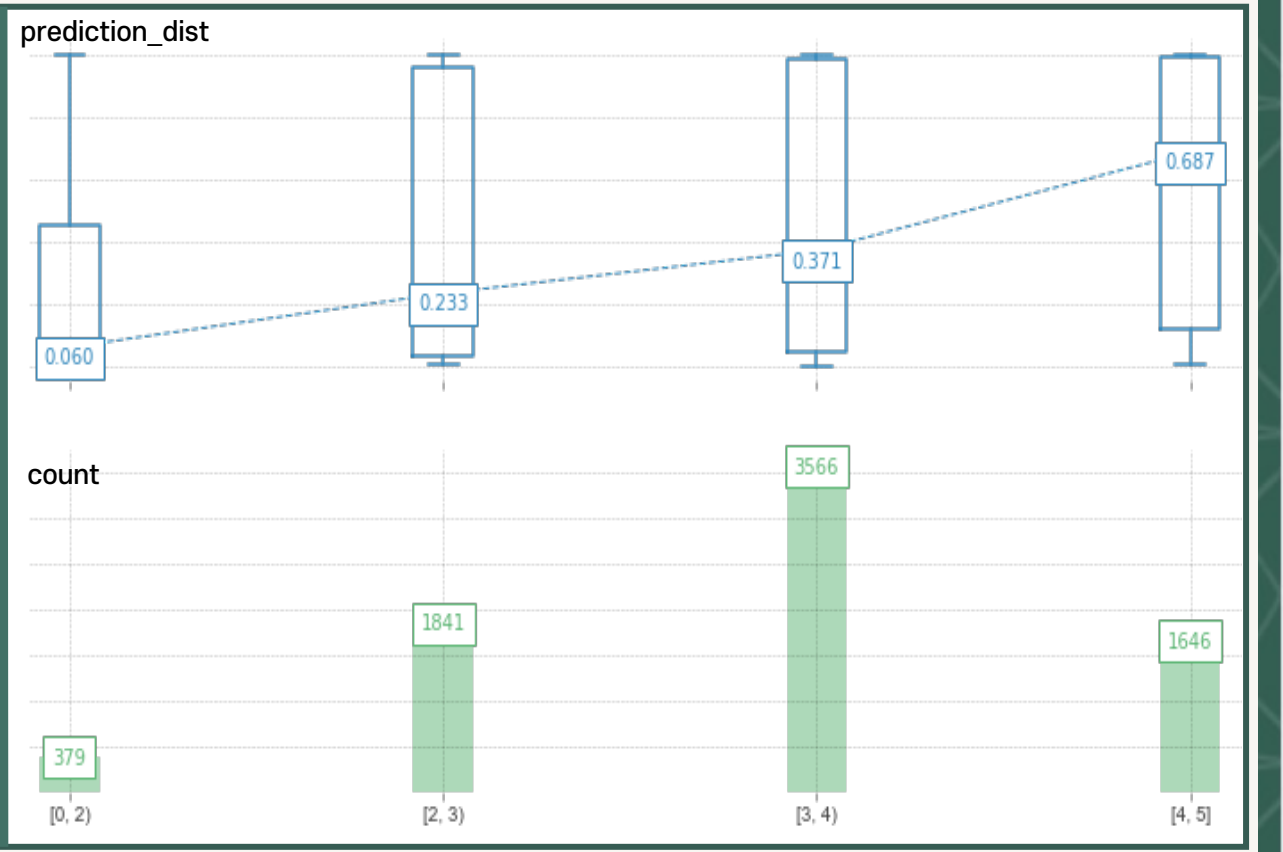
모델 해석 구간별 프라임 예측 확률

Actual predictions plot for *age_grp*

Distribution of actual prediction through different feature values



임직원

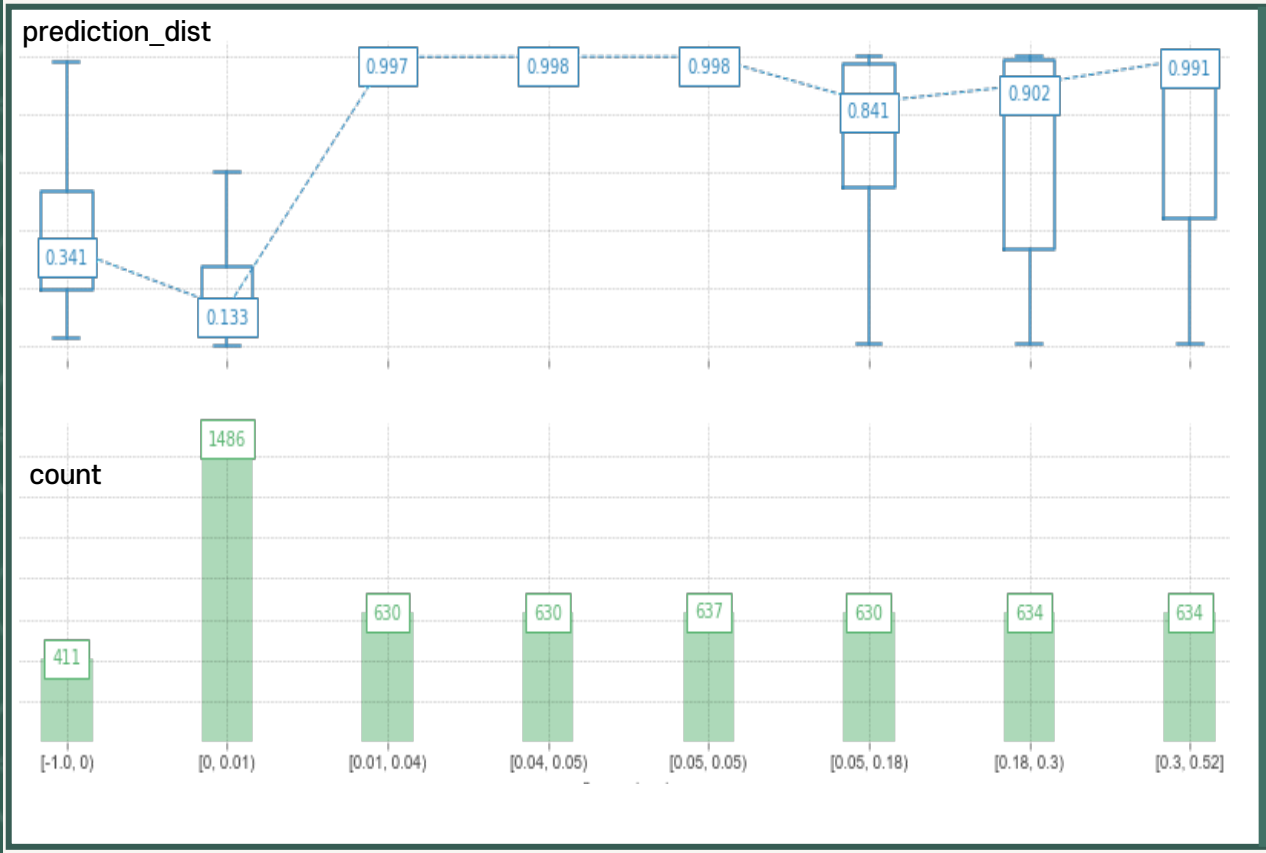


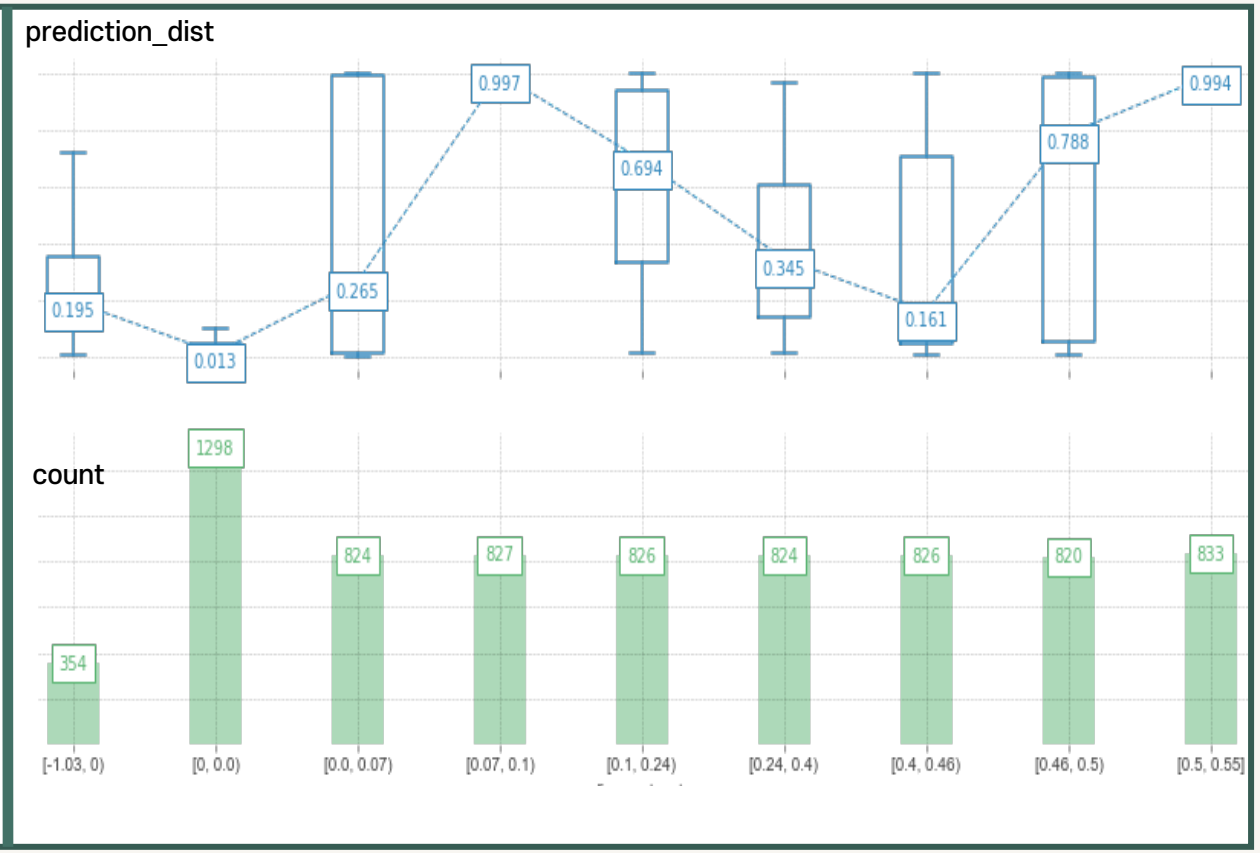
일반회원

모델 해석 구간별 프라임 예측 확률

Actual predictions plot for *discount_rate*

Distribution of actual prediction through different feature values

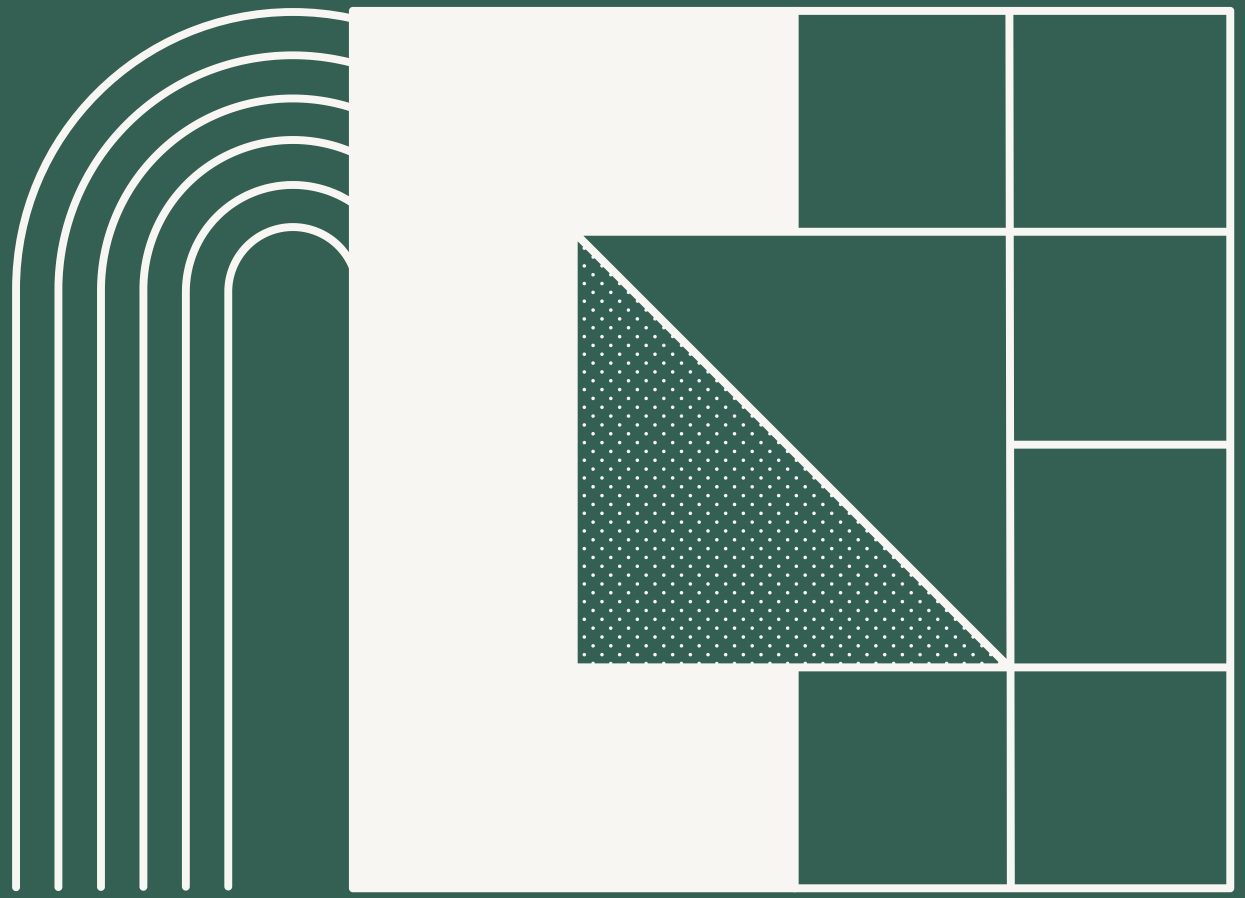




임직원

일반회원

모델 활용



모델 활용

2023 BDA 데이터 분석·활용 공모전

감사합니다

제일세당