식자재 쇼핑몰 재주문 예측

목차

데이터 설명

데이터 셋은 캐글에서 구했으며 Hunter's 슈퍼마켓(가상)의 약 200만 건의 온라인 구매기록으로 구성되어 있습니다.

200만건 이상의 데이터를 분석하기에는 시간과 장비의 한계가 있어서, 무작위 샘플링을 통해 201950개의 데이터를 사용했습니다.

kaggle

	order_id	user_id	order_number	order_dow	order_hour_of_day	days_since_prior_order	product_id	add_to_cart_order	reordered	department_id	department	product_name
0	2425083	49125	1	2	18	NaN	17	1	0	13	pantry	baking ingredients
1	2425083	49125	1	2	18	NaN	91	2	0	16	dairy eggs	soy lactosefree
2	2425083	49125	1	2	18	NaN	36	3	0	16	dairy eggs	butter
3	2425083	49125	1	2	18	NaN	83	4	0	4	produce	fresh vegetables
4	2425083	49125	1	2	18	NaN	83	5	0	4	produce	fresh vegetables
...	...	...	...	...	...	...	...	...	...	...	...	...
2019496	3390742	199430	16	3	18	5.0	83	8	0	4	produce	fresh vegetables
2019497	458285	128787	42	2	19	3.0	115	1	1	7	beverages	water seltzer sparkling water
2019498	458285	128787	42	2	19	3.0	32	2	1	4	produce	packaged produce
2019499	458285	128787	42	2	19	3.0	32	3	1	4	produce	packaged produce
2019500	458285	128787	42	2	19	3.0	123	4	1	4	produce	packaged vegetables fruits

2019501 rows × 12 columns

```
[ ]  df=df.sample(frac=0.1,random_state=2)  #2019501개의 데이터를 사용하고 싶었지만 시간관계상 201950개만 사용하려고 한다.
```

데이터 설명

데이터 셋에서 사용되는 특성은 다음과 같습니다.

order_id – (주문을 식별하는 고유 번호)
user_id - (사용자를 식별하기 위한 고유 번호)
order_number – (주문 번호)
order_dow – (주문 요일)
order_hour_of_day – (주문 시간)
days_since_prior_order - (주문 내역)
product_id – (제품의 ID)
add_to_cart_order – (장바구니에 추가된 항목 수)
reordered – (재주문의 발생 여부)
department_id - (각 부서에 할당된 고유 번호)
department – (부서 이름)
product_name – (제품 이름)

#	Column	Non-Null Count	Dtype
0	order_id	201950 non-null	int64
1	user_id	201950 non-null	int64
2	order_number	201950 non-null	int64
3	order_dow	201950 non-null	int64
4	order_hour_of_day	201950 non-null	int64
5	days_since_prior_order	189383 non-null	float64
6	product_id	201950 non-null	int64
7	add_to_cart_order	201950 non-null	int64
8	reordered	201950 non-null	int64
9	department_id	201950 non-null	int64
10	department	201950 non-null	object
11	product_name	201950 non-null	object

데이터 설명

결측치: days_since_prior_order 컬럼에서 12,567개의 결측치가 있는 것을 확인.

- 이후 모델링 과정에서 대치 예정

중복값: 0

```
df.isna().sum()
```

```
order_id                    0
user_id                     0
order_number                0
order_dow                   0
order_hour_of_day           0
days_since_prior_order    12567
product_id                  0
add_to_cart_order           0
reordered                   0
department_id               0
department                  0
product_name                0
dtype: int64
```

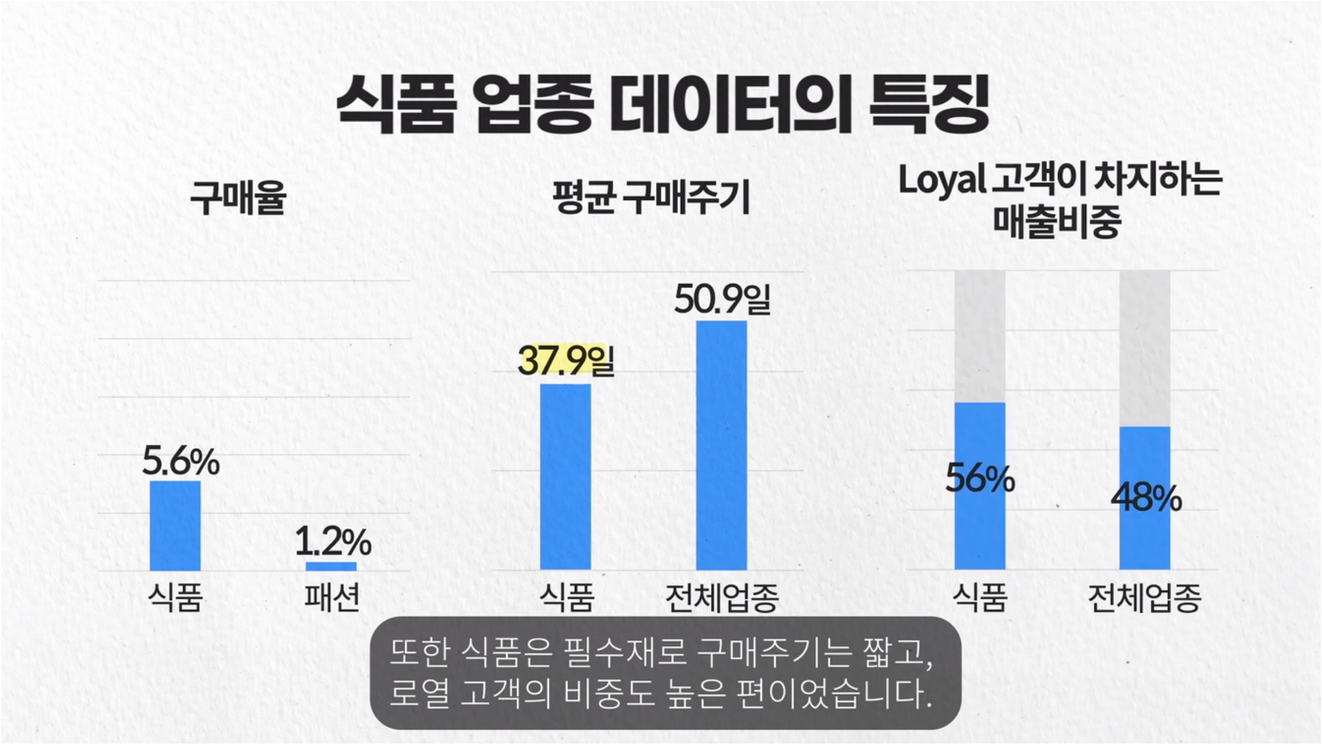
문제 정의

식품 쇼핑몰에서는 충성고객의 매출이 상당히 큰 영향을 미친다. 따라서 충성고객 유치를 위해 재주문을 예측한 뒤 재구매를 하지 않을 것으로 예측되는 고객에게 문자를 통한 쿠폰지급을 하려 한다.

지급 시기는 주문량이 많은 요일과 시간대로 정한다.

가설:

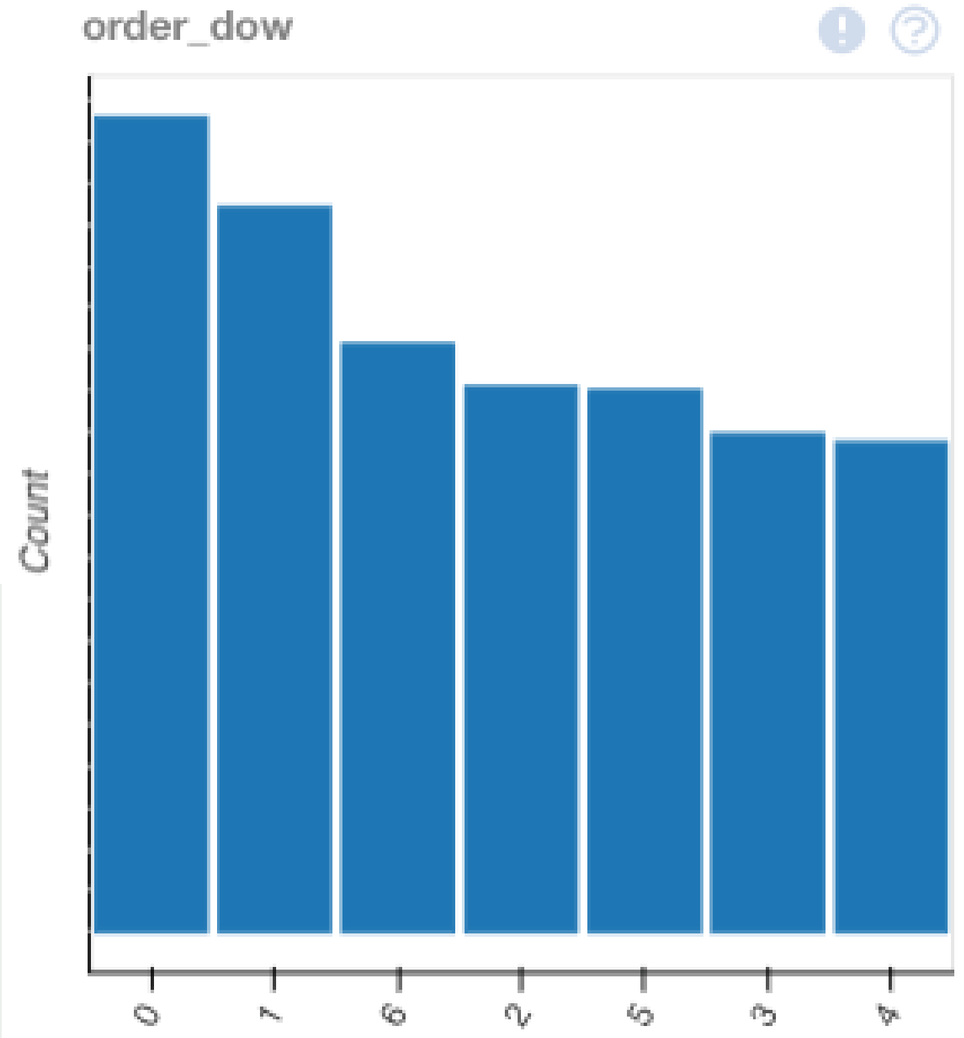
1. 장바구니에 넣은 물건 수의 양이 재주문 여부와 관계가 있을 것이다.
2. 구매한 제품의 카테고리에 따라 재주문 여부와 관계가 있을 것이다.



출처: Dighty

EDA

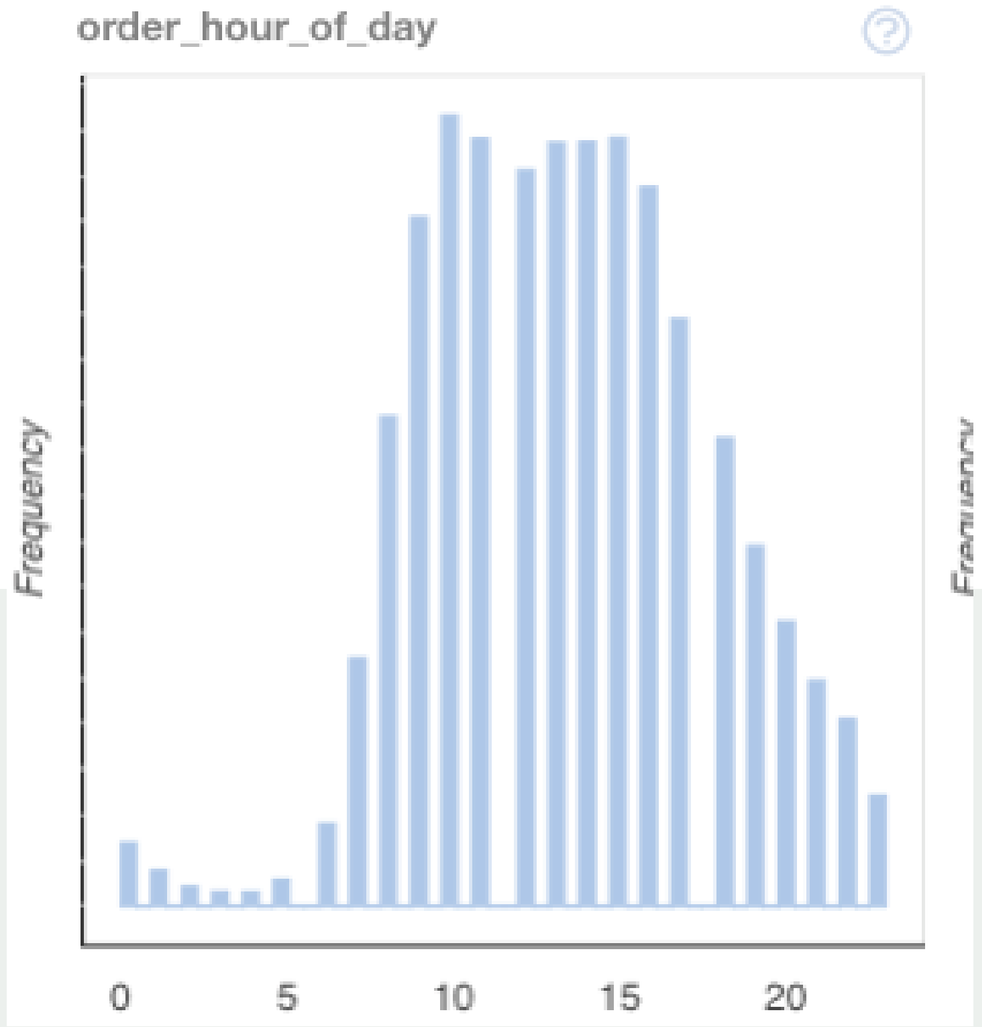
그래프는 요일 별 주문량을 나타내고 있다.



- 다음 그래프를 통해 월요일(0)에 가장 주문이 많다는 것을 알 수 있다.

EDA

그래프는 시간별 주문량을 나타내고 있다.



- 다음 그래프를 통해 오전 10시부터 오후 5시 까지 주문이 많다는 것을 알 수 있다.
- 오전 9시 지급

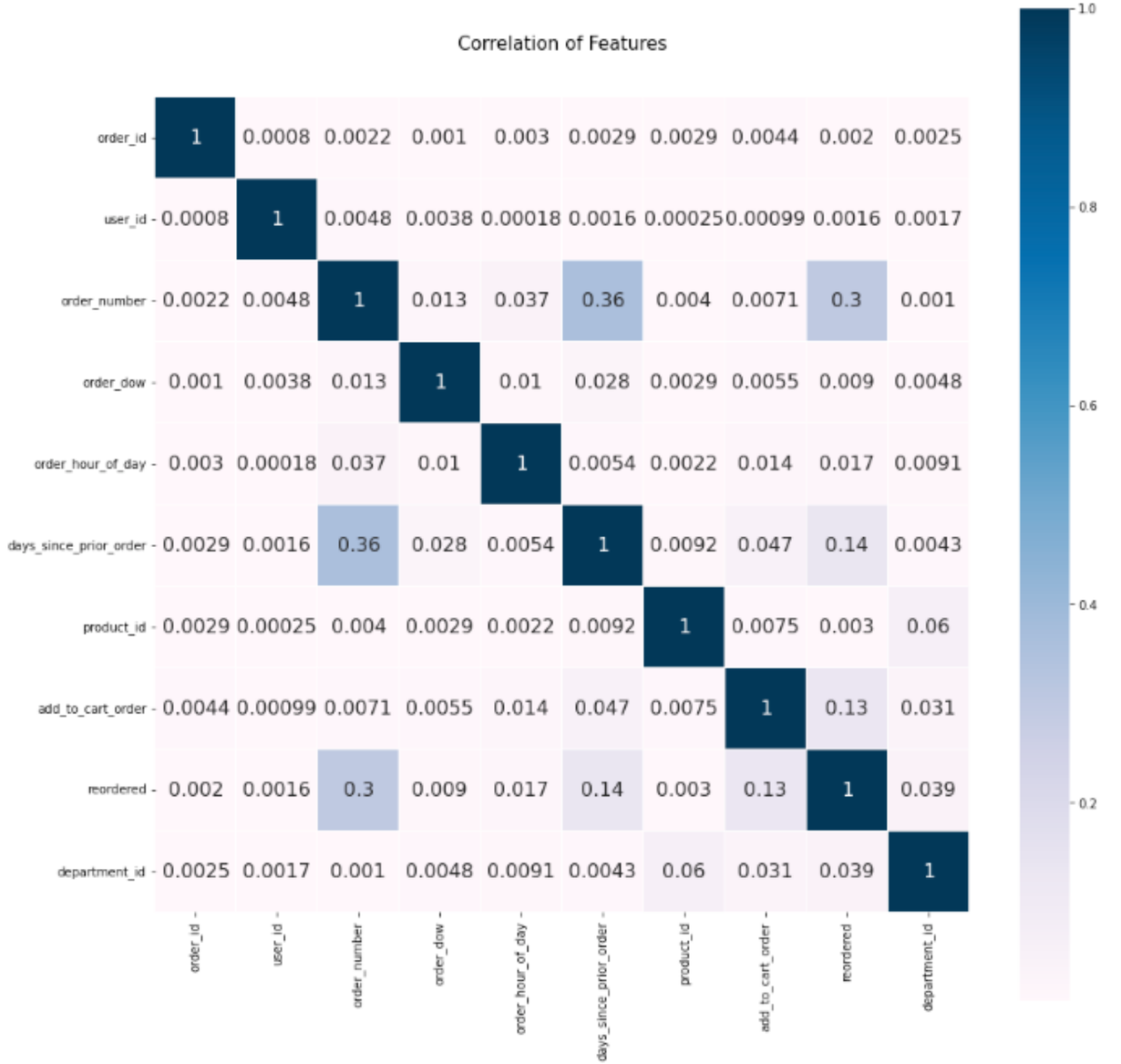
EDA

상관관계 히트맵

데이터셋의 상관계수를 히트맵으로 나타낸 결과 reordered와 상관계수가 높은 특성은 'order_number', 'days_since_prior_order', 'add_to_cart_order'로 추려볼 수 있다.

가설

"장바구니에 넣은 물건 수의 양이 재주문 여부와 관계가 있을 것이다".
reordered와 add_to_cart_order의 상관계수를 보아 장바구니에 넣은 물건 수와 재주문 여부는 관계가 있다는 것을 알 수 있다.



출처: 여기에 참고 문헌을 추가하세요.

EDA

카테고리별 평균 재주문율

다음 표는 쇼핑몰의 상품을 대분류한 카테고리별로 평균 재주문율을 보여준다.

department	reordered
dairy eggs	0.669549
beverages	0.656226
produce	0.648215
bakery	0.630761
deli	0.607944
pets	0.600726
alcohol	0.587459
meat seafood	0.581369
babies	0.573090
snacks	0.572303
breakfast	0.562613
bulk	0.556054
frozen	0.536709
other	0.468619
dry goods pasta	0.467962
canned goods	0.456889
household	0.405493
missing	0.402105
international	0.367262
pantry	0.341758
personal care	0.317976

가설

"구매한 제품의 카테고리에 따라 재주문 여부와 관계가 있을 것이다."

dairy eggs, beverages를 산 고객의 재주문율은 pantry,personal care를 산 고객의 재주문율의 두 배가량 차이가 나는 것을 확인할 수 있다.
따라서 유제품,계란,음료 등 금방 소비되는 신선식품 제품군을 구매하는 고객은 재주문을 할 확률이 높다고 할 수 있다.

모델 성능 비교


```
df['reordered'].value_counts(normalize=Tru
```

```
1    0.58927
0    0.41073
Name: reordered, dtype: float64
```

Baseline

```
학습 0.7412168853676653
검증 0.7332508046546174
```

	precision	recall	f1-score	support
0	0.72	0.56	0.63	16457
1	0.74	0.85	0.79	23933
accuracy			0.73	40390
macro avg	0.73	0.71	0.71	40390
weighted avg	0.73	0.73	0.73	40390

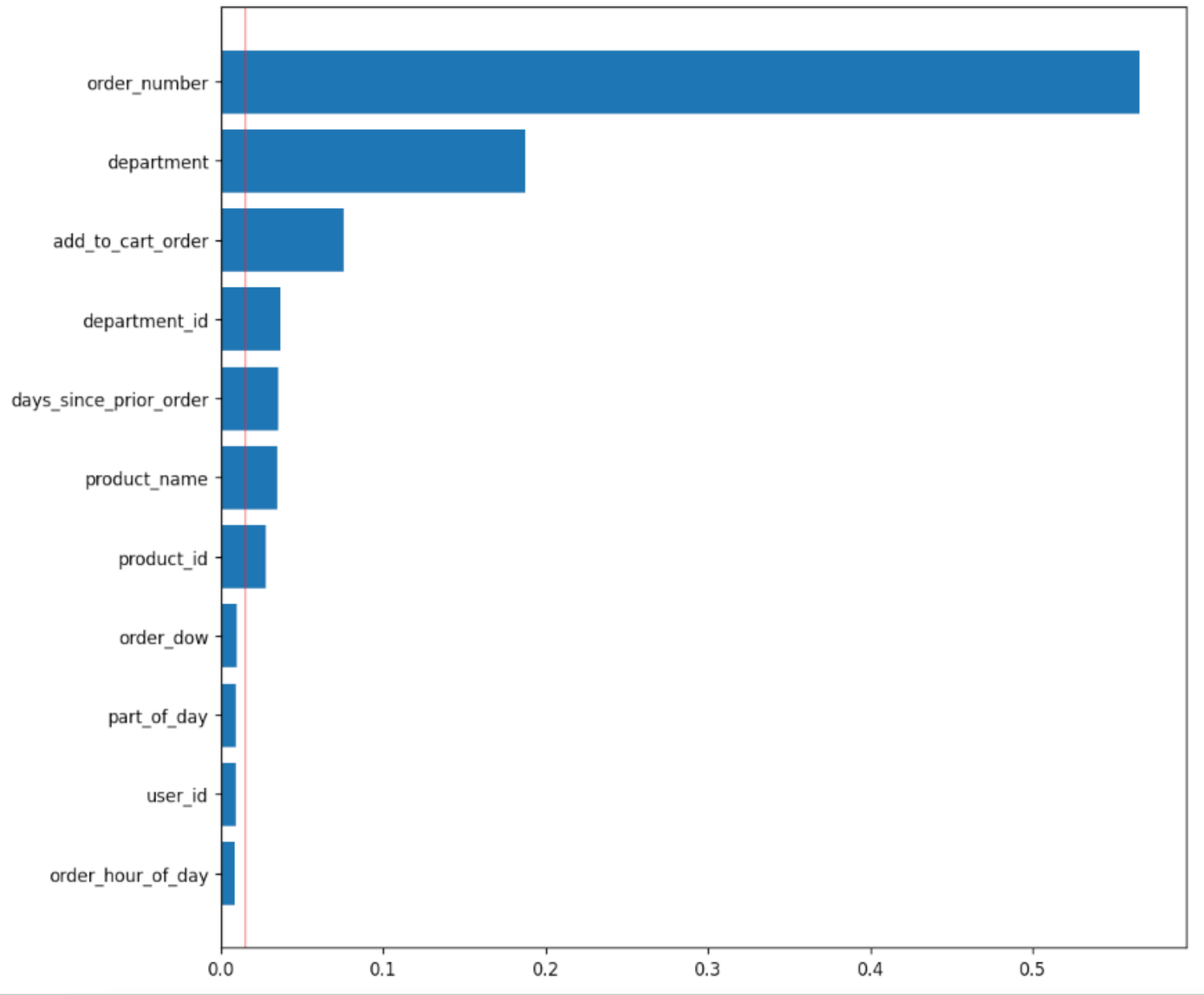
Xgboost

```
학습 0.6831455805892548
검증 0.686803664273335
```

	precision	recall	f1-score	support
0	0.85	0.28	0.42	16457
1	0.66	0.97	0.79	23933
accuracy			0.69	40390
macro avg	0.75	0.62	0.60	40390
weighted avg	0.74	0.69	0.64	40390

Randomforest

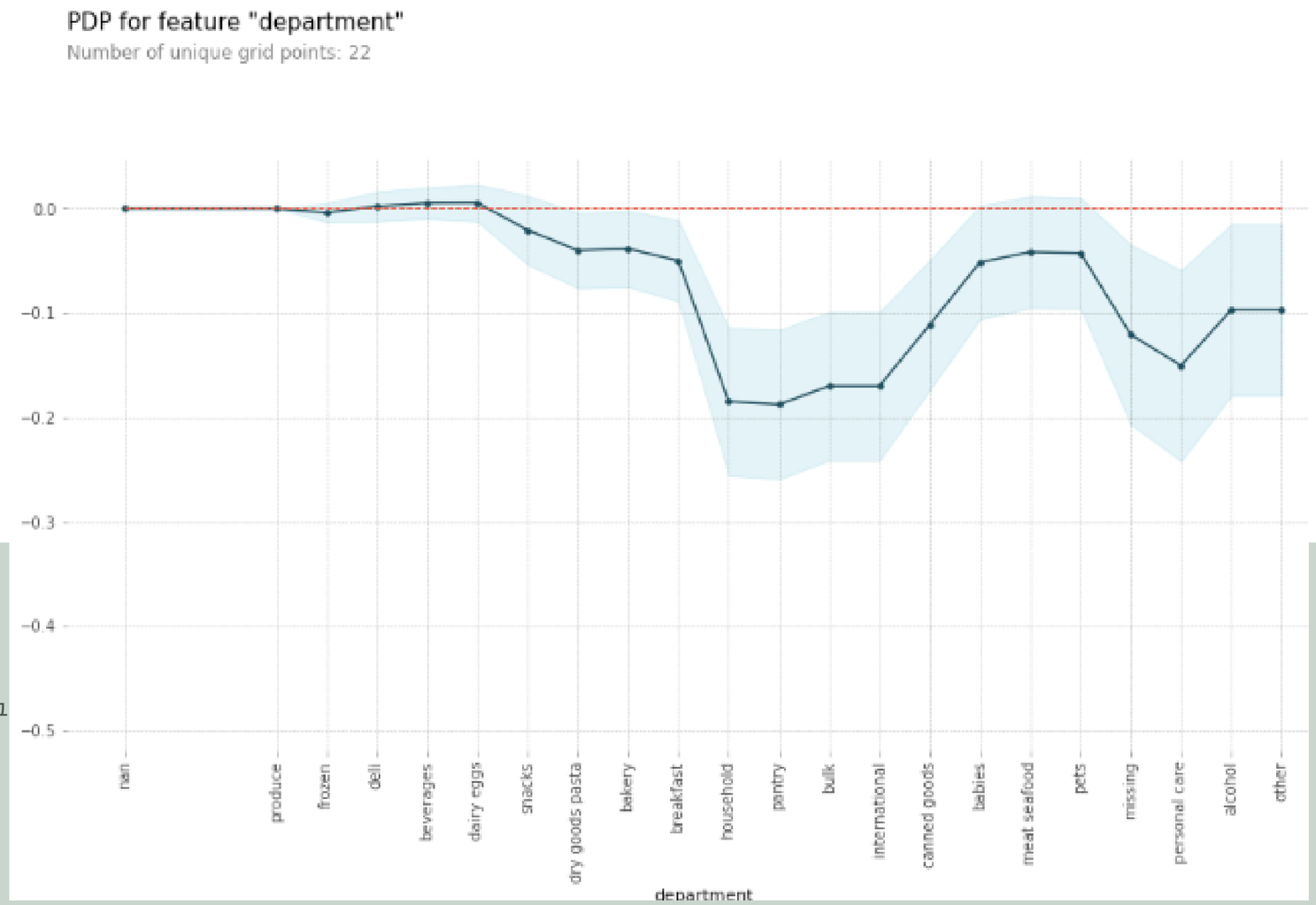
Feature Importance



하위 4개 특성을 drop한다.

order_number 특성을 사용하고 싶지만, 어떤 정보를 갖고있는지 알아내지 못했다.

PDP plot



household, pantry, bulk, international, canned goods, personal care 의 제품군들은 produce, frozen, deli, beverages, dairy eggs 제품군들보다 재구매를 할 확률이 15%에서 20% 차이난다고 할 수 있다.

인사이트 도출

쿠폰지급 대상

- household, pantry, bulk, international, canned goods, personal care 등 신선식품이 아닌 제품을 구매하는 고객

쿠폰지급 시기

- 월요일
- 오전 9시