




장바구니 재구매 상품 추천 알고리즘 개발

안경은

목차

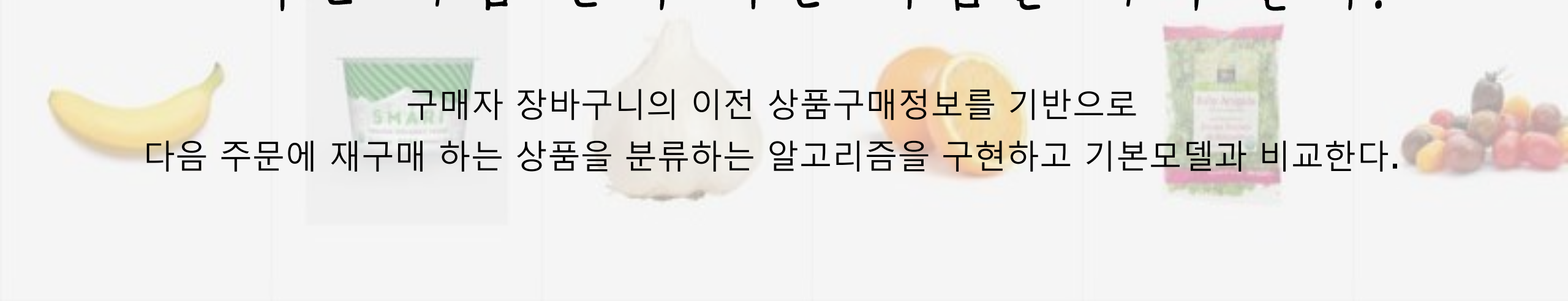


- 프로젝트 목표
- 문제 해결의 주요 아이디어
- 가정1 기초통계 및 기초모델 설명
- 가정2,3 비교모델 설명
- F1 score 결과값 비교



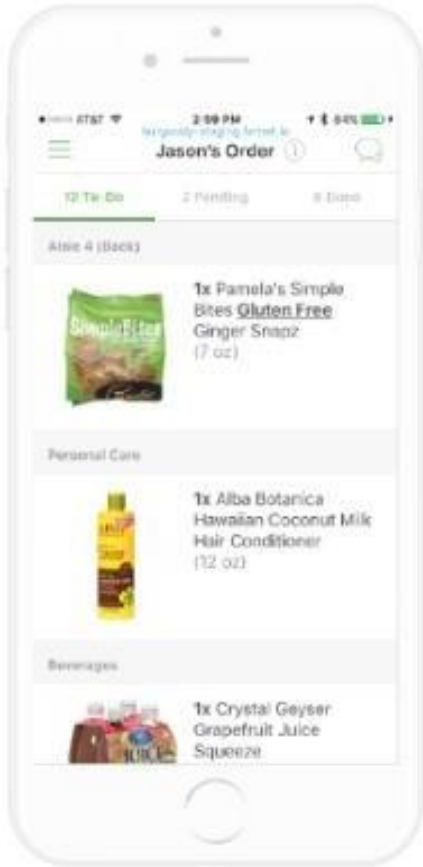
프로젝트 목표

“고객은 다음 번에 어떤 제품을 다시 살까?”



구매자 장바구니의 이전 상품구매정보를 기반으로
다음 주문에 재구매 하는 상품을 분류하는 알고리즘을 구현하고 기본모델과 비교한다.

식품배송서비스 instacart



100s of retailers



1,000s of stores



10,000s of shoppers



1,000,000s of products



총 10^{15} s의
선택지

고객들은 엄청난 수의 선택의 기로에 놓인다

상품 추천의 의의

125 million US households  5% market share  1x trips / week  1 minute saved per trip  618 years of shopping

상품추천으로 고객의 1분의 시간을 줄일 때 마다
총618년의 쇼핑시간이 절약된다(미국시장 기준)

목차



- 프로젝트 목표
- 문제 해결의 주요 아이디어
- 가정1 기초통계 및 기초모델 설명
- 가정2,3 비교모델 설명
- F1 score 결과값 비교



주요 아이디어(기초 가정)

1. 데이터의 기초통계를 통한 insight에서 나온, 재구매 여부와 상관도가 높은 변수 생성 (날것의 데이터에서 시각화해서 확인하기 쉽다.)

- 빨리 상하거나, 보관 기간이 짧은 식품류는 재구매율이 높을 것 이다
- 특정시간대 구매자들과 재구매율은 상관관계가 높을 것이다
- 특정 기간의 구매자들과 재구매율은 상관관계가 높을 것이다 등등

2. 구매횟수와 재구매 여부는 상관관계가 높을 것이다

- 구매횟수가 100위 안에 드는 물건들이 재구매 여부와 상관관계가 높을 것이다.

3. 고객은 장바구니의 목록 중 구입 빈도가 높은 상품을 재구매 할 것이다.



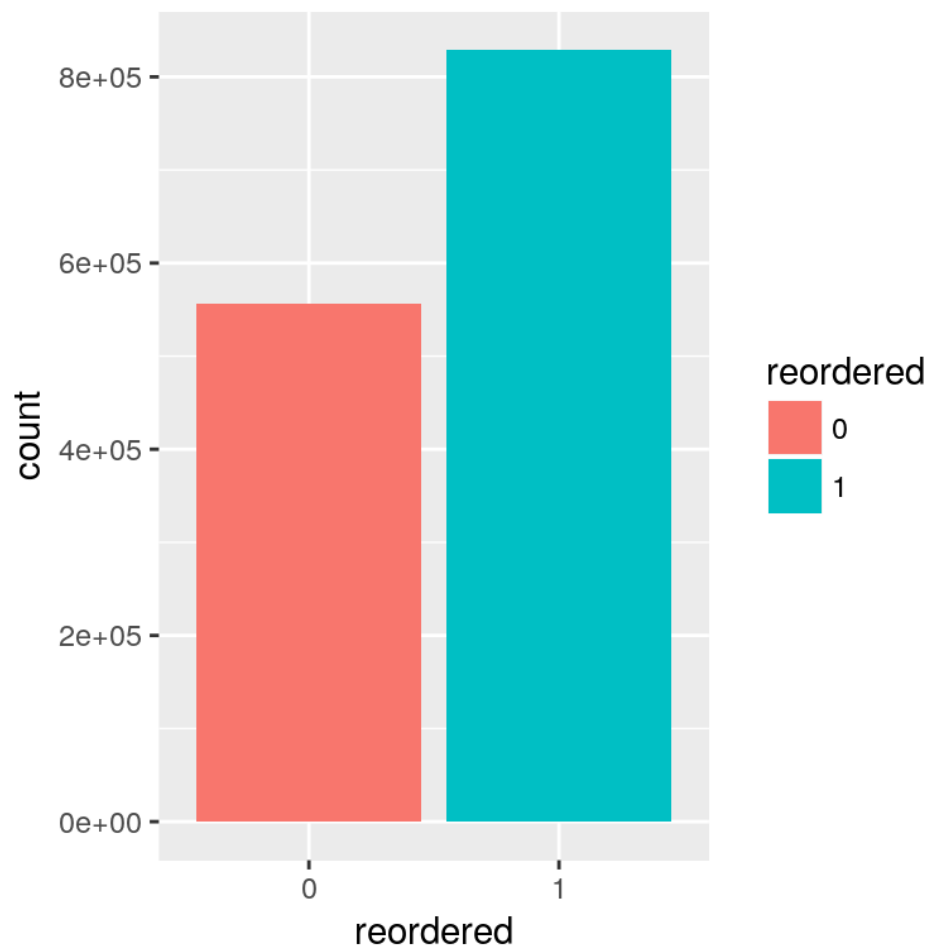
목차



- 프로젝트 목표
- 문제 해결의 주요 아이디어
- 가정1 기초통계 및 기초모델 설명
- 가정2,3 비교모델 설명
- F1 score 결과값 비교



재구매 비율 - 추천을 할만한 가치가 있을까?



재구매 여부	count	비율
0	555793	0.4014056
1	828824	0.5985944

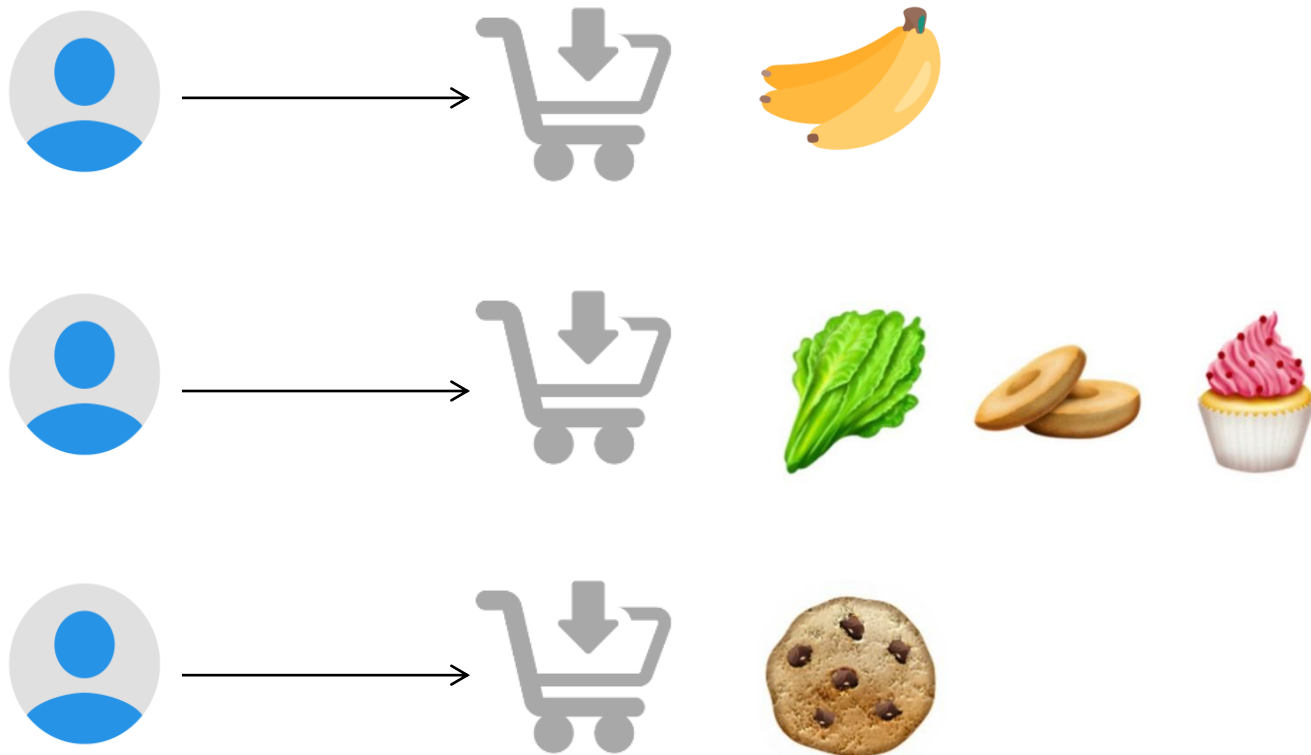
재구매 비율 약 60%로
과반수가 과거의 구매제품을
재구매한다.



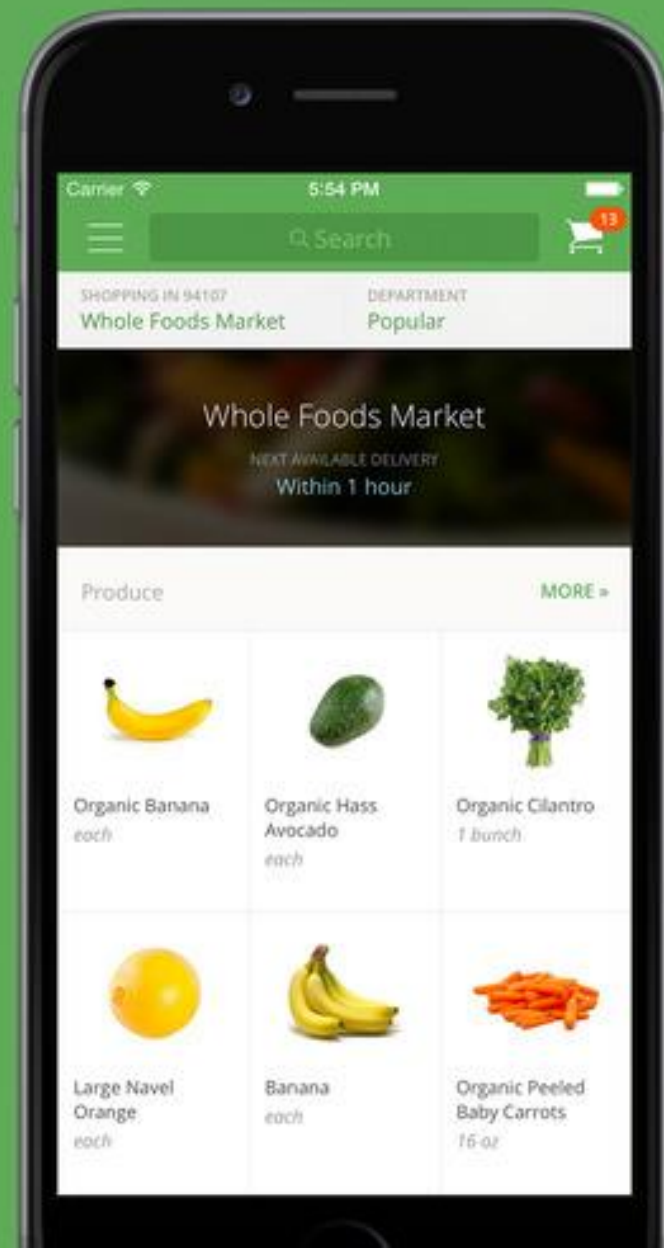


데이터

Instacart(식품배송 서비스)
고객들의 식품 구매정보



Delivery from Local Grocery Stores
Sign up to shop in your area!





데이터



고객(구매자) 20만명	고객이 구매한 상품 데이터 5만개	고객의 총 주문수 320만개
주문 ID	주문 ID	주문 ID
고객 ID	고객 ID	장바구니에 넣은 순서
주문번호 - 주문횟수 추정가능	상품 ID	주문번호 - 주문횟수 추정가능
주문 요일	상품 소분류 카테고리	주문 요일
주문 시간	상품 대분류 카테고리	주문 시간
주문일로부터 지난 날	상품 이름	이전 상품 재구매 여부





데이터 기초통계

- TOP 10 상품

상품 ID	순위	상품명
24852	1	바나나
13176	2	유기농 바나나
21137	3	유기농 딸기
21903	4	유기농 새싹 시금치
47626	5	레몬
47766	6	유기농 아보카도
47209	7	유기농 해스 아보카도
16797	8	딸기
26209	9	라임
27966	10	유기농 라즈베리

- TOP 10 재구매 상품

상품 ID	순위	상품명
1729	1	유당제거 우유
20940	2	유기농 저지방 우유
12193	3	100% 플로리다 오렌지 주스
21038	4	유기농 스펠트 토르띠야
31764	5	유기농 스파클링 생수캔
24852	6	바나나
117	7	과일 요거트
39180	8	유기농 저지방 1% 우유
12384	9	유기농 유당제거 1% 우유
24024	10	저지방 우유

- 제일먼저 장바구니에 담는 제품

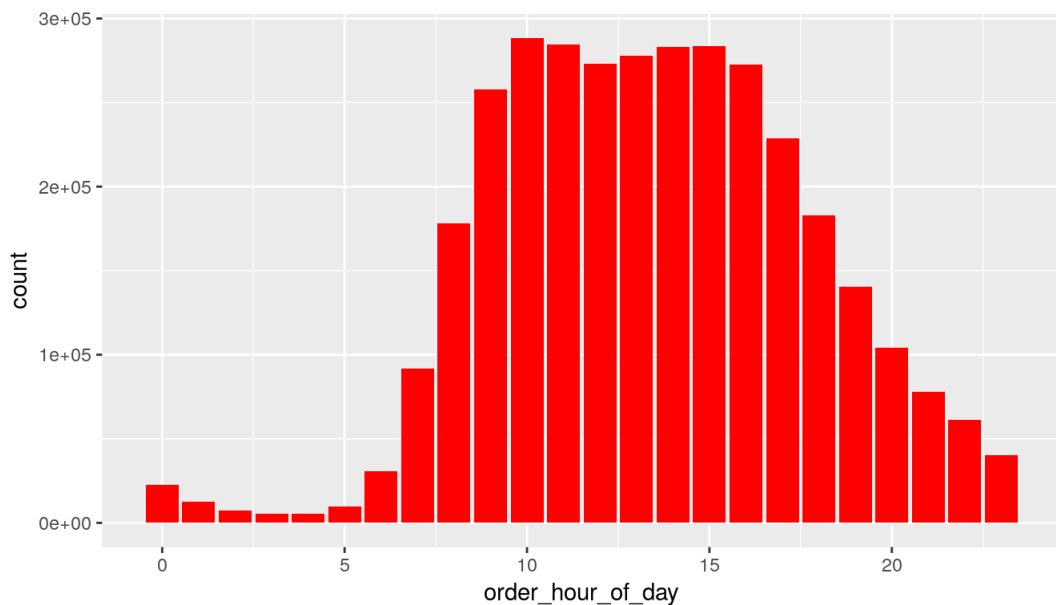
상품 ID	순위	상품명
45004	1	흰색 다층 휴지
11885	2	스파클링 워터
13128	3	아카킨 미네랄 워터
4100	4	유기농 에스프레소 커피콩
1729	5	2% 유당 제거 우유
6729	6	쿠키 쟁반
9285	7	뼈없는 돼지고기
6848	8	파티용 텀블러
12640	9	Natural Spring Water
26405	10	두루마리 휴지





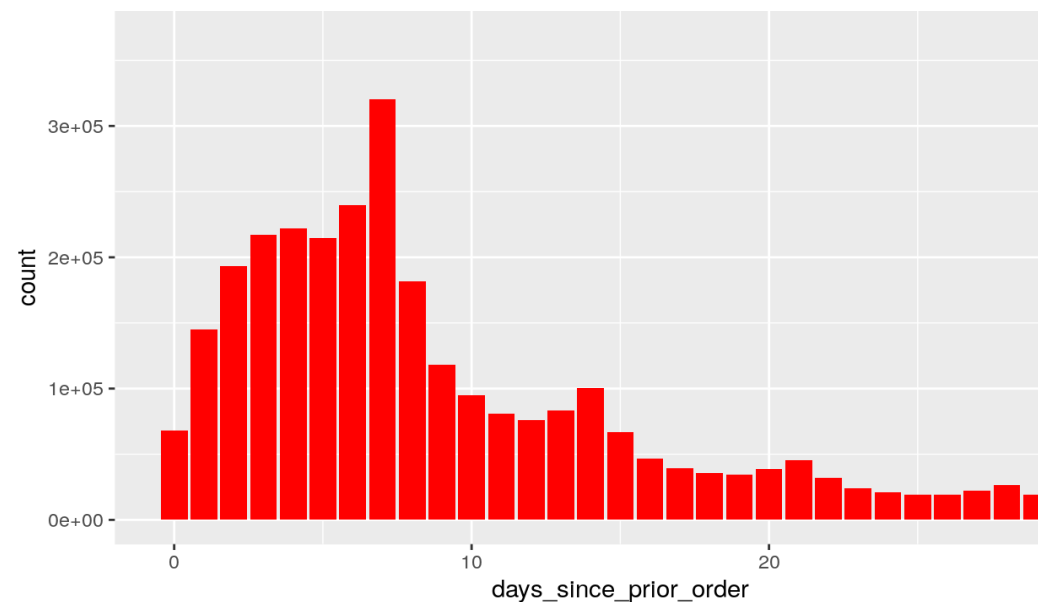
데이터 기초통계

- 하루에 주문율이 가장 높은 시간대



8:00-18:00까지의 주문율이 가장 높다

- 고객은 마지막 주문 이후 언제 다시 구매 하는가?

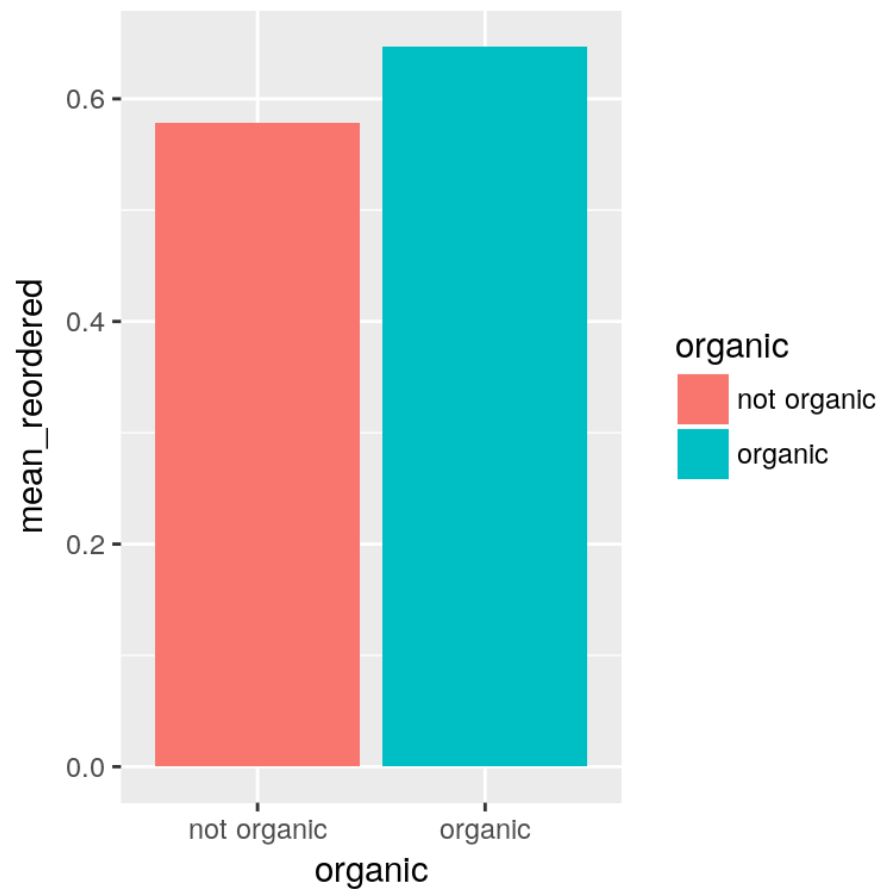


1주일 간격으로 식료품을 구매하는 고객이 가장 많다





데이터 기초통계 - 유기농 제품 재구매 비율



유기농 제품 재구매율은 약
64%

유기농 제품 구입자

구매율

not organic

0.5784985

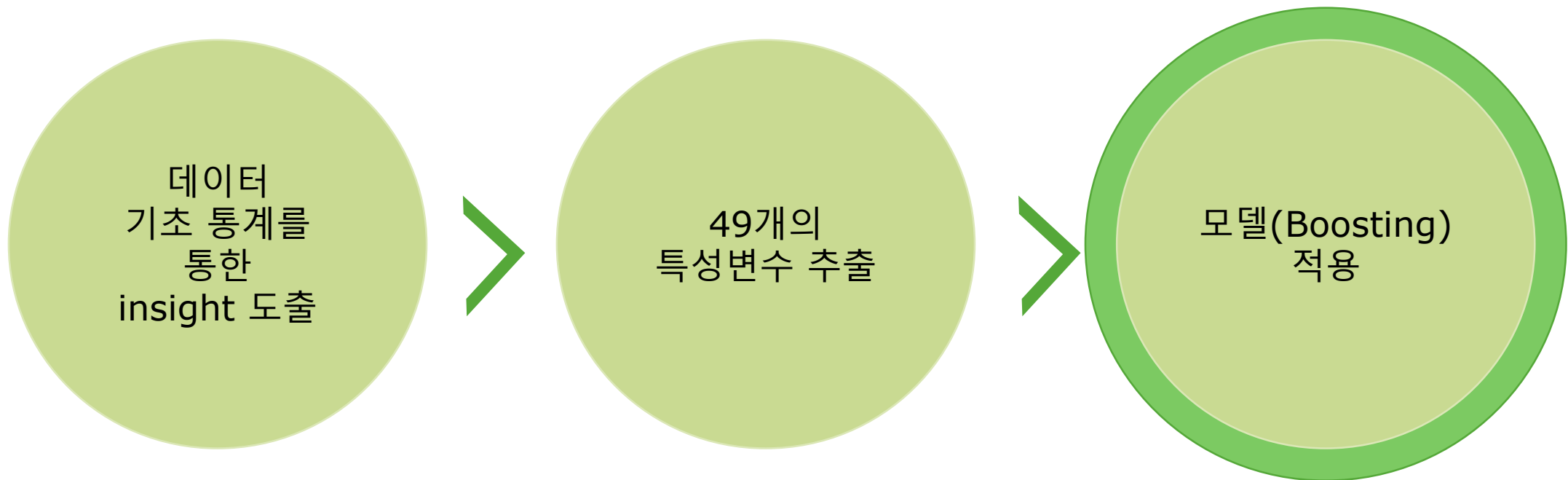
organic

0.6470981





기본 모델 분석 과정



기본 모델 설명

- 기본 모델은 가정1을 기반으로 한 모델이다.

1. 기초 통계를 통해 얻은 insight로 재구매 여부와 연관도가 높은 변수들만 뽑고 정제하여 특성변수로 만들었다.(49개)
2. 이 변수들이 정말 재구매 여부와 연관도가 높은지 정확도를 검증한다.(boosting의 역할)



특성변수 생성

기초통계를 바탕으로 총 49개의 특성변수 생성

1	user_orders	고객의 총 주문수	
2	user_period	고객의 첫 주문부터 마지막 주문이 일어난 때 까지의 기간	
3	user_mean_days_since_prior	주문 기간 평균	
4	user_total_products	주문한 총 상품수	
5	user_reorder_ratio	$\text{sum}(\text{reordered}=1)/\text{sum}(\text{order_number}>1)$	
6	user_distinct_products	$\text{n_distinct}(\text{product_id})$	
7	user_average_basket	$\text{user_total_products}/\text{user_orders}$	
8	user_last2	최근 2번째 주문 후 지난 일수	
9	user_last3	최근 3번째 주문 후 지난 일수	
10	user_interval_mean_last3	최근 1,2,3번째 주문 후 지난 일수 평균	
11	user_mean_order_dow	고객의 주문 요일 평균	
12	user_mean_order_hour_of_day	고객의 주문 시간 평균	



모델 적용 (Boosting)



목차



- 프로젝트 목표
- 문제 해결의 주요 아이디어
- 가정1 기초통계 및 기초모델 설명
- 가정2,3 비교모델 설명
- F1 score 결과값 비교



비교모델1 분석 - 구매횟수 주요변수 모델

구매횟수
상위 1000
개 상품 추출



상품에 대한
고객별 구매
횟수와
product_id
변수화



모델
(Boosting)
적용



상위 1000개
품목
F1 score
예측 값 구하
기



비교 모델1 설명

- 비교 모델1은 가정2를 기반으로 한 모델이다.

1. 구매율이 높은 상위 100개 상품에 대한 구매 횟수를 주요한 (재구매 여부와 상관도가 높은) 변수로 보고 있기 때문에 횟수에 의미를 담는 모델링을 한다.(one-hot-encoding)
2. 구매 횟수가 정말 재구매 여부와 연관도가 높은지 정확도를 검증한다 (boosting의 역할)



비교모델2 분석 - LDA모델 사용

LDA 모델을
통한 변수
추출



기본모형의
특성변수+
LDA 모델을
통한 변수



모델
(Boosting)
적용



상위 1000개
품목
F1 score
예측 값
구하기



비교 모델2 설명

- 비교 모델2는 가정3를 기반으로 한 모델이다.

1. 고객 i 의 j 번째 장바구니의 상품 목록을 하나의 문서로 간주하고
상품을 단어로 하여 LDA 적용
2. 유저들의 상품목록에서 뽑아낸 최다빈도수 상품(토픽)이 실제 재구매
상품과 얼마나 합치하는지를 검증한다(boosting의 역할)

목차



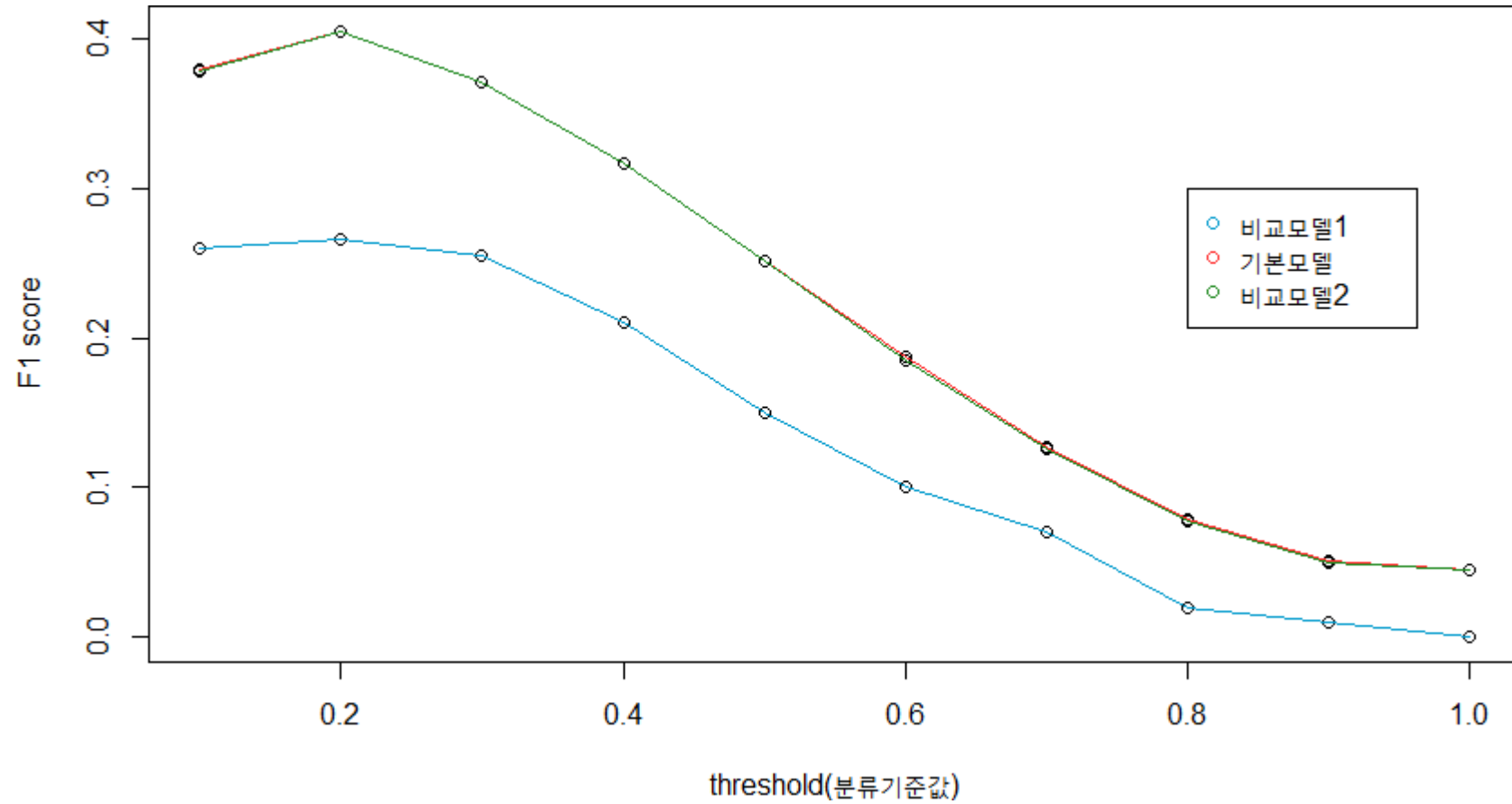
- 프로젝트 목표
- 문제 해결의 주요 아이디어
- 가정1 기초통계 및 기초모델 설명
- 가정2,3 비교모델 설명
- F1 score 결과값 비교

F1 SCORE에 대한 설명

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- Recall : 재현율
 - 찾아야 할 것 중에 실제로 찾은 비율은?
 - 컴퓨터가 True라 한 것 중에 실제 True의 비율
- Precision : 정밀도
 - 예측한 것 중에 정답의 비율은?
 - 실제 True중 컴퓨터가 True라 한 것의 비율
- F1 Score : recall과 precision의 조화평균

F1 최적값을 얻기 위한 분류 기준값 그래프



Threshold에 대한 설명

- Threshold는 재구매 여부(0 혹은 1)를 가르는 기준이다.
- 각 가정에 대한 정확도가 재구매로 연결 될 것이다 혹은 아니다로 판별하는 기준이 필요하기 때문에 threshold라는 분류 기준값을 두는 것이다.
- 위의 그래프는 이 분류 기준값에 따른 f1(결과값)을 나타낸 것이다.
 - F1값이 가장 높게 나오는 분류 기준값을 사용한다.

상위 1000개 상품에 대한 모델별 validation 값 비교

기본모델 - 최적의 F1 SCORE(49개 변수)

	f1	preci	recall	n.recom	n.real	n.collect
train	0.3953946	0.3712652	0.5016135	8.988684	6.374700	3.527749
valid	<u>0.4045941</u>	0.3786994	0.5205823	6.157477	4.171669	2.411833

비교모델1 - 최적의 F1 SCORE(구매횟수 변수화 모델)

	f1	preci	recall	n.recom	n.real	n.collect
valid	<u>0.26646464</u>	0.17385137	1.948614	30.22617	4.171669	4.0140160

비교모델2 최적의 F1 SCORE(LDA 잠재변수)

	f1	preci	recall	n.recom	n.real	n.collect
train	0.3958211	0.3708507	0.5013715	8.986504	6.374700	3.538044
valid	<u>0.4054196</u>	0.3794870	0.5217257	6.157609	4.171669	2.414694

결과값

- 비교모델2의 f1 결과값이 가장 높게 나왔다.
- 비교모델2의 알고리즘이 재구매 여부 판단에 가장 높은 정확성을 보인다고 볼 수 있다.