Instacart 장바구니 자료 분석

서울대학교 통계학과

김보영

2017. 10. 21

1. Validation set seed

- 제출 결과

eval	고객수
Train	81,209
Validation	50,000
Test	75,000
합계	206,209

훈련	시험	train F1	Val F1	Test F1
Train+Val	Test	0.40976	NA	0.40155
Train	Test	0.41464	0.398972	0.40097

- 모형 :
  None -> main(45개 + LDA 40개 입력변수)
  -> F1 optimization

- Seed 조정 (총 10번 시도)

훈련	시험	Train F1	Val F1
Train	Val	0.41355	0.40045

baseline

Important Findings for reorder - 1

- user_id: 54035

order_number	1	2	3	4	5	6	7	8	9	10	11	12
order_dow	6	6	0	1	0	6	6	0	6	1	6	6
order_hour_of_day	18	14	10	11	11	11	11	12	13	10	12	14
days_since_prior_order	-1	28	15	22	13	13	30	30	30	9	30	7
Cola	1	1	1	1	1	1	1	0	1	1	1	1
Reduced-Fat-Swiss-Deli-Thin-Slices-Cheese	0	0	0	0	0	0	0	1	0	0	0	1
Low-Fat-Strawberry-Banana-on-the-Bottom-Greek-Yogurt	0	0	1	0	1	0	1	1	1	0	1	0
24/7-Performance-Light-Weight-Cat-Litter	0	1	0	0	0	0	0	1	1	0	0	0
Fridge-Pack-Cola	0	0	0	0	0	0	0	1	0	0	0	0
Lowfat-1%-Milk	0	0	0	0	0	0	1	1	0	0	0	0
Original-Whole-Grain-English-Muffins	0	0	0	0	0	0	0	1	0	0	0	0
Strawberry-on-the-Bottom-Nonfat-Greek-Yogurt	1	1	1	0	1	0	1	1	0	0	0	0

- 고객이 한 상품을 여러 번 산다 할지라도, 사지 않는 시점이 있다.
- 이 고객은 8번째 주문에 Cola 대신 Fridge-pack-cola를 구매

2. Cola and Fridge pack cola

- Prior set

전체주문수	전체고객수
3,214,874	206,209

- Cola와 다른 상품들 비교

Item A	주문수	주문고객수
Cola	11,585	3,900

A와 B를 둘다 구매한 고객들 중에 A와 B를 동시에 구매한 적이 있는 고객 수

Item B	B주문수	주문교	주문합	주문교/합	B주문고객수	고객교*	고객합	고객교/합	*중 동시어/고객교	
Fridge Pack Cola	18,269	258	29,596	0.00872	5,202	590	8,512	0.06931	133	0.22542
Soda	35,791	362	47,014	0.00770	8,000	442	11,458	0.03858	174	0.39367
Ginger Ale	12,536	266	23,855	0.01115	6,002	368	9,534	0.03860	160	0.43478
Coke Classic	10,772	204	22,153	0.00921	3,709	423	7,186	0.05886	128	0.30260
Zero Calorie Cola	8,558	1	20,142	0.00005	1,605	5	5,500	0.00091	1	0.20000
Cola, Coke Life	117	16	11,686	0.00137	53	9	3,944	0.00228	5	0.55556

2. Cola and Fridge pack cola

- 따라서 U_A : A를 구매한 고객 집합 이라 할 때 상품 A에 대해

$$M_{A,B} = \frac{|U_A \cap U_B\text{의 원소이면서 동시에 } A\text{와 } B\text{를 주문한 적이 있는 고객의 집합}|}{|U_A \cap U_B|}$$

가 큰 상품 B와 작은 상품 C를 뽑는다.

이때 $|U_A| > 500, |U_B| > 500, |U_A \cap U_B| > 30$인 B 로 한정.

- 고객과 B, 고객과 C에 대한 입력값(25*2개) 추가

  - (user, product) 관련 입력값 14개
  - (user, aisle) 관련 입력값 3개
  - product 관련 입력값 8개

- A와 같은 Department 에 있는 상품들 중에 B와 C추출

  - Coke 안살때 Fridge Pack Cola를 사는 것은 대체품이라 생각할 수 있지만
  - Cola 안살때 Original Whole Grain English Muffins 을 산다고 Cola의 대체품이라고 하기는 설득력이 떨어진다.

2. Cola and Fridge pack cola

- Department별 고려 상품수

500명 이상이 구매한 상품수

id	department	전체 상품수	500고객 상품수	예외1	예외2	최종상품수
1	frozen	4,007	464	0	0	464
2	other	548	2	2	0	0
3	bakery	1,516	222	0	1	221
4	produce	1,684	590	0	0	590
5	alcohol	1,054	23	0	0	23
6	international	1,139	79	1	0	78
7	beverages	4,365	399	95	46	258
8	pets	972	2	2	0	0
9	dry goods pasta	1,858	204	0	1	203
10	bulk	38	9	1	3	5
11	personal care	6,563	57	2	7	48
12	meat seafood	907	123	0	0	123
13	pantry	5,371	533	0	0	533
14	breakfast	1,115	164	0	0	164
15	canned goods	2,092	227	0	0	227
16	dairy eggs	3,449	715	0	0	715
17	household	3,085	207	0	0	207
18	babies	1,081	106	0	0	106
19	snacks	6,264	581	0	0	581
20	deli	1,322	169	0	0	169
21	missing	1,258	NA	NA	NA	NA
	합계	49,688	4,876	103	58	4,715

	고려	전체	비율
전체상품수	4,715	49,688	0.094892
전체 (고객,상품) 쌍 수	6,348,419	8,474,661	0.749106

- 예외1 : 상품A에 대해 $|U_A \cap U_B| > 30$ 인 B가 0개
- 예외2 : 상품A에 대해 $|U_A \cap U_B| > 30$ 인 B가 1개

- 그 외의 상품들에 대한 확률값은 baseline 모형에서 추정한 것 사용한다.

2. Cola and Fridge pack cola

- 결과

	Train F1	Val F1
New	0.43716	0.40042
baseline	0.41355	0.40045