R과 Python을 이용한 마케팅 데이터 머신러닝 분석

엘포인트 매출 데이터 분석 – 개인화 상품추천

엄영범 박찬호 임서경 마근영 이종은

R과 Python을 이용한
마케팅 데이터 머신러닝 분석

1. 프로젝트 개요

팀 구성원 및 역할
주제 선정 배경
프로젝트 수행 일정

2. 프로젝트 추진내역

프로젝트 분석방법
SEMMA - Sample
SEMMA - Exploration
SEMMA - Modification
SEMMA - Modeling
SEMMA - Assessment

3. 데이터분석

상품 추천 분석 절차
머신러닝 모델 – Matrix Factorization
머신러닝 모델 – SVD
머신러닝 모델 – XGBoost
모델 평가

4. 결 론

결론
프로젝트 리뷰
Q & A

개인화 상품 추천

L.point

마케팅 데이터

R과 Python을 이용한
마케팅 데이터 머신러닝 분석

1. 프로젝트 개요

팀 구성원 및 역할
주제 선정 배경
프로젝트 수행 일정

팀 구성원 및 역할

엄영범

Project Manager

- 프로젝트 총괄
- Matrix Factorization 활용한 데이터 분석
- 최종 상품추천 파일 구현

박찬호

Engineer

- Python을 이용한 Matrix Factorization, XGBoost, SVD 분석
- 구매주기 필터링 구현

임서경

Analyst

- 프로젝트 일정 관리
- Matrix Factorization 활용한 데이터 분석
- 포트폴리오 제작

마근영

Engineer

- SVD 모델 구축 및 이를 활용한 데이터분석
- Matrix Factorization 활용한 데이터 분석

이종은

Analyst

- 데이터 전처리
- R을 이용한 XGBoost데이터 분석
- 최종 상품추천 파일 구현

주제선정배경

마케팅 빅데이터 분석은 고객 정보 데이터나 고객의 구매 이력 데이터 분석에서 나오는 관계, 규칙을 통해 고객의 욕구를 보다 정확하게 예측하여 고객이 원하는 상품과 서비스를 제공하는 데에 이용될 수 있다.

마케팅 빅데이터 분석은 고객과의 밀접한 관계 유지와 기업의 수익성 증대의 필수적인 요소이다.

분석 주제

롯데 고객의 2년간의 소비 활동 및 경쟁사 활동, 소비패턴 데이터를 활용하여

1) 고객의 선호 상품 추천

2) 고객의 새로운 구매 위한 상품 추천

목 표

고객이 선호하거나 선호 할 것으로 예상되는 상품 추천

롯데그룹 4개 제휴사 고객 구매 데이터 ▶ 머신러닝 기법 활용한 데이터 분석 ▶ 개인화 상품 추천

프로젝트 수행 일정

비고	1주차					2주차					3주차					4주차					5주차			
	4	5	6	7	8	11	12	13	14	15	18	19	20	21	22	25	26	27	28	29	2	3	4	5
주제선정	■	■	■	■																				
외부/ 내부 데이터 탐색				■	■																			
데이터 전처리				■	■																			
파생 변수 추가(구매 생명 주기)				■	■	■	■	■	■	■														
데이터 분석(MF, XGBoost)						■	■	■	■	■	■	■	■	■	■	■								
SVD 모델 구축 및 분석													■	■	■	■								
ML 모델 정교화														■	■	■								
예측															■	■								
책 제작																■	■	■	■	■	■	■		
PPT 작성																	■	■	■	■	■	■		
발표 준비																				■	■	■	■	

6월 4일 부터 7월 4일까지
총 23일(작업일) 동안
상황과 인원에 맞게 프로젝트를 수행

프로젝트 발표일

<	July					>
SUN	MON	TUE	WED	THU	FRI	SAT
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

R과 Python을 이용한
마케팅 데이터 머신러닝 분석

2. 프로젝트 추진내역

프로젝트 분석 방법

SEMMA - **S**ample

SEMMA- **E**xploration

SEMMA - **M**odification

SEMMA - **M**odeling

SEMMA - **A**ssessment

프로젝트 분석방법

데이터 생성/추출 [Sample]
- 모델링 및 모델 평가를 위한 데이터 준비

데이터 탐색 [Exploration]
- 데이터 조망을 통한 데이터 오류 검색
- 모델의 효율 증대
- 데이터 현황을 통해 비즈니스 이해
- 아이디어를 위해 이상 현상, 변화 등을 탐색

데이터 수정/변환 [Modification]
- 데이터가 보유한 정보의 표현 극대화
- 최적의 모델을 구축할 수 있도록 다양한 형태로 변수를 생성, 선택, 변형

모델 구축 [Modeling]
- 데이터의 숨겨진 패턴 발견
- 다양한 모델과 알고리즘 적용

모델 평가 및 검증[Assessment]
- 모델의 검증
- 서로 다른 모델을 동시에 비교
- 추가 분석 수행 여부 결정

SEMMA
[Sample]

데이터 설명

[데이터 셋 6종]

고객정보
01
고객번호
성별
연령대
거주지

구매내역
02
제휴사
영수증번호
대분류코드
중분류코드
...

상품분류
03
제휴사
상품코드
중분류명
소분류명
....

경쟁사이용
04
고객번호
제휴사
경쟁사
이용년월

멤버십여부
05
고객번호
멤버십명
가입년월

채널이용
06
고객번호
제휴사
이용횟수

SEMMA [Exploration 1]

연령별 구매내역 탐색

- 연령 별 판매 상위 10개 품목을 분석
- 모든 연령대에서 **종량제 봉투 / 재사용 봉투** 판매 순위가 높음
- 개인화 상품 추천이라는 분석 주제에 어긋난다고 판단하여 최종 상품 추천에서 배제

연령별 베스트 셀러 TOP 10

순위	19세이하	20~24세	25~29세	30~34세	35~39세	40~44세	45~49세	50~54세	55~59세	60세이상
1	종량제봉투	일반스낵	일반스낵	종량제봉투	종량제봉투	일반스낵	일반스낵	일반흰우유	청과	청과
2	두부류	종량제봉투	종량제봉투	일반스낵	일반스낵	종량제봉투	종량제봉투	일반스낵	채소	유기농채소
3	일반스낵	쿠키	커피음료	국산맥주	국산맥주	국산맥주	어묵	종량제봉투	유제품	유제품
4	감자스낵	혼합탄산	쿠키	커피음료	일반흰우유	어묵	일반흰우유	어묵	유기농채소	채소
5	어묵	국산맥주	국산맥주	일반흰우유	일반우유	일반흰우유	일반우유	청과	일반흰유우	농산가공
6	고추	과일음료	콜라	재사용봉투	떠먹는 요구르트	일반우유	두부류	두부류	종량제봉투	기능성우유
7	일반우유	생수	혼합탄산	떠먹는 요구르트	어묵	감자스낵	국산맥주	일반우유	농산가공	일반흰우유
8	바나나	공병/공박스	일반흰우유	감자스낵	재사용봉투	두부류	감자스낵	떠먹는 요구르트	두부류	종량제봉투
9	국불 봉지라면	일반소주	감자스낵	어묵	감자스낵	떠먹는 요구르트	떠먹는 요구르트	유제품	어묵	두부류
10	생수	감자스낵	과일음료	일반우유	두부류	크래커	마시는 요구르트	채소	떠먹는 요구르트	떠먹는 요구르트

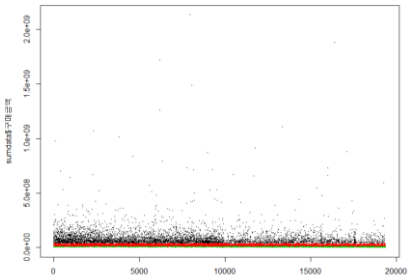
개인 상품 추천에서
종량제 봉투/재사용봉투 제외

SEMMA [Exploration 2]

구매금액 기준 분류 분석

- 롯데 제휴사를 이용한 고객의 2년간의 구매금액 합산을 통해 고객의 소득 수준을 추정 할 수 있을 것이라 여김
- 소득 수준이 비슷한 집단 간 구매패턴이 유사성을 가질 것이라 가정
- 구매 금액을 기준으로 그룹을 나누어 분석

구매금액 총합

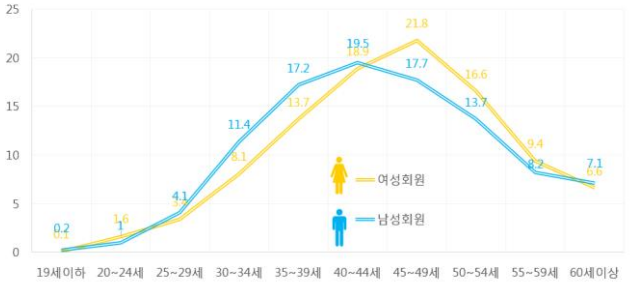


개인 상품 추천 알고리즘
Matrix Factorization 적용

나이 & 성별 기준 분류 분석

- 전통적인 통계 분석 기준인 인구통계학적 변수 탐색
- 30대 후반에서 50대 초반의 연령대가 구매 점유율이 높았음
- 인구통계학적 변수인 성별 / 연령대 기준으로 분류 분석

연령대 점유율



개인 상품 추천 알고리즘
SVD 적용

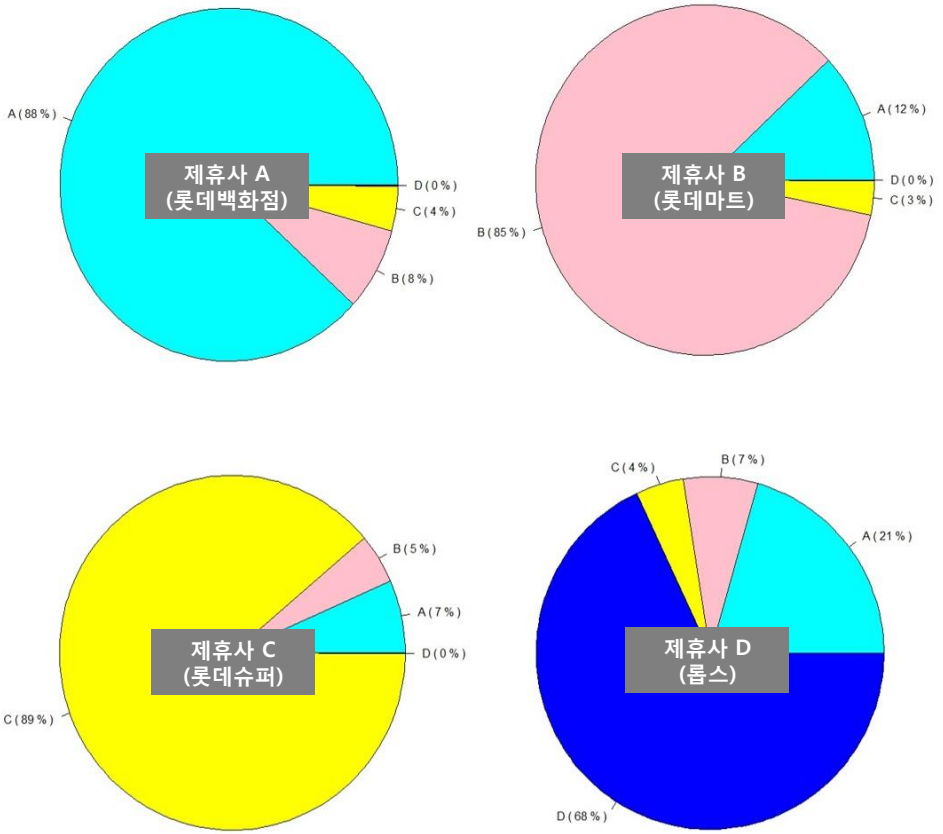
SEMMA
[Exploration 3]

제휴사 기준 분류 분석

- 제휴사 A : 롯데백화점 추정
- 제휴사 B : 롯데마트 추정
- 제휴사 C : 롯데슈퍼 추정
- 제휴사 D : 롭스 추정
- 롯데 그룹의 4개의 제휴사별로 매출 베스트셀러 품목들이 확연히 다른 것 확인
- 각 제휴사별 구매내역을 탐색한 결과 실제로 고객마다 자주가는 제휴사가 존재하는 것으로 확인됨
- 각 제휴사별 고객들의 구매 물품과 소비 패턴이 다르다고 여김
- 각 자주 가는 제휴사를 기준으로 나누어 분류 분석

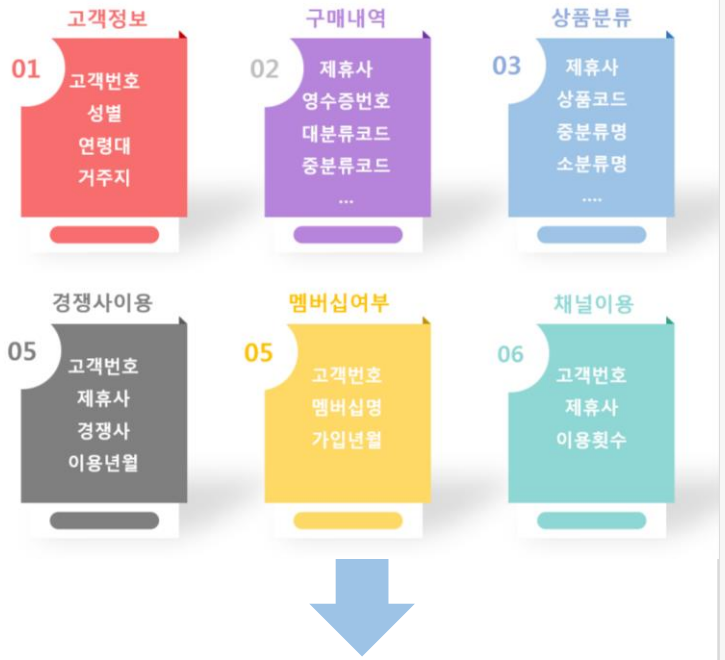
개인 상품 추천 알고리즘
XGBoost 적용

제휴사별 충성고객



SEMMA [Modification]

데이터셋 6종 – 결측치 탐색 및 제거



6개의 데이터셋 중 고객정보 테이블의
거주지 변수에서 결측치가 발견되어
결측치를 0으로 대체
(총 178명)

데이터 전처리 & 파생변수

채널클러스터 (구매금액&채널)

고객의 총 구매 금액과 채널 이용 횟수 기준
kmeans를 이용 6개 군집으로 클러스터링

고객 등급

제휴사 A : 롯데백화점 B : 롯데마트 C : 롯데슈퍼
D : 롭스의 우수 고객 등급 부여

구매 시간

고객 구매 시간을 2시간 단위로 나눠서 그룹화

구매 클러스터

고객의 총 구매금액과 총 구매횟수를 기준으로
클러스터링

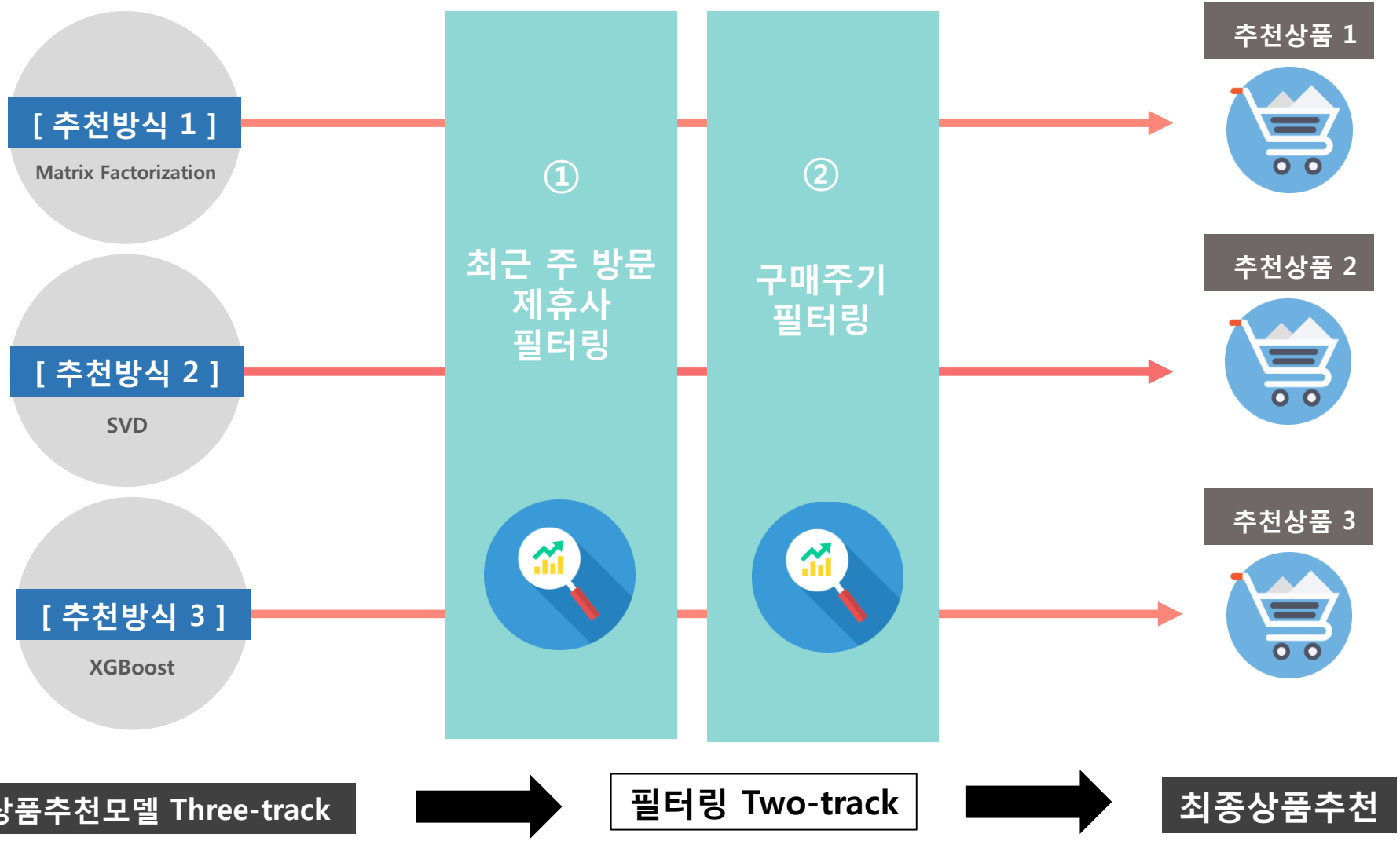
자주 방문하는 제휴사

최근에 자주 방문했던 제휴사

경쟁사 1 / 2 / 3 이용횟수

예) 제휴사X 충성고객 중 X의 경쟁사 이용한 횟수

SEMMA [Modeling]



상품추천모델 Three-track → 필터링 Two-track → 최종상품추천

SEMMA
[Assessment]

고객번호	성별	연령대	추천상품1	추천상품1 코드명	추천상품2	추천상품2 코드명	추천상품3	추천상품3 코드명
10283	M	50세~54세	A010402	채소	A010613	전문베이커리	A010302	유제품
10284	M	60세이상	A010402	채소	A010401	청과	A010403	유기농채소
10285	F	45세~49세	C030901	일반계란	C030204	오이	C030310	양파
10286	F	55세~59세	C030401	두부류	C070102	기능성우유	C030403	묵류
10287	F	55세~59세	B140101	감자스낵	B740101	봉지라면	B100504	떠먹는요구르트
10288	F	35세~39세	B100305	일반흰우유	B430101	어묵	A010302	유제품
10289	F	40세~44세	C030401	두부류	C150301	크래커	C170631	일반문구/사무용품
10290	F	60세이상	B100502	어린이/액상요구르트	B140103	일반스낵	B080601	주유소
10291	F	50세~54세	B100305	일반흰우유	B080601	주유소	B100402	커피/초코우유
10292	F	50세~54세	B160201	국산맥주	B160402	막걸리	B080601	주유소
10293	F	40세~44세	C030401	두부류	C070102	기능성우유	C030603	양송이버섯류
10294	F	40세~44세	B200602	주방균일가	B100305	일반흰우유	B140202	크래커
10295	F	40세~44세	A010602	일식델리	A010401	청과	A010603	서양델리
10296	F	50세~54세	B050117	친환경채소(특약)	B140103	일반스낵	B050901	친환경쌈채소
10297	M	50세~54세	C030901	일반계란	C060406	베이커리일반빵	C070101	일반우유
10298	F	55세~59세	B100305	일반흰우유	B100501	마시는요구르트	B050501	파프리카
10299	F	45세~49세	C030401	두부류	C070101	일반우유	C150301	크래커
10300	F	35세~39세	B100305	일반흰우유	B140104	쌀스낵	B120202	용기면

추천상품1 : Matrix Factorization

추천상품2 : SVD

추천상품3 : XGBoost

R과 Python을 이용한
마케팅 데이터 머신러닝 분석

3. 데이터분석

상품 추천 분석 절차
머신러닝 모델 – MF
머신러닝 모델 – SVD
머신러닝 모델 – XGBoost

상품 추천 분석 절차

추천 방식 1 : Matrix Factorization

각 고객의 상품구매 빈도를 Rating(평가 점수)활용 (고객과 소분류코드로 구성 된 테이블)

고객의 구매 금액에 따라 (상위 25% 중위 25~75% 하위 25%) 고객군 나누어 Matrix Factorization 알고리즘 적용

추천 방식 2 : SVD

고객 인구통계학적 변수(성별, 연령대) 바탕의 고객군 분류

18개의 고객군에 행이 고객번호, 열이 소분류코드, 행렬값이 구매빈도인 행렬 만들어 SVD 알고리즘적용

추천 방식 3 : XGBoost

제휴사 이용고객별 고객군 분류

① 최근 주 방문 제휴사 필터링

최상위 카테고리인 대분류코드 Target 변수 선정

XGBoost 모델을 구축하여 대분류코드 기준의 구매예측 수행

대분류코드 내 자주 구매했던 소분류코드 상품 추천

예측 Rating 높은 순으로 소분류코드 선정

① 최근 주 방문 제휴사 필터링

최근 자주 방문한 제휴사의 상품을 추천

② 구매주기 필터링

추천상품 1

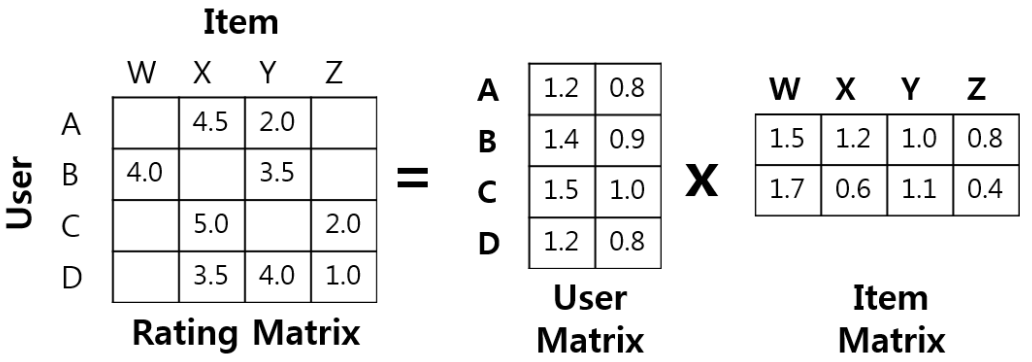
추천상품 2

추천상품 3

Matrix Factorization

Matrix Factorization란?

- 협업필터링의 희소성 문제를 해결할 수 있는 머신러닝 기법
- 행렬을 유저행렬과 아이템 특성 행렬로 분해하여 실제 행렬의 근사행렬을 만들 수 있음
- 근사된 행렬과 기존 행렬의 차이를 손실함수로 정의하여 손실 함수값 줄이는 수식 적용
- 오차 최소화하는 최적화 과정 거쳐 최종 예측값 산출



상품 추천 방향 설정

개인 맞춤형 상품 추천 1

개인 고객 소득 수준 -> 개인 고객 소비 패턴
2년간의 구매데이터 활용해 고객별 구매액 합산
구매금액 수준별 고객군 분류

행렬 q : 고객번호, 행렬 p : 소분류코드
Rating(행렬값) : 개인 고객별 구매 빈도

개인 상품 추천 알고리즘

Matrix Factorization 적용

Matrix Factorization

Matrix Factorization 적용 코드

```
> r = Reco()
> tr <- data_file("trainset.txt")
> test <- data_file("testset.txt")
> system.time(opts <- r$tune(tr, opts = list(lrate = c(0.1,0.2),
+ costp_l1 = 0, costq_l2 = 0, nthread=4, niter=10)))
> system.time(r$train(tr, opts= c(opts$min, nthread = 1, niter = 10)))
> pred_rvec <- r$predict(test, out_memory())
> pred_dt <- as.data.table(pred_rvec)
> write.csv(pred_dt, "recommend_result.csv", row.names = FALSE)
> id_product <- fread("testset.txt")
> result <- pred_dt
> result_bind <- cbind(id_product, result)
> write.csv(result_bind, "id_result_bind.csv", row.names=FALSE)
```

	A	B	C
1	V1	V2	pred_rvec
2	1	9010302	37.30334
3	2	9010302	58.08261
4	7	9010302	15.0014
5	8	9010302	183.7474
6	9	9010302	23.49161
7	11	9010302	41.12875
8	17	9010302	38.11887
9	18	9010302	41.03048
10	19	9010302	51.01313
11	22	9010302	16.13621
12	23	9010302	39.00629
13	25	9010302	6.3184

상품 추천 1 Matrix Factorization

- 설정한 파라미터에 따른 최적 모델 탐색
- 최적화된 모델로 train set 훈련
- 최종 예측값 추출 후 test set과 합쳐 추출
- 고객번호, 소분류코드, 해당 점수의 형태

Matrix Factorization

MF 추천 알고리즘 적용

구매금액 상위 25%

구매금액 중위 25%~75%

구매금액 하위 25%

고객의 구매 금액에 따라
(상위 25%, 중위 25~75%, 하위 25%)
고객군 나누어
Matrix Factorization
알고리즘 적용

자체 평가 모델 방법

추천한 상품과 실제로 12월에
구매한 상품 중 일치한 개수 구하기

실제 12월 구매 영수증

날짜	고객번호	소분류코드
2015/12/..	1	A1021
2015/12/..	2	C1151
2015/12/..	...	...

비교

MF알고리즘의 추천 상품

고객번호	소분류코드
1	A1021
2	B1031
...	

56%

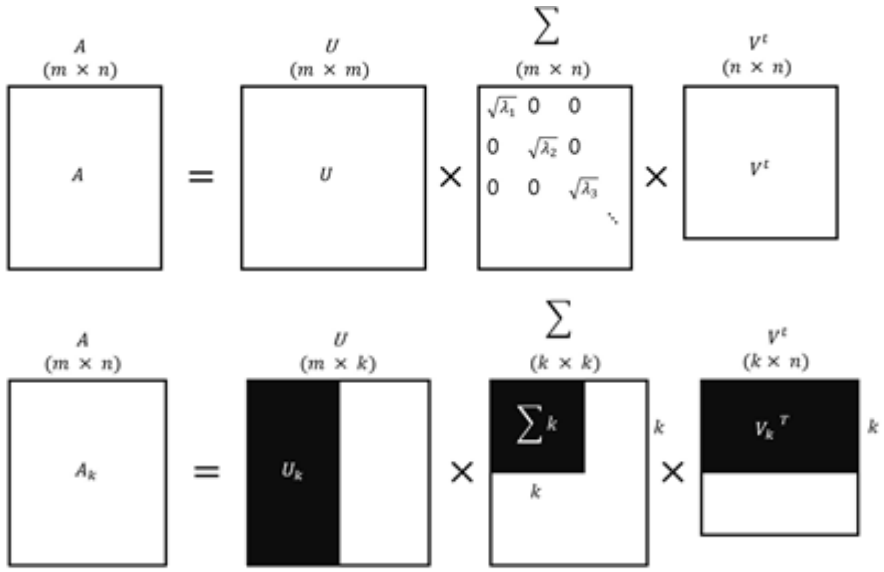
알고리즘 정확도 자체 평가 결과

10,669명(알고리즘일치하는고객)
/19,119명(12월에 구매가 발생한 고객의 수)

SVD (Single Value Decomposition)

SVD란?

- 머신러닝 기법으로 선형대수가 적용된 기법
- 행렬 A를 고유값분해 후 직교대각화 한 직교행렬 U, 직교행렬 V 그리고 A의 주성분 분석 값을 대각성분으로 가지는 직사각 대각행렬 S로 분해
- 데이터들을 행렬화시켜 차원축소를 통해 원래 행렬에 대한 근사행렬을 만들 수 있음
- 현재까지의 데이터(축소된 데이터)를 집어넣어 미래를 예측(축소 전 데이터)할 수 있음
- 잠재의미분석의 핵심 알고리즘



상품 추천 방향 설정

개인 맞춤형 상품 추천 2

인구통계학적 기준 활용
성별과 연령대로 분류

행렬 q : 고객번호, 행렬 p : 소분류코드
Rating(행렬값) : 개인 고객별 구매 빈도

고객 그룹군

총 18개 그룹

24세이하		25세~29세		30세~34세	
남	여	남	여	남	여
25세~39세		40세~44세		45세~49세	
남	여	남	여	남	여
55세~59세		50세~54세		60세 이상	
남	여	남	여	남	여

개인 상품
추천알고리즘

SVD 적용

SVD (Single Value Decomposition)

SVD 적용 코드

```
> tr <- read.table("trainset.txt", sep=" ", header=T)
> svd.cast <- dcast(tr, 고객번호 ~ 소분류코드, value.var="value", fun=NULL)
> svd.cast <- svd.cast[c(order(svd.cast$고객번호)),]
> svd.cast <- subset(svd.cast, select=-c(고객번호))
> svd.cast[is.na(svd.cast)] <- 0
> svd1 <- svd(svd.cast, nrow(svd.cast), ncol(svd.cast))
> U <- svd1$u[1:nrow(svd1$u), 1:round(length(svd1$d)*0.5)]
> D <- diag(svd1$d)[1:round(length(svd1$d)*0.5),
  1:round(length(svd1$d)*0.5)]
> TV <- t(svd1$v)[1:round(length(svd1$d)*0.5), 1:ncol(t(svd1$v))]
> svd2 <- U %*% D %*% TV
> svd3 <- melt(svd2)
> pred_dt <- subset(svd3, select=c(value))
> id_product <- fread("testset.txt")
> result_bind <- cbind(id_product, pred_dt)
> names(result_bind) <- c("V1", "V2", "pred_rvec")
> write.csv(result_bind, "id_result_bind.csv", row.names=FALSE)
```

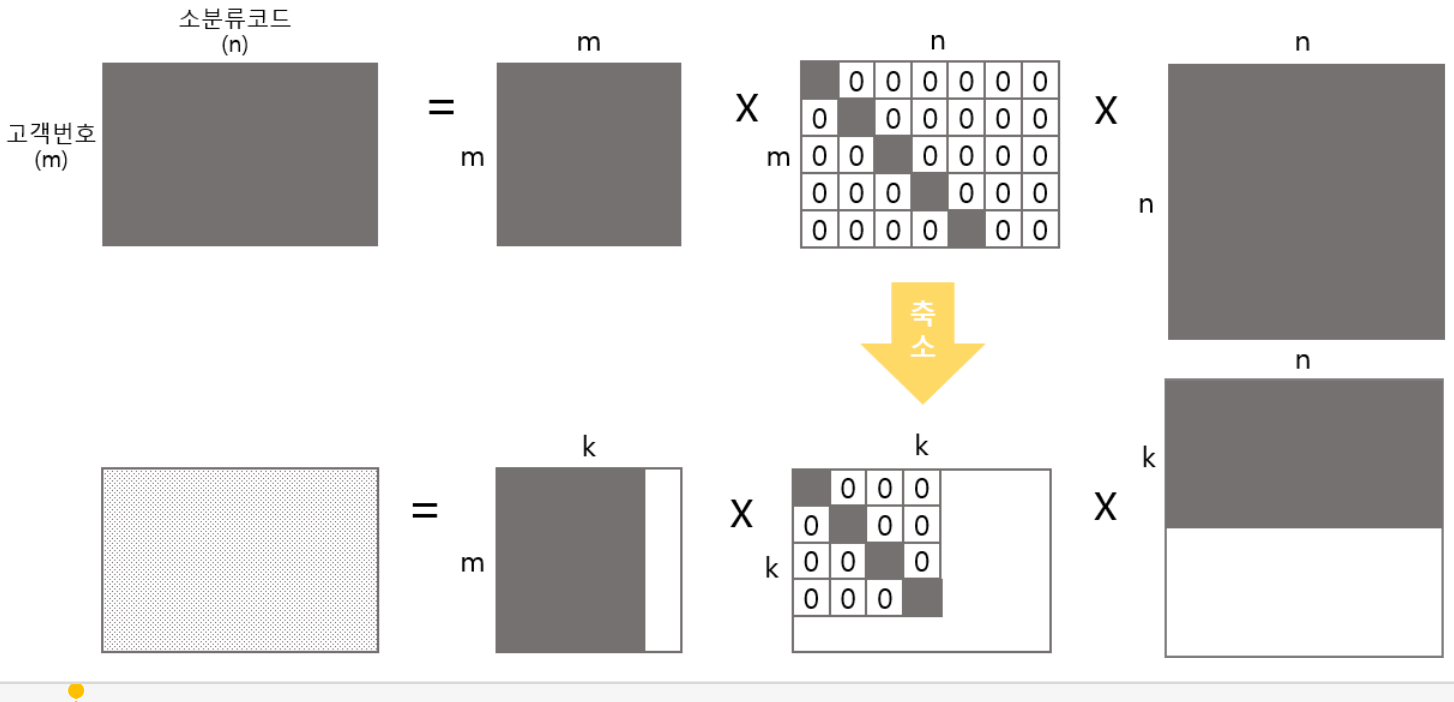
	A	B	C	D
1	V1	V2	pred_rvec	
2	1	6010101	0.159471	
3	2	6010101	-0.06382	
4	3	6010101	0.005794	
5	5	6010101	0.026665	
6	8	6010101	-0.0359	
7	11	6010101	-0.05398	
8	12	6010101	-0.01792	
9	17	6010101	0.043479	
10	20	6010101	0.074541	
11	21	6010101	-0.0095	
12	23	6010101	-0.04005	
13	24	6010101	-0.00872	

상품 추천 2 SVD

- 행이 고객번호, 열이 소분류코드, 행렬값은 구매빈도인 행렬을 생성(NA값은 0으로 대체)
- svd 함수를 이용해 행렬을 분해
- svd.cast = svd1$u %*% diag(svd1$d) %*% t(svd1$v)
- 위 식을 만족하는 $m \times m$, $m \times n$, $n \times n$ 3개의 행렬로 분해됨
- 행렬을 50%까지 축소해 예측한 후 소분류코드의 점수들을 저장하고 파일로 추출

SVD (Single Value Decomposition)

SVD 추천 알고리즘 적용



고객의
성별/나이
기준
18개 고객군
SVD 적용

78%

알고리즘 정확도 자체 평가 결과

14,897명(알고리즘일치하는고객)
/19,119명(12월에 구매가 발생한 고객의 수)

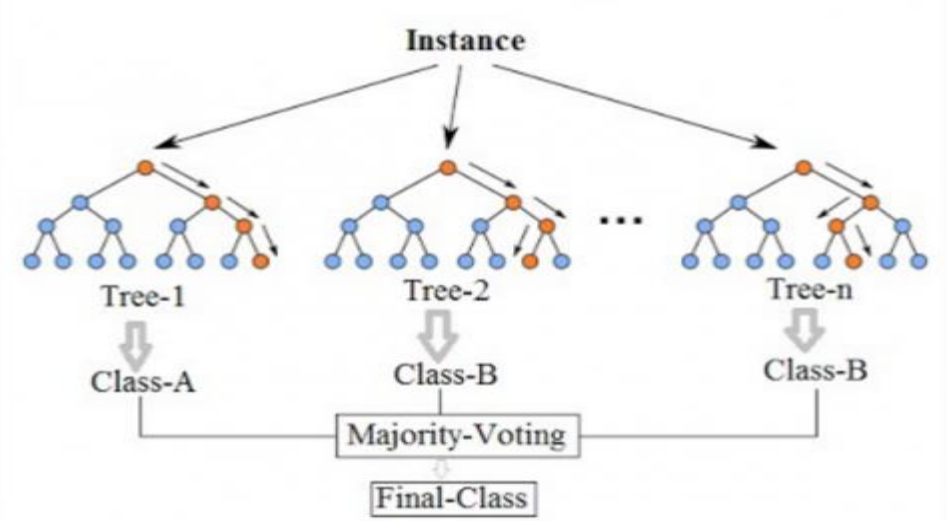
자체평가모델 방법

추천한 상품과 실제로
12월에 구매한 상품 중
일치한 개수 구하기

XGBoost

XGBoost란?

- **eXtreme Gradient Boosting의 줄임말**
- **앙상블 모형 : 여러 분류 모형의 결과를 조합하여 분류의 정확도 높이는 방법**
- **부스팅 : 잘못 분류된 데이터에 집중해서 새로운 분류 규칙을 만드는 단계를 반복하는 것**
- **일반적으로 의사 결정 트리의 앙상블 형태로 예측 모형 생성**



분류모델변수

기존변수		
	제휴사	구매시간
	구매금액	하이마트 보유
	성별	다둥이 보유
	연령대	롭스 보유
	거주지역	더영 보유

파생변수		
	채널클러스터 (구매금액&채널)	고객 등급 (제휴사별)
	구매 시간 (2시간 단위)	구매클러스터 (구매금액&횟수)
	경쟁사1 이용횟수 (e.g. 제휴사A의 충성고객 중 A의 경쟁사1을 이용한 횟수)	경쟁사2 이용횟수
		경쟁사3 이용횟수
	자주 방문하는 제휴사	

XGBoost

XGBoost 적용 코드

```
> for (product in 1:9){
>   product_merge <- fread("D:/final/Dummy_A.csv", select = c(product))
>   data_merge <- cbind(dataA, product_merge)
>   test <- filter(data_merge, (구매월 == 12 | 구매월 == 11) & 구매년도 == 15)
>   test <- as.data.table(test)
>   test <- test[,-c('구매월','구매년도')]
>   test_id <- test[,c('고객번호')]
>   train1 <- filter(data_merge, (구매월 != 12 & 구매월 != 11) & 구매년도 ==15)
>   train2 <- filter(data_merge, 구매년도 ==14)
>   train <- rbind(train2, train1)
>   train <- as.data.table(train)
>   train <- train[,-c('구매월','구매년도')]
>   labels <- train[[colnames(product_merge)]]
>   ts_label <- test[[colnames(product_merge)]]
>   labels <- as.numeric(labels)
>   ts_label <- as.numeric(ts_label)
>   train <- train[,-c(1,21)]
>   test <- test[,-c(1,21)]
>   dtrain <- xgb.DMatrix(data = as.matrix(train),label = labels)
>   dtest <- xgb.DMatrix(data = as.matrix(test),label=ts_label)
>   params <- list(booster = "gbtree", objective = "binary:logistic", eta=0.05, gamma=0,
+ max_depth=4, min_child_weight=1, subsample=1, colsample_bytree=1)
>   xgb1 <- xgb.train(params = params, data = dtrain, nrounds = 20, watchlist =
+ list(train=dtrain), early_stop_round = 50, maximize = F, eval_metric = "auc")
>   result <- data.table(test_id, predict(xgb1,dtest))
>   colnames(result) <- c("고객번호", colnames(product_merge))
>   write.csv(result, paste0("D:/final/xgboost/A제휴사/test_",
+ colnames(product_merge), "_1월", ".csv"), row.names = FALSE)
> }
```

	A	B	C	D	E
1	고객번호	더미_A1	더미_A2	더미_A3	더미_A4
2	1	0.79636	0.186427	0.043758	0.185453
3	1	0.602651	0.244511	0.049268	0.236336
4	1	0.79636	0.186427	0.043758	0.185453
5	1	0.602651	0.244511	0.049268	0.236336
6	1	0.732235	0.214341	0.043758	0.204079
7	1	0.232177	0.31552	0.083575	0.34924
8	1	0.218532	0.264171	0.069092	0.384591
9	1	0.79636	0.185868	0.043758	0.185453
10	1	0.379833	0.297832	0.083575	0.320839
11	1	0.732235	0.214341	0.043758	0.204079
12	1	0.181647	0.181647	0.040356	0.181647
13	1	0.79636	0.185868	0.043758	0.185453
14	1	0.732235	0.214341	0.043758	0.204079
15	1	0.181647	0.181647	0.040356	0.18165
16	1	0.218532	0.31552	0.083575	0.366099
17	1	0.732235	0.202431	0.043758	0.202147
18	1	0.79636	0.186427	0.043758	0.185453
19	1	0.719108	0.20987	0.046578	0.208892
20	1	0.79636	0.186427	0.043758	0.185453
21	1	0.615094	0.244511	0.064508	0.236336

상품 추천 3 XGBoost

- 저장해둔 더미변수 데이터를 한 열 씩, 즉 대분류코드 한 개씩 가져와서 기존 데이터와 병합
- 병합데이터에서 최근 2개월의 데이터를 추출하여 test set으로 설정 후 나머지는 train set으로 설정
- 각각의 데이터 셋의 더미변수 값을 label로 두고 더미변수와 고객번호 제거
- train set과 test set을 matrix 형태로 바꿔줌
- XGBoost에 사용할 parameter 생성 후 XGBoost 적용
- 적용하여 나온 모델에 test set을 적용시켜 나온 결과를 result에 저장하고 csv 파일로 저장

XGBoost

XGBoost 추천 알고리즘 적용

1 고객별 최근 자주 이용한 제휴사 선정

- 고객 구매내역데이터에서 최근 50%의 거래내역 추출
- 가장 자주 이용한 제휴사를 선정
- 해당 제휴사의 상품만을 대상으로 추천

2 Train set / Test set 구성

- 최근 구매내역을 고려
- Train set : 2014년01월~2015년10월
- Test set : 2015년 11월, 12월

3 제휴사별 대분류코드 Target 변수선정

- 최상위 카테고리인 제휴사 대분류코드 기준
- 선호상품(대분류코드) Target변수로 선정
  -> 중분류코드나 소분류의 경우
  학습량 부족 & 불필요한 비용 소모

4 XGBoost 적용

- 대분류코드 기준 고객별 상품구매확률도출
- 각 대분류코드 상품구매확률 계산
- 선호상품(대분류코드) 도출

자체평가모델 방법

- 대분류 코드 구매 확률 예측
- 실제 고객의 구매내역 중 가장 많이 구매한 대분류와 대조하여 일치하는 갯수 구하기

알고리즘 정확도 자체 평가 결과

10,609명(알고리즘일치하는고객)
/19,274명(11월, 12월에 구매가 발생한 고객의 수)

55%

R과 Python을 이용한
마케팅 데이터 머신러닝 분석

4. 결 론

개인화 상품 추천

결론

프로젝트 리뷰

Q & A

개인화 상품 추천

고객번호	성별	연령대	추천상품1	추천상품1 코드명	추천상품2	추천상품2 코드명	추천상품3	추천상품3 코드명
10283	M	50세~54세	A010402	채소	A010613	전문베이커리	A010302	유제품
10284	M	60세이상	A010402	채소	A010401	청과	A010403	유기농채소
10285	F	45세~49세	C030901	일반계란	C030204	오이	C030310	양파
10286	F	55세~59세	C030401	두부류	C070102	기능성우유	C030403	묵류
10287	F	55세~59세	B140101	감자스낵	B740101	봉지라면	B100504	떠먹는요구르트
10288	F	35세~39세	B100305	일반흰우유	B430101	어묵	A010302	유제품
10289	F	40세~44세	C030401	두부류	C150301	크래커	C170631	일반문구/사무용품
10290	F	60세이상	B100502	어린이/액상요구르트	B140103	일반스낵	B080601	주유소
10291	F	50세~54세	B100305	일반흰우유	B080601	주유소	B100402	커피/초코우유
10292	F	50세~54세	B160201	국산맥주	B160402	막걸리	B080601	주유소
10293	F	40세~44세	C030401	두부류	C070102	기능성우유	C030603	양송이버섯류
10294	F	40세~44세	B200602	주방균일가	B100305	일반흰우유	B140202	크래커
10295	F	40세~44세	A010602	일식델리	A010401	청과	A010603	서양델리
10296	F	50세~54세	B050117	친환경채소(특약)	B140103	일반스낵	B050901	친환경쌈채소
10297	M	50세~54세	C030901	일반계란	C060406	베이커리일반빵	C070101	일반우유
10298	F	55세~59세	B100305	일반흰우유	B100501	마시는요구르트	B050501	파프리카
10299	F	45세~49세	C030401	두부류	C070101	일반우유	C150301	크래커
10300	F	35세~39세	B100305	일반흰우유	B140104	쌀스낵	B120202	용기면

결 론

Matrix Factorization

순위	상위25%	소분류명	중위50%	소분류명	하위25%	소분류명
1	B080601	주유소	B080601	주유소	B100305	일반흰우유
2	A010401	청과	B100305	일반흰우유	C030401	두부류
3	A010302	유제품	C030401	두부류	B200801	다이소
4	A010402	채소	A010401	청과	B140103	일반스낵
5	B100305	일반흰우유	C070101	일반우유	B160201	국산맥주
6	A040902	디자이너부틱	B140103	일반스낵	B080601	주유소
7	A010403	유기농채소	A010302	유제품	A011003	원두커피
8	A010613	전문베이커리	A010402	채소	C070101	일반우유
9	A020302	기초 화장품	A010613	전문베이커리	B100301	기능성우유
10	A030114	아동놀이시설	B100504	떠먹는요구르트	C170701	생활잡화균일가

구매금액 기준 분류 분석

- 상위 25% 고객들의 경우 디자이너 부티크, 유기농 채소 등이 추천됨
- 중위 50%와 하위 25% 고객들의 경우 추천품목 대부분이 식료품
- 상대적으로 소득 수준이 낮은 고객들의 소비가 주로 필수재에 집중

56%

결 론

SVD

순위	24세이하,남	24세이하,여	25세~29세,남	25세~29세,여	30세~34세,남	30세~34세,여
1	다이소	페이셜팩류	쿠키	일반흰우유	일반흰우유	일반흰우유
2	국산담배	쿠키	커피음료	쿠키	일반우유	청과
3	국물용기라면	일반스낵	일반흰우유	기초 화장품	커피음료	유기농채소
…	…	…	…	…	…	어린이/요구르트(8)

순위	35세~39세,남	35세~39세,여	40세~44세,남	40세~44세,여	45세~49세,남	45세~49세,여
1	일반흰우유	일반흰우유	일반흰우유	일반흰우유	일반흰우유	일반흰우유
2	일반우유	청과	청과	일반우유	청과	일반우유
3	청과	다이소	일반우유	청과	유제품	청과
…	…	어린이/요구르트(10)	…	…	아웃도어(4)	…

순위	50세~54세,남	50세~54세,여	55세~59세,남	55세~59세,여	60세이상,남	60세이상,여
1	일반흰우유	일반흰우유	청과	청과	청과	청과
2	청과	청과	일반흰우유	일반흰우유	채소	일반흰우유
3	채소	일반우유	다이소	유기농채소	일반흰우유	채소
…	아웃도어(4)	…	아웃도어(5)	…	…	…

성별 / 연령대 기준 분류 분석

- 20대 초반 여성 : 화장품
- 20대 초반 남성 : 국물용기라면
- 30대 여성 : 어린이/요구르트
- 45세 이상 남성 : 아웃도어
- 성별과 연령대를 기준으로 나누어진 그룹의 구매품목이 서로 다른 특징을 가짐

78%

결 론

XGBoost

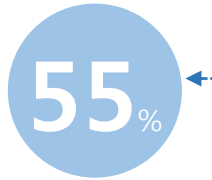
순위	A(롯데백화점)	B(롯데마트)	C(롯데슈퍼)	D(롭스)
1	청과	일반흰우유	두부류	일반스낵
2	채소	떠먹는요구르트	일반계란	과일음료
3	유제품	두부류	일반흰우유	하드캔디
4	농산가공	청과	청과	판초콜릿
5	전문베이커리	기능성시유	일반우유	페이셜팩류
6	유기농채소	일반스낵	일반스낵	혼합탄산
7	일반흰우유	다이소	바나나	채소
8	기초 화장품	주유소	감자스낵	일반흰우유
9	두부류	기능성우유	바아이스크림	쿠키
10	수입식품	어린이/액상요구르트	마시는요구르트	떠먹는요구르트

자주 가는 제휴사 기준 분류 분석

- A : 청과, 채소, 농산가공 등의 농산품
- B : 유제품과 일상용품
- C : 유제품과 스낵, 과일
- D : 페이셜팩류, 스낵, 캔디, 초콜릿

- A와 D는 각 그룹군이 명확한 특징을 보임
- B와 C는 비슷해 보이지만 유제품의 성격과 할인 매장의 유무가 다름

- 제휴사별로 다른 구매패턴 -> 추천내역도 달라짐



프로젝트 리뷰

01 구매금액

MatrixFactorizaion

56%

02 자주 가는 제휴사

SVD

78%

03 나이 / 연령대

XGBoost

55%

프로젝트 평가

- 집단별로 구매 패턴이 상이할 것이라 가정
- 구매금액 / 자주 가는 제휴사 / 나이 및 성별별로 데이터 셋을 나눔
- 각 데이터 셋에 Matrix Factorization, XGBoost, SVD 모형을 적용하여 모델을 생성
- 고객이 구매할 확률이 높은 품목을 후보로 선정 후 구매주기 필터링을 거쳐 최종 추천 품목들을 결정
- 구매금액 / 자주 가는 제휴사 / 나이 및 성별별로 상이 패턴을 보이는 것을 확인
- 자체 평가한 결과, 56%(Matrix Factorization), 55%(XGBoost), 78%(SVD)의 적중률(평균 63%)을 기록
- 집단별로 나누어 적용한 머신러닝 모델들이 충분히 유의미한 결과를 도출

R과 Python을 이용한
마케팅 데이터 머신러닝 분석

"Our focus should be not on emerging technologies but on emerging cultural practice"

우리가 주목해야 하는 것은 새롭게 떠오르는 기술이 아니다.
새롭게 나타나는 문화적 현상이다.

- Henry Jenkins -

R과 Python을 이용한
마케팅 데이터 머신러닝 분석