SAS 분석 챔피언십

분석 결과 보고서

SA296

김유경 신소윤 최지수

목 차

1. 분석 주제에 대한 이해

‘나도 몰랐던 나의 취향을 찾다’

비디오 포털 사이트를 이용하는 고객들에게 원하는 맞춤 상품을 추천해주는 것을 통하여, 고객의 해당 사이트에 대한 선호도를 높여줌. 고객도 몰랐던 취향을 찾는다는 건 고객이 지금까지 접하지 않았던 새로운 취향을 발견할 수 있도록 고객의 소비 이력에 기반한 추천 결과 뿐 아니라, 그동한 소비 하지 않았던 새로운 상품도 추천 목록에 포함시켜 주는 걸 의미한다고 생각

2. DATA 준비 과정

- 분석에 사용할 변수와 사용하지 않을 변수 구분
- 분석에 사용할 변수의 결측치/특이값 대체
- 변수 범주화
- 파생변수 생성

2. DATA 준비 과정

■ 분석에 사용할 변수와 사용하지 않을 변수 구분

- **Contents_meta (메타 테이블)**
- **Customer (사용자 데이터)**
- **History (시청/구매/찜 이력)**

사용

- History_detail (조회 이력)
- Search (검색어 데이터)

미사용

2. DATA 준비 과정

■ 분석에 사용할 변수와 사용하지 않을 변수 구분

- **Contents_meta (메타 테이블)**

  - 사용하지 않는다고 공지된 변수들+같은 정보를 담고 있다고 판단되는 변수들

    + 사용하고 싶었으나 사용하지 못한 변수들을 빼고 나머지 변수들을 분석에 사용

- **Customer (사용자 데이터)**

  - 모든 변수 사용

- **History (시청/구매/찜 이력)**

  - 모든 변수 사용

2. DATA 준비 과정

■ 분석에 사용할 변수와 사용하지 않을 변수 구분

- **Contents_meta (메타 테이블)**

: 비슷한 정보를 담고 있다고 판단되는 변수들 정리

-날짜 관련 변수

- 장르 관련 변수

- 제작국가 관련 변수

- 배우 관련 변수

- PD 관련 변수

} 미사용

2. DATA 준비 과정

■ 분석에 사용할 변수와 사용하지 않을 변수 구분

네이버 영화 ⓘ

다른 사이트를 보시려면 여기를 누르세요.

조폭 마누라 3 (My Wife Is A Gangster III, 2006)

네티즌 ★★★★☆ 7.54 (2,378) | 기자·평론가 ★★★☆☆ 6.00 (3) 평점주기

코미디, 액션 | 2006.12.28 개봉 | 115분 | 한국 | 15세 관람가

감독 조진규

내용 홍콩 최고의 명문 조직 화백련 보스의 외동딸 아령(서기), 보스 ... 더보기

다운로드 ♡80

날짜 변수들 중
일반적으로 영화 정보를 제공할 때 사용하는 날짜인
개봉날짜를 대표로 사용.

조폭마누라3 (My Wife is a Gangster 3 : HK Edition)

기본정보 | 통계정보 | 상영현황정보

최종수정: 2015-12-24 14:55:42 수정요청

코드 20060357

A.K.A (My Wife is a Gangster 3)

요약정보 장편 | 일반영화 | 액션, 코미디 | 115분 | 15세이상관람가 | 한국, 홍콩

개봉일 2006-12-28 | 제작연도 2006년 | 제작상태 개봉

크랭크인/업 해당정보 없음 | 촬영회차 해당정보 없음

상영타입 필름 : 필름(20060357N)

2. DATA 준비 과정

■ 분석에 사용할 변수의 결측치/특이값 대체

FREQ 프로시저

ok	빈도	백분율	누적 빈도	누적 백분율
0	8875	44.52	8875	44.52
1	11058	55.48	19933	100.00

합계 행: 7 합계 칼럼: 2

	release_date	broad_date
1	20140225	2월25일(화)
2	20140811	8월11일(월)
3	20140918	20140917
4	20140901	9월1일(월)
5	20150911	9월11일(금)
6	20151209	12월9일(수)
7	20151202	12월2일(수)

Release_date=broad_date일 때 ok=1

즉, 약 56% 정도 Release_date=broad_date 이다.

이 때, ok=0인 8875개의 값들 중 Release_date랑 broad_date **둘 중 하나만 결측인 경우를 제외하면** 다음의 **7개만 남음.**

따라서, 두 변수는 같은 정보를 담고 있다고 취급하고,

결측률이 더 작은 release_date를 사용하되,

만약 **release_date가 결측**이고 broad_date 가 값이 있으면 그 때는 **broad_date값을 가져옴.**

2. DATA 준비 과정

■ 분석에 사용할 변수의 결측치/특이값 대체

변수명	라벨	입력형태	결측률
album_date	앨범등록일	20140731093855	0%
release_date	앨범출시일	20090618	49.92%
broad_date	방영일자	NA	94.36%
update_date	업데이트일	13OCT2014	4.32%
create_date	생성일	20090807	0%
year	연도	NA	99.24%
make_year	제작연도	2009년	86.15%

가장 빠른 날짜부터 오름차순으로 점점 진한 색 칠함.

같은 의미의 변수로 취급

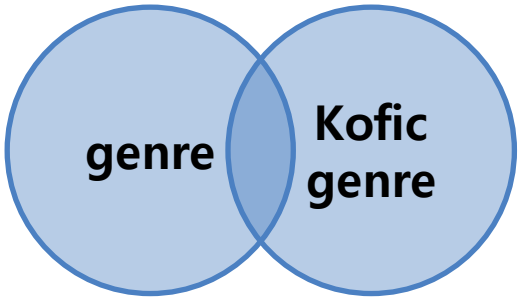
Release_date에서 board_date 를 이용해도 대체 할 수 없는 결측값은,

Release_date와 가장 가까운 **날짜인 create_date를 이용해 대체함.**

2. DATA 준비 과정

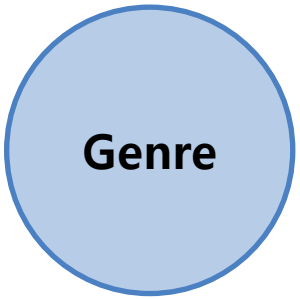
■ 분석에 사용할 변수와 사용하지 않을 변수 구분

Contents_cateogy="movie"



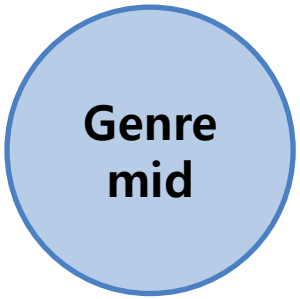
장르와 코픽장르를 합치고,
중복되는 부분은 한 번만 입력.

Contents_cateogy="tv_drama"



장르를 정리해 주고 사용

Contents_cateogy="variety"



중장르를 정리해 주고 사용

하나로 묶어서,
GENRE_NAME 변수를 만듦.

2. DATA 준비 과정

■ 분석에 사용할 변수의 결측치/특이값 대체

-genre_name (앞에서 여러 장르변수를 하나로 묶어 만든 변수)

장르_네임에 대한 결측값은 총 **20개.** 직접 검색해서 대체해줌.

2. DATA 준비 과정

■ 분석에 사용할 변수와 사용하지 않을 변수 구분

-제작국가 관련 변수들 정리

Country_of_origin 과 Make_nation 모두 "국가" 에 대한 정보를 담고 있음.

Country_of_origin의 결측률은 34.41%

Make_nation의 결측률은 86.14%

따라서, 결측률이 더 작은 **Country_of_origin을 기본으로 사용하고,**

Country_of_origin이 **결측인 경우 Make_nation 값을 대체해줌.**

2. DATA 준비 과정

■ 변수 범주화

관측빈도가 2200이하인 값들은
대륙을 기준으로 묶음.

2. DATA 준비 과정

■ 변수 범주화 / 결측치 처리

FREQ 프로시저

제작국가

country	빈도	백분율	누적 빈도	누적 백분율
NA	6791	34.07	6791	34.07
기타	301	1.51	7092	35.58
미국	3820	19.16	10912	54.74
영국	361	1.81	11273	56.55
유럽	980	4.92	12253	61.47
일본	2939	14.74	15192	76.22
중국	1576	7.91	16768	84.12
한국	2982	14.96	19750	99.08
홍콩	183	0.92	19933	100.00

-> NA를 제외하면, 8개의 범주로 축소됨.

-제작국가가 결측인 경우 결측값 대체 과정

1차로는 방송사 정보(category)를 통해

확실히 제작국가를 알 수 있는 상품들에 제작국가 기입

(방송사가 kbs면 제작국가는 한국)

2차로 나라가 결측인 상품 중 감독 이름이 6bite면

나라는 한국으로 함.

마지막으로, 나머지 상품들은 인터넷에 검색해서 결측값을 채워넣음.

2. DATA 준비 과정

■ 분석에 사용할 변수 전처리

- Contents_meta (메타 테이블) 의 **close_yn(종영여부)에 대한 결측값 처리**

1차로 종영여부에 따른 **상품 출시 년도의 간단한 통계량 구해봄.**

종영한 컨텐츠 중 가장 오래전에 출시한 컨텐츠의 년도(2007년)보다 더 전에 나온 컨텐츠는 종영한 것으로 구분함.

종영하지 않은 컨텐츠 중 가장 최근에 출시한 컨텐츠의 년도(2017)보다 더 후에 나온 컨텐츠는 종영하지 않은 것으로 구분함.

MEANS 프로시저

분석 변수 : 상품출시년도				
N	평균	표준편차	최솟값	최댓값
12100	2014.94	1.5467869	2007.00	2017.00

MEANS 프로시저

분석 변수 : 상품출시년도				
N	평균	표준편차	최솟값	최댓값
1351	2014.14	1.3229989	2011.00	2017.00

2. DATA 준비 과정

■ 분석에 사용할 변수와 사용하지 않을 변수 구분

- **close_yn(종영여부)에 대한 결측값 처리**

1차 과정을 통해 NA는 6482->4720로 감소함.

2차로 대장르가 영화인 경우 결측을 제외하면, 99%이상의 상품이 종영함.

따라서, **대장르가 영화인 경우 남은 결측은 전부 종영Y로 분류**

대장르가 방송인 경우 결측을 제외하면, 97%이상이 종영함.

따라서, **대장르가 방송인 경우 남은 결측은 전부 종영Y로 분류.**

2차에서 NA는 4720->63로 감소함.

그래도 분류안 된 상품들은 대장르가 애니, 다큐, 교육, 라이프.

직접 검색해보고 전부 종영으로 분류함.

2. DATA 준비 과정

■ 변수 범주화

- 앞에서 진행한 과정으로 결측을 없앤 release_date 를 년도만 끊어서 정리한 후 release_year로 변수명 저장함.

20세기 초중반	1927~1960
20세기 중후반	1961~1999
현대	2000~2009
최근	2010~2015
최신	2016~2017

- Price는 분위수에 기반해서 범주화 하고 price_level로 변수명 저장

분위수(정의 5)	
레벨	분위수
100% 최댓값	10000
99%	2500
95%	1500
90%	1500
75% Q3	1200
50% 중위수	700
25% Q1	500
10%	0
5%	0
1%	0
0% 최솟값	0

price	price_level
0	1
0<…≤500	2
500<…≤700	3
700<…≤1200	4
1200<…≤10000	5

- 관객수는 상위 약 1%만 인기있음(pop) 으로 코딩하고.

그 외 모든 상품은 unpop으로 코딩함.

2. DATA 준비 과정

■ 파생변수 생성

① 구매/시청/찜 이력 정리 & P변수 생성

B(구매)	W(시청)	F(찜)	Case	P(선호도)
0	0	1	찜만 한 경우.	0.5
0	1	0	찜없이 바로 시청한 것과 찜을 하고 시청한 것을 같은 경우로 침.	1
0	1	1		
1	0	0	구매만 한 것과 찜을 하고 구매한 것과 구매를 하고 시청한 것과 찜을 하고 구매 후 시청한 것은 같은 경우로 침.	1
1	0	1		
1	1	0		
1	1	1		

2. DATA 준비 과정

■ 파생변수 생성

② 앞에서 만든 P변수를 이용해, P_... 생성

P_... = 변수 내 해당 카테고리의 P값 총합/고객의 P값 총합

ex) P_미국 = 해당 고객이 본 미국 컨텐츠의 P값 총합 / 해당 고객이 본 전체 컨텐츠의 P값 총합

③ prefer_... 생성

P_...변수를 이용해 만든 변수. 0~5까지 6점 측도.

④ prefer_exist 생성.

어떤 변수의 prefer_... 중 하나라도 3이상인 값이 있으면,

prefer_exist는 Y 아니면 N

<단, > 범주가2개 인 경우,
prefer_...값이 같게 나왔을 경우는 prefer_exist="N"

2. DATA 준비 과정

■ 파생변수 생성

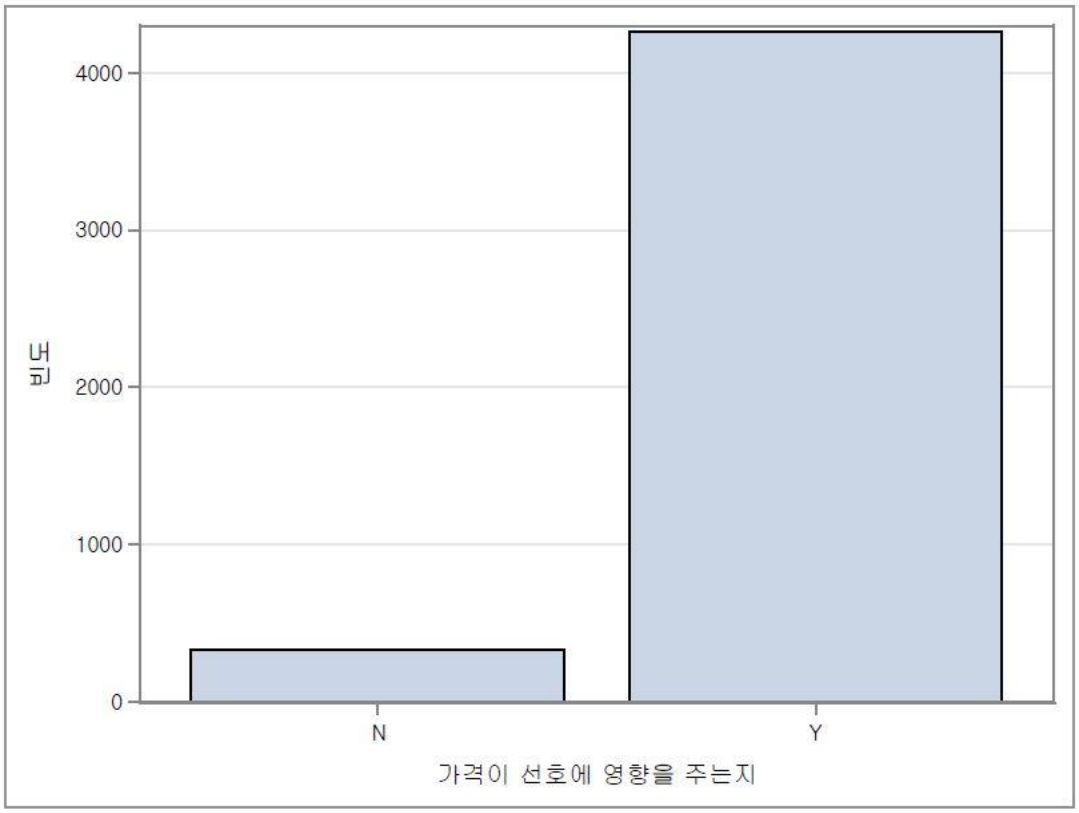
-prefer변수 생성 기준 & 파생변수 생성 예시

P_...	Prefer_...	level
0	0	선호 없음
0<...≤0.2	1	아주약한선호
0.2<...≤0.4	2	약한선호
0.4<...≤0.6	3	선호
0.6<...≤0.8	4	강한 선호
0.8<...≤1	5	아주 강한 선호

고객	상품	제작국가	구매/시청/찜	P(선호도)	P_한국	P_일본	P_미국	P_sum	Prefer_한국	Prefer_일본	Prefer_미국	prefer_exist
1	A	한국	구매	1	0.35	0.35	0.28	**2.8**	2	2	2	N
1	B	일본	구매	1	0.35	0.35	0.28	2.8	2	2	2	N
1	C	미국	시청	0.8	0.35	0.35	0.28	2.8	2	2	2	N
2	A	한국	구매	1	0.7	0.29	0	**3.4**	4	2	0	Y
2	B	일본	구매	1	0.7	0.29	0	3.4	4	2	0	Y
2	A	한국	시청	0.8	0.7	0.29	0	3.4	4	2	0	Y
2	D	한국	찜	0.6	0.7	0.29	0	3.4	4	2	0	Y

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색

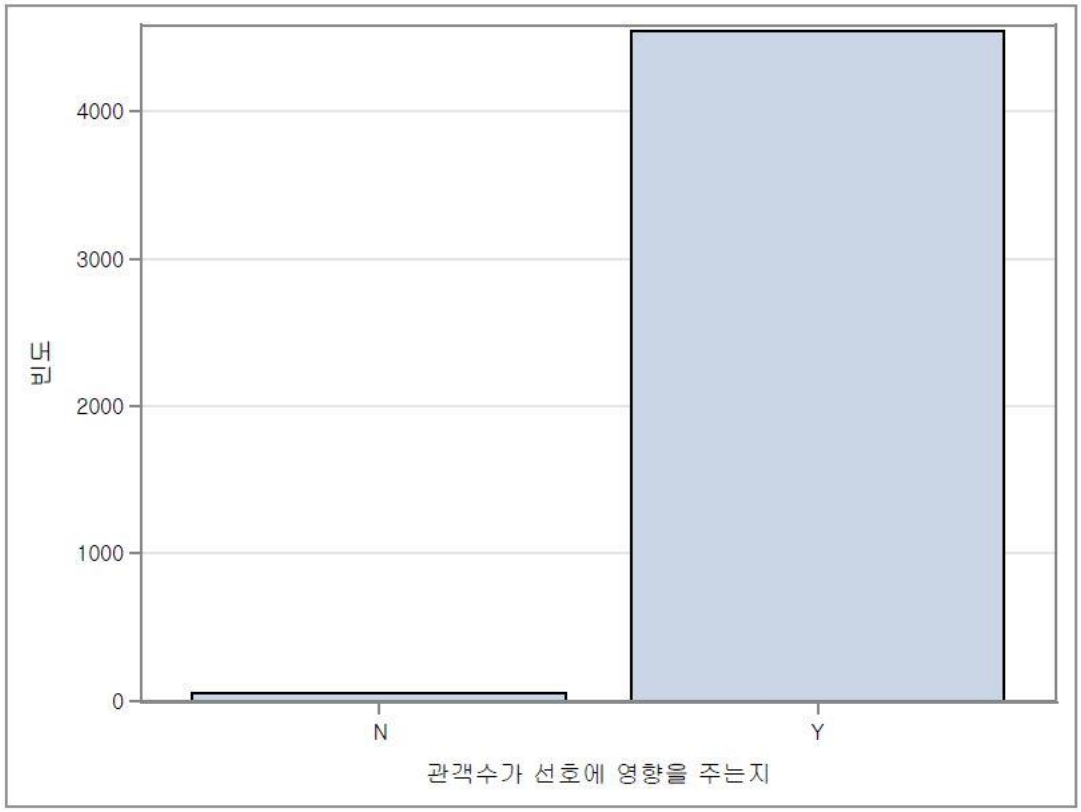


가격에 따른

고객별 Prefer_exist 막대그래프

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색

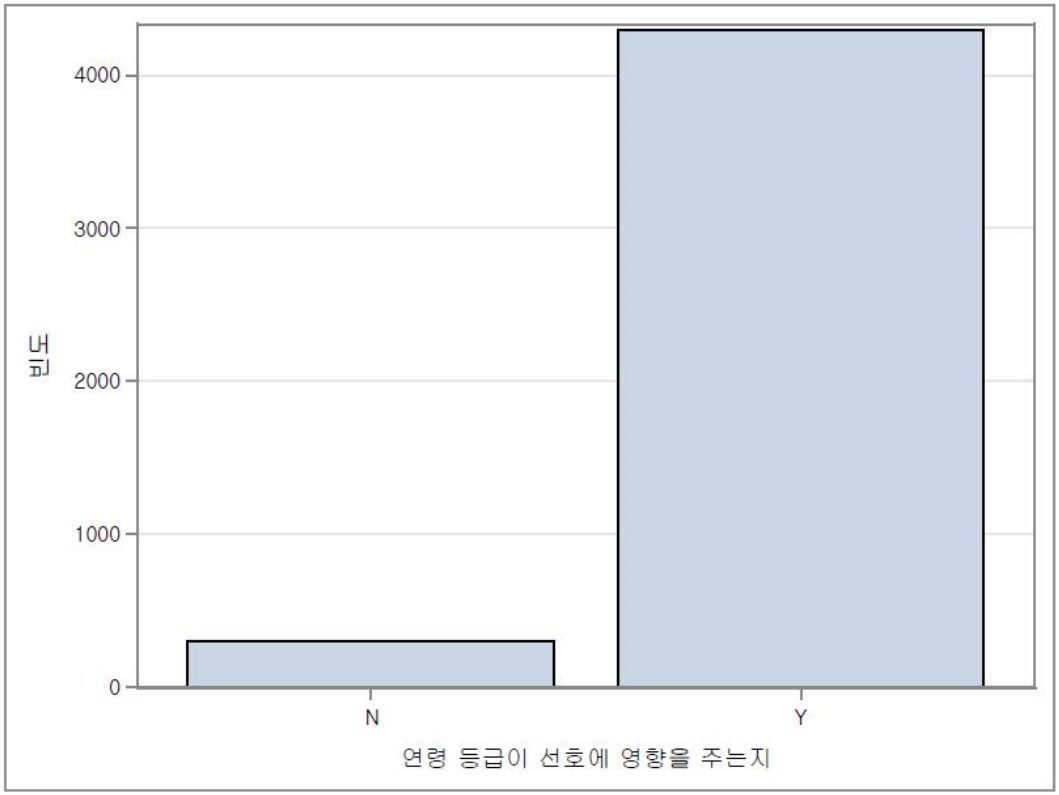


관객수에 따른

고객별 Prefer_exist 막대그래프

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색

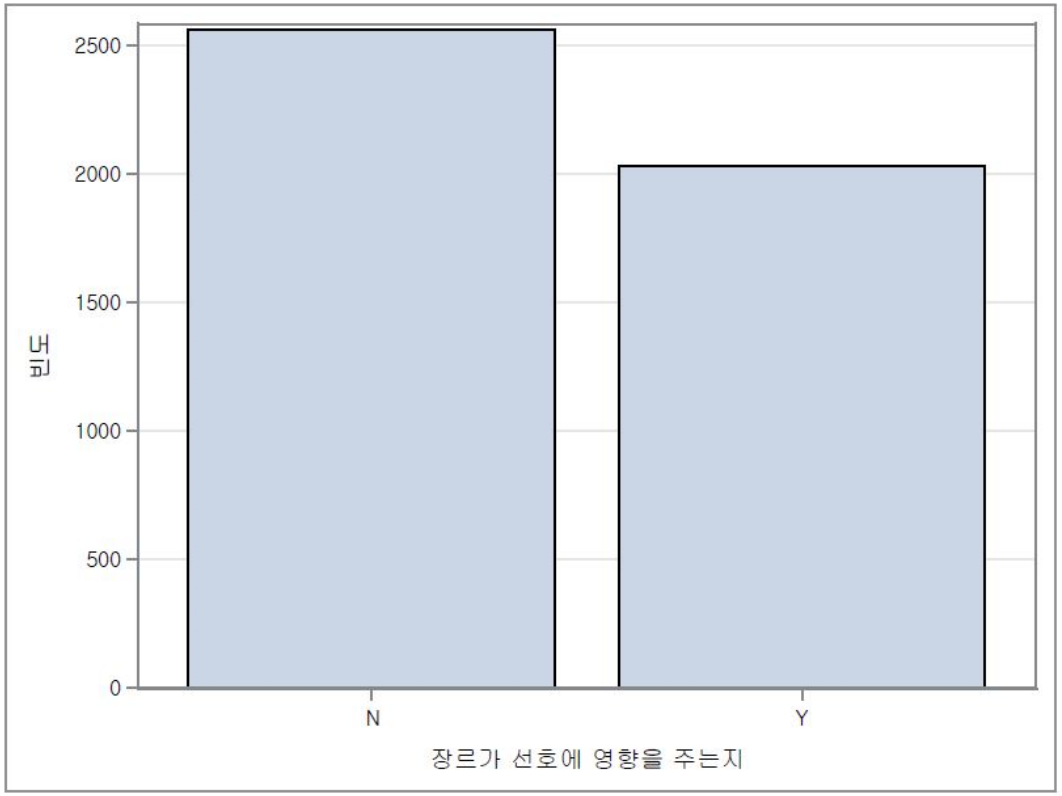


연령등급에 따른

고객별 Prefer_exist 막대그래프

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색

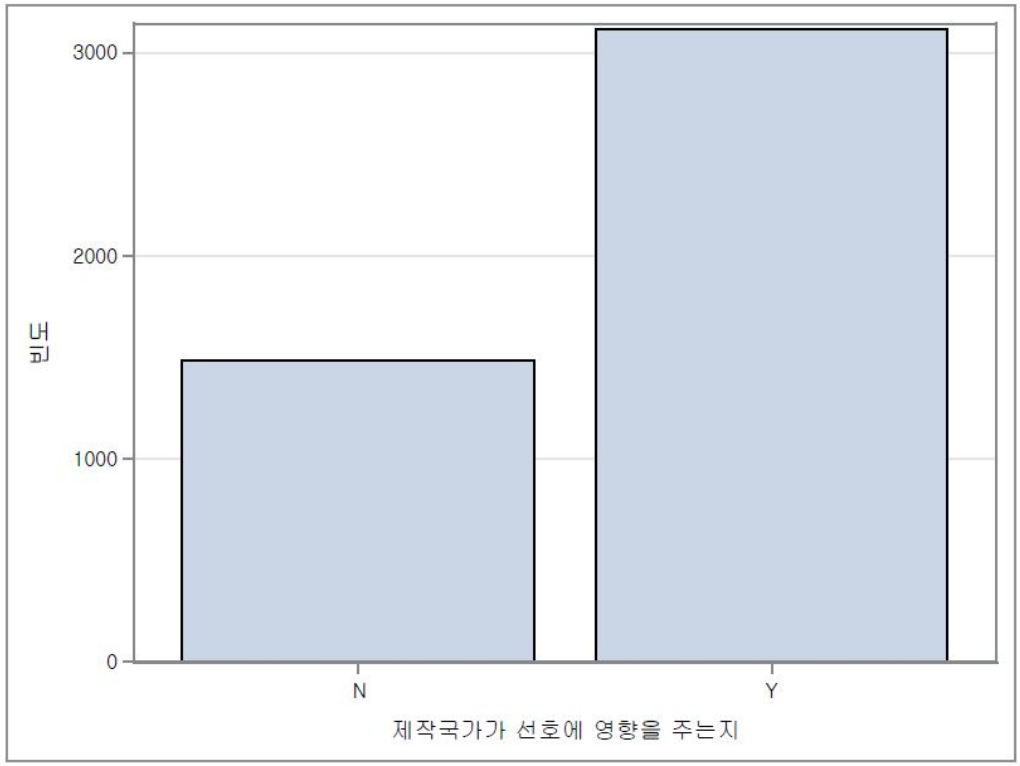


장르에 따른

고객별 Prefer_exist 막대그래프

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색

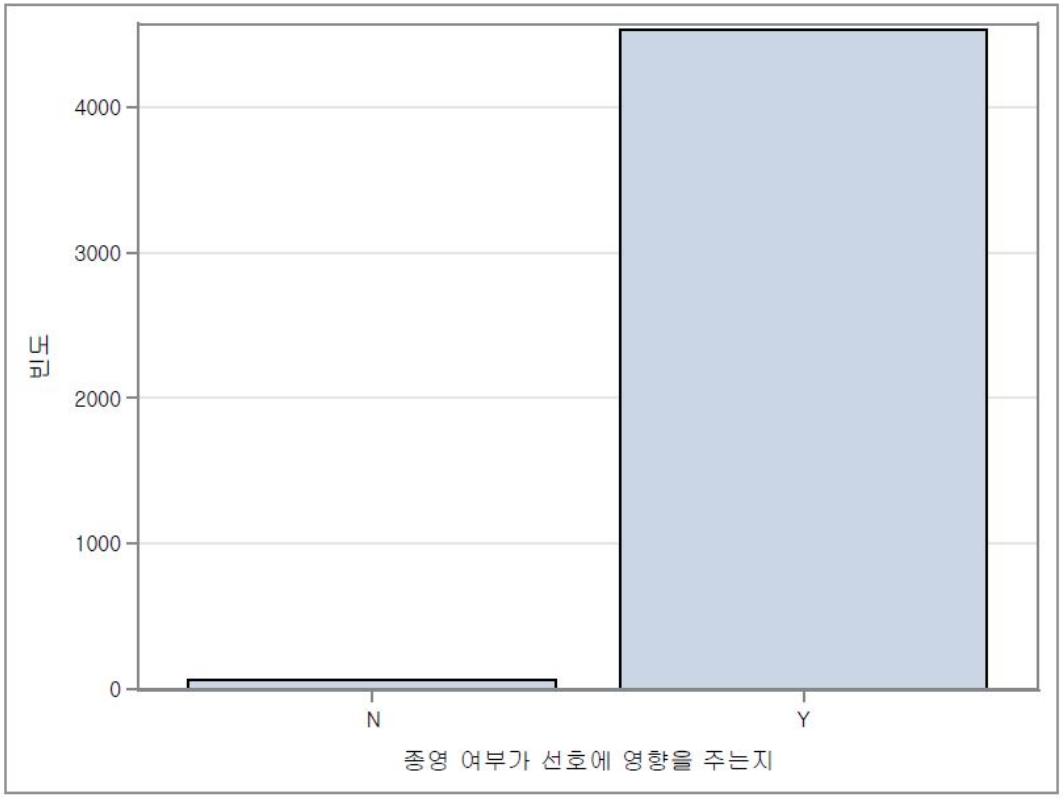


제작국가에 따른

고객별 Prefer_exist 막대그래프

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색

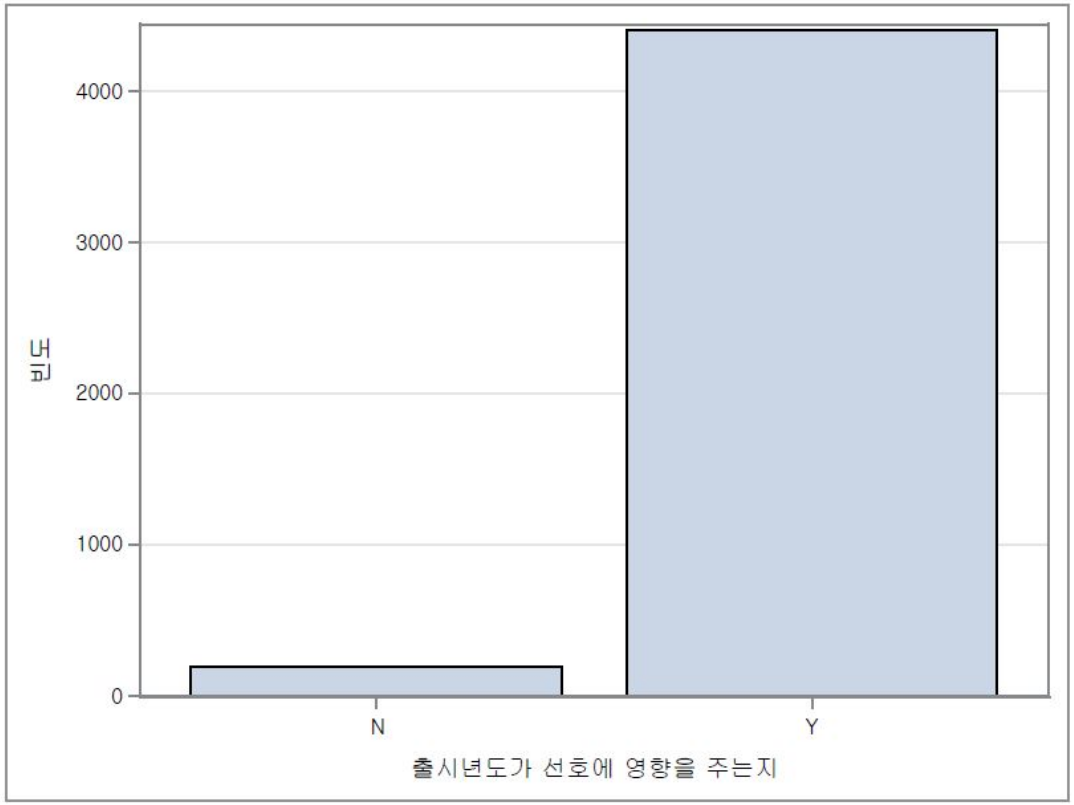


종영여부에 따른

고객별 Prefer_exist 막대그래프

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색

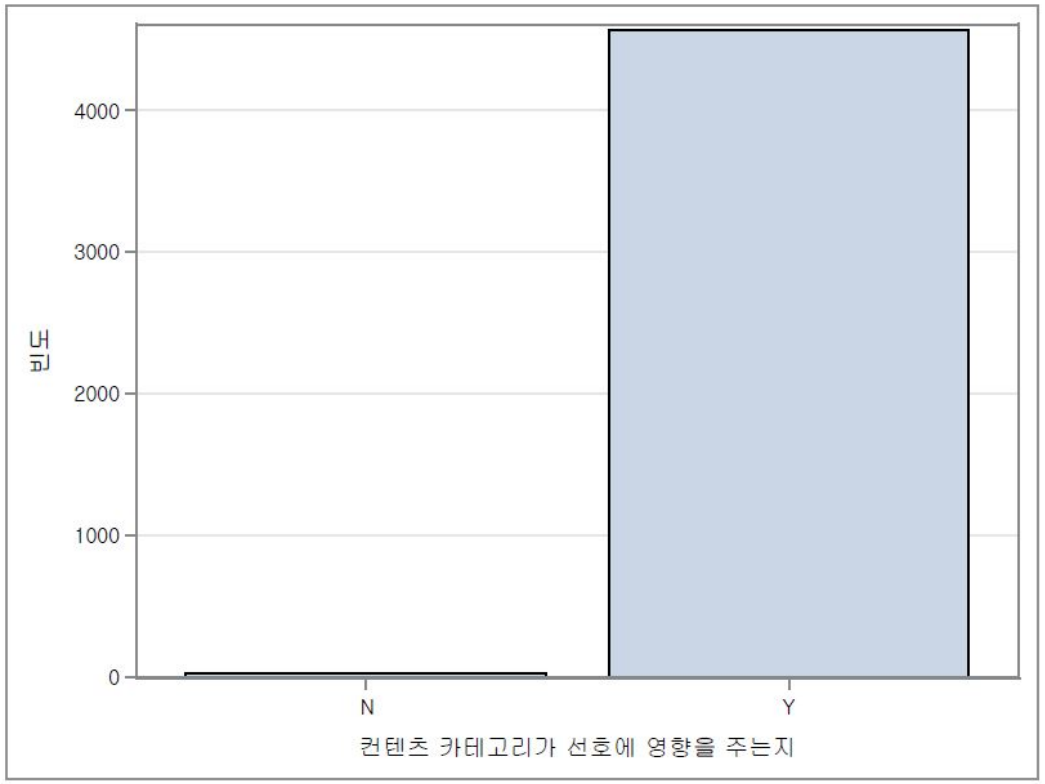


출시시기에 따른

고객별 Prefer_exist 막대그래프

2. DATA 준비 과정

■ 파생변수 생성을 통해 고객별로 선호도에 영향력 있는 세그먼트 탐색



컨텐츠 카테고리에 따른

고객별 Prefer_exist 막대그래프

3. 모델링

- 추천 알고리즘을 이용한 모델
- 로지스틱 회귀분석을 이용한 모델

3. 모델링

■ 추천 알고리즘을 이용한 모델

1) 고객 개인별 별점을 예측한 파생변수 생성 (implicit rating)

- 별점을 계산할때는 prefer_exist=N인 경우(Prefer_...값이 2이하 인것)은 별점에 포함시키지 않음

(ex. 고객이 제작국가에 대한 prefer_exist가 N인경우. 즉, 제작국가가 상품선호도에 영향을 미치지 않는 경우, 제작국가는 별점에 반영되지 않음)

- rating은 행 별로 하늘색 칸 값들의 평균값 (뒷 장에 그림 첨부)
- implicit_rating은 예측별점1을 소수점 둘째자리에서 반올림한 값

(implicit rating로 추천 알고리즘 실행)

3. 모델링

■ 추천 알고리즘을 이용한 모델

-별점 생성 예시

고객	상품	제작국가	장르	Prefer_한국	Prefer_일본	Prefer_미국	Prefer_로맨스	Prefer_공포	Prefer_코미디	rating	implicit_rating
2	A	한국	로맨스,코미디	4	2	0	4	0	3	3.666667	3.7
2	B	일본	로맨스	4	2	0	4	0	3	4	4
2	A	한국	코미디	4	2	0	4	0	3	3.5	3.5
2	D	한국	로맨스	4	2	0	4	0	3	4	4

3. 모델링

■ 추천 알고리즘을 이용한 모델

2) 아쉬운 점 : 왓챠 별점의 결측률이 높음.

우리가 생성한 예측 별점은 시청이력에 기반하였기 때문에,

시청하였다는 사실만으로 좋았다 또는 좋지 않았다(별점)을

알아내기 힘들다고 판단하여 왓챠별점과 평균을 내고 싶었으나,

왓챠 별점의 결측률이 84%가 되어 정보를 왜곡시킬 수 있다는 판단 하에

사용하지 않음.

3. 모델링

■ 추천 알고리즘을 이용한 모델

3) SAS에서 제공한 추천 알고리즘 코드 이용 (CF알고리즘)

- 컨텐츠 정보, 고객 정보, **생성한 예측 별점 정보**를 이용하여 추천 알고리즘 knn, svd, ensemble(knn+svd) 적용 후 고객별로 추천목록 5개씩 생성

- 컨텐츠 정보 중 키워드 정보를 이용해 유사한 아이템을 묶어주는 proc imstat(text parse옵션)을 통해 아이템기반 클러스터링 적용 후 고객별로 추천목록 5개씩 생성

3. 모델링

- 추천 알고리즘을 이용한 모델 성능 비교
  - knn
  - SVD
  - ensemble (knn+SVD)
  - cluster
  - ensemble (cluster+knn)
  - ensemble (cluster+SVD)
  - ensemble (knn+cluster+SVD)

3. 모델링

■ 추천 알고리즘을 이용한 모델

3) SAS에서 제공한 추천 알고리즘 코드 이용 (CF알고리즘)

- 성능 평가

PP기간에 구매/시청/찜 이력이 있는 고객에게 알고리즘을 통해 추천목록을 5개씩 생성하여 추천 성공률을 비교한 다음 가장 성공률이 좋은 알고리즘으로 결정

3. 모델링

■ 추천 알고리즘을 이용한 모델

3) SAS에서 제공한 추천 알고리즘 코드 이용 (CF알고리즘)

- 성능 평가

(1) knn(pc) vs knn(cv)

성공률 (단위 %)

seed	knn(pc)	knn(cv)
1	3.33	3.33
2	3.33	3.33
3	0	0
4	0	0
5	0	0
6	0	3.33
7	0	0
8	0	0
9	0	0
10	0	0

(seed1), (seed2)를 통해 랜덤으로 뽑은 30명에서
같은 고객에게 상품을 추천한 결과 1개가 일치하여 성공인데
pc, cv는 같은 고객에게 같은 상품을 추천함.

(seed6)를 통해 랜덤으로 뽑은 30명에서
sa_id = 12270 고객에게 cv로 추천한 결과만 성공함.

3. 모델링

■ 추천 알고리즘을 이용한 모델

3) SAS에서 제공한 추천 알고리즘 코드 이용 (CF알고리즘)

- 성능 평가

(2) cluster(15 개) vs cluster(30 개)

성공률 (단위 %)

seed	cluster(15)	cluster(30)
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	3.33	0
9	0	0
10	0	0

(seed8)를 통해 랜덤으로 뽑은 30명에서

sa_id=11433 고객에게 cluster(15)만 성공함.

3. 모델링

■ 추천 알고리즘을 이용한 모델

3) SAS에서 제공한 추천 알고리즘 코드 이용 (CF알고리즘)

- 성능 평가 (결론 : knn 모형 선택)

(3) knn이 사용될 경우(cv로 결정) 와 cluster가 사용될 경우(15개로 결정)

성공률 (단위 %)

seed	knn	svd	(knn + svd)	clust	(clust+knn)	(clust+svd)	(clust+knn+svd)
1	3.33	0	0	0	0	0	0
2	3.33	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	6.66	0	0	0	0	0
5	0	0	0	0	0	0	0
6	3.33	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	3.33	3.33	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0

성공률을 보았을 때,
knn 알고리즘이 가장 여러 번
성공횟수가 있었다.

svd 알고리즘의 경우
6.66% 의 성공이 있었으나,
우연히 맞았을 경우로 생각하여

추천 알고리즘에서는
knn 모형을 선택한다.

3. 모델링

■ 로지스틱 회귀분석을 이용한 모델

- 고객 별로 모든 상품에 대한 매트릭스를 생성함.
- buy 변수를 만듦. 고객이 해당 상품에 대한 bwf이력이 있으면 buy=1 아니면,

  buy=0 으로 코딩함. (X=close_yn, pr_info, contents_category, country, release_year,

  price_level ,audlevel, genre_name(한번에 여러값이,로 묶여있기 때문에 각각의 장르에 대해 변수를 만들어줌.

  Gerne_name을 돌리지 않고, ex. 스릴러(0,1)로 코딩되어 있는 변수를 만들고 돌림. Y=buy)
- 고객 별로 로지스틱 회귀모형을 만들고, 스코어링 데이터셋을 생성해서

  buy가 1이 나올 확률이 높은 상품 순으로 5개를 추천함.

  (op 또는 pp기간에 bwf이력이 있는 고객은 여기서 나온 추천 결과를 그대로 적용)
- 모든 고객을 성별/연령/사는지역 으로 클러스터링을 한 후에 각 군집별로 로지스틱

  회귀모형을 만들고, 나온 결과를 op pp 기간 모두 구매 이력이 없는 고객에게 추천.

4. 최종 모델 채택 및 결론

로지스틱 회귀분석을 이용한 모델이 더 성공률이 높으므로.

최종모형은 로지스틱 회귀분석을 통한 모형으로 채택

5. 마케팅 활용 방안

<마블 시청 이벤트> 등의 이벤트 진행 시
마블 히어로 영화 콘텐츠들이 추천 리스트에는 있으나
상위(5개 이내)에는 있지 않은 고객들 대상으로
<시청(or구매)을 하면 1000포인트 증정!>등의 이벤트를 진행하여
고객에게 알리면(App push알림 등의 방법으로)
자극을 한 고객들이 영상을 볼 확률이 높아질 것 이다.

(마블히어로 시청/구매 가능성이 많 높은 고객은 자극을 하지 않아도 콘텐츠만 노출을 해준다면 잘 보기 때문에 1000원을 주면서까지 유도할 필요는 없다.)