제5회 L.POINT Big DATA Competition
"Be the L.BA!"

분석 리포트
Digital Trend Analyzer

Team 포돌이
Members 김이영, 박희지, 정다연

Contents

온라인 행동 기반 트렌드 예측

I. 분석 주제 이해
분석 방향 수립

1.1 분석목적 및 주제

분석 목적

온라인 행동 기반
트렌드 예측

분석 주제

1) 주요 상품군별
온라인 선호지수 생성

2) 상품군별 수요 트렌드 예측 및
인사이트 도출

3) 1), 2)를 활용한 신규 서비스 제안

1.1 분석목적 및 주제

L.POINT | L.pay

회원수
3800만

엘포인트 결제액
5000억원
(2017년)

엘포인트 유관매출
40조원

롯데멤버스 엘페이가 급성장한 배경에는 고객 기반 **빅데이터 분석 능력**이 있다.
엘포인트 분석을 통해서 **소비자가 자주 결제하는** 숍과 상품 등에 대해
엘페이를 집중적으로 확대해왔다.

1.1 분석목적 및 주제

L.POINT | L.pay

고객들의 소비성향과
선호도가 빠르게
변화하는 시대

디지털 기술을
다루고 트렌드를
읽는 능력이 요구됨

엘포인트
시장의 확대

NTT도코모와 협약
일본 서비스 론칭 추진

고객들의 **소비 성향**과 **선호도**가 빠르게 변화하는 시대인 만큼
디지털 기술을 다루고 트렌드를 읽는 능력, 정교한 빅데이터 분석력이 매우 중요하다.
엘포인트 시장의 확대로 이러한 역량이 더욱 요구될 것으로 전망된다.

1.2 분석방향 수립

Step 0

계절성을 파악
분석하기에는
한계가 존재

Step 1

모형생성을 위한
군집화

Step 2

Data 분할 및
회귀분석

Step 3

훈련데이터를
바탕으로
검증 및 예측

Data 기간이 총 6개월

1) 이분법
2) K-fold Cross-validation방식

1.2 분석방향 수립

Step1. 모형 생성을 위한 군집화

1) 데이터셋 구성 → 2) 군집화 절차 → 3) 군집 결과 및 해석

1) 데이터셋 구성

✓ 엘포인트 월별 판매량 4-9월 판매량을 6개월 총 판매량으로 나눈 평균
→ 6개월 기준 평균 판매량

✓ 네이버 쇼핑 인사이트에서 2017년 8월 ~ 2018년 12월까지
총 1년 4개월의 상품군별 클릭수를 API로 불러들임

✓ 네이버 소분류 기준(약 1500개)과 & 엘포인트 소분류(898개)를 매칭시켜서
엘포인트 테이블의 계절성 파악 등의 요소를 보완시킴
→ 16개월 기준 평균 판매량으로 스케일링

1.2 분석방향 수립

Step1. 모형 생성을 위한 군집화

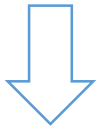
2) 군집화 절차

1. 월 평균 구매수가 50건이 안되는 상품군 소분류 제외
   → 898개에서 718개로 축소

2. Standard scaling 스케일링 실시

3. 주성분 분석(PCA) → 5개의 component로 축소

4. 병합군집 실시 → Agglomerative Clustering



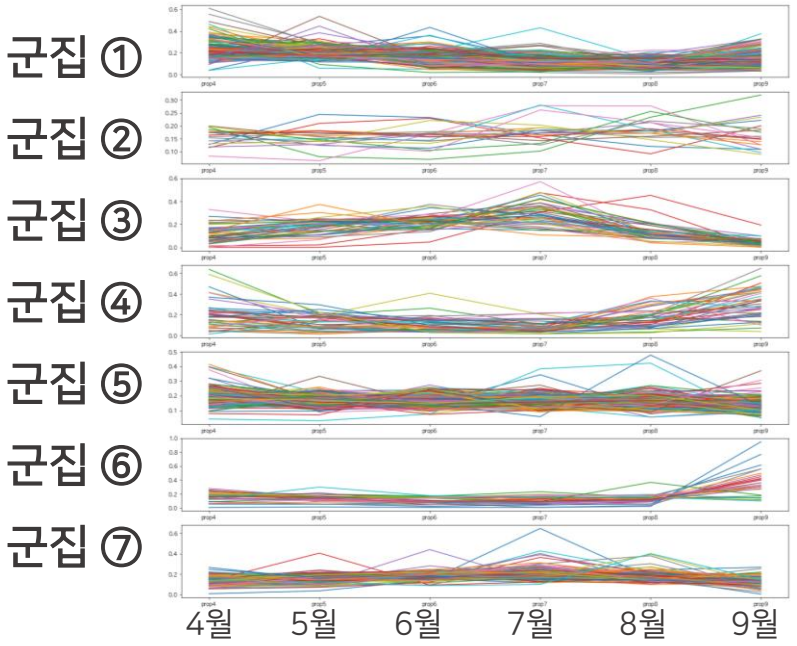
최종 군집(Cluster) 개수 = 7개

1.2 분석방향 수립

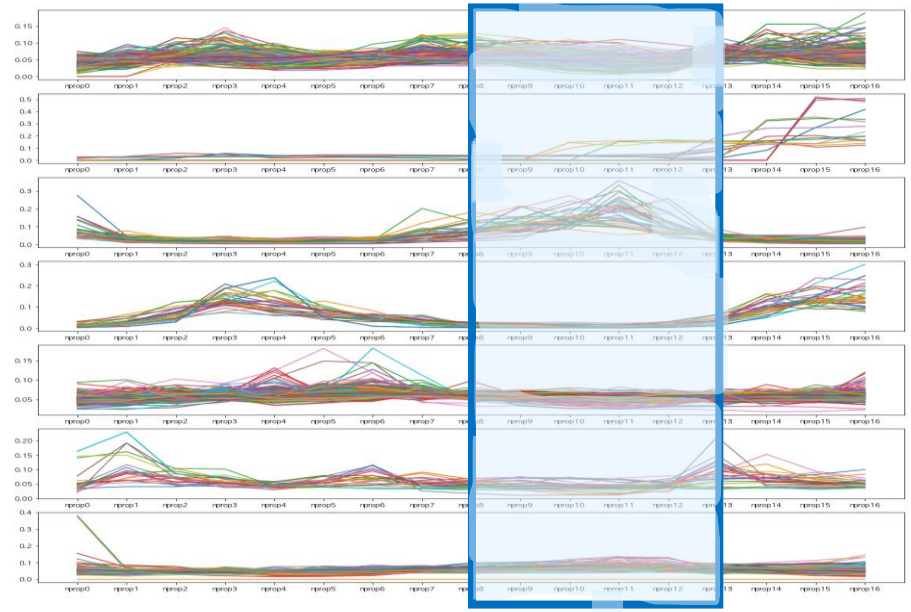
Step1. 모형 생성을 위한 군집화

3) 군집 결과

엘포인트 군집별 월별 판매량
(201년 4월 ~ 2017년 9월, **총 6개월**)



네이버 쇼핑인사이트 군집별 월별 클릭 수
(2017년 8월 ~ 2018년 12월, **총1년 4개월**)

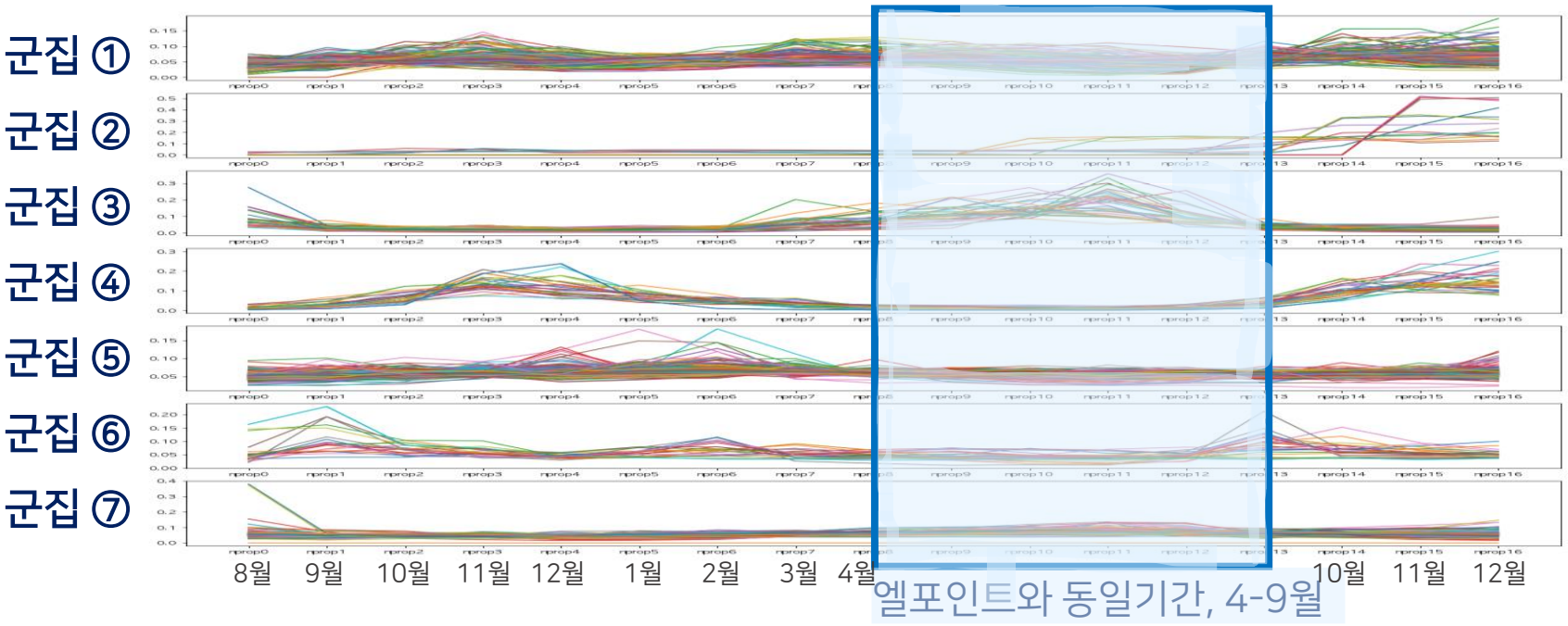


엘포인트와 동일기간, 4-9월

1.2 분석방향 수립

Step1. 모형 생성을 위한 군집화

네이버 쇼핑인사이트 군집별 월별 클릭 수
(2017년 8월 ~ 2018년 12월, **총1년 4개월**)



3) 군집 결과 및 해석

군집 ① → 대체적으로 11월, 4월, 겨울에 변동이 큼
군집 ② → 2018년 가을부터 판매량 상승
군집 ③ → 2-8월 상품군으로, 봄여름 상품군
군집 ④ → 10-2월 상품군으로, 늦가을부터 겨울 계절성 상품군
군집 ⑤, ⑦ → 대체적으로 변동이 적은 상품군
군집 ⑥ → 8,9월 , 2월에 증가하는 것으로 보아, 휴가철 상품군

1.2 분석방향 수립

Step 2. 모형 생성을 위한 Data 분할 및 회귀분석
Step 3. 훈련데이터를 바탕으로 검증 및 예측

1) 이분법

- 직전 2개월 데이터로 다음달(3개월 째 달) 판매량 예측

- 데이터 분할
  .TRAIN DATA: 4월+5월 데이터로 INPUT DATA 생성하고 6월 판매량으로 TARGET 생성
  .TEST DATA: 7월+8월 데이터로 INPUT DATA 생성하고 9월 판매량으로 TARGET 생성

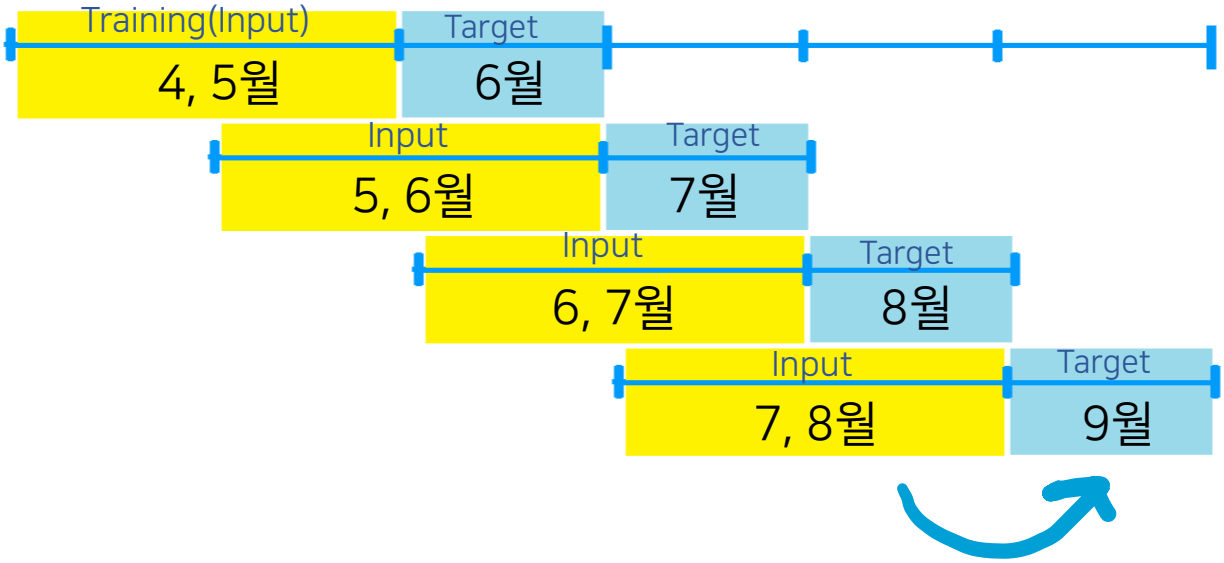
- 4~6월과 7~9월은 계절이 바뀌는 시기라서 생성한 모형을 평가하면 성능이 많이 저하될 것으로 예상

- 결과 확인 후 Test data 기준 모형의 성능이 만족스럽지 못한 경우 보완작업 실시
  (K-fold cross-validation 시도)

1.2 분석방향 수립

Step 2. 모형 생성을 위한 Data 분할 및 회귀분석
Step 3. 훈련데이터를 바탕으로 검증 및 예측

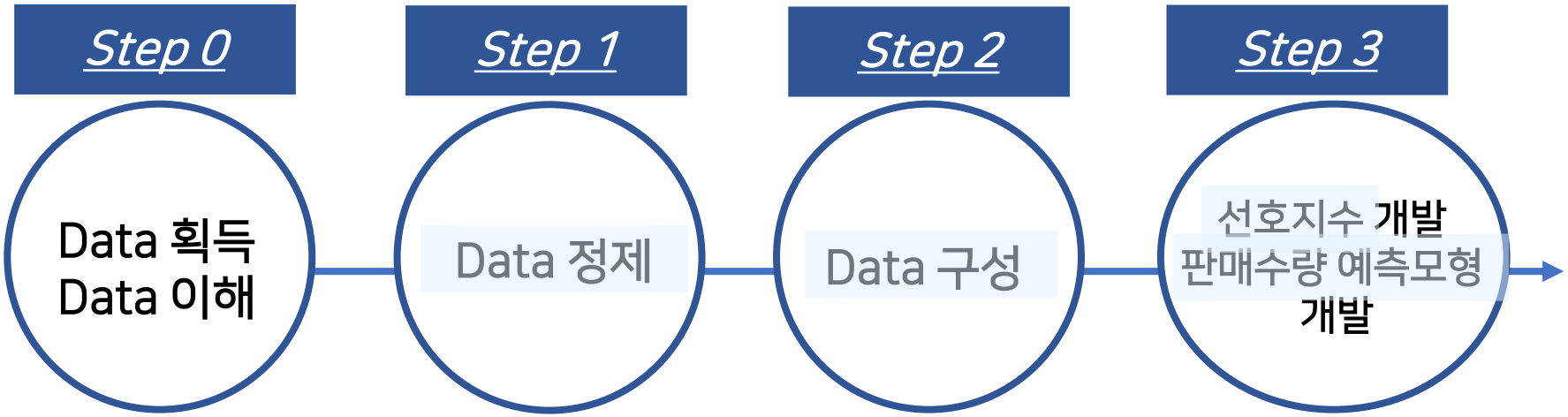
2) 4-fold cross validation 방식 (1안을 보완하는 방안)



- Dataset을 위와 같이 4개 구성 후 4-fold cross validation 실행 후 결과 확인하기
- 1) 에서 1개의 dataset으로 모형생성하고 또 다른 1개의 dataset으로 모형검증의 단점 보완 가능

II. 분석 프로세스

II. 분석 프로세스



1) 선호지수 개발용
2) 판매수량 예측모형 개발용

2.1 Data 획득 및 이해

- Dataset

Dataset = Table
괄호( ) 안은 UNIQUE 개수

Product	Master	Custom
상품구매	상품분류	회원
Session	**Search1**	**Search2**
세션	검색어1	검색어2

- 기준 변수에 따른 Table Join

회원(671,679개)
클라이언트ID(671679개)
성별
연령대

검색어1(2,884,943개)
클라이언트ID(511477개)
세션ID(1160589개)
검색키워드명
검색건수

세션(2,712,907개)
클라이언트ID(922737개)
세션ID(2425886개)
세션일련번호
세션일자
총페이지조회건수
총세션시간값
기기유형
지역대분류
지역중분류
지역소분류

상품구매(5,024,906개)
클라이언트ID(922737개)
세션ID(2425886개)
히트일련번호
상품코드(847652개)
상품추가정보
상품브랜드
단일상품금액
구매건수

상품분류(847,652개)
상품코드(847652개)
상품명
상품 대분류명
상품 중분류명
상품 소분류명

검색어2(8,051,172 개)
세션일자
검색키워드명
검색건수

2.1 Data 획득 및 이해

- 기준 변수에 따른 table join

① **Product** 상품구매 — 기준 변수 PD_C — **Master** 상품분류

② **Session** 세션 — 기준 변수 CLINT_ID SESS_ID — **Search1** 검색어1

③ **Product** 상품구매 / **Master** 상품분류 — 기준 변수 CLINT_ID SESS_ID — **Session** 세션 / **Search1** 검색어1

④ **Custom** 회원 — 기준변수 CLINT_ID — **Product** 상품구매 / **Master** 상품분류 — **Session** 세션 / **Search1** 검색어1

2.1 Data 획득 및 이해

전체 분석과정에서 적용한 가정

구매 행위
에 대한 정의

한 클라이언트가 같은 세션 안에서 히트 시퀀스가 같으면
= CLNT_ID + SESS_ID + SESS_SEQ + HITS_SEQ
를 구매 행위로 정의함. 즉, 히트 시퀀스가 달라지면
두번째 개별적인 구매행위로 간주함.

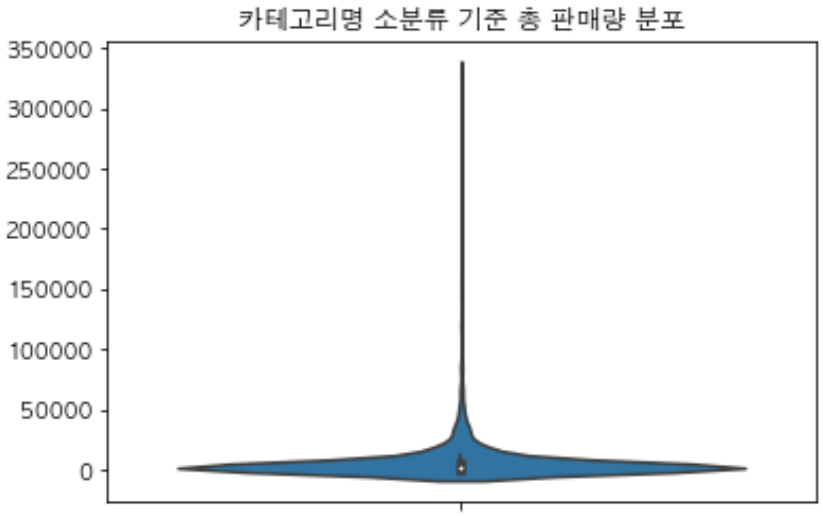
히트 (HITS)
에 대한 정의

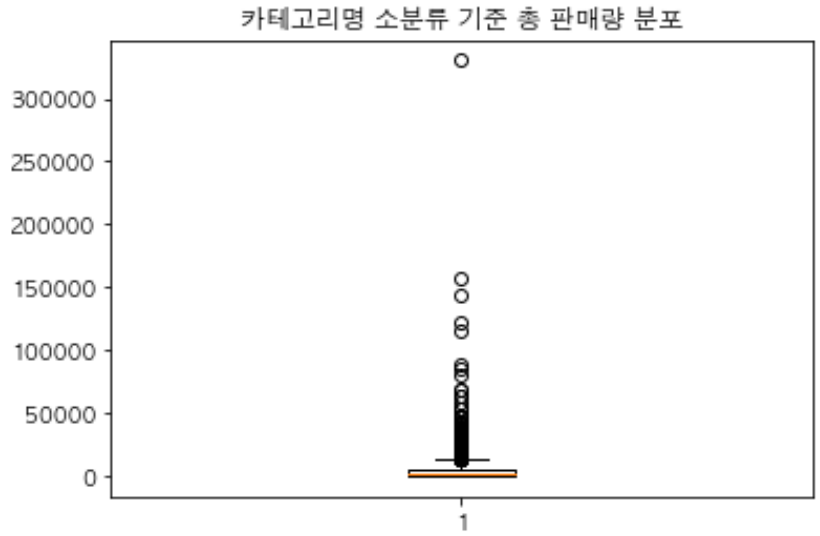
한 구매 당 카테고리별로 평균 히트 시퀀스
이를 통해, 상품군마다 구입에 이르기까지
얼마나 고민했는지 알 수 있음

※ 구입과 상관없는 행위도 포함되기 때문에 Total page view, Total hour를 활용하지 않음
※ 선호지수의 경우 판매량, 검색량, 평균히트수 순으로 중요도를 정의함

2.1 Data 획득 및 이해

① 기술통계 : 총 판매량





총 판매량의 분포가 왼쪽의 그래프와 같은 분포를 이루므로, 선호지수의 상·중 ·하의 기준을 아래와 같이 정의한다.

	TOT_SALES
count	898.00
mean	6590.74
std	17413.85
min	2.00
25%	413.25
50%	1781.00
75%	5533.00
max	330276.00

선호지수	
상	75%~100%
중	50%~75%
하	0~50%

2.1 Data 획득 및 이해

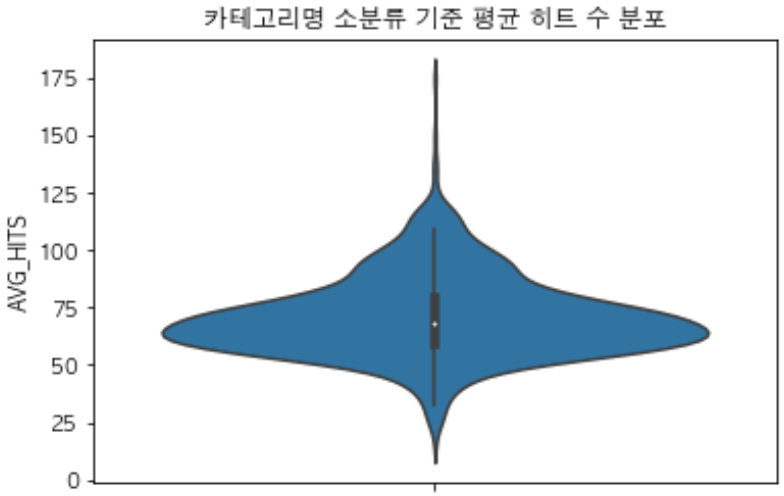
① 기술통계 : 총 판매량

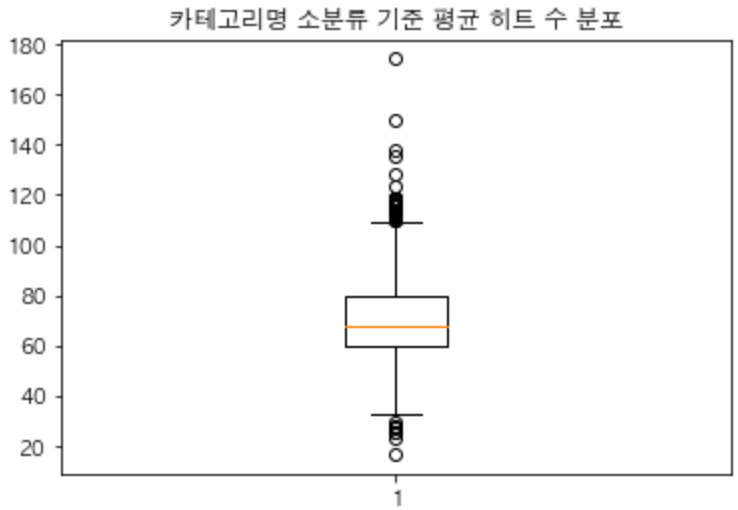
총 판매량이 높은 상위 10개 카테고리

	CLAC3_NM	CLAC2_NM	CLAC1_NM	TOT_SALES
184	남성티셔츠	남성의류상의	남성의류	330276
552	여성원피스	여성의류전신	여성의류	157240
573	여성티셔츠/탑	여성의류상의	여성의류	144633
500	여성남방셔츠	여성의류상의	여성의류	122838
176	남성캐주얼바지	남성의류하의	남성의류	115926
347	블라인드/버티컬	커튼/블라인드류	인테리어/조명	88429
515	여성바지	여성의류하의	여성의류	84845
412	스킨케어세트	스킨케어	화장품/뷰티케어	80956
350	블러셔/쉐이딩/하이라이터	메이크업	화장품/뷰티케어	69472
2	BB/파운데이션/컴팩트류	메이크업	화장품/뷰티케어	69417

2.1 Data 획득 및 이해

② 기술통계 : 평균 히트





	AVG_HITS
count	898.00
mean	70.96
std	17.82
min	16.75
25%	59.48
50%	68.03
75%	79.51
max	174.12

평균 히트의 분포가 왼쪽의 그래프와 같은 분포를 이루므로, 선호지수의 상·중 ·하의 기준을 아래와 같이 정의한다.

선호지수	
상	75%~100%
중	50%~75%
하	0~50%

2.1 Data 획득 및 이해

② 기술통계 : 평균 히트

평균 히트수가 작은 하위 10개 카테고리

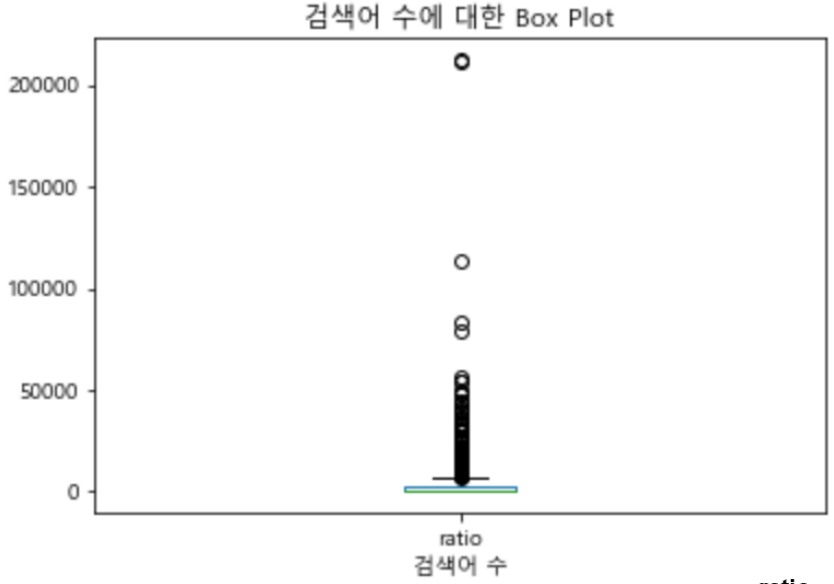
	CLAC3_NM	CLAC2_NM	CLAC1_NM	AVG_HITS
160	남성양말선물세트	남성양말류	속옷/양말/홈웨어	33.93
250	레저모바일상품권	모바일상품권	상품권	32.61
298	미용거울	미용소품	화장품/뷰티케어	29.20
358	산림욕기	공기청정/가습/제습	생활/주방가전	28.00
309	반죽기/제면기	주방가전	생활/주방가전	27.67
623	영화/문화모바일상품권	모바일상품권	상품권	27.04
727	자연/과학완구	교육완구	완구	25.89
251	로만쉐이드/벌룬쉐이드	커튼/블라인드류	인테리어/조명	25.33
4	DIY완구	여아완구	완구	23.00
46	과실주병	밀폐/보관용기	식기/조리기구	16.75

평균 히트수가 많은 하위 10개 카테고리

	CLAC3_NM	CLAC2_NM	CLAC1_NM	AVG_HITS
231	닭안심	닭고기류	축산물	174.12
121	남성등산스웨트셔츠/후드/집업	남성등산/아웃도어의류	스포츠패션	149.67
359	살구	국산과일	과일	138.00
289	문구세트	일반문구/사무용품	문구/사무용품	135.42
87	기타유아동양말류	유아동양말류	속옷/양말/홈웨어	128.11
147	남성스포츠베스트	남성일반스포츠의류	스포츠패션	123.19
581	여아남방셔츠	여아의류상의	유아동의류	118.89
613	영유아스웨트셔츠/후드/집업	유아의류상의	유아동의류	118.85
586	여아스웨터/풀오버	여아의류상의	유아동의류	118.55
68	기타남성화	남성화	패션잡화	118.50

2.1 Data 획득 및 이해

③ 기술통계 : 검색어 수



	ratio
count	898.000000
mean	4209.568608
std	13441.425073
min	0.000574
25%	169.655909
50%	681.699953
75%	2989.421622
max	212790.962979

검색어 수의 분포가 왼쪽의 그래프와 같은 분포를 이루므로, 선호지수의 상·중 ·하의 기준을 아래와 같이 정의한다.

선호지수	
상	75%~100%
중	50%~75%
하	0~50%

2.2 Data 정제

- 각 table에서의 수행했던 기본적인 정제작업 내용

① Product + Master 테이블

1) 브랜드명 전처리
Product 테이블의 ' 상품브랜드(PD_BRA_NM) ' 명 전처리 [] 키워드 제거함.

2) 상품 오분류 관련
여성재킷이 남성의류로 분류 되어있는 등의 오분류를 인식함. 이를 제대로 분류할 경우, 모든 카테고리에 일괄 적용해야 하는데, 오분류 및 정분류의 판단을 내리기엔 한계가 존재했기에 제공받은 데이터대로 분석을 진행함.

② Search 1 테이블

- 롱 포맷(long Format)을 와이드 포맷(wide Format)으로 변환함.

③ Custom 테이블

- 연령 60대, 70대, 80대는 '60대 이상'으로 하나로 묶음.

④ Session 테이블

- city zone 16개를 6개의 지역으로 묶음. (수도권, 전라도, 경상도, 충청도, 강원도, 제주도)
ex) 서울/인천/경기도를 '수도권'으로 묶음. 충청북도, 충청남도를 '충청도'로 묶음.

2.3 Data 구성

1) 선호지수 개발용 분석 Data 구성 프로세스

Step 1

상품군 소분류 기준
네트워크 분석
Network

Step 2

상품군 소분류 기준
세그멘테이션

H H L
H H L
L L L
H M M
L H L
…

상품군별 선호지수 개발

2.3 Data 구성

1) 선호지수 개발용 분석 dataset 구성

① 상품군 소분류 기준 네트워크 Network 분석

상품 네트워크 데이터 구성

CLAC3	SESS_ID	SESS_SEQ	HITS_SEQ	PD_C	PD_ADD_NM	PD_BUY_CT
156	7112398	181	231	218347	네이비	1
156	7112398	181	231	218347	블랙	1

구매 건당 상품의 거래를 나타내는 네트워크

네트워크 분석은 개별 구성인자들 사이에 형성되고 있는
관계적 속성을 분석의 대상으로 삼고 있다.

✓ 가장 자주 구매된 상품ID 쌍
상품코드 207981 제품과 208785 제품이 640번 같이 구매됐다.

✓ 상품 ID를 노드로 한 네트워크 → 상품군 네트워크
- node = 상품군 소분류 ID
- edge = 함께 구매된 상품군 거래
- weight = 같이 구매된 구매 건수

	node1	node2	count
763614	207981.0	208785.0	640
911367	242021.0	242022.0	466
2544236	737743.0	737745.0	274
2645361	809108.0	833515.0	269
1986925	495506.0	537192.0	245

가장 많이 구매된 상품ID 쌍

2.3 Data 구성

1) 선호지수 개발용 분석 dataset 구성

① 상품군 소분류 기준 네트워크 Network 분석

	Unnamed: 0	BC
34	기타일반문구/사무용품	0.022957
198	남성티셔츠	0.018084
22	여성티셔츠/탑	0.013752
487	여성원피스	0.013716
501	남성캐주얼바지	0.010408
509	여성남방셔츠	0.010019
47	BB/파운데이션/컴팩트류	0.009904
168	여성바지	0.009411
157	스킨케어세트	0.008669
6	남성팬티	0.008061

✓ 이 네트워크에서 Betweenness Centrality 로 상품군간의 영향도를 지수로 산출
✓ 상품 거래의 영향도를 나타내므로, 판매량이 아닌 색다른 관점의 영향도 산출

"**기타일반문구/사무용품**이 다른 상품군간에 영향력이 가장 크다.
나머진 대부분 **의류**가 함께 구매된다."

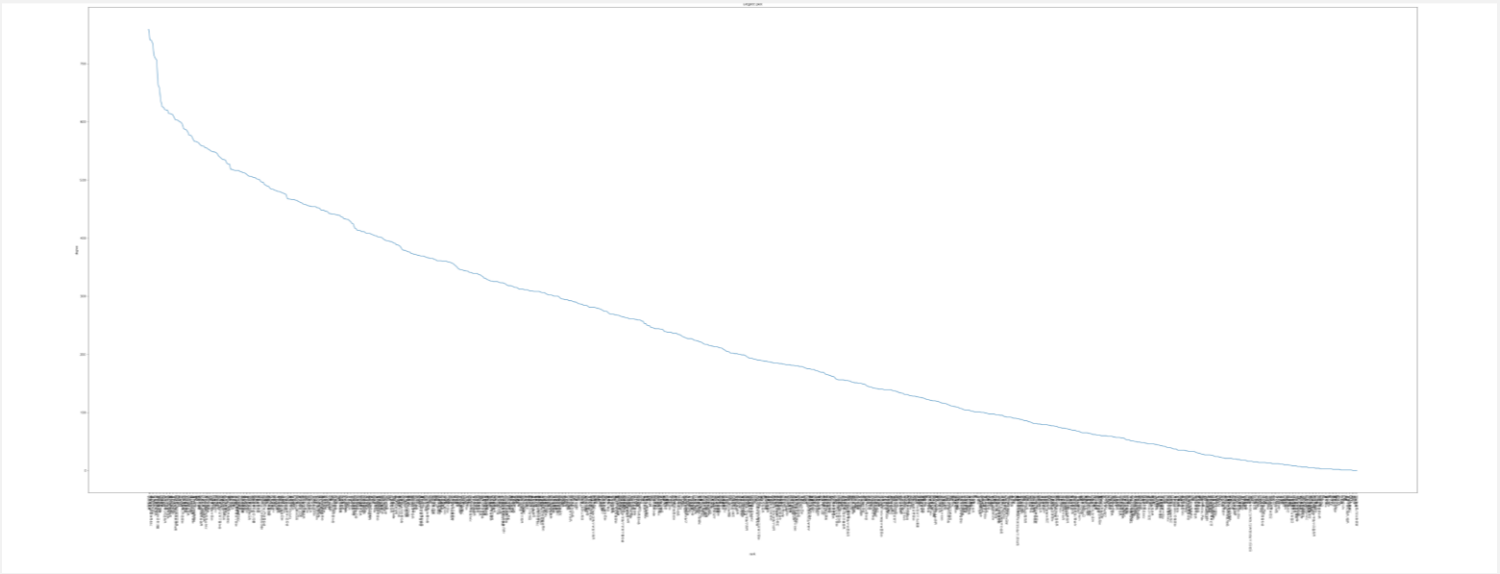
※ GEPHI 툴을 활용

2.3 Data 구성

1) 선호지수 개발용 분석 dataset 구성

① 상품군 소분류 기준 네트워크 Network 분석

상품군별 degree 분포 그래프



- ✓ 상품군별 degree 분포 그래프의 Average Degree = 228.759
- ✓ connected component 개수 =8
  8개 밖에 없으므로 굉장히 응집되어 있는 케이스

2.3 Data 구성

1) 선호지수 개발용 분석 Data 구성

② 상품군 소분류 기준 세그멘테이션(segmentation)

※ 선호지수의 경우 판매량, 검색량, 평균히트수 순으로 중요도를 정의함

CLAC3	판매량	검색량	평균히트수
소분류명1	상	중	중
소분류명2	중	상	중
소분류명3	중	중	하
소분류명4	하	중	하
소분류명5	상	상	상
.	.	.	.
.	.	.	.
.	.	.	.
898개			

상품군별 선호지수 도출

2.3 Data 구성

1) 판매수량 예측모형 개발용 분석 Data 구성 프로세스

Step 1

상품군 소분류 기준
군집분석
Clustering

Step 2

상품군 소분류 기준
회귀분석
regression

판매수량 예측모형 개발

외부데이터

NAVER DataLab.

검색어트렌드 쇼핑인사이트

API를 통해 데이터를 수집했으며,
우리 Dataset의 상품군과 매칭시킴

네이버 쇼핑인사이트 Data,
기상청 기온, 미세먼지 data

2.3 Data 구성

2) 판매수량 예측모형 개발용 분석 Data구성

① 상품군 소분류 기준 군집분석

앞서 군집 분석을 통해 도출한 7개의 군집 → 군집별로 모델 적용

한 모델별 216개 컬럼

CLAC3	HITS, 판매량, 여성, 남성(성별) 10대여성, 10대남성, 20대여성, 30대남성 (성별 + 연령조합) … 수도권 충청도 경상도(지역별) … 4월, 5월, 4·5월(월별) … 등의 조합 경우의 수와 지역별 월평균 날씨(기온, 미세먼지), 당월의 네이버 카테고리 클릭 수
소분류명1	
소분류명2	
소분류명3	
소분류명4	
소분류명5	
…	

2.4 선호지수 산정1

첫번째, 선호 지수 산정

상품군별 소분류
총 판매량

X 5

상품군별 소분류
총 검색량

X 3

상품군별 소분류
구매 연관성

X 2

선호지수

2.4 선호지수 산정1

두번째, 선호 지수 산정

	TOT_SALES	AVG_HITS	ratio
count	898.000000	898.000000	898.000000
mean	6590.744989	70.963196	4209.568608
std	17413.845231	17.823578	13441.425073
min	2.000000	16.750000	0.000574
25%	413.250000	59.485000	169.655909
50%	1781.000000	68.035000	681.699953
75%	5533.000000	79.512500	2989.421622
max	330276.000000	174.120000	212790.962979

① LOW

0% ~ 50%

② MIDDLE

50% ~ 75%

③ HIGH

75% ~ 100%

2.4 선호지수 산정1

첫번째, 선호지수산정 결과

선호지수산정 결과,
선호 지수 **상위 5개의 상품군**을 살펴보면,
가장 기본적인 **'의류 아이템'**이 많이 선호됨

CLAC3_NM	PREFRENC_IDX
남성티셔츠	158.707675
여성원피스	104.424890
여성티셔츠/탑	78.781269
남성캐주얼바지	59.857901
여성남방셔츠	55.524373
여성바지	48.823582

2.4 선호지수 산정2 (총 판매량 / 총 검색량 / 평균 히트 수)

(총판매량, 총 검색량, 평균 히트 수)

1) (H, H, H)

CLAC3_NM	TOT_SALES_LEVEL	TOT_SEARCH_LEVEL	AVG_HITS_LEVEL
남성티셔츠	H	H	H
여성원피스	H	H	H
여성티셔츠/탑	H	H	H
남성캐주얼바지	H	H	H
여성남방셔츠	H	H	H
여성바지	H	H	H
스킨케어세트	H	H	H
영유아티셔츠/탑	H	H	H

[판매량 / 검색량 / 히트 수가 모두 높은 상품군]
→ 가장 **기본적인 의류 상품군** (티셔츠, 바지, 셔츠) 등으로 구성됨.

2) (L, L, L)

드럼세탁기	L	L	L
탁자	L	L	L
스탠드형김치냉장고	L	L	L
시공가구	L	L	L
LED	L	L	L
바디슬리밍/리프팅	L	L	L
데오도란트	L	L	L
스피커	L	L	L
기타냉방가전	L	L	L
장식장/진열	L	L	L

[판매량 / 검색량 / 히트수가 모두 저조한 상품군]
→ '**전자제품**'이나 '**가구**' 등 과 같이 **부피가 크고 가격이 높은 제품**들로 구성됨

2.4 선호지수 산정2 (총 판매량 / 총 검색량 / 평균 히트 수)

(총 판매량, 총 검색량, 평균 히트 수)

3) (H, H, L)

BB/파운데이션/컴팩트류	H	H	L
블러셔/쉐이딩/하이라이터	H	H	L
페이셜클렌저	H	H	L
크림/밤/오일	H	H	L
선크림류	H	H	L
생수	H	H	L
페이셜팩류	H	H	L

[판매량과 검색량이 많고 히트가 적은 제품군]
→ 즉, 빠르게 검색해 구매되는 제품들은 대체적으로 '화장품'들로 구성됨.

4) (H, L, H)

```
mg4[(mg4["TOT_SALES_LEVEL"] == "H") & (mg4["TOT_SEARCH_LEVEL"] == "L") & (mg4["AVG_HITS_LEVEL"] == "H")]
```

CLAC3_NM	TOT_SALES_LEVEL	TOT_SEARCH_LEVEL	AVG_HITS_LEVEL	TOT_SALES	AVG_HITS	TOT_SEARCH_CNT	BC	SCALE_TOT_SALES

[판매량과 히트수는 많으나 검색량이 적은 제품군]
→ 이러한 경우는 존재하지 않았다.
→ 이를 통해, 검색량은 총 판매량과 평균 히트 수와 연관이 있음을 알 수 있다.

2.4 선호지수 산정2 (총 판매량 / 총 검색량 / 평균 히트 수)

(총 판매량, 총 검색량, 평균 히트 수)

5) (L, H, H)

CLAC3_NM	TOT_SALES_LEVEL	TOT_SEARCH_LEVEL	AVG_HITS_LEVEL
유아동속옷 세트	L	H	H
소품가방	L	H	H
여성레깅스	L	H	H
영유아재킷	L	H	H
여성스포츠 점퍼/재킷	L	H	H
영유아레깅스	L	H	H
영유아패딩	L	H	H

[판매량은 적고 검색수와 히트수는 높은 제품군]
→ 주로 **'유아용제품'** 이나 **'환불 불가 상품'** (속옷, **레깅스류**)들로 구성됨.

6) (H, H, M)

CLAC3_NM	TOT_SALES_LEVEL	TOT_SEARCH_LEVEL	AVG_HITS_LEVEL
남성런닝/트레이닝화	H	H	M
남성팬티	H	H	M
여성런닝/트레이닝화	H	H	M
남성스포츠 샌들/슬리퍼	H	H	M
여성샌들	H	H	M
브래지어	H	H	M
스포츠가방	H	H	M
유아동샌들	H	H	M

[판매량과 검색량이 많고 히트수 보통인 제품군]
→ **'스포츠상품'**이나 **'속옷'** 등과 같이 일상 생활에서 많이 사용되는 **생활용품**들로 구성됨.

2.4 선호지수 산정2 (총 판매량 / 총 검색량 / 평균 히트 수)

(총 판매량, 총 검색량, 평균 히트 수)

7) (L, H, L)

CLAC3_NM	TOT_SALES_LEVEL	TOT_SEARCH_LEVEL	AVG_HITS_LEVEL
벽걸이형에어컨	L	H	L
남성스포츠속옷	L	H	L
스탠드형에어컨	L	H	L
호일/랩/기름종이	L	H	L
기타에어컨	L	H	L
기저귀크림/파우더	L	H	L
이불/옷커버류	L	H	L

[검색량은 많으나 판매량과 히트수가 적은 제품군]
→ 주로 '**시즌상품(에어컨)**' 으로 구성됨.

8) (L, L, M)

여성골프베스트	L	L	H
남아가디건	L	L	H
머플러	L	L	H
넥워머	L	L	H
여성골프패딩	L	L	H
여아스웨터/풀오버	L	L	H
남성골프니트/가디건	L	L	H
여아코트	L	L	H

[높은 히트수이나 판매량과 검색량이 적은 제품군]
→ 주로 **등산, 골프**와 같이 **스포츠**에 관련된 상품들로 구성되어 있고 **계절성**이 있는 **'방한화', ' 머플러', '넥워머', '코트'** 등으로 구성됨

2.4 선호지수 산정2 (총 판매량 / 총 검색량 / 평균 히트 수)

(총 판매량, 총 검색량, 평균 히트 수)

9) (H, M, M)

CLAC3_NM	TOT_SALES_LEVEL	TOT_SEARCH_LEVEL	AVG_HITS_LEVEL
기타일반문구/사무용품	H	M	M
블라인드/버티컬	H	M	M
여성덧신류	H	M	M
기타조리도구	H	M	M
욕실소품	H	M	M
성인침구세트	H	M	M
성인패드/스프레드	H	M	M

[판매량이 많고 검색량과 히트수는 보통인 경우]
→ 일반 문구나 조리도구, 욕실도구 등의 '**생활용품**'으로 주로 구성된다.

10) (L, M, L)

오븐/전자레인지	L	M	L
마우스	L	M	L
전기면도기	L	M	L
식탁의자	L	M	L
성인요/요커버	L	M	L
일반세탁기	L	M	L
스포츠두건/머플러/마스크	L	M	L
멀티형에어컨	L	M	L

[판매량과 히트수는 작고 검색량은 보통인 제품군]
→ '**전자제품**'이나 '**가구**'와 같이 부피가 크거나 가격이 높은 제품들로 구성됨.

2.5 판매수량 예측모델 개발

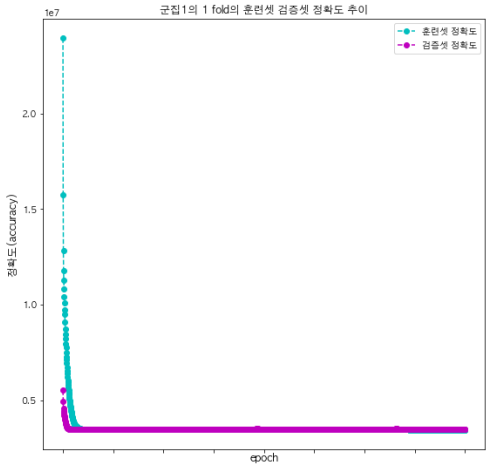
신경망

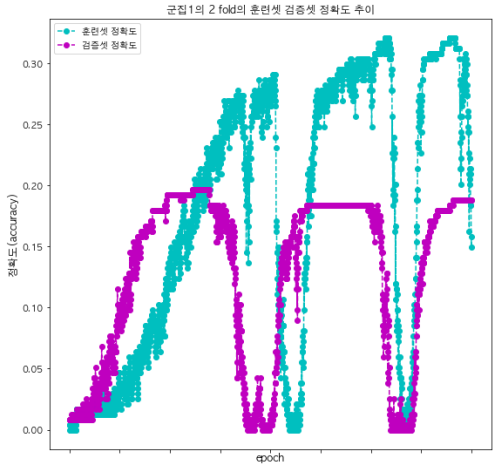
- 입력 = 직전 2개월의 data

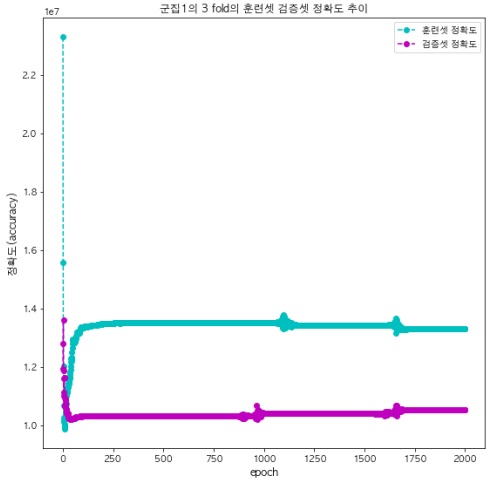
- relu함수사용 — tanh 함수는 sigmoid 함수의 문제점인 모형 훈련 지연을 해결 가능하고
- relu는 tanh와 비슷한 성능 가지면서 빠르게 수렴하기 때문에 relu선정

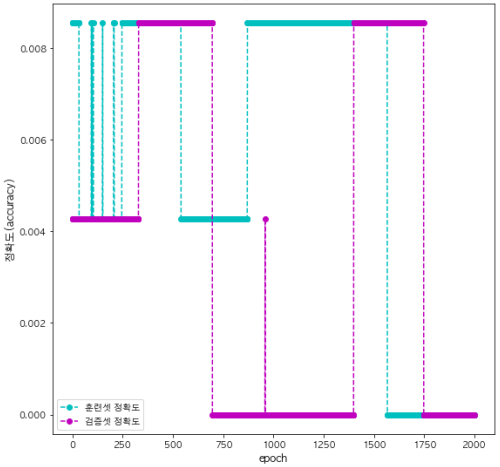
2.5 판매수량 예측모델 개발

군집 1의 fold별 딥러닝학습 훈련셋 검증셋 정확도 추이



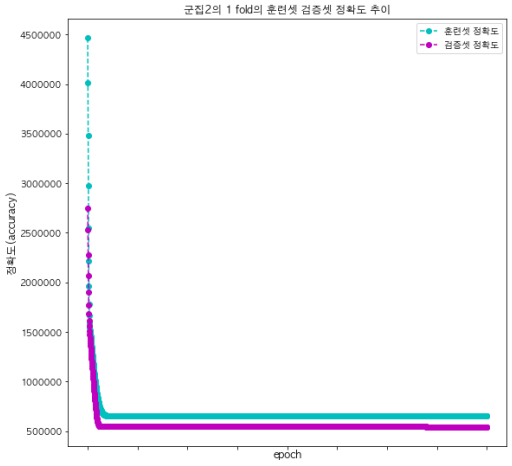


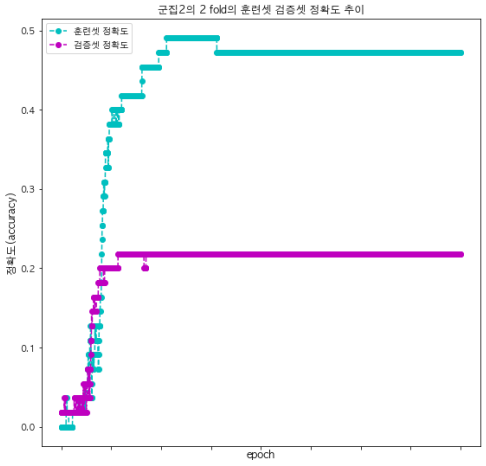


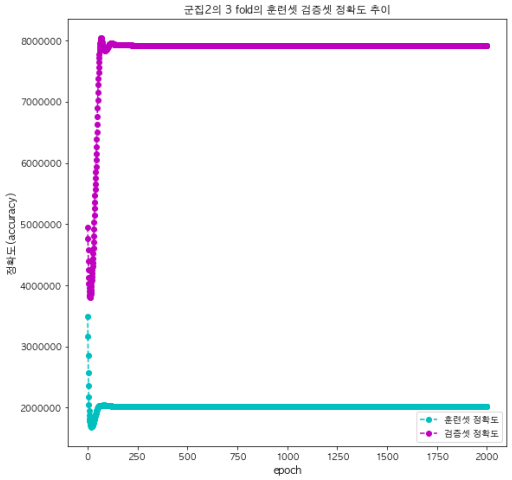


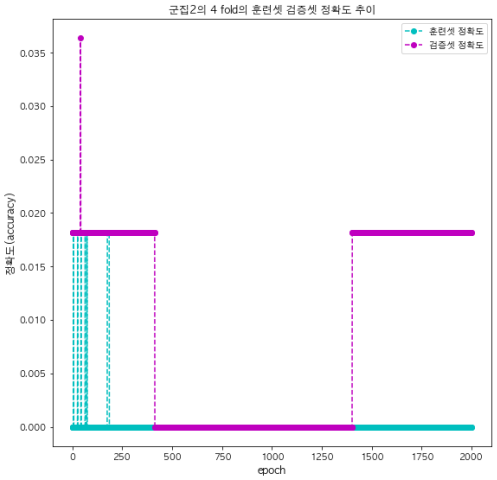
2.5 판매수량 예측모델 개발

군집 2의 fold별 딥러닝학습 훈련셋 검증셋 정확도 추이



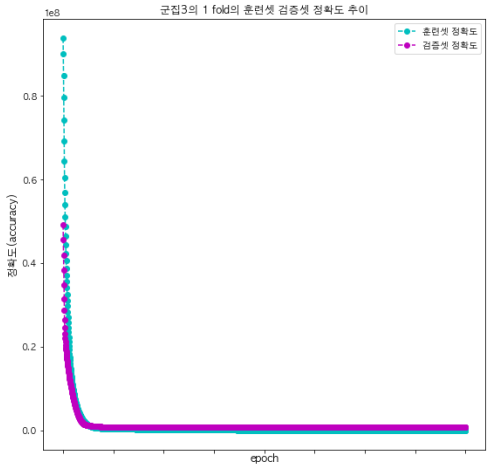


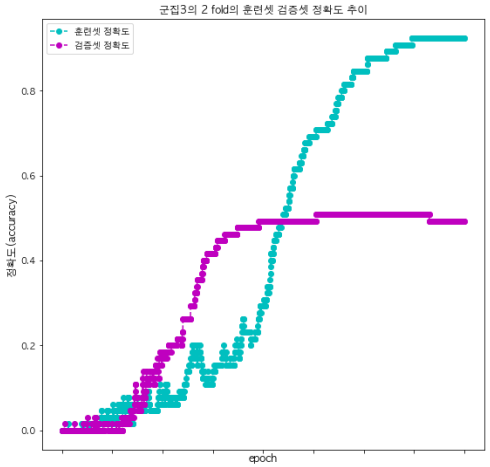


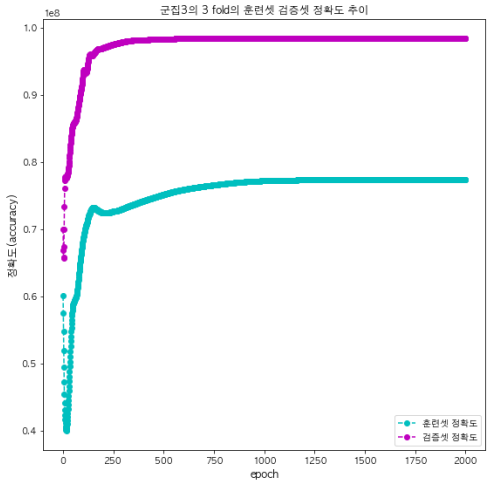


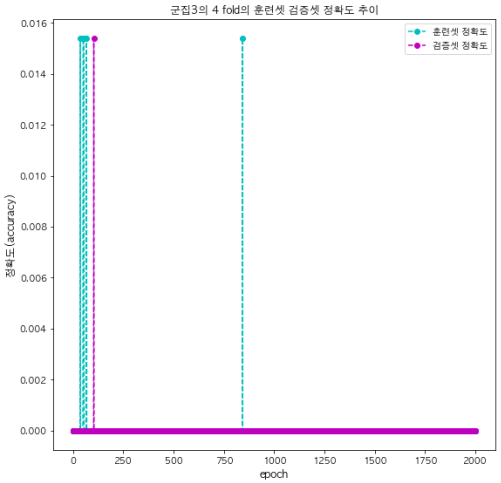
2.5 판매수량 예측모델 개발

군집 3의 fold별 딥러닝학습 훈련셋 검증셋 정확도 추이



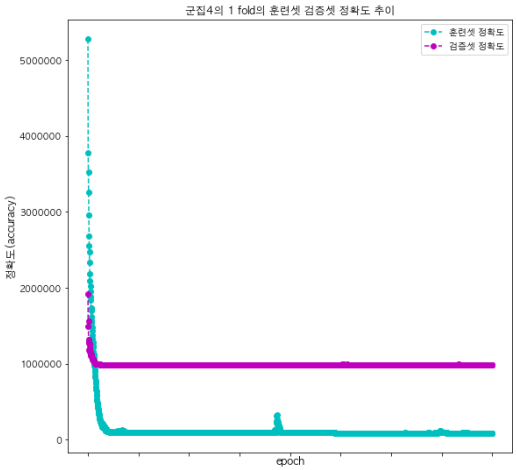


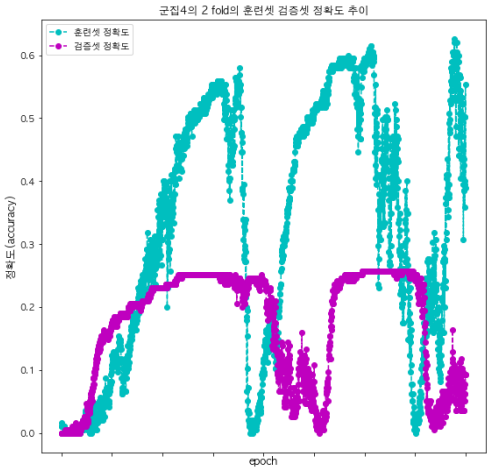


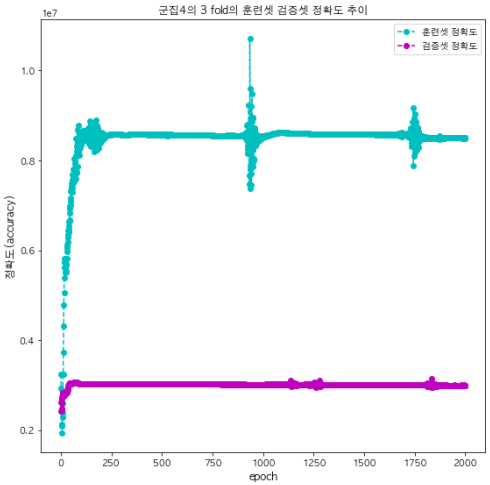


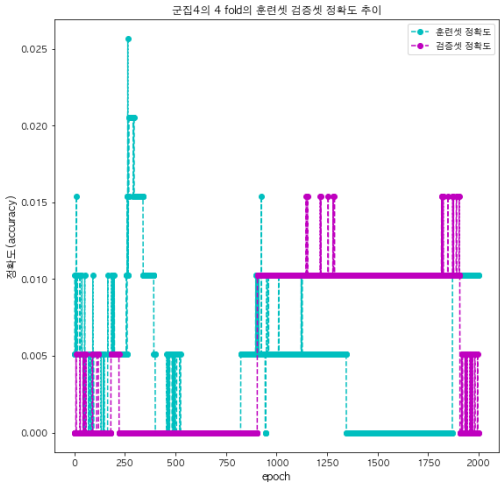
2.5 판매수량 예측모델 개발

군집 4의 fold별 딥러닝학습 훈련셋 검증셋 정확도 추이



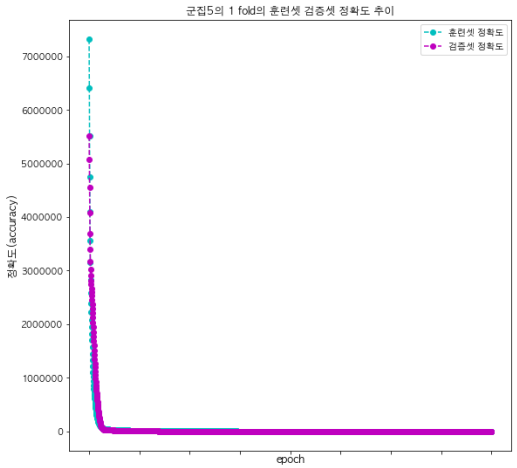


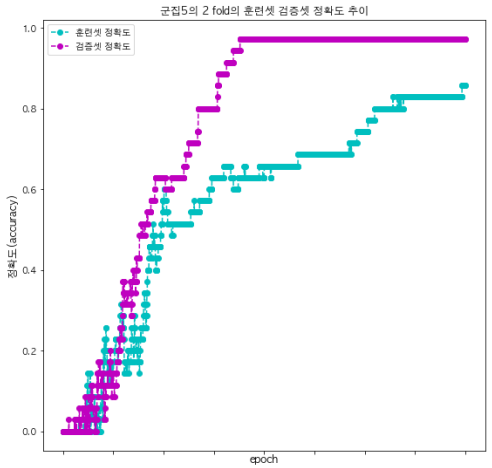


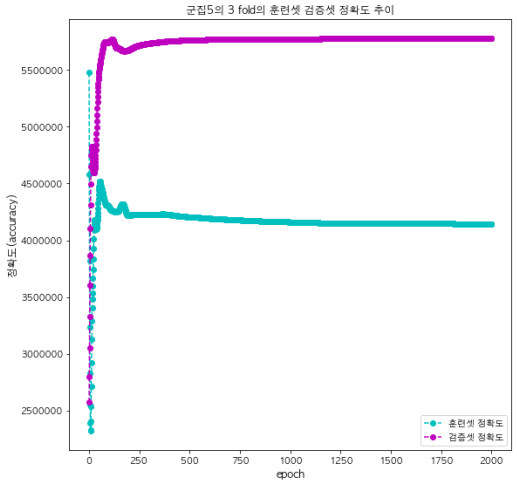


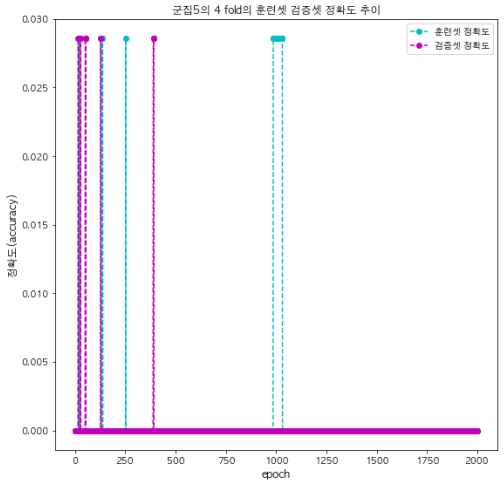
2.5 판매수량 예측모델 개발

군집 5의 fold별 딥러닝학습 훈련셋 검증셋 정확도 추이



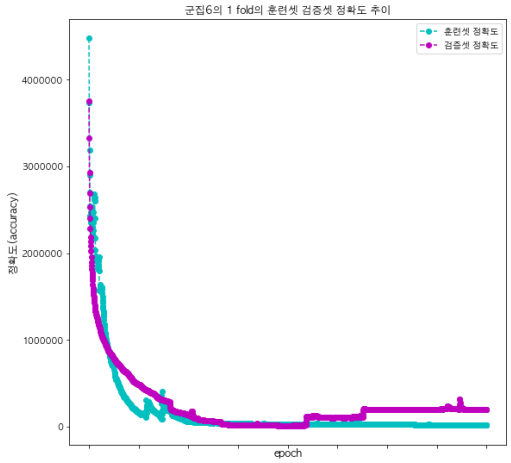


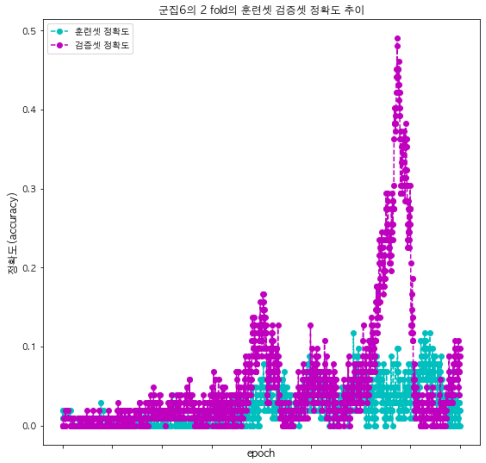


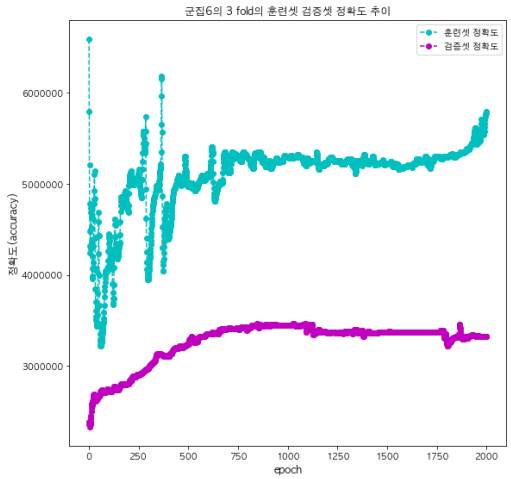


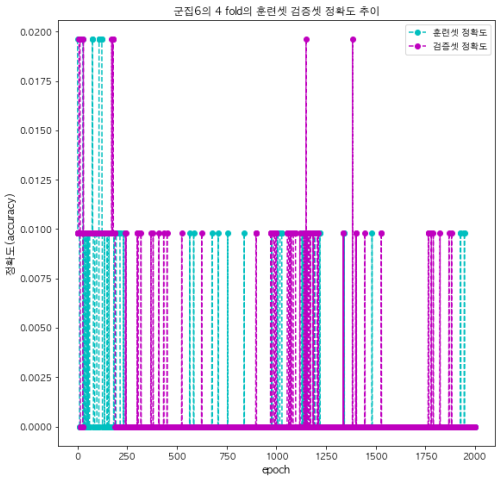
2.5 판매수량 예측모델 개발

군집 6의 fold별 딥러닝학습 훈련셋 검증셋 정확도 추이



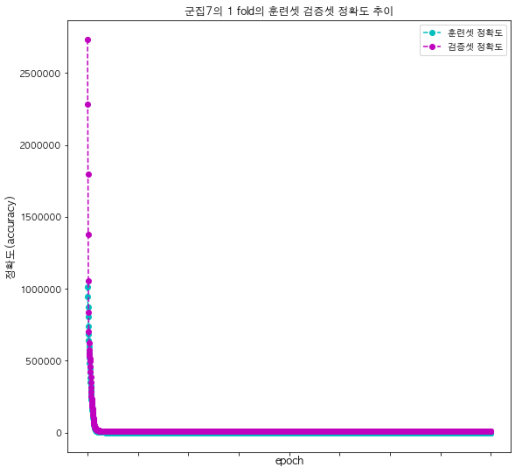


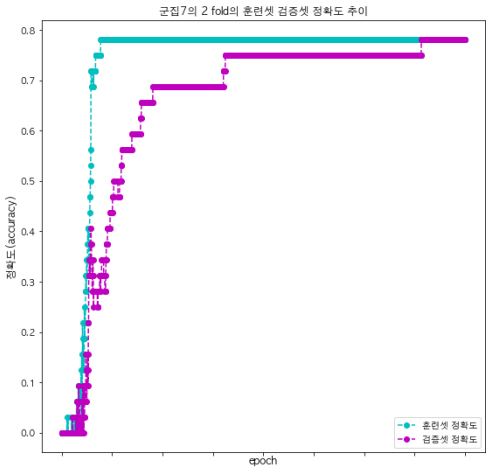


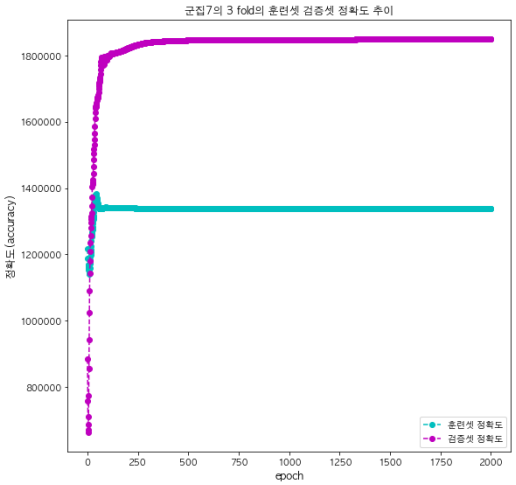


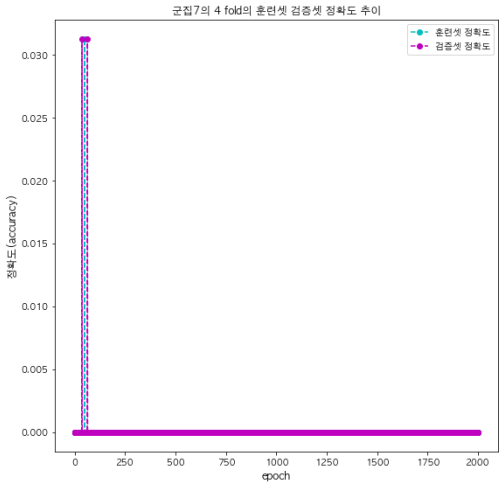
2.5 판매수량 예측모델 개발

군집 7의 fold별 딥러닝학습 훈련셋 검증셋 정확도 추이









III. 분석결과 활용 IDEA 제안

3.1 선호지수 활용방안

개발한 선호지수 분포 및 결과
프로파일 분석에서 찾아진
Insight를 기반으로 활용방안 도출

화장품 등
"사는 건 더 많이 사게끔"

H H L
더 많은 구매 유도 가능한 상품

< 방안 >
화장품 소분류 내에서 얼마 이상 구입 시 할인 쿠폰/화장품 샘플 브랜드 선택권 부여. COSMETIC DAY등 이벤트 날에 특정 브랜드 세일 및 메이크업 서비스 쿠폰 제공 등

레깅스, 영유아 의류 등
"간접 체험의 기회 부여로 신뢰성 높여야"

L H H
관심은 많은데 구매로 이어지지 않는 상품

< 방안 >
상세페이지 강화, 재질 표현력 다양화, 제품 zoom in 사진 수 증대, 영유아 의류 상품의 경우 주부들이 주로 쓰는 생필품 사은품을 추가로 제공

계절성이 담긴 방한화, 머플러 등
"고객이 상품을 비교하는 경우의 수를 줄여주자"

L L H
상품간 비교를 많이 하는 상품

< 방안 >
평균히트수가 높은 것으로 보아 상품간 비교를 많이 하는 것으로 판단됨. LLH 상품을 검색한 사람들에게는 Best 판매량, Best 평점인 제품을 우선적으로 노출시킴

3.2 Lesson Learned

본 경시대회 참여를 통해 학습한 사항들

예측관련	주어진 6개월의 한정된 data로 판매수량 예측의 어려움 → 이분법, K-fold 방식 적용
한글 텍스트 데이터 정제	인간의 생각의 흐름대로 표현된 검색어(비정형데이터)를 정제 및 분류하는 어려움, 물리적 시간과 다양한 조합의 로직과 패키지가 요구됨 Fuzzywuzzy[speedup]: 동일한 token 체크 ast : string을 딕셔너리 형태로 변경 CategoricalDtype : 카테고리형으로 변경
상품분류 카테고리 정의	상품분류 카테고리 정의 및 해석의 어려움 → 군집, 네트워크, 회귀 등 다양한 분석법 학습

3.3 향후 개선방안

본 분석에서 시도하지는 못했지만 향후 시도해보고 싶었던 사항

고객 세그먼테이션	고객데이터와 상품군별 선호지수를 접목하여 발전된 고객 세분화(세그멘테이션) 모델을 개발해보고 싶다.
행동정보 로그 탐색	데이터가 6개월만 제공 되었는데 좀 더 충분한 양의 데이터에서 행동정보 로그를 통해 고객 행동 기반을 더 자세히 분석해보고 싶다.

IV. 첨부

Ⅳ. 첨부

1. 전체 분석과정에서 적용한 **정제 로직**

2. 전체 분석과정에서 기타 **참고한 내용**
   - 온라인 선호지수 리서치
   - 예측모델 리서치

1. 전체 분석과정에서 적용한 정제 로직

온라인 선호지수

1. 소분류명

소분류명으로 분류 가능한 데이터 분류

키워드명	소분류명
남성수영복	남성수영복
여성가디건	여성가디건

2. 브랜드명 + 소분류명

브랜드명 + 소분류명으로 붙여서 분류

키워드명	소분류명
헤지스남성티셔츠	남성티셔츠
머렐 등산화	등산화

3. 전처리한 소분류명

'/'을 나누어서 데이터 분류
'류' 제거 전처리

키워드명	소분류명
옷커버	이불/옷커버류
아모스 트리트먼트	트리트먼트/팩

4. 중분류명

중분류명별로 구매한 소분류명의 비율을 구한 다음,
중분류 이름과 동일한 키워드 이름 분류

키워드명	소분류명	비율
여행용가방	여행용가방	"{'기타여행용가방': 0.04353688239387613, '여행용소품': 0.15200939457202506, '캐리어': 0.30445372303409884}"

1. 전체 분석과정에서 적용한 정제 로직

온라인 선호지수

5. 브랜드명 + 중분류명

브랜드명 + 중분류명으로 붙여서 분류

키워드명	소분류명	비율
샤넬향수	향수	{'남녀공용향수': 0.1741984430663676, '남성향수': 0.15760654439899724, '여성향수': 0.5959889167436337, ... }"

중분류명별 구매한 소분류명 비율 삽입

6. 대분류명

대분류명별로 구매한 소분류명의 비율을 구한 다음, 대분류 이름과 동일한 키워드 이름 분류

키워드명	소분류명	비율
상품권	상품권	"{'레저모바일상품권': 0.12560062182023743, '생활모바일상품권': 0.07684426229508197, '식음료모바일상품권': …. }"

대분류명별 구매한 소분류명 비율 삽입

7. 브랜드명 + 대분류명

브랜드명 + 대분류명으로 붙여서 분류

키워드명	소분류명	비율
핑골프남성의류	남성의류	"{'남성가디건': 0.011669331054822082, '남성남방셔츠': 0.04976778479093741, '남성베스트남성패딩': ….}"

대분류명별 구매한 소분류명 비율 삽입

8. 브랜드명

브랜드명으로 붙는 데이터 분류

키워드명	브랜드명	비율
맥	맥	"{'BB/파운데이션/컴팩트류': 0.1325361690475104, '기름종이': 0.011227237460636207, '립글로즈/틴트': 0.13267308657751814...}"

브랜드명별 구매한 소분류명 비율 삽입

1. 전체 분석과정에서 적용한 정제 로직

온라인 선호지수

9. 전처리한 브랜드명

브랜드명 전처리(괄호 및 공백 / 특수문자 삭제)
하여 데이터 분류

키워드명	브랜드명	비율
조르지오 아르마니	조르지오 아르마니	"{'BB/파운데이션/컴팩트류': 0.3303484734059769, '기름종이': 0.0009170352788866114, '남성용스킨케어류': ….}"

브랜드명별 구매한 소분류명 비율 삽입

10. 소분류명 세분화 ('티셔츠'가 들어간 소분류 비율)

9. 의 방법으로도 붙지않는 오타 혹은 세분화
(여성→여자 / 아동→키즈) 등 수동으로 붙임

키워드명	소분류명	비율
엠엘비키즈티셔츠	영유아티셔츠/탑	
롱티셔츠		"{'남성골프티셔츠': 0.016013983937769993, '남성등산티셔츠': 0.0522022475246055..}"

11. 브랜드명 + 성별

대체로 검색할 때, '르꼬끄골프여성'처럼 브랜드명 뒤에
성별을 붙이는 경우가 많았기에 포함해서 분류

키워드명	브랜드명	비율
네파남성	네파	"{'남성골프티셔츠': 0.0008767315448009822, '남성골프패딩': 0.0008767315448009822, '남성등산바지': …}"

브랜드명별 구매한 소분류명 비율 삽입

12. 전처리한 브랜드명 + 성별

11.과 같은 이유

키워드명	브랜드명	비율
캘빈클라인 진 남성	캘빈클라인진남성	"{'남성가디건': 0.010641728474685586, '남성남방셔츠': 0.003547242824895195 5, '남성베스트': …}"

브랜드명별 구매한 소분류명 비율 삽입

2. 전체 분석과정에서 기타 참고한 내용

온라인 선호지수 리서치

- ✓ 온라인 선호지수에 대한 정의
- ✓ 시장 세분화의 변수
- ✓ 유튜브 시청자 선호지수
- ✓ 하나의 상품군을 다양한 고객으로 타깃 세분화
- ✓ 한국의 소비생활지표
- ✓ 온라인쇼핑 동향조사
- ✓ 소비자물가조사

예측 모델 리서치

- ✓ 수요 예측 체계
  - 1) 시계열 분석
  - 2) 인과분석, 회귀분석
- ✓ 계절성 있는 데이터 회귀 예측
  - 회귀에서 계절성이 있는 변수들이 있는 경우 여러 개의 dummy variables 활용
- ✓ 계절성 있는 데이터 파악
  - ARIMA 모형을 활용

2. 전체 분석과정에서 기타 참고한 내용

참고문헌 & URL

✓ 온라인 선호지수에 대한 정의
http://www.wiselog.com/nethru/solution/customerfocus/preference.do

✓ 시장 세분화의 변수
https://brunch.co.kr/@flyingcity/57

✓ 유튜브 시청자 선호지수

✓ 하나의 상품군을 다양한 고객으로 타깃 세분화
https://post.naver.com/viewer/postView.nhn?memberNo=22377429&volumeNo=7711097

✓ 한국의 소비생활지표 : 한국소비자원

✓ 온라인쇼핑 동향조사 : 통계청

✓ 소비자물가조사 : 통계청

✓ 수요 예측 체계
https://www.lgeri.com/uploadFiles/ko/pdf/man/LGBI1005-19_20080908100702.pdf

✓ 계절성 있는 데이터 회귀 예측
http://www.real-statistics.com/multiple-regression/multiple-regression-analysis/seasonal-regression-forecasts/

✓ 계절성 있는 데이터 파악
http://www.statsoft.com/textbook/time-series-analysis

THANK YOU

봐 주셔서 감사합니다.

Team 포돌이