검색어 기반 업체별(브랜드) 주요 상품군 수요예측 및 개인 선호지수 시스템 분석개발

HigiCard
김윤태, 진교훈, 강경수

목 차

서 론

주요 과제 소개

주요 상품 군별
온라인 선호지수 생성

상품 군별 수요 트렌드
예측 및 인사이트 도출

새로운 아이디어 제안

주제의 필요성

주요 상품 군별 온라인 선호지수 생성

필요성

- 현재 고객들은 다른 고객들의 리뷰 평점에만 의존하여 상품을 구매하고 있다.
- 이들에게 선호지수가 주어진다면 더 명확하게 구매로 연결 지어질 수 있다.

목표

- 고객들의 편의성을 헤칠 수 있는 설문지와 같은 방식을 탈피하여 기존의 정보만으로 효율적인 선호지수 생성

고객의 선호지수 예측을 통한 고객의 상품 구매 유도 및 효율적인 재고관리를 위한 수요 예측의 한 지표로 사용

주제의 필요성

상품 군별 수요 트렌드 예측 및 인사이트 도출

필요성

- 고객은 제품이 품절된 것을 알면 바로 다른 구매처를 탐색한다. 온라인 쇼핑몰은 그 시장이 크기 때문에 대안도 그만큼 많기 때문이다.
- 제품을 공급하는 업체들에게 수요를 예측하여 전달할 수 있다면 재고 관리를 효율적으로 할 수 있다.

목표

- 기존 고객의 이탈 방지를 위한 효율적인 재고 관리
- 공급사슬을 최적화하고 최적화된 재고를 통한 추천 시스템으로의 연결

공급사슬관리 관점에서 문제를 접근: 채찍효과를 완화

검색어 기반으로 한 업체별(브랜드) 대표 상품군 수요예측 시스템

데이터 전처리

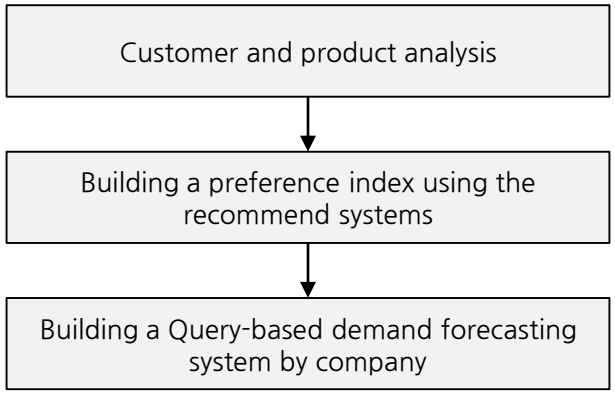
◆ 분석목적에 적합한 방식으로 데이터를 처리

불필요한 데이터 삭제

- 재고관리 기법 중 ABC 분석을 통해 우선 분석해야 할 대상 선정
- 올바르게 작성되지 않은 구매 정보이력은 효율적인 모델링을 저해하므로 삭제

필요에 따른 파생변수 생성

- 최신 트렌드 반영과 효과적인 모델링을 위해 검색어 관련 파생변수 생성
- 분석 결과 해석을 용이하게 해줄 외부 공공데이터 사용



〈그림 1〉 전체적인 분석 절차

고객 및 상품 분석

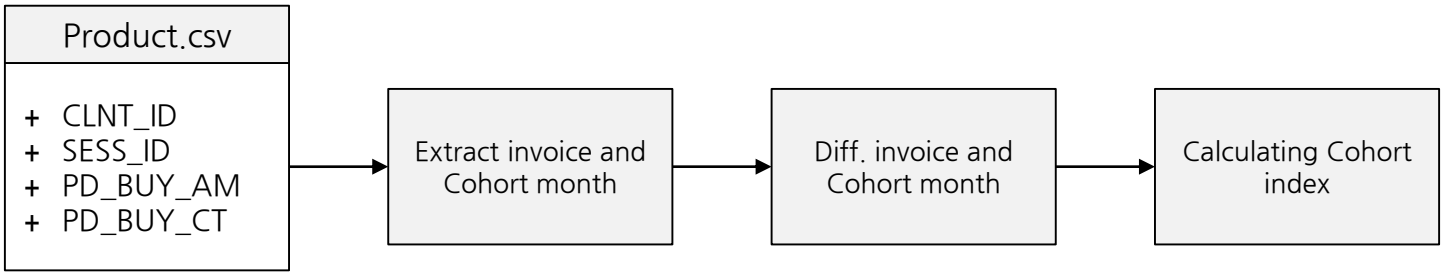
고객 유지의 중요성

Gartner	쇼핑몰 매출액의 80%는 기존 고객의 20%가 올려준다.
Marketing Metrics	신규고객에게 제품을 판매할 확률은 5~20%에 불과하다.
Forrester Research	신규고객을 유치하기위해서 기존고객에 5배의 비용이 추가된다.
Harvard Business School	고객유지율이 5%만 상승해도 25~95%까지 영업이익이 상승한다.

고객 유지와 이탈

코호트 분석

코호트 분석(cohort analysis)이란 특정기간동안 공통된 특성이나 경험을 갖는 사용자 집단을 구분하는 것으로 본 대회에서는 해당 데이터에서 나타나는 고객 유지율과 이탈율을 분석하였음.



〈그림 1〉 코호트 분석 절차도

코호트 분석 결과

〈그림 2〉 코호트 분석 시각화

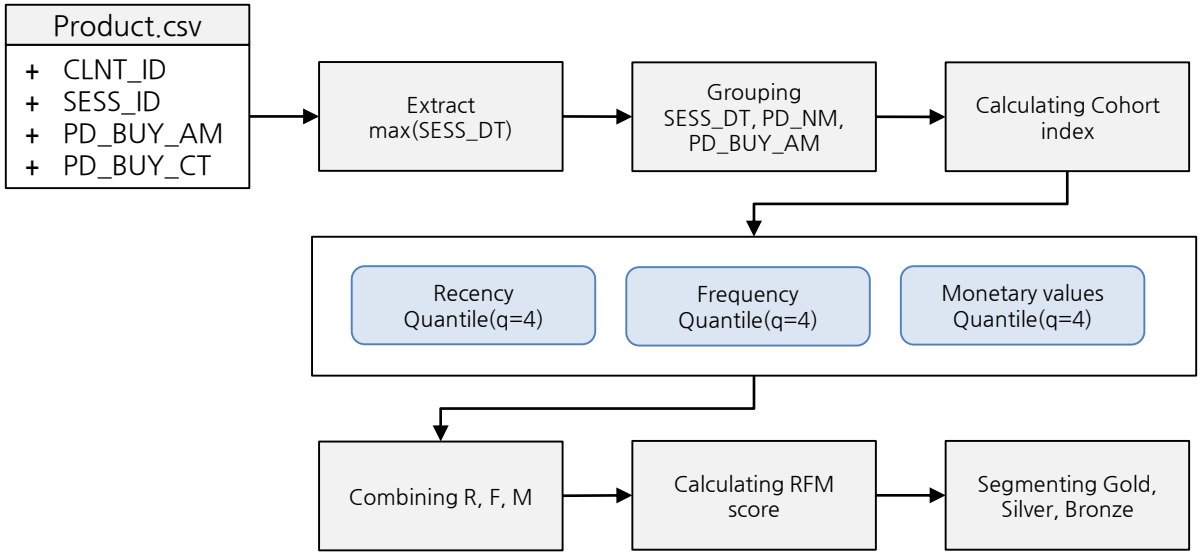
분석 결과

4월 1일부터 9월 30일까지 롯데를 이용한 고객의 유지율은 코호트 지수 2단계에서 평균 72.8%가 줄어든 27.2%로 유지되며 지수가 높아질 수록 유지율은 점차 떨어짐.

거래 특징에 따른 고객 세분화

RFM 세분화

각 고객의 R(recency), F(frequency), M(monetary)를 도출하여 다양한 고객의 특징을 도출된 값에 따라 세분화하는 방법으로 본 대회에서는 골드, 실버, 브론즈 군으로 세분화하여 각 등급에서 나타나는 특징을 분석하였음.



〈그림 3〉 RFM 세분화 절차도

고객 세분화 결과

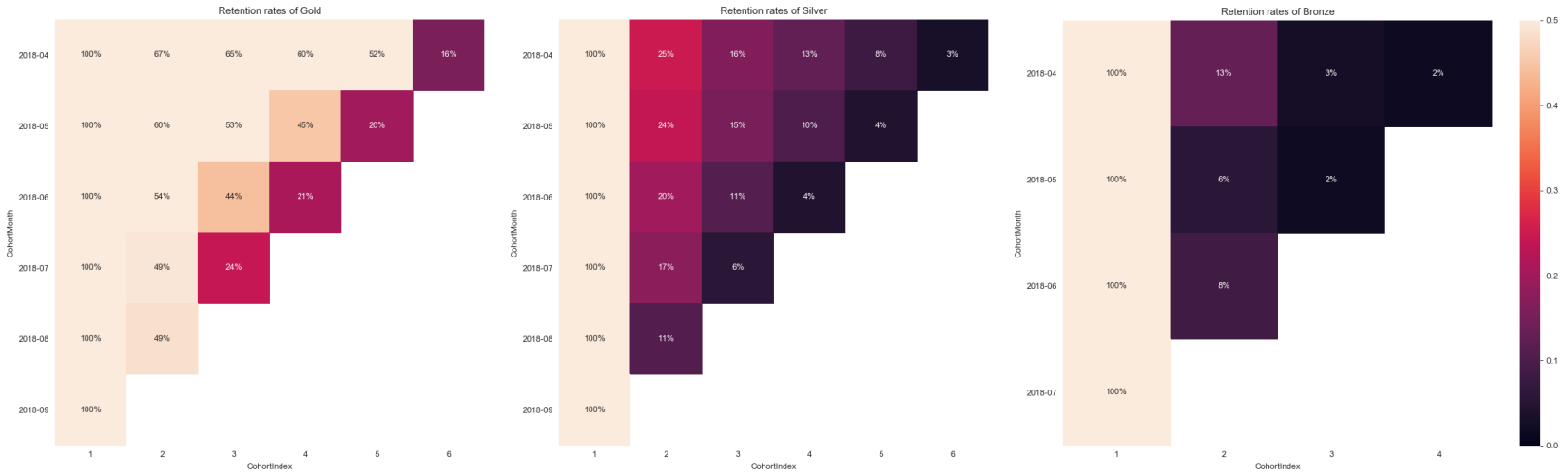
〈표 1〉 RFM 세분화 결과 요약

Segment	Recency	Frequency	Monetary	Count
Gold	46.5	11.5	613,132.1	285,327
Silver	81.7	2.9	154,137.4	493,809
Bronze	136.0	2.0	48,527.6	143,601

분석 결과

Gold 등급의 최근 거래일은 평균 46.5일이며, 거래빈도는 평균 11.5회 그리고 사용한 평균 구매액은 613,132.1원이며 총 285,327명이 해당 등급으로 세분화되었음. 등급이 내려갈수록 언급한 수치들이의 변화가 큰 편임.

세분화한 코호트 분석 결과



Gold | Silver | Bronze

〈그림 4〉 RFM 세분화 등급에 따른 코호트 분석 시각화

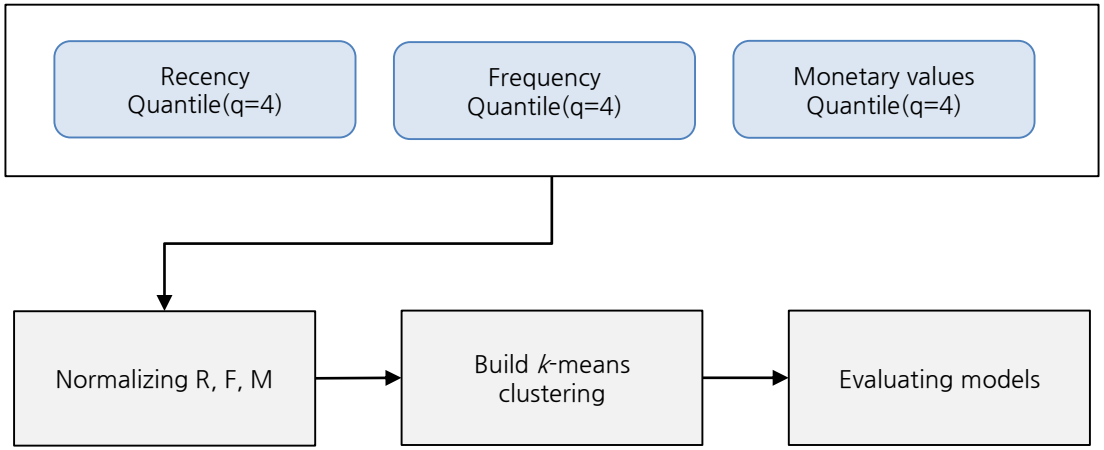
분석 결과

RFM 세분화에서 도출된 각 고객 등급을 활용하여 코호트 분석을 실시하였고 Gold 등급은 꾸준한 고객 유지율을 보이지만, Silver와 Bronze는 급격하게 하락하는 것으로 나타났음. 특히 Bronze 등급은 8월과 9월의 거래량이 없는 것으로 보이며 이는 신규 고객이 없는 것으로 판단됨.

거래 특징에 따른 고객 군집화

k-평균 군집화

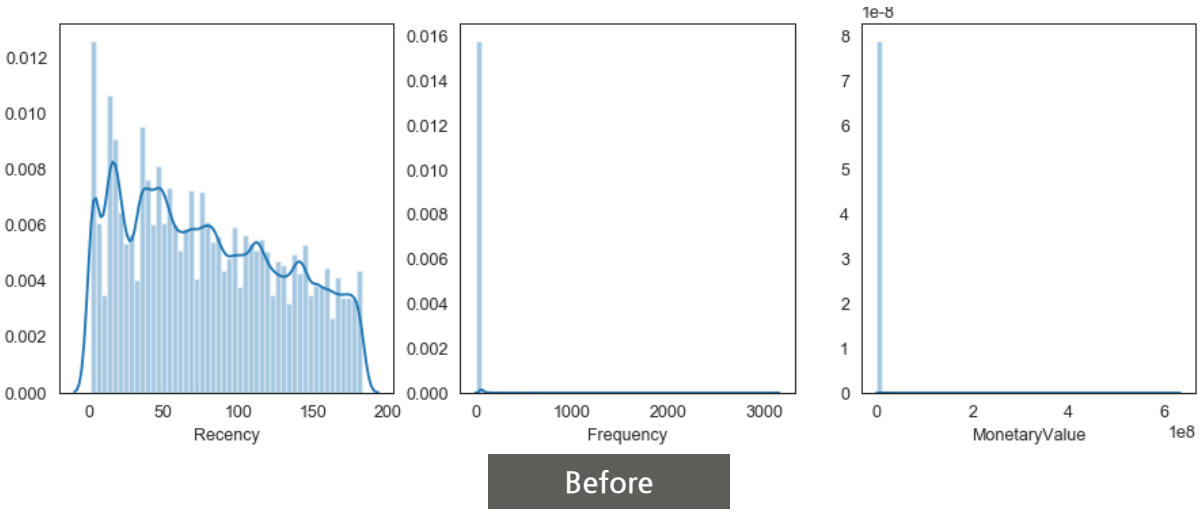
RFM 세분화를 위해 처리된 데이터를 가지고 *k*-평균 군집화를 실시, *k*-평균 군집화는 *k*개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작함. 비지도 학습으로 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행함. 해당 데이터를 통해 몇 개의 군집으로 군집가능한지 판단하기 위함.



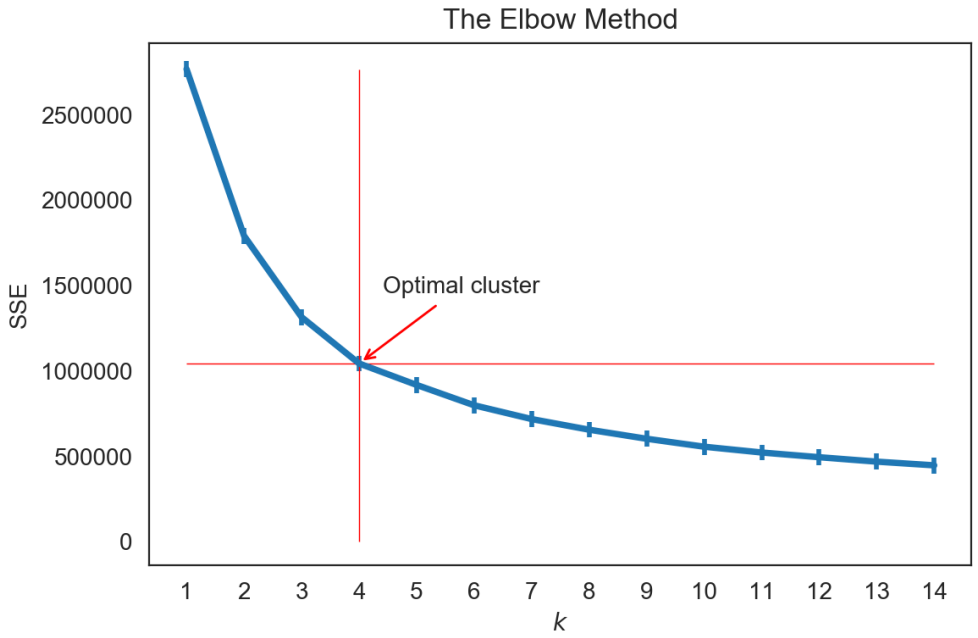
〈그림 5〉 *k*-평균 군집화 분석 절차도

거래 특징에 따른 고객 군집화

〈그림 6〉 *k*-평균 군집화 분석을 위한 데이터 표준화



거래 특징에 따른 고객 군집화



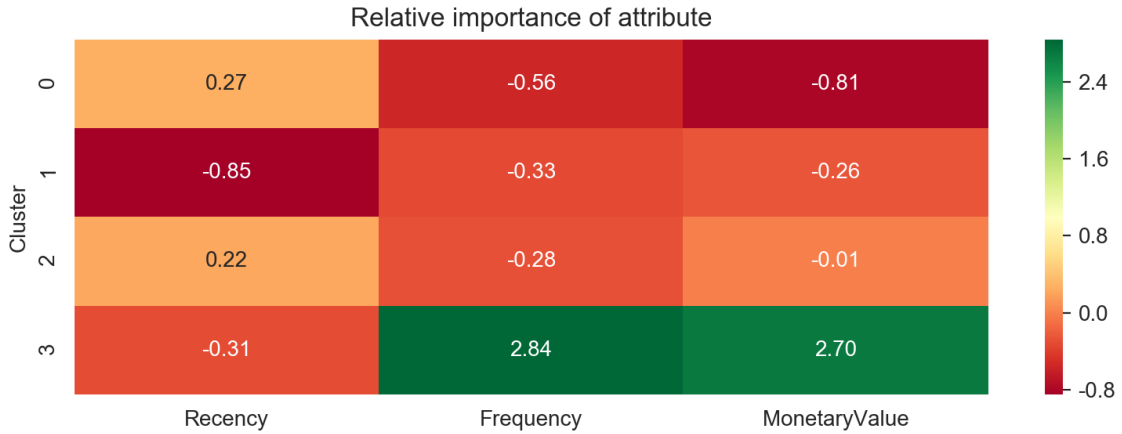
〈그림 7〉 최적의 군집개수 평가

분석 결과

휴리스틱 평가방법인 Elbow 기법을 사용하여 학습된 모형에서 최적의 군집 개수를 결정 하기위해 평가하였으며, 해당 데이터에서 군집할 수 있는 군집의 개수는 총 4개로 도출됨.

고객 군집화 결과

〈그림 8〉 *k*-평균 군집화 분석 결과 시각화



〈표 2〉 *k*-평균 군집화 분석 결과 요약

Cluster	Recency	Frequency	Monetary	Count
C_0	100.0	2.0	53,723.0	336,688
C_1	12.0	4.0	207,070.0	147,132
C_2	97.0	4.0	277,454.0	323,210
C_3	55.0	21.0	1,035,339.0	115,707

고객 세분화 및 군집 분석 결과 비교

FRM Segment

Segment	Recency	Frequency	Monetary	Count
Gold	46.5	11.5	613,132.1	285,327
Silver	81.7	2.9	154,137.4	493,809
Bronze	136.0	2.0	48,527.6	143,601

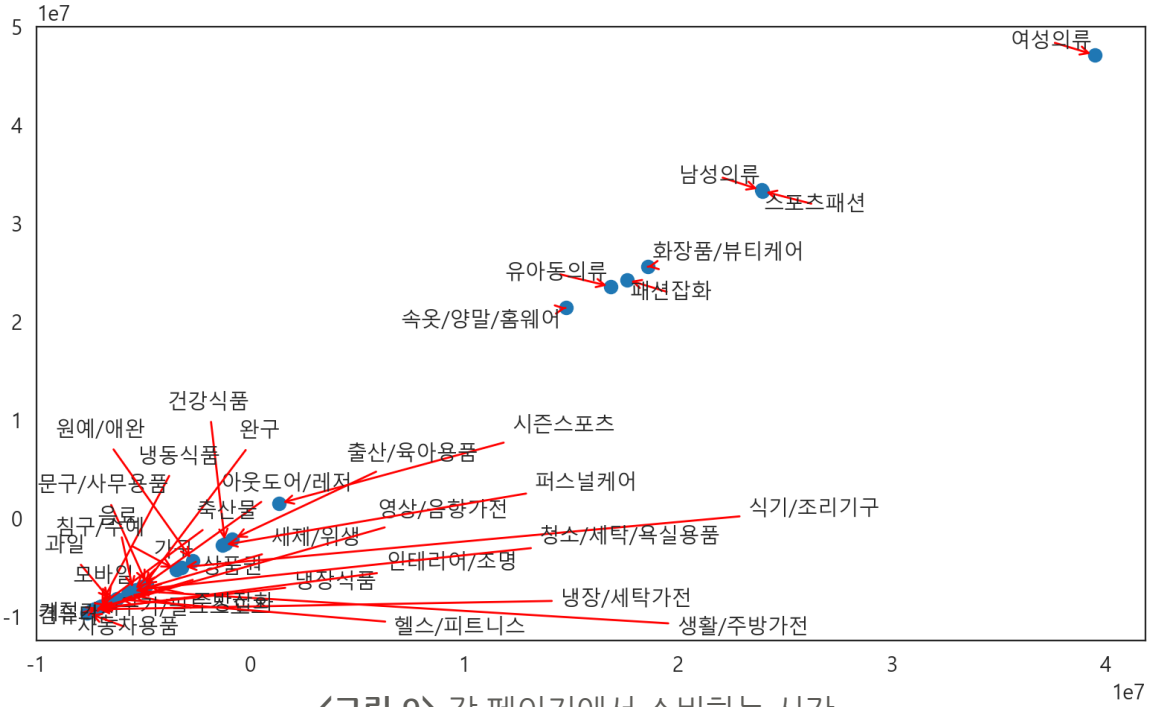
***k*-means clustering**

Cluster	Recency	Frequency	Monetary	Count
C_0	100.0	2.0	53,723.0	336,688
C_1	12.0	4.0	207,070.0	147,132
C_2	97.0	4.0	277,454.0	323,210
C_3	55.0	21.0	1,035,339.0	115,707

분석 결과

각 군집에 대한 분석 결과 RFM 세분화와 비슷한 결과를 끌어낼 수 있음. 이를 정리하면 Gold 및 C_0에 해당하는 고객은 지속적인 활동과 매출을 발생하며, 나머지 고객군에 대해 관리가 필요함.

각 페이지에서 소비하는 시간



〈그림 9〉 각 페이지에서 소비하는 시간

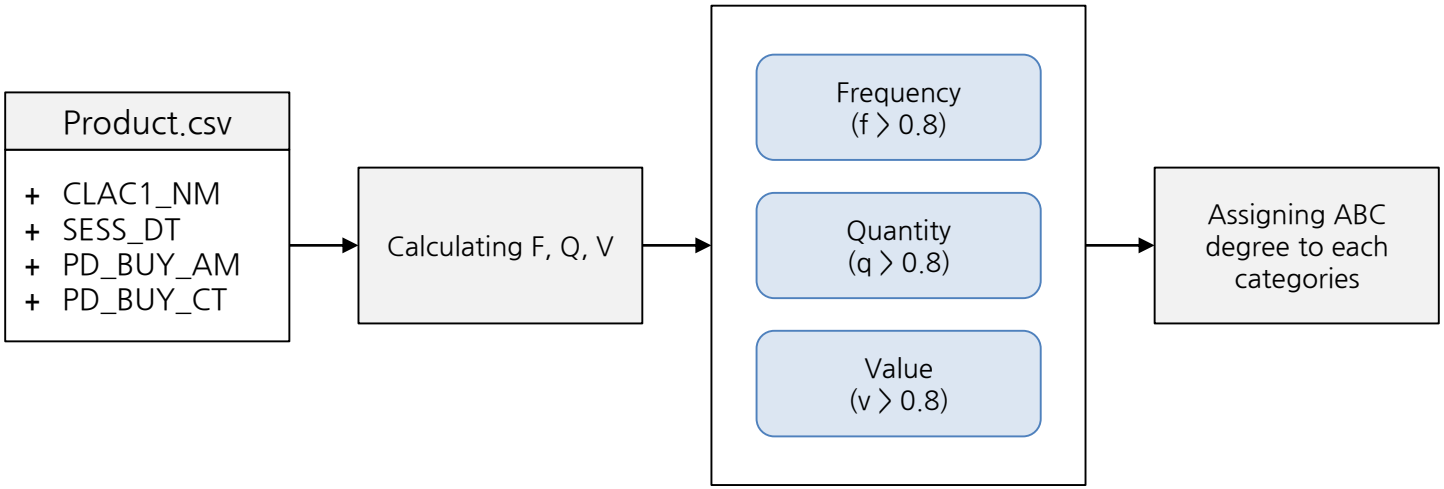
분석 결과

상품 대분류 페이지에서 고객이 모바일과 데스크톱에서 소비하는 시간을 MDS(다차원 척도법)를 활용하여 분석한 결과, 〈여성의류〉가 다른 상품군에 비해 매우 많은 시간을 소비하며,
기타 의류 및 화장품을 구매하기 위해 많은 두번째로 많은 시간을 소비하는 것으로 나타났음.
나머지 상품군에 대해서 고객은 매우 빠른 시간 내에 구매를 선택하는 것으로 도출됨.

분석대상의 축소

ABC 분석

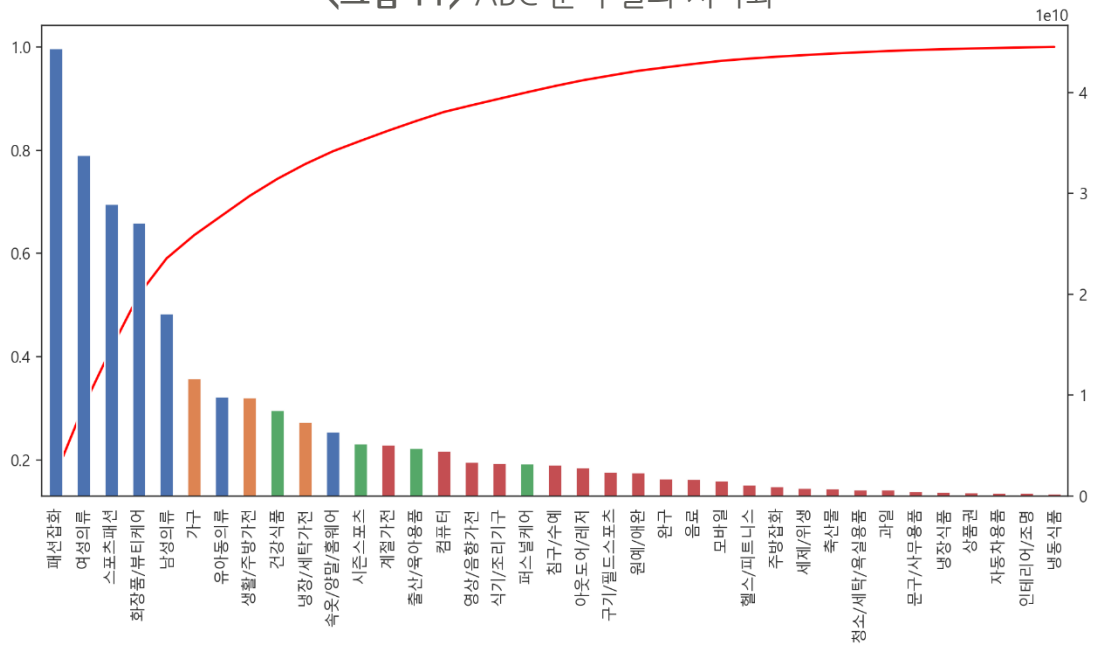
- ABC 분석은 취급하는 상품의 중요도에 따라 A등급, B등급, C등급으로 나누어 효율적으로 관리하기 위한 재고관리의 기법 중 하나. **파레토 법칙**에 따라 상위 20%의 상품이 매출의 80%를 발생시키며 이를 근거로 관리와 통제가 수준이 변함.
- 제공받은 데이터의 상품군 대분류는 총 37개, 중분류 128개, 소분류 898개로 매우 많은 상품이 존재하며, 이를 효율적으로 분석하고 전달하기위해서는 데이터를 축소시켜야 함.
- 따라서 ABC 분석을 통해 도출된 등급에 따라 우선 분석되어야 할 대상을 선정하기로 함.



〈그림 10〉 ABC 분석 절차도

ABC 분석 결과

<그림 11> ABC 분석 결과 시각화



등급	상품(대분류)
A	여성의류
	회장품/뷰티케어
	스포츠패션
	남성의류
	패션잡화
	유아동의류
B	건강식품
	퍼스널케어
	출산/육아용품
	시즌스포츠
C	생활/주방가전
	냉장/세탁가전
	가구

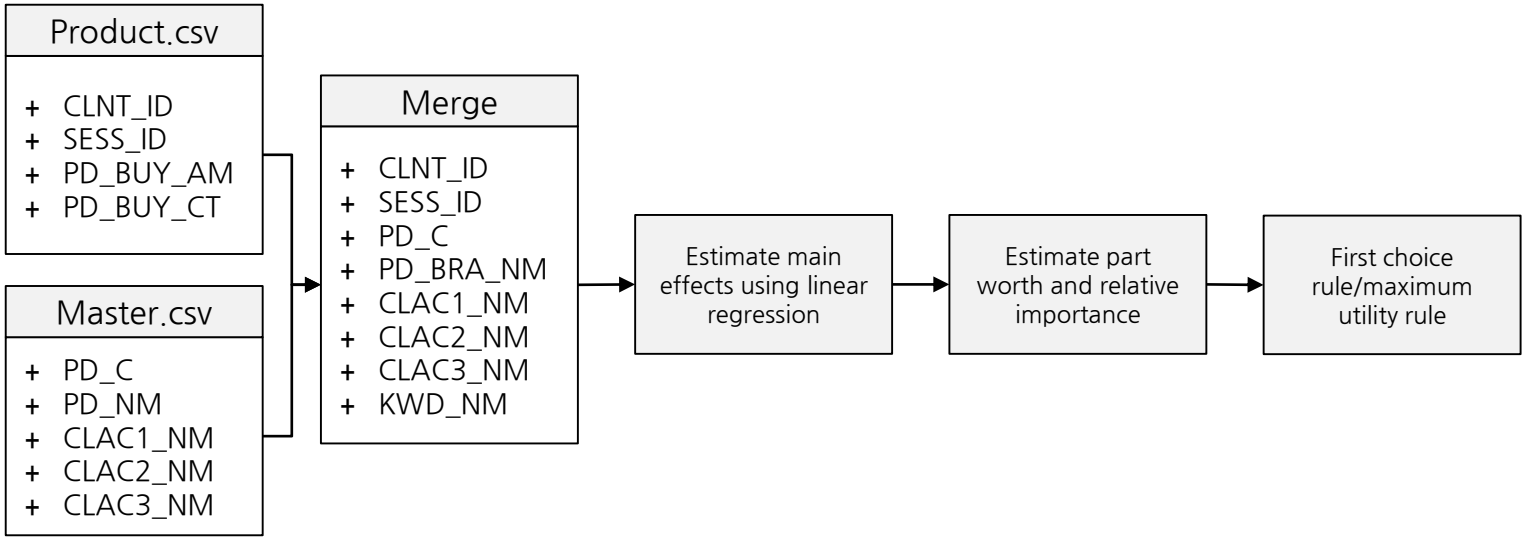
분석 결과

단순한 매출만을 고려하지 않고 거래빈도, 수량까지 고려하여 총 13개의 상품군만 우선 분석하기로 결정하였으며, 특히 A등급으로 부여된 6개의 대분류에 대해 집중적으로 분석하였음.

선호하는 제품에 대한 특징

컨조인트 분석

컨조인트 분석(conjoint analysis)은 각 제품의 속성(기능, 디자인, 등)을 체계적으로 변화시켜가면서 개인의 선호 체계를 측정하는 기법으로 여러 가지 다른 속성을 가진 제품이나 서비스에 대한 사람들의 평가를 확인하기 위해 활용하는 통계 기법임. 본 데이터에서 도출되는 속성에 따라 해당 서비스를 이용하는 고객의 선호를 알아보기 위해 분석을 진행함.



〈그림 12〉 컨조인트 분석 절차도

컨조인트 분석 결과

〈표 3〉 컨조인트 분석 결과

Category	Preference color	Preference size	Preference part	R^2	p-value
여성의류	브라운	M	여성점퍼	0.407	9.64e-38
남성의류	블루	2XL	남성정장셔츠	0.415	1.39e-14
스포츠패션	블루	M	남성등산점퍼/재킷	0.415	1.39e-14

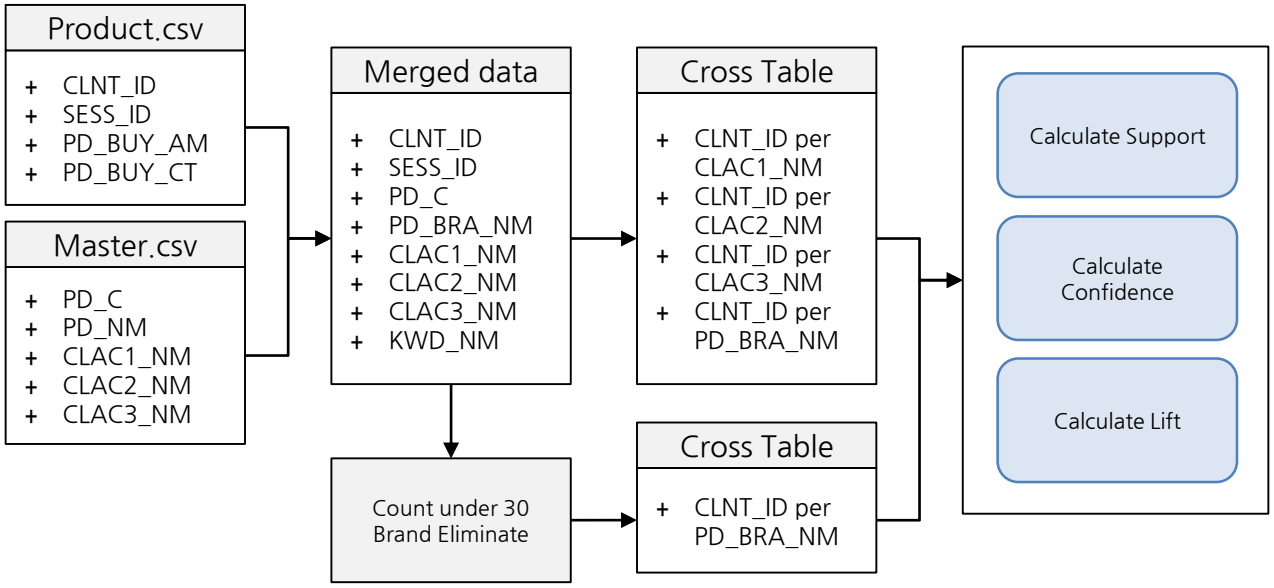
분석 결과

PD_NM에서 각 속성을 추출하여 여성의류, 남성의류, 스포츠 패션에 대한 선호 색상, 사이즈, 소분류를 도출하였고 결과는 〈표 3〉과 같음. 결과를 이해하기 어려워 추가적인 분석을 진행함.

장바구니 분석

연관규칙 분석

연관규칙 분석(association rules)는 아이템 집합에서 발생하는 일련의 규칙들을 생성하는 알고리즘. 경영학에서 장바구니 분석으로 알려져 있으며, 소비자들의 구매이력을 통해 “A 아이템을 구매하는 고객은 B 아이템 구매할 가능성이 높다”는 결론을 내는 알고리즘. 다른 상품을 제안하거나 추천 시스템 컨텐츠 기반(contents-based) 기본이 되는 방법론



〈그림 13〉 연관규칙 분석 절차도

장바구니 분석 결과

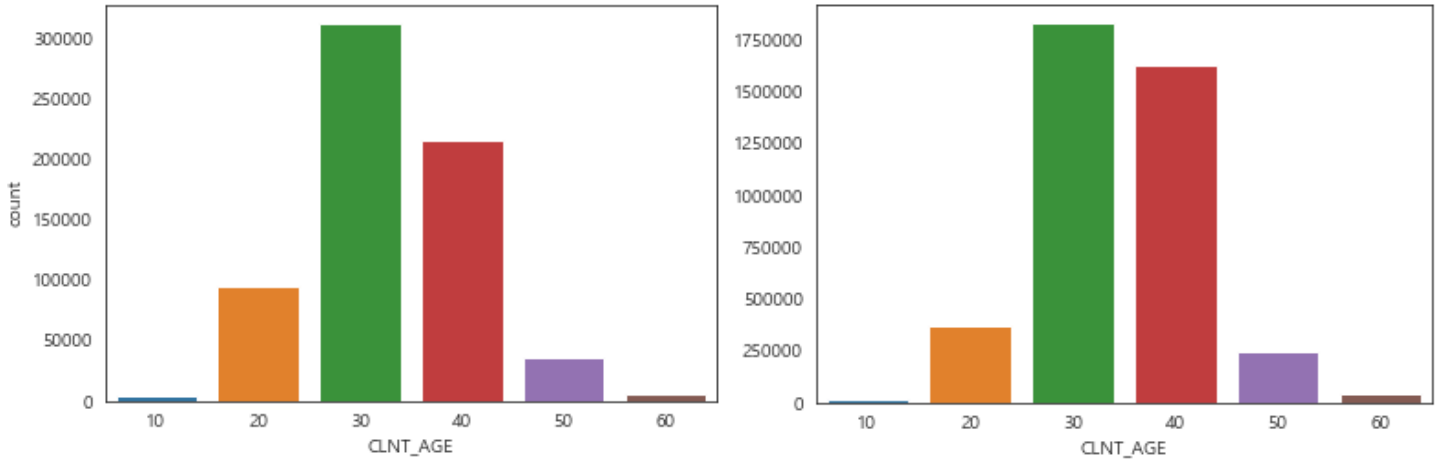
〈표 4〉 장바구니 분석 결과 요약

Antecedent	Consequent	confidence	lift
여성남방셔츠	남성티셔츠	1.561	42.079
BB/파운데이션/컴팩트류	남성티셔츠	1.509	40.668
남성티셔츠	여성원피스	1.451	36.922
BB/파운데이션/컴팩트류	여성원피스	1.176	29.917
여성남방셔츠	여성원피스	1.163	29.611
BB/파운데이션/컴팩트류	여성남방셔츠	1.007	31.763

분석 결과

신뢰도(confidence)를 기준으로 정렬하였으며, **여성남방셔츠**를 구매하는 고객들이 **남성티셔츠**를 구매하는 것으로 도출됨. 이를 좀 더 세부적으로 분석하기위해 **외부데이터**와 **1차원적인 분석**을 실시함.

장바구니 분석 결과

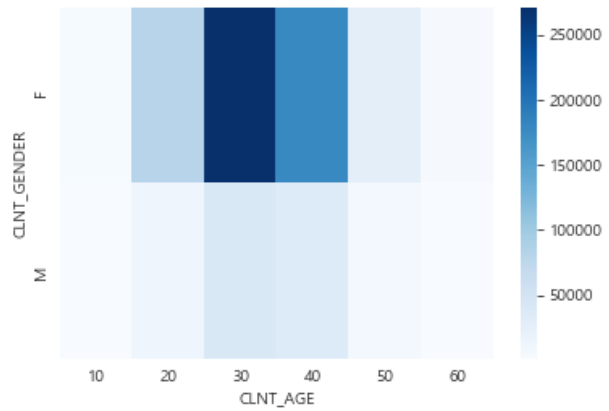


〈그림 14〉 구매액과 구매빈도 분석 결과

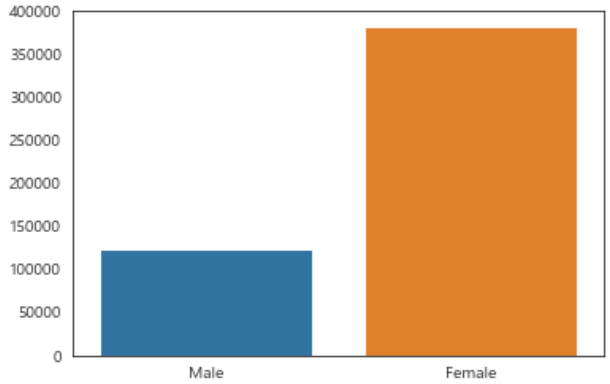
분석 결과

60, 70, 80대는 전체에서 1%의 비율로 하나로 통합하였으며, 연령대의 분포를 통해 전체 구매액을 30와 40대가 압도적으로 발생시키며, 구매빈도 또한 30와 40대가 압도적임.

장바구니 분석 결과



〈그림 15〉 나이와 성별에 따른 구매 차이

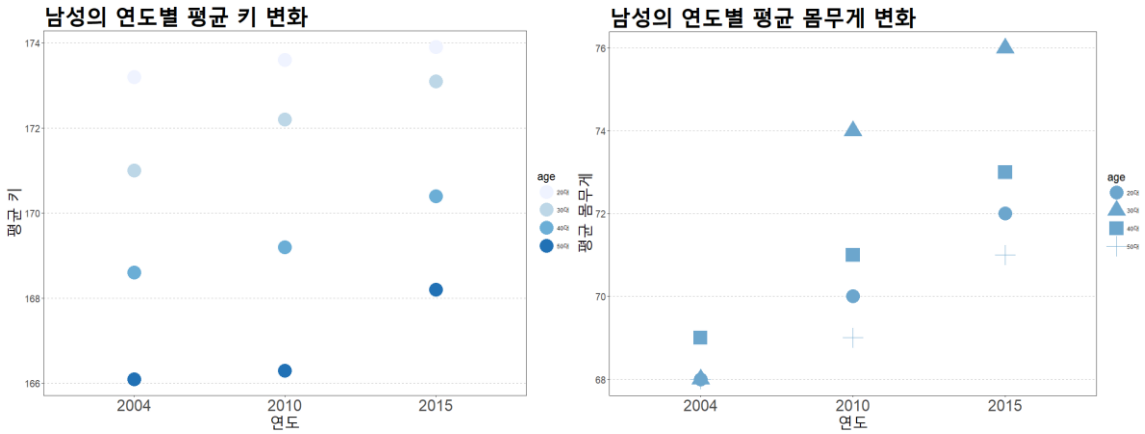


〈그림 16〉 남성의류 구매 성별

분석 결과

나이와 성별에 따른 구매 분포를 살펴보면 **30대 여성**이 매우 많음을 알 수 있으며, 실제 남성의류의 경우에서 남성보다 오히려 여성들이 많은 구매를 하는 것으로 나타남. 장바구니 분석의 결과를 통해 이는 여성이 자신의 옷을 구매할 때 남성의 옷을 같이 구매하는 것으로 볼 수 있으며, 세부적인 분석을 통해 유아용의류를 구매한 내역과 연령대를 통해 추론하면 **기혼자**일 가능성이 높다고 할 수 있음.

장바구니 분석 결과



〈그림 17〉 ABC 분석 결과 시각화

분석 결과

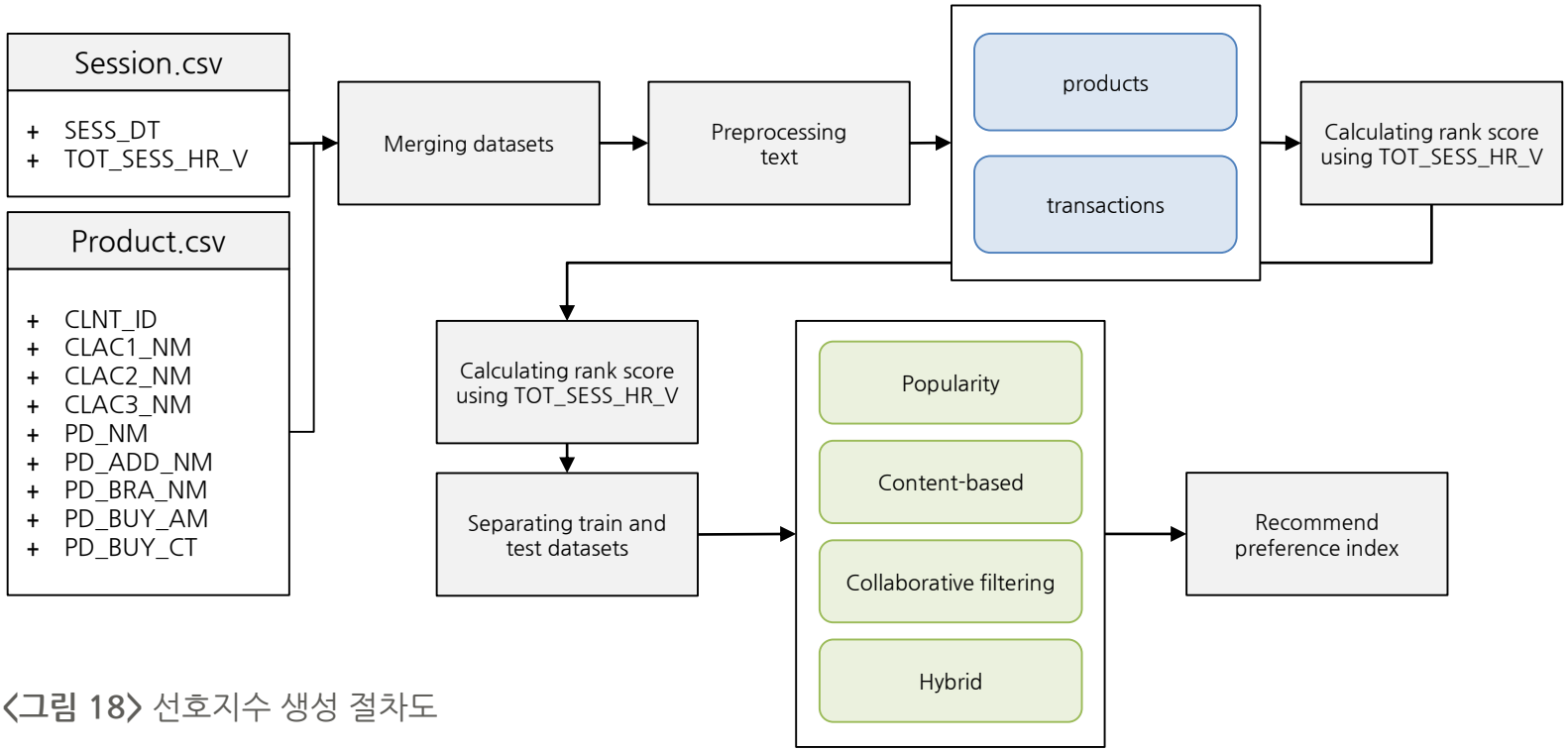
컨조인트 분석 결과, 남성의류의 경우 큰 사이즈의 남성정상셔츠 효용도가 가장 높게 도출됨. 추가적인 분석을 위해 산업통상자원부에서 제공하는 한국인 인체치수조사 데이터를 분석한 결과, 한국 남성의 키의 변화는 미미하지만 평균 몸무게의 변화는 해를 거듭할수록 급격히 증가하는 추세임을 알 수 있음. 특히 젊은 층의 몸무게 증가율이 인상적일 정도. 즉, 비만도 증가를 감안해보았을 때 큰 사이즈의 옷을 구매하는 것이라고 추론됨.

개인화 선호지수 시스템

개인화 선호지수 시스템

추천 시스템

선호지수를 도출하기 위해 Popularity, Content-based, Collaborative filtering 세 개의 추천 시스템 모형을 구축하였음. 리뷰 점수가 없기 때문에 페이지에서 머문 시간이 많을 수록 신중한 고민을 했을 것 이라 판단하고 4분위수를 활용하여 4점 척도로 평가하였음.



〈그림 18〉 선호지수 생성 절차도

개인화 선호지수 시스템 결과

	_CLNT_ID	hits@10_count	hits@5_count	interacted_count	recall@10	recall@5
162	5302905	28	10	201	0.139303	0.049751
318	4006944	20	10	188	0.106383	0.053191
2955	4931190	15	5	140	0.107143	0.035714
3877	5142305	39	28	87	0.448276	0.321839
868	2245619	50	36	78	0.641026	0.461538
720	5091280	9	1	76	0.118421	0.013158
2283	5453018	24	15	74	0.324324	0.202703
7650	6366359	13	10	61	0.213115	0.163934
674	3803223	18	10	51	0.352941	0.196078
6308	2426355	27	24	51	0.529412	0.470588

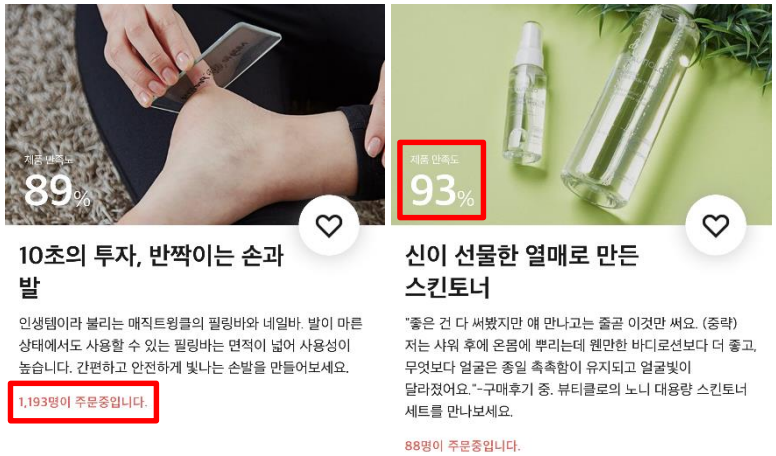
〈표 5〉 모형에서 도출된 평가 결과

분석 결과

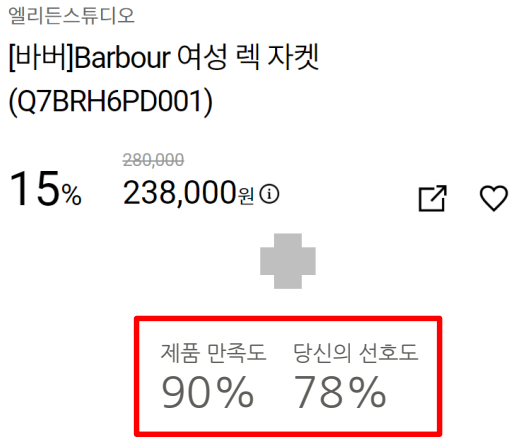
모형에서 도출된 결과값으로 구매빈도가 많다고 해서 정확성이 증가하지 않음. 학습된 모형을 통해 추천할 수 있는 아이템과 새로운 아이템을 클릭할 시, 선호지수를 백분율로 나타내어 제안하기로 함.

새로운 아이디어 제시: 당신의 선호지수

〈그림 19〉 K사 모바일 쇼핑몰 UI



〈그림 20〉 롯데백화점 온라인 몰



분석 결과 및 제안

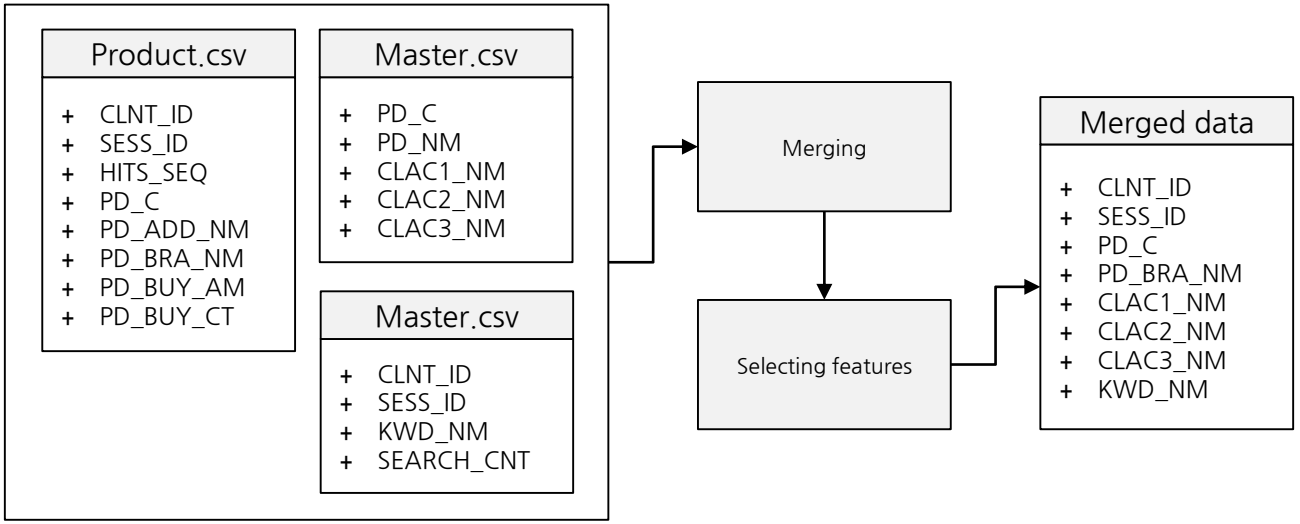
왼쪽은 K사의 모바일 쇼핑몰이며 제품 만족도를 백분율로 표시하여 해당 제품의 만족도를 전면에 내세우고 있지만, 롯데백화점 온라인몰은 썸네일 그리고 페이지 내용에서 고객이 원하는 정보를 보기 위해서는 상당한 체류시간이 발생함. 따라서 오른쪽 그림과 같이 제품 만족도와 선호도를 백분율로 나타내어 고객이 원하는 정보를 실시간으로 보여주는 새로운 아이디어를 제안함.

수요예측시스템

검색어 키워드 자연어 처리(1)

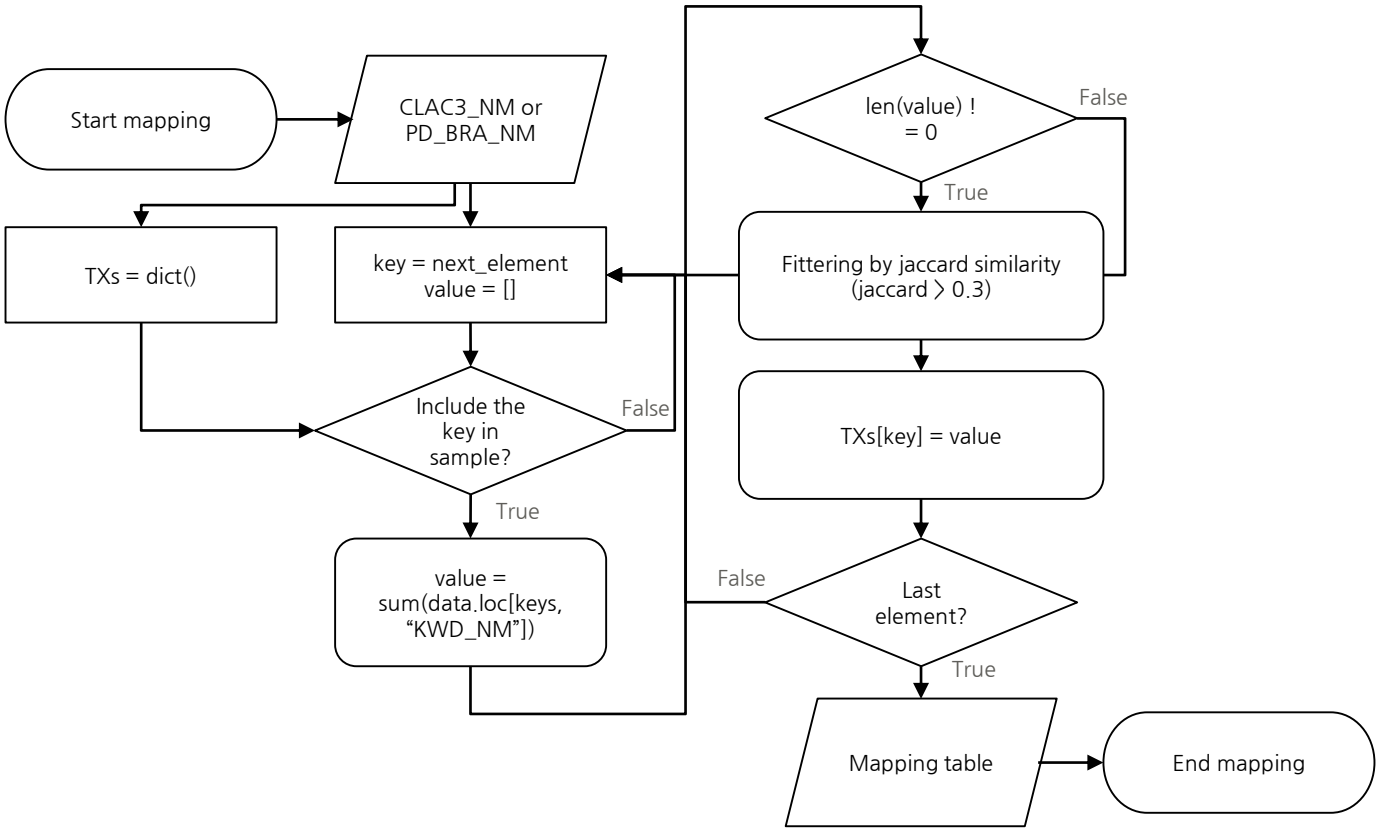
키워드 매핑

주어진 데이터 내에서 사용자의 의도가 직접적으로 드러나는 정보는 검색 키워드이다. 검색 키워드를 통해 사용자가 보고자 하는 상품의 분류 카테고리와 브랜드를 파악할 수 있도록 사전에 검색된 기록을 기반으로 새로운 키워드를 검색할 때, 사용자가 원하는 바를 추측할 수 있도록 하고자 함.



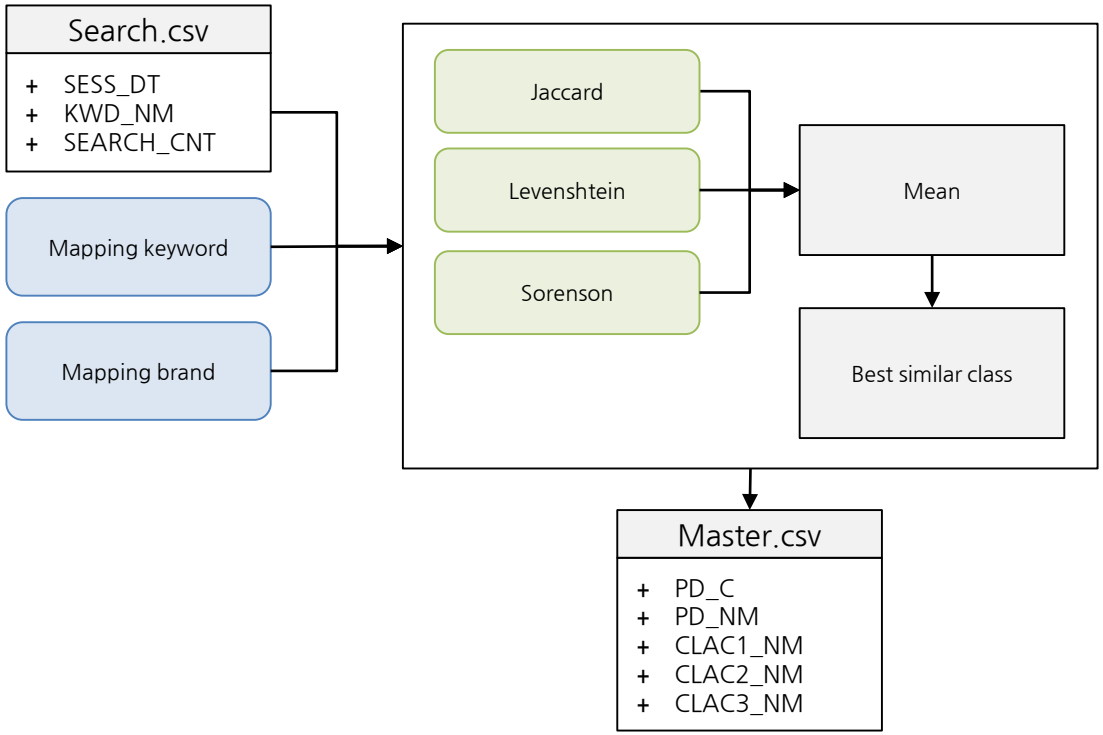
〈그림 21〉 키워드 매핑 알고리즘 절차도 (1)

검색어 키워드 자연어 처리(2)



〈그림 22〉 키워드 매핑 알고리즘 절차도 (2)

검색어 키워드 자연어 처리(3)

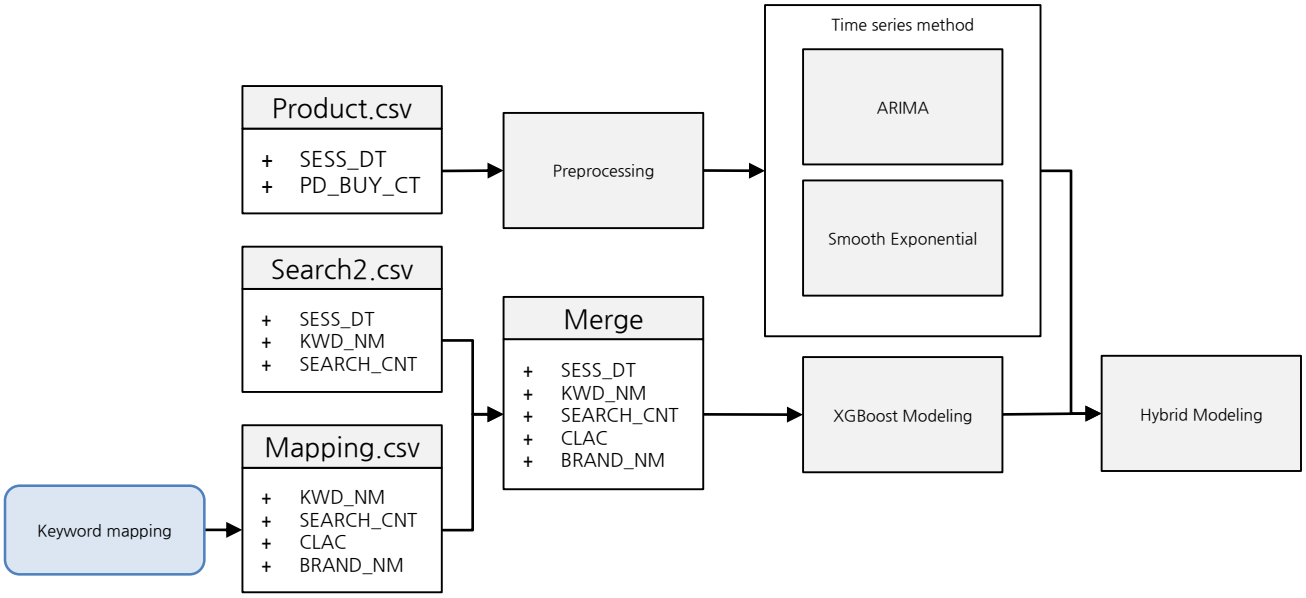


〈그림 23〉 키워드 매핑 알고리즘 절차도 (3)

검색어 기반 수요예측 시스템

수요예측 기법

- 지수평활법: 가장 최근 데이터를 가지고 가중치를 부여하여 과거의 시간에 대해서는 고려하지 않는 기법으로 전통적인 시계열 기법 중 하나
- ARIMA: 시계열 분석 기법의 한 종류로, 과거의 관측 값과 오차를 사용해서 현재의 시계열 값을 설명하는 ARMA 모형을 일반화 한 것
- XGBoost: Gradient Boosting 중 한 기법으로 빠르고 유연한 특징인 Boosting 기법 중 하나



〈그림 24〉 검색어 기반으로 한 업체별(브랜드) 주요 상품군 수요예측 시스템 절차도

검색어 기반 수요예측시스템

공급사슬관점에서의 수요예측

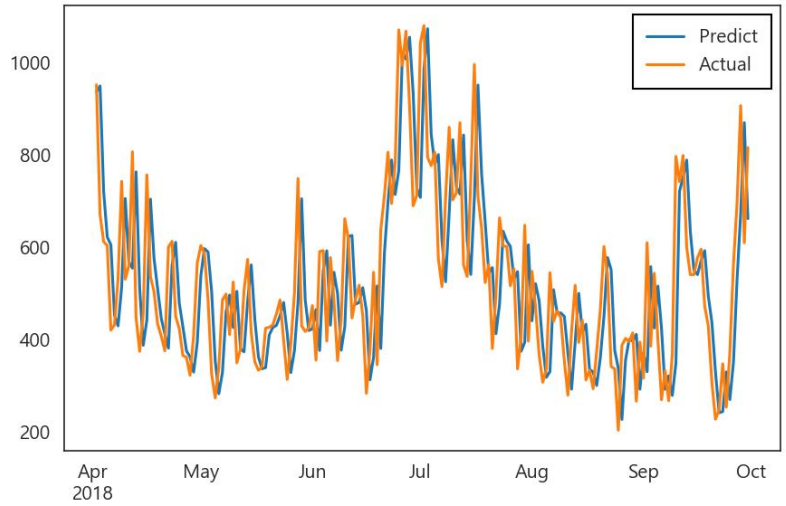
- 전체 공급사슬의 최적화, **채찍효과** 완화 그리고 비용절감을 위해 단순하게 상품군을 예측하기보다 **업체별**(브랜드) 주요 **상품군**에 대해 수요를 예측하기로 함.
- 롯데백화점의 대다수의 제품은 롯데의 브랜드가 다양한 공급업체가 모인 하나의 **플랫폼**으로 상품군만 예측한다면 의미가 없을 것으로 판단
- 따라서 브랜드를 기준으로 각 브랜드의 주요 상품군을 선정하여 해당 상품군에 대한 수요를 예측하고 9월 **마지막 주(7일)**에 해당하는 수요를 예측하기로 함.
- 7일을 선택한 이유는 공급업체(브랜드)가 재고를 확인하고 **보충**할 **리드타임**을 고려하여 선택

수요예측시스템은 소비자의 관점이 아닌 공급사슬관점에서 접근

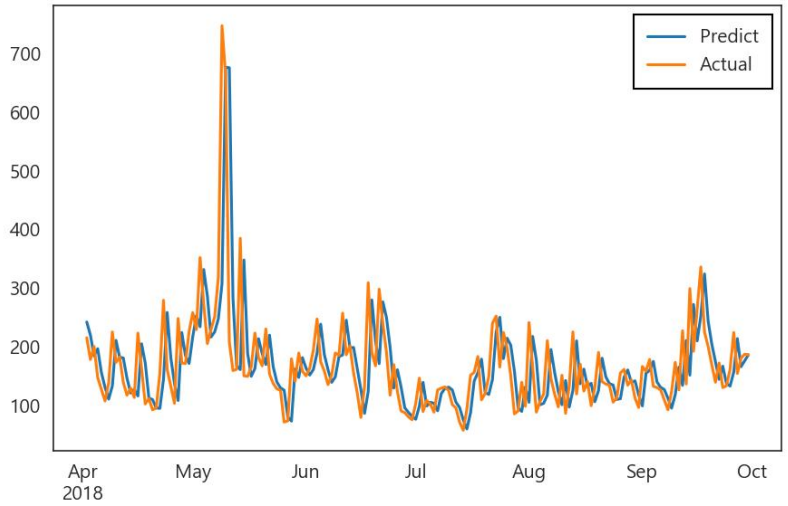
Source: https://www.ironsystems.com/services/supply-chain-management

수요예측시스템 시각화

나이키



입생로랑



〈그림 25〉 검색어 기반으로 한 업체별(브랜드) 주요 상품군 수요예측 결과

분석 결과

ARIMA모형에 적합 시켜본 결과 P-Value는 유의수준 0.05하에서 모형은
유의하다고 볼 수 있으며 지수평활법 또한 같은 이유로 모형이 유의하다고 볼 수 있다.
XGBoost의 경우 비모수적인 기법이므로 검색어와 같은 자연어 처리 결과를 확인하기 좋은 기법이므로 사용하였다.

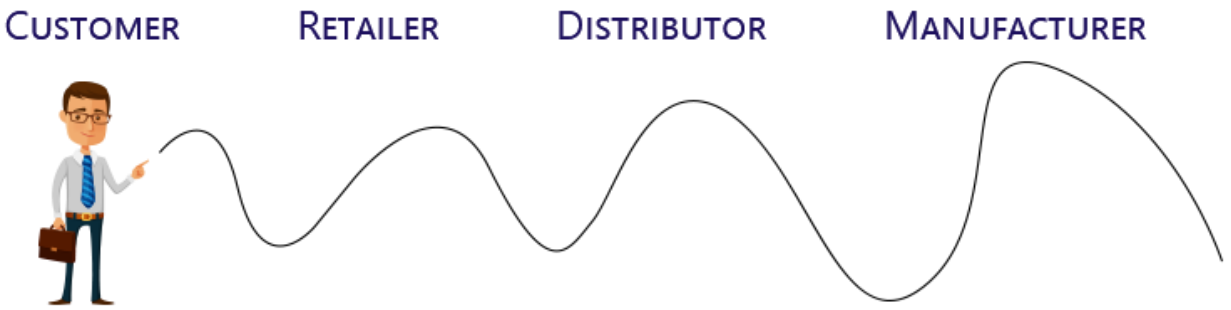
예측한 값과 실제 값을 도식화 시켜서 살펴보았을 때도 예측정확성이 좋은 것으로 판단함.

수요예측 시스템 평가

〈표 1〉 각 모형에 따른 RMSE와 RMSLE 평가

Model	Brand	RMSE	RMSLE
Smoothing Exponential	지오다노	182.38	0.491
	나이키	241.61	0.464
	입생로랑	34.31	0.196
ARIMA	지오다노	182.13	0.491
	나이키	240.88	0.463
	입생로랑	34.15	0.195
XGBoost	지오다노	63.82	0.161
	나이키	129.65	0.197
	입생로랑	34.21	0.213
Hybrid	지오다노	70.77	0.154
	나이키	138.11	0.204
	입생로랑	31.44	0.181

새로운 아이디어 제시



〈그림 24〉 수요가 상류로 올라갈수록 왜곡되는 채찍효과

Source: https://blog.arkieva.com/what-is-bullwhip-effect/

분석 결과

검색어를 기반으로 한 업체별(브랜드) 수요예측시스템은 다음과 같다:

- 검색어 기반의 목적은 유행을 타는 상품군을 반영하기 위함.
- 업체별(브랜드)를 기준으로 수요예측은 공급사슬관점에서 채찍효과 완화로 인한 전체 비용 절감을 목표로 구현함.
- 추세를 고려한 기법과 검색어를 기반으로 한 모형의 하이브리드가 안정적인 결과를 도출하는 것으로 나타남.

결론

최종 결론

고객 및 상품 분석 결과

→ ABC분석을 기반으로 상품군 분류
→ 기존 고객 유지와 등급에 따른 마케팅을 통해 매출 향상 도모
→ 구매자의 대다수가 기혼 여성으로 추론됨

개인 선호지수 결과

→ 신규고객 유치를 위해서 선호지수 개발
→ 기존 구매방식은 다른 사람들의 리뷰에만 전적으로 의존
→ 선호지수는 백분율로 나타내어 이해성을 높임
→ 선호지수를 통한 수요예측 시스템의 정확성 향상을 위한
→ 고객의 리뷰점수가 있으면 더욱 정확한 모델링 가능

수요예측시스템 모델링 결과

→ 최신 트렌드를 반영하기 위해 모델에 검색어 내용 추가
→ 평균 RMSLE 0.2대의 준수한 성능
→ 채찍효과 완화를 통해 공급업체의 효율적인 재고관리 향상
→ 팀 내 컴퓨터 성능으로 인해 복잡한 모형 구현이 어려웠음.