

CAKD 5 1st PROJECT

4분기 매출 극대화를 위한 유의 고객 예측 및 솔루션

TEAM 1

김민성 어정호 조경윤 조남현 최지원

목차

- 01 기획 동기 및 주제 선정
- 02 데이터 분석과 모델링
- 03 결과 해석 및 인사이트 도출
- 04 타겟 별 솔루션 제안

제 1 장

기획 동기 및 주제 선정

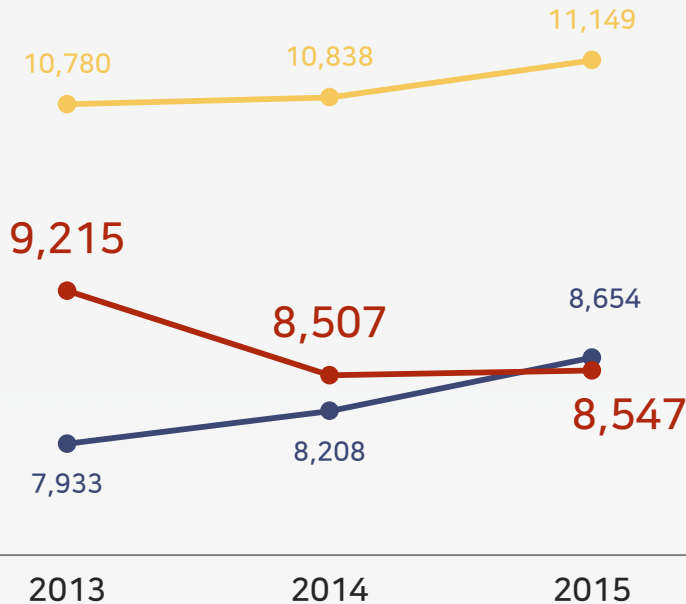
- 기획 배경
- 대상 고객 탐색
- 주제 선정

기획 배경

대형마트 3사 매출 추이

(단위: 십억 원)

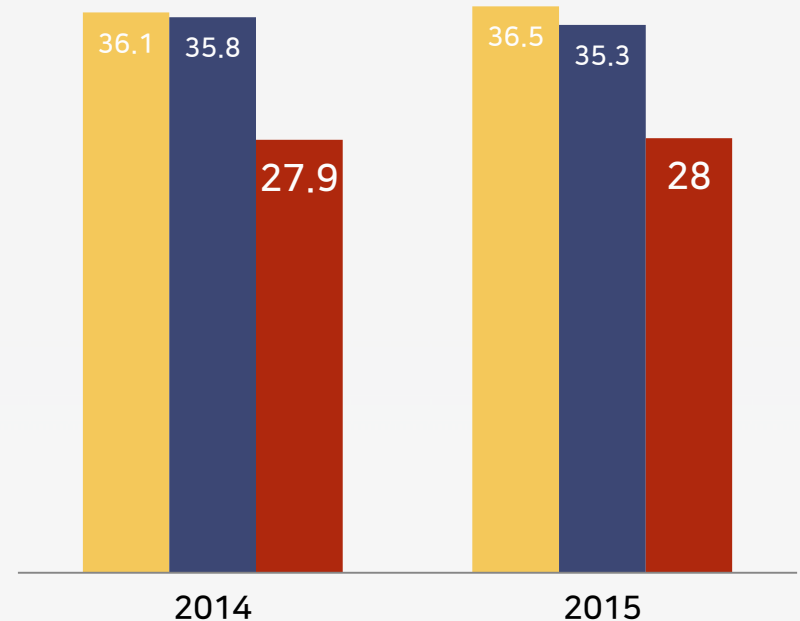
● E사 ● H사 ● L사



대형마트 3사 시장점유율

(단위: %)

■ E사 ■ H사 ■ L사

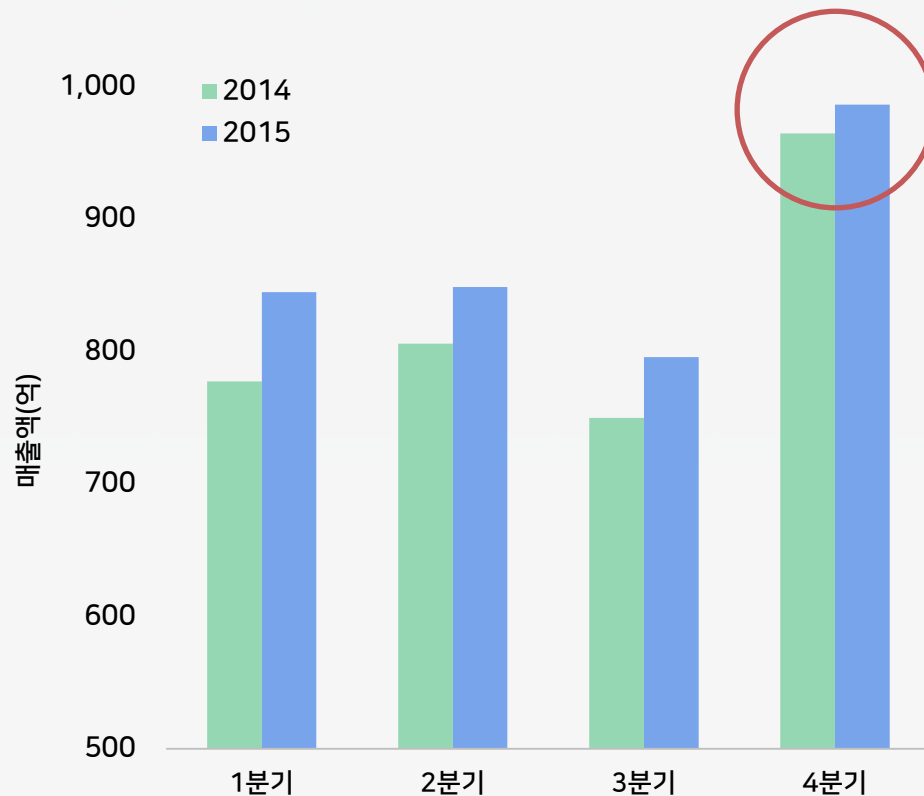


L사 2013년 대비 2015년 매출 **7.25% 하락**
오프라인 시장 점유율 **대형마트 3사 중 최저**



새로운 유통 활로 개척(온라인/모바일)
고객 맞춤 서비스 필요성 대두

기획 배경



2014년 매출액 3,296억 원
2015년 매출액 3,474억 원 (5.4% 신장)

전년도 대비 분기별 신장률

1분기 8.7%

2분기 5.3%

3분기 6.1%

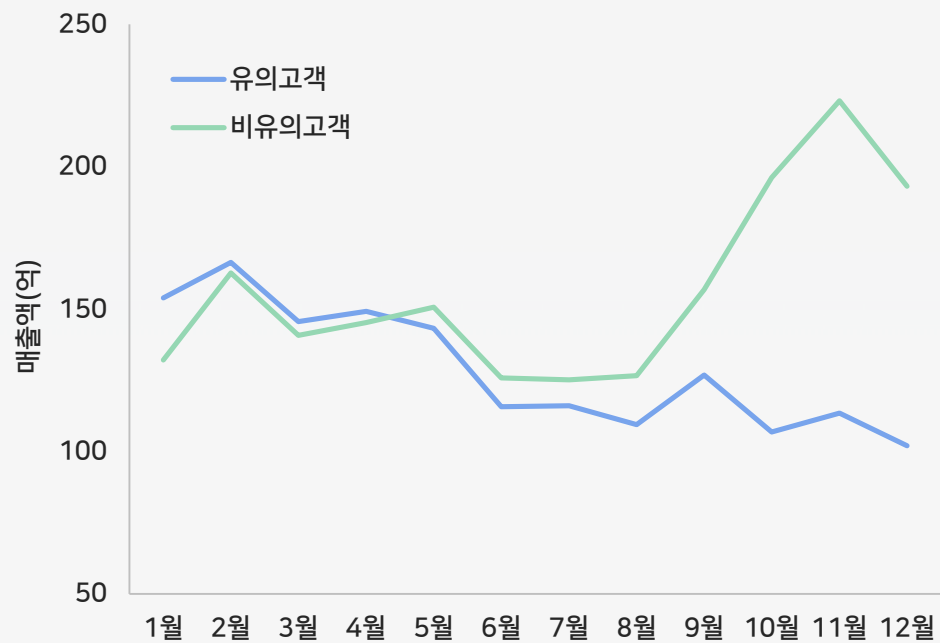
...

4분기 2.3%

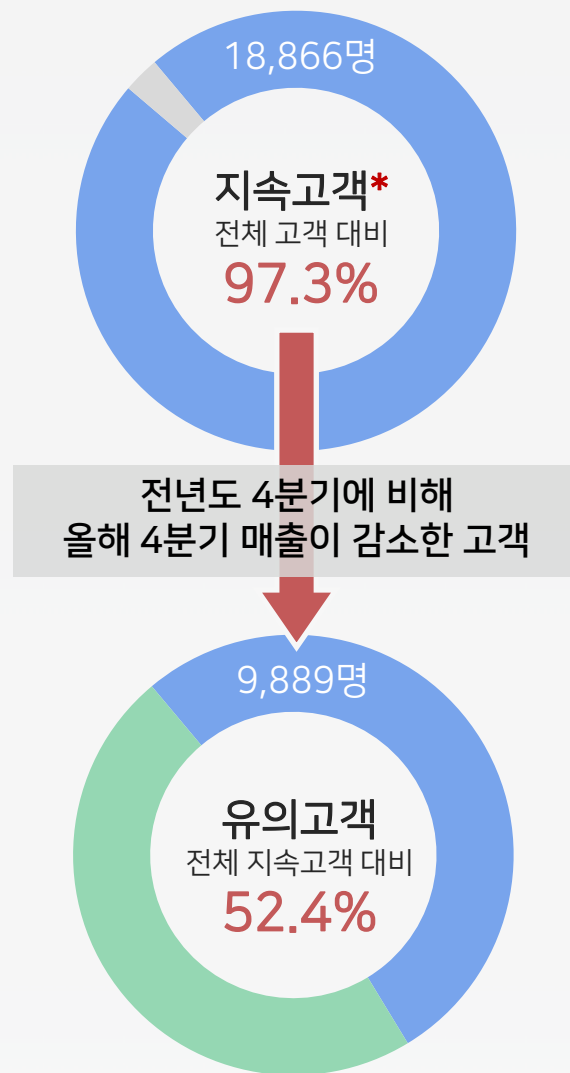
(1~3분기 평균 6.7%)

대상 고객 탐색

지속고객* 18,866명 대상
2015년 월별 매출



※ 계절성을 제거한 매출액



유의고객 핵심 ISSUE

1. 유의고객 중 과반수 이상 차지(52.4%)
2. 동년 1분기 대비 매출 **95억 하락**
3. 2015년 유의고객 매출 **294억 하락**

➡ 2015년 총 매출의 약 8.5%
(비유의고객의 경우 매출 441억 상승)



단위: (억)

	총 매출	전년 대비 증감액	전년 대비 신장률
1분기	430.6	28.5	7.1%
2분기	410.6	-18.5	-4.3%
3분기	361.3	-37.5	-9.4%
4분기	334	-266.7	-44.4%
계	1536.5	-294.2	-16.1%

지속적인 매출 하락세
잠재적 이탈 가능성 多

제 2 장

데이터 분석 및 모델링

- 활용 데이터 소개
- 모델링 환경
- 전체 프로세스 요약
- 변수 개발
- 모델 성능 테스트

L사 통합 고객 데이터	
Demo.	통합 고객 19,838명에 대한 성별, 연령 등 속성 데이터
구매내역 정보	2014~2015년 L사의 4개 계열사에서 이뤄진 구매 기록
상품분류	4개 계열사의 개별 상품분류 코드와 상품명 목록
경쟁사 이용	통합 고객 중 경쟁사 이용 내역이 있는 고객의 이용 기록
멤버십 이용	L사의 멤버십에 가입한 통합 고객에 대한 가입 정보
채널이용	통합 고객의 L사 온라인/모바일 유통 채널 이용 기록

L 그룹 계열사 정보

백화점, 마트, 슈퍼 등
다양한 상품을 취급하는
L 그룹의 계열사
(각 A사, B사, C사, D사로 표기)

데이터 탐색 및
전처리



python



유익고객
예측 모델 개발



XGBoost

 LightGBM

군집별 솔루션 제안



surpr!se

전체 프로세스 요약

1. 데이터 전처리 및 Feature Engineering



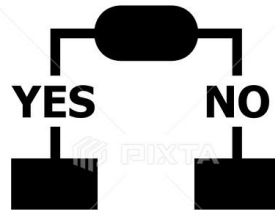
- 구매내역 데이터로부터 분류모델에 필요한 변수 데이터로 변환
- Oracle DB로 데이터 관리
- Python 라이브러리로 전처리



2. 분류 모델 구현 및 타겟 예측



- 기준일자 이전의 데이터로 유의고객이 될 고객들을 예측
- 일반적으로 많이 사용되는 분류 모델 중 성능이 좋은 모델을 선택



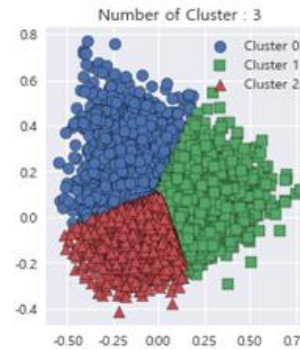
Decision tree

pixtastock.com - 40543714

3. 예측된 타겟의 군집화 및 특성 파악



- 유의고객이 될 고객들을 비슷한 특성을 가지도록 군집화
- 묶인 군집 별로 구매패턴을 파악



4. 군집 별 특성에 맞는 솔루션 제시



- 군집 특성에 맞는 솔루션 도출
- 추천 시스템을 이용한 개인화된 상품 추천 (아이템 기반 KNN 협업 필터링, 잠재요인 분석 협업 필터링)



정적 독립변수

단순지표 / 속성 데이터

분기별 구매금액

마지막 구매 후 경과일

주 구매 시기(월, 시간)

구매 시간대 별 평균 객단가

고객 속성 요소(연령대, 거주지)

...

동적 독립변수

변동 / 증감을 반영한 동적 지수

전년 대비 구매액 변화
(카테고리 별)

분기별 신장율

주말 방문 비율 변화

...

종속변수

2014년 4분기 대비

2015년 4분기 매출 감소 고객

9,889명

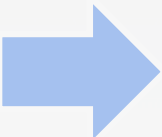


변수 개발 - 정적변수(1)

제휴사 통합 상품 카테고리 분류

네이버 쇼핑 카테고리 및
니스(NICE) 국제상품분류 체계를 참조

가공식품
신선식품
의약품/의료기기
교육/문화용품
디지털/가전
가구/인테리어
의류
전문 스포츠/레저
...



패션/의류
잡화
도서/문구/완구
시설/서비스
취미/스포츠/레저
화장/세정/세면
식료품
귀금속
생활
가전제품
명품
가구
공구류



고가 제품

카테고리 별 평균가의
2배 이상인 제품



저가 제품

카테고리 별 평균가의
0.5배 이하인 제품



중가 제품

고가와 저가를 제외한
나머지 금액대의 제품

소비 동향을 파악하기 위한 추가 분류 작업

변수 개발 - 정적변수(2)

시간1~4

폐점시간 (0시~8시) / 오전(9시~12시) / 오후(13시~20시) /
백화점 마감 후(21시~23시)로 범주화한 후 객단가 추출

경과일

14년 1분기~15년 3분기 사이의 구매일 중 가장 최근의 구매일자
기준일은 15년 10월 1일(15년 4분기의 시작일)

거주지역

구 우편번호 기준, 서울(0~99) 제외 나머지 지역은 지역별로 범주화(Label
Encoding)

변수 개발 - 동적변수

카테고리 별 15년 매출 비중

(편의상 '매출 비중' 용어를 CV로 통일)

카테고리 및 가격대 별 15년도 구매 금액 총합 ÷ 2년간 전체 구매 금액 총합
0.5 이상이면 15년 매출 \geq 14년 매출, 0.5 미만이면 15년 매출 $<$ 14년 매출

주말 방문 비율 변화

구매일자 항목에서 요일 추출

→ 분기 별 구매가 발생한 주말의 총합 ÷ 분기 내 주말의 일수 = 분기 별 주말 방문 비율

→ 7분기 주말 방문 비율 - 1분기 주말 방문 비율

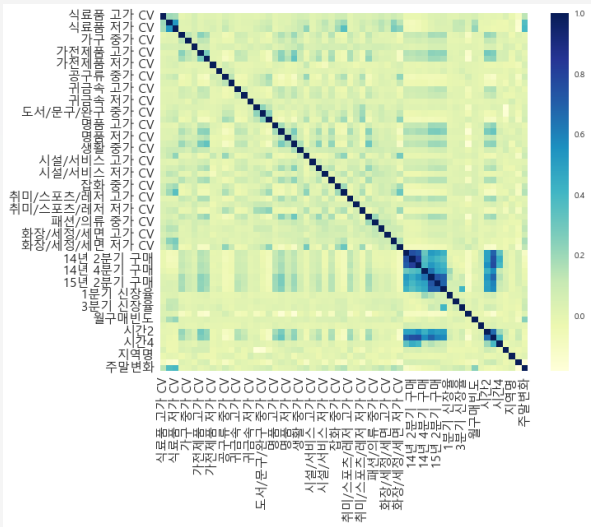
카테고리 별 상대등급

분기 별 구매액에 따른 등급화(1~5등급)

Q3 - 1.5*IQR 이상의 금액은 별도 분리 후 5개 등급으로 균등 분할

(분리한 고액 구매자는 모두 5등급으로 책정)

모델 설계



- 1. 히트맵을 통해
변수 간 상관관계 비교
→ 상관관계가 높은 변수는 통합
- 2. 변수의 추가/삭제를 통해
정확도 개선 확인
→ 정확도에 영향을 주지 않는
일부 변수는 삭제 처리
- 3. 1~7분기 학습, 8분기 예측
(학습 데이터를 검증용으로 한 번 더 분리)

최종 데이터셋

정적변수	동적변수
14년 4분기 구매	3분기 신장율
15년 2분기 구매	생활 고가 CV
시간3	패션/의류 고가 CV
14년 3분기 구매	공구류 증가 CV
월 구매 빈도	취미/스포츠/레저 저가 CV
14년 2분기 구매	취미/스포츠/레저 증가 CV
시간2	1분기 신장율
14년 1분기 구매	귀금속 고가 CV
15년 1분기 구매	화장/세정/세면 증가 CV
15년 3분기 구매	생활 저가 CV
시간4	가구 저가 CV
시간1	도서/문구/완구 저가 CV
연령대	잡화 저가 CV
시간구매빈도	시설/서비스 고가 CV
	귀금속 증가 CV
	화장/세정/세면 고가 CV

모델 성능 테스트(1)

Random Forest Classifier	
Accuracy	0.734
Precision	0.719
Recall	0.722
F1 Score	0.720
ROC-AUC Score	0.822
주요 변수(Feature Importance) TOP 3	
3분기 신장률	
14년 4분기 구매금액	
15년 3분기 구매금액	

Logistic Regression	
Accuracy	0.735
Precision	0.763
Recall	0.640
F1 Score	0.696
ROC-AUC Score	0.809
주요 변수(Feature Importance) TOP 3	
3분기 신장률	
14년 4분기 구매금액	
시간3(오후 시간대 구매 금액)	

LightGBM	
Accuracy	0.785
Precision	0.770
Recall	0.779
F1 Score	0.775
ROC-AUC Score	0.864
주요 변수(Feature Importance) TOP 3	
14년 4분기 구매금액	
3분기 신장률	
14년 3분기 구매금액	

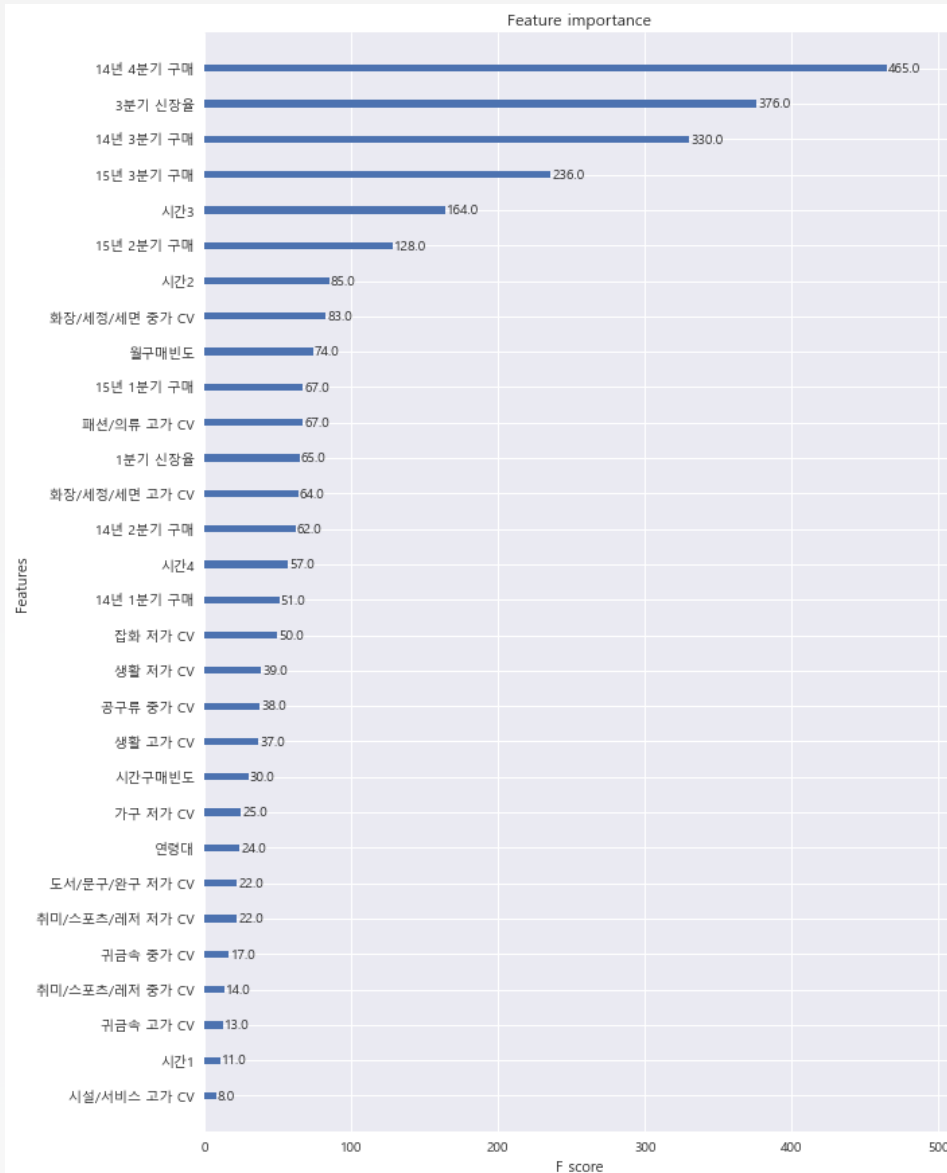
공통사항

3분기, 4분기 구매금액에 직접적인 영향 많이 받음
변수 하나에 중요도가 몰려 균일한 군집화가 어려움

차이점

동적변수에서의 중요도가 각각 다름
RF: 화장/세정/세면 증가 CV
LR: 생활 저가 CV
LGBM: 패션/의류 고가 CV

모델 성능 테스트(2)



XGBoost

Accuracy 0.785

Precision 0.772

Recall 0.775

F1 Score 0.774

ROC-AUC Score 0.866

주요 변수(Feature Importance) TOP 3

14년 4분기 구매금액

3분기 신장률

14년 3분기 구매금액

채택 사유

특정 변수의 중요도 편중이 비교적 완화

→ 군집 별 특성 파악에 용이

※ 예측한 유의고객 수 9,973 명

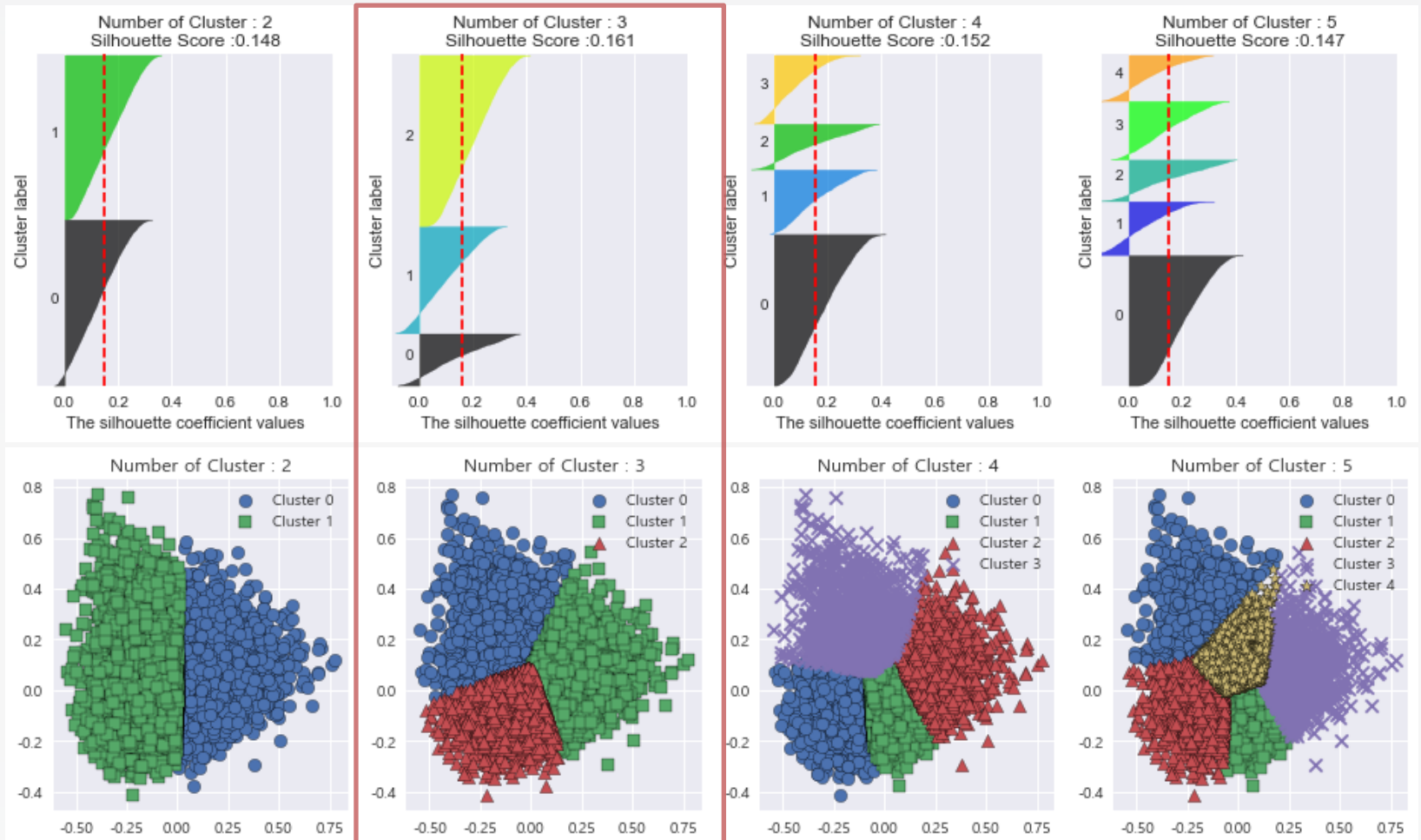
제 3 장

결과 해석 및 인사이트 도출

- K-평균 군집화
- 군집 분석 및 해석
- 마케팅 솔루션 및 상품 추천

K-평균 군집화

군집 별 균일하게 분배된 정도와 실루엣 계수를 고려하여 군집의 개수를 **3개**로 선정



군집 분석 및 해석

군집 0 (5,172명)

고객
구매
특성

- 672만, 448만
- ① 분기별 매출 평균이 타 군집 대비 36% 낮음
 - ② 타 군집 대비 인원수가 가장 많고, 구매 금액 감소폭이 대체로 적음
 - ③ 해당 군집은 중, 저가 상품군과 고가 상품군의 감소폭의 차이가 큰 편
예외적으로 식품의 경우 구매 금액이 증가하는 추세를 보임
 - ④ 군집 중 유일하게 경과일이 3일 미만 → 타 군집보다 내점 확률 높음

군집 변수 특성

잡화 고/중/저가 CV	0.248 / 0.522 / 0.585
패션·의류 고/중/저가 CV	0.318 / 0.489 / 0.497
식품 고/중/저가 CV	0.541 / 0.526 / 0.534
분기별 매출 평균	359만 7천 원
경과일	2.96일

전년 대비 3분기 감소율이 1%밖에 되지 않아,
조금 더 관리하면 매출 증가를 기대할 수 있는
잠재적 충성고객층

마케팅 솔루션

단번에 큰 금액대의 제품 권유보다는
지속적인 구매 습관 들이기가 중요하므로
1+1, N회 구매 이벤트 등 구매 횟수를
점차 늘릴 수 있는 단계별 마케팅 필요

‘살 만한’ 상품을 주로 사는 고객층으로.
평소에 자주 구매하던 제품을 추천

군집 중 가장 많은 인원수를 가졌고,
가장 특징이 희미한 군집이므로 멤버십 특전 등
고유한 소속감을 지니게 할 수 있는
제휴 서비스 활용

군집 분석 및 해석

군집 1 (3,233명)

고객 구매 특성

- 485만, 277만
- ① 타 군집 대비 1분기 구매액이 약 54% 높았으나 7분기 구매액은 34% 낮아짐
(군집 모두 7분기에 감소하는 경향이 있으나, 군집 1은 그 폭이 가장 크다)
 - ② 평균 구매액이 지속적으로 감소
1분기에서 7분기로 갔을 때 50% 이상 감소
 - ③ 주력 상품 중 하나인 패션/의류 에서 타 군집 대비 큰 감소폭을 보임

군집 변수 특성

패션 중/저가 CV	(타 군집 0.485 / 0.487) 0.357 / 0.395
14년 1분기 구매액	588만
15년 3분기 구매액	274만
분기별 신장률	0.82 / 0.47 / 0.68 (전 분기 신장률 감소)

3 군집 중 유일하게
전년 동분기 신장율이 감소로 하락세
이탈 고객이 될 가능성이 가장 큰 고객층

마케팅 솔루션

이탈 동향을 가장 크게 보이므로
상품 중심 마케팅 외의 할인, 이벤트 등
추가적 서비스 수단 마련

L사의 이용 전력이 충분히 있으므로
자주 구매한 상품 위주로 추천하여
구매 유도 및 서비스 이용 장려

번거롭게 내점하지 않아도 구매할 수 있는
온라인 채널 홍보(할인 행사와 연계)

군집 2 (1,568명)

고객
구매
특성

- ① 비교적 타 군집보다 고가 상품을 선호하는 편
- ② 14년 4분기 구매 평균 최고: 1458만 원
(타 군집 495만, 428만 / 군집 2가 3배 가량 높다)
- ③ 전체적으로 높은 분기 별 구매 금액대
- ④ 비수기라 할 수 있는 7월에도 구매 빈도 높음

군집 변수 특성

오전 9시 ~ 오후12시 구매 금액	12만 8천 원(타 군집 대비 2배)
오후 1시 ~ 오후 8시 구매금액	7만 2천 원(타 군집 대비 2배)
명품 고가/귀금속 고가/ 패션/의류 고가 CV	0.048 / 0.098 / 0.428 (타 군집 대비 2배)
최다 구매 월	7월
1분기 신장률	9.01%

시간대에 구매받지 않고
여유롭게 쇼핑하는 것을 선호하는
구매력 있는 고정 고객층

마케팅 솔루션

촉박한 타임세일보다는
구매 시 주어지는 지속적인 이익을 강조
구매 금액에 따른 부가적 혜택 제공

객단가를 높게 뽑아낼 수 있으므로
VIP와 같은 등급제를 통해
고가 상품군으로의 구매 유도

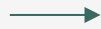
충성도가 비교적 높으므로
계절 상품을 개발하여 7월의 구매력을
4분기까지 견인할 수 있는 방안 마련

상품 추천 프로세스

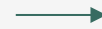
BASE - 개인화된 추천시스템



전체 고객의 구매내역을 통해
상품별(소분류명 기준) 구매지수 산정



고객 번호와 해당
군집 특성을 기반으로
상품 중분류 채택



두 가지 방식의 협업 필터링 방식으로
추천 리스트 생성
(아이템 기반 KNN 협업 필터링 /
잠재요인 분석 협업 필터링)

ADDITIONAL - 월별 인기상품



특정 월의 과거 구매내역을 통해
해당 월의 인기상품을 리스트화



Base 추천 리스트에서
인기 상품에는 가중치를 주어
추천리스트의 상단에 오도록 함

상품 추천 예시 - 군집 1, BASE

군집1

주력 상품 중 패션/의류 중고가 상품과 잡화 증가에서
타 군집 대비 큰 감소폭을 보이므로
해당 중분류 위주로 추천 리스트를 구성

```
target_id = "'00047'"
uid = change_target_id(target_id)
prodcl_list=['패션/의류 고가','패션/의류 증가','잡화 증가']
date = 20151001
```

예시) 군집1 구성원 중 한 명인 47번 고객

RMSE: 0.9837
RMSE: 0.9315
SVD의 RMSE가 Knn Baseline의 RMSE보다 낮아 SVD 기반 추천 시스템을 적용합니다.

	uid	iid	r_ui	est	중분류명	소분류명
0	00047	A040902	3.874683	3.571420	패션/의류 고가	디자이너부틱
1	00047	C160201	NaN	2.008890	잡화 증가	롤티슈
2	00047	A040214	NaN	1.923473	패션/의류 고가	캐릭터캐주얼
3	00047	A041001	NaN	1.621504	패션/의류 고가	모피
4	00047	A040215	NaN	1.427547	패션/의류 고가	수입캐릭터
5	00047	A040302	NaN	1.333262	패션/의류 고가	커리어
6	00047	B300601	NaN	1.292490	패션/의류 증가	교복
7	00047	B290405	NaN	1.156989	패션/의류 증가	여성캐주얼
8	00047	A060114	0.669042	1.150178	패션/의류 증가	L/C 아웃도어
9	00047	A040224	0.351357	1.113335	패션/의류 증가	global SPA

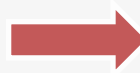
BASE 만 적용한 추천 상품 리스트

핵심 키워드: 10월의 인기상품을 결합

감소될 것이 예상되는 4분기의 첫 달인 10월 추천 상품으로
작년 10월 인기상품에 가중치를 주어
추천 리스트의 상단으로 끌어올린다.

ADDITIONAL 적용 전 추천 상품 리스트

	uid	iid	r_ui	est	중분류명	소분류명
0	00047	A040902	3.874683	3.571420	패션/의류 고가	디자이너부틱
1	00047	C160201	NaN	2.008890	잡화 증가	롤티슈
2	00047	A040214	NaN	1.923473	패션/의류 고가	캐릭터캐주얼
3	00047	A041001	NaN	1.621504	패션/의류 고가	모피
4	00047	A040215	NaN	1.427547	패션/의류 고가	수입캐릭터
5	00047	A040302	NaN	1.333262	패션/의류 고가	커리어
6	00047	B300601	NaN	1.292490	패션/의류 증가	교복
7	00047	B290405	NaN	1.156989	패션/의류 증가	여성캐주얼
8	00047	A060114	0.669042	1.150178	패션/의류 증가	L/C 아웃도어
9	00047	A040224	0.351357	1.113335	패션/의류 증가	global SPA





ADDITIONAL 적용한 추천 상품 리스트

	uid	iid	r_ui	est	중분류명	소분류명
0	00047	A040902	3.874683	3.458691	패션/의류 고가	디자이너부틱
2	00047	A040222	NaN	2.424069	패션/의류 고가	영 캐릭터
3	00047	A040215	NaN	2.390606	패션/의류 고가	수입캐릭터
1	00047	C160201	NaN	2.078152	잡화 증가	롤티슈
5	00047	A040214	NaN	1.958152	패션/의류 고가	캐릭터캐주얼
11	00047	B180205	NaN	1.710365	잡화 증가	가정용화장지
4	00047	A040401	NaN	1.520550	패션/의류 고가	트래디셔널
18	00047	A020201	NaN	1.386906	패션/의류 증가	선글라스(특정)
22	00047	A040225	NaN	1.310689	패션/의류 고가	영 컨템포러리
6	00047	A060166	NaN	1.278738	패션/의류 증가	스포츠의류

조원 소개



CEO 어정호

 <https://github.com/fish-ho>
 fishho97@gmail.com



이슈 탐색과 그에 따른 솔루션을 고안함에 있어서 변수 추출과 생성의 중요성을 크게 실감했습니다. 여러 시행착오를 겪으며 데이터 탐색 및 예측 모델링의 요소 기술과 당위성 판단력을 상승시킬 수 있는 값진 시간이었다고 생각합니다.



수석 개발자 조남현

 <https://github.com/MiddleJo>
 chonh0531@gmail.com



하나하나 뜯어 보기엔 그 크기가 너무나 방대한 빅데이터이지만, 머신 러닝을 통해 간편화하여 자세하게 분석할 수 있었습니다. 더 정확하고 자세한 분석을 위해 다양한 skill을 익히는 것이 무척 중요함을 다시금 깨달았습니다.



개발 및 디자인 최지원

 <https://github.com/JadeWednesday>
 plasticmelody@gmail.com



내가 올바르게 하고 있는 게 맞는지 의구심이 들 때에 곁에서 격려해 주는 사람들이 있다는 건 정말 큰 복이라는 걸 새로이 깨닫게 되었습니다. 여러모로 뜻깊은 3주였습니다.



데이터마이너 김민성

 <https://github.com/nycticebus0915>
 nycticebus0915@gmail.com



처음 해 본 프로젝트인 만큼 저에겐 조금 힘든 과정이었고, 데이터 분석 작업은 흐름을 이해하는 것도 어려웠지만 조원들의 도움으로 이해하고 넘어갈 수 있었습니다. 이번 과제에서 익힌 역량을 바탕으로 다음 과제엔 더 잘 해 보고 싶습니다.

조경윤

 <https://github.com/kkyxxn>
 kkyxxn@gmail.com



잘 모르는 분야에 대해 이해하게 되는 계기가 되어서 좋은 경험이었다고 생각합니다. 또한 데이터 모델링에 대해 부족함을 느껴서 좀 더 공부해야 할 것 같습니다.



감사합니다.

질문사항이 있다면 말씀해 주세요.