매출 감소 예측 모델 활용

맞춤 솔루션

디.매버릭스

Data Mavericks

팀원 : 김용훈, 옥유리, 임수현, 장인혁, 정슬기

목차

(index)

0. 개요

0_1 팀원소개 및 담당업무

0_2 프로젝트 기간별 수행 절차

0_3 분석환경 및 활용 라이브러리

Introduction of D.Mavericks

정 슬기
총괄, 발표, 데이터 전처리

김 용훈
군집화 작업 및 특징 정의, 데이터 전처리

옥 유리
탐색적 분석 및 도메인 정보수집, PPT

임 수현
머신러닝 개발, 데이터 전처리, 시각화

장 인혁
마스터 테이블 개발(SQL), 머신러닝 성능개선

프로젝트 기간 별 수행 절차

구분	기간	활동	비고
사전기획	5/22(월) ~ 6/2(금)	- 프로젝트 기획 및 주제 선정 - **기획안 작성**	+ 6/2 1차 기획안 발표
데이터 가공	5/29(월) ~ 6/9(금)	- 기준정보 정의 및 **데이터 정제** - 마스터 데이터 테이블 완성	
머신러닝 개발	6/12(월) ~ 6/16(금)	- **머신러닝** 결과 확인 및 피드백 - 피처 엔지니어링 보수 작업	
군집화	6/19(월) ~ 6/23(금)	- 군집화 작업 - 군집 특징 정의 및 **맞춤 서비스 개발**	
서비스 개발	6/26(월) ~ 6/28(수)	- **고객 맞춤 서비스 개발 및 마무리**	
총 개발기간	**약 6주** (5/22(월) ~ 6/29(목))		

분석 환경 및 활용 라이브러리

분석 언어 : **파이썬 , SQL**
분석 환경 : **주피터 , 오라클**

활용 라이브러리

데이터 전처리 및 분석 : **Numpy, Pandas**

데이터 시각화 : **Matplotlib, Seaborn**

머신러닝 : **사이킷런**

대대분류 통합 기준 – 유통물류진흥원

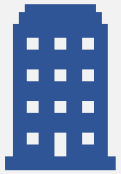
1. 데이터탐색

1_1 목적과 활용데이터, 기대효과

1_2 외부요인 분석

1_3 탐색적 분석

목적과 활용 데이터 설명



기업중심

고객중심에 착안

고객 구매 패턴 파악

감소고객 예측

맞춤형 솔루션 제공

목적과 활용 데이터 설명

L사의 **14~15년도 데이터**를 분석하여 고객의 **구매패턴 파악**.
구매 감소 고객 **예측모델을 통해** 고객을 패턴별로 분류 후,
구매 패턴 별 요구되는 니즈를 해결하기 위해 **고객 중심 마케팅 솔루션을 제언**.

- 고객 정보 관련 데이터
  - 고객번호
  - 성별
  - 연령대
  - 거주지
  - 멤버십 가입년월
  - 계열사별 관련 모바일/APP
  - 온라인 쇼핑몰 이용 횟수

- 고객 구매 데이터
  - 계열사
  - 영수증번호
  - 구매일자
  - 구매시간
  - 구매금액
  - 대분류코드
  - 중분류코드
  - 소분류코드
  - 중분류명
  - 소분류명

목적과 활용 데이터 설명 4개사 통합에 대한 이유

	제휴사	영수증번호	대분류코드	중분류코드	소분류코드	고객번호	점포코드
1	A	01254637	1	0101	A010103	04325	031
2	A	01254647	1	0101	A010101	00177	031
3	A	01254653	1	0101	A010103	04325	031
4	A	01254657	1	0101	A010101	12808	031
5	A	01252839	1	0101	A010101	05824	012
6	A	01200942	1	0101	A010101	07887	020
7	A	02757986	1	0101	A010101	15148	010
8	A	02757988	1	0101	A010101	02043	010
9	A	02757989	1	0101	A010101	02231	010

	제휴사	영수증번호	대분류코드	중분류코드	소분류코드	고객번호	점포코드
1	B	08725275	1	0101	B010108	13714	044
2	B	08725190	1	0101	B010105	16204	044
3	B	08724743	1	0101	B010108	15910	044
4	B	08750670	1	0101	B010102	02767	044
5	B	08750677	1	0101	B010108	02041	044
6	B	07095291	1	0101	B010102	14913	013
7	B	08431316	1	0101	B010106	01412	034
8	B	08919529	1	0101	B010108	13293	048
9	B	08919750	1	0101	B010110	08057	048

제휴사 : A,B,C,D로 구분

대분류코드, 중분류코드가 같지만

소분류코드가 다름

=〉 제휴사 4개를 통합할 수 있는

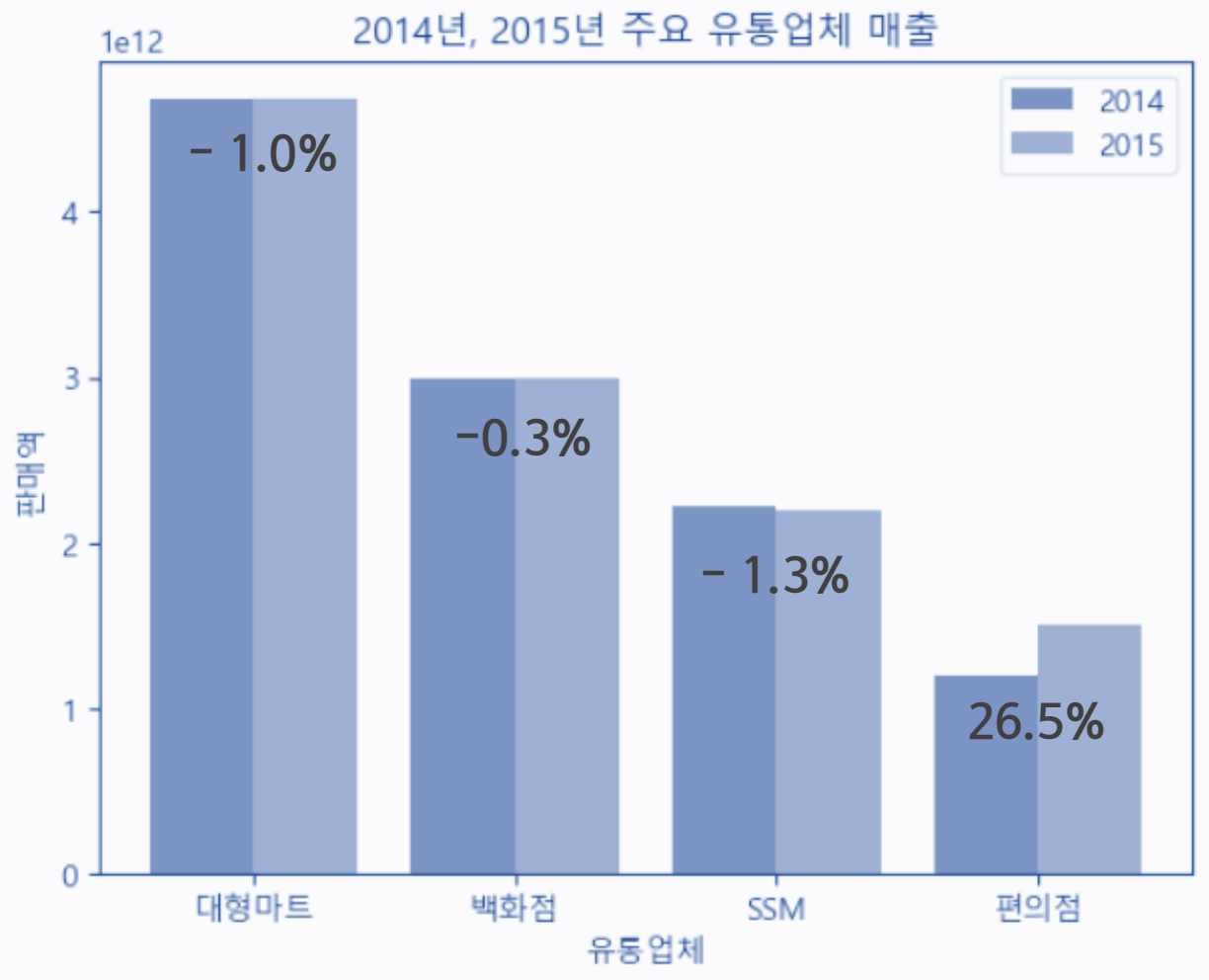
통합분류체계 마련

기대효과

감소고객 중 유의 구간을 설정하여 **10% 이상의 명확한 감소**가 확인 되는 **감소고객을 정의**.

-> 감소 고객 뿐만 아니라 유지 고객도 확인가능.

-> **맞춤형 솔루션 제공**하여 확실한 매출 상승 기대.



SSM: 기업형 슈퍼마켓

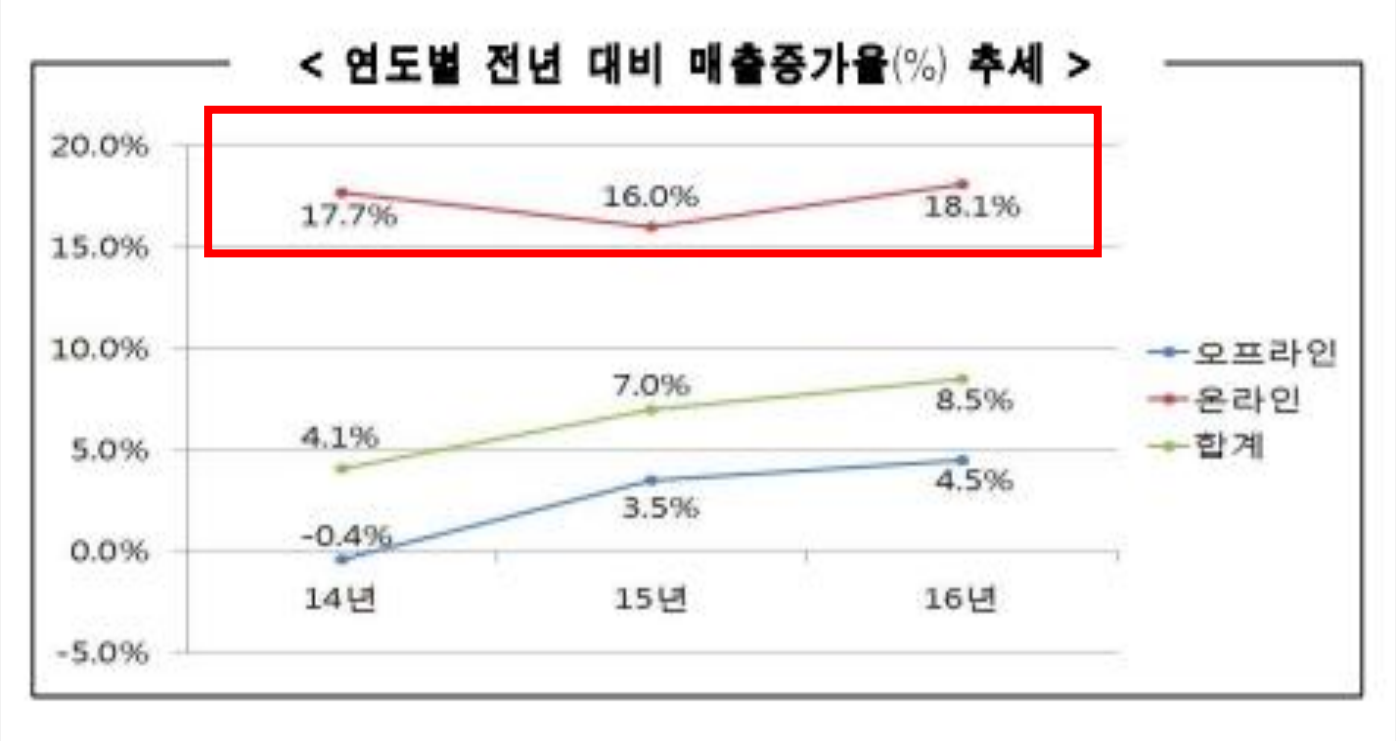
외부요인 분석 i

주요 유통업체 매출
전체적인 **감소 추세**

외부적 감소요인

- 메르스
- 전 년 대비 높은 기온으로 인한 겨울상품 판매 부진
- 경쟁 업태 성장 (**온라인 쇼핑**, 아울렛…)

출처: 산업통상자원부



출처: 산업통상자원부

외부요인 분석 ii

주요 유통업체 매출 추세

오프라인 유통업체와 상반된

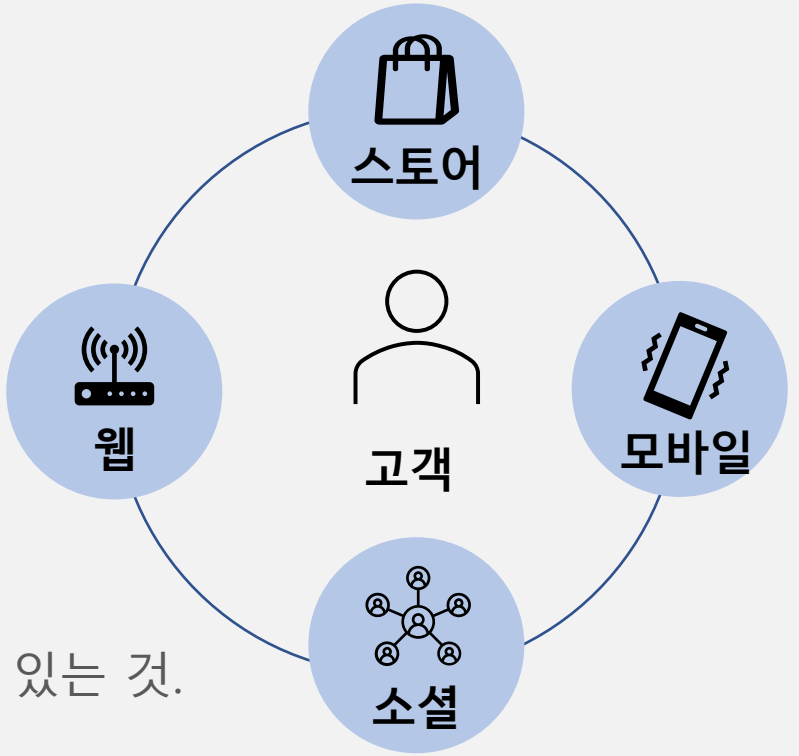
온라인 유통업체의 높은 비율과 비중 증가

외부요인 분석 iii

옴니채널

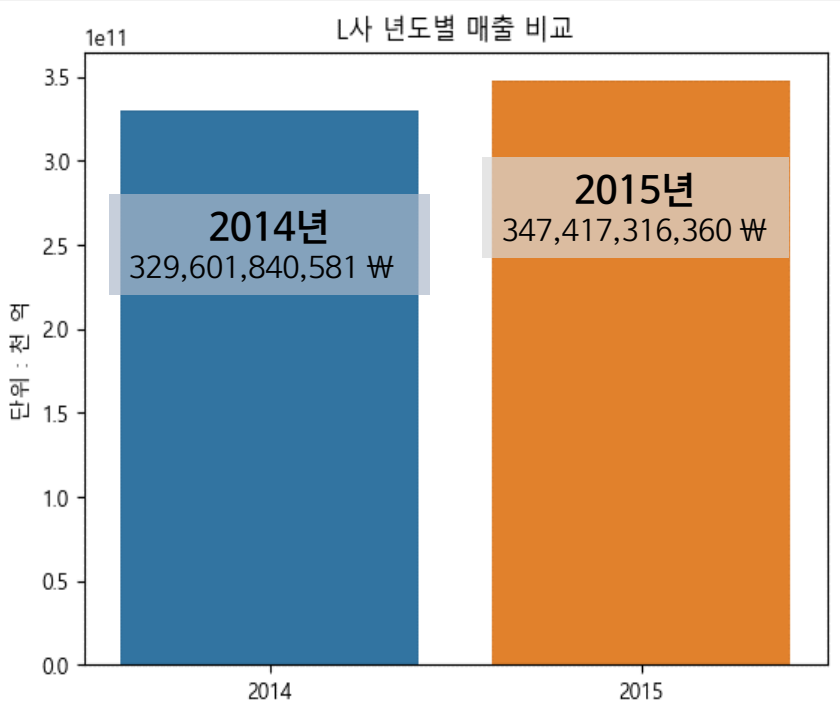
- 모든 것을 뜻하는 옴니(Omni) 와
제품의 유통경로를 의미하는 채널(channel)의 합성어.

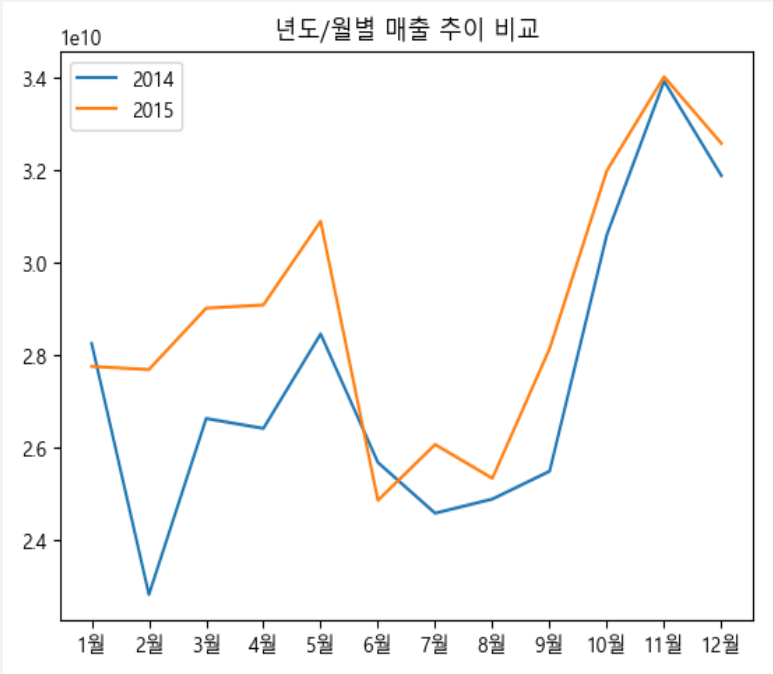
소비자들이 **시간과 장소에 구애받지 않고 서비스를 이용**할 수 있는 것.

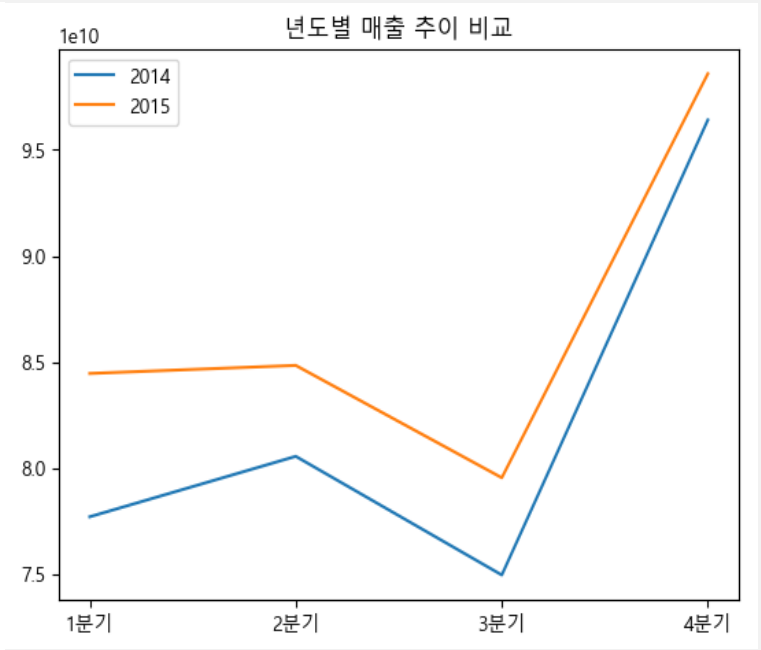


옴니채널의 중요성 더욱 확대

탐색적 분석 ⅰ – 매출





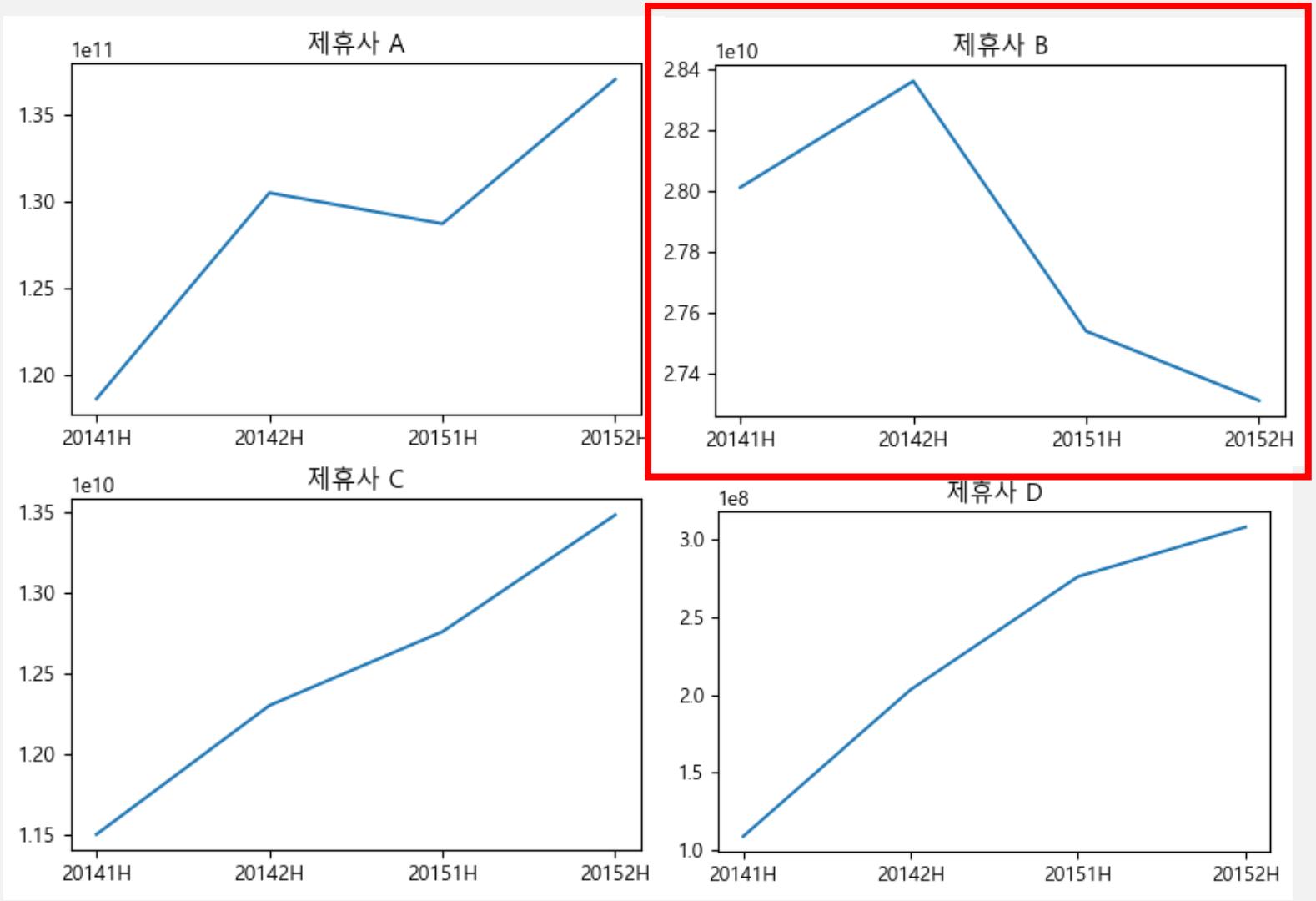


- 2014년 대비 2015년 매출 : **약 5.4% 상승**
- 2014, 2015년도 총 매출 합 : 677,019,156,941원 (약 6천 7백억 원)

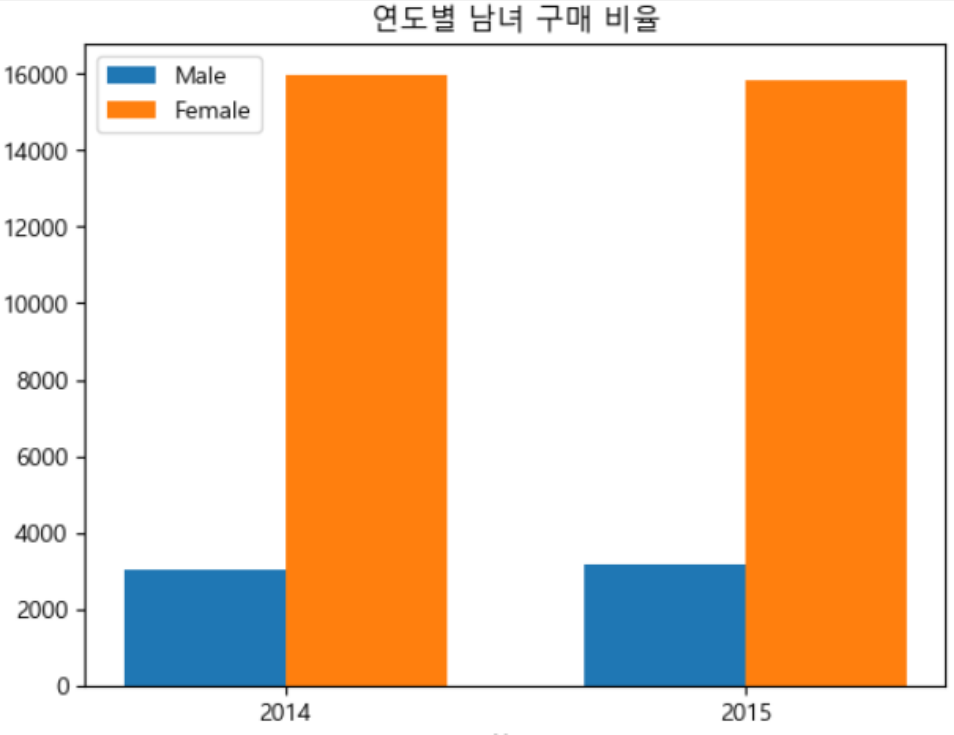
탐색적 분석 ii – 제휴사

반기별 매출 추세

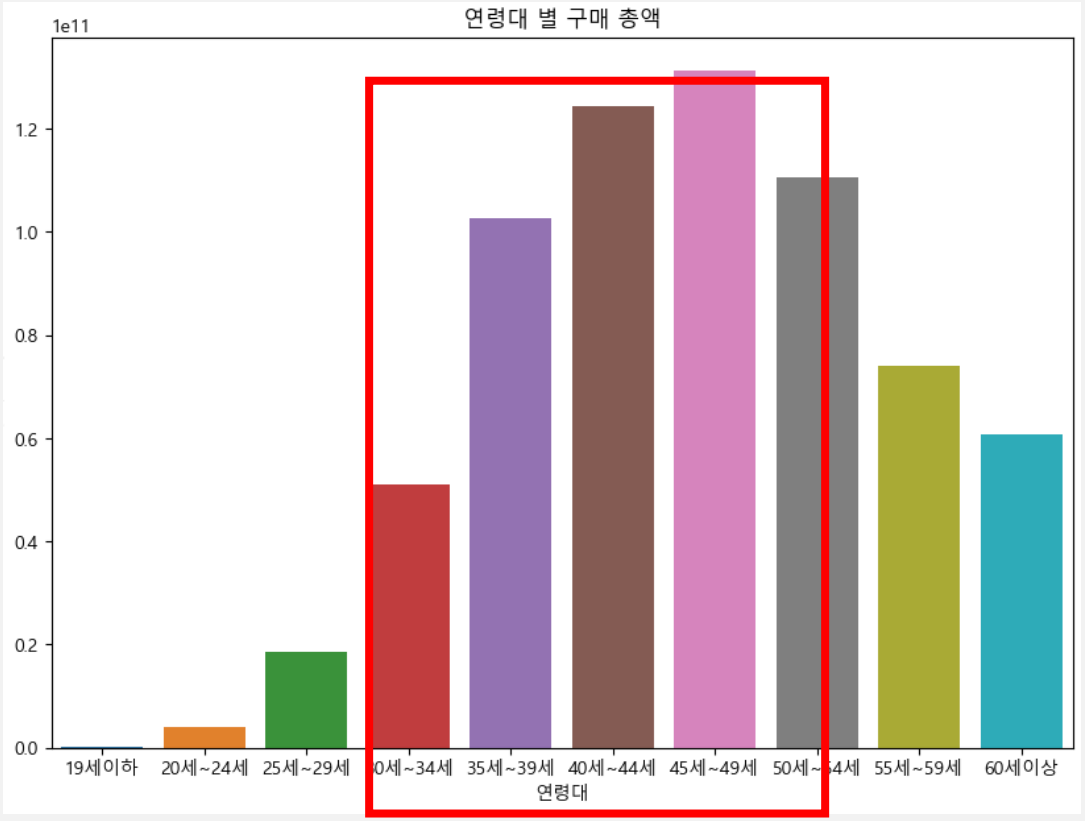
제휴사 **B만** 유일하게
매출 하락 추세



탐색적 분석 iii – 고객



2014년, 2015년 모두 남녀 비율 2:8



35세 ~ 54세의 연령대의 높은 구매 추세

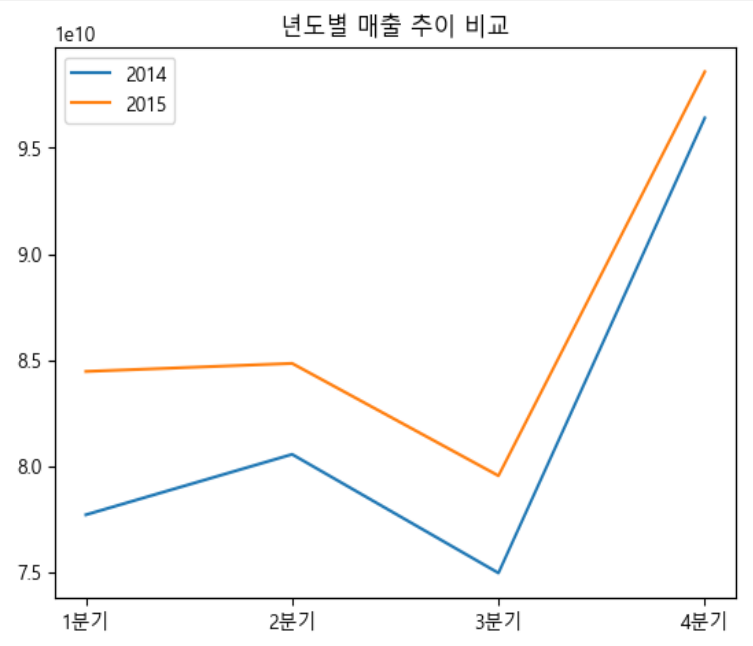
2. 과제 정의

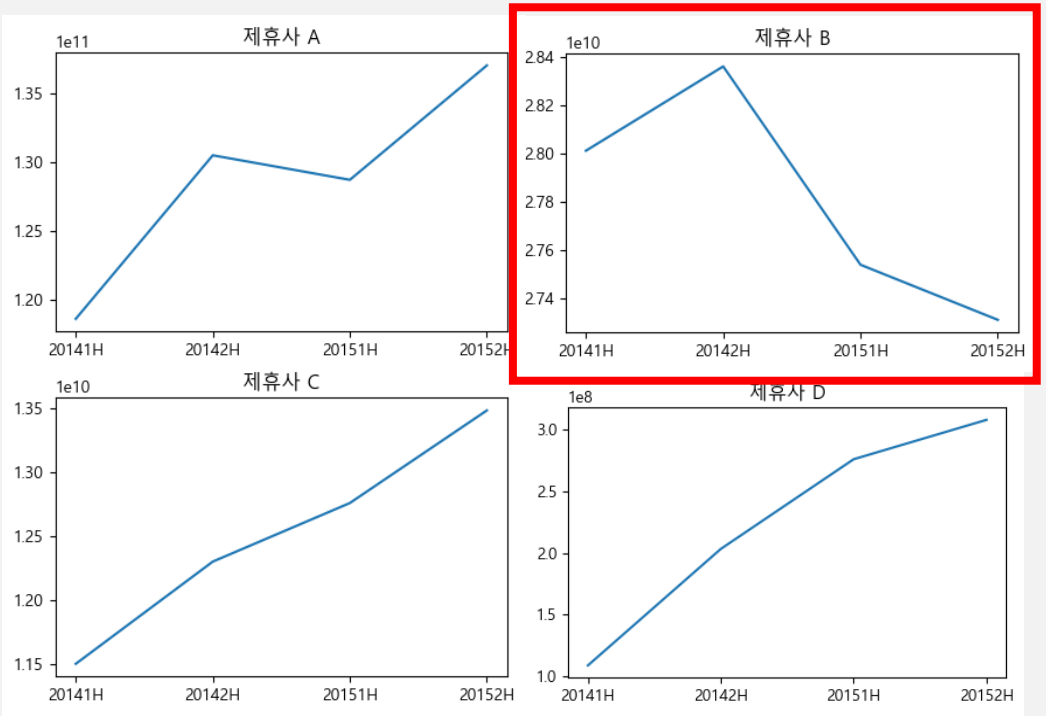
2_1 기준정보 정의

2_2 피처 엔지니어링

2_3 학습을 위한 데이터셋 분리와 변수 설정

감소고객 확인


년도별 매출 추이 비교
1e10
2014
2015
9.5
9.0
8.5
8.0
7.5
1분기
2분기
3분기
4분기


제휴사 A
1e11
1.35
1.30
1.25
1.20
20141H
20142H
20151H
20152H
제휴사 B
1e10
2.84
2.82
2.80
2.78
2.76
2.74
20141H
20142H
20151H
20152H
제휴사 C
1e10
1.35
1.30
1.25
1.20
1.15
20141H
20142H
20151H
20152H
제휴사 D
1e8
3.0
2.5
2.0
1.5
1.0
20141H
20142H
20151H
20152H

?

감소고객 – 분기별 증감 수치화, 계절성 제거 테이블

		2014_1	2014_2	2014_3	2014_4	2015_1	2015_2	2015_3	2015_4
	고객번호	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1	00001	9141590	10783765	10781550	23334762	9302985	8595380	7806580	5295938
2	00002	14123470	10199760	14841180	7080920	18481480	9310130	13842600	12001220
3	00003	302782	548433	648188	394366	279310	312970	421017	258101
4	00004	1309643	3732463	2434410	1823100	1737292	1919645	2127112	1610145
5	00005	3145330	2314820	1827290	398440	2124430	810500	0	2616260
6	00006	2166006	3115284	3163273	2950816	3619234	2769821	3518094	3203889

	고객번호	Q2_DIFF	Q3_DIFF	Q4_DIFF	Q5_DIFF	Q6_DIFF	Q7_DIFF	Q8_DIFF	구매차이합	구매감소여부
1	00001	1642175	-2215	12553212	-14031777	-707605	-788800	-2510642	-3845652	1
2	00002	-3923710	4641420	-7760260	11400560	-9171350	4532470	-1841380	-2122250	1
3	00003	245651	99755	-253822	-115056	33660	108047	-162916	-44681	1
4	00004	2422820	-1298053	-611310	-85808	182353	207467	-516967	300502	0
5	00005	-830510	-487530	-1428850	1725990	-1313930	-810500	2616260	-529070	1
6	00006	949278	47989	-212457	668418	-849413	748273	-314205	1037883	0

기준정보

	고객번호	Q1	Q8	구매감소여부	증감 %
1	00001	9141590	5295938	1	57.9%
2	00002	14123470	12001220	1	84.9%
3	00003	302782	258101	1	85.2%
4	00004	1309643	1610145	0	122.9%
5	00005	3145330	2616260	1	83.1%
6	00006	2166006	3203889	0	147.9%

1분기 금액을 100으로 설정.

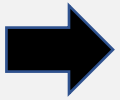
10% 구간을 둬서 90 이하를 감소고객으로 정의

피처 엔지니어링 핵심내용

- 코드 통합작업
- 선호제휴사 선정
- 연령대 통합
- 지역 통합

피처 엔지니어링 i

베이직케어,
클렌징,
남성케어,
여성케어,
유제품,
조미료,
통조림,
두유,
차,
스낵, 비스킷,
명품
주방용품
스포츠
...



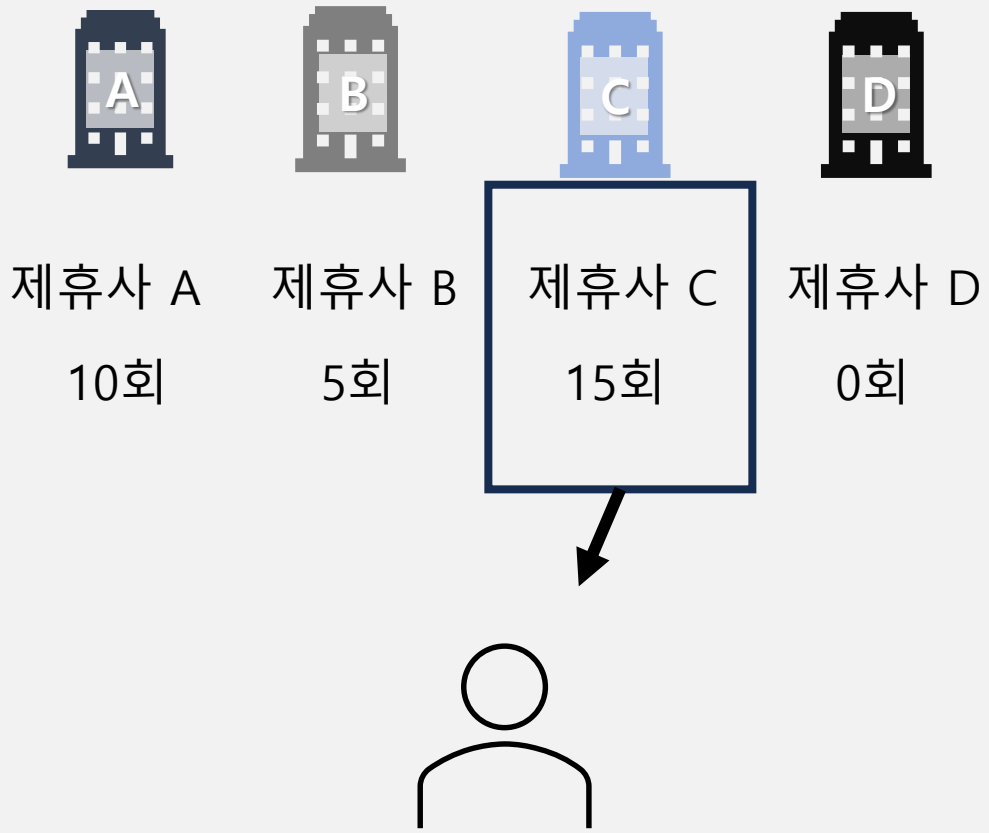
가공식품
신선식품
일상용품
의약품/의료기기
교육/문화용품
디지털/가전
가구/인테리어
의류
전문스포츠/레저
패션잡화
기타

유통물류진흥원의 상품분류표준 참고

중분류코드 700개

→11개의 대대분류 통합

피처 엔지니어링 ii



선호 제휴사

→ **최다 빈도 이용** 제휴사

피처 엔지니어링 iii

연령대 → **5개로 통합**

구분	연령
학생층	24세 이하
청년층	25세- 34세
중년층	35세 - 44세
장년층	45세 - 54세
노년층	55세 이상

지역 → **17개로 통합**

1. 서울 010 ~ 109
2. 경기 110 ~ 209
3. 인천 210 ~ 239
4. 강원 240 ~ 269
5. 충북 270 ~ 299
6. 세종 300 ~ 309
7. 충남 310 ~ 339
8. 대전 340 ~ 359
9. 경북 360 ~ 409
10. 대구 410 ~ 439
11. 울산 440 ~ 459
12. 부산 460 ~ 499
13. 경남 500 ~ 539
14. 전북 540 ~ 569
15. 전남 570 ~ 609
16. 광주 610 ~ 629
17. 제주 630 ~ 639

데이터 셋 분리 iii

<table>
<tr><td>1분기</td><td>2분기</td><td>3분기</td><td>4분기</td><td>5분기</td><td>6분기</td><td>7분기</td><td>8분기</td></tr>
<tr><td colspan="7">학습-검증 데이터 세트</td><td></td></tr>
<tr><td colspan="6">독립변수</td><td>종속변수</td><td></td></tr>
<tr><td colspan="6">1분기 - 6분기 데이터</td><td>1분기-7분기
구매감소유무</td><td></td></tr>
<tr><td></td><td colspan="7">평가 데이터 세트</td></tr>
<tr><td></td><td colspan="6">독립변수</td><td>종속변수</td></tr>
<tr><td></td><td colspan="6">2분기 - 7분기 데이터</td><td>2분기-8분기
구매감소유무</td></tr>
</table>

※독립변수

- **고객 속성 변수** : 고객이 **고유하게 가지고 있는 속성**을 의미하는 변수
- **구매 패턴 변수** : 분기별 **변화하는 구매 패턴**을 의미하는 변수.

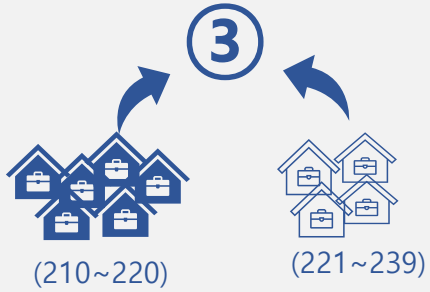
고객 속성 변수

성별

남녀 성별 구분

연령대

유사한 연령대를 묶어 5개로 구분



지역

각 지역코드에 할당된 지역을 배정.

결측치는 점포코드를 참고하여 채움.

멤버십 가입 유무

고객별 제휴사 멤버십 가입유무를 확인

채널 이용 횟수

고객별 온라인 채널 총 이용횟수를 확인

최근방문일

마지막 분기의 **최종날짜를 기준**으로

최근 방문일까지의 일수를 계산

구매 패턴 변수 등급화 방법

기준	14_1 가공	14_2 가공	14_3 가공	14_4 가공	14_5 가공	14_6 가공
금액	417,324 ₩	582,003 ₩	421,417 ₩	187,660 ₩	544,202 ₩	550,251 ₩
랭크화(등수)	11,094	7,193	12,654	17,781	9,936	9,307
등급화	**6등급**	**4등급**	**7등급**	**10등급**	**6등급**	**5등급**
분기 등급 변동	0	-2	3	3	-4	-1

변화율 : 등급 변동 절댓값의 합 → 12

순증감율 : 마지막 분기와 첫 분기의 등급 차 → -1

3. 머신러닝 활용

3_1 머신러닝 피처

3_2 모델링 및 성능 평가
(RF, DT, LR, LGBM, XGBoost)

3_3 군집화(K-Means Clustering)

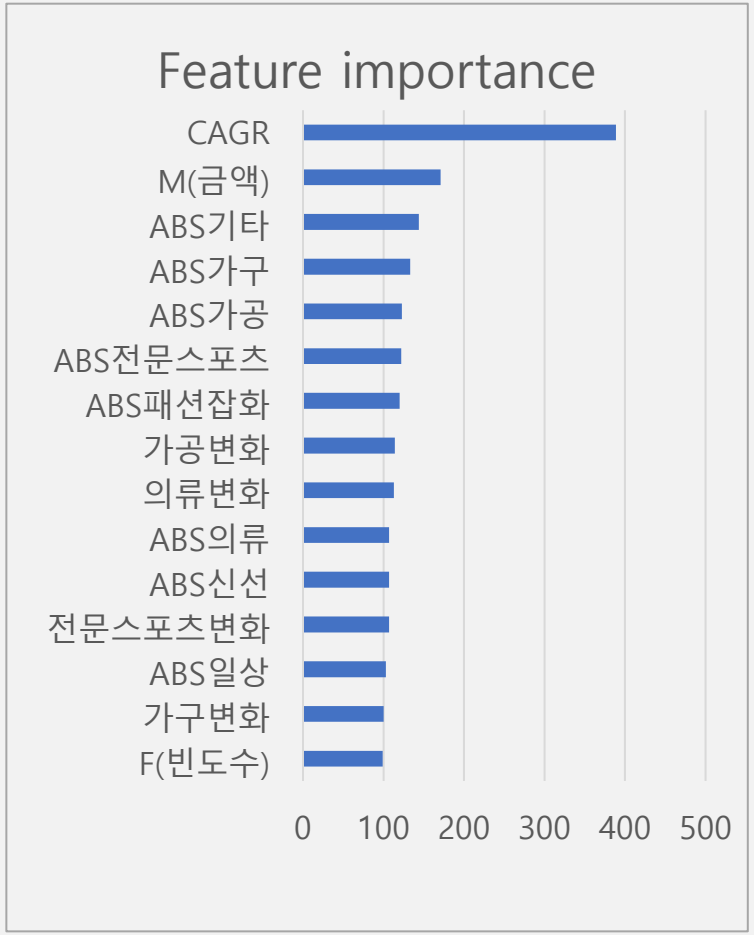
머신러닝 피처

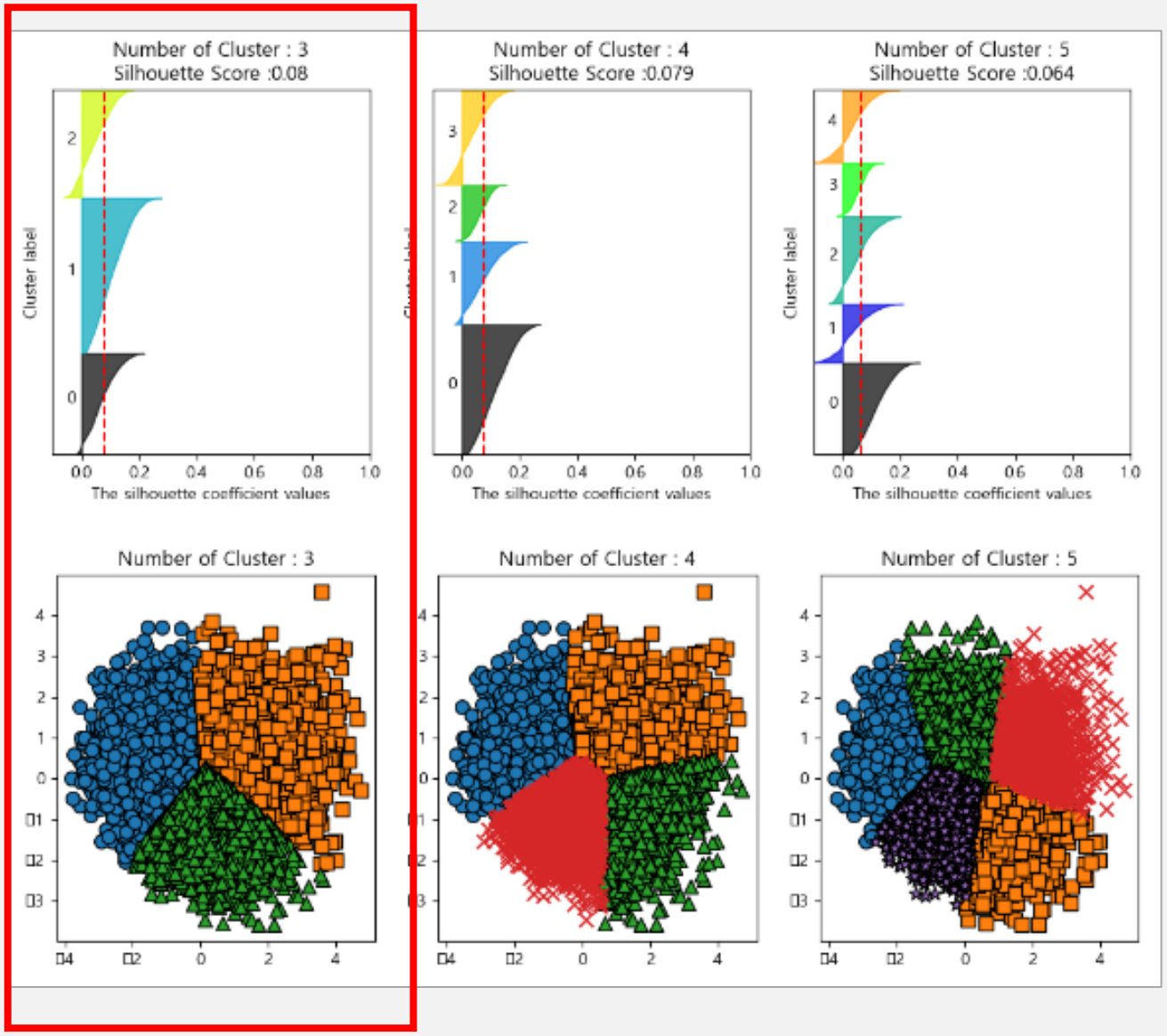
- CAGR (연평균 성장률)
- 제품별 변화율
- 제품별 순증감률
- RFM지수

모델링 및 성능 평가

	Random Forest	Decision Tree	Logistic Regression	Light GBM	XGBoost
Accuracy	0.7231	0.7209	0.7137	**0.7260**	0.7256
Precision	0.6582	0.6530	0.6372	0.6639	0.6669
Recall	0.6091	**0.6128**	**0.6244**	0.6077	0.5976
F1 score	0.6327	0.6128	0.6307	0.6346	0.6304
ROC AUC	0.7021	0.7016	0.6978	**0.7049**	0.7028

군집분석 / 특징


Feature importance
CAGR
M(금액)
ABS기타
ABS가구
ABS가공
ABS전문스포츠
ABS패션잡화
가공변화
의류변화
ABS의류
ABS신선
전문스포츠변화
ABS일상
가구변화
F(빈도수)
0
100
200
300
400
500


Number of Cluster : 3
Silhouette Score :0.08
Number of Cluster : 4
Silhouette Score :0.079
Number of Cluster : 5
Silhouette Score :0.064
Cluster label
The silhouette coefficient values
Number of Cluster : 3
Number of Cluster : 4
Number of Cluster : 5

4. 마케팅 제언

고객 군집별 마케팅 제언

고객 개인화 상품 추천

고객 특성별 마케팅 제언

구매 감소 고객 예측
6,920 명

군집 1 2,685명	신선 일상 가공 위주 (편의품) M등급 8,9,10등급 비중 높음
군집 2 2,700명	의약 비중이 조금 더 높다.(편의품) M등급 4,5,6,7 등급 비중 높음
군집 3 1,535명	가구 디지털 명품 위주(선매품, 전문품) M등급 1,2,3등급의 비중 높음

고객 군집 i

고객 군집 ii

고객 군집 iii & 개인화 맞춤 시스템

Surprise~!

고객 개인화 상품 추천

Q & A

Github_

정슬기
https://github.com/wjdtmfrl

김용훈
https://github.com/dydgns94

옥유리
https://github.com/yul77

임수현
https://github.com/fortis001

장인혁
https://github.com/jang-in-hyeok

디.매버릭스

Data Mavericks