

재구매율 예측을 통한 구매물품 예측 및 솔루션 제안



CONTENTS,

01 02 03 04 03

개요

자료구성
사용 모듈
데이터 목록

주제선정

EDA
주제 선정
분석 과제

재구매 예측

변수속성
모델링

Clustering

사용 변수 및 속성
Clustering 특성
군집별 특성 분석

마케팅 제안

추천 시스템
장바구니 분석

01

개요

자료구성, 사용 모듈, 사용 데이터

01 개요

자료 구성

- 분석대상(Analysis Target) : 2014년 ~ 2015년(2년간) L그룹 4개 계열사에서 구매한 고객 일부 발취
- 제공범위(Scope of Offer) : L그룹 4개 계열사의 구매이력(계열사 정보 비공개), 고객 성향을 파악할 수 있는 데이터 제공

사용 모듈

- Google Colab



- mljar-supervised



- surprise



사용 데이터 목록

구매내역 정보 / 채널(온 오프라인) 이용
경쟁사 이용 / 멤버십 여부
데모(고객정보) / 상품 분류

01 개요

구매내역 정보

제휴사(ABCD)/영수증번호/대분류코드/중분류코드/소분류코드/
고객번호/점포코드/구매일자/구매시간/구매금액

채널(온/오프라인) 이용

2015년 10월 ~ 12월, 총 3개월 분

경쟁사 이용

L사 4개 계열사 별 경쟁사 이용년월 정보

멤버십여부

멤버십에 가입한 고객

데모(고객정보)

19,383명의 고객 정보

상품분류

대분류, 중분류, 소분류로 나뉘어진 상품 분류 중 소분류 데이터 사용



고객의 구매 패턴 파악을 위해
고객 중심 데이터로 가공

데이터 가공

기존 대분류 코드/대분류명 및 중분류명 통합 및 보완

→ '유통상품지식뱅크' 상품분류기준

→ 대분류: 12 / 중분류: 127 / 소분류: 3519

원래_중분류명 수정_중분류명

샌들슬리퍼	신발
액세서리	헤어액세서리
모자	헤어액세서리
핸드백	가방
피혁로탈	신발
피혁브랜드	패션의류
피혁로탈	패션의류
장갑	장갑
가방브랜드	가방
잡화멀티샵	가방
핸드백	가방
준보석	패션소품
준보석	패션소품
우산/양산	신발
준보석	패션소품
우산/양산	기타의류
우산/양산	패션소품
썬글라스	안경
여행용품	여행용가방/소품
가방브랜드	여행용가방/소품
우산/양산	패션소품
액세서리	패션소품
액세서리	패션소품



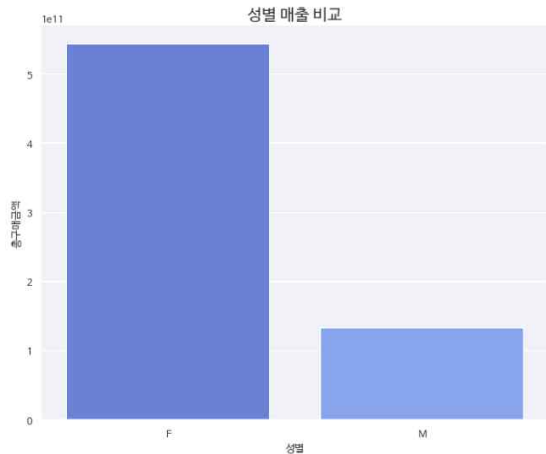
대분류코드(12)	대분류명	중분류명(127)
1	가공식품	가공식품세트류, 건강식품, 과자류, 기타가공식품, 냉동, 농산가공, 대용식재료, 분말류, 빵류/떡류, 수산가공, 어육/연식품, 영유아식품, 유제품, 음료, 절임/발효, 조미료, 주류, 즉석/편의, 차류, 축산가공, 커피/코코아, 통조림/병
2	신선식품	기타신선식품, 농산물, 신선식품세트류, 육류, 축산물
3	일상용품	건강의료용품, 공구, 기타일상용품, 기타주방용품, 생활용품, 세탁용품, 수집품, 애완용품, 영유아용품, 위생용품, 의약품, 일용잡화, 자동차용품, 주방용품, 청소/욕실용품, 화장품, DIY용품
5	의약품/의료기기	의료기기, 의료용품, 의약품
6	교육/문화용품	게임기/타이틀, 기타교육/문화용품, 기타인테리어, 문구/사무용품, 비디오, 서적, 수집품, 악기, 영상물, 영유아용품, 완구, 음반
7	디지털/가전	가전/주변기기, 건강/생난방가전, 게임기/타이틀, 기타디지털/가전, 네트워크장비, 생활가전, 영상가전, 음향가전, 저장장치, 주방가전, 카메라/캠코더용품, 컴퓨터/주변기기, 통신가전, 학습기기, 휴대폰액세서리,
8	가구/인테리어	거실가구, 기타인테리어, 수납가구, 아동/주니어가구, 원예용품, 인테리어소품, 장식가구, 주방가구, 침구단품, 침구세트, 침실가구, 커튼/블라인드, DIY용품
9	의류	기타의류, 스포츠웨어, 언더웨어, 유아동의류, 패션의류, 홀웨어
10	전문스포츠/레저	골프, 기타레저용품, 낚시, 농구, 등산, 물러/보드, 배구, 배드민턴, 수영, 스케이트, 스키/보드, 스포츠웨어, 야구, 요가/필라테스, 자전거, 축구, 캠핑, 탁구, 테니스/스쿼시, 헬스
11	패션잡화	가방, 모자, 벨트, 스카프, 시계, 신발, 안경, 양말, 여행용가방/소품, 유아동패션잡화, 장갑, 지갑, 패션소품, 헤어액세서리,
12	명품	명품
99	기타	기타, 식품, 행사

02

주제 선정

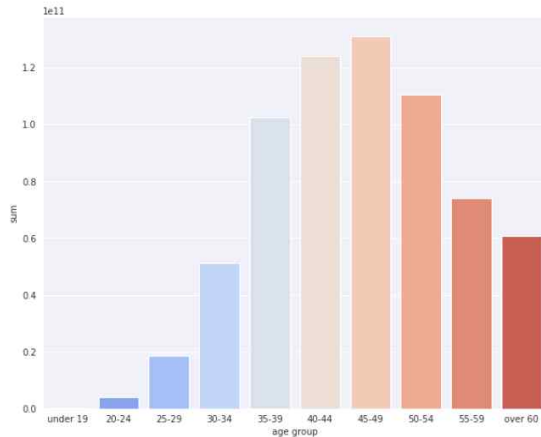
EDA/주제선정

성별 매출 비교



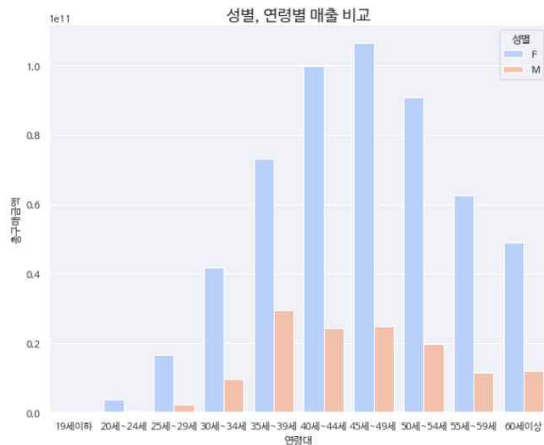
남성에 비해 **여성고객**의 총매출액이 압도적으로 높음

연령별 매출 비교



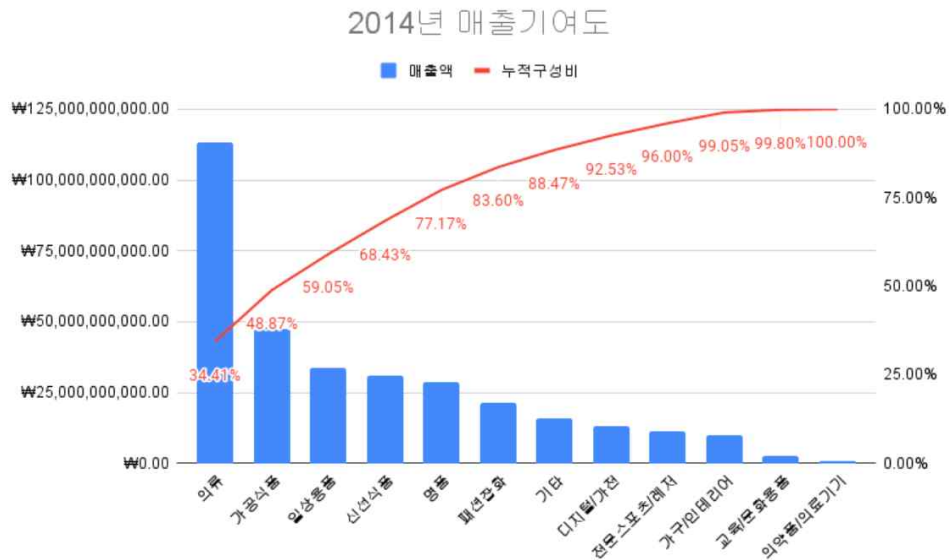
40대 고객의 매출액이 높은편이며,
연령대가 높을수록 매출액이 큰편

≧ 성별_연령별 매출 비교



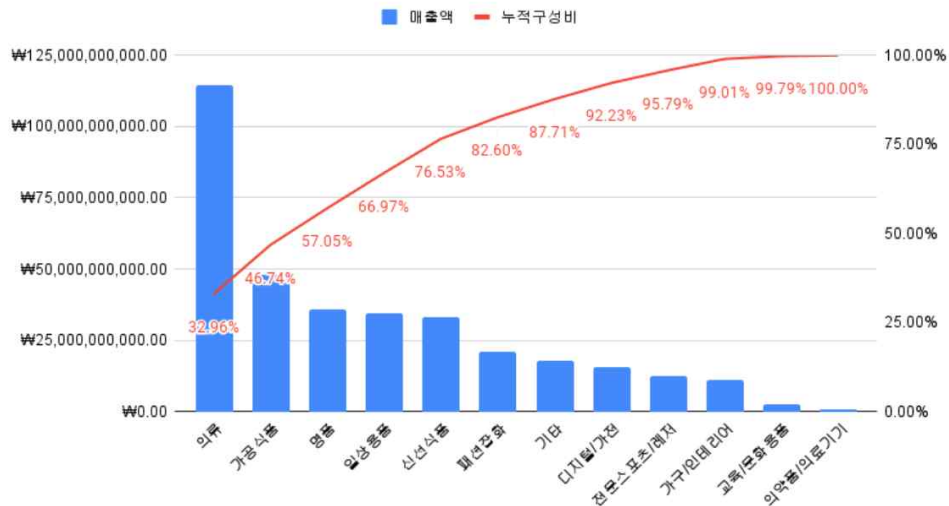
- 남성고객의 경우에는 연령별 매출액 그래프와 다르게 35세부터 39세의 매출액이 가장 크다는 특징이 있음
- 여성고객의 경우에는 40세부터 49세의 고객군의 매출이 큼

2014년 카테고리별 매출 기여도

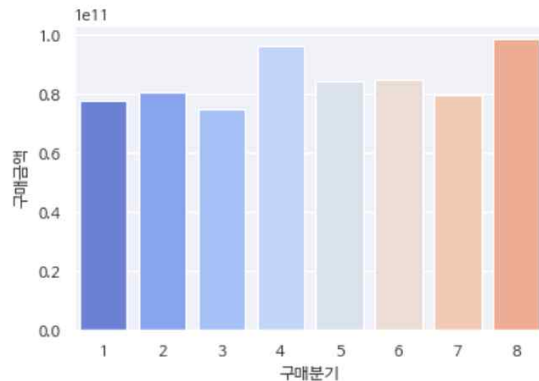
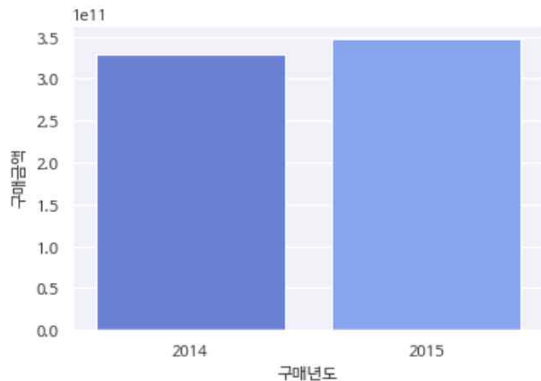


> 2015년 카테고리별 매출 기여도

2015년 매출기여도



➤ L사의 반기별 총 매출액 + 년도별 총 매출액



계절(분기)에 따른 증감의 변동이 있으나 14년도 보다 15년도의 총 매출액이 높음

“매출 증가 요인 탐색”

> 신규(혹은 복귀) 고객의 수



분기별_고객분류

14 2 유입고객 141

14 3 유입고객 73

14 4 유입고객 44

15 1 유입고객 14

15 2 유입고객 14

15 3 유입고객 8

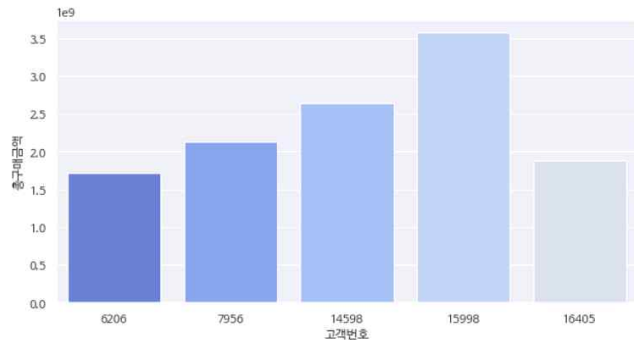
15 4 유입고객 3

dtype: int64

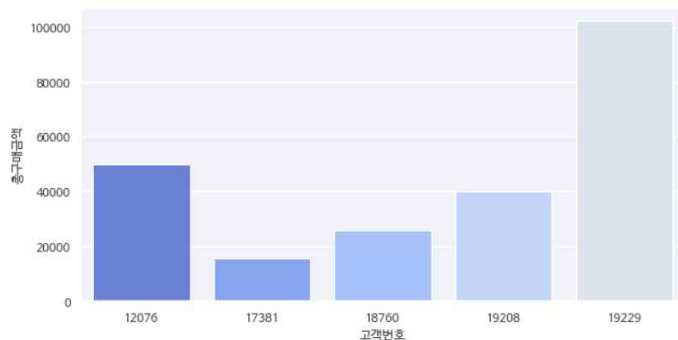
14년도 2분기의 유입 고객의 수(141명)는 유의미 할 수 있는 수치이나,
2년간의 전체 신규고객은 총 297명으로 총 고객 대비 신규고객의 수가 적음

» 기존고객 분석

영수증을 비교하여 같은 상품을 반복 구매하는 횟수를 통해 고객별 매출 기여도 확인



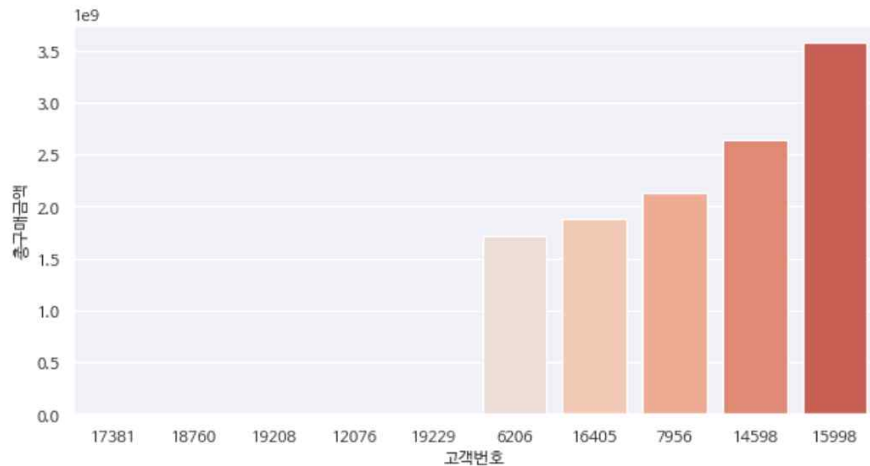
같은 상품을 반복 구매하는 횟수가 많은 고객 중
임의로 선택된 5명의 매출



같은 상품을 반복 구매하는 횟수가 적은 고객 중
임의로 선택된 5명의 매출

➤ 기존고객 분석

영수증을 비교하여 같은 상품을 반복 구매하는 횟수를 통해 고객별 매출 기여도 확인



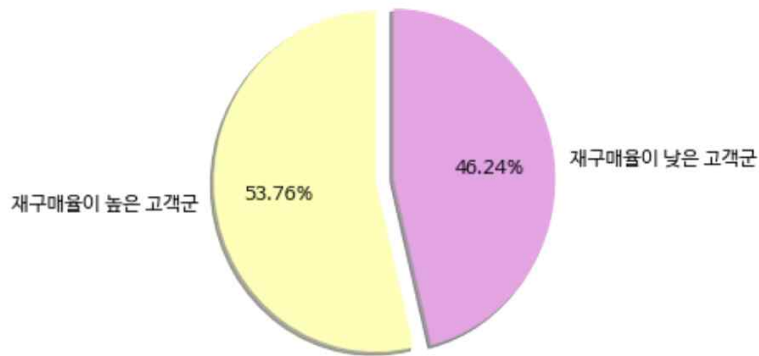
반복 구매 횟수가 적은 고객과 많은 고객(총 10명)을 함께 그린 그래프

➤ 같은 상품의 반복적 구매로 인한 매출액과 총매출액 비교

같은 상품을 반복적으로 구매하는 것이 전체 매출에서 어느정도의 영향을 끼치는가?



2년간의 총 매출



“기존 고객유지와
재구매율을
높이는 것이
중요과제”

반복구매가 많은 고객들의 매출기여도는 53.76%로
반복구매가 적은 고객들에 비해서는 높은 편

➤ 실제 마케팅 기법을 탐사: **고객유지 마케팅**

고객 유지 마케팅(Retention Marketing)이란 무엇인가?

고객 유지 마케팅은 그저 스쳐지나가는 수많은 마케팅 유행어 중 하나로 들릴 수 있겠지만 모든 비즈니스가 성장과 지속가능성 차원에서 매우 중요한 전략입니다.

즉, 고객 유지 마케팅은 기존 고객의 반복 구매에 중점을 둡니다. 따라서 유지 마케팅의 주요 목표는 반복 구매 고객이 더 자주 구매를 하여 고객 생애 가치(Customer Lifetime Value, CLV)를 높이는 것입니다.

고객유지 마케팅이 필요한 이유?

신규 고객을 확보하는 데는 기존 고객을 유지하는 것보다 훨씬 더 많은 비용이 소요
반복 구매 고객은 또한 매출의 상당 부분을 차지
많은 연구에 따르면 반복 구매 고객이 기업 매출의 40% 이상을 창출할 수 있으며,
고객 유지율을 5% 증가시킬 때 수익이 25-95% 증가할 수 있다고 한다.



주제 선정

재구매율 예측을 통한
구매물품 예측 및 솔루션 제안

주제 선정

- 재구매율을 분석하여
재구매 가능성이 낮은 고객과 높은 고객을 예측
- 군집화를 통하여 각 고객 군의 구매 패턴 파악 및 상품 추천을 제공

03

재구매 예측

변수속성 / 모델링

03 변수 속성

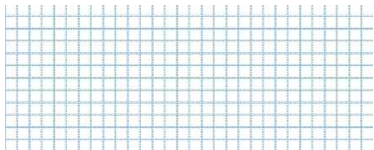
> 정리한 변수속성

cust id	고객번호
total ordered goods	총 구매 상품
days between first and last order	첫구매와 마지막 구매 사이의 간격
days between orders	구매일간의 간격
number of orders	구매 횟수
kinds of ordered goods	구매 상품 종류
mean of goods in order	구매 상품 갯수의 평균
sex	고객 성별
age	고객 나이
address	고객의 주소
label	재구매율

03 변수 속성

» 재구매율이란?

제품을 구매한 고객이 다시 해당 쇼핑물을 찾아 구매하는 비율

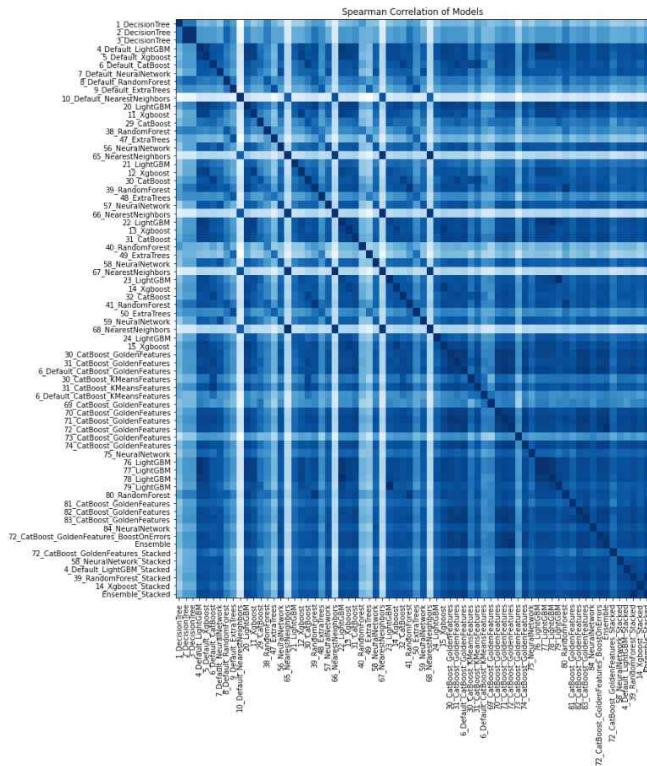


» 재구매율 라벨링

- 고객의 구매이력을 분석하여 구매 이력이 이어지는가 아닌가로 라벨링
- 이를 활용해 고객별 물품별로 구매한 재구매율을 모두 더한 뒤 전체 대비 재구매한 것의 비율을 재구매율로 삼는다.



03 재구매율 예측을 위한 모델링



Train Dataset

2014년 ~ 2015년 3분기 데이터

Test Dataset

2015년 4분기 데이터

Model

AutoML 중 **mljar-supervised** 패키지 사용

- Iterated Tasks 자동화
- 알고리즘 선택 및 파라미터 최적화
- 튜닝 자동화
- 최적 모델 추천

03 Train-Val 검증 Dataset(상위 5개)

» 총 84개의 모델의 성능을 비교

	Accuracy	Precision	recall	F1-score	auc
CatBoost (depth : 6)	0.795078	0.764211	0.722012	0.744486	0.842025
CatBoost (depth : 8)	0.792602	0.757708	0.725038	0.740299	0.858716
LightGBM	0.791312	0.761258	0.713944	0.73771	0.860846
Xgboost	0.787597	0.781183	0.667045	0.729744	0.860801
Forest	0.782232	0.755966	0.690872	0.721279	0.838168

Accuracy가 아닌 F1-Score를 중점으로 모델을 선택한 이유

- 고객이 연속해서 물건을 구매하는 경우가 전체 구매 대비 매우 적은 수로 데이터의 불균형이 발생
- 데이터의 불균형시 쉽게 과적합 현상을 보이는 Accuracy보다는 정밀도¹와 재현율²의 수치가 적절하게 조합된 평가지표인 F1-Score를 중점으로 삼는 것이 좋겠다는 결과를 도출

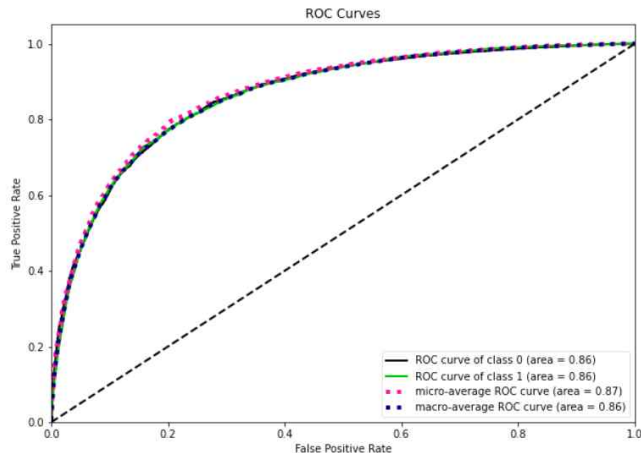
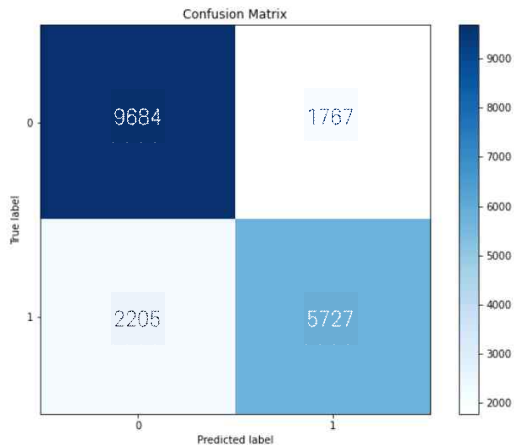
정밀도¹(Precision) : 모델이 Positive라고 예측한 것들 중 실제 Positive였던 확률 / 재현율²(Recall) : 실제 Positive 중 모델이 Positive라고 예측한 확률

03 Train-Val 검증 Dataset(상위 5개)

➤ **CatBoost** 0.744486 (depth : 6)

범주형 변수의 예측모델에 최적화된 모델
(범주형 데이터를 처리하는 새로운 방법 제시)

- 다른 GBM에 비해 과적합이 적음
- 범주형 변수에 대해 모델의 정확도와 속도가 높음
- encoding 작업 없이 모델의 input 가능

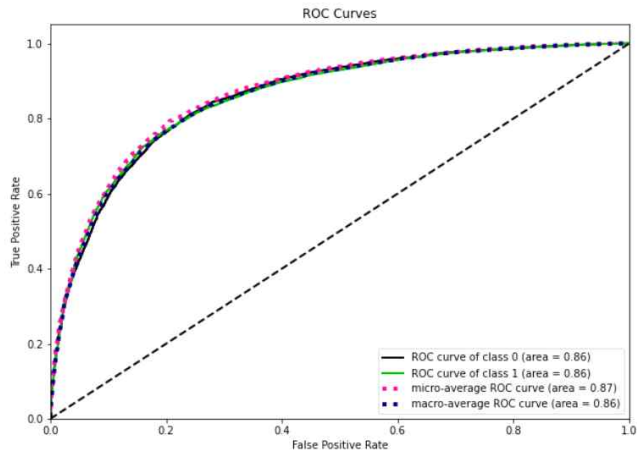
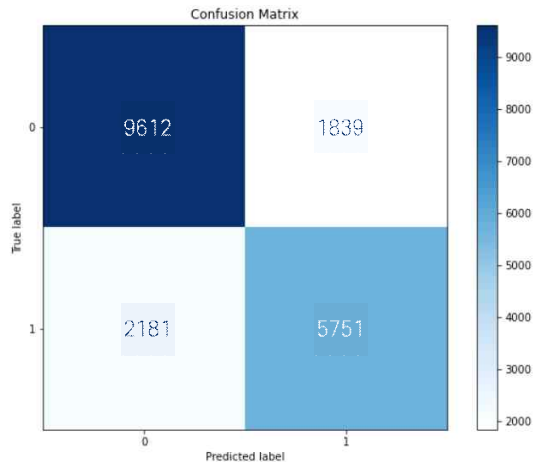


03 Train-Val 검증 Dataset(상위 5개)

➤ **CatBoost** 0.740299 (depth : 8)

범주형 변수의 예측모델에 최적화된 모델
(범주형 데이터를 처리하는 새로운 방법 제시)

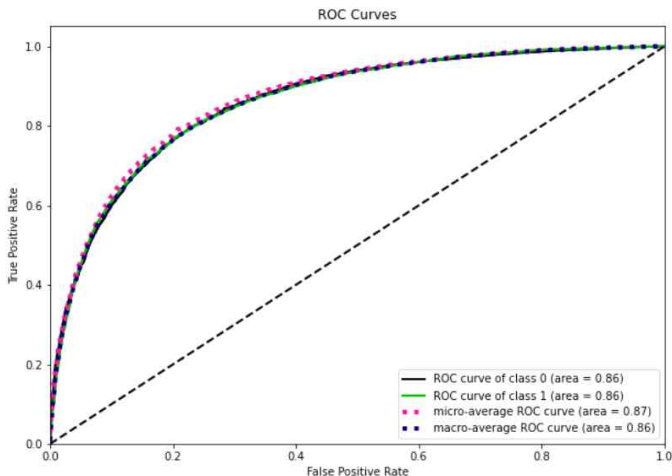
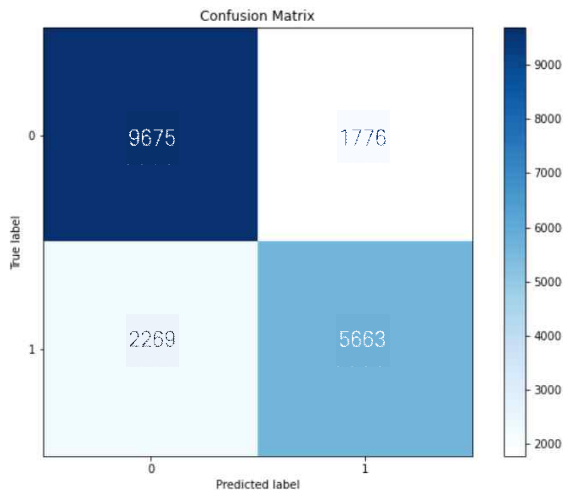
- 다른 GBM에 비해 과적합이 적음
- 범주형 변수에 대해 모델의 정확도와 속도가 높음
- encoding 작업 없이 모델의 input 가능



03 Train-Val 검증 Dataset(상위 5개)

➤ LightGBM 0.73771

- XGBoost보다 학습시간과 메모리 사용량 적음
- 기능의 다양성
- 카테고리형 피처의 자동 변환이 가능하고 최적 분할이 가능

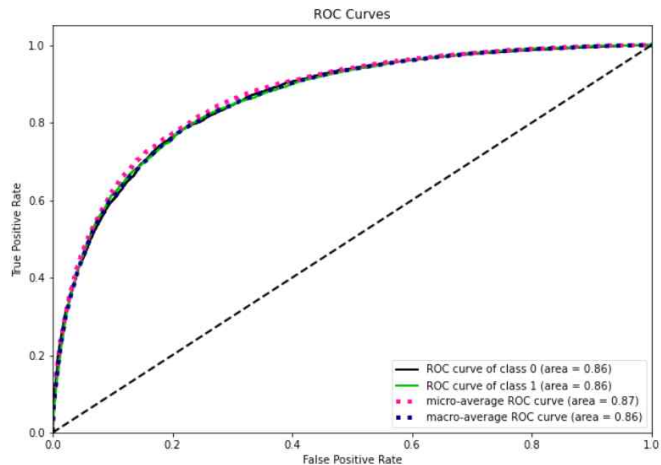
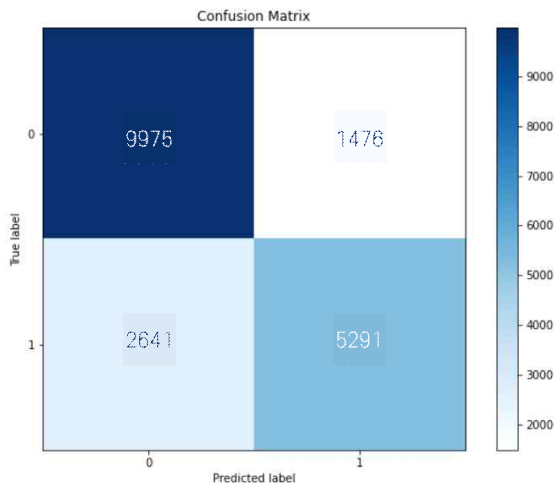


03 Train-Val 검증 Dataset(상위 5개)

➤ Xgboost 0.729744

트리 기반의 앙상블 학습모델

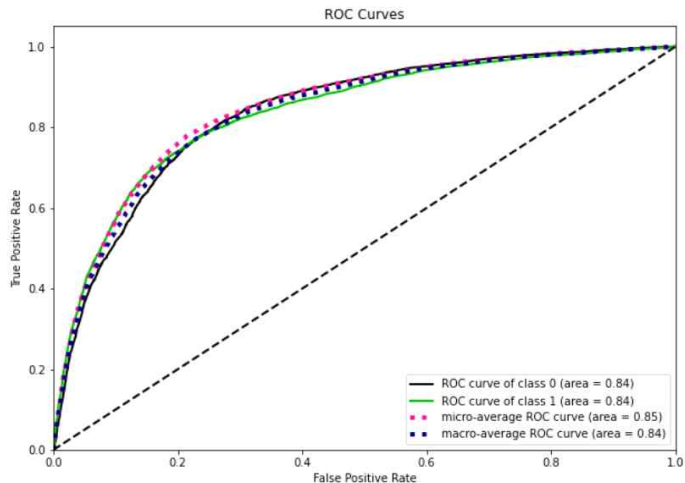
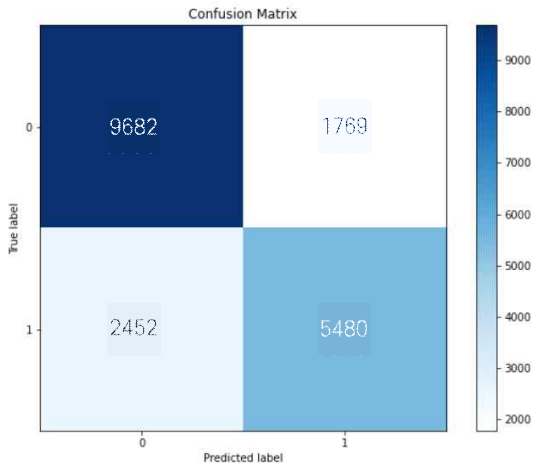
- 뛰어난 예측 성능
- GBM 대비 빠른 수행시간
- 과적합 규제 기능 / 결손값 자체 처리



03 Train-Val 검증 Dataset(상위 5개)

> Random Forest 0.721279

- 과대 적합(overfitting) 을 방지하기 위해, 최적의 기준 변수를 랜덤하게 선택
- 일반화 및 성능 우수
- 파라미터 조정 용이 / scale 변환 불필요



03 Test-Dataset 결과

> 최적화 후 Train-Val, Test 결과

```
1 f1_score(preds_class, test_target)
```

```
0.817089624156762
```

```
1 accuracy_score(preds_class, test_target)
```

```
0.8531187122736419
```

```
1 confusion_matrix(preds_class, test_target)
```

```
array([[10177, 1545],  
       [ 1302, 6359]])
```

f1_score

약 82%

accuracy

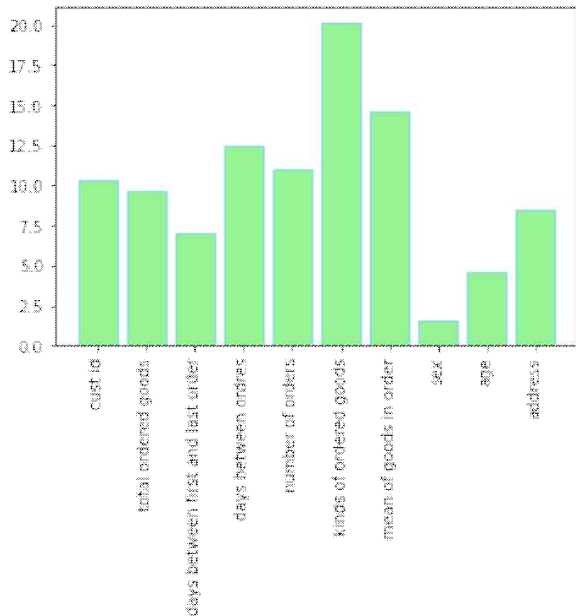
약 85%

coufusion matrix

10177	1545
1302	6359

03 Test-Dataset 결과

> Feature Importance



“ 재구매율 예측 결과 ”

재구매율이 높은 고객 예측 7,589명

재구매율이 낮은 고객 예측 11,722명

04

Clustering

사용 변수 및 속성 / 군집별 특성, 군집별 특성 분석

04 Clustering

» 사용 변수 및 속성

중요변수 12개를 추출하여
군집화 진행

구매상품수	고객이 구매한 상품의 갯수
첫-마지막	첫구매와 마지막 구매 사이의 간격
구매건수	고객이 구매한 횟수(영수증 갯수)
구매항목수	고객이 구매한 상품의 항목 수
영수증평균개수	고객이 구매한 횟수의 평균
구매금액평균	고객이 구매한 금액의 평균
구매간격	고객의 구매 간격 평균
구매_평균항목수	고객이 한번 구매시 평균적으로 몇가지의 상품을 구매했는가
총구매금액	
총영수증개수	고객의 총 영수증 갯수
최다구매물품	고객이 가장 많이 구매한 상품
재구매율	영수증 분석으로 얻어진 고객의 평균 재구매율

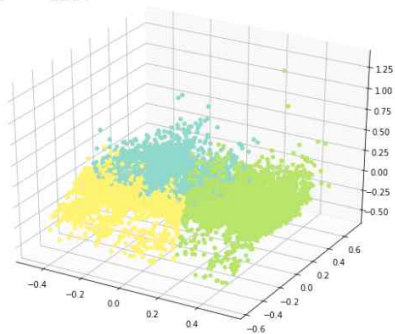
04 Clustering 모델

> 모델 비교

K_Means

- 군집 3개일 경우 분포가 고름
- 평균 실루엣 score: 0.7521

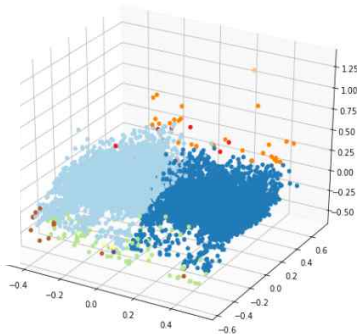
1	6795
0	6301
2	6287



Mean_Shift

- 최적의 bandwidth로 구할 경우 군집이 13개가 나오며 편차가 심함
- 평균 실루엣 score: 0.4062

0	12011
1	7212
2	80
7	28
4	14
12	11
5	9
10	8
3	4
11	3
6	1
8	1
9	1



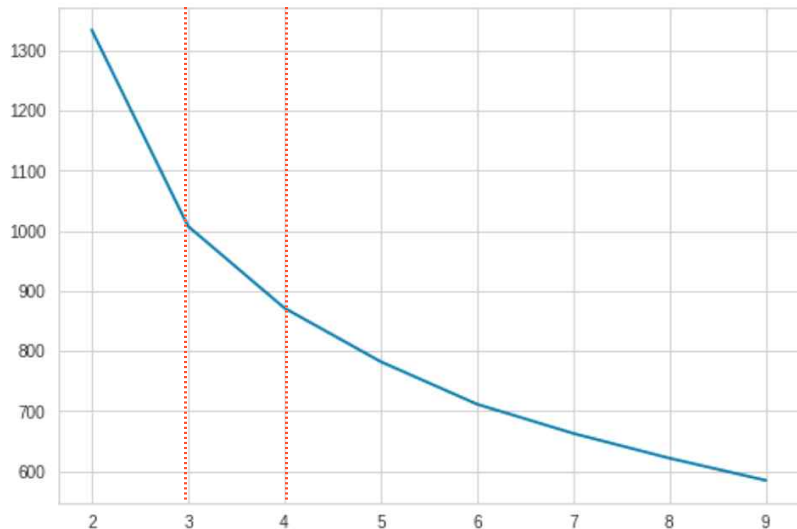
DBSCAN

- 군집이 제대로 나누어 지지 않음
- 평균 실루엣 score: 0.6381

0	19381
-1	2

04 Elbow-Method

› K_Means 모델에서 엘보우 기법사용

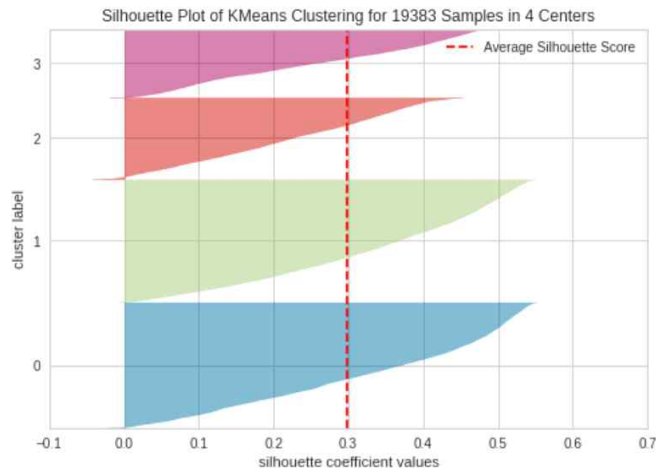
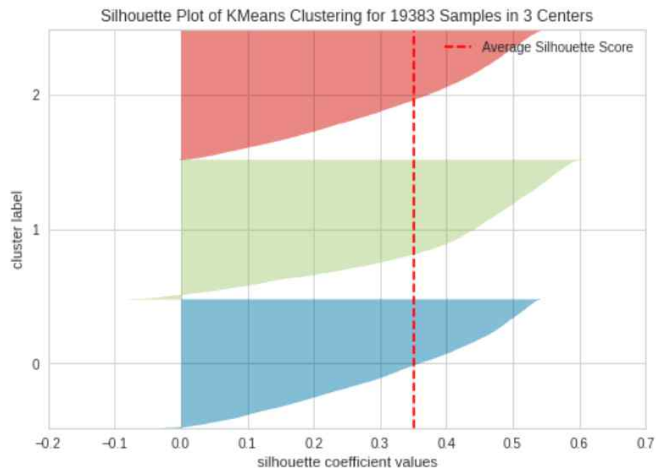


최적의 군집 수를 구하기 위해 사용

Cluster 간의 거리 합을 나타내는 inertia가 급격히 떨어지는 지점의 K값을 군집의 개수로 사용해볼 수 있는데, 3 또는 4의 위치에서 꺾이는 모습

04 Silhouette Coefficient

➤ 실루엣 계수 확인

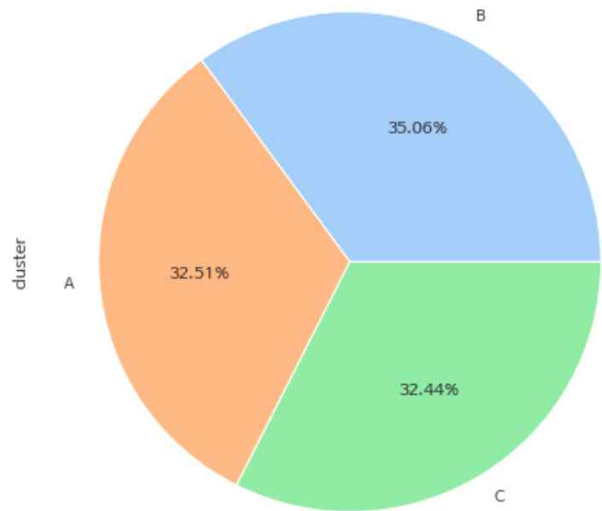


K=3으로 군집화한 것이 k=4에 비하여 데이터 포인트들의 실루엣 계수 값이 완만하며 고른것을 확인

➡ “ 최종 군집 3개 ”

04 군집특성

➤ 최종 군집 확인 및 재구매 고객비율 확인

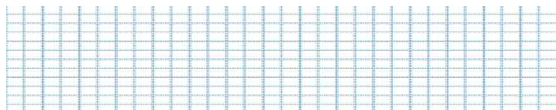


	총 고객수	재구매율 높음	재구매율 낮음
A	6301	707	5594
B	6795	2380	4415
C	6287	4574	1713

04 각 군집별 특성

	AOV	구매상품수	구매항목수	구매간격	재구매율	상품구매비율
A	15756.32	1951	469	2.24	0.05	<div>식품 <div><div></div><div></div><div></div></div></div> <div>가전/가구 <div><div></div><div></div><div></div></div></div> <div>의류/패션/명품 <div><div></div><div></div><div></div></div></div>
B	33481.57	1582	349	2.79	0.07	<div>식품 <div><div></div><div></div><div></div></div></div> <div>가전/가구 <div><div></div><div></div><div></div></div></div> <div>의류/패션/명품 <div><div></div><div></div><div></div></div></div>
C	60122.11	881	189	4.23	0.09	<div>식품 <div><div></div><div></div><div></div></div></div> <div>가전/가구 <div><div></div><div></div><div></div></div></div> <div>의류/패션/명품 <div><div></div><div></div><div></div></div></div>

AOV: 평균 주문 가치



04 각 군집별 특성

➤ 각 군집의 특성 정리

A군집

- 구매간격이 짧고
구매항목 수는 많으나,
전체적인 재구매율은 낮은편
- 가격이 저렴한 것들을
자주 구매하는 고객군
- 카테고리 중 식품과
일상용품의 구매율이 높은편

B군집

- 구매간격과 재구매율은
보통이나, 구매항목 수가
많은편
- 가격 민감도가 낮은편의 고객군
- 카테고리 중 식품과
- 의류/패션의 구매율이 높은편

C군집

- 구매간격이 비교적 길고
구매항목수는 적으나,
재구매율은 높은편
- 높은 가격에 크게 영향을
받지 않는 고객군
- 카테고리중 의류/패션/명품의
구매율이 높은편

05

군집별 마케팅 제안

각 군집에 사용될 추천 시스템

재구매 가능성이 높은 물품을 예측하는 모델

마케팅 제안

기대효과

05 A, B군집에 사용될 추천 시스템

> A, B 군집 특성을 반영한 추천시스템

"C군집에 비하여 재구매율이 낮은 편"

현재 구매 물건으로 추천을 하기보다
조금이라도 재구매율을 높일 수 있게,
재구매 가능성이 높게 예측된 고객군의
구매물품을 분석(예측)하여 추천하는 것이
더 효과적일 것이라 판단



→ 즉, 사용자를 기반으로 물품 추천

A군집 :

대표적인 추천시스템인
SVD(특이값 분해) 활용

재구매 가능성이 높게 예측된 고객이
낮게 예측된 고객보다 훨씬 다수

B군집 :

KNN(최근접 이웃기반 협업필터링) 활용

재구매율 가능성이 높게 예측된 고객과
낮게 예측된 고객이 고르게 분포

05 재구매 가능성이 높은 물품을 예측하는 모델

> 사용한 변수 속성

고객번호	고객 id
영수증번호	구매한 영수증 번호
구매상품수	해당 상품 구매 횟수
첫.마지막	첫 구매와 마지막 구매사이의 기간
구매간격	연속된 두 구매사이의 기간의 평균
구매건수	구매한 전체 상품수
구매항목수	구매한 상품 항목 수
구매_평균항목수	구매 당 평균 항목 수
reorder_rate	상품이 처음 구매 되어지고 두번째에 다시 재구매 되어진 비율
소분류코드	물품의 소분류 코드
상품당_총구매수	각 상품당 총 구매 수
처음_두번째주문비용	처음 구매 되어지고 두번째에 다시 재구매 되어진 비율
각 상품 먼저 구매고객 수	가장 먼저 카트에 넣은 고객 수
전체판매량/첫번째구매	전체 판매량 대비 첫번째로 구매 되어진 비율
평균구매량	각 상품을 구입한 고객의 평균 구매 횟수
cnt	총 재구매 수
goods_reorder_rate	구매한 전체 상품 중 재구매한 상품의 비율
특정상품구매횟수	특정 상품 구매 횟수
특정상품포함비율	전체 구매횟수 중 특정 상품을 포함하는 비율
처음구매횟수	각 고객이 특정 상품을 첫번째로 구매한 주문회차
마지막구매횟수	각 고객이 특정 상품을 마지막으로 구매한 주문회차
전체구매횟수/특정상품포함	특정 상품을 마지막으로 구매하고 난 후의 구매횟수
특정상품구매후_특정상품구매비율	특정 상품을 사고 난 후 특정 상품을 포함 하는 구매의 비율

Train Dataset

2014년 ~ 2015년 3분기 데이터

Test Dataset

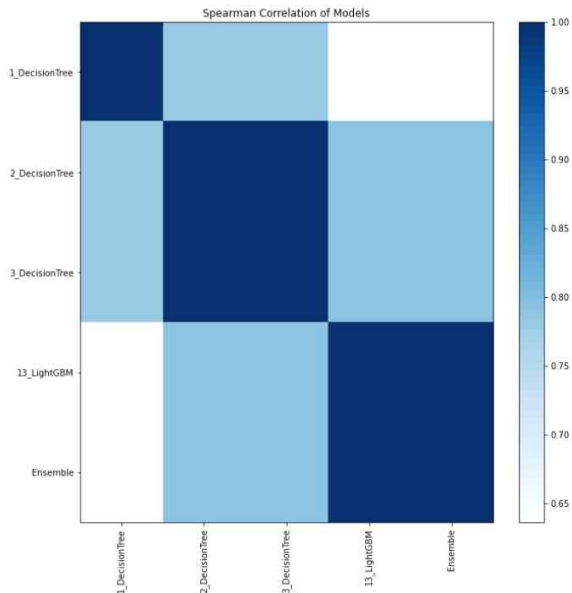
2015년 4분기 데이터

Model

A와 B군집에서 재구매 가능성이 높게
예측된 고객들을 대상으로
재구매 가능성이 높은 물품을 예측하는
모델 제작

05 Model

> 모델간의 correlation_heatmap

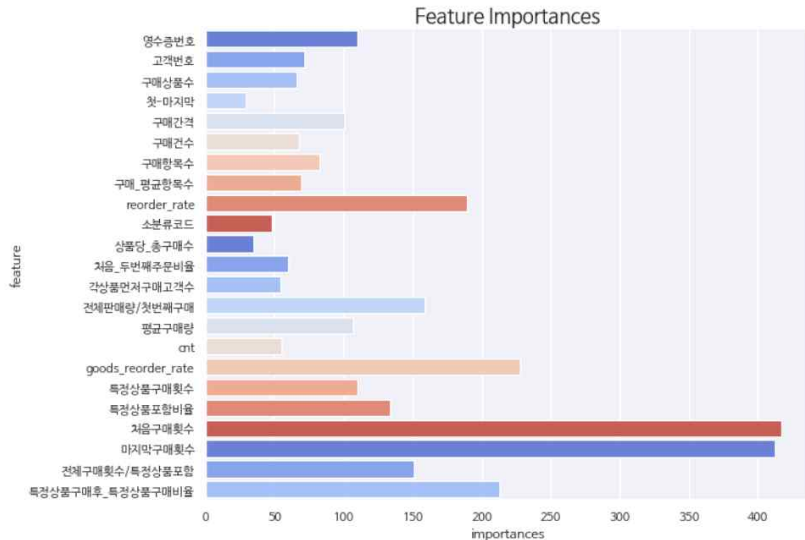


f1-Score 결과

가장 f1_score가 높게 나온 LightGBM으로
최적화(learning_rate = 0.5, num_leaves = 15) 후
학습을 재진행하기로 결정

name	model_type	metric_type	metric_value	train_time
1_DecisionTree	Decision Tree	f1	0.1403553299	529.09
2_DecisionTree	Decision Tree	f1	0.1527617405	583.29
3_DecisionTree	Decision Tree	f1	0.1527617405	592.64
13_LightGBM	LightGBM	f1	0.1604323221	22517.6

> Feature Importances



Test score

auc	accuracy	f1	precision
0.651	0.913	0.351	0.354

한계점 :

- 구매내역(행)의 수가 너무 많아 사용하는 Colab 환경 내 RAM과 런타임으로는 복잡한 모델에서는 학습진행 및 결과 도출에 있어 문제가 발생.
- 맞춰야 할 정답 레이블인 소분류의 상품 수가 3519개로 많아 Score가 낮게 나온 경향

02 A군집 마케팅 제안

items	score	item_name
797	2.0787740175206024	니트단품
1516	1.9845713627854549	선글라스(단기행사)
1067	1.9761525750857167	명성골프
2844	1.892080725865676	청백
3233	1.885756575667724	푸드애플라이어
3233	1.9654588494650207	푸드애플라이어
1067	1.959859843791608	명성골프
1516	1.941969683697214	선글라스(단기행사)
797	1.9014701342089653	니트단품
2891	1.8318316513350248	친환경메론
1516	2.166723426504819	선글라스(단기행사)
3233	2.166225463516816	푸드애플라이어
797	2.1153938962973946	니트단품
1067	1.9507190979724576	명성골프
2891	1.899880986354843	친환경메론
2327	3.365931491351394	원목학생철재
916	3.297697162317689	디자이너너부틱
275	3.1279923238352403	공병
1004	3.1168969786996383	마에스트로

마케팅 제안

- 서프라이즈 패키지의 SVD를 활용 예측된 물품 중 식품과 일상용품 카테고리의 물품을 선별해 추천을 진행한다.
- 식품의 경우 가격 변동이 크지 않고 수요의 탄력성이 높지 않은 상품군이며, 저렴한 가격의 상품들을 위주로 소비하는 패턴이 발견되었기 때문에 해당 상품들의 할인 행사, 이벤트 등으로 수요를 높일 수 있을 것이라 예상됨

02 B군집 마케팅 제안

score	user_id_y	소분류명
5.0	1	Bag&Bag
2.3184011942101708	1	L.B
2.029740868253141	1	L/C 아웃도어
1.96136962142236	1	MP3 外
1.862210162420264	1	N.B
5.0	1	PB바나나
2.2634668763748467	1	VIC마켓피자
2.003369297987843	1	VIC피자
1.996850175673184	1	global SPA
1.8429305728060423	1	가공우유
5.0	1	가공행사
2.251492762534169	1	가방
1.994758376226165	1	가정용화장지
1.845529766547853	1	간장
1.8422854965338784	1	간편과일
5.0	1	감
2.451232833738816	1	감자스낵
2.304261771850624	1	건강식품(홍삼)
1.99480938586193	1	건미역
1.871106919577656	1	건생선

마케팅 제안

- B군집의 경우 재구매 가능성이 높게 예측된 고객군(4415명)과 재구매 가능성이 낮게 예측된 고객군(1308명)이 고른 분포
- KNN(최근접 이웃기반 협업필터링)을 활용해 구매 가능성이 높게 예측된 물품 중 식품과 의류/패션 카테고리의 물품을 선별해 추천을 진행시 수요를 높일 수 있을 것이라 예상됨

02 C군집에 사용될 추천 시스템

> C군집의 특성을 반영한 연관상품 추천 시스템

A, B군집에 비하여 재구매율이 높은 편이므로 구매물품의 연관규칙을 분석하여 추천
즉, 아이템 기반 추천 시스템으로 장바구니 분석(연관규칙분석; apriori algorithm;) 활용

장바구니분석이란

연관 분석(Association Analysis) 이라고도 불리는 장바구니 분석은
대량의 정보로부터 개별 데이터 사이의 연관규칙을 찾는 것
예를 들어 마켓의 구매내역 중 특정 물건의 판매 발생 빈도를 기반으로
'A물건을 구매하는 사람들은 B물건을 함께 구매하는 경향이 있다' 라는
규칙을 찾아 활용하게 된다.



02 사용 알고리즘 / C군집 마케팅 제안

> Apriori 알고리즘

추천시스템의 1세대라고 할 수 있는
Apriori 알고리즘은 빈발항목집합을
추출하는 것이 원리

- 장점:

이해하기 쉬운 간단한 원리

상품간의 많은 연관규칙을 발견할 수 있음

- 단점:

상품 수가 많을 수록 그 계산량이
기하급수적으로 늘어남.



0.0516014234875444	L/C 아동복
0.0518234165067178	유아복
0.0519125683060109	L/C 아동복
0.0519125683060109	유아복
0.0520361990950226	수입 아동복
0.0520361990950226	N/B 아동복
0.0520361990950226	N/B 아동복
0.052064631956912	N/B 아동복
0.0520833333333333	L/C 아동복
0.0521739130434782	L/C 아동복
0.0522727272727272	유아복
0.0522727272727272	유아복
0.0523917995444191	L/C 아동복
0.0526315789473684	재사용봉투, '열군가공우유', '산딸기', '생화', '전도복송아', '감자스낵
0.0526315789473684	유부, '재사용봉투', '생화', '전도복송아', '감자스낵', '파일/바인더
0.0526315789473684	유부, '열군가공우유', '산딸기', '전도복송아', '감자스낵', '파일/바인더
0.0526315789473684	다이소, '유부', '산딸기', '생화', '전도복송아', '감자스낵
0.0526315789473684	다이소, '유부', '재사용봉투', '전도복송아', '감자스낵', '일반빵
0.0526315789473684	다이소, '산딸기', '생화', '전도복송아', '감자스낵', '일반빵
0.0526315789473684	다이소, '유부', '전도복송아', '감자스낵', '파일/바인더', '일반빵
0.0526315789473684	다이소, '재사용봉투', '산딸기', '생화', '전도복송아', '감자스낵
0.0526315789473684	다이소, '유부', '산딸기', '전도복송아', '감자스낵', '파일/바인더
0.0526315789473684	다이소, '산딸기', '생화', '전도복송아', '감자스낵', '파일/바인더
0.0526315789473684	다이소, '유부', '산딸기', '전도복송아', '감자스낵', '일반빵
0.0526315789473684	다이소, '유부', '재사용봉투', '산딸기', '전도복송아', '감자스낵

마케팅 제안

- 연관규칙분석 알고리즘인 Apriori를 활용하여 분석된 물품을 활용하여 고객별 맞춤 추천 서비스 제공
- 가격 민감도가 낮고 명품의 구매율도 높은 편이기 때문에 사치품에 대한 마케팅을 진행

02 기대효과

➤ 상관계수를 통해 구해본 결과

재구매율 상승



해당 마케팅제안으로 재구매율이 10% 정도 상승함을 가정한다면

매출 상승



최종 예상 매출은 2년간 전체매출
대비 2~3% 증가할 것으로 계산



재구매율 예측을 통한 구매물품 예측 및 솔루션 제안

감사합니다

Q&A