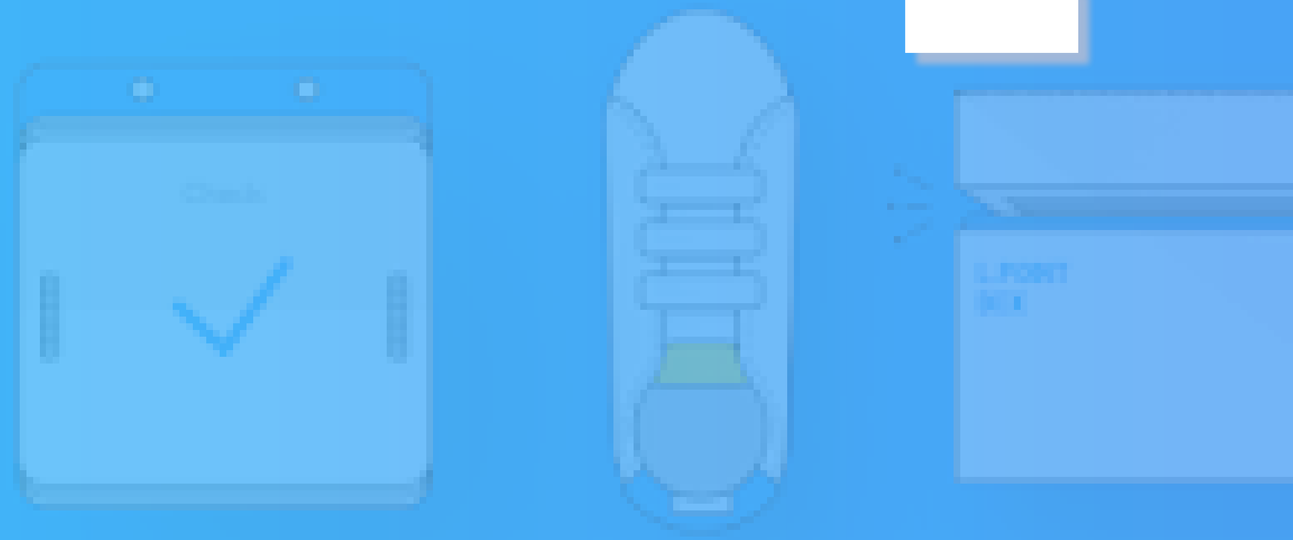
L.POINT

엘포인트로 일상을 즐겨허다



제7회 롯데멤버스 빅데이터 경진대회

Team - L(ove).POINT

제7회 롯데멤버스 빅데이터 경진대회

Team - L(ove).POINT

CONTENTS

제7회 롯데멤버스 빅데이터 경진대회

LPOINT_BIG_COMP_01_DEMO.csv

	cust	ma_fem_dv	ages	zon_hlv
0	M000034966	여성	40대	Z07
1	M000059535	여성	30대	Z12
2	M000136117	여성	30대	Z11
3	M000201112	여성	50대	Z17
4	M000225114	여성	40대	Z17

LPOINT_BIG_COMP_02_PDDE.csv

	cust	rct_no	chnl_dv	cop_c	br_c	pd_c	de_dt	de_hr	buy_am	buy_ct
0	M430112881	A01000001113	1	A01	A010039	PD0290	20210101	10	15000.0	1
1	M646853852	A01000002265	1	A01	A010025	PD1389	20210101	10	79700.0	1
2	M430112881	A01000003148	1	A01	A010039	PD0290	20210101	10	19000.0	1
3	M430112881	A01000003148	1	A01	A010039	PD0290	20210101	10	19000.0	1
4	M430112881	A01000004946	1	A01	A010039	PD0290	20210101	10	19000.0	1

LPOINT_BIG_COMP_03_COP_U.csv

	cust	rct_no	cop_c	br_c	chnl_dv	de_dt	vst_dt	de_hr	buy_am
0	M839993508	21102612B015763935	B01	B010012	1	20211026	20211026	12	60000
1	M839993508	21110610B014219744	B01	B010012	1	20211106	20211106	10	17100
2	M839993508	21021112B013419710	B01	B010012	1	20210211	20210211	12	136500
3	M839993508	21092010B012637545	B01	B010012	1	20210920	20210920	10	34200
4	M839993508	21101009D015920171	D01	D010614	1	20211010	20211010	9	2500

LPOINT_BIG_COMP_04_PD_CLAC.csv

	pd_c	pd_nm	clac_hlv_nm	clac_mcls_nm
0	PD0001	소파	가구	거실가구
1	PD0002	스툴/리빙의자	가구	거실가구
2	PD0003	탁자	가구	거실가구
3	PD0004	장식장/진열장	가구	거실가구
4	PD0005	기타가구	가구	기타가구

LPOINT_BIG_COMP_05_BR.csv

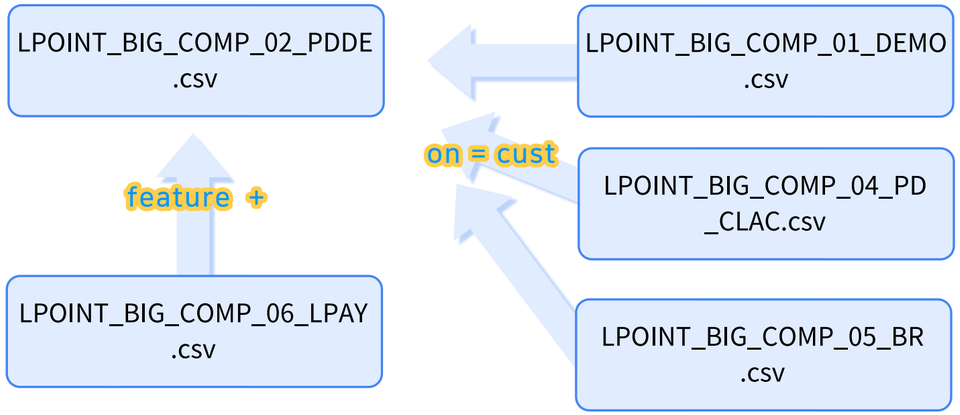
	br_c	cop_c	zon_hlv	zon_mcls
0	A010001	A01	Z17	Z17024
1	A010002	A01	Z17	Z17018
2	A010003	A01	Z17	Z17011
3	A010004	A01	Z16	Z16007
4	A010005	A01	Z17	Z17005

LPOINT_BIG_COMP_06_LPAY.csv

	cust	rct_no	cop_c	chnl_dv	de_dt	de_hr	buy_am
0	M629656521	210803210311226	A03	1	20210803	21	10900
1	M216016456	210803130167542	L01	2	20210803	13	6860
2	M205142844	210803140275112	A02	1	20210803	14	9000
3	M737010483	210803040637594	A06	2	20210803	4	36740
4	M707775545	210803140675502	A06	2	20210803	14	138500

롯데멤버스 빅데이터 경진대회측 제공 데이터

제7회 롯데멤버스 빅데이터 경진대회



-L.POINT_BIG_CO MP_02 데이터에 cust를 기준으로 L.POINT_BIG_COMP_01, 04, 05 을 병합

-L.POINT_BIG_COMP_06 데이터에서 cop_c, most_chnl은 cust 별 최빈값으로 feature 추출

-L.POINT_BIG_COMP_06 데이터에서 point 총 사용량을 확인하기 위해 cust 별 buy_am - sum() feature 추출

-L.POINT_BIG_COMP_06 데이터에서 point 사용 횟수를 확인하기 위해 cust 별 count() feature를 추출

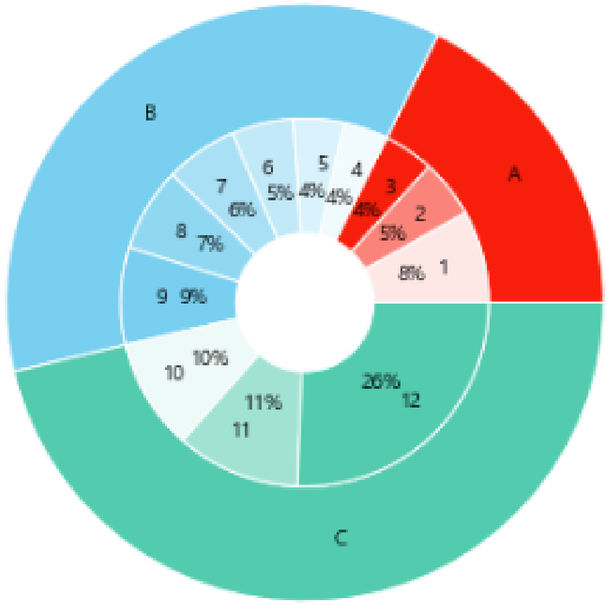
chnl_dv	cop_c	br_c	pd_c	de_dt	de_hr	buy_am	buy_ct	pd_nm	clac_hlv_nm	clac_mcls_nm	zon_mcls	ma_fem_dv	ages	zon_hlv
1	A01	A010039	PD0290	20210101	10	19000.0	1	남성티셔츠	남성의류	남성의류상의	Z10042	여성	50대	Z17
1	A01	A010039	PD0290	20210101	10	19000.0	1	남성티셔츠	남성의류	남성의류상의	Z10042	여성	50대	Z17

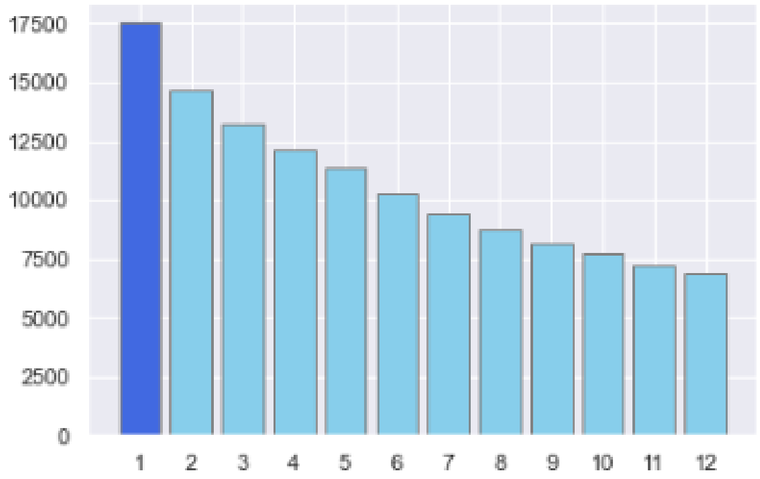
chnl_dv	cop_c	br_c	pd_c	de_dt	de_hr	buy_am	buy_ct	pd_nm	clac_hlv_nm	clac_mcls_nm	zon_mcls	ma_fem_dv	ages	zon_hlv	중복수
1	A01	A010039	PD0290	20210101	10	38000.0	2.0	남성티셔츠	남성의류	남성의류상의	Z10042	여성	50대	Z17	2.0

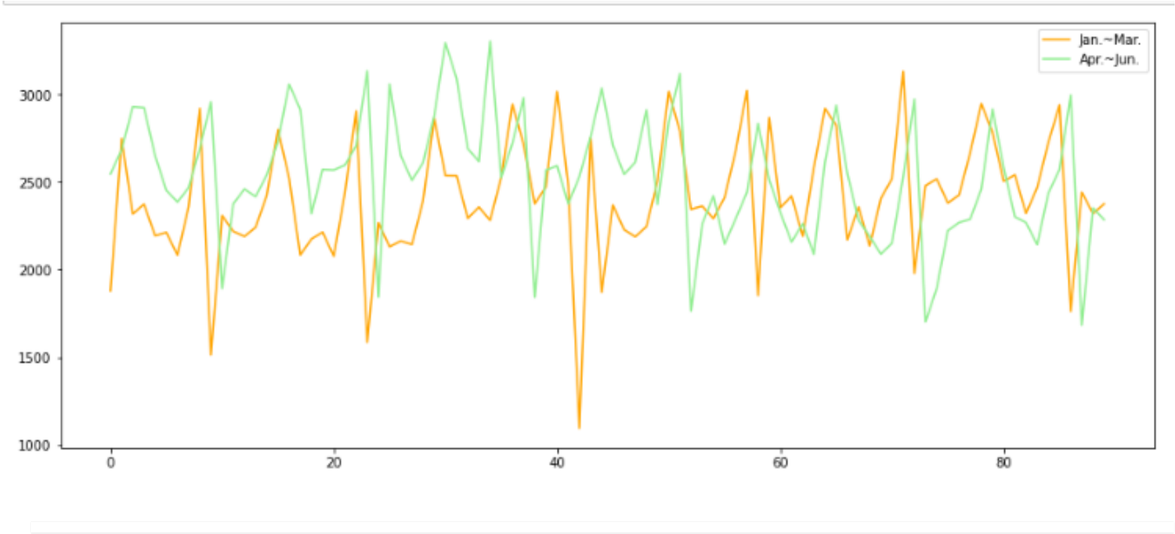
데이터 행 중에서 중복 데이터가 발견됐는데, 중복된 이유가 있다고 판단

따라서 데이터를 그냥 제거하는 것이 아닌 buy_am, buy_ct는 더하고, 한 행만 남기고 나머지 제거 후 '중복 수' 라는 column을 추가하여 줄인 행 수를 알 수 있게 해주었다.

제7회 롯데멤버스 빅데이터 경진대회







좌측 상단 그래프는 고객별 방문 달 수를 count 하여 각 달마다 비율을 나타낸 것이다. 1~3개를 적은 방문(A)로, 4~9개를 중간(B), 10~12개를 많은 방문(C)로 분류하여 시각화를 진행하였다.

좌측 하단 그래프는 1월 달에 온 고객이 2월,3월~12월까지 연속적으로 존재하는 비율에 대해 시각화를 진행

1->2->3월로 이동하면서 계속해서 존재하는 cust 수를 시각화한 것인데, 이를 12월달까지 이동하면서 각 달마다 연속적으로 온 고객 수를 시각화했다.

좌측 상단 그래프를 통해 우리가 분류한 C그룹의 수는 50%에 미치지 못하고, 좌측 하단의 그래프를 통해 12월까지 연속적으로 온 고객 또한 50%가 되지 못함을 파악할 수 있다.

우리는 이에 C와 같은 충성 고객을 늘리는 전략 제시를 하고자 한다.

추가로 위 좌측 위 그래프는 1~3월(1분기), 4~6월(2분기) 총 방문수를 일자별로 시각화 한것으로 어느정도 주기성이 나타난다는 사실을 시각화를 통해 확인하였고, 추가적으로 날씨 데이터와 같은 시계열 데이터를 통해 더 자세히 분석할 것이다.

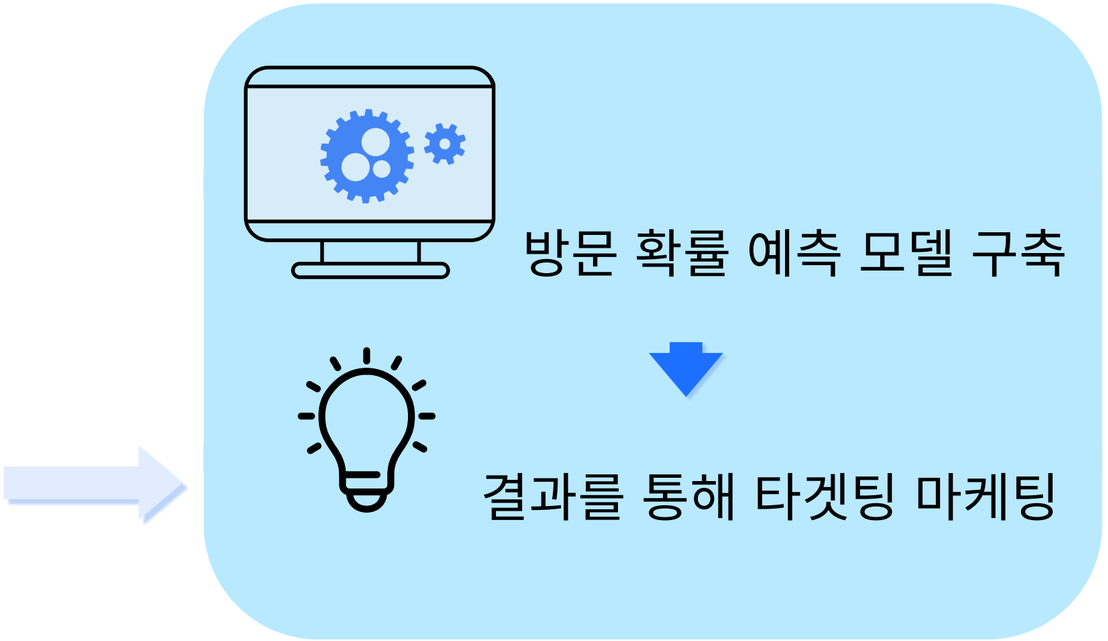
제7회 롯데멤버스 빅데이터 경진대회

왼쪽 기사들로 기업이 충성고객을 확보하고 늘려가는 것이
상당히 중요하다는 것을 알 수 있다.

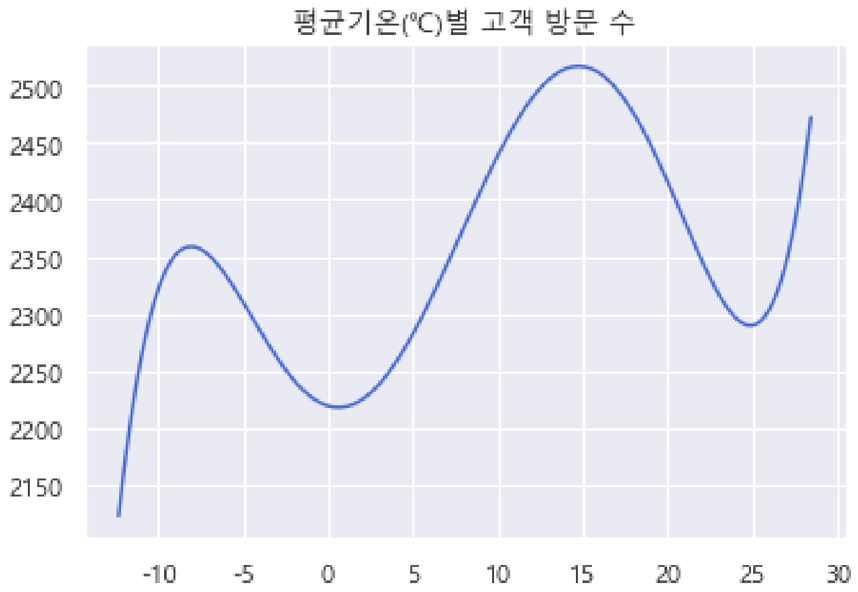
기업들은 실제로 '충성고객'을 타겟팅하는 다양한 마케팅
전략을 사용중이며, 여기에서 우리의 주제가 나오게 되었다.

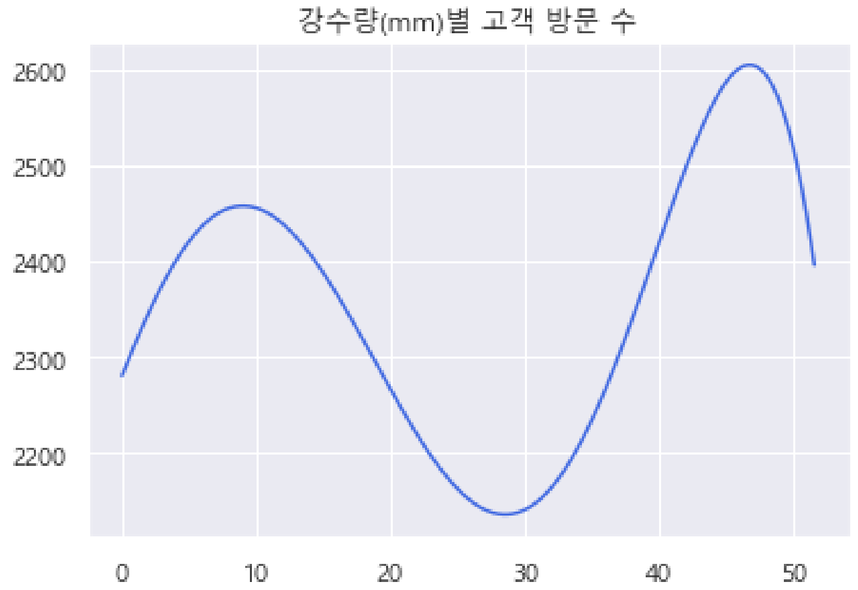
이에 우리는 기존에 있는 충성 고객 전략에 덧붙여, 충성 고객의 반대 그룹인 취약 고객 층의 방문 유도 전략을 생각하게 되었다

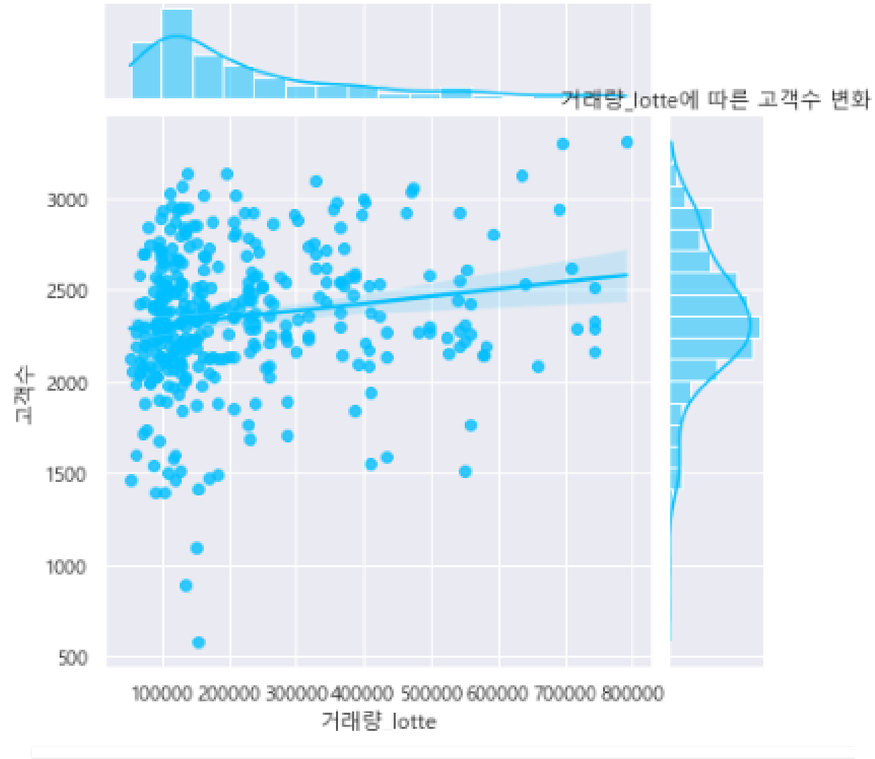
따라서 고객별 다음 달 방문 확률을 예측하는 모델을 구축하고,
이후 모델 예측 결과를 통해 예측 방문 확률에 따라 구분하여 타켓팅 마케팅을 진행 해볼까한다.



제7회 롯데멤버스
빅데이터 경진대회







좌측 위 그래프는
기상청_2021일자별_기온(출처_기상청_기상자료개방포털)와 방문객과의 추세 그래프

좌측 아래 그래프는
기상청_2021일자별_기온(출처_기상청_기상자료개방포털)와 방문객과의 추세 그래프

우측 그래프는 pykrx에서 불러온 2021 네이버 금융 데이터에서 거래량_lotte에 따른 고객수 변화를 나타낸 그래프

위와 같은 그래프를 통해서 외부 데이터가 방문 수를 예측하는데 유의미함을 보임

제7회 롯데멤버스
빅데이터 경진대회

목표 : 다음달에 방문할 확률에 대한 개인화 마케팅

ㅇ 고객 별 예측을 위한 고객 별 피쳐 생성

- 총구매액, 구매건수, 평균구매액, 최대구매액, 최소구매액
- 시간적 특성별 방문 비율, 내점일수, 구매주기, 공휴일
- 주방문요일, 주구매 특징, 상품가격별 구매비율, 구매 다양성, 연령대별 선호도
- 날씨에 따른 구매 특징, 주가에 따른 구매 특징, 온오프라인 구매 비율
- L 포인트 사용 정도 등

Word 2 Vec

- 고객 별 월간 구매한 대분류를 분석하여 피쳐로 활용

skip-gram 방식을 사용하여 단어마다 주변 3개의 단어와 관계를 파악하여 총, 20개의 피쳐를 생성

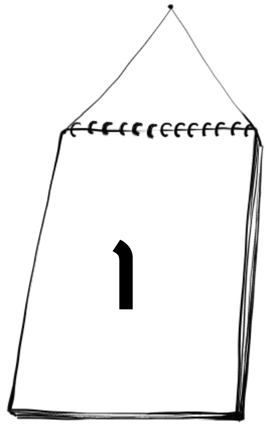
KMeans

- Standard Scaler를 활용하여 피쳐간 정규화를 해준 후, KMeans를 활용하여 클러스터링 진행

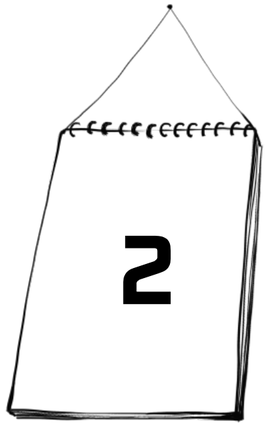
그룹 수는 성능을 점검하며 설정. 최종적으로 5개의 그룹으로 고객을 나눔.

제7회 롯데멤버스 빅데이터 경진대회

ㅇ 학습 방식

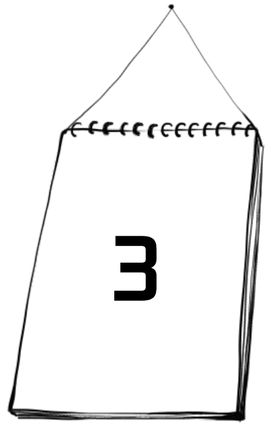


X_train



Y_train

X_test



Y_test

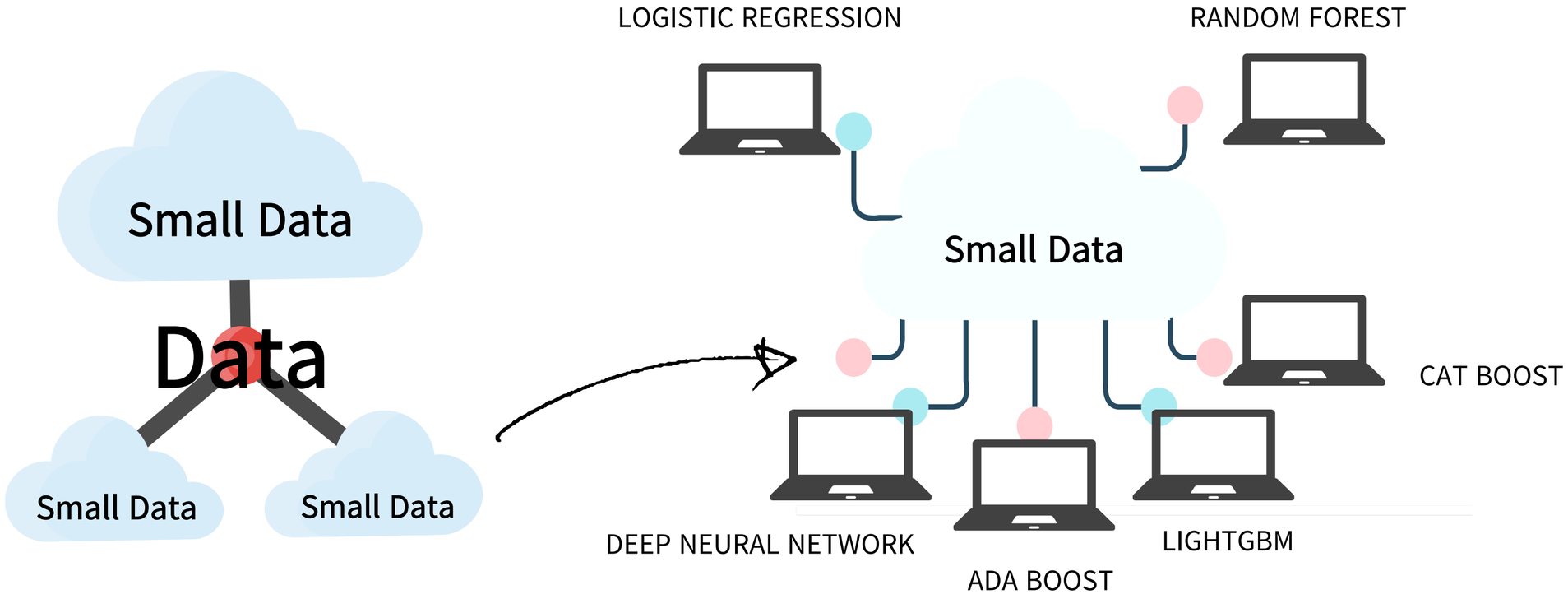
모델 학습을 위해서는 train data와 다음 달에 올지 안올지에 대한 정답 label이 필요. 따라서, 다음 달의 존재하는 고객일 경우 1, 아닐 경우 0으로 사용

예를 들어, 1월을 학습시 정답 레이블로 2월에 있는 고객이 1월에 존재할 경우 1로 설정 아닐 경우 0으로 설정하여 학습 진행
1, 2월을 이용해 train을 진행하고 2, 3월을 이용하여 점수를 판단

전략 제시에 고객 별 다음 달의 올 확률을 사용하기에, 정확한 확률을 위해 튜닝 및 점수는 log_loss를 사용

제7회 롯데멤버스 빅데이터 경진대회

ㅇ 모델 설정



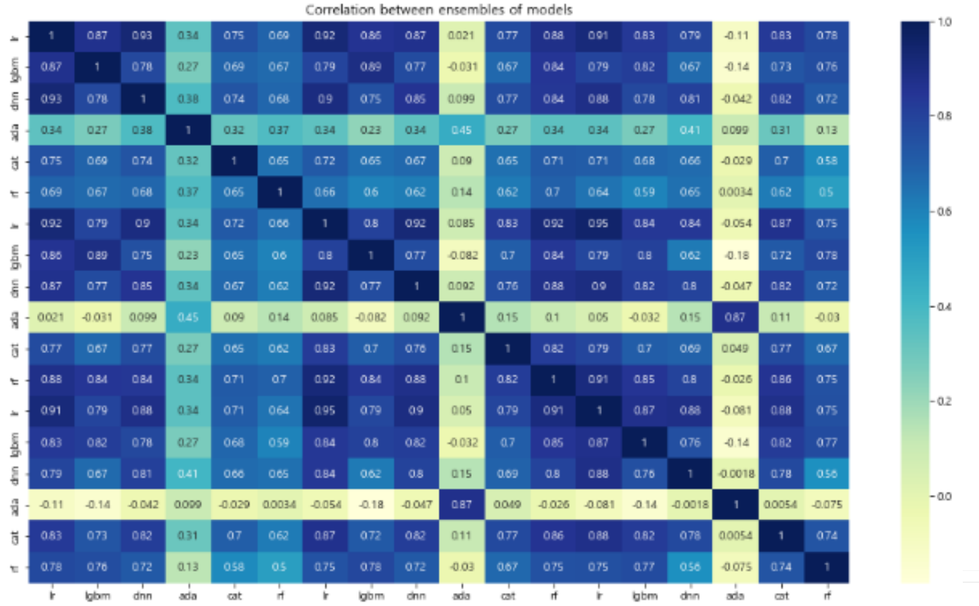
데이터프레임에서 랜덤으로 피쳐 50%를 추출하여 작은 데이터 프레임 3개를 생성

이후, 각각의 작은 데이터 프레임에서 개별적인 6개의 모델 튜닝 후, 학습을 통한 결과 도출

제7회 롯데멤버스 빅데이터 경진대회

ㅇ 모델 앙상블

X 18



18개의 SUBMISSION(결과)를 각각 상관계수 파악 히트맵

결과물을 앙상블하는 방식은 산술평균, 기하평균, 멱평균을 사용

총 18개의 결과물 중 2~5개의 조합으로 결과물을 만듦

모든 결과물 중 성능이 가장 좋은 결과물을 최종 결과물로 선정

제7회 롯데멤버스 빅데이터 경진대회


```python
model_dict={}; model_dict1={} ; model_dict2={}   # 학습한 모델 저장
score_lst=[]; score_lst1=[]; score_lst2=[]  # validation 점수 저장   딕셔너리 형태 key 값 : log_loss, roc_auc, accuracy
result_lst=[]; result_lst1=[]; result_lst2=[]


lst_e=[[model_dict, score_lst, result_lst], [model_dict1, score_lst1, result_lst1], [model_dict2, score_lst2, result_lst2]]

t = X_train
k = X_test
a_1=random.sample(list(t.columns),int(len(list(t.columns))//2))
b_1=random.sample(list(t.columns),int(len(list(t.columns))//2))
c_1=random.sample(list(t.columns),int(len(list(t.columns))//2))
```

데이터 셋을 3개의 작은 데이터 셋으로 나누는 과정

검증 및 결과 (예시:9월 예측 성능)

Feature_set_1 :

- Logistic : {'log_loss': 0.4334431508746004, 'roc_auc': 0.7591655847679943, 'accuracy': 0.8145581492288454}
- Lgbm : {'log_loss': 0.5215584359393713, 'roc_auc': 0.7582082013557917, 'accuracy': 0.8137244685285535}
- DNN : {'roc_auc': 0.7540811984336081}
- AdaBoost : {'log_loss': 0.6526985341001599, 'roc_auc': 0.7618755471165111, 'accuracy': 0.8138807836598583}
- CatBoost : {'log_loss': 0.4714018329801267, 'roc_auc': 0.7475109194235701, 'accuracy': 0.81382867861609}
- RandomForest : {'log_loss': 0.4330249135176414, 'roc_auc': 0.7384679557119316, 'accuracy': 0.8141413088786995}

Feature_set_2 :

- Logistic : {'log_loss': 0.45020810998980926, 'roc_auc': 0.7422975950837396, 'accuracy': 0.8154960400166736}
- Lgbm : {'log_loss': 0.4397032925416829, 'roc_auc': 0.7451053600149986, 'accuracy': 0.8134118382659441}
- DNN : {'roc_auc': 0.7253124995594875}
- AdaBoost : {'log_loss': 0.66635784533895, 'roc_auc': 0.7466063798895124, 'accuracy': 0.8129949979157982}
- CatBoost : {'log_loss': 0.4565424332287267, 'roc_auc': 0.7215160483985785, 'accuracy': 0.8111192163401417}
- RandomForest : {'log_loss': 0.4496144000845651, 'roc_auc': 0.716944929595532, 'accuracy': 0.8153397248853689}

Feature_set_3 :

- Logistic : {'log_loss': 0.42109018318579716, 'roc_auc': 0.7650824251727868, 'accuracy': 0.8159649854105877}
- Lgbm : {'log_loss': 0.44145255025386737, 'roc_auc': 0.74339352848389, 'accuracy': 0.8102334305960817}
- DNN : {'roc_auc': 0.7354728231987268}
- AdaBoost : {'log_loss': 0.6655855739264643, 'roc_auc': 0.7610115611621635, 'accuracy': 0.8134118382659441}
- CatBoost : {'log_loss': 0.46080659879615593, 'roc_auc': 0.729310220876486, 'accuracy': 0.810754481033764}
- RandomForest : {'log_loss': 0.4512236343448503, 'roc_auc': 0.70752808531423, 'accuracy': 0.8096081700708628}

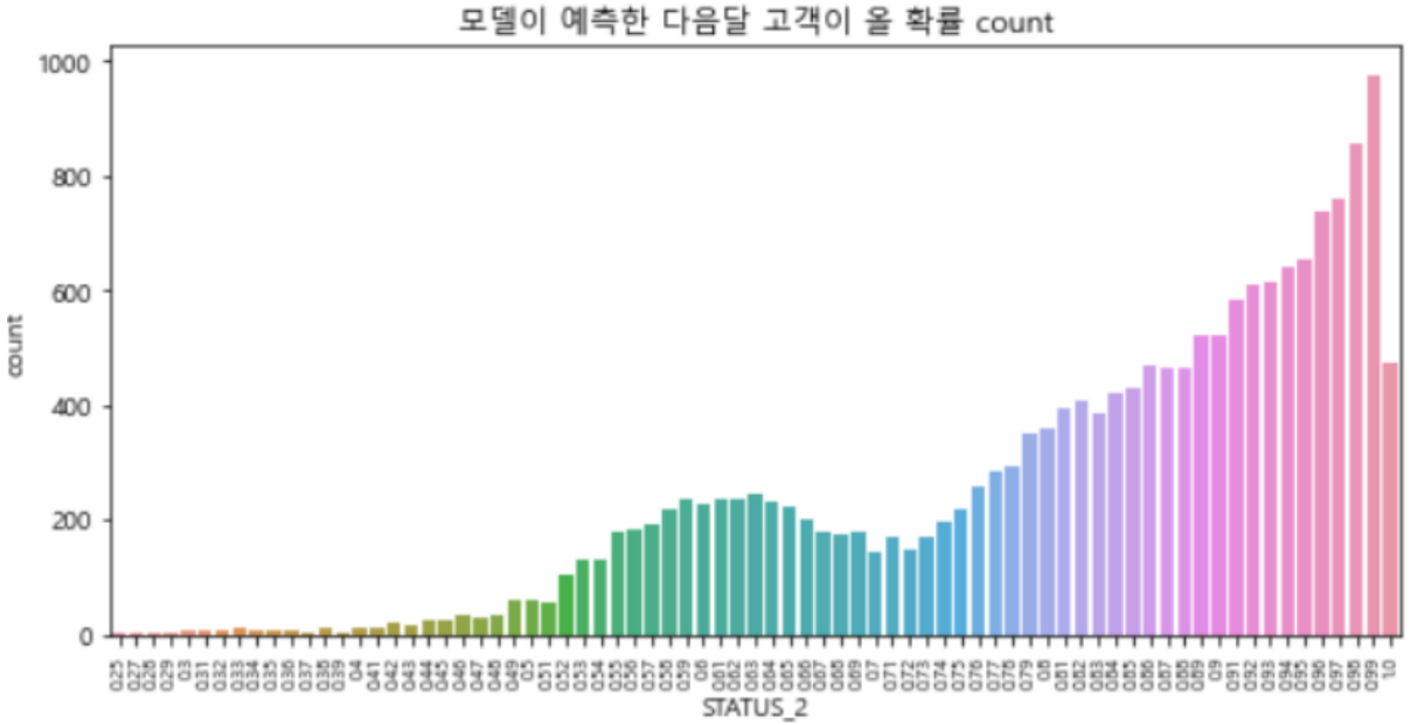
각각의 데이터 셋 3개마다 6개의 모델을 사용하여 예측한 점수

제7회 롯데멤버스 빅데이터 경진대회

최종 결과

	ID	STATUS
0	M000136117	0.945766
1	M000261625	0.608945
2	M000350564	0.758195
3	M000419293	0.829459
4	M000494848	0.992146
...	...	...
17945	M999599111	0.719990
17946	M999673157	0.854782
17947	M999770689	0.992443
17948	M999849895	0.698876
17949	M999962961	0.958260

17950 rows × 2 columns



ACCURACY SCORE : 0.84974930362117
LOG_LOSS : 0.3536696727294783
앙상블 시 사용한 모델 : 'lr', 'lr2'
멱평균 P수치 : 1.5

최종적으로 정확도가 약 85%의 수치를 보였으며, 이는 1월을 예측하기 위해서 11월, 12월을 통해 모델을 학습하면 12월에 있는 고객들이 1월 재방문 여부를 85%의 정확도로 예측할 수 있음.

예측한 결과를 COUNT PLOT을 통해 시각화를 하면 위와 같은 분포를 확인할 수 있음

제7회 롯데멤버스 빅데이터 경진대회

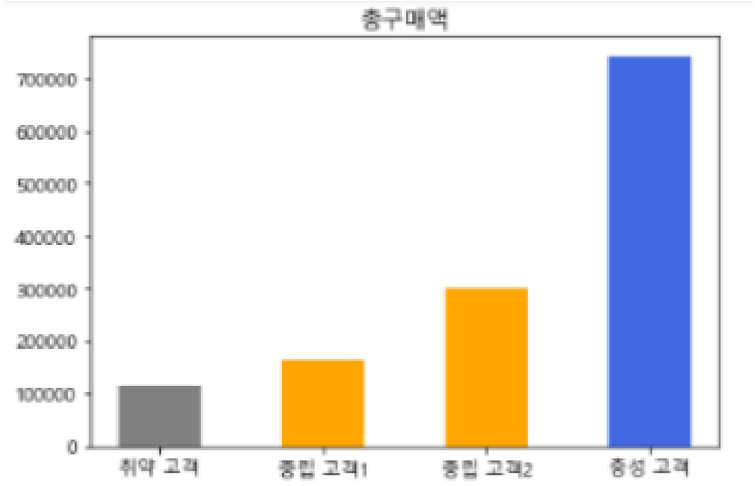
고객 별 다음 달에 올 확률을 이용하여 다음과 같은 그룹을 설정한다.

다음 달에 올 확률 0.5 이하인 그룹 -> 취약 고객

다음 달에 올 확률 0.5 이상, 0.7 미만인 그룹 -> 중립 고객 1

다음 달에 올 확률 0.7 이상, 0.9 미만인 그룹 -> 중립 고객 2

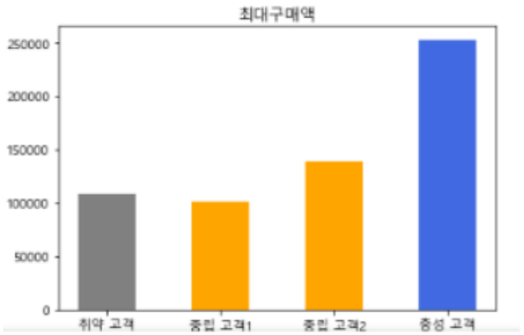
다음 달에 올 확률 0.9 이상인 그룹 -> 충성 고객

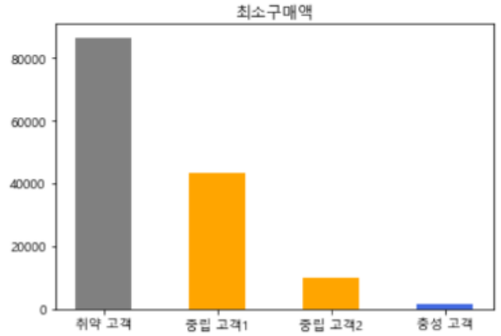


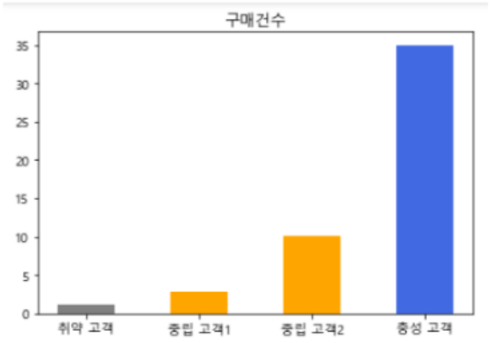
각 그룹별, 총구매액을 살펴본 결과 다음달에 올 것이라 예측한 고객들일수록 총 구매액이 큰 것을 확인할 수 있다.

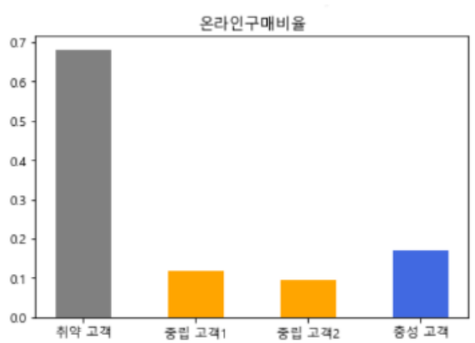
또한, 충성 고객의 특징과 취약 고객의 특징을 분석하여 취약 고객이 충성 고객의 특징을 가질 수 있도록 유도하는 방향의 전략을 제시하여 기업의 수익 창출 극대화를 기대해 볼 수 있다.

제7회 롯데멤버스 빅데이터 경진대회









충성 고객과 나머지 고객을 비교해보았을 때, 충성 고객은 최대 구매액이 가장 높으며 최소구매액은 가장 낮은 것을 볼 수 있다.

그리고 구매건수도 가장 높은데 이는 충성 고객의 경우 생활 용품, 식료품에서부터 전자제품, 의류까지 다양한 물품을 구매함을 알 수 있다.

다양한 물품을 구매하지 않는 취약, 중립 고객에 대하여 한 물품을 살 시, 다른 종류의 매장에서의 혜택을 제공하는 연계방식을 사용한다면 마케팅 효과를 기대할 수 있을 것이다.

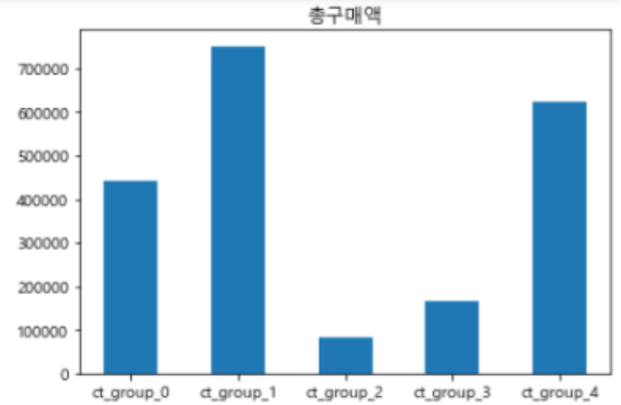
또한, 가장 고객 그룹 별 가장 두드러진 특징으로 온라인 구매비율을 확인할 수 있다. 중립 고객과 충성 고객의 경우 온라인 구매비율이 낮고, 취약 고객의 경우에는 온라인 구매비율이 높다.

이에 취약 고객을 좀 더 충성 고객으로 변환시키기 위해서는 오프라인 방문 유도가 필요할 것으로 보인다.

이를 위해 온라인 구매 시 오프라인 상품권을 증정, 온라인 사이트에 오프라인 매장의 장점을 보여주는 홍보를 하는 등의 마케팅을 제시한다면 기존 취약 고객을 충성 고객으로 유도할 수 있을 것으로 보인다.

제7회 롯데멤버스 빅데이터 경진대회

세부적인 마케팅 제시 방안

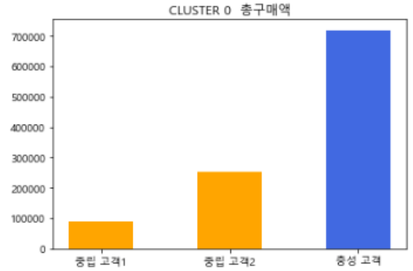


K-MEANS 클러스터링 방식을 이용한 집단의 고객별 총구매액 평균을 시각화한 자료이다.

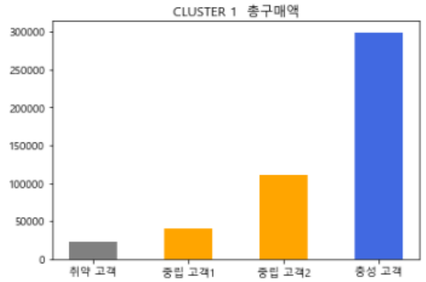
클러스터링 방식을 사용하여 집단별로 이질적으로 만들어 각 집단 별 맞춤 마케팅 전략을 제시 할 수 있다.

각 집단 별 다시 한번 다음 달의 올 확률 별로 세부 그룹화를 진행하여 더욱 더 고객 개인화 마케팅 전략에 이용할 수 있을 것이다.

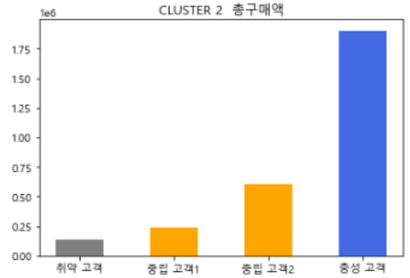
취약 고객: 0 중립 고객1: 15 중립 고객2: 1079 충성 고객: 4317



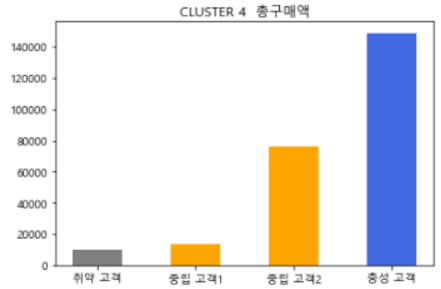
취약 고객: 4 중립 고객1: 495 중립 고객2: 1927 충성 고객: 1262



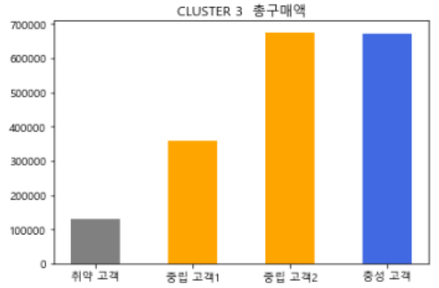
취약 고객 : 42 중립 고객1 : 1577 중립 고객2 : 2189 충성 고객 : 997



취약 고객 : 46 중립 고객1 : 628 중립 고객2 : 1580 충성 고객 : 810



취약 고객: 278 중립 고객1: 246 중립 고객2: 257 충성 고객: 201



THANK YOU

Team - L(ove).POINT