빅데이터 분석을 통한
자동차 공유 플랫폼 마케팅 전략 제안

솔트룩스&새싹 2차 프로젝트 – 오프라인1 조

오프라인1조
작성자: 김도균, 반위홍, 오해윤, 윤태양

Index

분석 주제 이해

분석 배경

한국경제

'데이터'로 타다금지법·코로나19 위기 극복

쏘카, 데이터 기반 마케팅으로 유니콘 등극 렌터카 이미지를 주중 이동 수요로 확장 카셰어링 경쟁에 구독 '록인전략'으로 대응.

2021. 9. 2.

과거 방식

Why?

기존 업계 실무자→ 엑셀을 활용한 고객 데이터 관리, 설문조사 방식의 전통적 마케팅 활용

최근

"빅데이터, 머신러닝, 인공지능 기법들을 통해 고객정보와 이용내역 데이터를 분석하고 모델을 만들어 최적의 마케팅 전략을 제안"

분석 목표

1. EDA
- 다양한 데이터 탐색 기법을 사용
- 이용 내역 및 고객 데이터 특징 및 경향성 파악

2. 이용 내역 분석
- FA (요인 분석)
- K – means clustering

3. 휴면 고객 예측 모델
- 아이디 별 그룹화 및 파생변수 생성
- 의사결정나무를 통한 예측 모델 구축 및 변수중요도 파악

고객 마케팅 전략 제안

데이터 이해 및 전처리

데이터 이해

탑승 내역

데이터 사이즈: **(83083,48)**
데이터 기준 날짜: **2018-01-08**

고객 정보

데이터 사이즈: **(24523,27)**
데이터 기준 날짜: **2018-01-30**

아이디 별 그룹화, 파생변수 생성

이용 내역 분석 데이터 셋

*취소 내역을 제외한 탑승내역 67,929개의 데이터 사용

휴먼 고객 예측용데이터 셋

* 탑승내역에 존재하는 6,850개 고객 데이터 분석

병합 데이터 전처리

● **불필요 변수 제거**

단지구분,단지유형,멤버쉽 등 불필요한 22개의 변수 제거

● **차량 정비로 의심 되는 아이디 'eunji7805' 제거**

아이디	생년월일	연령	등급	…	예약상태	사용시작	사용종료
eunji7805	1960-01-01	52	GRDA02	…	정비	2017-01-13 15:49:00	2017-01-13 15:49:00
eunji7805	1960-01-01	52	GRDA02	…	취소	NaT	2017-01-13 15:56:00

● **거주단지가 관리자, 탈퇴회원인 데이터 제거**

아이디	생년월일	연령	등급	…	거주단지	예약상태	단지유형
sasty131	NaN	NaN	NaN	…	☆탈퇴회원	NaN	NaN
msyoo64	1964-02-14	52	NaN	…	★관리자	세차	업무용

● **등급, 최초탑승일시, 최근탑승일시 결측치 제거**

아이디	생년월일	연령	등급	…	예약상태	최초탑승일시	최근탑승일시
taeyang1234	1989-11-29	32	NaN	…	종료	NAN	NAN
hyun12	1982-06-29	39	NaN	…	취소	NAN	NAN

● **이용시간, 예약대기시간 (시), 총요금 입력값 –(마이너스)인 데이터 제거**

아이디	생년월일	연령	등급	…	이용시간	예약대기시간 (시)	총요금
euna7308	1973-08-26	43	GRDA02	…	5.75	12.18	-20100
aoaoqw	1990-12-21	26	GRDA02	…	–16.65	57.51	28810

병합 데이터 전처리

● 연장요금, 반납지연요금, 최소패널티, 거리요금 결측치를 0으로 대치

아이디	생년월일	연령	…	연장지연요금	반납지연요금	취소페널티
jjy2837	1989-11-29	28	…	NAN	NAN	NAN

jjy2837	1989-11-29	28	…	0	0	0

● 쿠폰 변수의 결측치를 “사용안함”으로 변경

아이디	생년월일	연령	…	연장지연요금	쿠폰	취소페널티
youmi0217	1988-11-27	29	…	0	NAN	5000

youmi0217	1988-11-27	29	…	0	사용안함	5000

● 생년월일 NaN, 연령이 0으로 표기된 데이터 → 해당 대여존과 차종의 사용자 평균연령으로 대체

아이디	생년월일	연령	대여존	차종
sneakersa	NaN	0	고려대하나스퀘어	모닝
qmffnzosel	NaN	0	평택소사벌이곡6	모닝

아이디	생년월일	연령	대여존	차종
sneakersa	1991-01-01	26	고려대하나스퀘어	모닝
qmffnzosel	1985-01-01	32	평택소사벌이곡6	모닝

병합 데이터 전처리

● 잘못 입력된 성별 변수의 입력 값 변경

아이디	연령	성별	…	이용거리	총요금	누적이용금액
gthasu	38	1900-02-15 00:00:00	…	57	22390	247460
skykoko1385	49	1901-06-07 00:00:00		59	44340	280910

gthasu	38	남	…	57	22390	247460
skykoko1385	49	여		59	44340	280910

● 예약요금 입력값 “59080”는 0으로 변경

아이디	연령	성별	…	이용거리	예약요금	누적이용금액
yumki68	48	남	…	35	“59080””	466450

yumki68	48	남	…	35	0	466450

● 반납지연요금 입력값 “75680”에서 “ ”(큰따옴표) 제거

아이디	연령	성별	…	이용거리	반납지연요금	누적이용금액
k11111008	37	남	…	32	“75680””	1073340

k11111008	37	남	…	32	75680	1073340

병합 데이터 전처리

- 예약상태는 이용중을 제외한 취소와 종료만 사용

아이디	연령	성별	…	예약상태	사용시작	사용종료
ko4ko4	39	남	…	이용중	2018-01-02 13:31:00	NaT
hjjh2	37	여	…	이용중	2018-01-05 18:54:00	NaT

- 이용거리가 있지만 이용시간이 0인 데이터 제거

아이디	연령	성별	…	이용거리	이용시간(시)	사용시작	사용종료
campos2	33	여	…	1.6	0	NaT	2017-01-01 03:00:00
pk1021	33	남	…	0.015	0	2017-01-01 07:45:00	2017-01-01 07:46:00

- 이용거리가 있지만 예약상태가 취소인 데이터 종료로 변경

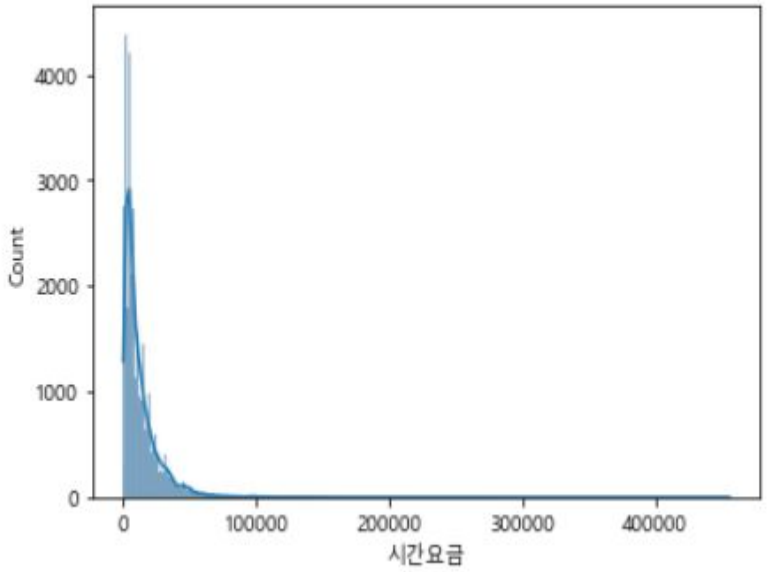
아이디	연령	성별	…	이용거리	예약상태	사용시작	사용종료
sy2018	41	여	…	8	취소	2017-07-18 15:52:00	2017-07-18 17:06:00
bks116	28	남	…	34	취소	2017-08-10 04:57:00	2017-08-10 11:26:00

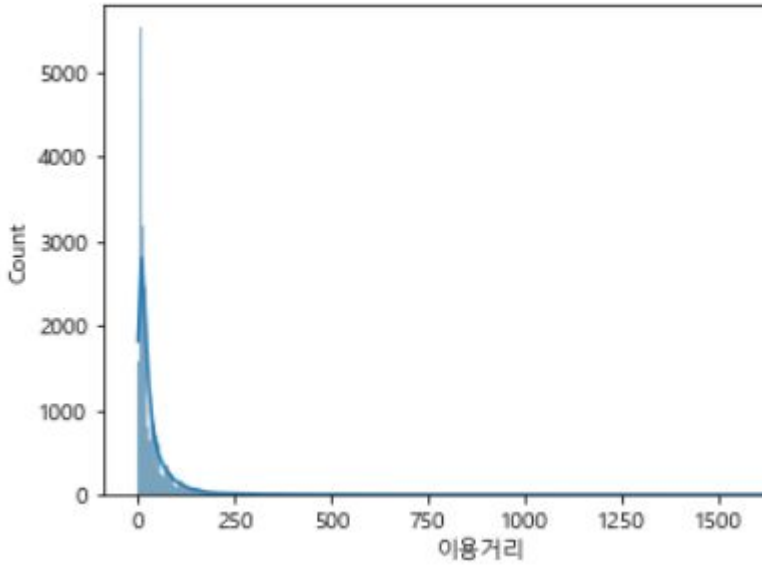
sy2018	41	여	…	8	종료	2017-07-18 15:52:00	2017-07-18 17:06:00
bks116	28	남	…	34	종료	2017-08-10 04:57:00	2017-08-10 11:26:00

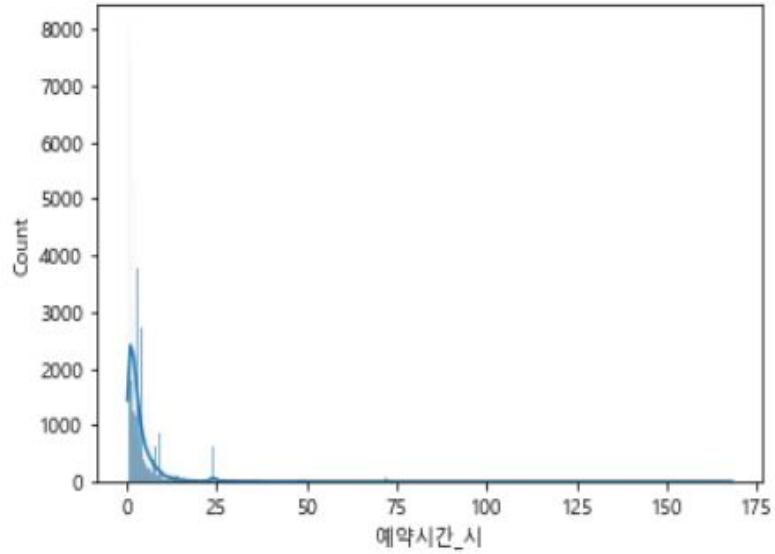
이용 내역 분석

EDA : 이용 내역 분석 데이터

	시간요금	이용거리	총요금	예약요금	연장요금	반납지연요금	거리요금	누적이용금액	예약시간_시	예약대기시간_시
count	67934.000000	67934.000000	67934.000000	67934.000000	67934.000000	6.793400e+04	67934.000000	6.793400e+04	67934.000000	67934.000000
mean	12892.058984	37.278888	18602.878897	11100.349972	1674.424883	9.047570e+02	5711.024818	1.065576e+06	4.038020	11.446865
std	15401.436681	66.751319	23538.425259	13871.708544	4343.490994	2.014893e+04	10236.458663	1.368812e+06	7.054832	41.371344
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000e+00	0.166667	0.000000
25%	3540.000000	7.000000	5420.000000	3540.000000	0.000000	0.000000e+00	1050.000000	2.176500e+05	1.000000	0.133333
50%	7800.000000	17.000000	11270.000000	6560.000000	0.000000	0.000000e+00	2560.000000	5.738600e+05	2.166667	0.566667
75%	16627.500000	41.000000	23060.000000	14160.000000	1760.000000	0.000000e+00	6290.000000	1.314710e+06	4.333333	6.866667
max	455560.000000	1772.000000	502820.000000	455560.000000	141480.000000	2.710400e+06	265800.000000	9.446410e+06	168.000000	799.350000

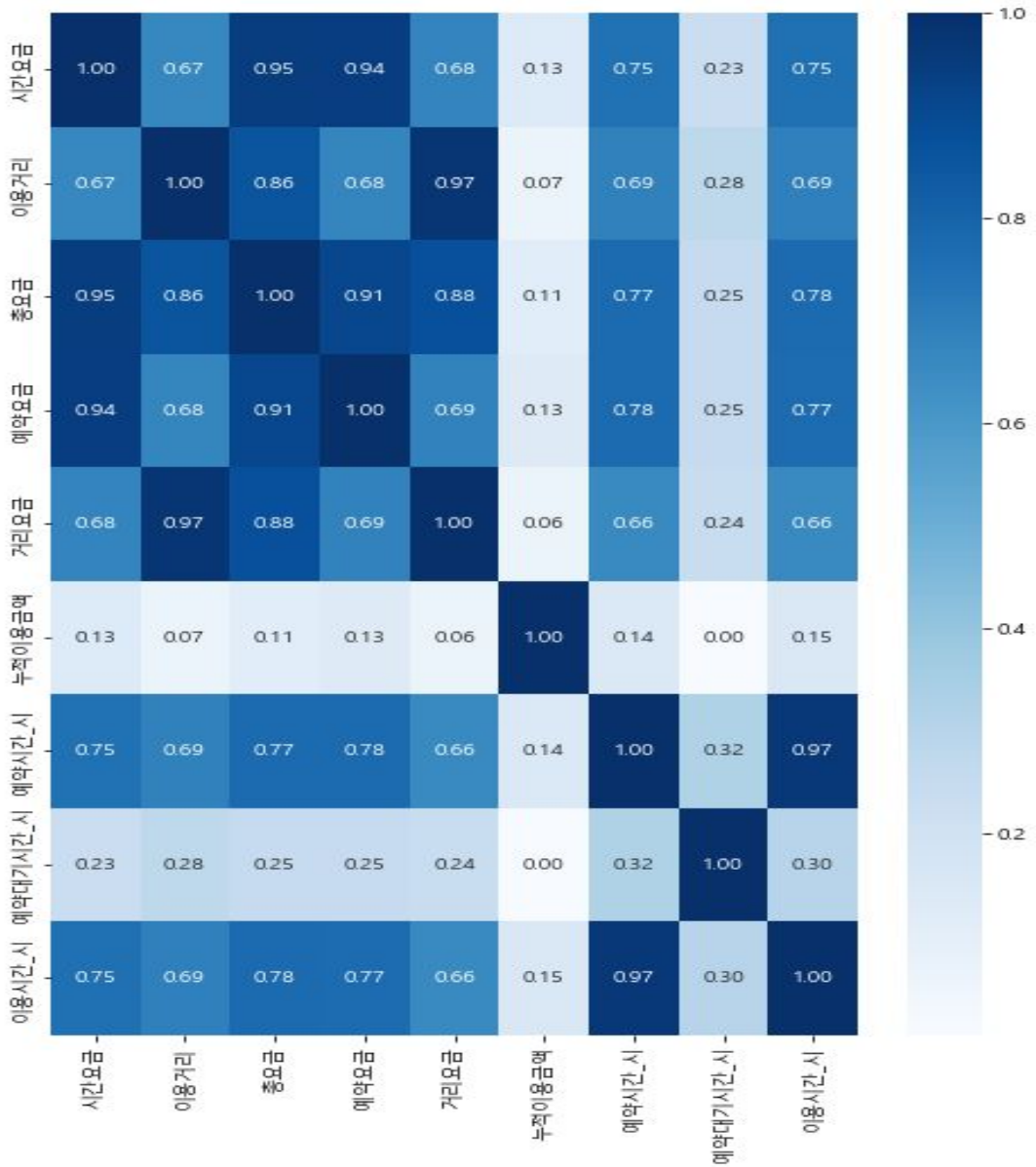




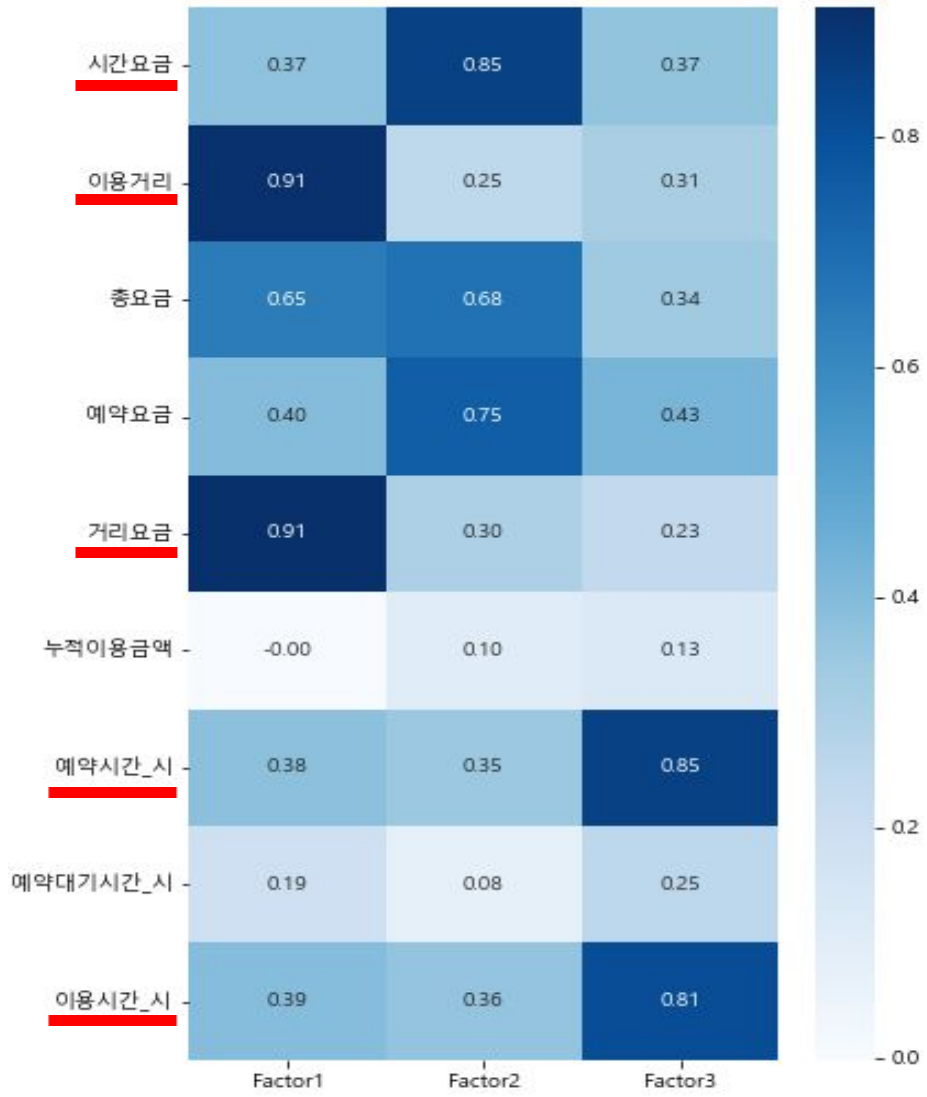


EDA : 이용 내역 분석 데이터

● 연속형 변수간의 상관관계



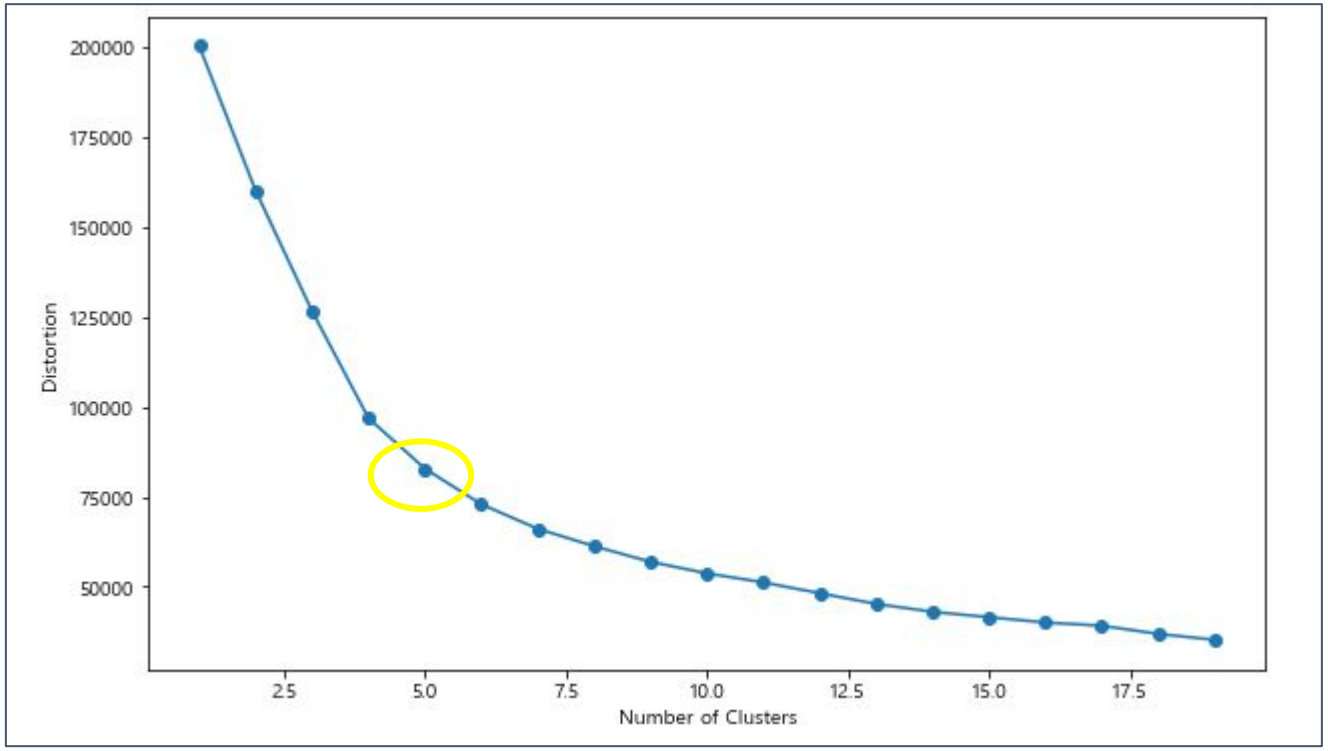
● FA 요인 분석



Factor 정의 **Factor1** : 거리 요인 **Factor2** : 요금 요인 **Factor3** : 시간 요인

k - means clustering

- **elbow method를 통해 최적의 군집 수 결정**



k_means_cluster	Factor1	Factor2	Factor3
0	-0.265247	-0.395122	-0.184400
1	-0.126905	0.992935	-0.115622
2	0.268988	-6.057446	14.409380
3	5.080134	0.262572	-0.637829
4	-0.068533	1.517965	3.064734

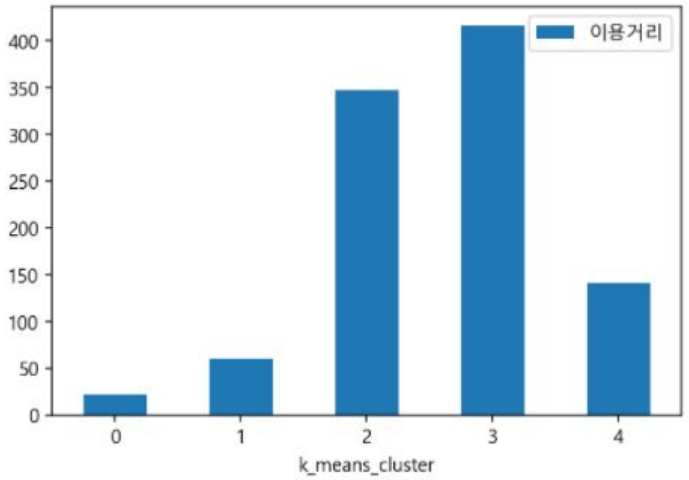
k_means_cluster	Factor1	Factor2	Factor3
0	0.425907	0.371128	0.278353
1	0.779585	0.921200	0.500055
2	4.493125	4.515455	4.654004
3	2.992957	2.099027	2.030452
4	1.541457	2.265370	1.730562

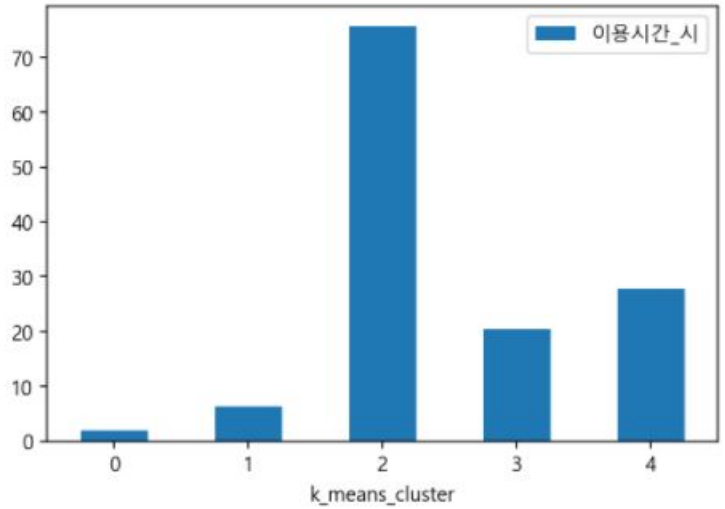
*군집별 데이터량

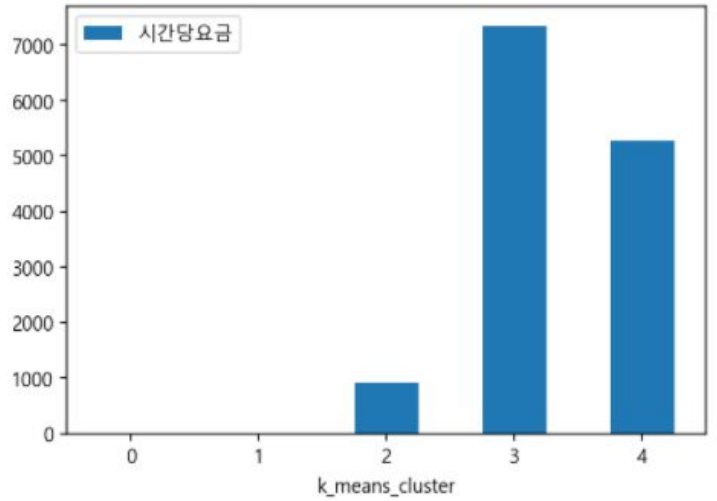
0	52,519
1	12,895
2	142
3	911
4	1,462

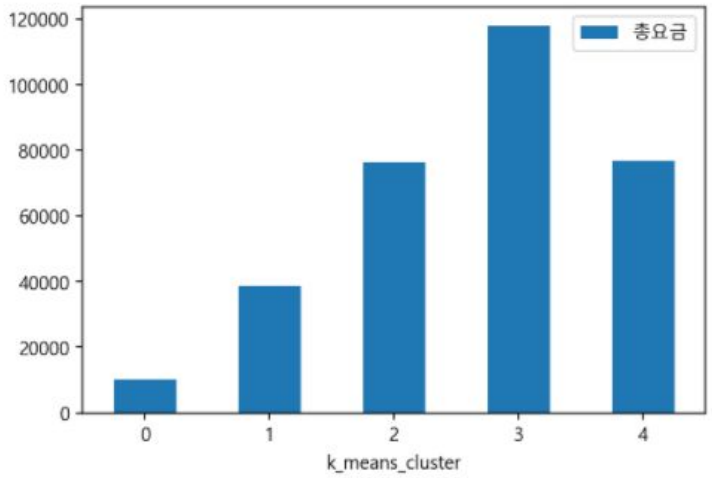
시각화 및 해석

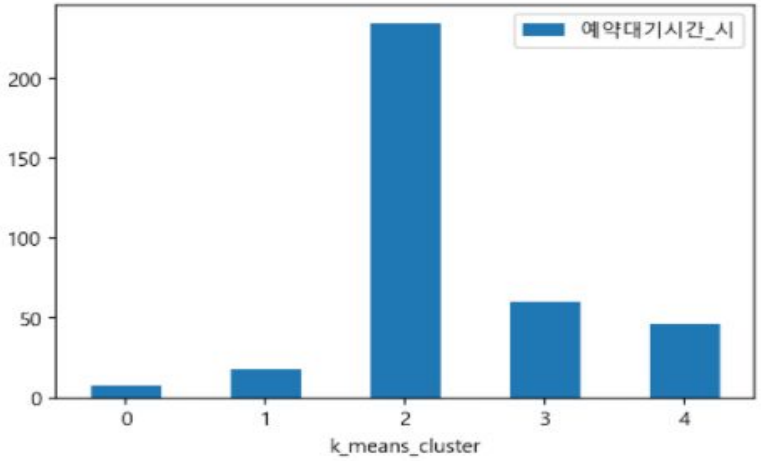
- 군집별 특성







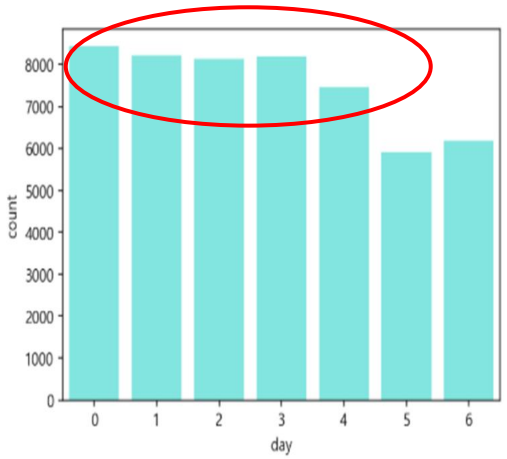


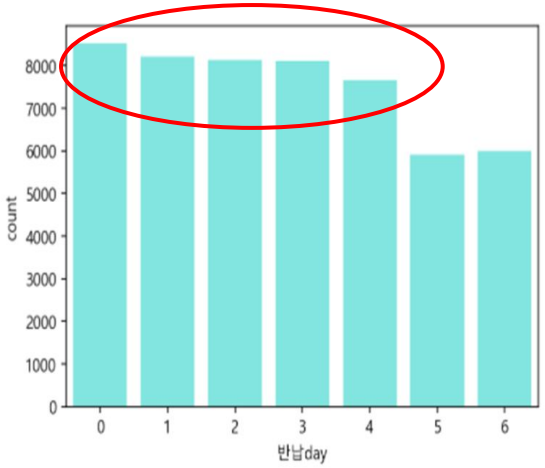


시각화 및 해석

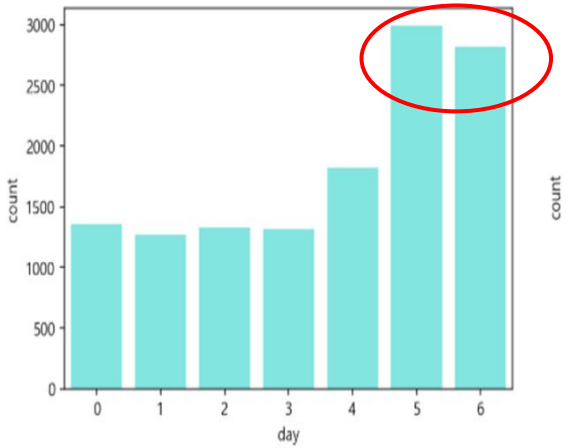
- 군집별 사용 시작 및 반납 요일

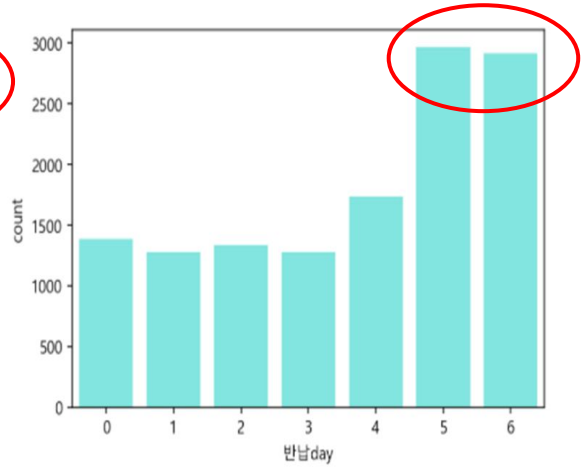
cluster_0



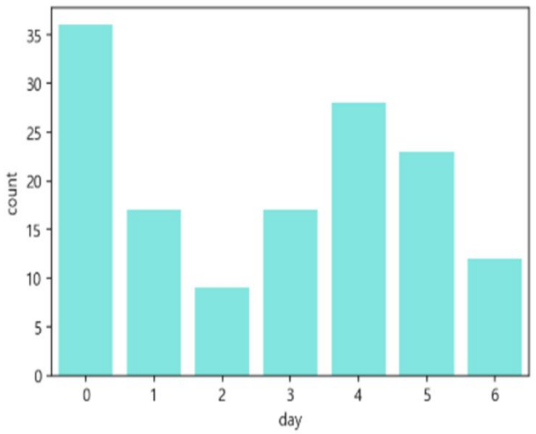


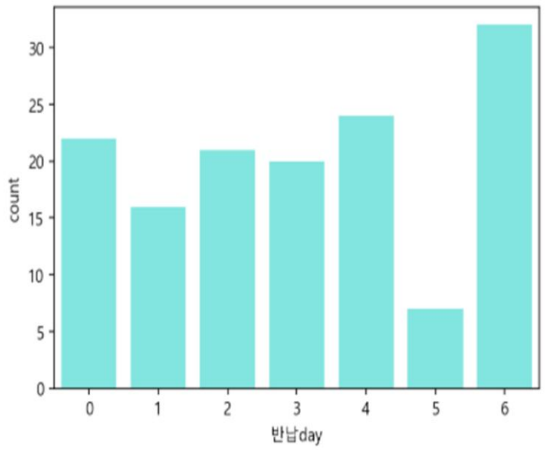
cluster_1





cluster_2

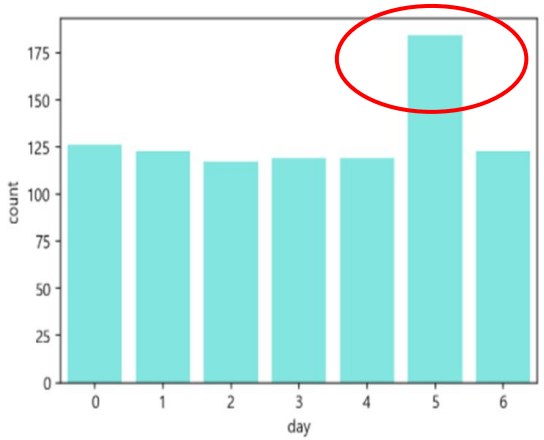


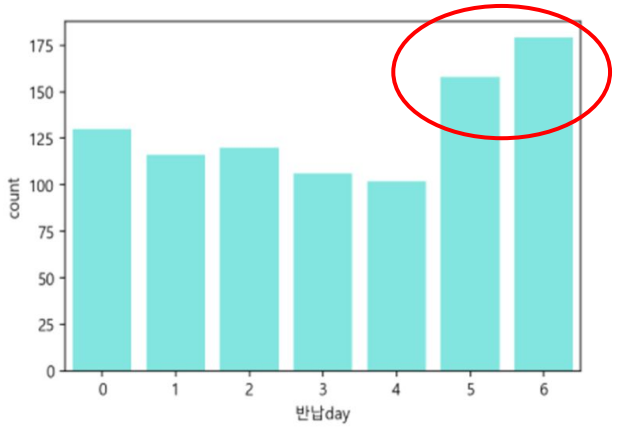


x축의 0~6은 요일의 순서 (월요일~일요일)를 의미

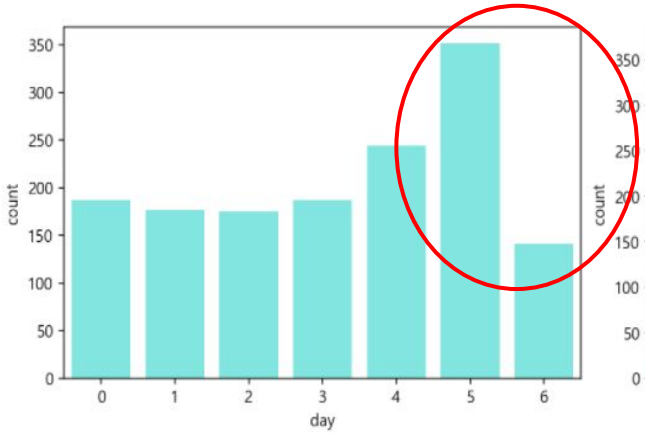
시각화 및 해석

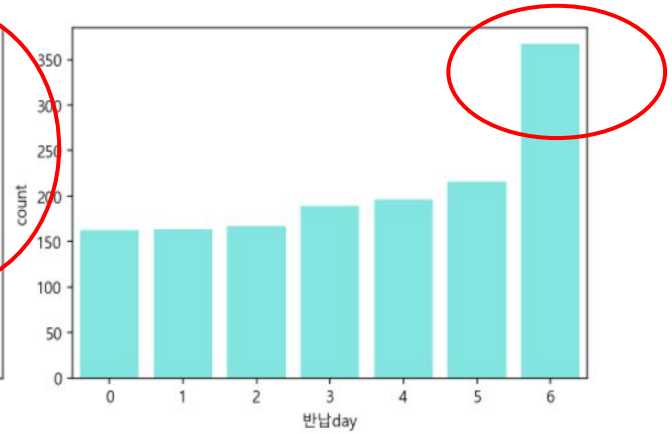
cluster_3





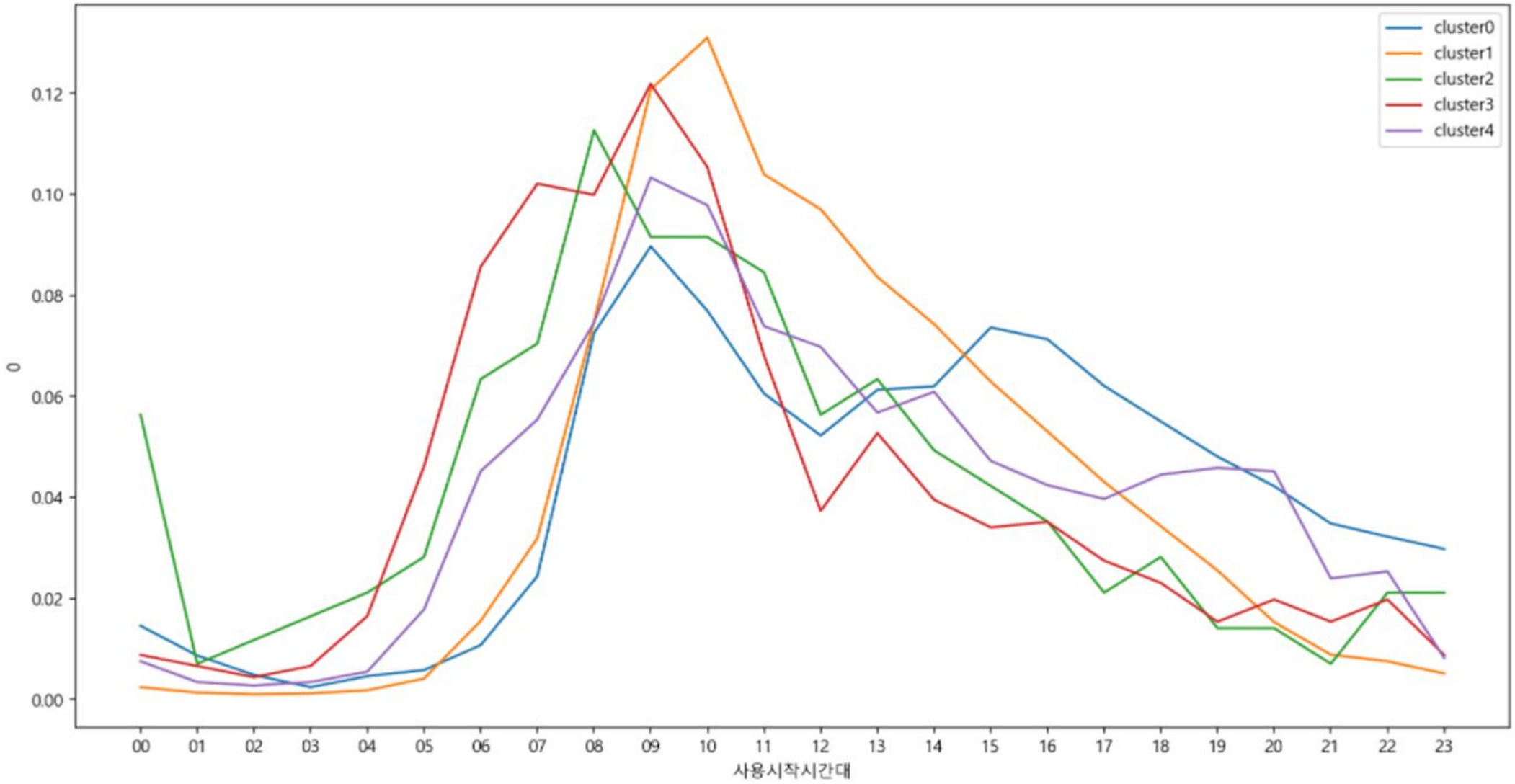
cluster_4





시각화 및 해석

- 군집별 사용시작 시간대 비율



이용내역 분석 결론 및 해석

- 군집별 특성

0번 : 가까운 거리를 이용하는 업무나 일상생활 사용 목적으로 짧게 이용하는 클러스터

1번 : 가까운 거리를 여가를 목적으로 요금을 신경쓰지 않고 사용하는 클러스터

2번 : 장기적으로 대여하는 클러스터

3번 : 장거리를 당일치기이나 1박2일 여행 및 드라이브을 목적으로 사용하는 클러스터

4번 : 여행을 목적으로 하지만 짧은 거리를 교통수단으로만 사용하는 클러스터

휴면 고객 예측 모델

휴면 고객 예측 데이터 파생변수 생성

● 날짜, 시간 관련 변수

가입후경과일수	카셰어링 서비스 회원 가입 후 경과한 일 수	**주사용시간대**	아이디별 가장 많은 사용시작 시간대
최초탑승후경과일수	카셰어링 서비스를 처음 사용한 후 경과한 일 수	**출퇴근시간대사용 비율**	출퇴근시간대(8~10시,17~19시)에 사용한 비율
최근탑승후경과일수	카셰어링 서비스를 마지막으로 사용한 후 경과한 일 수	**예약일시 count**	아이디별 예약한 날의 횟수
예약시간 합계, 평균, 최대	아이디별 예약시간의 합계, 평균, 최대	**주말사용률**	전체 서비스 사용 횟수 중 주말에만 사용한 비율
이용시간 합계, 평균, 최대	아이디별 이용시간의 합계, 평균, 최대	**사용주기**	아이디별 카셰어링 서비스 사용주기 (사용시작날짜최대 - 사용시작날짜최소) / 사용횟수
예약대기시간 합계, 평균, 최대	아이디별 예약대기시간의 합계, 평균, 최대		

● 요금 및 거리 관련 변수

누적이용금액	아이디별 총 누적이용금액	**반납지연요금합계, 반납지연율**	아이디별 반납지연요금 합계 및 반납 지연 비율
총요금 합계, 평균, 최대	아이디별 총요금의 합계, 평균, 최대	**연장요금합계, 연장빈도율**	아이디별 연장요금 합계 및 연장 비율
이용거리 합계, 평균, 최대	아이디별 이용거리의 합계, 평균, 최대	**취소패널티합계**	아이디별 취소패널티 요금 합계

휴면 고객 예측 데이터 파생변수 생성

● 그 외 변수

등급	고객 등급(GRDA01~06)	**종료**	사용 종료한 횟수
성별	고객 성별(남/여)	**취소**	예약을 취소한 횟수
주소	고객의 주소지(서울,인천,평택,남양주…)	**취소비율**	전체 사용횟수에서 예약을 취소한 비율
연령	고객의 연령(25,37….)	**쿠폰사용률**	전체 사용횟수에서 쿠폰을 사용한 비율
차종 count	아이디별 어떤 차종에 탑승했는지 count	**타도시방문비율**	주소와 대여존이 다른 내역의 사용 비율
대여존 count	아이디별 어떤 대여존을 사용했는지 count		

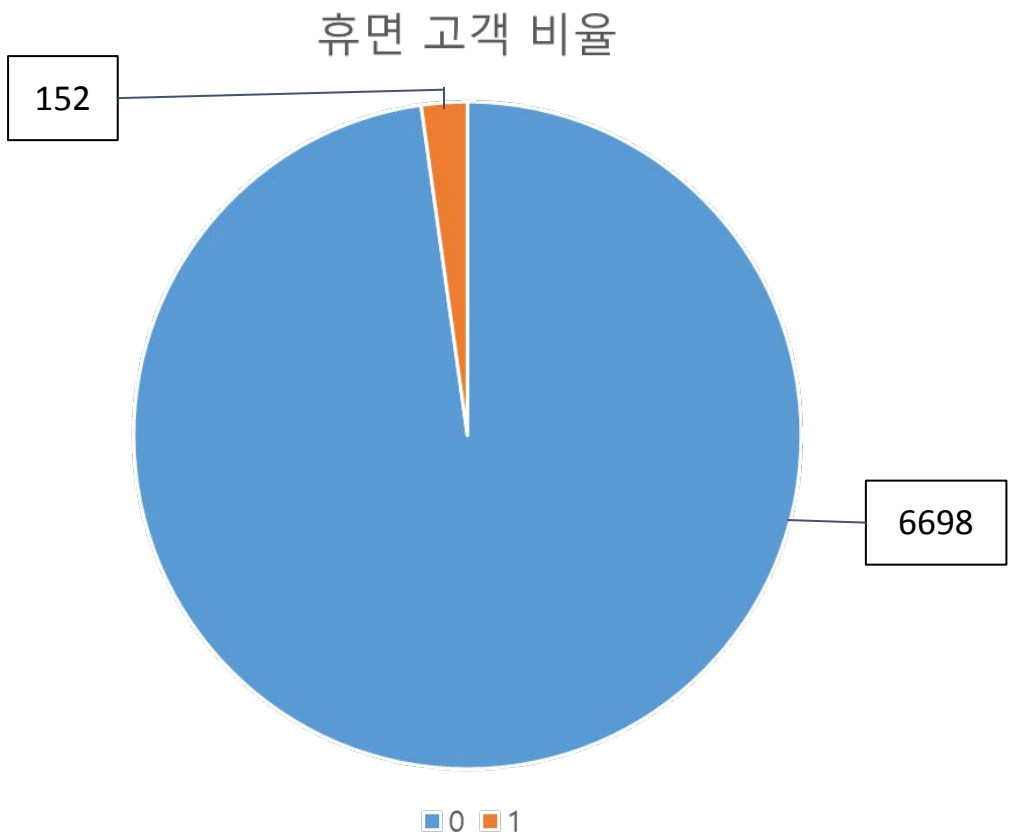
● 파생변수 생성 결과

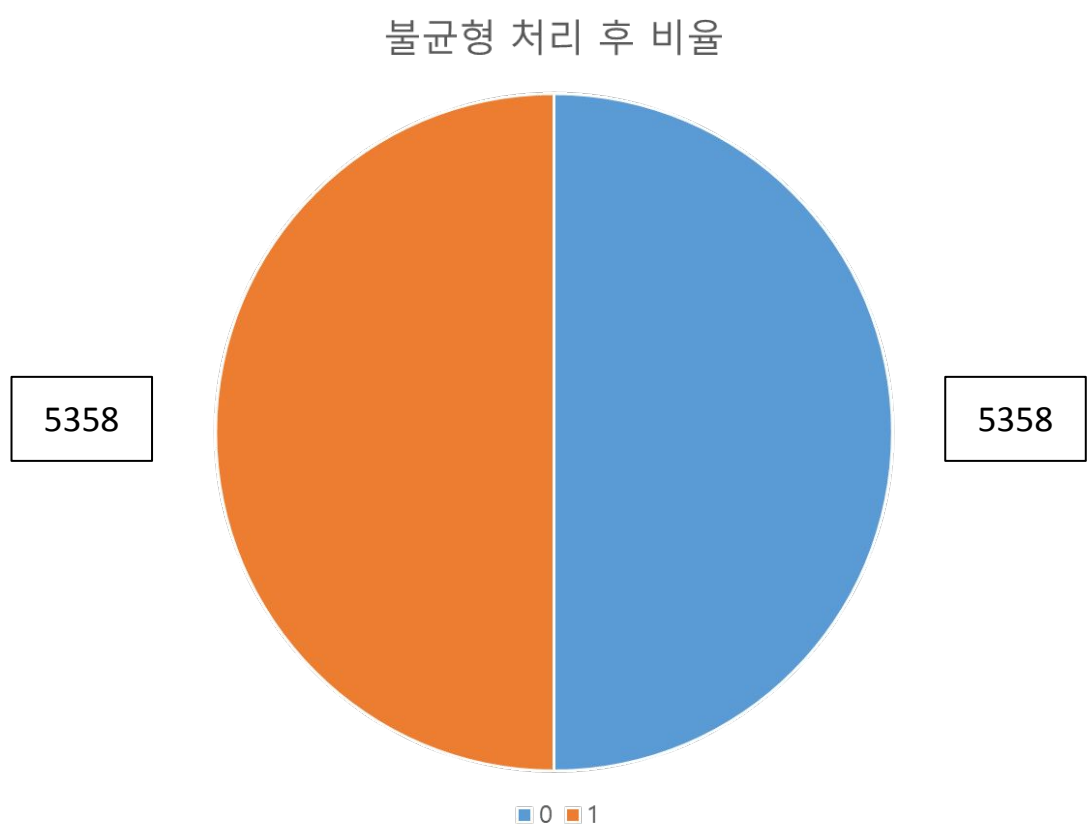
아이디	성별	연령	주소	…	이용거리_총합	이용거리_최대	이용거리_평균
jjy2837	남	48	서울	…	5013.0	398.0	111.4
chlong119	여	28	서울	…	493.0	233.4	123.2

(6851,41)

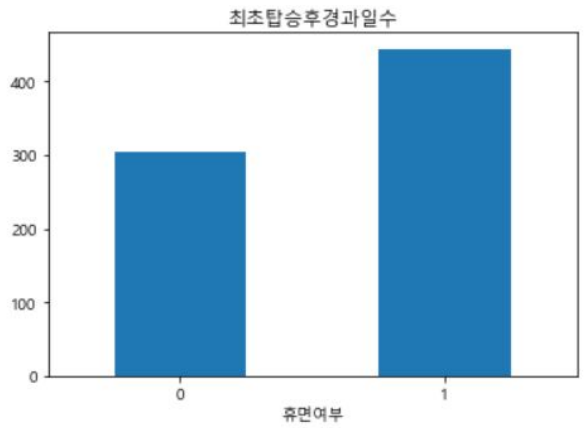
휴먼 고객 예측 데이터 비율

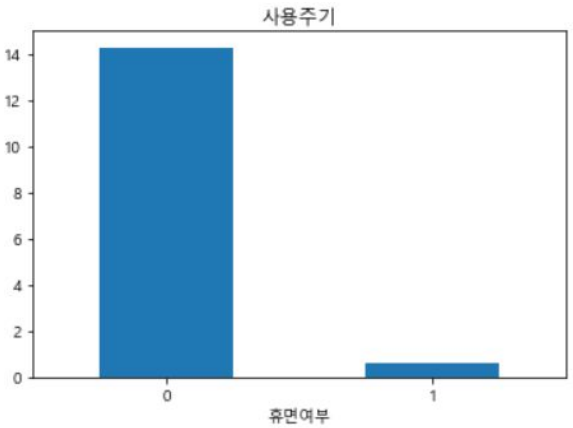
- 탑승내역에 존재하는 6850명의 고객 분석
- 데이터 날짜 기준(2018-01-30) 최근 1년간 사용 기록이 없는 고객을 휴면 고객으로 분류

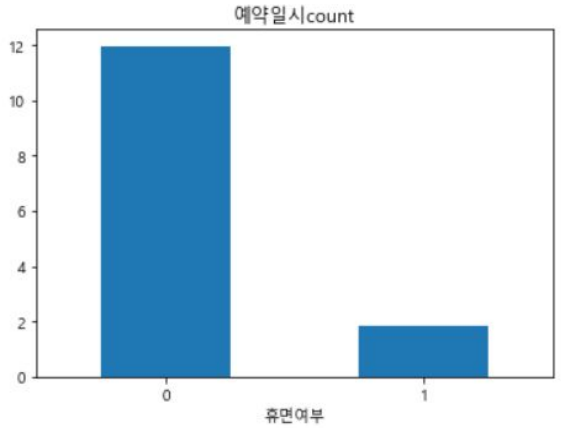


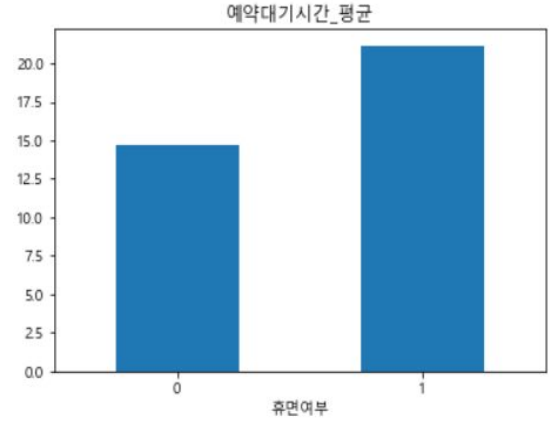


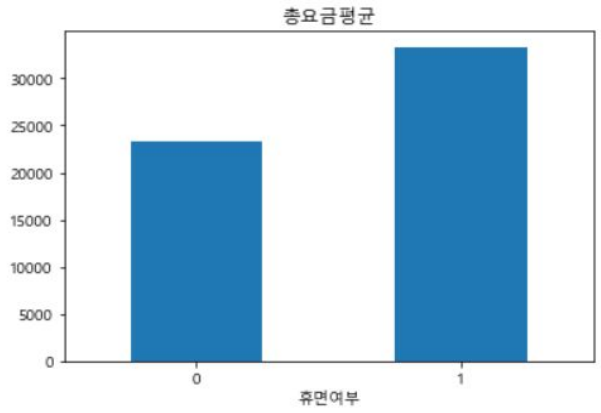
EDA : 휴면 고객 예측 데이터

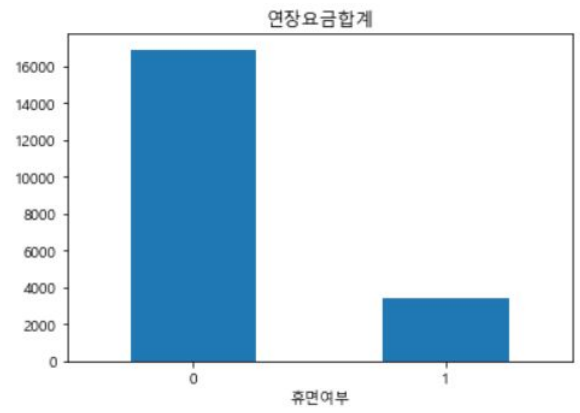


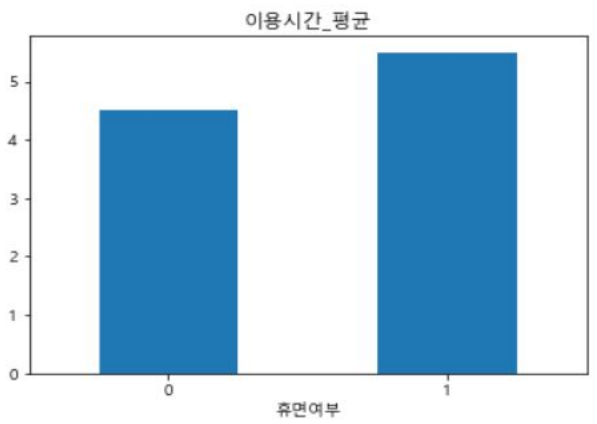


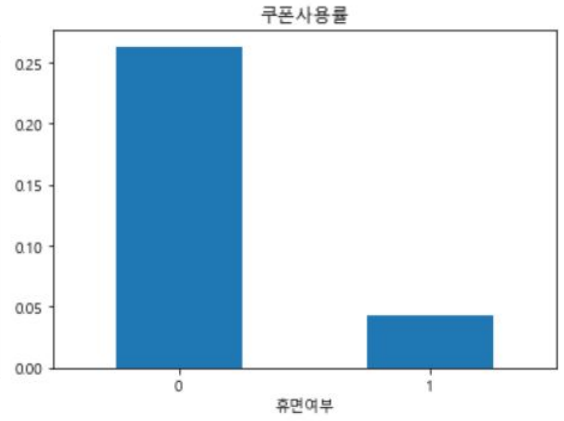






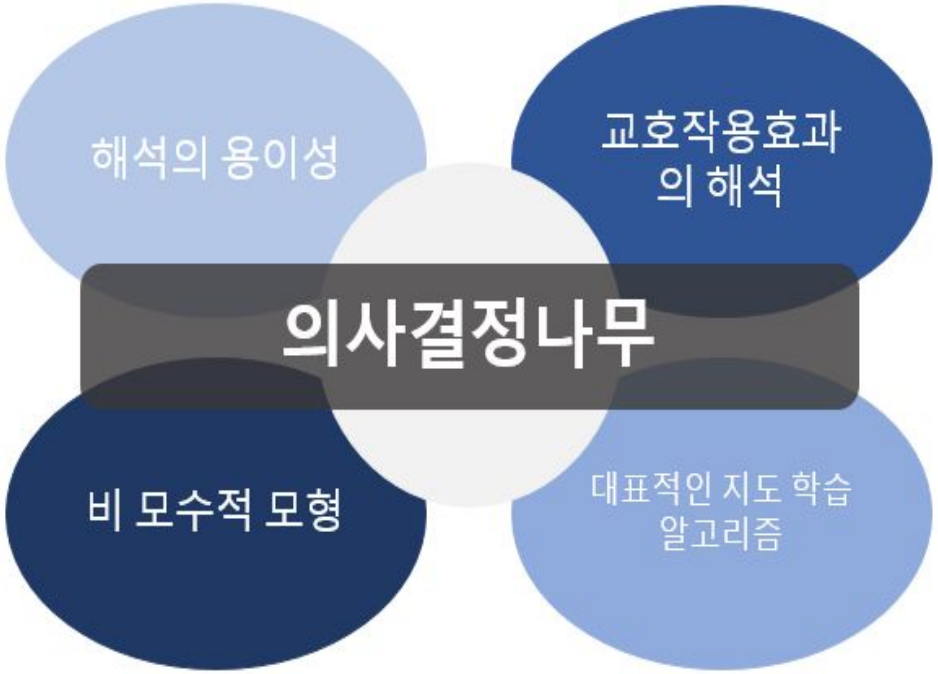


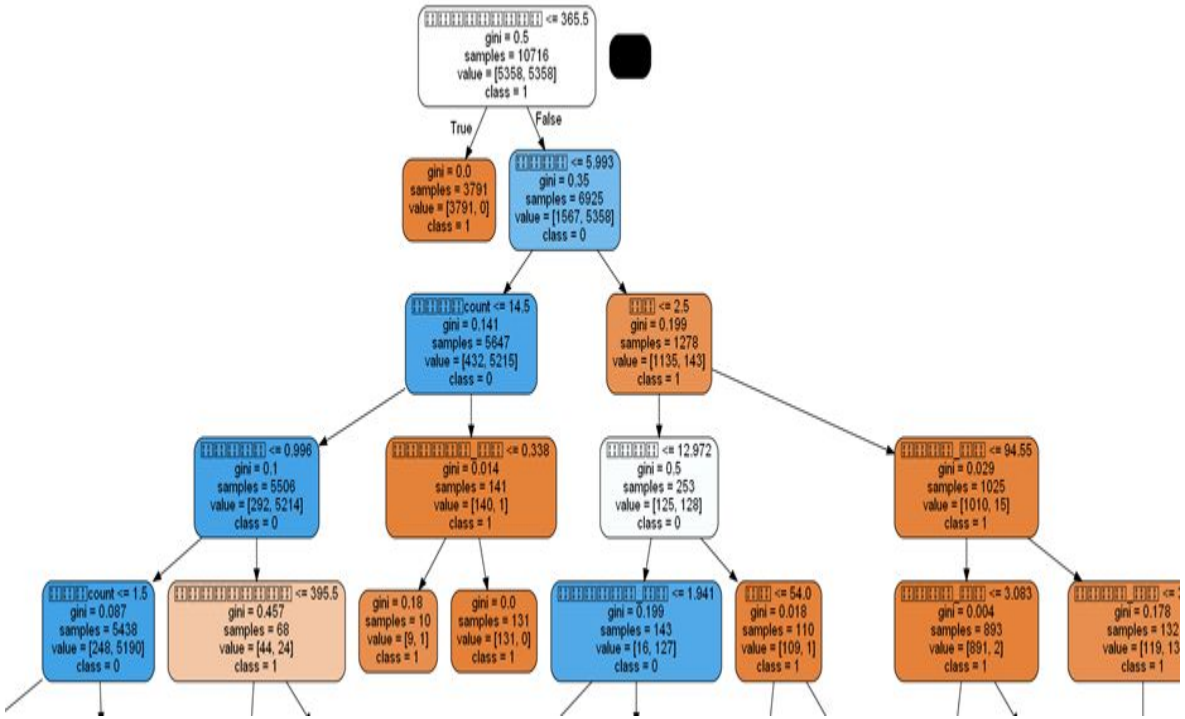




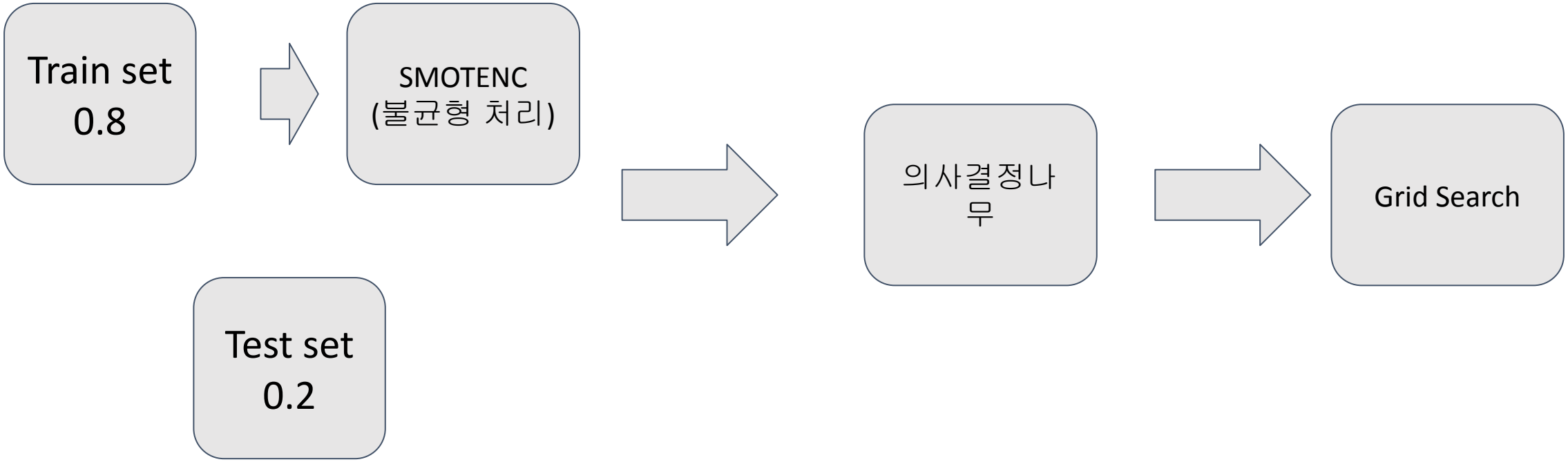
(0 : 일반고객 1 : 휴면고객)

모델링


해석의 용이성
교호작용효과의 해석
의사결정나무
비 모수적 모형
대표적인 지도 학습 알고리즘

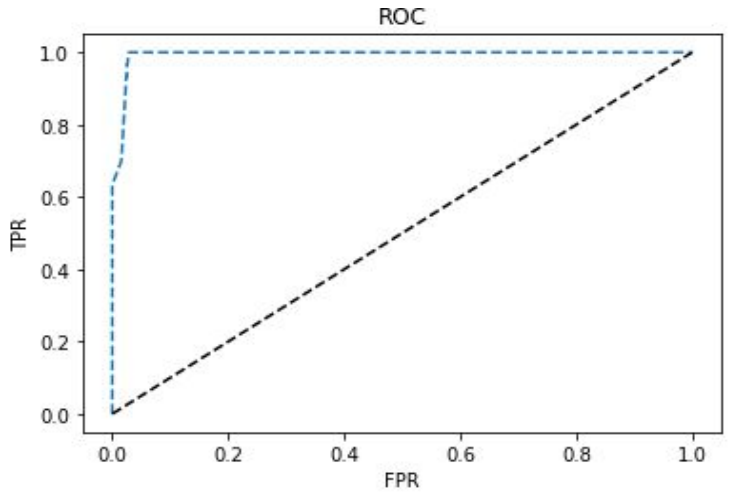

gini = 0.5
samples = 10716
value = [5358, 5358]
class = 1
True
False
gini = 0.0
samples = 3791
value = [3791, 0]
class = 1
gini = 0.35
samples = 6925
value = [1567, 5358]
class = 0
count <= 14.5
gini = 0.141
samples = 5647
value = [432, 5215]
class = 0
gini = 0.199
samples = 1278
value = [1135, 143]
class = 1
gini = 0.1
samples = 5506
value = [292, 5214]
class = 0
gini = 0.014
samples = 141
value = [140, 1]
class = 1
gini = 0.5
samples = 253
value = [125, 128]
class = 0
gini = 0.029
samples = 1025
value = [1010, 15]
class = 1
count <= 1.5
gini = 0.087
samples = 5438
value = [248, 5190]
class = 0
gini = 0.457
samples = 68
value = [44, 24]
class = 1
gini = 0.18
samples = 10
value = [9, 1]
class = 1
gini = 0.0
samples = 131
value = [131, 0]
class = 1
gini = 0.199
samples = 143
value = [16, 127]
class = 0
gini = 0.018
samples = 110
value = [109, 1]
class = 1
gini = 0.004
samples = 893
value = [891, 2]
class = 1

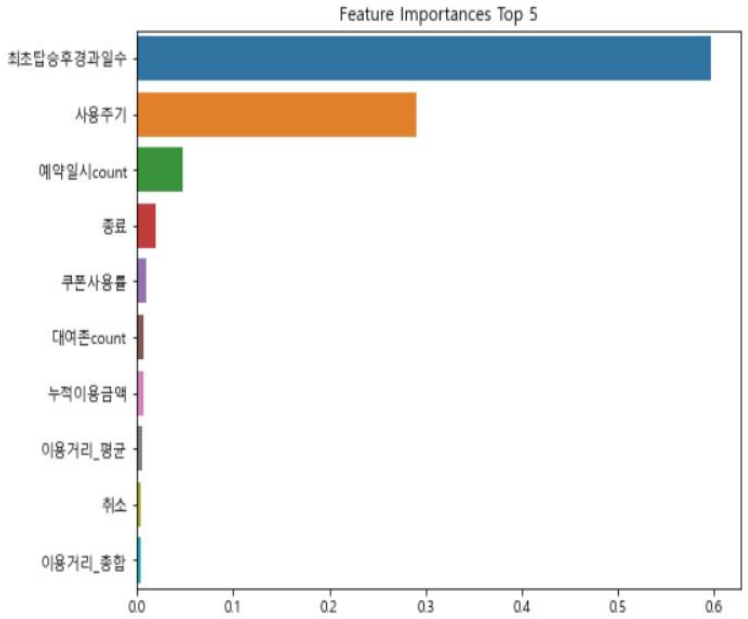
모델링



성능 평가 및 해석

accuracy	**0.974**
recall	**0.967**
precision	**0.453**
F1	**0.617**
AUC	**0.97**





모델 해석 및 결론

변수중요도에서 “최초탑승후경과일수” 변수가 가장 높은 수치를 보이는 것을 알 수 있다, 트리를 시각화하여 해석하였을 때, 가장 처음으로 보이는 규칙은 “최초탑승후경과일수 < 365.5”이다.

따라서 휴면고객을 분리하는 데 있어서 최초탑승후경과일수가 1년이 초과하였는지의 여부가 가장 중요함을 알 수 있다. 이를 통해 최초탑승후경과일수, 사용주기 등을 통해 잠재적인 휴면 고객들을 찾아볼 필요가 있다.

더불어 잠재휴면고객이라고 판단된다면 이 고객의 CLV(고객생애가치)에 따라 이 고객을 유지하는 전략과 포기하는 전략 중 하나를 선택할 수도 있다.

활용방안 및 기대효과

활용방안 및 기대효과

고객정보

사용 목적

탑승 내역

이용 내역

잠재 휴면 고객

활용방안 및 기대효과

● 이용 내역 분석

1. 예약시에 사용 목적을 선택하여 이용 군집을 예측하고 맞춤형 혜택 제공
(ex, 단순 이동의 경우 0번 또는 1번 군집으로 판단하여 일정 횟수 이상을 사용할 경우 혜택 제공,
사용 목적을 여행으로 선택했을 경우 3번 또는 4번 군집으로 판단하여 여행지와 관련된 상품 제공)

0번, 1번군집 : 단순이동 → 일정 횟수 이상 사용시 혜택 제공, 업무의 경우 사업자 혜택
2번 군집 : 장기대여 → 일정km이상 이용시 할인 혜택
3번 군집 : 드라이브 관련 상품 제공
4번 군집 : 여행 관광지, 숙박 관련 혜택 제공

맞춤형 서비스 제공으로 충성 고객 형성

● 휴면 고객 예측

1. 최초탑승후 1년이 지날시 혜택 제공
2. 사용주기가 짧고 최근 탑승이 오래된 고객에게 알림 적용

휴면 고객으로 예상되는 고객에게 맞춤형 서비스를 제공하여 기존의 고객 유지

감사합니다