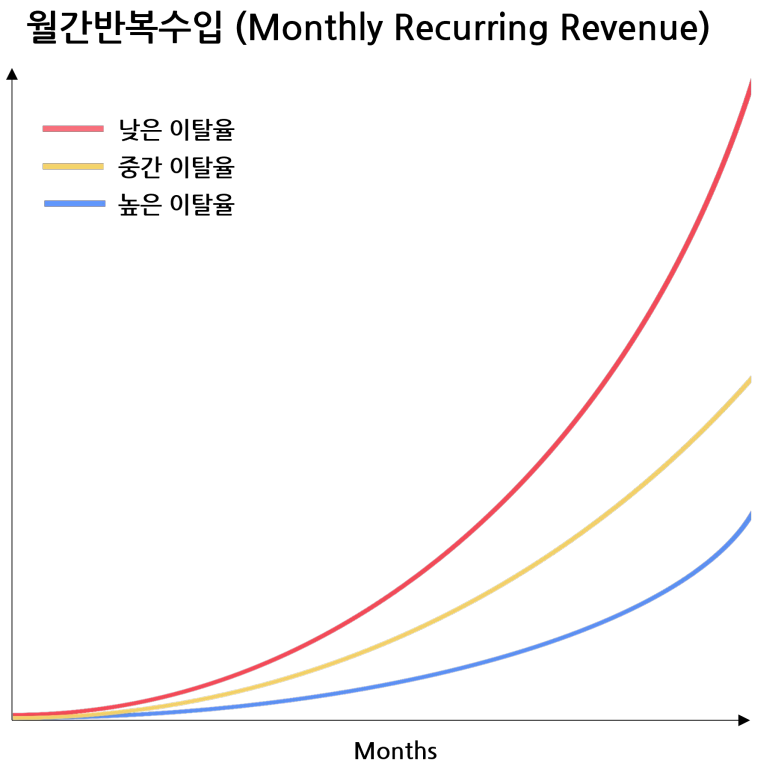
Machine Learning Project

음원 스트리밍 서비스 이탈 유저 예측

DS 11기 4팀 (박무연, 박지호, 안태진, 천정은)

1 분석의 필요성



매월 요금을 지불하는 구독 서비스에서

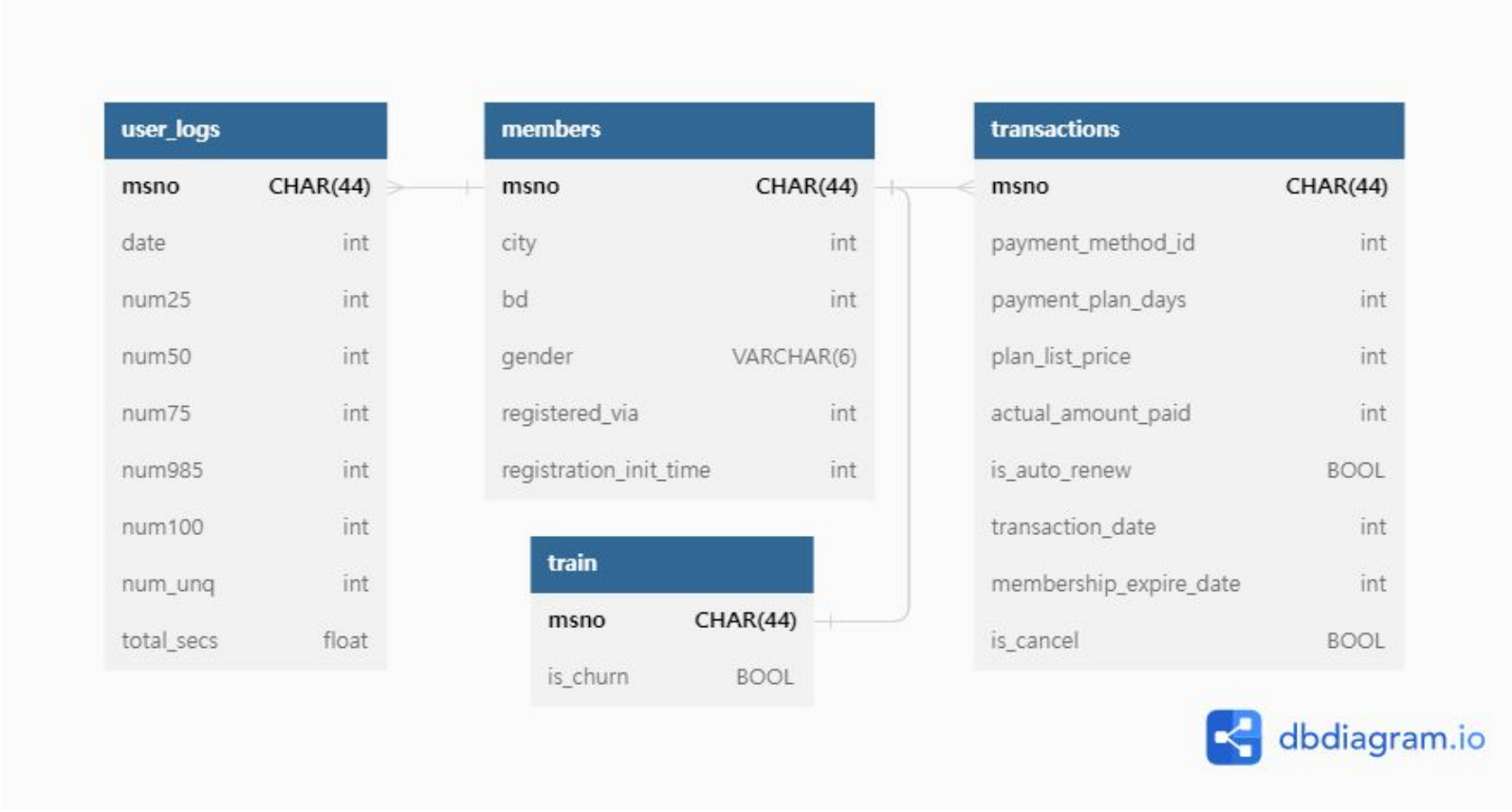
고객의 수 ∝ 매출

고객 이탈 → 수익

⬇️

2 데이터 소개

Data Source : Kaggle [KKBox's Churn Prediction Challenge]



2 데이터 소개

데이터 샘플링

User ID Mapping

컬럼 타입변경
(Object > Date)

EDA

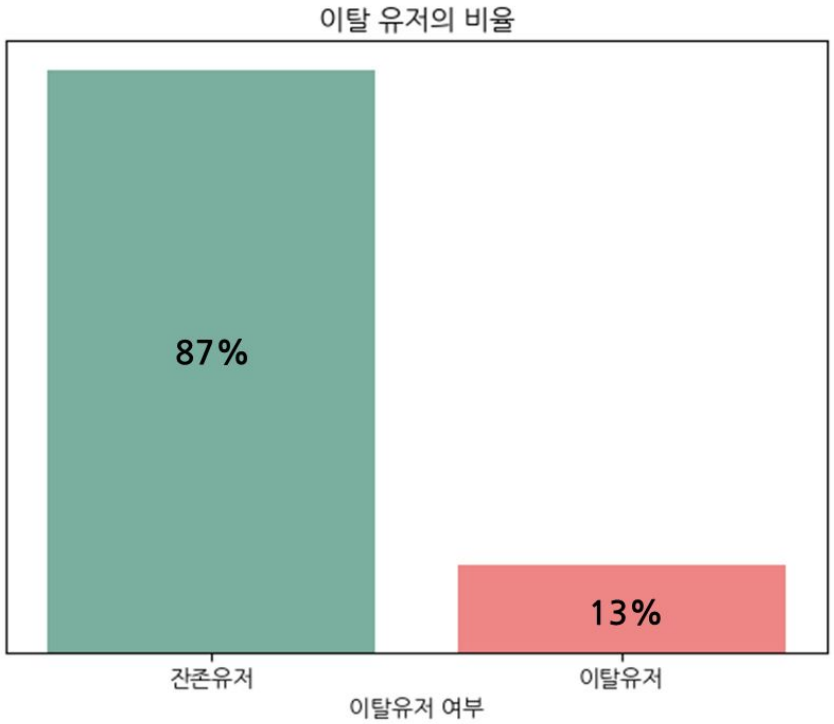
데이터 필터링

데이터 전처리

모델링

Source
https://pandas.pydata.org/about/citing.html
https://spark.apache.org/images/spark-logo-trademark.png

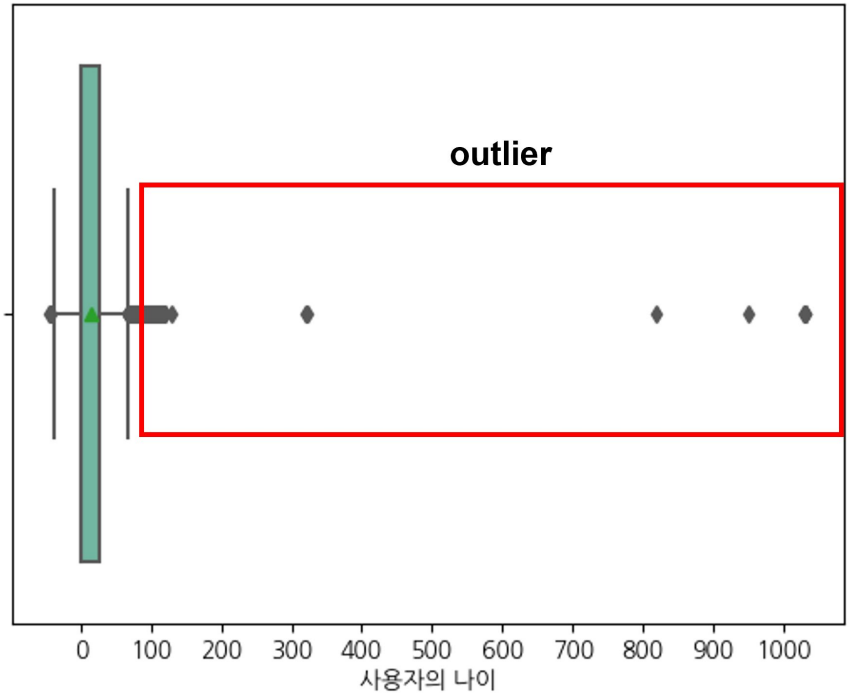
3 EDA - Train dataset



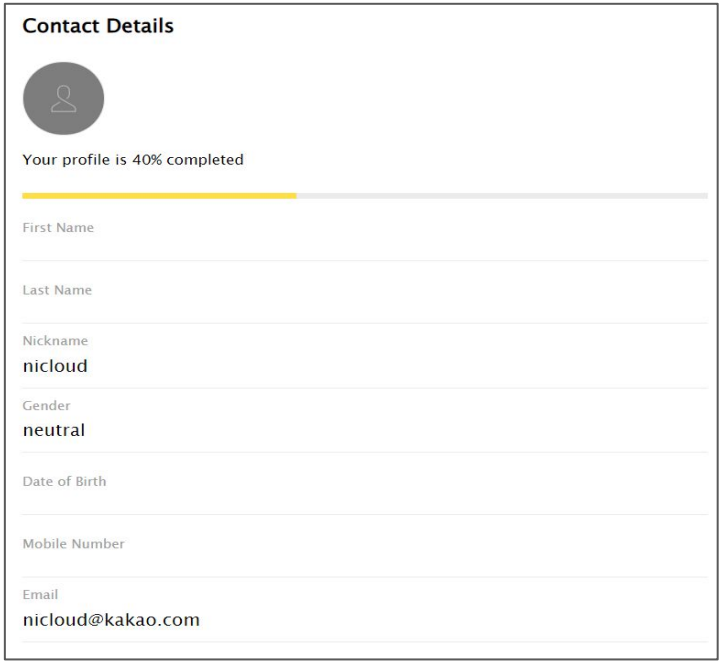
“ 이탈 유저(1)와 잔존 유저(0)의 비율에서 큰 차이를 보여 모델링에 유의 할 필요성을 확인”

3 EDA - Members Dataset

사용자의 나이 분포

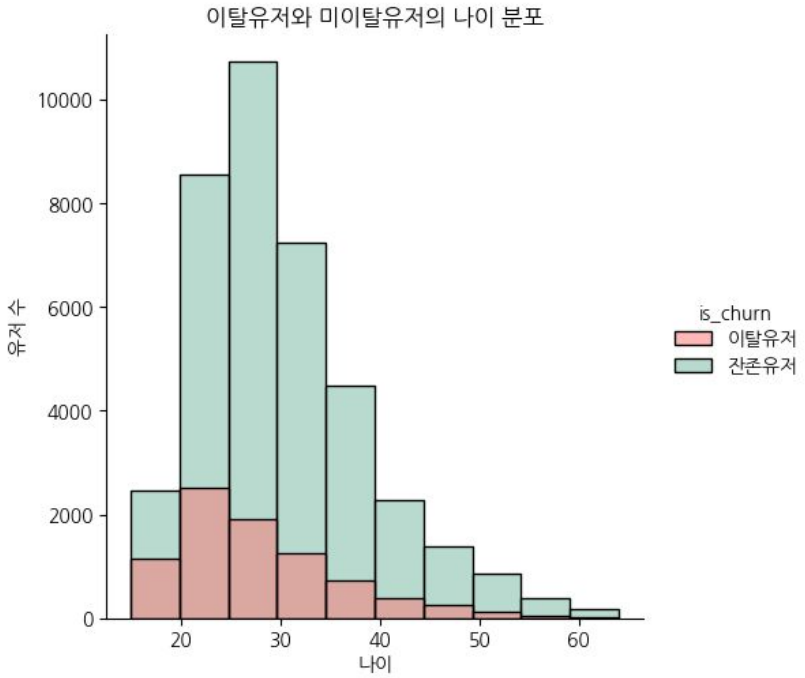


서비스 프로필 등록 화면

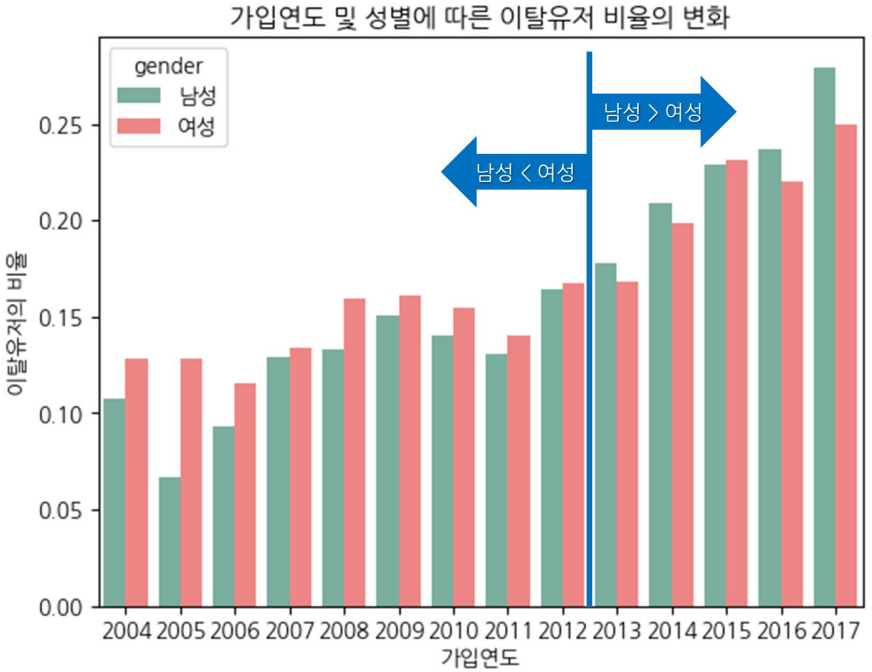


" 성별 / 나이 컬럼에 이상치를 확인. 서비스 확인 결과 고객이 직접 입력하기 때문. "

3 EDA - Members Dataset

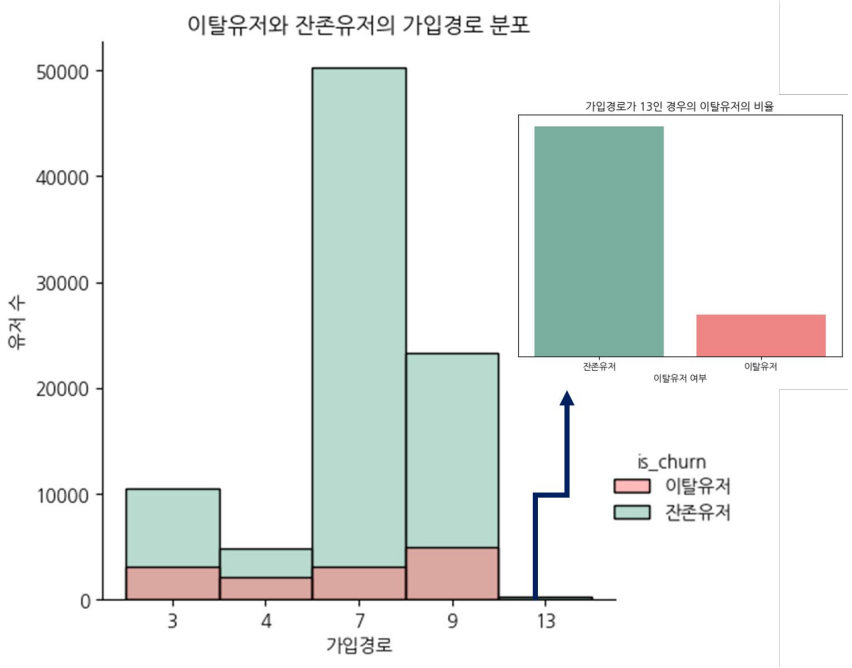


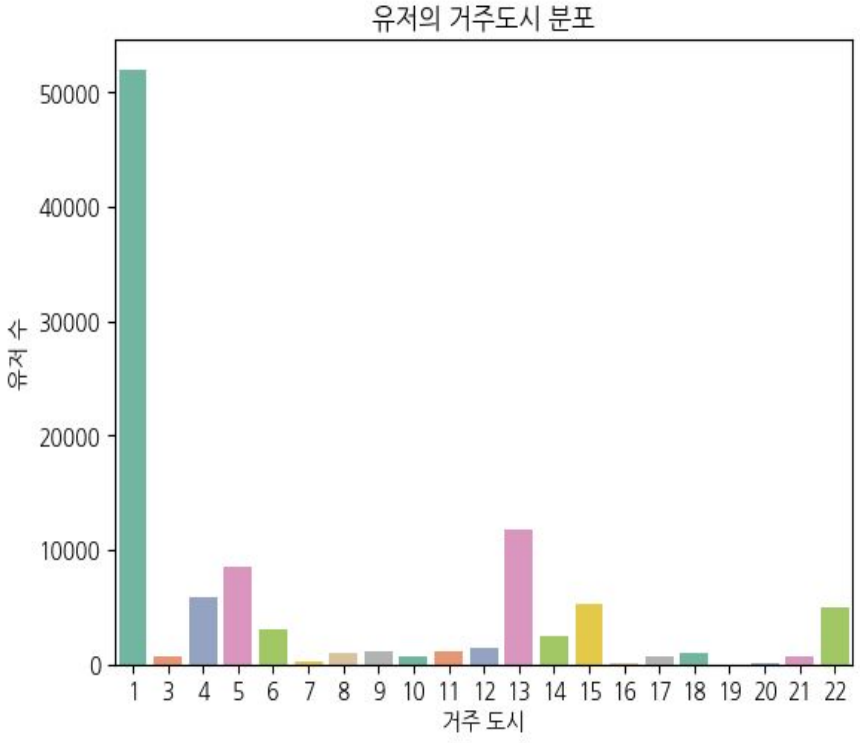
"10대 후반에서 20대 초반의 연령대가 다른 연령대보다 이탈률이 높음"



"최근에 가입한 유저일수록 이탈률이 높음"

3 EDA - Members Dataset

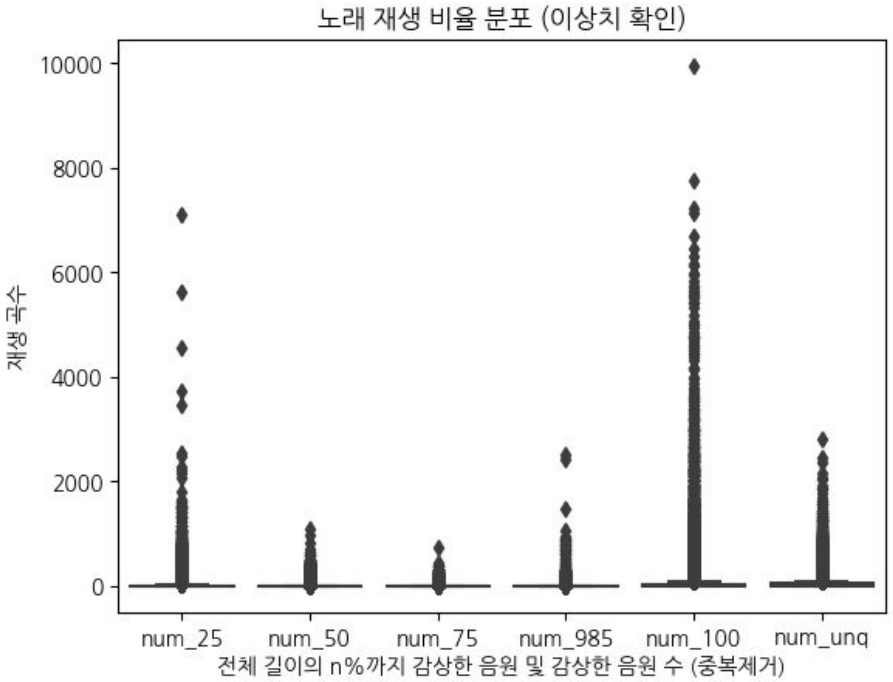


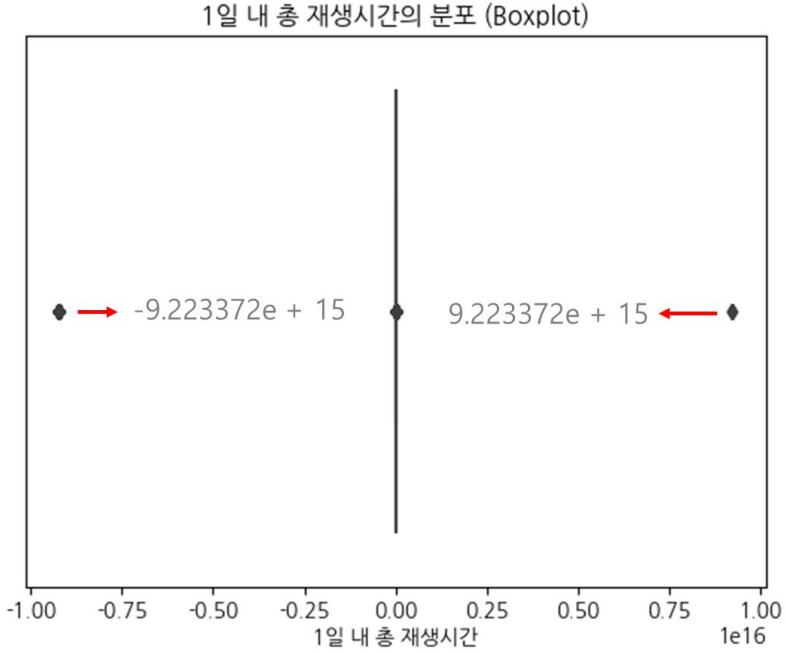


"가입경로가 7인 경우, 유저의 잔존률이 가장 높았음"
"가입경로가 9인 경우, 유저의 이탈률이 가장 높았음"

"1번 도시에 유저들이 많이 거주함"

3 EDA - User_logs Dataset





" 노래 재생 수의 경우 어디까지가 이상치인지 생각해볼 필요가 있었다."

3 EDA 및 전처리 - Transaction Dataset : 삭제대상

ID	…	총 재생 시간(sec)
1	…	**-9.22372e+15**
2	…	**130,561**
3	…	12,064

총 재생 시간

총 재생 시간 < 0 삭제

총 재생 시간 > 86,400 삭제

ID	num_25	num_50	num_75	…
1	**-1**	0	25	…
2	0	0	0	…
3	15	**-12**	3	…
4	**7049**	0	0	…

재생 횟수

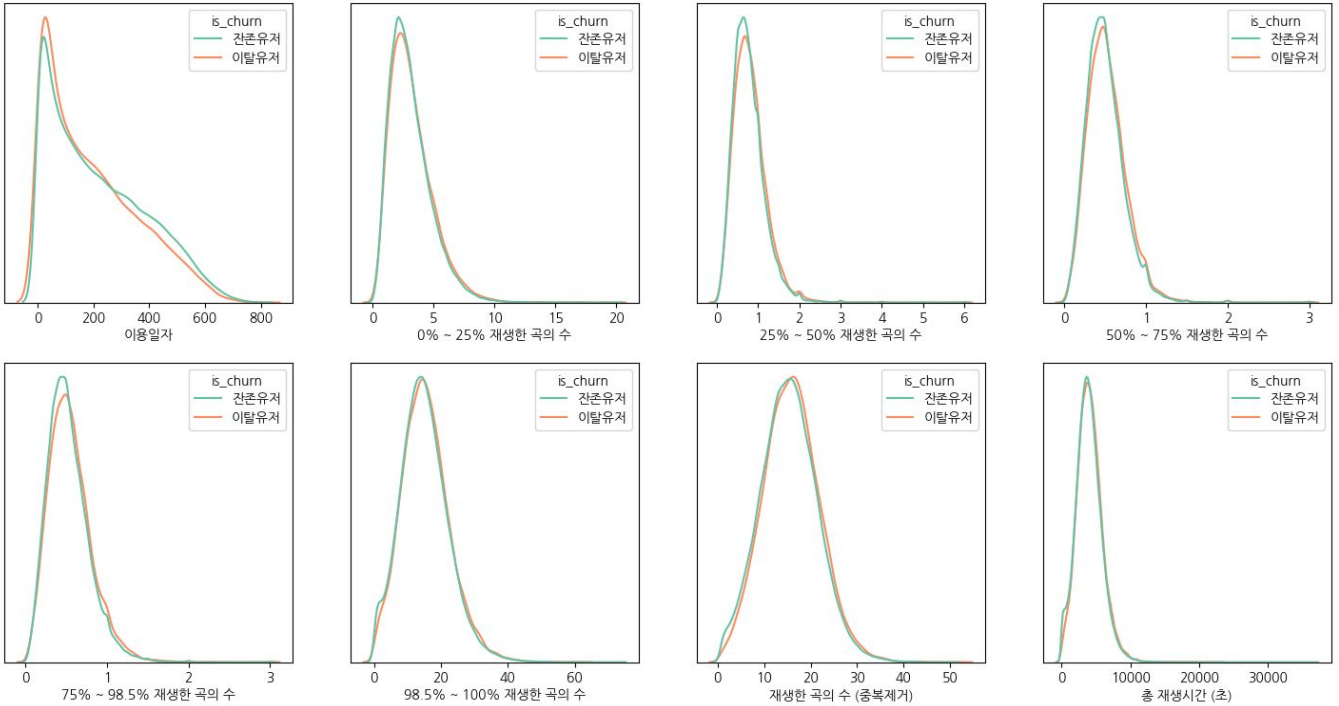
각 재생 횟수 < 0 삭제

재생 횟수

IQR * 1.5 외 삭제

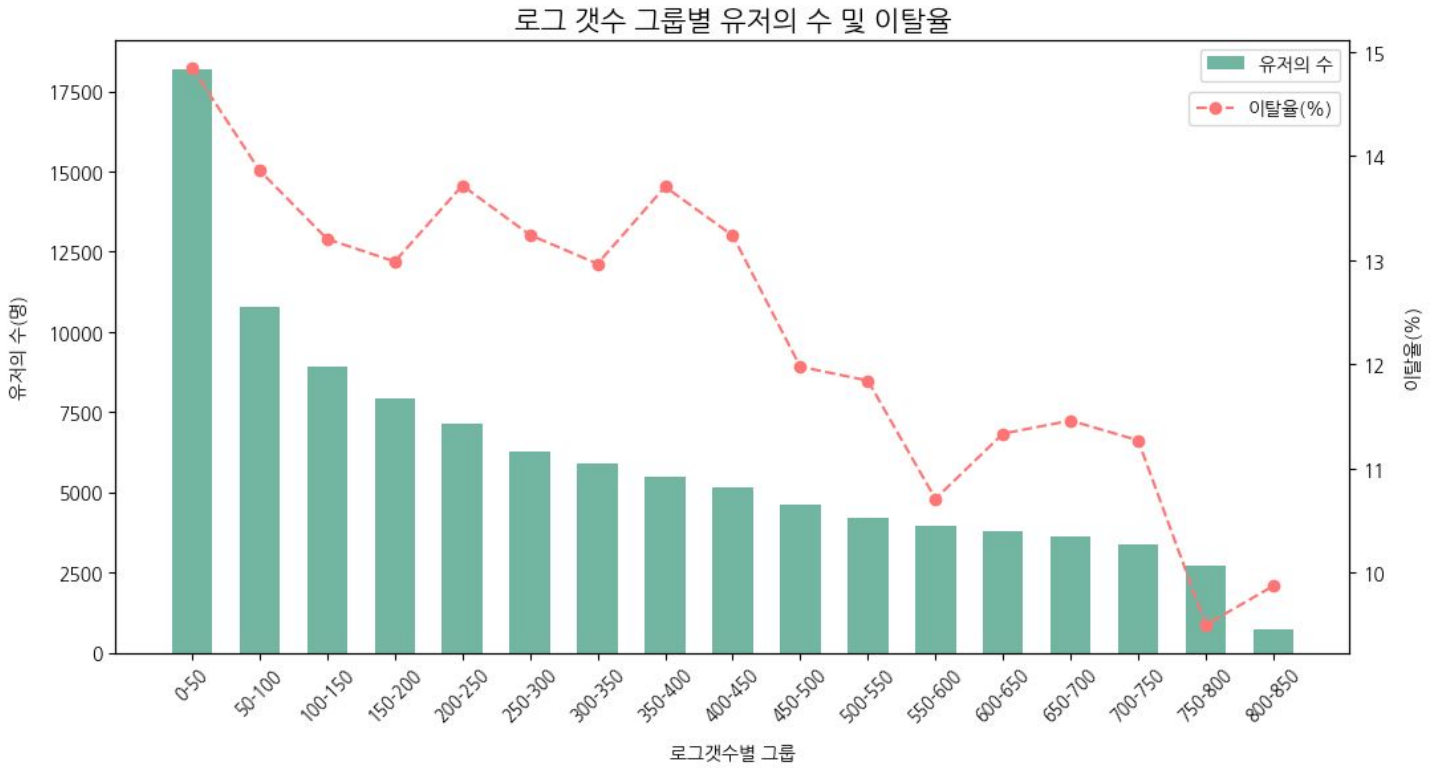
3 EDA - User_logs Dataset

이탈유저/잔존유저의 로그 데이터 분포



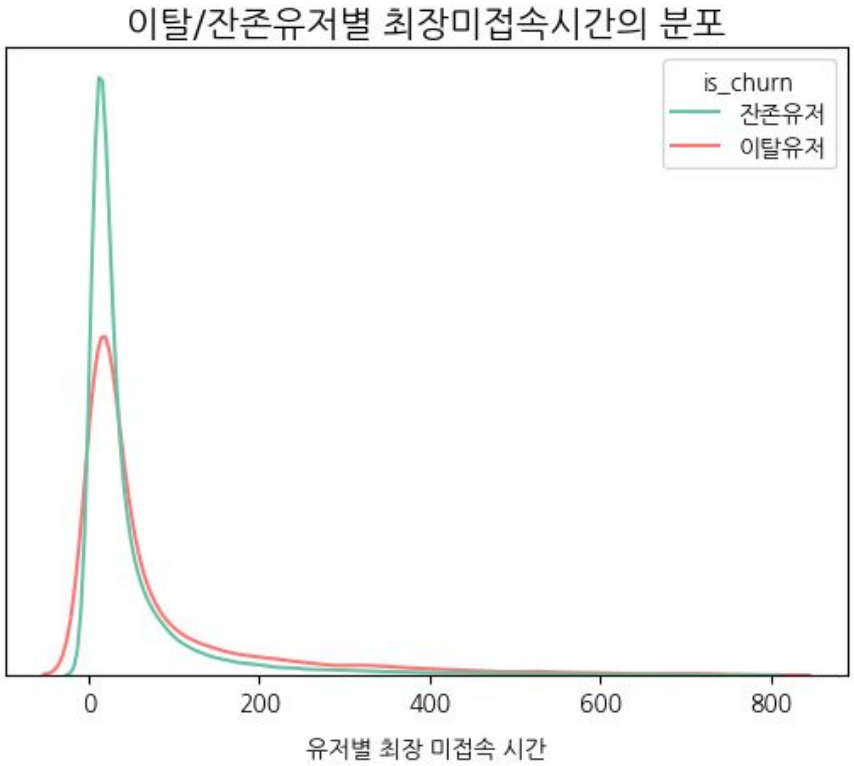
“ 이탈 유저와 잔존 유저의 로그 데이터 분포에 큰 차이가 없음을 확인.”

3 EDA - User_logs Dataset



“사용로그 개수가 적은 유저와 많은 유저간의 이탈률의 차이를 확인할 수 있음.”

3 EDA - User_logs Dataset



잔존 유저 중 30일 이상 미접속 경험

“37%”

이탈 유저 중 30일 이상 미접속 경험

“46%”

“ 이탈 유저의 장기간 미접속 경험 비율이 높음.”

3 EDA - Transaction Dataset : 삭제대상

ID	거래일	구독 만료일
1	2015-07-24	2015-08-03
1	2015-07-24	2015-08-23

ID, 거래일이 중복된 데이터

마지막 데이터 외 삭제

ID	거래일	구독 만료일
2	2015-07-24	2015-07-03
3	2015-07-24	2015-08-23

거래일보다 구독 만료일이 낮은 경우

삭제

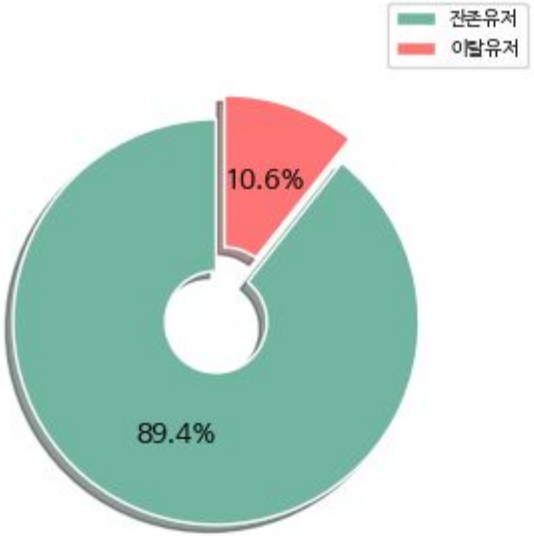
ID	거래일	구독 만료일
4	2015-07-24	2015-08-03
5	2016-12-24	2018-09-23

구독 만료일이 이상치로 보이는 경우

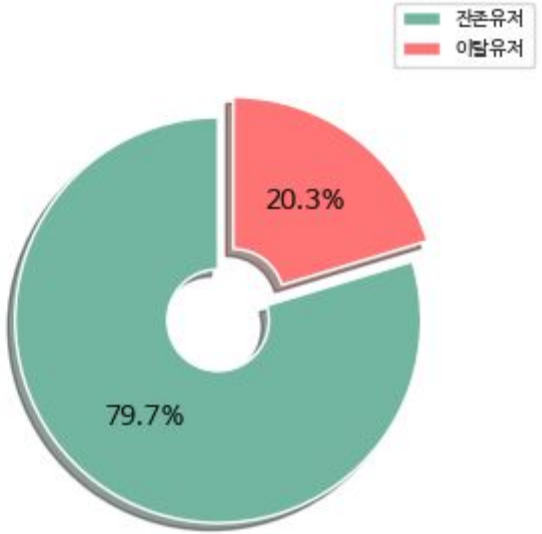
삭제

3 EDA - Transaction Dataset

서비스 이용 취소 경력이 없는 경우의 이탈유저 분포

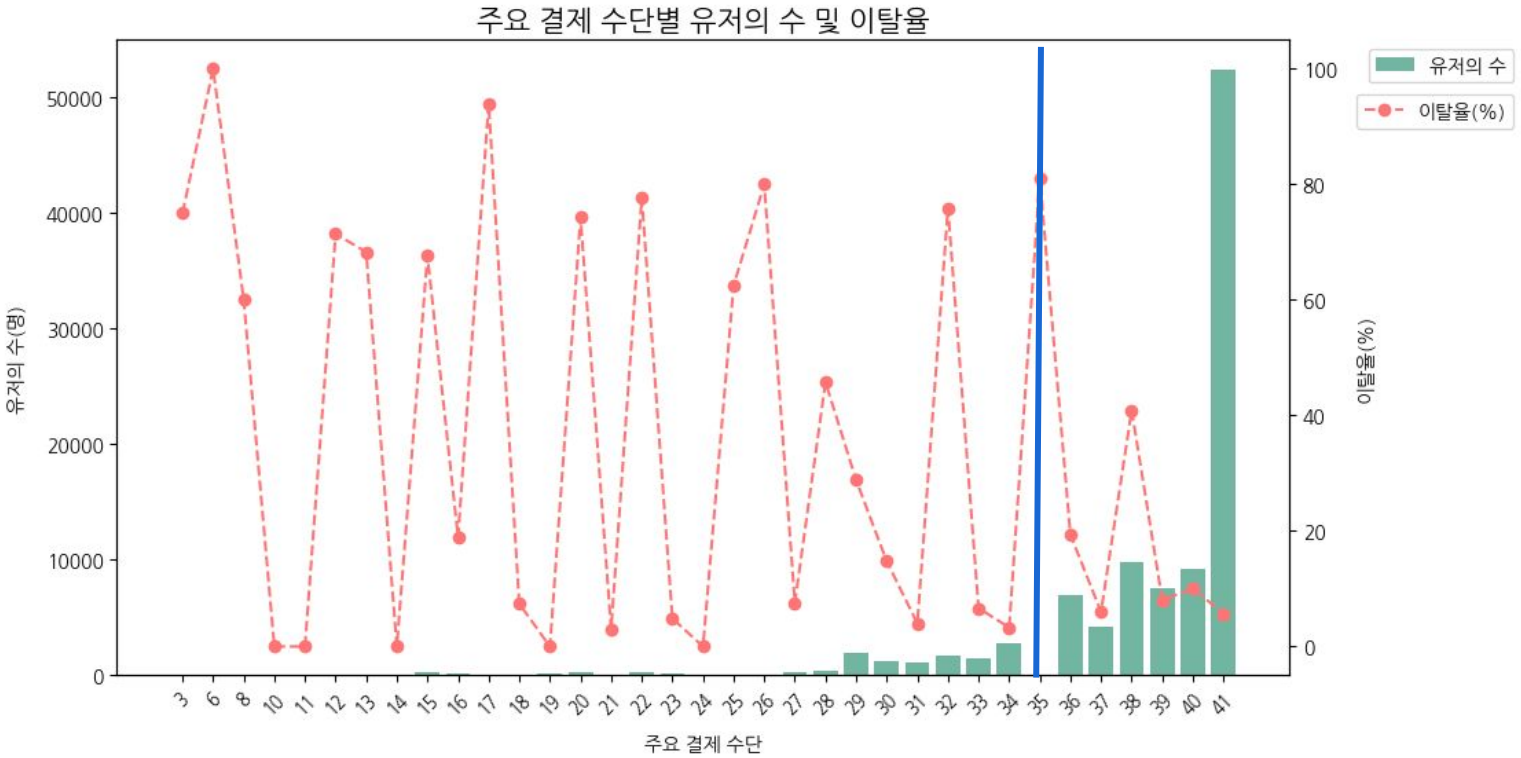


서비스 이용 취소 경력이 있는 경우의 이탈유저 분포



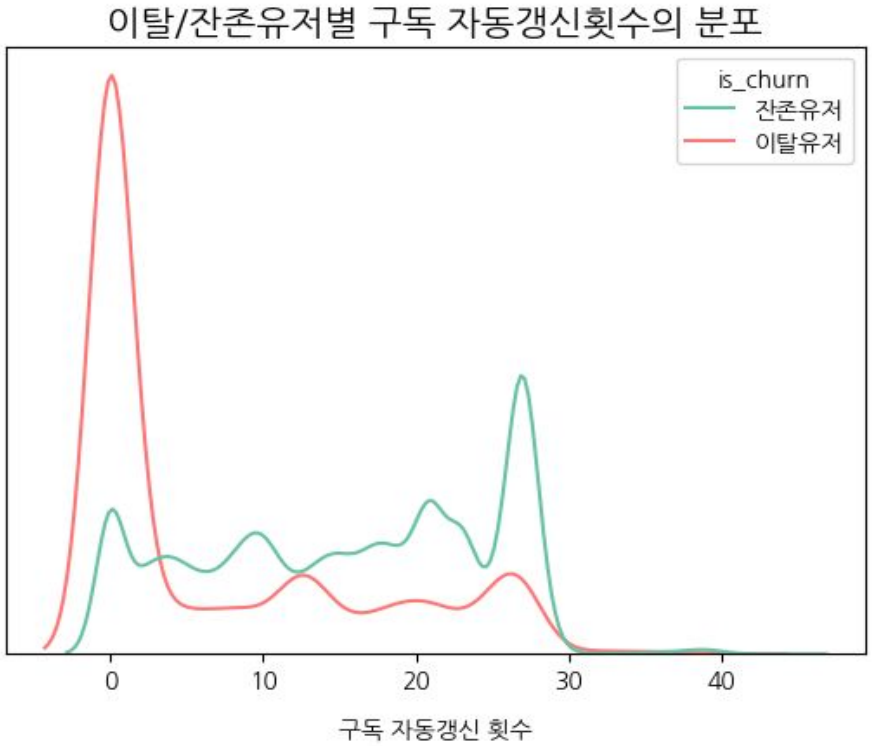
“ 서비스 이용 중 구독을 취소해본 이력이 있는 사용자의 이탈률이 높은 것을 확인 ”

3 EDA - Transaction Dataset



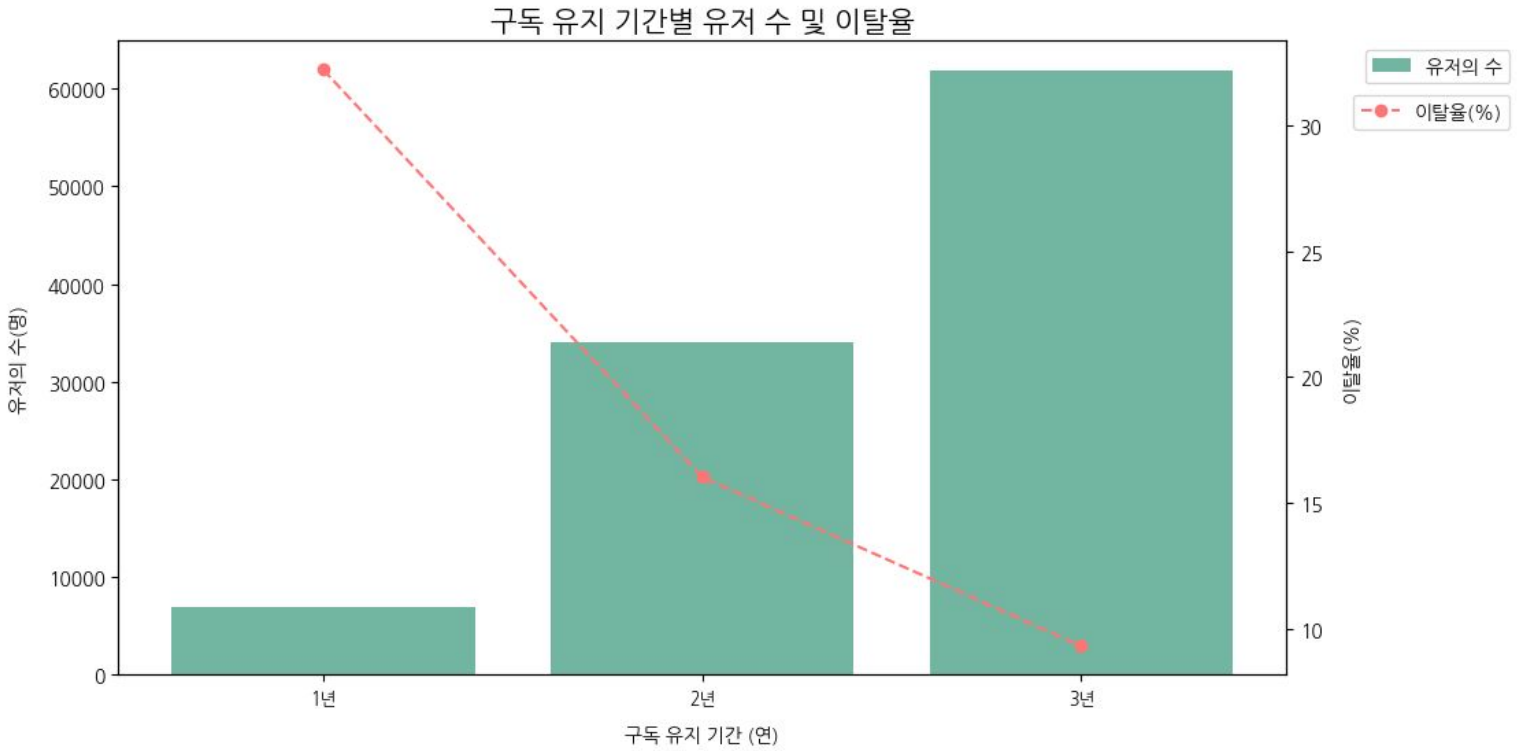
" 결제 수단의 표본이 적은 경우 이탈률에 큰 차이를 보이며, 38번 결제수단에서 높은 이탈률을 보임 "

3 EDA - Transaction Dataset



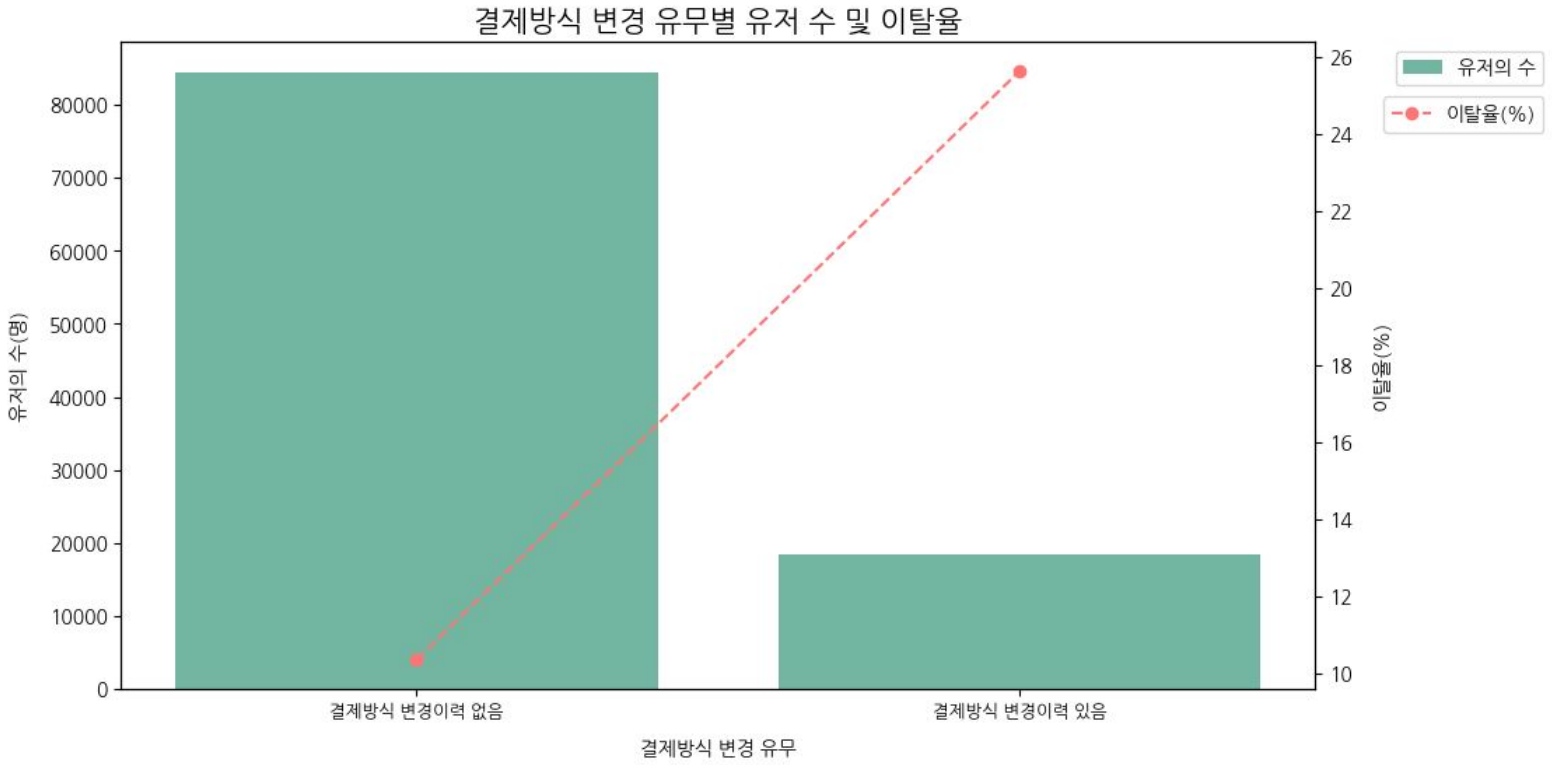
“ 구독의 자동 갱신 여부에 따라 이탈 유저와 잔존 유저의 분포에 차이가 있음을 확인 ”

3 EDA - Transaction Dataset



“구독기간 1년 미만인 유저의 이탈률이 높은 것을 확인”

3 EDA - Transaction Dataset



“ 결제 방식을 변경한 경험이 있는 유저의 이탈률이 높은 것을 확인.”

4 모델링

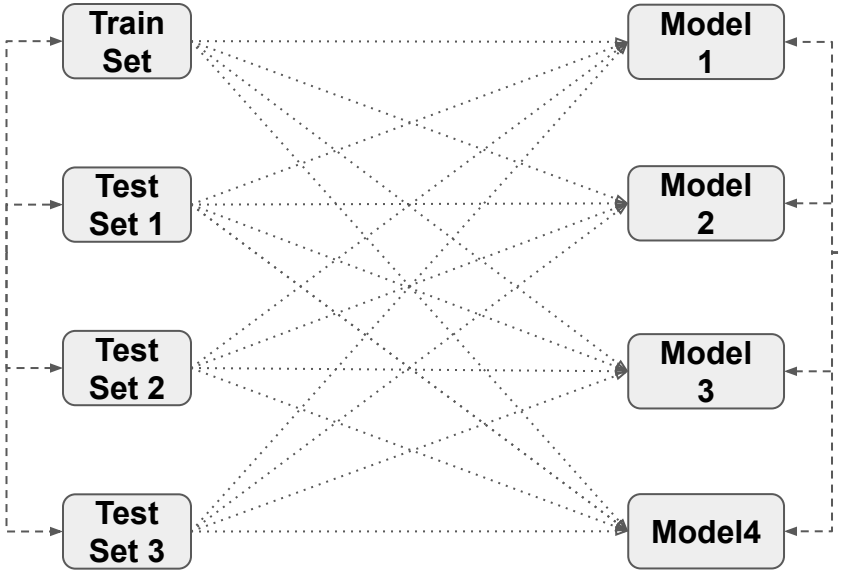
Final_Modeling.ipynb

Class 호출 및 데이터 학습

training_module.py

데이터 생성 및
전처리 클래스

- Data Load
- Preprocessing
- Feature Processing



modeling_module.py

데이터 학습 클래스

- Scaler
- Oversampling
- Modeling

4 모델링

Features

유저 정보	gender	성별
	bd	나이
	city	서비스 이용 도시
	register_init	최초 가입일

유저 거래	member_duration	마지막 멤버쉽 만료일 - 최초 멤버쉽 가입일
	Is_cancel	구독 취소 경험 1회 이상
	Is_auto_renew	자동 구독 여부
	discount	할인 기록 (멤버쉽 플랜 가격 != 실제 지불 가격)
	after_regit_to_buy	최초 멤버쉽 가입일 - 계정 생성일

유저 로그	per_25	25% 이하 감상한 노래의 비율
	per_25_75	25~75% 감상한 노래의 비율
	per_100	75% 이상 감상한 노래의 비율
	seconds_per_song	총 재생 시간 / 사용자가 재생한 음악의 수
	mean_seconds_mon	월 평균 노래 재생 시간
	max_log_term	서비스 사용 로그 간격 중 가장 큰 값

4 모델링

Preprocessing

Scaler	Oversampling
O	O
O	X
X	O
X	X

Feature 가공

나이	
숫자 그대로 사용	연속형
	범주형(10세 단위)
	범주형(5세 단위)
	범주형(25세<)
15~64세 외에는 Null 처리	연속형
	범주형(10세 단위)
	범주형(5세 단위)
	범주형(25세 <)

성별
Drop
그대로 사용

최초 가입일
연도로 구분
연월로 구분

delay
연속형
범주형

city
1 or else
범주형(전체)

Model

모델 선택
RandomForest
Catboost
LightGBM
Logistic Regression
Naive Bayes

5 모델링 결과

Precision , Recall 비교

구분	Model	Train Precision	Test Precision	Train Recall	Test Recall
Scaler X & Over-Sampling X	**LGBM**	**0.77**	**0.73**	**0.41**	**0.39**
	Catboost	**0.83**	**0.74**	**0.47**	**0.41**
Scaler O & Over-Sampling X	LGBM	0.77	0.70	0.41	0.39
	Catboost	0.83	0.70	0.47	0.42
Scaler X & Over-Sampling O	LGBM	0.91	0.49	0.89	0.56
	Catboost	0.93	0.52	0.91	0.54
Scaler O & Over-Sampling O	LGBM	0.94	0.60	0.91	0.52
	Catboost	0.97	0.68	0.92	0.46

Scaler

결과에 유의미한 차이 없음

Oversampling

precision 저하 / 과적합 심화

5 모델링 결과 - 변수선택, 변수추출

Model	k	Train Precision	Test Precision	Train Recall	Test Recall
LGBM	14	1.00	0.71	1.00	0.41
Catboost	14	0.85	0.72	0.52	0.44
LGBM	5	1.00	0.58	1.00	0.36
Catboost	5	0.78	0.67	0.39	0.35

selectKBest

성능 개선 / 과적합 개선

성과 없음

Model	k	Train Precision	Test Precision	Train Recall	Test Recall
LGBM	7	0.69	0.61	0.36	0.32
Catboost	7	0.79	0.61	0.41	0.31
LGBM	10	0.72	0.62	0.38	0.35
Catboost	10	0.89	0.69	0.44	0.34

PCA
(k=7 / 85%, k=10 / 96%)

성능 개선 / 과적합 개선

성과 없음

5 모델링 결과 - 데이터 추가 (220000개)

Model	Train Precision	Test Precision	Train Recall	Test Recall
LGBM	0.74	0.72	0.38	0.37
Catboost	0.78	0.74	0.44	0.44

데이터 추가(220000개)

과적합 개선

성능 성과 없음

5 모델링 결과 - 변수선택

로그 데이터만 사용하여 모델링

Model	Train Precision	Test Precision	Train Recall	Test Recall
LGBM	0.60	0.58	0.22	0.21
Catboost	0.62	0.56	0.26	0.22

고객 정보, 거래 데이터만 사용하여 모델링

Model	Train Precision	Test Precision	Train Recall	Test Recall
LGBM	0.63	0.59	0.24	0.22
Catboost	0.62	0.53	0.26	0.22

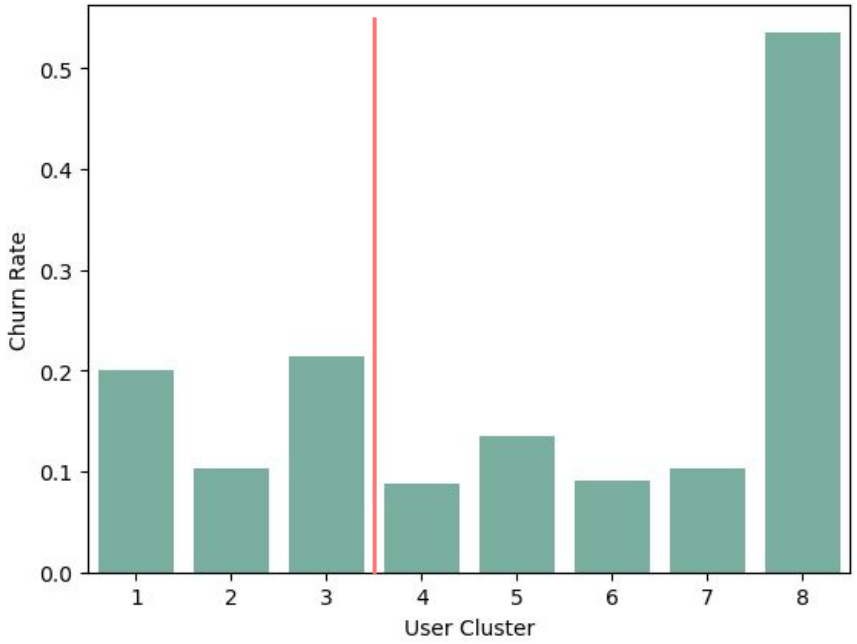
Feature를 나누어 테스트 진행

성능 개선 성과 없음

5 모델링 결과

K-means Clustering

Cluster	이탈률	최단 구독 개월	최장 구독 개월
1	0.20	0.03	5.83
2	0.10	5.56	10.63
3	0.21	10.4	14.87
4	0.08	14.77	18.87
5	0.13	18.76	24.00
6	0.09	22.43	31.50
7	0.10	23.83	31.23
8	0.54	30.93	40.43



5 모델링 결과

Model	Train Precision	Test Precision	Train Recall	Test Recall
LGBM	0.73	0.59	0.23	0.18
Catboost	0.73	0.62	0.23	0.20

User segmentation

성능 개선 성과 없음

6 결론

활용방안

EDA를 통해 이탈유저와 비이탈 유저간에 유의미한 차이가 보이는 컬럼을 확인.

→ 이를 활용한다면 서비스 개선에 도움이 될 듯

클러스터링을 통해 유저 그룹을 구분하여 확인한 결과 멤버쉽 유지 기간에 따른 이탈률에 차이를 확인.

한계점

모델의 성능 낮아 머신러닝 모델로서의 가치가 떨어지는 측면이 있음.

→ 성별, 나이와 같은 개인정보 데이터는 유저가 직접 입력하는 데이터로 신뢰하기 어려웠다.

생각보다 유저의 서비스 이용 패턴(유저 로그)에 유의미한 차이를 보이는 컬럼이 적었다.