Machine Learning Project

은행 데이터를 이용한
이탈 원인 및 고객 이탈 예측

평일 오후 3조
김승민 박영주 한상현 한정연

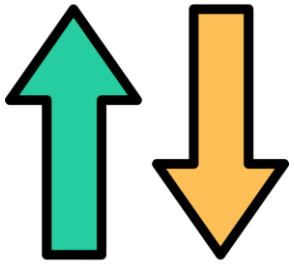
PROJECT INDEX

01

프로젝트 소개

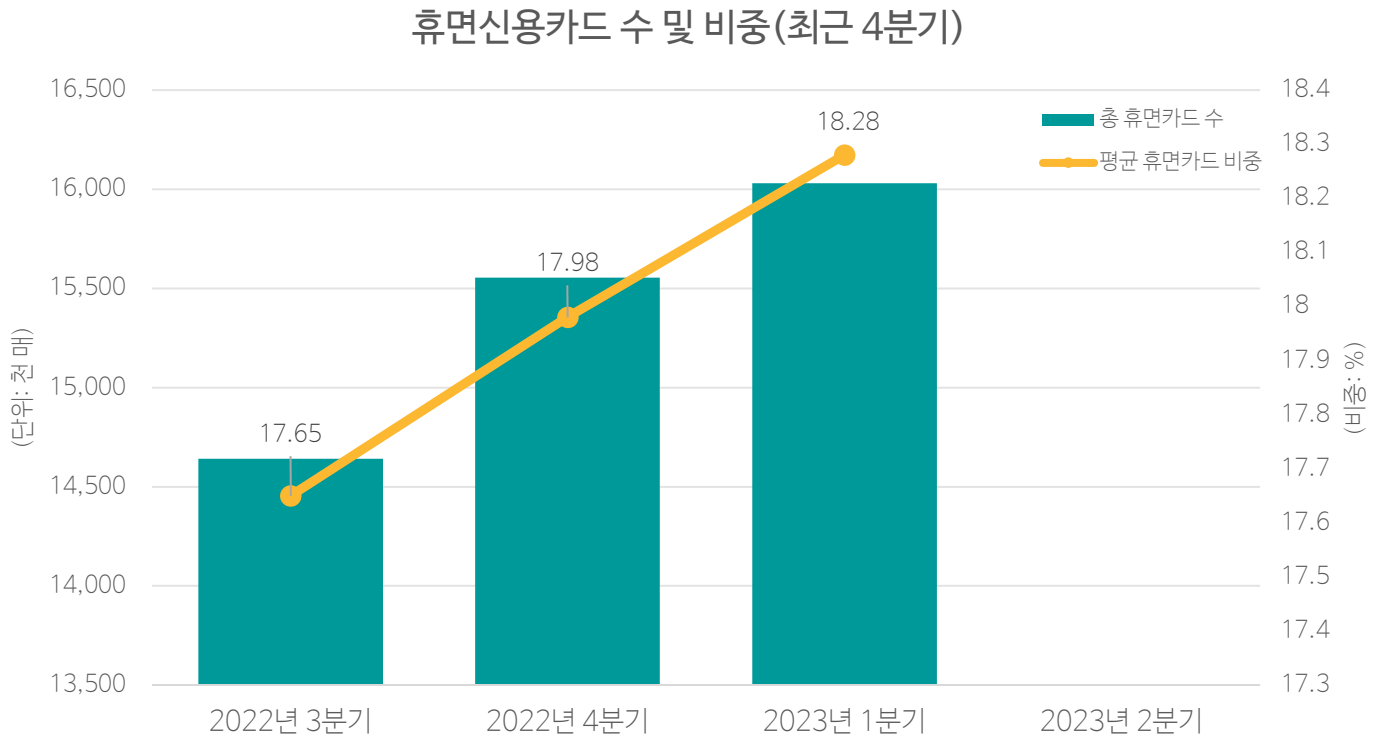
카카오, 네이버, 토스 등 금융 사업에 뛰어드는 플랫폼 증가

은행, 카드사의 신규 고객 유치 경쟁 증가



기존 카드 이용률과 거래 횟수 감소

휴면 카드 고객 증가



출처: 여신금융협회

"1000만장을 깨워라"... 카드사 리텐션 경쟁

휴면카드 1037만장... 8%↑
10장 중 2장은 장롱에
페이 늘자 더 늘어

2금융 > 카드

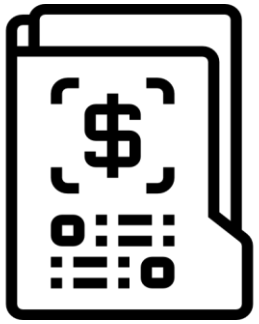
우리카드, 리텐션 마케팅·비용절감 효과 톡톡

신규 고객을 유치하는 비용이 기존 고객 유치 비용보다 **높다.**
국내 은행 중 신규 고객 유치보다 **리텐션 마케팅**에 집중하여 영업이익을 낸 사례가 있다.

이를 바탕으로 기존/휴면 고객의 이탈을 방지할 수 있도록 **이탈 예측 모델을 구축**해보려고 한다.

02

데이터 분석 및 전처리



은행 고객 정보와 이탈 여부를 포함한 데이터

10,127 rows * 23 columns

Numeric Variables (14):

Customer_Age, Dependent_count, Months_on_book, Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon, Credit_Limit, Total_Revolving_Bal, Avg_Open_To_Buy, Avg_Utilization_Ratio, Total_Trans_Amt, Total_Trans_Ct, Total_Amt_Chng_Q4_Q1, Total_Ct_Chng_Q4_Q1

Text Variables (6):

Attrition_Flag, Gender, Education_Level, Marital_Status, Income_Category, Card_Category

Other (3):

CLIENTNUM, Naive_Bayes_Classifier_....1, Naive_Bayes_Classifier_...2

출처: 캐글(Kaggle)
https://www.kaggle.com/datasets/thedevastator/predicting-credit-card-customer-attrition-with-m

02 데이터 분석 및 전처리

1. 데이터 탐색

1차 변수 선정
- 불필요한 컬럼 드롭
- 컬럼명/순서 변경

EDA
- 데이터 시각화
- 기본 정보 확인

2. 데이터 전처리

기초 전처리
- 결측치 없음
- 이상치 없음

모델링 준비
- 변수 인코딩
- Train/Test 분리
- 데이터 스케일링

3. 모델 학습

Feature Selection
- RFECV을 이용

파라미터 튜닝
- Grid Search CV
- 베이지안 최적화
- Over Sampling

4. 성과 평가

사용 알고리즘
- 로지스틱 회귀
- 랜덤 포레스트
- KNN
- XGBoost

모델 평가 지표 비교

5. 모델 학습

변수의 영향력 확인
- Feature Importance
- SHAP Value

6. 프로젝트 결론

분석 의의

분석 한계점

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status
0	768805383	Existing Customer	45	M	3	High School	Married
1	818770008	Existing Customer	49	F	5	Graduate	Single
2	713982108	Existing Customer	51	M	3	Graduate	Married
3	769911858	Existing Customer	40	F	4	High School	Unknown
4	709106358	Existing Customer	40	M	3	Uneducated	Married
...	...	...	...	...	...	...	...
10122	772366833	Existing Customer	50	M	2	Graduate	Single
10123	710638233	Attrited Customer	41	M	2	Unknown	Divorced
10124	716506083	Attrited Customer	44	F	1	High School	Married
10125	717406983	Attrited Customer	30	M	2	Graduate	Unknown
10126	714337233	Attrited Customer	43	F	2	Graduate	Married

10127 rows × 23 columns

1차 변수 드롭

데이터를 불러온 후 고객 식별 번호와
Naive_Bayes_Classifier 컬럼 두 개 제거

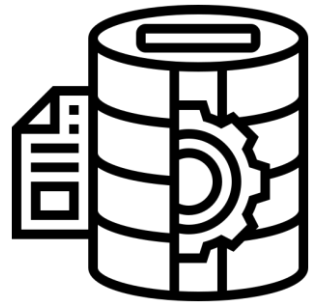
총 3개의 컬럼 제거 후 프로젝트 진행

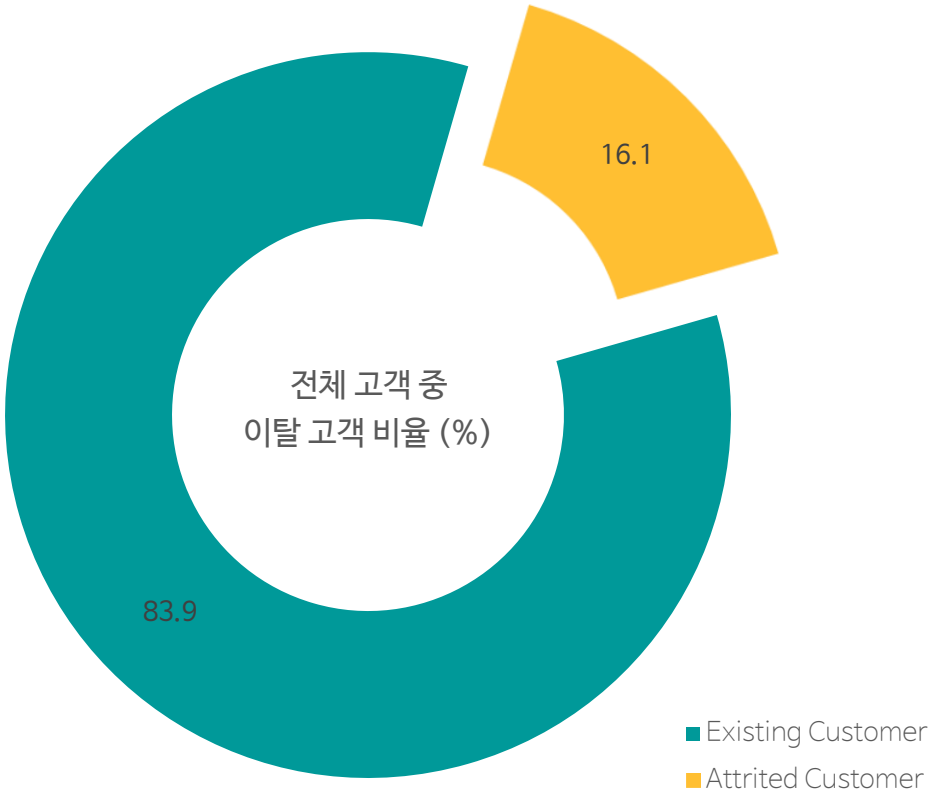
컬럼 이름 변경

```
df.rename(columns={
    'Attrition_Flag' : 'Exited',
    'Customer_Age' : 'Age',
    'Dependent_count' : 'Dependents',
    'Education_Level' : 'Education',
    'Marital_Status' : 'Marital',
    'Income_Category' : 'Income',
    'Card_Category' : 'Card_Type',
    'Months_on_book' : 'Tenure',
    'Total_Relationship_Count' : 'Product_Cnt',
    'Months_Inactive_12_mon' : 'Inactive_Months',
    'Contacts_Count_12_mon' : 'Contacts_Cnt',
    'Total_Revolving_Bal' : 'Revolv_Bal',
    'Avg_Open_To_Buy' : 'Avg_OTB',
    'Total_Amt_Chng_Q4_Q1' : 'Amt_Chng_Q4_Q1',
    'Total_Trans_Amt' : 'Trans_Amt',
    'Total_Trans_Ct' : 'Trans_Cnt',
    'Total_Ct_Chng_Q4_Q1' : 'Cnt_Chng_Q4_Q1',
    'Avg_Utilization_Ratio' : 'Avg_Util_Ratio'
}, inplace=True)
```

컬럼 순서 변경

```
df = df[['Exited', 'Age', 'Gender', 'Dependents',
         'Education', 'Marital', 'Income’,
         'Card_Type', 'Tenure', 'Product_Cnt',
         'Inactive_Months', 'Contacts_Cnt',
         'Credit_Limit','Revolv_Bal', 'Avg_OTB',
        'Avg_Util_Ratio', 'Trans_Amt', 'Trans_Cnt',
         'Amt_Chng_Q4_Q1', 'Cnt_Chng_Q4_Q1']]
```



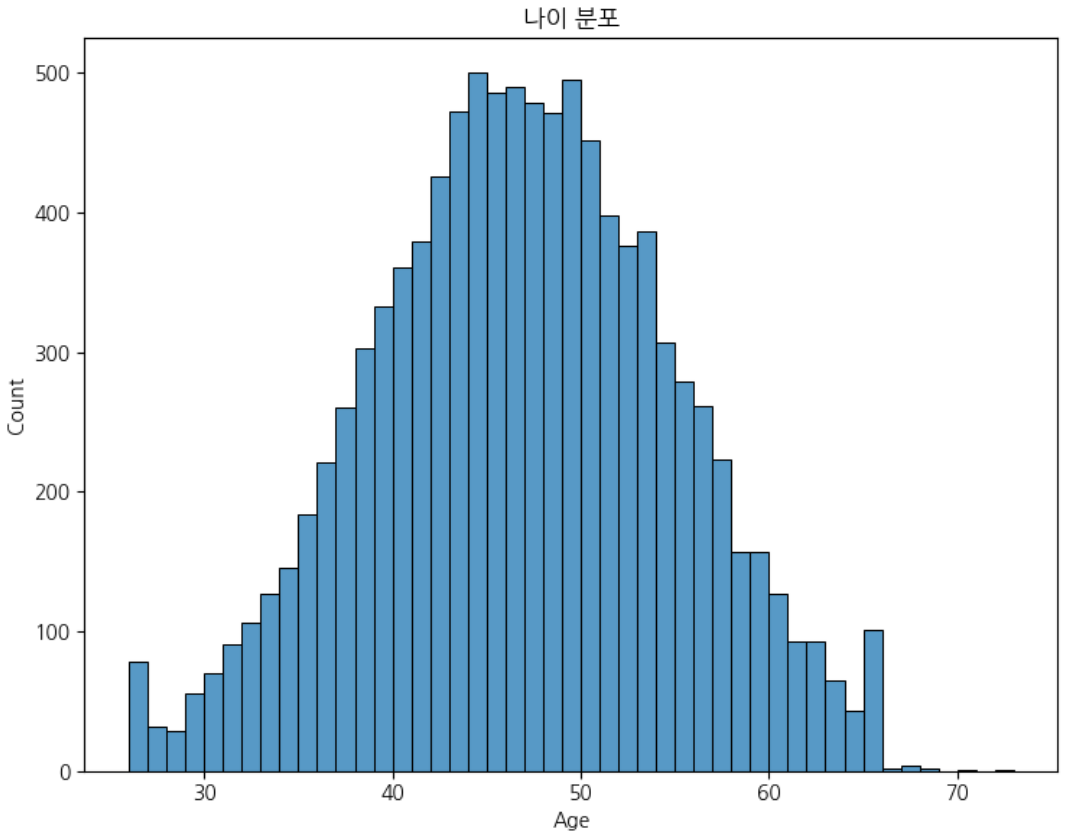


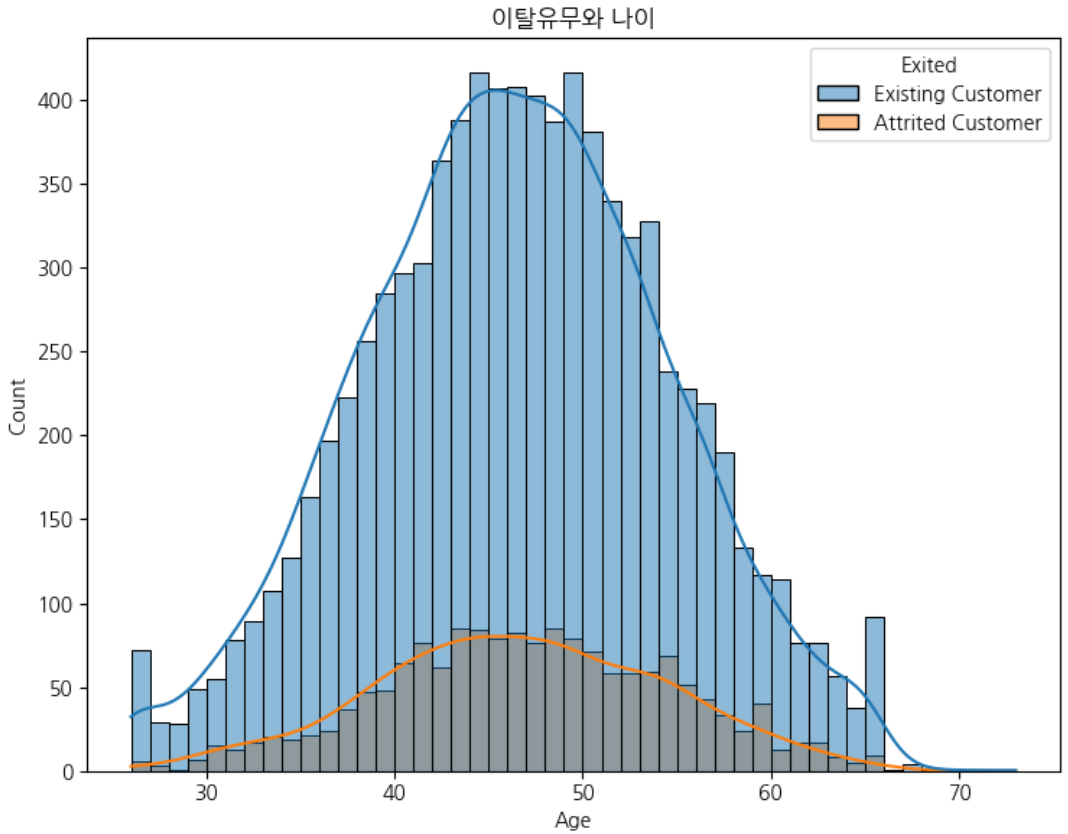
정확한 고객 이탈 유무를 분석하기 위해서는

Attired Customer(이탈 고객) 데이터를 분석하는 것이 중요

은행 고객들의 나이 분포와 그에 따른 이탈 유무

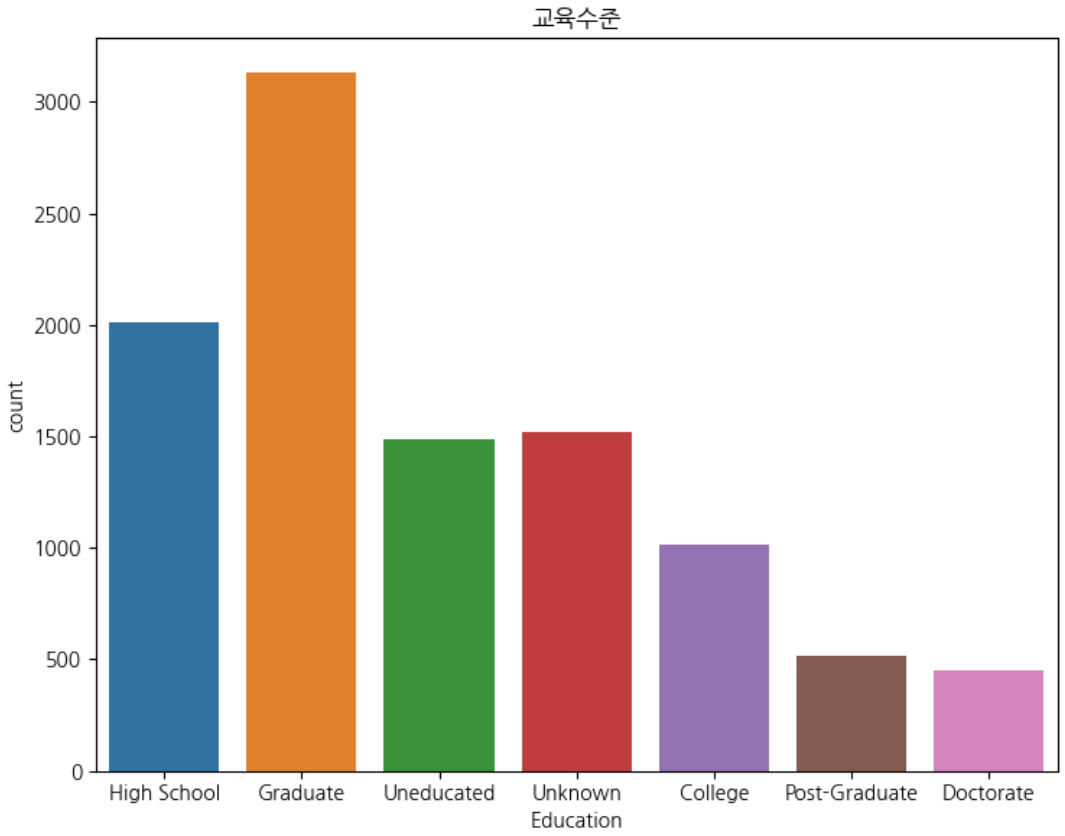
은행을 이용하는 주 연령층이 40대와 50대인 것을 알 수 있다

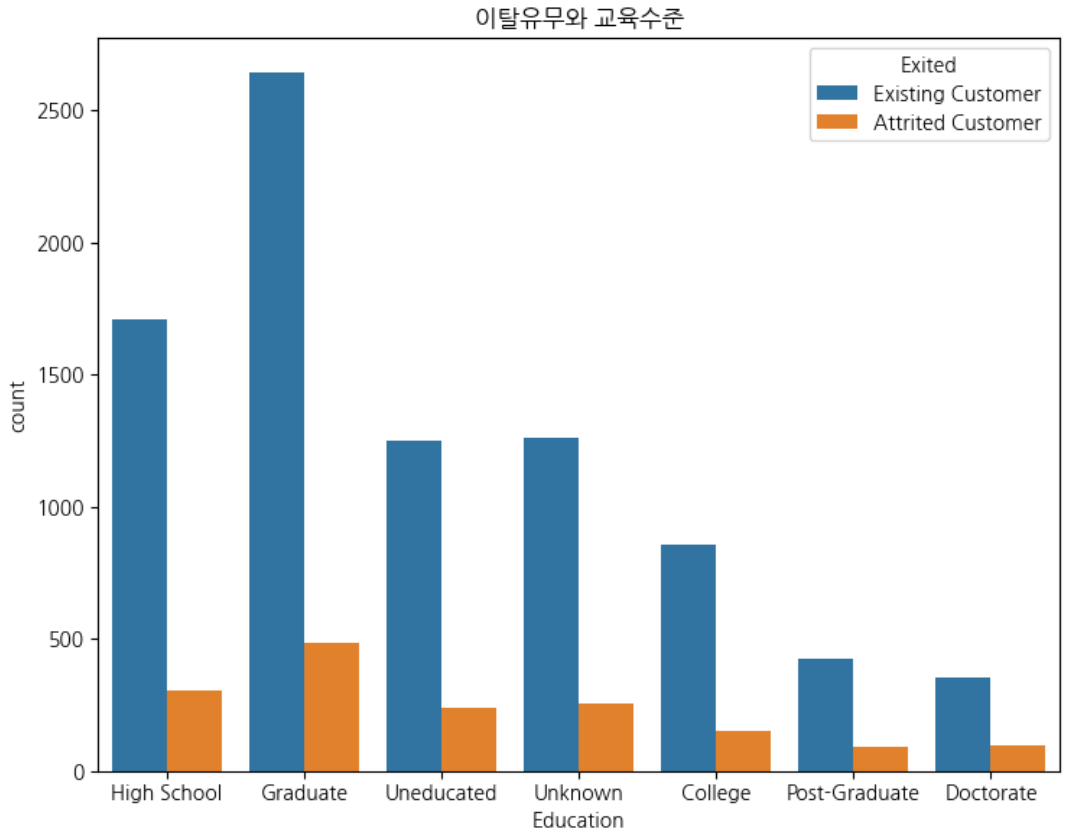




은행 고객들의 학력과 그에 따른 이탈 유무

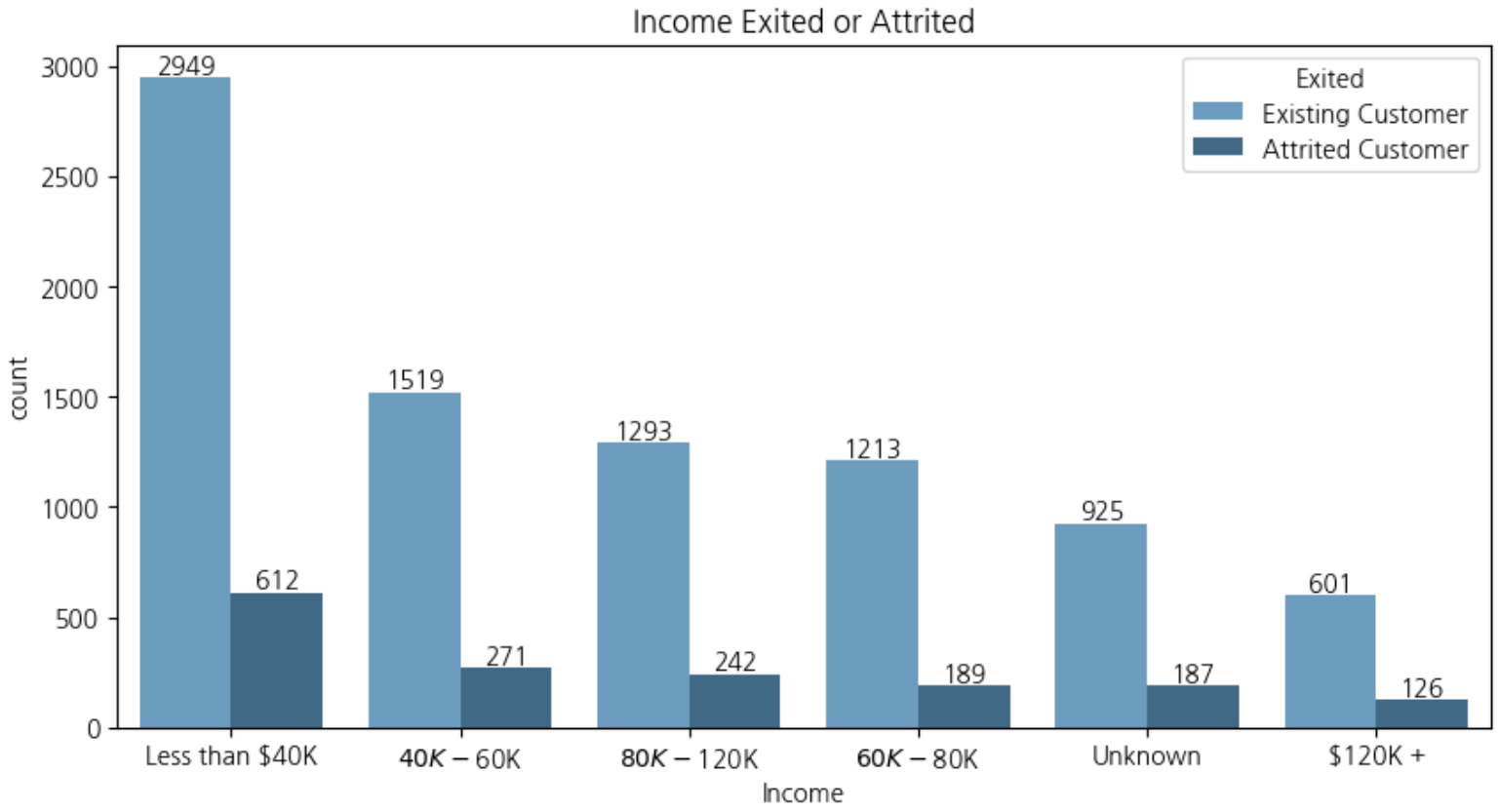
고객의 약 30%가 대학을 졸업한 것을 알 수 있으며 학력은 이탈률과 크게 관계가 없는 것으로 나타난다





은행 고객들의 수입(연봉)과 그에 따른 이탈 유무

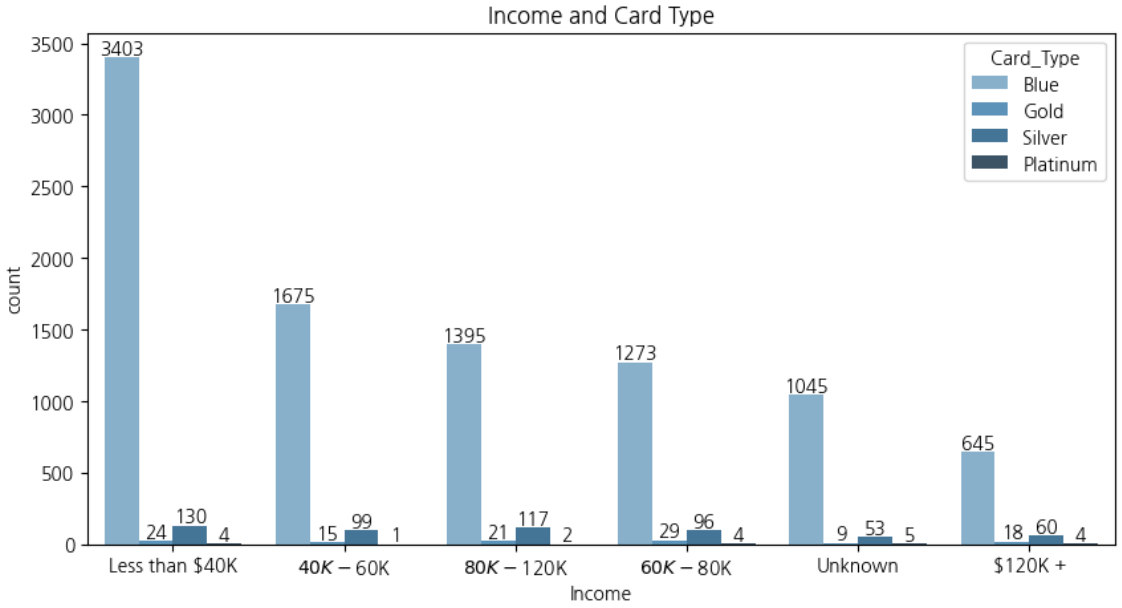
고객의 약 35% 정도의 연봉이 4만 달러 이하인 것을 알 수 있으며 연봉에 관계없이 각 항목마다 약 13~17%의 이탈률을 가지고 있다.

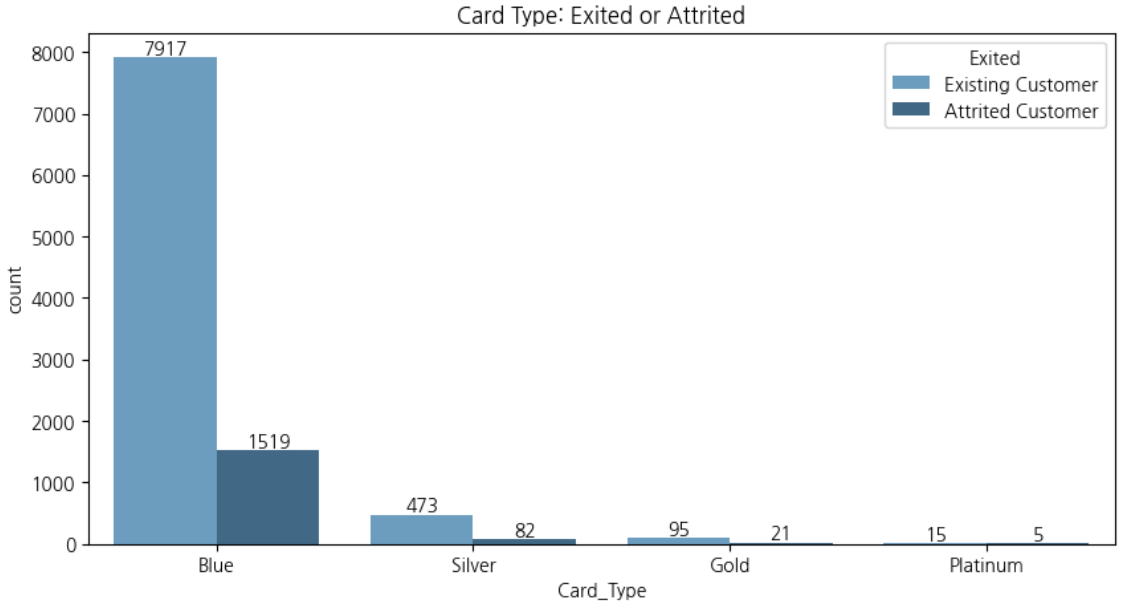


은행 고객들이 소지하고 있는 카드 종류와 그에 따른 이탈 유무

Blue 카드를 소지한 고객의 수가 가장 높다.

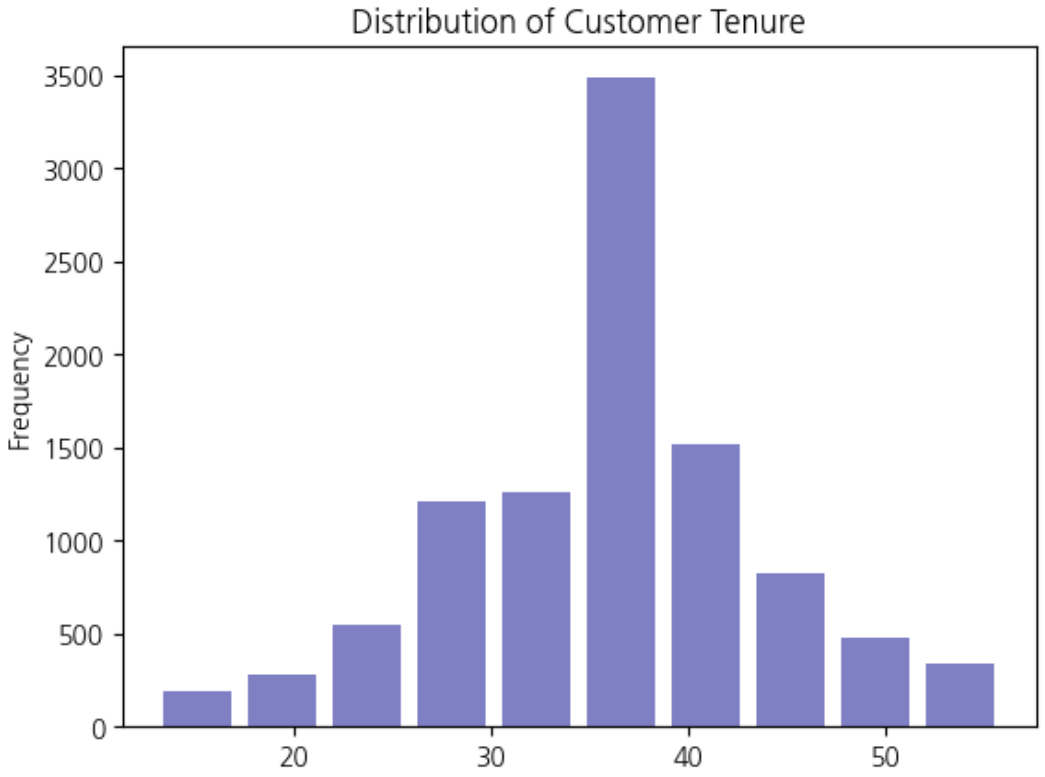
연봉과 카드 종류별 가입자의 상관 관계로 은행이 카드를 등급별로 발행하는 것을 알 수 있다.

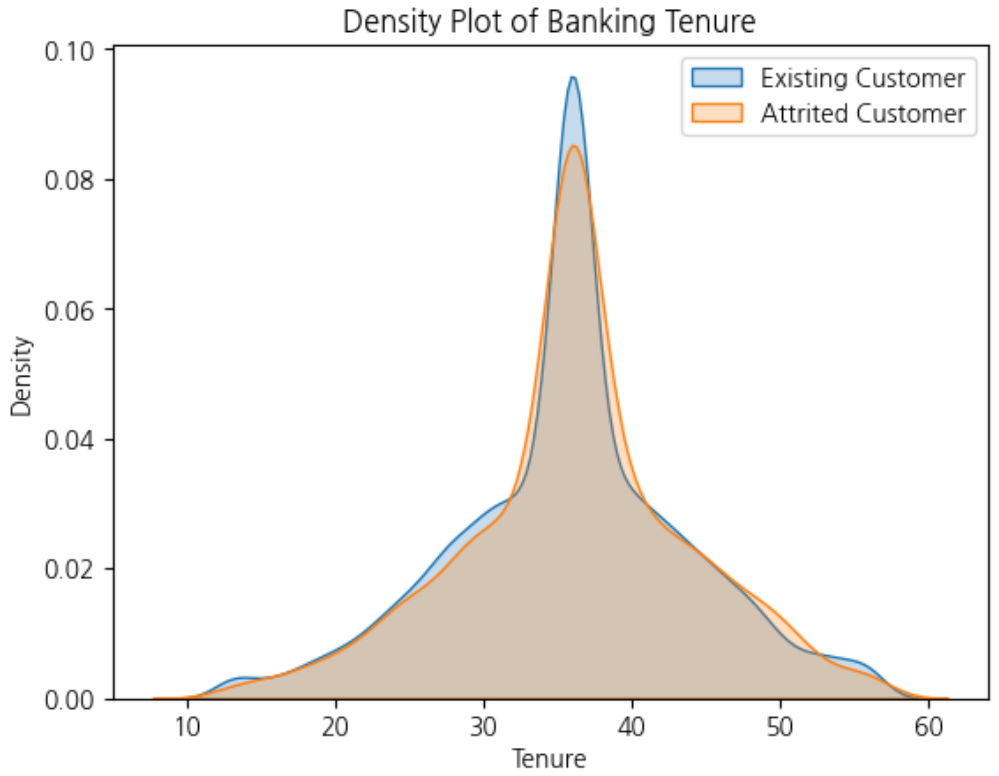




은행 고객들이 카드를 보유한 기간과 그에 따른 이탈 유무

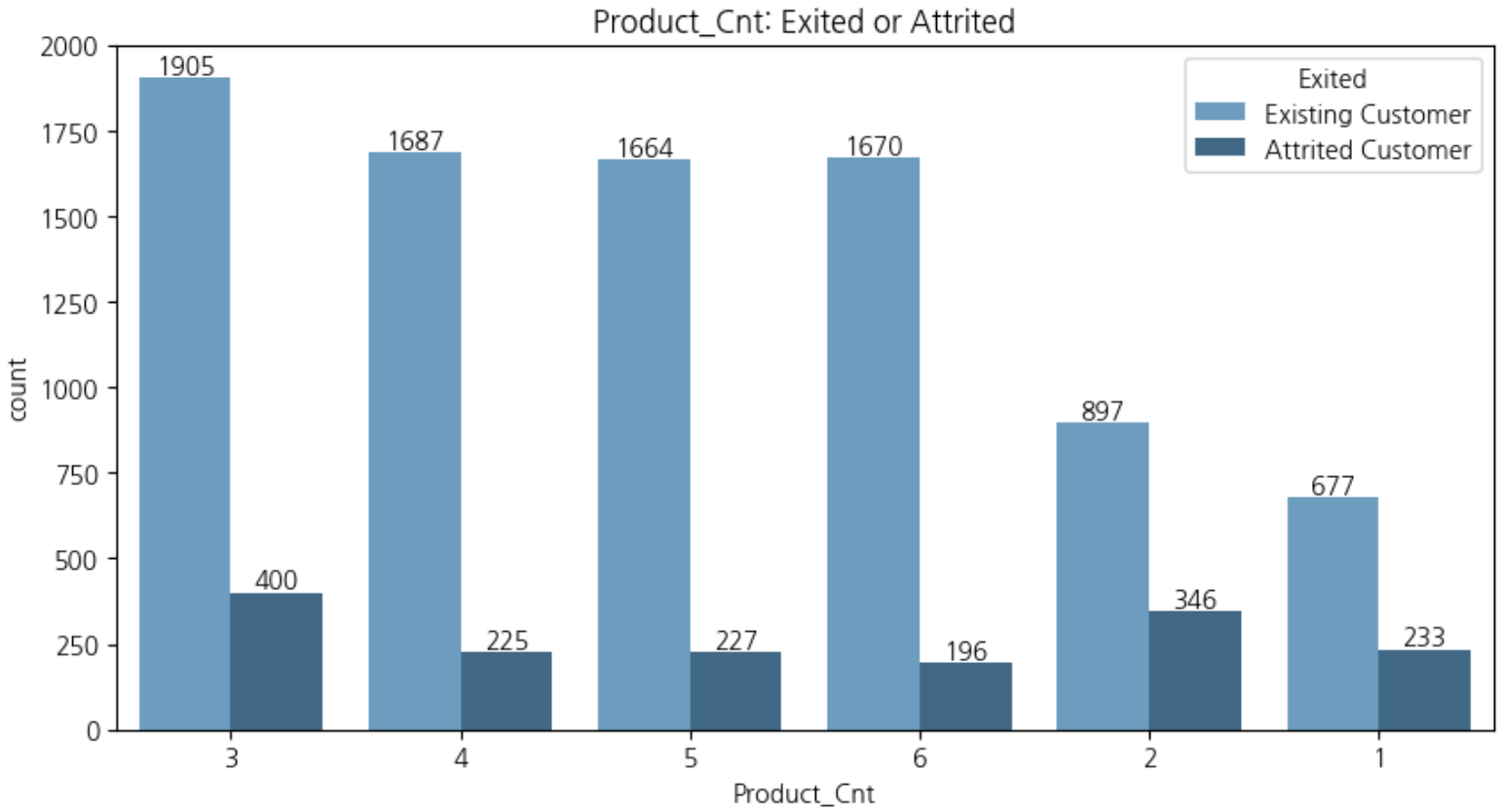
가입 기간이 36개월인 고객의 수가 가장 많으나 현재 고객과 이탈한 고객의 밀도를 비교했을 때, 가입기간과 이탈은 크게 관계가 없는 것을 확인할 수 있다.





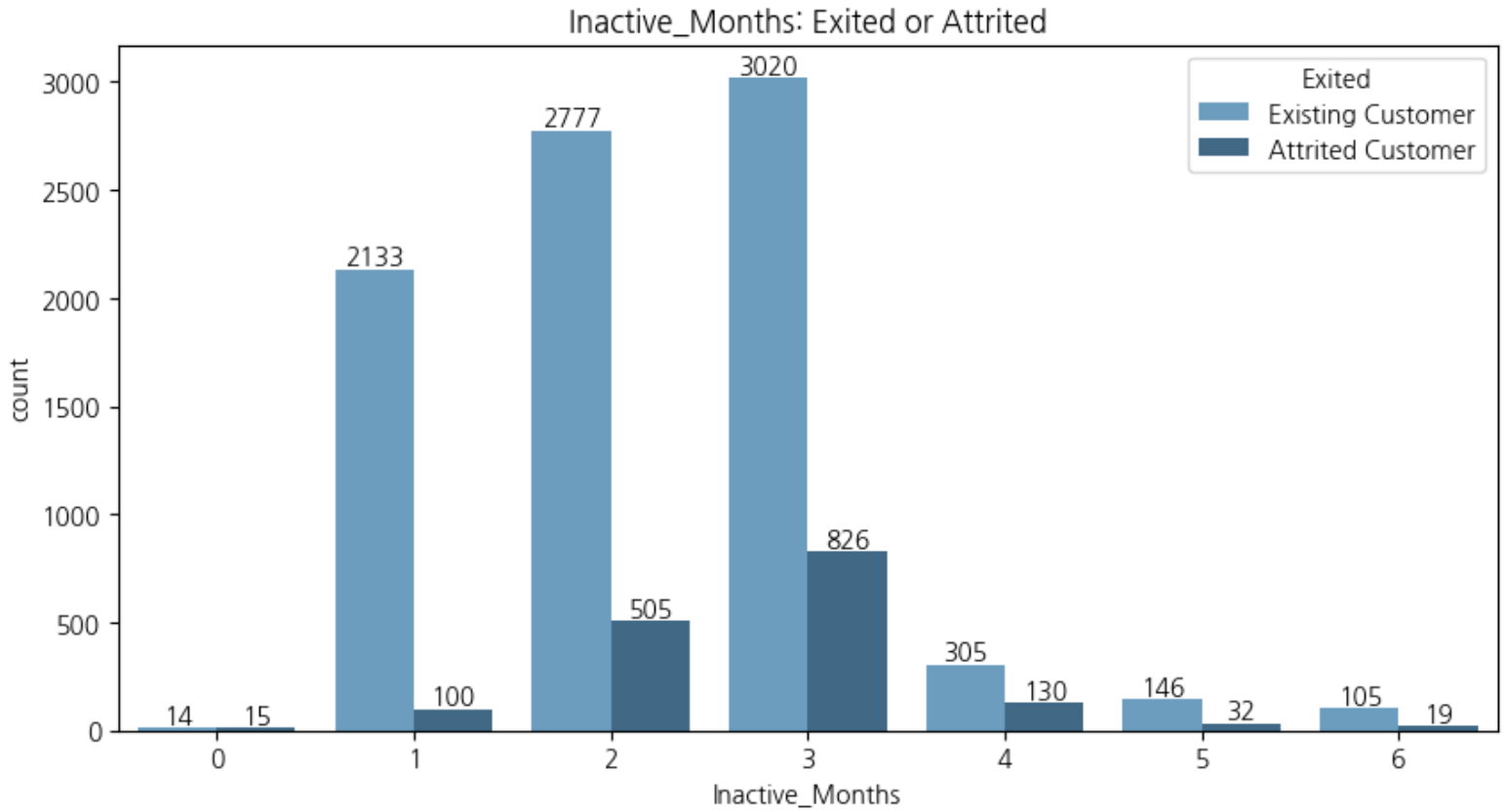
은행 고객들이 가입한 상품수와 그에 따른 이탈 유무

가입한 상품이 적은 고객들이 비교적 높은 이탈률을 보인다.



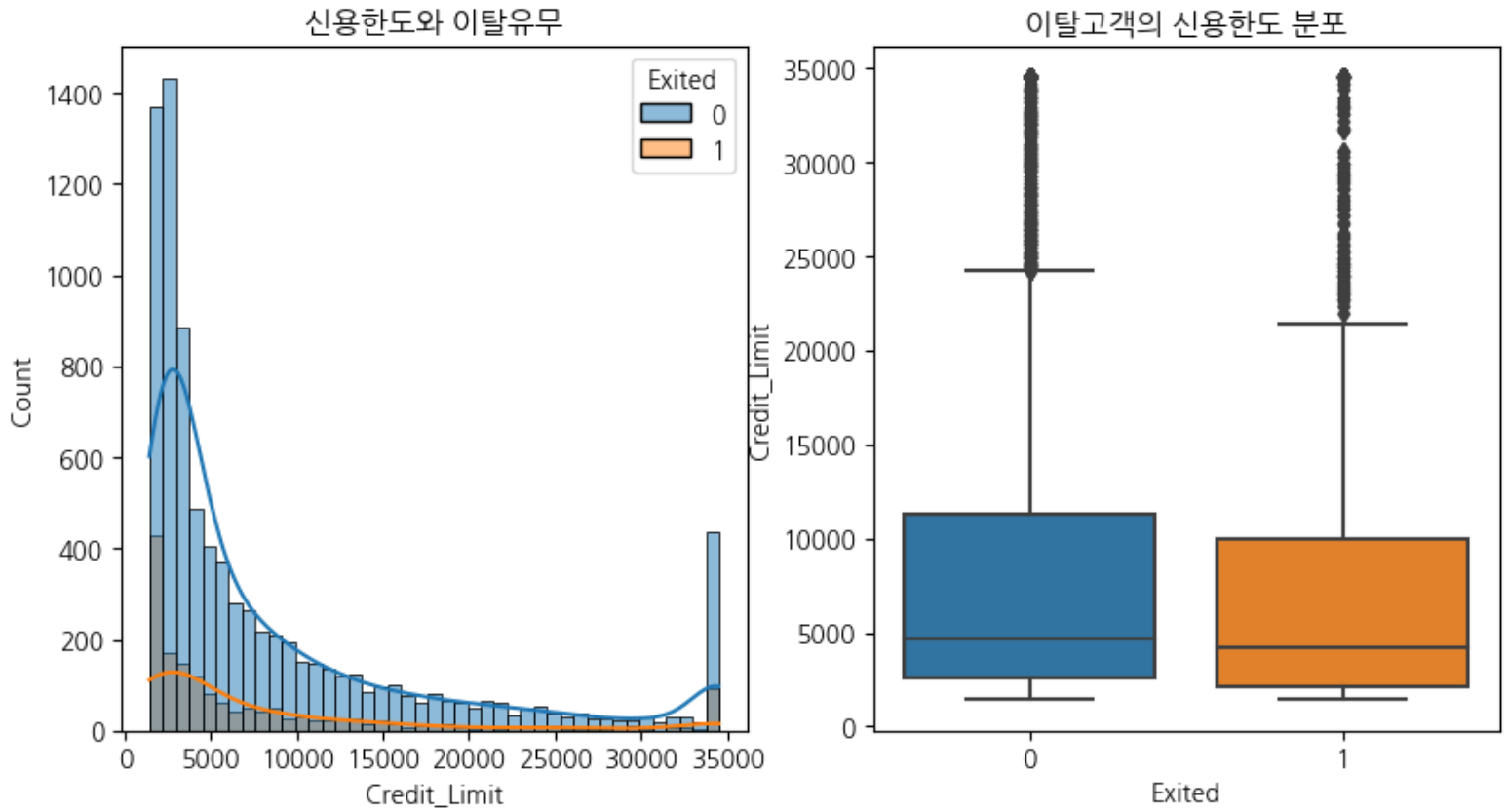
은행 고객들의 비활성화 기간과 그에 따른 이탈 유무

비활성화 기간이 길어질수록 이탈하는 고객이 많을 것으로 예상했으나 휴면 기간이 4개월인 고객의 이탈률이 가장 높은 것으로 나타났다.



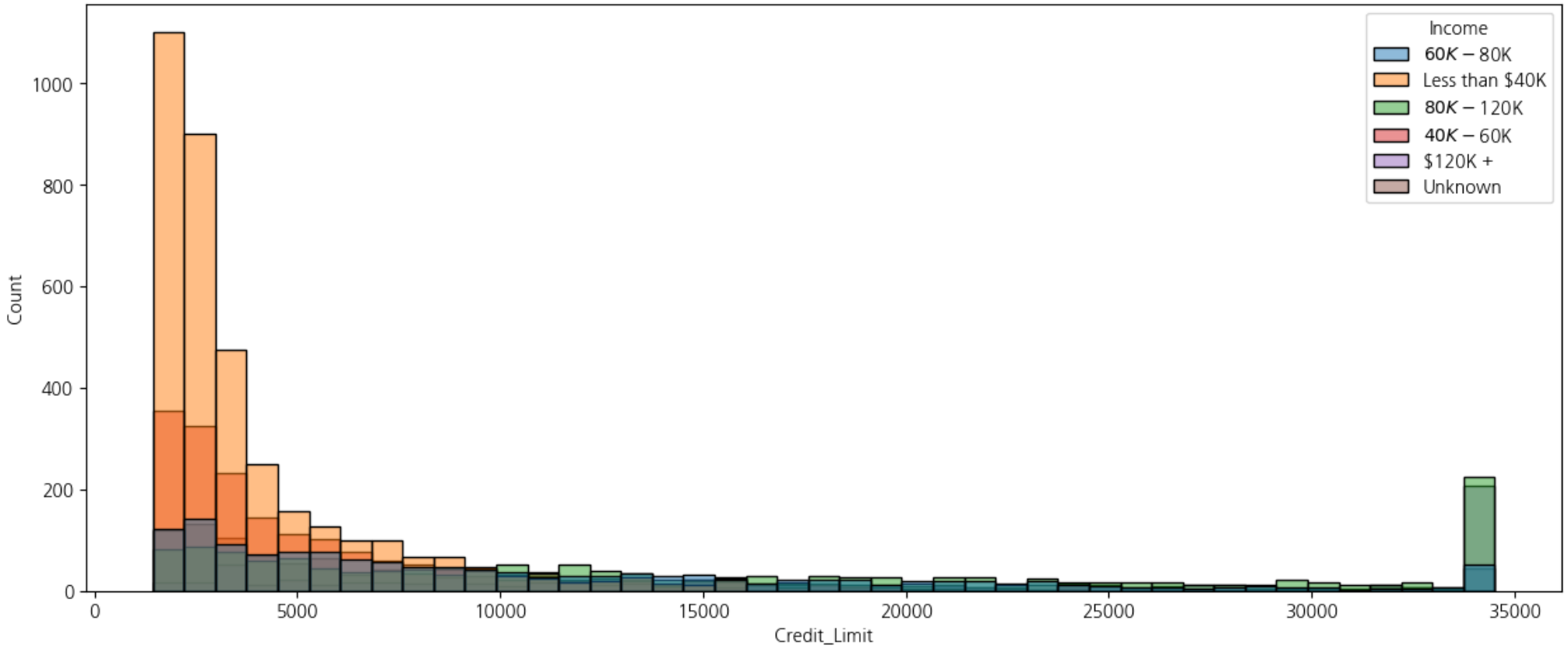
은행 고객들의 신용한도와 그에 따른 이탈 유무

신용한도가 낮은 고객의 수가 많은 것을 확인하였으며 이는 앞의 고객 연봉과도 상관 관계가 있는 것으로 보여진다.



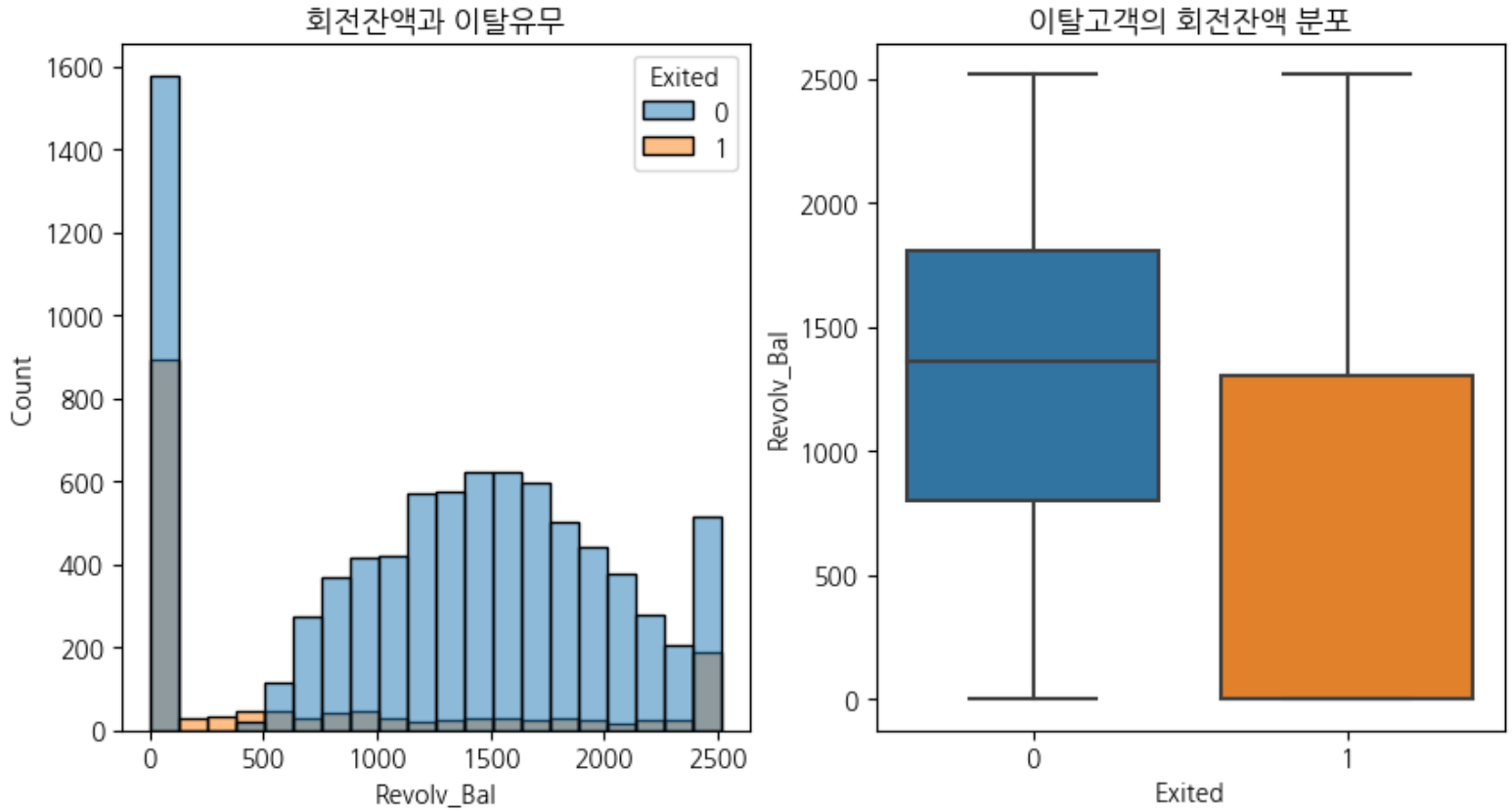
은행 고객들의 신용한도와 그에 따른 이탈 유무

신용한도가 낮은 고객의 수가 많은 것을 확인하였으며 이는 앞의 고객 연봉과도 상관 관계가 있는 것으로 보여진다.



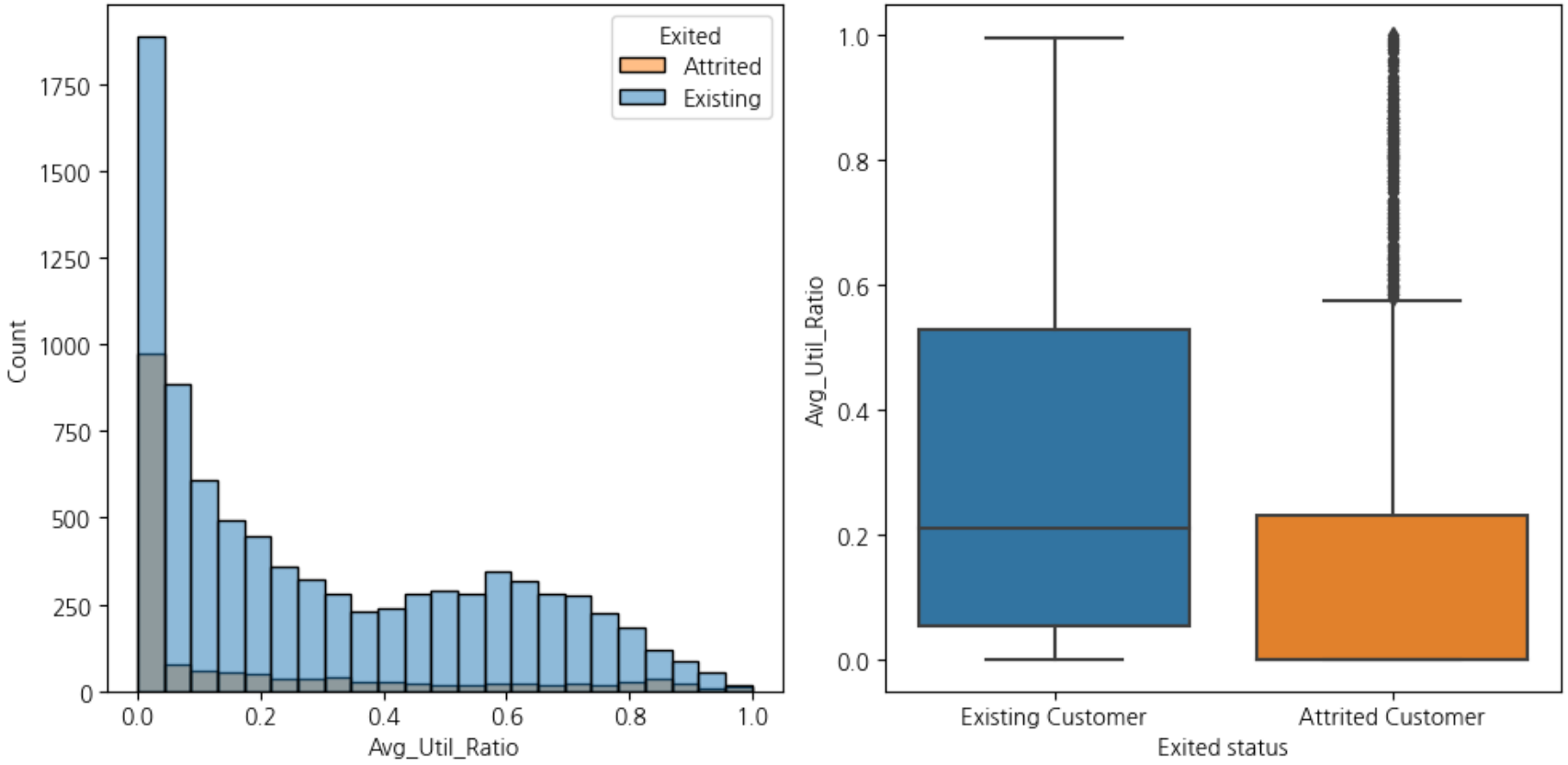
은행 고객들의 리볼빙 잔액과 그에 따른 이탈 유무

부채가 0이거나 2,500 이상의 높은 부채를 가지고 있는 고객의 이탈률이 높은 것으로 보여진다.



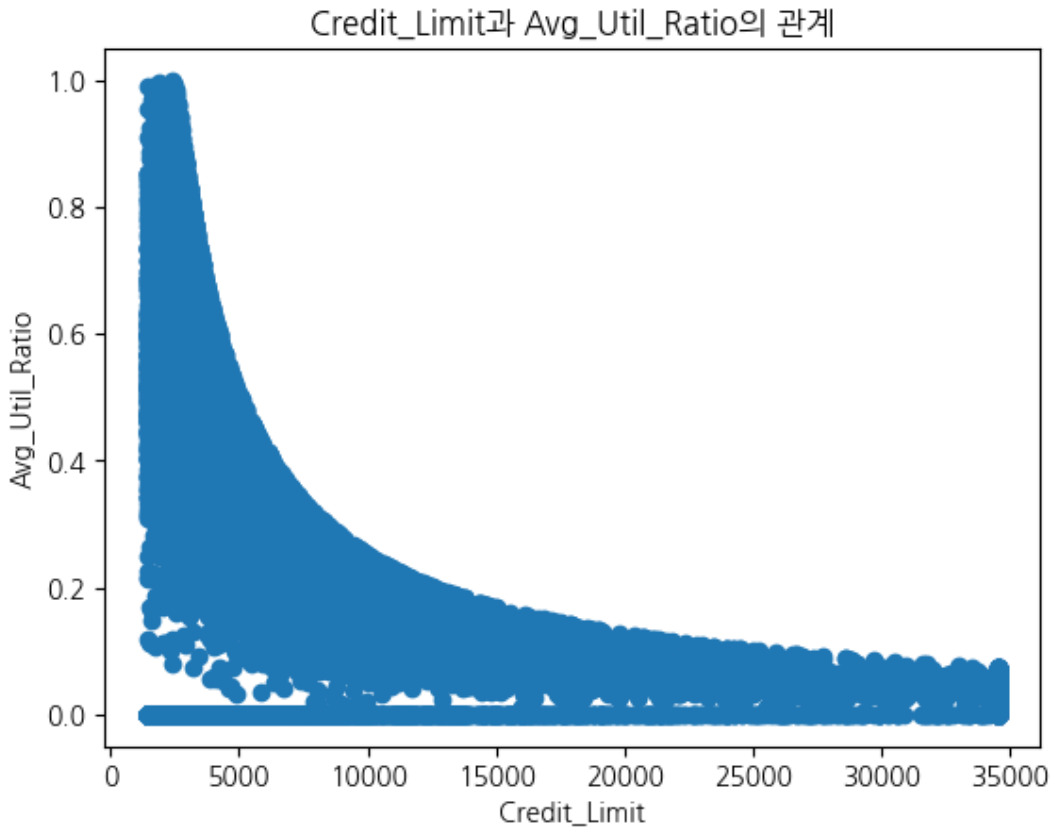
은행 고객들의 평균 이용률과 그에 따른 이탈 유무

신용카드 이용률이 0에 가까울 수록 이탈률이 높은 것으로 확인되었다.



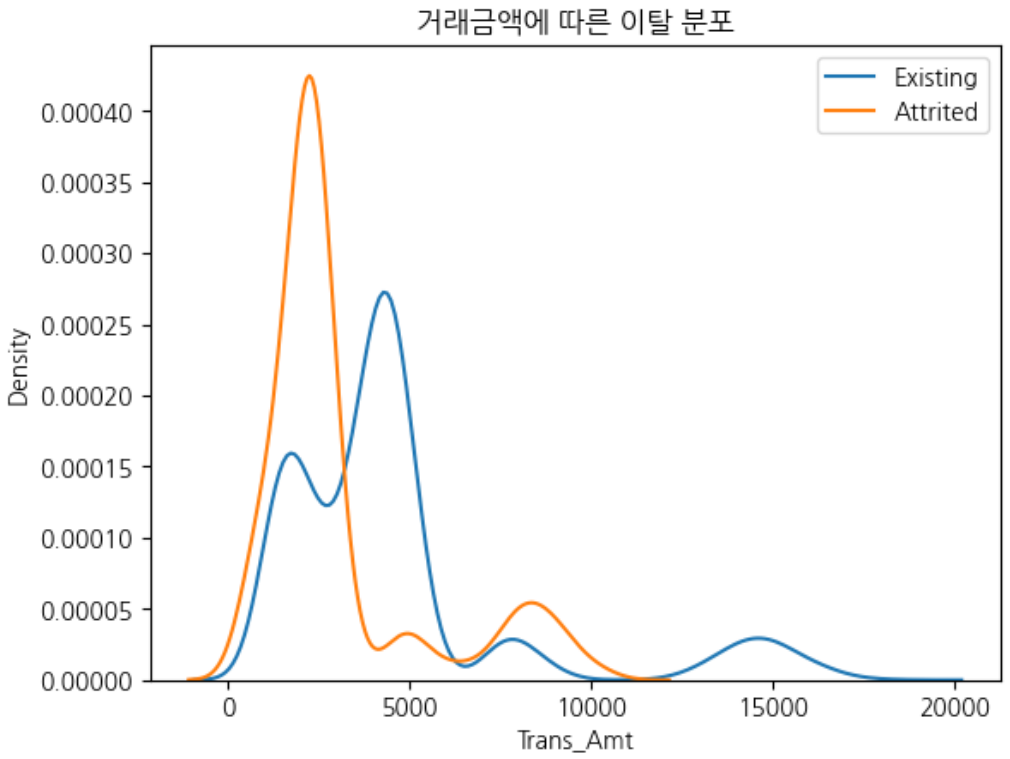
신용 한도와 평균 이용률과의 상관 관계

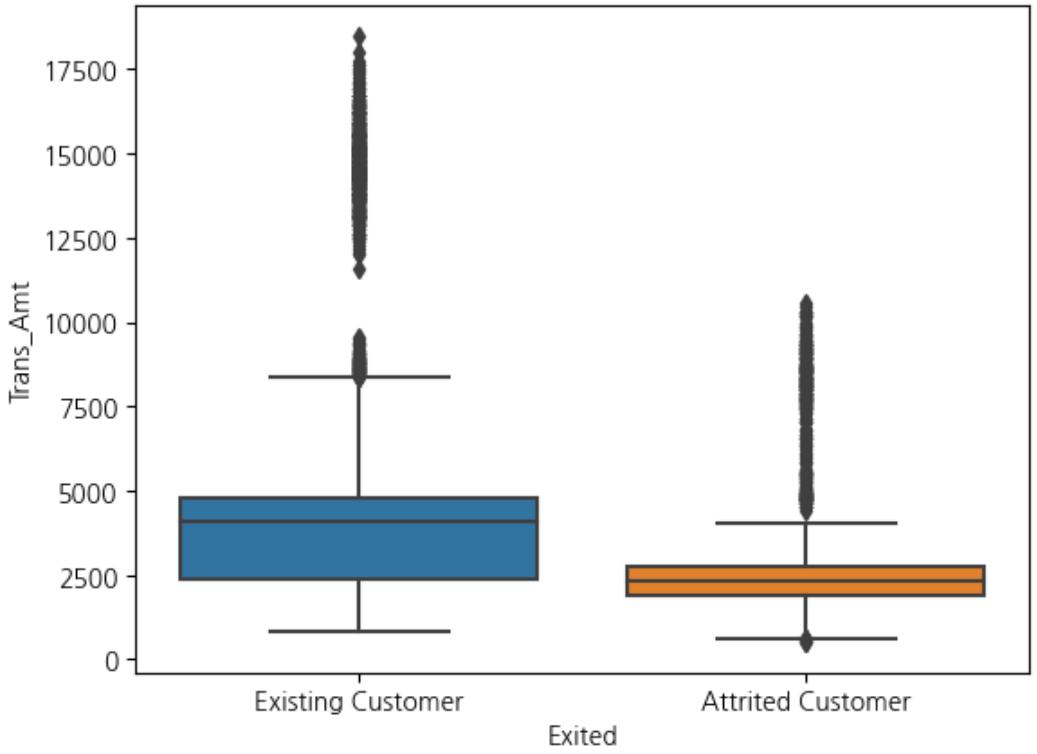
신용 한도가 낮을수록 평균 이용률이 높아지는 것을 확인하였다.
이는 한도가 낮을수록 신용 카드를 많이 사용하는 고객이 많다는 것을 의미하며
고객의 부채가 크거나 고객의 신용 카드 이용률이 높다는 것을 나타낸다.



은행 고객들의 거래 금액과 그에 따른 이탈 유무

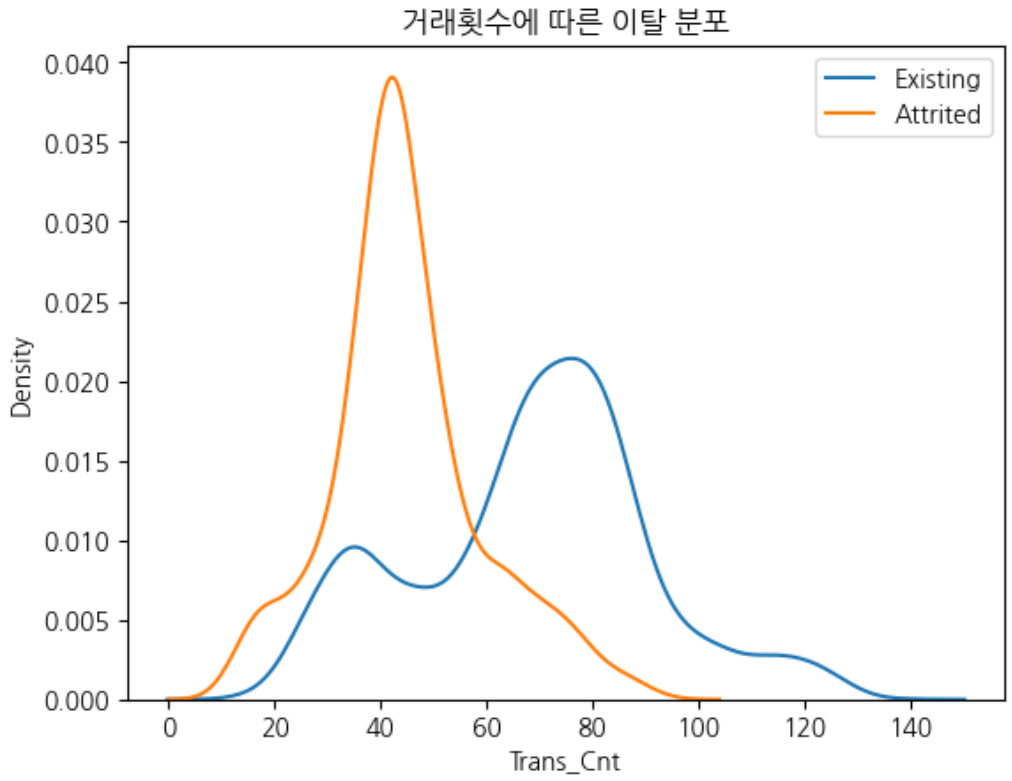
신용카드 거래 금액이 높은 고객은 낮은 이탈률을 가지고 있음을 확인할 수 있다.

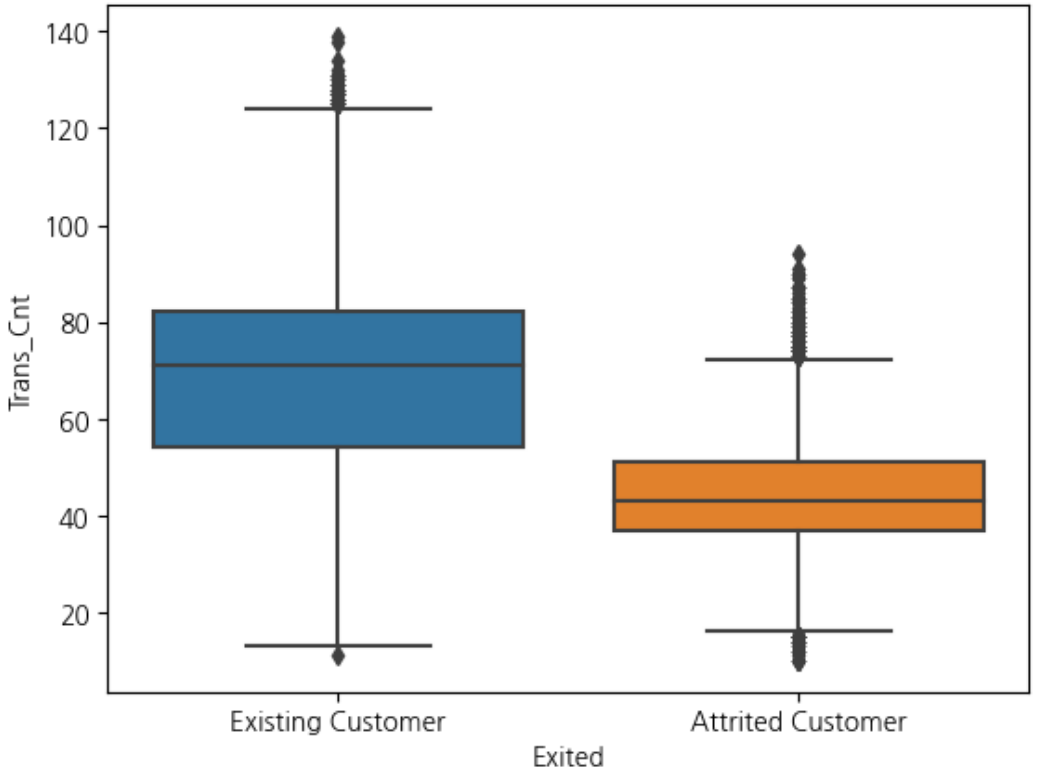


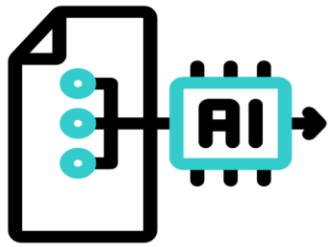


은행 고객들의 거래 횟수와 그에 따른 이탈 유무

고객의 거래 횟수가 많을수록 이탈률이 낮은 것을 확인할 수 있다.







데이터 전처리 과정

결측치/이상치 확인

변수 인코딩

Train/Test Split

데이터 스케일링

결측치 확인 및 제거

```
df_new.isnull().sum()
```

```
Exited 0 Age 0 Gender 0 Dependents 0
Education 0 Marital 0 Income 0 Card_Type 0
Tenure 0 Product_Cnt 0 Inactive_Months 0
Contacts_Cnt 0 Credit_Limit 0 Revolv_Bal 0
Avg_OTB 0 Avg_Util_Ratio 0 Trans_Amt 0
Trans_Cnt 0 Amt_Chng_Q4_Q1 0 Cnt_Chng_Q4_Q1 0
dtype: int64
```

변수 인코딩 (Label Encoding)

‘Exited’: 유지 0, 이탈 1
‘Gender’: 남자 0, 여자 1
‘Education’: unedu 0, high 1, col 2,
gradu 3, post 4, doc 5, unknown 6
‘Marital’: single 0, married 1, divorce 2,
unknown 3
‘Income’: less40 0, 40~60 1, 60~80 2, 80~120 3,
120+ 4, unknown 5
‘Card_Type’: blue 0, silver 1, gold 2, platinum 3

Train/Test Split

```
from sklearn.model_selection import train_test_split
```

데이터를 트레인:테스트 (80:20)으로 분리

데이터 스케일링 (Standard Scaler)

```
from sklearn.preprocessing import StandardScaler
std = StandardScaler()
```

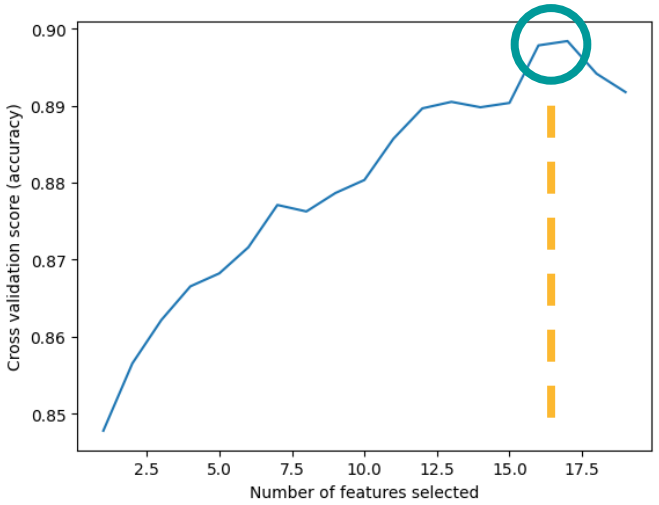
Train 데이터 스케일링

03

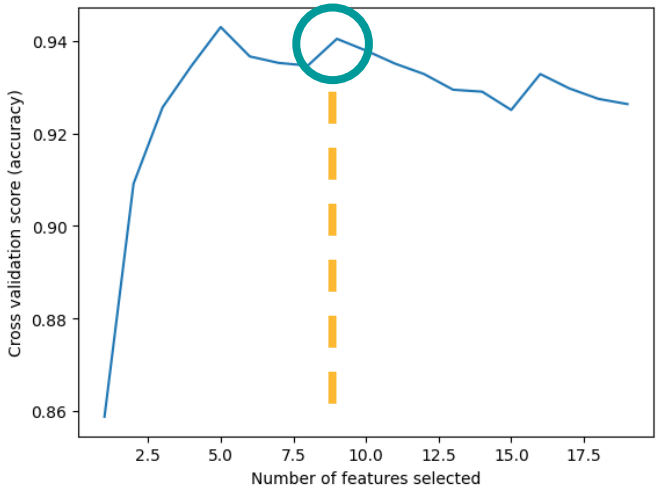
모델 학습 및 모델 평가

Feature Selection

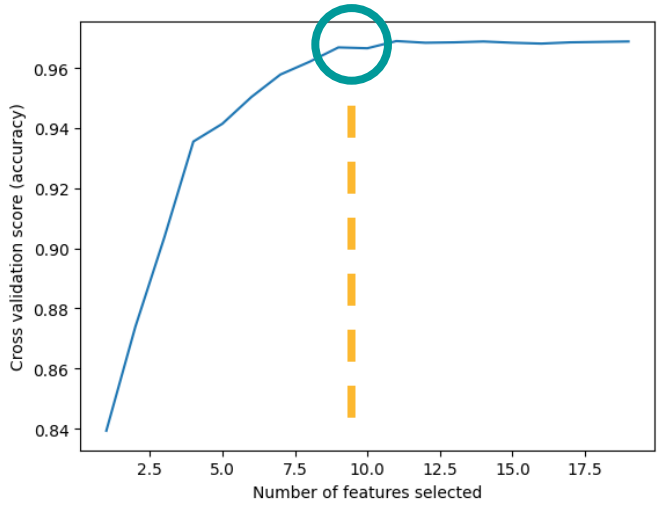
RFECV(Recursive Feature Elimination with Cross Validation)을 이용하여 2차 변수 선정을 진행 (8~9개)



LogisticRegression



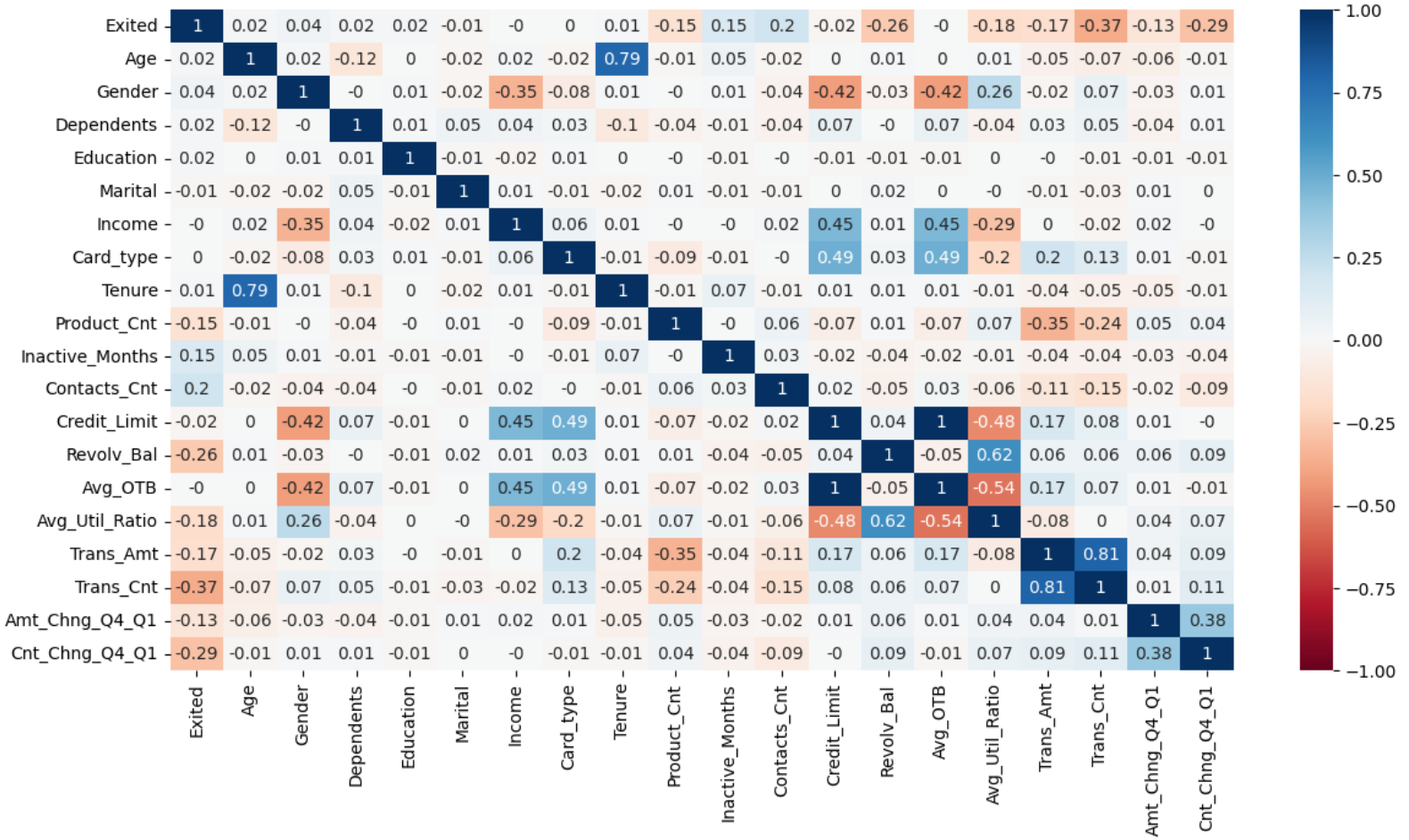
RandomForest



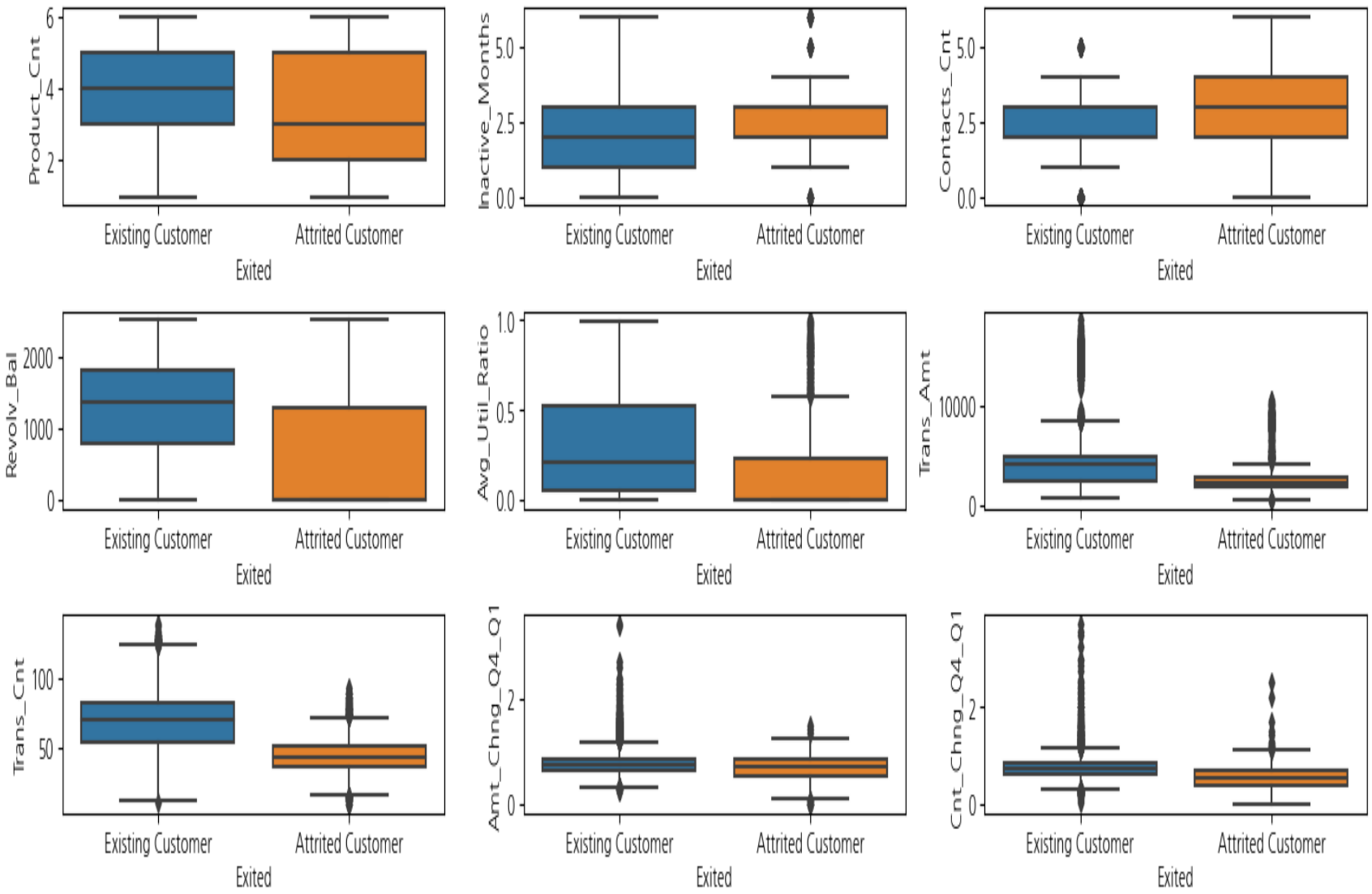
XGBClassifier

Heat Map: Correlation

'Product_Cnt'
'Inactive_Months'
'Contacts_Cnt'
'Revolv_Bal'
'Avg_Util_Ratio'
'Trans_Amt'
'Trans_Cnt'
'Amt_Chng_Q4_Q1'
'Cnt_Chng_Q4_Q1'



BoxPlot


Product_Cnt
Inactive_Months
Contacts_Cnt
Revolv_Bal
Avg_Util_Ratio
Trans_Amt
Trans_Cnt
Amt_Chng_Q4_Q1
Cnt_Chng_Q4_Q1
Existing Customer
Attrited Customer
Exited

Feature Selection

8개 컬럼을 이용한 모델 평가

'Product_Cnt', 'Inactive_Months', 'Revolv_Bal', 'Avg_Util_Ratio', 'Trans_Amt', 'Trans_Cnt', 'Amt_Chng_Q4_Q1','Cnt_Chng_Q4_Q1'

Name	Accuracy	Precision	Recall	F1
LogisticRegression	0.896218	0.739647	0.548838	0.629356
RandomForestClassifier	0.934433	0.893580	0.672389	0.767165
KNeighborsClassifier	0.930385	0.848893	0.690822	0.761249
XGBClassifier	0.962279	0.902100	0.858630	0.879711

Feature Selection

9개 컬럼을 이용한 모델 평가

'Product_Cnt', 'Inactive_Months', 'Contacts_Cnt', 'Revolv_Bal', 'Avg_Util_Ratio', 'Trans_Amt', 'Trans_Cnt', 'Amt_Chng_Q4_Q1','Cnt_Chng_Q4_Q1'

Name	Accuracy	Precision	Recall	F1
LogisticRegression	0.898292	0.742063	0.564819	0.640557
RandomForestClassifier	0.940752	0.889831	0.720934	0.796117
KNeighborsClassifier	0.928311	0.840770	0.685297	0.754543
XGBClassifier	0.964748	0.908250	0.868459	0.887775

03 모델 학습 및 모델 평가

Model Selection

Grid Search CV와 베이지안 최적화를 이용해 최적의 파라미터를 설정

```
from sklearn.model_selection import GridSearchCV
from hyperopt import fmin, tpe, hp, Trials,
space_eval
```

Logistic Regression

이진 분류에 적합한 선형 모델로, 계수들의 해석이 용이

Random Forest

높은 예측 성능을 보이며 결정 트리들의 조합으로 과대적합을 줄이고
안정적인 예측을 제공

KNN

간단하고 직관적인 모델

XGBoost

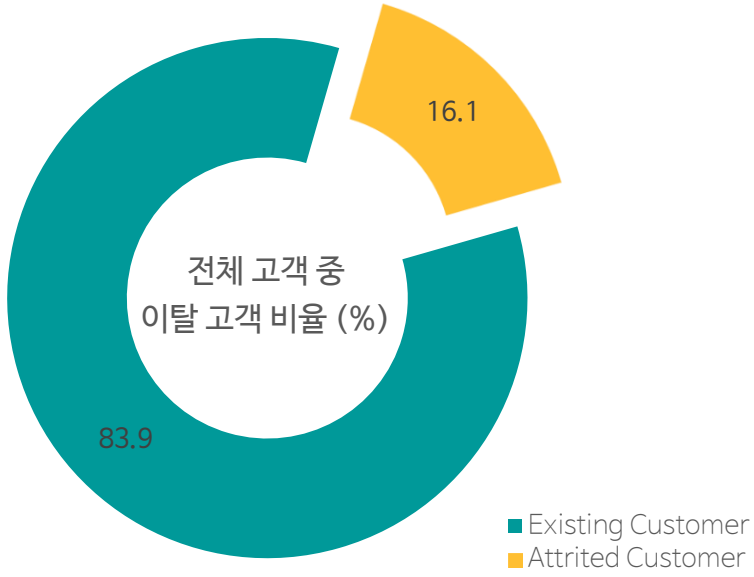
부스팅 알고리즘으로, 랜덤 포레스트보다 높은 예측 성능
과적합에 강하고, 높은 일반화 성능을 제공

Best Parameters

Model	Parameter
Logistic Regression	{'C': 3, 'penalty': 'l1'}
Random Forest	'clf__max_depth': 9, 'clf__min_samples_leaf': 1, 'clf__min_samples_split': 5, 'clf__n_estimators': 100
KNN	'metric': 'manhattan', 'n_neighbors': 10, 'weights': 'distance'
XGBoost	‘gamma': 2, 'learning_rate': 0.05, 'max_depth': 9, 'min_child_weight': 1, 'n_estimators': 300, 'subsample': 0.7

모델 비교

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.901777	0.752	0.578462	0.653913
Random Forest	0.955084	0.882352	0.830769	0.855784
KNN	0.886969	0.6666667	0.590769	0.626427
XGBoost	0.970385	0.926045	0.886154	0.905660



Over Sampling

불균형한 클래스 분포를 가진 데이터에서 소수 클래스의 샘플을 증가시킴

Random Oversampling

소수 클래스의 샘플을 중복해서 늘림

SMOTE

소수 클래스의 샘플을 이용하여 합성된 샘플을 생성

ADASYN

SMOTE의 변형으로, 소수 클래스 샘플에 가중치를 부여하여 합성된 샘플을 생성

Borderline-SMOTE

소수 클래스의 경계에 위치한 샘플을 주로 증가

Over Sampling 비교

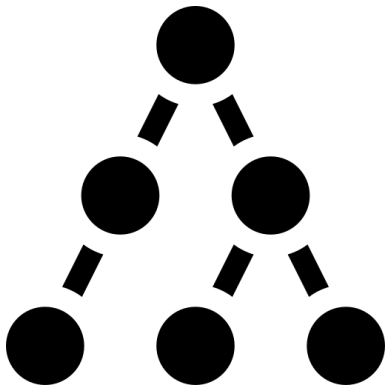
Method	Accuracy	Precision	Recall	F1
Random Oversampling	0.956071076011846	0.8172043010752689	0.9353846153846154	0.8723098995695839
SMOTE	0.9565646594274433	0.8264462809917356	0.9230769230769231	0.872093023255814
Borderline-SMOTE	0.9521224086870681	0.8	0.9353846153846154	0.8624113475177304
ADASYN	0.9501480750246791	0.7931937172774869	0.9323076923076923	0.8571428571428572

최종 모델 - XGBoost (SMOTE oversampling)

'gamma': 2, 'learning_rate': 0.05, 'max_depth': 9,
'min_child_weight':1, 'n_estimators': 300, 'subsample': 0.7

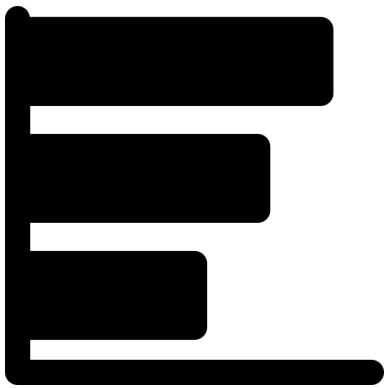
04

프로젝트 결론



Feature Importance

XGBoost 모델의 트리 구조에서
각 특성이 사용된 빈도수를 기준으로 중요도를 계산



SHAP Value

모든 특성 조합의 경우를 고려하여 각 특성의 기여도를 계산

Feature Importance

Trans_Cnt 고객의 총 거래 횟수

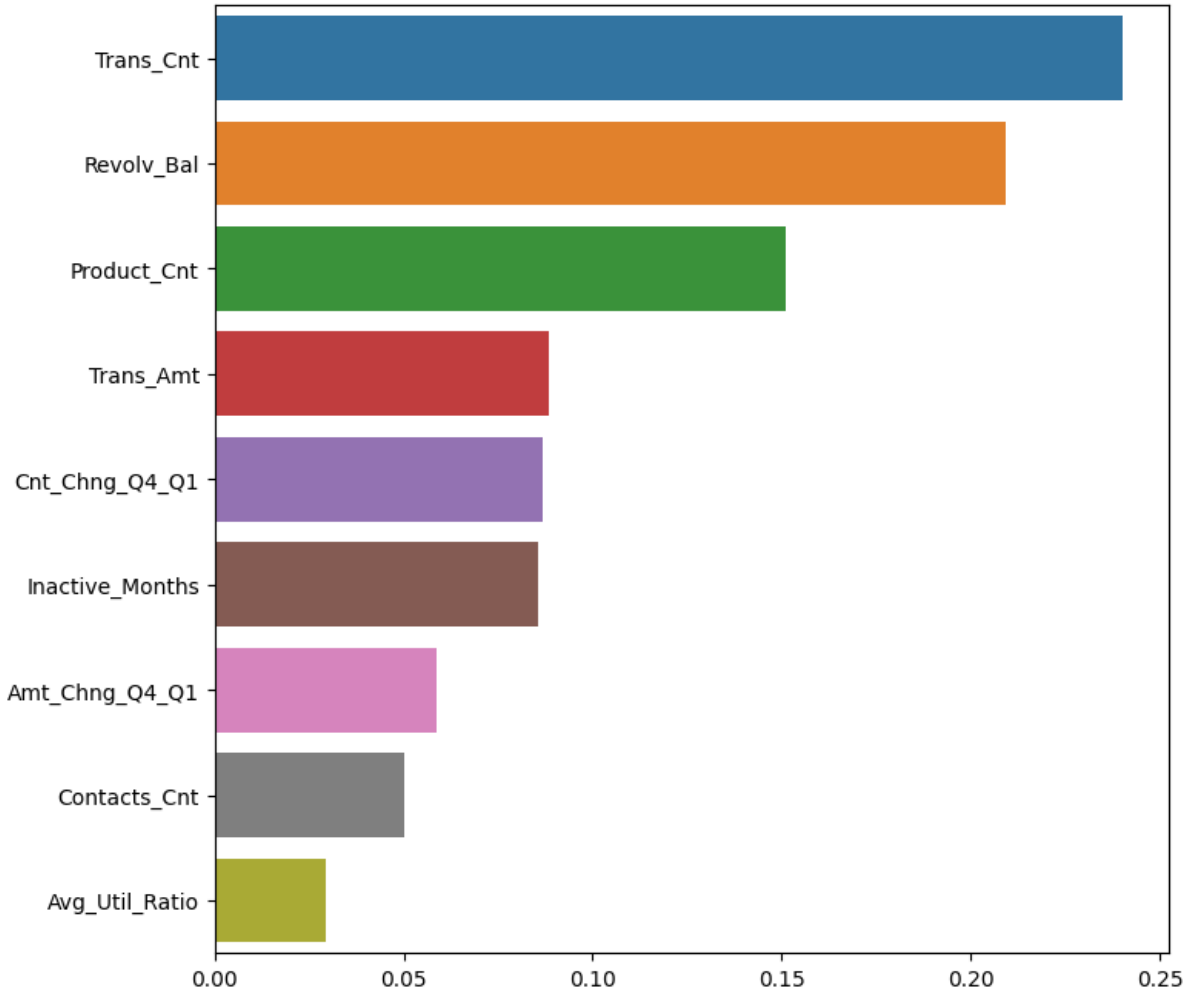
거래 횟수가 적은 고객은 거래를 자주 하는 고객보다 이탈할 확률이 높습니다.

Revolv_Bal 리볼빙 잔액

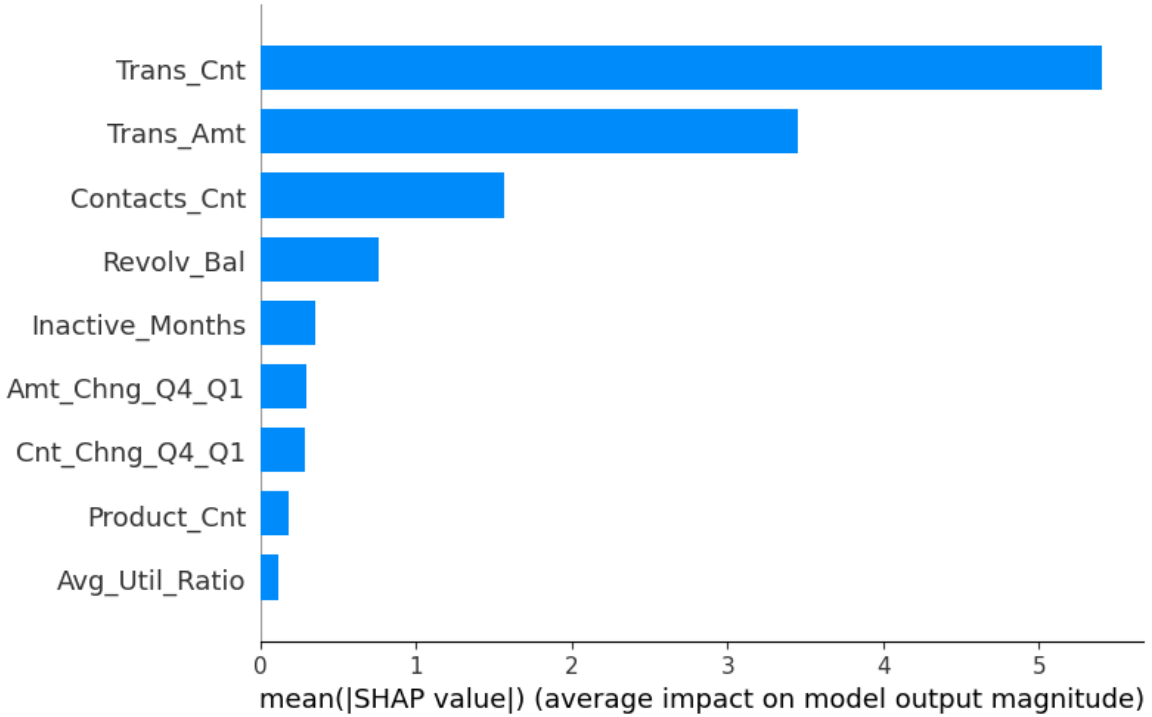
리볼빙 잔액은 고객의 부채 상태를 나타내는 지표로 리볼빙 잔액이 작을수록 이탈할 확률이 높습니다.

Product_Cnt 가입된 상품의 갯수

고객이 가입한 상품의 개수로 많은 상품에 가입되어 있는 고객은 그렇지 않은 고객보다 이탈률이 낮습니다.



04 프로젝트 결론



SHAP Value

Trans_Cnt 고객의 총 거래 횟수

거래 횟수가 적은 고객은 거래를 자주 하는 고객보다 이탈할 확률이 높습니다.

Trans_Amt 고객의 총 거래량

거래 금액의 양이 클수록 이탈 확률이 작습니다.

Inactive_Months 비활성 기간

고객이 은행을 이용하지 않은 기간으로 해당 변수의 중요도가 높을 것으로 예상했으나 비교적 낮은 중요도를 가진 것으로 확인하였습니다

04 프로젝트 결론

분석 의의

은행을 이용하는 고객들의 이탈에 영향을 주는 요인 파악

높은 이탈률을 가진 고객을 대상으로 마케팅 전략 기획
고객 유지를 위한 고객 관리 프로그램 실시

신규 고객 유치에 들어가는 마케팅 비용 감소

충성도 높은 고객을 확보하여 매출 증대

분석 한계점

비식별화 된 가상 데이터

데이터가 비식별화 되어있으며 모델 학습을 위한 추가 변수 수집이 불가능

가상의 데이터로 실제 상황에 바로 적용하기에는 부적합
데이터 이해와 모델 적용에 한계

감사합니다