서울대학교 빅데이터 핀테크 과정 기계학습과 딥러닝 팀 프로젝트

신용카드 서비스 고객 이탈 예측

4조 박석훈 박태준 윤성규 임동건

31 August 2023

Contents

데이터 소개

Explaining the Features in the Dataset

데이터 소개

Credit Card customers

-Predict Churning Customers(6기 데이터)
카드를 이용하는 고객 중 서비스를 이용하지 않을 고객 예측

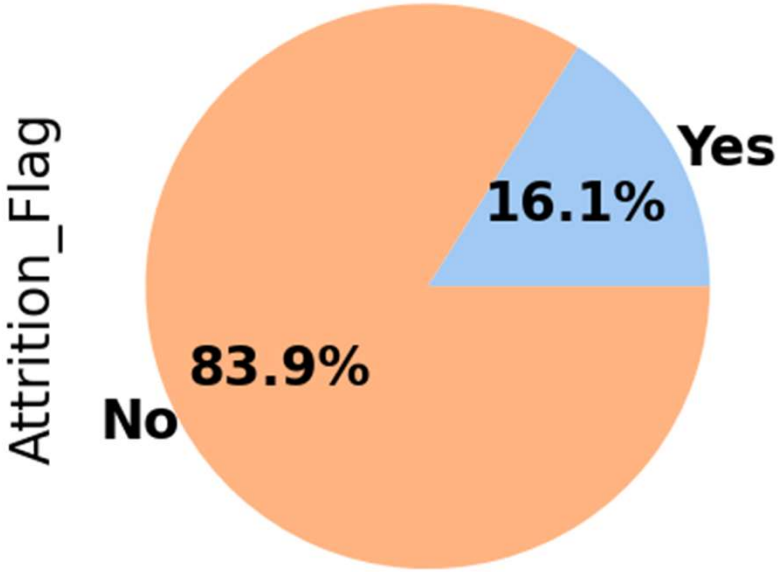
총 10,217명 고객의 연령, 성별, 소득 수준 등 개인 정보와 거래금액, 신용한도, 리볼빙 잔액 등 21가지 항목에 대한 정보로 구성

데이터 출처:https://www.kaggle.com/sakshigoyal7/credit-card-customers

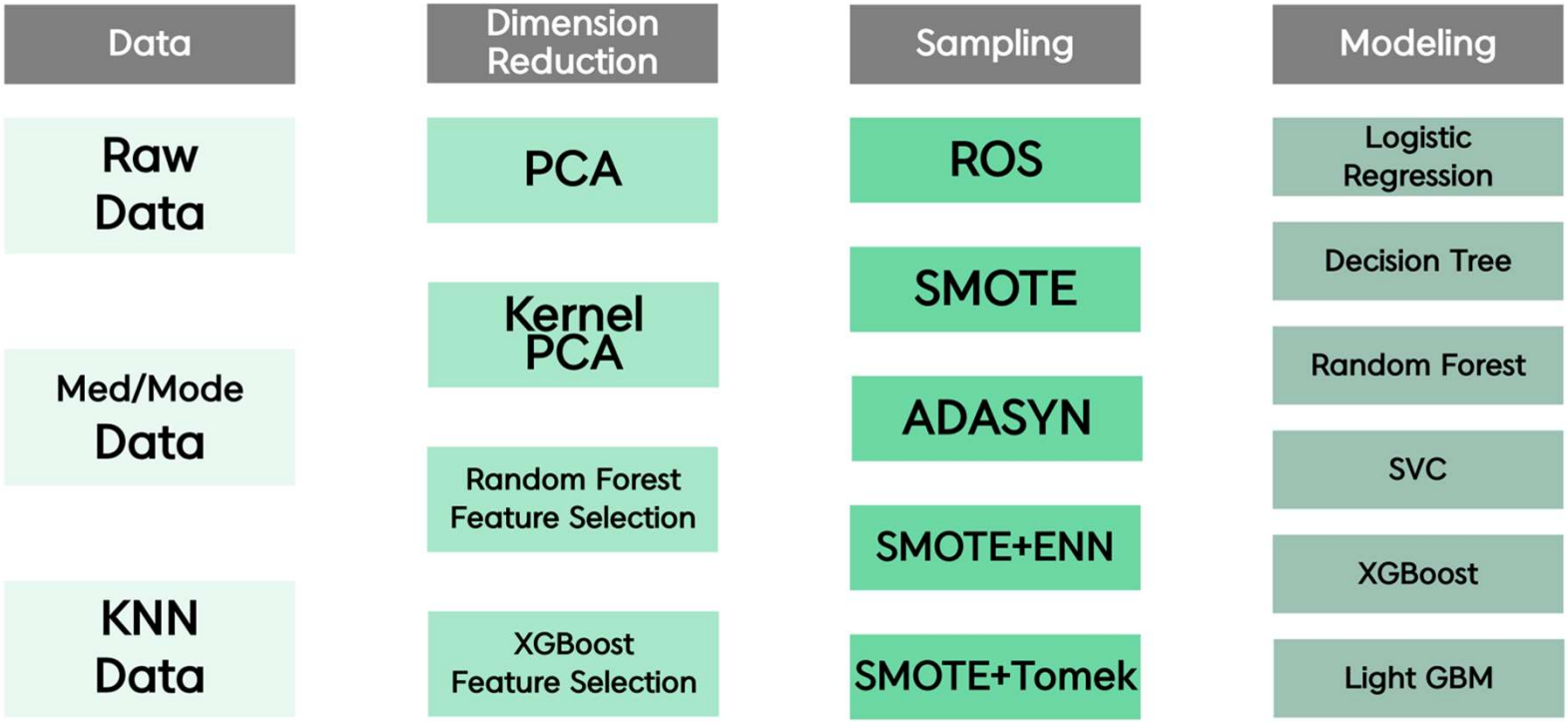
반응 변수 : Attrition_Flag

계좌 해지 여부를 나타내는 변수로, 1은 YES, 0은 NO를 의미

아래 그림에서 알 수 있듯이 1 비율이 적은 imbalanced data이다.



기존 모델 분석 프로세스



출처 : 6기 발표 자료

기존 모델 결과

Model	Condition		Accuracy	Recall
Ridge	Lambda 5 / Threshold 0.45		0.8179	0.9035
SVC	PCA / C 10 / Gamma 1 / Kernel kbf		0.7937	0.6527
Decision Tree	Max_Depth 10 / Max_Features None / Min_sam_split 4	PCA	0.7586	0.7717
		FS	0.9235	0.8457
RandomForest	Max_Depth 10 / Max_Features Auto / n_estimators 100	Min_sam_split 4	0.8223	0.7749
		Min_sam_split 4	0.9235	0.8457
XGBoost	Gamma 0.1 / learning_rate 0.2 / n_estimators 100	PCA / Max_depth 10	0.8948	0.5723
		PCA / Max_depth 6	0.9294	**0.9035**
LGBM	Learning_rate 0.2 / n_estimators 100	PCA / Max_Depth 12 / num_leaves 50	0.8850	0.5981
		PCA / Max_Depth 10 / num_leaves 30	**0.9368**	0.9164

데이터 탐색 및 전처리

EDA - Exploratory Data Analysis

범주형 변수 설명

변수명	데이터 타입	내용
Attrition_Flag	**범주형**	**서비스 이탈 여부** **(Attrited Customer : 이탈함. Existing Customer : 계속 사용)**
Gender	범주형	성별 (M : Male. F : Female)
Education_Level	범주형	학력수준 (Doctorate, Post-Graduate, College, High School, Uneducated, Unknown)
Marital_status	범주형	결혼 여부 (Married, Single, Divorced, Unknown)
Income _Category	범주형	연간 소득 범주 (Unknown, <$40k, $40k, $-60k, 60k, -$80k,$80k, -$120k,>$120k)
Card_category	범주형	카드 타입 (Blue, Silver, Gold, Platinum)

수치형 변수 설명 - 1

변수명	데이터 타입	내용
CLIENTUM	수치형	계좌 보유 고객에게 부여한 번호
Customer_Age	수치형	고객의 나이
Dependent_count	수치형	가구 구성원 수
Months_on_book	수치형	고객의 은행과의 거래 개월 수
Total_Relationship_Count	수치형	고객이 보유하는 총 금융 상품 수
Months_Inactive_12_mon	수치형	최근 12개월 간 카드 거래가 일어나지 않은 달의 수
Contacts_Count_12_mon	수치형	최근 12개월 간 연락 횟수
Credit_Limit	수치형	신용카드의 신용한도

수치형 변수 설명 - 2

변수명	데이터 타입	내용
Total_Revolving_Bal	수치형	신용카드의 총 리볼빙 잔액 (할부 구매 잔액)
Avg_Open_To_Buy	수치형	최근 12개월 평균 신용 대출 한도
Total_Amt_Chng_Q4_Q1	수치형	1/4분기 대비 4/4분기 거래 금액 변화량
Total_Trans_Amt	수치형	최근 12개월 총 거래대금
Total_Trans_Ct	수치형	최근 12개월 총 거래횟수
Total_Ct_Chng_Q4_Q1	수치형	1/4분기 대비 4/4 분기 거래횟수 변화량
Avg_Utilization_Ratio	수치형	총 평균 이용률 (총 리볼빙 잔액 / 신용 한도)
Months_Inactive_12_mon	수치형	최근 12개월 간 카드 거래가 일어나지 않은 달의 수

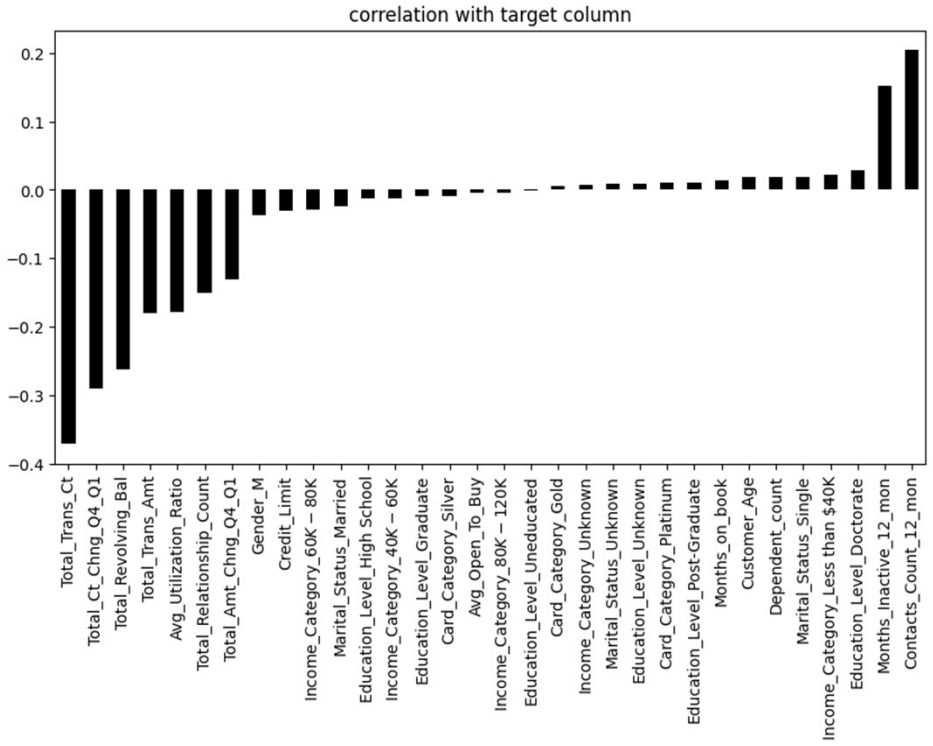
범주형 변수 처리 - One-Hot-Encoding

인코딩 된 변수 간 공선성으로 인해 범주 하나씩 지우고 진행 ex) Gender - Female

변수명	데이터 타입	내용
Attrition_Flag	**범주형**	**'Attrition_Flag'**
Gender	범주형	'Gender_M'
Education_Level	범주형	'Education_Level_Doctorate', 'Education_Level_Graduate', 'Education_Level_High School', 'Education_Level_Post-Graduate', 'Education_Level_Uneducated', 'Education_Level_Unknown',
Marital_status	범주형	'Marital_Status_Married', 'Marital_Status_Single', 'Marital_Status_Unknown',
Income _Category	범주형	'Income_Category_$40K - $60K', 'Income_Category_$60K - $80K', 'Income_Category_$80K - $120K', 'Income_Category_Less than $40K', 'Income_Category_Unknown',
Card_category	범주형	'Card_Category_Gold', 'Card_Category_Platinum', 'Card_Category_Silver'

반응변수에 따른 변수들 상관계수

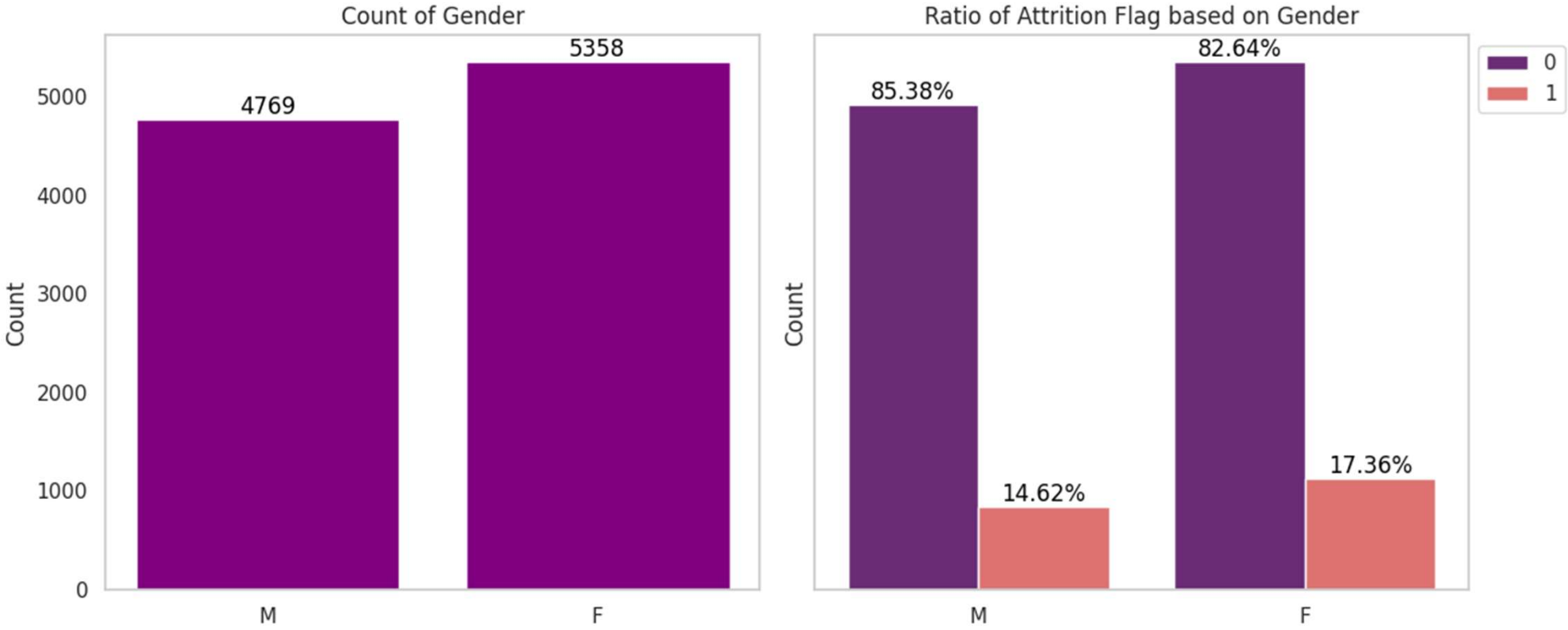
반응변수에 따른 변수들 상관계수 비교



- 카드를 **활발히 이용하는 경우**
  반응 변수와의 상관 계수 **낮아짐**
  - Total_Trans_Ct
  - Total_Amt_Chng_Q4_Q1
  - Total_Revolving_Bal
  - Total_Trans_Amt

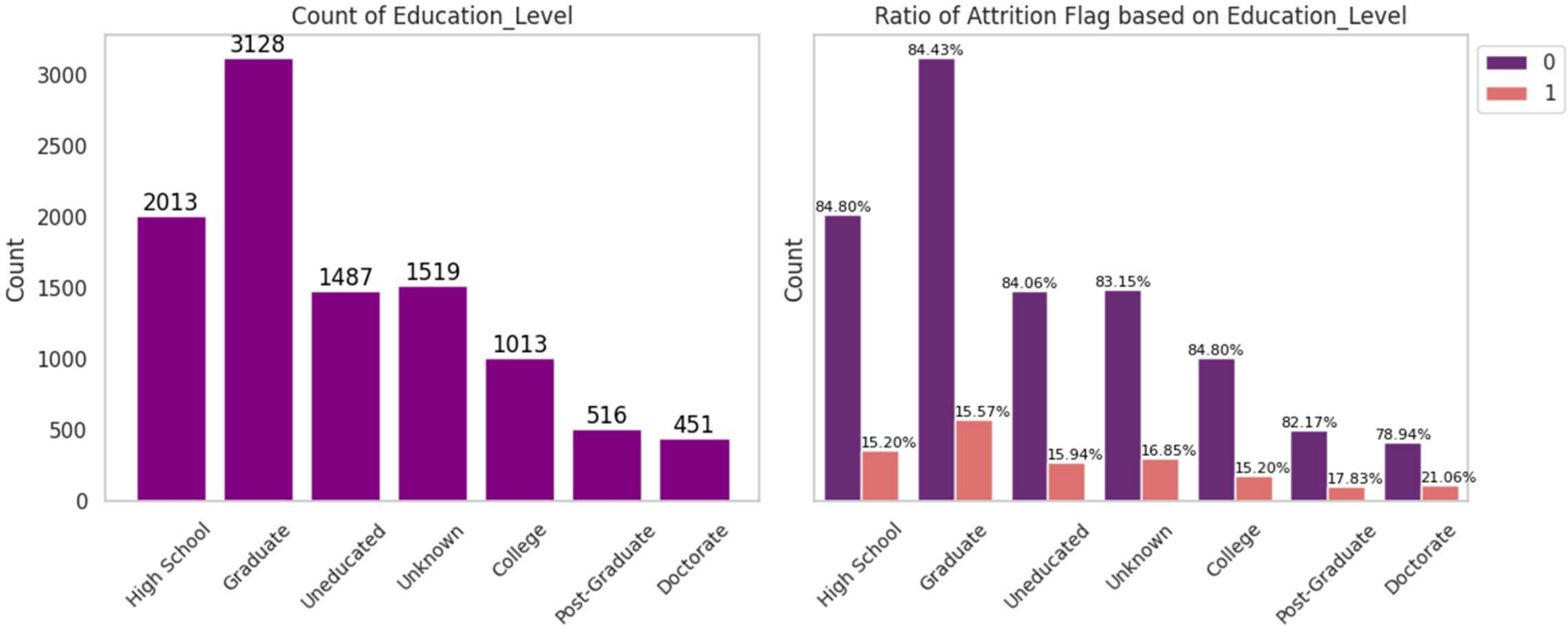
- 카드를 **활발히 이용하지 않는 경우**
  반응변수와의 상관 계수가 **높아짐**
  - Months_Inactive_12_mon
  - Contacts_Count_12_mon

범주형 변수 시각화 - 성별



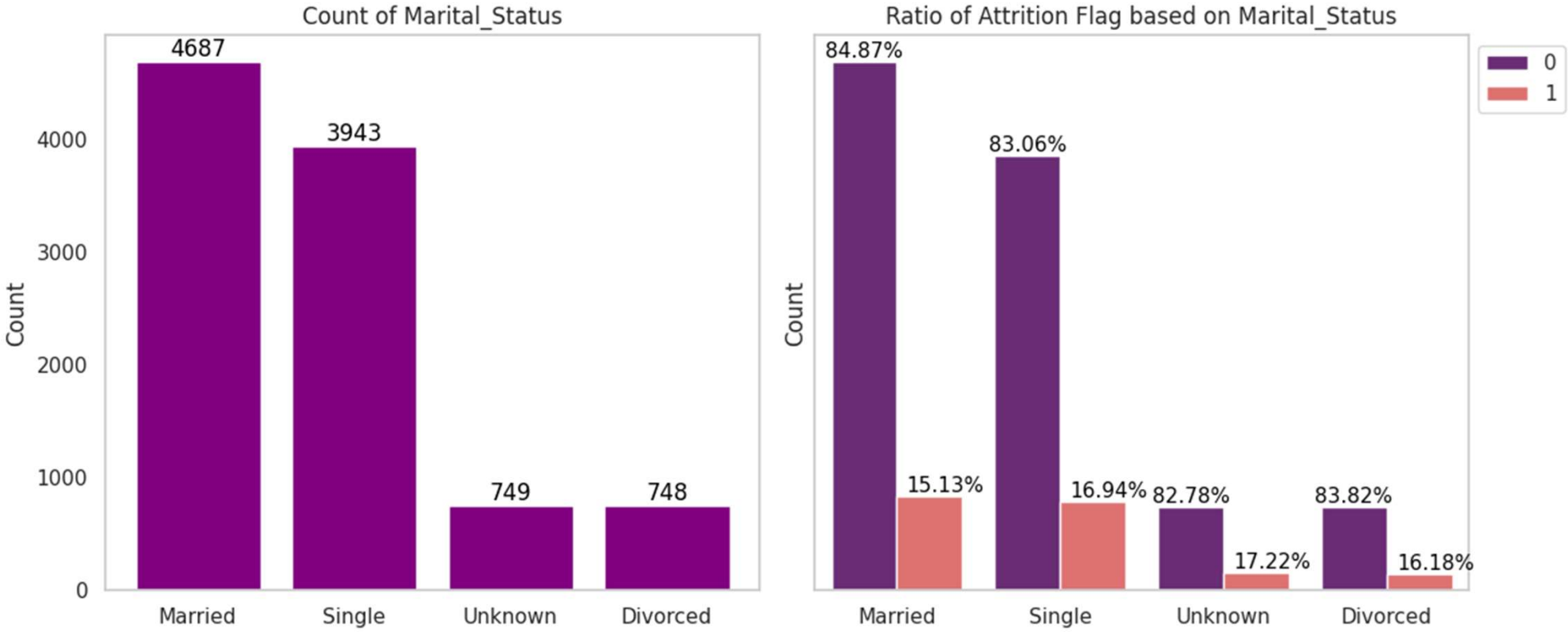
여성이 남성에 비해 2.74%p의 높은 이탈율을 보이고 있습니다.

범주형 변수 시각화 - 교육 수준



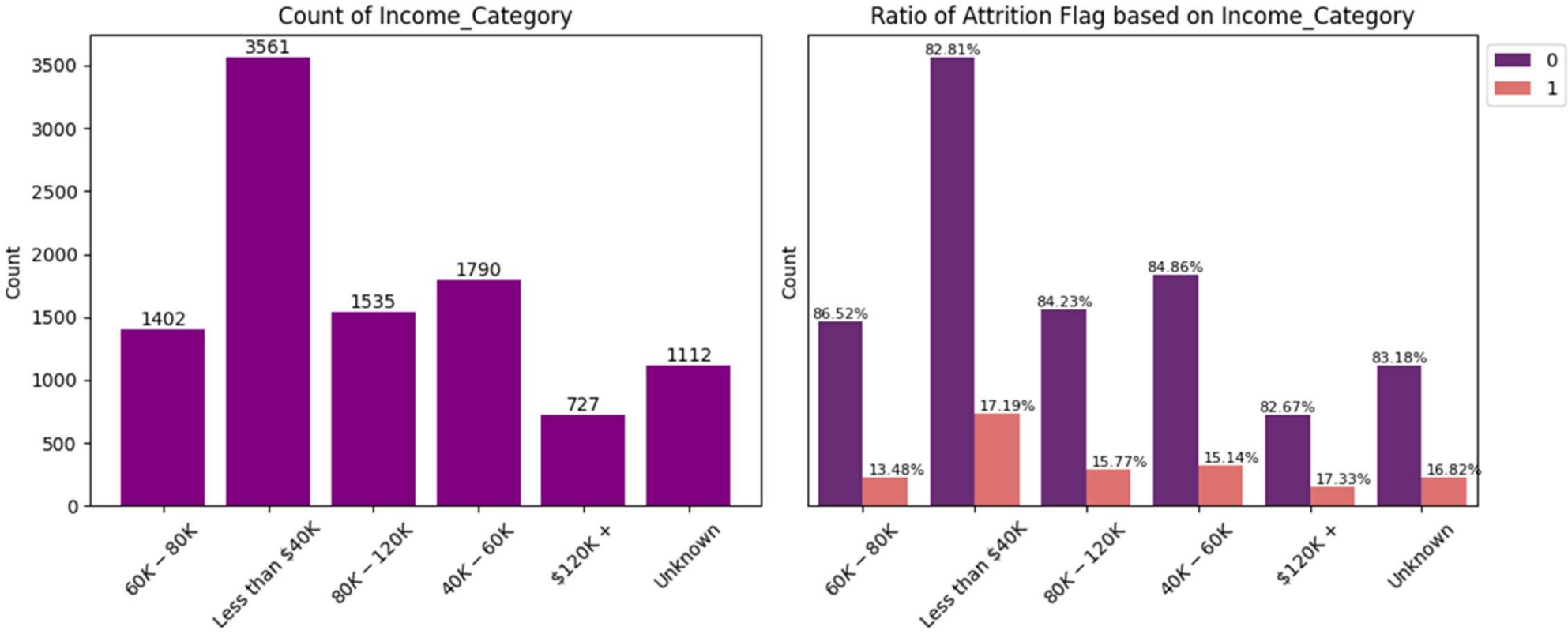
교육 수준이 높은 Doctorate (21.06%), Post-Graduate (17.83%)에서 다른 수준 보다 비교적 높은 이탈율을 보임

범주형 변수 시각화 - 결혼 여부



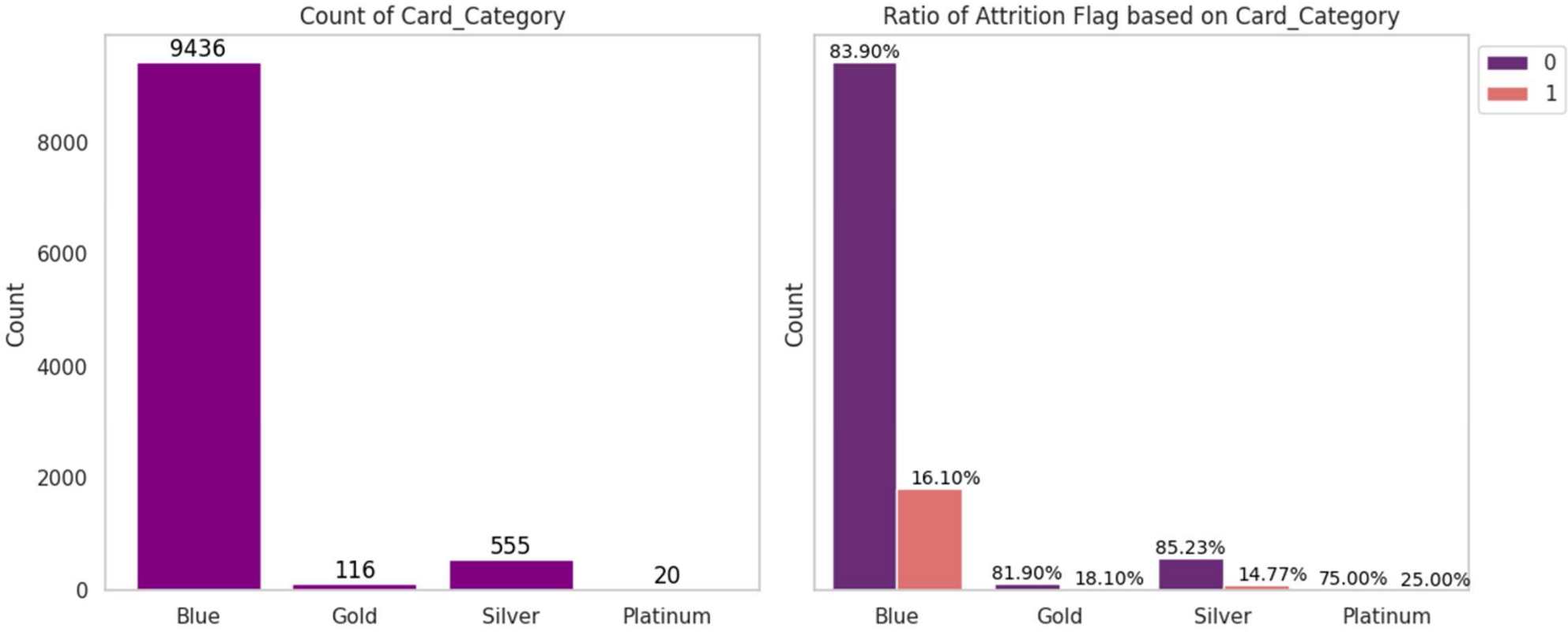
Married 상태(15.13%)에서 다른 상태보다 이탈율이 조금 낮은 것을 볼 수 있음

범주형 변수 시각화 - 소득 수준



40K 미만 부분(17.19%)에서 가장 높은 이탈율을 보임

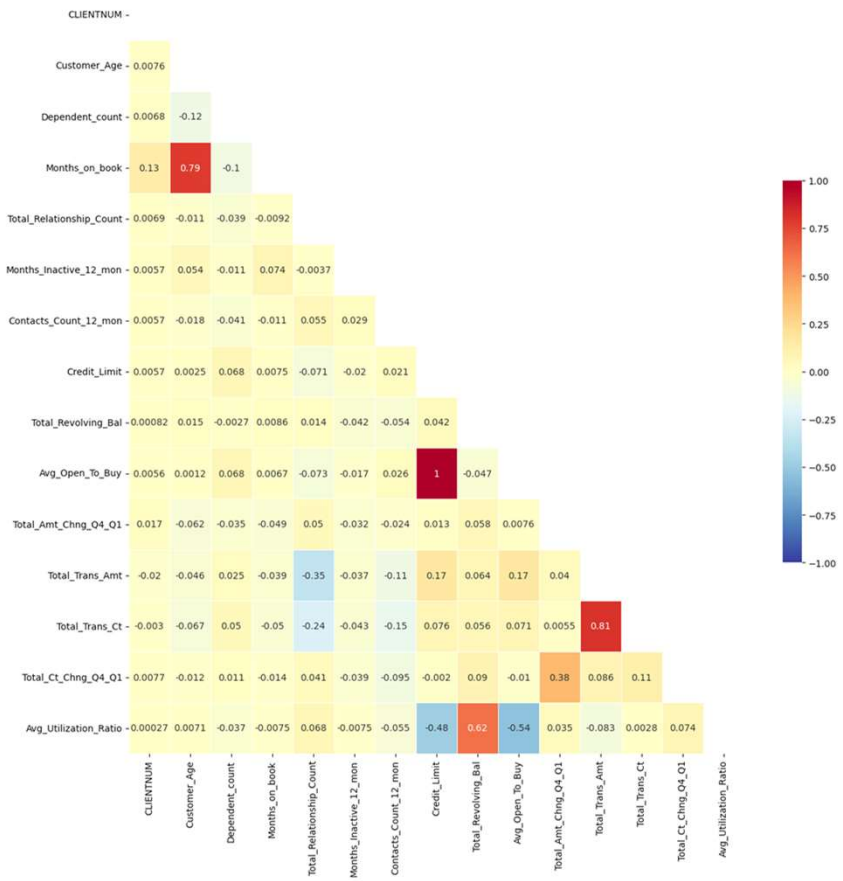
범주형 변수 시각화 - 카드 타입



Platinum (25%) 등급에서의 가장 이탈율이 높았으며, 대부분의 사람들은 Blue 등급의 카드를 사용하고 있음을 확인할 수 있음.

수치형 변수 간 상관관계 분석

Correlation Matrix 활용 수치형 변수 간 상관관계 확인



Correlation 의 절댓값이 큰 상황

- **Avg_Open_To_Buy - Credit_Limit : 1**
  - 최근 12개월 평균 신용 대출 한도와 신용카드의 신용한도는 고객의 신용정도와 매우 연관이 크므로 두 변수 중 하나만 사용해도 무방하다고 생각됨

- **Months_on_book - Customer_Age : 0.81**
  - 고객의 은행과의 거래 개월 수와 고객의 나이는 고객의 나이가 많으면 거래한 개월 수가 클 확률이 높으나 변수를 삭제하면 나이 많은 신규 고객의 이탈에 대한 정보를 얻기 힘들다고 판단하여 두 변수 모두 사용하기로 함

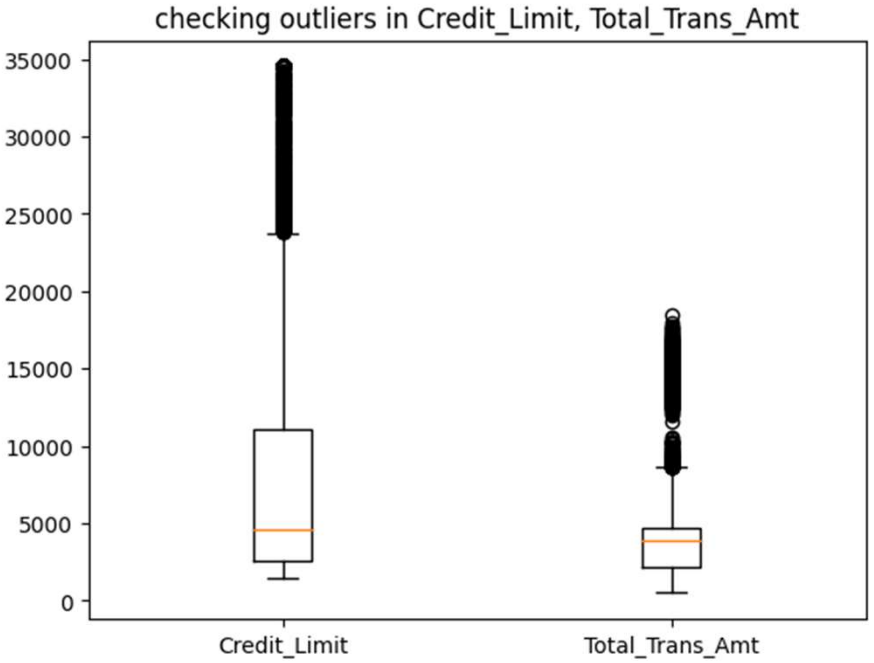
- **Total_Trans_Amt - Total_Trans_Ct : 0.79**
  - 최근 12개월 총 거래대금 과 최근 12개월 총 거래횟수는 거래횟수가 크면 거래대금이 클 확률이 높으나 변수를 삭제하면 소액 위주의 결제 고객의 이탈에 대한 정보를 얻기 힘들다고 판단하여 두 변수 모두 사용하기로 함

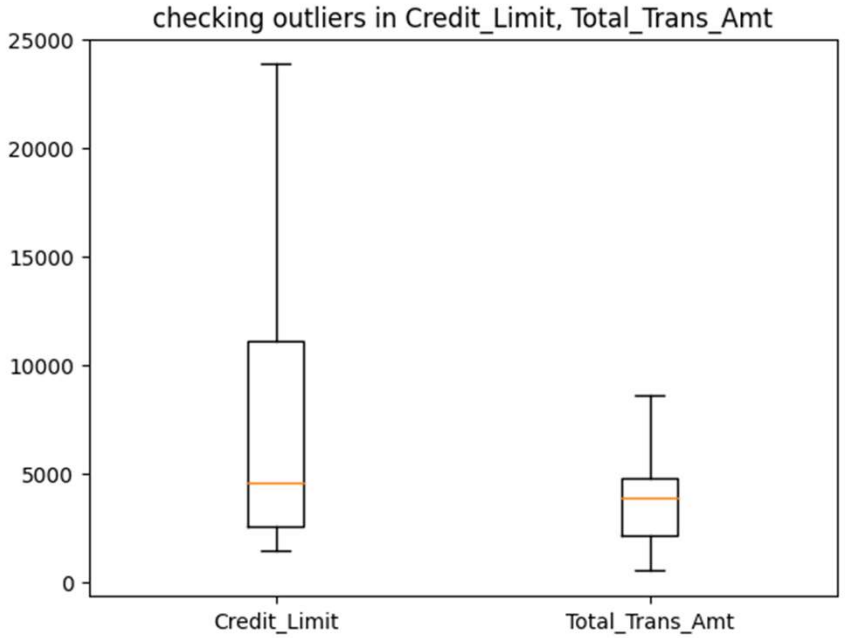
Avg_Open_TO_Buy 변수는 사용하지 않기로 함

이상치 처리

범위가 큰 금액 관련 Outlier 처리하였었으나, 수치적으로 큰 차이가 없음.

새로운 평가 지표를 투입 하여 이상치 처리 없이 진행



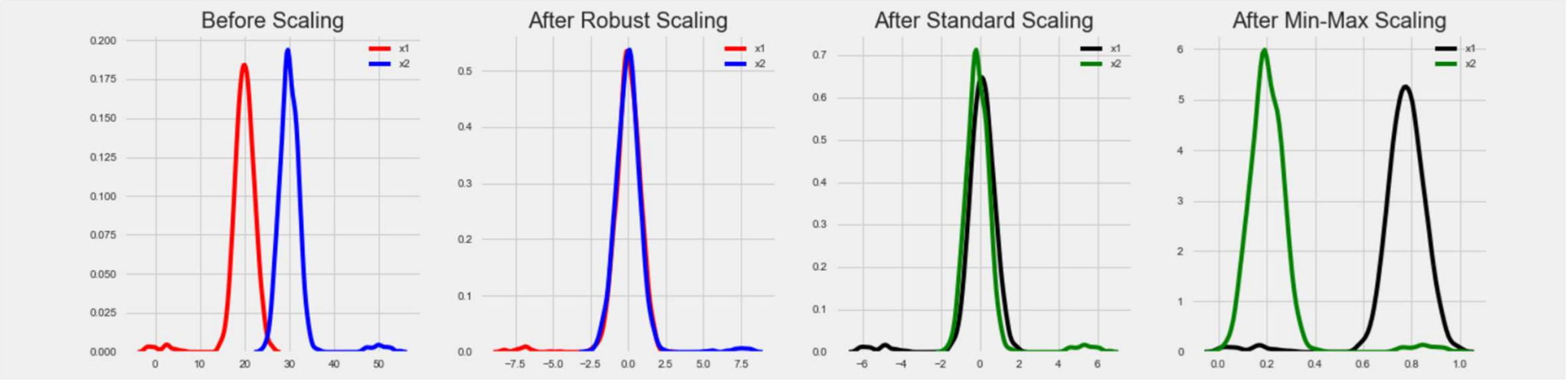


Scaling

StandardScaler : 데이터 평균 0, 표준편차 1, 다양한 변수 균일 조정 유리

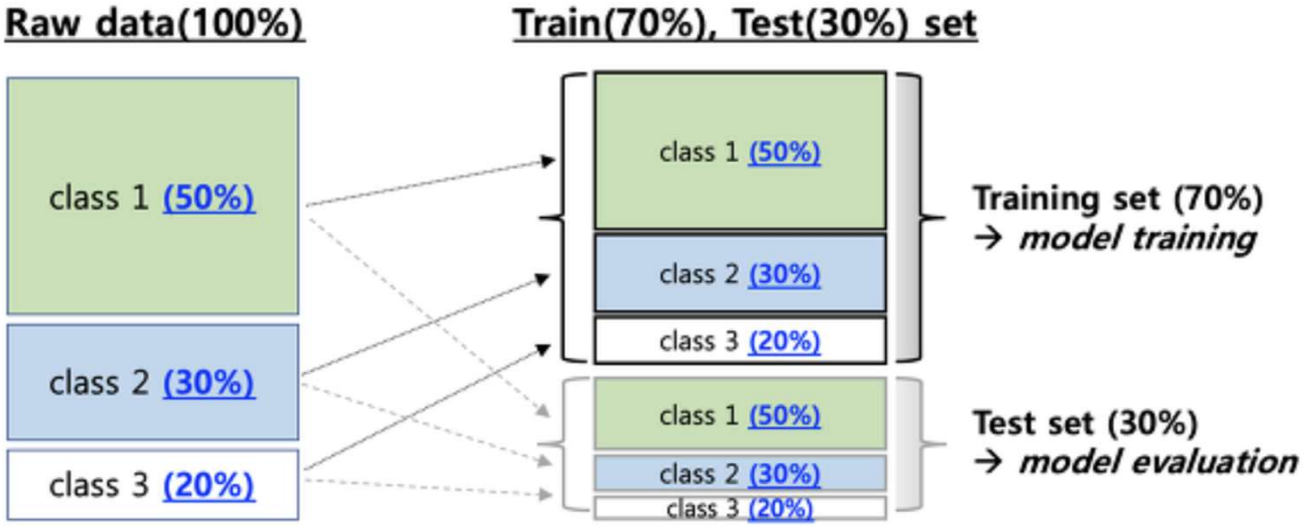
MinMaxScaler : 데이터를 0 ~1, 이상치에 민감

RobustScaler : 중앙값과 사분위 수 사용. Outlier 영향 최소화



Stratify

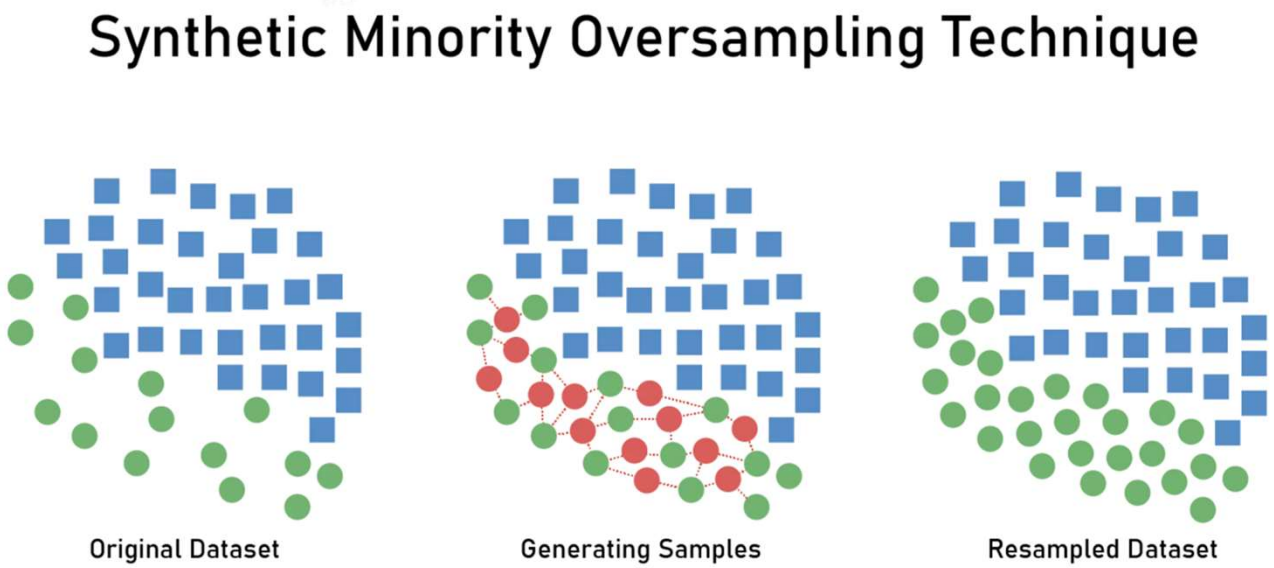
전체 데이터에 대한 편향을 줄이기 위해 하위 그룹별 일부 샘플을 추출하는 stratify 진행



기존 데이터를 나누는 것에 그치지 않고 분포 비율까지 맞추어 Overfitting 방지

SMOTE

데이터 불균형 문제를 해결하기 위해 분류 모델에 따라 클래스를 생성하는 SMOTE 시도



추가 평가 지표 : **Revenue_Risk**

가정 1 : 카드사가 고객으로부터 얻는 순이익은 총 거래 금액에 비례함
가정 2 : 이탈할 것으로 예상되는 고객들을 대상으로 카드사가 순이익의 1/3을 추가 서비스 비용으로 제공함
가정 3 : 추가 서비스를 제공한다면 고객들은 전부 해당 카드를 계속 사용함
가정 4 : 나머지 고려할 수 있는 조건들은 동일함

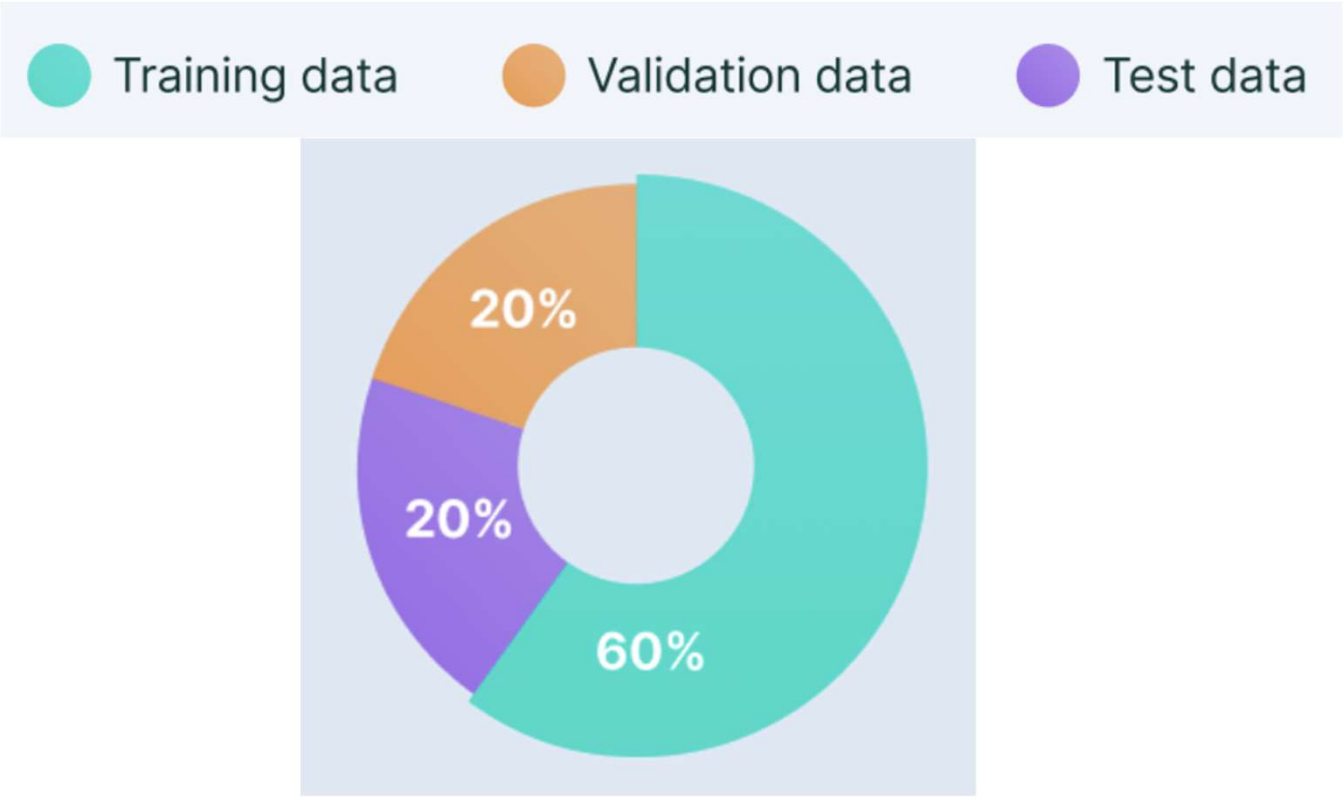
이때 카드사가 추가적인 서비스를 시행함으로써 얻는 리스크 :
FP에 해당하는 사람들에게 추가 서비스를 시행함으로써 불필요한 지출을 하는 것 +
FN에 해당하는 사람들에게 추가 서비스를 시행하지 않음으로써 더 많은 이윤을 얻지 못하는 것

Revenue_Risk = 0.33 * FP[Total_Trans_Amt] + 0.67 * FN[Total_Trans_Amt]

모델링 및 평가

Modeling

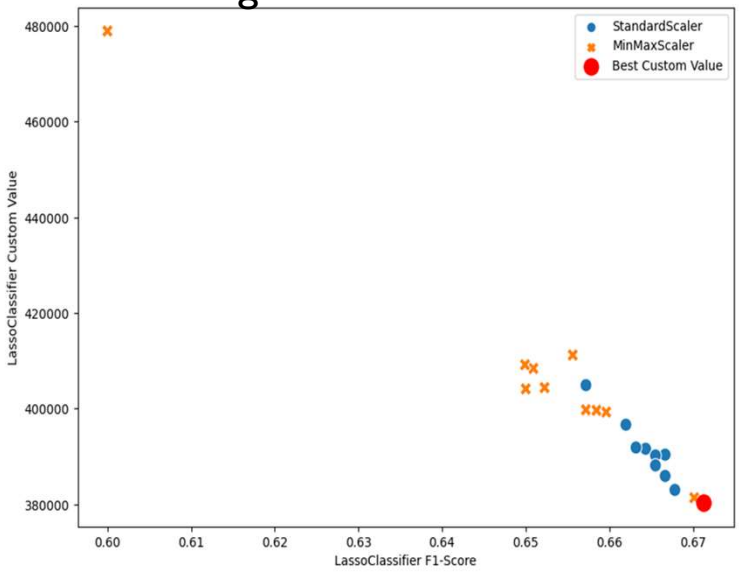
Model



- Lasso
- Ridge
- ElasticNet
- SVM
- KNN
- DecisionTree
- RandomForest
- XGBoost
- LGBM

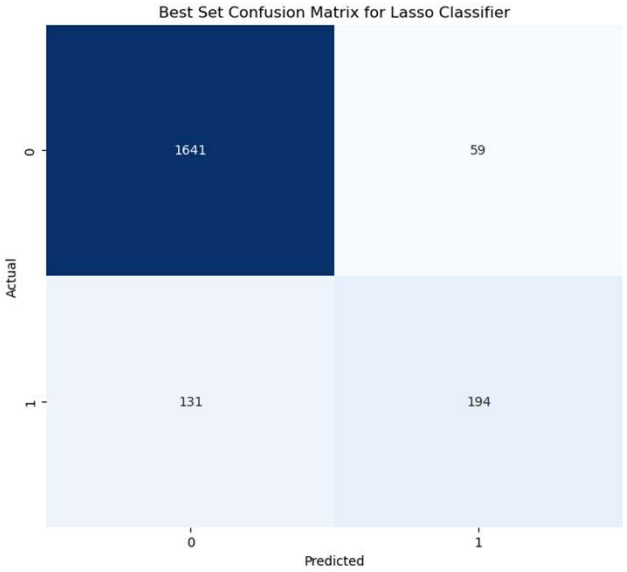
Lasso Classifier - Stratify

HyperParameter 선정



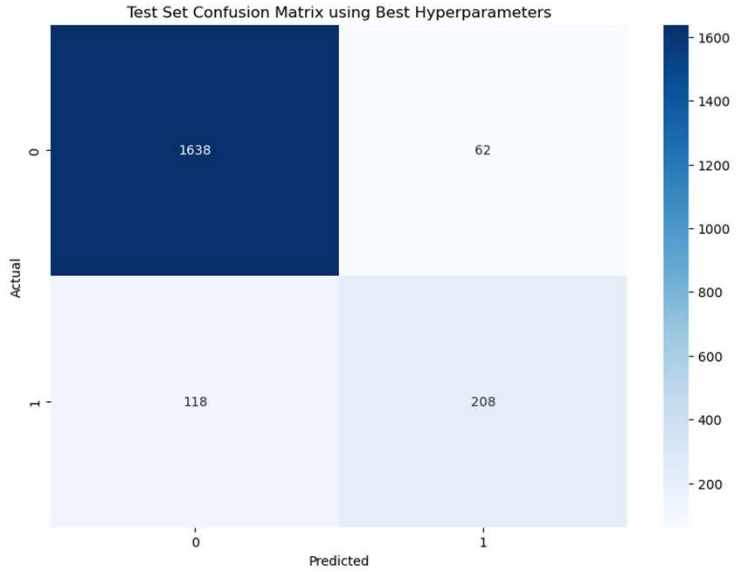
Best hyperparameters:
C : 10.0

Valid Set



Best set Accuracy : 0.9062
Best set F1-Score : 0.6713
Best set Precision : 0.7618
Best set Recall : 0.5969
Best set Custom Value : 380317.37

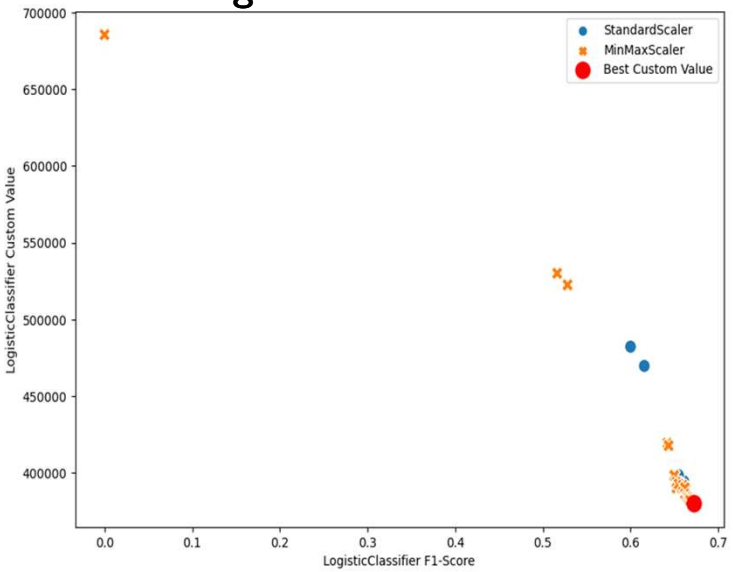
Test Set



Test set Accuracy : 0.9112
Test set F1-Score : 0.6980
Test set Precision : 0.7704
Test set Recall : 0.6380
Test set Custom Value : 359026.31

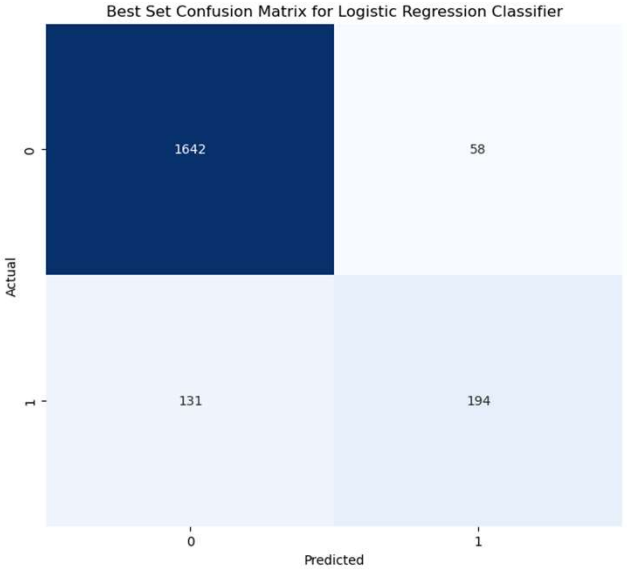
Ridge Classifier - Stratify

HyperParameter 선정



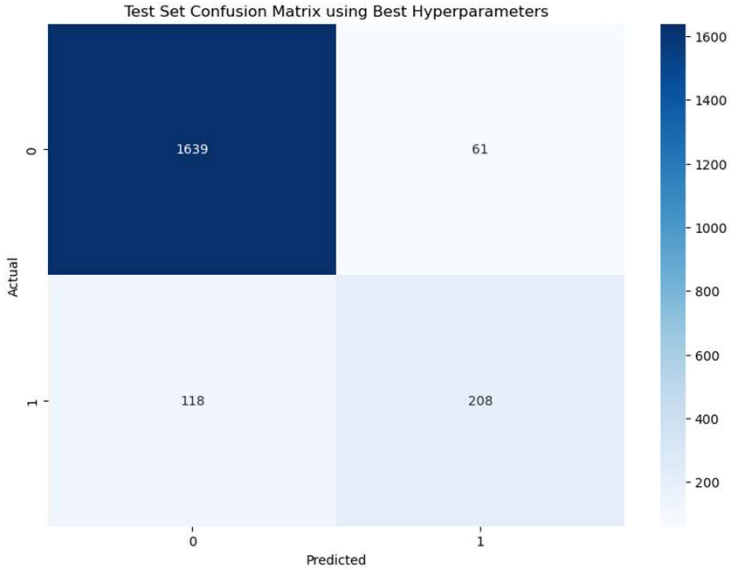
Best hyperparameters:
C : 5.0

Valid Set



Best set Accuracy : 0.9067
Best set F1-Score : 0.6724
Best set Precision : 0.7698
Best set Recall : 0.5969
Best set Custom Value : 379839.53

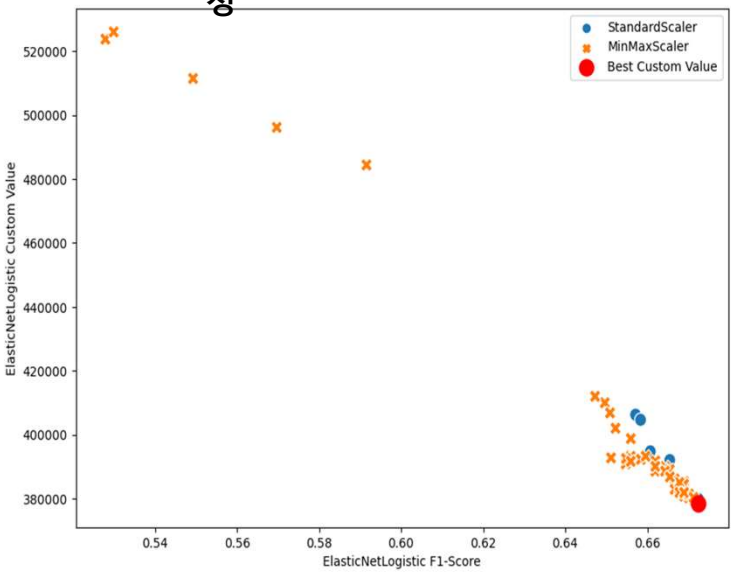
Test Set



Test set Accuracy : 0.9116
Test set F1-Score : 0.6992
Test set Precision : 0.7732
Test set Recall : 0.6380
Test set Custom Value : 358539.23

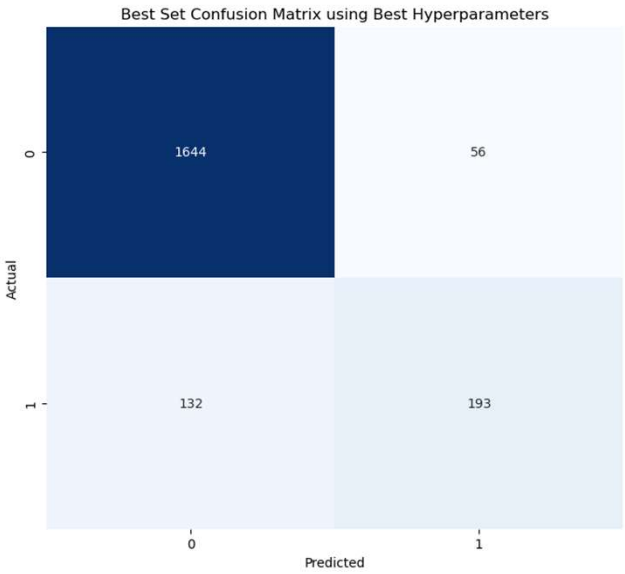
ElasticNet - Stratify

HyperParameter 선정



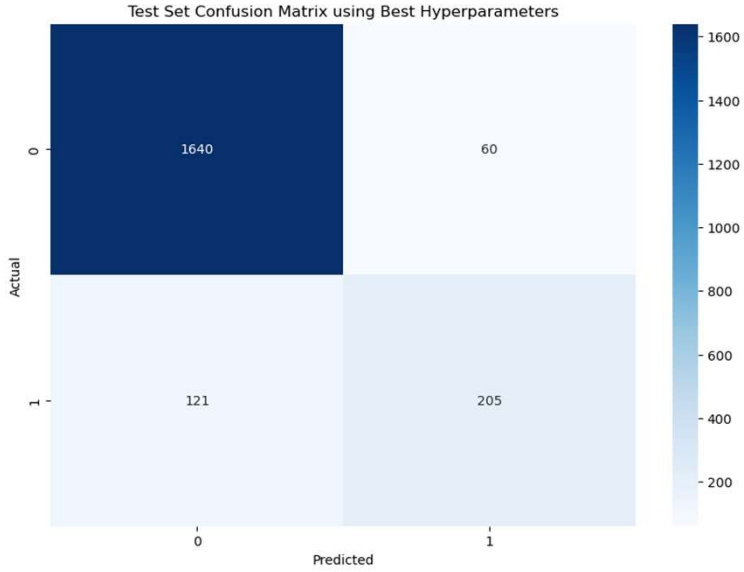
Best hyperparameters:
C : 5.0
l1-ratio : 0.7

Valid Set



Best set Accuracy : 0.9072
Best set F1-Score : 0.6725
Best set Precision : 0.7751
Best set Recall : 0.5938
Best set Custom Value : 378402.22

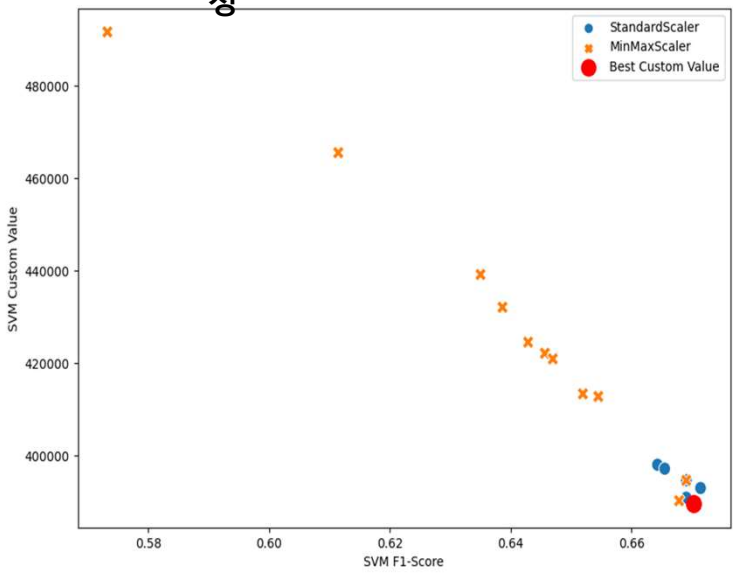
Test Set



Test set Accuracy : 0.9107
Test set F1-Score : 0.6937
Test set Precision : 0.7736
Test set Recall : 0.6288
Test set Custom Value : 367103.11

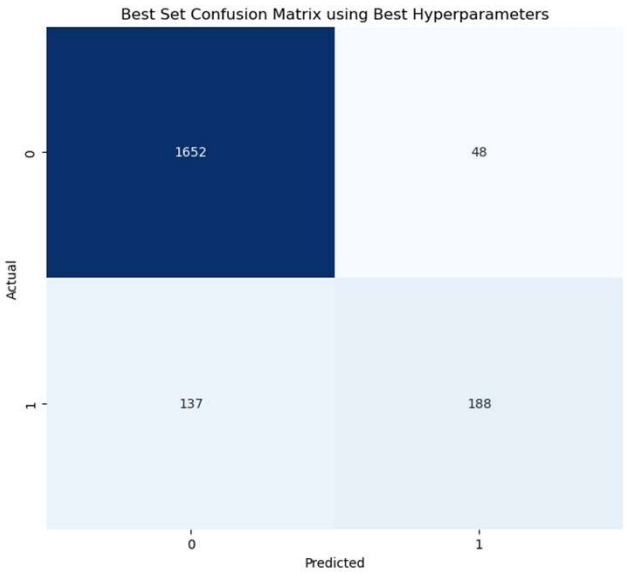
SVM - Stratify

HyperParameter 선정



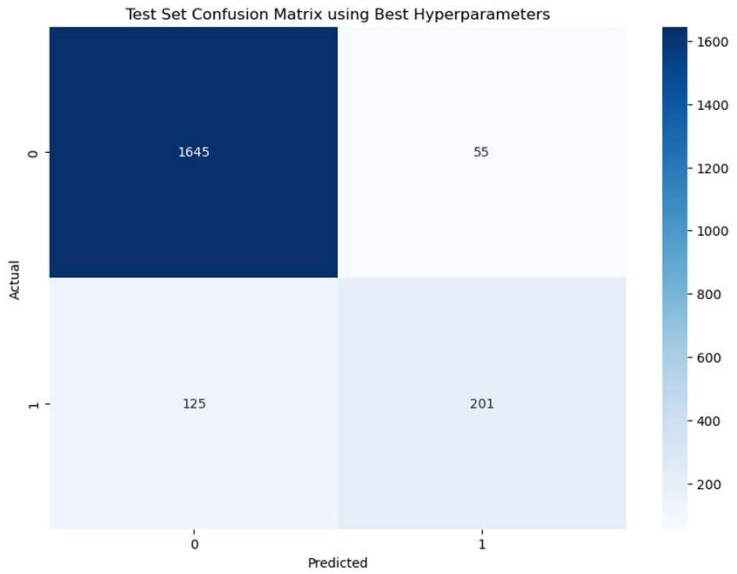
Best hyperparameters:
C : 0.4

Valid Set



Best set Accuracy : 0.9086
Best set F1-Score : 0.6702
Best set Precision : 0.7966
Best set Recall : 0.5785
Best set Custom Value : 389590.29

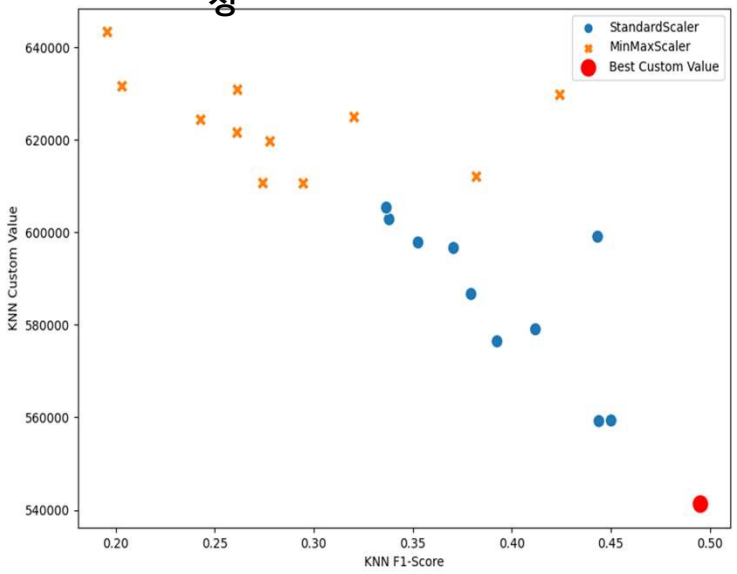
Test Set



Test set Accuracy : 0.9112
Test set F1-Score : 0.6907
Test set Precision : 0.7852
Test set Recall : 0.6155
Test set Custom Value : 369969.83

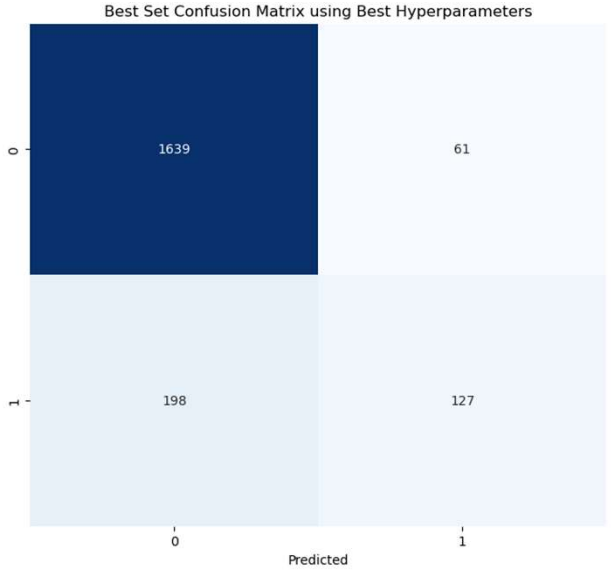
KNN - Stratify

HyperParameter 선정



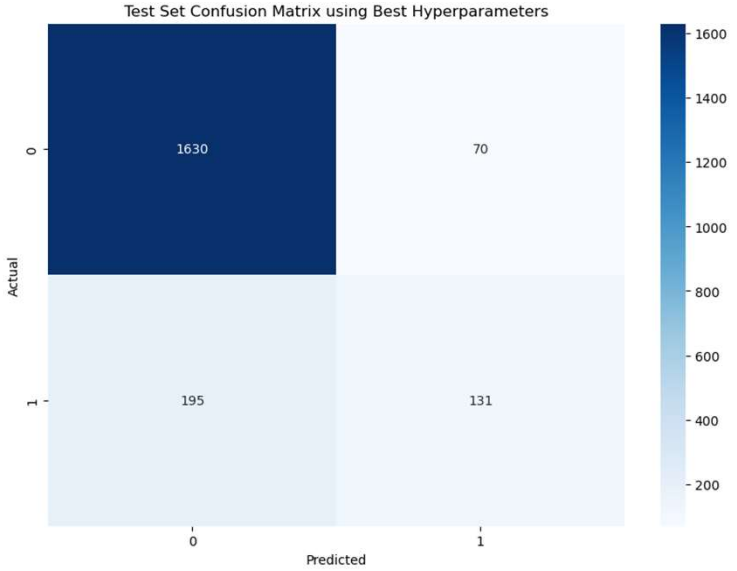
Best hyperparameters:
n_neighbors : 3

Valid Set



Best set Accuracy : 0.8721
Best set F1-Score : 0.4951
Best set Precision : 0.6755
Best set Recall : 0.3908
Best set Custom Value : 541320.10

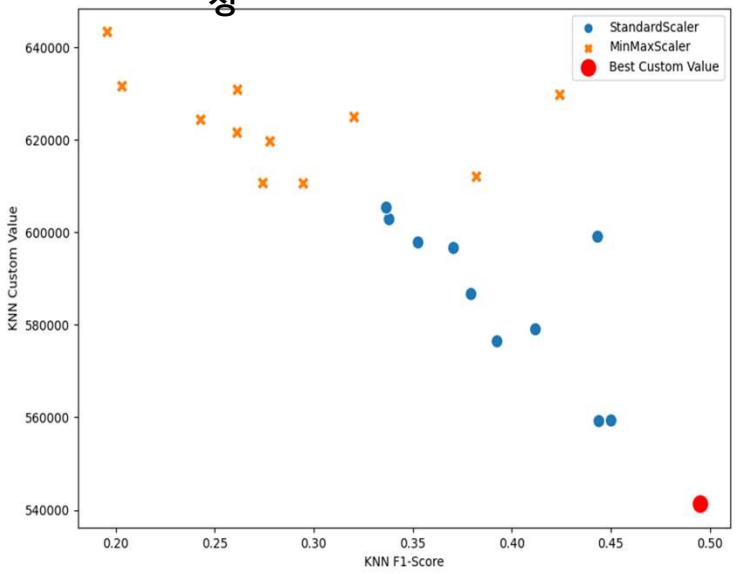
Test Set



Test set Accuracy : 0.8692
Test set F1-Score : 0.4972
Test set Precision : 0.6517
Test set Recall : 0.4018
Test set Custom Value : 530796.47

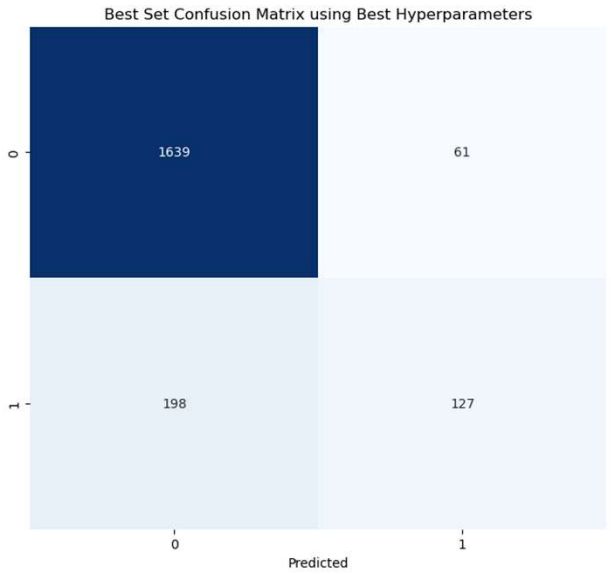
Decision Tree - Stratify

HyperParameter 선정



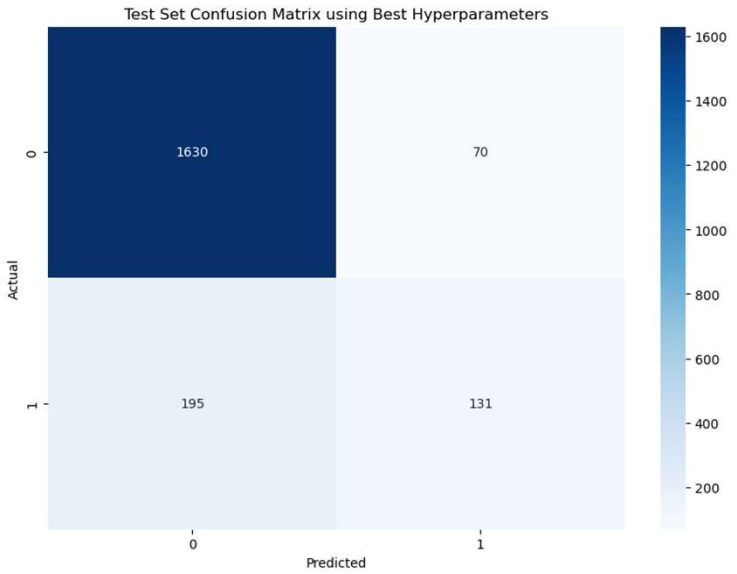
Best hyperparameters:
max_depth : 10
min_samples_split : 5
min_samples_leaf : 1

Valid Set



Best set Accuracy : 0.9353
Best set F1-Score : 0.7956
Best set Precision : 0.8070
Best set Recall : 0.7846
Best set Custom Value : 211809.27

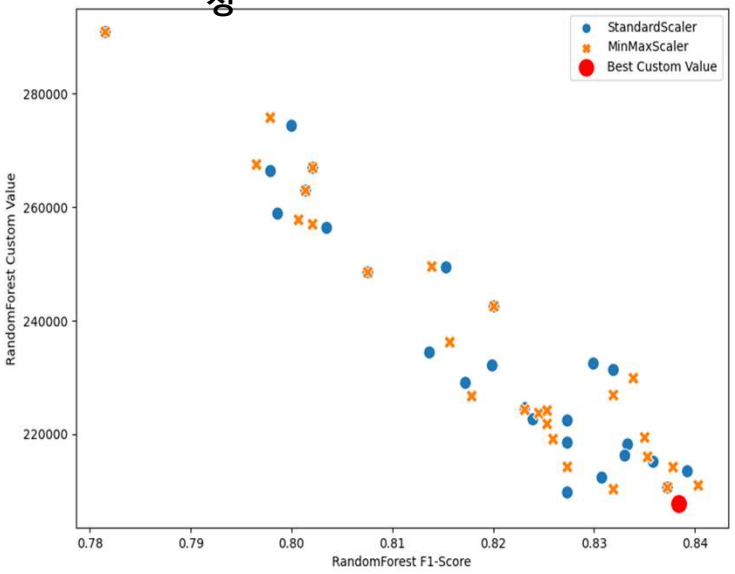
Test Set



Test set Accuracy : 0.9457
Test set F1-Score : 0.8297
Test set Precision : 0.8375
Test set Recall : 0.8221
Test set Custom Value : 174101.75

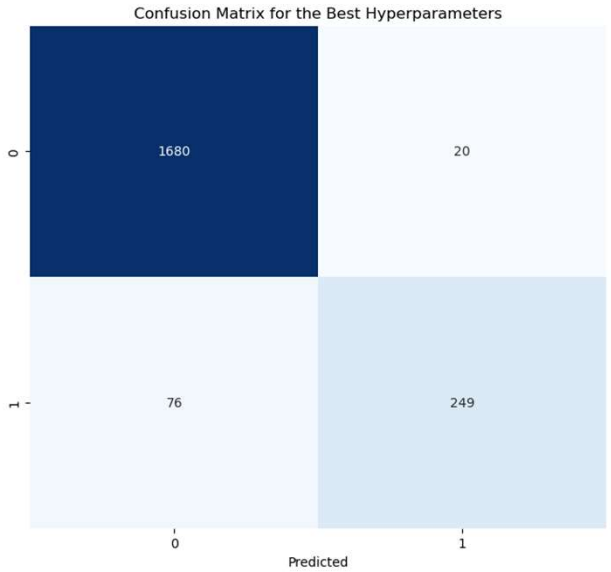
Random Forest - Stratify

HyperParameter 선정



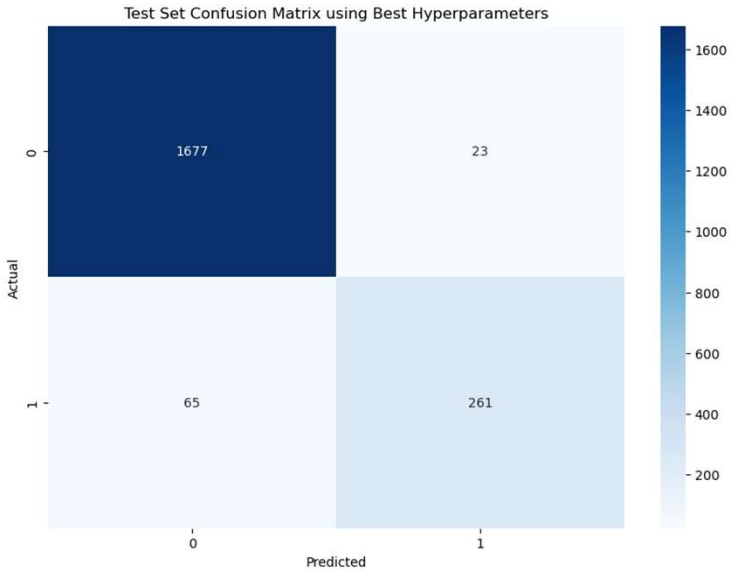
Best hyperparameters:
n_estimators : 200
max_depth : None
min_samples_split : 2

Valid Set



Best set Accuracy : 0.9526
Best set F1-Score : 0.8384
Best set Precision : 0.9257
Best set Recall : 0.7662
Best set Custom Value : 207663.90

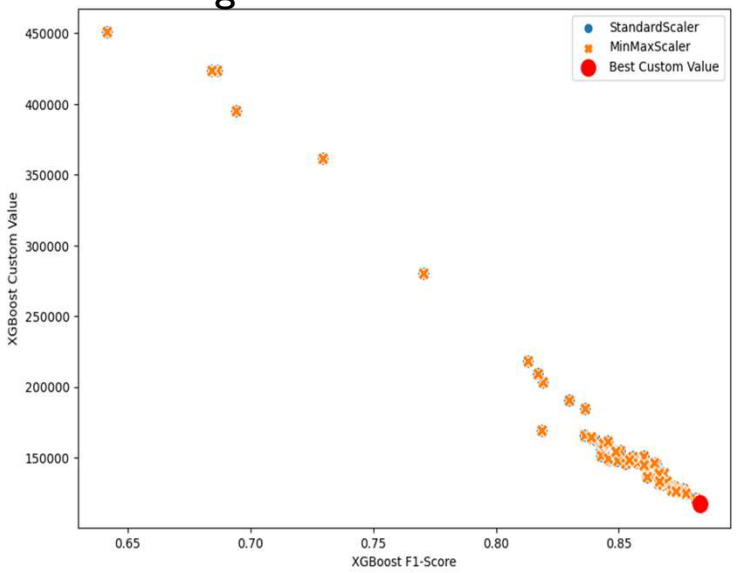
Test Set



Test set Accuracy : 0.9566
Test set F1-Score : 0.8557
Test set Precision : 0.9190
Test set Recall : 0.8006
Test set Custom Value : 204410.01

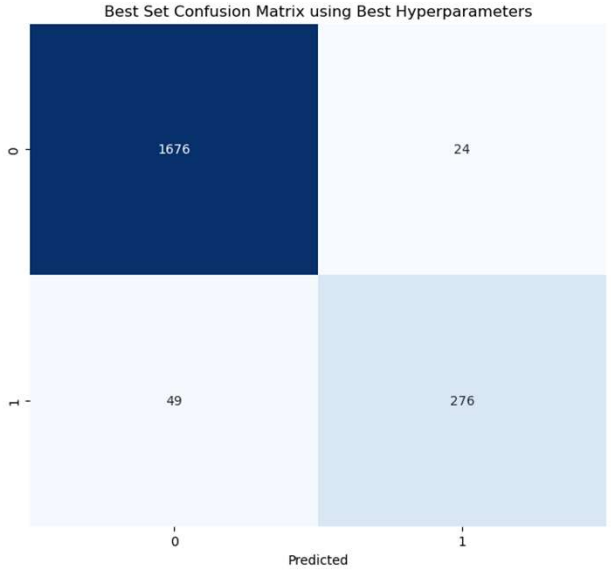
XGBoost - Stratify

HyperParameter 선정



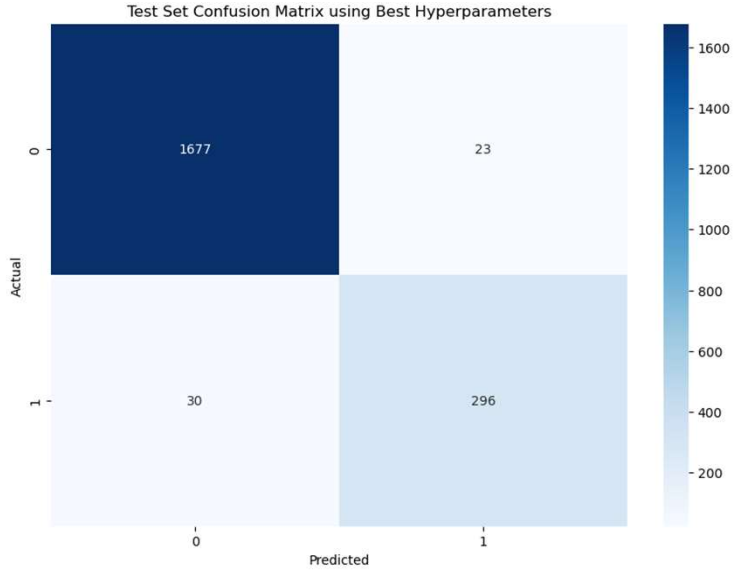
Best hyperparameters:
learning_rate : 0.07
n_estimators : 150
max_depth : 5
gamma : 0

Valid Set



Best set Accuracy : 0.9640
Best set F1-Score : 0.8832
Best set Precision : 0.9200
Best set Recall : 0.8492
Best set Custom Value :117267.17

Test Set

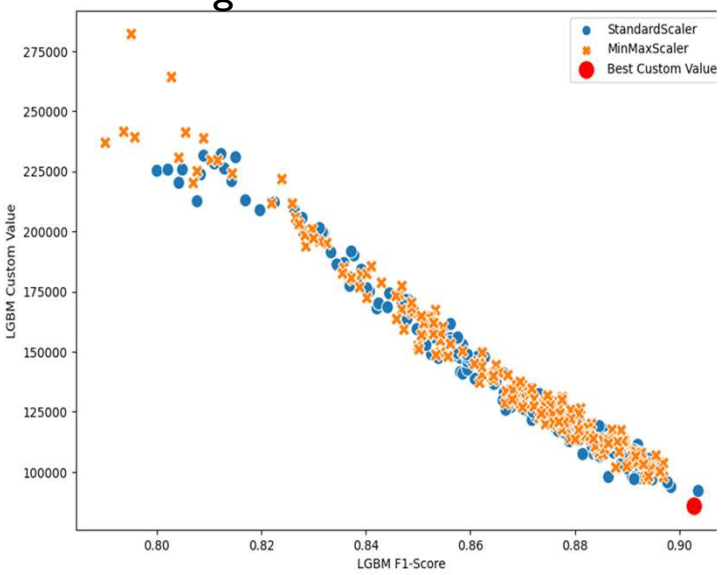


Test set Accuracy : 0.9738
Test set F1-Score : 0.9178
Test set Precision : 0.9279
Test set Recall : 0.9080
Test set Custom Value : 87092.58

LGBM <FeedBack> - Stratify

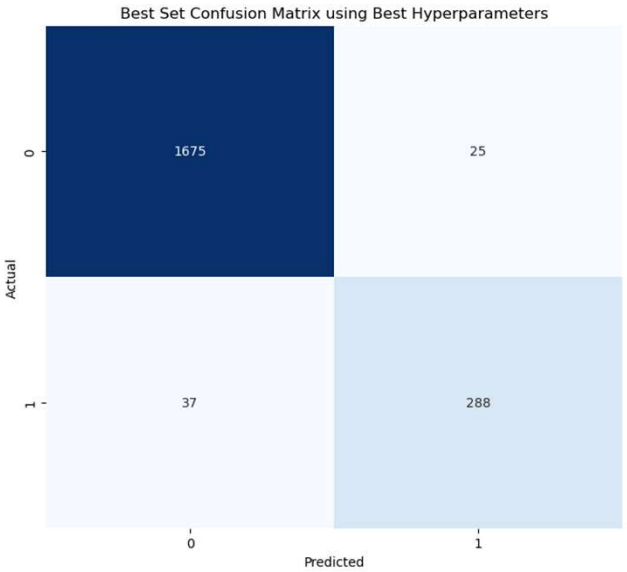
하이퍼 파라미터 조정

HyperParameter 선정



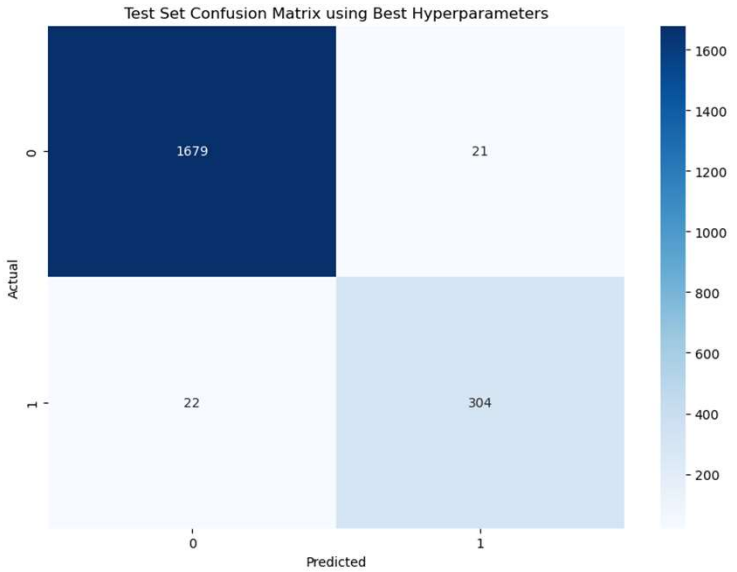
Best hyperparameters:
learning_rate : 0.1
n_estimators : 200
max_depth : 7
num_leaves : 14
min_data_in_leaf : 20

Valid Set



Best set Accuracy : 0.9694
Best set F1-Score : 0.9028
Best set Precision : 0.9201
Best set Recall : 0.8862
Best set Custom Value : 85816.94

Test Set



Test set Accuracy : 0.9778
Test set F1-Score : 0.9339
Test set Precision : 0.9354
Test set Recall : 0.9325
Test set Custom Value : 66446.02

평가 지표 확인(Stratify)

Model	Accuracy	F1-Score	Precision	Recall	Revenue_Risk
Logistic	0.9111	0.6979	0.7703	0.6380	359026.31
Lasso	0.9111	0.6979	0.7703	0.6380	359026.31
Ridge	0.9111	0.6991	0.7732	0.6380	358539.23
Elastic	0.9106	0.6937	0.7735	0.6288	367103.11
SVM	0.9111	0.6907	0.7851	0.6165	369969.83
KNN	0.8692	0.4971	0.6517	0.4018	530796.47
DecisionTree	0.9457	0.8297	0.8375	0.8220	174101.75
RandomForest	0.9565	0.8557	0.9190	0.8006	204410.01
XGBoost	0.9738	0.9178	0.9279	0.9079	87092.58
LGBM Feedback	**0.9787**	**0.9339**	**0.9353**	**0.9325**	**66446.02**

평가 지표 확인(SMOTE)

Model	Accuracy	F1-Score	Precision	Recall	Revenue_Risk
Logistic	0.8771	0.6103	0.6230	0.5982	485522.47
Lasso	0.8801	0.6173	0.6343	0.6012	460658.52
Ridge	0.8771	0.6103	0.6230	0.5982	485522.47
Elastic	0.8776	0.6113	0.6250	0.5982	483857.62
SVM	0.8766	0.6044	0.6242	0.5859	489353.31
KNN	0.8371	0.4795	0.4935	0.4663	591463.85
DecisionTree	0.9299	0.7936	0.7541	0.8374	199723.64
RandomForest	0.9580	0.8670	0.8850	0.8497	159218.01
XGBoost	0.9689	**0.9032**	**0.9046**	0.9018	88648.58
LGBM	**0.9684**	0.9027	0.8946	**0.9110**	**86806.76**

결론 및 해석

Data

모델 성능 비교 (Feedback 이전)

Model	Accuracy	F1-Score	Precision	Recall	Revenue_Risk
LGBM (Stratify)	**0.9763**	**0.9241**	**0.9542**	0.8957	**75398.73**
XGBoost (Stratify)	0.9738	0.9178	0.9279	0.9079	87092.58
LGBM (SMOTE)	0.9689	0.9032	0.9046	0.9018	88648.58
XGBoost (6기)	0.9294	0.7974	0.7131	0.9035	-
LGBM (6기)	0.9368	0.8172	0.7366	**0.9164**	-

- **우수 예측 모델 비교** : LGBM (stratify) vs LGBM(6기)
- Accuracy : 0.039
- F1-Score : 0.107
- Precision : 0.217
- **추가 평가지표 :** Revenue_Risk 산출 결과 LGBM(stratify) 가장 우수

모델 성능 비교 (Feedback 이후)

Model	Accuracy	F1-Score	Precision	Recall	Revenue_Risk
LGBM(stratify) Feedback	**0.9787**	**0.9339**	**0.9353**	**0.9325**	**66446.02**
XGBoost (Stratify)	0.9738	0.9178	0.9279	0.9079	87092.58
LGBM (SMOTE)	0.9689	0.9032	0.9046	0.9018	88648.58
XGBoost (6기)	0.9294	0.7974	0.7131	0.9035	-
LGBM (6기)	0.9368	0.8172	0.7366	0.9164	-

- **전년도와 우수 예측 모델 비교** : LGBM(stratify) Feedback vs LGBM(6기)
- Accuracy : 0.0419
- F1-Score : 0.1167
- Precision : 0.1987
- Recall : 0.0161
- LGBM(6기) 모델보다 더 높은 Recall을 갖는 모델을 하이퍼 파라미터를 조정하여 Feedback 이후 기존에 더 낮았던 Recall이 낮은 문제를 해결

결론 및 해석

Model	Accuracy	F1-Score	Precision	Recall	Revenue_Risk
LGBM(stratify) Feedback	**0.9787**	**0.9339**	**0.9353**	**0.9325**	**66446.02**
XGBoost (Stratify)	0.9738	0.9178	0.9279	0.9079	87092.58
LGBM (SMOTE)	0.9689	0.9032	0.9046	0.9018	88648.58
XGBoost (6기)	0.9294	0.7974	0.7131	0.9035	-
LGBM (6기)	0.9368	0.8172	0.7366	0.9164	-

- 전체적으로 모델의 성능을 향상
  - 이탈 고객에 대한 예측의 정확도가 높아짐
- **추가 평가지표 :** Revenue_Risk를 도입함
  - 추가적인 서비스를 지급하는 비용과 이탈한 고객에 의해 줄어들 이익금을 반영한 수치로 계산
  - 단순한 정확도보다 기업의 이익을 파악하는데 도움이 될거라 생각됨

고객 이탈 방지 전략

- 카드사용의 거래횟수에 따라 ‘Attrition Flag’와의 상관계수 변화

  → 거래 횟수가 증가할 수록 이탈 확률 이 감소

- 'Months_Inactive_12_mon’(최근 12개월간 카드 거래가 일어나지 않은 달의 수)가 높을 수록

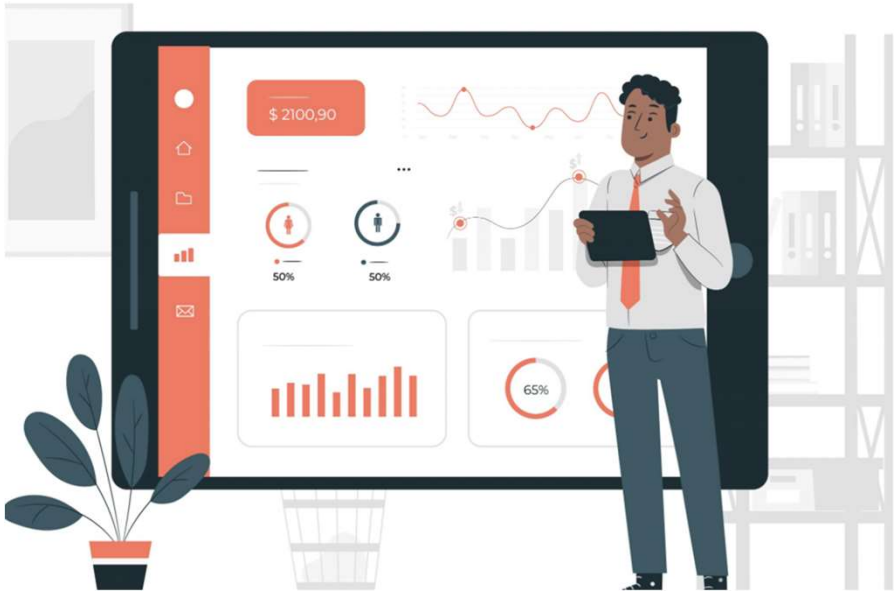
  → ‘Attrition Flag’와의 상관계수 변화

  → 거래빈도가 낮을 수록 이탈 확률 증가

- Card category에 따라 등급이 높을 수록 이탈 확률이 증가함

  → Blue 16% , Gold 18%, Silver 14%, Platinum 25%

고객 이탈 방지 전략



고객 Segment별 맞춤형 마케팅 전략 수립

- 반응변수(Attrition_Flag)인 서비스 해지율과 상관계수가 높은
  최근 12개월 간 무거래 고객(Months_Interactive_12_mon)을 대상으로
  카드 사용을 유도할 수 있는 다양한 서비스 마련
  ex) 무거래 고객 대상 Target - 이메일, TM 마케팅 진행
  실적 충족 시 할인 / 캐시백 서비스 제공, 무이자 할부 서비스 등

- Platinum 타입 고객의 서비스 해지율이 타 등급 대비 가장 높은 비율(25.00%)임을
  고려하여 해당 그룹의 이탈을 방지할 수 있는 차별화된 혜택 제공
  ex) 등급이 높을 수록 등급별 차별화된 프리미엄 서비스 및 멤버십 혜택 제공