

B.a.f Winter Vacation Team Project

이상거래탐지


Creditcard Fraud Detection

team 3

최솔 변지형 신수빈 박지현



CONTENTS



01 데이터 소개
Data Introduction

02 데이터 탐색 및 전처리
EDA & Data Pre-Processing

03 변수 선택 및 모델링
Feature Engineering
& Modeling

04 모델 평가 및 한계점
Model Evaluation & Limits



01

데이터 소개

Data Introduction

데이터 소개

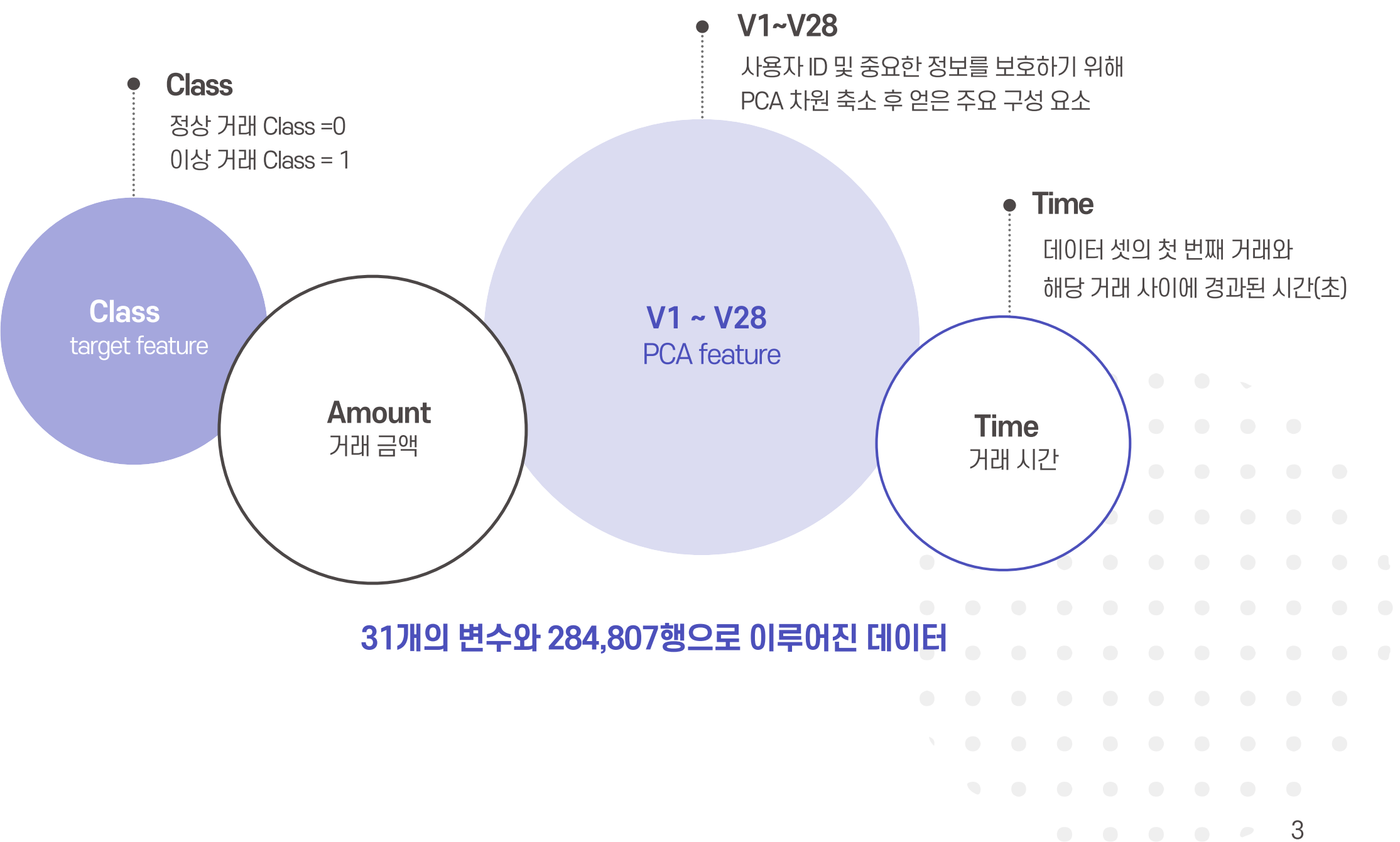
Data Introduction

분석 배경

카드사들은 사기성 신용카드 거래를 인지하여
고객이 구매하지 않은 품목에 대한 비용을 청구하지 않는 것이 중요

데이터 설명

2013년 9월 유럽 카드 소유자들의 신용카드 거래
이틀 동안 발생한 거래를 나타내며, 284,807개의 거래중
492개의 부정 거래 존재 : 매우 불균형한 데이터



02

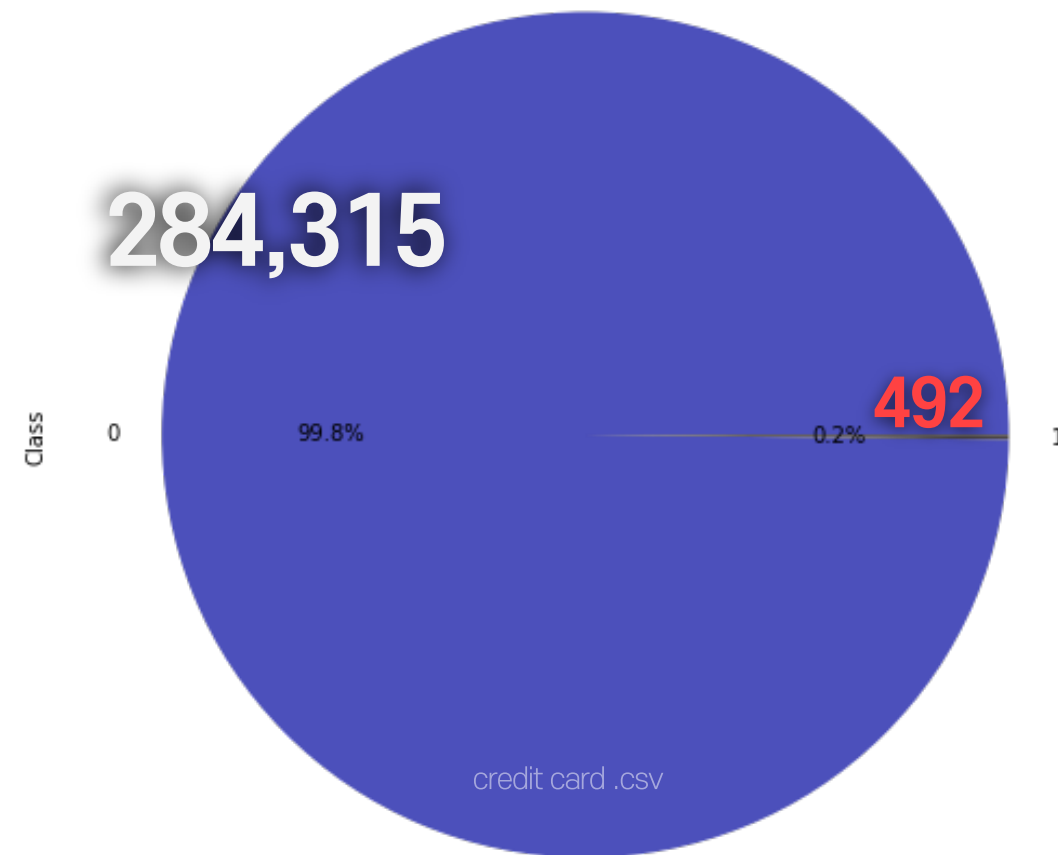
데이터 탐색 및 전처리

EDA & Data Pre-Processing

데이터 탐색

EDA

01 불균형한 데이터



Class 변수의 불균형 → 샘플링 뒤 전처리 진행

02 신용카드 결제 금액이 \$0인 1,825개의 데이터

전체 데이터에서 0.6% 차지, 오류 의심

- 카드 번호 확인 서비스
: 카드가 정상인지 확인하기 위해 \$0의 요금 부과
- \$0.01 미만의 소액 거래
: \$0.01, \$0.02와 같은 0.01단위의 거래 존재하므로
\$0.01 미만의 금액은 \$0으로 표시될 가능성 존재
- 할인된 금액 / 이벤트 응모성 결제

Amount = 0 데이터 → 유의미하다고 판단

데이터 탐색

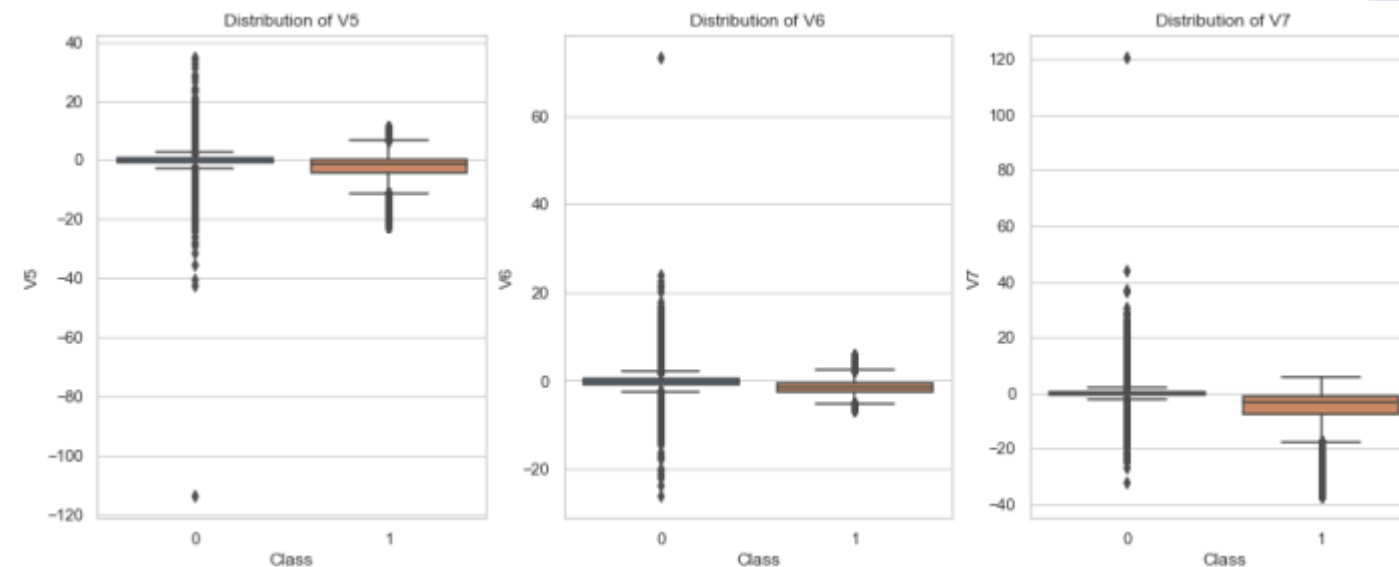
EDA

03 결측치 없음

```
df.isnull().any()
```

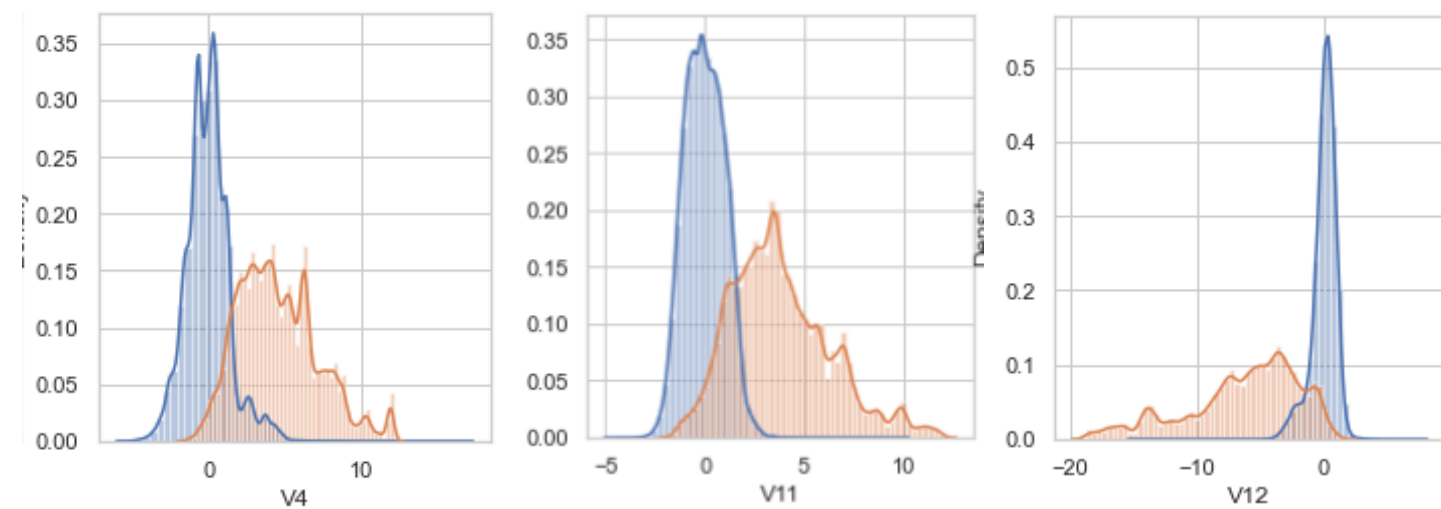
Class	False
Time	False
Amount	False
V1	False
V2	False
V3	False
V4	False
V5	False
V6	False
V7	False
V8	False
V9	False
V10	False
V11	False
V12	False
V13	False
V14	False
V15	False
V16	False
V17	False
V18	False
V19	False
V20	False
V21	False
V22	False
V23	False
V24	False
V25	False
V26	False
V27	False
V28	False

04 샘플링 후 변수 시각화



V5, V6, V7
: 극심한 이상치 존재

V4, V11, V12
: Class별로 다른 분포를 보임



데이터 전처리

Sampling

01 train / test set 분할

전체 데이터의 20%를 test set 으로 지정
train set : 227,84 5행 / test set : 56,962 행

02 1,251개의 중복 데이터 처리

31개 변수의 값이 모두 일치하는 데이터 존재
무의미한 데이터라 판단 → 처음 행 보존, 중복 행 717개 제거

03 Sampling 방법 선택

| SMOTE(Synthetic Minority Over-sampling Technique)

오버샘플링 방법

소수 클래스의 샘플을 가져온 뒤 임의의 값을 추가하여 새로운 샘플을 데이터에 추가하는 방식

.....● Class 0 : 226758
Class 1 : 370 → 226758

| SMOTETomek

복합 샘플링 방법으로 SMOTE와 Tomek's link를 결합한 방법

* Tomek's link: 서로 다른 클래스가 있을 때 서로 다른 클래스끼리 가장

가까운 데이터들이 토멕링크로 묶여서 토멕링크 중 분포가 높은 데이터를 제거

.....● Class 0 : 226758 → 226105
Class 1 : 370 → 226105



| SMOTEENN

복합 샘플링 방법으로 SMOTE와 ENN을 결합한 방식

* ENN(Edited Nearest Neighbours) : 소수 클래스 주변의 다중 클래스 값 제거

.....● Class 0 : 226758 → 217858
Class 1 : 370 → 208253

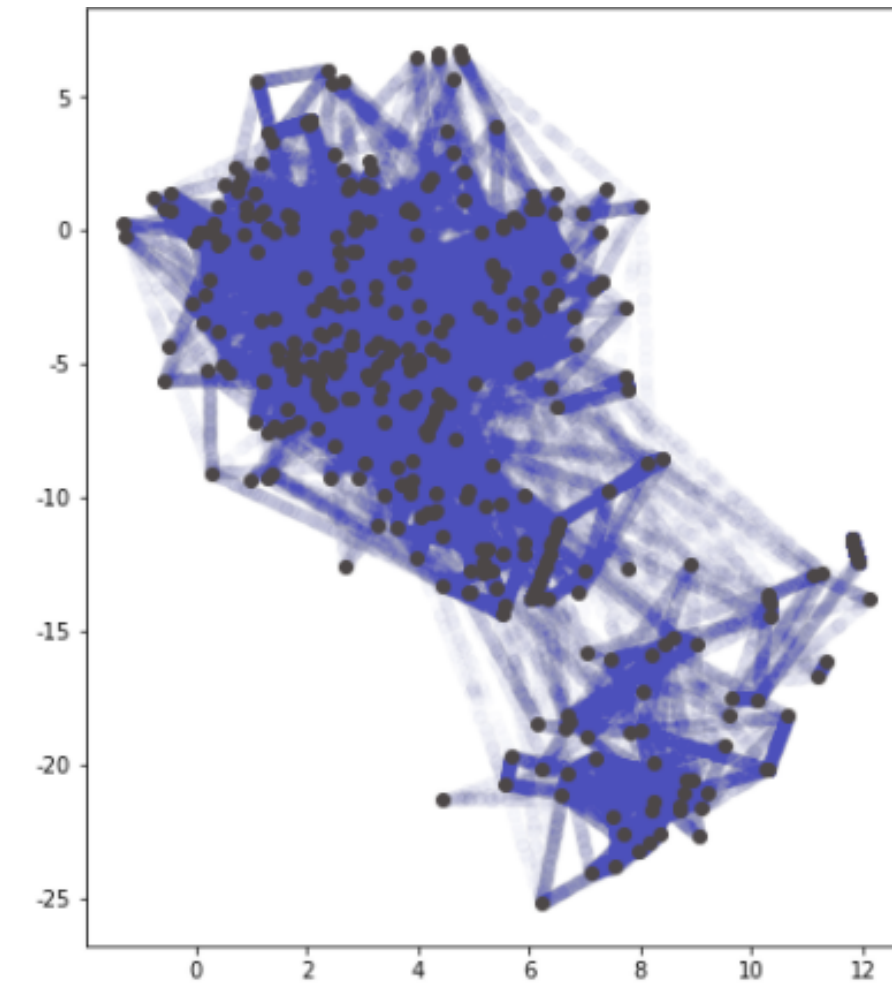
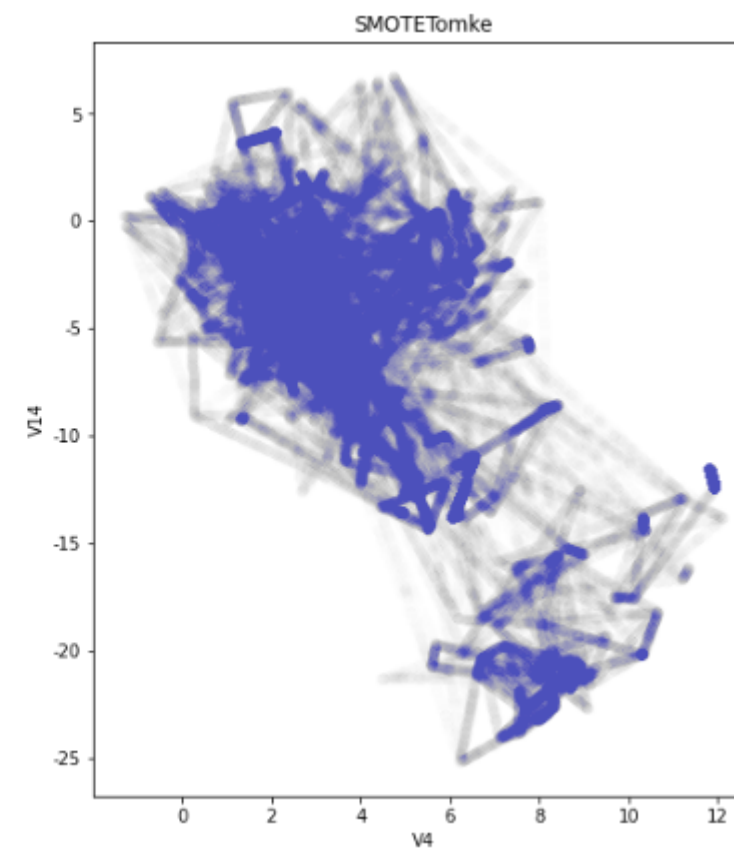
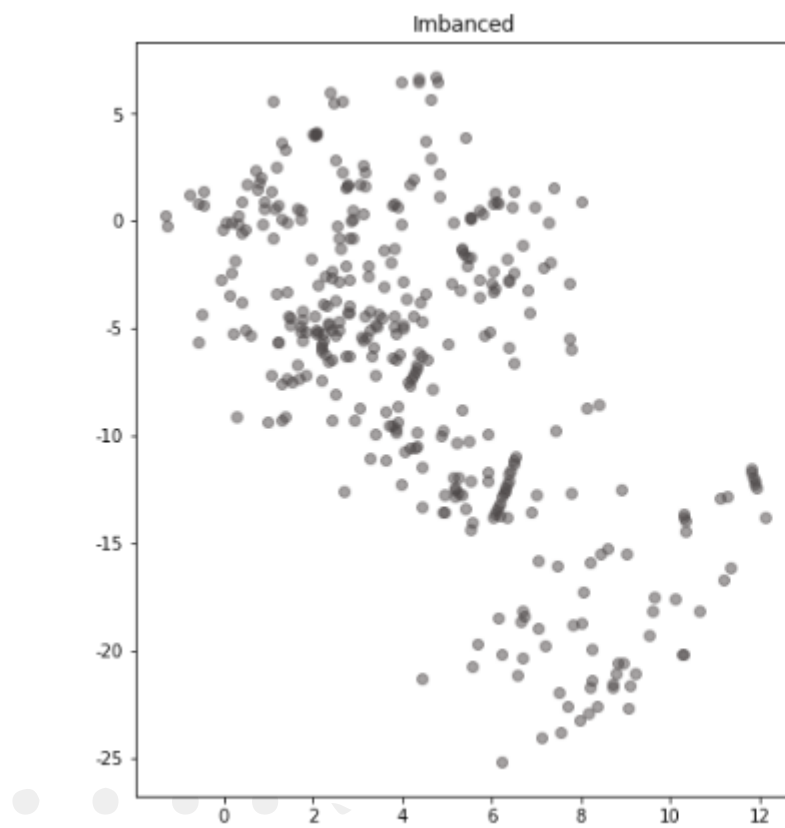
데이터 전처리

SMOTETomek

Class 0 : 226758 → 226105

Class 1 : 370 → 226105

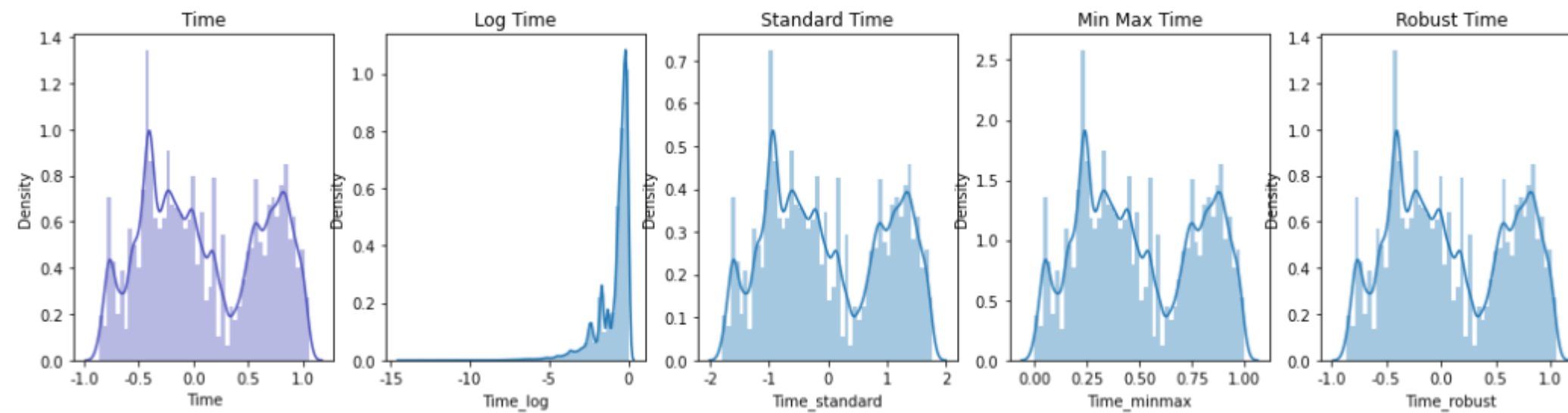
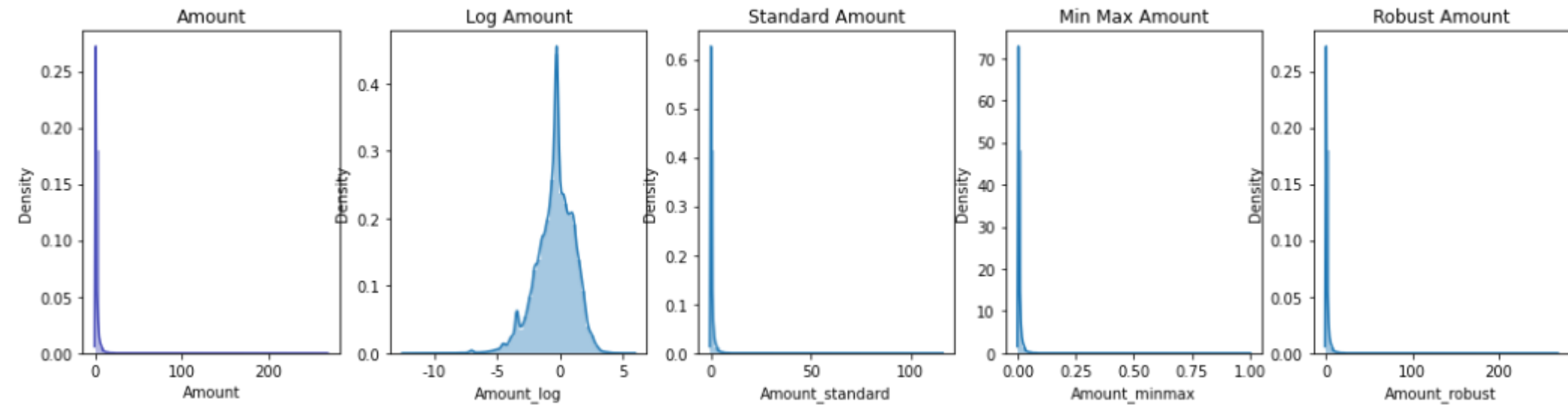
Class = 1의 분포 변화



데이터 전처리

Scaling

PCA 처리된 변수 V1~V28에 비해 Amount와 Time은 데이터 분포가 크며, 특히 Amount는 0에 편향된 비대칭 데이터
→ 스케일링을 통해 원본 데이터 변환



Robust Scaler 선택

: 중앙값과 분위수를 사용하여 이상치에 크게 영향 받지 않음

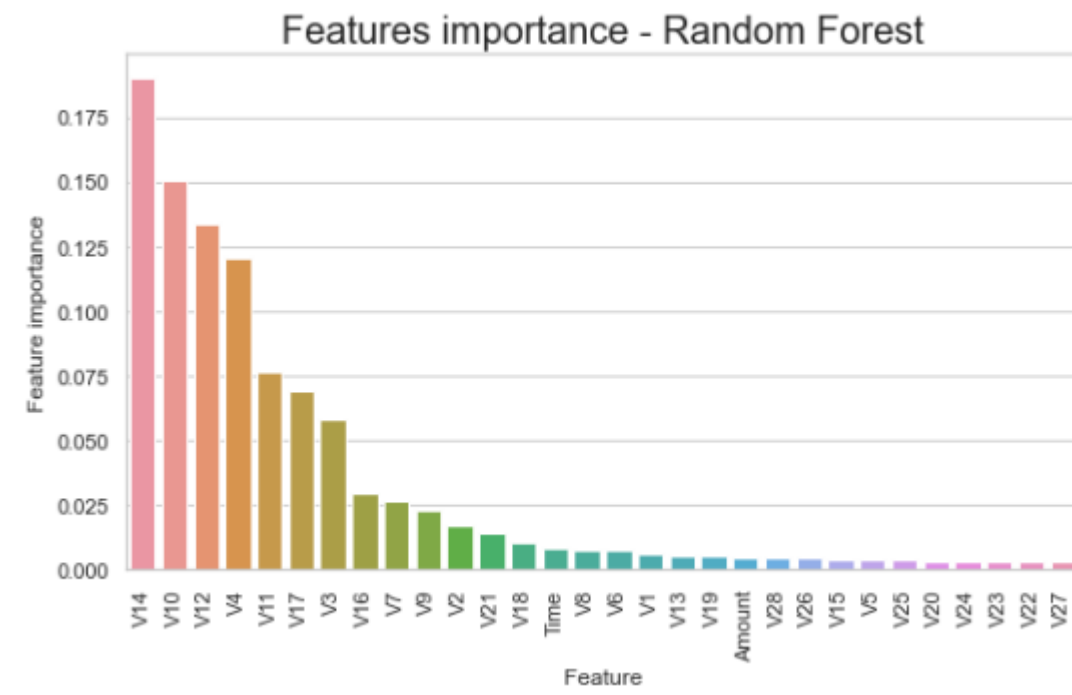
데이터 전처리

이상치 제거

	문제 상황	최종 선택
01 변수 선택	상관관계 높은 변수 타겟 변수와 상관관계 높은 4개 변수	변수 중요도 높은 변수 모델 별 변수 중요도 높은 4개 변수
02 처리 순서	샘플링 전 이상치 제거 후 샘플링으로 성능 저하	샘플링 후 샘플링으로 저하된 성능 이상치 제거로 개선
03 처리 방법	Z-score / Winsorization 수정된 Z-score/ (1,99%), (5,95%)	IQR IQR*3 기준으로 이상치 확인
04 삭제 / 대체	삭제 데이터 손실 우려	대체 최소값과 극소값으로 대체

데이터 전처리

이상치 제거

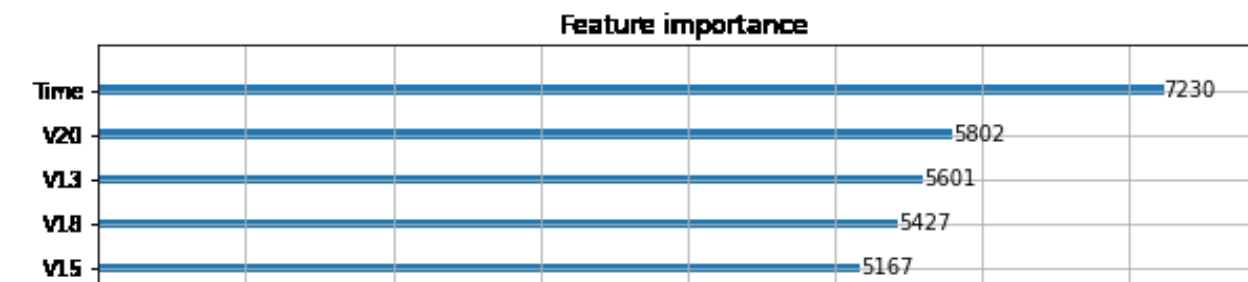
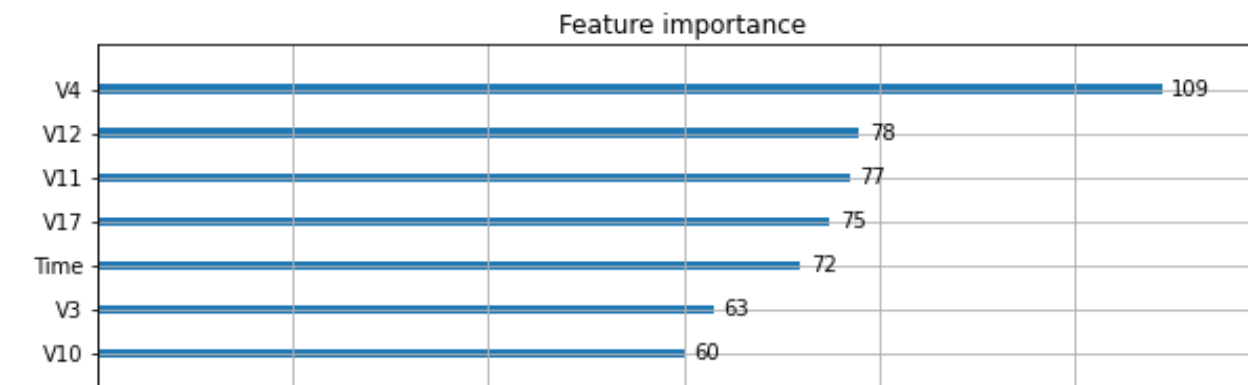
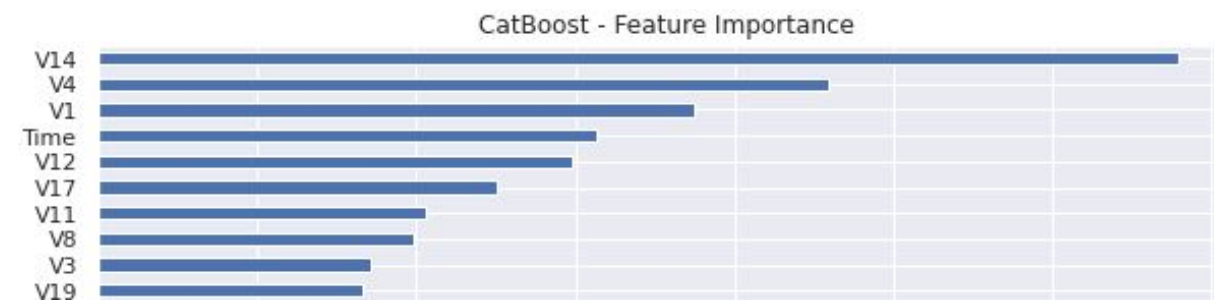


| Random Forest : V4, V10, V12, V14

| CatBoost : V1, V4, V12, V14

| XGBoost : V4, V11, V12, V17

| LGBM : V13, V15, V18, V20



.....● CatBoost

.....● XGBoost

.....● LGBM

데이터 전처리

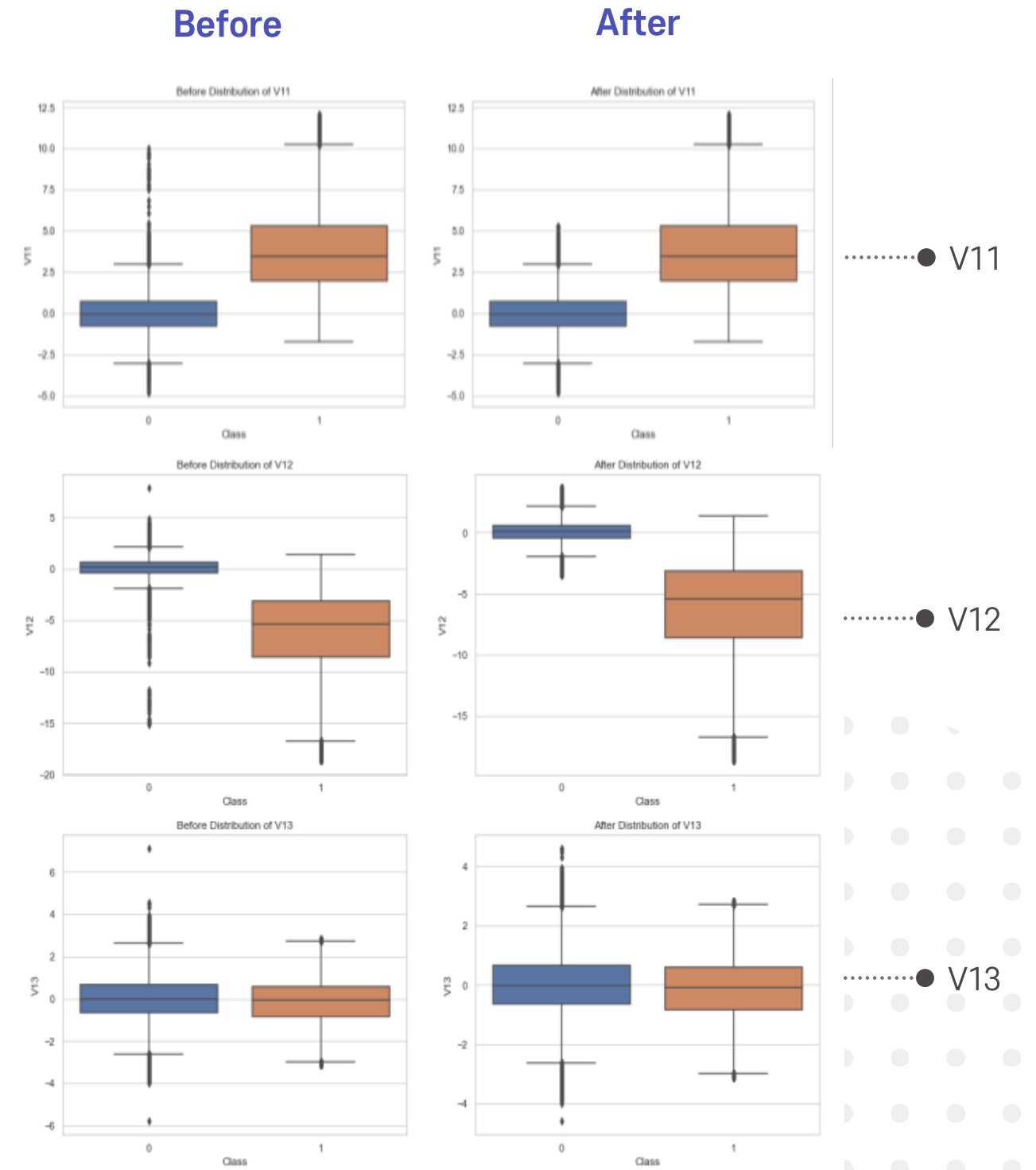
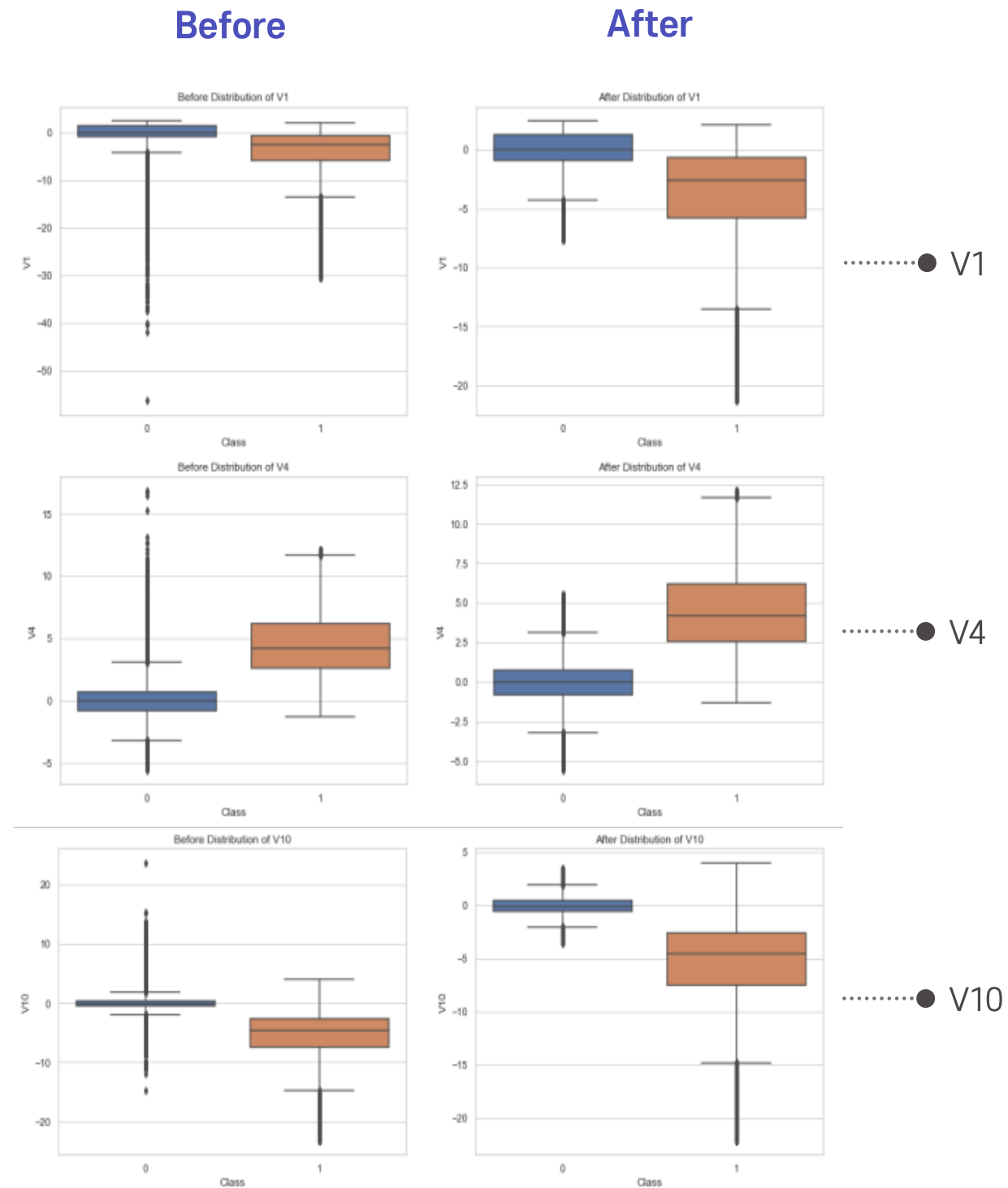
이상치 제거

| Random Forest : V4, V10, V12, V14

| CatBoost : V1, V4, V12, V14

| XGBoost : V4, V11, V12, V17

| LGBM : V13, V15, V18, V20



데이터 전처리

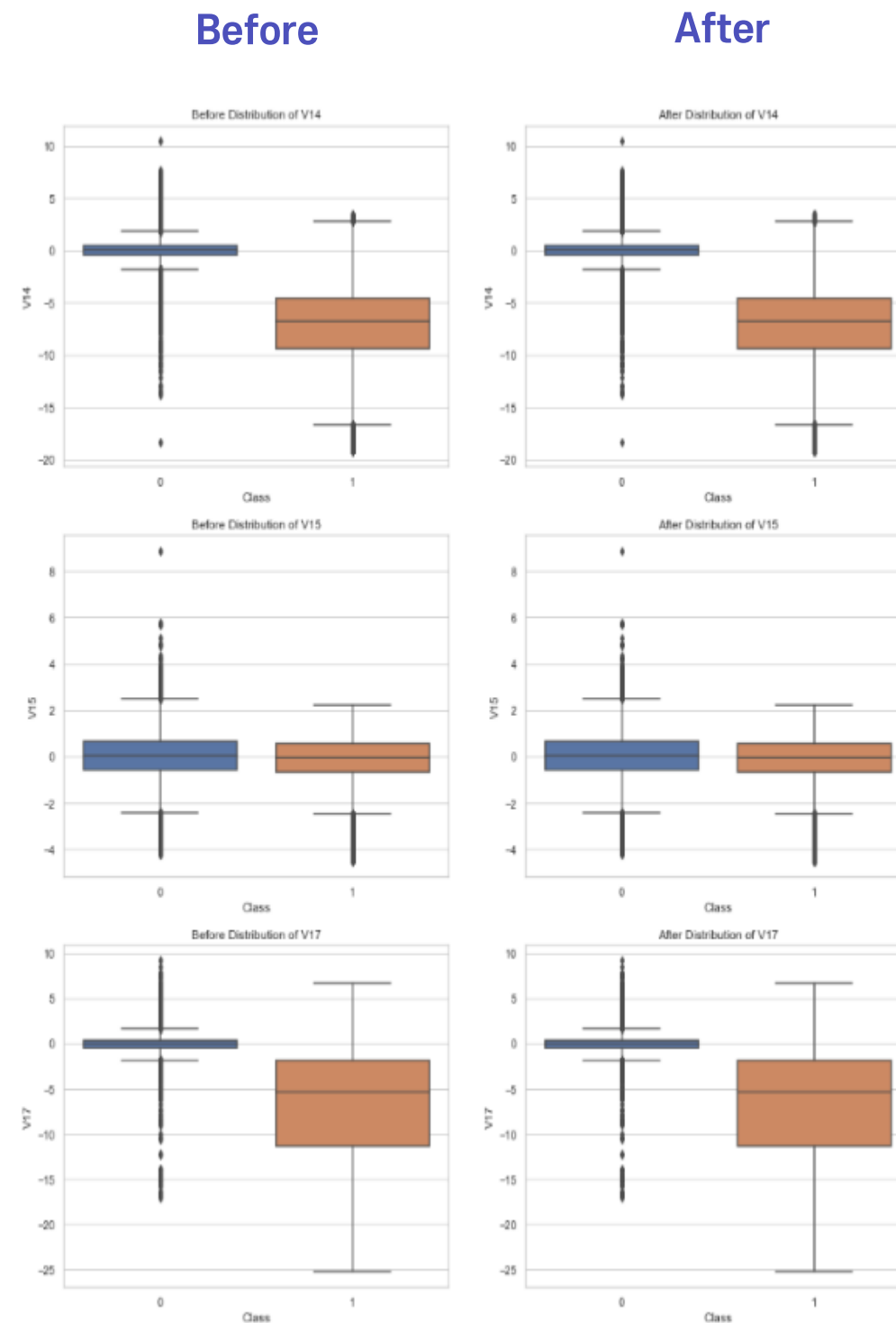
이상치 제거

| Random Forest : V4, V10, V12, V14

| CatBoost : V1, V4, V12, V14

| XGBoost : V4, V11, V12, V17

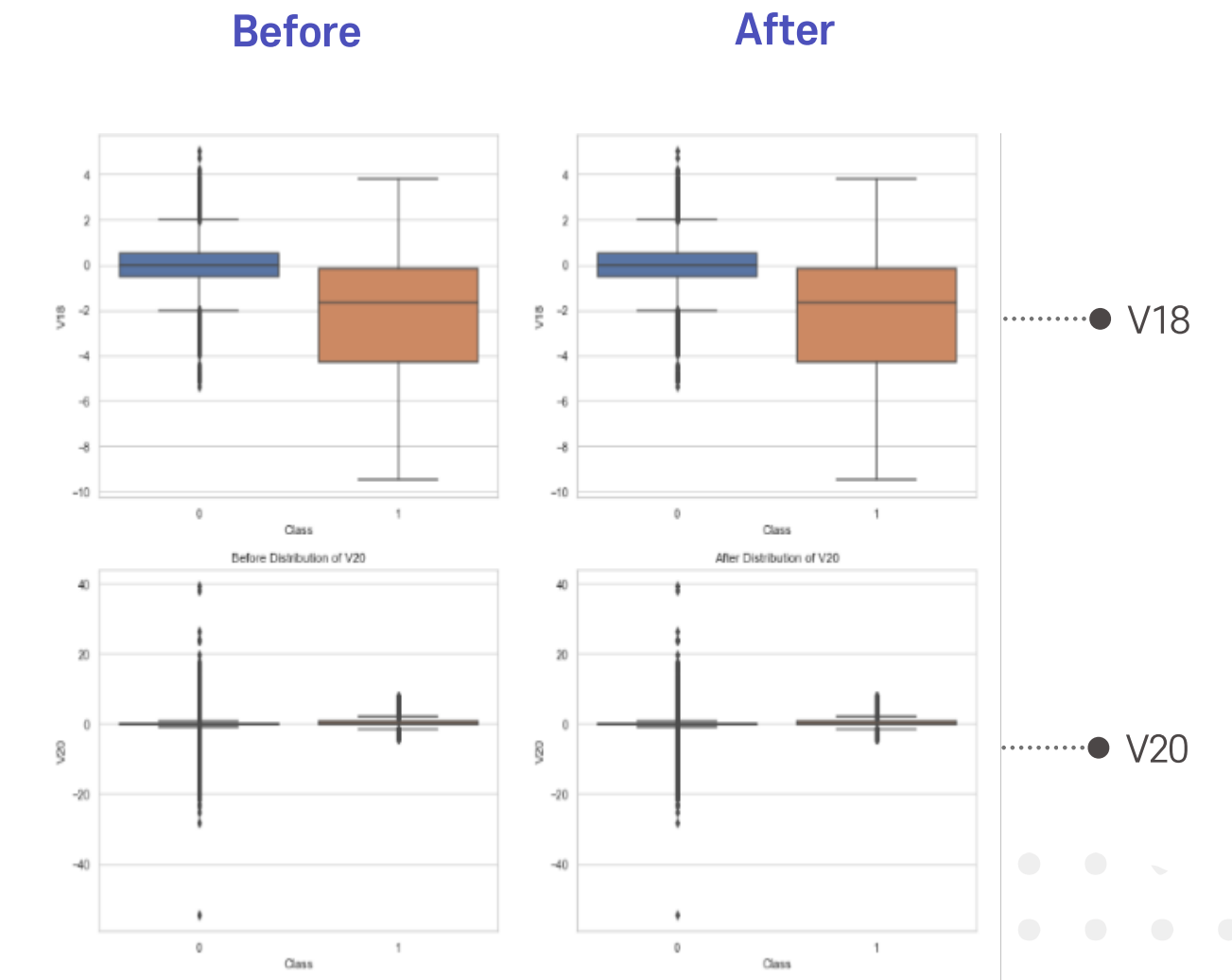
| LGBM : V13, V15, V18, V20



.....● V14

.....● V15

.....● V17



.....● V18

.....● V20

데이터 전처리

다중공선성 제거

01 변수 제거 기준



02 변수 제거 과정

Feature	VIF Factor
V7	53.11
V17	35.06
V3	30.83
V12	26.23
V16	26.00
V5	24.78
V10	22.77
V14	19.71
V2	19.57
V11	14.10
V18	13.88
V1	13.70



Feature	VIF Factor
V17	34.8
V12	26.2
V16	25.91
V3	24.17
V10	19.61
V14	18.96
V5	15.88
V11	14.04
V18	13.88
V12	13.29



Feature	VIF Factor
V12	24.85
V16	22.42
V3	22.36
V10	19.36
V14	18.96
V5	14.56
V11	13.84
V18	13.00
V12	10.60



Feature	VIF Factor
V12	24.46
V16	22.00
V3	20.73
V10	19.14
V14	18.30
V11	13.64

RF : V12 V3 V16 V10 V14
 CatBoost : V12 V16 V10 V3 V14 V11
 LGBM: V12 V16 V14 V3 V10 V11

03

변수 선택 및 모델링

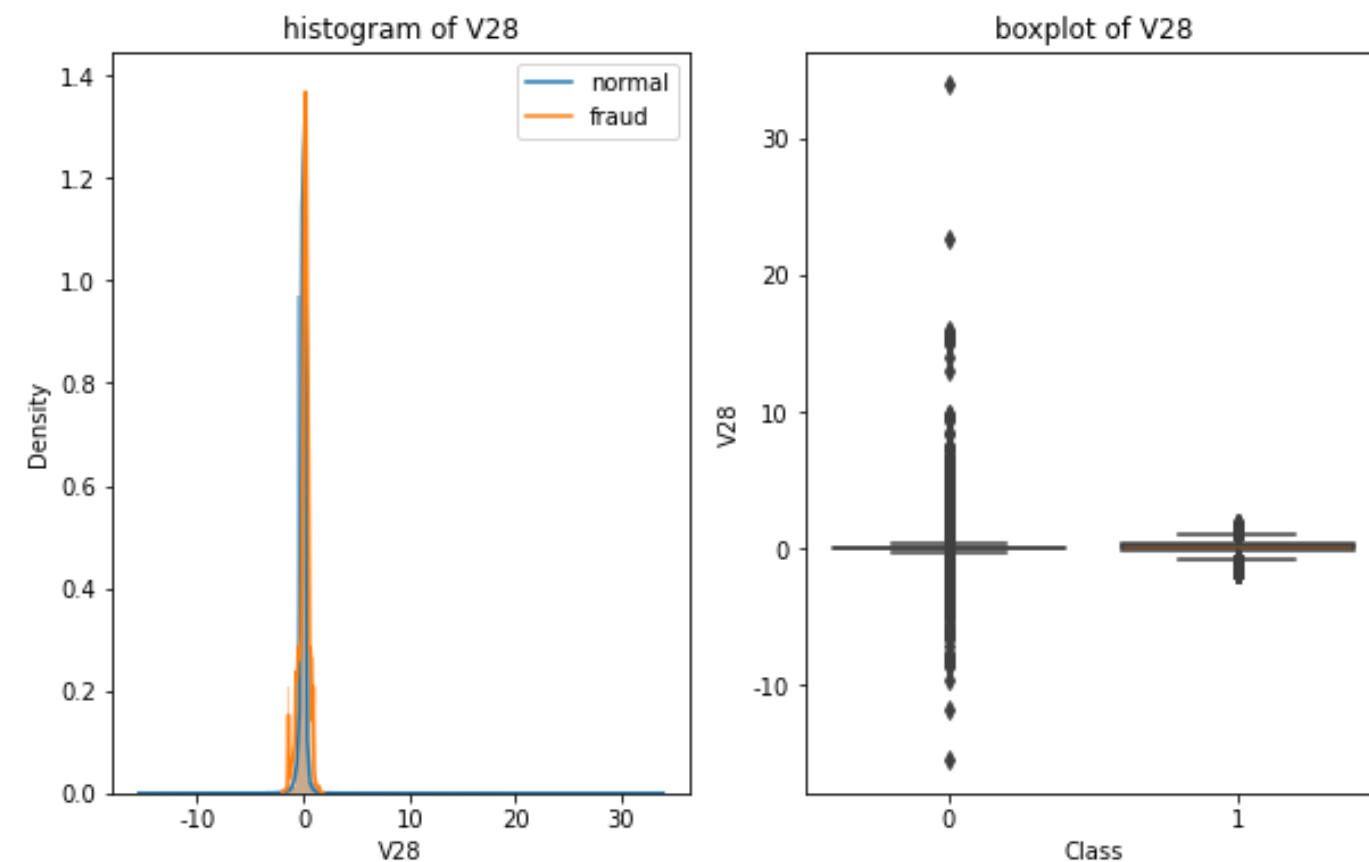
Feature Selction & Modeling

변수 선택

Feature Selection

01 Near - Zero Varinace

변수를 선택하는 기법 중 가장 단순한 방법, 분산이 거의 0인 변수 제거
V28 : 가장 낮은 분산 (0.2 이하)을 가진 변수



02 RFE (Recursive feature elimination) - 재귀적 특성 제거

모든 특성으로 시작하여 모델을 만들고, **변수 중요도**가 가장 낮은 특성을 제거하는 방식
제거한 변수는 제외하고 나머지 변수 전체로 새로운 모델 생성
지정한 특성개수가 남을 때까지 이 과정을 반복, 모델 지정 필요

| **Random Forest** V25 / V27 / V20 / V22

| **CatBoost** V27 / V22 / V20

| **XGBoost** V20 / V26 / V24 / V21

| **LGBM** V12 / V14 / V10

.....● 성능 비교 후 제거 변수 선택

변수 선택

최종 변수 비교



모델 학습 Hyperparameter Tuning



Random Forest

CatBoost

Manual Search

사용자가 꼽은 조합 내에서 최적의 조합을 찾는 방법

RandomForest : 성능을 올리는 방향으로 튜닝한 뒤, rfe로 변수선택 후에는 과적합을 방지하는 방향으로 튜닝함

CatBoost : 모델 특성 상 튜닝 없이 기본값으로도 좋은 성능을 보여줌.

또한 튜닝을 통해서 얻을 수 있는 효과는 크지 않음

Bayesian optimization

어느 입력값을 받는 미지의 목적 함수를 상정하여, 그 함수값을 최대로 만드는 최적해를 찾음
즉, 목적 함수(탐색대상함수)와 해당 하이퍼파라미터 쌍을 대상으로 Surrogate Model(대체 모델)을 만들고, 평가를 통해 순차적으로 업데이트해 가면서 최적의 하이퍼파라미터 조합을 탐색



XGBoost

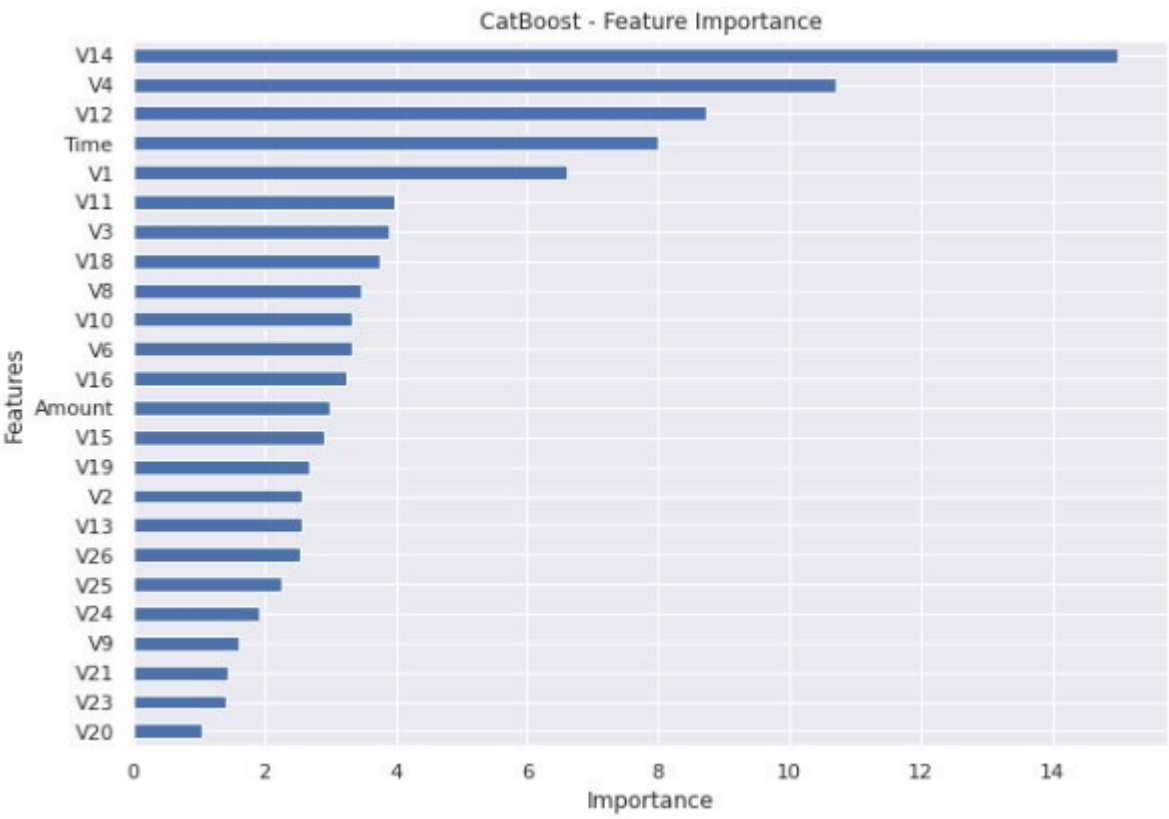
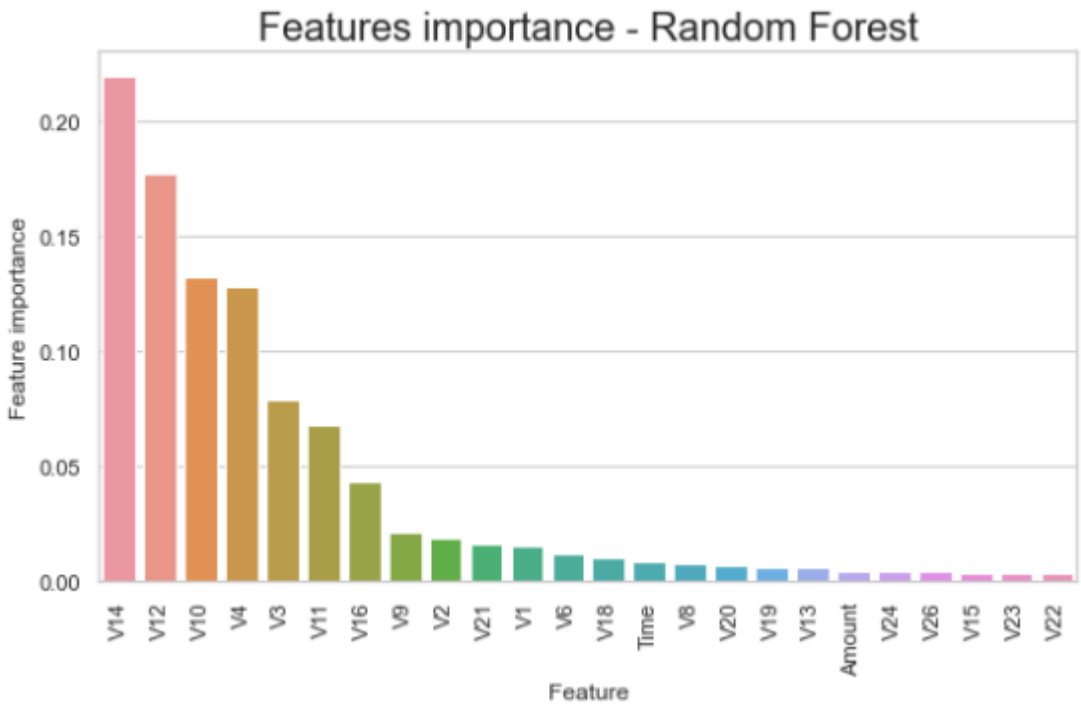
LGBM

04

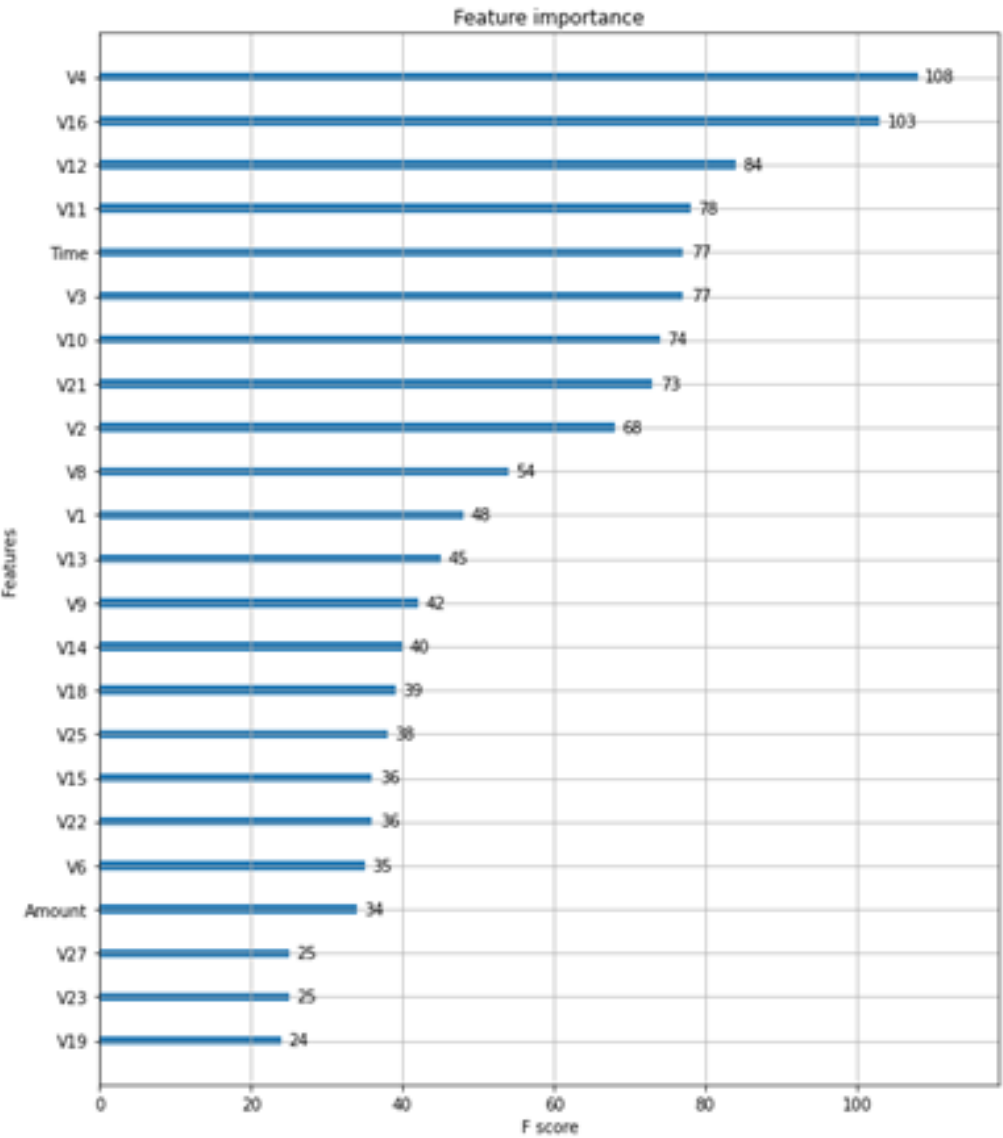
모델 평가 및 한계점

Model Evaluation & Limits

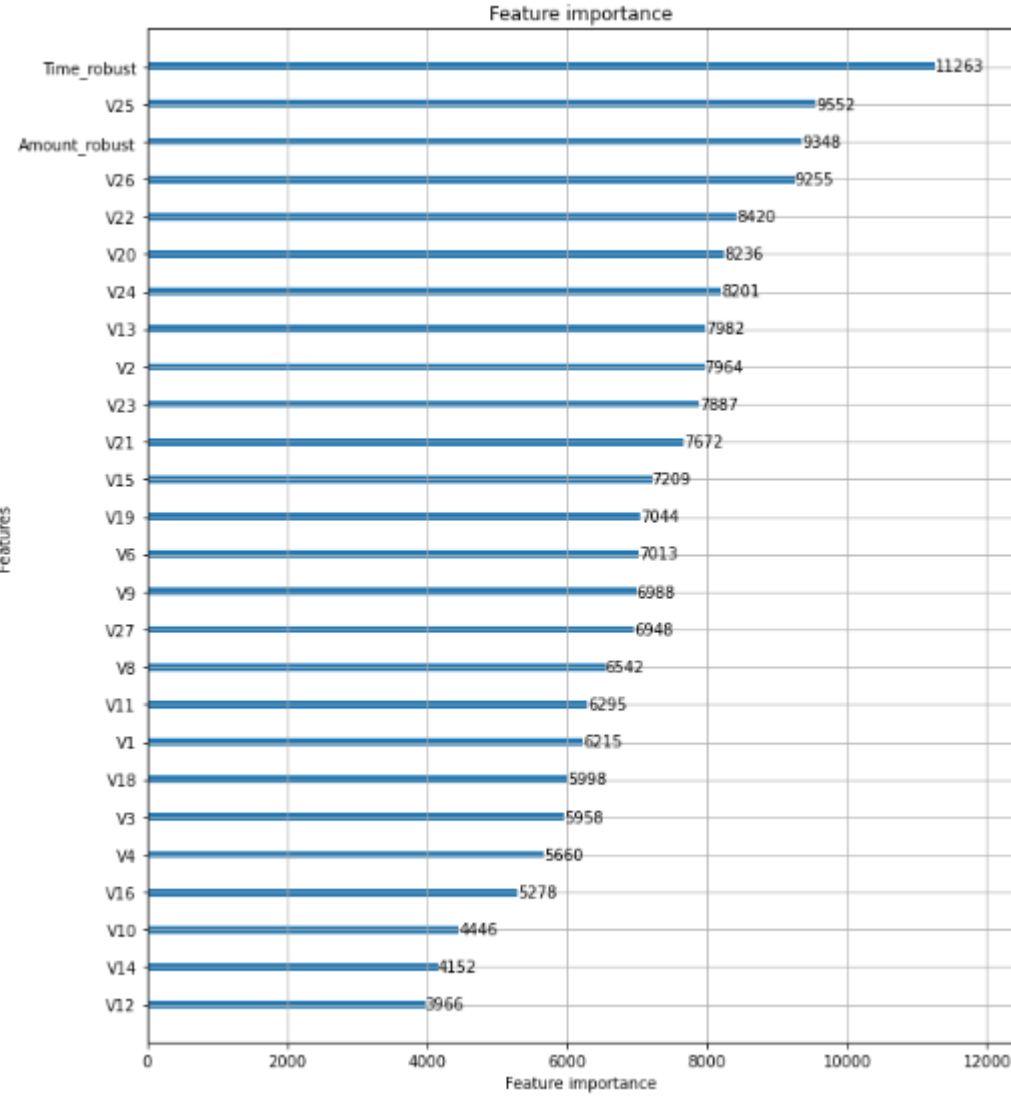
변수 중요도



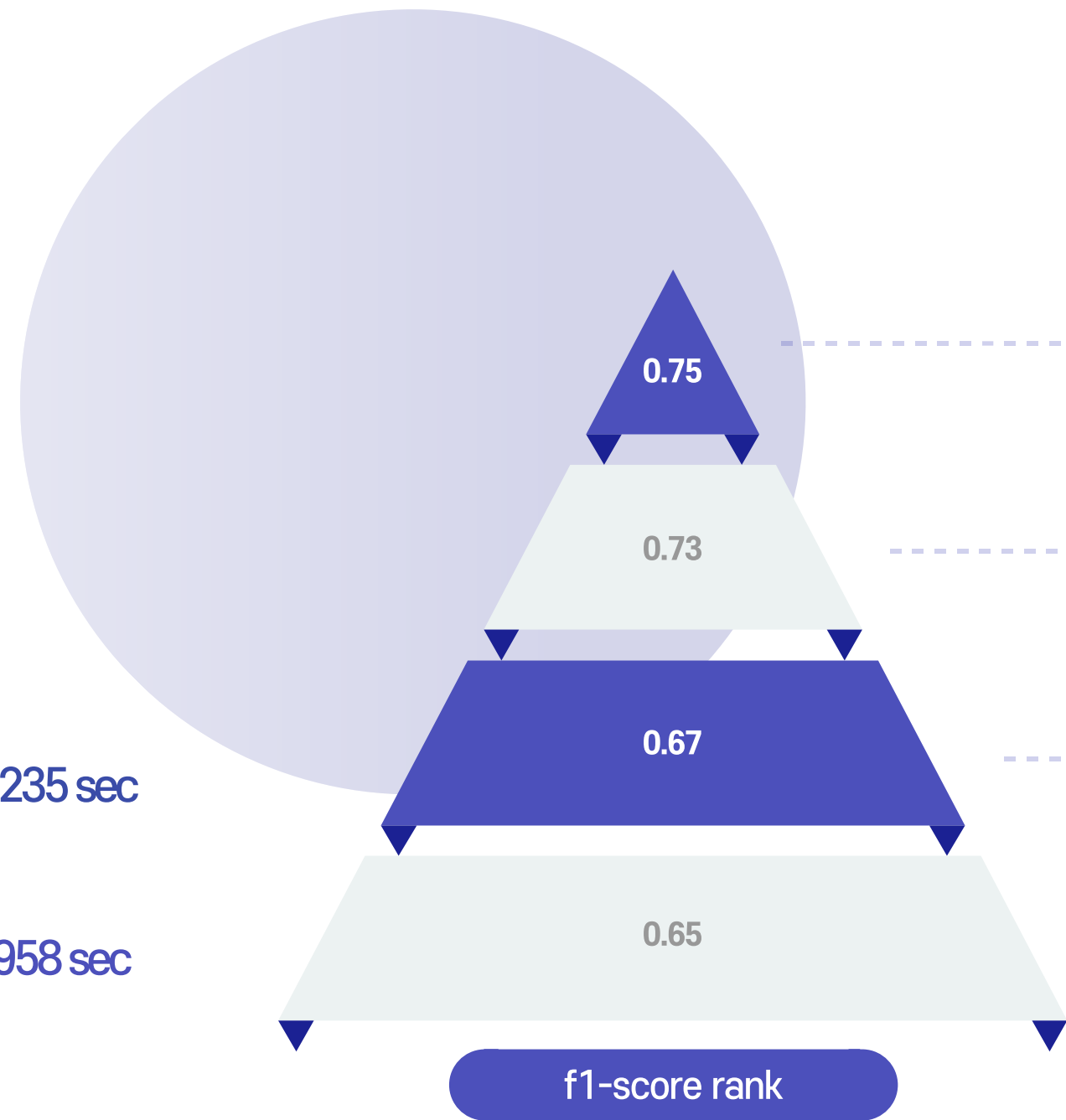
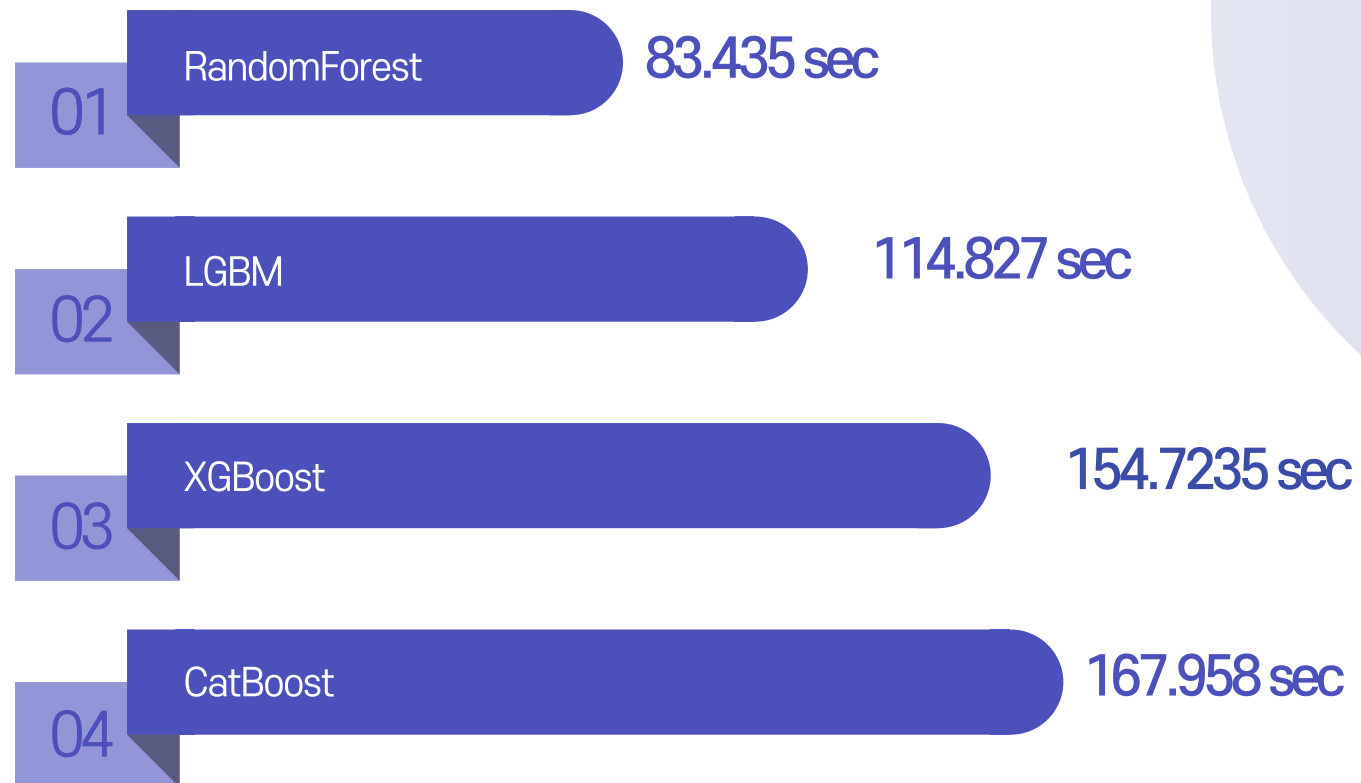
XGBoost



LGBM



모델 평가 성능 비교



LGBM

정확도: 0.9990 정밀도: 0.6989 재현율: 0.8092
f1-score: 0.7500 auc: 0.9043

CatBoost

정확도 : 0.9988 정밀도 : 0.6259 재현율 : 0.8679
f1-score : 0.7273 auc : 0.9335

XGBoost

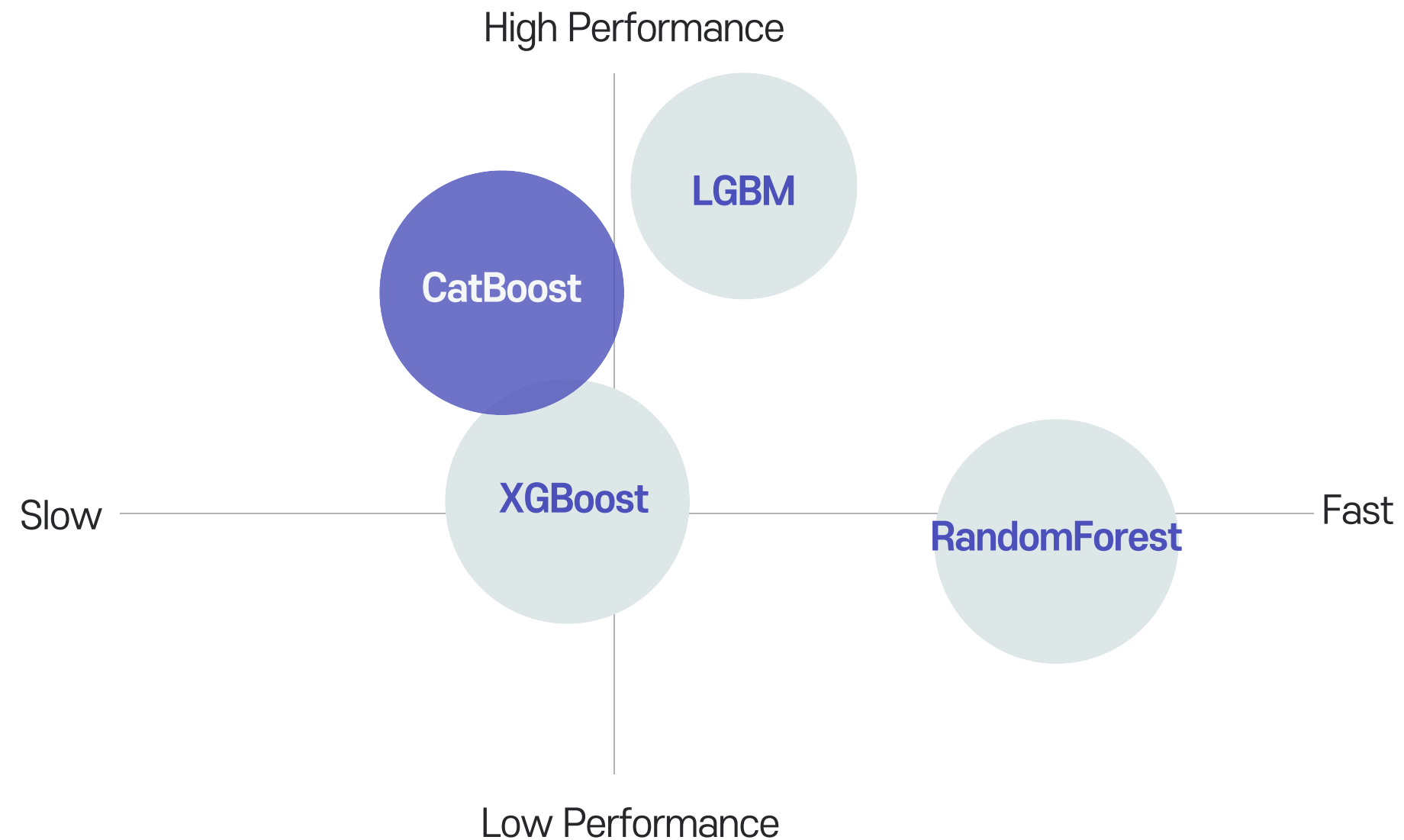
정확도 : 0.9985 정밀도 : 0.5772 재현율 : 0.8113
f1-score : 0.6745 auc : 0.9051

RandomForest

정확도 : 0.9982 정밀도 : 0.5140 재현율 : 0.8680
f1-score : 0.6456 auc : 0.978

모델 평가

성능 비교



- RandomForest : 제일 빠른 속도지만 상대적으로 낮은 성능
- LGBM: 높은 정확도와 빠른 속도를 보였지만 변수 중요도에서 다른 모델과 다소 큰 차이를 보임.
(Class별로 수치가 차이가 보였던 변수들(V12, V14 등)의 낮은 중요도 , 다른 모델에서는 공통적으로 중요하다고 판단된 변수들)
- XGBoost : 속도와 성능 면에서 다른 모델에 비해 두드러지지 못함
- CatBoost : 튜닝 없이 높은 정확도 + 상대적으로 속도는 느리지만 튜닝 시간 절약 가능

한계점

- 데이터 불균형으로 인한 분석의 어려움
- PCA처리 된 자료라 해석 및 전처리 과정에서 어려움
- SMOTETomek 과정에서 Class = 0인 데이터 수를 많이 줄이지 못함
- 다중공선성 위험을 완전히 없애지 못했음
- SVM 시간 너무 오래 걸려서 사용하지 못함
- 로지스틱 회귀, 의사결정나무의 성능이 낮아서 최종 모델에서 사용하지 못함
- 전처리 과정에서 성능이 낮아지는 경우 존재
- 과적합 가능성 존재

B.a.f Winter Vacation Team Project

THANK
YOU

team 3

최솔 변지형 신수빈 박지현