ML 과제 Report

1. 데이터 로드 및 기본 탐색
[원본 데이터 Class 비율]
Class
0 284315
1 492
Name: count, dtype: int64
Class
0 0.998273
1 0.001727
Name: proportion, dtype: float64
-> 전체 데이터 중 사기 거래(Class 1)의 비율이 매우 낮은 불균형 데이터임을 확인

2. 샘플링 결과
[2] 샘플링 후 데이터 개수: 10492
Class
0 10000
1 492
Name: count, dtype: int64
-> 10000개로 추려진 정상 거래와, 492건의 기존 사기 거래가 잘 합쳐짐

3. 데이터 전처리
(코드 셀 결과물은 따로 없음)
-> 데이터의 Amount(거래 금액) 변수는 값의 범위가 넓어 모델 학습 시 편향을 유발할 수 있습니다. 이를 방지하기 위해 StandardScaler를 적용하여 평균 0, 분산 1을 갖는 정규 분포 형태로 변환하였습니다.
변환된 값은 Amount_Scaled라는 새로운 변수로 추가하였으며, 기존의 Amount 변수와 학습에 불필요한 Time 변수는 제거하였습니다. 전처리가 완료된 데이터프레임을 독립변수(X)와 종속변수(y, Class)로 분리하였습니다.

4. 학습 데이터와 테스트 데이터 분할
[4] 분할 후 학습 데이터 비율:
Class
0 0.953056
1 0.046944
Name: proportion, dtype: float64
-> 불균형 데이터의 특성을 고려하여 stratify=y 옵션을 설정, 원본 데이터의 정상/사기 비율이 분할된 데이터셋에서도 동일하게 유지되도록 하였습니다.

5. SMOTE 적용
[5] SMOTE 적용 후 데이터 개수:
{0: 7999, 1: 7999}
-> 기존의 사기 건수 492건에서, 정상 거래와의 갯수가 맞춰지도록 7999건까지 늘렸습니다.
기존 데이터는 정상 거래가 사기 거래보다 많은 불균형 상태입니다. 이를 그대로 학습 시, 모델은 다수 클래스인 '정상' 쪽으로 편향(Bias)되어 사기 거래를 제대로 탐지하지 못할 위험이 있습니다. 단순히 데이터를 복제하는 방식은 과적합(Overfitting)을 유발할 수 있습니다.
따라서 소수 클래스 사이의 특성을 반영한 가짜 데이터를 생성하는 SMOTE를 이용합니다.

6. 모델 학습 및 성능 평가
결과에 앞서 학습 전략부터 설명드리겠습니다.
데이터 샘플링(Under-sampling)과 SMOTE(Over-sampling)를 통해 클래스 불균형을 완화한 후, 성능을 극대화하기 위한 3단계 최적화(Optimization) 전략을 취했습니다.

1. 임계값 조정 (Threshold Tuning)
Precision-Recall Curve를 분석하여 F1-score가 최대가 되는 최적의 임계값을 찾아냄
<- 학습 데이터는 SMOTE를 통해 1:1 비율이 되었으나, 테스트 데이터는 여전히 불균형(Imbalanced) 상태이기 때문
-> Precision(정밀도)와 Recall(재현율)의 트레이드오프(Trade-off) 관계를 고려

2. 하이퍼파라미터 튜닝 (Criterion & Estimators)
불순도 지표를 Gini에서 Entropy로 변경하고, 트리의 개수(n_estimators)를 100개에서 300~500개로 늘림
-> Entropy는 클래스 불순도에 더 민감하게 반응
-> 트리 개수 증가는 과적합을 방지하고 일반화 성능을 높이는 데 기여

3. 모델 고도화
최종 모델로 RandomForestClassifier 대신 ExtraTreesClassifier (Extremely Randomized Trees)를 선정
-> 변동성 감소, SMOTE 노이즈에 대한 강건성(Robustness)

해당 전략에 따라 학습된 모델을 사용하여 테스트셋(X_test)에 대한 예측값(predict)과 사기일 확률(predict_proba)을 산출하였으며, 주요 성능 지표를 확인하였습니다.
그 결과는 다음 페이지 내용과 같습니다:

>> [ExtraTrees 하이퍼파라미터 + Threshold 조정] 성적표
PR-AUC Score: 0.9529
Best Threshold: 0.4920
예상 F1: 0.9305 (Precision: 0.9775, Recall: 0.8878)

	precision	recall	f1-score	support
0	0.99	1.00	1.00	2001
1	0.98	0.89	0.93	98
accuracy			0.99	2099
macro avg	0.99	0.94	0.96	2099
weighted avg	0.99	0.99	0.99	2099

-> class 0,1 둘 다 Recall ≥ 0.80, F1 ≥ 0.88, PR-AUC ≥ 0.90을 만족하였습니다.