4.9 분류 실습 - 캐글 산탄데르 고객 만족 예측

1) XGBoost 기본모델

기본 설정으로 XGBoost 를 학습했을 때,
검증 AUC 로그를 출력하며 조기 종료 조건을 통해 과적합 방지
이때 **테스트 데이터 ROC-AUC 는 약 0.8429** 로 출력

2) Hyperopt 를 이용한 자동 튜닝

Hyperopt 라이브러리를 사용해 XGBoost 의 주요 하이퍼파라미터 최적화

탐색한 변수: 트리의 깊이, 자식 노드 최소 가중치, 학습률, 컬럼 샘플 비율 등
5 겹 교차검증을 사용해 평균 AUC 를 기준으로 탐색 진행

탐색 결과, 최적의 파라미터 조합을 적용했을 때 **ROC-AUC 가 0.8457 로 상승**
이는 기존보다 약 **0.0028 개선**된 수치

3) XGBoost 의 피처 중요도

튜닝된 모델의 피처 중요도를 시각화했을 때, var38, var15 이 두 피처가 모델의 결정에 압도적인 영향을 미쳤고
saldo_, num_var, imp_var 계열의 변수들도 상당히 중요한 역할을 하는 것으로 나타남.

즉, 고객의 잔고(saldo)나 거래 횟수(num_var) 같은 재무적 활동 지표가 불만족 여부 예측에 핵심적

4) LightGBM 실험

LightGBM 은 같은 데이터로 비교 실험을 진행
기본 설정에서 **ROC-AUC 는 0.8380** 으로 XGBoost 보다 약간 낮은 성능을 보임

하이퍼파라미터를 동일하게 Hyperopt 로 튜닝했지만 성능 향상 폭은 크지 않았고
최종적으로 XGBoost 보다 소폭 낮은 수준에 머묾

다만 **훈련 속도는 훨씬 빠름.**

<Decision Tree and Random Forest Classifier Models>

1) 주요 주제 & 목적

의사결정 나무(Decision Tree) 와 **랜덤 포레스트(Random Forest)** 분류 모델을 비교하고 두 모델의 작동 원리, 장단점, 그리고 실험 결과를 통해 어떤 상황에서 유리한지 분석함.

2) 다룬 모델

1. Decision Tree (의사결정 나무, DT)
    - 데이터를 분할하면서 규칙을 만들어 결정 경로 생성
    - 설명력이 좋고 직관적
    - 과적합에 약함
2. Random Forest (랜덤 포레스트, RF)
    - 여러 개의 결정 나무를 **앙상블**해 평균 혹은 투표로 예측
    - 각 나무는 **부트스트랩 샘플링** + **무작위 피처 선택** 조합
    - 과적합 완화, 예측 성능 안정성 증대

3) 실험 구성 & 데이터 처리 흐름

- 특정 데이터셋을 불러와서 전처리
- 학습/검증/테스트 분할
- Decision Tree 와 Random Forest 모델 각각 학습
- 하이퍼파라미터 튜닝
- 평가 지표: 정확도, 정밀도, 재현율, F1, ROC-AUC

4) 실험 결과 요약 & 비교

- Decision Tree 결과
    - 학습 데이터에서 높은 정확도
    - 검증/테스트에서 성능 하락 → 과적합 조짐
    - 특정 피처가 기준이 되어 분할되는 경로들이 명확히 시각화 가능
- Random Forest 결과
    - Decision Tree 대비 **일반화 성능이 더 안정적**
    - 변동성 낮고 과적합 덜함
    - 개별 나무의 성능보다는 앙상블의 평균 효과가 중요

<Beginner Friendly CATBOOST with OPTUNA>

1) 주요 목표 & 주제

CatBoost 모델을 Optuna 라는 하이퍼파라미터 최적화 라이브러리와 결합해 사용하는 방법 안내. 즉, CatBoost 의 특장점을 살리면서 자동 탐색을 통해 최적 파라미터를 찾는 흐름을 보여줌.

2) 다룬 주요 내용 & 흐름

1. **CatBoost 소개 및 장점**
    - 범주형 변수 처리에 강점
    - 과적합에 대한 내성 존재
    - 빠른 학습 및 예측 가능
2. **데이터 전처리 및 준비**
    - 데이터셋을 불러오고 결측치 처리
    - 범주형 변수 인덱스 지정
    - 학습/검증 데이터 분리
3. **Optuna 를 통한 하이퍼파라미터 탐색**
    - Optuna 의 Trial 객체로 탐색 공간 정의
    - 각 Trial 마다 CatBoost 모델을 학습하고 검증 AUC 또는 손실을 기준으로 평가
    - 최적 파라미터를 찾아내고 기록
4. **최적 파라미터로 모델 재학습 & 평가**
    - Optuna 탐색으로 얻은 파라미터를 사용해 CatBoost 모델을 재학습
    - 검증/테스트 데이터로 성능 평가 (예: ROC-AUC)
5. **해석 및 중요도 분석**
    - CatBoost 자체 기능이나 별도의 도구로 **피처 중요도 시각화**
    - 어떤 변수들이 모델 예측에 기여했는지 분석