**인공지능 학습 결과서**

# **1. 기본 정보**

* **프로젝트명 :** 가입 고객 이탈 예측 – 고객 특성에 따른 이탈률 예측
* **작성자 :** SKN14-2Team
* **작성일 :** 2025.06.05
* **모델 목적/용도 :** 고객의 이탈률을 예측하여 고객 이탈을 방지하기 위한 마케팅 비용 등을 최소화 하고 고객을 유지하기 위한 척도로 활용
* **사용 데이터셋 :** [**Credit Card Customer Churn Prediction**](https://www.kaggle.com/datasets/rjmanoj/credit-card-customer-churn-prediction/code)
* **학습 대상(task) :** 고객 데이터 및 이탈여부( 이진분류 )

**2. 데이터 요약**

* **학습 데이터 크기 :** 약 8,000 건
* **검증 데이터 크기 :** 약 2000 건
* **데이터 전처리 내용 요약 :** 오버샘플링과 언더샘플링으로 성능 개선 시도
* **클래스 분포 (분류 문제일 경우) :** 잔류(0), 이탈(1)
* **클래스 샘플 수 비율 :** 잔류 대 이탈 비율 4 : 1

**3. 모델 구조 및 설정**

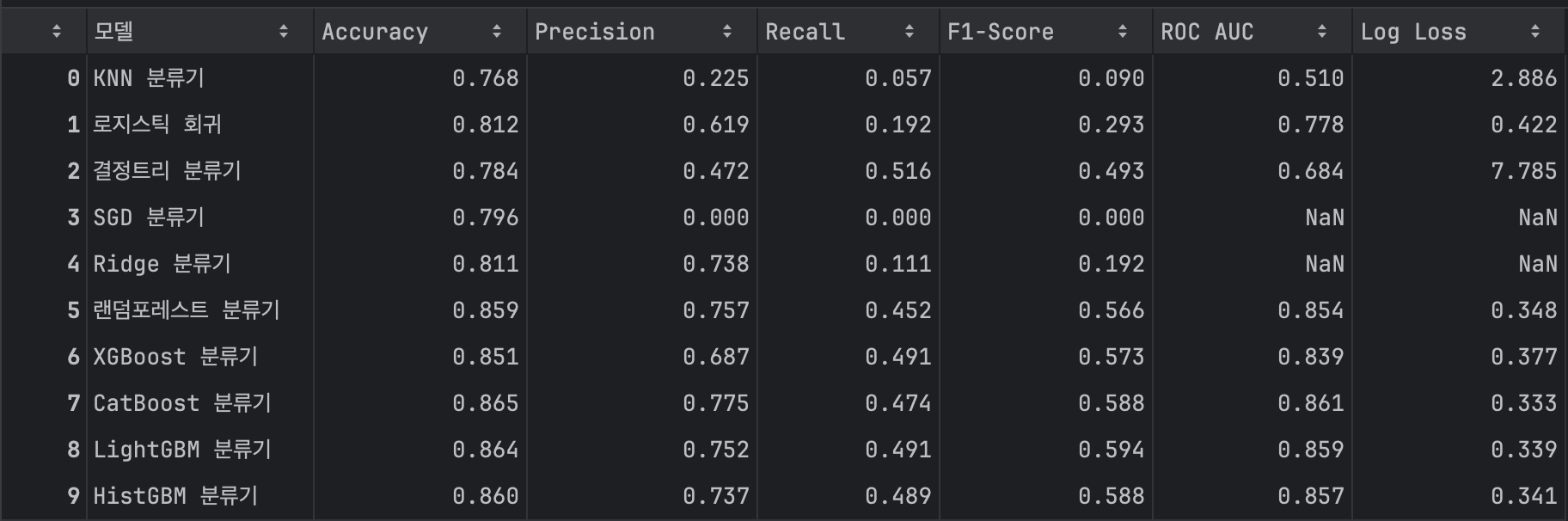
* **사용 모델:**
  + SGDClassifier
  + DecisionTree
  + SVC
  + MLP
  + TabNet
  + RandomForestClassifier
  + XGBoostClassifier
  + HistGradientBoostClassifier
  + LightBGM
  + CatBoost
* **프레임워크/라이브러리 :**
  + PyTorch
  + scikit-learn
  + pandas
* **하이퍼파라미터 설정 :** 데이터셋의 영향이 커서 하이퍼파라미터로 조정하는 정도로는 성능에 변화가 없거나 오히려 하락하는 모습을 보임.

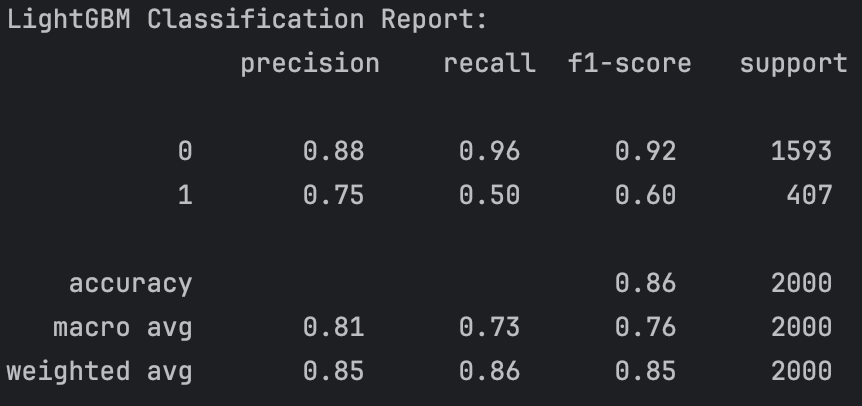
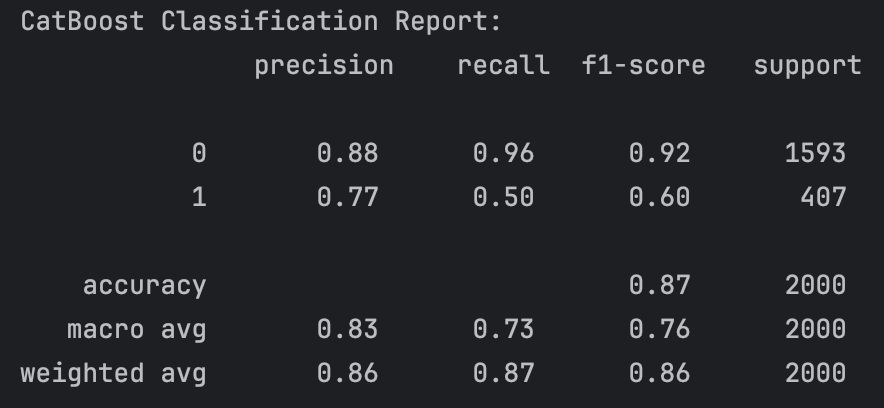
**4. 학습 환경**

* **OS / 플랫폼 :** Win11, MacOS
* **GPU / CPU 사양 :** 2.40GHz 4Core CPU, M2
* **RAM / Storage :** 16GB
* **소프트웨어 버전:** (Python, 라이브러리 등)
  + python : 3.12
  + black : 25.1.0
  + numpy : 2.2.5
  + pandas : 2.2.3
  + torch : 2.7.0
  + matplotlib : 3.10.3
  + seaborn : 0.13.2
  + streamlit : 1.45.1
  + jupyter : 1.1.1
  + xgboost : 3.0.2
  + lightgbm : 4.6.0
  + catboost : 1.2.8
  + scikit-learn : 1.6.1
  + pytorch-tabnet : 4.1.0

**5. 성능 결과**

* **모델 성능 지표 :** 잠재이탈확률이 높은 고객의 pool 을 예상하는 것이 주 목적이므로, 정확도, 재현률, F1-Score 를 기준으로 Top3 선정
  + LightGBMClassifier( F1 = 0.594 )
  + CatBoostClassifier( Accuracy = 0.865, Precision = 0.775 )
  + HistGBMClassifier( F1 = 0.588, 빠른 학습/추론 )





**7. 주요 해석 및 분석**

* **모델 성능 요약 해석 :** 작은 데이터셋, 불균형비 4:1 등 성능향상에 제약
* **오류 사례 분석 :** 실제 이탈률 예측에 사용되는 컬럼의 개수가 절반 이하
* **모델 한계 및 개선점 :** 상관관계가 높은 특성들을 추가하고, 데이터셋 크기를 키우되 불균형비가 너무 커지지 않도록 데이터를 증류할 방법이 필요함