**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
* **행:**10000개
* **열:** 14개
* **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.

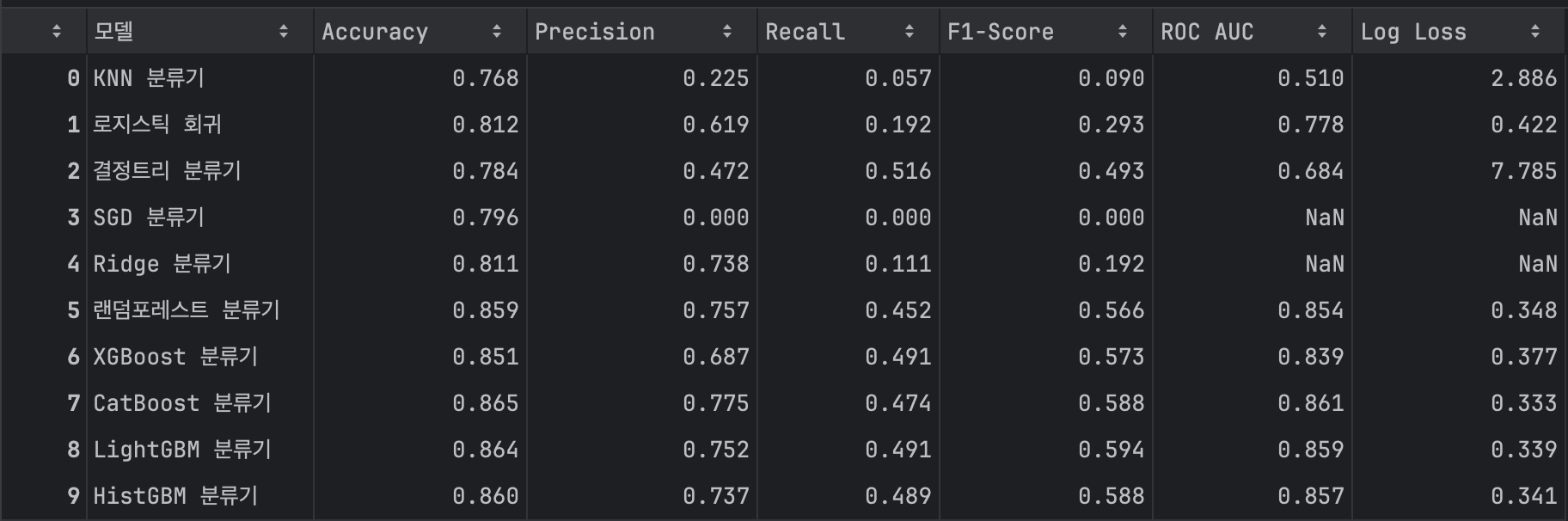
**4. 학습 환경**

* **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. 주요 모델 학습 결과**

* **모델 성능 지표 :** 먼저 10개의 머신러닝 모델을 **하이퍼파라미터 조정 없이 학습**시켜 기본 성능을 비교하였음. 이후 상위 3개의 모델을 선정하여 **하이퍼파라미터 튜닝을 통해 최적화**함.
  + LightGBMClassifier이탈 고객 예측 성능 F1 = 0.60, 정밀도 0.75, 재현율 0.50)
  + CatBoostClassifier( F1 = 0.60, 정밀도 0.79, 재현율 0.48)
  + HistGBMClassifier( F1 = 0.588, 빠른 학습/추론 )

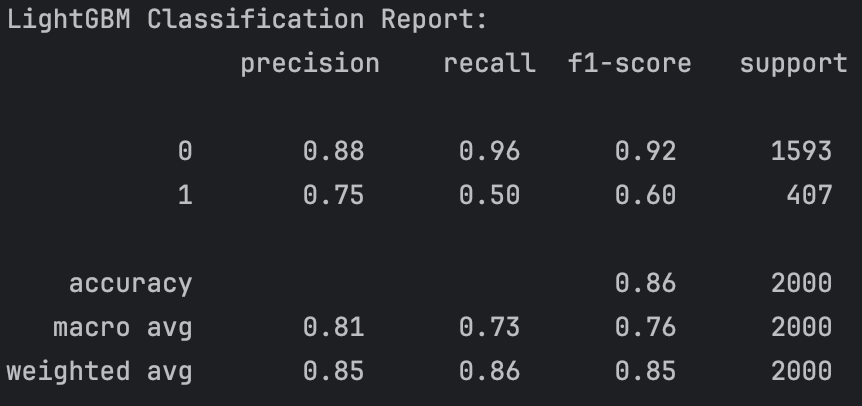
**[하이퍼파라미터 전 모델 평가]**



**[Top2 모델 평가]**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.



**5 -1 acuuracy가 90%가 넘지 않는 이유**

**(1️)데이터 불균형의 영향**

초기 데이터는 **잔류 고객(음성)이 약 4배 많고, 이탈 고객(양성)은 소수**인 불균형 상태였습니다. 이는 머신러닝 모델이 이탈 고객을 충분히 학습하기 어렵게 만드는 주요 원인.

**(2️)인위적 밸런싱을 사용하지 않은 이유**

* SMOTE 등 오버샘플링/언더샘플링 기법도 시도해 보았지만, 성능이 오히려 하락하거나 과적합이 발생함.
* 무엇보다 **실제 운영 환경에서도 이탈 고객은 소수일 가능성이 높기 때문에**, **불균형 상태 그대로 학습하고 평가**하는 것이 현실적인 모델 성능을 반영한다고 판단함.

**(3️)현실에 기반한 평가 전략**

* 비록 모델 지표가 아주 높진 않지만, \*\*실제 업무에 적용할 수 있는 기준선(Baseline)\*\*을 확보하는 것이 더 중요하다고 보았음.

**5-2 비교분석**

| **평가 지표** | **LightGBM** | **CatBoost** | **우세 모델** |
| --- | --- | --- | --- |
| 정확도 Accuracy | 0.86 | 0.87 | CatBoost |
| 정밀도 Precision(이탈) | 0.75 | 0.79 | CatBoost |
| 재현율 Recall(이탈) | 0.50 | 0.48 | LightGBM |
| F1 점수 F1-score(이탈) | 0.60 | 0.60 | 동일 |

**5-3 사용자 모델 선택**

**(1) CatBoost가 적합한 경우**

* 이탈 고객 예측을 **정확하게** 해야 하는 비즈니스
* 잘못된 예측(오탐)으로 **불필요한 비용**이 발생하는 경우 (예: 마케팅 비용, 상담 리소스 낭비)

**(2) LightGBM이 적합한 경우**

* 이탈 고객을 **놓치지 않아야** 하는 경우 (Recall 중요)
* 실제 이탈 시 **매출 손실이 큰 VIP 고객**이 존재하는 경우

**(3) HistGBM이 적합한 경우**

* 모델을 자주 학습하거나 빠른 응답이 필요한 경우

**6. 주요 해석 및 분석**

* **모델 성능 요약 해석 :** 작은 데이터셋, 불균형비 4:1 등 성능향상에 제약에도 불구하고 현실적인 데이터의 핏한 모델을 제작
* **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

**7. 결론 및 향후 방향**

* 모델 성능이 아주 뛰어나진 않지만, 현실적인 상황을 반영하여 **신뢰할 수 있는 기준선**을 확보함.
* 앞으로는 운영 데이터가 누적됨에 따라 **threshold 조정**, **정밀 튜닝**, **특성 확장(feature engineering)** 등을 통해 점차 개선해나갈 수 있음.

**초기 모델은 완벽함보다는 현실 반영이 중요함.**

우리가 처한 데이터 상황과 실제 업무 환경을 고려하여 전략적으로 판단함.

**8.부록**

**텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

**텍스트, 스크린샷, 그래프, 도표이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

**텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.텍스트, 스크린샷, 라인, 번호이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.텍스트, 라인, 그래프, 도표이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.텍스트, 라인, 그래프, 도표이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**