머신러닝 및 딥러닝 모델 학습 보고서

Ⅰ. 머신러닝

본 프로젝트에서는 인터넷 서비스 고객의 이탈(churn) 여부를 예측하기 위해 데이터의 특성과 문제 구조에 적합한 다섯 가지 머신러닝 모델을 선정하였다.
각 모델은 서로 다른 학습 방식과 특성을 가지며, 이를 통해 모델 간 성능 비교와 변수 중요도 해석이 가능하도록 설계하였다.

1. Logistic Regression (로지스틱 회귀)

- 특징:
  로지스틱 회귀는 이진 분류 문제에서 가장 기본적이면서 해석력이 높은 선형 모델이다.
  본 데이터의 타깃 변수 *churn* 은 0(유지) 또는 1(이탈)로 구성되어 있어, 각 고객의 이탈 확률을 로짓(logit) 형태로 추정하였다.
- 활용:
  전처리 과정에서 연속형 변수(bill_avg_log, download_avg_log, upload_avg_log)와
  원-핫 인코딩된 범주형 변수(contract_type, subscription_label)를 함께 활용하였다.
  이를 통해 각 요인의 회귀계수를 통해 이탈에 영향을 미치는 주요 요인을 명확히 해석할 수 있다.
- 역할:
  복잡한 트리 기반 모델의 결과를 보완하는 해석 지표로 활용되었다.

2. Random Forest (랜덤 포레스트)

- 특징:
  여러 개의 Decision Tree 를 앙상블하여 성능을 향상시키는 모델로,
  데이터의 비선형 관계를 잘 포착하고 이상치와 스케일링에 강건하다.
- 데이터 적합성:
  본 데이터는 bill_avg, download_avg 등의 연속형 변수와
  subscription_label, contract_type 과 같은 범주형 변수가 함께 존재하는 혼합형 데이터로,
  트리 기반 구조인 Random Forest 에 적합하다.
- 활용:
  변수 중요도를 시각화하여 이탈에 영향을 주는 핵심 요인을 제시하는 데 활용되었다.

3. XGBoost

- 특징:
  Gradient Boosting 알고리즘 기반의 고성능 트리 앙상블 모델로, 학습 속도와 정확도가 우수하다.

- 활용 이유:
  약 7 만 개의 행과 다수의 파생 변수를 포함한 데이터의 복잡한 비선형성을 반영하기 위해 사용하였다.
  결측치 처리와 정규화 부담이 적고, L1·L2 정규화 기능을 내장하여 과적합 방지에 효과적이다.
- 성과:
  실제 서비스 환경의 노이즈 데이터에도 안정적인 성능을 보였다.

4. LightGBM

- 특징:
  XGBoost 의 구조를 개선한 모델로, 대규모 데이터 처리 속도와 메모리 효율성이 뛰어나다.
- 활용 이유:
  고객 수가 약 71,892 명에 달하는 대규모 데이터와 다수의 범주형 변수를 효율적으로 처리하기 위해 채택하였다.
  Label Encoding 된 범주형 변수를 내부적으로 최적화하여 별도의 원-핫 인코딩이 불필요하다.
- 성과:
  예측 성능과 학습 효율성 간의 균형이 가장 우수한 모델로 평가되었다.

5. CatBoost. 특징:
범주형 변수 처리에 최적화된 부스팅 모델로, Order Encoding 과 Target Statistics 방식을 사용한다.

- 활용 이유:
  subscription_label, contract_type, 구독 연수 등 범주형 특성이 다수 존재하는 데이터 구조에 적합하였다.
- 장점:
  데이터 불균형과 결측치를 자동으로 처리하여 복잡한 전처리 없이도 높은 성능을 확보하였다.
- 성과:
  고객 세분화(segment)별로 이탈 위험군 탐지에 안정적인 결과를 보였다.

6. 모델 선정의 종합적 타당성

- 데이터 구조 측면:
  수치형·범주형 혼합 데이터로, 트리 기반 모델이 적합하며 Logistic Regression 은 해석력 제공한다.

- 모델 다양성 측면:
  단순 선형 → 랜덤 포레스트 → 부스팅 계열(XGBoost, LightGBM, CatBoost)로
  점진적인 성능 및 복잡도 비교 가능하다.
- 실무 활용성 측면:
  LightGBM 과 CatBoost 는 실제 고객 이탈 예측 서비스(텔레콤, 구독 서비스 등)에서
  널리 활용되는 모델로, 본 프로젝트의 비즈니스 목적과 일치한다.

Ⅱ. 딥러닝

MLP (다층 퍼셉트론)

1) Sklearn MLPClassifier 기반 Baseline 구축

- 목적:
  전통적인 ML 환경에서 MLP 의 기본 성능을 빠르게 검증하기 위함이다.
- 내용:
  GridSearchCV 를 활용해 하이퍼파라미터를 탐색하고,
  최적 조합을 통해 Baseline 성능을 확보하였다.
  이후 PyTorch 및 TensorFlow 모델과의 비교 기준으로 사용되었다.

2) PyTorch 기반 MLP 구축

- 목적:
  Sklearn 에서 도출된 최적 구조가 딥러닝 프레임워크에서도 재현되는지 검증하였다.
- 구성:
  Batch Normalization, Dropout, EarlyStopping 등 학습 안정화 기법을 적용하였다.
  학습률(0.1, 0.01, 0.001)을 변경하며 민감도 분석을 수행하였다.
- 결과:
  학습률 변화에도 성능 차이가 거의 없어, 데이터 품질이 충분히 확보된 경우
  학습률 조정이 성능에 큰 영향을 주지 않을 수 있음을 확인하였다.

3) TensorFlow 기반 MLP 구축

- 목적:
  다른 프레임워크에서도 동일 구조의 MLP 성능을 비교하기 위함이다.
- 내용:
  Keras API 를 활용하여 모델을 설계하였으며, 은닉층 수와 노드 수를 변경하며 성능
  변화를 분석하였다.
  이를 통해 최적의 은닉층 구성을 도출하였다.

4) RandomSearch 기반 하이퍼파라미터 탐색

- 특징:
  GridSearch 의 전수 탐색 대비 연산 비용이 적은 무작위 탐색 방식을 사용하였다.
- 내용:
  n_iter 와 교차검증 횟수에 따른 효율을 비교하여
  최소한의 자원으로 최적 성능에 근접한 하이퍼파라미터를 찾았다.
- 결과:
  딥러닝 환경에서는 GridSearch 보다 RandomSearch 가 효율적인 탐색 전략임을
  검증하였다.

Ⅲ. 결론 (모델 선정 중심 보고서용)

본 프로젝트에서는 고객 이탈 예측이라는 이진 분류 문제를 해결하기 위해,
머신러닝과 딥러닝의 대표적인 모델들을 폭넓게 비교·분석하고 각각의 적용 목적과 역할을 명확히 구분하였다.

1. 머신러닝 모델 요약
    - Logistic Regression은 변수 해석에 강점을 가진 선형 모델로, 고객 이탈에 영향을 미치는 주요 요인을 파악하는 기준 모델로 사용되었다.
    - Random Forest는 비선형 관계를 포착하고 변수 중요도를 시각화할 수 있어, 모델 비교의 중간 수준 역할을 수행하였다.
    - XGBoost, LightGBM, CatBoost는 모두 Gradient Boosting 기반의 고성능 트리 앙상블로,
      데이터의 복잡성과 대규모 특성을 반영하기 위한 핵심 모델로 선정되었다.
      특히 LightGBM과 CatBoost는 연산 효율성과 범주형 변수 처리 측면에서 우수성을 보여,
      실무 적용 가능성이 가장 높은 모델로 평가되었다.
2. 딥러닝 모델 요약
    - **MLP (다층 퍼셉트론)**은 Sklearn, PyTorch, TensorFlow 등 다양한 프레임워크를 통해 구현하여,
      동일 구조의 모델이 환경에 따라 어떻게 동작하는지를 비교하였다.
    - 딥러닝 모델은 복잡한 데이터 패턴 학습에 적합하지만,
      본 프로젝트에서는 데이터 특성이 구조화되어 있어, 머신러닝 모델과의 상호 비교 분석용으로 활용되었다.

- RandomSearch를 통해 하이퍼파라미터 탐색 전략의 효율성을 검증하였으며,
  딥러닝 모델의 구조적 이해 및 학습 안정화 기법(Batch Normalization, Dropout 등)을 적용하였다.

3. 모델 구성의 의의

   - 본 프로젝트는 단일 모델의 성능 향상보다는,
     모델 간의 구조적 차이와 데이터 처리 방식의 이해를 목표로 하였다.
   - 머신러닝과 딥러닝 모델을 함께 사용함으로써,
     "예측력(Performance)"과 "해석력(Interpretability)"을 균형 있게 고려할 수 있었다.
   - 특히 LightGBM과 CatBoost는 대규모 데이터 환경에 적합한 구조로,
     실제 비즈니스 환경에서의 고객 이탈 예측 시스템 구축에 바로 적용 가능한 모델로 평가된다.

4. 향후 적용 및 확장 방향

   - 본 연구에서 구축한 모델들은 추후 시계열 데이터 분석, 고객 행동 예측, 추천 시스템 등
     다양한 응용 분야로 확장될 수 있다.
   - 또한, **모델 해석 도구(SHAP, LIME)**를 결합하여 예측 결과에 대한 설명 가능성을 높이고,
     실무 환경에서 활용 가능한 AI 기반 고객 관리 시스템으로 발전시킬 수 있다.