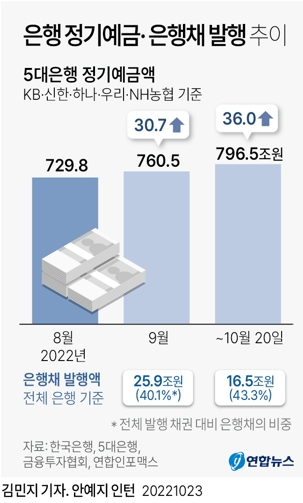
**실제 은행 데이터를 통한 기계학습 이상탐지 성능 비교 및 평가**

신규철1, 이현수1, 황세진2

1광운대학교 정보융합학부

2광운대학교 수학과

**초 록**

실험을 진행한 데이터는 은행 마케팅 캠페인에 대한 정보를 담고 있는 데이터이다. 최근들어 갈등이 고조되고 있는 국제 정세와 가파른 인플레이션으로 인해 자금 유동성이 떨어지고 있다. 또한, 기준금리 인상과 유동성 규제 비율 충족, 기업대출 자금 조달 수요 등에 따라 최근 은행들은 일제히 예·적금 금리를 크게 올리고 은행채를 대거 발행하면서 시중 자금과 채권시장 자금을 사상 최대 규모로 빨아들이고 있는 점도 유동성 감소의 주요 원인으로 꼽힌다.[[1]](#footnote-0) 정기 예금에 대한 사회적 관심이 높아짐에 따라, 각 은행은 고객 유치를 위해 다양한 시도를 하고 있다. 

본 실험에서는 데이터의 기본적인 특징을 이해하고 정상을 정의한 후, 이상탐지 알고리즘을 활용하여 해당 데이터의 정상을 탐지하는 학습모델을 구축하고자 한다. 특히, 강의를 통해 학습한 알고리즘 외 DBSCAN, Isolation forest, SGD 알고리즘을 스스로 학습하고 실험에 추가하였다.

실험에서는 머신러닝 알고리즘을 적용하여 정확도 측면의 비교 실험을 수행했으며, 실험 결과 X가 TPR X.XX%, FPR X.XX%, 그리고 F1 X.XX%만큼 좋은 성능을 보였다. 본 실험을 통해서 이상탐지를 실제 현장에 적용해 볼 수 있음을 깨닫고, 본 데이터를 기

**1. 서론**

본 실험에서 선정한 데이터는 포르투갈의 한 은행 기관의 다이렉트 마케팅 캠페인(전화 통화) 데이터셋으로 UCI Machine Learning Repository에서 가져왔다. 분류 목표는 고객의 정기예금 가입의사를 예측하는 것이다. 해당 데이터셋의 고객 정기예금 가입 비율은 **가입 : 미가입 = 15 : 85**의 비율이며, 데이터의 불균형이 관찰되었다. 따라서 이상탐지 모델을 통한 데이터 분류를 시도해 보았다.

직접 마케팅을 하여도 정기 예금에 가입하지 않는 사람이 대부분이므로, 이상탐지 측면에서 이들을 정상으로 분류하고, 정기 예금에 가입한 사람을 이상이라고 구분하여 적용하였다. 따라서 본 실험은 정기예금을 가입할 확률이 높은 사람들을 가입하지 않을 것으로 예측하는 것을 예방하여 고객 유치를 하는 것을 목적으로 한다.

**2. 데이터 기초 분석 및 전처리**

2.1 EDA

본 데이터는 입력변수가 총 16개로, 수치형 변수 7개, 범주형 변수 9개로 구성되어 있다. 출력변수는 정기예금 가입 여부에 따라 yes/no로 구분이 되어 있다. 총 관측치 개수는 45211개였으며, 출력변수에 불균형 문제가 존재하였다(15 : 85의 비율).

수치형 변수는 크게 정규분포를 따른다고 보기 어려웠으며, 왜도 측정 결과 age(나이), day(날짜)를 제외한 모든 변수의 왜도는 3 이상이었다. 범주형 변수는 각 변수의 count를 측정하였고, 특이한 점으로 month(고객과의 마지막 연락 달)의 값 중 5월의 값이 다른 달에 비해 2배 정도의 관측치 수를 차지한 것과(13766개), 10월, 9월, 3월, 12월은 상대적으로 저조한 데이터 양을 보였다 (1000개 이하).

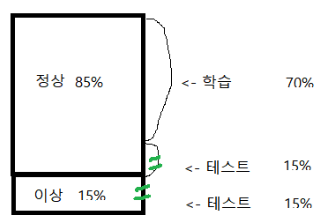
2.2 전처리

데이터에 ‘unknown’ 문자열 값이 존재하였고, 해당 실험에서는 ‘unknown’ 값을 null값으로 판단하여 데이터를 제거하였다. ‘unknown’값이 포함된 데이터 인스턴스를 삭제하였고, 특히 ‘poutcome’ 변수는 해당 값이 대부분을 차지하였기에(약 80%) 변수 자체를 제거하였다.

표준편차가 큰 수치형 변수들은 Robust Scaling을 적용하였고, 추가로 모든 수치형 변수에 MinMax Scaling을 진행하였다. 범주형 변수는 Binary Encoding, One Hot Encoding의 인코딩 기법을 이용하여 적절하게 수치형 변수로 변환해주었다.

마지막으로, 변수 간 선형 상관성을 조사하여 선형 상관성이 유의미한 변수들을 제거하였다(상관계수 0.3이상).

정상과 이상 비율이 85:15 에서 이상 4513개와 동일하게 정상에서 랜덤 샘플링하여 4513개를 뽑아 테스트 데이터로 만들었다.

즉, 

**3. 추가 적용 알고리즘**

본 실험에서 적용해본 추가적인 알고리즘은 다음과 같다. **DBSCAN**은 Clustering 알고리즘 중 하나로, 밀도 기반 Clustering을 진행한다. 비교적 다양한 형태의 데이터 분포에 효율적이라고 알려져 있다. 특정 거리를 지정하여 관측치 간 이웃을 정하며, parameter로는 이웃을 지정하는 거리(eps), 군집을 정의하기 위한 최소 관측치 수(minPts)가 있다. 관측치는 Core, Border, Outlier로 분류되며, 해당 보고서에서는 Outlier로 분류된 관측치들을 정기예금에 가입하였다고 판단하여 실험을 진행하였다.

**Isolation Forest** : 데이터셋을 의사결정나무 형태로 표현하여, 정상 데이터를 분리하기 위해서는 의사결정나무의 깊은 곳을 확인해야 하고, 이상치는 의사결정나무 상단부에서 분리할 수 있다는 것을 이용한 알고리즘이다. 해당 특성을 이용해 의사결정나무를 몇 회 타고 내려가야 분리되는가를 기준으로 정상과 이상을 분리한다.

**1-svm with gradient descent(SGD)** :

SVM은 목적식의 optimization 과정에서 gradient descent 를 통해 optimal solution을 찾는다. 그런데 알고리즘 특성상 optimization 과정에서 많은 계산량이 요구되므로, 이를 보완하기 위해 SGD(stochastic gradient descent) 방식을 적용한다. SGD는 전체 데이터 중 일부인 mini-batch를 사용하여 loss function을 계산하는데, 이 점 덕분에 훨씬 빠른 계산 속도를 보인다. 오히려 빠른 속도 덕분에 BGD보다 좋은 local optima를 찾을 가능성도 올라간다는 장점이 있다.

**Parzen Window Density Estimation**은 Mixture of Gaussian 의 가정이 완화된 것으로 특정한 분포 가진다는 가정 없이(non-parametric) 주어진 데이터로부터 주변부의 밀도를 추정하는 방식이다.

가우시안 커널을 사용한 커널 density Estimation 이 Parzen Window

데이터 하나하나가 가우시안 모달 개수

오버 피팅 생길 수 있다는 단점이 있다.

**5. 실험 결과 및 해석**

|  | **Threshold** | **FRR** | **FAR** | **F1 Score** |
| --- | --- | --- | --- | --- |
| **GDE** | **상위**  **15%** | **0.1354** | **0.7266** | **0.3882** |
| **MOG** | **상위**  **15%** | **0.117** | **0.7682** | **0.3437** |
| **LOF** | **상위**  **5%** | **0.0567** | **0.9617** | **0.07** |
| **Parzen Window** | **상위**  **15%** | **0.0682** | **0.7456** | **0.3847** |
| **Auto**  **encoder**  **(MLPRegressor)** | **상위**  **15%** | **0.1387** | **0.6911** | **0.4268** |
| **Auto**  **encoder**  **(tensorflow)** | **상위**  **15%** | **0.1431** | **0.7188** | **0.3948** |
| **kNN** | **상위 11%** | **0.0441** | **0.6409** | **0.4421** |
| **k-Means** | **상위 15%** | **0.0906** | **0.5028** | **0.4905** |
| **DBSCAN** | **상위 15%** | **0.1517** | **0.821** | **0.1733** |
| **Isolation Forest** | **상위 15%** | **0.1373** | **0.7758** | **0.2212** |
| **PCA** | **상위 15%** | **0.1373** | **0.7758** | **0.2212** |
|  |  |  |  |  |

# 본 실험에서 정기예금을 가입할 확률이 높은 사람들을 가입하지 않을 것으로 예측하는 것을 예방하여 고객 유치를 하는 것을 목적으로 한다. 따라서 실험 목적에 맞는 지표인 FAR가 낮은 모델이 성능이 좋은 모델이라고 판단하였다. 해당 기준에 따르면 성능이 좋다고 판단되는 모델은 1.~~~~ 2.~~~~ 3.~~~~으로 확인되었다.

poutcome

1. https://www.yna.co.kr/view/AKR20221022044000002?input=1195m [↑](#footnote-ref-0)