**[Project #1] Detecting Card Transactions Fraud with Machine Learning**

**(Classification; Anomaly Detection)**

**주제:** 카드 사기로 인한 손실과 사기 탐지 시스템 운영 비용의 최소화를 위한 머신러닝 모델 개발.

**목표:** 데이터를 변수의 확률/통계, 가시화 등 다각적으로 분석하여 머신러닝 모델을 제안하고 검증함.

**1. 내용**

* 데이터 분석, 가시화를 통한 분석
* 선형 분류 모델의 이해 및 구현
* PCA, EM 군집화 algorithm 등을 활용하여 feature selection 혹은 추론
* Logistic regression, tree 구조, 모델의 설명성 등을 활용하여 feature selection 혹은 추론
* 결측값 등의 데이터 전처리
* 신경망 (Neural Network)구조는 활용 금지.

**2. 보고서**

* 1. 개요, 2. 방법론 및 구현 방법, 3. 실험 결과 및 분석, 4. 결론 형태로 기술 (표지 포함 15p 이하)
* 가설 및 분석을 포함하여 개발 과정의 단계를 기술 (e.g., 가시화, 상관관계 분석 등)
* 분석에 기반하여 제안한 알고리즘 별 정당성 기술 (trial-and-errors, trouble shooting)
* 비용 외 다양한 평가 지표를 활용한 모델 성능을 종합적으로 평가   
  (e.g., AUC-ROC, F1-score, Precision, Recall 등)

**3. 비용 가정 (참고)**

**a)** 사기 거래 미탐지 비용:   
- 거래 금액의 100%   
- 예: 100만원 사기 거래 미탐지 시 100만원 손실

**b)** 정상 거래 오탐지 비용:   
- 건당 고정 비용: 5,000원 (고객 응대, 조사, 해명 비용)   
- 거래 금액의 0.1% (거래 지연 또는 취소로 인한 기회 비용)   
- 예: 100만원 정상 거래 오탐지 시 5,000원 + 1,000원 = 6,000원 손실

**c)** 모델 운영 비용:   
- 검사 건당 100원 (컴퓨팅 리소스, 유지보수 비용 등)

* 총 비용은 상기된 비용의 총합(**a + b + c**)이며 이외의 비용은 고려하지 않음.

**4. 데이터 및 구현 언어 (language):**

* LMS에 첨부된 파일(‘**2024\_DS\_project.csv’**)을 참조할 것
* Python 또는 C++을 사용할 것

**5. 제출 프로그램 구현 및 명세 (형식을 반드시 엄수):**

* 제출은 아래의 파일을 ‘**데이터사이언스\_##.zip**’로 묶어 LMS(lms.ssu.ac.kr) 과제 란에 한 명이 제출할 것 (##은 자신의 팀 번호로, 두 자리 자연수임. e.g., 데이터사이언스\_01.zip)

◼ 사용자들의 2019년 데이터가 담긴 거래 파일에 예측값 추가 (‘submission\_exmaple.csv’를 참조)

◆ 파일명: ‘submission\_##.csv

◼ 모델의 학습을 위한 코드

◆ 파일명: source\_##.py 또는 source\_##.cpp 또는 source\_##.zip

◼ 소스코드 설명서

◆ 파일명: source\_spec\_##.pdf

◆ 작성한 코드의 개요를 클래스 또는 기능 별로 자유롭게 기술

◆ 소스코드 설명서의 첫 페이지는 보고서의 양식 첫 페이지와 동일하게 사용

◼ 보고서

◆ 파일명: report\_##.pdf

◆ 주어진 양식 (보고서 양식.docx)을 참조하여, 아래의 구성으로 기술할 것

▸ “개요 - 방법론 및 구현 방법 - 실험 결과 및 분석 – 결론” 중심으로 자유로이 기술

▸ 제안한 알고리즘 별 정당성 기술

◼ 발표자료 (ppt, pdf)

**6. 평가 방법**:

- 성실도 (trial-and-errors, algorithm contents 활용도, 발표 등) (40%)

- 제안한 방법의 차별성 및 정당성 발표 (30%)

- 상대적 성능 순위 평가 (**3.**의 가정 기준 총 비용) (10%)

- 보고서, 소스코드 및 결과 (20%); **구성원 기여도 보고서 표지에 기재**

◼ Copy & paste 체크

◼ 결과의 유효성 체크 (e.g., 테스트 데이터에서의 유효성, 분석과 결과의 인과성 등)

◼ 참조한 source code가 있을 시, citation 필수

**제출 기한:**

**2024.10.27 (일) 23:59**

**발표일:**

**2024.10.28-29 (월-화) 오프라인 수업시간 (303호)**

**발표 순서는 10월 28일 수업 시간에 랜덤 배정**

**유의 사항:**

* **컴파일 및 실행이 안되면 0점 처리**
* **보고서, 소스코드 등의 Copy & Paste는 0점 처리**
* **기간을 반드시 지킬 것 (늦은 제출은 불허하며, 0점 처리)**
* **자신의 팀이 제안하는 방법의 핵심부분을 중점적으로 이야기할 수 있도록 합니다.**
* **발표 시간은 최대 10분입니다 (Q&A 포함).**