

금융 이상 거래 탐지 파이프라인 구축 보고서

1. 프로젝트 개요

은행 거래 데이터를 기반으로 이상 거래를 자동으로 탐지하는 파이프라인을 구축함.

전처리, 모델 적용, 결과 저장까지의 전 과정을 자동 실행 가능하도록 설계함.

- 출처: [Kaggle - Bank Transaction Dataset for Fraud Detection]
- 구성: 총 2,512건의 거래 데이터와 16개 변수로 구성됨

2. 분석 및 파이프라인 구성 단계

2-1. 데이터 전처리

- 공백 포함 컬럼명을 밑줄(_)로 변경해 처리 용이성 확보
- `TransactionDate`, `PreviousTransactionDate`를 `datetime` 형식으로 변환
- `TransactionAmount` 등 5개 수치형 변수는 `StandardScaler`로 정규화
- 예측 결과 매칭을 위해 `TransactionID`는 유지

2-2. 모델 구성 및 예측

- `IsolationForest`를 사용해 이상 거래 탐지 모델 구성
- 정규화된 수치형 변수만 입력, `TransactionID`는 제외
- `contamination=0.1` 설정으로 전체 거래 중 약 10%를 이상 거래로 간주
- 예측 결과: 정상 2,260건 / 이상 252건

2-3. 결과 저장 및 필터링

- 이상 거래로 분류된 252건만 필터링해 별도 저장
- 원본 거래 데이터와 `TransactionID` 기준으로 병합
- 전체 거래에 대해 이상 여부(0/1)를 포함한 파일도 함께 저장

2-4. 자동 실행 구조화

- 전체 과정을 함수 단위로 분리하고 `main()` 블록으로 자동 실행 구조 구성
- `load → preprocess → model predict → save` 순서로 실행되도록 설계
- 파일 경로만 지정하면 한 번에 결과 도출 가능
- 파이썬 스크립트로 저장해 CLI나 작업 스케줄러에 연동 가능

3. 결과 요약

- 총 2,512건의 거래 데이터를 기반으로 이상 거래 탐지 파이프라인을 구축함
- `IsolationForest` 모델을 통해 **252건의 이상 거래**를 탐지 (약 10%)
- 탐지 기준은 IsolationForest가 통계적으로 다른 패턴을 탐지한 결과로 분류됨
- 전체 과정은 파이썬 스크립트로 자동화되어 반복 실행 가능