[LG U+ WHY NOT SW CAMP]

군집화&분류
모델 정의서 & 성능평가서

소비 행동 패턴 기반 나만의 AI 금융 코치

팀명 : Moni

팀원 : 박시하(팀장), 박소현, 김단하, 홍예은

1. 모델 정의서 (Model Definition Document)

1.1. 개요 (Overview)

- 모델명: 소비 패턴 기반 고객 유형 군집화 및 분류 모델 (v1.0)

- 개발 목적:
  - 고객들의 소비 내역을 분석하여 유사한 성향을 가진 군집(Cluster)으로 정의 (군집화)
  - 신규 고객 유입 시, 별도의 복잡한 분석 없이 즉시 해당 고객의 유형(Group)을 판별하여 맞춤형 서비스를 제공 (분류)

- 적용 알고리즘 **(Hybrid Approach):**
  - **Step 1 (**유형 정의**):** K-Means Clustering
  - **Step 2 (**유형 판별**):** Random Forest Classifier

- 개발 환경: Python 3.12, Scikit-learn, Pandas, NumPy

1.2. 데이터 구조 및 전처리 (Data Structure & Processing)

원천 데이터(Transaction)를 모델 학습용 데이터(User Profile)로 변환하는 과정

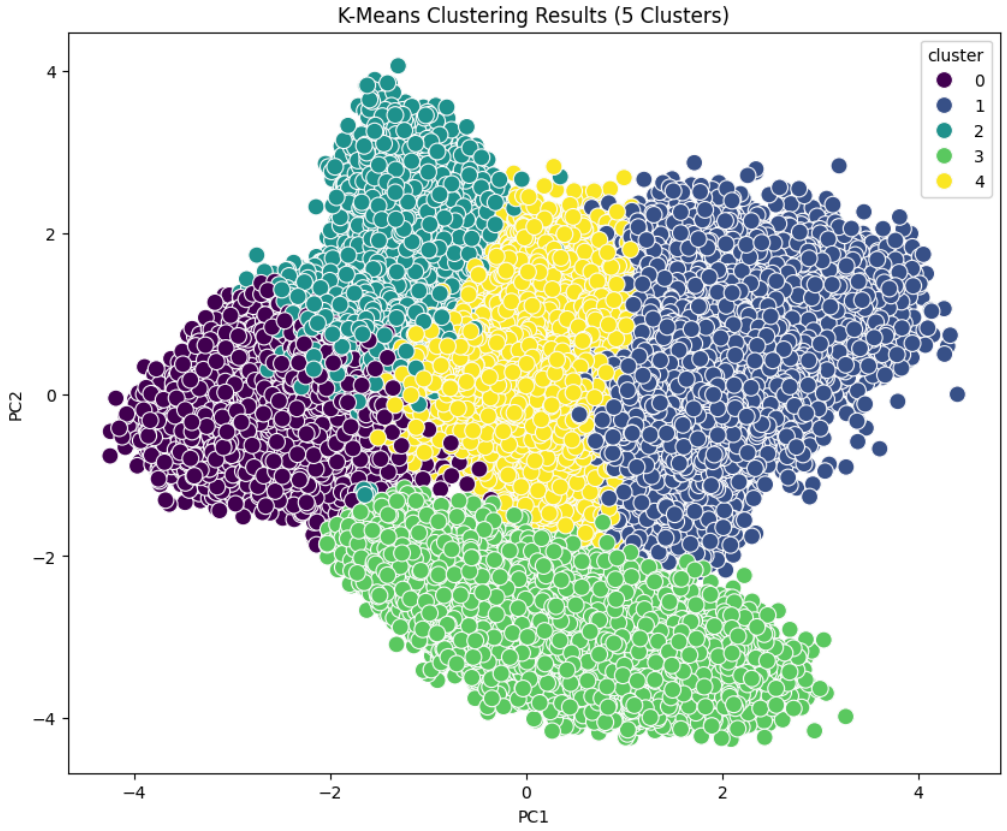
- 원천 데이터: [사용자ID, 거래일시, 가맹점명, 거래금액]

- 파생 변수 **(Feature Engineering):** 사용자 1명당 1개의 행(Row)을 가지도록 집계
  - *saving_rate(*저축 비율*):* ('savemoney' 카테고리 거래금액) / (총 지출)
  - *remain_ratio(*월말 잔액 비율*):* (월말 잔액 평균) / (총 지출)
  - *invest_ratio(*투자 지출 비율*):* (총 투자 지출) / (총 지출)
  - *spend_volatility(*지출 변동성*):* (지출 표준편차) / (지출 평균)
  - *peak_spend_months(*지출 폭발 월*):* 특정 월에서 지출이 평균의 1.5배 이상 폭발한 횟수
  - *fixed_cost_ratio(*고정비 비율*):* (고정비) / (총 지출)

- 최종 입력 데이터 **(Input X):** 위에서 생성된 사용자별 특징 벡터 (User Feature Vector)

- 타겟 변수 **(Target y):** 군집화 과정을 통해 할당된 군집 라벨 **(Cluster)** (예: 0, 1, 2, 3, 4.)

1.3. 모델 아키텍처 **(Model Architecture)**



[Phase 1: 군집화 **(Clustering)]**

- 목적**:** 타겟 라벨(y) 생성
- 설정**:**
  - 군집 수 결정 방법: K=5개로 설정
- 정의된 군집 특성**:**
  - *Group 0(*절약형*):* 저축률과 잔액비율이 모두 높고 지출 변동성이 낮음
  - *Group 1(*목표요정형*):* 지출 정점 개월 수 가 가장 빈번, 고정비 비율 높음, 지출변동성 큼
  - *Group 2(*안정형*):* 저축율은 낮지만 잔액비율이 매우 높음
  - *Group 3(*진격의 투자형*):* 투자 비율이 압도적으로 높음
  - *Group 4(YOLO*형*):* 지출변동성이 다소 높고, 저축이나 잔액비율이 낮음

[Phase 2: 분류 **(Classification)]**

- 목적**:** 신규 사용자의 그룹 예측
- 모델**:** Random Forest Classifier
- 주요 하이퍼파라미터**:**
  - n_estimators: 200 (트리 개수)
  - max_depth: 7 (과적합 방지)
  - min_samples_split: 60

- min_samples_leaf: 30
- ccp_alpha: 0.004
- n_jobs: -1

2. 성능 평가서 (Performance Evaluation Document)

2.1. 평가 개요

- 평가 일시: 2025.12.16
- 데이터셋 구성:
  - 전체 데이터 [120,000]건 중 Train(70%) / Test(30%) 분할
- 평가 지표:
  - 군집화 품질:
    - Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score
  - 분류 정확도:
    - Precision, Recal, F1-Score:

2.2. 정량적 성능 결과

2.2.1 군집화(Clustering) 품질 평가

모델	**Silhouette Score**	**Calinski-Harabas Score**	**Davies-Bouldin Score**
K-Means	0.3162	58925.5248	1.1275

- Silhouette Score(실루엣):데이터가 자신의 군집내에서 얼마나 가깝게 뭉쳐있고, 다른 군집과 얼마나 멀리 떨어져 있는지를 나타내는 지표
- Calinski-Harabasz Score(칼린스키-하라바즈): 군집간의 거리(분산)는 멀고, 군집내의 거리(분산)는 가까울수록 높아지는 지표
- Davies-Bouldin Score(데이비즈-볼딘): 내 군집과 가장 비슷한(가까운) 다른 군집과의 유사도

2.2.2 분류(Random Forest) 모델 성능

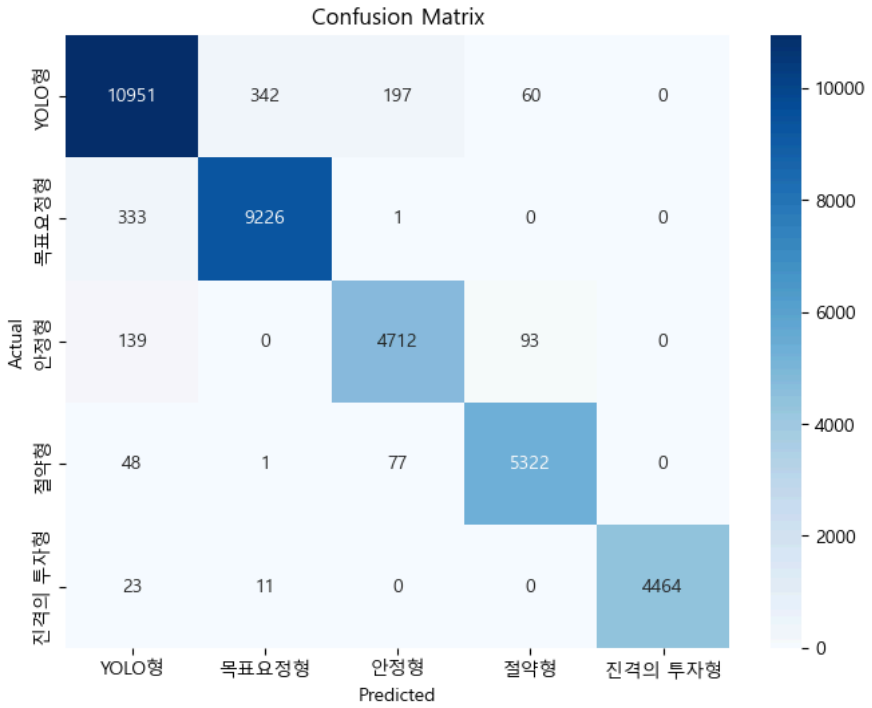
군집화된 라벨을 Random Forest가 얼마나 잘 학습하고 맞추는지에 대한 평가

Group (군집)	**Precision (정밀도)**	**Recall (재현율)**	**F1-Score**
Group 0 (절약형)	0.9721	0.9769	0.9745
Group 1 (목표요정형)	0.9630	0.9651	0.9641
Group 2 (안정형)	0.9449	0.9531	0.9489
Group 3 (진격의 투자형)	1.000	0.9924	0.9962

Group 4 (YOLO형)	0.9528	0.9481	0.9504
Weighted Avg	**0.9632**	**0.9632**	**0.9632**

- Precision(정밀도):모델이 True라고 분류한 것 중에서 실제 정답인 비율
- Recall(재현율): 실제 정답인 것 중에서 모델이 True라고 찾아낸 비율
- F1-Score: Precision과 Recall의 조화 평균

2.2.3 Confusion Matrix(혼동행렬)



2.3. 주요 변수 중요도 (Feature Importance)

Random Forest 모델이 사용자의 유형을 판별할 때 어떤 특성을 중요하게 보았는지 분석한 결과입니다.

순위	변수명	중요도
1	**remain_ratio**	0.199334
2	**saving_rate**	0.195727
3	**peak_spend_months**	0.192015
4	**invest_ratio**	0.173093
5	**spend_volatility**	0.154333

6	**fixed_cost_ratin**	0.085498

2.4. 종합 의견

- 본 모델은 비지도 학습을 통해 고객을 5개의 유의미한 그룹으로 정의하였으며, 이를 기반으로 한 Random Forest 분류 모델은 약 **96.3%**의 높은 정확도를 보임.
- 데이터가 추가되면 주기적인 재군집화(Re-clustering)를 통해 최신 트렌드를 반영해야 함.