# Version 4 - Generated Data + SMOTE + Optuna - Feature Engineering Process

เอกสารอธิบายกระบวนการสร้าง feature แบบอ่านเข้าใจง่าย พร้อม code/artifact ที่เกี่ยวข้อง

## Process Overview

- generate synthetic order/return data เพื่อทดลองเพิ่มปริมาณข้อมูล
- clean missing/outlier/duplicate และสร้าง clean_dataset_v4_generated.csv
- ทำ EDA category/channel/payment/correlation
- สร้าง feature จำนวนมาก รวม interaction/encoded features
- ใช้ SMOTE สำหรับ imbalance และ Optuna สำหรับ tuning
- ประเมิน XGBoost/LightGBM/RandomForest/Logistic และเลือก XGBoost_SMOTE_Optuna

## Code / Artifact Reference

| Code / Artifact | Path |
|---|---|
| clean_dataset_v4.py | docs\version 4\scripts\clean_dataset_v4.py |
| run_v4_generated_end_to_end_pipeline.py | docs\version 4\scripts\run_v4_generated_end_to_end_pipeline.py |

## Input / Output

| Stage | Artifact |
|---|---|
| Input clean data | docs\version 4\data\processed\clean_dataset_v4_generated.csv |
| Engineered dataset | docs\version 4\data\features\df_engineered_v4_generated.csv |
| Used feature list | docs\version 4\reports\model_evaluation\v4_used_features.csv |

## Feature Engineering Logic

- เริ่มจาก clean dataset ของ version นั้น แล้วสร้าง numeric/encoded/interaction features
- ใช้ train/test split เดียวใน version เพื่อเปรียบเทียบ model อย่างยุติธรรม
- drop leakage/identifier fields ก่อน train
- ประเมินด้วย Accuracy, Recall, Precision, F1, AUC และ Cost Matrix

## Performance Result

| Accuracy | Recall | Precision | F1 | AUC | Cost | Rating | Used Features |
|---|---|---|---|---|---|---|---|
| 83.45% | 46.39% | 45.00% | 45.69% | 85.38% | 31,650 | C | 180 |

## Decision

ใช้เป็น benchmark ของ generated data แต่ยังไม่เลือก final เพราะ V2 ให้ balance ดีกว่า

## Important Notes

- เอกสารนี้ไม่ทับ model artifact เดิม แต่รีเจน documentation ให้ตรงกับสถานะล่าสุดของโปรเจ็กต์

- ถ้า metrics/model เปลี่ยน ให้รัน scripts/generate_current_project_docs_v1_to_v4.py ใหม่เพื่อ sync เอกสาร
- สำหรับ production จริงควร validate ด้วยข้อมูลจริงแบบ time-based holdout ก่อน deploy