Data Dictionary – Flat User Credit Scoring Dataset (Labeled)

This document describes the final flat, labeled dataset used for training and evaluating AI-based Credit Scoring models in the Vietnamese fintech context.

1. Dataset Overview

Dataset name: flat_user_credit_scoring_labeled.csv
Grain: One row per user
Purpose: Model training and evaluation for credit risk classification.

2. User Identifier and Demographics

- user_id: Anonymous user identifier, randomly generated.
- user_age: Age of the user, ranging from 18 to 65.
- user_region: Region of residence (HN, HCM, DN, CT, OTHER).

3. Telco Aggregated Features

- telco_account_age_days: Number of days the telecom account has been active.
- telco_avg_revenue_mean: Average monthly telecom spending.
- telco_avg_revenue_std: Volatility of telecom spending.
- telco_recharge_count_mean: Average monthly recharge count.
- telco_recharge_amount_mean: Average monthly recharge amount.
- telco_mobile_data_mb_mean: Average monthly mobile data usage.
- has_telco_data: Flag indicating availability of telco data.

4. Academic Features

- edu_highest_level: Highest education level achieved.
- edu_gpa_band: GPA category band.
- edu_graduation_status: Graduation status.
- edu_institution_tier: Quality tier of the institution.
- edu_major_group: Broad field of study.
- has_academic_data: Flag indicating availability of academic data.

5. E-Wallet Aggregated Features

- wallet_txn_count: Total number of wallet transactions.
- wallet_avg_amount: Average transaction amount.
- wallet_max_amount: Maximum transaction amount.

- wallet_total_amount: Total transaction amount.
- wallet_large_txn_ratio: Ratio of large transactions above 2 million VND.
- wallet_failure_rate: Ratio of failed transactions.
- has_ewallet_data: Flag indicating availability of e-wallet data.

6. Target Label Definition

target: Binary credit risk label.
0 = Good customer (low risk).
1 = Bad customer (high risk).

The label is synthetically generated using stability, capacity, and behavioral risk indicators. No future information is used, and the bad rate is approximately 35 percent.