Data Dictionary – Credit Scoring Dataset (Vietnam Context)

This document describes the structure, meaning, and data generation patterns of the synthetic datasets used for training a Credit Scoring model in the Vietnamese fintech context.

1. base_users.csv

Purpose: Base population table representing all potential users. All other datasets are joined to this table using **user_id**.

- Grain: 1 row per user
- user_id: Anonymous user identifier (random string, non-sequential)
- user_age: Age of user (18–65)
- user_region: Living region (HN, HCM, DN, CT, OTHER)

2. telco_features.csv

Purpose: Proxy for financial stability and lifestyle through telecom usage behavior. Coverage is approximately 85% of users.

- Grain: 1 row per user
- telco_account_age_days: Number of days the telecom account has been active (0.5–10 years)
- telco_avg_revenue_month_6–9: Monthly telecom spending (80k–600k VND, right-skewed)
- telco_outgoing_call_minutes_month_6–9: Outgoing call duration (gamma distributed)
- telco_incoming_call_minutes_month_6–9: Incoming call duration
- telco_recharge_count_month_6–9: Number of top-ups per month
- telco_recharge_amount_month_6–9: Total recharge amount (50k–300k denominations)
- telco_mobile_data_mb_month_6–9: Mobile data usage (~1–20GB/month)

3. academic_features.csv

Purpose: Proxy for human capital and long-term earning potential. This dataset does not contain internal academic records and reflects self-declared or publicly available information.

- Grain: 1 row per user (coverage ~50%)
- edu_highest_level: Highest education level (high school, college, university, postgraduate)
- edu_gpa_band: GPA band (average, good, excellent)
- edu_graduation_status: Graduated or ongoing
- edu_institution_tier: Institution quality tier (tier_1 is rare)
- edu_major_group: Broad field of study

4. ewallet_transactions.csv

Purpose: Raw transactional data reflecting day-to-day spending behavior. Each user can have multiple transactions.

- Grain: 1 row per transaction (coverage ~60% of users)
- wallet_transaction_id: Unique transaction identifier (UUID)
- wallet_transaction_datetime: Timestamp within the last 12 months
- wallet_transaction_category: Spending category (food, transport, shopping, bill, education)
- wallet_transaction_amount_vnd: Transaction amount (20k–10M VND, with long-tail)
- wallet_payment_method: Linked bank or wallet balance
- wallet_transaction_status: Success (~98%) or failed

5. Cardinality and Join Logic

base_users (1) → telco_features (0..1), academic_features (0..1), ewallet_transactions (0..N). E-wallet transactions must be aggregated to user level before joining into a flat table.