**Comprehensive Guide to Your Dataset Columns**

The columns can be grouped into three categories:

1. **Base Profile Features:** Static or slowly changing characteristics of the user.
2. **Transactional & Behavioral Features:** Features derived from simulated user activity.
3. **Calculated Risk & Target Features:** The final risk scores and the loan default label.

**1. Base Profile Features**

These columns define *who* the user is in a broad sense.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **What it Signifies** | **Limits / Typical Range** |
| **user\_id** | A unique number to identify each person. | Integer from 0 to 49,999. |
| **age** | The applicant's age in years. | Integer from 18 to 65. |
| **income\_tier** | A simplified category of the user's financial standing. **(0=Weak, 1=Mid, 2=Strong)**. Influences how much income they are simulated to receive. | Integer: 0, 1, or 2. |
| **employment\_tenure** | How long the user has been employed, in months. A proxy for job stability. | Float, from 0 to (age - 20) \* 12. |
| **device\_tier** | The quality of their smartphone. A proxy for wealth/digital fluency. **(0=High-end, 1=Mid, 2=Low-end)**. | Integer: 0, 1, or 2. |
| **app\_diversity** | The number of apps on their phone. A proxy for digital engagement. | Integer, typically from ~5 to ~50. |
| **clickstream\_volatility** | A score representing how erratic their digital behavior is. High volatility can be a sign of instability. | Float, from 0.0 to 1.0. |
| **peer\_default\_exposure** | A score representing the proportion of their social circle that has defaulted on loans. A social risk indicator. | Float, from 0.0 to 1.0. |
| **financial\_coping\_ability** | A self-reported score of how well they can handle a financial emergency. A psychometric measure of resilience. | Integer from 1 to 5. |
| **asset\_diversity** | A count of the different types of assets they have (e.g., savings, vehicle, etc.). A measure of wealth. | Integer, typically from 0 to 5. |
| **earner\_density** | The number of earning individuals in their household. A measure of household financial stability. | Float, typically from 1.0 to 4.0. |
| **urbanization\_score** | A score representing how developed or urbanized their location is. Higher score means more urban. | Float, from 0.0 to 1.0. |
| **local\_unemployment\_rate** | The local unemployment rate in their area. A key macroeconomic risk factor. | Float, typically from 0.0 to 0.4. |

**2. Transactional & Behavioral Features**

These columns are calculated from the simulated UPI transaction history.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **What it Signifies** | **Limits / Typical Range** |
| **bnpl\_repayment\_rate** | "Buy Now, Pay Later" repayment consistency. A direct measure of financial discipline for a specific credit product. | Float, from 0.0 to 1.0. (0 if not a BNPL user). |
| **debt\_burden** | The ratio of monthly loan payments to monthly income. A critical measure of over-indebtedness. | Float. 0 for no debt. Can go above 1.0 for highly indebted users. |
| **utility\_payment\_ratio** | A score representing the consistency of paying utility bills on time. A strong indicator of reliability. | Float, from 0.0 to 1.0. |
| **ott\_spending\_tier** | Categorized spending on OTT services (Netflix, etc.). **(0=Low, 1=Mid, 2=High)**. | Integer: 0, 1, or 2. |
| **food\_delivery\_tier** | Categorized spending on food delivery. **(0=Low, 1=Mid, 2=High)**. | Integer: 0, 1, or 2. |
| **ride\_hailing\_tier** | Categorized spending on ride-hailing services. **(0=Low, 1=Mid, 2=High)**. | Integer: 0, 1, or 2. |
| **skill\_spend** | Total amount spent on upskilling courses (Coursera, Udemy) over the observation period. A signal of career growth ambition. | Float, from 0 upwards. |

**3. Calculated Risk & Target Features**

These are the final outputs of your data engineering, which the model will learn from, and the final answer key.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **What it Signifies** | **Limits / Typical Range** |
| **...\_risk** (e.g.,debt\_risk) | A standardized risk score **(0=Low, 1=Mid, 2=High)** for a specific feature, calculated based on thresholds defined in your logic. | Integer: 0, 1, or 2. |
| **total\_risk** | The final, combined risk score for a user, calculated by summing up all the individual ...\_risk scores. **This is a key summary feature.** | Float, no hard limits but will have a practical range based on your number of risk features. |
| **default\_prob** | The calculated probability that a user will default, derived from their total\_risk score via a logistic function. | Float, strictly between 0.0 and 1.0. |
| **loan\_default** | **The Target Variable.** The final "ground truth" answer. **(0 = Did NOT Default, 1 = DID Default)**. This is what your machine learning model will try to predict. | Integer: 0 or 1. |