**. Data Source**

This dataset is a synthetic version inspired by the original [Credit Risk dataset on Kaggle](https://www.kaggle.com/datasets/laotse/credit-risk-dataset) and enriched with additional variables based on [Financial Risk for Loan Approval data](https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval). SMOTENC was used to simulate new data points to enlarge the instances. The dataset is structured for both categorical and continuous features.

**2. Metadata**

The dataset contains 45,000 records and 14 variables, each described below:

| **Column** | **Description** | **Type** |
| --- | --- | --- |
| person\_age | Age of the person | Float |
| person\_gender | Gender of the person | Categorical |
| person\_education | Highest education level | Categorical |
| person\_income | Annual income | Float |
| person\_emp\_exp | Years of employment experience | Integer |
| person\_home\_ownership | Home ownership status (e.g., rent, own, mortgage) | Categorical |
| loan\_amnt | Loan amount requested | Float |
| loan\_intent | Purpose of the loan | Categorical |
| loan\_int\_rate | Loan interest rate | Float |
| loan\_percent\_income | Loan amount as a percentage of annual income | Float |
| cb\_person\_cred\_hist\_length | Length of credit history in years | Float |
| credit\_score | Credit score of the person | Integer |
| previous\_loan\_defaults\_on\_file | Indicator of previous loan defaults | Categorical |
| loan\_status (target variable) | Loan approval status: 1 = approved; 0 = rejected | Integer |

**3. Data Usage**

The dataset can be used for multiple purposes:

* **Exploratory Data Analysis (EDA)**: Analyze key features, distribution patterns, and relationships to understand credit risk factors.
* **Classification**: Build predictive models to classify the loan\_status variable (approved/not approved) for potential applicants.
* **Regression**: Develop regression models to predict the credit\_score variable based on individual and loan-related attributes.

**Mind the data issue from the original data, such as the instance > 100-year-old as age.**

This dataset provides a rich basis for understanding financial risk factors and simulating predictive modeling processes for loan approval and credit scoring.