**Credit Risk Analysis Dataset**

**About the Dataset**

This dataset is a synthetic version inspired by the original Credit Risk dataset on Kaggle, enhanced with additional variables focused on financial risk for loan approvals. To expand the instances, SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) was used, creating a dataset suitable for classification and regression modeling with both categorical and continuous features.

**Key Points:**

* **Records**: 45,000
* **Variables**: 14
* **Types of Features**: Categorical and Continuous

**Dataset Structure**

The dataset contains 14 key variables related to applicant demographics, financial status, loan details, and credit history, with a focus on assessing credit risk for loan approval.

|  |  |  |
| --- | --- | --- |
| Column | Description | Type |
| person\_age | Age of the applicant | Float |
| person\_gender | Gender of the applicant | 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) | Loan approval status: 1 = approved; 0 = rejected | Integer |

**Note**: The dataset includes instances with unusual values, such as ages above 100, due to synthetic data generation.

**Potential Use Cases**

This dataset supports various analyses and modeling approaches for credit risk and loan approval:

1. **Exploratory Data Analysis (EDA)**:
   * Analyze distributions, correlations, and patterns in credit risk factors.
   * Identify key trends among demographic and loan variables.
2. **Classification Modeling**:
   * Build predictive models for loan approval (loan\_status as target).
   * Train classifiers to identify approved vs. rejected loan applications.
3. **Regression Analysis**:
   * Predict credit\_score based on individual and loan-related features.

**Important Considerations**

* **Data Quality**: The synthetic data may include implausible values, such as applicants over 100 years old, reflecting limitations in the original data source.
* **Usage Scope**: This dataset serves as a foundation for educational purposes, training, and testing machine learning models in financial risk prediction.