Credit Risk Assessment Data Dictionary

LendSmart - Loan Application Dataset

File: `credit_risk_data.csv` **Observations**: 2,500 loan applications **Time Period**: 2022-01-01 to 2024-12-29 (3 years) **Good Loan Rate**: 73.44% (1,836 good loans) **Default Rate**: 26.56% (664 defaults)

Variable Descriptions

Identification Variables

Variable	Type	Description	Values
`application_id`	String	Unique application identifier	APP_001 to APP_2500
`application_date`	Date	Loan application submission date	YYYY-MM-DD format
`loan_amount`	Float	Requested loan amount	$5,000 - $500,000

Financial Indicators

Income and Employment

Variable	Type	Scale	Description
`annual_income`	Float	$15,000-$149,930	Annual gross income
`employment_years`	Float	0-19.3	Years in current employment
`job_stability_score`	Float	0.011-0.999	Employment stability indicator (higher = more stable)

Credit History

Variable	Type	Scale	Description
`credit_score`	Integer	334-850	FICO credit score
`credit_utilization`	Float	0.004-0.998	Credit utilization ratio
`payment_history_score`	Float	0.029-1.000	Payment history quality (higher = better)
`open_credit_lines`	Integer	0-11	Number of open credit accounts

Debt and Assets

Variable	Type	Scale	Description
`debt_to_income_ratio`	Float	0.009-0.979	Total debt payments / gross income
`savings_ratio`	Float	0.000-0.893	Savings / annual income

Variable	Type	Scale	Description
asset_value	Float	$551-$1,000,000	Total asset value (home, investments, etc.)

Demographic Variables

Variable	Type	Description	Values
age	Integer	Applicant age	18-75
education_level	Categorical	Highest education completed	High School (444), Associates (596), Bachelors (834), Masters (442), Doctorate (184)
marital_status	Categorical	Marital status	Single (580), Married (1311), Divorced (439), Widowed (170)
residential_stability	Float	Years at current address	0.0-16.4

Outcome Variable

Variable	Type	Description	Values
loan_status	Binary	Loan performance outcome	0 = Good (paid in full), 1 = Default (90+ days past due)

Data Quality Notes

Completeness

- **Missing Data**: No missing values in the dataset (100% complete data)
- **Data Quality**: All 2,500 observations have complete information across all 18 variables

Variable Distributions

- **Credit Scores**: Approximately normal distribution (mean = 681.7, median = 700)
- **Income**: Right-skewed distribution (mean = $67,708, range: $15,000-$149,930)
- **Loan Amounts**: Wide range from small personal loans to large mortgages (mean = $155,716, range: $5,000-$500,000)
- **Default Rate**: 26.56% overall (664 defaults out of 2,500 applications), varies by credit score segments

Group Differences

- **Default vs Good Loans**: Significant differences in credit scores, debt ratios, employment stability
- **Covariance Structure**: Different covariance matrices between groups (heteroscedasticity)
- **Multivariate Normality**: Approximately normal within groups after transformation

Analytical Considerations

Discriminant Analysis Suitability

- **Sample Size**: N=2,500 exceeds minimum requirements (Good: 1,836, Default: 664)
- **Class Balance**: 73.44% good, 26.56% default (acceptable imbalance)
- **Variable Types**: Mix of continuous and categorical predictors (15 continuous, 2 categorical)
- **Group Separation**: Clear separation expected based on financial indicators

Recommended Analyses

1. **Exploratory Data Analysis**: Examine group differences, correlations, distributions
2. **Assumption Testing**: Multivariate normality, homoscedasticity, multicollinearity
3. **Model Development**: Linear and Quadratic Discriminant Analysis
4. **Variable Selection**: Stepwise methods for optimal predictor set
5. **Model Validation**: Cross-validation, ROC analysis, confusion matrices

Business Context Variables

- **Industry**: Fintech lending platform (peer-to-peer and direct lending)
- **Loan Types**: Personal loans, small business loans, debt consolidation
- **Risk Framework**: Traditional credit scoring with machine learning enhancement
- **Regulatory Environment**: Consumer lending regulations, fair lending requirements

File Format

- **Encoding**: UTF-8
- **Delimiter**: Comma (,)
- **Missing Values**: Represented as empty cells (pandas default)
- **Header**: First row contains variable names
- **Index**: No row index included (use application_id as primary key)

Note: This dataset is synthetically generated for educational purposes. All applicant identifiers and financial data are fictional but based on realistic lending industry patterns and statistical relationships commonly found in credit risk modeling.