Credit Card Fraud Detection Dataset Overview

Introduction

Credit card fraud detection is a critical problem in the finance industry. Fraudulent transactions can cause significant financial losses and undermine customer trust. In this project, we develop a machine learning model to detect fraudulent credit card transactions using a dataset of anonymised transactions.

Dataset Summary

- **Total Rows:** 100,000
- **Total Features:** 21 (including the target variable)
- **Purpose:** The dataset is designed for **credit risk assessment**, specifically to predict whether a borrower will **default** on a loan based on their financial and personal information.

1. Dataset Composition

The dataset includes **20 independent variables (features)** and **1 target variable** (default). These features can be categorized into three main groups:

1. **Personal & Demographic Information** – Age, employment status, education level, home ownership, marital status, number of dependents.
2. **Financial Features** – Income, loan amount, loan term, interest rate, debt-to-income ratio, monthly expenses, annual savings, retirement savings.
3. **Credit History Features** – Credit score, credit history length, number of credit lines, late payments, bankruptcies.

Target Variable (default**)**

- **Binary Variable:**
  - 0 = Loan was repaid successfully (No Default).
  - 1 = Borrower failed to repay the loan (Default).
- The dataset is **balanced**, with approximately **50% defaults and 50% non-defaults**.

2. Features (Independent Variables)

Demographic Features

1. **age** (integer, 21–65 years)- Age of the borrower. Older individuals may have a longer credit history, impacting risk assessment.
2. **num_of_dependents** (integer, 0–4)- Number of dependents (e.g., children, spouse, elderly parents). More dependents can impact financial stability.
3. **education_level** (categorical: "high_school", "bachelor", "master", "phd") -Highest level of education attained. Higher education may be correlated with higher income and financial stability.
4. **marital_status** (categorical: "single", "married", "divorced") - Marital status of the borrower. Married individuals may have joint income but also higher expenses.

Financial Features

5. **income** (integer, $20,000–$100,000)- Annual income of the borrower. Higher income generally reduces the risk of default.
6. **loan_amount** (integer, $1,000–$50,000)- Amount of money borrowed. Larger loans may carry a higher risk of default.

7. **loan_term** (categorical: 12, 24, 36, 48, 60 months)- Duration of the loan in months. Longer terms may increase risk due to financial uncertainty.
8. **interest_rate** (float, 3.5%–15.0%)- Annual interest rate on the loan. Higher rates increase repayment burden.
9. **debt_to_income_ratio** (float, 10%–50%)- Ratio of total debt payments to income. Higher values indicate financial stress.
10. **home_ownership** (categorical: "own", "rent", "mortgage")- Indicates whether the borrower owns a home, rents, or has a mortgage. Homeowners may have more financial stability.

Credit History Features

11. **credit_score** (integer, 300–850)- Numerical representation of the individual's creditworthiness. Higher scores indicate lower risk.
12. **credit_history_length** (integer, 1–30 years)- Number of years the borrower has had a credit history. Longer history generally means better creditworthiness.
13. **num_credit_lines** (integer, 1–20)- Number of active credit lines (e.g., credit cards, loans). Too many or too few can be risky.
14. **late_payments** (integer, 0–9)- Number of late payments in the borrower's credit history. More late payments increase default risk.
15. **bankruptcies** (integer, 0–2)- Number of past bankruptcies. Even a single bankruptcy significantly increases the risk of default.

Savings & Expense Features

16. **annual_savings** (integer, $500–$50,000)- Amount saved annually. Higher savings indicate financial stability.
17. **retirement_savings** (integer, $1,000–$200,000)- Money saved for retirement. Indicates long-term financial planning.
18. **monthly_expenses** (integer, $500–$10,000)- Total monthly expenses. Higher expenses relative to income may increase risk.

Employment Features

19. **employment_status** (categorical: "employed", "unemployed", "self-employed")

- Employment status of the borrower.
- **Employed:** Steady income source, lower risk.
- **Unemployed:** No stable income, high risk.
- **Self-employed:** Variable income, moderate risk.

---

3. Dataset Challenges & Opportunities

- **Predictive Modeling**: Can be used to train machine learning models to classify borrowers as high or low risk.
- **Feature Engineering**: Some categorical variables (like employment status and education level) need to be encoded for machine learning models.
- **Handling Class Imbalance**: The dataset is already balanced, which simplifies model training.
- **Financial Decision-Making**: Useful for lenders to assess borrower risk and make informed lending decisions.

---

4. Use Cases

- **Loan Approval Systems:** Automate decision-making based on borrower risk.
- **Credit Score Modeling:** Identify key factors that impact loan defaults.
- **Customer Segmentation:** Classify borrowers into risk groups for personalized financial products.