Loan Defaulter Dataset

Project title: Loan Default Prediction
Data source: Kaggle — Loan Defaulter dataset (uploaded by **GAURAV DUTTA**).
Prepared by: Mohamed Sheta / Ahmed Ashraf / Tasneem Hussein / Ossama Ayman / Malak Ahmed

---

1. Introduction

This document outlines the dataset for the Loan Default Prediction project, its source, the team involved, and a detailed explanation of each file. The project aims to build predictive models estimating the probability of loan default (binary target). The dataset includes three files: **application_data.csv**, **previous_application.csv**, and **columns_description.csv**. It first provides an overall overview, then explains each file individually with suggested preprocessing, feature engineering, and sample code for merging historical data.

---

2. Objective

The objective of this case study is to apply Exploratory Data Analysis (EDA) techniques in a real-world business scenario within the banking and financial services domain. The focus is to develop a basic understanding of risk analytics by analyzing customer data and their previous loan applications. By combining demographic, financial, and historical loan information, the study aims to identify key factors that influence the risk of default. This will help in minimizing financial losses for the bank by improving lending decisions and building a stronger risk assessment framework.

---

3. Dataset overview

The dataset comprises information about loan applicants and their credit histories. Use the files together to perform thorough EDA (exploratory data analysis), create aggregated historical features, and build robust predictive models.

Primary responsibilities for each file

- `application_data.csv` — main data: client-level features and the **TARGET** label (0 = non-default, 1 = default). This is the file you will use for preprocessing, training and evaluating models.
- `previous_application.csv` — historical loan applications from the same clients. Use it to create additional features (counts, averages, ratios) summarizing past behaviour for each `SK_ID_CURR`.
- `columns_description.csv` — column dictionary: definitions and explanations of columns across the dataset. Use this as a reference throughout analysis.

Main workflow

1. Start with application_data.csv for cleaning, encoding, and baseline modeling.
2. Aggregate previous_application.csv by SK_ID_CURR to create historical features, then merge them into the main table for improved predictions.
3. Use columns_description.csv as the authoritative reference for column meanings and units.

---

4. File-by-File Breakdown

4.1 application_data.csv — Main Application Data

Purpose: Core dataset of current loan applications with client attributes and default label.
Target column: TARGET — 1 = default, 0 = repaid.

Top 10 important columns:

Column	Description
SK_ID_CURR	Unique client identifier.
TARGET	Default indicator (1 = default, 0 = repaid).
AMT_INCOME_TOTAL	Total annual income of the client.
AMT_CREDIT	Credit amount of the loan.
AMT_ANNUITY	Loan annuity (installment) amount.
NAME_CONTRACT_TYPE	Type of contract (Cash loans, Revolving loans).
CODE_GENDER	Client gender.
NAME_EDUCATION_TYPE	Education level of the client.
NAME_FAMILY_STATUS	Marital status of the client.
EXT_SOURCE_3	External risk score (very predictive feature).

4.2 `previous_application.csv` — Historical Applications

Purpose: Contains all previous loan applications of the same clients. Use to create historical features per `SK_ID_CURR`.

Top 10 important columns:

Column	Description
SK_ID_CURR	Client identifier (link to application data).
SK_ID_PREV	Previous application identifier.
AMT_APPLICATION	Amount of credit applied for.
AMT_CREDIT	Amount of credit approved.
NAME_CONTRACT_TYPE	Type of previous loan.
NAME_CONTRACT_STATUS	Status of the application (Approved, Refused…).
PRODUCT_COMBINATION	Product combination offered.
DAYS_DECISION	Days relative to current application when the decision was made.
RATE_DOWN_PAYMENT	Rate of down payment for the loan.
CHANNEL_TYPE	Channel through which the loan was applied (e.g., branch, internet).

4.3 columns_description.csv — Column Dictionary

Purpose: This file provides a human-readable description of each column across all files. It is not used for modeling directly but is essential for understanding the meaning and units of the variables.

Top 10 example columns from the dictionary:

Column name	Meaning
SK_ID_CURR	Client identifier used across files.
TARGET	Binary indicator of default.
AMT_CREDIT	Amount of credit requested/approved.
AMT_INCOME_TOTAL	Client’s total annual income.
NAME_CONTRACT_STATUS	Application status in previous_application.csv.
EXT_SOURCE_3	External risk score 3.
CODE_GENDER	Gender of the client.
NAME_FAMILY_STATUS	Marital status.
NAME_EDUCATION_TYPE	Education level.
CHANNEL_TYPE	Channel type of previous applications.