Loan Defaulter Risk Analysis

Exploratory Data Analysis (EDA) Summary Report (GitHub)

Project Goal

- Apply EDA techniques in a real-world business scenario.
- Understand risk analytics in the banking and financial sector.
- Analyze how data can minimize lending risks.
- Identify patterns and indicators associated with loan defaults.

Dataset Overview

Dataset Name	Records	Fields	Categorical	Numeric	Description
application_data.csv	307,511	122	16	106	Customer-level info at loan application (demographics, income, credit, employment).
previous_application.csv	1,670,214	37	16	21	Historical loan application data (approved, refused, ongoing).

Source:
Kaggle Loan Defaulter Dataset

- application_data.csv
- previous_application.csv

Missing Data Handling & Feature Engineering

Objective: Improve data quality and model readiness by addressing missing values.

- **Feature selection:**
  - Dropped columns having **>40% null values** and **low correlation** with the TARGET variable.
- **Categorical features:**
  - If **missing >25%**, filled proportionately using the **most frequent category**.
  - If **missing <25%**, filled using **mode**.
- **Numerical features:**
  - Filled missing values with **mean**, **median**, or **mode** as appropriate.
- **Continuous variables:**
  - Applied **binning** based on **quantile analysis** to categorize continuous features.

❖ Key EDA Insights

A. Key Findings (Categorical Variables)

- **Loan Type:** The Majority of customers have taken **cash loans**, which are comparatively **more risky** compared to **revolving loans** in terms of default.
- **Gender:**
  - Most loans were taken by **female customers**.
  - Female default rate ≈ **7%**, lower than males — indicating a **safer segment**.
- **Type of Suite:**
  - Majority of applicants are **unaccompanied**, with a default rate of **~8.2%**, which remains acceptable.
- **Income Type:**
  - **Working professionals**, **commercial associates**, and **pensioners** are the **safest income groups**.
- **Education:**
  - **Higher education** group has the **lowest default rate (~5.4%)**.
- **Family Status:**
  - **Married customers** show better repayment behavior (default rate ≈ **7.6%**).
- **Housing Type:**
  - Customers owning a **house/apartment** are **more reliable** (default rate ≈ **8%**).
- **Occupation:**
  - **Low-skill laborers** and **drivers** are **highest defaulters**.
  - **Accountants, core staff, and managers** are **safer segments** (default ≤ 7.5–10%).
- **Organization Type:**
  - **Transport Type 3** organizations show **high default risk**.
  - **Others**, **Business Entity Type 3**, and **Self-employed** are **relatively safe** (default ≈ 10%).

B. Univariate Analysis (Numerical Variables)

- Majority of **loan amounts**, **credit amounts**, and **goods prices** fall between **0–1 million**.
- Most customers have **income ≤ 1 million**.
- **Annuity payments** typically range between **0–50K**.

C. Bivariate Analysis

- **AMT_CREDIT ↔ AMT_GOODS_PRICE** are **positively correlated**; higher credit amounts correspond with **lower default rates**.
- Customers with **income ≤ 1M** and **loan < 1.5M** show **higher default probability**.
- Applicants with **1–4 children** are **safer** borrowers.
- Customers who can **pay annuities around 100K** and **loan amount < 2M** are **low risk**.

D. Analysis on Merged Data

- **Repair purpose** loans were the most common in previous applications — also with **highest cancellations**.
- **80–90%** of previously **refused/canceled** applicants are **repayers** in the current dataset — a potential **re-engagement opportunity**.
- **Unused previous offers** show **lowest default risk (~8.3%)** containing customers with **high-income**

❖ Business Recommendations

A. Target Customer Profile (Low Risk)

- **Income:** Below **1 million**
- **Organization Type:** *Others, Business Entity Type 3, Self-employed*
- **Occupation:** *Accountants, Core staff, Managers, Laborers*
- **Demographics:** *Married*, *female*, *highly educated*, with *≤5 children*
- **Housing:** Owns *house/apartment*
- **Lifestyle:** *Unaccompanied* individuals are acceptable (~8.2% default)

B. Recommended Loan Segments

- **Credit Amount:** ≤ 1 million
- **Annuity:** Around 50K (based on eligibility)
- **Income Bracket:** Below 1 million
- **Previous Applicants:** Reconsider those *previously unused offers*, as **above 90% are now repayers**

C. Precautions / High-Risk Segments

- Avoid applicants from **Transport Type 3** organizations.
- Avoid **low-skill laborers** and **drivers**.
- Avoid **previously refused offers** (unexpectedly high default rates, **~12%**).

❖ Tools & Libraries Used

• Python • Pandas • NumPy • Matplotlib • Seaborn

❖ Conclusion

This EDA reveals actionable insights for minimizing loan default risks. By targeting specific customer profiles and refining loan strategies, financial institutions can improve repayment rates and reduce exposure to high-risk segments.