EDA Credit Analysis

Made by: Navneet Kumar

Introduction

- When evaluating a loan application, financial institutions must decide whether to approve or reject the request.

- Two key risks in this process:
  - Denying credit to a reliable applicant results in lost business opportunities.
  - Granting credit to a high-risk applicant may lead to financial losses due to non-repayment.

Goal & Data Overview

- **Objective:** Examine loan application data to determine the main factors influencing credit approval and default likelihood.

- Dataset Utilized:
  - Active Applications: application_data
  - Historical Applications: previous_application

Analytical Approach

1. Importing and examining data
2. Identifying missing values and inconsistencies
3. Performing exploratory data analysis (EDA)
4. Checking for class imbalance
5. Analyzing categorical variables (Univariate)
6. Analyzing numerical variables (Univariate)
7. Comparing numerical & categorical data (Bivariate)
8. Evaluating relationships between numerical variables
9. Assessing previous loan applications
10. Drawing final conclusions and insights

Exploratory Data Analysis (EDA)

- Dataset Examination:
  - Files analyzed: application_data and previous_application
  - Dataset characteristics reviewed (.shape, .info(), .dscribe())

Data Preprocessing:

- Handling missing values:
  - Columns with excessive missing values (>35%) were removed.
  - Moderate missing values (≤19%) were filled (categorical: mode, numerical: median).
- Rectifying incorrect data types.
- Addressing inconsistencies in date-related columns by converting negative values.
- Ensuring categorical variables have standardized labels.

Data Preparation & Adjustments

- Transformed categorical attributes into numerical formats for better analysis.
- Standardized binary variables (e.g., Converted ‘Y/N’ values into 1/0).
- Imputed missing data based on logical assumptions:
  - Reclassified unknown organization types based on applicant income.
  - Created meaningful groupings for income, loan amounts, and age brackets.

Checking Data Imbalance for Target Variable

- Since there is a huge imbalance between the TARGET variables 0 and 1, it makes more sense to divide data frame into two sub datasets then continue our analysis.

- I have splits data frame as follows:
  - Target0 : (Non-Defaulted Population) Clients without Payment Difficulties.
  - Target1 : (Defaulted Population) Clients with Payment Difficulties.

Categorical Variable Analysis – Univariate

- **Demographics & Loan Applicants:**
  - More female applicants than male.
  - Middle-aged individuals (35–60) have the highest default tendencies.

- **Loan Purpose & Employment Type:**
  - Majority of applicants seek cash loans.
  - Most applications come from salaried professionals, retirees, and business associates.
  - Lower participation from students, unemployed, and entrepreneurs.
  - Working professionals exhibit the highest risk of default.

Additional Categorical Insights

- **Education & Marital Status:**
  - Secondary education holders form the largest applicant base and have the highest default risk.
  - Married individuals apply the most but struggle with repayments more often than other groups.
  - Widowed applicants have the lowest representation.

- **Employment & Income Groups:**
  - Pensioners and manual laborers are frequent loan seekers and have higher default rates.
  - Middle-income earners are the largest borrower group and show noticeable default patterns.

Numerical Variable Analysis – Univariate

- **Loan Amount & Payment Trends:**
    - Loan repayment amounts show significant variation among defaulters.
    - The loan amount distribution remains largely similar for both defaulters and non-defaulters.

- **Income & Product Pricing Insights:**
    - The income spread among defaulters is more scattered than non-defaulters.
    - The pricing of goods purchased using credit aligns similarly for both groups.

Cross-Analysis of Numerical & Categorical Data

- **Impact of Income, Education, and Family Background:**
  - Widowed applicants with higher education levels show lower default rates.
  - Married individuals with advanced education borrow less and repay reliably.

- **Credit Limits & Demographics:**
  - A majority of loan applicants receive relatively smaller credit amounts.
  - Borrowers with stable family structures and higher education tend to secure larger loan approvals.

Cross-Analysis of Two Categorical Variables

Loan default risk is influenced by employment type and income:

- Salaried professionals have high application rates but lower default probability.
- Pensioners and small business owners have higher default risks.
- Unskilled workers face the highest likelihood of default.
- Individuals with higher education generally demonstrate better repayment behavior.

Correlation Findings

Key Relationships Identified:

- Credit amount strongly correlates with goods price.
- Higher income is associated with fewer dependents.
- Loan amounts tend to be higher for individuals with valuable assets.
- Applicants from densely populated areas generally have higher credit approvals.

Loan Types & Approval Trends

- **Customer Segments & Loan Outcomes:**
  - 80.7% of applicants have taken loans before.
  - 14.5% are first-time borrowers.
  - Approval rate: ~38.8% | Rejection rate: ~58.5%.

- **Loan Purposes & Success Rates:**
  - Applications for home repairs have the highest rejection rates.
  - Loans for education and medical expenses have balanced approval-rejection trends.
  - Car and debt consolidation loans experience higher rejection rates.

Property Type & Loan Defaults

- Applicants residing in office apartments receive higher credit limits with lower default rates.
- Those living in cooperative housing exhibit higher instances of loan defaults.
- Banks should be cautious when approving large loans for co-op apartment residents and instead prioritize stable housing applicants.

Key Insights

•Default rates among pensioners are declining, whereas working professionals show increasing risk.
•Married and widowed individuals experience fewer repayment difficulties compared to unmarried or civil-married applicants.
•Secondary education holders default more frequently than those with higher education.
•Unskilled laborers and lower-secondary education applicants are at greater risk of non-repayment.
•The count of ‘Low skilled Laborers’ in ‘OCCUPATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 17%. Hence, client with occupation type as ‘Low skilled Laborers’ are the driving factors for Loan Defaulters.
•The count of ‘Lower Secondary’ in ‘NAME_EDUCATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 11%. Hence, client with education type as ‘Lower Secondary’ are the driving factors for Loan Defaulters.
•Banks should focus more on contract type Student ,pensioner and Businessman with housing type other than Co-op apartment, Office apartment for successful payments.
•Banks should focus less on income type Working as they are having the greatest number of unsuccessful payments.

Recommendations for Lenders:

- Prioritize applicants with stable employment and housing conditions.
- Carefully assess working professionals with high rejection rates before approval.
- Favor individuals living with family or in long-term residential arrangements.

THANK YOU