Bank Loan Default risk analysis

By Faizan Quazi

Introduction:

The loan providing companies find it hard to give loans to the people due to their insufficient or nonexistent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample
- All other cases: All other cases when the payment is paid on time

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- **Approved**: The Company has approved loan Application
- **Canceled**: The client canceled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- **Refused**: The company had rejected the loan (because the client does not meet their requirements etc.)
- **Unused** offer: Loan has been canceled by the client but on different stages of the process

Business Objective:

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Approach:

I. Getting Jupyter Ready

Objective:

- Set up the analysis environment by importing essential libraries.

Steps:

- Import NumPy, Pandas, Matplotlib, and Seaborn.
- Reading & Understanding the Data

Objective:

- Establish a foundation for analysis by importing and exploring the datasets.

Steps:

- Import input files.
- Thoroughly inspect DataFrames to gain an initial understanding of the data structure.

III. Data Cleaning & Manipulation

Objective:

- Ensure data quality and prepare it for analysis.

Steps:

- Calculate null values.
- Streamline the dataset by removing unnecessary columns.
- Standardize values by converting negative days to positive.
- Transform days into age and create categorical age bins.
- Categorize amount variables into bins.
- Convert selected columns to categorical types.

IV. Data Type Conversion

Objective:

- Harmonize data types to align with analysis requirements.

Steps:

1. Perform necessary data type conversions.

V. Null Value Data Imputation

Objective:

- Enhance dataset completeness by addressing missing data.

Steps:

- Impute null values in categorical variables using the mode.
- Create a new category for categorical variables with higher null percentages.
- Impute null values in numerical variables using the median.

VI. Identifying and Handling Outliers

Objective:

- Identify and mitigate outliers to improve analysis accuracy.

Steps:

- Identify outliers in specific columns.
- Implement outlier handling, particularly in AMT_INCOME_TOTAL and DAYS_EMPLOYED.

VII. Data Analysis Strategy

Objective:

- Establish a robust foundation for data exploration and insights generation.

Steps:

- Address data imbalance.
- Conduct categorical data analysis through segmentation and bivariate analysis.
- Analyze numeric data through segmentation and bivariate analysis.
- Perform a bifurcation of the database based on the TARGET variable.
- Explore correlation matrices.

VIII. Conclusions

Key Insights:

- Certain client attributes strongly correlate with loan repayment or default.

Decisive Factors for Repayers:

- Education type, income source, client rating, organization type, age, employment duration, income level, loan purpose, and number of children.

Decisive Factors for Defaulters:

- Gender, family status, education type, income source, client rating, occupation, organization type, age, employment duration, number of children, number of family members, and goods price.

Risk Mitigation Strategies:

- Implement higher interest rates for specific categories like housing type, credit amount, income, children count, and loan purpose.

High Default Risk Categories:

- Purpose of loan for repairs has the highest default rate.
- Consider historical rejection data for loans with the purpose of repair for future risk assessment.

Tech-Stack Used:

1. **Programming Language:**
   - Python 3.8
2. **Data Analysis and Manipulation:**
   - **NumPy**: Used for numerical computations and array operations.
   - **Pandas:**

   Employed for data manipulation, cleaning, and analysis using DataFrames.
3. **Visualization:**
   - **Matplotlib:**

   Utilized for basic plotting and visualization.
   - **Seaborn:**

Implemented for enhanced data visualization and statistical graphics.

4. **Data Exploration and Analysis:**
   - **Jupyter Notebook:**

   Integrated development environment for interactive computing and documentation.

7. **Version Control:**
   - **Git**: Used for version control and collaboration.

Insights:

Green Lights for Reliable Borrowers:

- Education Matters: Higher academic degrees signal lower default risk.
- Stable Income is Key: Students and businessmen historically repay loans successfully.
- Reliable Reputation: Clients with a "RATING 1" credit rating are a safer bet.
- Experience Pays Off: Longer careers (40+ years) translate to fewer defaults.
- Financial Strength Matters: Higher income (above 700,000) indicates a lower chance of default.
- Family Factor: Individuals with 0-2 children tend to be dependable borrowers.
- Responsible Borrowing: Loans for hobbies or garage purchases show high repayment rates.

Red Flags for Potential Defaults:

- Gender Gap: Men exhibit a higher tendency to default compared to women.
- Marital Status Matters: Single individuals or those in civil marriages have higher default risks.
- Education Matters: Lower secondary and secondary education levels correlate with increased risk.
- Income Instability: Maternity leave or unemployment can lead to defaults.
- High-Risk Professions: Avoid certain occupations like low-skill laborers, drivers, and waitstaff due to higher default rates.
- Job Hopping: Individuals with less than 5 years of employment show increased default risk.
- Large Families: Clients with 9+ children have a 100% default rate, necessitating loan rejection.
- High Loan Amounts: Credit exceeding 3 million tends to attract more defaulters.

Smart Strategies for Responsible Lending:

- Targeted Interest Rates: Apply higher interest rates for specific categories with higher default risk (e.g., young borrowers, self-employed individuals, larger loan amounts).
- Conditional Approvals: Consider loan approvals for certain riskier categories (e.g., rented apartments, 300-600 k loans) with stricter terms and higher interest rates to minimize losses.
- Loan Purpose Scrutiny: Approach loan applications for repairs with caution, as they display high risk and frequent rejections/refusals.

.

Results:

1. Risk Identification: Pinpointed decisive factors for repayers and defaulters, enhancing risk assessment.

2. Decision Foundations: Established a basis for informed loan approvals.

3. Strategic Recommendations: Proposed tailored strategies, including higher interest rates for specific risk segments.

4. Loss Prevention: Contributed to preventing losses by addressing risky applicant categories.

5. Future Risk Guidance: Provided insights for future risk assessments, especially for high-default loan purposes.

6. Improved Understanding: Enhanced insight into loan dynamics, considering demographics and finances.

7. Data-Driven Framework: Established a data-driven decision framework for continuous loan approval improvement.