Bank Loan Case Study Report

Analyst – Jeenal Bolia

Excel sheet – application data, previous application data

Project Description

In this project, we seek to identify the optimal indicators of loan default risk via EDA on a loan application dataset. The dataset has client financial data, credit amounts, demographics and a target column for default status. The analysis will be directed towards exploring univariate, segmented, and bivariate relationships to find patterns and characteristics that identify likely clients from not reliable clients.

Approach

To accomplish this, I followed a systematic data analytics workflow:

- Cleared and prepared the data, dealing with outliers and nulls.
- Conducted univariate analysis to look at the columns independently, like income, credit amount, etc.
- Conducted segmented analysis, breaking the data into defaulters and non-defaulters to analyze which patterns were representative and related.
- Conducted bivariate analysis to analyze two variables together, and especially with the target (default).
- Highlighted the key correlations for each segments to illustrate the most important financial variables.

Tech Stack used

Microsoft Excel 2022 Used for data preprocessing, statistical analysis, creating PivotTables, applying conditional formatting, and charts (bar, pie, histogram, etc.) to visualize trends and distributions.

Insights

- The loan amount (AMT_CREDIT) had a strong association with the goods price (AMT_GOODS_PRICE) among all clients.
- Income and credit had a reasonable linkage among non-defaulters but no relationship among defaulters — suggesting no reasonable checks would be take place in these risky contexts.
- It would appear that by restricting approval guidelines in regards to income and credit association, the reasons for defaults could potentially be avoided.

- The data also showed class imbalances, in which non-defaulters were the majority within the dataset.

Results

Task A: Identifying and Handling Missing Data in Excel

In application data excel and previous application data excel –

Column type	Missing values replaced with
Category	Unknown or blank
Numeric	Mean or 0

Task B: Outliers Handling

To check outliers, formula –

- =QUARTILE.INC(RANGE, QTR)
- IQR = Q3 – Q1
- Lower Limit: Q1 - 1.5 * IQR
- Upper Limit: Q3 + 1.5 * IQR

Checked outliers in numerical columns, which can probably contribute to risk analysis such as –

- AMT_INCOME_TOTAL, AMT_CREDIT, DAYS_EMPLOYED, AMT_ANNUITY, etc.
- Done using conditional formatting
- Marked them red

Task C: Class Distribution in Loan Dataset

- Count of class 0 = 45,973 -> leading to 91.95%
- Count of class 1 = 4,026 -> leading to 8.05%

Class Imbalance Assessment:

The data demonstrates a significant imbalance between class distributions, with 91.95% of applications labeled Class 0 (Non-Default) and only 8.05% labeled Class 1 (Default). With a skewed class distribution of a 92:8 split, it is possible we may obtain biased predictions in our model since the overwhelming constructive class is favored. To validly evaluate model performance, we should consider the class imbalance in the modeling stage and/or training stage, with approaches like resampling or including advanced performance metrics (i.e., F1-Score, ROC-AUC) to ensure that everyone is treated fairly and accounted for with regard to the biased performance on Class 1.

Task D: Univariate, Bivariate & Segmented Analysis

1. Segmented Univariate Analysis

Objective: To determine the distribution of categorical variables compared to different segments of another variable (e.g., level of education versus gender).

Steps Taken:

- Created Pivot Tables with:
- NAME_EDUCATION_TYPE in Rows
- CODE_GENDER in Columns
- Count of TARGET as Value

Findings:

- From this analysis we see that many applicants had an educational level of "Secondary / Secondary Special".

- Female applicants were more prevalent than male applicants in the lowest educational levels.

- Very few applicants reported having "Academically Degreed" or "Incomplete Higher" educations, both male and female positions.

2. Segmented Bivariate Analysis

Objective: To determine how the relationship between a "explanatory" variable and the target variable (default or not), varies across segments of another explanatory variable.

Steps Taken:

- Used Pivot Tables with:
- NAME_EDUCATION_TYPE in Rows
- CODE_GENDER and TARGET in Columns
- Count of TARGET in Values

Key Findings:

- Default rates were only slightly higher in males at most levels of education.

- Default rates for applicants with Lower Secondary education was higher than all other levels committed, especially among males.

- Analyses showed that the default rates for those with Higher education, and academic degree were significantly lower and therefore better "credit risk".

Task E Summary: Identifying Top Correlations in Segmented Data

In order to comprehend the reasons for loan default better, the dataset segmented based on TARGET variable into:

- Segment 0 → TARGET = 0
- Segment 1 → TARGET = 1

Key Findings of Analysis Correlation

Something That Correlated Strongly in Both Segments:

- AMT_GOODS_PRICE vs AMT_CREDIT
- With Correlation of ≈ 0.98 across both segments

Therefore, there is a strong linear relationship between how much credit was issued and the value of goods purchased, regardless of ability to repay.

Segment 0 (TARGET = 0):

- Moderate Correlation:
- AMT_INCOME_TOTAL vs AMT_CREDIT → 0.377
- Indicates that amount of income may have some influence on credit amount in reliable clients, but others factors may also to be assessed.

Segment 1 (TARGET = 1):

- Very Low Correlation:
- AMT_INCOME_TOTAL vs AMT_GOODS_PRICE → 0.0133
- Shows that there is near 0 relationship between income and value of purchases for defaulters — indicating that credit assessments made in these cases have no regard for ability to repay.

These trends indicate that:

Credit assessment is typically made in relation to product value, but not always full consideration for total ‘worth’ of customers, particularly in high-risk credits.

Defaulting clients appear to misjudge affordability, where income does not possess significant power over loan amount / purchase amount.