Bank Loan Case Study – Final Project Report

1. Project Description

This project involves analyzing bank loan data to identify patterns that predict loan defaults. As a data analyst at a financial services company, your goal is to minimize the risks of approving loans for customers who are likely to default, and maximize approvals for reliable customers, especially those without rich credit histories. You've performed a complete Exploratory Data Analysis (EDA) using Excel.

2. Approach

The project was divided into five core tasks:
- Task A: Missing Value Analysis
- Task B: Outlier Detection
- Task C: Data Imbalance Assessment
- Task D: Univariate, Segmented Univariate, and Bivariate Analysis
- Task E: Correlation Analysis by Scenario

Each task was conducted using Excel functions, pivot tables, statistical formulas, and visualizations.

3. Tech Stack Used

- GOOGLE SHEETS
- Functions: COUNTBLANK, IF, CORREL, QUARTILE, COUNTIF, etc.
- Tools: Pivot Tables, Charts, Conditional Formatting, Filters

4. Key Insights

- Missing values were mostly found in columns like OCCUPATION_TYPE and EXT_SOURCE_1, which were imputed or flagged.
- Outliers were identified in income, annuity, and credit columns using the IQR method.
- TARGET was highly imbalanced (~92% non-defaulters), requiring careful treatment in predictive modeling.
- Applicants with low external scores and short employment history showed higher default risk.
- Approved vs refused applications in past loans correlated with certain loan purposes and channels.

Task A: Missing Value Analysis

In Task A, I analyzed missing values in both application and previous application datasets. We used COUNTBLANK and COUNTIF in Excel to calculate missing values and percentages. Columns with over 40% missing were flagged for dropping, while others were marked for imputation or retention. Charts were created to show the missing percentage per column.

Task B: Outlier Detection

Outliers were detected using the Interquartile Range (IQR) method. I focused on numeric variables like AMT_INCOME_TOTAL, AMT_CREDIT, and DAYS_EMPLOYED in the application data, and AMT_CREDIT, RATE_INTEREST_PRIMARY, and CNT_PAYMENT in previous applications. Outlier thresholds were calculated, and records falling outside were flagged. Box plots and bar charts visualized these distributions.

Task C: Data Imbalance Analysis

We found that the TARGET variable in application data was highly imbalanced (~92% non-defaulters). This was visualized using bar and pie charts. In previous applications, NAME_CONTRACT_STATUS was used to examine imbalance, showing a majority of approved cases. This justified the need for balancing strategies in modeling.

Task D: EDA (Univariate, Segmented, Bivariate)

Univariate analysis was performed on income, credit amount, and days of birth using histograms and summary statistics. Segmented analysis compared income type and housing type across defaulters vs non-defaulters. In previous data, we segmented loan purposes by approval status. Bivariate analysis included scatter plots and correlation matrices for EXT_SOURCE scores, AMT_CREDIT vs AMT_ANNUITY, and more. Charts clearly showed risk patterns and feature relationships.

Task E: Correlation Analysis by Scenario

Correlation of numeric variables with TARGET was calculated for the entire dataset. Since TARGET is constant in filtered subsets, segment-wise correlation was not computed directly but top correlated features were highlighted. EXT_SOURCE_2, DAYS_BIRTH, and EXT_SOURCE_3 showed strongest relationships with default. A separate correlation analysis with STATUS_BINARY (Approved = 1, others = 0) was conducted on previous applications, highlighting RATE_INTEREST_PRIMARY and CNT_PAYMENT as significant indicators.

5. Final Result

Based on the detailed analysis performed using Excel across both application and previous loan datasets, the following actionable results were achieved:

1. Clear Risk Indicators Identified:
   - Low EXT_SOURCE scores strongly correlate with default (TARGET = 1)
   - Clients with shorter employment duration (DAYS_EMPLOYED) tend to default more
   - Self-employed and unemployed applicants had a noticeably higher default rate

2. Imbalanced Data Distribution:
   - Only ~8% of clients had payment difficulties, confirming the dataset is imbalanced
   - Approved loan status in historical data heavily dominated, highlighting risk of overfitting

3. Key Features for Predictive Modeling:
   - Variables like AMT_CREDIT, AMT_ANNUITY, and DAYS_BIRTH showed meaningful trends and outlier clusters
   - Past loan purposes (e.g., mobile phones, furniture) and sales channels impacted approval rates

4. Data Cleaning Framework Established:
   - Missing values handled using median or logical imputation
   - Outliers flagged using IQR for risk-based treatment or capping
   - Feature segmentation added clarity to creditworthiness insights

These insights will enable the business to:
- Approve reliable applicants faster
- Reduce losses from high-risk loans
- Train models with cleaner, targeted features for scoring applicants

6. Drive Link to Files

XLSX: Bank Loan Case Study (1).xlsx

PDF: Bank Loan Case Study – Final Project Report.pdf

VIDEO: video4184958008.mp4