Cairo University
Faculty of Engineering
Computer Engineering Department

Bank Loan Default Risk Analysis

Project Proposal

Idea

The project aims to analyze risk factors in loan applications to help banks minimize losses that happen due to defaults. As there are some applications with no credit history or missing guarantees this leads to high risks to banks and lending companies. This results in increased default rates.

By analysing previous and current loan applications, this project will aim to provide insights to how banks and lending companies can assess risk and improve lending strategies.

Dataset

The dataset contains information about loan applications, including details about clients (gender, age, income, own car, property or not, number of children, ..), clients' financial history (income, source of income, ..), and loan status. Also includes features like whether the client had payment difficulties before or not, loan amount, interest rate, repayment behavior, etc..

It also covers property details, social circle influence, and loan application metadata like which document did he deliver and which not.

https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter/data

The dataset consists of two parts:

- **Current Applications:** +300K records with 122 attributes.
- **Previous Applications:** +1.6M records with 37 attributes.

Due to the **large size of the dataset**, we will use **PySpark** to efficiently handle distributed data processing and analysis.

Planned Approach

Language and Tools:

- Python
- PySpark (for scalable processing)

1. Data Preprocessing

- Handle missing values appropriately.

- Encode categorical variables.
- Merge and align both current and previous application datasets when needed.

2. Exploratory Data Analysis (EDA)

- Use statistical summaries (mean, median, min, max, percentages, etc.)
- Visualizations like histograms, bar charts, box plots, and heatmaps (for correlation).
- Identify trends and key risk indicators.

3. Feature Selection

- Identify the most relevant features using techniques like correlation analysis, feature importance, and mutual information.
- Apply **association rule mining** to discover hidden patterns and relationships in applicant behavior and repayment likelihood.

4. Model Selection and Training

- Test a variety of classification models (e.g., Logistic Regression, Random Forest, XGBoost, etc.)
- Evaluate using metrics such as **accuracy, precision, recall, F1-score, and ROC-AUC**.

5. Prediction

- Predict the **probability of loan default** for new applicants.
- Rank applicants based on risk and provide actionable insights for decision-making.