Predicting Loan Default Risk with Machine Learning

Overview

This project focuses on building a machine learning model to predict the probability of loan default using historical data from loan applicants. The model will be trained using attributes of applicants and their payment behavior to classify applicants as 'good' or 'bad' clients.

Data

The dataset consists of two tables that can be merged by ID:

Attributes Table:

- Client ID
- Gender
- Car ownership
- Property ownership
- Number of children
- Annual income
- Income type
- Education level
- Marital status
- Housing type
- Age (derived from 'DAYS_BIRTH')
- Employment duration (derived from 'DAYS_EMPLOYED')
- Mobile phone, work phone, phone, email availability
- Occupation
- Family size

Payment Behavior Table:

- Client ID
- Months since the last installment
- Loan status (paid or in arrears)

Project Steps

Label Creation:

- Determine 'good', 'bad', and potentially 'arrear' labels based on payment behavior data.

Data Preprocessing:

- Clean, manipulate, and aggregate the data.
- Handle missing values and outliers.

Sampling:

- Use SMOTE to handle class imbalance, especially for 'bad' labels.

Modeling:

- Train and evaluate classification models:
    - Logistic Regression
    - Random Forest
    - Gradient Boosting
    - Support Vector Machines

Cross-Validation:

- Perform cross-validation to assess model performance and generalization ability.

Model Deployment:

- Build a basic interface to allow users to input applicant data and obtain predictions.