Predicting Loan Payback Using Python

Assignment Overview

This project focuses on developing a production-ready machine learning system that predicts the probability of a loan repayment outcome. The system combines data science techniques with practical software engineering to create a tool that financial institutions can use for credit risk assessment. The project demonstrates the use of Python for handling datasets, preparing data for analysis, and building a predictive model in a structured and reproducible manner, as well as building a custom interactive app for easy usage.

The deliverable is a complete Python application that:

- Trains and evaluates machine learning models on historical loan data
- Provides an interactive interface for real-time loan approval decisions

Assignment Specifications

Loan default prediction is critical for financial institutions to assess credit risk and make informed lending decisions. By analysing patterns in historical loan data, we can identify factors that distinguish successful borrowers from those likely to default. This project uses a dataset from Kaggle containing borrower information, loan characteristics, and repayment outcomes. The goal is to build a system that not only achieves high predictive accuracy but also provides actionable insights for loan officers depending on the probability of the loan payback.

Technical Requirements

The solution must be:

- Accurate: Achieve high accuracy
- User-friendly: Allow non-technical staff to make predictions easily
- Maintainable: Include clear documentation and modular code

Data Description

The dataset consists of two main files:

- Training data: Contains loan records along with the known loan repayment outcome.
- Test data: Contains loan records without the outcome variable.

Each row corresponds to one loan. Columns include borrower-related attributes (such as income and credit indicators) and loan-specific information (such as loan amount and interest rate).

Program Functionality

The Python program performs the following steps:

1. Data Loading

The program reads the provided CSV files and verifies that the data has been loaded correctly. If the data cannot be loaded, an appropriate error message is displayed, and the program terminates.

2. Data Exploration and Preprocessing

Summary statistics are generated to understand the structure of the data. Missing values and inconsistent entries are handled to prepare the dataset for modelling.

3. Feature Preparation

Relevant variables are selected and transformed where necessary to make them suitable for machine learning algorithms.

4. Model Training and Evaluation

A classification model is trained on the training dataset. Model performance is evaluated using appropriate metrics, such as accuracy and ROC-AUC.

5. Prediction Generation

The trained model is used to generate predictions for the test dataset.

6. Interactive Prediction Tool

A user-friendly interface allows user to input loan application details and receive instant approval or rejection recommendation with risk assessment and probability.

Data source: “Predicting Loan Default” Kaggle Competition, from https://www.kaggle.com/competitions/playground-series-s5e11