Project Proposal: Loan Approval Classification Analysis

Introduction

This project aims to predict loan approval status by analyzing key applicant and financial risk factors, using the dataset. This dataset offers a balance of categorical and continuous features, making it ideal for classification.

Research Questions

The project addresses these main questions:

1. What demographic and financial features most influence loan approval decisions?

2. How accurately can a classification model predict loan approval based on an applicant's financial profile?

Dataset and Variables

The dataset includes 45,000 records and 14 variables related to loan applicants and their financial profiles. Key variables are person_age (applicant's age), person_gender (applicant's gender), person_education (highest education level), person_income (annual income), person_emp_exp (years of employment experience), and person_home_ownership (home ownership status, e.g., rent, own). Loan-related details include loan_amnt (requested loan amount), loan_intent (purpose), loan_int_rate (interest rate), loan_percent_income (loan amount as a percentage of annual income), and cb_person_cred_hist_length (credit history length in years). Additional financial indicators are credit_score (applicant's credit score) and previous_loan_defaults_on_file (indicator of past loan defaults), with loan_status serving as the target variable indicating loan approval status (1 = approved, 0 = rejected).

The project will involve several key steps, starting with **Data Cleaning and Preprocessing** to address outliers, normalize numeric data, and encode categorical variables. Next, **Exploratory Data Analysis (EDA**) will be conducted to explore feature distributions and relationships, allowing us to identify significant risk factors for loan approval. In the **Modeling** phase, we will develop classification models like Logistic Regression, Decision Tree, and Random Forest which will predict loan approval status. Finally, the **Evaluation** step will assess model performance using accuracy, precision, recall, and F1-score for classification, and R-squared for regression, to ensure the models' effectiveness and reliability.

Sample Code to Read and Preview the Dataset

```
import pandas as pd

loan_df = pd.read_csv('loan_data.csv')

loan_df.head()

missing_values = loan_df.isnull().sum()

print("Missing Values:\n", missing_values)

print("Data Types:\n", loan_df.dtypes)

print("Data Summary:\n", loan_df.describe())
```