**Project: Automated Loan Approval System**

**Made By: Siddhartha Patra**

**1. Introduction**

This project aims to develop an automated loan approval system for XYZ Bank, one of Australia's largest banks. The current manual loan processing system has become increasingly challenging due to the subjective decision-making process and the growing number of loan applications. The objective is to build a machine learning solution that analyzes various applicant factors and predicts whether a loan should be granted or not.

**2. Problem Description**

The project focuses on a classification problem wherein a model needs to be developed to predict loan approval for applicants. By considering factors such as credit score, past history, and additional applicant information, the model will determine the loan granting status. The dataset will be explored, cleansed, and preprocessed to ensure optimal model performance. The output will include a probability score and a loan status label indicating approval or refusal.

**3. Loan Approval/Eligibility Problem**

To make informed decisions about loan approvals, statistical models are required to assess the likelihood of loan repayment. This problem involves implementing a model that predicts loan approval based on complex factors present in the provided dataset. The dataset used in this project is an anonymized synthetic dataset designed to resemble genuine loan data characteristics.

**4. Dataset Description**

The dataset consists of over 100,000 loan records and includes the following fields:

* Loan ID: Unique identifier for loan information.
* Customer ID: Unique identifier for each customer, who may have multiple loans.
* Loan Status: Categorical variable indicating if the loan was granted to the customer.
* Current Loan Amount: Loan amount that was either fully paid off or defaulted (pertaining to previous loans).
* Term: Categorical variable indicating if it is a short-term or long-term loan.
* Credit Score: Riskiness of the borrower's credit history (0 to 800).
* Years in Current Job: Categorical variable indicating the number of years the customer has been in their current job.
* Home Ownership: Categorical variable indicating home ownership status (Rent, Home Mortgage, Own).
* Annual Income: Customer's annual income.
* Purpose: Description of the purpose of the loan.
* Monthly Debt: Customer's monthly payment for existing loans.
* Years of Credit History: Years since the first entry in the customer's credit history.
* Months since Last Delinquent: Number of months since the last delinquent loan payment.
* Number of Open Accounts: Total number of open credit cards.
* Number of Credit Problems: Number of credit problems in the customer's records.
* Current Credit Balance: Customer's current total debt.
* Maximum Open Credit: Maximum credit limit across all credit sources.
* Bankruptcies: Number of bankruptcies.
* Tax Liens: Number of tax liens.

**5. Evaluation Criteria**

The model's accuracy must achieve a minimum threshold of 70% to meet the passing grade criteria.

**6. Approach**

The project was approached systematically with the following steps:

1. Understanding the problem statement and business objectives.
2. Familiarizing ourselves with relevant libraries and their functionalities.
3. Conducting in-depth exploratory data analysis (EDA) of each feature, identifying patterns, outliers, and relationships.
4. Performing feature engineering to create new meaningful features and enhance the predictive power of the model.
5. Data cleansing and preparation, including handling missing values, outliers, and data formatting.
6. Creating custom functions for machine learning models to streamline the modeling process.
7. Implementing appropriate data imputation techniques to fill missing values.
8. Defining an approach to solve the classification problem, considering various algorithms suitable for loan approval predictions.
9. Preparing the data for machine learning models by performing feature scaling and encoding categorical variables.
10. Training and testing multiple models using cross-validation techniques to assess their performance.
11. Building statistical models such as Gradient Boosting, XGBoost, etc., and fine-tuning hyperparameters for optimal performance.
12. Balancing the dataset using SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance issues.
13. Evaluating model performance using metrics such as accuracy, precision, recall, and F1-score.
14. Understanding additional evaluation metrics like ROC Curve and MCC (Matthews Correlation Coefficient).
15. Selecting the best-performing model based on various evaluation metrics.
16. Creating pickle files to save the trained model for future reusability and deployment.