# Bank Loan Classification - Brief Report

By Dipesh Mahato

## Contents

## Approach

In this task, I aimed to build a machine learning model that can accurately classify whether a personal loan was accepted or not based on the provided information. The dataset consists of 15 columns, including features such as age, gender, income, education level, and more, with the "Personal Loan" column as the target variable. The dataset was in xlsx format, I converted it to csv format for faster execution of the data.

My approach involved several key steps:

### Step 1: Data Pre-processing and Exploratory Data Analysis (EDA)

I started by loading the dataset and performing data pre-processing to handle any missing values. EDA was then conducted to gain insights into the data's distribution and correlations between variables. The categorical variables were encoded into numerical representations using the LabelEncoder from sklearn.

### Step 2: Feature Engineering

Feature engineering was performed to create new features that could potentially improve the model's performance. I performed feature scaling using MinMaxScaler and created interaction features like "Income_CCAvg" by multiplying "Income" and "CCAvg." Additionally, I used Polynomial Features and PCA for dimensionality reduction.

### Step 3: Handle Class Imbalance

To handle the class imbalance in the target variable, I checked the class distribution and applied Synthetic Minority Over-sampling Technique (SMOTE) to create synthetic samples for the minority class.

### Step 4: Model Building and Training

I built and trained multiple machine learning models, including Random Forest, Logistic Regression, and AdaBoost. I performed hyperparameter tuning using GridSearchCV with Stratified K-Folds cross-validation to select the best-performing models.

### Step 5: Model Evaluation

The final models were evaluated using accuracy, confusion matrices, and classification reports on the test set. The models' performance was compared to select the best deployment model.

### Step 6: Save the Model

The best model, as determined by its performance on the test set, was saved using the joblib library for future use.

## Key Findings and Insights

- The dataset contains missing values in columns – Gender, Income, Home Ownership, and Online. These missing values were handled using appropriate imputation techniques such as filling with mode for categorical variables and mean for numerical variables.
- Feature engineering, such as creating interaction features and handling outliers, improved the model's performance.
- Class imbalance in the target variable was addressed using SMOTE, leading to better predictions on the minority class.
- The Random Forest classifier achieved the highest accuracy and F1-score among the evaluated models, making it the selected model for deployment.

## Observations and Recommendations

- Class imbalance in the target variable "Personal Loan" was addressed effectively using SMOTE, leading to more balanced predictions.
- To further enhance model performance, more advanced techniques like feature selection or ensembling could be explored.
- Additional data may provide further insights and improve model generalization.
- Continuous monitoring and periodic updates of the model with new data are recommended to ensure ongoing accuracy.

Overall, the Random Forest model outperformed the other two models in terms of overall accuracy and balanced precision-recall scores for both classes. It demonstrated better performance in predicting loan acceptance and rejection.