

## 1 Problem Understanding

Probability of default measures the degree of likelihood that the borrower of a loan or debt will be unable to make the necessary scheduled repayments on the debt, thereby defaulting on the debt.

## 2 Data Understanding

The given dataset consists of user details and the classification goal is to predict whether the loan applicant will default (1/0) on a new debt (variable y).

## 3 Data Preparation

To handle the imbalance in data up-sample is performed using SMOTE algorithm (Synthetic Minority Oversampling Technique). SMOTE creates synthetic samples from the minor class (default) instead of creating copies. Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked, new observations.

## 4 Modeling

Model is implemented using Logistic Regression. Logistic Regression is used when the dependent variable(target) is categorical. In this case a Binary Logistic Regression is used i.e a categorical response with only two 2 possible outcomes. Example: default or Not