**Problem statement:**

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan. Upon receiving the loan application, the company need to validate the customer eligibility for loan. Company wants to automate the loan eligibility process (real time) based on customer details provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. The data (**loan\_prediction\_train.csv**) corresponds to a set of loan applications associated with individual customers. The data has 614 observations and 11 features.  Each observation is independent. The data has missing feature values. The analyst can implement a strategy to fill the gaps if desired. There may be outliers in the data. For each observation, it was recorded whether the loan was approved or not. If the loan was approved the loan status is denoted by ‘Y’, otherwise by ‘N’.

You are asked to predict the loan status for each observation in the test set(**loan\_prediction\_test.csv).**

**Data description:**

Given below is the description for each variable.

Variable ----- Description

Loan\_ID -->> Unique Loan ID

Gender -->> Male/ Female

Married -->> Applicant married (Y/N)

Dependents -->> Number of dependents

Education -->> Applicant Education (Graduate/ Under Graduate)

Self\_Employed -->> Self employed (Y/N)

ApplicantIncome -->> Applicant income

CoapplicantIncome -->> Coapplicant income

LoanAmount -->> Loan amount in thousands

Loan\_Amount\_Term -->> Term of loan in months

Credit\_History -->> credit history meets guidelines

Property\_Area -->> Urban/ Semi Urban/ Rural

Loan\_Status -->> Loan approved (Y/N)

**Suggested analysis guidelines:** This is a standard supervised classification task. In such problems, the labels are included in the training data and the goal is to train a model for learning to predict the labels from the features. The analyst is suggested to roughly follow the steps mentioned below while dealing with the problem:

* **Exploratory data analysis**- To understand the data. It may involve both univariate and bivariate analysis of data.
* **Missing value and outlier treatment**.
* **Model building**: classification algorithms such as logistic regression and other suitable algorithm
* **Evaluation metrics**: Comparative results from at least two algorithms
* **Feature Engineering** – to check if the model performs better with new features.