**Data Exploration Insights**

Predict ‘Bad\_label’ target feature with a model built on training set, then evaluating the model on the given test set.

This is a binary classification problem, with ‘Bad\_label’ having ‘0’ and ‘1’, denoting ‘accept’ and ‘reject’ loan application response respectively.

**Data Pre-Processing and Analysis**

Pre-processing was done remove duplicate samples, correct erroneous data, set features to the correct data type and remove features that contribute minimally to the predictive power of the model to be built.

The training set contains information for 23896 unique applications for loans and it is split into three subsets named ‘Data’, ‘Account’ and ‘Enquiry’.

**Data**

**23896** observations

Initially containing 87 features of current information for the loan applications, most importantly the response feature ‘bad\_label.’ It mostly consists of demographic information like gender, place of residence, occupation, email address, etc. Prominent features only include ‘cibil\_score’, marital\_status,’customer\_no’

## Account’

**186039** observations

The ‘Account’ subset consists of the most important data, mainly of the applicant’s financial background, mostly relating to loans and credit payment.

Features describing payment history seem to have the most significance. However, they are stored as a series of characters and further processing must be done to interpret it better

## Enquiry’

**404034** observations

The ‘Enquiry’ subset contains information regarding credit enquiries. Information pertaining to the type of loan enquiry would be useful.

# Feature engineering

Features relating to repayment history, duration of credit servicing, credit limit and credit enquiries are good metrics for determining loan risk.

# Model calculation

GINI was calculated as Gini = 2 \* Area Under ROC – 1

Max. GINI: 24.73.