**Loans default will cause huge loss for the banks, so they pay much attention on this issue and apply various methods to detect and predict default behaviours of their customers**

The data is given for 12 months and is meant to be predicted for the 13th month. It is a classification problem in which I had to predict if the loan is repaid in the 13th month with the respective given loan\_id.

First imported the sets of some basic and advanced python libraries and loaded the training data file. Then checked for the Shape, Correlation and Description of all the columns. For further analysis replaced the Categorical values like 'Source', 'Loan Purpose' and 'Financial Institution into integer with the help of labelling them to some certain int. values.

Variables like 'Loan\_id','Origination\_Date' and 'First\_Payment\_Date' can be dropped since they were not providing relevant information for the prediction.

'm13' column was biased with 99% major class and 1% minor class. To overcome this problem of Imbalanced data I used SMOTE to balance the data.

SMOTE does this by selecting similar records and altering that record in a column at a time by a random amount within the difference to the neighbouring records.

Used train-test-split to split data into Training and Testing and then applied Logistic Regression followed up by Confusion Matrix, F-Test on training data.

Then to predict ‘m13’ I had used PCA, an unsupervised statistical technique used to examine the inter-relations among a set of variables.

Then applied Decision Tree which gives the better results.