**Data Exploration**:

Exploratory data Analysis is done using pandas profiling and the nature of variables is checked as well as correlation of feature variables and a imbalance in dataset is noticed. With a huge imbalance of 99.45 : 0.55 in ratio.The dataset is highly imbalance

SMOTE is done on the dataset to balance it.

Feature Engineering:

Pandas profiling showed two Boolean variables insurance type and m13 as well as 20 categorical variables and 6 numeric variables.

It has rejected ‘number\_of\_borrowers’ due to high correlation with ‘coborrowers\_credit score’

.Instead of using m1 ,m2 ,…,m13 .I have used m1- m12,m1-m2,m2-m3,…..,m11-m12 i.e the difference in the features as I noticed that if there are more ‘0’ in one particular row for m1 ,m2…..m12 then the chance for m13 to be ‘0’ is more .

I have further used pearsonr correlation to find out corealtion between different features.I have found out that few other features in this which can be neglected .

I Converted Source ,Financial Instituition,loan purpose into categorical data.And then calculated the first payment time to get the response given by the customer.

**Model Building and Selection -**

* I have initially used random forest with Grid search CV using hyper parameter tuning
* Then tried lgbm which gave me a great private score but bad public score
* Then I tried Adaboost with smote then it gave me greatest private score but least public score .
* Finally I settled in XGB Boost by doing hypertuning of the parameters gamma qand alpha by fixing the learning rate and I got my highest public score and a good private score.

Please find the three documents

* 1. Final code file
  2. Code file containing cumulative of all approaches used
  3. Code file with highest private score.