MICRO CREDIT LOAN

Submitted by:

Ria Maitra

Abstract

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

This sector gives loans without collateral which increases the risk of bad debts. Bad debts is a cause of concern for MFI's as it negatively affects their portfolio. It also makes them vulnerable to credit risk and increases the cost of monitoring.

This paper examined the factors predicting microfinance credit default of a MFI. The dataset consists of 209593 microcredit beneficiaries along with details of their spending habits, main balance, last recharge etc. The dataset is analyzed properly both with univariate analysis and bivariate analysis. It is been observed that features, daily_decr30, daily_decr90, rental30, rental90, last_rech_amt_ma, cnt_ma_rech30, sumamnt_ma_rech30, medianamnt_ma_rech30, cnt_ma_rech90, sumamnt_ma_rech90, medianamnt_ma_rech90, cnt_loans30, amnt_loans30 and amnt_loans90 are highly correlated with target variable that is label (0-default,1-non-default). It is also been observed that certain independent variable are highly correlated with each other. We have used Logistic Regression, GaussianNB, Knn, Decision Tree and RandomForest models for prediction. RandomForest model has given the best performance with test-score 94%.

This model could be used by microfinance institutions to screen prospective loan applicants.

Introduction

Milton Friedman once said that “the poor stay poor, not because they are lazy, but because they have no access to capital”. Till date, a huge section of population remains outside the formal banking system. Microfinance institutions are the only ones equipped to reach the ‘unbankable’ or ‘unbanked’ masses.

Poverty is the main cause of concern in improving the economic states of developing countries. A microfinance institution is an organization that offers financial services to low income populations. Almost all give loans to their mebers and many offer insurance, deposit and other services.

A great scale of organization is regarded as microfinance institutions. They are those that offer credits and other financial services to the representatives of poor strata of populations.

Microfinance is increasingly being considered as one of the most effective tools of reducing poverty by enabling microcredit to the financial poor. Microfinance has a significant role in bridging the gap between the formal financial institutions and the rural poor. The MFI accesses financial resources from the banks and other mainstream financial institutions and provide financial and support services to the poor.

It enables people expand their present opportunities, the income accumulation of poor households has improved due to the presence of microfinance institutions that offer funds for their business.

It provides easy access to credit-Microfinance opportunities provide people credit when it is needed the most. Banks do not usually offer small loans to customers, MFI providing microloans bridge this gap.

It makes future investments possible, microfinance makes more money available to the poor sections of the economy. So, apart provide them with credit for constructing better houses improving their healthcare facilities, and exploring better business opportunities.

It serves the under- financed section of the society, majority of the microfinance loans provided by MFIs are offered to women. Unemployed people and those with disabilities are also beneficiaries of microfinance. These financing options help people take control of their lives through the betterment of their conditions.

Microfinance industry has to face quite a specific set of challenges, which cannot be addressed with solutions meant for commercial banks.

Around 60% of MFI agree that the geographic factors make it difficult to communicate with clients of far flung areas which create a problem in growth and expansion of the organization.

Over-indebtedness is major issue, also credit risk. As mostly customers are uneducated, high rate of interest also makes it difficult for customer to pay back eventually they become bad debt. Over-indebtedness makes the MFIs vulnerable to credit risk and increases the cost of monitoring that they have to incur to stay profitable in the long run.

Problem Statement:

Though Microfinance Institutions enables people expand their present opportunities by providing easy access to credit but still MFI faces the challenges of credit risk. As the rate of interest id high than any commercial bank, often people fail to pay their loan. Also, because of geographic factors makes it difficult to communicate with clients of far-flung areas which create a problem in growth and expansion of the organization.

Review of Literature:

In the literature, the terms microcredit and microfinance are often used interchangeably, but it is important to highlight the difference between them because both terms are often confused. Sinha (1998, p.2) states “microcredit refers to small loans, whereas microfinance is appropriate where NGOs and MFIs1 supplement the loans with other financial services (savings, insurance, etc)”. Therefore microcredit is a component of microfinance in that it involves providing credit to the poor, but microfinance also involves additional non-credit financial services such as savings, insurance, pensions and payment services (Okiocredit, 2005).

Microcredit and microfinance are relatively new terms in the field of development, first coming to prominence in the 1970s, according to Robinson (2001) and Otero (1999). Prior to then, from the 1950s through to the 1970s, the provision of financial services by donors or governments was mainly in the form of subsidised rural credit programmes. These often resulted in high loan defaults, high loses and an inability to reach poor rural households (Robinson, 2001).

Robinson states that the 1980s represented a turning point in the history of microfinance in that MFIs such as Grameen Bank and BRI2 began to show that they could provide small loans and savings services profitably on a large scale. They received no continuing subsidies, were commercially funded and fully sustainable, and could attain wide outreach to clients (Robinson, 2001). It was also at this time that the term “microcredit” came to prominence in development (MIX3 , 2005). The difference between microcredit and the subsidised rural credit programmes of the 1950s and 1960s was that microcredit insisted on repayment, on charging interest rates that covered the cost of credit delivery and by focusing on clients who were dependent on the informal sector for credit (ibid.). It was now clear for the first time that microcredit could provide large-scale outreach profitably.

The 1990s “saw accelerated growth in the number of microfinance institutions created and an increased emphasis on reaching scale” (Robinson, 2001, p.54). Dichter (1999, p.12) refers to the 1990s as “the microfinance decade”. Microfinance had now turned into an industry according to Robinson (2001). Along with the growth in microcredit institutions, attention changed from just the provision of credit to the poor (microcredit), to the provision of other financial services such as savings and pensions (microfinance) when it became clear that the poor had a demand for these other services (MIX, 2005).

The importance of microfinance in the field of development was reinforced with the launch of the Microcredit Summit in 1997. The Summit aims to reach 175 million of the world's poorest families, especially the women of those families, with credit for the self-employed and other financial and business services, by the end of 20154 (Microcredit Summit, 2005). More recently, the UN, as previously stated, declared 2005 as the International Year of Microcredit.

Objectives:

- To know the relationship between the independent variables with the dependent variable.
- To build a module which best classifies default and non-default customers.

Significance of the Study:

This study helps to understand the relationship between independent variables and dependent variable. It also helps to identify default (loan) customers which will be beneficial in further investment and improvement in selection of customers.

Scope of the Study:

Through this study it will be easy to identify which features plays important role in predicting default customers.

Methodology

This research has used quantitative analysis including:

- Univariate analysis
- Bivariate Analysis
- Descriptive Statistics

To study the relationship between independent variable and dependent variables

- Machine learning Algorithms: to predict default customers.

Research Aim:

The research aim is to build a model to predict default customers on the basis of independent variables.

Research Design:

The research has used the following steps:

- Exploratory Data Analysis:
- Pre-processing Pipeline
- Building Machine learning Models
- Conclusion

Data Collection:

Secondary source of data is used provided by a MFI client.

Format of data is csv file.

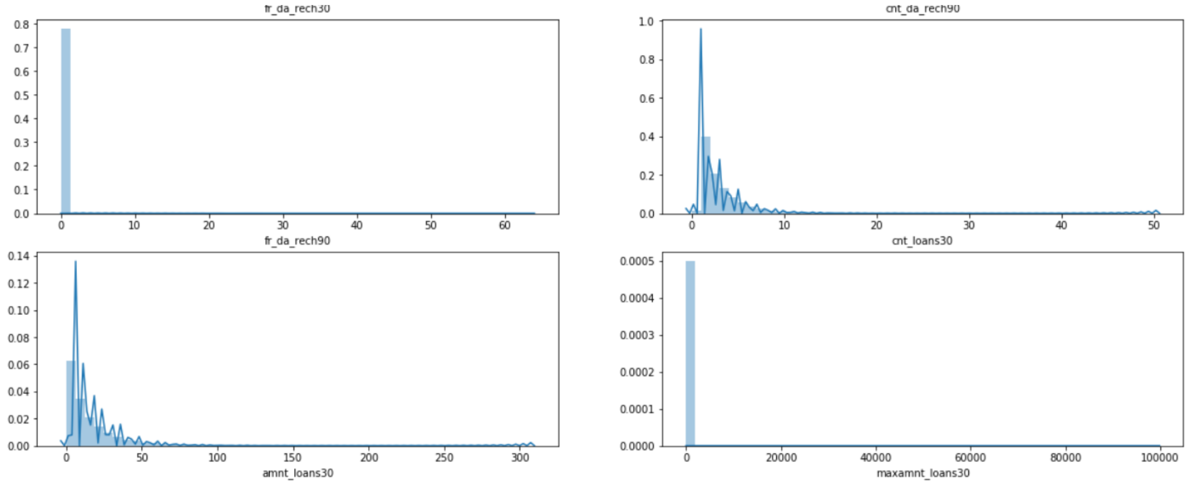
Data set contains 209593 rows and 33 features.

Data Preprocessing:

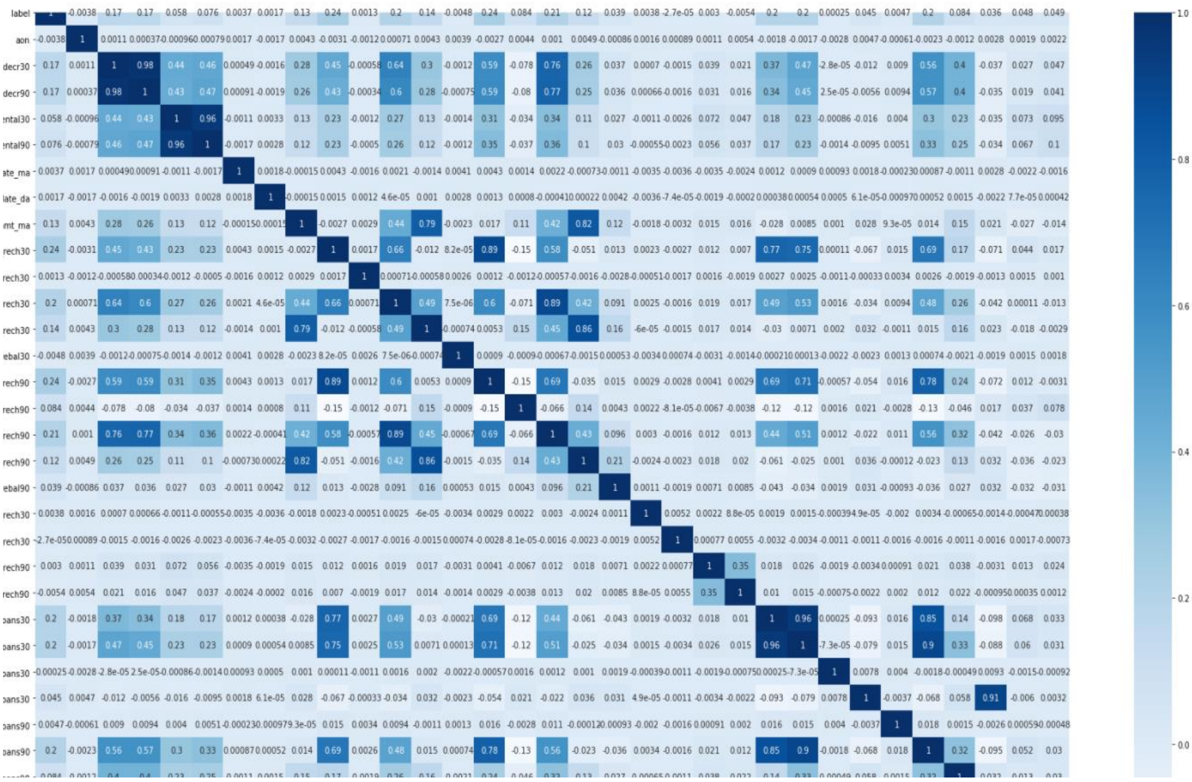
- Data type is checked.

```
label                          int64
msisdn                        object
aon                          float64
daily_decr30                 float64
daily_decr90                 float64
rental30                     float64
rental90                     float64
last_rech_date_ma            float64
last_rech_date_da            float64
last_rech_amt_ma               int64
cnt_ma_rech30                  int64
fr_ma_rech30                 float64
sumamnt_ma_rech30            float64
medianamnt_ma_rech30         float64
medianmarechprebal30         float64
cnt_ma_rech90                  int64
fr_ma_rech90                   int64
sumamnt_ma_rech90              int64
medianamnt_ma_rech90         float64
medianmarechprebal90         float64
cnt_da_rech30                float64
fr_da_rech30                 float64
cnt_da_rech90                  int64
fr_da_rech90                   int64
cnt_loans30                    int64
amnt_loans30                   int64
maxamnt_loans30              float64
medianamnt_loans30           float64
cnt_loans90                  float64
amnt_loans90                   int64
maxamnt_loans90                int64
medianamnt_loans90           float64
payback30                    float64
payback90                    float64
pcircle                       object
pdate                 datetime64[ns]
dtype: object
```

- Columns msisdn and pcircle has been drop as they are not giving any information.
- Univariate analysis is done to check the nature of data, if positive or negative skewed.



- All features are right skewed also few features has negative values too.
- Bivariate analysis is done to understand the relationship with dependent and independent variables.



- Dark shades represent high correlation.
- Following independent variables are highly correlated with dependent variable:

  * daily_decr30

  * daily_decr90

  * rental30

  * rental90

  * last_rech_amt_ma

  * cnt_ma_rech30

  * sumamnt_ma_rech30

  * medianamnt_ma_rech30

  * cnt_ma_rech90

  * sumamnt_ma_rech90

  * medianamnt_ma_rech90

  * cnt_loans30

  * amnt_loans30

  * amnt_loans90

- Treating Outliers
- We count days between 29 march 2007 and pdate column and nan those which are greater than aon.
- Then separately we treatment all columns one by one.
- We split our data into train and test and also stratify target variable to make it balanced

- We used StandardScaler and PCA to scale and select columns which are highly correlated with dependable variable.
- Below are details of each feature.

Variable	Definition	Comment
label	Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}	
msisdn	mobile number of user	
aon	age on cellular network in days	
daily_decr30	Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)	
daily_decr90	Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)	
rental30	Average main account balance over last 30 days	Unsure of given definition
rental90	Average main account balance over last 90 days	Unsure of given definition
last_rech_date_ma	Number of days till last recharge of main account	
last_rech_date_da	Number of days till last recharge of data account	
last_rech_amt_ma	Amount of last recharge of main account (in Indonesian Rupiah)	
cnt_ma_rech30	Number of times main account got recharged in last 30 days	
fr_ma_rech30	Frequency of main account recharged in last 30 days	Unsure of given definition
sumamnt_ma_rech30	Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)	
medianamnt_ma_rech30	Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)	
medianmarechprebal30	Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)	
cnt_ma_rech90	Number of times main account got recharged in last 90 days	
fr_ma_rech90	Frequency of main account recharged in last 90 days	Unsure of given definition
sumamnt_ma_rech90	Total amount of recharge in main account over last 90 days (in Indian Rupee)	
medianamnt_ma_rech90	Median of amount of recharges done in main account over last 90 days at user level (in Indian Rupee)	
medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level (in Indian Rupee)	
cnt_da_rech30	Number of times data account got recharged in last 30 days	
fr_da_rech30	Frequency of data account recharged in last 30 days	
cnt_da_rech90	Number of times data account got recharged in last 90 days	
fr_da_rech90	Frequency of data account recharged in last 90 days	
cnt_loans30	Number of loans taken by user in last 30 days	
amnt_loans30	Total amount of loans taken by user in last 30 days	
maxamnt_loans30	maximum amount of loan taken by the user in last 30 days	There are only two options: 5
medianamnt_loans30	Median of amounts of loan taken by the user in last 30 days	
cnt_loans90	Number of loans taken by user in last 90 days	
amnt_loans90	Total amount of loans taken by user in last 90 days	
maxamnt_loans90	maximum amount of loan taken by the user in last 90 days	
medianamnt_loans90	Median of amounts of loan taken by the user in last 90 days	
payback30	Average payback time in days over last 30 days	
payback90	Average payback time in days over last 90 days	
pcircle	telecom circle	
pdate	date	

- We have use 5 models for classification: They are:
- Logistic Regression
- GaussianNB
- Knn
- DecisionTree
- RandomForest

Model Deployment

After analyzing the problem statement, the best way to predict default customers is to use classifications model because classification models take inputs, analyse them and give predicted result.

To evaluate our models we will use :

- Confusion matrix
- F1 score
- Recall
- Precision
- True Negatives
- False Negative

We will not evaluate or model with accuracy score, since our data is imbalanced. Therefore we will focus on F1 score, Precision and Recall.

- **True Negatives:** Ratio of actual negative prediction over total negatives. Higher the value better it is.
- **False Negatives:** Ratio of actual positive predicted as negatives over actual positive.
- **Precision:** Out of all positives, how many are actually positive.
- **Recall:** Out of all actual positives how many have been predicted as positive.

Following are the models along with their evaluation.

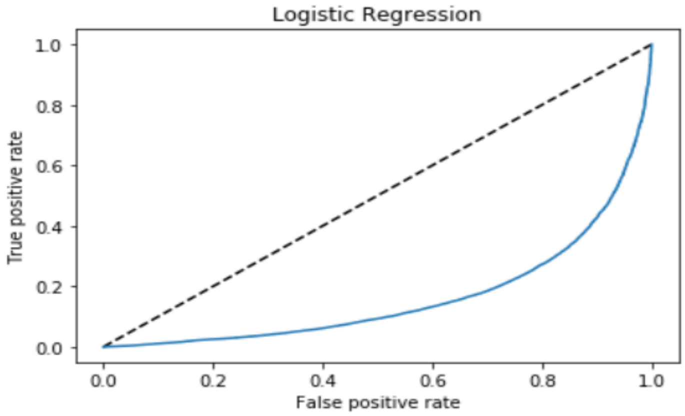
Logistic Regression:

```
print(classification_report(y_test,pred))
```

```
[1 1 1 ... 1 1 1]
F1_score: 0.934235373066779
[[  875  7758]
 [  670 59863]]
              precision    recall  f1-score   support

           0       0.57      0.10      0.17      8633
           1       0.89      0.99      0.93     60533

    accuracy                           0.88     69166
   macro avg       0.73      0.55      0.55     69166
weighted avg       0.85      0.88      0.84     69166
```

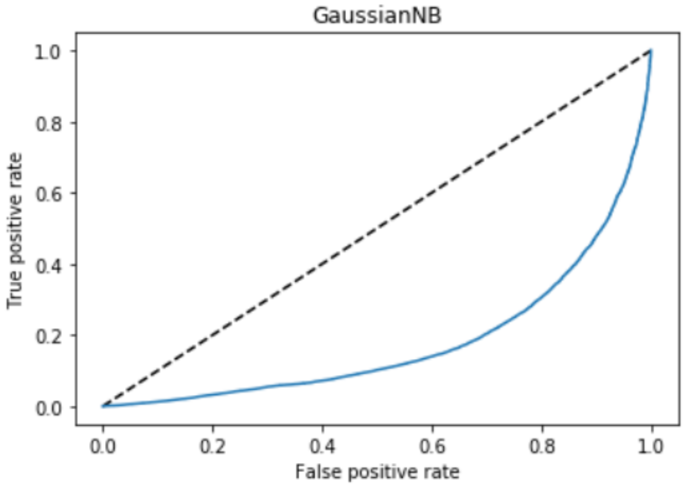


GaussianNB:

```
F1_score: 0.7901580900616579
[[ 7063  1570]
 [19973 40560]]
              precision    recall  f1-score   support

           0       0.26      0.82      0.40      8633
           1       0.96      0.67      0.79     60533

    accuracy                           0.69     69166
   macro avg       0.61      0.74      0.59     69166
weighted avg       0.88      0.69      0.74     69166
```

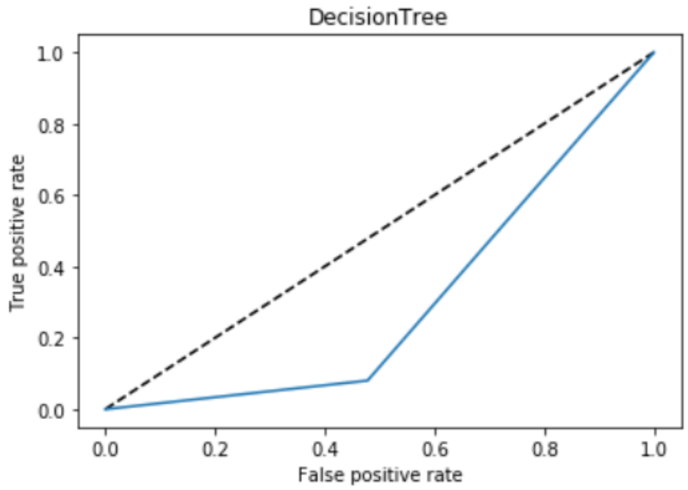


Decision Tree:

```
f1_score 0.9226832298136646
[[ 4123  4510]
 [ 4826 55707]]
              precision    recall  f1-score   support

           0       0.46      0.48      0.47      8633
           1       0.93      0.92      0.92     60533

    accuracy                           0.87     69166
   macro avg       0.69      0.70      0.70     69166
weighted avg       0.87      0.87      0.87     69166
```

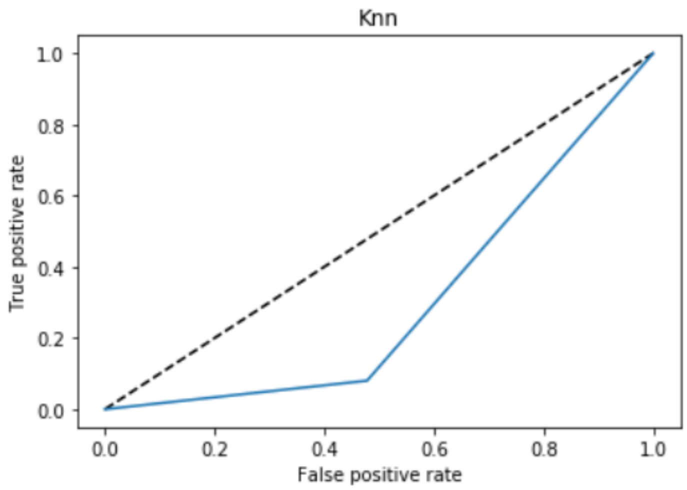


Knn:

```
f1_score 0.9165129227956453
[[ 3978  4655]
 [ 5391 55142]]
              precision    recall  f1-score   support

           0       0.42      0.46      0.44      8633
           1       0.92      0.91      0.92     60533

    accuracy                           0.85     69166
   macro avg       0.67      0.69      0.68     69166
weighted avg       0.86      0.85      0.86     69166
```

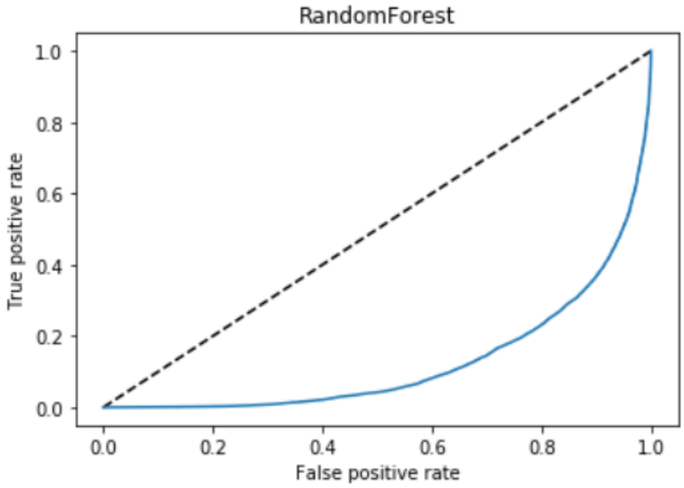


Random Forest:

```
0.9477564329765364
[[ 2368  6265]
 [  368 60165]]
              precision    recall  f1-score   support

           0       0.87      0.27      0.42      8633
           1       0.91      0.99      0.95     60533

    accuracy                           0.90     69166
   macro avg       0.89      0.63      0.68     69166
weighted avg       0.90      0.90      0.88     69166
```



Observation:

Model Name	**Test Score**	**True Negatives**	**False Negatives**
Logistic Regression	0.934	0.988	0.898
GaussianNB	0.790	0.670	0.181
Decision Tree	0.922	0.920	0.522
Knn	0.916	0.910	0.539
Random Forest	0.947	0.993	0.725

	Models	train_score	test_score	cross_val_score
0	Lg	0.879268	0.934235	93.485268
1	gnb	0.691640	0.790158	93.485268
2	dct	0.999665	0.922683	92.245070
3	Knn	0.909811	0.916513	91.830437
4	rf	0.905438	0.947756	94.813009

Random forest is performing best as it is not overfitted or underfitted also True Negative Score is high.

Conclusion

After analyzing the data set following features should be take more care to predict if the customer could be a default customer or not.

* daily_decr30
* daily_decr90
* rental30
* rental90
* last_rech_amt_ma
* cnt_ma_rech30
* sumamnt_ma_rech30
* medianamnt_ma_rech30
* cnt_ma_rech90
* sumamnt_ma_rech90
* medianamnt_ma_rech90
* cnt_loans30
* amnt_loans30
* amnt_loans90

Also, Random Forest is the best performing model with test score of 94% with high true negative value signifying the model is good in predicting default customer which was our main concern.

	Test Score	True Negatives	True Positives
Random Forest	0.947	0.993	0.725

References:

1. https://www.projectguru.in/challenges-indian-microfinance-industry/
2. https://www.nelito.com/blog/challenges-faced-microfinance-institutions.html
3. https://www.investopedia.com/terms/m/microfinance.asp
4. https://microfinanceinfo.com/micro-financial-institutions/
5. https://www.bankbazaar.com/personal-loan/microfinance-institutions.html
6. https://www.bankbazaar.com/personal-loan/microfinance-institutions.html