Micro-Credit Defaulter Model

Submitted by:
Aneesha B Soman

ACKNOWLEDGMENT

Acknowledgement The success and final outcome of this project required a lot of guidance and assistance from Sajid Choudhary Sir and I am Extremely fortunate to have got this all along the completion of my project work Whatever I have done is only due to such guidance and assistance and I would not forget to thank him.

I respect and thank Sajid Choudhary Sir, for giving me an opportunity to do the project work in Data Modelling and Analytics and providing us all support and guidance which made me complete the project on time . I am extremely grateful to him for providing such a nice support and guidance though he had busy schedule managing the company affairs.

I have also referred to various articles in Towards Data Science and Kaggle to obtain codes on various visualisation methods.

INTRODUCTION

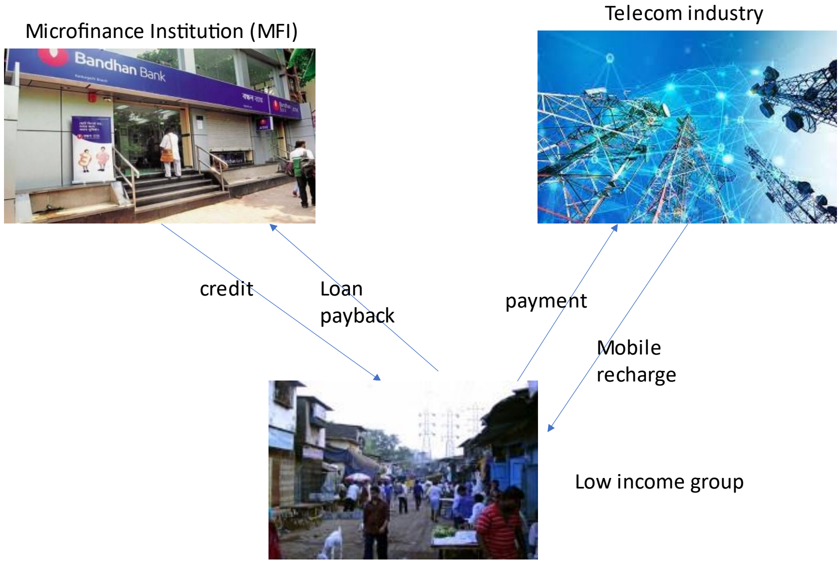
Business Problem Framing

The project deals in obtaining a better understanding the debtors behaviour and analysing the various parameters are more tending towards default. By understanding this, the Micro finance institution can focus on such debtors and also create marketing models for the other categories of creditors.

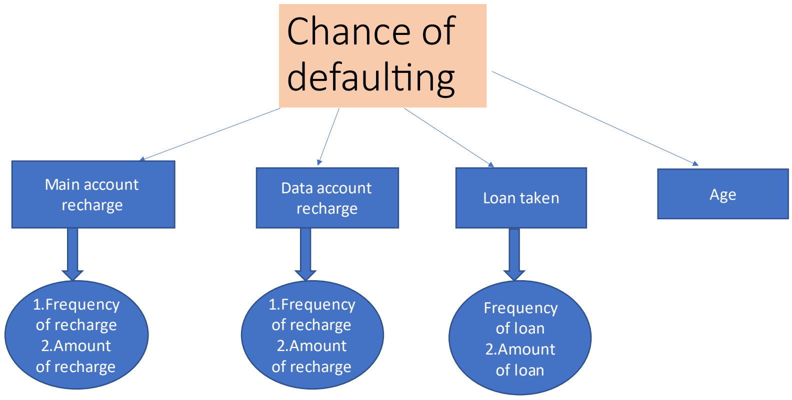
Conceptual Background of the Domain Problem

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 (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

A telecom industry is collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days



Review of Literature

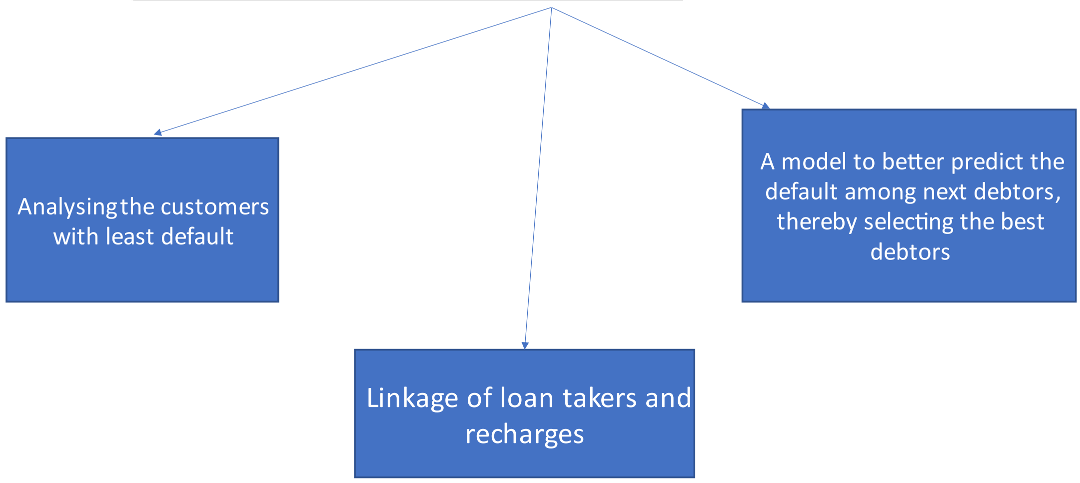


The chances of defaulting depends on these, where the highest correlation is seen with
Number of times main account got recharged
Total amount of recharge in main account over last 90 days

Motivation for the Problem Undertaken

Telecommunication has various advantages like Quick and accessible communication
saves time,Saves gasoline (do not need to drive distance),More than two people can communicate with at least one another at an equivalent time,Next “best thing” to being there,Easy to exchange ideas and knowledge via phone and/or fax,Worldwide access,Easy access to the people you would like to contact,Less effort in using transportation just to satisfy a private personally and many more. However this importance enlarges when it is for the Lower Income group as it becomes a critical factor in raising them out of poverty.
A MFI is partnering with the telecom company to provide credit. However the mission cannot go on for long if the customer defaults the payment to the MFI. Hence the Project aims to find the best debtors, whereby the social issue is addressed and the mission can continue for long

business task the project address



Monetary Benfits from the model

1. Focus on parameters on customers who rarely defaults
2. Improve Marketing towards customers who have good payment history
3. Charge higher interest rate on those who might default

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

Mathematical model	Data is analysed statistically Analysed through variance inflation factor Analysed through correlation and multicollinearity
Analytical graphs	Graphical modelling done through seaborn and matplotlib

Data Sources and their formats

1.Data orgin:

Data is obtained from Telecom operator who has a tieup with an MFI.

2.Description of data:

a.Data obtained was in csv format

b.Data had 37 columns and the column names were :

'unamed', 'label', 'msisdn', 'aon', 'daily_decr30', 'daily_decr90','rental30', 'rental90','last_rech_date_ma', 'last_rech_date_da', 'last_rech_amt_ma', 'cnt_ma_rech30','fr_ma_rech30','sumamnt_ma_rech30','medianamnt_ma_rech30','medianmarechprebal30', 'cnt_ma_rech90', 'fr_ma_rech90', 'sumamnt_ma_rech90','medianamnt_ma_rech90', 'medianmarechprebal90', 'cnt_da_rech30', 'fr_da_rech30', 'cnt_da_rech90', 'fr_da_rech90', 'cnt_loans30', 'amnt_loans30', 'maxamnt_loans30', 'medianamnt_loans30', 'cnt_loans90', 'amnt_loans90', 'maxamnt_loans90', 'medianamnt_loans90', 'payback30', 'payback90', 'pcircle', 'pdate'

	unamed	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	last_rech_amt_ma
0	1	0	21408I70789	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	1539
1	2	1	76462I70374	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	5787
2	3	1	17943I70372	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	1539
3	4	1	55773I70781	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	947
4	5	1	03813I82730	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	2309

fr_ma_rech30	sumamnt_ma_rech30	medianamnt_ma_rech30	medianmarechprebal30	cnt_ma_rech90	fr_ma_rech90	sumamnt_ma_rech90
21.0	3078.0	1539.0	7.50	2	21	3078
0.0	5787.0	5787.0	61.04	1	0	5787
0.0	1539.0	1539.0	66.32	1	0	1539
0.0	0.0	0.0	0.00	1	0	947
2.0	20029.0	2309.0	29.00	8	2	23496

medianamnt_ma_rech90	medianmarechprebal90	cnt_da_rech30	fr_da_rech30	cnt_da_rech90	fr_da_rech90	cnt_loans30	amnt_loans30	maxamnt_loans30
1539.0	7.50	0.0	0.0	0	0	2	12	6.0
5787.0	61.04	0.0	0.0	0	0	1	12	12.0
1539.0	66.32	0.0	0.0	0	0	1	6	6.0
947.0	2.50	0.0	0.0	0	0	2	12	6.0
2888.0	35.00	0.0	0.0	0	0	7	42	6.0

medianamnt_loans30	cnt_loans90	amnt_loans90	maxamnt_loans90	medianamnt_loans90	payback30	payback90	pcircle	pdate
0.0	2.0	12	6	0.0	29.000000	29.000000	UPW	20-07-2016
0.0	1.0	12	12	0.0	0.000000	0.000000	UPW	10-08-2016
0.0	1.0	6	6	0.0	0.000000	0.000000	UPW	19-08-2016
0.0	2.0	12	6	0.0	0.000000	0.000000	UPW	06-06-2016
0.0	7.0	42	6	0.0	2.333333	2.333333	UPW	22-06-2016

c.There are both numerical and categorical columns. There is also a date column.

d.label means Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}

<u>3. Data engineering</u>

a. renaming of data columns done:

	unamed	label	mobile number of user	age on cellular network in days	Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)	Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)	Mean main account balance over last 30 days	Mean main account balance over last 90 days	Number of days till last recharge of main account	Number of days till last recharge of data account	Total Amount of last recharge of main account (in Indonesian Rupiah)	Number of times main account got recharged in last 30 days	Frequency of main account recharged in last 30 days
0	1	0	21408I70789	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	1539	2	21.0
1	2	1	76462I70374	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	5787	1	0.0
2	3	1	17943I70372	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	1539	1	0.0
3	4	1	55773I70781	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	947	0	0.0
4	5	1	03813I82730	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	2309	7	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
209588	209589	1	22758I85348	404.0	151.872333	151.872333	1089.19	1089.19	1.0	0.0	4048	3	2.0
209589	209590	1	95583I84455	1075.0	36.936000	36.936000	1728.36	1728.36	4.0	0.0	773	4	1.0
209590	209591	1	28556I85350	1013.0	11843.111670	11904.350000	5861.83	8893.20	3.0	0.0	1539	5	8.0

Frequency of main account recharged in last 30 days	Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)	Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)	Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)	Number of times main account got recharged in last 90 days	Frequency of main account recharged in last 90 days	Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)	Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)	Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah)	Number of times data account got recharged in last 30 days	Frequency of data account recharged in last 30 days	Number of times data account got recharged in last 90 days	Frequency of data account recharged in last 90 days
21.0	3078.0	1539.0	7.50	2	21	3078	1539.0	7.50	0.0	0.0	0	0
0.0	5787.0	5787.0	61.04	1	0	5787	5787.0	61.04	0.0	0.0	0	0
0.0	1539.0	1539.0	66.32	1	0	1539	1539.0	66.32	0.0	0.0	0	0
0.0	0.0	0.0	0.00	1	0	947	947.0	2.50	0.0	0.0	0	0

Number of times data account got recharged in last 90 days	Frequency of data account recharged in last 90 days	Number of loans taken by user in last 30 days	Total amount of loans taken by user in last 30 days	maximum amount of loan taken by the user in last 30 days	Median of amounts of loan taken by the user in last 30 days	Number of loans taken by user in last 90 days	Total amount of loans taken by user in last 90 days	maximum amount of loan taken by the user in last 90 days	Median of amounts of loan taken by the user in last 90 days	Mean payback time in days over last 30 days	Mean payback time in days over last 90 days	telecom circle	date
0	0	2	12	6.0	0.0	2.0	12	6	0.0	29.000000	29.000000	UPW	20-07-2016
0	0	1	12	12.0	0.0	1.0	12	12	0.0	0.000000	0.000000	UPW	10-08-2016
0	0	1	6	6.0	0.0	1.0	6	6	0.0	0.000000	0.000000	UPW	19-08-2016
0	0	2	12	6.0	0.0	2.0	12	6	0.0	0.000000	0.000000	UPW	06-06-2016

b.Dropped the unique value and unnecessary columns-unnamed, telecom circle,year and mobile number of user

c.Engineering date made into usable format

d.The label is imbalanced, hence balanced the data by oversampling it

4.Data Preprocessing Done

a.The data is assumed to be linear, Homogeneity of variances, Normality and Independence. Eda is done to remove the outliers to make data normal and linear.

b.EDA done is using Zscore. Only 8% data is allowed to be dropped

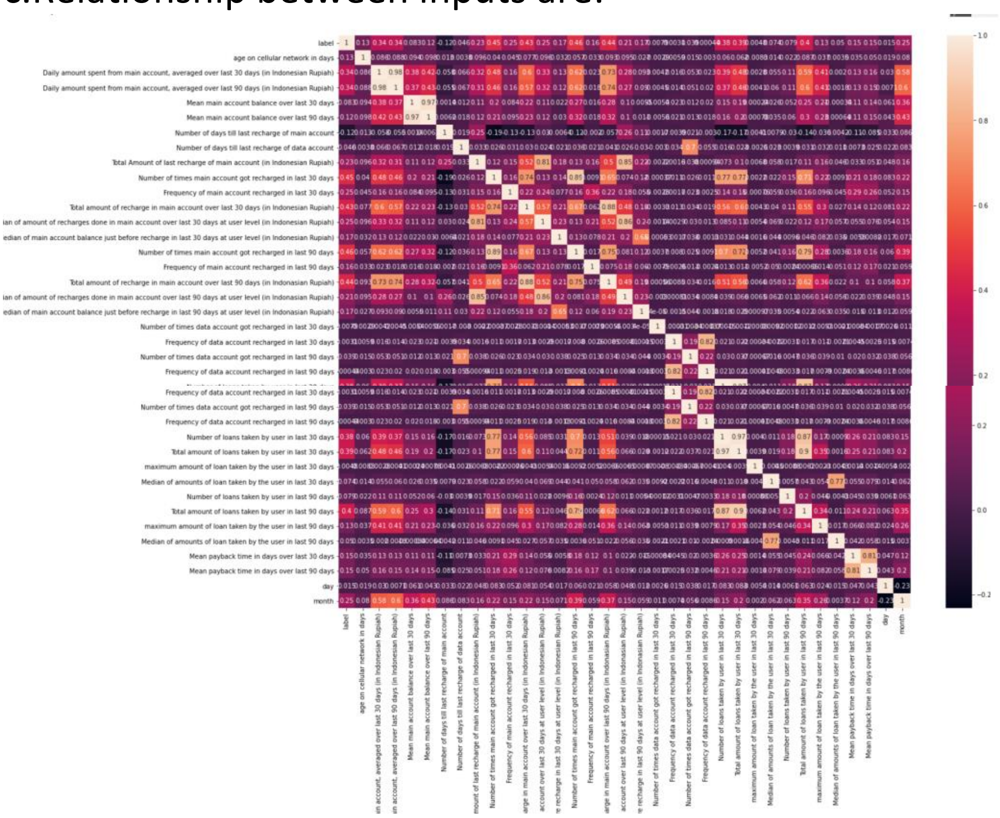
5.Data Inputs- Logic- Output Relationships

a.Input is numerical format and output is categorical format

b.Input and ouputs relationship is:

```
label                                                                                      1.000000
Number of times main account got recharged in last 90 days                                 0.462423
Number of times main account got recharged in last 30 days                                 0.448417
Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)          0.437448
Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)          0.425084
Total amount of loans taken by user in last 90 days                                        0.401482
Total amount of loans taken by user in last 30 days                                        0.385494
Number of loans taken by user in last 30 days                                              0.378190
Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)    0.342357
Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)    0.337201
Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)    0.249749
Frequency of main account recharged in last 30 days                                        0.247957
month                                                                                      0.247002
Total Amount of last recharge of main account (in Indonesian Rupiah)                       0.232737
Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)    0.212249
Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)     0.169449
Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah)     0.166191
Frequency of main account recharged in last 90 days                                        0.156048
Mean payback time in days over last 30 days                                                0.152427
Mean payback time in days over last 90 days                                                0.150471
maximum amount of loan taken by the user in last 90 days                                   0.131697
age on cellular network in days                                                            0.126196
Mean main account balance over last 90 days                                                0.118125
Mean main account balance over last 30 days                                                0.083477
Number of loans taken by user in last 90 days                                              0.078526
Median of amounts of loan taken by the user in last 30 days                                0.073765
Median of amounts of loan taken by the user in last 90 days                                0.050284
Number of days till last recharge of data account                                          0.046307
Number of times data account got recharged in last 90 days                                 0.038601
day                                                                                        0.014918
Number of times data account got recharged in last 30 days                                 0.007902
maximum amount of loan taken by the user in last 30 days                                   0.004796
Frequency of data account recharged in last 30 days                                        0.003093
Frequency of data account recharged in last 90 days                                        0.000436
Number of days till last recharge of main account                                         -0.117551
Name: label, dtype: float64
```

c.Relationship between inputs are:



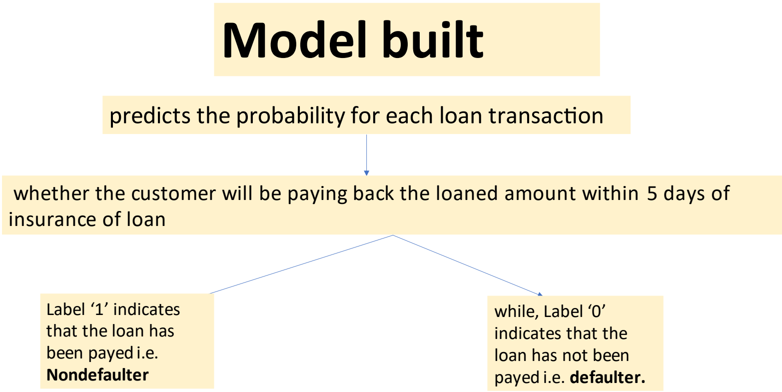
6.Hardware and Software Requirements and Tools Used

Library	**Used in the project**
Pandas library	1.Read the csv file, describe it,count values,converting date into usable format,dropping duplicates
Numpy library	Using zscore
Seaborn and matplotlib	For visualization
sklearn	Model building
GridSearchCV	hyperparameter tuning
pickle	saving data
Ridge,Lasso	Regularisation

Hardware: Windows 10

Softwares: Jupyter notebook

Model/s Development and Evaluation



1.Identification of possible problem-solving approaches

The data set was was analysed both statistically and graphically.
The statistical analysis showed that data to have outliers, to have no null values and that datas independent variable had numerical datas alone.Hence the datas outliers were removed(8%) and made more normalised
The Graphical analysis showed the dependent variable to be highly imbalanced and hence the need was there to sample it.Hence the label was undersampled and the Number of Yes and No were made equal to 26162

2. Testing of Identified Approaches (Algorithms)

The label was categorical hence classification algorithms were used, which were
- logistic regression
- K-nearest neighbours
- Random forest
- DecisionTreeClassifier

3.Run and Evaluate selected models

1.Models used:

```python
#modelling

from sklearn.linear_model import LogisticRegression

LR=LogisticRegression()
LR.fit(x_train,y_train)
predlr=LR.predict(x_test)
print(accuracy_score(y_test,predlr))
print(confusion_matrix(y_test,predlr))
print(classification_report(y_test,predlr))
```

```python
#modelling
from sklearn.tree import DecisionTreeClassifier

dt=DecisionTreeClassifier()
dt.fit(x_train,y_train)
preddt=dt.predict(x_test)

print(accuracy_score(y_test,preddt))
print(confusion_matrix(y_test,preddt))
print(classification_report(y_test,preddt))
```

```python
#modelling
from sklearn.ensemble import RandomForestClassifier

rf=RandomForestClassifier()
rf.fit(x_train,y_train)
predrf=rf.predict(x_test)

print(accuracy_score(y_test,predrf))
print(confusion_matrix(y_test,predrf))
print(classification_report(y_test,predrf))
```

```python
#modelling
from sklearn.svm import SVC

svc=SVC()
svc.fit(x_train,y_train)

ad_pred=svc.predict(x_test)

print(accuracy_score(y_test,ad_pred))
print(confusion_matrix(y_test,ad_pred))
print(classification_report(y_test,ad_pred))
```

2.Accuracy score,f1 score,precision,classification matrix ,f1 score and recall was obtained for each of it:

a.Logistic regression

```
[[4786 1121]
 [1573 4054]]
              precision    recall  f1-score   support

           0       0.75      0.81      0.78      5907
           1       0.78      0.72      0.75      5627

    accuracy                           0.77     11534
   macro avg       0.77      0.77      0.77     11534
weighted avg       0.77      0.77      0.77     11534
```

b.Decision tree classifier

```
[[4737 1170]
 [1132 4495]]
              precision    recall  f1-score   support

           0       0.81      0.80      0.80      5907
           1       0.79      0.80      0.80      5627

    accuracy                           0.80     11534
   macro avg       0.80      0.80      0.80     11534
weighted avg       0.80      0.80      0.80     11534
```

c.RandomForestClassifier

```
[[4970  937]
 [ 786 4841]]
              precision    recall  f1-score   support

           0       0.86      0.84      0.85      5907
           1       0.84      0.86      0.85      5627

    accuracy                           0.85     11534
   macro avg       0.85      0.85      0.85     11534
weighted avg       0.85      0.85      0.85     11534
```

d. SVC

```
[[4947  960]
 [1202 4425]]
              precision    recall  f1-score   support

           0       0.80      0.84      0.82      5907
           1       0.82      0.79      0.80      5627

    accuracy                           0.81     11534
   macro avg       0.81      0.81      0.81     11534
weighted avg       0.81      0.81      0.81     11534
```

Modelling was tried on 4 classification techniques

	Logistic regression model	DecisionTreeClassifier model	RandomForest Classifier	SVC
Accuracy score	78	80	85	81
Cross validation score	74.98	79.84	85.14	76.19
roc_auc_score	76.53	7996	85.32	78.42

ii.Cross validation score of each of the algorithm:

Logistic Regression	0.751175915619696
Decision Tree Model	0.7940953503336645
Random forest Model	0.8521881712147857
SVC Model	0.7633797817322845

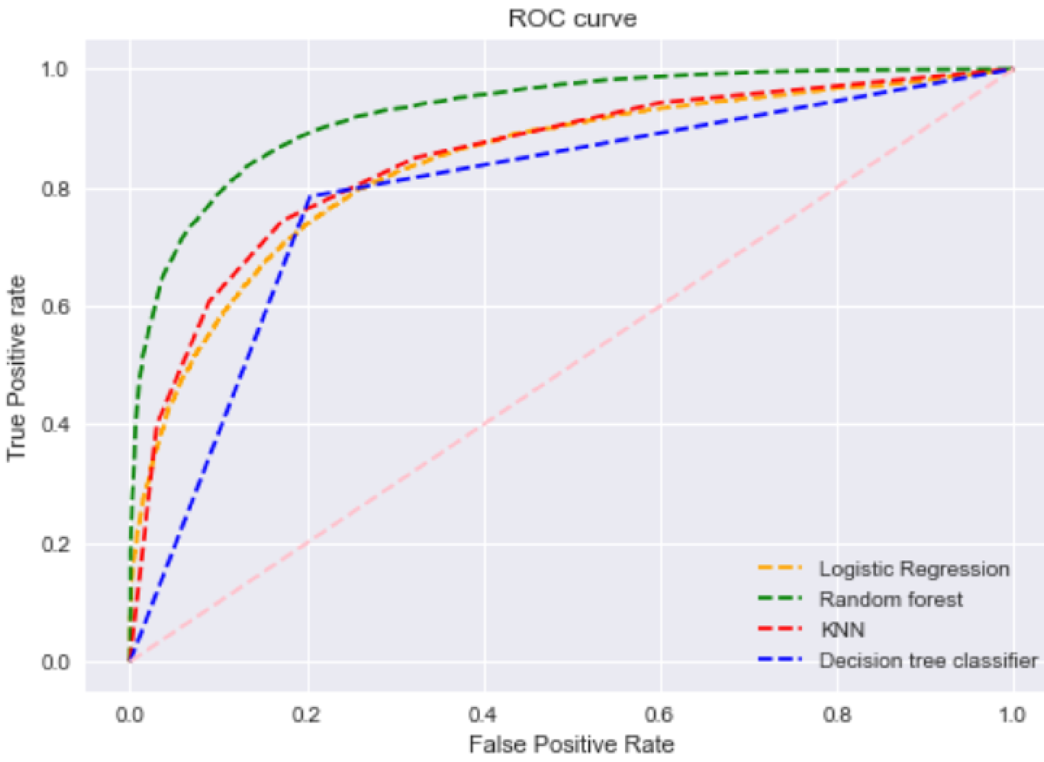
iii.ROC AUC Score

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'.

Logistic Regression	0.7653400529726067
Decision Tree Model	0.7996377504539964
Random forest Model	0.8532703109921091
SVC Model	0.7842754869182716

iv.ROC AUC curve

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.
The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

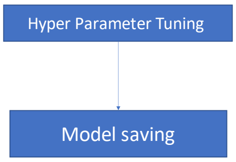


4.Key Metrics for success in solving problem under consideration

A.The **accuracy score, ROC AUC score and ROC AUC curve** was used **Accuracy is used when the True Positives and True negatives are more important** while F1-score is used when the False Negatives and False Positives are crucial.
However since we are more concerned about the True Positive and True Negative we would be preferring the Accuracy score over F1 score

B.Since the Accuracy score and ROC AUC score of Randomforest Classifier is highest, we would be selecting it for hyperparamter tuning

C.Hyperparameter tuning



```
#HyperParameter tuning
from sklearn.model_selection import GridSearchCV
paramters={'n_estimators':[2,3,4,5],
           'criterion':['gini','entropy'],
           'min_samples_split':[2,3,4,5],
           'min_samples_leaf':[2,3,4,5,6],
           'max_leaf_nodes':[2,3,4,5,10],
}

GCV=GridSearchCV(RandomForestClassifier(),paramters,cv=5)

GCV.fit(x_train,y_train)
```

```
98]: GridSearchCV(cv=5, estimator=RandomForestClassifier(),
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_leaf_nodes': [2, 3, 4, 5, 10],
                         'min_samples_leaf': [2, 3, 4, 5, 6],
                         'min_samples_split': [2, 3, 4, 5],
                         'n_estimators': [2, 3, 4, 5]})
```

```
GCV.best_params_
```

```
{'criterion': 'entropy',
 'max_leaf_nodes': 10,
 'min_samples_leaf': 3,
 'min_samples_split': 5,
 'n_estimators': 4}
```

```
mod=RandomForestClassifier(criterion='gini',max_leaf_nodes=10,min_samples_leaf=2,min_samples_split=4,n_estimators=5)
```

```
mod.fit(x_train,y_train)
pred=mod.predict(x_test)
print(accuracy_score(y_test,pred)*100)
```

```
79.44030497314158
```

```
classifier=RandomForestClassifier()
classifier.fit(x_train,y_train)
```

```
]: RandomForestClassifier()
```

D.Model scores

```
scr=cross_val_score(classifier,x,y,cv=5)
print("Cross validation score of Random forest model :", scr.mean())
```

```
Cross validation score of Random forest model : 0.8537853352106575
```

```
classifier.fit(x_train,y_train)
pred=classifier.predict(x_test)
print("accuracy score of the Random Forest model is",accuracy_score(y_test,pred)*100)
```

```
accuracy score of the Random Forest model is 85.69571997920637
```

```
classifier.fit(x_train,y_train)
print("ROC AUC Score of the Random forest model is",roc_auc_score(y_test,classifier.predict(x_test)))
```

```
ROC AUC Score of the Random forest model is 0.8551406183986445
```

```
print(confusion_matrix(y_test, y_pred))
```

```
[[5019  868]
 [ 807 4848]]
```

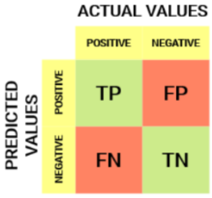
Model scores obtained are:

Cross validation score	85.37
Accuracy score	85.69
ROC AUC score	85.51

Summary of Classification scores

Classification table :

```
[[5019  868]
 [ 807 4848]]
```



1 means loan has been payed(Non defaulter)
0 means loan was not payed within the time frame(Defaulter)

```
print(classification_report(y_test, y_pred))
```

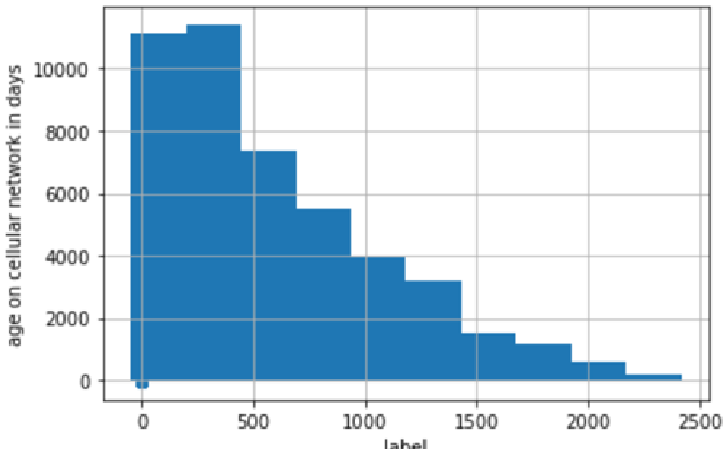
```
              precision    recall  f1-score   support

           0       0.86      0.85      0.86      5887
           1       0.85      0.86      0.85      5655

    accuracy                           0.85     11542
   macro avg       0.85      0.85      0.85     11542
weighted avg       0.85      0.85      0.85     11542
```

The summary of the classification report is presented below.

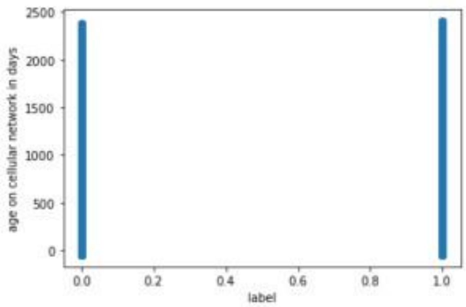
	Formula	Definition	Value
Sensitivity (recall of Non defaulter)	derived from: True positive/(True positive + False negative)	Sensitivity summarizes our true positive rate, which is how many we got correct out of all the positive cases.	85%
Specificity (recall of defaulter)	True negative/(True negative + False positive)	Specificity summarizes our true negative rate, which is how many we got correct out of all the negative cases.	86%
Precision of Non defaulter cases	True positive/(True positive + False positive)	Precision of non defaulter cases summarize the accuracy of fraud cases detected. That is, out of all that I predicted as fraud, how many are correct.	86%
Precision of defaulter cases	True negative/(True negative + False negative)	Precision of defaulter cases summarize the accuracy of non-fraud cases detected. That is, out of all that I predicted as non-fraud, how many are correct.	85%
F1 scores	(2 x recall x precision)\(recall + precision)	As we are interested in defaulter cases, only the F1 scores on fraud cases are reported.	86%

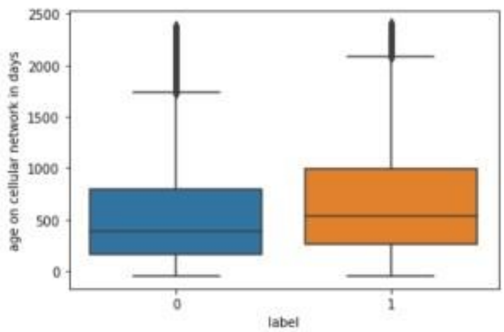
5.Visualizations

1. age on cellular network in days



Ages on the celllar network varies between 0 to 2400 days





The default has been seen

1.Median is 480days

2.Minium is 0days

3.Maximum is 1750days

2. A person can recharge two accounts

- Main account recharge
- Data account recharge

Main Account

Number of times recharge done in last 30

Median of main account balance just before recharge in last 30 days at user level

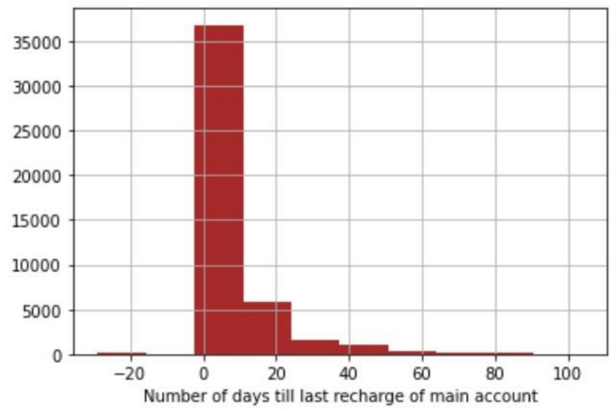
Median of amount of recharges done in main account over last 30 days at user level

Daily amount spent from main account, averaged over last 30 days

Total amount of recharge in main account over last 30 days

Mean of main account balance over last 30 days

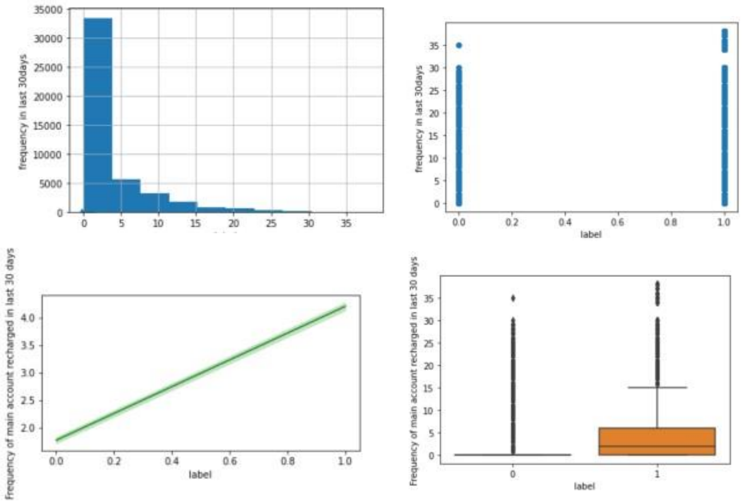
Number of days till last recharge of main account



Most of them recharged within last 10days of taking the sample

A.For Last 30days:

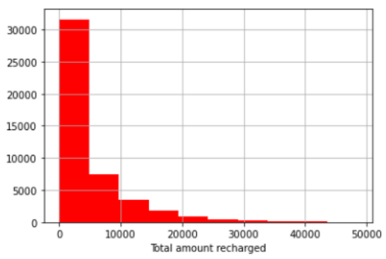
Number of times recharge done in last 30



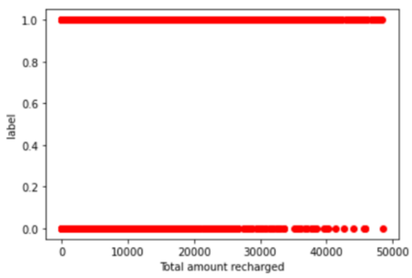
1.Maximum times recharge was done was between 1 and 4

2.Default is less if the number of recharge is above 30
i.e a person with more recharging potential will not default much

Total amount recharged

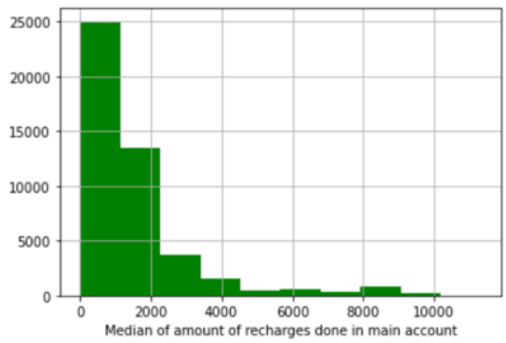


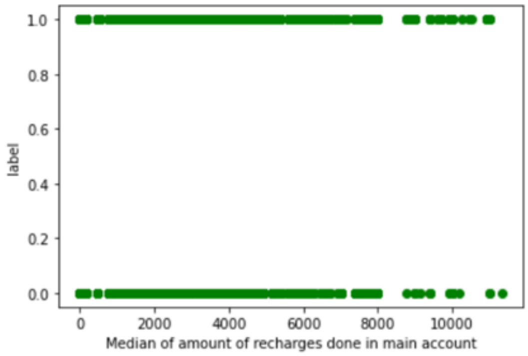
The total amount recharged is maximum between 1 and 9500



The people who have recharged above 30,000 have defaulted less

Median of amount of recharges done in main account

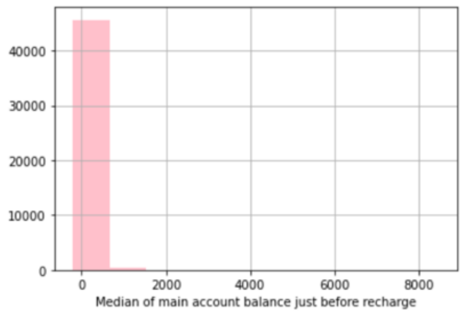


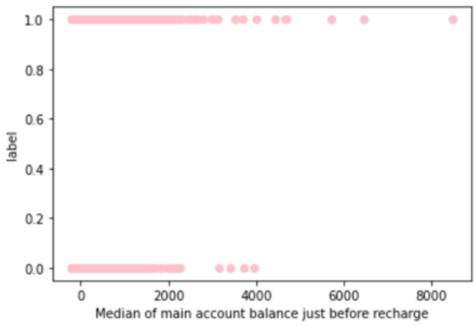


Median of amount of recharge is maximum between 0 and 1000

The median of amount of recharge is not much dependent on the label

Median of main account balance just before recharge



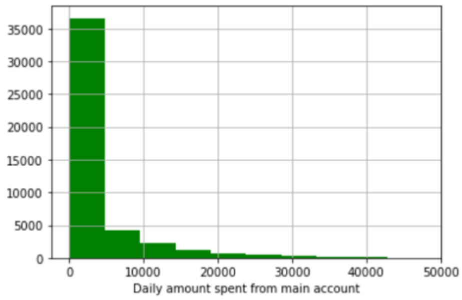


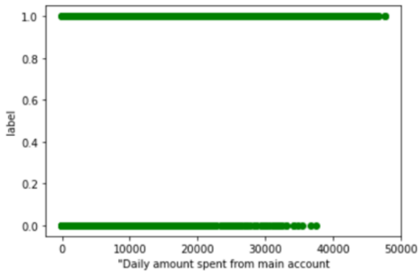
The maximum account balance median before recharge was between 0 and 700

Median of balance before recharge when it was above 2500, there was lesser chances to default

B.LAST 90days

Daily amount spent from main account

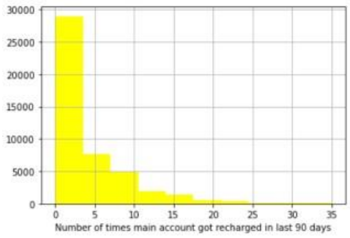


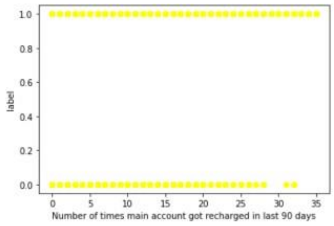


Maximum daily spent in last 90 days is between 0 and 5000

The default rate is lower when the person uses daily amount above 39,000

Number of times main account got recharged in last 90 days

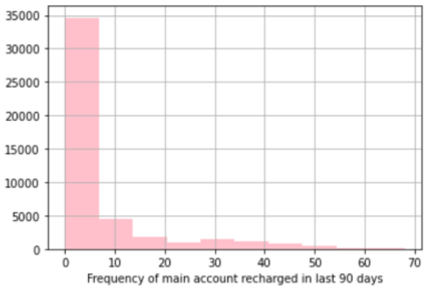


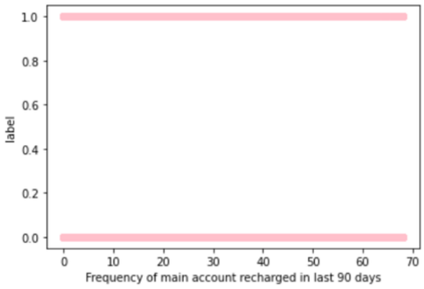


Maximum number of times is between 1 and 4

Above 28 times, the default rate is less

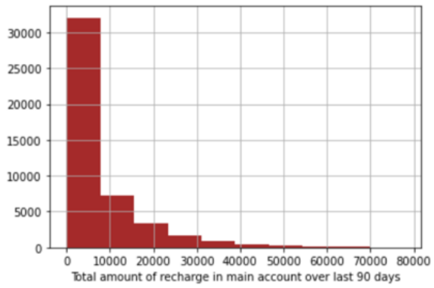
Frequency of main account recharged in last 90 days

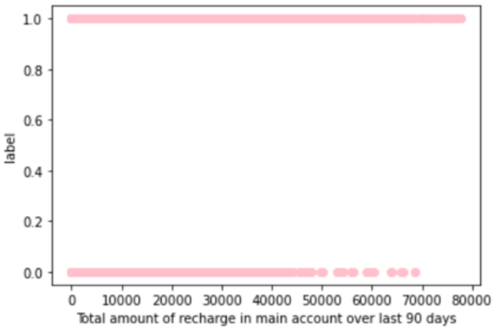




Frequency does not impact the label

Total amount of recharge in main account over last 90 days

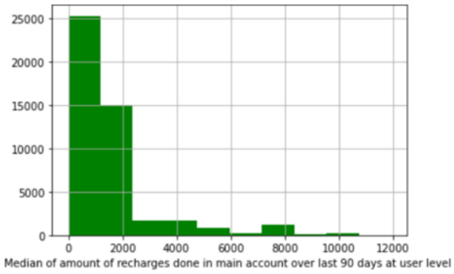


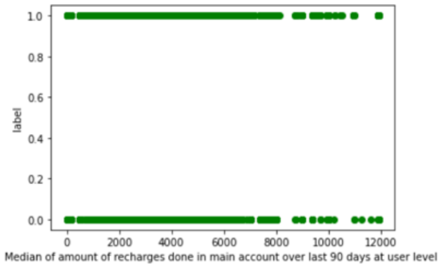


Total amount of recharge is maximum upto 9,000

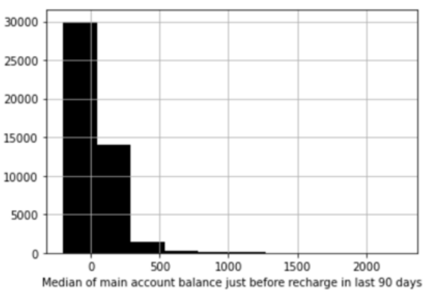
The default rate decreases for the customers who recharged their main account in last 90 days above 50,000

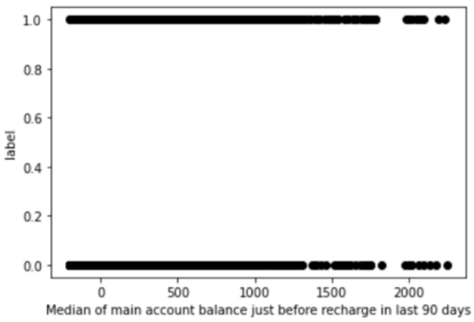
Median of amount of recharges done in main account over last 90 days at user level





Median of main account balance just before recharge in last 90 days

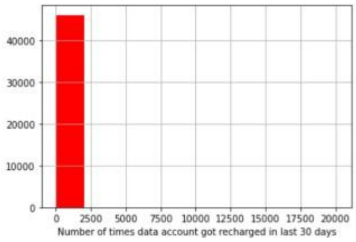


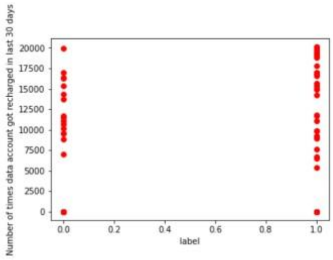


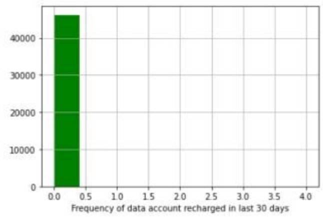
B.DATA ACCOUNT

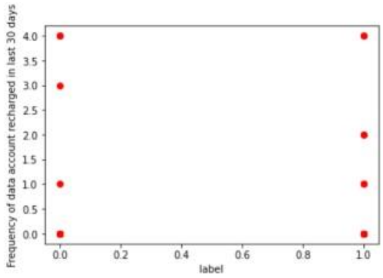
A.For Last 30days:

Number of times data account got recharged in last 30 days





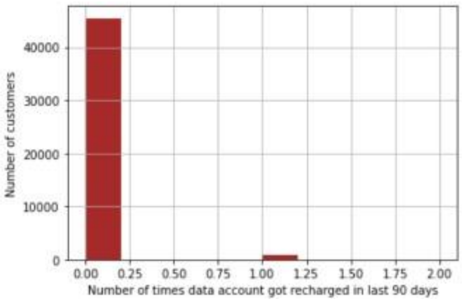


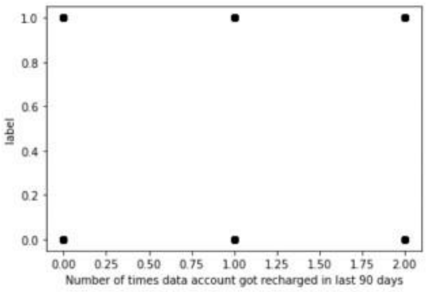


The people who had recharged the data account between 7,500 and 17500 have shown defaulting

B.For Last 90days

Number of times data account got recharged in last 90 days





Number of times is 40,000 and there is no impact on the label

3.LOAN

MFI bank

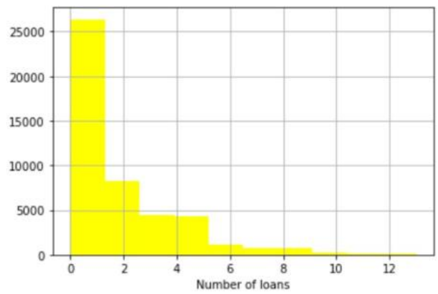
loan amount of 5

payback amount should be 6

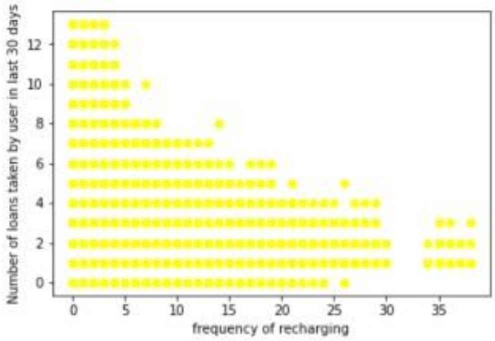
loan amount of 10

payback amount should be 12

Number of loans taken by user in last 30 days

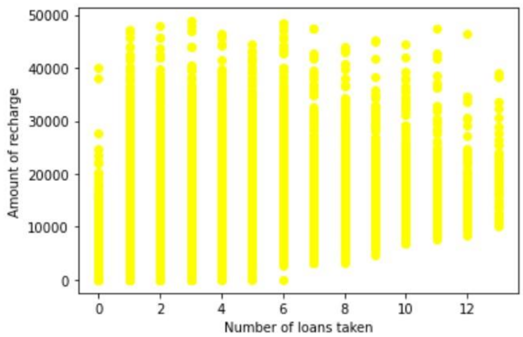


Between 0 and 2 is the maximum loans taken

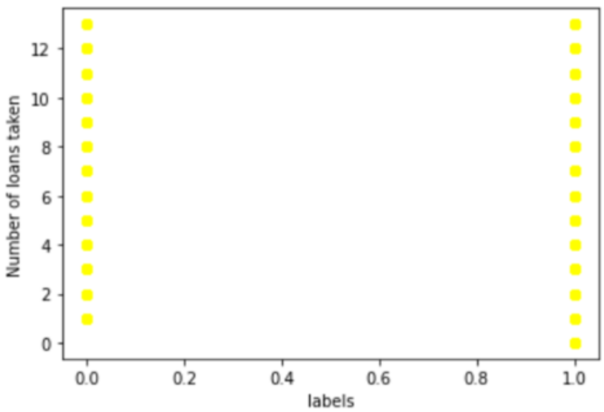


	Number of loans taken	Frequency of recharge
	1	0 to 24
	2	0 to 30
	3	0 to 29
	4	0 to 29
	5	0 to 20
	6	0 to 20
	7	0 to 14
	8	**0 to 9**
	9	0 to 6
	10	0 to 6
	11	0 to 5
	12	0 to 5

It can be seen the higher the loan taken,
The frequency of recharge is less

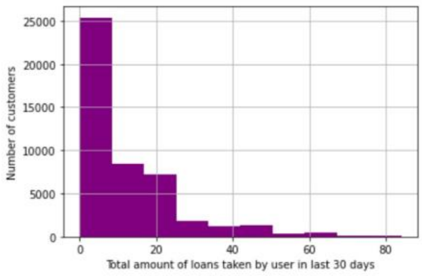


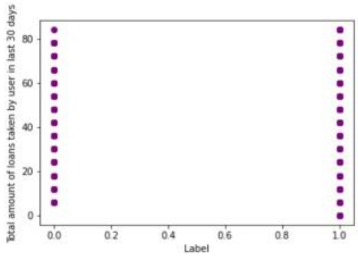
People who have taken loans 2 times,3 times and 6 times are seen to recharge with maximum amounts



The number of loans taken does not impact the Defaulting much

Amount of loan taken



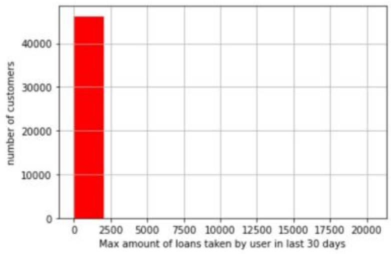


The total amount of loans taken has been between 0 and 7

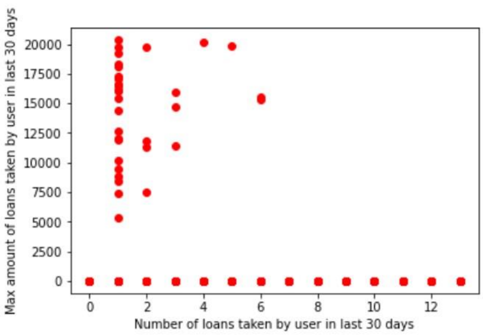
The amount of loan taken and default rate has not much interdependence

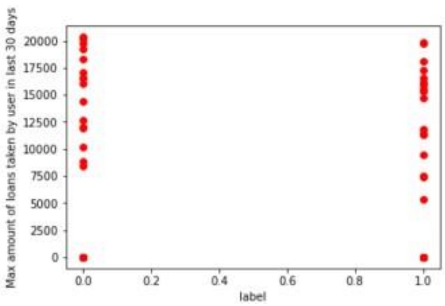
maximum amount of loan taken by the user in last 30 days

There are only two options: 5 & 10 Rs., for which the user needs to pay back 6 & 12 Rs. respectively



The highest loan amounts areupto 2,400

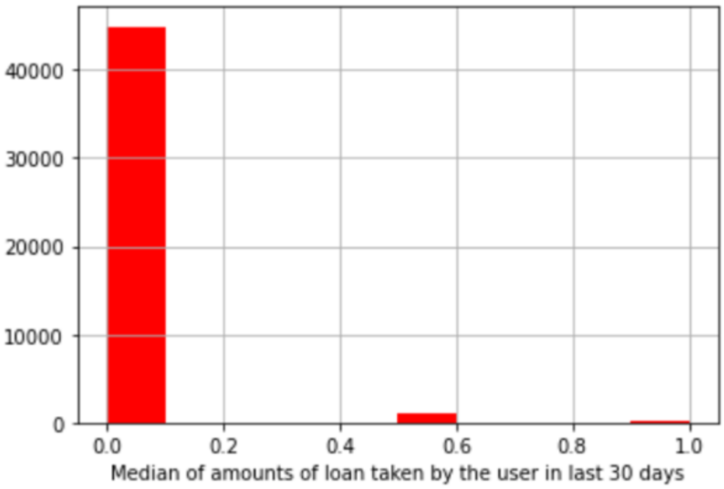




Maximum amount of loans are when number of loan taken is 1

Default rate is higher when max loan taken is above 17500

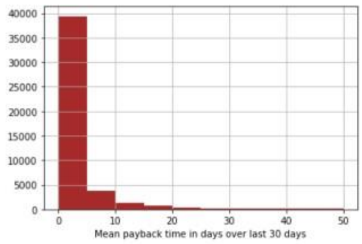
Median of amounts of loan taken by the user in last 30 days

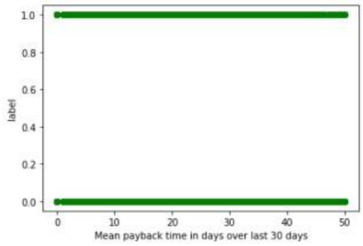


Median is within 0.1

4.Mean payback time

For 30days

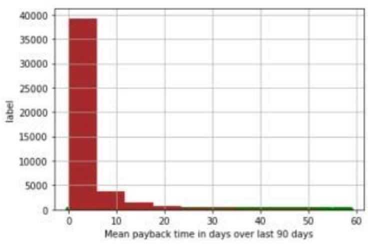


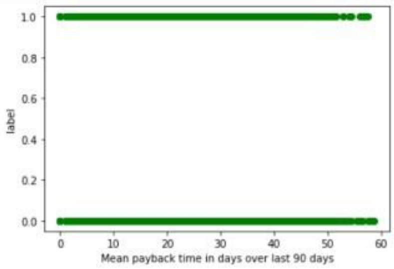


Mean payback varies between 1 and 9

It does not impact the labels

For 90 days





Mean payback is inside 7days mostly

It does not impact the label

CONCLUSION

1.Key Findings and Conclusions of the Study

This project has built a model that can detect Credit default. In doing so, the model can reduces loses for Micro Finance companies. The challenge behind credit default in machine learning is that credit default are far less common as compared to legit insurance claims.

Five different classifiers were used in this project: - logistic regression, K-nearest neighbours, Random forest and DecisionTreeClassifier. The best model for the dataset chosen is Random forest model. The Model can predict with an accuracy of 85.69 %, if the customer will default or not

Inferences from the Problem are:

a.the highest factors that determined the defaulting or not are:

- Number of times main account got recharged in last 90 days
- Number of times main account got recharged in last 30 days
- Total amount of recharge in main account over last 90 days (in Indonasian Rupiah
- Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
- Total amount of loans taken by user in last 90 days

b.Details:

i. Ages on the celllar network varies between 0 to 2400 days
ii. The default has been seen Minium is 0days and Maximum is 1750days
iii. **For Last 30days in main account**
    - Most of them recharged within last 10days of taking the sample
    - Maximum times recharge was done was between 1 and 4
    - Default is less if the number of recharge is above 30i.e a person with more recharging potential will not default much

- The total amount recharged is maximum between 1 and 9500
- The people who have recharged above 30,000 have defaulted less
- Median of amount of recharge is maximum between 0 and 1000
- The median of amount of recharge is not much dependent on the label
- The maximum account balance median before recharge was between 0 and 700
- Median of balance before recharge when it was above 2500, there was lesser chances to default

iv **For Last 90days in main account**

- Maximum daily spent in last 90 days is between 0 and 5000
- The default rate is lower when the person uses daily amount above 39,000
- Maximum Number of times main account got recharged is between 1 and 4 and Above 28 times, the default rate is less
- Frequency of main account recharged in last 90 days does not impact the label
- Total amount of recharge in main account over last 90 days maximum upto 9,000
- The default rate decreases for the customers who recharged their main account in last 90 days above 50,000

iv. Data account

- 30 days: The people who had recharged the data account between 7,500 and 17500 have shown defaulting
- 90 days: Number of times data account got recharged in last 90 days is 4000 and it has no impact on the label

v. Loans taken:

- Number of loans taken by user in last 30 days is maximum Between 0 and 2

Number of loans taken	Frequency of rechar
1	0 to 24
2	0 to 30
3	0 to 29
4	0 to 29
5	0 to 20
6	0 to 20
7	0 to 14
8	0 to 9
9	0 to 6
10	0 to 6
11	0 to 5
12	0 to 5

It can be seen the higher the loan taken,The frequency of recharge is less

- People who have taken loans 2 times,3 times and 6 times are seen to recharge with maximum amounts
- The number of loans taken does not impact the Defaulting much
- The total amount of loans taken has been between 0 and 7
- The amount of loan taken and default rate has not much interdependence
- There are only two options: 5 & 10 Rs., for which the user needs to pay back 6 & 12 Rs. respectively The highest loan amounts are upto 2,400
- Maximum amount of loans are when number of loan taken is 1
- Default rate is higher when max loan taken is above 17500
- Median of amounts of loan taken by the user in last 30 days is within 0.1

vi. **Mean payback time:**

- For 30days Mean payback varies between 1 and 9 but It does not impact the labels
- For 90days Mean payback is inside 7days mostly and does not impact the labels

2. Key Challenges in the Problem

The Problem with the datas of credit is that most of the customers pay back and only few default(due to which the banks are not going into Bank Run). Hence there comes class imbalance problem in the dataset. However the challenge is addressed by balancing the class before modelling.

3.Limitations of the problem

A problem in the dataset is that the data is for the urban sector as well, making the model inefficient for rural sector alone. The dataset should consider also a column which shows the income group which they belong to making the model more efficient in addressing the social sector problem of insuffient funds for the Poor to recharge there telecom numbers

Another challenge is the prevalence of only a small time frame in the data, making it insufficient for a large time frame. The more number of calls could be because of a festival or college holidays. Hence more larger time frame needs to be considered.