

MICRO CREDIT DEFAULTER PROJECT

Submitted by:

Harshitha K.S

**ACKNOWLEDGMENT**

1.“**Predicting the Microfinance Credit Default: A study of Nsoatreman Rural Bank, Ghana” Journal of Advances in Mathematics and Computer Science” Journal by the author Ernest Yeboh Boateng, Franchis T Oduro.**

This paper examined the factors predicting the micro-finance credit default in Northern Ghana. Data was collected from the 409 micro-credit beneficiaries of Nsoatreman Rural Bank who were located In urban , semi-urban and rural areas. L ogistic Regression was used to analyse the data. It was evident from study factors such as educational level, number of dependents, type of loan, adequacy of loan facility, duration for repayment of loan, number of years in business, cost of capital, penod within a year the loan was advanced to client had a significant effect on credit default. To enhance the efficient management of micro-credit,it is encouraged that the Micro-finance institutions(MFIs) adopt the group loan policy as the main mode of advancing micro loans to clients rather than individual loan policy. Again the MFI’s should team up with the Ministry of education through the Non-Formal Education Division to Organize function literacy workshops for micro-credit beneficiaries so as to equip them with the required knowledge to do successful business. Also MFI’s should consider giving loans with repayment duration of atleast12 months and atmost 24 months.

2**. “Predicting the micro-credit default among micro-borrowers in Ghana” by author Kwame Simpe ofori, Eli Fianu, Kayode Omoregie, Nii Afotey Odai, Oduro Gyimah**.

In this paper, Micro-finance plays an important role in economic development in many developing countries. However many of these micro-finance institutions are faced with the problem of default because non-formal nature of business and individual they lend money too. This study seeks to find the determinants of credit default in micro-finance institutions. with data on 2631 successful loan applicants from micro-finance institutions with branches all over country we proposed Binary Logistic regression model to predict the probability of default. They found that following variables are significant in determining default. Age, Gender, Martial Status, Income level, Residential status, number of dependents, loan Amount, tenure. We also found to be more among the younger generation and in males. They however,found loan purpose to be significant in determining credit default. Micro-finance Institutions could use this model to screen prospective loan applicants in order to reduce the level of default.

3.“**An Empirical Analysis of the Loan Default Rate of Micro-Finance Institutions” by author Anthony Siaw, Evans Brako Ntiamoah, Emmanuel Oteng**

In this paper, Microfinance institutions have been extending loans to different defict units in Ghana and this study aimed at addressing the following issues. Identifying the causes of loan default and the processes involved in granting loan by Micro-Finance institutions in Ghana. The convenient and Purposive sampling techniques were employed to select respondants to provide answers to questionnaires. The population of survey constituted the management and non-management staff and customers of some selected micro-finance institutions in Ghana. Hypothesis of the study will be analysed using correlation and regression. Results of study show that there are high positive correlation between the constructs of loan default causes and how loans are granted.

4.“**Prediction of Loan Defaulters Micro-finance using Network Data David Murphy” by author David Murphy**

In these paper, they are faced with the growing competition in micro-financing market and has higher operational risk, it is ever more important for an MFI to be able to leverage less conventional customer data to improve the efficiency of their lending models. Most MFIs are active in developing countries where financial history is generally non –existent on their user base which increases the difficulty in assessing the credit worthiness of individuals. Instead, an alternative source of data such as mobile phone call and SMS logs can be utilized to assist with this problem. In this study, call and SMS logs from the borrowers of a MFI operating in the Kenyan market place are featurized and used to train various classification models. The results shows how such data is valuable commodity in predicting the default class, particularly when relationship tie-strength features are introduced. The influence of an existing borrower’s loan outcome on a new loan applicant withi their social Network is also modelled using the spreading activation method as an alternative approach to traditional classification, but results indicate that they are not effective.

5. “**Assessing Institutional Characterstics on Micro-Credit default in Kenya: a Comparative Analysis of Microfinance and financial intermediaries” by the author Muturi Phyllis Muthoni**

In these paper, a major common concern on micro-credit repayment remains a major obstacle to Micro Financial Instiutions(MFIs) and financial Intermediaries(FIs) in Kenya. The health of MFI sector in sub sahara Africa is acause of concern due to the increased portfolio at risk(PAR). This region records the highest risk globally with its PAR 30 greater than 5%. This study sought to investigate causes of loan default with MFIs and Financial Intermediaries(FIs) in Kenya and specifically to evaluate the influence of institutional characterstics on loan default in MFIs and FIs . this study was based on pecking order theory and Grameen bank model and on positivism philosophy which adopts a quantitative approach to investigate the phenomena and uses descriptive survey design to investigate the phenomena and uses descriptive survey design to investigate the phenomena and uses descriptive survey design to investigate the populations by selecting samples to analyse and dicover occurrences. A target population of 294 MFIs institution and 76 Financial Institutions was used. A multistage sampling procedure was used and a sample of 106 MFIs and 40 FIs selected.Random sampling was used to select the respondants since each participant had an equal opportunity to be selected. Primary data was collected by use of questionanaire with closed and open responses presented on a five likert scale, making it easy for respondent to fill. Data was analysed by quantitative methods

by the use of SPSS; version 21. Descriptive statistics and Inferential statistics and some tests were carried out a 95 percent confidential such as F-test, t-test to examine parameters that were significant with the p-value less than 5 Percent being considered significant. Data was presented in form of frequency tables, bar charts and pie-charts for easy interpretation of results. A multiple regression model and pearson correlation were used to establish relationships among the variables. The findings of the study indicated that institutional characterstics wre significant among MFIs and FIs but with some differences in the parameters measured. The findings of the study will be of the significance to policy makers, MFIs, FIs, small business, universities and the general public as a source of knowledge for future reference

**INTRODUCTION**

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.

Many microfinance institutions (MFI), experts and donorsare supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be6(in Indonesian Rupiah), while, for the loan amount of 10(in Indonesian Rupiah), the payback amount should be 12(in Indonesian Rupiah).

The sample data is provided to us from our client database. . In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

**Analytical Problem Framing**

Mathematical/ Analytical Modeling of the Problem.

* When we analyse the d-types we get the following output

Unnamed: 0 int64

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 object

dtype: object

* There are no null values present in the data

Unnamed: 0 0

label 0

msisdn 0

aon 0

daily\_decr30 0

daily\_decr90 0

rental30 0

rental90 0

last\_rech\_date\_ma 0

last\_rech\_date\_da 0

last\_rech\_amt\_ma 0

cnt\_ma\_rech30 0

fr\_ma\_rech30 0

sumamnt\_ma\_rech30 0

medianamnt\_ma\_rech30 0

medianmarechprebal30 0

cnt\_ma\_rech90 0

fr\_ma\_rech90 0

sumamnt\_ma\_rech90 0

medianamnt\_ma\_rech90 0

medianmarechprebal90 0

cnt\_da\_rech30 0

fr\_da\_rech30 0

cnt\_da\_rech90 0

fr\_da\_rech90 0

cnt\_loans30 0

amnt\_loans30 0

maxamnt\_loans30 0

medianamnt\_loans30 0

cnt\_loans90 0

amnt\_loans90 0

maxamnt\_loans90 0

medianamnt\_loans90 0

payback30 0

payback90 0

pcircle 0

pdate 0

dtype: int64

* When we check the information we get output as follows:-

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 209593 entries, 0 to 209592

Data columns (total 37 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 209593 non-null int64

1 label 209593 non-null int64

2 msisdn 209593 non-null object

3 aon 209593 non-null float64

4 daily\_decr30 209593 non-null float64

5 daily\_decr90 209593 non-null float64

6 rental30 209593 non-null float64

7 rental90 209593 non-null float64

8 last\_rech\_date\_ma 209593 non-null float64

9 last\_rech\_date\_da 209593 non-null float64

10 last\_rech\_amt\_ma 209593 non-null int64

11 cnt\_ma\_rech30 209593 non-null int64

12 fr\_ma\_rech30 209593 non-null float64

13 sumamnt\_ma\_rech30 209593 non-null float64

14 medianamnt\_ma\_rech30 209593 non-null float64

15 medianmarechprebal30 209593 non-null float64

16 cnt\_ma\_rech90 209593 non-null int64

17 fr\_ma\_rech90 209593 non-null int64

18 sumamnt\_ma\_rech90 209593 non-null int64

19 medianamnt\_ma\_rech90 209593 non-null float64

20 medianmarechprebal90 209593 non-null float64

21 cnt\_da\_rech30 209593 non-null float64

22 fr\_da\_rech30 209593 non-null float64

23 cnt\_da\_rech90 209593 non-null int64

24 fr\_da\_rech90 209593 non-null int64

25 cnt\_loans30 209593 non-null int64

26 amnt\_loans30 209593 non-null int64

27 maxamnt\_loans30 209593 non-null float64

28 medianamnt\_loans30 209593 non-null float64

29 cnt\_loans90 209593 non-null float64

30 amnt\_loans90 209593 non-null int64

31 maxamnt\_loans90 209593 non-null int64

32 medianamnt\_loans90 209593 non-null float64

33 payback30 209593 non-null float64

34 payback90 209593 non-null float64

35 pcircle 209593 non-null object

36 pdate 209593 non-null object

dtypes: float64(21), int64(13), object(3)

memory usage: 59.2+ MB

* Data Sources and their formats

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

rental90:-Average main account balance over last 90 days

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

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

sumamnt\_ma\_rech90:-Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)

medianamnt\_ma\_rech90:-Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)

medianmarechprebal90:-Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah)

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

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

* Data Pre-processing :-
* We have dropped three object that is msisdn, pcircle, pdate. Here pcircle is categorical variable we have to convert to float otherwise. I had done the following function

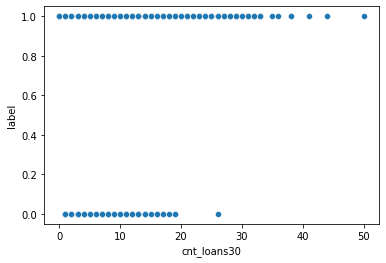
df = pd.get\_dummies(df, columns=["pcircle"], drop\_first=True)

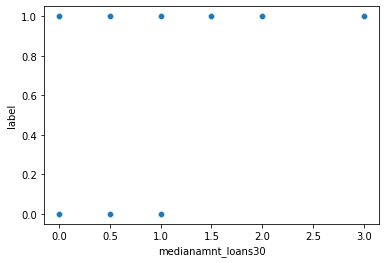
df = pd.get\_dummies(df, columns=["pdate"], drop\_first=True)

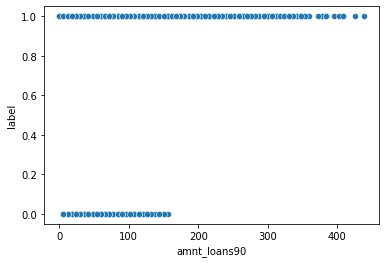
df.drop(['msisdn'], axis=1)

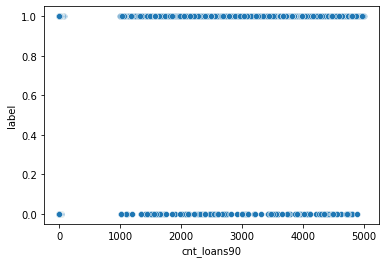
* Data Inputs- Logic- Output Relationships:-

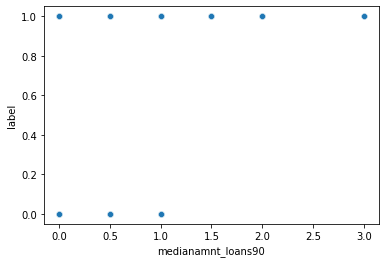
Here the output is taken as label and input are taken as following variable such as amount\_loans30, amount\_loans90, cmtloans30, cmtloans90,mdnamtloans30, mdnamtloans90, maxamtloans30, maxamntloans90

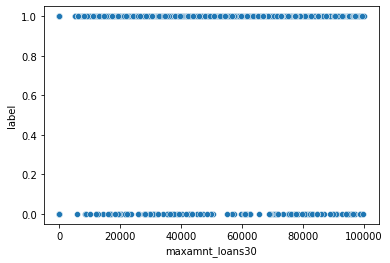
****

****

****

****







Assumptions:-

1. Positive examples :- the flag indicating the user paid back the credit amount within 5 days of issuing the loan(1)- success

2. Negative examples:- the flag indicating the user have not paid credit amount within 5 days of issuing the loan(0)- Failure

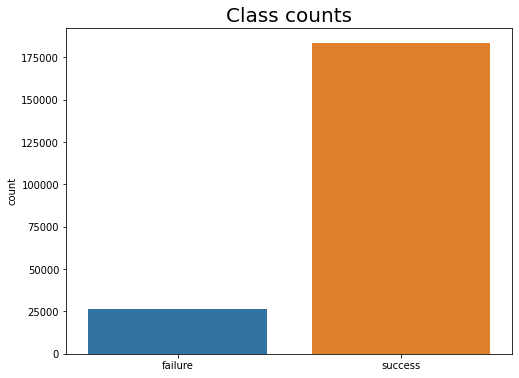
Positive examples = 183431

Negative examples = 26162

Proportion of positive to negative examples = 70.11%

Observation:-

* there are 1,83,431 user paid back the credit amount within a 5 days of issuing a loan
* there are 26,162 user not paid back the credit amount within a 5 days of issuing a loan



* Software Requirements and Tools Used:-

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

**Getting Up and Running With Jupyter Notebook**

There are many distributions of the Python language. This will focus on just two of them for the purposes of installing Jupyter Notebook. The most popular is [CPython](https://realpython.com/cpython-source-code-guide/), which is the reference version of Python that you can get from their [website](https://www.python.org/). It is also assumed that you are using **Python 3**.

### Installation

If so, then you can use a handy tool that comes with Python called **pip** to install Jupyter Notebook like this:

The next most popular distribution of Python is [Anaconda](https://www.anaconda.com/). Anaconda has its own installer tool called **conda** that you could use for installing a third-party package. However, Anaconda comes with many scientific libraries preinstalled, including the Jupyter Notebook, so you don’t actually need to do anything other than install Anaconda itself.

### Starting the Jupyter Notebook Server

Now that you have Jupyter installed, let’s learn how to use it. To get started, all you need to do is open up your terminal application and go to a folder of your choice. I recommend using something like your Documents folder to start out with and create a subfolder there called Notebooks or something else that is easy to remember.

This will start up Jupyter and your default browser should start (or open a new tab) to the following URL: <http://localhost:8888/tree>

## Creating a Notebook

Now that you know how to start a Notebook server, you should probably learn how to create an actual Notebook document.

All you need to do is click on the New button (upper right), and it will open up a list of choices. On my machine, I happen to have Python 3 installed, so I can create a Notebook that uses these. For simplicity’s sake, let’s choose Python 3.

### Naming

You will notice that at the top of the page is the word Untitled*.* This is the title for the page and the name of your Notebook. Since that isn’t a very descriptive name, change it!

Just move your mouse over the word Untitled and click on the text. You should now see an in-browser dialog titledRename Notebook. rename this one to project name

### Running Cells

A Notebook’s cell defaults to using code whenever you first create one, and that cell uses the kernel that you chose when you started your Notebook.

In this case, you started yours with Python 3 as your kernel, so that means you can write Python code in your code cells. Since your initial Notebook has only one empty cell in it, the Notebook can’t really do anything.

Thus, to verify that everything is working as it should, you can add some Python code to the cell and try running its contents

**Model/s Development and Evaluation**

* Observation :- 1.There are no null values present in the data 2. there are 2 data types present int and float

1. there are 1,83,431 user paid back the credit amount within a 5 days of issuing a loan

2.there are 26,162 user not paid back the credit amount within a 5 days of issuing a loan.

* Testing of Identified Approaches (Algorithms)

Algorithm I used here Random forest classifier and Decision tree

Random forest classifier

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

It works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.

**Advantages:**

* Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
* It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
* The algorithm can be used in both classification and regression problems.
* Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
* You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

**Disadvantages:**

* Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming.
* The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

## Finding important features

Random forests also offers a good feature selection indicator. Scikit-learn provides an extra variable with the model, which shows the relative importance or contribution of each feature in the prediction. It automatically computes the relevance score of each feature in the training phase. Then it scales the relevance down so that the sum of all scores is 1.

This score will help you choose the most important features and drop the least important ones for model building.

Random forest uses gini importance or mean decrease in impurity (MDI) to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables.

**Decision Tree :**Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

**Construction of Decision Tree :**  
A tree can be “learned”by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning*.* The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

**Decision Tree Representation :**  
Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree,testing the attribute specified by this node,then moving down the tree branch corresponding to the value of the attribute as shown in the above figure.This process is then repeated for the subtree rooted at the new node.

**Strengths and Weakness of Decision Tree approach**  
The strengths of decision tree methods are:

* Decision trees are able to generate understandable rules.
* Decision trees perform classification without requiring much computation.
* Decision trees are able to handle both continuous and categorical variables.
* Decision trees provide a clear indication of which fields are most important for prediction or classification.

The weaknesses of decision tree methods :

* Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
* Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
* Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

.

**CONCLUSION**

Microfinance has been globally accepted as the preferred medium to reach out to the rural and productive poor with banking services which includes micro credit to help alleviate poverty Micro credit default has been identified to be one of the major drawbacks of this laudable initiative as it depletes these revolving funds and reduces investors’ confidence. Therefore, it is important to understand the factors that influence a loan beneficiary to default so that appropriate counter measures can be developed to prevent and reduce the incidents of default.

This analysis shows that label, aon, daily\_dec\_30, daily\_dec\_90, rental30, rental90, medianamnt-ma-rech90,medianamnt-rech -30, cnt\_da\_rech30, cnt\_da\_rech\_90, maxamnt\_loans30,maxamnt\_loans90,medianamnt\_loans30, medianamnt\_loans90 were important determinants of default.