NAME OF THE PROJECT

MICRO-CREDIT DEFAULTER MODEL

Submitted by:

ASHISH KUMAR SAMAL

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

I wish to express my sincere gratitude to DataTrained Education, Khusboo Garg,SME for providing me an opportunity to do my internship and project work in “FLIP ROBO”.

It gives me immense pleasure in presenting this project report on “Micro Credit Defaulter Model”. It has been my privilege to have a team of project guide who have assisted me from the commencement of this project. The success of this project is a result of sheer hard work, and determination put in by me with the help of You Tube videos, references taken from Github.com,Kaggle.com, skikit-learn.org.

I hereby take this opportunity to add a special note of thanks for Miss, Khusboo Garg, who undertook to act as my mentor despite his many other professional commitments. His wisdom, knowledge and commitment to the highest standards inspired and motivated me. Without his insight, support this project wouldn’t have reached fruitfulness.

The project is dedicated to all those people of Fliprobo, Datatrained who helped me while doing this project.

INTRODUCTION

- Business Problem Framing

The company is 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. The company know the importance of communication so they have also focused on providing product and services to low-income families. To do this, they have a collaboration with MFI to provide micro credit on mobile balances to be paid back in 5 days. The customer is considered as defaulter if he fails to pay the sum of money within the stipulated period of time. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah). Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. In order to make sure this undeserved population has a positive loan experience; company makes use of a variety of alternative data include transactional information--to predict their clients' repayment abilities.

- Conceptual Background of the Domain Problem

We can see there is huge population with no financial record is there to access all that remote areas to help them we need to come with more ground projects to help such population.In this whole scenario we encounter that there is a must use of Artificial Intelligence because as we can see that there is high variation of defaulters and non-defaulters.By the use of AI we can analyse patterns of peoples who are taking micro credit by the help of their history of recharge, date of recharge, daily usage, there payment history of loans etc. After putting all these constrains in AI and modelling with different models we will come to a point where we can suggest a point of view what more improvement is needed or who all are the ones which will be defaulter and should be stop from credit facility.

- ## Review of Literature

  This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done. Micro credit has been the term which refers to the formal and informal arrangements of providing financial services to the poor for the upliftment. It is microfinance which over the past decades has changed the perception of poor from non-bankable to bankable and recommending various methodologies to provide financial services. Also, Microfinance over the years has not only tried to alleviate poverty across the world but also shown glimpses of sustaining themselves from profit earned in the process.

  Asian Development Bank (ADB) defines microfinance as “the provision of a broad range of financial services as deposits, loans, money transfers, insurance to small enterprise and households, and providing credit in telecommunication services.CGAP (2003) defines microfinance as “a credit methodology that employs effective collateral substitutes to deliver and recover short-term working capital loans to micro entrepreneurs.”

  I went and felt that Microfinance Institution purpose is to fulfil the financial needs of the poor either through informal or flexible approach. There is no single model that fits in all the circumstances. Number of microfinance models emerged in different countries/states according to the suitability to their local conditions. Broadly, the microfinance delivery methods can be classified into six parts:

  - Grameen Bank Model
  - Joint Liability Group Model

- Individual Lending Model
- The Self-Help Group Model
- Village Banking Model
- Credit Unions and Cooperatives

We got to know after studying about micro finance that this facility is to curb down and help poor who are short form money. So, by the help of this they provide small loans to them and a deadline of paying back. But sometimes these low sound people become defaulter they cannot repay the loan taken so that is why these small amounts are given in loan because if they fail to repay then the micro finance company don't face big losses. As well as for more safety they are using Artificial Intelligence to speculate and understand the patterns where the person is failing to repay the credit taken. By varieties of model's data scientist and researcher are putting their expertise knowledge for minimizing and overcoming from this problem.

An attempt has been made in this article to review the available literature in the area of microfinance.

- Motivation for the Problem Undertaken

  The initiative of the company to provide Micro credit is very noble to help the low-income group of people but there are certain people who take advantage of this noble idea and don't bother to repay the money and become defaulter. So, it is necessary to stop this type of practice. Sometime people with good intension remain deprived of getting loan from the financial institution due to some dishonest people. Hence Machine Learning can be used to predict the defaulter and non-defaulter by using different parameters.

  Problem facing issue we saw in this whole project that people are not well sound in terms of money they require money for their basic small needs. Micro finance is a hand from which they can take as a help. In today world everyone needs a basic requirement to carry out daily life. A middle-class sector is still not targeted by micro finance sector.I also believe that a 5-day time given to repay the credit taken is less for a low-income family it may get difficult to repay. Micro finance should come up with more new flexible strategies to uplift this low- and middle-class income groups.Motivation behind this whole project is that we will analyse the whole low-income Indonesia population dataset who have

taken a credit facility for telecom. We will train our model by using AI methods to check the patterns when a person gets defaulter and who all are the persons who pay on time and are non-defaulters. I hope this model will help in analysing the patterns for micro credit institutions while making their policies for giving credit in telecom sector.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

The micro credit defaulter dataset which is from Indonesia consists of 37 columns and 209593 rows. In our dataset we have datatypes of various type. There are 21 columns of float64, 13 columns of int64 & 3 columns of object. There are no null values present in our dataset. It is a huge imbalanced data and for moving further we have to do data cleansing before moving to next step.There is high standard deviation from the mean value. The difference between the third quantile and maximum value was huge in many cases which was quite abnormal and hence I decided to replace them with Q3+1.5(IQR) if it is more than Q3+1.5(IQR). In some places the minimum values were negative which also seem to be abnormal in that case. Hence, it was replaced by Q1-1.5(IQR) if it is below the minimum value. It was found in some variables that, the maximum value was abnormally high which was replaced by a normal high number of that variable. The visualization also helped to identify the skewness present in the data. Those skewness were

also corrected using square root transformation. At last, after data pre-processing we come the model building section, were I used Logistic Regression, K-Nearest Neighbour, XG-Boost , Decision Tree and Random Forest Classifier .

- ## Data Sources and their formats

This data is been provided by a Telecom company. 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. The company shared around 2 lakh data of their customer with different transaction behaviour to understand and to predict their future behaviour. The data is been provided in CSV format with 37 different variables in different columns and 209593 rows.

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 users
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 Preprocessing Done

  Next moving to column name msisdn, pcircle and pdate which are of object type and cannot give any help in best performance of model. Pcircle means a code given to certain areas which are not showing any relevance with population mobile credit taking.Msisdn,pcircle was removed along columns and from pdate the p month and pday was extracted using datetime functions.

Most of the data in the dataset was full of outliers. Those outliers were corrected by replacing them with Q3+1.5(IQR) if it is more than Q3+1.5(IQR). The data was also skewed. Some of them were negatively whereas some are positively skewed. All the skewed data was corrected using square root transformation where ever applicable. In some places the minimum values were negative which also seem to be abnormal in that case. Hence, it was replaced by Q1-1.5(IQR) if it is below the minimum value.Columns having negative values were converted to absolute values. The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records.So,I used SMOTE technique to handle the imbalance in the dataset.And the data was scaled using standard scaler before resampling.

- Data Inputs- Logic- Output Relationships

The input data provided, helps to understand the behaviour of the customer, their various transaction records, their frequency of transaction during a period of time etc, all these helps to predict the customer’s intension toward the repayment of loan.

- Hardware and Software Requirements and Tools Used

Data Science task should be done with sophisticated machine with high end machine configuration. But unfortunately, the machine which I’m currently using is powered by intel core i5 processor with 8GB of RAM. With this above-mentioned configuration, I managed to work with the data set in Jupyter

Notebook which help us to write Python codes. The library used for the assignment are Numpy, Pandas, Matplotlib, Seaborn, Scikit learn and other modelling libraries for Logistic Regression, K-Nearest Neighbour, XG-Boost , Decision Tree and Random Forest Classifiers.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  The dataset is imbalanced. Label 1 has 87.5% of data whereas label 0 has approximately 12.5%. As I went through the dataset, I found lot of outliers and skewness are present in the dataset. The outliers were corrected by replacing them with Q3+1.5(IQR) if it is more than Q3+1.5(IQR). In some places the minimum values were negative which also seem to be abnormal in that case. Hence, it was replaced by Q1-1.5(IQR) if it is below the minimum value. The skewness was also reduced using square root transformation wherever applicable. There were certain columns which had least importance with our target variable, hence those were dropped.

  To handle the imbalance data we use SMOTE technique and scaled it using standard scaler. After data cleaning and data transformation, data visualization was done to represent data graphically. At last, the most important part was to build model for the data set.

- Testing of Identified Approaches (Algorithms)

  Listing down all the algorithms used for the training and testing.

  a. Logistic Regression
  b. K-Nearest Neighbour
  c. Random Forest Classifier
  d. XG-Boost
  e. Decision Tree Classifier

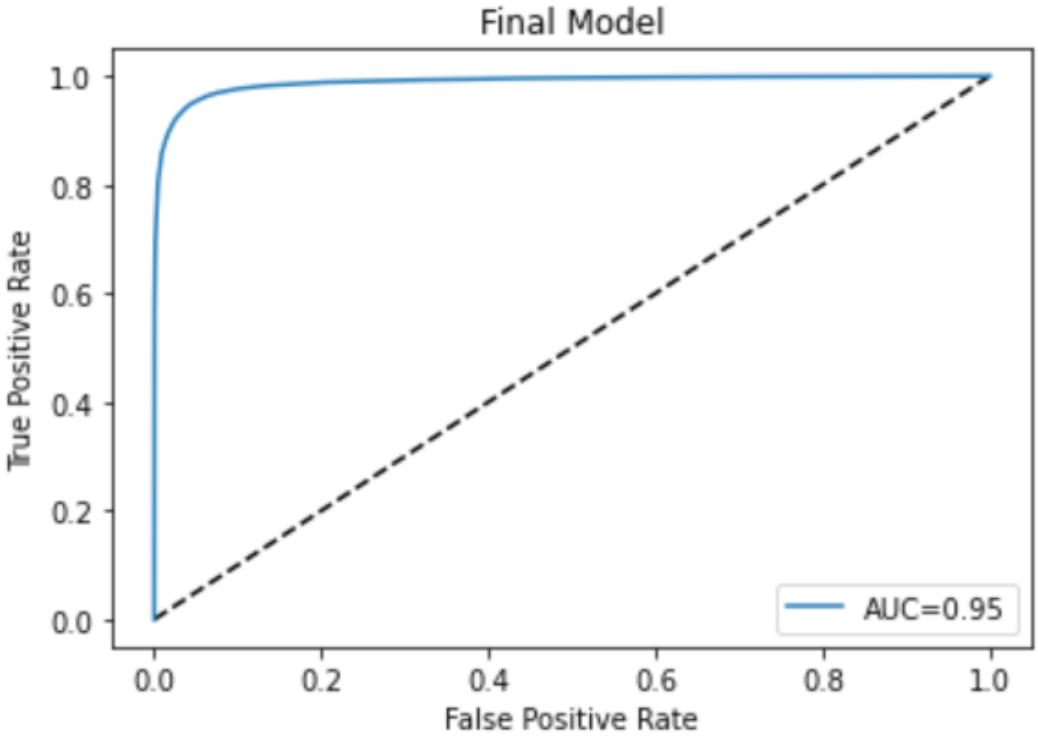
- Run and Evaluate selected models

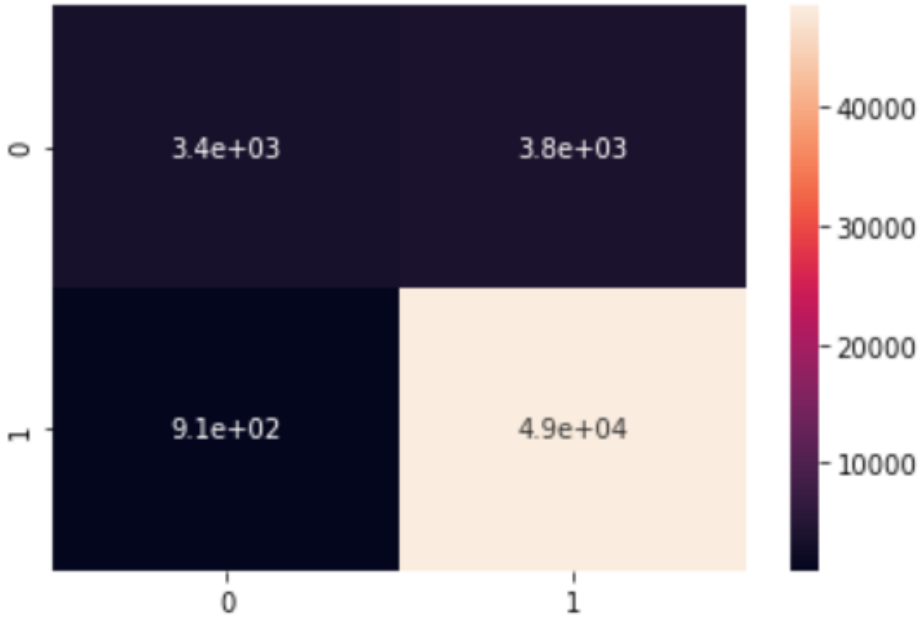
```
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score,accuracy_score,confusion_matrix,classification_report
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import cross_val_score
from xgboost import XGBClassifier
```

- Key Metrics for success in solving problem under consideration

```
final_mod=RandomForestClassifier(n_estimators=300,max_features= 'auto',criterion='gini',class_weight='balanced')
final_mod.fit(x_train,y_train)
pred=final_mod.predict(x_test)
fin_CM = confusion_matrix(y_test,pred)
fin_CR = classification_report(y_test,pred)
fin_acc = accuracy_score(y_test,pred)

print('confusion metrics :', '\n', fin_CM)
print('classification report ', '\n', fin_CR)
print('accuracy score: ', '\n' , fin_acc)
```

```
confusion metrics :
 [[47183  2218]
 [ 2512 47140]]
classification report
              precision    recall  f1-score   support

         0.0       0.95      0.96      0.95     49401
         1.0       0.96      0.95      0.95     49652

    accuracy                           0.95     99053
   macro avg       0.95      0.95      0.95     99053
weighted avg       0.95      0.95      0.95     99053

accuracy score:
 0.9522477865385198
```

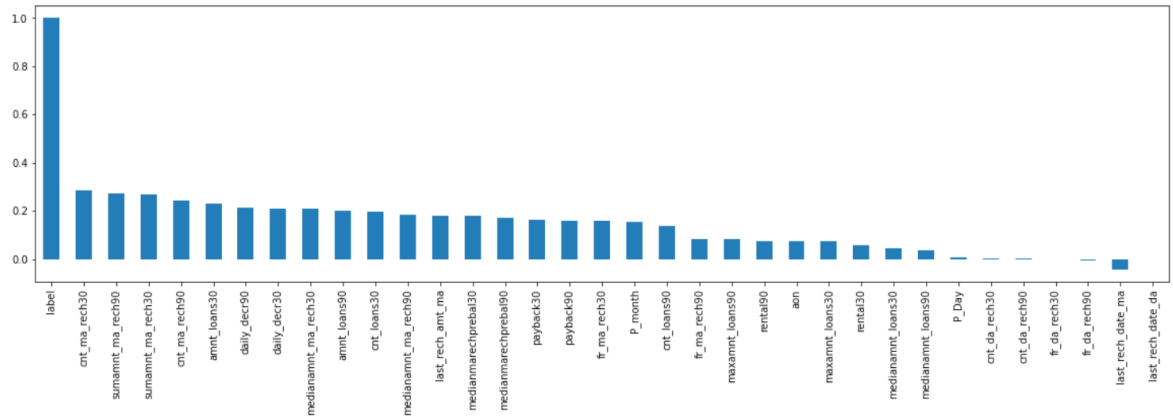




- Visualizations

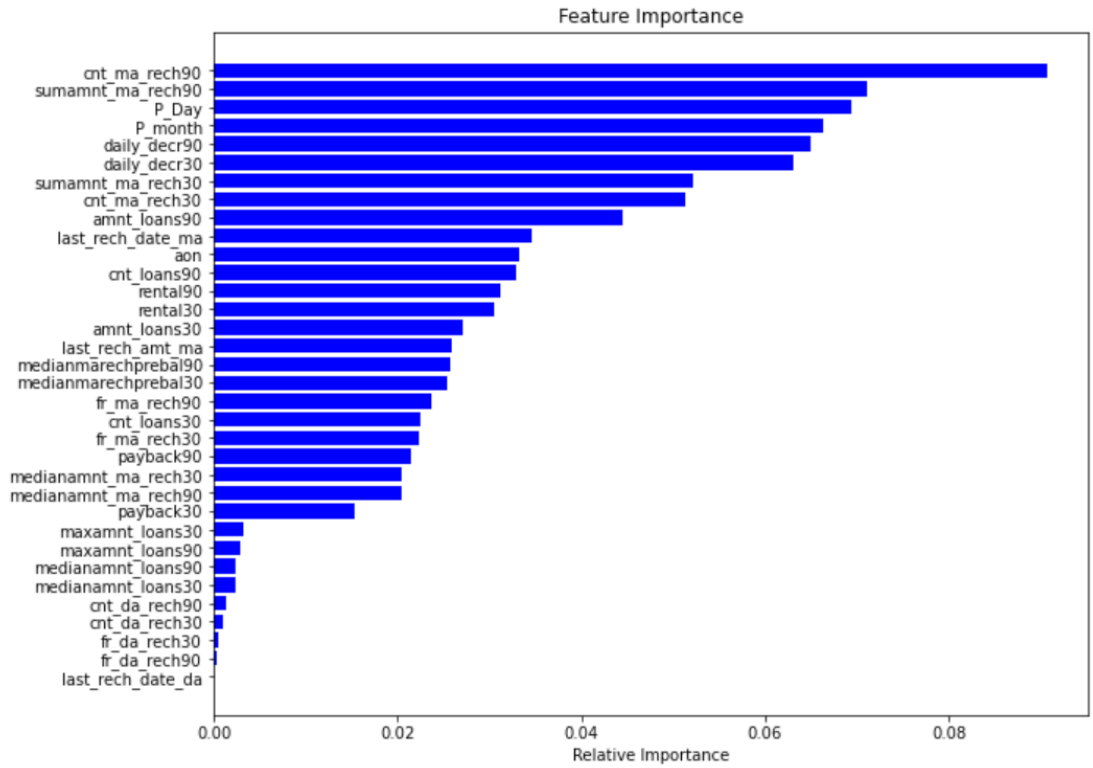
Percentage of both label outputs in our dataset

We can check how much each of the features are important for the label prediction.



The correlation heatmap :

- Interpretation of the Results



CONCLUSION

- Key Findings and Conclusions of the Study

In past, micro credit and their institution were less as it could not cover the whole population who were facing issues in credit taking. There was no proper financial map planned to cover these low-income groups. Now various financial institutions, banks, NGO's are coming up with great

credit facilities in many sectors such as agriculture, small-scale business, telecom service etc. In the given dataset we notice that mostly, the customers have the intension of repaying. There are certain cases, when the customers have no intension of repayment but the number of such customers are few. With the model built, we can certainly determine customers having intension of repayment or not.

Role of Government

MFI's should be controlled and monitored by government and some laws against them.

- The ministry must enforce strong policies, strategies, laws and regulations that enhances introduction of enough microfinance institutions to influence companions in financial sectors in order to lower interest rate and other cost of borrowing.

- The ministry should take up regular monitoring and evaluation in MFIs on credit compliance and loan disbursement.

-The ministry must reinforce effective loan application and processing procedures to Influence timely loans disbursement for instance from the current two-week period to a minimum of three days so that SMEs can access such facilities that require capital to accomplish.

Recommendations to MFI's

MFI's should ensure that they promote timely disbursement of loans. Timely access to capital is very crucial to the growth of sme's because the challenge of capital underlies most of the challenges rural SMEs face.This micro credit facility is very good for the upliftment of the society and give the needy ones an opportunity to stand and carry on their life with the help of such small loan provided by them. It is still in growing phase there must be more capital and respected sectors for recording and providing these facilities to humongous low-income population.

- Learning Outcomes of the Study in respect of Data Science

The dataset was full of outliers, skewness and unbalanced data which was the biggest challenge to overcome. Hence data cleaning was very important to get proper prediction.Feature scaling was done by help of standard scaler.And the imbalance in the dataset was handled using SMOTE

technique.For cross validation of the model K-Fold cross validation was used. I have used Logistic Regression, K-Nearest Neighbour,XG-Boost, Decision Tree and Random Forest Classifier. Among the five algorithms Random Forest Classifier gave the best outcome. We used RandomizedSearchCV for hyper parameter tuning of the best model as it is fast as compared to GridSearchCV.

- Limitations of this work and Scope for Future Work

The solution can be applied to the customer having a transaction history but the model may not perform well with customer having new profile and no transaction history. Nevertheless, the model will perform well with customer having transaction history and can predict whether a person will be a defaulter or non-defaulter. Hence, we can say that this statistical model will be helpful in future for the prediction of micro credit defaulter and non-defaulter customer.