FLIP ROBO

PROJECT REPORT

Submitted By

Aswini Arun Patil

NAME OF THE PROJECT

Micro Credit Defaulter

ACKNOWLEDGMENT:

- Primarily I would like to thank God to being able to complete this project with success. Then I would like to express my special thanks of gratitude to my SME,

- And I am thankful I am part of flib rob technology of employee, who given me the golden opportunity to do this wonderful project on the given topic which is also help me in doing a lot of research and I came to know about so many new things, I am really thankful to flip robo.

- Here all the data set was been provided to me and on that bases the EDA, data visualization, analysis has been taking place

- we have been taken much more references while surfing on the websites too

- Below in the report everything is mentioned from introduction of the project till the conclusion how it worked and finally how we got the best accuracy score for the given data set

DATE:28/08/2022

ASWINI A. PATIL
Data Science course
Institute: Data trained education
Internship: Flip Robo technology
@Bangalore

INTRODUCTION

- ## Business Problem Framing

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 donors are 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.

- ## Conceptual Background of the Domain Problem:

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 underserved population has a positive loan experience, company makes use of a variety of alternative data include transactional information-- to predict their clients' repayment abilities

- # Review of Literature

1. What is Microfinance?

"Microfinance" is often seen as financial services for poor and low-income clients. In practice, the term is often used more narrowly to refer to loans and other services from providers that identify themselves as "microfinance institutions" (MFIs). Microfinance can also be described as a setup of a number of different operators focusing on the financially under-served people with the aim of satisfying their need for poverty alleviation, social promotion, emancipation, and inclusion. Microfinance institutions reach and serve their target market in very innovative ways. Microfinance operations differ in principle, from the standard disciplines of general and entrepreneurial finance. This difference can be attributed to the fact that the size of the loans granted with microcredit is typically too small to finance growth-oriented business projects. Some unique features of microfinance as follows:

i. Delivery of very small loans to unsalaried workers.

ii. Little or no collateral requirements.

iii. Group lending and liability.

iv. Pre-loan savings requirement.

v. Gradually increasing loan sizes.

Implicit guarantee of ready access to future loans if present loans are repaid fully and promptly Microfinance is seen as a catalyst for poverty alleviation, delivered in innovative and sustainable ways to assist the underserved poor, especially in developing countries.

2. Default in Microfinance

Default in microfinance is the failure of a client to repay a loan. The default could be in terms of the amount to be paid or the timing of the payment.

- Motivation for the Problem Undertaken

Our main objective of doing this project is to build a model to predict whether the users are paying the loan within the due date or not. We are going to predict by using Machine Learning algorithms.

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:

Starting with the dataset, when I looked through the statistical description, we come to see that most of the data are unbalanced. 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 Log transformation and square root transformation. At last, after data pre-processing we come the model building section, were I used Logistic Regression, Gaussian NB and Random Forest Classifier.

- ## Data Sources and their formats

The data is been provided by one of our clients from telecom industry. They are a fixed wireless telecommunications network provider and 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 data is been given by Indonesian telecom company and they gave it to us in a CSV file, with data description file in excel format. They also had provided the problem statement by explaining what they need from us and also the required criteria to be satisfied.

Let's check the data now. Below I have attached the snapshot below to give an overview.

```
import numpy as np
import pandas as pd

import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import zscore
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import cross_val_score
import warnings
warnings.filterwarnings('ignore')
```

```
ds=pd.read_csv(r"C:\Users\Ashwini\Documents\Data file.csv")
```

```
ds
```

	Unnamed: 0	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	...	maxamnt_loans30	mec
0	1	0	21408I70789	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	...	6.0	
1	2	1	76462I70374	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	...	12.0	
2	3	1	17943I70372	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	...	6.0	
3	4	1	55773I70781	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	...	6.0	
4	5	1	03813I82730	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	...	6.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
209588	209589	1	22758I85348	404.0	151.872333	151.872333	1089.19	1089.19	1.0	0.0	...	6.0	
209589	209590	1	95583I84455	1075.0	36.936000	36.936000	1728.36	1728.36	4.0	0.0	...	6.0	
209590	209591	1	28556I85350	1013.0	11843.111670	11904.350000	5861.83	8893.20	3.0	0.0	...	12.0	
209591	209592	1	59712I82733	1732.0	12488.228330	12574.370000	411.83	984.58	2.0	38.0	...	12.0	
209592	209593	1	65061I85339	1581.0	4489.362000	4534.820000	483.92	631.20	13.0	0.0	...	12.0	

209593 rows × 37 columns

It shows that we have a total of 209593 rows and 37 columns present in our data frame. We have the label column that stores the defaulter and non-defaulter values marked with 0 and 1 making this a Classification problem.

- # Data Preprocessing Done:

Checked for missing values to confirm the information of no null values present provided in the problem statement.

```
df.isna().sum()#checking null values
```

```
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
```

- # Data Inputs- Logic- Output Relationships

The input data provided, helps to understand the behavior 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.

- State the set of assumptions (if any) related to the problem under consideration

  No as such assumption been done related to the circumstances.

- 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 i3 processor with 4GB 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. As I'm using low configuration machine so it took more time then usual to execute codes. The library used for the assignment are Numpy, Pandas, Matplotlib, Seaborn, Scikit learn.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  The data set contain more than 2 lakh data with no null values related to the customer. 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). The skewness was also reduced using Log transformation and square root transformation wherever applicable. There were certain columns which had least importance with our target variable,

hence those were dropped. 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)

  Following are the algorithms used for the training and testing: - a. Logistic Regression b. Gaussian NB c. Random Forest Classifier.

- ## Run and Evaluate selected models:

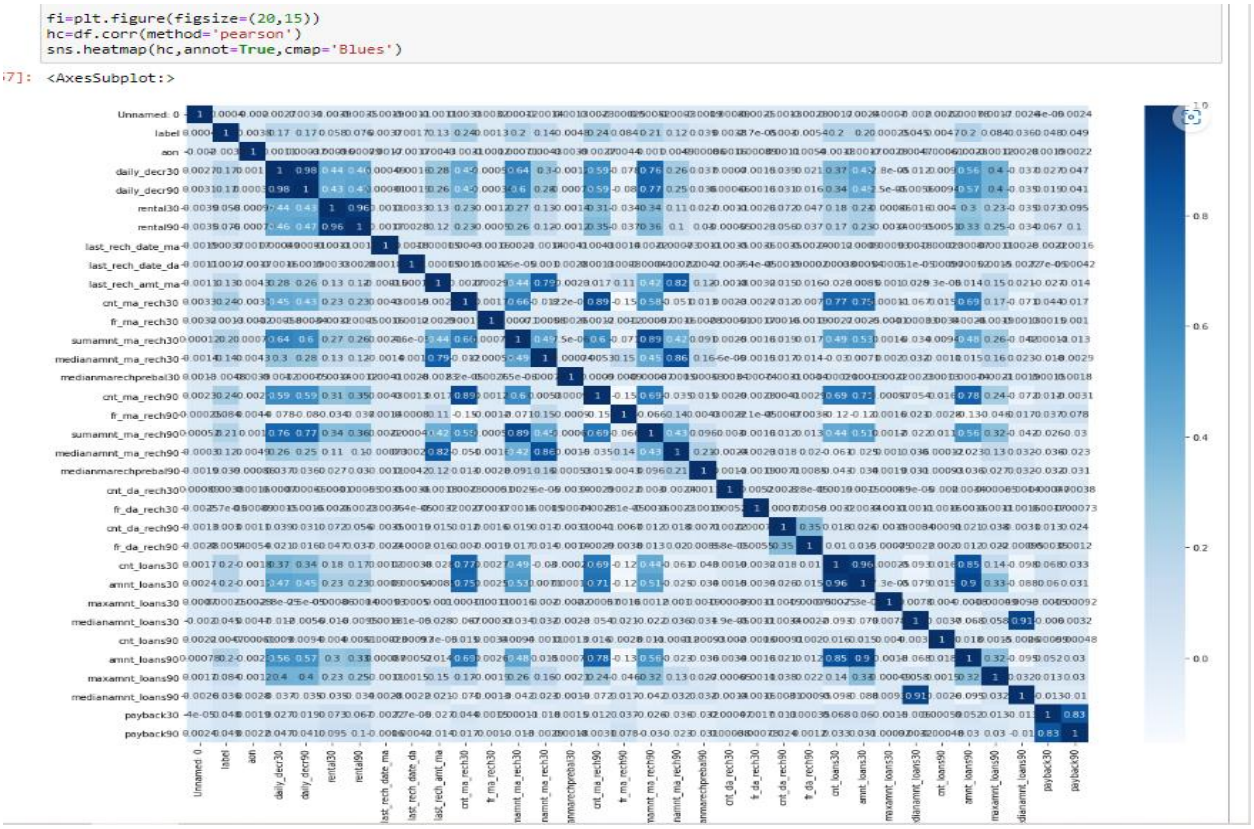
  I created a Classification Model function incorporating the evaluation metrics so that we can get the required data for all the models

- ## Key Metrics for success in solving problem under consideration

  As mentioned earlier, the dataset is unbalanced with 87.5% of label 1 and 12.5% of label 0, which made it clear that, we cannot blindly rely on accuracy score for the prediction as it can lead to biasness. Hence, I have used confusion matrix and AUC ROC curve to determine the accuracy of the model

- ## Visualizations

  Now, we will see the different plots done with this dataset in order to know the insight of the data present



CONCLUSION

- ## Key Findings and Conclusions of the Study

  From the final model MFI can find if a person will return money or not and should an MFI provide a load to that person or not judging from the various features taken into consideration.

- ## Learning Outcomes of the Study in respect of Data Science

  I built multiple classification models and did not rely on one single model for getting better accuracy and using cross validation comparison I ensured that the model does not fall into overfitting

and underfitting issues. I picked the best one and performed hyper parameter tuning on it to enhance the scores.

- ## Limitations of this work and Scope for Future Work

  Limitation is it will only work for this particular use case and will need to be modified if tried to be utilized on a different scenario but on a similar scale. Scope is that we can use it in companies to find whether we should provide loan to a person or not and we can also make prediction about a person buying an expensive service on the basis of their personal details that we have in this dataset like number of times data account got recharged in last 30 days and daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah) so even a marketing company can also use this.