

Micro Credit Loan

Submitted by:

Biranchi Narayan Nath

**ACKNOWLEDGMENT**

The dataset and to do instructions of the project was provided by my internship Company FlipRobo . During working with the project I have taken some help from my SME Nitin Mishra and from notes and previous projects to complete the project.

**INTRODUCTION**

* Business Problem Framing

A Telecom Industry collaborating with an MFI provides micro-credit on mobile balance to poor families living in remote areas with not much source of income. The loan repayment have a fixed time of 5 days to return back to the company. Here the problem is to create a model to Identify defaulter and non-defaulter users.

The data is similar to real world data like company churn prediction, bank loan approval.

* Conceptual Background of the Domain Problem

The binary classification problem can be better understand by studying the graphs too.

* Review of Literature

During working on the project I have researched about the users why they are not paying the loan amount that should be return back to the company and what are the causes behind it.

Those peoples having less cellular network in the day time, are failure to pay the loan because they are staying at more remote areas with very low income.

Peoples those are taking the loan more number of times and more amount from the company in the last 90 and 30 days are mostly non-defaulters.

All the 40673 new users taking loan from the company in the month of August 2016 have no defaulter.

* Motivation for the Problem Undertaken

To build a model to know the user behaviour and predict in terms of a probability for each loan transaction whether the customer will be paying back the loaned amount within 5 days of issuance of loan.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

During working in the project mathematics and statistics are used behind most of the technique and algorithms.

In the correlation chat pure mathematics is used to study the relationship among the features and their correlation value. It is very necessary to study the correlation among the features to know how the features are correlated with each other and also by knowing the correlation matrix we can drop some redundant features to save some computational cost and time.

In .describe() method we find the statistical summary of the dataset so that one can know the data, in the technique we find the count, mean, standard deviation, minimum, maximum and percentiles like 25th,50th & 75th of all the features. Knowledge of statistical summary is highly required before working with any project.

Data Sources and their formats

The data is provided by my internship Company FlipRobo the origin of the data is the client Telecom Company. The dataset have 37 features and 209593 rows/observations.In the 37 features the datatype of 1 feature is datetime, 21 features are float, 13 features are integer and 2 features are object. This is a binary classification problem. The features are

'Unnamed: 0', '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',

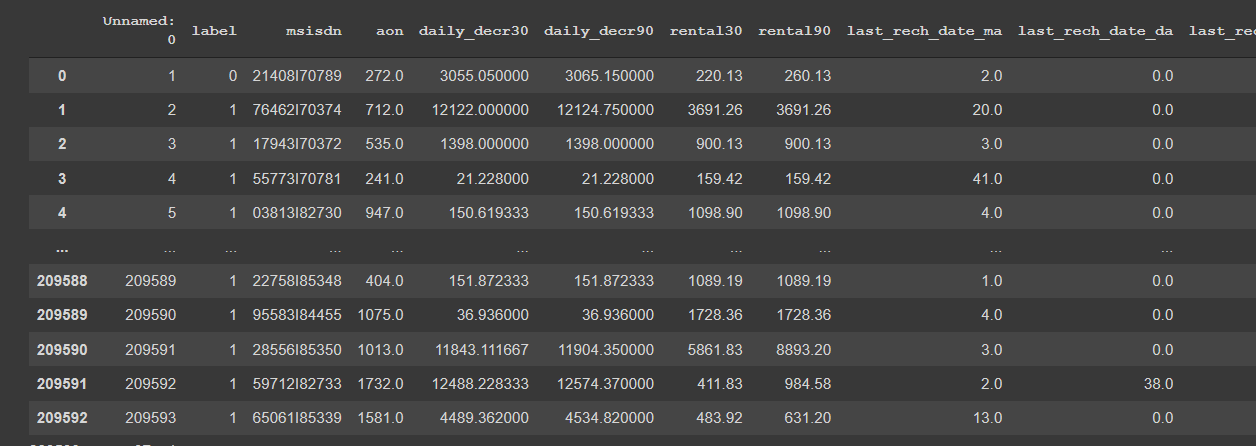
'medianamnt\_ma\_rech90', 'medianmarechprebal90', 'cnt\_da\_rech30',

'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'



* Data Preprocessing Done

Data is almost cleaned. So the features not required for model building is dropped. Firstly the ‘Unnamed: 0’ and mobile number feature are dropped. Then after checking the correlation matrix some highly correlated features are dropped.

Some assumptions concluded are:

1. All the users are in the same telecom circle.

2. It is the dataset for the months of June, July and August 2016.

In June the defaulter and non- defaulter are 13261 and 69893 respectively. In July 12901 are defaulter and 72864 are non-defaulter and in the month of August all the 40674 users are non-defaulter.

3. with increase in daily\_decr30, daily\_decr90 increases.

4. with increase in rental30 rental90 increases.

5. payback30 and payback 90 are directly propersional.

* Hardware and Software Requirements and Tools Used

**Hardware:** Laptop and Mouse.

**Software:**

i. Jupyter Notebook: All the libraries, packages and tools are called by it.

ii. Pandas: for reading csv file, using dataframe and dataframe related operations like checking shape, information and statistical summary. Saving the prediction in csv file.

iii.matplotlib & seaborn: both are used for plotting

iv.sklearn: It is used for splitting the dataset into train test sets, scaling training dataset, getting algorithms used for model building , metrics used for checking model capacity and saving the model for further use.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

For solving the problem the approaches are

i. Data Preprocessing:

In this step EDA is performed to know the dataset. So that to find clear vision about the data and perform as per required. In this step data is cleaned for model building.

ii. Dataset splitting: Dataset is splitted into train test split.

iii. Model Building: multiple models are created.

iv. Model Evaluating: Among all the models the best model is choose for Final model.

* Testing of Identified Approaches (Algorithms)

The algorithms are:

i.LogisticRegression

ii.KNeighborsClassifier

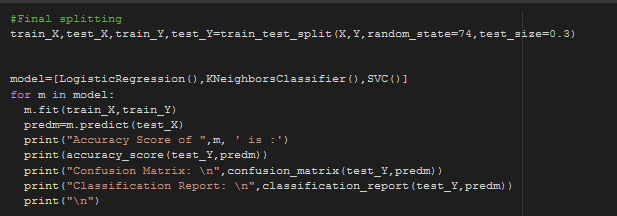
iii.SVC

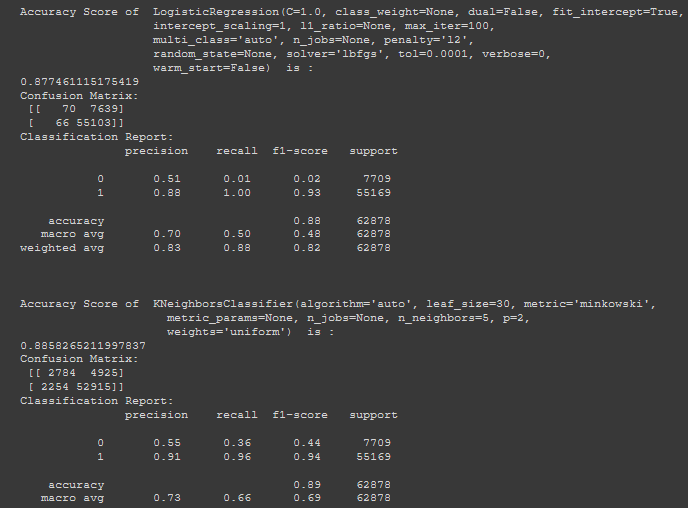
iv.RandomForestClassifier

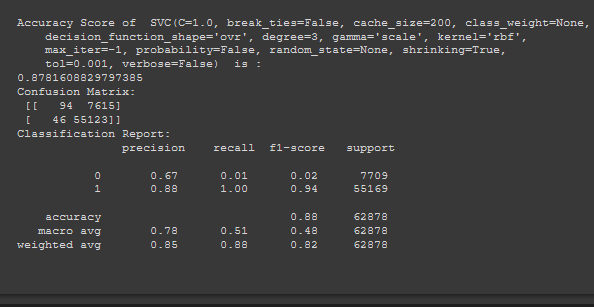
v.GradientBoostingClassifier

* Run and Evaluate selected models

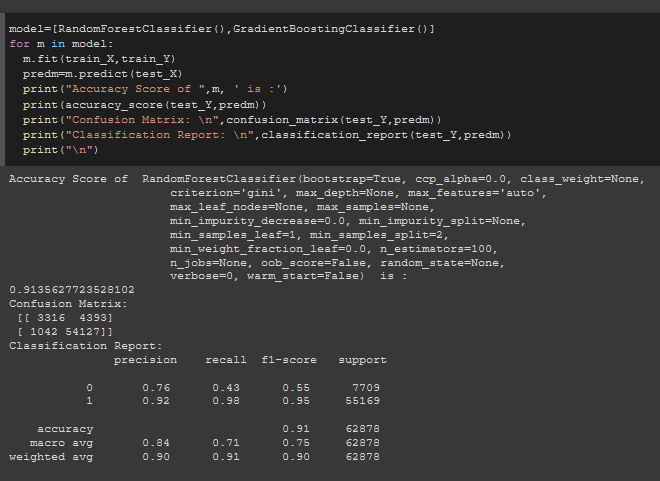
LogisticRegression, KNeighborsClassifier & SVC

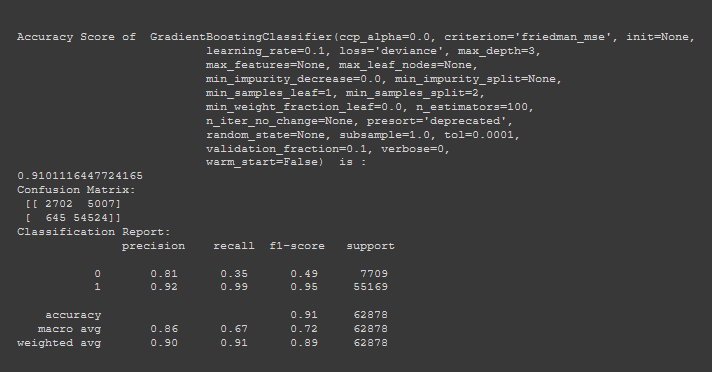






RandomForest,gradientBoosting





* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

i.accuracy\_score: As it is a classification problem accuracy score is required to check the accuracy score so it is used.

ii.confusion\_matrix: To see the number tp,fp,tn,fn.

iii)Classification\_report: To check all the metrics like precision,recall,f1-score it is used.

**CONCLUSION**

* Key Findings and Conclusions of the Study
* Those peoples having less cellular network in the day time, are failure to pay the loan because they are staying at more remote areas with very low income.
* Peoples those are taking the loan more number of times and more amount from the company in the last 90 and 30 days are mostly non-defaulters.