

**Micro Credit Loan Defaulter Model**

**Submitted by:**

***Saurabh Chandra***

**ACKNOWLEDGMENT**

In the present world of competition there is a race of existence in which those are having will to come forward succeed. Project is like a bridge between theoretical and practical working. With this willing I joined this particular project.

I have taken efforts in this project. However it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I would like to express my special thanks to my SME Miss Astha Mishra and Ankur Srivastava who gave me the golden opportunity to do this wonderful project on the topic Micro Credit Loan Defaulter, which also helped me in doing a research and I came to know about so many new things.

I am really thankful to them.

I would aldo thankful to the online platforms who help me a lot in finishing this project within the limited time.

THANKS AGAIN TO ALL WHO HELPED ME.

**INTRODUCTION**

* Business Problem Framing

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

The company provide micro credit loan to their customers. 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.

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.

* Conceptual Background of the Domain Problem

So we build a machine learning model that helps to understand the company that who is defaulter customer and who is not based on the sample data. And Exploratory Data Analysis help them to visualize the data graphically which is esay to understand.

* Review of Literature

In this project the sample data is provided to us from our client database. 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.

The client wants some predictions that could help them in further investment and improvement in selection of customers. So we build a machine learning model to help them.

First we do all the data preprocessing steps and do EDA to visualize the data graphically and after that we make a machine learning model in order to improve the selection of customers for the credit.

* Motivation for the Problem Undertaken

Every investigation begins with ideas that are further developed and inspired to address a variety of situations and circumstances.

The client wants some predictions that could help them in further investment and improvement in selection of customers. So to help them we make this project.

My motivation behind this project is to do the proper research because research as a process for finding a solution to a problem after making a deep analysis and conducting studies of relevant factors. In general, research is a method designed to ensure that the information obtained is reasonable and supported by the quantitative and qualitative data, and that involves a systematic process. It includes the process of designing research methods, collecting and describing.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

I have use describe function from pandas library to see all the statistical model of our dataset.



This function gives all the statistics summary of our dataset that is mean, median, minimum and maximum values of our dataset. It also tells whether data is right skewed or left skewed based of the diference between mean and median.

From that part we also able to analyze that whether outliers are present in our dataset or not based of the difference between 75% and maximum value. The describe function also gives the counting of our data for the particular column.



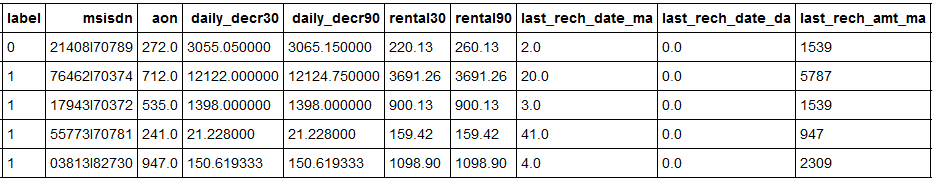
From this function we know the statistics of column which is object or datetime format.

We do data exploration to see the insight of our data and if any duplicates row is present then we have to remove it.

Before going to model training we have to change the data type of the columns into numeric datatype by using encoding technique.

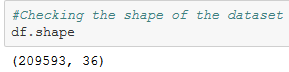
* Data Sources and their formats

In this project the sample data is provided to us from our client database. The dataset is in csv (comma seprated values) format.



Here you see the sample of our dataset.

This dataset contains 209593 rows and 36 columns.



From the above function we can see the shape of our dataset.

Dataset contains mixed type of columns. Most of the columns are numeric data type but some columns are object and only one column is in datetime datatype.



We can see the datatype of our dataset by using this function.

* Data Preprocessing Done

**Deleting duplicates rows**:

In data cleaning steps first we remove the duplicates rows .



In our data set msisdn is a mobile number of user and mobile number is unique for every customers. There are only 186243 unique number out of 209593 so rest of the data is duplicates entry so we have to remove those entry.

**Checking the skewness***:*

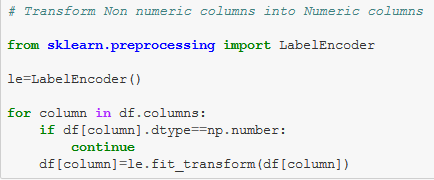
A variable is considered, highly skewed if its absolute value is greater than 1. A variable is considered,moderately skewed if its absolute value is greater than 0.5.

**

This function tells the skewness of our dataset. Most of the columns is right skewed which we handel it via scaling because if we remove skewness via log method or boxcox then NaN value occures in these columns.

**PreProcess Categorical Features:**

If we take a look at the columns, we can see that there are some features like ‘msisdn’ or ‘pdate’ that are not numerical, they are categorical. Machine learning algorithms expect to work with numerical values, so these categorical features should be transformed. One of the most popular categorical transformations is called ‘**Label Encoder**’ as below



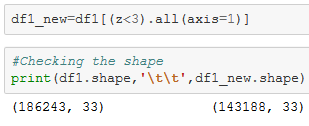
***Detecting and Removing outliers:***

The outliers is an observation point that is distant from other observation. The data science projects starts with collection of data and that’s when outliers first introduced to the population. Though we will not know about the outliers at all in the collection phase.The outliers can be a result of a mistake during data collection or it can be just an indication of variance in our data.

There are two ways of analysis we will follow to find out the outliers-Uni-variate (one variable outlier analysis) with the help of box plot and Multi-variate (two or more variable outlier analysis) with the help of scatter plot.

**

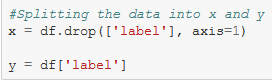
**

**

So if we remove outliers, 23% of our data is removed as a outliers. So we are not going to remove outliers.

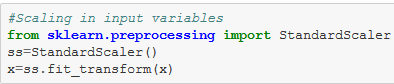
**Splitting the data into input and target features:**

Before building a machine learning model, data is always split into two different parts that are called Training and Testing. For the training purpose of the model, we only expose the training data and never allow testing data to be exposed. Once the model gets trained using that data, we make use of the model to compute the predictions over the testing data, which is stored in a single variable known as y\_pred. We can store it in a different variable as well. We will first define the independent variable and dependent variable x and y respectively.



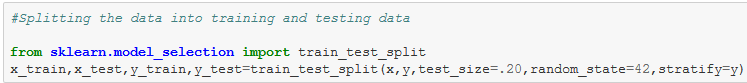
**Feature Scaling:**

It is recommended to perform some type of scaling on numerical features. It is used to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.



***Splitting the data:***

Now when all categorical variables are transformed and all numerical features are normalized, we need to split our data into training and test sets. We split 80% to training and 20% for testing.

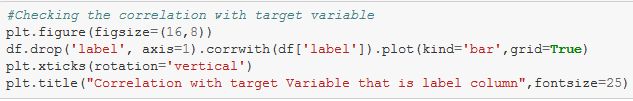


* Data Inputs- Logic- Output Relationships

 Computers learn from experience that they are presented with in terms of data. The data to enable learning exists relative to a specific task and consists of several parameters:

* *T*, a task to accomplish, e.g. predict the labels which is our target variable
* *E*, some value of experience, e.g. label observed
* *P*, some value of performance, e.g. how many labels are predicted

Once a program has learnt from these inputs, it can take a new, previously unseen experience and from that predict, in our example, the specific labels. The labels might be strongly correlated to say age on mobile network, maximum number of recharge and maximum amount of loans taken, and the average main account balance and daily amount spend from main account.



This code will show the correlation of the columns with target column which is label. I am going to shown the above code result in visualization part.

* Hardware and Software Requirements and Tools Used

Hardware:

Processor—Intel (R) Core(TM) i5-2430M CPU @ 2.40GHz

Installed Memory(RAM)—8.00 GB

System type—64-bit Operating System

Software: Windos 10 Pro

We have used Python Package because it is powerful and general purpose programming language.

NumPy—It is a math library to work with ndimensional arrays. It enables us to do computation effectively and regurarly. For working with arrays, dictionary, functions data type we need to know NumPy.

Pandas—It is high level Python library and easy to use for data importing , manipulation and data analysis.

Matplotlib—It is a plotting that provide 2D and 3D plotting.

Seaborn-- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

SciPy—It is a collection of numerical algorithm and domain specific tool boxes including optimization, statistics and much more.

Scikit-learn—It is a collection of tools and algorithm for machine learning. It works with NumPy and SciPy and it is easy to implement machine learning models.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

We know that this is classification problem so we use accuracy score, classification report and confusion matrix as our eveluation matrix. We also see the AUC score and also plot the AUC\_ROC curve for our final model.

As we know this dataset is imbalance so we don’t too much focus on accuracy score . We see the precision and recall value along with f1\_score.

* Testing of Identified Approaches (Algorithms)

As we know that this dataset is imbalance so first we see the result without doing any sampling and for that I use Logistic Regression with KFold cross validation and hyperparameter tuning.

We also use Random Forest Classifier as our evaluation model without using hyperparameter tuning because our dataset is too large and it takes more than hour to give the result.

After that I use sampling technique and use our Random Forest model with sampled training data and see the results.

There are three sampling technique which I have used to handle the imbalance problem of our dataset.After using Sampling technique to balance our dataset and then we apply Random Forest classifier model and see the results.

* **Under Sampling:** By using the under sampling we are trying to reduce the points of maximum labels.
* **Over Sampling:** By using the over sampling we are trying to increase the points of minimum labels.
* **SMOTETomek:** SMOTETomek is a hybrid method which uses on under sampling method (Tomek) in with over sampling method (SMOTE).
* Run and Evaluate selected models

### Logistic Regression by using Cross Validation like KFold and Hyperparameter tunning:

### 

### 

### Output:

### 

### Random Forest Classifer:

### 

### 

### Output:

### 

### We clearly see that precision and recall is very good in Random Forest as compare to Logistic Regression.

### Now we are using some sampling technique and use Random Forest Model and see the result.

### Under Sampling: By using the under sampling we are trying to reduce the points of maximum labels.

### 

### Output:

### 

### Over Sampling: By using the over sampling we are trying to increase the points of minimum labels.

### 

### Output:

### 

### SMOTETomek: It is a method of imblearn. SMOTETomek is a hybrid method which uses on under sampling method (Tomek) in with over sampling method (SMOTE).

### 

### Output:

### 

From above all results we see that under sampled model gives very poor result. So don't use under sampling unless, untill your dataset is very samll.

### The Precision and recall value is high when we use over sampling. So we choose over sampled model as our final model.

* Key Metrics for success in solving problem under consideration

### Although it is a classification problem so we use accuracy score, classification report and confusion matrix as our evaluation metrics along with AUC ROC score.

### As we know that our dataset is imbalance so we do not much focus upon the accuracy score. We mainly see the precision and recall value of our model.

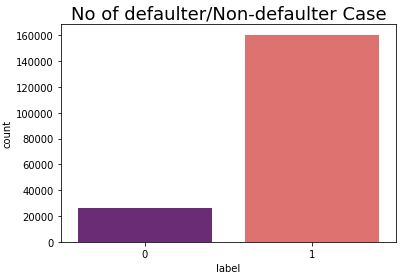
### Precision talks about all the correct predictions out of total positive predictions. Recall means how many individuals were classified correctly out of all the actual positive individuals.

### Along that we also consider AUC ROC score as our key metrix. Because 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.

* Visualizations

It is the graphical representation of data that is used to check about the presence of outliers, patterns, distribution of the data, etc. There are different data visualisation libraries in python that include matplotlib, seaborn, etc. We will make use of the seaborn and matplotlibtolibrary to visualise the dataset.

***Checking the count of Defaulter and Non-Defaulter Customers:***

****

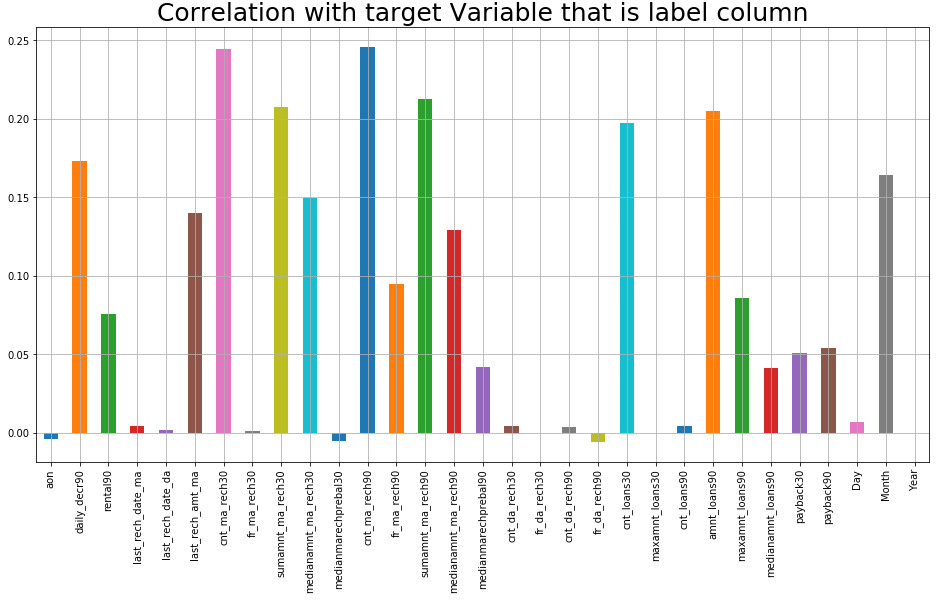
Label 1 indicates loan has been payed i.e Non-Defaulter and label 0 indicates indicates that the loan has not been payed i.e. defaulter.

**1 160383**

**0 25860**

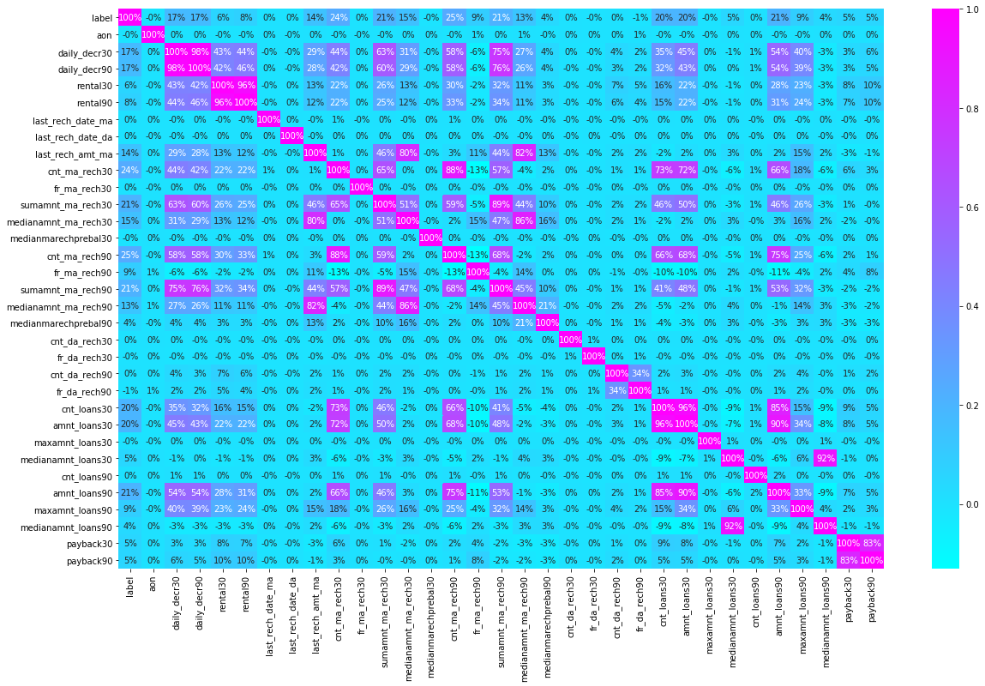
86% of the sample are in Non defaulter category, and 14% of the sample is in defaulter category. We will take this into account when splitting the data into a training and test set. We also set a seed to make this blog reproducible.

**Correlation with target feature:**

****

In the above graph we can see the correlation of the columns with respect to target column that is label.

Correlation Matrix:



1-daily\_decr30 and daily\_decr90 features are highly correlated with each otheer.

2-rental30 and rental90 features are highly correlated with each other.

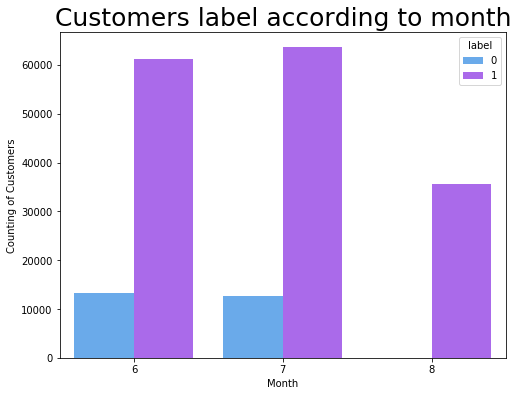
3-cnt\_loans30 and amount\_loans30 columns are highly correlated with each other.

4-amount\_loans30 is also highly correlated with amount\_loans90 column.

5-medianamnt\_loans30 and medianamnt\_loans90 is highly correlated with each other.

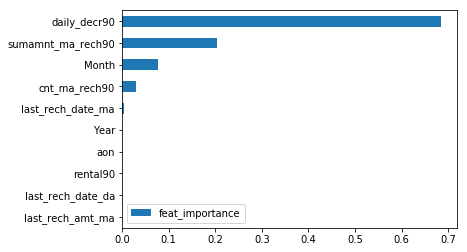
6-We have to drop one of the features which are highly correlated with other feayures. And if we dont do this then our model will face multicolinearity problem.

**Customers Label according to month:**



There are severals customers at June and July month who did not pay their loan.

**Feature Importance**:



By looking at the daily\_decr90 which is Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah), it seems that this feature helps to discriminate the data indeed. This feature can bring insights for company when analyzing a customers.

* Interpretation of the Results

From the above interpretation we come to know that this is classification based problem so we have learned to build a complete machine learning model for classification based problem. msisdn, aon and maxamnt\_loans90 features are good for prediction level. Label ‘1’ indicates that the loan has been payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

We will also check the most important features according to decision tree. This information can bring insights by applying feature engineering to improve the classification of the target.

We also visualize the data and see the outcomes of our result that what percentage of customers are in defaulter case, which column is most correlated with target column and much more.

**CONCLUSION**

* Key Findings and Conclusions of the Study
* There are no null values in the dataset.
* The dataset is imbalanced. Label ‘1’ has approximately 86% records, while, label ‘0’ has approximately 14% records.
* maxamnt\_loans90 columns gives information about customers with no loan history.
* msisdn and aon features some values which might not be realistic. So drop the row which contain not realistic value
* There are some rows which is repeated means duplicate entries are present in our dataset.
* Sampling data gives the better precision and recall value along with auc score.
* The given dataset has only three months of data and it is also not the present year data.
* The collected data is only for one area circle.

* Learning Outcomes of the Study in respect of Data Science

In this project we learn how to build a machine learning model for classification based problem.

We also learn how to handle the imbalanced dataset for machine learning model. Because when we over sampling and use this over sampled data to build a ML model then it gives the better result.

The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.

On doing this project the biggest problem I have faced is that I am not able to use GridSearchCV. Because when I use GridSearchCV then my system takes too much time to give the result as our dataset is too large . So If I uses GridSearchCV then our result improves.

So based on all the learning and outcomes our Random Forest Classifier Model with over sampled data gives the best result so we save this model as our final model by using Joblib as a pickle file.

* Limitations of this work and Scope for Future Work

The dataset is belongs to only one area circle so our model trained for the data of only one area circle and also this data is too old it is from 2016 and contains only three months of data.

As Today I uses i5 processor computer so I am not able to use GridSearchCV and cross validation because my system takes too much time to give the result. If some how I uses GridSearchCV and Cross validate the model then our model result will be surely improved.

I would conclude the project report by hoping that now you have understood every step that is required to be done to build a machine learning model. We have built the classification model for classifying the labels that is which customers is in label 1 that is Non defaulter and which one is in label 0 that is defaulter customer and then evaluated it using different error metrics.