

**MICRO CREDIT DEFAULTER MODEL**

**SUBMITTED BY:**

**RITU RANI**

**INTRODUCTION**

MFI (Microfinance Institution) is an organization that provide financial services to low income groups. Microfinance services becomes very useful when targeting especially the unbanked poor families which are living in remote areas and they don’t much sources of income. MFS provided by Microfinance Institution are Group Loans, Individual Business Loans, Agricultural Loans and so on.

The MFI industry is mainly focusing on low income families and are very useful in such areas, the implementation of Microfinances services 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.

**PURPOSE:**

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 a Microfinance Institution 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 be 6 (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.

**PROBLEM STATEMENT:**

We will build a model which can be used to predict in terms of a probability for each loan transaction, whether the consumer will be able to pay back the loaned amount within 5 days of insurance of loan.

In this case, 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.

In this notebook, I'll build a model to predict if an applicant get approval for loan or not. Machine Learning techniques are very useful in predicting outcomes for large amount of data. In this paper five machine learning algorithms, Logistic Regression (LR), Support Vector Classifier (SVC), Decision Tree Classifier (DTC), K Nearest Neighbour (KNN) and Random Forest Classifier (RF) are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Random Forest Classifier machine learning algorithm is better as compared other machine learning approaches.

Our task is to predict whether the consumer will be able to pay back the loaned amount within 5 days of insurance of loan approval or not using various classification algorithms. Exploratory data analysis is done on the dataset to achieve insights and the pre-processing pipeline is done to get the data ready for the training.80% of the data is used for training purpose and 20% for the testing purposes. Five Classification models are trained and their performances are compared with various performance metrics like confusion metrics, accuracy score, cross validation score and the Receiver operating characteristic curve.

**LOADING DATASET**

The data set is of shape (209593,37) i.e. It has 37 attributes and 209593 rows.

The dataset provides 36 input variables and 1 target variable that are a mixture of ordinal, categorical and numerical data types. Following are the variables is our dataset:

1. **Label:** Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan {1: success, 0: failure}.
2. **Msisdn**: mobile number of users.
3. **Aon:** age on cellular network in days.
4. **daily\_decr30:** Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
5. **daily\_decr90:** Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
6. **rental30:** Average main account balance over last 30 days
7. **rental90:** Average main account balance over last 90 days
8. **last\_rech\_date\_ma:** Number of days till last recharge of main account
9. **last\_rech\_date\_da:** Number of days till last recharge of data account
10. **last\_rech\_amt\_ma:** Amount of last recharge of main account (in Indonesian Rupiah)
11. **cnt\_ma\_rech30:** Number of times main account got recharged in last 30 days
12. **fr\_ma\_rech30:** Frequency of main account recharged in last 30 days
13. **sumamnt\_ma\_rech30:** Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
14. **medianamnt\_ma\_rech30:** Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)
15. **medianmarechprebal30:** Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)
16. **cnt\_ma\_rech90:** Number of times main account got recharged in last 90 days
17. **fr\_ma\_rech90:** Frequency of main account recharged in last 90 days
18. **sumamnt\_ma\_rech90:** Total amount of recharge in main account over last 90 days (in Indonesian Rupiah)
19. **medianamnt\_ma\_rech90:** Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah)
20. **medianmarechprebal90:** Median of main account balance just before recharge in last 90 days at user level (in Indonesian Rupiah)
21. **cnt\_da\_rech30:** Number of times data account got recharged in last 30 days
22. **fr\_da\_rech30:** Frequency of data account recharged in last 30 days
23. **cnt\_da\_rech90:** Number of times data account got recharged in last 90 days
24. **fr\_da\_rech90:** Frequency of data account recharged in last 90 days
25. **cnt\_loans30:** Number of loans taken by user in last 30 days
26. **amnt\_loans30:** Total amount of loans taken by user in last 30 days
27. **maxamnt\_loans30:** maximum amount of loan taken by the user in last 30 days
28. **medianamnt\_loans30:** Median of amounts of loan taken by the user in last 30 days
29. **cnt\_loans90:** Number of loans taken by user in last 90 days
30. **amnt\_loans90:** Total amount of loans taken by user in last 90 days
31. **maxamnt\_loans90:** maximum amount of loan taken by the user in last 90 days
32. **medianamnt\_loans90:** Median of amounts of loan taken by the user in last 90 days
33. **payback30:** Average payback time in days over last 30 days
34. **payback90:** Average payback time in days over last 90 days
35. **pcircle:** telecom circle
36. **pdate:** date

We have three kinds of data types:

**Object:** It means variables are categorical.

Following are the Categorical variables in our dataset: msisdn, pcircle, pdate.

**int64:** It represents the integer variables.

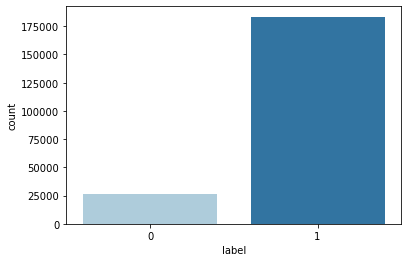
Following are the integer variables in our dataset: label, last\_rech\_amt\_ma, cnt\_ma\_rech30, cnt\_ma\_rech90, fr\_ma\_rech90, sumamnt\_ma\_rech90, cnt\_da\_rech90, fr\_da\_rech90, cnt\_loans30, amnt\_loans30, amnt\_loans90, maxamnt\_loans90

**float64:** It represents the variable that has some decimal values. They are also numerical. Following are the float64 variables in our dataset: aon, daily\_decr30, daily\_decr90, rental30, rental90, last\_rech\_date\_ma, last\_rech\_date\_da, fr\_ma\_rech30, sumamnt\_ma\_rech30 medianamnt\_ma\_rech30, medianmarechprebal30, medianamnt\_ma\_rech90, medianmarechprebal90, cnt\_da\_rech30, fr\_da\_rech30, maxamnt\_loans30, medianamnt\_loans30, cnt\_loans90, medianamnt\_loans90, payback30, payback90.

**Exploratory Data Analysis**

Now, we will do exploratory data analysis to get the insight about the data and how target variable depends on various attributes.

First, we are analyzing our target variable i.e. “label”.

 Figure.1

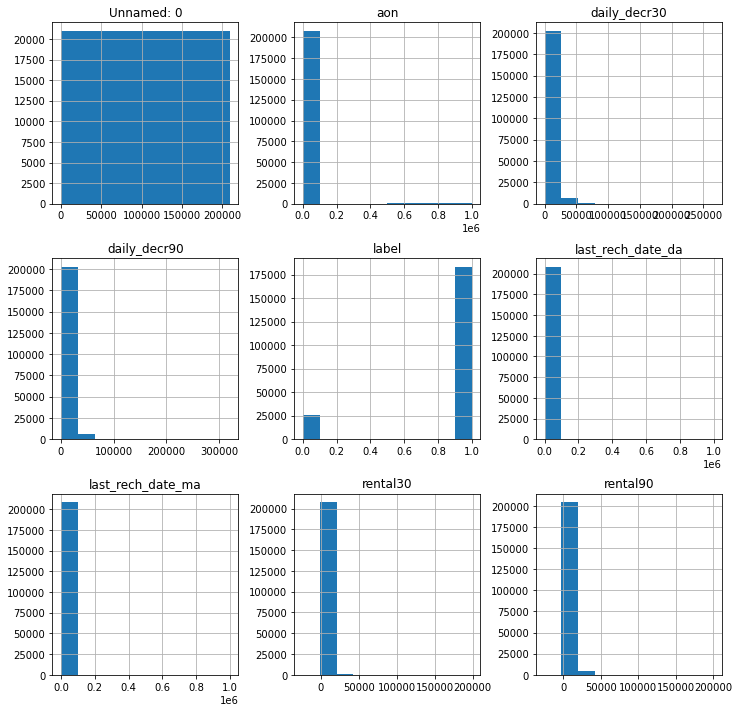
This figure shows the analysis of our target variable.

1=183431 (No. of users paid back the credit within next 5 days)

0 = 26162 (No. of users who failed to pay back the credit within next 5 days).

**ANALYSIS OF CONTINOUS VARAIBLES:**

First, analysing top 10 features:

  
 Figure.2

We will drop our first attribute named as “Unnamed” because it is just the number of rows of the dataset.

Except “label” which is ordinal in nature, all other features are right-skewed distributed.

A right-skewed distribution has a long right tail and also called positive-skew distributions. This is because there is a long tail in the positive direction on the number line.

Now, we analyse next 10 attributes:

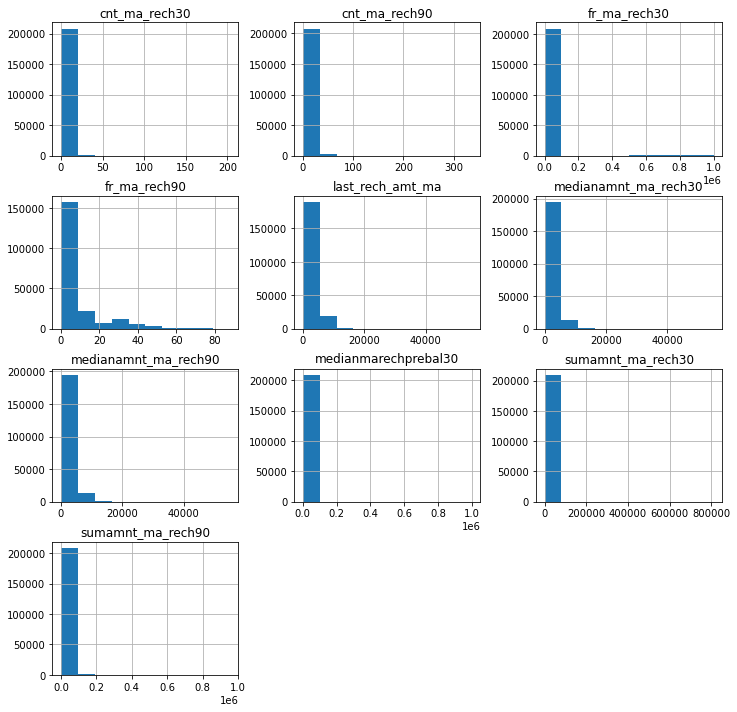


Figure.3

We can analyse that all features are right-skewed. We will remove skewness by power-transform later.

Analysing next 10 attributes:

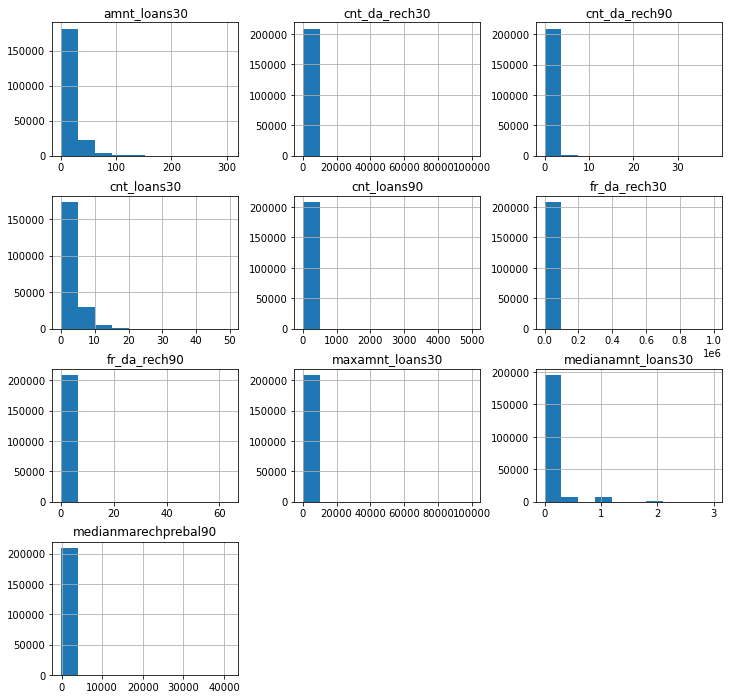
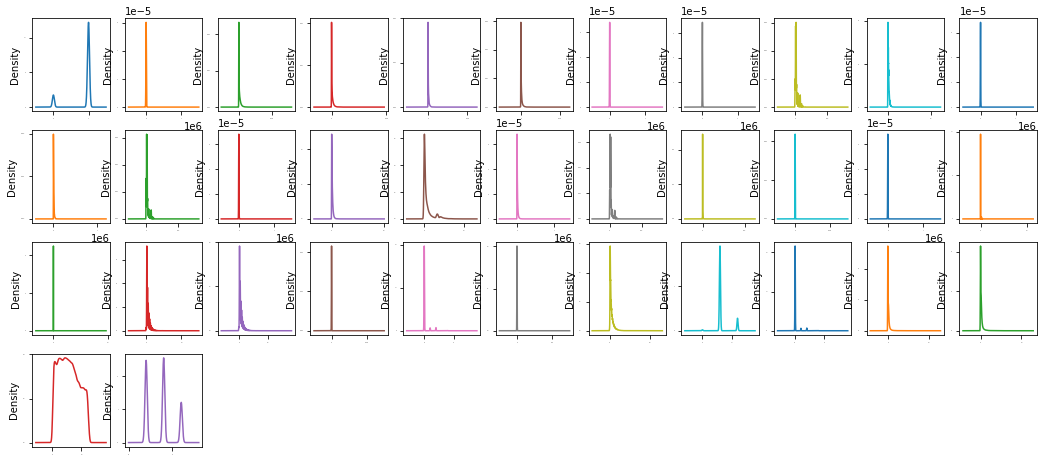


Figure.4

All these attributes are right-skewed distributed.

**OBSERVATION:** Figure.5

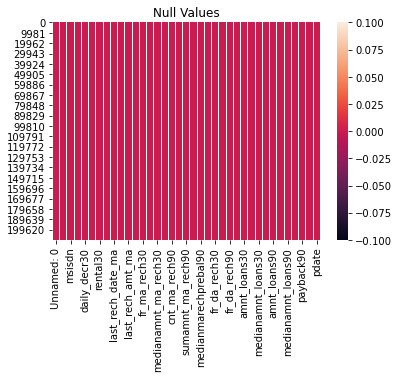
****

We can observe from above graphs that in all the independent attributes right skewness is present.

**DATA PRE-PROCESSING:**

**Data pre-processing** is very essential step in any **data** mining process. It directly impacts the predictions of the model. If data is unclean, have missing vales, missing attributes or contains outliers, if skewness is present, then all these factors degrade the quality of our results and our predictions will be biased.

First, we will check for missing or NaN values through heat-map:

 Figure.6

We can see that there is no NaN or missing values present in out dataset.

**Handling date column:**

We have converted the pdate attribute into three different features named as “day”, “month”, “year”.

**Drop irrelevant columns:**

We have dropped the irrelevant features which are not contributing to our predictions.

The features which we dropped are 'Unnamed: 0', 'msisdn', 'pcircle', 'pdate', 'year'.

**Correlation matrix heatmap:**

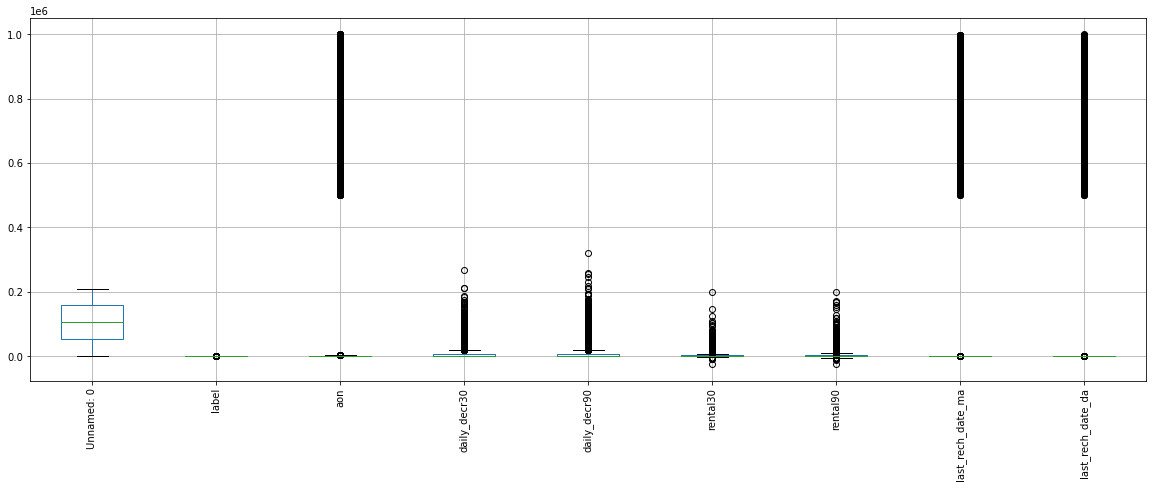
Checking correlations is very important to analyse data. A heatmap has been plotted to check the correlation between the attributes, if there is positive or negative relationship. This is one of the methods to decide which attributes affect the target variable the most.

OBSERVATIONS:

* daily\_decr30 & daily\_decr90 are strongly correlated.
* daily\_decr30 & sumamnt\_ma\_rech90 are strongly correlated.
* rental30 & rental90 are strongly correlated.
* medianamnt\_ma\_rech30 & last\_rech\_amt\_ma is strongly correlated.
* amnt\_loans90 & amnt\_loans30 are strongly correlated.

**OUTLIERS:**

An outlier means an observation that falls outside the overall pattern or we can say an abnormal distance from other values in a random sample from a population.



We have outliers present in all attributes. From scipy.stats we imported Z-score and drop all the rows in which threshold value is greater than 3. But by dropping these rows we lost our 21 percent data and the results came from our predictions will be biased. Therefore, we will not drop these outlier values as these values are important for our predictions.

**TREATING SKEWNESS:**

If the skewness is between -0.5 and 0.5 then the data is fairly symmetrical and represent normal distribution. If the skewness is between 0.5 and 1 or -1 and -0.5 then the data is moderately skewed. If the skewness is less than -1 or greater than 1then the data is highly skewed.

As we earlier analysed that skewness is present in all the attributes. Therefore, we will treat this with power transformation.

**SPLITTING DATASET:**

The shape of the dataset after dropping of the irrelevant columns is (209593,35). We split the dataset where 80% is used for training the model and 20% for testing the model. Hence out of 209593 data entries, 167675 are used for training and 41918 are used for testing the model.

**FINDING BEST RANDOM STATE**

Our model best performs at random state 77 and we are achieving 0.88 accuracy score.

Five Classification Algorithms are used.

1. Logistic Regression
2. RandomForestClassifer
3. SupportVectorClassifier

**Logistic Regression**

We are achieving 89% accuracy with Logistic Regression.

**Support Vector Classifier**

We are achieving 89.3% accuracy with SVC

**Random Forest Classifier**

We are achieving 92.2% accuracy with Random Forest Classifier.

**K Nearest Neighbour**

We are achieving 90% accuracy with K Nearest Neighbour.

**Decision Tree Classifier**

We are achieving 88.3% accuracy with Decision Tree Classifier.

Therefore, we proceed with Random Forest Classifier as it is giving highest accuracy score.

**TUNNING WITH BEST PARAMETERS:**

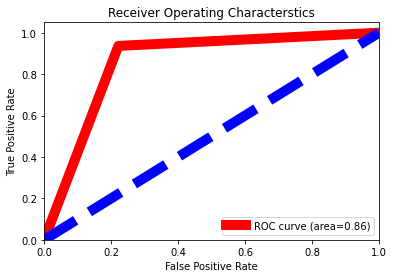
Imported RandomizedSearchCV from sklearn.model\_selection and find out the best parameters of RandomForestClassifier which performed best on our model.

Following are the best parameters for our model:

{'max\_features': 'log2', 'criterion': 'entropy'}

**ROC (Receiver Operating Characteristic) Curve**

Receiver Operating Characteristic curve or ROC curve represents a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the TPR against the FPR at various threshold settings. The true-positive rate (TPR) is also known as recall, sensitivity, or probability of detection in machine learning. The false-positive rate (FPR) is also known as the probability of false alarm or fall-out.

**** Figure.8

The area under the curve is 86%

**SAVE THE MODEL**

We finally save our best model by importing pickle. The use of pickle is widespread as they allow us to easily transfer data from one server or system to another and then store it in a file or database.

**CONCLUSION**

We successfully predict whether the consumer will be able to pay back the loaned amount within 5 days of insurance of loan using various classification algorithms. Exploratory data analysis is done on the dataset to achieve insights and the pre-processing pipeline is done to get the data ready for the training.80% of the data is used for training purpose and 20% for the testing purposes. Five Classification models are trained and their performances are compared with various performance metrics like confusion metrics, accuracy score, cross validation score and the Receiver operating characteristic curve. The RandomForestClassifier comes out to be the best performing algorithm above all other models with an accuracy of 92.3% and over all generalizing well.

**DOWNLOAD JUPTYER NOTEBOOK**

Click on the below link to find my juypter notebook in GitHub:

<https://github.com/riturani2403/FlipRobo-Internship/blob/main/Micro%20Credit.ipynb>