

**NAME OF THE PROJECT**

Micro-Credit Defaulter ML

**Submitted by:**

Yogesh.C.Mudliar

**ACKNOWLEDGMENT**

It is a genuine pleasure to express my deep sense of thanks and gratitude to my guide, Ms.Khushboo Garg, for allowing me to work on this project. It was a great way to expose myself to the actual research environment.

I thank FLIPROBO for permitting me to work with them.

I take this opportunity to say heartfelt thanks to Dr. Deepika Sharma, VP-learning And development DataTraind for her overall dedication, devotion, and support towards me. I convey my sincere regards to all the DataTraind team thanks for supporting me during academic years of my post-graduation course in data science.

I express my profound sense of gratitude to my mentor Ms.Khushboo Garg, FLIPROBO for her guidance at every step of my research work.

Apart from the project, I learned a lot from her, she gave me valuable thought- "To think"; that I will benefit from, for a long time to come. I am indebted to her more than she knows.

**References:**

1) <https://www.investopedia.com/terms/m/microcredit.asp>

2) www.Google.com

3) <https://app.grammarly.com/>

4) https://www.youtube.com/watch?v=NeOxIiV\_ikw

5)https://www.researchgate.net/publication/265161200\_Predicting\_Credit\_Default\_among\_Micro\_Borrowers\_in\_Ghana

6) https://en.wikipedia.org/

**INTRODUCTION**

* **Business Problem Framing**

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.

**Exercise:**

Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label ‘1’ indicates that the loan has been paid i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been paid i.e. defaulter.

* **Conceptual Background of the Domain Problem**

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.

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

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.

The client wants some predictions that could help them in further investment and improvement in selection of customers.

**Exercise:**

We are building a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label ‘1’ indicates that the loan has been paid i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been paid i.e. defaulter.

* **Review of Literature**

Microcredit, also called micro banking or microfinance, a means of extending credit , usually in the form of small loans with no collatera, to non-traditional borrowers such as the poor in rural or undeveloped areas.

Microfinance institutions play a major role in economic development

in many Developing countries. However many of these microfinance institutions are faced with the problem of default because of the non-foral nature of the business and individuals they lend money to. This study

seeks to find the determinants of credit default in microfinance

institutions. With data on 209592 rows and 37 Columns loans from a

microfinance institution with braches all over the country we proposed a Random Forest Classifier model to predict the probability of default. Microfinance institutions could use this model to screen prospective loan applicants in order to reduce the level of default.

* Motivation for the Problem Undertaken

Microfinance institutions play a major role in economic development in many developing countries. However many of these microfinance institutions are faced with the problem of default because of the non-formal nature of the business and individuals they lend money to. This study seeks to find the determinants of credit default in microfinance institutions.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

We are building a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan.

There are no null values in the dataset. There may be some customers with no loan history. The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records.

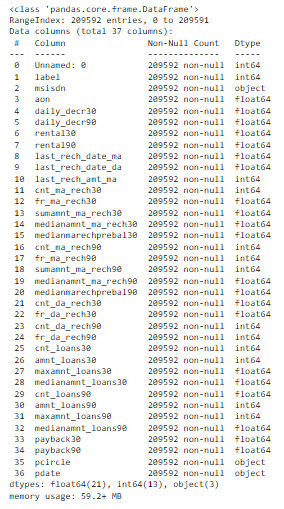
We have checked the correlation of data. In most of the columns Outliers are present, we have removed the outliers using the zscore method. We have used the various visualization plot for all features. In the distribution plot we observed that there is a skewness present in data. We have removed the skewness using the yeo-johnson method. Then we have built the model, checked their accuracy score, confusion matrix and classification report.

I done Extensive EDA have to be performed to gain relationships of important variable

* Data Sources and their formats

The data was provided by the Flip Robo Technologies and it is in the .csv format. The data is huge it has 209592 rows and 37 columns.

-The below image shows data format



**Data Description:**

|  |  |
| --- | --- |
| Variable | Definition |
| 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 user |
| 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

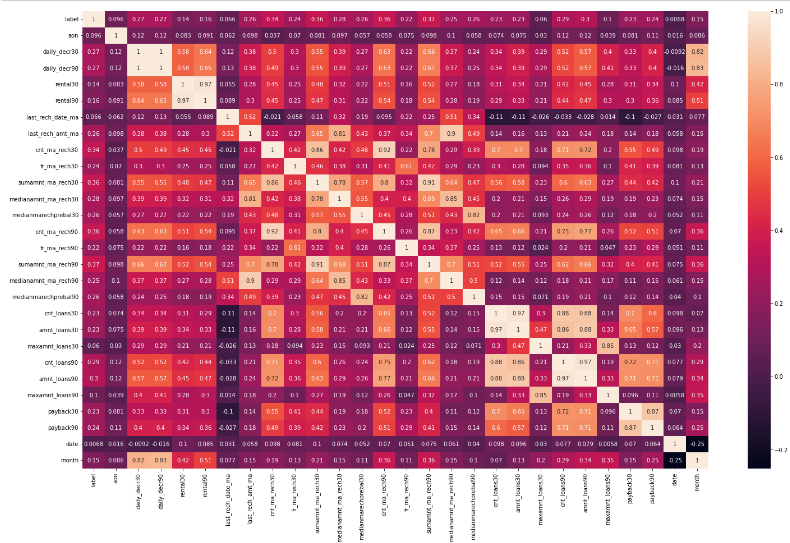
After collecting a data, the next step is to get it ready for analysis. This means cleaning, or ‘scrubbing’ it, and is crucial in making sure that you’re working with high quality data.

* **Removing unwanted data points** — extracting irrelevant observations that have no bearing on intended analysis.
* **Bringing structure to data** — fixing layout issues, which will help to map and manipulate this data more easily.
* **Filling in major gaps** — this data contains null values and I notice that important data are missing. Once we identified gaps, we can go about filling them.

During cleaning the data we have to do EDA(Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations). This helps identify initial trends and characteristics, and can even refine our hypothesis.

* Data Inputs- Logic- Output Relationships

With the correlation plot we can understand the relationship between each feature with the target variables.



There are some inputs which is important to find our outputs like the price of houses with the available independent variables.

'aon', 'daily\_decr30', 'daily\_decr90', 'rental30', 'rental90', 'last\_rech\_date\_ma', '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\_loans30', 'amnt\_loans30', 'maxamnt\_loans30', 'cnt\_loans90', 'amnt\_loans90', 'maxamnt\_loans90', 'payback30', 'payback90'

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

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. MFIs can sustain and increase deployment of loans to stimulate the poverty reduction goal if repayment rates are high and consistent. MFIs are able to reduce interest rates and processing fees if repayment rates are high, thus increasing patronage of loans. A high repayment rate is a catalyst for increasing the volume of loan disbursements to various sectors of the economy. Borrowers that do not have formal education are likely to have inadequate knowledge of loan acquisition and management, thereby making them unable to repay the loans given to them. Literate peoples will pay off their loans better than illiterate peoples because they understand the advantages of prompt loan repayment.

* **Hardware and Software Requirements and Tools Used**

Here is the hardware and software used in the project.

Processor – core i3

RAM – 12 GB

SSD – 250 GB

**Software requirements:**

**numpy** : library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**pandas** : software library written for the Python programming language for data manipulation and analysis

**sklearn**: Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines

**seaborn**: Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

**matplotlib.pyplot** : is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

Worked on these models:

1) Logistic Regression

2) Random Forest Classifier

3) Decision Tree Classifier

4) Gradient Bossting Classifier

5) KNeighbors Classifier

**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**

Cleaning, or ‘scrubbing’ the data, and is crucial in making sure that we are working with high quality data.

-**Removing unwanted data points** — extracting irrelevant observations that have no bearing on intended analysis.

-**Bringing structure to data** — fixing layout issues, which will help to map and manipulate this data more easily.

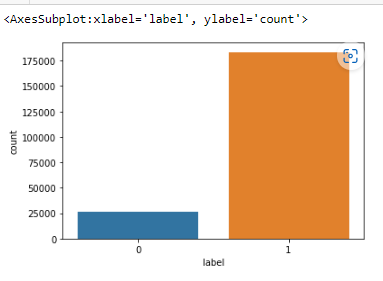
-**Filling in major gaps** — This data contains null values and I notice that important data are missing. Once we identified gaps, we can go about filling them.

During cleaning the data we have to do EDA (Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations). This helps identify initial trends and characteristics, and can even refine our hypothesis.

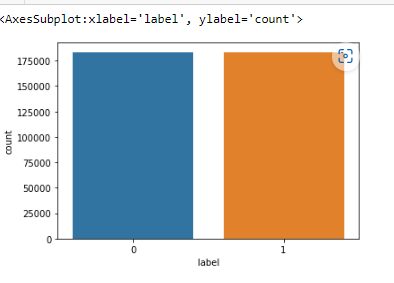
I done Predictive analysis (Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyse current and historical facts to make predictions about future or otherwise unknown events)

Predictions about which variables are important to predict the price of variable, and how do these variables describe to know the person defaulter or non-defaulter.

Balancing our target variable:



**↓**



* **Testing of Identified Approaches (Algorithms)**

1. **Logistic Regression:**

**-** Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

**2) Random Forest Classifier:**

-The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

**3) Decision Tree Classifier:**

**-** The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

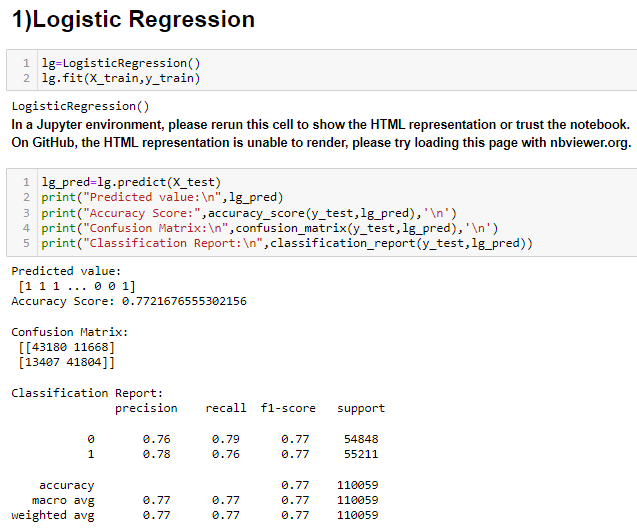
**4) Gradient Bossting Classifier:**

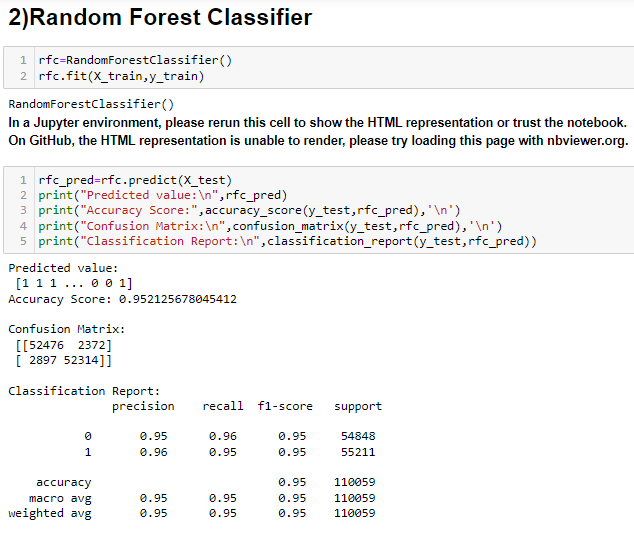
-Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.

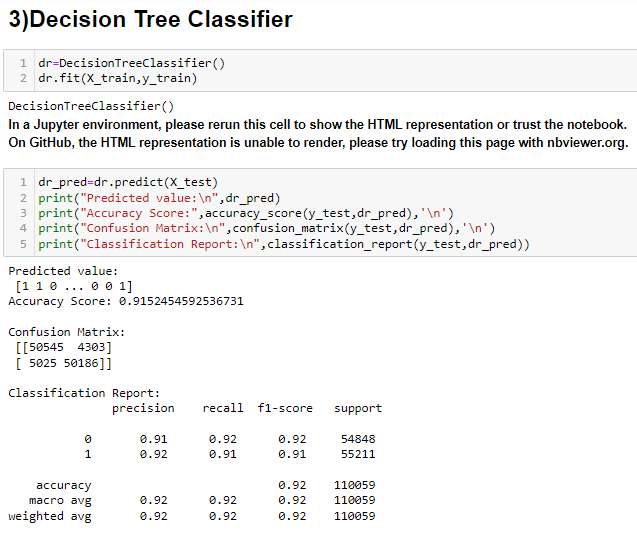
**5) KNeighbors Classifier:**

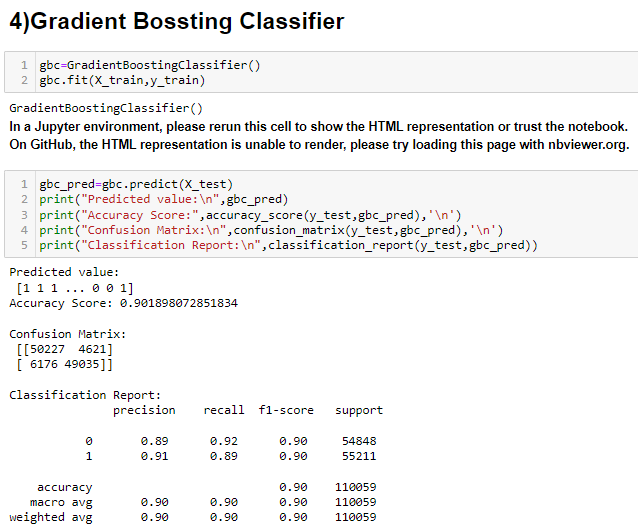
**-** KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

* **Run and Evaluate selected models**











* **Key Metrics for success in solving problem under consideration**

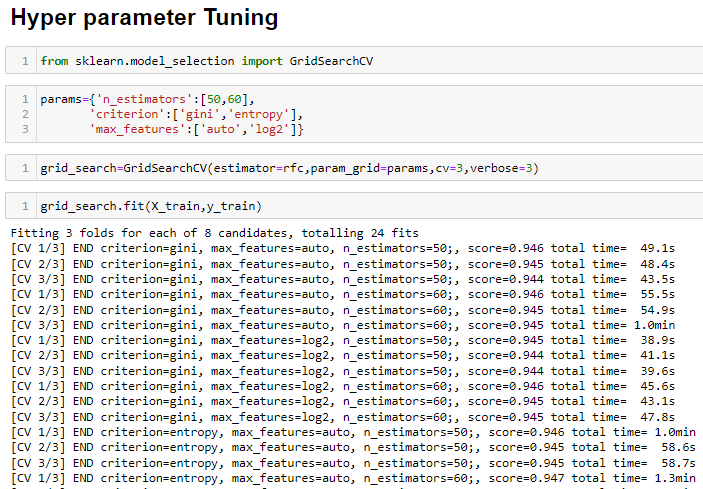
During cleaning the data we have to do EDA (Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations). This helps identify initial trends and characteristics, and can even refine our hypothesis.

I done Predictive analysis (Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyse current and historical facts to make predictions about future or otherwise unknown events)

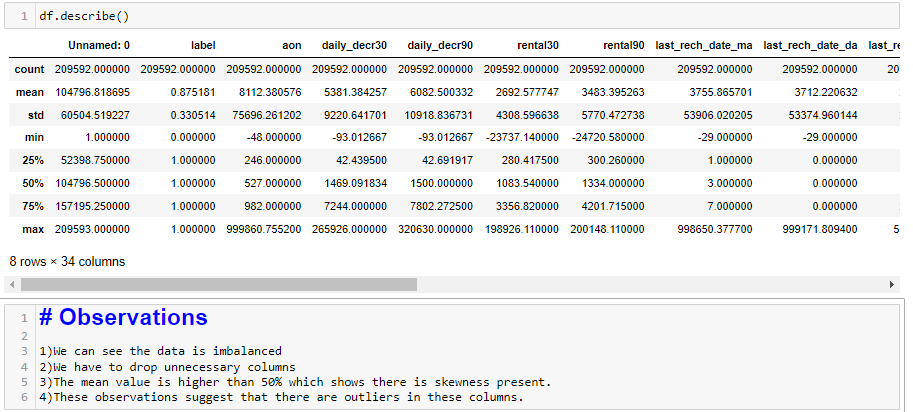
Predictions about which variables are important to predict the price of variable, and how do these variables describe to know the person defaulter or non-defaulter.

**Hyperparameter tuning:**

Consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.



* Visualizations



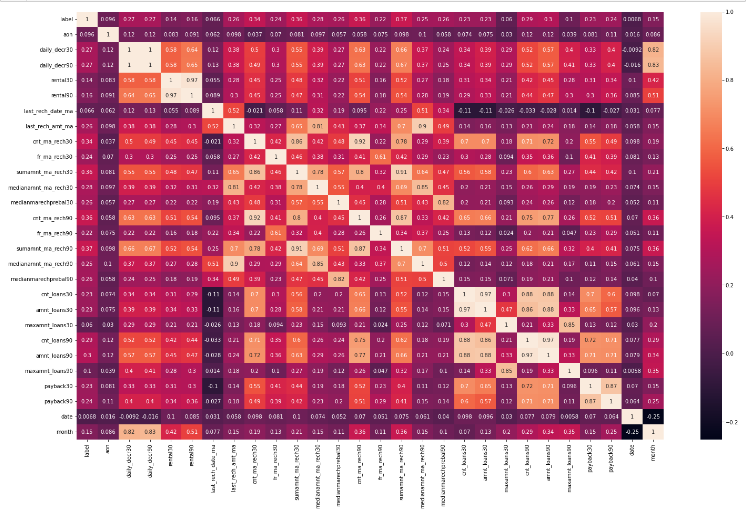
1) We can see the data is imbalanced

2) We have to drop unnecessary columns

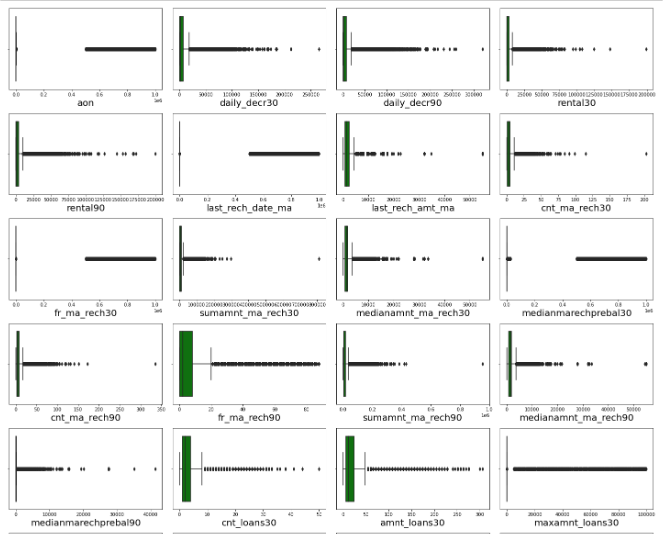
3) The mean value is higher than 50% which shows there is skewness present.

4) These observations suggest that there are outliers in these columns.

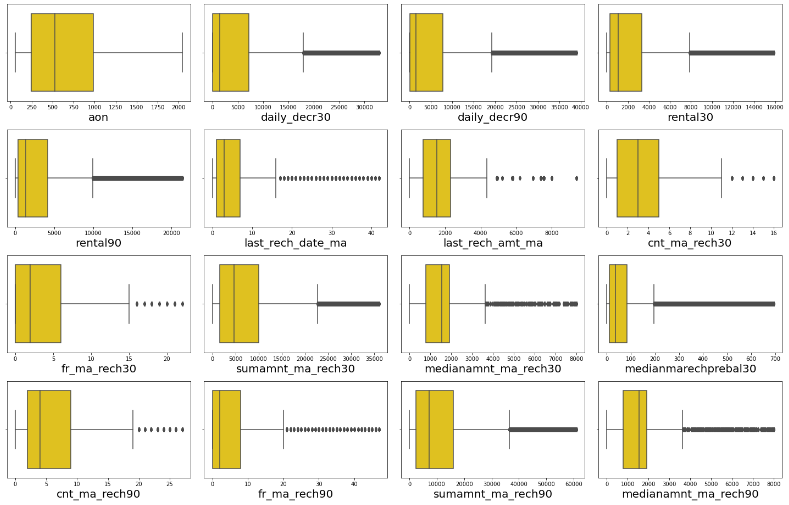
* **Correlation plot**



Removing the outliers from the given datasets



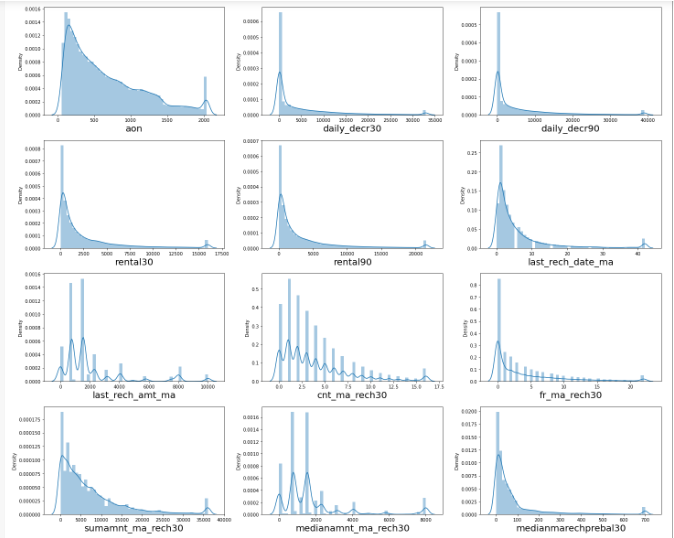
**↓**

****

-We can observe the dataset is imbalenced

- Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images. A distribution can have right (or positive), left (or negative), or zero skewness.

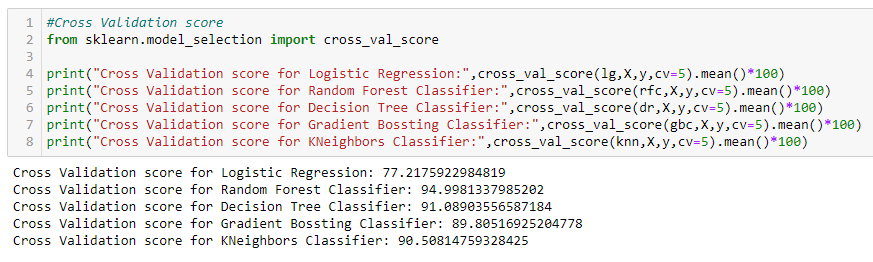
- We have positive skew , the tail of a distribution curve is longer on the right side. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left. Skewness does not inform on the number of outliers; it only communicates the direction of outliers.



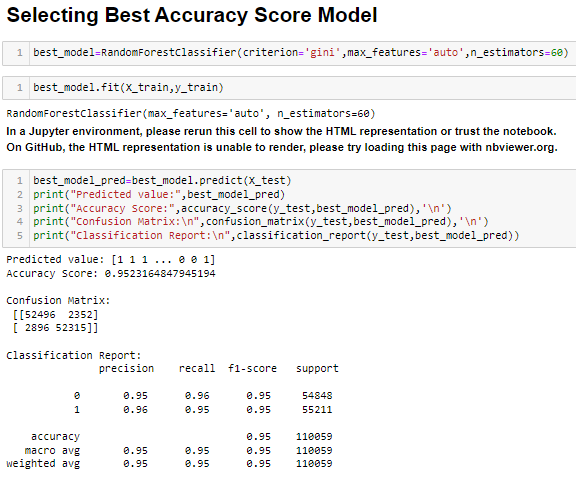
-Balancing the target



* **Interpretation of the Results**



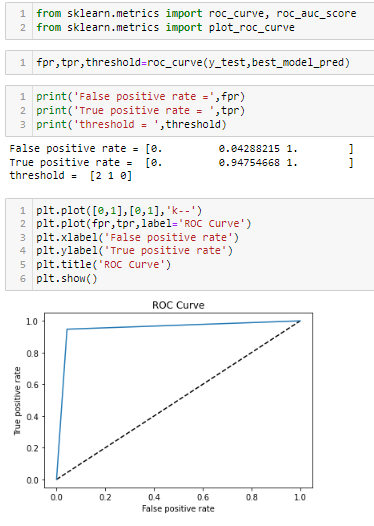
* We got Cross Validation score for Random Forest Classifier: 94.9981337985202



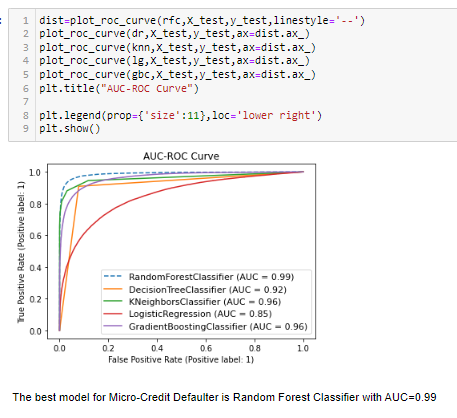
-I select Random Forest Classifier model for Micro-Credit Defaulter, I also checked AUC-ROC curve.

AUC-ROC CURVE:

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.



-I plot graph between all of the models and we found the best model for Micro-Credit Defaulter is Random Forest Classifier with AUC=0.99



**CONCLUSION**

* Key Findings and Conclusions of the Study

Microfinance institutions play a major role in economic development in many developing countries. However many of these microfinance institutions are faced with the problem of default because of the non-formal nature of the business and individuals they lend money. This study

seeks to find the determinants of credit default in microfinance institutions. With data on 209592 rows and 37 Columns we proposed a best model for Micro-Credit Defaulter is **Random Forest Classifier** with AUC=0.99 model to predict the probability of default. Microfinance institutions could use this model to screen prospective loan applicants in order to reduce the level of default.

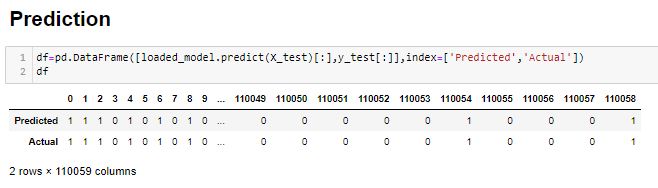
* Learning Outcomes of the Study in respect of Data Science
* First I [define the business](https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/#step-one-defining-the-question) problem and done EDA(Exploratory data analysis ), this dataset having Skewness (measure of the asymmetry of a distribution.) and having outliers(observation that lies an abnormal distance from other values in a random sample from a population.)
* I keep all necessary dataset and drop all unnecessary data.
* [Cleaning the data](https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/#step-three-cleaning-the-data)

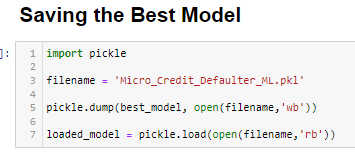
**Removing unwanted data points** — extracting irrelevant observations that have no bearing on intended analysis.

**Bringing structure to data** — fixing layout issues, which will help to map and manipulate this data more easily.

**Filling in major gaps** — this data contains null values and I notice that important data are missing. Once we identified gaps, we can go about filling them.

* [Analysing the data](https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/#step-four-analyzing-the-data)set
* [Sharing the](https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/#step-five-sharing-your-results) predictions





* Limitations of this work and Scope for Future Work

-The data is interesting, as it contains more feature and the volume is huge. It takes time for visualization of distribution plot of all features.

-I used various visualization plots which helped me to understand the data.

- Dropped the unnecessary columns which having the more zero values. It helps to avoid the Multicollinearity (statistical concept where several independent variables in a model are correlated.)

- I have used the 5-machine learning algorithm to make the prediction for Micro Credit Defaulter project

- We can also analyze the dataset by observing age, gender, gross monthly income, and tenure with the current employer, loan amount, and the tenor of loan, number of dependents, other income, and other deductions were important determinants of default.