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.

***MICRO-CREDIT DEFAULTER PREDICTION***



****

**SUBMITTED BY : ARUNPRASAD.K**

**GITHUB LINK : <https://github.com/arunprasad-k>**

# **CONTENT**

**ABSTRACT**

**GENERAL TERMS**

**KEY WORDS**

**1.INTRODUCTION**

**2.Analytical Problem Framing**

**3.PROCESS FLOW**

**4.BASIC STATISTICS ABOUT OUR DATASET**

**5.DATA CLEANING**

**6. EXPLORATORY DATA ANALYSIS**

**6.1 Uni variate Analysis - Categorical Features**

**6.2 Displacement plot**

**6.3. Bi-variate Analysis**

**7.METHODOLOGY**

**7.1Feature Engineering**

**7.1.1correlation**

**7.1.2 Looking for Outliers**

**7.1.3 Scaling Input**

**7.2 Machine Learning Models**

**7.2.1 Logistic Regression**

**7.2.2 Decision Tree**

**7.2.3 Random Forest**

**7.2.4 Support Vector Machine**

**7.2.5 k-nearest neighbors**

**7.2.6 Gradient boosting**

**8. MODEL EVALUATION**

**8.1 Confusion Matrix**

**8.1.1Confusion Matrix for Logistic Regression**

**8.1.2Confusion Matrix for Decision Tree**

**8.1.3Confusion Matrix for SVC**

**8.1.4Confusion Matrix for Random Forest**

**9.PREDICTION**

**9.1 Hyper Tuning using GridSearchCV**

**9 .2Further Evaluation**

**9.2.1 Cross-validation**

**9.2 .2 Auc\_roc\_Curve**

**10. CONCLUSION**

**11.FUTURE WORK**

**12. REFERENCE**

**Abstract**

Micro-credit is the extension of very small loans (microloans) to impoverished borrowers who typically lack [collateral](https://en.wikipedia.org/wiki/Collateral_(finance)" \o "Collateral (finance)), steady employment, or a verifiable [credit history](https://en.wikipedia.org/wiki/Credit_history" \o "Credit history). It is designed to support entrepreneurship and alleviate poverty. Many recipients are illiterate, and therefore unable to complete paperwork required to get conventional loans. Micro-credit is part of [microfinance](https://en.wikipedia.org/wiki/Microfinance" \o "Microfinance), which provides a wider range of financial services, especially savings accounts.

**General Terms**

Data Analytics, Exploratory Data Analytics, Machine Learning, Model Evaluation, Data Science.

**Keywords**

Data mining, Logistic Regression, Random Forest, Feature Engineering, confusion Matrix,F1 score, auc roc curve

1. **Introduction**

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.

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.

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.

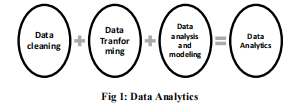
They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback 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).

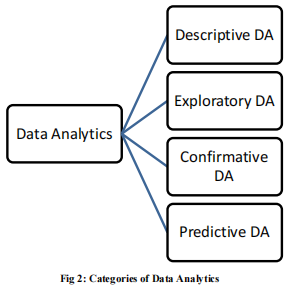
**Objective:**

Our objective is to 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.

1. **Analytical Problem Framing**

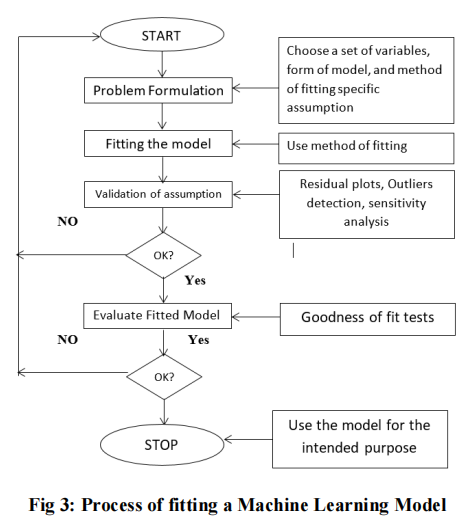
**DATA ANALYTICS AND ITS CATEGORIES**

****

****

1. **PROCESS FLOW**

There is a step by step approach to choose a particular model for the current problem. We need to decide whether a particular machine learning model is suitable for our problem or not. Here we can see process flow being followed.

****

**4)BASIC STATISTICS ABOUT OUR DATASET**

**Data formats**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 209593 non-null int64

1 label 209593 non-null int64

2 msisdn 209593 non-null object

3 aon 209593 non-null float64

4 daily\_decr30 209593 non-null float64

5 daily\_decr90 209593 non-null float64

6 rental30 209593 non-null float64

7 rental90 209593 non-null float64

8 last\_rech\_date\_ma 209593 non-null float64

9 last\_rech\_date\_da 209593 non-null float64

10 last\_rech\_amt\_ma 209593 non-null int64

11 cnt\_ma\_rech30 209593 non-null int64

12 fr\_ma\_rech30 209593 non-null float64

13 sumamnt\_ma\_rech30 209593 non-null float64

14 medianamnt\_ma\_rech30 209593 non-null float64

15 medianmarechprebal30 209593 non-null float64

16 cnt\_ma\_rech90 209593 non-null int64

17 fr\_ma\_rech90 209593 non-null int64

18 sumamnt\_ma\_rech90 209593 non-null int64

19 medianamnt\_ma\_rech90 209593 non-null float64

20 medianmarechprebal90 209593 non-null float64

21 cnt\_da\_rech30 209593 non-null float64

22 fr\_da\_rech30 209593 non-null float64

23 cnt\_da\_rech90 209593 non-null int64

24 fr\_da\_rech90 209593 non-null int64

25 cnt\_loans30 209593 non-null int64

26 amnt\_loans30 209593 non-null int64

27 maxamnt\_loans30 209593 non-null float64

28 medianamnt\_loans30 209593 non-null float64

29 cnt\_loans90 209593 non-null float64

30 amnt\_loans90 209593 non-null int64

31 maxamnt\_loans90 209593 non-null int64

32 medianamnt\_loans90 209593 non-null float64

33 payback30 209593 non-null float64

34 payback90 209593 non-null float64

35 pcircle 209593 non-null object

36 pdate 209593 non-null object

###### **Variable - Definition**

* ****label** -** 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 Unsure of given definition
* ****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 Unsure of given definition
* ****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 There are only two options: 5 & 10 Rs., for which the user needs to pay back 6 & 12 Rs. respectively
* ****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** - A**verage payback time in days over last 30 days
* ****payback90** -** Average payback time in days over last 90 days
* ****pcircle** -** telecom circle
* ****pdate** -** date

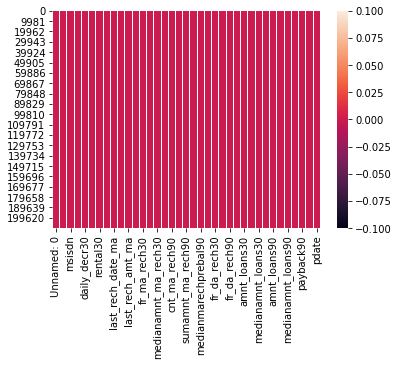
**Data Dimensions**

The data provided consistof **209593 rows and 37 columns**

Now let us explore our data set by knowing the influence of each attribute on the column label. We will create histograms, Bar plots to achieve this.

**5.DATA CLEANING**

Before applying any type of data analytics on the data set, the data should be first cleaned.Here we find there are no missing values in the data set which needed to be handled in case if any present**.**

****

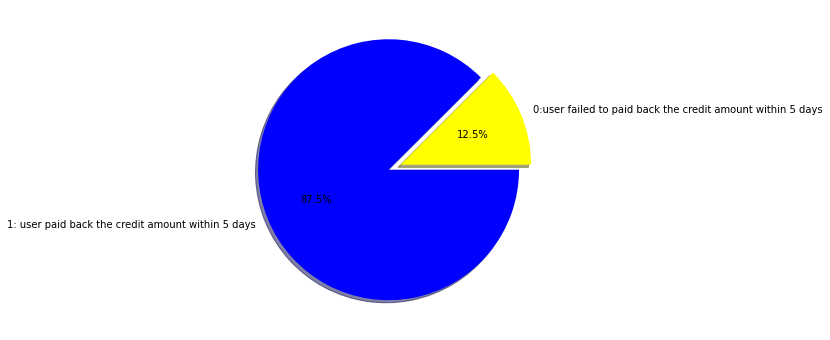
**6. EXPLORATORY DATA ANALYSIS**

We are going to perform exploratory data analysis for our problem in the first stage. In exploratory data analysis data set is explored to figure out the features which would influence the survival rate. The data is deeply analyzed by finding a relationship between each attribute and our target label

### **6.1 Uni variate Analysis - Categorical Features**

Now let's find the Unique values of all the categorical features we have , starting with the targeted variable.

**Column Label**

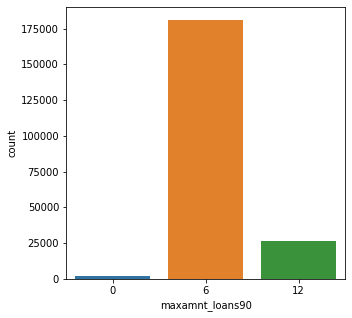
****

Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}

The above pie plot shows

* that the number of users who pay back the credit within 5 days i.e [1] is around 80%
* that the number of users who failed to pay back the credit within 5 days i.e [0] is around 20%

**Column maxamnt\_loans90:**

****

**Plot Insight:**

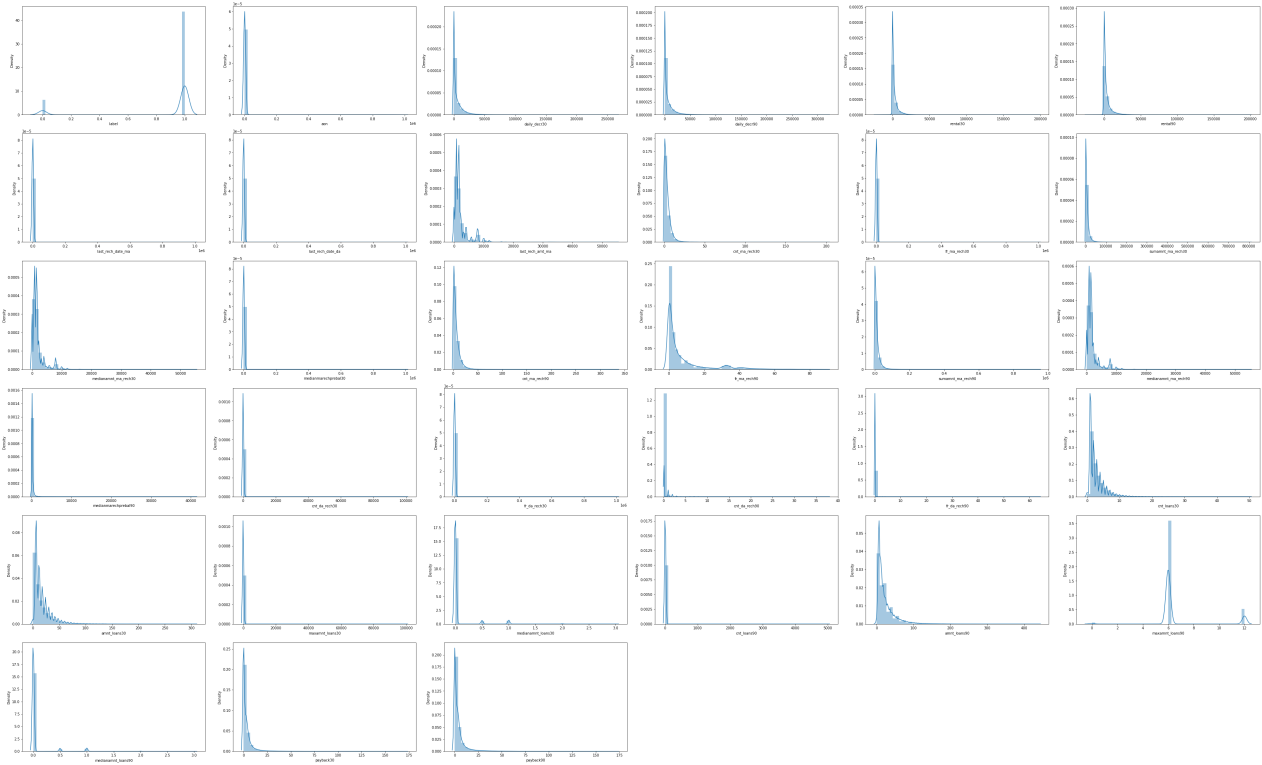
"maxamnt\_loans90"- maximum amount of loan taken by the user in last 90 days

1. users who took loan of 5 Indonesian Rupiah is more( around 175000 users)

users who took loan of 10 Indonesian Rupiah is less compared to users who took 5 Indonesian Rupiah( around 175000 users)

1. users who didn't took any loan were very low

**6.2 Displacement plot**

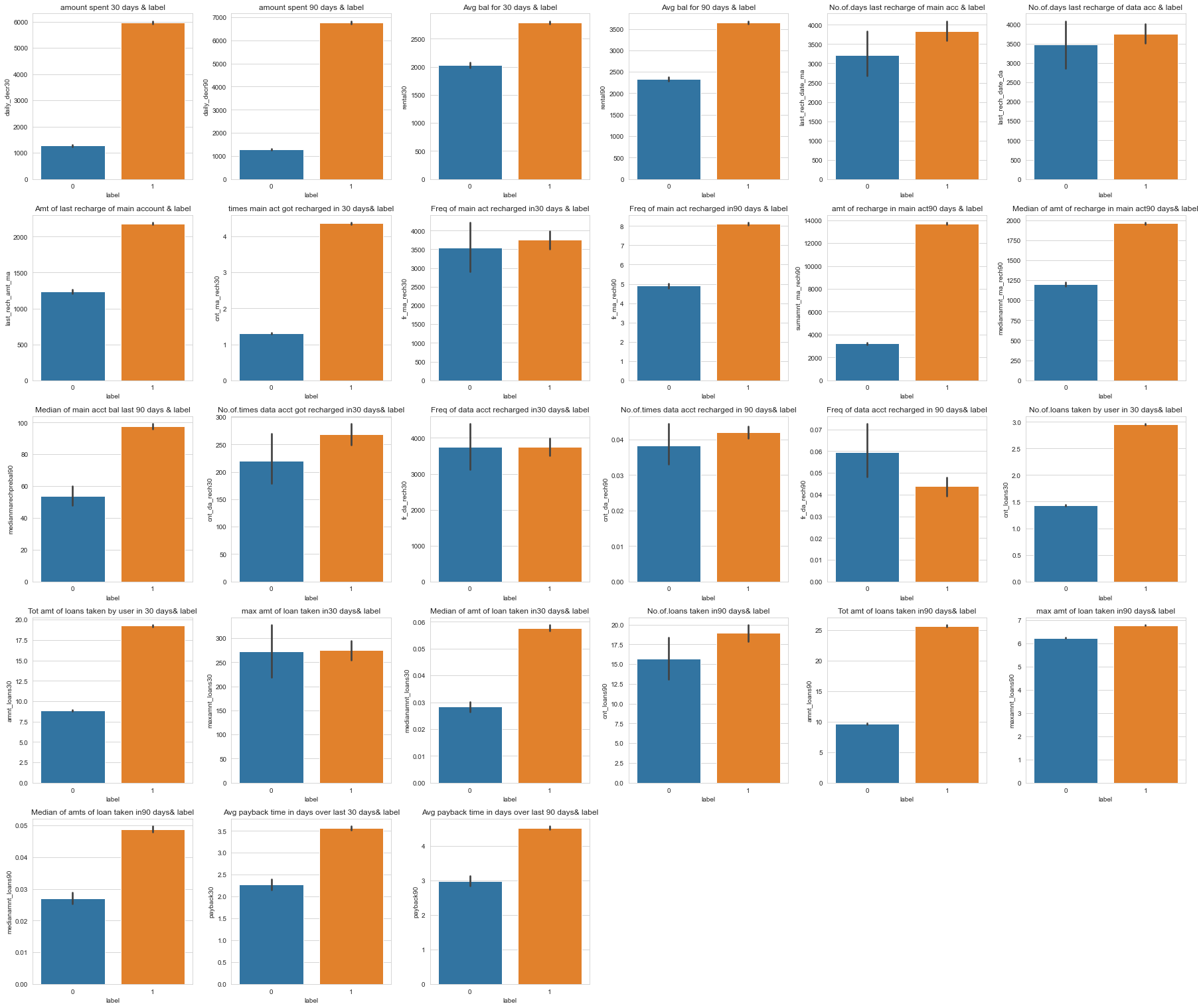
****

From the displacement graph we infer that all the columns shows Right Skewed Distribution. Hence all the attributes must have Positive skewness.

# **6.3. Bi-variate Analysis**

Finding a relationship between each attribute and our target column label.

**Attributes verses Label(Target)**

****

**Plot Insight:**

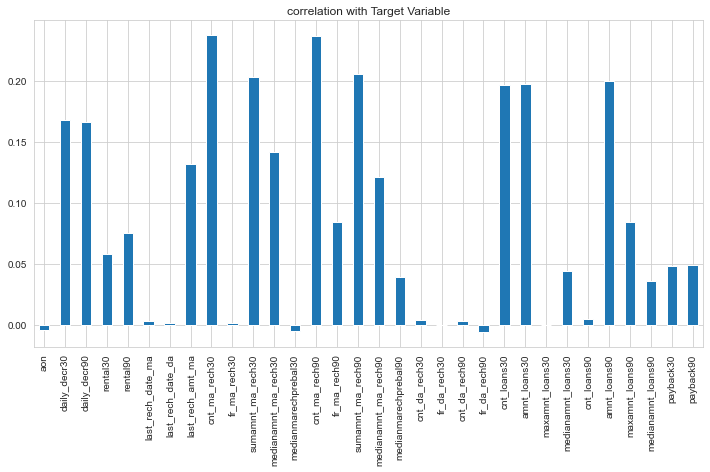
* **label vs daily\_decr30** - user who paid back the credit amount ,spend more amount to recharge their main account for a month in total when compared with the user who failed to pay back the loan amount
* **label vs daily\_decr90** - user who paid back the credit amount ,spend more amount to recharge their main account for a three months in total when compared with the user who failed to pay back the loan amount
* **label vs rental30** - user who paid back the credit amount used to keep more balance in their main account(avg around 3700) for a months when compared with the user who failed to pay back the loan amount(avg around 2000)
* l**abel vs rental90** - user who paid back the credit amount used to keep more balance in their main account(avg around 3600) for 3 months when compared with the user who failed to pay back the loan amount(avg around 2250)
* **label vs last\_rech\_date\_ma** -user who repay the loan used to recharge their mobile main account balance often(in total 3750)
* l**abel vs last\_rech\_date\_da** -user who repay the loan used to recharge their mobile data account balance(in total 3750)and user who repay the loan used to recharge their mobile data account balance(in total 3750)
* **label vs last\_rech\_amt\_ma**  - there is no much difference b/w the user who repay the loan and the user who failed to pay with respect to their recharge on their mobile data account balance
* **label vs cnt\_ma\_rech30** - user who repay their loan used to recharge main account atleast 4 times an average in a month whereas the user who failed to repay their the loan ,recharge 1.1 times an average in a month
* **label vs fr\_ma\_rech30** - The frequency of main account recharged main account for a month who repay their loan is around 3750 in approx whereas the frequency of main account recharged main account for a month who failed to repay their loan is around 3500 in approx
* **label vs fr\_ma\_rech90** - The frequency of main account recharged for 90 who repay their loan is around 8 in approx whereas the frequency of main account recharged for a month who failed to repay their loan is around 4.9 in approx
* l**abel vs sumamnt\_ma\_rech90** - Total amount of recharge in main account for 3 month by users who repay the loan amount amount is around 14,000(approx) and Total amount of recharge in main account for 3 month by users who failed to repay the loan amount amount is around 3,000(approx)
* **label vs medianamnt\_ma\_rech90** - median of amount of recharge for 90 days by users who repay the loan amount amount is around 1900(approx) and median of amount of recharge for 90 days by users who failed to repay the loan amount amount is around 1200(approx)
* **label vs medianmarechprebal90** - median of recharge in main account for 90 days by users who repay the loan amount amount is around 90(approx) and median of recharge in main account for 90 days by users who failed to repay the loan amount amount is around 55(approx)
* **label vs cnt\_da\_rech30** - user who repay their loan used to recharge data account atleast 280 times an average in a month whereas the user who failed to repay their the loan ,recharge 220 times an average in a month
* **label vs fr\_da\_rech30** - The frequency of data account recharged for 30 days is around 3800 in approx by both the users who paid and failed to repay their loan
* **label vs cnt\_da\_rech90** - user who repay their loan used to recharge data account atleast 0.04 times an average in 3 month whereas the user who failed to repay their the loan ,recharge 0.03 times an average in 3 month
* **label vs fr\_da\_rech90** - The frequency of data account recharged for 90 days is around 0.04 in appox by the users who repay their loan and The frequency of data account recharged for 90 days is around 0.06 in appox by the users who failed repay their loan
* **label vs cnt\_loans30** - no.of loans taken by the user who repay is 3 in average whereas no.of loans taken by the user who failed to repay is 1.4 in average in one month
* **label vs amnt\_loans30** - total amount of loans taken by users who repay it is 19 indonesian rupee in average whereas total amount of loans taken by users who failed to repay it is 8.7 indonesian rupee in average in a month
* **label vs maxamnt\_loans30** - The maximum amount of loan taken by both the defaulter and the non-defaulter where almost equal i.e 270(approx) for a period of 1 month
* **label vs medianamnt\_loans30** - the median of amount of loan taken by the defaulter is around 0.03 whereas the median of amount of loan taken by the user who repaid their loan amount is around 0.06 for a period of 1 month
* **label vs cnt\_loans90** - no.of loans taken by the user who repay is 18.5 in average whereas no.of loans taken by the user who failed to repay is 15.5 in average for a period of 3 month
* **label vs amnt\_loans90** - total amount of loans taken by users who repay it is 25 indonesian rupee in average whereas total amount of loans taken by users who failed to repay it is 10 indonesian rupee in average for a period of 3 month
* **label vs maxamnt\_loans90** - The maximum amount of loan taken by the defaulter is 6.8 and the non-defaulter is 6.2 for a period of 3 month
* **label vs medianamnt\_loans90** - the median of amount of loan taken by the defaulter is around 0.05 whereas the median of amount of loan taken by the user who repaid their loan amount is around 0.03 for a period of 3 month
* **label vs payback30** - the defaulter takes 23 time to pay back the loan as an average whereas the non-defaulter takes 36 times to repay the loan in a period of one month
* **label vs payback90** - the defaulter takes 3 time to pay back the loan as an average whereas the non-defaulter takes 4.5 times as an average to repay the loan in a period of 3 month

**7.METHODOLOGY**

**7.1Feature Engineering**

Feature engineering is the most important part of data analytics process. It deals with, selecting the features that are used in training and making predictions. In feature engineering the domain knowledge is used to find features in the dataset which are helpful in building machine learning model. It helps in understanding the dataset in terms of modeling. A bad feature selection may lead to less accurate or poor predictive model. The accuracy and the predictive power depend on the choice of correct features. It filters out all the unused or redundant features.

**7.1.1correlation**

****

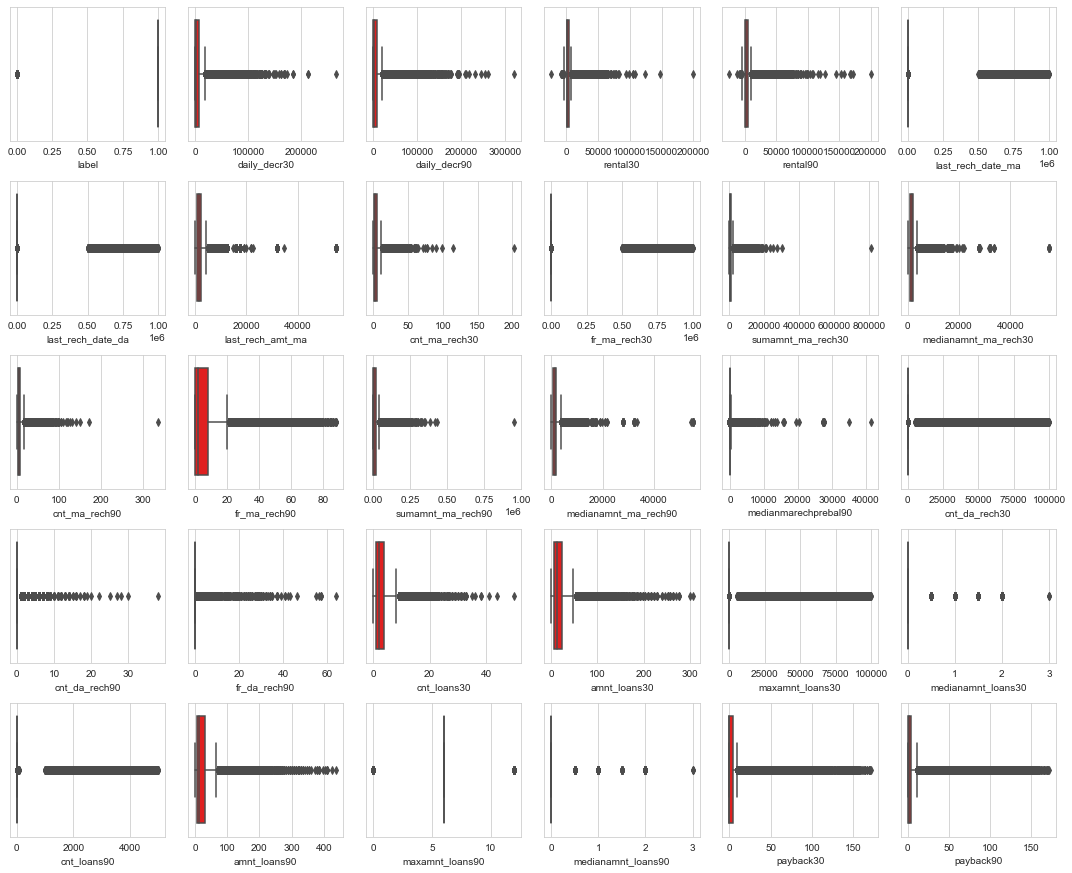
Based on the exploratory analysis above, following features are used i.e 'daily\_decr30', 'daily\_decr90', 'rental30', 'rental90', 'last\_rech\_date\_ma', 'last\_rech\_date\_da','last\_rech\_amt\_ma', 'cnt\_ma\_rech30', 'fr\_ma\_rech30', 'sumamnt\_ma\_rech30', 'medianamnt\_ma\_rech30','cnt\_ma\_rech90', 'fr\_ma\_rech90', 'sumamnt\_ma\_rech90', 'medianamnt\_ma\_rech90', 'medianmarechprebal90', 'cnt\_da\_rech30','fr\_da\_rech30', 'cnt\_da\_rech90', 'cnt\_loans30','amnt\_loans30', 'maxamnt\_loans30', 'medianamnt\_loans30', 'cnt\_loans90', 'amnt\_loans90', 'maxamnt\_loans90', 'medianamnt\_loans90', 'payback30', 'payback90'.

**.**

Label column is chosen as response column. These features are selected because their values have an impact on the prediction of loan defaulters. These features will be the value of “y” in the bar-plots as shown above. If wrong features where selected then even the good algorithm may produce the bad predictions. Therefore, feature engineering acts like a backbone in building an accurate predictive model.

**7.1.2 Looking for Outliers**

The difference between a good and an average machine learning model is often its ability to clean data. One of the biggest challenges in data cleaning is the identification and treatment of outliers. In simple terms, outliers are observations that are significantly different from other data points. Even the best machine learning algorithms will underperform if outliers are not cleaned from the data because outliers can adversely affect the training process of a machine learning algorithm, resulting in a loss of accuracy.

****

# **Percentage of Data Loss**

**loss\_of\_data=(209593-164713)/209593\*100**

**loss\_of\_data=21.41%**

* when we try to remove outliers it will results in loss of data around 22%. so it is not best practice to remove outliers with high percentage loss of data.
* From the document Provided we know that data is expensive and we cannot lose more than 7-8% of the data.
* so proceeding without outlier removal.

**7.1.3 Scaling Input**

Many machine learning algorithms perform better when numerical input variables are scaled to a standard range. This includes algorithms that use a weighted sum of the input, like logical regression, and algorithms that use distance measures, like k-nearest neighbors.

**7.2 Machine Learning Models**

Various machine learning models are implemented to validate and predict the loan defaulters.

**7.2.1 Logistic Regression**

Logistic regression is the technique which works best when dependent variable is dichotomous (binary or categorical). The data description and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables is done with the help of logistic regression. It is used to solve binary classification problem, some of the real life examples are spam detection- predicting if an email is spam or not, health-Predicting if a given mass of tissue is benign or malignant, marketing- predicting if a given user will buy an insurance product or not.

**7.2.2 Decision Tree**

Decision tree is a supervised learning algorithm. This is generally used in problems based on classification. It is suitable for both categorical and continuous input and output variables. Each root node represents a single input variable (x) and a split point on that variable. The dependent variable (y) is present at leaf nodes. For example: Suppose there are two independent variables, i.e. input variables (x) which are height in centimeter and weight in kilograms and the task to find gender of person based on the given data. (Hypothetical example, for demonstration purpose only).

There are two types of decision tree based on the type of target variable.

**1. Categorical Variable Decision Tree:** The tree in which target variables have categorical values.

**2. Continuous Variable Decision Tree:** The tree in which the target variable has continuous values.

**7.2.3 Random Forest**

Random forest algorithm is supervised classification algorithm. The algorithm basically makes forest with large number of trees. The higher the number of trees in the forest gives the higher accuracy results. Random forest algorithm can be used for both classification and regression problems. For instance, it will take random samples of 100 observation and 5 randomly chosen initial variables to build a model. The same process is repeated a number of times, then the final prediction is made according to the observations. Final prediction is a function (mean) of each prediction.

**7.2.4 Support Vector Machine**

Support Vector Machine (SVM) falls in supervised machine learning algorithm. This algorithm is used to solve both classification and regression problems. The classification is performed by constructing hyper planes in a multidimensional space that separates cases of different class labels. For categorical data variables a dummy variable is created with values as either 0 or 1. So, a categorical dependent variable consisting three levels, say (A, B, C) can be represented by a set of three dummy variables:

**A: {1, 0, 0}; B: {0, 1, 0}; C: {0, 0, 1}**

**7.2.5 k-nearest neighbors**

In [statistics](https://en.wikipedia.org/wiki/Statistics" \o "Statistics), the k-nearest neighbors algorithm (k-NN) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics" \o "Non-parametric statistics) [classification](https://en.wikipedia.org/wiki/Classification" \o "Classification) method first developed by [Evelyn Fix](https://en.wikipedia.org/wiki/Evelyn_Fix" \o "Evelyn Fix) and [Joseph Hodges](https://en.wikipedia.org/wiki/Joseph_Lawson_Hodges_Jr." \o "Joseph Lawson Hodges Jr.) in 1951 and later expanded by [Thomas Cover](https://en.wikipedia.org/wiki/Thomas_M._Cover" \o "Thomas M. Cover). It is used for [classification](https://en.wikipedia.org/wiki/Statistical_classification" \o "Statistical classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis" \o "Regression analysis). In both cases, the input consists of the k closest training examples in [data set](https://en.wikipedia.org/wiki/Data_set" \o "Data set). The output depends on whether k-NN is used for classification or regression:

**In k-NN classification,** the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive [integer](https://en.wikipedia.org/wiki/Integer" \o "Integer), typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

**In k-NN regression,** the output is the property value for the object. This value is the average of the values of k nearest neighbors.

**7.2.6 Gradient boosting**

Gradient boosting is a [machine learning](https://en.wikipedia.org/wiki/Machine_learning" \o "Machine learning) technique for [regression](https://en.wikipedia.org/wiki/Regression_(machine_learning)" \o "Regression (machine learning)) and [classification](https://en.wikipedia.org/wiki/Classification_(machine_learning)" \o "Classification (machine learning)) problems, which produces a prediction model in the form of an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning" \o "Ensemble learning) of weak prediction models, typically [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning" \o "Decision tree learning).When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms [random forest](https://en.wikipedia.org/wiki/Random_forest" \o "Random forest).It builds the model in a stage-wise fashion like other [boosting](https://en.wikipedia.org/wiki/Boosting_(machine_learning)" \o "Boosting (machine learning)) methods do, and it generalizes them by allowing optimization of an arbitrary [differentiable](https://en.wikipedia.org/wiki/Differentiable_function" \o "Differentiable function) [loss function](https://en.wikipedia.org/wiki/Loss_function" \o "Loss function).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model**  **\**  **score** | **Logistic**  **regression** | **KNeigborsClassifier** | **DecisionTree**  **Classifier** | **SVC** | **Random**  **Forest**  **Classifier** | **Gradient**  **Boosting**  **Classifier** |
| **Training accuracy** | 87.66 | 91.73 | 99.72 | 87.79 | 99.72 | 90.98 |
| **Testing accuracy** | 87.53 | 89.14 | 86.77 | 87.65 | 91.23 | 90.81 |

**8. MODEL EVALUATION**

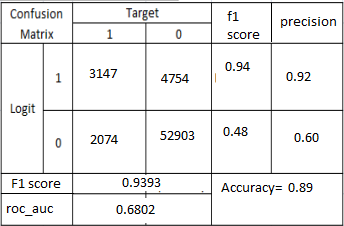
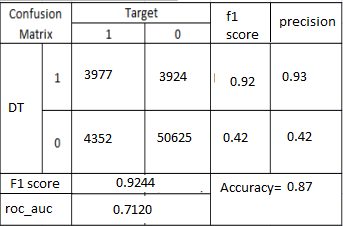
The accuracy of the model is evaluated using “confusion matrix”, ”classification report”, ”F1 score”,”roc\_auc\_score”. A confusion matrix is a table layout that allows to visualize the correctness and the performance of an algorithm.

**8.1 Confusion Matrix**

A confusion matrix is a method to verify how accurately the classification model works. It gives the actual number of predictions which were correct or incorrect when compared to the actual result of the data. The matrix is of the order N\*N, here N is the number of values. Performance of such models is commonly evaluated using the data in the matrix.

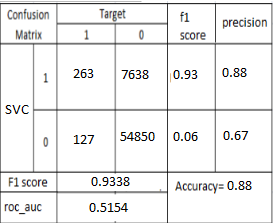
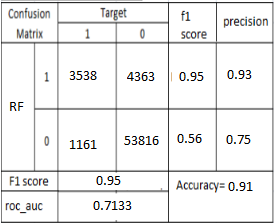
**8.1.1Confusion Matrix for 8.1.2Confusion Matrix for**

**Logistic Regression Decision Tree**

** **

**8.1.3Confusion Matrix for 8.1.4Confusion Matrix for**

**SVC Random Forest**

** **

**9.PREDICTION**

Since we have evaluated all models by using confusion matrix we will predict by using model which has highest accuracy.Here we can choose Random Forest models to predict the micro credit defaulter

**9.1 Hyper Tuning using GridSearchCV**

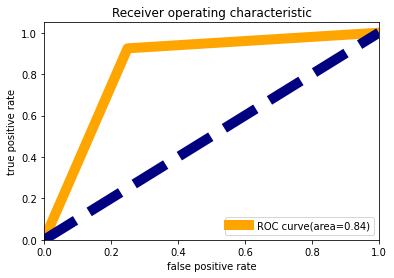
Hyperparameters are crucial as they control the overall behavior of a machine learning model. The ultimate goal is to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results. Here we first find the best parameters for the Random Forest Model and indulge it to the model to improve our predicting accuracy. There are several ways for finding the parameters, here we use the most powerful and most commonly used method named GridSearchCV.

**9 .2Further Evaluation**

**9.2.1 Cross-validation**

The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

**9.2 .2 Auc\_roc\_Curve**

****

**10. CONCLUSION**

Data cleaning is the first step while performing data analysis. Exploratory data analytics helps one to understand the dataset and the dependency among the attributes. EDA is used to figure out the relationship between the features of the dataset.This is done by using various graphical techniques. The one used above barplot ,countplot and heatmaps. By applying EDA some conclusions are drawn and facts are found. There is high influence of number of times main account got recharged in last 30 days .

In feature engineering the actual parameters to be used while designing the training model and prediction model is found out on the basis of exploratory data analytics process.

We used Machine Learning models 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

The confusion matrix gives the accuracy of all the models, the Random

Forest Classifier is proves to be best among all with an accuracy of **91.165431470466.** This means the predictive power of Random Forest Classifier in this dataset with the

chosen features is very high.

It is clearly stated that the accuracy of the models may vary when the choice of feature modelling is different. Ideally logistic regression and Random Forest Classifier are the models which give a good level of accuracy when it comes to classification problem.

**11.FUTURE WORK**

The future of microfinance seems bright as it will continue to expand beyond the traditional institutions. New actors like distribution networks and mobile operators will assist in offering financial products and services at cheaper costs to underprivileged and isolated population.

In future the idea can be extended by making more advanced graphical user interface with the help of newer libraries . An interactive page can be made, i.e. if the value of a attribute is changed on the scale the values corresponding to its graph (barplot or heatmap or countplot) will also change. We can also draw much focused conclusions by combining results we obtained.

**12. Reference**

* Dr. Neeraj Bhargava, Girja Sharma, Decision Tree Analysis on J48 Algorithm for Data Mining. Volume 3, Issue 6, June 2013.
* Machine Learning Benchmarks and Random Forest Regression, Segal, Mark R, 2004.
* https://en.wikipedia.org/wiki
* Proceedings of Student-Faculty ,PG Program in Data Science, Machine Learning and Neural Networks , DataTrained.com