

**Micro-Credit Defaulter**

Submitted by:

Amritesh Kumar

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shubam Yadav as well as Flip Robo Technologies who gave me the opportunity to do this project on **‘**Micro-Credit Defaulter’ & also helping me to gain in-depth knowledge of Machine Learning and DataScience to derive insights for organizational goals or to meet business needs.

Also, I have utilized a few external resources that helped me to complete this project. All the external resources that were used in creating this project are listed below:

<https://stackoverflow.com/questions>

<https://medium.com/>

<https://www.kaggle.com/>

<https://www.geeksforgeeks.org/>

<https://www.codegrepper.com/>

<https://www.analyticsvidhya.com/>

<https://towardsdatascience.com/>

<https://github.com/>

**INTRODUCTION**

Business Problem Framing

**Problem Overview**

Micro-Credit Defaulter

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).

The sample data is provided to us from our client database. It is hereby given to you for this exercise. 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.

Prediction:

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 payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

Conceptual Background of the Domain Problem

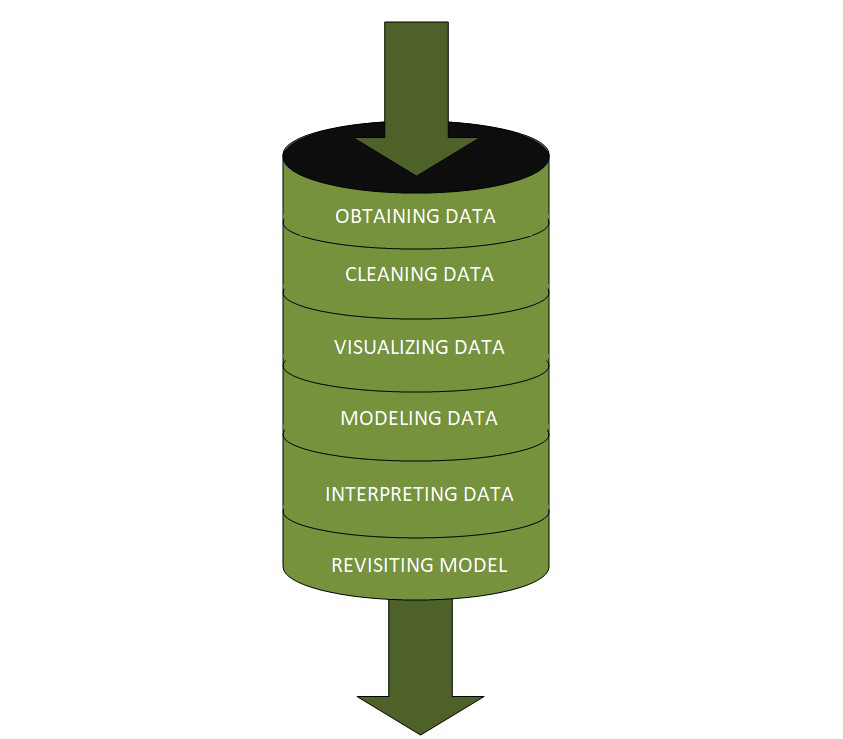
**MACHINE LEARNING AND DATA SCIENCE FOR BUSINESS:**

Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn from experience, make predictions and gradually improving its accuracy. It is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data science will increase, requires to assist in the identification of the most relevant business questions and subsequently the data to answer them. Following are the ways Data science can add value to Business :

* Empowering management and officers to make better decision
* Directing actions based on trends—which in turn help to define goals
* Challenging the staff to adopt best practices and focus on issues that matter
* Identifying opportunities
* Decision making with quantifiable, data-driven evidence
* Testing these decisions
* Identification and refining of target audiences

**DATASCIENCE PIPELINE:**

The data science pipeline is a collection of connected tasks that aims at delivering an insightful data science product or service to the business organization. The responsibilities include collecting, cleaning, exploring, modeling, interpreting the data, and other processes of the launching of the product. This final product can be used for to achieve Business Goals.



**Exploratory Data Analysis:**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

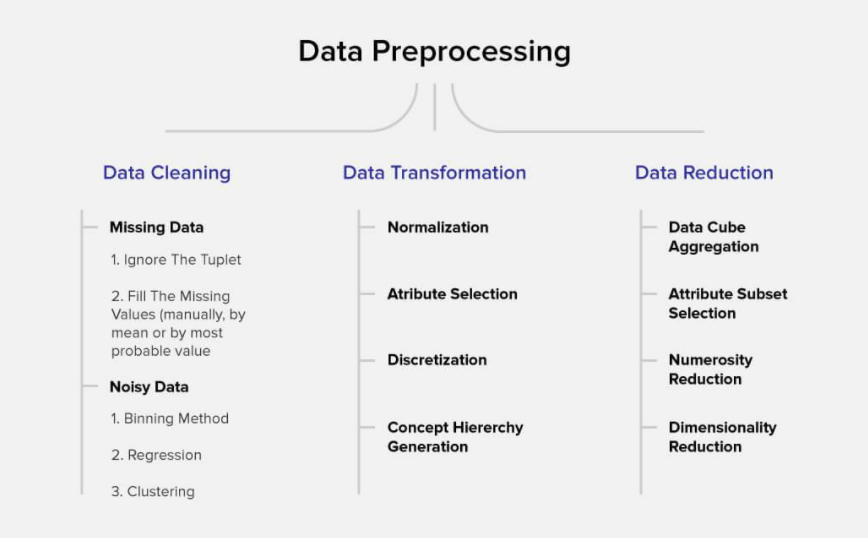
Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions

**TYPES OF EXPLORATORY DATA ANALYSIS:**

* Univariate Non-graphical
* Multivariate Non-graphical
* Univariate graphical
* Multivariate graphical

**DATA PRE-PROCESSING & FEATURE ENGINEERING:**

Preprocessing simply refers to perform series of operations to transform or change data. It is transformation applied to our data before feeding it to algorithm. When creating a machine learning project, and doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.



Data pre-processing is a very vital input to machine learning models, It is to prepare the raw data & make it suitable for efficient machine learning model. These are the methods of data preprocessing and we are going to use the required ones in our project.

**FEATURE ENGINEERING:**

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. As you may know, a “feature” is any measurable input that can be used in a predictive model.

Feature engineering**, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.** It can produce new features for both supervised and unsupervised learning, with the goal of**simplifying and speeding up data transformations**while also**enhancing model accuracy.**

**Feature Engineering Techniques for Machine Learning**

* **Imputation**
* **Handling Outliers**
* **Log Transform**
* **One-hot encoding/Label Encoding**
* **Scaling**

**Data Transformation:**

**Label Encoding:**

**As we mentioned above in library installation,** Label Encoder is used to encode labels by assigning them numbers. It is used to encode single or multiple columns. Thus, if the feature is color with values such as [‘white’, ‘red’, ‘black’, ‘blue’]., using Label Encoder may encode color string label as [0, 1, 2, 3]

**Handling Outliers:**

The most important phase in Feature Engineering is handling outliers because it ensures that our model is trained on accurate data which leads to accurate models. An outlier may occur due to the variability in the data. It may indicate an experimental error or heavy skewness in the data(heavy-tailed distribution). We have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

Below are some of the techniques of detecting outliers

* Boxplots
* Z-score
* Quantile method
* Percentile method

## Variance Inflation Factor (VIF)

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library, Let us import the package.

**SKEWNESS REMOVAL-(POWER-TRANSFORM):**

Key step prior to initiating Machine learning models, optimizing, scaling the data to provide it as a input to start the modelling.

A power transform will make the probability distribution of a variable more Gaussian. This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution. The log transform is a specific example of a family of transformations known as power transforms. The power\_transform library present in the Sklearn. Pre-processing package.

**MINMAX SCALER:**

MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

Before scaling we have to train test split the data.since we have to do skewness removal and scaling only on input data.

**TRAIN TEST SPLIT:**

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train\_test\_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train\_test\_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

**PERCENTILE METHOD:**

The IQR can then be calculated as the difference between the 75th and 25th percentiles. We can then calculate the cutoff for outliers as **1.5 times the IQR and subtract this cut-off from the** 25th percentile and add it to the 75th percentile to give the actual limits on the data. Use percentile-based approach. For Example, Data points that are **far from 99% percentile and less than 1 percentile** are considered an outlier.

Review of Literature

**ABSTRACT:**

Micro finance Solution has been seen very beneficial to unbanked poor families living in remote areas by providing them with Group, Agricultural, Individual business loans. We are working with a telecommunication client out of Indonesia, that is Telecom Industry (Indonesia). They understand the importance of communication & how it will affect a person’s life in this modern era. They are providing a plan in tie-up with MFI. i.e. to provide micro-credit on mobile balances which needs to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates this duration. Our purpose is to build a model which can enable them to predict the customers & decide on further investments & improvements with customer selection.

Motivation for the Problem Undertaken

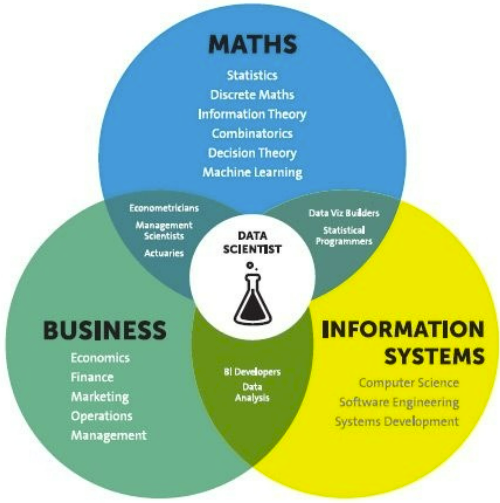
**Business Goal:**

We are provided with data from out client database with clear details on customer transactions done over such period. We need to build a model & understand the factors which defines defaulter or non-defaulter based on the criteria of repaying the loan amount in 5 days timeframe. This will help them to decide on further investment & improvements with customer selection.

**Analytical Problem Framing**

Mathematical/ Statistical /Analytical Modeling of the Problem

Mathematics, Statistics and Analytics are three of the most important concepts of Data Science. Data Science revolves around these three fields and draws their concepts to operate on the data.we will explore its practical usages in this field.  So let’s first explore how much these three are required for data science.



**Mathematical Modelling**

Mathematical models are important, selecting the right one to answer the business question can bring tremendous value to the organization. Machine Learning is a field that focuses on computers having the ability to learn/operate without being programmed to do so.

Mathematics is playing an essential role in the latest technologies like Machine Learning, Artificial Intelligence, Data Science and Deep Learning, etc., It is because every algorithm built in the latest technologies has a mathematical function behind it and aid in identifying patterns.

The understanding of various notions of Statistics and Probability Theory are key for the implementation of such algorithms in data science. Notions include: Regression, Maximum Likelihood Estimation, the understanding of distributions (Binomial, Bernoulli, Gaussian (Normal)) and Bayes’ Theorem.

The main reason for a greater significance of mathematics is because of its various concepts like: –

· Linear Algebra

· Probability

· Calculus

· Statistics

**Linear Algebra & Calculus**

Deep learning requires us to understand linear algebra & calculus, to understand how it works, for example forward propagation, backward propagation, parameters setting etc. For linear algebra, there are matrix operations (plus, minus, times, divide), scalar product, dot product, eigen-vectors and eigenvalues.

It is a branch of  Mathematics for studying systems of equations. it can be one, two, and multi-dimensional equations. it helps us to solve numerical data or relations between two or more variables by establishing relations or equations between them. for example,

here' one basic algebraic equation:

    y = a + bx + cx2

linear-algebra has a wide range of applications such as statics and matrices calculations, linear regression equations, descriptive statistics, graphic image vectors, Fourier series, graphs, and network establishment.

machine-learning algorithms like linear regression, logistic regression uses linear algebra to solve our target variables with given inputs/attributes or feature vectors given in the data set.

**Calculus**

 Calculus is used essentially in optimization techniques. Using calculus, you can carry out mathematical modeling of artificial neural networks and also increase their accuracy and performance. For calculus, the data scientist need to understand various differentiation (to second-order derivative), integration, partial differentiation.

**Differential Calculus**

  Differential Calculus studies the rate at which the quantities change. Derivates are most widely used for finding the maxima and minima of the functions. Derivates are used in optimization techniques where we have to find the minima in order to minimize the error function.

**Integral Calculus**

It is the mathematical study of the accumulation of quantities and for finding the area under the curve. Integrals are further divided into definite integrals and indefinite integrals.

**Probability**

The probability theory is very much helpful for making the prediction and Estimation.With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.

Probability is a very important mathematical concept for data science, used in validating hypothesis, bayes theorem and interpreting outputs in machine learning.

Bases on these we try to estimate various events, and the likelihood of the outcome. sometimes we wat graphical representations of probable outcomes which we call probability density functions or density curves.

Concepts of probability help us estimate expected value from given variables, to solve confusion matrix in classification algorithms, information entropy, evidence of particular attributes in naive Bayes classification, and even in statistics for hypothesis testings.

**Statistics**

A statistical model is a mathematical representation (or mathematical model) of observed data. When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically.

So the areas in statistics are simple statistics like measurement of centrality, distributions and different probability distributions (Weibull, Poisson etc), Baye’s Theorem

statistics is divided into two –

* Descriptive Statistics
* Inferential Statistics

#### **Descriptive Statistics**

Descriptive Statistics or summary statistics is used for describing the data. It deals with the quantitative summarization of data. This summarization is performed through graphs or numerical representations.

### Descriptive Statistics:

1) Mean, Median, Mode

2) IQR, percentiles

3) Std deviation and Variance

4) Normal Distribution

5) Z-statistics and T-statistics

6) correlation and linear regression

**Inferential Statistics**

It is the procedure of inferring or concluding from the data. Through inferential statistics, we make a conclusion about the larger population by running several tests and deductions from the smaller sample.

### Inferential Statistics:

1) Sampling distributions

2) confidence interval

3) chi-square test

4) Advanced regression

5) ANOVA

The mathematical concepts noted above are key in understanding/implementing the following Machine Learning techniques.

* Supervised learning, including regression and classification models.
* Unsupervised learning, including clustering algorithms and association rules.

### **Regression Models**

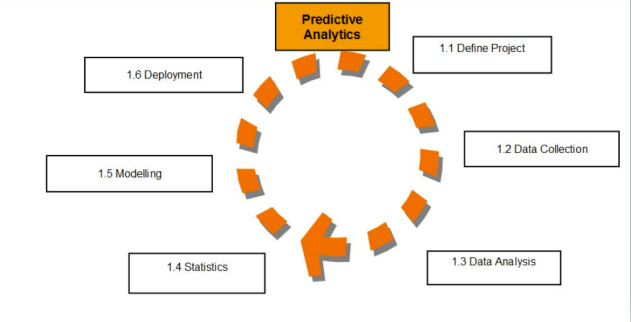
Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential [business decisions](https://www.northeastern.edu/graduate/blog/data-driven-decision-making/).

### **Classification Models**

**Classification** is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

**Analytical Models:**

An analytical model estimates or classifies data values by essentially drawing a line through data points. When applied to new data or records, a model can predict outcomes based on historical patterns.



.  An analytical model is quantitative in nature, and used to answer a specific question or make a specific design decision. Different analytical models are used to address different aspects of the system, such as its performance, reliability, or mass properties.Data analysis comes with the fundamental types of data analytics encounter in data science: Descriptive, Diagnostic, Predictive, and Prescriptive.

* Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.
* Descriptive analysis is often used when reviewing any past or present data. This is because raw data is difficult to consume and interpret, while the metrics offered by descriptive analysis are much more focused.
* The example of descriptive statistics or analytics is to calculate the mean, median mode, standard deviation, and similar kinds of statistical calculation on finance or sales data.
* Diagnostic analytics takes it a step further to uncover the reasoning behind certain results. Diagnostic analytics is usually performed using such techniques as data discovery, drill-down, data mining, and different type of bivariant data analysis like  correlations.etc.,
* Predictive Analytics is a **statistical method that utilizes algorithms and machine learning to identify trends in data and predict future behaviors**. Predictive Analytics can take both past and current data and offer predictions of what could happen in the future.
* Predictive models typically utilize variability in data to make the correct prediction and more variability of ingredient data that shows the relationship with what is possible to predict that united together into a prediction or valid score.
* Prescriptive analytics automatically synthesizes big data, mathematical sciences, business rules, algorithms, and machine learning to make predictions and then suggests decision options to take advantage of the predictions. Prescriptive means (optimization and simulation).

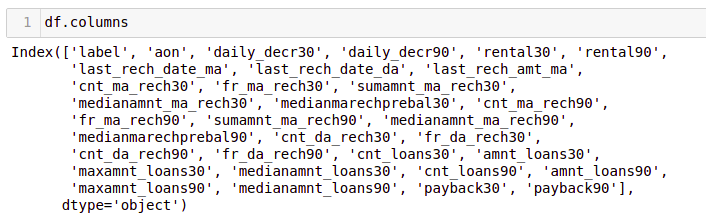
Data Sources and their formats

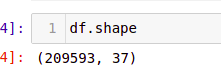
**Technical Requirements**:

* 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.
* For some features, there may be values which might not be realistic. We may have to observe them and treat them with a suitable explanation.
* We might come across outliers in some features which we need to handle as per our understanding. Keep in mind that data is expensive and we cannot lose more than 7-8% of the data.
* Data contains 209593 entries each having 37 variables.
* Data set doesn’t contains Null values. We treated them using the domain knowledge and our own understanding.
* Extensive EDA has been performed to gain relationships of important variable and labels.
* Data contains one numerical and all others as categorical variable. We handled them accordingly.
* We built Machine Learning models, applied regularization and determined the optimal values of Hyper Parameters.
* We found important features which affect the labels positively or negatively.
* The dataset is enclosed in notebook file.
* The dataset is provided to us by FlipRobo Technologies.And the dataset is in excel file format.

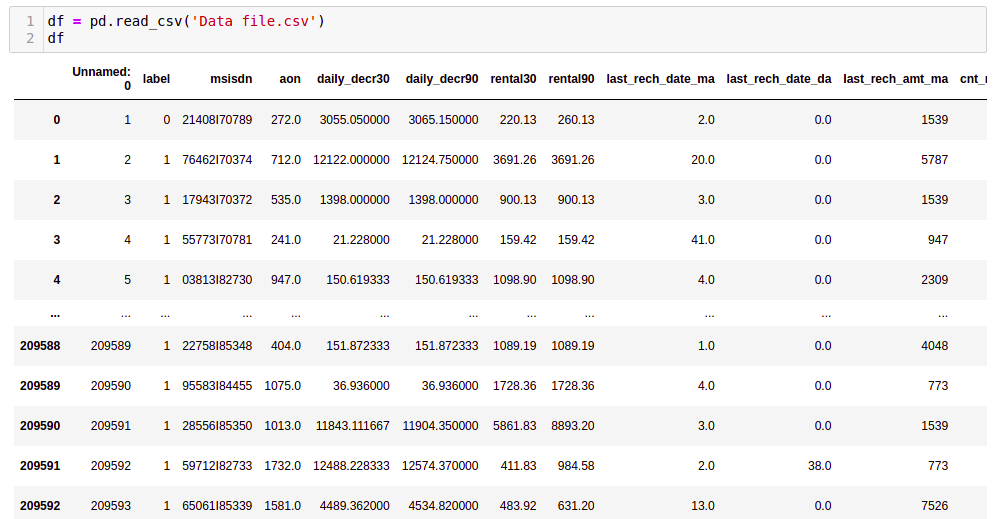
**Data Description:**

Following are the descriptions for data variables.





**DATA ACQUISITION**



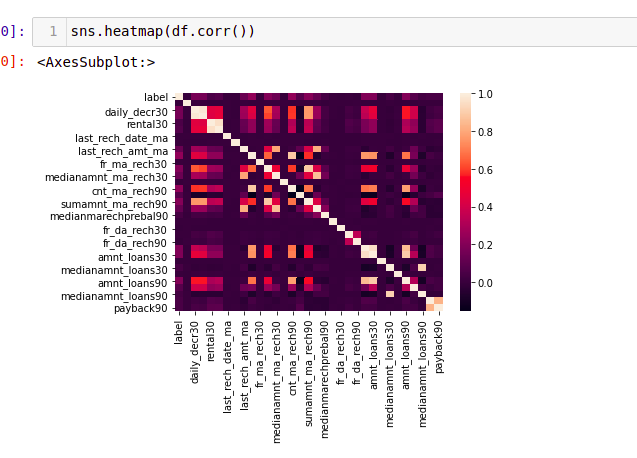
**FEATURE DESCRIPTION:**

Following are the features and their data type details,

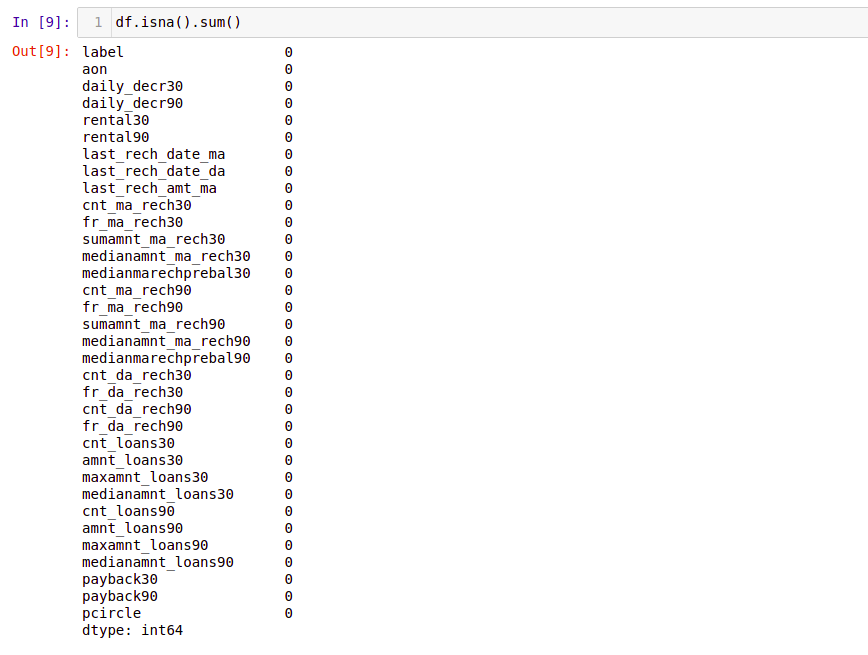
# 

# 

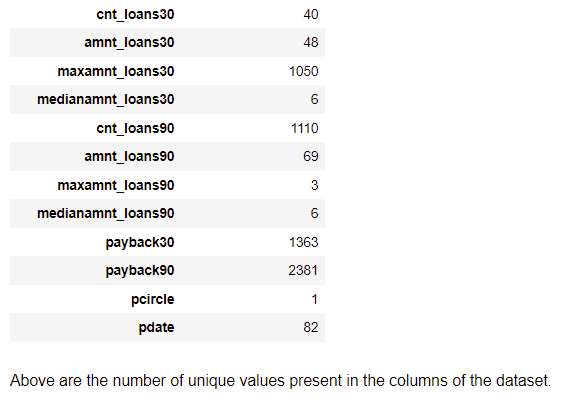
This info() method gives the information about the dataset which includes indexing type, column type, no-null values and memory usage.The dataset contains 3 different types of data, integer, object and float. There are 3 columns with object datatype.Here all the columns are numeric excecpt pcircle,pdate and msisdn.pcircle and msisdn features are object datatype and pdate feature is in datetime datatype.The dataset consist of 209593 rows and 36 columns. We will encode the object datatypes using appropriate encoding techniques before building machine learning models.

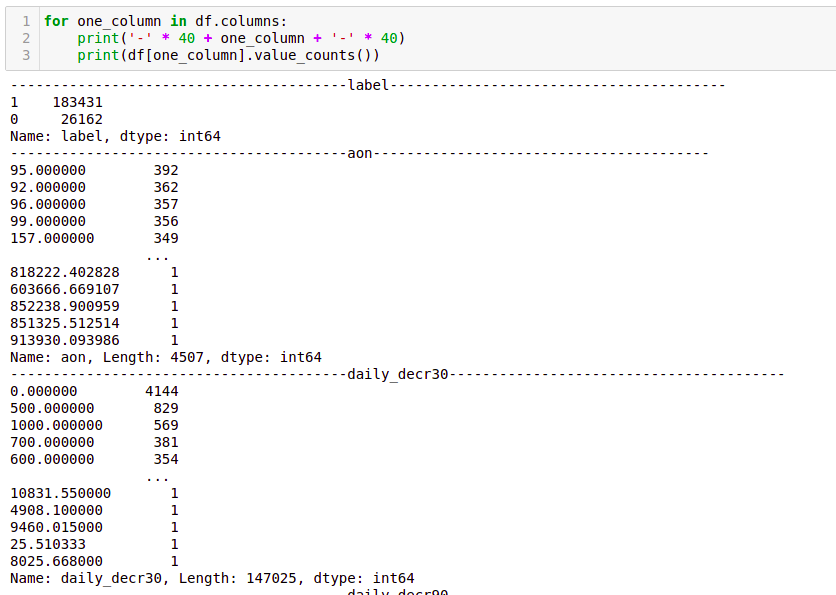


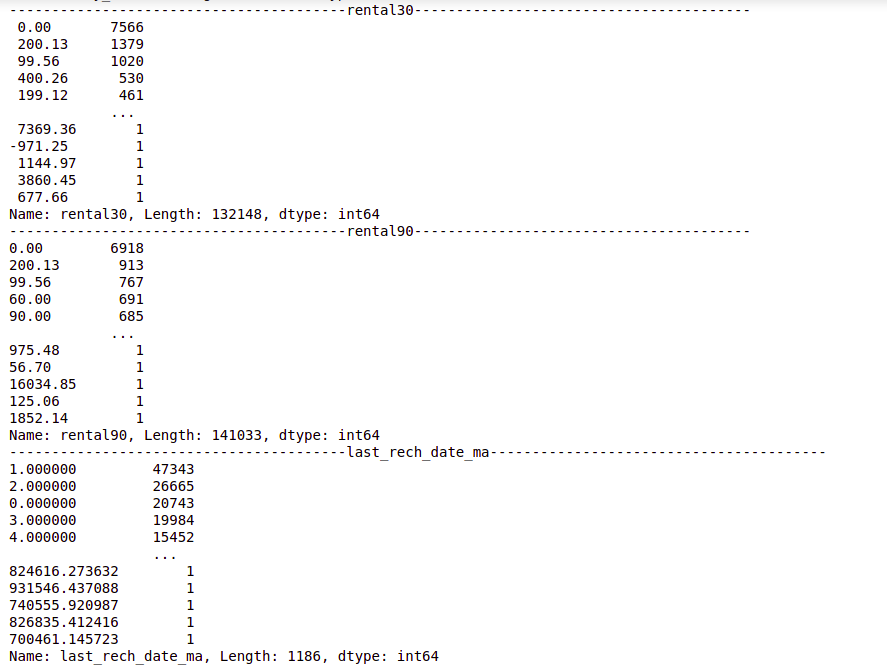
Exploratory Data Analysis and Preprocessing Done

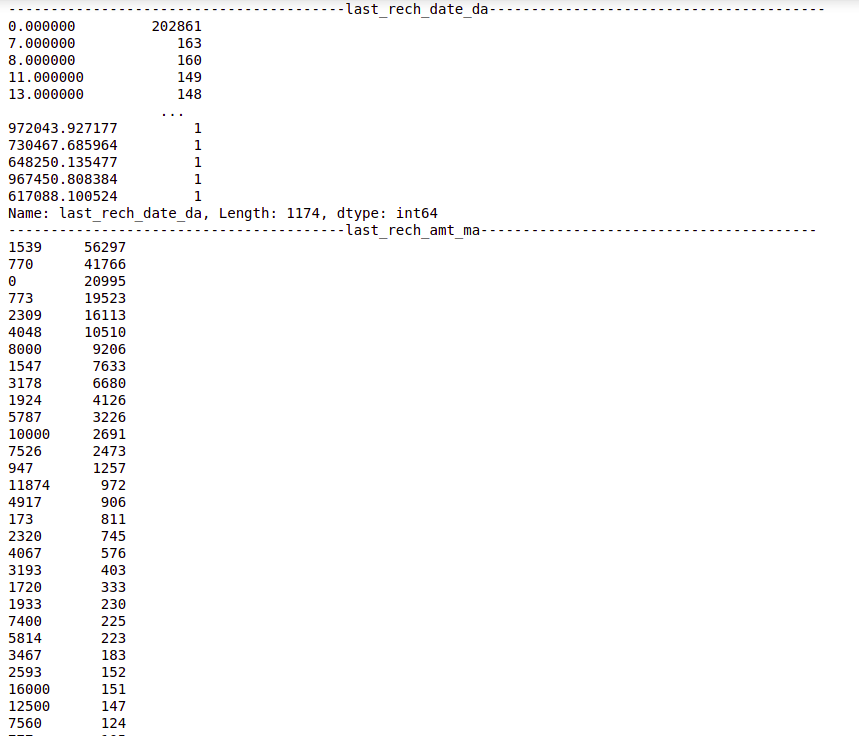


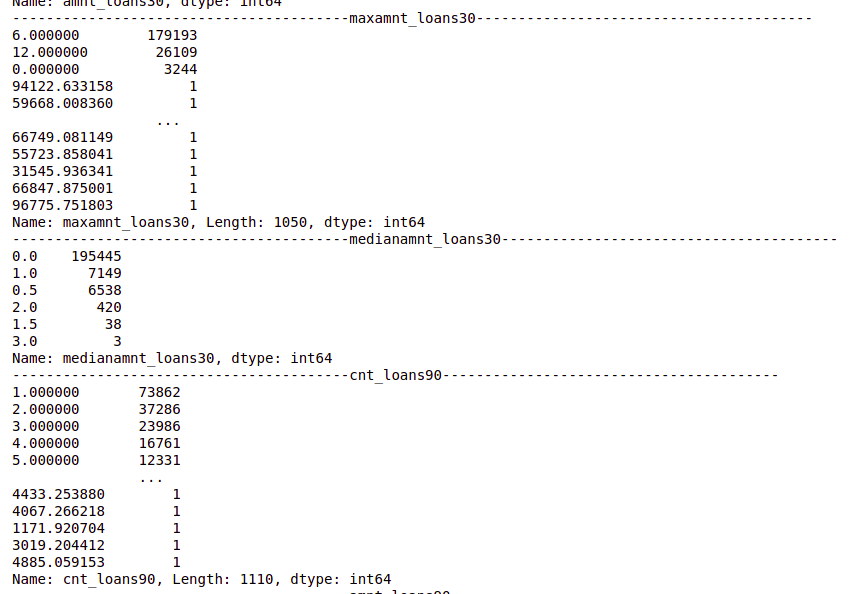
There are no null values in the dataset.



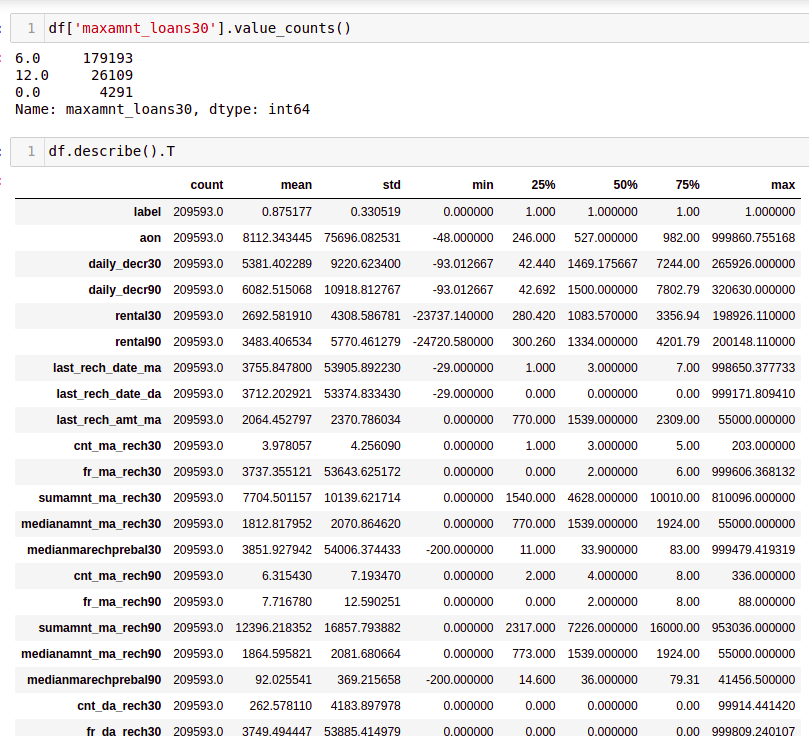












Data Inputs- Logic- Output Relationships

In Classification, the output variable must be **a discrete value**.

In classification, **inputs are divided into two or more classes**, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) or more of these classes. This is typically tackled in a supervised way.

In machine learning, classification refers to **a predictive modeling problem where a class label is predicted for a given example of input data**. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. The output variables are often called **labels or categories**. ... A classification can have real-valued or discrete input variables. A problem with two classes is often called a two-class or binary classification problem. A problem with more than two classes is often called a multi-class classification problem. A classification algorithm, in general, is a function that weighs the input features so that the output **separates one class into positive values and the other into negative values**. Classification is a data mining function that **assigns items in a collection to target categories or classes**. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

Classification analysis is a data analysis task within data-mining, **that identifies and assigns categories to a collection of data to allow for more accurate analysis**. ... Classification analysis can be used to question, make a decision, or predict behavior through the use of an algorithm.

DATA PREPROCESSING AND FEATURE ENGINEERING

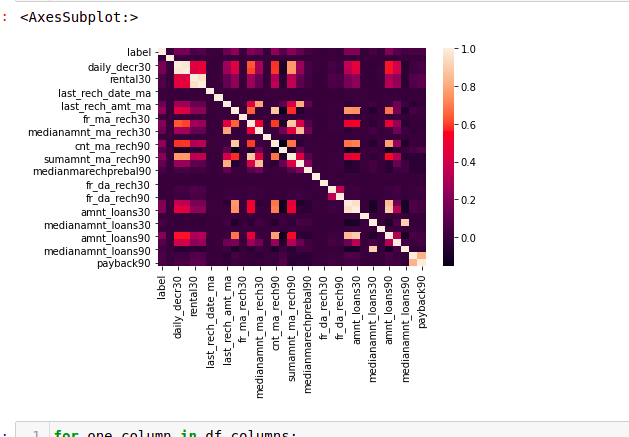
**Correlation with Heatmap:**

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect [negative correlation](https://www.investopedia.com/terms/n/negative-correlation.asp), while a correlation of 1.0 shows a perfect [positive correlation](https://www.investopedia.com/terms/p/positive-correlation.asp). A correlation of 0.0 shows no linear relationship between the movement of the two variables.Correlation statistics can be used in finance and investing. Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.

It can also be defined as the measure of dependence between two different variables. If there are multiple variables and the goal is to find correlation between all of these variables and store them using appropriate data structure, the **matrix data structure**is used. Such matrix is called as **correlation matrix.**

Correlation heatmap is graphical representation of **correlation matrix**representing correlation between different variables.

**For to do feature selection and make feature ready for the model building.we check correlation of variables using heatmap.And describe method for the census data set.**



These are the most positively correlated column with the target column 'label'

cnt\_ma\_rech30  
cnt\_ma\_rech90  
sumamnt\_ma\_rech90  
sumamnt\_ma\_rech30  
amnt\_loans90  
amnt\_loans30  
cnt\_loans30

**Correlation model:**

Graph depicts clearly the positive and negative correlation of each variables with target column, justifies the outcome outlined in Multivariate analysis, that higher the education higher the gain & vice-versa

Hardware and Software Requirements and Tools Used

**HARDWARE & Software Tools, Libraries and Packages Used:**

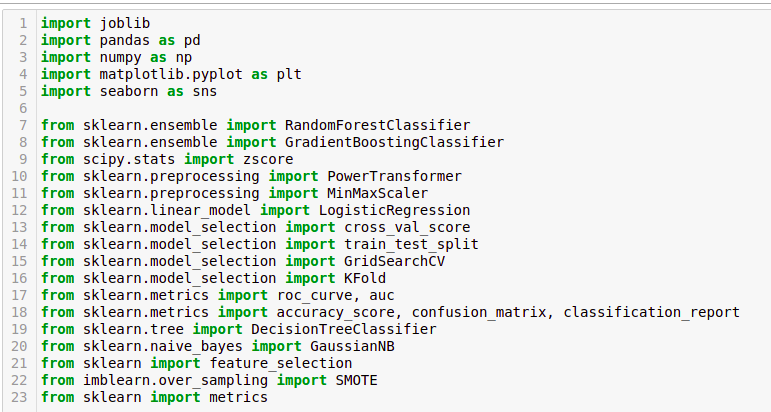
Hardware :Intel i7,RAM 16GB used.

Software: Jupyter Notebook (Anaconda 3)

Language: Python

Libraries:

1. Pandas
2. Numpy
3. Matplotlib
4. Seaborn
5. Sklean
6. Scipy
7. Statsmodels
8. Pip-Package install Manager



|  |  |  |
| --- | --- | --- |
| **Category** | **Tool** | **Function** |
| Data loading and analysis | Import pandas as pd | Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. Pandas is built on top of numpy. So, numpy gets some superpower with pandas. It offers data structures and operations for manipulating numerical tables and time series. |
| Import numpy as np | NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.It has Quantile method too for removing outliers. It is the fundamental package for scientific computing with Python |
| Data visualization | Import matplotlib.pyplot as plt | Matplotlib is a plotting library used for data visualization. |
| Import seaborn as sns | Seaborn is also a plotting library. It is more advanced than matplotlib but works with matplotlib |
| Scikit Learn Preprocessing Libraries | Sklearn.preprocessing | Package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.Has power transformer to remove skewness. In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate.  It has MinMaxScaler to scale the data. |
| Sklearn.preprocessing import LabelEncoder | Label Encoding in Python can be implemented using the Sklearn Library. Sklearn furnishes a very effective method for encoding the categories of categorical features into numeric values. Label encoder encodes labels with credit between 0 and n-1 classes where n is the number of diverse labels. |
| Import statistics | Import statsmodels.api as sm | From scipy import stats This module provides functions for calculating mathematical statistics of numeric (Real-valued) data. This library provides a number of common functions and types useful in statistics. It focus on high performance, numerical robustness, and use of good algorithms |

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library and we can import this library.

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train\_test\_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train\_test\_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

Many learning algorithms have been proposed. It is often valuable to assess the efficacy of an algorithm. In many cases, such assessment is relative, that is, evaluating which of several alternative algorithms is best suited to a specific application.

People even end up creating metrics that suit the application. In this article, we will see some of the most common metrics in a classification setting of a problem.

Choice of metrics influences how the performance of machine learning algorithms is measured and compared.

**The most commonly used Performance metrics for classification problem are as follows,**

* Accuracy.
* Confusion Matrix.
* Precision, Recall, and F1 score.
* ROC AUC.
* Log-loss.

Accuracy:

Accuracy is a good measure when the target variable classes in the data are nearly balanced. Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made. It is the simple ratio between the number of correctly classified points to the total number of points.

Confusion Matrix:

 It is used for Classification problem where the output can be of two or more types of classes. *For a sensible model, the principal diagonal element values will be high and the off-diagonal element values will be below i.e., TP, TN will be high.*

# Precision:

**Precision** is the fraction of the correctly classified instances from the total classified instances. Precision helps us understand how useful the results are.

Recall or Sensitivity:

**Recall** is the fraction of the correctly classified instances from the total classified instances.  Recall helps us understand how complete the results are.

F1 Score:

The F-score is often used in the field of information retrieval for measuring search, document classification, and query classification performance.

The F-score has been widely used in the natural language processing literature, such as the evaluation of named entity recognition and word segmentation

**Specificity:**

Specificity is the exact opposite of Recall.

Log Loss

Logarithmic loss (or log loss) measures the performance of a classification model where the prediction is a probability value between 0 and 1. Log loss increases as the predicted probability diverge from the actual label.

Lower the log-loss value, better are the predictions of the model.

# ROC AUC

# **A Receiver Operating Characteristic curve** or **ROC curve** is created by plotting the True Positive (TP) against the False Positive (FP) at various threshold settings. The ROC curve is generated by plotting the [cumulative distribution function](https://medium.com/towards-artificial-intelligence/pdf-cdf-in-machine-learning-5a10f7abeef6) of the True Positive in the y-axis versus the cumulative distribution function of the False Positive on the x-axis.The area under the ROC curve (ROC AUC) is the single-valued metric used for evaluating the performance.

# *The higher the AUC, the better the performance of the model at distinguishing between the classes.*

Grid search is used as an approach to hyper-parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. GridSearchCV helps us combine an estimator with a grid search preamble to tune hyper-parameters.

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows :

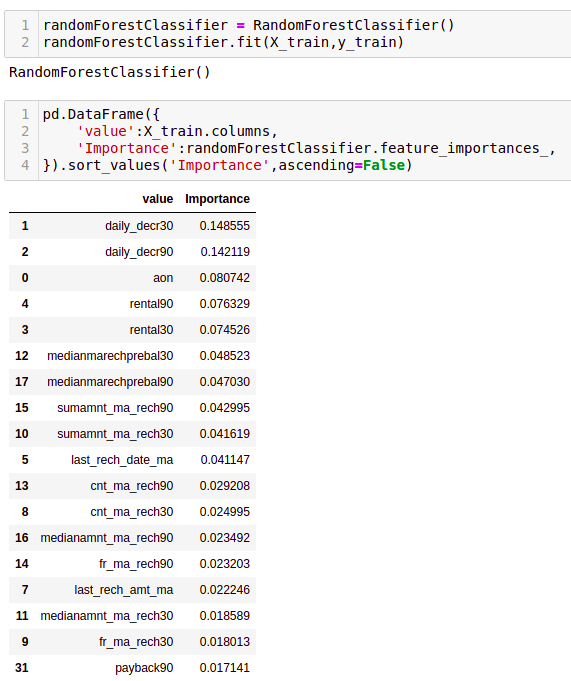
1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

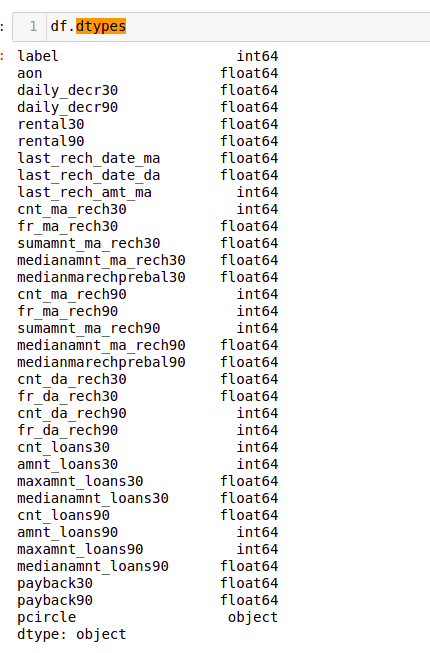
The most important use of PCA is to **represent a multivariate data table as smaller set of variables** (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables.

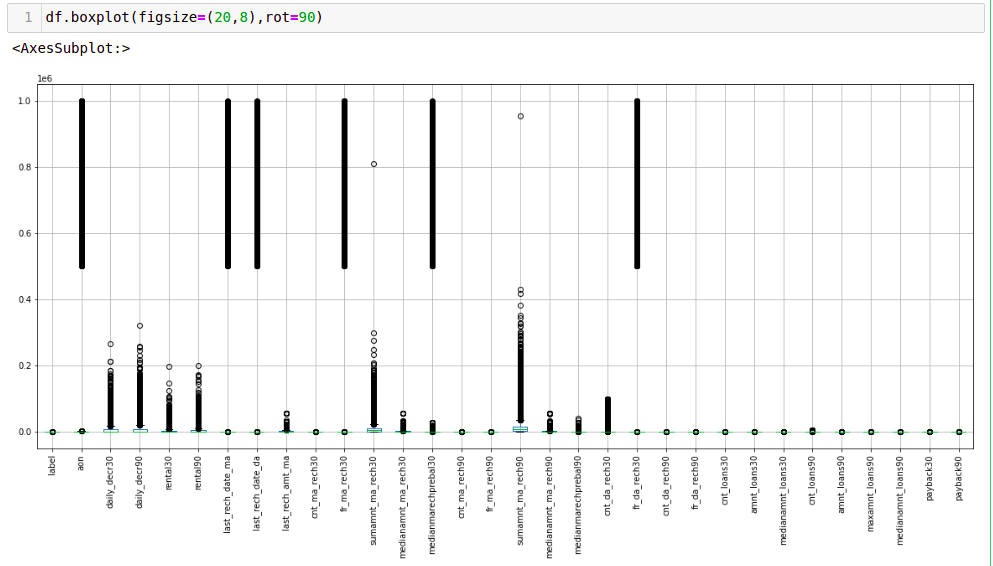
Also import all the required algorithms for classification purpose below.

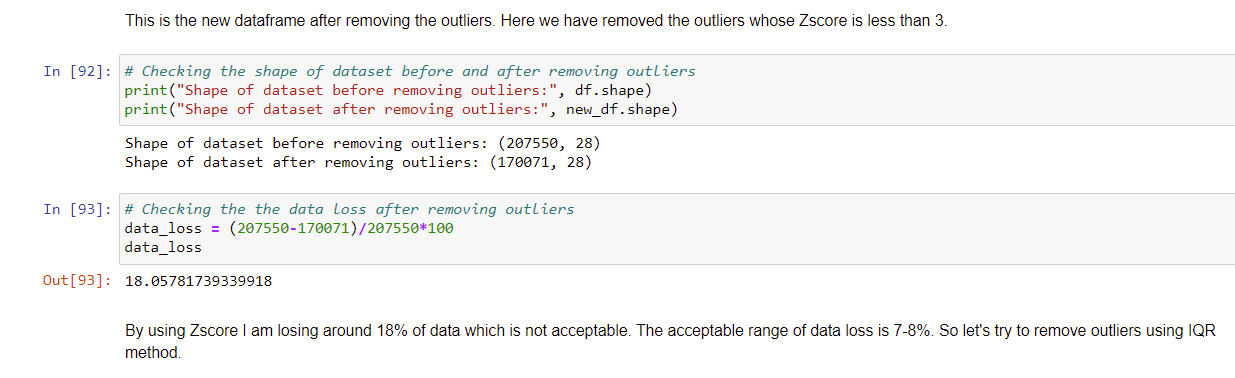
**Model/s Development and Evaluation**

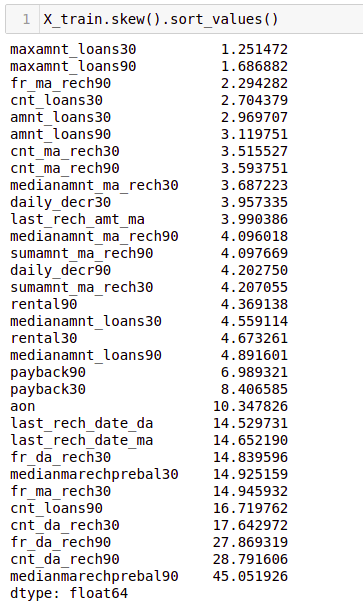
Identification of possible problem-solving approaches (methods)

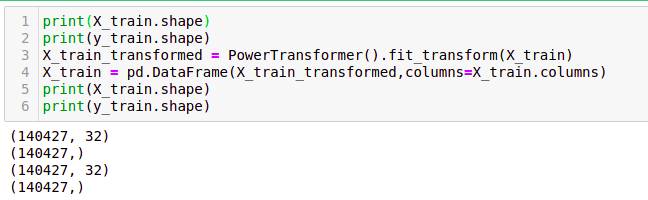


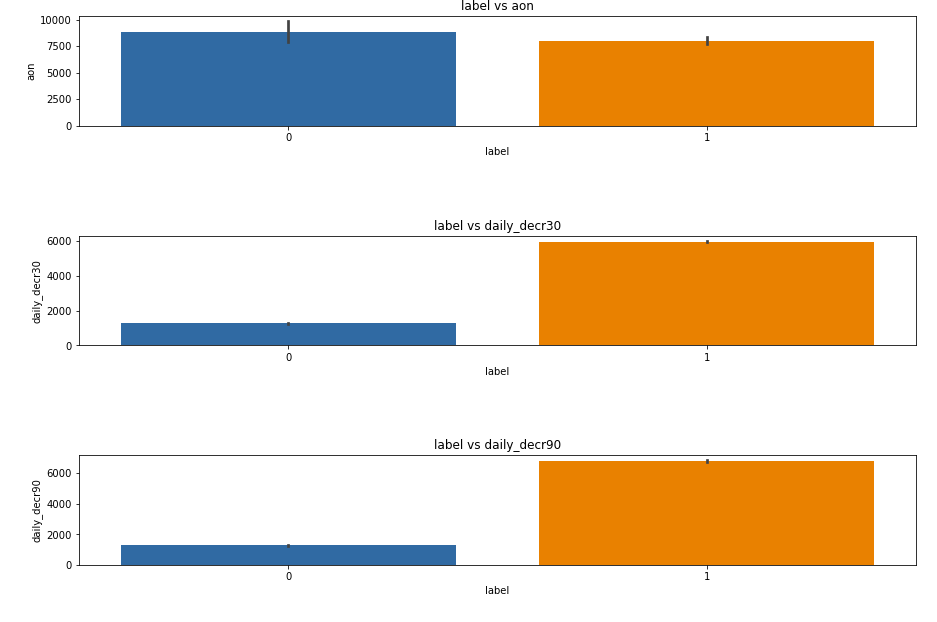


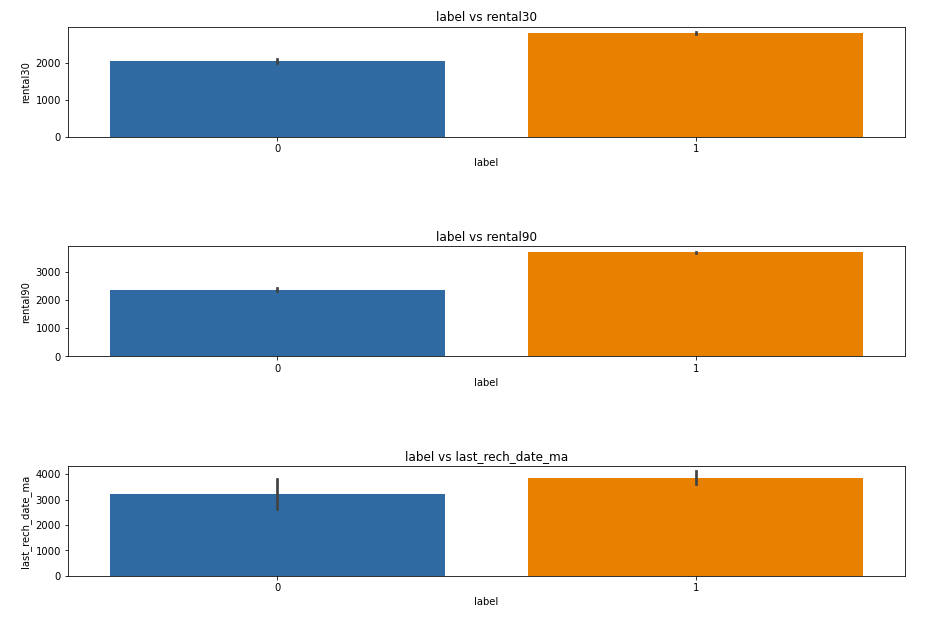


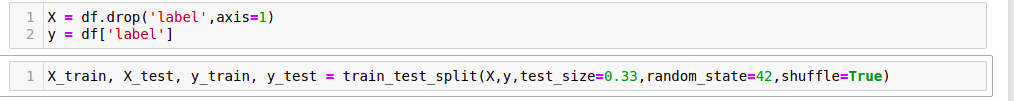












Testing of Identified Approaches (Algorithms)

These are all the Algorithms used for Model Building and Prediction.We did Hyper Parameter Tuning with these algorithms using the GridSearchCV.

RandomForestClassifier

GaussianNB

LogisticRegression

DecisionTreeClassifier

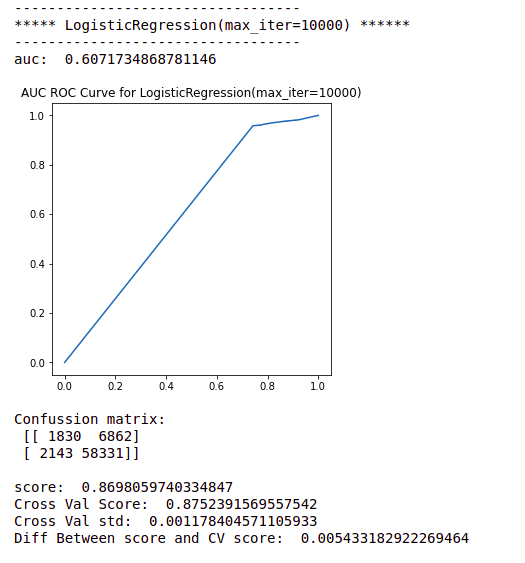
Gradient Boosting Classifier

**These algorithms has been used for both Training and Testing purpose and got evaluated with classification metrics such as f1score,confusion matrix,precision,recall and AUC ROC curve etc.,**

Run and Evaluate selected models

Logistic Regression:

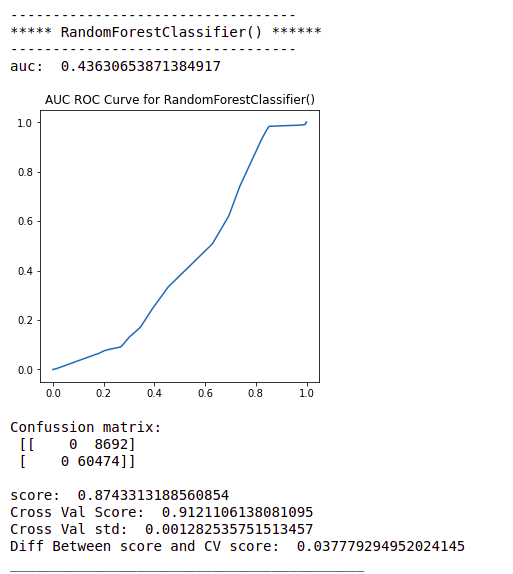
Logistic regression is used **to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables**. Logistic regression is a **simple and more efficient method for binary and linear classification problems**. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification.Logistic Regression is used **when the dependent variable(target) is categorical**.



RANDOM FOREST CLASSIFIER

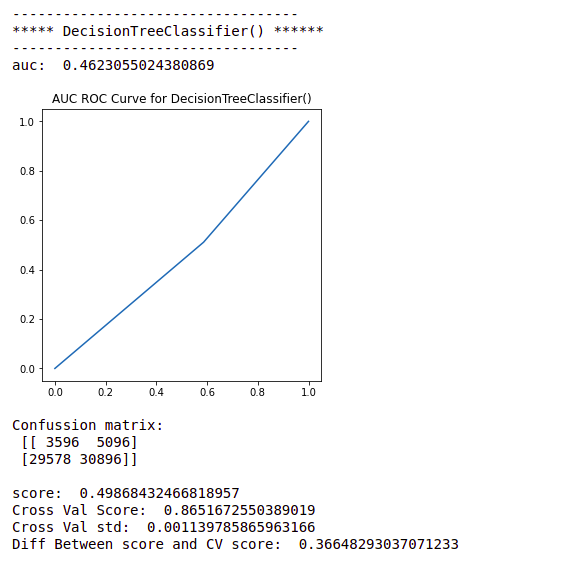
The random forest is a **classification algorithm consisting of many decisions trees**. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. It **can perform both regression and classification tasks**. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

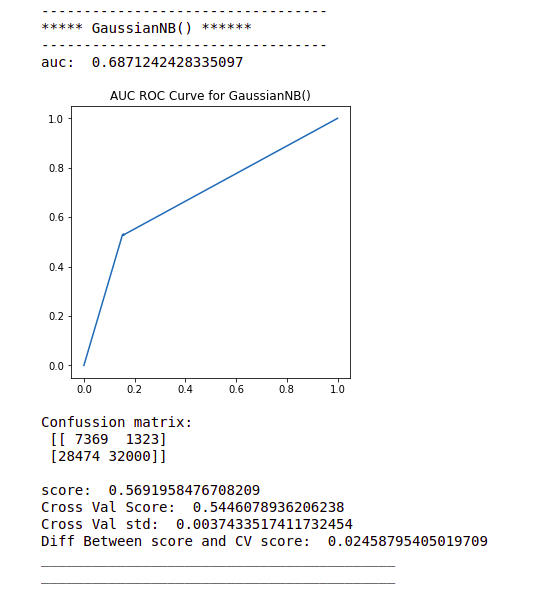
Random forest **adds additional randomness to the model, while growing the trees**. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

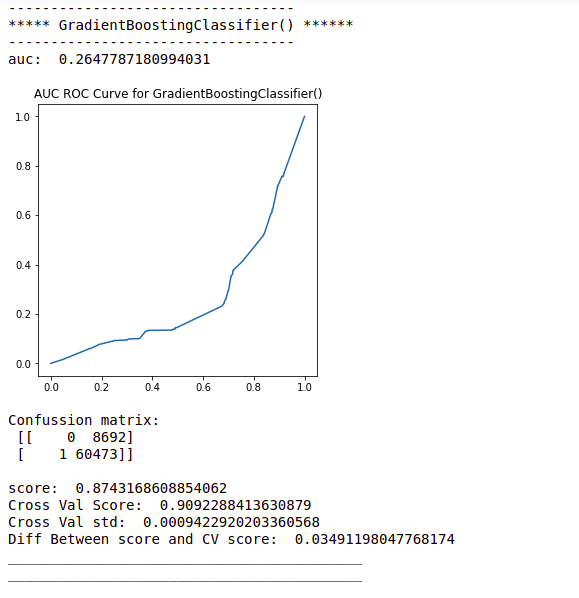


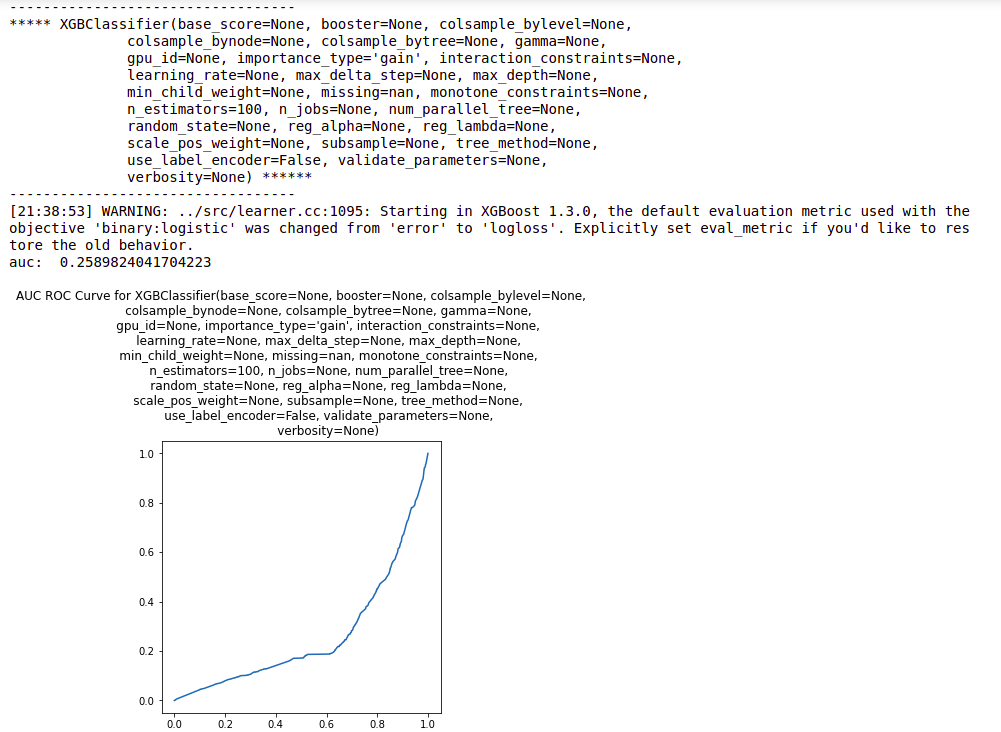
DECISION TREE CLASSIFIER

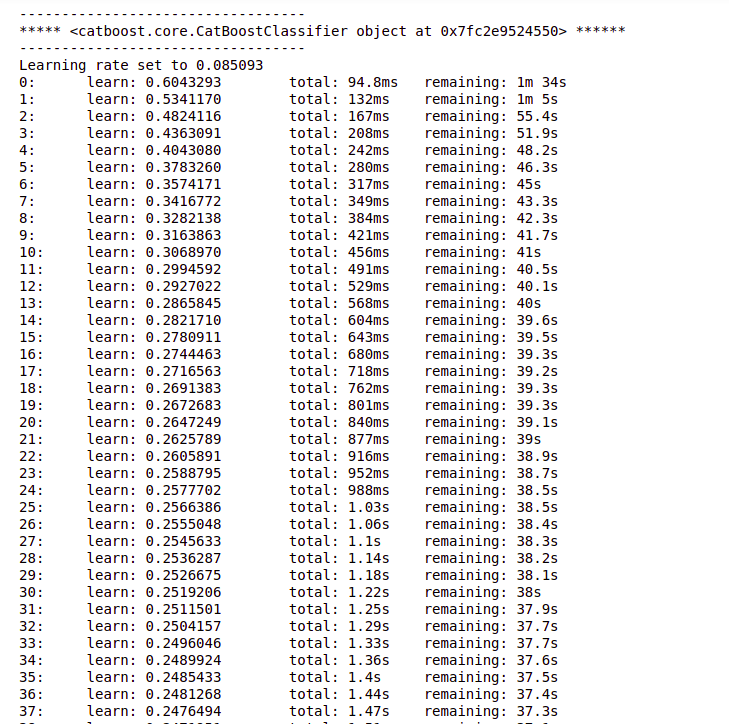
The main advantage of the decision tree classifier is **its ability to using different feature subsets and decision rules at different stages of classification**. Decision tree often involves higher time to train the model. Decision tree training is relatively expensive as the complexity and time has taken are more. The Decision Tree algorithm **is inadequate for applying regression and predicting continuous values**.In this, **the data is continuously split according to a certain parameter**. The tree can be explained by two entities, namely decision nodes and leaves.

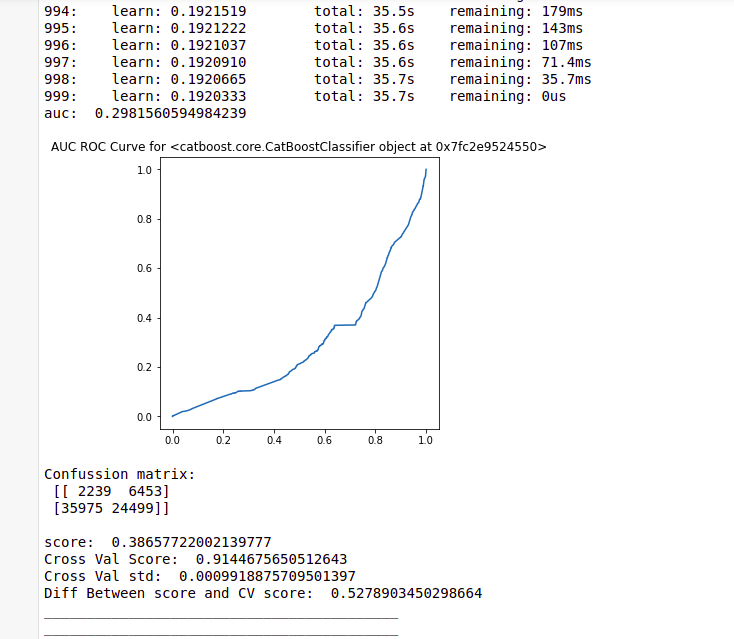












These are some of the algorithms used and it described here with the snapshot of their code and the results observed over different evaluation metrics are also mentioned.

The evaluation metrics used here is classification metrics.

Key Metrics for success in solving problem under consideration

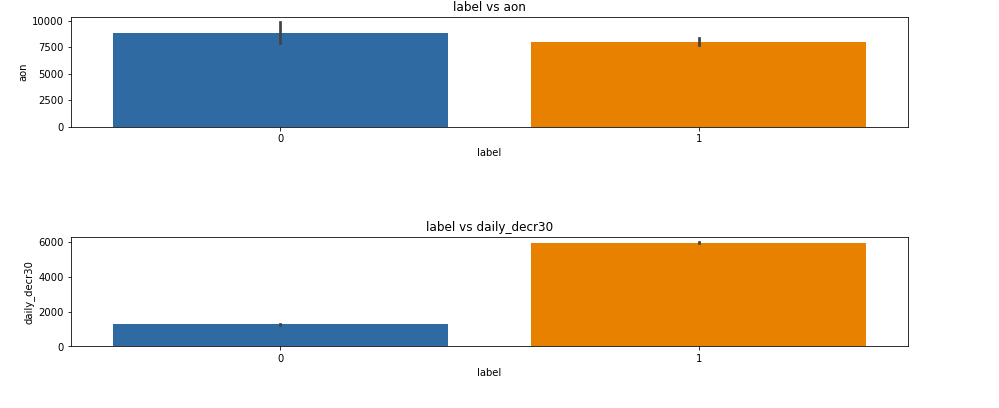
An evaluation metric **quantifies the performance of a predictive model**. This typically involves training a model on a dataset, using the model to make predictions on a holdout dataset not used during training, then comparing the predictions to the expected values in the holdout dataset.

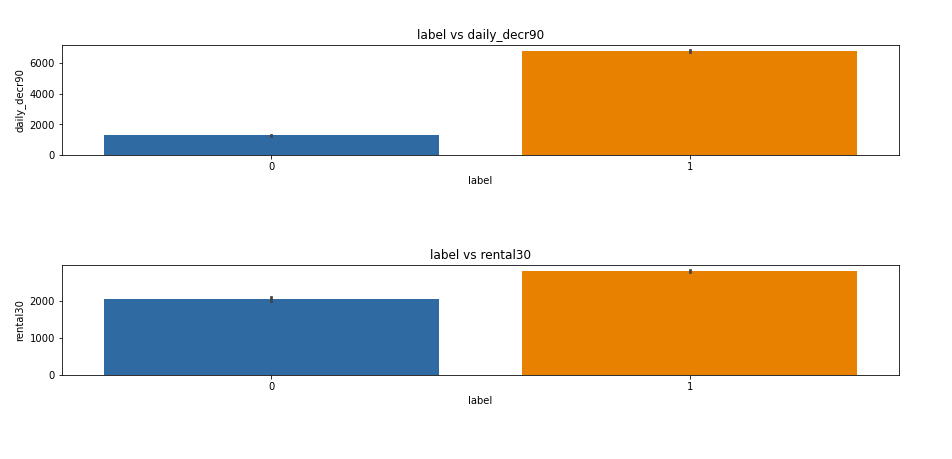
We got Good accuracy with Cat Boost Classifier when comparing with other model’s performance.

CatBoost is **based on gradient boosted decision trees**. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees. The number of trees is controlled by the starting parameters.

Visualizations







MODEL SAVING:



Inferences:

The Defaulter cases of Micro card credit company can be monitored and can improve them as Non Defaulters status by concentrating and changing the following factors mainly,

cnt\_ma\_rech30 ----Number of times main account got recharged in last 30 days

cnt\_ma\_rech90 ----Number of times main account got recharged in last 90 days

sumamnt\_ma\_rech90 ----Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)

sumamnt\_ma\_rech30 ---Total amount of recharge in main account over last 30 days (in Indonasian Rupiah)

amnt\_loans90 ----Total amount of loans taken by user in last 90 days

amnt\_loans30 ----Total amount of loans taken by user in last 30 days

cnt\_loans30 ----Number of loans taken by user in last 30 days

The following cases is more related with defaulter cases.Deep Analysis can be made and Improve the results on by considering the following factors too:

cnt\_loans90 ----Number of loans taken by user in last 90 days

rental30 and rental90------ Average main account balance over last 30 and 90 days

last\_rech\_amnt\_ma------ Amount of last recharge of main account (in Indonesian Rupiah)

fr\_amnt\_rech\_30 and fr\_amnt\_rech\_90----- Frequency of main account recharged in last 30 and 90 days

median\_amnt\_loans\_30 and median\_amnt\_loans\_90-------- Median of amounts of loan taken by the user in last 30 and 90 days

In aug why payback is low than other two months?

Why after calendar day 14, we foresee less payback?

Why all take for 5 Rupiah pay back loan (6 Rupiah).

The company’s customer retention and growth can also be improved mainly by considering the factors

daily\_decr90 ------ Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)

fr\_ma\_rech90------ Frequency of main account recharged in last 90 days

**CONCLUSION**

Key Findings and Conclusions of the Study

Customer selection & Improvement can be clearly understood from outcome of the model.

Key factors which needs to be overlooked to define a customer turning defaulter or non-defaulter, is listed as below,

* 1. Number of loans taken in last 90 days
  2. Rental amount spent across last 30 & 90 days.
  3. Last recharge amount
  4. Frequency of recharge done in last 30 & 90 days.
  5. Median amount of loans taken in last 30 & 90 days.

Critical elements to be better understood & suitable solution to be developed for improvement of credit scheme is as below,

Reason for

* Less customer payback post 14th calendar day of every month?
* Less customer payback frequency in Aug month?
* Frequency of more people opting for 5 Rupiah credit over 10 Rupiah credit plan.

Learning Outcomes of the Study in respect of Data Science

From the above models CatBoostClassifier performs well. Because,CatBoost is the only boosting algorithm with very less prediction time. It is comparatively 8x faster than XGBoost while predicting.It uses symmetric trees, that makes it to have a fast inference. Its boosting schemes helps to reduce over fitting and improves quality of the model. It supports sophisticated categorical features. So we save this model for prediction

CatBoost is that it is easy to use, efficient, and works especially well with categorical variables. As the name implies, CatBoost means ‘**categorical**’ boosting. It is quicker to use than, say, XGBoost, because it does not require the use of pre-processing your data, which can take the most amount of time in model building process.

 In classification, a permutation is performed randomly, then a calculation is performed from a standard formula unique to CatBoost (**ordered target encoding**):

target\_average = countInClass + prior / totalCount + 1

CatBoost has some useful benefits, with easy implementation. Some of the main features of this competitive library are that even without parameter tuning the default parameters provide for great results, categorical features do not need preprocessing, quick computation, increase in accuracy with less overfitting, and lastly, efficient predictions.

Thus this Cat Boost Classifier Model performs well for this dataset.so we saved this model.

Limitations of this work and Scope for Future Work

* It needs to build deep decision trees to recover dependencies in data in case of features with high cardinality. ...
* Doesn't work for unknown category values, i.e., the values that don't exist in the learn dataset.
* Understanding in depth of variable dependencies and handling can solve this problem.