**MICRO CREDIT LOAN USE CASE**

**SUBMITTED BY:**

**Karishma Yadav**

**ACKNOWLEDGMENT**

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

**INTRODUCTION**

The main objective of this paper is to predict whether assigning the loan to a particular person will be safe or not. We have implemented this loan prediction problem using Decision tree algorithm and data cleaning in Python as there are missing values in the dataset. We use map function for the missing values. The aim of this paper is to apply machine learning technique on dataset which has 1000 cases and 7 numerical and 6 categorical attributes. The creditability of a customer for sanctioning loan depend on several parameters,

**ANALYTICAL PROBLEM FRAMING**

* Mathematical/ Analytical Modelling of the Problem

Every classification algorithm is built up with strong mathematical models and logic. Though the decision tree is one of the classification algorithms every algorithm is bound to make one decision before starting of making the trained model. The logic of the SVM is for finding out the optimal separating hyper plane that maximizes the margin of training data. There is a key challenge to find out the appropriate algorithm for each application in machine learning. The application depends on training time, prediction speed and accuracy, number of parameters, number of features, validation strategies, estimation of uncertainty condition, etc. Hence, in machine learning, the amount of math like probability, calculus, linear algebra, complex algorithm, etc. is used to justify the algorithm. The usage of algebra or linear algebra for Eigen value calculation, principal component analysis and probability is used for prediction and choice of the classes as this is a critical part in machine learning algorithm.

* Data Sources and their formats

The provided dataset corresponds to all loans issued to individuals .The dataset has 209593 observations and  37 features. The data contains the indicator of default, payment information, credit history, etc. Customers under 'current' status have been considered as non-defaulters in the dataset. We have also been provided with a Data dictionary that best describes the features.

The dataset has no missing values and the figures can be considered as ground truth, but lots of columns are either irrelevant, very sparse or non-informative. Moreover, the dataset is unbalanced, with approximately 6% of loans considered as defaulted.

* Data Pre-processing Done

The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

* Data Inputs- Logic- Output Relationships

**Correlation:**

### Finding the correlation between variables

We will now look at the correlation structure between our variables that we selected above. This will tell us about any dependencies between different variables and help us reduce the dimensionality a little bit more

* Hardware and Software Requirements and Tools Used
* Tools: Python 3.8.5, Jupyter Notebook, Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, Scipy
* Techniques: classification, Decision Tree classifier, Random Forest Classifier, Ada-Boosting Classifier
* Hardware: I3 processor, 4GB RAM
* Model/s Development and Evaluation

I have used the below models for classification:

**Random forest Classifier**

**Decision Tree classifier**

**Ada-Boosting classifier**

* Identification of possible problem-solving approaches (methods)

1. Read the data (from csv, json etc)
2. Identify the dependent and independent variables.
3. Check if the data has missing values or the data is categorical or not.
4. If yes, apply basic data preprocessing operations to bring the data in a go to go format.
5. Now split the data into the groups of training and testing for the respective purpose.
6. After splitting data, fit it to a most suitable model. (How to find a suitable model is answered below)
7. Validate the model. If satisfactory, then go with it, else tune the parameters and keep testing. In a few cases, you can also try different algorithms for the same problem to understand the difference between the accuracies.
8. From step 7 one can also learn about accuracy paradox.
9. Visualize the data.

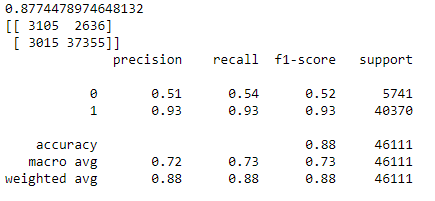
* Testing of Identified Approaches (Algorithms)

1. KNN
2. Naive Bayes
3. Cross validation
4. Confusion matrix
5. Accuracy score

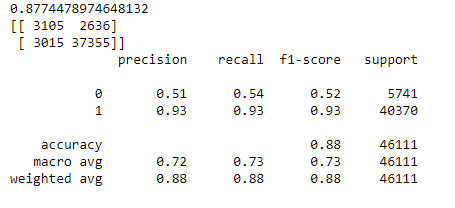
* Run and Evaluate selected models

I have used the below models for classification:

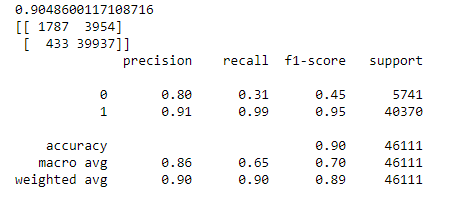
* **Random forest Classifier**

****

* **Decision Tree classifier**

****

* **Ada-Boosting classifier**



* Key Metrics for success in solving problem under consideration

1. Classification Matrix

(Accuracy,precision,recall,F1 score ,ROC, AUC)

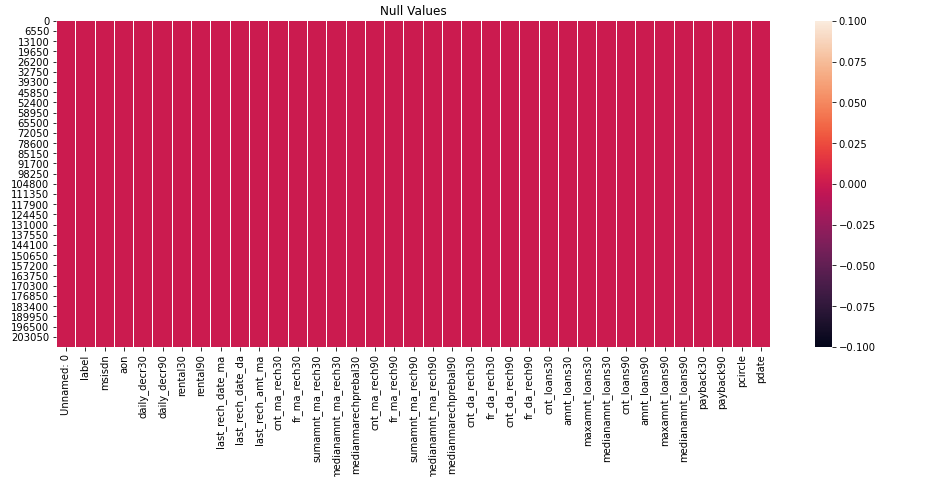
A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

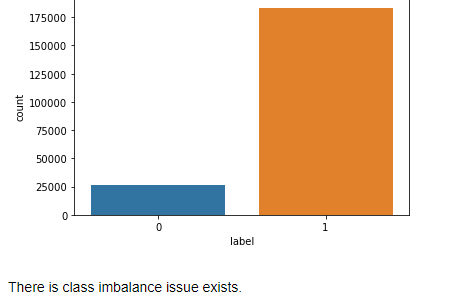
1. Statistical matrix

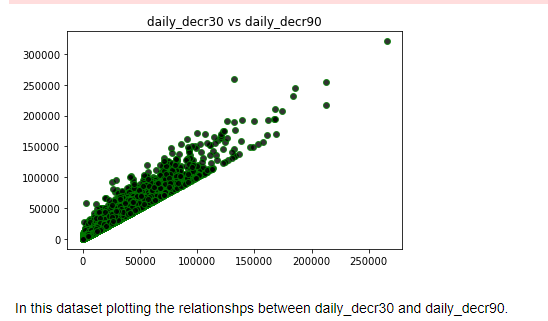
(Correlation)

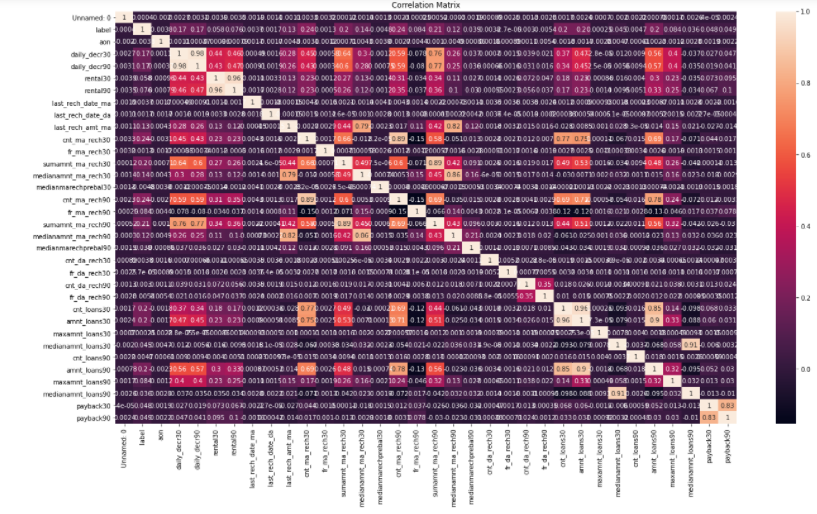
In **statistics** and in particular in regression analysis, a design **matrix**, also known as model **matrix** or repressor **matrix** and often denoted by X, is a **matrix** of values of explanatory variables of a set of objects. ... The design **matrix** is used in certain **statistical** models

* Visualizations









* Interpretation of the Results

Comparing all the above model, looks like the third model has a decent Accuracy and Precision score. Since our data is imbalanced we are getting results that has huge variation between the models. Various sampling techniques can be used in order to balance the data and make predictions but since we have limited time we have not applied the sampling techniques.

**CONCLUSION**

Three ML algorithms are used to predict the loan approval status of customers for bank loans. The results shown that the prediction accuracy is 0.90, 0.87 and 0.87 for ada-boosting, DT algorithm RF algorithms respectively. Among three the accuracy of ada-boosting algorithm is best for prediction of loans. In future the ada-boosting algorithm can be applied on other data sets available for loan approvals to further investigate its accuracy. A rigorous analysis of other machine learning algorithms other than these three can also be done in future to investigate the power of machine learning algorithms for loan approval prediction.