

***Micro-Credit Defaulter Model***

Submitted by:

Anu Sindhuja Galidevara

**ACKNOWLEDGMENT**

<https://www.kaggle.com/selener/prediction-of-credit-card-default>

<https://stackoverflow.com/questions/40758562/can-anyone-explain-me-standardscaler>

**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.

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).

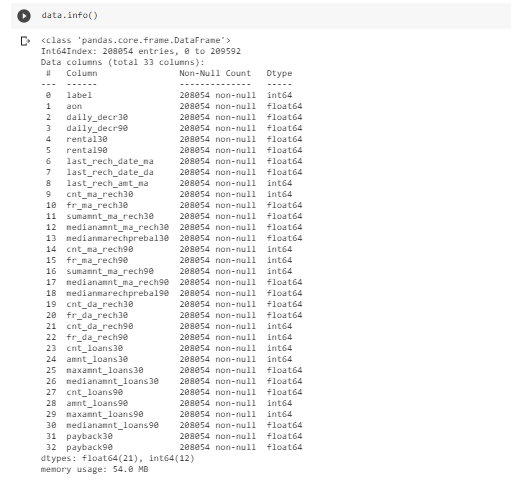
This problem aims to apply multiple machine learning algorithms to analyze the terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. We will evaluate and compare the performance of the model candidatesin order to choose the most robust model. Moreover, we will also decide which are important features in our best predictive model.

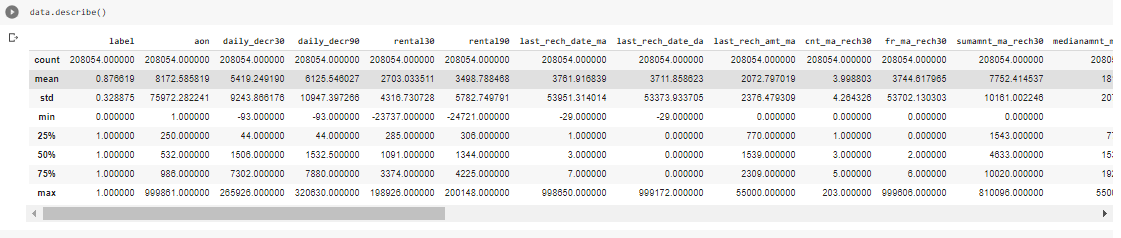
**Analytical Problem Framing**

This document is structured as follows:

* First exploration: just to see what we have.
* Cleaning: time to make choices about undocumented labels
* Feature engineering: time to be creative
* Final result and lessons learned

As a first step, we have to look if there are missing or anomalous data





There are no missing values but a few anomalous things

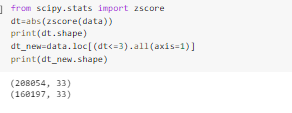
We apply outlier and extreme values elimination

based on interquartile ranges.

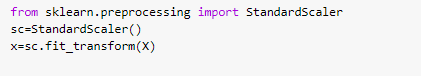
We apply outlier and extreme values elimination

based on interquartile ranges.

For outlier and extreme value elimination we use zscore.



Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data.



**Model/s Development and Evaluation**

The supervised machine learning can be further divided into two sub-groups: classification and regression. The problem with categorical outputs are grouped into classification problem, while the outputs of a regression problem are numerical. In this research, the output is either credit card defaulter or non-defaulter. The problem, hence, should be grouped into classification problem.

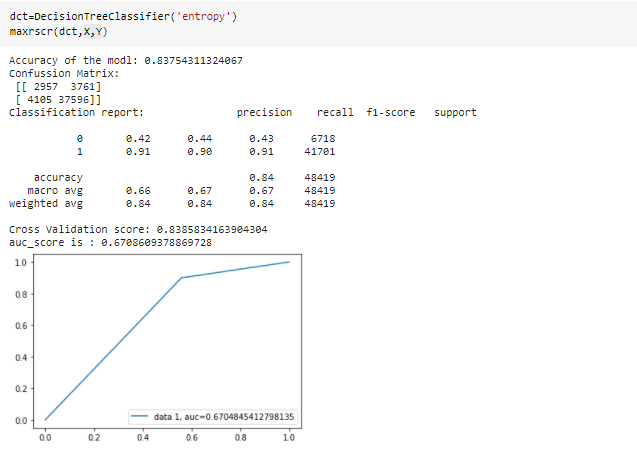
* Logistic Regression:

The logistic regression algorithm is based on the concept of probability of the predicted output which lies within 0 and 1 range:



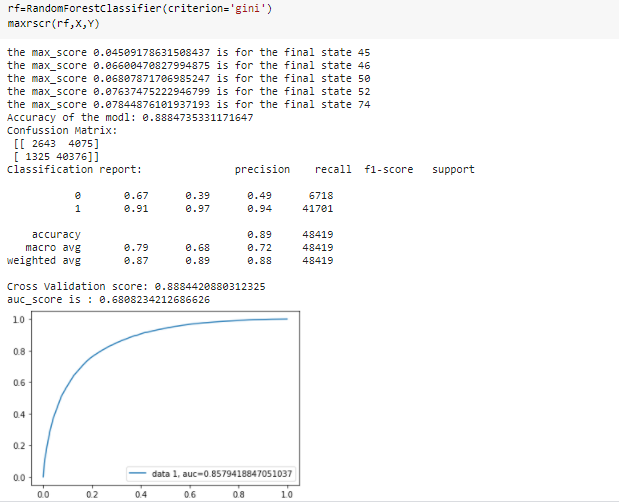
* Decision tree

The Tree is very easy to use and imagine, it can handle both numerical and categorical variables, but it is easy to fall into the trap of overfitting (i.e. creating a very complex model that perfectly describes your training set but fails in predicting). It is also said to be senstive to small changes in the data, leading to different results. For this reason, it is common to use ensembles such as Random Forest to avoid this risk.



* Random Forest

While the decision tree algorithm has many advantages, it also has many drawbacks. For example, the decision tree that is returned by greedy algorithm may not be the globally optimal tree. In this case, building multiple trees is in demand. The samples are randomly selected and replaced for the individual tree and the final model is built by merging all trees together so that the result should be more accurate and stable than one decision tree. The process is also named as “bagging” method



According to the confusion matrix and performance index of random forest model, there are 40376 true positive values and 2643 true negative values. The accuracy classification score for test set is 0.88,which means that about 91% of data instances in the test set are predicted correctly by our random forest model, which is trained based on the training set.

**CONCLUSION**

The objective of this paper is to train multiple supervised learning algorithms to predict customers behavior on paying off credit card balance. We first investigated the data by using the exploratory data analysis techniques including cleaning missing or invalid values and exploring the relationship between different features. The correlation plot helped us to visualize these relationships and important features. We started with the logistic regression algorithm, then built a decision tree model which has a better performance than the former model. Next, we experimented with an Randomforest model. The prediction accuracy rate of decision tree model is higher than the Randomforest model.