

**Micro-Credit Defaulter**

Submitted by:

Prabhat Chauhan

**ACKNOWLEDGMENT**

I would like to mention my special thanks to flip robo for constantly providing us the opportunity to work on this projects which helped me to learn new things, and I also pay my sincere thanks to sajid Chaudhary sir for providing the support that is needed in between the projects, with his guidance I learned a lot and will continue to do so.

**INTRODUCTION:**

* Business Problem Framing

The problem statement is to predict whether the customer will be paying loan amount in 5 days or takes more than 5 days to payback loan. The telecom and DTH industry has been lending such short term loans to all the customers in general, with this kind of project business can decide whether it is beneficial to offer such loan, for which we will analyse loan + interest received from this service and loss due to default in payment can be compared and derive a conclusion finally predict future expected losses and future expected additional income to decide loan offer is a feasible decision

* Conceptual Background of the Domain Problem

The challenge to lenders to predict the future financial status of customer or borrower based on current conditions, but with the help of Machine Learning and handful information one can build a predictive model, which helps the businesses to evaluate the lendable loan amounts, target customers, decision about interest chargeable and so on.

Data science is the solution for such future accountable predications.

* Review of Literature:

Microfinance, according to our project is an offer provided by a telecommunication network in collaboration with micro financial institution, allows customers to take small loans in a manner that is consistent with ethical lending practices, with low income families and poor customers as their main target.

The offer(loan) details is that loan borrowed should be paid back in 5 days, failing which Consumer is considered as defaulter, since he is not paid back within given time period of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount is 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount is 12 (in Indonesian Rupiah).

The telecom service provider wants to make a decision if the current plan is feasible and understand the future profits out of this offer, hence we have used machine learning as a tool to predict the expected defaulters and non-defaulters on the basis of given attributes as parameters for forecasting.

Since it is categorical type of prediction (defaulter & non- defaulter) we have used classification algorithms

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken

The data shape is 2, 09,593 which is divided into 37 columns, understanding the columns and its impact on the target column was the challenge and motivation to accomplish the goal of predicting the target.

**Analytical Problem Framing:**

* Mathematical/ Analytical Modeling of the Problem

1. We have used percentage formula for analysing the data we will have to forego while removing outliers.
2. Statistical tool – with the help of correlation we divided columns into three category :-

Highly positive correlation of attributes with label, Positively correlated columns

Negative correlation.

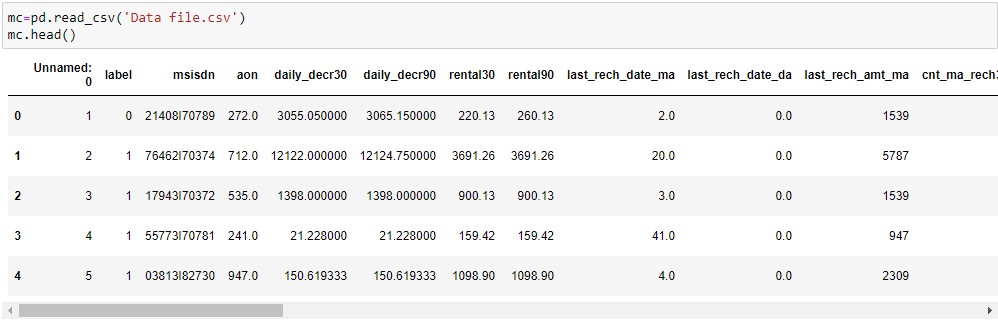
1. Feature engineering has been used to analyse Pdate column, visualize label column with other attributes & analyse categorical column.
2. While building the model, we understood all algorithms gave different accuracy, which could be result of over fitting the model hence cross validated and used grid search CV to extract the best parameter of analysis

* Data Sources and their formats

The data was provided by Fliprobo Company with a detailed project history which contained. Detailed description of each column.

➢ Dataset

* Problem statement of the project.
* An over view of data after uploading in notebook



* Detailed description of all attributes in our dataset



* Data Pre-processing Done:

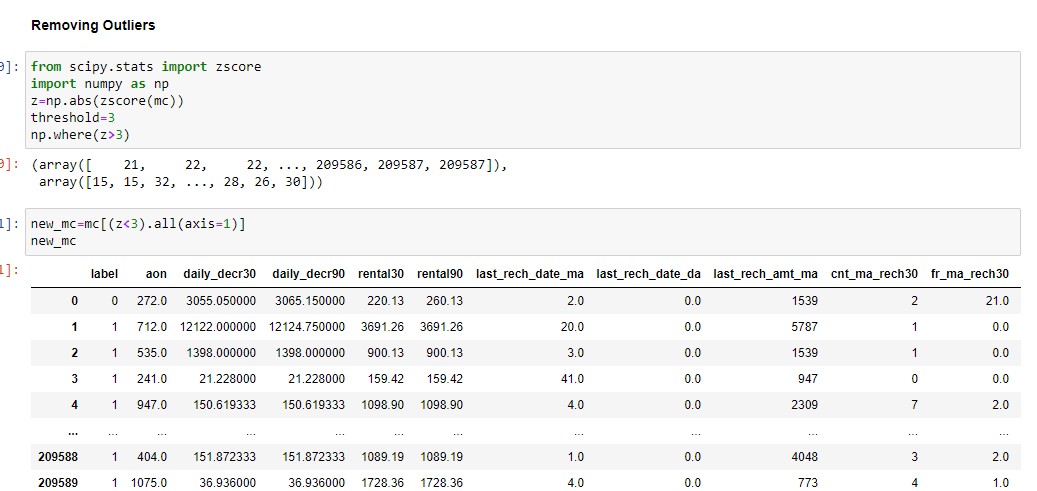
Unnamed column is similar to index hence we cannot make decision with this column so we will drop the column.

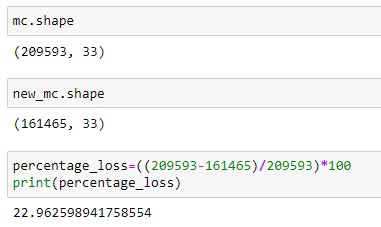
Pcircle column has ‘UPW’ in all the columns hence we cannot make any productive result analyzing this column, so we will drop the column. Msisdn column again has unique id of the customers hence we remove this column.

Pdate column is in DDMMYYYY format, so we will drop the column. Understood there is outliers in all columns after visualizing in bar plot and analyzing extent of data being skewed?

Checked outliers removal process and found Zscore will remove 22.96%.

Outliers removing process is not recommended as we will lose substantial data.





* Data Inputs- Logic- Output Relationships

In the process of understanding the data and processing the same we have selected all the columns except: - unnamed: 0, label, msisdn, pcircle & pdate. are removed because we already have median amount column for the same data.

* Hardware and Software Requirements and Tools Used In case of hardware requirements 4 -8 GB RAM, i3 processor. Whereas software

Requirements are:- Python, Jupyter notebook.

Libraries :Pandas, Numpy, Scipy, Seaborn, matplotlib.pyplot, zscore, sklearn.model\_selection( train\_test\_split,cross\_val\_score), sklearn.tree for decision tree classifier, sklearn.neighbors for Kneighbors classifier, sklearn.ensemble for randomforest, Adaboost, Gradient boosting and Bagging classifiers.

sklearn.metrics (confusion\_matrix, accuracy\_score, classification\_report, roc\_curve, auc, recall\_score, precision\_score), sklearn.externals (joblib)

Model/s Development and Evaluation:

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

For understanding the distribution of data and any deviations present, which part of the data contains outliers we used Describe as statistical tool and observed the data on the basis of difference between :- mean and standard deviation, minimum and Q1, Q3 and maximum data. Correlation as used to understand the extent attributes bear impact on our target column ‘label’.

Used sclera to train and test the model and cross\_val\_score to cross validate the outcome in choose the best, then applied GridsearchCV to choose best parameters for the highest scoring models, metrics and auc\_roc curve to derive the best fit model.

* Testing of Identified Approaches (Algorithms):

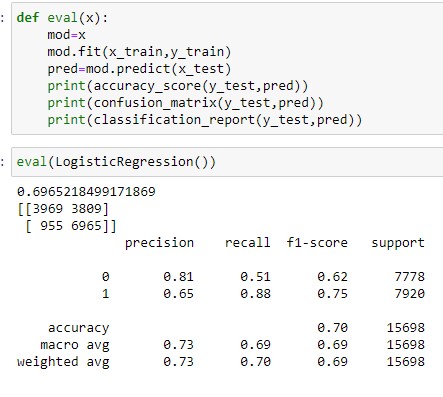
The classification algorithms used to predict micro credit loan defaulters are:-

* 1. Logistic regression
  2. SVC
  3. Decision tree
  4. Kneighbors
  5. Randomboosting
  6. Adaboosting
  7. Gradientboosting
  8. Bagging
* Run and Evaluate selected models:

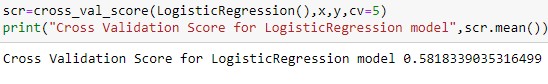
Classifiers adopted and results:

Logistic Regression: It is an appropriate analysis to conduct when the dependent variable is binary type, keeping target column positive heart disease or negative heart disease which is converted to binary type we choose logistic regression for predictive analysis.

Accuracy score

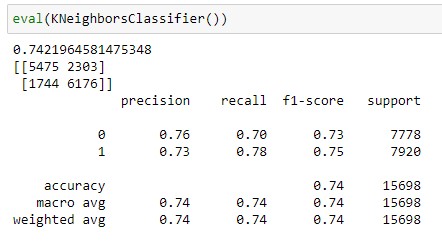


Cross validated the result



K-Nearest Neighbors: The classifier which derives value of k from root of total number of samples and classification is made by giving majority votes to its neighbors which is more apt for our dataset as we can observe for similar age group, sex, type of chest pain etc., how people have been effected.

Accuracy Score

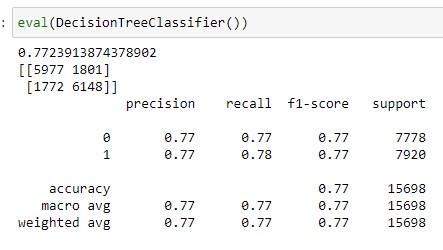


Cross validated the result

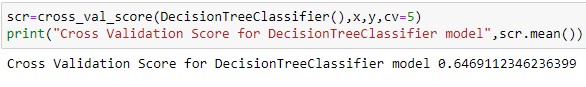


Decision Tree Classifier: It is basically analyzing data in a tree like structure, which is widely used in medical Dataset and best on categorical data.

Accuracy score

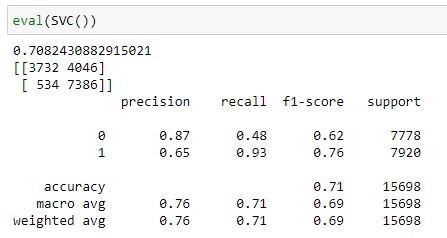


Cross validated the result

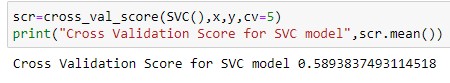


Support Vector Classifier: The main objective of support vector classifier is to the data in hyperplane it leads in categorizing the data which helps to make predictions accurately.

Accuracy score

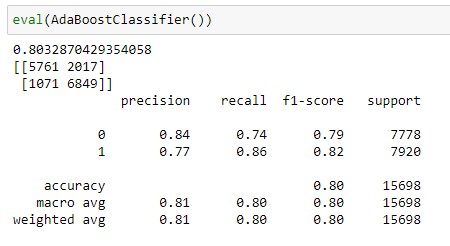


Cross validated score

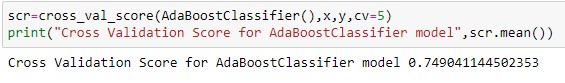


AdaBoost Classifier: AdaBoost Classifier is best used to boost the performance of decision trees on binary classification in our target data we have defaulter and nondefaulter to be predicted which is in binary form and we have used AdaBoost to increase the efficiency of binary classifiers.

Accuracy score

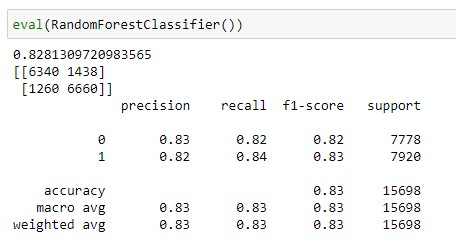


Cross validated the result

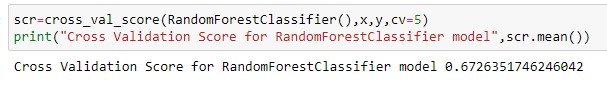


Random Forest Classifier: Random forest creates decision trees on randomly selected data samples, derives solution or predicts for each tree and finally selects the best solution out of it, so I can be considered as collection of decision trees.

Accuracy score

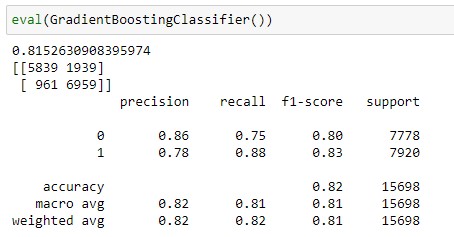


Cross validated score

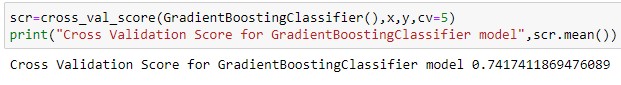


Gradient Boosting: Gradient Boosting is a machine learning boosting that fits boosted decision trees by minimizing an error gradient, as the model relies on the intuition that the best model possible next model, when combined with the previous models and hence minimizes the overall prediction error.

Accuracy score

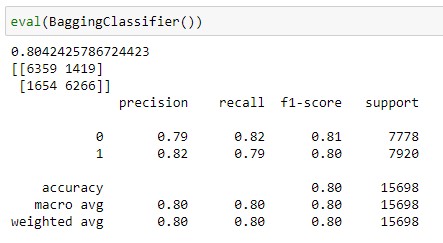


Cross validated result

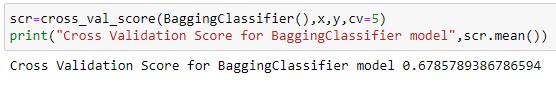


Bagging: Bagging is one among the best classifiers as it created using multiple estimators which can be trained using different sampling techniques.

Accuracy score



Cross validated result



• Key Metrics for success in solving problem under consideration

After understanding the data, data cleaning, Feature engineering, data visualization and summarising the outcome from it, training model & hyper parameter tuning next step is to apply evaluation metrics to finalize the model, we have used

**Confusion Matrix:**

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

**Accuracy:**

Accuracy is the proportion of true results among the total number of cases examined.

Accuracy = (TP+TN)/(TP+FP+FN+TN)

**Precision:**

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, it is a valid choice of evaluation metric when we want to be very sure of our prediction.

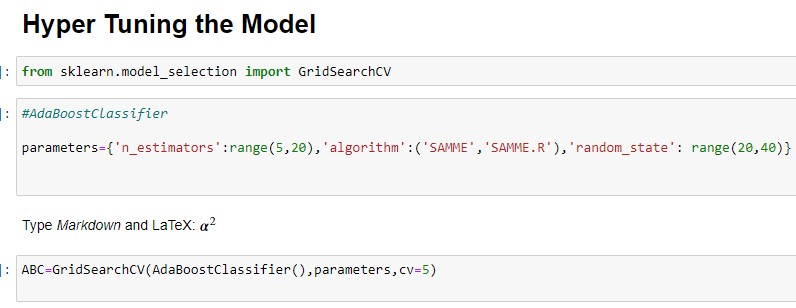
Precision = (TP)/(TP+FP)

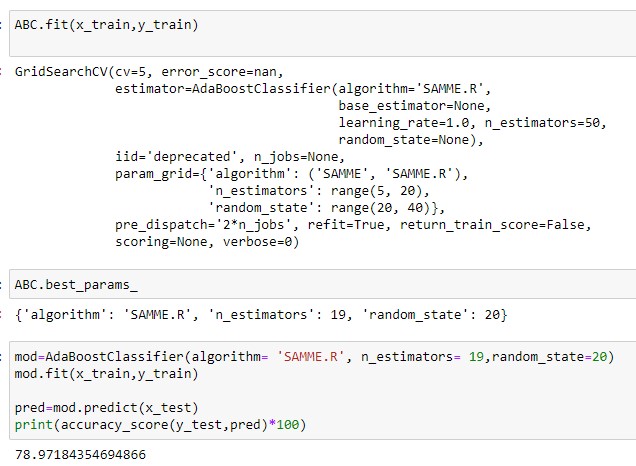
**Recall:**

Recall is a measure that tells us what proportion of customers actually failed to make payment within 5 days by the algorithm as having defaulter.

**Hyper Parameter Tuning:**

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

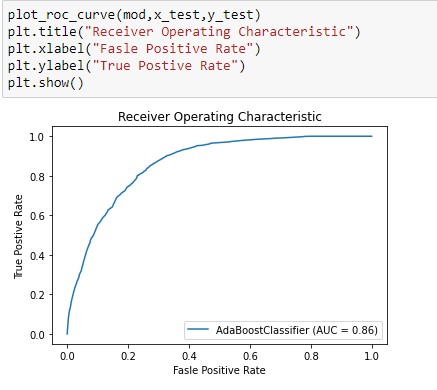




**AUC\_ROC:**

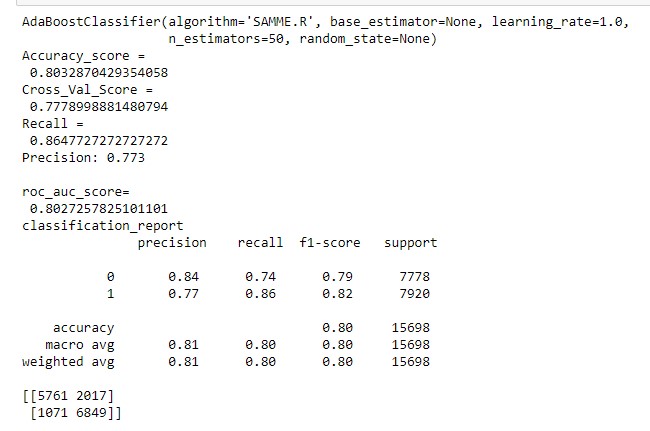
The “Area under the Curve” (AUC) of “Receiver Characteristic Operator” (ROC)

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems, it is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.



Interpretation of the Results:

The best results interpreted after pre-processing and model building are as follows



We can conclude Ada Boost Classification has the best predictable accuracy score hence we have dumped Ada Boost Classification in the pickle file.

**CONCLUSION:**

* Key Findings and Conclusions of the Study

We understood Pdate column was important but not based on year hence we extracted Pmonth and Pdays from Pdate column and deleted Pdate. We could observe from 37 columns there were few columns repeated like for example data column, data with minimum values, data with maximum values and data with median column such information has deleted and reserved only the apt column for analysis.

* Learning Outcomes of the Study in respect of Data Science

With the help of exploratory data analysis we could visualise the large data and make interpretation of it and also summary statistics helped us understand how data is spread, where outliers are present , to what extent deviation is present and understanding the relationship between target column and other attributes, repeated columns was eliminated to making right prediction.

* Limitations of this work and Scope for Future Work

Limitation of the study is, we could have made better data analysis if the population had equal number defaulters and non – defaulters.

Once we make analysis based on above specification we can understand columns better in order to delete or divide the important columns categorically for better analysis.