

MICRO CREDIT LOAN – DEFAULTER PREDICTION

Submitted by:

NARENDRAN SURESH

**ACKNOWLEDGMENT**

I would like to take the opportunity to thank the organisation ‘FlipRobo Technologies”, “Datatrained “and my mentor Ms. Astha Mishra for their immense support and guidance for making this feat achievable. I would also like to thank my dad Mr. Suresh Kumar S who is a statistics subject expert for guiding me along the right path and techniques to be used.

FlipRobo Technologies who are specialised in making ML / AI models provided me the Data set as a part of my internship.

References were made to several articles among Medium, KdNuggets, towardsdatascience, realpython, machinelearningmastery, python and sklearn documentation for the successful completion of the project.

**INTRODUCTION**

* Micro-Credit Defaulter – Business Problem

The Client (a telecom provider in Indonesia ) is providing mini credit loans in the form of main account balance in the mobile account which has to be repaid by the users over a period of 5 days. The objective is to create a machine learning model that can predict if the user will return the loan within the timeframe of 5 days or if he will be a defaulter.

* Conceptual Background

The problem is similar to a typical classification problem (bank loan defaulter / customer churn ) where the dataset is imbalanced. The number of instances of defaulters in the entire dataset is limited to 12.5% . SMOTE technique needs to be applied to balance the dataset to avoid the model being biased. The confusion matrix, roc-auc curve and f1 score are crucial metrics in evaluating the model performance.

* Review of Literature

The Client is a mobile network provider in Indonesia. Mobile communication is very important in today’s world and effective communication opportunities is a key to development among the citizens of the country. Unfortunately, the situation is not the same among all the income classes in any country. Our Client, hence are aimed at providing their services to low income and poor population. They are operating a model where the user can avail a loan of either 5 or 10 Indonesian Rupiah into their mobile accounts. The customer then has to return 6 and 12 Indonesian Rupiah in the next 5 days. The customer is considered a defaulter if the customer fails to pay back the loan amount with the accrued interest within the period.

* Motivation for the Problem Undertaken

The project was the first provided to me by FlipRobo as a part of the internship programme. The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary motivation. Further diving into the dataset, the motive is to help the poor or low-income band to have continuous access to their mobile accounts, and to make emergency calls even when they do not have account balance making use of the loan facility. Alternatively, it is also important to ensure that the client/provider does not incur a loss for providing the facility. Hence, we are focused on building a model that predicts defaulters by making use of the historical data which would help in approval process of the loans to end users with a clean sheet.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

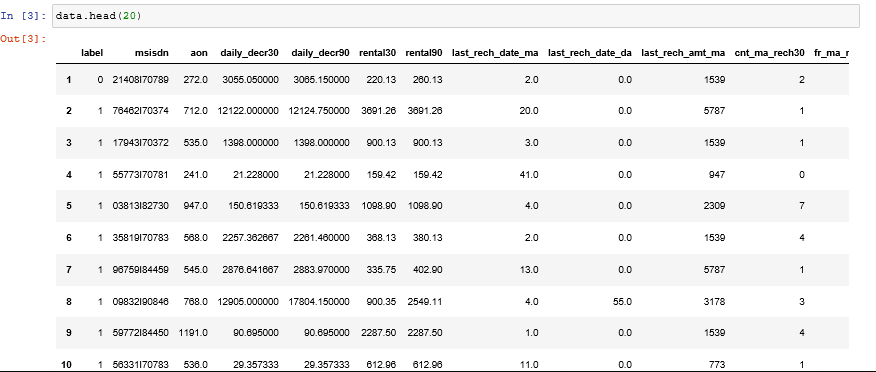
The dataset is a csv file with 37 attributes (36 features and 1 target). The target variable is either 1 or 0 which means non defaulter and defaulter respectively. The other key attributes are the account balances, days since last recharge, age on network, median recharge balance for 30 and 90 days and many more. The similar attributes for 30 and 90 days are highly correlated and conveys the same. Hence for the purpose of the project, highly correlated attributes needs to be removed.

* Data Sources and their formats

The dataset contains 37 columns with each row explaining features and characteristic of users and their defaulter history. FlipRobo, as a part of the internship provided the data for analysis and modelling.

Apart from the two features with object datatype and one with date, all the others are either int or float.

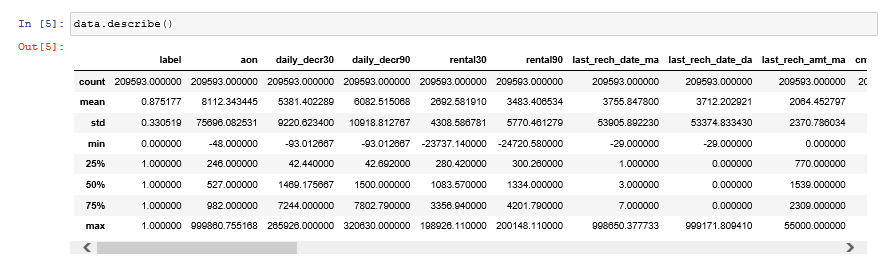
Below is a snapshot of the data and its datatypes.



* Data Preprocessing

From an initial glance, it is understood that the msisdn(the phone numbers) are unique values and does not add value to our analysis. Hence the column can be dropped.

Also the ‘pcircle’ which is the telecom circle, which has only one unique value across the dataset, ‘UPW’ can also be dropped.

From the describe function (image below, it is clear that the dataset has outliers.

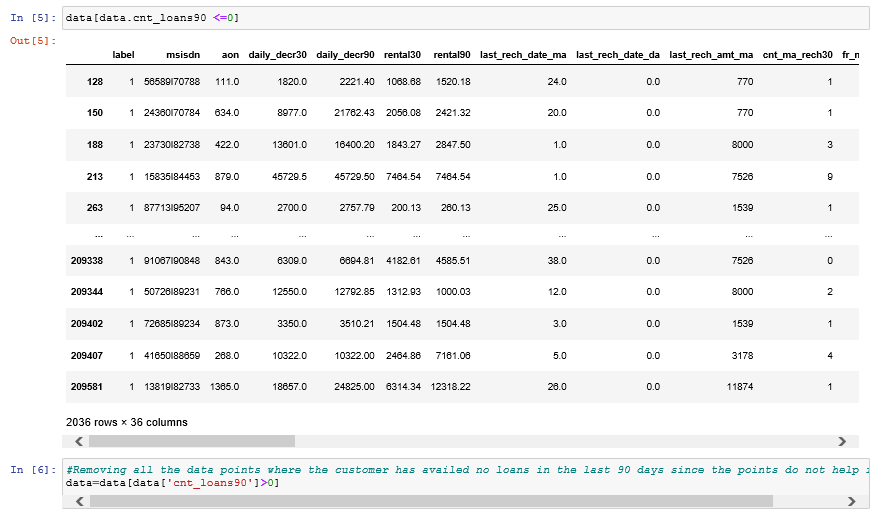
Before using generalised techniques, a univariate analysis is done to find and omit any insensible data.

Univariate Analysis as follows:

* aon (Age on network) has a mean of 8112 and the 75th percentile lies at 982. The maximum value in years is 2739 most likely to be an error.
* daily\_decr30, daily\_decr90 max values also suggest the presence of outliers, the values suggest that 265926 and 320630 Rupiah were spent on a day, mostly likely to be an error rental30 and rental90 also have outliers.
* Last\_rech\_date\_ma and last\_rech\_date\_da also have outliers since the max values in years is 2739 which is impossible last\_rech\_amt\_ma has the max value at 55000 whereas the mean and 75th percentile at 2064 and 2309 respectively.
* Presence of outlier detected maxamnt\_loans30 and cnt\_loans90.

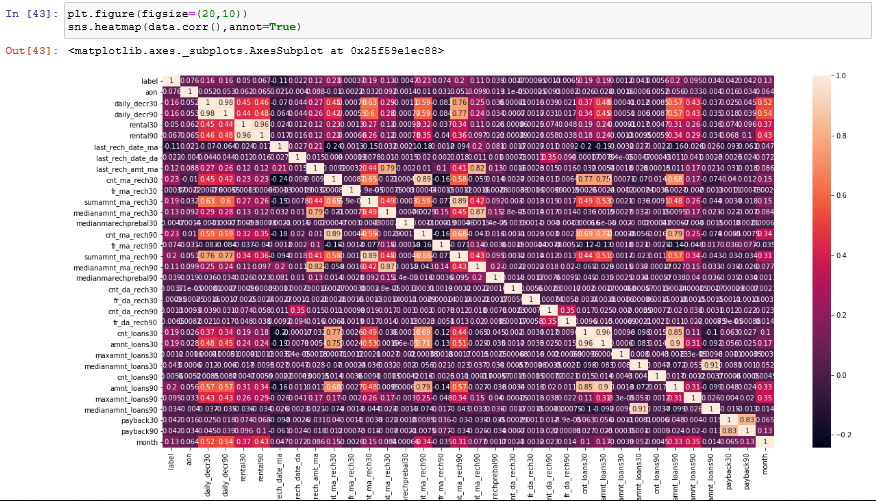
Assumptions made:

* Data points with age on network (aon) >14600 were removed since it has not been 40 years since mobile phones were discovered
* Data points with cnt\_loans90 < = 0 were removed since the customers with no loans in the last 90 days do not benefit the case study.



All the columns which mentioned no of days where the days were greater than 14600 has been removed for the reasons mentioned above.

A check for correlation can now be performed:



From the above heatmap, it can be noticed that quite a number of variables have a very high correlation with others.



The above columns were dropped since it had a correlation of > 0.9.

A similar analysis was done for all the columns.

Now that univariate analysis has been taken care of, we can now apply zscore and remove the outliers in the data.

* Data Inputs- Logic- Output Relationships

The relations are evident as per the correlation heatmap. All features are positively correlated to the target.

Apart from that there are variables are related as below.

cnt\_loans\_30 and amnt\_loans\_30 have a correlation of 0.96, so it is safe to remove either.

#rental\_30 and rental\_90 have a correlation of 0.96, so either can be removed.

#daily\_decr\_30 and daily\_decr\_90 have a correlation of 0.98, so either can be removed.

#amnt\_loans\_30 and amnt\_loans\_90 have a correlation of 0.9, so either can be removed .

#median\_amt\_loans\_30 and median\_amt\_loans\_90 have a correlation of 0.91, so either can be removed.

* Assumptions
* Data points with age on network (aon) >14600 were removed since it has not been 40 years since mobile phones were discovered.
* Data points with cnt\_loans90 < =0 were removed since the customers with no loans in the last 90 days do not benefit the case study.
* Hardware and Software Requirements and Tools Used

Hardware:

Inter Core (i7) – 5500U, clock speed at 2.40GHz

RAM – 12.0 GB

Software:

Jupyter Notebook (Anaconda 3) – Python 3.7.6

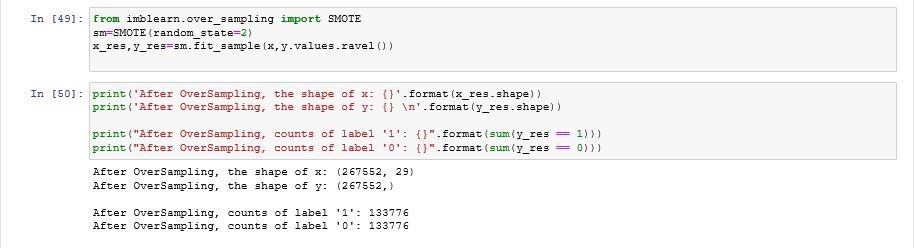
Microsoft Excel

Libraries & Packages used – Pandas, numpy, sklearn, matplotlib, seaborn, sklearn, scipy, imblearn

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

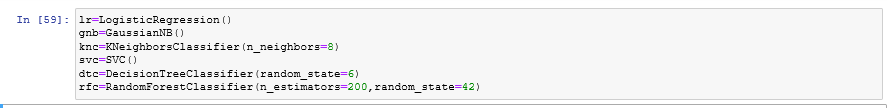
After having removed the illogical data and the outliers, data can be split into features and the target variable. But the real problem we have here is that the data is still imbalanced.

To sort this, the SMOTE technique is applied so that the data becomes balanced.

Just one more step and we can get to training the models. We need toscale the data using standard scaler prior to the data being fed into the models.

* Testing of Identified Approaches (Algorithms)

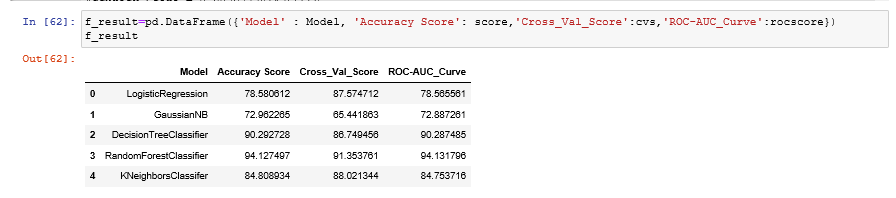
The problem in hand is a typical classification problem, hence we use the following algorithms and find the one that suits our dataset the best:

* Logistic Regression
* GaussianNB
* KNeighborsClassifier
* SVC
* DecisionTreeClassifier
* RandomForestClassifier
* Run and Evaluate selected models

Since there are only two outcomes we are trying to predict, it is worth trying logistic regression as well.

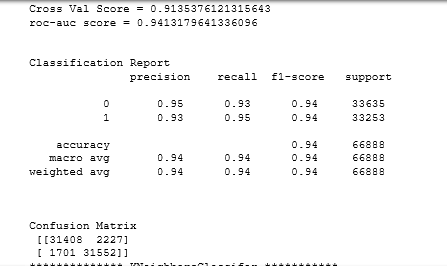
The algorithms can be run in loop using the below code:



The results of the above code has been summarised into a table below:

From the above, it is clear that Random Forest Classifier is performing really well.

Lets now look at the other performance metrics of the model



Key Metrics for success in solving problem under consideration

Accuracy score at 94.12% was considered as the base metric, but inorder to ensure that the model is not overfitting, other metrics are to be considered as well.

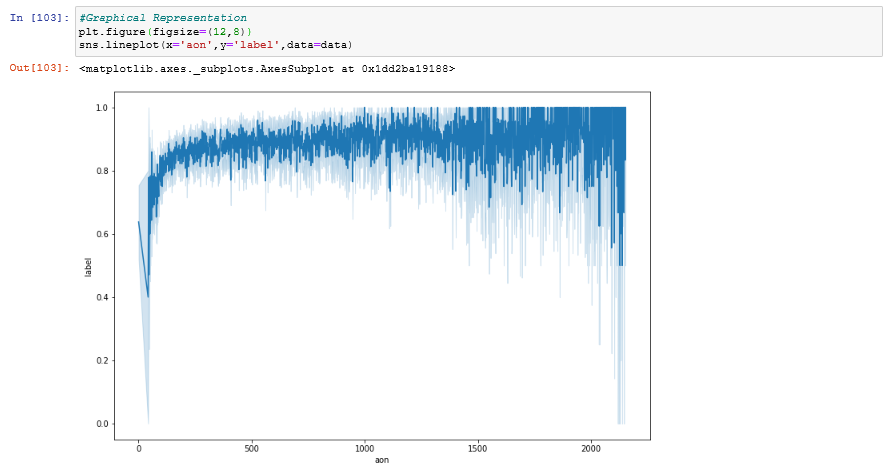
The cross – validation was done to check the model for overfitting and it gave a good 91.3%

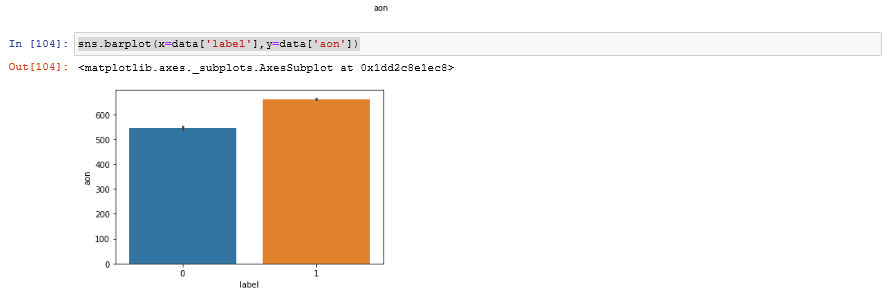
The ROC-AUC Curve was also checked for and it was evaluated to be 94.1%.

The confusion matrix,f-1 score and precision were also evaluated for each models.

* Visualizations

A plot was done between the age on network and label and it was observed that either new or very old customers had a tendancy to be a defaulter, but this could not be confirmed since data is all over the place.



Now, a bar plot was tried between the same variables

Inferences were not much different, hence other variables like dependence of the label on month was tried for.



* Interpretation of the Results

From the results, it is clear that the model performance is above 90% which is a desirable score for any model. To further enhance the results, we can try hyperparameter tuning using GridsearchCV.

The precision,recall and f1 score are all above 90% and the model can make a good prediction for the future data.

**CONCLUSION**

* Key Findings and Conclusions of the Study

From the dataset, it was observed that mostly new customers are more likely to be defaulters.

The defaulters tend to recharge their accounts less frequently compared to that of the other customers.

* Learning Outcomes of the Study in respect of Data Science

The data set was large and contained many outliers. Several outlier and insensible data removal techniques were used.

The dataset being imbalanced, got the opportunity to work on balancing techniques SMOTE.

* Limitations of this work and Scope for Future Work

The company details like company set up date and mobile network initialization date would have given the analyst / scientist a better picture on many of the features and I need not have kept the dataset to the date of discovery of mobile phones.

The pcircle feature had only one value, that is one network circle, a wider data range would have given a better opportunity to predict the dynamics of people within different regions.

Data entry mistakes if reduced would have made the analysis process a lot easier.

In addition, additional data for defaulters would have enabled us to better predict the scenario without having to duplicate the imbalanced dataset.