

**Micro Credit Loan Project**

Submitted by:

Parth Rupavatiya

**ACKNOWLEDGMENT**

Some of the articles and research papers, I find useful for completion of this project.

REFERENCES:

1. Zoran Ereiz, “Predicting default loans using machine learning (optiML).” 27th Telecommunications forum TELFOR 2019.
2. Michael A. Turner, Patrick Walker,” Predicting Financial Account Delinquencies with Utility and Telecom Payment Data” PERC May 2015.
3. Philippe Bracke, Anupam Datta, Carsten Jung and Shayak Sen, “Machine learning explainability in finance: an application to default risk analysis”, Staff Working Paper No. 816.
4. Alexandru C., Monica M., Crisan ALBU, “PREDICTIVE MODELS FOR LOAN DEFAULT RISK ASSESSMENT”, Economic Computation and Economic Cybernetics Studies and Research, Issue 2/2019;

**INTRODUCTION**

**Business Problem Framing:**

* Micro Credit loan risk is one of the most important risks to be managed by financial entities. Without payback of loan there is no profit, hence the problem of credit risk management is relevant to all financial entities. Micro credit risk is economic loss that originates from the failure of the counterparty to fulfil its contractual obligations. For example, timely payment of interest.
* One of our client in telecom collaborates with a microfinance institution (MFI) to provide micro credit on mobile balances to be paid back in 5 days. The consumer is believed to be delinquent if he deviates from the path of paying back the loaned amount within 5 days.
* Here, we need to find whether the customer payback the micro credit loan in 5 days or not.

**Conceptual Background of the Domain Problem:**

* In this project, we have dataset of micro credit loan and purpose of this dataset is to predict the customer payback the micro loan within five days or not. Based on that prediction financial institution get to know about customers profile and decide to provide loan or not of particular customers.

**Review of Literature:**

* This project includes data cleaning, data analysis and predictive methods based on machine learning algorithms. The aim is to allocate a particular client a probability of default according to his profile and payment behaviour.
* Zoran Ereiz (2019) conducts a study on Predicting default loans using machine learning (optiML). AI and machine learning is used for create a credit risk management model. Different scoring models are introduced to evaluate certain parameters that could affect the loan payment.
* Michael A. Turner, Patrick Walker (2015) conducts a study on Predicting Financial Account Delinquencies with Utility and Telecom Payment Data. This study re-examines the data used for the 2012 PERC report, a new pathway to financial inclusion with a focus on the relation between bill payment behaviour on non- financial accounts and future delinquencies on financial accounts.
* Alexandru C., Monica M., Crisan ALBU (2019) conducts a research on predictive models for loan default risk assessment. The aim of this study is to analyse data from the Lending Club platform, which contains a number of clients who could not repay the credit in full, thus entering into default. The study was designed to apply a series of machine learning algorithms to develop four predictive models able to explain the studied event through classifiers such as: LightGBM, XGBoost, Logistic Regression and Random Forest.

**Motivation for the Problem Undertaken:**

A fundamental issue in the field of credit operations is assessment of client’s creditworthiness, based on his ability to pay back the loan. Customers with delay in pay back loan may lead to a high chance of default. Thus, companies need to examine their databases in order to discover patterns for customer behaviour and reduce the risk of micro credit loan. However, how these processes interact in customer’s non pay back loan probability remains a major challenge in terms of model performance validation and essential before model deployment so that it can be used in daily decision making process.

**Analytical Problem Framing**

**Mathematical/ Analytical Modelling of the Problem:**

* We use some mathematical, statistical and analytics approaches in this project. Which is described below:

1. **Synthetic minority oversampling technique (SMOTE):** In our dataset target variable is imbalanced and it provides inaccurate results during machine learning part. So, we used SMOTE method to cop up with that issue. SMOTE is an oversampling method and one of the most commonly used oversampling method to solve the imbalance classification problem. This method creates synthetic samples (not duplicate) of minority class. These synthetic records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class.
2. **Power-Transform:** In this dataset every attributes are skewed and skewness is affect our machine learning model so, it is necessary to remove skewness. For solving this issue we use power transform method. This method used to stabilize variance and make the data like normal distribution.
3. **Standard scaler:** After removing skewness, we need to scale our data. For this we use standard scaler method. This method normalizes our data and essential for machine learning algorithms that calculate distance between data. For instance, most of the classifiers calculate the distance between two points by the distance. If one of the features has large value, then distance consider that particular feature. This method is necessary, where large and small values present in our data. This method transform our data with mean = 0 and standard deviation = 1.
4. **Principle Component Analysis (PCA):** Principle component analysis is a technique of reduce the dimensionality of the variables. PCA is reducing the dimensionality of the large dataset, by transforming large set of variables into small set of variables. Reducing the number of variables affect our accuracy, But the trick is trade a little accuracy for simplicity. We use this method because smaller dataset are easier to explore and visualize, also make process faster for machine learning algorithms. So, idea behind the PCA is reduce the number of variables of a dataset, while preserving as much information as possible. We already have 32 independent variables but by using PCA we convert those 32 variables into 16 variables.

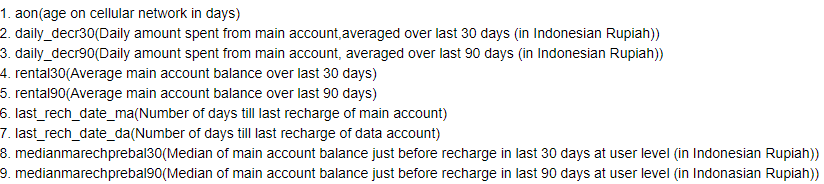
**Data Sources and their formats:**

* We obtain our data from the microfinance institution (MFI) of Indonesia through one of our client. The data is about 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 data set contains 209593 records and 37 variables. All the variables are related to customers finance regularity.

**Data Pre-processing Done:**

For cleaning or pre-processing the data we use some techniques which are described below:

* We check null values present in the data, but there is no null value found in the dataset.
* Drop unnamed: 0, msisdn, pcircle and pdate attributes, because this variables not useful for our machine learning model and don’t affect our target variable. Unnamed: 0 variable contain only index of the data, msisdn contain mobile no. of the user, pcircle contain only one value through the data and pdate contains dates.
* After that we check columns which have negative value and found that 9 columns have negative value which we can see below:



Above all columns can’t be negative so, we convert negative values with positive values.

* Then, we convert age on cellular network (aon) variable in days to years for ease in visualization and machine learning calculations.
* At last, we use SMOTE, power-transform, standard scaler and PCA techniques for pre-processing the data, which we already discussed.

**Data Inputs- Logic- Output Relationships:**

* We used classification machine learning models because our target variable loan delinquency is binary (0 and 1).
* There are many classification models but here we used some of them models.
* First we split our dataset into two segments: training and testing. We take 80% data for training and 20% data for testing. For splitting data we use train test split method. Below is the code for splitting the data:



1. 80% of the observation as training set-x\_train
2. The associated target for each observation in x\_train - y\_train
3. 20% of the observation as test set- x\_test
4. The target associated with the test set-y\_test.

* After splitting data we passed training data to machine learning models. The fitted model will first be used to generate prediction on the test set (x\_test). Next, the predicted class labels are compared to the actual observed class label (y\_test).

**Libraries Used:**

* We used many libraries used in this project, which is described below:

1. Numpy: This library is used for scientific computing. It supports multidimensional arrays and matrices.
2. Pandas: This library is used for data analysis and modelling convenient in python. Pandas simplify analysis by converting CSV, JSON, and TSV data files or a SQL database into a data frame with rows and columns.
3. Matplotlib and Seaborn: Both libraries are used for data visualization.
4. Imbalanced-learn: This library is used for imbalanced classification of target variable.
5. Scikit-learn: The Python library, [Scikit-Learn](https://scikit-learn.org/), is built on top of the Matplotlib, Numpy, and SciPy libraries. It has wide range of algorithms.

**Models Development and Evaluation**

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

1. Data preparation
2. Data cleaning
3. Data analysis
4. Over sampling
5. Handling skewness
6. Scaling
7. Principle component analysis
8. Train test split
9. Machine learning algorithms

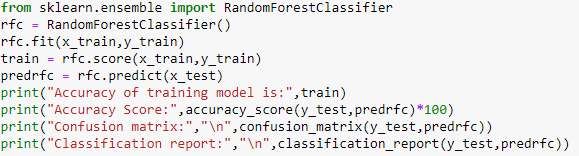
**Testing of Identified Approaches (Algorithms):**

The classification algorithm that we used is:

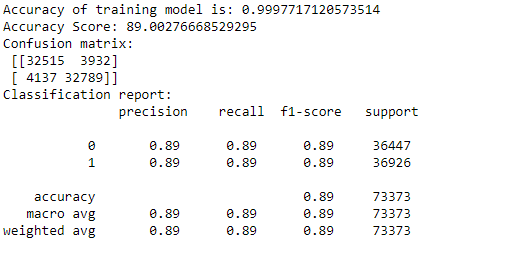
1. Logistic Regression
2. Decision tree classifier
3. GaussianNB
4. K-Neighbors Classifier
5. Random Forest Classifier
6. Gradient Boosting Classifier
7. Extra Trees Classifier
8. Catboost Classifier
9. Xgboost Classifier

**Building machine learning models:**

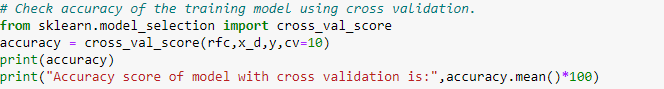
* We use many algorithms to find best model, but here we describe only best model.
* We find random forest classifier as a best model. It is supervised learning algorithm. This algorithm generates many individual trees, often hundreds or thousands.
* The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.
* This classifier produces multiple decision trees and merges them together to produce accurate result. Below is the code of our model with random forest classifier:



Output:



* We get 99% training model accuracy and 89% test data accuracy with best confusion matrix and classification report.
* Then, we check random forest classifier training model accuracy using cross validation to confirm that our model is not going through underfitting or overfitting. Below is the code for cross validation:



* Here, we use CV = 10, that means our training set is divided into 10 parts and provide mean accuracy of those 10 parts.

Output:



* We get 89% accuracy using cross validation that means our model is not underfitted or overfitted.
* Now, check accuracy of all used algorithms:

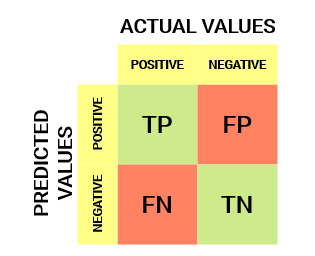


**Key Metrics for success in solving problem under**

**Consideration:**

* We use three types of metrics for solving problem. Which is described below:

1. Accuracy score: Accuracy score is a metric for evaluating classification model. It is calculated by number of correct prediction made divided by the total number of prediction made.
2. Confusion matrix: A confusion matrix is an nxn matrix used for evaluating performance of classification model, where n is the number of target classes. The matrix compares the target values with predicted values which are predicted by machine learning model. This matrix gives us a view of how good our classification model work and what kind of errors t is making. Below is the simple figure of confusion matrix.

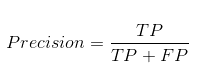


* If the model predicts that a client will enter default, and client really enters into default, then it is true positive situation (TP).
* When the model classifies a case as positive, but it actually negative, then it is a false positive observation (FP).
* If the model predicts that a client will not enter into default, and the client didn’t entered into default in the test set, then it is true negative situation (TN).
* The prediction of the model may be negative, while the client may actually enter into default, then it is false negative classification (FN).
* Below is the figure of our best confusion matrix with random forest classifier:



1. Classification Report: A classification report is used to measure the quality of predictions from a classification algorithm. This report shows the metrics like precision, recall and f1- score. The metrics are calculated by using true and false positives, true and false negatives.

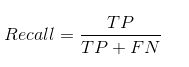
* Precision depicts how many of the correctly predicted cases actually turned out to be positive. Below is the formula of precision and calculation of precision with random forest classification model :



Precision = 32515/ (32515+3932)

Precision = 0.89

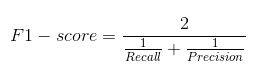
* Recalls depicts how many of actual positive cases we were able to predict correctly with our model. Below is the formula of recall:



Recall = 32515/ (32515+4137)

Recall = 0.89

* F1- score gives combined idea about precision and recall metrics. Below is the formula of f1-score:

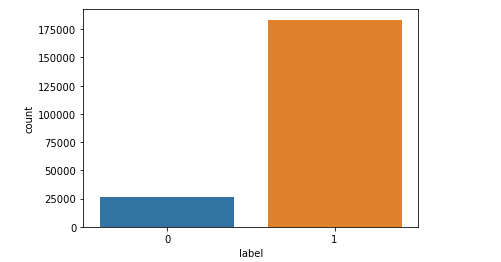


F1 –score = 2/ ((1/0.89) + (1/0.89))

F1- score = 0.89

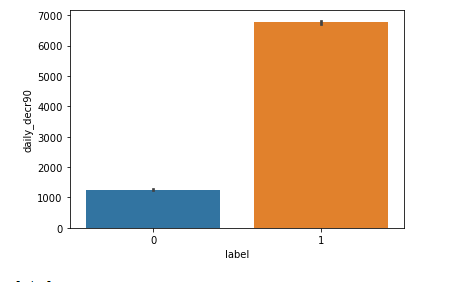
**Visualizations:**

* Now, let’s look at the analysis of the data.

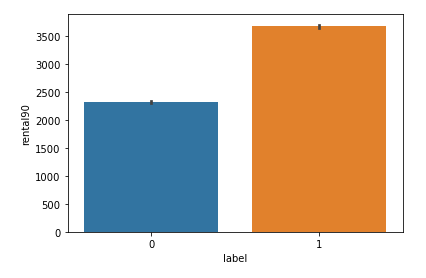


Label is our target variable, Where 0 depicts user does not able to paid back the credit amount within 5 days of issuing the loan and 1 depicts user paid back the credit amount within 5 days of issuing the loan.

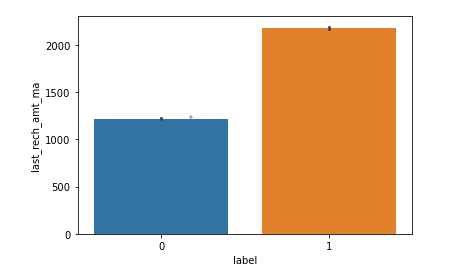
From the above plot, we can see that there is imbalanced classification of our target variable and for solve that issue we used SMOTE technique, which we already discussed.



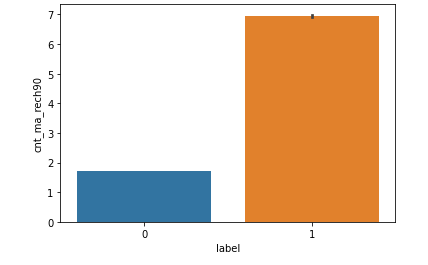
From the above plot, we can say that daily amount spent from main account, averaged over last 90 days is around 1200 Rs. for label 0 customers and around 6700 Rs. for label 1 customers (Balance is in Indonesian Rupiah). Customers who paid back loan within five days (Label 1) is spent more on daily basis than label 0 customers.



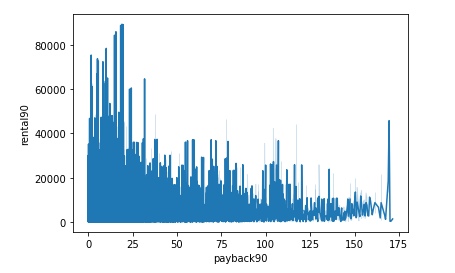
Above plot is average main account balance over last 90 days vs. label. Average main account balance over last 90 days is 2400 Rs. for label 0 customers and above 3500 Rs. for label 1 customers.



Above plot is average amount of last recharge of main account vs. label (0 and 1). From that we can say that average amount of last recharge of main account is approx. 1200 Rs. for label 0 clients and above 2000 Rs. for label 1 clients.



Above plot is average number of times main account got recharged in last 90 days vs. label. Average number of times main account got recharged in last 90 days is nearby 2 times for label 0 customers and nearby 7 times for label 1 customers. From above plot we can say that clients who pay back loan within 5 days are spent more from their main accounts and do recharge frequently of their main accounts compared to the label 0 customers.



Above plot is average main account balance over last 90 days vs. average pay back time in days over last 90 days. Customers with high main balance account took less time for paying back loan and vice versa. Customers with low balance take more time or not paying loan within 5 days creating major loss to the entities.

**CONCLUSION**

**Key Findings and Conclusions of the Study:**

* The purpose of this article was twofold: to understand the pattern of micro credit loan risk and make predictive model able to effectively classify observation in the two classes, i.e. customers pay back loans within 5 days and customers fail to pay back loans in 5 days.
* Our target variable is imbalanced, so we use SMOTE technique to get a good result.
* We use many classifiers to find best model and best result were observed for the random forest classifier with 89% accuracy.
* In order to decrease loss of the company, the company should start using of machine learning model and start some marketing strategies like notification or sms on mobile to pay back the loan within 5 days.

**Future Work:**

For future work, the model should be further improved by trying to prepare even better dataset (perhaps by selecting only specific and relevant attributes). Moreover, try other sampling technique to get best results.