

**Micro -Credit Defaulter Model**

Submitted by:

PUSPRAJ KUMAR

**ACKNOWLEDGMENT**

I mentioning of all the references, data sources give is given to me by Flip Robo, towardsdatascience.com website and Marco Peixeiro YouTube channel resources that helped me and guided me in completion of the project.

**INTRODUCTION**

* Business Problem Framing

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. Microfinance services (MFS) becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The MFS provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

MFI are working with one such client that is in Telecom Industry from Indonesia. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

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).

* Conceptual Background of the Domain Problem

Firstly, I insert the data set ‘Data file .csv’ for performing this project and after that I proceeded step by step. I dropped the ‘unnamed’ column as to why these columns remain or not, will not put any effect on the result, check the null values have or not and plot by heatmap, in this data set the outlier have been tested through function such as boxplot, scatterplot etc. In this, I have shown the data correlation by heatmap, that means how much columns have been correlated with each other.

Through the count plot function have been show how much customer are defaulters and non-defaulter. After doing these all thing I did drop the ‘pcircle’ columns, because this column does not affect our result.

* Review of Literature

I have selected the classification model for this data set because the label column has the form of values is ​​0 and 1. First of all I have only used several classification models for testing only accuracy scores.

After this, I have only selected the model with the highest accuracy score, for the different random state. After selecting, I used Remaining Evaluation metrics on the Particular model such as Confusion matrix, Classification report and auc\_roc. I have not used any hyper tuning method for, because of predictive accuracy value as probably good.

* Motivation for the Problem Undertaken

Inspiration for the problem because of the I have also wanted to help some people in a remote area, and by the way my roots also belong to the remote area. so that's why for me motivation of this problem.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In this project I have been using the evaluation matrix-like accuracy score, confusion matrix, classification report and roc\_auc. I am going explain mathematically formula.

Accuracy score=No. of correct prediction/total no. of input sample.

TP +TN/TP+TN+FP+FN

=43985+41230/43985+41230+2752+80

= 0.9678

Confusion matrix:

Precision = TP/TP + FP

= 43985/43985+2752

=0.9411

Recall=TP/TP+FN

=43985/43985+80

=0.9981

Specificity =TN/TN+FP

= 41230/41230+2752

=0.9370

And also, I have used VIF for showing highly correlation feature.

* Data Sources and their formats

Data Sources are given by Flip Robo Technology, in .csv format. This data set very easy to understand because all feature is described in the very well-mannered and that column are clearly mentioned which have to use for the testing purpose clearly mentioned. The name of column is 'label'.

* Data Pre-processing Done

In this data set have no null value.so I have checked with the heatmap function. I have also drop those columns which are not use full. Like ‘unnamed’ and ‘pcircle’ after dropping these columns not effect on our result. After doing these all thing I have to move to the EDA process for showing easily outlier, correlation etc.

* Data Inputs- Logic- Output Relationships

For this data set, I have used the classification model because target values are in the form of 0 and 1. However, the classification model converts row data to numeric form but in this data set, the target column value is already converted to 0 and1.

If any columns we have drop from this data set then it affects the output result.

* Hardware and Software Requirements and Tools Used

I have done hardware window 10 operating with 4 GB ram and 1 TV memory space in this project. I used Jupiter notebook software to complete this project. In Jupiter Notebook has Multiplot, Seaborne, Pandas used for this project. For understand data set I have also use excel.

**Model/s Development and Evaluation**

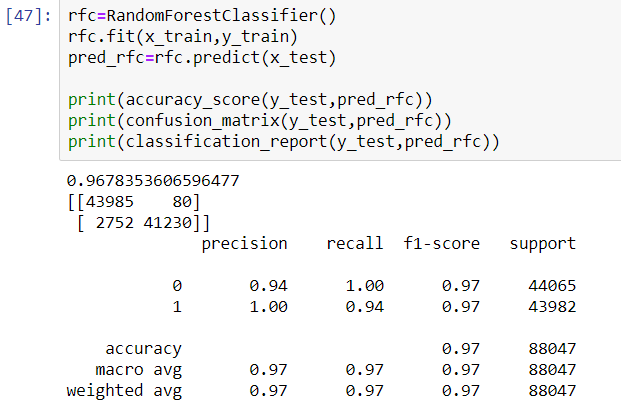
* Identification of possible problem-solving approaches (methods)

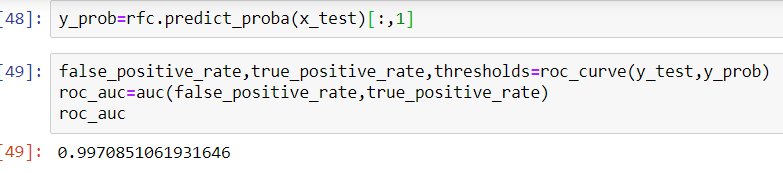
For solving this problem, I have followed these steps for statistical and analytical is, these customers are defaulter or non-defaulter. Checking these things by the help of multiple classification model that I used Logistic regression, Decision Tree, Random forest and GaussianNB. From these model I have check only accuracy score and after this those model give maximum accuracy score for a particular random state then for those model I used other evaluation matrix like confusion matrix, classification report.

* Testing of Identified Approaches (Algorithms)

I have used classification algorithms for training and testing is Logistic Regression, K-Neighbours classifier ,Decision Tree Classifier and Random Forest Classifier.

* Run and Evaluate selected models





* Key Metrics for success in solving problem under consideration

For this data set I have used the classification models and for classification model we use evaluation matrix is accuracy score, confusion matrix, classification report, and auc\_roc is used.

* Visualizations

Firstly, I have use data visualization box plot, scatter plot from matplotlib, cat plot, dist plot, count plot from seaborn. Generally, we use a box plot for plotting outlier, cat plot can be used for multiple kinds like a bar, scatter, violin, box, boxen, point. Count plot counts all feature that has any particular column.

* Interpretation of the Results

In this project I have used these things for visualization, pre-processing and modelling. I have used EDA for visualization I have used Eda for visualization in which for performing EDA used library like pandas, matplotlib, seaborn. In pre-processing I have checked in pre-processing, I have checked how many types of data have been in this data set, what is the size of the data set file, what is the shape of the data set and through Describe method, I have distributed all number type columns.in these, all model random forest classification model has given best result on 40 random state. Accuracy score is 0.967835360659647 the best score for this data set.

**CONCLUSION**

* Key Findings and Conclusions of the Study

I have used EDA for all the features. On mobile number user feature I have used encoding because this feature data type which is object type for the next process it needs to convert, so I did. I have also used this here because I wanted to get my Mathematically show done to the columns who are highly correlated.

* Learning Outcomes of the Study in respect of Data Science

I have learnings obtained about the power of visualization, data cleaning and various algorithms used something like that, box plot, scatter plot, cat plot, distribution plot, count plot etc. In this data set have not duplicate values and no null values so for this data set data cleaning is not perform. I have used label encoding, variance inflation factor (VIF), principle component analysis (PCA), standard scaler and resample the data set.

* Limitations of this work and Scope for Future Work

The limitation for using this data set is that if we do not use resampling in it, then we will get an accuracy score almost exactly, but cannot provide a good score for recall and precision. I have used these algorithms which give the best result and after I have used hyper tuning methods but the result after the hyper tuning accuracy has not been affected.

I followed the steps:

Check the null values have or not.

Check the correlation

Used PCA, Standard Scaler, resampling for get better result.

Check the how much customer are defaulter and non-defaulter.

After doing these things go for testing.