

**MICRO CREDIT PROJECT**



Submitted by:

PRADEEP KUMAR SINGH

**ACKNOWLEDGMENT**

In preparation of this Project, I had to take the help and guidance of some video tutorials, who deserve my deepest gratitude. As the completion of this Project gave me much pleasure, I would like to show my gratitude Kris Naik videos on YouTube, articles on Kdnuggets and medium for giving me a good guidelines for this Project throughout start to end using some sample data. I would also like to expand my gratitude to all those who have directly and indirectly who helped me to use wide thinking to explore here in dataset.

In addition, a thank you to my mentor Astha Mishra, who introduced me to the Methodology of work, and who support always wherever I stuck ever. I also thank FlipRobo Technologies for providing such an opportunity to work on various types of projects which gradually improve my vision to apply for the datasets.

**INTRODUCTION**

Business Problem Framing

The low income families in rural and slums area have always the issue to fulfilling the needs they have and the abilities to complete them sometimes happen and sometimes not. And the reason is non other than finance. A Microfinance Institution is an organization that offers financial services to these low income populations. The financial services becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income.

Conceptual Background of the Domain Problem

Microfinance is a category of financial services targeting individuals and small businesses who lack access to conventional banking and related services. Microfinance services are designed to reach excluded customers, usually poorer population segments, possibly socially marginalized, or geographically more isolated, and to help them become self-sufficient.

Review of Literature

Microfinance was defined initially as the provision of microloans to poor entrepreneurs and small businesses lacking access to credit. The two main mechanisms for the delivery of financial services to such clients were: (1) relationship-based banking for individual entrepreneurs and small businesses and (2) group-based models, where several entrepreneurs come together to apply for loans and other services as a group. Over time, microfinance has emerged as a larger movement whose object is: *"a world in which as everyone, especially the poor and socially marginalized people and households have access to a wide range of affordable, high quality financial products and services, including not just credit but also savings, insurance, payment services, and fund transfers.”*

Motivation for the Problem Undertaken

The Project here is also one of the data collected by Micro finance company with respect to telecom industry. The data presented the various types of customers taking the loans from the company who paus on time and some are not paying it and never even taking an excuse. So using the above data, I have to see the person is defaulter or non defaulter by exploring his previous 30 and 90 days record represented from the data statistics.

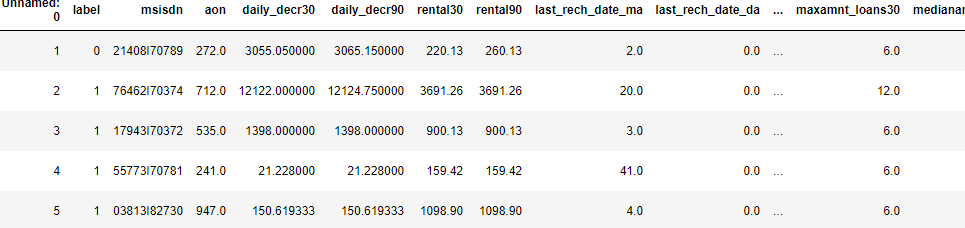
**Analytical Problem Framing**

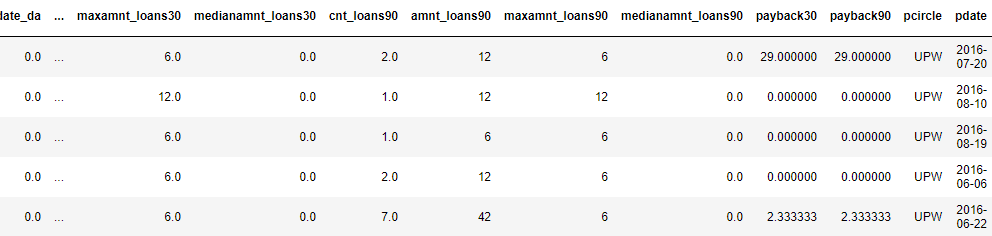
* Mathematical/ Analytical Modelling of the Problem

The whole dataset has checked for the null value at first and checked for the datatype of each feature because sometimes it seems to be a numeric value but appears as a object datatype. The target column which is ‘Label’ here shows the data which is imbalanced. On displaying the statistical summary for the data, it is observed that each and every feature here has the outliers so applying z-score to cut it off.

* Data Sources and their formats

I received the dataset for this project through my mentor on this project and is in the .csv file. Some of the info such as series number, msisdn are seems unusable with respect to the problem. When we have the data for 30 & 90 days previous activities then “Date” column seems to be not necessary.Here are some of the screenshots of data.





* Data Preprocessing Done

Some of the columns that are not adding any value to the predicted variable are dropped off such as Unnamed:0,msisdn,pcircle,pdate.On verifying the correlation among the data, some of the feature seems to be highly correlated through the data,so I drop it to remove the ambiguity while operations.

* Data Inputs- Logic- Output Relationships

The output of the data is completely depends on the input we provide. So,its best the scale the input correctly and remove the clutter to get the data that adds value to the prediction.

* Hardware and Software Requirements and Tools Used

Hardware And Software required for this project:

* Laptop with I5 processor and 4 GB Ram
* Anaconda (Jupyter Notebook)
* Python Pandas for processing
* Scikit learn library and imblearn for model deployment

**Model/s Development and Evaluation**

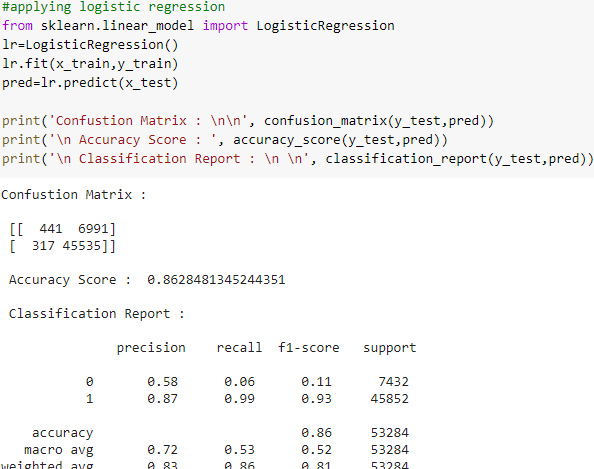
* Identification of possible problem-solving approaches

Before taking any further steps, I checked for the datatype first and then dropping some unnecessary columns. On displaying the statistical summary for the data, it shows the huge number of outliers on all the features, so applying z-score to minimise it. The label ratio is showing that data is imbalanced. So for one attempt, I checked by Linear model and then applying other models to make the prediction better with some hyperparameter tuning.

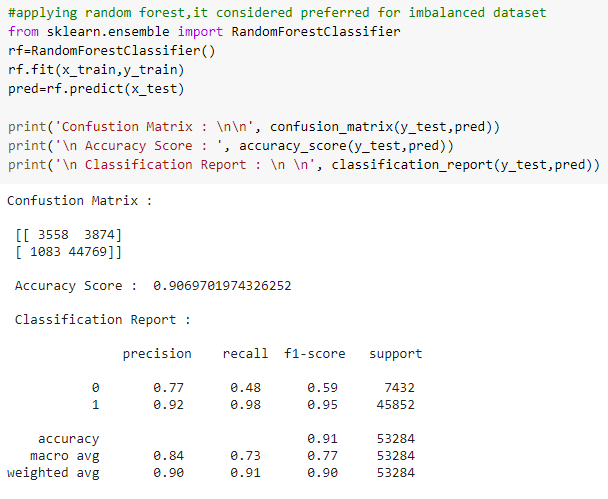
* Testing of Identified Approaches (Algorithms)

Following are the algorithms that I applied on this dataset

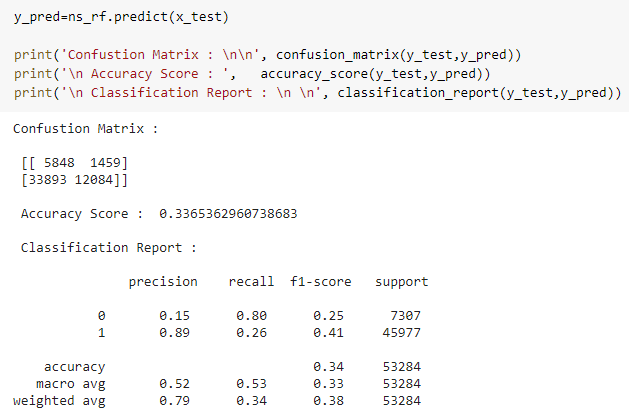
* Logistic Regression
* Random Forest without hyperparameter tuning
* Random Forest with undersampling
* SMOTEtomek
* Snipped Screenshot of Applied algorithms



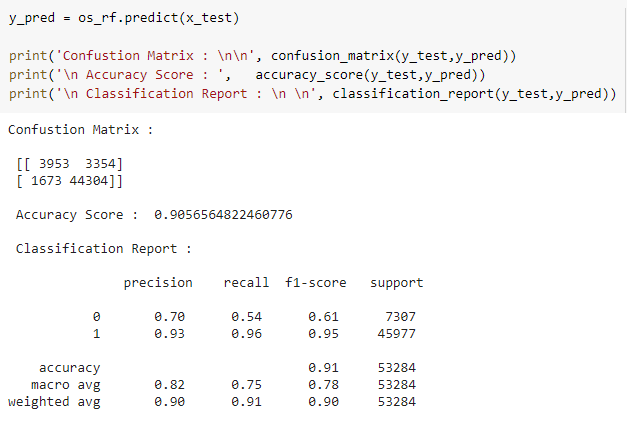
Logistic Regression



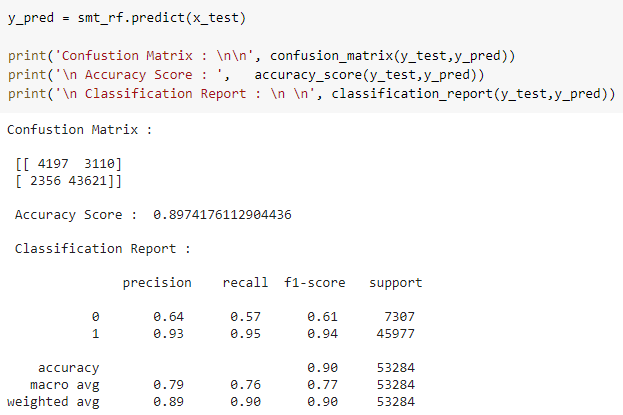
Random Forest without tuning



Random forest with under-sampling



Random Forest with Oversampling

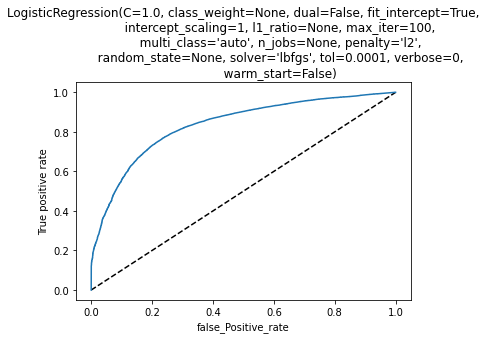


SMOTETomek

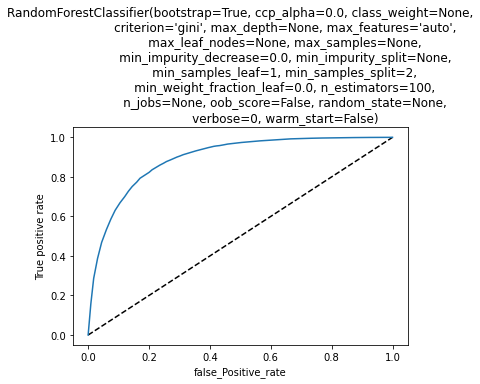
* Key Metrics for success in solving problem under consideration

As it is clear from above observations that we have bias in the data labels or imbalanced data. So the performance metrics I consider for this data on all the applied algorithms are F1-score, Precision, Recall and Roc-Auc Score.

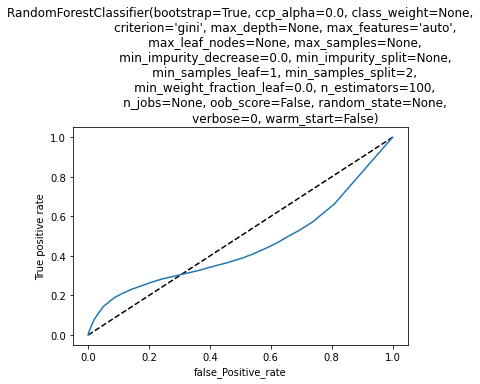
* Visualizations



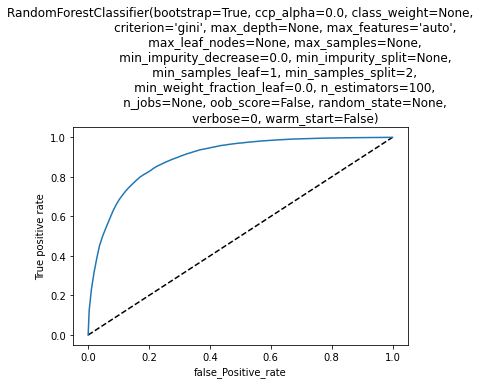
The above is the representation of roc-auc curve of Logistic Regression having score of 0.52.This is not a good model to deploy. Precision and recall value is very low too.



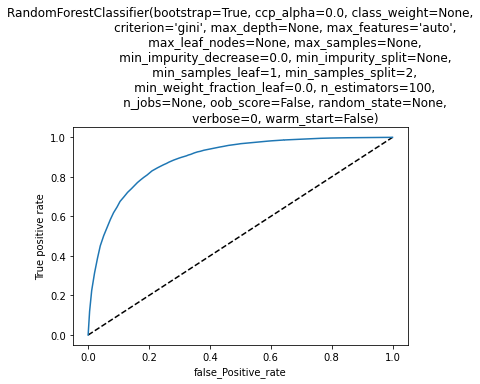
This curve shows the roc auc of Random Forest without any hyperparameter tuning with a score of 0.76.Precision and Recall are also upto the mark.



The above graph is of under-sampling characteristics and it don’t need any explanation why it is not a good model to work. It has score of 0.53,precision of 0.15 and Recall of 0.25.



The oversampling graph above show the roc auc curve of oversampling and has a very good score of 0.72 throughout all the roc, precision and recall value.



The SMOTE has too a goof roc-auc curve and roc score of 0.76 and has fair precision and recall value.

* Interpretation of the Results

Oversampling and SMOTETomek are really good performing model according the above observations done. Random Forest without any hyperparameter tuning gives such a good results also shows it is also above the acceptance margin.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Some of the observations I infer from the whole project are as follows:

* The Same data is presented on these dataset using various different features but shows high correlation and can be processed using only one features. Some of these features are “Date”,”rental30”,”rental90”.
* Heatmap of correlation of all the features comprising together shows fr\_da\_rech30 and fr\_da\_rech90 are not proper features to used for predictions.
* The value counts of target variable clearly shows the given dataset is imbalanced and in the improper ratio.
* Applying normal algorithms without any hyperparameter tuning and under-sampling don’t perform well and give biased result.
* Learning Outcomes of the Study in respect of Data Science

It is clear from data processing and after statistical summary that the whole dataset is imbalanced in nature. And it need a little regularisation or operations to handle it all. We first check it using Linear model and random forest, on this the random forest performs well even without any tuning. So it better to process the random forest with some tuning with oversampling and Under-sampling. Oversampling gives the best performance over under-sampling. SMOTE Tomek is also preferable model as per result and gives good performance.