

Micro-Credit

Submitted by:

Preeti Kumari

**ACKNOWLEDGMENT**

I am really grateful to FlipRobo Technology for providing me with this project. I am also thankful to ur mentor Mr. Harsh Ayush for providing us the guidance to complete the project.

Also, below are some blogs which helped me to work on the project in a better way.

* <https://stackabuse.com/applying-filter-methods-in-python-for-feature-selection/>
* <https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/>

**INTRODUCTION**

* Business Problem Framing

Telecom Industry is 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

We need to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan

* Conceptual Background of the Domain Problem

Microfinance allows people to take on reasonable small business loans safely, and in a manner that is consistent with ethical lending practices

The majority of micro financing operations occur in developing nations, such as Uganda, Indonesia, Serbia, and Honduras.

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.

A Lot of microfinance institute also got together with telecom partners to provide financial assistance to people with less income and for doing that they keep into account various details as last recharge done, last data recharge, maximum amount of loan taken, number of days till last recharge of main account, frequency of data account recharged and many more and on the basis of this criteria , the organisation decided if the will be paid within 5 days or not.

* Motivation for the Problem Undertaken

This project is a part of our curriculum and will provide us with more insights to how can we solve business problems using data science.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The statistical method used was correlation, absolute method, and standard scalar to scale all the input variables at same level for better output. Also performing under sampling on output to avoid over fitting.

* Data Sources and their formats

Data source: The sample data is provided to us from our client database.

The data has 209593 rows and 36 columns, and most of the data features are of integer or float data type.

The target or to be predicted column’s name is “label” which has two values 0 and 1, Label ‘1’ indicates that the loan has been payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

Data Pre-processing Done

1. Dropping column named "Unnamed: 0" as it contains index number only.
2. Converting date from object data type to date-time and then splitting into year and month.
3. Created a function with only one unique value throughout the column and dropping that column as it won’t be of any good,so dropped column ‘pcircle’ and ‘year’.
4. Created a function with creating a loop to find out the highly correlated variables and then drop them as the data then will be redundant we have set the correlation value as 0.97 ,if two columns have correlation more than this then that will be printed. Dropping those columns as it wont be of much use.
5. Some of the columns have negative columns which need to be treated as no negative value should be present, e created a for loop for the same and used absolute function to get the negative value converted to positive.

* Data Inputs- Logic- Output Relationships

.

Here we used correlation function to find the relation between the input columns and the output columns.

A positive correlation suggests with the increase in the value of a particular feature the output increases, whereas negative correlation suggests with the increase of input feature, the output decreases.

Most of the columns have positive correlation except columns like “fr\_da\_rech30”,” aon”, “medianmarechprebal30” and ”fr\_da\_rech90”

* Hardware and Software Requirements and Tools Used

We have made the project on Jupyter notebook.

The libraries used are pandas, numpy, matplotlib, seaborn.

1. Pandas are used
2. Matplotlib- This library helps us to build multiple plots at a time.
3. Scipy is an open-source python library that is used for both scientific and technical computation.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

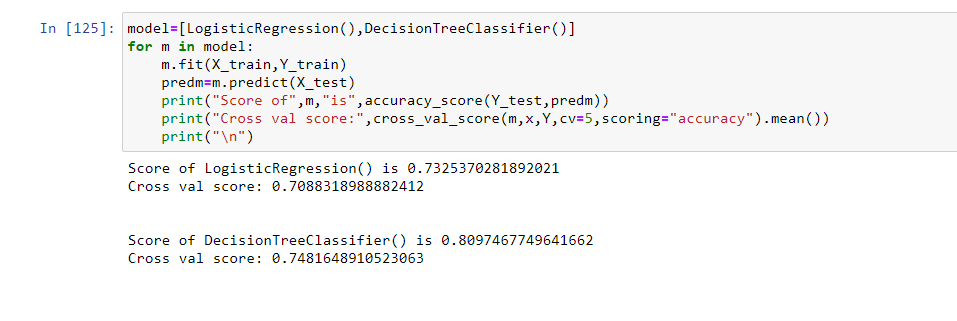
We split the data into x and y , where x contains all the input features and y has output column.

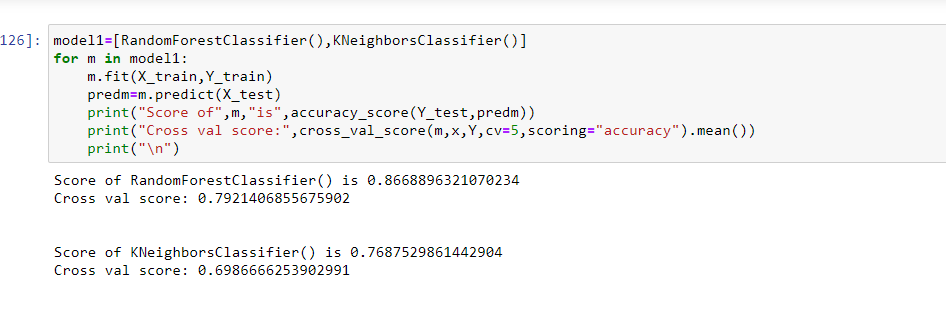
The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records so to avoid overfitting we used Under-Sample which will make the dataset balance.

Standard Scalar- This will help o bring down all the columns of x at the same scale and will be useful for prediction.

We are running a for loop to run various models to find out which is better, we have split the model into two for loops so that it will

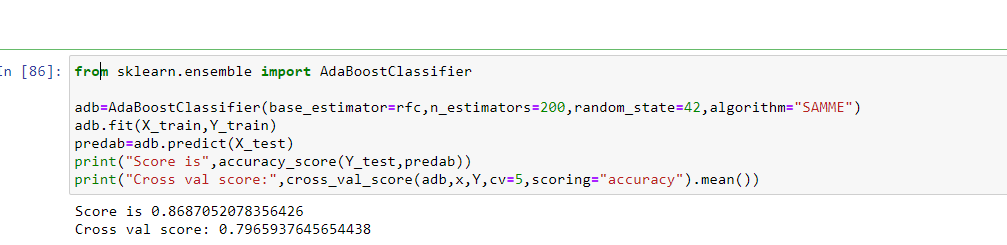
Run faster.





We have used four classifiers and then choosing Random Forest Classifier as it has the best score.

Also to increase the score we used Ada boost classifier. Ada Boost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier.



We chose base estimator as rfc which was defined as Randm Forest Classifier followed by training the dataset, predicting the trained dataset and then printing the scores.

* Run and Evaluate selected models

For evaluating the model we have used confusion matrix and classification report

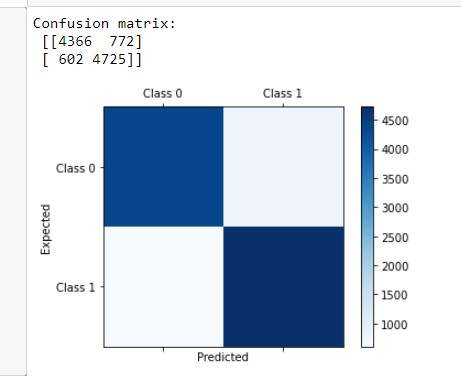
Confusion matrix consist of True positive, true negative, false positive and false negative.

True Positives (TP) - These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes. Here the count is 4366.

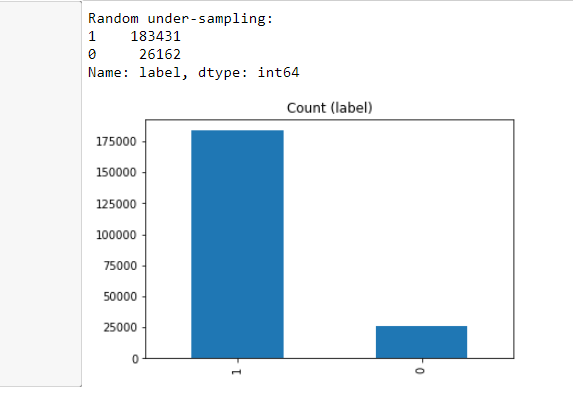
True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also. Here the count is 4725.

False Positives (FP) – When actual class is no and predicted class is yes. The count here is 2.False Negatives (FN) – When actual class is yes but predicted class in no. The count is 772.

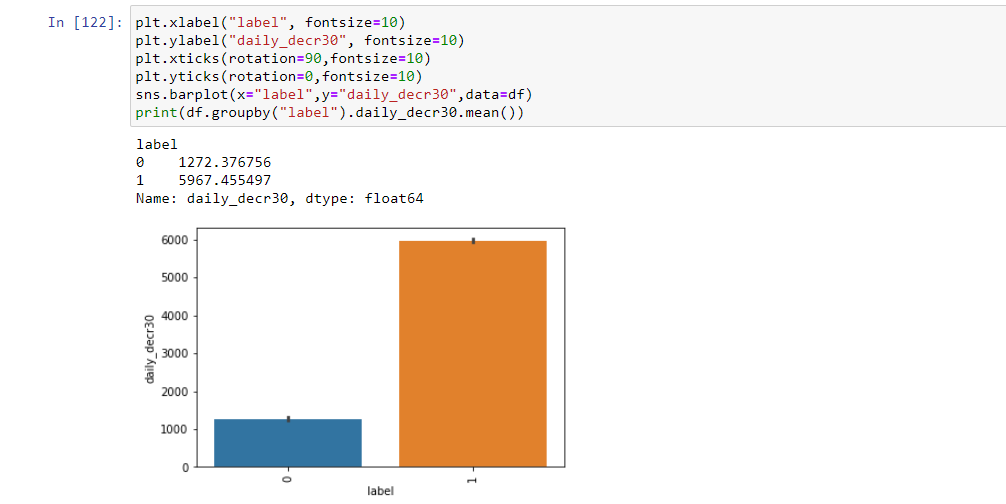
False Negatives (FN) – When actual class is yes but predicted class in no. The count is 602.



* Visualizations

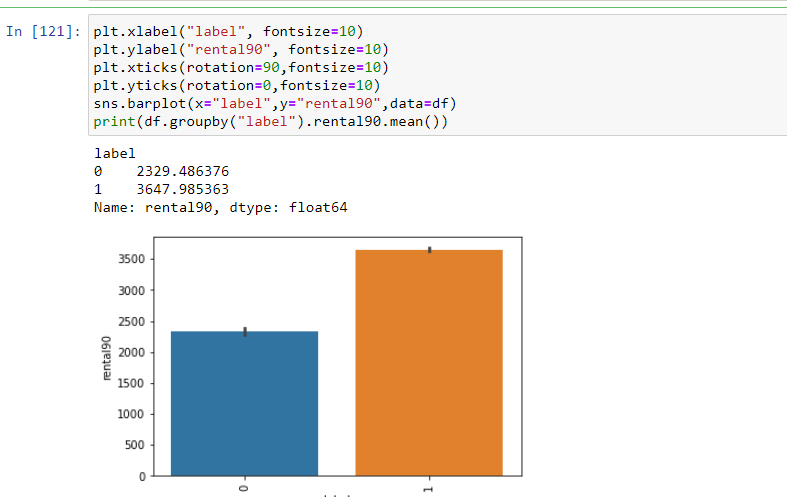
1. 

We visualised the label column and can see that the data is highly imbalanced.



Observation:

The people who have daily amount spent from main account frequently have paid the loan more easily.

3.

Observation:

The people with more average main account balance over last 90 days have higher capacity of paying the loan easily.

4.

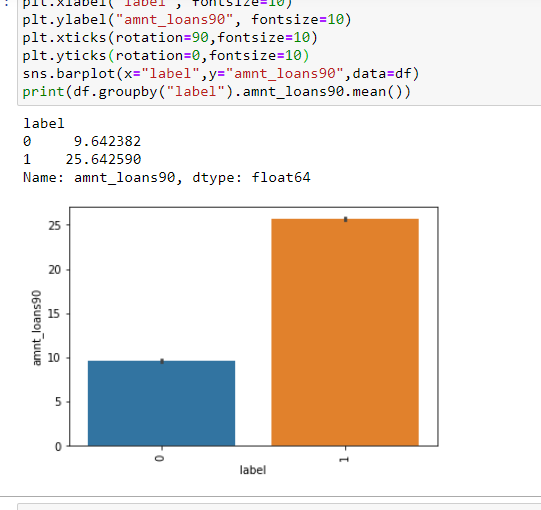


5



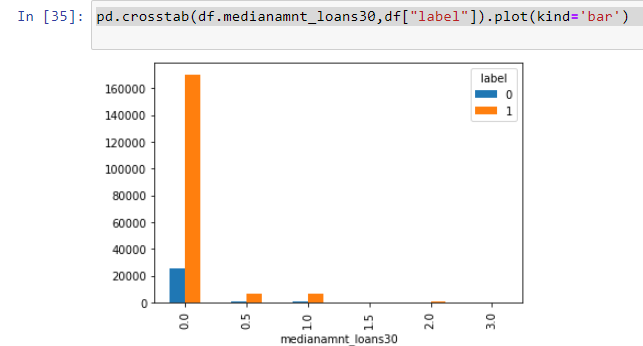
Observation:

The one paying loan and the one not paying have both have same frequency for the maximum amount of loan user took as a lot of people had not taken any loan.

6.

Observation:

This depicts that people who have taken loan within the last 90 days are more prone of returning the loan within the required time.

7.

Observation:

For the one paying loan easily, the median amount of loan taken within 30 days is more.

Interpretation of the Results

Visualization - With the help of visualization we understood that there are few columns which are highly related to the output and there are some columns which are not that much effected.

Pre-processing- With this we dropped few columns which won’t be useful for our model building, also we treated all the negative values followed by dropping which were highly correlated and then scaling input columns.

**CONCLUSION**

* Key Findings and Conclusions of the Study

1. The target variable was indeed depended on lots of columns and change in one could affect the other.
2. Visualizing helped us to provide a better understanding of the relationship.

* Learning Outcomes of the Study in respect of Data Science

The only problem faced during the assignment was that losing of data as it was the database provided by the client so we could not lose data as such, but to avoid we consider most of the columns and dropped columns only with redundant values and with just one unique value as it won’t be much useful.

Working with large dataset was also a new experience.

* Limitations of this work and Scope for Future Work

After working on this large dataset we can work on various dataset with large also learnt how to work on imbalanced dataset as working on imbalanced dataset will result in overfitting.