

NAME OF THE PROJECT

MICRO-CREDIT-PROJECT

Submitted by:

HARSHITA SEHGAL

**ACKNOWLEDGMENT**

<https://elitedatascience.com/imbalanced-classes>

From this site i was able to understand the concept of imbalanced classes and how to use that to my project

And other sites such as medium,towards data science were also a great help for the purpose of reference

Apart from this mentors of fliprobo guided me with the concepts and motivated me to complete this project .

**INTRODUCTION**

* Business Problem Framing

Herein, we have dataset of telecom industry they have several plans for users. Their agenda is to provide these services to everyone including the lower income group so they have introduced loans in collaboration with MFI ,A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on,now with respect to telecom industry they payback period has been set up as 5 days,if a user fails to pay the loan then he is a defaulter.In real life scenario ,this problem is a logical approach to handle the classification between a defaulter and non defaulter,this problem becomes more special and unique because it’s just not simply a credit loan problem,but it is more associated with telecom industry and the loan here is for communication purpose( the mobile data packages),So surely this is a smart move in this era wherein communication plays a vital role .

* Conceptual Background of the Domain Problem

Talking about the domain, main domain here is the financial domain as the main focus is credit (loan taken for mobile plans).Also this data belongs to telecom industry,So their plans and usage frequency of recharge done by the user and all such features are included in the dataset,so some vital knowledge regarding telecom sector is also required.

* Review of Literature

.

Analysing the dataset,

shape of dataset is as follows (209593, 36)

Almost all are numerical values (int and float)

No null values are present

Target variable is label

Ratio among them is- counts for label 1 is 183431 and counts for 0 label is 26162(ratio 87.5% and 12.5%)

So I have to solve a classification problem which is imbalanced and have presence of outliers,almost all the columns apart from pcircle and phone number are of vital importance

* Motivation for the Problem Undertaken

.

All thanks to the mentors and team of fliprobo who were present throughout the duration of project and provided me an opportunity to try my hands on this particular problem statement which made my logics and clarity about the credit (finance) how it works for a telecom industry and identifying the defaulters.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This problem is a classification problem,the target variable is itself a stastistical parameter.we have 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. In this case, Label ‘1’ indicates that the loan has been payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed .for a loan amount of 5 payback amount should be 6,and for loan amount of 10 payback amount is 12.

* Data Sources and their formats

The source of the data is from a telecom industry having 36 columns and 209593 rows,the data was in a excel file for which i saved a csv copy and uploaded that on my jupyter notebook,

Dtypes,

label int64

msisdn object

aon float64

daily\_decr30 float64

daily\_decr90 float64

rental30 float64

rental90 float64

last\_rech\_date\_ma float64

last\_rech\_date\_da float64

last\_rech\_amt\_ma int64

cnt\_ma\_rech30 int64

fr\_ma\_rech30 float64

sumamnt\_ma\_rech30 float64

medianamnt\_ma\_rech30 float64

medianmarechprebal30 float64

cnt\_ma\_rech90 int64

fr\_ma\_rech90 int64

sumamnt\_ma\_rech90 int64

medianamnt\_ma\_rech90 float64

medianmarechprebal90 float64

cnt\_da\_rech30 float64

fr\_da\_rech30 float64

cnt\_da\_rech90 int64

fr\_da\_rech90 int64

cnt\_loans30 int64

amnt\_loans30 int64

maxamnt\_loans30 float64

medianamnt\_loans30 float64

cnt\_loans90 float64

amnt\_loans90 int64

maxamnt\_loans90 int64

medianamnt\_loans90 float64

payback30 float64

payback90 float64

pcircle object

pdate object

dtype: object

df.describe() shows a lot of vital info

df.corr() also shows the most correlated variables with target variable here our the results,

cnt\_ma\_rech30 0.237331

cnt\_ma\_rech90 0.236392

sumamnt\_ma\_rech90 0.205793

sumamnt\_ma\_rech30 0.202828

amnt\_loans90 0.199788

amnt\_loans30 0.197272

cnt\_loans30 0.196283

daily\_decr30 0.168298

daily\_decr90 0.166150

medianamnt\_ma\_rech30 0.141490

last\_rech\_amt\_ma 0.131804

medianamnt\_ma\_rech90 0.120855

fr\_ma\_rech90 0.084385

maxamnt\_loans90 0.084144

rental90 0.075521

rental30 0.058085

payback90 0.049183

payback30 0.048336

medianamnt\_loans30 0.044589

medianmarechprebal90 0.039300

medianamnt\_loans90 0.035747

cnt\_loans90 0.004733

cnt\_da\_rech30 0.003827

last\_rech\_date\_ma 0.003728

cnt\_da\_rech90 0.002999

last\_rech\_date\_da 0.001711

fr\_ma\_rech30 0.001330

maxamnt\_loans30 0.000248

fr\_da\_rech30 -0.000027

aon -0.003785

medianmarechprebal30 -0.004829

fr\_da\_rech90 -0.005418

* Data Preprocessing Done

Now before moving towards splitting my data into training and testing,I observed abnormalities in my dataset like for example if we talk about daily recharge a user has done in last 30 days,daily\_decr30 column we have values as low as -93 and as high as 2,00,000 whereas the average range is somewhat between 1000 to 10,000 so there is a huge gap between these numbers and similarly the same applies to other columns too,

Other than this msidn,pdate has no such relevance msidn is the phone number and pdate also holds no relevance because the data itself is talking with relation to no of days.so i have dropped these two columns

And applied normalisation (normalizer) and scaling the data so that i can have a uniform data.I also applied z score but that resulted in loss if data as the data shape changed to .So i will prefer normaliser method as data is expensive and loosing this much data will result in loss of valuable information.Other thing I faced was my dataset was highly imbalanced

Normalisation outliers imbalanced dataset

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

Now my output is the label class and rest 32 columns serve as the input,as discussed this is a classification problem with two classes 1 and 0( 0 is defaulter unable to pay the loan within 5 days and 1 is non defaulter that is the user had payed the loan),so now this depends on these 32 columns,comprising of recharge(usage) and loan taken and payback that is the days in which it has been paid back,and other columns are somewhat extension to these like median,frequency,counts etc.

* State the set of assumptions (if any) related to the problem under consideration

No such assumptions,rather i have 3 approaches to solve this problem which are discussed here,

Because of imbalanced dataset,and abnormalities present I have applied the metrics three times,first before upsampling and downsampling,then second after applying upsampling and down sampling for the dataset after transforming and scaling(normaliser) and third for dataset after applying z score.As my dataset is imbalanced 1st approach was a fail with accuracy as high as 80 and roc auc as 50 whereas second and third approach gave a good accuracy of 75 and roc auc score as 74 ,The detailed predictions regarding all the metrics are discussed ahead in this report.

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Jupyter notebook(Anaconda framework)

Libraries used are listed below-

For eda purpose-

Matplotlib and seaborn

And the common library pandas, numpy

For preprocessing

from sklearn.preprocessing import Normalizer

For evaluation metrics

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import confusion\_matrix,classification\_report

from sklearn.metrics import accuracy\_score

from sklearn.naive\_bayes import MultinomialNB

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import roc\_curve

from sklearn.metrics import roc\_auc\_score

**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.

Because of imbalanced dataset,and abnormalities present I have applied the metrics three times,first before upsampling and downsampling,then second after applying upsampling and down sampling for the dataset after transforming and scaling(normaliser) and third for dataset after applying z score.As my dataset is imbalanced 1st approach was a fail with accuracy as high as 80 and roc auc as 50 whereas second and third approach gave a good accuracy of 75 and roc auc score as 74 ,The detailed predictions regarding all the metrics are discussed ahead in this report.

* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

Algorithms used are as below-

 Logistic Regression

 gnb

 Ada boost classifier

 K-Nearest Neighbours

 Decision Tree

 Random Forest

* Gradient boosting classifier
* Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

Logistic Regression

It is the mostly used algorithm across industry ,we will import libraries such as Logistic Regression from sklearn.linear\_model, importing confusion matrix and classification report we get the following results

from sklearn.linear\_model import LogisticRegression

lg=LogisticRegression()

lg.fit(train\_x,train\_y)

pred=lg.predict(test\_x)

print("accuracy\_score:",accuracy\_score(test\_y,pred))

print(classification\_report(test\_y,pred))

accuracy\_score: 0.7461184526145929

precision recall f1-score support

0 0.74 0.75 0.75 45713

1 0.75 0.74 0.75 46003

accuracy 0.75 91716

macro avg 0.75 0.75 0.75 91716

weighted avg 0.75 0.75 0.75 91716

2. Decision Tree Classifier

A decision tree classifier is a tree in which internal nodes are labelled by features,We have two criterion ‘gini’ and ‘entropy’

Importing DecisionTreeClasifier from sklearn.tree

Here are the results

dct=DecisionTreeClassifier()

dct.fit(train\_x,train\_y)

preddct=dct.predict(test\_x)

print(classification\_report(test\_y,preddct))

precision recall f1-score support

0 0.91 1.00 0.95 45713

1 1.00 0.91 0.95 46003

accuracy 0.95 91716

macro avg 0.96 0.95 0.95 91716

weighted avg 0.96 0.95 0.95 91716

3.

from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier()

rf.fit(train\_x, train\_y)

rf\_predict=rf.predict(test\_x)

rf\_conf\_matrix = confusion\_matrix(test\_y, rf\_predict)

rf\_acc\_score = accuracy\_score(test\_y, rf\_predict)

print(rf\_conf\_matrix)

print(rf\_acc\_score)

[[45588 125]

[ 2084 43919]]

0.975914780409089

4. from sklearn.ensemble import AdaBoostClassifier

ad=AdaBoostClassifier(n\_estimators=7)

ad.fit(train\_x,train\_y)

ad\_pred=ad.predict(test\_x)

print(accuracy\_score(test\_y,ad\_pred))

print(confusion\_matrix(test\_y,ad\_pred))

print(classification\_report(test\_y,ad\_pred))

0.748037419861311

[[36405 9308]

[13801 32202]]

precision recall f1-score support

0 0.73 0.80 0.76 45713

1 0.78 0.70 0.74 46003

accuracy 0.75 91716

macro avg 0.75 0.75 0.75 91716

weighted avg 0.75 0.75 0.75 91716

5.knn

precision recall f1-score support

0 0.80 0.97 0.88 45713

1 0.96 0.76 0.85 46003

accuracy 0.87 91716

macro avg 0.88 0.87 0.86 91716

weighted avg 0.88 0.87 0.86 91716

6. GaussianNB

0.6519800252954774

precision recall f1-score support

0 0.81 0.40 0.53 45713

1 0.60 0.90 0.72 46003

accuracy 0.65 91716

macro avg 0.70 0.65 0.63 91716

weighted avg 0.70 0.65 0.63 91716

7.With hypertuning the parameters for gradient boosting classifier I observed these results,

Learning rate: 0.05

Accuracy score (training): 0.744

Accuracy score (validation): 0.742

Learning rate: 0.1

Accuracy score (training): 0.748

Accuracy score (validation): 0.746

Learning rate: 0.25

Accuracy score (training): 0.762

Accuracy score (validation): 0.760

Learning rate: 0.5

Accuracy score (training): 0.769

Accuracy score (validation): 0.768

Learning rate: 0.75

Accuracy score (training): 0.769

Accuracy score (validation): 0.768

Learning rate: 1

Accuracy score (training): 0.770

Accuracy score (validation): 0.770

Learning rate: 1.25

Accuracy score (training): 0.766

Accuracy score (validation): 0.763

Learning rate: 1.75

Accuracy score (training): 0.733

Accuracy score (validation): 0.733

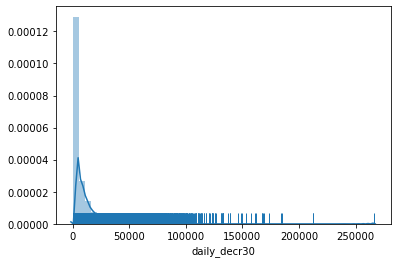
For hypertuning for logistic regression,I got the following result,

Best: 0.745132 using {'dual': False, 'max\_iter': 100}

Execution time: 48.619722843170166 ms

* Key Metrics for success in solving problem under consideration
* The best key metric was random forest classifier and decision tree classifier both giving a high accuracy as well as high roc auc score,coming to the approach ,upsampling was the best logical approach to balance the data and give meaningful results.
* Visualizations

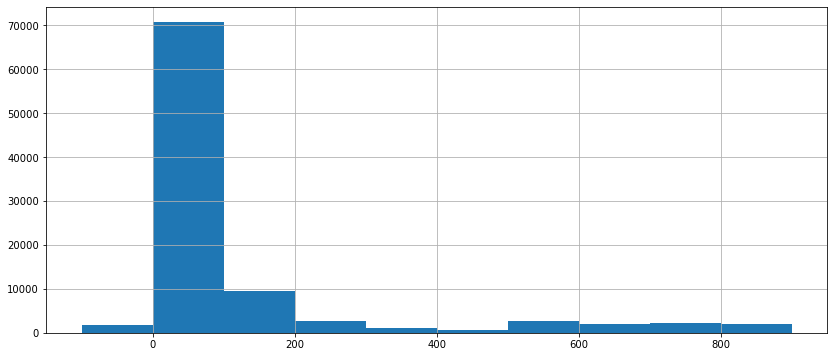
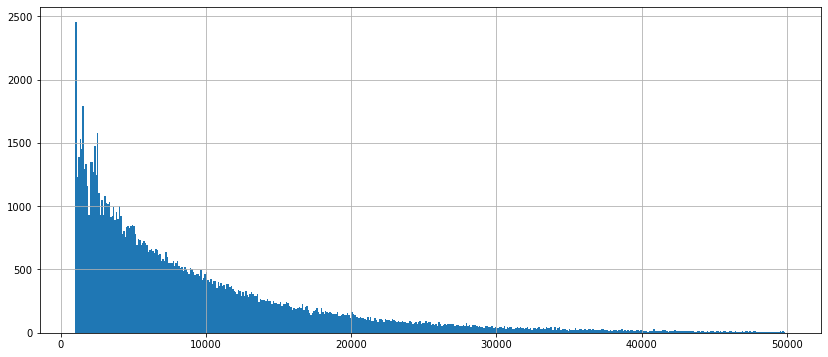
So this is the dist plot for my first column daily amount spent on main account over last 30 days



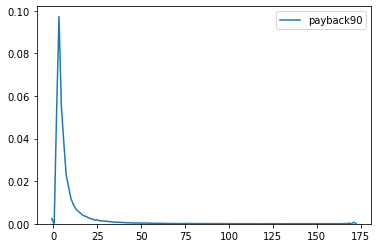
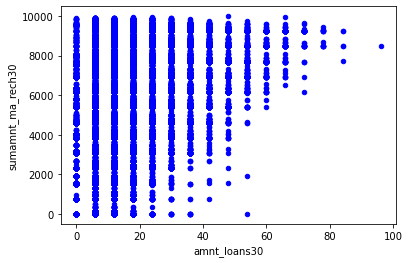
After further coding i can zoom in and find a more clear numbers as the range is too large,by setting the in values

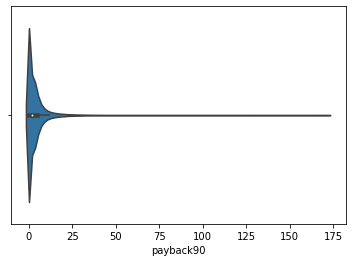
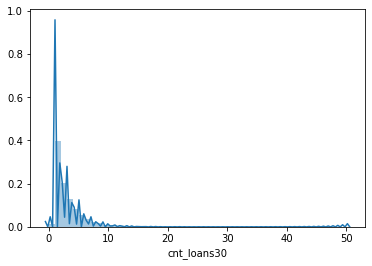
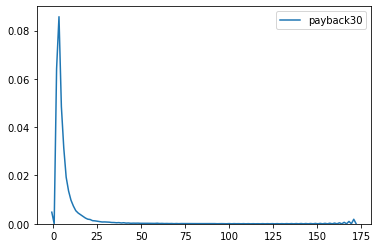
I can observe from -100 to 1000 values lie more till 200 and then for 1000 to 50,000 (max value lie till 10,000)

Similar is the case with daily\_decr90



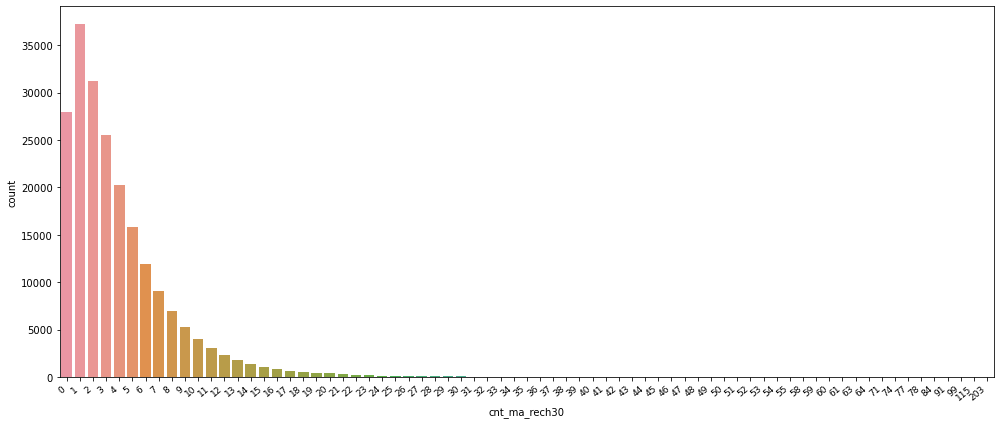
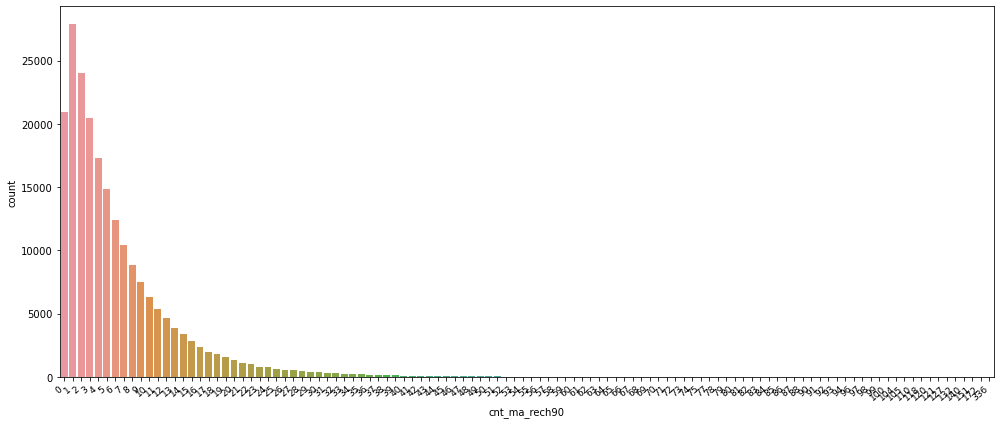
Lets observe the scatter plot between total amount of loans over 30 days and total amount of recharge over last 30 days we can see the relation values are not very scattered indicating more the recharge more the amount of loan

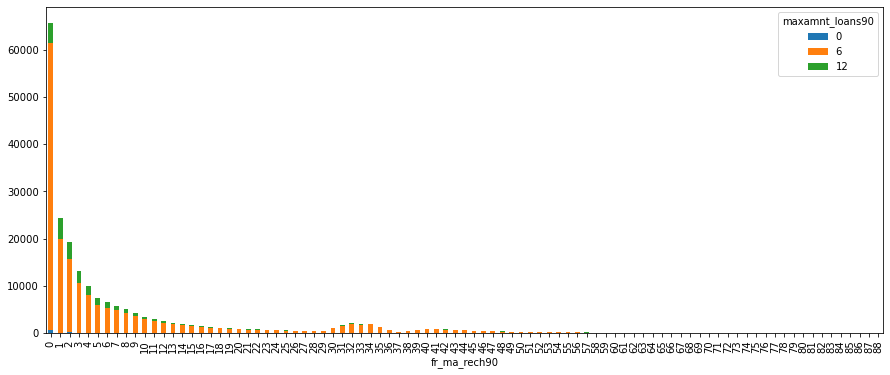
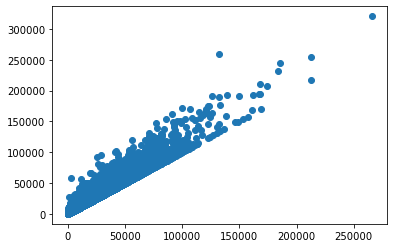


The payback 30 that is the avg days is more between 0 to 25

For payback 90 too the days are more from 0 to 25,no of recharge done over last 30 and 90 days in main account,for 90 its more between (0 to 38)

For 30(0 to 23) after that the no of times recharge done gets decreasing





This is for understanding the frequency of recharge done over 90 days with relation to max amount of loans(0,6,12).so for 12 no of loans this gets on decreasing with the increase in frequency,at o recharge we see 0 loans taken afterwards the average loan taken is 6.

* Interpretation of the Results

So from the above ,

The recharge frequency and loans for 30 days is similar to the pattern of 90 days in most of the cases

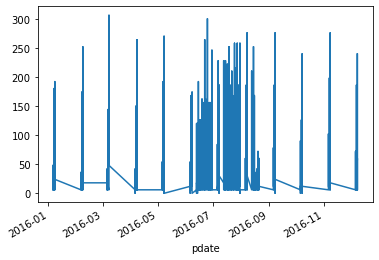
There cannot be seen much direct relation to label class,but there are many relationships between other variable the most important being the usage that is the recharge and no of loans taken

The data is imbalanced so downsampling and upsampling methods have been used

The data has basically two important things recharge and laons for 30 and 90 days ,other columns are somewhat a extension to these like median,frequency and count revolvling around these two

Apart from this there is payback column and the average days lue between 0 to 25

Also concerning the pdate

’

Pdate mostly lies from 7 th month to 9th month of 2016.

**CONCLUSION**

* Key Findings and Conclusions of the Study

By looking at the auc roc score for the upsampled model ,the roc auc score came out to be at 74 percent whereas before upsampling the auc roc score was 54. Whereas earlier accuracy was of 80 percent and after upsampling it was around 75.So precision recall f1 score,and auc roc score was the decision criteria for my model’s performance .Among all the metrics used decision tree and random forest worked really well.

* Learning Outcomes of the Study in respect of Data Science

So from this project I learned how apart from a traditional loan problem,there are credit realted problems revolving around other sectors too like in this case it was telecom industry,this model classified the loans given to the users were paid back within 5 days or not,so this was a great move and with data science the telecom company can study the behaviour of users ,how much loan they can take and thereby identifying there potential users also providing these types of credit to all incomes groups especially poor.

The major limitation faced was imbalance dataset which was solved by resampling and generating upsamples,for future work and rnd purposes i would like to explore more on this and the perfect ratio for two classes (because in this scenario i generated equal samples for both the classes with upsampling) for a classification problem ,so if ratio varies than how the results will differ will be another task.