

MICRO CREDIT DEFAULTER LOAN PREDICTION

Submitted by:

Sudhakar sasum

**ACKNOWLEDGMENT**

Sources are datatrained lms machinelearningmastery.com,

scikitlearn.org,

github.com

**INTRODUCTION**

* Business Problem Framing

**Problem Statement:**

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.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

We are working with one such client that is in Telecom Industry. 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 understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

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).

The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

* Conceptual Background of the Domain Problem

The concepts we have to use in this project is diffferent classification techniques such as random forest classifier ,adaboost classifier etc and for data imbalance we use over sampling techniques

* Review of Literature

Our main motive is 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. 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 i.e. defaulter.

First we load the data and description of data consists of standard deviation which is high shows outliers in the data is high, so we should not delete the outliers which would cause data loss,and we should check for null values which are not present in this data,now we split the data for training and testing,and we should do PC analysis like standard scaler for standardisation of data. Here the data is imbalanced so check the data by exploratory data analysis now we

Should use different imbalancing techniques such as oversampling,smote analysis etc with different classifiers like randomforest classifier and adaboost classifier and check the accuracy score,roc curve and predict the data by testing.conclude the best data that gives best results

* Motivation for the Problem Undertaken

The main objective of this project is to , whether the customer will be paying back the loaned amount within 5 days of insurance of loan.this imbalanced data gives experience to deal this imbalanced techniques for classification.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Main theme of this project to build classification model with data imbalance techniques.work flow of this project

1)gathering data csv file

2)data preprocessing:checking for messy data,checking outliers,

3)data visulaisation:checking data imbalance by using different visualisation techniques.

4)correlation:checking correlation by heatmap

5))spilitting data: peforming train test split

6)standardisation:performing standard scaler

7)model fitting:performing model fitting by classification techniques such as random forest classifier ,adaboost classifiers with data imbalance techniques such as oversampling,smote

8)testing:predicting the output of test data

9)comparing accuracy score,roc scores of different classifications

10)conclude the best results with the above results

* Data Sources and their formats

Machine Learning is so fundamental that it could be used to redo almost everything in industry. Understanding the four main data types is the the key you can use to open the door to Machine Learning.

When you work with different data types, you need to gain an understanding of them from a machine learning perspective, especially if you’re working on visualizations and data storytelling. Alina Zhang, our data superhero, shares insights into a machine learning perspective on different data types. Learn from the best to make better decisions about data conversions and encoding.

Alina explains that machine learning generally categorizes data into one of four main data types. These are:

1. **Numerical data –** This can be discrete or continuous data, but it always uses exact numbers that are not ordered in time. It’s also called quantitative data.
2. **Categorical data –** This is data that expresses characteristics, so it is also called the “class label” in a super classification context. Although categorical data can be represented using numbers, the numbers don’t have a mathematical meaning.
3. **Time series data –** This data consists of numbers that were collected across a period of time.
4. **Text data –** This is essentially words, which you might want to turn into numbers as soon as possible.

Data we used in this project is text data .csv

* Data Preprocessing Done

Check for any outliers in the data ,and then check null values, but this project consists of more outliers and zero null values

* Data Inputs- Logic- Output Relationships

Ou main motive is 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. 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 i.e. defaulter

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

Here, you can describe any presumptions taken by you.

* Hardware and Software Requirements and Tools Used

Hardware:any normal configuration

Software:jupyter notebook,in windows ,mac or linux

Tools:pandas library ,matplotlib,seaborn,classification techniques such as randomforest classifier adaboost classifier.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Primary approch is to load data with all libraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

%matplotlib inline

import seaborn as sn

import warnings

warnings.filterwarnings('ignore')

Reading data files

import pandas as pd

df=pd.read\_csv('Data file.csv')

from IPython.display import display

pd.options.display.max\_columns = None

display(df)

* Testing of Identified Approaches (Algorithms)

#### Independent and Dependent Features

X=df.drop("label",axis=1)

y=df.label

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

X\_train,X\_test,y\_train,y\_test=train\_test\_split(X,y,train\_size=0.75)

from sklearn.preprocessing import StandardScaler

import numpy as np

scaler = StandardScaler()

X = scaler.fit\_transform(X)

from sklearn.linear\_model import LogisticRegression

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

from sklearn.model\_selection import KFold

import numpy as np

from sklearn.model\_selection import GridSearchCV

clf=GridSearchCV(log\_class,grid,cv=cv,n\_jobs=-1,scoring='f1\_macro')

RANDOMFORESTCLASSIFIER

from sklearn.ensemble import RandomForestClassifier

classifier=RandomForestClassifier(criterion='gini', max\_depth=10, min\_samples\_split=5, min\_samples\_leaf=1)

classifier.fit(X\_train,y\_train).fit(X\_train,y\_train)

RANDOMFORESTCLASSIFIER WITH UNDER SAMPLING

from collections import Counter

Counter(y\_train)

from collections import Counter

from imblearn.under\_sampling import NearMiss

ns=NearMiss(version=1, n\_neighbors=3)

X\_train\_ns,y\_train\_ns=ns.fit\_sample(X\_train,y\_train)

print("The number of classes before fit {}".format(Counter(y\_train)))

print("The number of classes after fit {}".format(Counter(y\_train\_ns)))

ADABOOST CLASSIFIER

from sklearn.ensemble import AdaBoostClassifier

# Create adaboost classifer object

abc = AdaBoostClassifier(n\_estimators=50,

learning\_rate=1)

# Train Adaboost Classifer

model = abc.fit(X\_train, y\_train)

#Predict the response for test dataset

y\_pred = model.predict(X\_test)

RANDOMFOREST WITH SMOTE OVERSAMPLING

print("Before OverSampling, counts of label '1': {}".format(sum(y\_train == 1)))

print("Before OverSampling, counts of label '0': {} \n".format(sum(y\_train == 0)))

# import SMOTE module from imblearn library

# pip install imblearn (if you don't have imblearn in your system)

from imblearn.over\_sampling import SMOTE

sm = SMOTE(random\_state = 2)

X\_train\_res, y\_train\_res = sm.fit\_sample(X\_train, y\_train.ravel())

* Run and Evaluate selected models

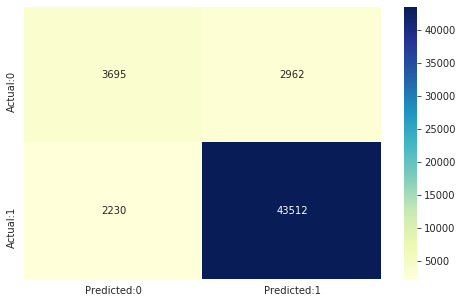
# confusion Matrix

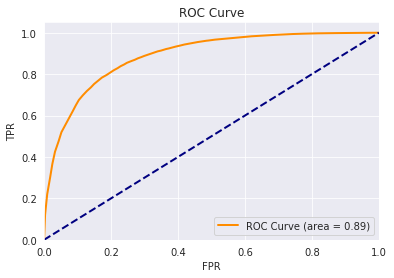
cm=confusion\_matrix(y\_test,y\_pred)

conf\_matrix=pd.DataFrame(data=cm,columns=['Predicted:0','Predicted:1'],index=['Actual:0','Actual:1'])

plt.figure(figsize = (8,5))

sns.heatmap(conf\_matrix, annot=True,fmt='d',cmap="YlGnBu");

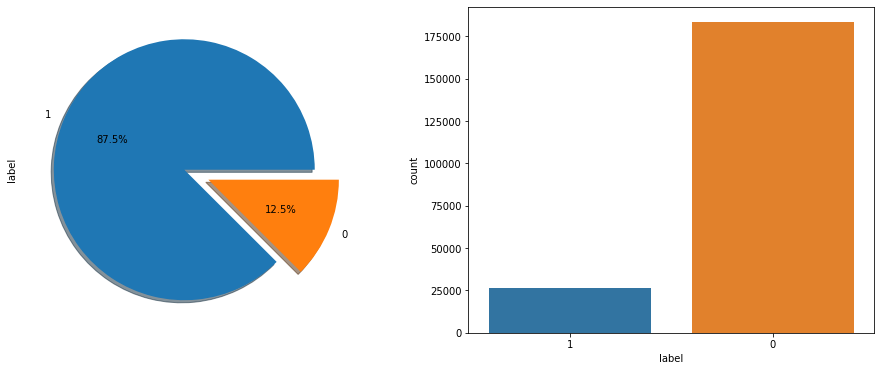


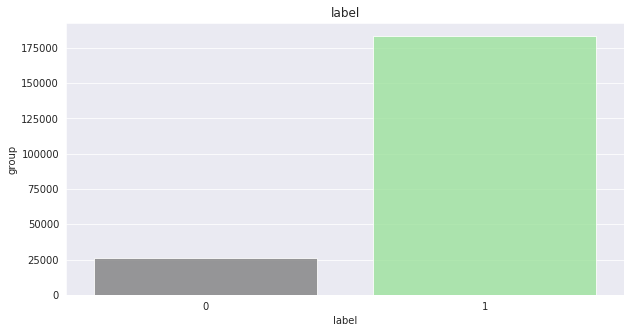


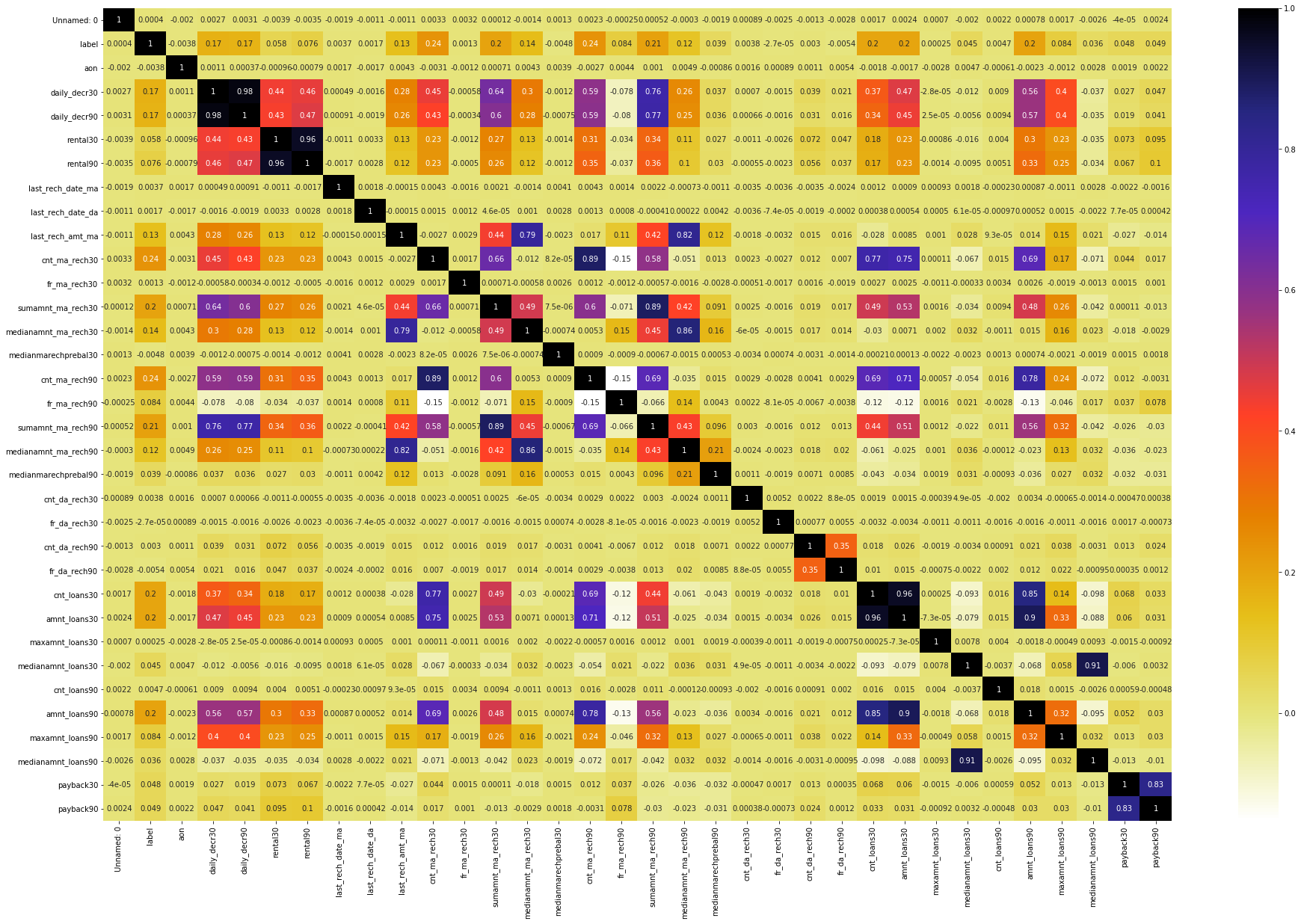
* Key Metrics for success in solving problem under consideration

Key metrics is to find best classification by accuracy

* Visualizations







If different platforms were used, mention that as well.

* Interpretation of the Results

RandomforestClassifier gives best results with smote oversampling

**CONCLUSION**

* Key Findings and Conclusions of the Study

1)data more outliers

2)data is imbalanceed

3)randomforestclassifier performs well

* Learning Outcomes of the Study in respect of Data Science

By checking the heat map we can clearly find the highly correlated features and by fitting different classification techniques we conclude randomforest with smote oversampling works well

* Limitations of this work and Scope for Future Work

Since my hardware limitations i can perform with only limit number of techniques. You can use further data imblanced techniques such as

### SMOTETomek

### SMOTETomek is somewhere upsampling and downsampling. SMOTETomek is a hybrid method which is a mixture of the above two methods, it uses an under-sampling method (Tomek) with an oversampling method (SMOTE). This is present within imblearn.combine module

### SMOTEEENN

SMOTEENN another library present within imblearn.combine module. This performs similar to SMOTETomek, there is some difference in results between the two methods.

These are best methods to overcome class imbalance but we needs higher GPUs to perform since normal hardware consume more time .