

Micro-Credit Defaulter Model

Submitted by:

HIMANI UNIYAL

**ACKNOWLEDGMENT**

I am really grateful for this project opportunity and would sincerely thank to “Flip Robo” for trusting me with this project.

Besides this, I would like to express my gratitude towards Srishti Mann Ma’am (SME) to provide us quick revert on the queries raised and continuous support throughout the project.

**INTRODUCTION**

* Business Problem

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.

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. 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.

* Conceptual Background of the Domain Problem

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. 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.

Telecom Industry have launched various products and have developed its business and organization based on the budget operator model. 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.

* Review of Literature

The data description and problem statement are given with the problem statement have been analyzed. The research has been done by Google.

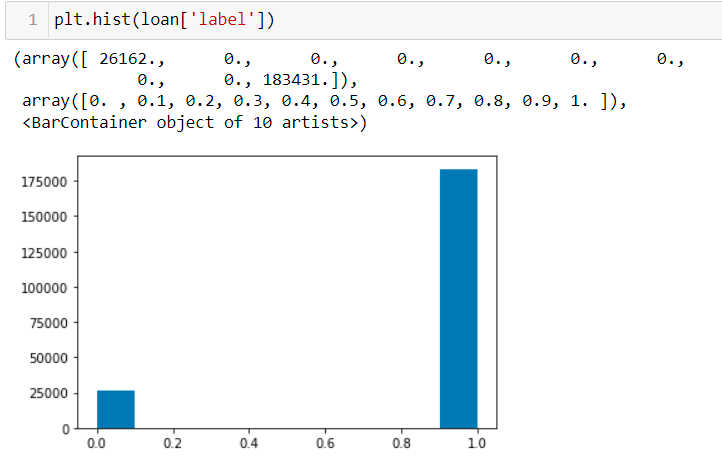
* Motivation for the Problem Undertaken

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. 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.

**Analytical Problem Framing**

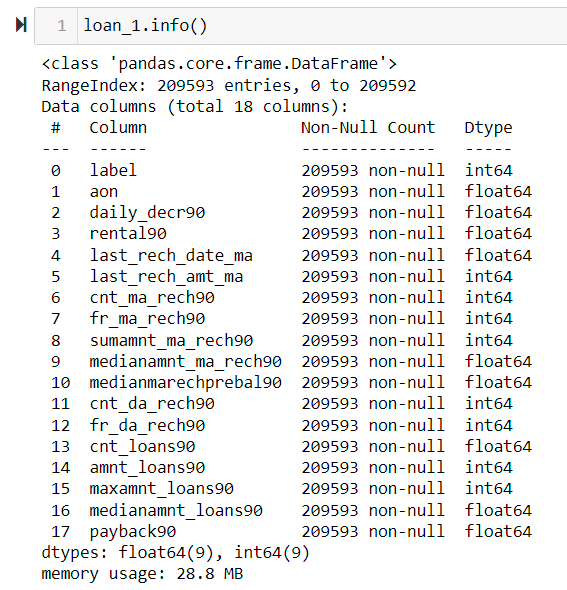
* Mathematical/ Analytical Modeling of the Problem

After analysing the data set, the following observation has been made:

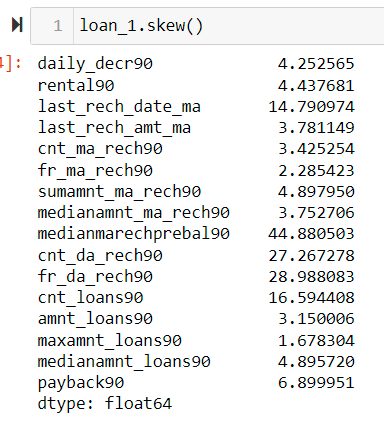
* The shape of data set 209593 rows × 37 columns, which shows that the number of attributes is 37 in number.
* There are no null values in the data set.
* The data set is imbalanced, which was resolved by SMOTE technique.
* 
* After analysing the data, it has been seen that the data set is distributed for 30 days and 90 days and representing majorly same relationship.
* pcircle has same value so it can be dropped. And date parameter as no direct relationship with label so it can be dropped
* Data Sources and their formats

It has been seen that the data set is distributed for 30 days and 90 days The columns are 'Unnamed: 0', 'label', 'msisdn', 'aon', 'daily\_decr30', 'daily\_decr90', 'rental30', 'rental90', 'last\_rech\_date\_ma', 'last\_rech\_date\_da', 'last\_rech\_amt\_ma', 'cnt\_ma\_rech30', 'fr\_ma\_rech30', 'sumamnt\_ma\_rech30', 'medianamnt\_ma\_rech30', 'medianmarechprebal30', 'cnt\_ma\_rech90', 'fr\_ma\_rech90', 'sumamnt\_ma\_rech90', 'medianamnt\_ma\_rech90', 'medianmarechprebal90', 'cnt\_da\_rech30', 'fr\_da\_rech30', 'cnt\_da\_rech90', 'fr\_da\_rech90', 'cnt\_loans30', 'amnt\_loans30', 'maxamnt\_loans30', 'medianamnt\_loans30', 'cnt\_loans90', 'amnt\_loans90', 'maxamnt\_loans90', 'medianamnt\_loans90', 'payback30', 'payback90', 'pcircle', 'pdate'] dtype='object').

After observing the data I have concluded that, 90 days data can be considered for further evaluation.

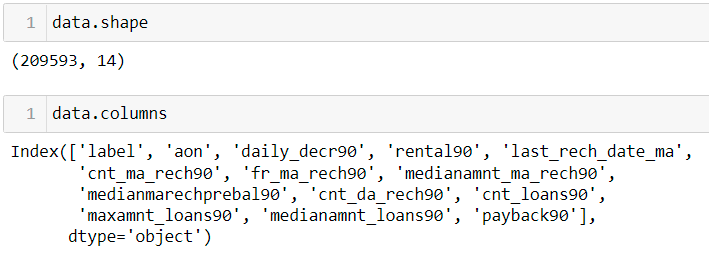


There are lot of outliers in the dataset.

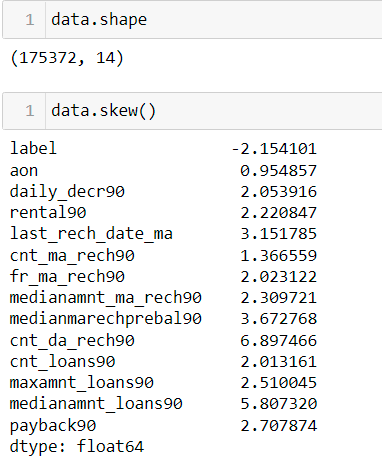


* Data Preprocessing Done

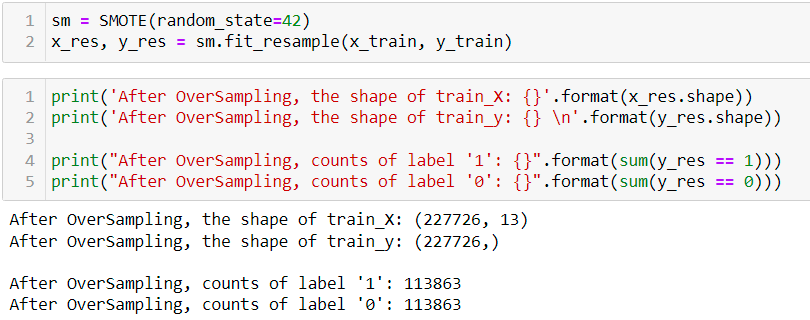
Firstly, I have analysed, each attribute and seen that the 90 days data can be considered for futher analysis and thus reduced the dimension of data set. Also delete the data set with no relation with the label prediction. Also the attributes having high correlation shown in the heat map.



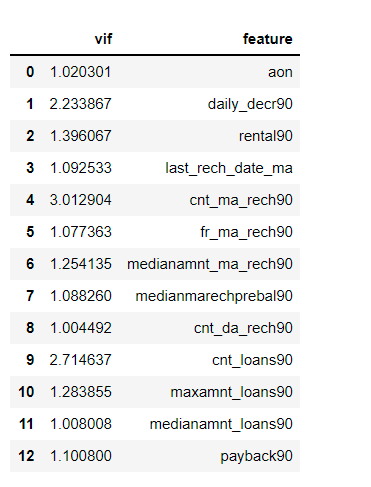
There were lot of outliers so, z score has been used and fixed the threshold at 3, and reduced the skewness.



Now for balancing the dataset, standard scaler is used to scale the data. SMOTE technique is used to balance it.



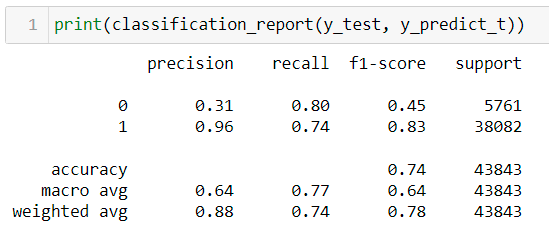
Also the varience has been checked:



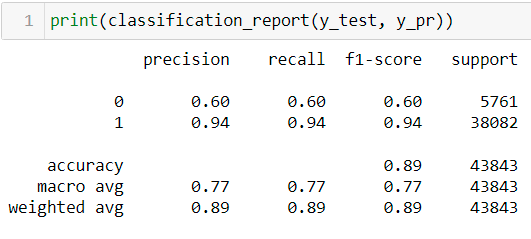
And found perfect to go ahead.

* Data Inputs- Logic- Output Relationships

The problem in the statement is a classification problem. So logistics regression has been applied and 74% accuracy was evaluated.



Than Random Forest ensemble technique has been used and it gives the accuracy of 89%.



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

The date attribute has no contribution in the paying the loan.

* Hardware and Software Requirements and Tools Used

Laptop, router, Python, Machine learning algorithms. Following libraries:

import seaborn as sns

from matplotlib import pyplot as plt

import pandas as pd

from imblearn.over\_sampling import SMOTE

from sklearn.preprocessing import StandardScaler

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

from sklearn.linear\_model import LogisticRegression

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

from sklearn.linear\_model import LogisticRegression

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

from sklearn.model\_selection import cross\_val\_score

from sklearn.model\_selection import GridSearchCV

from sklearn.ensemble import RandomForestClassifier

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Firstly the attributes has been reduced, after data set was balanced, reduced the skewness of data set.

The following algorithms used:

Linear Regression

Random Forest

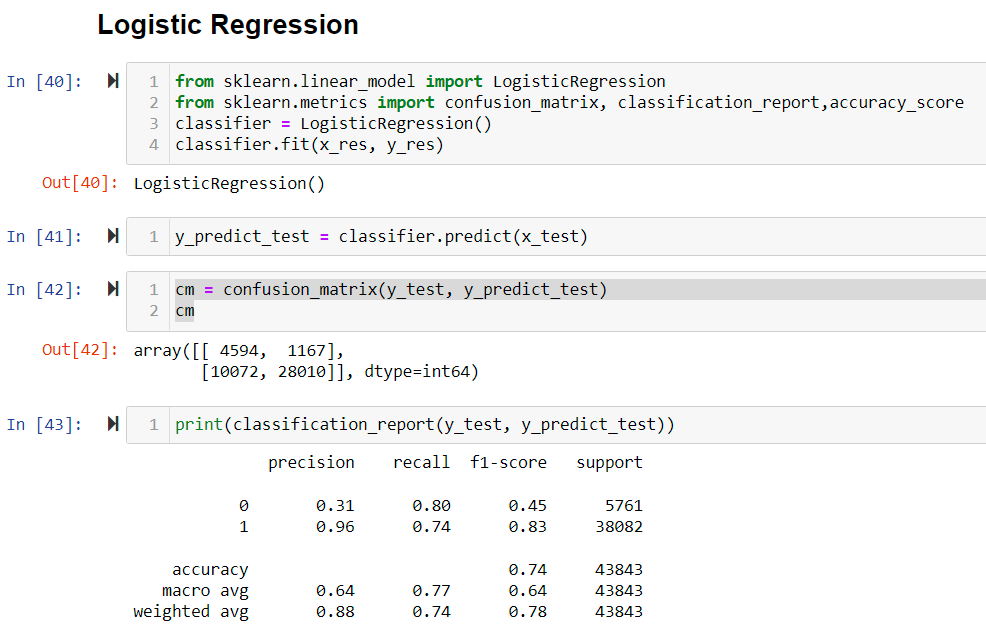
* Testing of Identified Approaches (Algorithms)

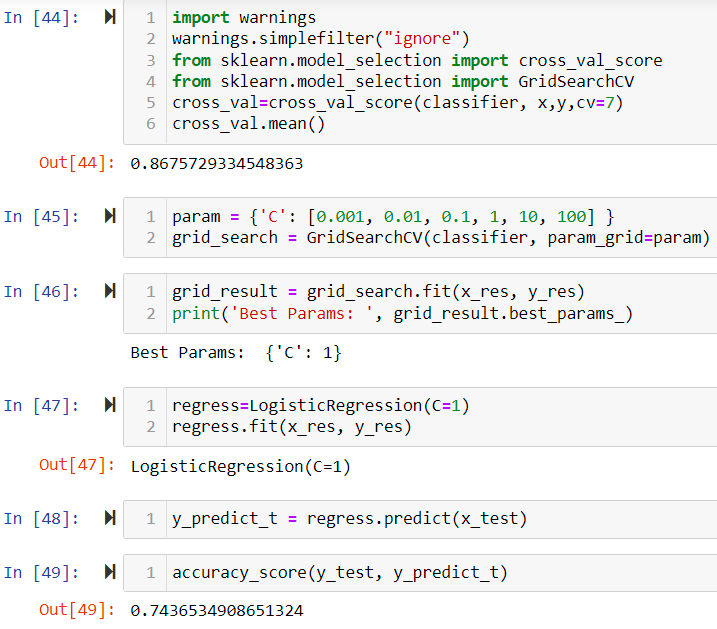
Linear Regression

Random Forest

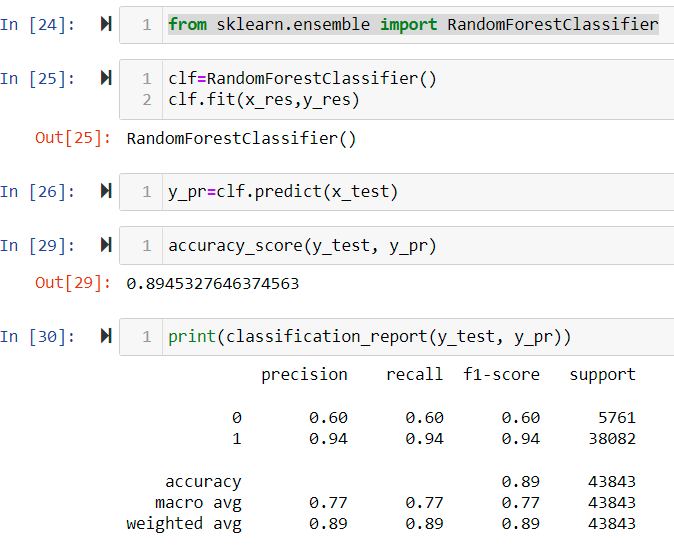
* Run and Evaluate selected models

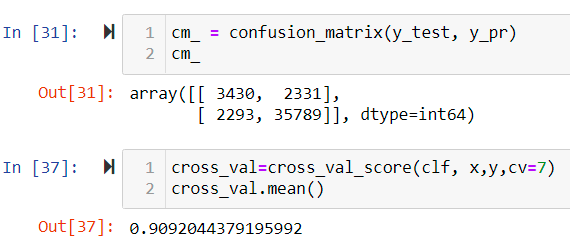
Logistic Regression:

.



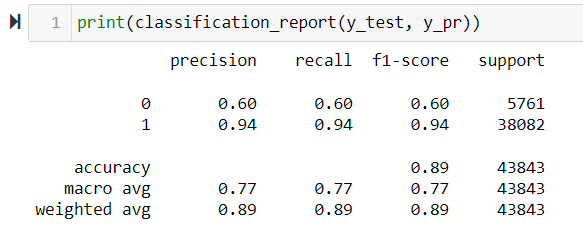
Random Forest:



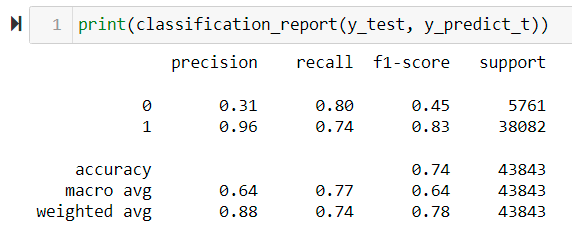


* Key Metrics for success in solving problem under consideration

Random Forest



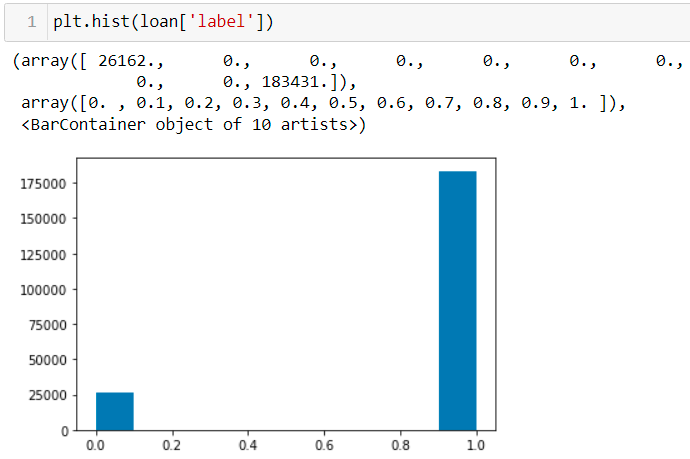
Logistic Regression:



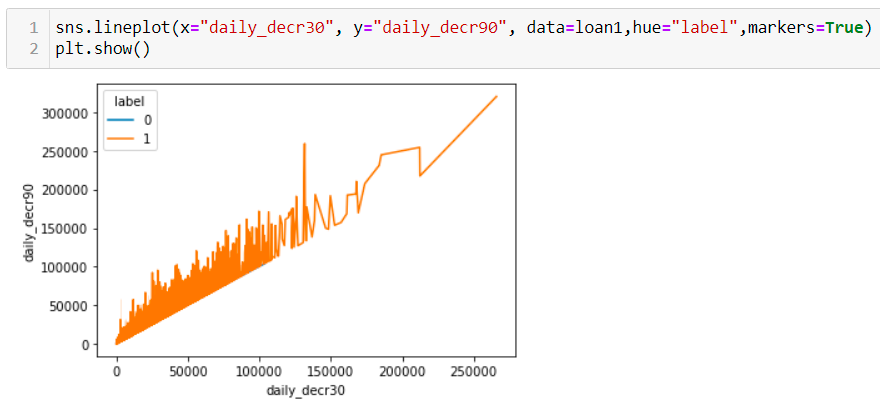
Also confusion matrics has been obtained for both.

* Visualizations

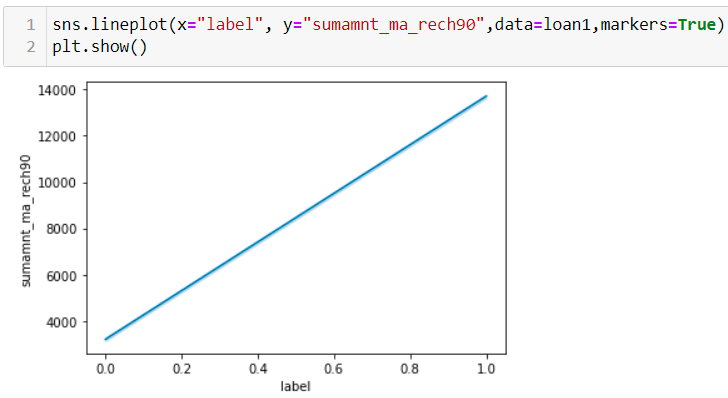
To check te distribution of label histogram was plotted:



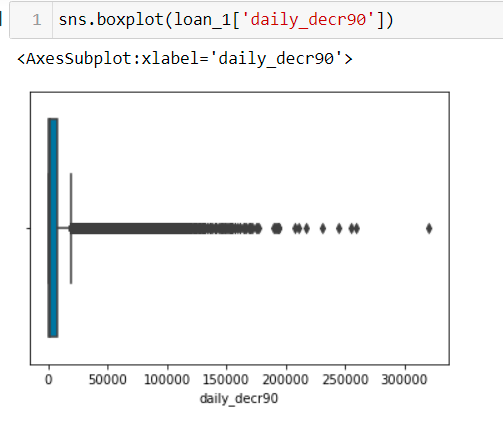
Line plot has been plotted to check the relation between the variables.



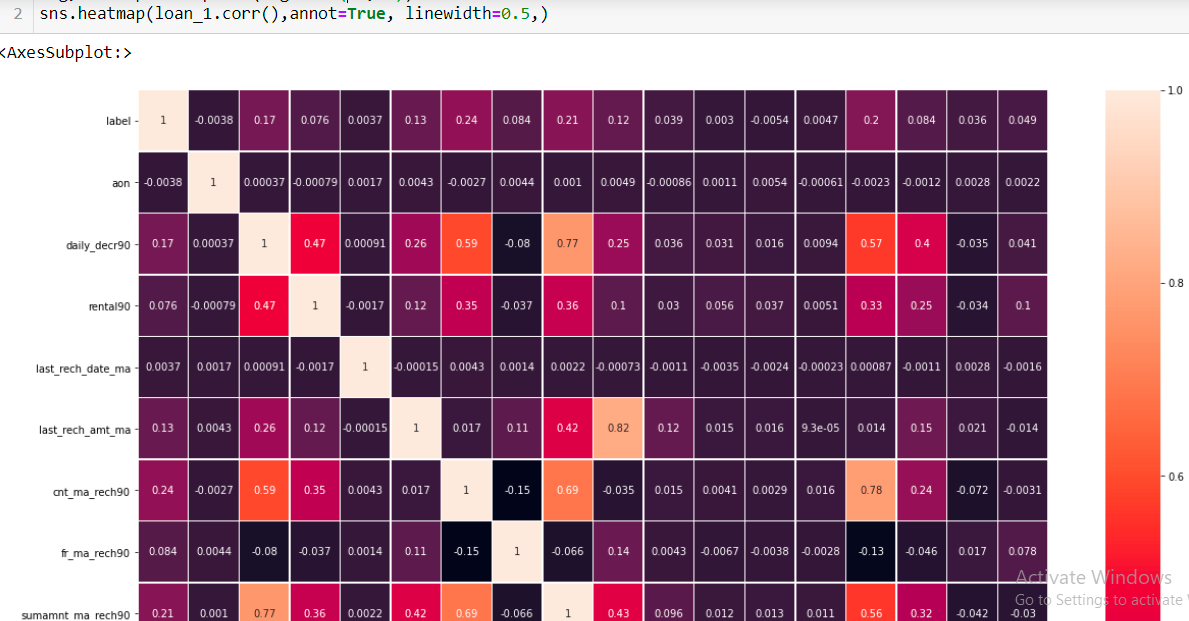
Line plot has been plotted to check relation with the label.



The box plot has been plotted to check the outliers.

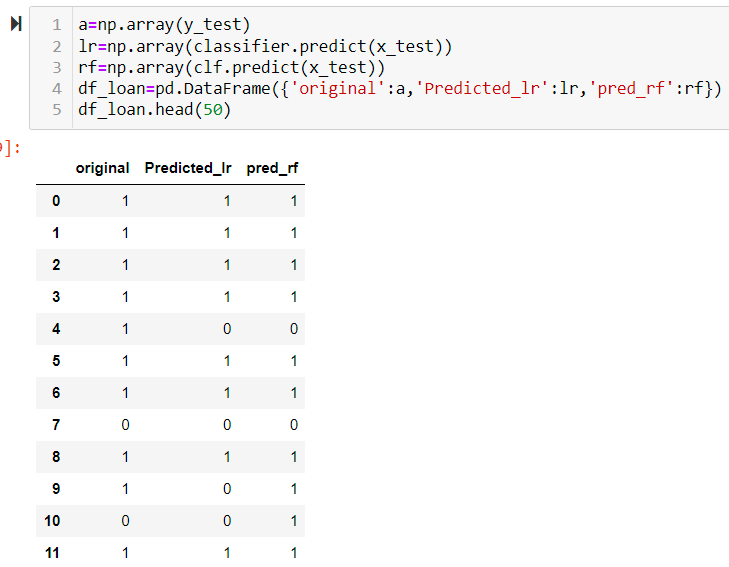


The heat map is plotted to check the correlation between all the attributes.



.

* Interpretation of the Results

.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The Random Forest gives the better prediction of the problem.

* Learning Outcomes of the Study in respect of Data Science

The size of data set is too big and took long time in execution of the code. Also because of size of data set, individual analysis needs to be done.

Dimension reduction is the key problem in the data set. Although some direct techniques can be used but after visualization and studying the variables some similarity has been found. And decision has been made.

Also the unbalanced data is big problem, so result may vary from ground reality.

Outliers in data set has been a problem, although we have implemented the z score, with threshold 3.

* Limitations of this work and Scope for Future Work

The more algorithm can be used to improve the accuracy. Other techniques can be used for outliers and dimension reduction.

Also hyper tuning can be applied in Random forest, but it is taking too long so did not apply to it.