

Micro Credit Loan

Submitted by:

C. Anilkumar Reddy

**INTRODUCTION**

* Business Problem Micro Finance Loan

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 or not . In real life we can observe Airtel 10 rs recharge loan.

* Conceptual Background of the Domain Problem

We can clearly see from the output column that it is a classification problem since the values are either 0 or 1. Also given that the data is imbalanced.

* Review of Literature

The main goal of this project is does the loaned amount is returned back in 5 days or not for micro finance loan .The data has huge number of outliers and also given we cant loose much of the data and also i took help from geek for geeks and other youtube channel

Named kirh naik .Since it is and unbalanced data we need to take care of it.

* Motivation for the Problem Undertaken

The main objective of this project is to predict whether the loaned amount is returned back within 5 days.

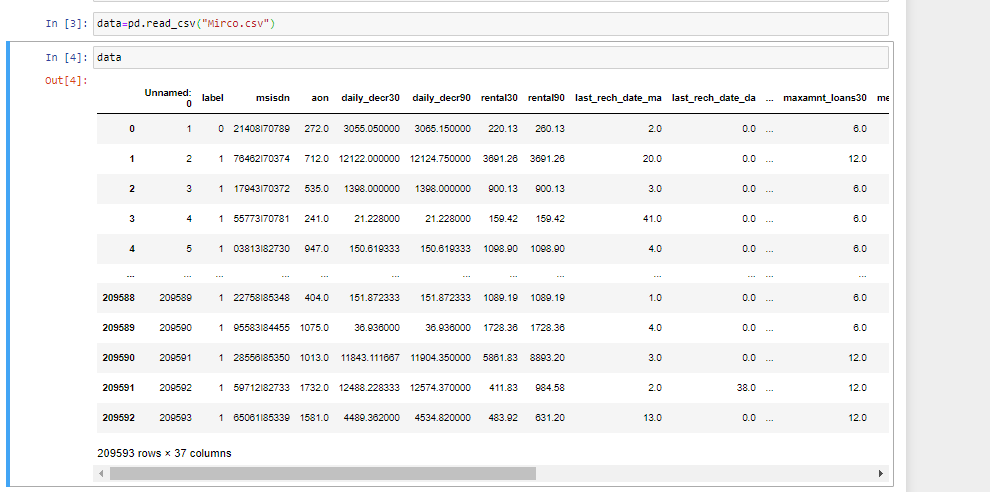
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

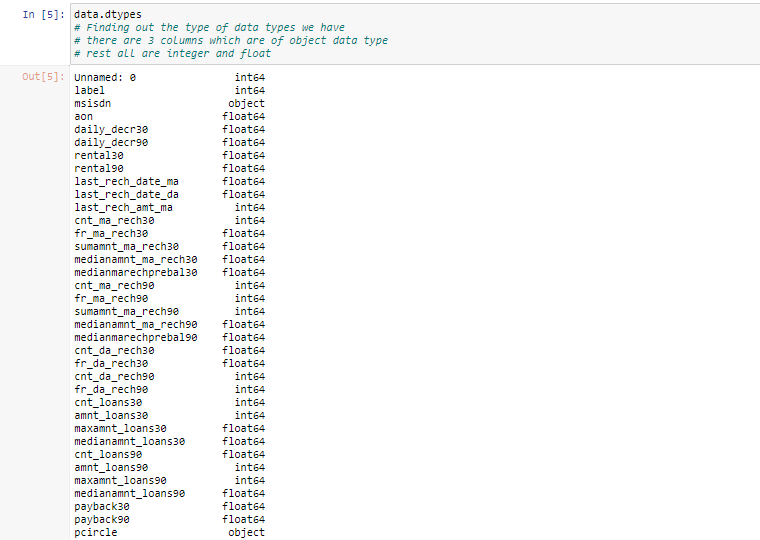
From the dataset it is clear that the output column label is classified as 0 and 1 so it is a classification problem. .Label 1 is 87% and 0 is 13% so our data is imbalanced .data stats gives us an idea that there are huge outliers are present cause max values is more than 75% also from the distribution plots the data is slightly right skewed. Co-relation map give insight the columns that are co-related and also by plotting the loan taken amount rs 5 about 80%

Of the label data .The frequency of paying back rs 5 loan is more compared to other loan .Using a logistic regression isn’t a good idea cause our data is imbalanced so we need to try different ML-algorithm and check for recall, precision, roc accuracy score .

* Data Sources and their formats

With the help of pandas we are reading the micro finance csv file which is

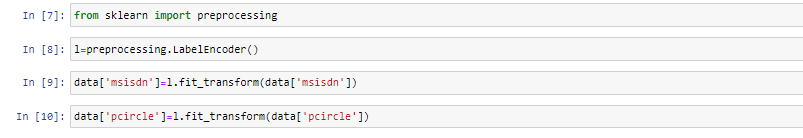
Checking the data types there are 3 object dtypes and rest all are int,float so we need to change those data types by using pre-processing



There is pdate which is in date format and also 2 columns which are object .

* Data Preprocessing Done

We have objects in our data so we are converting with sklearn-preprocess lib.

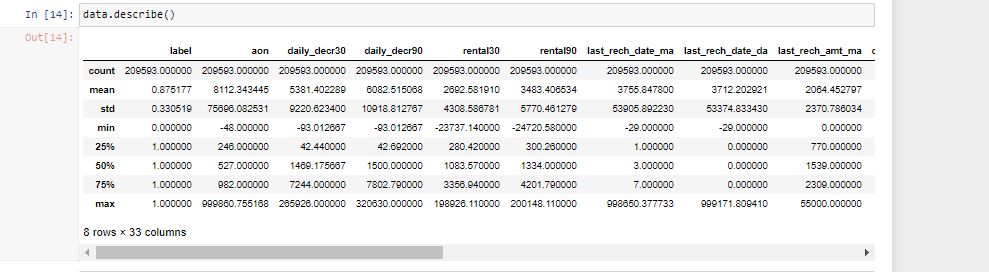


Then I am dropping unnecessary columns

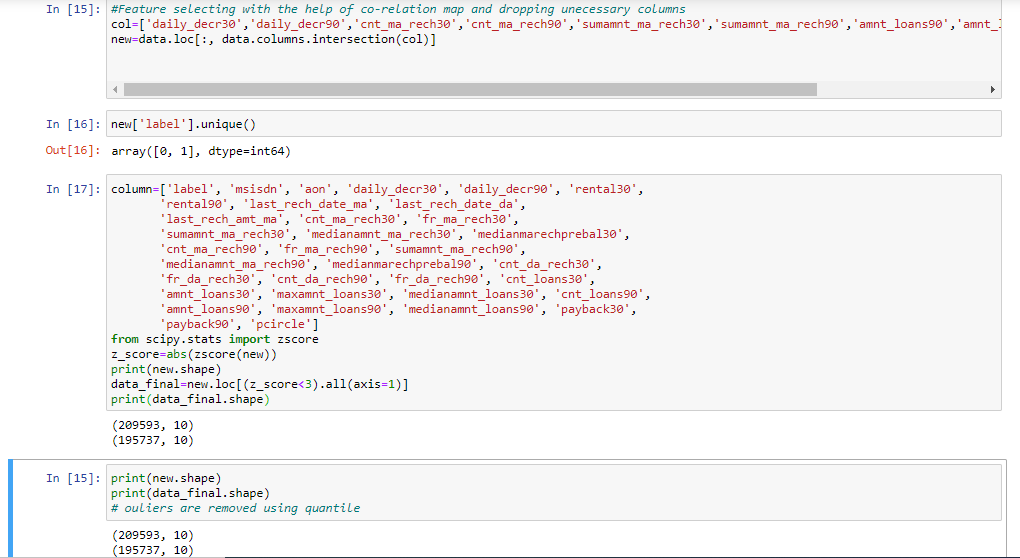
4.PNG

5.PNG

From the describe stats it is clear than we have huge outliers in our data cause max values in daily desr 30 have more than 75% which we need to remove them.



We don’t want to loose huge amount of data so we using zscore technique to remove outliers. The new data shape tells us that we didn’t loose much data. We are using a new data frame were only the columns which are highly co-related are used with the help of co-relation heatmap.



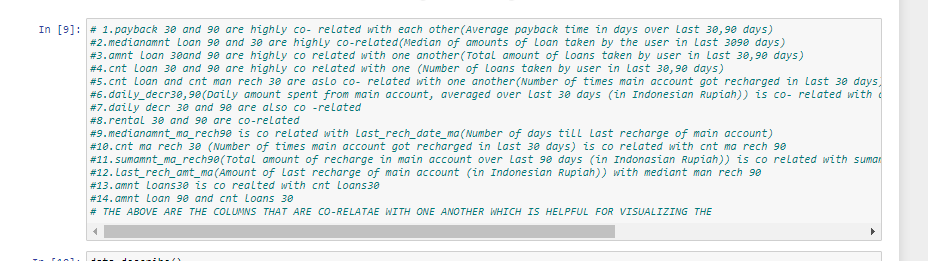
* Data Inputs- Logic- Output Relationships

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

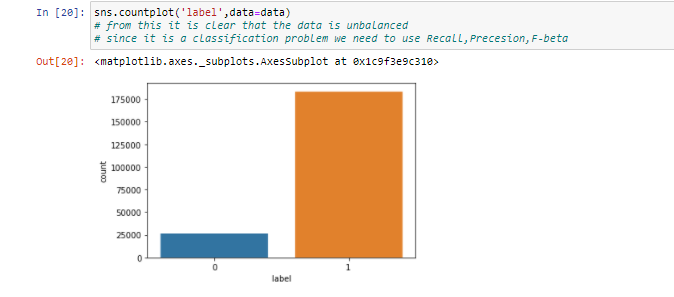
To understand the relationship between the input and o/p we can observe from co-relation map.



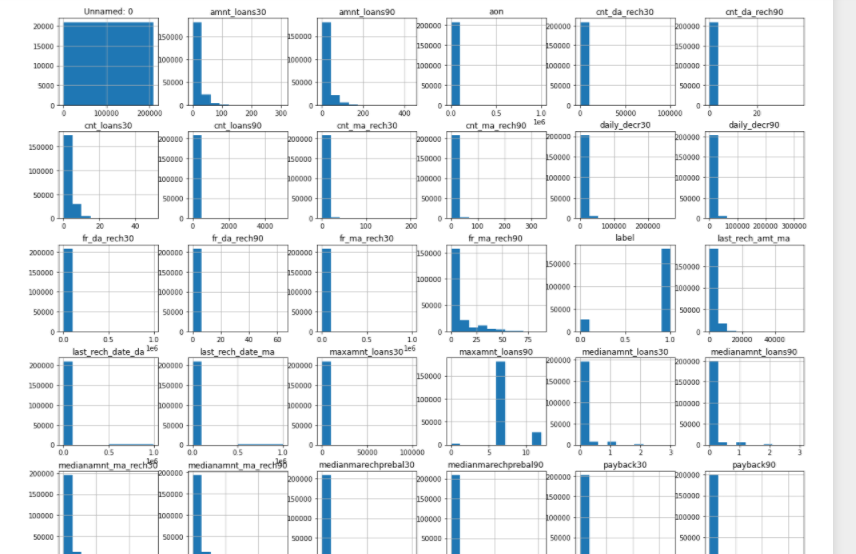
The below picture tells the columns that are co-related with each other so we can plot with the output column

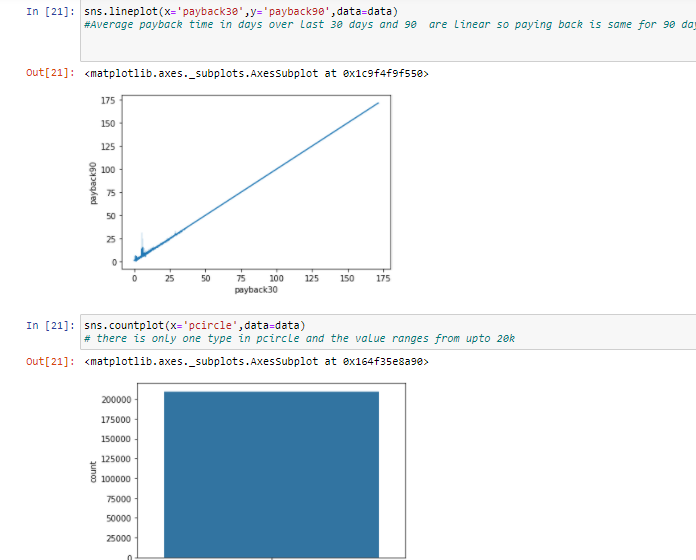
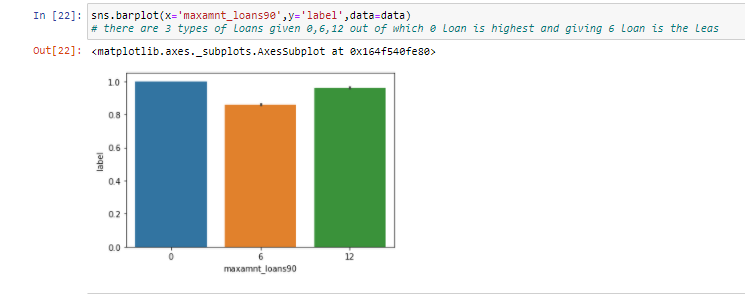


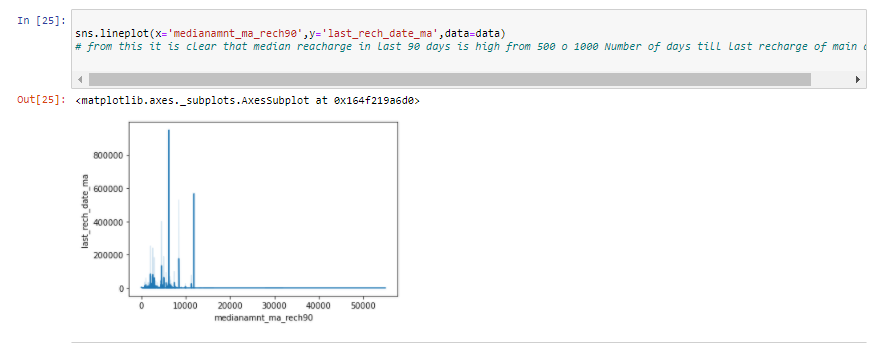
Plotting the o/p column with countplot and the data is distributed uneven .1 has the around 87% and 0 has 13%



We can check the histogram plot to understand the data



Here are the following plots. 



The above plots are giving info that 5rs loan is returned back most than 10 rs and pay back time for 30 and 90 days are linearly proportional to one another

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

I am assuming that using Logistic regression isn’t effective cause the

Model suggests the o/p to be 1 cause most of the data is 1 so using

Other models would we great and also need to check the recall,

Precession, auc roc score and also for low false negative rate.

* Hardware and Software Requirements and Tools Used

We are using python language .Jupyter Notebook is used cause it is

Easy visualization tool. Pandas, Numpy, sklearn ,Pre-processing, metrics, model selection,ml-lib(from sklearn.neighbors import KNeighborsClassifier,from sklearn.svm import SVC,from sklearn.tree import DecisionTreeClassifier,from sklearn.naive\_bayes import GaussianNB,from sklearn.linear\_model import LogisticRegression,

from sklearn.ensemble import RandomForestClassifier,from sklearn.ensemble import AdaBoostClassifier,from xgboost import XGBClassifier)

Matplot and seaborn are used for visualization the data. The above following are the library used in our project.

The basic hardware requirements are 4gb ram,i3 processor and 500gb space.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I am assuming that using Logistic regression isn’t effective cause the

Model suggests the o/p to be 1 cause most of the data is 1 so using

Other models would we great and also need to check the recall,

Precession, auc roc score and also for low false negative rate.

There are also outliers present in our data so we are cleaning them with iqr .

* Testing of Identified Approaches (Algorithms)

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

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC

from sklearn.tree import DecisionTreeClassifier

from sklearn.naive\_bayes import GaussianNB

from sklearn.linear\_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import AdaBoostClassifier

from xgboost import XGBClassifier

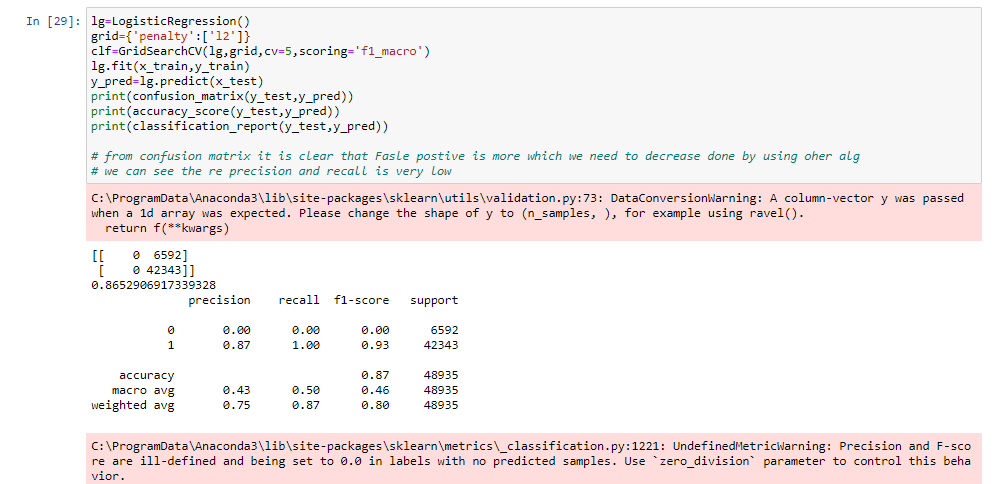
The following are the machine learning algorithms used in our

model

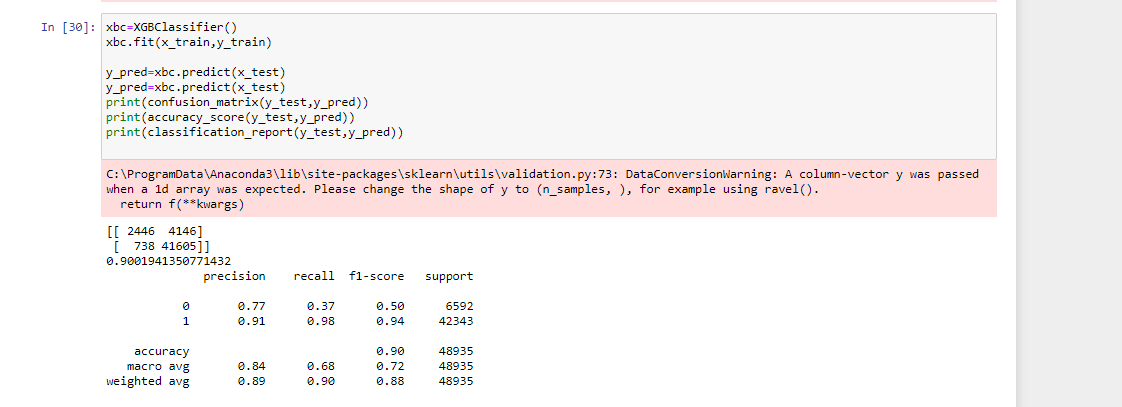
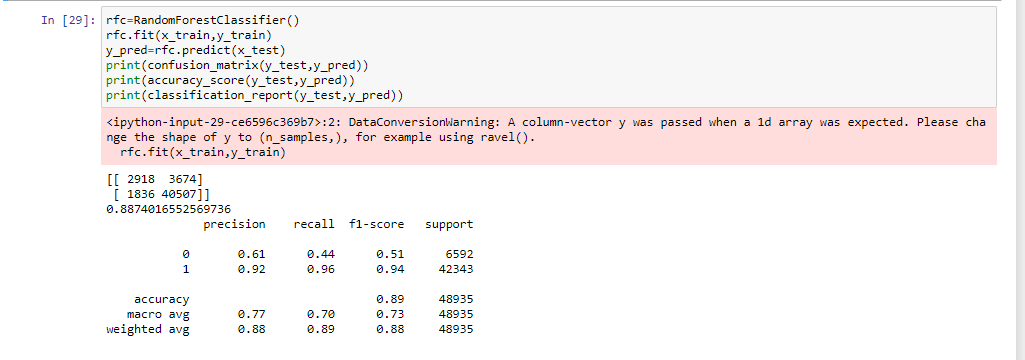
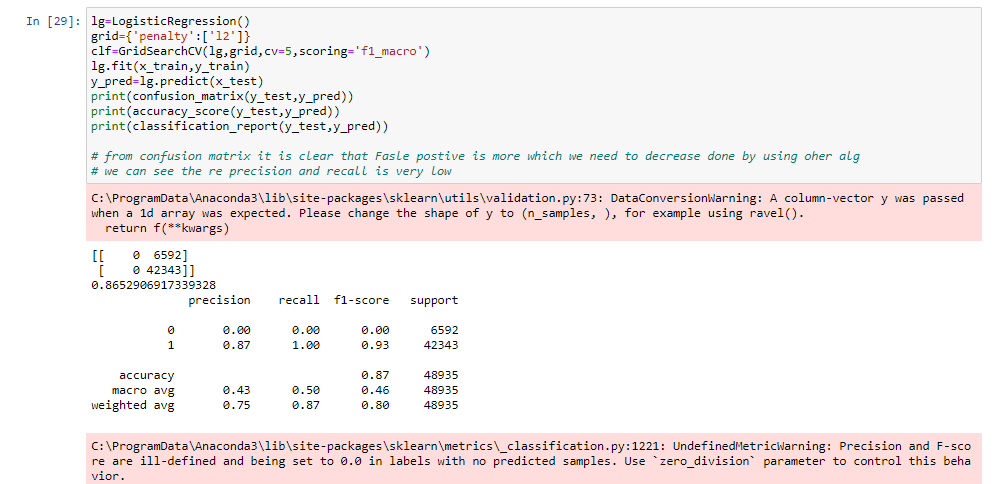
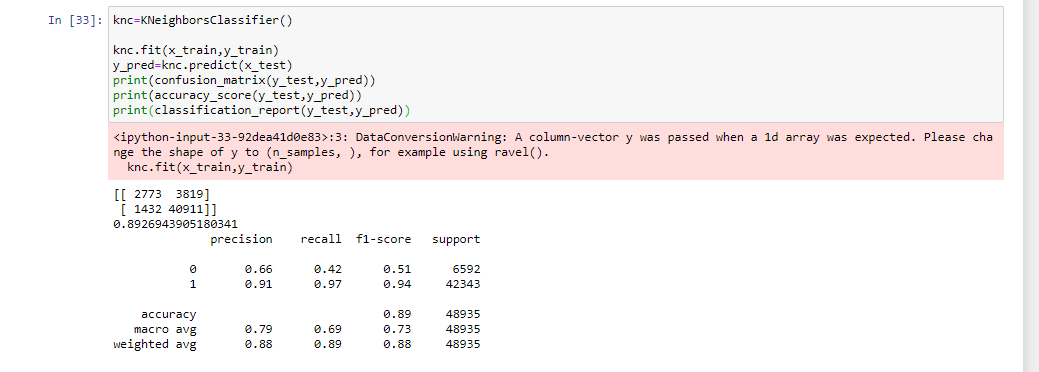
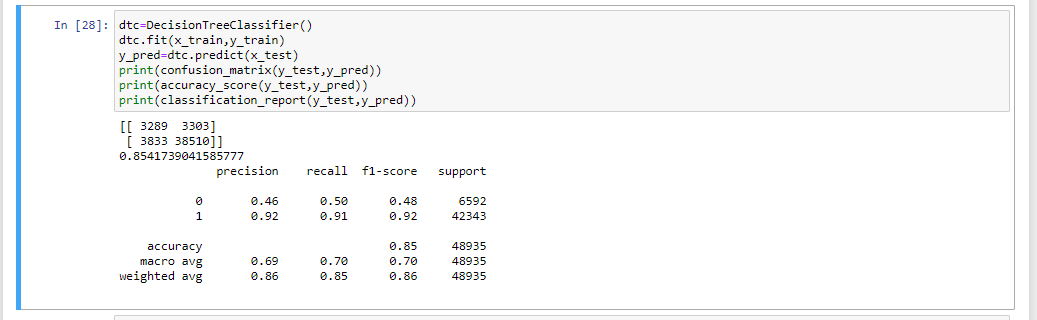
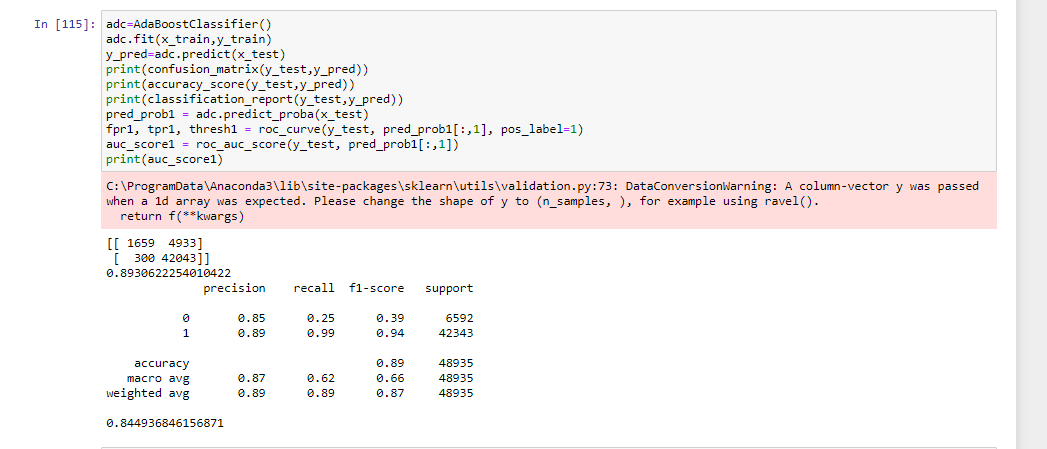
* Run and Evaluate selected models

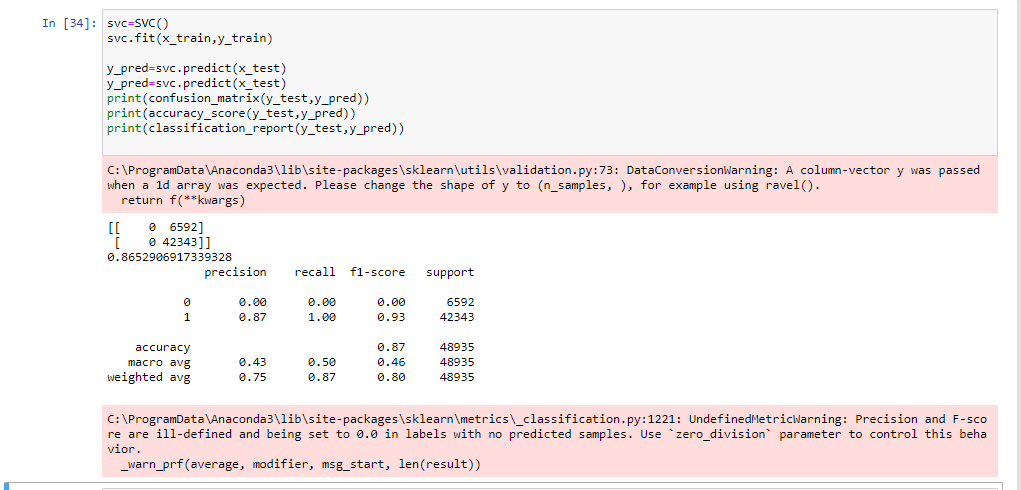
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

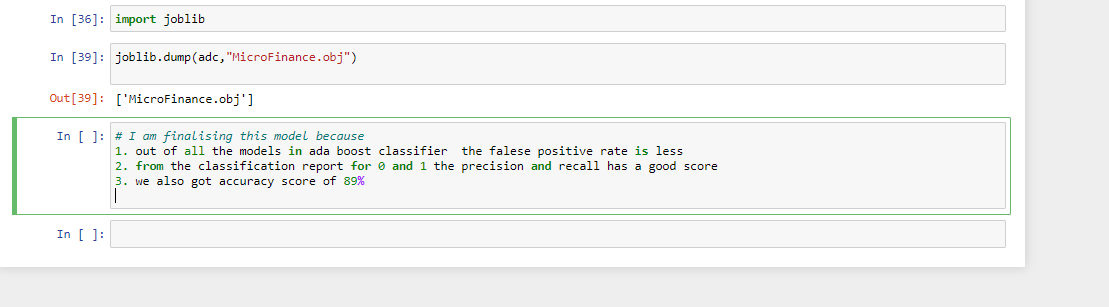
The 1st alg used is logistic regression we can observe that we have 0 fp and fn rate total which is imbalanced and recall is also low



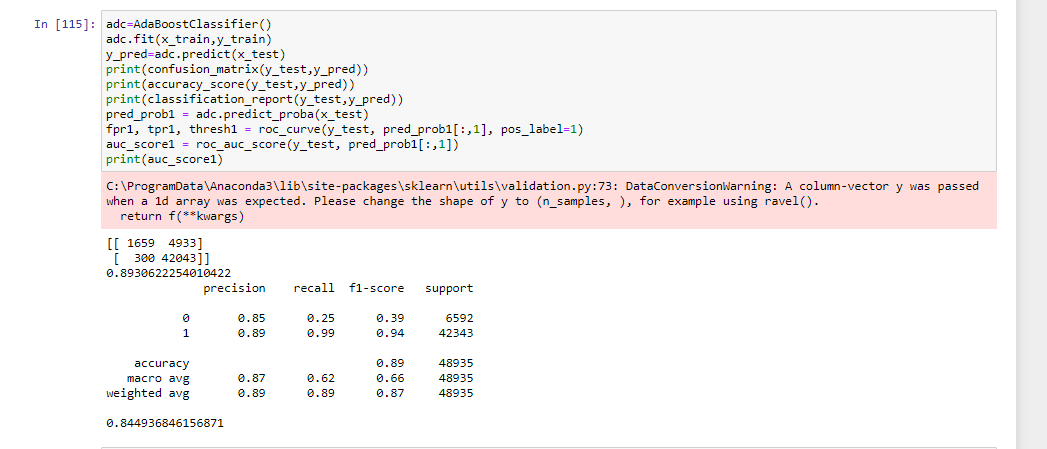
The below are the list of alg used in sequence out of which svc is working best and also have a low false negative rate which is required and we have accuracy of 86%







* Key Metrics for success in solving problem under consideration



The above is the best metrics cause we have less false positive rate and also accuracy score is 89% we can check the precession and recall to be 0.85 and 0.25 which is good compared all other algorithm

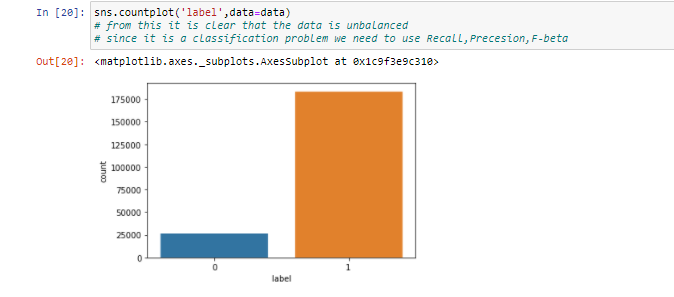
* Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

Firstly we are checking the o/p column y which says that data is

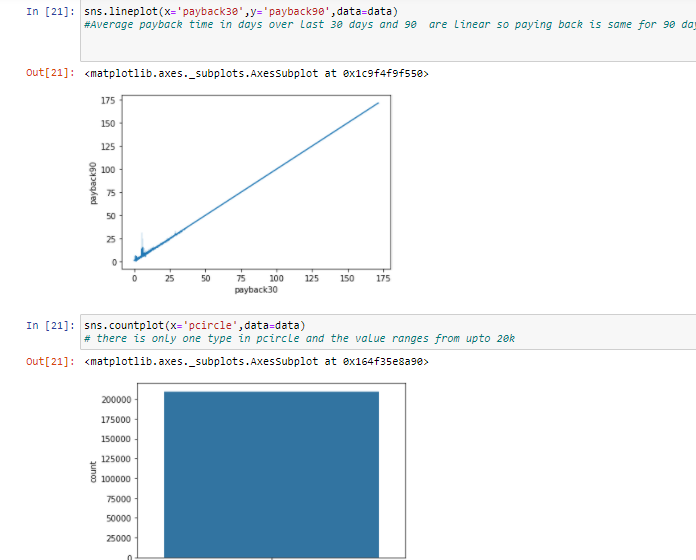
Imbalanced 1 is more and 0 is less using countplot from sns lib.



The below 2 graphs show are linear relationship between the

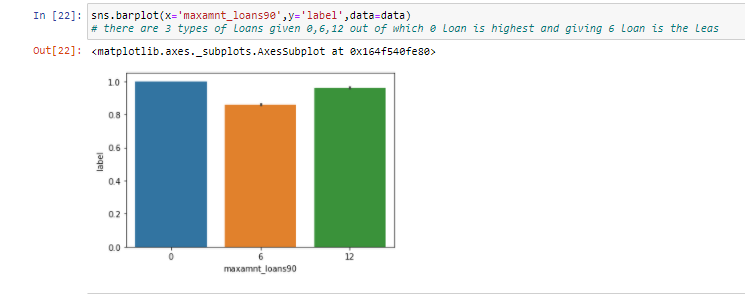
Amount payback in 30 and 90 and also the total amount of

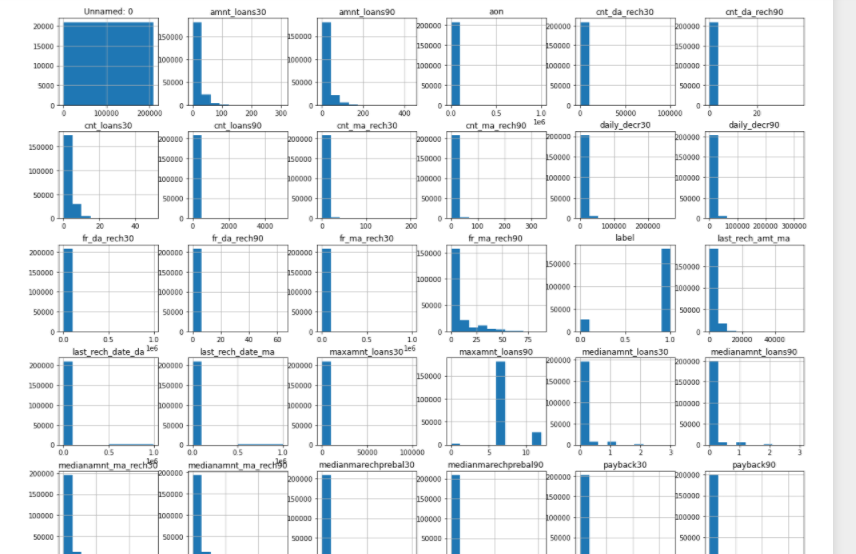
Recharge done .



We can also check from this plot that the highest taking loan is 12

And 2nd is 0 and last comes the rs6 .



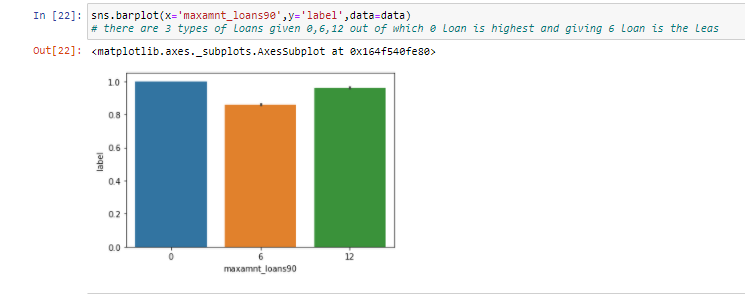


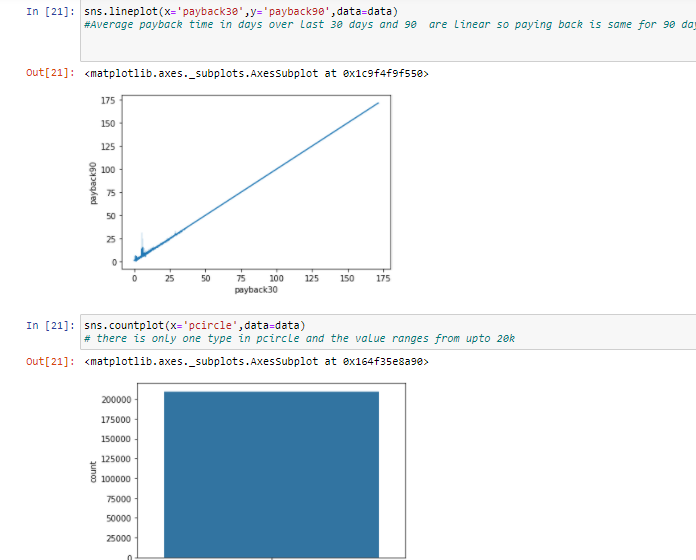
From the above histogram it is clear that our data is rightly skewed

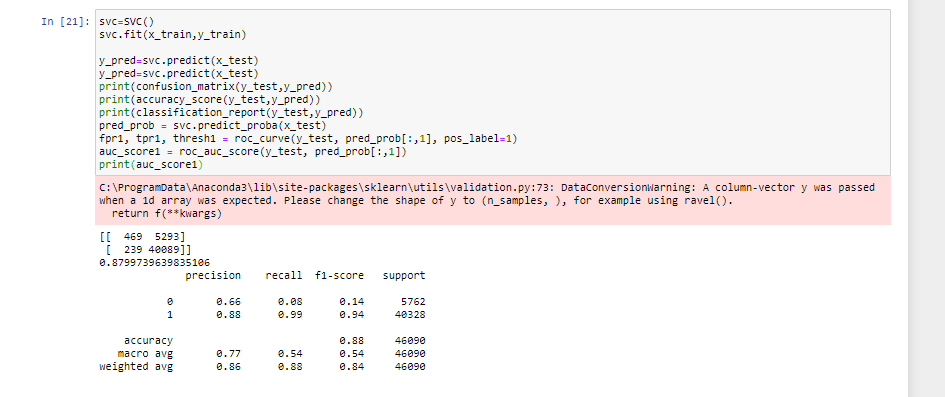
And we need remove it with log.

* Interpretation of the Results

From the visualizations, preprocessing and modelling we can observe that there are 3 types of loans and paying back are directly proportional and out of all svc is the best model cause we have less false negative rate.







**CONCLUSION**

* Key Findings and Conclusions of the Study

Since the data is imbalanced we need to choose the correct algorithm .we cant use logistic cause it is sensitive and we will

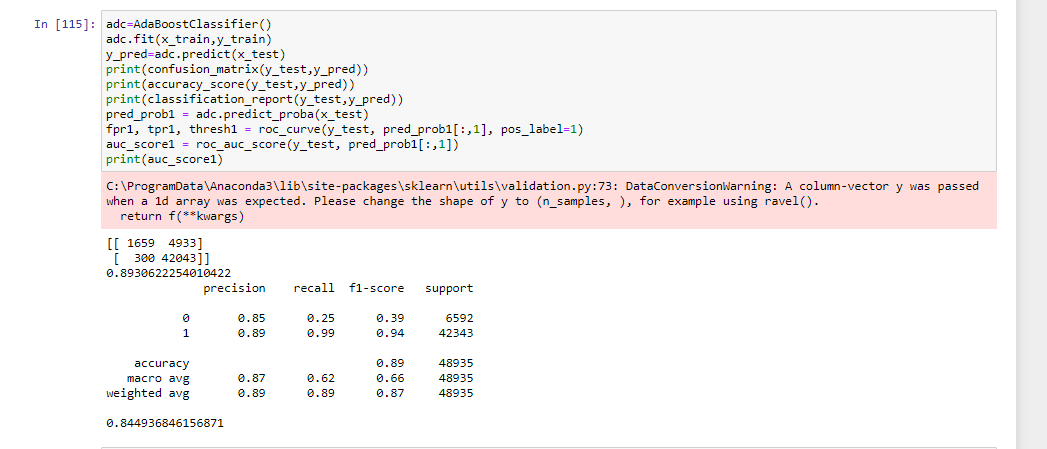
Get 1 o/p most of time .We need to remove outliers and also

Scale the data before training and testing.

* Learning Outcomes of the Study in respect of Data Science

I main problem I faced with the dataset is removing outliers i tried with zscore we can also use it.

The best algorithm used is svc cause it has less false negative rate and also it has as good precession, recall compared to all alg.



* Limitations of this work and Scope for Future Work

To improve further results we need to decrease the false negative rate and also improve the roc auc score .