NAME OF THE PROJECT

Micro credit defaulter project

Submitted by:

Munish kumar

ACKNOWLEDGMENT

I took the help of google and website of miscrofinance institution (MFI) and problem statement also helped me to understand the problem and basics of data.

Data description helped me to find the meaning of each and every column it was very helpful to me to understand all the columns of data and what is the meaning of each coumn.

INTRODUCTION

- Business Problem Framing

in the data set our target is to find or predict a person that he or she is a defaulter or not

- Review of Literature

  In this project, the main problem is this whether a person is a defaulter or not we have to predict that the person who took the loan returned the loan or not if he returned the loan mean he or she is not defaulter if he did not return the loan means that person is defaulter.

Motivation for the Problem Undertaken

My objective behind making this project is that we avoid giving the loan to those people who may be defaulters or who will not return the loan because if we will give the loan continuously to those people who will not return money than our business is suffering from lose.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

for better understanding of data I plot a lot of graphs like bar graph, count plot and scatter plot for outliers detection I plot the box plot and for distribution or skewness I plot the distplot

- Data Preprocessing Done

There is not a lot of requirement to clean this data because it is also cleaned I del only those columns which are not relevant or do not have a good correlation with the target column or shows multicollinearity

- Hardware and Software Requirements and Tools Used

I complete this project with the help of pyhton library ,matplotlib ,seaborn library these library are inside the pyhton, machine learning algorthim like support vactor classifier , decision tree classifier , random forest classifier

Model/s Development and Evaluation

When I solved this project my approach is like this first I try to understand the meaning of data and later I understand the meaning of column when I understand the meaning of all the column than I check which column is usefull to me and is there any null value is present in that column or not and what is the datatype of that column and I delete those column which is not usefull for data like there is a unnamed:0 column in the data which is not usefull and do not give any information so I delete that column and I also delete those column which shows multicollinearity

- Testing of Identified Approaches (Algorithms)

I use support vactor classifier, decision tree classifier , random forest classifier, k neighbours classifier but all the algorthims is not giving me the output my notebook is in hanging position so I make a comment of the code.

Run and Evaluate selected models

```
In [228]: dtc= DecisionTreeClassifier()
          dtc.fit(x_train,y_train)
          dtc.score(x_train,y_train)
          preddtc = dtc.predict(x_test)
          print(accuracy_score(y_test,preddtc))
          print(confusion_matrix(y_test,preddtc))
          print(classification_report(y_test,preddtc))
```

```
0.9099875820471882
[[25924  2349]
 [ 2725 25372]]
              precision    recall  f1-score   support

           0       0.90      0.92      0.91     28273
           1       0.92      0.90      0.91     28097

    accuracy                           0.91     56370
   macro avg       0.91      0.91      0.91     56370
weighted avg       0.91      0.91      0.91     56370
```

Decision tree classifier gives 90 per accuracy

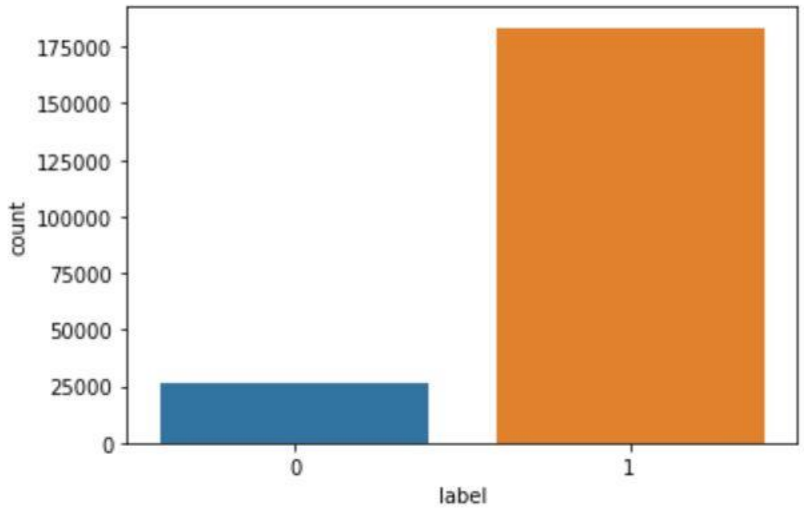
- Key Metrics for success in solving problem under consideration

I use confusion matrix and classification report

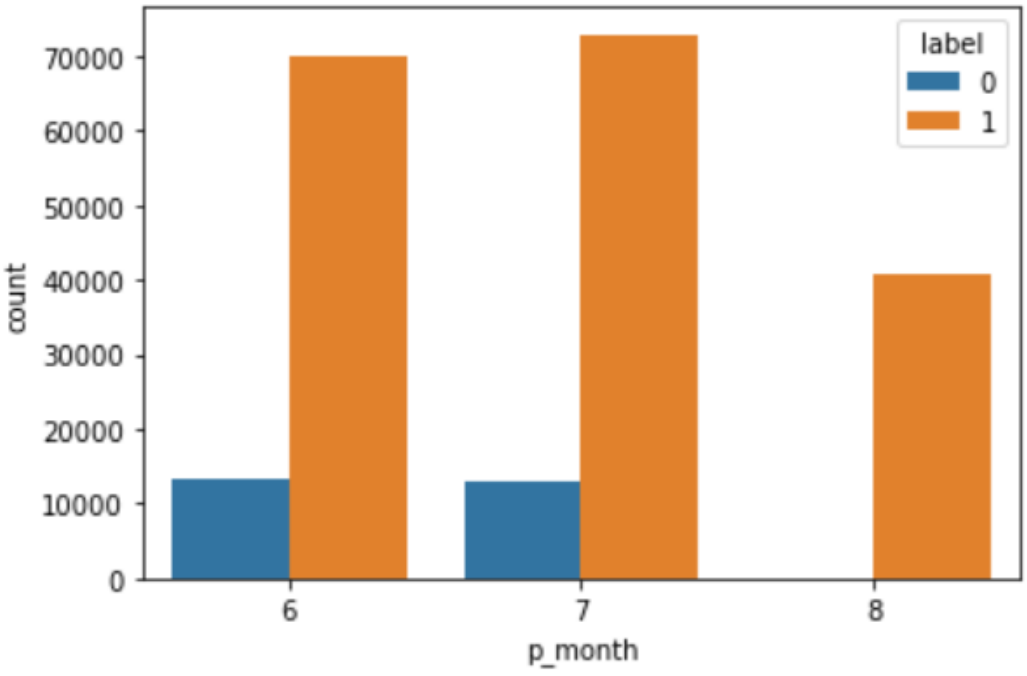
- Visualizations
- I use seaborn library and matplotlib to visualize the data and find the meaning of the data if I see the relation between label and aon with the help of scatterplot than I see most of the data is distributed from 0.5 to 1.0
- And between daily decr_90 and label most of the data is distributed from 0_75000.
- And between the rental_90 and label data is distributed from 0_70000
- And between the fr_ma_rech30 and label most of the data is distributed from 0.5 to 1.0 .
- And between the cnt_ma_rech90 and label a data is distributed from 0 to 50 .

- And with the help of countplot I can also see that there is no failure in 8 month and also see that success rate is higher than failure

I use matplotlib and seaborn library to visualize the data with the help of matplotlib I plot the box plot and with the help of seaborn library I plot the scatterplot and find the relation between the columns



With the help of count plot I can see that success rate is higher than failure, 1 represents success and 0 represents failure.



I can see no failure in 8 month

- Learning Outcomes of the Study in respect of Data Science

Decision tree classifier gives 90 per accuracy rest model is not working in my system so I am unable to define the accuracy pf all the model