MICRO CREDIT LOAN USE CASE

Submitted by:

AVINASH PATEL

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my teacher as well as my intern company FLIPROBO who gave me this golden opportunity to do this wonderful project on the topic MICRO CREDIT LOAN USE CASE, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly I want to mention some sites which helped me when I got stuck somewhere while completing my projects.

Those sites are :

http://scikit-learn.org

https://www.w3schools.com

https://www.youtube.com

https://www.kaggle.com

INTRODUCTION

- ## Business Problem Framing

  Our client provided us with data which contains many factors which could decide that a customer who took any micro loan would return back the loan on time or not at all.

  So, 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

  I think we have to get some knowledge on how this loan giving works ,to which kind of people can give the loan, if someone does not return the loan then what happens, how much loan we can give to people and on how much interest rates ,how frequently can we give loan to the same person etc.

- ## Review of Literature

  Microfinance Institutions
  Microfinance institutions (MFIs) are financial companies that provide small loans to people who do not have any access to banking facilities. The definition of "small loans" varies between countries. In India, all loans that are below Rs.1 lakh can be considered as microloans.

  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.

- Motivation for the Problem Undertaken

  As I am an intern I want to work on as many project as possible for me. So ,I am highly motivated to do this project and to learn new thing and also learn from my mistakes.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

  First of all I checked that how many of the column are categorical columns and how many of them are numerical columns. By doing so I find out that there are only two columns with categorical data i.e. msisdn, pdate . So I have to convert these data into numerical column.

  Now I started to visualise the column individually.

  On creating a Distplot for the columns that I have treated I find out that many of the columns are right skewed i.e. the bell curve is elongated parallel to the x-axis on the right side.

  After that I created a boxplot again to see how much outliers are still left and some of the columns still have many outliers.

  Then I proceed to create a heatmap for each column and find out that the maximum correlation with our target variable is 24% which is of cnt_ma_rech90 column.

  Then I plotted a countplot of the target variable and saw that the label 1 is way higher than label 0.

- Data Sources and their formats

  The data is provided by our client who is in Telecom Industries. The data has 209593 rows and 37 columns. The data is provided in csv(comma separated value) format. There are 2 categorical columns and 35 numerical columns.

- Data Pre-processing Done

For msisdn:As it is the phone number s of the users so it is not useful for the training of model so I dropped it.

For pdate: As for this column I separated the month and day and year and created a separate column for the day and month. As for the year we don't need that data so we deleted it.

Then after doing so we are left with all the columns as numerical columns. Now, I see that many columns have two separate columns for the data of 30 days and 90 days so as we already have the da ta for the 90 days so we deleted all the columns with same data for 30 days. Now I plotted a boxplot to see how many outliers are there in the columns and some of the columns have high amount of outlier so I treated each column according to their unique attributes.

Treating aon column: As this column is "age on cellular network in days" so the average life expectancy in Indonesia is 71.74 years so I took 72 and multiplies it with 365 to get maximum days a person could have the same number and then I compared all the value of this column and if any value exceeds this value then I stored the index for that and deleted all those columns.

Treating last_rech_date_ma column: All the values in this columns are less then 100 and all the above values are too high to be the number of days. Also some of the values are in negative and days could not be in negative.

Treating last_rech_date_da column: By looking at the xml file this column has mostly 0 as values.

Hence I am checking how many 0 there are and 96% of data are zeros. So, I am going to delete this column.

Treating cnt_da_rech90 and fr_da_rech90 columns: According to the df.describe() method we can safely say that more than 75% data is 0 which is of no use. So we are going to drop these columns as well. So we are going to delete this column as well.

Treating cnt_ma_rech90 and fr_ma_rech90 column: As both of these columns have the same definition according to the Data_Descriprion xml file so we are going to drop one of the column. Here cnt_ma_rech90 has a higher correlation than the other so we are going to keep this column and delete the other one.

Treating cnt_loans90 column: We can see there is a abnormal maximum value 4997.517944 in this column and 30 is the 75% of the values so we are going to delete any value higher than 100 just to be on the safe side.

Treating medianamnt_loans90 column: As we can see from ths table that more than 75% of this column's data is 0 so we are going to delete this column.

- Data Inputs- Logic- Output Relationships

  The Remaining Data after all the pre-processing acts as an input data for the Model(in our case Ridge Classifier) the already trained model gives us an out put of an array of 0 and 1 for the given input.

- State the set of assumptions related to the problem under consideration

  1:The average life expectancy of a person in Indonesia is 72 years according to google data.
  2: All the values in this columns are less then 100 and all the above values are too high to be the number of days.
  3:There could not be values of days in negative.

- Hardware and Software Requirements and Tools Used

  Ram:8GB

  ROM:200MB

  Processor : Ryzen 5 3600X(6 cores 12 threads)

  Tool : Jupiter Notebook

  Language Used: Python

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  First I imported all the basic libraries of python used in the problem solving. Then I read carefully the data and the description given to me. After that I imported the provided csv file into the Jupiter notebook. Now first I checked for any nan values in the dataset after that I checked how many number of rows and columns I am dealing with and I also done some more analysis of the raw data to get a general idea of there max and mins and there inter quartile ranges etc.

  After analysing the data I started cleaning the data as explained above.

  After cleaning I did some visualisation of the data to get some insights from our cleaned data.

  I removed skewness and some excess outliers. On the completion of all these steps I moved on to the model selection part. Here I tested many models and choose one of the model to proceed.

  After I selected one of the models I did hyperparameter tuning and finally I saved my model.

- Testing of Identified Approaches (Algorithms)
  - DecisionTreeClassifier
  - RidgeClassifier
  - RandomForestClassifier
  - KNeighborsClassifier
  - ExtraTreesClassifier
- Run and Evaluate selected models

  After running all the models the lowest difference between the accuracy score and the cross validation is for Ridge Classifier.

```
DTC = DecisionTreeClassifier()
classify(DTC,x,y)
```

```
Accuracy: 84.30535521919333
CV Score: 96.09461311325283
Diffrence: 11.789257894059503
```

```
RC = RidgeClassifier()
classify(RC,x,y)
```

```
Accuracy: 76.75838521791525
CV Score: 76.83520099210044
Diffrence: 0.07681577418519225
```

```
RFC = RandomForestClassifier()
classify(RFC,x,y)
```

```
Accuracy: 88.46391206704523
CV Score: 97.78520553902767
Diffrence: 9.321293471982443
```

```
KNN = KNeighborsClassifier()
classify(KNN,x,y)
```

```
Accuracy: 71.32319377750187
CV Score: 85.99082490381737
Diffrence: 14.667631126315499
```

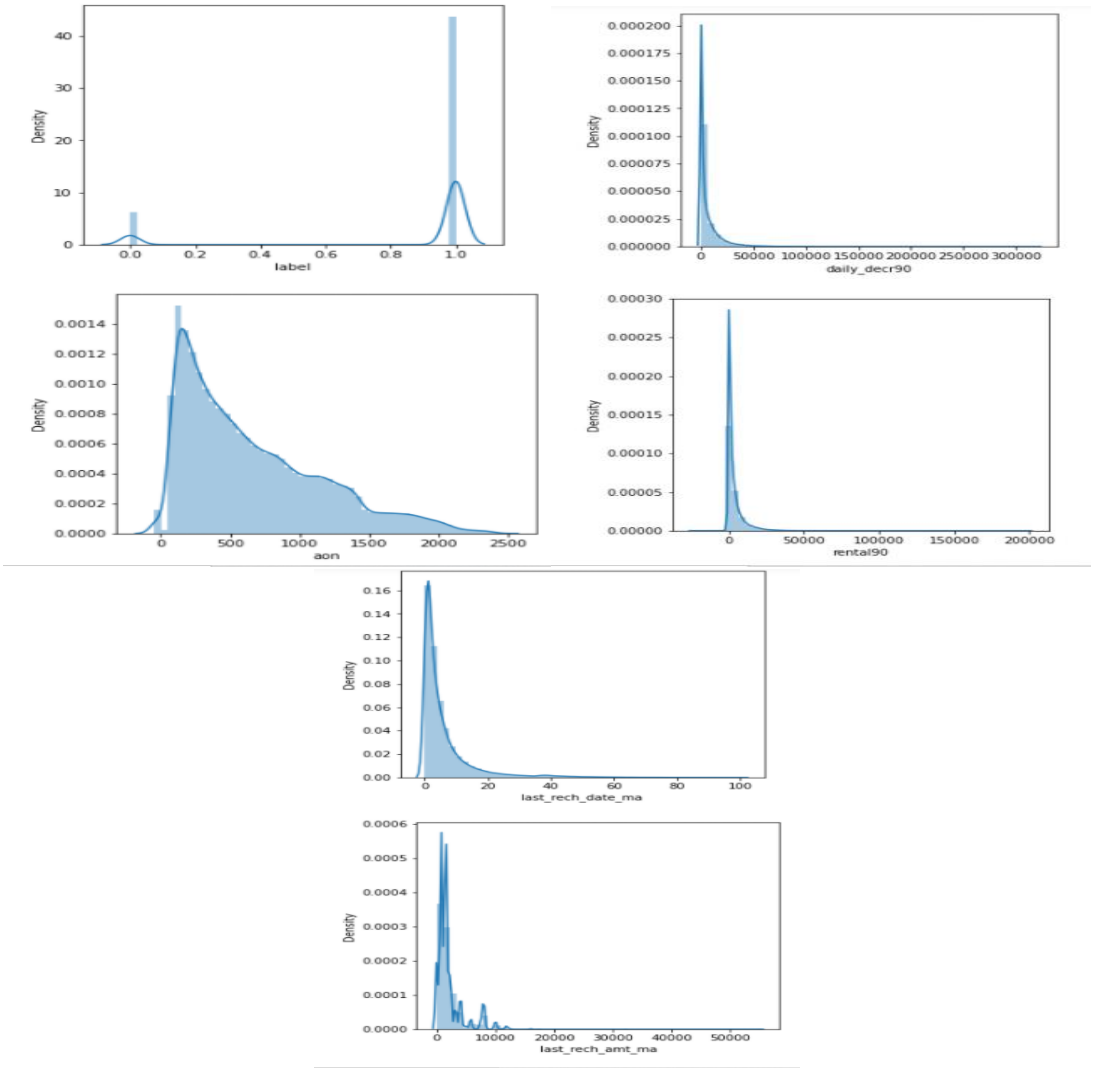
```
ETC = ExtraTreesClassifier()
classify(ETC,x,y)
```

```
Accuracy: 87.83947123372711
CV Score: 98.48523694397377
Diffrence: 10.645765710246664
```

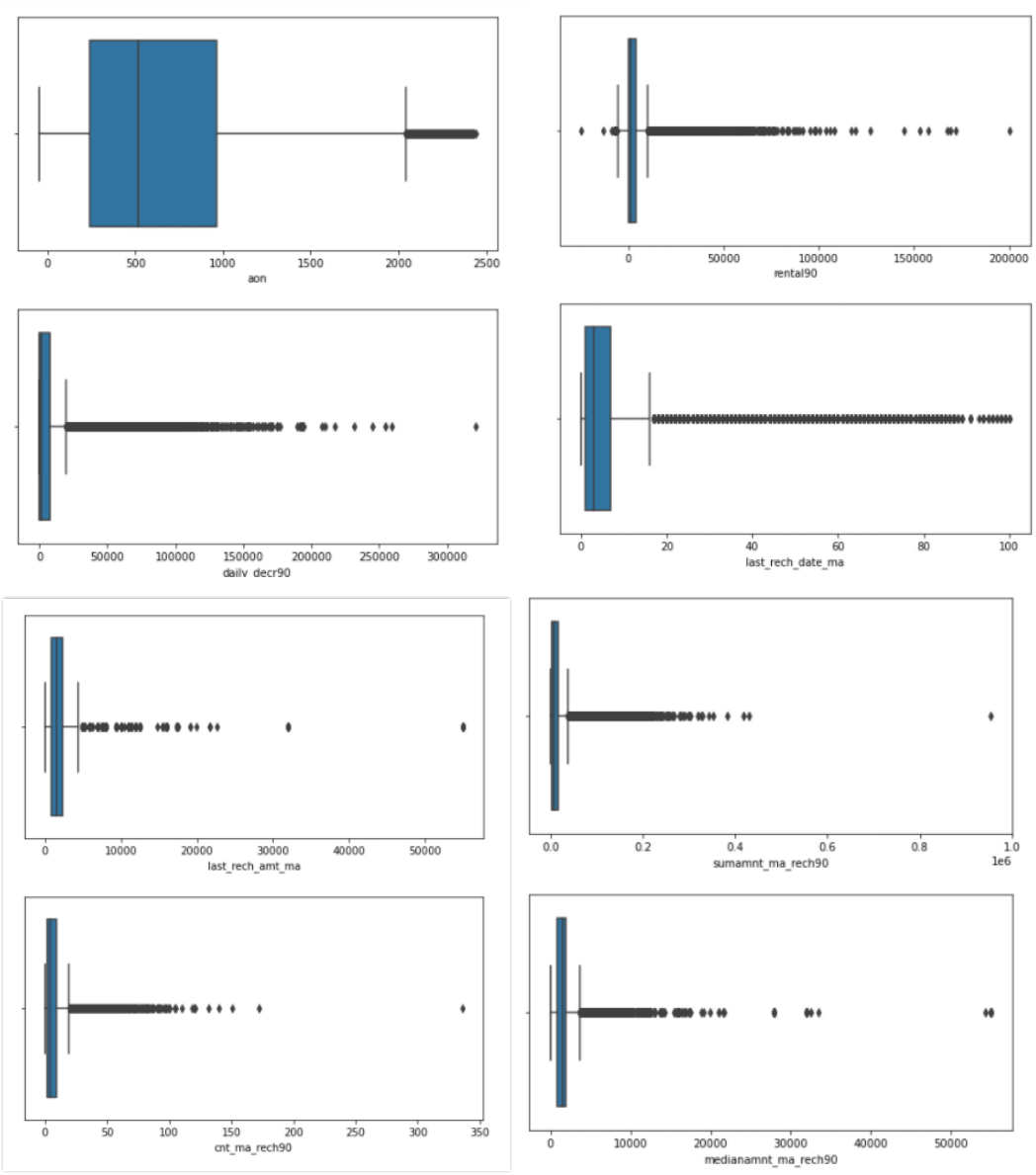
That means ridge classifier is not overfitted nor underfitted.

- Visualizations

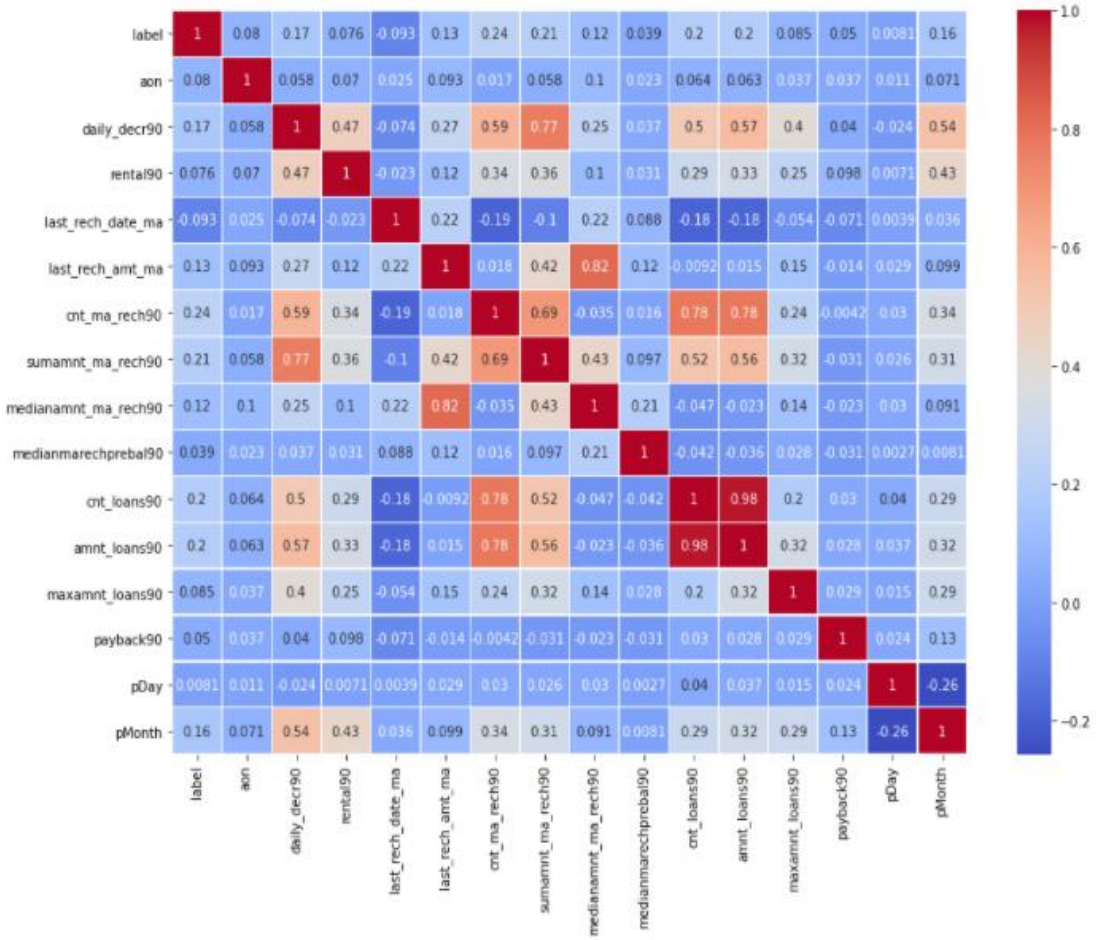
  Distplot:



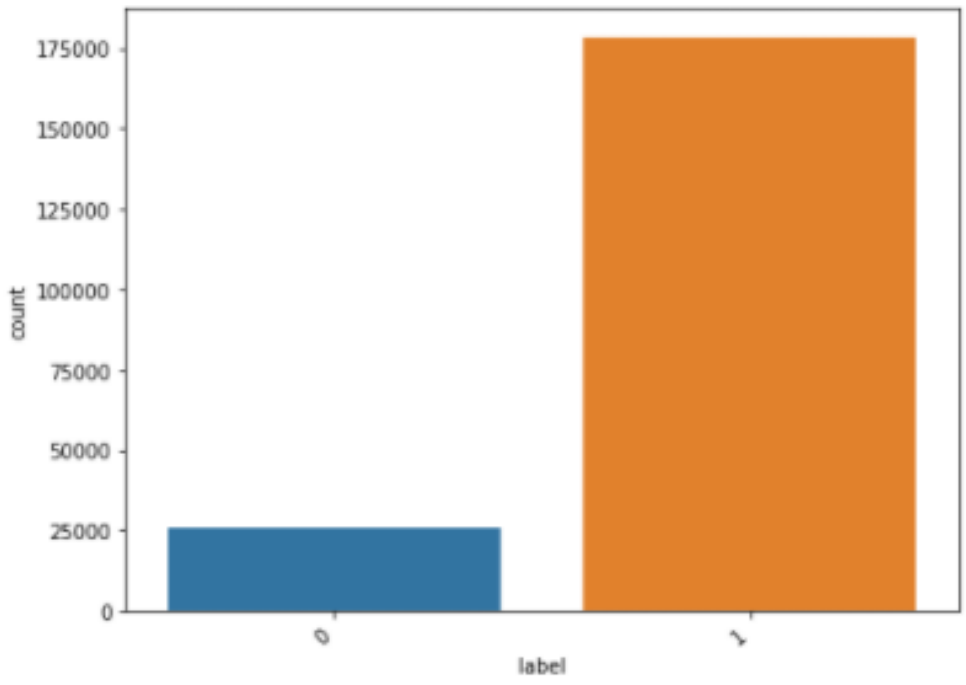
Boxplot:



Heatmap:


label 1 0.08 0.17 0.076 -0.093 0.13 0.24 0.21 0.12 0.039 0.2 0.2 0.085 0.05 0.0081 0.16
aon 0.08 1 0.058 0.07 0.025 0.093 0.017 0.058 0.1 0.023 0.064 0.063 0.037 0.037 0.011 0.071
daily_decr90 0.17 0.058 1 0.47 -0.074 0.27 0.59 0.77 0.25 0.037 0.5 0.57 0.4 0.04 -0.024 0.54
rental90 0.076 0.07 0.47 1 -0.023 0.12 0.34 0.36 0.1 0.031 0.29 0.33 0.25 0.098 0.0071 0.43
last_rech_date_ma -0.093 0.025 -0.074 -0.023 1 0.22 -0.19 -0.1 0.22 0.088 -0.18 -0.18 -0.054 -0.071 0.0039 0.036
last_rech_amt_ma 0.13 0.093 0.27 0.12 0.22 1 0.018 0.42 0.82 0.12 -0.0092 0.015 0.15 -0.014 0.029 0.099
cnt_ma_rech90 0.24 0.017 0.59 0.34 -0.19 0.018 1 0.69 -0.035 0.016 0.78 0.78 0.24 -0.0042 0.03 0.34
sumamnt_ma_rech90 0.21 0.058 0.77 0.36 -0.1 0.42 0.69 1 0.43 0.097 0.52 0.56 0.32 -0.031 0.026 0.31
medianamnt_ma_rech90 0.12 0.1 0.25 0.1 0.22 0.82 -0.035 0.43 1 0.21 -0.047 -0.023 0.14 -0.023 0.03 0.091
medianmarechprebal90 0.039 0.023 0.037 0.031 0.088 0.12 0.016 0.097 0.21 1 -0.042 -0.036 0.028 -0.031 0.0027 0.0081
cnt_loans90 0.2 0.064 0.5 0.29 -0.18 -0.0092 0.78 0.52 -0.047 -0.042 1 0.98 0.2 0.03 0.04 0.29
amnt_loans90 0.2 0.063 0.57 0.33 -0.18 0.015 0.78 0.56 -0.023 -0.036 0.98 1 0.32 0.028 0.037 0.32
maxamnt_loans90 0.085 0.037 0.4 0.25 -0.054 0.15 0.24 0.32 0.14 0.028 0.2 0.32 1 0.029 0.015 0.29
payback90 0.05 0.037 0.04 0.098 -0.071 -0.014 -0.0042 -0.031 -0.023 -0.031 0.03 0.028 0.029 1 0.024 0.13
pDay 0.0081 0.011 -0.024 0.0071 0.0039 0.029 0.03 0.026 0.03 0.0027 0.04 0.037 0.015 0.024 1 -0.26
pMonth 0.16 0.071 0.54 0.43 0.036 0.099 0.34 0.31 0.091 0.0081 0.29 0.32 0.29 0.13 -0.26 1
label
aon
daily_decr90
rental90
last_rech_date_ma
last_rech_amt_ma
cnt_ma_rech90
sumamnt_ma_rech90
medianamnt_ma_rech90
medianmarechprebal90
cnt_loans90
amnt_loans90
maxamnt_loans90
payback90
pDay
pMonth
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Countplot:


175000
150000
125000
100000
75000
50000
25000
0
count
0
1
label

- Interpretation of the Results

  After all the pre processing and then visualising the data I find out that although there are many columns to the dataset ,but most of the columns are useless and the remaining columns do not have a significant correlation with our Target column.

CONCLUSION

- Key Findings and Conclusions of the Study

  I find out that few columns are correlated to our target variable that means predicting our target variable is difficult for any model because there is no direct correlation between them. Also, many of the columns have values which are unrealistic. Many of the columns have highly skewed values and many columns also have high amounts of outlies in them.

- Learning Outcomes of the Study in respect of Data Science

  As I have already told that this dataset has a lot of outliers and I have to look into each and every column to get some of the data cleaned. As per the visualisation part I plotted boxplots to see the number of outliers and in which column they exist and in how much amount.

  I have also plotted Distplot to get a general idea of the columns which have skewness to further treat them.

  I plotted heatmap to see which columns are in a correlation with our target variable.

  I have plotted countplot to see that our target variable is balanced or not.

- Limitations of this work and Scope for Future Work

  As I am not able to fully understand the meaning of all the columns and if I co7uld get someone who is already working then the Telecom company then I could get a better idea of how all the column are working and which outer columns needs to be in the dataset. As per the future scope, yes there is definitely more room to grow as I could get the model efficiency to 76% it could be more on further research.