

Micro-Credit Defaulter Project

Submitted by:

PAVANKUMAR BOLL

**ACKNOWLEDGMENT**

I would like to express my special thanks to flip robo technlogies for given me such a wonderfull of apportunity to explore and insite work on a data science projects.

And and also thanks to Datatraiend institute’s mentors to motivate and understand the concepts of Data science. I have got lot of help from their resourses and concept to complete these project.

Link that helps me to complete this project

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html>

**INTRODUCTION**

Business Problem Framing:-

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 purpose of this sample data is used to improve the selection of customers for the credit, to help them in further investment and improvement in selection of customers.

* Conceptual Background of the Domain Problem

Data is gather from Microfinance Institution (MFI) the aim of this is institutions to 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.

* Motivation for the Problem Undertaken

Motive of this project is to determine the customer is defaulter or not it means customer would pay intrest and repay the loan or not. Motive is to collect the customer those are legitimate to repay the loan with intrest.

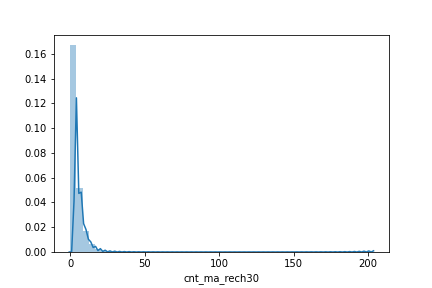
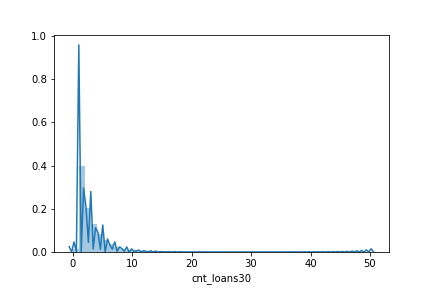
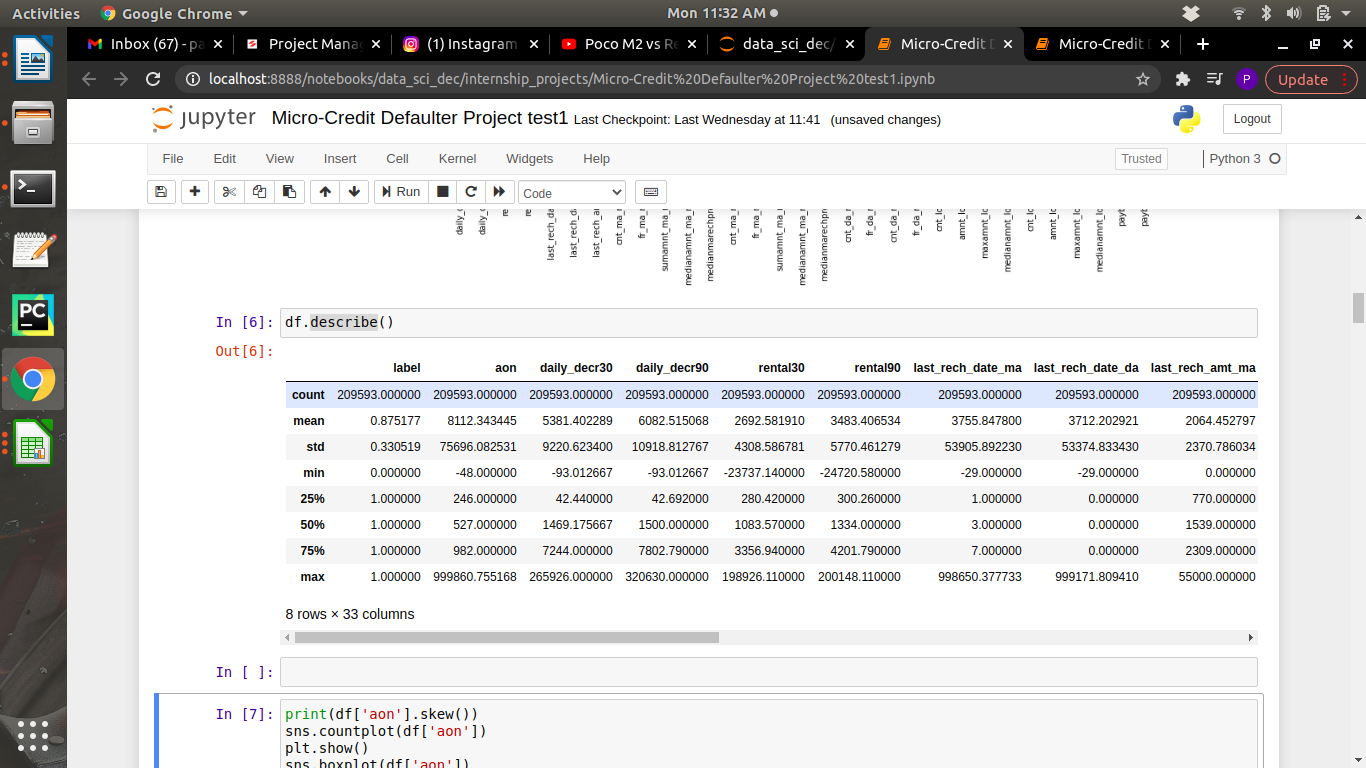
**Analytical Problem Framing**

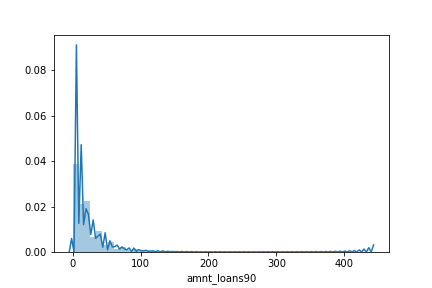
* Mathematical/ Analytical Modeling of the Problem

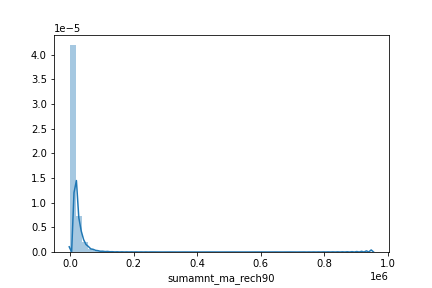
In this sample data gives lot’s of information to predict a customer is defaulter or not. Data contain 209593 rows and 36 Columns. Data has no missing values all the coulmn having some data. Some of the columns are in integer and some are in Object. There is msisdn column type is object each row having a number with one alphabet ‘I’ in middle of the string to convert is object to interger i converted alphabet ‘I’ to blank. Then column is in string number so i typecast each rows to integer same as these i replaced UPW to 1 in pcircle column. I have handel date time column to make day, week and Year column differently so datetime column information will give in three different column with data type integer.

* Data Sources and their formats

The data is dipsersed with an heigh variance on continues data. Minimum value and mean value having large difference so we can say that data have some outliers or having heigh skewness in the data i.e (Right skewed and Left Skewed).





* Data Preprocessing Done

In this data all the columns having data in different diffenret range. So not to bias on columns i have preprocess the data to scale all the columns in a similar range. I have to used the StandardScaler class from preprocessing to Scale all columns in a similar range to do not biase a machine while model generation. By Scaling on Lable data dataset become in a similar range and machine cannot be bias.

* Data Inputs- Logic- Output Relationships

Inputs of the data contains the activities of customer’s in bank account. How each customer frequently debiting and crediting money into their respective bank. How minimum balance is being holding in last 30 days and last 90 days. Logic behind this is that if any customer is not holding minimum amount in last 2 to 3 month’s or he’s fequently debiting amount these type of candidates come’s under defaulter i.e failure or not a good customer.

* Hardware and Software Requirements and Tools Used

Tools and libraries that i have used to solve project:

* Software: Anaconda, Jupyter Notebook, Python3
* Libraries: Numpy, Pandas, Matplotlib, Seaborn,

Sklearn.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The data set contain lable data and target data lable has cleaned by some EDA steps. Target data is in the form of classified factor’s 1 / 0 i.e (success and failure). By seeing the our target column we can conclude that this problem comes under supervised and Classification Problem. Target column having in factor logistic regression can be used for predictions because it is best when the target column having in factors. For more testing SVC, Decision Tress Classifier, KNN algorithms have used. Rather testing on one by one algorithm and choosing best alogithm and hyper tuning that best model. I have hyper tunned all the algorithm and added into pipeline so i will get best algorithm along with the best parameters.

* Testing of Identified Approaches (Algorithms)

For testing how good model is generated. I have splitted data set into two parts (70% 30%) and each part is splited into further two part that are lable data and target data. First 70% splited data is used for training the model and another rest of 30% used for Prediction testing the model how good model is predicted in terms of percentage.

List of Algorithms that i have used:-

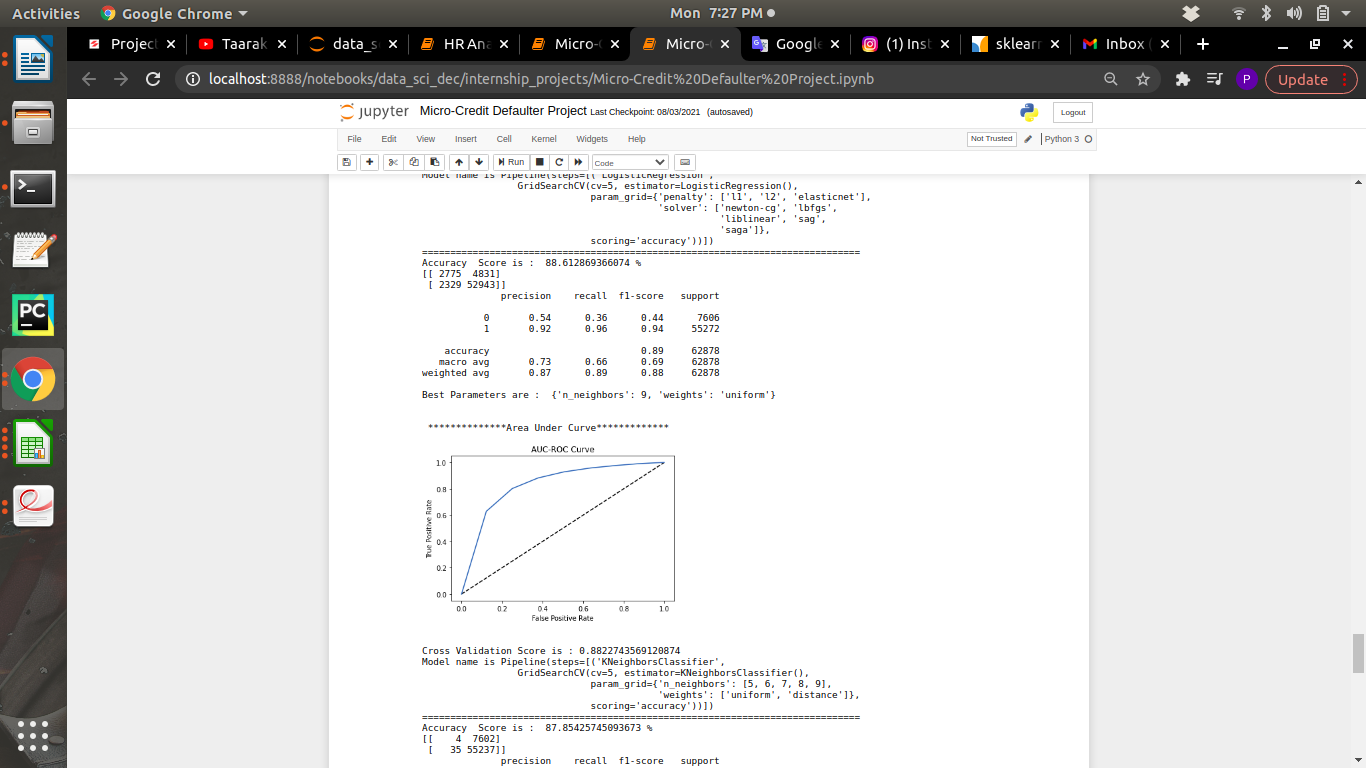
1. Logistic Regression
2. Decision Tress Classifier
3. KNN
4. SVC

* Run and Evaluate selected models

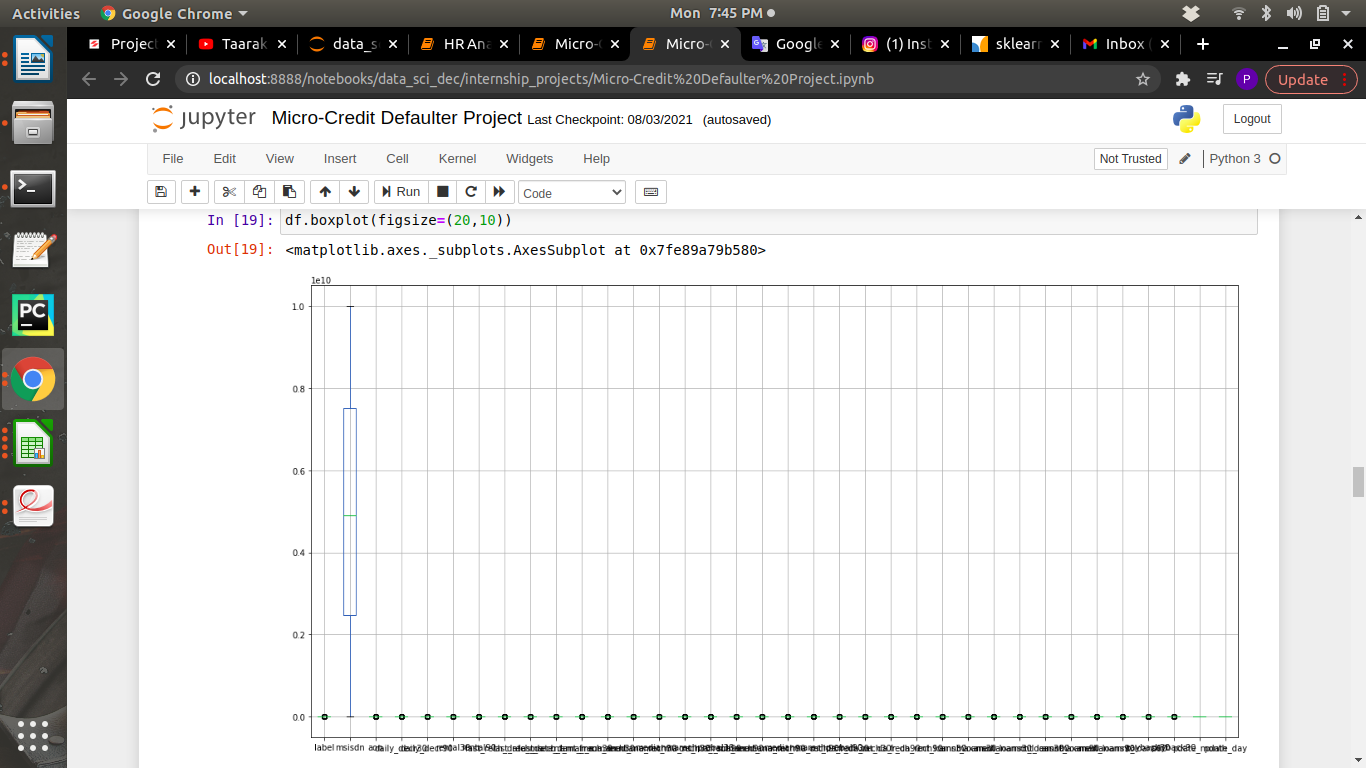
After running Piple it has gave best model is KNN Classifier with a 88% accuracy. For to represent accuracy in graphycal manner i have plot AUC ROC curve. KNN Classifier Algorithm is used collect n naighbors to that point and rank that points.

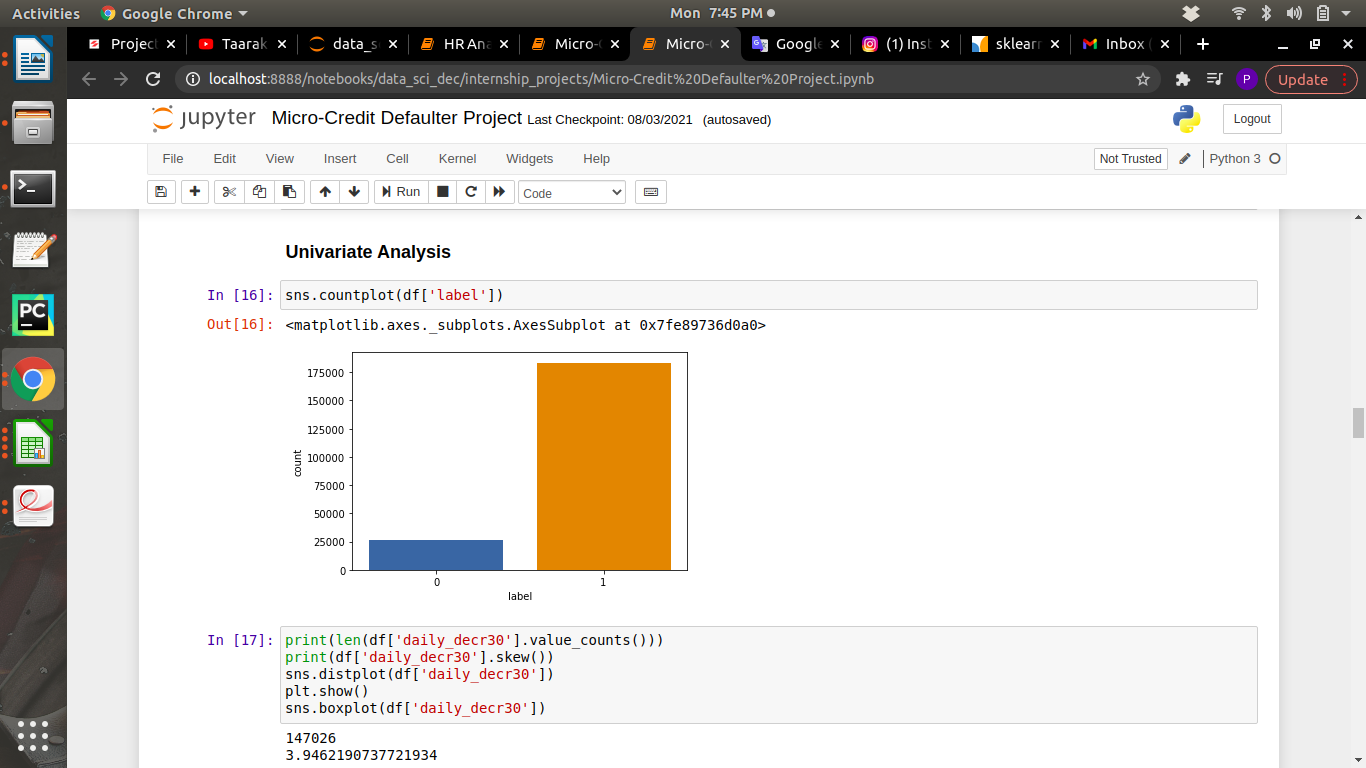
E.g n\_naighbors=7

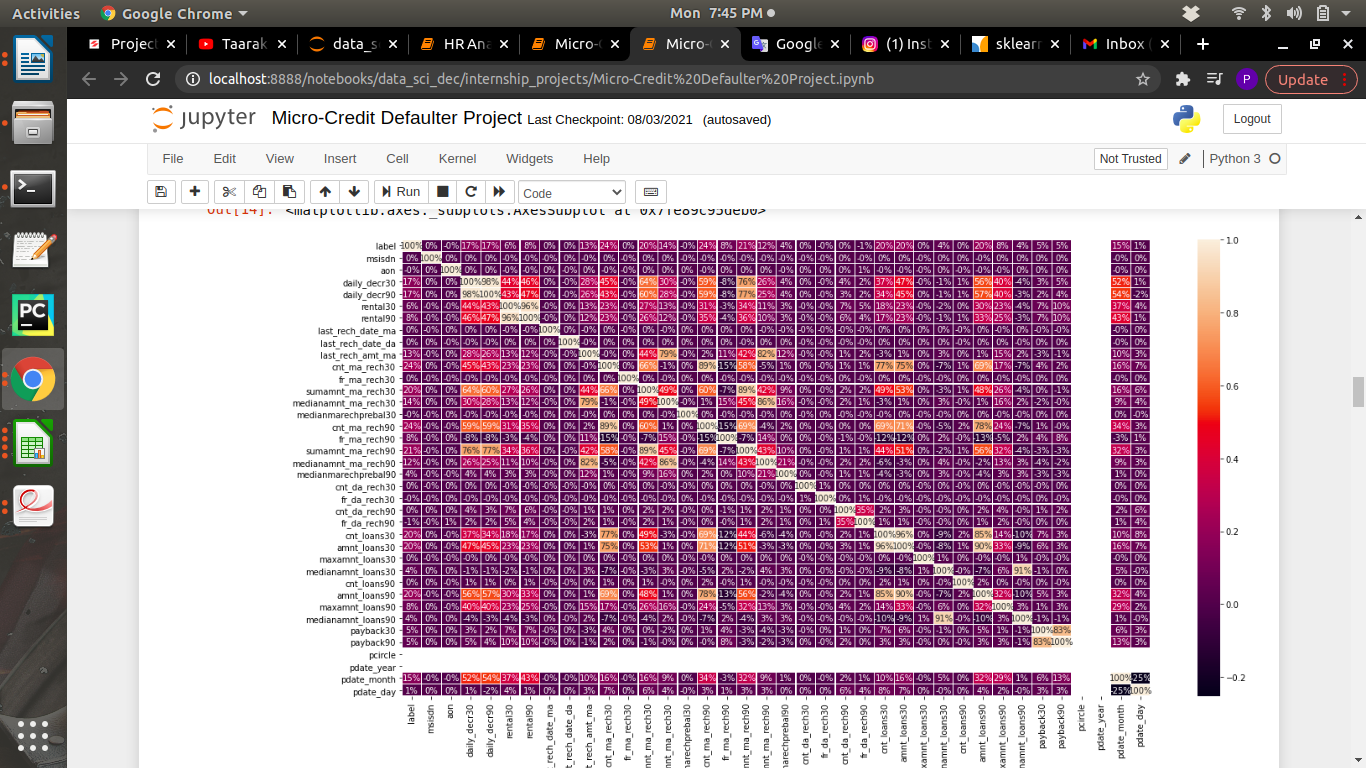
Any rondom point is used to collect 7 points neartest to the random ponit and rank that points If out of 7points 4points represents 1(success) and rest 3 are 0(failure). Heighest points are 1 near to that random point it will predict that random point is 1(success).



* Visualizations

Box plot show tt there is no outliers present in the dataset.

Countplot showing the count of 0’s and 1’s. It show that 1’s have more than 75% and 0’s have near to 25%

Above heatmap shows percentage of correlating with taget column i.e label. It helps a lot to understand which column heigh correlating with target column wich affects more while predecting.

* Interpretation of the Results

From visualization i got to know that data set have no null values, how much correalating with target column and how the data is skewd and distributed. And count of defaulter and non defaulter. By preprocessing and scalling the data data set become unbiased while model generating. Lastly by applying some algorithm in pipeline it has given the best model with the best parameter. Best model has given with the best parameter so the problem of over fitting and underfitting is solved.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

I have learn some inbuild funtion of python at the time of data cleaning. Object columns to Integer are done by using python map function over on lambda function. The main challenging is that to run all model once and get the best model name with best parameters also with a AUC ROC curve plot. This is done by using pipeline. This will helps me to get out from overfitting and underfitting.