Micro_Cerdit_Defaulter Predicition Model

Submitted by:

Ganesh Prasad Bhandari

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

For this problem micro loan defaulter, I have done R & D via internet and did google and got the knowledge regarding the problem faced by these types of financial institution after providing loan to small, impoverished people who even don’t have basic needs fulfilled i.e. food, clothes and shelters as well. Here, Wikipedia and other different articles including newspaper I have read for understanding the problem and completion of the project with more than 98% accuracy.

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 sample data is provided to us from our client database. It is hereby given to you for this exercise. 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

  Describe the domain related concepts that you think will be useful for better understanding of the project.

MFI can only be survived if they can get a model which can predict the loan borrower would be more likely to default or repay the loan in the given time period. In these type of financial institution if more than 2% or 3 % default happens then its cause repercussion to the health of these institution and would not be survive in long run because of bad debts. So, for understanding we should take help of the financial analyst as well.

- Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

To overcome this type of problem, their have done no. of researches from different institutions and by private agencies. These are non profit organisation which servers millions of pennyless persons to make them self dependent via starting or doing their own work i.e. doing agriculture, making hand baskets, handicraft items, soap making, aggarwatti (fragrance) making, crispy papad making at home or in a small room etc. So, to support these type of institution government should also do some aid and motivation to the them in different type of rebate forms like rebate in tax rates, discount in different type of institutional assets purchasing etc.

- Motivation for the Problem Undertaken

  Describe your objective behind to make this project, this domain and what is the motivation behind.

  I was motivate because of the TechRobo Technology's project to make a robust model which can generalize the prediction of defaulter of loan in future via understanding the patterns and behaviour of the person via their data available with these institution or data collection by 3rd party make available to us for making new predictive models . We want to make help to these institution do their work properly and in large scale so every poor will get benefit from these type of institutes and become self dependent for their earnings and in turn reduce poverty and all.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

  For making this predictive modelling we used lot of mathematical and stastical analytical functions and model. Like for outliers finding we use box plot , pairplot, correlation coefficient and find the skewness of the data i.e. Is it skewed or not, if yes then right skewed or left and then I decided to use IQR (inter quartile range) method to find the outliers and set the threshold to get rid of it instead of Z score method with 3 std deviation etc. It all depends upon the problem statement what we have to use and how much for threshold to declare the outliers or not.

- Data Sources and their formats

  What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

  Data sources were internet, Kaggle and google for all most cases but some time its provided by some private agencies and companies itself, but I got it from my company i.e. TechRobo Technology, Bangalore. Description of the data is given below. It was in csv formate.

Variable	Definition	Comment
label	Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}	
msisdn	mobile number of user	
aon	age on cellular network in days	
daily_decr30	Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)	
daily_decr90	Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)	
rental30	Average main account balance over last 30 days	Unsure of given definition
rental90	Average main account balance over last 90 days	Unsure of given definition
last_rech_date_ma	Number of days till last recharge of main account	
last_rech_date_da	Number of days till last recharge of data account	
last_rech_amt_ma	Amount of last recharge of main account (in Indonesian Rupiah)	
cnt_ma_rech30	Number of times main account got recharged in last 30 days	
fr_ma_rech30	Frequency of main account recharged in last 30 days	Unsure of given definition
sumamnt_ma_rech30	Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)	
medianamnt_ma_rech30	Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)	
medianmarechprebal30	Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)	
cnt_ma_rech90	Number of times main account got recharged in last 90 days	
fr_ma_rech90	Frequency of main account recharged in last 90 days	Unsure of given definition
sumamnt_ma_rech90	Total amount of recharge in main account over last 90 days (in Indonesian Rupiah)	
medianamnt_ma_rech90	Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah)	
medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level (in Indonesian Rupiah)	
cnt_da_rech30	Number of times data account got recharged in last 30 days	
fr_da_rech30	Frequency of data account recharged in last 30 days	
cnt_da_rech90	Number of times data account got recharged in last 90 days	
fr_da_rech90	Frequency of data account recharged in last 90 days	
cnt_loans30	Number of loans taken by user in last 30 days	
amnt_loans30	Total amount of loans taken by user in last 30 days	
maxamnt_loans30	maximum amount of loan taken by the user in last 30 days	There are only two options: 5 & 10 Rs., for which the user needs to pay back 6 & 12 Rs. respectively
medianamnt_loans30	Median of amounts of loan taken by the user in last 30 days	
cnt_loans90	Number of loans taken by user in last 90 days	
amnt_loans90	Total amount of loans taken by user in last 90 days	
maxamnt_loans90	maximum amount of loan taken by the user in last 90 days	
medianamnt_loans90	Median of amounts of loan taken by the user in last 90 days	
payback30	Average payback time in days over last 30 days	
payback90	Average payback time in days over last 90 days	
pcircle	telecom circle	
pdate	date	

- Data Preprocessing Done

  What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

  Here in data preprocessing I first looked at the data in a tabular form and then also with histogram and other correlation matrix and pairplot etc. We also did 5 point summary to see the mean, median and mode and 25percentile to 75 percentile etc. Null values, Missing values, data types etc as well. We also did feature importance via decision tree to find the top most predictable features which we used in modelling and drop some useless features too. We used up sampling also to make both the classes

proportionally equal for unbiased prediction. Then I did scaling of the dataset also for training data set.

- 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.

  Data have 209593 records and 37 features (col) and from these we had 3 object columns and remaining integer and float types. We delete all object columns as they have of no use to this prediction and we checked for null or missing values and luckily we didn't have any missing values. But in this dataset we have lot of outliers in many features and we had treated well with IQR method. After that we scaled the data and then drop some of the features which were useless or have no impact in prediction.

  We did all EDA and data pre-processing and feature engineering and feature selection part which took me almost 60% time.

  After that we split the data into X and Y and then train and test split which was already scaled data but without feature extraction and reduction make ready for used in modelling or fitting.

  Then I used different models like Logistic Regression, Naïve Bayes, Knn, XGBoostClassifier and SVM as well but I didn't get good accuracy.

  Then I moved to other ensembled models and got the accuracy more than 98% which I was targeting to get good predicitive model

and I did in the end. Then I did prediction and now this model is available for productionalization and deployment .

Here, the most important thing is that we have to be very careful at the time of EDA and data pre-processing and feature engineering and feature selection because as we all know that if our input data is good then our output will be good, so we have to work deeply in data pre-processing part with lot of time and efforts.

In future we can do hyperparameter tunning as well to get more accuracy as well.

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

  Here, you can describe any presumptions taken by you.

Here, In this model making I thought that data can be more impure and with lot of noise also but in this dataset I got no missing values, no categorical features which we had to make to numerical variables with label encoder or one hot encoder etc. So, this data was good to process after doing some data pre-processing.

- ## Hardware and Software Requirements and Tools Used

  Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

  For this project, I used Jupyter notebook with all python and open source libraries like pandas for data frame, NumPy for numerical calculation, SciPy for scientific calculation, scikit learn, TensorFlow, keras, os and other useful libraries.

  In this project, I faced lot of problems related to hardware and software like for running the pairplot , heatmap it took lot of time and still not complete and sometime system hanged as well. I can't run SVM and other algorithm because of the limitation of the hardware gpu processor, ram etc. I didn't do properly this project because of lack of proper hardware and software infrastructure.

  To conclude, I want to say that for good predicitive modelling we should run lot of algorithm and for the same we should have lot of resources and to overcome lack of resources problem we can use cloud basis like docker container and Kubernetes enginee in AWS or with other cloud service provider like Azure, GCP etc.

  Hope, in future our company will provide the cloud platform for running and deployment of the model to its fullest level.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving of this problem.

  Well in this project of finding micro credit defaulter prediction I firstly used the data cleaning and pre-processing step i.e.

  1) To look deeper at the dataset first, to know what type of data we have and what data want to tell us.
  2) We used different visualization technique via matplotlib and seaborn libraries to understand the corelation and hidden pattern of the data.
  3) Understand the collinearity matrix and know that Is their any collinearity within the independent features or with dependent features, it is because if we have any multicollinearity features then above our set threshold, we should remove those features otherwise it will confused our model and get impact to our prediction.
  4) Then I checked missing value or null values if any but luckily we didn't face this problem in this dataset otherwise we have to fill those null values with some statistical methods like mean of that column or median or mode etc.
  5) We used statistical method of finding outliers i.e. IQR to set the threshold to get the outliers and then removed all those. We can also use Z score method etc. We also visualize outliers through box plot here but we can use other plot like scatter plot etc
  6) Then after treating outliers now I checked the imbalanced data and I found out that data is totally imbalance and the ration is about 1: 7 almost. So, I used up sampling technique to balanced

the both classes because if we don't do data balancing then it will be biased towards majority classes and our model will predict wrongly.

7) I did other feature engineering technique as well to make the data clean and noise free to inject in our model. We used feature importance method and dropped some of the features which were not so important to our model prediction as well.

- Testing of Identified Approaches (Algorithms)

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

For any model we have to use different algorithms via spot selection of algo using pipeline and get the best also from the list of algos. Here, I used some algo's like

Logistic Regression

Naïve Bayes

KNN

SVM

XGBoost Classifier etc

LinearDiscriminantAnalysis

DecisionTreeClassifier (CART)

AdaBoostClassifier

GradientBoostingClassifier

RandomForestClassifier

ExtraTreesClassifier

I used all these algo's for training and testing with different hit and trial method with sometimes up sampling and features dropped and sometime without up sampling and without dimensionality reduction i,e, dropping of features as well.

- 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.

These algo's mentioned above I used without up-sampling and feature engineering but with scaling to know how much accuracy and F1 score I will get for this. I didn't get good accuracy or F1 Score more than

87% and this is not good model for our business problem to give loan to someone who may default at the time of repayment of it. We at least need 95% or more accuracy or F1 score to make our model to make correct or less wrong prediction. I did scaling and feature dropping and up sampling and used different algo's here also to find the best model. Models I used are given below. i.e.

Logistic Regression

Naïve Bayes (GaussianNB)

LinearDiscriminantAnalysis

KNeighborsClassifier

DecisionTreeClassifier (CART)

After using this here I got 95% accuracy, so I again tried with other ensemble techniques i.e.

AdaBoostClassifier

GradientBoostingClassifier

RandomForestClassifier

ExtraTreesClassifier

I evaluate the model on the basis of Accuracy score and F1 score and with precision and recall as well and in the last after using ensemble technique model I got the best accuracy without tuning hyperparameter but we can do hyperparameter tuning in future as well.

Now I got the best score for my model i.e. 98% with ExtraTreesClassifier and I did prediction with this model. Now we can use to deploy this model in generalised way for our problem.

- Key Metrics for success in solving problem under consideration

  What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

I used Accuracy score, F1 score and Precision recall metrics to know the success of my model. Some times only rely on accuracy metrics is not good . for instance

We have a dataset consist of 95% non-defaulter vs % 5% defaulter ration and our model's accuracy score would be 95% and it predict all person as a non-defaulter but here is 5% who did default and it got a big loss to the financial institution, so how we can reply on this accuracy of model. So, sometime we should considered F1 score which include both precision and recall and do almost great prediction as well.

We used confusion matrix to see the True positive, false positive, false negative and true negative ration. For loan default prediction we have to considered more on type 1 error.

This is the blog from minitab.com for type1 and type 2 errors

THE NULL HYPOTHESIS AND TYPE 1 AND 2 ERRORS

When statisticians refer to Type I and Type II errors, we're talking about the two ways we can make a mistake regarding the null hypothesis (Ho). The null hypothesis is the default position, akin to the idea of "innocent until proven guilty." We begin any hypothesis test with the assumption that the null hypothesis is correct.

We commit a Type 1 error if we reject the null hypothesis when it is true. This is a false positive, like a fire alarm that rings when there's no fire.

A Type 2 error happens if we fail to reject the null when it is not true. This is a false negative—like an alarm that fails to sound when there *is* a fire.

It's easier to understand in the table below, which you'll see a version of in every statistical textbook:

Reality	*Null (H_0) not rejected*	*Null (H_0) rejected*
Null (H_0) is true.	Correct conclusion.	**Type 1 error**
Null (H_0) is false.	**Type 2 error**	Correct conclusion.

These errors relate to the statistical concepts of risk, significance, and power.

REDUCING THE RISK OF STATISTICAL ERRORS

Statisticians call the risk, or probability, of making a Type I error "alpha," aka "significance level." In other words, it's your willingness to risk rejecting the null when it's true. Alpha is commonly set at 0.05, which is a 5 percent chance of rejecting the null when it is true. The lower the alpha, the less your risk of rejecting the null incorrectly. In life-or-death situations, for example, an alpha of 0.01 reduces the chance of a Type I error to just 1 percent.

A Type 2 error relates to the concept of "power," and the probability of making this error is referred to as "beta." We can reduce our risk of making a Type II error by making sure our test has enough power—which depends on whether the sample size is sufficiently large to detect a difference when it exists.

- 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.

For the visualizations of the data I only used limited plot and descriptions it is because of lack of hardware resources and gpu's otherwise I was first planning to plot pairplot, heatmaps, barplots and other visualization graphs. Here, I just used correlation matrix, 5 point summary and small heatmap which is not able to give any insights about the features at all. Yes box plot did give us little bit about outliers.

So, cloud platform is very important for these types of ML projects.

```
df1.corr()
```

	label	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	last_rech_amt_ma	cnt_ma
label	1.000000	-0.003785	0.168298	0.166150	0.058085	0.075521	0.003728	0.001711	0.131804	0
aon	-0.003785	1.000000	0.001104	0.000374	-0.000960	-0.000790	0.001692	-0.001693	0.004256	-0
daily_decr30	0.168298	0.001104	1.000000	0.977704	0.442066	0.458977	0.000487	-0.001636	0.275837	0
daily_decr90	0.166150	0.000374	0.977704	1.000000	0.434685	0.471730	0.000908	-0.001886	0.264131	0
rental30	0.058085	-0.000960	0.442066	0.434685	1.000000	0.955237	-0.001095	0.003261	0.127271	0
rental90	0.075521	-0.000790	0.458977	0.471730	0.955237	1.000000	-0.001688	0.002794	0.121416	0
last_rech_date_ma	0.003728	0.001692	0.000487	0.000908	-0.001095	-0.001688	1.000000	0.001790	-0.000147	0
last_rech_date_da	0.001711	-0.001693	-0.001636	-0.001886	0.003261	0.002794	0.001790	1.000000	-0.000149	0
last_rech_amt_ma	0.131804	0.004256	0.275837	0.264131	0.127271	0.121416	-0.000147	-0.000149	1.000000	-0
cnt_ma_rech30	0.237331	-0.003148	0.451385	0.426707	0.233343	0.230260	0.004311	0.001549	-0.002662	1
fr_ma_rech30	0.001330	-0.001163	-0.000577	-0.000343	-0.001219	-0.000503	-0.001629	0.001158	0.002876	0
sumamnt_ma_rech30	0.202828	0.000707	0.636536	0.603886	0.272649	0.259709	0.002105	0.000046	0.440821	0
medianamnt_ma_rech30	0.141490	0.004306	0.295356	0.282960	0.129853	0.120242	-0.001358	0.001037	0.794646	-0
medianmarechprebal30	-0.004829	0.003930	-0.001153	-0.000746	-0.001415	-0.001237	0.004071	0.002849	-0.002342	0
cnt_ma_rech90	0.236392	-0.002725	0.587338	0.593069	0.312118	0.345293	0.004263	0.001272	0.016707	0
fr_ma_rech90	0.084385	0.004401	-0.078299	-0.079530	-0.033530	-0.036524	0.001414	0.000798	0.106267	-0
sumamnt_ma_rech90	0.205793	0.001011	0.762981	0.768817	0.342306	0.360601	0.002243	-0.000414	0.418735	0
medianamnt_ma_rech90	0.120855	0.004909	0.257847	0.250518	0.110356	0.103151	-0.000726	0.000219	0.818734	-0
medianmarechprebal90	0.039300	-0.000859	0.037495	0.036382	0.027170	0.029547	-0.001086	0.004158	0.124646	0
cnt_da_rech30	0.003827	0.001564	0.000700	0.000661	-0.001105	-0.000548	-0.003467	-0.003628	-0.001837	0
fr_da_rech30	-0.000027	0.000892	-0.001499	-0.001570	-0.002558	-0.002345	-0.003626	-0.000074	-0.003230	-0
cnt_da_rech90	0.002999	0.001121	0.038814	0.031155	0.072255	0.056282	-0.003538	-0.001859	0.014779	0

fr_da_rech30	-0.000027	0.000892	-0.001499	-0.001570	-0.002558	-0.002345	-0.003626	-0.000074	-0.003230
cnt_da_rech90	0.002999	0.001121	0.038814	0.031155	0.072255	0.056282	-0.003538	-0.001859	0.014779
fr_da_rech90	-0.005418	0.005395	0.020673	0.016437	0.046761	0.036886	-0.002395	-0.000203	0.016042
cnt_loans30	0.196283	-0.001826	0.366116	0.340387	0.180203	0.171595	0.001193	0.000380	-0.027612
amnt_loans30	0.197272	-0.001726	0.471492	0.447869	0.233453	0.231906	0.000903	0.000536	0.008502
maxamnt_loans30	0.000248	-0.002764	-0.000028	0.000025	-0.000864	-0.001411	0.000928	0.000503	0.001000
medianamnt_loans30	0.044589	0.004664	-0.011610	-0.005591	-0.016482	-0.009467	0.001835	0.000061	0.028370
cnt_loans90	0.004733	-0.000611	0.008962	0.009446	0.004012	0.005141	-0.000225	-0.000972	0.000093
amnt_loans90	0.199788	-0.002319	0.563496	0.567204	0.298943	0.327436	0.000870	0.000519	0.014067
maxamnt_loans90	0.084144	-0.001191	0.400199	0.397251	0.234211	0.251029	-0.001123	0.001524	0.148460
medianamnt_loans90	0.035747	0.002771	-0.037305	-0.034686	-0.035489	-0.034122	0.002771	-0.002239	0.021004
payback30	0.048336	0.001940	0.026915	0.019400	0.072974	0.067110	-0.002233	0.000077	-0.027369
payback90	0.049183	0.002203	0.047175	0.040800	0.095147	0.099501	-0.001583	0.000417	-0.014260

33 rows × 33 columns

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	209593.0	104797.000000	60504.431823	1.000000	52399.000	104797.000000	157195.00	209593.000000
label	209593.0	0.875177	0.330519	0.000000	1.000	1.000000	1.00	1.000000
aon	209593.0	8112.343445	75696.082531	-48.000000	246.000	527.000000	982.00	999860.755168
daily_decr30	209593.0	5381.402289	9220.623400	-93.012667	42.440	1469.175667	7244.00	265926.000000
daily_decr90	209593.0	6082.515068	10918.812767	-93.012667	42.692	1500.000000	7802.79	320630.000000
rental30	209593.0	2692.581910	4308.586781	-23737.140000	280.420	1083.570000	3356.94	198926.110000
rental90	209593.0	3483.406534	5770.461279	-24720.580000	300.260	1334.000000	4201.79	200148.110000
last_rech_date_ma	209593.0	3755.847800	53905.892230	-29.000000	1.000	3.000000	7.00	998650.377733
last_rech_date_da	209593.0	3712.202921	53374.833430	-29.000000	0.000	0.000000	0.00	999171.809410
last_rech_amt_ma	209593.0	2064.452797	2370.786034	0.000000	770.000	1539.000000	2309.00	55000.000000
cnt_ma_rech30	209593.0	3.978057	4.256090	0.000000	1.000	3.000000	5.00	203.000000
fr_ma_rech30	209593.0	3737.355121	53643.625172	0.000000	0.000	2.000000	6.00	999606.368132
sumamnt_ma_rech30	209593.0	7704.501157	10139.621714	0.000000	1540.000	4628.000000	10010.00	810096.000000
medianamnt_ma_rech30	209593.0	1812.817952	2070.864620	0.000000	770.000	1539.000000	1924.00	55000.000000
medianmarechprebal30	209593.0	3851.927942	54006.374433	-200.000000	11.000	33.900000	83.00	999479.419319
cnt_ma_rech90	209593.0	6.315430	7.193470	0.000000	2.000	4.000000	8.00	336.000000
fr_ma_rech90	209593.0	7.716780	12.590251	0.000000	0.000	2.000000	8.00	88.000000
sumamnt_ma_rech90	209593.0	12396.218352	16857.793882	0.000000	2317.000	7226.000000	16000.00	953036.000000
medianamnt_ma_rech90	209593.0	1864.595821	2081.680664	0.000000	773.000	1539.000000	1924.00	55000.000000
medianmarechprebal90	209593.0	92.025541	369.215658	-200.000000	14.600	36.000000	79.31	41456.500000
cnt_da_rech30	209593.0	262.578110	4183.897978	0.000000	0.000	0.000000	0.00	99914.441420
fr_da_rech30	209593.0	3749.494447	53885.414979	0.000000	0.000	0.000000	0.00	999809.240107
cnt_da_rech90	209593.0	0.041495	0.397556	0.000000	0.000	0.000000	0.00	38.000000
fr_da_rech90	209593.0	0.045712	0.951386	0.000000	0.000	0.000000	0.00	64.000000

fr_da_rech90	209593.0	0.045712	0.951386	0.000000	0.000	0.000000	0.00	64.000000
cnt_loans30	209593.0	2.758981	2.554502	0.000000	1.000	2.000000	4.00	50.000000
amnt_loans30	209593.0	17.952021	17.379741	0.000000	6.000	12.000000	24.00	306.000000
maxamnt_loans30	209593.0	274.658747	4245.264648	0.000000	6.000	6.000000	6.00	99864.560864
medianamnt_loans30	209593.0	0.054029	0.218039	0.000000	0.000	0.000000	0.00	3.000000
cnt_loans90	209593.0	18.520919	224.797423	0.000000	1.000	2.000000	5.00	4997.517944
amnt_loans90	209593.0	23.645398	26.469861	0.000000	6.000	12.000000	30.00	438.000000
maxamnt_loans90	209593.0	6.703134	2.103864	0.000000	6.000	6.000000	6.00	12.000000
medianamnt_loans90	209593.0	0.046077	0.200692	0.000000	0.000	0.000000	0.00	3.000000
payback30	209593.0	3.398826	8.813729	0.000000	0.000	0.000000	3.75	171.500000
payback90	209593.0	4.321485	10.308108	0.000000	0.000	1.666667	4.50	171.500000

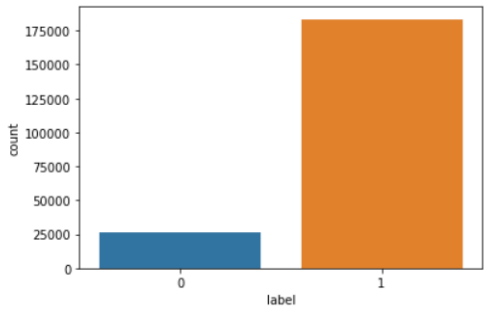
Insights:-

In the above table, you can see the operation between the data in a certain period such as max, min, std. It describe the data how data is distributed.

```
sns.countplot(df2['label'])
```

```
C:\Users\Ganesh\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword
arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit k
eyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='label', ylabel='count'>
```



Insight:-

In the above countplot X axis 0 represents loan defaulters persons and 1 shows Non defaulters persons who have paid loans.

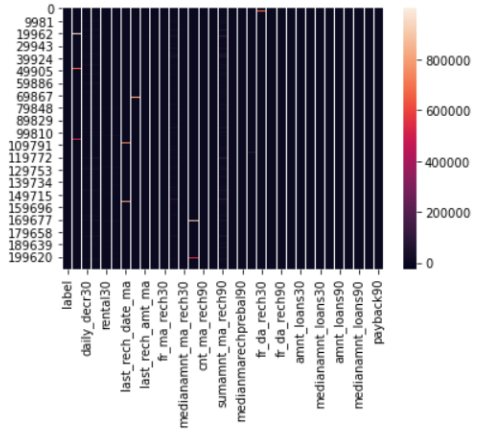
```
df2['label'].value_counts().sort_index()
```

```
0     26162
1    183431
Name: label, dtype: int64
```

```
plt.figure(figsize=(18,18))
sns.heatmap(df2.corr(),annot=True,cmap='plasma')
```

```
#sns.heatmap(df1,annot=True, fmt='f')
#sns.heatmap(df1,annot=True)
sns.heatmap(df1)
```

```
<AxesSubplot:>
```



- Interpretation of the Results

  Give a summary of what results were interpreted from the visualizations, pre-processing and modelling.

  After doing this project, We interpret that data set didn't have missing values but lot of features with observations. Lot of feature columns have outliers to whom I treated very well and also did scaling of the training features. Then I did feature engineering to do dimensionality reduction via feature importance method from decision tree and drop some of the useless features. Here we can do ridge or lasso or PCA as well for the same purpose. I also did

upsampling to balance the imbalanced classes for unbiased prediction of the model.

Than we fit the data into different type of models and did evaluation with the help of confusion matrix and F1 and Precision and recall method. I did focus on type 1 error to minimize to reduce loan defaulters on the same time I tried to maintain balance between non defaulter get loan as well.

In the last I got the best accuracy which I wanted to achieve i.e. 98% which indicates this is our good model and do great prediction and can do deploy in production environment for regular use.

CONCLUSION

To conclude all, I just wanted to let you know that its all depend up the business problem and then understand the data what we have or what will gather or collect related to that for accurate solution or prediction modelling. As we listened that what data comes in modelling as a input, output comes out as per input. So, for best predictive modelling we have to make our data clean, without any bias and noisy and at the same time balanced sometimes as well. In this modelling I did lot of data pre-processing

and up sampling and scaling before data fit to model. I tried lot of models and then finally got the best model as per our business problem. But due to lack of cloud infrastructure and other hardware and some of the software which wouldn't install in my system

I can't use whole things like hyperparameter tuning and other algorithms, but in future we can use cloud computing for deployment of the model in real time and also do tuning of the model. We can also use tensor board for the visualization of the modelling how it works with all parameter and hyperparameters and we can tweak and play with all to make our model more robust and do real-time prediction with 99.9% accuracy.

Thanks