MICRO-CREDIT DEFAULTER LIST

Submitted By :

ASHUTOSH CHAUDHARY

ACKNOWLEDGMENT

I would like to express my gratitude towards my internship mentor Ms. Srishti Maan for helping me in completion of the project.

BUSINESS PROBLEM

From my understanding the problem is about making prediction that whether a person would be a loan defaulter or not on the basis of other variables given so that we could minimize the losses without taking risk.

OBJECTIVE FOR PROBLEM UNDERTAKEN

We have to study every feature's behaviour present in the dataset & make observation from its behaviour about how every feature is giving the signs about that whether person will return loan within time limit or not & building predictive model on the basis of those features information to reduce the loan defaulters.

ANALYTICAL PROBLEM FRAMING

- **Origin of dataset & data types of every features**

Dataset is provided by the company & we have to import it using various libraries necessary for the project to get completed. Also data types of most features are continuous except three features which contains categorical data.

- **Mathematical/Analytical modelling of the problem**

For visualization we only use three plots most of the times that were countplot, distplot & scatterplot & For model building we use Logistic Regression, Decision Tree Classification, Random Forest Classification & Bagging Classification models to opt best out of them to work on dataset.

- **Assumptions related to problem statement**

No assumptions were made while working on the dataset.

- **Libraries & Tools used**

We used numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickle & warnings libraries for this task

STEPS TAKEN FOR THE TASK

1. **Importing Libraries for the task**

   Numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickles & warnings were imported for task to get completed.

2. **Importing Dataset using libraries**

   Imported the dataset using pandas library in jupyter notebook

3. **Checking Dimension of dataset**

   By checking dimensions of dataset we get to know that it contains 209593 rows & 37 columns

4. **Checking Description of dataset**

   From description we find the mean, min value, max value, etc of every column which contains continuous data in them & we get to know that only 6 columns have categorical data in them

5. Checking for presence of null values in dataset

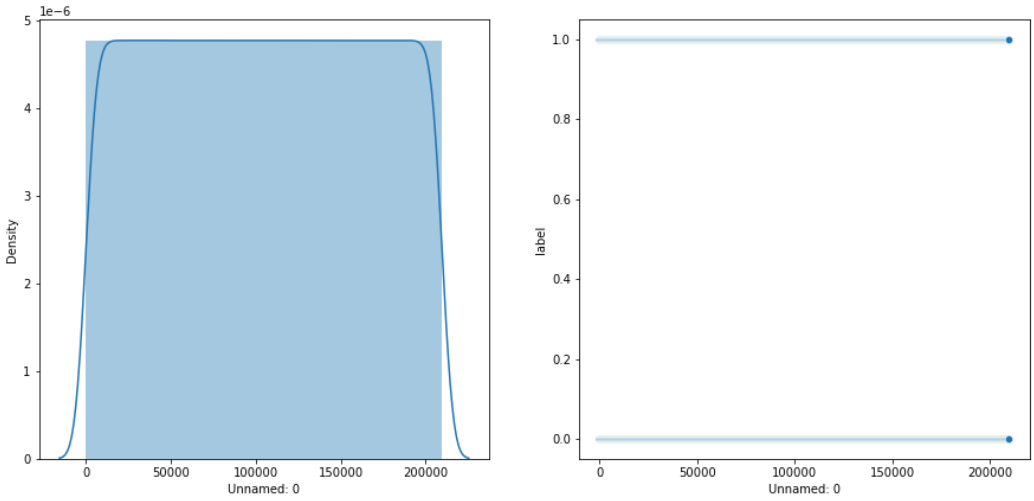
We checked for the presence of null values in every column of dataset by doing it repeatedly for every column as doing it for whole dataset does not show every columns null value value_counts & we concluded that there are no null values present in the dataset.

6. Identifying Target variable

By looking at dataset we identified target variable which is named label

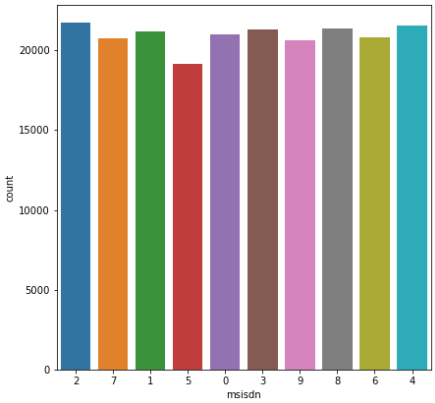
7. Performing EDA on whole dataset

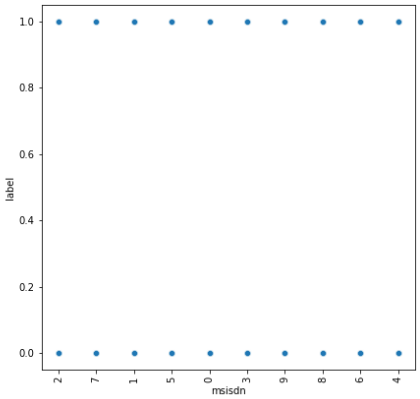
Unnamed: 0



By looking at unique value it was giving numbers like it is an index & from visualization we could not conclude anything

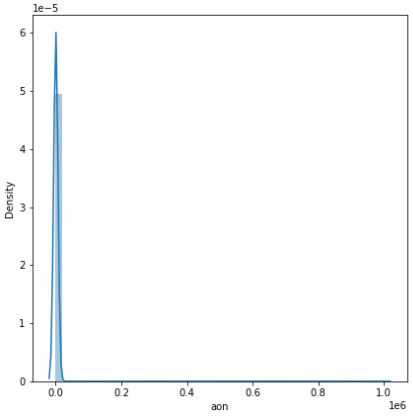
Msisdn (mobile number of user)

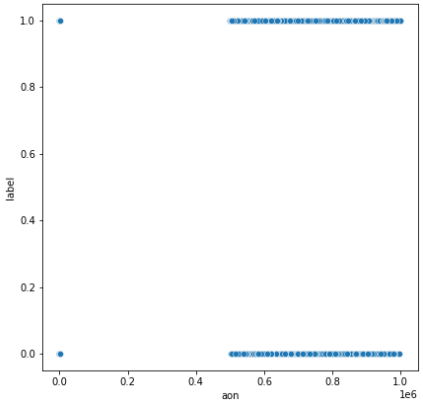




From visualization we conclude that most users use mobile number starting with digit 2 & every category of mobile number user is present in both categories of label.

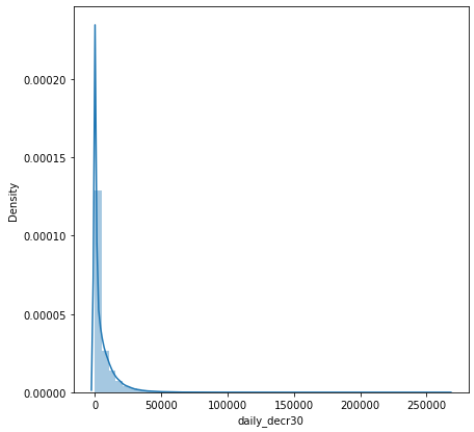
aon (age on cellular network in days)

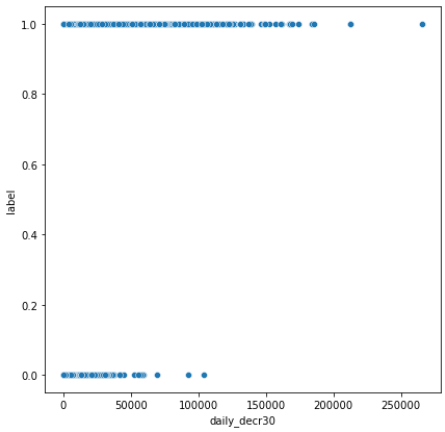




From visualization we conclude that we couldnot determine skewness of data with having highest density around 1 or 2 & data of aon is present in both categories of label over same range of aon's data.

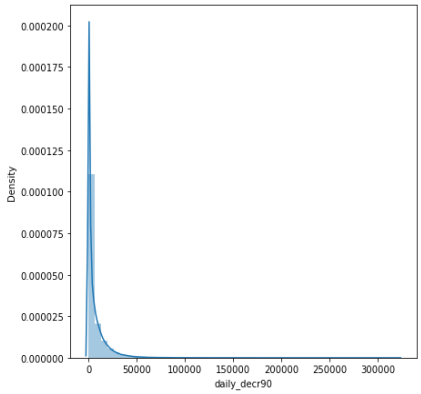
daily_decr30 (daily amount spent from main account avg over last 30 days)

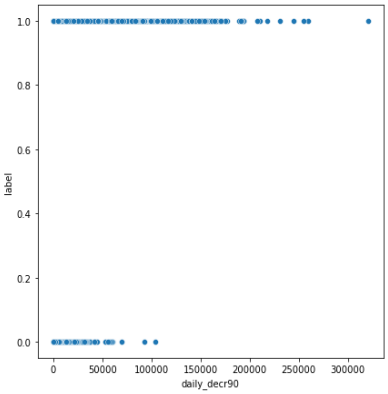




From visualization we concluded that data is positively skewed with having highest density at around 50 & users in category 0 of label are present till 10000's range where as users in category 1 of label are present well over 10000's range of daily_decr30

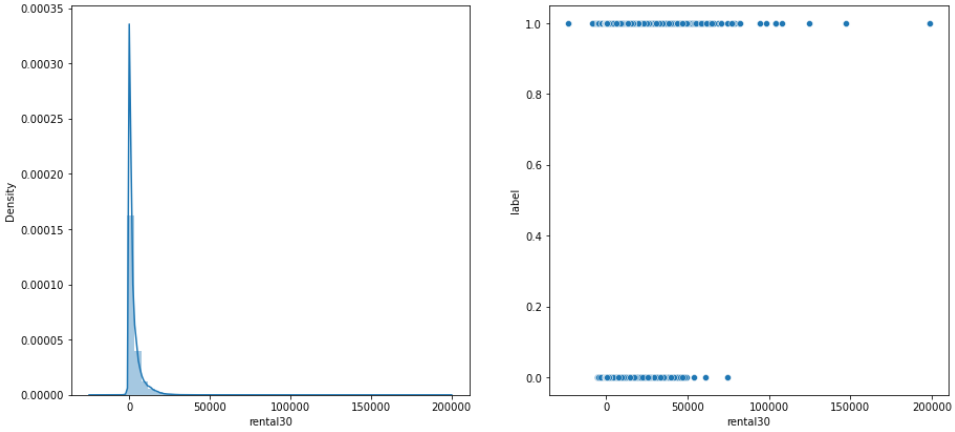
daily_decr90 (daily amount spent from main account avg over last 90 days)





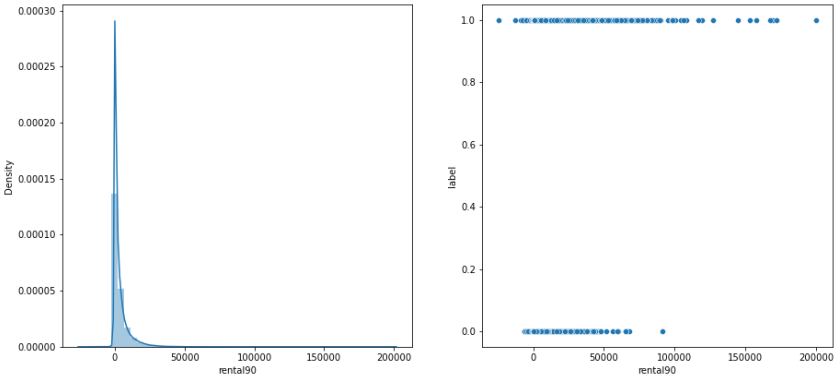
From visualization we concluded that data is positively skewed with having highest density at around 50 & users in category 0 of label are present till 10000's range where as users in category 1 of label are present well over 10000's range of daily_decr90

rental30 (avg main account balance over last 30 days)



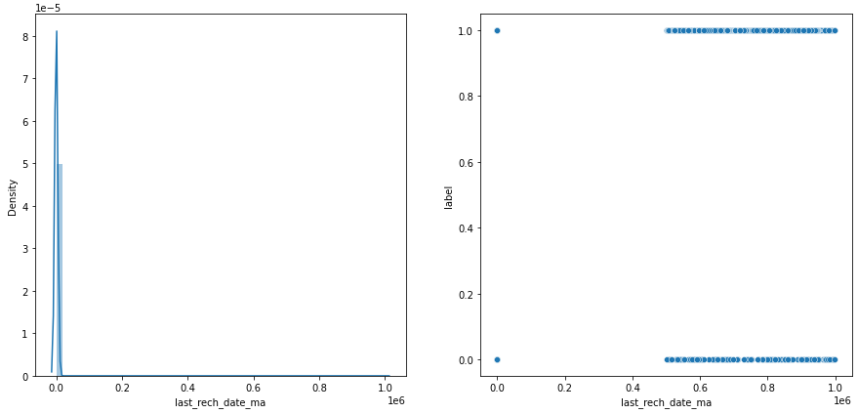
From visualization we concluded that data is positively skewed with having highest density at around 50 & users in category 0 of label are present till 10000's range where as users in category 1 of label are present well over 10000's range of rental30

rental90 (avg main account balance over last 90 days)



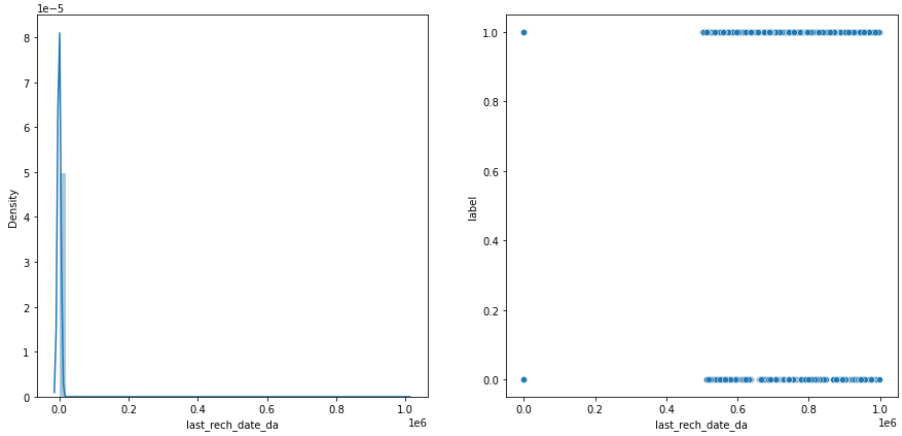
From visualization we concluded that data is positively skewed with having highest density at around 50 & users in category 0 of label are present till 10000's range where as users in category 1 of label are present well over 10000's range of rental90

last_rech_date_ma (number of days till last recharge of main account)



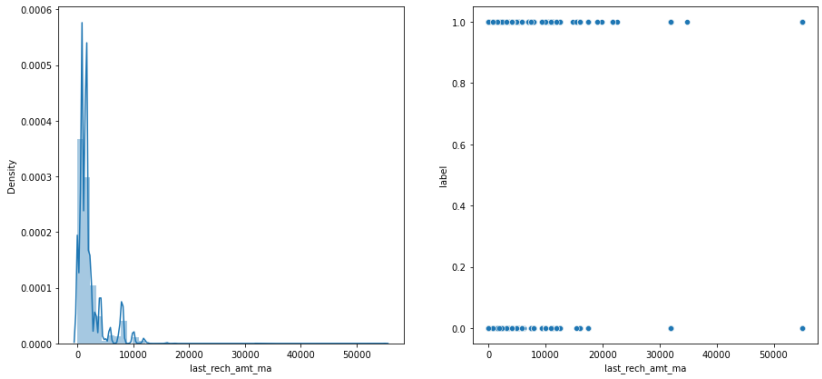
From visualization we concluded that we could not determine skewness of data with data having highest density at around 1 or 2 & data is present in both categories of label over same range of last_rech_date_ma's data.

last_rech_date_da (number of days till last recharge of data account)



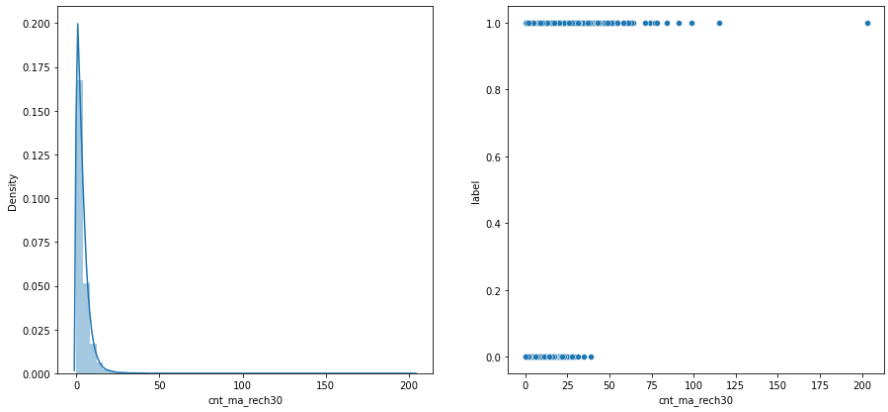
From visualization we concluded that we could not determine skewness of data with data having highest density at around 1 or 2 & data is present in both categories of label over same range of last_rech_date_da's data.

last_rech_amt_ma (amount of last recharge of main account)



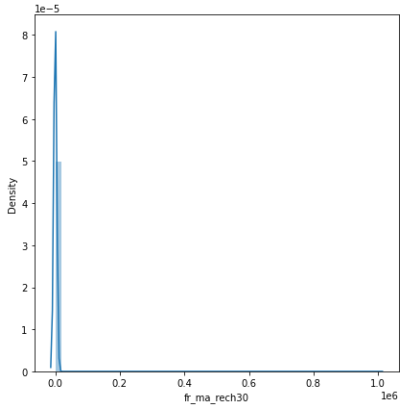
From visualization we concluded that data is positively skewed with data having highest density at around 30 & users in category 0 of label are present till 20000's range with being 1 or 2 over it where as users in category 1 of label are present well over 20000's range of last_rech_amt_ma.

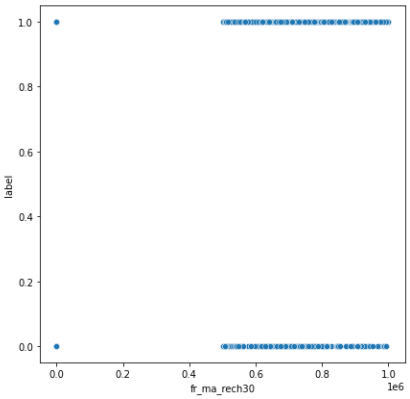
cnt_ma_rech30 (number of times main account got recharged in past 30 days)



From visualization we concluded that data is positively skewed with having highest density at around 10 & users in category 0 of label are present till 50's range where as users in category 1 of label are present well over 50's range of cnt_ma_rech30

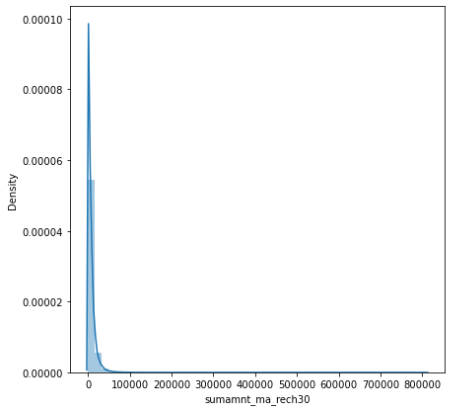
fr_ma_rech30 (Frequency of main account recharged in last 30 days)

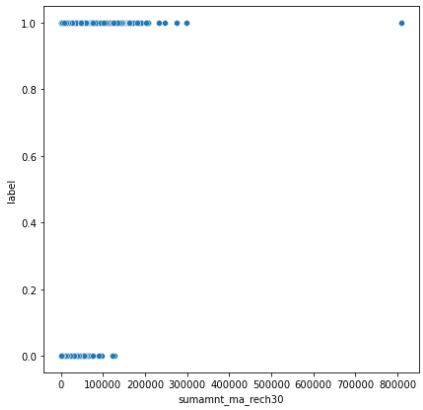




From visualization we concluded that we couldnot determine skewness of data with data having highest density at around 0 & data is present in both categories of label over same range of fr_ma_rech30's data.

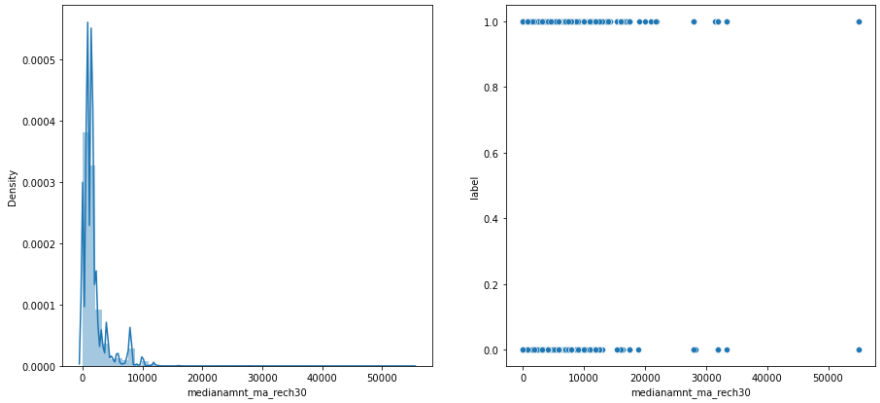
sumamnt_ma_rech30 (Total amount of recharge in main account over last 30 days)





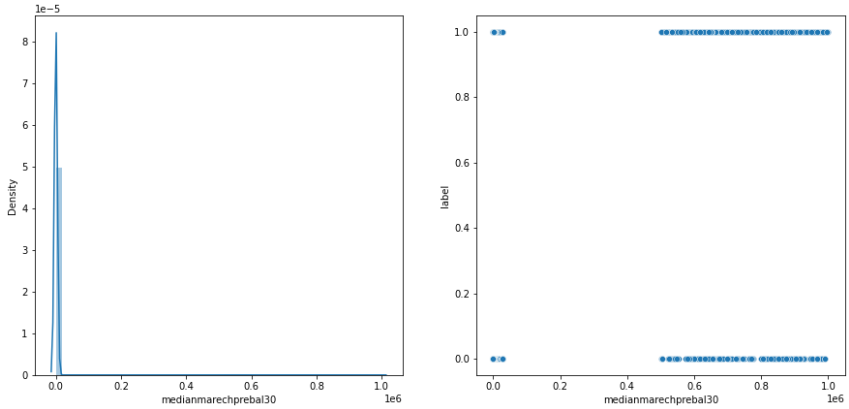
From visualization we concluded that data is positively skewed with having highest density at around 2000 & users in category 0 of label are present till 100000's range where as users in category 1 of label are present well over 100000's range of sumamnt_ma_rech30.

medianamnt_ma_rech30 (Median of amount of recharges done in main account over last 30 days at user level)



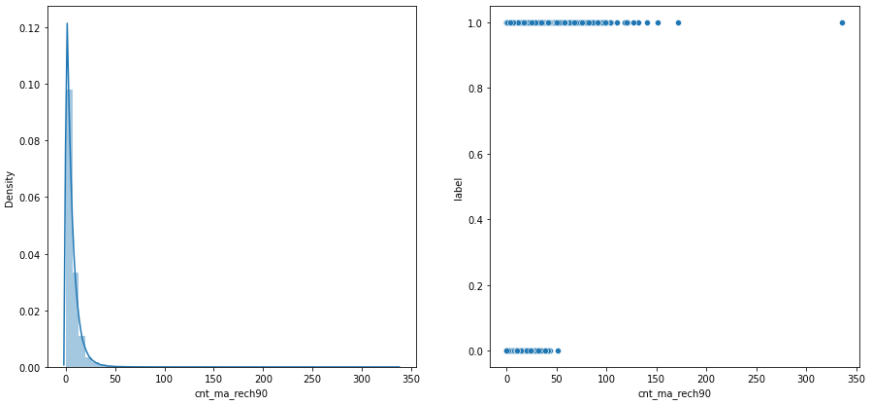
From visualization we concluded that data is positively skewed with having highest density at around 1000 & users of medianamnt_ma_rech30 in category 1 of label are present more then category 0 of label.

medianmarechprebal30 (Median of main account balance just before recharge in last 30 days at user level)



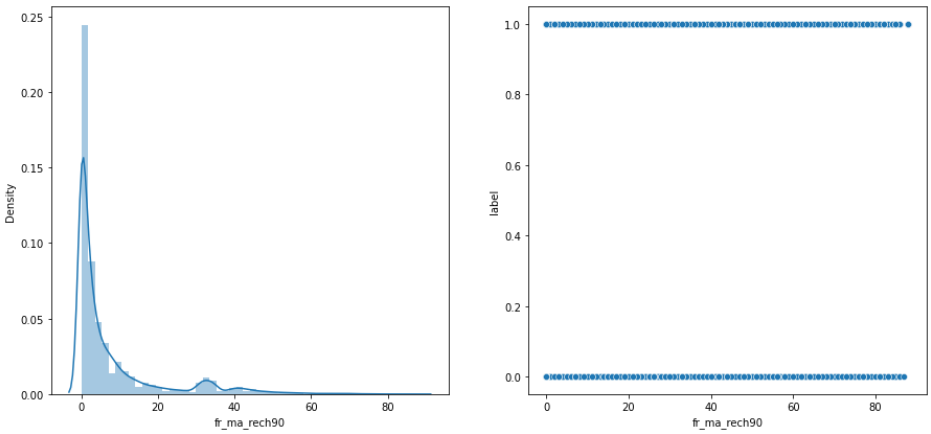
From visualization we concluded that we couldnot determine skewness of data with data having highest density at around 0 & data is present in both categories of label over same range of medianmarechprebal30

cnt_ma_rech90 (Number of times main account got recharged in last 90 days)



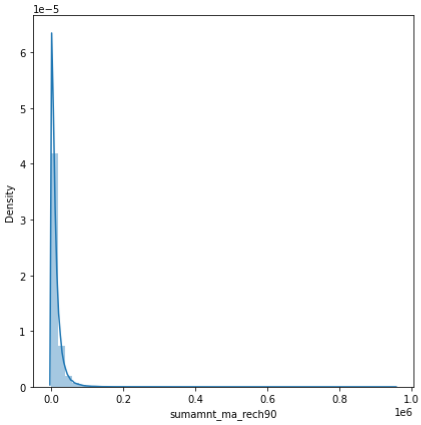
From visualization we concluded that data is positively skewed with having highest density at around 10 & users in category 0 of label are present till 50's range where as users in category 1 of label are present well over 50's range of cnt_ma_rech90.

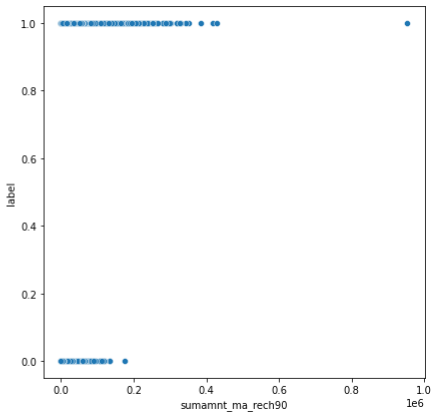
fr_ma_rech90 (Frequency of main account recharged in last 90 days



From visualization we conclude that data is positively skewed with having highest density at around 5 & data is present in both categories of label over same range of fr_ma_rech90.

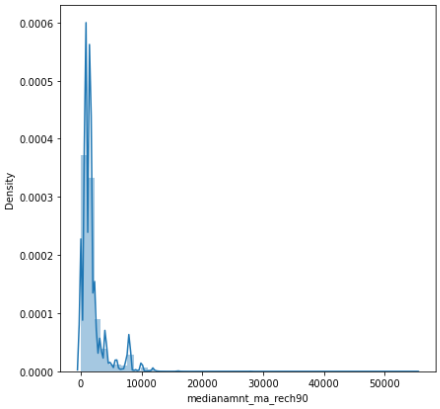
sumamnt_ma_rech90 (Total amount of recharge in main account over last 90 days)

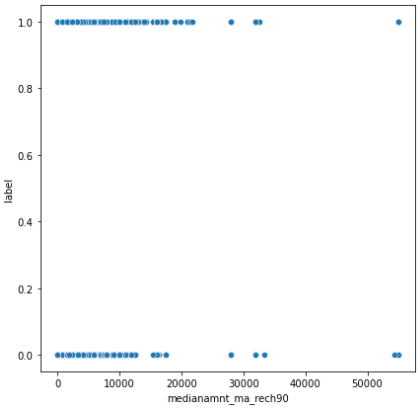




From visualization we concluded that data is positively skewed with having highest density at around 0.03 & users in category 0 of label are present till 0.2's range where as users in category 1 of label are present well over 0.2's range of sumamnt_ma_rech90.

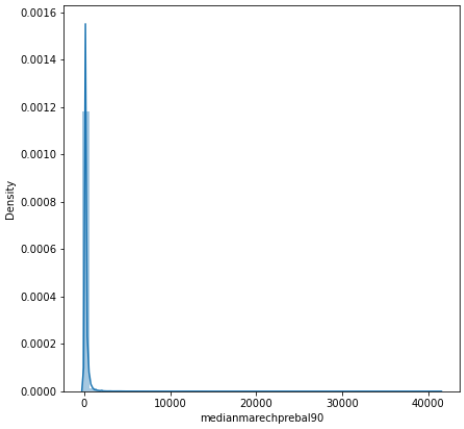
medianamnt_ma_rech90 (Median of amount of recharges done in main account over last 90 days at user level)

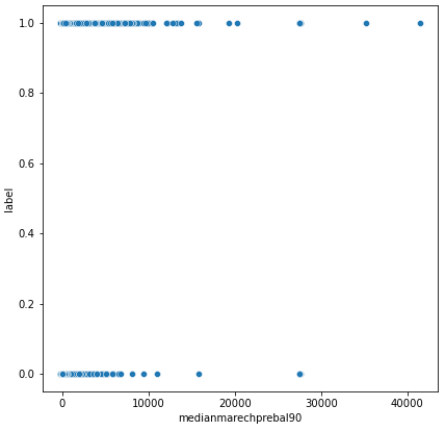




From visualization we concluded that data is positively skewed with having highest density at around 2000 & users of medianamnt_ma_rech90 are present more in category 1 then in category 0 of label.

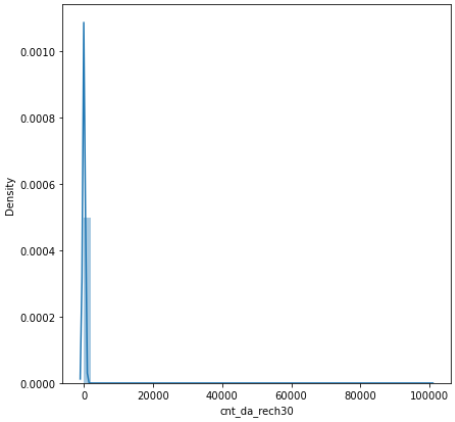
medianmarechprebal90 (Median of main account balance just before recharge in last 90 days at user level)

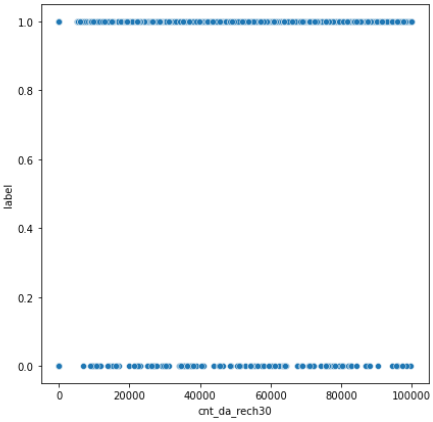




From visualization we concluded that we could not determine skewness of data with data having highest density at around 100 & users of medianmarechprebal90 are present more in category 1 then category 0 of label.

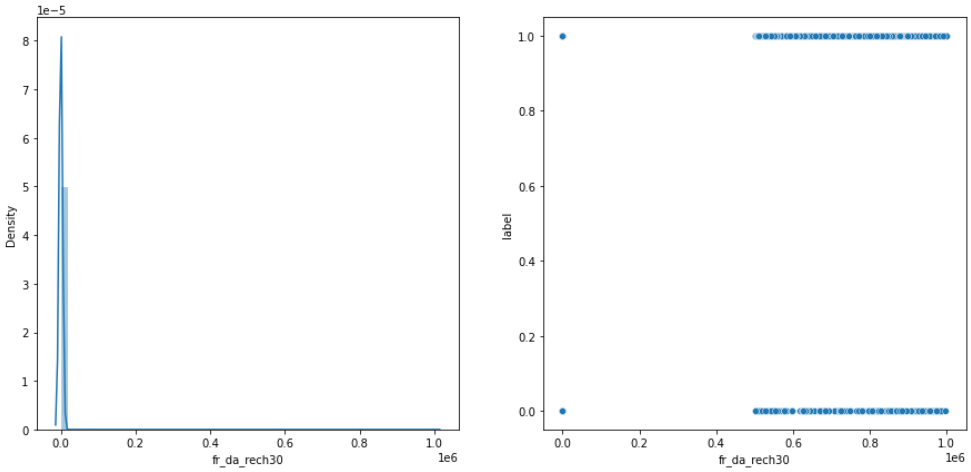
cnt_da_rech30 (Number of times data account got recharged in last 30 days)





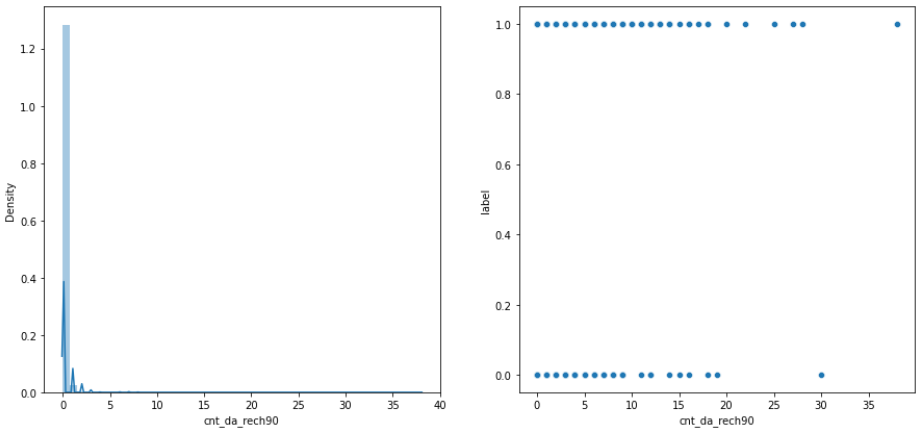
From visualization we conclude that we could not determine skewness of data with having highest density at around 300 & data is present in both categories of label over same range of cnt_da_rech30.

fr_da_rech30 (Frequency of data account recharged in last 30 days)



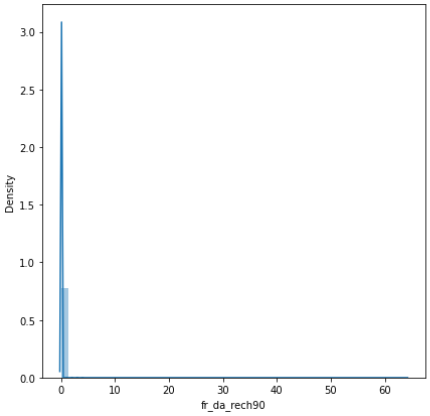
From visualization we conclude that we could not determine skewness of data with having highest density at around 300 & data is present in both categories of label over same range of fr_da_rech30.

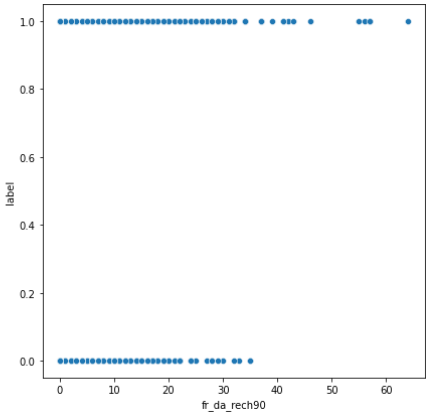
cnt_da_rech90 (Number of times data account got recharged in last 90 days)



From visualization we concluded that data is positively skewed with having highest density at around 2 & cnt_da_rech90 users are present more in category 1 then category 2 of label.

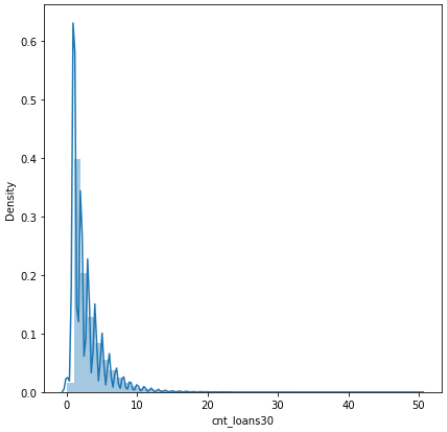
fr_da_rech90 (Frequency of data account recharged in last 90 days)

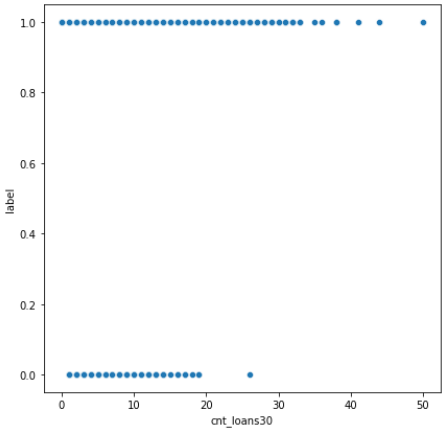




From visualization we concluded that data is positively skewed with having highest density at around 2 & users in category 0 of label are present till 40's range where as users in category 1 of label are present well over 40's range of fr_da_rech90.

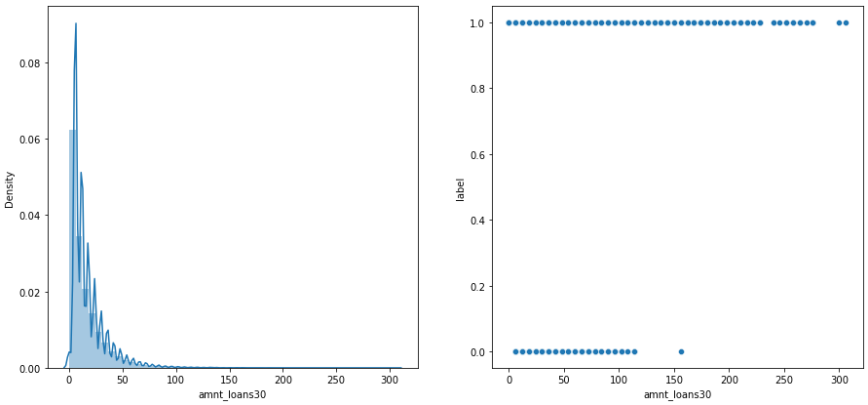
cnt_loans30 (Number of loans taken by user in last 30 days)





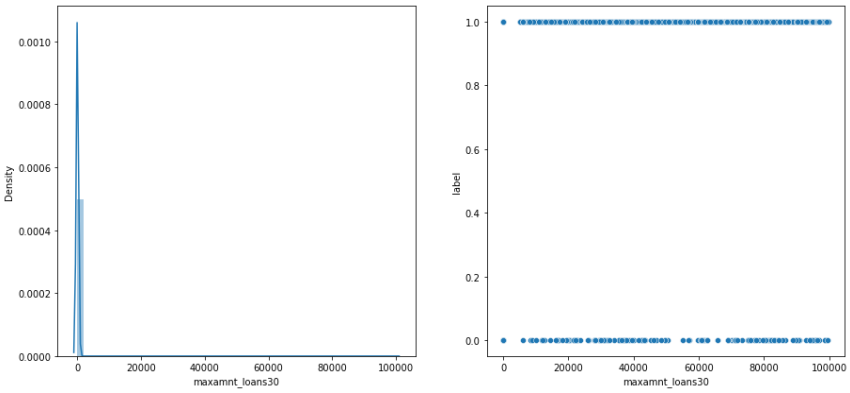
From visualization we concluded that data is positively skewed with having highest density at around 3 & users in category 0 of label are present till 30's range where as users in category 1 of label are present well over 30's range of cnt_loans30.

amnt_loans30 (Total amount of loans taken by user in last 30 days)



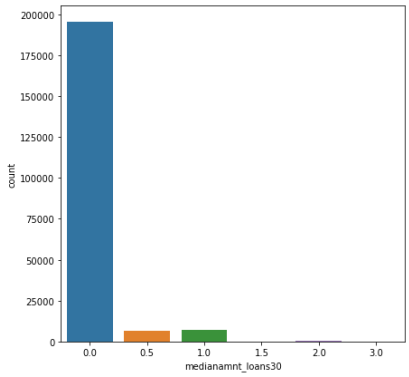
From visualization we concluded that data is positively skewed with having highest density at around 5 & users in category 0 of label are present till 160's range where as users in category 1 of label are present well over 160's range of amnt_loans30.

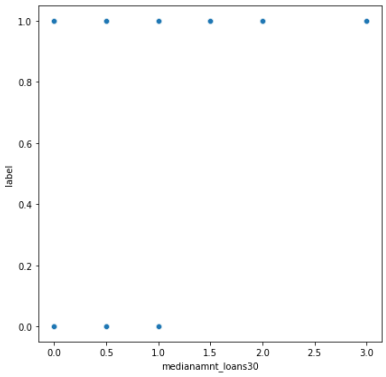
maxamnt_loans30 (maximum amount of loan taken by the user in last 30 days)



From visualization we could not determine skewness of data with having highest density at around 100 & data is present in both categories of label over same range of maxamnt_loan30 with data present a bit more in category 1 then category 0.

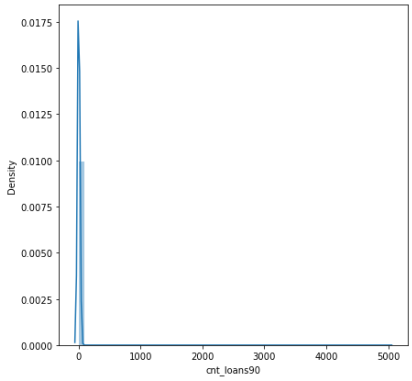
medianamnt_loans30 (Median of amounts of loan taken by the user in last 30 days)

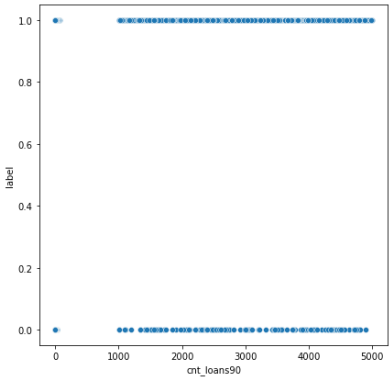




From visualization we conclude that most users are present in 0.0 category of medianamnt_loans30 & only 0, 0.5, 1 categories of medianamnt_loans30 are present in both categories of label, else every other category of medianamnt_loans30 is present in only category 1 of label except for category 2 of medianamnt_loans30 which is not present in any category of label.

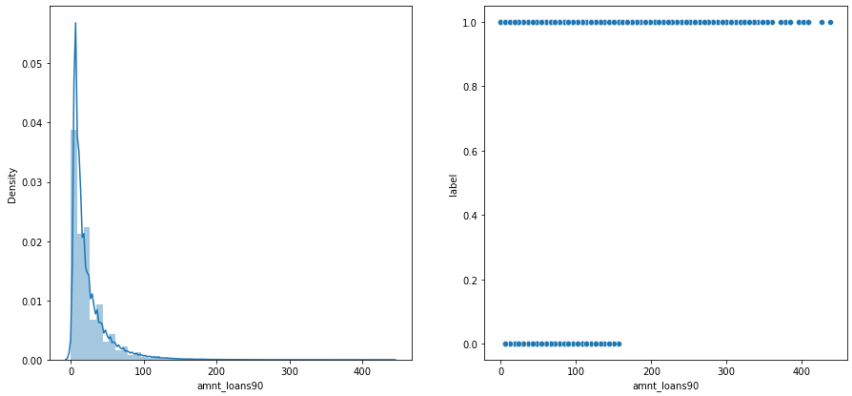
cnt_loans90 (Number of loans taken by user in last 90 days)





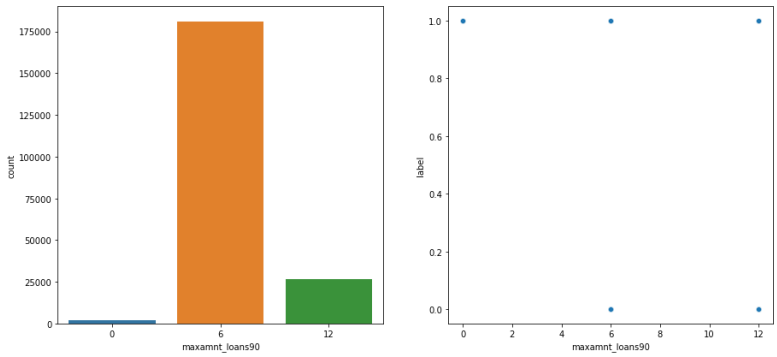
From visualization we conclude that we could not determine data skewness with data having highest density at around 100 & data is present in both categories of label over same range of cnt_loans90.

amnt_loans90 (Total amount of loans taken by user in last 90 days)



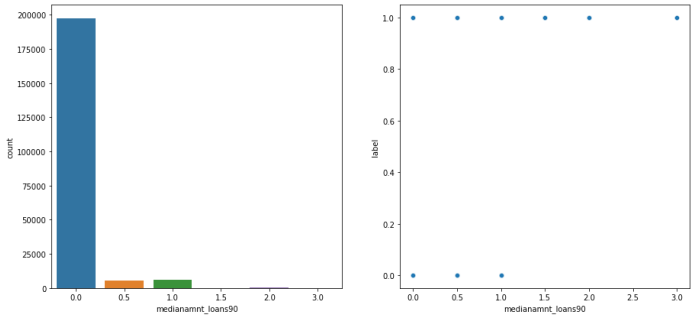
From visualization we conclude that data is positively skewed with having highest density at around 15 & users in category 0 of label are present till 180's range where as users in category 1 of label are present well over 180's range of amnt_loans90.

maxamnt_loans90 (maximum amount of loan taken by the user in last 90 days)



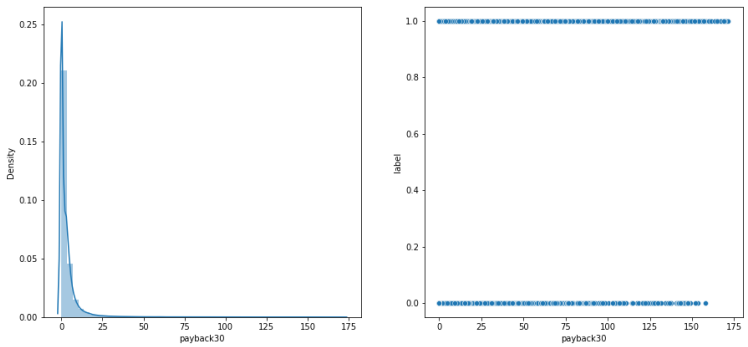
From visualization we concluded that most users are present in category 6 of maxamnt_loans90 & every category of maxamnt_loans90 is present in both categories of label except for only category 0 of maxamnt_loans90 which is present in only category 1 of label.

medianamnt_loans90 (Median of amounts of loan taken by the user in last 90 days)



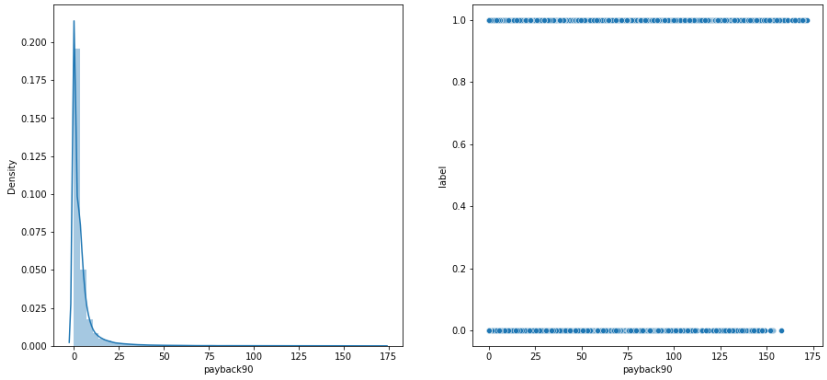
From visualization we conclude that most users are present in category 0 of medianamnt_loans90 & every category of medianamnt_loans90 is present in both categories of label till category 1 of medianamnt_loans90 & beyond that category every category of medianamnt_loans90 is present in only category 1 of label except for category 2.5 which is not present in any category of label.

payback30 (Average payback time in days over last 30 days)



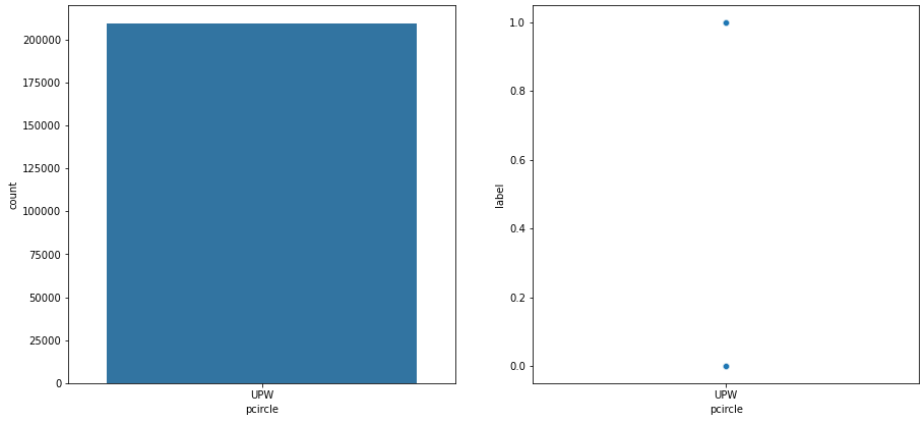
From visualization we conclude that data is positively skewed with having highest density at around 4 & data is present in both categories of label over same range of payback30 with data being present little bit more in category 1 then category 0 of label.

payback90 (Average payback time in days over last 90 days)



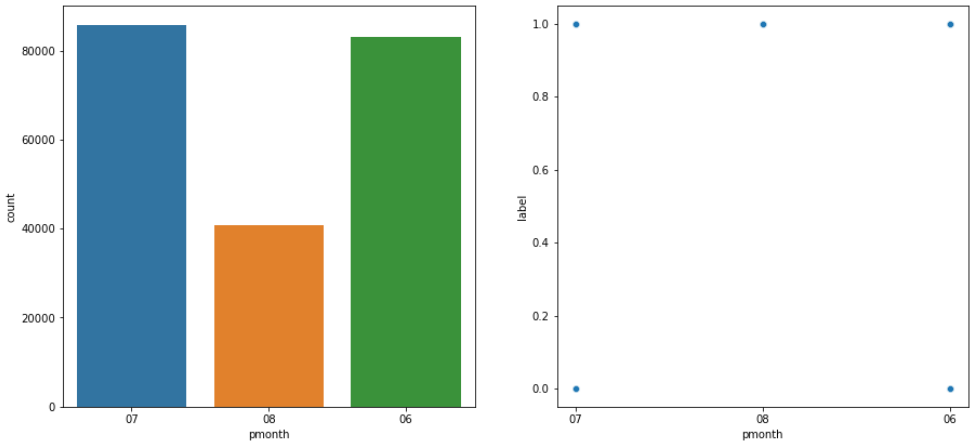
From visualization we conclude that data is positively skewed with having highest density at around 4 & data is present in both categories of label over same range of payback30 with data being present little bit more in category 1 then category 0 of label.

pcircle (telecom circle)



From visualization we could not conclude Anything as there is only one same value present in every row of the whole & it is present in both categories of label.

pdate (date)



Refined data to extract month information from it as the data given was of a single year.

Then from visualization we concluded most users are present in category 7 of pmonth & every category of pmonth is present in both categories of label except for category 8 which is present only in category 1 of label.

Changed data type

After visualization we checked data types of every column & then we changed data type of every column to int or float data type.

8. Plotting Heatmap

We plotted heatmap to check correlation of every column with target column & we concluded that following (pmonth, amnt_loans90, amnt_loans30, cnt_loans30, medianamnt_ma_rech90, sumamnt_ma_rech90, cnt_ma_rech90, medianamnt_ma_rech30, sumamnt_ma_rech30, cnt_ma_rech30, daily_rech90 & daily_rech30) columns have high positive correlation with target column & we also dropped following (pcircle, payback, medianamnt_loans90, cnt_da_rech90, fr_ma_rech90, rental30) columns to reduce multicollinearity as these column does not having high correlation with target column but were having high correlation with other columns.

9. Splitting & Scaling data

Then we splitted data into X & y in which y contains target variable & X contains other variables. Then we further splitted X & y into X_train, X_test, y_train & y_test sets using train_test_split & after that we scaled X_train & X_test set using StandardScalar to remove outliers if present any in the columns.

10. Applying GridSearchCV

We chose 4 classification models (Logistic Regression, Decision Tree Classification, Random Forest Classification, Bagging Classification) to check performance of them on the dataset to find out the best suited model for the dataset. Then one by one we applied GridSearchCV on each model to find best hyper parameter tuning while model is working on dataset. After applying GridSearchCV on every model we concluded that Decision Tree Classification model is giving the best Test accuracy score which was 0.9064 with hyper parameter combination (max_depth : 5, min_samples_leaf : 1, min_samples_split : 2)

11. Applying Model on dataset

Chose Decision Tree Classification Model & made it ready by tuning hyper parameter. After tuning hyper parameter we applied model on dataset to make prediction. After making prediction we made a dataframe which contained actual & predicted values side by side to compare them.

12. Applying metrices

After making prediction we checked performance of the model through various metrices :

- Accuracy score : 0.9064
- Cohen Kappa score : 0.4763
- Confusion matrix : [3236, 4672
  1212, 53758]

By looking at above all the 3 metrices score we concluded that model is performing quite well.

13. Saving Model

After checking performance of model we saved it using pickle library.

CONCLUSION

In the end we can conclude that most people are non defaulters by looking at plots of various columns & model that has been build is performing quite good on the dataset & would work efficiently on the future dataset for making prediction & reducing the losses of organization.