FLIP ROBO

Micro Credit Defaulter

Submitted by:

Prateek AP

INTRODUCTION

Business Problem:

MFI is a financial institution that gives financial aid to the economically low-class people. As an initiative of this cause, they are tying up with mobile financial services which is telecom provider which provides low-cost services to its customers as part of this cause. Partnership is mainly to facilitate low economy background people to get credit on the service availed from their client. So, in current situation, client has to predict who are potential customers who are credit defaulters.ie, fail to pay the principal amount within specified period of time.

Conceptual Background of the Domain Problem

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. In order to make sure this underserved population has a positive loan experience; company makes use of a variety of alternative data include transactional information--to predict their clients' repayment abilities.

Review of Literature

This is problem related to financial services & banking sector.Microfinance institutions play a major role in economic development in many developing countries. However, many of these microfinance institutions are faced with the problem of default because of the non-formal nature of the business and individuals they lend money to. This study seeks to find the determinants of credit default in microfinance institutions. With data on 2631 successful loan applicants from a microfinance institution with branches all over the country we proposed a Binary logistic regression model to predict the probability of default. We found the following variables significant in determining default: Age, Gender, Marital Status, Income Level, Residential Status, Number of Dependents, Loan Amount, and Tenure. We also found default to be more among the younger generation and in males. We however found Loan Purpose not to be significant in determining credit default. Microfinance institutions could use this model to screen prospective loan applicants in order to reduce the level of default.

Motivation for the Problem Undertaken

Since the financial services are providing micro credit to the telecom users through telecom partners, they expect the customers to return the offered credit with in certain span of time. If they failed to pay back the same within stipulated time as given as a deadline,they should be labeled as defaulters. So, the objective behind giving credit is to provide financial assistance to the financially degraded individuals. The main motivation behind this initiation is to help the poorer section of the society.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

On analysis of the dataset, we got an unbalanced dataset with more than 80% of the target variables as non-defaulters and remaining are defaulters. So before balancing the data we have to look into the dataset for any abnormalities in the data. We have most of the features having right or left skewness. So, we will first interpret the data for our Analysis by getting some statistical parameters like mean, median & quantiles which decides the factors that will help predict the target variable. So, we will apply the formula to identify & delete the outliers for our better model building.

Data Sources and their formats

Data set is of the Indonesian financial services which has multiple independent features like Age on cellular network, Daily amount spend, Average main account balance etc. The dataset file we are using is of the excel format.

Data Pre-processing Done

1. Loading Dataset
2. Shape of our dataset having 209593 rows and 37 columns
3. Looking at statistical parameters
4. No Null values in our dataset
5. Duplicate data entries within msisdn removed
6. Finding Unrealistic values in the dataset and imputing them with median.
7. Dataset is imbalanced.

Data Inputs- Logic- Output Relationships

The input data provided, helps to understand the behaviour of the customer, their various transaction records, their frequency of transaction during a period of time etc, all these helps to predict the customer's intension toward the repayment of loan.

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

The dataset had a lot of unrealistic values, eg. Aon had values of people who were customers over 100 years ago. The dataset was for 2016, we have dropped the values which were less than 0 days as a customer and the values who were customers from more than 25 years.

Hardware and Software Requirements and Tools Used

Minimum Hardware Requirements:

- 8Gb+ of RAM
- 128 GB or more (256 GB recommended)
- 10Gb+ of free hard drive space
- Working keyboard
- Trackpad/Mouse
- Display
- A power Adapter.

Software's and Tool's Used:

- Jupyter Notebook
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Plotly
- Scikit Learn

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

The data set contain more than 2 lakh data with no null values related to the customer. The dataset is imbalanced. Label 1 has 87.5% of data whereas label 0 has approximately 12.5%. As I went through the dataset, I found lot of outliers and skewness are present in the dataset. The outliers were corrected by replacing them with median to avoid data loss. The skewness was also reduced using power transformation. There were a lot of columns high correlation, to reduce this PCA was done. After data cleaning and data transformation, data visualization was done to represent data graphically. At last, the most important part was to build model for the data set.

Testing of Identified Approaches (Algorithms)

LogisticRegression :

The Accuracy Score is 76.43054182064228
Confusion Matrix:/
[[38580 9669]
[12973 34843]]

	precision	recall	f1-score	support
0	0.75	0.80	0.77	48249
1	0.78	0.73	0.75	47816
accuracy			0.76	96065
macro avg	0.77	0.76	0.76	96065
weighted avg	0.77	0.76	0.76	96065

Cross Validation Score is 76.1258025832563

Difference between accuracy score and cv is 0.30473923738597364

KNeighborsClassifier :

Report for model KNeighborsClassifier()
The Accuracy Score is 86.3134336126581
Confusion Matrix:
[[46335 1914]
[11234 36582]]

	precision	recall	f1-score	support
0	0.80	0.96	0.88	48249
1	0.95	0.77	0.85	47816
accuracy			0.86	96065
macro avg	0.88	0.86	0.86	96065
weighted avg	0.88	0.86	0.86	96065

Cross Validation Score is 84.82399380418218

Difference between accuracy score and cv is 1.4894398084759217

DecisionTreeClassifier :

Report for model DecisionTreeClassifier()
The Accuracy Score is 84.22005933482538
Confusion Matrix:
[[41649 6600]
[8559 39257]]

	precision	recall	f1-score	support
0	0.83	0.86	0.85	48249
1	0.86	0.82	0.84	47816
accuracy			0.84	96065
macro avg	0.84	0.84	0.84	96065
weighted avg	0.84	0.84	0.84	96065

Cross Validation Score is 82.869063382217

Difference between accuracy score and cv is 1.350995952608372

RandomForestClassifier :

Report for model RandomForestClassifier()
The Accuracy Score is 91.18409410295114

Confusion Matrix:
[[44732 3517]
[4952 42864]]

	precision	recall	f1-score	support
0	0.90	0.93	0.91	48249
1	0.92	0.90	0.91	47816
accuracy			0.91	96065
macro avg	0.91	0.91	0.91	96065
weighted avg	0.91	0.91	0.91	96065

Cross Validation Score is 89.77565143528119

Difference between accuracy score and cv is 1.408442667669945

ExtraTreesClassifier :

Report for model ExtraTreesClassifier()
The Accuracy Score is 92.23130172279186
Confusion Matrix:
[[45259 2990]
[4473 43343]]

	precision	recall	f1-score	support
0	0.91	0.94	0.92	48249
1	0.94	0.91	0.92	47816
accuracy			0.92	96065
macro avg	0.92	0.92	0.92	96065
weighted avg	0.92	0.92	0.92	96065

Cross Validation Score is 91.15690658805306

Difference between accuracy score and cv is 1.0743951347387934

AdaBoostClassifier :

Report for model AdaBoostClassifier()
The Accuracy Score is 76.70535574871181
Confusion Matrix:
[[38585 9664]
[12714 35102]]

	precision	recall	f1-score	support
0	0.75	0.80	0.78	48249
1	0.78	0.73	0.76	47816
accuracy			0.77	96065
macro avg	0.77	0.77	0.77	96065
weighted avg	0.77	0.77	0.77	96065

Cross Validation Score is 76.28507007769755

Difference between accuracy score and cv is 0.4202856710142555

GradientBoostingClassifier :

Report for model GradientBoostingClassifier()
The Accuracy Score is 78.12418674855567
Confusion Matrix:
[[38716 9533]
[11482 36334]]

	precision	recall	f1-score	support
0	0.77	0.80	0.79	48249
1	0.79	0.76	0.78	47816
accuracy			0.78	96065
macro avg	0.78	0.78	0.78	96065
weighted avg	0.78	0.78	0.78	96065

Cross Validation Score is 77.85276188572713

Difference between accuracy score and cv is 0.27142486282853895

CatBoostClassifier :

Report for model <catboost.core.CatBoostClassifier object at 0x0000026F87BA29D0>
The Accuracy Score is 84.1773798990267
Confusion Matrix:
[[41596 6653]
[8547 39269]]

	precision	recall	f1-score	support

0	0.83	0.86	0.85	48249
1	0.86	0.82	0.84	47816
accuracy			0.84	96065
macro avg	0.84	0.84	0.84	96065
weighted avg	0.84	0.84	0.84	96065

Cross Validation Score is 83.3437429734929

Difference between accuracy score and cv is 0.8336369255338099

Run and evaluate selected models

The algorithm used for hyper parameter tuning and fitting train & test dataset are

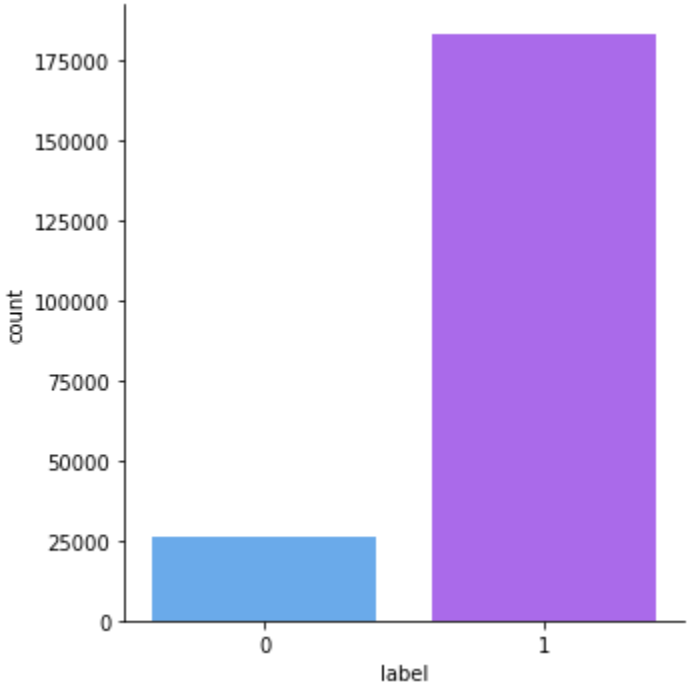
a. ExtraTreesClassifier
b. Logistic Regression
c. GradientBoostingClassifier

Key Metrics for success in solving problem under consideration

As mentioned earlier, the dataset is unbalanced with 87.5% of label 1 and 12.5% of label 0, which made it clear that, we cannot blindly rely on accuracy score for the prediction as it can lead to biasness. Hence, I have used confusion matrix and AUC ROC curve to determine the accuracy of the model and finalizing a model based on these parameters.

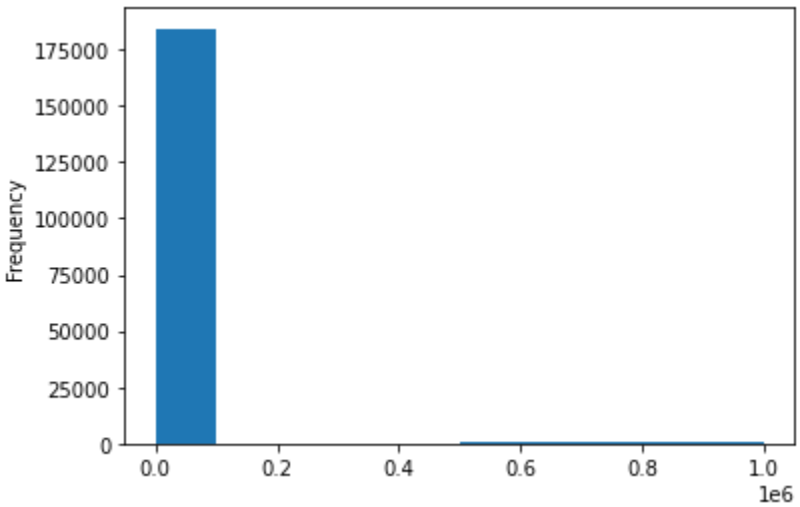
Visualizations

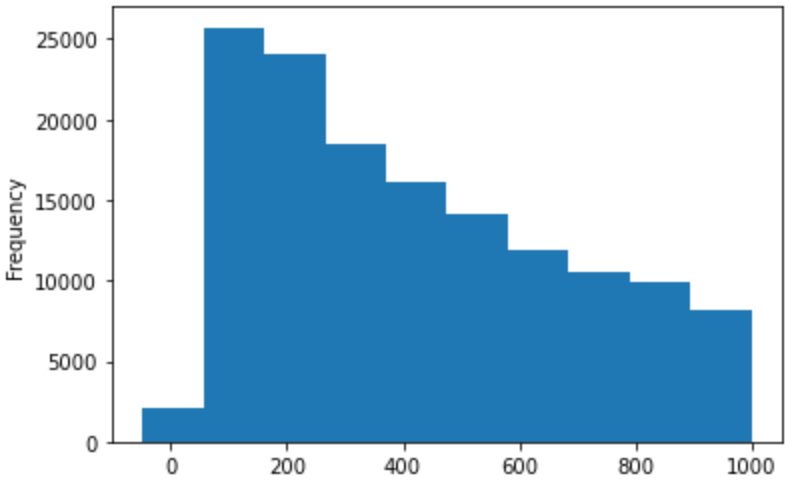
Label :



We see that the data is highly imbalanced and would have to be balanced.

Aon:

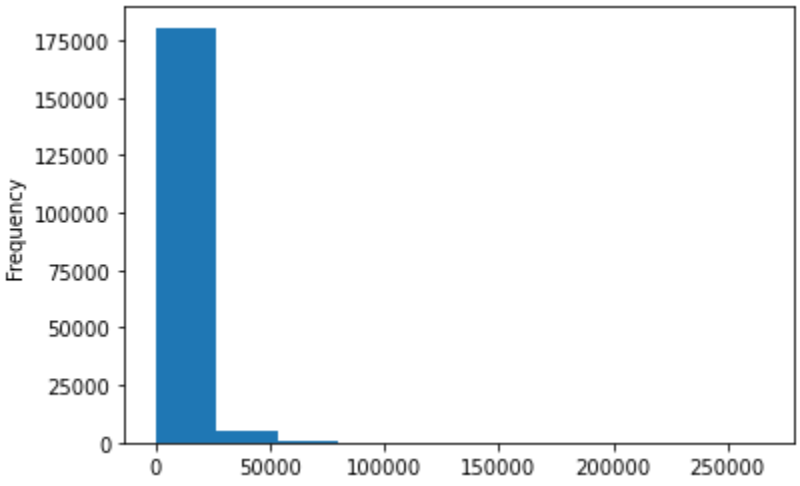


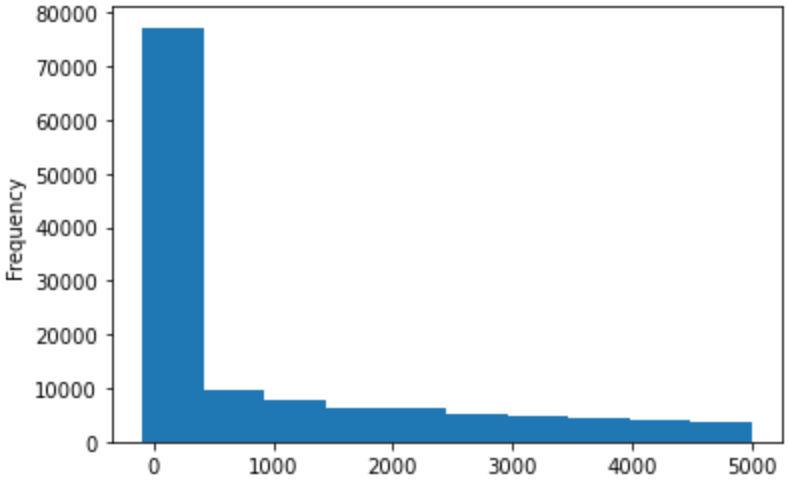


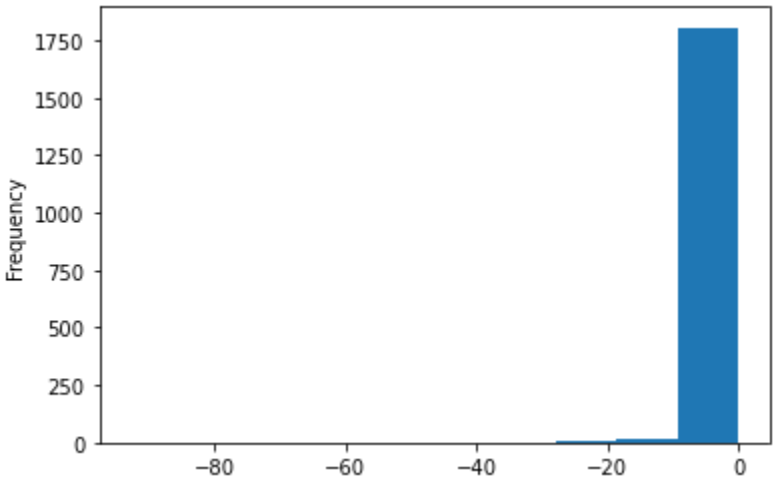
From above, we see the majority of the customers have been with telecom provider between 70-1000 days old. We also see that there are a few records for aon as negative, which cannot be possible.

We cannot have age on cellular network to be less than 0 as if a person is not a customer how would// why would a company loan him? Hence, we could drop all the values which are negative but that would lead to data loss hence we go ahead with median imputation of the data.

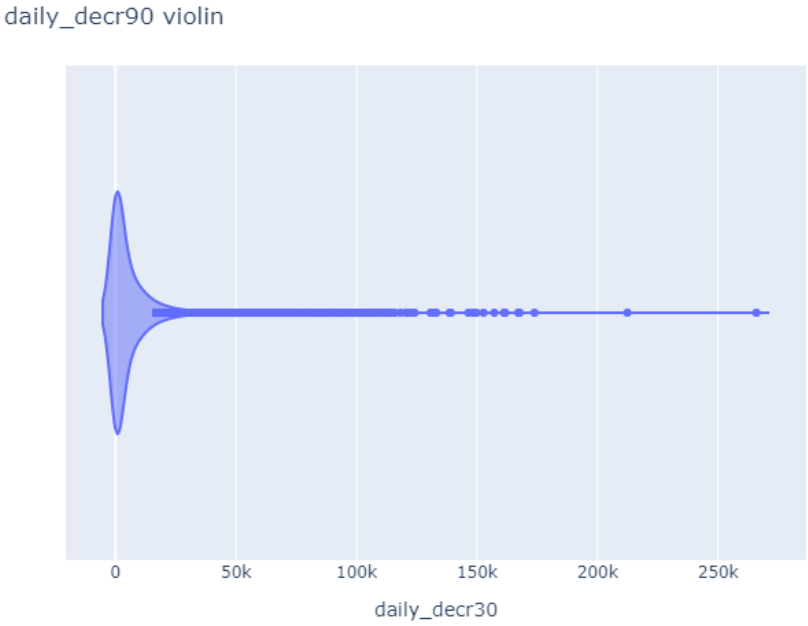
Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah): daily_decr30




Frequency
80000
70000
60000
50000
40000
30000
20000
10000
0
0
1000
2000
3000
4000
5000

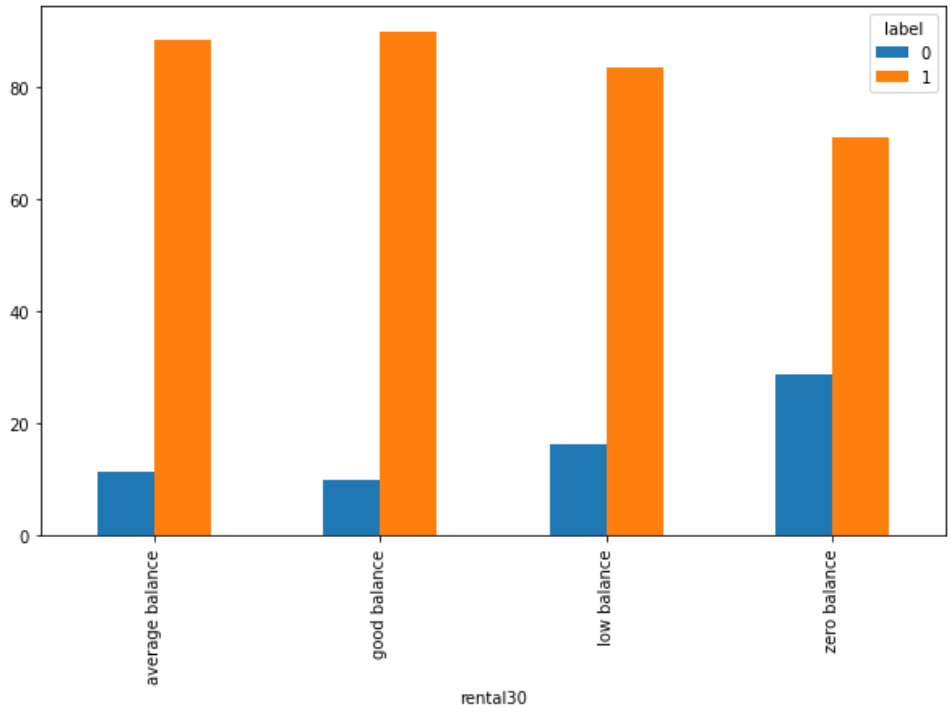

Frequency
1750
1500
1250
1000
750
500
250
0
−80
−60
−40
−20
0

We see that the Daily amount spent from main account cannot be a negative number hence let's drop all the negative values.



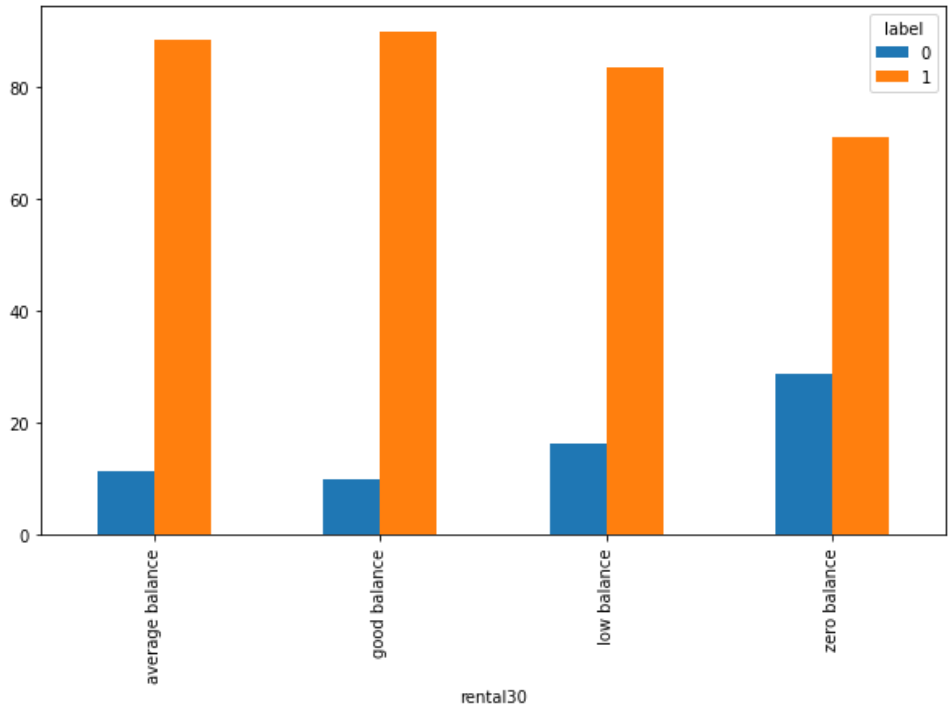
we see a large amount of outliers present.

Average main account balance over last 30 days: rental30



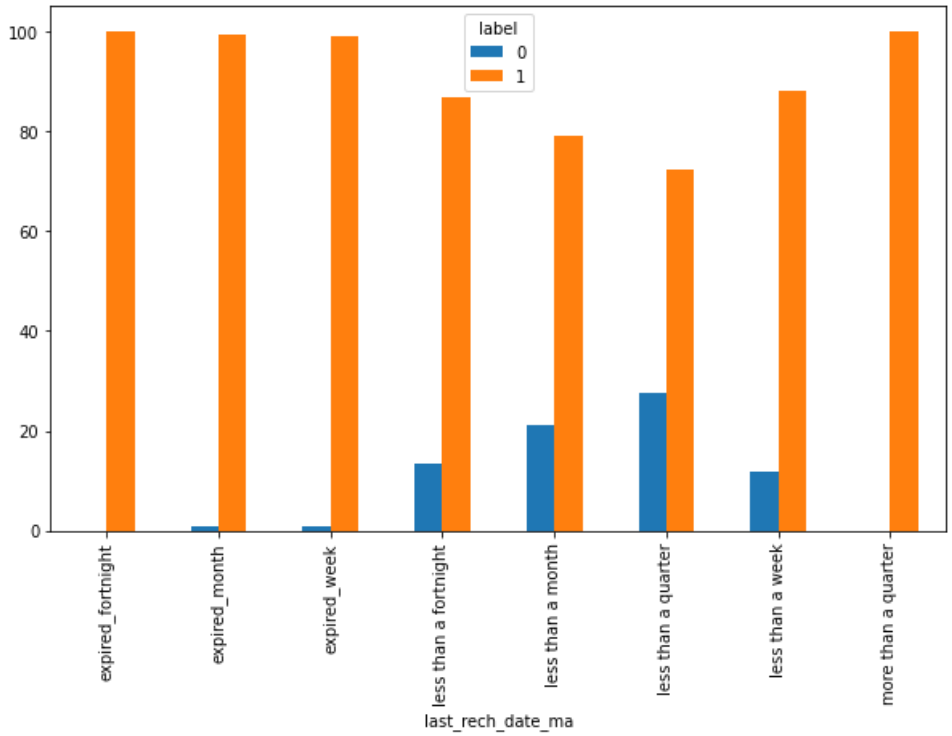
we see that the people who maintain good and average balance have a higher chance of paying back the credit amount, were as people with zero balance default the payments the most.

Average main account balance over last 90 days: rental90



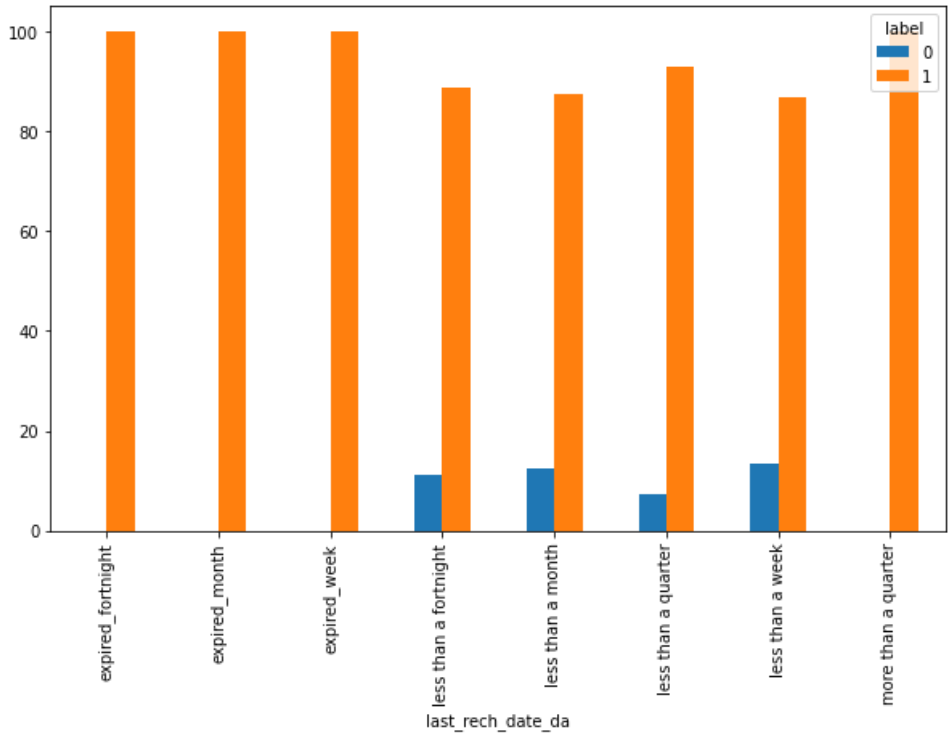
We see a similar trend where the people with zero or low balance, tend to default the payments the most and people with good balance the least.

Number of days till last recharge of main account: last_rech_date_ma



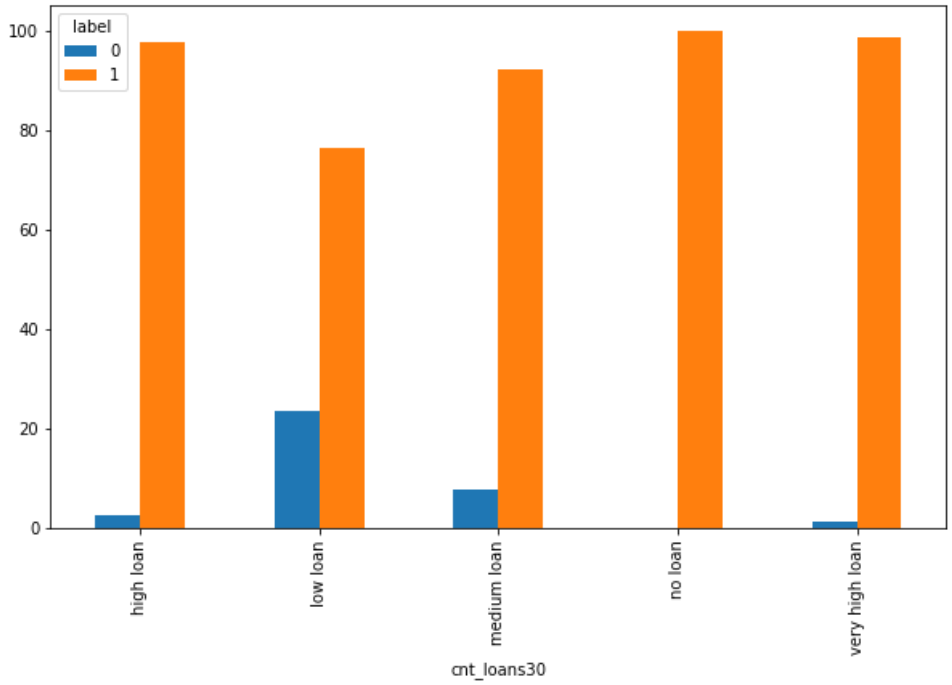
From above we see customers whose accounts are expiring between a week to a quarter are the once who default the payments the most.

Number of days till last recharge of data account: last_rech_date_da



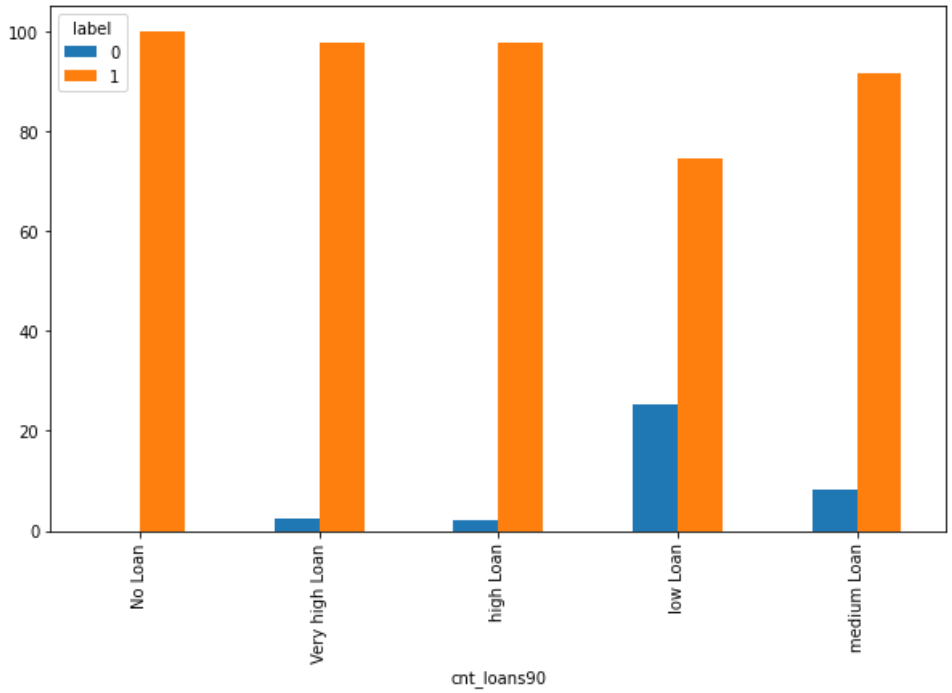
We see a similar trend as we saw in last_rech_date_ma.

Number of loans taken by user in last 30 days: cnt_loans30



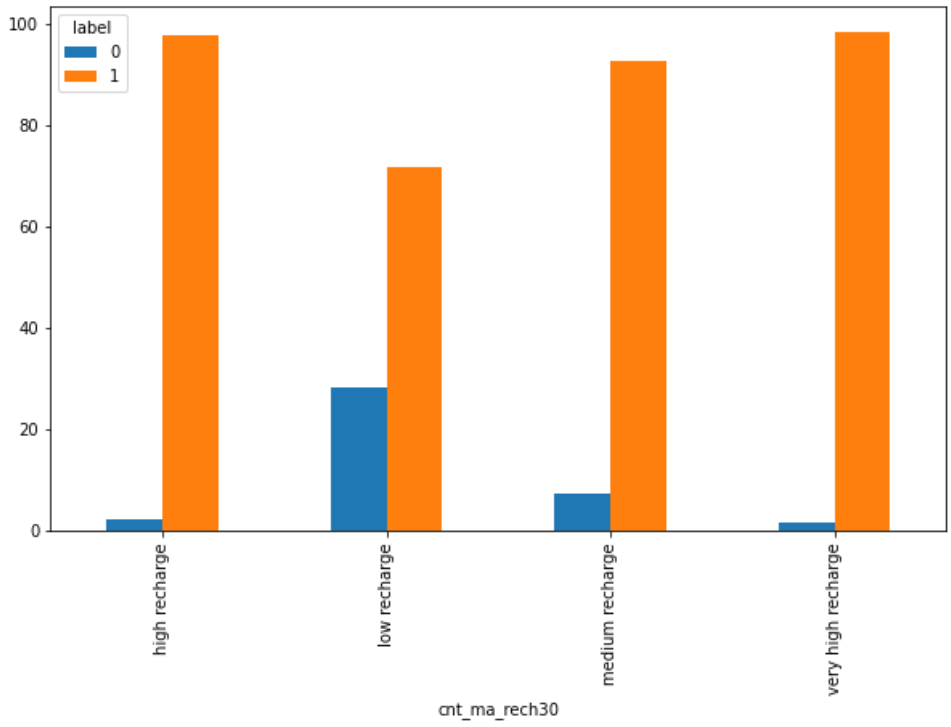
we see that people with low loans tend to default the payments the most.

Number of loans taken by user in last 90 days: cnt_loans90



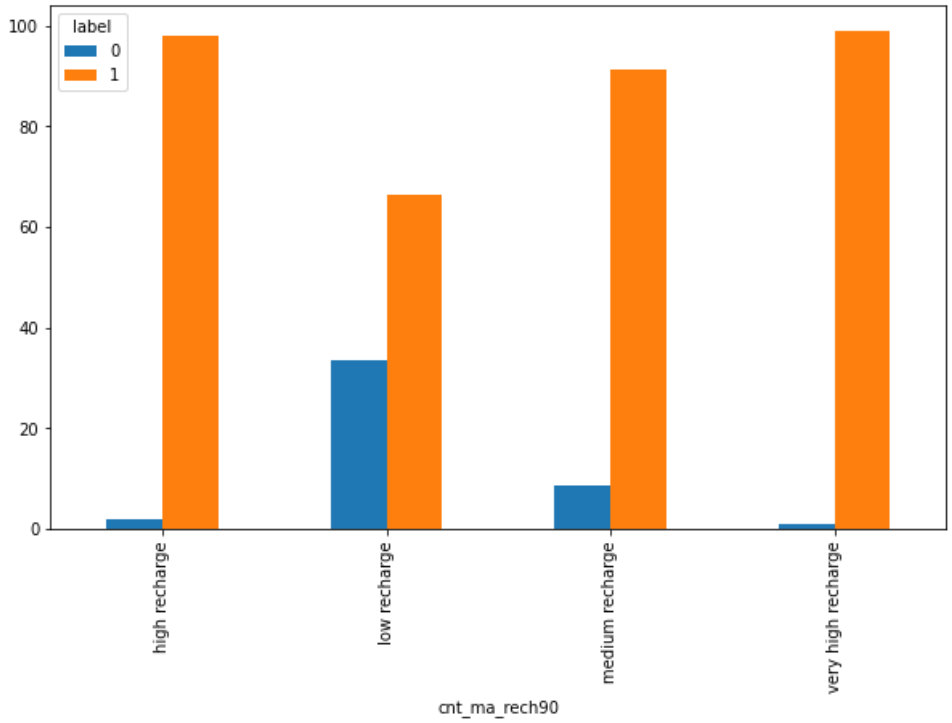
We see a similar trend where in people with low loans tend to default the payments the most.

Number of times main account got recharged in last 30 days: cnt_ma_rech30



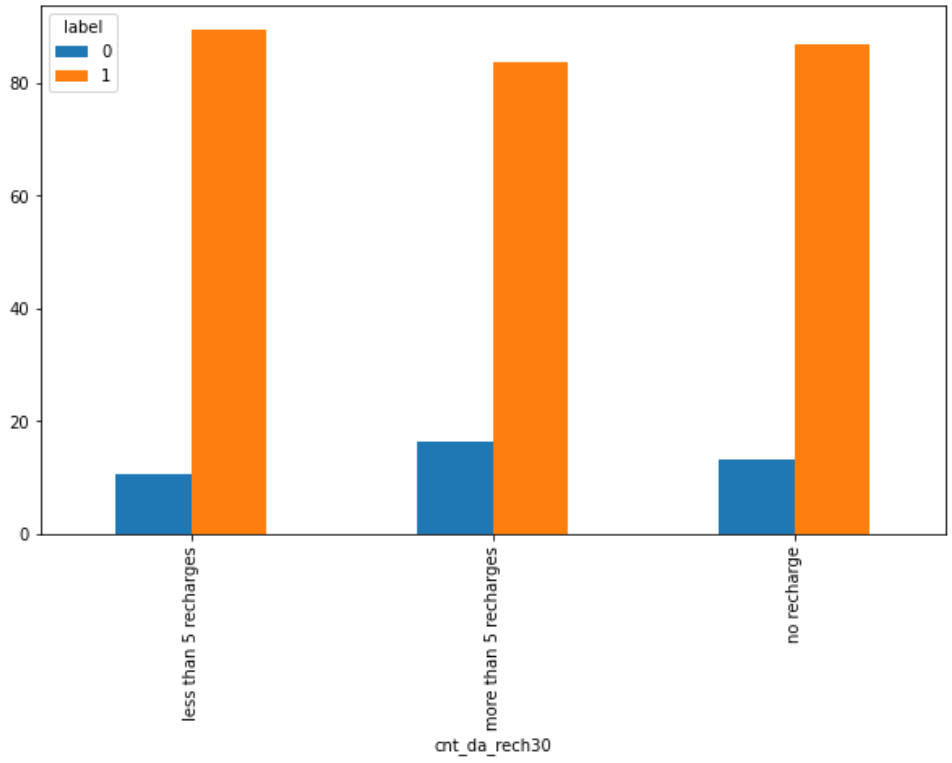
We see that the people who recharge less are the once who default the payments the most.

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



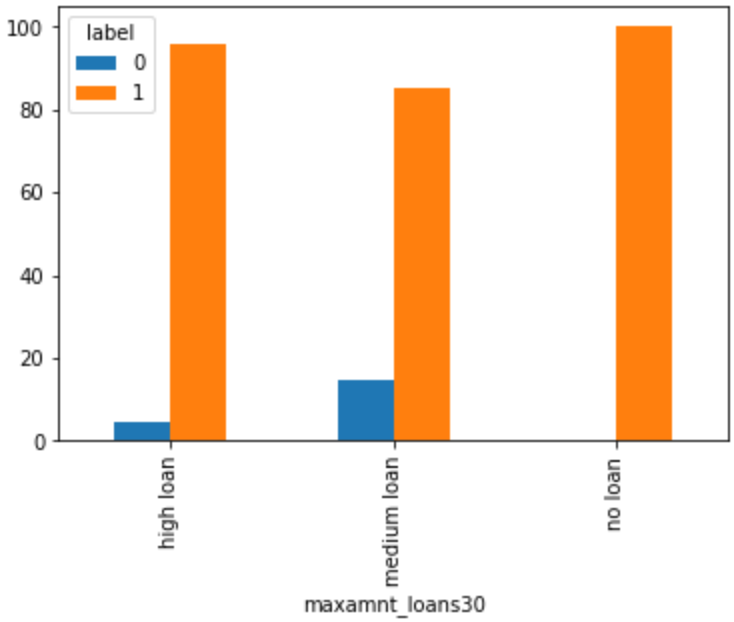
We see a similar trend, where in people who recharge less tend to default the most.

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



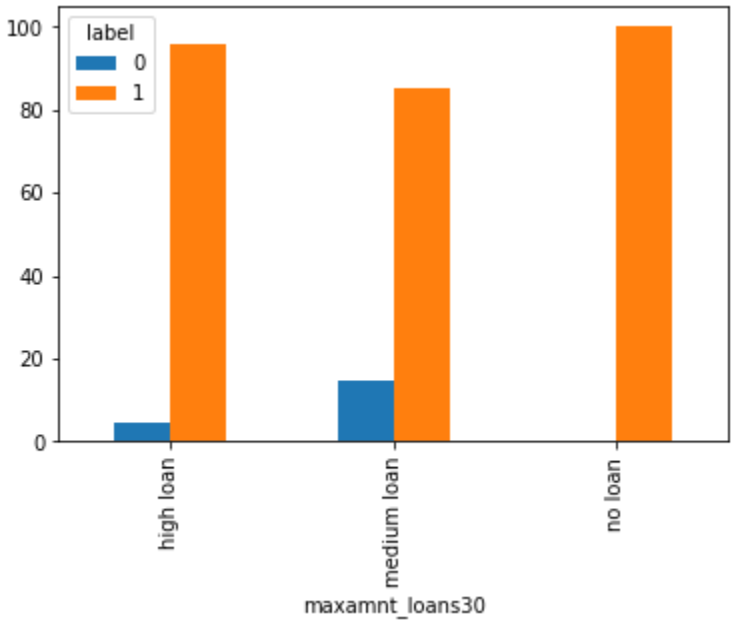
We see people with more than 5 recharges tend to default the most.

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



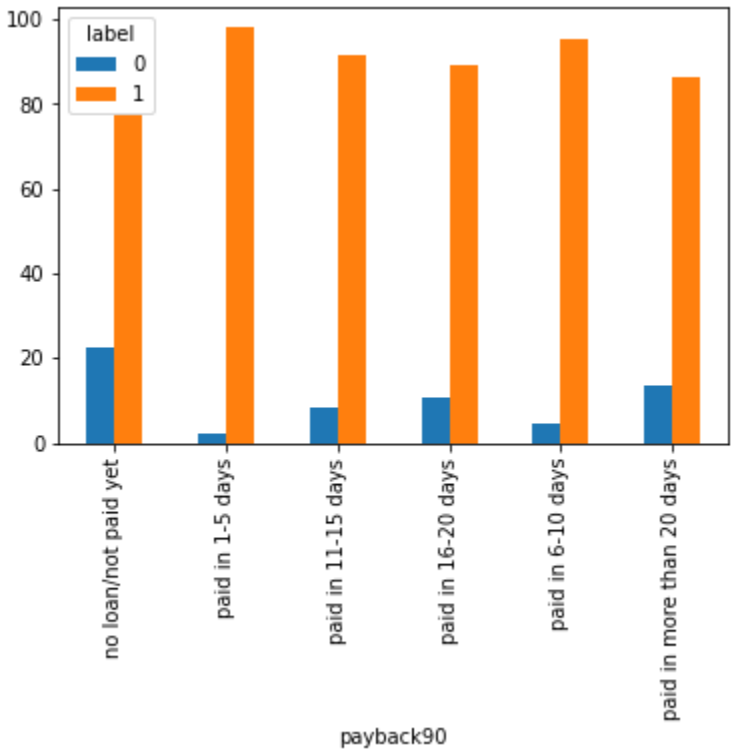
we see that the customers who take medium loans are the once, who default the payments most and people who do not take a loan do not default the payments.

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

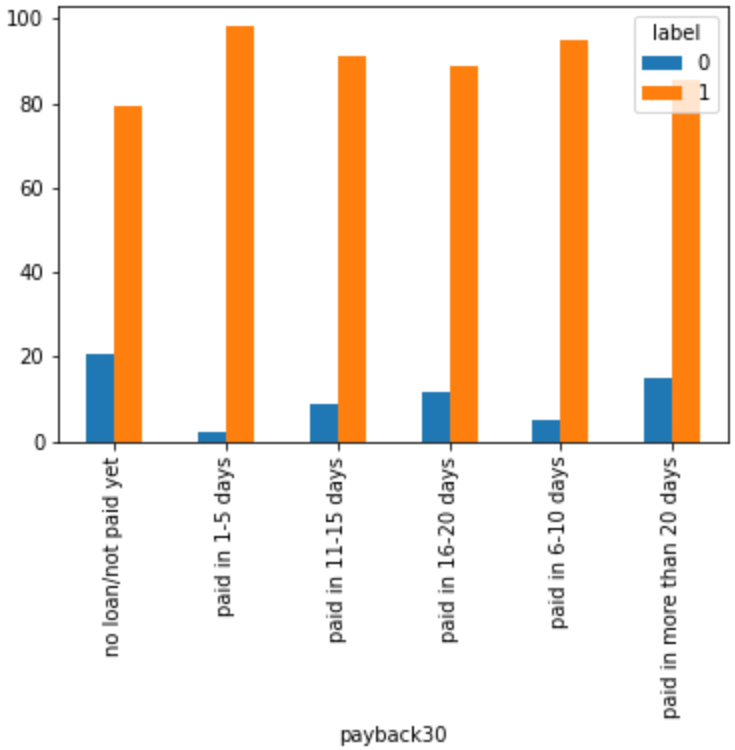


We see a similar trend where medium loan takers default the payments the most and people with no loans do not.

Average payback time in days over last 90 days: payback90

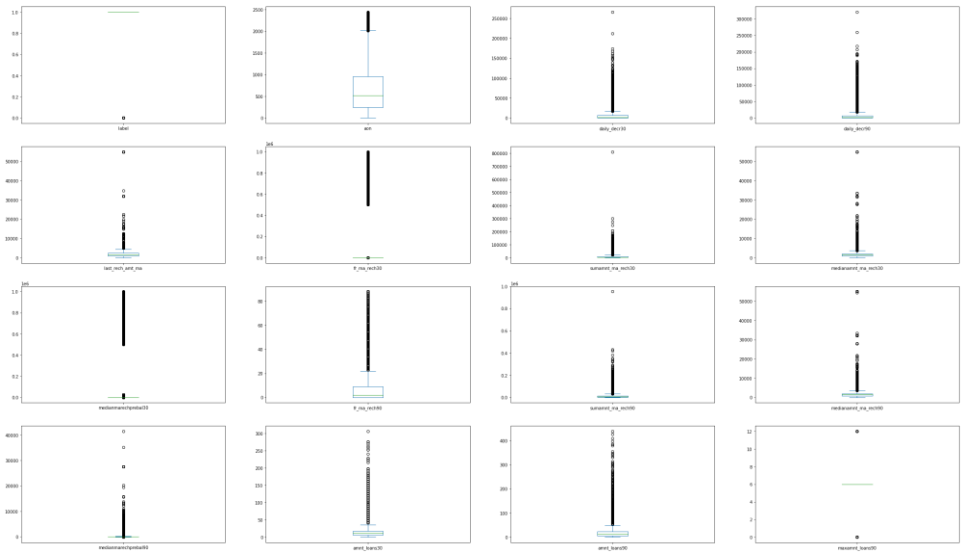


Average payback time in days over last 30 days: payback30



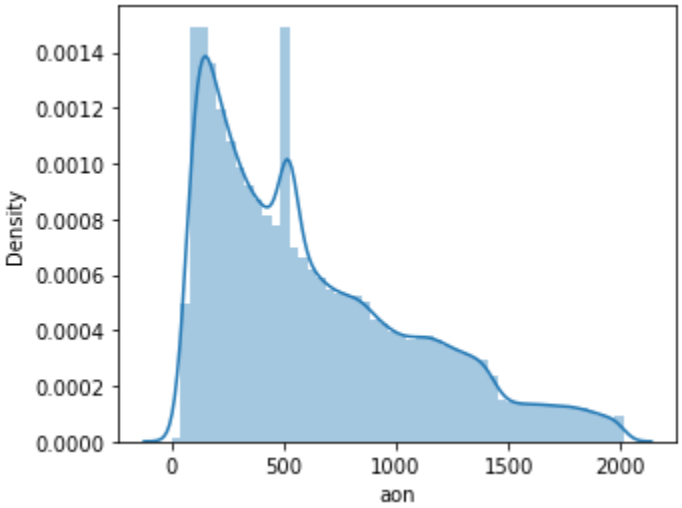
We see a similar trend where in 0 days or no loan/not paid yet contains the highest number of defaulters and paid in 1-5 the least.

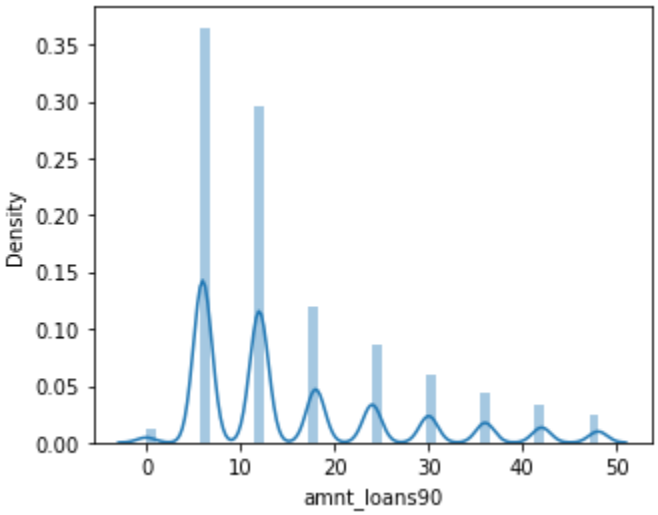
Outliers:

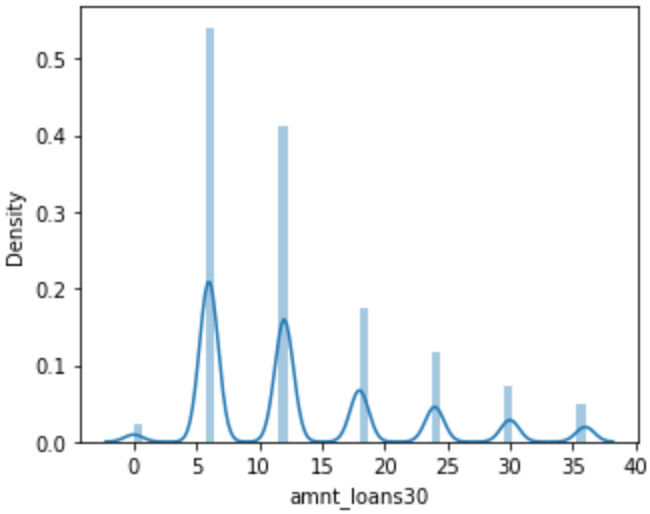


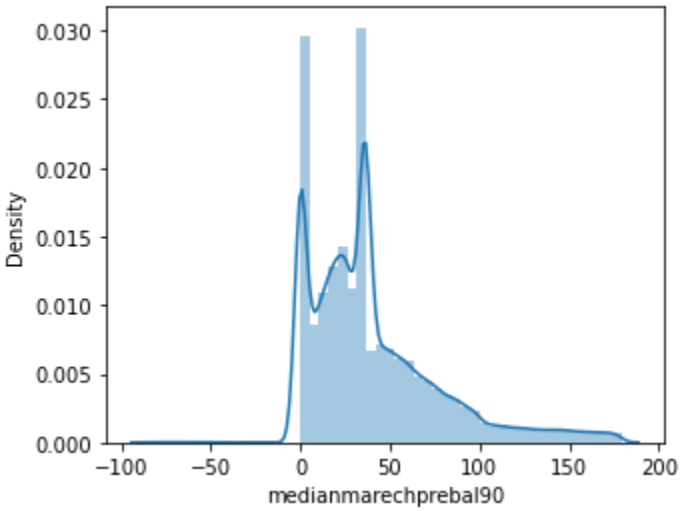
We see presence of outliers in all the columns. Deleting outliers from the data would result in a huge data loss, hence we are going ahead with the median imputation of the outliers.

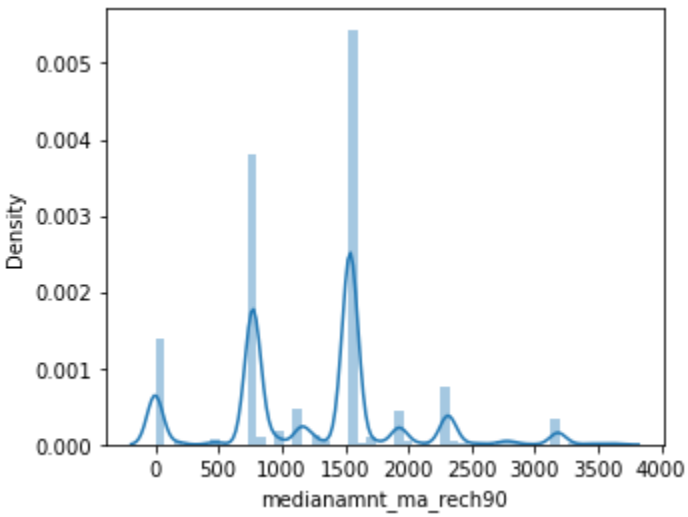
Skewness:

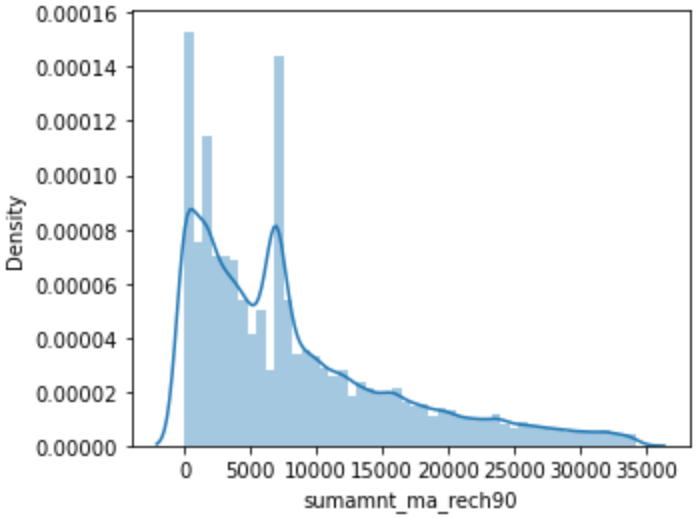


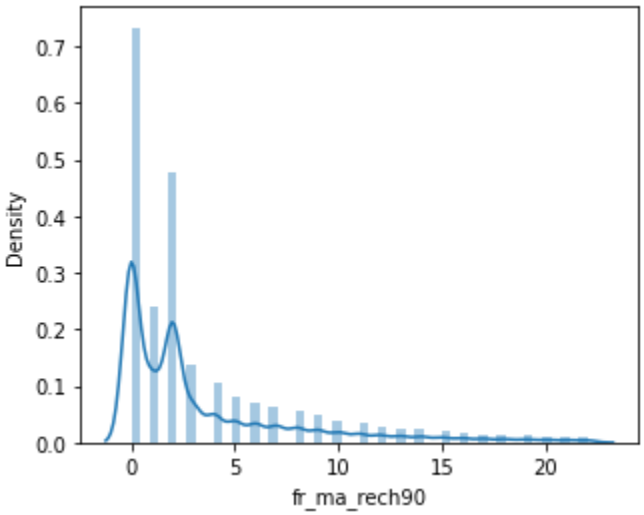


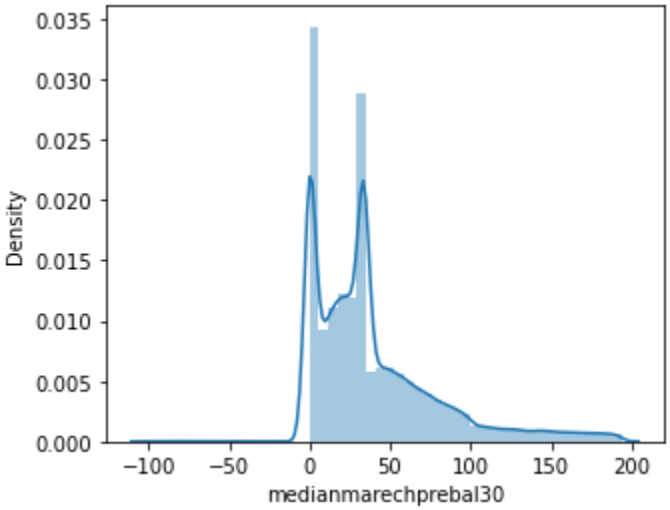

Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
30
35
40
amnt_loans30

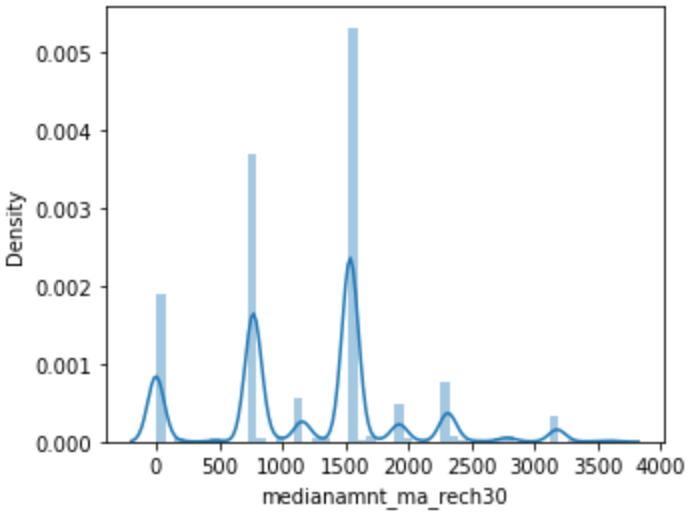

Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
−100
−50
0
50
100
150
200
medianmarechprebal90

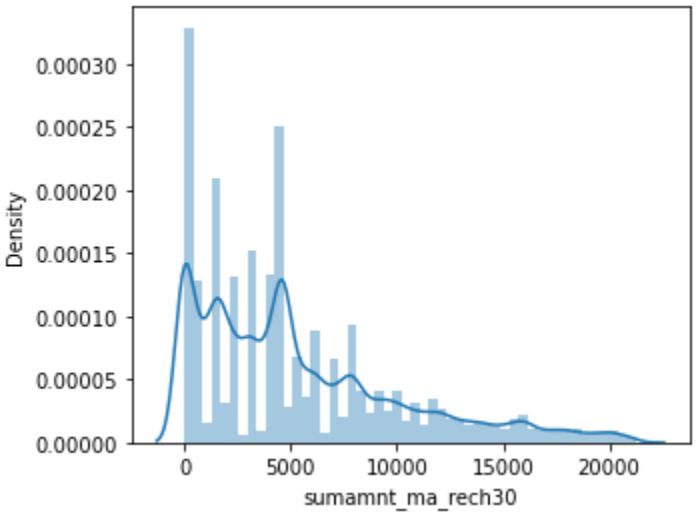

Density
0.000
0.001
0.002
0.003
0.004
0.005
0
500
1000
1500
2000
2500
3000
3500
4000
medianamnt_ma_rech90

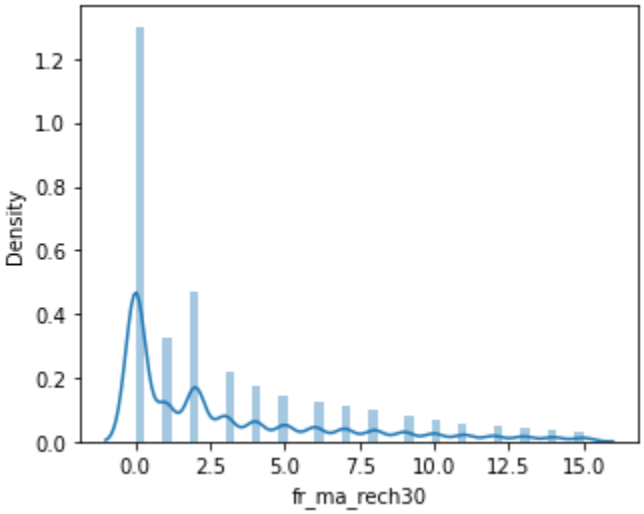

0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
Density
0
5000
10000
15000
20000
25000
30000
35000
sumamnt_ma_rech90

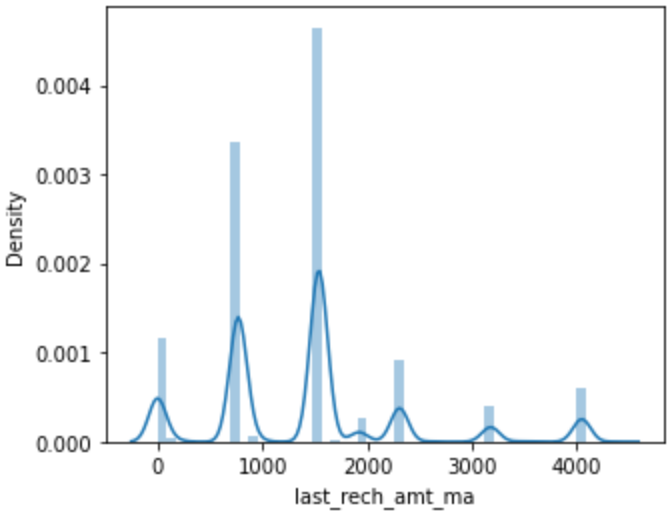

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
0
5
10
15
20
fr_ma_rech90

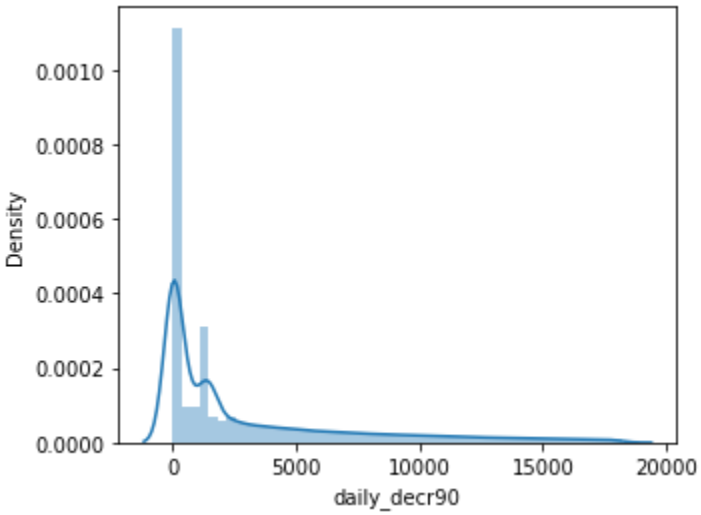

0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
−100
−50
0
50
100
150
200
medianmarechprebal30

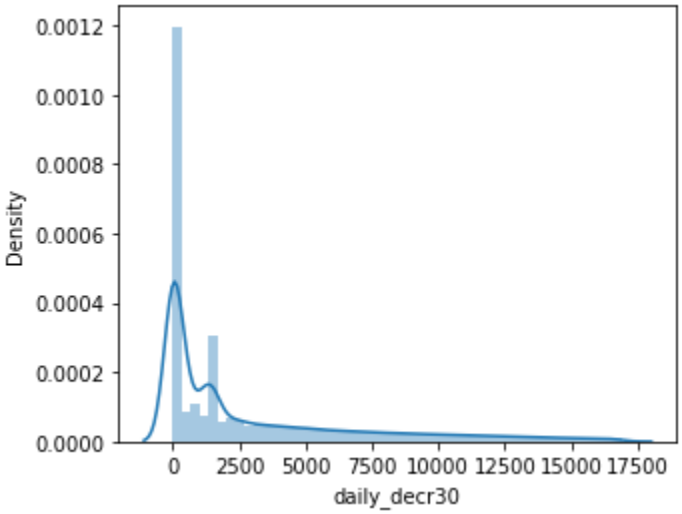

0.005
0.004
0.003
0.002
0.001
0.000
Density
0
500
1000
1500
2000
2500
3000
3500
4000
medianamnt_ma_rech30


0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Density
0
5000
10000
15000
20000
sumamnt_ma_rech30

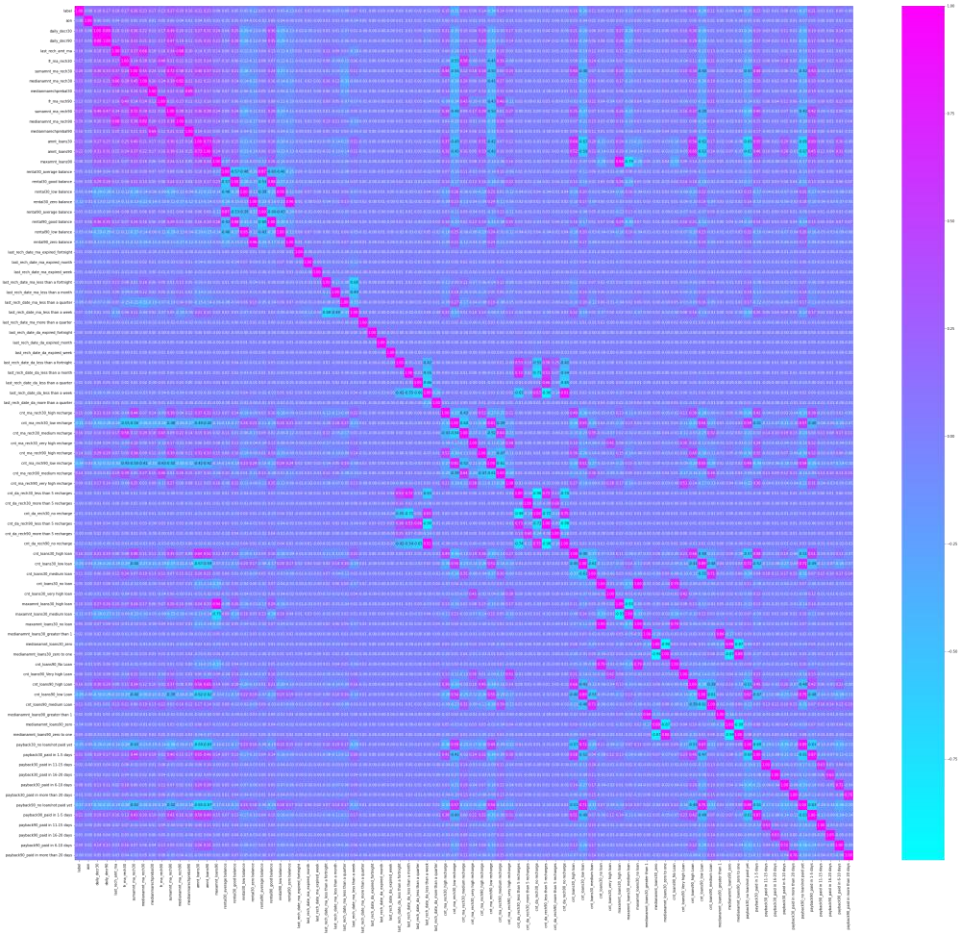

1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
0.0
2.5
5.0
7.5
10.0
12.5
15.0
fr_ma_rech30


0.004
0.003
0.002
0.001
0.000
Density
0
1000
2000
3000
4000
last_rech_amt_ma


0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
5000
10000
15000
20000
daily_decr90


0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
2500
5000
7500
10000
12500
15000
17500
daily_decr30

Correlation:



We see a strong correlation between a lot of features from above, we would go ahead with dimensionality reduction technique PCA to reduce multicollinearity and reduce the dimensions as well.

Interpretation of the Results

We have achieved an accuracy of 92% which is pretty good with a difference in cross-validation of 1.14.

Report for model ExtraTreesClassifier(max_features=50, min_samples_leaf=2, min_samples_split=3, n_jobs=-1)

The Accuracy Score is 92.06058397959714

Confusion Matrix:

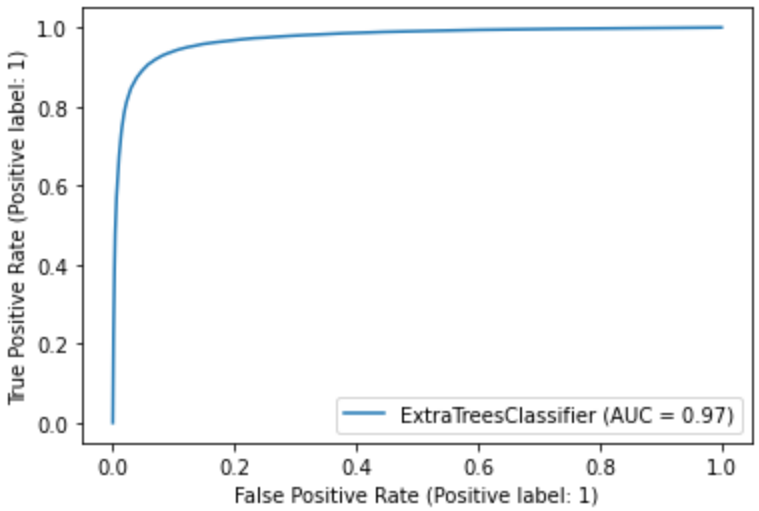
[[45166 3083]

[4544 43272]]

	precision	recall	f1-score	support
0	0.91	0.94	0.92	48249
1	0.93	0.90	0.92	47816
accuracy			0.92	96065
macro avg	0.92	0.92	0.92	96065
weighted avg	0.92	0.92	0.92	96065

Cross Validation Score is 90.91144727309066

Difference between accuracy score and cv is 1.1491367065064821



CONCLUSION

Key Findings and Conclusions of the Study

Mostly, the customers have the intension of repaying. There are certain cases, when the customers have no intension of repayment but the number of such customers are few. With the model built, we can certainly determine customers having intension of repayment or not.

Learning Outcomes of the Study in respect of Data Science

The dataset was full of outliers, skewness and unbalanced data which was the biggest challenge to overcome. Hence data cleaning was very important to get proper prediction. I have used Logistic Regression, GradientboostingClassifier and ExtraTreeClassifier. Among the three algorithms ExtraTreeClassifier gave the best outcome.

Limitations of this work and Scope for Future Work

The solution can be applied to the customer having a transaction history but the model may not perform well with customer having new profile and no transaction history. Nevertheless, the model will perform well with customer having transaction history and can predict whether a person will be a defaulter or non-defaulter. Hence, we can say that this statistical model will be helpful in future for the prediction of micro credit defaulter and non-defaulter customer.