MICRO CREDIT DEFAULTER PROJECT

Submitted by:

John Tojo

Data science intern

ACKNOWLEDGMENT

It gives me immense pleasure to deliver this report. Working on this project was a great learning experience that helped me attain indepth knowledge on data analysis process. Flip Robo Technologies (Bangalore) provided all of the necessary information and datasets, required for the completion of the project. I express my gratitude to my SME, Gulshana Chaudhary, for providing the dataset and directions for carrying out the case study procedure

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).

- Conceptual Background of the Domain Problem
    a. The borrowers are generally from low-income backgrounds.
    b. Loans availed under microfinance are usually of small amount, i.e., Loans of value 5& 10 rupees are provided by our client (telecom operator) in collaboration with a Microfinance Institute (MFI).
    c. The loan tenure is short, 5 days
    d. High return (20%) as well as High risk as it doesn't require any collateral
    e. These loans are usually repaid at higher frequencies
    f. The purpose of most microfinance loans is income generation

Analytical Problem Framing

- Data Sources and their formats

  This data is given to FlipRobo from their client dataset. The information given in the dataset is in the CSV format. There are 209593 rows and 37 columns in the dataset provided.

- Data Preprocessing Done
    1. The .info() used to get info
    2. Check for nulls, drop column with more than 80% null datas.
    3. nunique() fuction was used to check for unique data present in each feature, drop column for those whose value are all unique or has only 1 unique value
    4. Check duplicates
    5. Divide data into three categories
        a. Numerical continuous data
        b. Numerical discrete data
        c. Categorical data

6. Imputation for numerical and categorical data
7. Visualize and statistical interpretation
8. Treat skewness
9. Check correlation and check for multicollinearity
10. Outliers check for train dataset

- Hardware and Software Requirements and Tools Used

Software Technology being Used:

- ➢ Programming language: Python
- ➢ Distribution: Anaconda Navigator
- ➢ Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

- ➢ Pandas, NumPy, matplotlib, seaborn and scikit-learn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

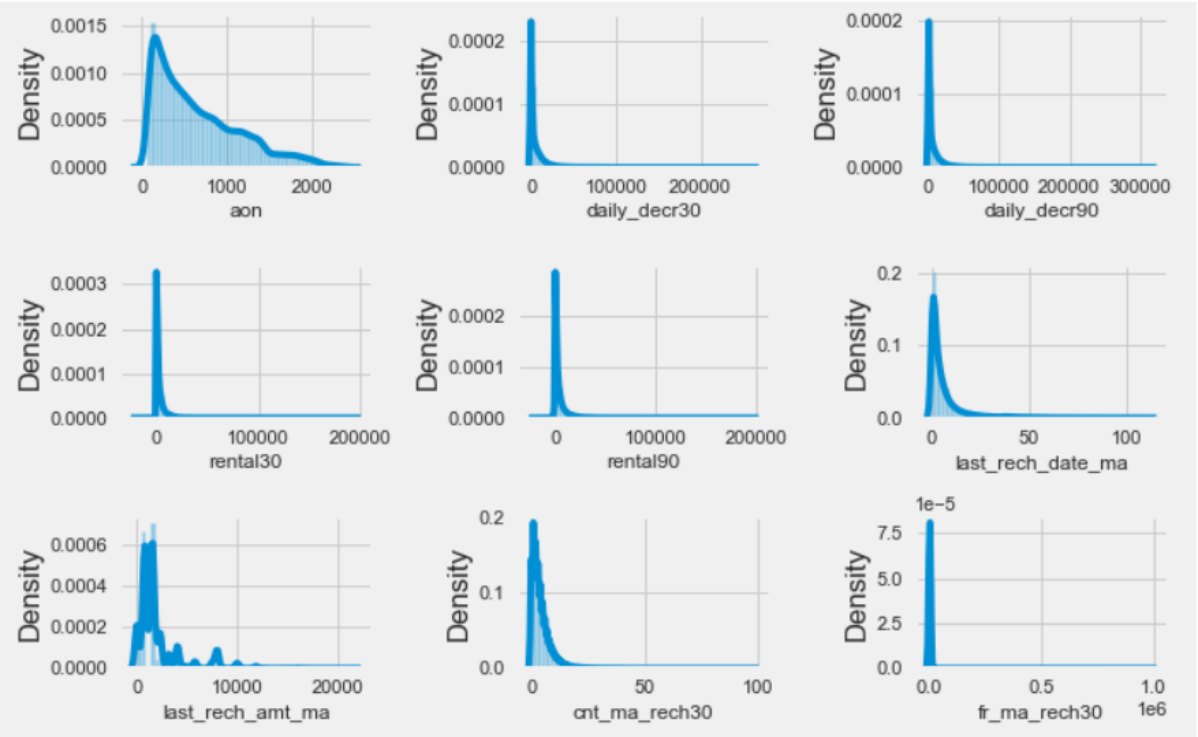
a. Data Pre-processing Done

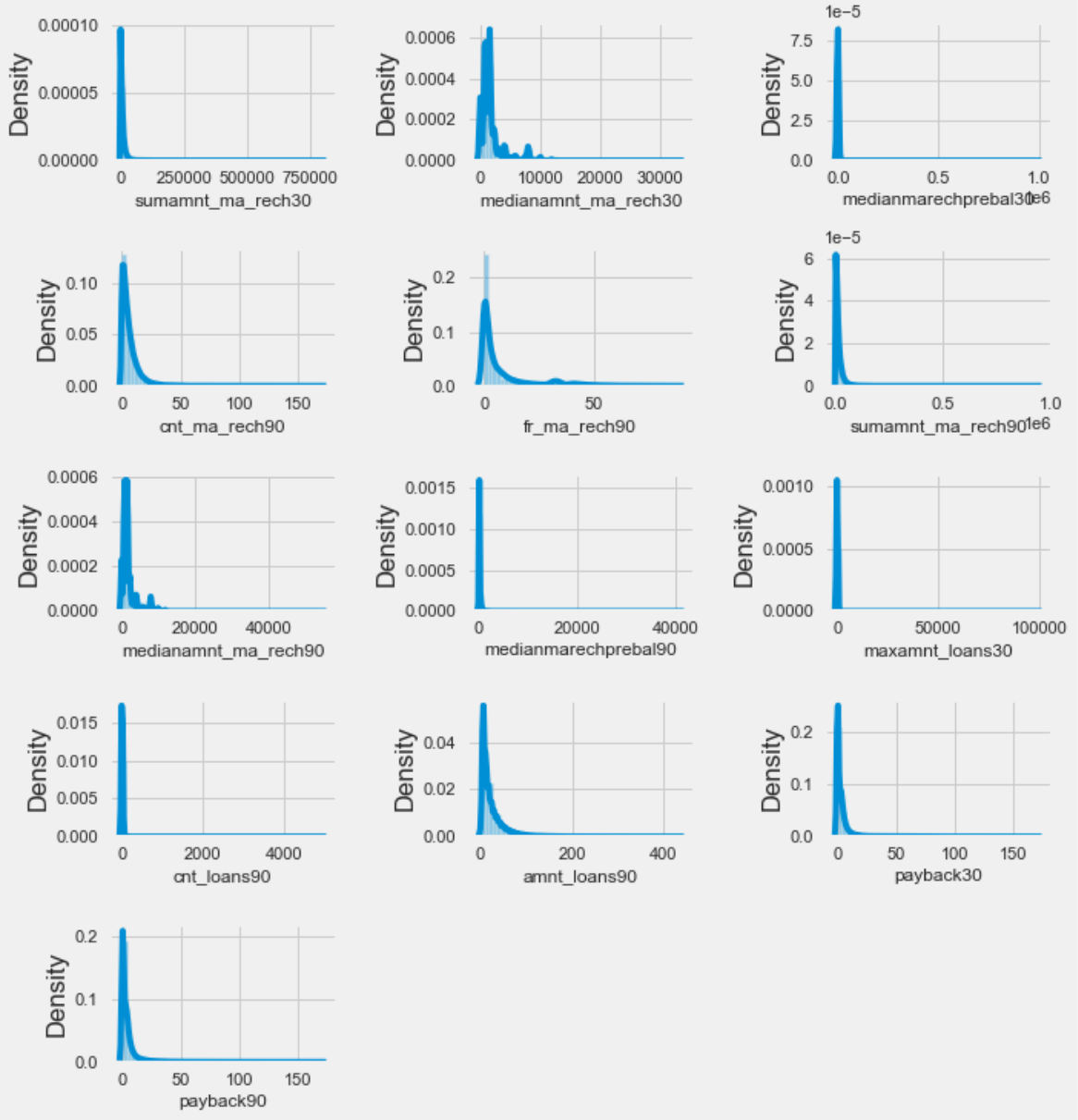
i. Pdate was split into 3 columns Day, Month and Year, Pdate was deleted
ii. The data type of features presents were checked and had two object datatype and rest all were numeric
iii. While checking for number of unique value present in each feature, it was found that pcircle and year had only 1 unique data and Unnamed: 0 had all unique data hence these columns were dropped in addition to which msisdn was dropped as it was phone number of the customers

iv. No nulls were present
v. The number of zero value present were checked and those features having more than 80% of value as zero were filtered out and dropped
vi. The features were divided into wo numerical continuous and numerical discrete. There were no object features after dropping the above mentioned features
vii. Features having negative values were filtered out, among this it was found that it was not possible for aon and last_rech_date_ma to have negative values, index where negative values present were filtered and rows were dropped
viii. Dropped extreme values which made no sense for aon and last_rech_date_ma
ix. From describe function it can be seen that the values are highly spread

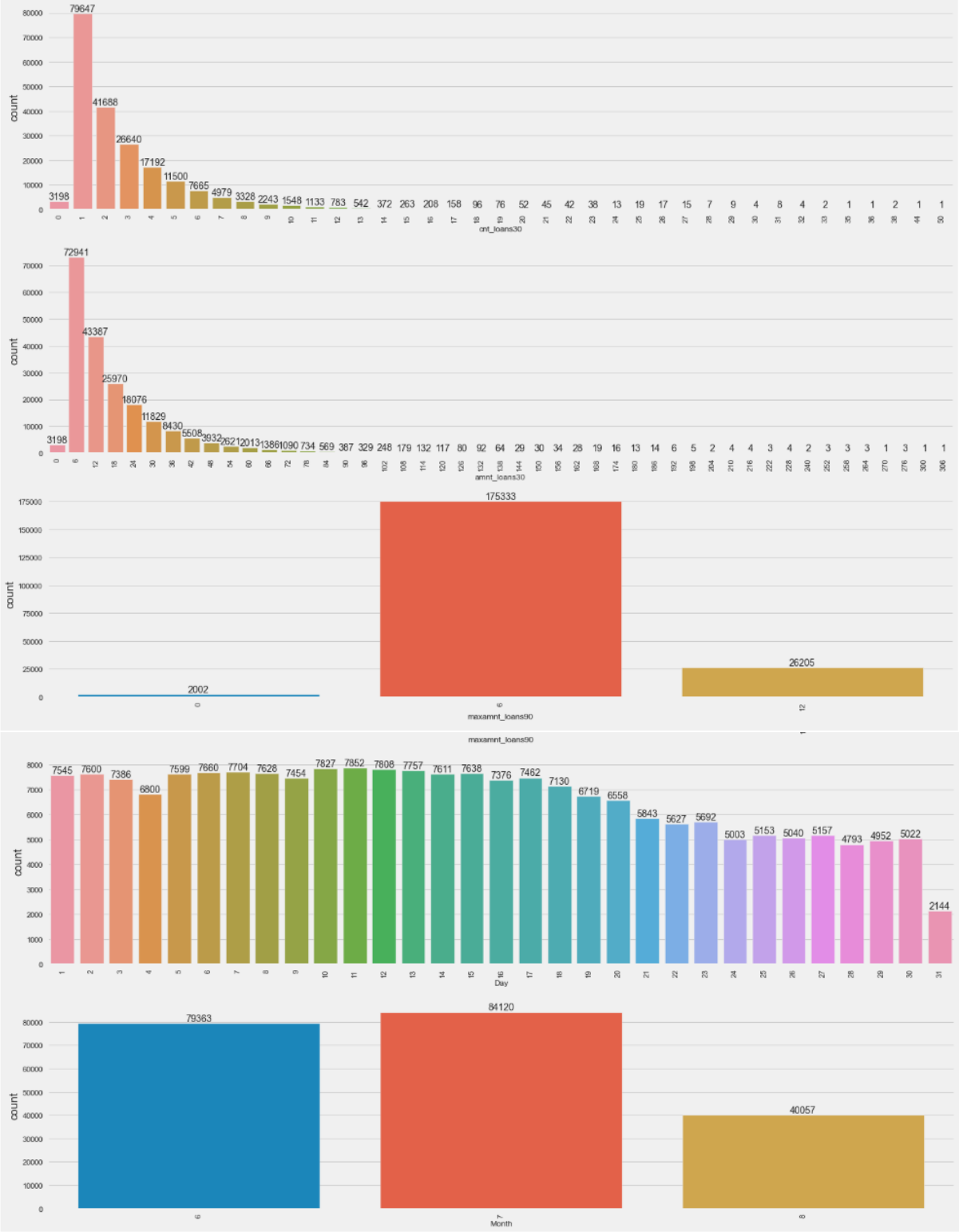
b. Visualisation

i. Univariate





From the graphs obtained it can be seen that data is highly skewed and outliers are present

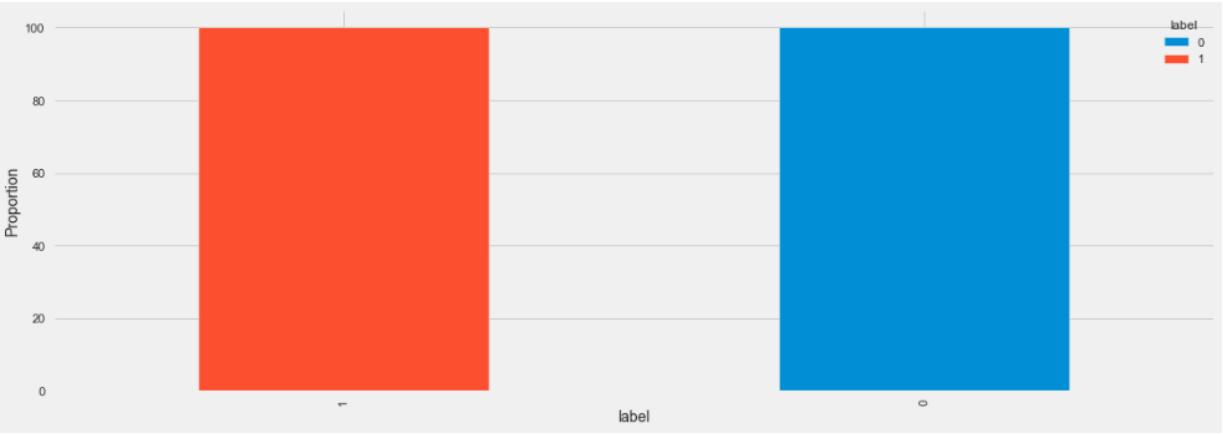


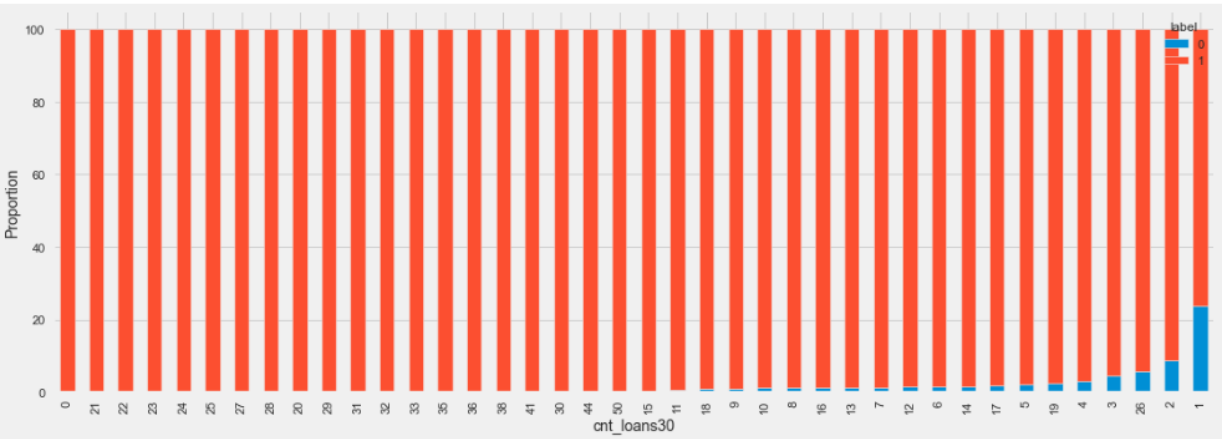
Observations

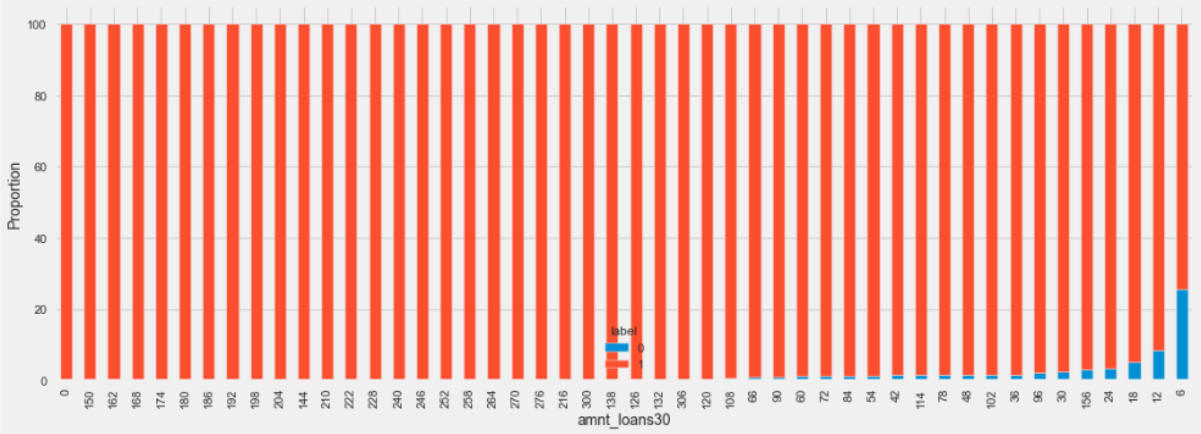
- cnt_loans30
  - most of the customers have taken 1 loan in past 30 days
  - there are some cases where no of loans taken are more than 20

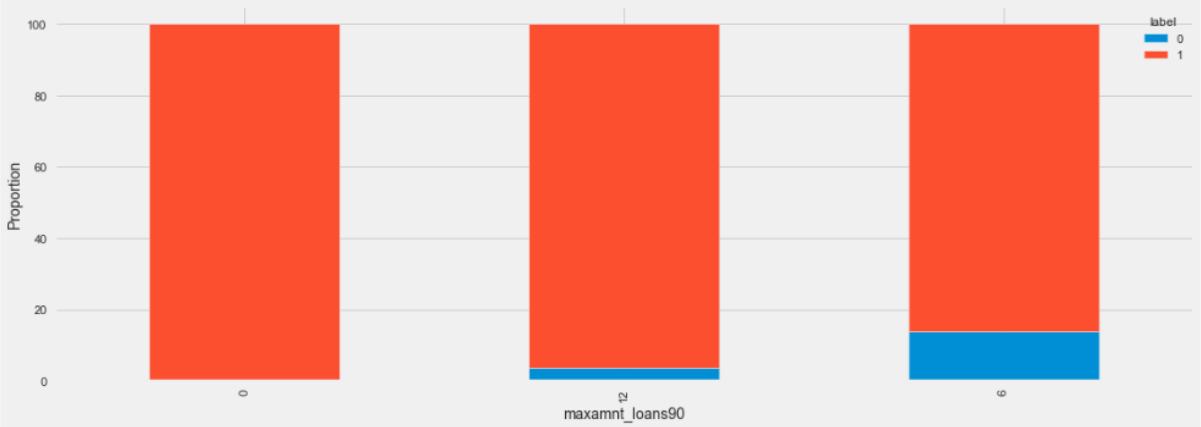
- amnt_loans30
  - total amount of loan taken by the customer in the past 30 days is mostly 6 followed by 12
- maxamnt_loans90
  - the max amount of loans taken by cusomter in 90 days is 6
- Day
  - days 6, 7, 10, 11, 12 has the largest amount of data
- Month
  - 7th month has most amount of data

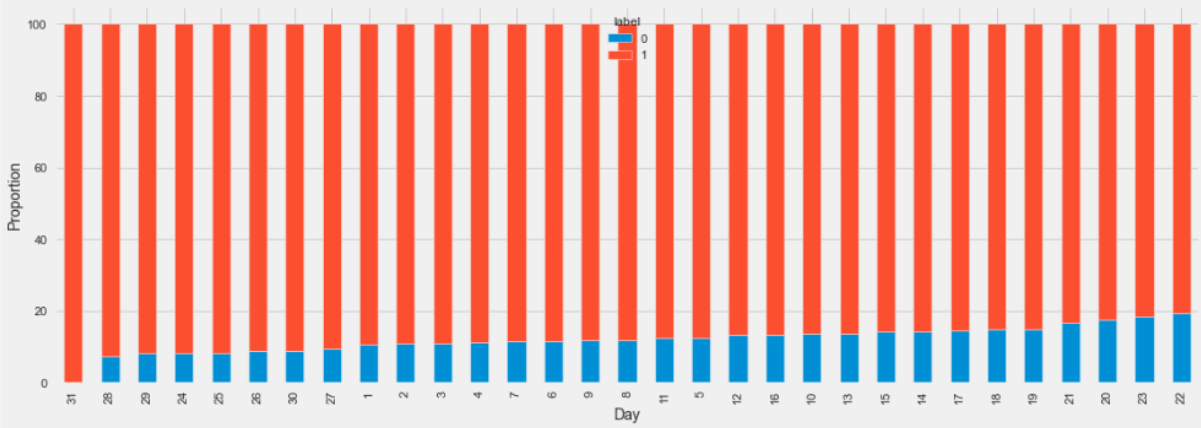
ii. Bivariate Analysis

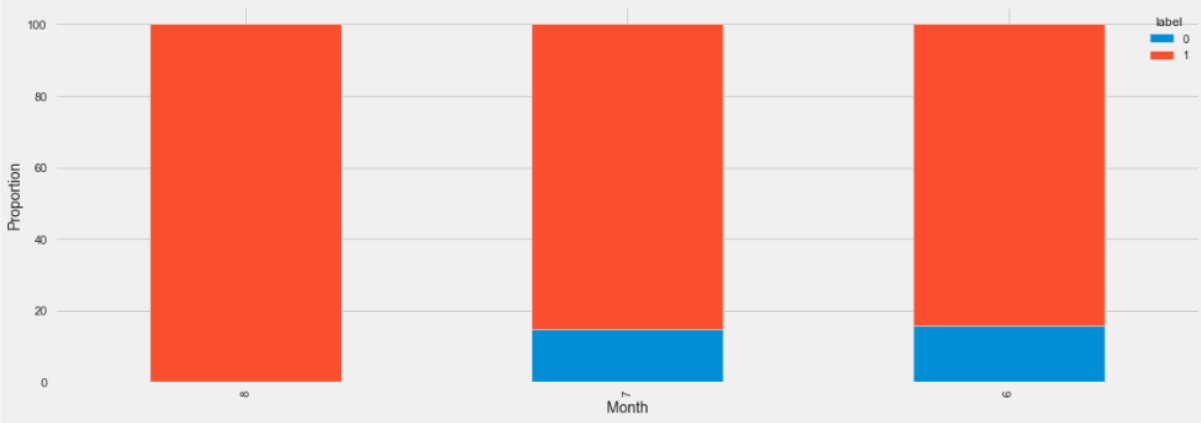


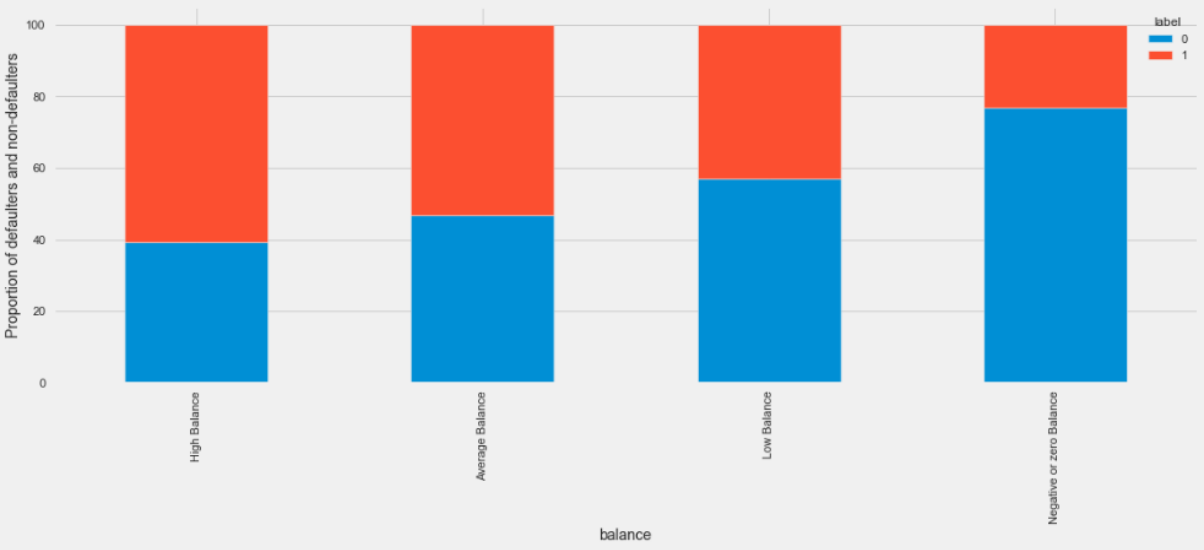



Proportion
100
80
60
40
20
0
label
0
1
0 150 162 168 174 180 186 192 198 204 144 210 222 228 240 246 252 258 264 270 276 216 300 138 126 132 306 120 108 66 90 60 72 84 54 42 114 78 48 102 36 96 30 156 24 18 12 6
amnt_loans30

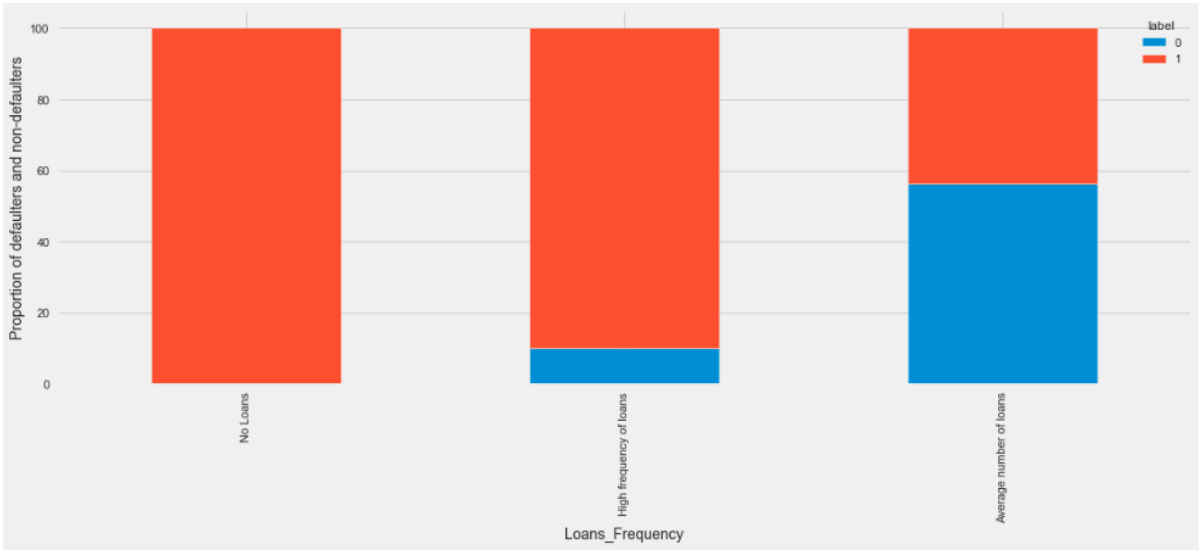

Proportion
100
80
60
40
20
0
label
0
1
0
12
6
maxamnt_loans90


Proportion
100
80
60
40
20
0
label
0
1
31 28 29 24 25 26 30 27 1 2 3 4 7 6 9 8 11 5 12 16 10 13 15 14 17 18 19 21 20 23 22
Day

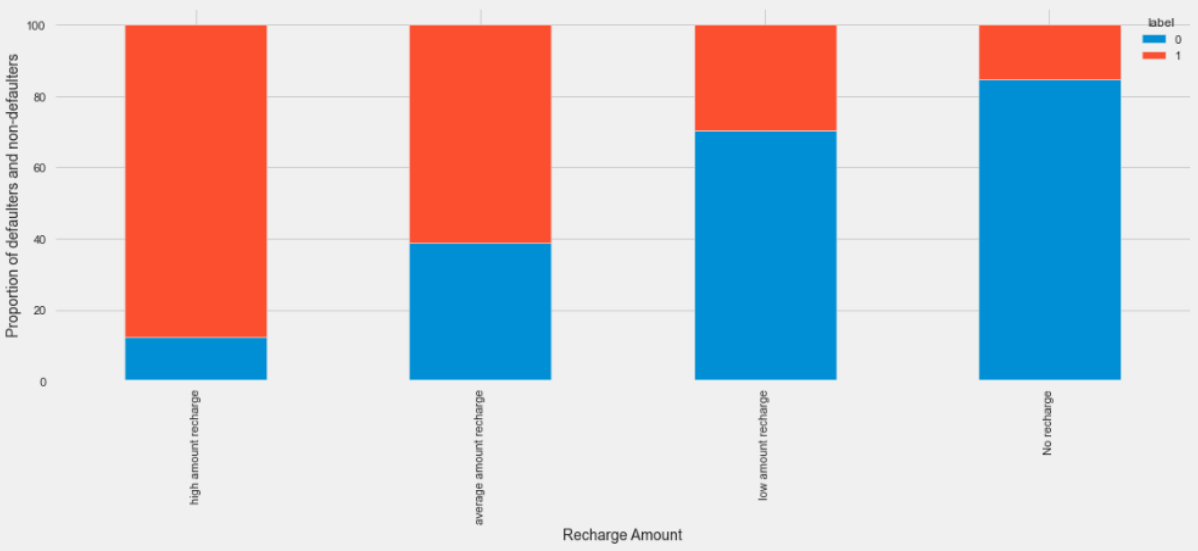

Proportion
100
80
60
40
20
0
label
0
1
8
7
6
Month



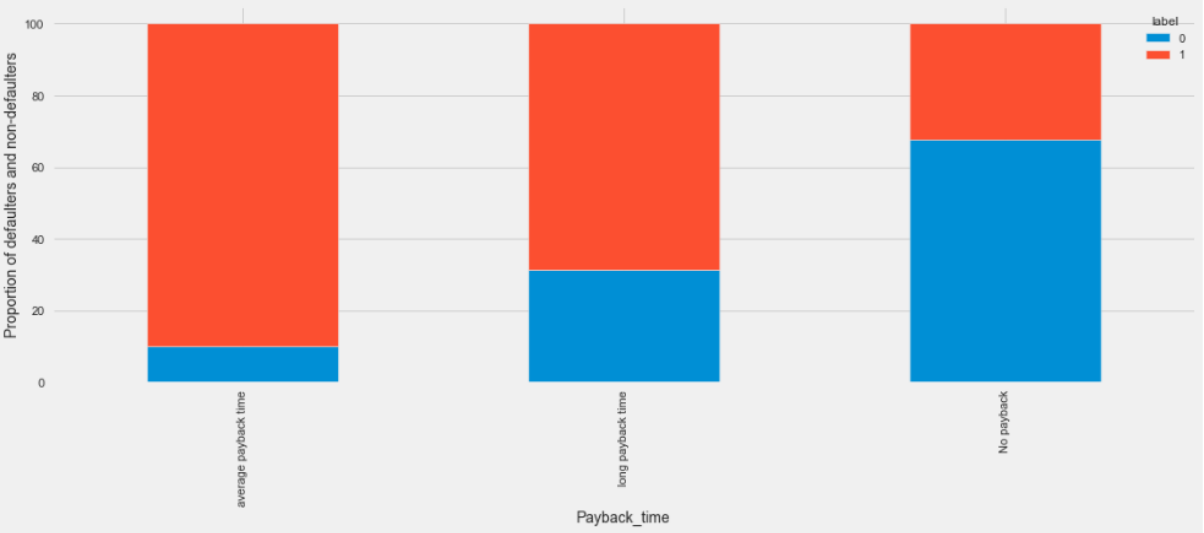
- negative or zero balance have high chance to default



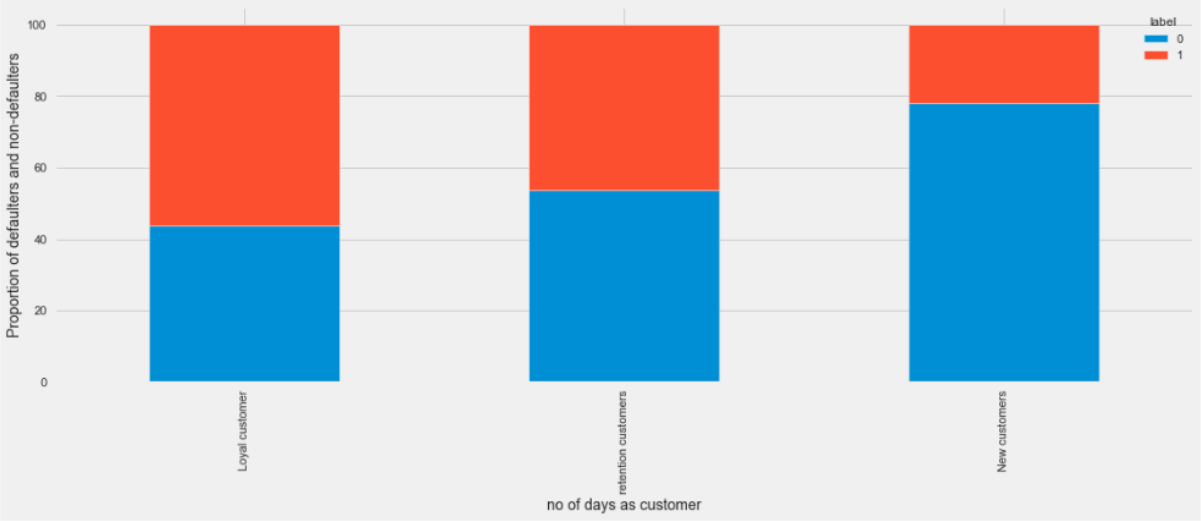
- customers who take loans in the range 0-5 have high chance to default



- customers those who dont recharge their accounts have a chance to default



- defaulters are high for those who have not paid back



- defaulters are highest among new customers

c. Data cleaning

I. Skewness

The skewness was checked and all those columns which were not within limits were first transformed using power transform and then cbrt method was used and the above mentioned step was repeated again and columns which were not within limits were dropped

II. Outliers

Outliers presence were checked using box plots it was found that outlies were present and treated using zscore, as there were limitation on data loss no further treated was done as the total data loss after removing rows and zscore method had accounted for 6.8%

III. Correlation
It can be seen that features were not highly correlated to the target and heatmap was plotted to see feature to feature correlation, it was found that some feature had high correlation with other features which meant that it could result in multicollinearity

IV. VIF
Vif method was adopted to treat multicollinearity and all those features which where not within limits were dropped

V. Balanced dataset
As the problem is classification type, it was needed to check if the amount of data present in class 0 and class 1 were almost equal, as it was not the case used smote to increase the data present in class 0

- Testing of Identified Approaches (Algorithms)

The target variable is label which is numerical discrete datatype which take 0 and 1 classification problem
. The following models were used for the analysis

1. Random Forest
2. Adaboost classifier
3. Gradientboost classifier
4. Decision Tree
5. KNN

	test accuracy	max_cv_score	diff	mse	mae	F1 score	roc_curve_area
DT	90.0	90.386366	0.000951	0.099783	0.099783	0.912317	0.897652
RF	94.2	94.471767	0.032831	0.057963	0.057963	0.949546	0.941774
adaboost	84.5	84.915290	0.044895	0.155186	0.155186	0.863800	0.841205
KNN	88.0	88.631245	0.074017	0.120492	0.120492	0.884662	0.884327
GRAD	89.7	89.425214	0.274786	0.103295	0.103295	0.909824	0.894660

- Random Forest is the best model
    1. highest roc_curve area
    2. highest test accuracy and cv_score
    3. 2nd highest in difference between cv_score and test accuracy
    4. least error compared to other model
    5. highest F1score, precision and recall
    6. importance is to reduced FN, which implies that in actual case its default but prediction is not defaulter, the least value of FN is for KNN, followed by RF

CONCLUSION

- Key Findings and Conclusions of the Study
    a. The customers who maintain a negative or zero balance or more likely to default followed by low balance accounts
    b. The customers who have moderate frequency of taking loans have higher chance to default
    c. The customers who fail to recharge their account could potentially end up defaulting followed by low amount recharge in their account

d. Customers who don’t payback or have longer duration of payback have very high chance to default
e. It can be noted that new customers are more likely to default, followed by retained customers
f. Most of the defaulting was done on month 6
g. Customers who take loan on 22nd have high chance to default

- Learning Outcomes of the Study in respect of Data Science
    a. The data was highly skewed with large number of outliers and unrealistic data was present. The most challenging part was to reduced the data loss and maintain in within a limit of 7%.
    b. The skewness couldn’t be reduced for all column, need to research on other treatment available

- Limitations of this work and Scope for Future Work
    a. The data had large number of unrealistic data present and outliers, while cleaning the data it was difficult to reduce the data loss within the specified limit. As the data was bought it should be taken care not to have such unrealistic data as much as possible
    b. The data given were highly skewed

c. More factors should be taken into account such as the **income** of the person who is taking the loan and assign a **credit score** so that it is possible to cluster the customers based on it and reduce the defaulting rate