Micro Credit Loan Prediction Project

Submitted by:
ANISH ANTONY

ABSTRACT:

One such client that is in Telecom Industry have a fixed wireless telecommunications network provider offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber. 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). In this project we will build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been paid i.e., non-defaulter, while, Label '0' indicates that the loan has not been paid i.e., defaulter

To predict this model, we need to have the data preprocessed, trained and tested using the classification algorithms. Then it is hypertuned and best algorithm with best parameters is obtained and finally the loan status is predicted.

Keywords: Loan, Data cleaning, payback amount, classification

CHAPTER I

INTRODUCTION

1.1 Business Problem Framing:

Problem Description:

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.

Business Objectives:

We have to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been paid i.e., non-defaulter, while, Label '0' indicates that the loan has not been paid i.e., defaulter.

1.3 Review of Literature

From the article “Predicting mortgage early delinquency with machine learning methods” published in European Journal of Operational Research.

This paper investigates the performance of thirteen methods for modelling and predicting mortgage early delinquency probabilities. These models include variants of logit models, some commonly used machine learning methods, and variants of ensemble models. We find that heterogenous ensemble methods lead other methods in the training, out-of-sample, and out-of-time datasets in terms of risk classification. Nonetheless, various predictive accuracy performance measures yield different rankings among the thirteen methods and no method consistently dominates in this performance dimension in the training, out-of-sample, and out-of-time data. Lastly, predictive accuracy is a major challenge facing all mortgage early delinquency models, even in the training data.

From the article “A Research Paper on Loan Delinquency Prediction”

Evaluating and forecasting the willingness of borrowers to repay is critical for banks to minimize the possibility of defaults in the payment of loans. For this purpose, there is a mechanism set up by banks to manage a loan request depending on the status of the lender, like job status, history of credits and many more. But the new assessment scheme may not be sufficient for assessing the willingness of such borrowers to repay, such as students or non-credit historians. To better determine the willingness of all sections of people to repay, we tried different machine learning models on the Kaggle data model and assessed the value of all the features used. This paper uses Machine Learning to identify factors that influence mortgage defaults. Our main purpose is to figure out the delinquency status of loans for the any ‘n’ next month given the delinquency status for the previous 12 months. We apply the adjusted Gradient Boosting (XGBoost) approach to data. First, we prove that the precision of XGBoost estimation is better than the logistic regression. Second, we use the Permutation Feature Importance approach to identify the most significant variables affecting delinquency. This paper uses loan statistics over an extension cycle and points out the value of the interest rate in the determination of the delinquency. Our findings vary from those found in the literature. In our Combined Loan-to-Value and Unemployment model, while significant, they are both dominated by the factors described above. This comparison in data draws attention to the importance of the market cycle in deciding the causes of delinquency

1.4 Motivation for the Problem Undertaken

Our main goal is to design and develop a system that assists in detecting and formulating the telecom delinquency in a detailed way that can be easily understood by telecom operators. Since it involves prediction work. Machine learning provides a vast range of classification problems and decision-making algorithms which has become our best way for this problem. Supervised learning approach which we are going to use provides a perfect solution since the program learns from the input data and uses the output results to analyse new observations. There are many algorithms such as Naïve Bayes Classifier algorithm, Logistic Regression model, Decision trees model and Random Forest algorithm, Support Vector machines and many more. Here we are providing a brief explanation of algorithms we have used.

CHAPTER II

Analytical Problem Framing

2.1 Mathematical/ Analytical Modeling of the Problem:

Machine Learning:

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists chooses to use depends on what type of data they want to predict.

Supervised learning:

In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Supervised learning can be separated into two types of problems when data mining—classification and regression

- Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, etc.
- Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

Unsupervised learning:

This type of machine learning involves algorithms that train on unlabelled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

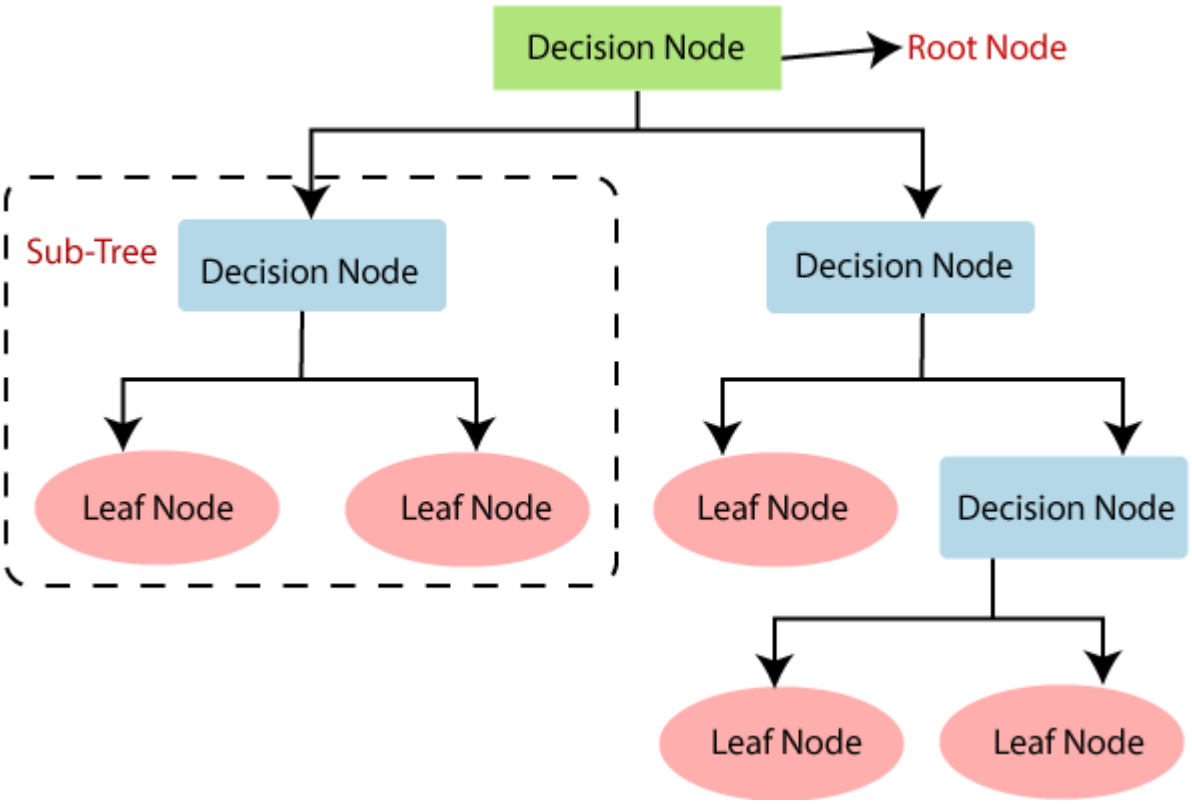
Popular un-supervised algorithms are K-means clustering, affinity propagation etc.

Semi-supervised learning:

This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

Reinforcement learning:

Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.



Decision Tree Algorithm:

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

As you can see from the diagram above, a decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes. The leaf nodes represent all the possible outcomes within the dataset.

Random forest Algorithm:

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type i.e., multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

Statistical Analysis:

Statistical analysis, or statistics, involves collecting, organizing and analysing data based on established principles to identify patterns and trends.

Predictive analysis:

Predictive analysis uses powerful statistical algorithms and machine learning tools to predict future events and behaviour based on new and historical data trends. It is important to note that predictive analysis can only make hypothetical forecasts and the quality of the predictions depends on the accuracy of the underlying data sets.

The terms used for statistical analysis are:

Mean	$\overline{x} = \frac{\sum x}{n}$	x = Observations given n = Total number of observations

Median	If n is odd, then
	$M = \frac{n+1}{2th}$ term
	If n is even, then
	$M = \frac{\left(\frac{n}{2}\right)th\ term + \left(\frac{n}{2}+1\right)th\ term}{2}$
		n = Total number of observations

Mode	The value which occurs most frequently	
Variance	$= \sigma^2 = \sum \frac{(x-x)^2}{n}$	x = Observations given = Mean n = Total number of observations

Standard Deviation	$S = \sigma = \sqrt{\sum \frac{(x-x)^2}{n}}$	x = Observations given $\bar{x}$ = Mean n = Total number of observations

$$\text{Z score} = \frac{x-\bar{x}}{\sigma}$$

Where,

x = Standardized random variable

$\bar{x}$ = Mean

σ = Standard deviation.

Quartile Formula:

When the set of observation is arranged in an ascending order, then the 25th percentile is given as:

$$Q_1 = (\frac{n+1}{4})^{th} \text{ term}$$

The second quartile or the 50th percentile or the Median is given as:

$$Q_2 = (\frac{n+1}{2})^{th} \text{ term}$$

The third Quartile of the 75th Percentile (Q3) is given as:

$$Q_2 = (\frac{3(n+1)}{4})^{th} \text{ term}$$

$$\text{IQR} = \text{Upper Quartile} - \text{Lower Quartile}$$

Regression:

Regression is a statistical technique used to find a relationship between a dependent variable and an independent variable. It helps track how changes in one variable affect changes in another or the effect of one on the other. Regression can show whether the relationship between two variables is weak, strong or varies over a time interval. The regression formula is:

$$Y = a + b(x)$$

Y represents the independent variable, or the data used to predict the dependent variable

x represents the dependent variable which is the variable you want to measure

a represents the y-intercept or the value of y when x equals zero

b represents the slope of the regression graph

Hypothesis testing:

Hypothesis testing is used to test if a conclusion is valid for a specific data set by comparing the data against a certain assumption. The result of the test can nullify the hypothesis, where it is called the null hypothesis or hypothesis 0. Anything that violates the null hypothesis is called the first hypothesis or hypothesis 1.

2.2 Data Sources and their formats:

1. Data Collection Phase

The dataset contains 209593 rows and 34 columns.

2. Data Cleaning:

Analysing the dataset, we find the following conclusions:

- 'Unnamed: 0 is unwanted feature and it has to be deleted
- Since, the dataset is about Loan prediction,' msisdn' has 186243 unique values so it can be deleted
- The feature pcircle has only 1 value for all rows it doesn't contribute to the prediction, so it can be deleted.
- The feature pdate has data attribute so it has to be converted to day, date and month, year
- Date feature can be converted to day name, no of days and month name.
- After deleting the unwanted features and adding the date features, we get 36 columns and stored into a new dataframe 'df1'.
- Now we analyse the dataset again, we find that we have some negative values and extreme positive values, so these values are to be removed and kept as nan values
- We need to need to clean the dataset since the data is collected from different websites.

The features used in the dataset are:

- label
- aon
- daily decr30
- daily decr90
- rental30
- rental90
- last rech date ma
- last rech date da
- last rech amt ma
- cnt ma rech30
- fr ma rech30
- sumamnt ma rech30
- medianamnt ma rech30
- medianmarechprebal30
- cnt ma rech90
- fr ma rech90
- sumamnt ma rech90
- medianamnt ma rech90
- medianmarechprebal90
- cnt da rech30
- fr da rech30
- cnt da rech90
- fr da rech90
- cnt loans30
- amnt loans30
- maxamnt loans30
- medianamnt loans30
- cnt loans90
- amnt loans90
- maxamnt loans90
- medianamnt loans90
- payback30
- payback90
- Month
- Day name
- No of days

The categorical features were standardized and made uniform for all data

2.3 Data Preprocessing Done:

Exploratory Data Analysis (EDA):

From the dataset we find that some of the features have negative values and values greater than 100000, in which both are unrealistic. These values were removed and replaced as na.

Now we have to fill the missing values

Imputation of missing values:

A basic strategy to use incomplete datasets is to discard entire rows and/or columns containing missing values. However, this comes at the price of losing data which may be valuable (even though incomplete). A better strategy is to impute the missing values, i.e., to infer them from the known part of the data.

Univariate vs. Multivariate Imputation:

One type of imputation algorithm is univariate, which imputes values in the i-th feature dimension using only non-missing values in that feature dimension (e.g. impute.SimpleImputer). By contrast, multivariate imputation algorithms use the entire set of available feature dimensions to estimate the missing values (e.g. impute.IterativeImputer).

Univariate feature imputation:

The SimpleImputer class provides basic strategies for imputing missing values. Missing values can be imputed with a provided constant value, or using the statistics (mean, median or most frequent) of each column in which the missing values are located. This class also allows for different missing values encodings.

Multivariate feature imputation:

A more sophisticated approach is to use the IterativeImputer class, which models each feature with missing values as a function of other features, and uses that estimate for imputation. It does so in an iterated round-robin fashion: at each step, a feature column is designated as output y and the other feature columns are treated as inputs X. A regressor is fit on (X, y) for known y. Then, the regressor is used to predict the missing values of y. This is done for each feature in an iterative fashion, and then is repeated for max_iter imputation rounds. The results of the final imputation round are returned.

In a real world dataset, there will always be some data missing. This mainly associates with how the data was collected. Missing data plays an important role

creating a predictive model, because there are algorithms which does not perform very well with missing dataset.

Fancyimput

fancyimpute is a library for missing data imputation algorithms. Fancyimpute use machine learning algorithm to impute missing values. Fancyimpute uses all the column to impute the missing values. There are two ways missing data can be imputed using Fancyimpute

1. KNN or K-Nearest Neighbor
2. MICE or Multiple Imputation by Chained Equation
3. K-Nearest Neighbor

To fill out the missing values KNN finds out the similar data points among all the features. Then it took the average of all the points to fill in the missing values.

Multiple Imputation by Chained Equation:

MICE uses multiple imputation instead of single imputation which results in statistical uncertainty. MICE perform multiple regression over the sample data and take averages of them.

Note:

Missing values can be generally filled by mean or mode methods. But for large datasets this may be inaccurate. So for that we use imput techniques.

When compared to univariate methods, multivariate methods are much more accurate, because they include all the features while filling the missing values.

So we go along with MICE imputer.

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

Assumptions for data collections:

In this dataset, we assumed that the dataset containing negative values and extremely positive values are removed as they are not realistic.

2.5 Hardware and Software Requirements and Tools Used:

Hardware – PC Windows 10, 4 GB Ram

Software – Google chrome, MS Excel, Python, Selenium webdriver

Libraries – Pandas, NumPy, Matplotlib, Seaborn, sklearn, SciPy. Stats, DecisionTreeClassifier, accuracy_score, IterativeImputer

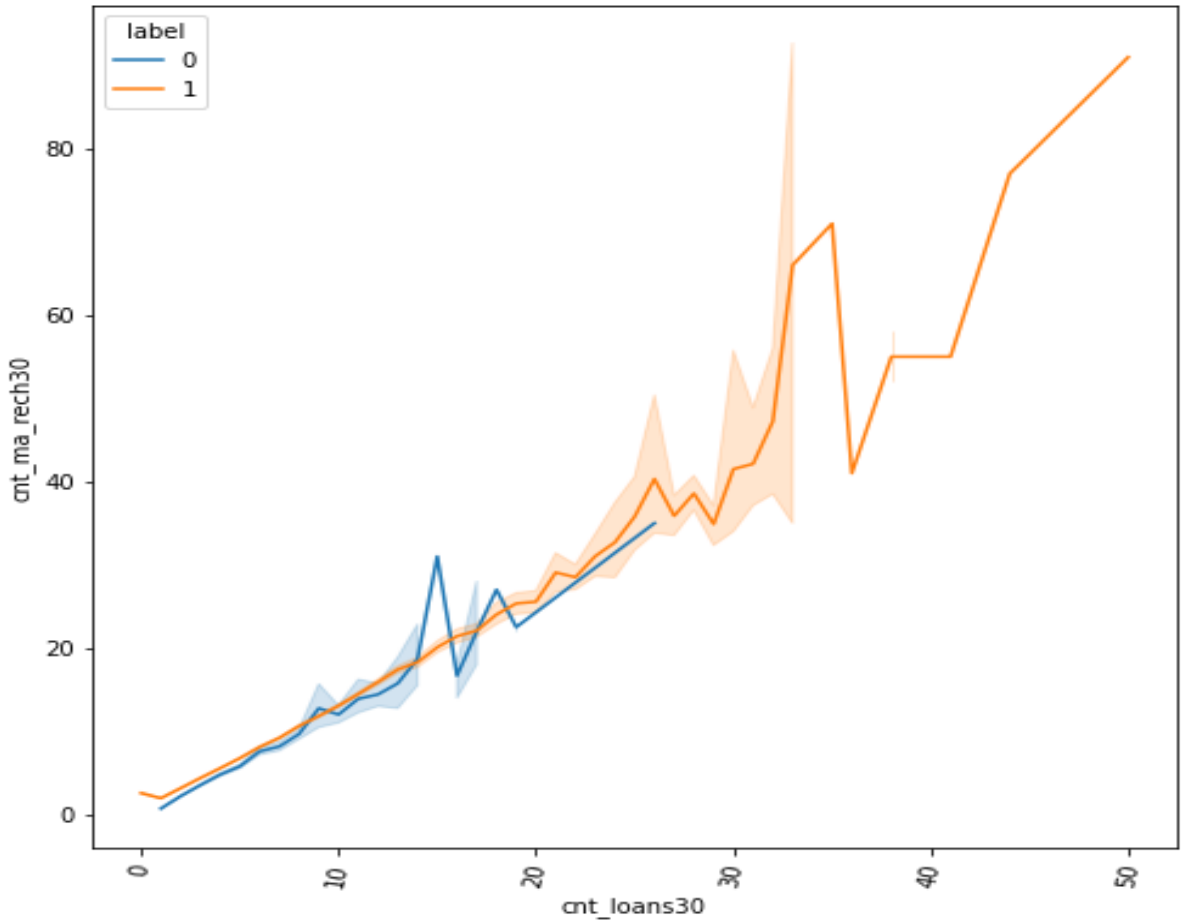
- Iterative Imputer– Python
- Data cleaning – Python, Pandas, NumPy & SciPy. Stats
- Data visualization – Matplotlib & Seaborn
- Machine learning – Sklearn

2.6 Data Inputs- Logic- Output Relationships:

Note: For data visualization we taken only the graphs referring to the loans. When comparing the features with the No of days we get the following conclusions:

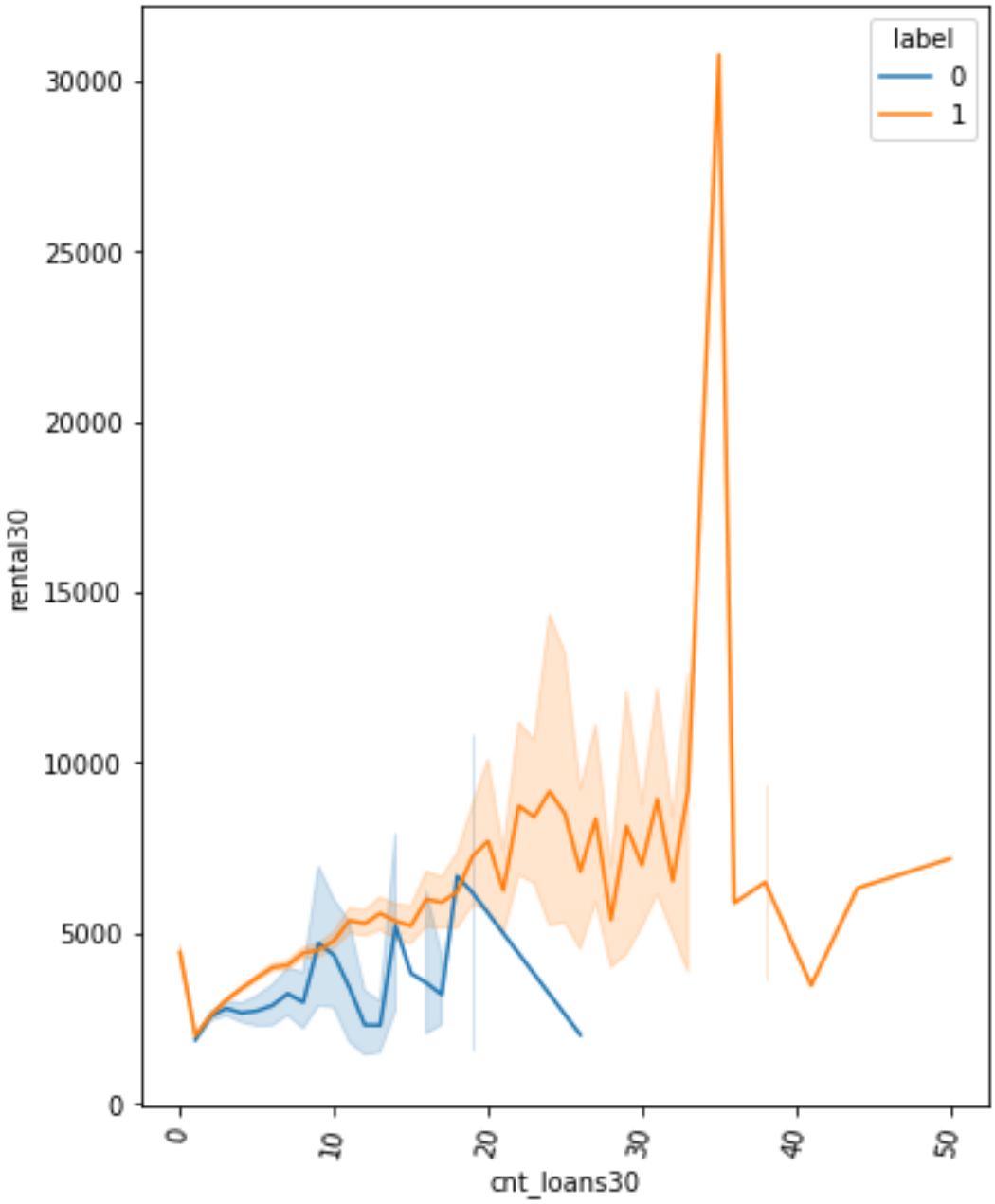
- For the cases whose loan is repaid: Almost in all the features, the features show an increase in trend as the no of days.
- For the cases whose loan is not repaid: Almost in all the features, the features show a similar trend but with lesser range
- As the no of days increases the loan repaid probability is also increased.

1. cnt_loans30 vs cnt_ma_rech30



For 30 days, the main account balance increases as no of loans increase. This is mainly because more people recharge 30 to 40 times take loans 20 to 30 times, and their loan was paid.
But for people who make less recharges take less loans and they didn't pay it.

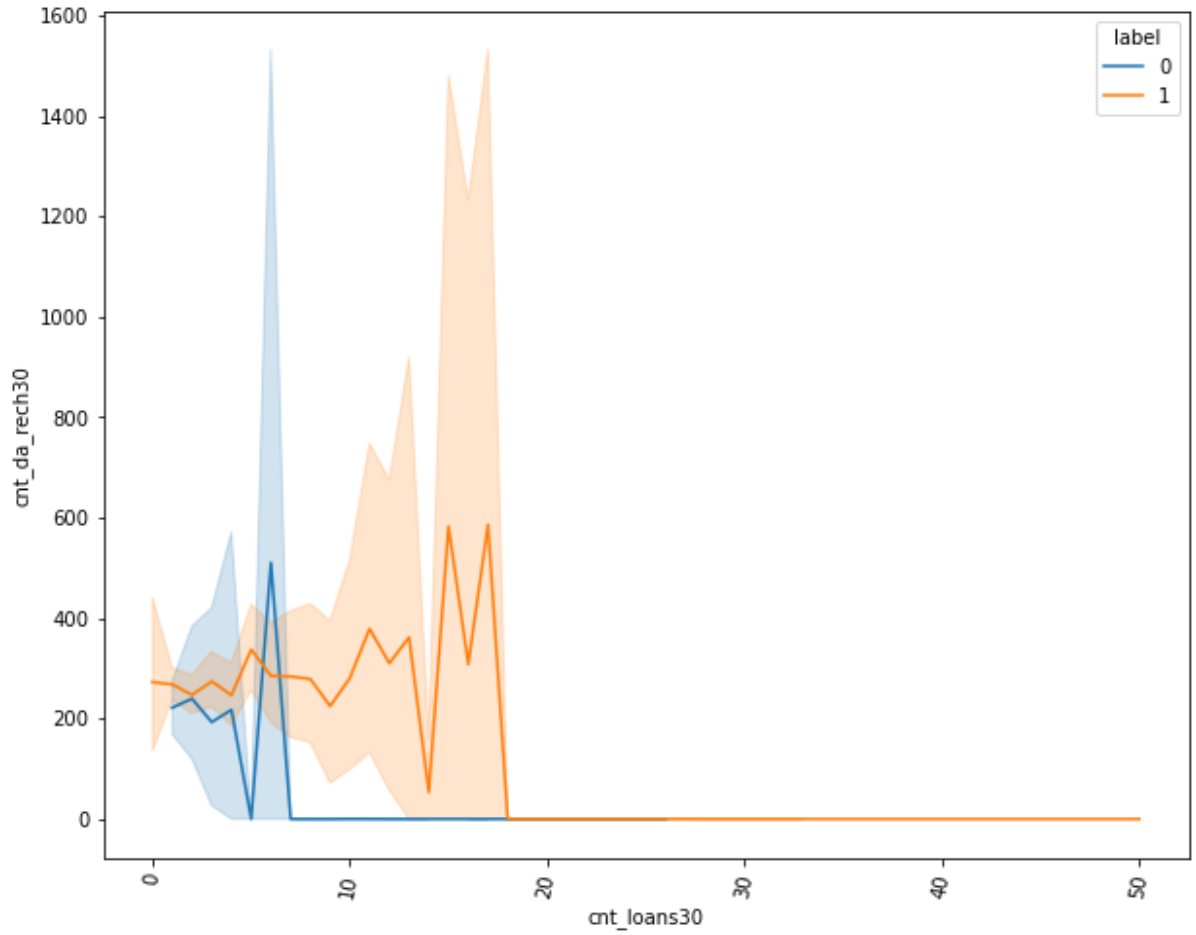
2. cnt_loans30 vs rental30:



For people as the average main balance increases, they take more loans between 20 to 30 times have average of 5000 to 10000 in their main account balance for the last 30 days.

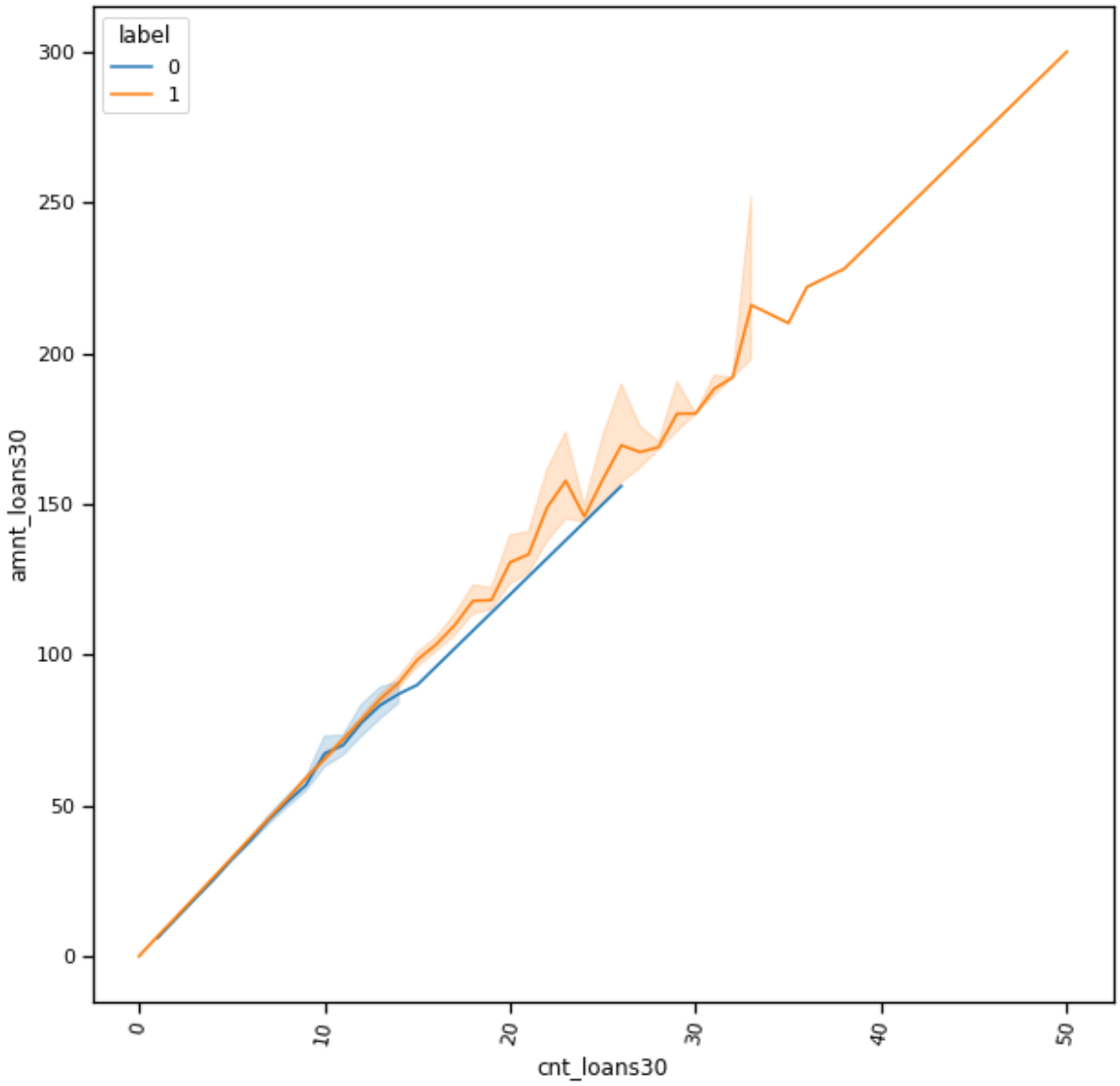
But for people who have less main account balance take mostly 10 loans and don't repay it.

3. cnt_loans30 vs cnt_da_rech30:



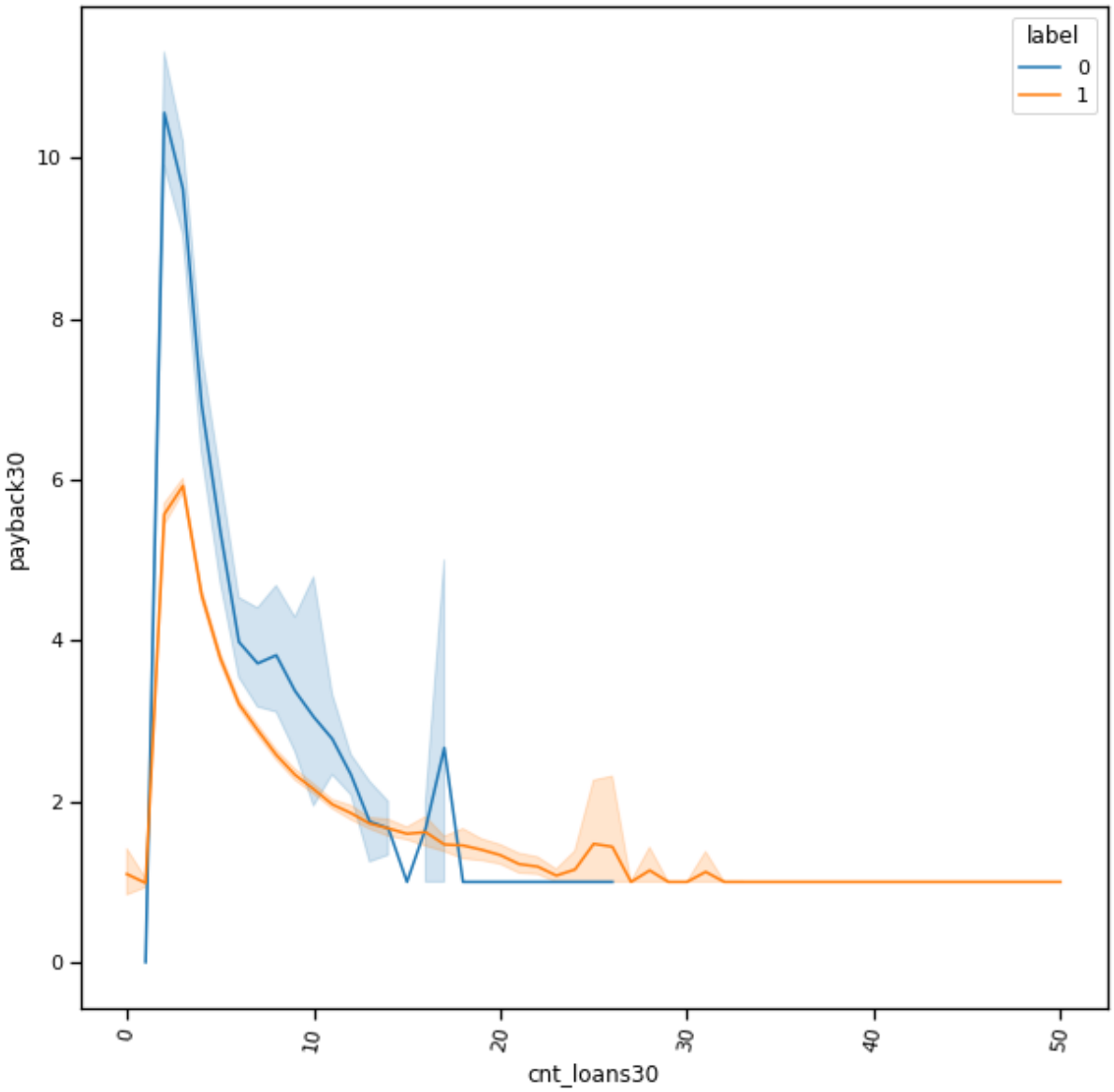
For people who recharged their data accounts for the last 30 days for 200 to 400 times approximately take 10 to 20 times loans and they repay it but those whose recharge less take less loan and not repay it.

4. cnt_loans30 vs amnt_loans30:



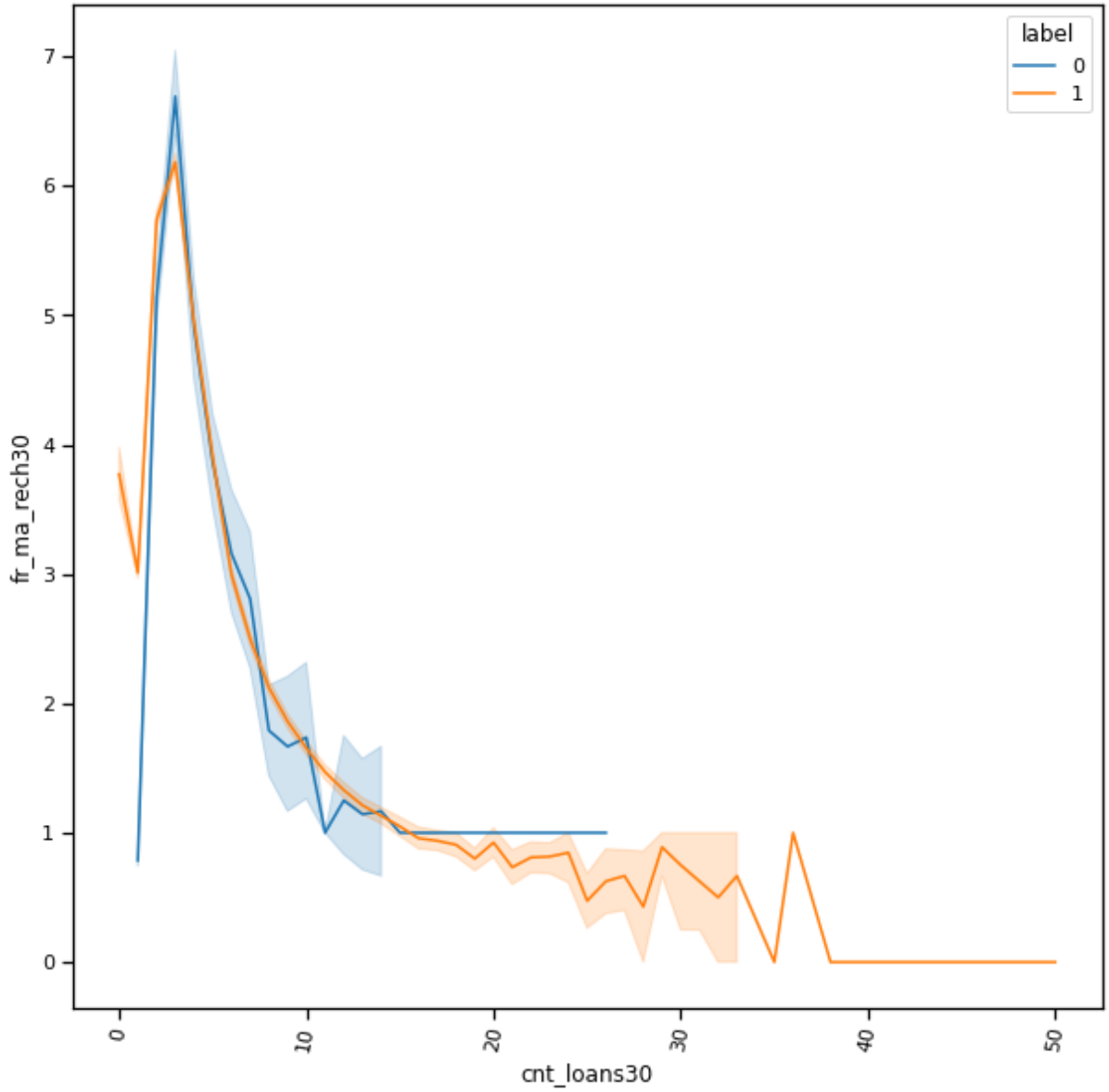
As count increases the amount also increase, and most people who take less loan amount of loan don’t repay.

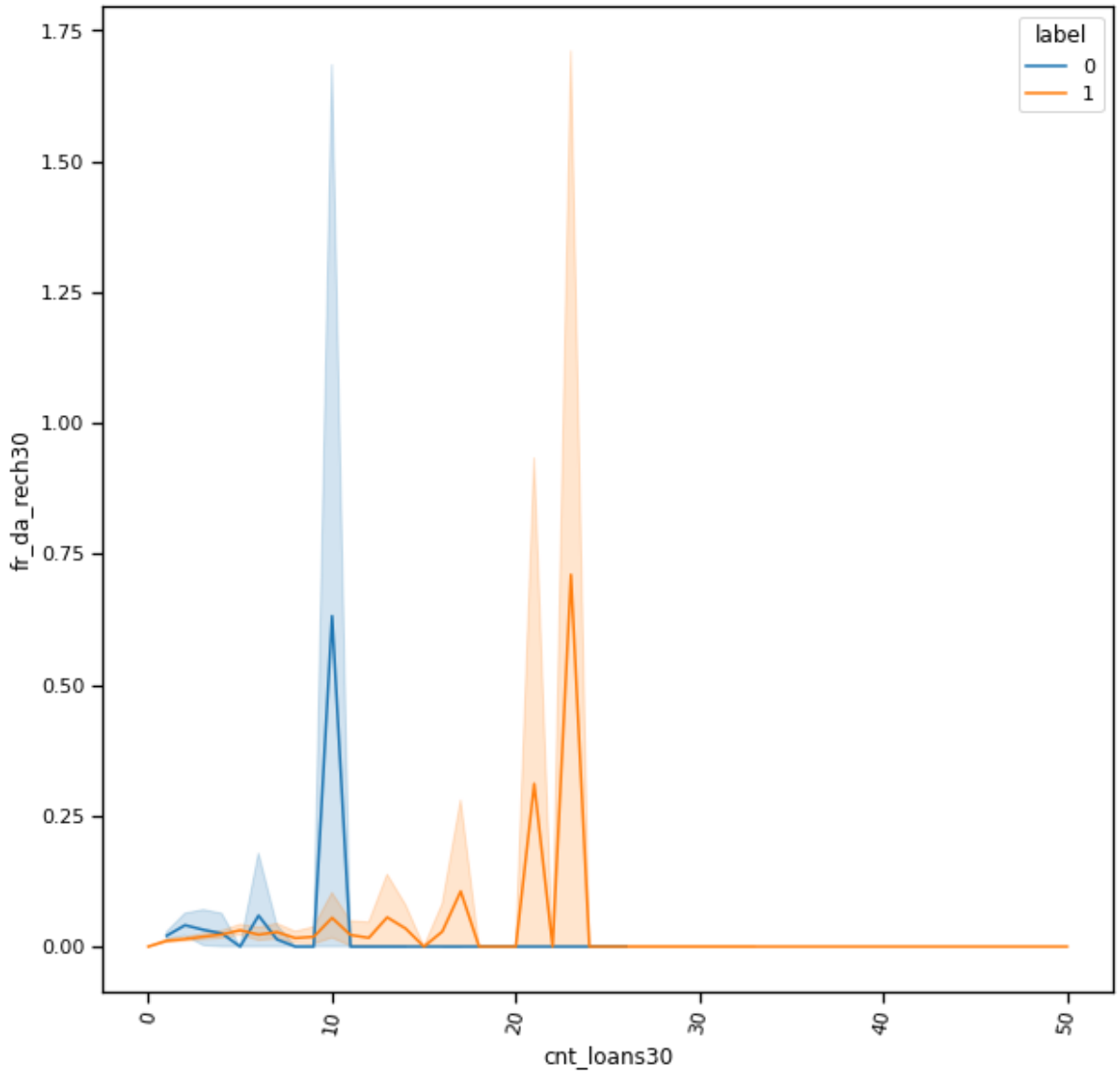
5. cnt_loans30 vs payback30:



Most people who take less no of loan payback time is more

6. fr_da_rech30 and fr_ma_rech30 vs count of loans:



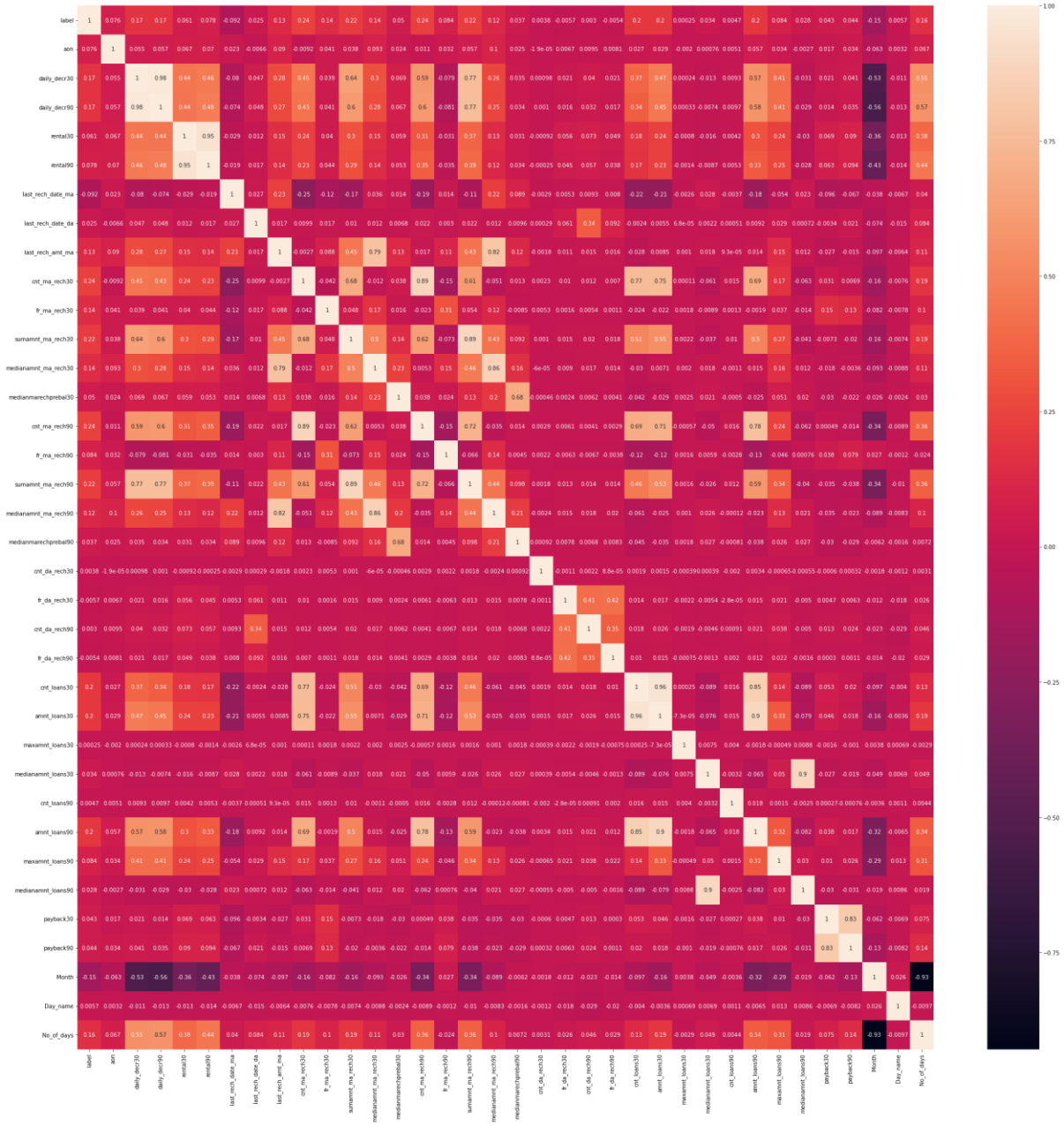

label
0
1
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
fr_da_rech30
0
10
20
30
40
50
cnt_loans30

S.No	Features	Customer who not paid (mean)	Customer who paid (mean)
1	label	0	1
2	aon	561.1006344	674.1321586
3	daily_decr30	1296.169239	5918.778737
4	daily_decr90	1302.77302	6639.236966
5	rental30	2070.1056	2875.80806
6	rental90	2377.271068	3763.293114
7	last_rech_date_ma	8.432867348	5.822581446
8	last_rech_date_da	0.470430108	1.001578151
9	last_rech_amt_ma	1237.04583	2182.462408
10	cnt_ma_rech30	1.30341717	4.359530287
11	fr_ma_rech30	1.836297292	4.189314617
12	sumamnt_ma_rech30	2249.502752	8366.892153
13	medianamnt_ma_rech30	1036.9419	1923.430522
14	medianmarechprebal30	50.533788	102.7526319
15	cnt_ma_rech90	1.812743674	6.957629844
16	fr_ma_rech90	4.903600642	8.118011677
17	sumamnt_ma_rech90	3168.42454	13015.55083
18	medianamnt_ma_rech90	1198.404633	1959.565804
19	medianmarechprebal90	56.55666305	98.99438708
20	cnt_da_rech30	220.1753306	268.6229699
21	fr_da_rech30	0.024584182	0.017105631
22	cnt_da_rech90	0.038338048	0.041944928
23	fr_da_rech90	0.059360905	0.043765776
24	cnt_loans30	1.431312591	2.948340248
25	amnt_loans30	8.873633514	19.24683396
26	maxamnt_loans30	271.8712637	275.053437
27	medianamnt_loans30	0.019876156	0.040974535
28	cnt_loans90	15.70120786	18.9202643
29	amnt_loans90	9.642382081	25.6425904
30	maxamnt_loans90	6.23438575	6.769989805
31	medianamnt_loans90	0.018538338	0.034503437
32	payback30	2.227276202	3.375558112
33	payback90	2.926037765	4.297735933
34	Month	1.506880208	1.159291505
35	Day_name	2.973358306	3.007877622
36	No_of_days	28.04001988	38.86368716

From the two graphs we find that people who take more frequent loan payback better.

Once preprocessing is completed, the categorical data is converted to continuous data using label encoding.

Now the dataframe is checked for correlation and heatmap is shown below:



Now the dataset is checked for presence of outliers. Here we use two methods

1. IQR Method

From this method we have more rows of outliers which may collapse the dataset

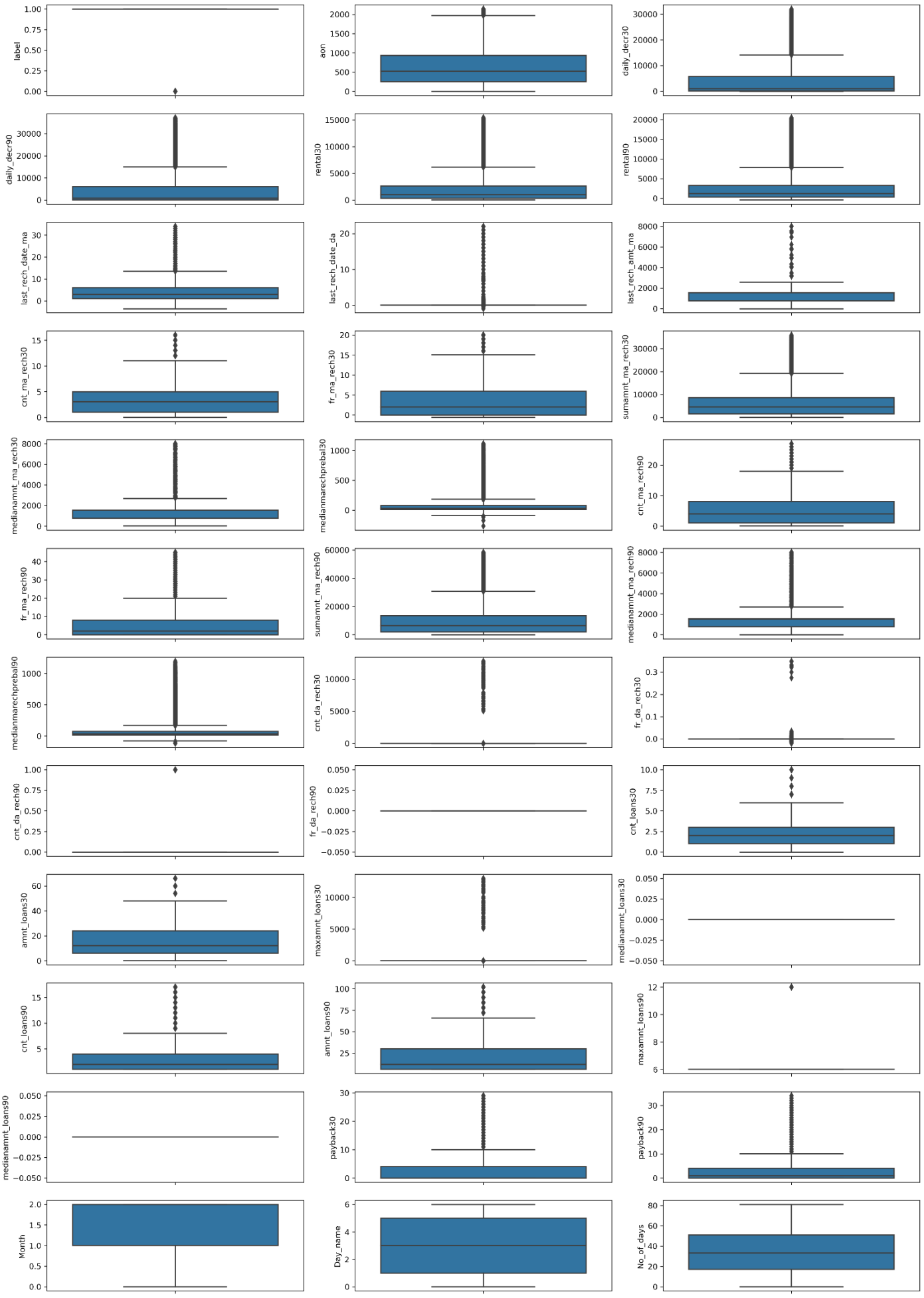
2. Z-Score Method

Using Z-Score method we detected less outliers

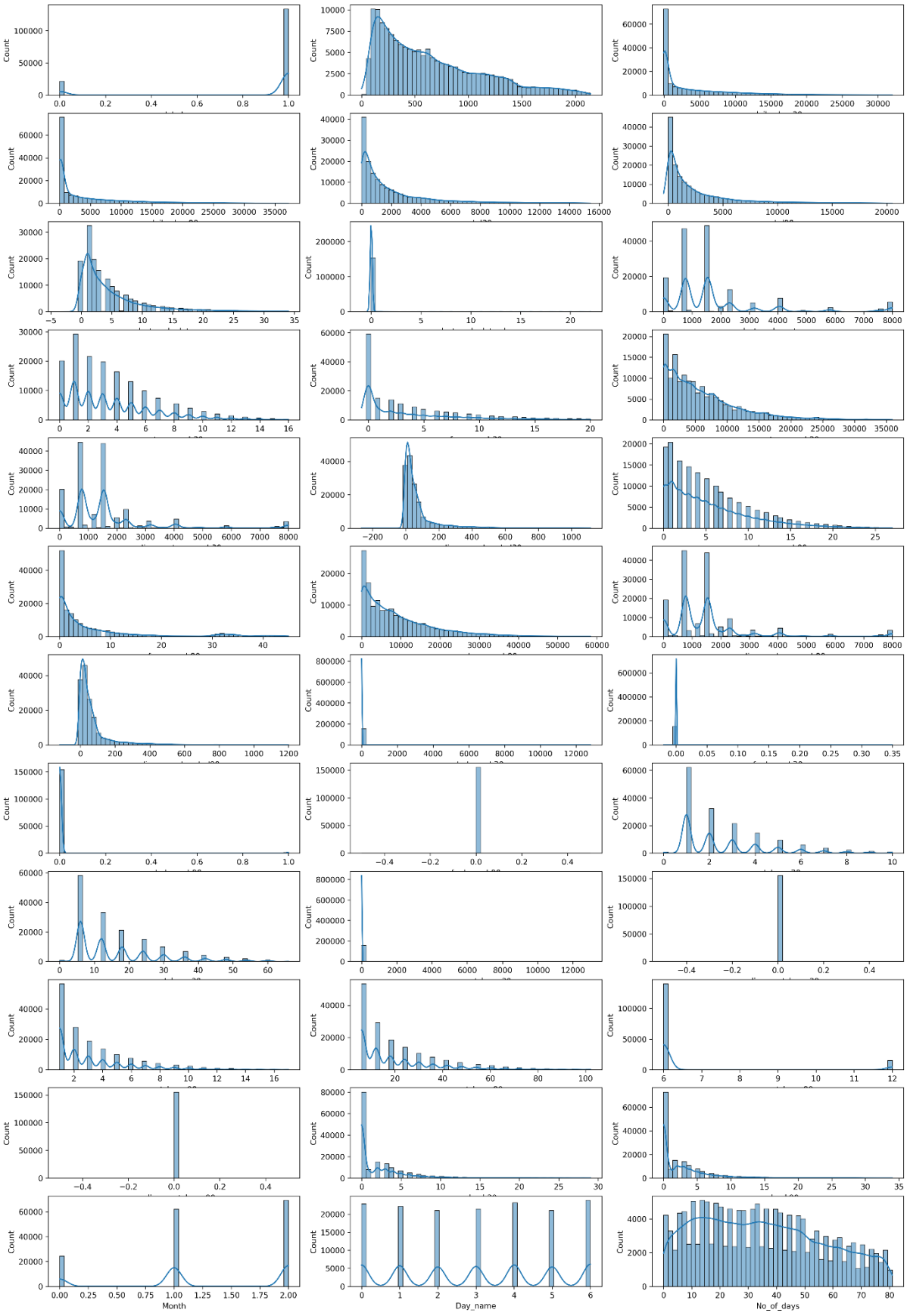
From the Z-Score method’s dataframe is used further.

The new dataframe has the shape of 155246rows and 36 columns

After removing outliers, we plotted the box-plot


label
aon
daily_decr30
daily_decr90
rental30
rental90
last_rech_date_ma
last_rech_date_da
last_rech_amt_ma
cnt_ma_rech30
fr_ma_rech30
sumamnt_ma_rech30
medianamnt_ma_rech30
medianmarechprebal30
cnt_ma_rech90
fr_ma_rech90
sumamnt_ma_rech90
medianamnt_ma_rech90
medianmarechprebal90
cnt_da_rech30
fr_da_rech30
cnt_da_rech90
fr_da_rech90
cnt_loans30
amnt_loans30
maxamnt_loans30
medianamnt_loans30
cnt_loans90
amnt_loans90
maxamnt_loans90
medianamnt_loans90
payback30
payback90
Month
Day_name
No_of_days

Skewness:



Histogram:

A histogram is plot to check whether the features are normally distributed or not.

CHAPTER III

Model/s Development and Evaluation

3.1 Identification of possible problem-solving approaches (methods)

Basic Parameters:

1. Standardization:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

2. Train Test data:

Train/Test is a method to measure the accuracy of your model. It is called Train/Test because we split the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing. We train the model using the training set. We test the model using the testing set.

3. Linear regression

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.

5. Random Forest Regressor

Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees. It executes by constructing a different number of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

3.2 Testing of Identified Approaches (Algorithms):

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

- Logistic Regression
- Decision Tree Classifier
- GaussianNB
- Decision Tree Classifier
- KNeighbors Classifier

3.3 Run and evaluate selected models:

From the above, the model is scaled using standard scaler, looped with the above methods and best model is obtained.

From this the best model is DecisionTreeClassifier from the Random state 44. Now this model is hypertuned and best parameters is obtained

3.4 Key Metrics for success in solving problem under consideration:

Accuracy Parameter:

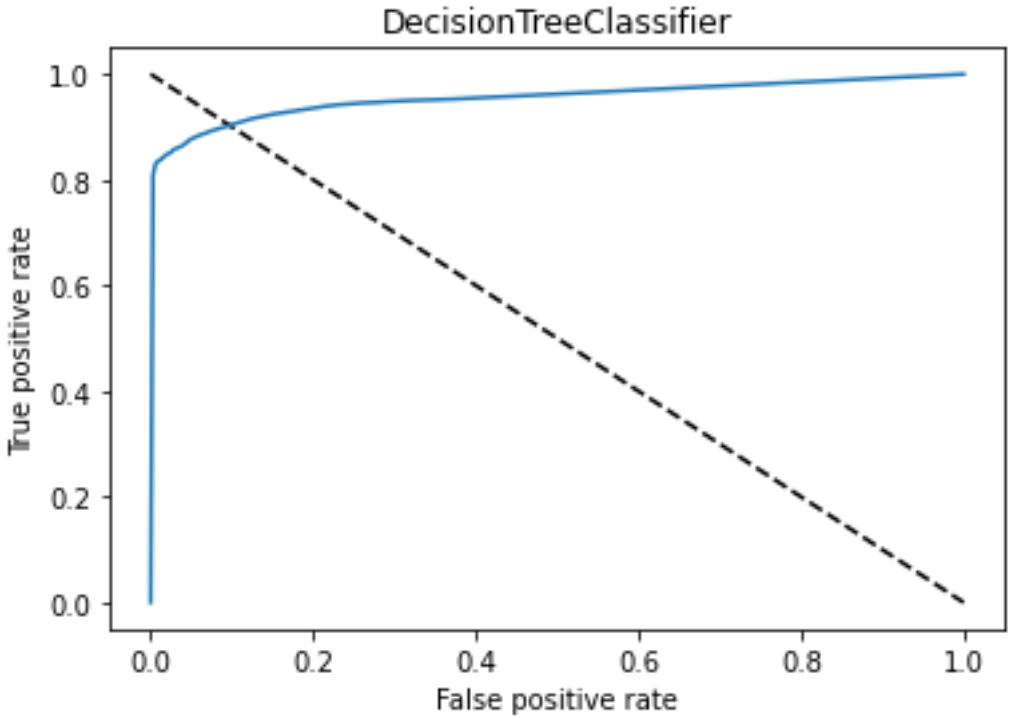
Accuracy Score: 91.39887113011413

The best model is obtained by Hypertuning the existing models.

3.5 Visualizations:

We get almost a linear plot

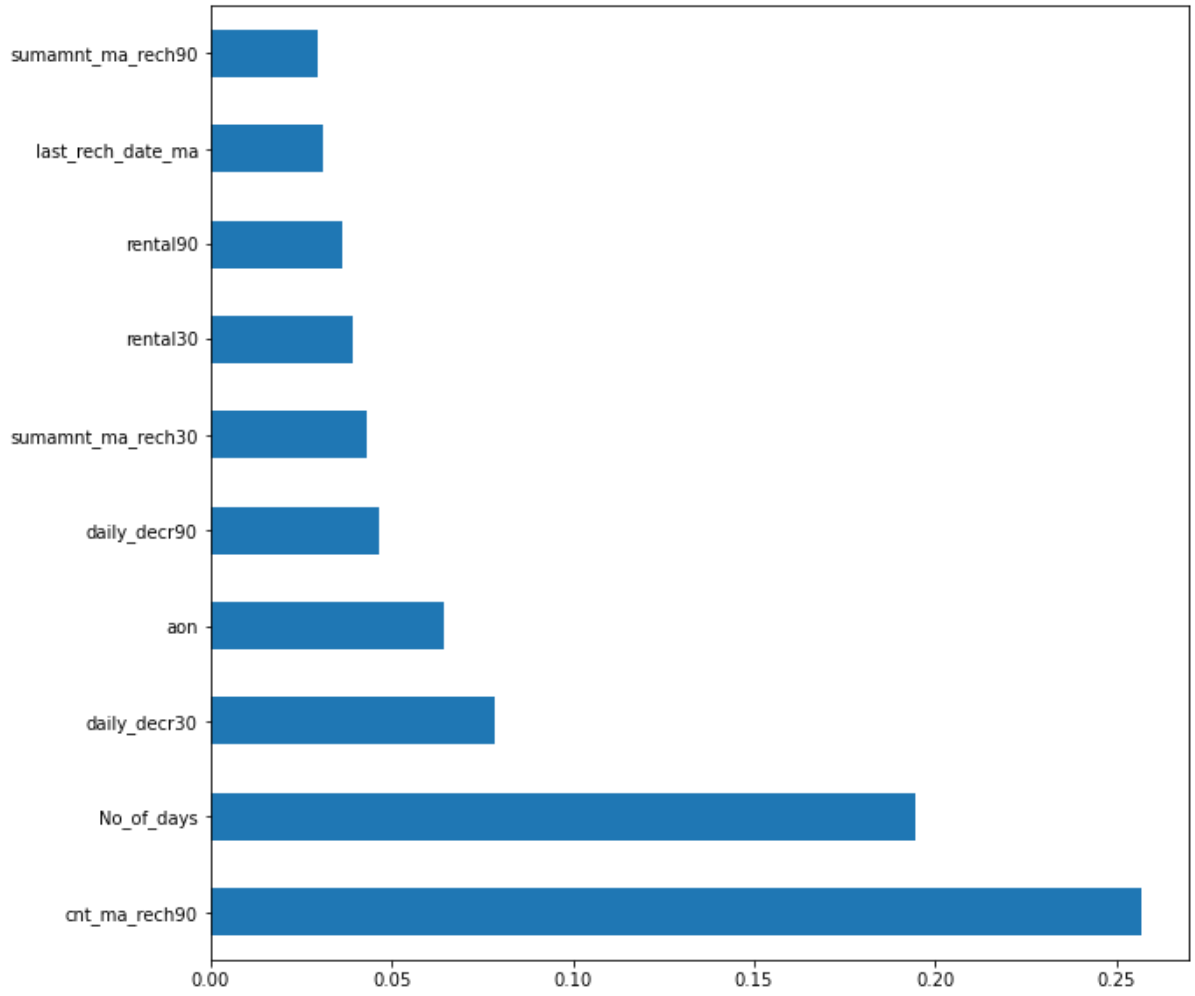
Distribution of Predicted Values:



confusion matrix:

```
[[42215,  2126],
       [ 5478, 38588]
```

Feature Importance:



Regarding the feature importance's, cnt_ma_rech90 dominates the loan predicton more.

3.6 Interpretation of the Results

- From the results, we find that this problem can be solved by a classification method and loan repay status can be predicted.
- The skewness is not reduced because it causes more damage to the data(more na)
- cnt_ma_rech90 dominates the loan prediction more.

CHAPTER IV
CONCLUSION

4.1 Key Findings and Conclusions of the Study:

- This dataset has been cleaned for unrealistic data such as negative and extreme positive values
- Since the target feature is categorical data, this problem can be solved by classification algorithms
- Decision tree algorithm gives an accuracy score 91.39
- cnt_ma_rech90 dominates the loan prediction more.

4.2 Learning Outcomes of the Study in respect of Data Science:

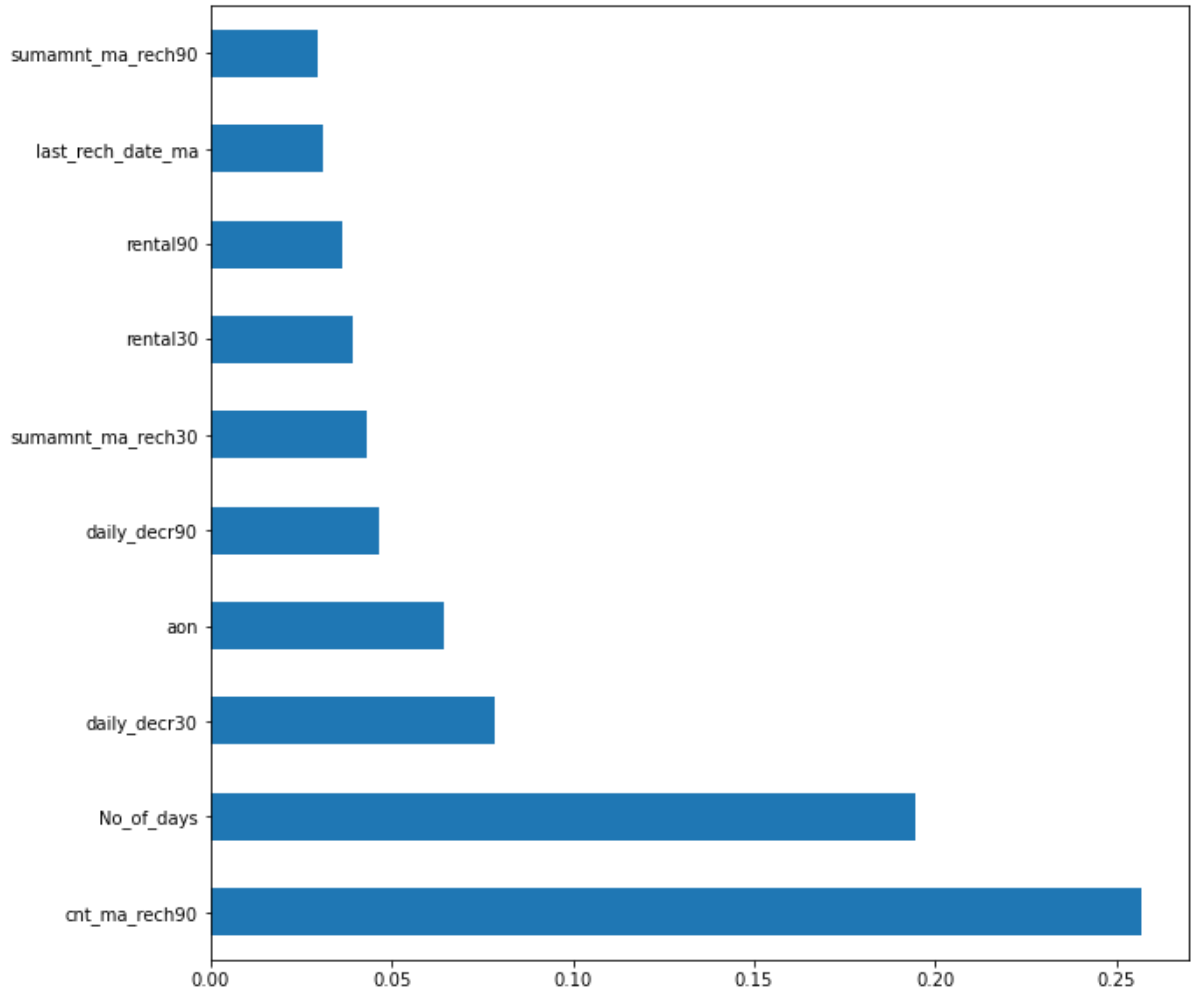
- It gives a deep learning of Data cleaning
- It emphasizes the importance of unrealistic data
- Data uniformity and the importance of features required
- How data collection affects the results
- Role of data cleaning, feature selection, etc.

4.3 Limitations of this work and Scope for Future Work:

- Here unrealistic values only removed, this dataset contains mostly zero which will be evaluated in feature
- Skewness have not been reduced

Data Analysis:

Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been payed i.e. Non- defaulter, while, Label '0' indicates that the loan has not been payed i.e. defaulter.



In general, from the dataset, we find that the customers who takes more loans will repay more.

From the feature importance, we find cnt_ma_rech90, dominates more

Loan has to be given based on the aoc, daily_decr30, sumamnt_ma_rech90

For people whose main account got recharged a greater number of times in last 90 days will get loan and repay it.

People whose Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah) is more, they have more possibility of getting and repaying loans more.

Also depends upon the Age on number.

For the features which have greater feature importances will have more control on the target variable, so the top features will predict whether the customer will be paying back the loaned amount within 5 days of insurance of loan