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

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.

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
- 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.

Dataset Description:

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

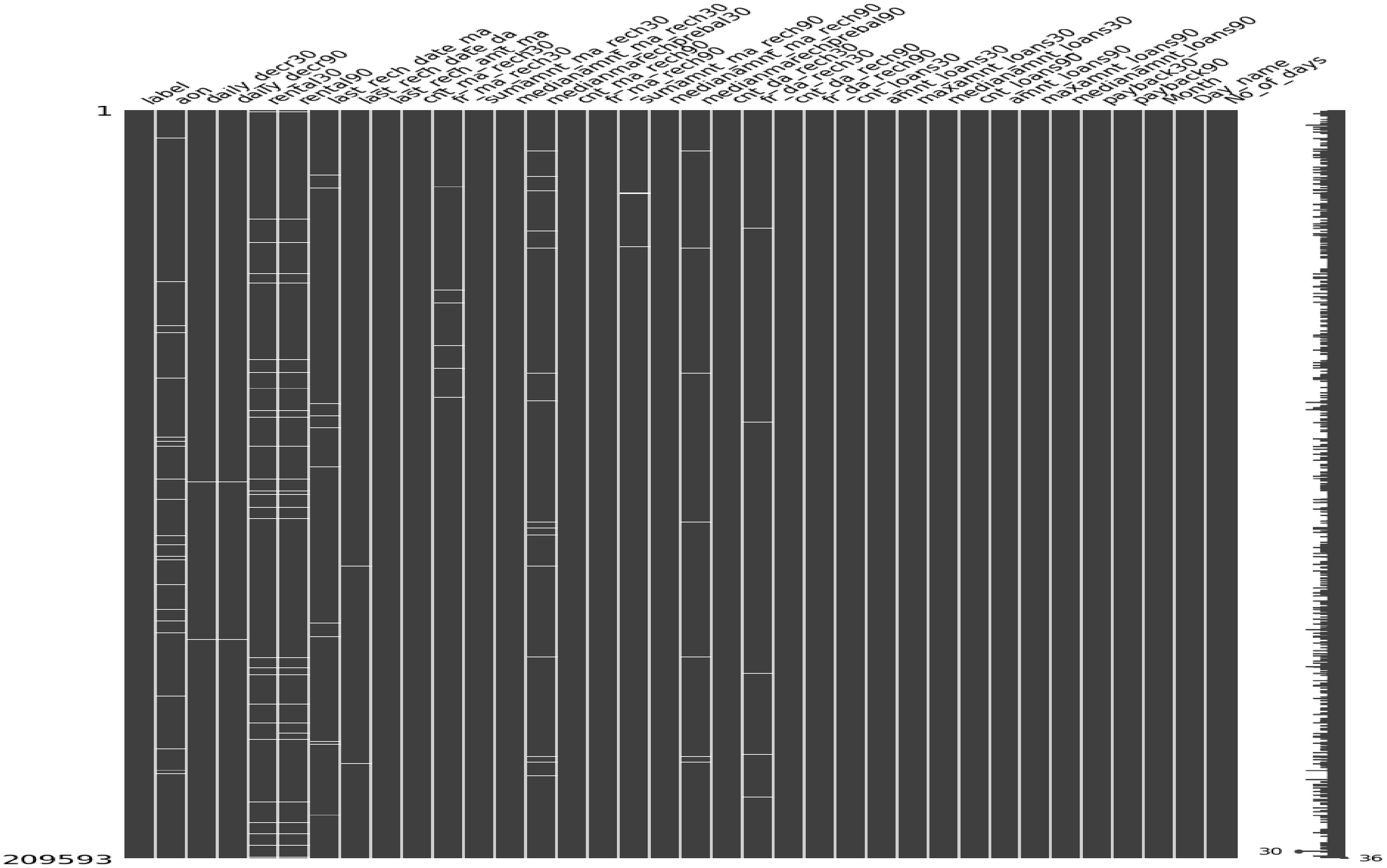
Cont'd…

S.No	Features	Customer who not paid (mean)	Customer who paid (mean)
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

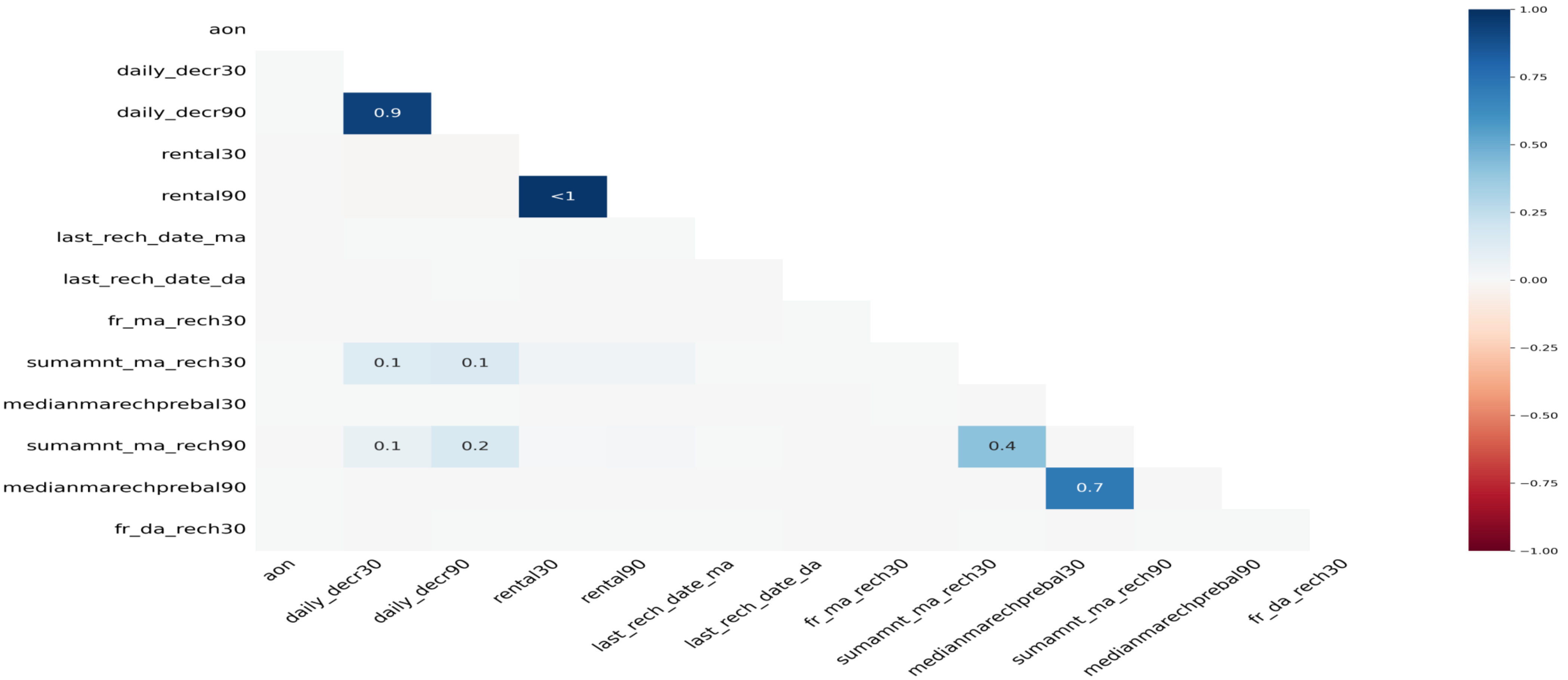
Hardware and Software Requirements and Tools Used:

- 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

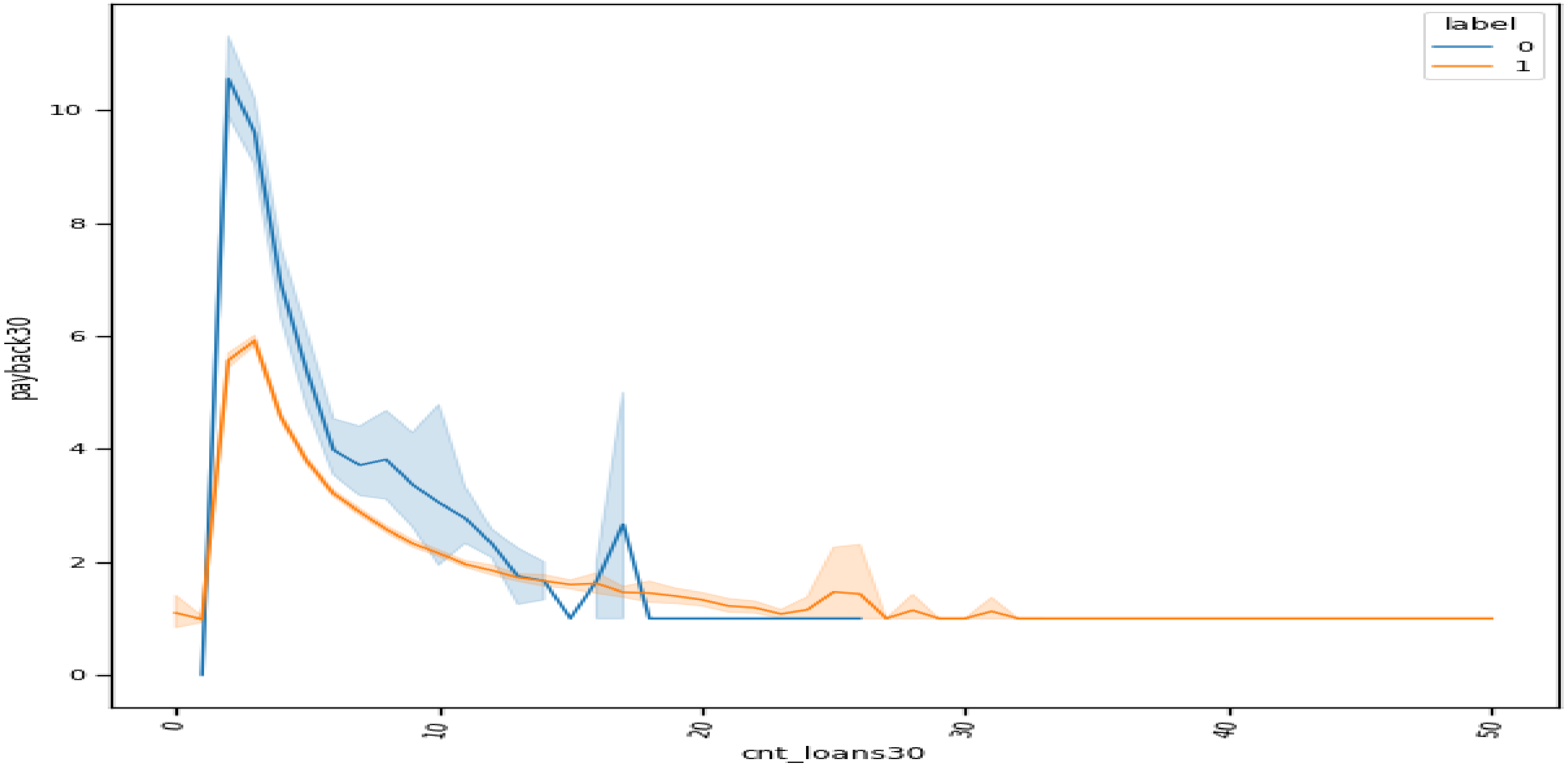
Randomness of removing unrealistic data:



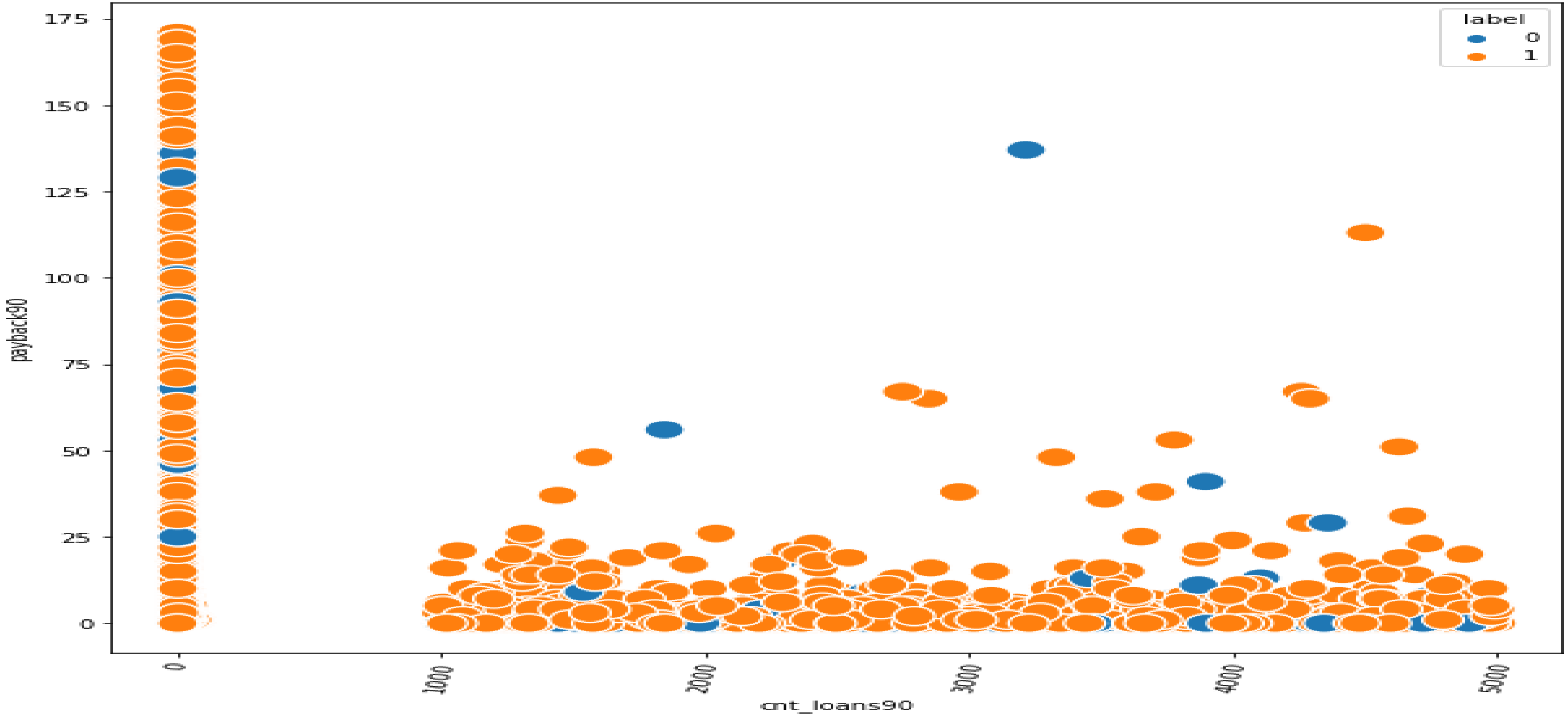
Heatmap of Data After Removing Unrealistic data



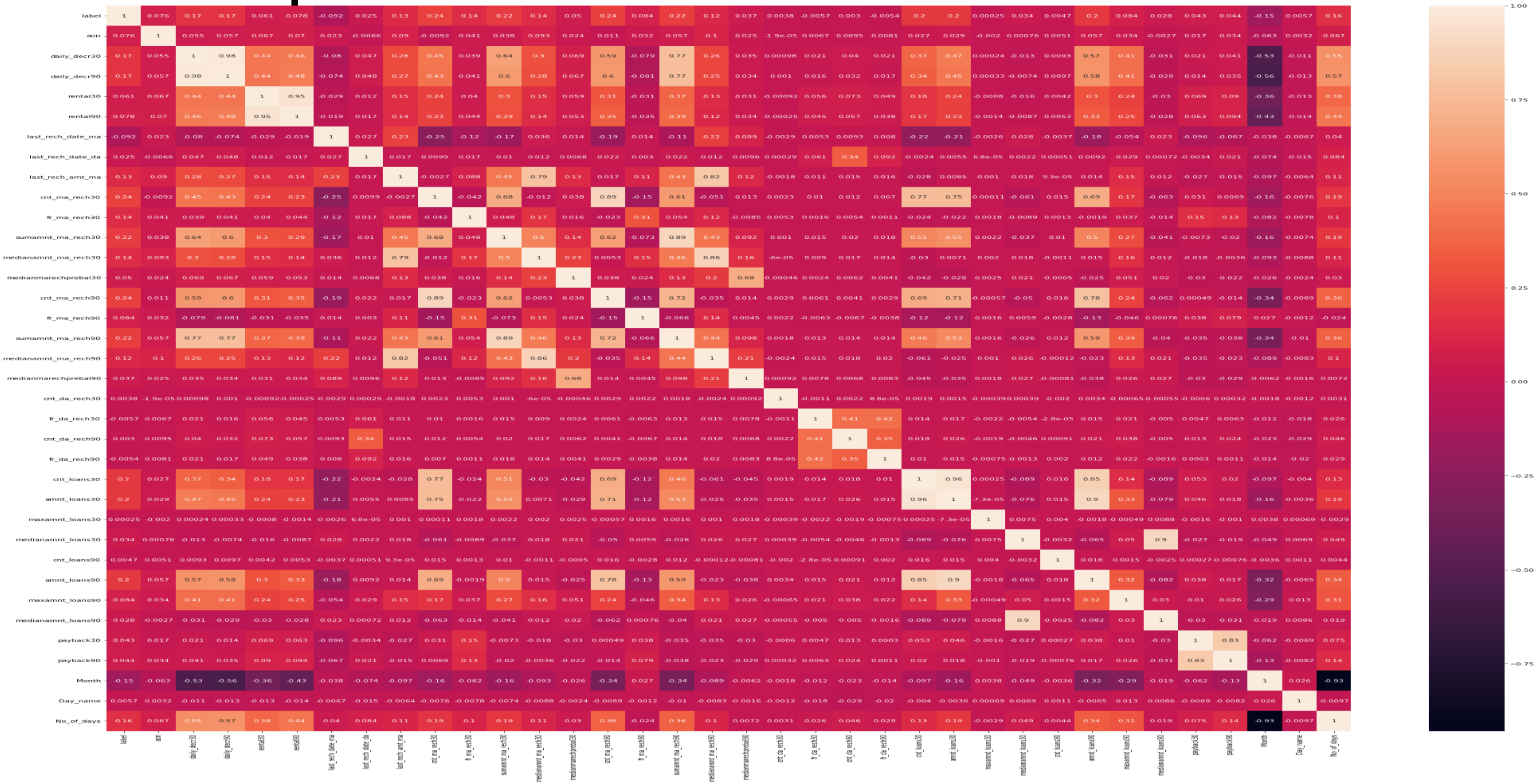
Cnt_loans_30 vs payback30



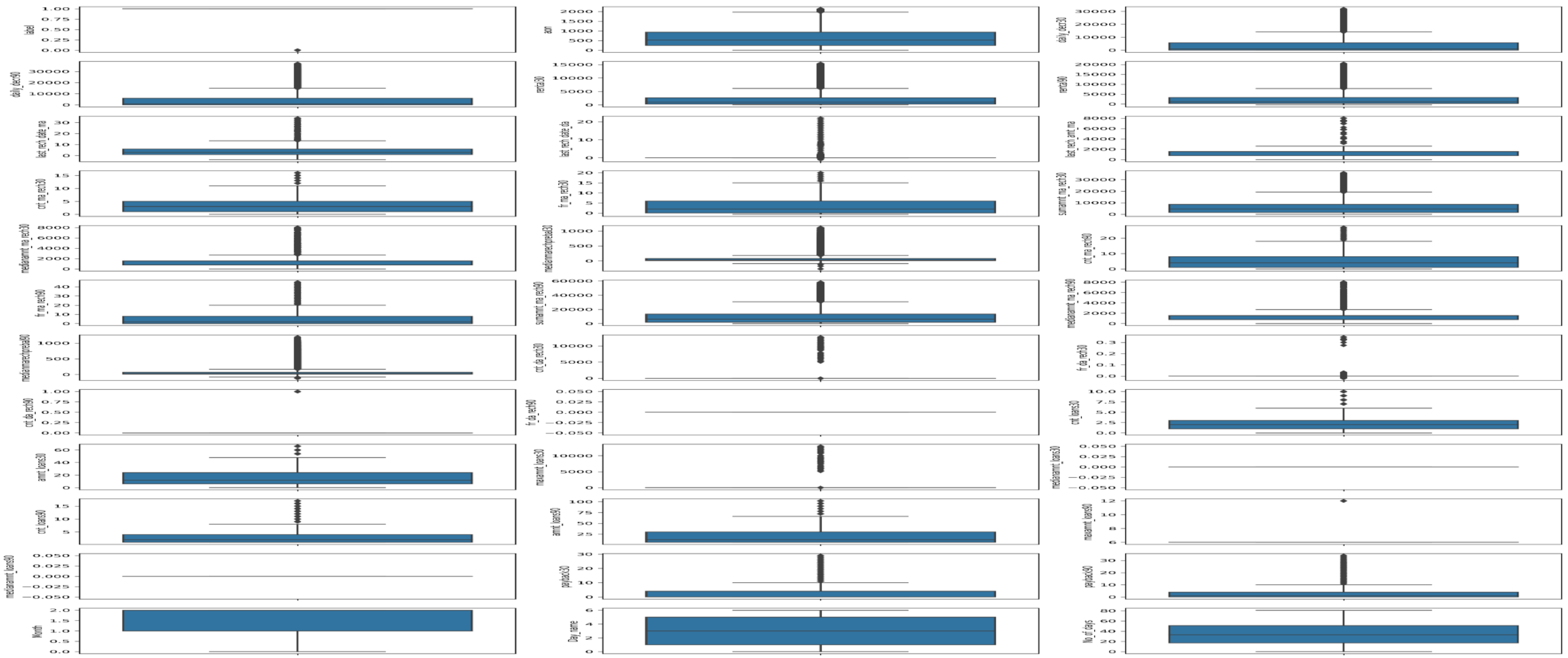
Cnt_loans90 vs payback90



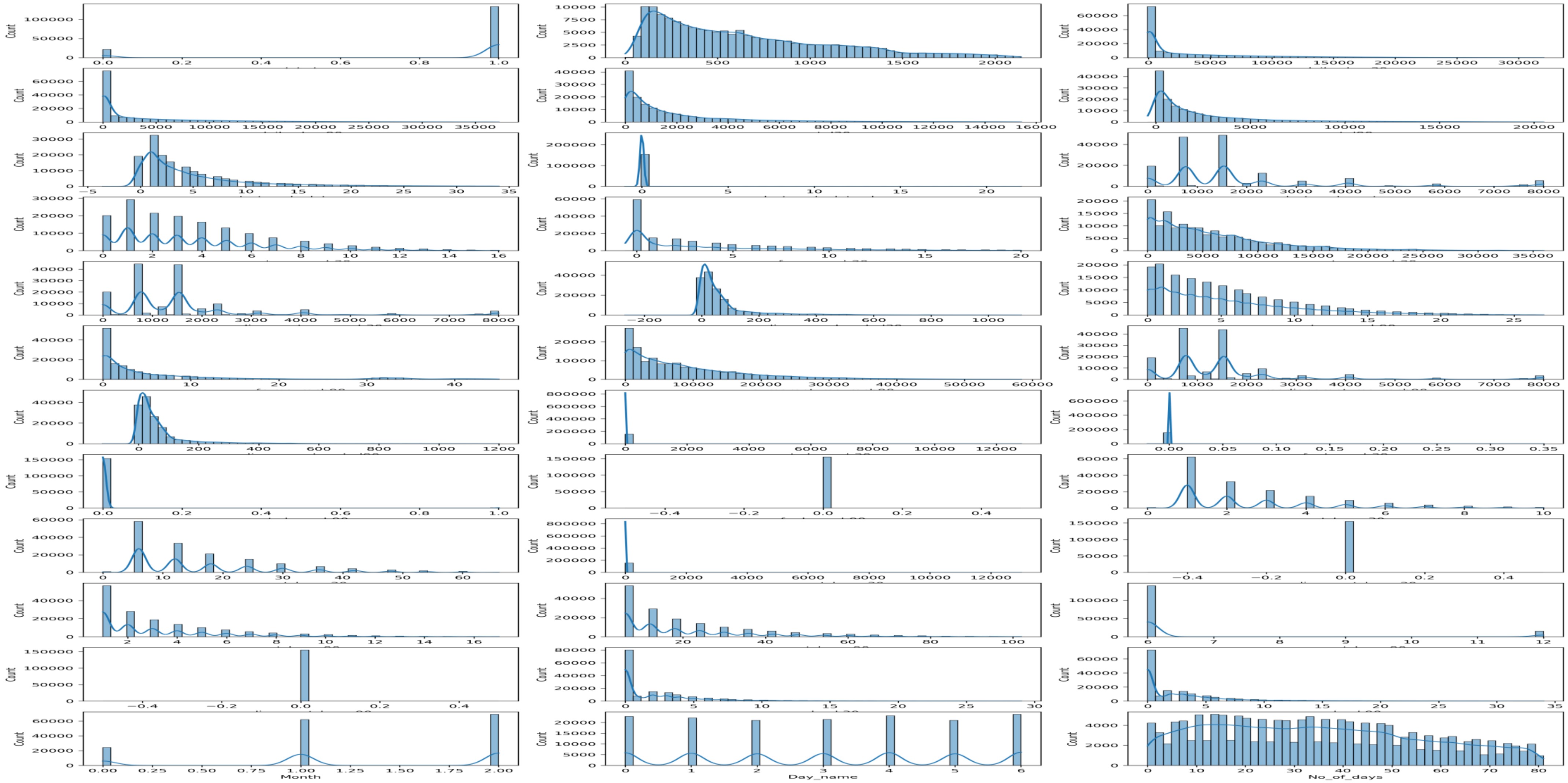
Heatmap:


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
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Boxplot:



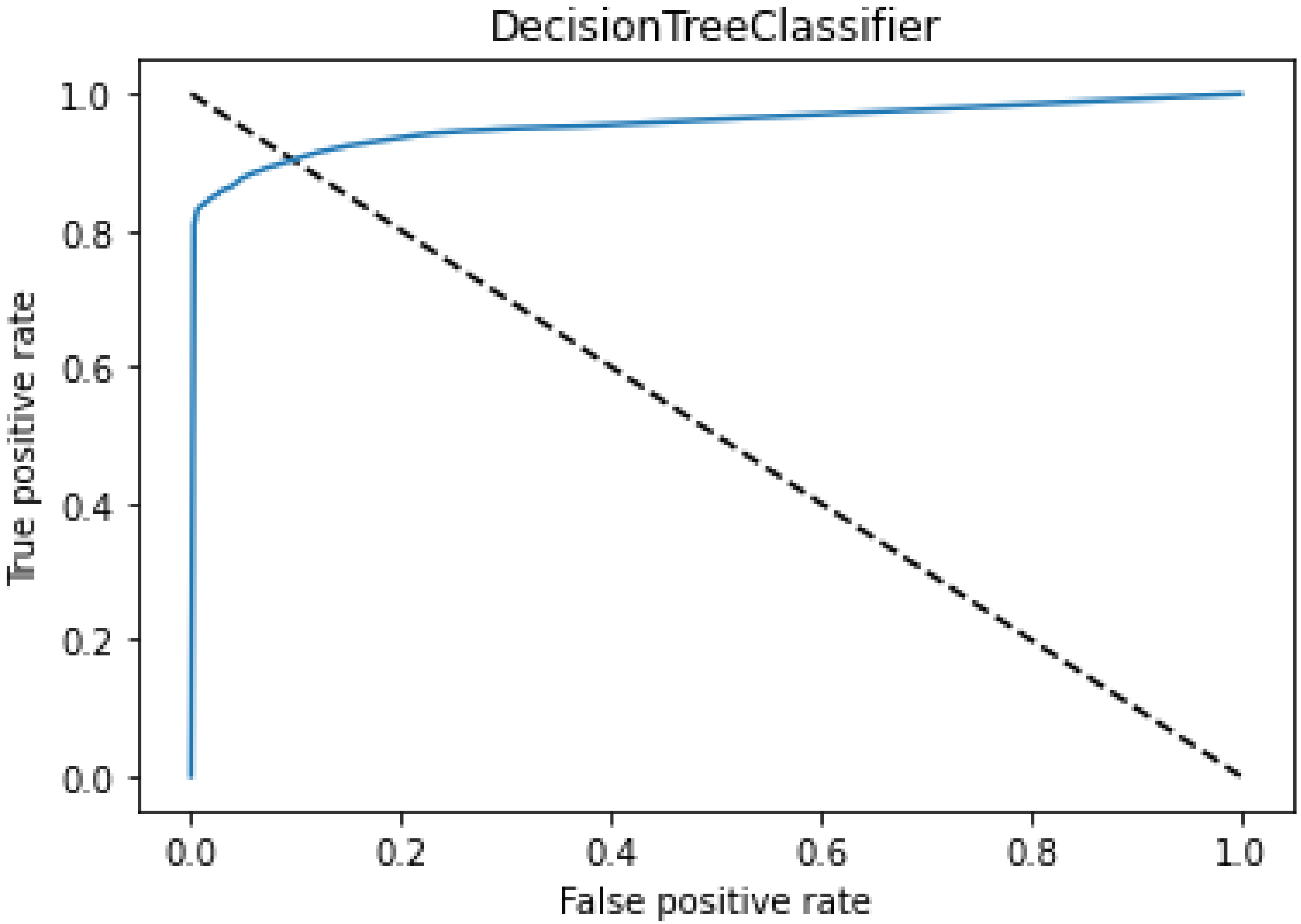
Histogram:



Testing of Identified Approaches (Algorithms):

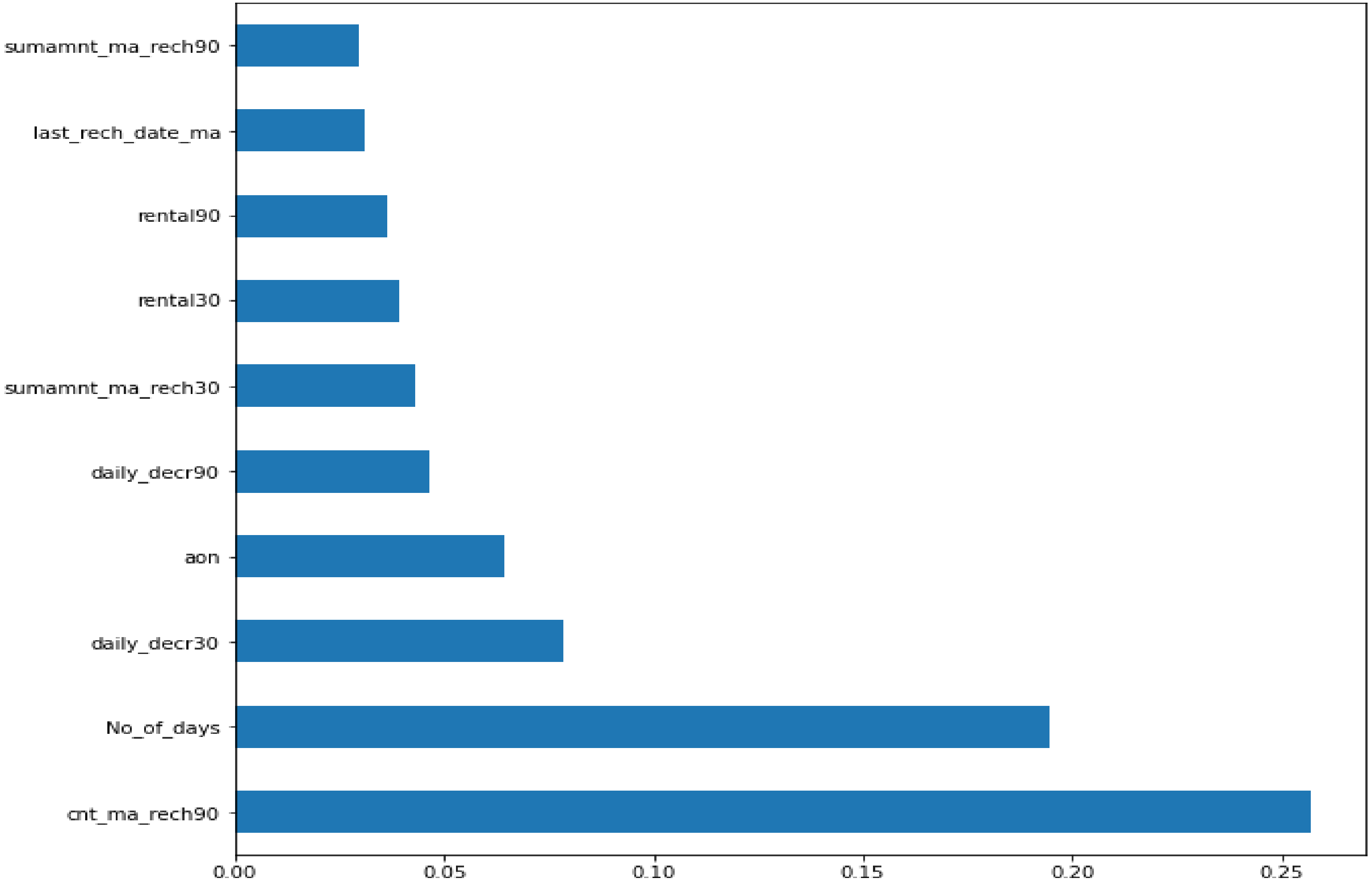
- Logistic Regression
- Decision Tree Classifier
- GaussianNB
- Decision Tree Classifier
- KNeighbors Classifier

Y test plot:



DecisionTreeClassifier
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Feature importance’s



Accuracy Parameter:

- Accuracy score: 91.38

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.