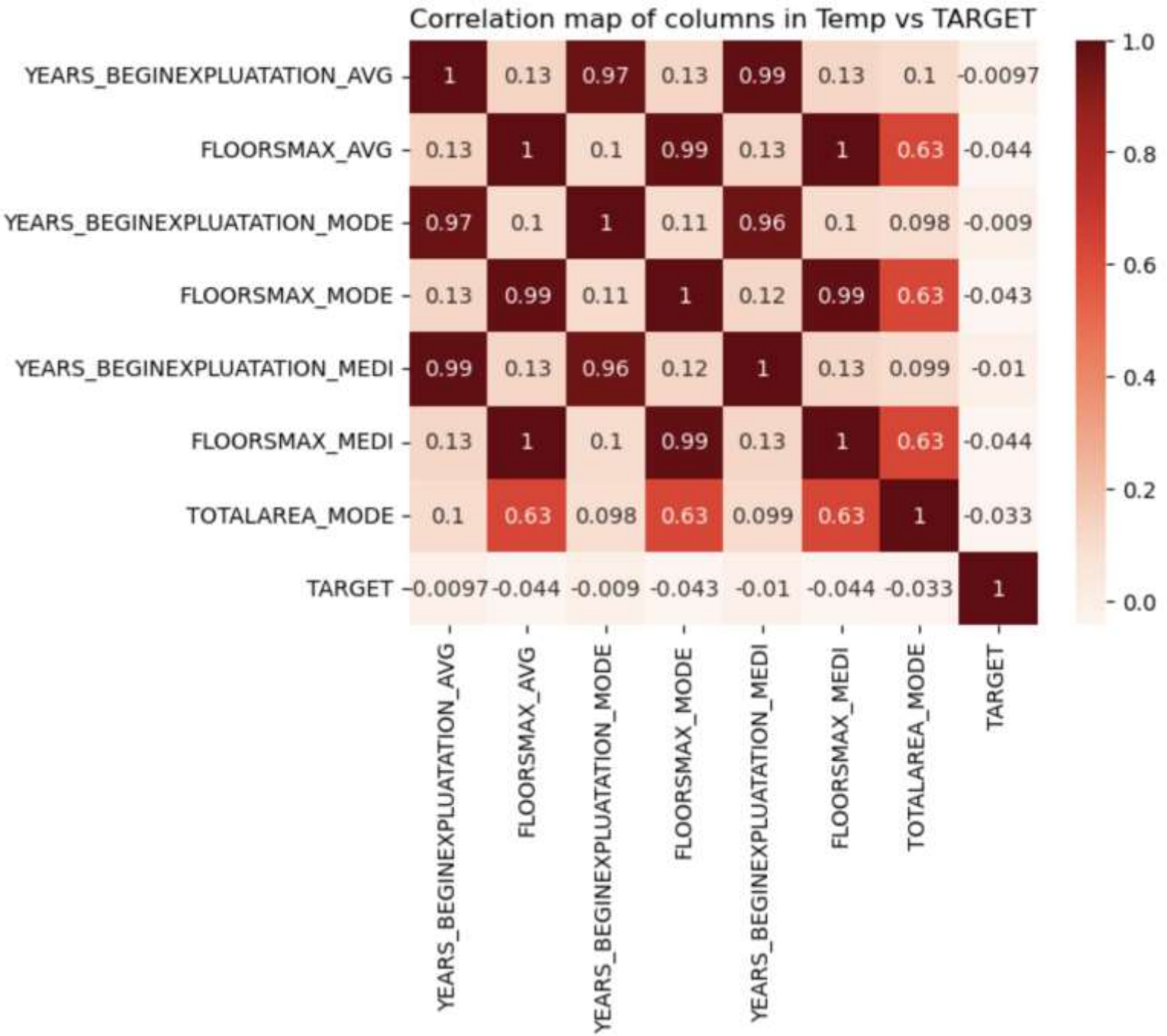
Credit EDA Assignment– Plots Presentation

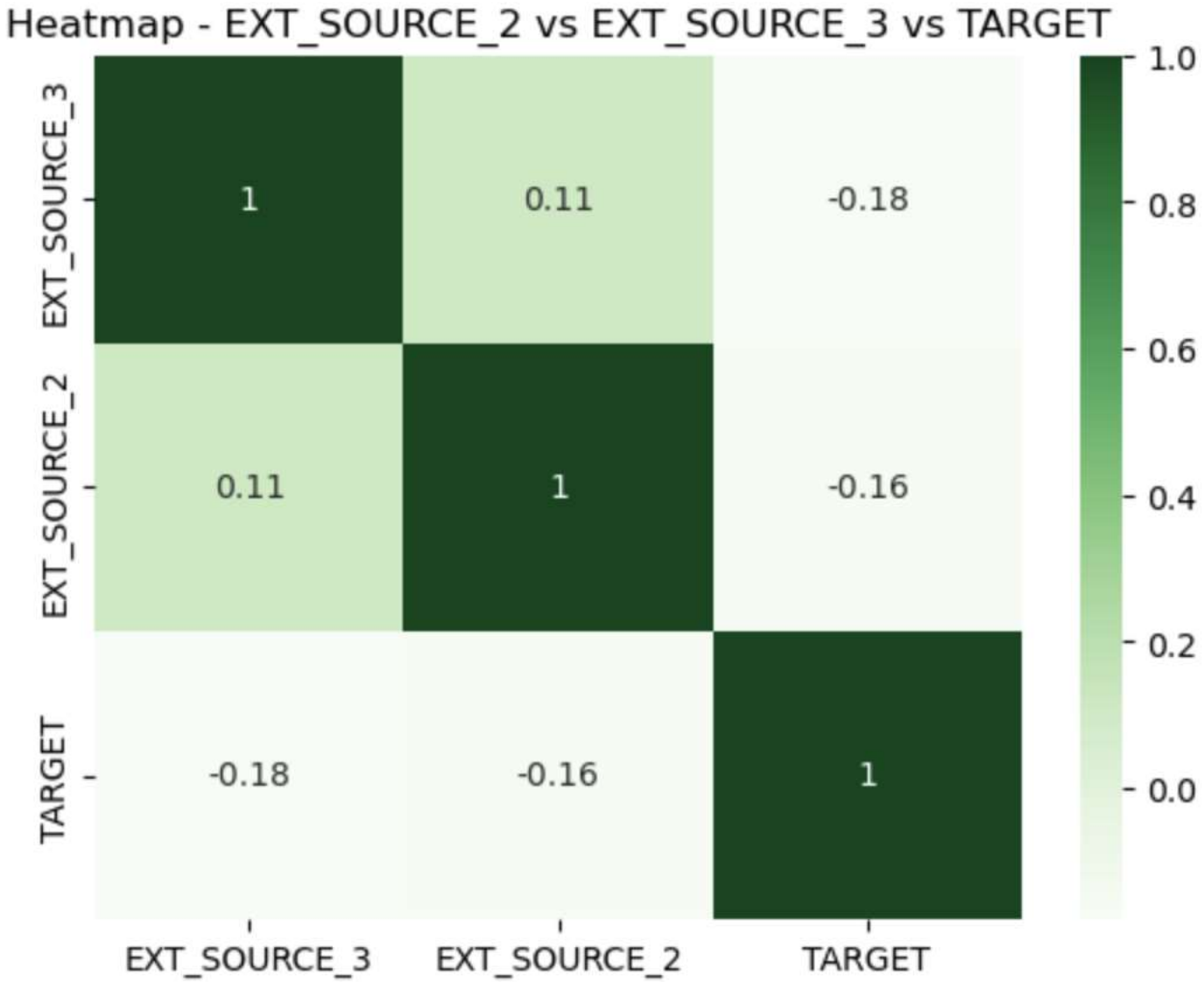
- by

Akella Venkata Koushik

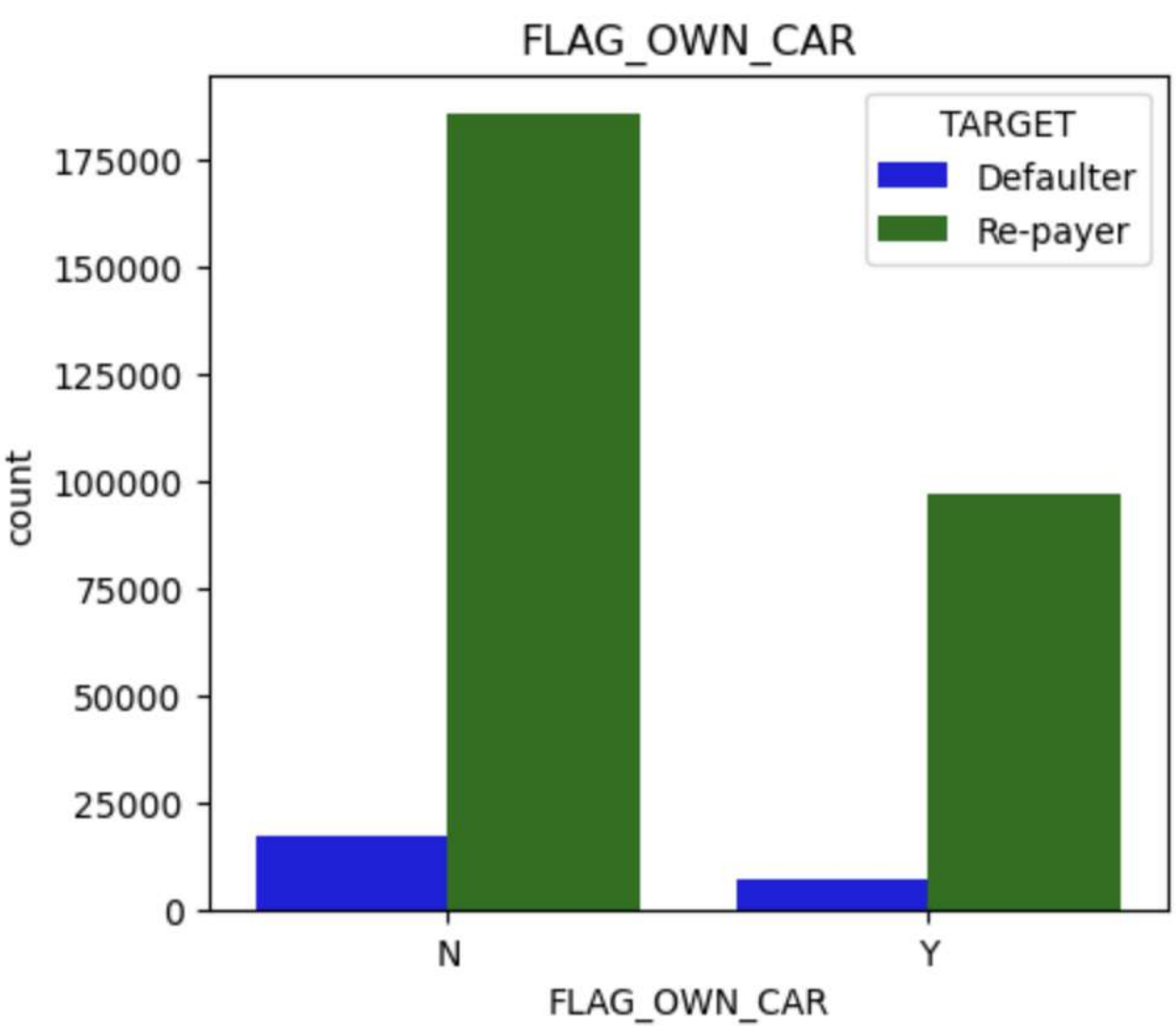
DS C49 – October 2022 batch

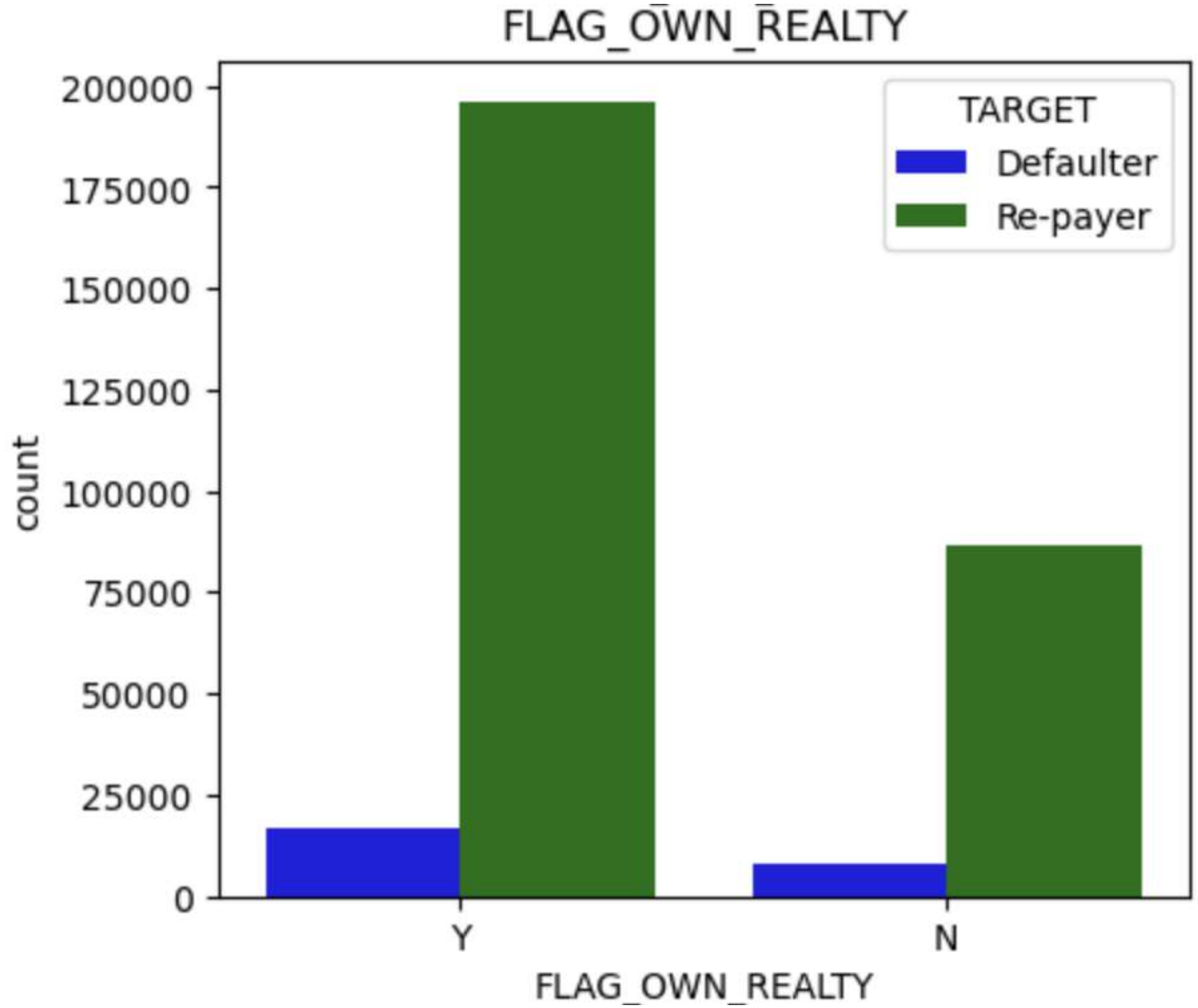


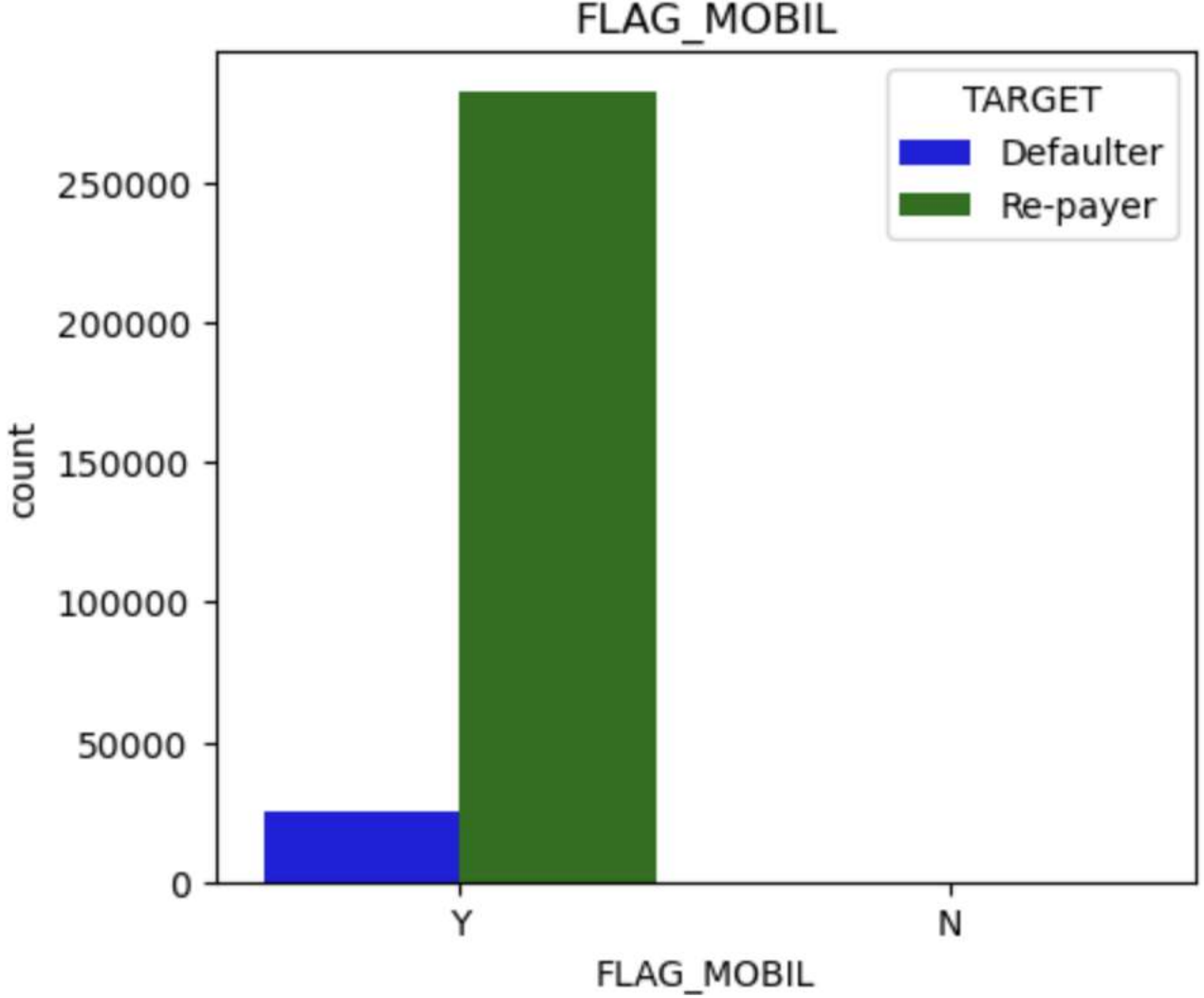
The correlation plot shows that the 8 columns,
['YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG', 'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE', 'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_MODE', 'EMERGENCYSTATE_MODE’]
have a non-linear relationship with the Target variable, and hence, even if we impute the missing values in these columns, they will not affect the TARGET value. hence these can be dropped comfortably.

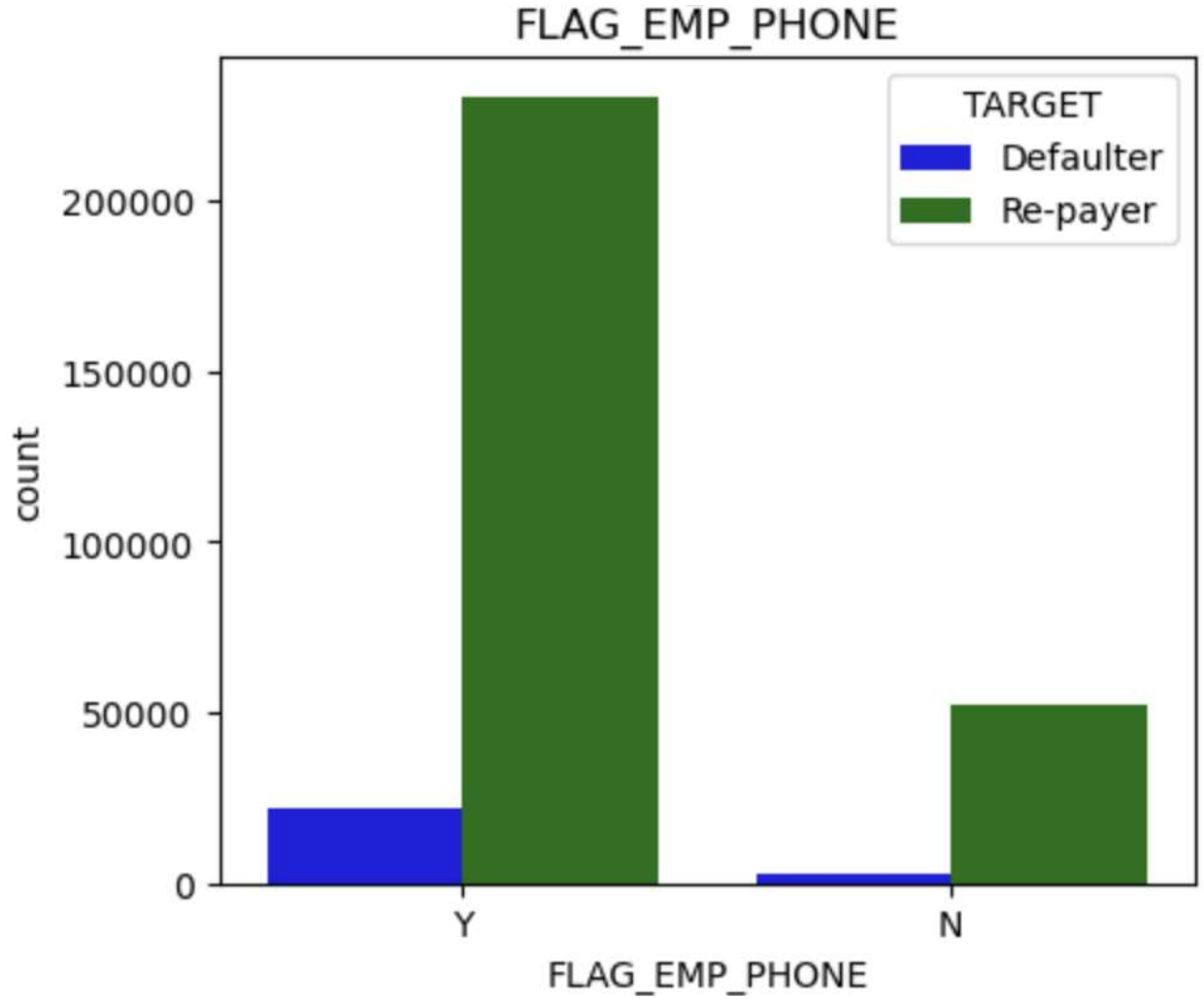


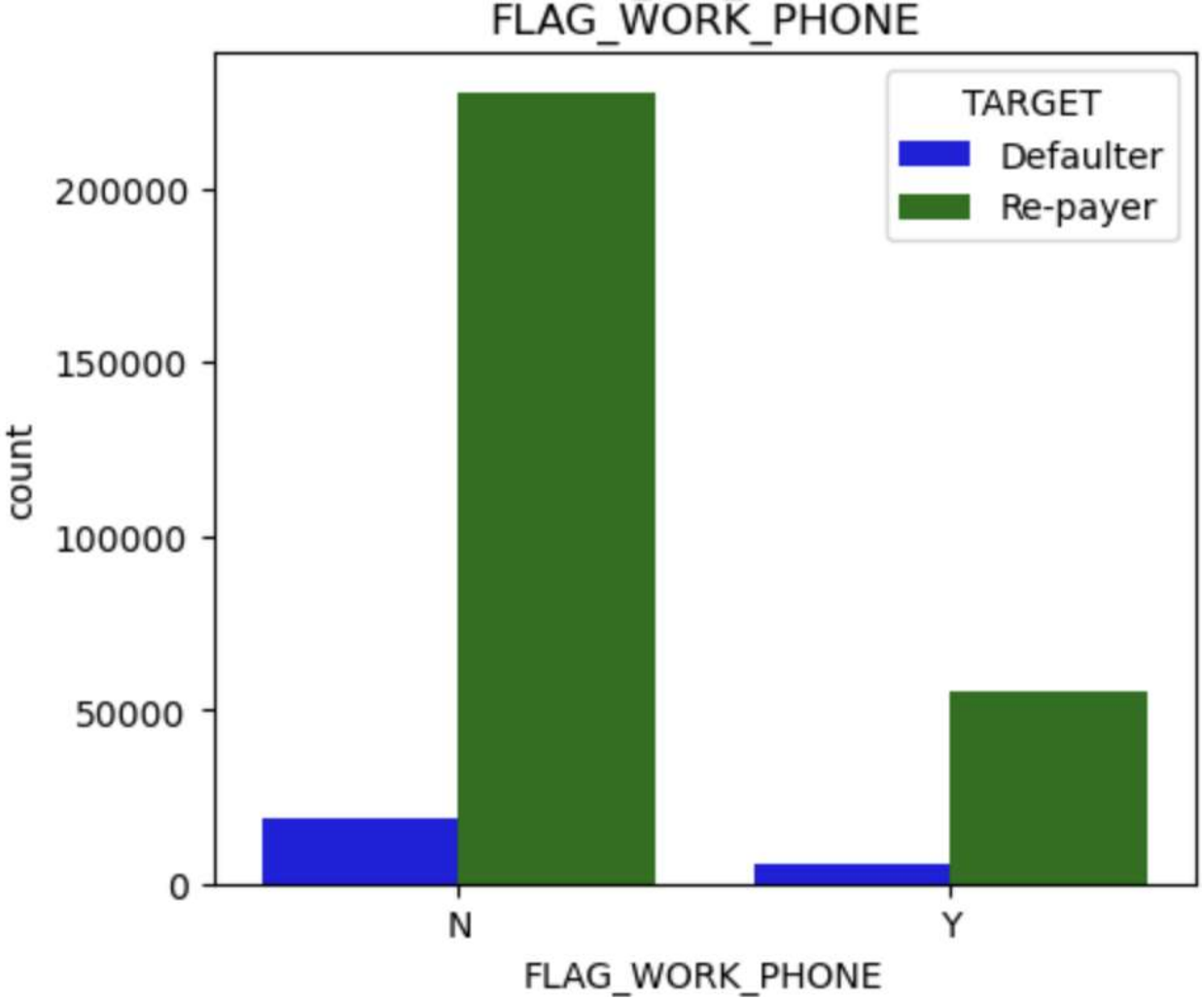
these two columns, 'EXT_SOURCE_3', 'EXT_SOURCE_2' also clearly have a non-linear relation with TARGET variable and hence I feel these have little to no effect on the TARGET variable. and hence, I am dropping these columns

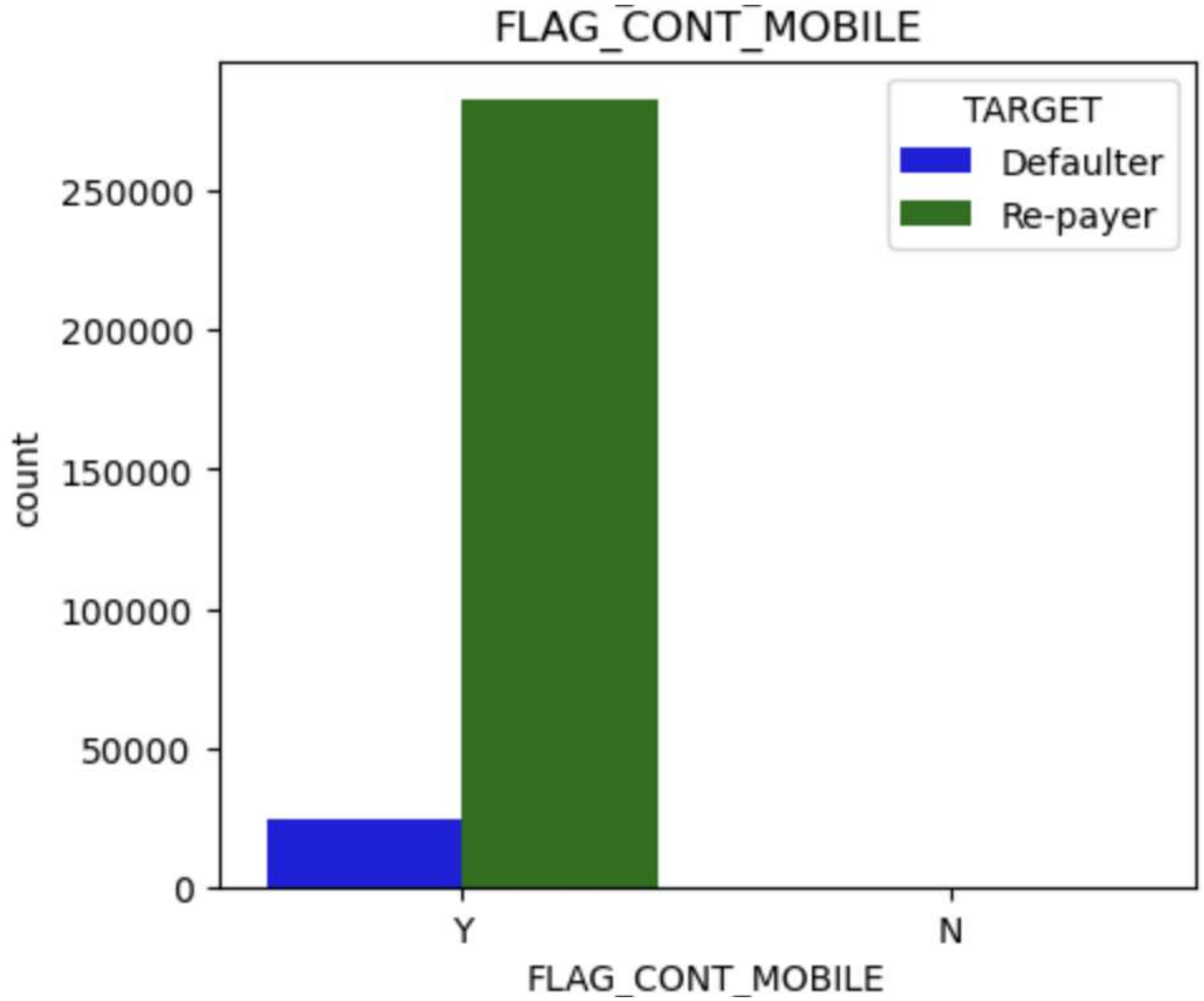

FLAG_OWN_CAR
TARGET
Defaulter
Re-payer
count
175000
150000
125000
100000
75000
50000
25000
0
N
Y
FLAG_OWN_CAR

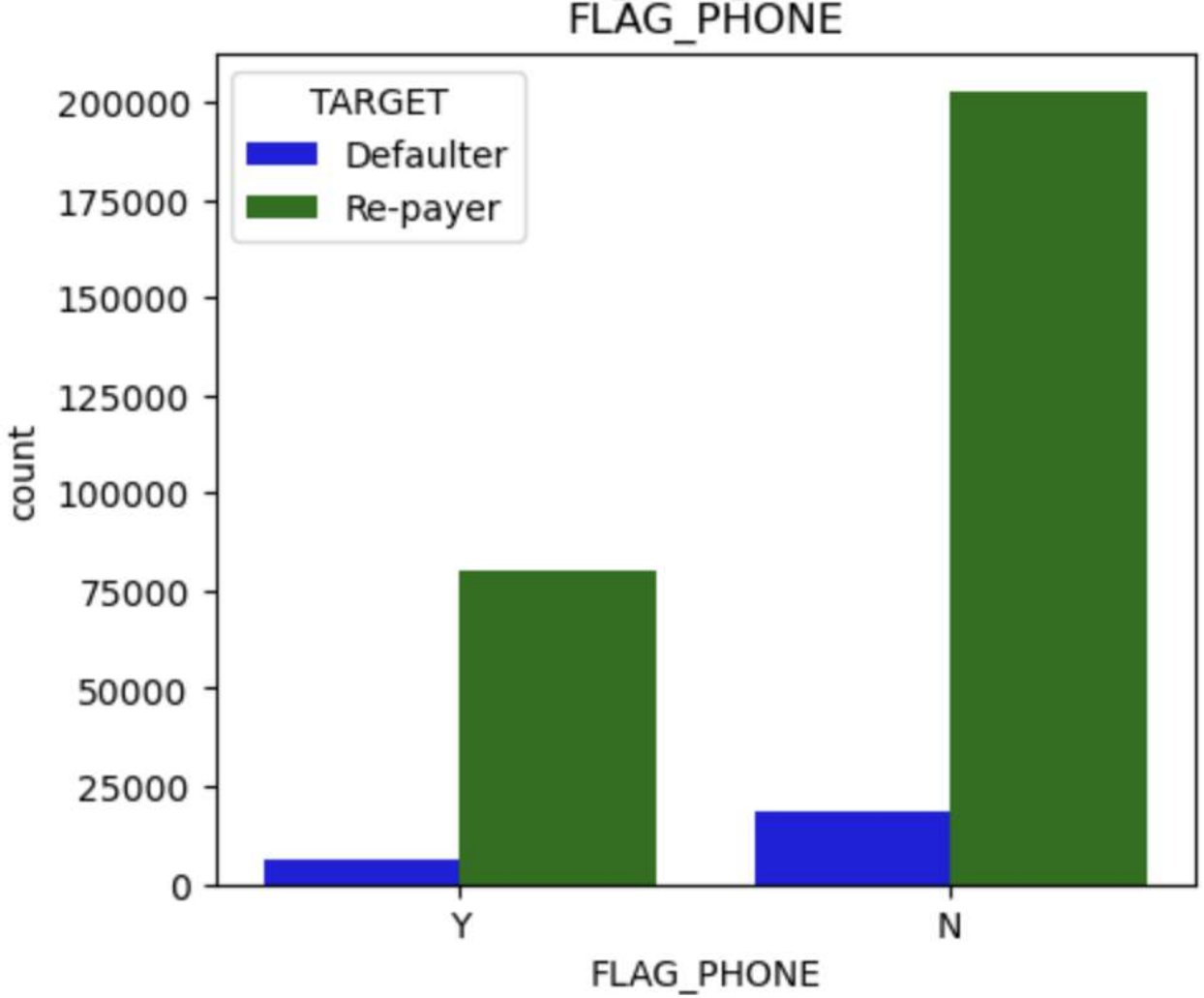

FLAG_OWN_REALTY
TARGET
Defaulter
Re-payer
count
200000
175000
150000
125000
100000
75000
50000
25000
0
Y
N
FLAG_OWN_REALTY

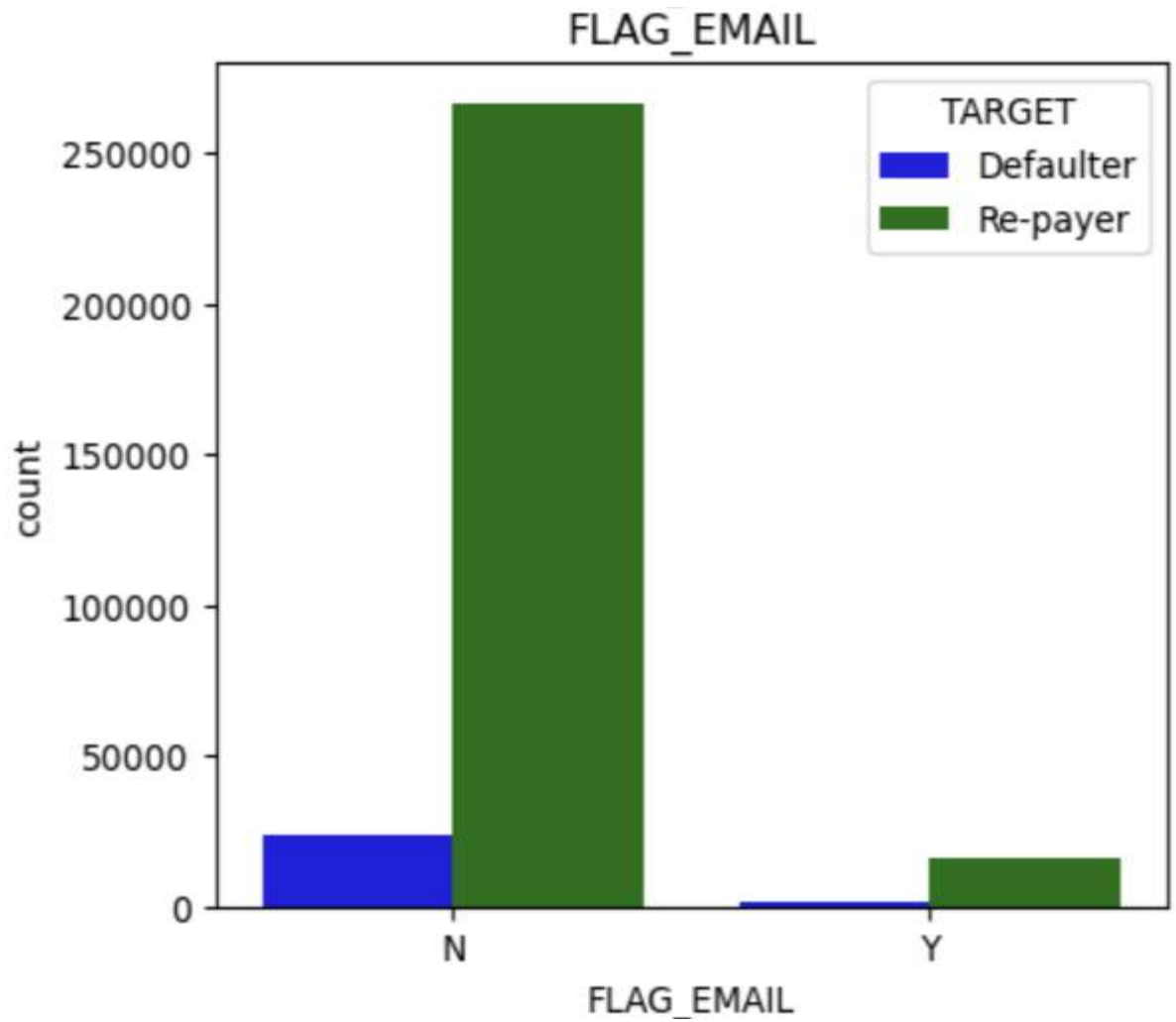

FLAG_MOBIL
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
Y
N
FLAG_MOBIL

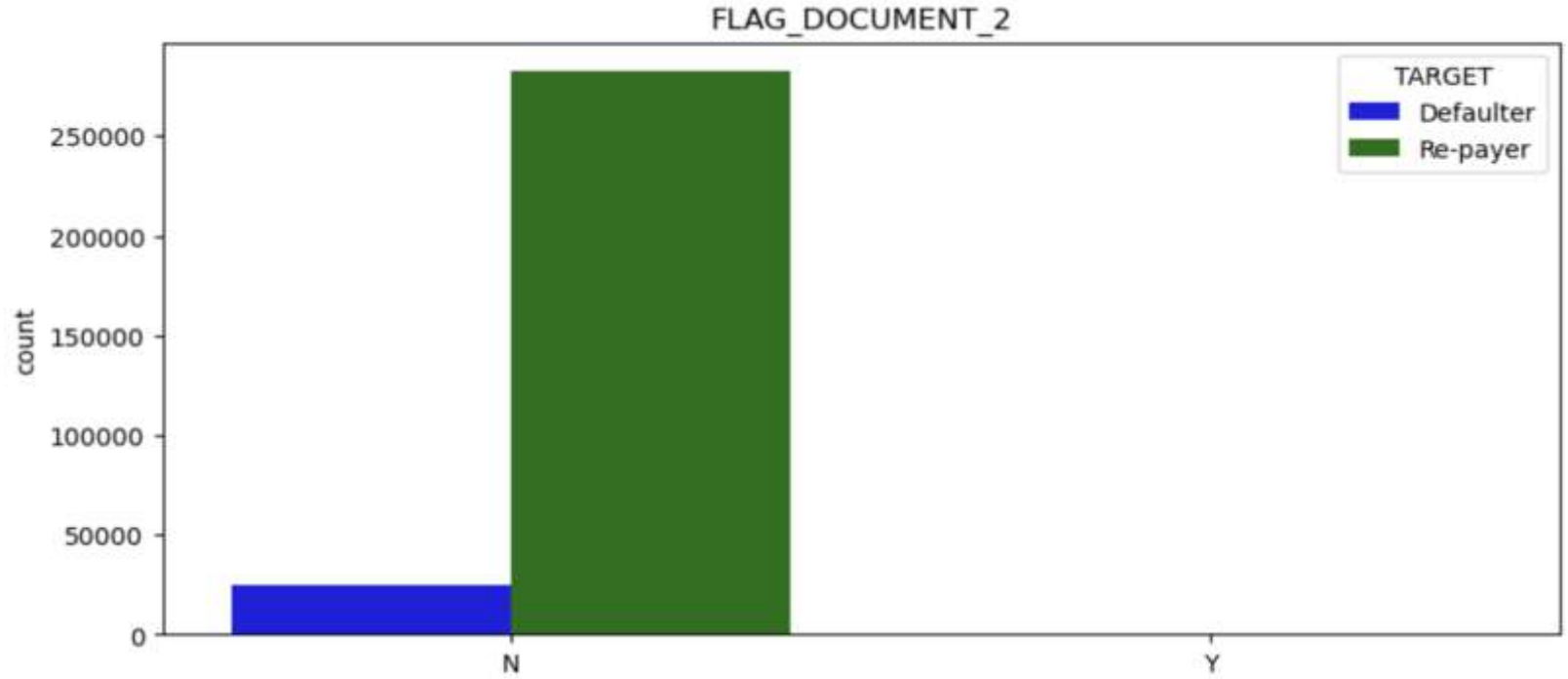

FLAG_EMP_PHONE
TARGET
Defaulter
Re-payer
200000
150000
100000
50000
0
count
Y
N
FLAG_EMP_PHONE

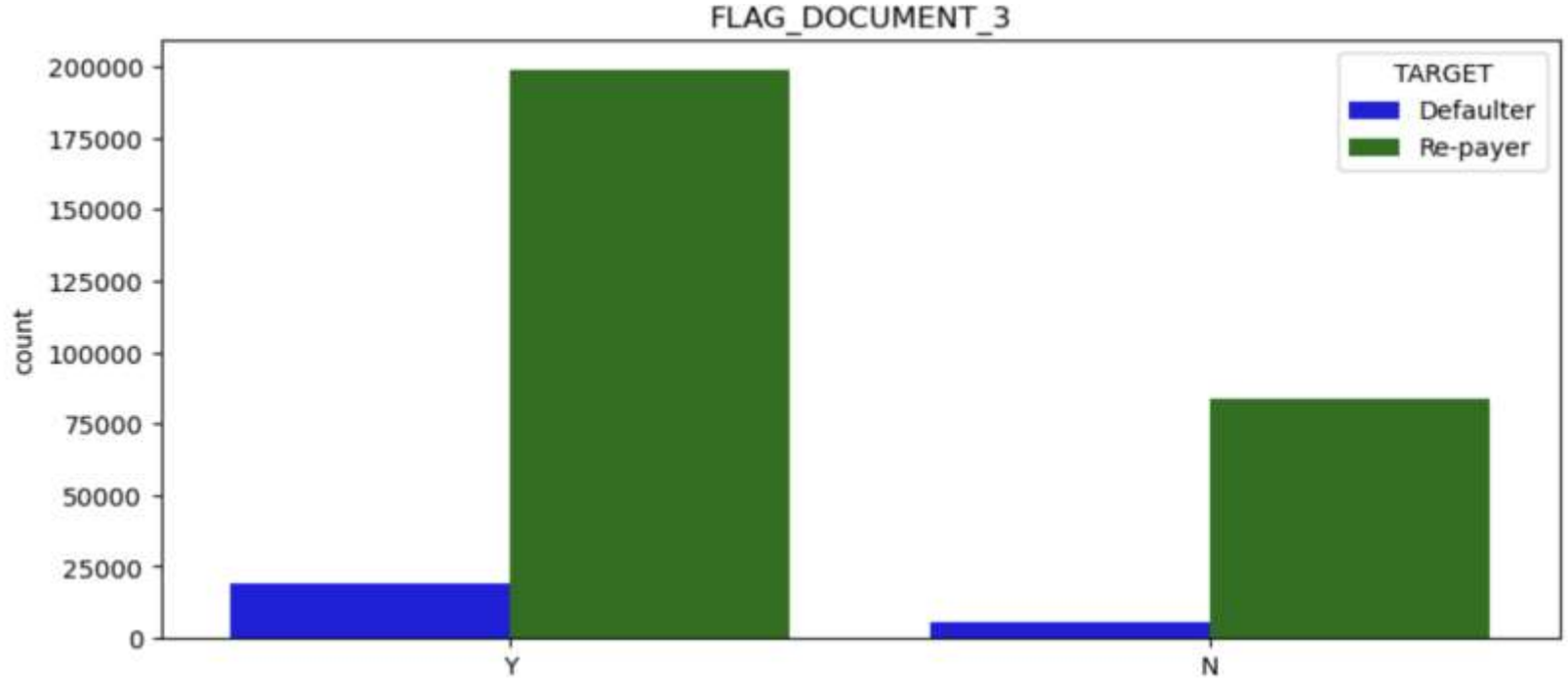

FLAG_WORK_PHONE
TARGET
Defaulter
Re-payer
200000
150000
100000
50000
0
count
N
Y
FLAG_WORK_PHONE

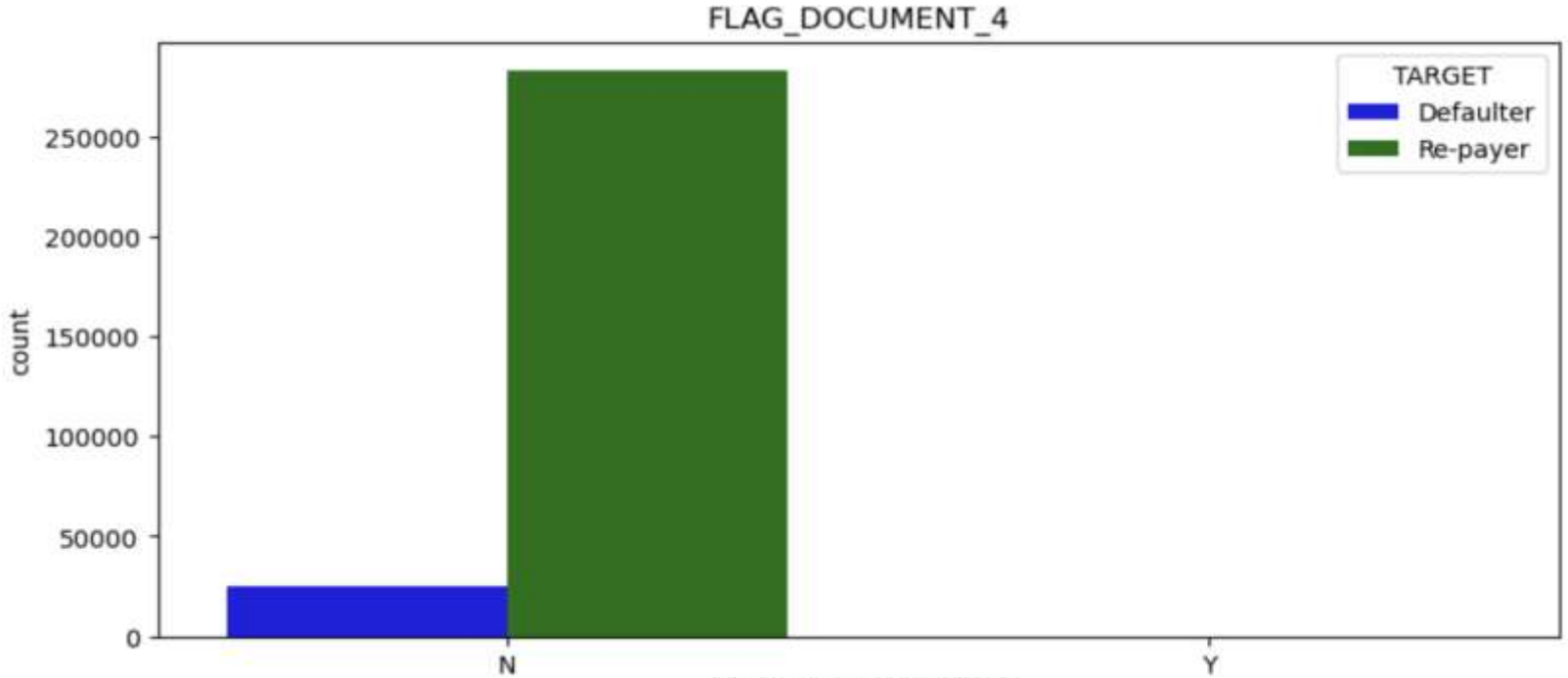

FLAG_CONT_MOBILE
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
Y
N
FLAG_CONT_MOBILE

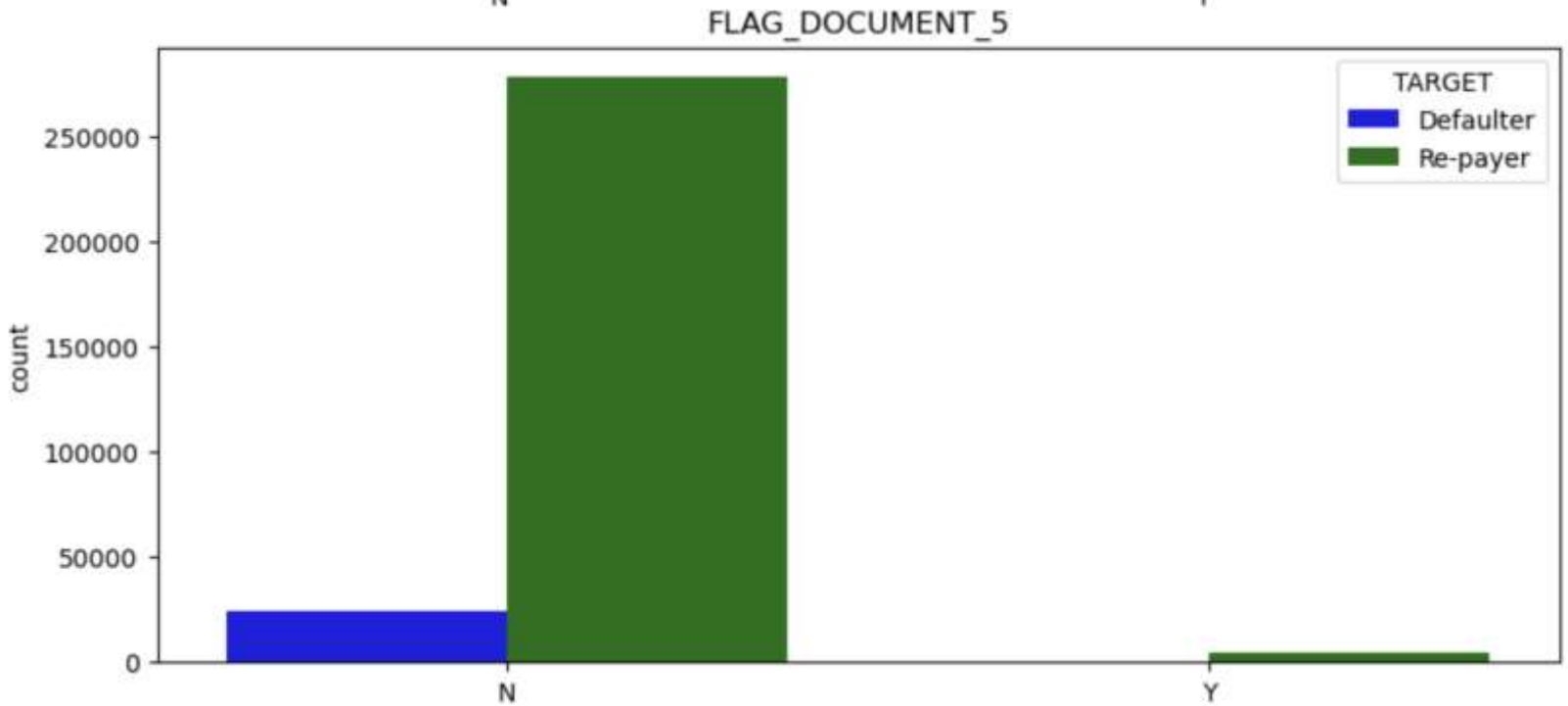

FLAG_PHONE
TARGET
Defaulter
Re-payer
200000
175000
150000
125000
100000
75000
50000
25000
0
count
Y
N
FLAG_PHONE

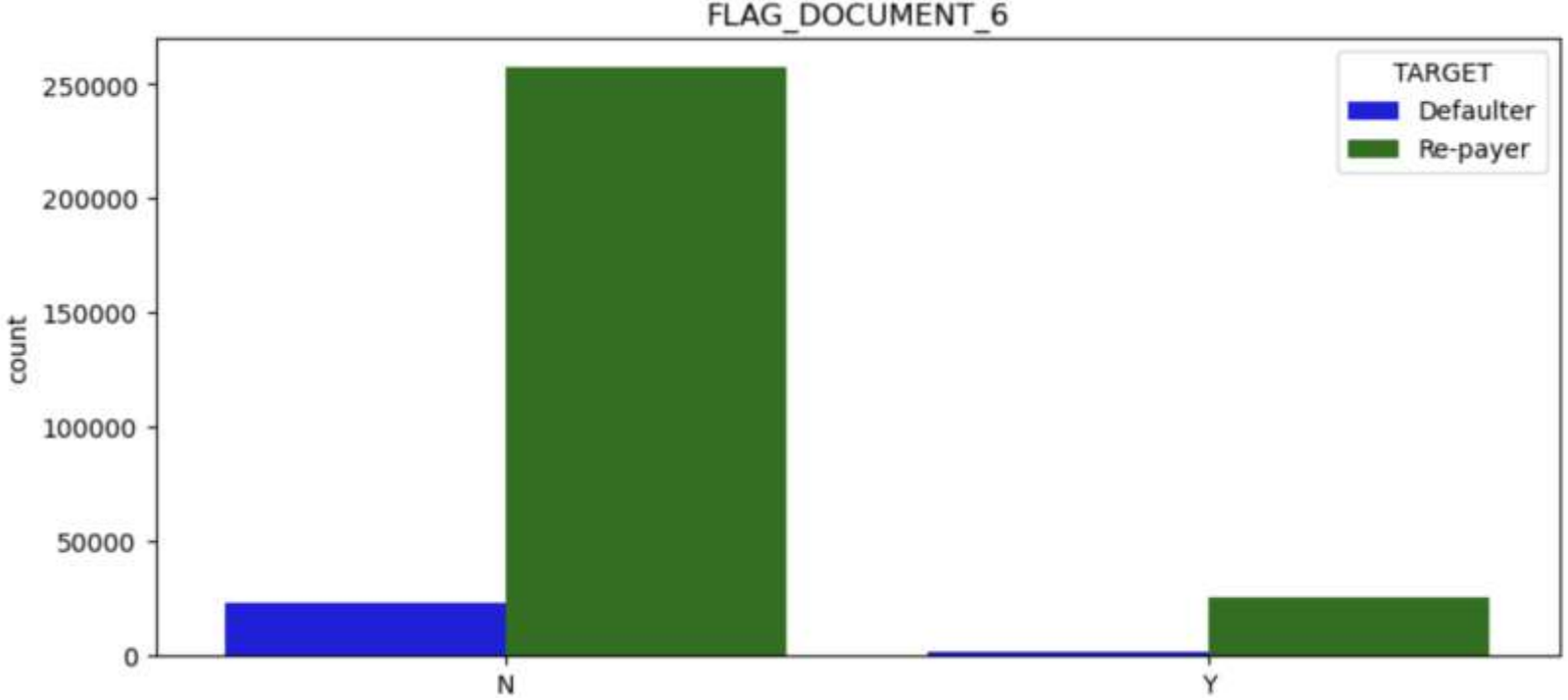

FLAG_EMAIL
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y
FLAG_EMAIL

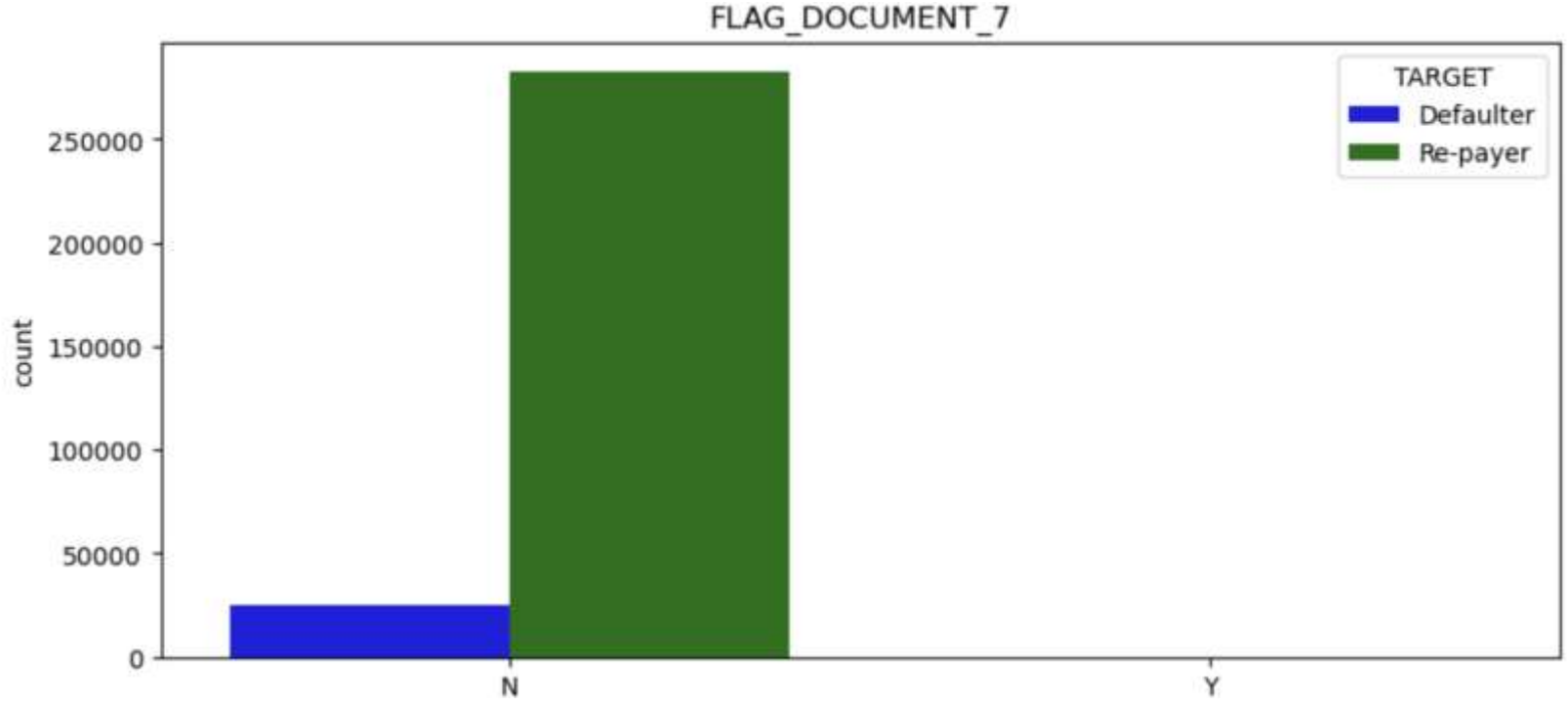

FLAG_DOCUMENT_2
TARGET
Defaulter
Re-payer
count
250000
200000
150000
100000
50000
0
N
Y

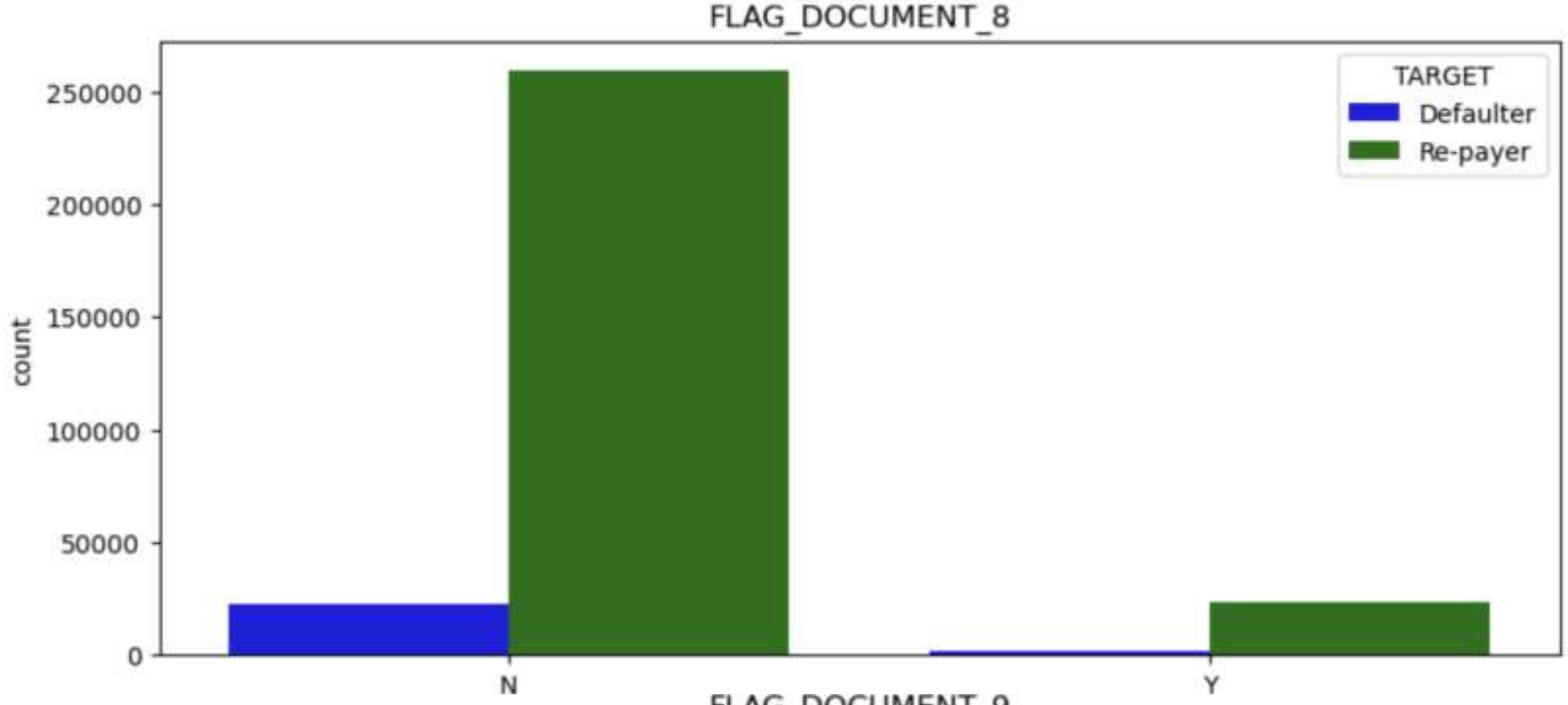

FLAG_DOCUMENT_3
TARGET
Defaulter
Re-payer
count
200000
175000
150000
125000
100000
75000
50000
25000
0
Y
N

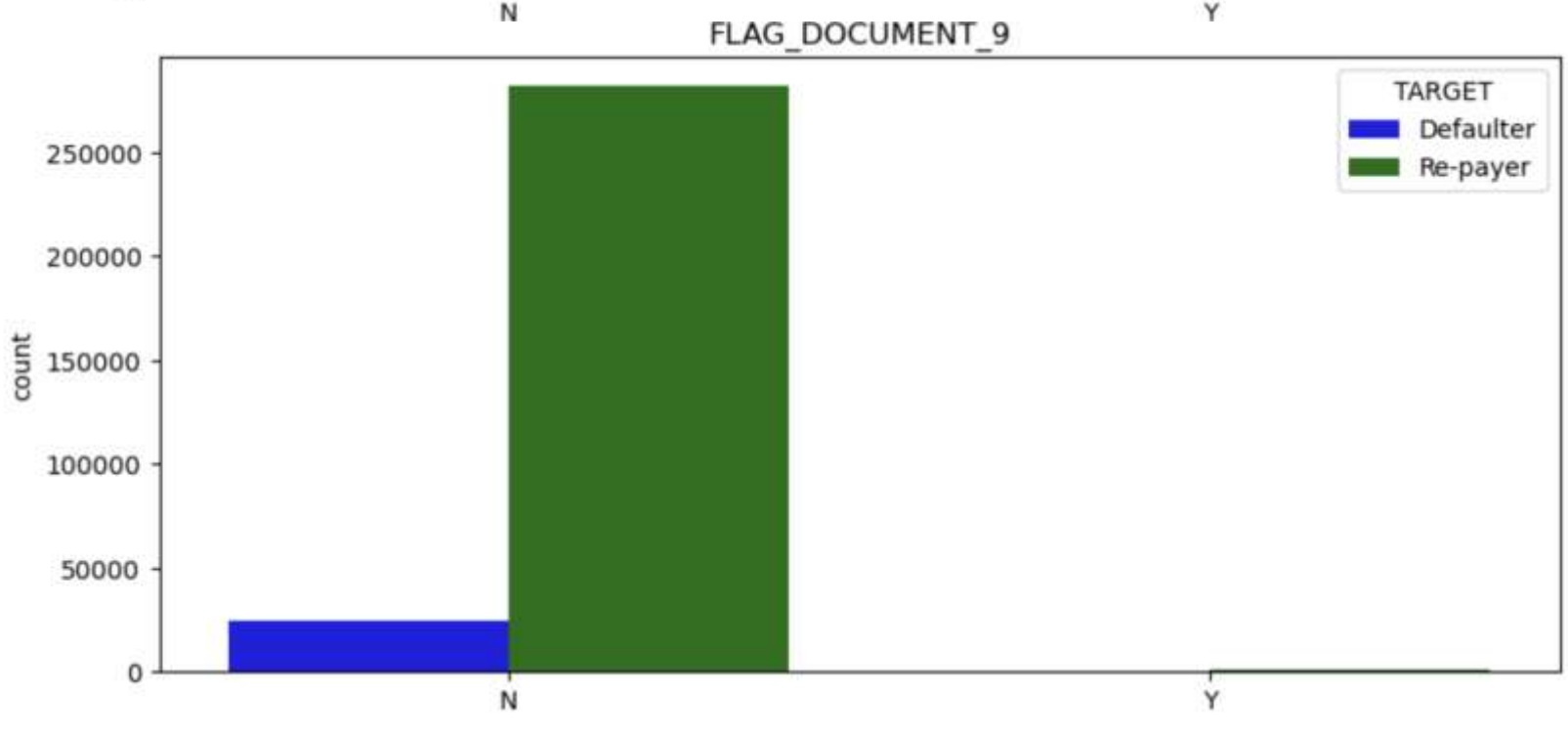

FLAG_DOCUMENT_4
TARGET
Defaulter
Re-payer
count
250000
200000
150000
100000
50000
0
N
Y

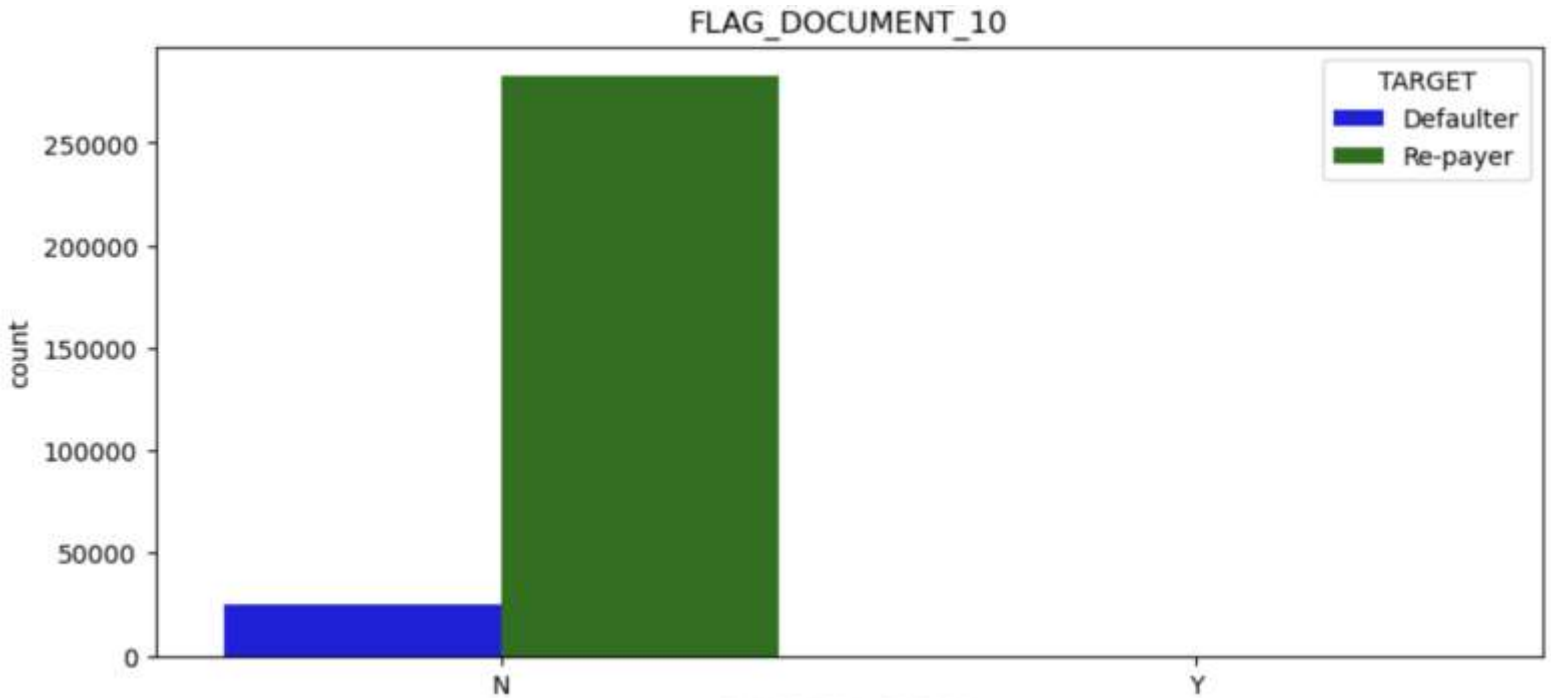

FLAG_DOCUMENT_5
TARGET
Defaulter
Re-payer
count
250000
200000
150000
100000
50000
0
N
Y

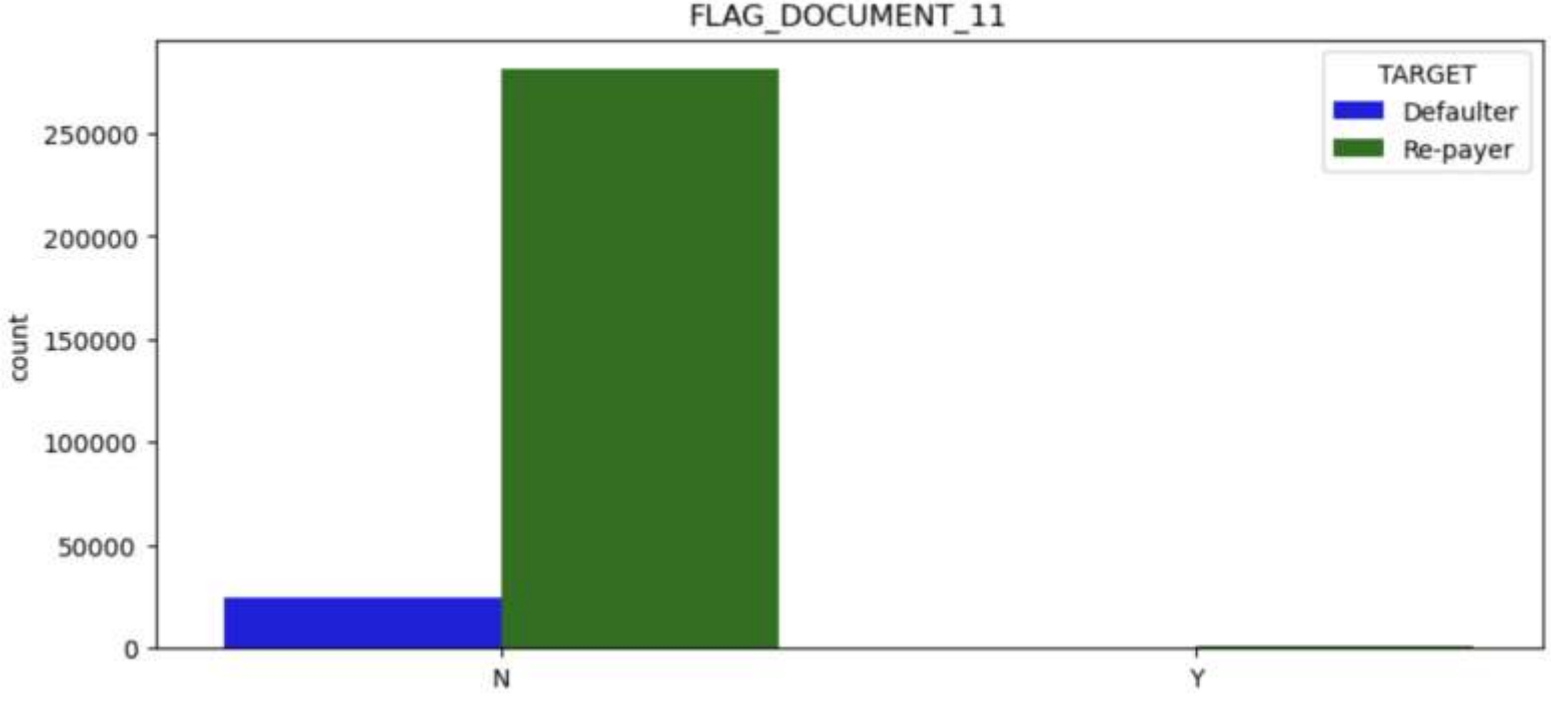

FLAG_DOCUMENT_6
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y

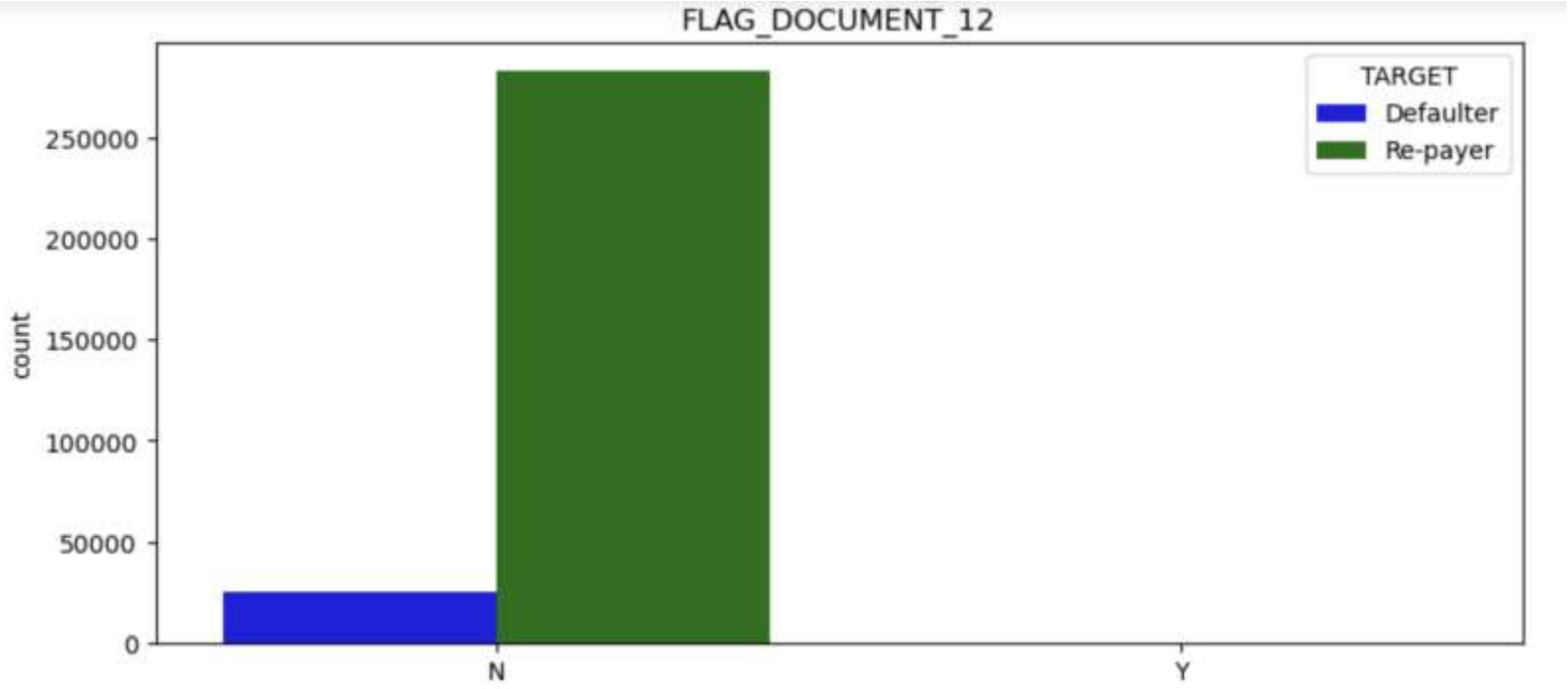

FLAG_DOCUMENT_7
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y

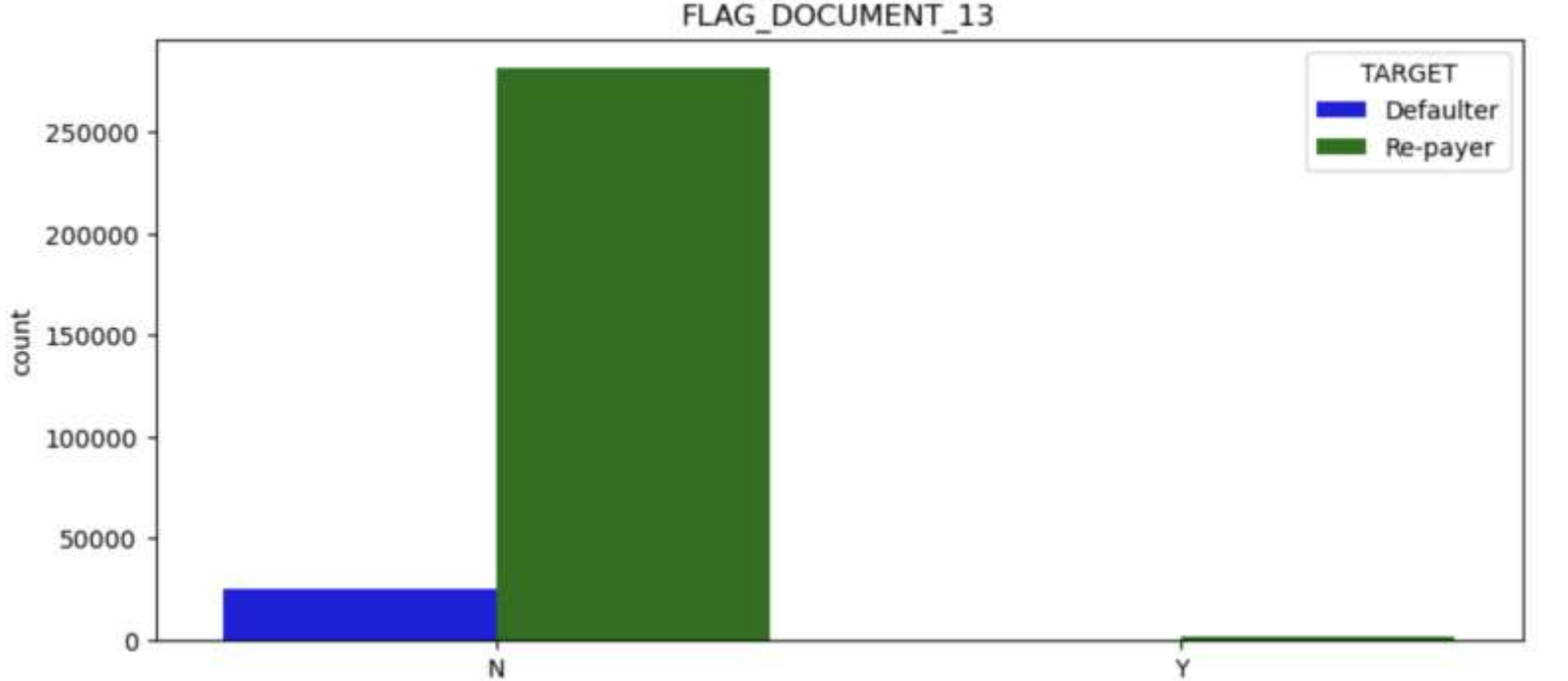

FLAG_DOCUMENT_8
TARGET
Defaulter
Re-payer
count
250000
200000
150000
100000
50000
0
N
Y

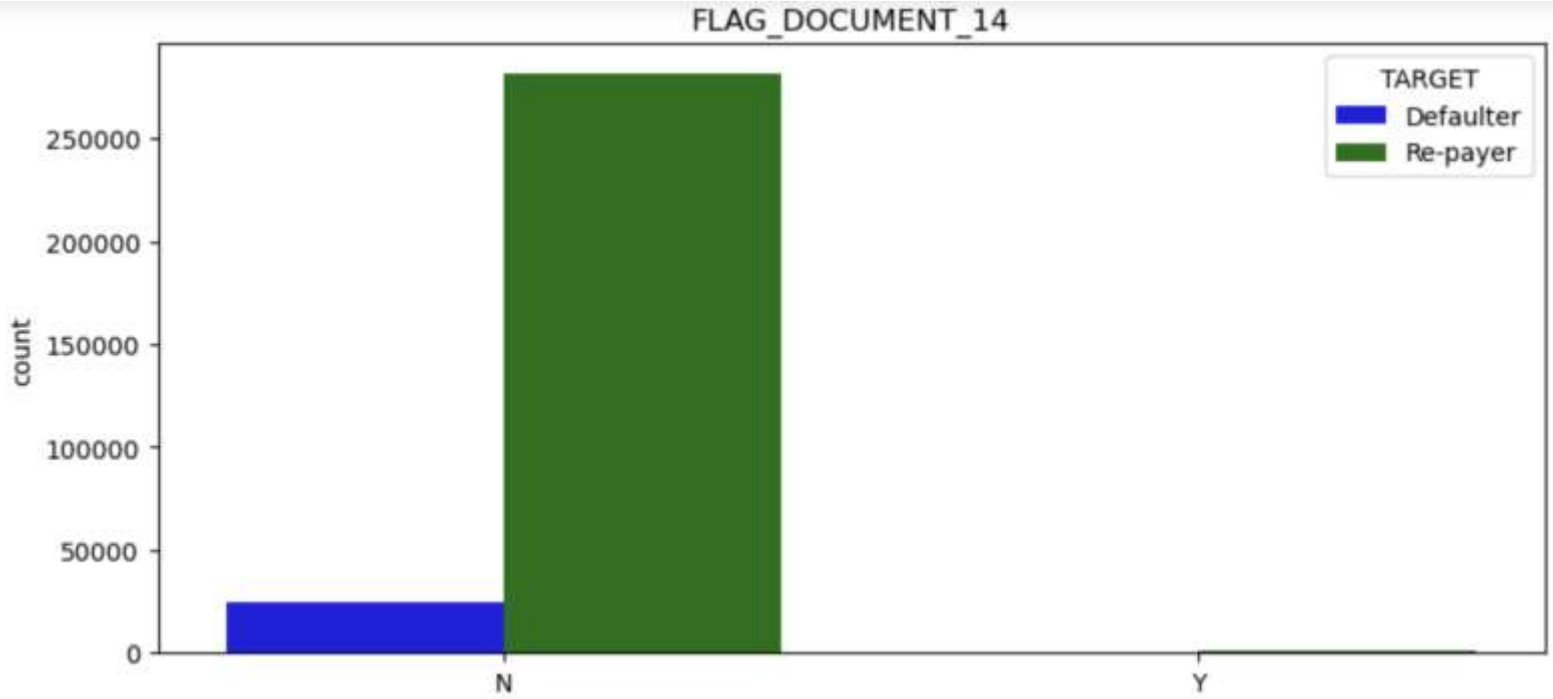

FLAG_DOCUMENT_9
TARGET
Defaulter
Re-payer
count
250000
200000
150000
100000
50000
0
N
Y

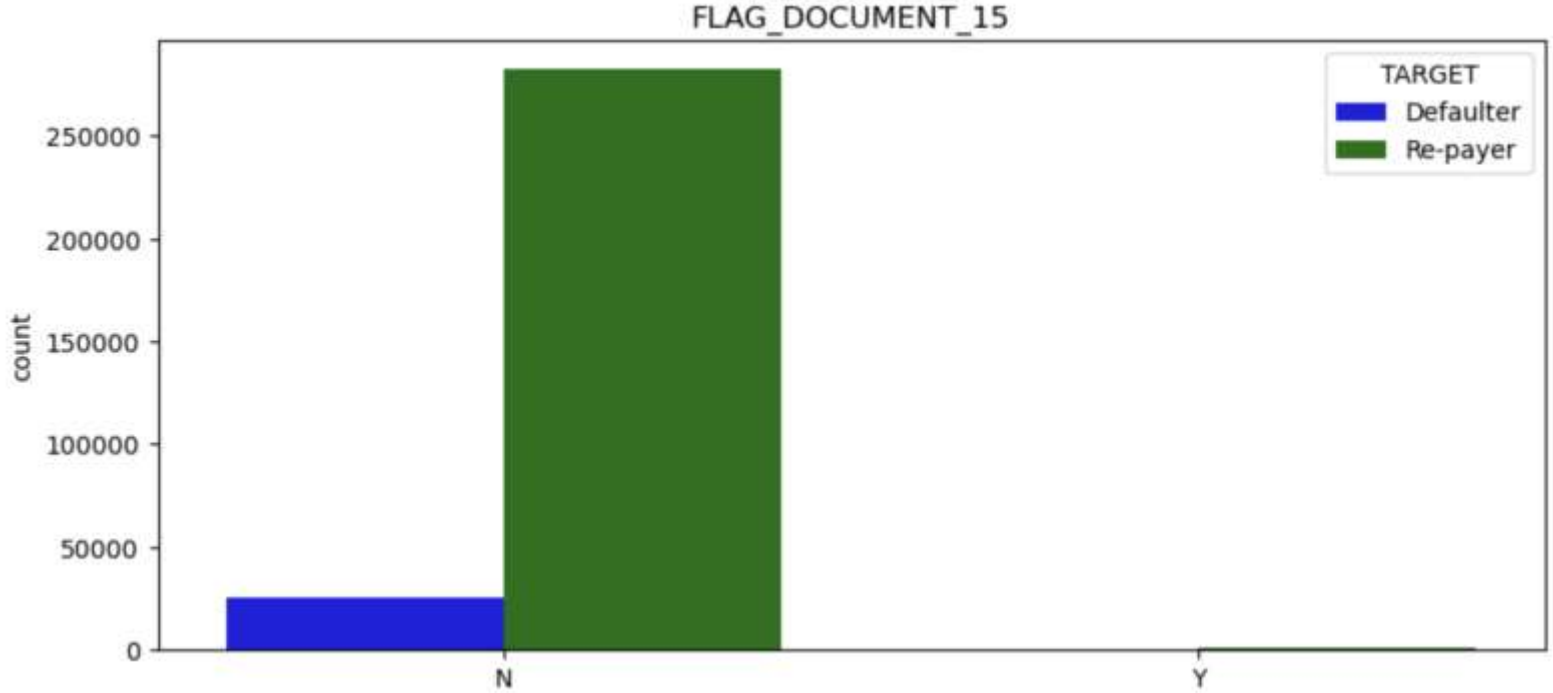

FLAG_DOCUMENT_10
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y
FLAG_DOCUMENT_10

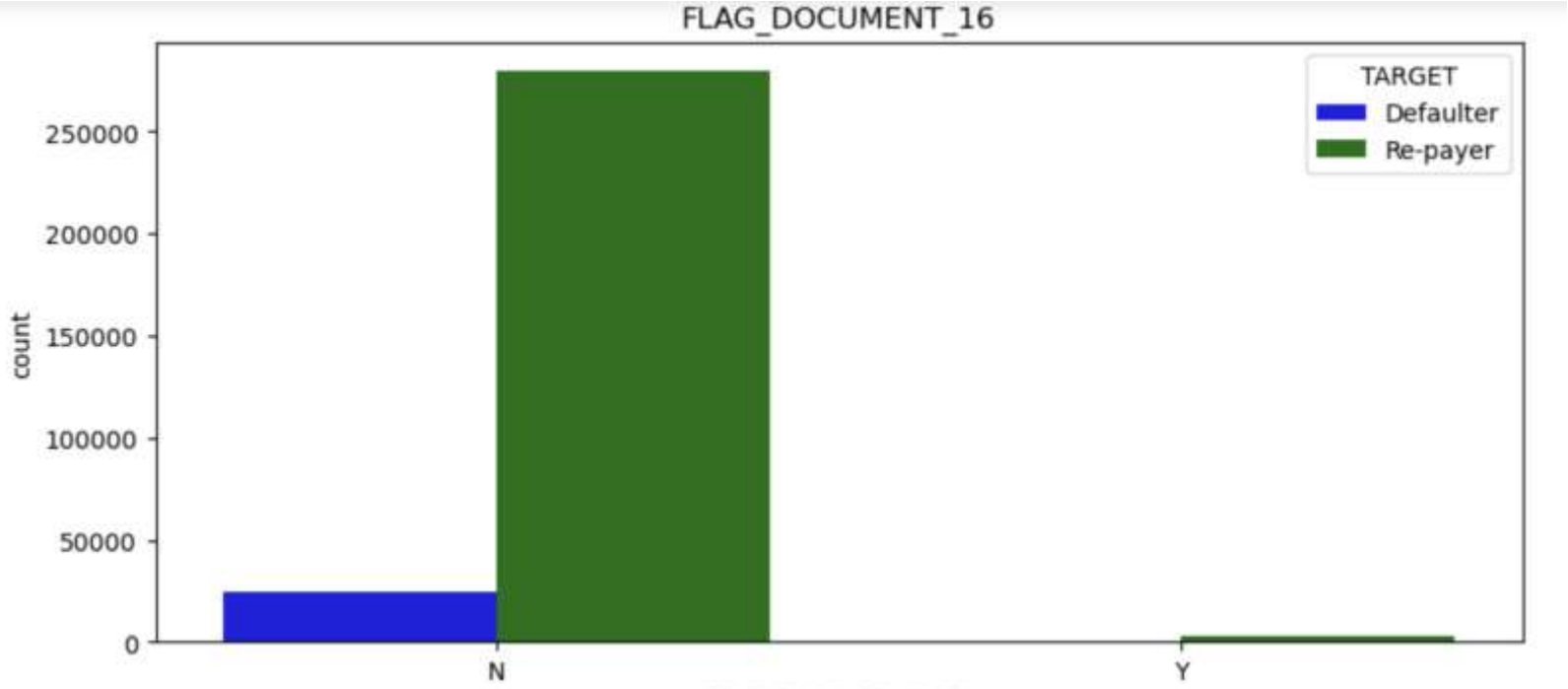

FLAG_DOCUMENT_11
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y


FLAG_DOCUMENT_12
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y
FLAG_DOCUMENT_12

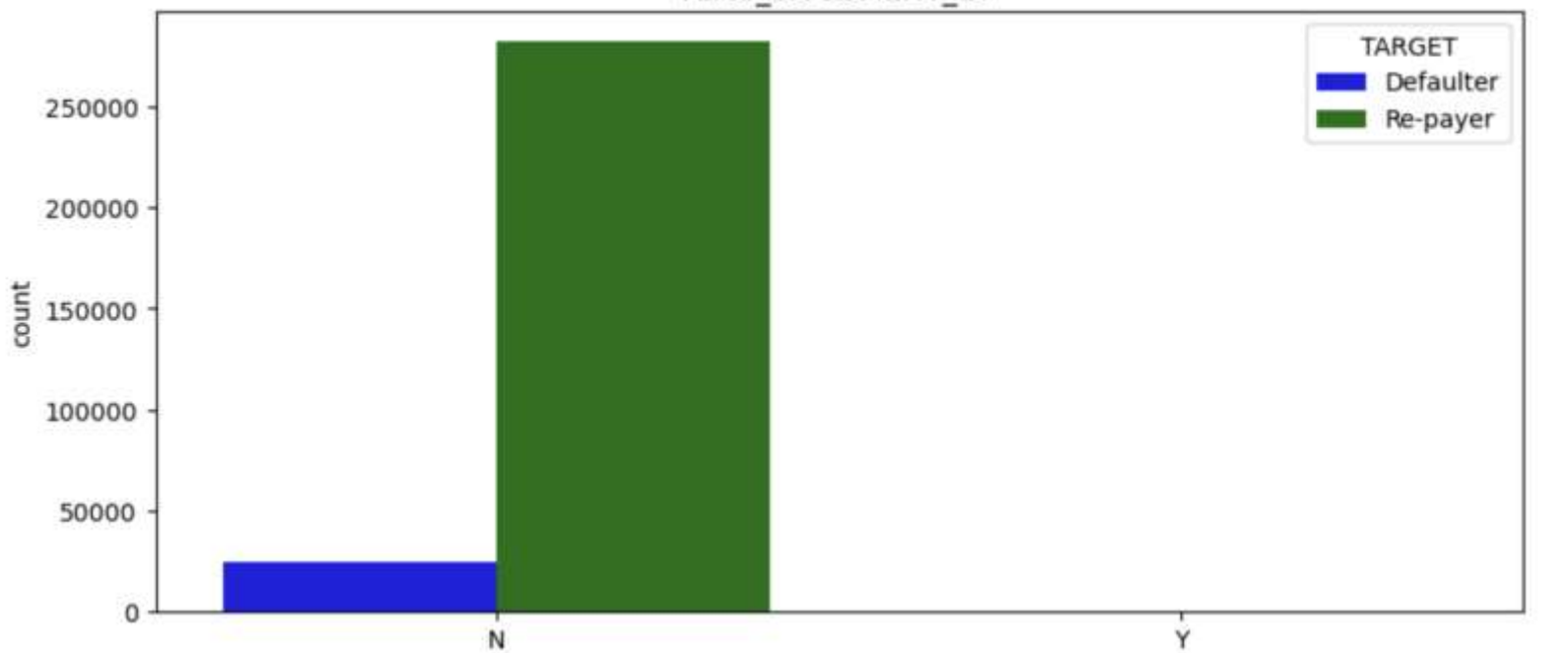

FLAG_DOCUMENT_13
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y

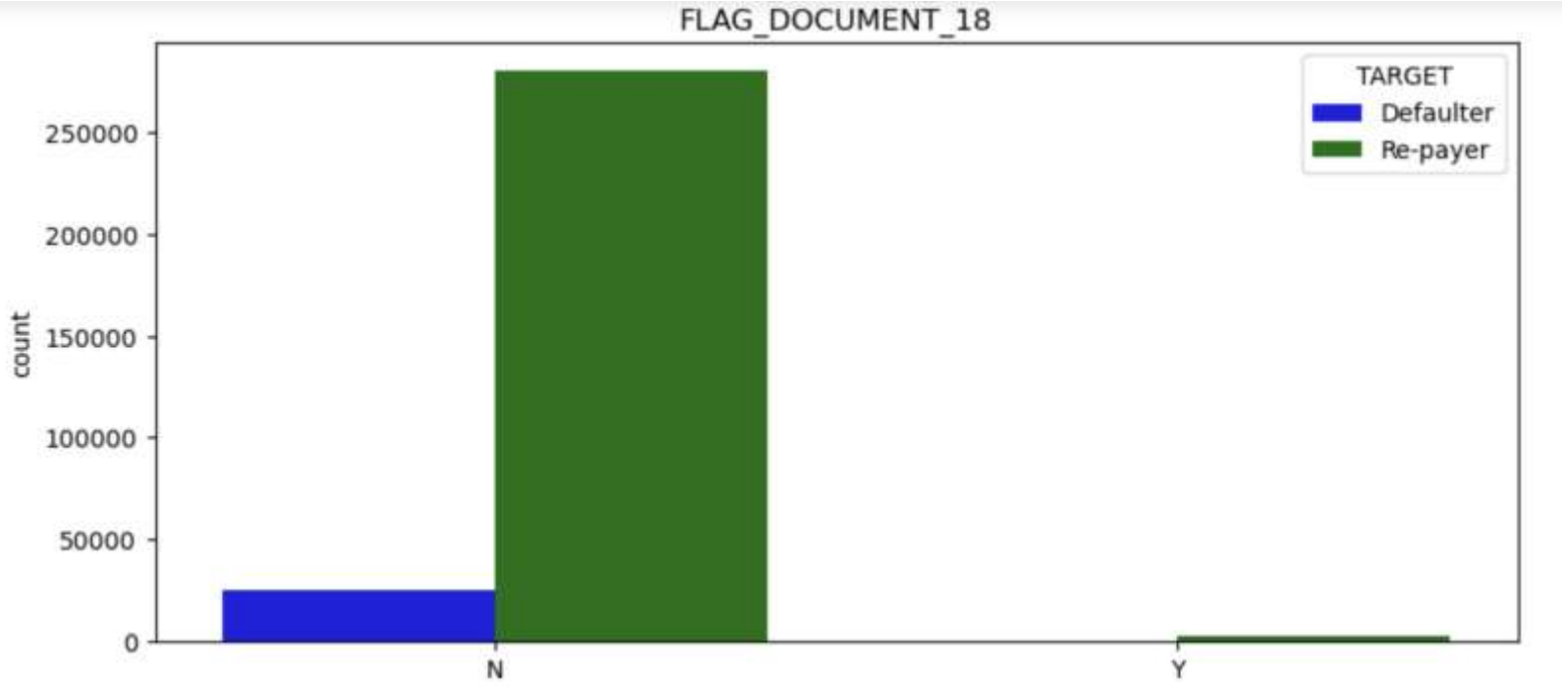

FLAG_DOCUMENT_14
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y
FLAG_DOCUMENT_14

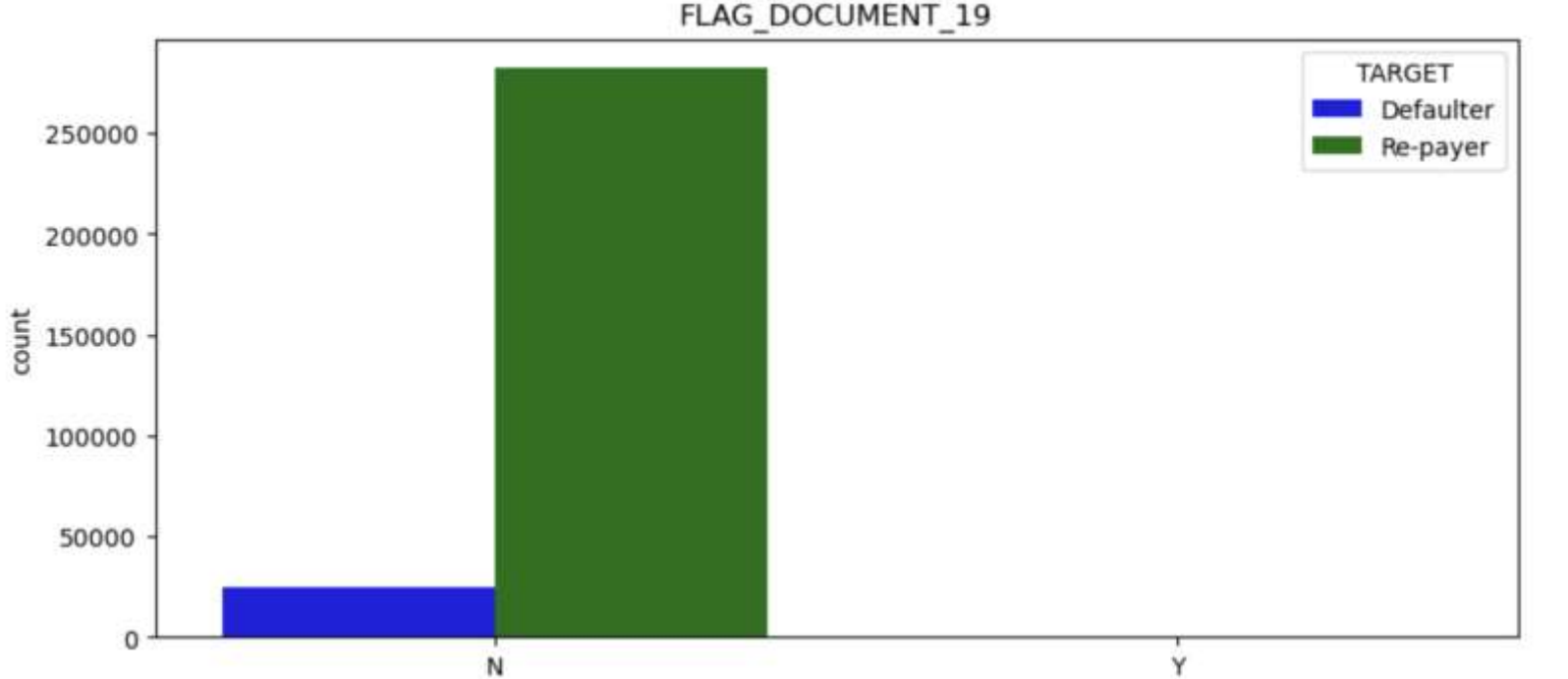

FLAG_DOCUMENT_15
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y

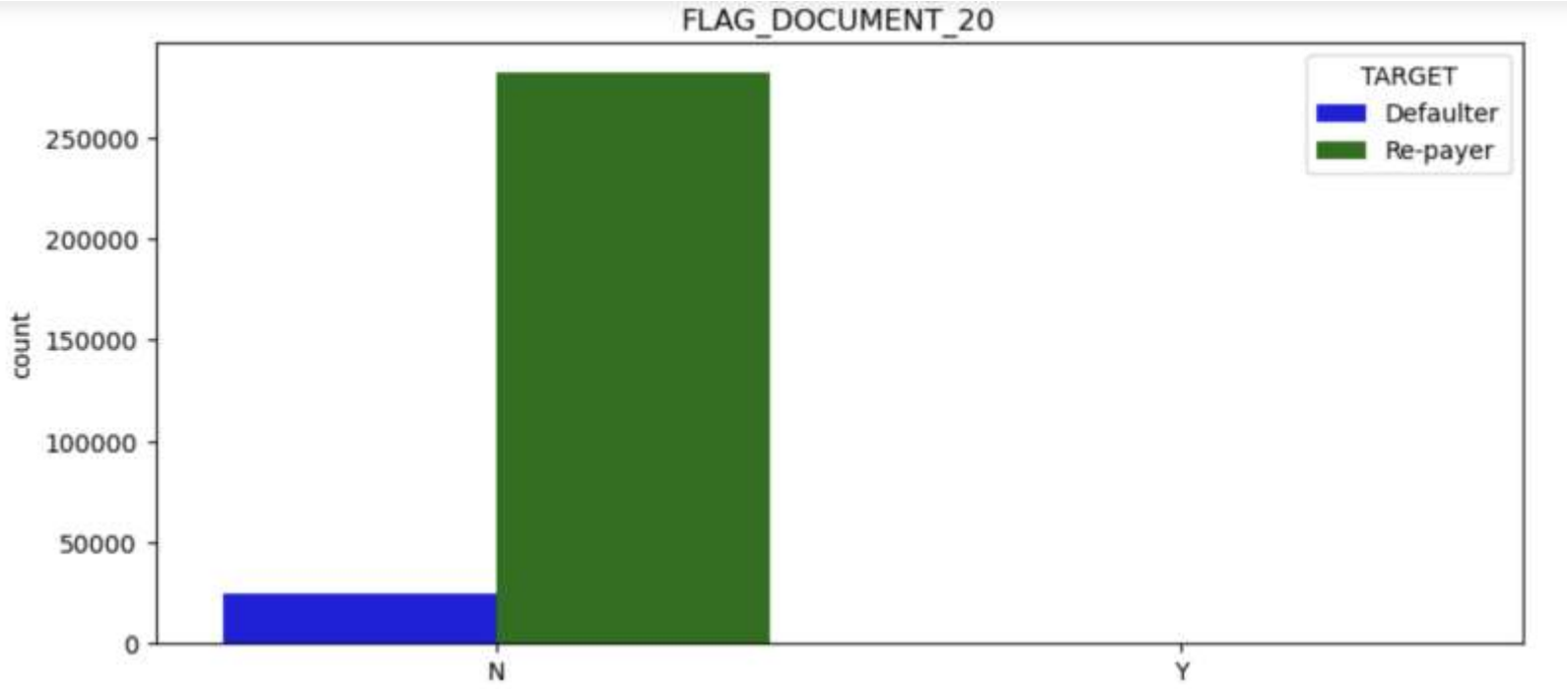


FLAG_DOCUMENT_17

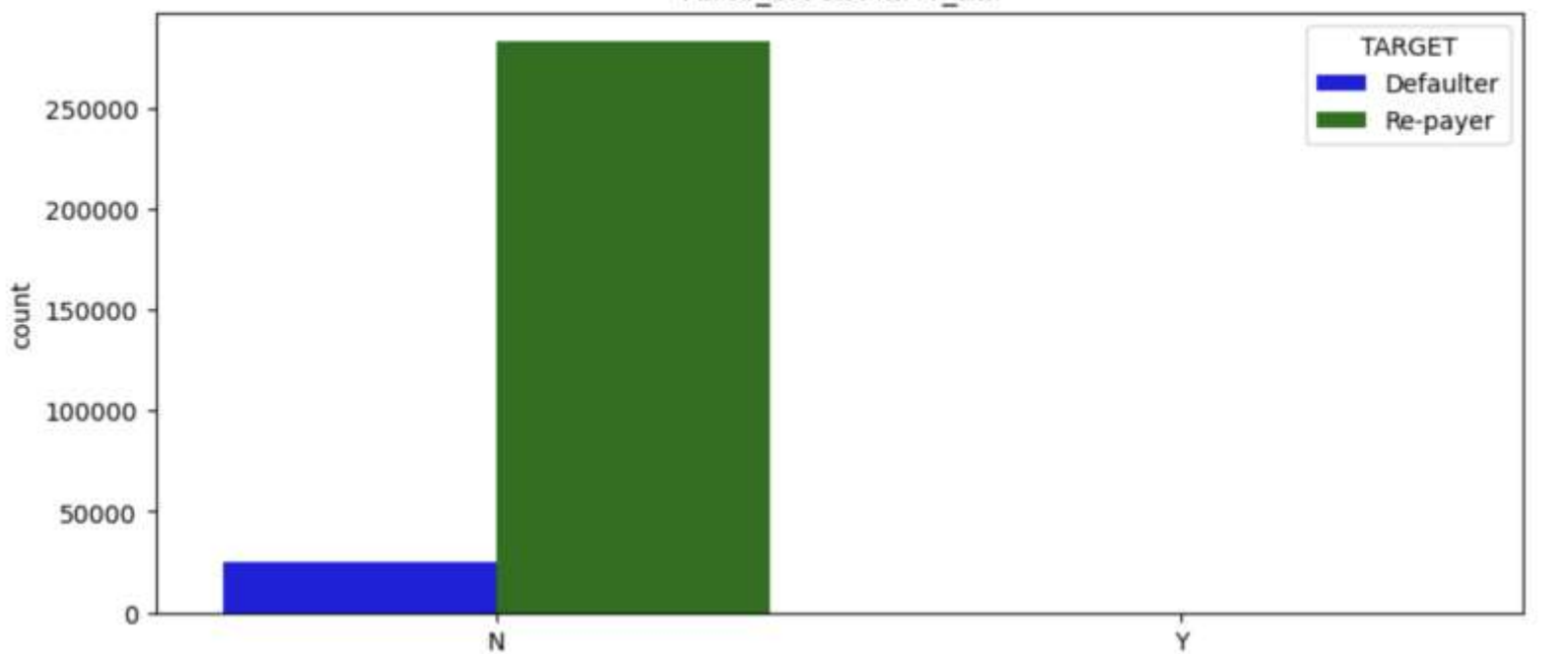



FLAG_DOCUMENT_18
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y
FLAG_DOCUMENT_18


FLAG_DOCUMENT_19
TARGET
Defaulter
Re-payer
250000
200000
150000
100000
50000
0
count
N
Y



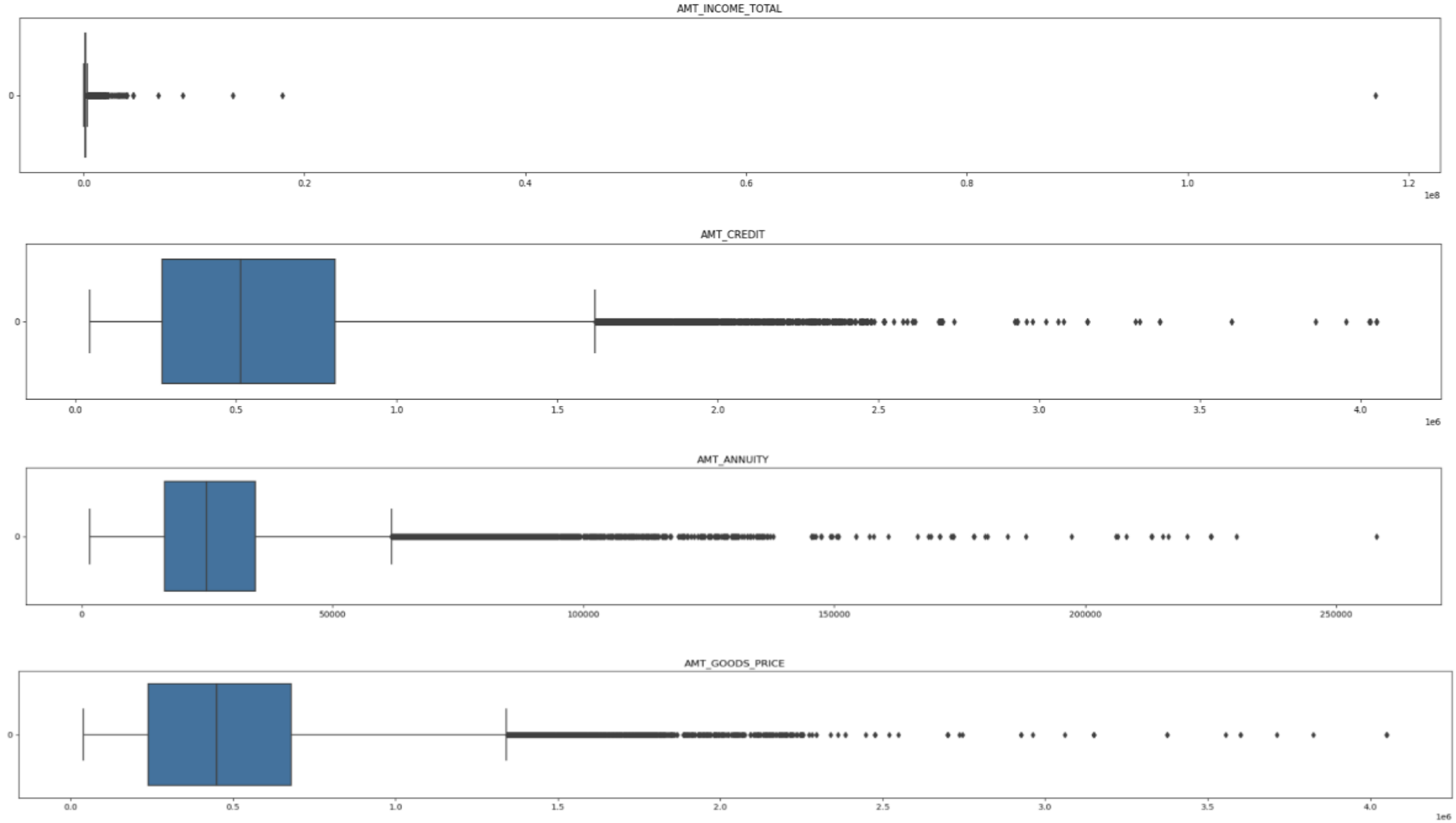
FLAG_DOCUMENT_21



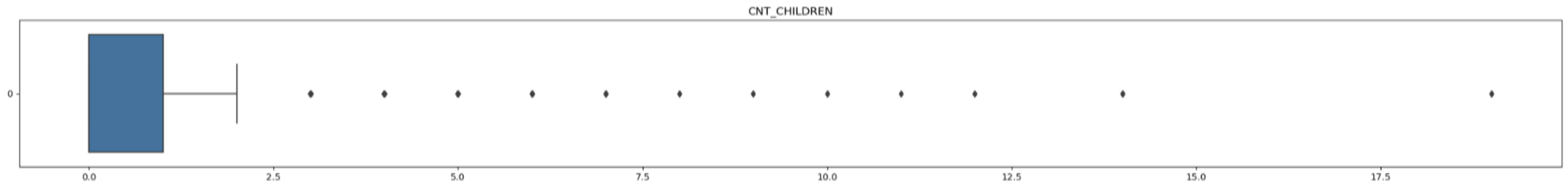
Observations from above plots in slides 4-17:

FLAG_OWN_REALTY, FLAG_MOBIL, FLAG_EMP_PHONE, FLAG_CONT_MOBIL, FLAG_DOCUMENT_3 for 'Y' has more re-payers than defaulters. i.e., applicants

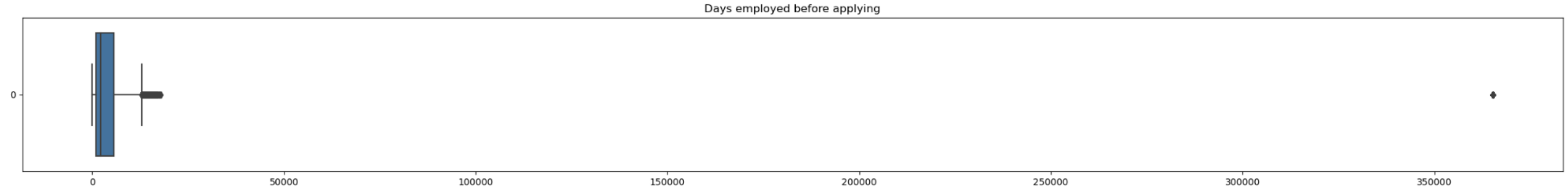
1. applicants having own house/flat have been re-payers than defaulters.
2. applicants who provided mobile number have more re-payers than defaulters.
3. applicants whose phones are reachable easily are re-payers than defaulters.
4. applicants who provided Document-3 are likely to be re-payers than defaulters


AMT_INCOME_TOTAL
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e8
AMT_CREDIT
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
AMT_ANNUITY
0
0
50000
100000
150000
200000
250000
AMT_GOODS_PRICE
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6

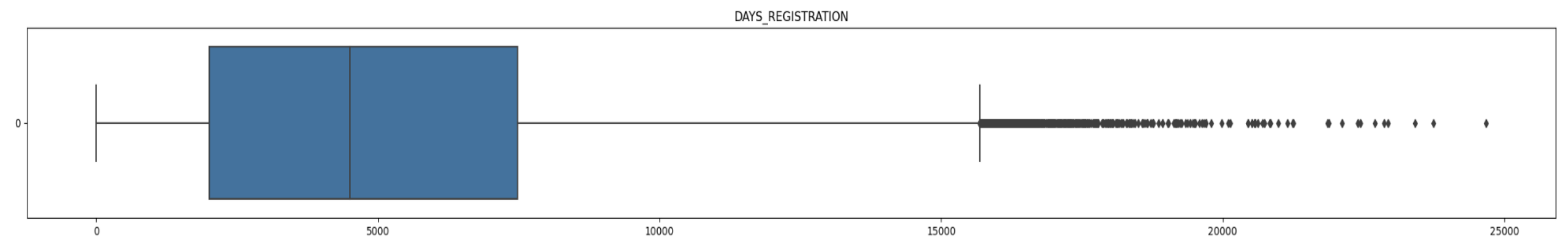
'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE’ have a lot of outliers in these columns, we can either drop the rows that corresponding to these outliers and proceed with the analysis, but in our case, since I think, these are some of the significant columns, I wish to keep them intact for our analysis.



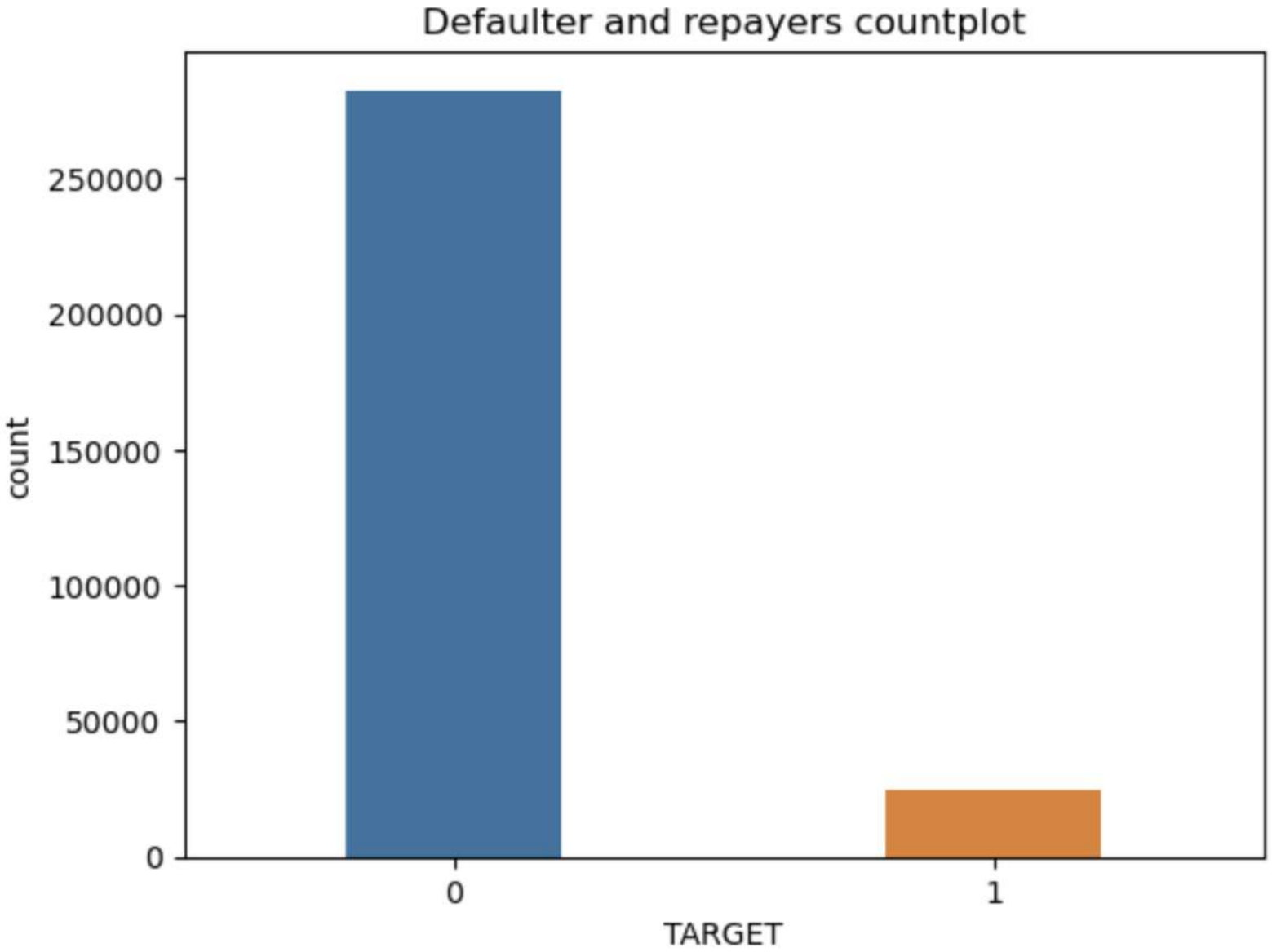
- As evident from the description of CNT_CHILDERN column and also its corresponding boxplot, 1st quartile is completely is completely missing from the boxplot as most of the data is present in the second quartile ie, above 75%.
- Also, as can be seen from the boxplot, there are a very few outliers corresponding to number of childern usually greater than 4 which is a rarity. hence this data can be termed as reliable



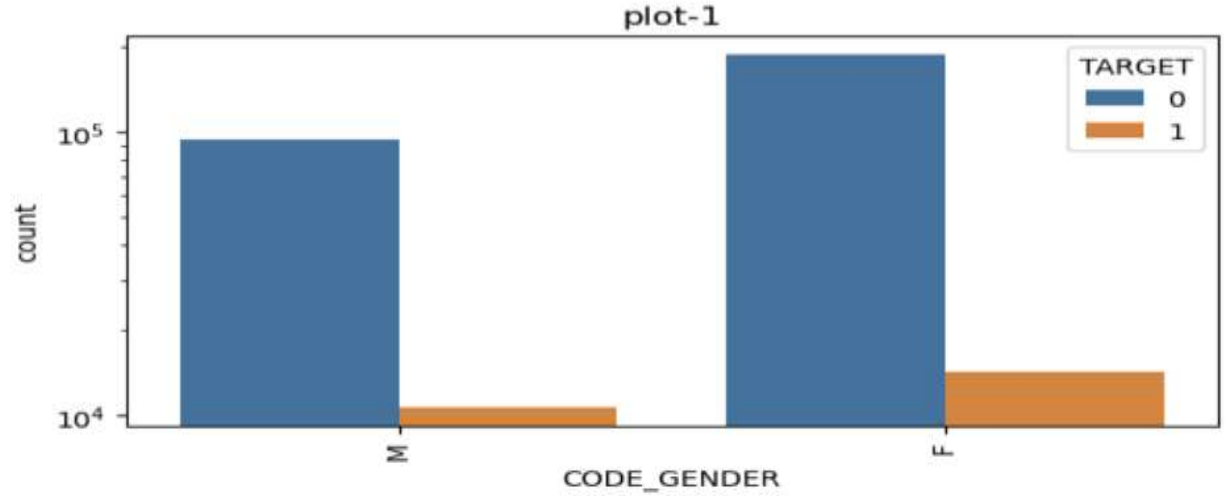
- one of the outlier point is above 350000 days which is roughly 958 years of employment which is not possible and hence we may conclude that it is a wrong data

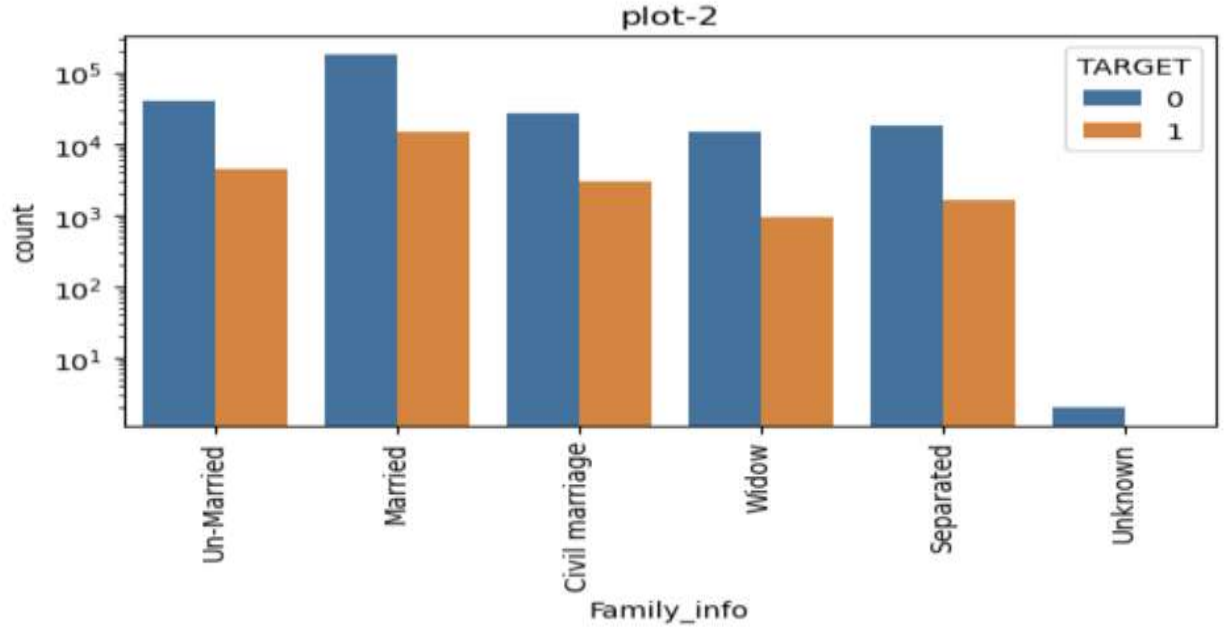


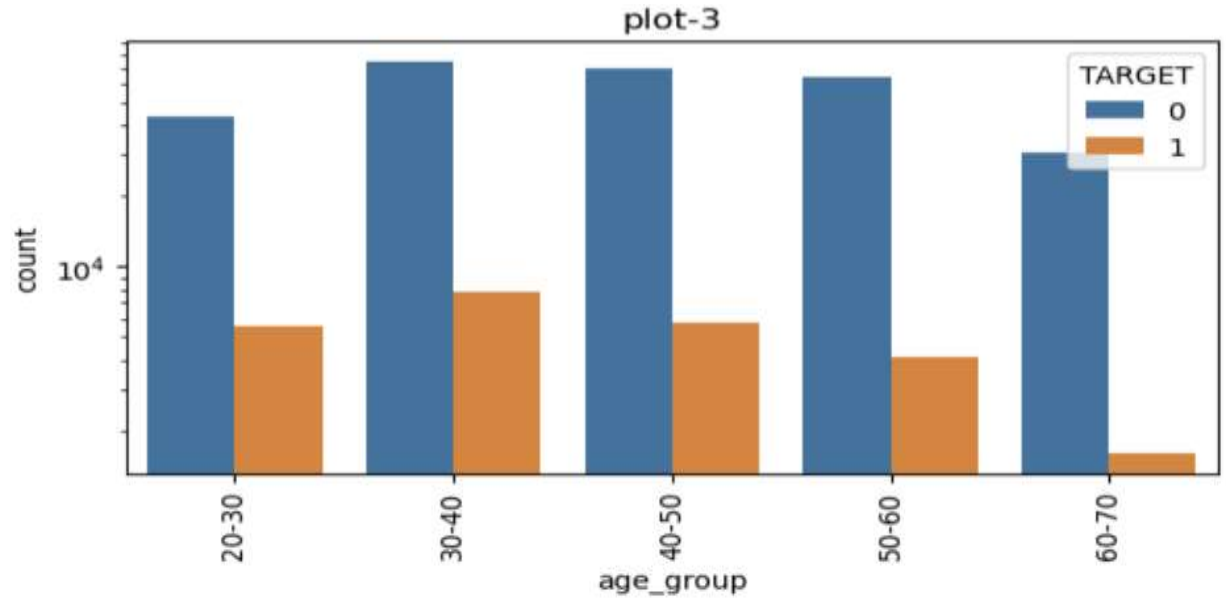
as the max age of applicant is 69 years, the years before which that applicant has changed his registration should be less than 68 years or 25000 days which is true as evident from the boxplot above, hence the data is reliable

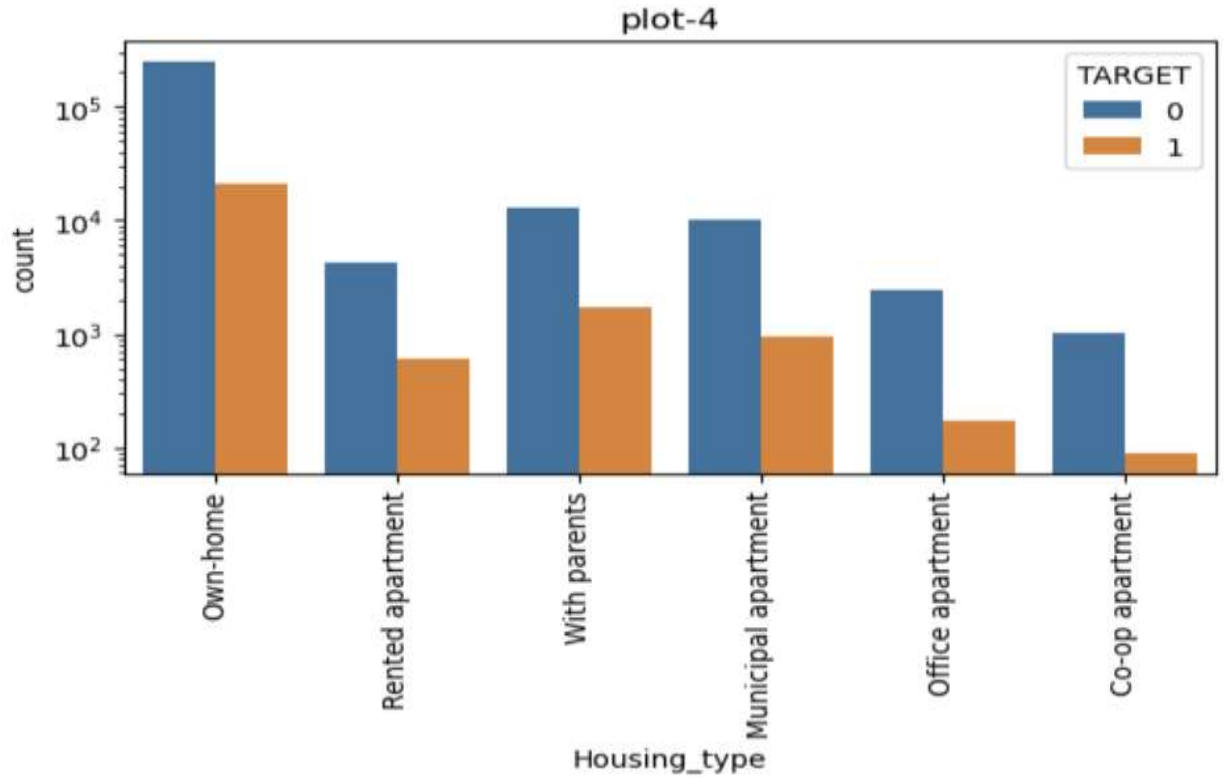


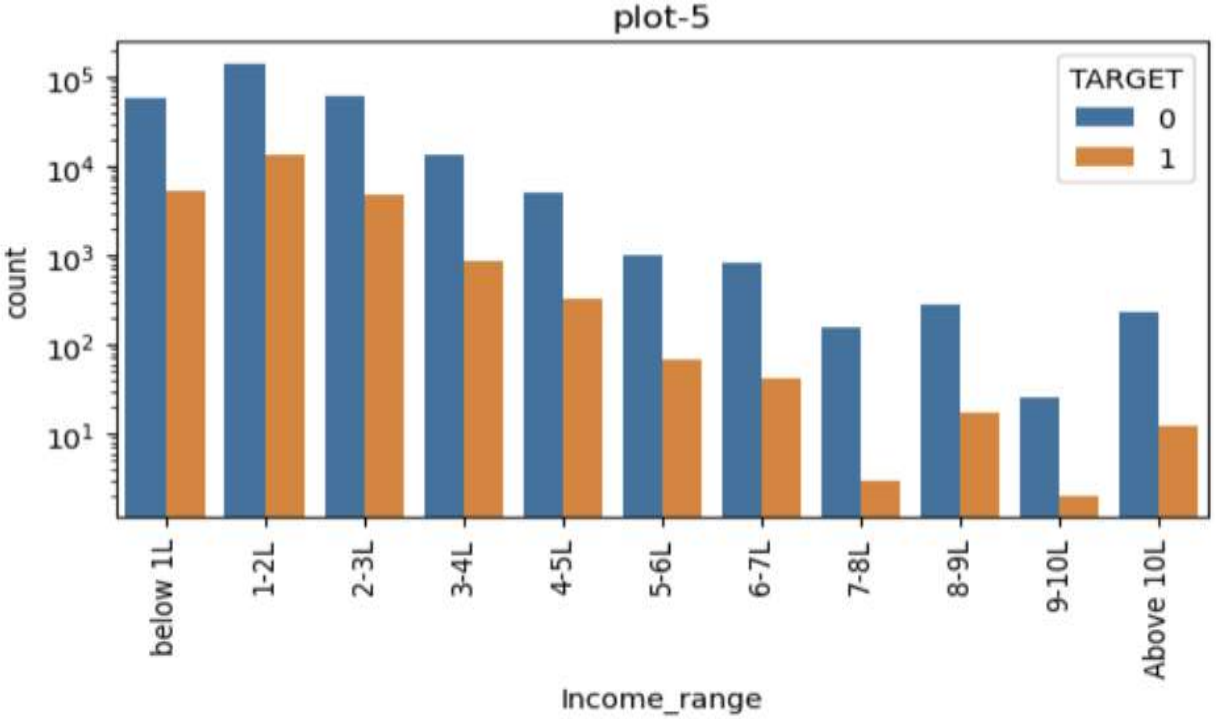
This count shows that there is a huge variation in number of defaulters and repayers

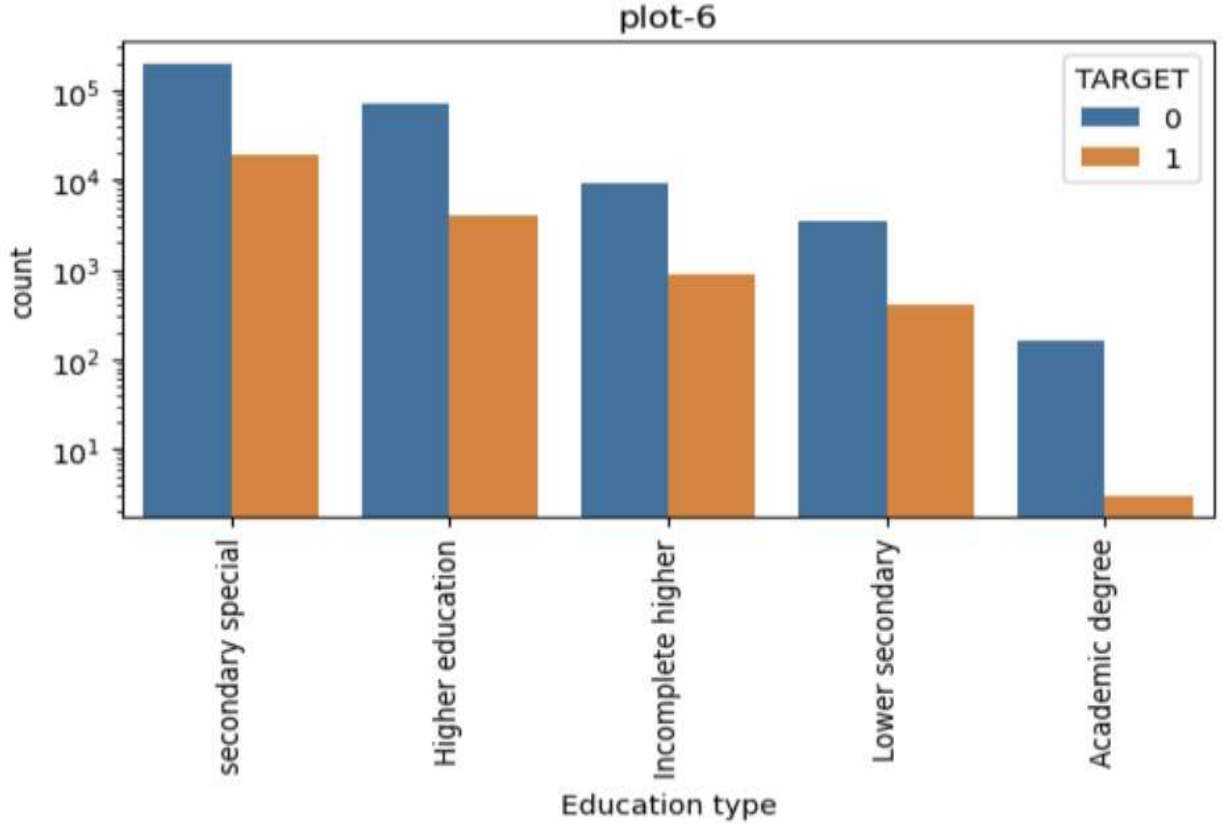

plot-1
TARGET
0
1
count
10^5
10^4
M
F
CODE_GENDER


plot-2
TARGET
0
1
count
10^5
10^4
10^3
10^2
10^1
Un-Married
Married
Civil marriage
Widow
Separated
Unknown
Family_info


plot-3
TARGET
0
1
count
10^4
20-30
30-40
40-50
50-60
60-70
age_group

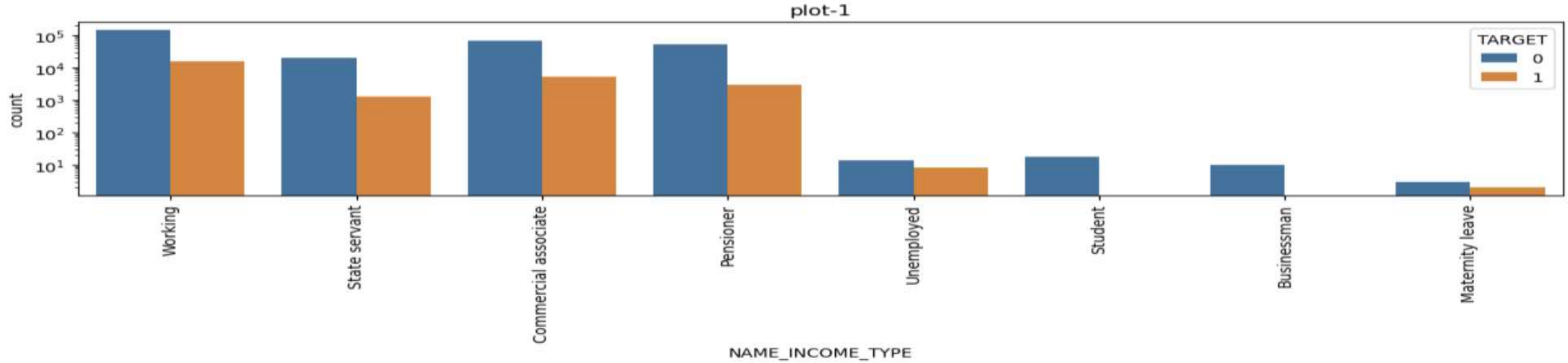

plot-4
TARGET
0
1
count
10^5
10^4
10^3
10^2
Own-home
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
Housing_type

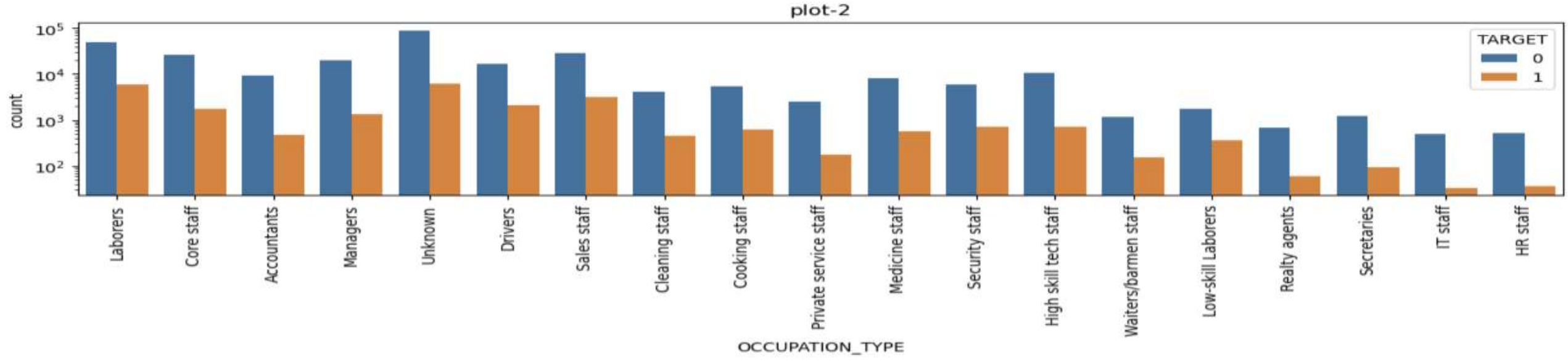




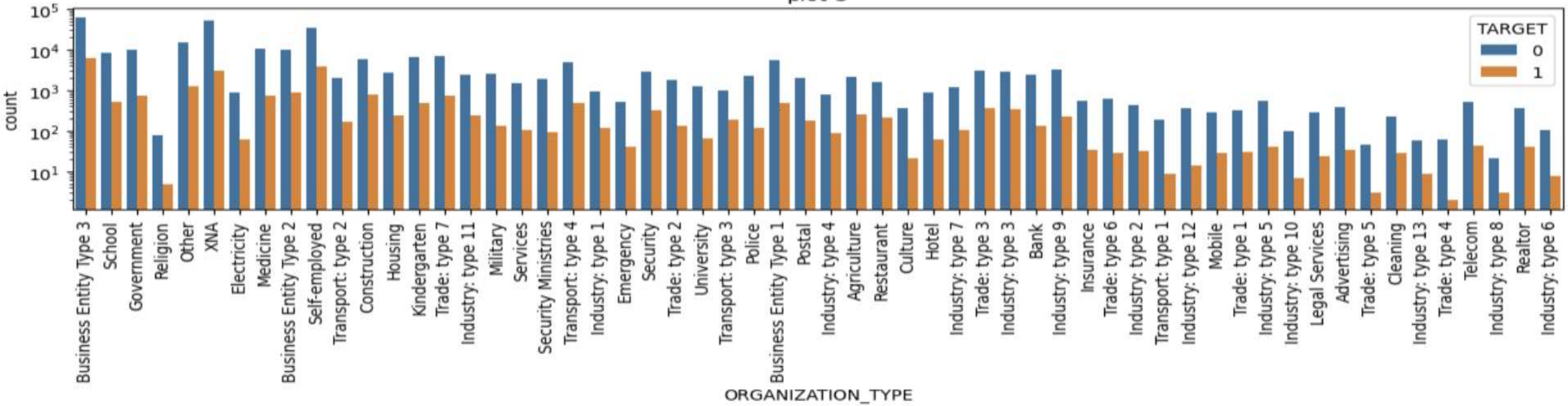
NOTE: inference from the plots above

- plot-1: female clients are likely to default than male clients (XNA - is present because as instructed, I haven't handled the missing values and left them as it is)
- plot-2: married clients are likely to default than other clients
- plot-3: clients with age-group between 30-40 are more defaulters, then comes 40-50, and then 50-60
- plot-4: clients with own house/apartment are likely to default than those who stay in rented house or with their parents
- plot-5: clients with income in the range 1-2L are likely to default than clients with other income groups
- plot-6: clients with secondary special education are likely to default.

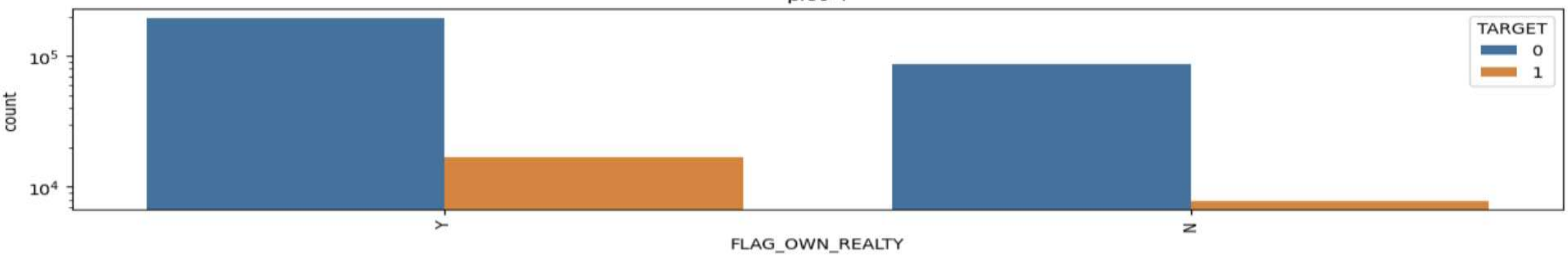

plot-1
count
TARGET
0
1
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


plot-2
count
TARGET
0
1
Laborers
Core staff
Accountants
Managers
Unknown
Drivers
Sales staff
Cleaning staff
Cooking staff
Private service staff
Medicine staff
Security staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE

plot-3


count
10^5
10^4
10^3
10^2
10^1
TARGET
0
1
Business Entity Type 3
School
Government
Religion
Other
XNA
Electricity
Medicine
Business Entity Type 2
Self-employed
Transport: type 2
Construction
Housing
Kindergarten
Trade: type 7
Industry: type 11
Military
Services
Security Ministries
Transport: type 4
Industry: type 1
Emergency
Security
Trade: type 2
University
Transport: type 3
Police
Business Entity Type 1
Postal
Industry: type 4
Agriculture
Restaurant
Culture
Hotel
Industry: type 7
Trade: type 3
Industry: type 3
Bank
Industry: type 9
Insurance
Trade: type 6
Industry: type 2
Transport: type 1
Industry: type 12
Mobile
Trade: type 1
Industry: type 5
Industry: type 10
Legal Services
Advertising
Trade: type 5
Cleaning
Industry: type 13
Trade: type 4
Telecom
Industry: type 8
Realtor
Industry: type 6
ORGANIZATION_TYPE

plot-4


count
10^5
10^4
TARGET
0
1
Y
N
FLAG_OWN_REALTY

NOTE: inferences from the above plots

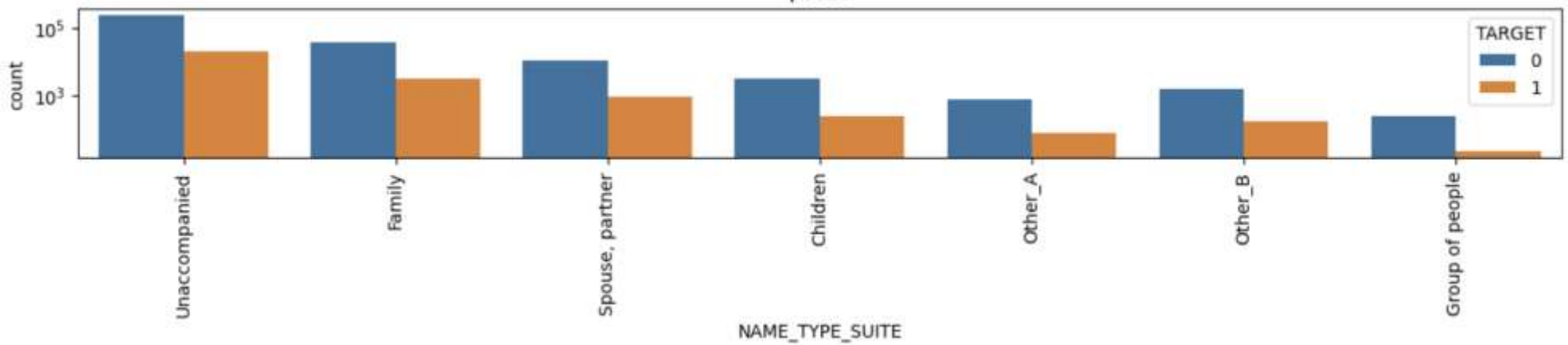
•plot-1: working clients are likely to default.

•plot-2: clients who haven't specified their occupation type are likely defaulters. of those who mentioned their occupation, labourers are likely defaulters.

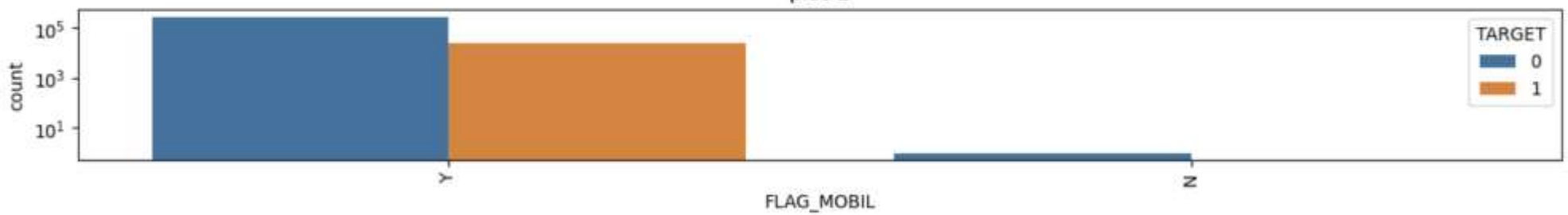
•plot-3: clients who have business entity of type-3 are likely defaulters apart from those who haven't specified their organization type.

•plot-4: clients who have their own Real-estate property are mostly turning defaulters than those who don't own any real-estate

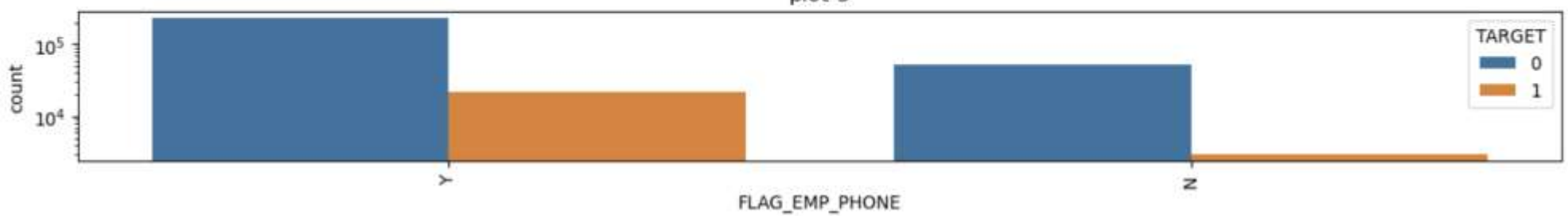
plot-1



plot-2



plot-3

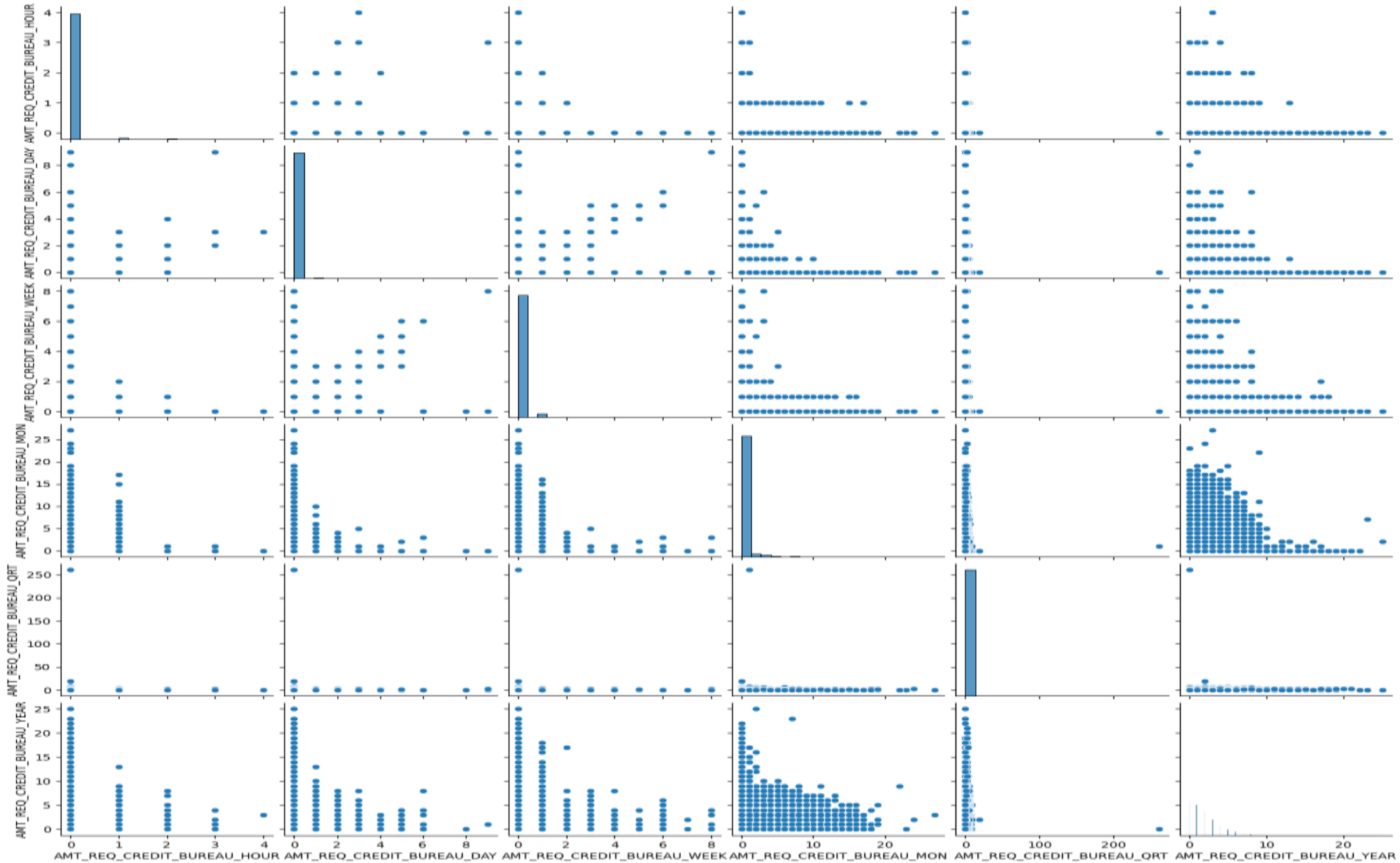


NOTE: inferences from the above plots based on contact info

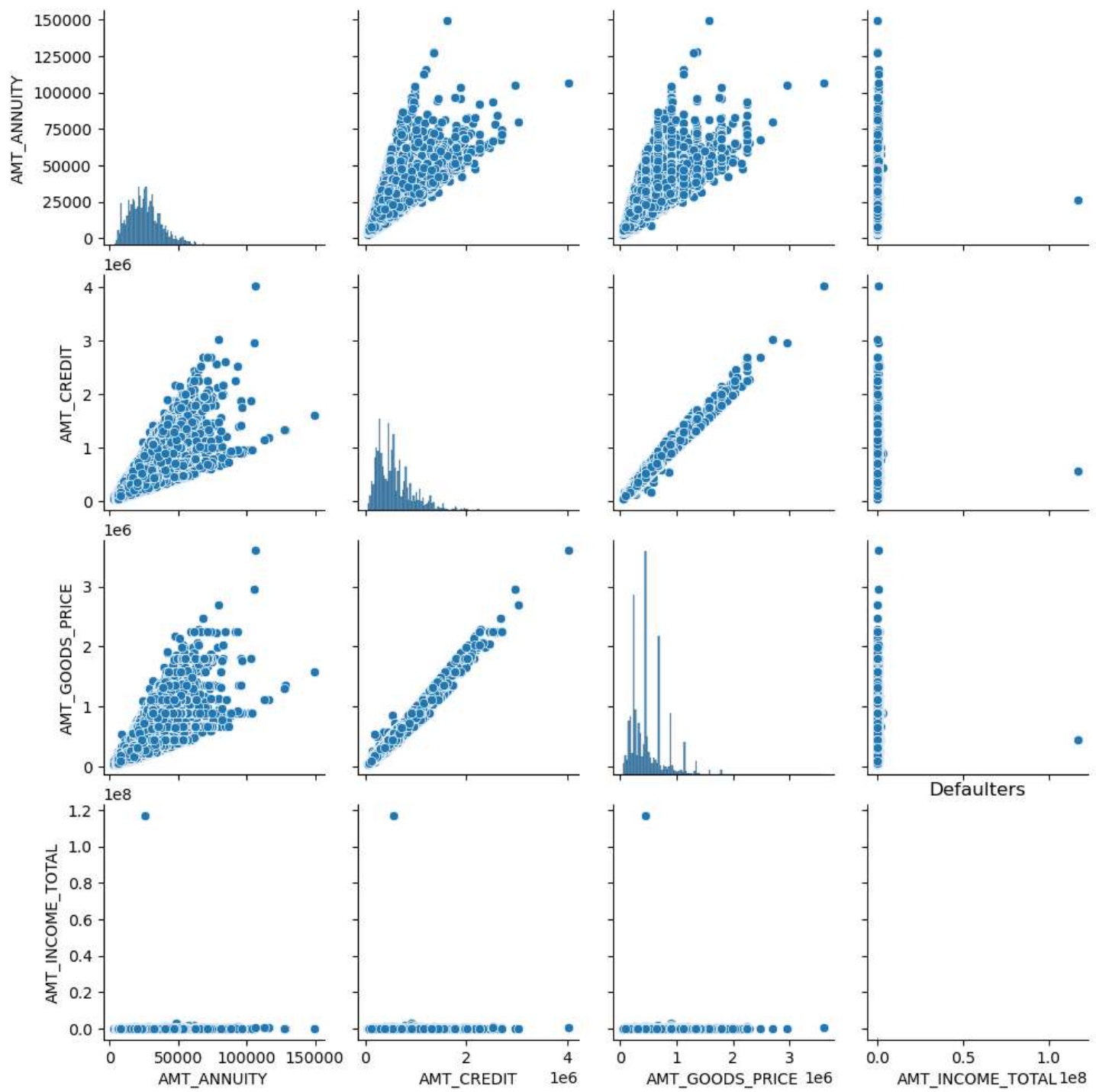
•plot-1: clients who were not accompanied by anyone are likely to become defaulters.

•plot-2: clients who provided their phone number are defaulting more than those who don't.

•plot-3: clients who provided their work phone number are defaulting more than those who don't.

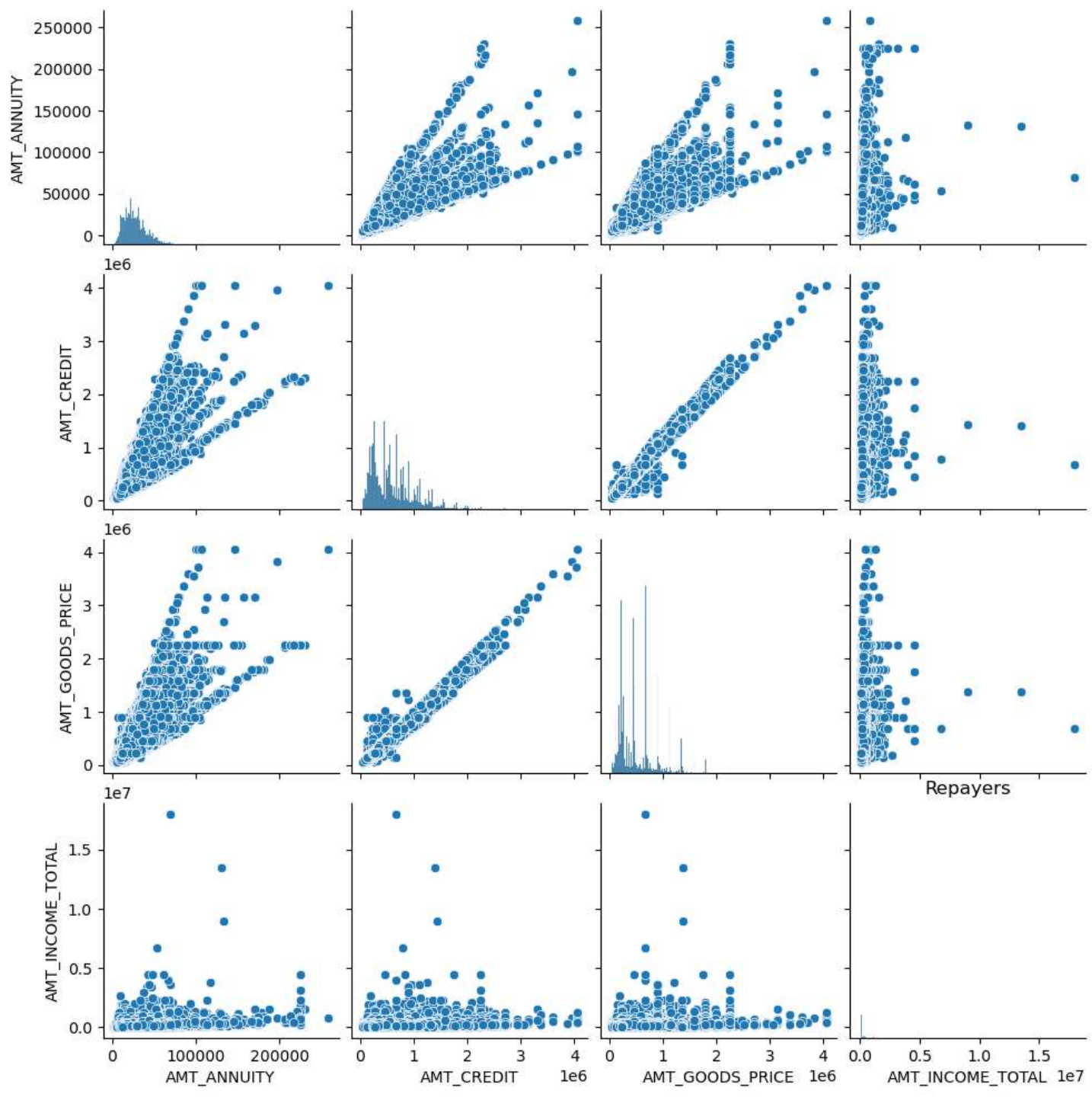


'AMT_REQ_CREDIT_BUREAU_HOUR','AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR', have no proper relation between them.



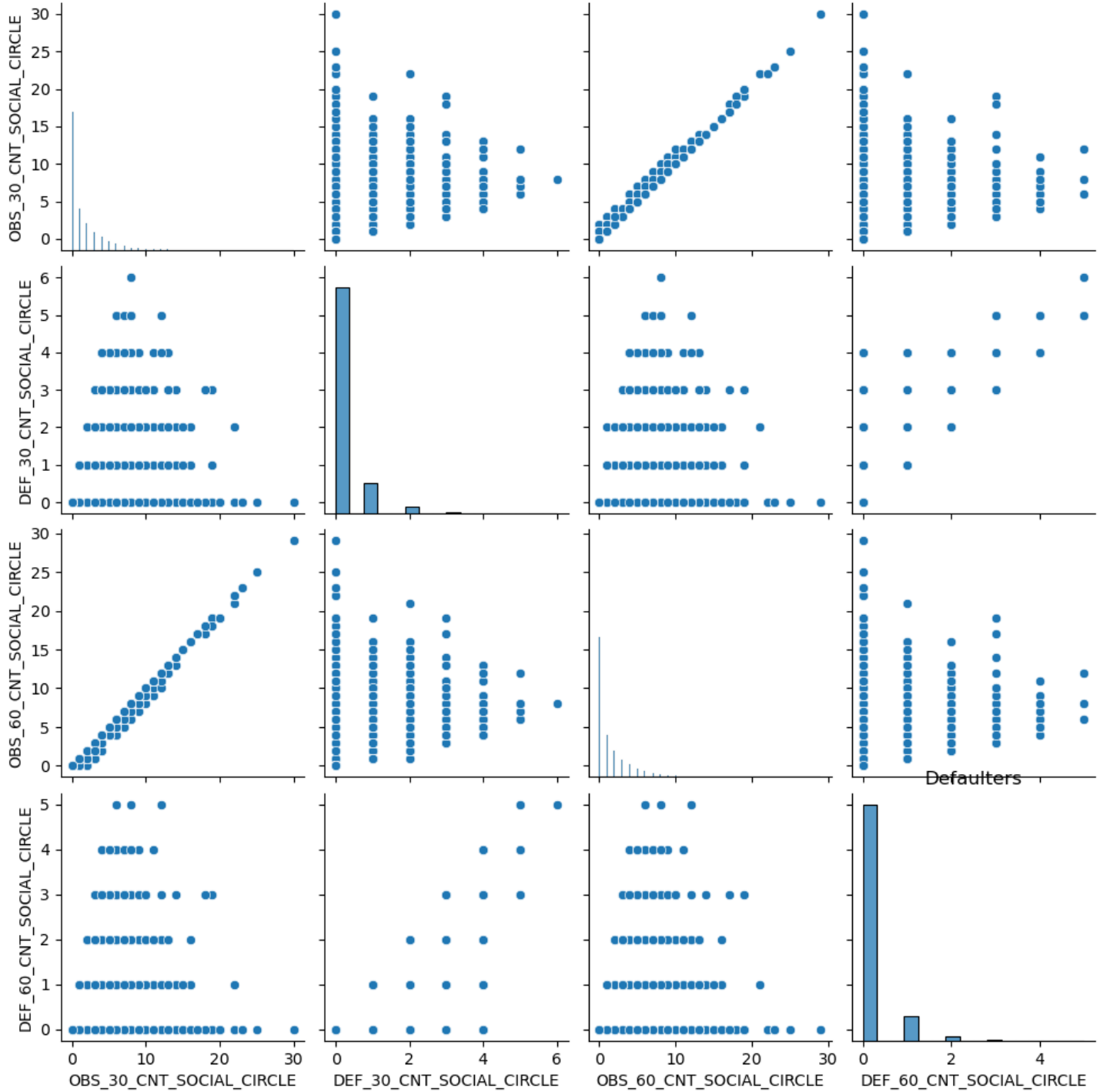
for Defaulters:

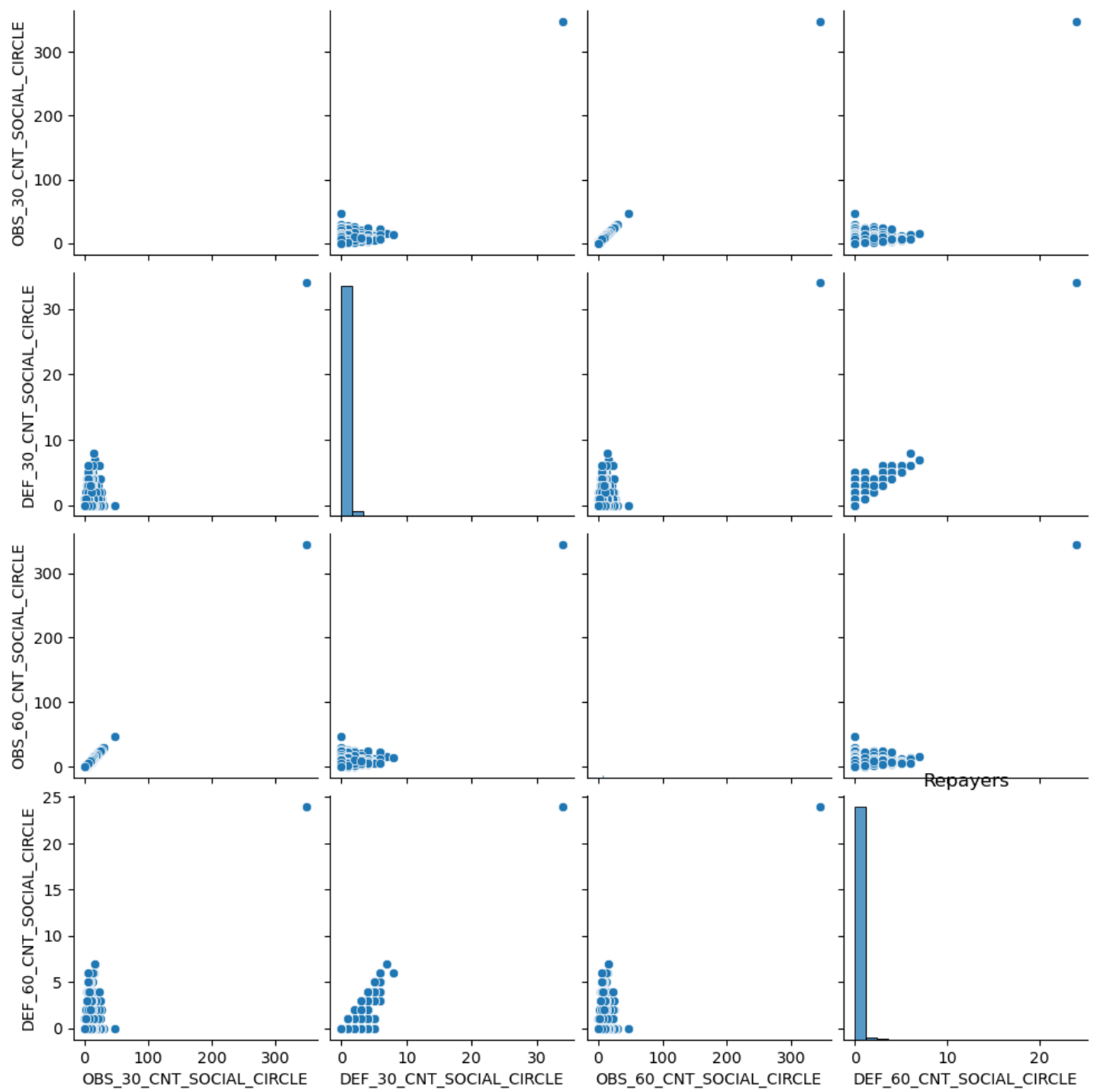
* goods_price and Credit amount have some linear relation among them
* goods-price and credit amount show some similar sort of relationship with Annuity amount



for repayers:

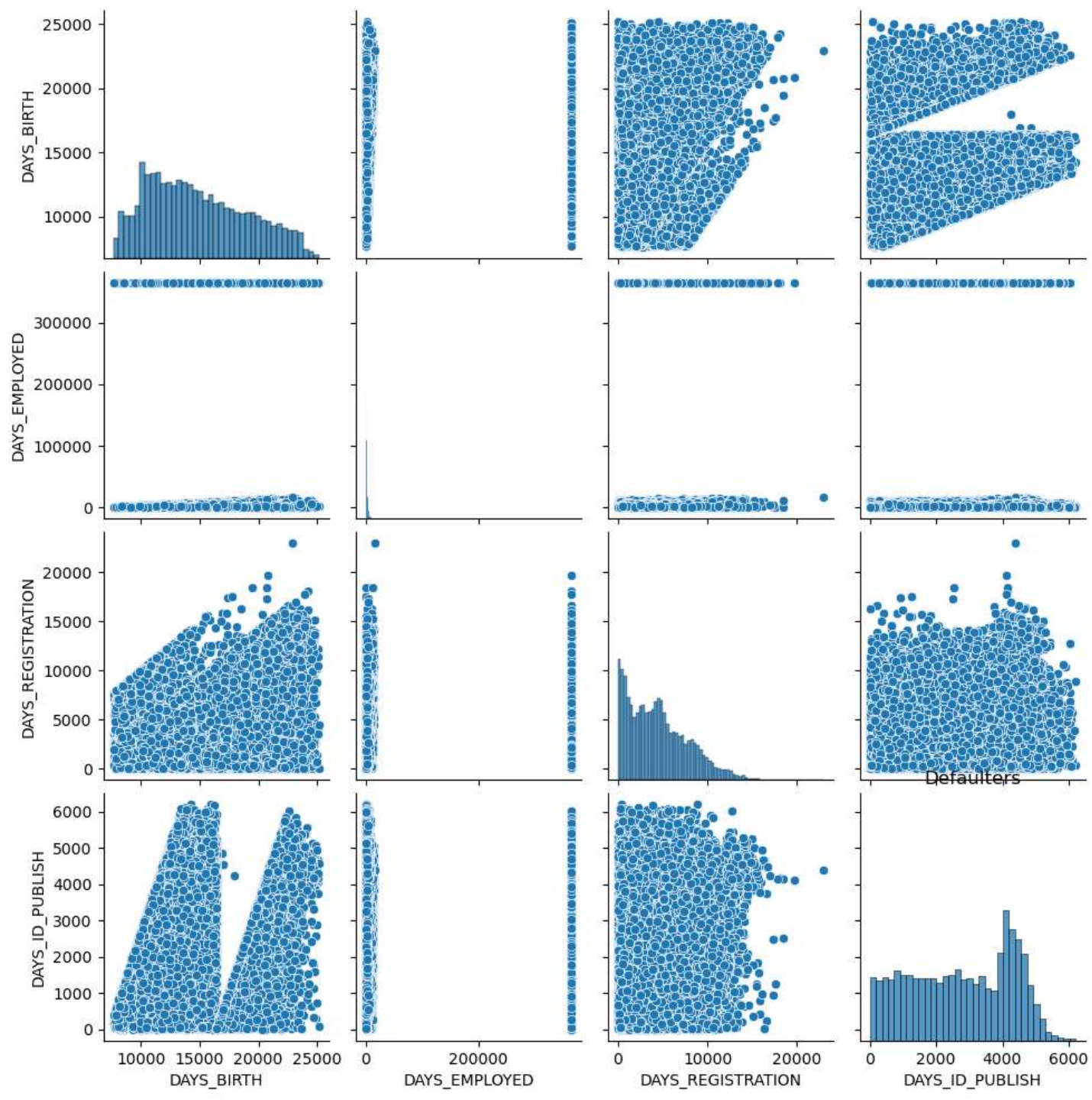
•similar kind of patterns are observed in case of both Repayers and Defaulters. but the scatter points in Repayers are more as the data is imbalanced.

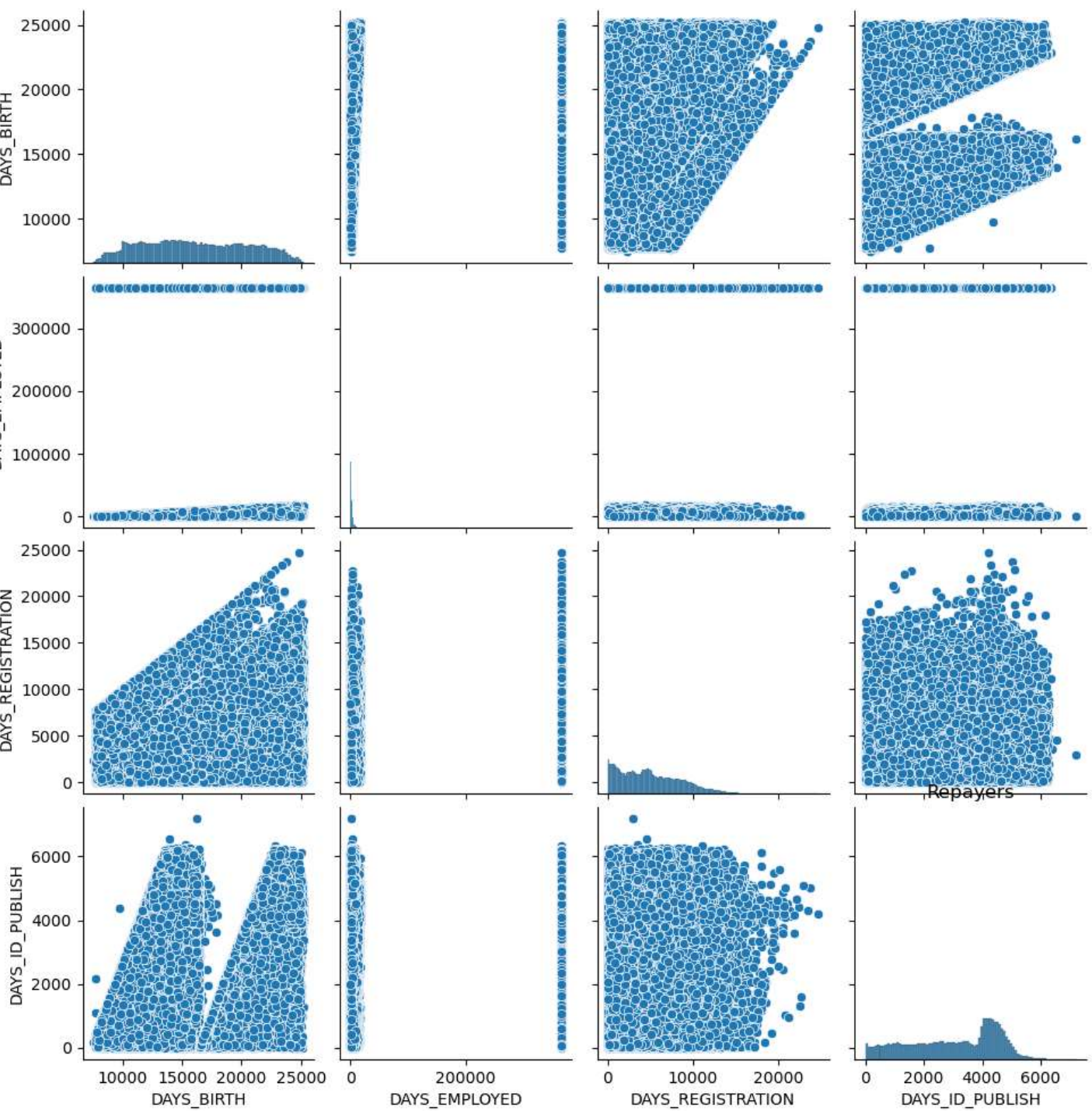

OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
Defaulters



NOTE:

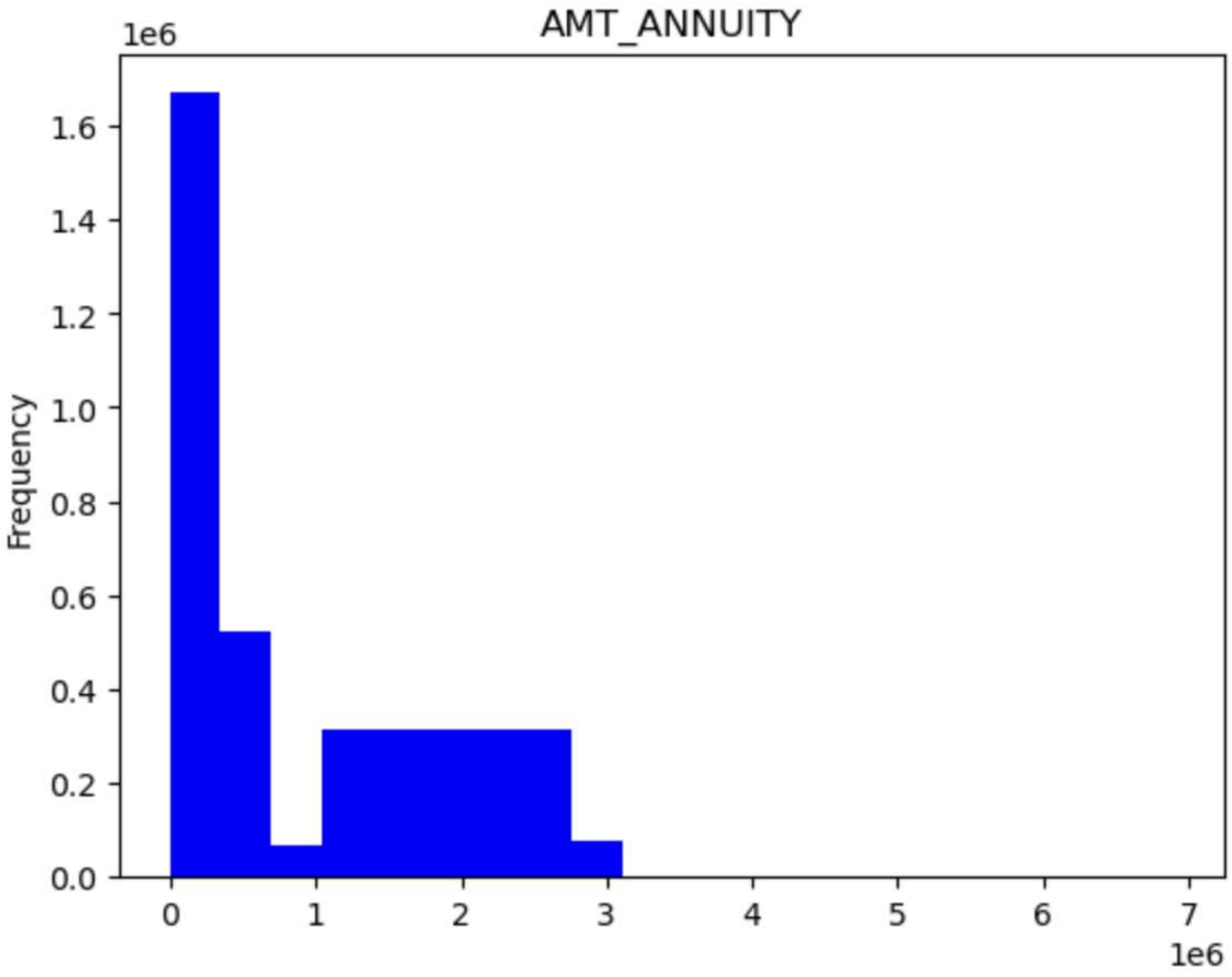
•in both the cases of Defaulters and Repayers, observable social surroundings of the client 30 DPD and 60 DPD show a linear relation.


DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
Defaulters



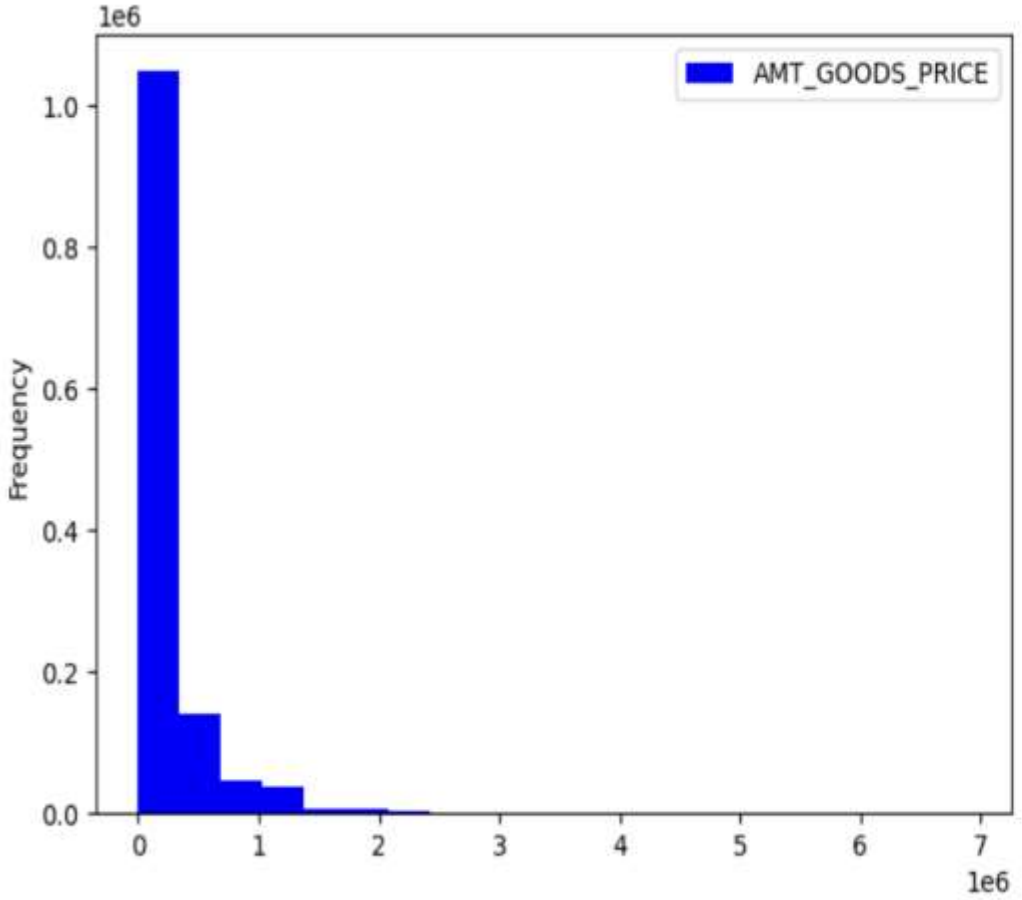
NOTE:

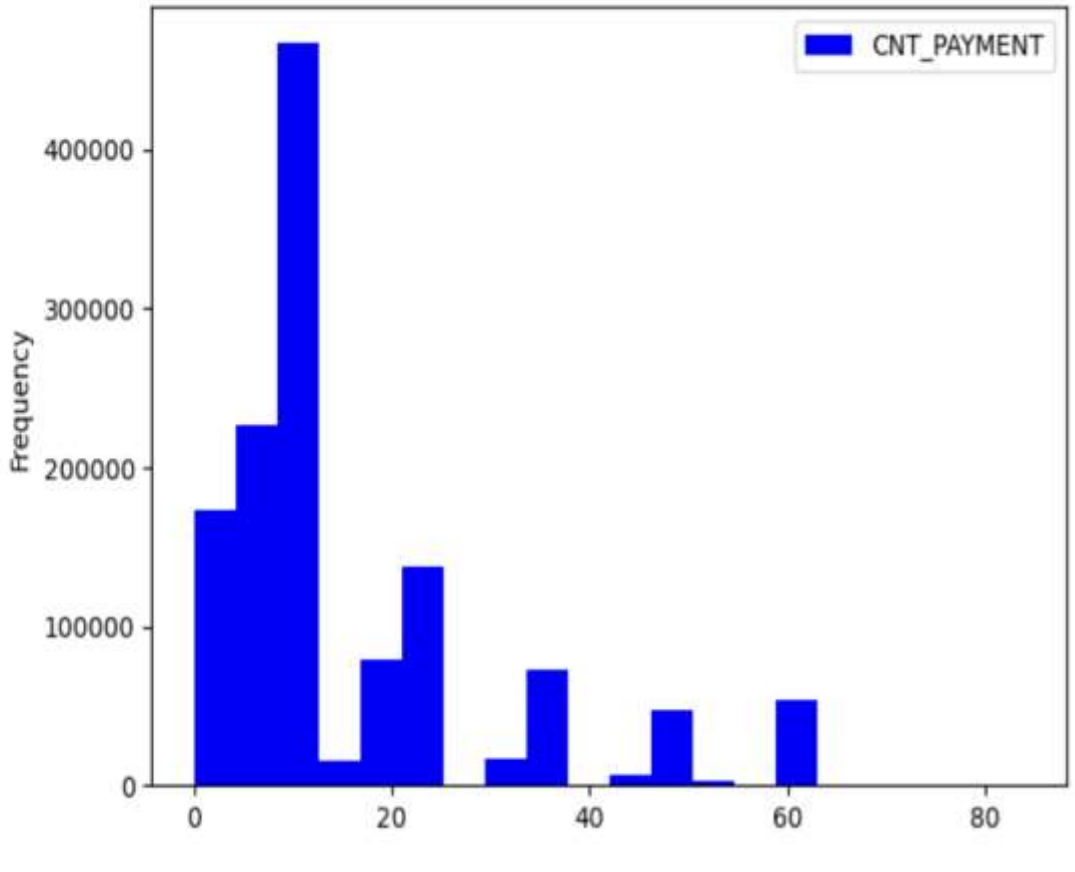
•from the above two pairplots for Defaulters and Repayers, it has been found that there is not specific relation between 'DAYS_BIRTH','DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH'



NOTE:

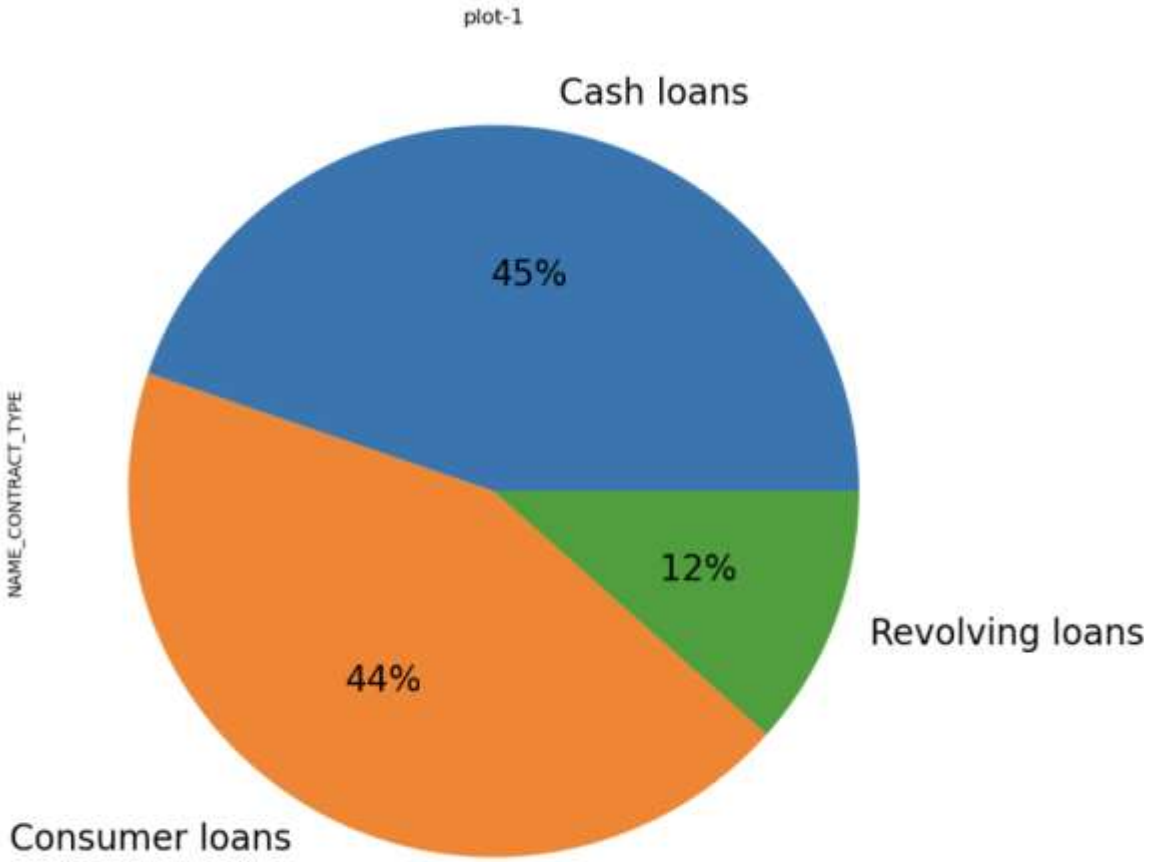
the hist plot of AMT_ANNUITY doesn't show any clear trend of Gaussian curve so there is no clear way that I am able to decide how to deal with null values in AMT_ANNUITY. hence I am also dropping this column.

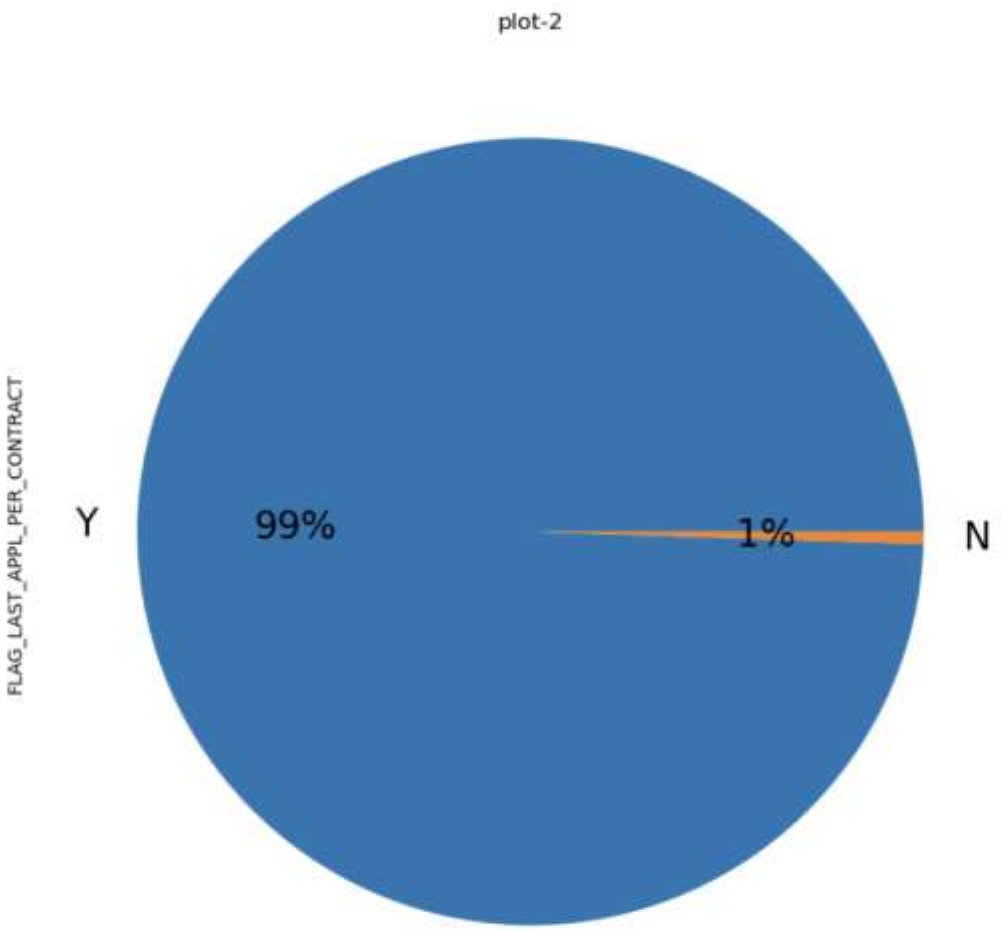




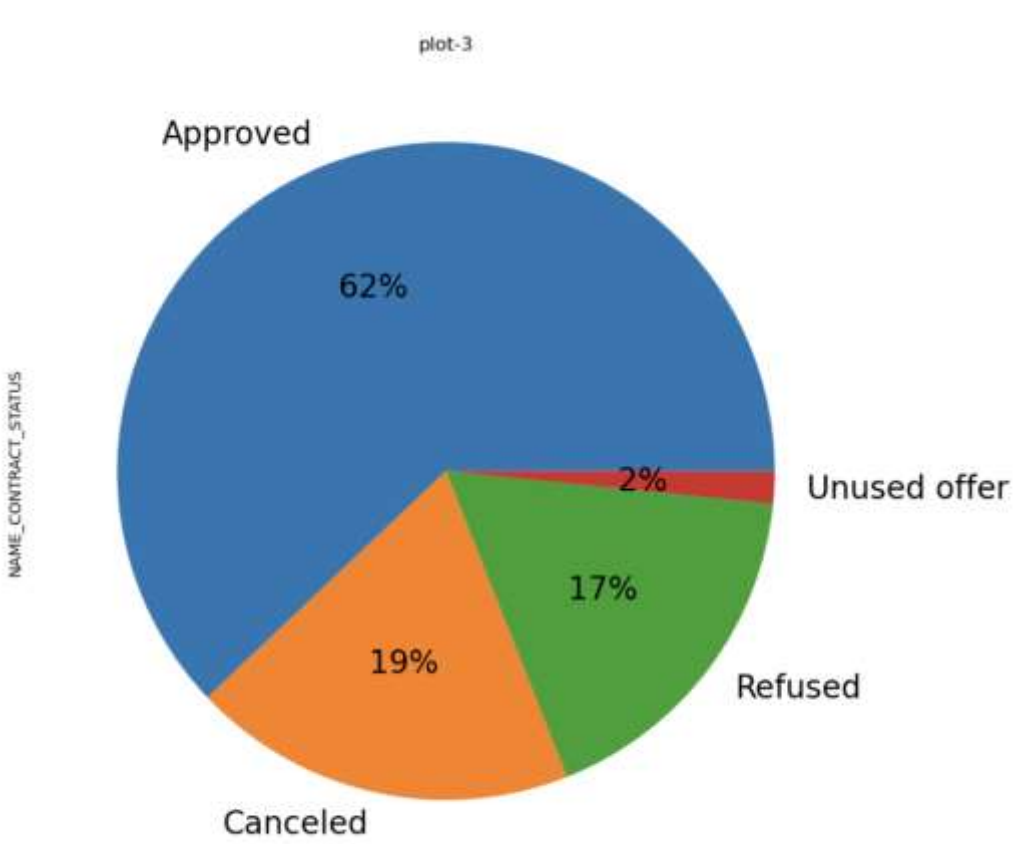
NOTE:

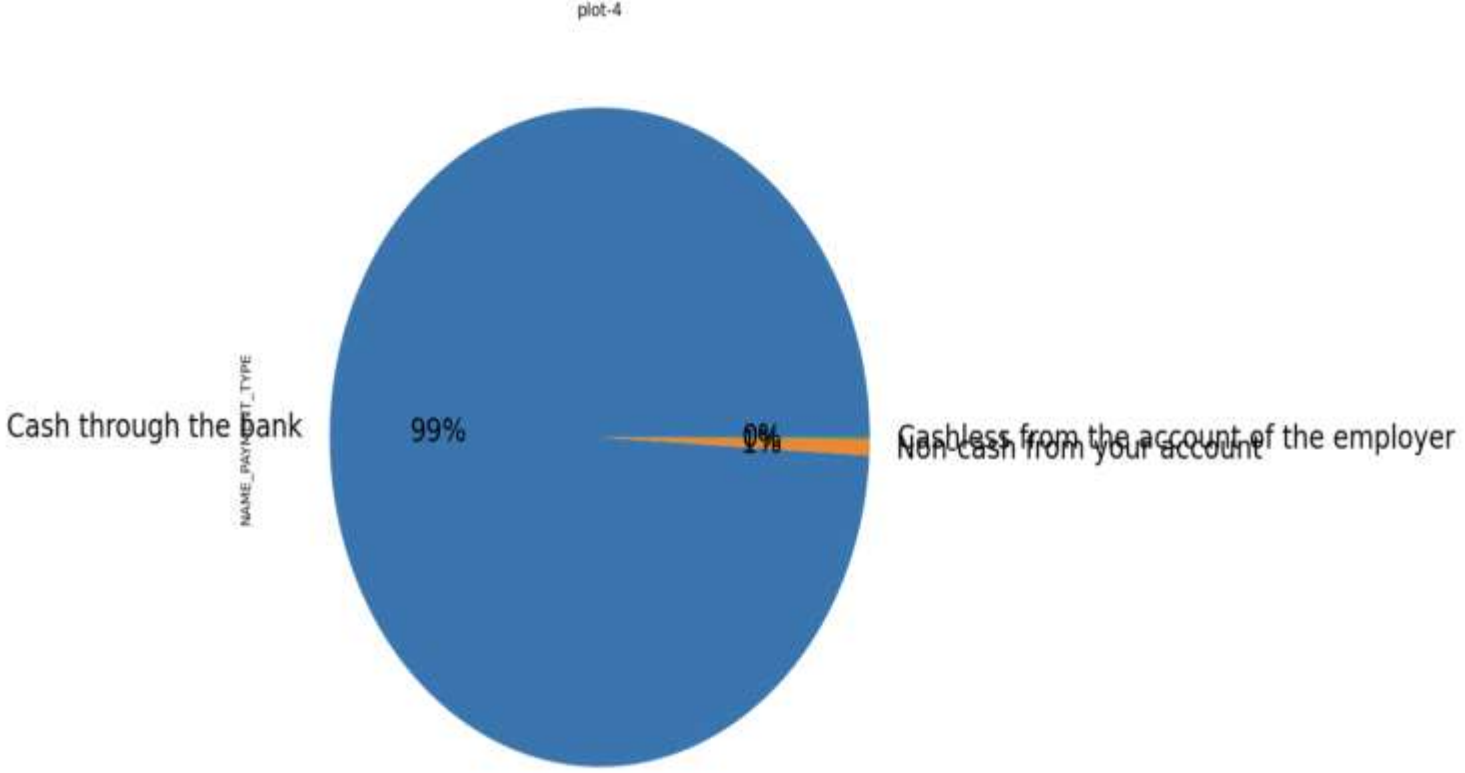
since a similar kind of trend is observed in case of AMT_GOODS_PRICE and CNT_PAYMENT, as in case of AMT_ANNUITY, we are going to drop these two columns as well





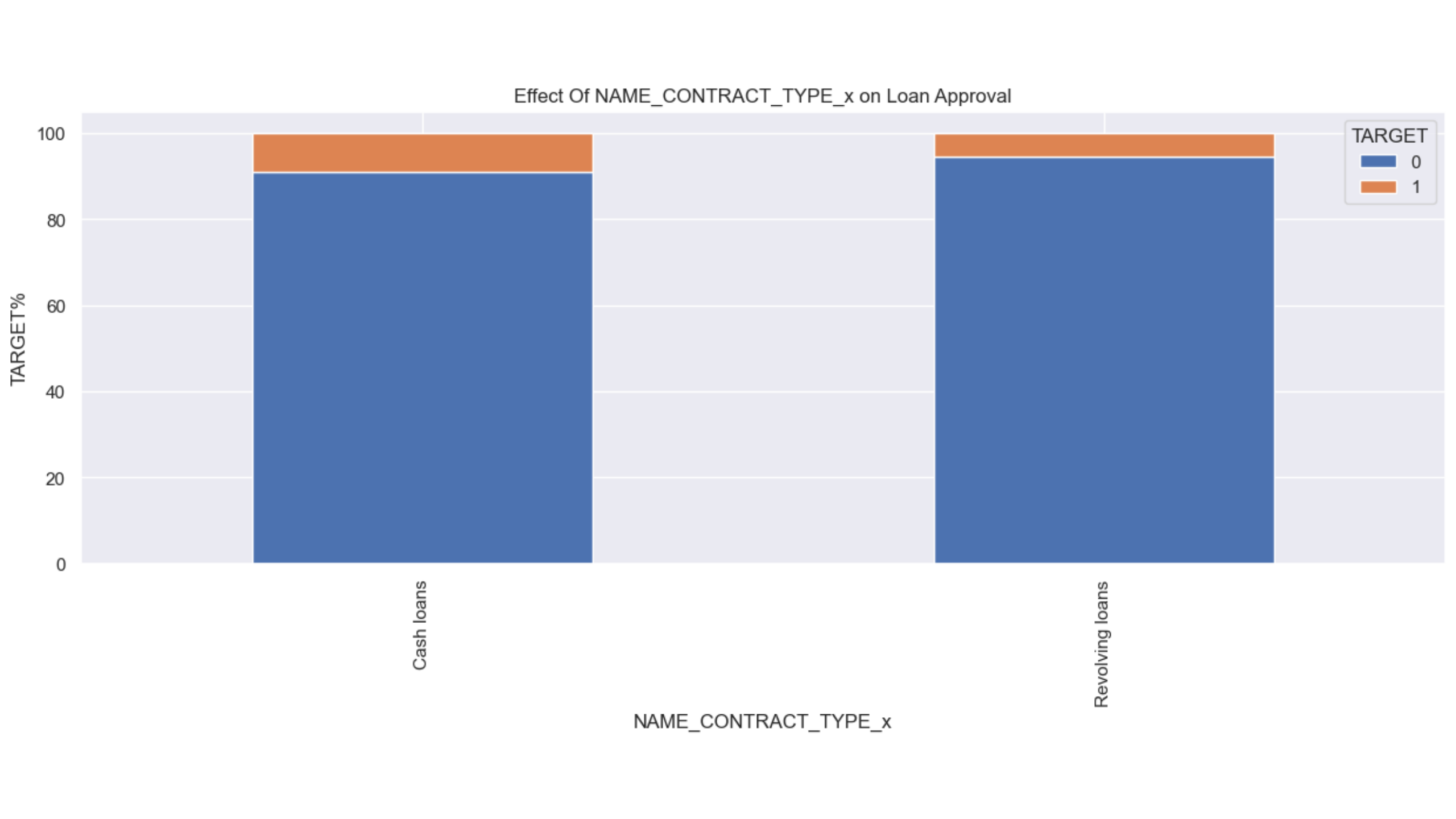
•plot-1: % of cash loans > consumer loans> revolving loans
•plot-2: FLAG_LAST_APPL_PER_CONTRACT is too unbalanced data and can be dropped

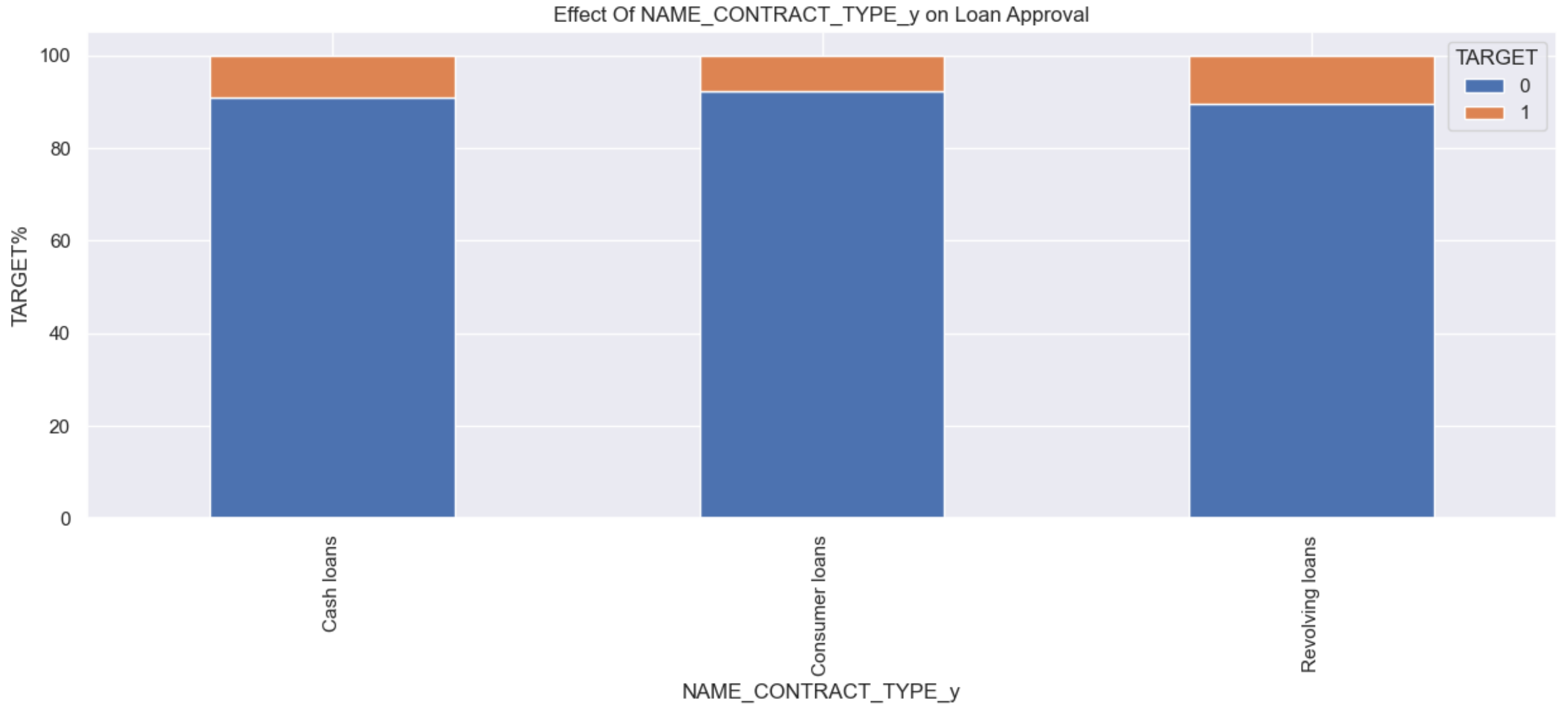




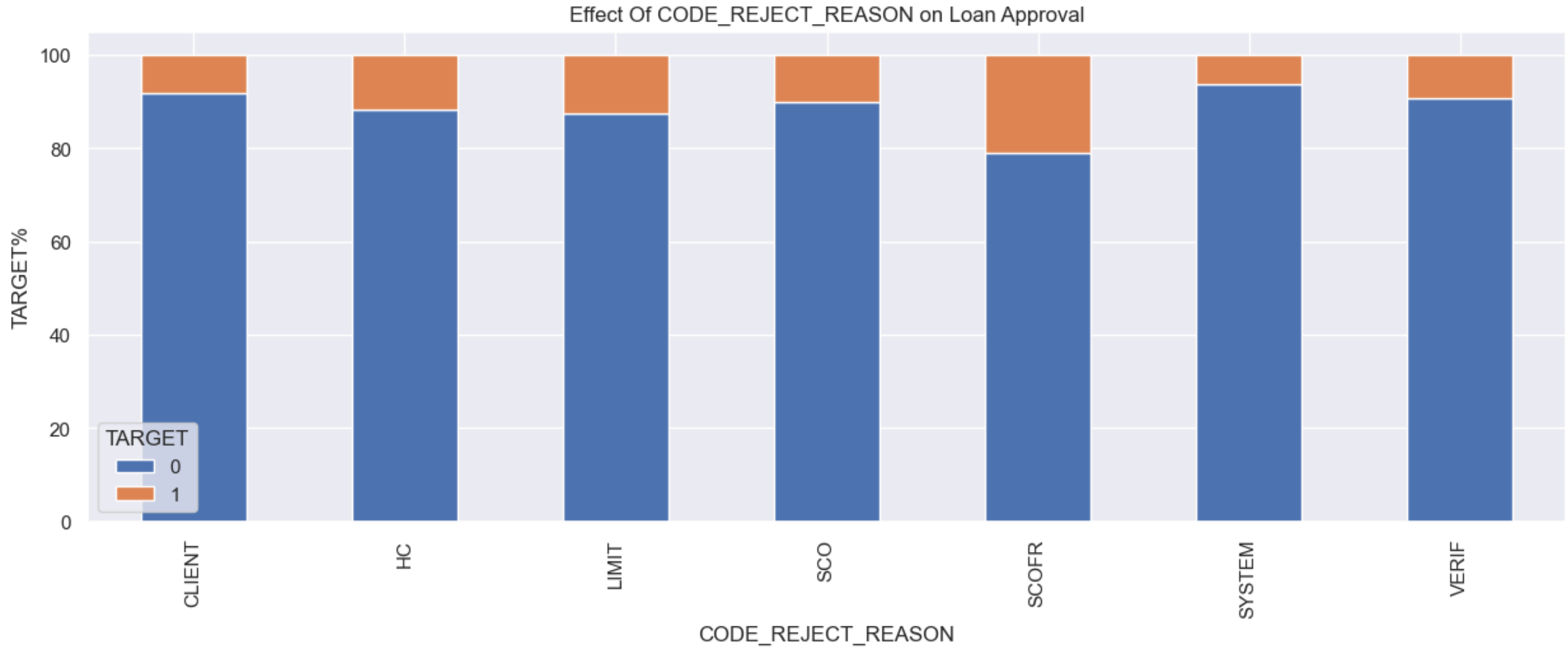
•plot-3: NAME_CONTRACT_STATUS - 62% of clients had their previous applications approved

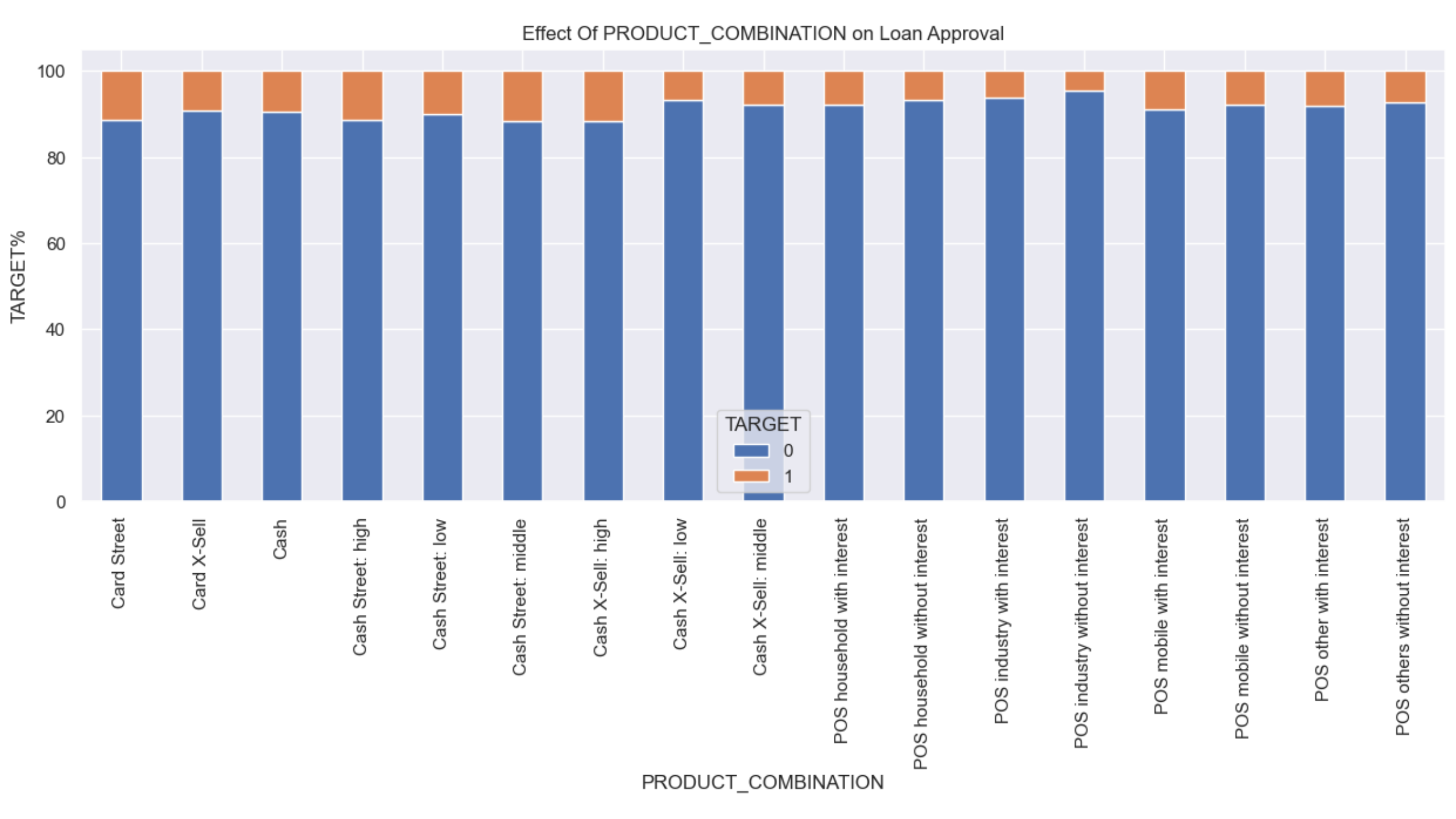
•plot-4: NAME_PAYMENT_TYPE is also too unbalanced and can be dropped


Effect Of NAME_CONTRACT_TYPE_x on Loan Approval
TARGET
0
1
100
80
60
40
20
0
TARGET%
Cash loans
Revolving loans
NAME_CONTRACT_TYPE_x


Effect Of NAME_CONTRACT_TYPE_y on Loan Approval
TARGET
0
1
TARGET%
100
80
60
40
20
0
Cash loans
Consumer loans
Revolving loans
NAME_CONTRACT_TYPE_y


Effect Of Goods_price_range on Loan Approval
TARGET%
100
80
60
40
20
0
TARGET
0
1
below 1L
1-2L
2-3L
3-4L
4-5L
5-6L
6-7L
7-8L
8-9L
9-10L
Above 10L
Goods_price_range


Effect Of CODE_REJECT_REASON on Loan Approval
TARGET%
100
80
60
40
20
0
TARGET
0
1
CLIENT
HC
LIMIT
SCO
SCOFR
SYSTEM
VERIF
CODE_REJECT_REASON


Effect Of PRODUCT_COMBINATION on Loan Approval
TARGET%
100
80
60
40
20
0
TARGET
0
1
Card Street
Card X-Sell
Cash
Cash Street: high
Cash Street: low
Cash Street: middle
Cash X-Sell: high
Cash X-Sell: low
Cash X-Sell: middle
POS household with interest
POS household without interest
POS industry with interest
POS industry without interest
POS mobile with interest
POS mobile without interest
POS other with interest
POS others without interest
PRODUCT_COMBINATION


Effect Of Income_range on Loan Approval
TARGET%
100
80
60
40
20
0
TARGET
0
1
below 1L
1-2L
2-3L
3-4L
4-5L
5-6L
6-7L
7-8L
8-9L
9-10L
Above 10L
Income_range

NOTE: as far as my understanding of the data set from the entire analysis is concerned, the factors that effect the loan approval for any appliant is driven by following attributes that greatly effect the target variable

- NAME_CONTRACT_TYPE_x
- NAME_CONTRACT_TYPE_y
- Income_range of client
- Goods_price_range for which the loan is being taken
- Reject reason for the previous loan
- Product combination of the client

THANK YOU