CREDIT EDA ANALYSIS

Datasets used:

application_data.csv

previous_application.csv

By:Shreejith.S.

Problem Statement

- The credit lending companies face difficulty in lending loans to clients due to non-existence of credit history.
- Because of this issue, companies can be under loss due to 2 scenarios.
- If the client is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan,then approving the loan may lead to a financial loss for the company.
- We are having 2 datasets of interest, application_data.csv which contains the information about if a client has payment difficulty or not.
- Another dataset previous_application.csv has the information about the previous loan data of client.

Approach of analysis

- We initially loaded the dataset into the notebook and checked the summary of the dataset using .info() function.
- We then dropped the columns with more than 40% null values. We selected the number 40% because this is an industry standard and columns with more than this amount(40%) of null values is not useful for drawing insight.
- We also dropped few more columns like series of FLAG_DOCUMENT columns because it’s significance was not clear and wouldn’t add much value to the analysis.
- Data imbalance was calculated by taking the following formula and found to be 11.39

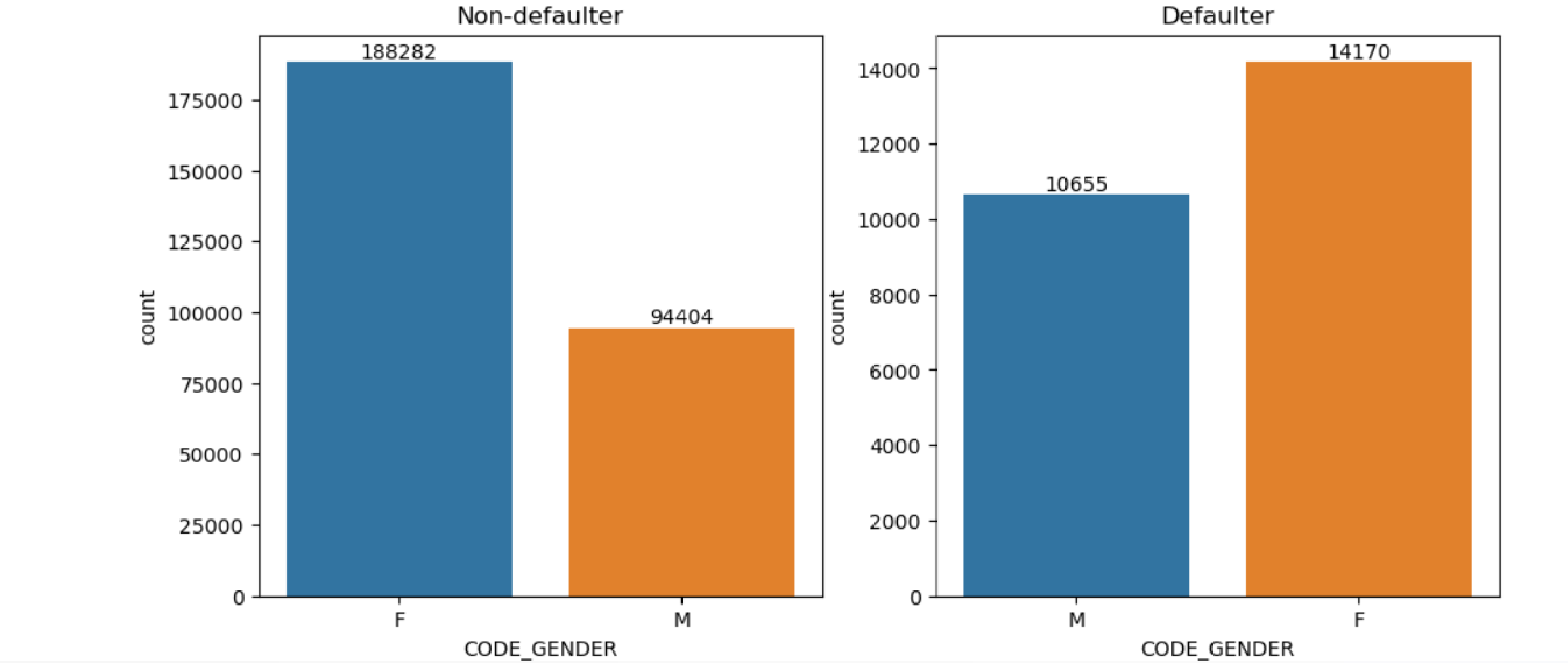
  Data imbalance = (count of non-defaulters)/(count of defaulters)

Approach of analysis

- Then, the DAYS_ column was converted into positive values using abs() function and then into YEARS_ columns.
- Missing values was handled by filling the numerical columns with median and categorical columns with mode of the column
- Outliers were handled by capping the outlier values to upper bound and lower bound values.
- Univariate,Bi-variate analysis and heatmaps were plotted to analyse the dataset.
- We then merged the application_data and previous_application dataset and performed few analysis on them.
- In our analysis we have used the term defaulters for dataframe or data associated with TARGET=1 and non-defaulters for dataframe or data associated TARGET=0.

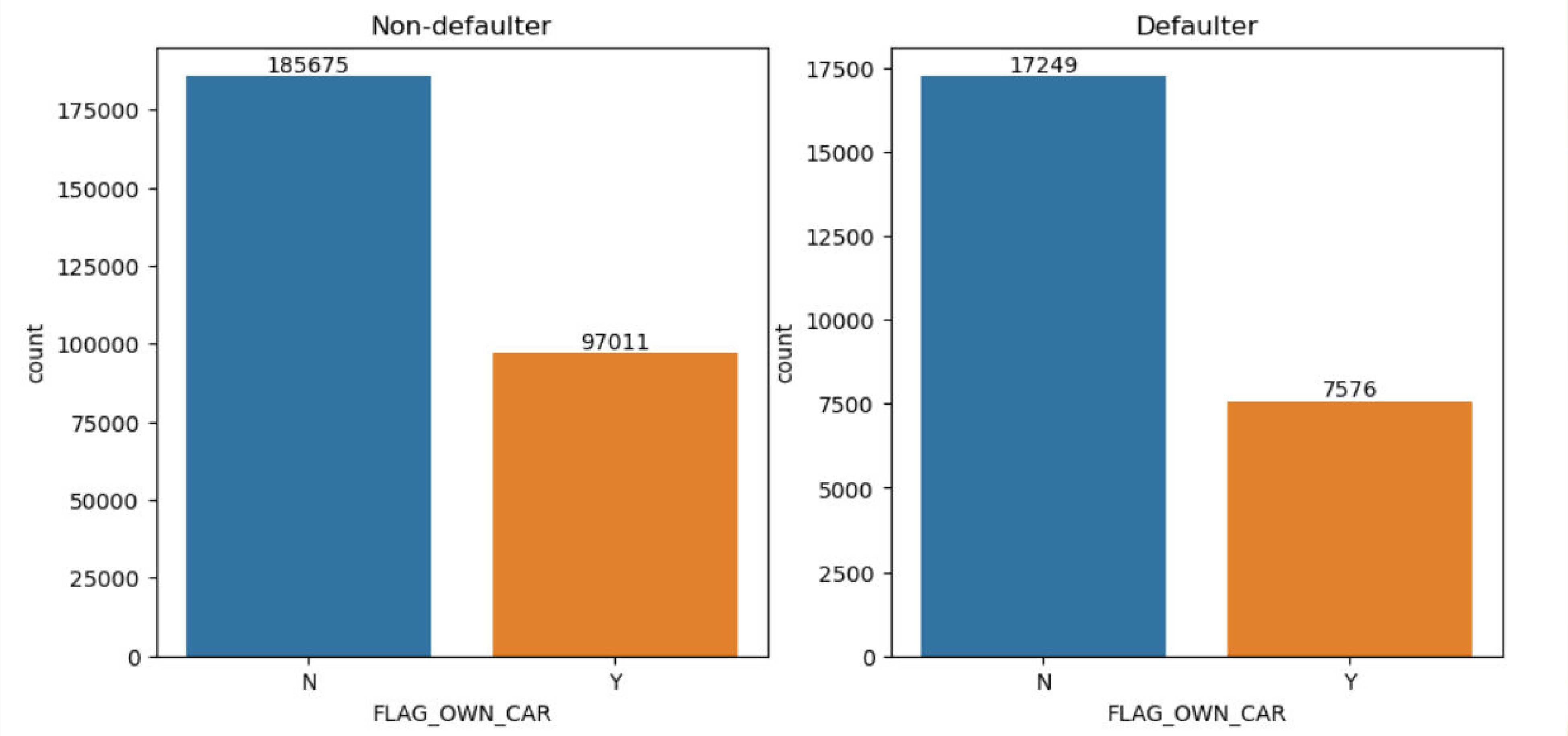
Segmented-Categorical univariate analysis

Distribution of CODE_GENDER



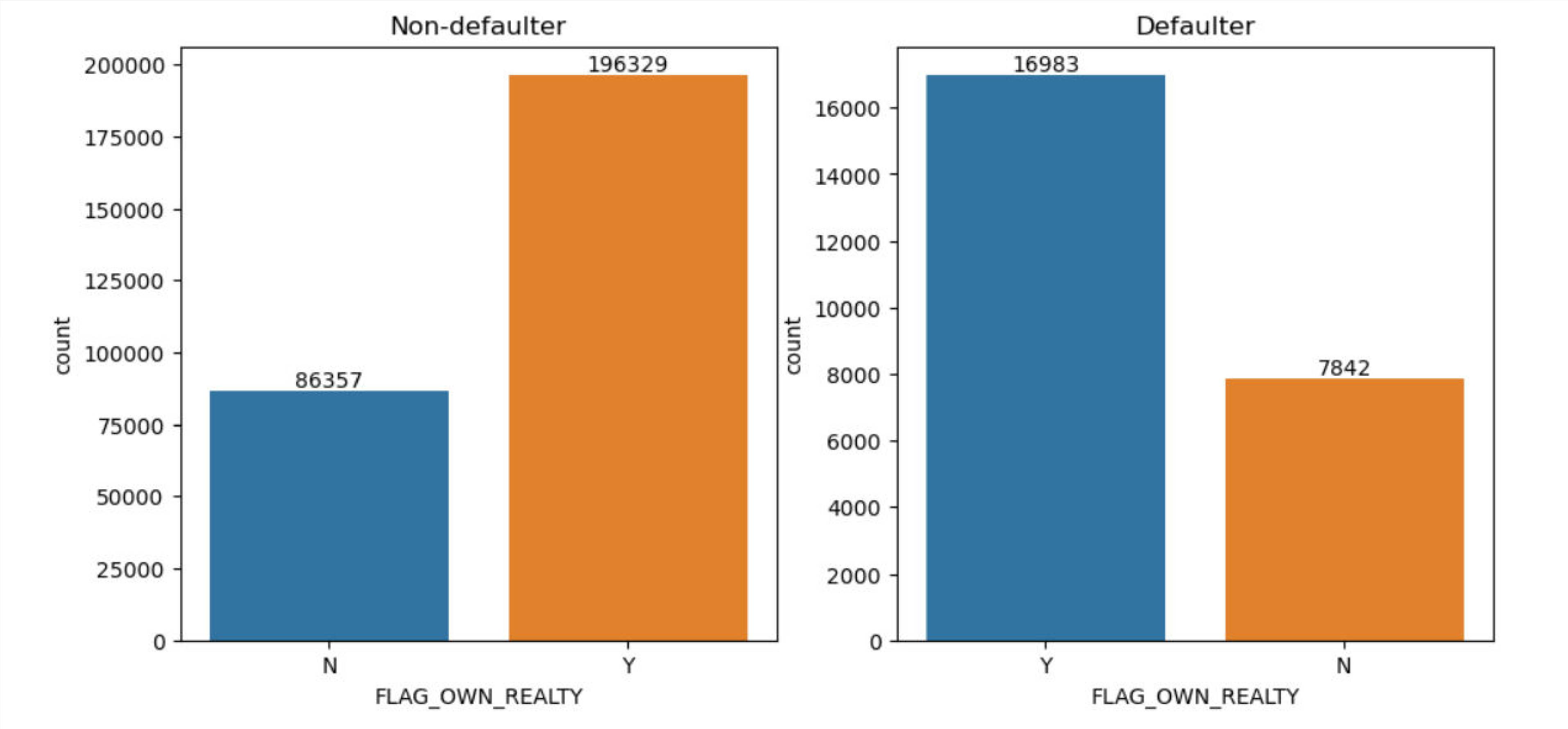
Among the loan applicants, we can see that there are slightly more female defaulters compared to male defaulters. But, the ladies are more serious about paying their loans on time than gents as the count of female non-defaulters is almost the double of male non-defaulters.

Distribution of FLAG_OWN_CAR



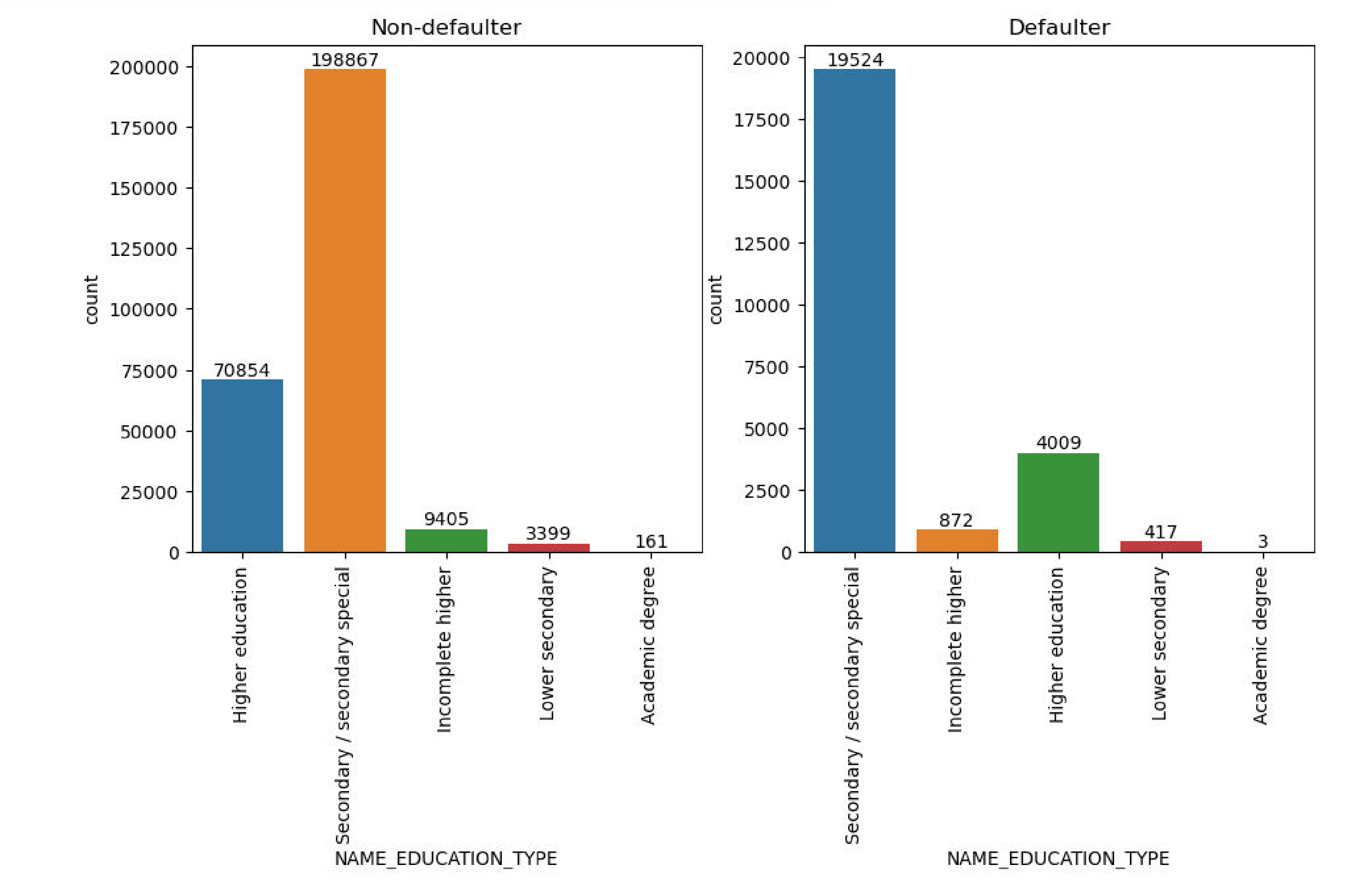
Among the loan applicants, the number of applicants without car are more than double the number of applicants with car. The trend is the same for both defaulters and non-defaulters.

Distribution of FLAG_OWN_REALTY



Both defaulters and non-defaulter applicants owning a house or a flat is much more than the applicants witout a house or flat.

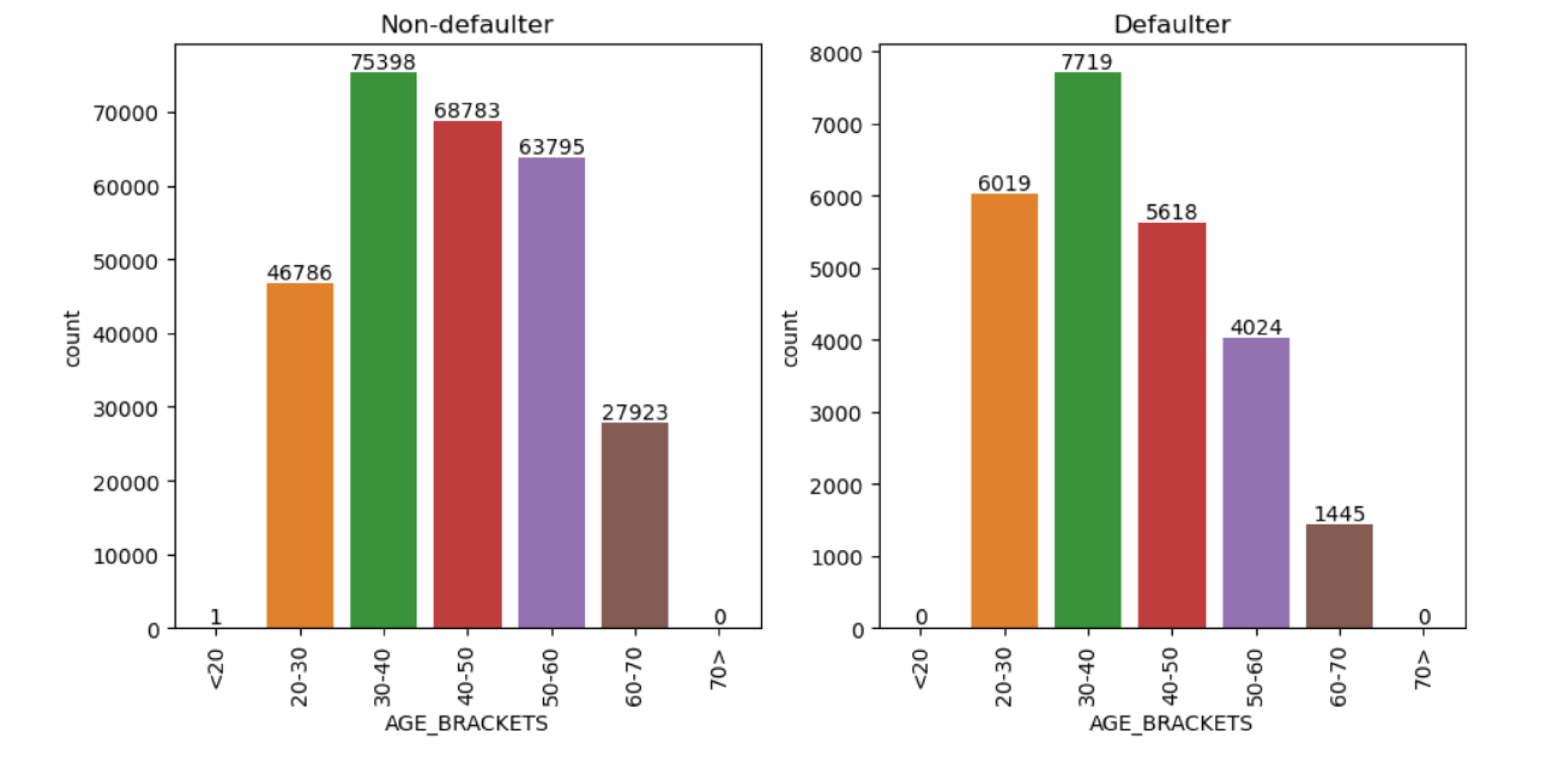
Distribution of NAME_EDUCATION_TYPE



Among the non-defaulters , the clients with secondary/secondary special education are more in number and the next set of major applicants are having higher education.

Among the defaulters, the trend is similar.

Distribution of AGE_BRACKETS

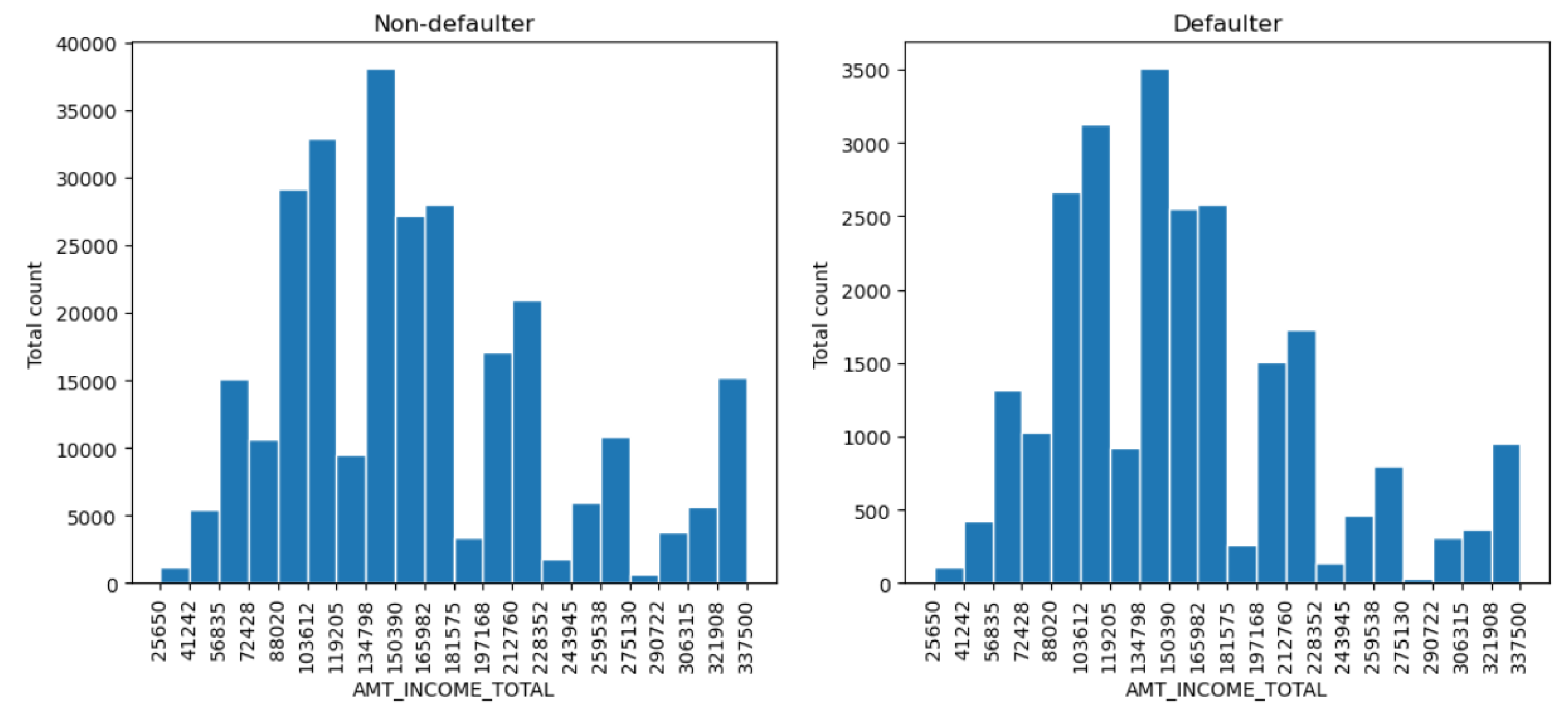


Most of the defaulters and non-defaulters belong to 30-40 age group, but among non-defaulters next major set of applicants belong to the age group of 40-50.

Age group of 20-30 are the second highest applicants who turned out to be defaulters.

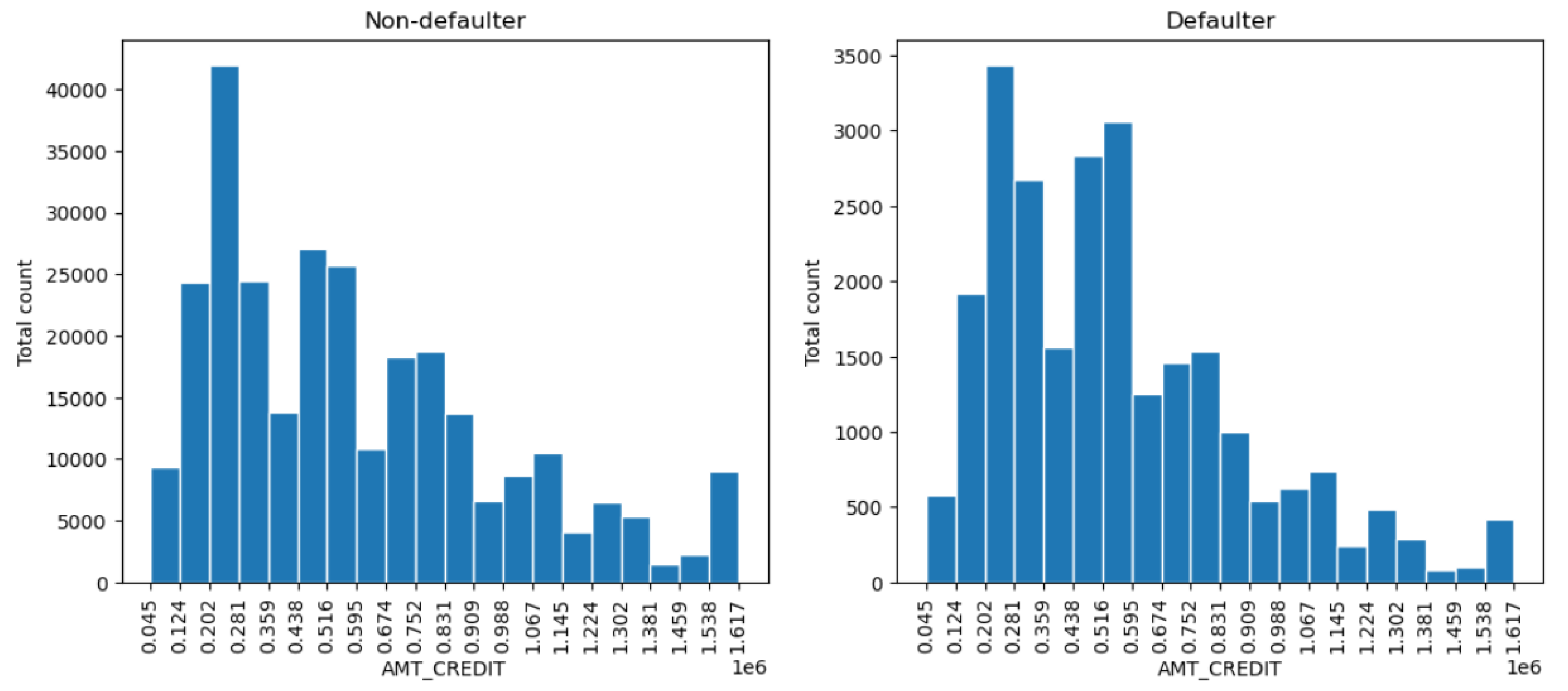
Segmented - Numerical univariate analysis

Distribution of AMT_INCOME_TOTAL



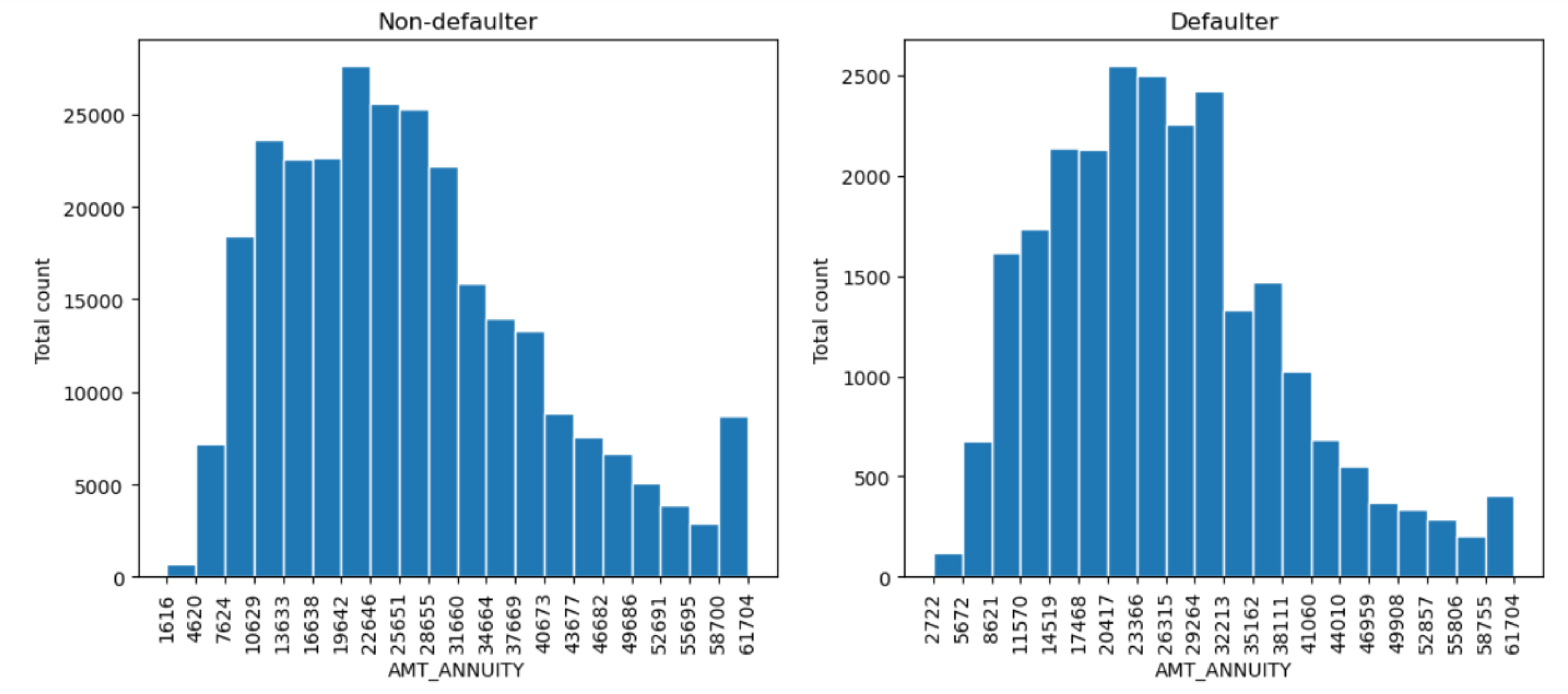
Majority of the defaulters and non-defaulters earn in the bracket of 1.34 lakhs to 1.5 lakhs.There are about 37000 non-defaulters belonging to the high income bracket, but about 3500 non-defaulters are earning this amount

Distribution of AMT_CREDIT



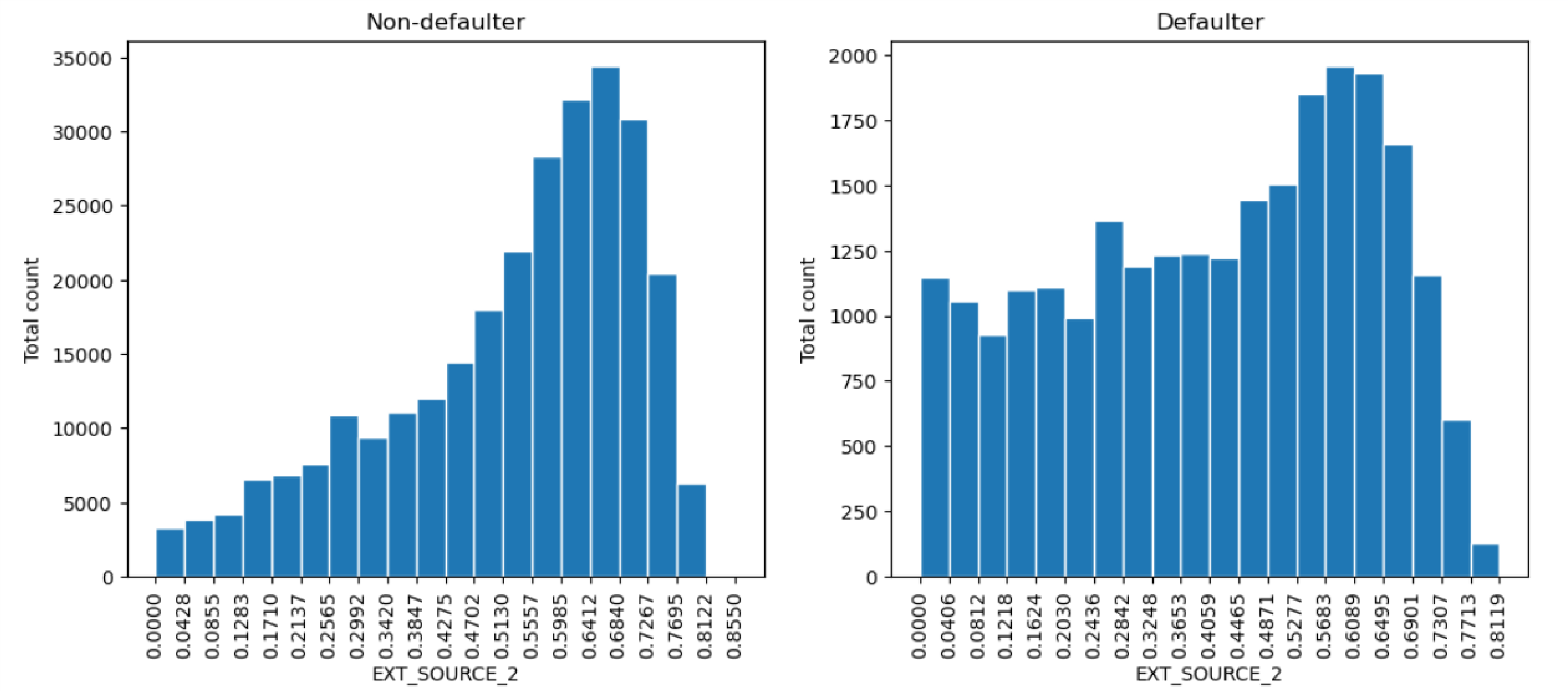
More than 40000 non-defaulters and close to 3500 defaultes are taking loans of credit amount in the group of 202000 to 281000

Distribution of AMT_ANNUITY



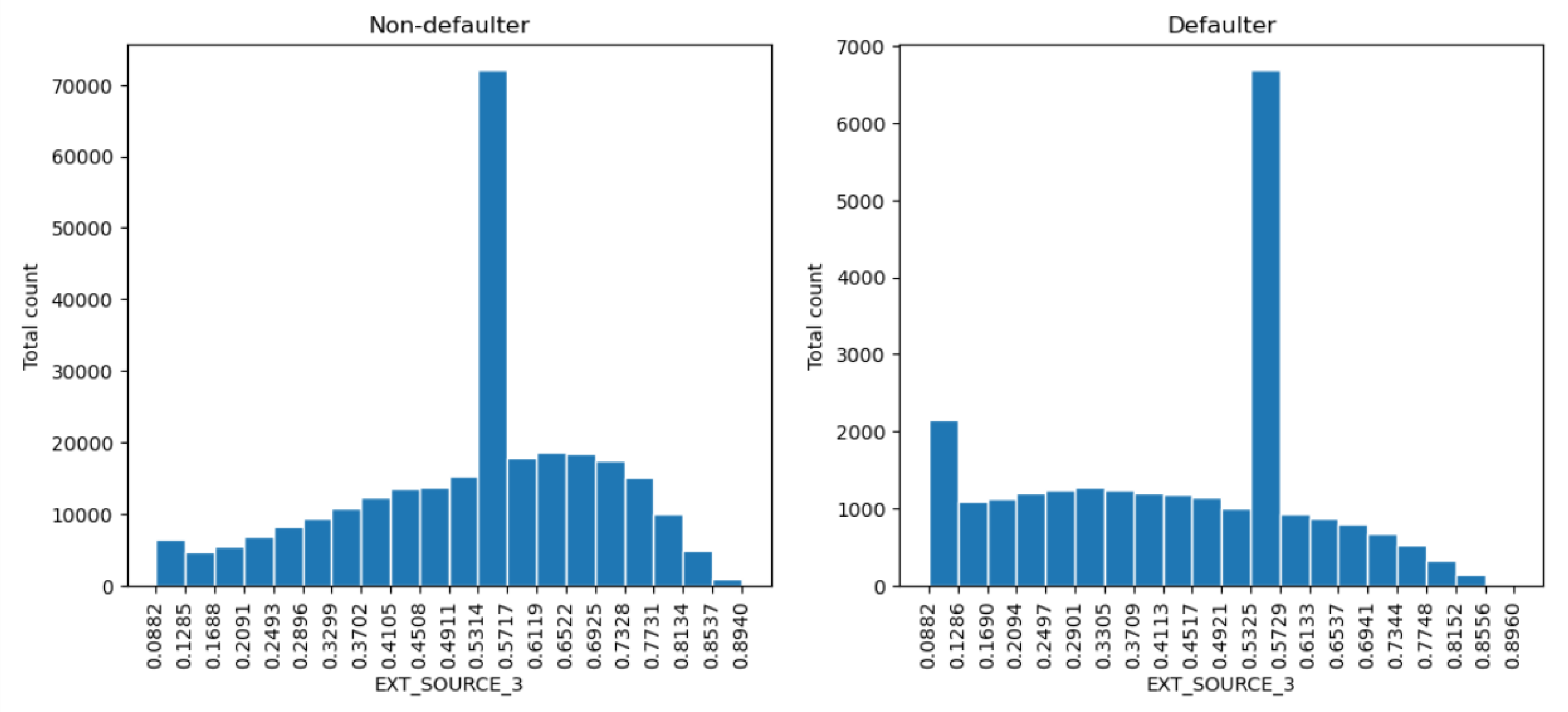
More than 25000 non-defaulters are having a loan annuity of about 19000 to 22000 while around 2500 defaulters are having a loan annuity of 20000 to 23000.

Distribution of EXT_SOURCE_2



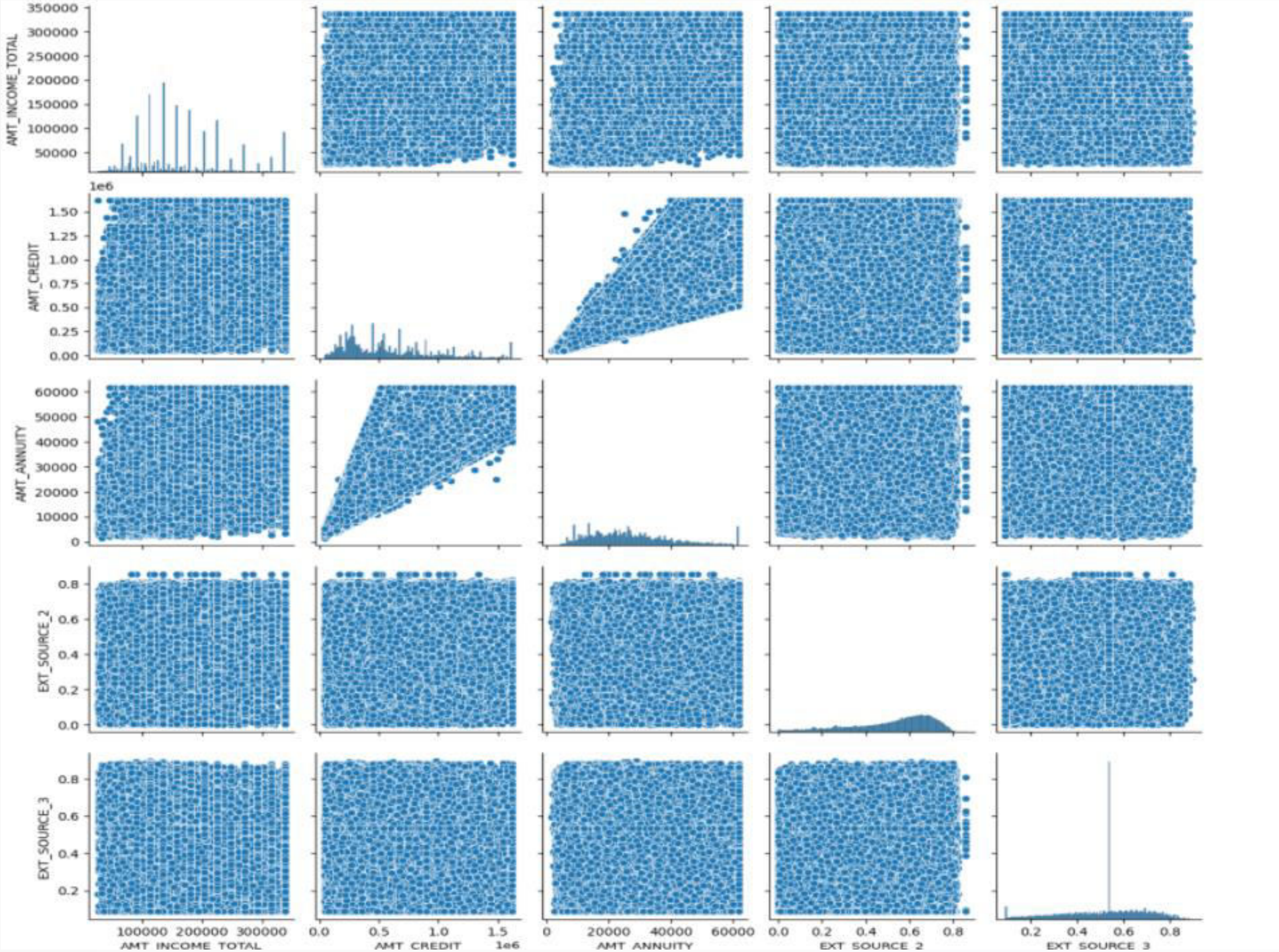
Most of the non-defaulters i.e. close to 35000 are having rating from external source_2 in the range of 0.64-0.68, while there are close to 2000 defaulters having rating of 0.56-0.6 .

Distribution of EXT_SOURCE_3



Majority of (About 70000) non-defaulters and majority of defaulters(close to 7000) are having a rating of 0.53-0.57

Bi-variate analysis- Scatter plot



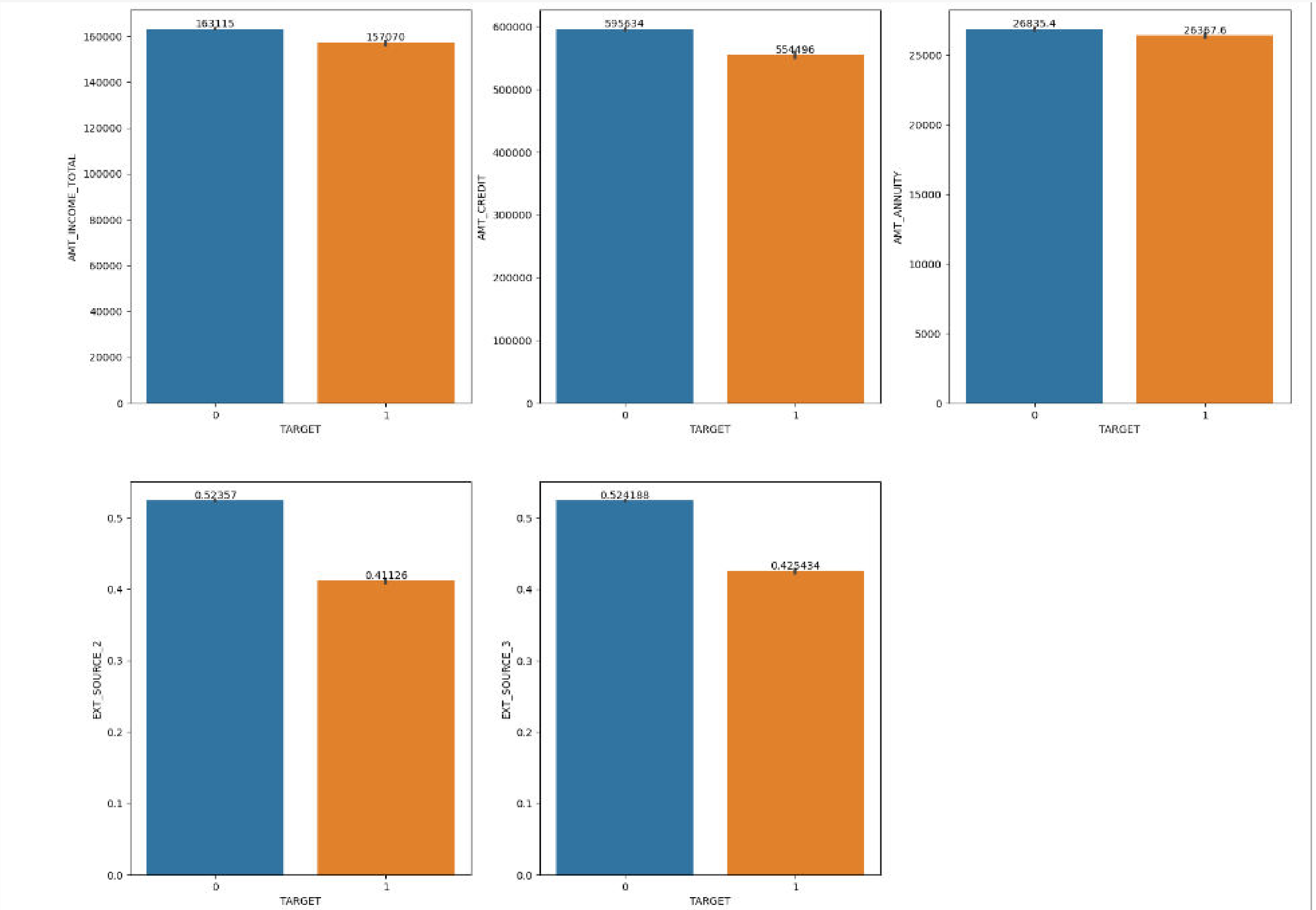
Bi-variate analysis- Scatter plot

- In the above analysis we can observe that there is linear positive relationship between AMT_ANNUITY and AMT_CREDIT.

- But we are not able to observe any strong linear relationship between other numerical variables

Bi-variate analysis- Bar plot

Target variable vs Numerical columns



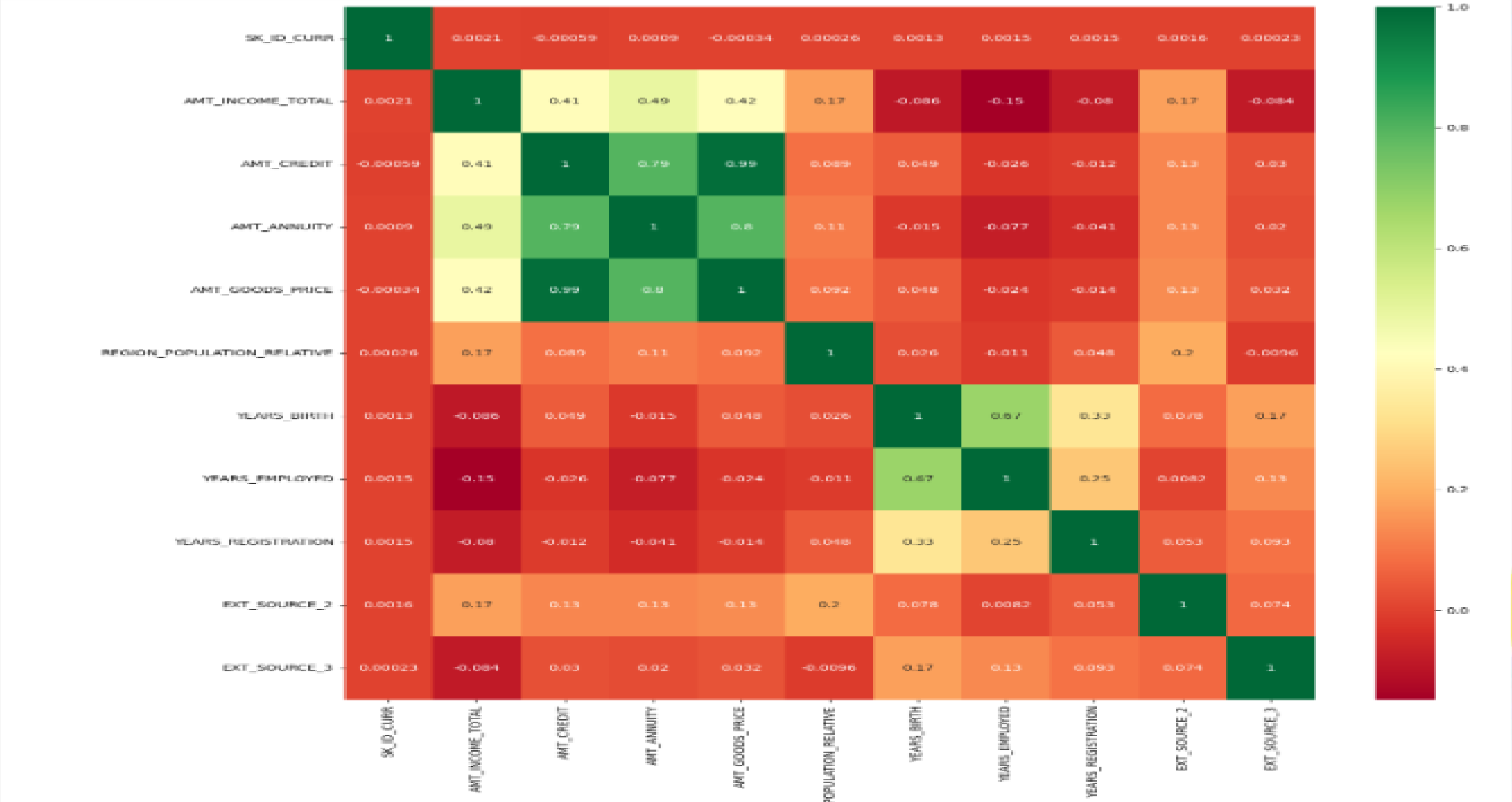
Bi-variate analysis- Bar plot

- Company can consider it safer to provide loans to the clients with higher income, are ready to accept and pay higher annual instalment.

- The rating given by external source 2 and external source 3 is much greater for non-defaulters(TARGET=0) compared to defaulters(TARGET=0).

- So, the company consider this as one of the important indicator while granting loans.

Correlation for non-defaulters



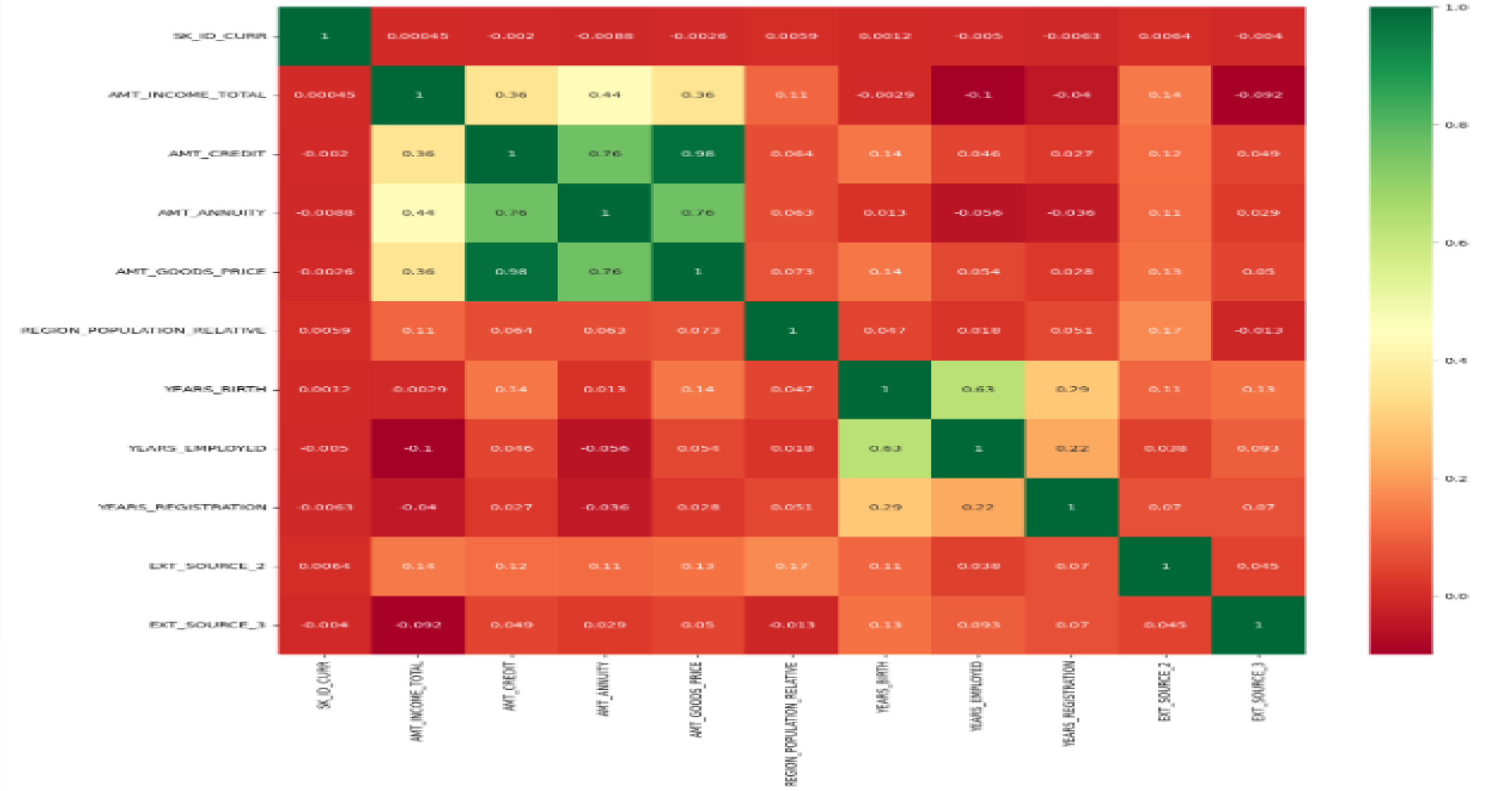
	SK_ID_CURR	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	EXT_SOURCE_2	EXT_SOURCE_3
SK_ID_CURR	1	0.0021	-0.00059	0.0009	-0.00034	0.00026	0.0013	0.0015	0.0015	0.0016	0.00023
AMT_INCOME_TOTAL	0.0021	1	0.41	0.49	0.42	0.17	-0.086	-0.15	-0.08	0.17	-0.084
AMT_CREDIT	-0.00059	0.41	1	0.79	0.99	0.089	0.049	-0.026	-0.012	0.13	0.03
AMT_ANNUITY	0.0009	0.49	0.79	1	0.8	0.11	-0.015	-0.077	-0.041	0.13	0.02
AMT_GOODS_PRICE	-0.00034	0.42	0.99	0.8	1	0.092	0.048	-0.024	-0.014	0.13	0.032
REGION_POPULATION_RELATIVE	0.00026	0.17	0.089	0.11	0.092	1	0.026	-0.011	0.048	0.2	-0.0096
YEARS_BIRTH	0.0013	-0.086	0.049	-0.015	0.048	0.026	1	0.67	0.33	0.078	0.17
YEARS_EMPLOYED	0.0015	-0.15	-0.026	-0.077	-0.024	-0.011	0.67	1	0.25	0.0082	0.13
YEARS_REGISTRATION	0.0015	-0.08	-0.012	-0.041	-0.014	0.048	0.33	0.25	1	0.053	0.093
EXT_SOURCE_2	0.0016	0.17	0.13	0.13	0.13	0.2	0.078	0.0082	0.053	1	0.074
EXT_SOURCE_3	0.00023	-0.084	0.03	0.02	0.032	-0.0096	0.17	0.13	0.093	0.074	1

Correlation for non-defaulters

- Top 10 correlations for dataset with TARGET = 0 are

Column 1	Column 2	Correlation
AMT_CREDIT	AMT_GOODS_PRICE	0.985582
AMT_ANNUITY	AMT_GOODS_PRICE	0.797315
AMT_ANNUITY	AMT_CREDIT	0.794808
YEARS_EMPLOYED	YEARS_BIRTH	0.674356
AMT_ANNUITY	AMT_INCOME_TOTAL	0.492921
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.417592
YEARS_BIRTH	YEARS_REGISTRATION	0.332781
YEARS_EMPLOYED	YEARS_REGISTRATION	0.254248
EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.195567

Correlation for defaulters



Correlation for defaulters

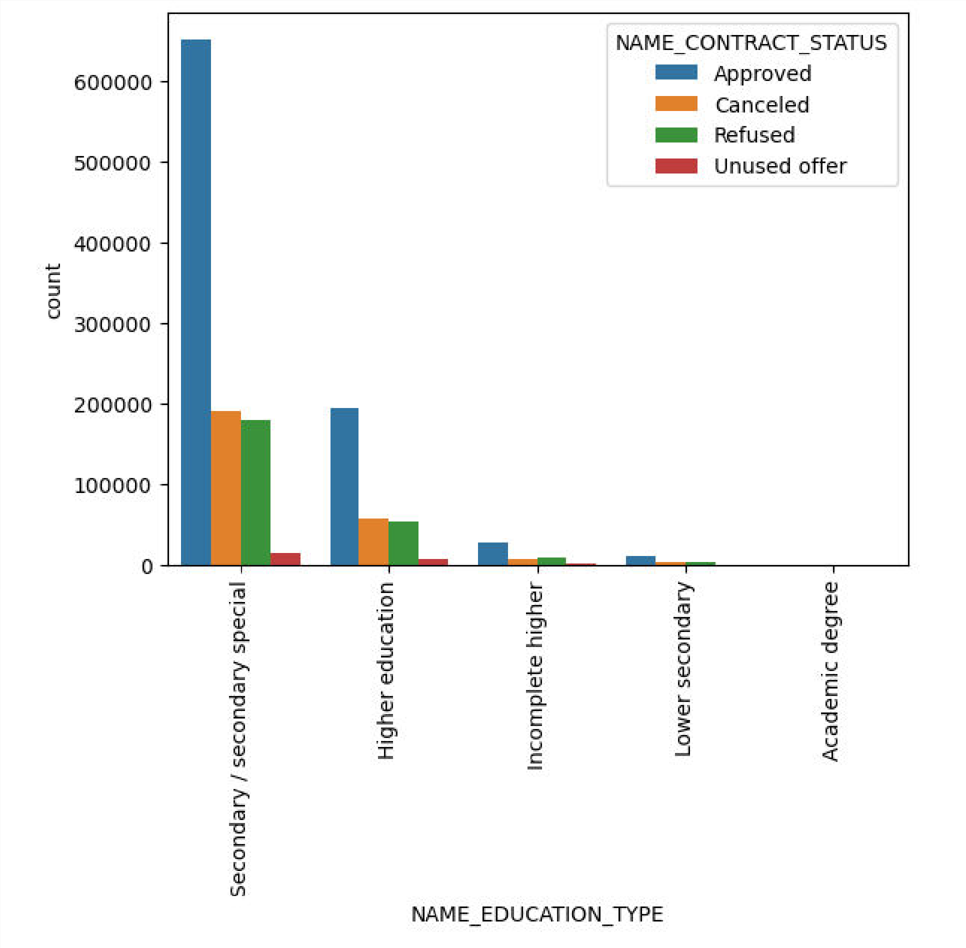
- Top 10 correlations for dataset with TARGET = 1 are

Column 1	Column 2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.981837
AMT_GOODS_PRICE	AMT_ANNUITY	0.760287
AMT_CREDIT	AMT_ANNUITY	0.760123
YEARS_BIRTH	YEARS_EMPLOYED	0.626650
AMT_ANNUITY	AMT_INCOME_TOTAL	0.436918
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.357696
AMT_CREDIT	AMT_INCOME_TOTAL	0.356199
YEARS_REGISTRATION	YEARS_BIRTH	0.288794
YEARS_REGISTRATION	YEARS_EMPLOYED	0.224188
REGION_POPULATION_RELATIVE	EXT_SOURCE_2	0.167309

Analysis on merged dataset

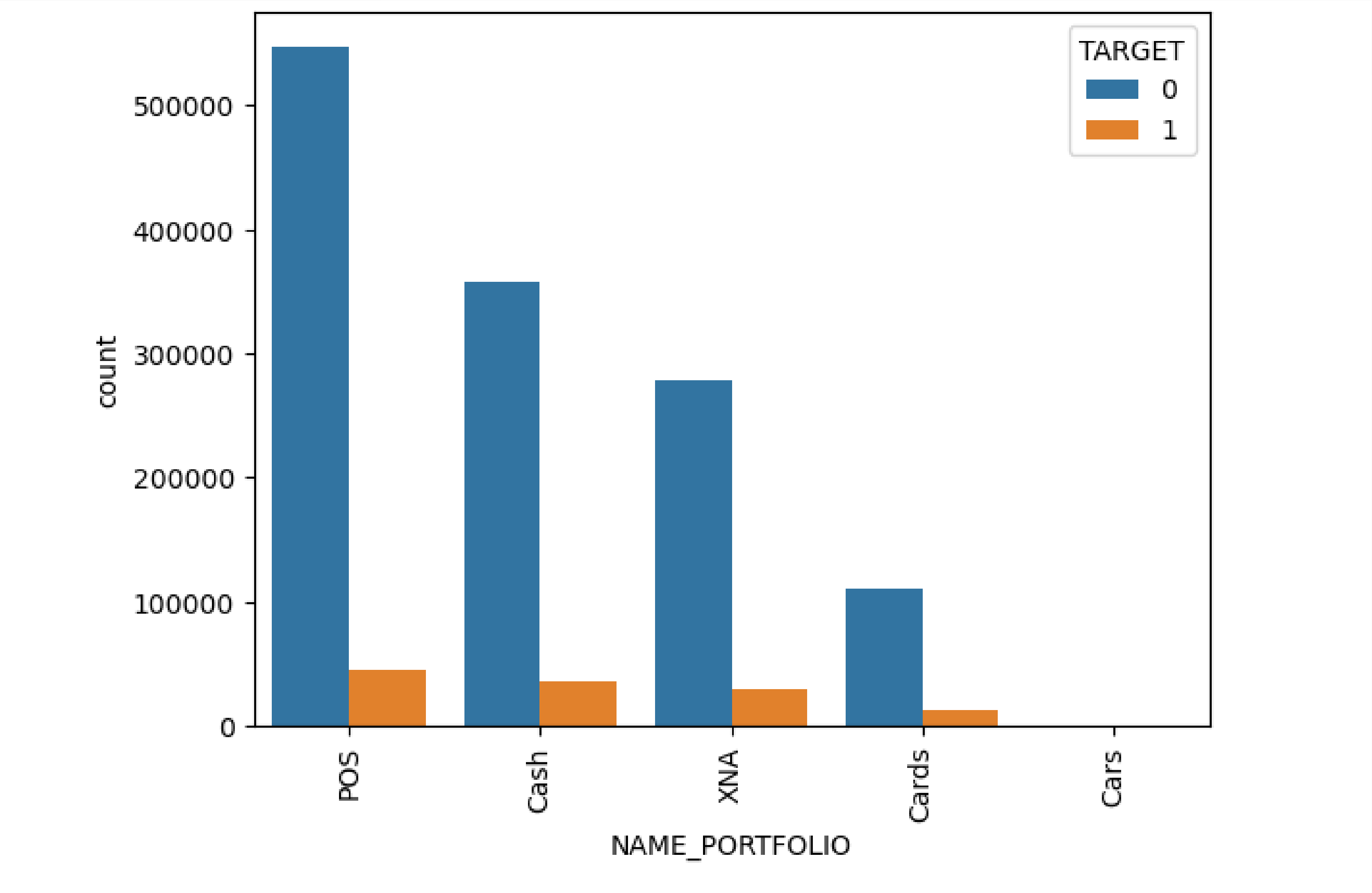
Bi-variate analysis

NAME_EDUCATION_TYPE vs NAME_CONTRACT_STATUS



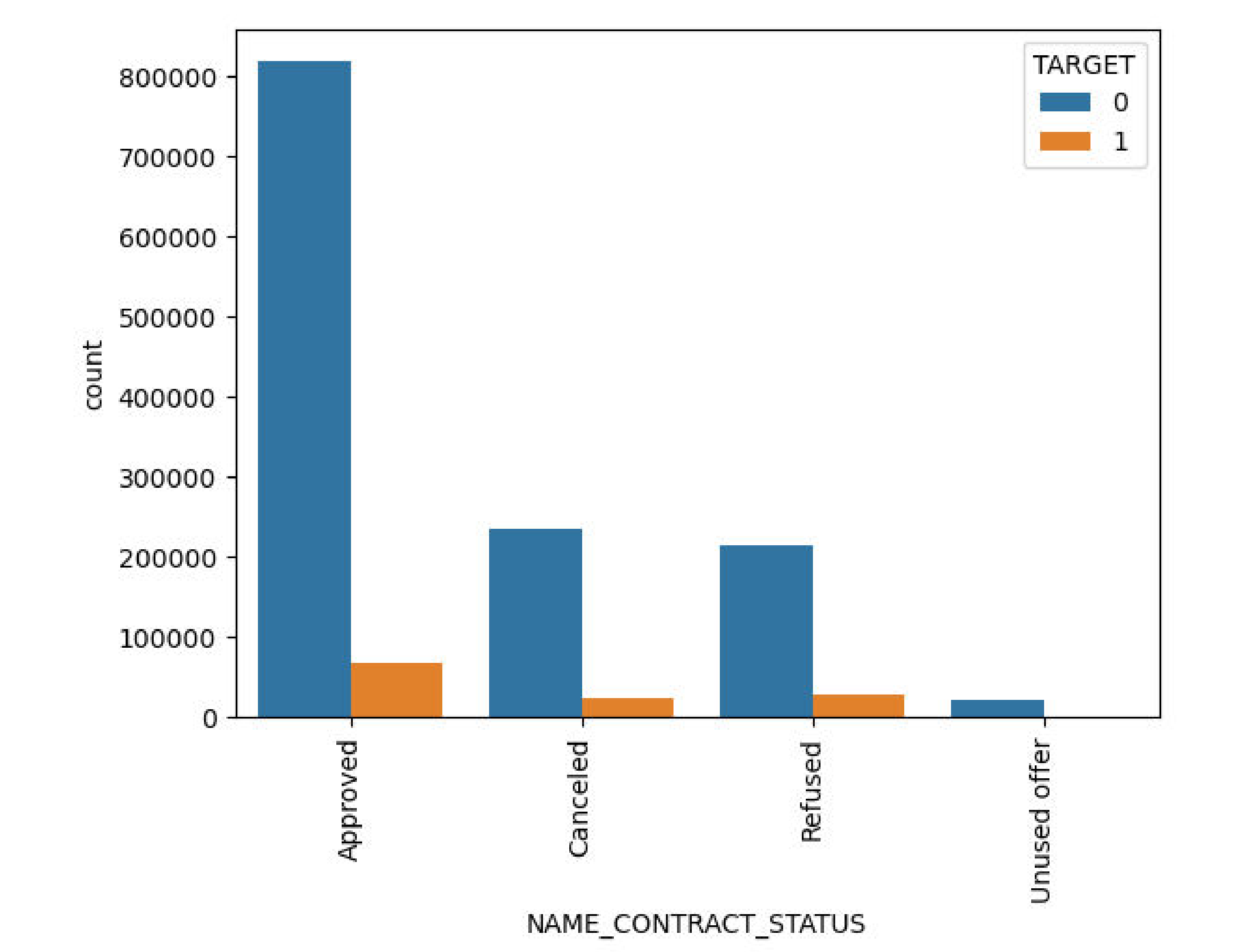
We can observe that clients with Secondary/secondary special education are having highest number of approved loans. The trend is similar in other levels of education.

TARGET vs NAME_PORTFOLIO



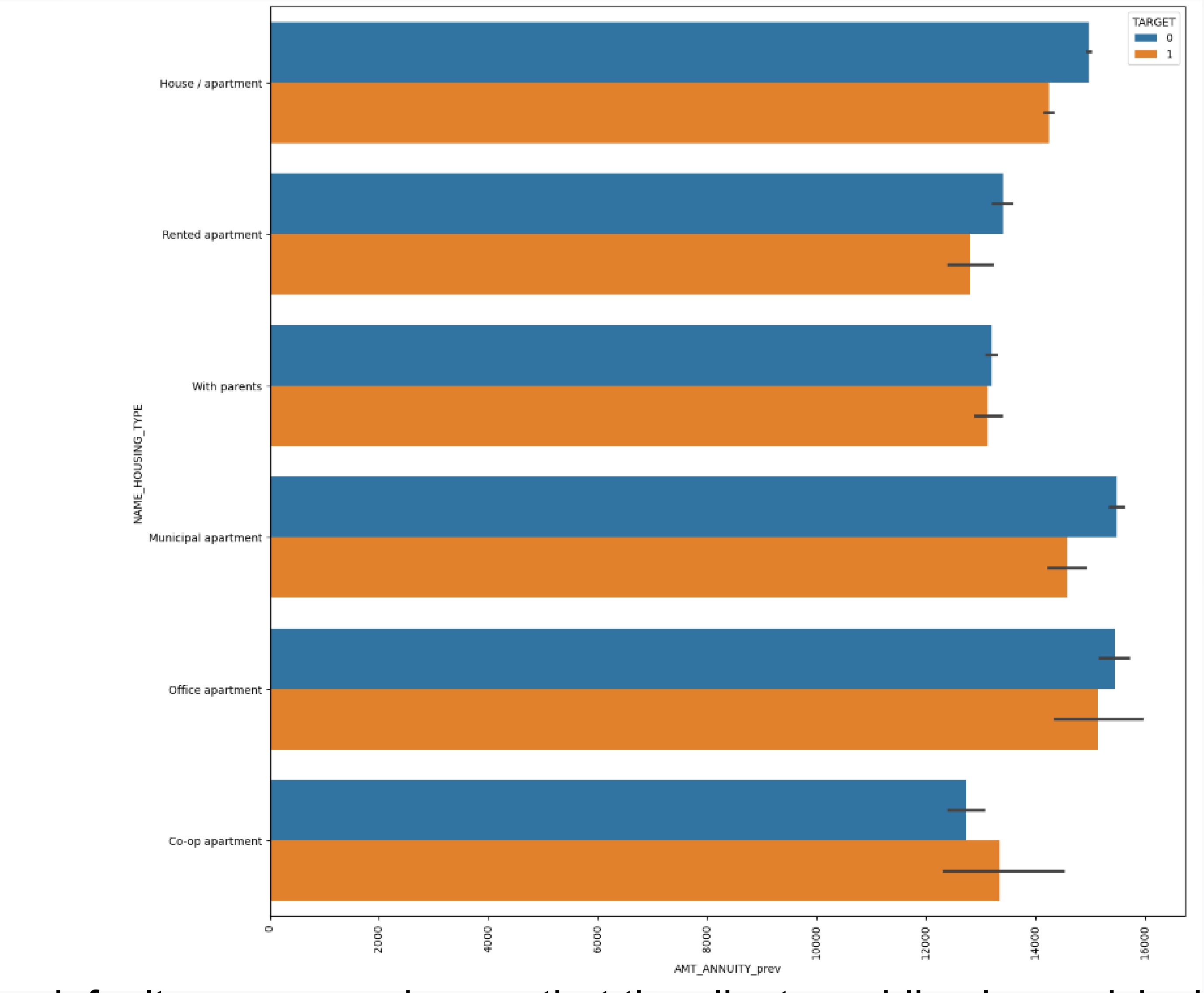
Portfolio count of non-defaulters is much more than the portfolio count of defaulters
This is true for all types of portfolios

TARGET vs NAME_CONTRACT_STATUS



Number of approved loans for non-defaulters is much more compared to the approved loans of defaulters. This is similar even in case of other contract status.

NAME_HOUSING_TYPE vs AMT_ANNUITY_prev



Among non-defaulters we can observe that the clients residing in municipal apartment are having highest installment amount among non-defaulters.
But among defaulters clients residing in office apartment are having highest installment amount.

Summary

- Company can consider external source 2 and external source 3 ratings as an important factor while granting loans.
- Company should rethink while providing loans to clients residing in municipal apartments as they are more among defaulters.
- Company should work on categorising portfolios better as there are a large number of people among ‘XNA’ categories and it’s not clear about it’s importance.
- Company should focus on clients having secondary and higher education as they are showing lesser patterns of defaulting.
- Company should be extra careful while considering income, credit amount and annual instalments of the clients as both defaulters and non-defaulters are showing similar trends.