Credit EDA Assignment

Credit Risk Analysis

Submitted by –
Mariam Abbas Zaidi

Problem Statement

- To identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- The aim is to identify the applicants who are capable of repaying the load and ensure such applicants are not rejected.
- To understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default for Risk assessment.

Approach Methodology

- Discovering – Loading the data frame and understanding Data dictionary.
- Structuring – Segregating features into Continuous, Categorical and ID(Unique Identifiers)
- Cleaning - a) Checking for Data inconsistency.
  b) Check for Missing values
  c) Check for Outliers
- Checking the data imbalance ratio in this scenario.
- Perform Univariate, Bivariate and Multivariate Analysis.
- Examine top 10 correlation for defaulter and non-defaulters.
- Merge the 2 data sets after cleaning and analysing each of them.
- Derive insights and conclusions from above analysis.

Assumptions

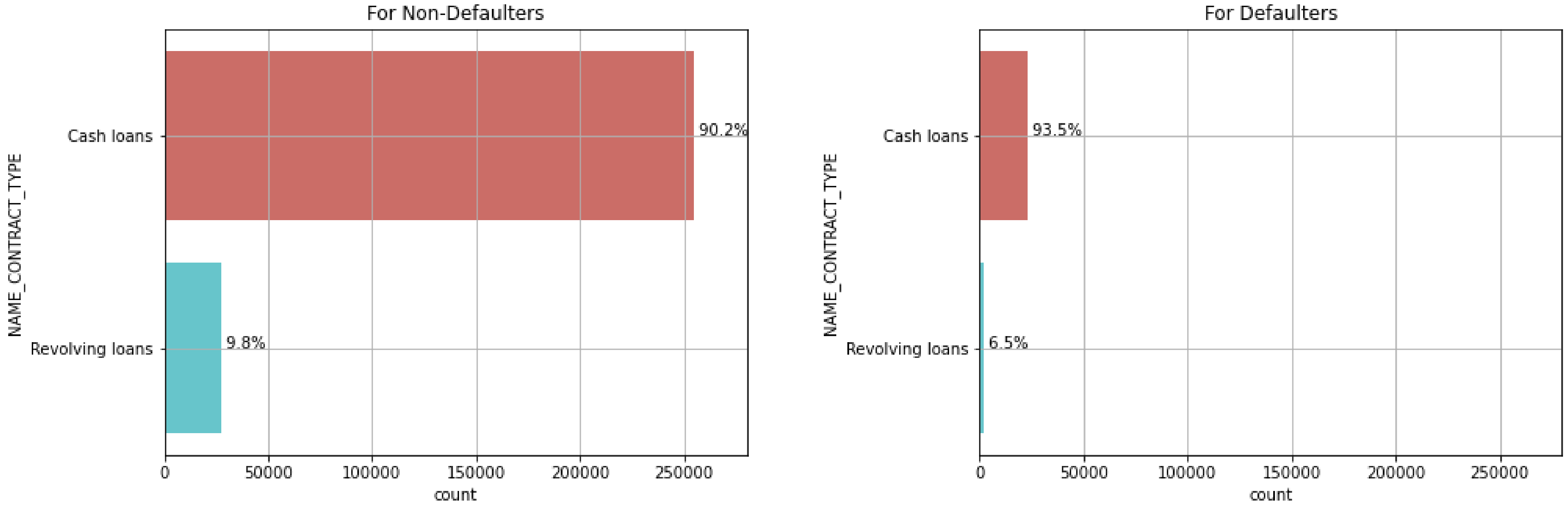
- Here I have assumed the cut-off percentage for Missing values as 45% and 35% respectively for Application Data and Previous Data respectively. Meaning all columns that has more than 45% and 35% of missing values have been dropped.
- All negative values in days were converted to positive values.

Data Imbalance

- Here we infer from TARGET variable that 8.07% people are Defaulters(clients with payment difficulties) and the other majority 91.92% are others who are non-Defaulters
- The Imbalance ratio is 11.38%
- The TARGET feature has been columnized into Non defaulters and Defaulters for further analysis

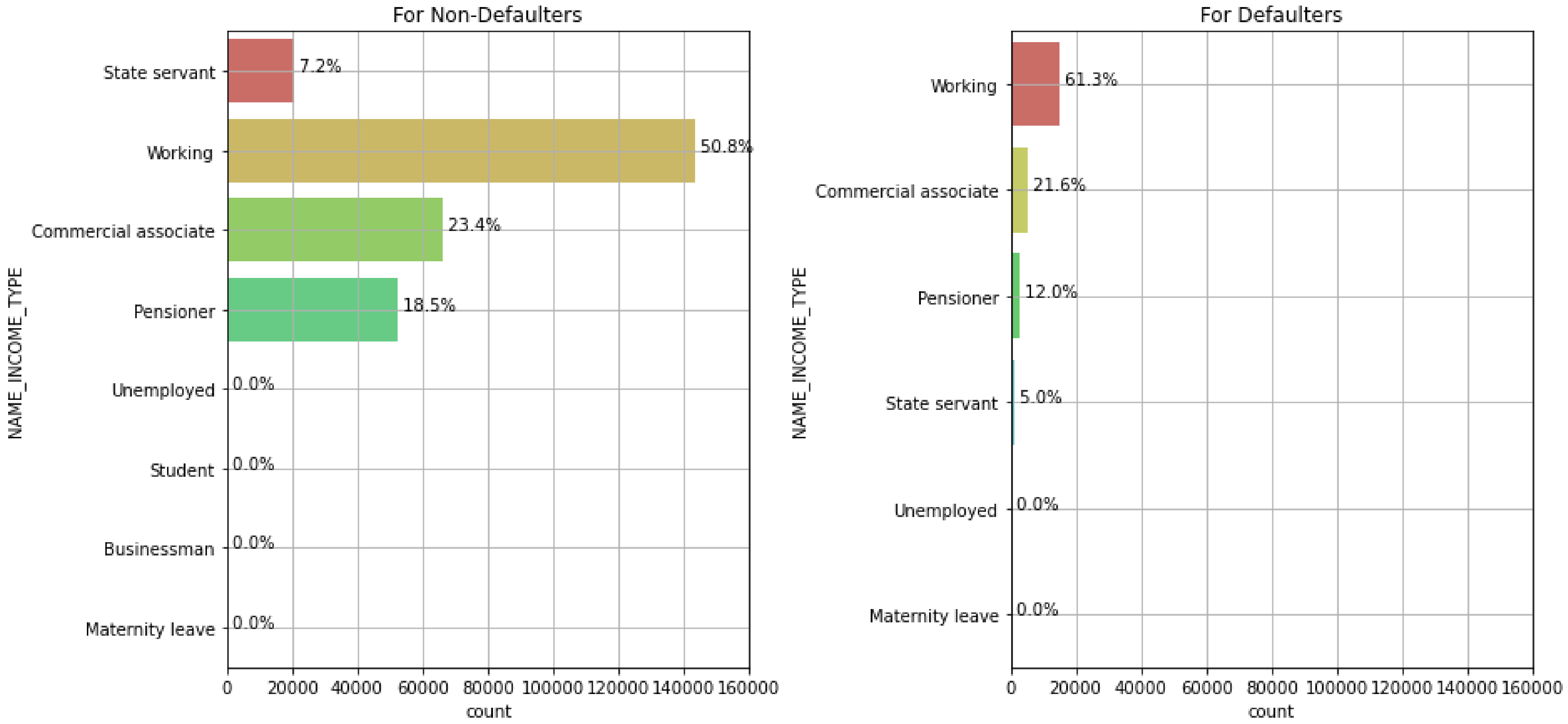
Segregated Univariate Analysis

Name Contract type analysis for Defaulters and Non defaulters



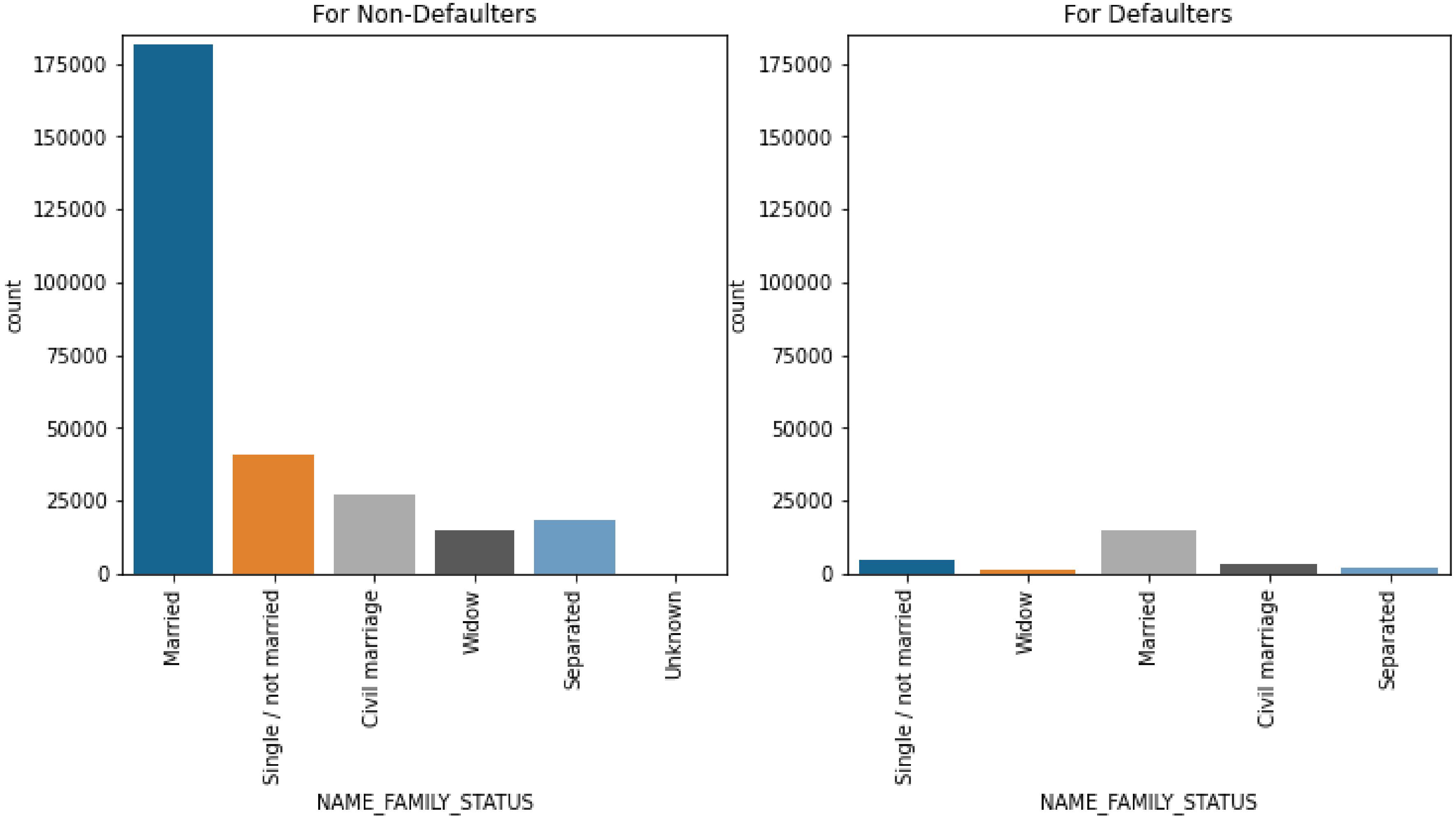
Inference: Cash loans are more likely to be given out by the bank and most likely yo be repaid showing 90% as compared to 9% in Revolving Loans

Income type analysis for Defaulters and Non defaulters



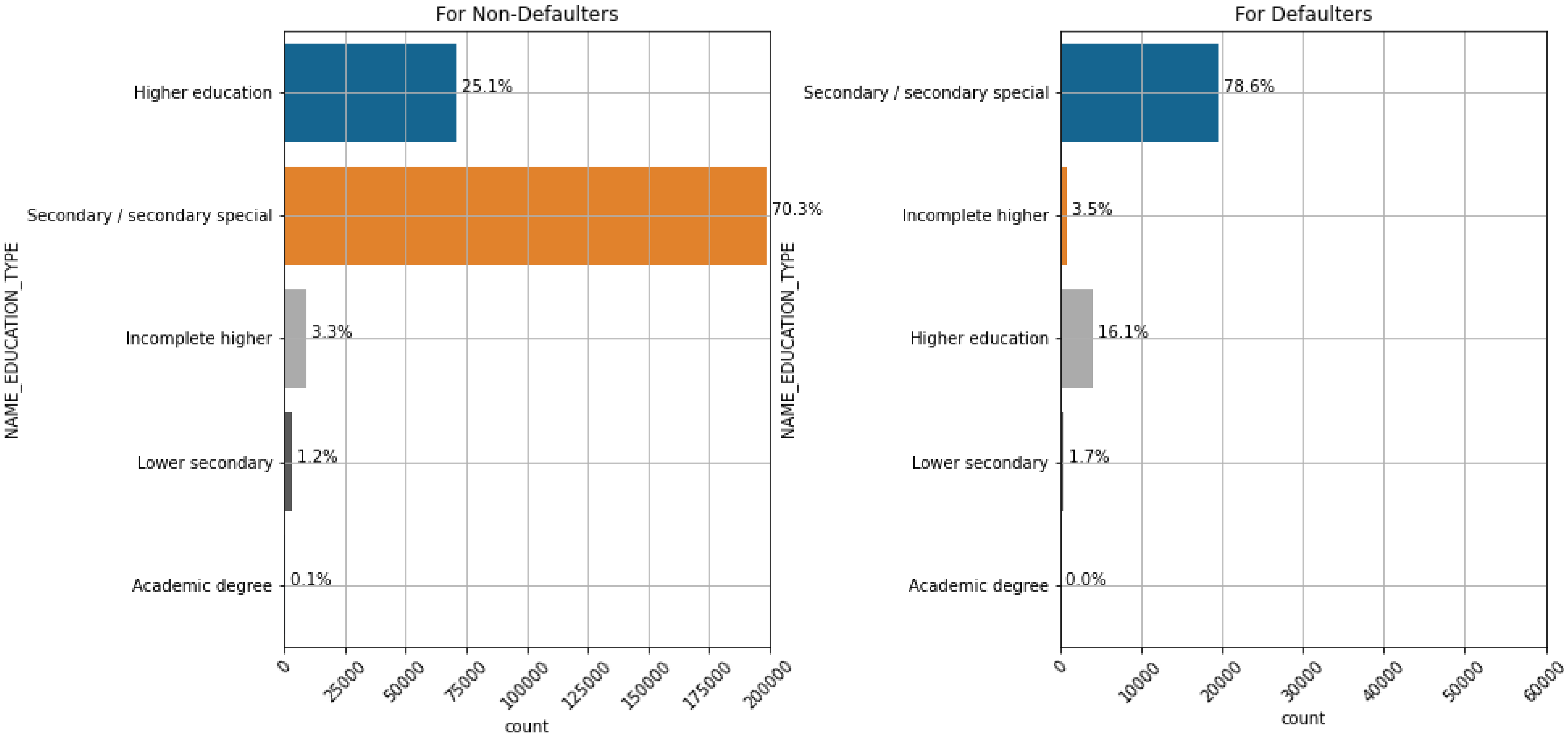
Inference: Most loans are given out to working professionals and Working Professionals are most likely to be non-defaulters in loans followed by Commercial associates.

Family Status analysis for Defaulters and Non defaulters



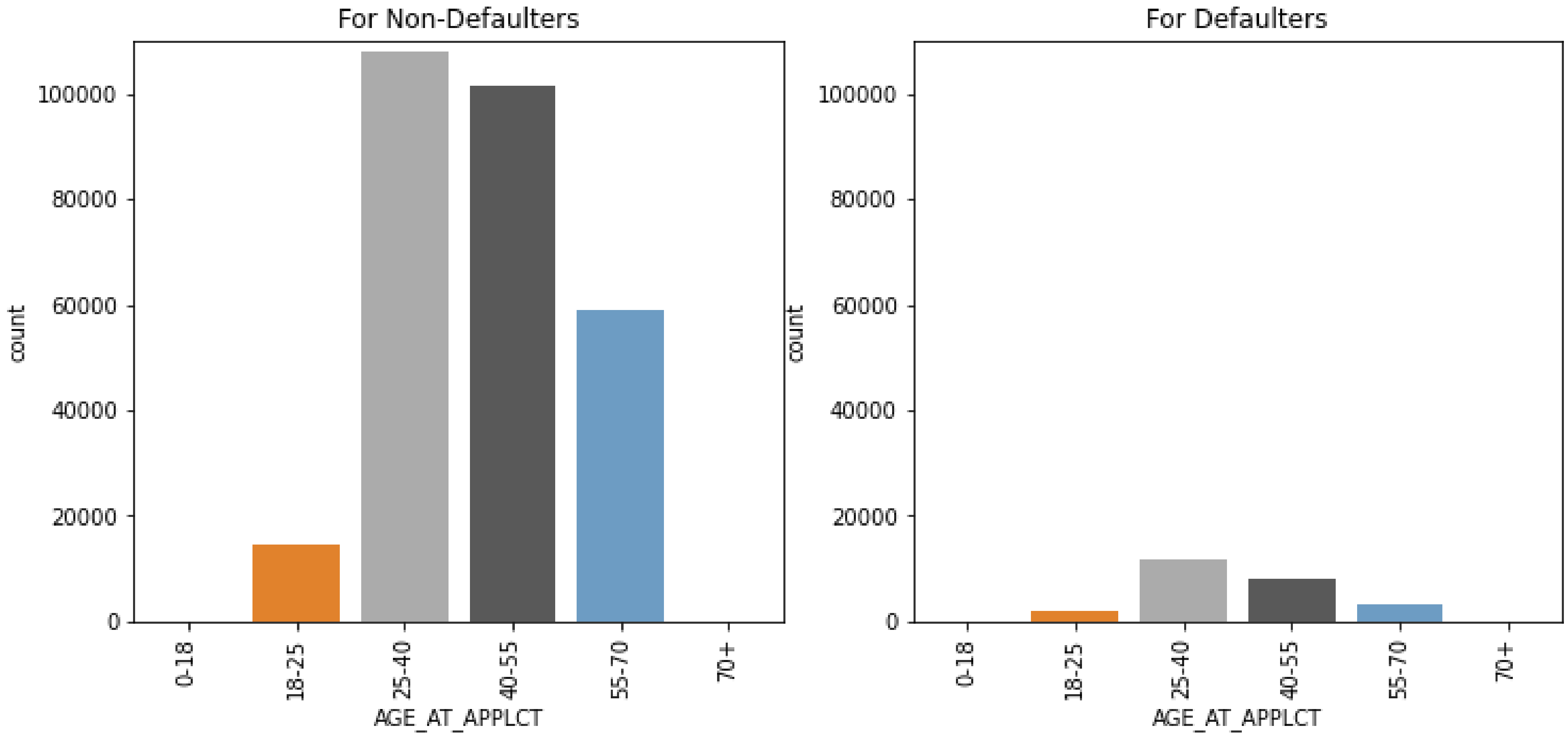
Inference : Married customers are most likely to repay loan amount as compared to any other category and customers in Separated and Widow category are least defaulters.

Education type analysis for Defaulters and Non defaulters



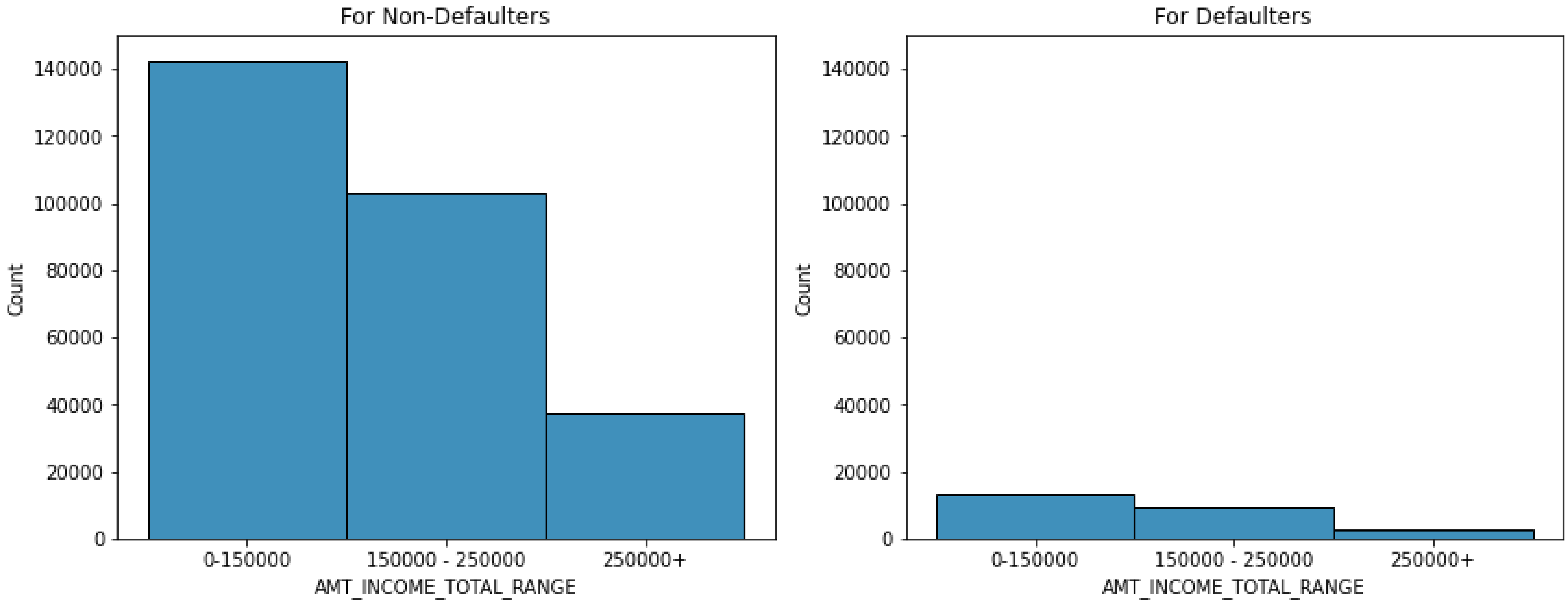
Inference: Customer having Secondary Education are most likely to pay back on time followed by customers with Higher Education. Very few percentage of loans have been given to Customers with Academic degrees

Age analysis for Defaulters and Non defaulters



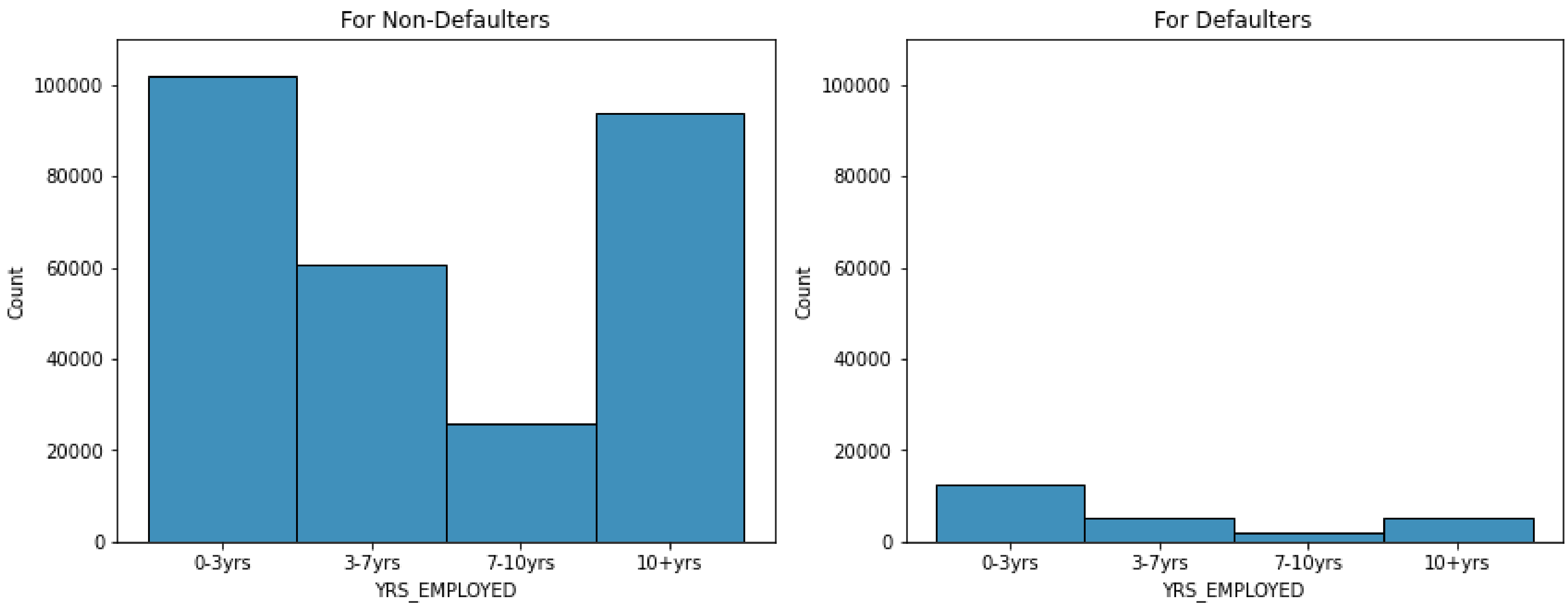
Inference: Customers aged between 25-40 are most likely to pay back loans. Customer aged 18-25 are least likely to be given loans by the bank since the defaulters and non defaulters are both in low percentages

Amount of Income analysis for Defaulters and Non defaulters



Inference: Customers with low income range(0-150000) have been given maximum loans and are more likely to pay them back

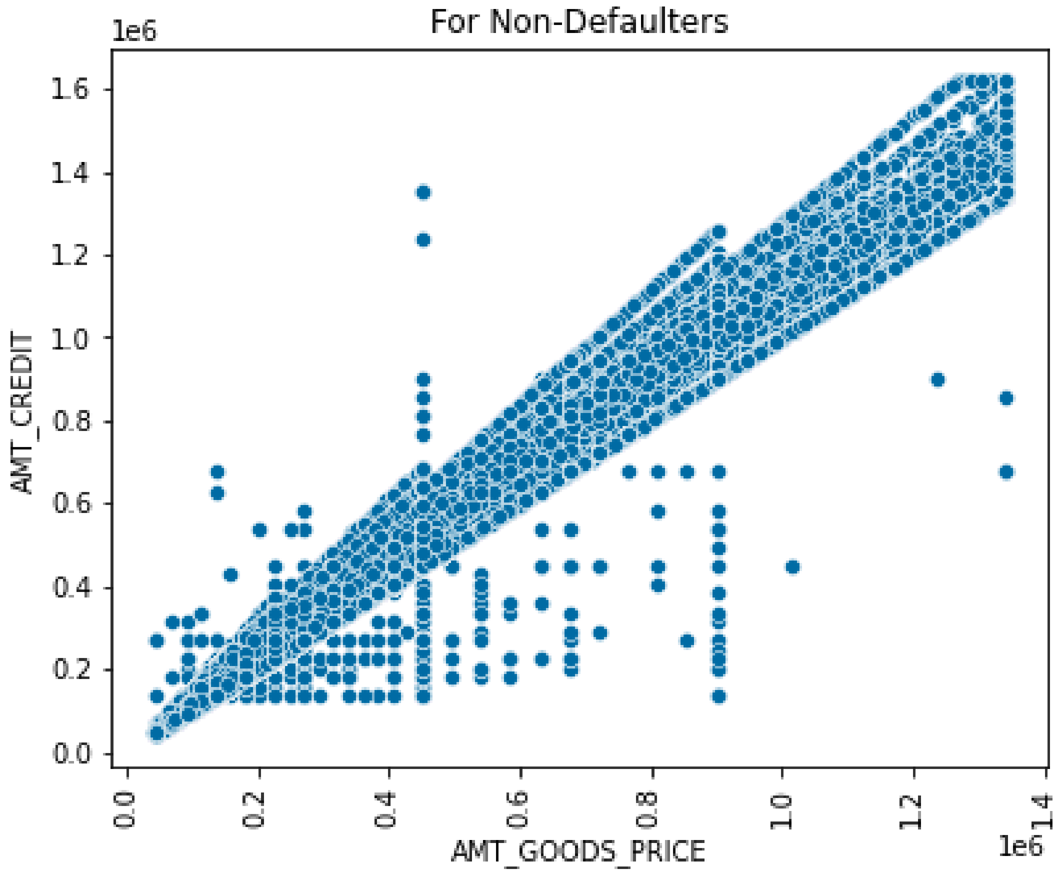
Years employed analysis for Defaulters and Non defaulters

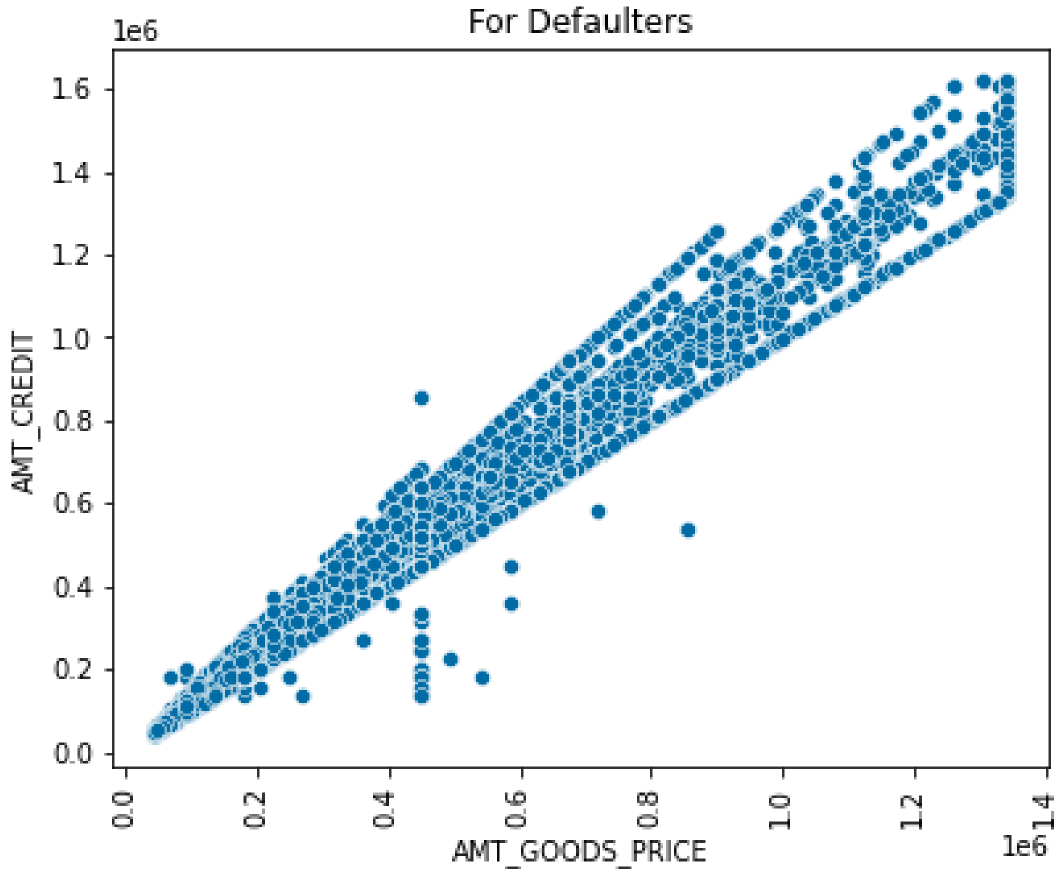


Inference: Customers with 0-3 yrs working are most likely to repay loans closely followed by people with 10+ yrs of experience

Bivariate Analysis

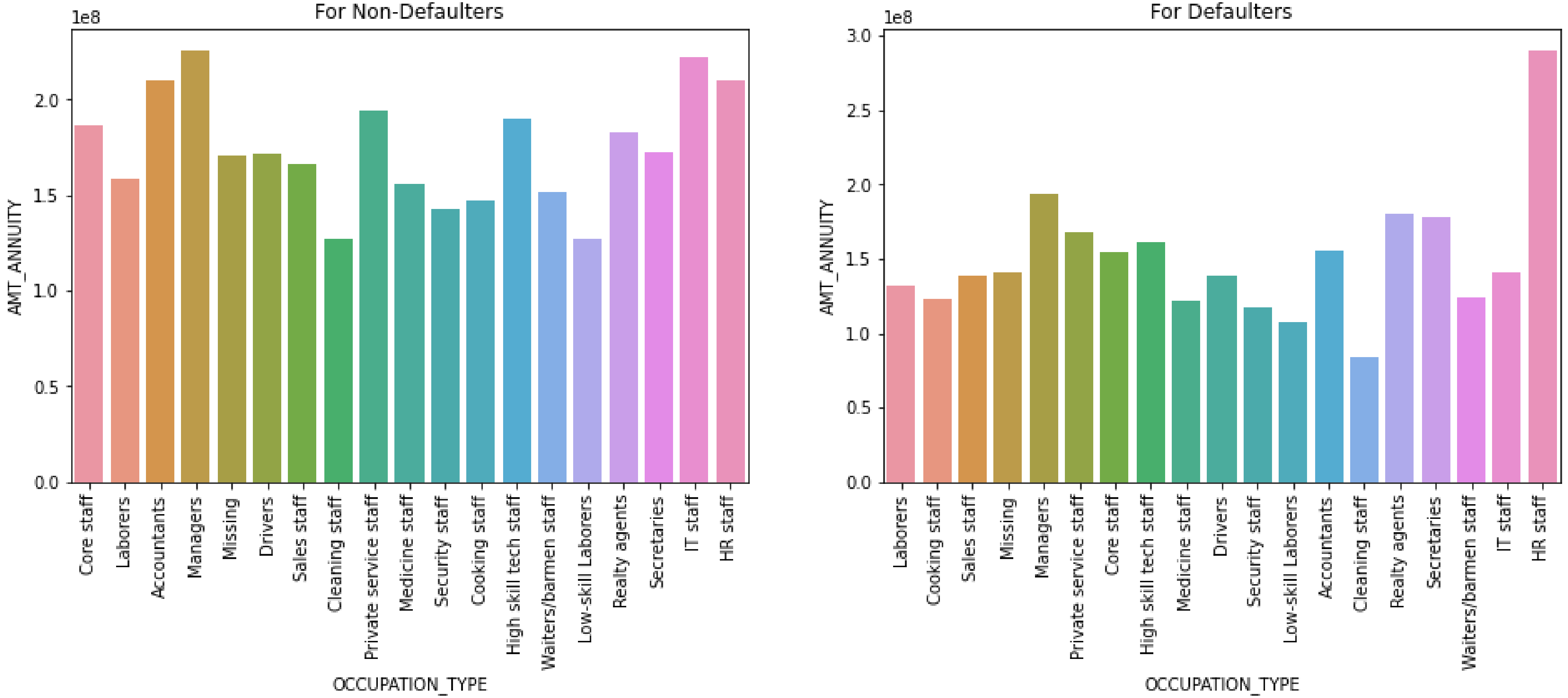
Credit amount VS Goods price amount





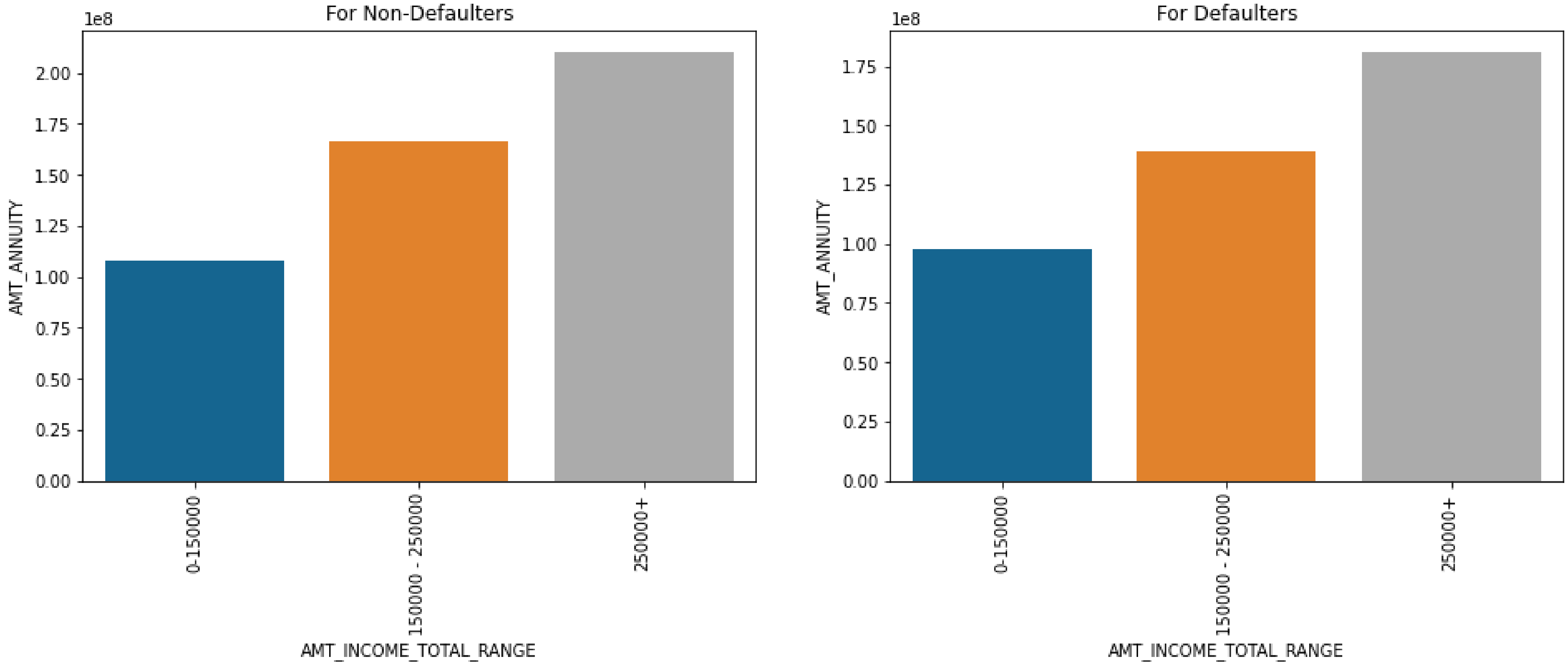
Inference: AMT_CREDIT and AMT_GOODS_PRICE show a positive linear correlation. Also the number of defaulters decreases with increasing AMT_CREDIT and AMT_GOODS_PRICE

Annuity amount VS Occupation type



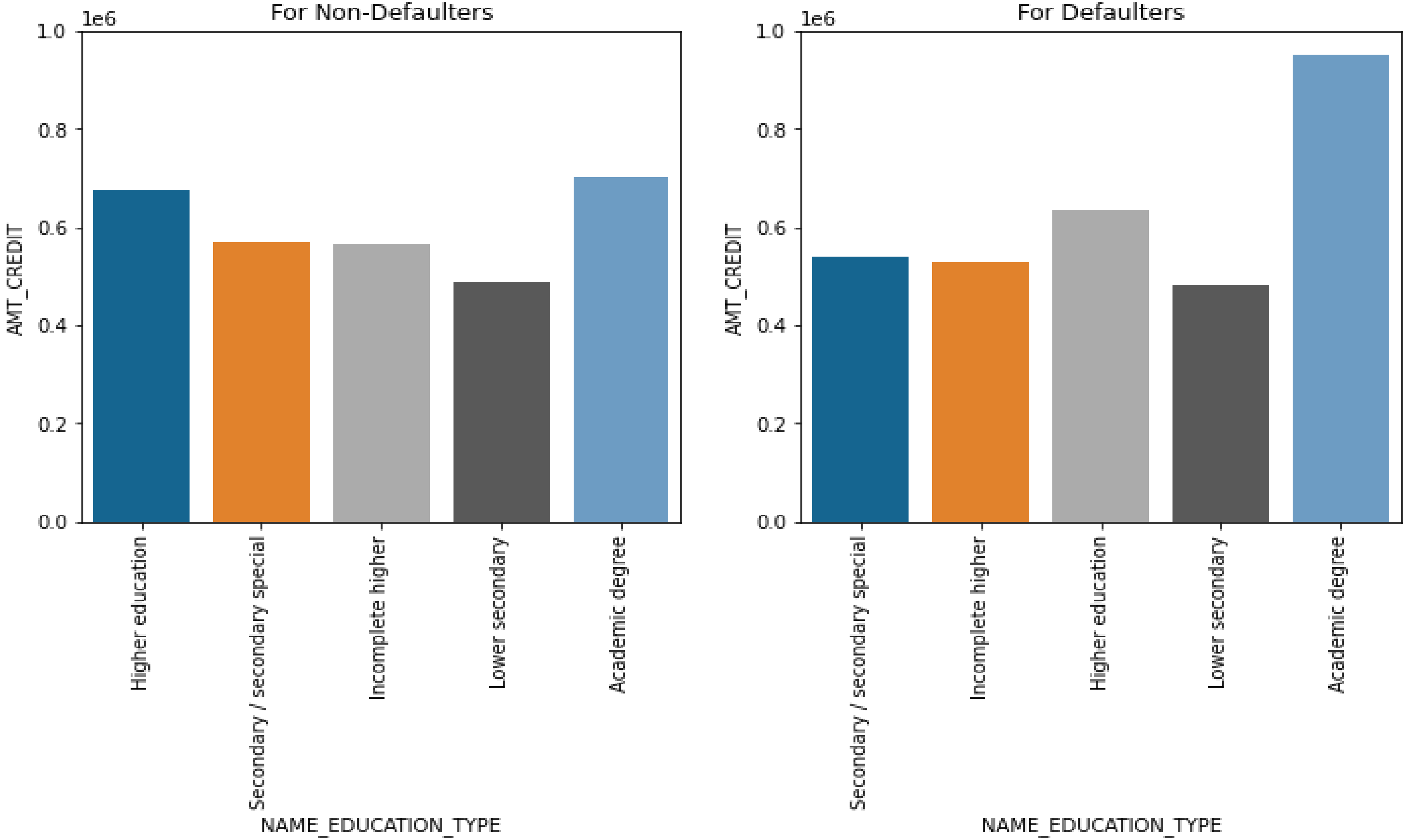
Inference: Majority of NON-defaulters are Managers who get high loan annuity amount. In defaulters we see HR staff getting high annuity loan amount.

Annuity amount VS Total Income



Inference: Customers with higher income are given high annuity amount. But we have seen in earlier graphs that customers with low income range are most Non-defaulters and maximum loans are given to them with low annuity amount.

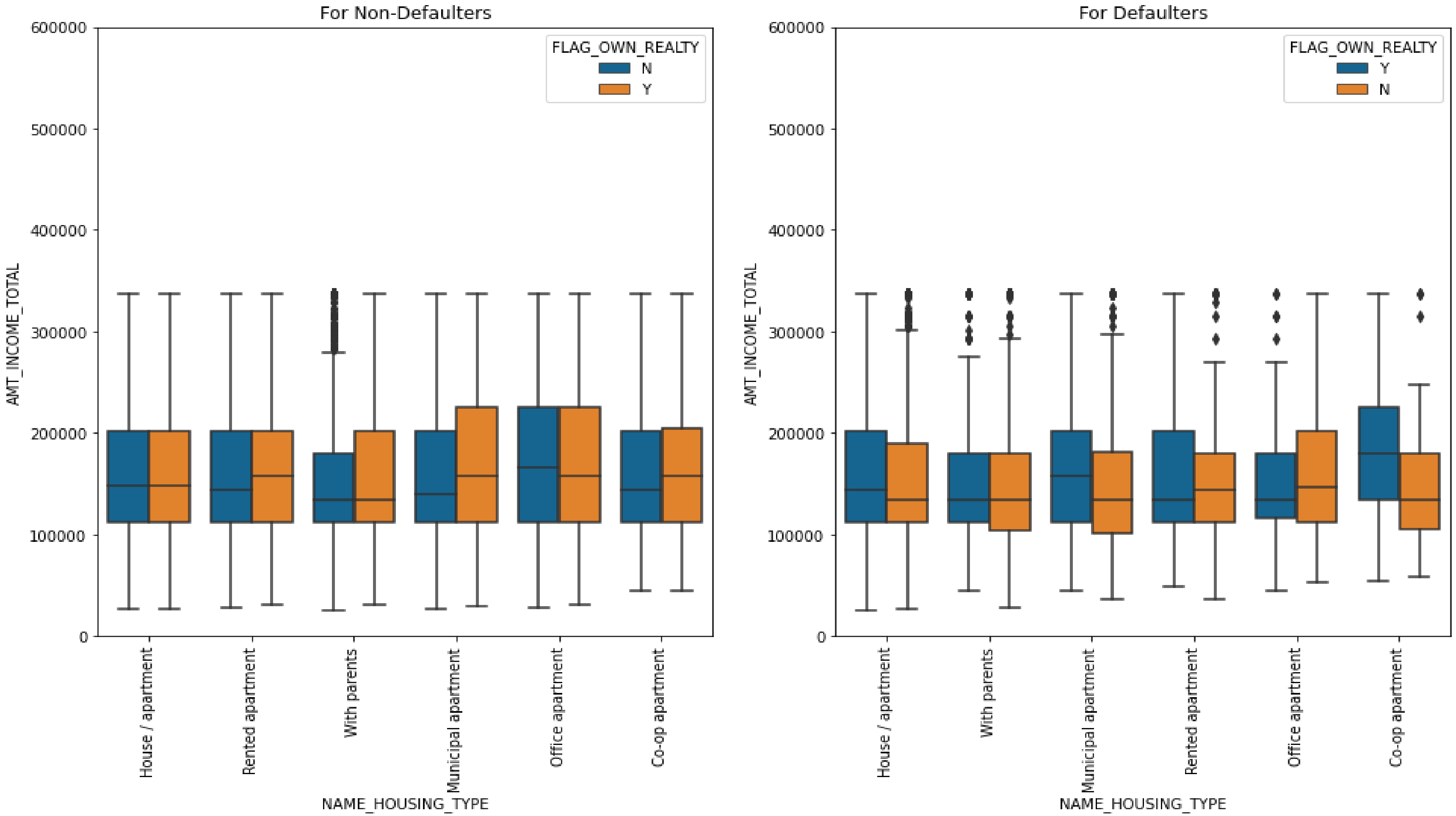
Credit amount VS Education type



Inference: Academic college degree holders are most likely to default.

Multivariate Analysis

Income type analysis for Defaulters and Non defaulters

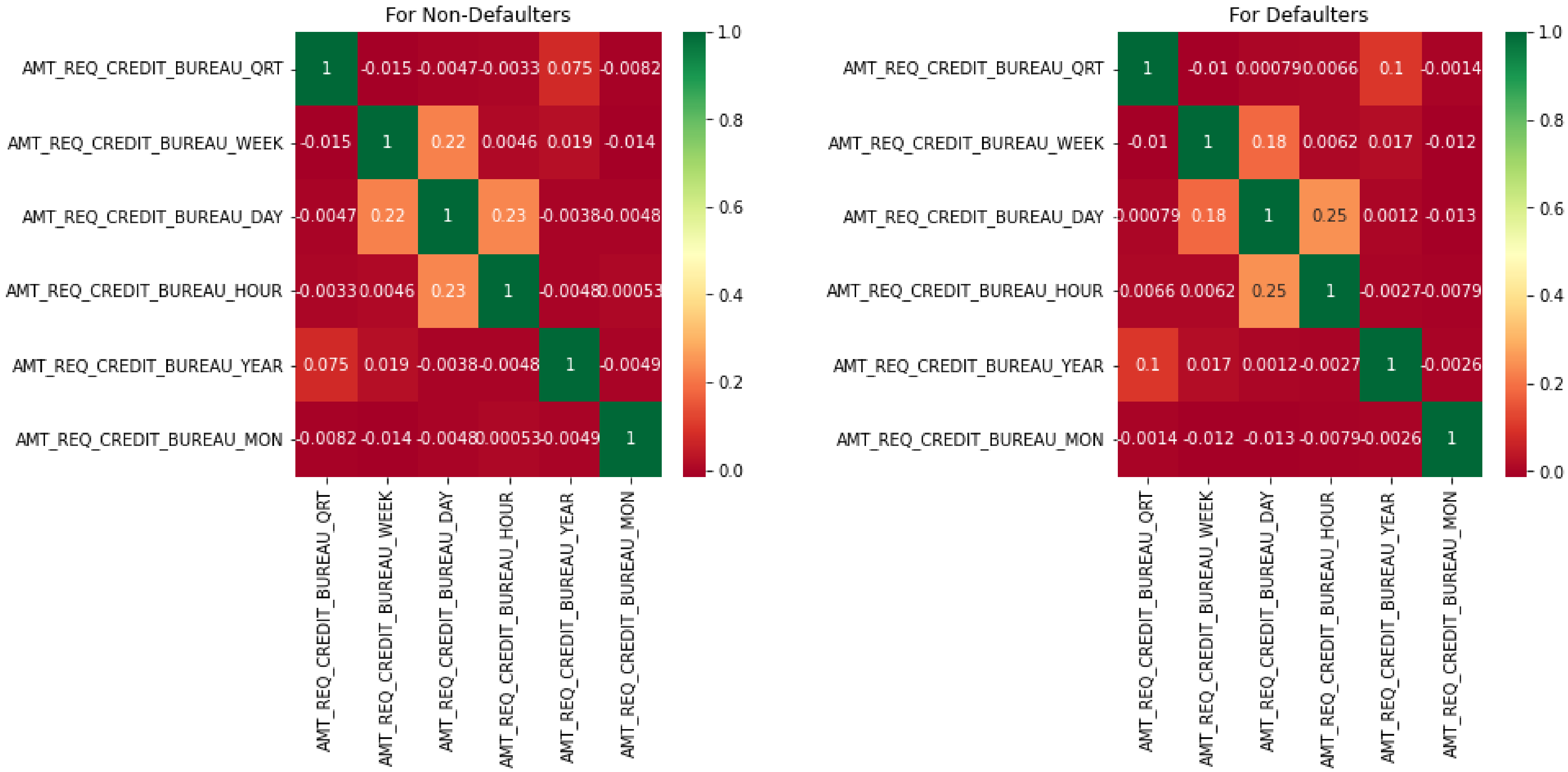


Inference: This graph doesnt not give us enough information. However, people staying in co-op apartments having lower income who do not own a realty are least defaulters. Most loans are given out to customers living in House / apartment having high total income.

Social circle Correlation analysis

Inference: There is a strong correlation between OBS_60_CNT_SOCIAL_CIRCLE & OBS_30_CNT_SOCIAL_CIRCLE and least between DEF_60_CNT_SOCIAL_CIRCLE & OBS_30_CNT_SOCIAL_CIRCLE for both Defaulters and Non-Defaulters

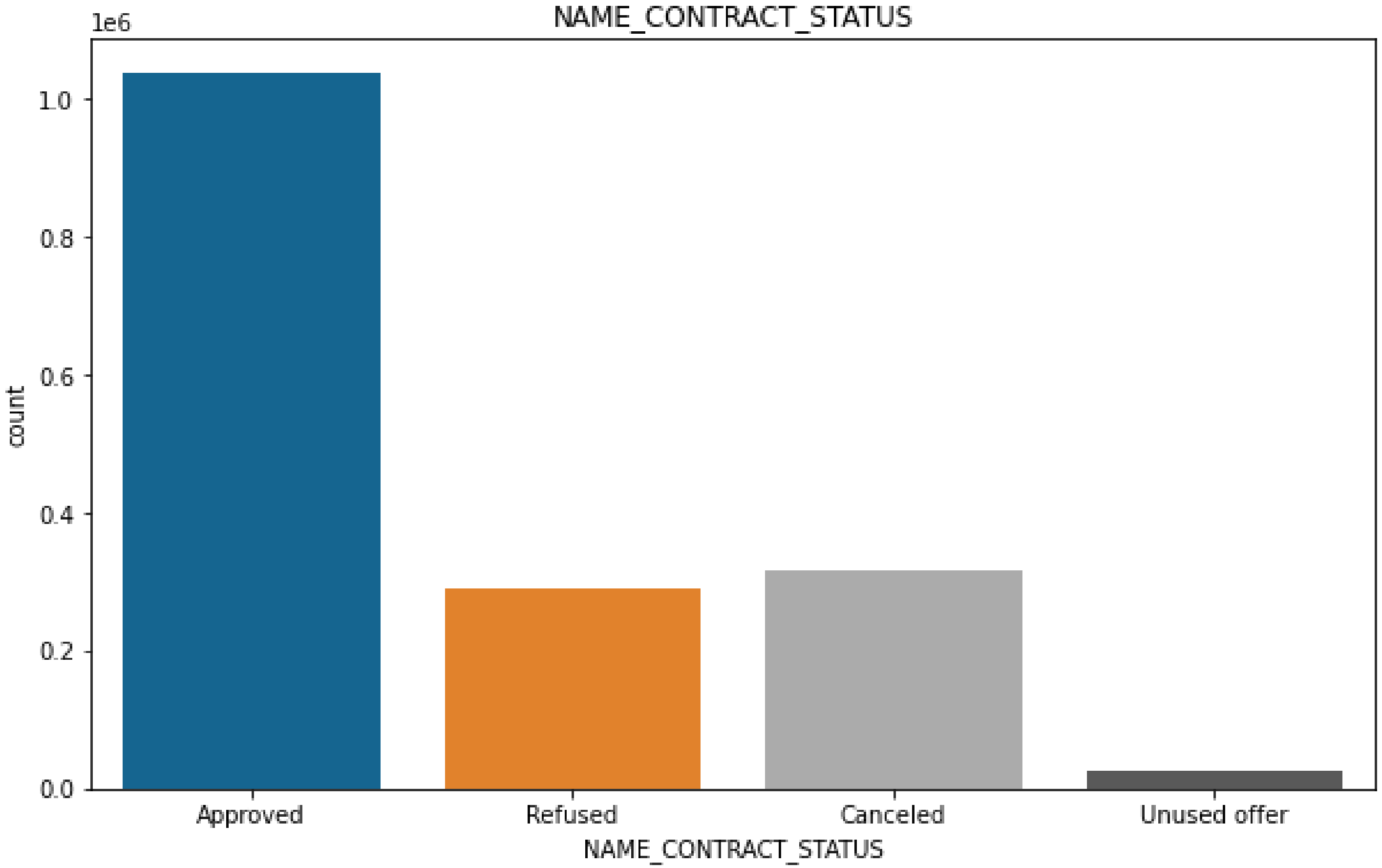
Number of Enquiries to Credit Bureau about the client correlation analysis



Inference: All these features show little or no correlation and hence can be ignored from analysis

Univariate Analysis

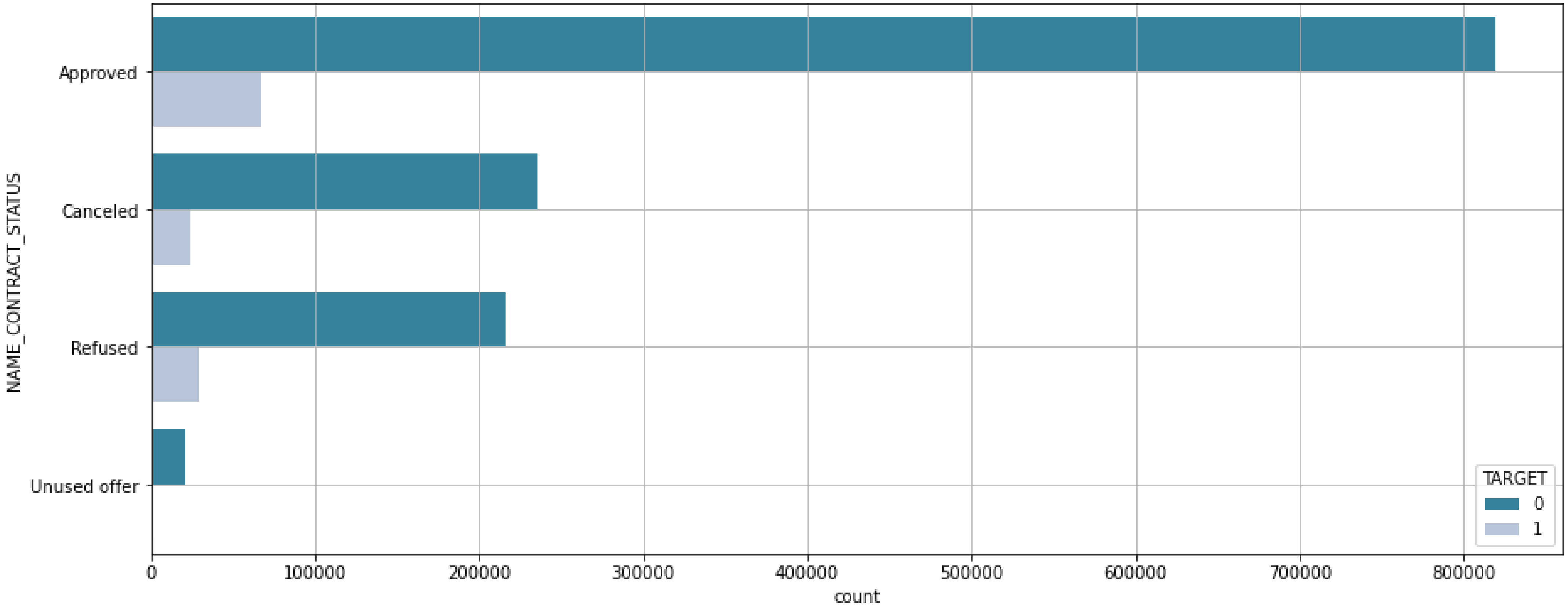
Status of Previous loans - Previous application Data



Inference: 62% of the loans had been approved and the rest have been rejected or cancelled.

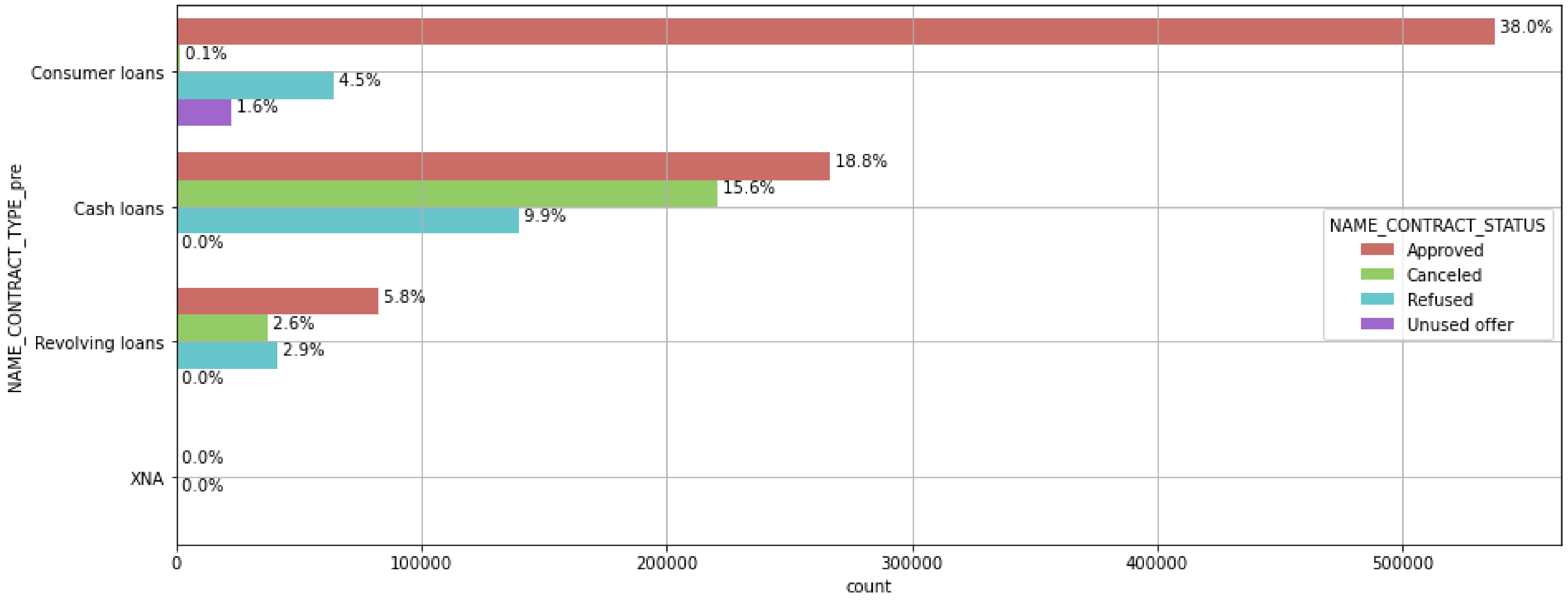
Segmented Univariate Analysis – Merged Data

Status of previous loan VS Current Defaulters and Non-Defaulters



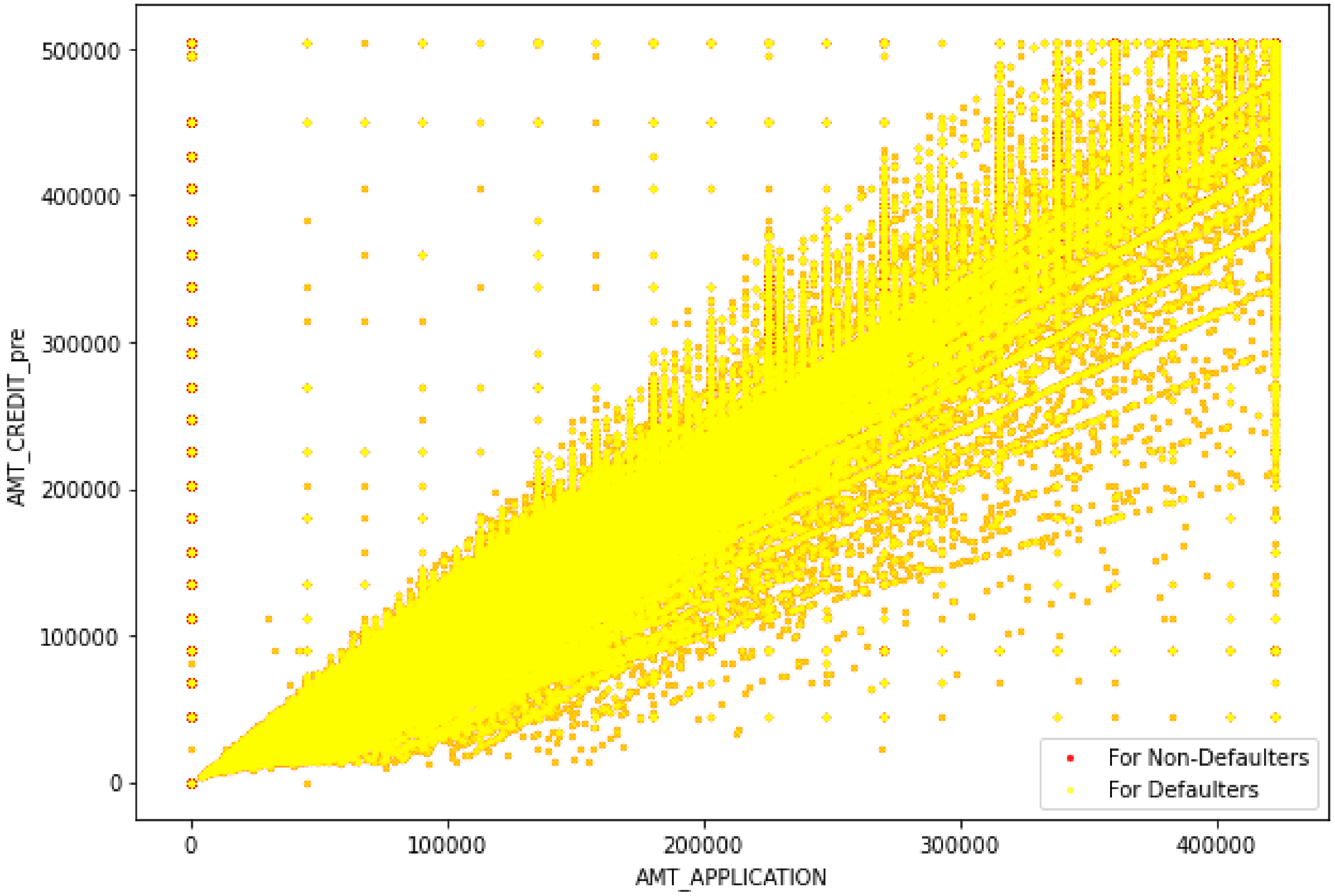
Percentage of previously approved loan applicants that paid loan amount on time : 92.41%
Percentage of previously approved loan applicants that defaulted in current loan : 7.59%
Percentage of previously refused loan applicants that paid loan amount on time : 24.37%
Percentage of previously refused loan applicants that defaulted in current loan : 3.32%
Inference: Study shows that customers who had previously been approved of loans are most likely to repay them again currently without delays
The customers who were rejected loans previously show they are more likely to be non-defaulters next time.

Status of previous loan VS Type of Loans



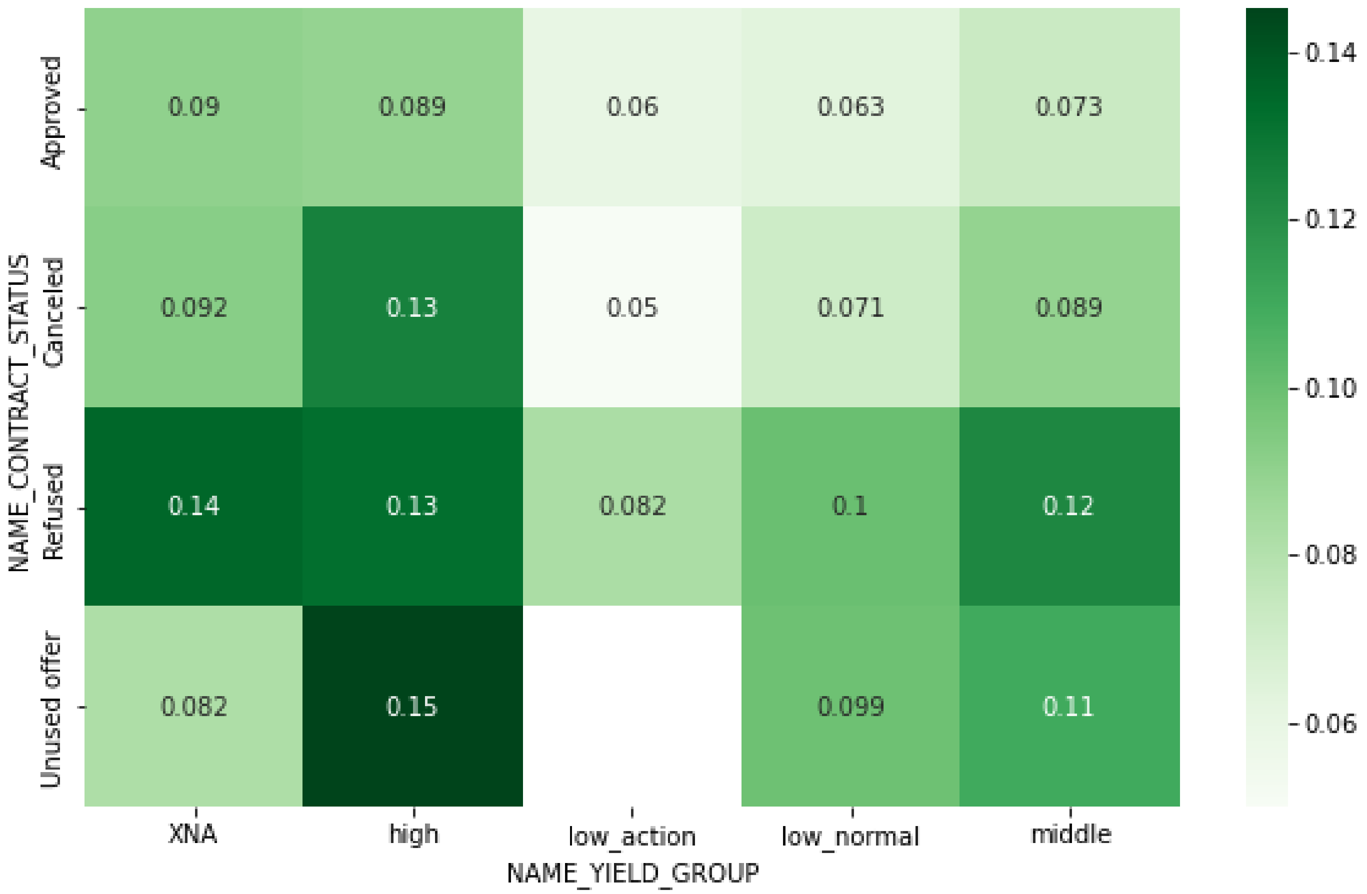
Inference: Mostly Consumer loans(38%) have been approved previously with little to no cancellations and very few have been refused.
Where as Cash loans have always faced many approvals(18.84%) with half of them being rejected(9.87%).
Cash loans have been cancelled the most at 15.6%.

Status of previous loan VS Type of Loans



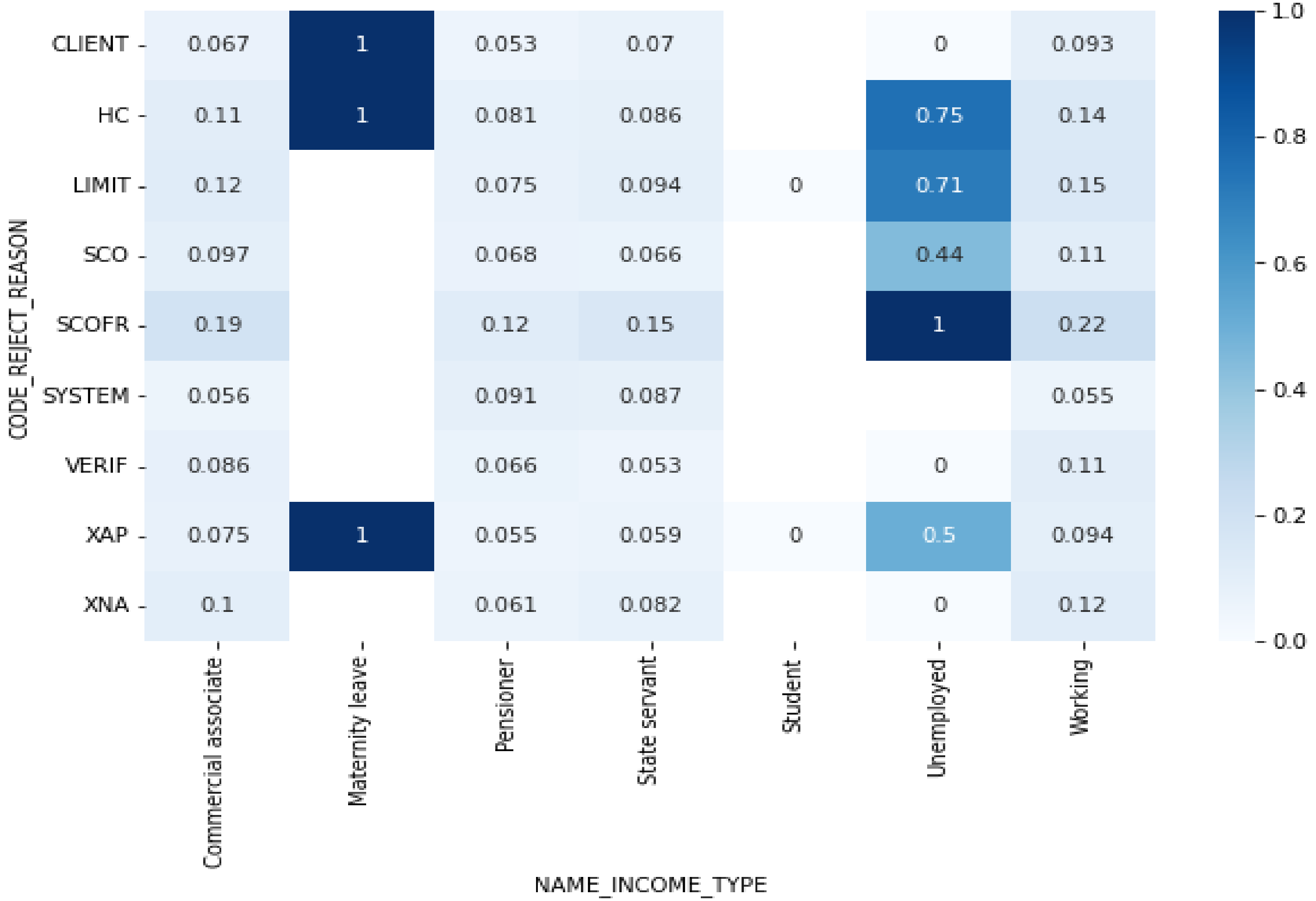
Inference: Shows a positive correlation between Amount Credit and Amount applied for loan in case of defaulters and also shows in areas increased credit amount. Whereas for Non- Defaulters the bank was right on mostly reducing the credit amount for high application amount.

Name yield group VS Status of Previous Loans



Inference: High yield group who had unused offer or refused or cancelled the loans due to any reason are seen to be most defaulters followed by middle yield group

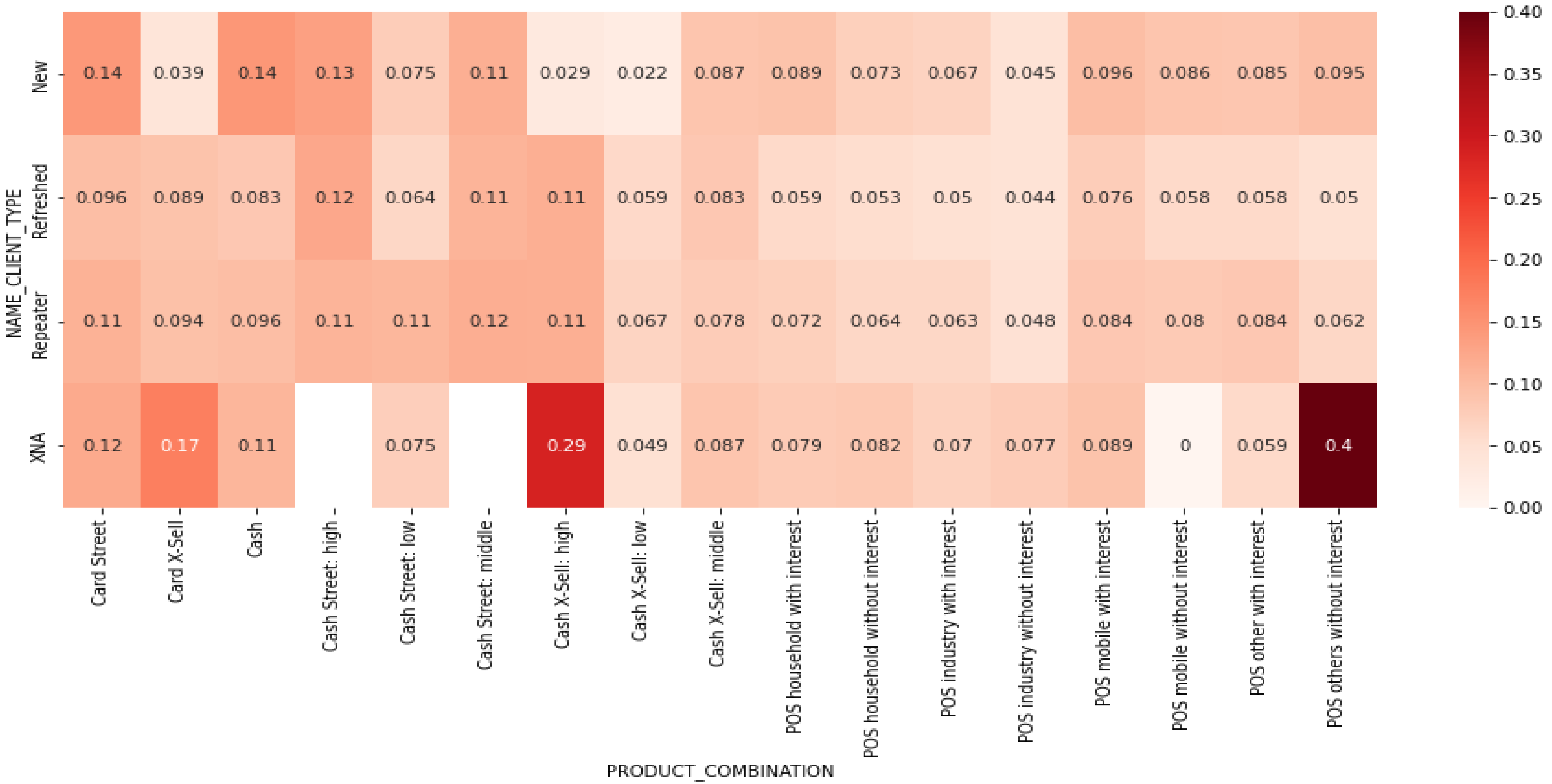
Income type VS Rejection reason for Previous Loans



CODE_REJECT_REASON	Commercial associate	Maternity leave	Pensioner	State servant	Student	Unemployed	Working
CLIENT	0.067	1	0.053	0.07		0	0.093
HC	0.11	1	0.081	0.086		0.75	0.14
LIMIT	0.12		0.075	0.094	0	0.71	0.15
SCO	0.097		0.068	0.066		0.44	0.11
SCOFR	0.19		0.12	0.15		1	0.22
SYSTEM	0.056		0.091	0.087			0.055
VERIF	0.086		0.066	0.053		0	0.11
XAP	0.075	1	0.055	0.059	0	0.5	0.094
XNA	0.1		0.061	0.082		0	0.12

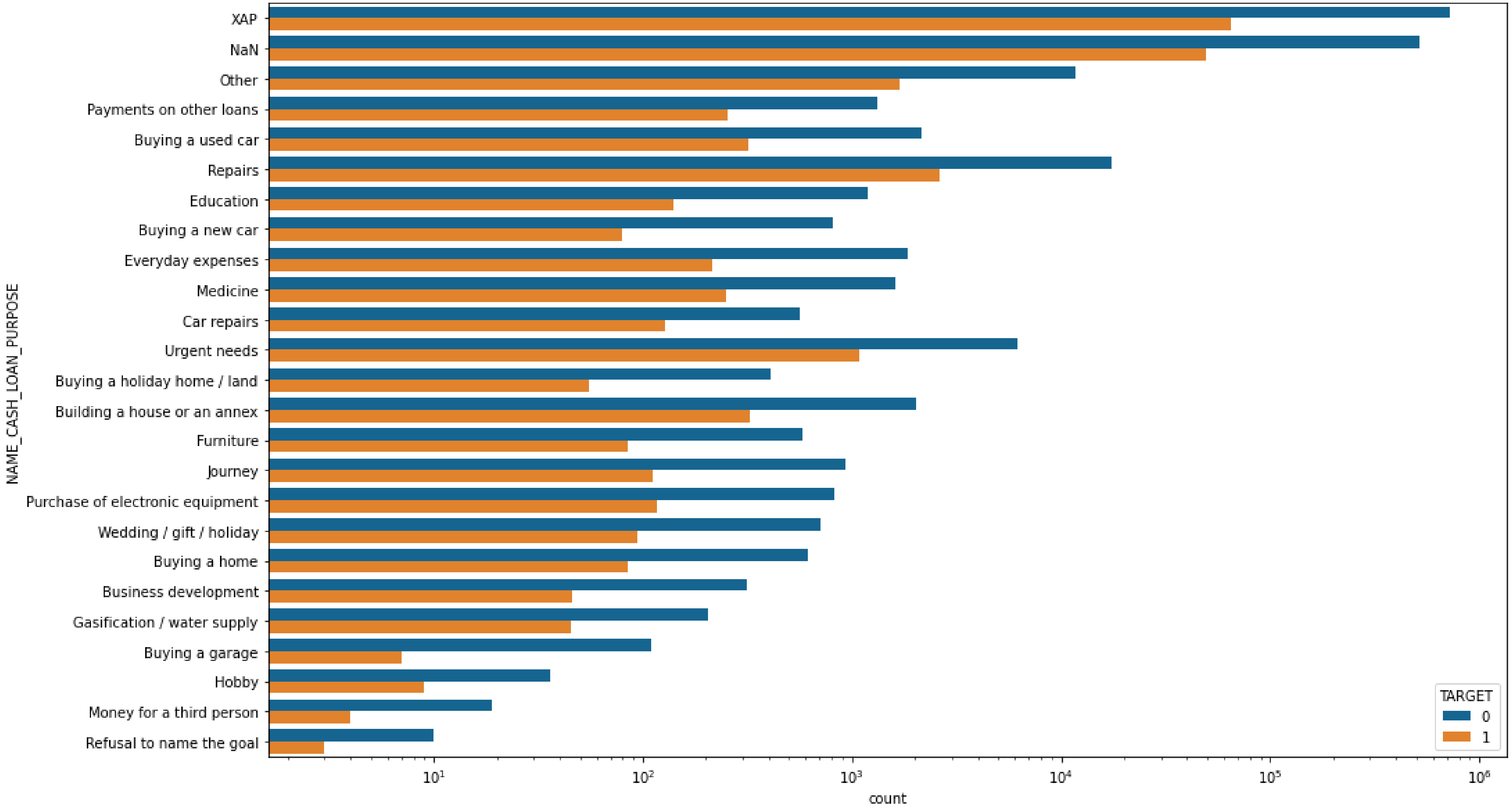
Inference: Most defaulters are customers on maternity leave whose loan have been previously rejected due to HC, CLIENT and XAP reason followed by Unemployed customers with most common reasons for rejection being HC, LIMIT and SCOFR. Tells us that people on maternity leave and Unemployed people have been most defaulters.

Client type VS Product Combination



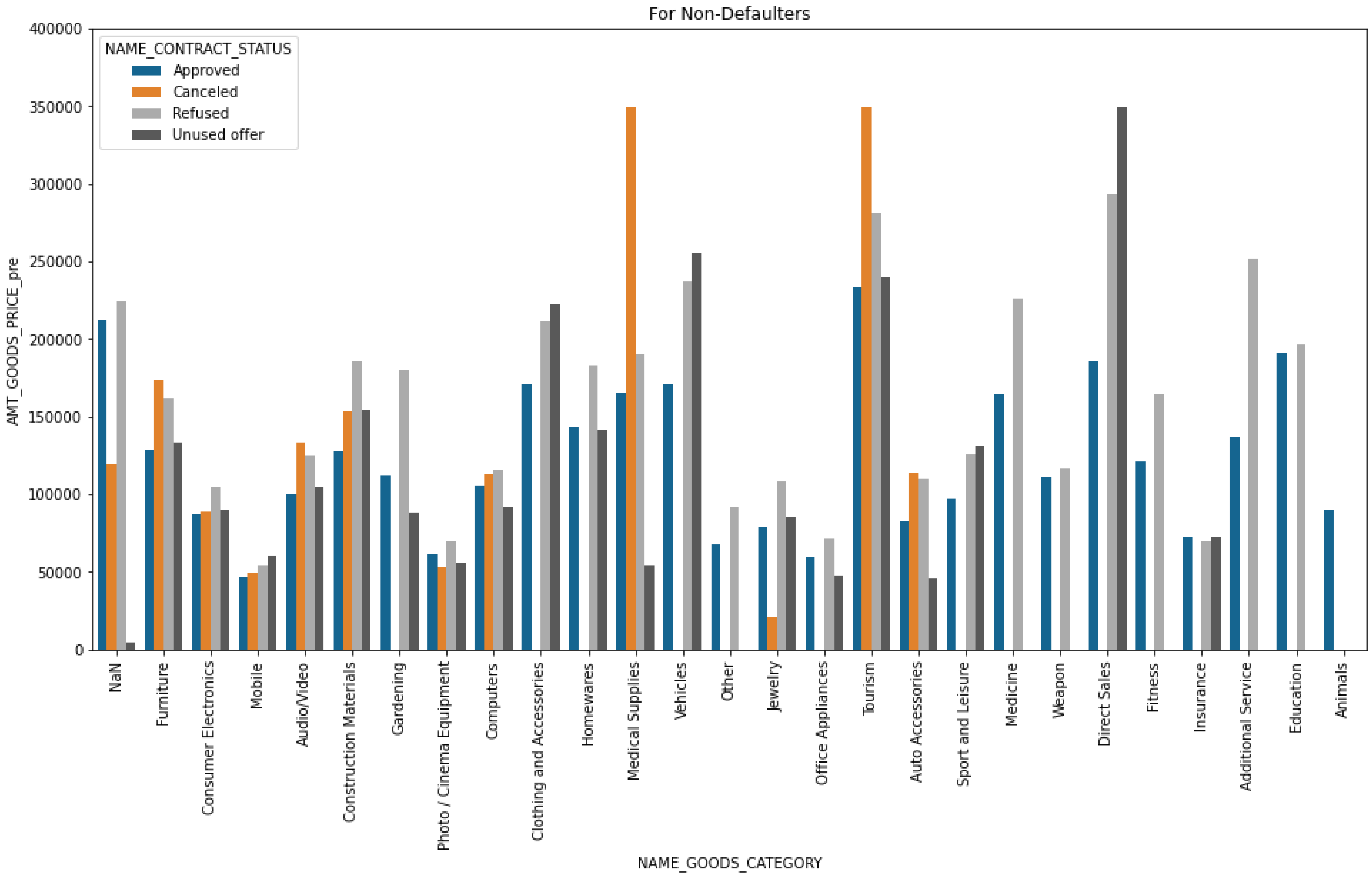
Inference: New Clients with Cash X-sell:low and Cash X-sell:high product combinations are the least likely to default. Highest default chances are in Card Street

Purpose of taking previous loans for defaulters & non-defaulters



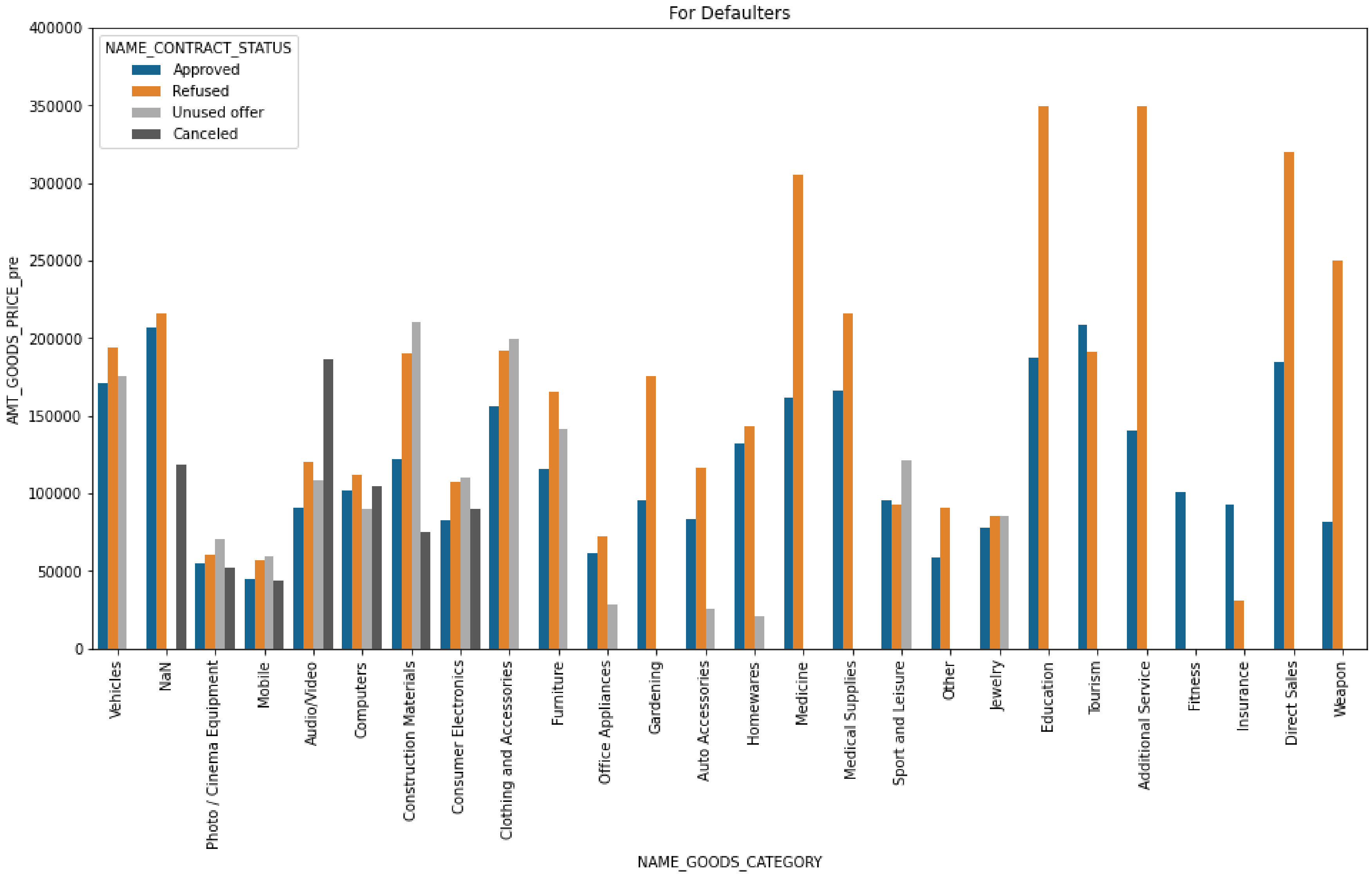
Inference: Customers taking Education Loans and Urgent needs and Car loans are most likely to payback on time.

Amount of Goods Price VS Goods category for Non- Defaulters in current application



Inference: Bank can approve loans for Tourism category and Medical supplies at high goods price as they are most likely to payback on time

Amount of Goods Price VS Goods category for Defaulters in current application



Inference: Bank is right in rejecting Education and Additional service loans at a high goods price as they are likely to default

Summary

- Study shows that customers who had previously been approved of loans are most likely to repay them again currently without delays .
- Mostly Consumer loans(38%) have been approved previously with little to no cancellations and very few have been refused. Where as Cash loans have always faced many approvals(18.84%) with half of them being rejected(9.87%). Cash loans have been cancelled the most at 15.6%. Safer for banks to give consumer loans.
- Most defaulters are customers on maternity leave whose loan have been previously rejected due to HC, CLIENT and XAP reason followed by Unemployed customers with most common reasons for rejection being HC, LIMIT and SCOFR. Tells us that people on maternity leave and Unemployed people have been most defaulters.
- New Clients with Cash X-sell:low and Cash X-sell:high product combinations are the least likely to default. Highest default chances are in Card Street
- Shows a positive correlation between Amount Credit and Amount applied for loan in case of defaulters and also shows in areas increased credit amount. Whereas for Non- Defaulters the bank was right on mostly reducing the credit amount for high application amount.
- Bank can approve loans for Tourism category and Medical supplies at high goods price as they are most likely to payback on time.
- Bank is right in rejecting Education and Additional service loans at a high goods price as they are likely to default.

Summary

- Cash loans are more likely to be given out by the bank and most likely to be repaid showing 90% as compared to 9% in Revolving Loans
- Working Professionals are most likely to be non-defaulters in loans followed by Commercial associates.
- Married customers are most likely to repay loan amount as compared to any other category.
- Customer having Secondary Education are most likely to pay back on time followed by customers with Higher Education. Very few percentage of loans have been given to Customers with Academic degrees and Bank is right on that since Academic college degree holders are most likely to default.
- Customers with low income range(0-150000) have been given maximum loans and are more likely to pay them back. Customers with low income range are most Non-defaulters and maximum loans are given to them with low annuity amount.
- Customers with 0-3 yrs working are most likely to repay loans closely followed by people with 10+ yrs of experience
- Majority of NON-defaulters are Managers who get high loan annuity amount. In defaulters we see HR staff getting high annuity loan amount.

Conclusion

Customers who can be safely given loan to:

- Working professionals Having Secondary or Higher Education aged between 25-55 yrs with about 3 yrs working experience.
- People who have more number of people in their family and are married.
- Managers can be given High annuity amount
- Previous customers with approved loans
- Providing Consumer loans
- Product combinations Cash X-sell: low and high
- Customers asking for low to medium application amount
- Goods category: Tourism and Medical Supplies at high goods price

Customers who should be avoided by the Bank:

- Academic college degree holders
- Hight annuity loan amount to HR staff as they are the most likely defaulters
- people on maternity leave and Unemployed people
- Providing Cash loans
- Product combination: Card street
- Education and Additional service loans at a high goods price