# Credit EDA Case Study

BY PATIBANDLA MANOGNYA

# PURPOSE

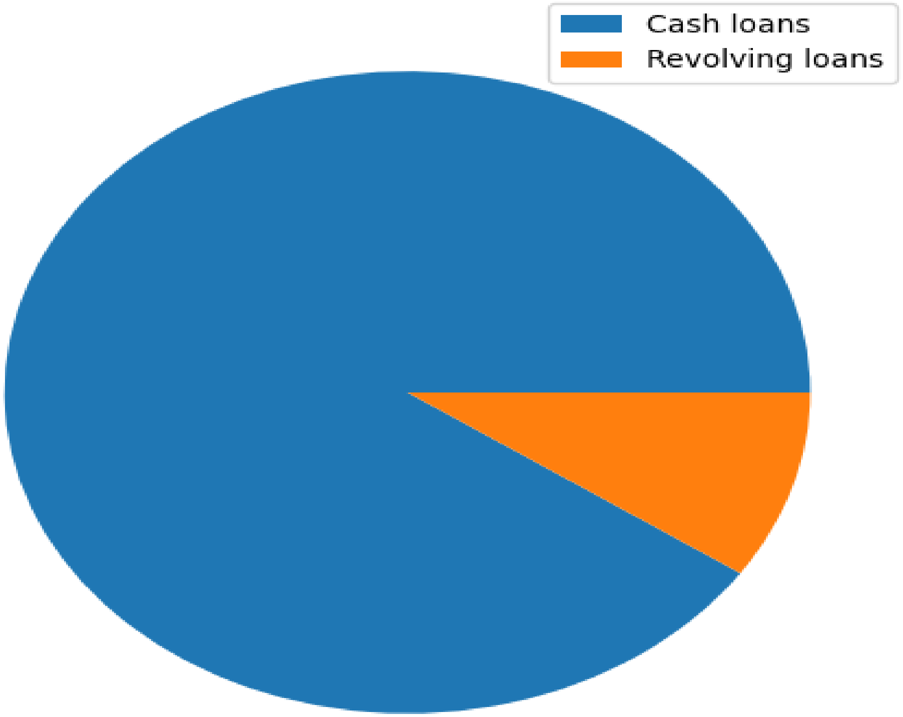
- Aim to identify patterns which indicate if a client has difficulty paying their instalments which may be used by bank for for taking actions such as
  - Denying the loan,
  - Reducing the amount of loan
  - Lending (to risky applicants) at a higher interest rate, etc.

- Driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

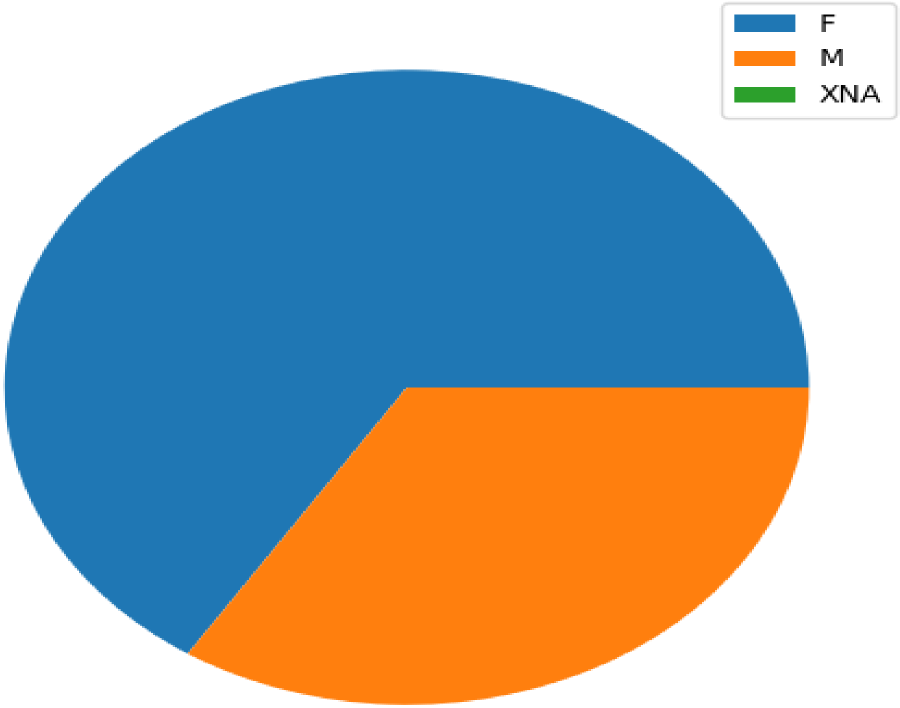
- To ensure customers that the customers capable of repaying the loans are not getting rejected.

# Current Application Data Analysis

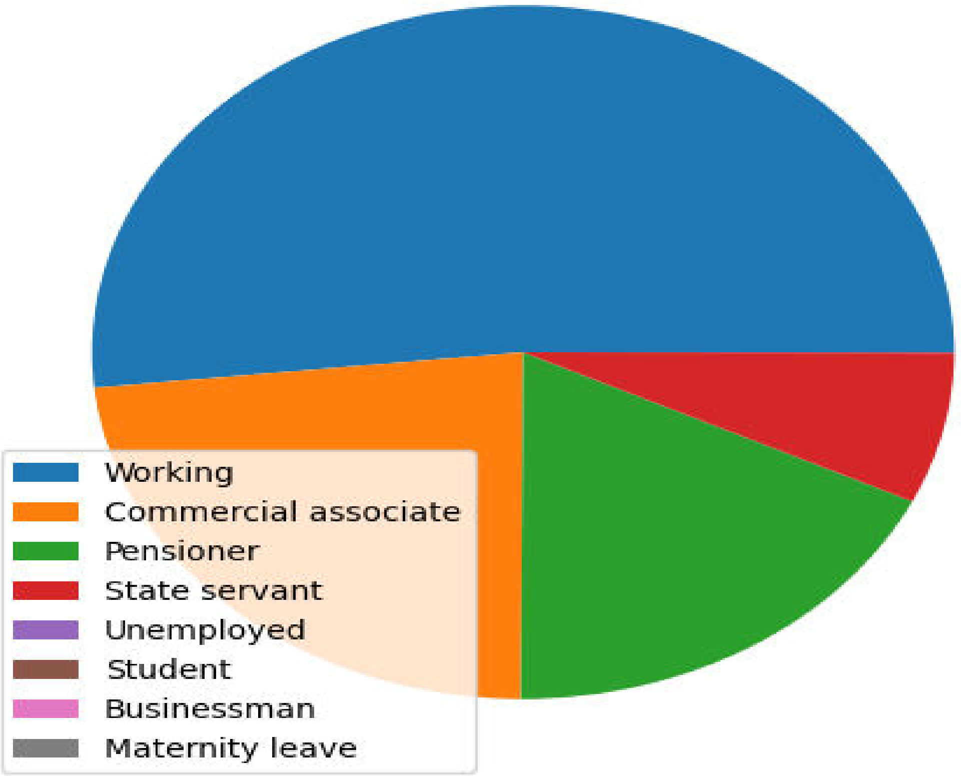
Contains all the information of the client at the time of application. The data is about whether a client has payment difficulties

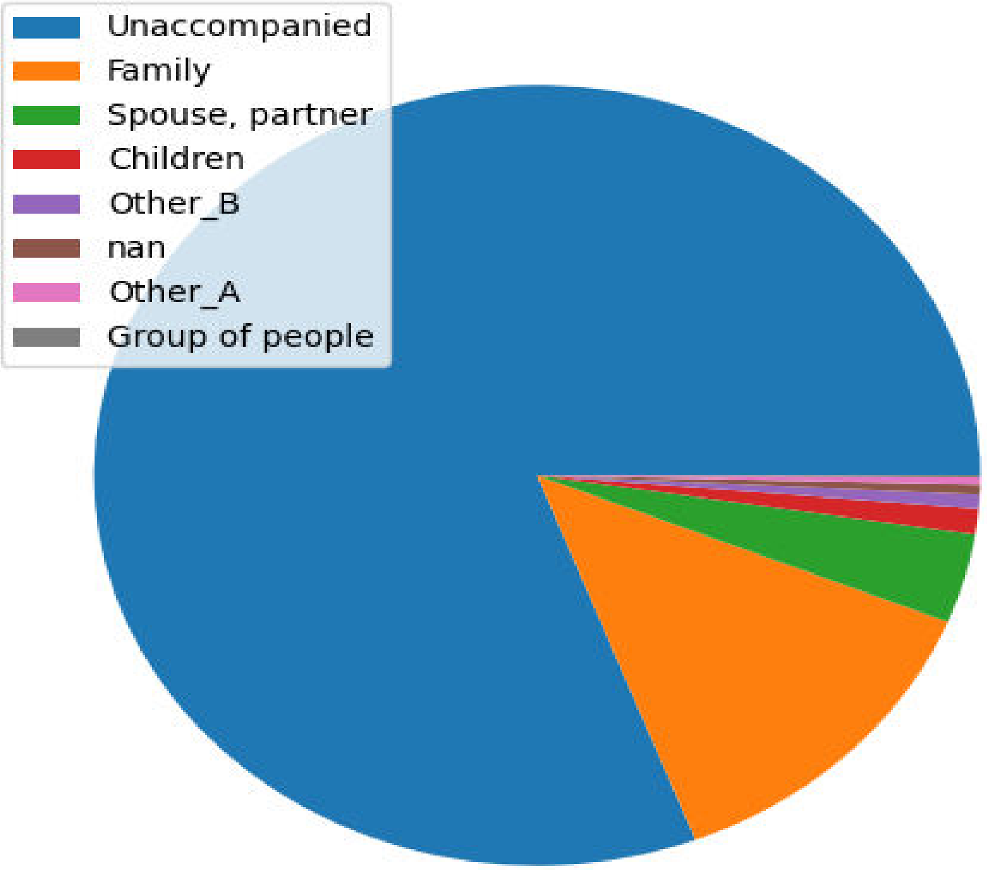


At 90%, cash loans are more common than revolving loans.



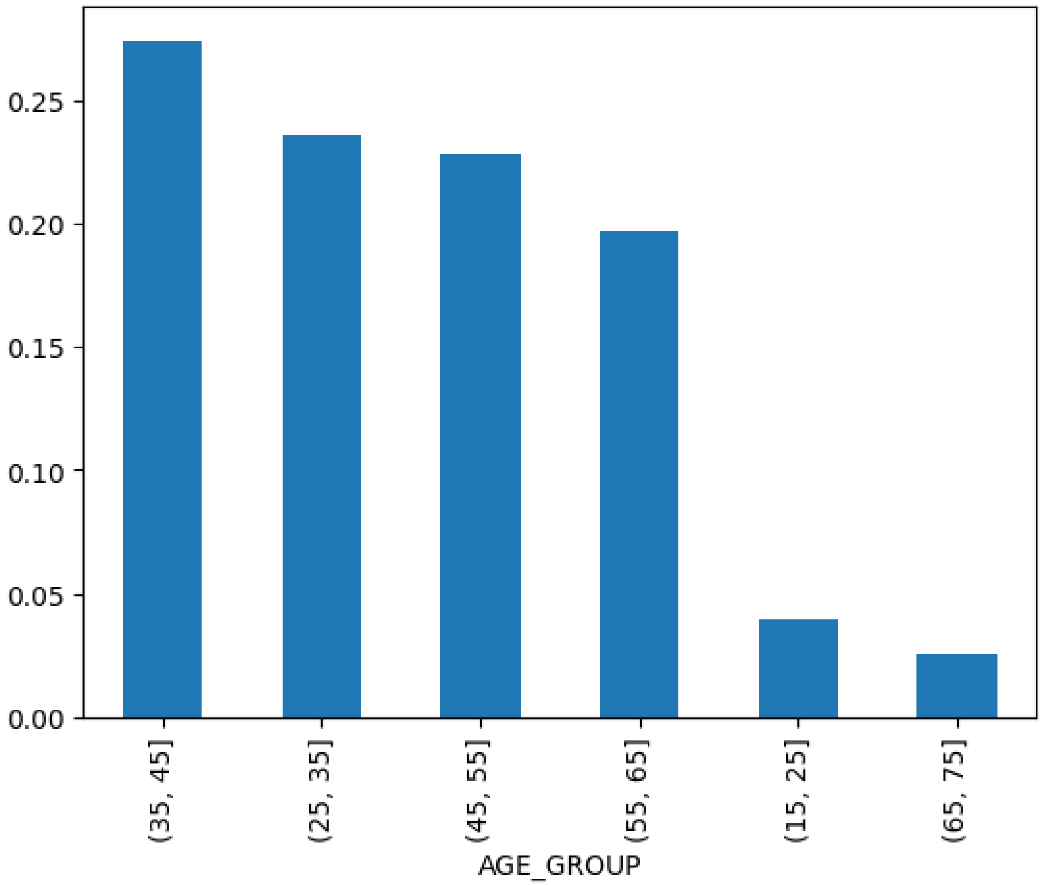
Of the total number of loan applicants, 65% are women and 34% are men.

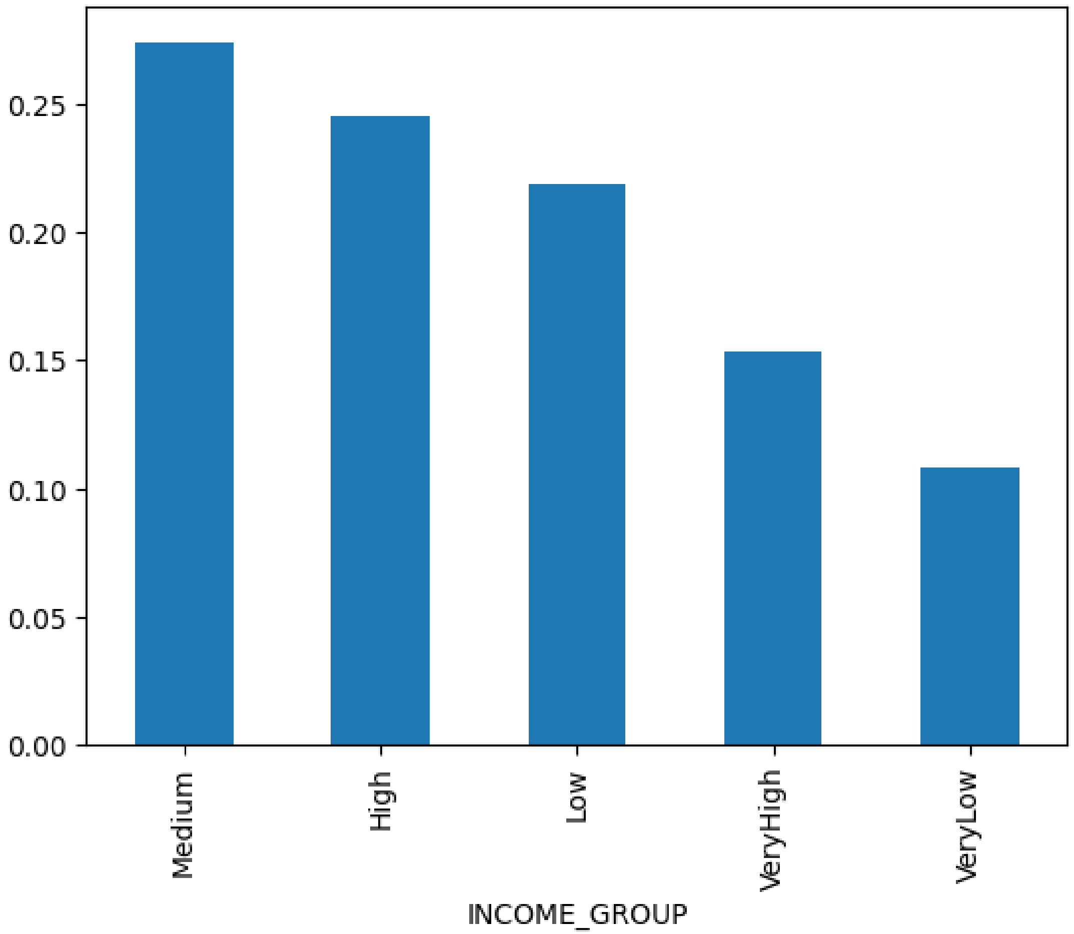




While working-class candidates make up the majority of the applicants, 18% are pensioners.

81% of applicants applied for loans with an accompanying person.

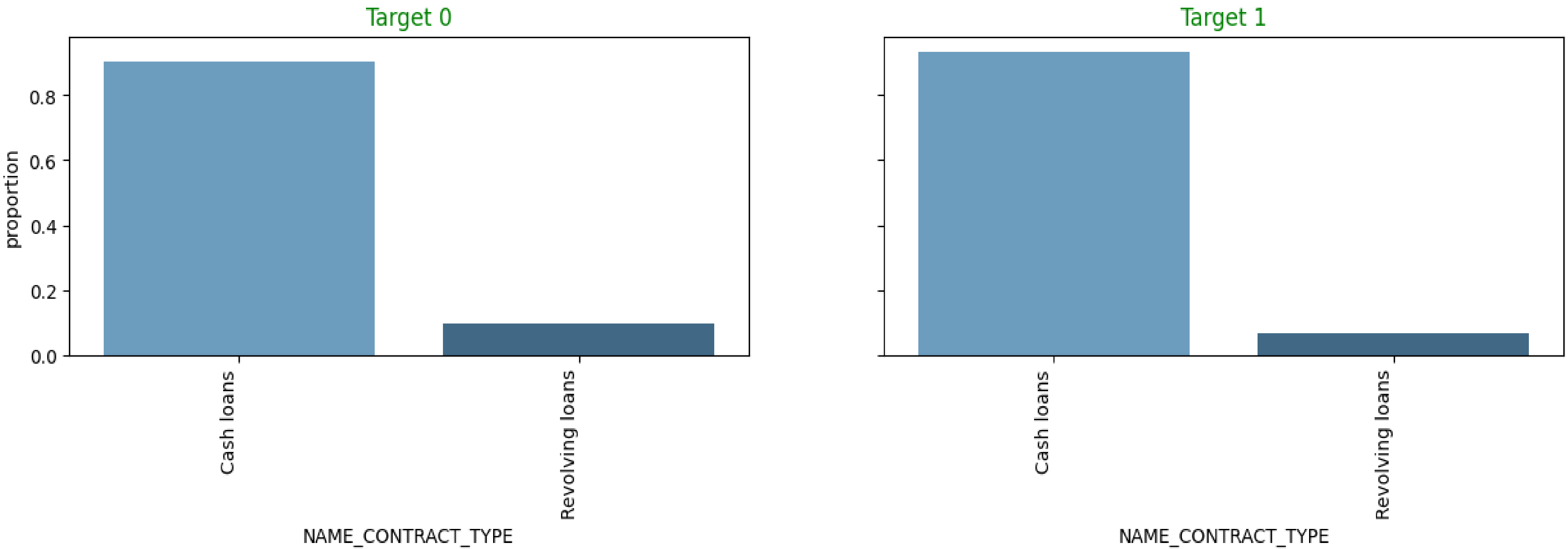




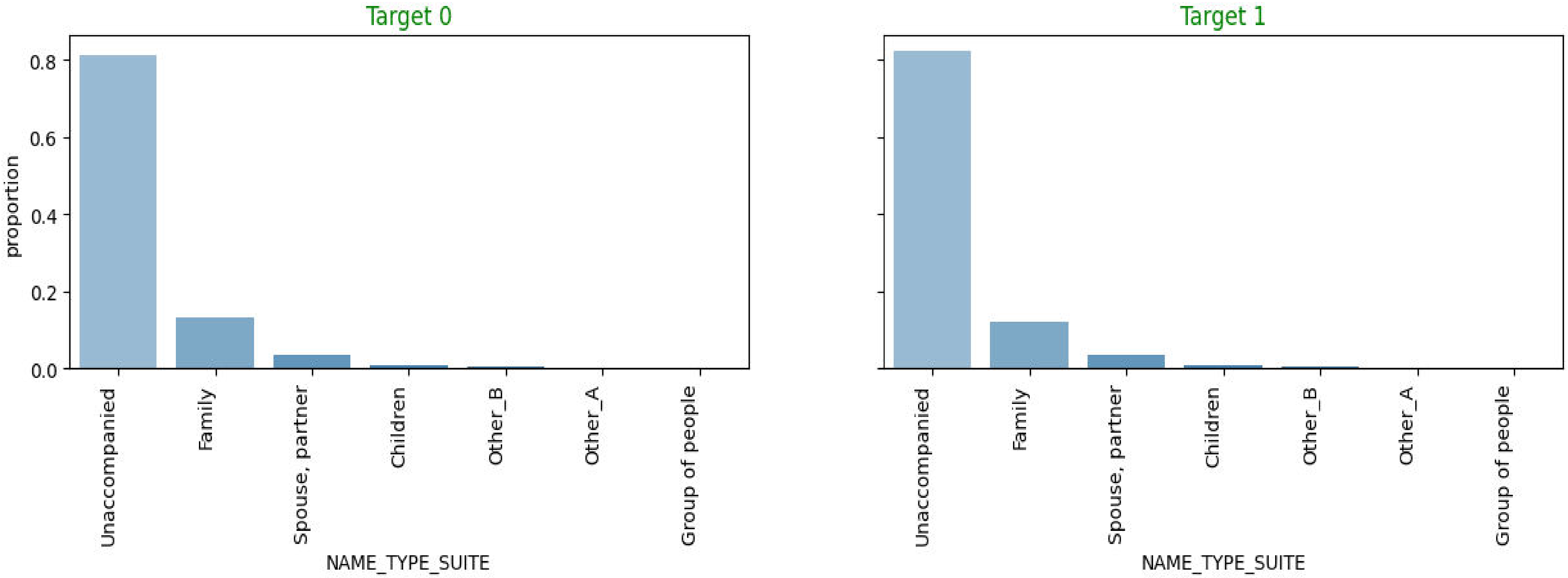
The age group that applies for loans the most frequently is 35–45. This could be explained by the age-related element of consumption.

The largest group seeking for loans is the group with medium incomes.

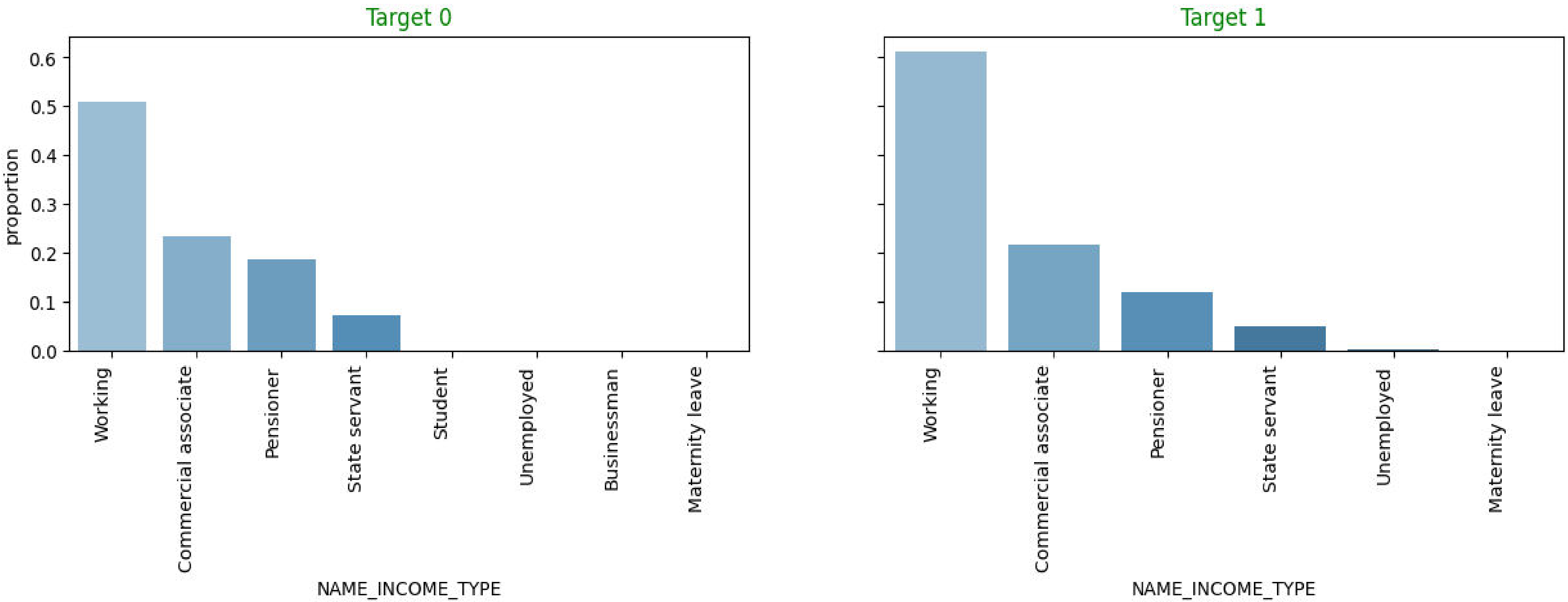
# Univariate Analysis on Categorical Nominal



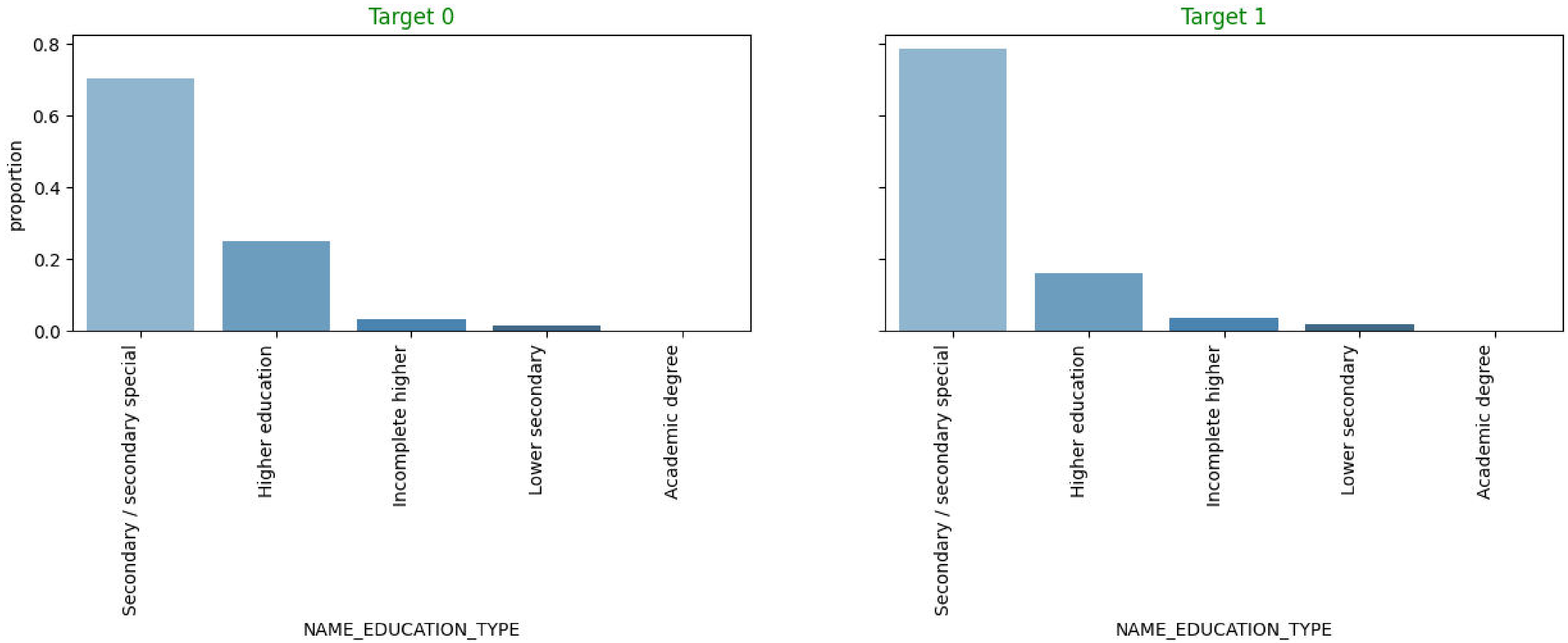
A sizable portion of the business's portfolio consists of cash loans. 85% for Target 0 and nearly 95% for Target 1.



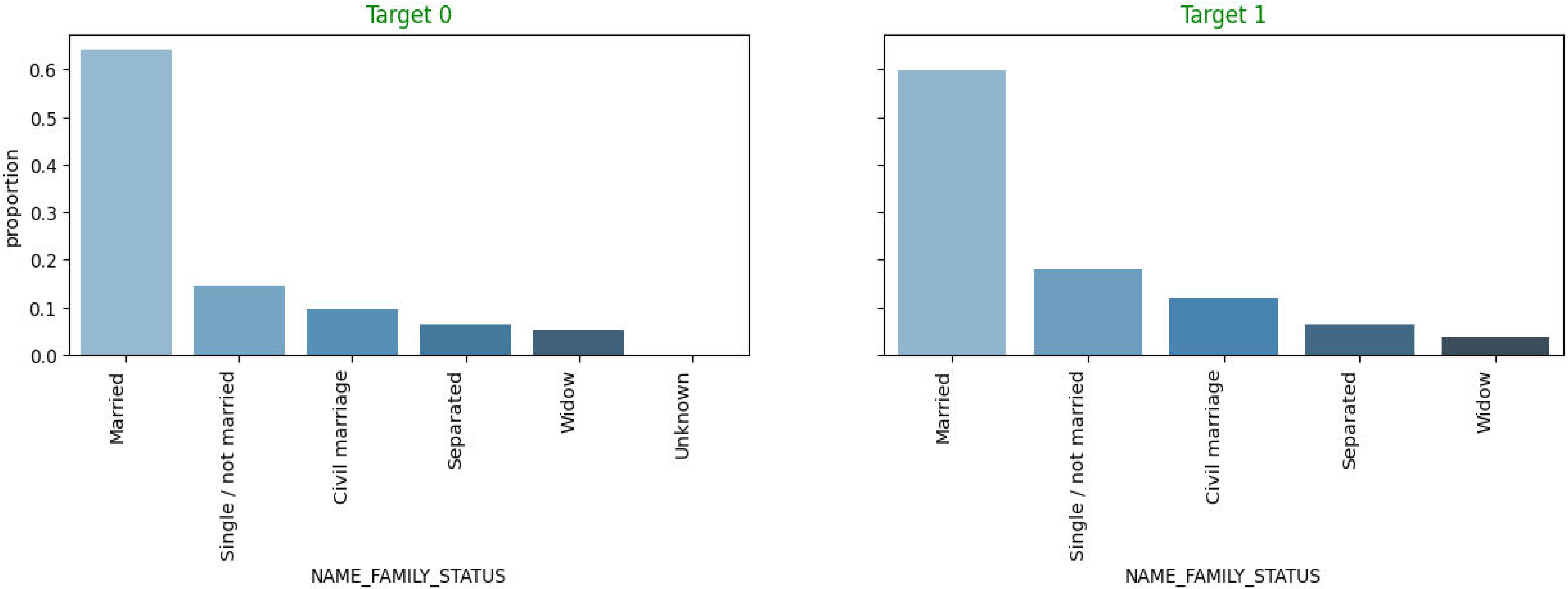
80–90% of Targets 0 and 1 are requesting unaccompanied loans. indicating that there is no way for this parameter to affect payment default.



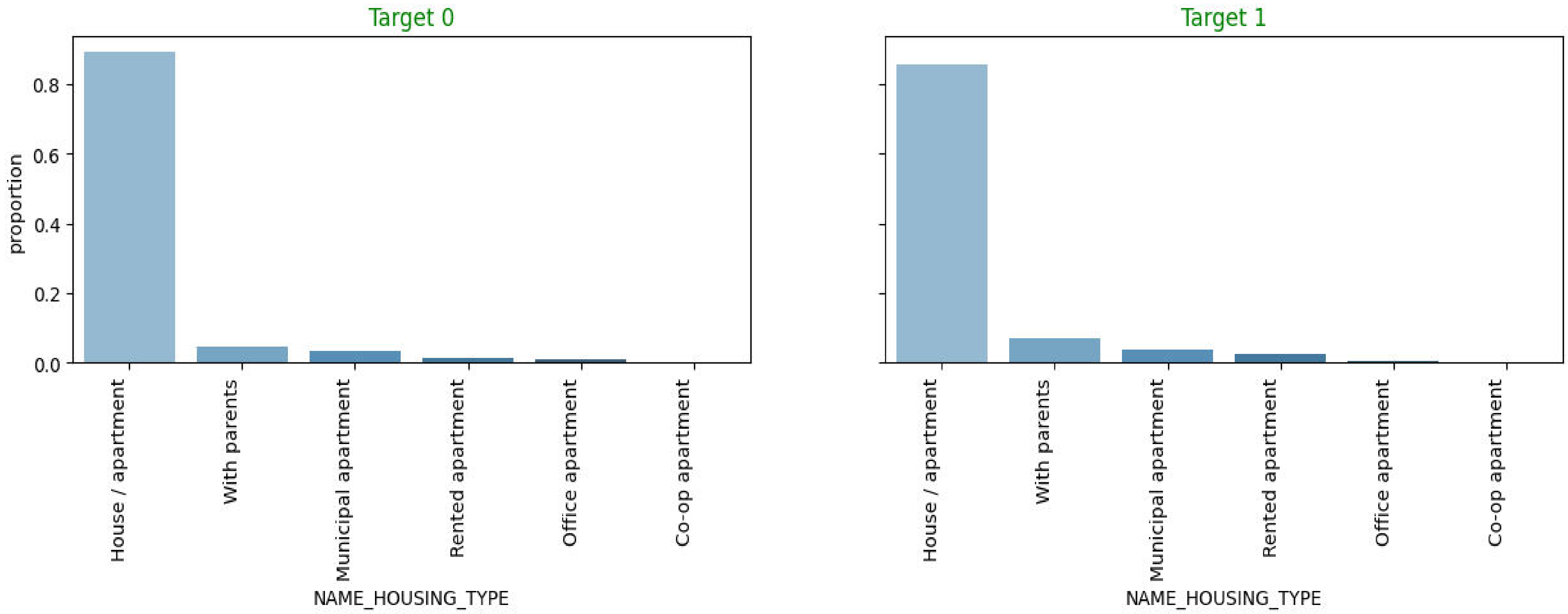
Working income types are 50% in the case of Target 0 and 60% in the case of Target 1.



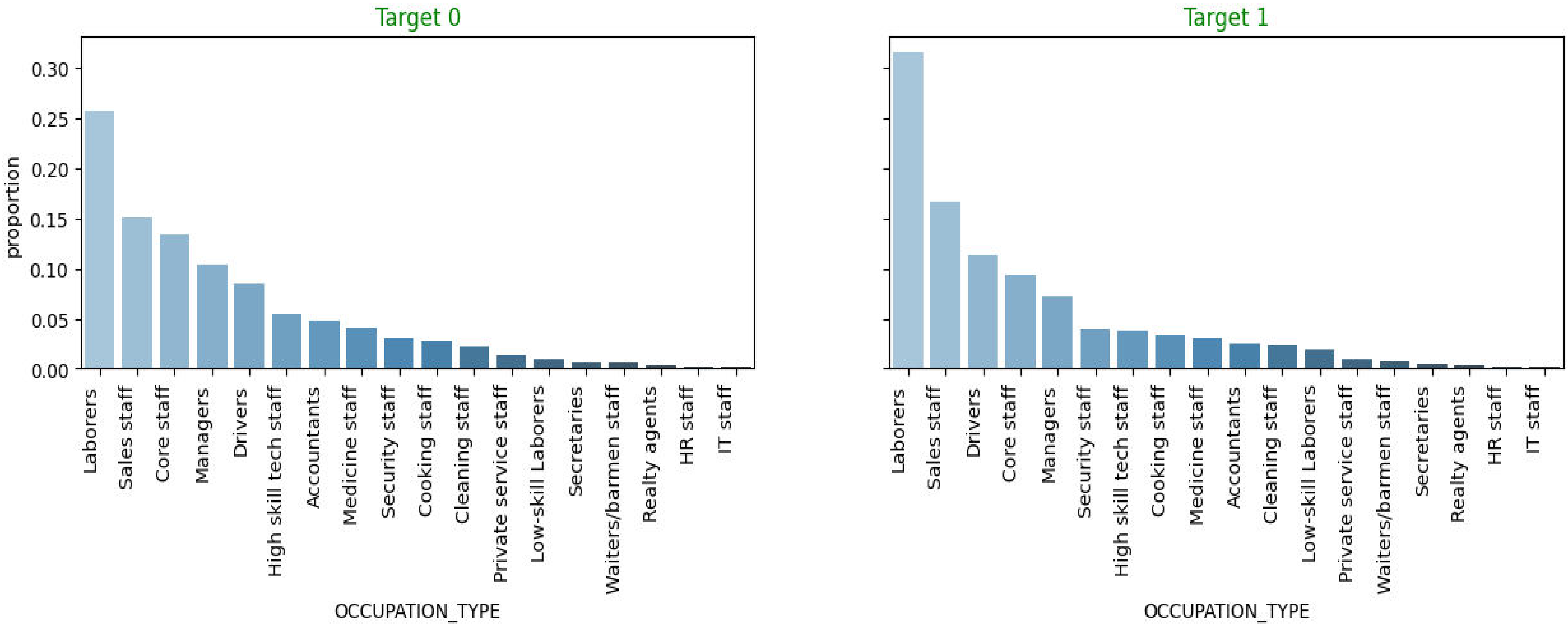
Applicants with a secondary education have sought for loans more frequently than applicants without one in Targets 0 and 1.



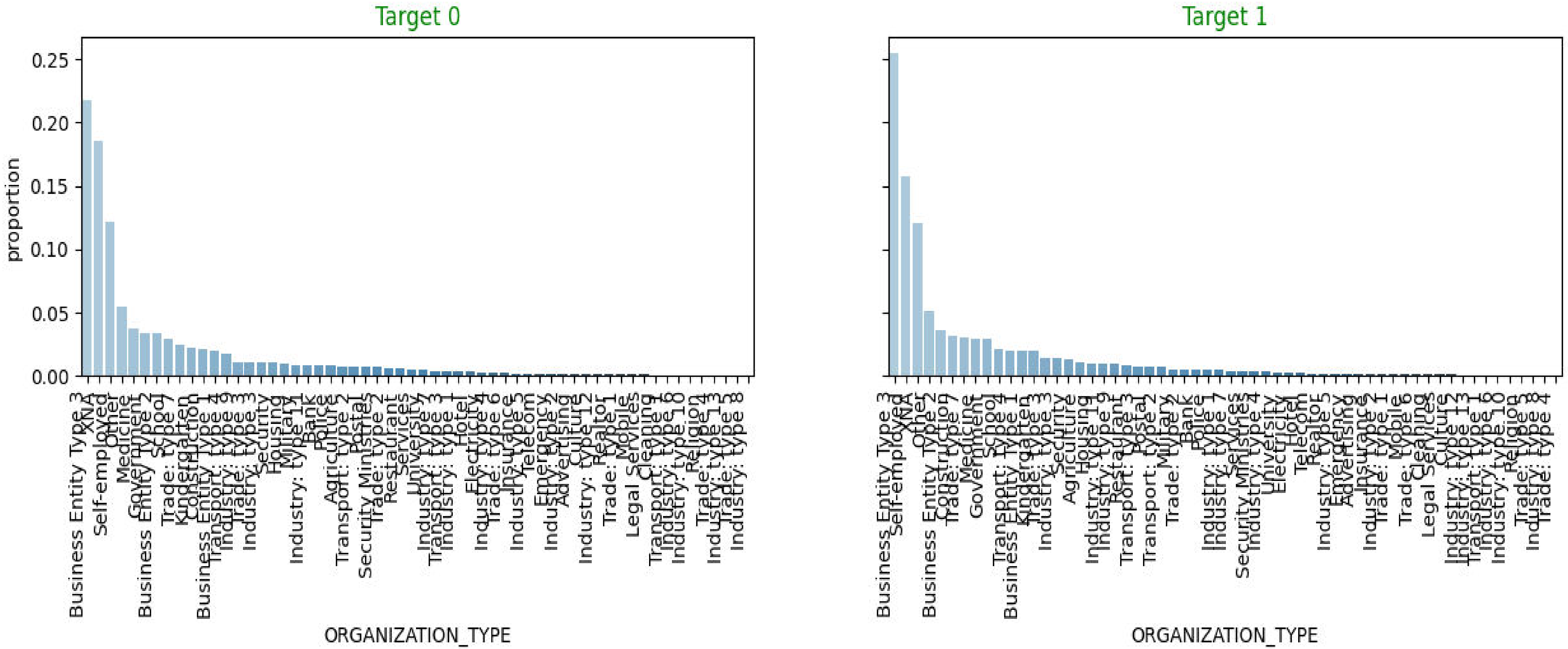
Of the Married applicants: Approximately 60% of them have not paid on time.



Of the applicants for Target 0 and Target 1, NAME_HOSUING_TYPE -85–90% reside in "House/apartment". Indicating that there is no way for this parameter to affect payment default.

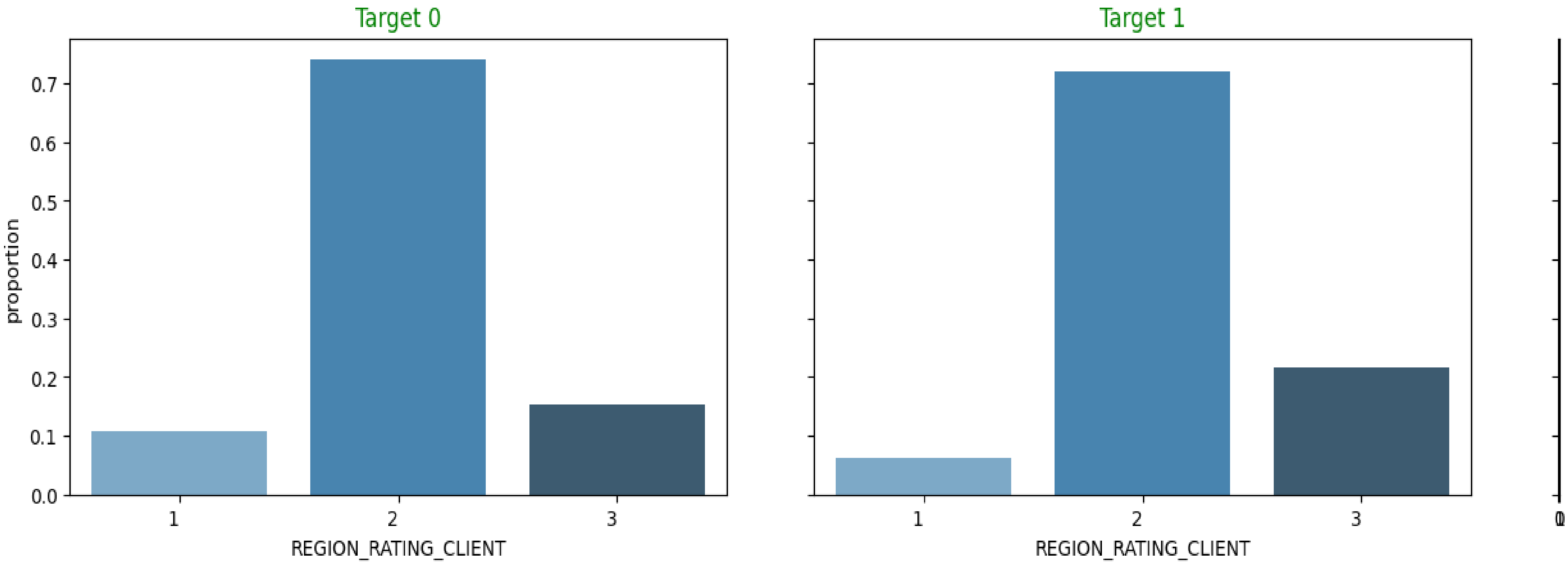


50% of defaulters are laborers, salespeople, core employees, and drivers. The largest percentage of applicants are also laborers.

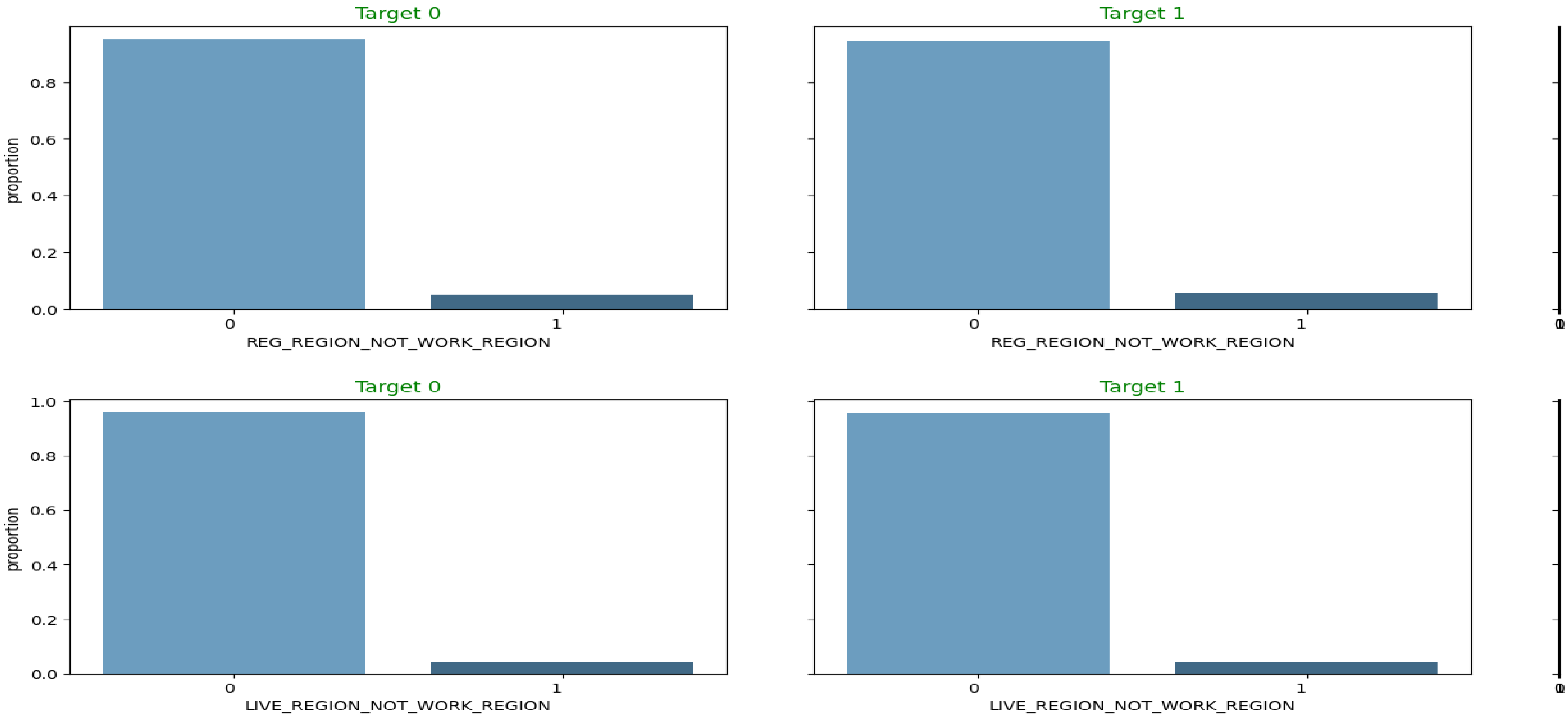


Self-employed and Business Entity Type 3 account for up to 40% of defaulters. This group also has the largest percentage of borrowers.

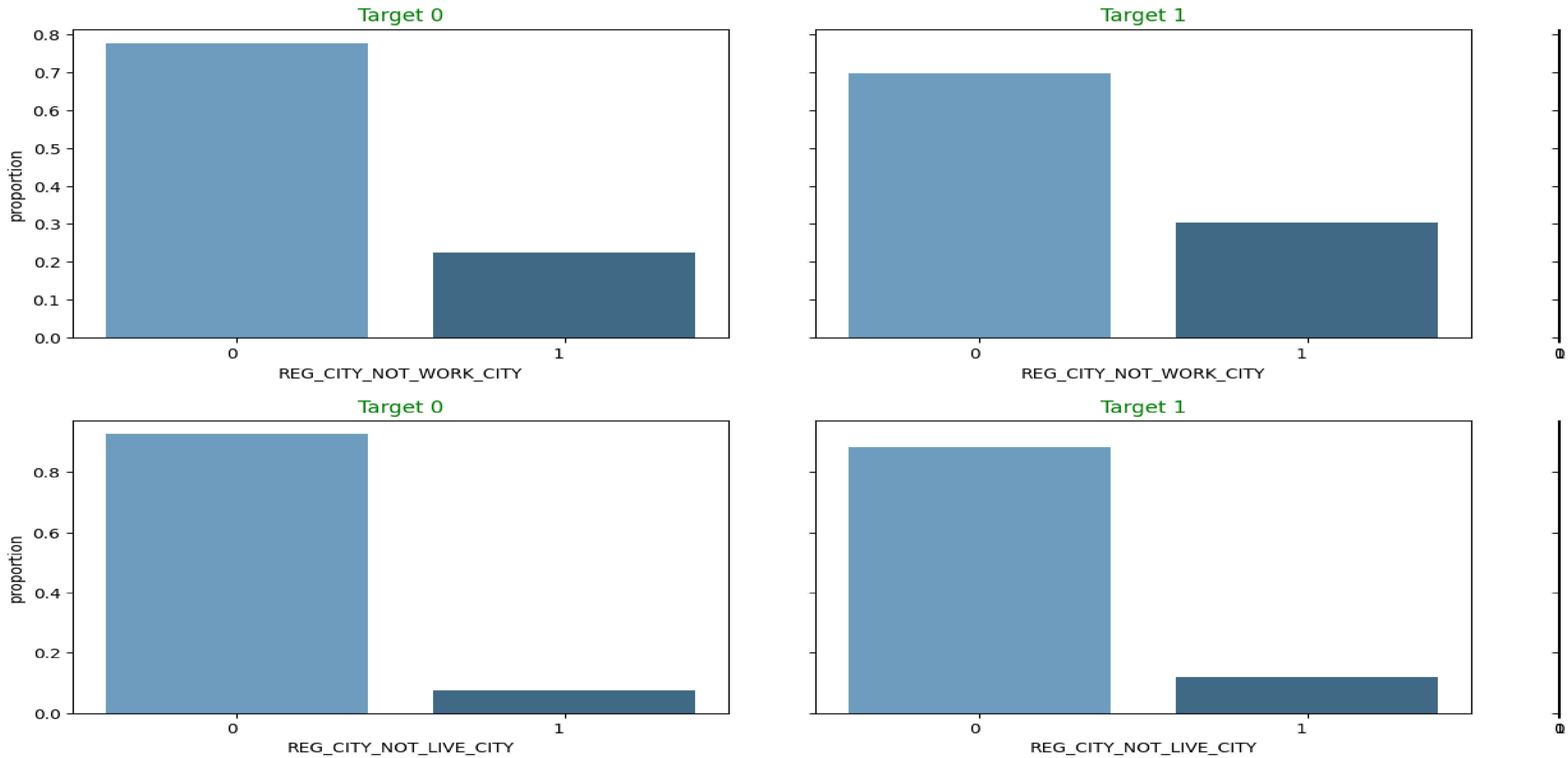
# Univariate Analysis on Categorical Ordered



Region 2 has the greatest percentage of applicants for both Target 0 and Target 1.

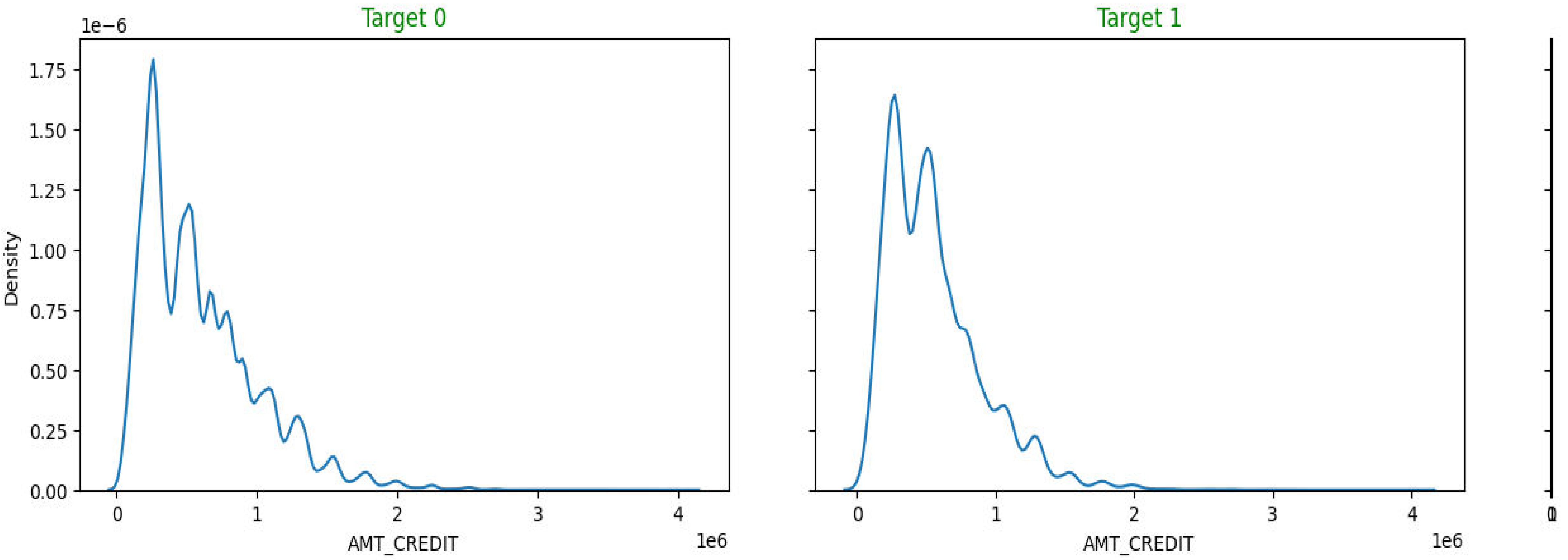


Target 1 is extremely low and does not appear to have an impact on the default rate for either Target 0 or Target 1 out of Region (REG_REGION_NOT_WORK_REGION, REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION).

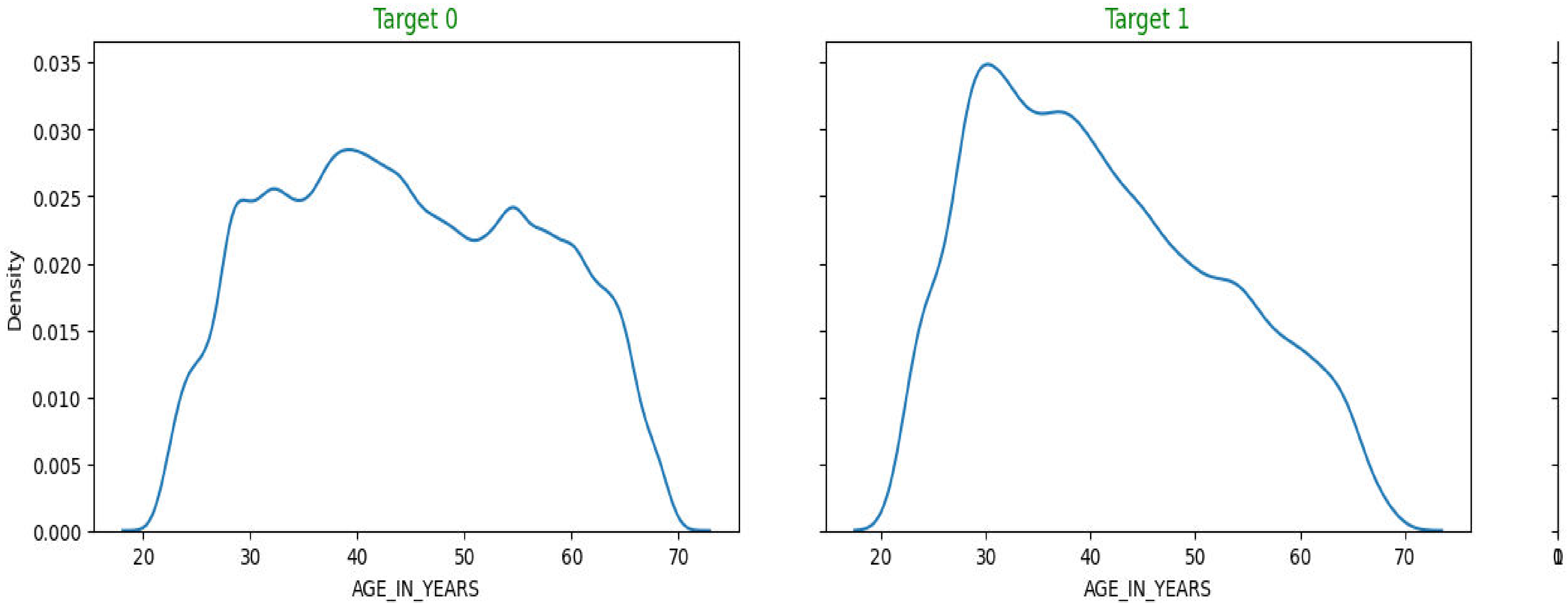


The default ratio for REG_CITY_NOT_WORK_CITY and REG_CITY_NOT_LIVE_CITY is higher at 1, meaning that they differ from permanent addresses.

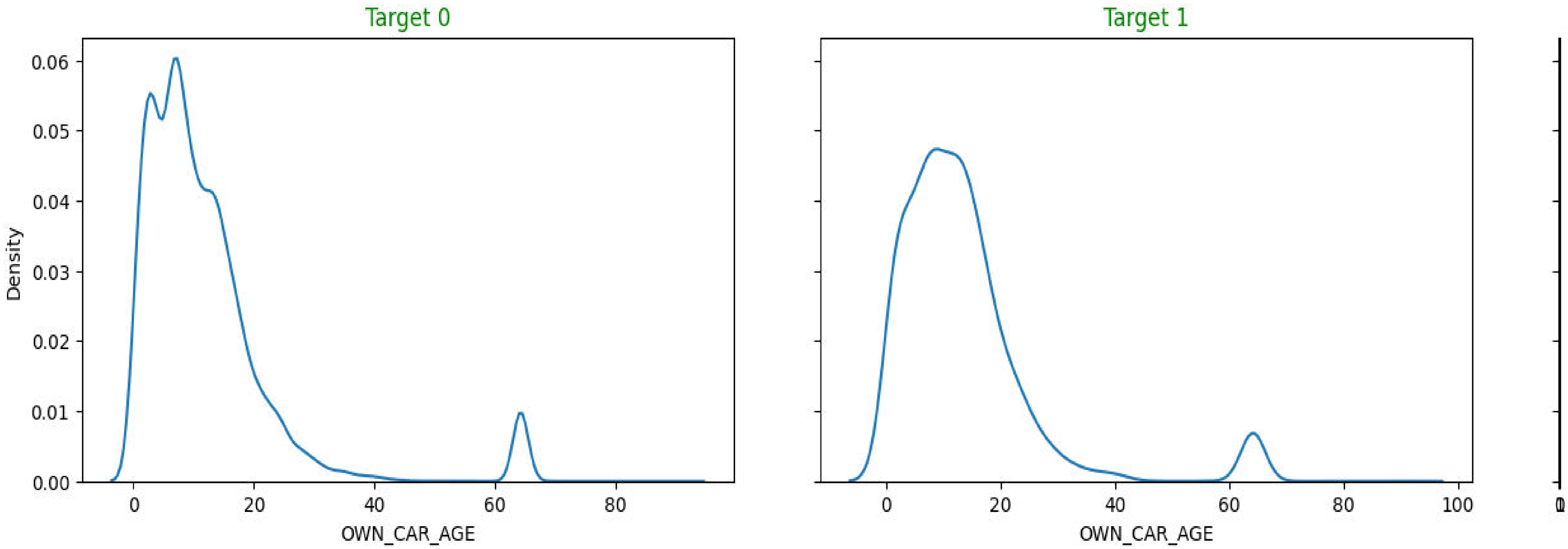
# Univariate Analysis on Continuous Variables



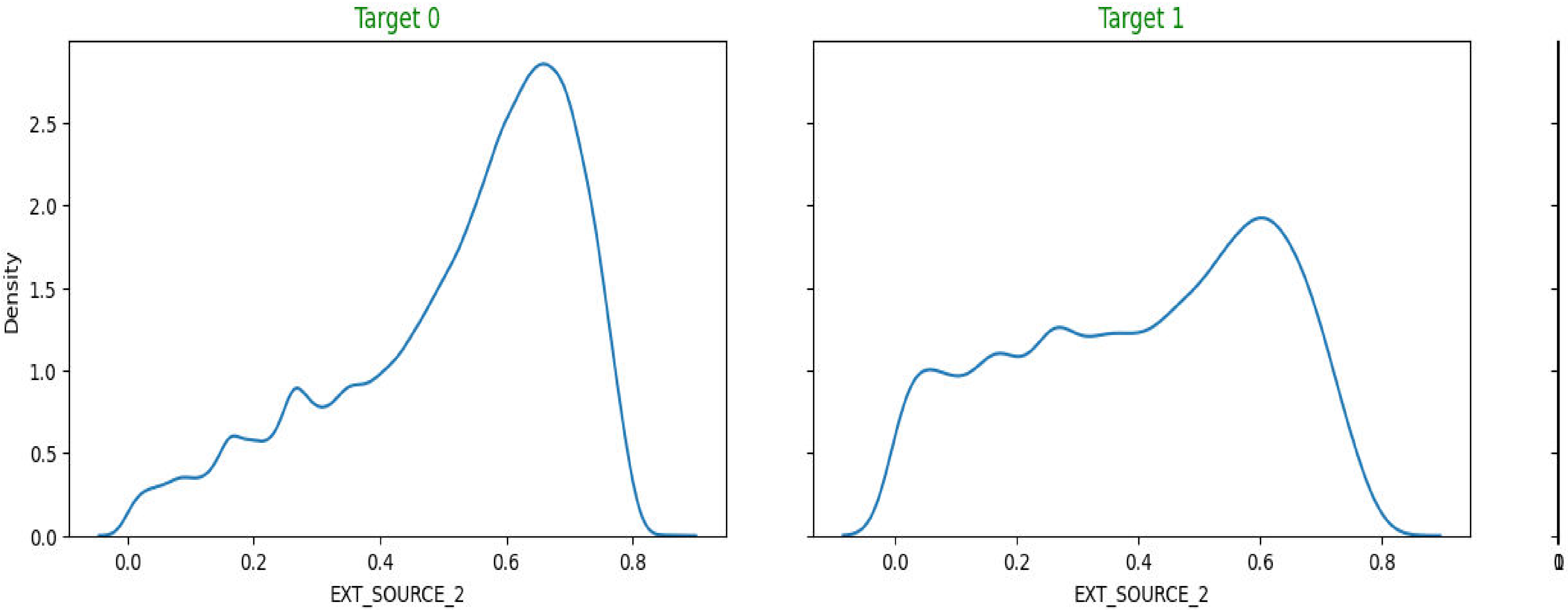
Seems to be lower for Target 1, indicating a smaller default loss to the business.



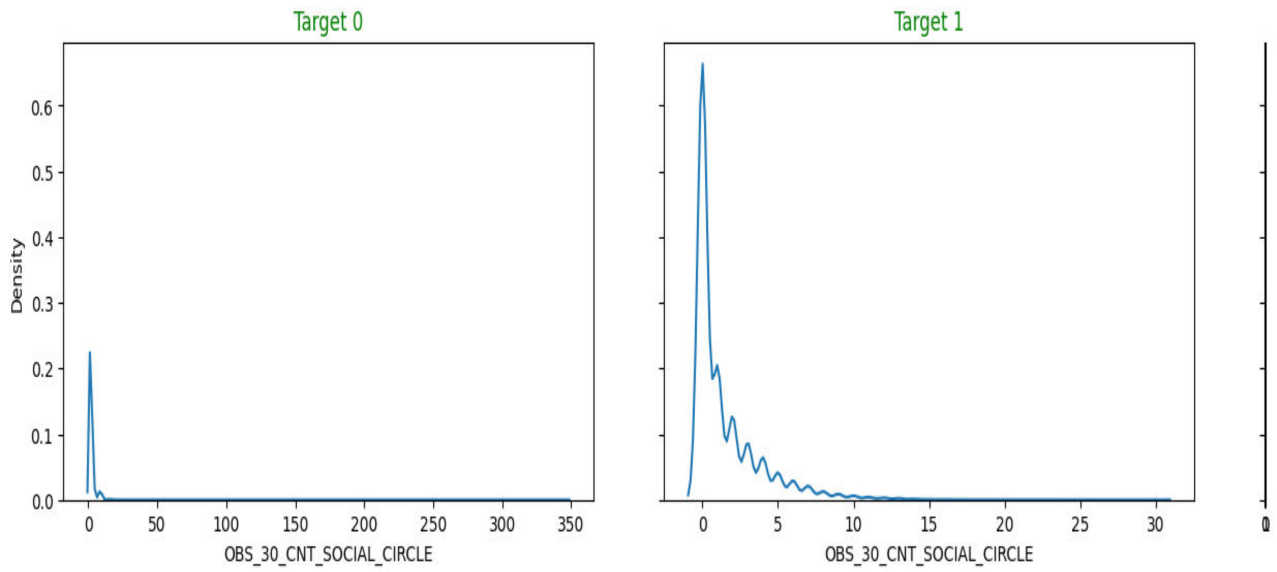
A higher AGE_IN_YEARS Density of 30 years in Target 1 indicates a higher default rate among younger people.

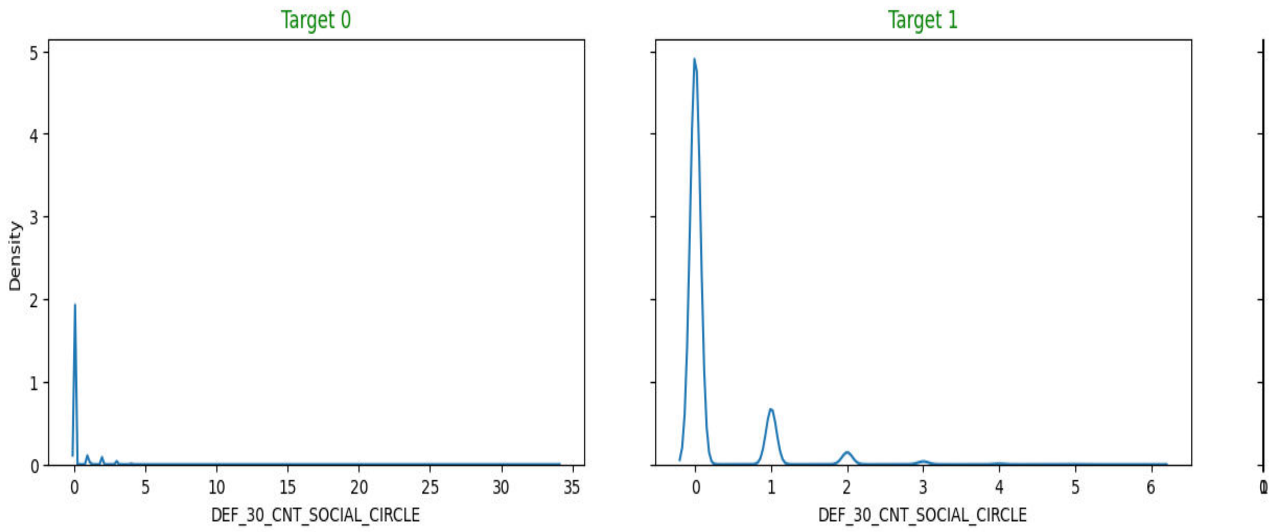


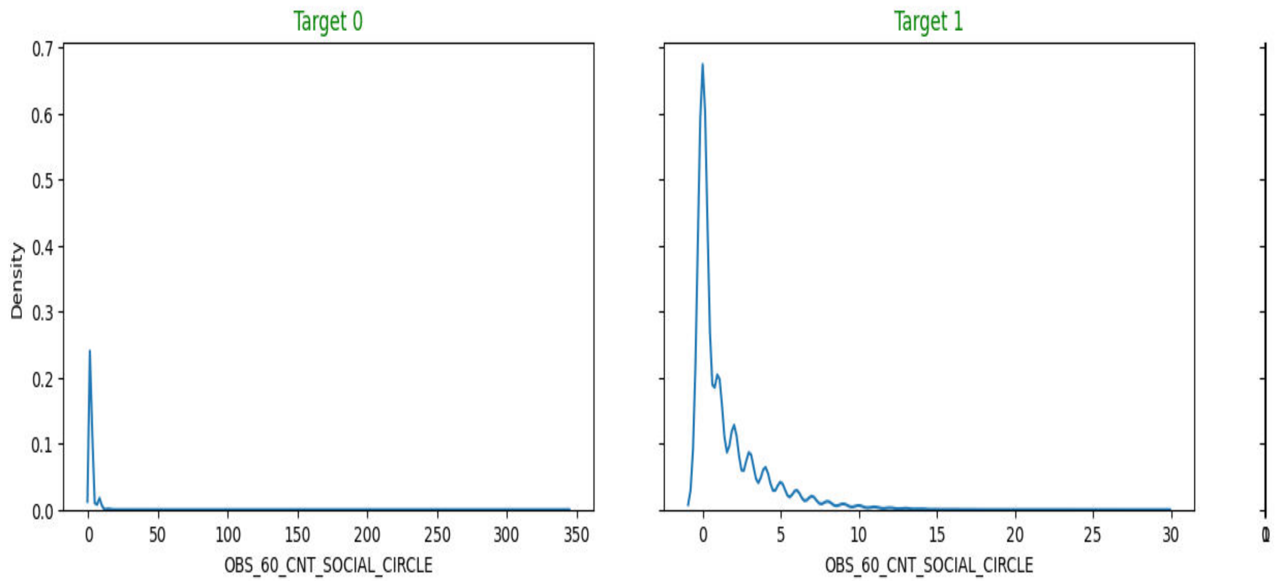
Fewer applicants  in Target 1 OWN vehicles

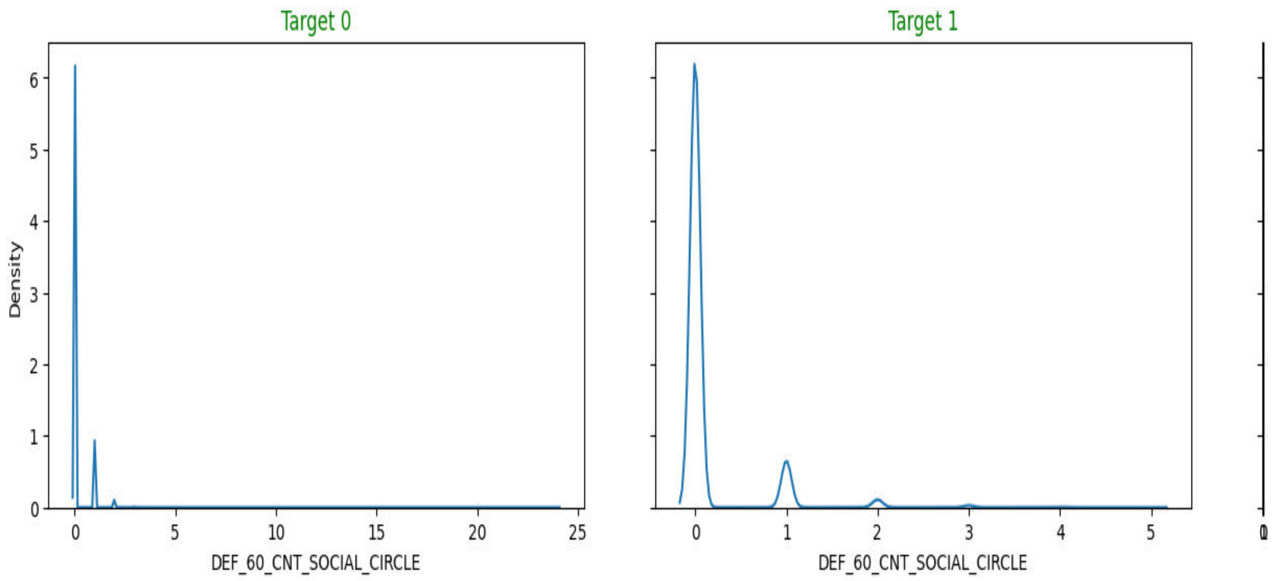


EXT_SOURCE_2 is clearing, indicating a larger density of better scores for TARGET 0.

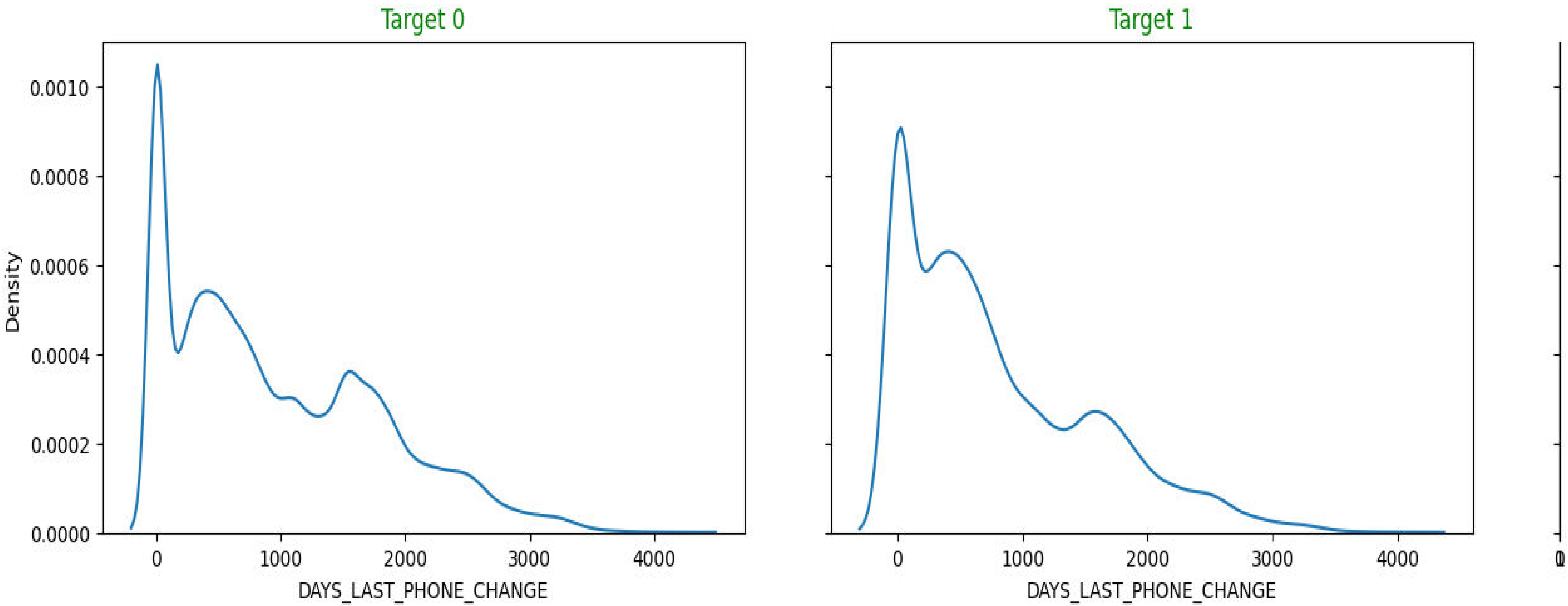




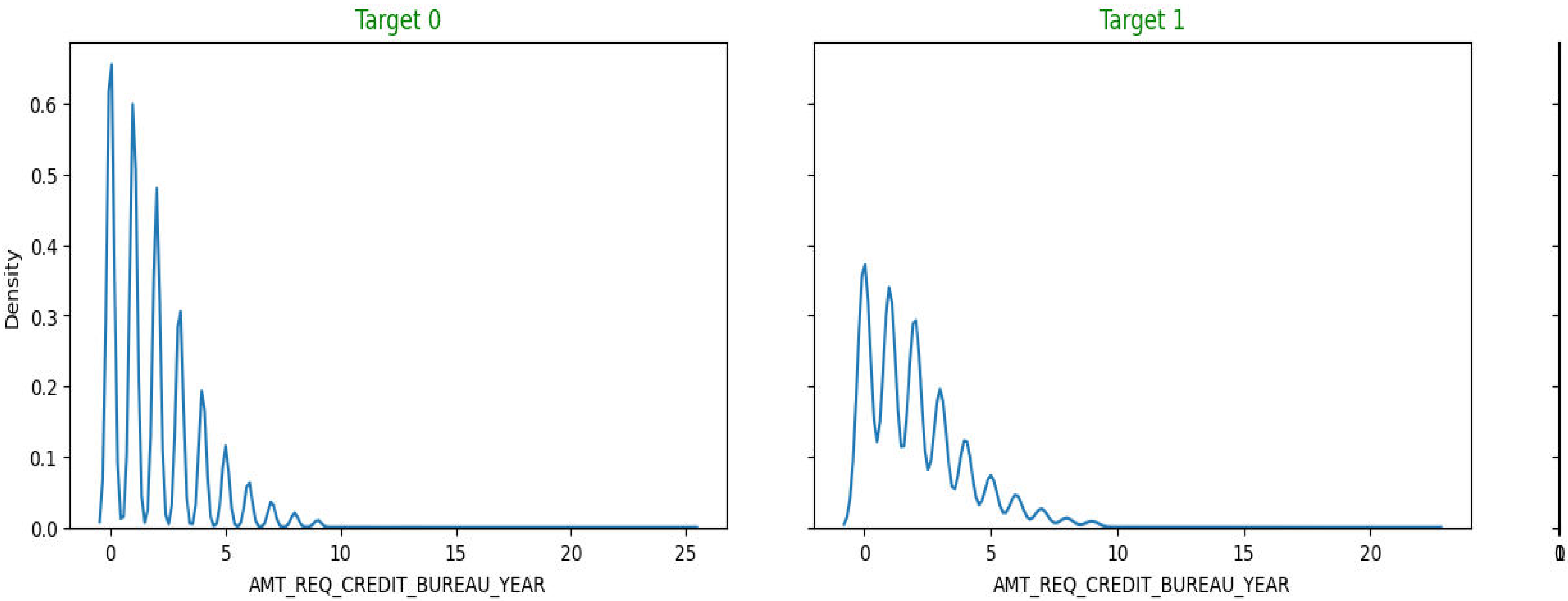




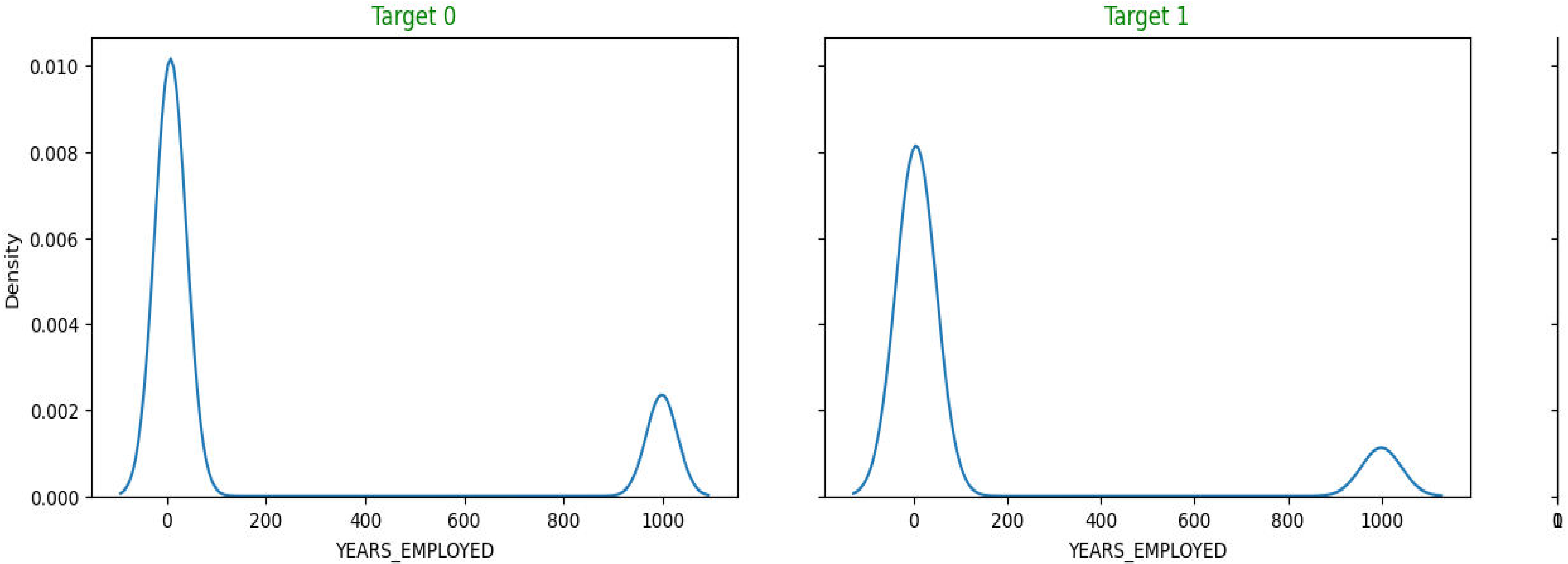
It is very evident that, for Target 1, the 30 DPD and 60 DPD observed in the social environment are greater (OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE).



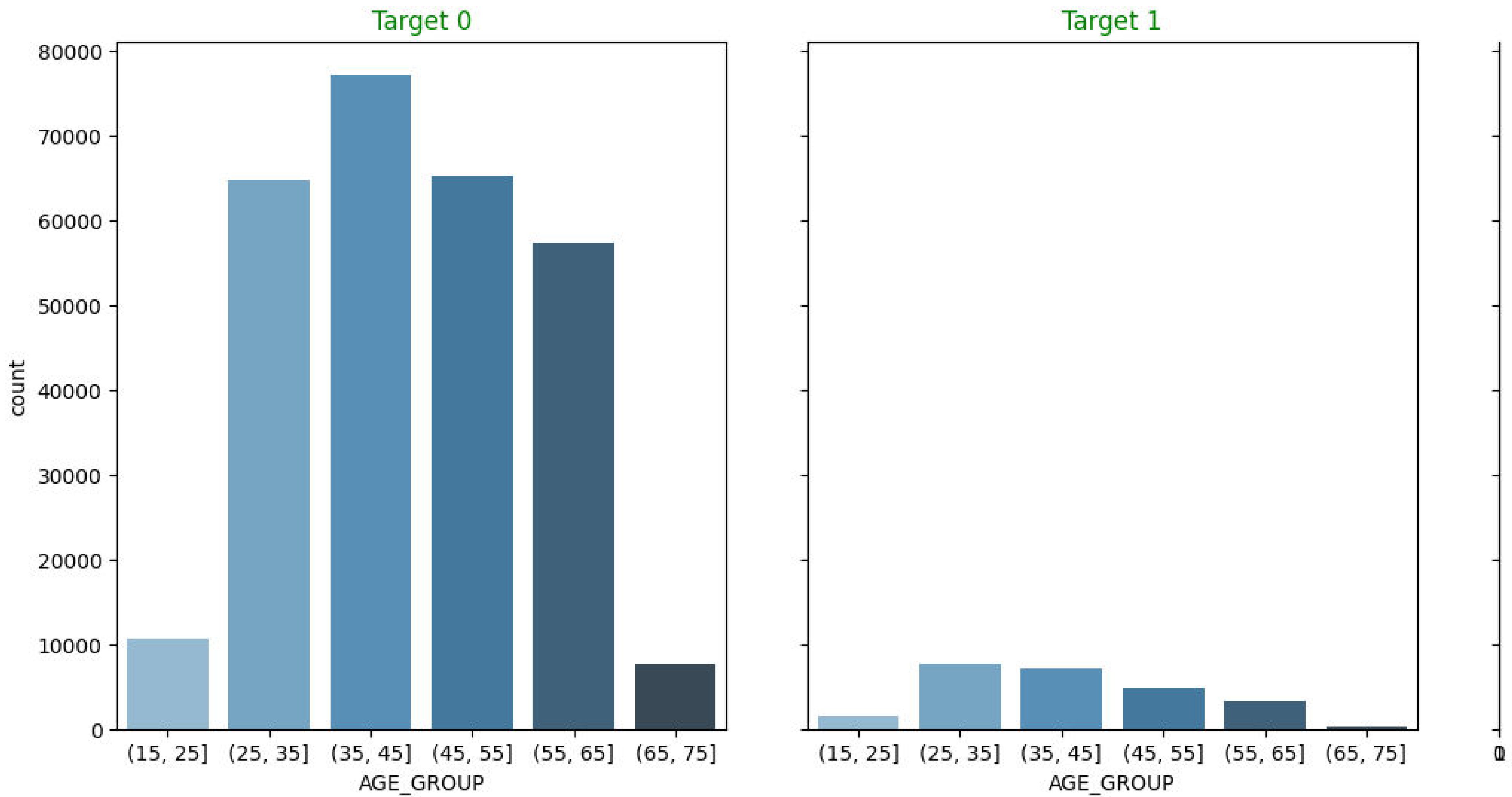
Compared to Target 1, more individuals from Target 1 have replaced their phone earlier. demonstrating problems with loan repayment intentions



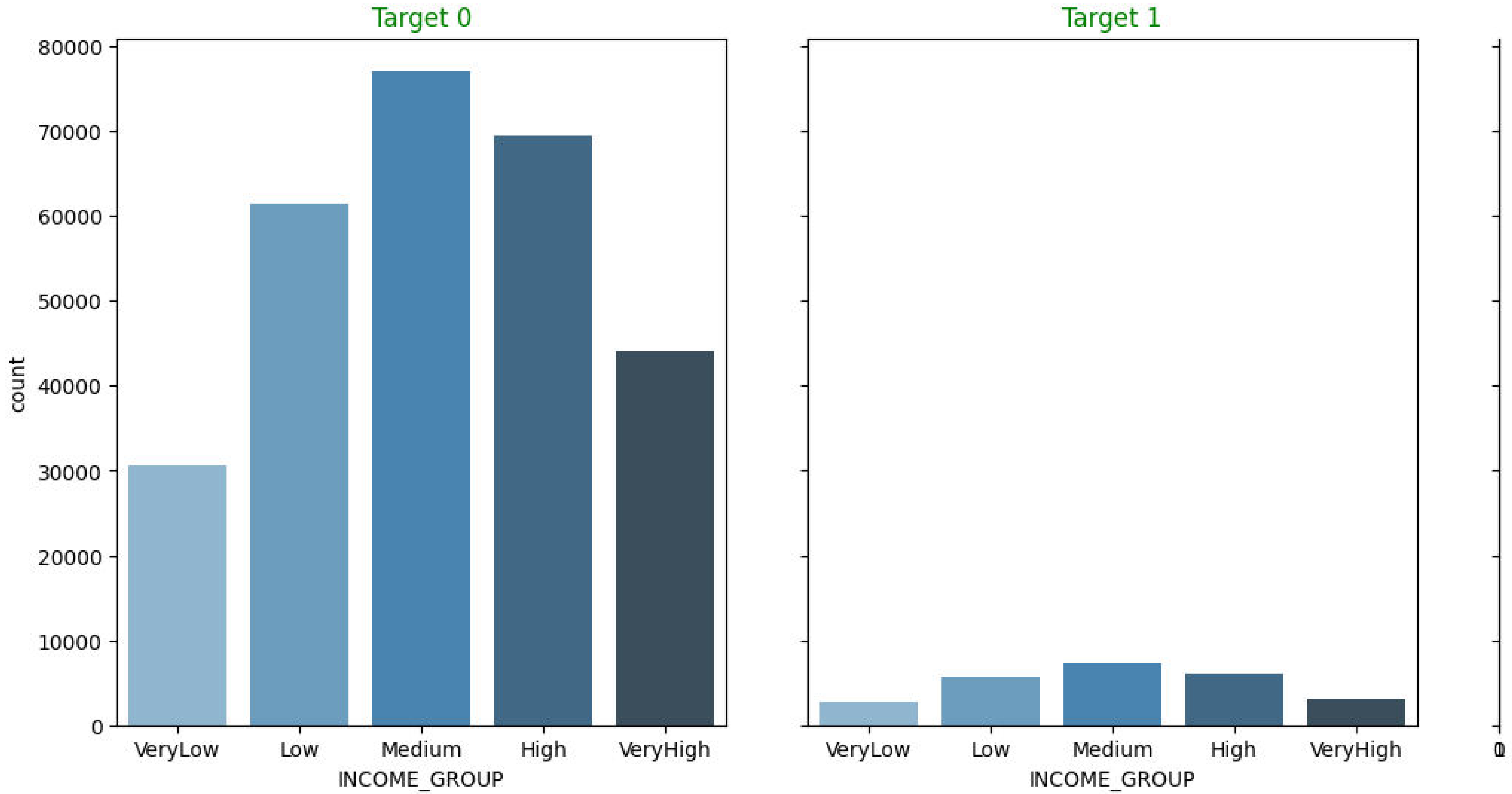
Target 1 is smaller than Target 0 in regards AMT_REQ_CREDIT_BUREAU_YEAR, month, week, and hour. This can mean they're looking to borrow money from several lenders.



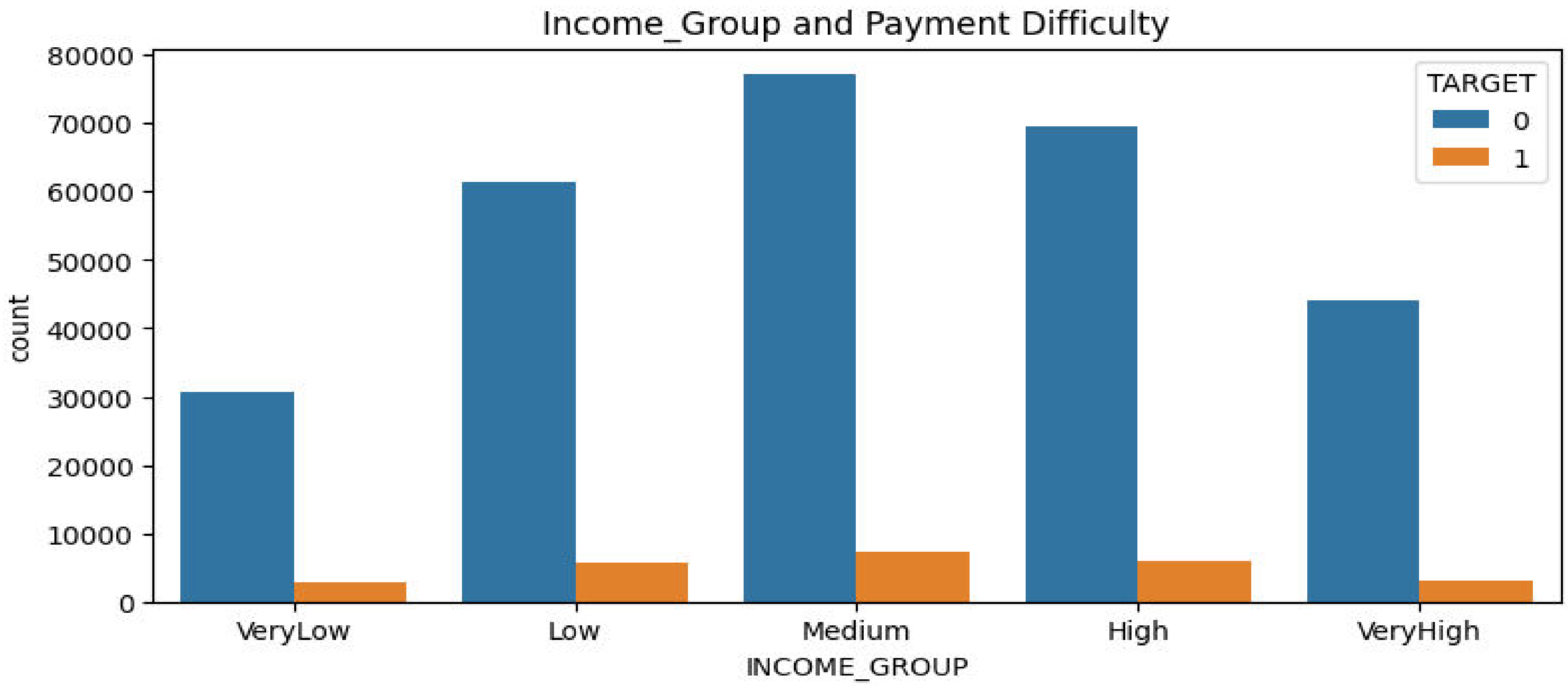
YEARS_EMPLOYED contains a significant number of erroneous entries, which results in an inaccurate data representation.



AGE_GROPUP -35-45 are more in Target 0. In Target 1- 25-25 have higher share. Age does seem like influencing default.



INCOME_GROUP - Medium income group have more count in Target 0 and Target 1

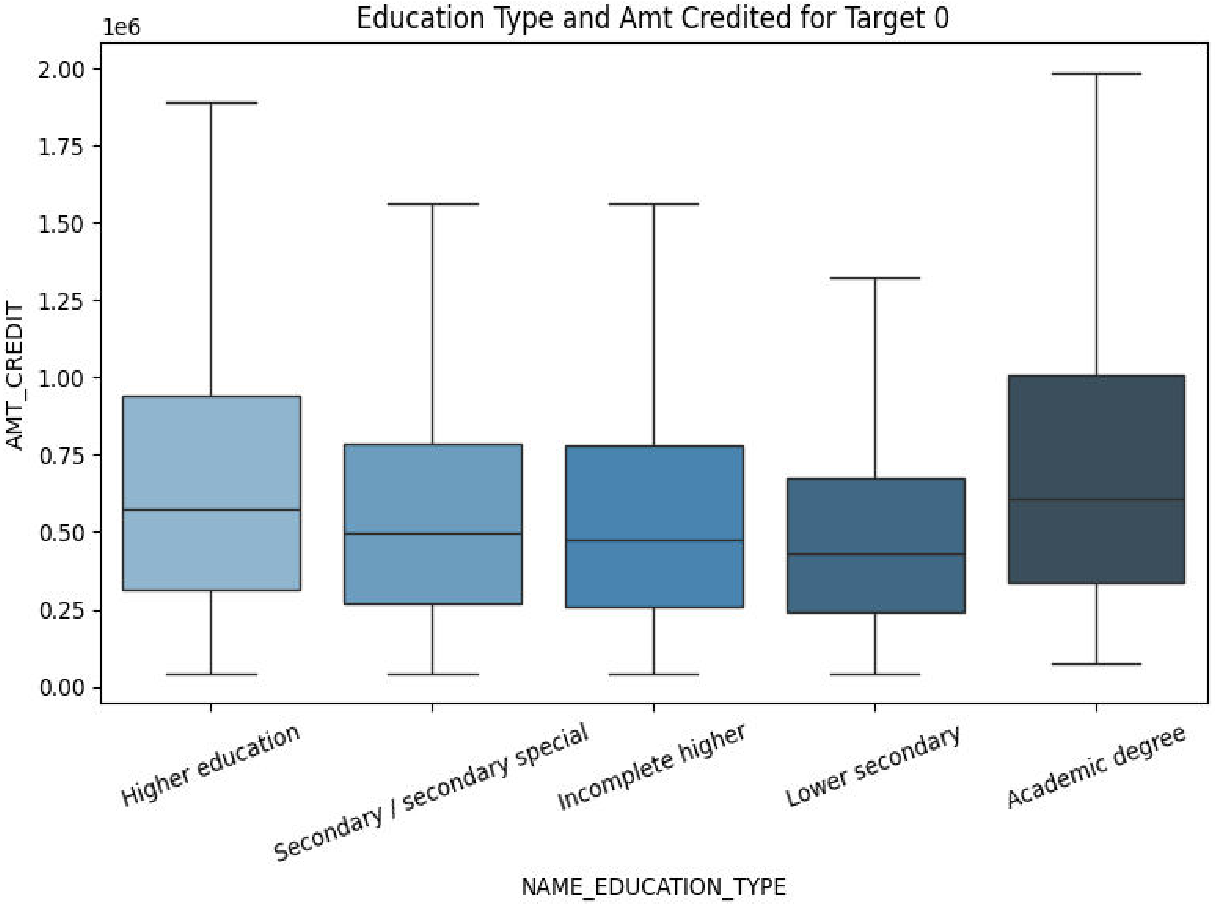


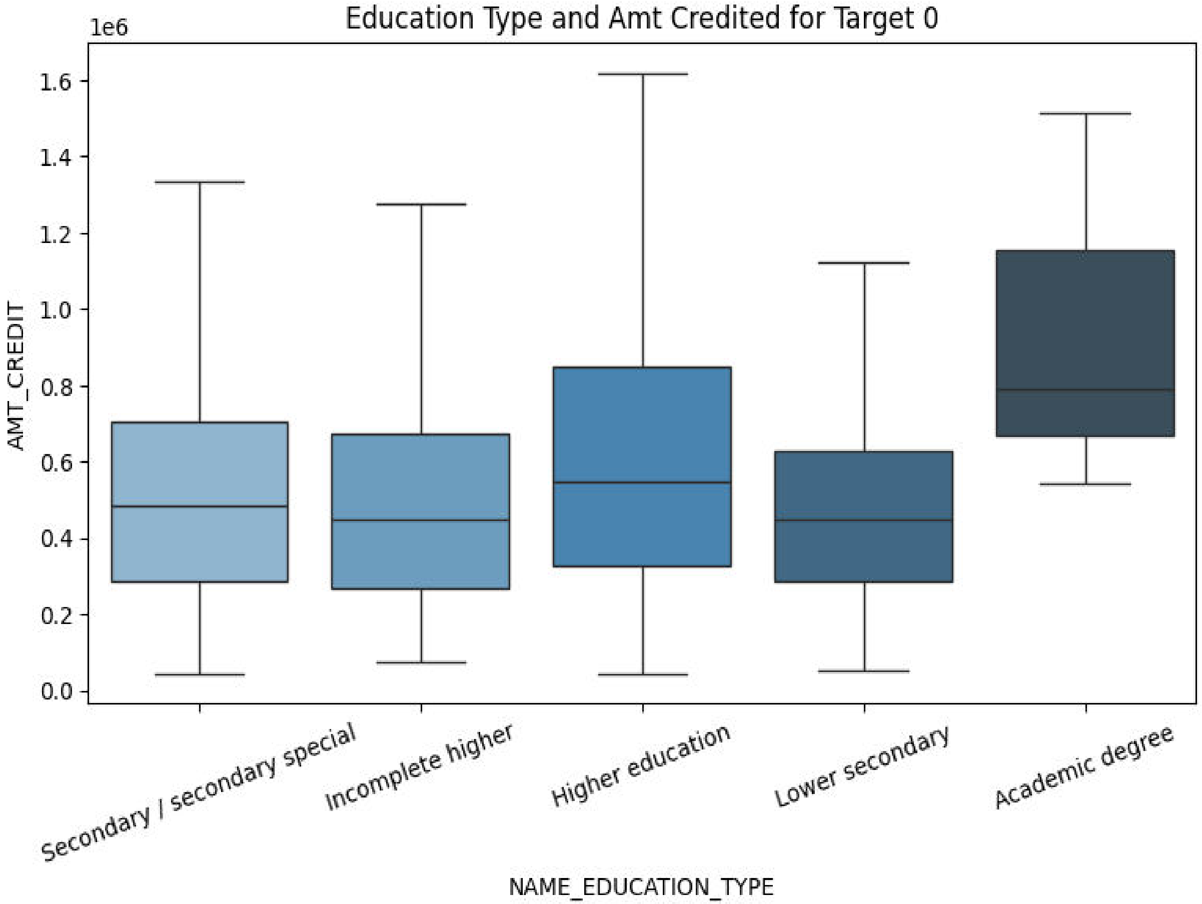
Even so, it may be deduced that the group with medium income receives the greatest number of loans. Due to increased AMT_CREDIT, the default value per loan is largest in the high income group.

Higher amounts that are not repaid may have an impact on the financial institution's loan book.

When approving loans to higher income groups, the organization needs to come up with new guidelines and procedures.

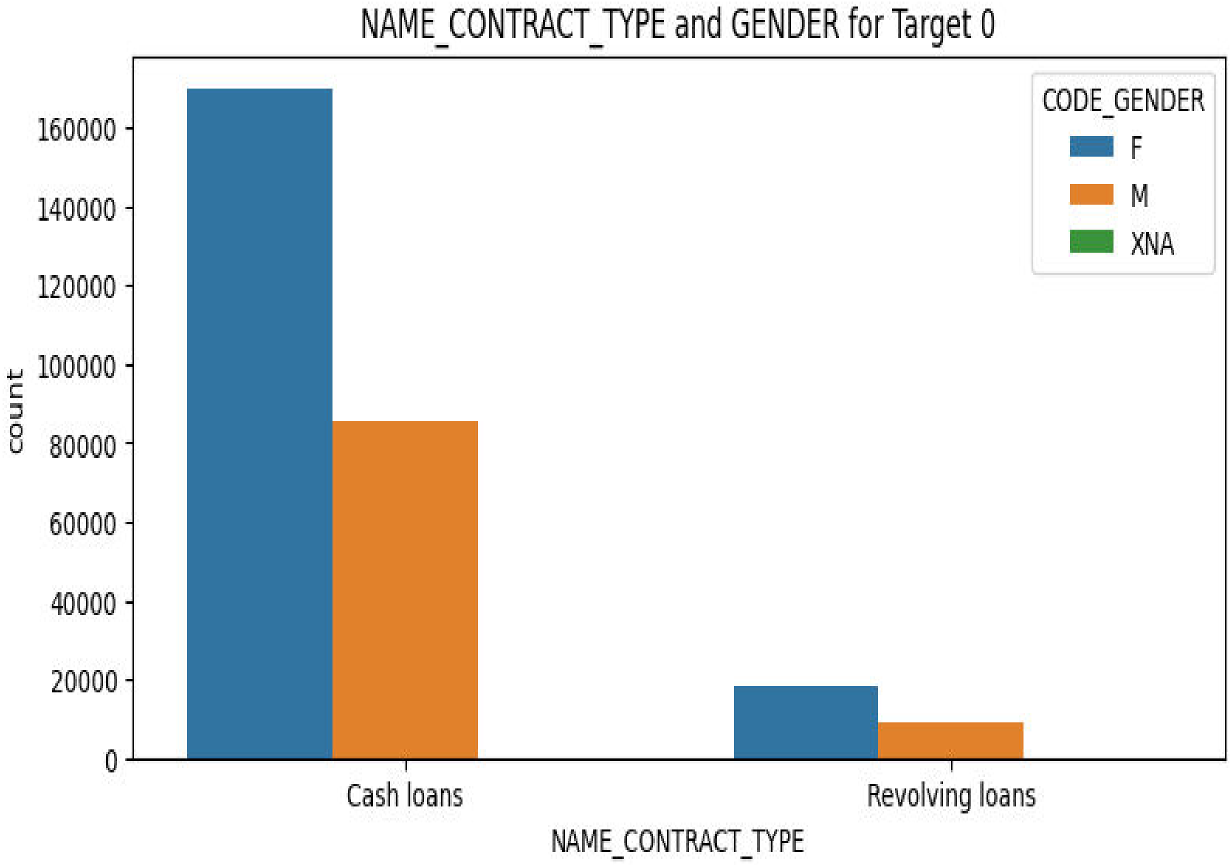
# Bivariate Analysis on Categorical and Continuous Variable

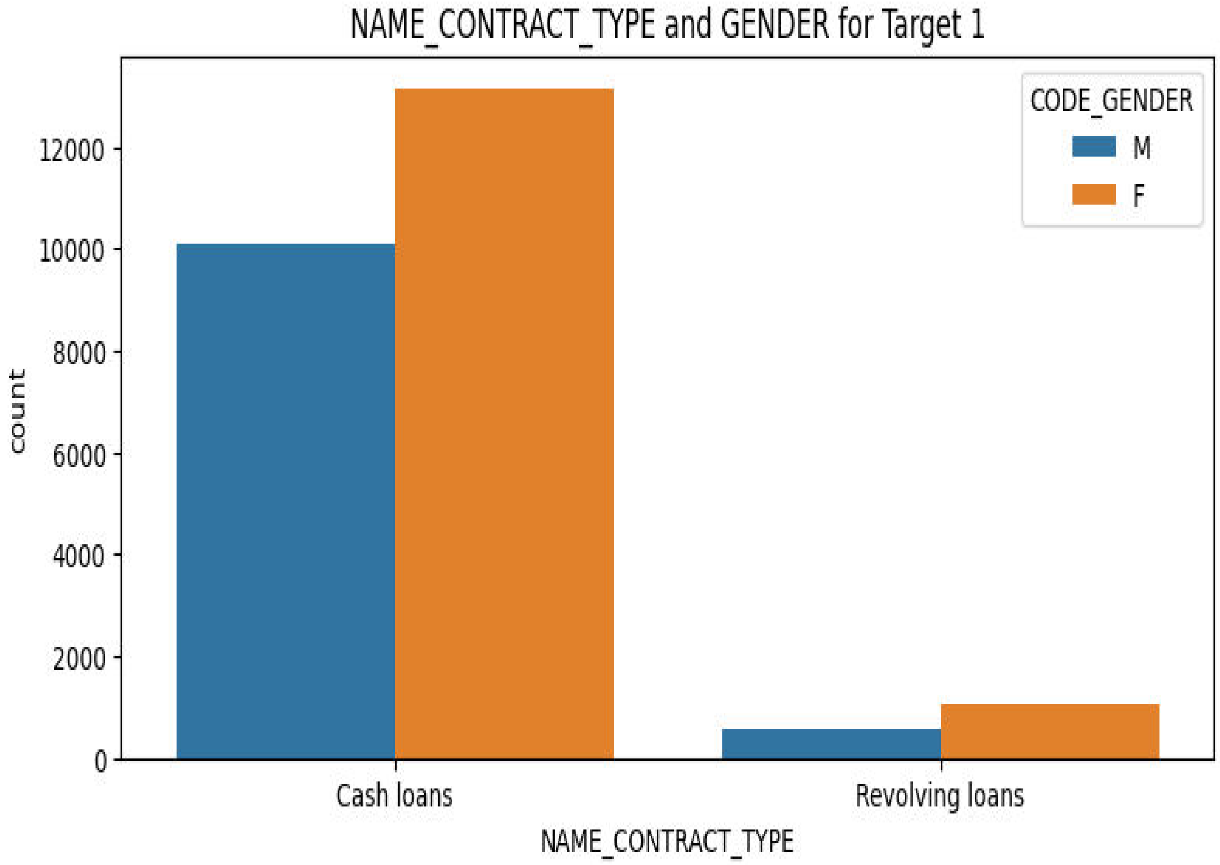




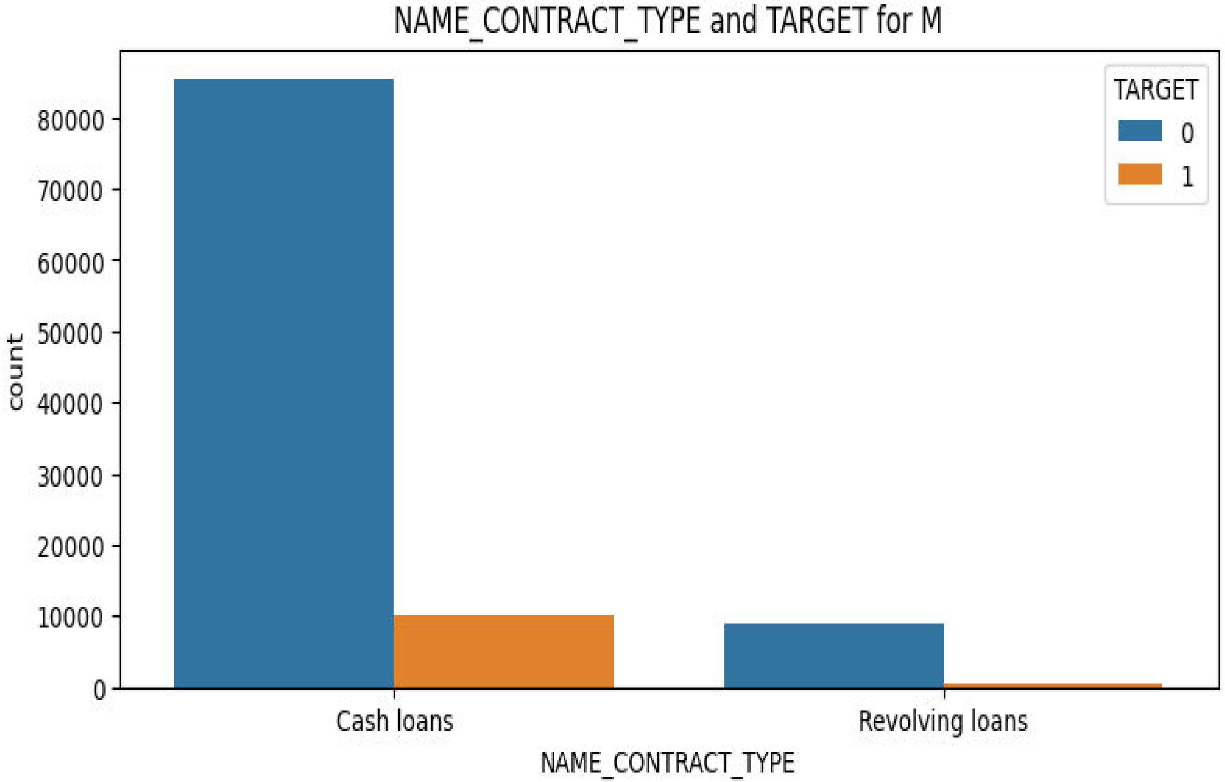
For applicants with academic degrees, the median loan default value is greater. However as the curve above illustrates, the number of candidates with an academic degree is minuscule.

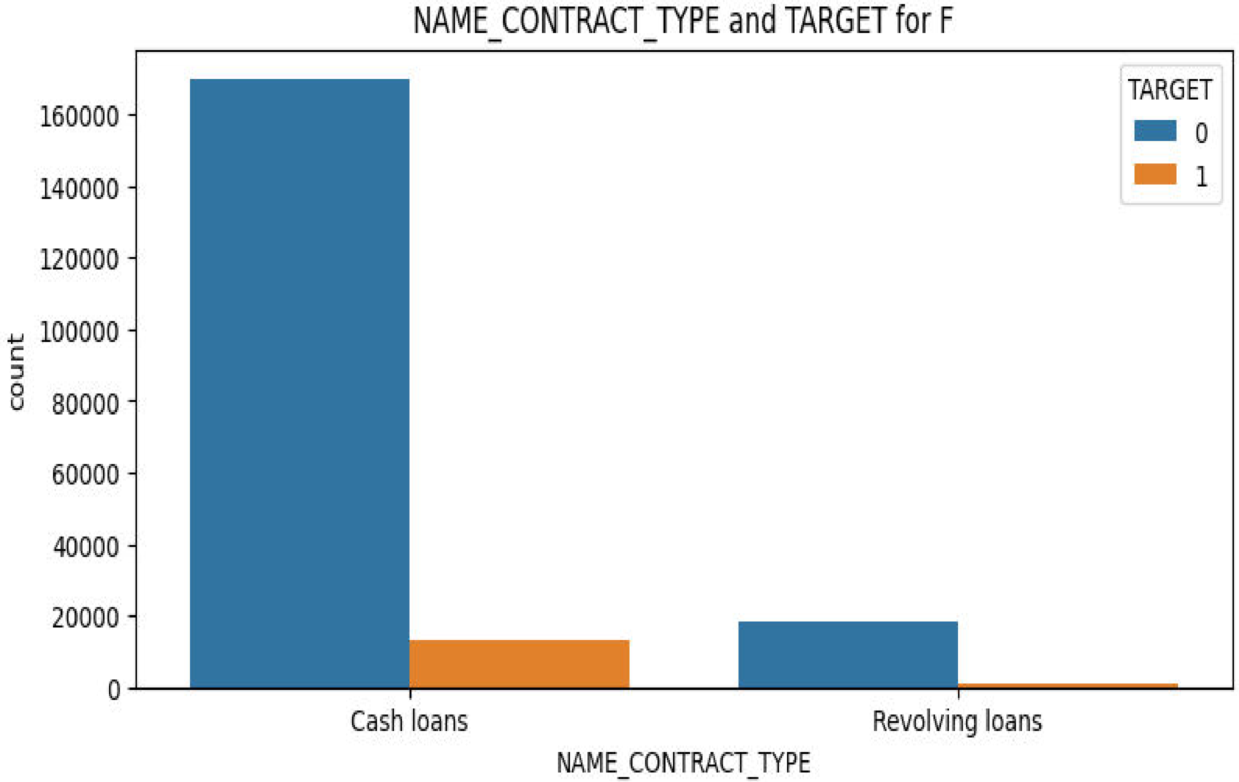
Nothing can be inferred from this analysis..



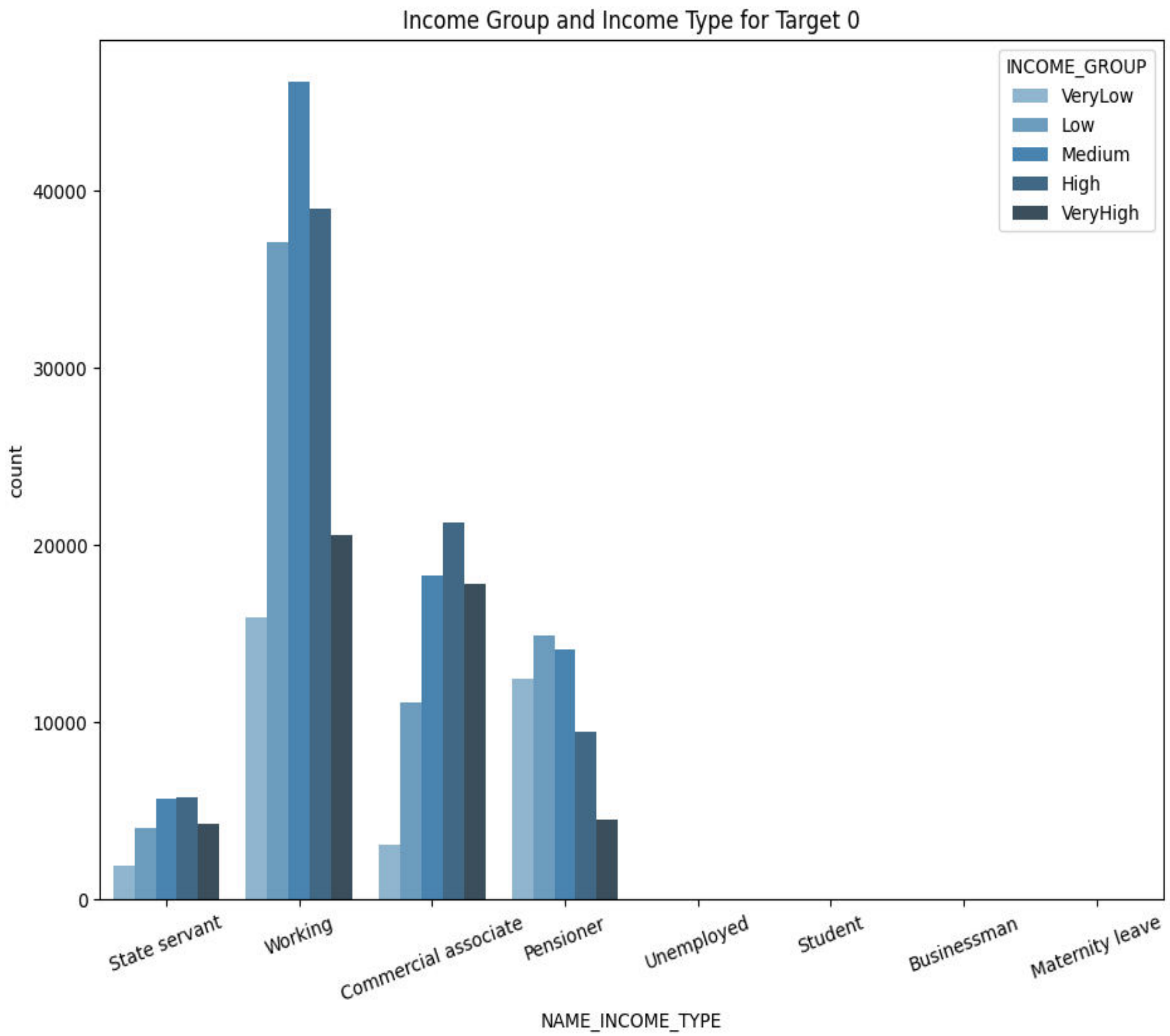


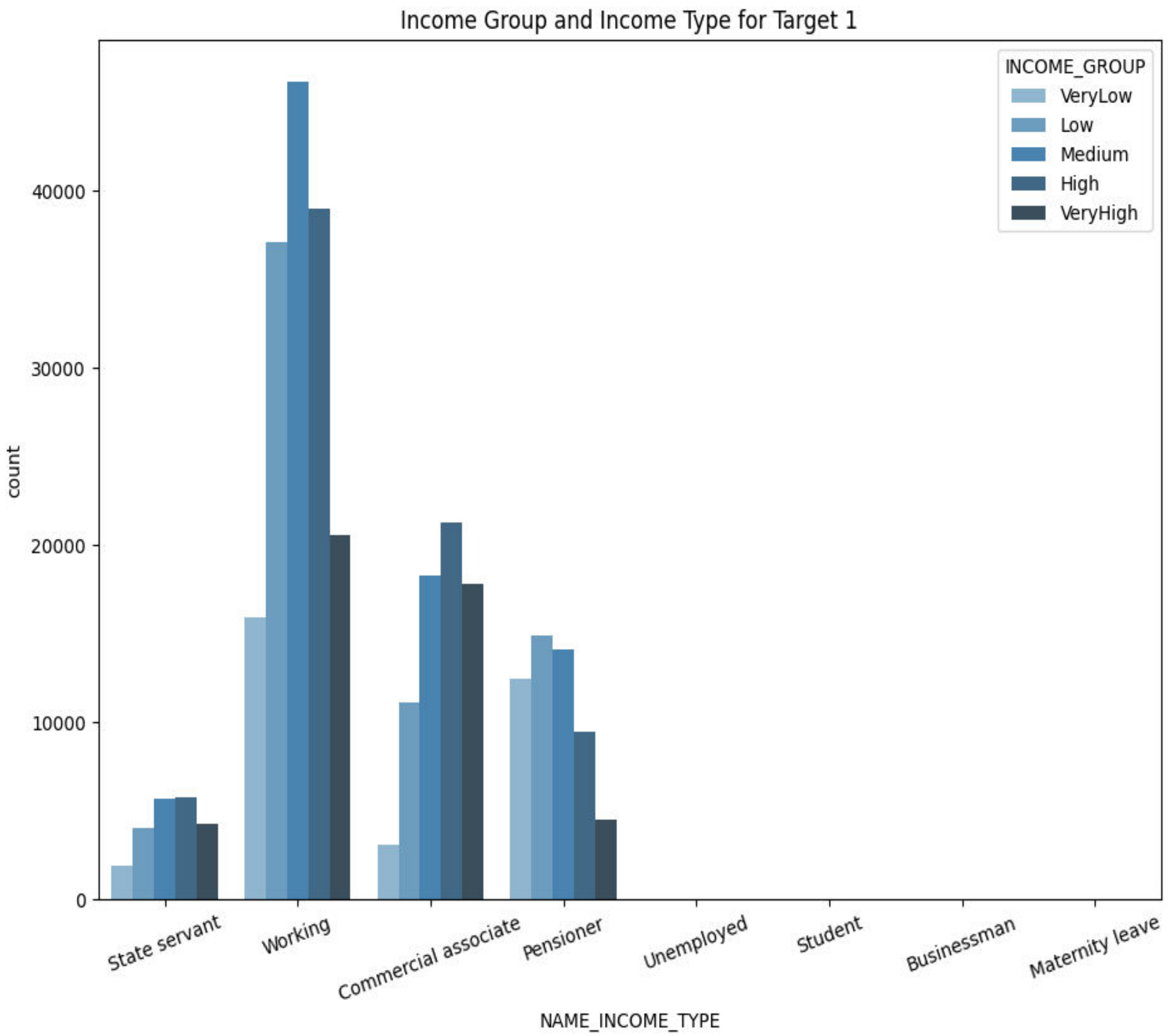
The statistics above indicates that there are more female loan applicants. As can be seen in the plot above, the percentage of male candidates who are denied is larger even if their number is smaller.



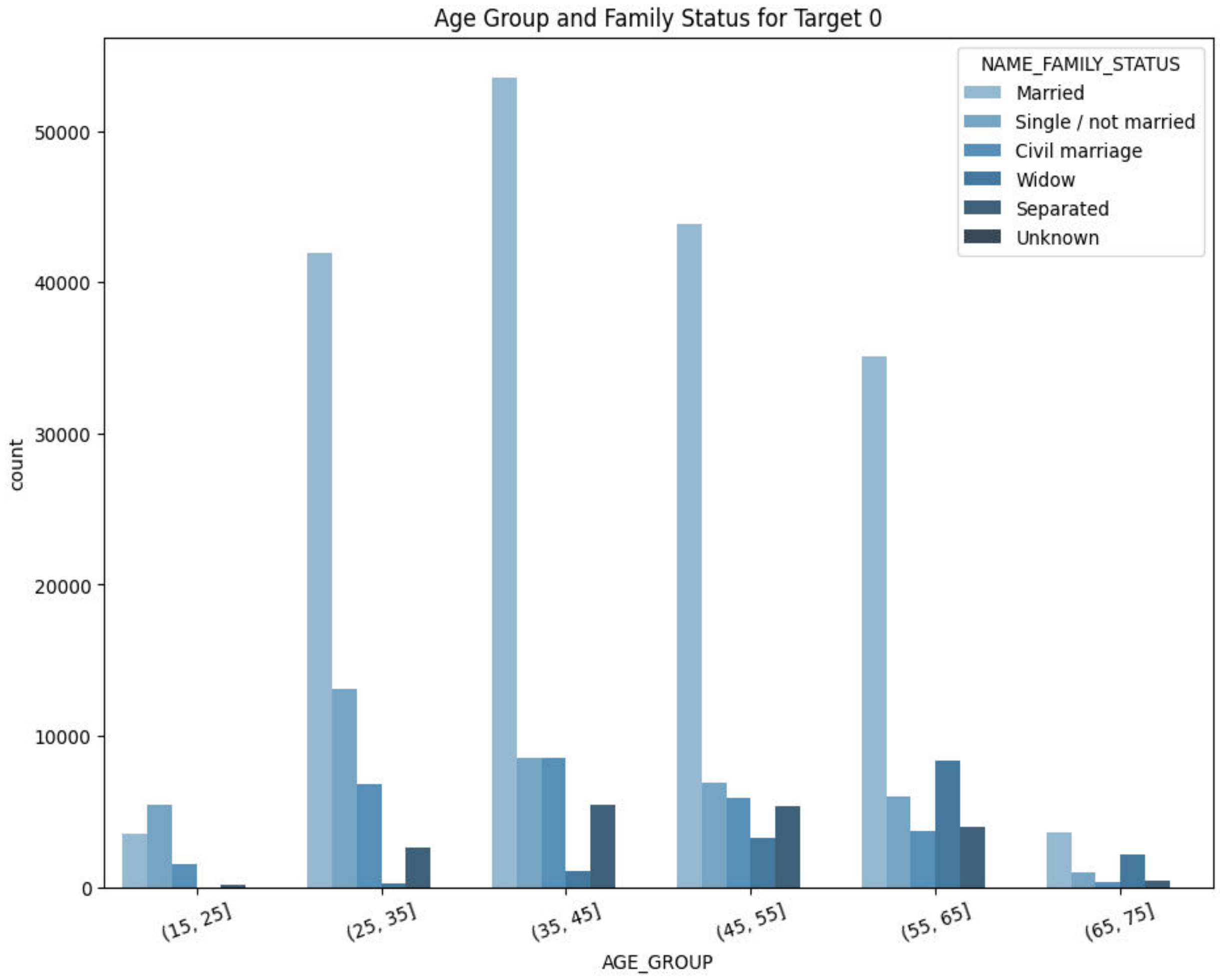


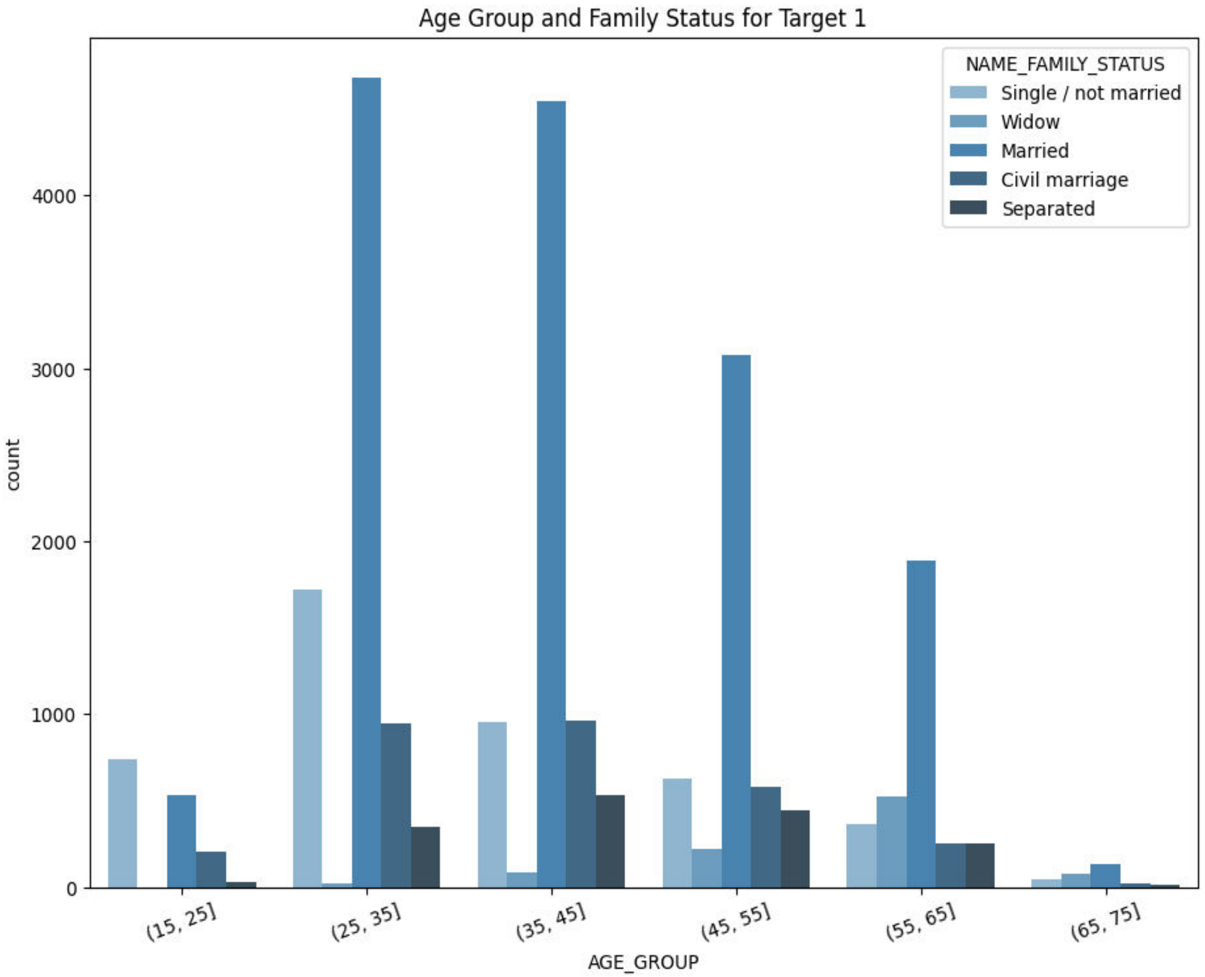
Male applicants are defaulting more that female applicants





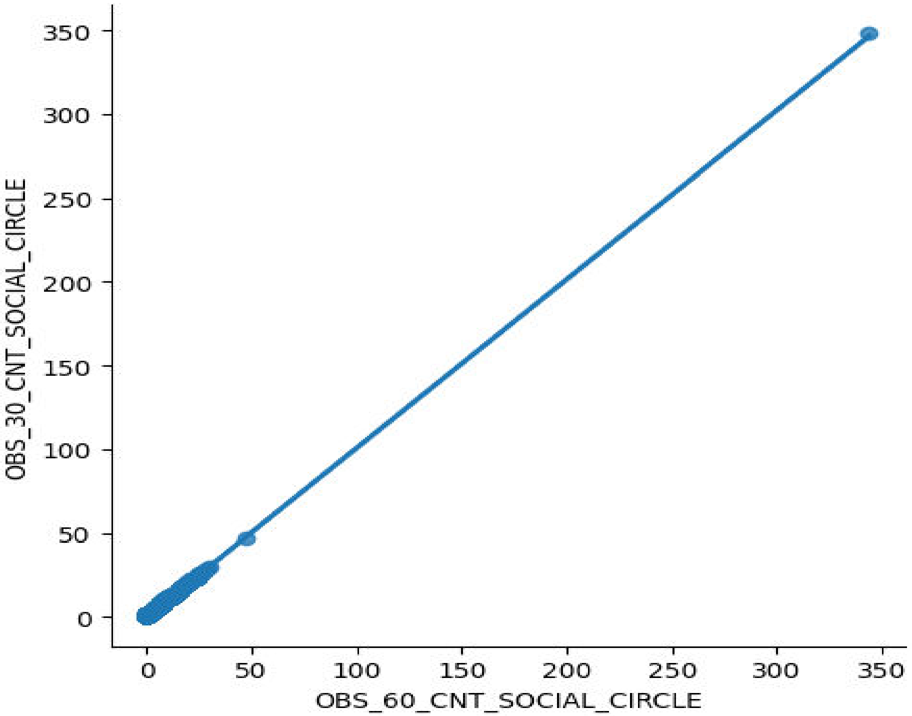
Approximately 1 in 12 defaults belong to the income type and medium income group. Above the average of 1 in 11 defaults

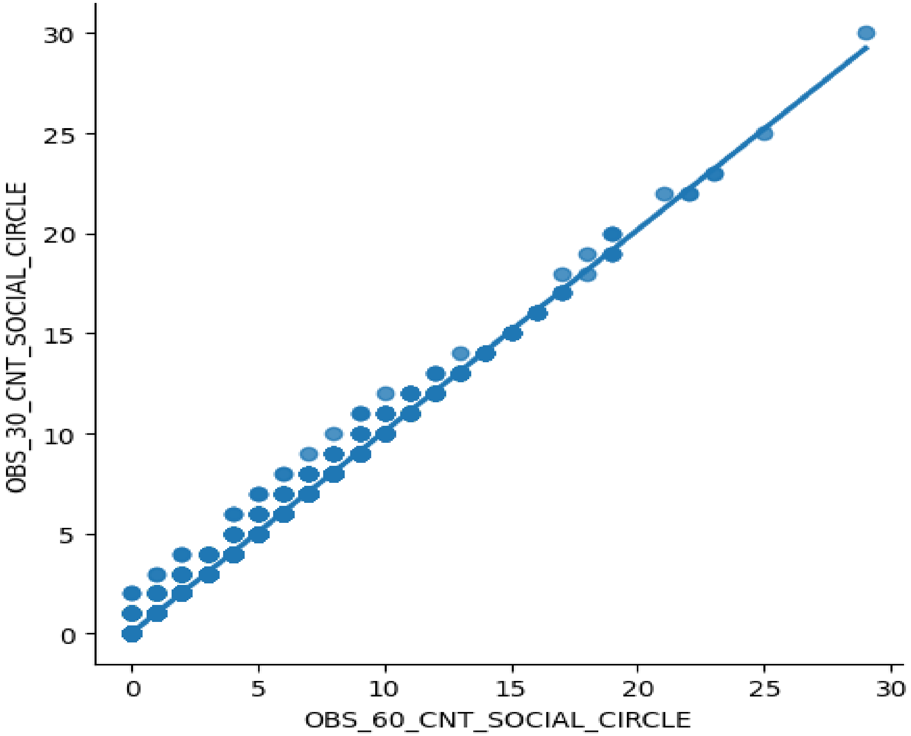




The majority of applicants experiencing payment difficulties are married applicants in the 25–35 and 35–45 age groups.

# Correlations - Analysis on Continuous Variables

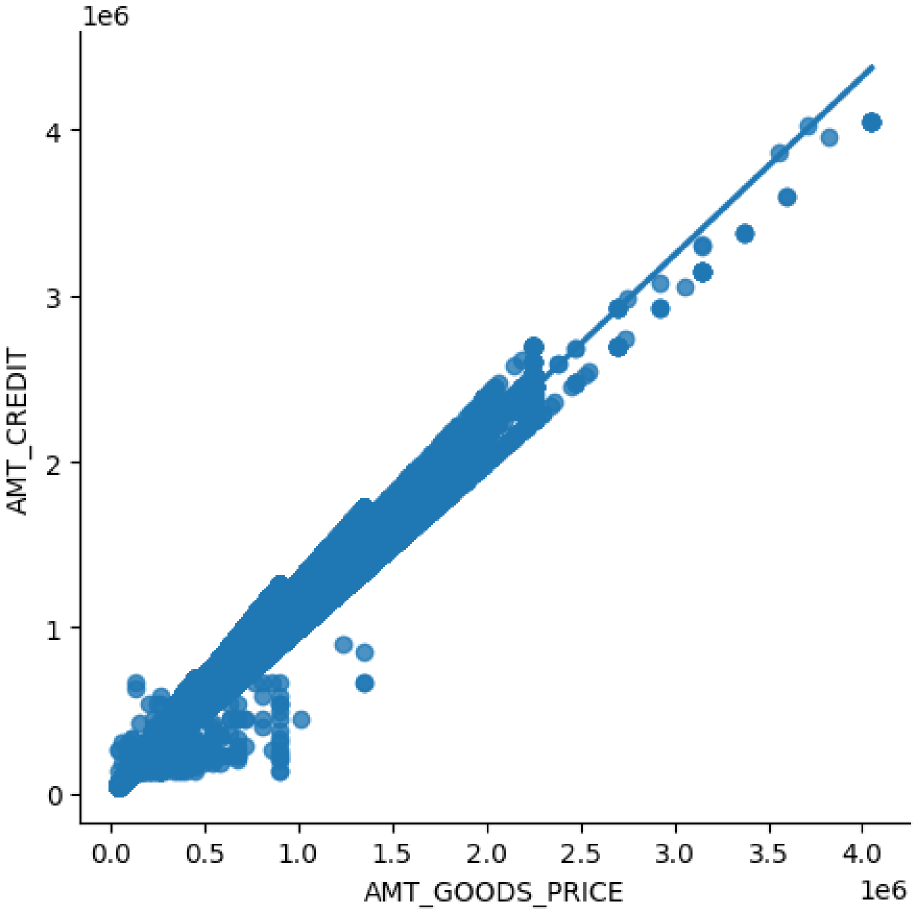


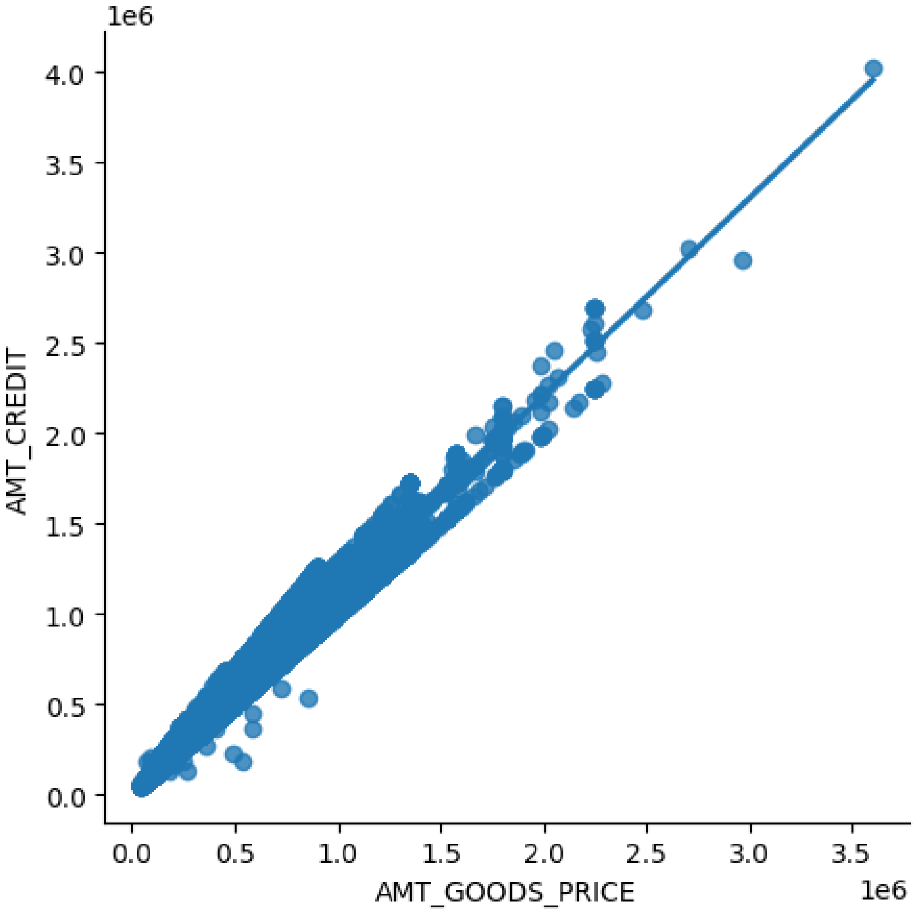


The client's social environment is indicated by the OBS_30_CNT_SOCIAL_CIRCLE' and OBS_60_CNT_SOCIAL_CIRCLE', which have observable 30/60 DPD.

Undoubtedly, these are connected. Additionally, we can observe that it is steeper and higher for Target 1, indicating that this parameter needs to be carefully examined during the clearance process.

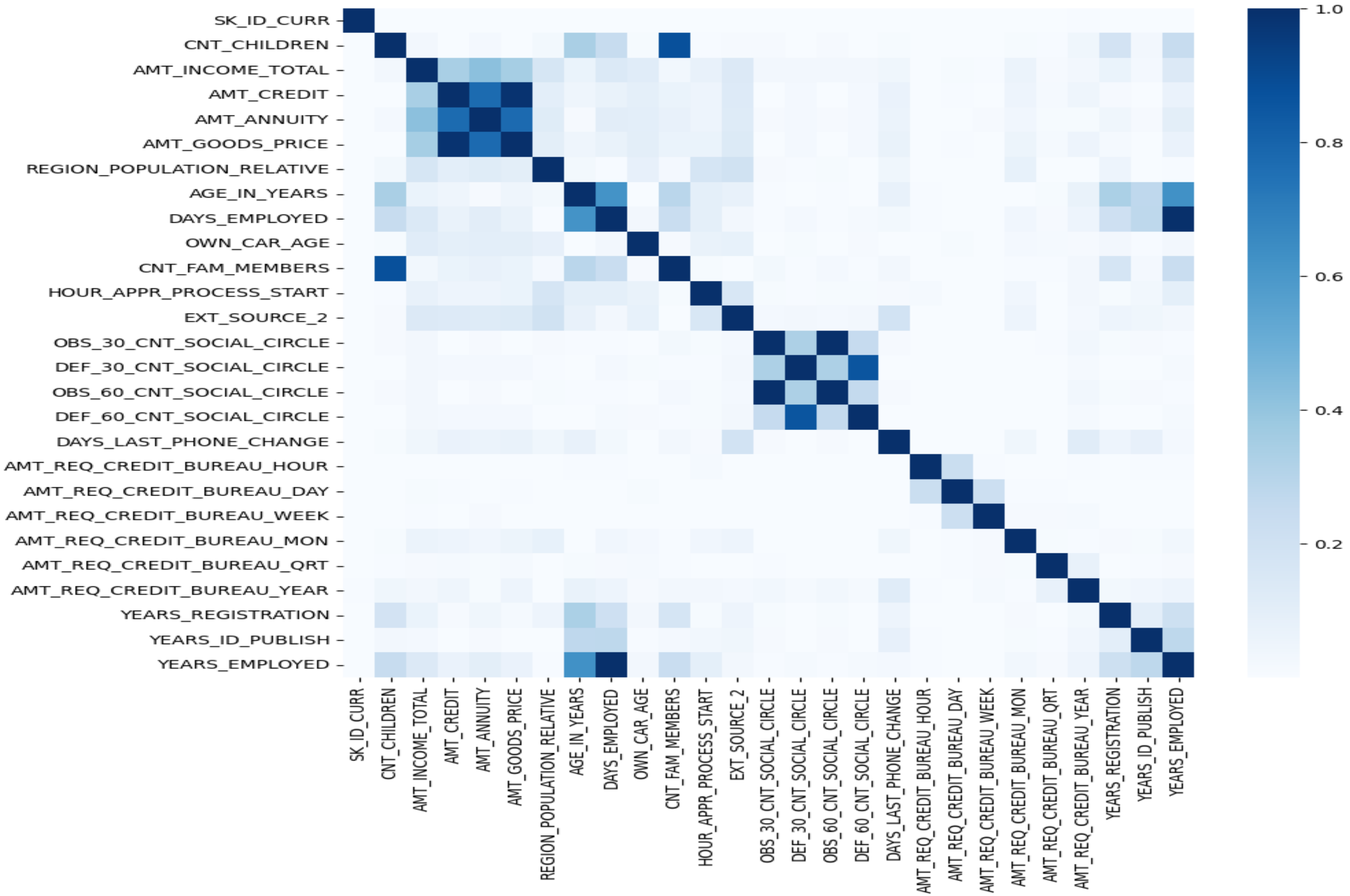
CNT_SOCIAL_CIRCLE_DEF_30 - The trend is rising. However, Target 1's graph is not dense because it includes less data

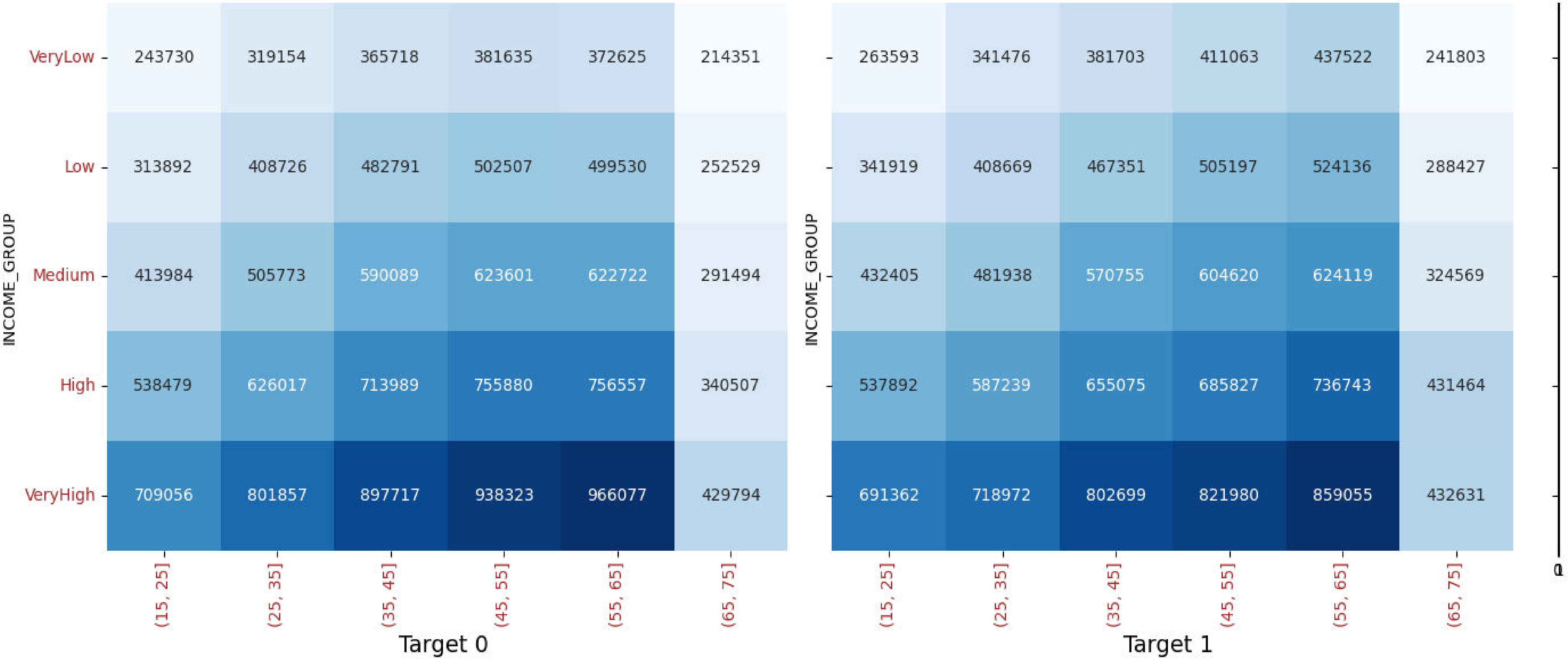




For Target 1, it appears that AMT_CREDIt and AMT_GOOD PRICE are not rising in line with AMT_INCOME, which could result in default.

# Multivariate Analysis





Target 0

| INCOME_GROUP | (15, 25] | (25, 35] | (35, 45] | (45, 55] | (55, 65] | (65, 75] |
|---|---|---|---|---|---|---|
| VeryLow | 243730 | 319154 | 365718 | 381635 | 372625 | 214351 |
| Low | 313892 | 408726 | 482791 | 502507 | 499530 | 252529 |
| Medium | 413984 | 505773 | 590089 | 623601 | 622722 | 291494 |
| High | 538479 | 626017 | 713989 | 755880 | 756557 | 340507 |
| VeryHigh | 709056 | 801857 | 897717 | 938323 | 966077 | 429794 |

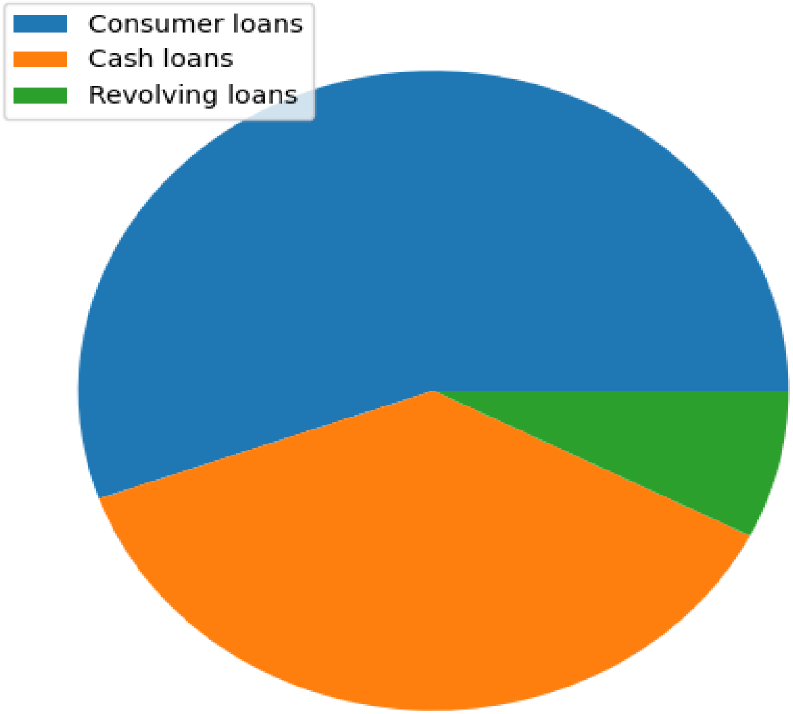
Target 1

| INCOME_GROUP | (15, 25] | (25, 35] | (35, 45] | (45, 55] | (55, 65] | (65, 75] |
|---|---|---|---|---|---|---|
| VeryLow | 263593 | 341476 | 381703 | 411063 | 437522 | 241803 |
| Low | 341919 | 408669 | 467351 | 505197 | 524136 | 288427 |
| Medium | 432405 | 481938 | 570755 | 604620 | 624119 | 324569 |
| High | 537892 | 587239 | 655075 | 685827 | 736743 | 431464 |
| VeryHigh | 691362 | 718972 | 802699 | 821980 | 859055 | 432631 |

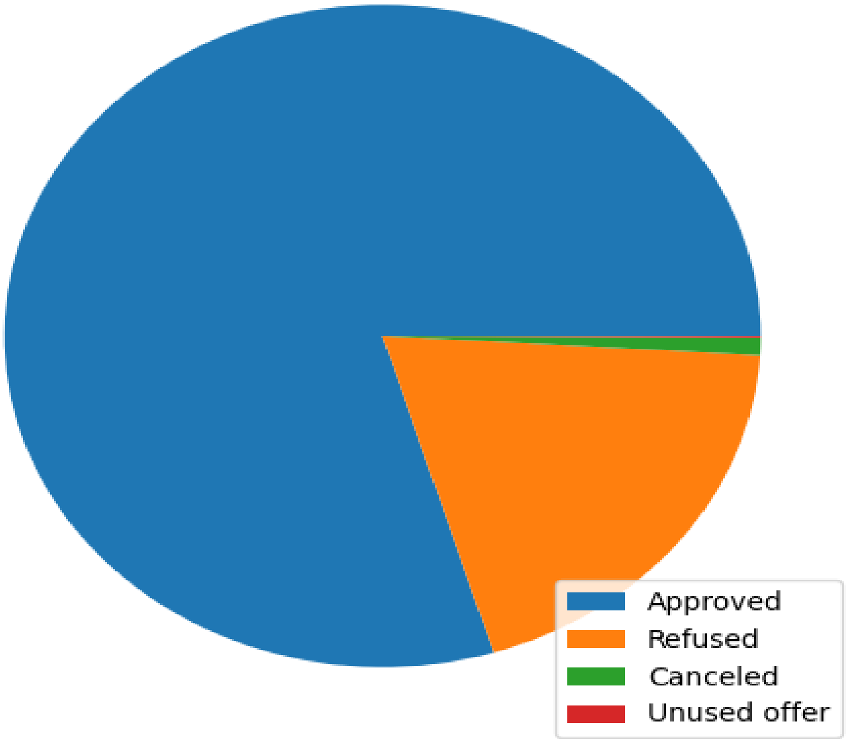
Age Group 55-65 in Very High income group has high amount credit. As explained above, this could result as loss in loan book

# Previous Application Data Analysis

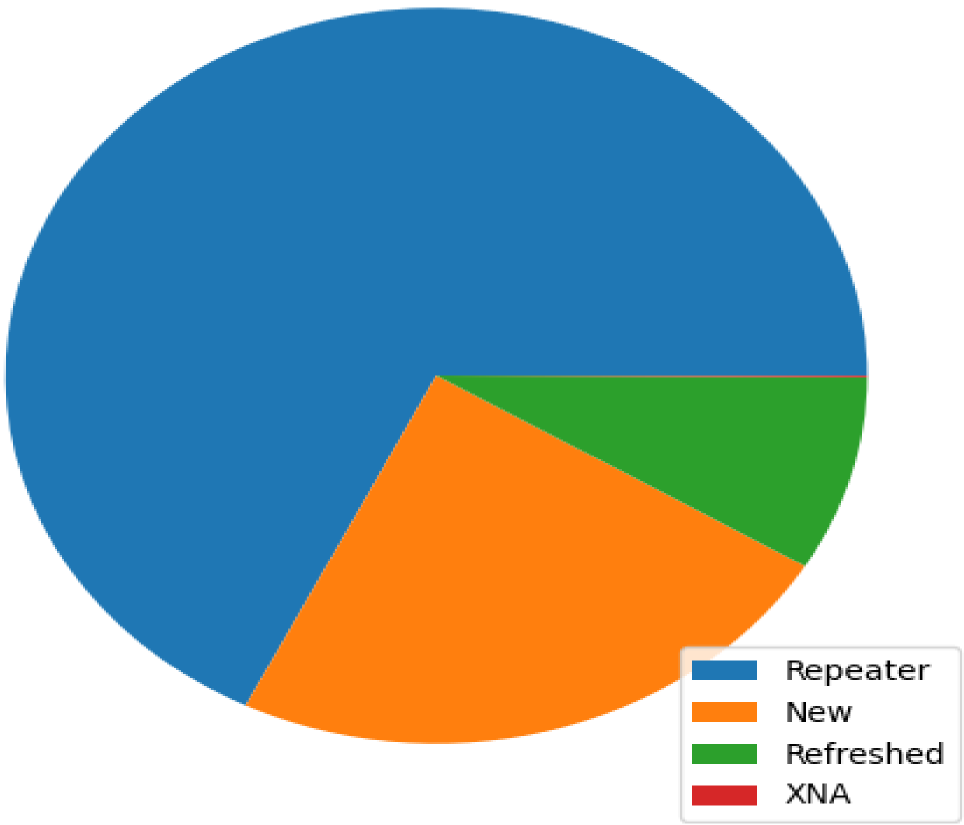
Contains information about the client’s previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

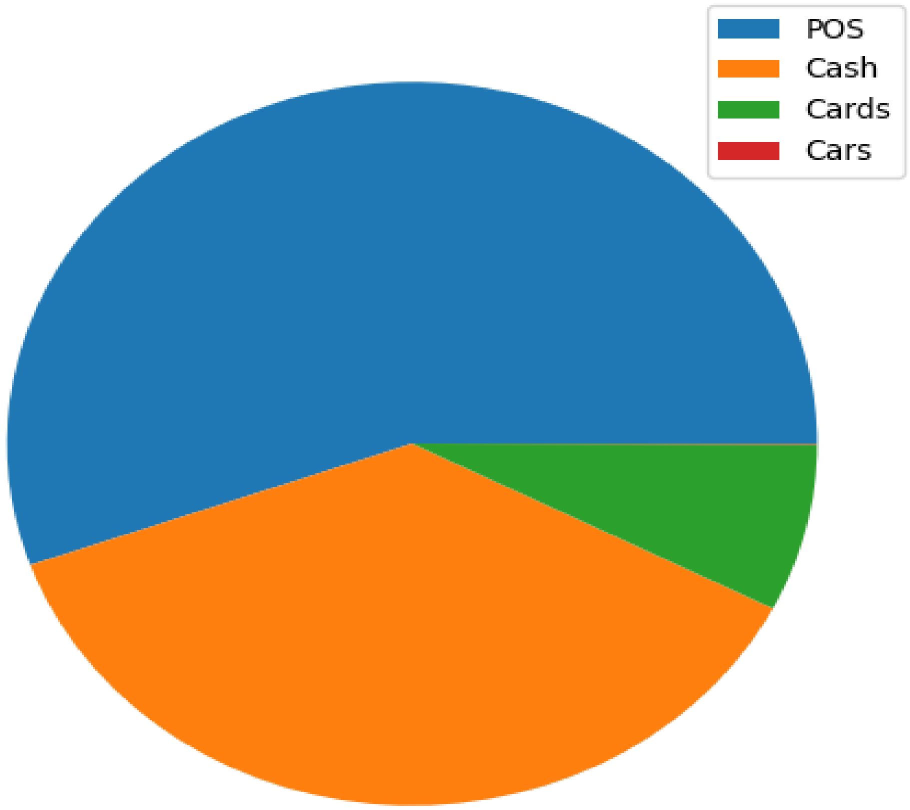


The Application data frame did not contain the Consumer Loan, a distinct form of loan found in this data frame. Consumer loans make up 55% of all loans. 37% of loans are cash, with the remainder being revolving



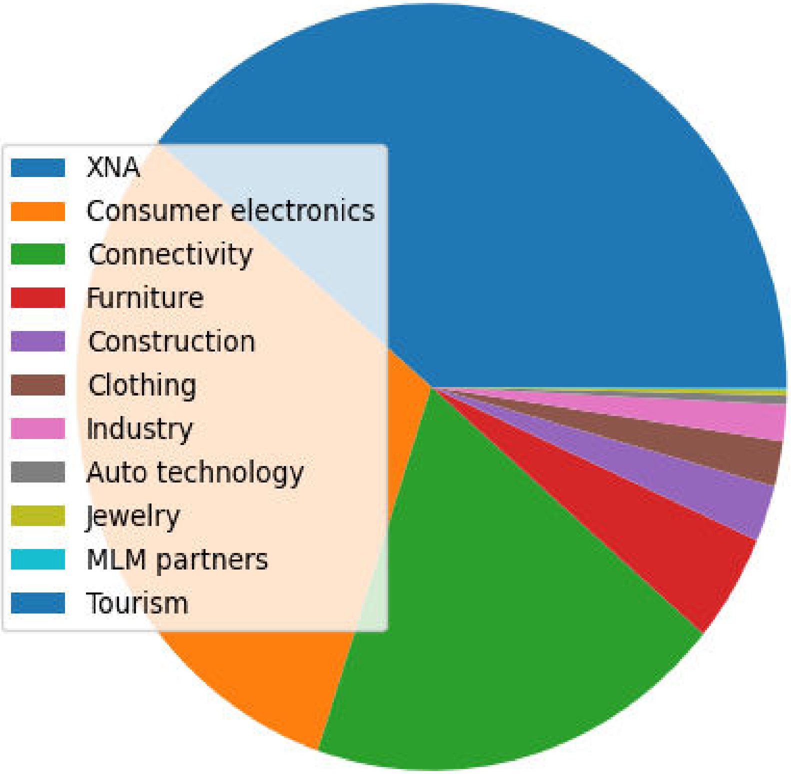
Of all loans, 79% are approved; the remainder are rejected, canceled, or unused.





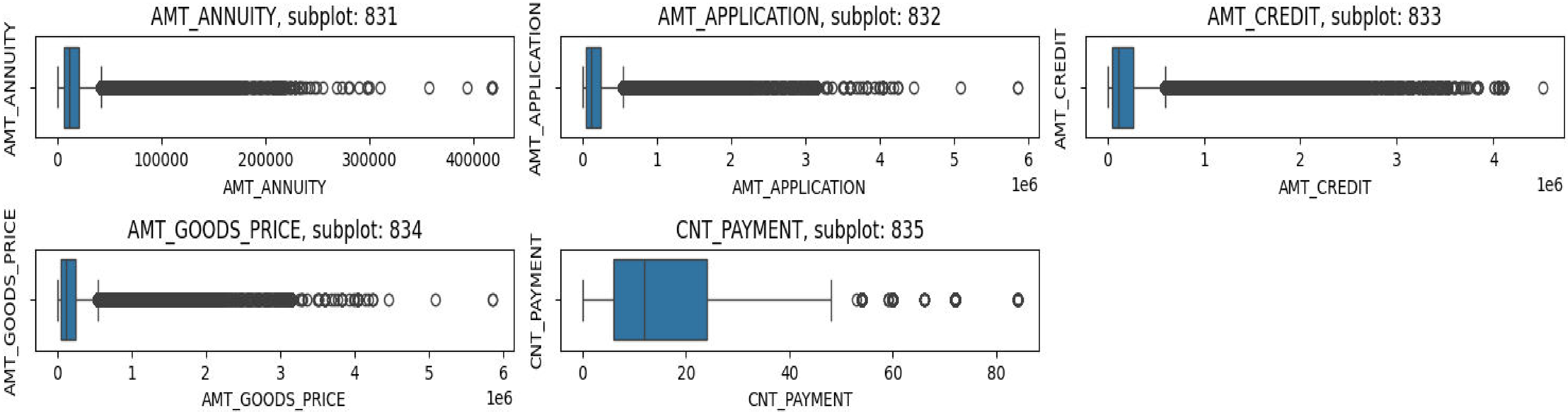
Repeaters make up 67%. Additionally, certain null values appear as XNA in NAME_CLIENT_TYPE.

Of the applicants, 55% obtained a loan to buy a point of sale system.

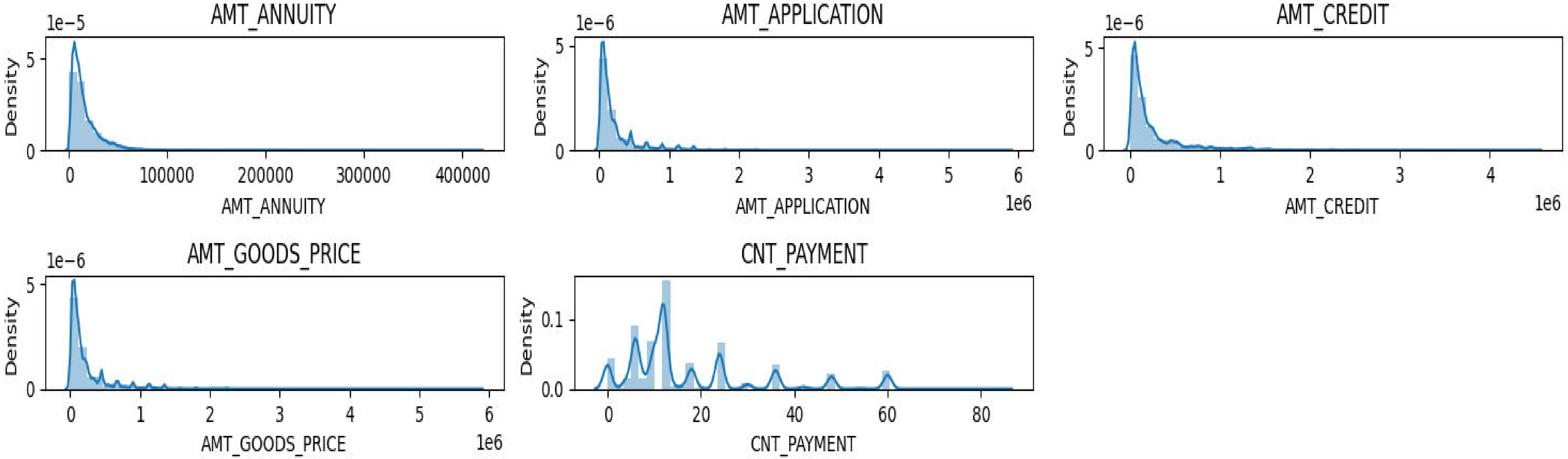


The name seller sector has 37% XNA values; the next-highest category is consumer electronics, at 30%.

# Numerical Variable

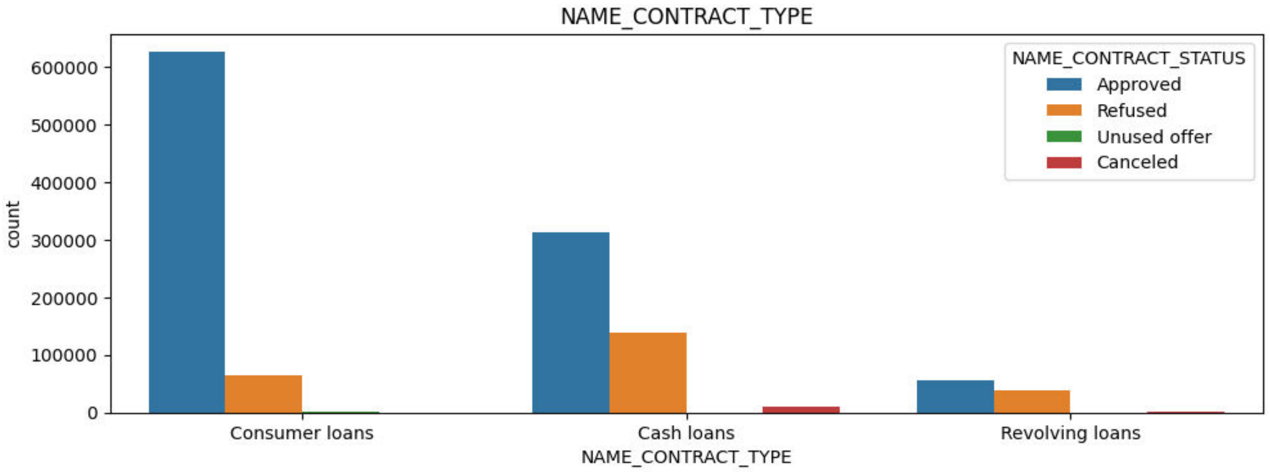


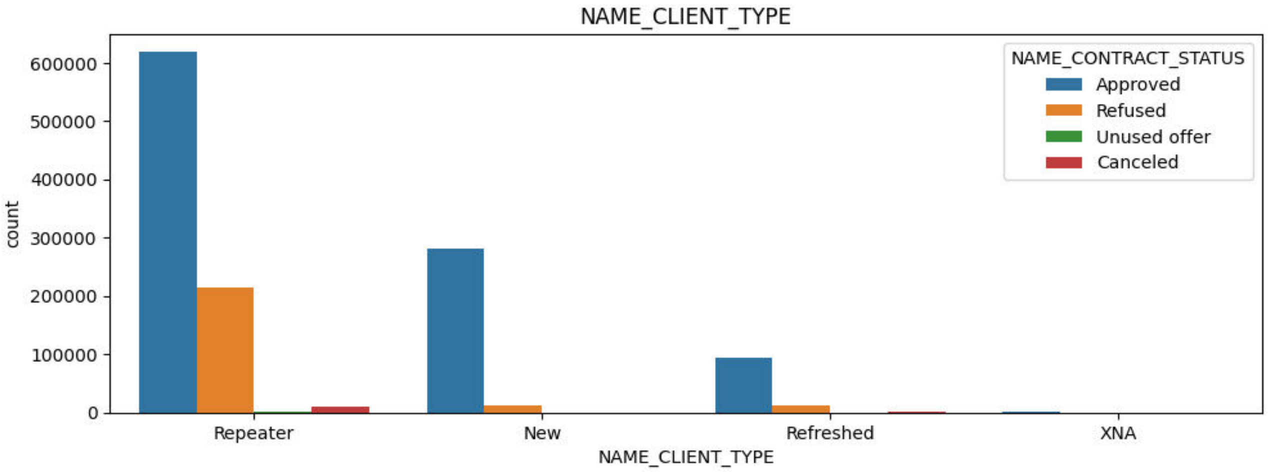
Continuous Variables seem to have high percentage of outliers.

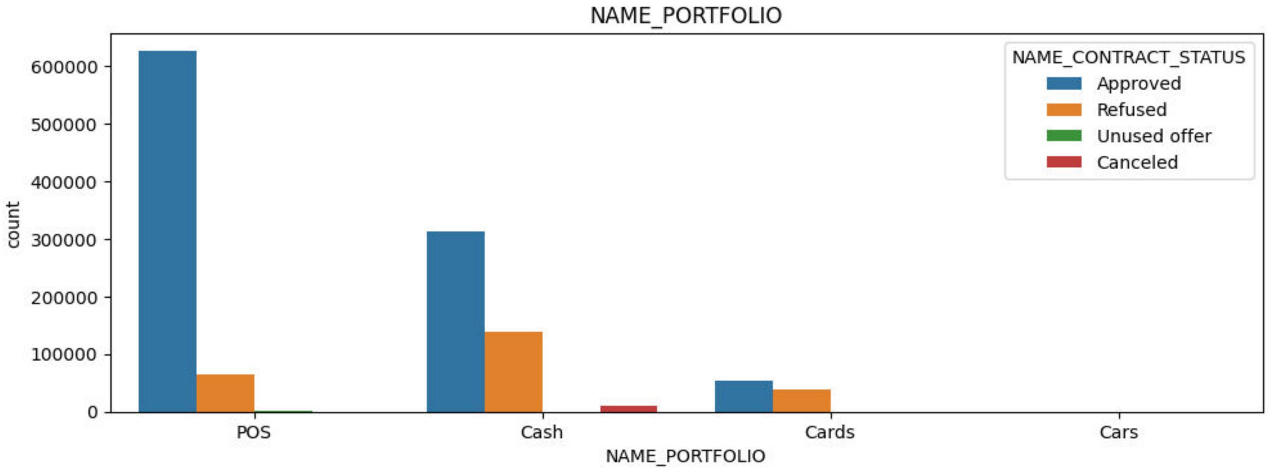


Continuous Variables seem to have high percentage of outliers. Box plot and distribution both indicate the same.

# Bivariate Analysis on Categorical and Continuous Variable







The highest number of applicants for authorized loans is for consumer loans.

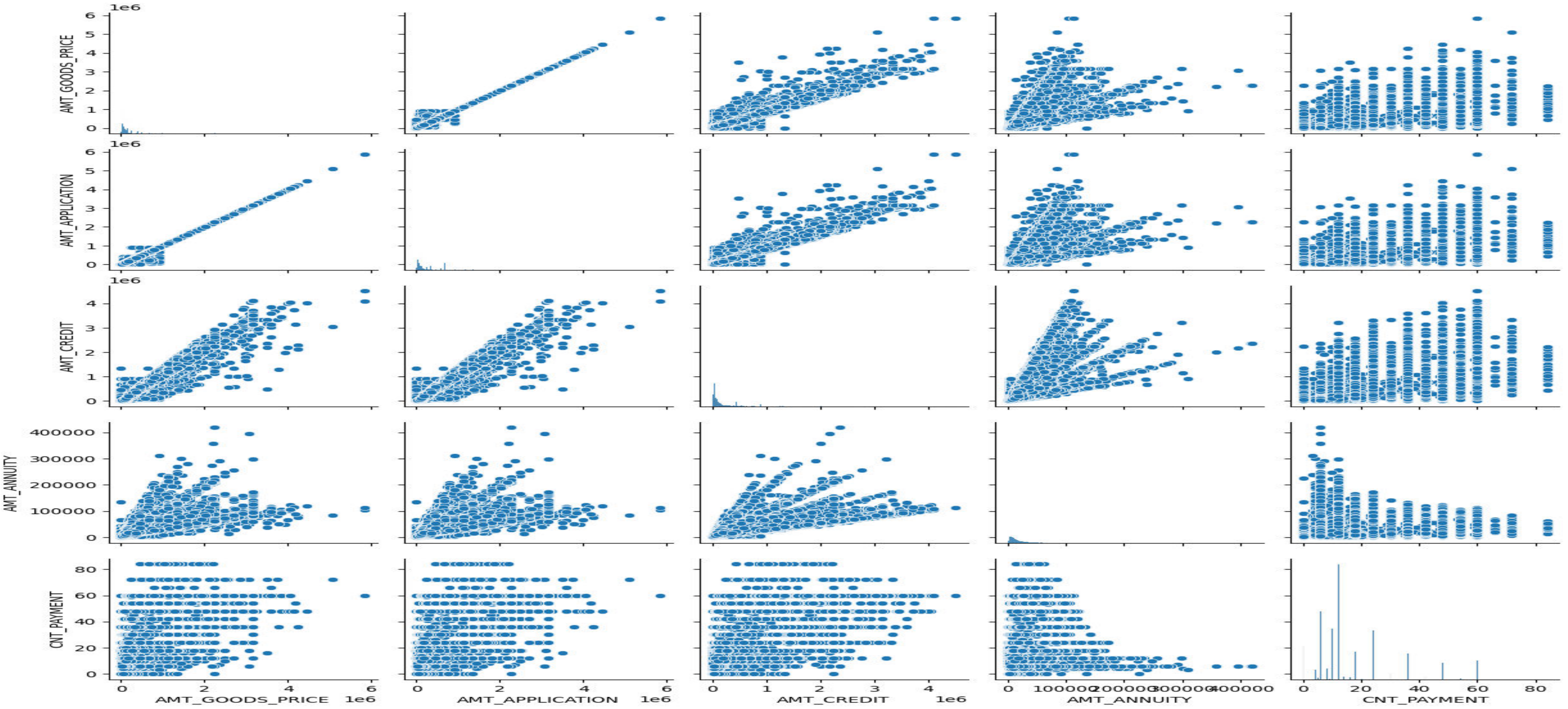
No loans appear to have been canceled in the cash loan category compared to consumer loans.

Compared to consumer loans, more cash loans have been turned down.

The bank has more repeat customers across all categories—approved, rejected, unused, and cancelled.

Like point 2, more cash loans have been denied than point 2; POS transactions appear to be consumer loans.

# Top Correlations



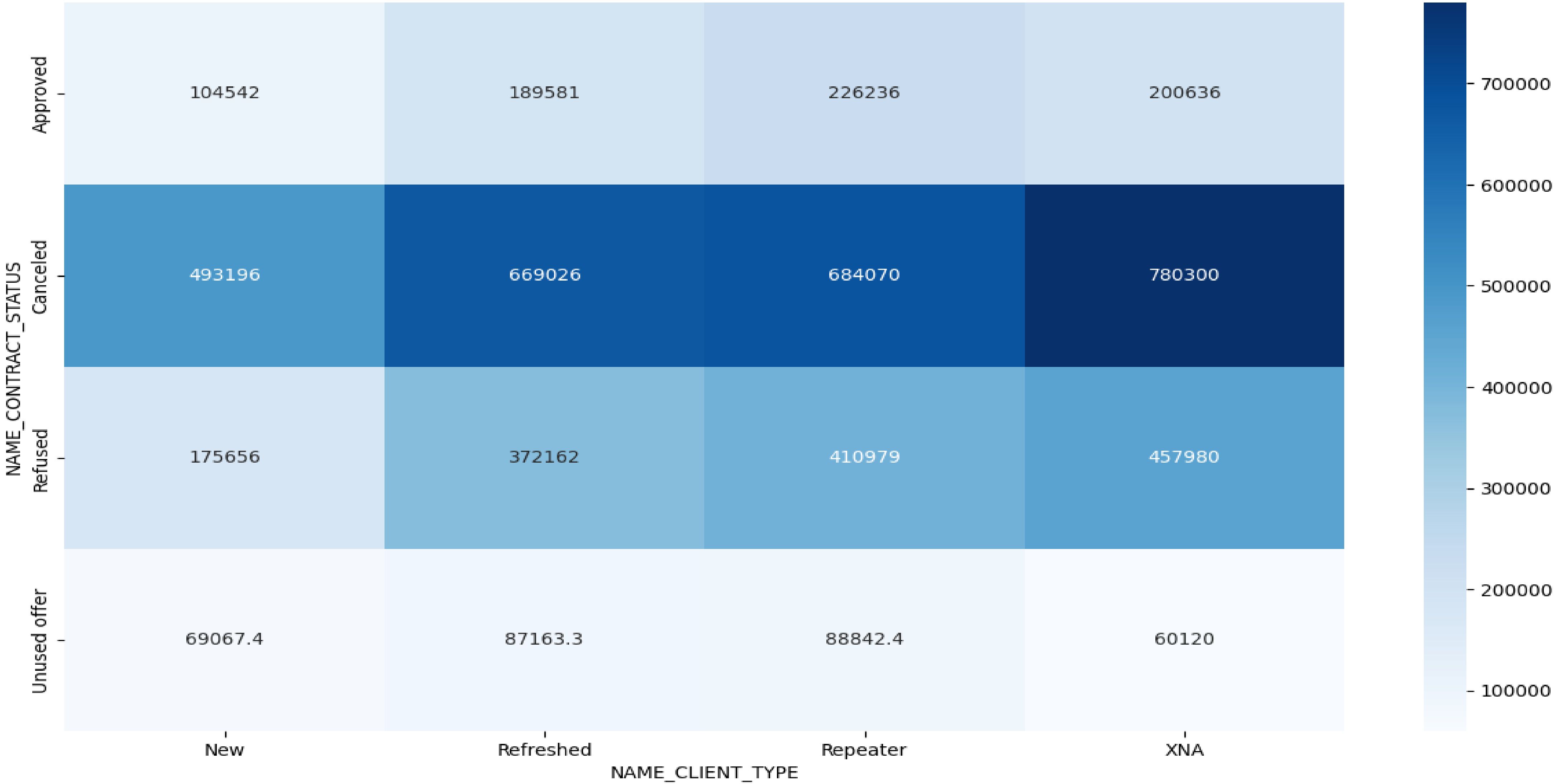
# Observations

-As predicted, there is a strong association between AMT_GOODS_PRICE, AMT_ANNUITY, and AMT_APPLICATION. Loan requirements will increase with the value of the goods purchased, and these factors will undoubtedly correspond.

-AMT_Credit and AMT_GOOD_PRICE are comparable. The link is strong as well.

-A strong association between Column CNT Payment and AMT credit—that is, a higher credit score and a longer term of loan. However, no such correlation is apparent.

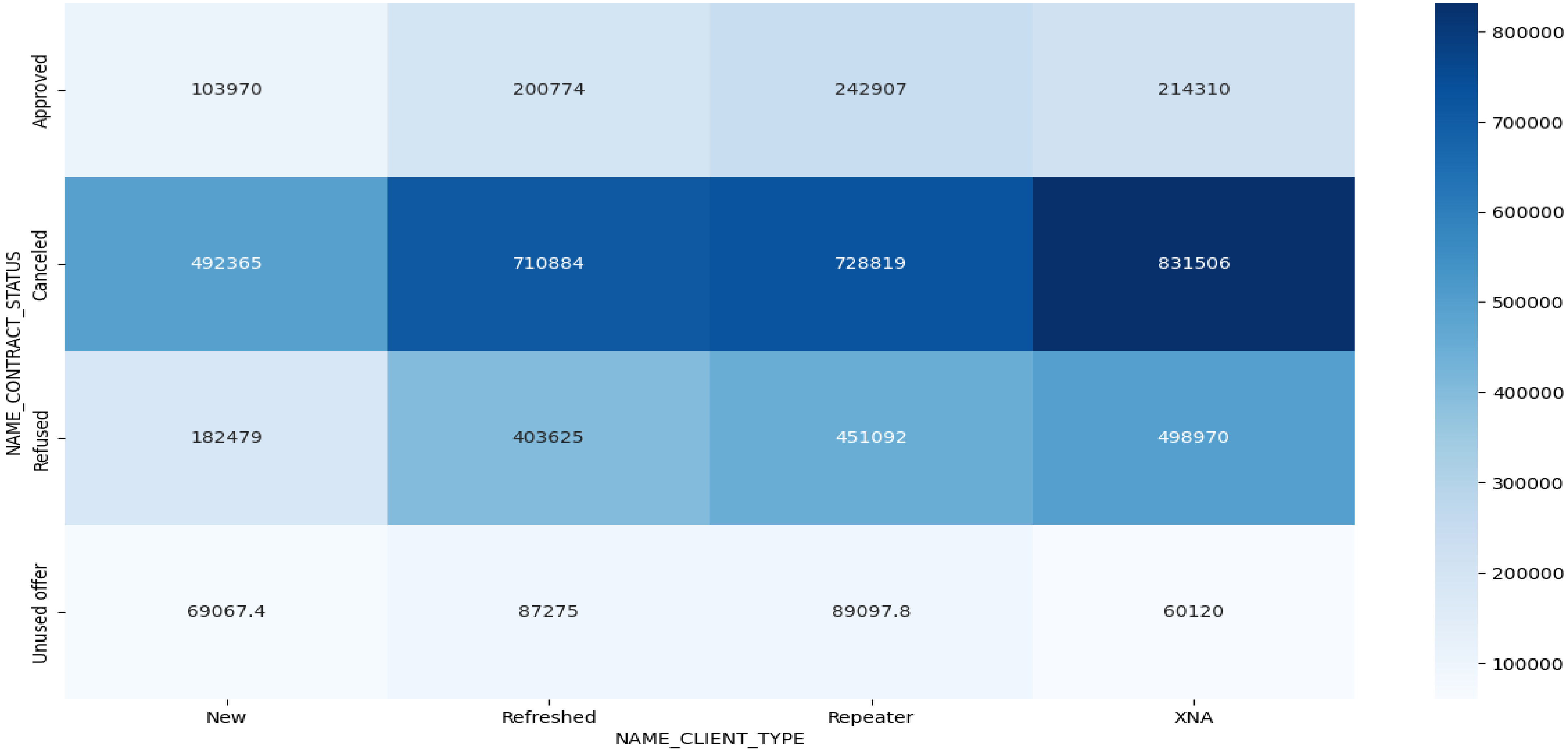
# Multivariate Analysis



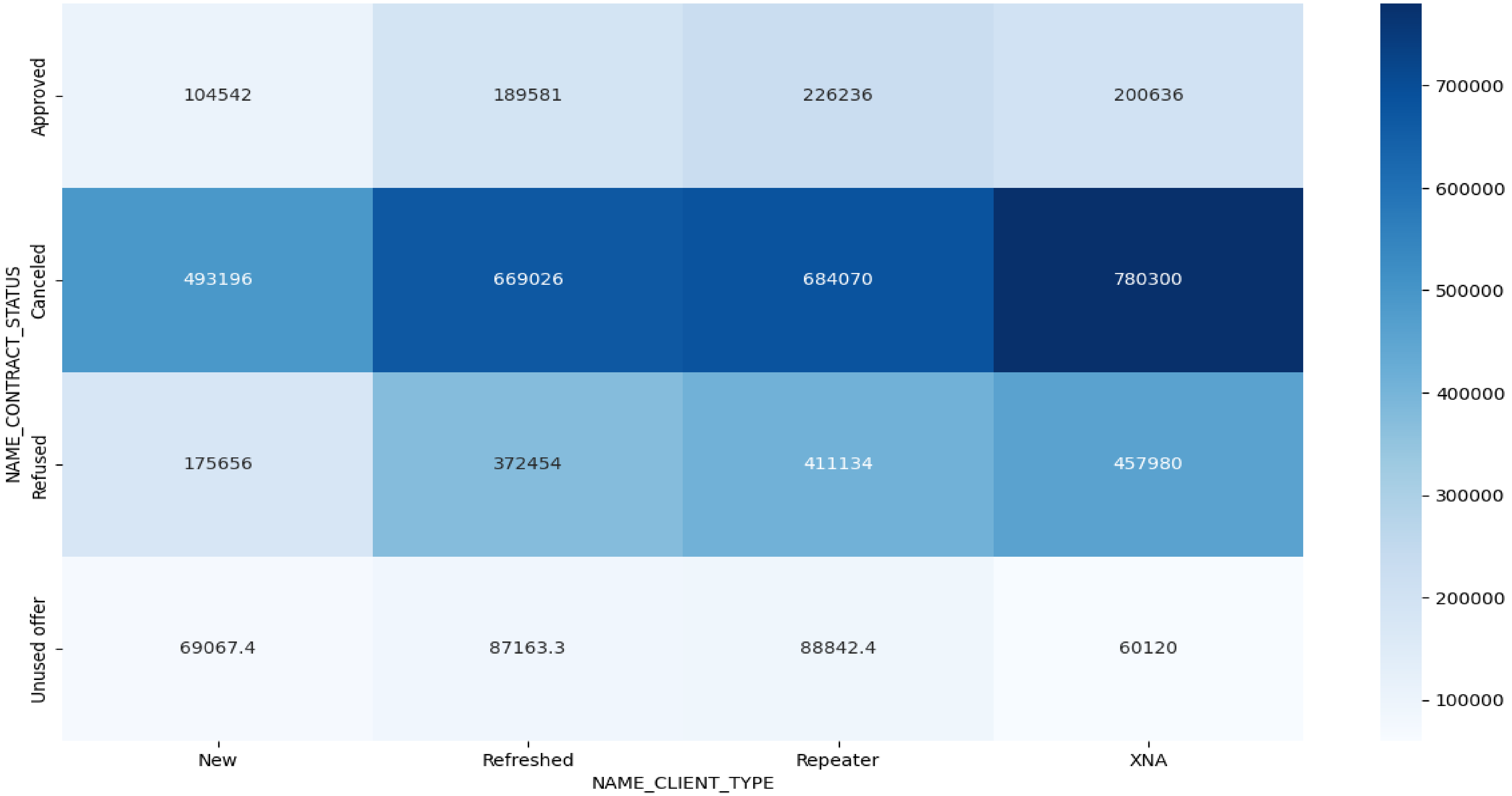
There are few unused offer applications

A large volume of applications were canceled. The bank might be rejecting these because of the potential for a high debt-to-income ratio resulting from the large sum and consequent credit default risk.

The volume of applications from resellers is lower than that of new clients. This could mean that the bank offers repeat applicants more benevolent policies, interest rates, etc.



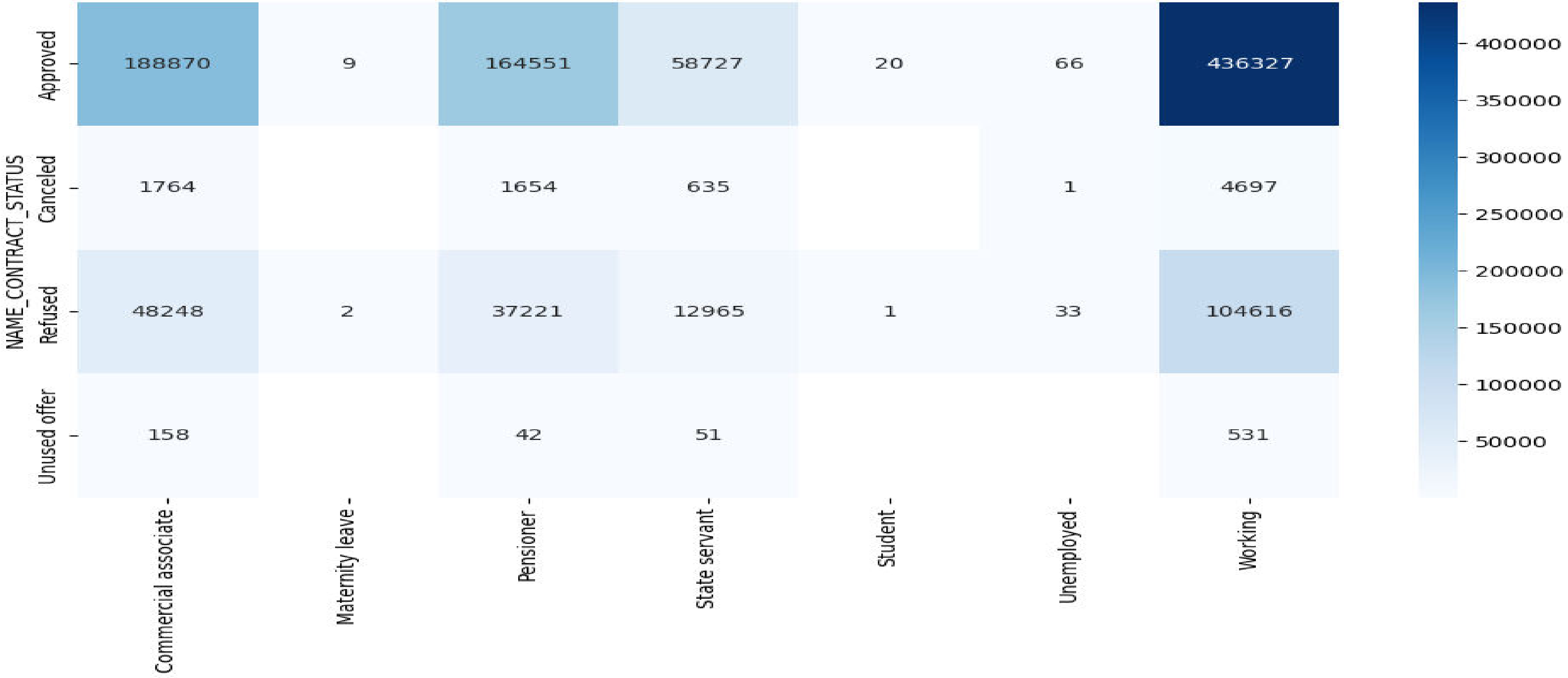
A unused deal There is hardly much credit. This could be the cause of the customer's lack of use. Unable to comprehend why a credit amount should be applied to cancelled and refused orders?.



All cancelled and refused cases have higher value of goods than other categories

# Merged Data frames Analysis

Merged Data Frame contains data from Application_Data file and Previous_Application Data file



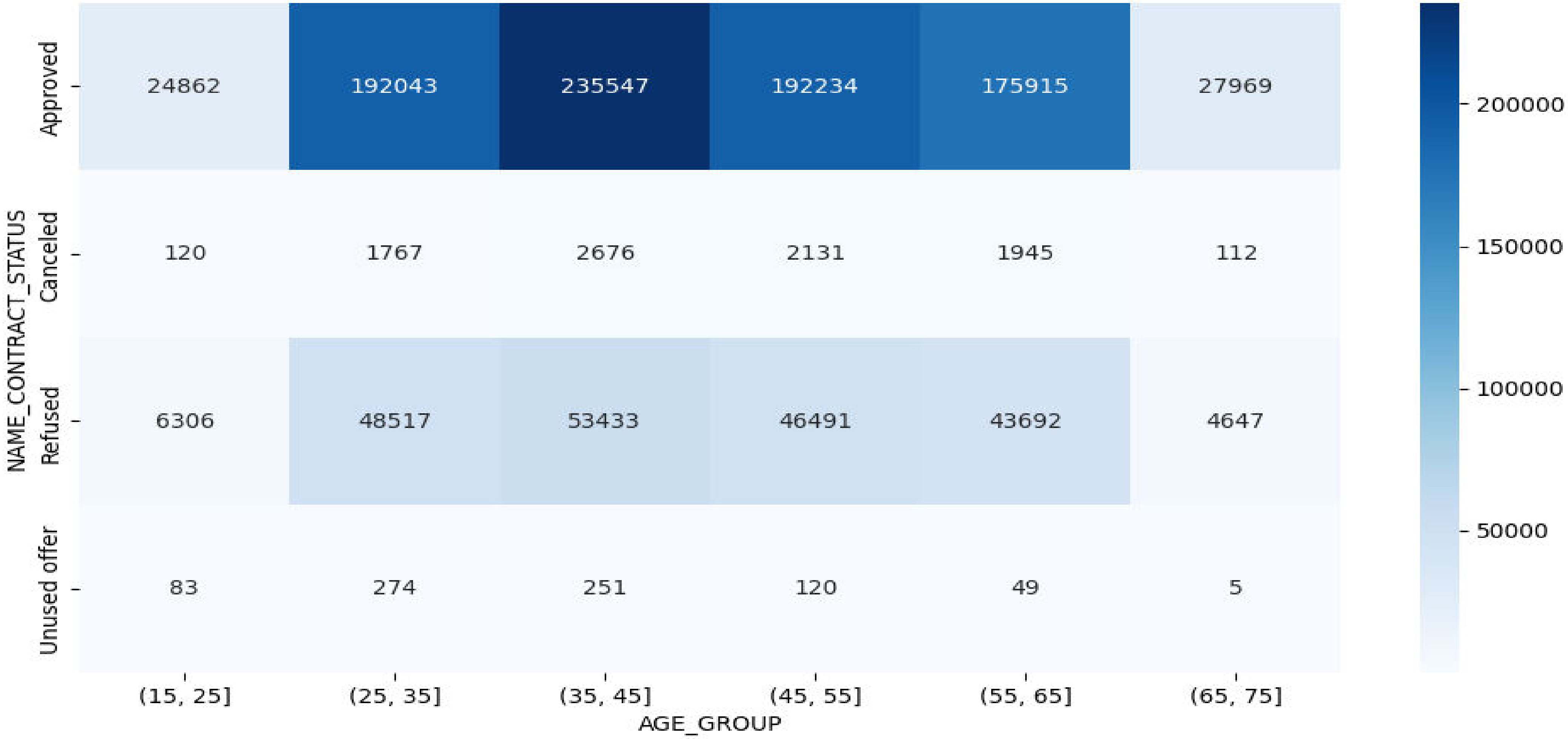
Observations

-Higher on the above matrix indicates correlation to default, since Target 1 is the default.

-The greatest majority of working applicants with approved status have defaulted.

-It's concerning that prior applications including rejected, canceled, and unused loans also had defaults. This shows that the financial institution granted the new application despite having rejected or canceled the prior one, and that they are now risking loan default.

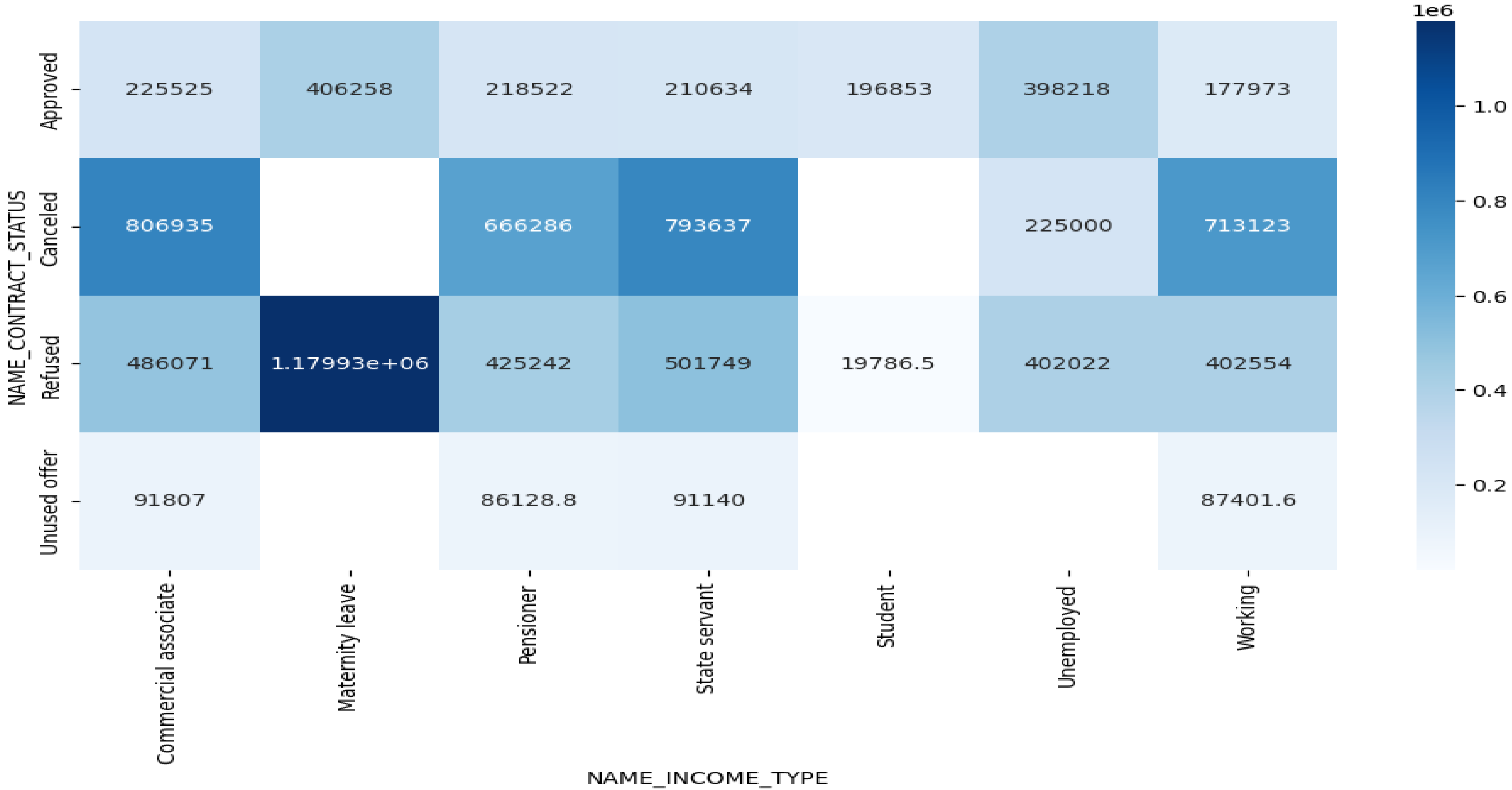
- 14,389 working-class applicants who were previously denied have since defaulted.



- Higher on the above matrix indicates correlation to default, since Target 1 is the default.

- The age groups of 25–35 and 35–45 had greater default rates on approved loans.

- Loans in the prior application were denied, canceled, and now defaulted.



- Greater respect given to jobless people; maternity leave is a significant component

- Unused offers have lower credit values, which may be the cause of the applicant's failure to use them.

## Default cases in approved applications

- The application data frame analysis determined that each of the characteristics listed below would result in a default.
- These were verified by comparison with the accepted applications and default cases, and the results are accurate.
- High Default
- "INCOME_GROUP": A group with a moderate income
- 'AGE_GROUPS': 25–35, then 35–45
- "NAME_INCOME_TYPE": Operational
- "OCCUPATION_TYPE" - 31% by Laborers
- 'ORGANIZATION_TYPE' - Type of business
- 'OWN_CAR_flag' - 31% do not own a vehicle
- 'OWN_REALTY_flag': 70% of people do not own a home

# Case Summary

The demographics of defaulters during the application data frame examination, all of the variables listed below were found to have default values.

These were verified by comparing them to the approved loans that have defaulted, and the results are accurate.

- Middle class
- 25–35 years old, then the 35–45 age group
- A man
- Jobless
- Workers, Salespeople, and Drivers
- Type of business 3
- Residence - None

Other crucial factors to take into account are:

- Days since the last phone number change; - Lower figure points that need to be addressed; - Number of bureau hits in the previous week. Zero hits in a month, etc., is good.

- Amount of money is not commensurate with the good purchased; this raises concerns about low income and high good value.

- There is cause for concern as prior applications involving loans that were declined, canceled, or unused also had default. This shows that the financial institution granted the latest application despite having rejected or canceled the prior one, and they are currently in default on these.

- Rejected credible applications had less loan amounts because they were not used.

- Since there are fewer defaults, female applicants should be given more consideration.

- Applicants who are employed make up 60% of defaulters. It is not necessary to reject applicants who are employed, however. Thorough examination of additional factors is required.

- There are examples in which payments are being made on schedule for the current application despite prior applications having loans that were declined, canceled, or unused. This suggests that such cases may have involved poor decision-making.

# Thank You!