CREDIT EDA ASSIGNMENT

Compiled by Shraavan Sridhar

PROBLEM STATEMENT I

- When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile.

Two types of risks are associated with the bank’s decision:

- If the applicant 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, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
- Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

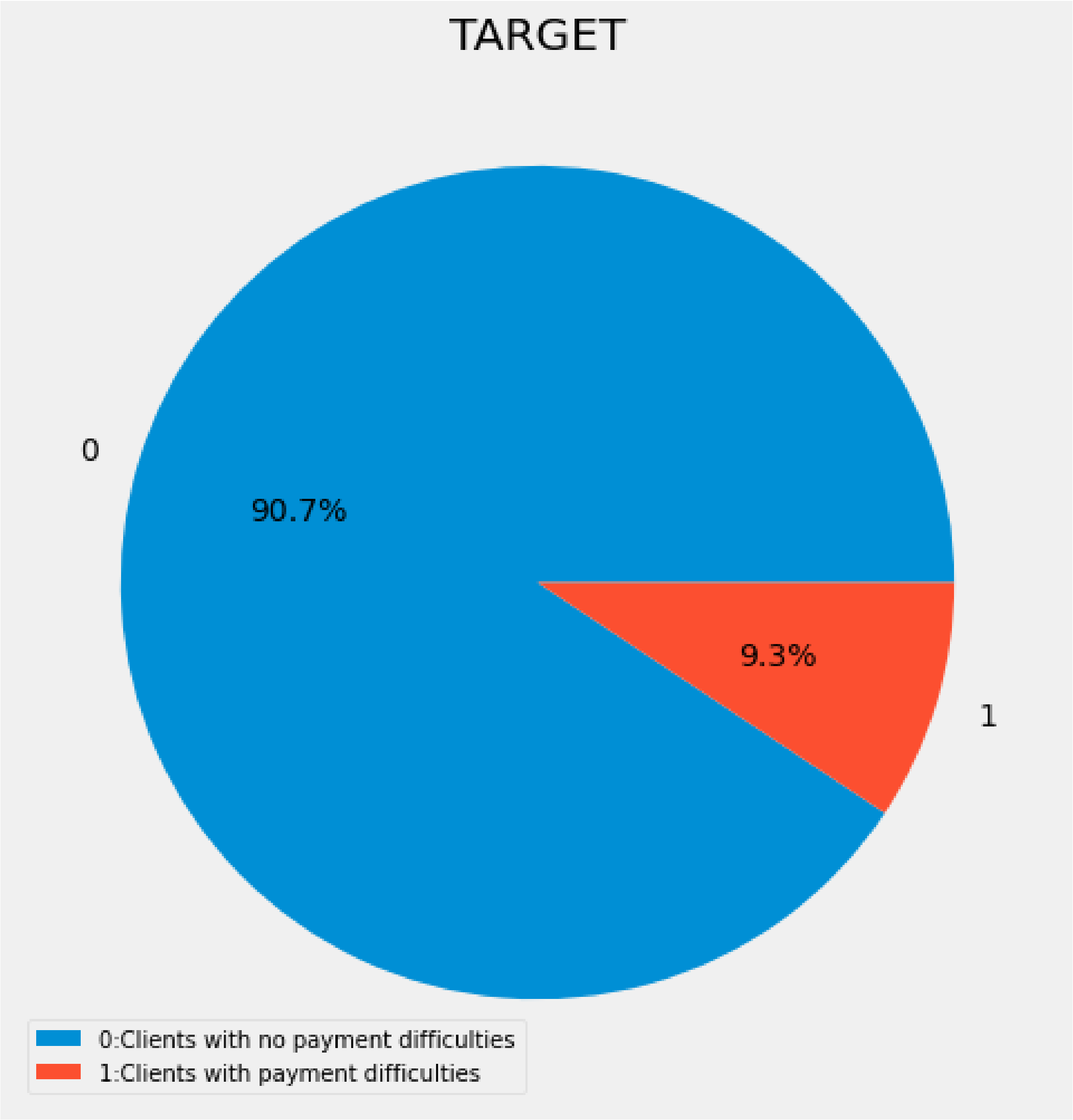
- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.

PROBLEM STATEMENT II

- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
- Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

HIGH DATA IMBALANCE - TARGET

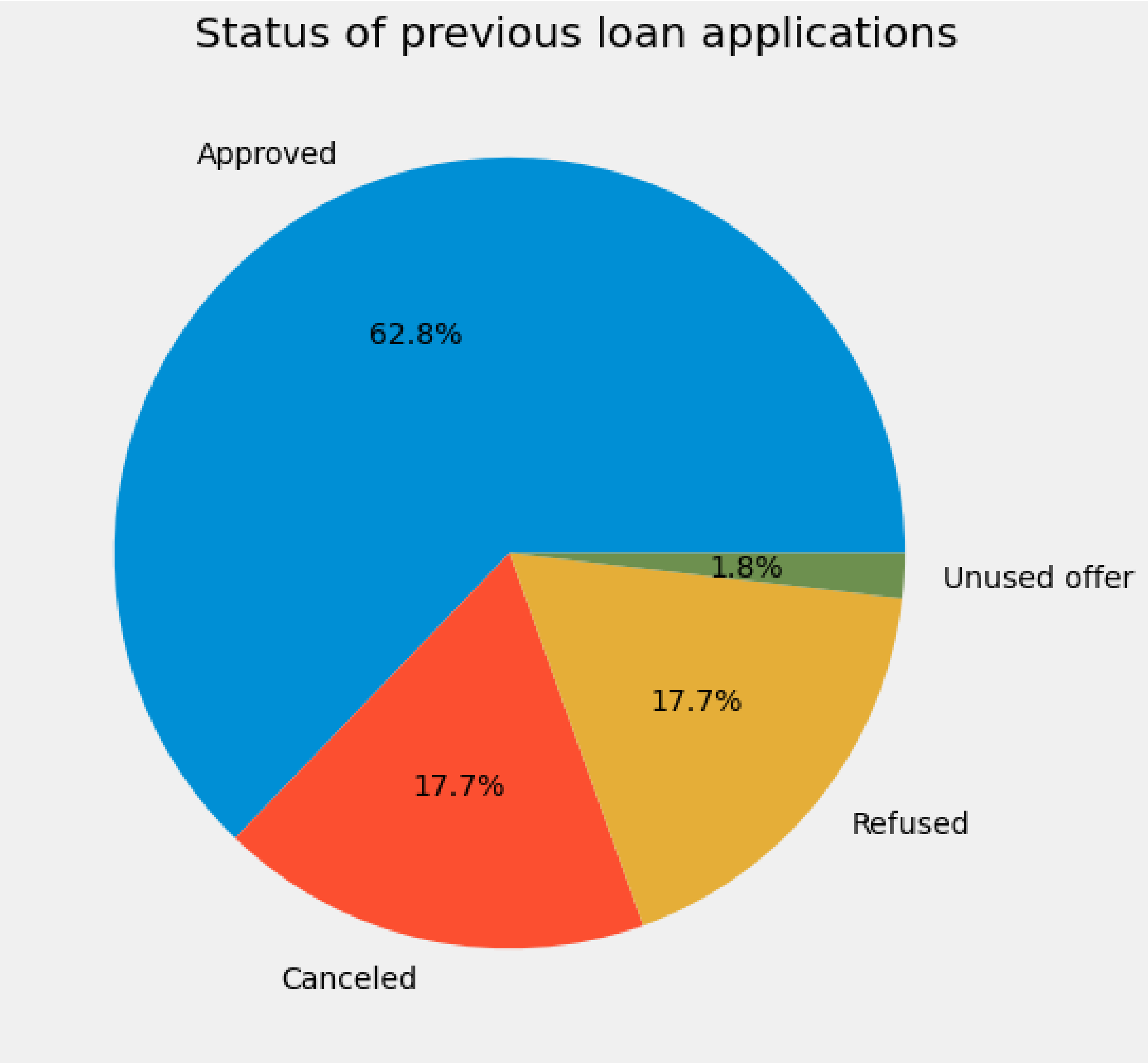
Clients who do not have payment difficulties
VS
Clients who have payment difficulties.



- Most of the clients who have applied for loans do not have payment difficulties about 91%.
- Only 9% of clients who have applied for loans have payment difficulties.
- The imbalance ratio is at 91:9.

UNIVARIATE ANALYSIS

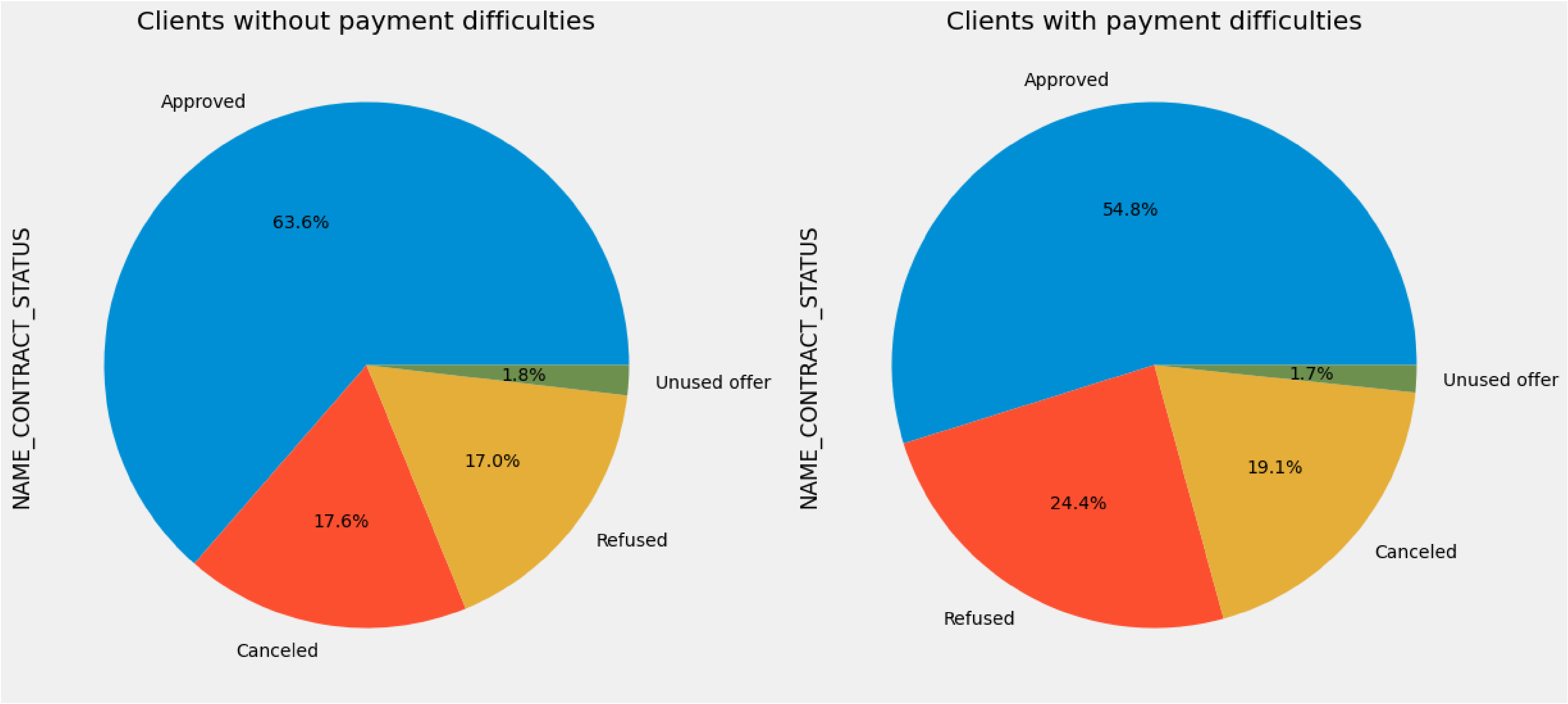
APPLICATION DISTRIBUTION



- Approval rate is approx. 63%. Majority of the loans get approved.
- Rejection Rate is at 17.7%

UNIVARIATE ANALYSIS

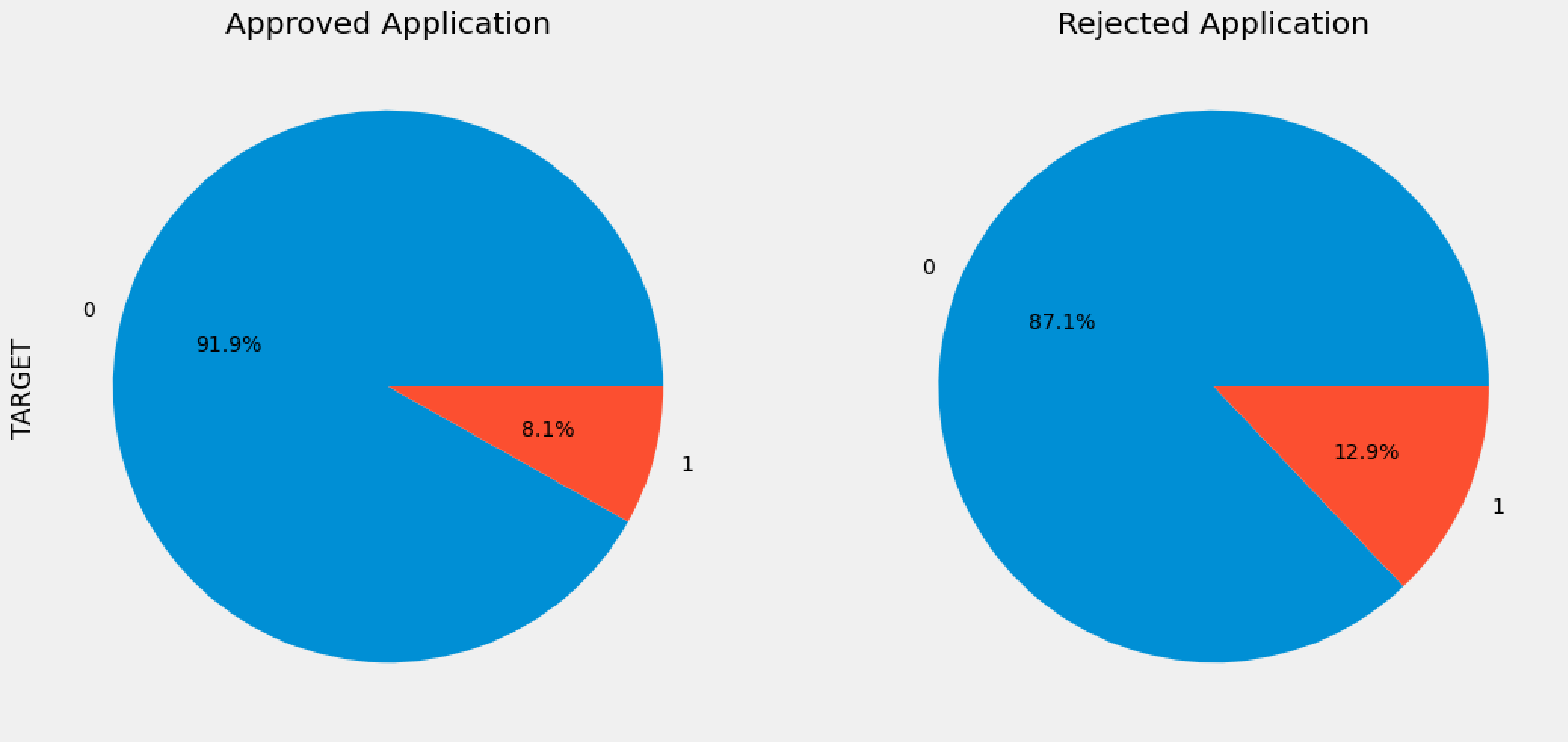
APPLICATION DISTRIBUTION AMONG TARGET GROUPS



- Clients without payment difficulties are more likely to get their loans approved.
- Also, their refusal percentage is lower compared to clients with payment difficulties.

UNIVARIATE ANALYSIS

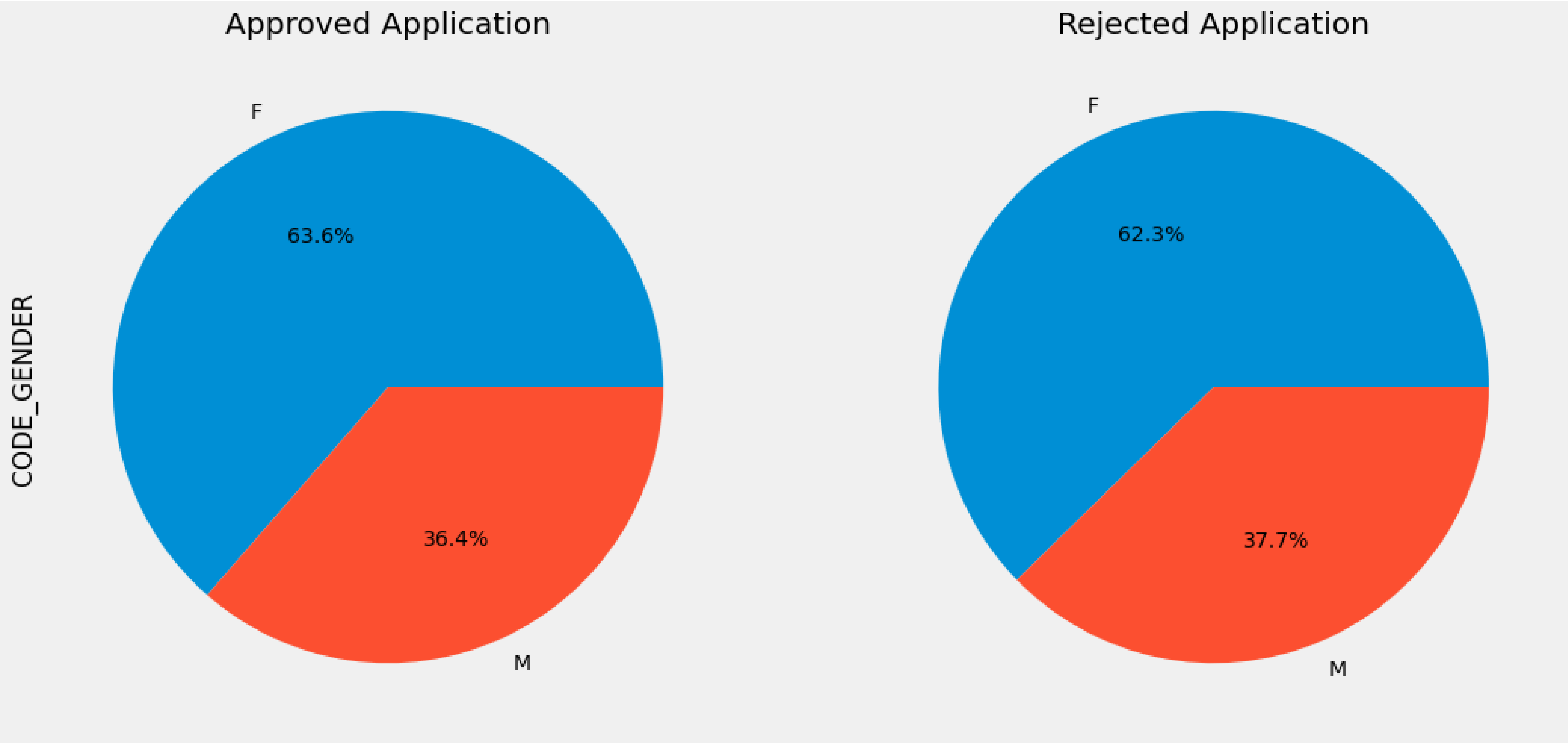
APPLICATION DISTRIBUTION AMONG TARGET GROUPS



- For Approved applications (0), clients who had payment difficulties is only 8.1%
- For Rejected applications (1), clients who had payment difficulties is higher at 12.9%.

UNIVARIATE ANALYSIS

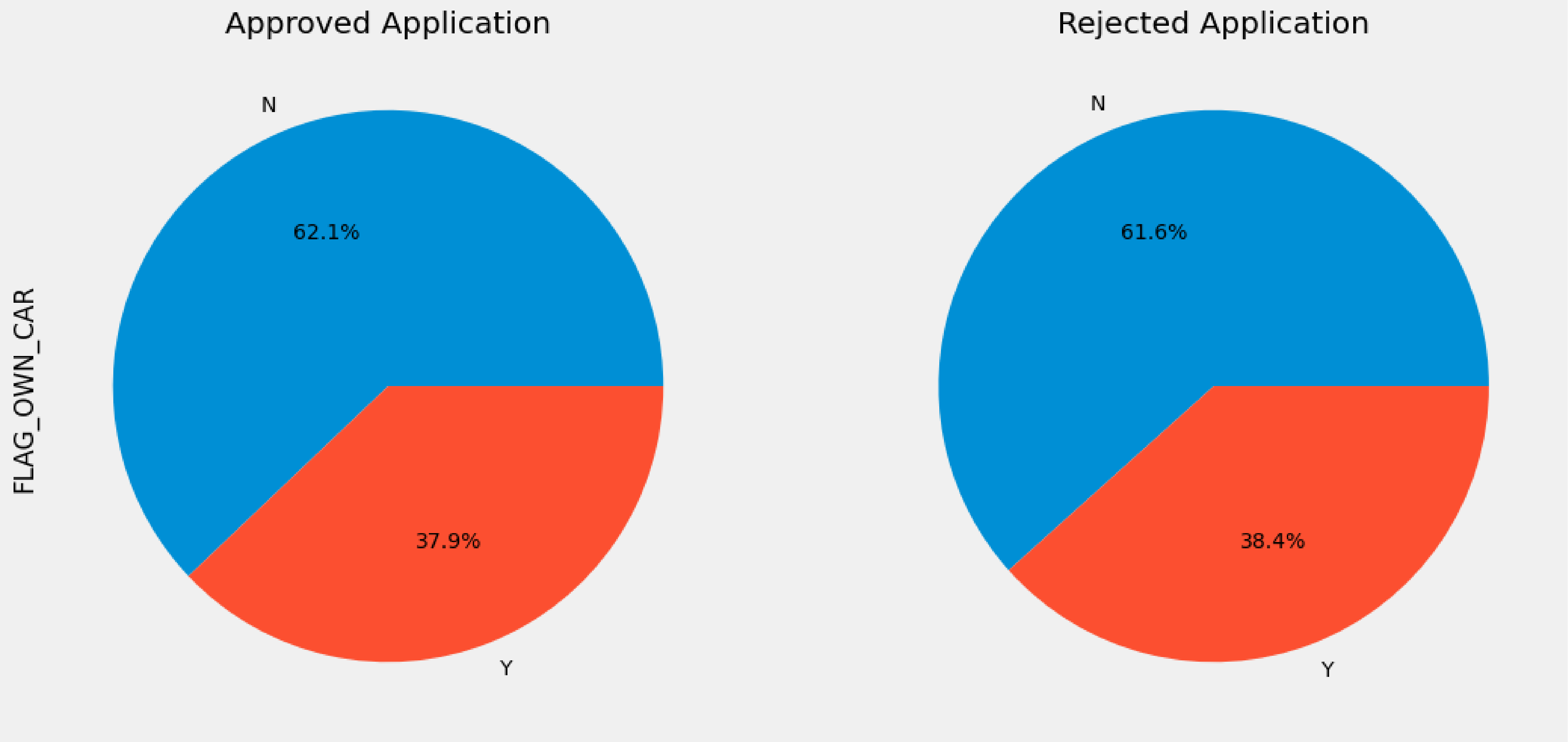
APPLICATION DISTRIBUTION AMONG GENDER



- The stats here are quite similar wherein both males and females face similar approvals and rejections.

UNIVARIATE ANALYSIS

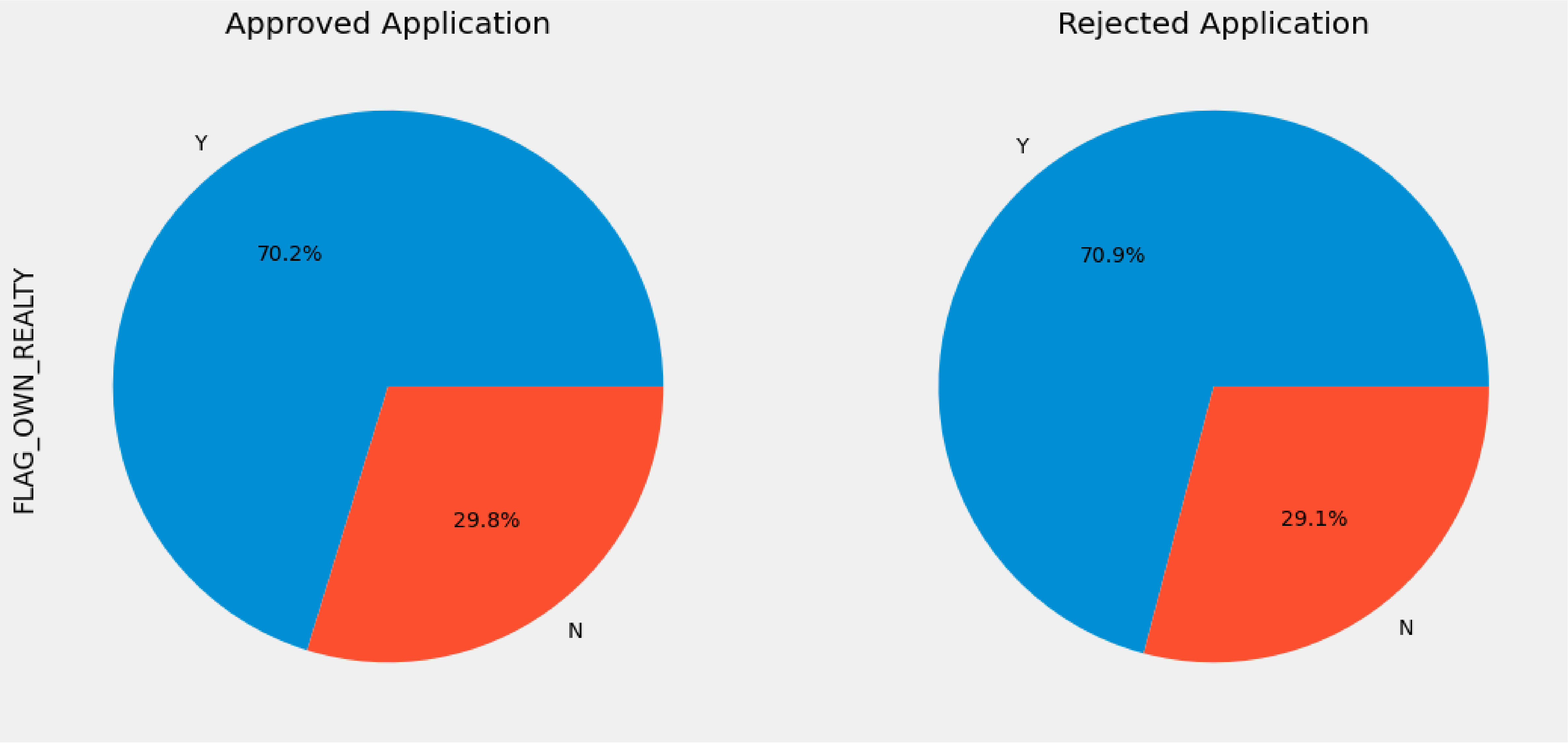
APPLICATION DISTRIBUTION - OWNING A CAR



- Most clients in both the categories do not own a car and owning a car does not affect the status of application.

UNIVARIATE ANALYSIS

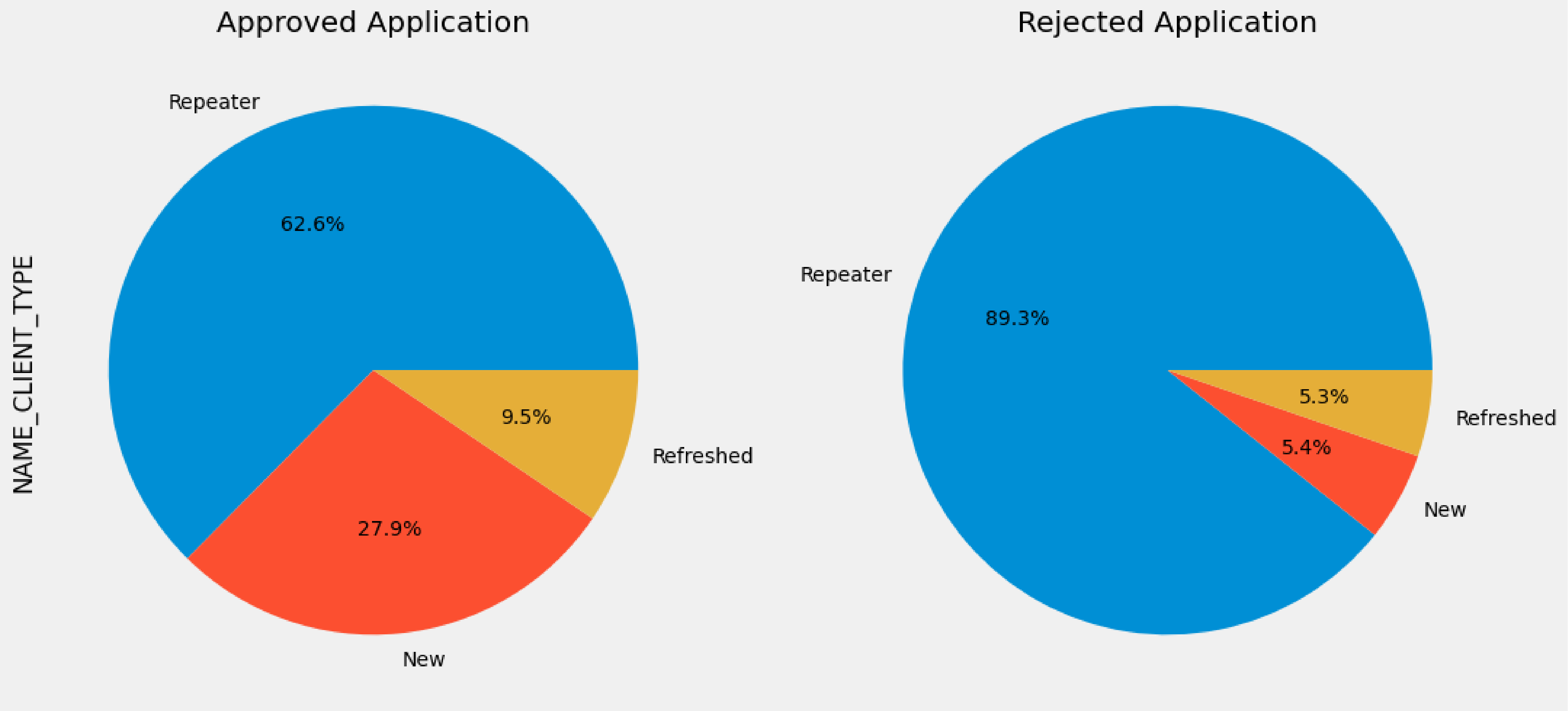
APPLICATION DISTRIBUTION OWNING REALTY



- Most clients own a house/apartment and again this does not have an impact on the approval or rejection.

UNIVARIATE ANALYSIS

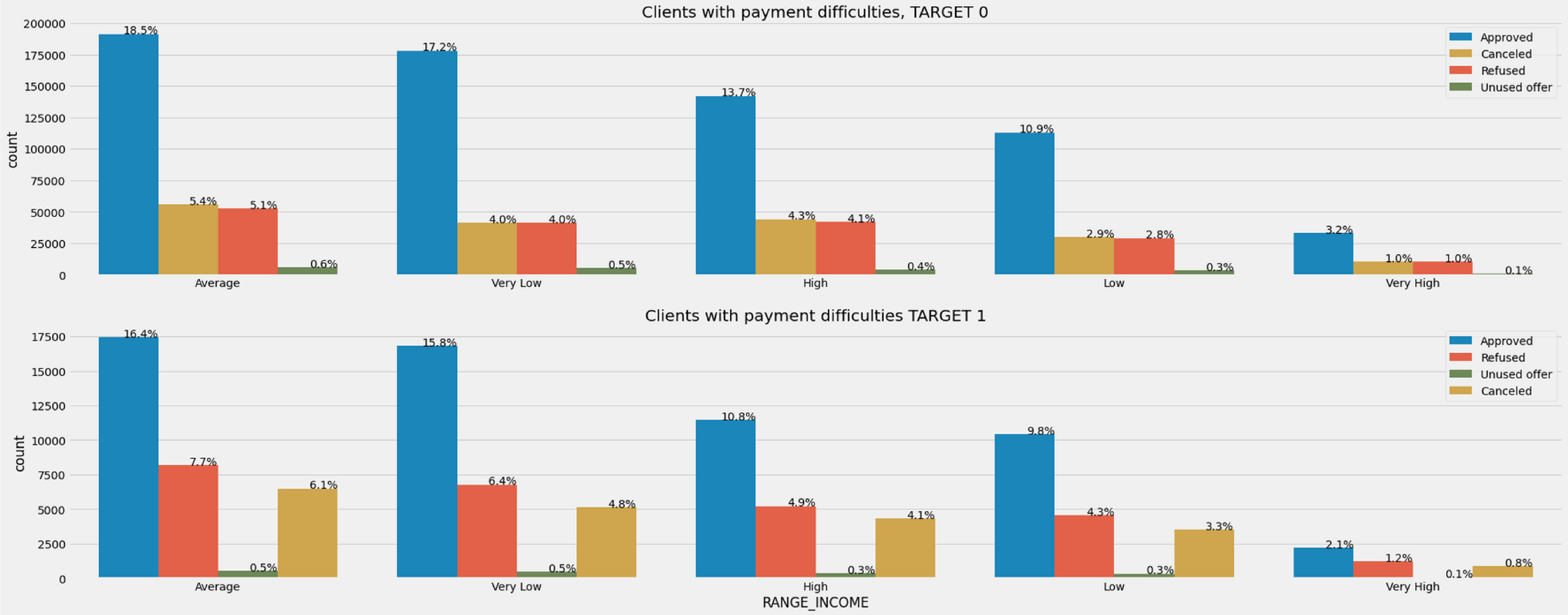
APPLICATION DISTRIBUTION AMONG CLIENT TYPES



- The applications that are rejected the most are of the "Repeaters".
- Approx. 28% of new application have been approved.

UNIVARIATE ANALYSIS

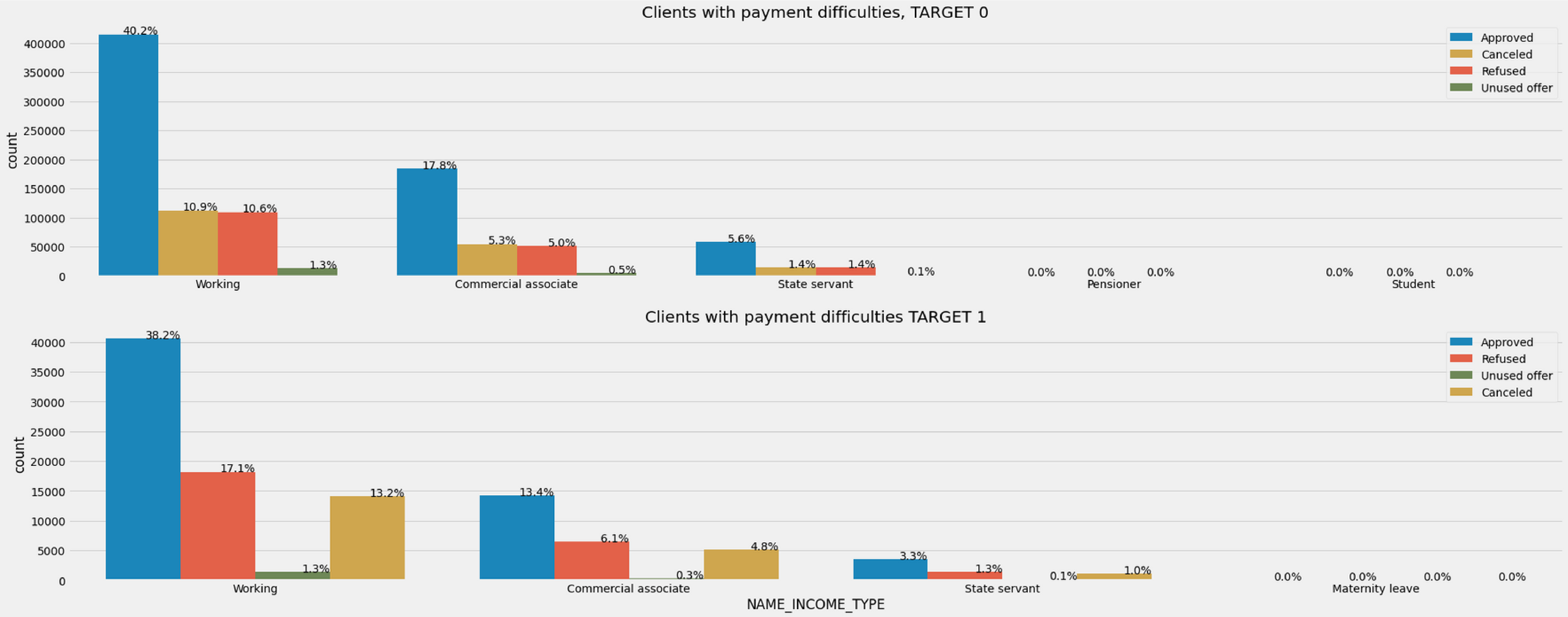
INCOME RANGE DISTRIBUTION



- For clients with payment difficulties, the rejection rate among all the income categories is higher compared to the clients with no payment difficulties.
- Rejection Rate for clients among the average income categories seem the highest (5.1% for TARGET0 and 7.7% for TARGET1).

UNIVARIATE ANALYSIS

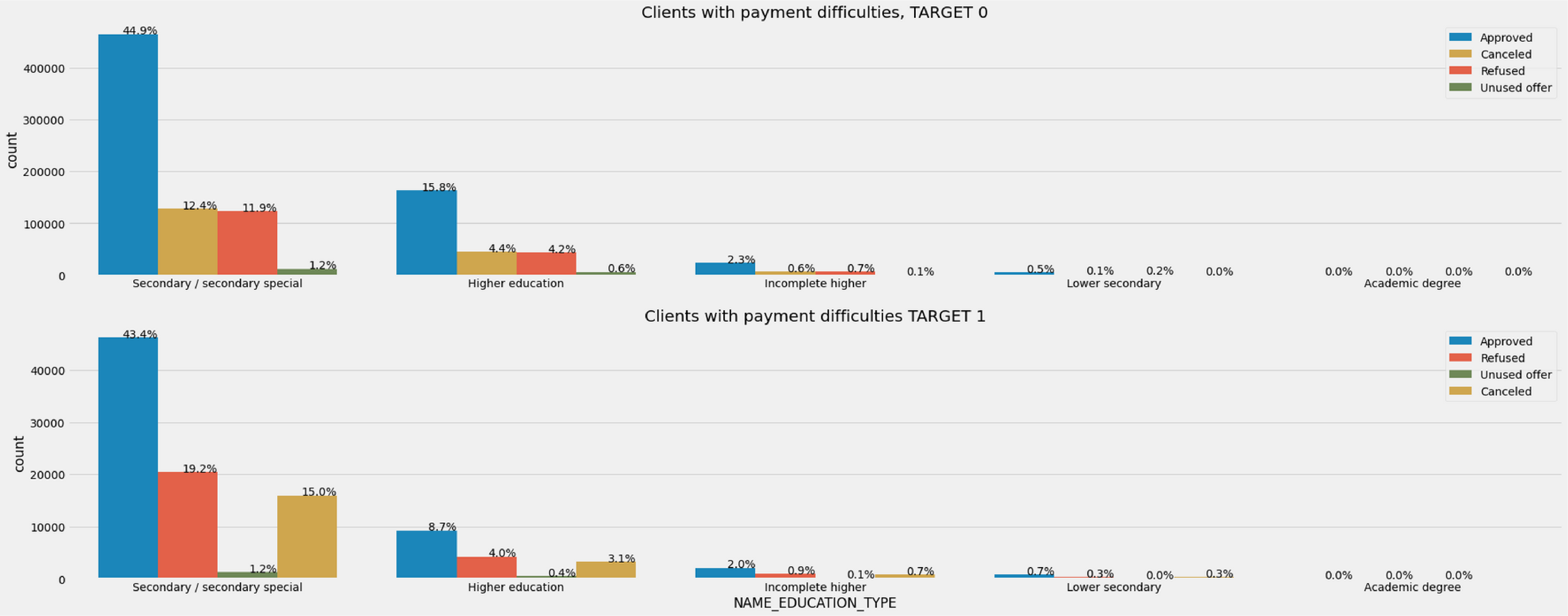
INCOME TYPE DISTRIBUTION



- State Servants have lesser payment difficulties.
- The rejection rate for working class with payment difficulties seem very high (17%) For clients belonging to this category, it is better to be cautious while approving loans.
- Unused offers are highest among working class.

UNIVARIATE ANALYSIS

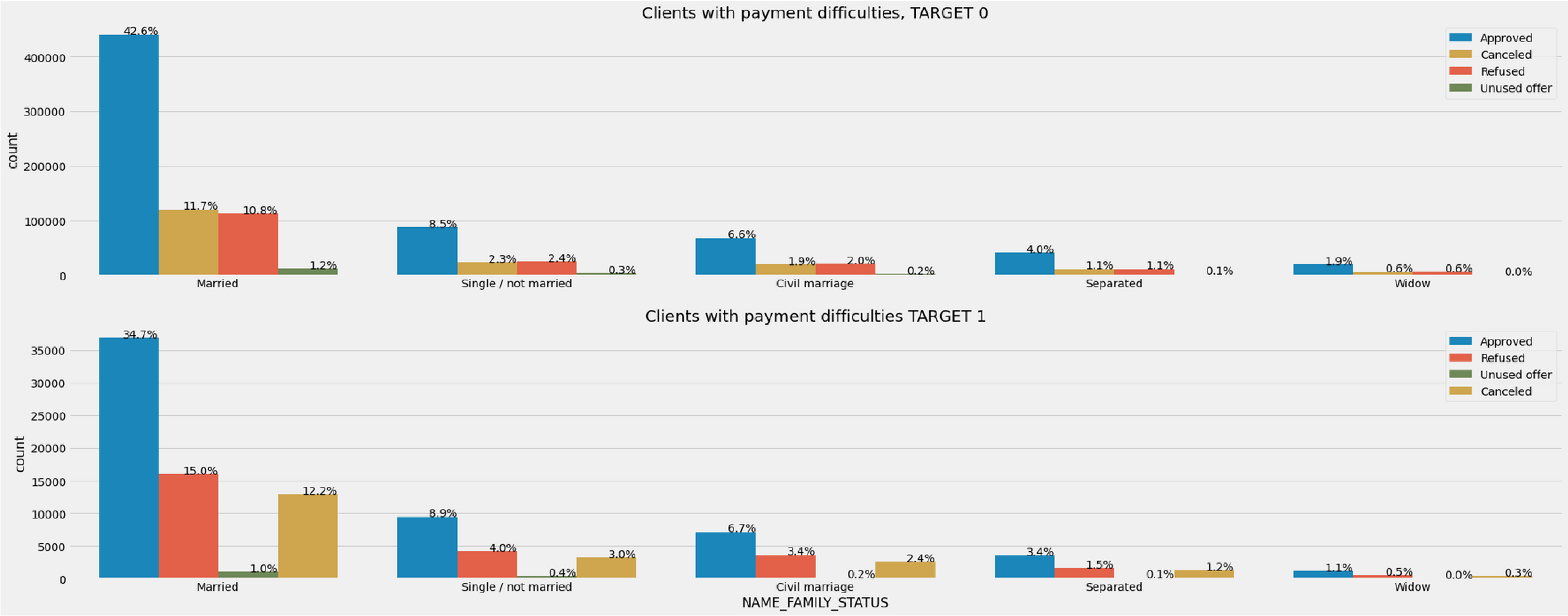
EDUCATION DISTRIBUTION



- For secondary special, applications rejections are the highest. For clients with payment difficulties the refusal or rejection stand at 19.2%
- For higher education, the rejections are less.

UNIVARIATE ANALYSIS

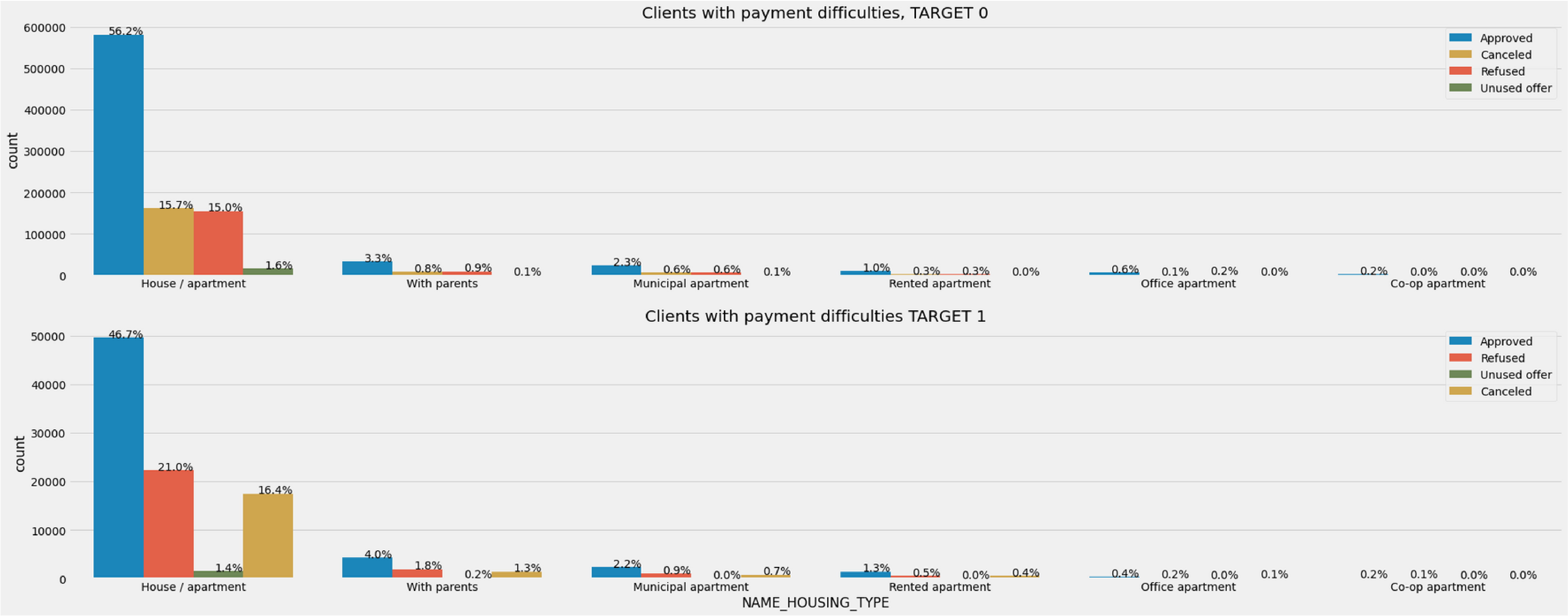
FAMILY STATUS



- The approval rates for FAMILY STATUS remains the same irrespective of the payment difficulties except for married clients wherein they have only 34% approval rating compared to 43% for TARGET0.
- The rate of rejection also seem high for Married clients with payment difficulties.

UNIVARIATE ANALYSIS

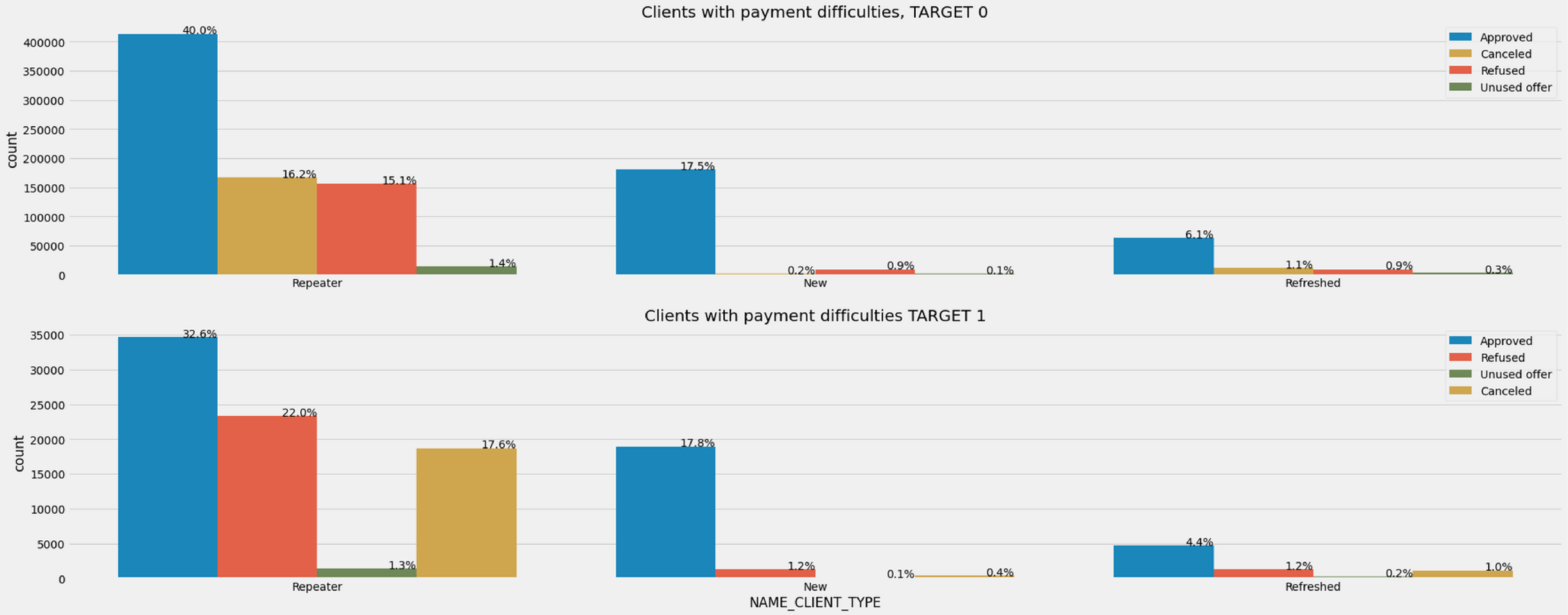
HOUSING TYPE



- Clients who do not stay in House/Apartments have less rejections.
- Rejections are highest for clients who stay in apartments.
- One must be cautious while approving loans to clients who stay in apartments with payment difficulties.

UNIVARIATE ANALYSIS

CLIENT TYPE

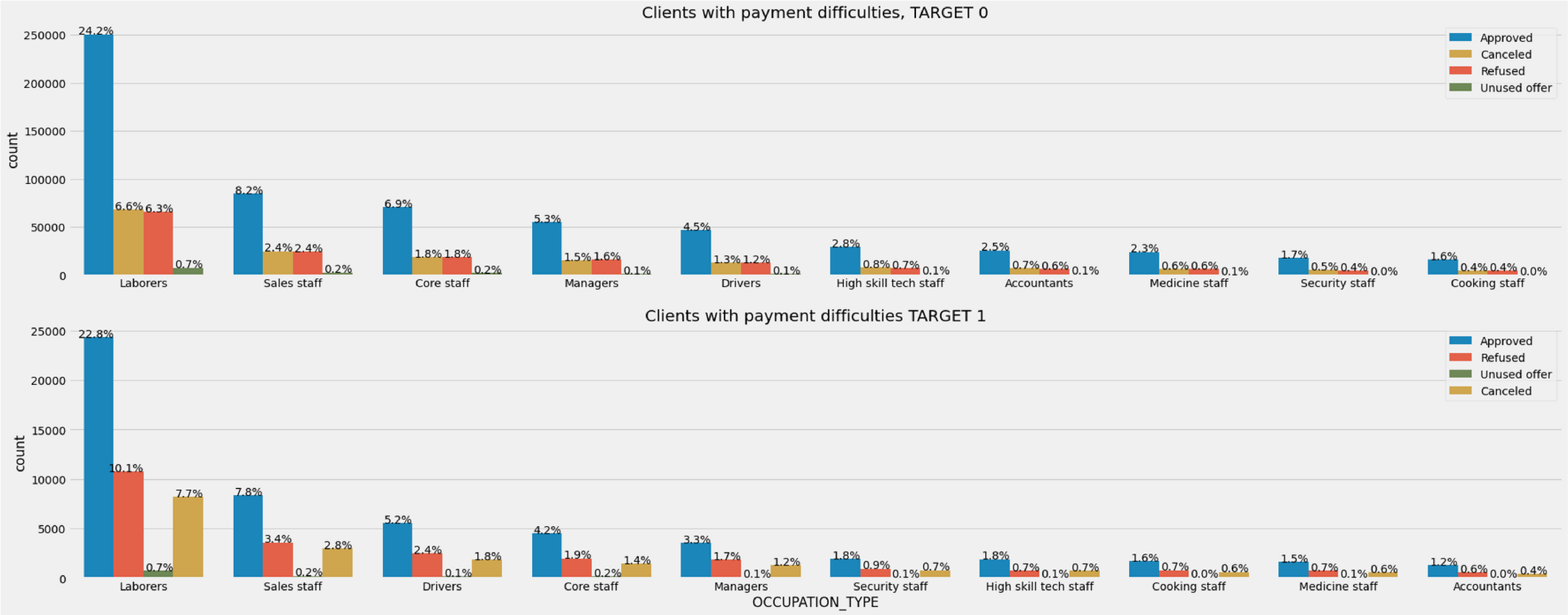

Clients with payment difficulties, TARGET 0

Clients with payment difficulties TARGET 1

- Among repeater clients, they face the most rejections.
- Refreshed and new clients do not have that many rejections.
- Approach with caution while approving loans to repeater clients.

UNIVARIATE ANALYSIS

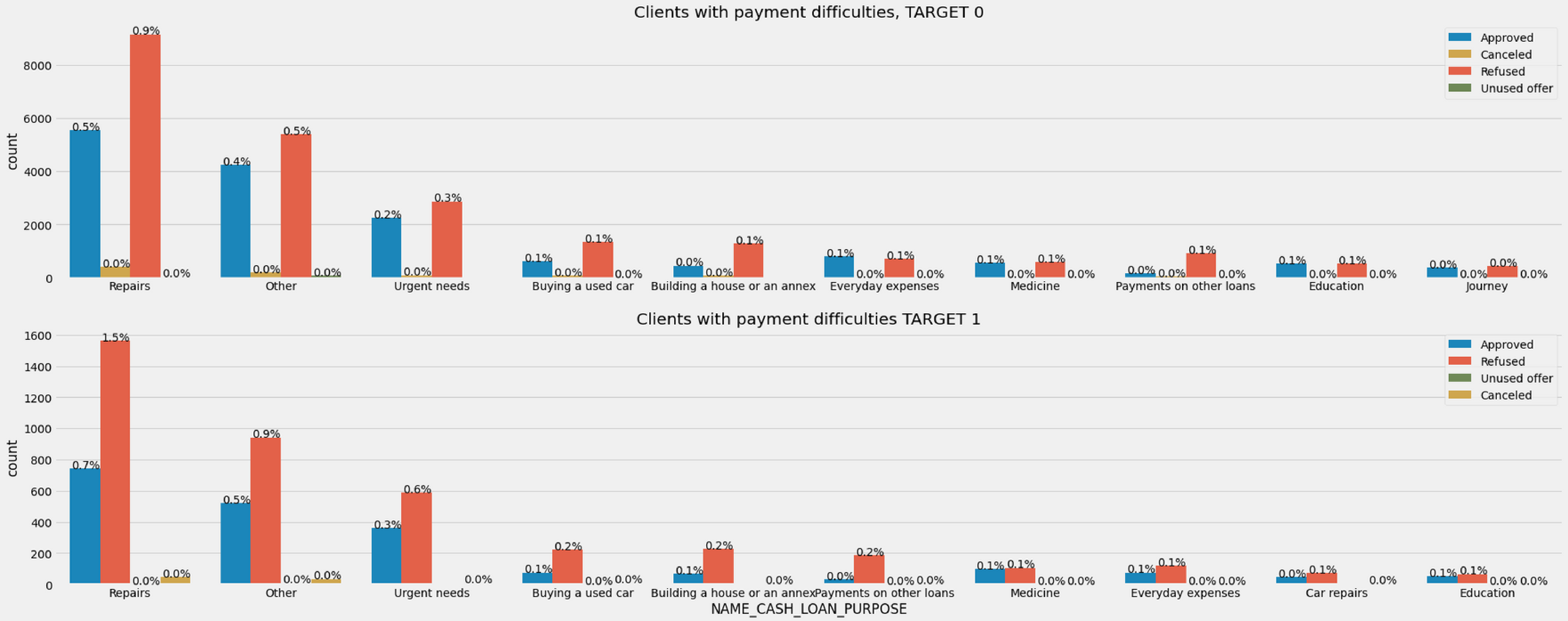
OCCUPATION TYPE



- Max payment difficulties, occur in Laborers and the rejection rates are higher for the ones with payment difficulties.

UNIVARIATE ANALYSIS

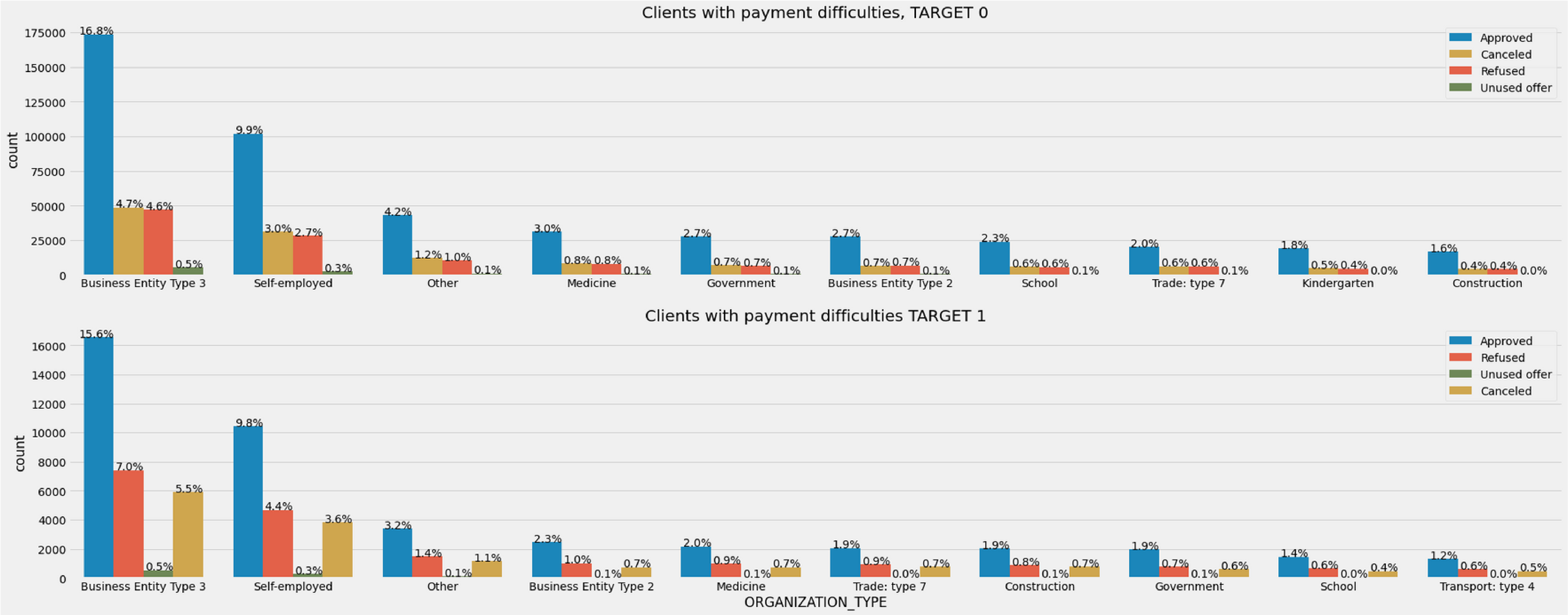
PURPOSE OF LOAN



- Repairs, others and loans with urgent needs see more rejections than approvals. Safe to saw that one should be very careful while approving as the payment difficulties are also high.

UNIVARIATE ANALYSIS

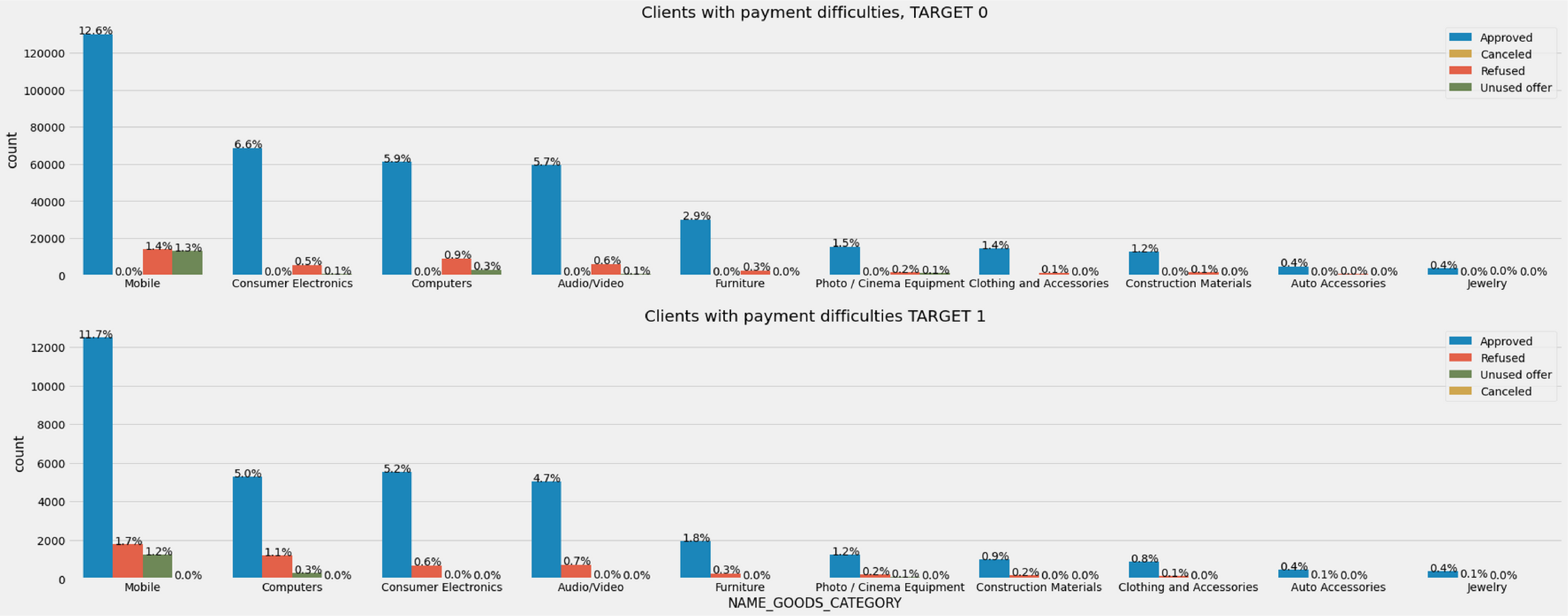
TYPE OF ORGANISATION



- Maximum payment difficulties are faced by Business Entity Type 3 and Self Employed and their rejections rates have also been high.

UNIVARIATE ANALYSIS

TYPE OF GOODS



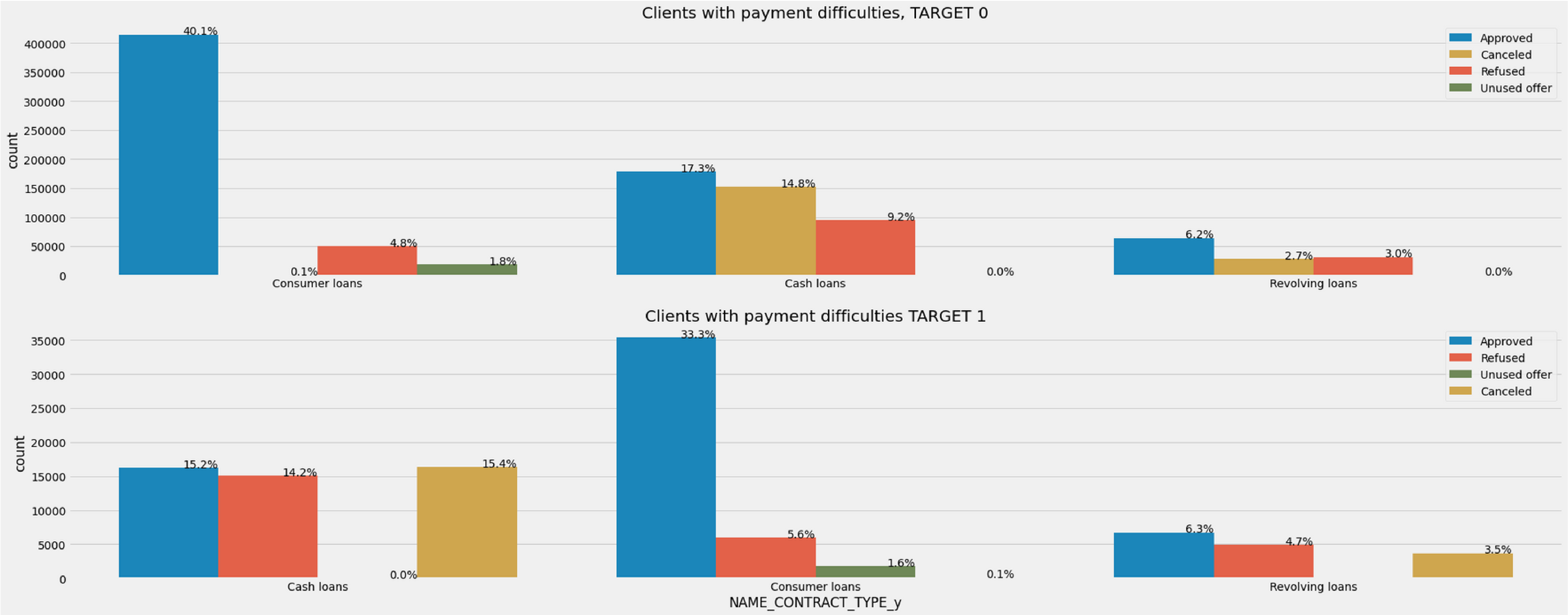
Clients with payment difficulties, TARGET 0

Clients with payment difficulties TARGET 1

- Most loans were procured for Mobile Phones and Cosumer Electronics.
- Payment defaults are more for Mobile phones compared to other entities irrespective of the target group.
- We hardly see any cancellations.

UNIVARIATE ANALYSIS

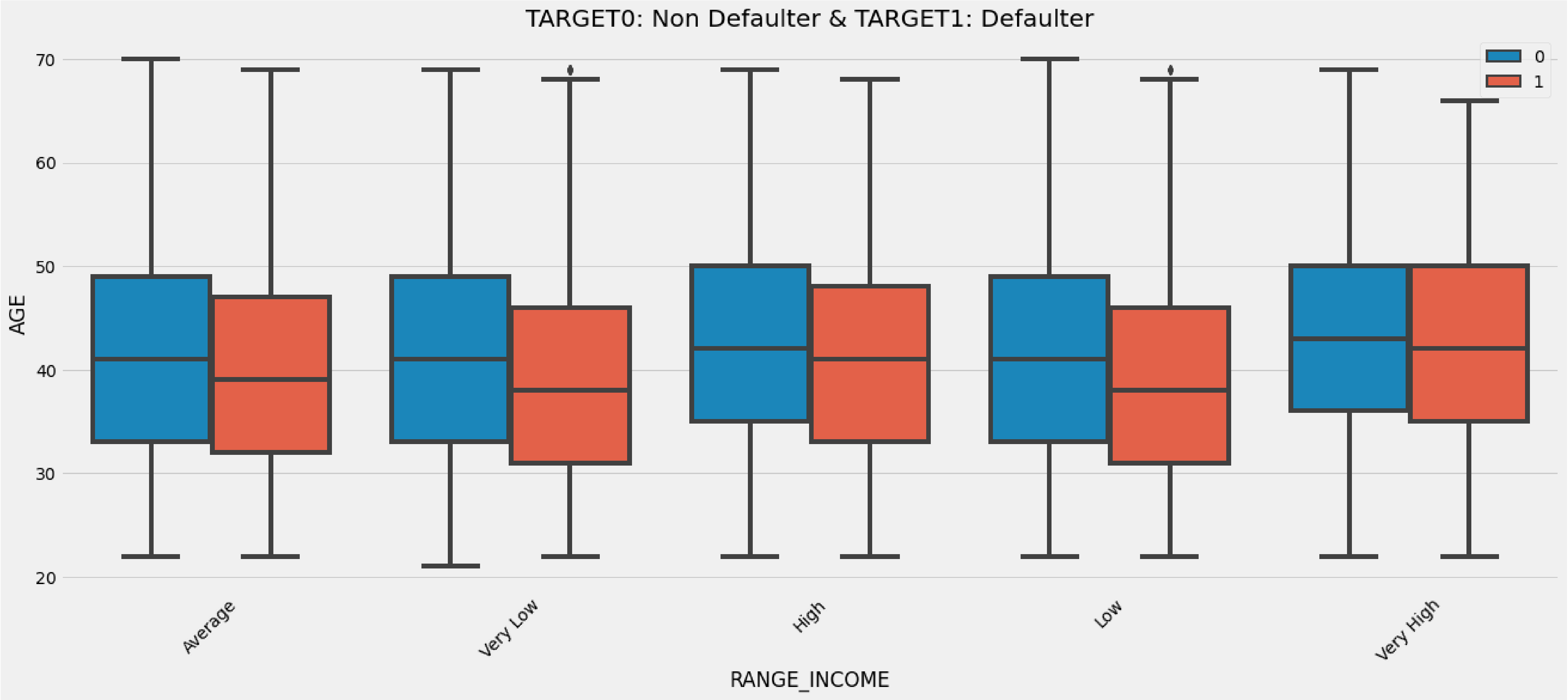
TYPE OF CONTRACT



- Cash loans see highest rejection rate.
- For cash loans, the rejections might be mainly due to higher payment difficulties.
- For cash loans with payment difficulties, we see that the rejection and approval rates are quite similar.

BI/MULTI-VARIATE ANALYSIS

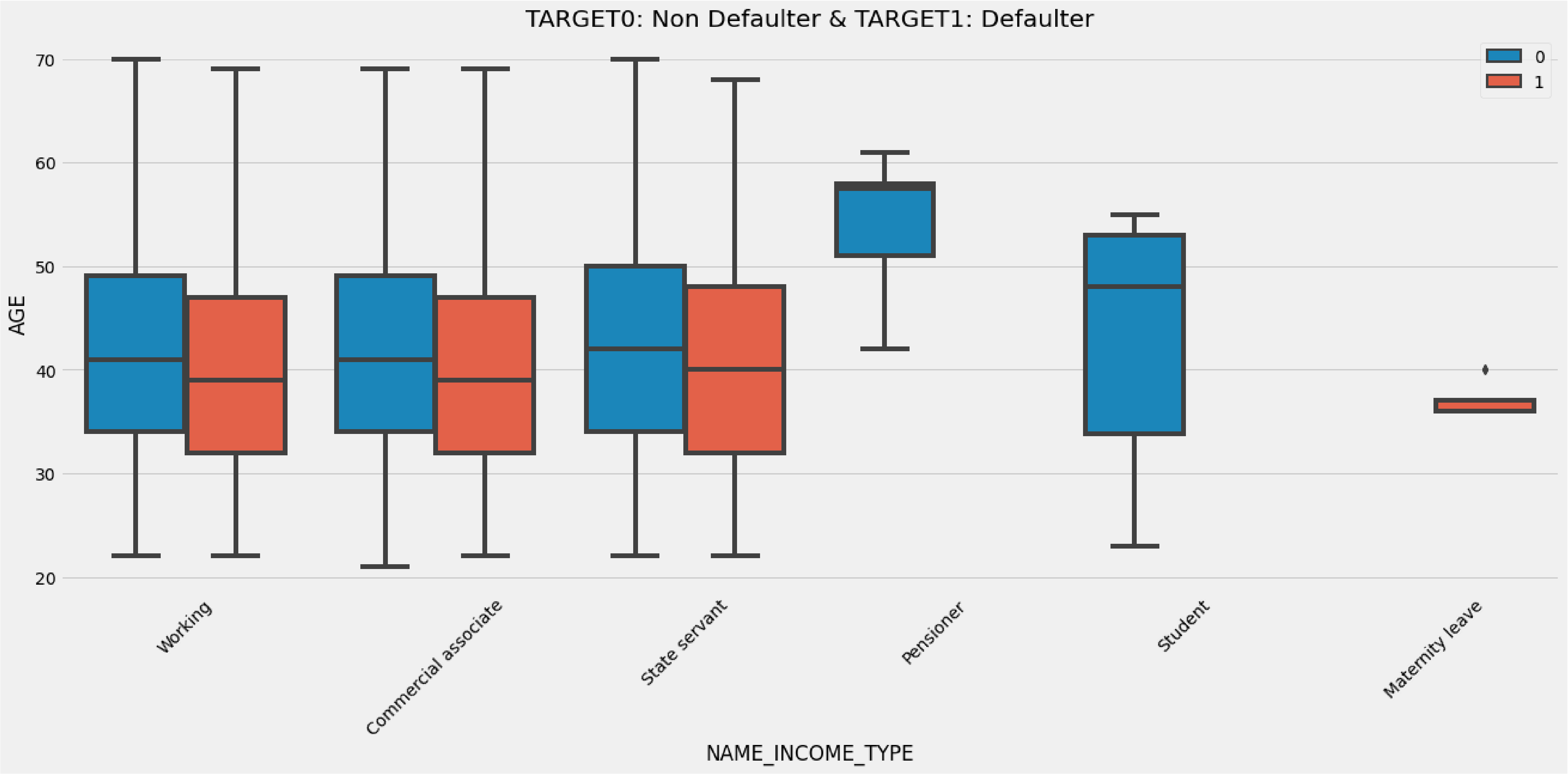
AGE VS INCOME RANGE



- The median age is between 40-45.
- For very high and high income groups we see that the median age is above 40 for both targets.
- For average, low and very low income groups below the age of 40, the payment difficulties seem to be prevailant and one must be cautious while approving loans to them .

BI/MULTI-VARIATE ANALYSIS

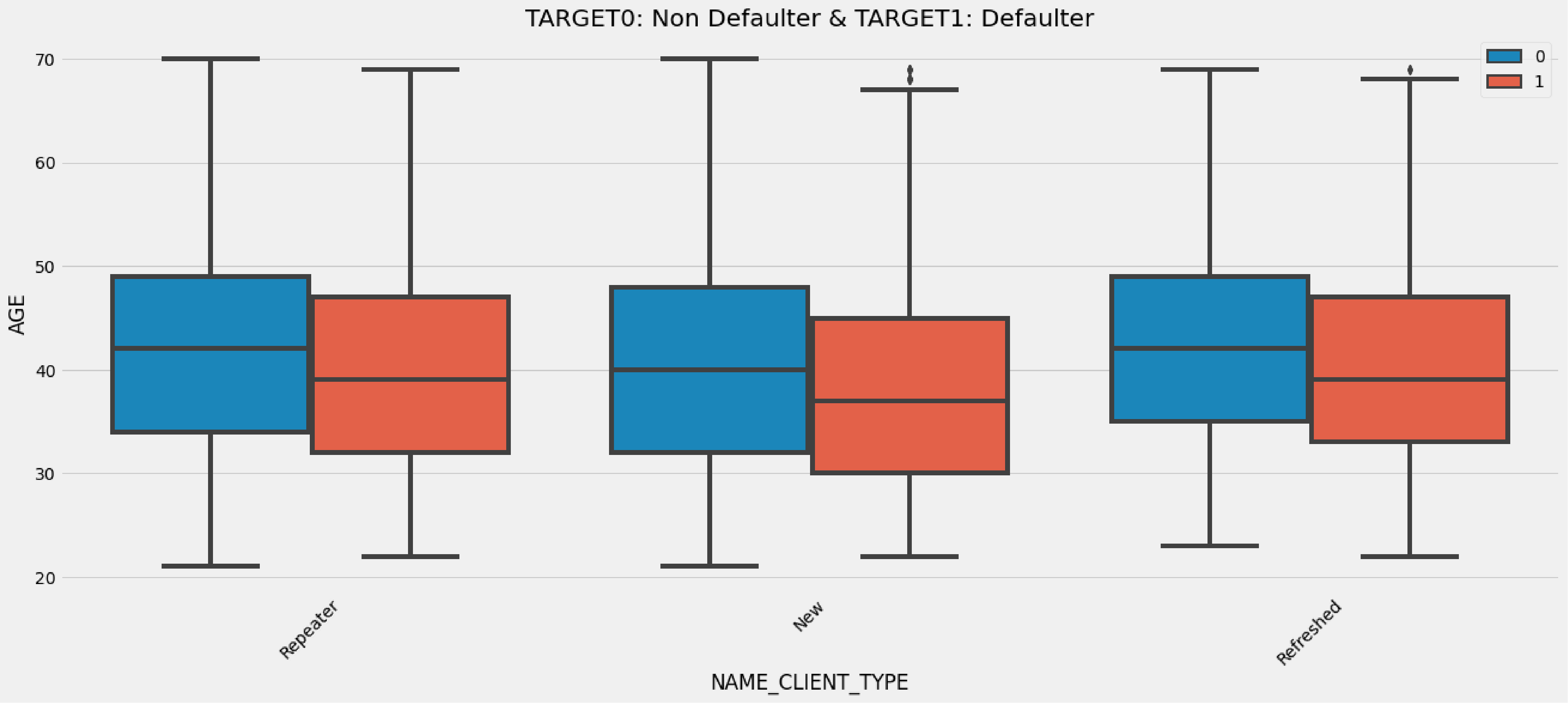
AGE VS INCOME TYPE



- Students and Pensioners do not show any payment difficulty.
- The working class, commercial associates below the age of 40 seem to have defaulted/had payment difficulties.
- Maternity leave group seem to have defaulted.

BI/MULTI-VARIATE ANALYSIS

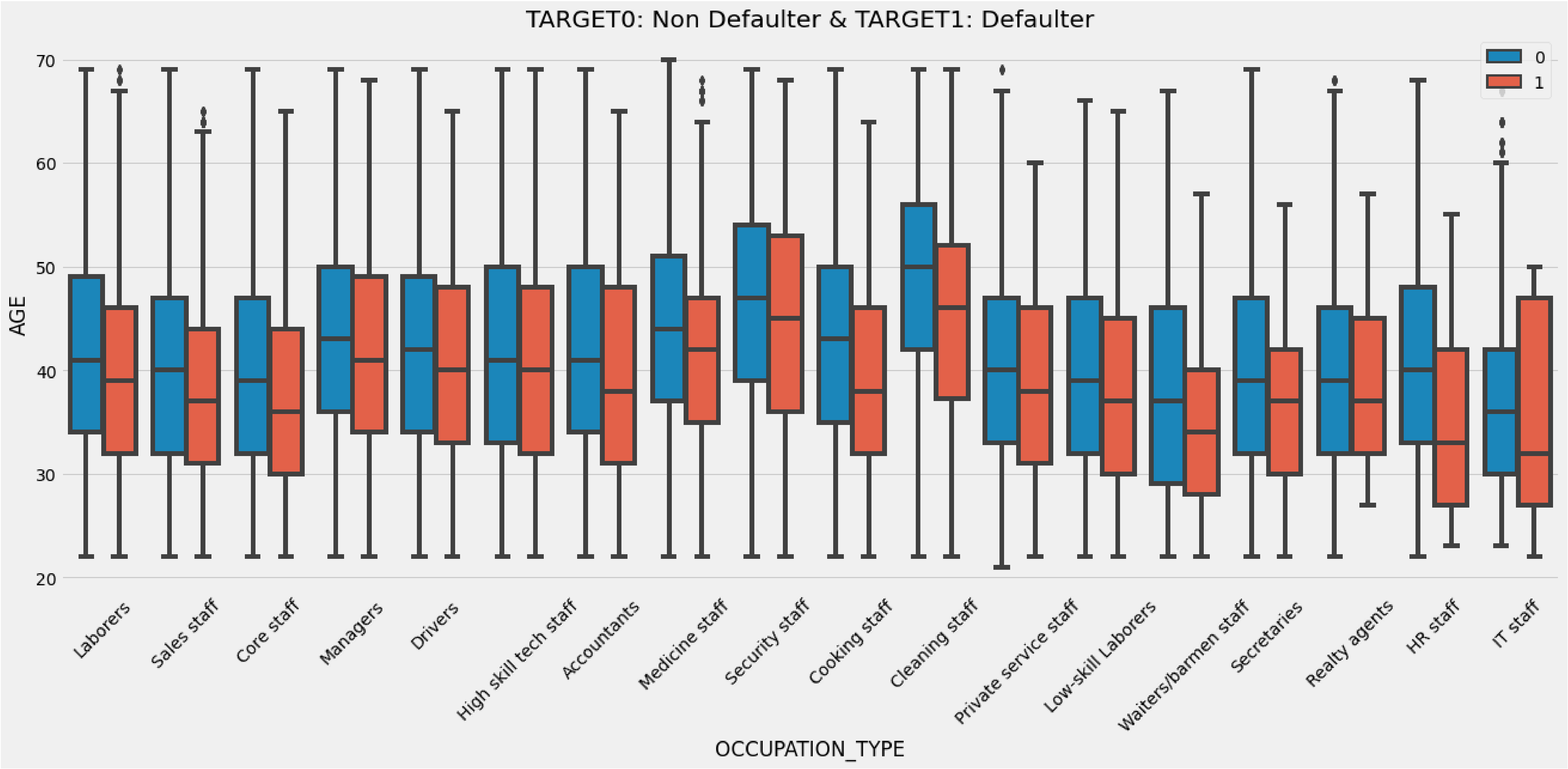
AGE VS CLIENT TYPE



- The median age for New Clients is lower compared to Repeater and Fresh clients.
- New clients who are younger, between 30-40 years of age seem to have defaulted.
- New and refreshed clients 40 and above have lesser defaults.

BI/MULTI-VARIATE ANALYSIS

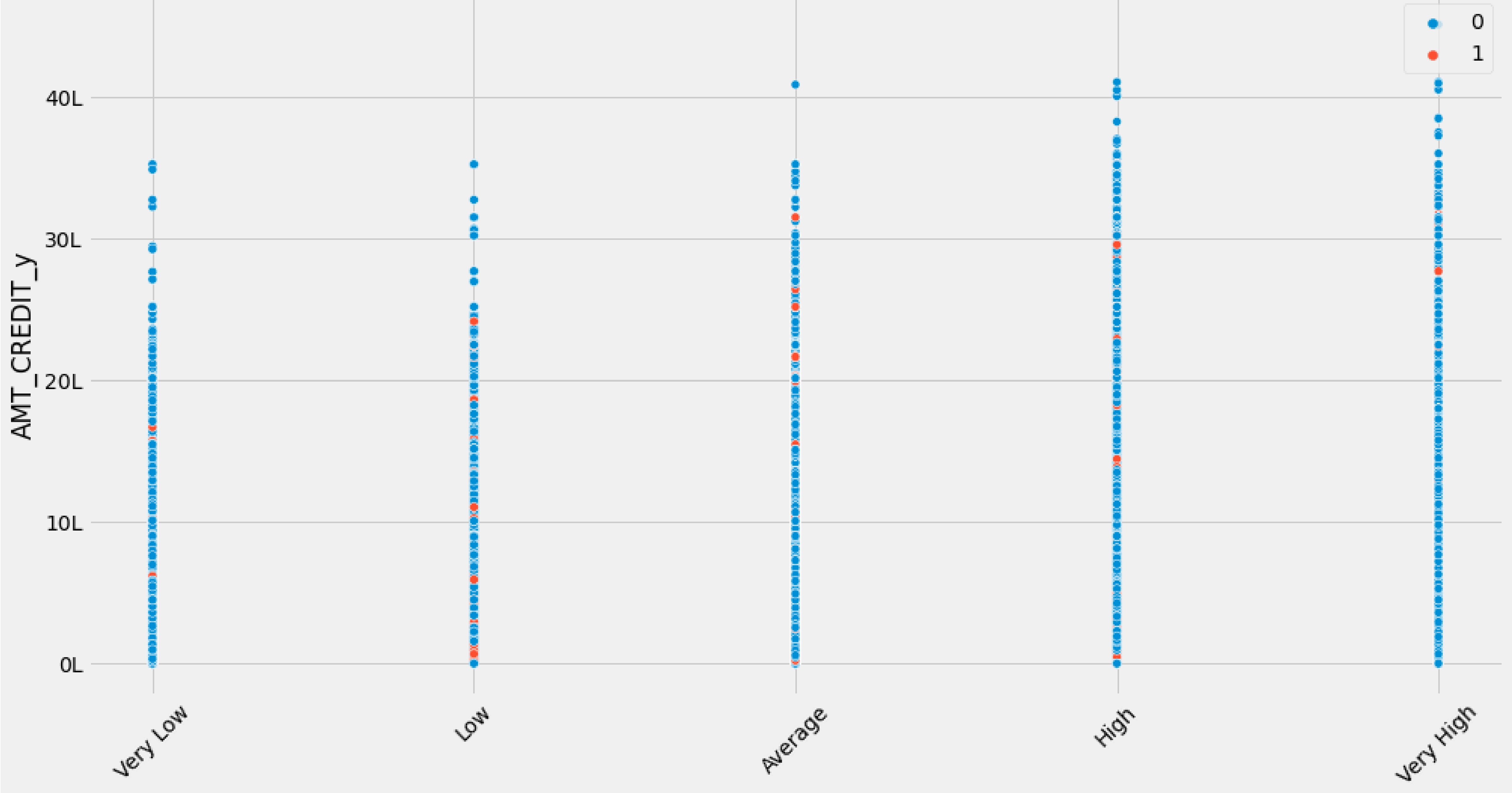
AGE VS OCCUPATION TYPE



- IT Staff, HR staff below the age of 35 are likely to default.

BI/MULTI-VARIATE ANALYSIS

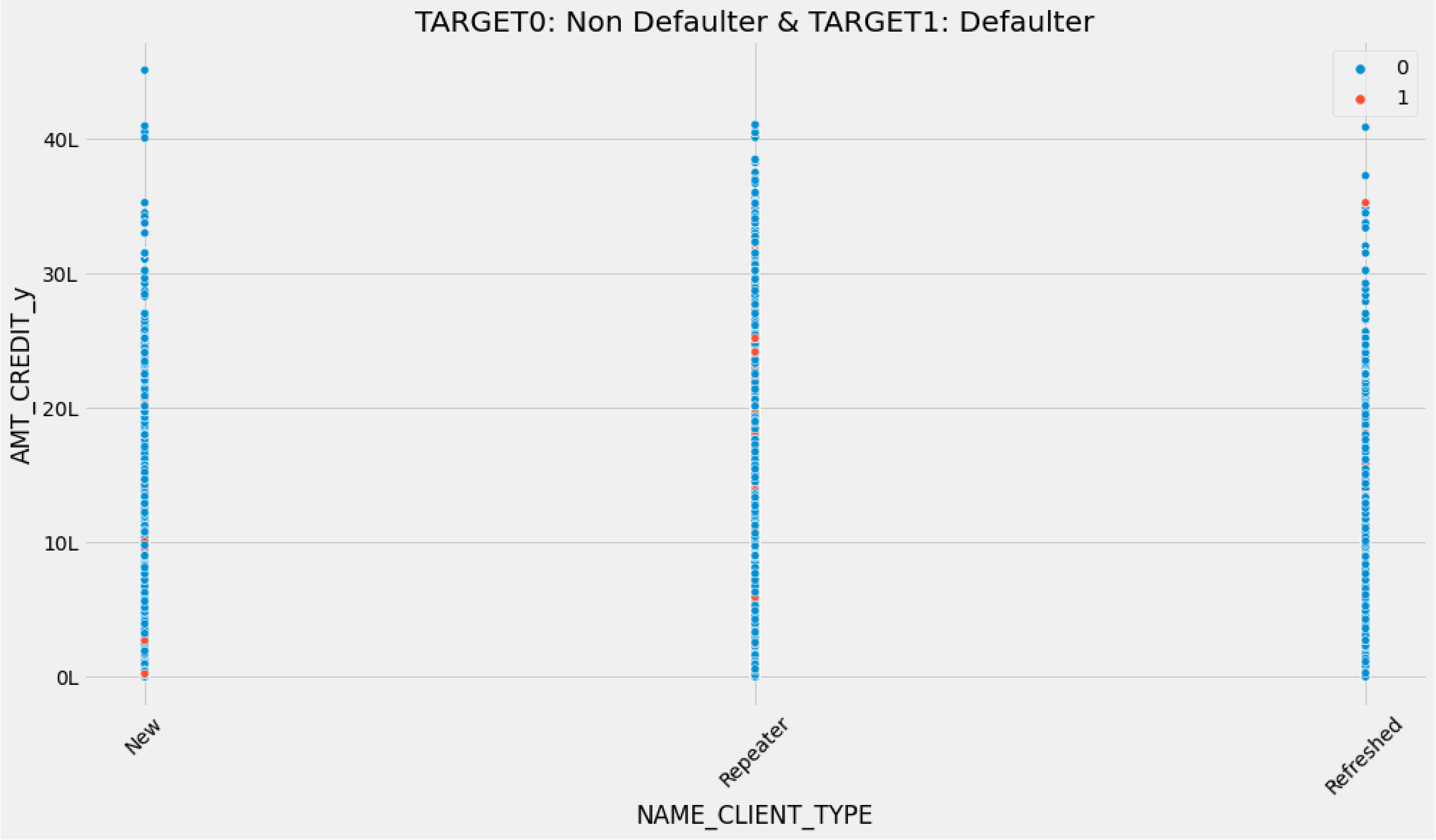
CREDIT_AMT VS INCOME GROUP



- Lower income groups with CREDIT_AMT 0-10 L seems to have max defaulters. One should be cautious while approving loans for this group in the same credit range.
- The average income group has payment difficulties between 20L-30L range.
- Very high and High income groups do not have much of a difficulty.

BI/MULTI-VARIATE ANALYSIS

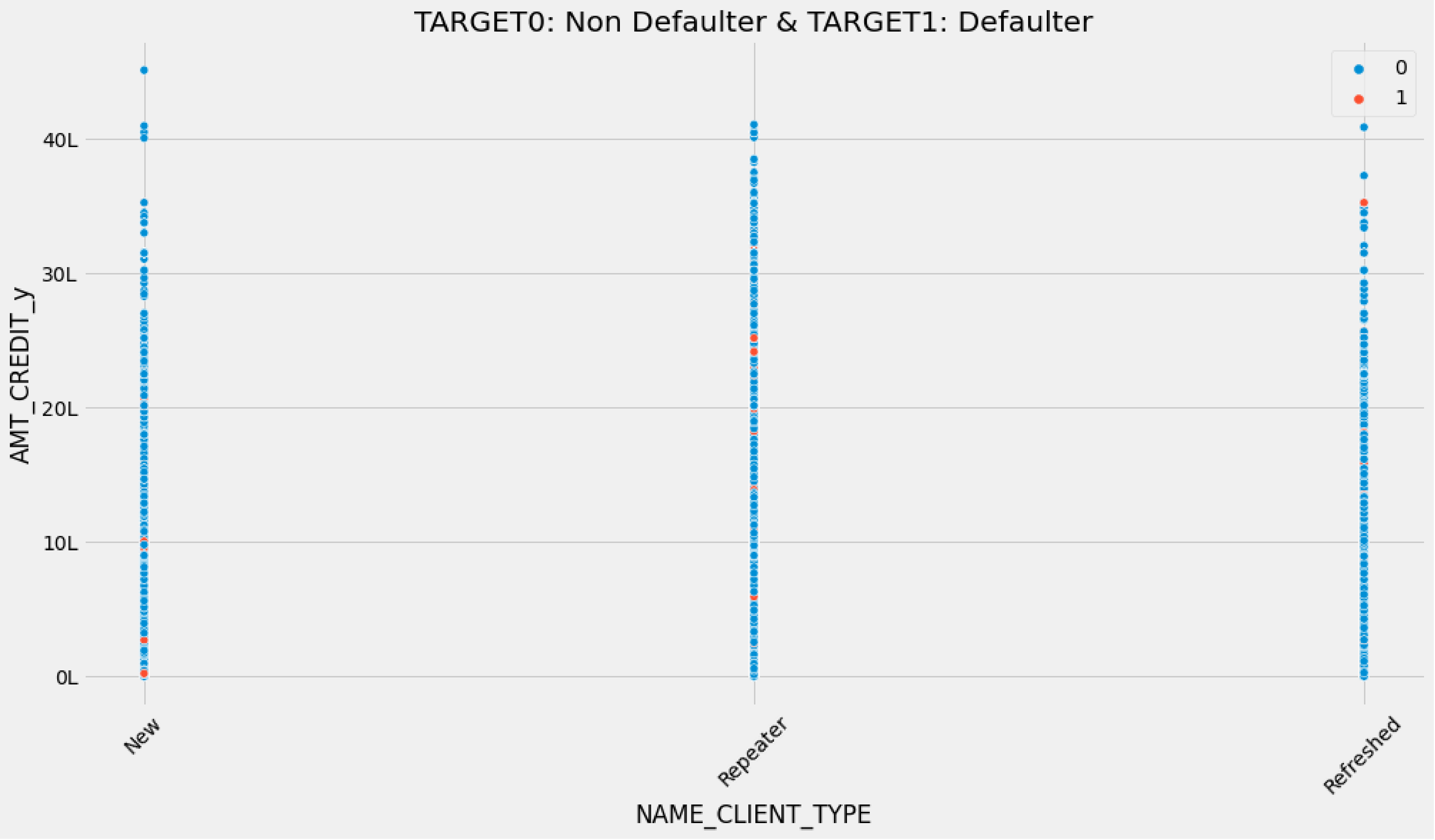
CREDIT_AMT VS INCOME TYPE



- Maternity Leave group are known to default. Better to avoid issuing loans to them.
- Students and Pensioners show no payment difficulties.

BI/MULTI-VARIATE ANALYSIS

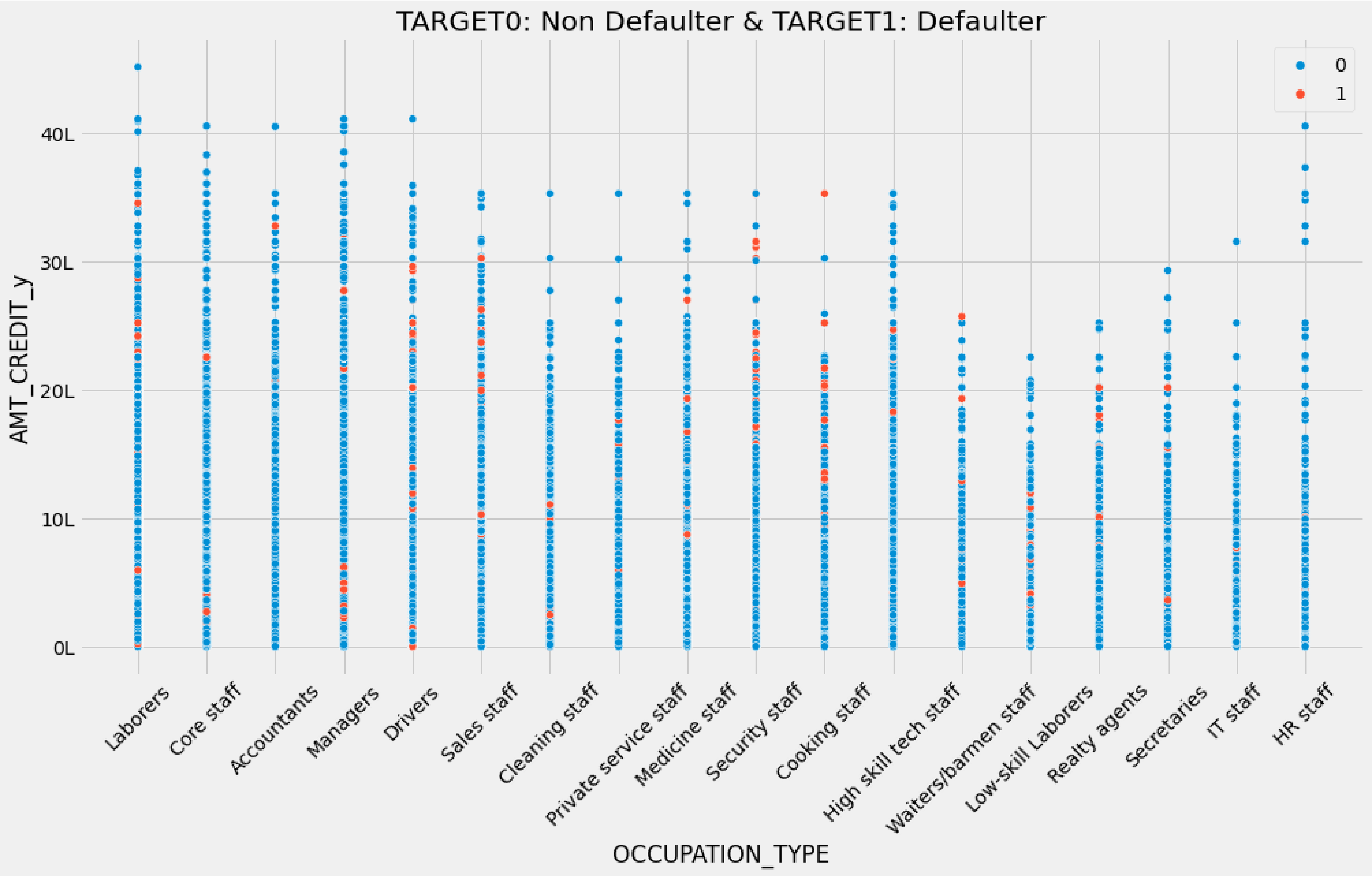
CREDIT_AMT VS CLIENT TYPE



- New clients seem to be defaulting in 0L-10L category.
- Repeater clients are defaulting more in the 20L-30L category.
- Refreshed clients hardly have any payment difficulties wrt CREDT_AMT.

BI/MULTI-VARIATE ANALYSIS

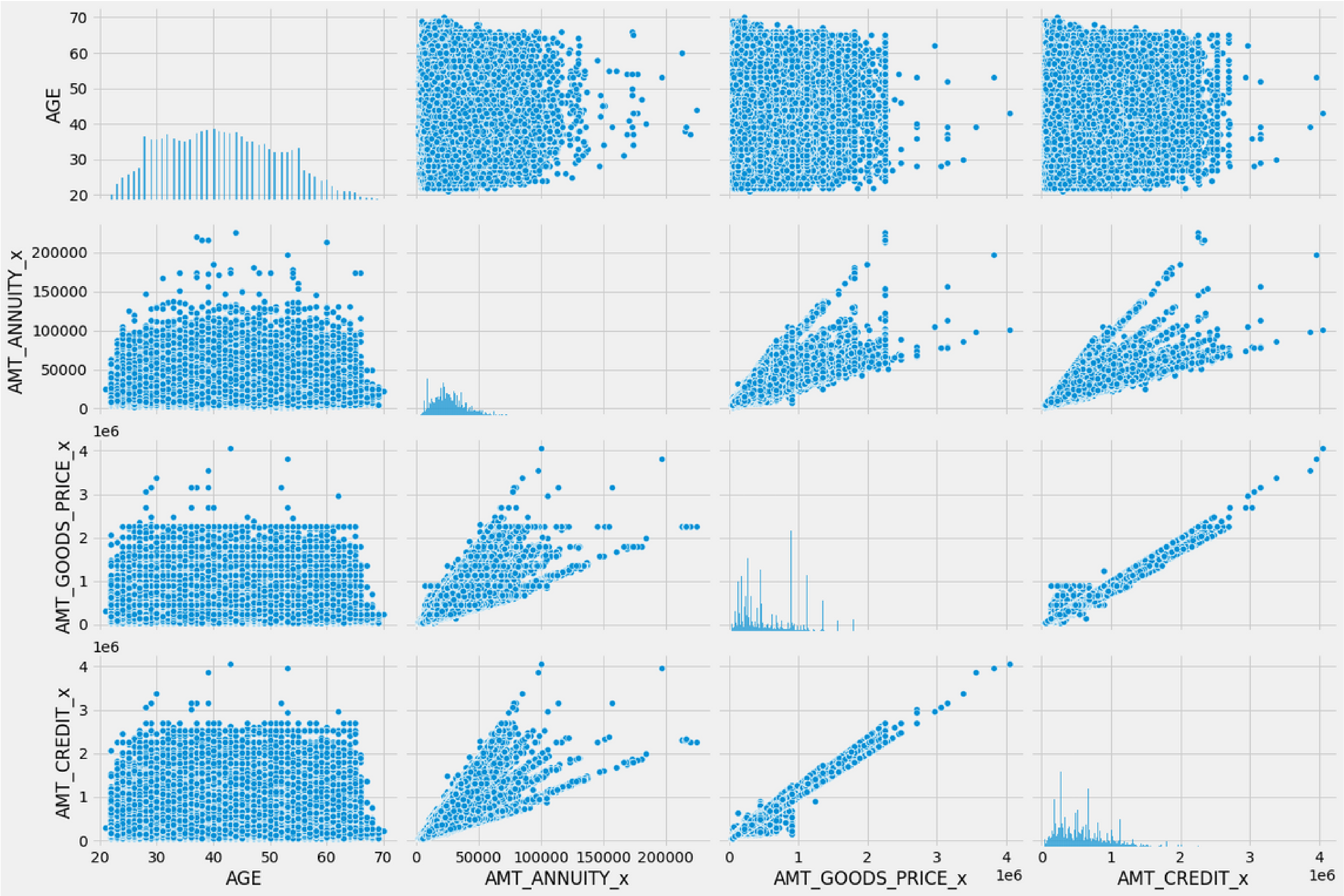
CREDIT_AMT VS OCCUPATION TYPE



- HR and IT staff do not have much of payment difficulties this maybe due the high income.
- Drivers, Sales Staff,Security Staff, Cooking Staff seem to have defaults in the 20L-30L category.
- In a general sense, payment difficulties are most for higher credit amounts.

BI/MULTI-VARIATE ANALYSIS

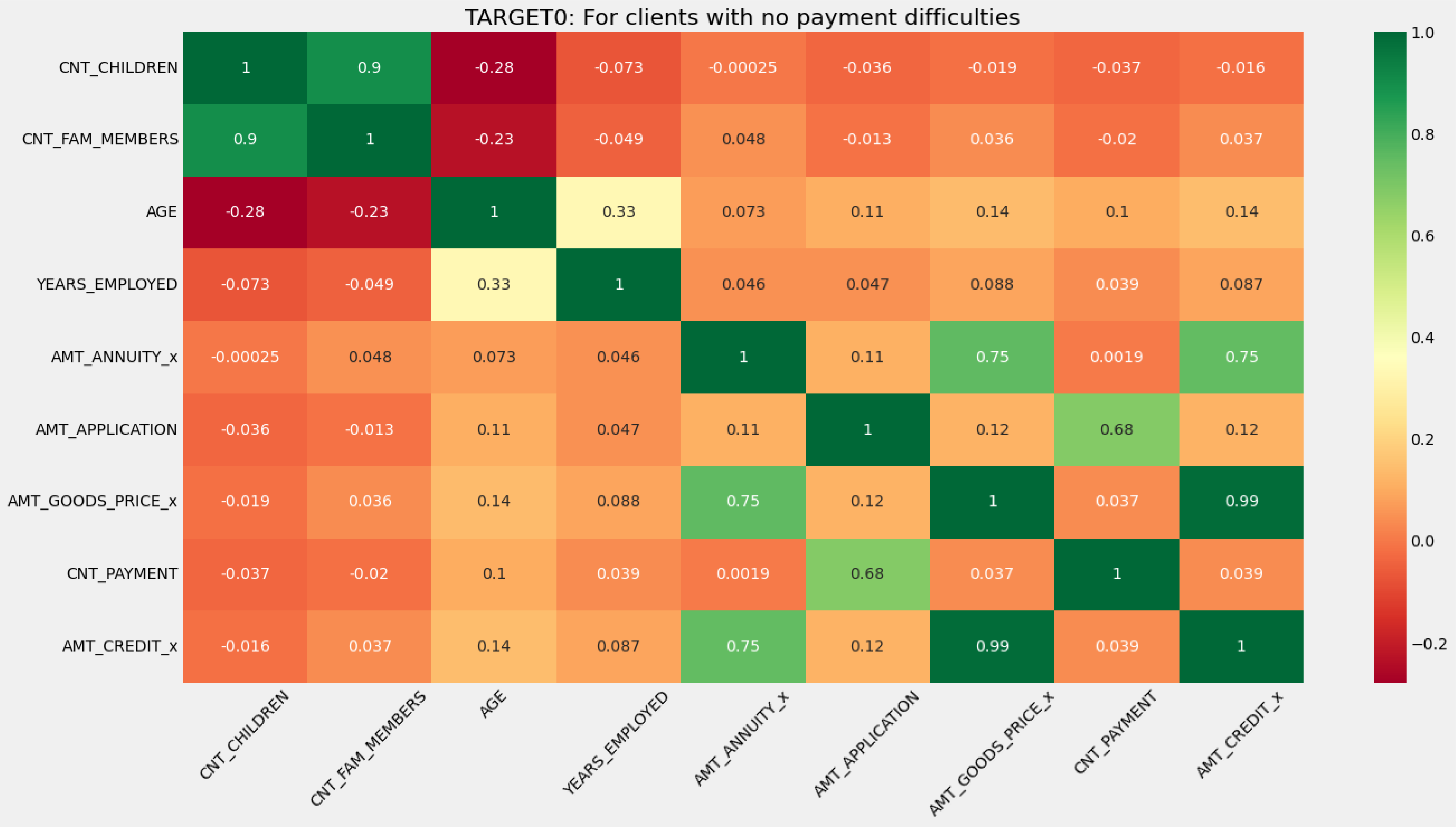
NUMERICAL COLUMNS ANALYSIS



- AMT_GOODS_PRICE vs AMT_CREDIT: As one increases, the other increases as well.
- AMT_GOODS_PRICE vs AMT_ANNUITY: As one increases, the other increases as well to an extent.
- AMT_CREDIT vs AMT_ANNUITY: As one increases, the other increases as well to an extent.
- AGE VS OTHERS: It is pretty much consolidated/dense apart from a few extreme values.

BI/MULTI-VARIATE ANALYSIS

CORRELATION HEATMAPS : TARGET0



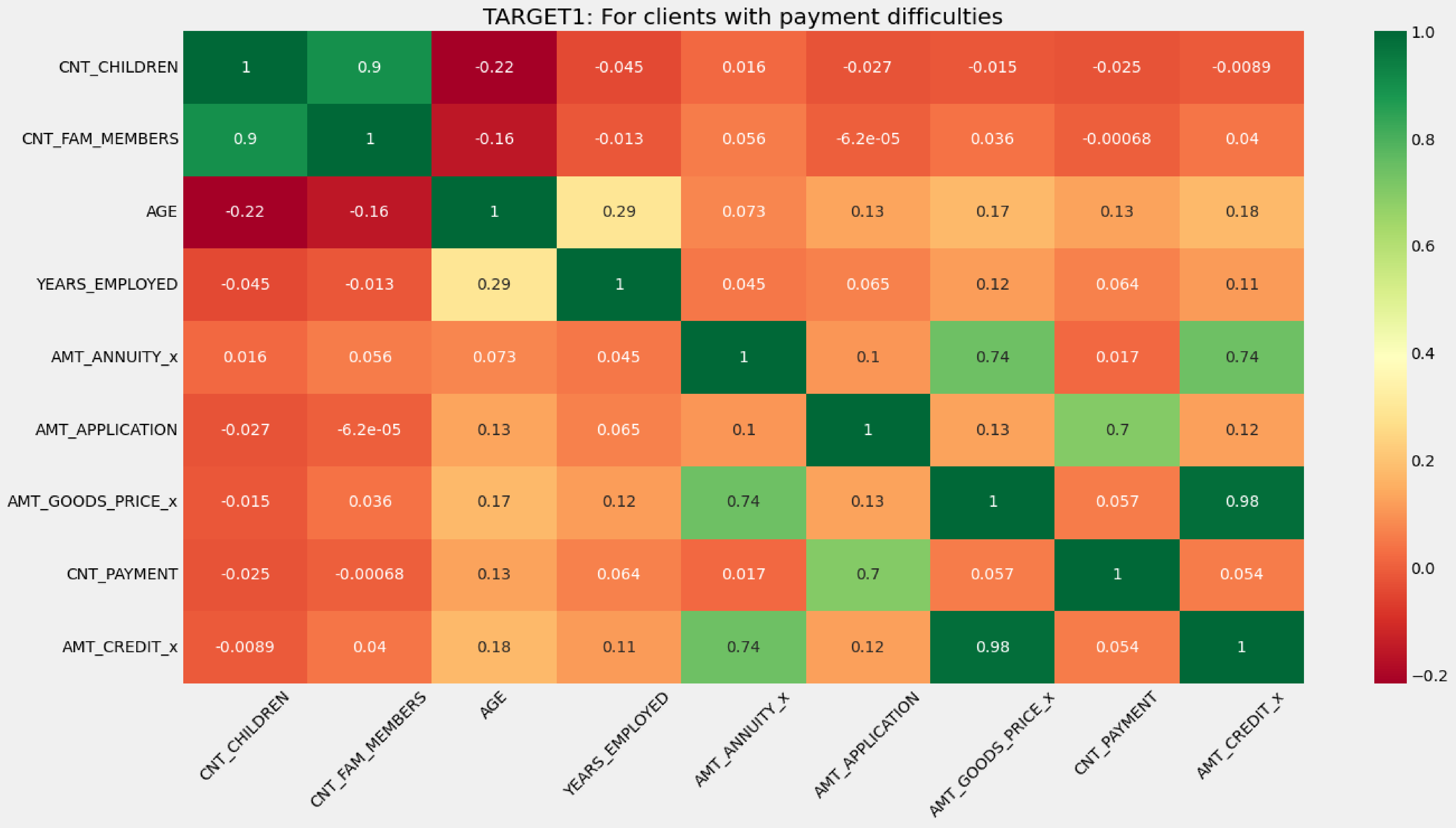
TARGET0: For clients with no payment difficulties

	CNT_CHILDREN	CNT_FAM_MEMBERS	AGE	YEARS_EMPLOYED	AMT_ANNUITY_x	AMT_APPLICATION	AMT_GOODS_PRICE_x	CNT_PAYMENT	AMT_CREDIT_x
CNT_CHILDREN	1	0.9	-0.28	-0.073	-0.00025	-0.036	-0.019	-0.037	-0.016
CNT_FAM_MEMBERS	0.9	1	-0.23	-0.049	0.048	-0.013	0.036	-0.02	0.037
AGE	-0.28	-0.23	1	0.33	0.073	0.11	0.14	0.1	0.14
YEARS_EMPLOYED	-0.073	-0.049	0.33	1	0.046	0.047	0.088	0.039	0.087
AMT_ANNUITY_x	-0.00025	0.048	0.073	0.046	1	0.11	0.75	0.0019	0.75
AMT_APPLICATION	-0.036	-0.013	0.11	0.047	0.11	1	0.12	0.68	0.12
AMT_GOODS_PRICE_x	-0.019	0.036	0.14	0.088	0.75	0.12	1	0.037	0.99
CNT_PAYMENT	-0.037	-0.02	0.1	0.039	0.0019	0.68	0.037	1	0.039
AMT_CREDIT_x	-0.016	0.037	0.14	0.087	0.75	0.12	0.99	0.039	1

- The highest correlation is between AMT_GOODS_PRICE of application data and AMT_CREDIT of application data : 0.99
- The least correlation is between AGE and CNT_CHILDREN : -0.28

BI/MULTI-VARIATE ANALYSIS

CORRELATION HEATMAPS :TARGET1



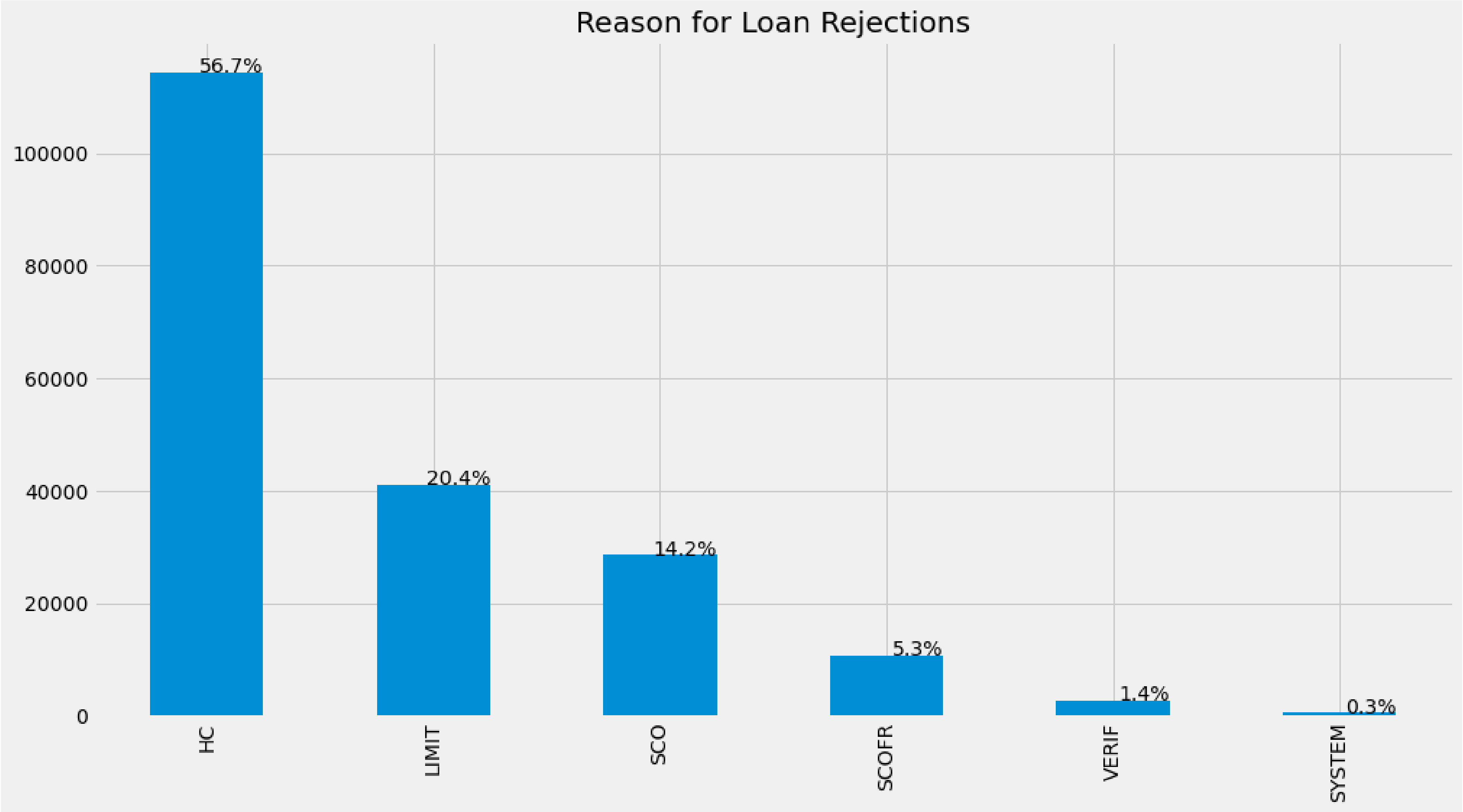
TARGET1: For clients with payment difficulties

	CNT_CHILDREN	CNT_FAM_MEMBERS	AGE	YEARS_EMPLOYED	AMT_ANNUITY_x	AMT_APPLICATION	AMT_GOODS_PRICE_x	CNT_PAYMENT	AMT_CREDIT_x
CNT_CHILDREN	1	0.9	-0.22	-0.045	0.016	-0.027	-0.015	-0.025	-0.0089
CNT_FAM_MEMBERS	0.9	1	-0.16	-0.013	0.056	-6.2e-05	0.036	-0.00068	0.04
AGE	-0.22	-0.16	1	0.29	0.073	0.13	0.17	0.13	0.18
YEARS_EMPLOYED	-0.045	-0.013	0.29	1	0.045	0.065	0.12	0.064	0.11
AMT_ANNUITY_x	0.016	0.056	0.073	0.045	1	0.1	0.74	0.017	0.74
AMT_APPLICATION	-0.027	-6.2e-05	0.13	0.065	0.1	1	0.13	0.7	0.12
AMT_GOODS_PRICE_x	-0.015	0.036	0.17	0.12	0.74	0.13	1	0.057	0.98
CNT_PAYMENT	-0.025	-0.00068	0.13	0.064	0.017	0.7	0.057	1	0.054
AMT_CREDIT_x	-0.0089	0.04	0.18	0.11	0.74	0.12	0.98	0.054	1

- The heat-maps of TARGET0 and TARGET1 seem quite similar.
- The highest correlation is between AMT_GOODS_PRICE of application data and AMT_CREDIT of application data : 0.98
- The least correlation is between AGE and CNT_CHILDREN : -0.22

BI/MULTI-VARIATE ANALYSIS

REASONS FOR LOAN REJECTION



- HC: Assuming that this stands for high credit, we see that most of the applications are being rejected due to HC requirements.
- LIMIT: This maybe a case where the client required higher amount as loan but the bank could not provide that amount due to various reasons.

CONCLUSION

TARGET0 - Clients with payment difficulties TARGET1- Clients without payment difficulties.

- About 91% of the clients do not experience any payment difficulty while about 9% of clients have defaulted.Imbalance Ratio - 91:9
- Most of the applications get approved(62.8%) while rejection rate and unused offers stand at 17.7%. A large number of offers are unused. Unused offers can be reduced based on the analysis provided.
- **Clients without payment difficulties are more likely to get their loans approved**. Also, their refusal percentage is lower compared to clients with payment difficulties. Hence, additional checks and filters need to be put in for approving applications with payment difficulties.

Owning a Car

- Most clients do not own a car. Only 38% of them do. However, this cannot be a factor for approving/rejecting loans.

Owning Realty

- Most clients own a house/apartment and again this does not have an impact on the approval or rejection.

Income

- Rejection Rate for clients among the average income categories seem the highest (5.1% for TARGET0 and 7.7% for TARGET1). **Additional caution to be taken while approving loans for TARGET1 in the average income range.** Very High and High income ranges have seen lesser rejections.

CONCLUSION CONTINUED

Income Type

- **State Servants seem to have lesser payment difficulties.** The rejection rate for working class with payment difficulties seem very high. For clients belonging to this category, it is better to be cautious while approving loans.
- Unused offers are highest among working class.

Marital Status

- Married clients with payment difficulties have lesser approval rates compared to the ones without payment difficulties.

Housing Type

- Clients who do not stay in House/Apartments have lesser rejection rates. Rejection rates highest for clients who stay in apartments. One must be cautious while approving loans to clients who stay in apartments with payment difficulties.

Client Type

- Among repeater clients, they face the most rejections. Refreshed and new clients do not have much of rejections. **Approach with caution while approving loans to repeater clients.**

Occupation Type

- **Maximum payment rejections, occur in Labourers** and the rejection rates are higher for the ones with payment difficulties.

CONCLUSION CONTINUED

Loan Purpose

- **Repairs, others and loans with urgent needs see more rejections than approvals.** Safe to say that one should be very careful while approving as the payment difficulties are also high.

Organisation Type

- Business Entity Type 3 and Self Employed have their rejections rates higher than most.

Goods Category

- Most loans were procured for Mobile Phones and Consumer Electronics. Rejections are more for Mobile phones compared to other entities irrespective of the target group.We hardly see any cancellations.

Contract Type:

- **Cash loans see highest rejection rate.**
- For cash loans, the rejections might be mainly due to higher payment difficulties.
- For cash loans with payment difficulties, we see that the rejection and approval rates are quite similar.

Client Type

- Very less proportion of New and Refreshed clients get rejected.
- Defaults/refusals are high for repeater clients with payment difficulties. The percentage of unused loans is also high. One must be cautious while approving loans to repeater clients.

CONCLUSION CONTINUED

AGE

- The median age is between 40-45.
- For very high and high income groups we see that the median age is above 40.
- **For average, low and very low income groups below the age of 40, the payment difficulties seem to be prevalent and one must be cautious while approving loans to them** .
- **Students and Pensioners do not show any payment difficulty**. They can be given loans.
- The working class, commercial associates below the age of 40 seem to have defaulted/had payment difficulties.
- **Maternity leave group seem to have defaulted. Better to avoid giving loans to them.**
- New clients who are younger, between 30-40 years of age seem to have defaulted.
- New, refreshed and repeater clients above the age of 40 have lesser defaults.
- Most Occupation types, below the age of 40 seem to defaulted.
- **IT Staff, HR staff below the age of 35 are likely to default. Rest of them seem to pay on time.**

CONCLUSION CONTINUED

CREDIT AMOUNT

- **Lower income groups with CREDIT_AMT 0-10 L seems to have max defaulters.** One should be cautious while approving loans for this group in the same credit range.
- The average income group has payment difficulties between 20L-30L range.
- Very high and High income groups do not have much of a difficulty.
- New clients seem to be defaulting in 0L-10L category.
- **Repeater clients are defaulting more in the 20L-30L category.**
- Refreshed clients hardly have any payment difficulties wrt CREDT_AMT.
- **HR and IT staff do not have payment difficulties this maybe due the high income.**
- Drivers, Sales Staff,Security Staff, Cooking Staff seem to have defaults in the 20L-30L category.
- In a general sense, payment difficulties are most for higher credit amounts.

CONCLUSION CONTINUED

CORRELATIONS

- The highest correlation is between **AMT_GOODS_PRICE of application data and AMT_CREDIT of application data 0.99 for TARGET0.**
- The least correlation is between AGE and CNT_CHILDREN : -0.28
- The heat maps of TARGET0 and TARGET1 seem quite similar.
- **The highest correlation is between AMT_GOODS_PRICE of application data and AMT_CREDIT of application data : 0.98 for TARGET1.**
- The least correlation is between AGE and CNT_CHILDREN : -0.22