upGrad

International Institute of Information Technology Bangalore

Credit EDA
Group Case Study

By

Regina Aboobacker

Pooja Pawani

Post Graduate Diploma in Data Science
Batch - C22 August 2020

PROBLEM STATEMENT

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. Using EDA analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

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.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

BUSINESS OBJECTIVE

This case study aims –
to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as

1. Denying the loan
2. Reducing the amount of loan
3. Lending (to risky applicants) at a higher interest rate, etc.

This will ensure that the consumers capable of repaying the loan are not rejected.
Identification of such applicants using EDA is the aim of this case study.

ANALYSIS APPROACH

For this case study we are using – EXPLORATORY DATA ANALYSIS.

Exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods

We will explore these Data sets and perform the exploratory data analysis. We have performed following steps of this approach and have addressed the problem statement:

- ❑ Descriptive Analysis
- ❑ Identified null values and Handled Missing value
- ❑ Removed duplicates
- ❑ Outlier Treatment
- ❑ Normalizing and Scaling(Numerical Variables)
- ❑ Encoding Categorical variables(Dummy Variables)
- ❑ Bivariate Analysis
- ❑ Multivariate Analysis

DATA UNDERSTANDING

This dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample

- **All other cases:** All other cases when the payment is paid on time.

DATA CLEANING

Before Cleaning

Application.csv -
No. of rows - 307511
No. of columns - 122

Previous_application.csv -
No. of rows - 1670214
No. of columns - 37

- **Dropped unwanted columns** - All normalised columns, Flag document columns, Car age column,
- **Handled Missing Values** – Columns having large number of null values are dropped and others are handled with mean/mode/median as required
- **Standardisation** –Required flag columns like Columns like flag_own_car, flag_own_realty, days are mapped to 1 and 0 from Y and N, Age(converting birth days to age)
- **Fixed Invalid values and data types** – eg - Occupation type, Gender, days_registration
- **Bucketing the numerical variables** - converted income amount to income group and age to age group for better analysis
- **Handled Outlier** – Income amount

After Cleaning

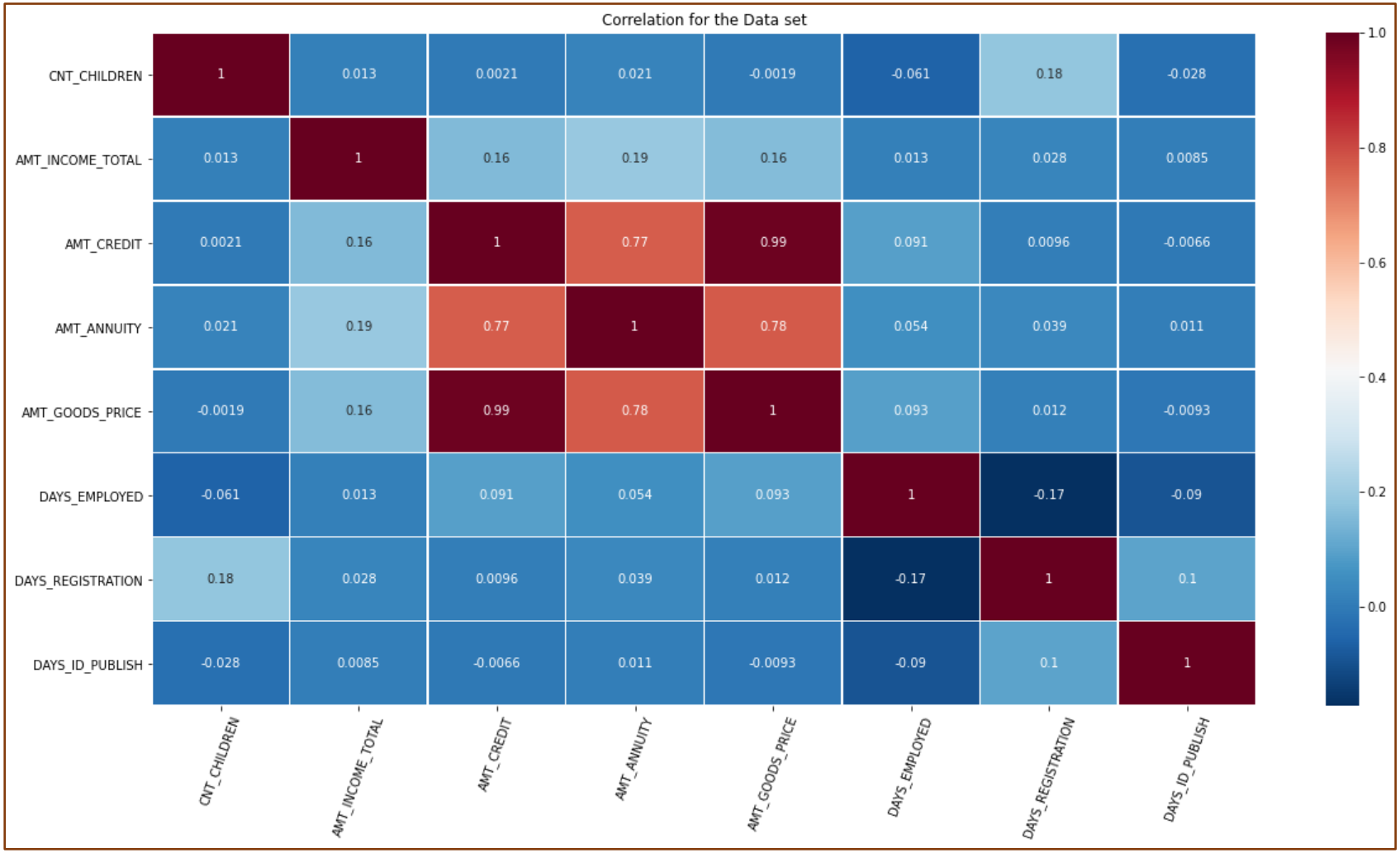
Application.csv -
No. of rows - 307510
No. of columns - 33

Previous_application.csv -
No. of rows - 1670214
No. of columns - 23

Merged data set -
No. of rows - 1413698
No. of columns - 55

Analysis of application.csv data

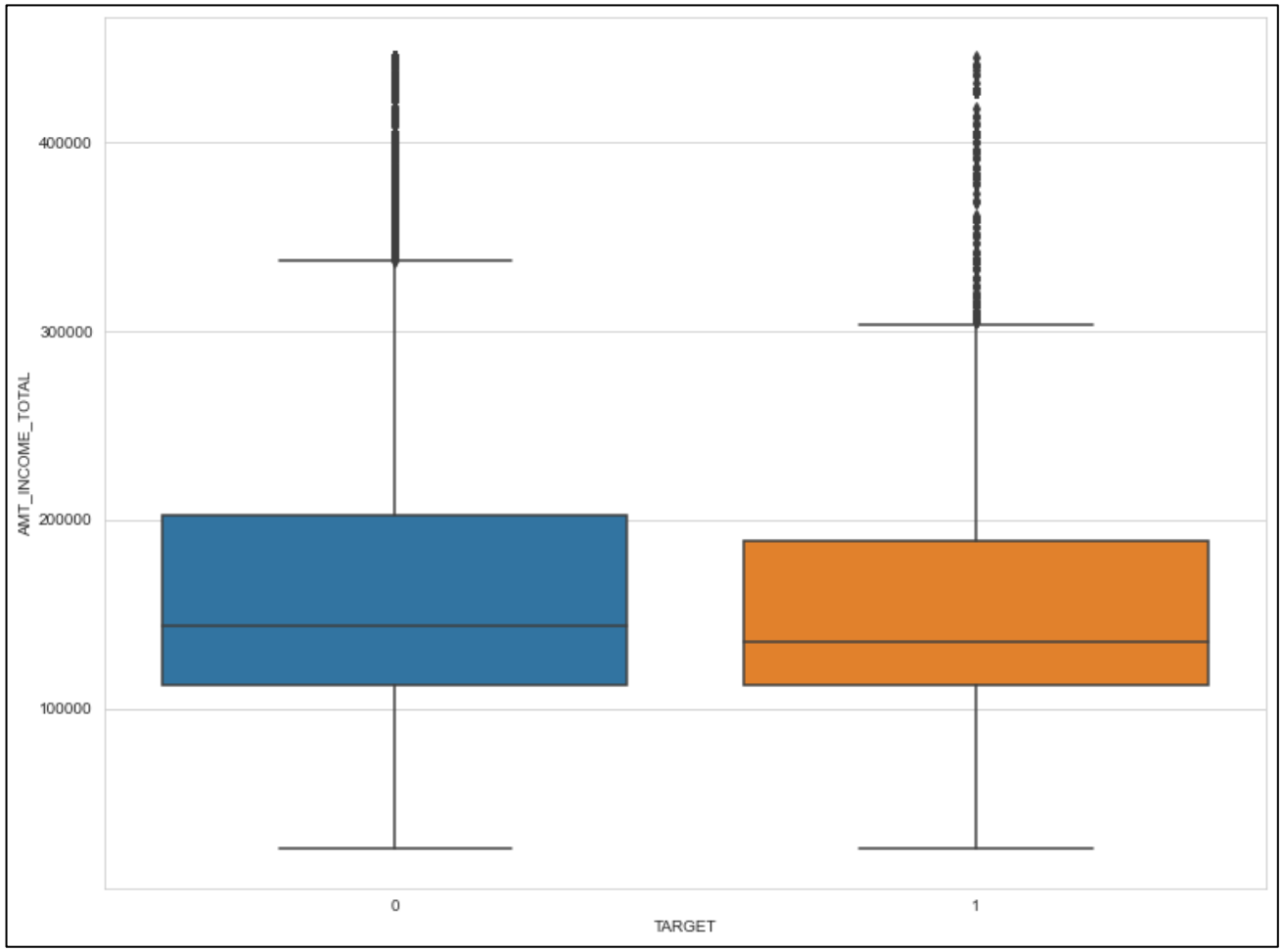
Correlation matrix



From the above matrix we can see AMT_CREDIT, AMT_GOODS_PRICE and AMT_ANNUITY are highly correlated

Univariate Analysis

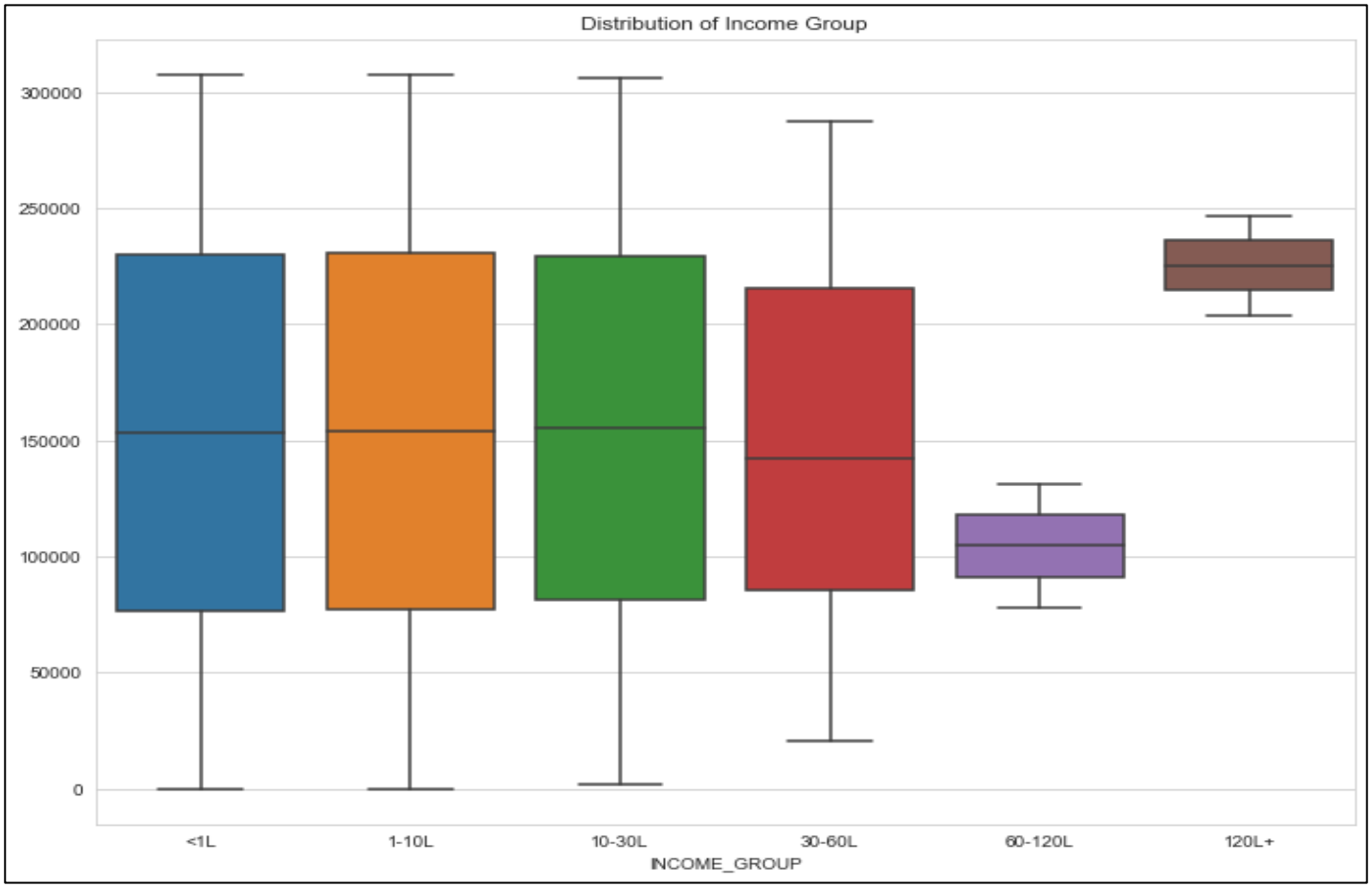
Distribution of Income for Target 0 and Target 1 group



Inference -

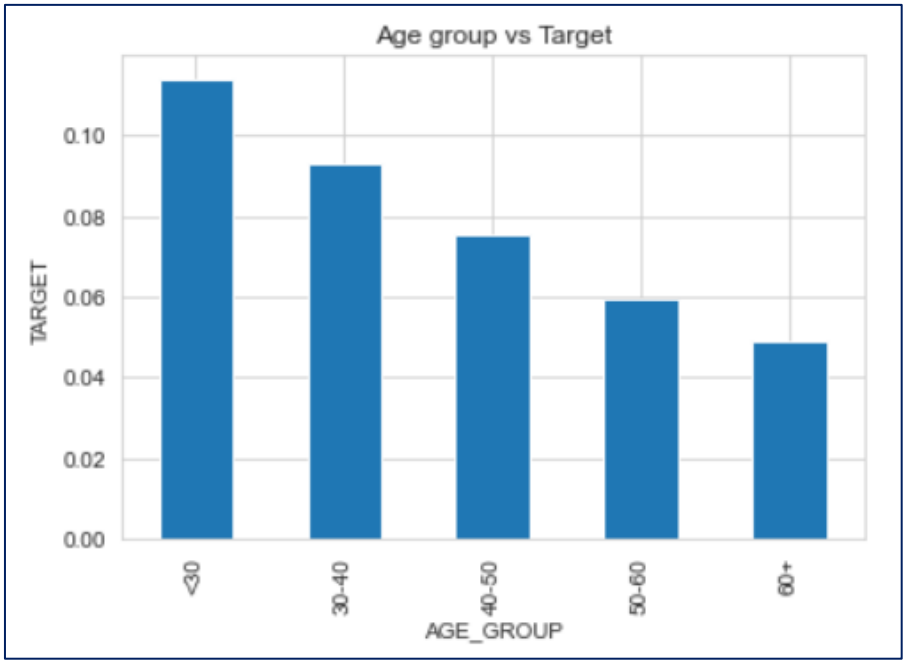
- Most of the clients are present above median , the third quartile is higher than the first quartile in both target 1 and target 0 case.
- There are some outliers

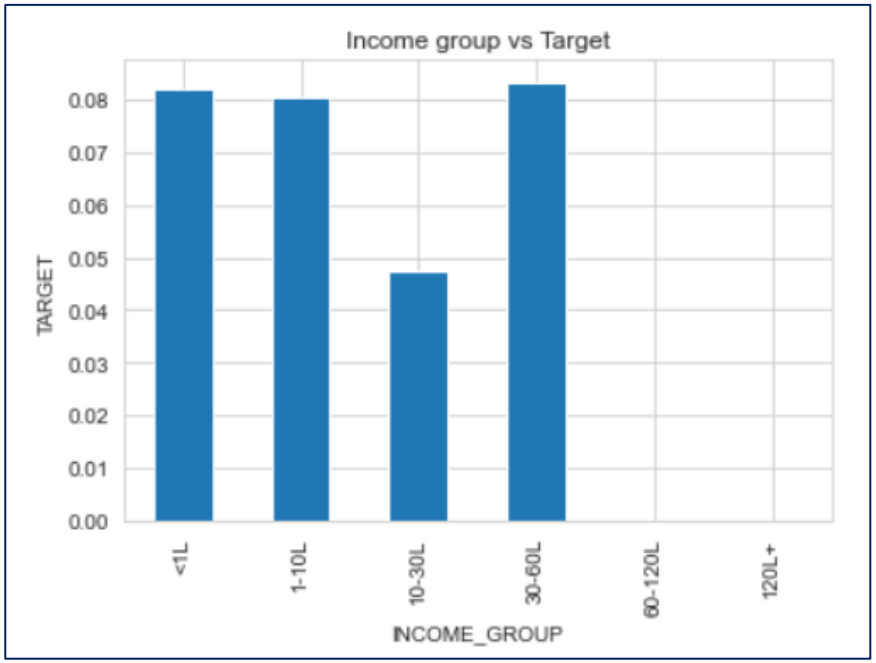
Cont.....



Inference - Majority of the applicants are upto the income group of 60L

Bivariate Analysis

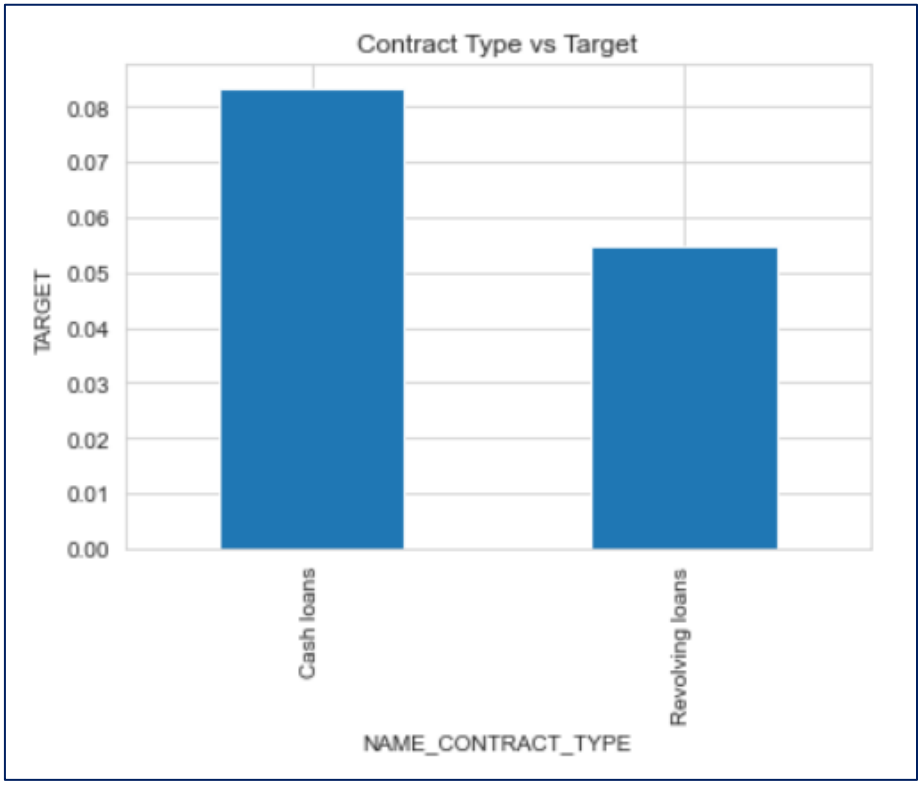




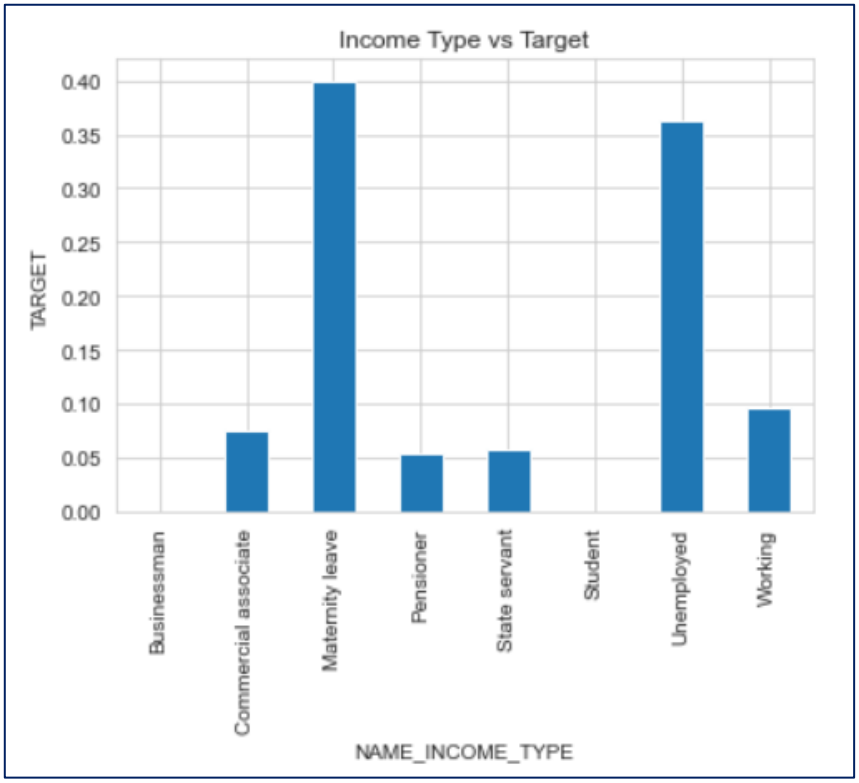
Inference: People with age group <30 seems to have more payment difficulties

Inference: People from Income group 60 L and above seems to be probable customers

Continued..

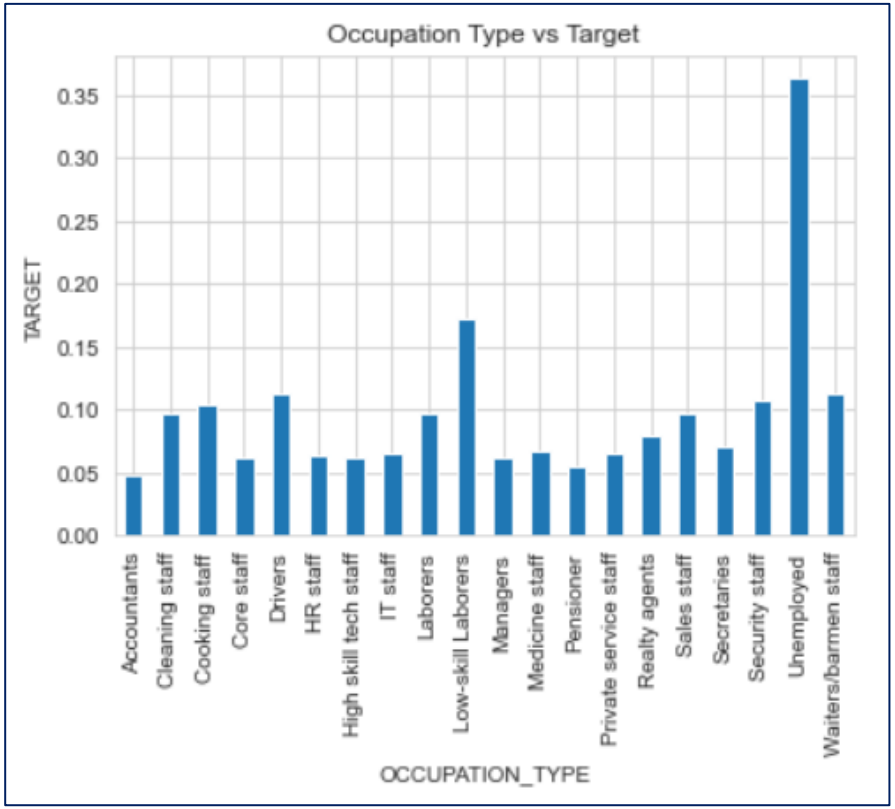


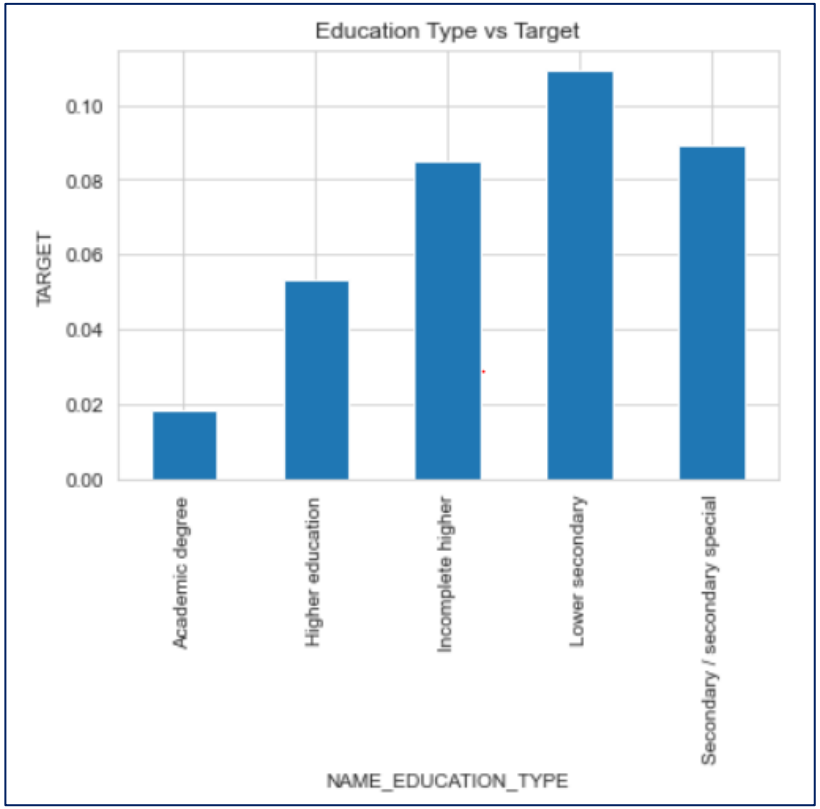
Inference: People who have taken cash loans seems to have more payment difficulties



Inference: Unemployed people or those on maternity leave seems to have payment difficulties

Continued..

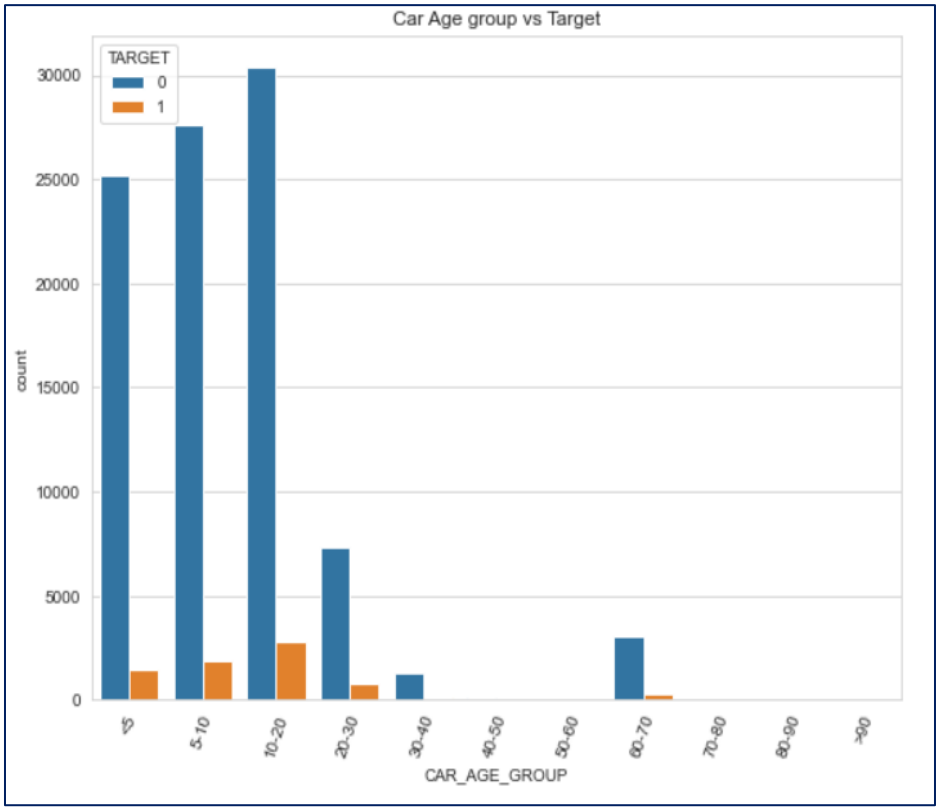




Inference: Unemployed people and low skill laborers seems to have payment difficulties

Inference: People with education type lower secondary seems to have more payment difficulties

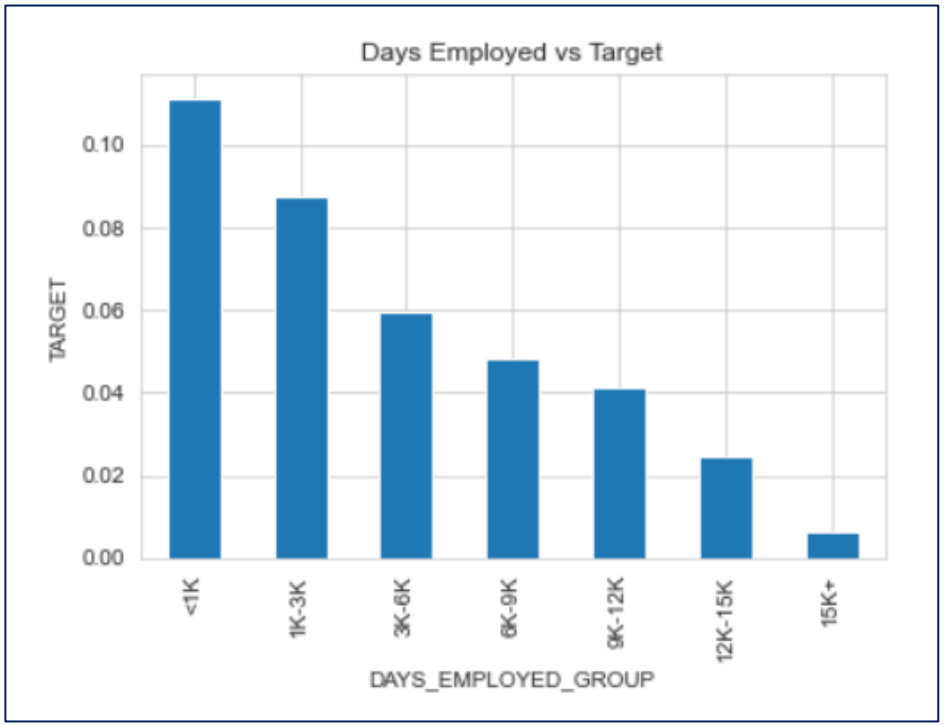
Continued..

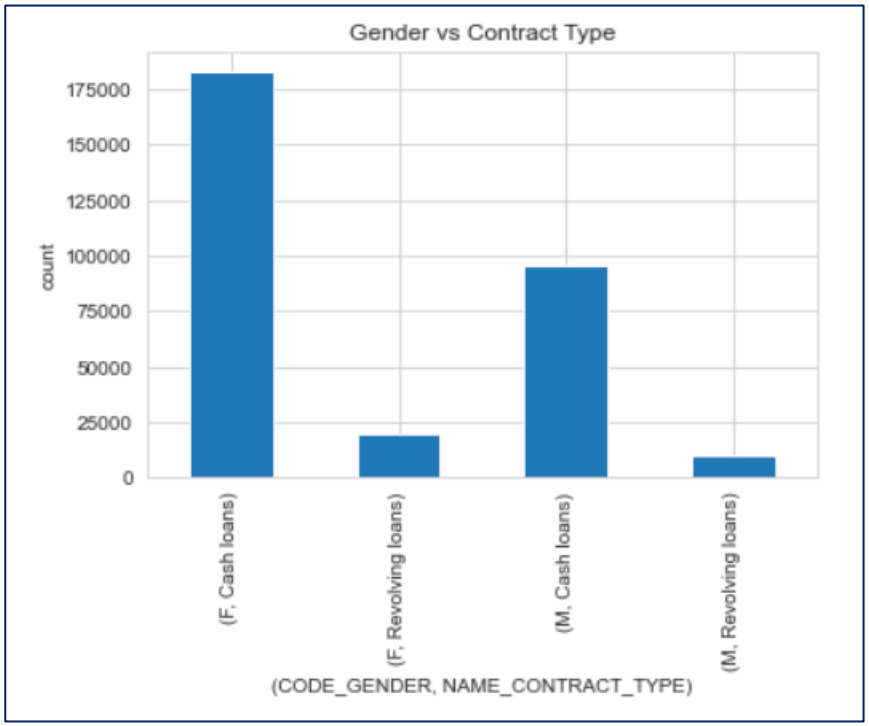


Inference: People who stay in rented apartments or with parents seems to have more payment difficulties

Inference: People having cars with age group below 30 seem to have payment difficulties

Continued..



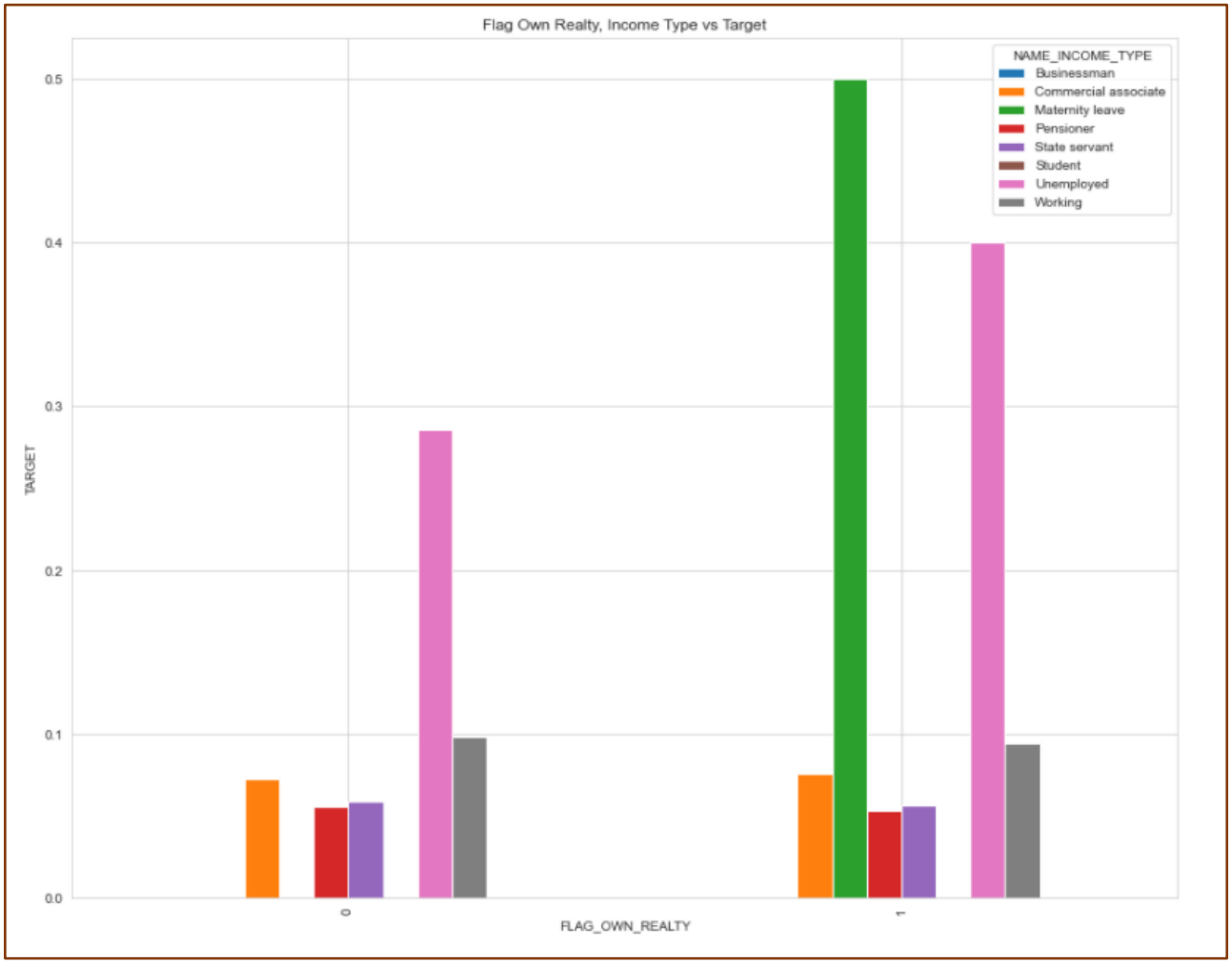


Inference: People having lesser days of employment seem to be having payment difficult

Inference:

- Cash loans are outnumbering revolving loans
- In both the categories of contract type, female applicants are outnumbering male applicants

Multivariate Analysis

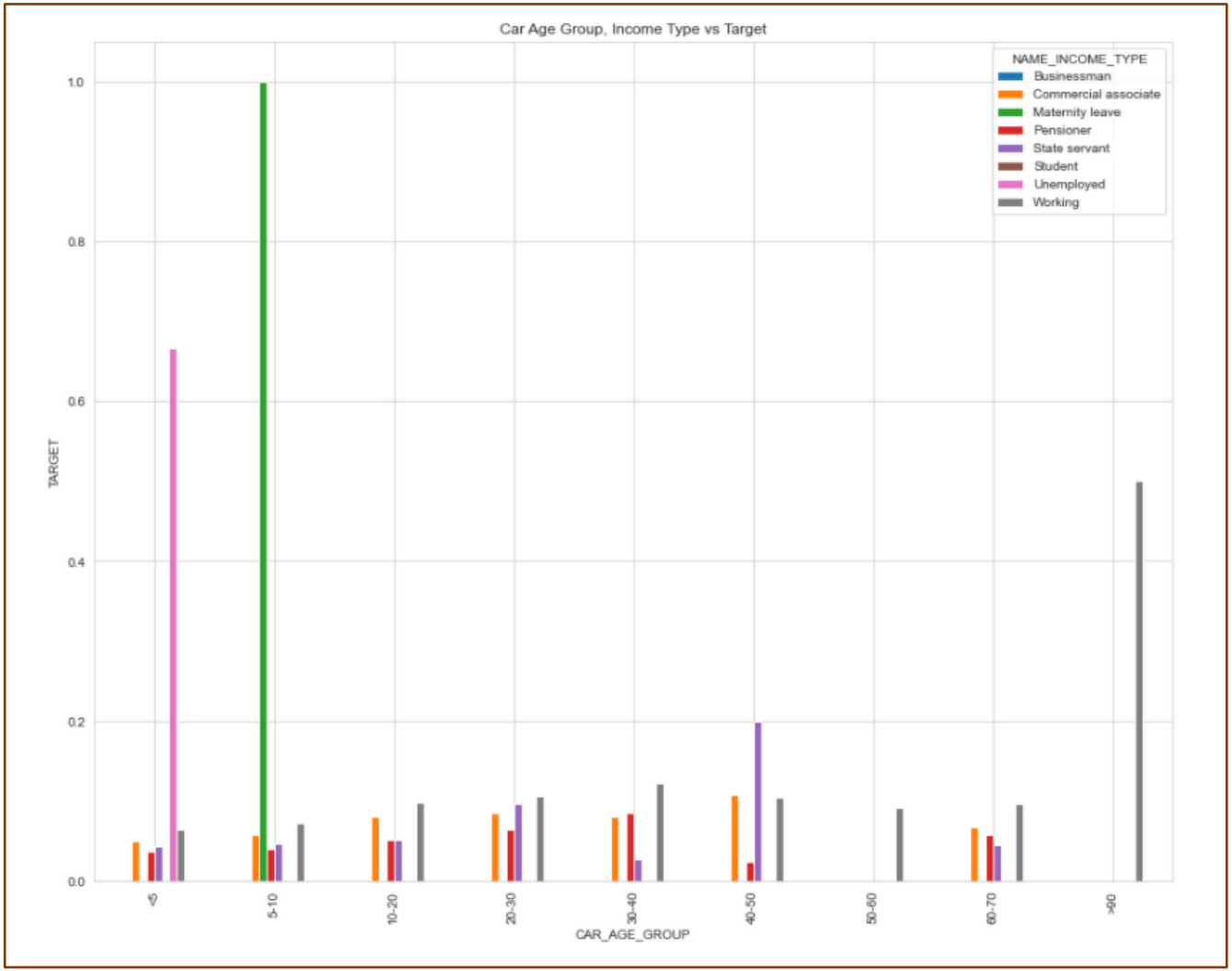


Inference:

Following people seem to have payment difficulties:

- People on Maternity leave owning realty
- Unemployed people irrespective of owning realty

Continued..

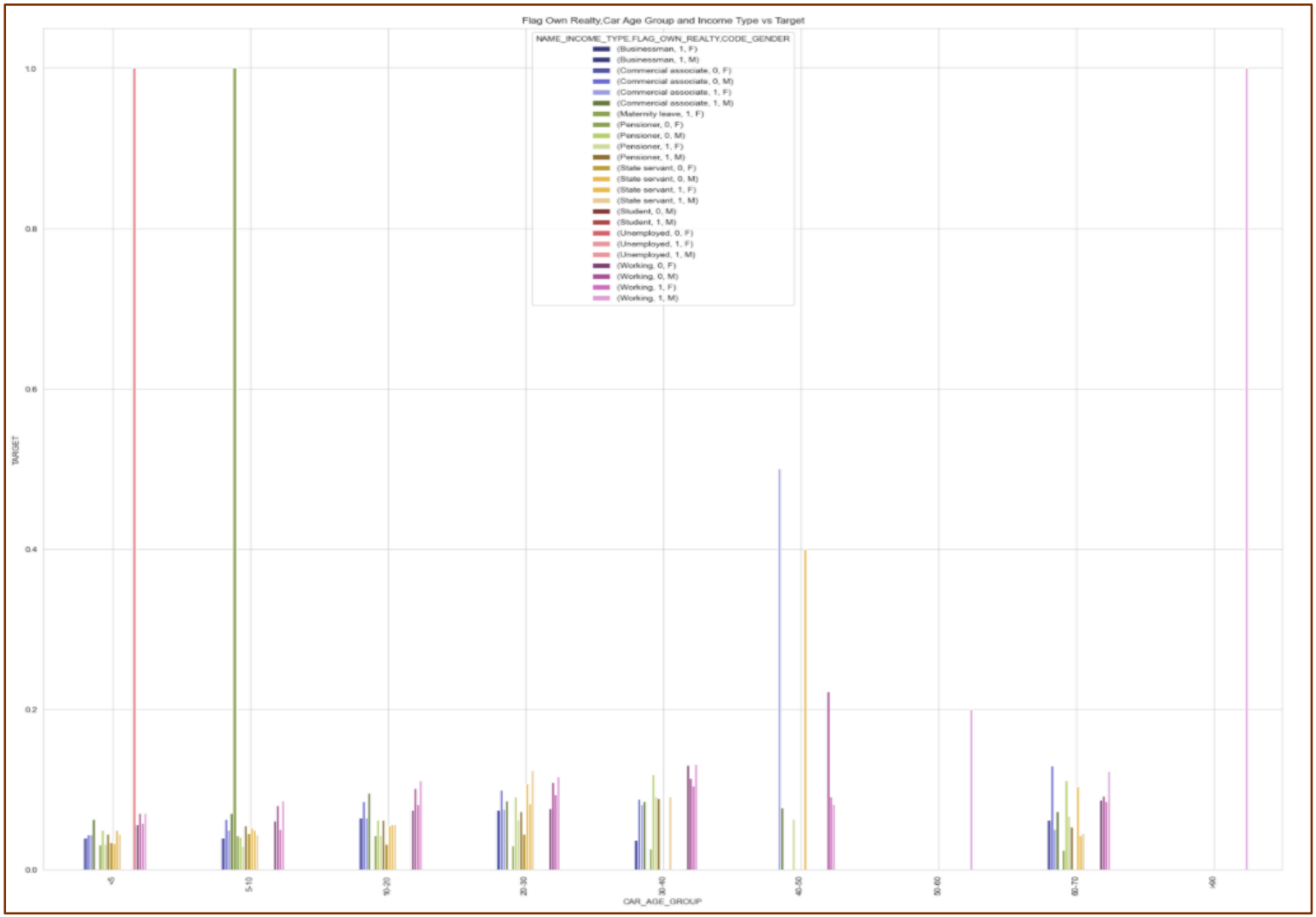


Inference:

Following people seem to have payment difficulties:

- People on Maternity leave having a car age less than 10 days
- Unemployed people having a car age less than 5 days

Continued..

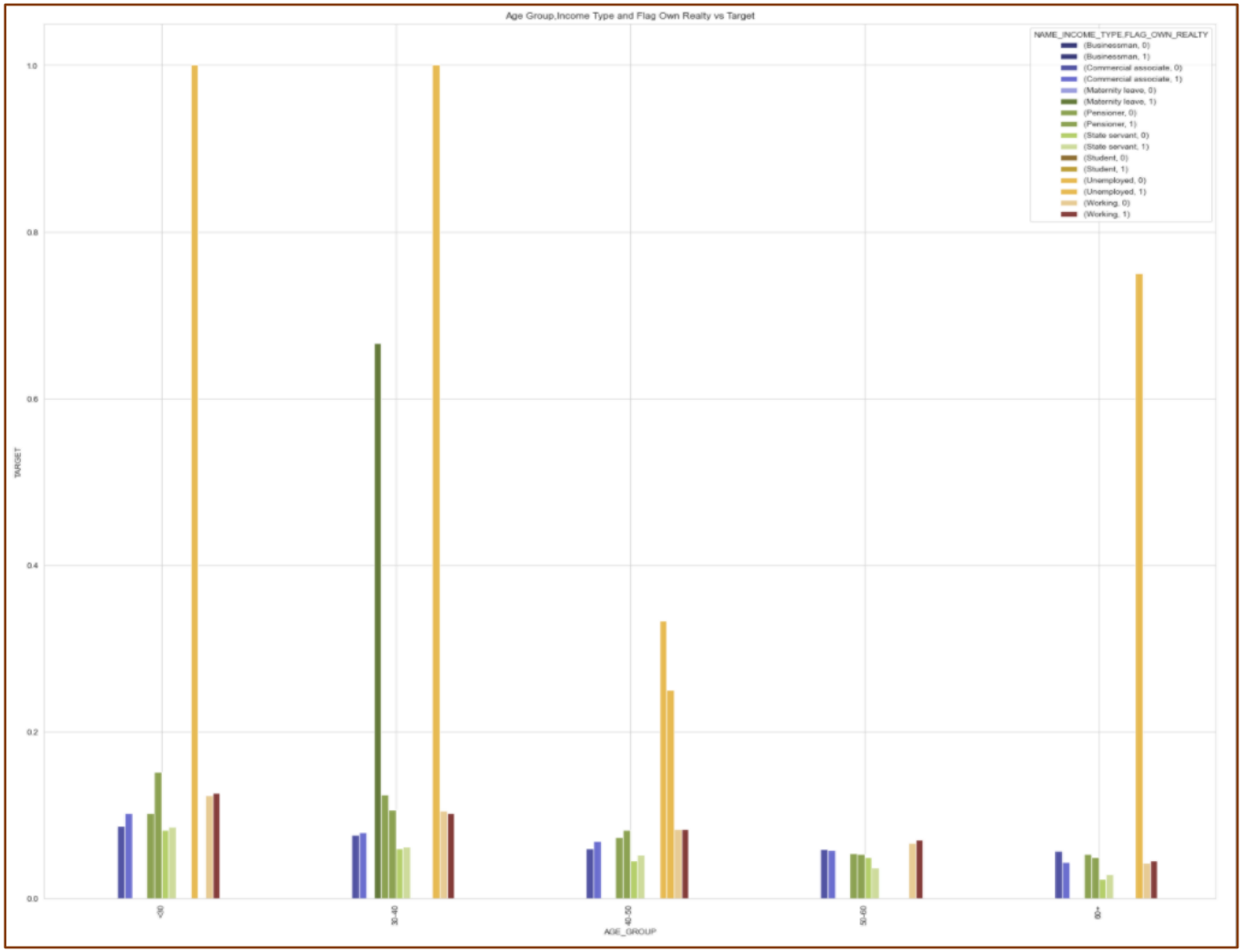


Inference:

Following people seem to have payment difficulties:

- People on Maternity leave owning realty and having a car age less than 10 days
- Unemployed people owning realty and having a car age less than 5 days

Continued..

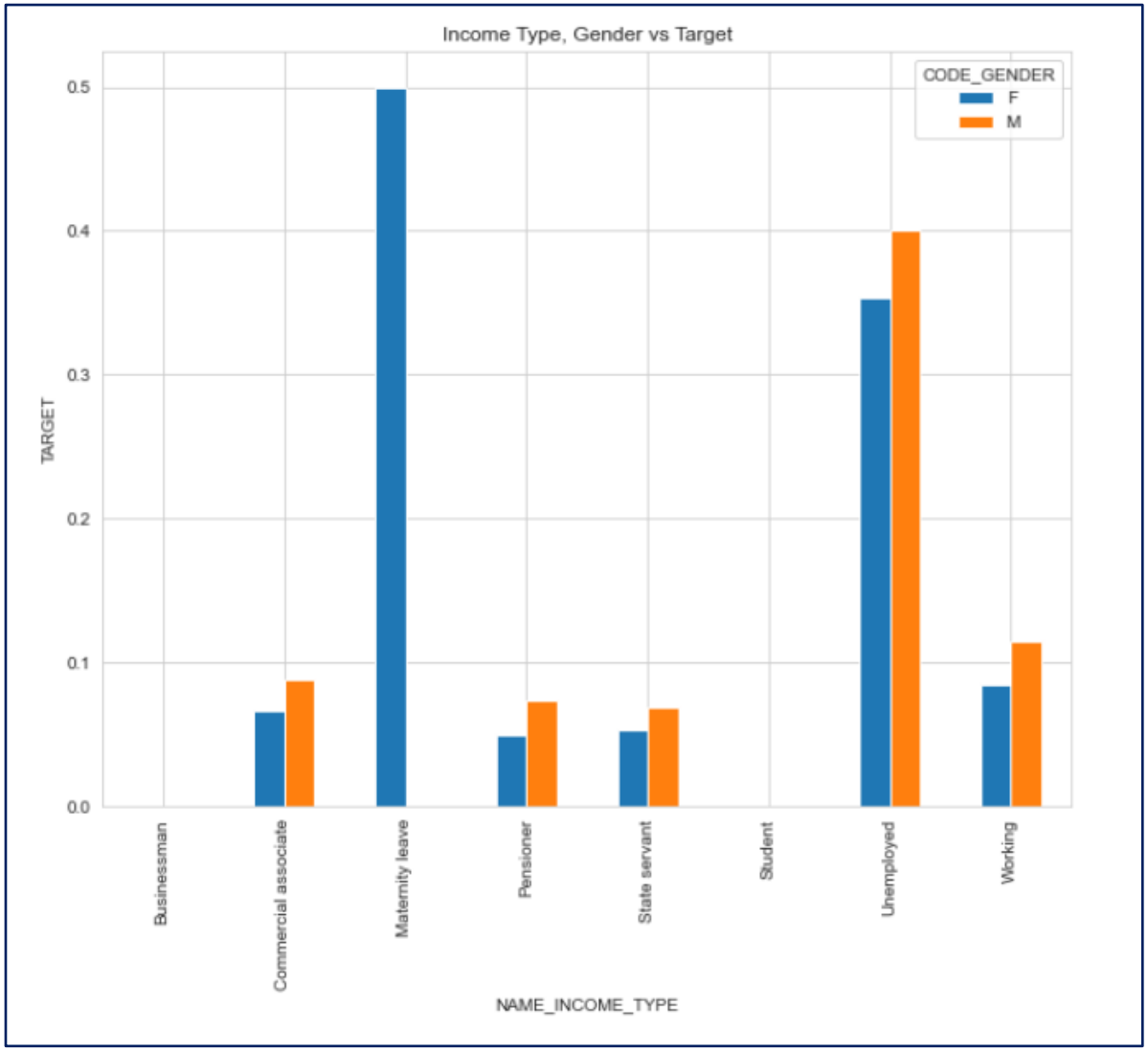


Inference:

Following people are having payment difficulties:

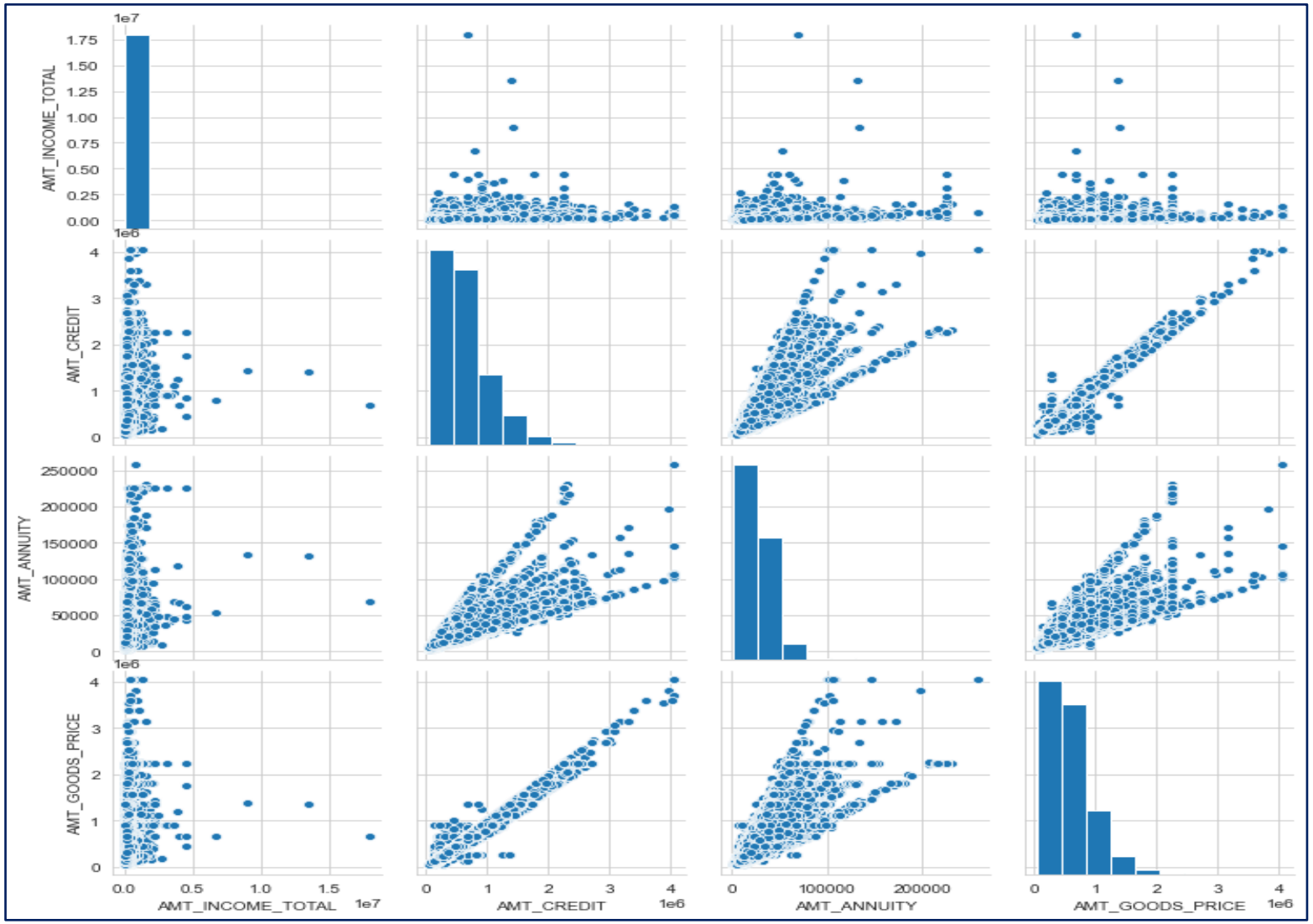
- Unemployed people of age group <30 not owning a realty
- Unemployed people of age group 30-40 owning a realty
- People on Maternity leave of age group 30-40 owning realty

Continued..



Inference: Females on maternity leave and Unemployed men seem to be having more payment difficulties

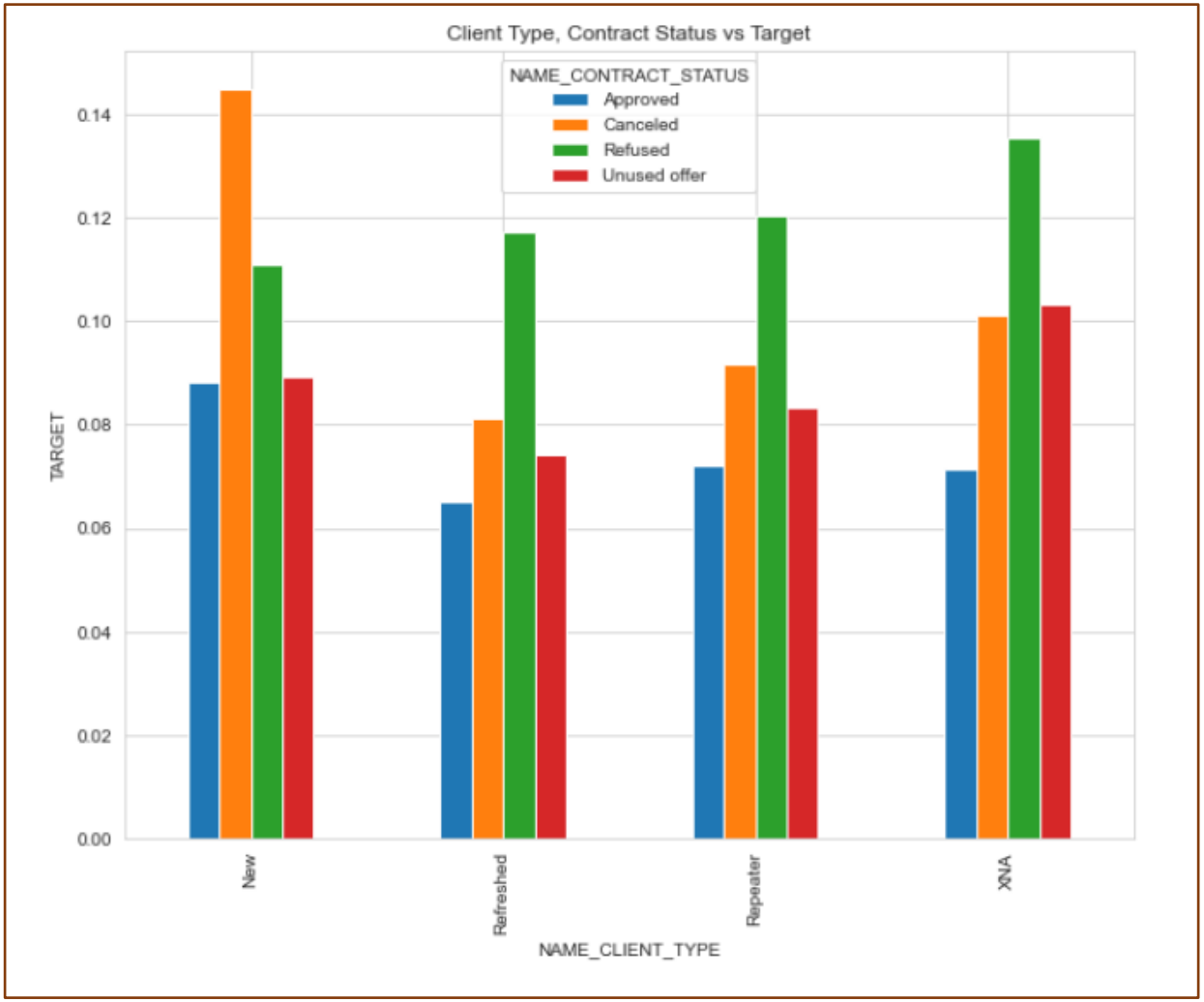
Pair Plot



Inference: Higher the good price , higher is the credit amount and higher the annuity. Goods price, Credit amount and Annuity are linearly related

Analysis of merged dataset (application.csv + previous_application.csv)

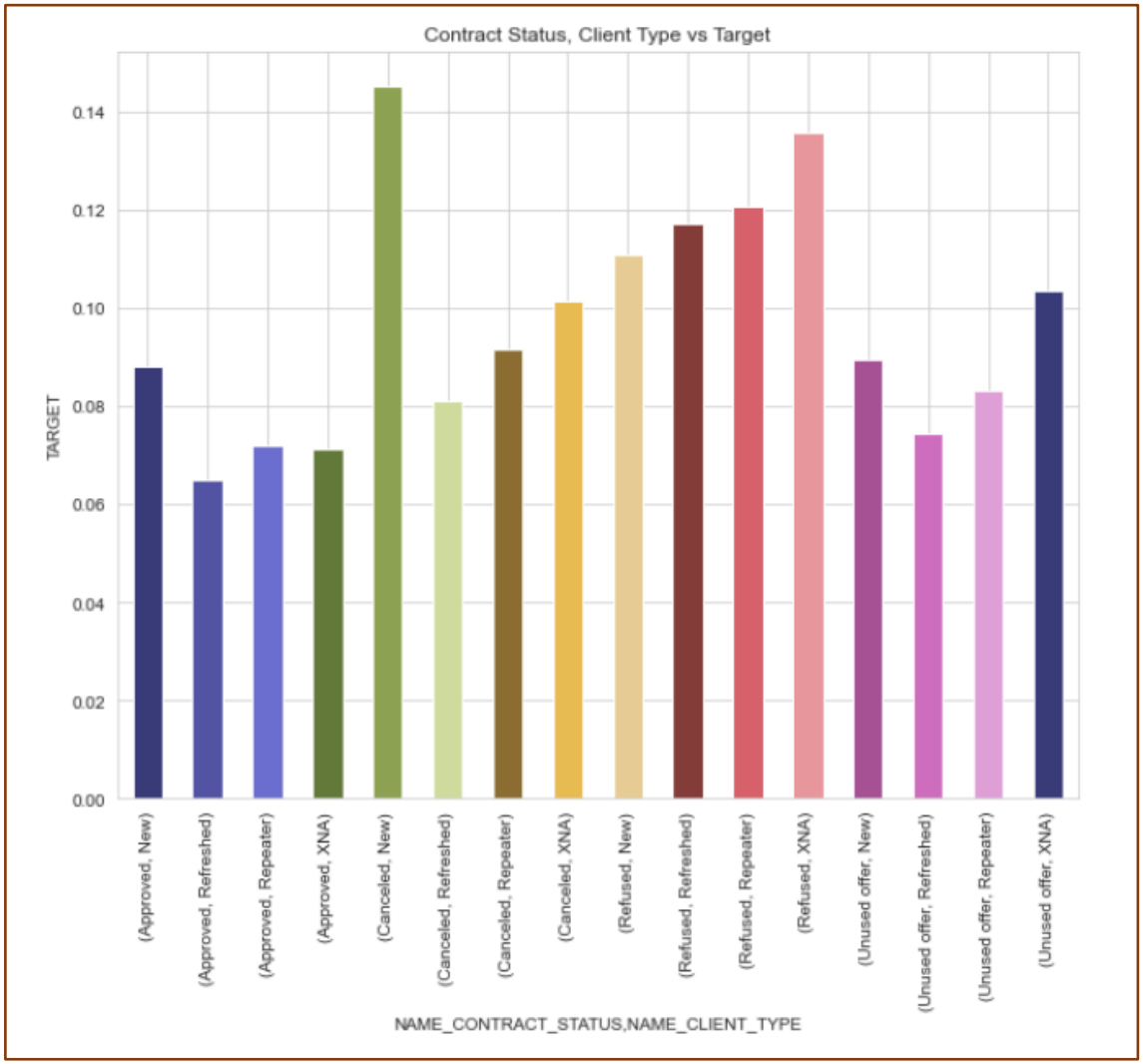
Multivariate analysis



Inference:

- Number of approved requests are lesser compared to other categories for all client types
- More number of cancelled and approved requests observed for new clients
- Number of refused requests seems comparable across all client types

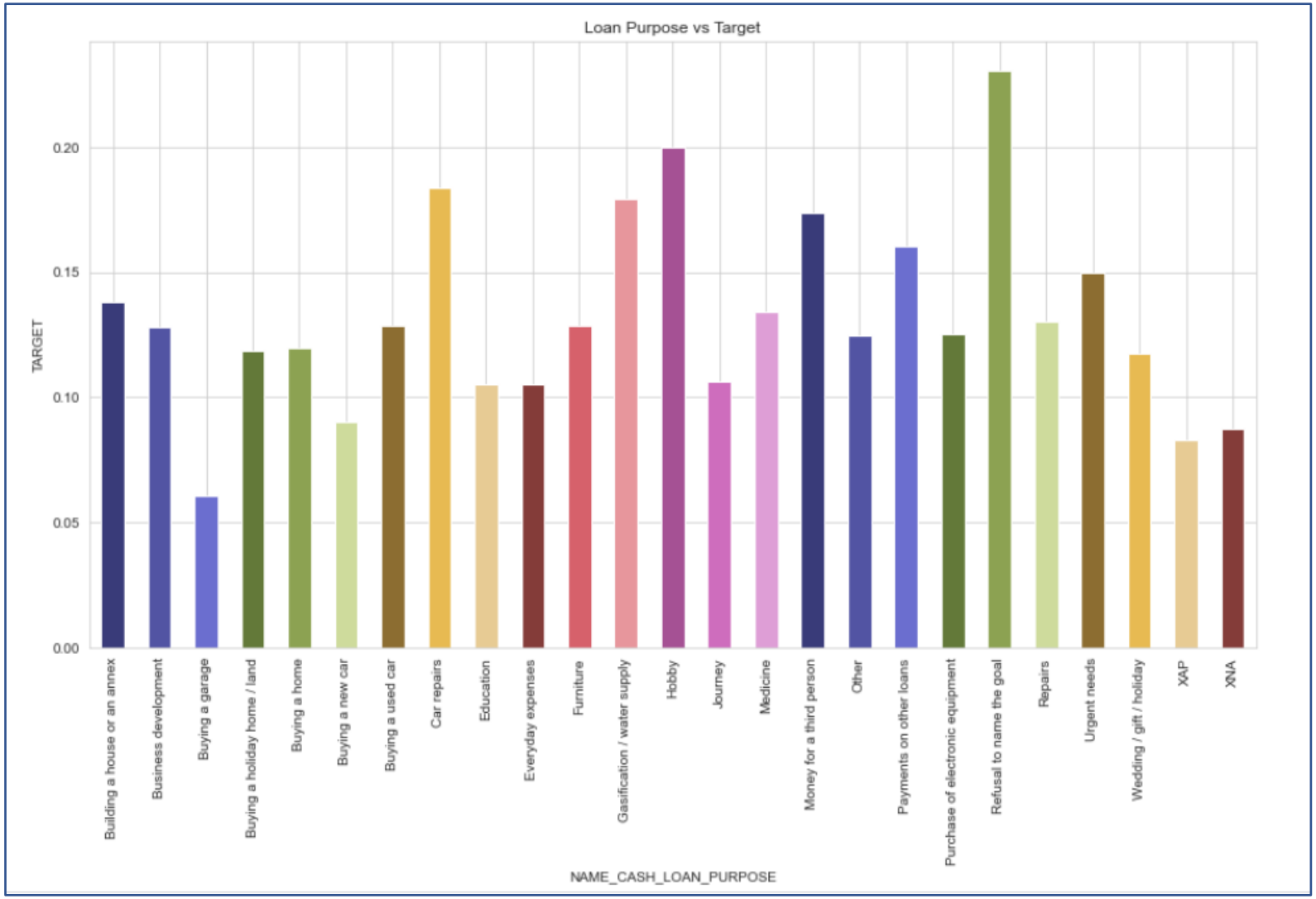
Continued..



Inference:

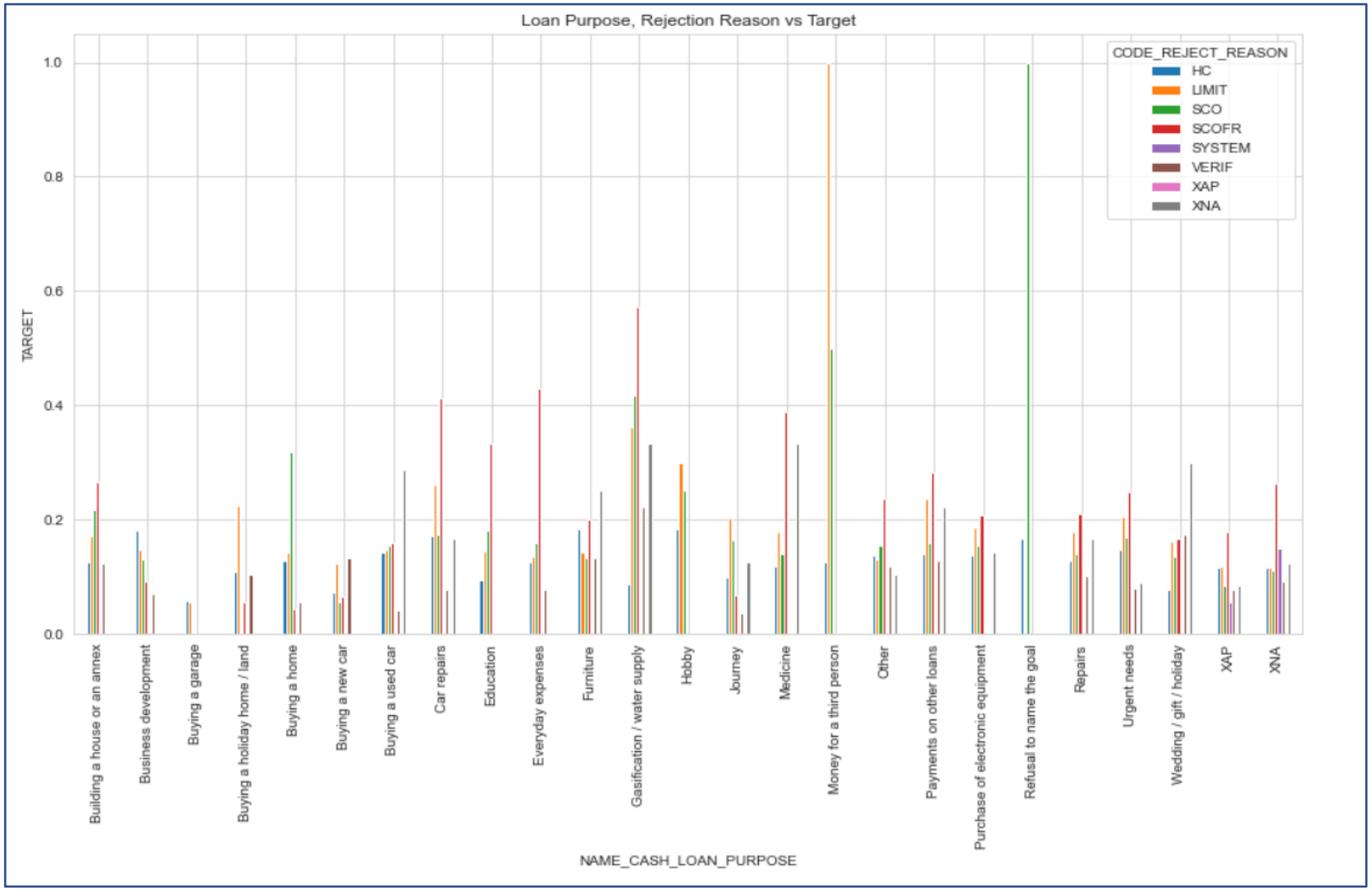
- New clients with cancelled application status seems to have payment difficulties
- Clients with refused application status seems to be having payment difficulties irrespective of client type

Continued..



Inference: People with loan purposes stated as 'Refusal to name the goal', 'Hobby' and 'Car repairs' seem to be having payment difficulties

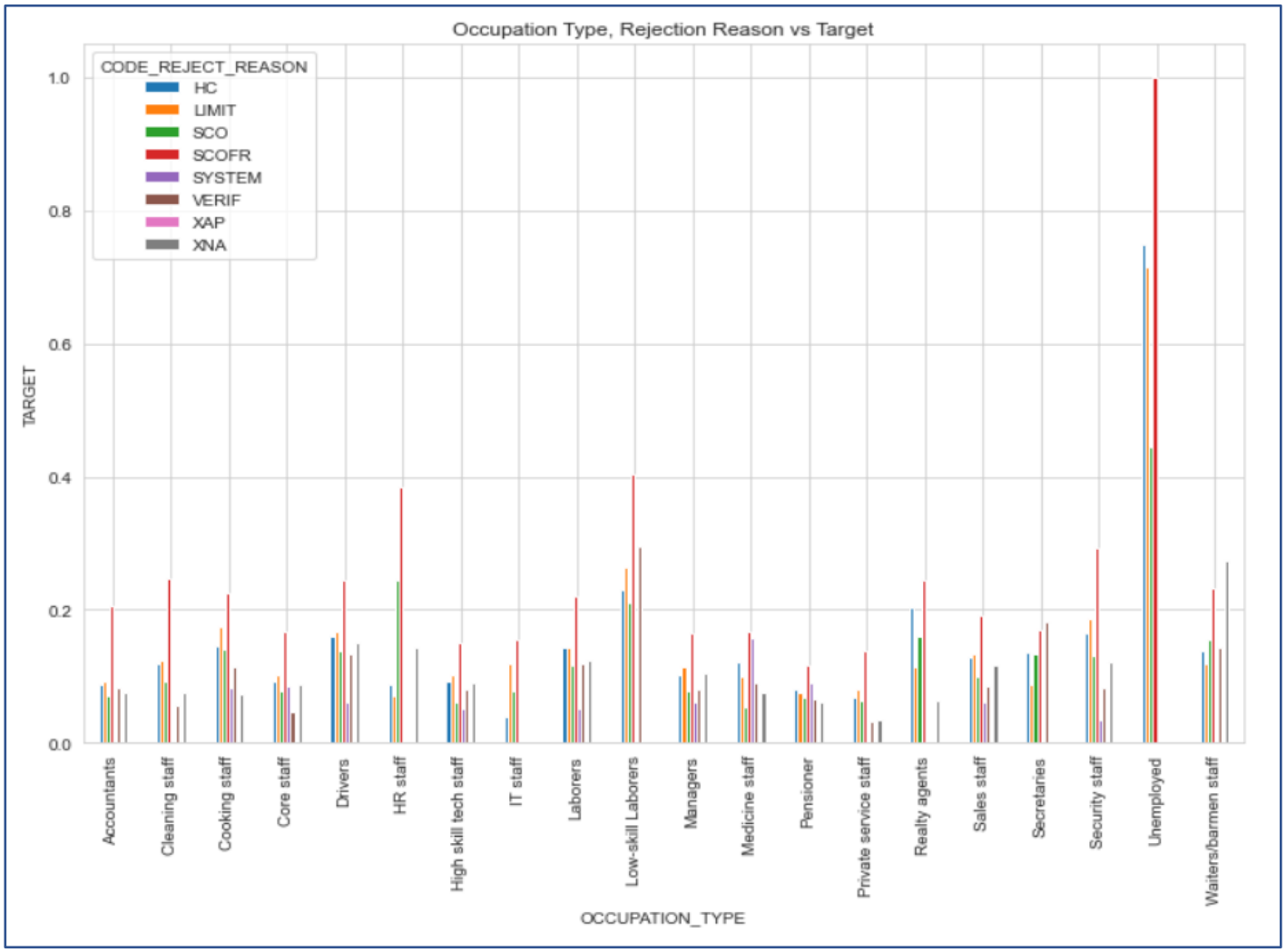
Continued..



Inference: Of the rejected applications, following seem to have payment difficulties:

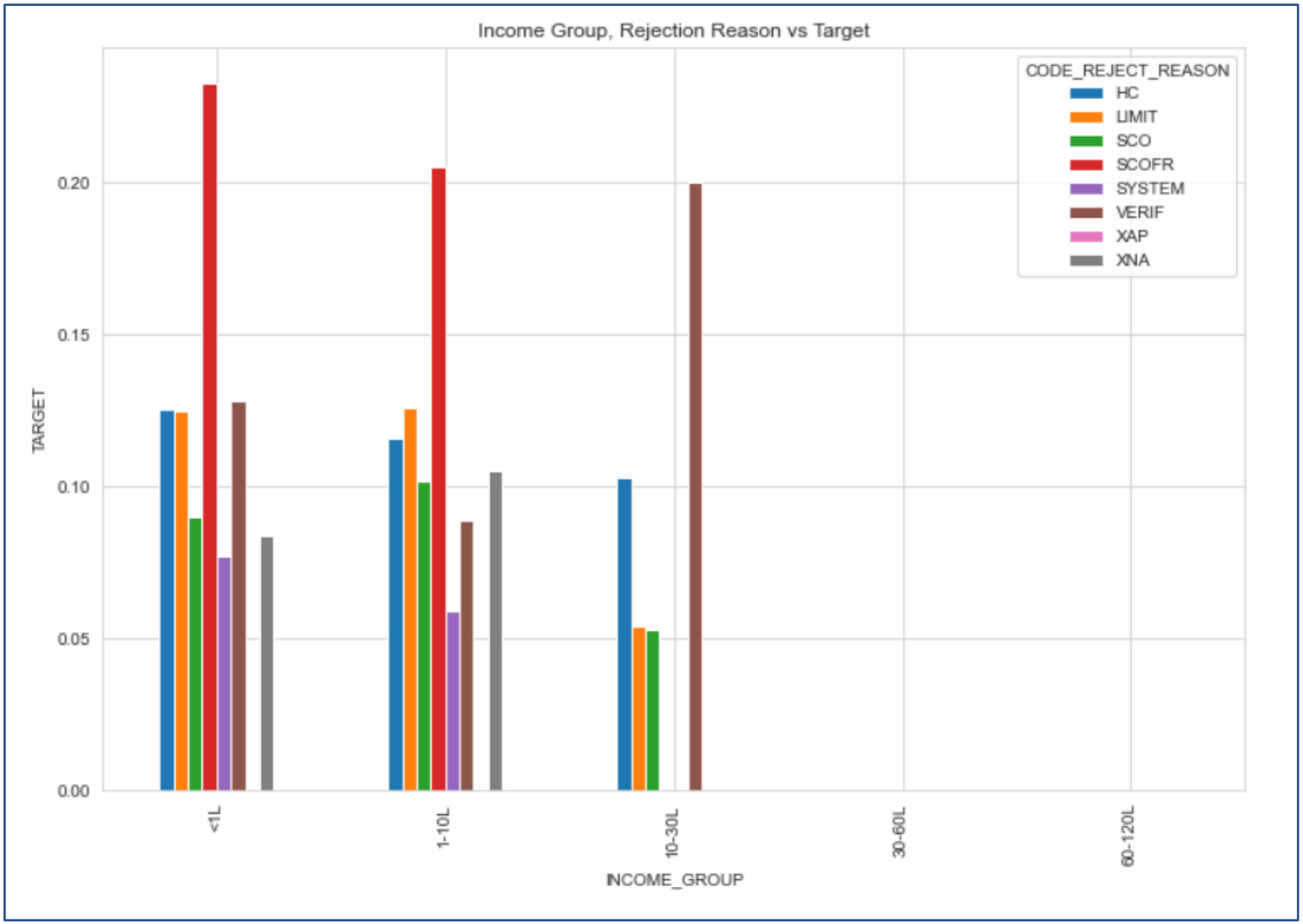
- Loan purpose as 'Refusal to name the goal' with rejection reason as 'SCO'
- Loan purpose as 'Money for a third person' with rejection reason as 'SCOFR

Continued..



Inference: Unemployed people with application rejection reason as HC and SCOFR seem to be having more payment difficulties

Continued..



Inference: Following seem to be having more payment difficulties:

- People having income less than 10 lakhs with application rejection reason as SCOFR
- People having income between 10-30 lakhs with application rejection reason as VERIF

Conclusion

Inferences

- Target value 1 implies payment difficulties and 0 indicates no payment difficulties. Imbalance of data noticed based on this variable. Ratio of imbalance based on Target value is found to be: **91.34 : 8.66**
- Bank should focus less on Unemployed people and people who are on maternity leave, since they are having payment difficulties and likely to default
- Bank should focus more on people having Income more than 60 Lacs, as they are the probable customers
- Bank should focus less on people who are applying for loan purpose as 'Refusal to name the goal', 'Hobby', 'Money for a third person' and 'Car repairs'
- People having lesser days of employment seem to be having payment difficulties