CREDIT EDA CASE STUDY

By Shibani Roy Choudhury

Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.

This case study aims to identify patterns, which indicate if a client has difficulty paying their instalments, that they may used for taking actions on loan applications, such as :

- denying the loan,
- reducing the amount of loan,
- 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.

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 utilize this knowledge for its portfolio and risk assessment.

Analysis Approach

Data Overview:

The dataset comprises various attributes such as applicant demographics, loan details, repayment history, credit scores, and contract statuses. It includes historical loan application data with information on approved, refused, cancelled, and unused contracts.

Analysis Approach:

1. Data Cleaning and Preprocessing:

- **Handle missing values**: Impute missing values or remove observations with significant missing data.
- **Data transformation:** Convert numerical variables into categorical representations and vice versa if necessary.
- **Outlier detection**: Identify and handle outliers that could skew the analysis results.

Descriptive Analysis:

- Explore summary statistics of key variables such as loan amount, applicant income, repayment status, etc.
- Visualize distributions and correlations between variables using histograms, box plots, and correlation matrices.
- Segment customers based on demographic attributes, income levels, and credit profiles.
- Analyze loan performance and default rates across different customer segments.

Loan Approval Analysis:

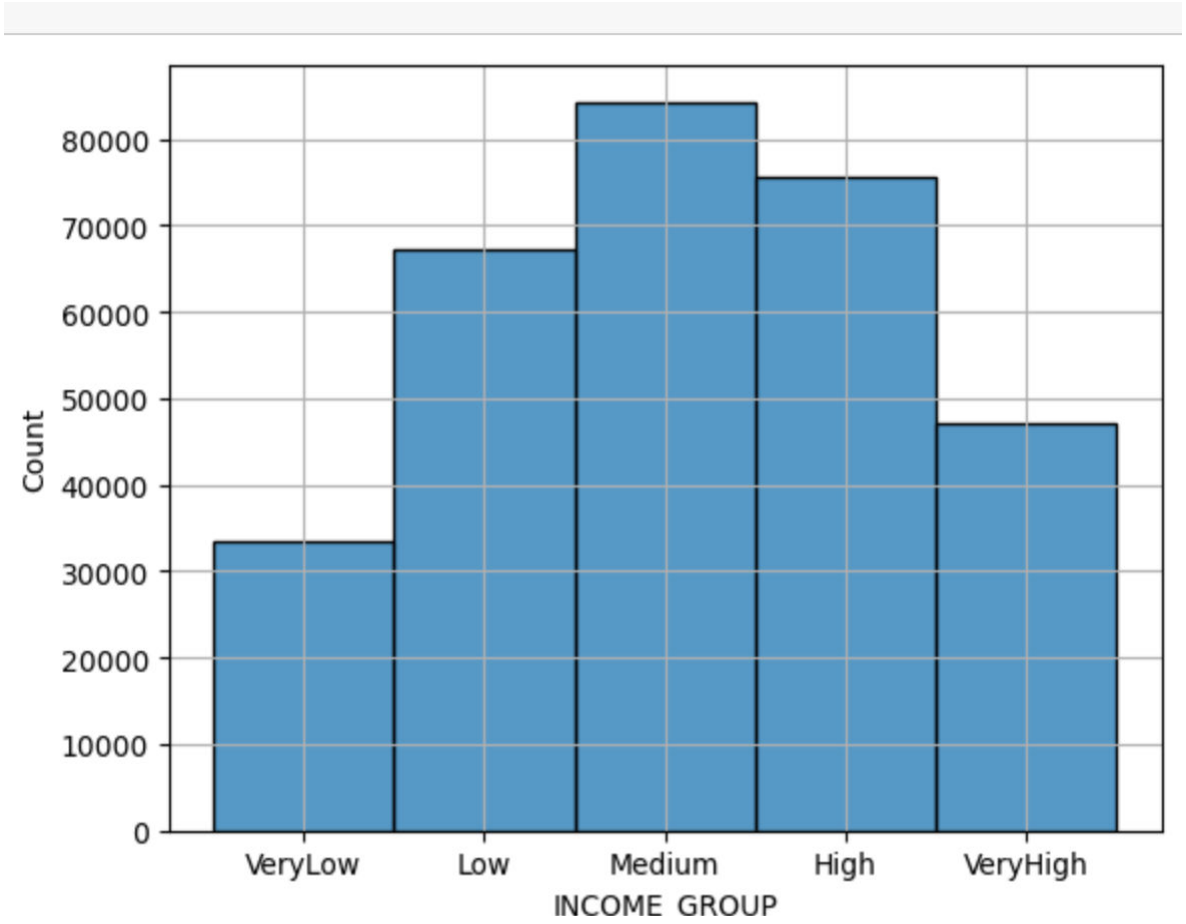
Analyze approval rates based on applicant demographics (e.g., age, gender, marital status), income levels, and credit scores.

Inferences : Identify factors which effects the default tendency of loan application.

UNIVARIATE ANALYSIS

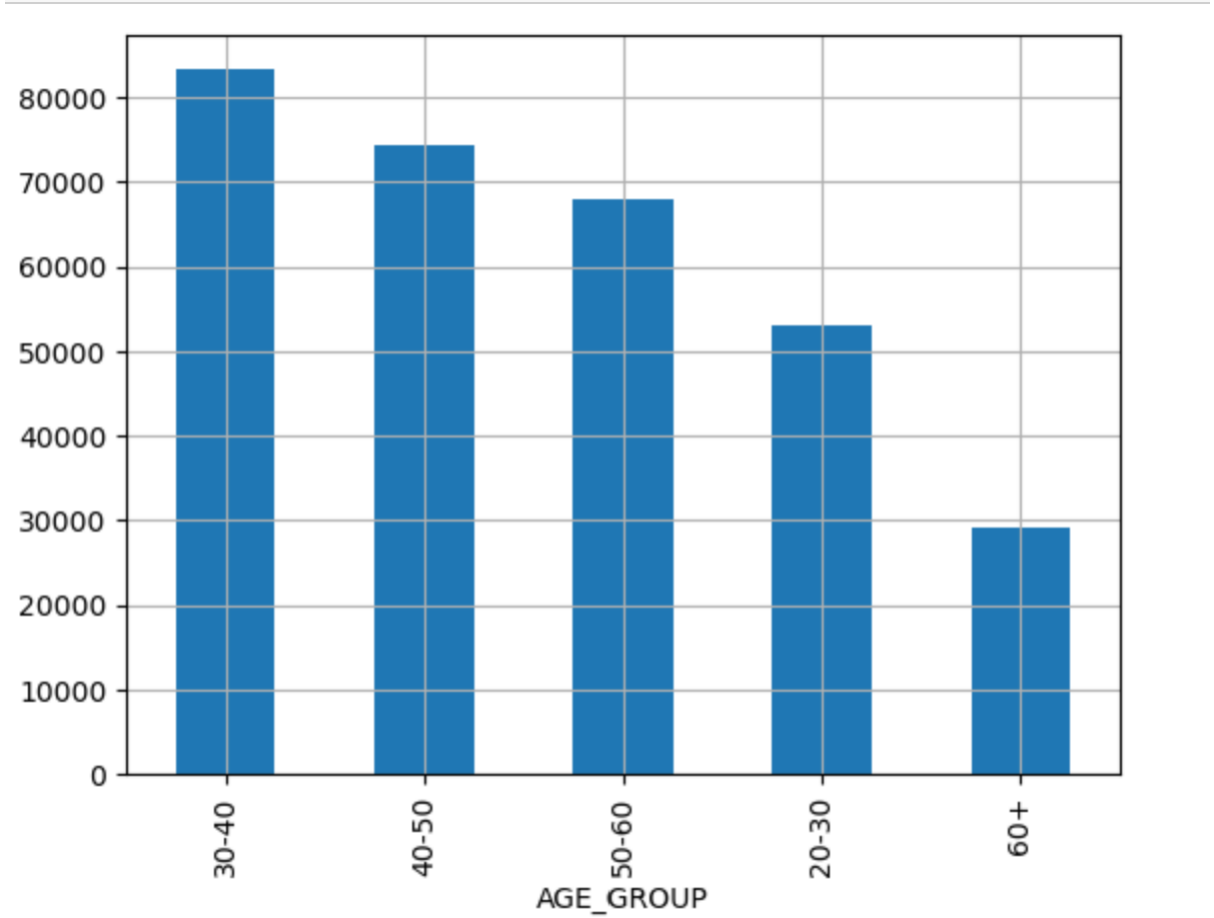
Points to be concluded from the graph on the right side.’

- More applicant belongs to Medium income group, then High income group



Points to be concluded from the graph on the right side.’

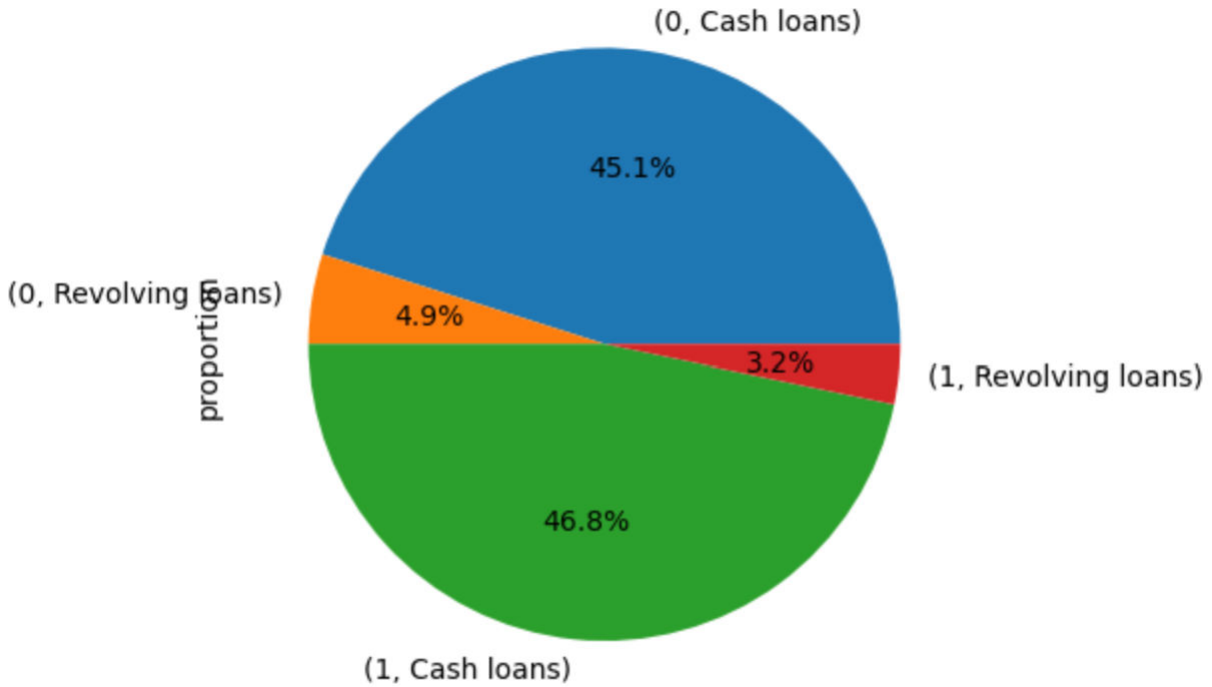
- Maximum applicants belongs to 30-40 range



ANALYSIS WITH TARGET

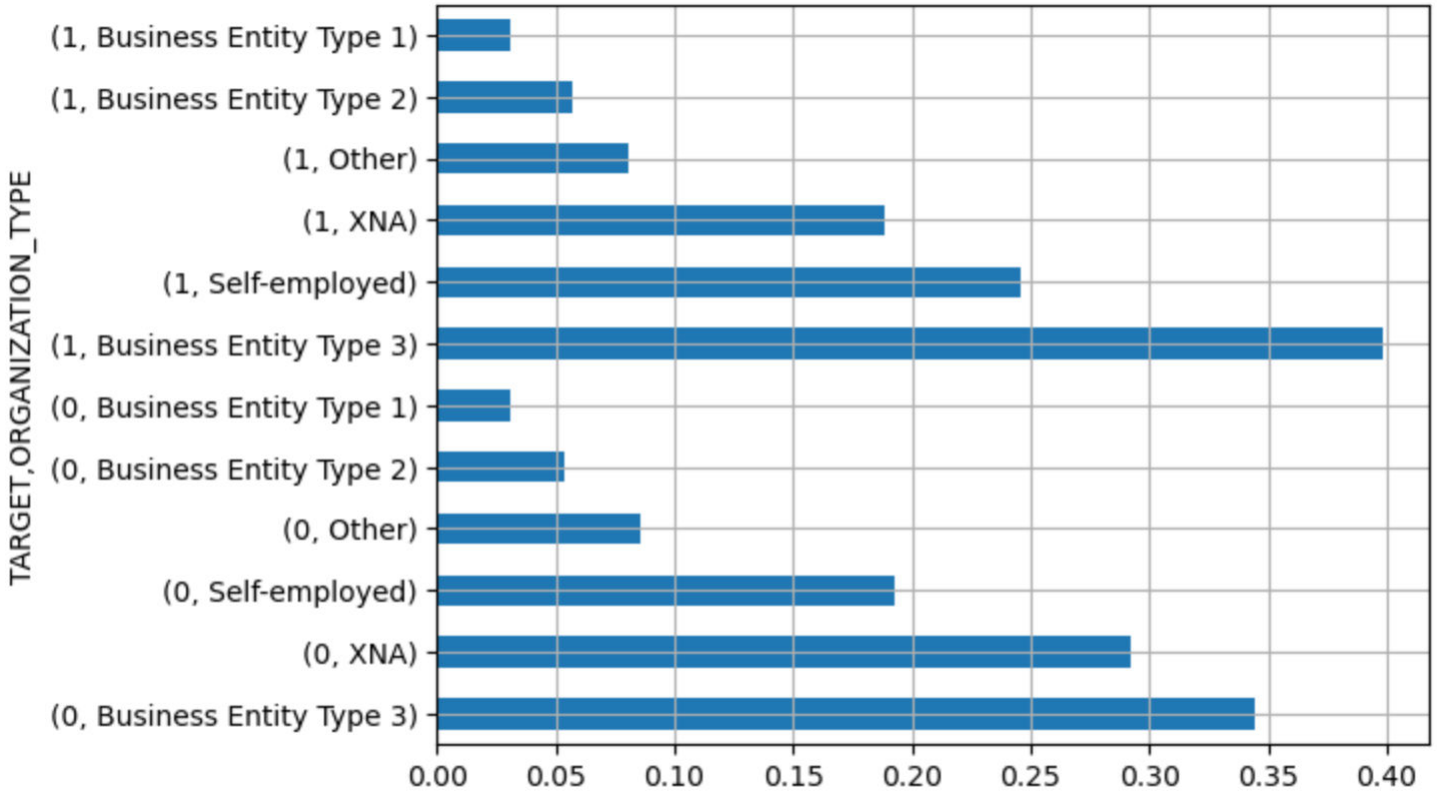
Points to be concluded from the graph on the right.

- Most loan applications as of Cash Loans.
- For Target 0 (No Payment difficulty) - 45.1% and almost 46.8% for Target-1(Payment difficulty).
- Most cash loan having payment difficulty as compare to Revolving loan’.



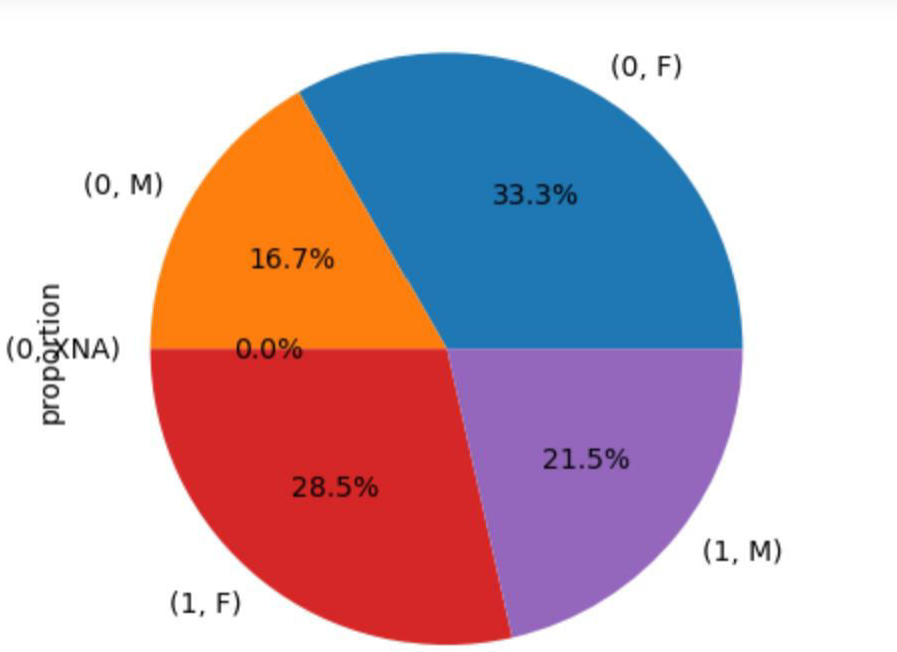
Points to be concluded from the graph on the right.

- Business ENTITY TYPE 3 AND SELF EMPLOYED add up to more than 40% defaulters..
- The highest % of loan defaulters are also seems this category. Most cash loan having payment



Points to be concluded from the graph on the right.

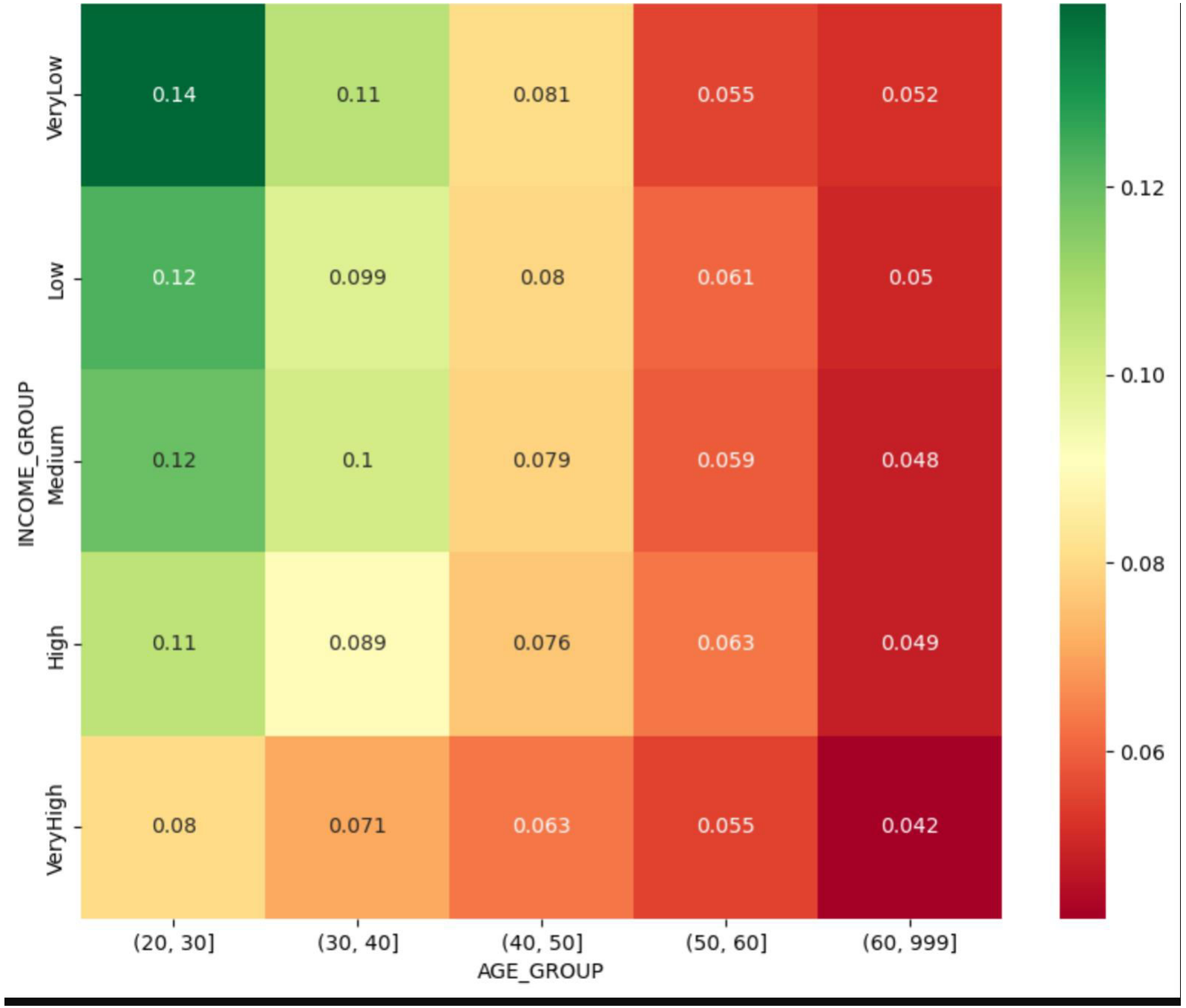
- Ratio defaulter case between Male and Female, shows that as compare to Female Male defaulters are more
- As shown in the plot data has more females as loan applicant.
- Similarly as per plot, though male applicants are lower, ratio of male applicants defaulting is higher.



MULTIVARIATE ANALYSIS - TARGET

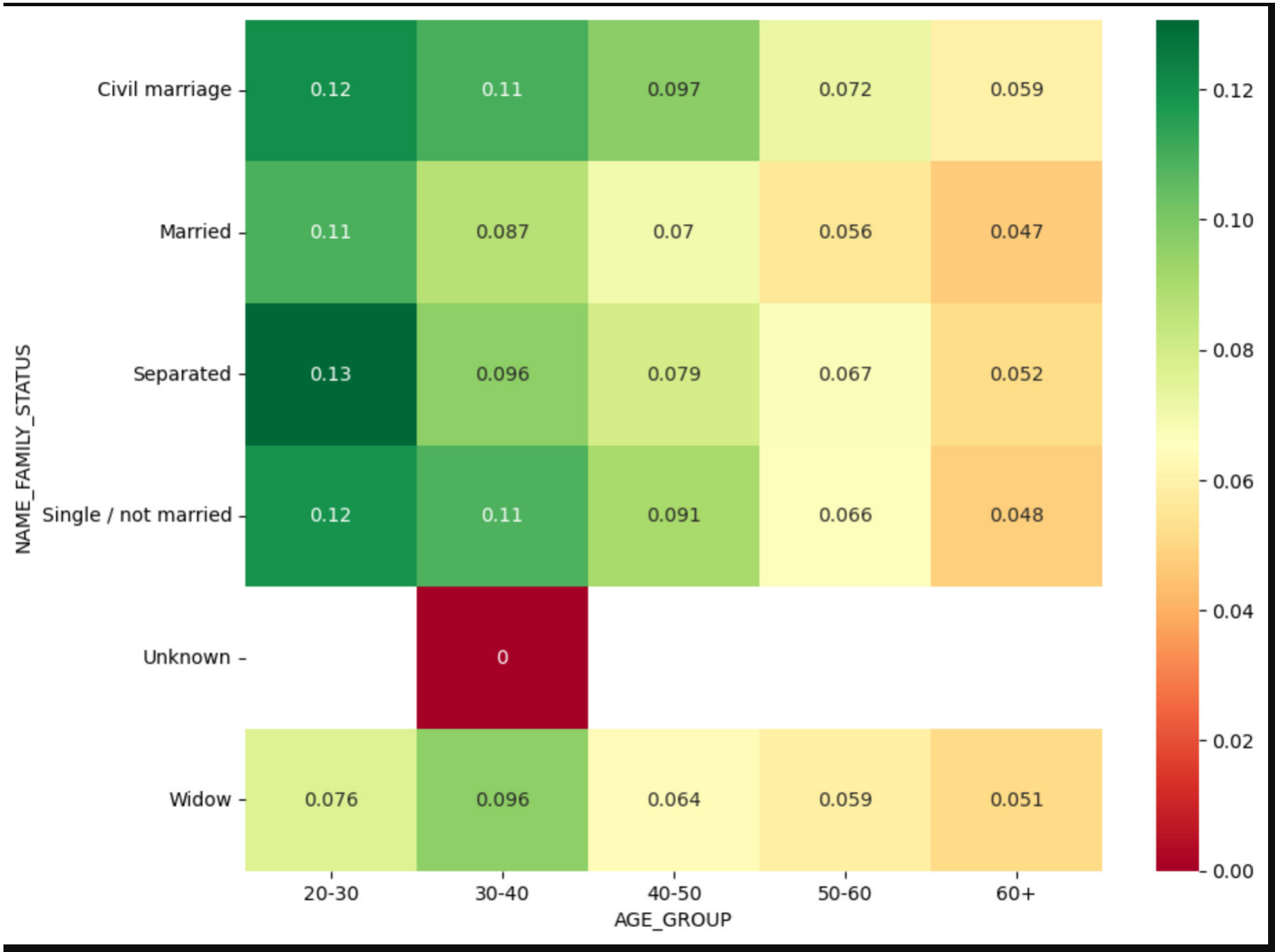
Points to be concluded from the graph on the right.

- AGE_GROPUP – 40+ are more in TARGET 0. In Target 1- 20-30 have higher share. Age does seem like influencing default as shown in earlier plot also.
- INCOME_GROUP - Medium income group, Low and very Low income group have more count in Target 1 then other income group



Points to be concluded from the graph on the right.

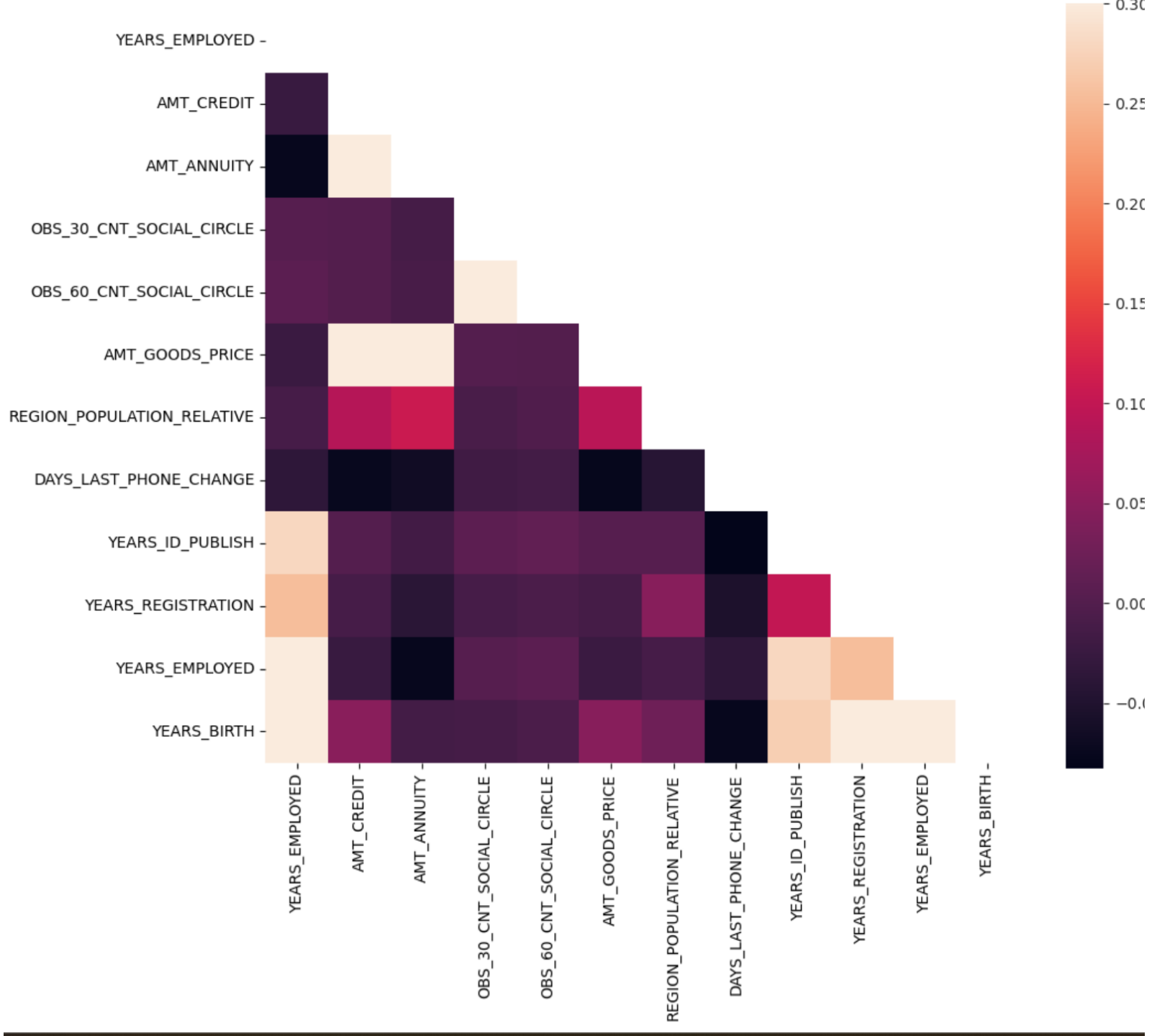
- In Married Applicant in the age group 20-30 and 30-40 is the largest group of applicant with payment difficulties then other age group.
- In all Marital status Age 20-30 having largest share in payment difficulty



Correlation Matrix

Points to be concluded from the graph on the right side.

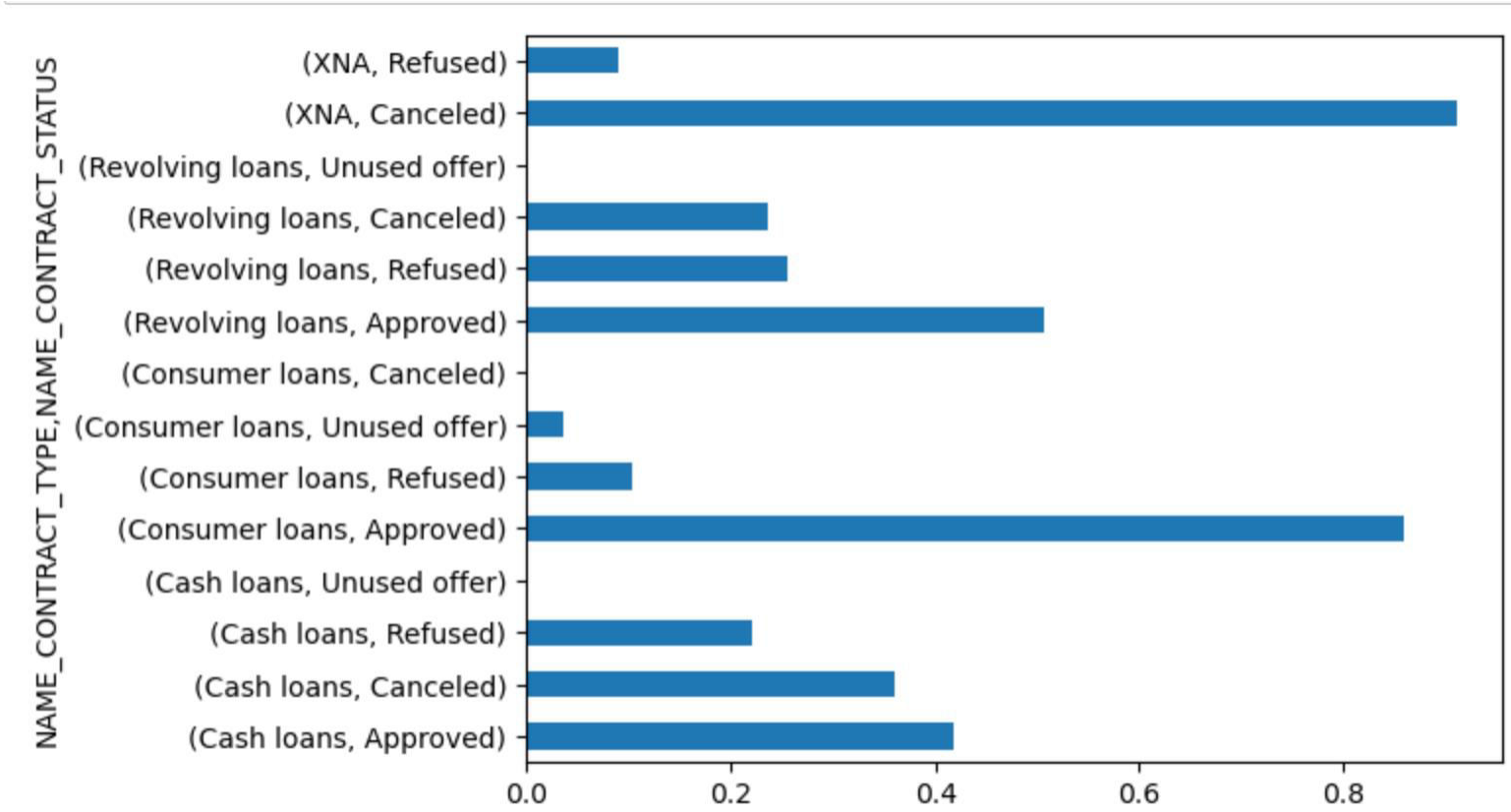
- AMT_Annuity is highly coorelated with AMT_Credit.
- AMT_Credit highly coorelated with AMT_Goods_Price
- AMT_CREDIT is highly coorelated with AMT_Goods PRice
- YEARS_EMPLOYED is highly coorelated with YEARS_BIRTH



BIVARIATE ANALYSIS WITH APPROVED STATUS IN PREVIOUS LOAN APPLICATION

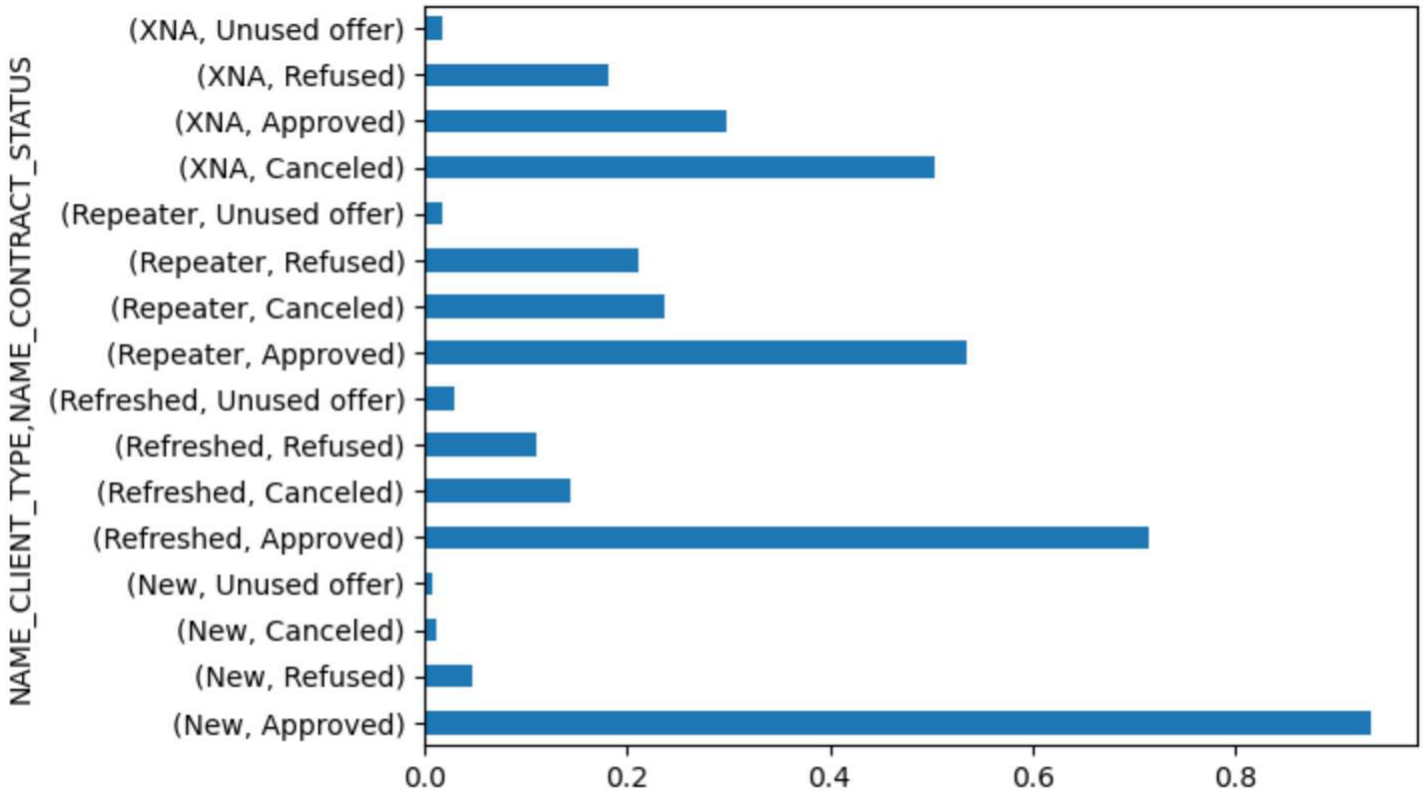
Points to be concluded from the graph on the right side.

- In approved category, consumer loan has largest no of applicants.
- There seem to be no cancelled loans in consumer category than cash loan.
- More cash loans have been refused than consumer loans.



Points to be concluded from the graph on the right.

- The bank has more repeaters in all refused, unused, cancelled categories.
- The bank approved more applications for new applicants

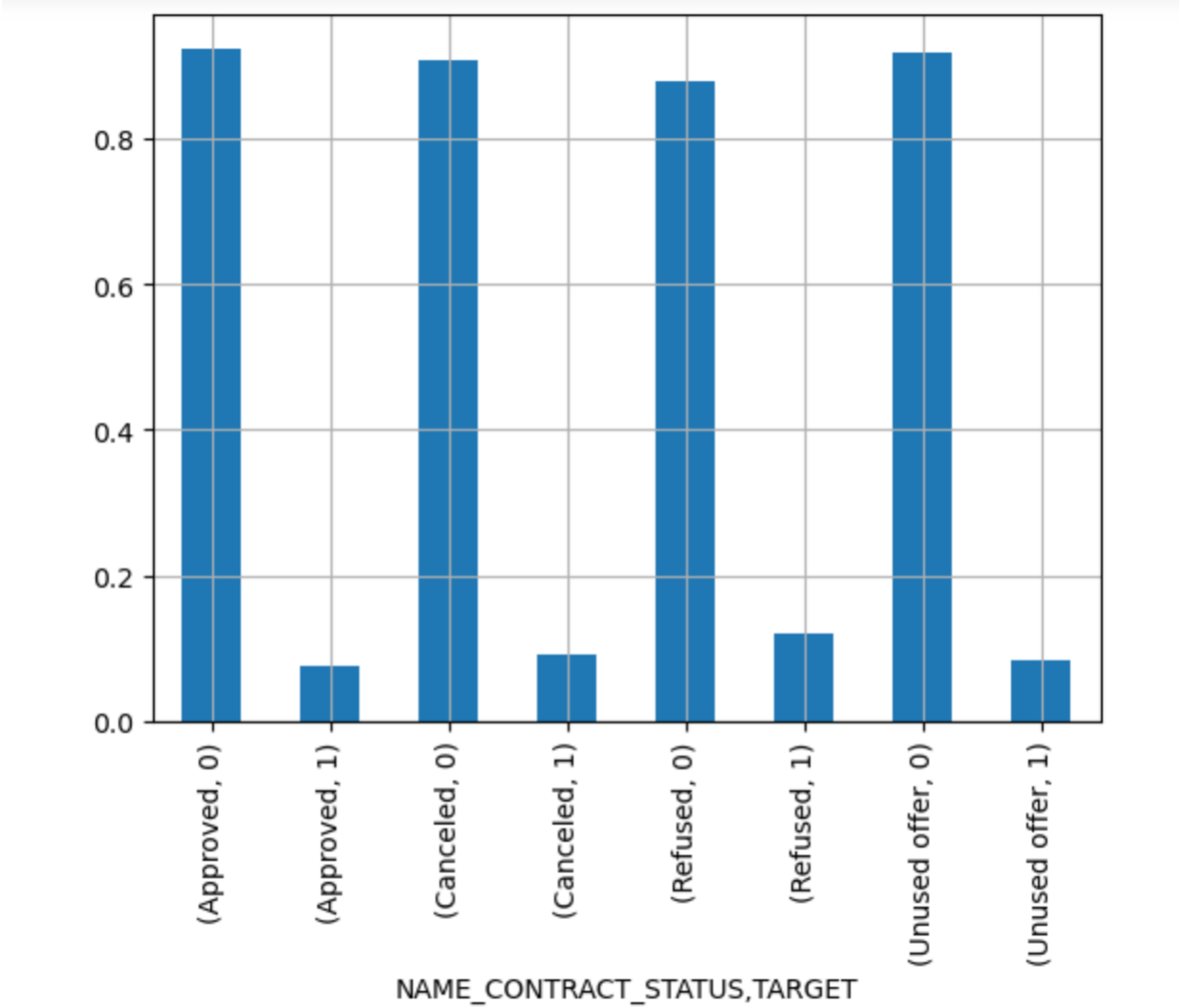


ANALYSIS WITH MERGE DATASET
IN
TARGET AND APPLICATION STATUS

Few points can be concluded from the graph.

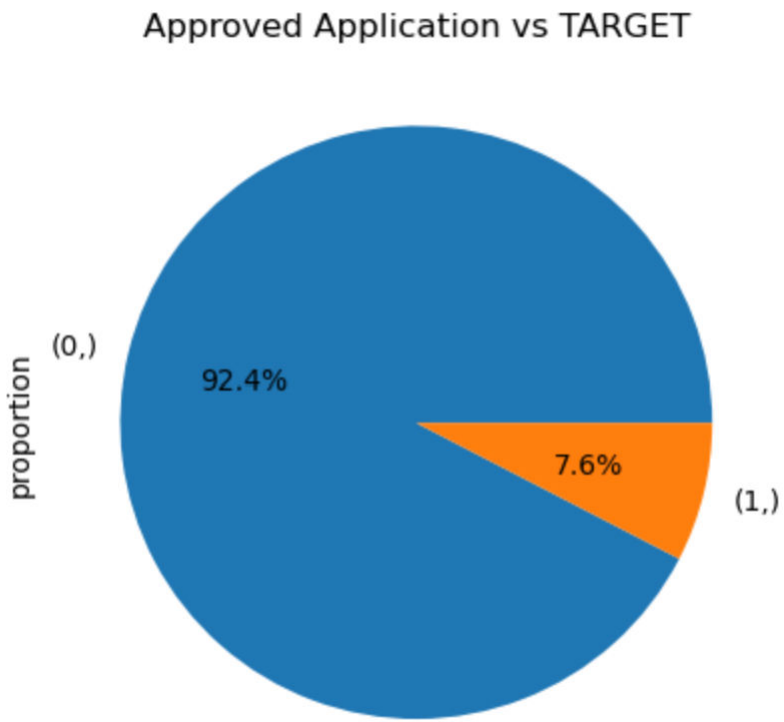
- Maximum application without having payment difficulty(TARGET = 0)- get cancelled, refused and unused : Loss of business.
- Similarly Previous cancelled, refused and unused application having payment Difficulty, need to think why previously not approved application, now approved and facing payment difficulty.

.



Few points can be concluded from the graph.

- 7.6% previously Approved cases having payment Difficulty



Summary

Below are the some factors / variables that effecting the default cases, that we found during the analysis of application data that also shown correct when analyze the same with previous applications against the Approved loans as well as other loans type, which having defaults, and it showing same trends

- Male having more default as compared to female
- Medium Income group .
- Age group 20-30 .
- Labourés Occupation Type.
- Organization Business-3 .
- Unemployed or Female on Maternity Leave having more default

Thank you