Credit EDA Assignment

Shayak Majumder
Data Science Program - July 2024

Problem Statement - Risk Analytics

In this case study, we were asked to analyze data from a loan-providing company to figure out:

- Identify patterns showing client difficulty in paying instalments
- Suggest taking actions like denying loans, reducing loan amounts, or offering higher interest rates to risky applicants
- Ensure capable consumers aren't wrongly rejected
- Use Exploratory Data Analysis (EDA) to identify such applicants

Datasets Provided

We were provided with 3 datasets:

- ***application_data.csv:*** Contains client information collected during the application process, focusing on payment difficulties
- ***previous_application.csv:*** Contains past loan application data
- ***columns_description.csv:*** Dictionary of variables/terms used in above two datasets

Analysis Approach

Based on what I learned so far in the UpGrad EDA module, I decided to take a 5-step approach:

Step 1: Read and inspect the data

Step 2: Handle missing values

Step 3: Handle Outliers which might contaminate my analysis

Step 4: Standardize data, add data columns for easier analysis

Step 5: Visualize data and draw observations

Step 1: Reading & Inspecting Data

Upon importing and inspecting the datasets, I found:

- Both *application_data* and *previous_data* loaded fine and there were no issues with headers as well
- I created dataframes ‘*app_data*’ based on *application_data* and ‘*prev_app*’ based on *previous_data*
- The third one, dictionary dataframe, was showing *UnicodeDecodeError*, so decided to open it in Google Sheets for quick reference on the side

Step 2: Handling Missing Values
Step 3: Handling Outliers

There were several null values in both dataframes. To understand things better, I took a simple approach:

- Figure out which columns had **over 40% null values**. These columns were dropped from the dataframes after quick inspection of what these columns contained and if the data was necessary for my EDA
- For columns with **less than 40% null values**, I used imputations based on median and mode (for numerical and categorical data, respectively) to handle missing values
- Since this is financial data, made sense to **use medians or modes** as there were several outliers which could be important for EDA, like ‘income’, ‘no. of children’, etc. Using mean would have created problems for my EDA
- After dropping the columns with over 40% missing values, I inspected remaining columns for major outliers **to understand if missing values could be replaced with median or mean**

Step 4: Standardizing Data, Adding Data Columns For Easier Analysis

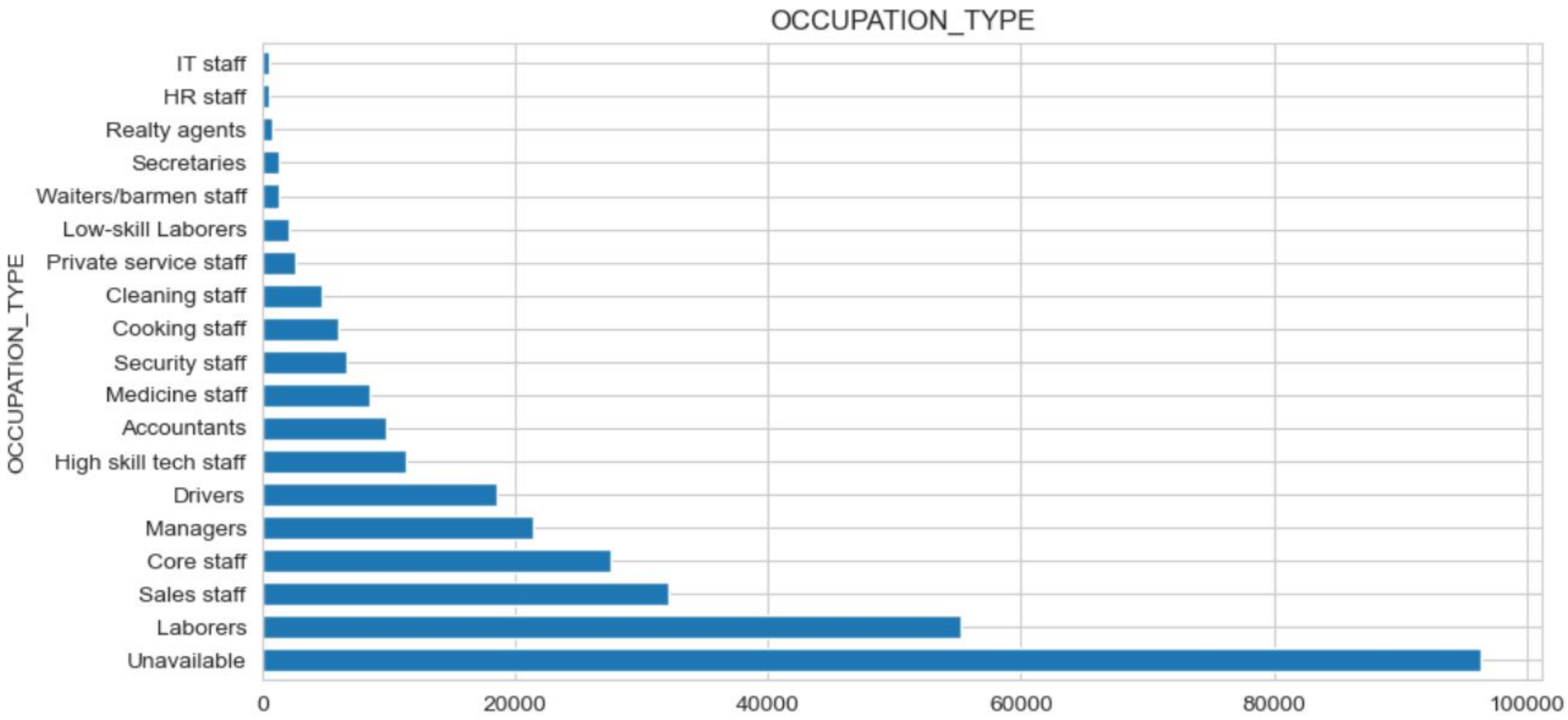
Top notes:

- Most of the *days* data were converted into *years* for better analysis
- Created buckets for several data, like employment years, income amount, and credit amount
- Created a new column for Credit Ratio to help understand if an applicant will have trouble in making payments
- Changed *flags* data from Yes or No to 1 and 0 to ease reading

Step 5: Visualising Data & Drawing Observations

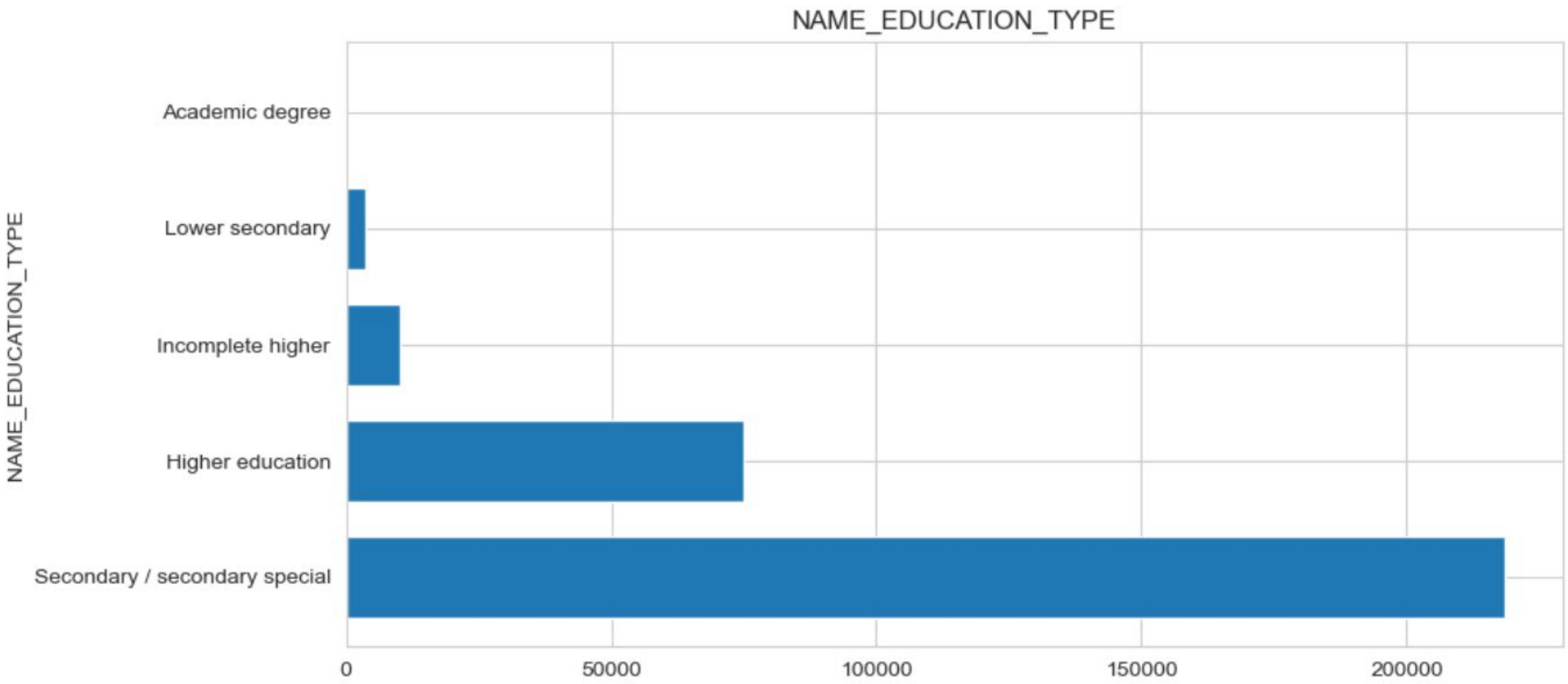
Observation 1:

Loan applicants have a greater female count than male



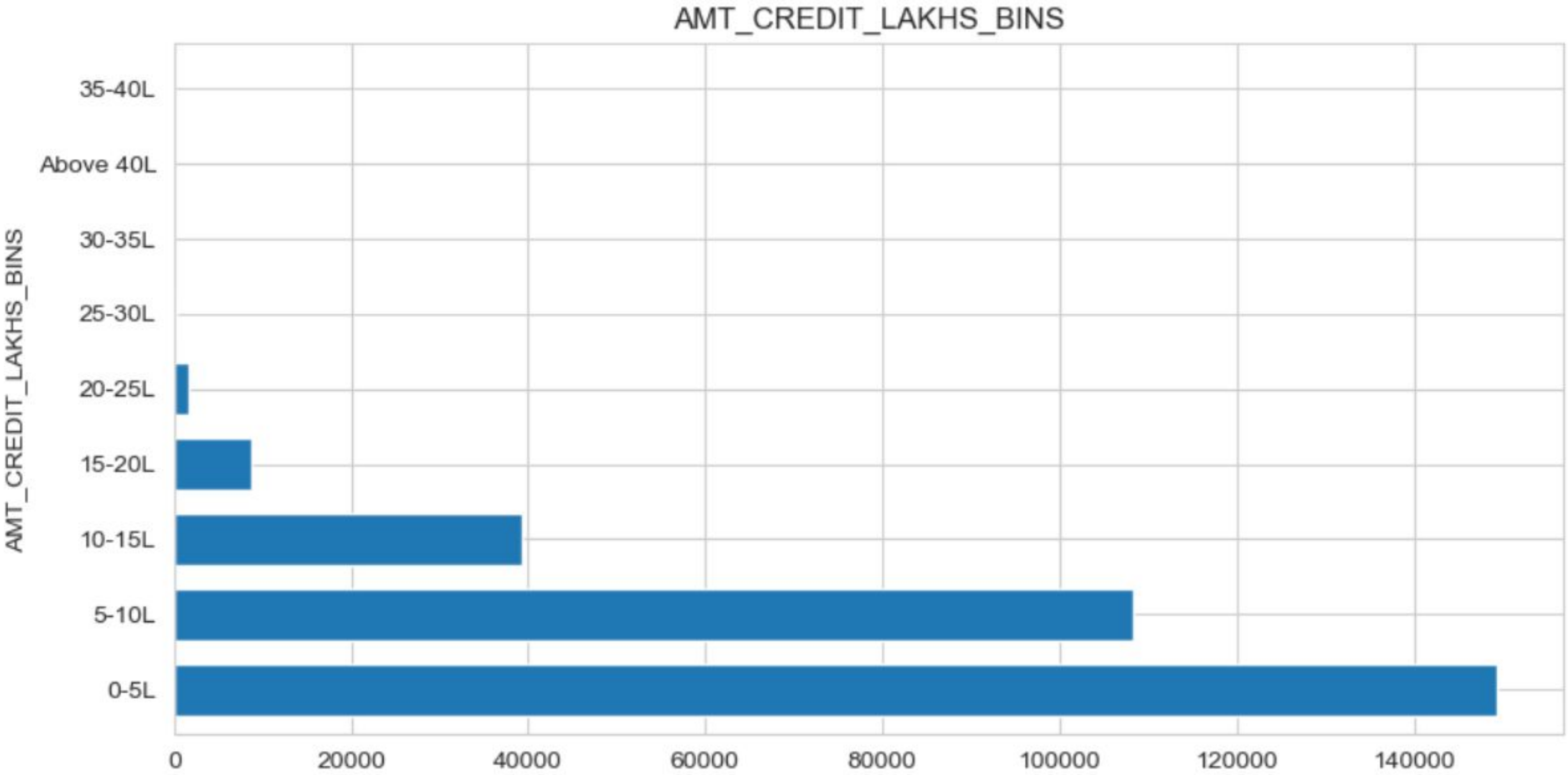
Observation 2:

Laborers are the second-highest applicants. However, most of the data is still unavialble. So, ideally, the bank should figure out the missing occupation as the number is pretty high.



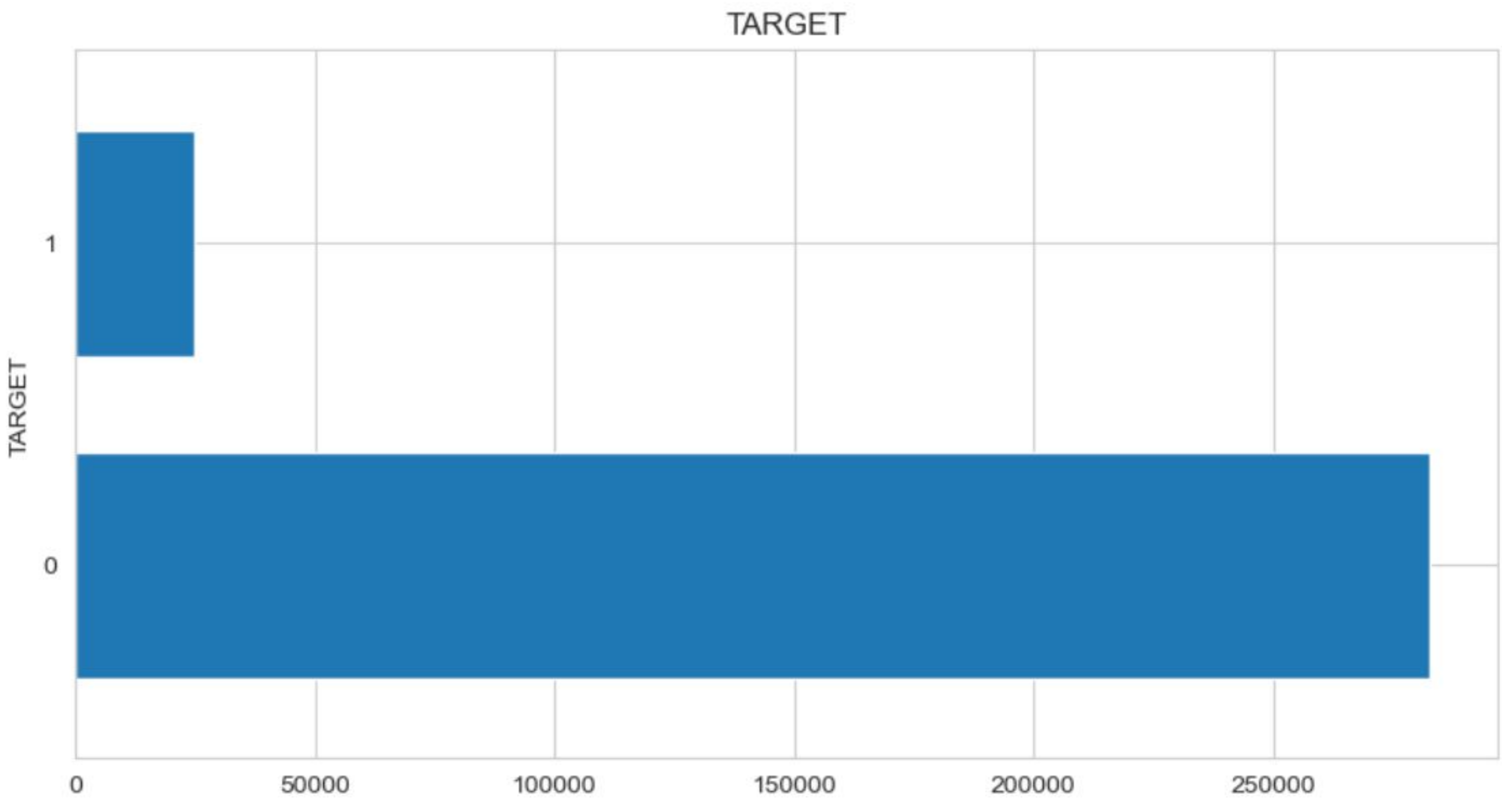
Observation 3:

Applications who have completed Secondary Education/Secondary Special are the highest in number



Observation 4:

Credit bucket of Rs 0-5 lakh have the highest number of applicants



Observation 5:

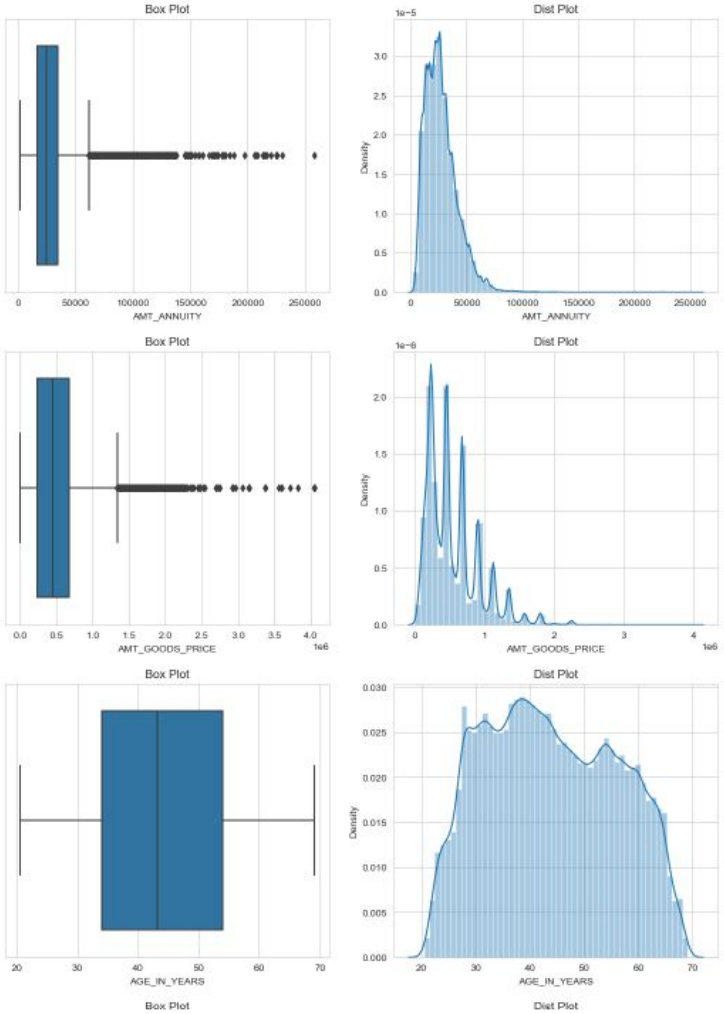
The number of applicants who had payment difficulties are noticeably lesser, between 0 - 50,000

The non-difficulty applicants are higher, well over 250,000

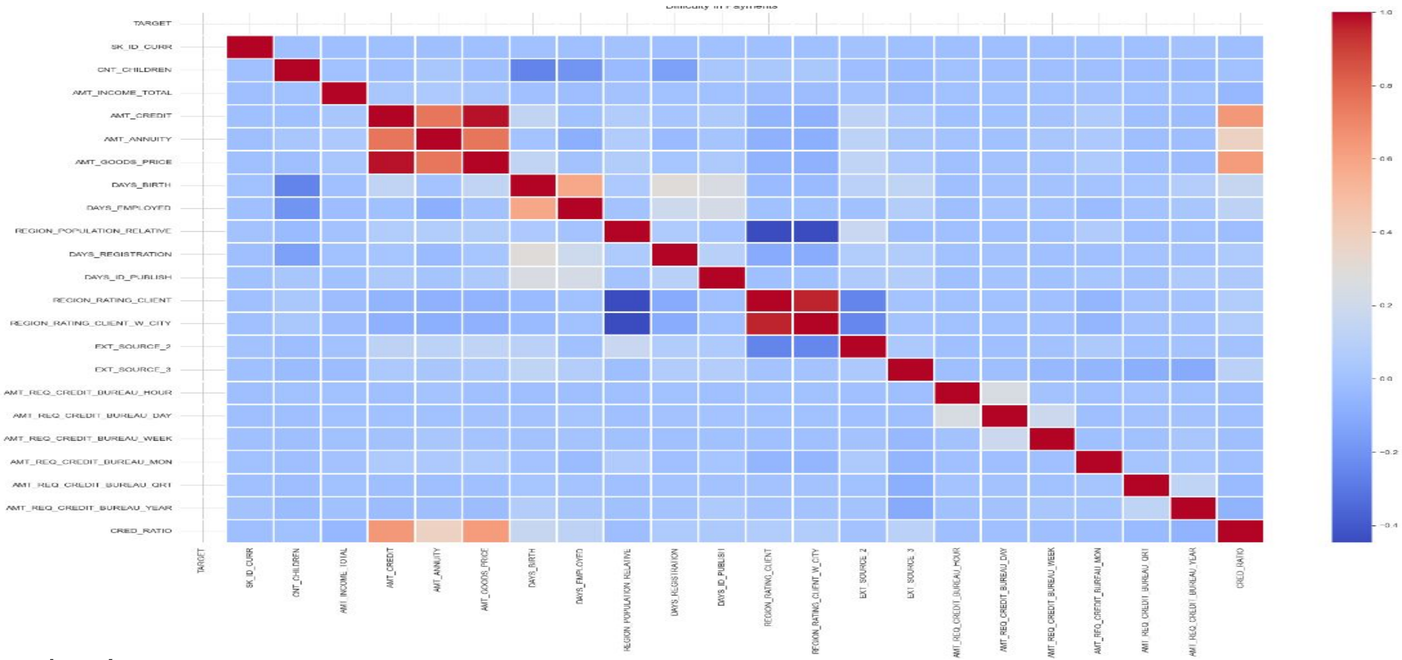
I then ran a loop of box plots and histograms for all numerical data to understand various variables a little better. Here are the observations:

Observation 6 (for all numerical data):

- AMT_ANNUITY: Mostly concentrated between 25,000 to 35,000
- AMT_GOODS_PRICE: Most goods prices are between 0.25 and 0.75
- AGE_IN_YEARS: Most applicants are aged between 35-46
- EMPLOYMENT_YEARS: Strange outlier of a data of 1,000 years of employment, which has to be an error. We should remove it! Otherwise, most applicants are in between 0-20 years
- AMT_INCOME_TOTAL_LAKHS: Most applicants earn less than 10-20 lakh. However, there is 1 outlier at 1200. That might not be an error as some applicant could have an income of 12 crore
- AMT_CREDIT_LAKHS: Mostly concentrated between 2-7.
- CNT_FAM_MEMBERS: 2.5 family members seem to be the highest. I don't think we need to change that
- CRED_RATIO: The most population has a credit ratio between 5 to 10



Correlation (For Payment Difficulty Targets)



Observation /

- Credit ratio has a strong correlation with AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE
- AMT_CREDIT has high correlation with AMT_ANNUITY and AMT_GOODS_PRICE

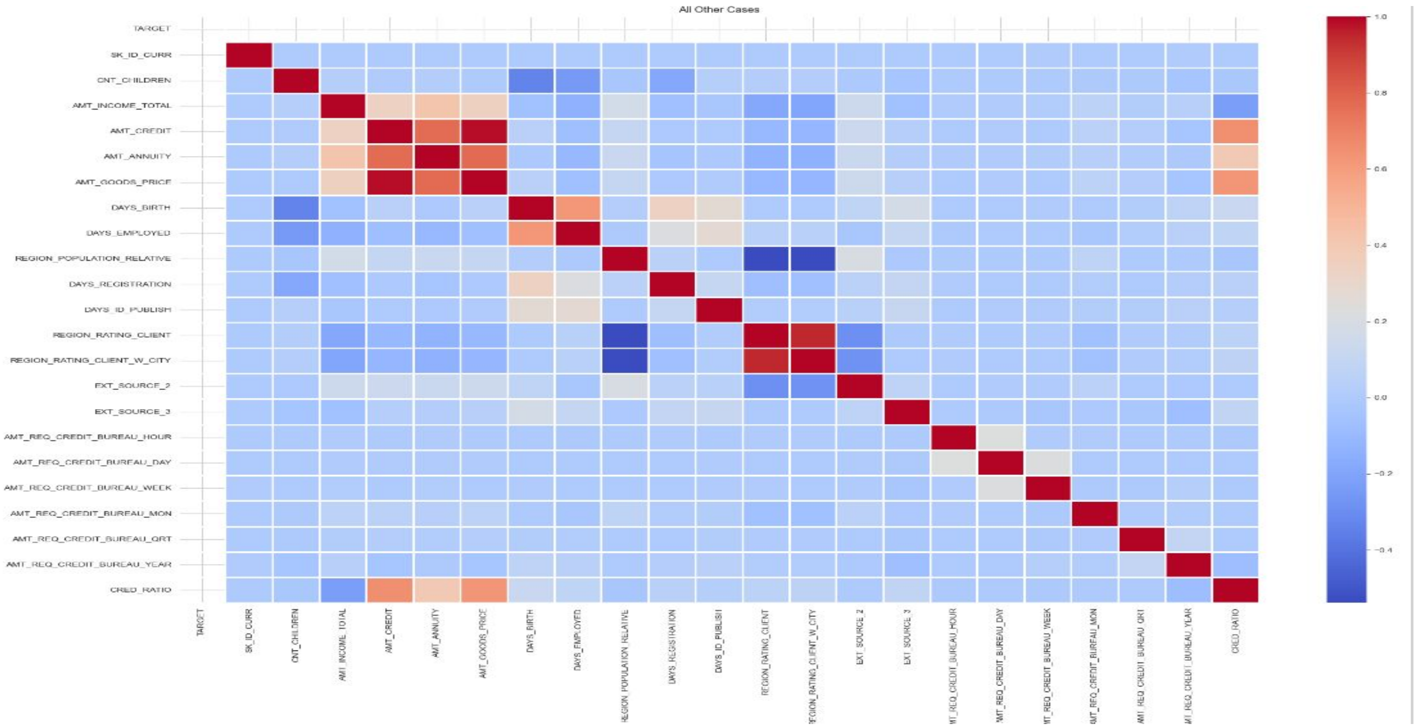
Top 10 Correlation (For Payment difficulty Targets)

	TARGET	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH
TARGET	True	True	True	True	True	True	True	True
SK_ID_CURR	True	False	True	True	True	True	True	True
CNT_CHILDREN	True	True	False	True	True	True	True	True
AMT_INCOME_TOTAL	True	True	True	False	True	True	True	True
AMT_CREDIT	True	True	True	True	False	True	True	True
AMT_ANNUITY	True	True	True	True	True	False	True	True
AMT_GOODS_PRICE	True	True	True	True	True	True	False	True
DAYS_BIRTH	True	True	True	True	True	True	True	False
DAYS_EMPLOYED	True	True	True	True	True	True	True	True
REGION_POPULATION_RELATIVE	True	True	True	True	True	True	True	True
DAYS_REGISTRATION	True	True	True	True	True	True	True	True
DAYS_ID_PUBLISH	True	True	True	True	True	True	True	True

Observation 8 (For Payment Difficulty Targets)

1. CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE , DAYS_BIRTH, DAYS_EMPLOYED, REGION_POPULATION_RELATIVE, DAYS_REGISTRATION , and DAYS_ID_PUBLISH are the top 10 variables.

Correlation (For All Other Targets)



Observation 9

- Credit ratio has a strong correlation with AMT_CREDIT, AMT_GOODS_PRICE, AMT_ANNUITY, and AMT_INCOME_TOTAL
- AMT_CREDIT_TOTAL has a high correlation with AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE
- DAYS_BIRTH also has high correlation with DAYS_EMPLOYED, which is sort of expected due to higher age count

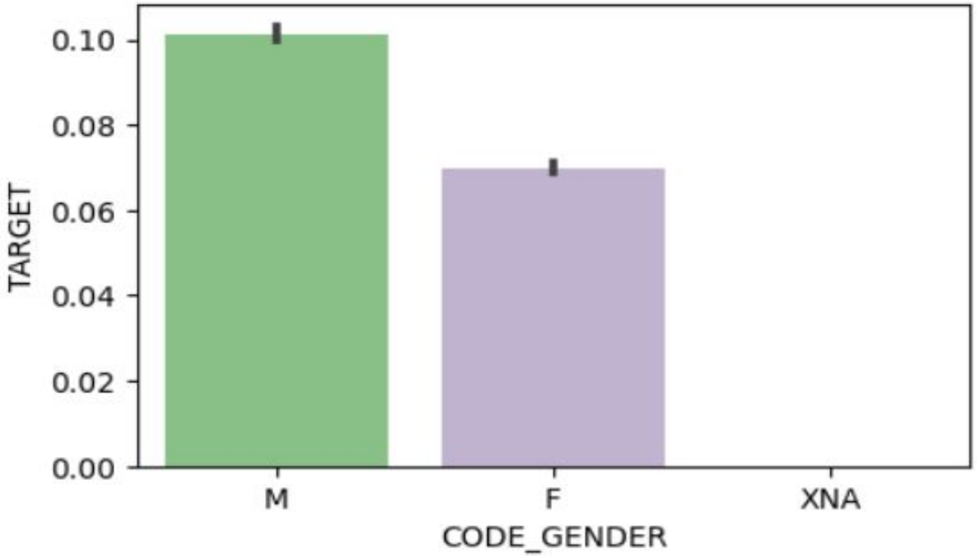
Top 10 Correlation (For All Other Targets)

	TARGET	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH
TARGET	True	True	True	True	True	True	True	True
SK_ID_CURR	True	False	True	True	True	True	True	True
CNT_CHILDREN	True	True	False	True	True	True	True	True
AMT_INCOME_TOTAL	True	True	True	False	True	True	True	True
AMT_CREDIT	True	True	True	True	False	True	True	True
AMT_ANNUITY	True	True	True	True	True	False	True	True
AMT_GOODS_PRICE	True	True	True	True	True	True	False	True
DAYS_BIRTH	True	True	True	True	True	True	True	False
DAYS_EMPLOYED	True	True	True	True	True	True	True	True
REGION_POPULATION_RELATIVE	True	True	True	True	True	True	True	True
DAYS_REGISTRATION	True	True	True	True	True	True	True	True
DAYS_ID_PUBLISH	True	True	True	True	True	True	True	True

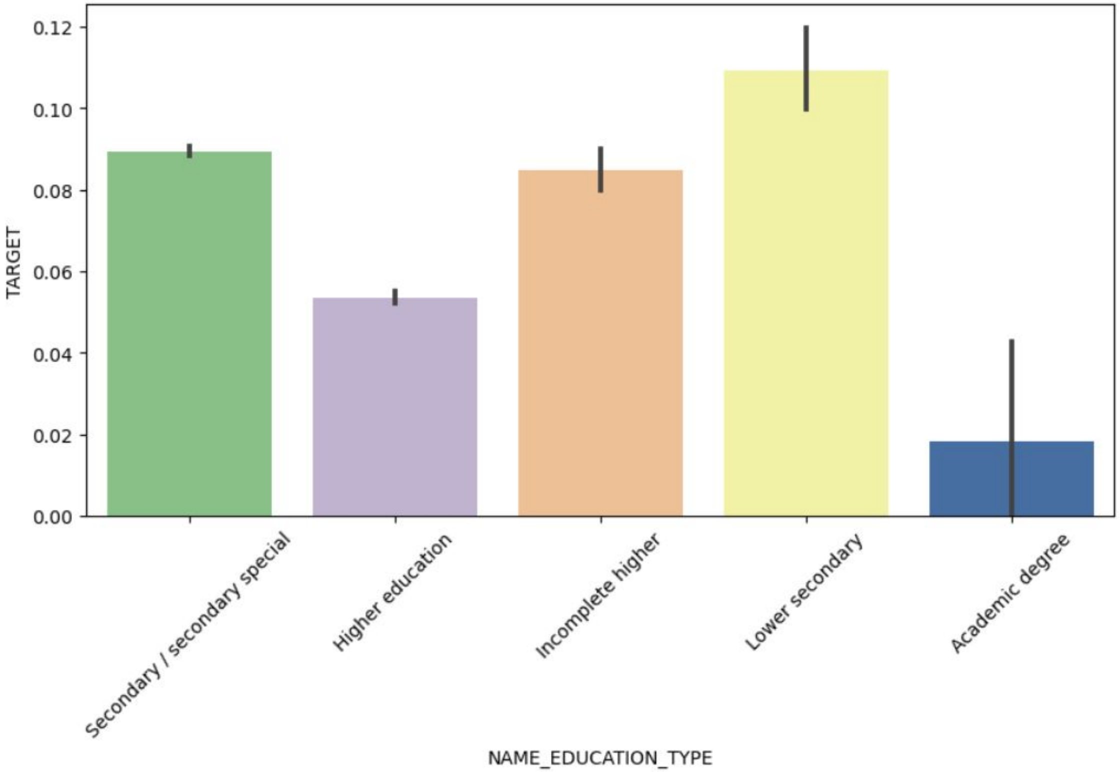
Observation 10 (For All Others TARGET)

1. CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE , DAYS_BIRTH, DAYS_EMPLOYED, REGION_POPULATION_RELATIVE, DAYS_REGISTRATION , and DAYS_ID_PUBLISH are the top 10 variables.
2. This is same as the previous observation.

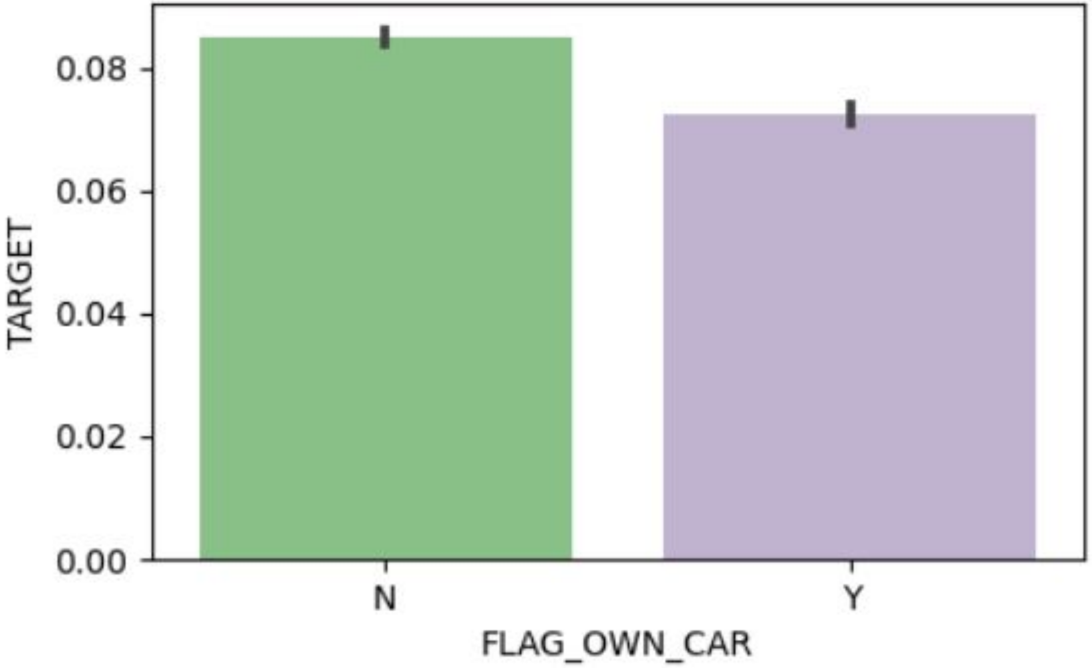
Top Recommendations/Drivers For Biz



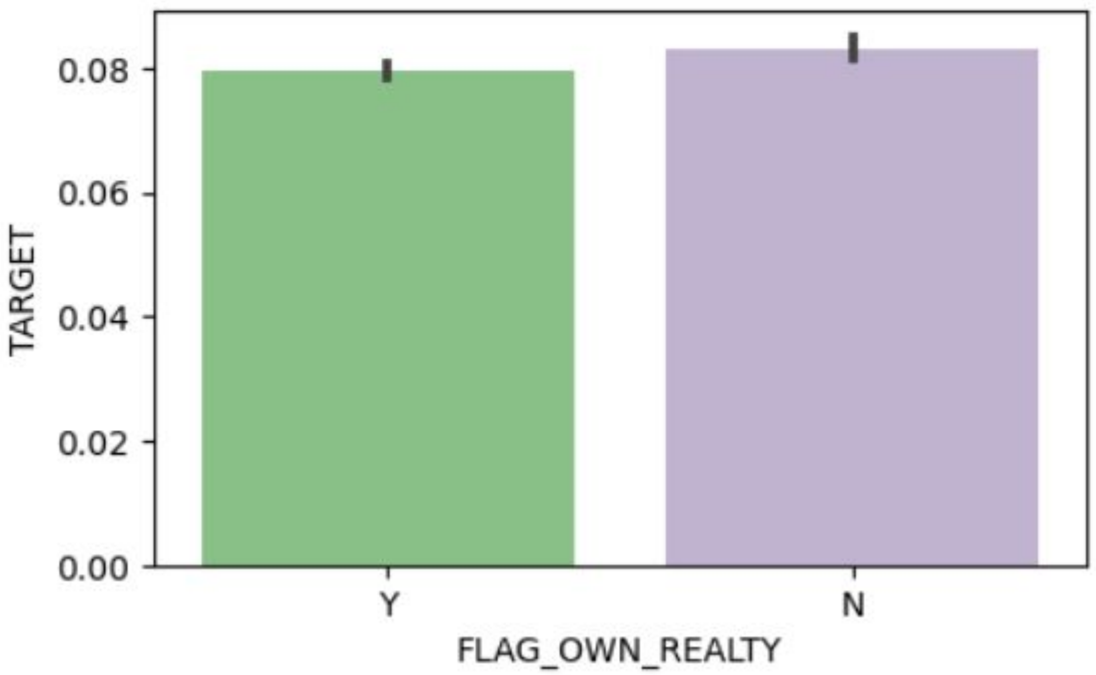
Male applicants have a high score in payment difficulties. Female applicants, on the other hand, are lesser. So, **men have a higher default rate**.



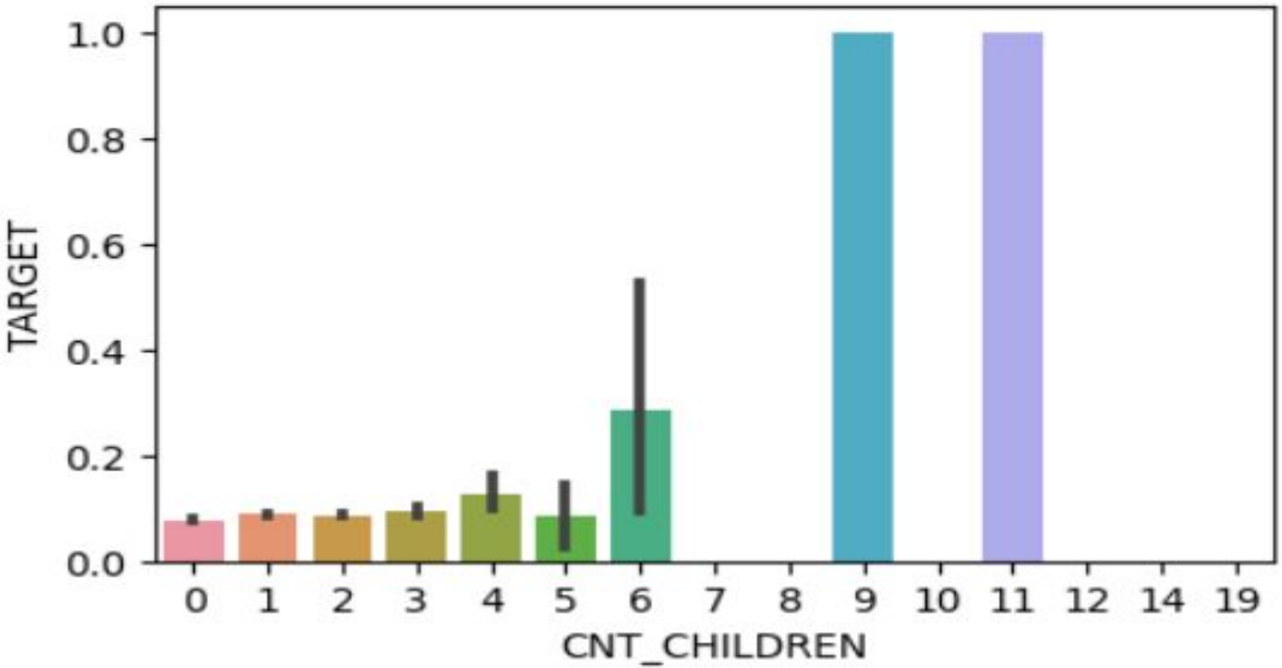
Applicants with **education levels up to Lower Secondary have more difficulties in paying than others**. So, chances are they will default more. If we have to put an order of risk of default, it will be ***Lower Secondary*** > ***Secondary/Secondary Special*** > ***Incomplete Higher*** > ***Higher Education*** > ***Academic Degree***



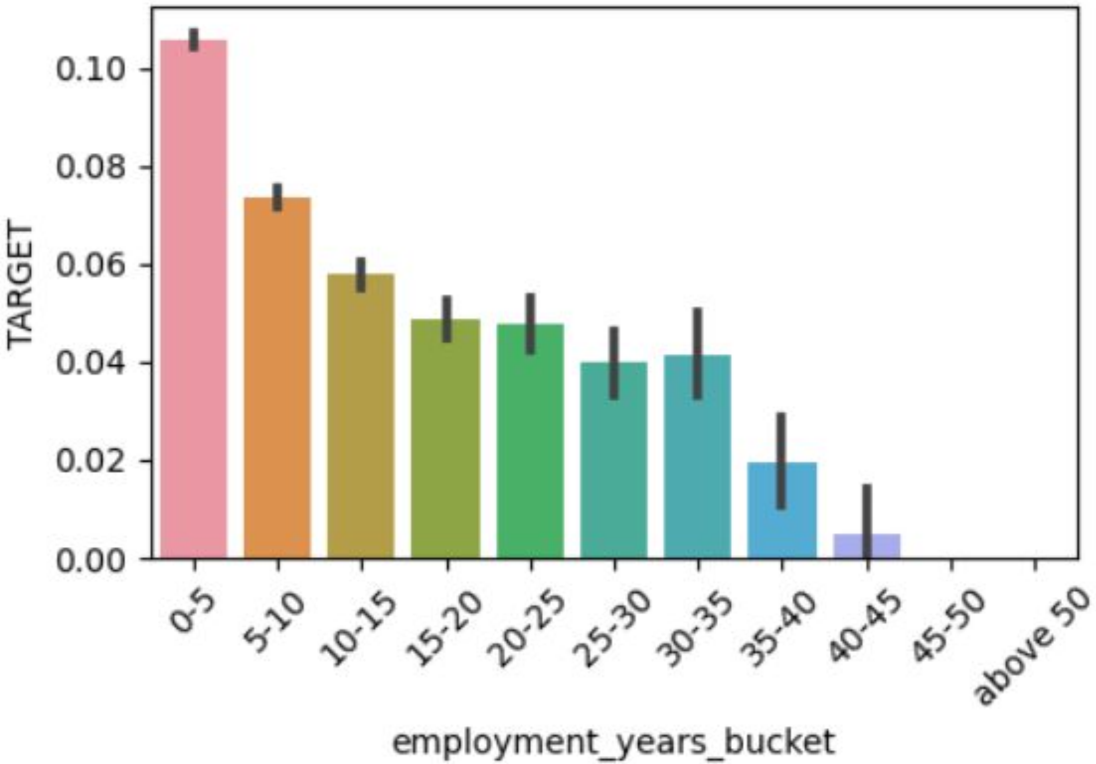
People with no cars are more in number when it comes to difficulty with payments. So, **people with cars stand less chance to be defaulters**.



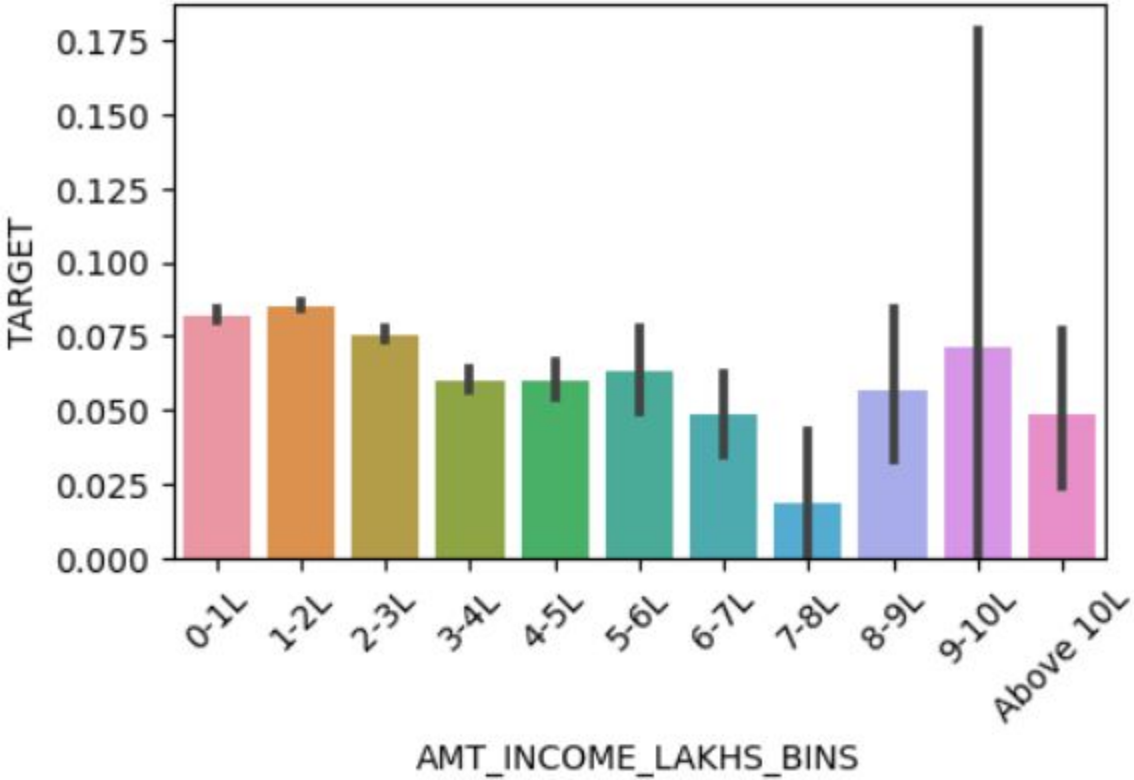
People with no personal real estate are more in number when it comes to difficulty with payments. **So, people with own homes stand less chance to be defaulters.**



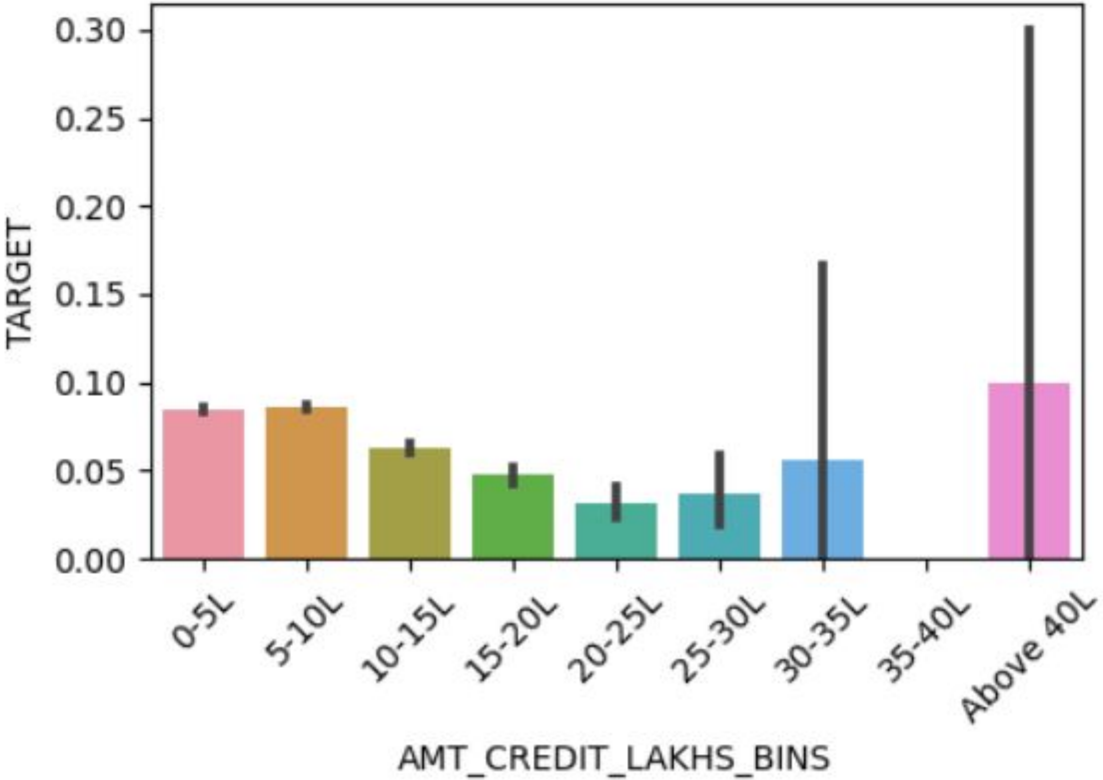
Nothing too surprising here. **More the number of children, more chances of defaulting.**



Applicants with 0 - 5 years of experience has **more difficulty in making payments**.



Applicants with annual income of **1 lakh to 2 lakh have the most trouble** completing payments. **The 7-8L income bracket has the least chances of defaulting.**



Applicants **with over 40 lakh of credit stand a higher chance of defaulting**. Interestingly, **those in the 20-25L bracket stand the lowest chance of defaulting**.

Thank You