EDA Credit Assignment

Presented by: Sonam Tripathi

Highlights

- Problem Statement
- Methodology
- Finding Data Imbalance Ratio
- Analytical results for application data : Univariate Analysis, Segmented Univariate , Bivariate and Multivariate Analysis
- Top Correlation between Target and other variables
- Analysing Previous Data Set :Data cleaning ,Handling outliers, Standardize values
- Merging of both data set
- Top Correlation after merging data sets
- Insights on the basis of previous and current applications data set
- Conclusion

Problem Statement

- Identifying patterns which indicate potential loan defaulters by analysing key influencers(driving factors) behind loan defaults like credit score, income lend, employment status, loan amount , borrowers credit history ,loan term ,debit to income ratio and loan type.

We need to analyse results w.r.t target 0 and target 1.

- Target 1:Defaulter(rejected)
- Target 0: Non-defaulter(all other cases)

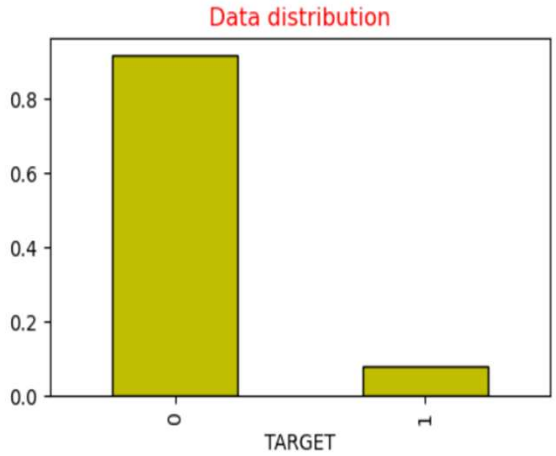
Methodology

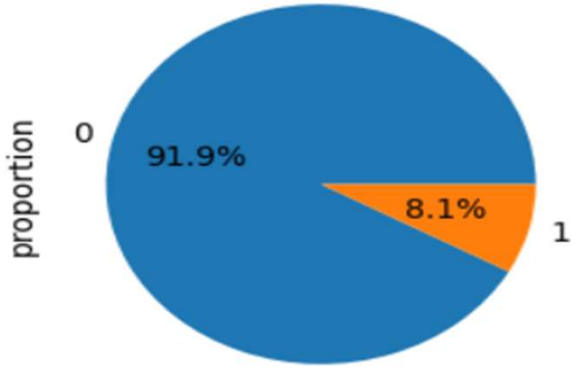
- To analyse data ,we first need to import essential libraries for data manipulation, analysis, and visualization. Then loading the application data and moving further for data cleaning.
- First, we will check basis information related to data like info(),describe(),shape then move further and finding the missing values in columns.
- Dropping columns with more than 40% of missing values and those were not affecting our analysis also.
- After dropping, imputing missing values by mean, median and mode if column is numeric then can choose mean or median depends on outliers and for categorical column ,imputing the values with mode.
- For XNA and XAP values ,we can impute with Unknown or Missing.
- After imputation, we will check data types of columns .if they are not in right type then typecast them into correct format like categorical columns should be of object type, numeric columns be in int /float type and feature engg. columns be of category type.

Methodology

- Handling Outliers : Use binning and capping as per requirement ,it will reduce the impact of outliers.
- Standardising the data : Some of columns have negative values like age .age cannot be negative so converting those kinds of values.
- Finding top correlation w.r.t target variable.
- Segmenting data frame w.r.t target 0 as non-defaulters and target 1 as defaulters .
- Finding data imbalance ratio.
- After this moving forward for finding insights from data so first performing the categorical and numeric univariate analysis.
- Then move further for segmented univariate analysis ,bivariate and multivariate analysis.
- Performing same steps for previous data set and merging previous data with application data for better insights .
- Comparing the results of previous data ,we can see clear insights and can conclude some parameters for considerations.

Data Imbalance Ratio



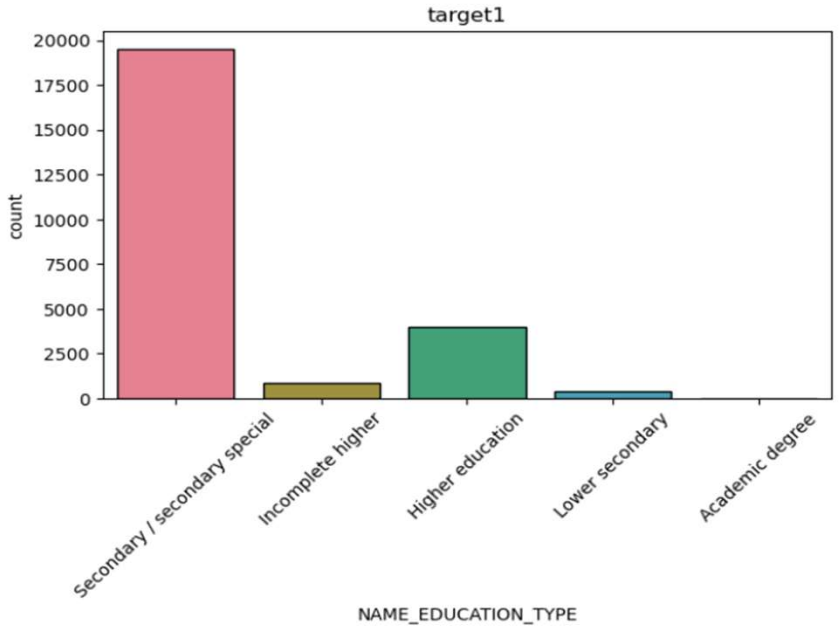


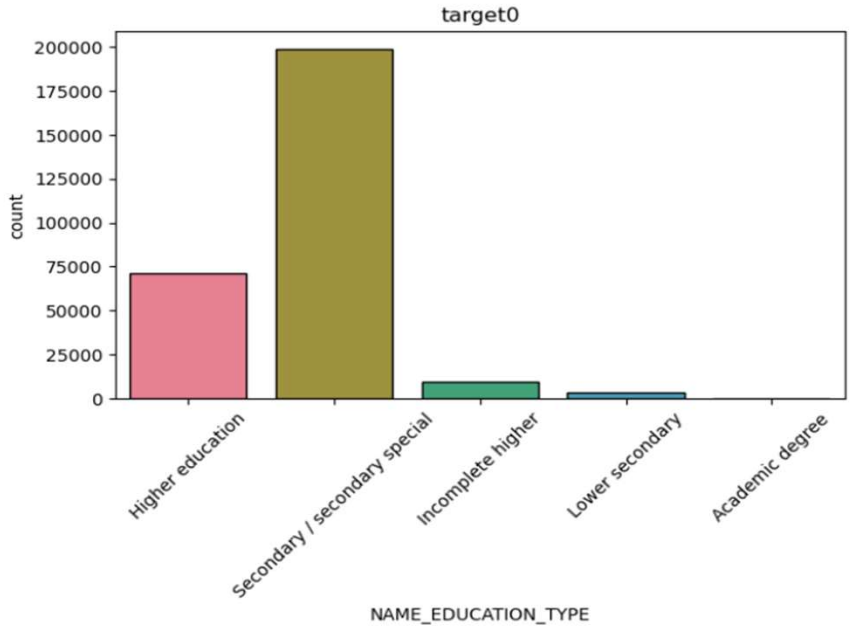
- We can see ,data is highly imbalanced that means number of non-defaulters are more than number of defaulter.
- so now we will segment our data w.r.t TARGET as target0 and target 1 to analyze the result who is going to be defaulter in each category.
- for target 0 ,percentage is 91.9 % and for target 1 it is 8.1%.
- As data is highly imbalanced so results may become biased.
- Target 1 : defaulter
- Target 0: non-defaulter

Univariate analysis : Education type

Univariate analysis w.r.t target 0 ,target 1:

- We can see that clients with education secondary /secondary special are more likely to be defaulted as their percentage is more in target1 data set as compered to target0 data set.

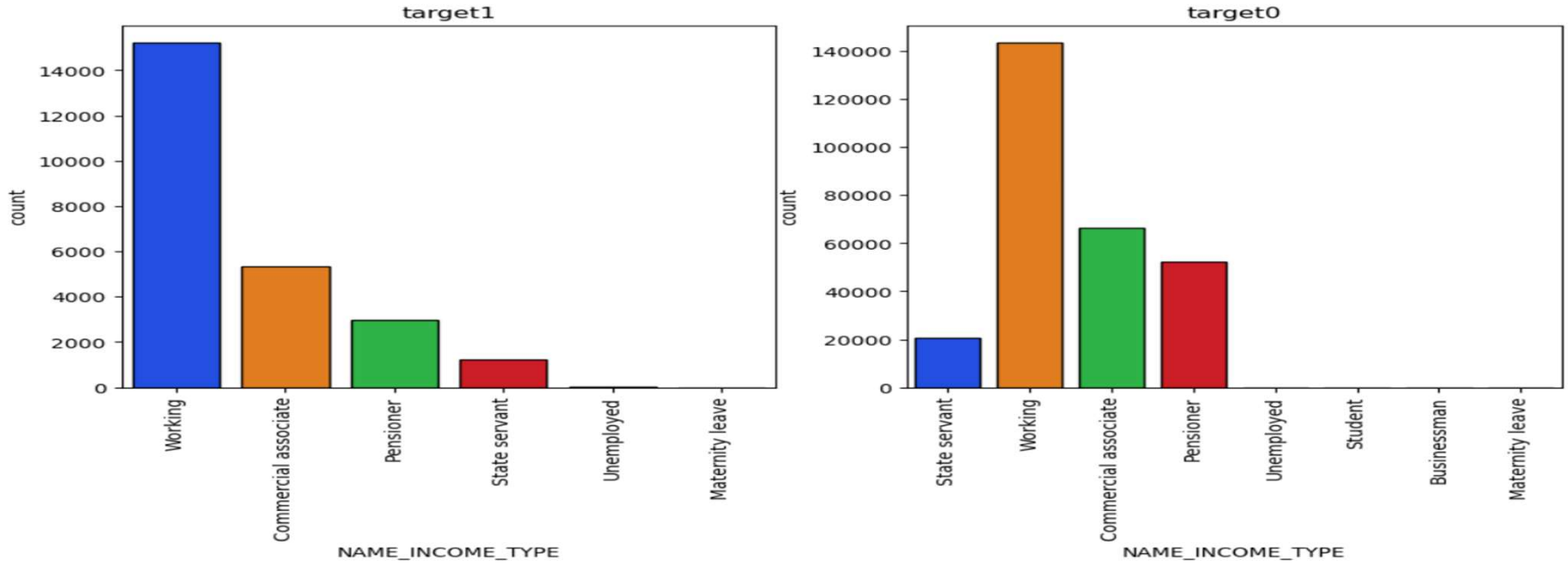




Univariate analysis : Income type

Univariate analysis w.r.t target 0 ,target 1:

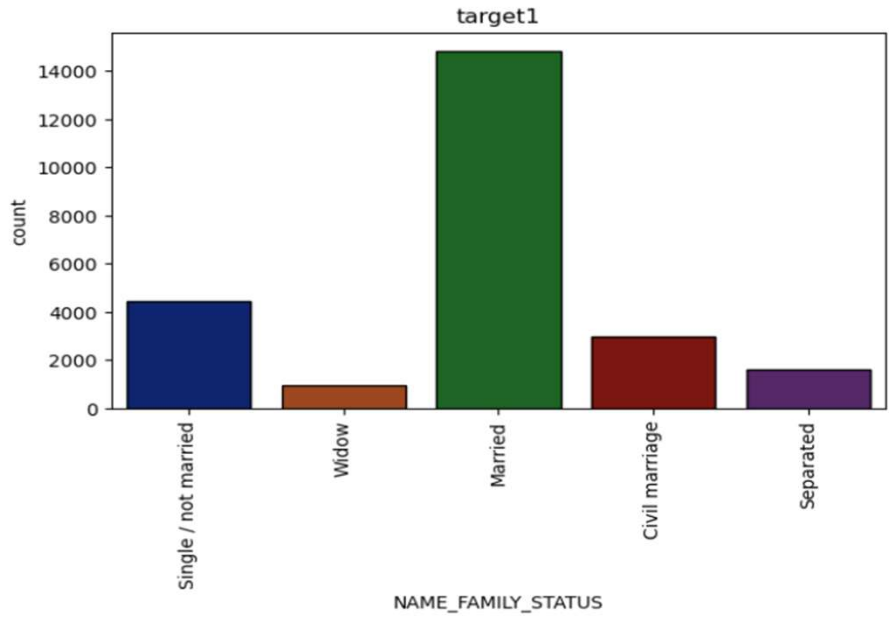
- Working income type clients are more in number to take loans.
- they are more likely to be defaulted as their percentage is more in target1 data set as compered to target0 data set.

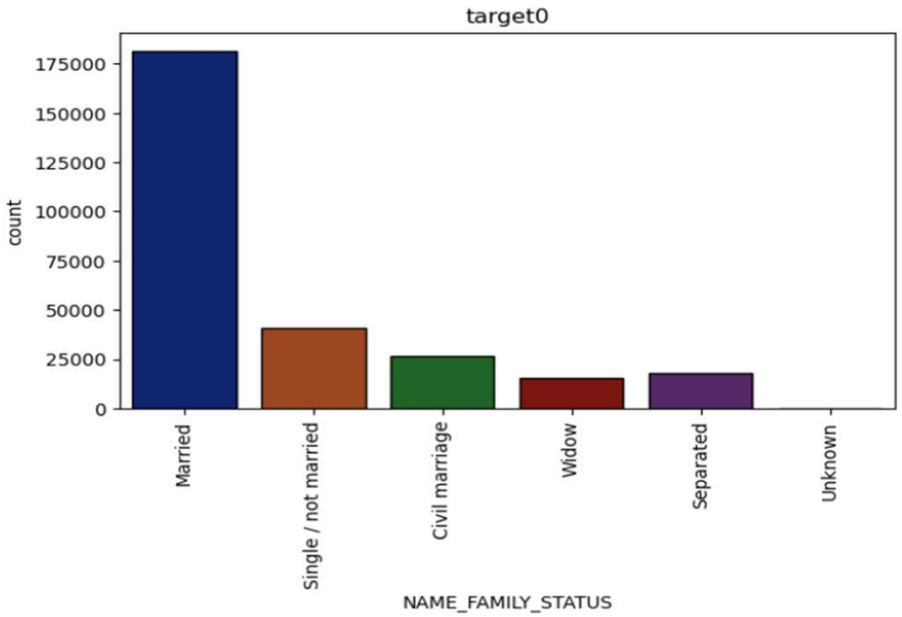


Univariate analysis : Family status

Univariate analysis w.r.t target 0 ,target 1:

- We can see that married clients are more likely to take loans and also more likely to be defaulted as their percentage is more in target1 data set as compered to target0 data set.

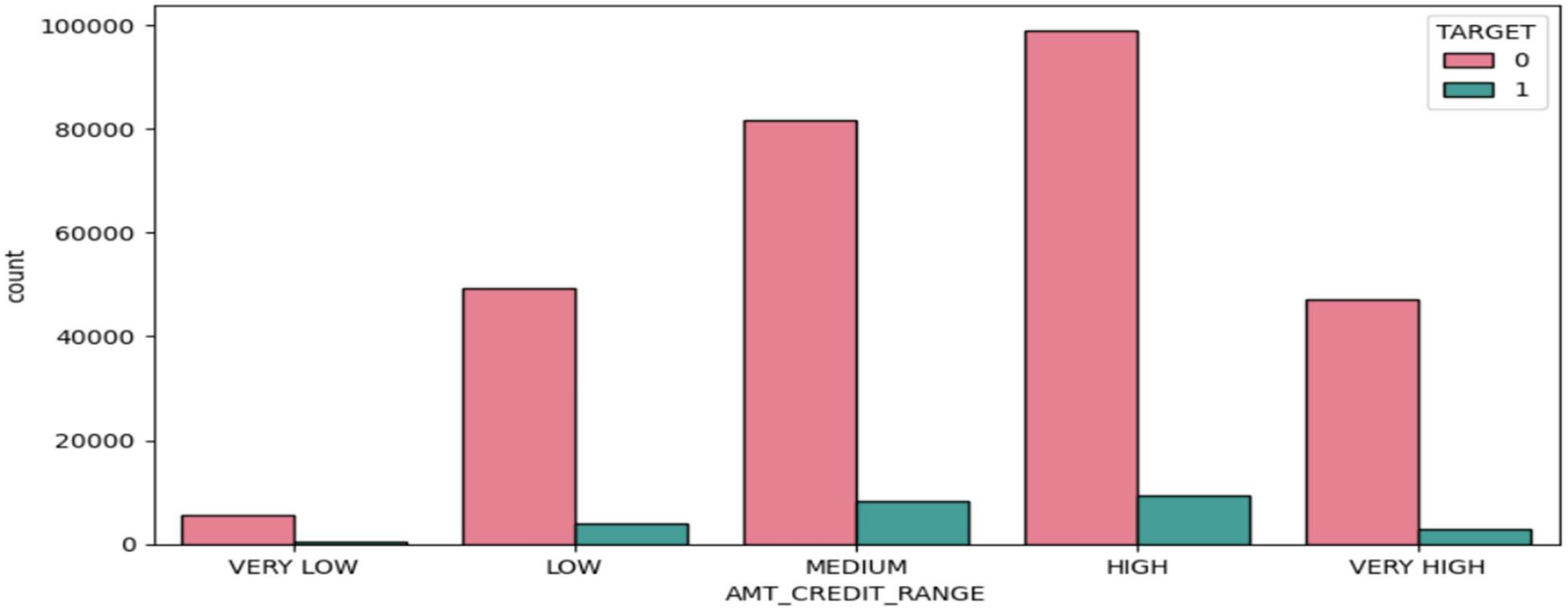




Bivariate analysis : Amount credit range

Insights:

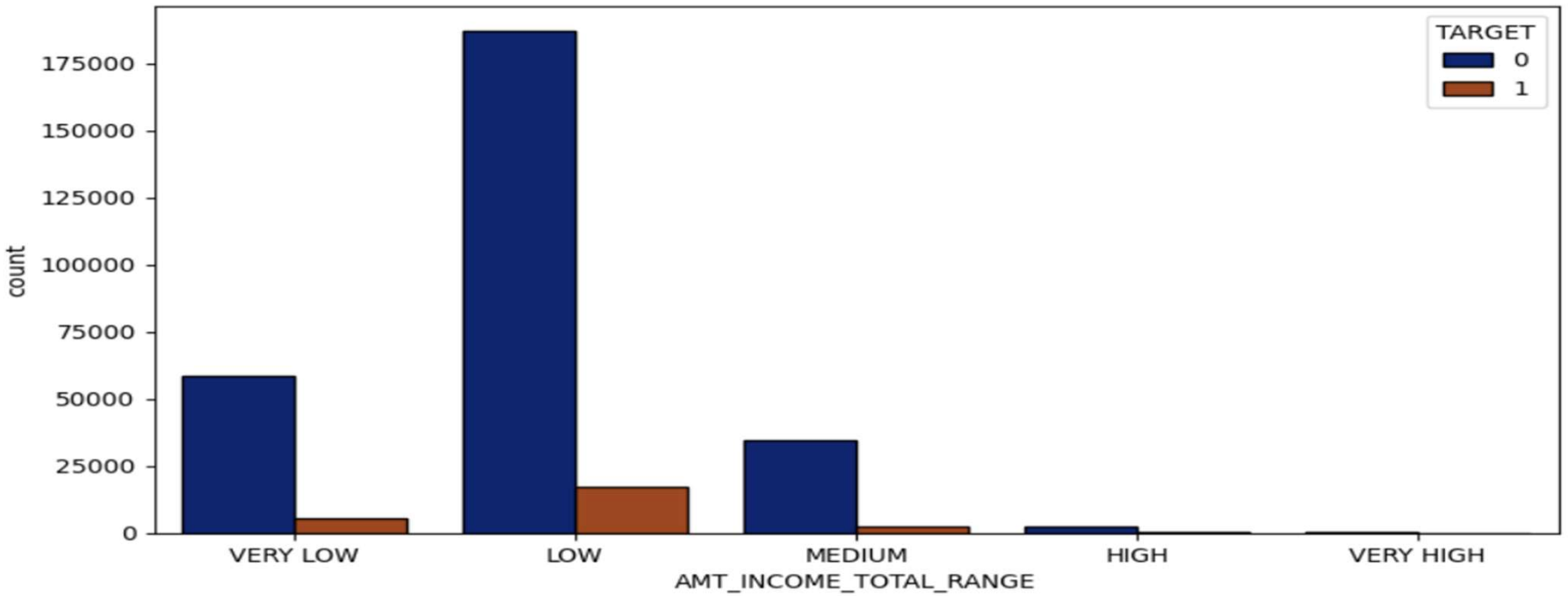
- Clients having HIGH credit range are more.
- Clients with HIGH & MEDIUM credit are likely to be defaulters than others.



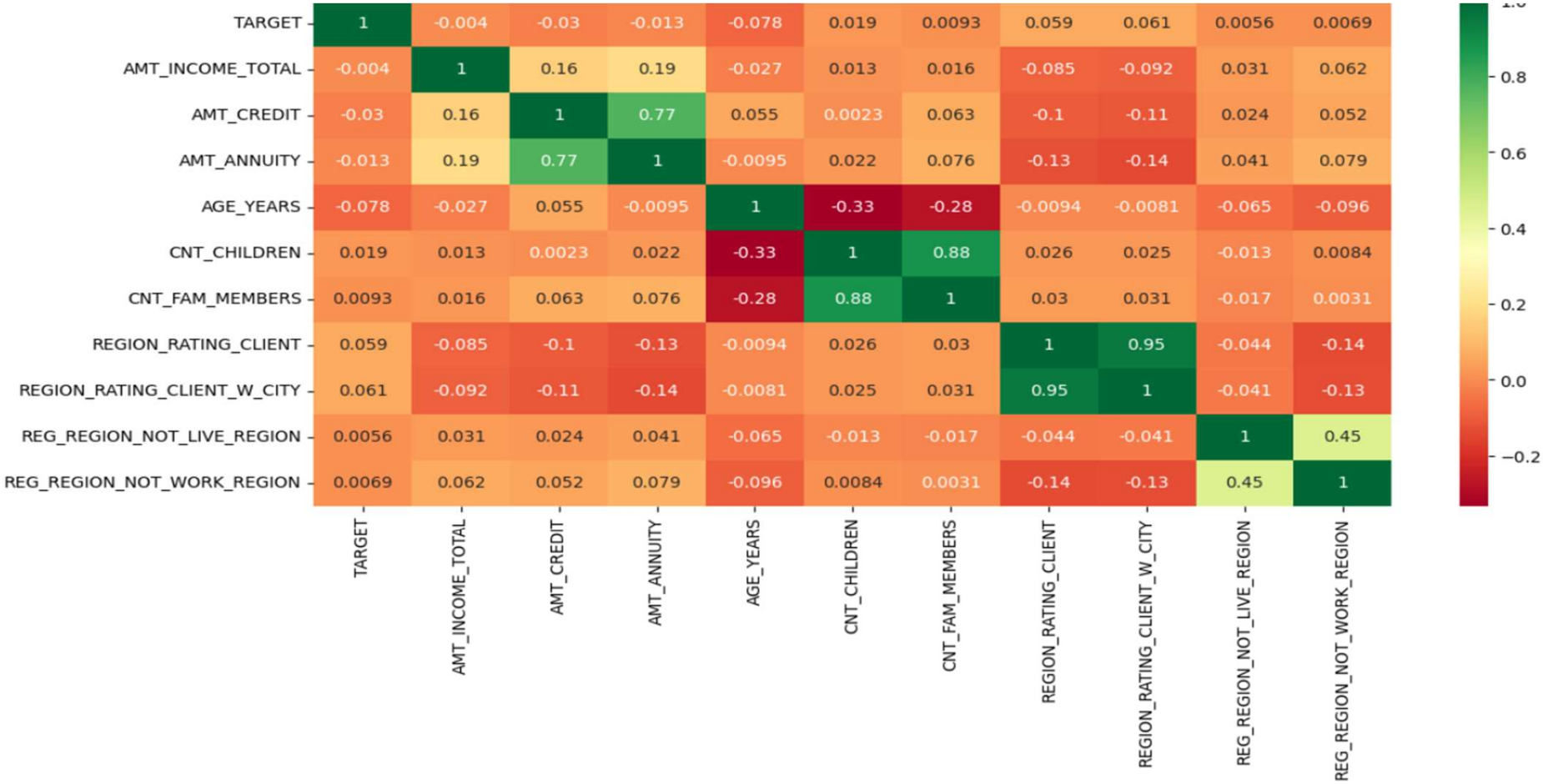
Bivariate analysis : Amount income range

Insights:

- clients with low-income range are likely to be defaulted as low-income range has more value.



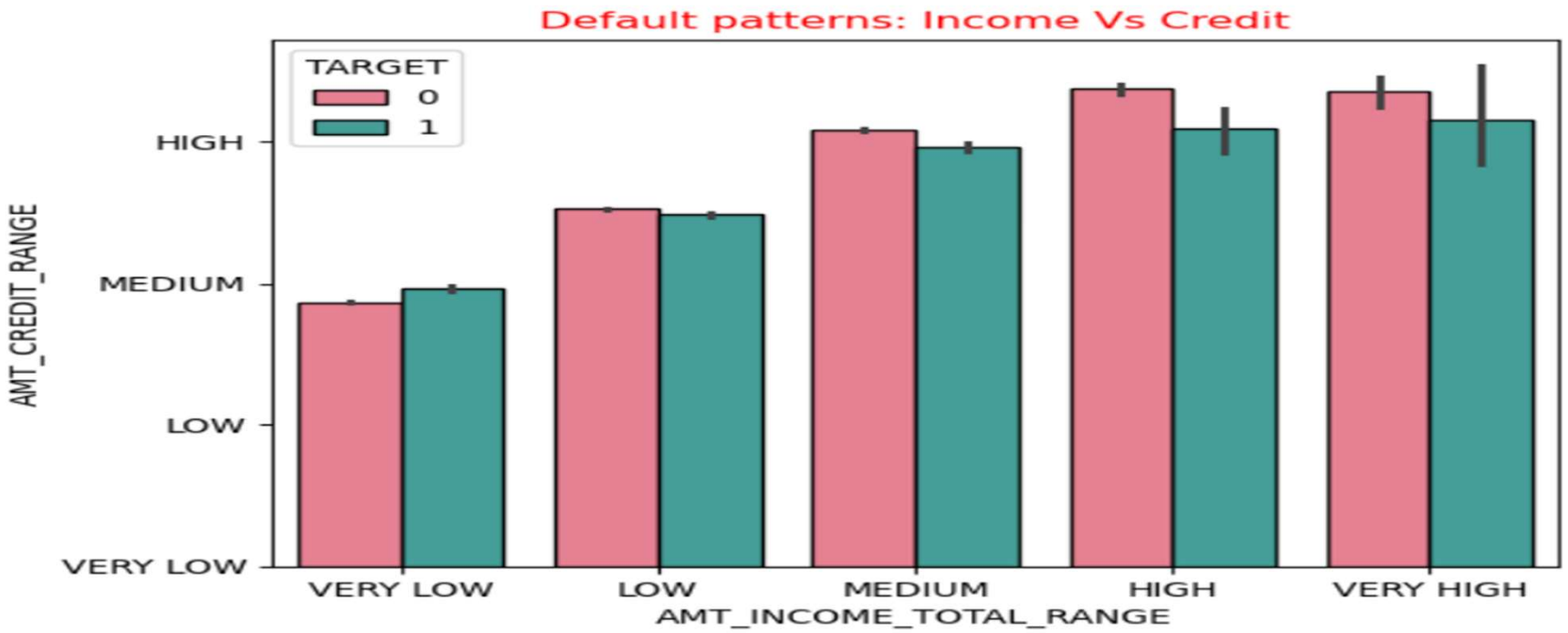
Top Correlation for Application data



Multivariate analysis: Income Vs Credit Vs Target

Insights:

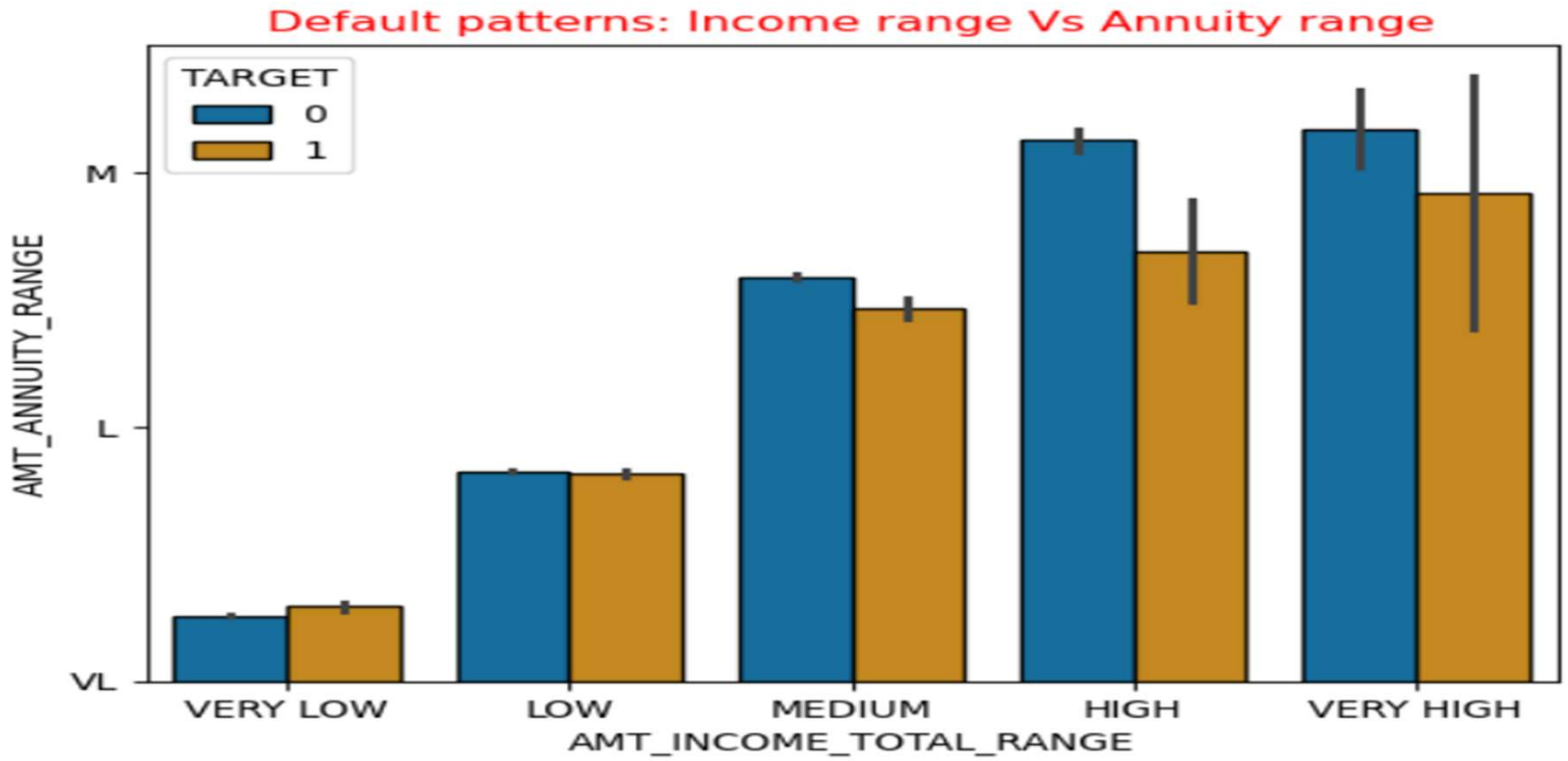
- We can see clients with higher income have higher credit also as income increases credit tends to increase.
- Clients with very low income range are likely to be defaulters.



Multivariate analysis: Income Vs Annuity Vs Target

Insights:

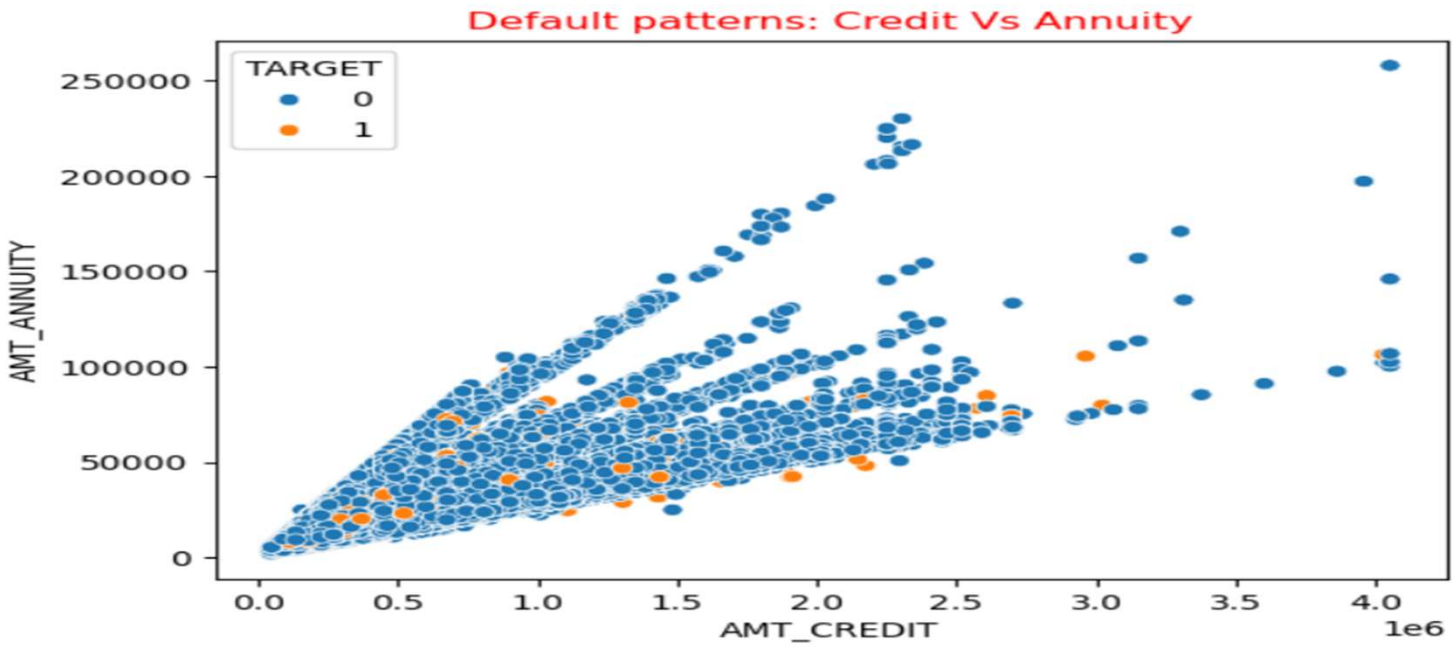
- Here we can observe that clients who have lower income range are likely to be defaulter.
- As income range increases, annuity tends to increase.



Multivariate analysis: Credit Vs Annuity Vs Target

Insights:

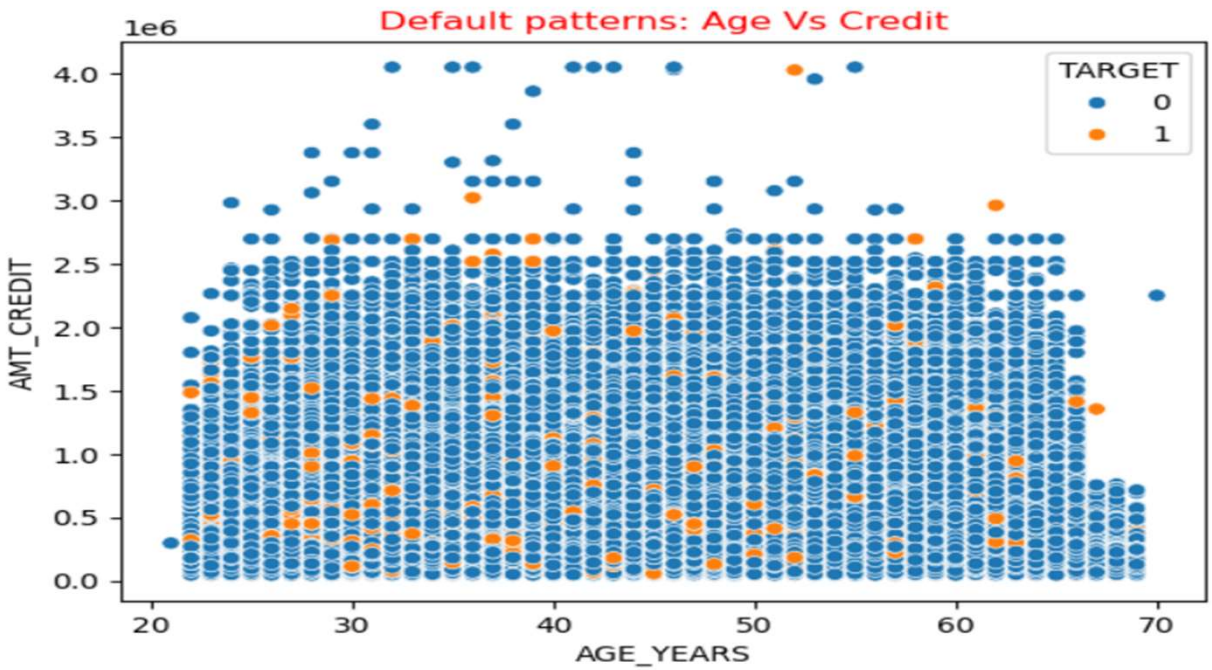
- We can see positive correlation between AMT_ANNUITY and AMT_CREDIT. if AMT_CREDIT increases then AMT_ANNUITY also increases.
- 2. We can see data is imbalanced as there are larger number of loan approvals than rejections.



Multivariate analysis : Age Vs Credit Vs Target

Insights:

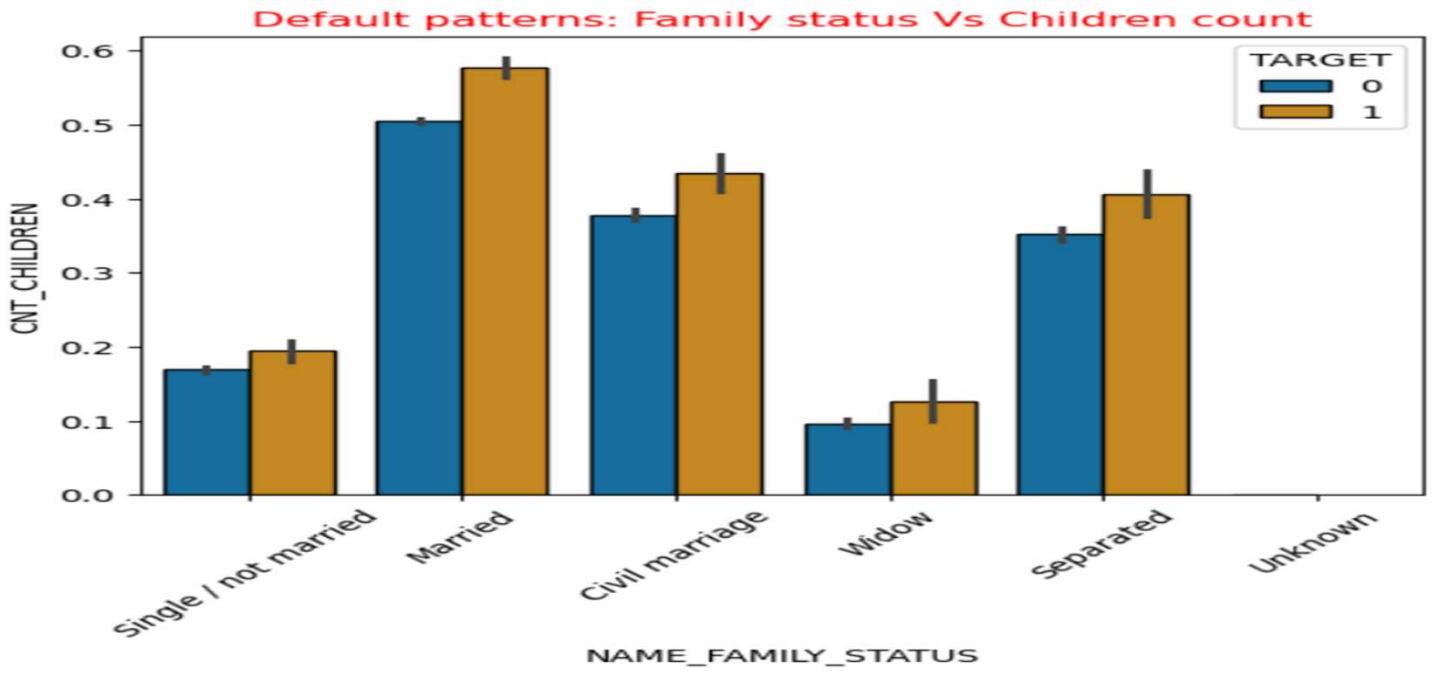
- Data points are scattered between 20 to 70 years of age that means clients are in range of 20 to 70 years.
- We can see that clients in age range of 25 to 45 are taking more credit amounts.



Multivariate analysis: Family status Vs Count children Vs Target

Insights :

- We can clearly see that married clients with more children are likely to be defaulters.
- Separated clients having more children are also .

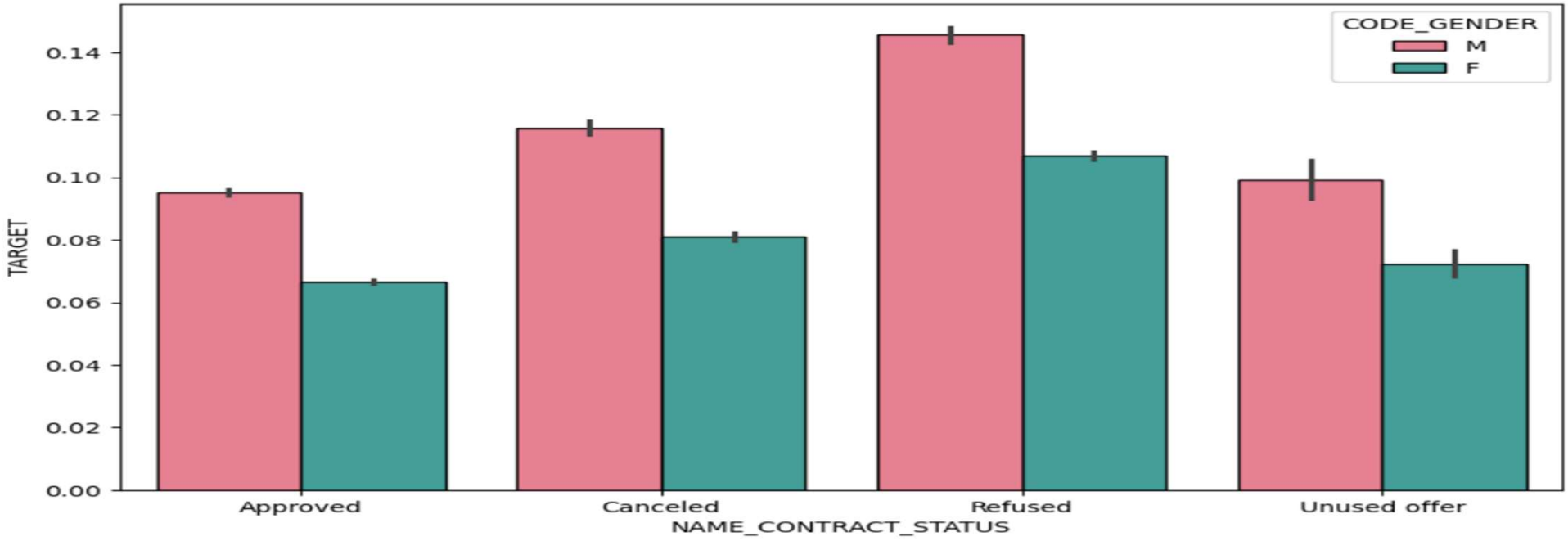


Top correlations after merging application data and previous data

Multivariate analysis of Name contract status Vs Gender Vs Target

Insights:

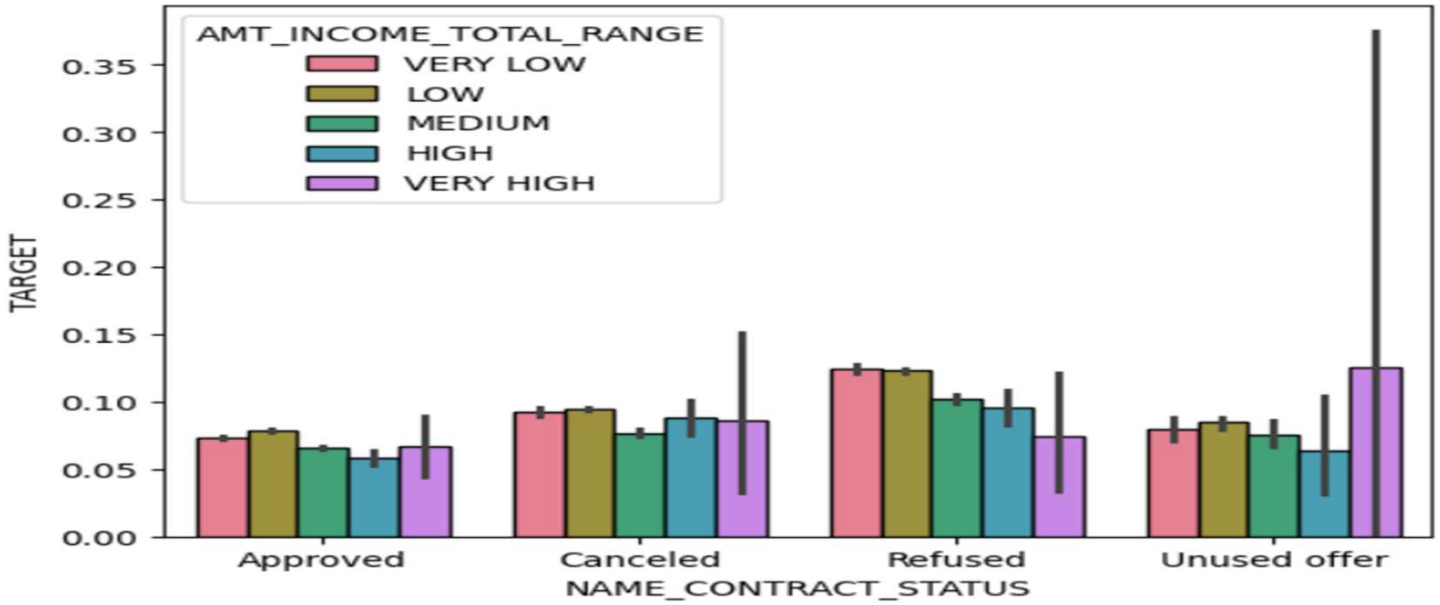
- It is clearly visible that previously male's applications are more refused than approved . That means males are most likely to be defaulter .their rejection ration is high as compared to females.
- Male ratio is high in every category.



Multivariate analysis of Name contract status Vs Income range Vs Target

Insights:

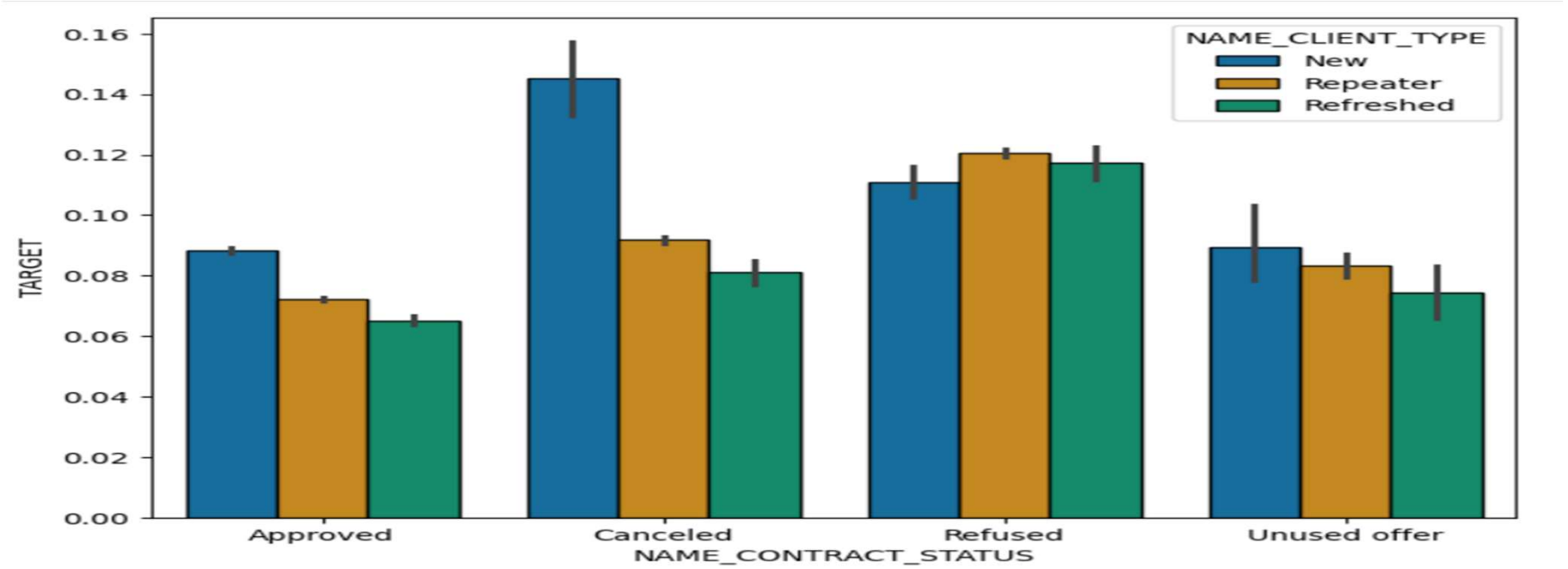
- We can see that clients who are previously canceled are in low income range. the percentage of low income range category is more than others. Clients who are previously rejected are in very low and low income range.
- A very high income range clients previously cancelled their loan at different stage of process(unused offer).



Multivariate analysis of Name contract status Vs Client type Vs Target

Insights:

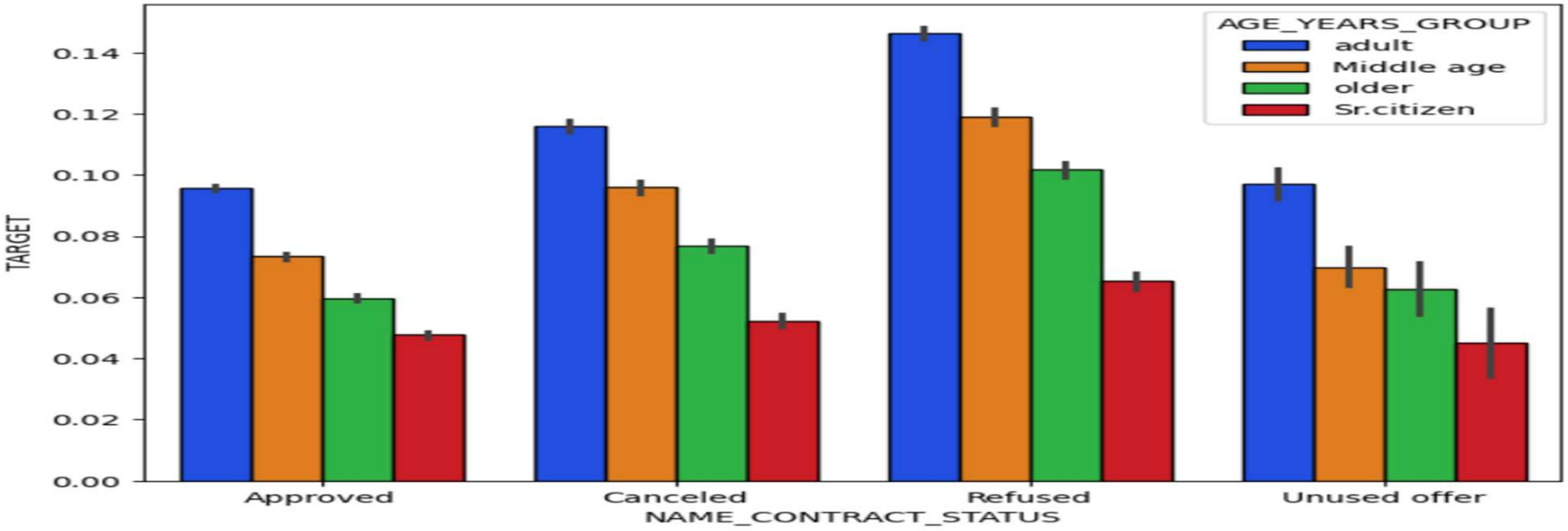
- In previously refused status, repeaters are more in numbers than others that means repeaters are likely to default.
- Mostly new client applications are approved.
- New clients have more cancelled application also.



Multivariate analysis of Name contract status Vs Age group Vs Target

Insights:

- Previously adults loan applications are refused more than others , it may be because adults are more in number who are applying for loan.
- Sr.citizen applications are less refused as applications are also less in this category.
- In each category, adult ratio is high than others.



Conclusion

As data is highly imbalanced so analysis can be biased . As per data given here are some observations:

- Clients with lower income group are likely to be defaulters. In other words, we can say that lower income salary people can not repay loans easily , they will face some difficulty.
- Banks should check those clients properly who were previously defaulted.
- Clients with higher education are less likely to default so bank should consider their application more as compared to secondary education people.
- Married clients with more number of children tends to default as per analysis. Banks should check clients who have more number of children.
- Banks should focus on sanctioning revolving loans as cash loans have more number of defaulters.
- Banks should focus on new clients more rather than repeaters.
- Clients in adult age group are likely to default so we can predict that may be at young age they are not settled much .after age 35 people are more likely to be settled .

Thank You