EDA CREDIT CASE STUDAY

BY

GAURAV SAINI

Purpose

There are 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.

Analysis of the data set has been done in python on a Jupiter Notebook

Approach

Steps :

→ Importing Module

→ Reading the Dataset into Pandas DataFrame

→ Checking missing values and handling them

→ Checking outliers, data imbalance, ration of imbalance

→ We have divided the features into small segments and analyzed segment wise using a smaller dataframe containing only relevant categories.

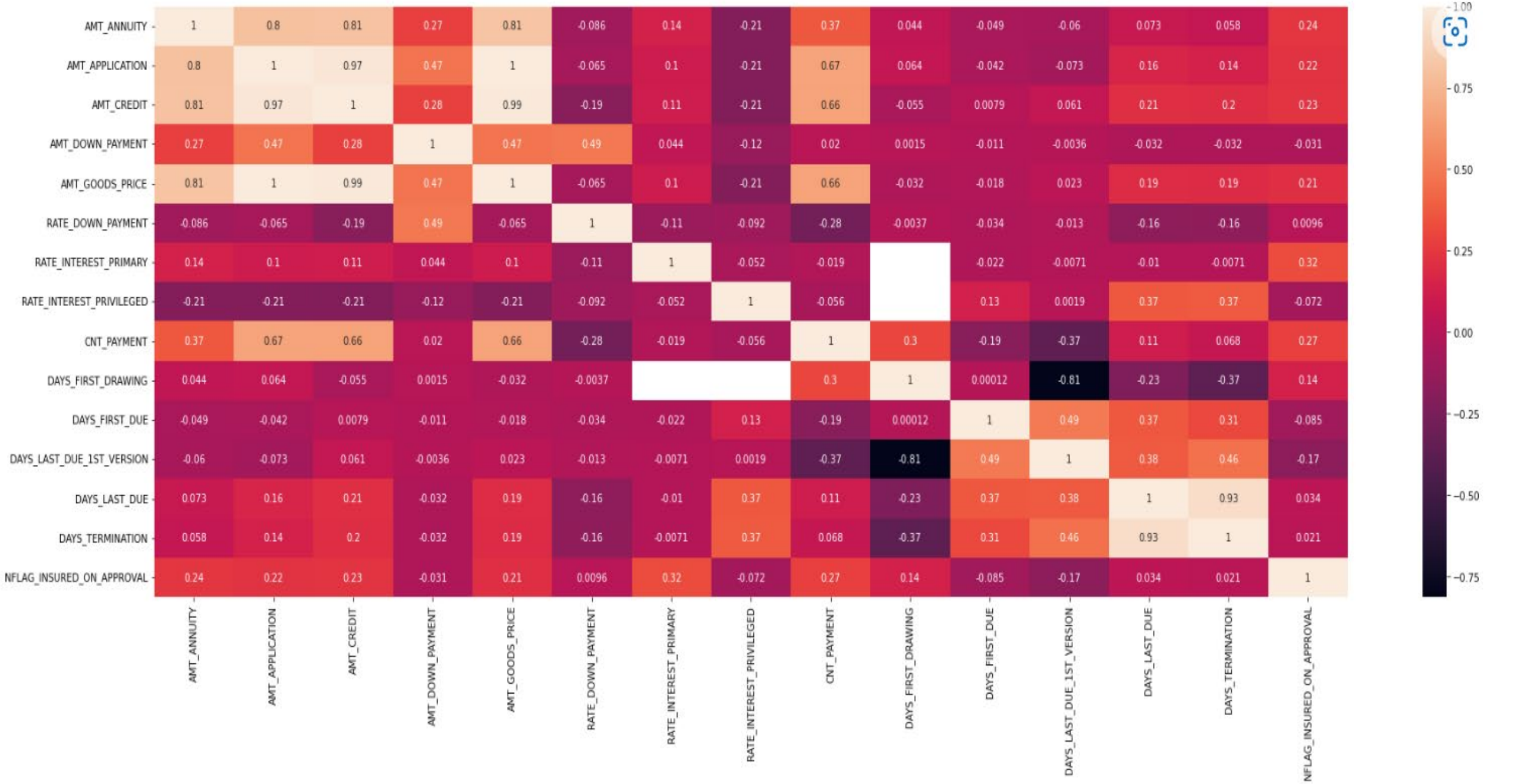
→ Plots and percentage wise Defaulters calculation are done segment wise as well.

Correlation Between Numeric Feature of Previous Application Data

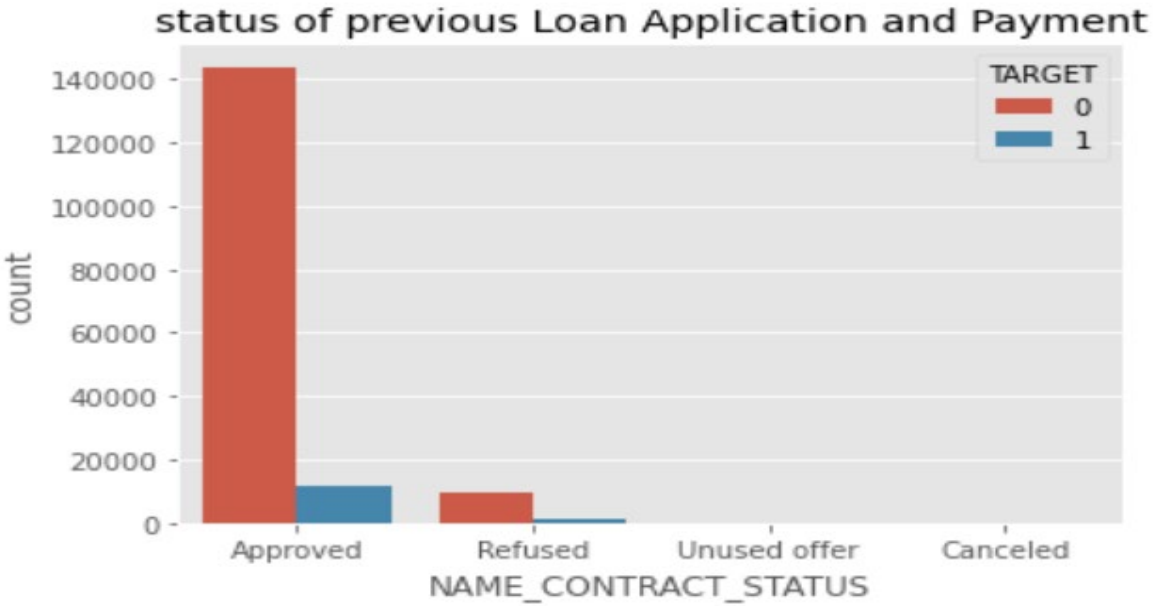
→ As you can see below(next slide) the lighter shade means very highly correlated, therefore DAYS_LAST_DUE and DAYS_TERMINATION are highly correlated

→ similarly AMT_ANNUTY, AMT_APPLICATION, AMT_CREDIT, AMT_GGODS_PRICE are highly correlated.

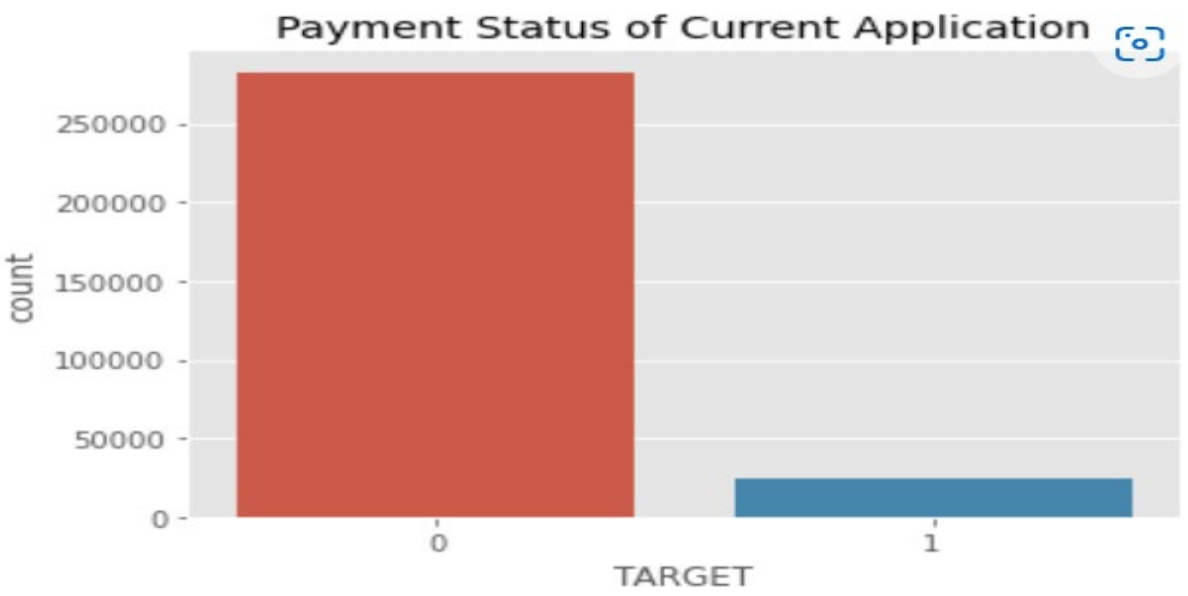
This kind of high correlation value is not accepted in linear model and this will cause multi collinearity and it will like increase number of feature we don't need this many feature so we safely drop it.


AMT_ANNUITY 1 0.8 0.81 0.27 0.81 -0.086 0.14 -0.21 0.37 0.044 -0.049 -0.06 0.073 0.058 0.24
AMT_APPLICATION 0.8 1 0.97 0.47 1 -0.065 0.1 -0.21 0.67 0.064 -0.042 -0.073 0.16 0.14 0.22
AMT_CREDIT 0.81 0.97 1 0.28 0.99 -0.19 0.11 -0.21 0.66 -0.055 0.0079 0.061 0.21 0.2 0.23
AMT_DOWN_PAYMENT 0.27 0.47 0.28 1 0.47 0.49 0.044 -0.12 0.02 0.0015 -0.011 -0.0036 -0.032 -0.032 -0.031
AMT_GOODS_PRICE 0.81 1 0.99 0.47 1 -0.065 0.1 -0.21 0.66 -0.032 -0.018 0.023 0.19 0.19 0.21
RATE_DOWN_PAYMENT -0.086 -0.065 -0.19 0.49 -0.065 1 -0.11 -0.092 -0.28 -0.0037 -0.034 -0.013 -0.16 -0.16 0.0096
RATE_INTEREST_PRIMARY 0.14 0.1 0.11 0.044 0.1 -0.11 1 -0.052 -0.019 -0.022 -0.0071 -0.01 -0.0071 0.32
RATE_INTEREST_PRIVILEGED -0.21 -0.21 -0.21 -0.12 -0.21 -0.092 -0.052 1 -0.056 0.13 0.0019 0.37 0.37 -0.072
CNT_PAYMENT 0.37 0.67 0.66 0.02 0.66 -0.28 -0.019 -0.056 1 0.3 -0.19 -0.37 0.11 0.068 0.27
DAYS_FIRST_DRAWING 0.044 0.064 -0.055 0.0015 -0.032 -0.0037 0.3 1 0.00012 -0.81 -0.23 -0.37 0.14
DAYS_FIRST_DUE -0.049 -0.042 0.0079 -0.011 -0.018 -0.034 -0.022 0.13 -0.19 0.00012 1 0.49 0.37 0.31 -0.085
DAYS_LAST_DUE_1ST_VERSION -0.06 -0.073 0.061 -0.0036 0.023 -0.013 -0.0071 0.0019 -0.37 -0.81 0.49 1 0.38 0.46 -0.17
DAYS_LAST_DUE 0.073 0.16 0.21 -0.032 0.19 -0.16 -0.01 0.37 0.11 -0.23 0.37 0.38 1 0.93 0.034
DAYS_TERMINATION 0.058 0.14 0.2 -0.032 0.19 -0.16 -0.0071 0.37 0.068 -0.37 0.31 0.46 0.93 1 0.021
NFLAG_INSURED_ON_APPROVAL 0.24 0.22 0.23 -0.031 0.21 0.0096 0.32 -0.072 0.27 0.14 -0.085 -0.17 0.034 0.021 1
AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_DOWN_PAYMENT
AMT_GOODS_PRICE
RATE_DOWN_PAYMENT
RATE_INTEREST_PRIMARY
RATE_INTEREST_PRIVILEGED
CNT_PAYMENT
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Imbalance Data



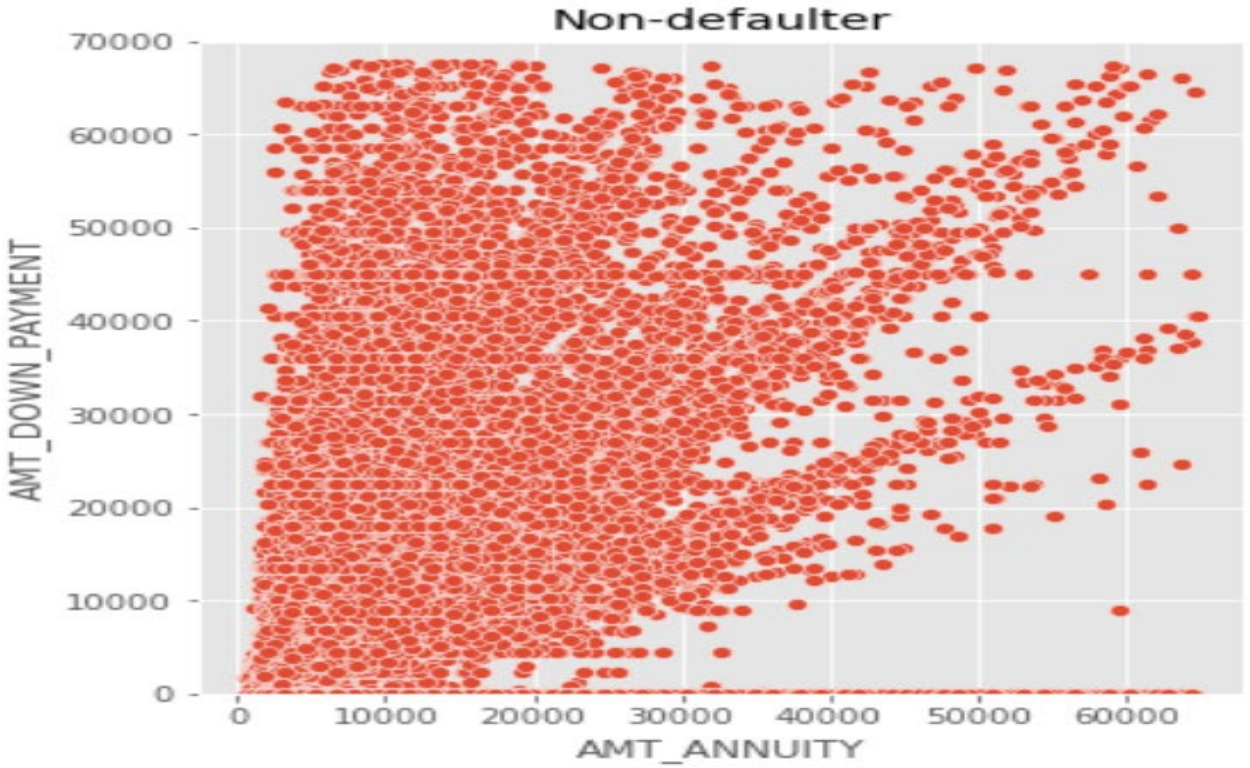
- The data is very highly imbalanced.
- Majority of loans approved some are refused.

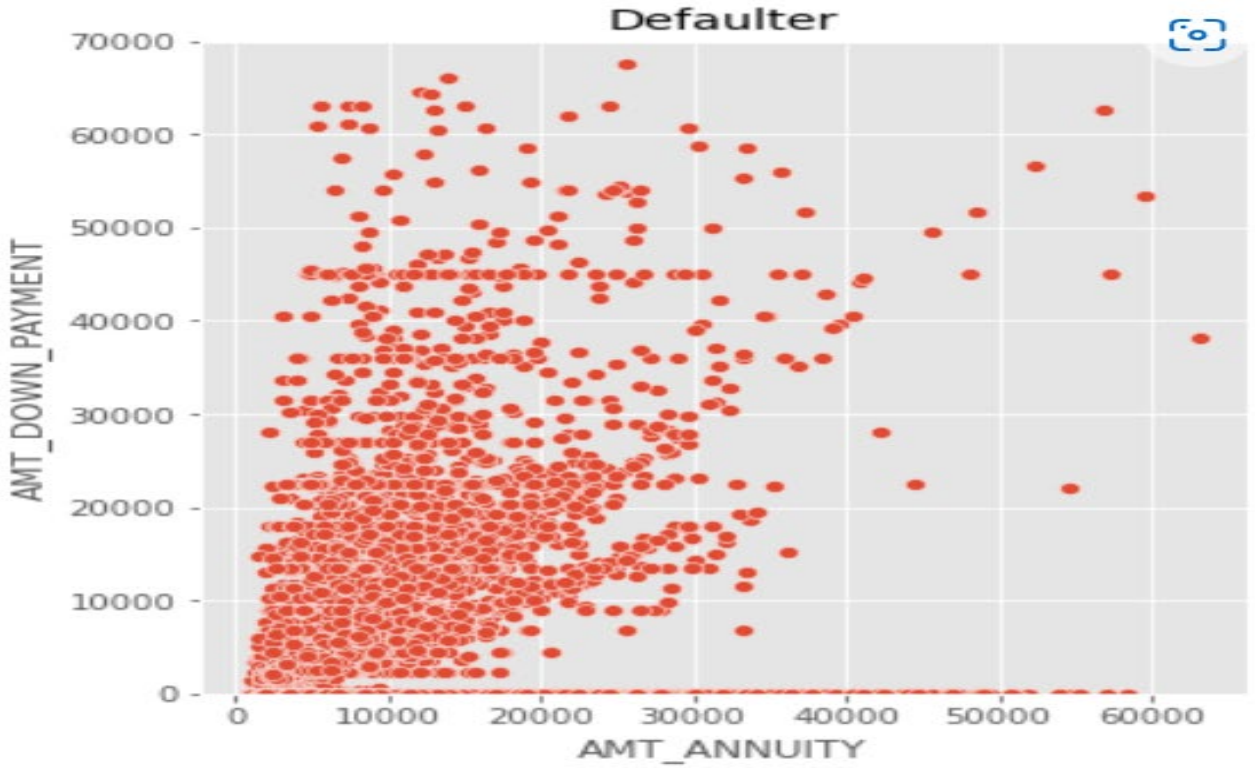


- As well as data is very highly imbalanced.
- Number of defaulter is very less in total application.

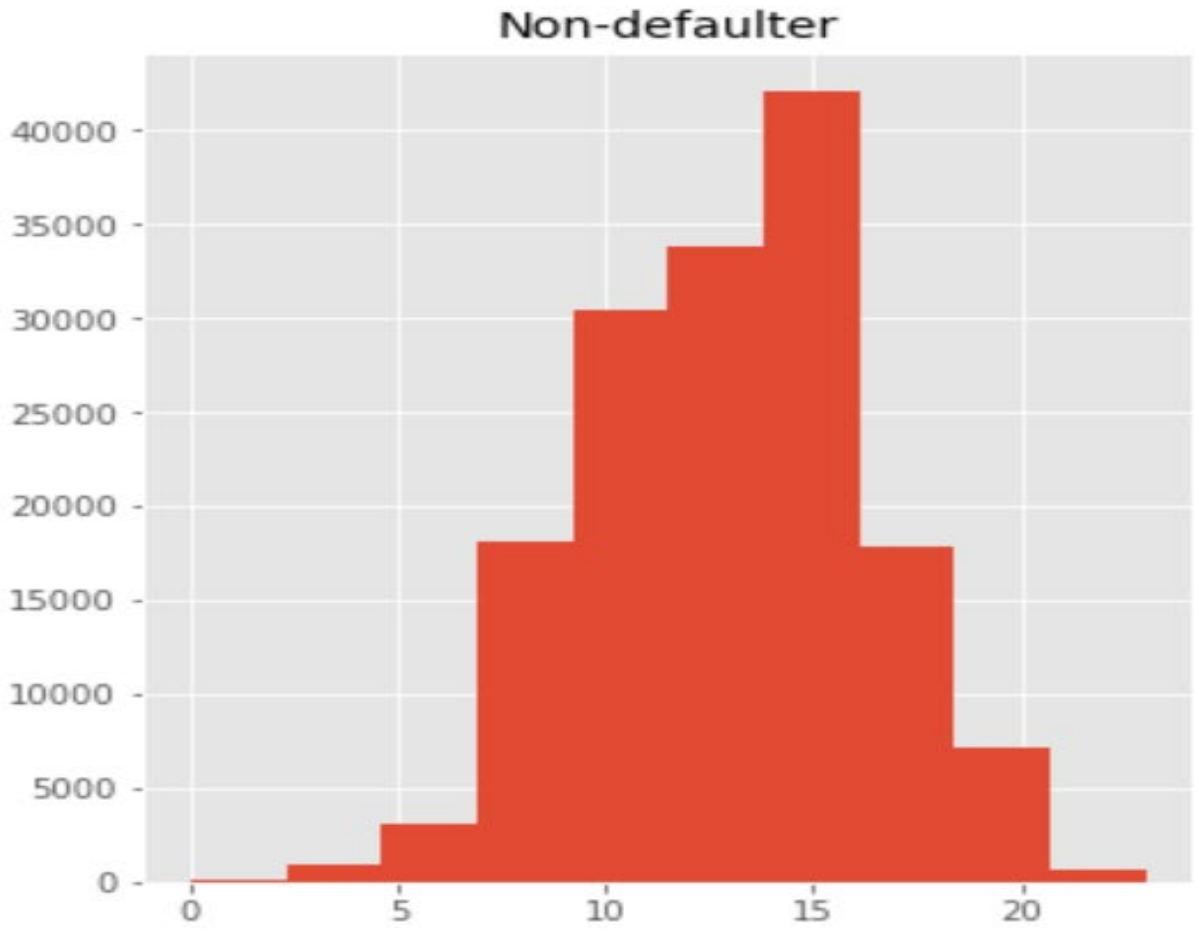
Numeric Feature of Previous Application Data

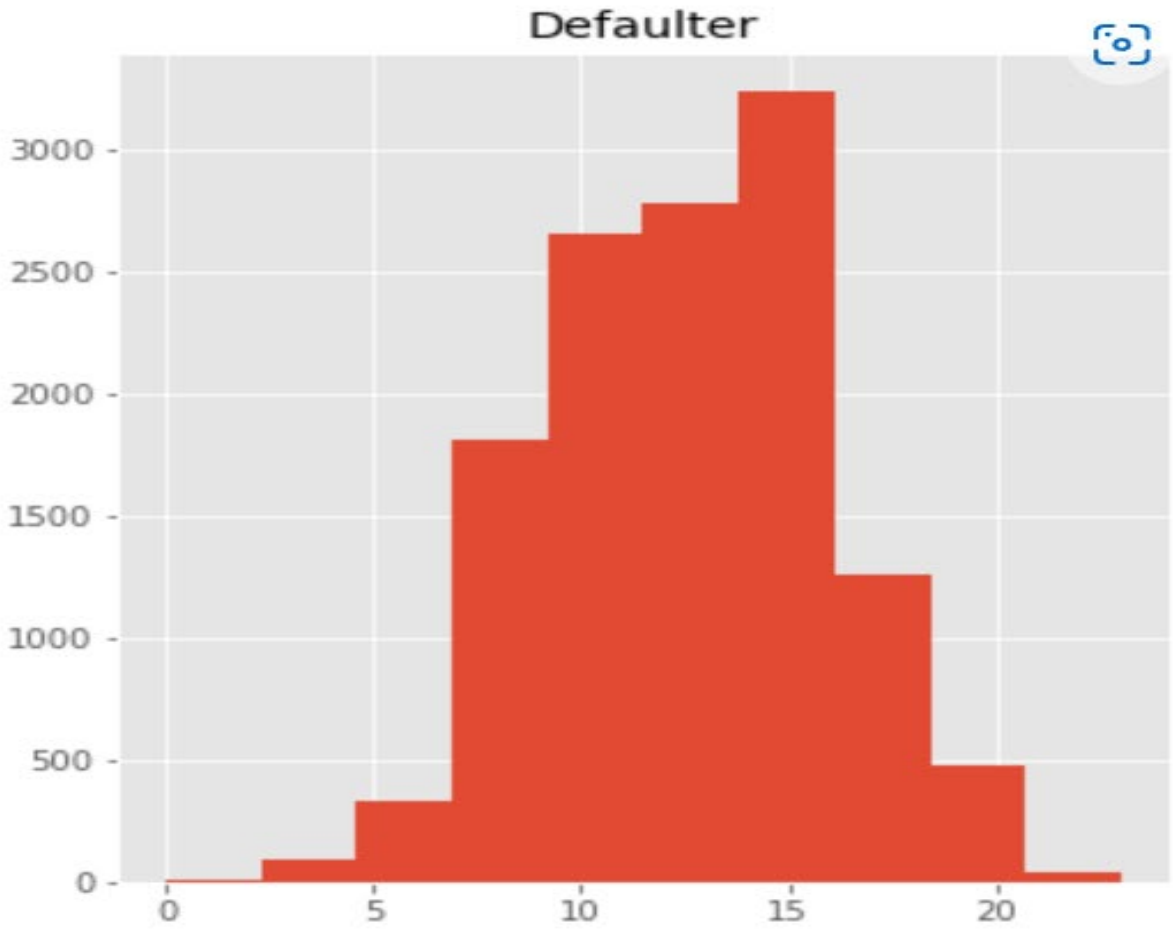
- NUMBER OF DEFAULTERS ARE LESS FOR LARGER AMOUNT OF ANNUITY OF PREVIOUS APPLICATION
- FOR HIGHER DOWN PAYMENT DEFAULTER CASES ARE LESS





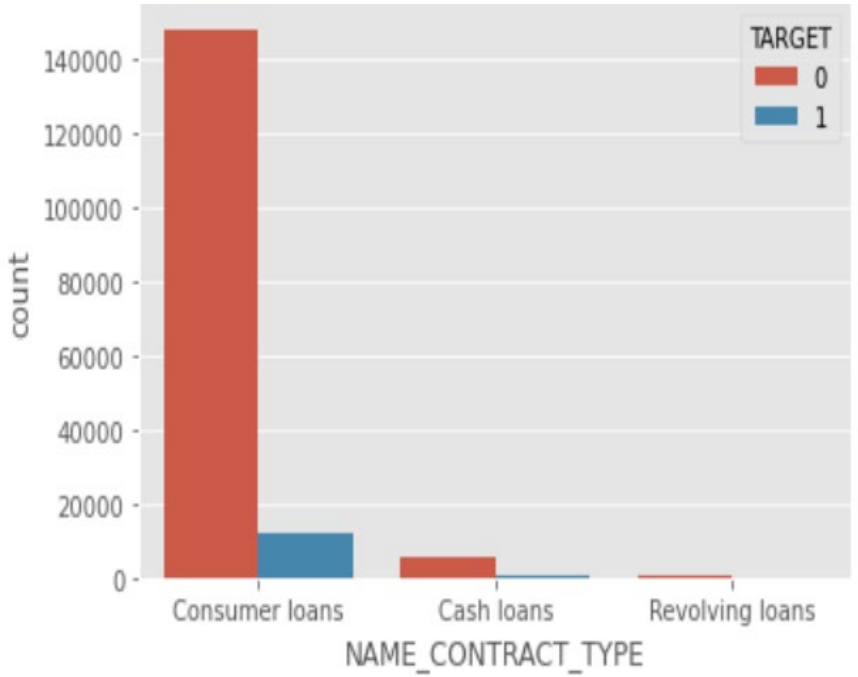
Visualization

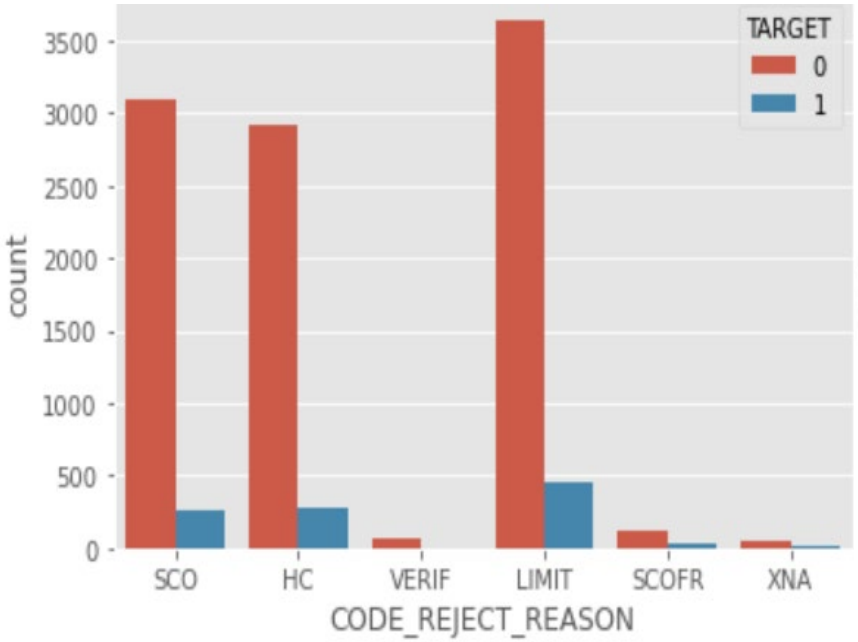


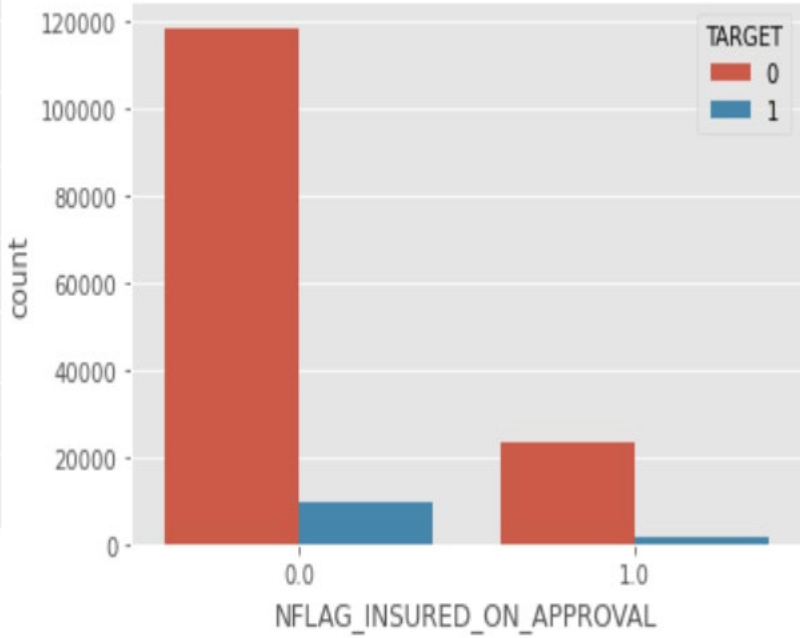


- Most of the loans are applied around 15:00 hours . This feature is does not have visible impact on TARGET variable

Applied, Rejection, Approval







→We can understand what kind of loans applied.
.

→Highest number of loans are applied for

Consumer loans.

→ This is the reason why . .
loans are rejected

→ As seen in the above plot .

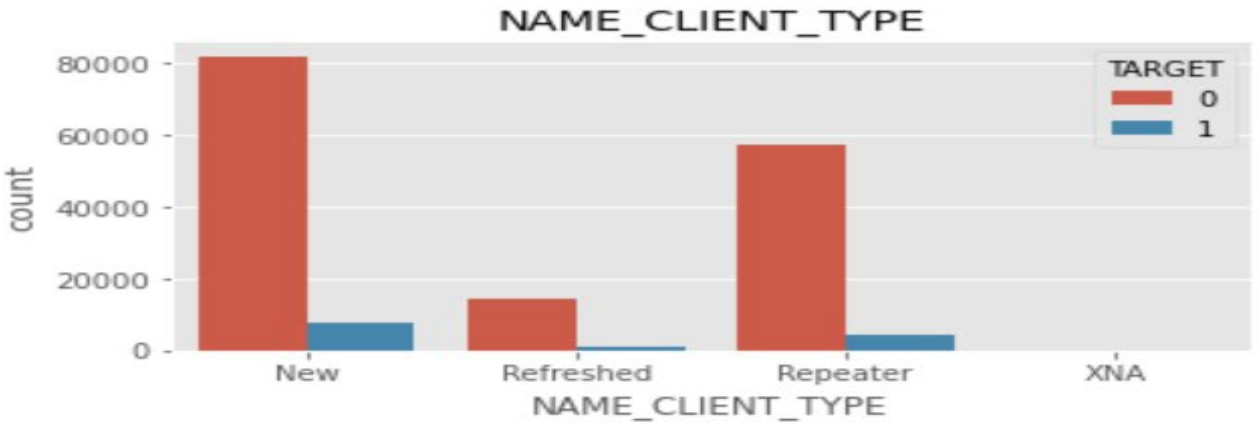
'SCO', 'LIMIT', ''HC are

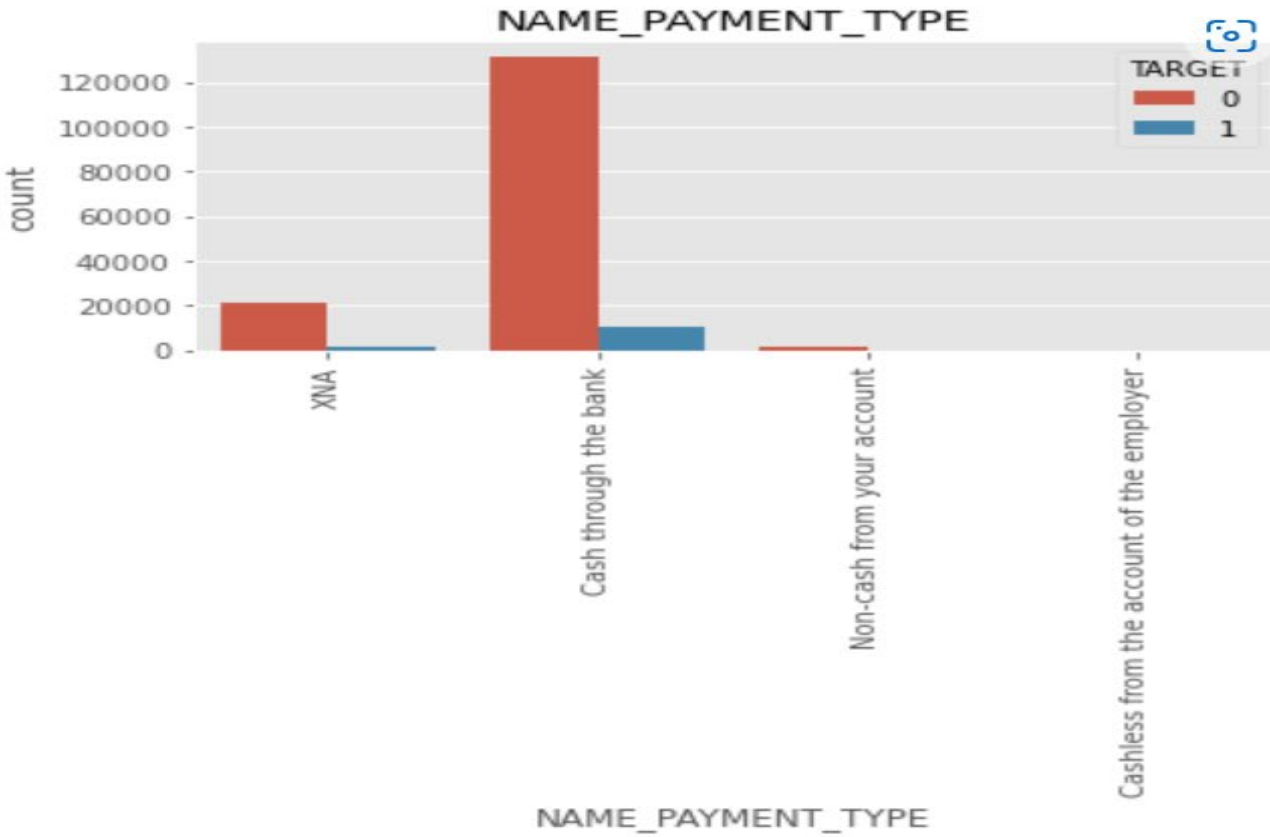
most common reason of rejection

→ Most of the people did not required insurance during previous loan application.

→ People are just braved they don't want insurance.

Visualization

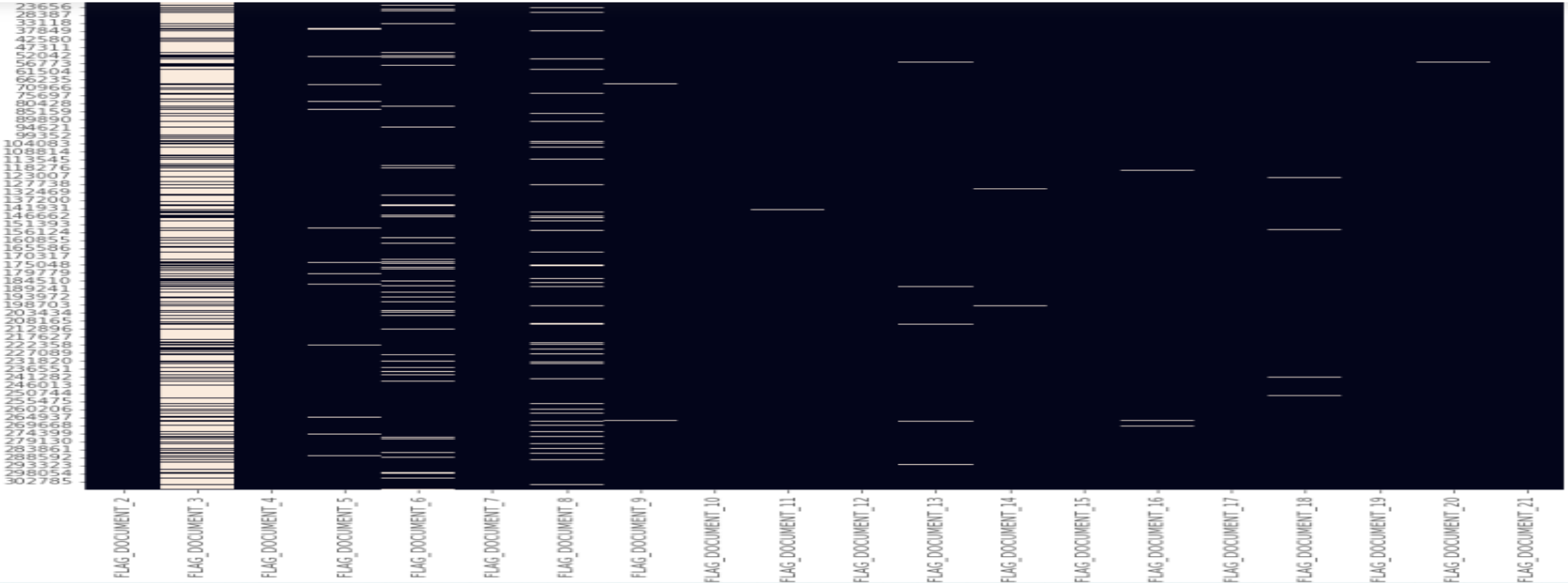




1. Most of the clients are repeaters

2. "Cash through the bank" is the most frequently use payment method

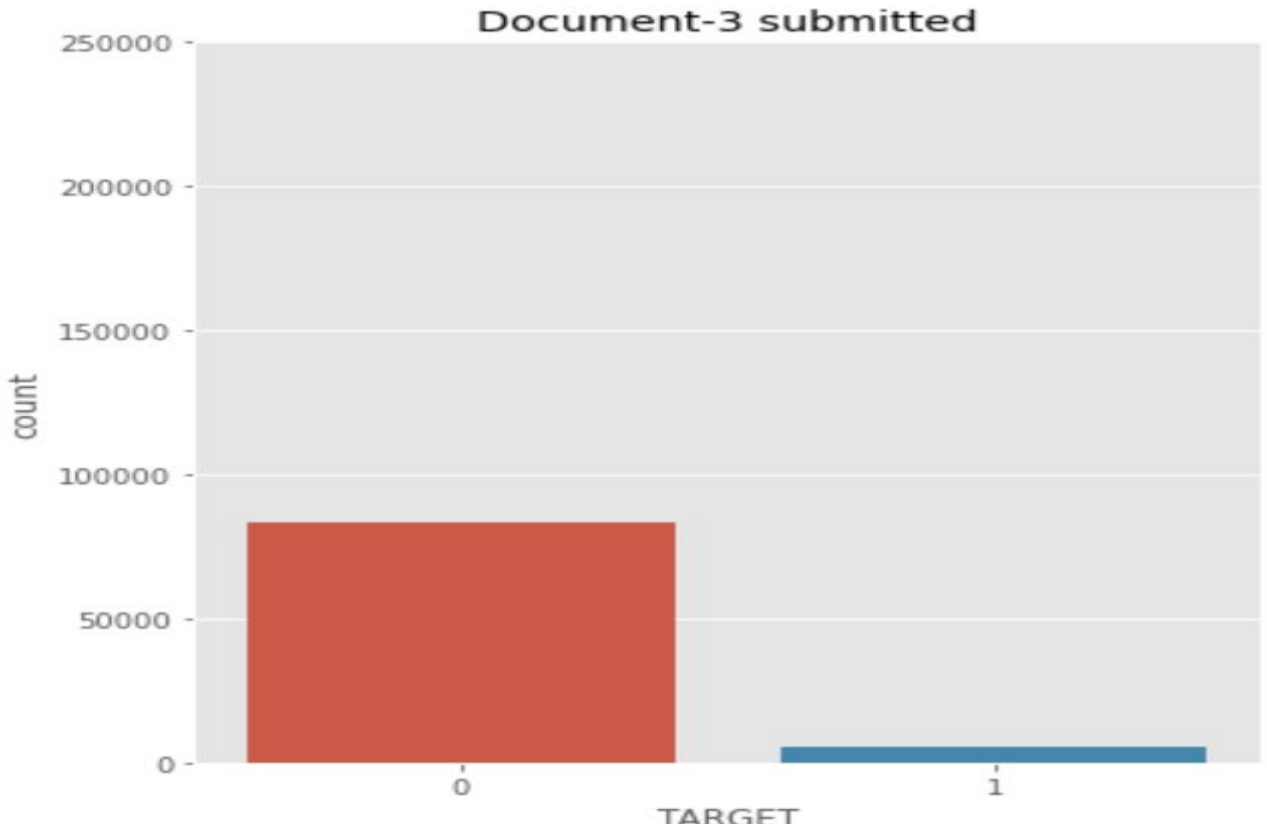
Documents Submitted By Clients

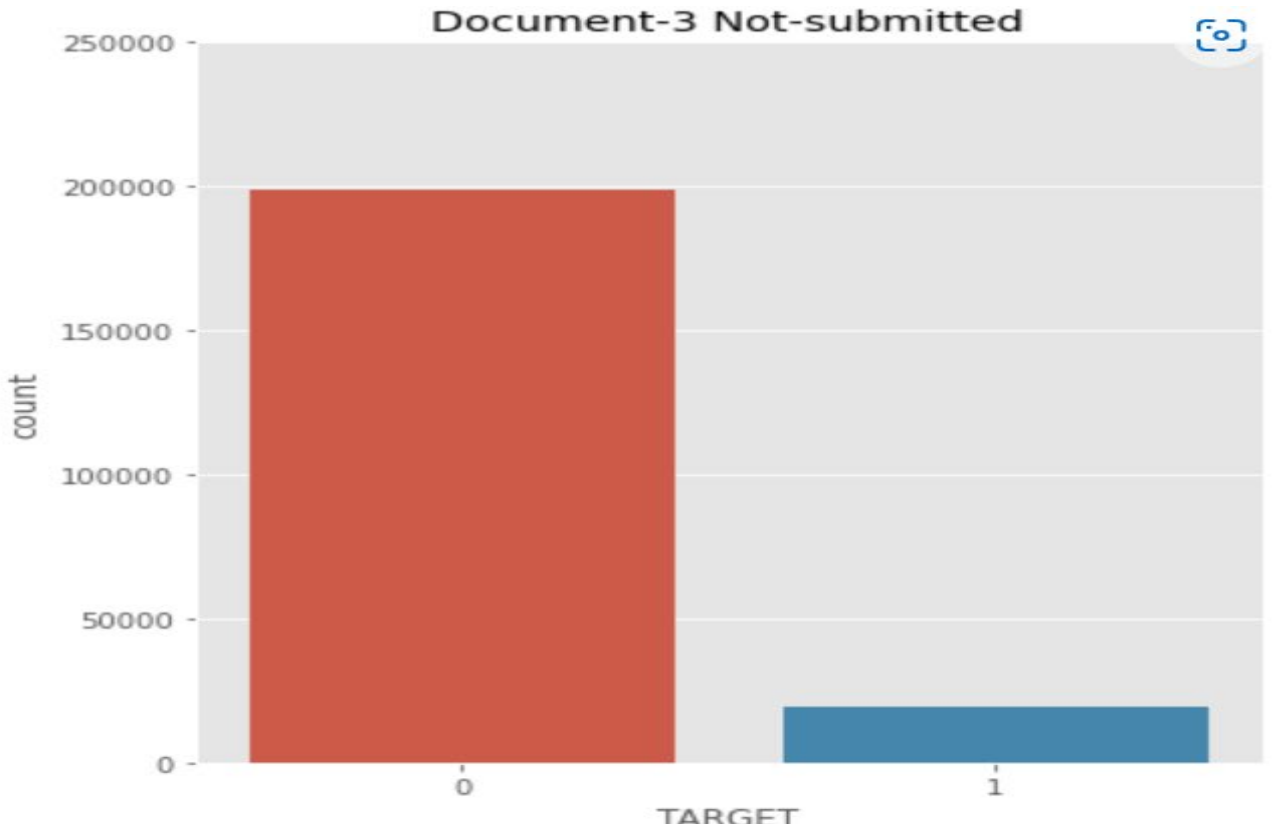


1. The heatmap suggests that all of the documents except Document 3 were not provided by clients in majority of the cases.

2. Hence we can assume all the document(except document 3) will not contribute towards analyzing the data. Hence columns can be dropped

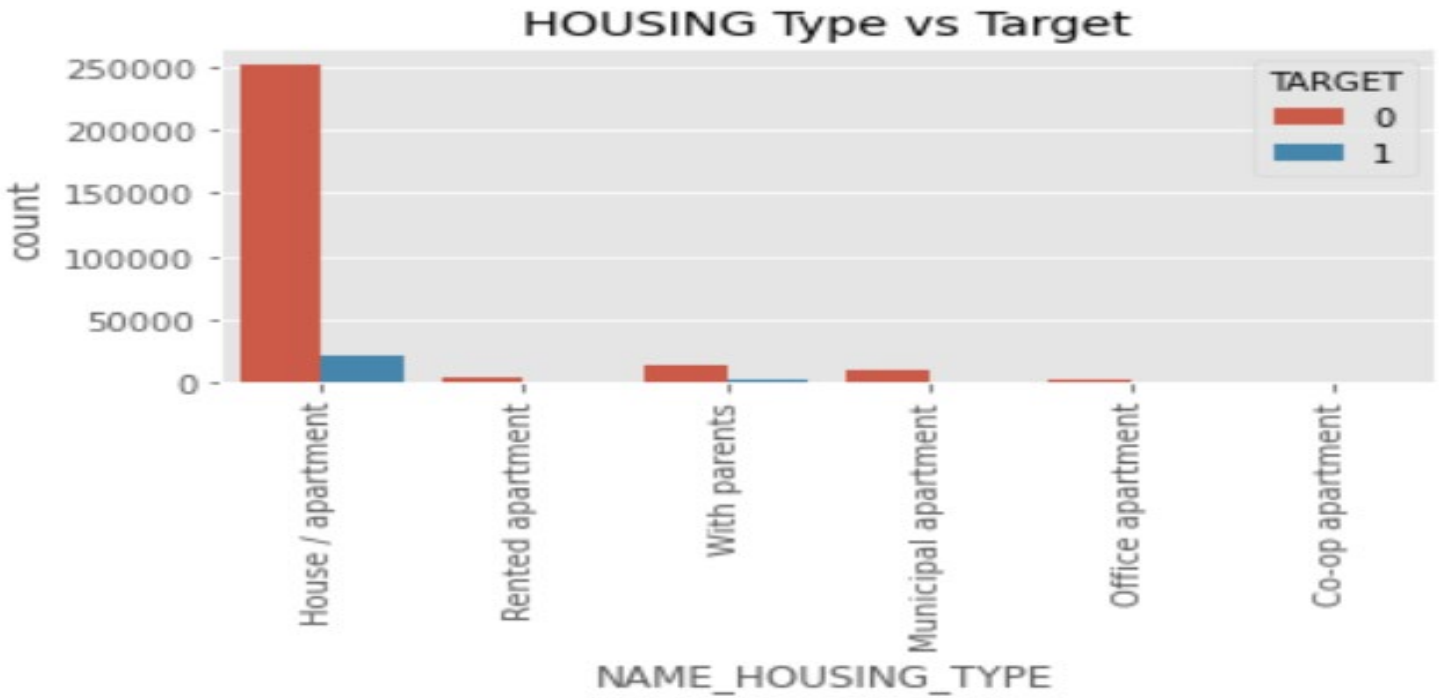
Default and Non-Default





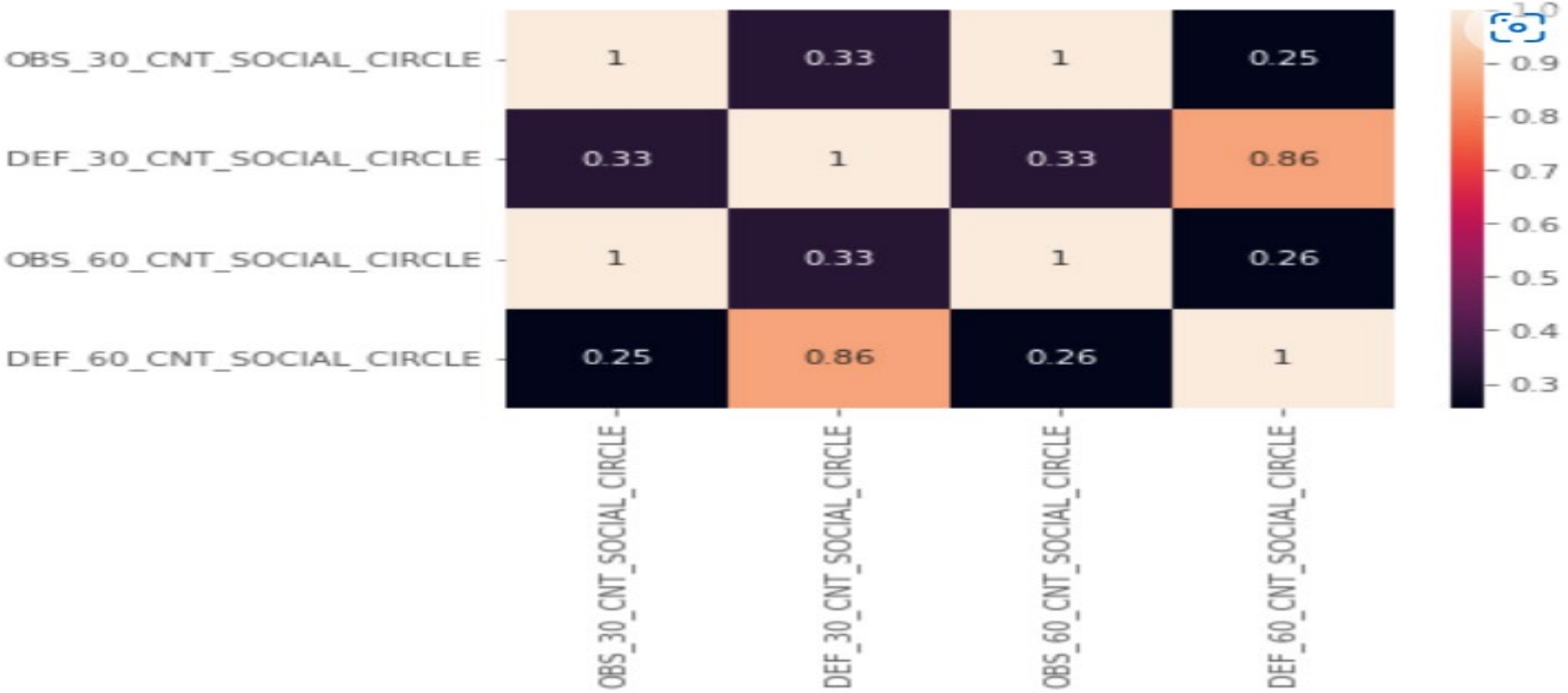
- FLAG_DOCUMENT_3 is showing similar trend for both non_defaulter and defaulter.
- Hence this column can be dropped.

Housing Information of Clients



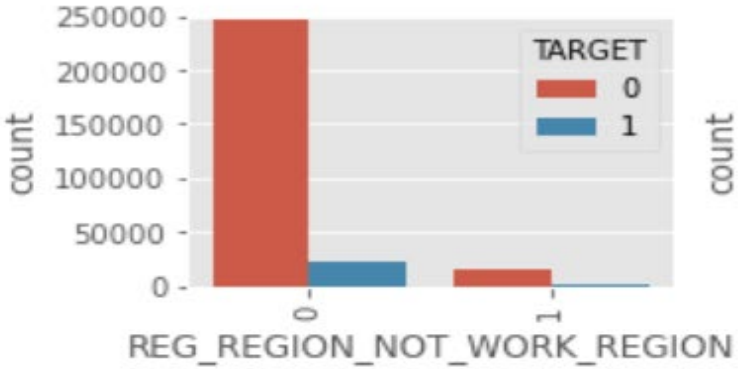
- Most of the clients live in House/Apartment
- Clients living with their parents or in rented apartment have higher rate of default.

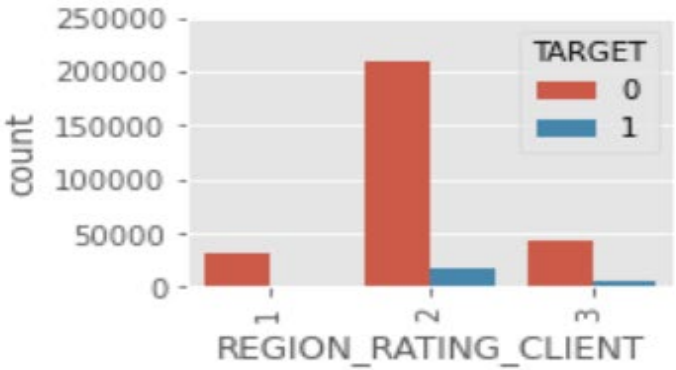
Social Circle Info



- DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE are highly correlated
- OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE are identical columns

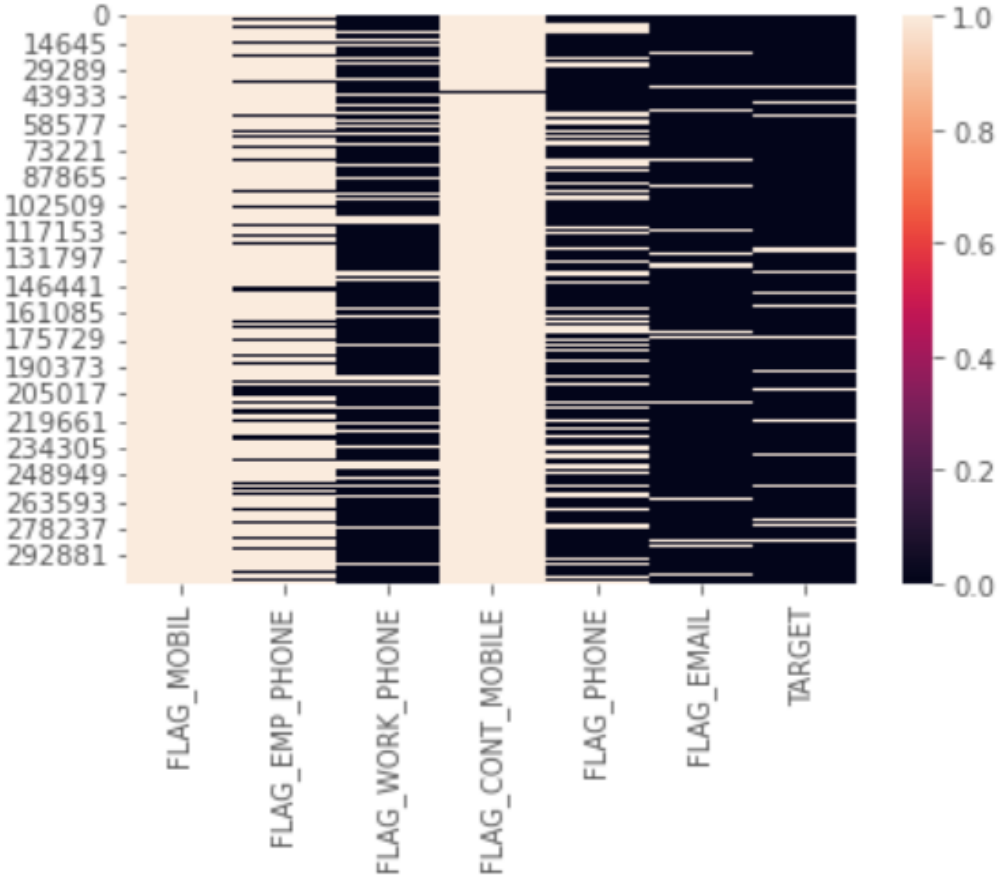
Region Related Data





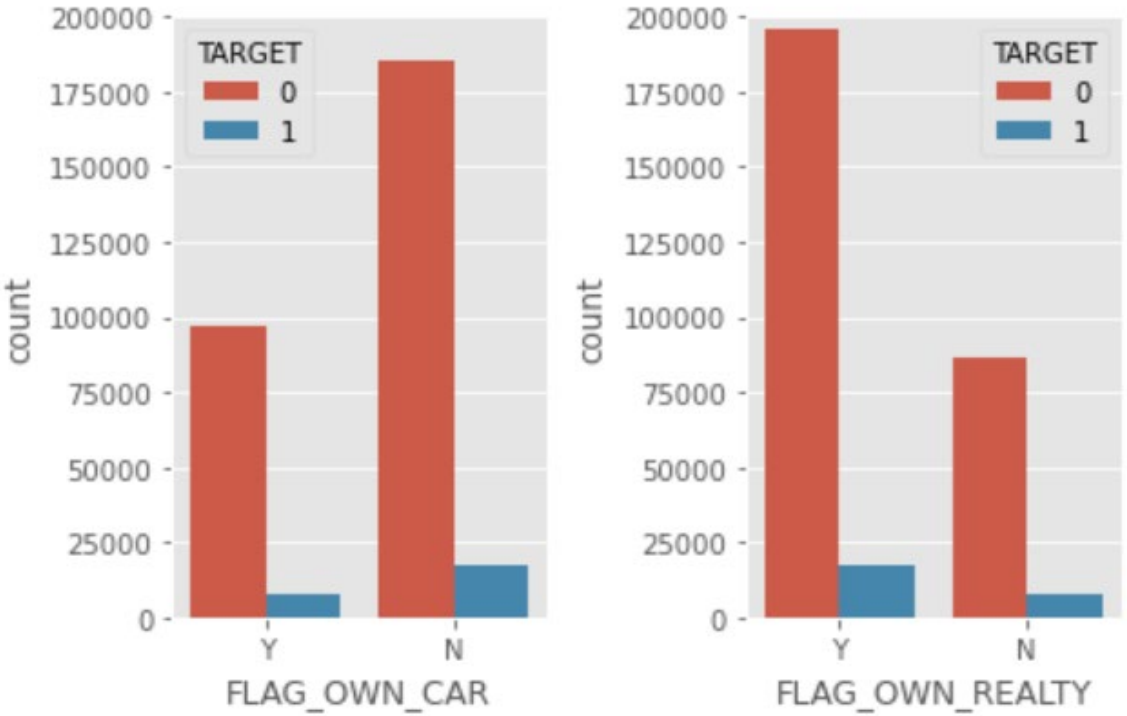
- Defaulter rate is highest when REG_REGION_NOT_WORK_REGION=0 i.e permanent address and working address is same
- Highest Application have Region rating of 2

Contact Related Info



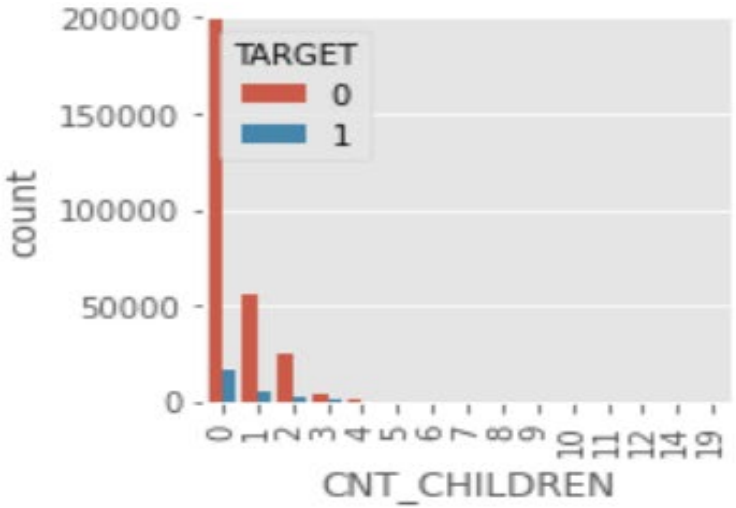
- All the features in contact_df are categorical (0 and 1)
- As there no similarity of patterns of TARGET value with the features, we are assuming the feature are not useful for analysis.
- Hence all of the features can be removed

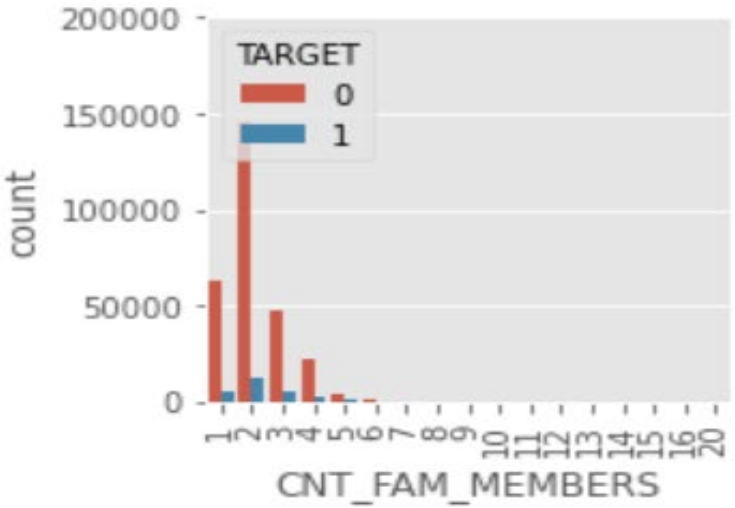
Assets Details

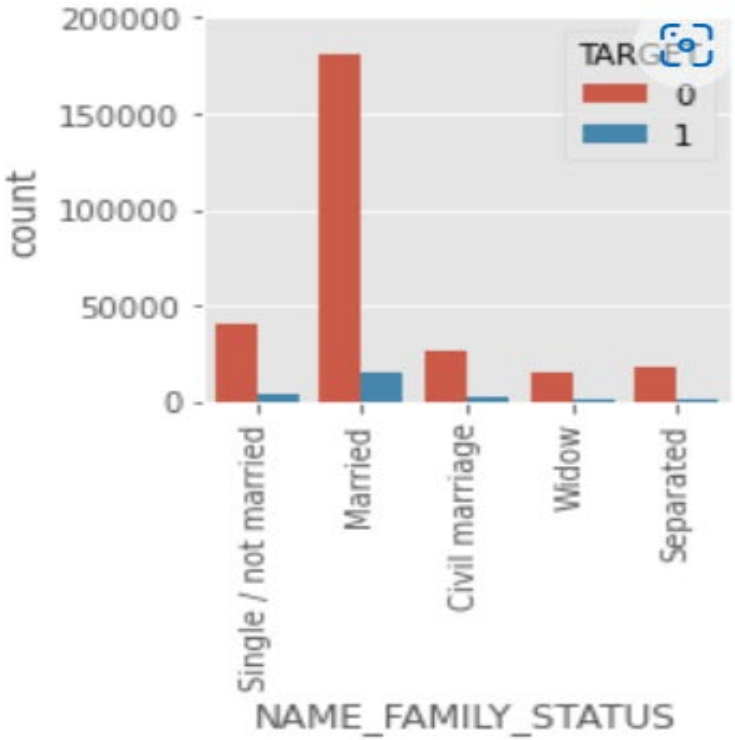


- Most of the clients own realty
- Most of the clients do not own car
- People not owning reality and car and have a slightly higher default tare than the people who own reality and car

Family Related Info

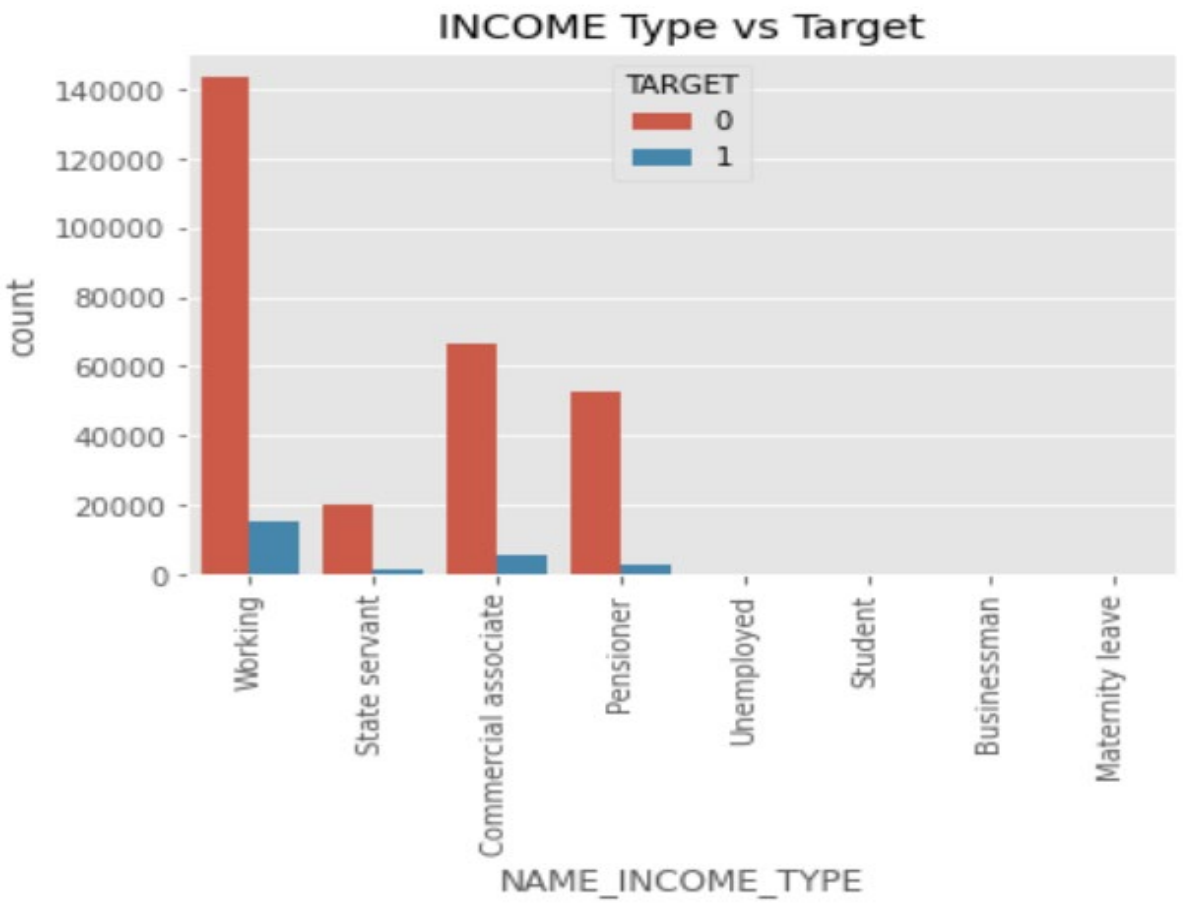






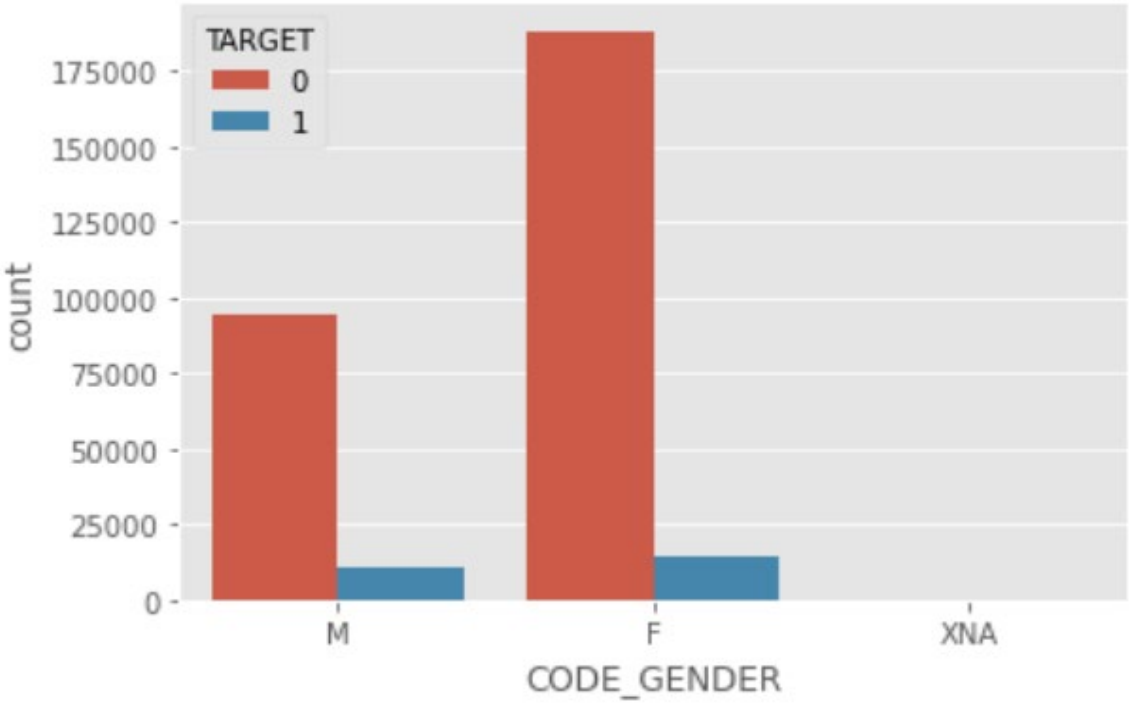
- Default rate is highest for Civil Manager and Single clients
- Most of the clients are married (and/or) no children (and/or) 2 family members
- Clients with relatively more number of children (and/or) family members have higher default percentage
- For some of the cases where count children/family members high, and the default rate very high or very low. This cases cannot be taken as a conclusion as number of clients having a large family is very low.

Education and Occupation Info



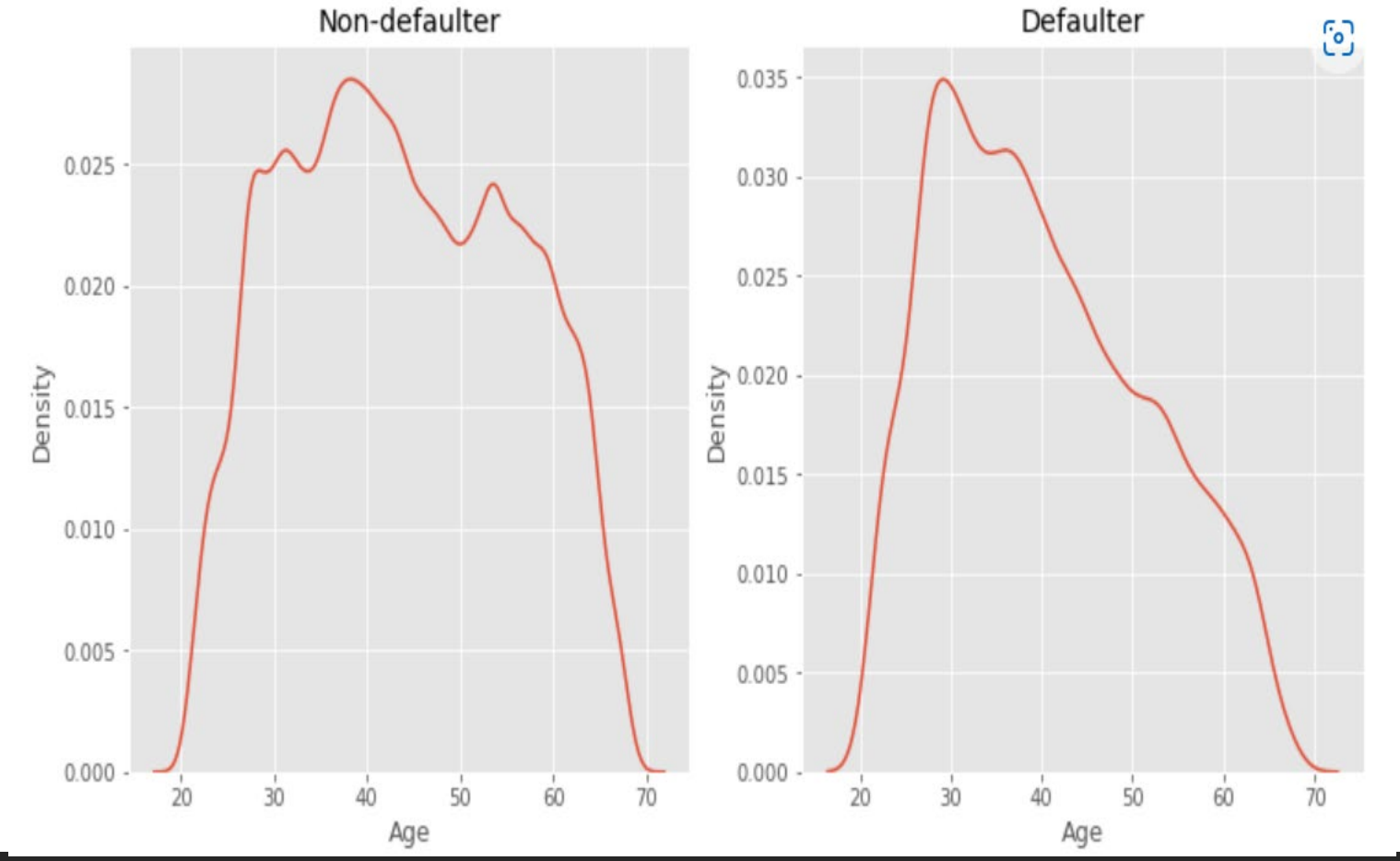
- Most of the clients are working.
- Clients on Matemity Leave and unemployed has highest percentage of Defaulter
- Businessman have lowest(0) percentage of<Defaulter However clients of income type("Unemployed", "Student", "Businessman", "Matemity leave") are very few in the dataset to contribute in the analysis.

Gender



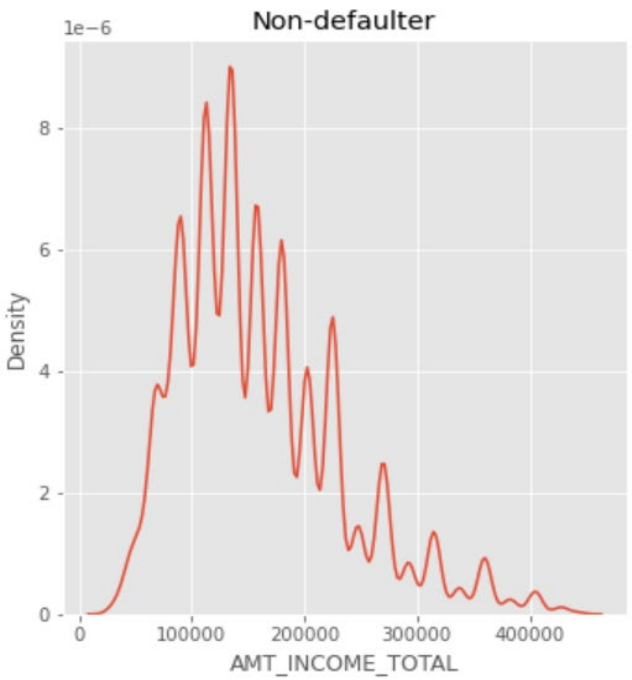
- Female clients are more than male clients
- Defaulter percentage for male clients

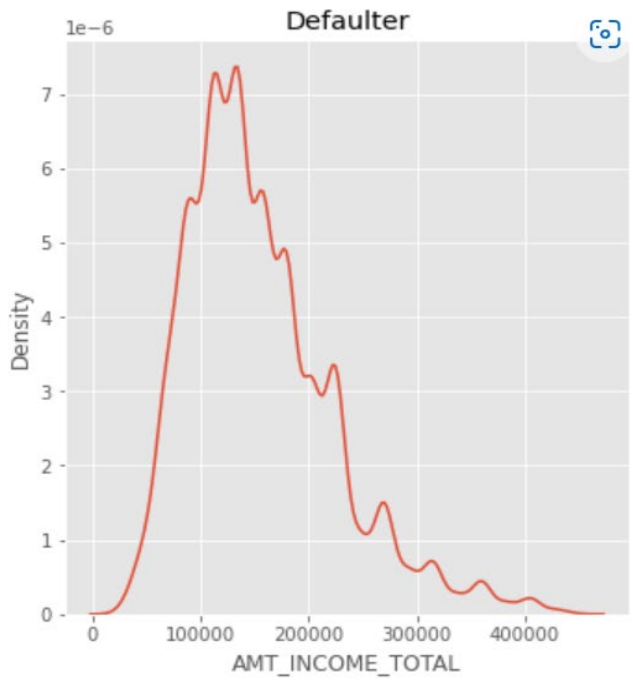
DAYS_BIRTH

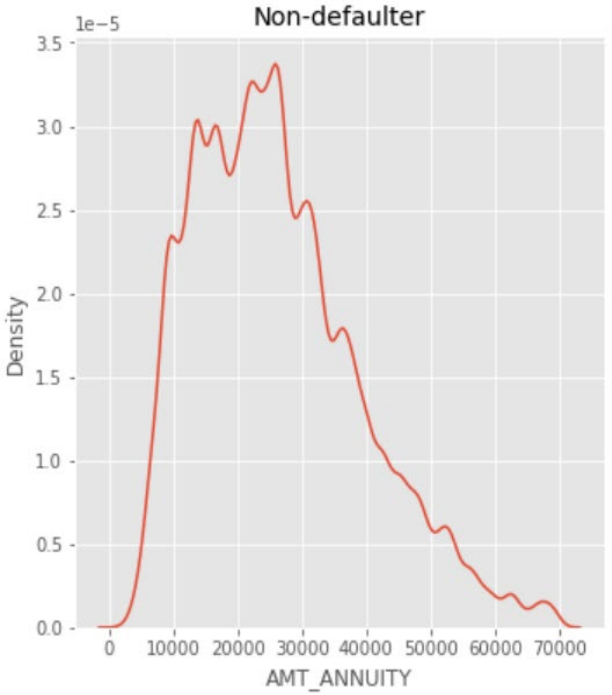


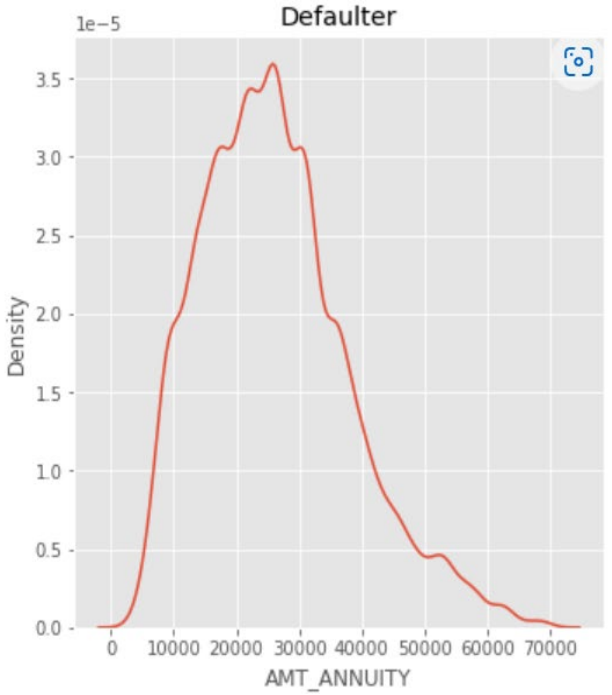
People of age 30 have higher default rate
Default cases are less for clients more than 40 years old.

Income and Annuity









For Defaulter and non-defaulter AMT_ANNUITY distribution is similar.