

Credit EDA Case Study

This case study aims in finding two types of risks that are associated with the bank's decision called credit risk :-

- 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

***An EDA(exploratory data analysis) have been used to analyse the patterns present in the data to avoid above mentioned risk associated in lending. The company can utilise this to build a healthy portfolio and thereby reduce business loses**

Data Understanding

The dataset has 3 files as explained below:

- *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**
- *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**
- *Columns_description.csv* is data dictionary which describes the meaning of the variables

Data preparation

Treating missing values



Preferred to drop the columns having more than 50% of missing values



Considered 35 columns for easy of analysis imputed missing values with 0 in columns

Changing default data type



Incorrect data type of columns are identified



Example integers are considered as object and vice versa corrected the same

Finding outliers and removing them

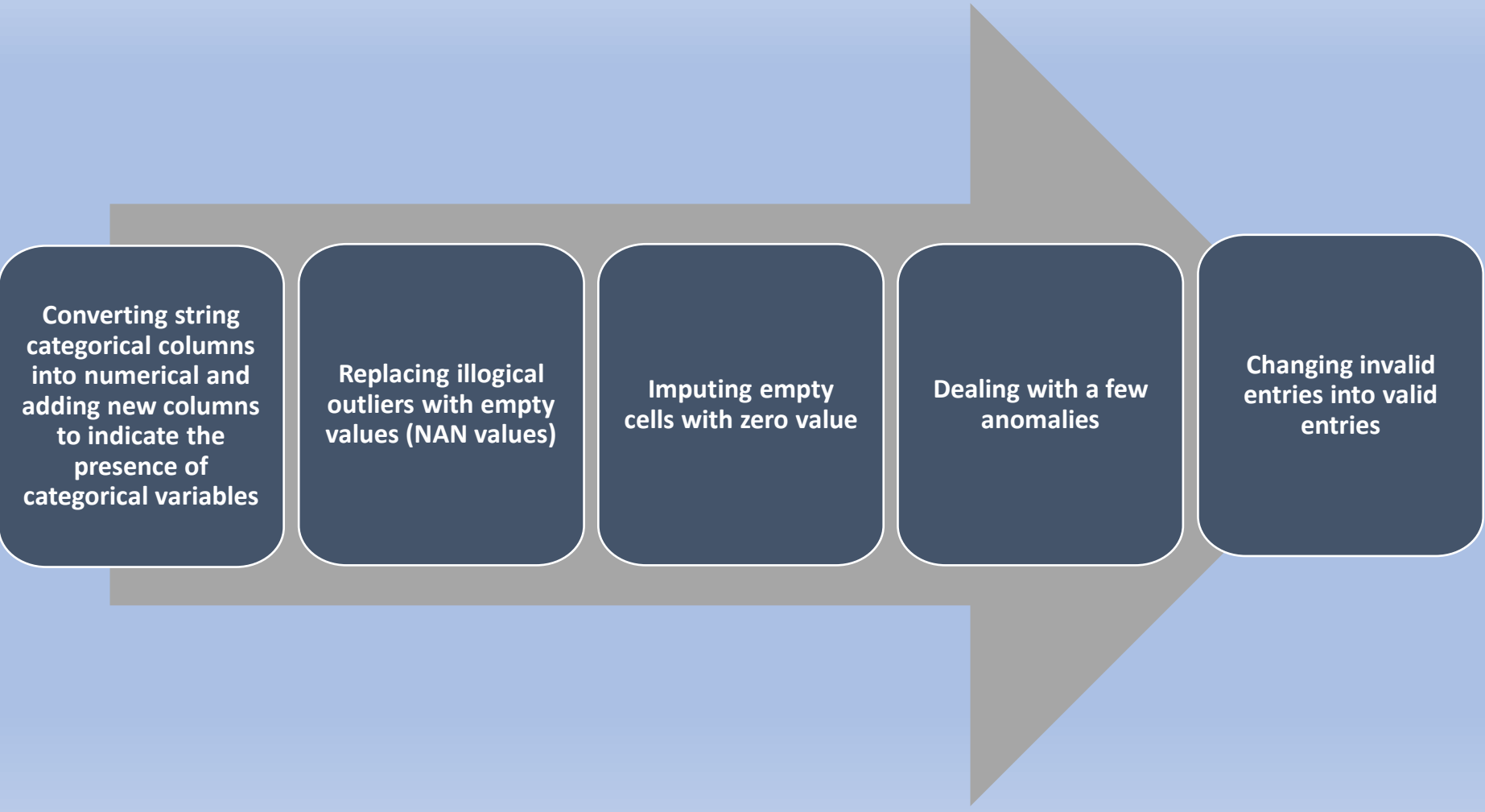


Checked entire float columns sets for outlier



Considered only data below 99% quartile

Treating Missing Values & Changing Data Type



Converting string
categorical columns
into numerical and
adding new columns
to indicate the
presence of
categorical variables

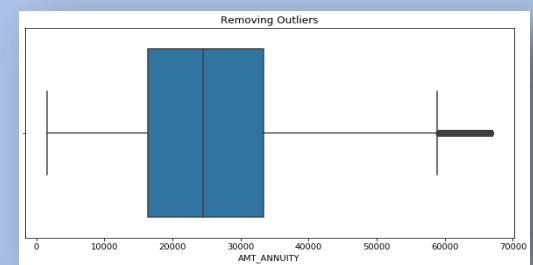
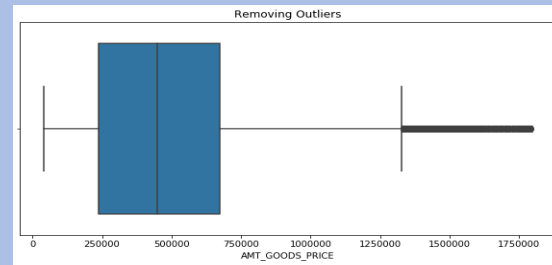
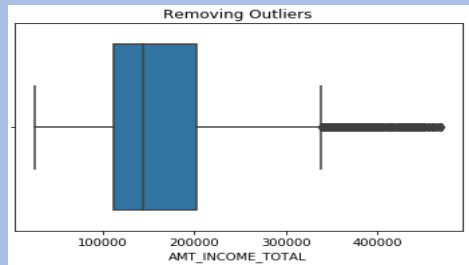
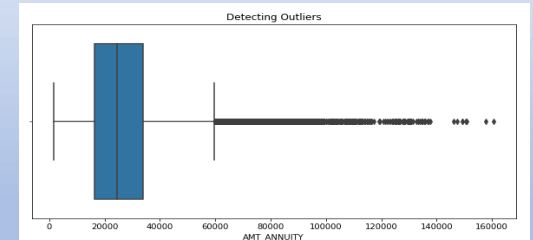
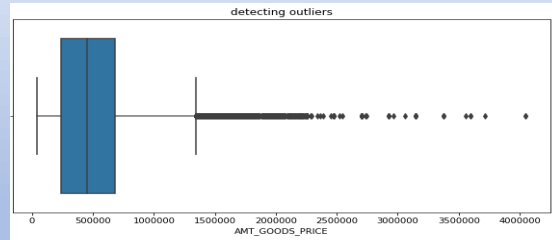
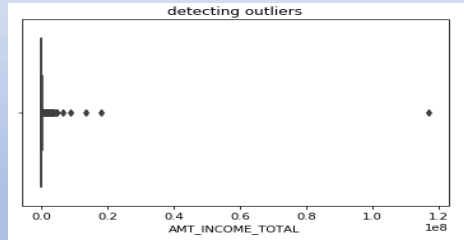
Replacing illogical
outliers with empty
values (NAN values)

Imputing empty
cells with zero value

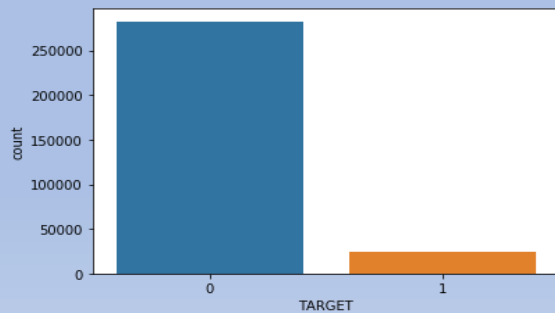
Dealing with a few
anomalies

Changing invalid
entries into valid
entries

Outliers and Data imbalance



Data above 99 percentile is removed in all above cases. The approach was to remove as minimum outliers as possible



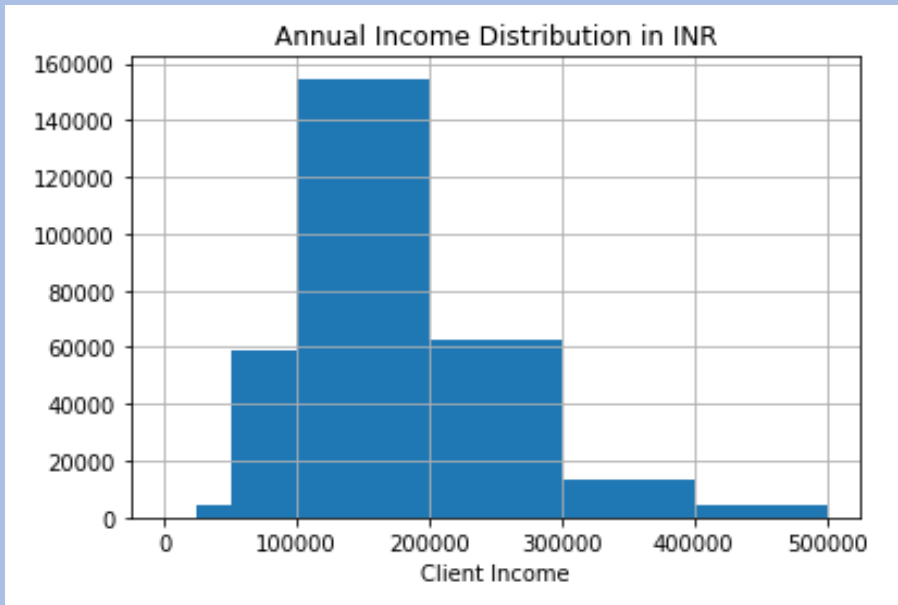
From this information, we see this is an imbalanced dataset. There are far more loans that were repaid on time than loans that were not repaid.

Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

Segmented Analysis

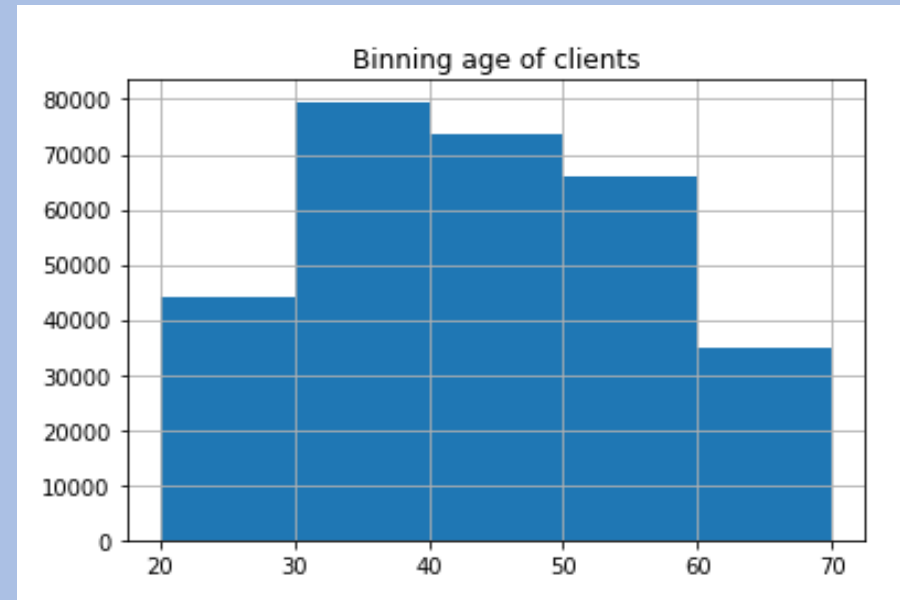
Income groups:

we can see from below graph that income segment between 10000-20000 is highest amongst all other segments



Age group:

Age group of 30-40 is highest followed by 40-50 and 50-60 This mean adults contribution among entire population is significant



UNIVARIATE ANALYSIS

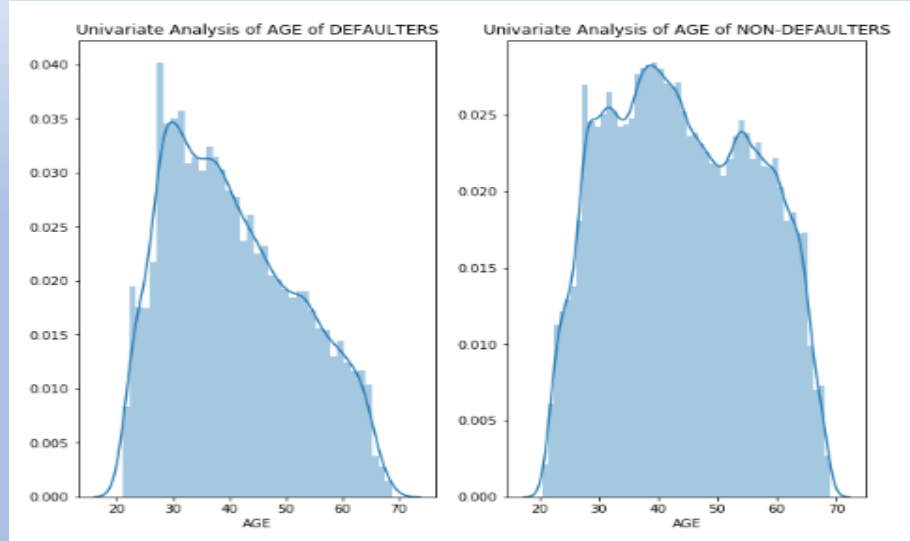


FIG (1)

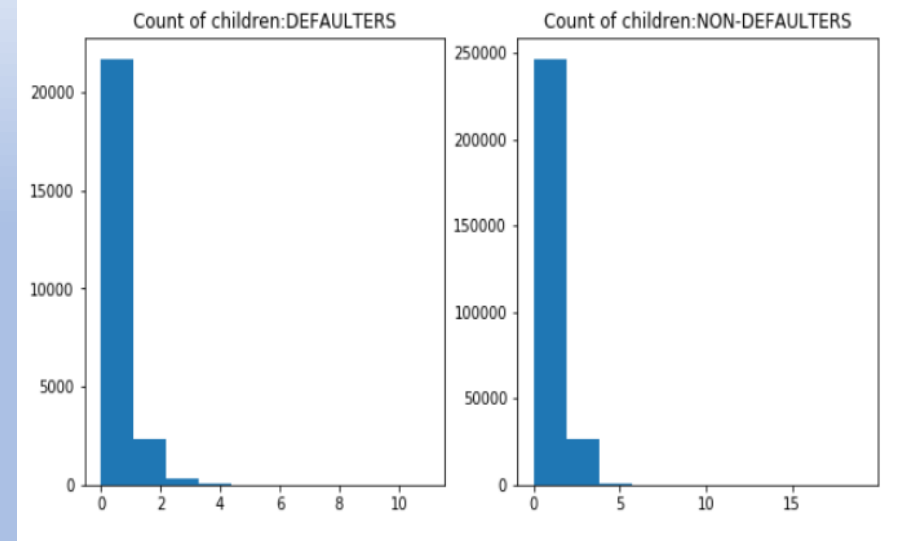


FIG (2)

Fig 1: shows there is high number of defaulter in the age between 20-30 and high number of non defaulter in age range 35-45
Fig 2: shows no such significant difference in the count of children for Defaulters and Non-defaulters

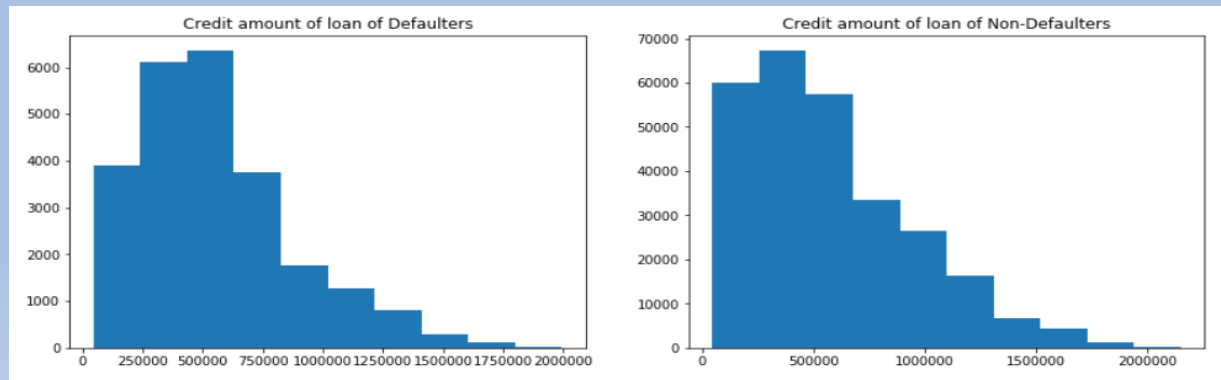


FIG (3)

Fig 3: shows The credit amount of the loan for the Defaulters is less than the Non-Defaulters. Low income group have low eligibility and hence low credit amount.

UNIVARIATE ANALYSIS

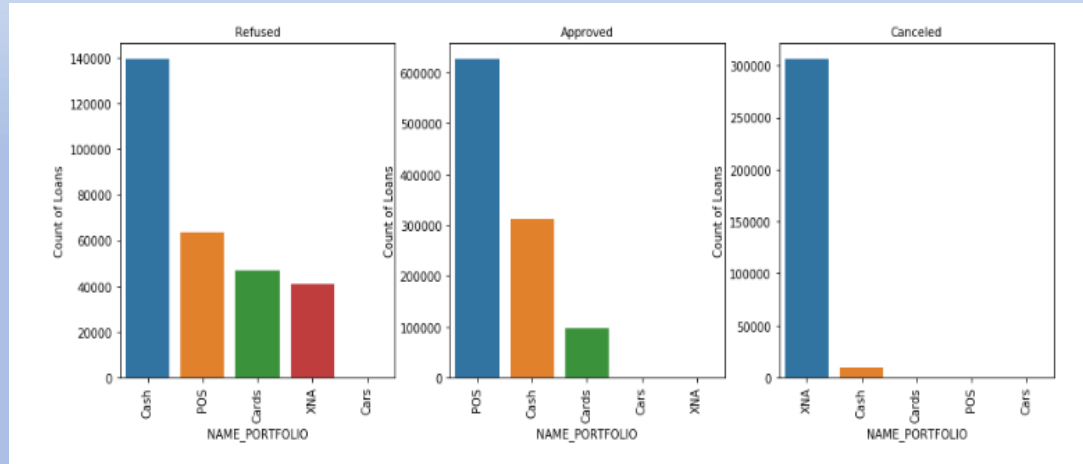


FIG (4)

Fig 4 shows
Most approved loans were POS
Most refused loans were Cash. This may be a good sign since additional exposure have not been increased, the loans were absorbed from existing loans

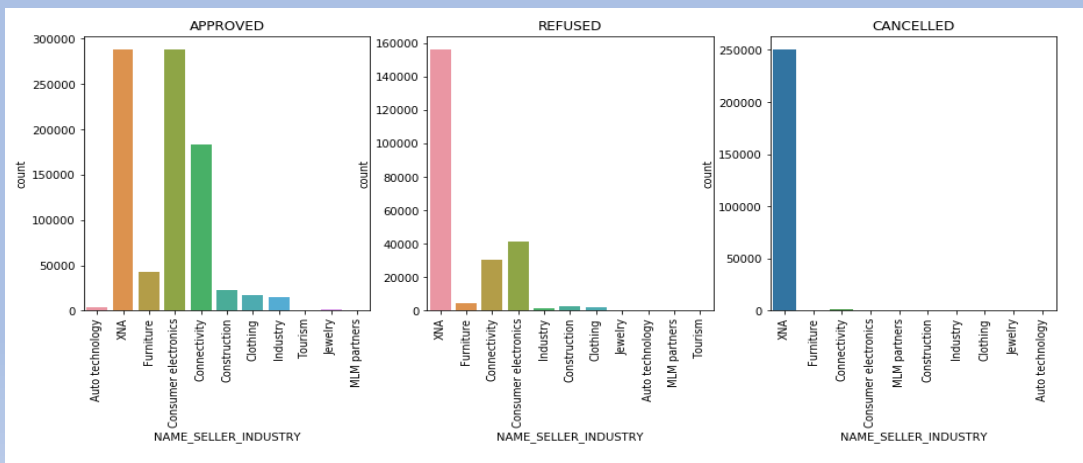
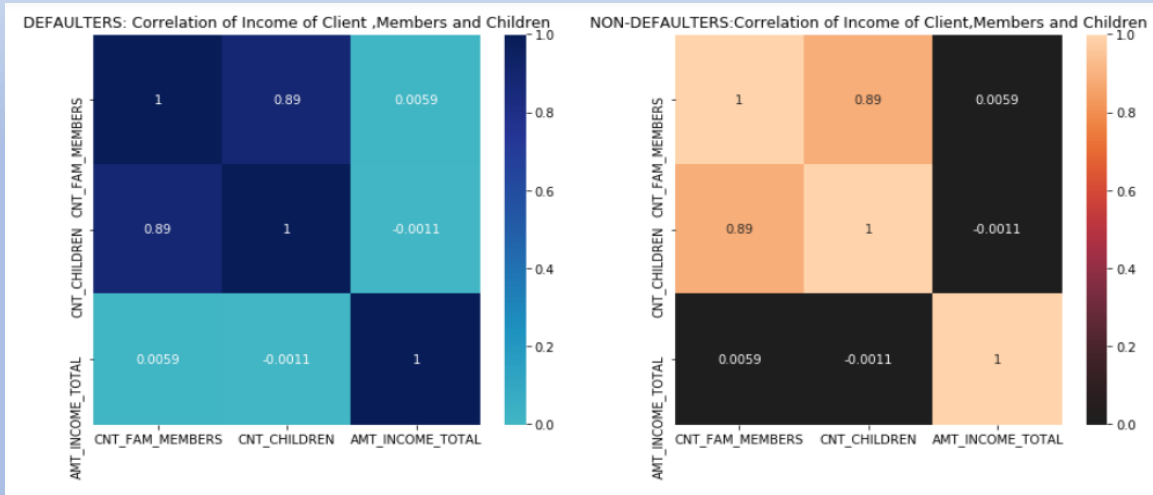


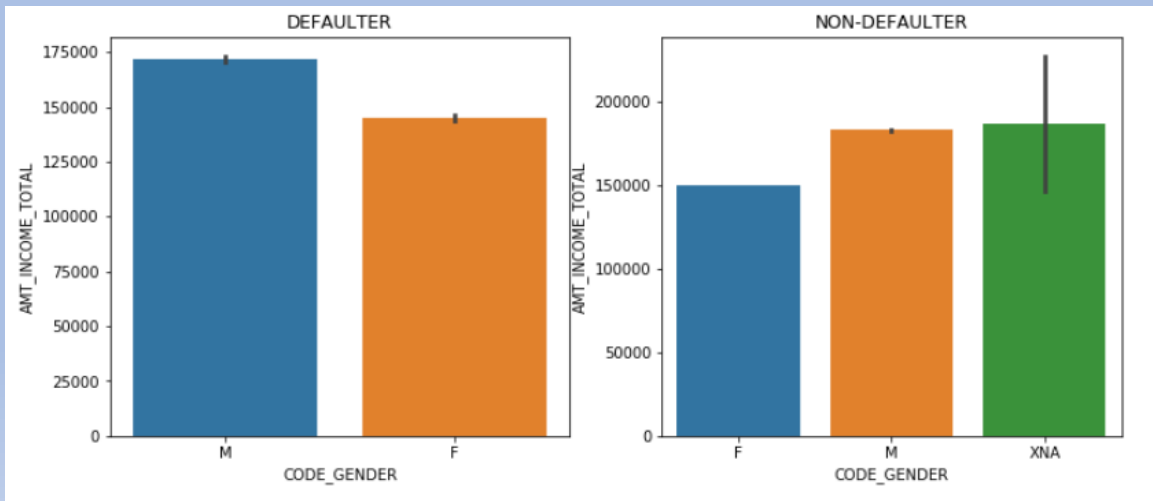
FIG (5)

Fig 5 shows
Most approved application from consumer electronics industry
no significant rejection in any industry

BIVARIATE ANALYSIS

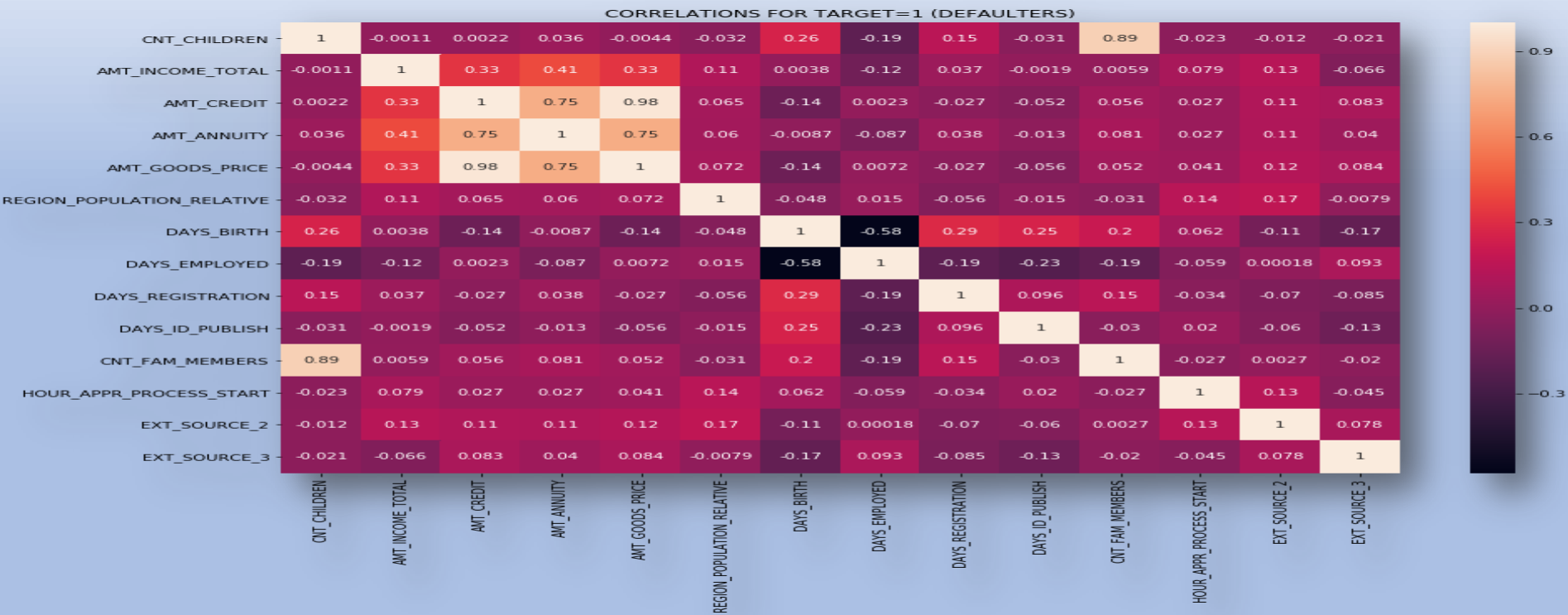


The correlation between the count of family members and the Income of client is almost same for both Defaulters and Non- Defaulters



The Males have high income than the females for both the categories

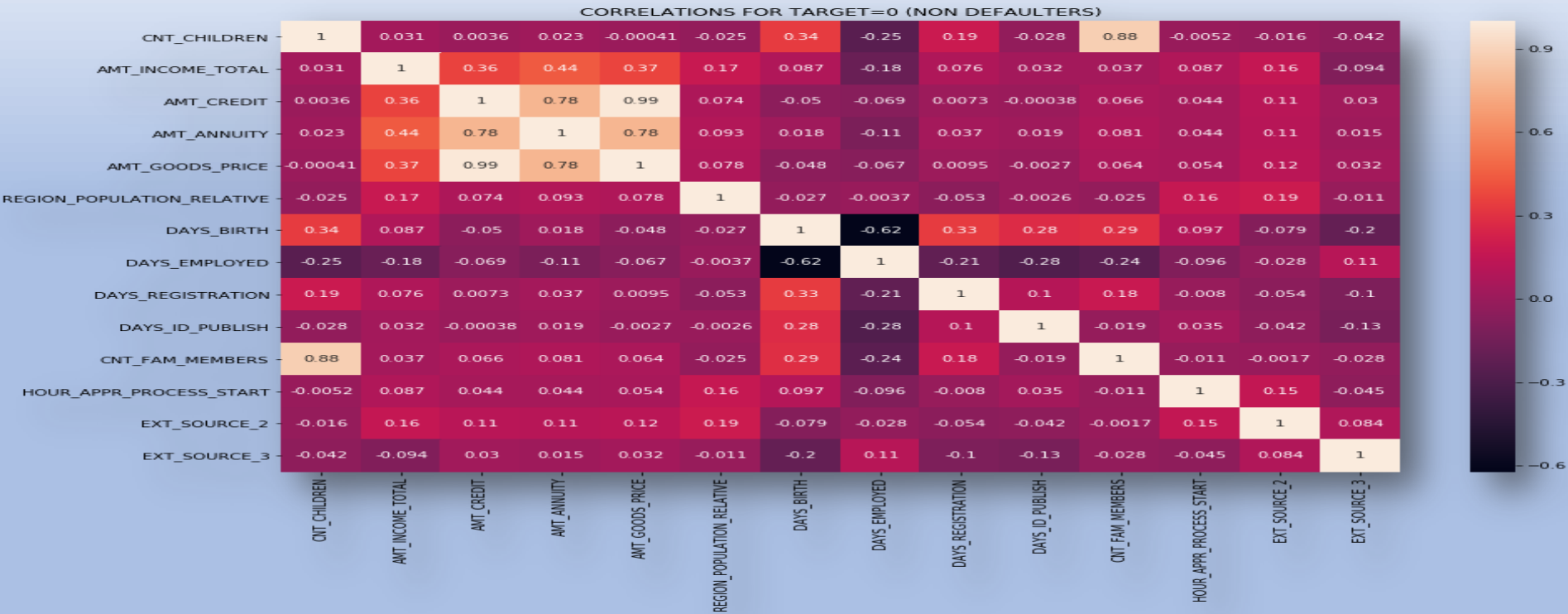
DEFAULTERS CORRELATION



TOP 10 CORRELATIONS FOR TARGET=1 (DEFAULTERS) ARE:

- 1. AMT_GOODS_PRICE & AMT_CREDIT
- 2. CNT_FAM_MEMBERS & CNT_CHILDREN
- 3. AMT_ANNUITY & AMT_CREDIT
- 4. AMT_GOODS_PRICE & AMT_ANNUITY
- 5. DAYS_EMPLOYED & DAYS_BIRTH
- 6. AMT_ANNUITY & AMT_INCOME_TOTAL
- 7. AMT_GOODS_PRICE & AMT_INCOME_TOTAL
- 8. AMT_CREDIT & AMT_INCOME_TOTAL
- 9. DAYS_REGISTRATION & DAYS_BIRTH
- 10. DAYS_BIRTH & CNT_CHILDREN

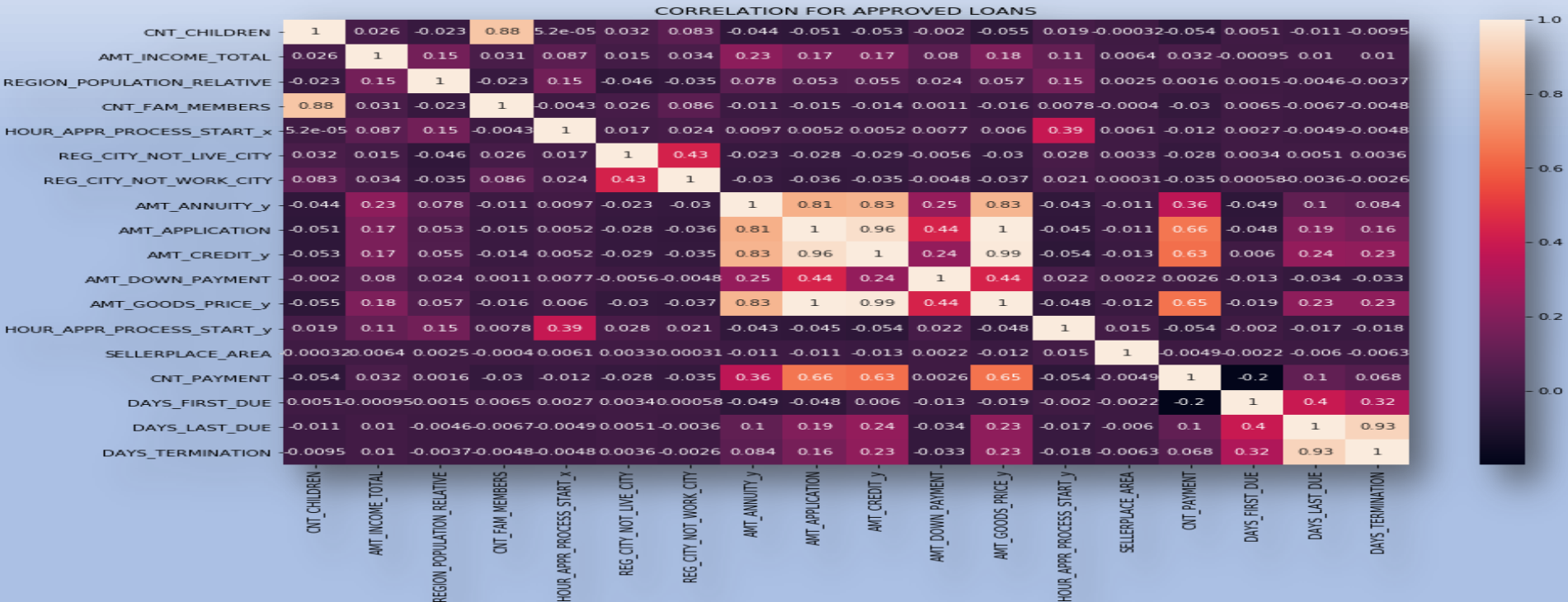
NON DEFAULTERS CORRELATION



TOP 10 CORRELATIONS FOR TARGET=0 (NON-DEFAULTERS) ARE:

- 1. AMT_GOODS_PRICE & AMT_CREDIT
- 2. CNT_FAM_MEMBERS & CNT_CHILDREN
- 3. AMT_ANNUITY & AMT_CREDIT
- 4. AMT_GOODS_PRICE & AMT_ANNUITY
- 5. DAYS_EMPLOYED & DAYS_BIRTH
- 6. AMT_ANNUITY & AMT_INCOME_TOTAL
- 7. AMT_GOODS_PRICE & AMT_INCOME_TOTAL
- 8. AMT_CREDIT & AMT_INCOME_TOTAL
- 9. DAYS_REGISTRATION & DAYS_BIRTH
- 10. DAYS_BIRTH & CNT_CHILDREN

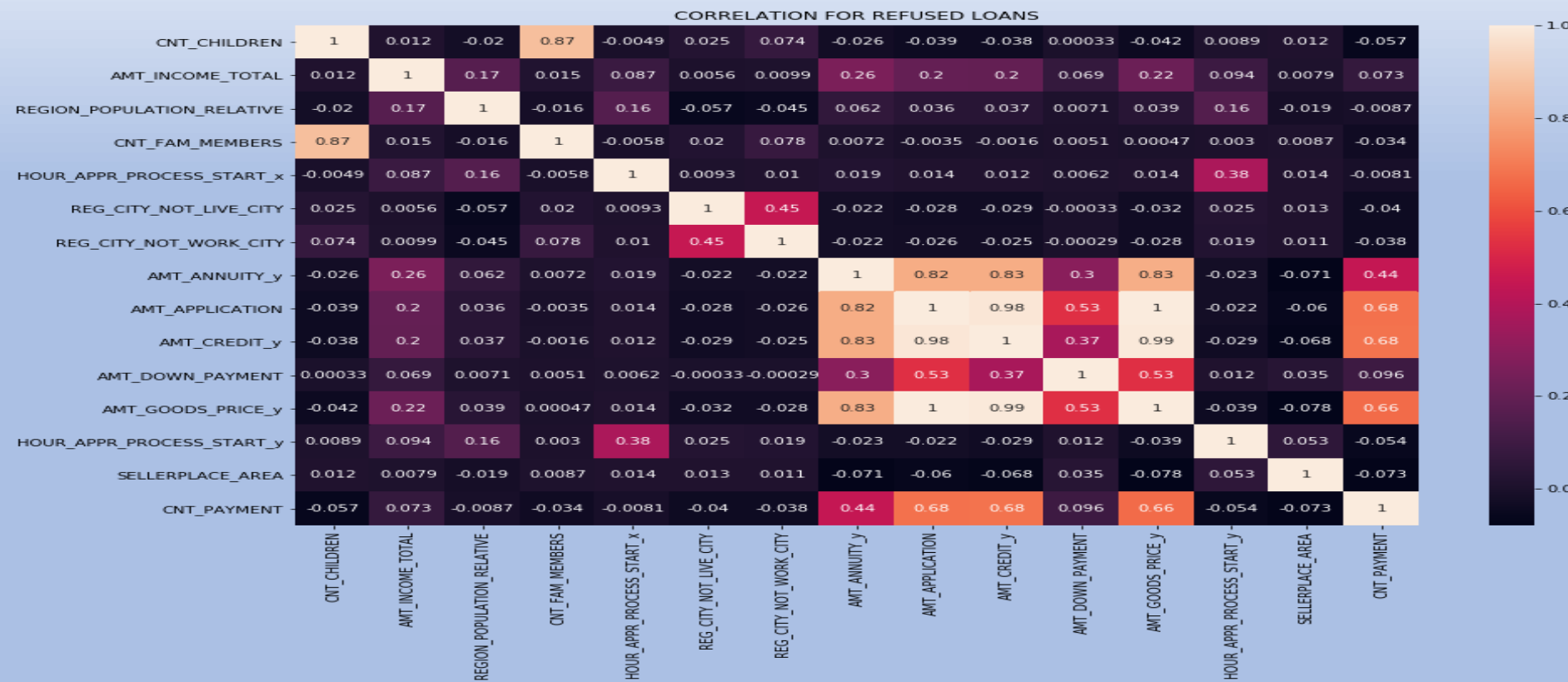
CORRELATION:- Approved Loans



TOP CORRELATED VARIABLES FOR APPROVED LOANS

1. AMT_ANNUITY_Y & AMT_APPLICATION
2. AMT_APPLICATION & AMT_CREDIT_Y
3. AMT_ANNUITY_Y & AMT_CREDIT_Y
4. AMT_GOODS_PRICE_Y & AMT_CREDIT_Y

CORRELATION : Rejected Loans



TOP CORRELATED VARIABLES FOR REFUSED LOANS

- 1. AMT_ANNUITY_y & AMT_CREDIT_y
- 2. AMT_APPLICATION & AMT_CREDIT_y
- 3. AMT_ANNUITY_y & AMT_GOODS_PRICE_y
- 4. AMT_GOODS_PRICE_y & AMT_CREDIT_y

Highlights

- The Percentage of people who have paid their loan is: 91.93 %
- The Percentage of people who have NOT paid their loan is: 8.07 %
- The Ratio of Data Imbalance is: 11.39.
- The number of **Cash loans** is much higher than the number of **Revolving loans** for both Target = 0 and Target = 1
- The number of **Females** taking loans is much higher than the number of **Males** for both Target = 0 and Target = 1.
- People with Academic Degree rarely take loans and are rarely defaulters. So they are potentially good customers.
- Higher educated clients have low defaults.
- High number of people having low years of employment are likely to default.
- Pensioner in between age 60-70 are good re-payers.
- Realty agents have high risk of default than skilled labours and drivers.
- Higher educated applicant with no of children more than 4 children has high defaults.