Credit EDA assignment

By Sayujya Vartak

Introduction

- This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. 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.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Business Objectives

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (too 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.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

EDA approach

1. Importing all important Libraries

2. Setting The Jupyter Notebook
   - Changing directory location
   - Setting view options for DataFrame
   - Suppressing unnecessary warnings

3. Loading Datasets

4. Data Understanding
   - Checking Structure of data sets
   - Checking other attributes like using info(), describe(), etc.

5. Data Cleaning
   - Deleting unnecessary columns and rows
   - Data Quality check - Analyzing missing values, improper data types, duplicated rows
   - Imputing Missing values
   - Changing improper data types

6. Binning

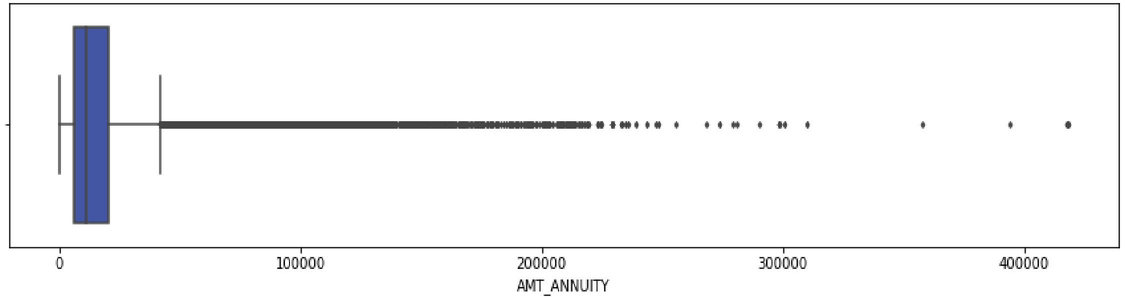
7. Handling Outliers

8. Data Analysis
   - Imbalance Analysis
   - Defining functions for plotting
   - Univariate/Bivariate and Multivariate

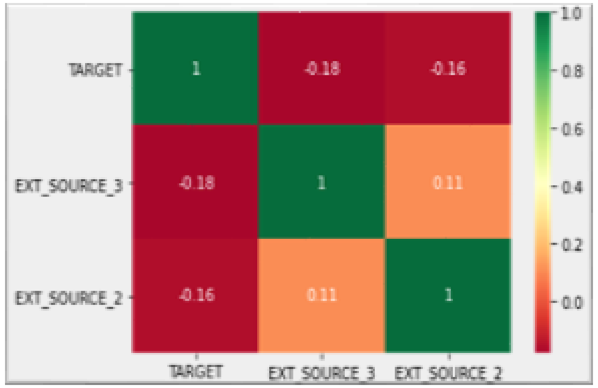
9. Merged dataset analysis
   - Imbalance Analysis
   - Defining functions for plotting
   - Univariate/Bivariate and Multivariate

Data Cleaning

- Both The datasets application.csv and previous_application.csv had many columns with greater than 40% of the data missing
- Deleting these columns is the best way to deal with them as imputing these many missing values will add bias to the data
- Plotting heatmaps of correlation matrix to determine the effect of a particular column on target variable, thus to understand which columns are useful for analysis and which to delete
- Remaining columns are imputed using appropriate methods
- Categorical columns are imputed using modal value of that particular column
- Numeric columns are imputed using mean, median or mode.
- Numeric columns with highly skewed data are imputed using median
- Using kdeplots to find best imputation value for numeric columns out of mean, median or mode if unsure of best imputation value
- Identifying outliers using boxplots
- Creating new columns by binning continuous variables into fixed size bins for range-wise analysis
- After binning, deleting the outliers which are extremely far away from the expected range



Imputing AMT_ANNUITY column with median because of highly skewed data

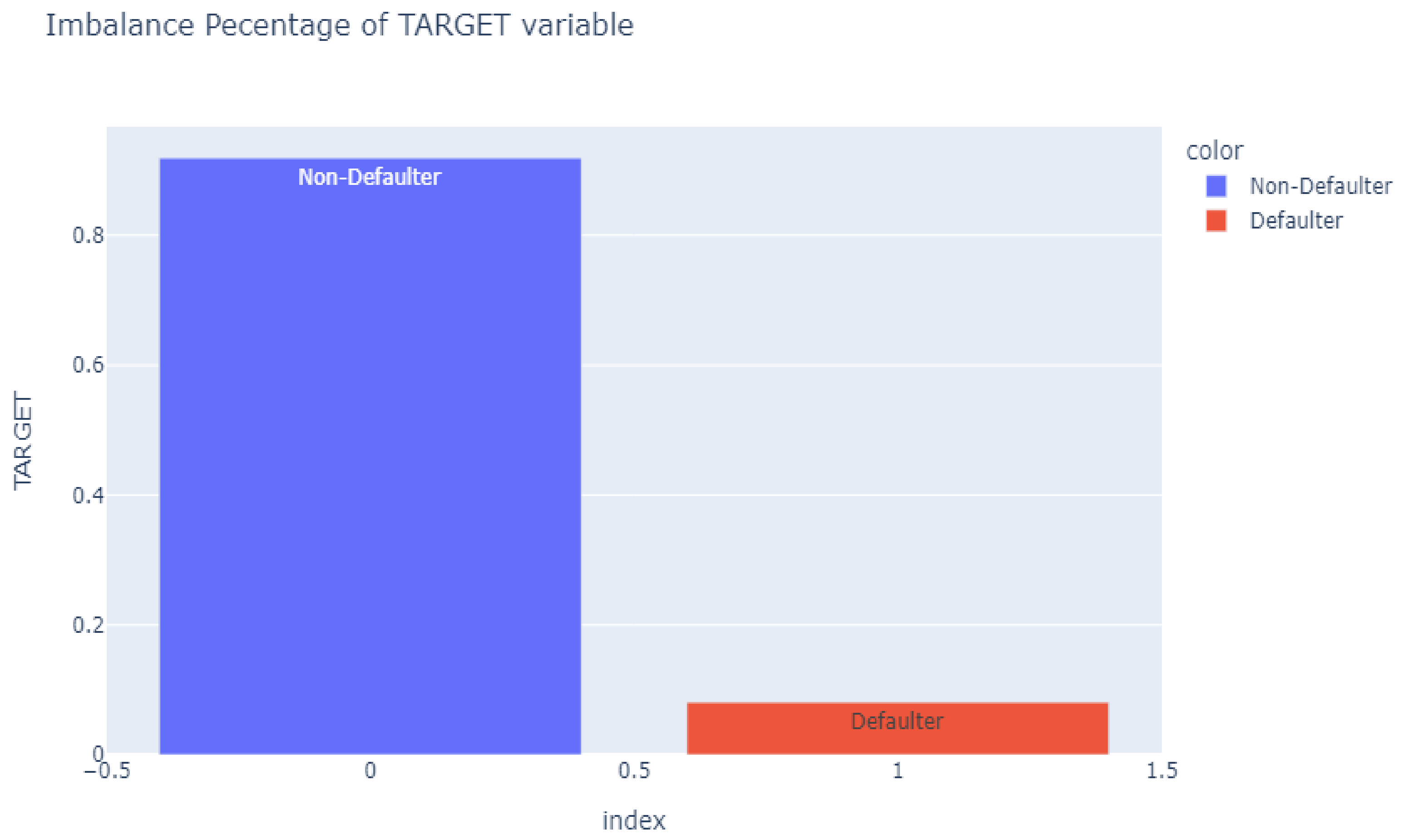


Heatmap used to determine if EXT_SOURCE_3 and EXT_SOURCE_2 are needed for analysis

Data Analysis

- Imbalance Analysis of TARGET variable
  - For application dataset and merged dataset(application + previous application)
- Segmented/Non-Segmented Univariate Analysis for categorical and numeric variables of both data set
  - For application dataset and merged dataset, segmented analysis using TARGET variable
  - For previous application dataset, segmented analysis using NAME_CONTRACT_STATUS variable
- Segmented/Non-Segmented Bivariate Analysis
  - For application dataset
- Plotting Heatmaps for application dataset (Segmented/Non-Segmented)

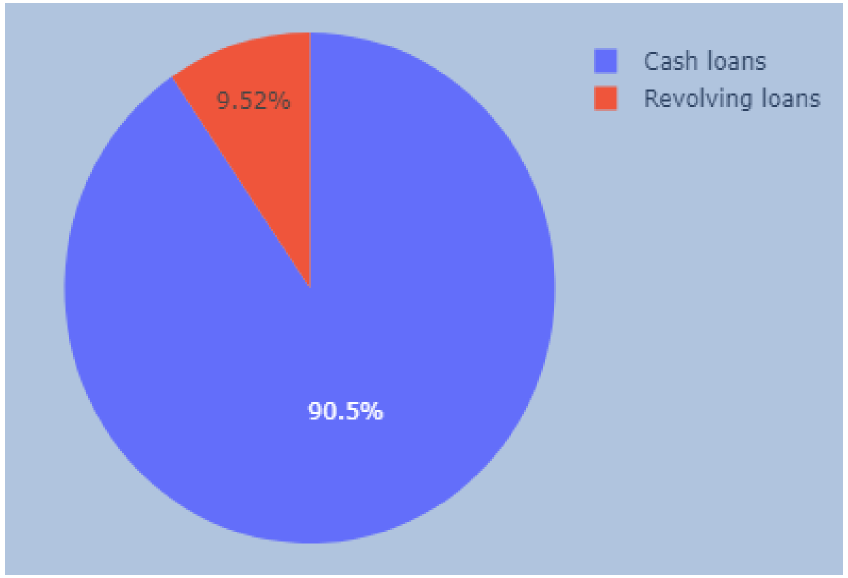
Imbalance Analysis

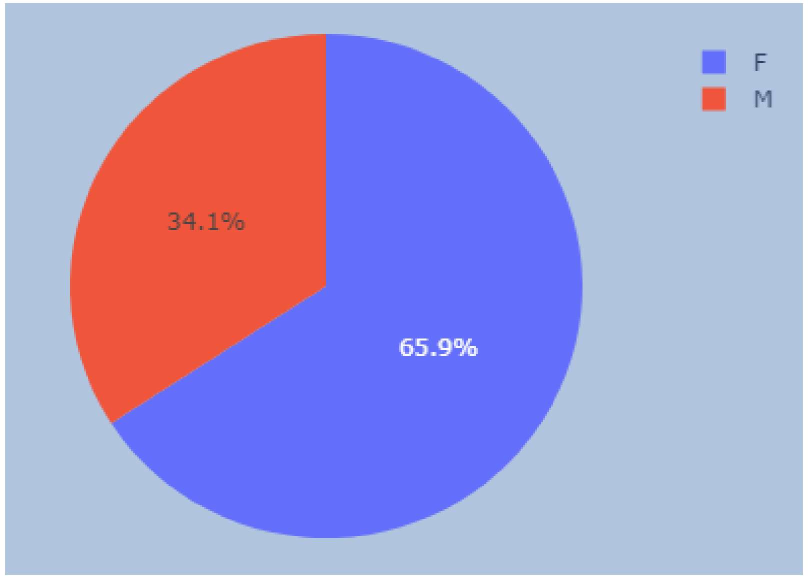


Inference

More than 91% of the applicants have paid the loan without any late payment

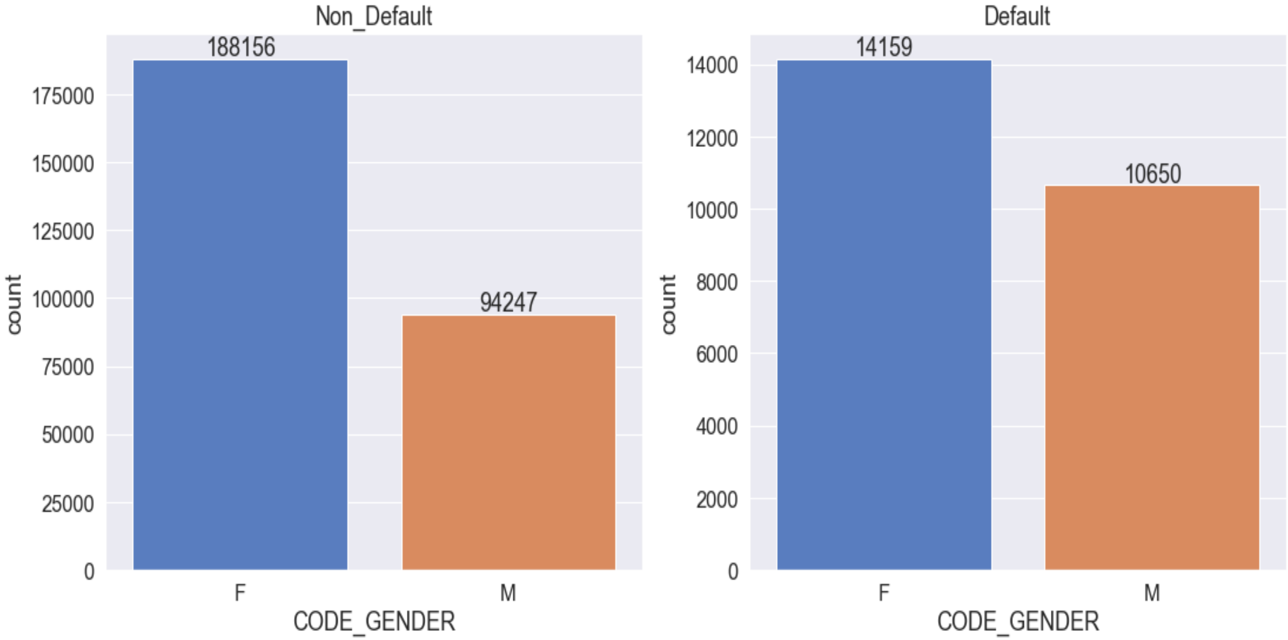
Univariate analysis on NAME_CONTRACT_TYPE and CODE_GENDER



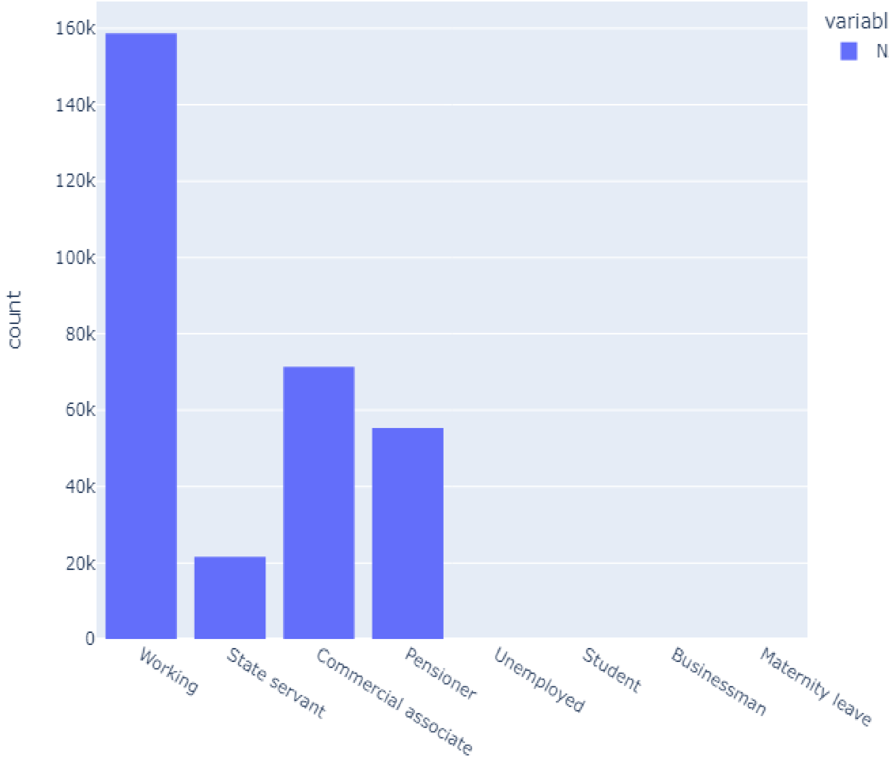


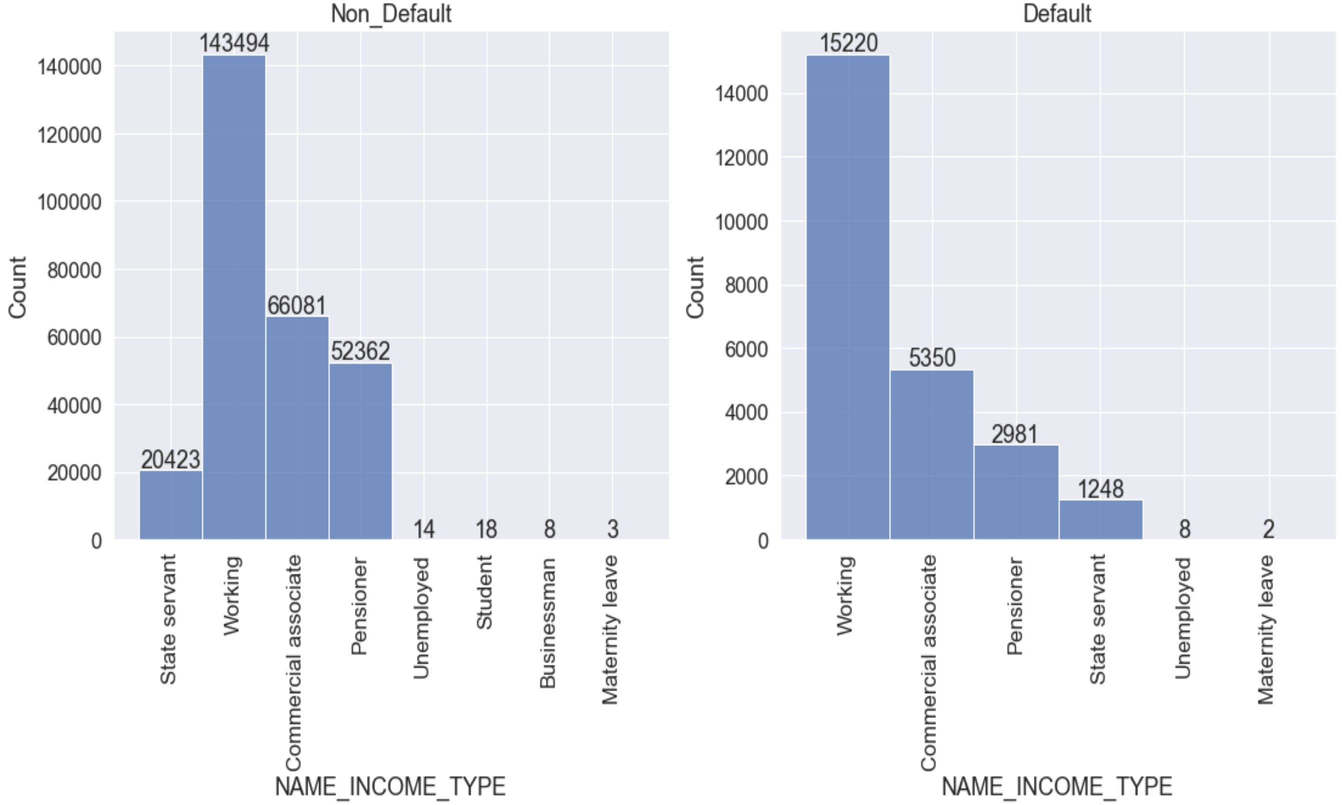
Insights

- Majority of Loans are Cash loans(~90%) and only a small percentage of values are of Revolving loans(~10%)
- Majority of current applicants are Female(65.9%) and with male applicants(34.1%) being about half the number of female applicants
- Number of Female applicants are higher in both Non Defaulting as well as Defaulting applicants
- The Male applicants are about half the number of Female applicants for non-defaulters
- While the Male applicants and Female applicants are close in numbers for defaulters



Univariate analysis on NAME_INCOME_TYPE

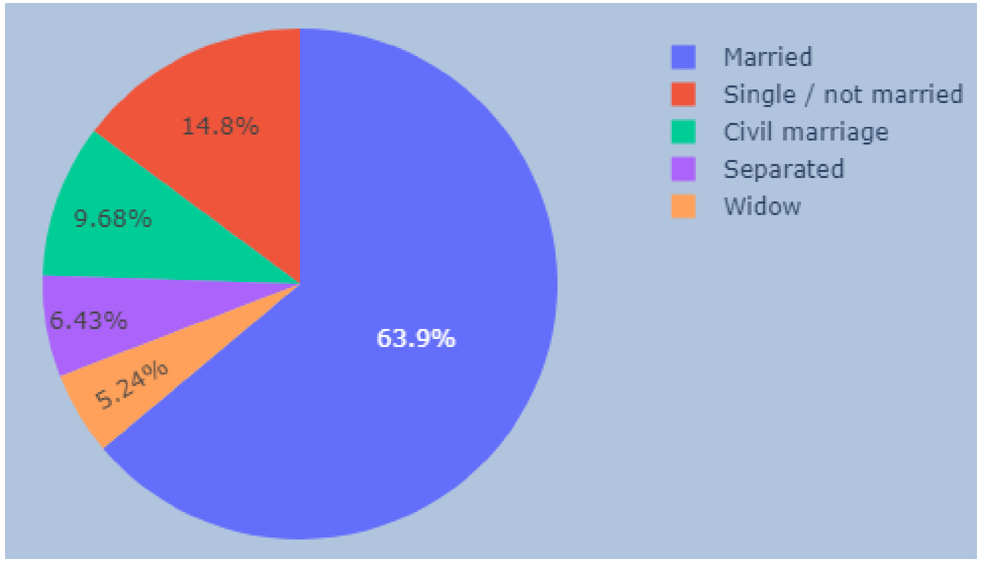


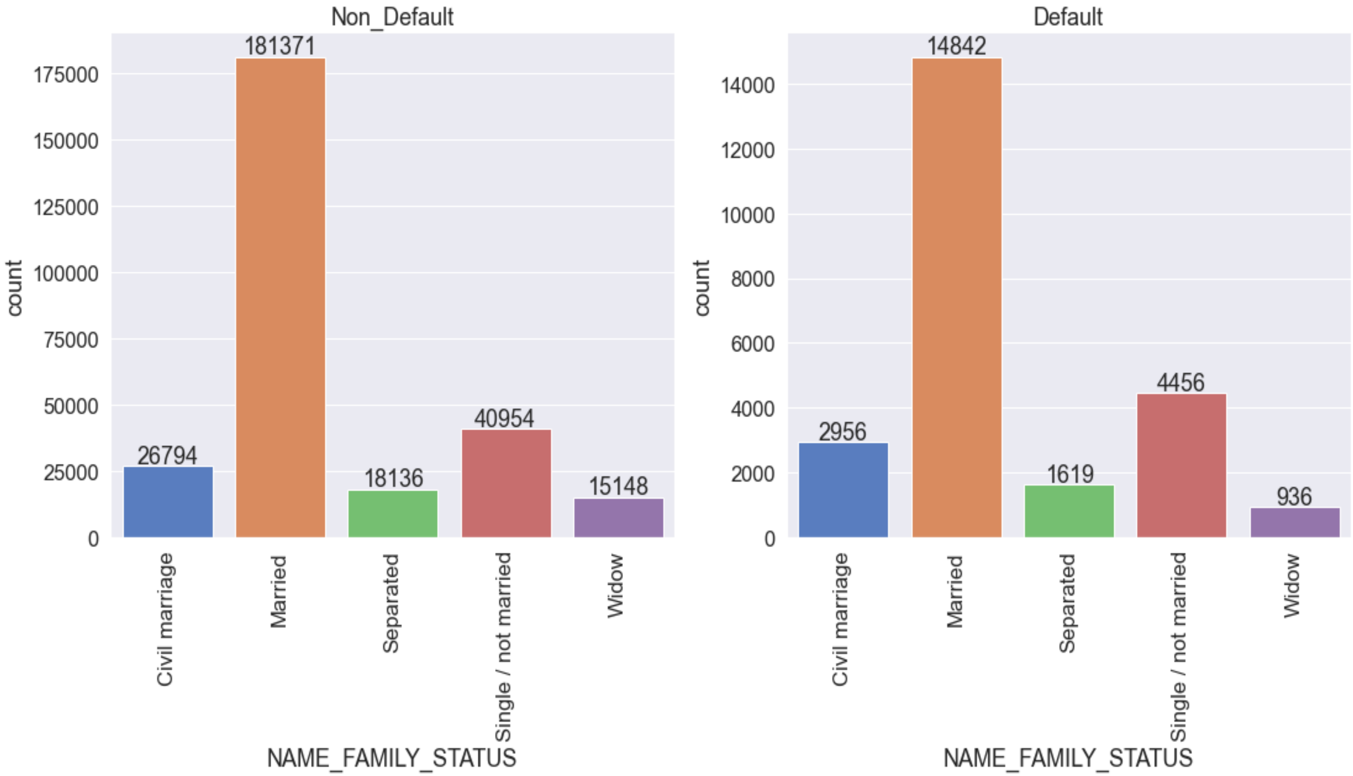


Insights

- Majority of applicants have Income type 'Working' with Commercial associates and Pensioners being 2nd and 3rd and State Servant being 4th
- ***Maternity leave*** and ***Unemployed*** though minimal in numbers have highest number of defaulter which is about 40% for ***Maternity leave*** and about 36% for ***Unemployed*** ; rest have less than 10% defaulters
- The applicants with income type as ***Pensioner*** and ***State servant*** though only 18% and 7% of all the income types has the least amount of defaulters(~5%)
- The applicants with income type ***Businessman*** and ***Student*** though minimal in numbers have no history of defaulting.

Univariate analysis on NAME_FAMILY_STATUS

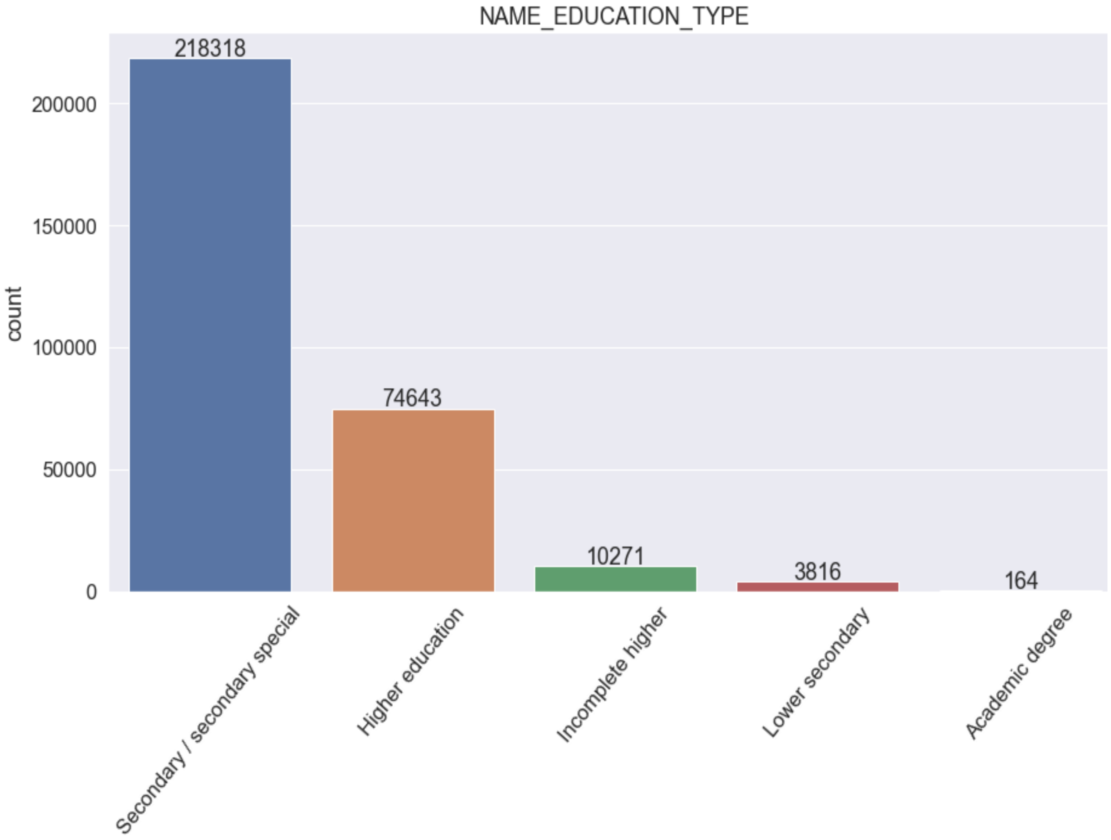


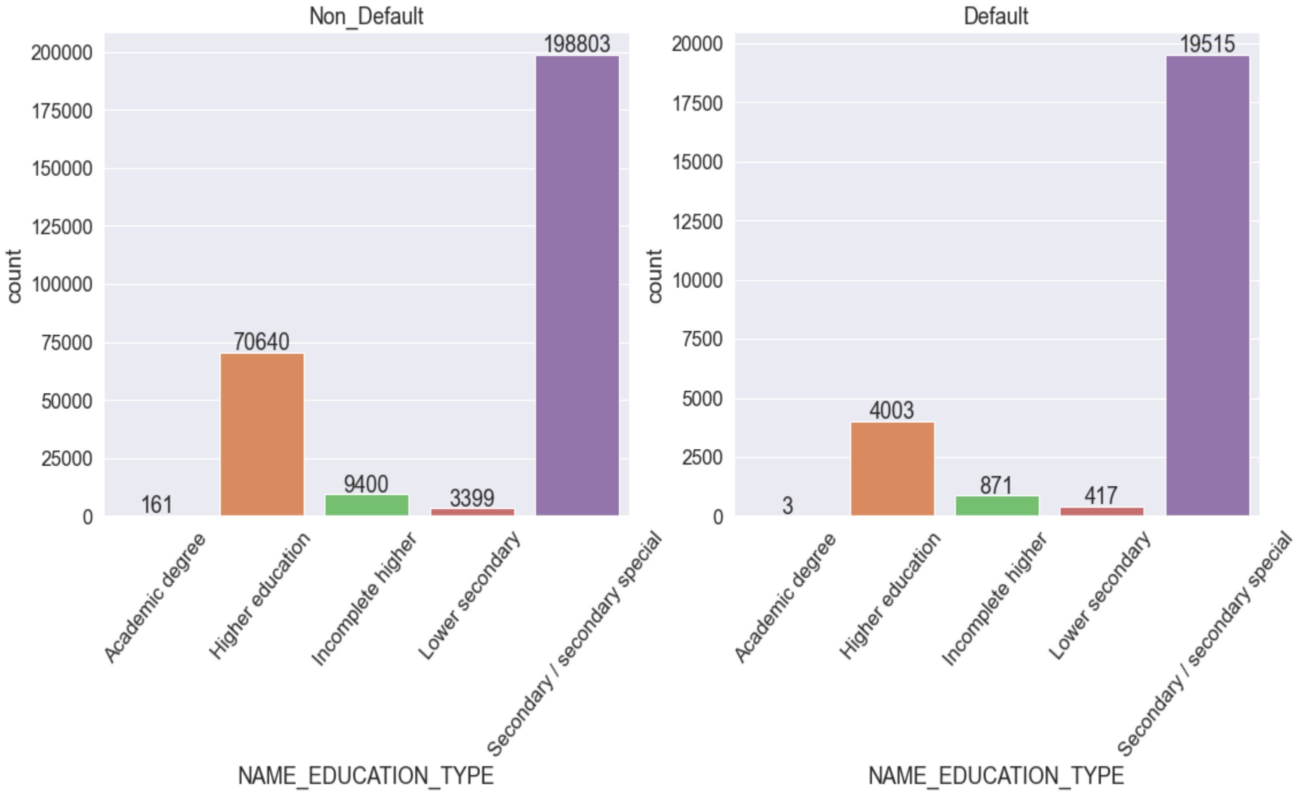


Insights

- Majority(~63%) of applicants are ***Married*** followed by ***Single/not married***, ***Civil marriage***, ***Separated*** and ***Widow*** being 14.8%, 9.68%, 6.43% and 5.24%
- ***Civil marriage*** applicants have highest percentage of defaulters (10%) and ***Widow*** have the least defaulting applicants(~6%)

Univariate analysis on NAME_EDUCATION_TYPE

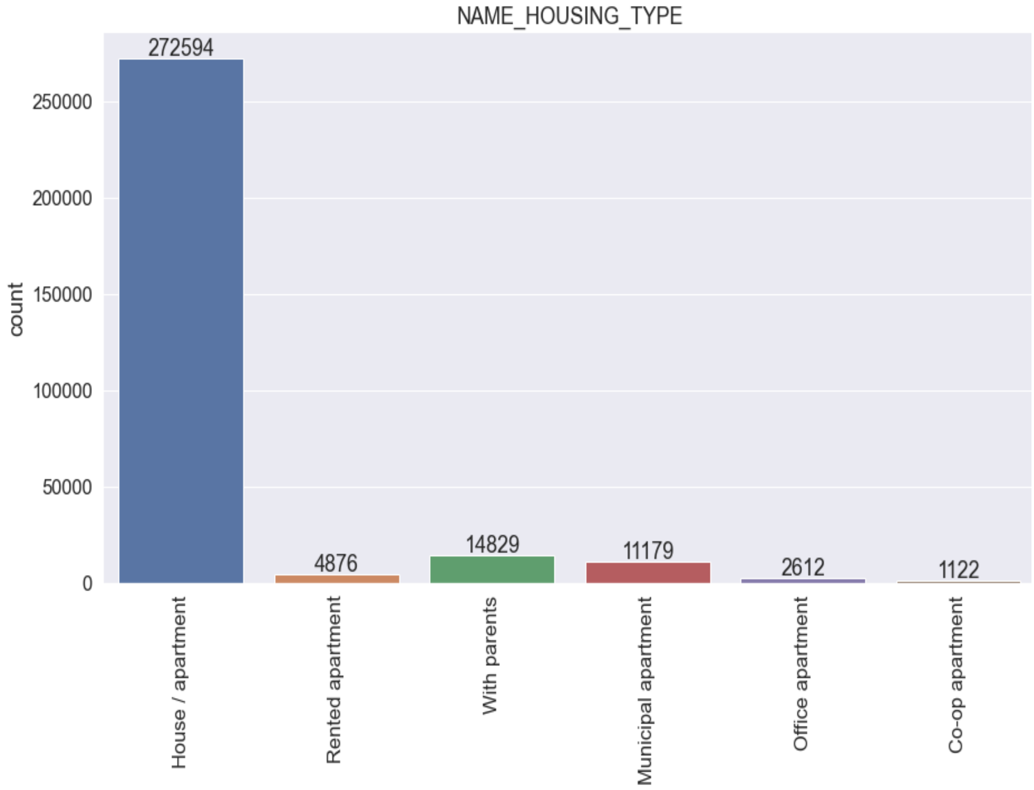


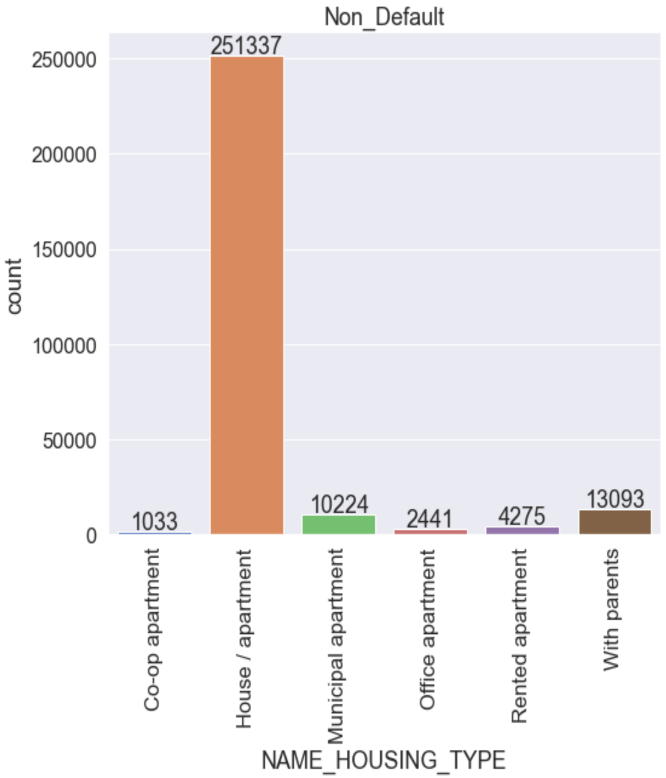


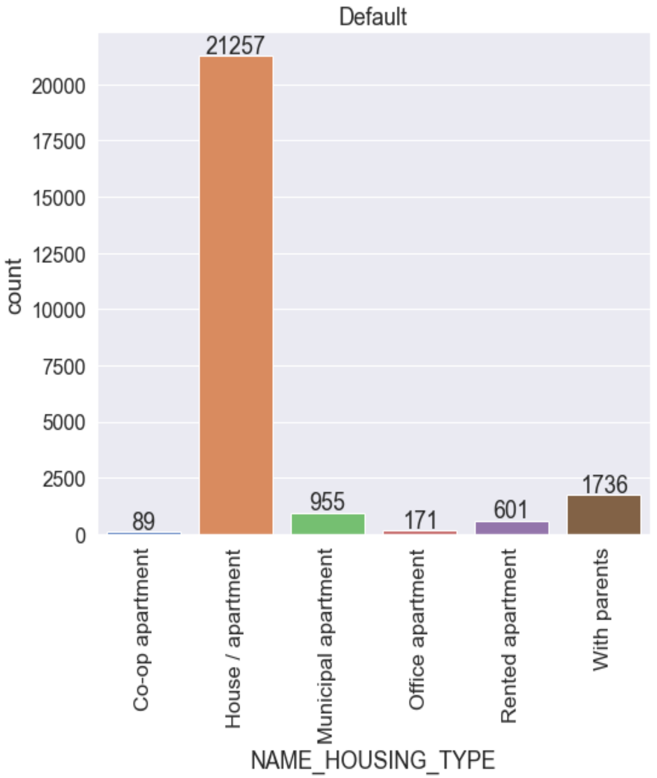
Insights

- Majority of total applicants have highest education level as ***Secondary/ secondary special***
- Applicants with highest level of education as ***Lower secondary*** have highest probability of defaulting (>10%)
- Applicants with highest level of education as ***Academic degree*** have the least probability of defaulting (<2%)

Univariate analysis on NAME_HOUSING_TYPE







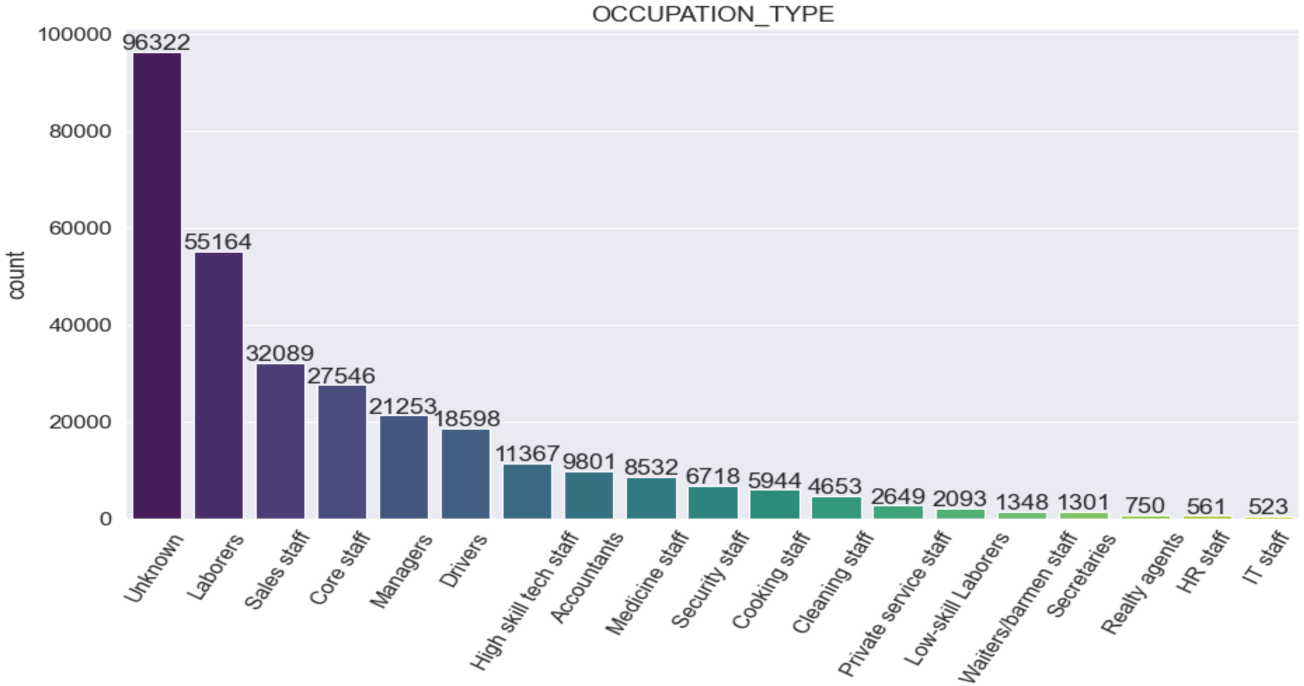
Insights

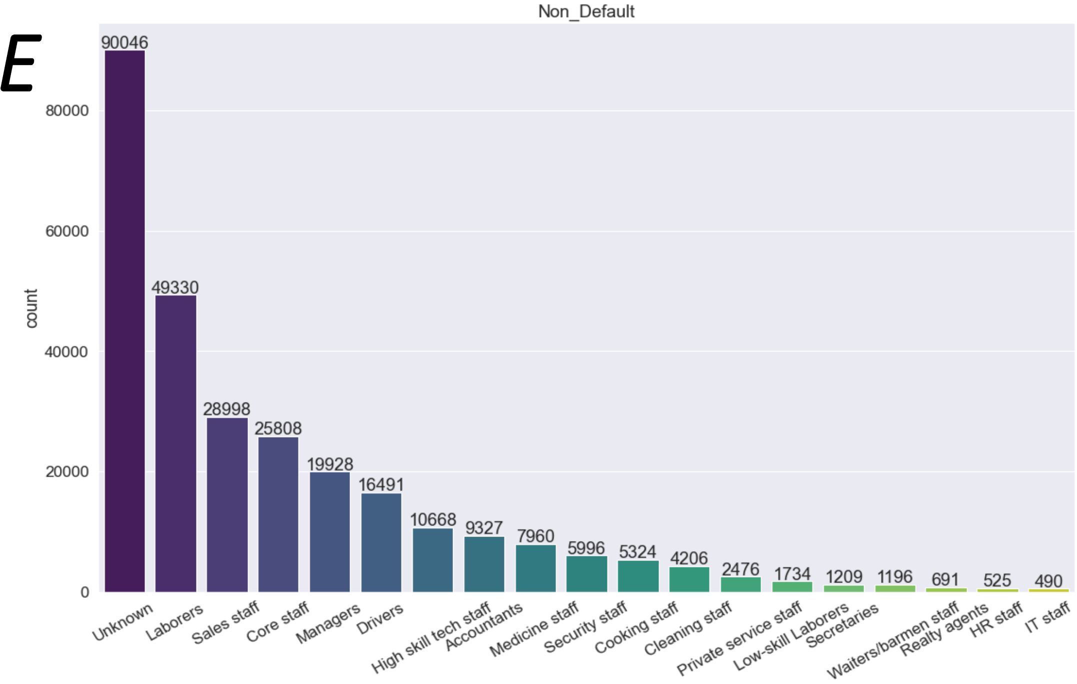
- About 88% of housing types is ***House/apartment***
- applicants living in ***Rented apartment*** and ***With parents*** have higher probability of defaulting which in about >12% for ***Rented apartment*** and about 13% for ***With parents***
- applicants living in ***Office apartments*** have lowest probability of defaulting(~6%)

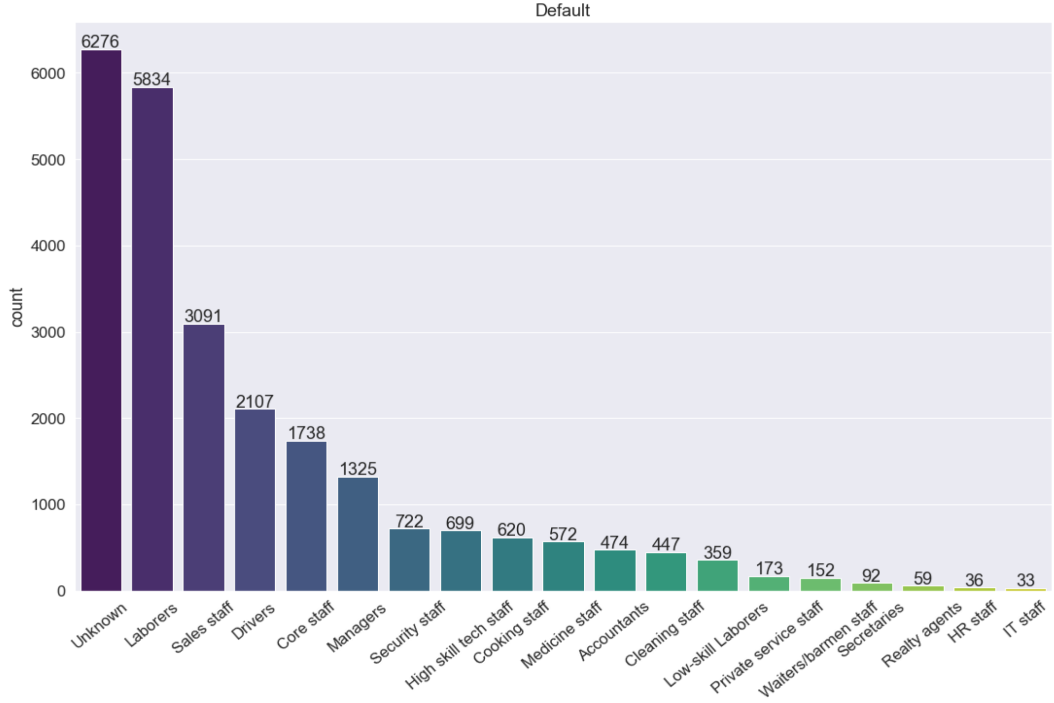
Univariate analysis on OCCUPATION_TYPE

Insights

- Majority of applicants have their occupation type ***Laborers*** or ***Sales staff***
- There are lot of applicants with their occupation type as ***Unknown***
- The most defaulting occupation type is ***Low-skill Laborers*** with about 17% of total applicants defaulting
- The other occupation types with high defaulting percentages in order are ***Drivers (>11%)***, ***Waiters/barmen staff (>11%)***, ***Security staff (>10%)***, ***Laborers (>10%)*** and ***Cooking staff (>10%)***
- Rest of the occupation types have less than 10% defaulters of the total number of applicants of that particular occupation type with the lowest defaulting occupation being ***Accountants*** with only 5% defaulting rate



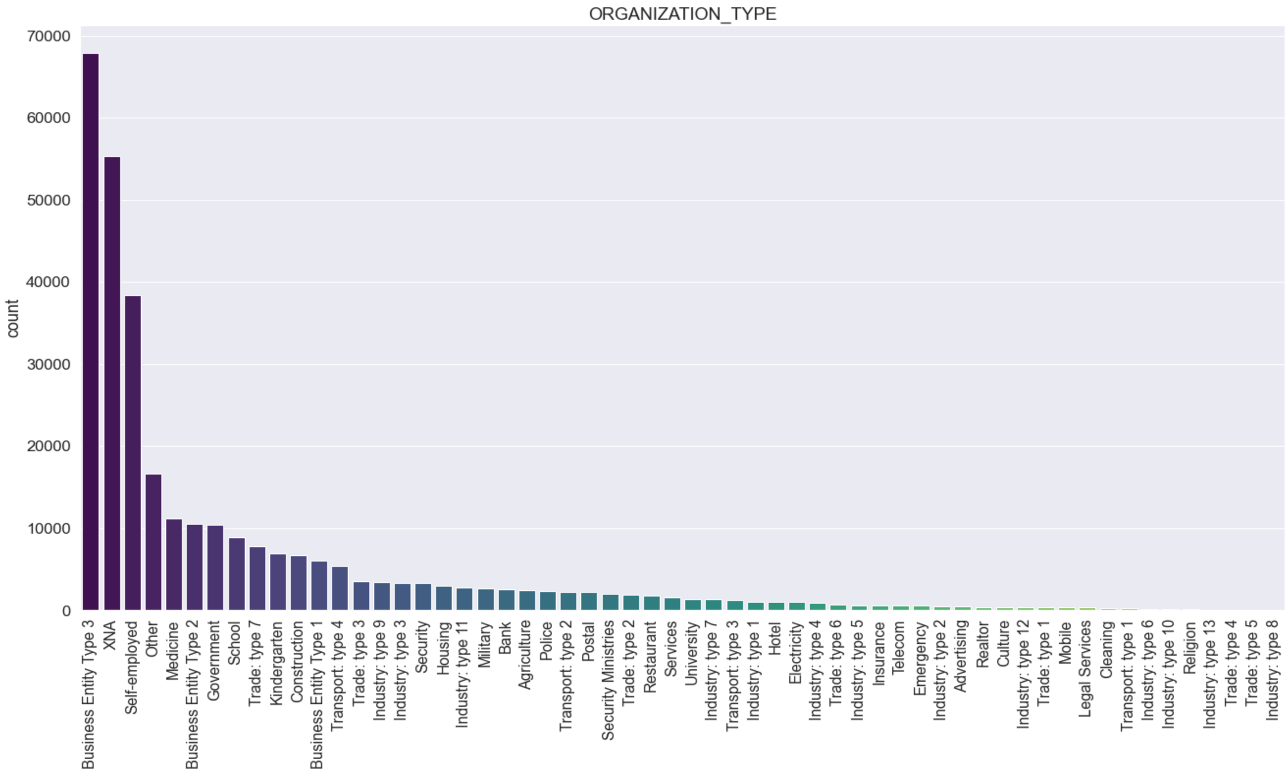


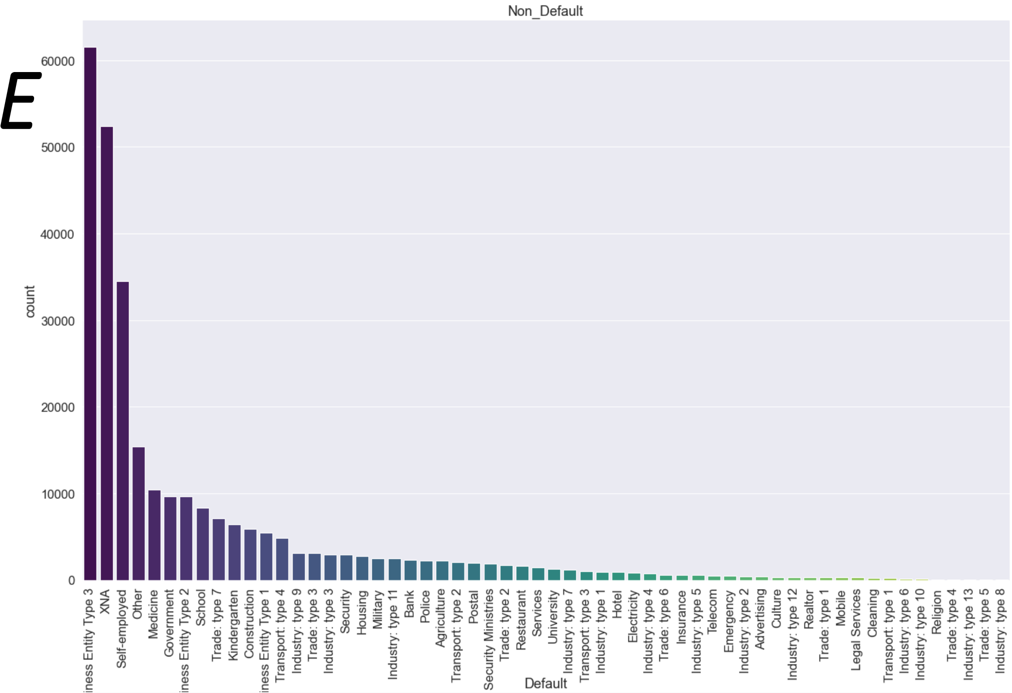


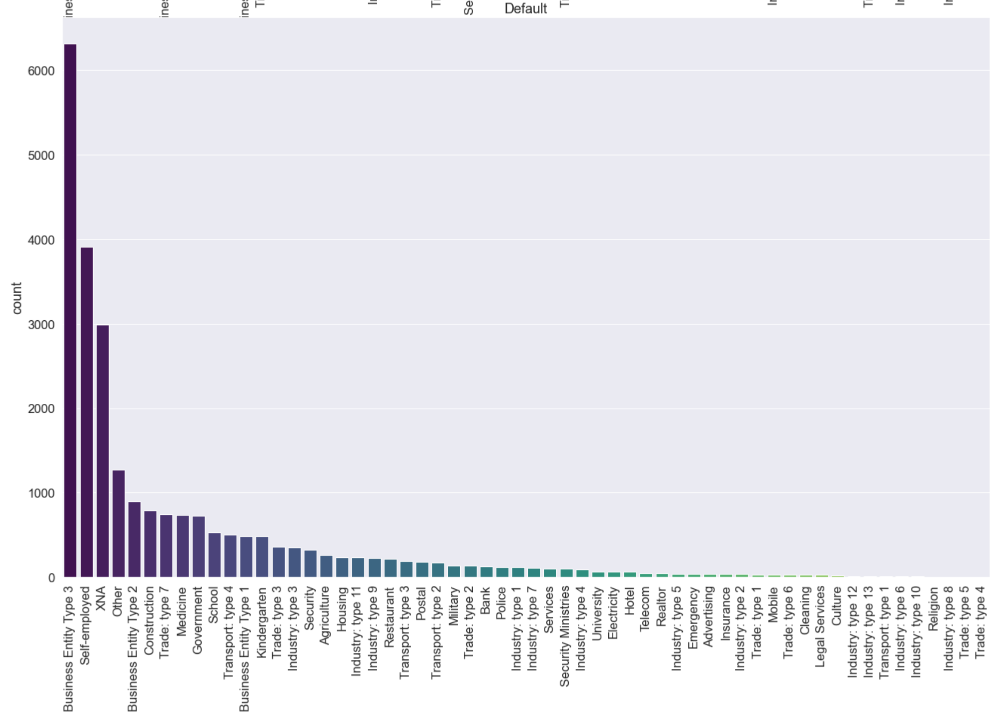
Univariate analysis on ORGANIZATION_TYPE

Insights

- The most defaulting Organization type is ***Transport: type 3*** with >15% defaulting rate
- The least defaulting Organization type is ***Trade: type 4*** with only about 3% applicants defaulting of the total applicants in Trade: type 4 Organization type



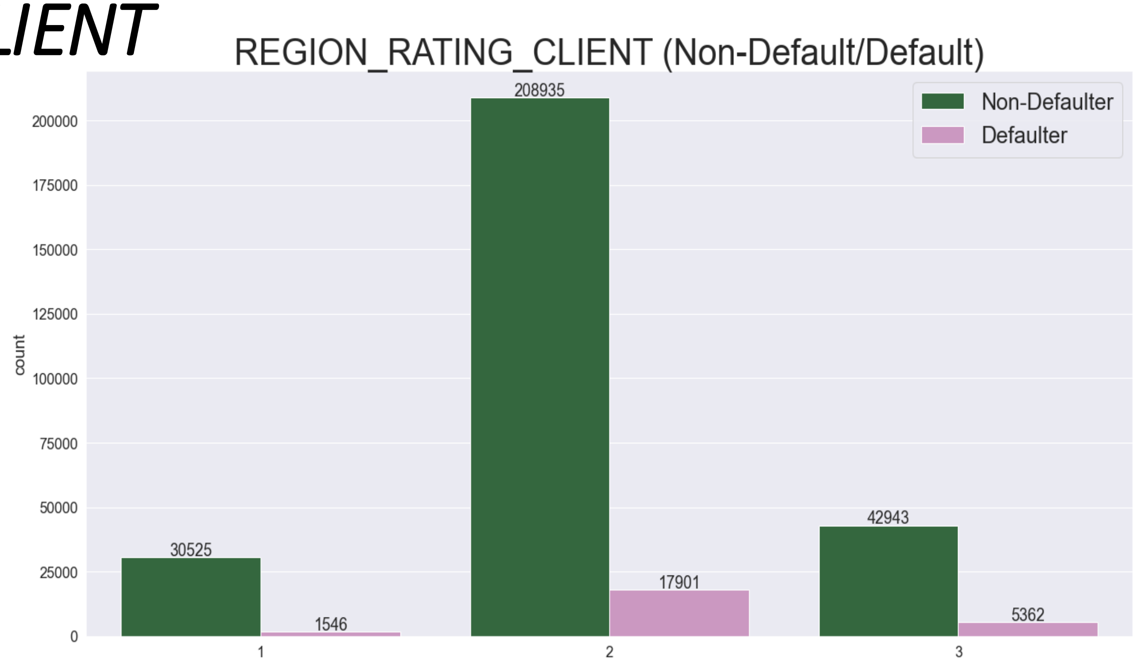


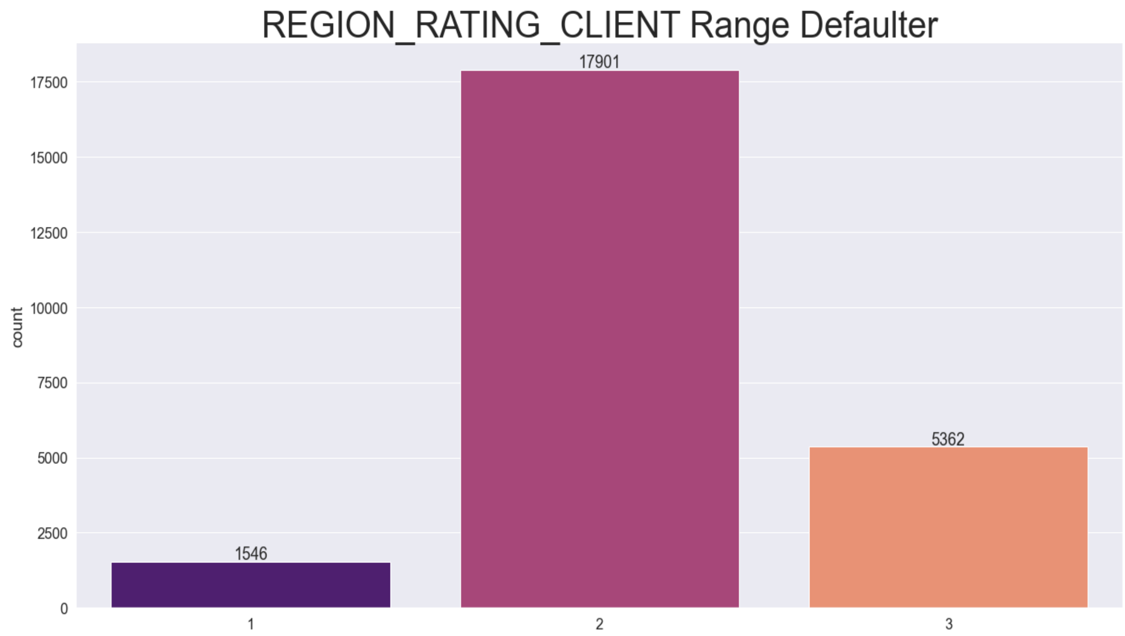


Univariate analysis on REGION_RATING_CLIENT

Insights

- clients residing in regions with rating 1 has the least probability of defaulting (<5%) while clients residing in regions with rating 3 has the highest probability of defaulting (>11%)



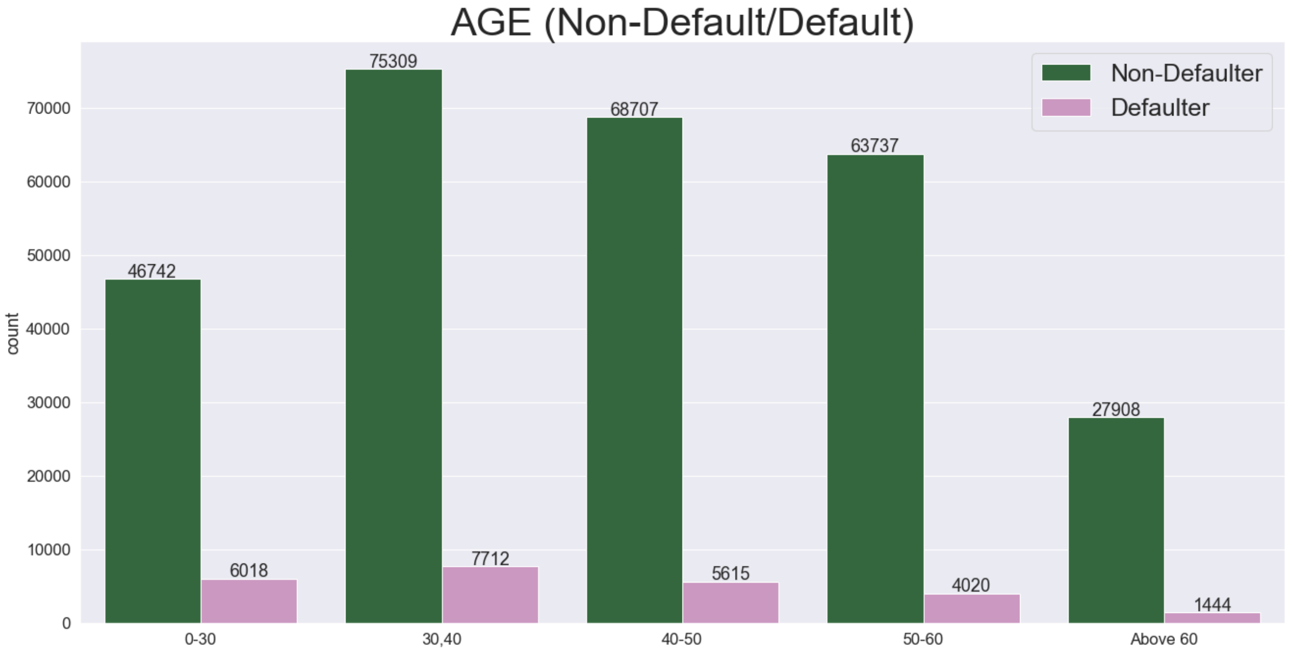


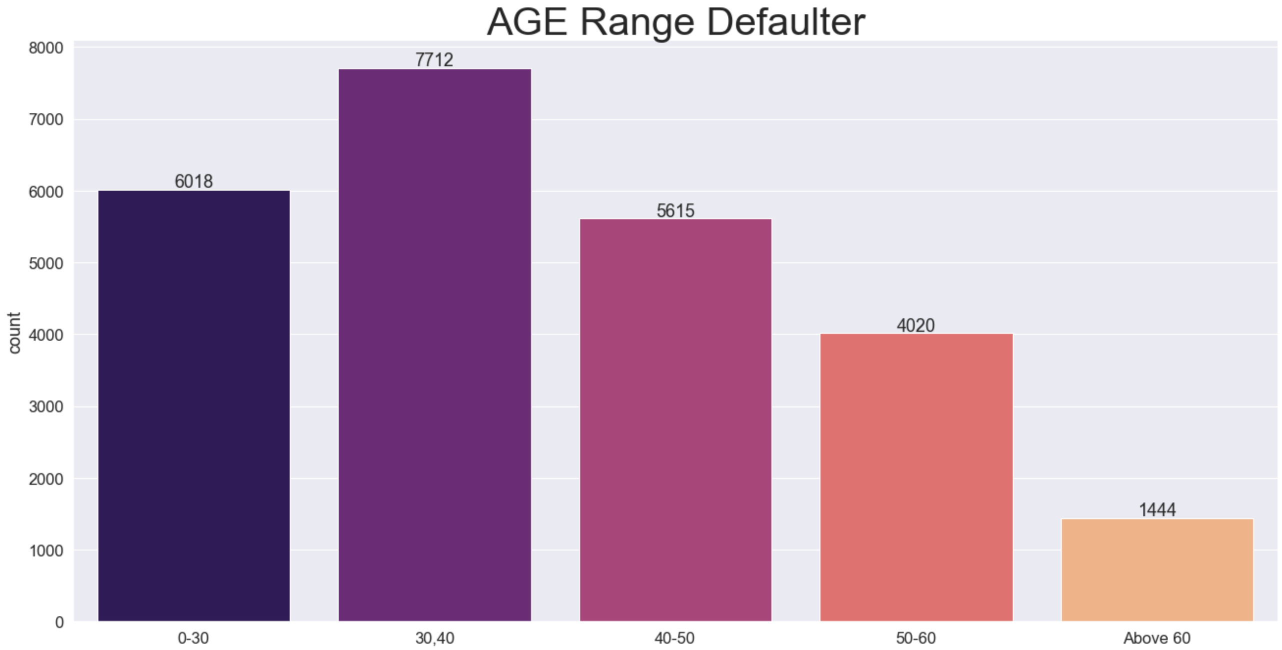
Univariate analysis on AGE

Insights

- Majority of applicants are between the age ***30-60***
- The Highest number of defaulters are in the age range ***30-40*** followed by ***0-30*** then ***40-50*** then ***50-60*** and finally ***Above 60***

- The Age range with higher probability of defaulting is ***0-30*** with about 11% of total applicants defaulting out of total applicants in this range
- The Least defaulting applicants are ***Above 60*** with only ~5% defaulting rate





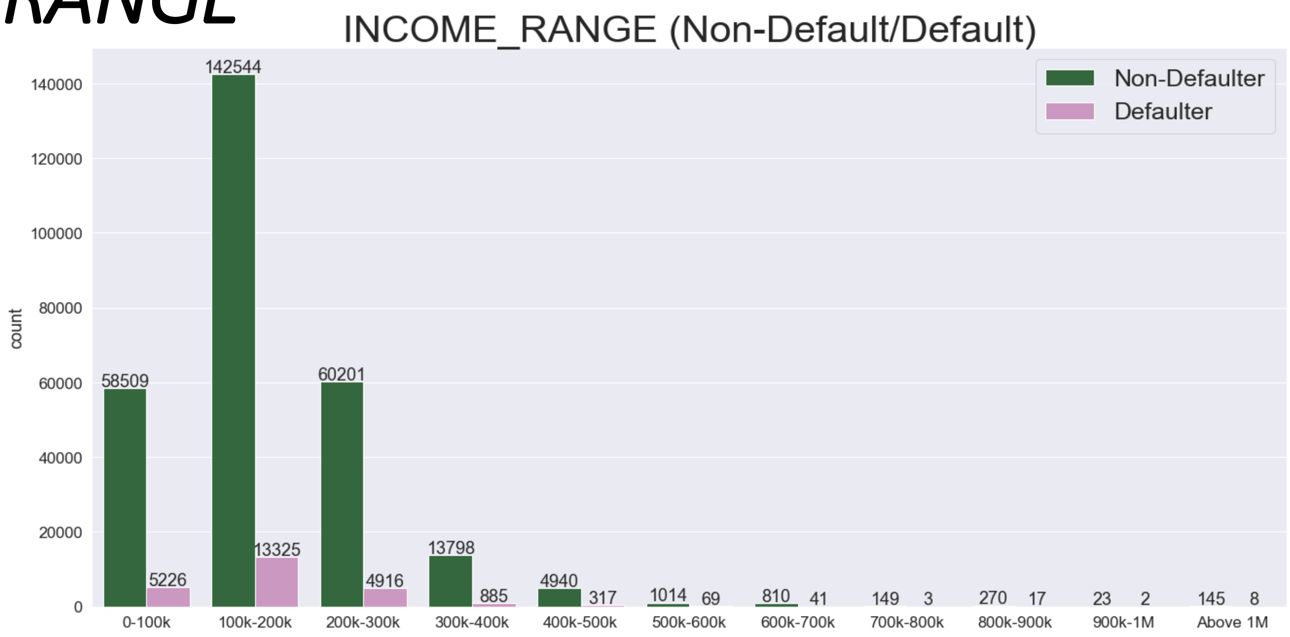
Univariate analysis on INCOME_RANGE

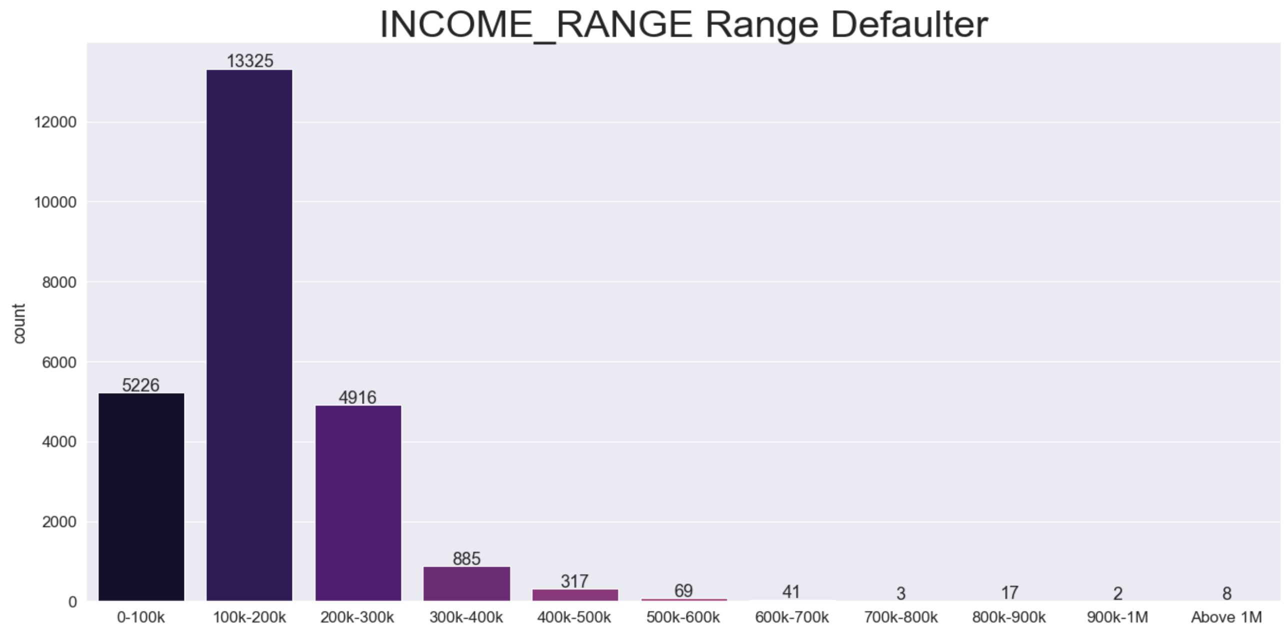
Insights

- Majority of applicants are between the income range ***0-300k***
- The Highest number of defaulters are in the income ranges ***100k-200k***(>8%), ***0-100k***(~8%) and ***900k-1M***(8%)
- applicants in the income range 700k-800k though very less in number have the least defaulters with less than 2% of applicants defaulting

Inference

- People with income less than 200k are more likely to default than people with income above 200k, the exception being income range ***900k-1M*** which also has high percentage of defaulters





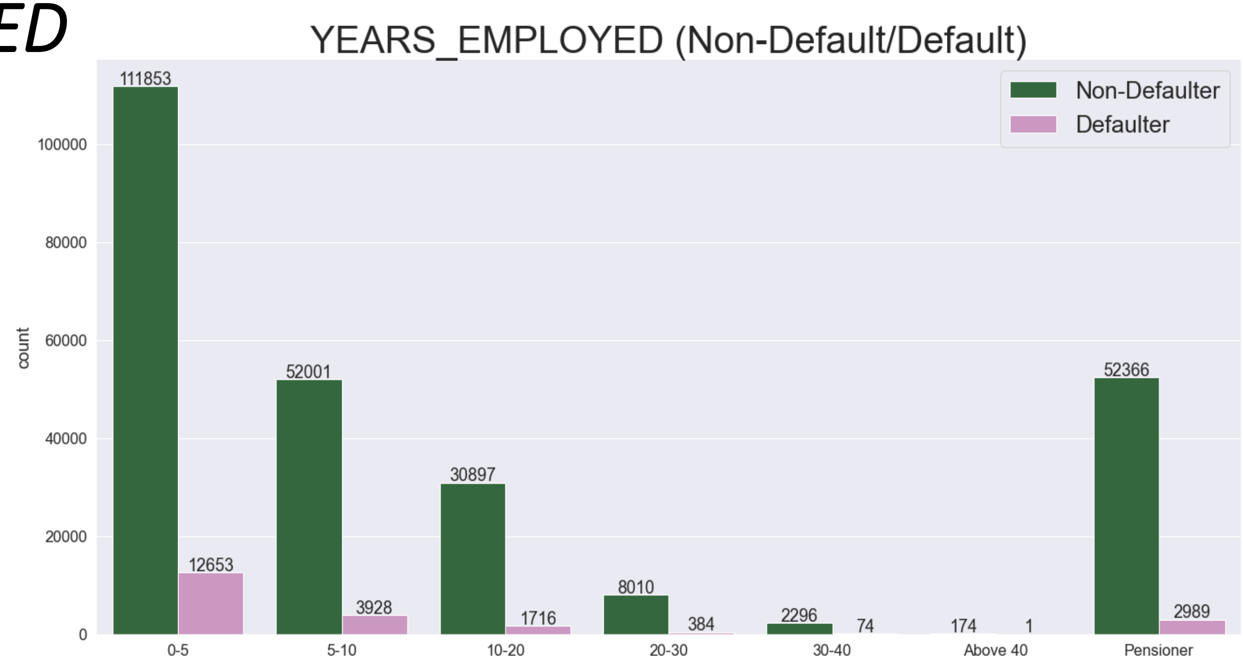
Univariate analysis on YEARS_EMPLOYED

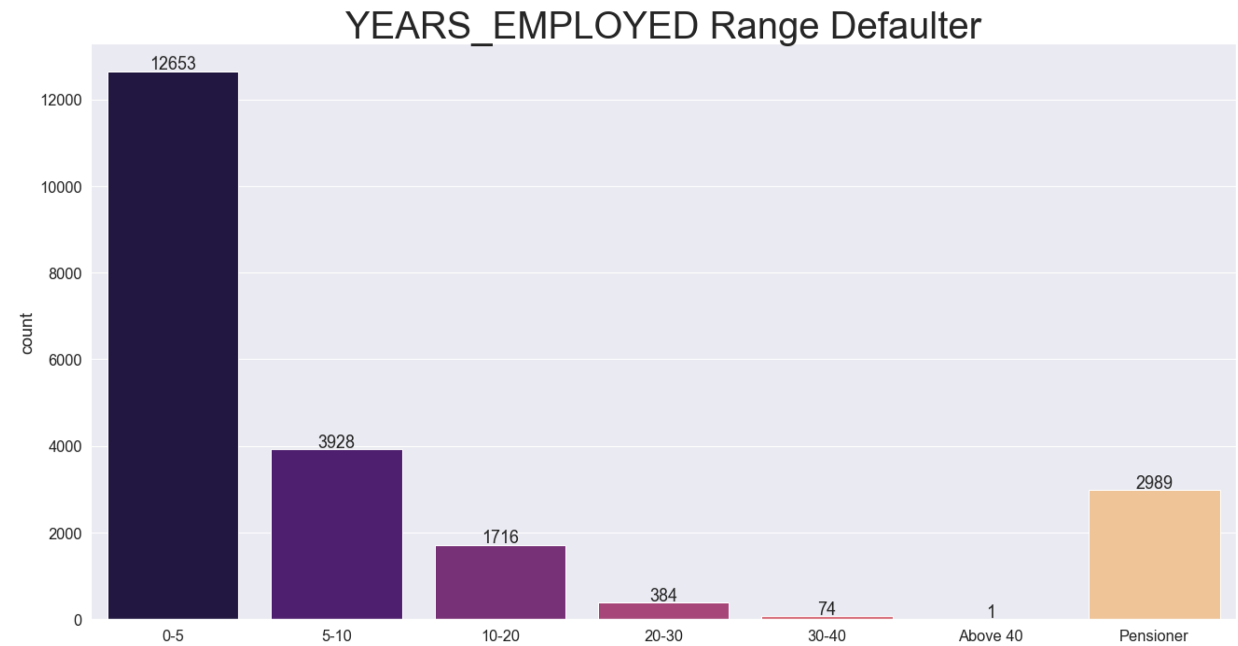
Insights

- Majority of applicants are either been employed for less than 10 years or are retired Pensioners
- The majority of defaulters are also either been employed for less than 10 years or are retired Pensioners with highest number of defaulters being in the ***0-5 (~10%)*** years ranges and lowest being ***Above 40 (Less than 1%)***

Inferences

- applicants who are employed for more than 40 years have the least probability of defaulting and are safer for loan approval
- As the years employed increases the probability of applicants being defaulter decrease

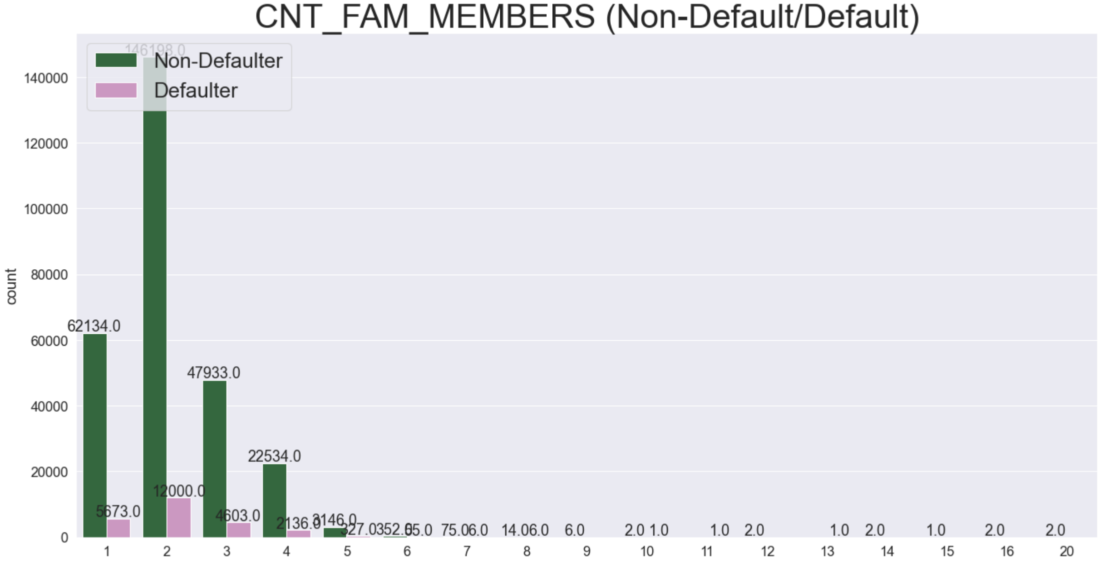


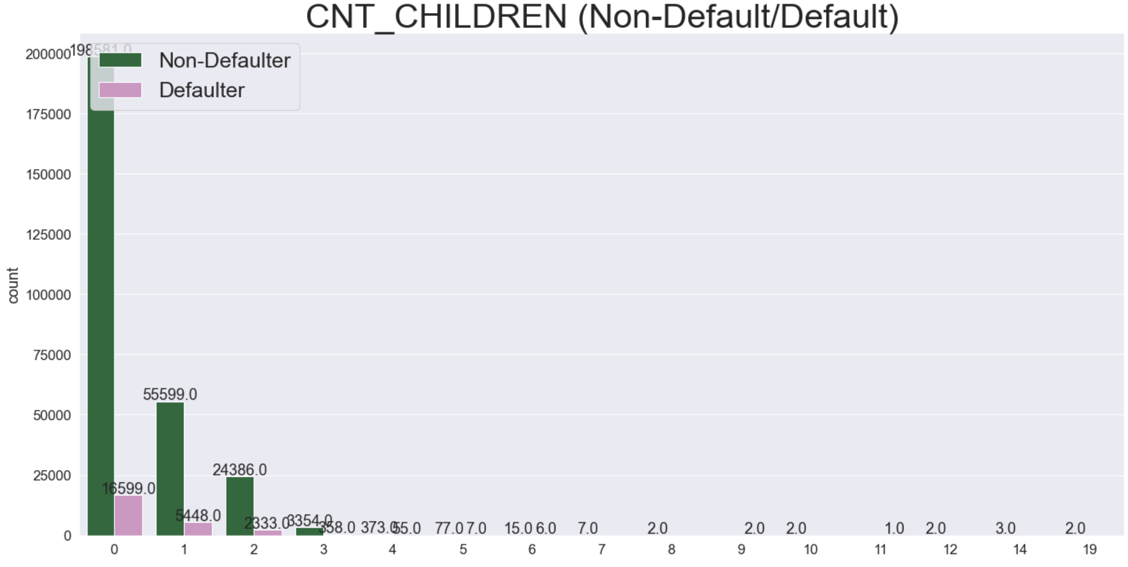


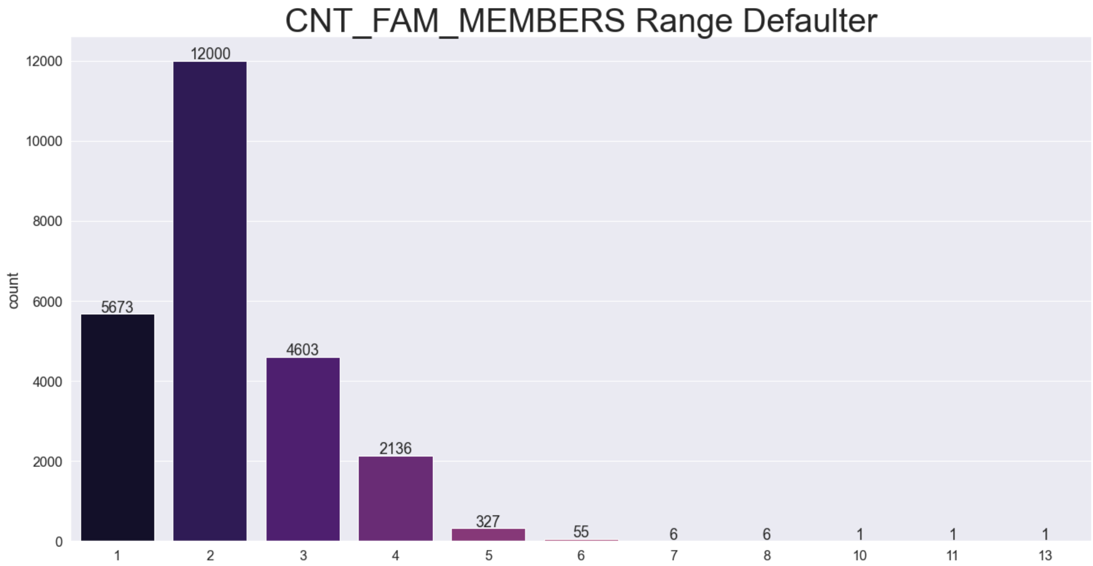
Univariate analysis on CNT_CHILDREN and CNT_FAM_MEMBERS

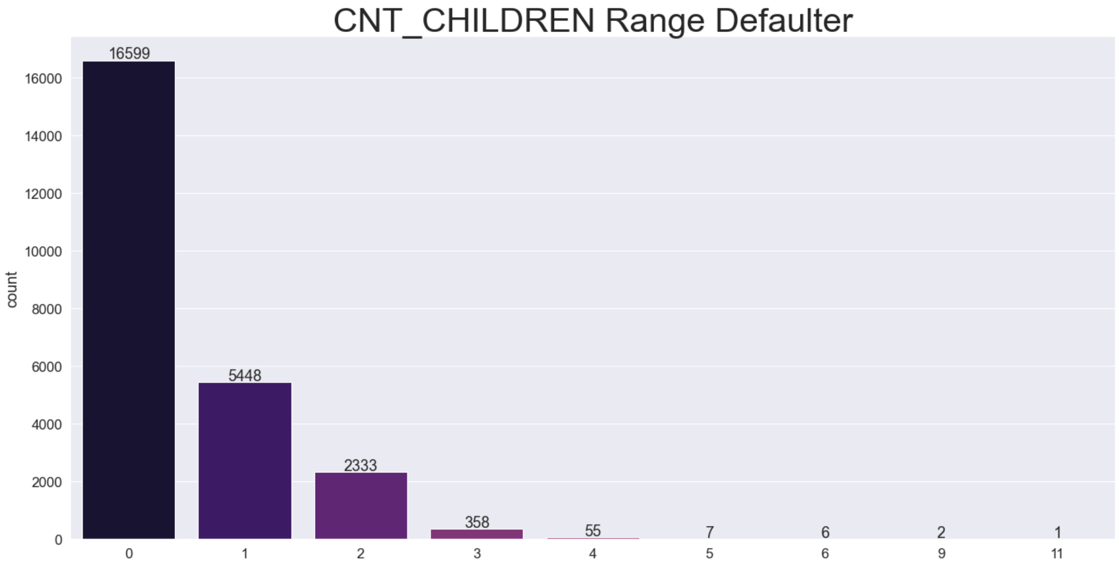
Insights

•As the number of family members increase, probability of applicants defaulting also increase and we can see the same trend fo r number of children of an applicant









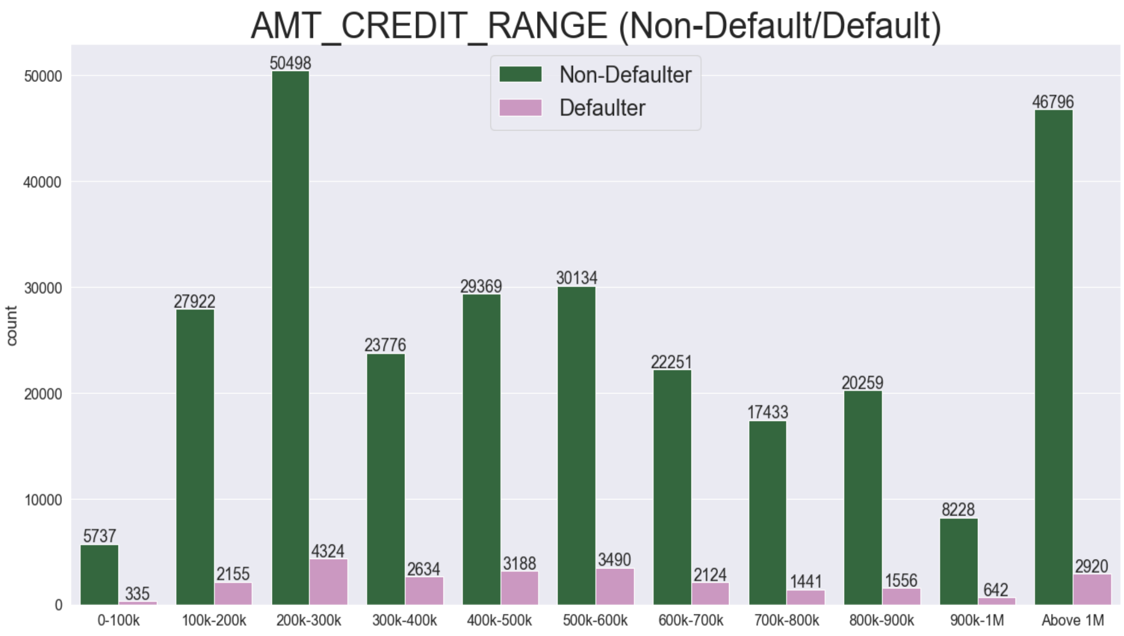
Univariate analysis on AMT_CREDIT_RANGE

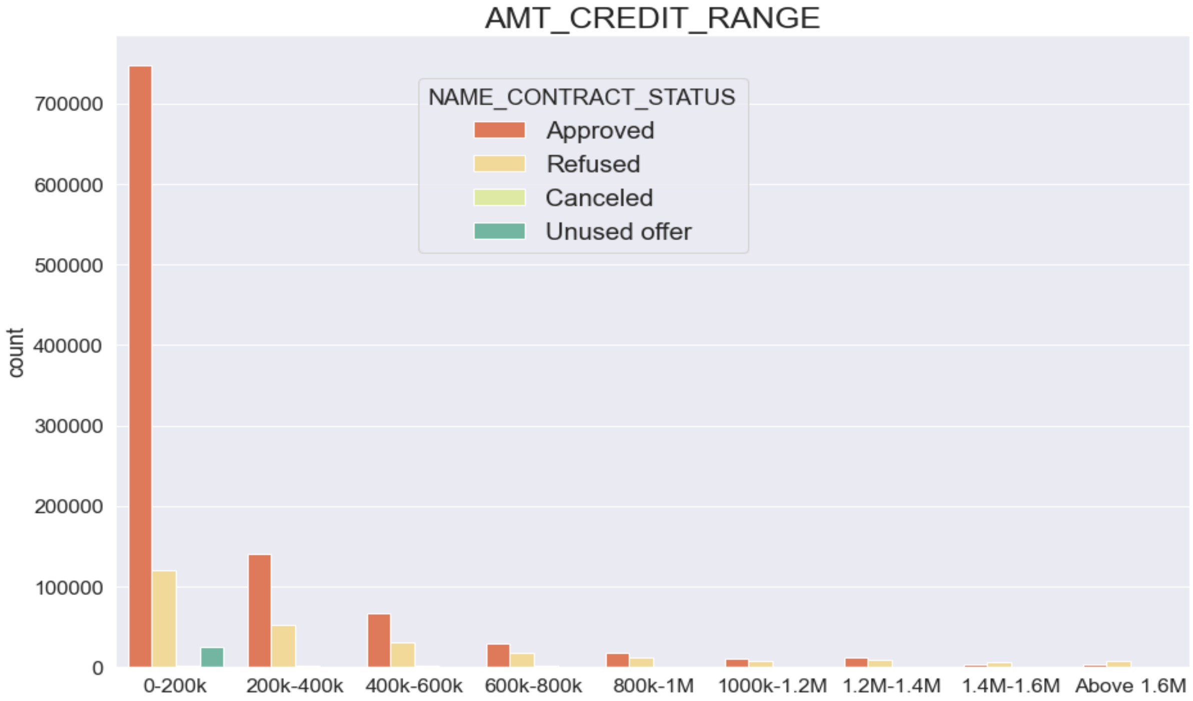
Insights

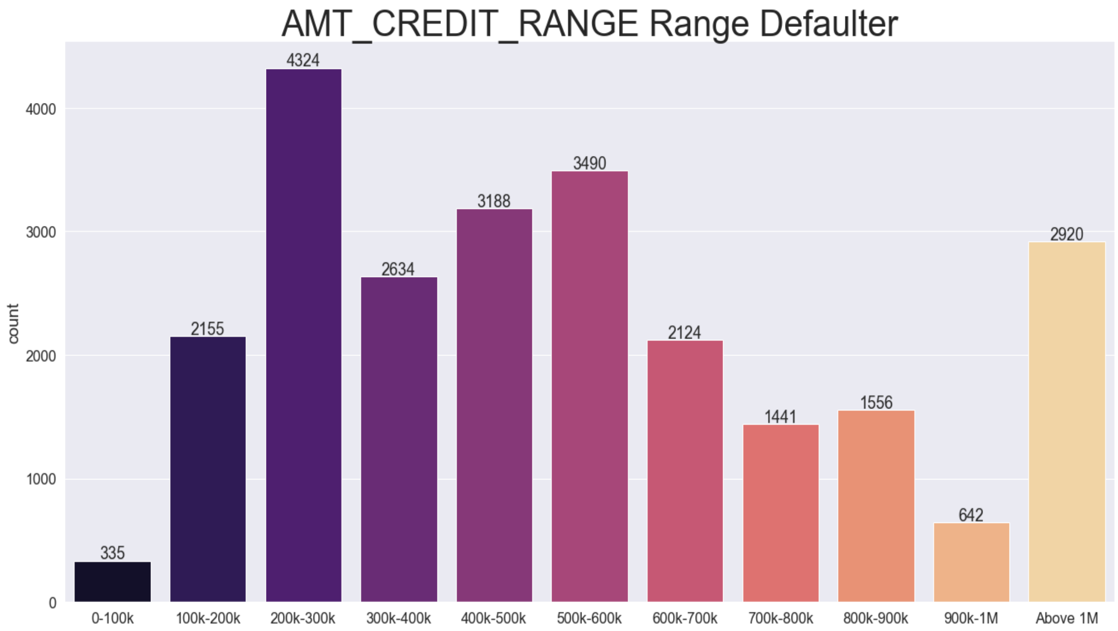
- Majority of applicants have AMT_CREDIT between 100k-600k
- The highest number of defaulters are in the AMT_CREDIT range ***200K-300K (>7%)*** but the highest percentage of defaulters are in the range ***500k-600k(>10%)***
- the AMT_CREDIT ranges with least percentage of defaulters are ***0-100k (less than 6%)*** and ***Above 1M(less than 6%)***

Inferences

- As the credit range increases the probability of loan getting rejected increases and loan getting approved decreases







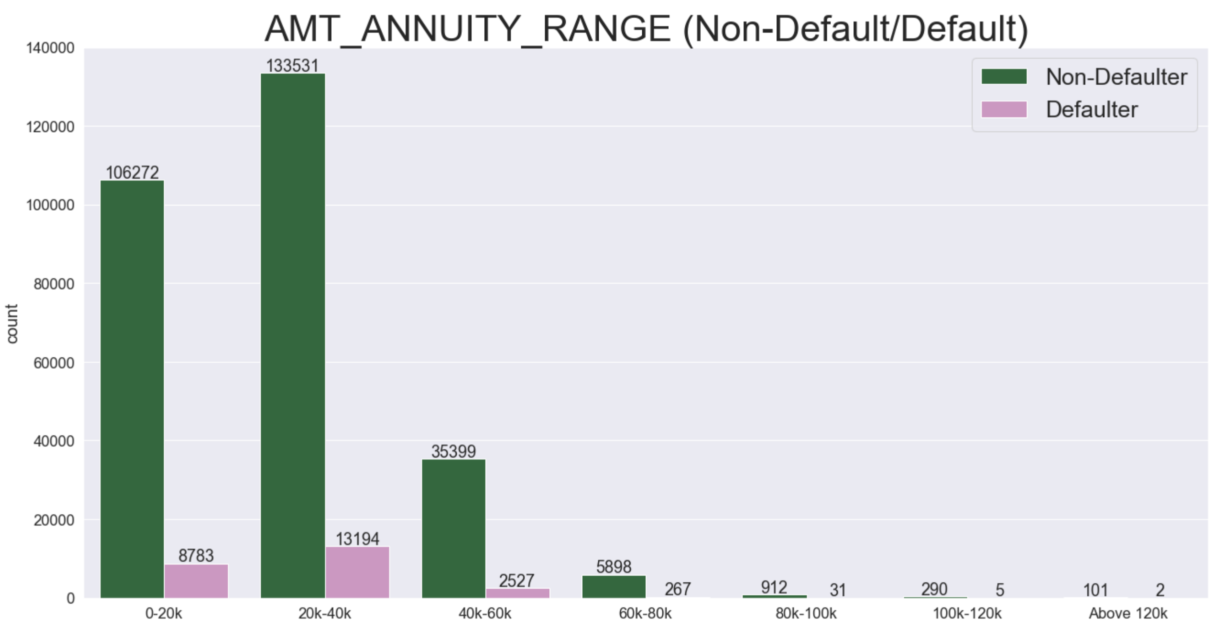
Univariate analysis on AMT_ANNUITY_RANGE

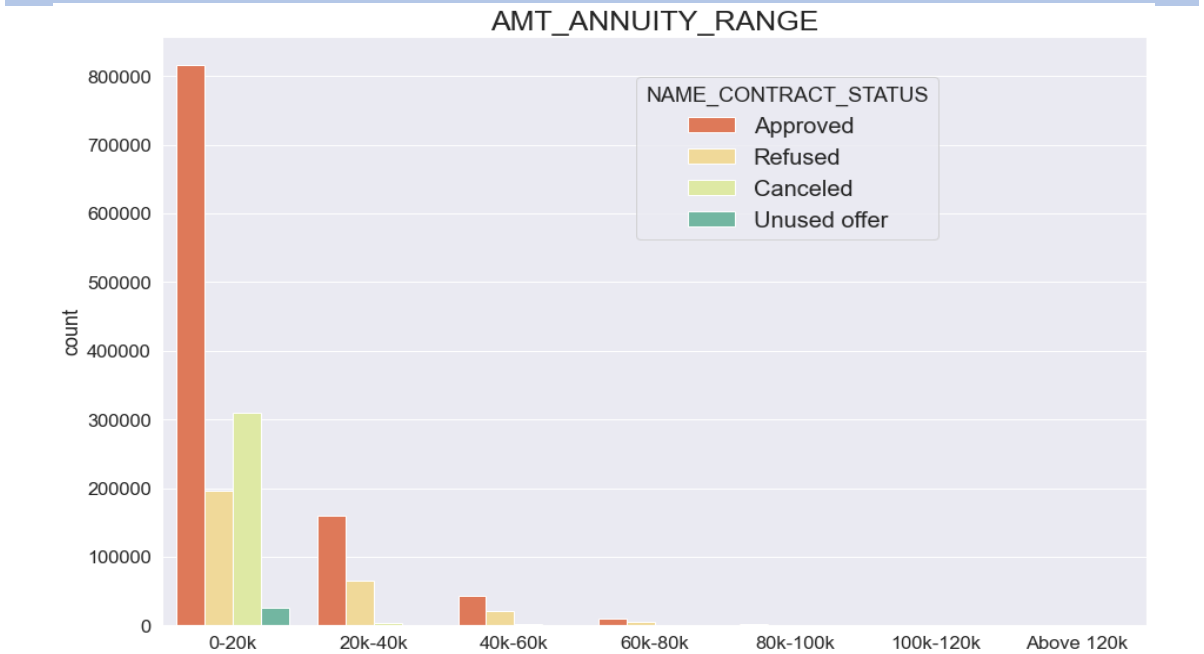
Insights

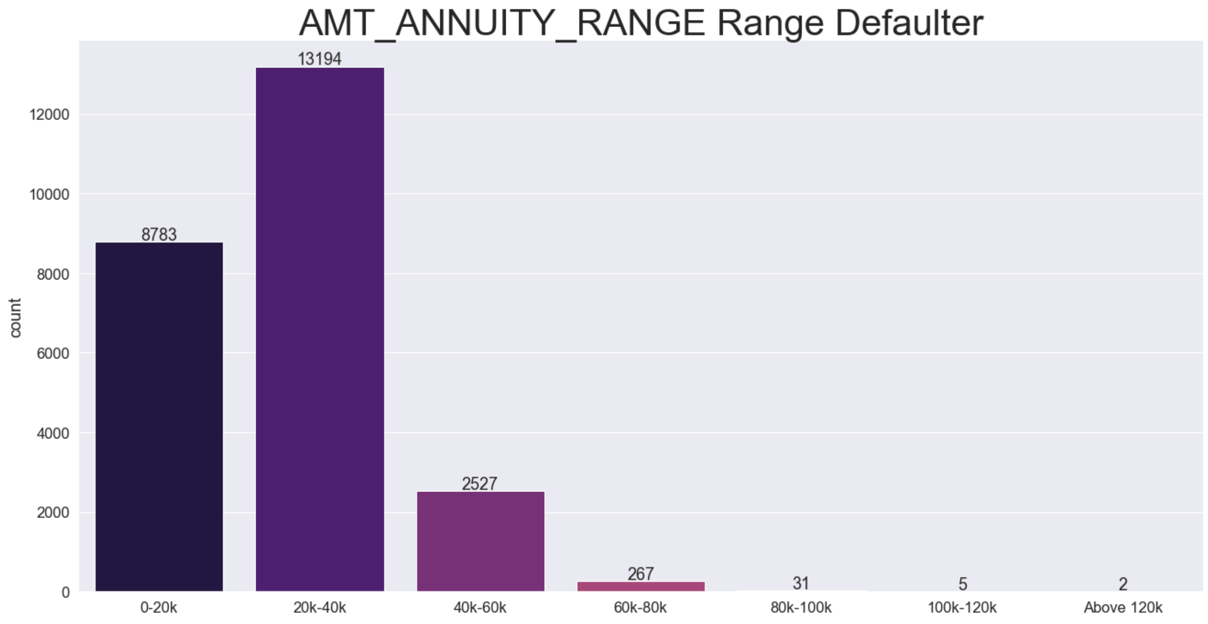
- Majority of applicants have the AMT_ANNUITY less than 60K
- The highest percentage of defaulters have AMT_ANNUITY between ***20k-40k (>7%)*** and the applicants having AMT_ANNUITY ***100k-120k (>1%)*** and ***Above 120k (>1%)***
- For AMT_ANNUITY between 0-20K, the percentage of loans refusal is lowest (less than 15%)
- Mid-range ANNUITY (between 40K-80K) has the highest loan refusal percentage
- 0-20k AMT_ANNUITY range also has the most percentage of loans getting cancelled

Inferences

- The percentage of defaulters decrease as the AMT_ANNUITY increase with the exception being the range ***20k-40k*** where there is an increase in defaulting applicants
- As the AMT_ANNUITY increases the probability of loans getting approved increase







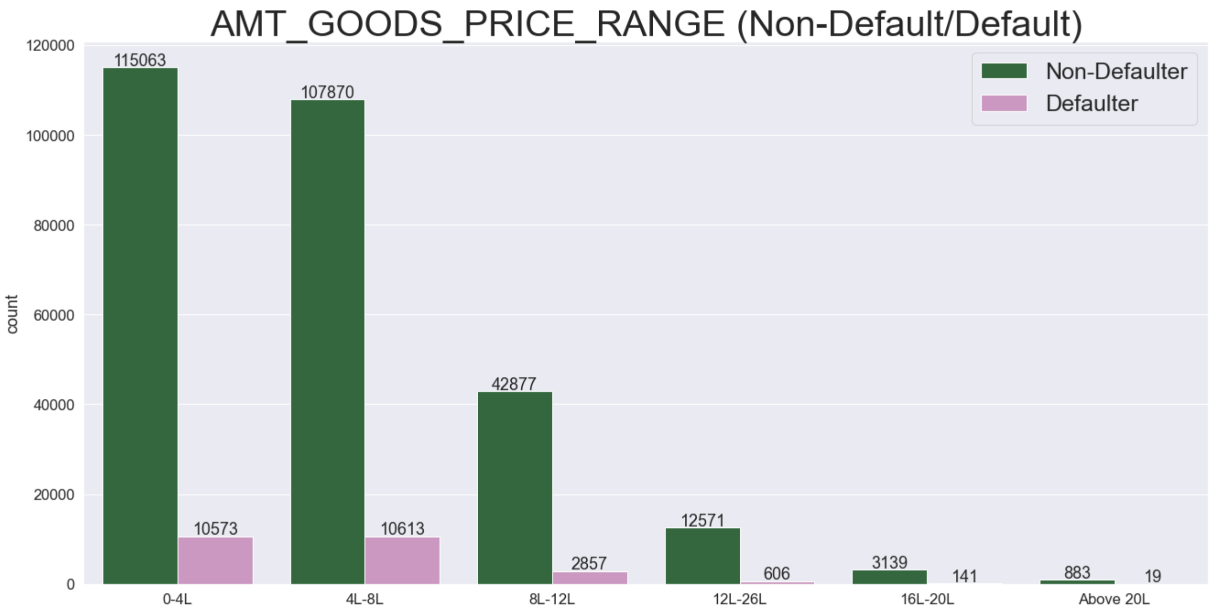
Univariate analysis on AMT_GOODS_PRICE_RANGE

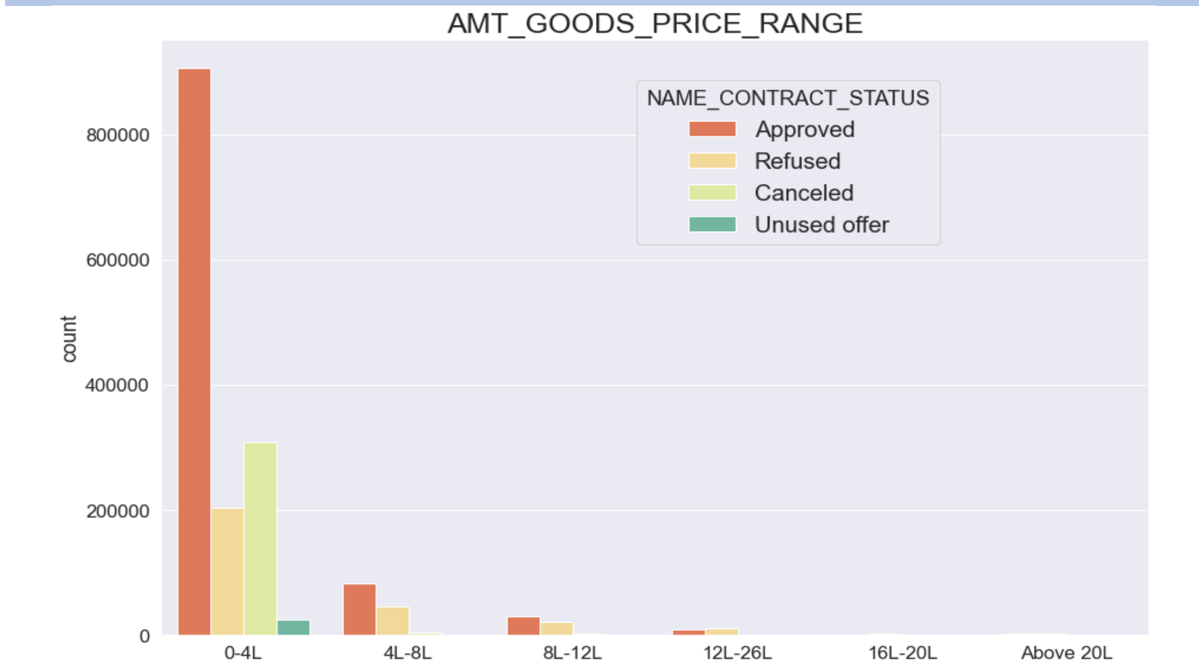
Insights

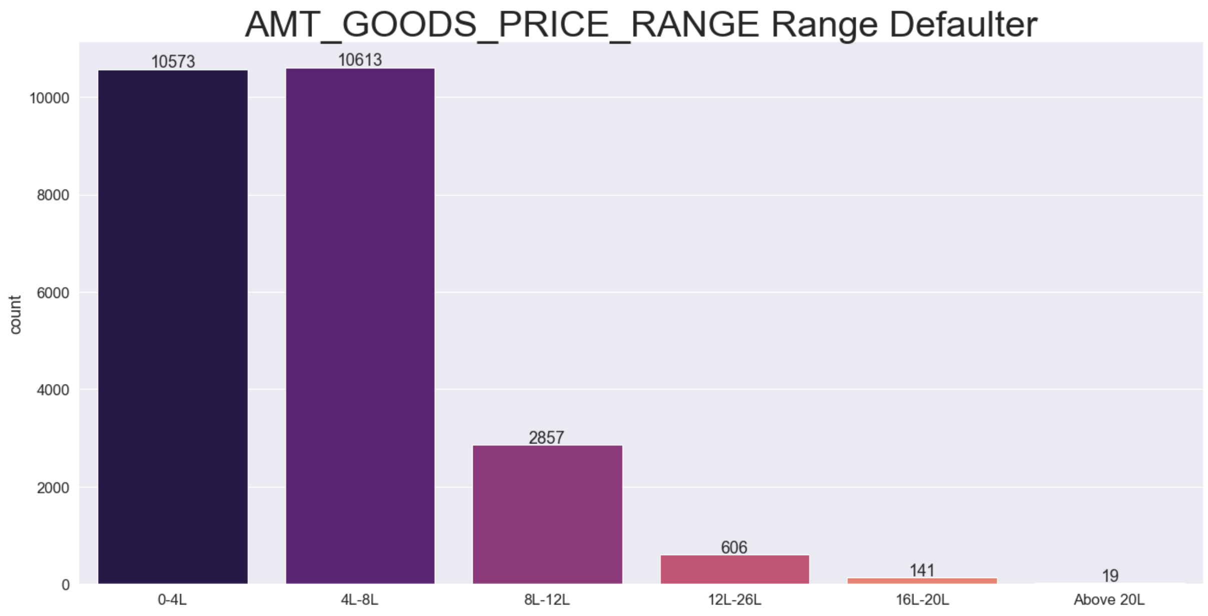
- Majority of applicants have the AMT_GOODS_PRICE less than 8L
- The highest percentage of defaulters have AMT_GOODS_PRICE between ***4L-8L (~9%)*** followed by ***0-4L (>8%)*** , the AMT_GOODS_PRICE range having least percentage of defaulters is Above ***20L (2%)***

Inferences

- As price of goods increase percentage of defaulters decrease
- As the price of goods increases the probability of loan getting rejected increases and loan getting approved decreases



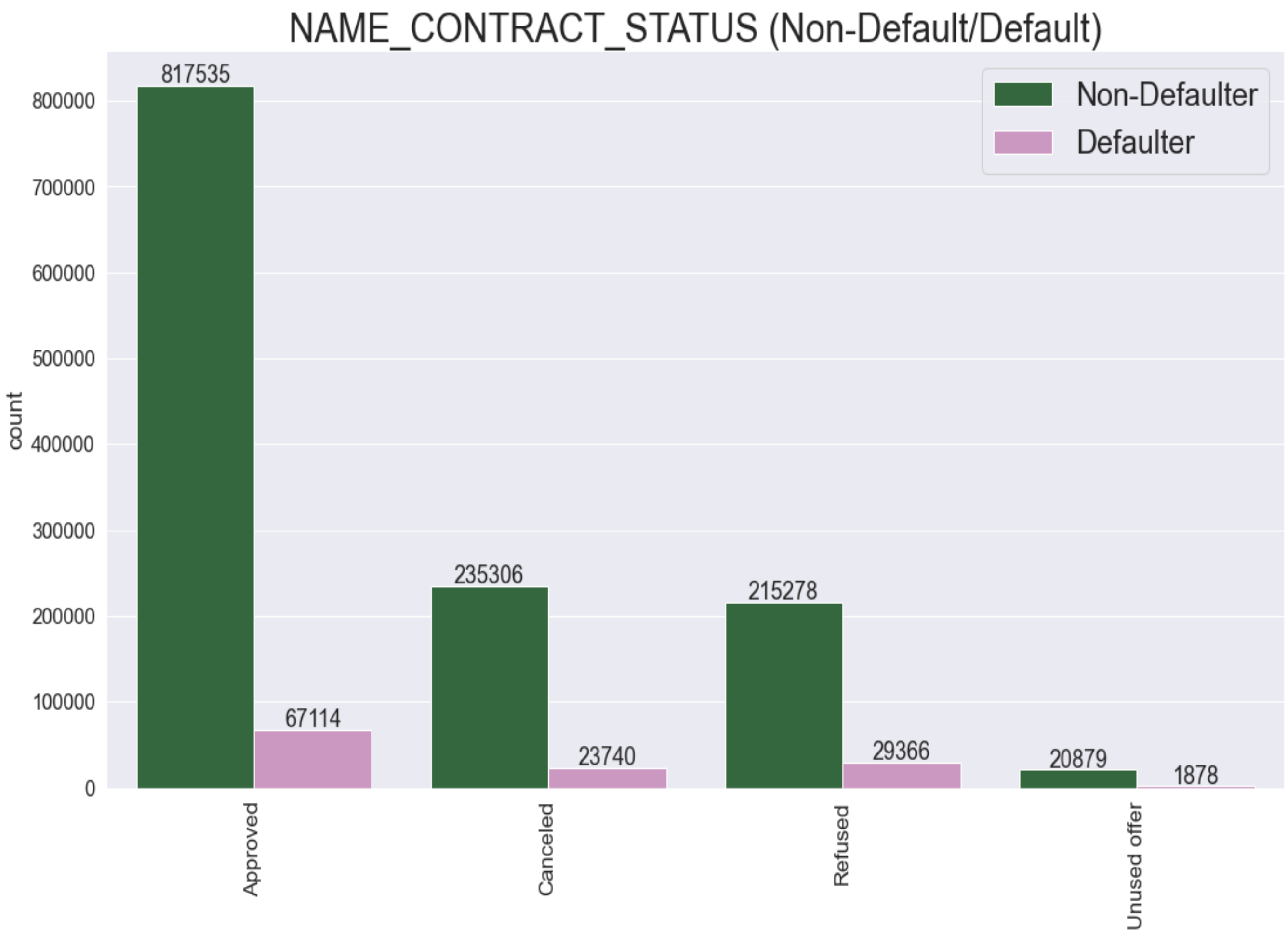




Univariate analysis on NAME_CONTRACT_STATUS

Insights

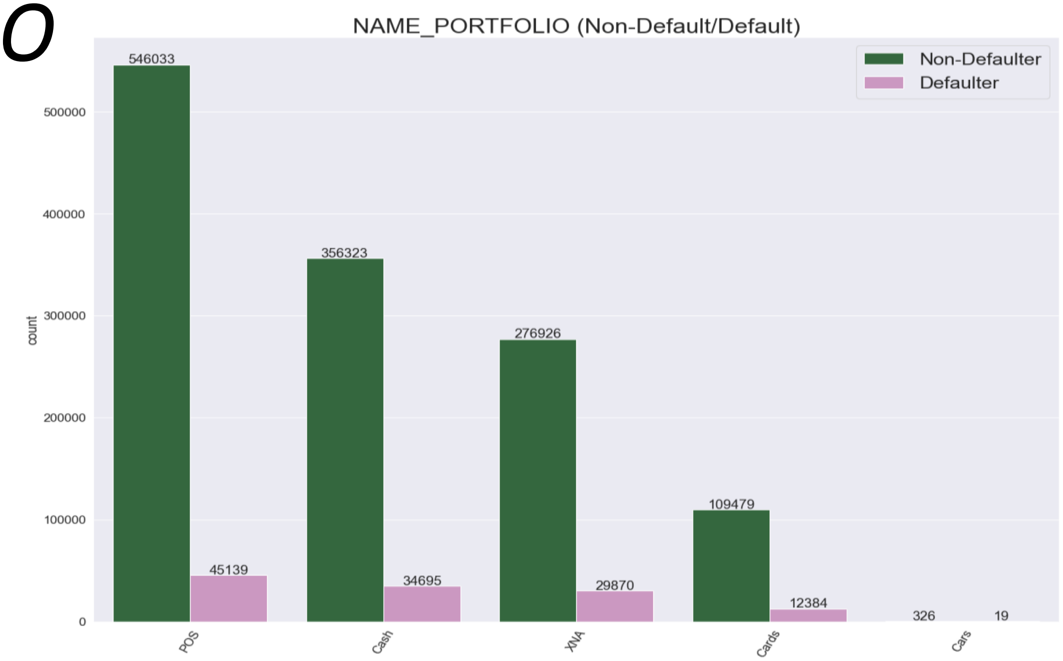
- About 90% of cancelled applications are of non-defaulting applicants
- For defaulting applicants about 8% of total applicants who defaulted in previous loans have the loan status as approved
- While about 88% of the total refused loans are of non-defaulting applicants
- 90% of the cancelled applicants belong to non-defaulting clients, so reducing the interest rates could result in more non-defaulting applicants not cancelling the loan
- 88% of the total refused applicants have actually repaid the loan without defaulting

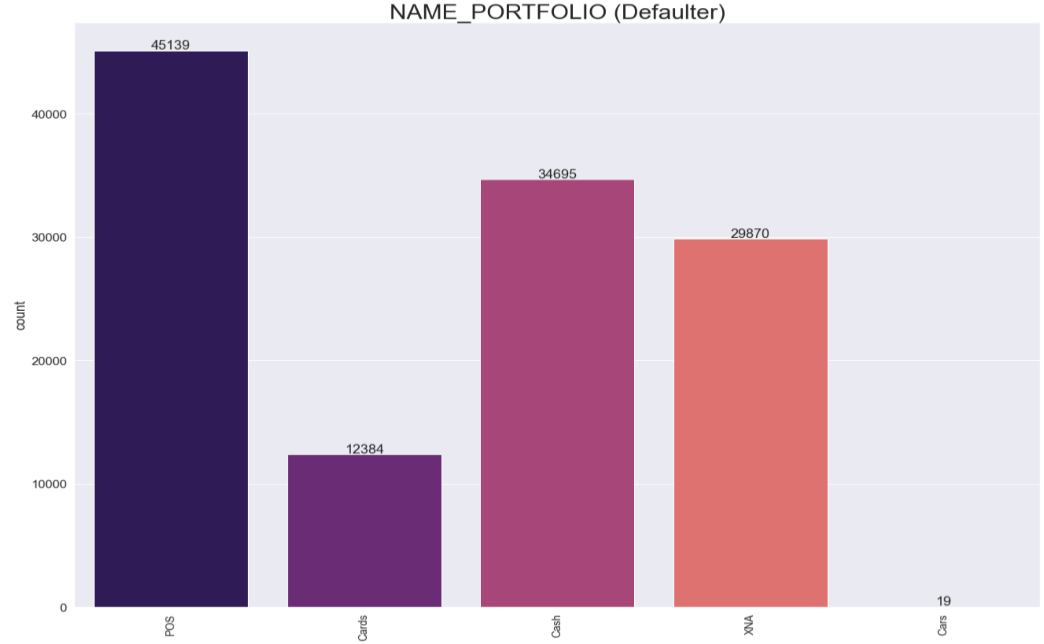


Univariate analysis on NAME_PORTFOLIO

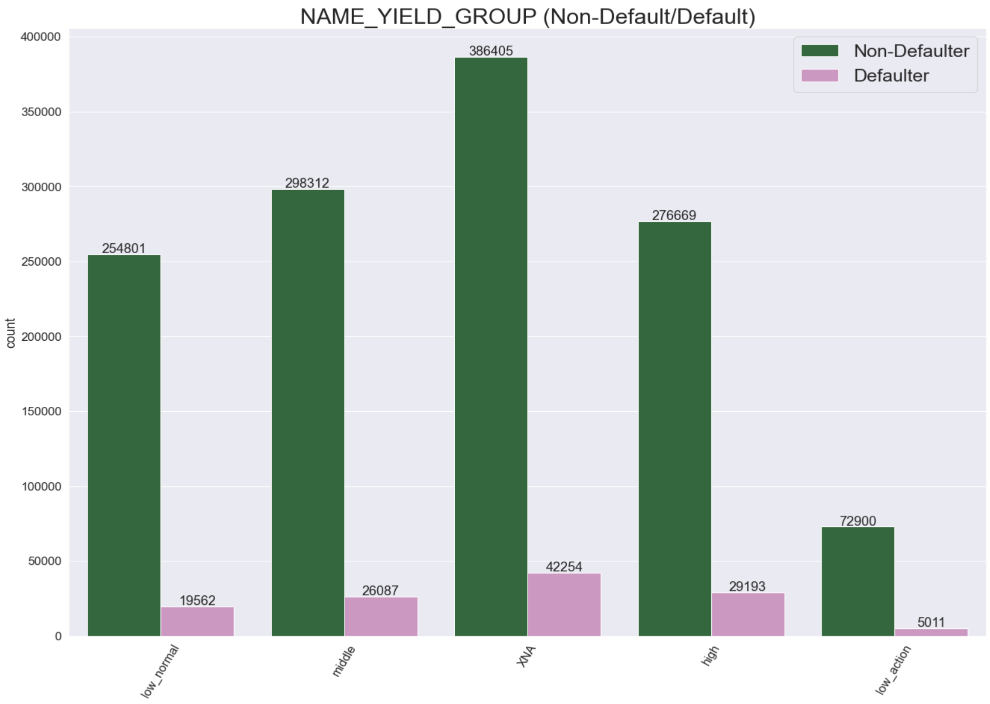
Insights

- There are some mis-spelled entries ***Cars*** which are actually ***Cards*** but, because the number of mis-spelled entries is very very low we can leave them as it is.
- In NAME_PORTFOLIO ***Cards*** have the highest percentage of defaulters
- Also there are significant amount of unknown values



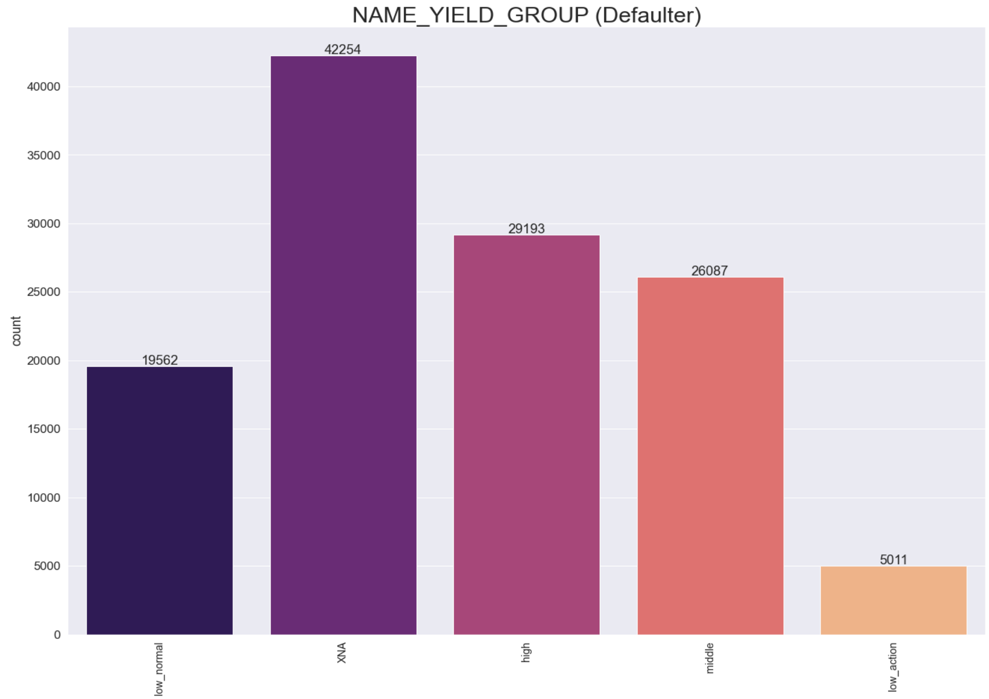


Univariate analysis on NAME_YIELD_GROUP



Insights

- Highest percentage of defaulters are in NAME_YIELD_GROUP ***High***(~9%) and the lowest in ***low_action***(<7%)
- There are significant amount of unknown values in NAME_YIELD_GROUP



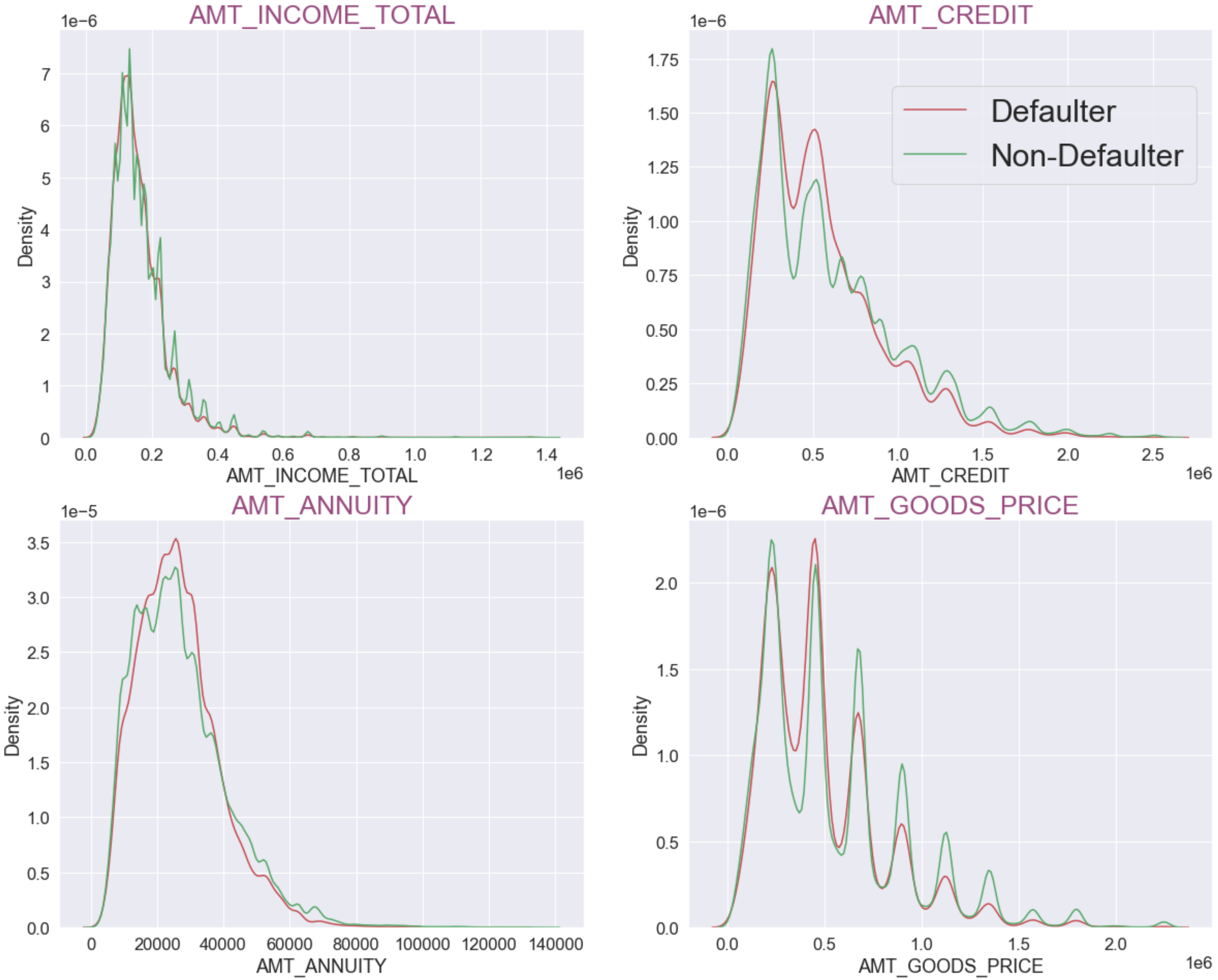
Inferences

- For application having high interest rates the probability of applicants defaulting is also higher while application having low_action interest rates have lower probability of applicants defaulting
- A non-defaulting applicants is more likely to default if interest rates are higher

Segmented Numeric analysis on AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE

Insights

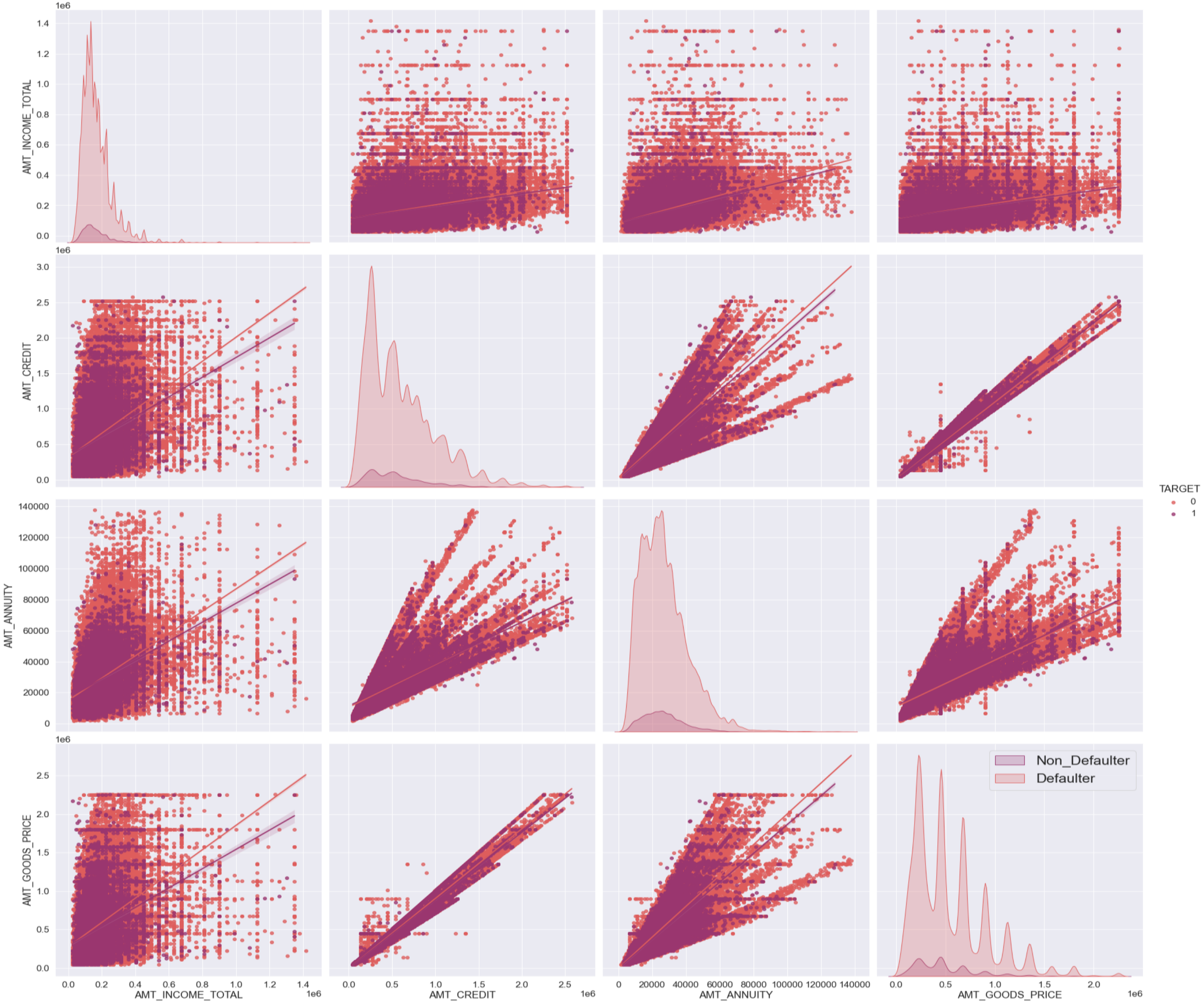
- As in all the plots the distribution of both Defaulters and Non-Defaulters overlap so we cannot infer much from this
- Most applicants pay annuity amount less than 60,000
- Most of credit amount is below 10,00,000
- Majority of applicants have total income below 4,00,000
- For the goods having price between the range 250K to 600K frequency of defaulters is slightly more than that of non-defaulters



Segmented Bivariate Numeric analysis on AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE

Insights

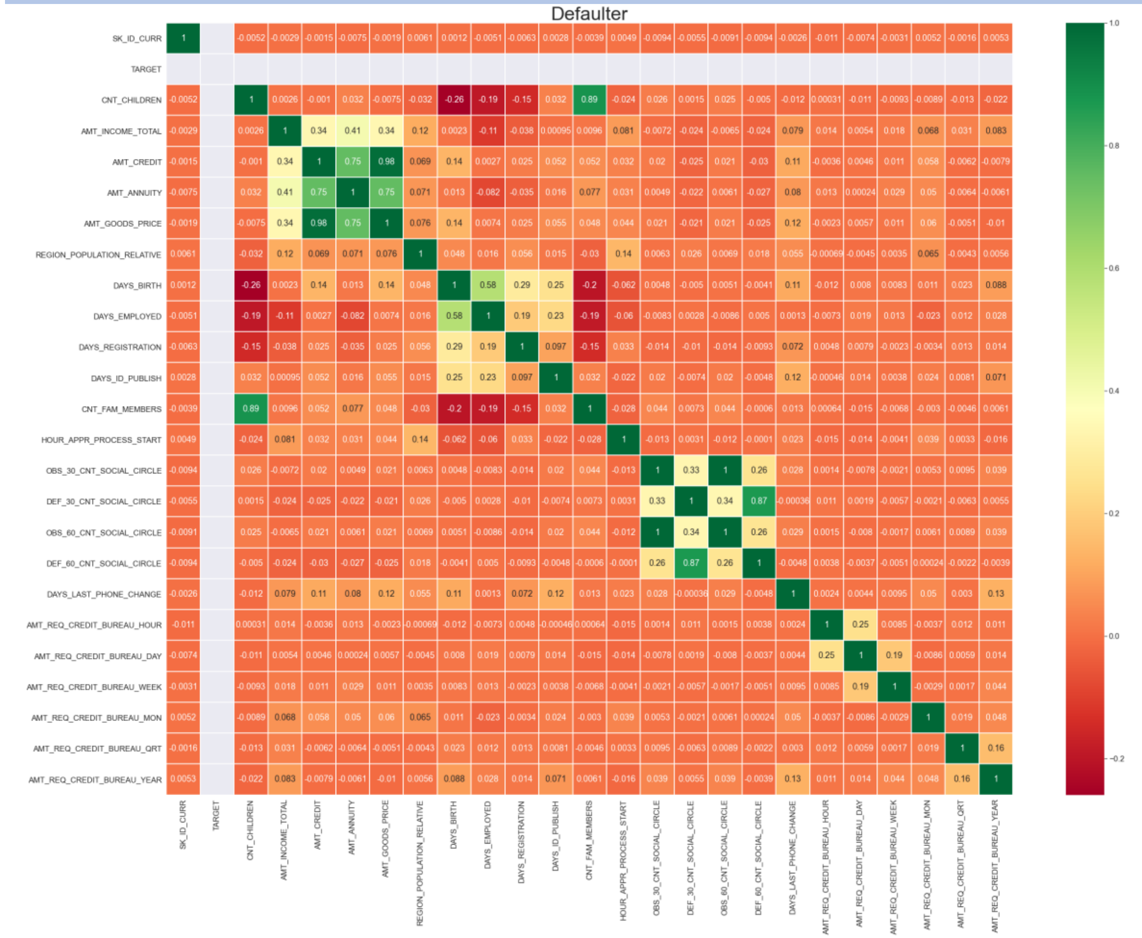
- AMT_ANNUITY and AMT_CREDIT have a very high positive correlation while AMT_CREDIT and AMT__GOODS_PRICE also have high positive correlation
- AMT_INCOME_TOTAL has a moderate positive correlation with AMT_CREDIT, AMT_GOODS_PRICE and AMT_ANNUITY

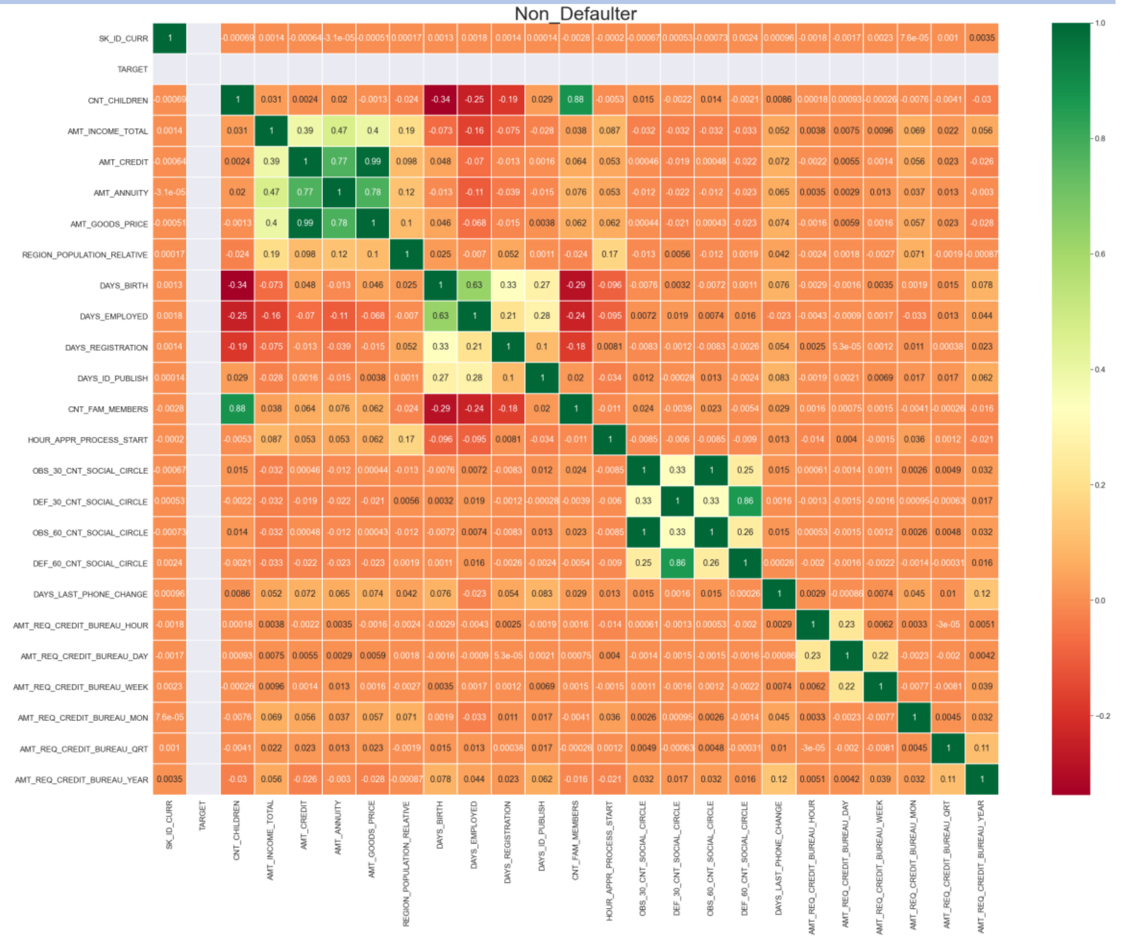


Plotting Segmented Heatmaps

Insights

- AMT_CREDIT, AMT_GOODS_PRICE and AMT_ANNUITY all have high correlation with each other, where AMT_CREDIT and AMT_GOODS_PRICE is very highly correlated(~0.99)
- Variables OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE have the highest correlation i.e. 1. DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE are also highly correlated(~0.86) which actually increases a bit from 0.86 to 0.87 for Defaulting applicants
- For Defaulters the AMT_ANNUITY and AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE correlating slightly decreases from 0.77 and 0.78 to 0.75 and 0.75.
- DAYS_BIRTH and DAYS_EMPLOYED have a moderately high correlation of 0.63 for Non-Defaulters which decreases to 0.58 for Defaulters
- There is a significant drop of correlation between AMT_INCOME_Total and AMT_CREDIT, AMT_INCOME_Total and AMT_ANNUITY, AMT_INCOME_Total and AMT_GOODS_PRICE from 0.39 to 0.34 for AMT_CREDIT, 0.47 to 0.41 for AMT_ANNUITY and from 0.4 to 0.34 for AMT_GOOD_PRICE





Conclusion

Attributes corresponding to applicants likely to default

- Applicants employed for less than 5 years
- Applicants with NAME_FAMILY_STATUS as civil marriage
- Applicants having highest level of education as Lower secondary
- Applicants living in ***Rented apartment*** and ***With parents*** have high probability of defaulting
- Applicants having OCCUPATION_TYPE as ***Low-skill Laborers***
- Applicants having generally less paying job like ***Drivers***, ***Security staff*** etc. are more likely to default
- Applicants with ORGANIZATION_TYPE as ***Transport: type 3***
- Young applicants between age less than 30 have more probability of defaulting
- Applicants with income less than 200k are more likely to default
- In NAME_PORTFOLIO ***Cards*** have the highest percentage of defaulters
- A non-defaulting applicants is more likely to default if interest rates are higher

Attributes corresponding to applicants likely to not default

- Applicants who are employed for *more than 40 years*
- Applicants having NAME_INCOME_TYPE as ***Businessman*** and ***Student*** have no history of defaulting and are more safer for loan approval
- Applicants having NAME_EDUCATION_TYPE ***Academic degree***
- applicants living in ***Office apartments*** have lower probability of defaulting
- Applicants residing in regions with rating **1** has the least probability of defaulting
- Applicants above age 60 have least probability of defaulting
- application having low_action interest rates have lower probability of applicants defaulting
- Applicants in higher income group, which are employed for more than 35-40 years are the safest for loan approval

Suggestions

- Recording the Purpose of loans can help in a better analysis of defaulting/non-defaulting clients
- If an applicant has no history of defaulting, then offering a middle or low interest rate would result in less probability of that applicants defaulting
- For applicants having less paying jobs, If loan is approved applying higher interest rate can minimize loss if client defaulted
- For applicants with CNT_CHILDREN greater than 6 have higher probability of defaulting so higher interest can be applied to minimize loss