CREDIT EDA ANALYSIS

By
Tuhin Mondal
(DS C40)
29.03.2022

Problem Statement

- It's a challenge for the financial institutes to provide loans to the people due to their insufficient credit data. The people who can not repay the loan they were approved are called the **"Defaulters"**. EDA is used to analyze the patterns present in the user 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.**

There are 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, hence called the **"Defaulters"**
- **All other cases:** All other cases when the payment is paid on time.

BUSINESS CONTEXT

- This analysis is 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 (to 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.

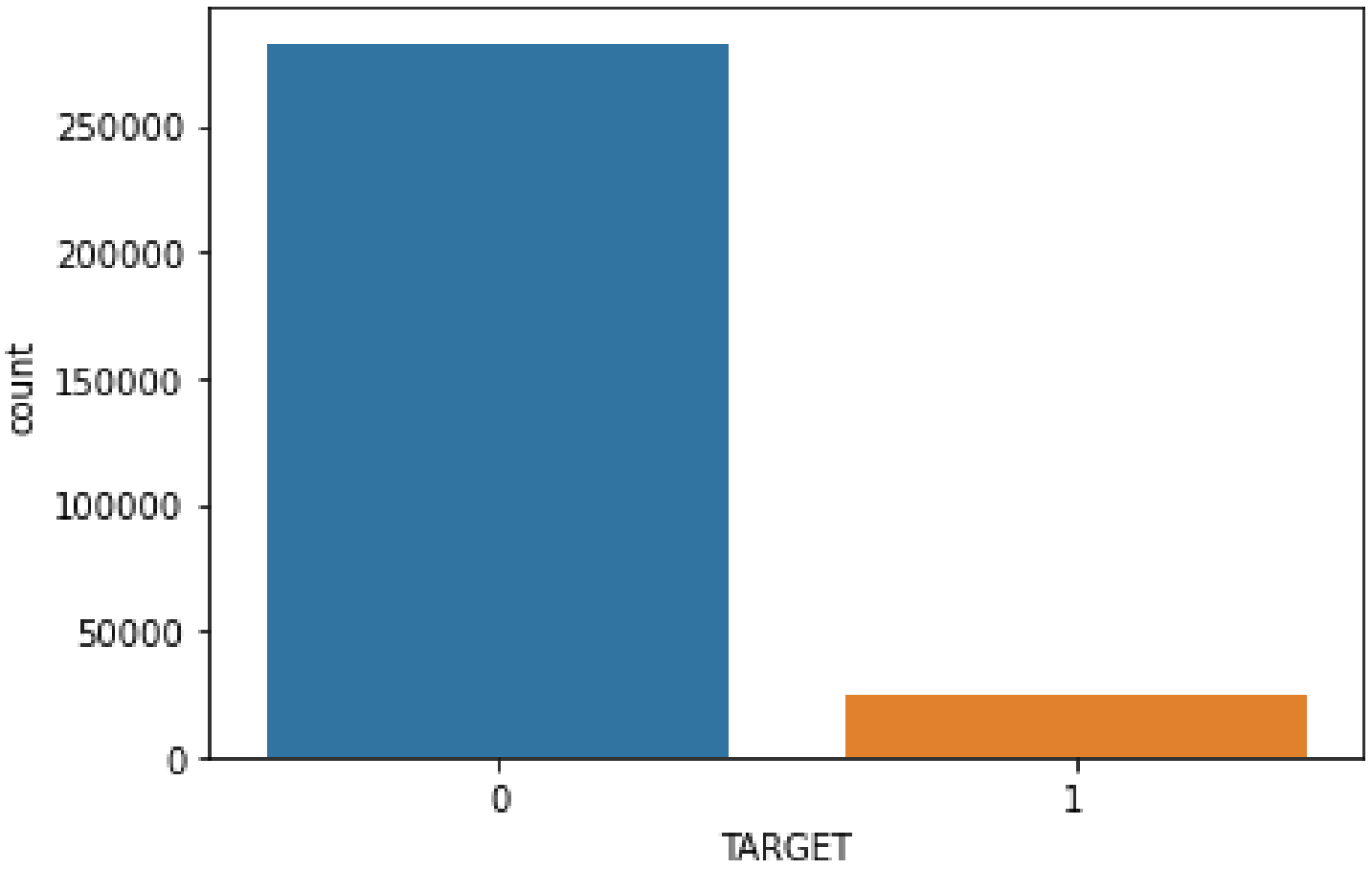
- So the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment. Customer demographic plays the key role here.

ANALYSIS APPROACH

The below steps have been taken in the whole EDA Analysis process :

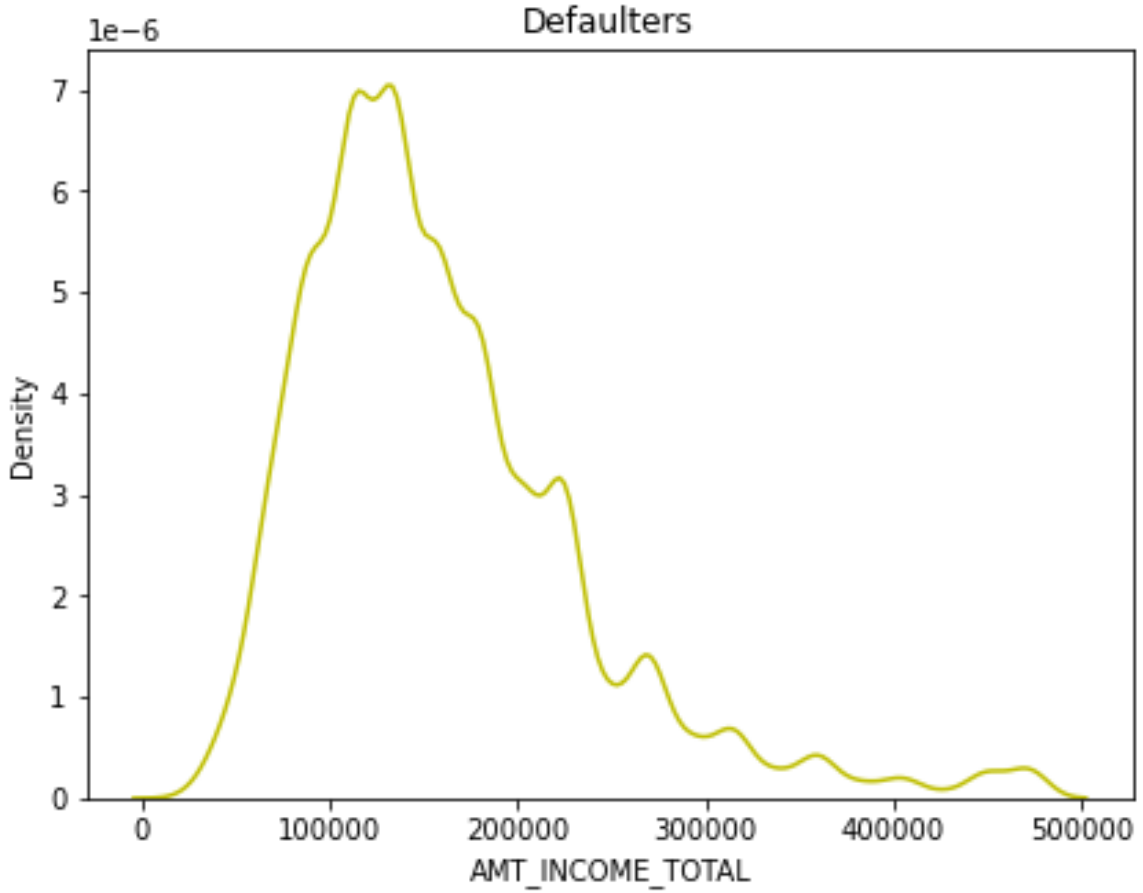
- First of all , the “application_data.csv” file was imported.
- The Missing values percentage was checked and variables with more than 45% missing data were dropped.
- The rest of the missing values were treated by imputing appropriate values.
- Outliers were handled with the proper methods.
- Variables were binned into categories.
- Imbalance percentage was checked.
- The data were partitioned into two data frames as Defaulters (Target=1) and non defaulters (Target=0).
- Univariate and Bivariate analysis were performed for continuous and categorical variables.
- The “previous_application.csv” was imported after that.
- The “defaulter” and “non defaulter” data were merged with the “previous_application.csv” file and the univariate and bivariate analysis were done for different status of the previously applied loan data.

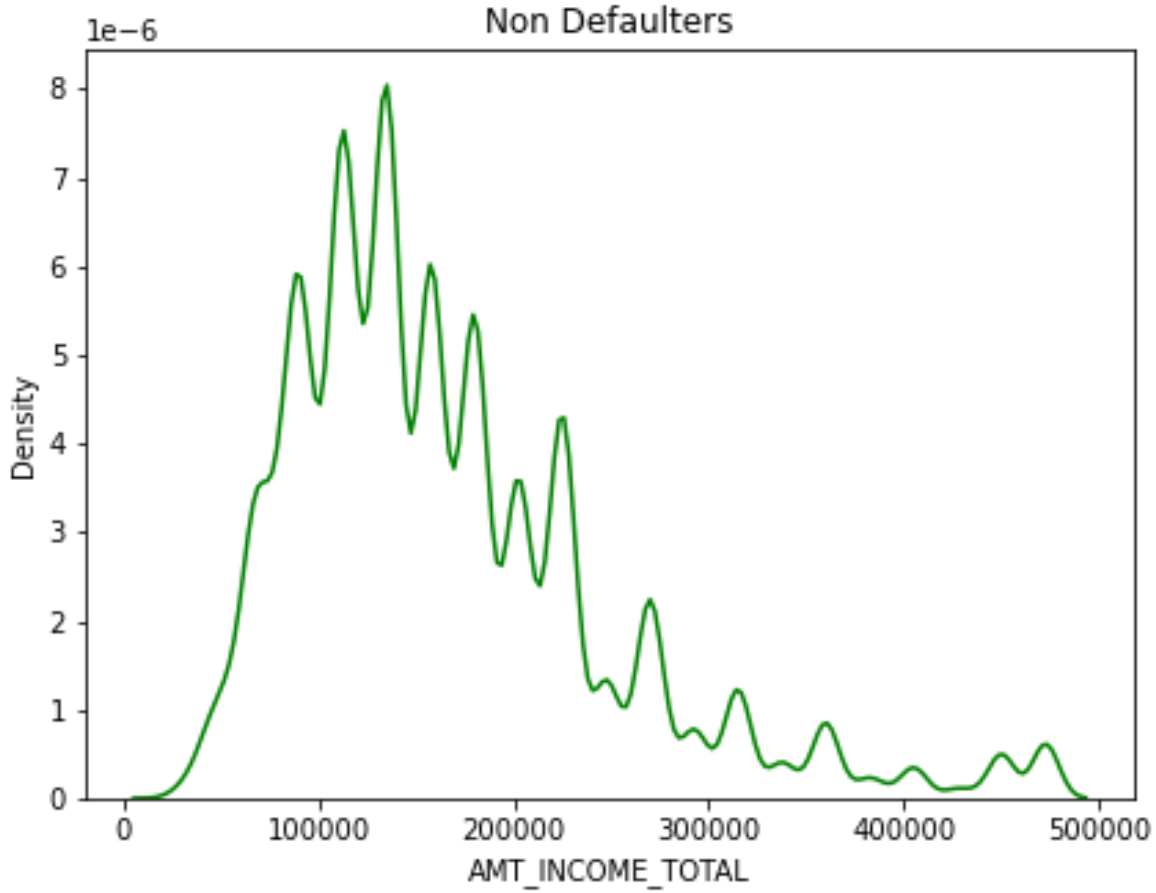
Data Imbalance



In the given data, the percentage of “Target=1” or default and ”target=0” or non default are 8% and 92% respectively, which implies data is heavily imbalanced in term of the target variable.

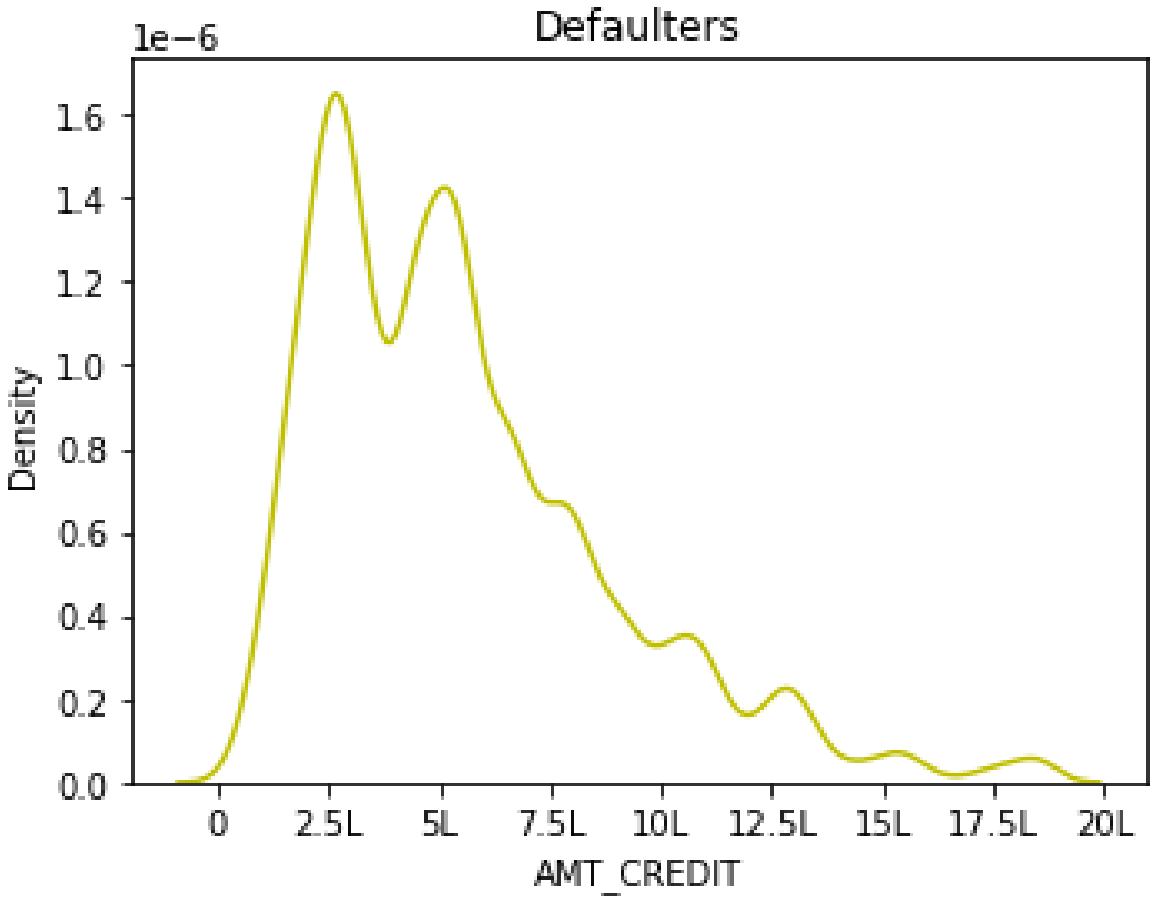
Customer Profiling : Univariate Analysis

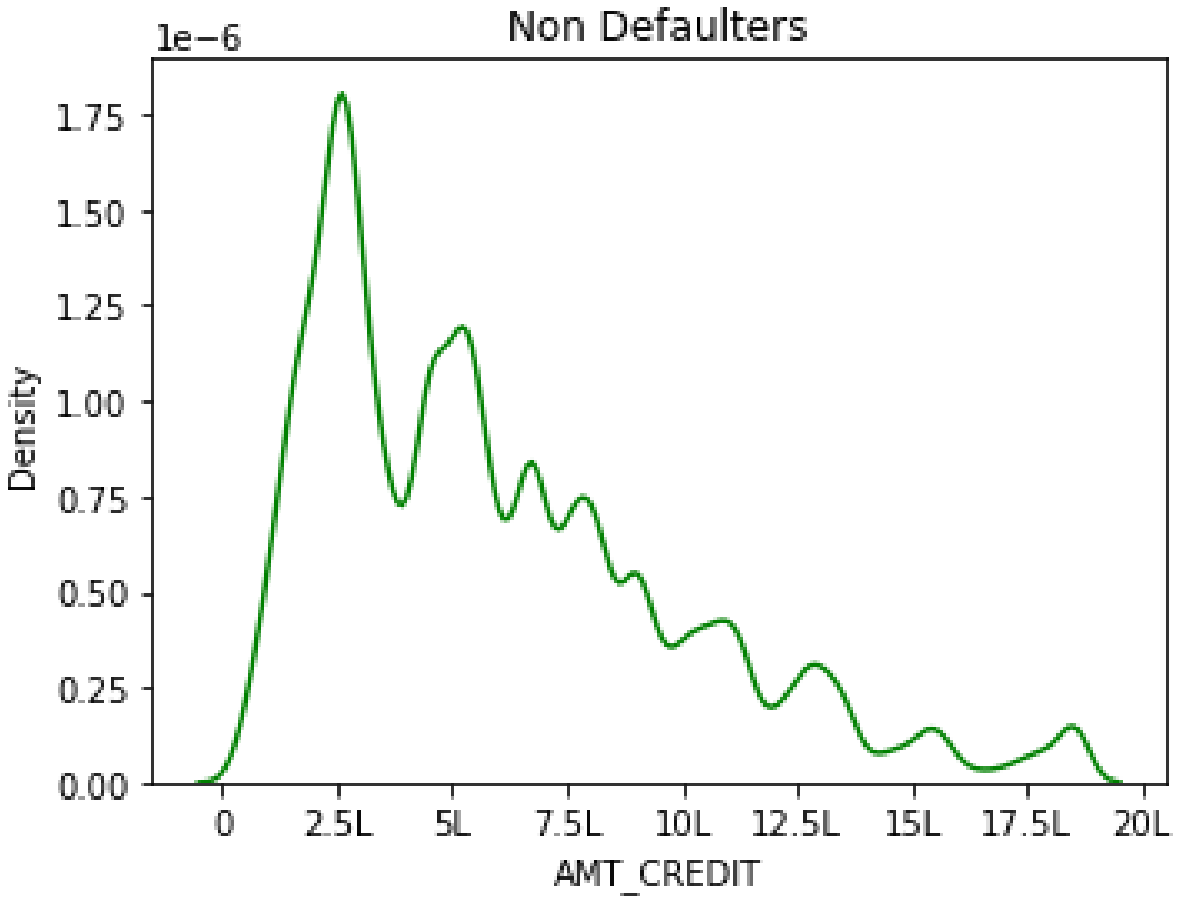




AMT_INCOME_TOTAL - Analysis : It can be seen that most of people to default have an income range of 10 to 15 lakh. Also people with higher income amount are less likely to default.

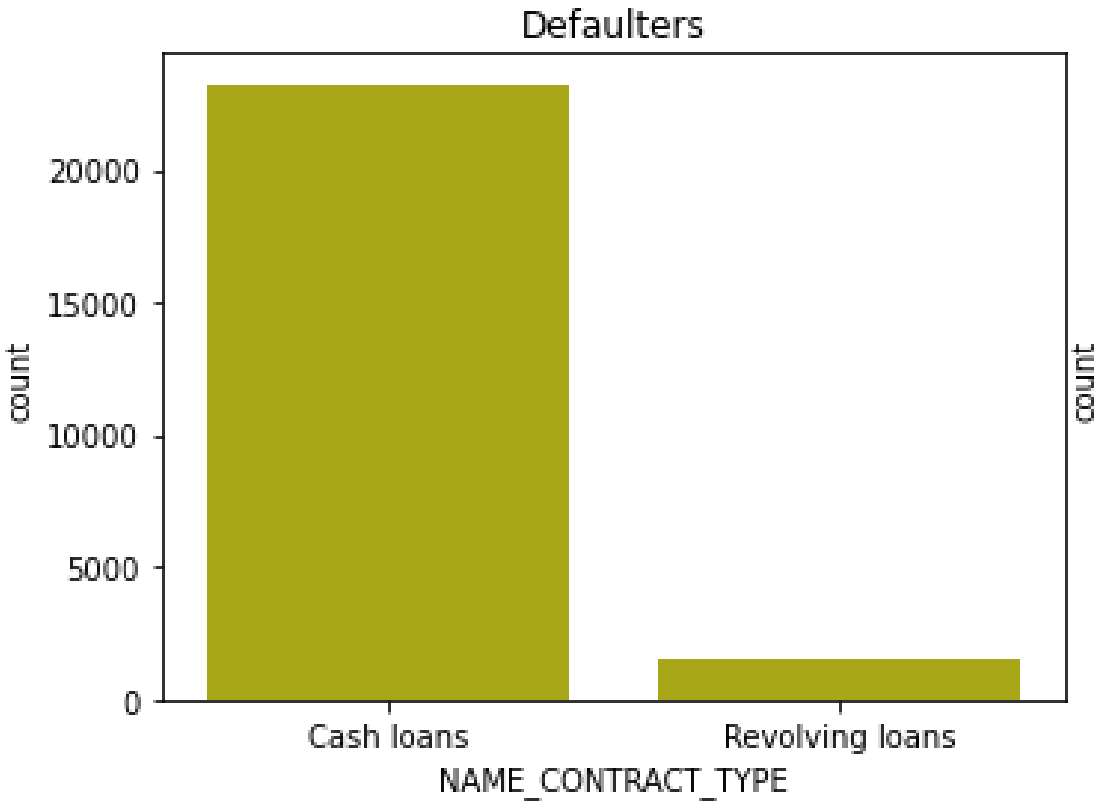
Customer Profiling : Univariate Analysis

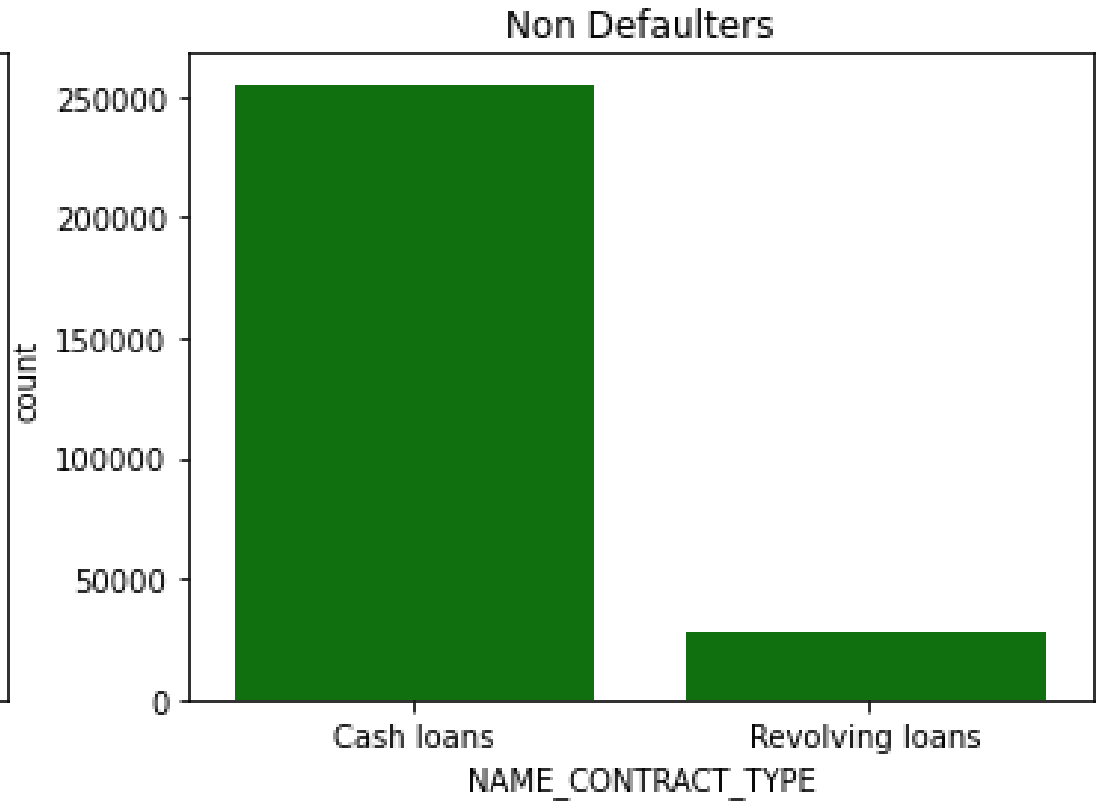




"AMT_CREDIT“ : it is visible that the people taking a loan amount of 10lakhs or less are more likely to make default.

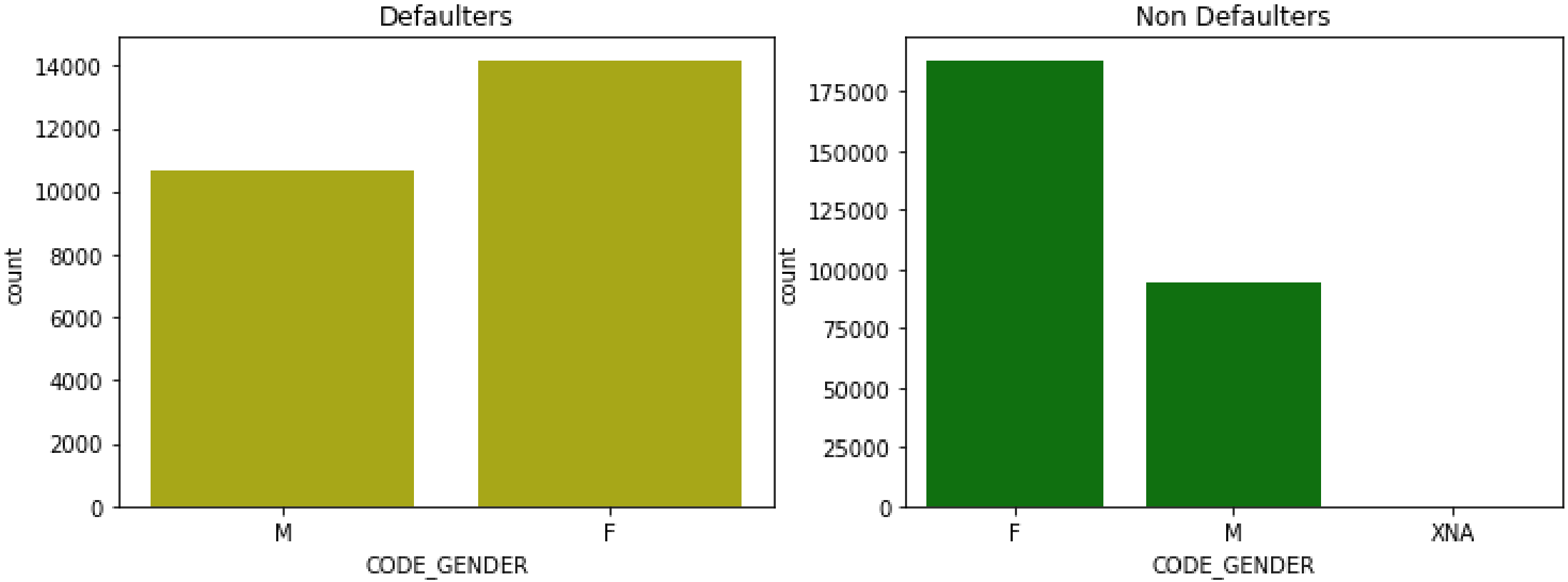
Customer Profiling : Univariate Analysis





“NAME_CONTRACT_TYPE“ : Ratio of revolving loans to cash loans is lower in the case of people who tend to default

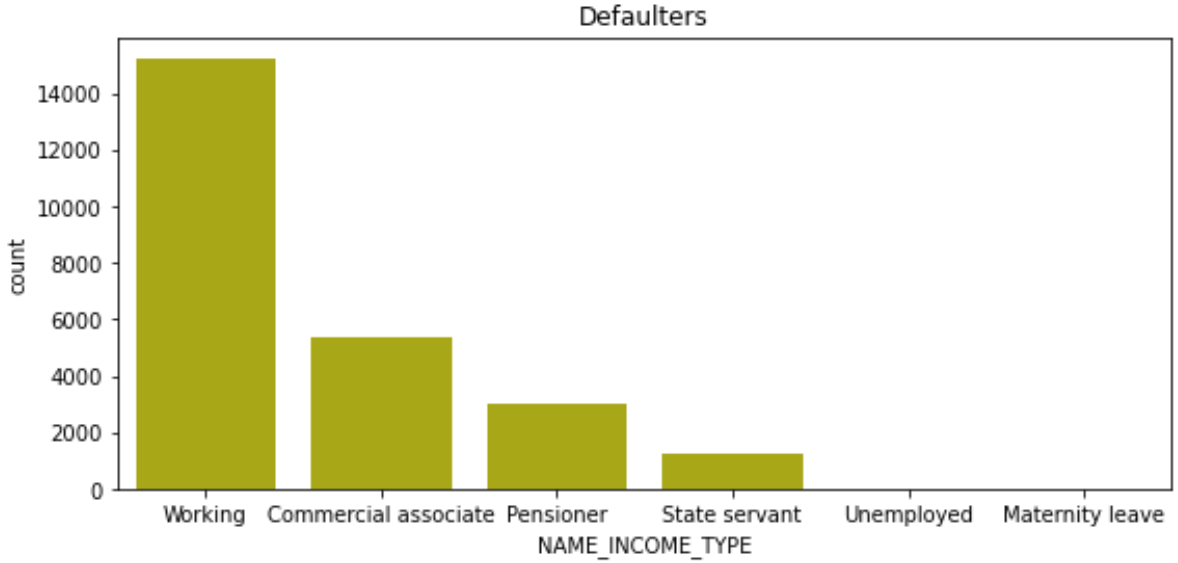
Customer Profiling : Univariate Analysis

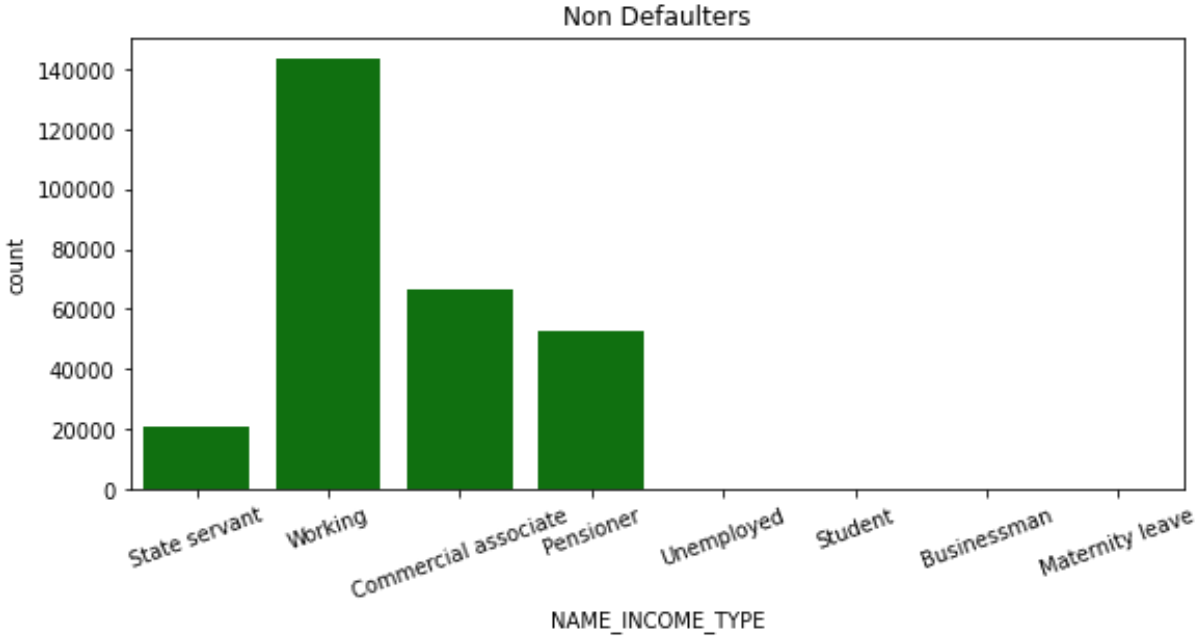


- CODE_GENDER - Analysis :

The plots show that it is the Females who make more defaulters

Customer Profiling : Univariate Analysis

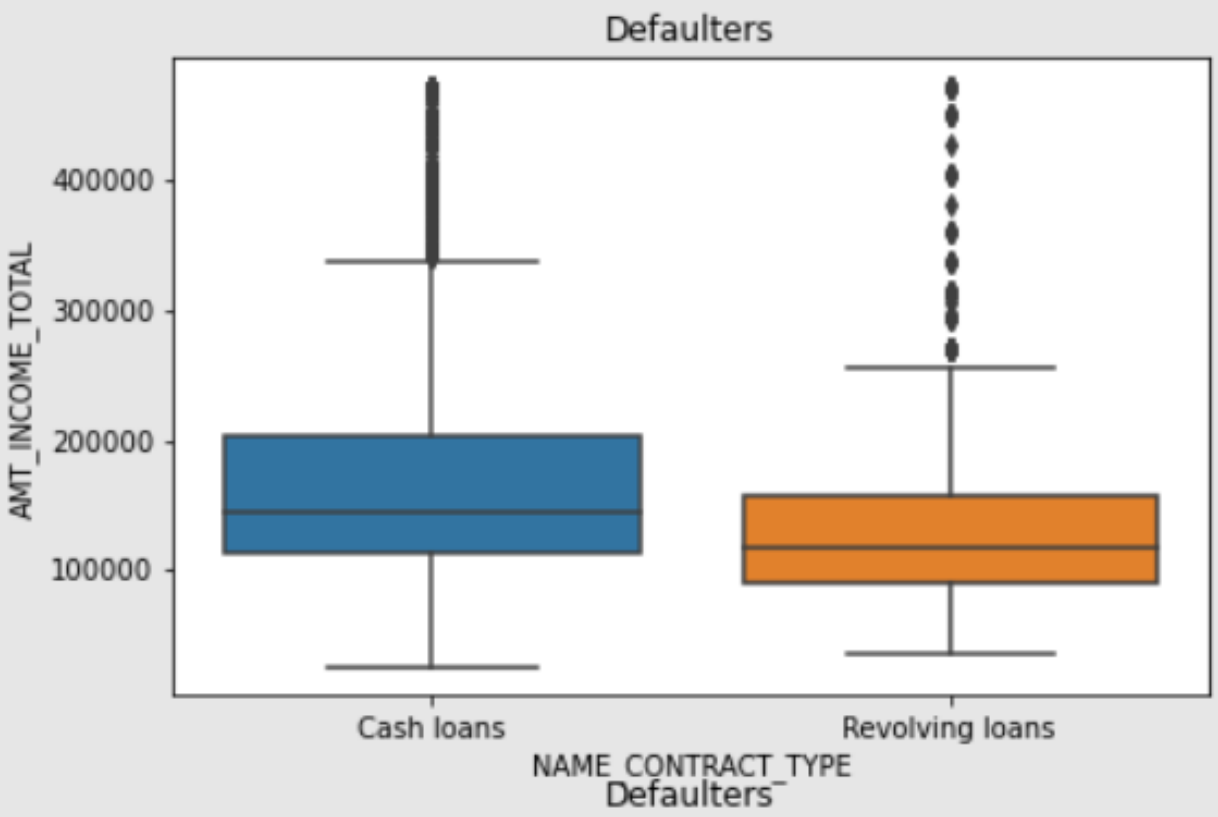


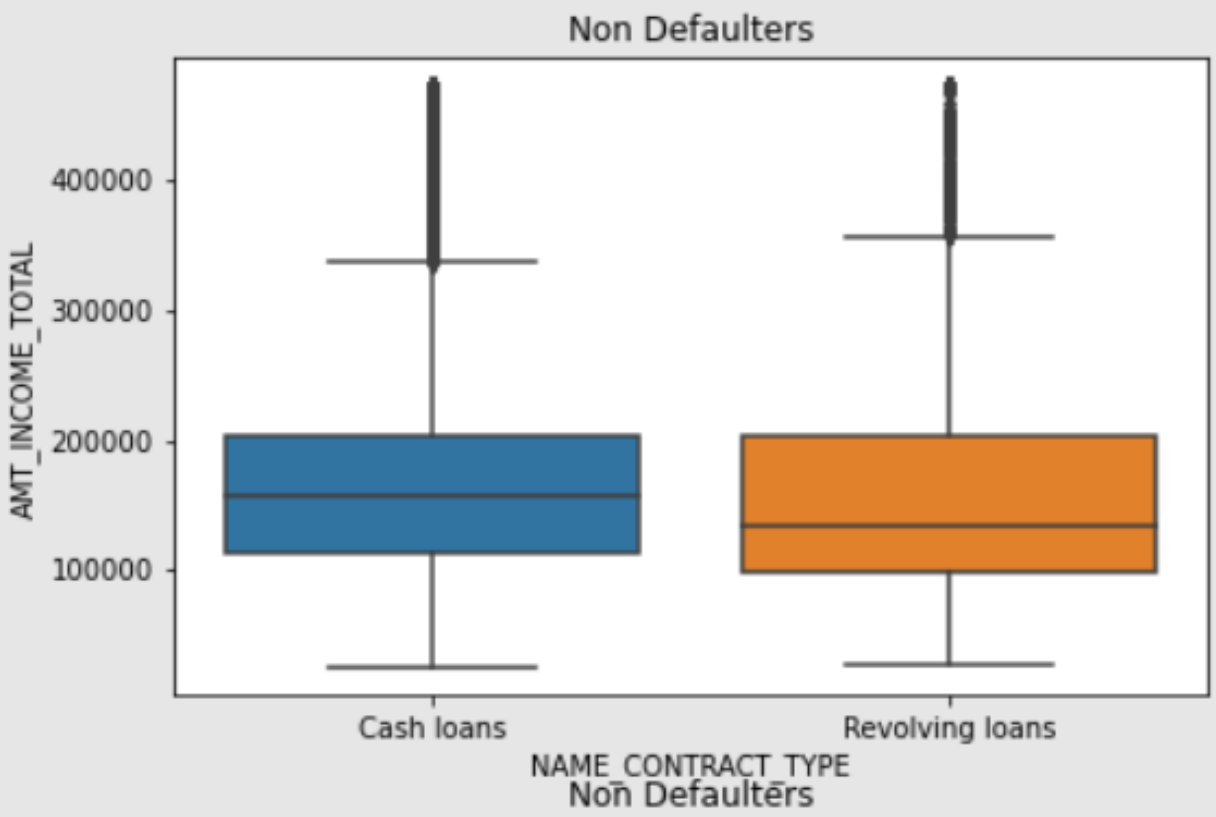


NAME_INCOME_TYPE - Analysis :

Plots above show that number of people in both Defaulter and non defaulter cases are same in the working profession.
Less people to work in commercial associate, pensioner and State servant are likely to default.

Customer Profiling : Bivariate Analysis

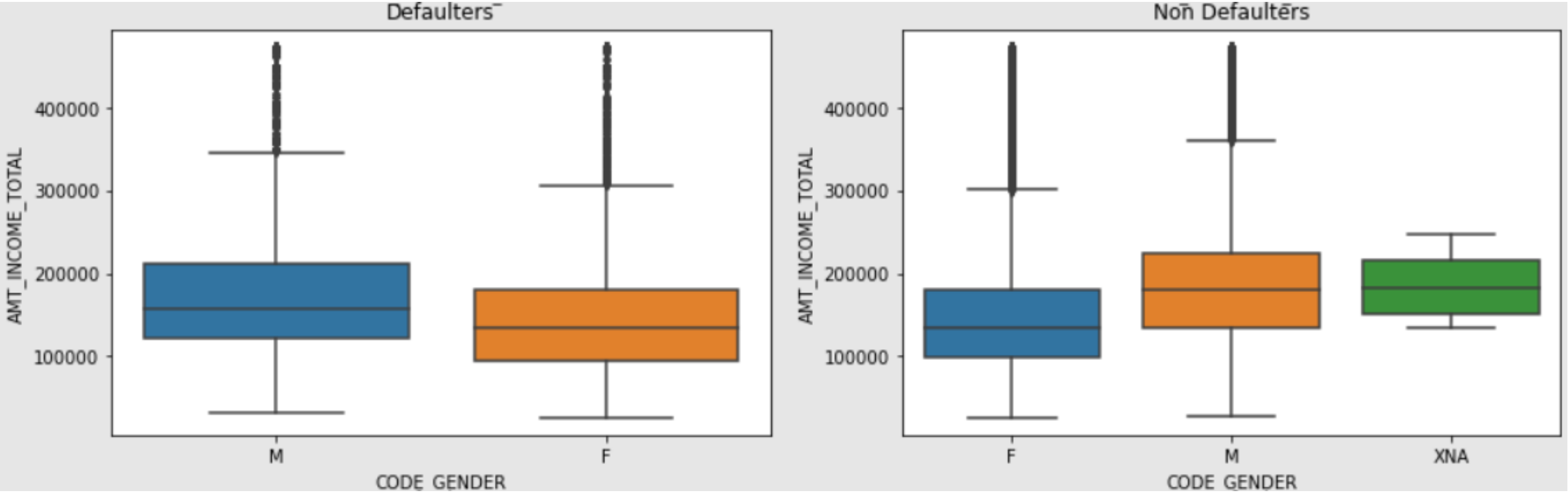




NAME_CONTRACT_TYPE :

In case of Non Defaulters, we can see people more people have salary more than the Median in Revolving loan. The distribution is rightly skewed, where as it is almost balanced in case of Defaulters.

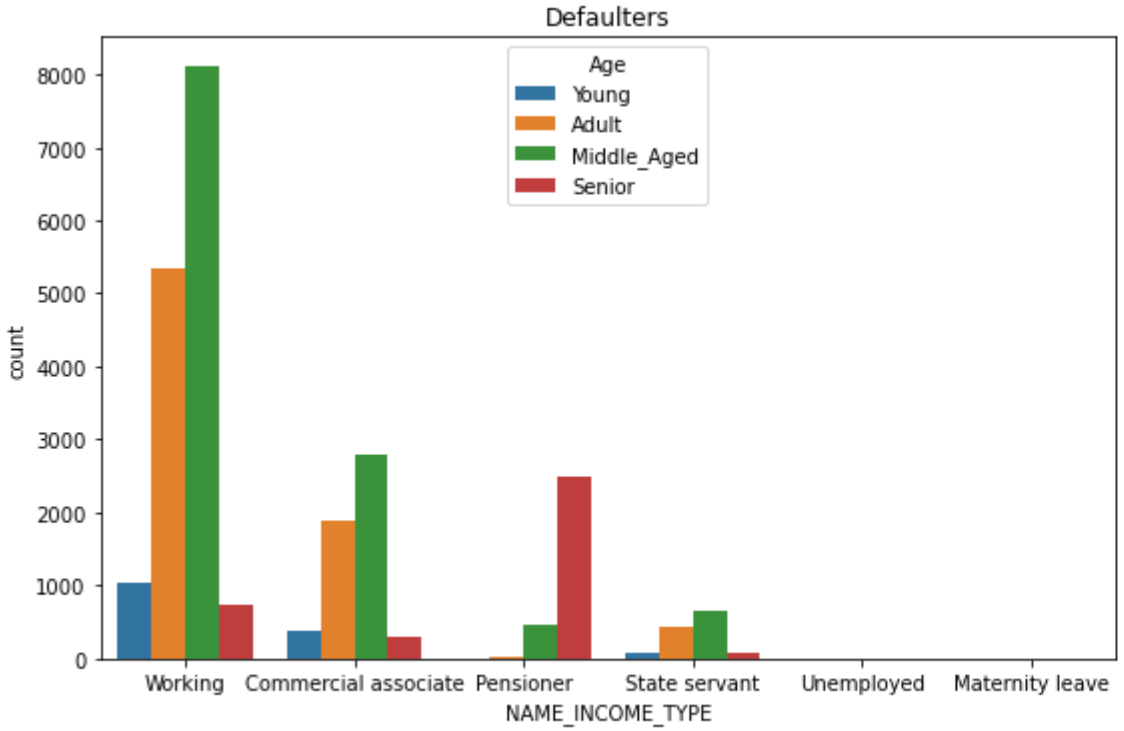
Customer Profiling : Bivariate Analysis

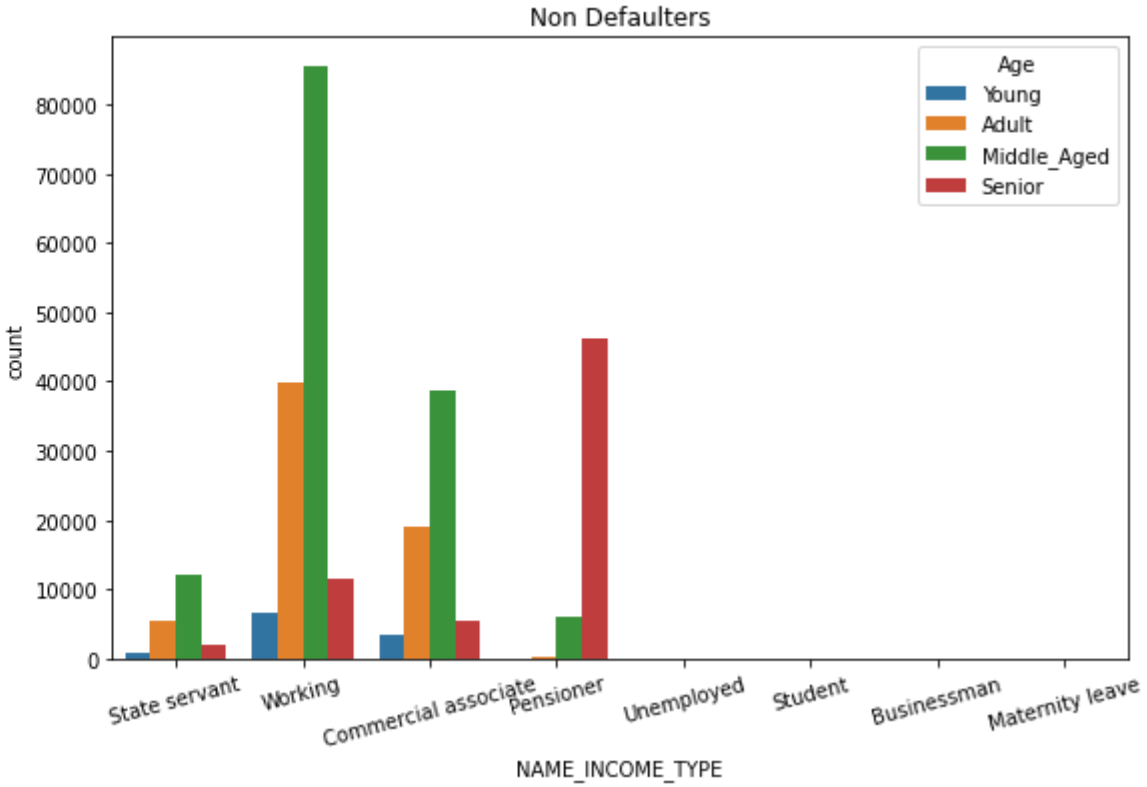


CODE_GENDER :

In case of Defaulters, males with income more than median are more likely to default. Income is balanced through all genders in non defaulters data set.

Customer Profiling : Bivariate Analysis





Analysis :

Middle aged people in the professions of Working and commercial associate are more likely to default. Whereas Senior aged persons who are pensioner are likely to default, which makes sense age wise, as the senior people generally are pensioners.

Customer Profiling : Bivariate Analysis

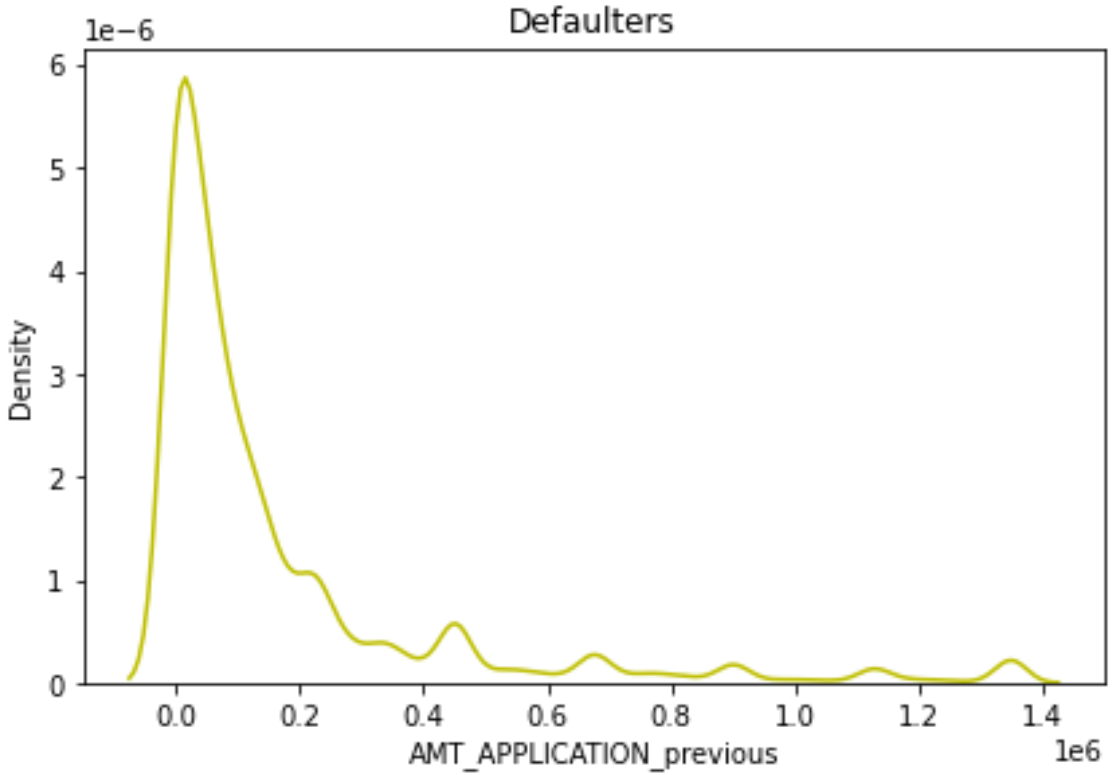
	Variable1	Variable2	Correlation_in_Defaulters
96	AMT_CREDIT	AMT_GOODS_PRICE	0.982110
137	AMT_ANNUITY	AMT_CREDIT	0.758001
135	AMT_ANNUITY	AMT_GOODS_PRICE	0.757306
13	CNT_FAM_MEMBERS	CNT_CHILDREN	0.737829
165	DAYS_BIRTH	DAYS_EMPLOYED	0.626753
136	AMT_ANNUITY	AMT_INCOME_TOTAL	0.427960
83	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.352716
97	AMT_CREDIT	AMT_INCOME_TOTAL	0.350124
167	DAYS_BIRTH	DAYS_REGISTRATION	0.288906
164	DAYS_BIRTH	DAYS_ID_PUBLISH	0.252863

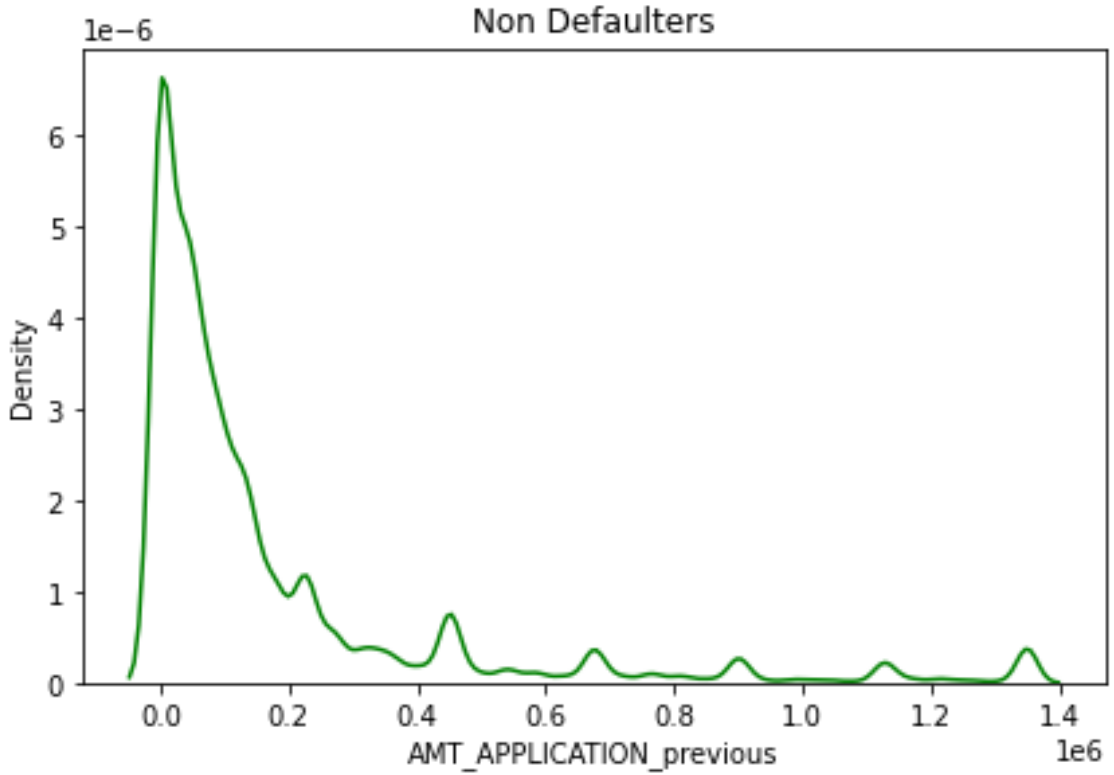
	Variable1	Variable2	Correlation_in_Non_Defaulters
96	AMT_CREDIT	AMT_GOODS_PRICE	0.986490
135	AMT_ANNUITY	AMT_GOODS_PRICE	0.792642
137	AMT_ANNUITY	AMT_CREDIT	0.789822
13	CNT_FAM_MEMBERS	CNT_CHILDREN	0.757436
165	DAYS_BIRTH	DAYS_EMPLOYED	0.674933
136	AMT_ANNUITY	AMT_INCOME_TOTAL	0.488400
83	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.417210
97	AMT_CREDIT	AMT_INCOME_TOTAL	0.410460
167	DAYS_BIRTH	DAYS_REGISTRATION	0.332997
125	DAYS_EMPLOYED	DAYS_ID_PUBLISH	0.280196

Analysis :

Top 9 rows show that the correlations coefficient are almost same among the same set of variables. It means these variables remain independent of either defaulting or non defaulting. The top five variable pair shows the most correlation among them.

Customer Profiling : Univariate Analysis of merged data

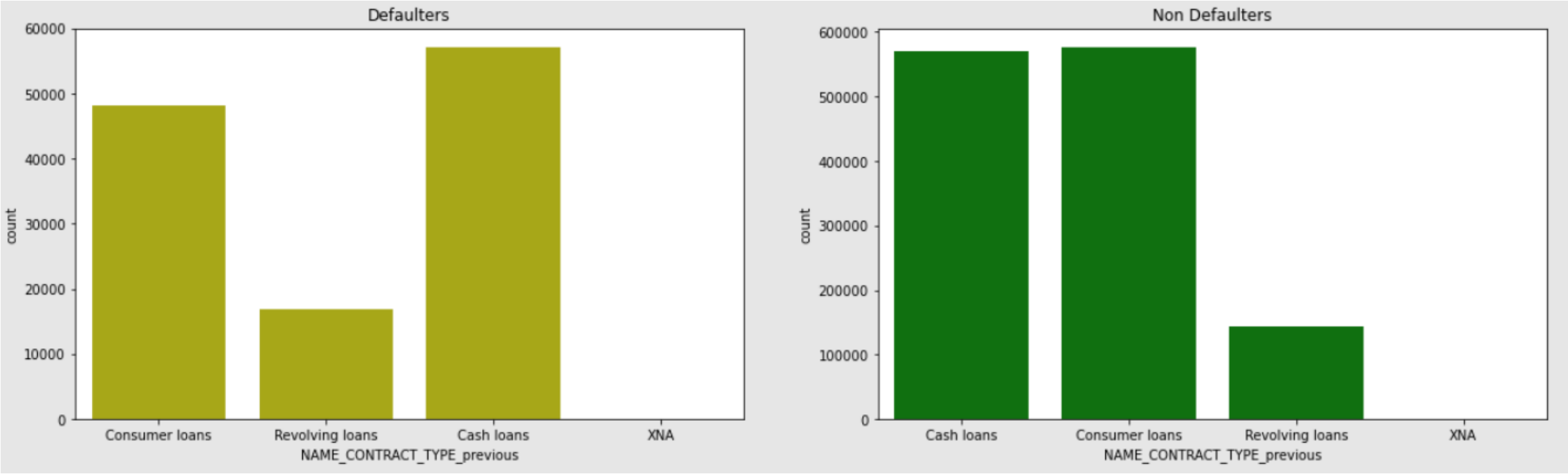




Analysis :

People who applied for less amount were more likely to default.

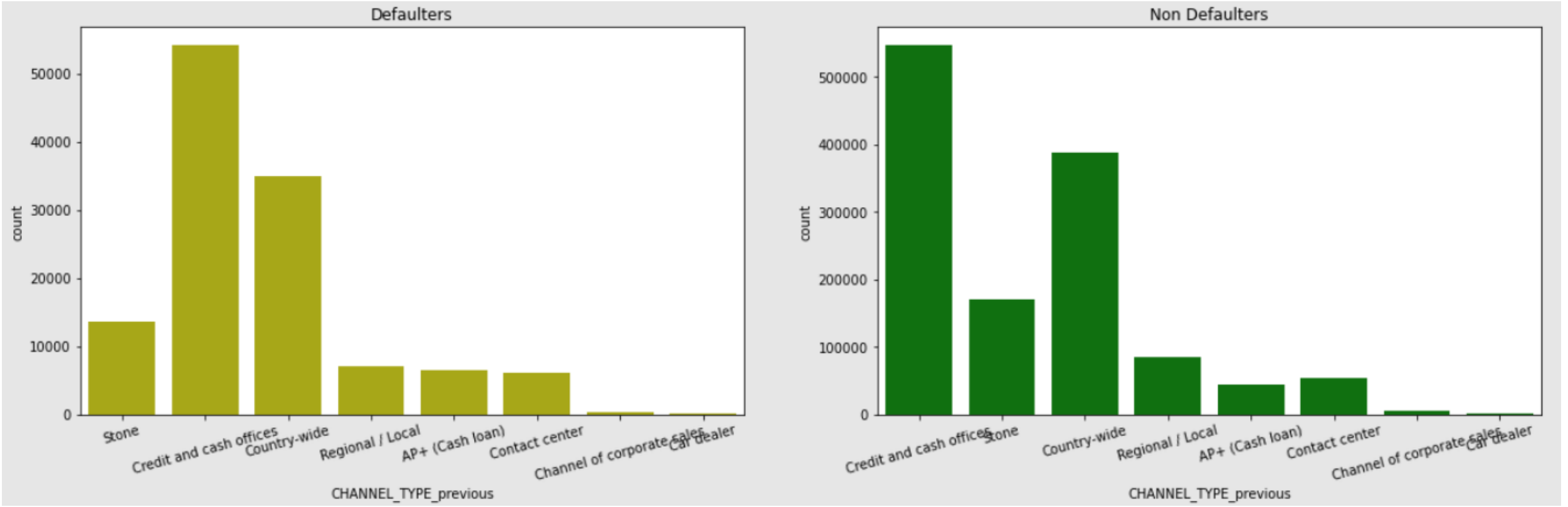
Customer Profiling : Univariate Analysis of merged data



NAME_CONTRACT_TYPE_previous :

people to have taken less consumer loan and more revolving loan are more likely to default

Customer Profiling : Univariate Analysis of merged data



CHANNEL_TYPE_previous : People to have used less Stone and more AP+Cash loan channel are likely to default

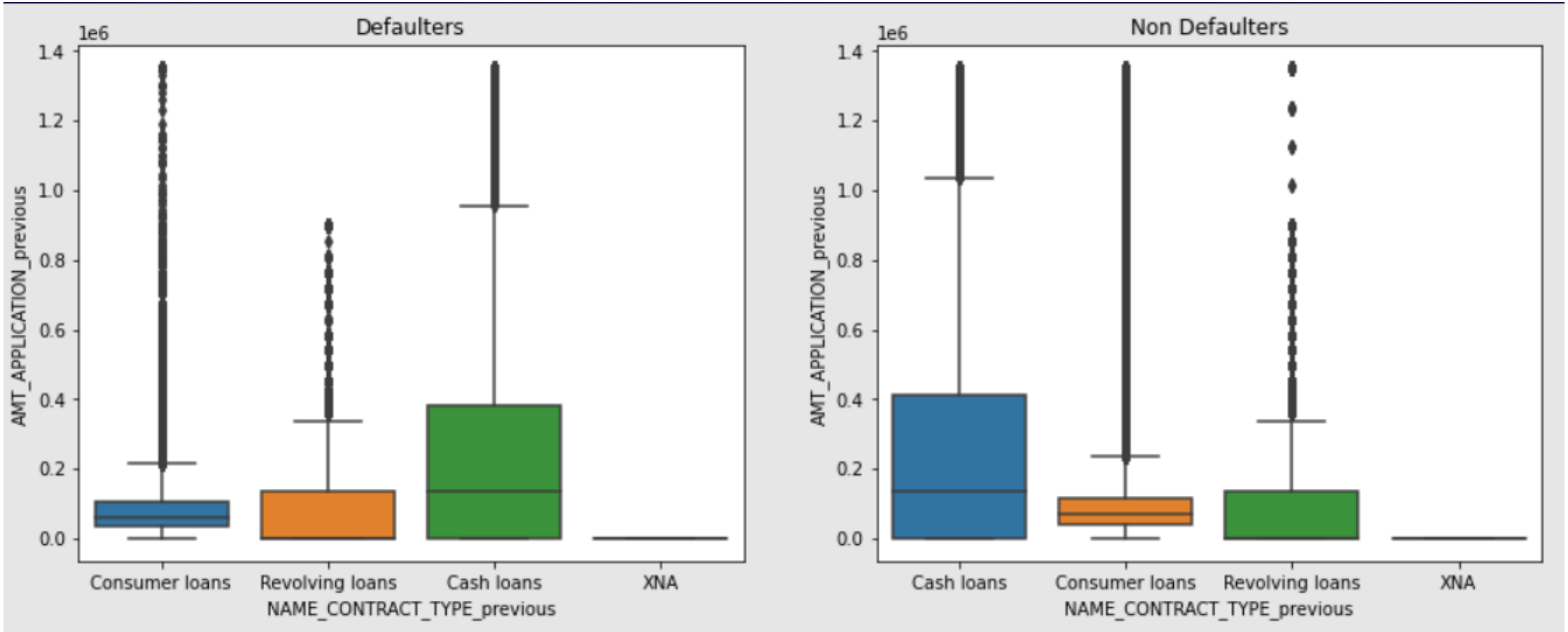
Customer Profiling : Bivariate Analysis of merged data

	Variable1	Variable2	Correlation_in_Defaulters
16	AMT_GOODS_PRICE_previous	AMT_APPLICATION_previous	0.999515
17	AMT_GOODS_PRICE_previous	AMT_CREDIT_previous	0.977044
11	AMT_CREDIT_previous	AMT_APPLICATION_previous	0.959106
15	AMT_GOODS_PRICE_previous	AMT_ANNUITY_previous	0.862848
10	AMT_CREDIT_previous	AMT_ANNUITY_previous	0.851443
5	AMT_APPLICATION_previous	AMT_ANNUITY_previous	0.845251
21	CNT_PAYMENT_previous	AMT_APPLICATION_previous	0.718685
23	CNT_PAYMENT_previous	AMT_GOODS_PRICE_previous	0.708855
22	CNT_PAYMENT_previous	AMT_CREDIT_previous	0.690539
20	CNT_PAYMENT_previous	AMT_ANNUITY_previous	0.496898

	Variable1	Variable2	Correlation_in_Defaulters
16	AMT_GOODS_PRICE_previous	AMT_APPLICATION_previous	0.999762
17	AMT_GOODS_PRICE_previous	AMT_CREDIT_previous	0.976259
11	AMT_CREDIT_previous	AMT_APPLICATION_previous	0.957324
15	AMT_GOODS_PRICE_previous	AMT_ANNUITY_previous	0.849559
5	AMT_APPLICATION_previous	AMT_ANNUITY_previous	0.834855
10	AMT_CREDIT_previous	AMT_ANNUITY_previous	0.831359
21	CNT_PAYMENT_previous	AMT_APPLICATION_previous	0.701767
23	CNT_PAYMENT_previous	AMT_GOODS_PRICE_previous	0.692109
22	CNT_PAYMENT_previous	AMT_CREDIT_previous	0.674575
20	CNT_PAYMENT_previous	AMT_ANNUITY_previous	0.424478

Analysis :
From the correlation charts above, it can be seen that these four variables:
AMT_ANNUITY_previous,AMT_APPLICATION_previous,AMT_CREDIT_previous,AMT_GOODS_PRICE_previous
have strong correlations among them, which means change in one variable would impact others significantly

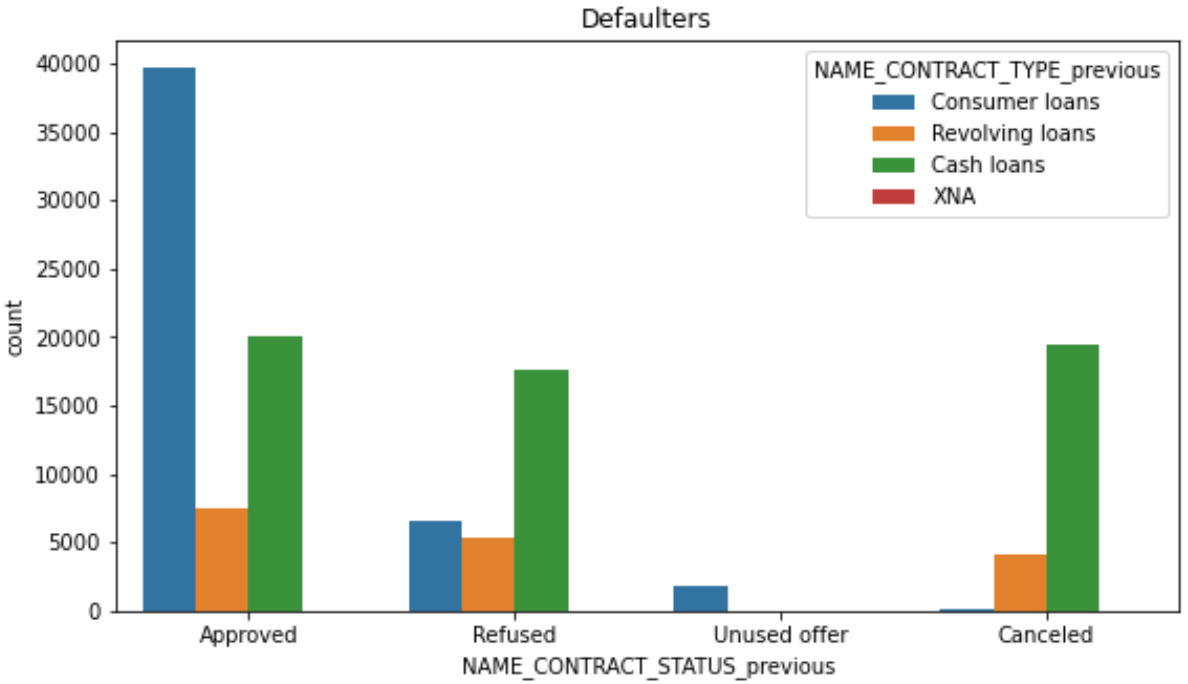
Customer Profiling : Bivariate Analysis of merged data

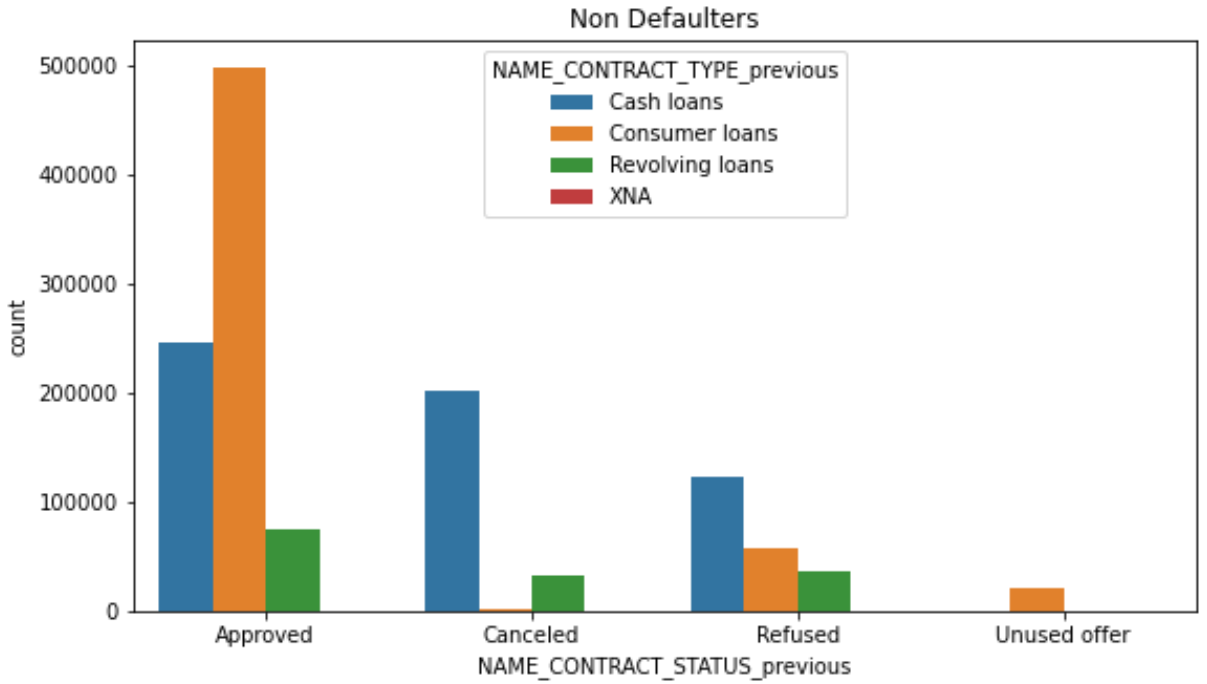


Analysis :

The type of loan taken in the previous cases do not have a distinguishing factor for identifying defaulters as the plots remain same

Customer Profiling : Bivariate Analysis of merged data

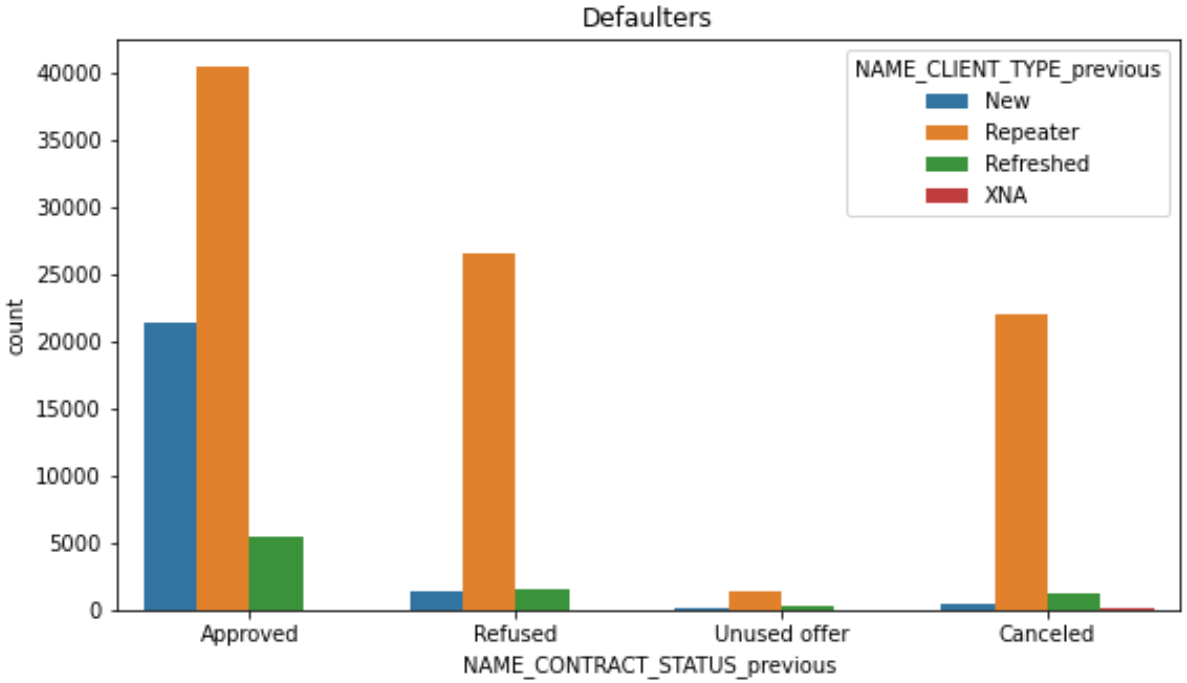


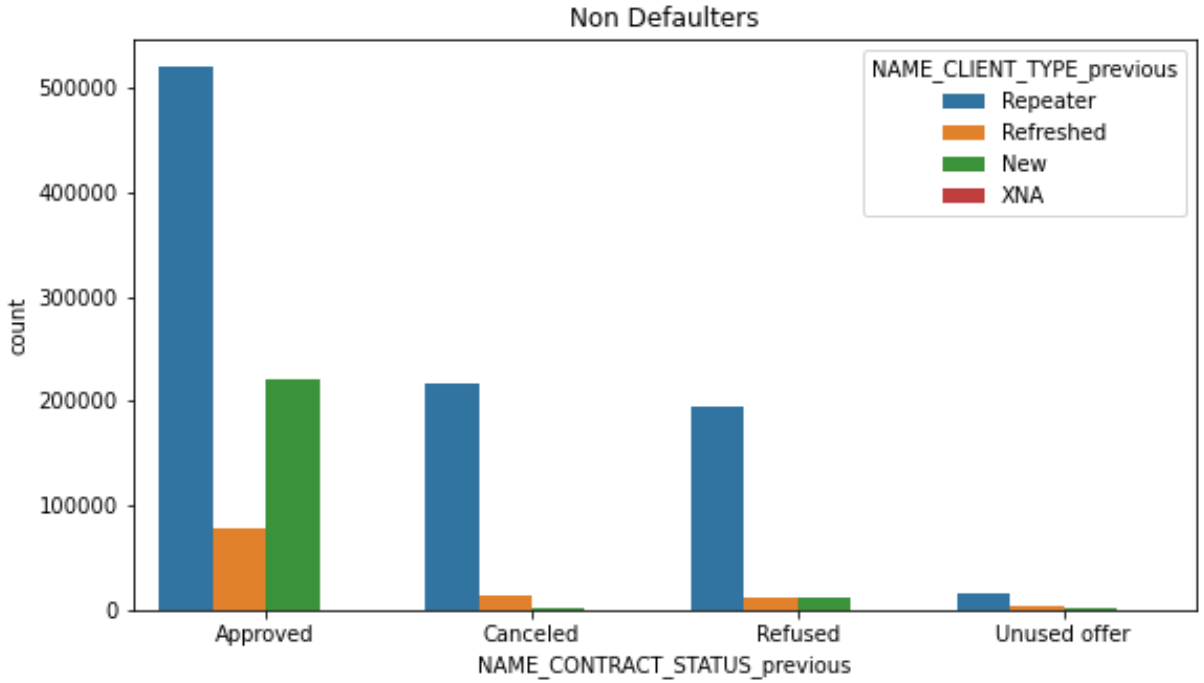


Analysis :

People whose applied loan was 'refused' or 'cancelled' with a cash loan previously, are more likely to default. Also people applied for Consumer loans and were approved previously, are more likely to default too.

Customer Profiling : Bivariate Analysis of merged data





Analysis :

Very evidently, the 'repeated' clients, whose loans were 'refused' or 'cancelled' previously are more likely to default

Recommendations

From the EDA analysis we can conclude the below points to figure out the profile of a person who is more likely to default :

- Lower income group & lower amount of loan taken - tends to default more.
- Customers who were ‘Repeaters’ and were ‘Refused’ in previous applications.
- Females are more likely to default.
- Cash loans are counted as default, revolving loan is safer option.
- People who received less Higher education are more likely to default.
- Single/unmarried people are more likely to default.
- people who tried to pay via cash and their payment was refused previously, are more likely to default

Conclusion

TARGET	NAME_CONTRACT_STATUS_previous	
0	Approved	0.579229
	Canceled	0.166684
	Refused	0.152756
	Unused offer	0.014778
1	Approved	0.047565
	Canceled	0.016835
	Refused	0.020823
	Unused offer	0.001329

From this table, we can see that previously 15% refused loan was turned out to be ‘Non defaulters’, hence that is a loss for the financial institution.

previously 4.75% approved loan was turned out to be ‘Defaulters’, hence that is a loss for the financial institution as well.

So the EDA process can be used to reduce this percentages in order to avoid financial loss in future.