Credit EDA Case Study

Mousham Kuri

Rithesh Kamath

Data Cleaning, Formatting & Imputations

- Number of columns in Application_data set with missing value more than 50 % : 41
  - Action Taken : Deleted all these columns from the dataset
- Number of columns in Previous_application dataset with missing value more than 50 % : 4
  - Action Taken : Deleted all these columns from the dataset
- Converted negative values to absolute values for the following columns :
  - DAYS_BIRTH
  - DAYS_EMPLOYED
  - DAYS_REGISTRATION
  - DAYS_ID_PUBLISH
  - DAYS_LAST_PHONE_CHANGE
- Data Imputation :
  - Categorical variable : Filled with “Unknown”
  - Numeric variables : Used mean/median/mode as per data

Outlier Detection & Analysis

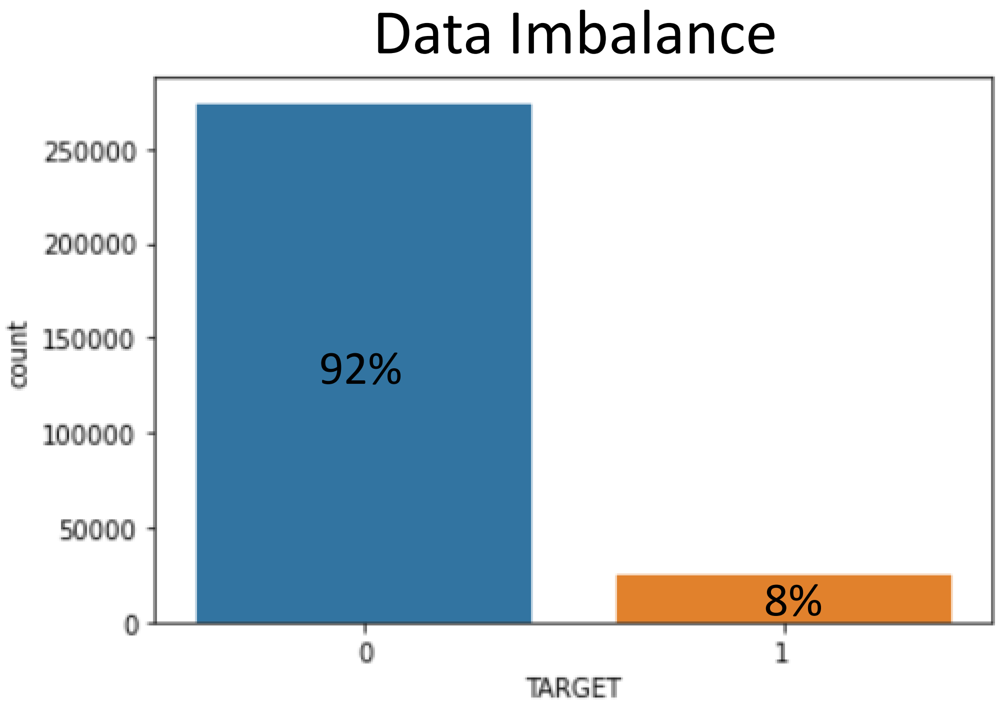
- Outlier detection has been done for the below columns in the Application_data set :
  - AMT_INCOME_TOTAL
  - AMT_ANNUITY
  - AMT_CREDIT

- Outlier detection has been done for the below columns in the Previous_application dataset :
  - AMT_GOODS_PRICE
  - AMT_APPLICATION
  - CNT_PAYMENT

- Outliers have been handled on the above mentioned columns of both the data sets :
  - Values outside of 99 percentile has been excluded

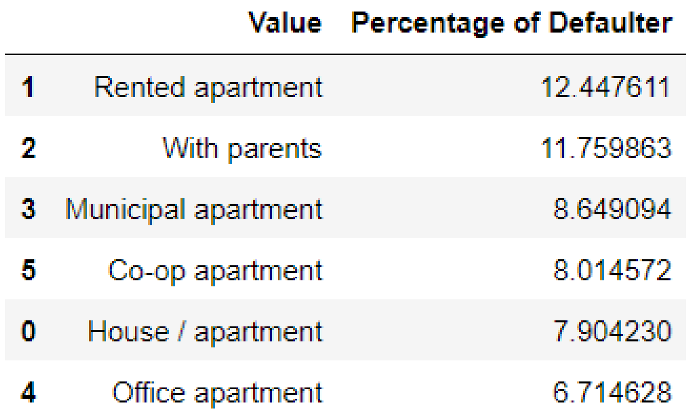
Data Imbalance Ratio

- Number of defaulters in the total population is very less which makes the data highly imbalanced.
- Number of defaulters : 24413
- Number of non-defaulters : 273926
- Percentage of defaulters : 8%
- Percentage of non-defaulters : 92%
- The ratio of defaulters to non-defaulters is 8:92 = 2:23

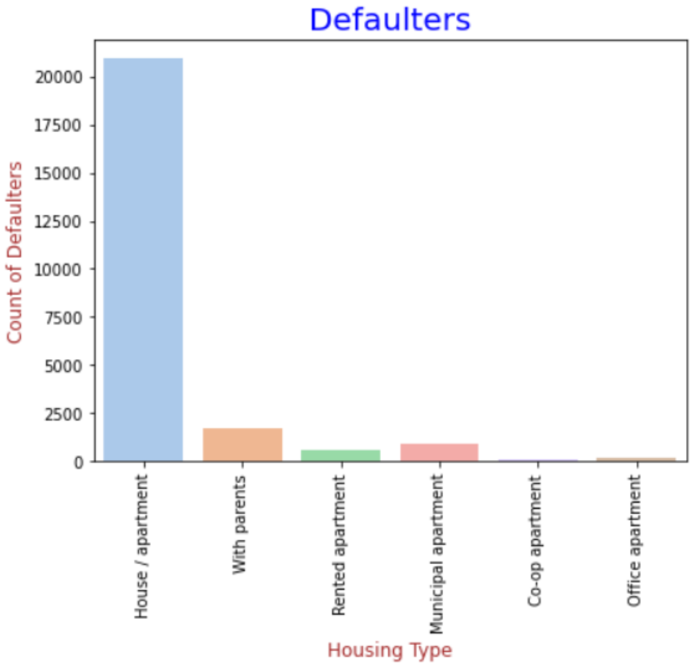


Housing Information of Applicant Univariate Analysis

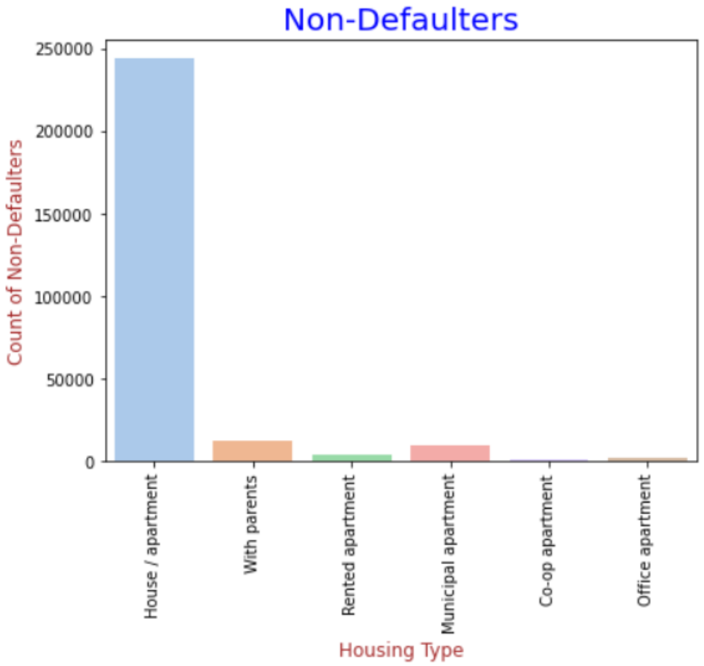
- Large number of people stay in House / apartment and very few stay in Co-op apartment
- People staying in Rented apartment and with parents have the highest chance of being defaulters whereas people staying in Office apartment have the least chance of defaulting



	Value	Percentage of Defaulter
1	Rented apartment	12.447611
2	With parents	11.759863
3	Municipal apartment	8.649094
5	Co-op apartment	8.014572
0	House / apartment	7.904230
4	Office apartment	6.714628



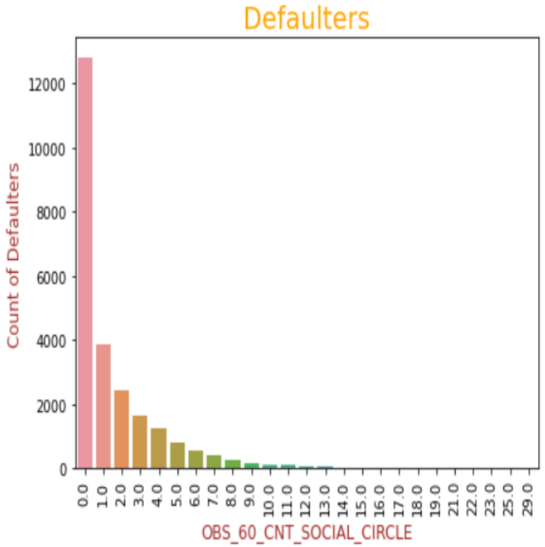
Defaulters

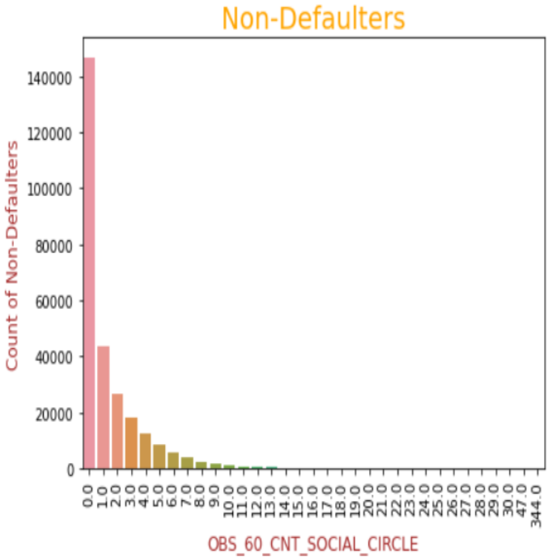


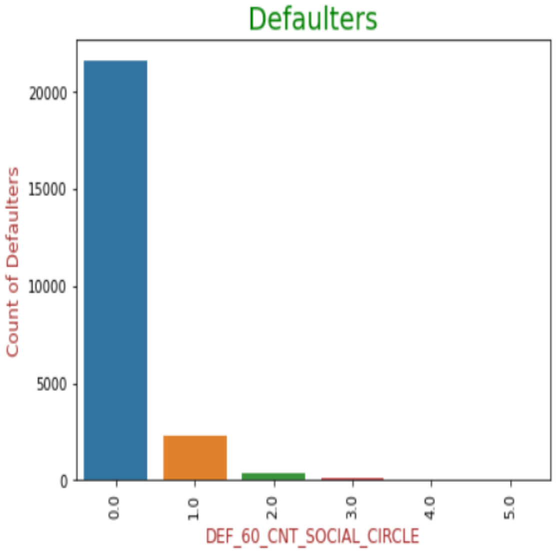
Non-Defaulters

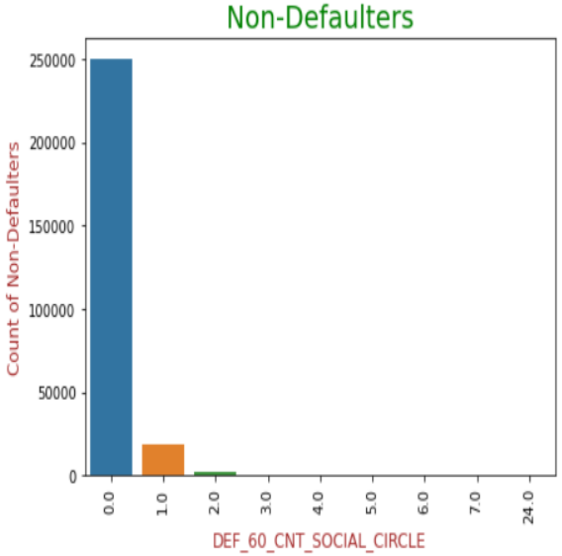
Social Circle Info Univariate Analysis

- DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE are highly correlated
- For both defaulters as well as non-defaulters 'DEF_60_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE' columns show a similar trend



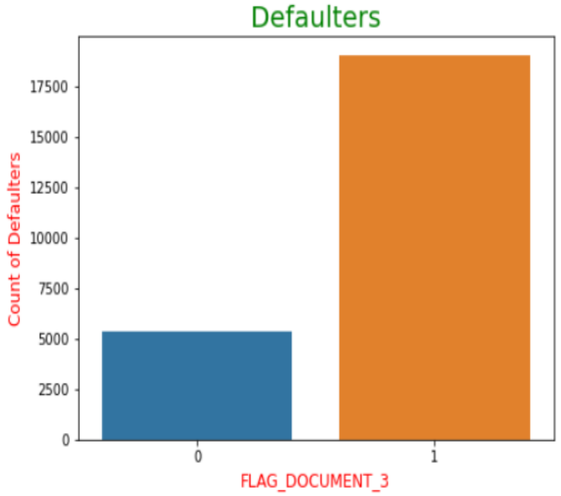


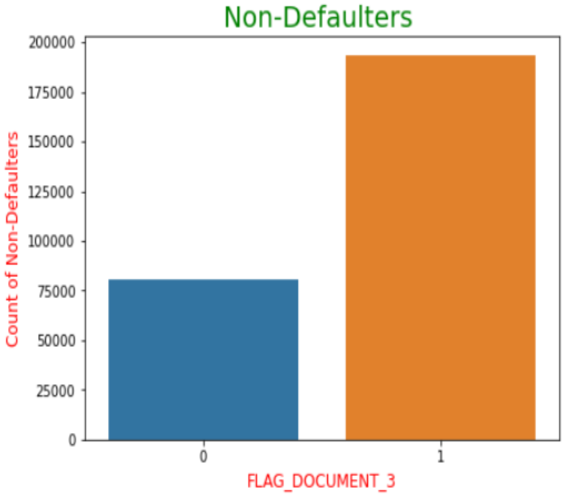


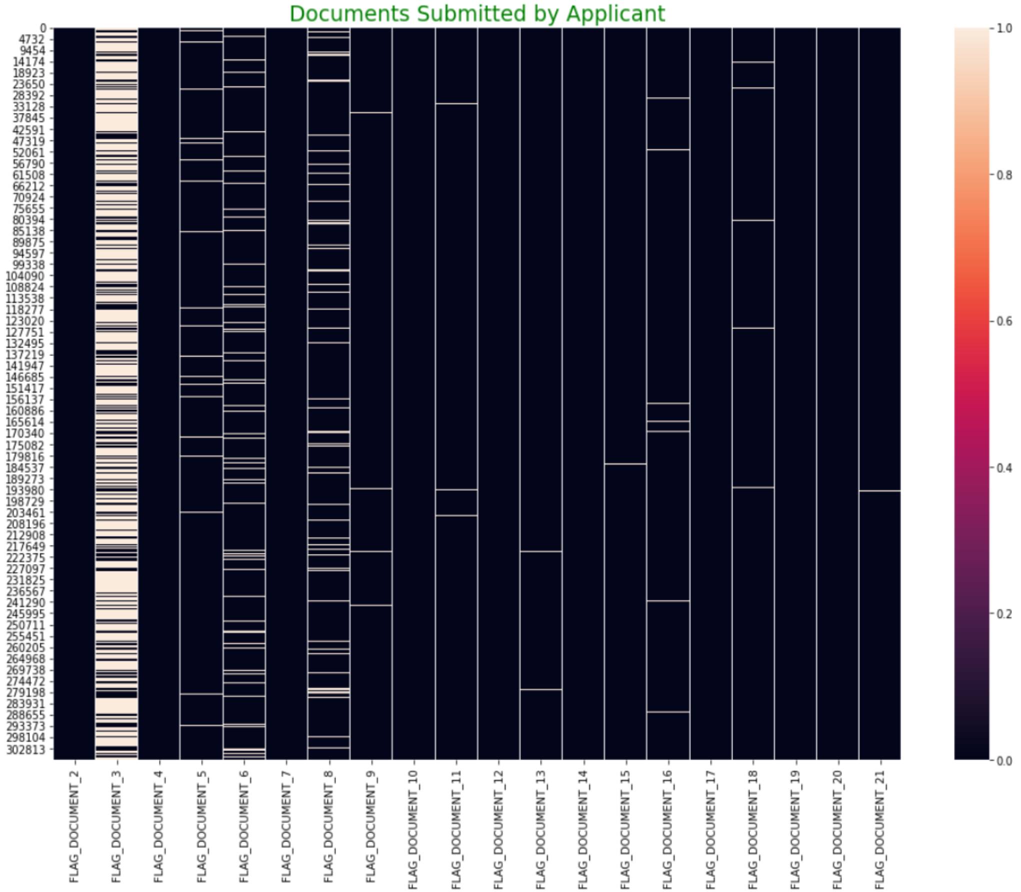


Documents Submitted by Applicants Univariate Analysis

- Apart from Document_3, majority of the applicants did not submit any other documents. However, the document_3 submission follows a similar trend for defaulters as well as non-defaulters.

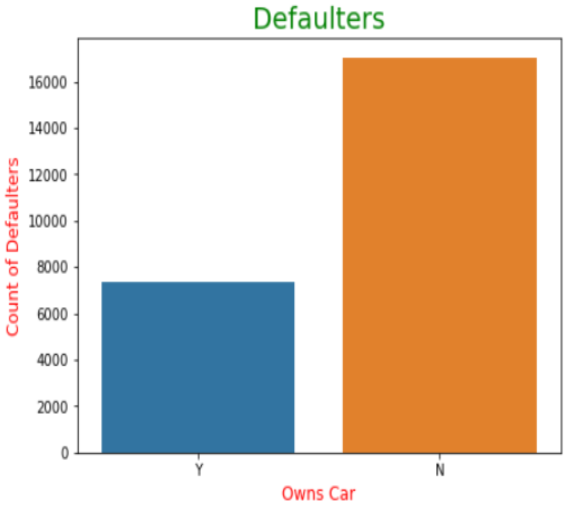


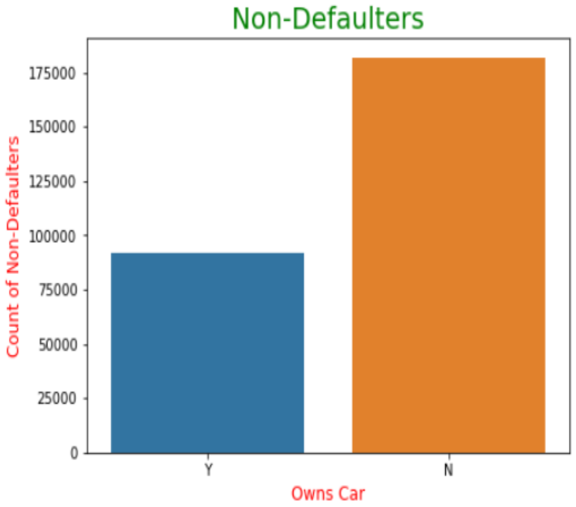


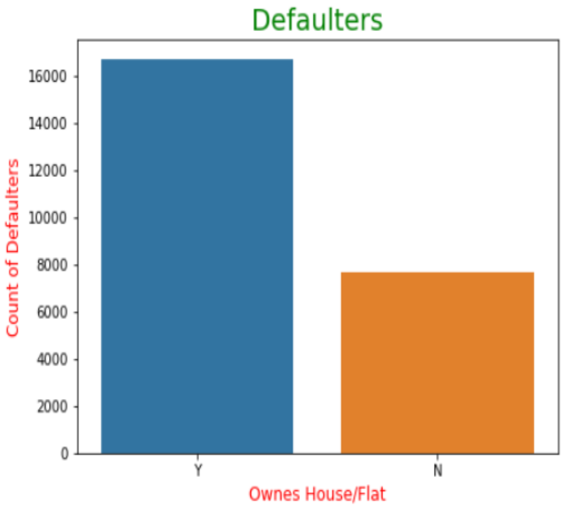


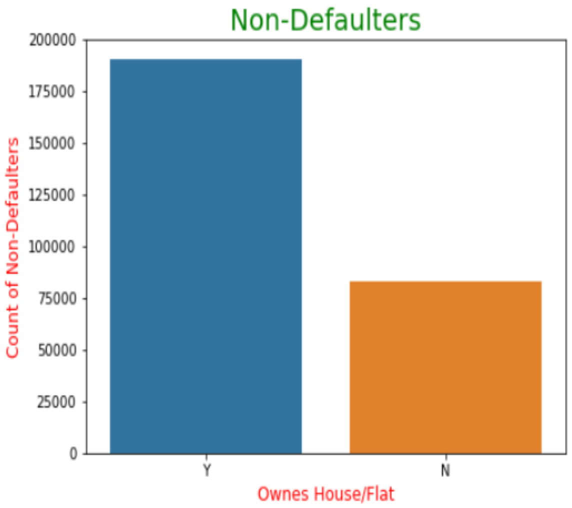
Asset Details Univariate Analysis

- More number of Non-Defaulters own cars & Houses/Flat compared to defaulters. Hence, it can be concluded that people not owning realty and cars have a slightly higher defaulting rate than the people who own realty and cars


Defaulters

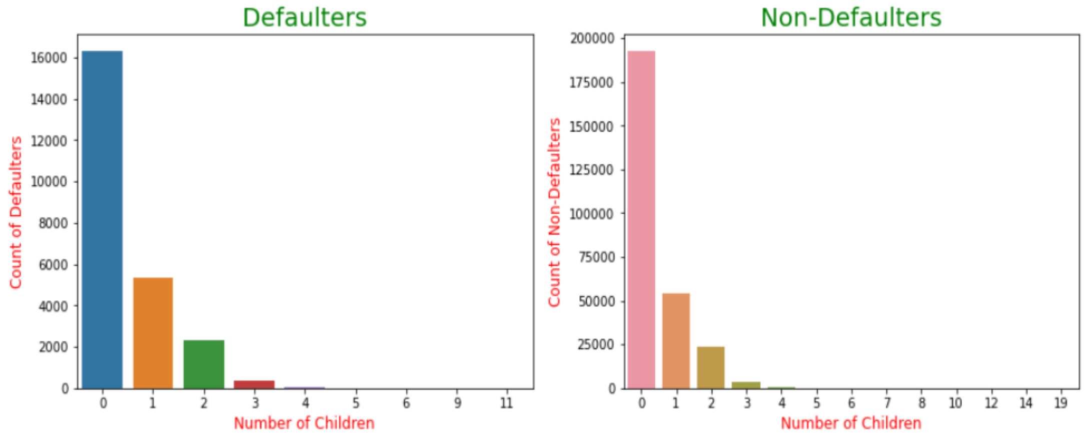

Non-Defaulters

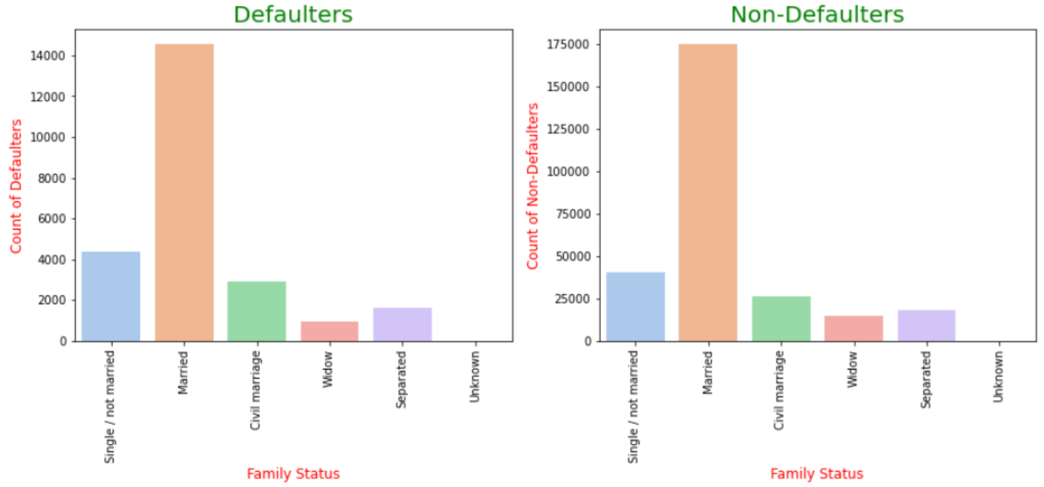

Defaulters

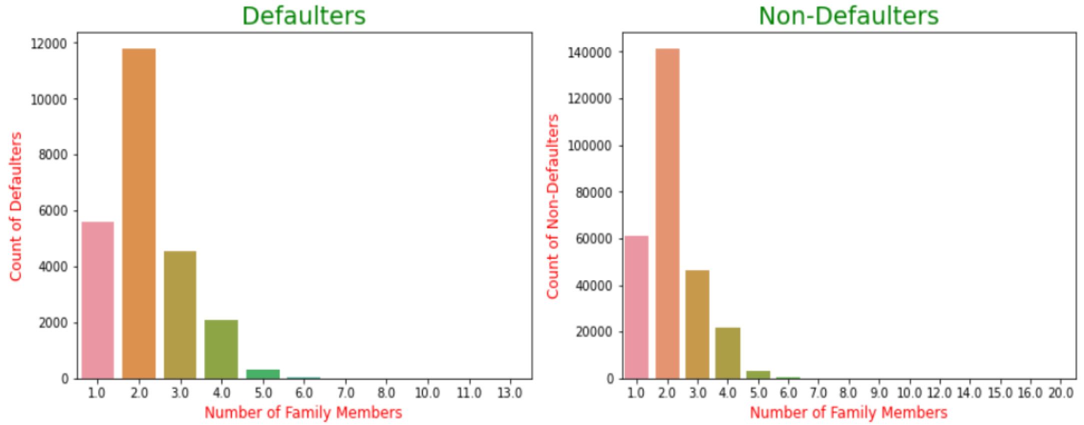

Non-Defaulters

Family Related Info Univariate Analysis

- Majority People in the dataset are married and have no children with 2 family members
- Civil Marriage and Single/not married people have the highest defaulters rate



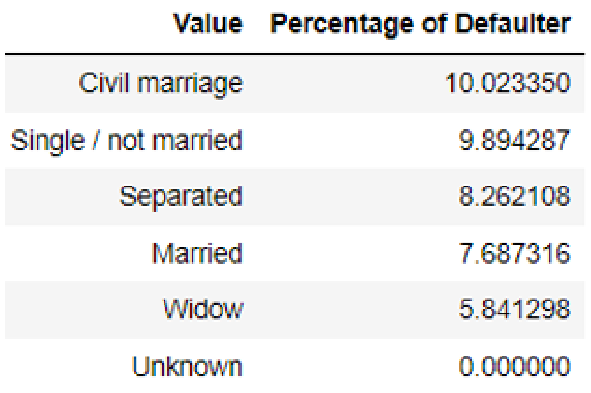




Family Related Info (Contd..)

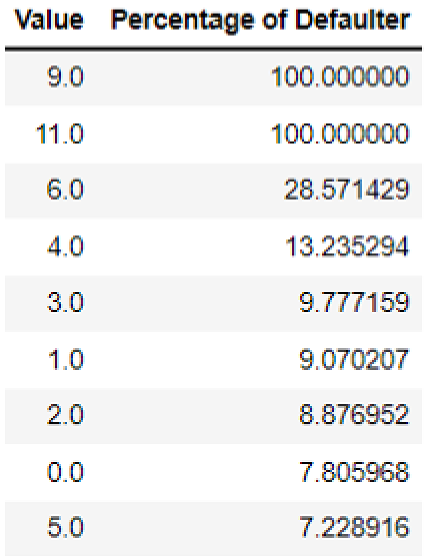
- People with more number of children and family members have higher defaulting percentage

Count of Family Status



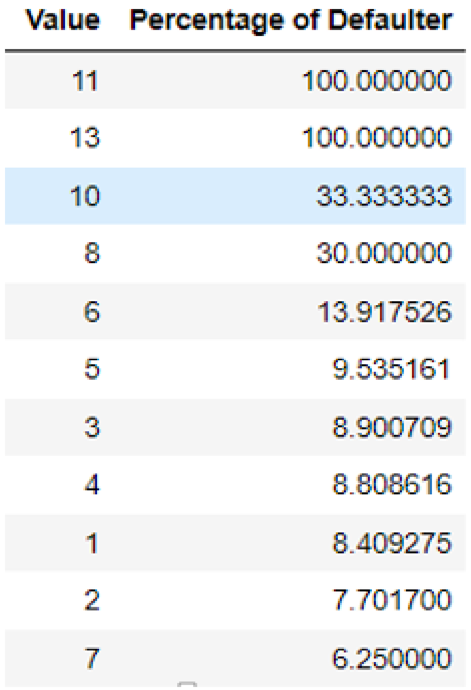
Value	Percentage of Defaulter
Civil marriage	10.023350
Single / not married	9.894287
Separated	8.262108
Married	7.687316
Widow	5.841298
Unknown	0.000000

Count of children



Value	Percentage of Defaulter
9.0	100.000000
11.0	100.000000
6.0	28.571429
4.0	13.235294
3.0	9.777159
1.0	9.070207
2.0	8.876952
0.0	7.805968
5.0	7.228916

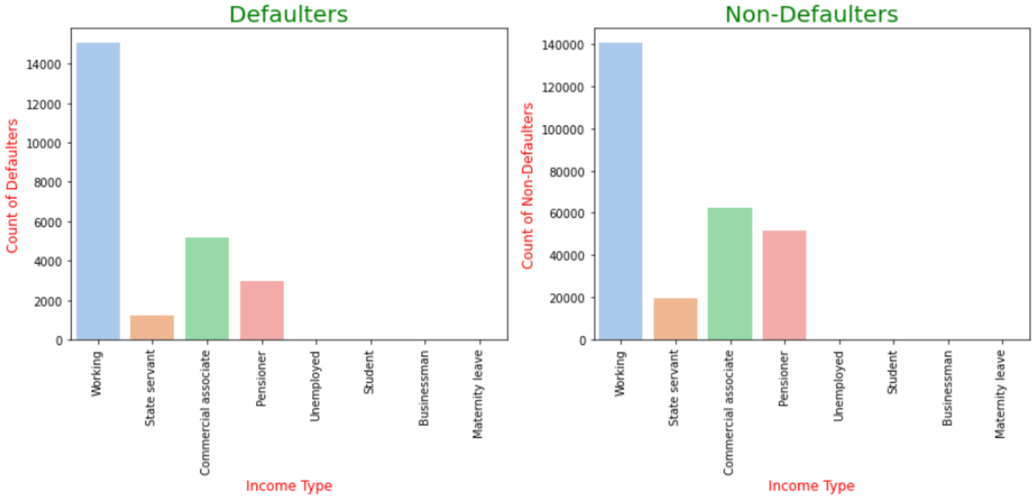
Count of Family Members

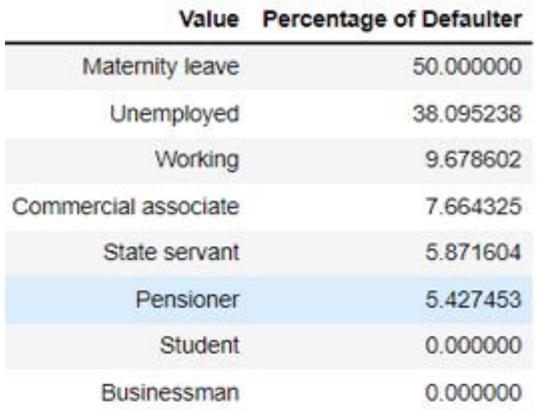


Value	Percentage of Defaulter
11	100.000000
13	100.000000
10	33.333333
8	30.000000
6	13.917526
5	9.535161
3	8.900709
4	8.808616
1	8.409275
2	7.701700
7	6.250000

Education and Occupation Info Univariate Analysis

- Most of the people part of this data set are working professionals
- People on Maternity Leave and Unemployed people have the highest defaulting percentage whereas Students & Businessmen have the least defaulting percentage
- People with Lower secondary as their education have the highest defaulting percentage whereas people with Academic degree have the least defaulting percentage



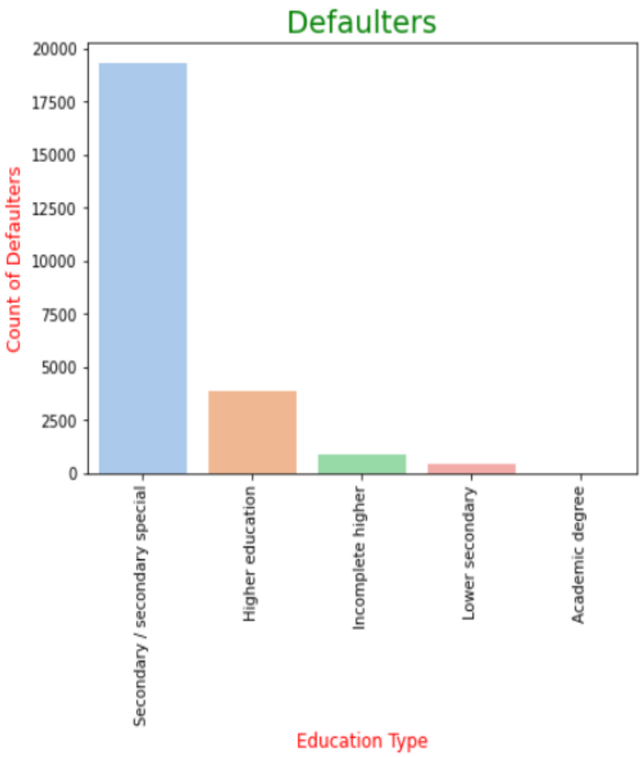


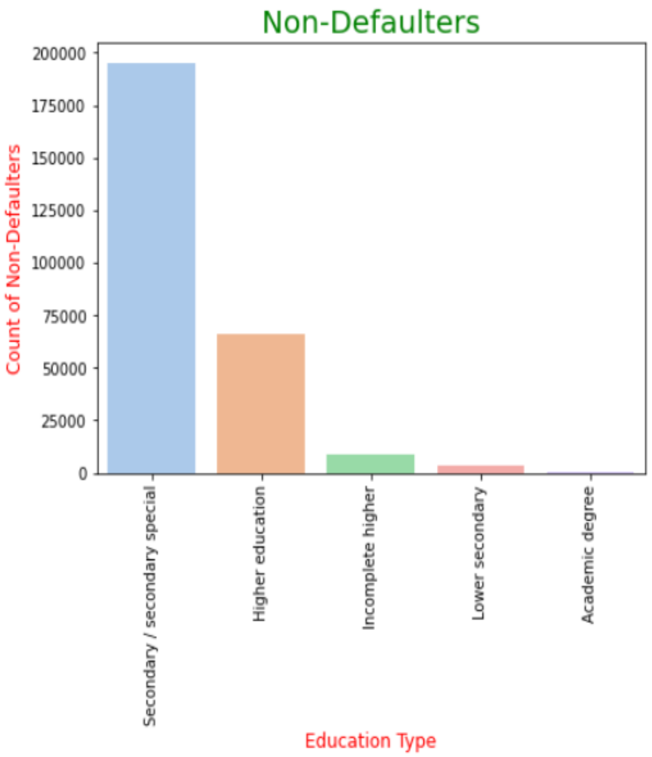
Value	Percentage of Defaulter
Maternity leave	50.000000
Unemployed	38.095238
Working	9.678602
Commercial associate	7.664325
State servant	5.871604
Pensioner	5.427453
Student	0.000000
Businessman	0.000000

Count of Income Type

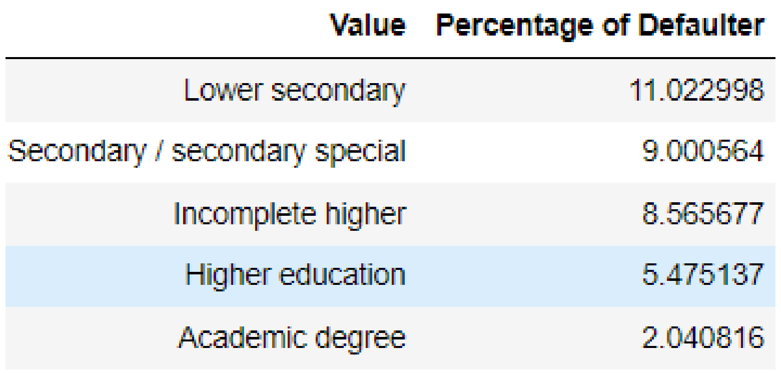
Education and Occupation Info (Contd..)

- Low-skill Laborers have highest defaulting percentage whereas Accountants have the least defaulting percentage


Defaulters

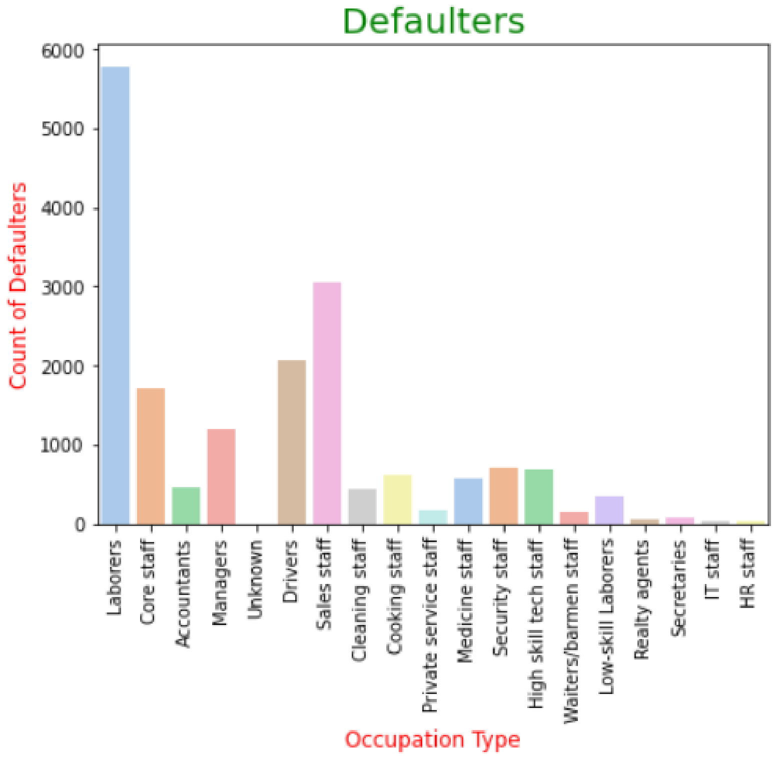

Non-Defaulters

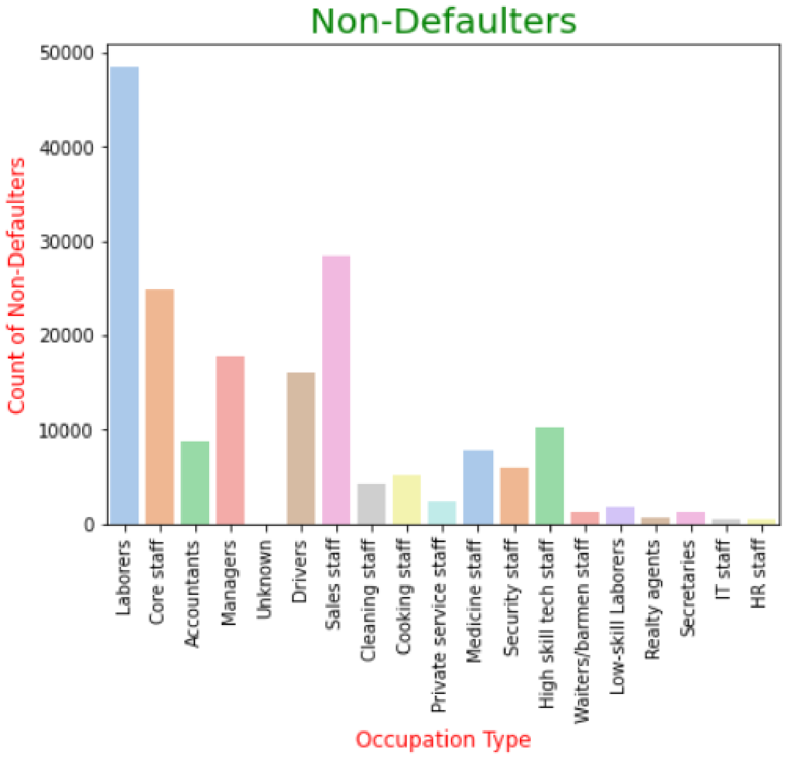
Count of Education Type



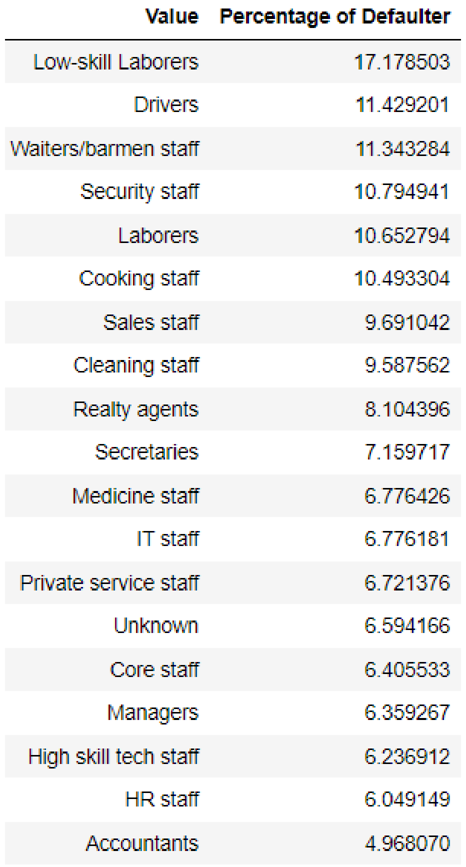
Value	Percentage of Defaulter
Lower secondary	11.022998
Secondary / secondary special	9.000564
Incomplete higher	8.565677
Higher education	5.475137
Academic degree	2.040816

Education and Occupation Info (Contd..)





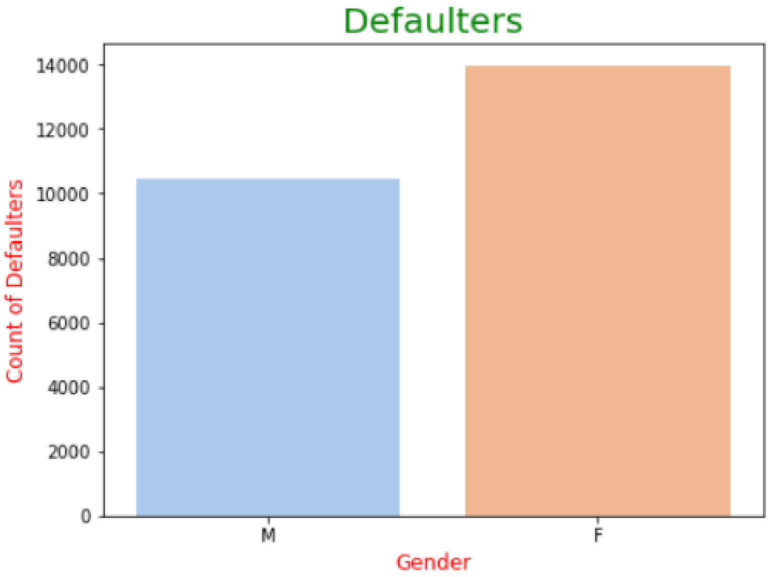
Count of Occupation Type

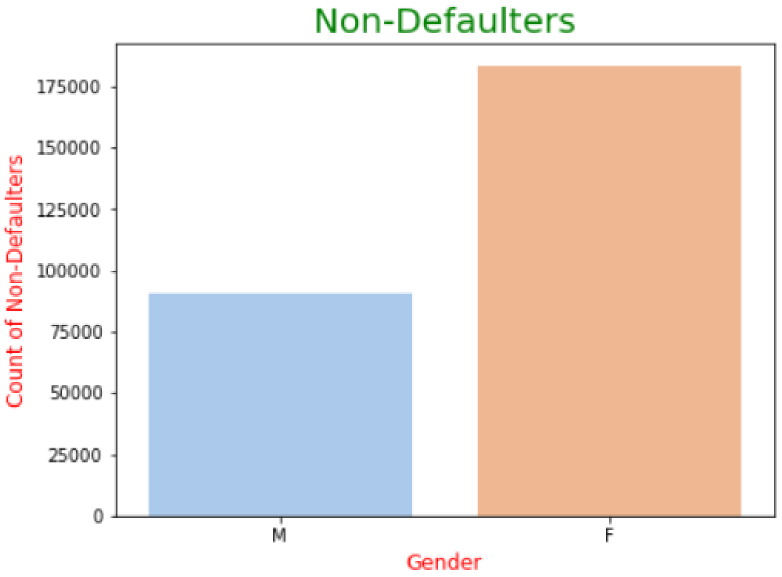


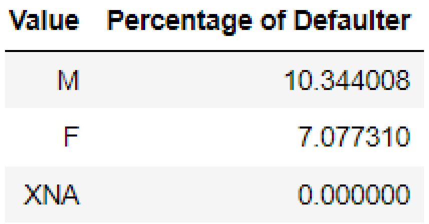
Value	Percentage of Defaulter
Low-skill Laborers	17.178503
Drivers	11.429201
Waiters/barmen staff	11.343284
Security staff	10.794941
Laborers	10.652794
Cooking staff	10.493304
Sales staff	9.691042
Cleaning staff	9.587562
Realty agents	8.104396
Secretaries	7.159717
Medicine staff	6.776426
IT staff	6.776181
Private service staff	6.721376
Unknown	6.594166
Core staff	6.405533
Managers	6.359267
High skill tech staff	6.236912
HR staff	6.049149
Accountants	4.968070

Gender Univariate Analysis

- Regardless of defaulters or non-defaulters, there are more number of females in the dataset than males
- Males have a higher defaulting percentage



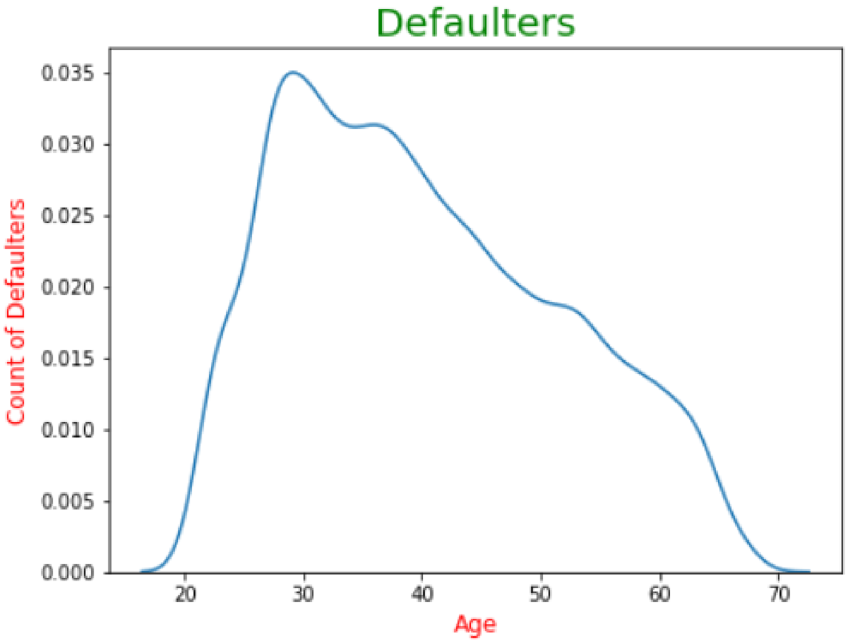


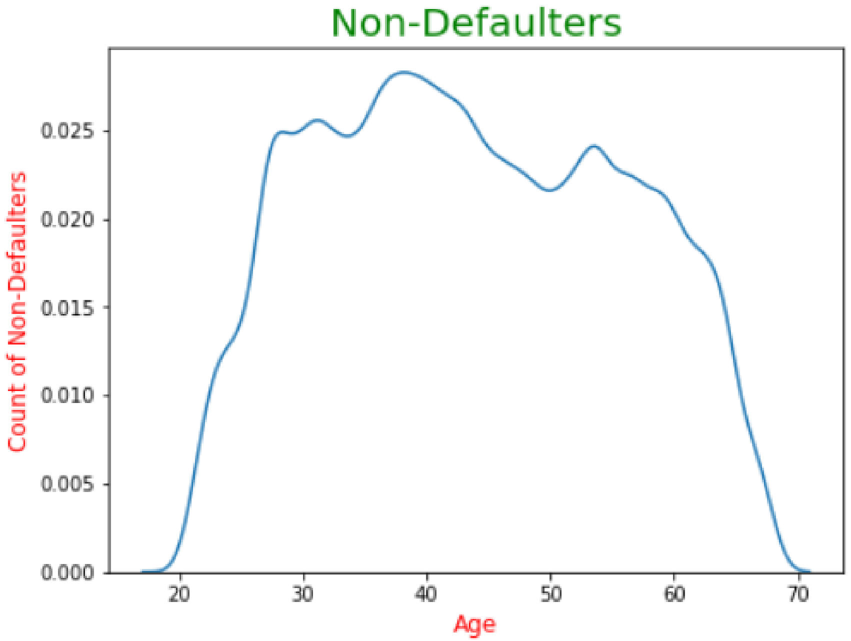


Value	Percentage of Defaulter
M	10.344008
F	7.077310
XNA	0.000000

Age Univariate Analysis

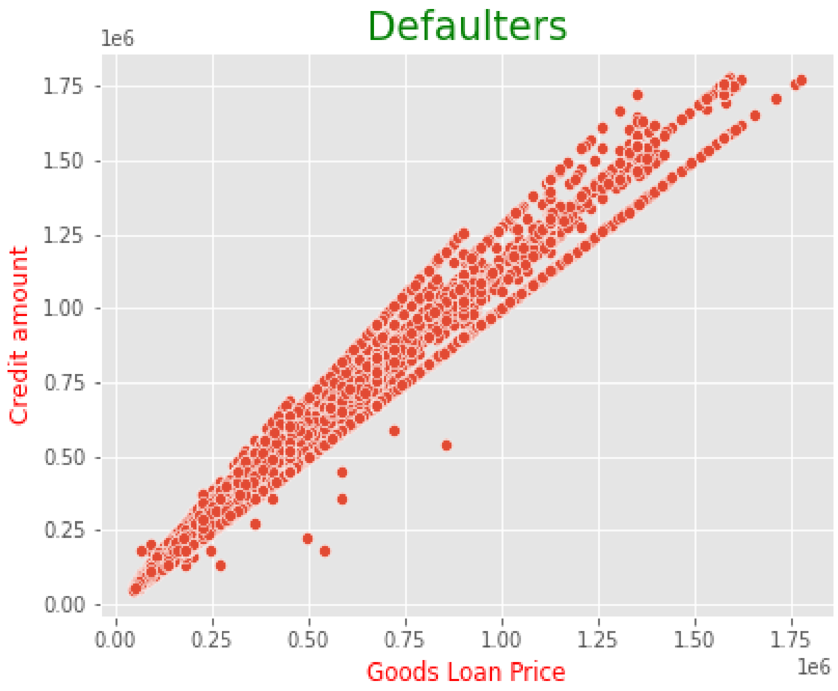
- People around the age of 30 years have highest defaulting rate
- Defaulting rate reduces as the age crosses 30 years

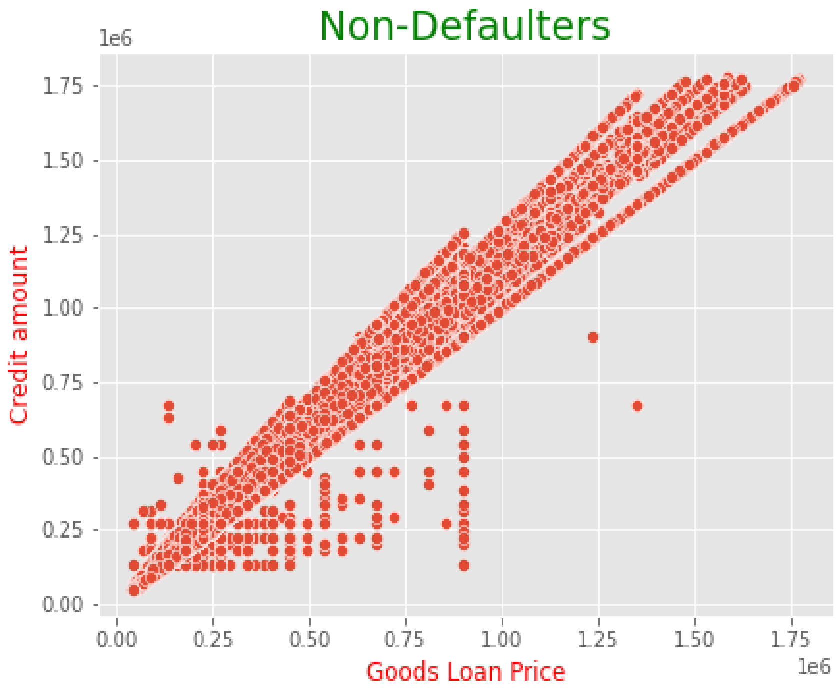




Credit amount and goods price Bivariate Analysis

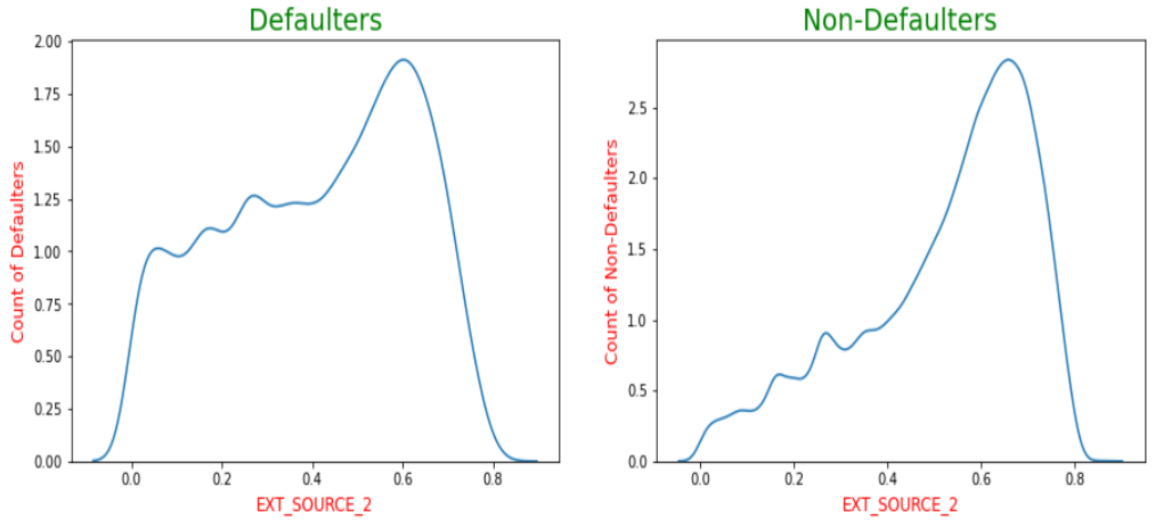
- AMT_CREDIT and AMT_GOODS_PRICE follow a linear relation where the credit amount of the loan increases with increase in the price of the goods for which the loan is given
- The number of defaulters is less than that of non-defaulters for lower range of credit amount and goods price



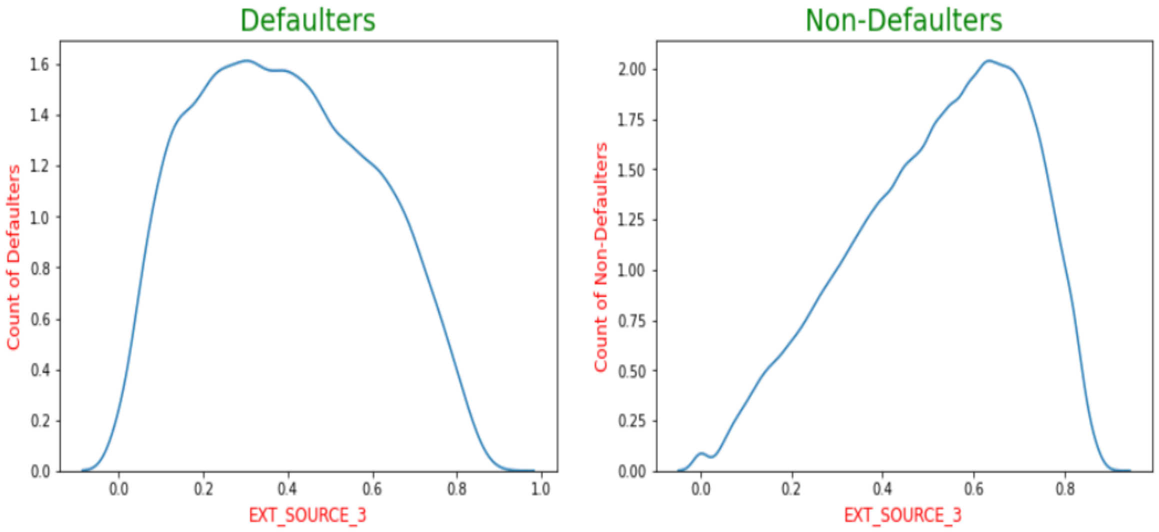


Normalized score Univariate Analysis

- EXT_SOURCE_3 has a very different distribution for defaulters and non-defaulters whereas EXT_SOURCE_2 is almost similar for both defaulters & non-defaulters



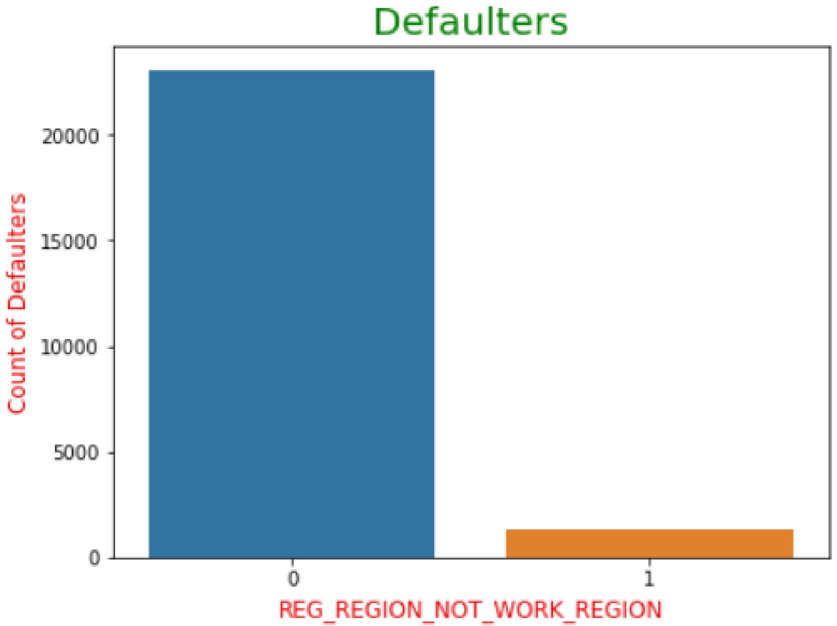
EXT_SOURCE_2

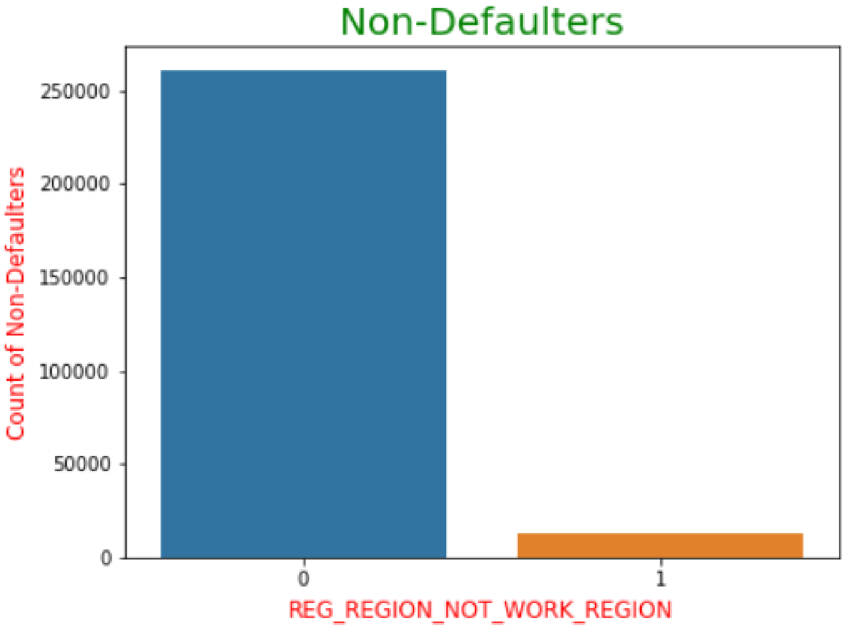


EXT_SOURCE_3

Address Univariate Analysis

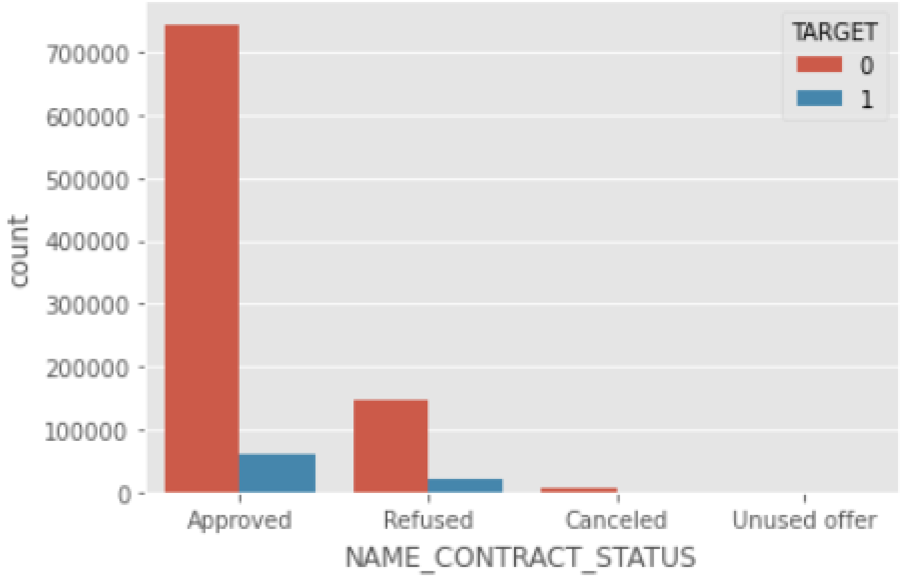
- Defaulter rate is highest when the permanent address is same as the working address which is plotted by REG_REGION_NOT_WORK_REGION = 0





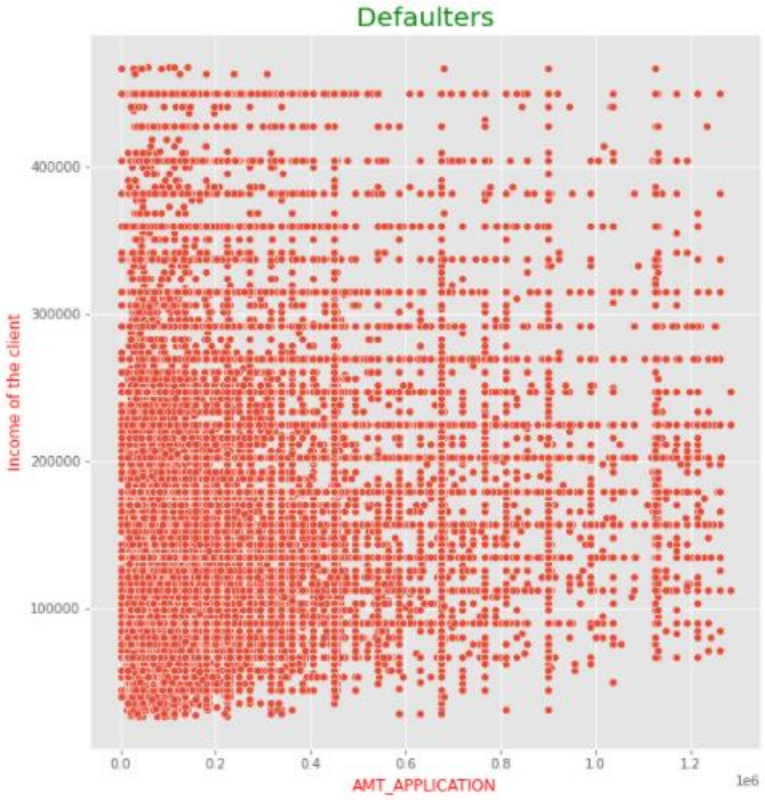
Previous Application Data Analysis

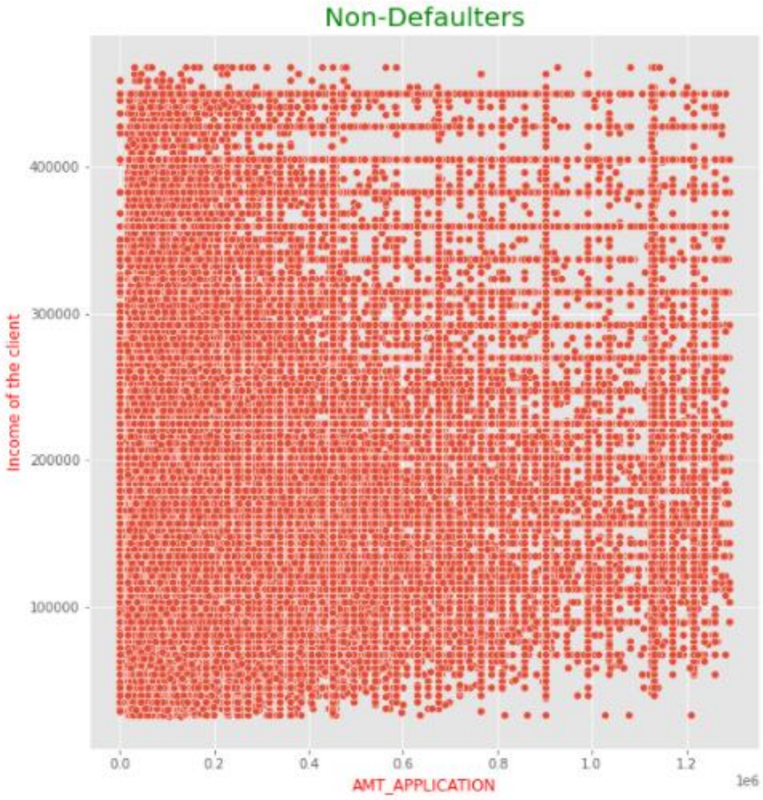
- 8% of the previously approved loan applicants defaulted in current loan
- 88% of the previously refused loan applicants were able to pay current loan
- The percentage of applicants whose current loan defaulted but their previous loans were approved is very less which means that these applicants are more likely to pay their current loan in time than the applicants whose previous loans were refused



Income Vs Credit amount
Previous Application Data Analysis
Bivariate Analysis

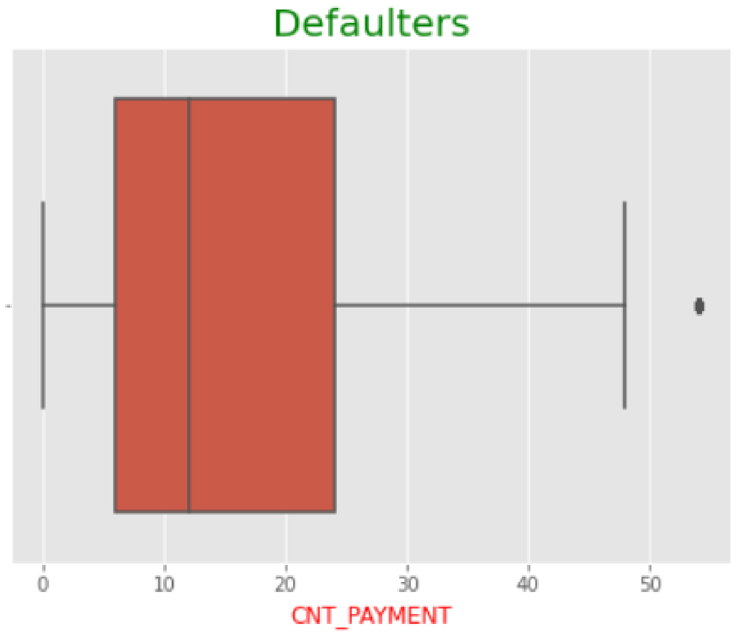
- Loan rejection rate is much lower if the income is higher than 500000. Also, loan requests higher than 200000 has a higher rejection rate

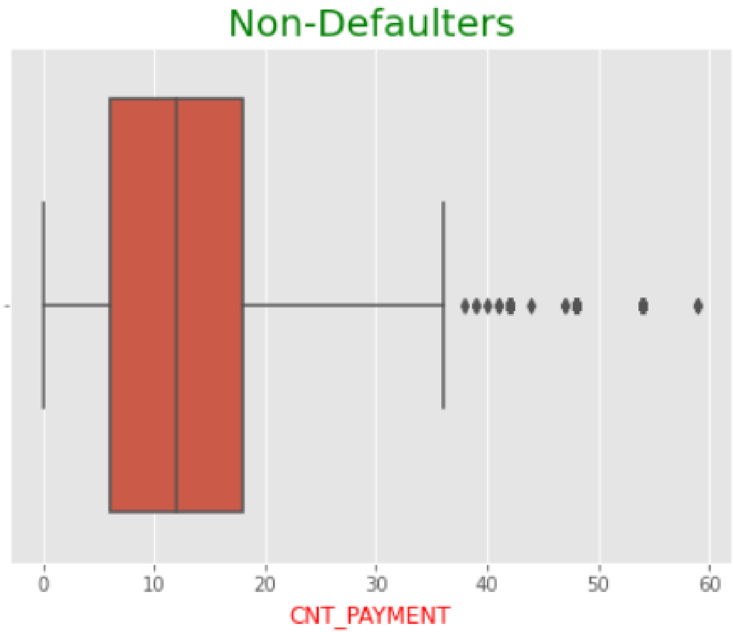

Defaulters


Non-Defaulters

Previous Credit term
Previous Application Data Analysis
Univariate Analysis

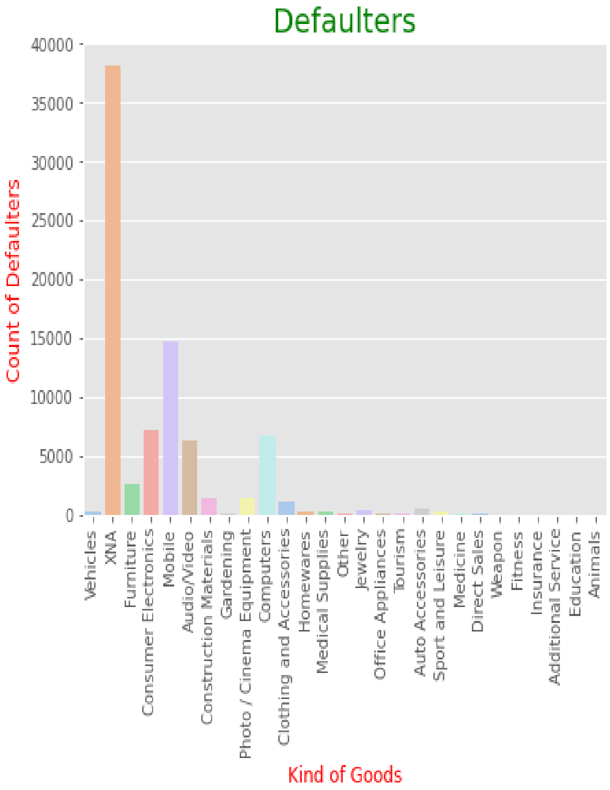
- 75 percentile for defaulters is more than that of non defaulters for CNT_PAYMENT
- The Maximum line for the Defaulters is more than that of the non-defaulters for CNT_PAYMENT
- For those who had lower CNT_PAYMENT in previous application, cases of default are higher

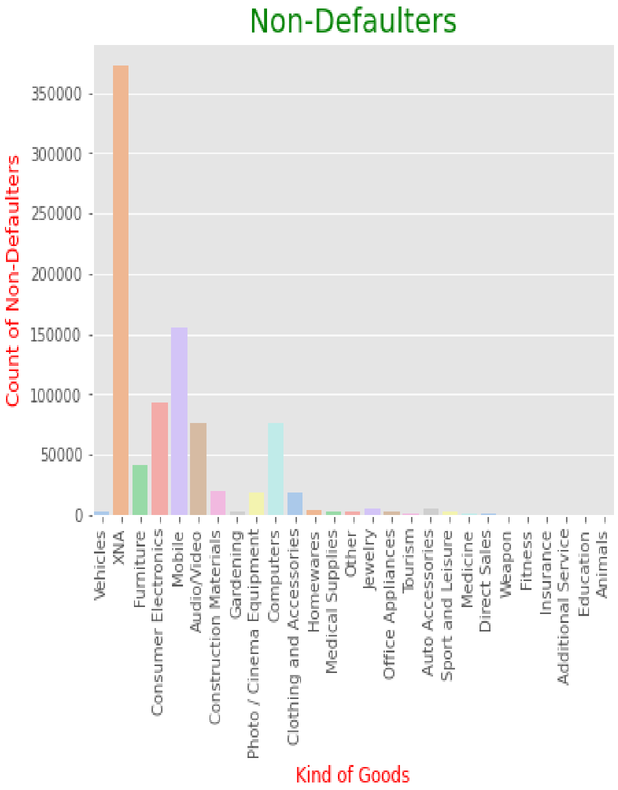


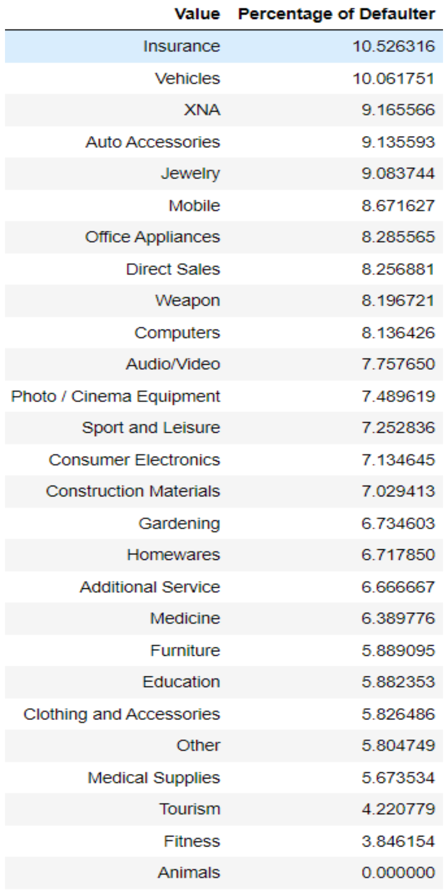


Goods category
Previous Application Data Analysis
Univariate Analysis

- People who had previously applied for Insurance and Vehicles have the highest percentage of defaulting cases



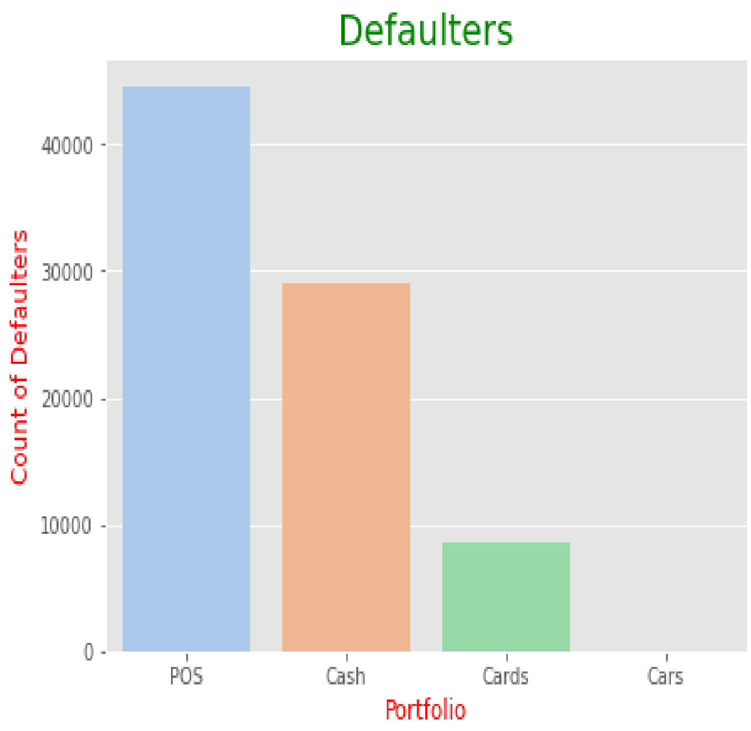


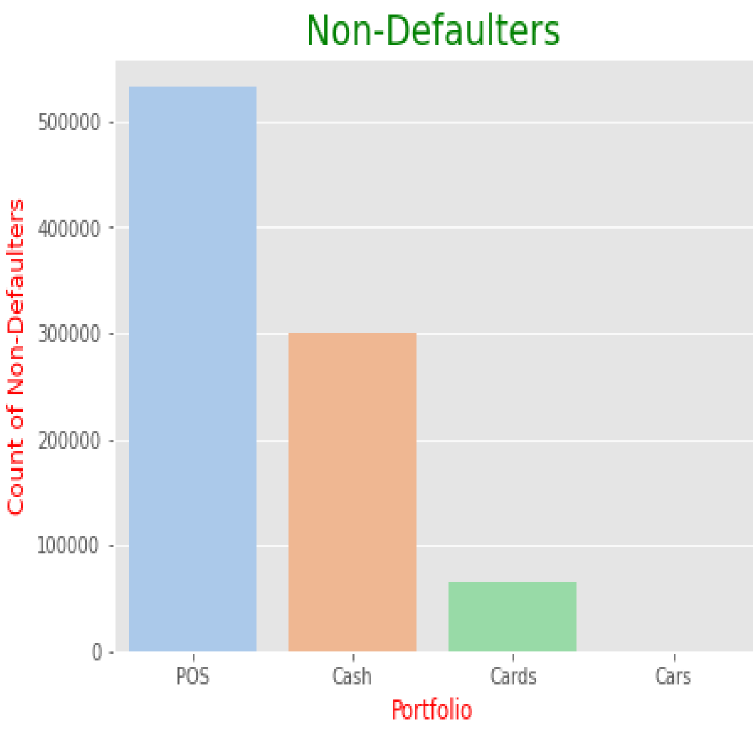


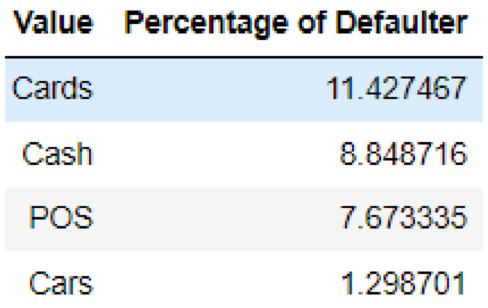
Value	Percentage of Defaulter
Insurance	10.526316
Vehicles	10.061751
XNA	9.165566
Auto Accessories	9.135593
Jewelry	9.083744
Mobile	8.671627
Office Appliances	8.285565
Direct Sales	8.256881
Weapon	8.196721
Computers	8.136426
Audio/Video	7.757650
Photo / Cinema Equipment	7.489619
Sport and Leisure	7.252836
Consumer Electronics	7.134645
Construction Materials	7.029413
Gardening	6.734603
Homewares	6.717850
Additional Service	6.666667
Medicine	6.389776
Furniture	5.889095
Education	5.882353
Clothing and Accessories	5.826486
Other	5.804749
Medical Supplies	5.673534
Tourism	4.220779
Fitness	3.846154
Animals	0.000000

Portfolio
Previous Application Data Analysis
Univariate Analysis

- The defaulter rate is highest where the previous application was for Cards



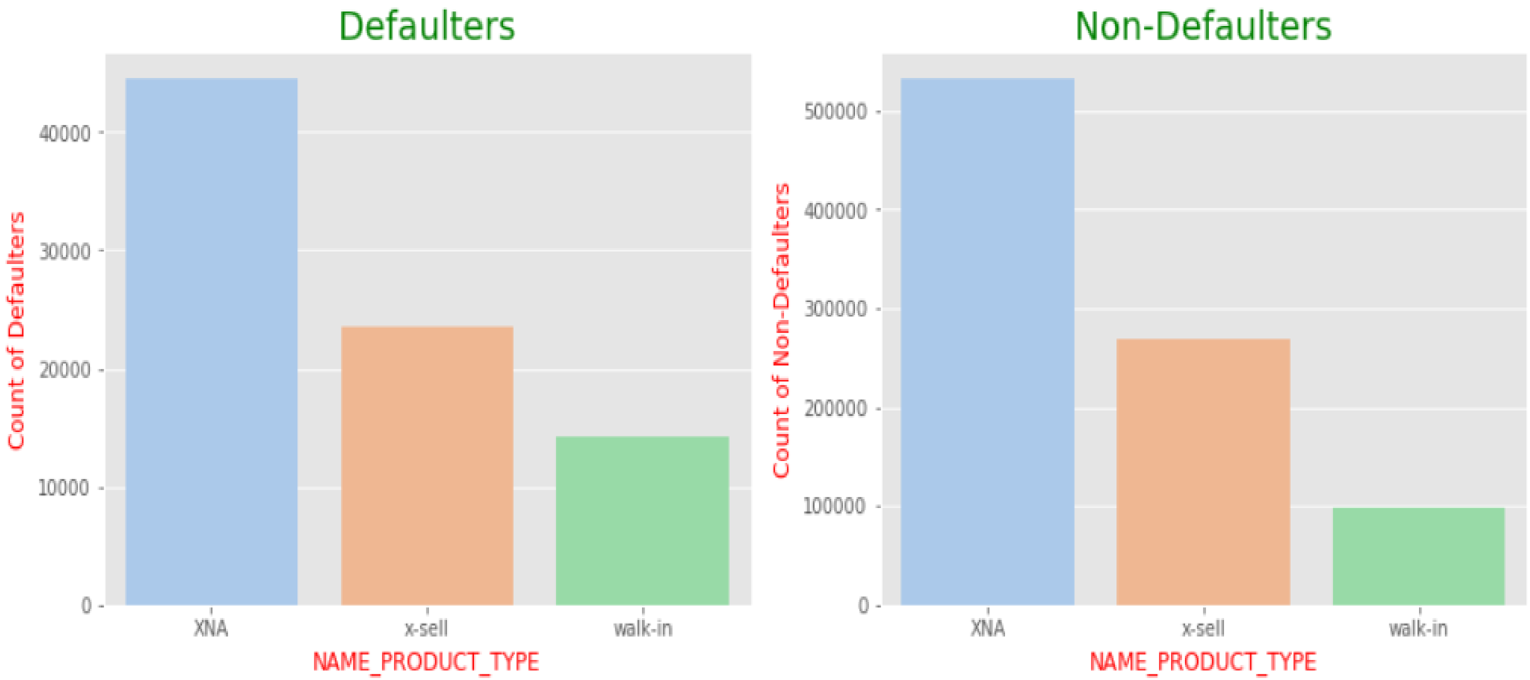


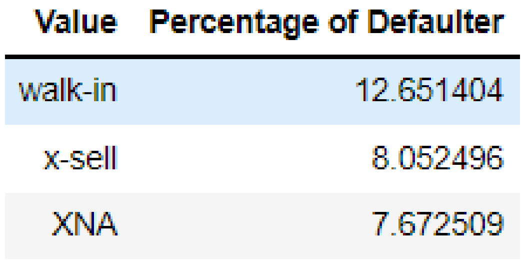


Value	Percentage of Defaulter
Cards	11.427467
Cash	8.848716
POS	7.673335
Cars	1.298701

Walk-in or X-sell
Previous Application Data Analysis
Univariate Analysis

- Out of all the previous applicants, walk-ins have defaulted 12% in current loan

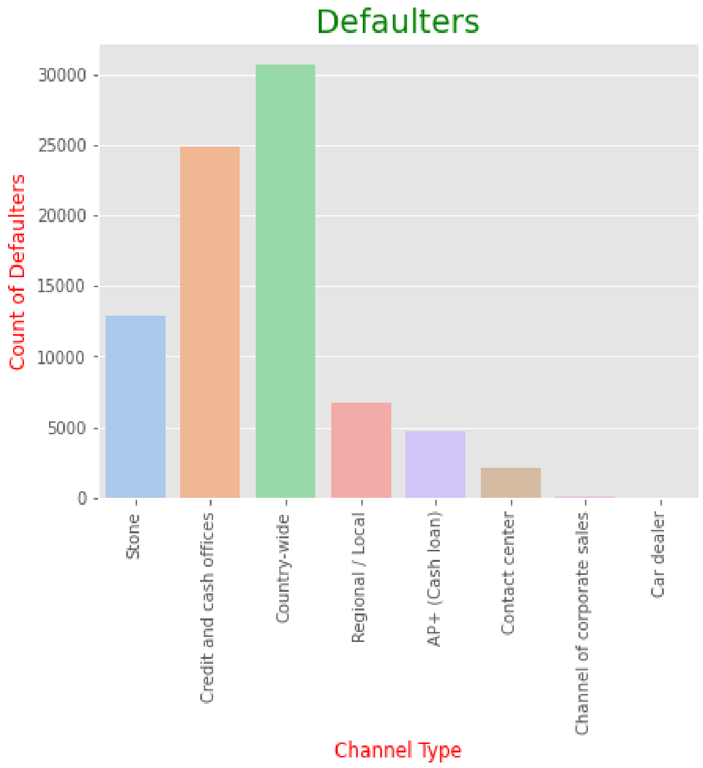


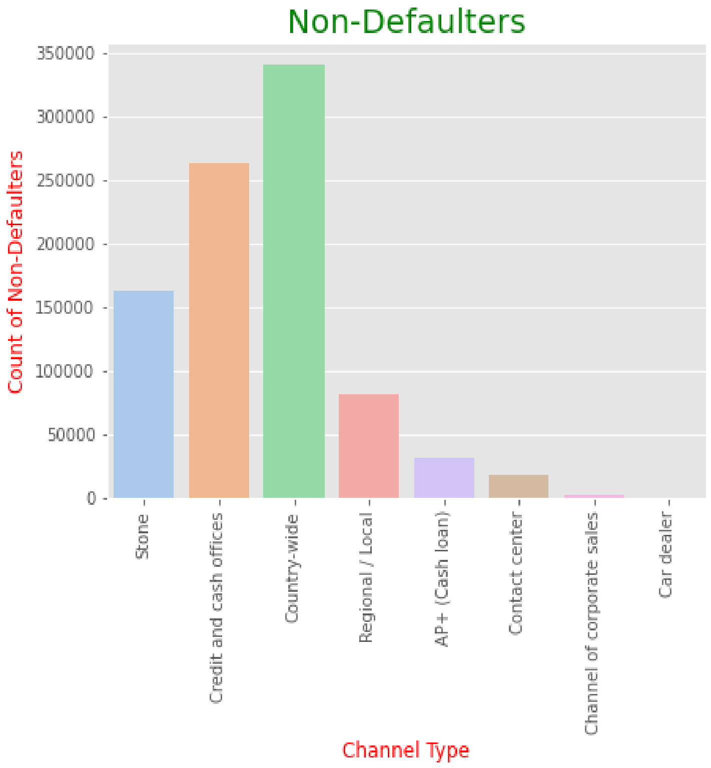


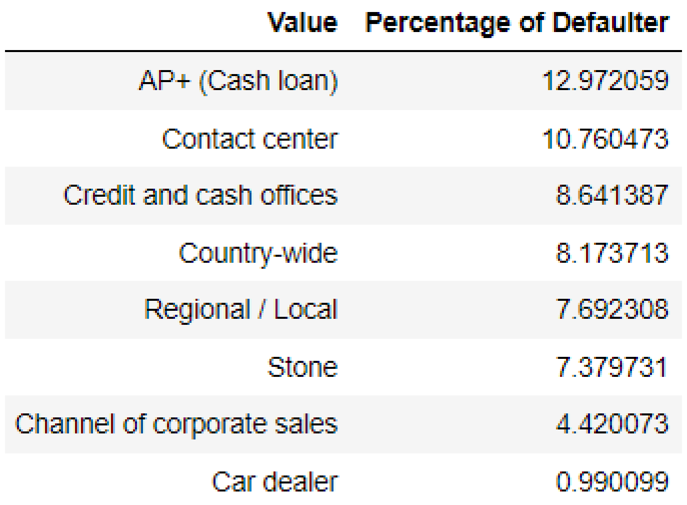
Value	Percentage of Defaulter
walk-in	12.651404
x-sell	8.052496
XNA	7.672509

Channel type
Previous Application Data Analysis
Univariate Analysis

- 13% loan applicants defaulted in the current application who were acquired via channel AP+ (Cash Loan) in their previous application



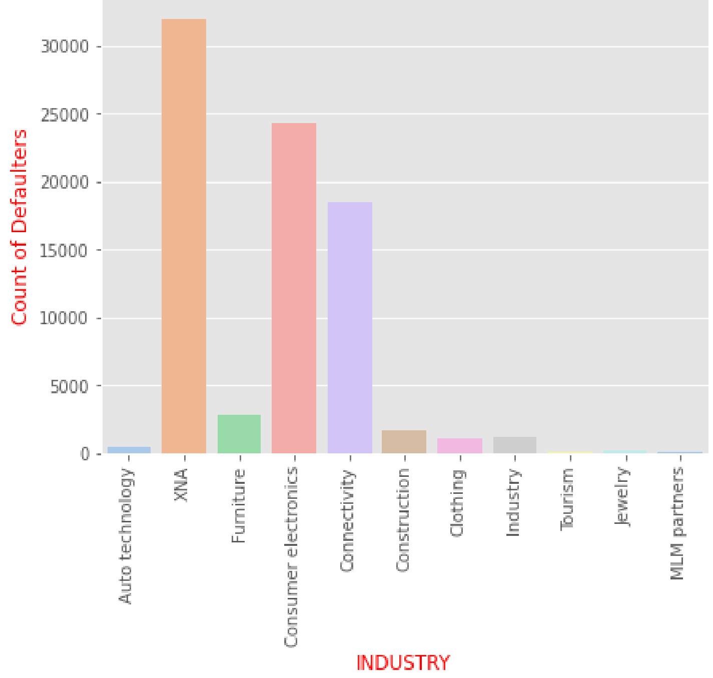




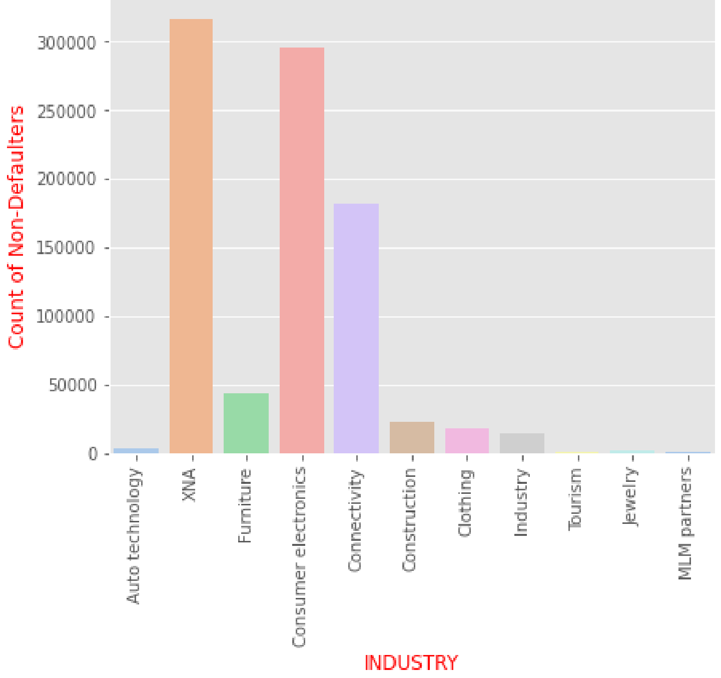
Value	Percentage of Defaulter
AP+ (Cash loan)	12.972059
Contact center	10.760473
Credit and cash offices	8.641387
Country-wide	8.173713
Regional / Local	7.692308
Stone	7.379731
Channel of corporate sales	4.420073
Car dealer	0.990099

Industry
Previous Application Data Analysis
Univariate Analysis

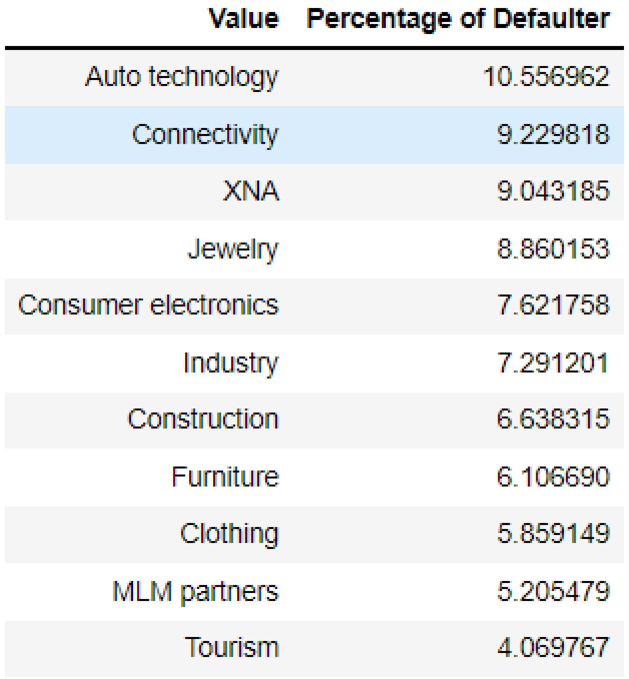
- In seller Industry Auto technology has the highest defaulting rate & Tourism has the lowest number of defaulters



Defaulters



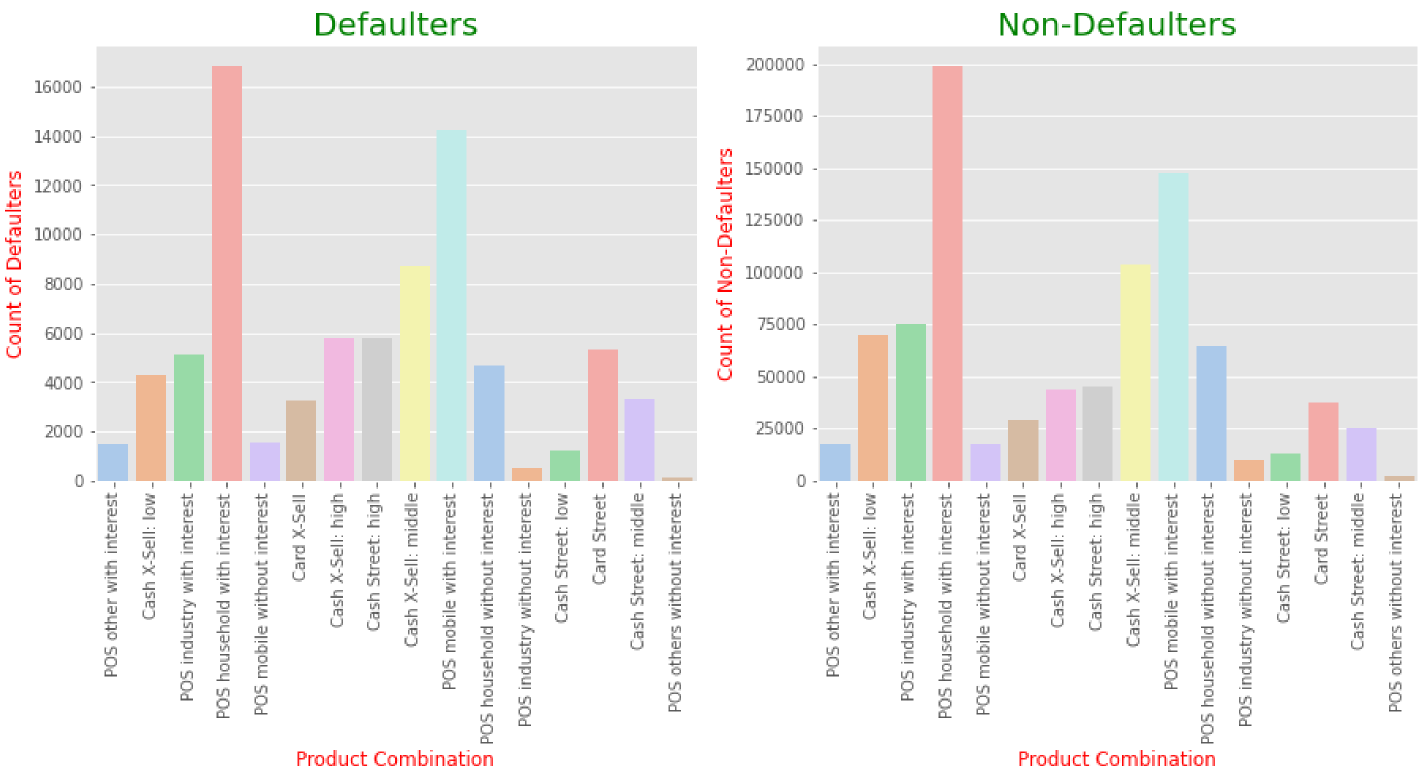
Non-Defaulters

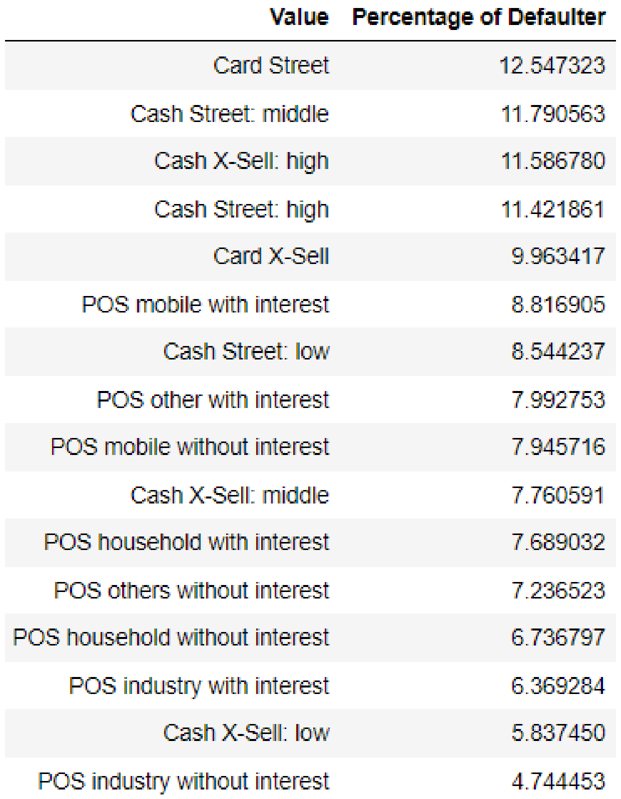


Value	Percentage of Defaulter
Auto technology	10.556962
Connectivity	9.229818
XNA	9.043185
Jewelry	8.860153
Consumer electronics	7.621758
Industry	7.291201
Construction	6.638315
Furniture	6.106690
Clothing	5.859149
MLM partners	5.205479
Tourism	4.069767

Product combination
Previous Application Data Analysis
Univariate Analysis

- Highest percentage of defaulting cases is for Card Street

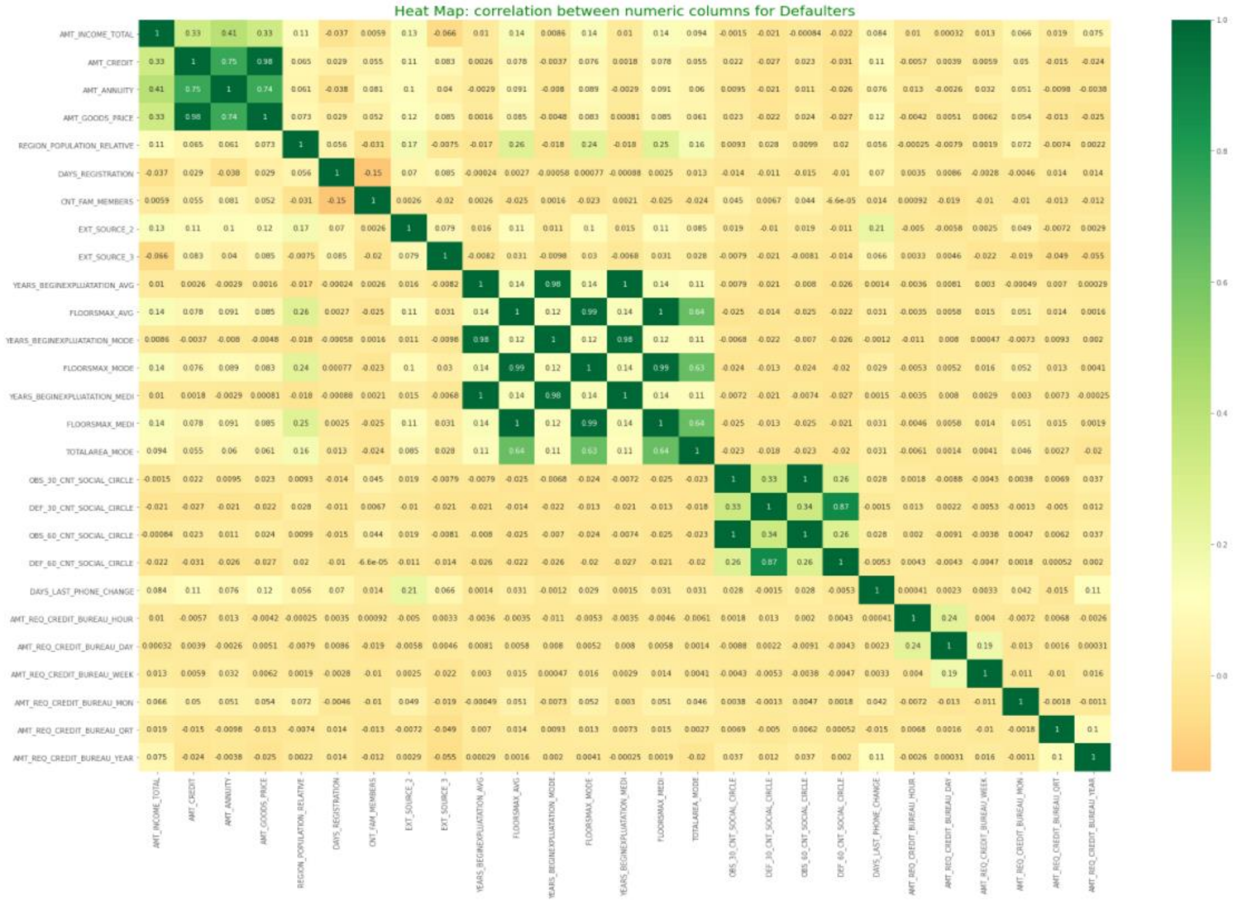




Value	Percentage of Defaulter
Card Street	12.547323
Cash Street: middle	11.790563
Cash X-Sell: high	11.586780
Cash Street: high	11.421861
Card X-Sell	9.963417
POS mobile with interest	8.816905
Cash Street: low	8.544237
POS other with interest	7.992753
POS mobile without interest	7.945716
Cash X-Sell: middle	7.760591
POS household with interest	7.689032
POS others without interest	7.236523
POS household without interest	6.736797
POS industry with interest	6.369284
Cash X-Sell: low	5.837450
POS industry without interest	4.744453

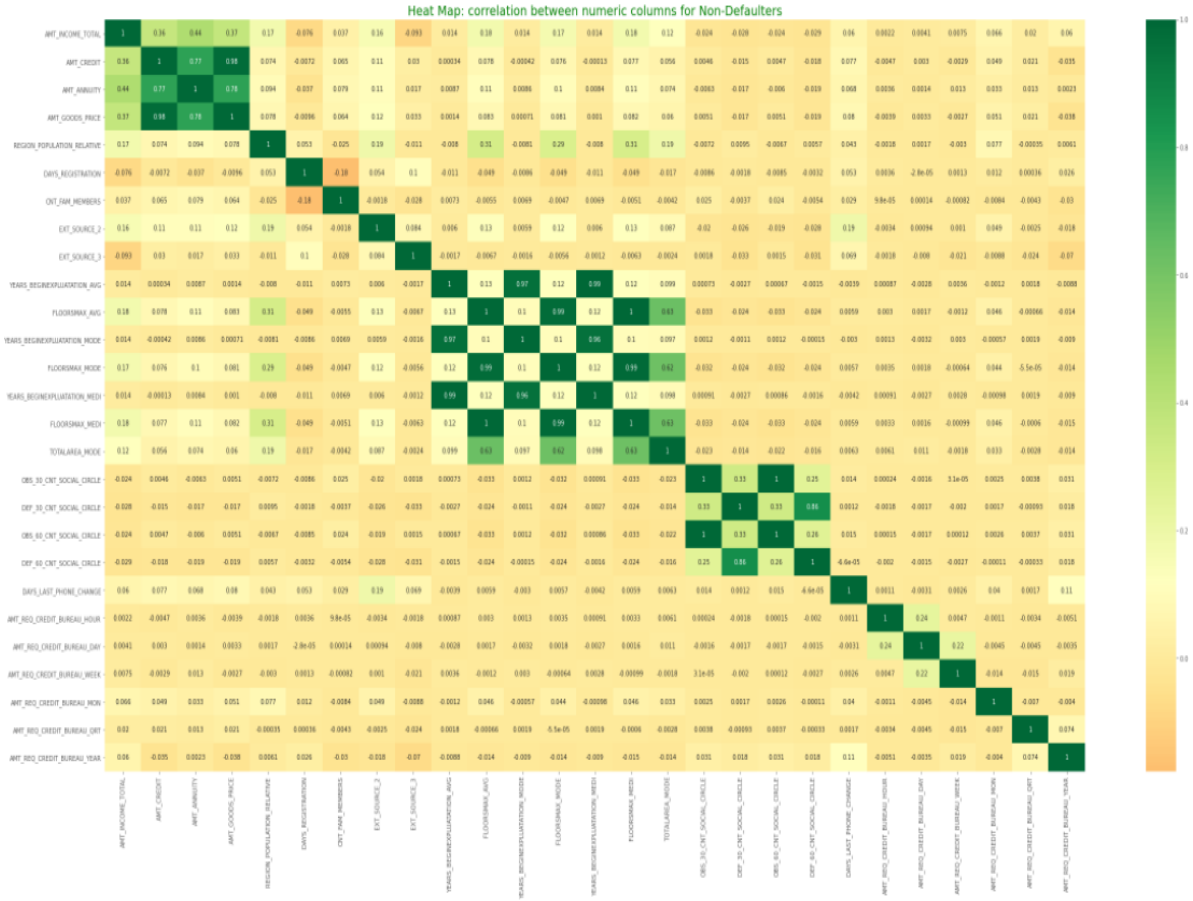
Application Data Correlation

- Top 10 Correlation between numeric columns for Defaulters:
  1. AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_YEAR
  2. OBS_60_CNT_SOCIAL_CIRCLE, OBS_30_CNT_SOCIAL_CIRCLE
  3. FLOORSMAX_MEDI, FLOORSMAX_AVG
  4. YEARS_BEGINEXPLUATATATION_AVG, YEARS_BEGINEXPLUATATATION_MEDI
  5. FLOORSMAX_MODE, FLOORSMAX_MEDI
  6. FLOORSMAX_AVG, FLOORSMAX_MODE
  7. AMT_CREDIT, AMT_GOODS_PRICE
  8. YEARS_BEGINEXPLUATATATION_MODE, YEARS_BEGINEXPLUATATATION_AVG
  9. YEARS_BEGINEXPLUATATATION_MEDI, YEARS_BEGINEXPLUATATATION_MODE
  10. REGION_RATING_CLIENT_W_CITY, REGION_RATING_CLIENT



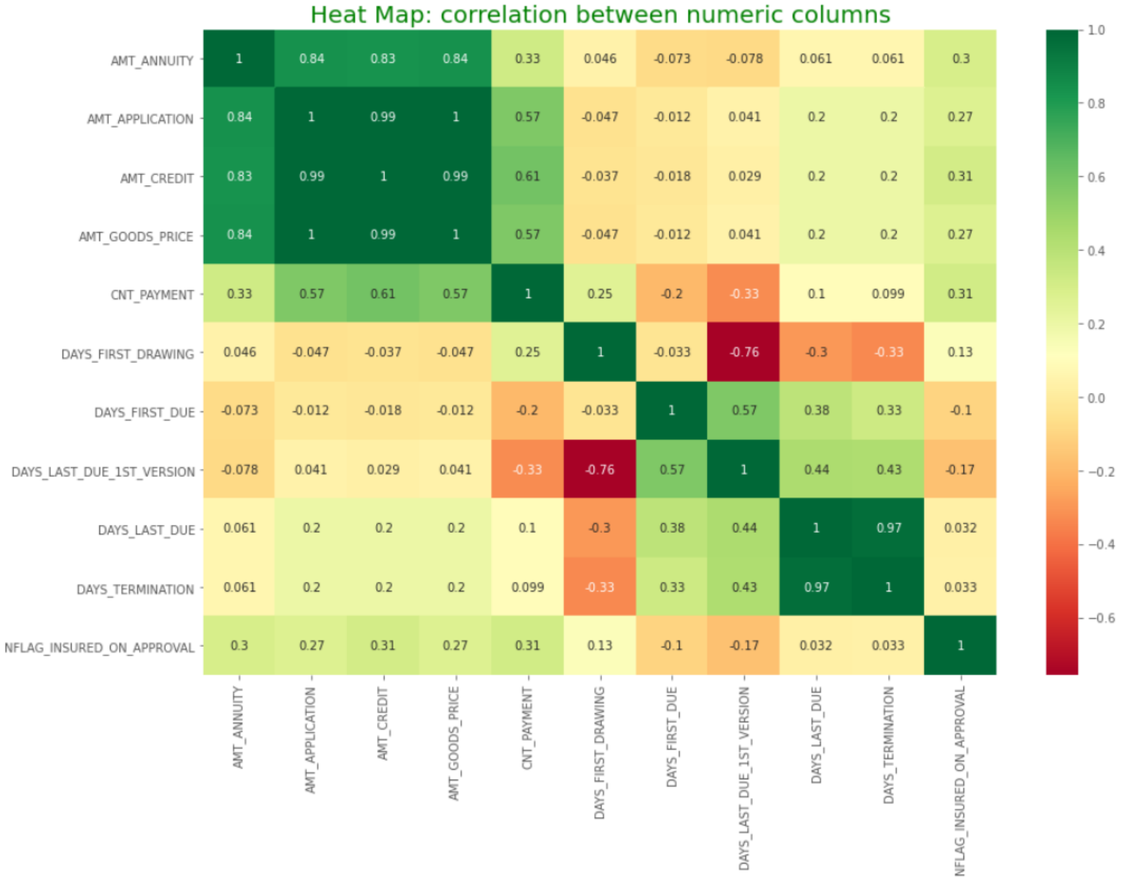
Application Data Correlation

- Top 10 Correlation between numeric columns for Non-Defaulters:
  1. AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_YEAR
  2. OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE
  3. FLOORSMAX_AVG, FLOORSMAX_MEDI
  4. YEARS_BEGINEXPLUATATATION_MEDI, YEARS_BEGINEXPLUATATATION_AVG
  5. FLOORSMAX_MEDI, FLOORSMAX_MODE
  6. FLOORSMAX_AVG, FLOORSMAX_MODE
  7. AMT_CREDIT, AMT_GOODS_PRICE
  8. YEARS_BEGINEXPLUATATATION_MODE, YEARS_BEGINEXPLUATATATION_AVG
  9. YEARS_BEGINEXPLUATATATION_MEDI, YEARS_BEGINEXPLUATATATION_MODE
  10. REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY



Previous Application Correlation

- Correlation between numeric columns:
  - 'AMT_ANNUITY', 'AMT_APPLICATION', 'AMT_CREDIT' & 'AMT_GOODS_PRICE' are highly correlated
  - 'DAYS_TERMINATION' & 'DAYS_LAST_DUE' are also highly correlated
  - 'DAYS_LAST_DUE_1st_VERSION' & 'DAYS_FIRST_DRAWING' have a high negative correlation



Heat Map: correlation between numeric columns

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	CNT_PAYMENT	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	NFLAG_INSURED_ON_APPROVAL
AMT_ANNUITY	1	0.84	0.83	0.84	0.33	0.046	-0.073	-0.078	0.061	0.061	0.3
AMT_APPLICATION	0.84	1	0.99	1	0.57	-0.047	-0.012	0.041	0.2	0.2	0.27
AMT_CREDIT	0.83	0.99	1	0.99	0.61	-0.037	-0.018	0.029	0.2	0.2	0.31
AMT_GOODS_PRICE	0.84	1	0.99	1	0.57	-0.047	-0.012	0.041	0.2	0.2	0.27
CNT_PAYMENT	0.33	0.57	0.61	0.57	1	0.25	-0.2	-0.33	0.1	0.099	0.31
DAYS_FIRST_DRAWING	0.046	-0.047	-0.037	-0.047	0.25	1	-0.033	-0.76	-0.3	-0.33	0.13
DAYS_FIRST_DUE	-0.073	-0.012	-0.018	-0.012	-0.2	-0.033	1	0.57	0.38	0.33	-0.1
DAYS_LAST_DUE_1ST_VERSION	-0.078	0.041	0.029	0.041	-0.33	-0.76	0.57	1	0.44	0.43	-0.17
DAYS_LAST_DUE	0.061	0.2	0.2	0.2	0.1	-0.3	0.38	0.44	1	0.97	0.032
DAYS_TERMINATION	0.061	0.2	0.2	0.2	0.099	-0.33	0.33	0.43	0.97	1	0.033
NFLAG_INSURED_ON_APPROVAL	0.3	0.27	0.31	0.27	0.31	0.13	-0.1	-0.17	0.032	0.033	1

Thank You!