CREDIT EDA CASE STUDY

BY

1. Lakshmi Ramya
2. Hemendra Sai

INTRODUCTION TO APPROACH

❑ Our approach to the case study involved 4 steps.

a) Data cleaning

b) Univariate Analysis

c) Segmented Univariate Analysis

d) Bivariate Analysis

❑ DATA CLEANING:

We have selected 10 columns which we thought will be relevant to our analysis. Following are the columns that we have selected

a) SK_ID_CURR

b) TARGET

c) NAME_CONTRACT_TYPE

d) CODE_GENDER

e) FLAG_OWN_REALTY

f) AMT_INCOME_TOTAL

g) AMT_CREDIT

h) AMT_ANNUITY

i) AMT_GOODS_PRICE

j) DAYS_EMPLOYED

- Out of these 10 variables, 5 are continuous variables(marked in red),4 are categorical variables(marked in blue) and the other one is the application id.
- Our data thus consisted of 307511 rows and 10 columns.
- We have initially checked for missing value in our data. These are results when we checked for missing values. Missing values for each variable is marked in red.

a. SK_ID_CURR- 0

b. TARGET - 0

c. NAME_CONTRACT_TYPE - 0

d. CODE_GENDER - 0

e. FLAG_OWN_REALTY - 0

f. AMT_INCOME_TOTAL - 0

g. AMT_CREDIT - 0

h. AMT_ANNUITY - 12

i. AMT_GOODS_PRICE - 278

j. DAYS_EMPLOYED - 0

- When compared with 307511 rows, only 12 and 278 rows having missing values for the variables ‘AMT_ANNUITY’ & ‘AMT_GOODS_PRICE’.
- Therefore, we have treated them individually.

TREATING MISSING VALUES 'AMT_ANNUITY':

- Annuity is the amount of repayment by the borrower per year.
- Most of the loans have a similar tenure and therefore we have assumed annuity as a fraction of credit to be fairly constant.
- We created new variable called 'amortization' = "AMT_AANUITY/AMT_CREDIT"
- When we have tested this variable, we found that standard deviation of this variable across all the rows is very small(0.022). Therefore, we have assumed mean value of amortization to fairly represent the value of amortization in the missing rows.
- Using this logic, we have calculated the missing values for AMT_ANNUITY by using (AMT_CREDIT*mean amortization).
- This way we have treated all the 12 rows which had the missing values.

TREATING MISSING VALUES 'AMT_GOODS_PRICE':

- AMT_GOODS_PRICE is defined the price of the asset for which the loan is being taken.
- Loan to value[Loan To Value] is amount of loan to the goods price.
- We created new variable called 'LTV' = "AMT_CREDIT/AMT_GOODS_PRICE"
- When we have tested this variable, we found that standard deviation of this variable across all the rows is very small(0.12). Therefore, we have assumed mean value of LTV to fairly represent the value of LTV in the missing rows.
- Using this logic, we have calculated the missing values for AMT_GOODS_PRICE by using (AMT_CREDIT/LTV).
- This way we have treated all the 278 rows which had the missing values.

Thus, we have tereated all the missing values in the data set.

MERGING PREVIOUS APPLICATION:

- We have merged previous application data. However, we have used only one variable from the previous application data for our analysis.
- We have selected Status of previous loan application as one of the variable that may affect the current application. Hence, following 3 columns are selected from the previous application data and are merged.

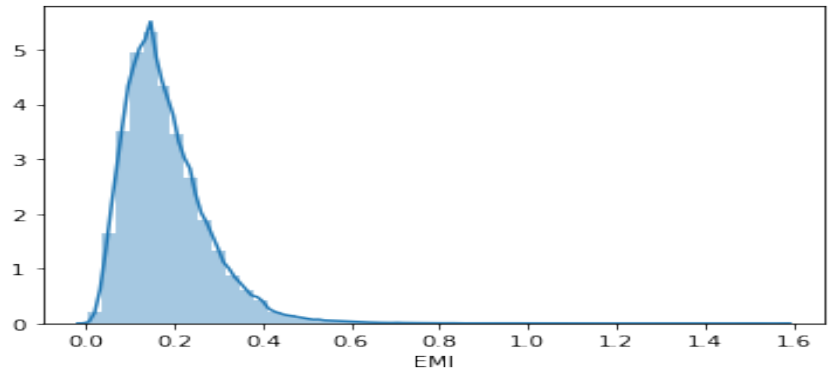
a. SK_ID_PREV

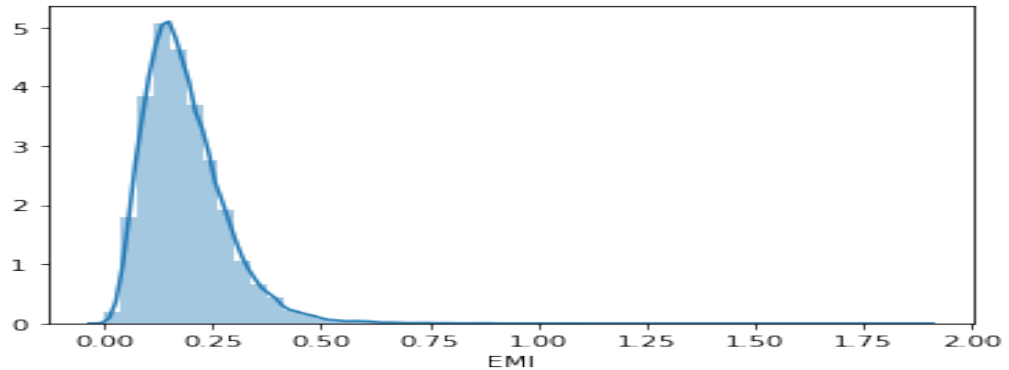
b. SK_ID_CURR

c. NAME_CONTRACT_STATUS

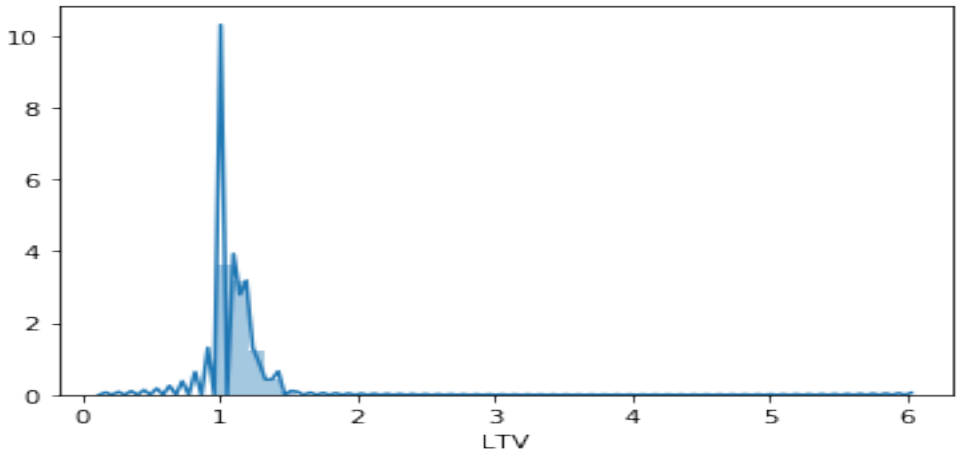
❑ UNIVARIATE ANALYSIS :

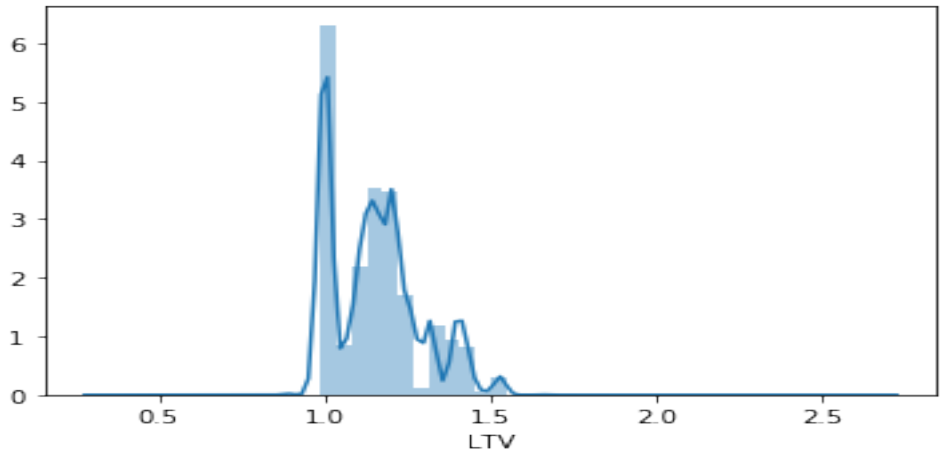
- We have plotted histograms and box plots for various numerical and categorical variables in jupyter notebook. Each of our observations have been mentioned in the jupyter notebook for every plot.
- Although we could not get any valuable insights on how much each variable affects default rates(mainly because the plots were mostly similar for Target 0 & Target 1), we did it to get feel for the data.
- We then tried to plot histograms for the variables created by us using the variables in the dataset.
- We have created a new column called EMI which is obtained by AMT_ANNUITY/AMT_IMCOME_TOTAL. Since every individual will have a limitation in what fraction of income, he can use to repay the loan, we thought that will throw out some pattern between Target 1 & Target 0. Following are the plots. Plot on the left side is for Target 0 and plot on the right side is Target 1.



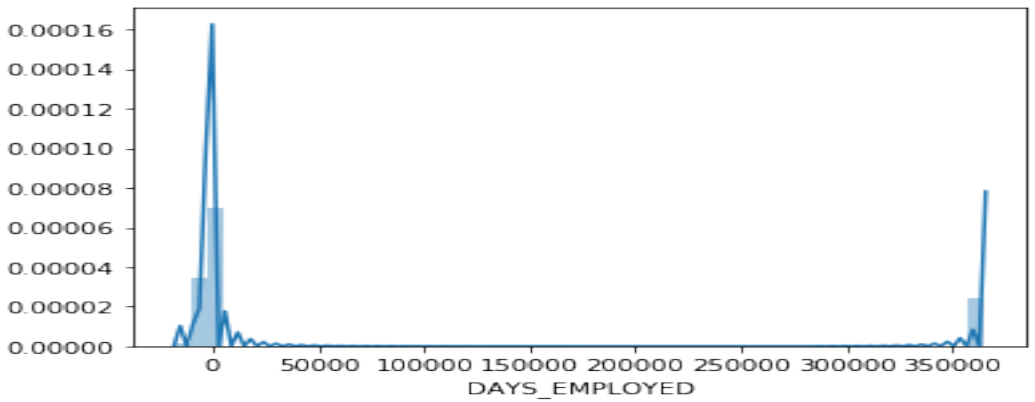


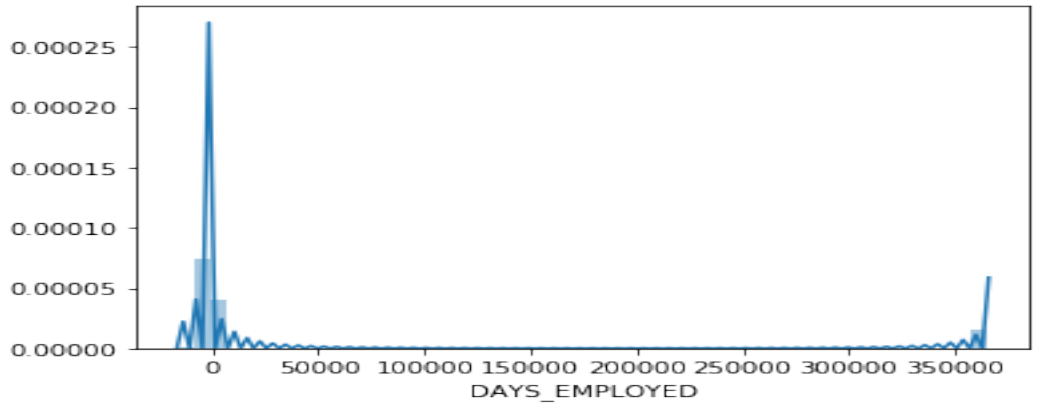
- From the two plots we can make an observation that individuals from Target 1 are paying 25% of their incomes as annuity as repayment whereas individuals from Target 0 are paying 20% of their incomes as repayment. This may be the reason for payment difficulties in target 1.
- Second plot is made for “LTV” which as explained earlier is the amount of loan as a fraction of goods price. Following are results of our plots. Left one is for Target 0 and right plot is for Target 1. It is usually preferred to keep the LTV as low as possible inorder to secure the loan well. We observed that LTV for target 0 is mostly concentrated around 1 or less than 1 while LTV of Target 1 has been beyond one for the most part. This suggests that people who have borrowed more compared to the value of goods are prone to default.



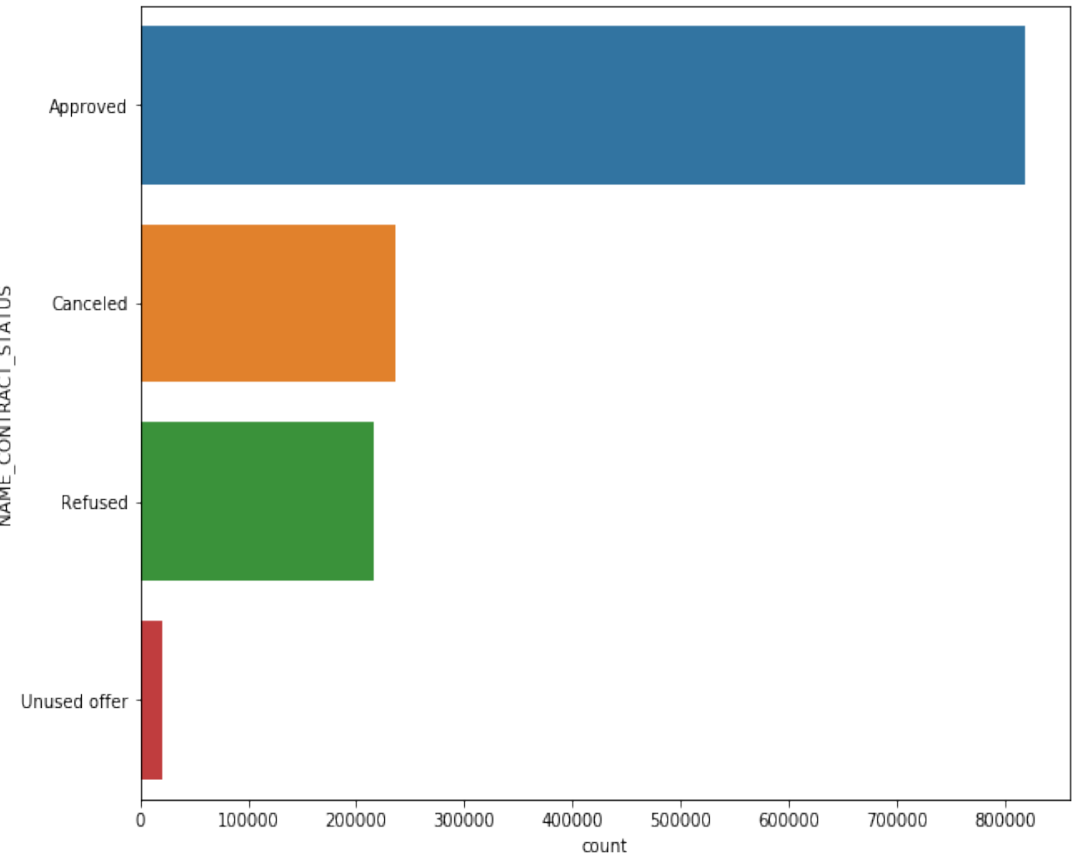


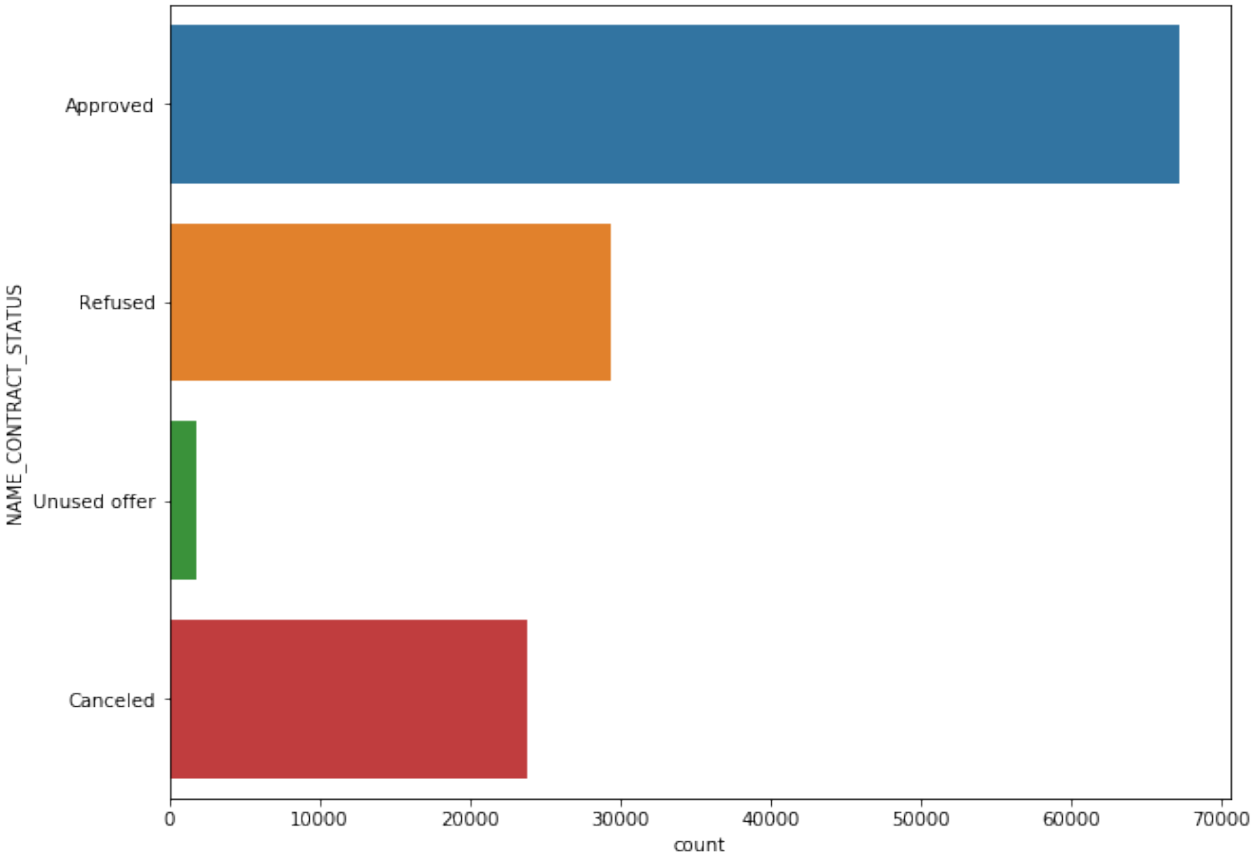
- Third plot has been made made for the days of employment and following are the results. Left one is for Target 0 and right one is for Target1.



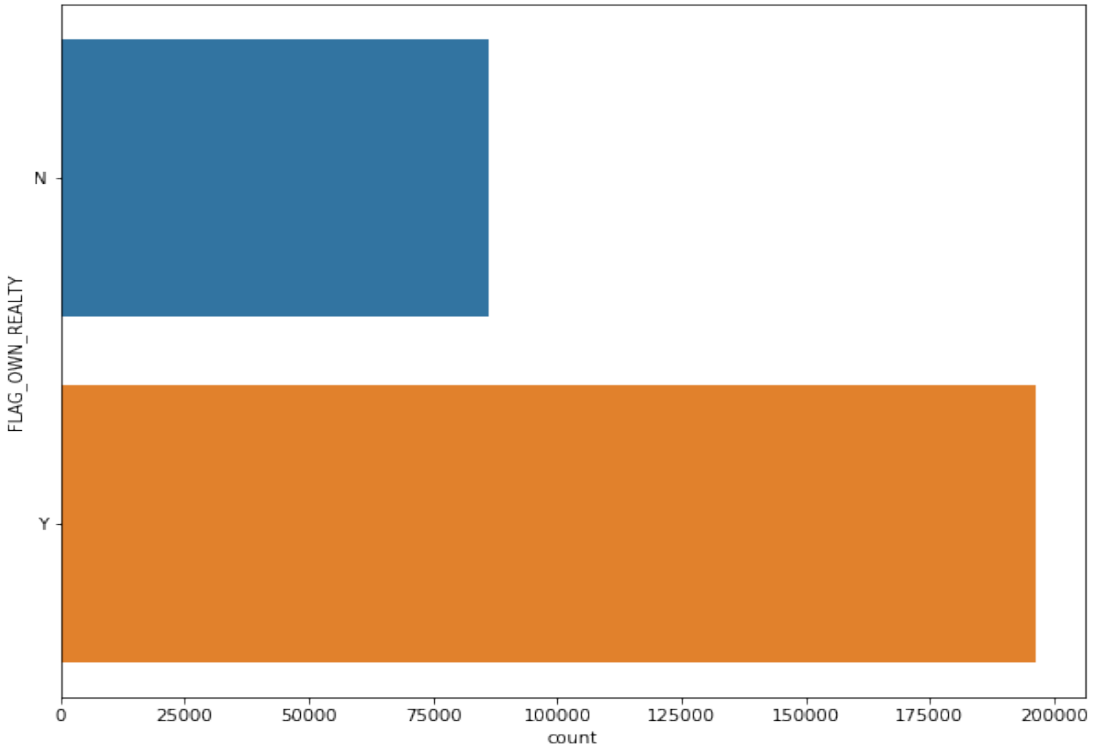


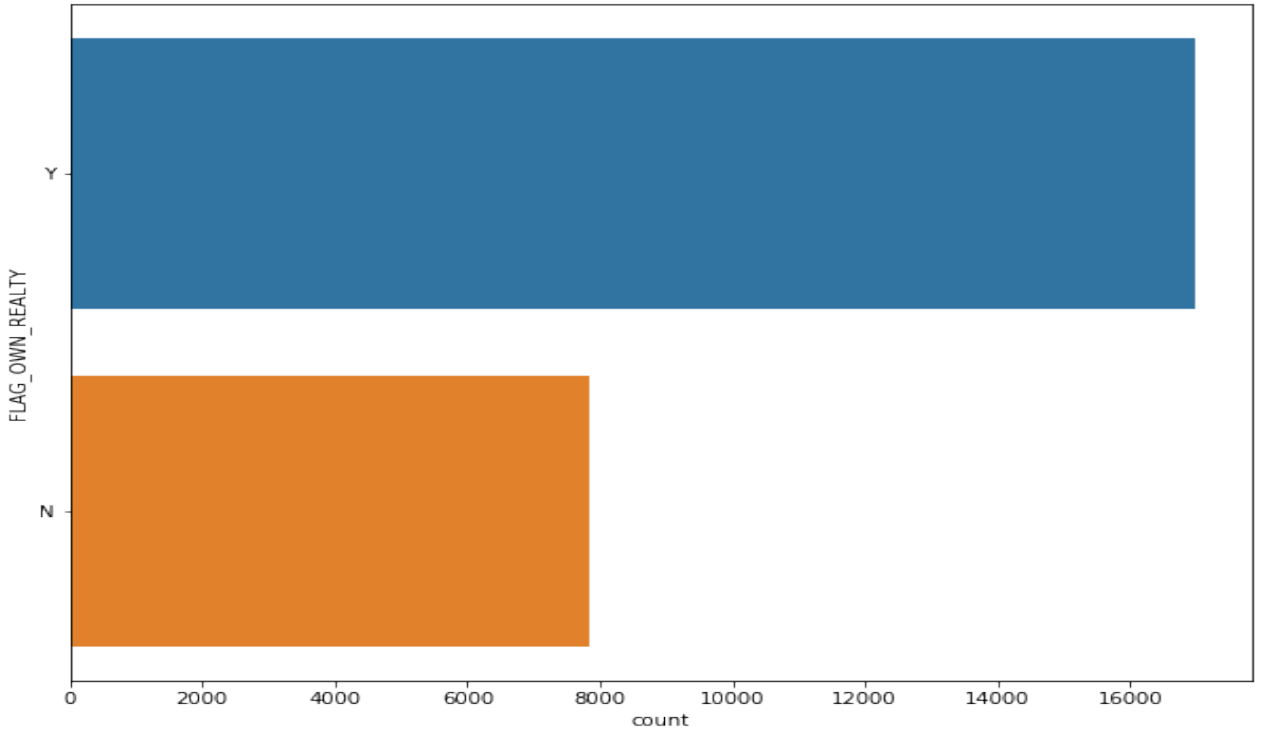
- Although it is not very concrete, people with high number of days employed are in Target 0 than in Target 1 which can be observed by the higher spike toward the right. This also suggest, no.of days employed also is affecting the default rates. Higher the number of days employed, the safer it is for the bank to lend.
- These are the observations made from univariate analysis of continuous variables.
- For Univariate analysis of categorical variables, we have initially selected the status of previous application as one metric. After merging the previous application with the current application, we have segmented the data frame into two parts of Target 1 and Target 0. Following are the plots we have observed. Plot on the left side belong to target 0 and plot on the right side belongs to target 1. We can observe that higher percentage of people from target 1 have been rejected for the loan in previous application. This throws a pattern between the rejection status in previous application and default rate in current application.



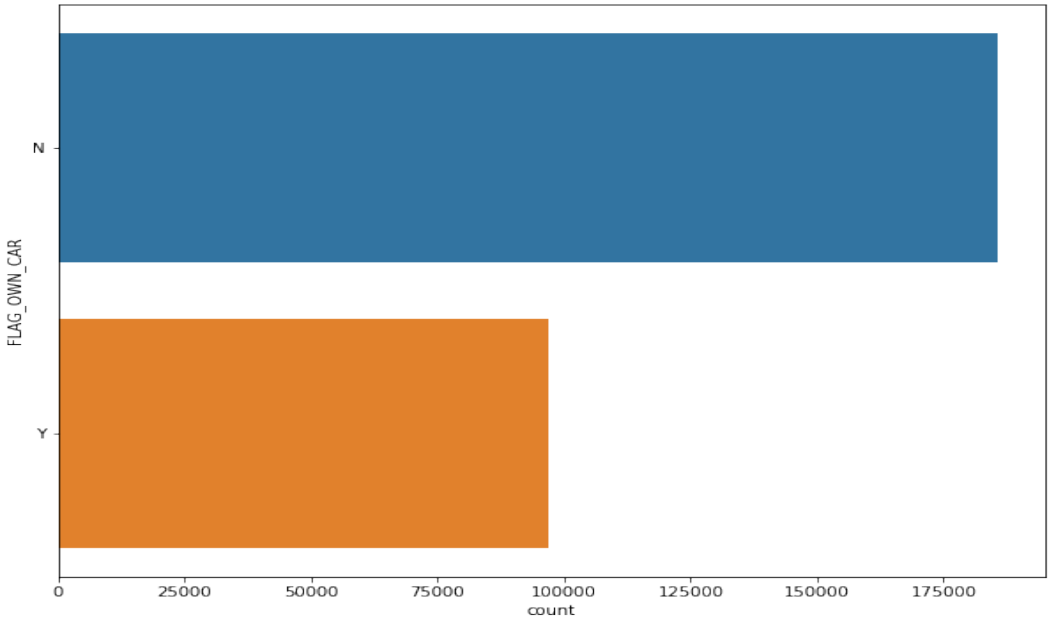


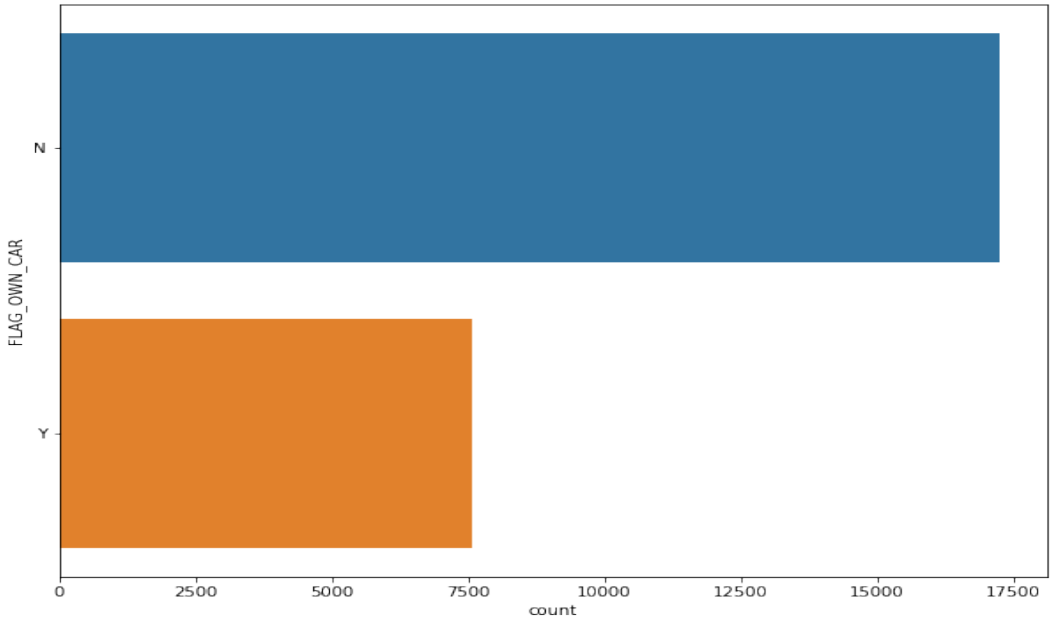
- Second observation from the categorical variables is made from the ownership of the house. Following are the results of the plots. Plot on the left belongs to target0 and plot on the right belongs to Target 1.
- Although we cannot make a concrete observation from this data, but we can observe that percentage of number of people who do not own a house is slightly lower in Target 0 which suggest people with house have less chance of defaulting.





- Third observation from the categorical variables is made from the ownership of the car. Following are the results of the plots. Plot on the left belongs to target0 and plot on the right belongs to Target 1.
- Although we cannot make a concrete observation from this data, but we can observe that percentage of number of people who own a car is slightly higher in Target 0 which suggest people with car have less chance of defaulting.





- With this we conclude our univariate analysis.

❑BIVARIATE ANALYSIS:

- We have also performed bivariate analysis of continuous variables. Following are the results of analysis. Table on the left shows correlation for Target 1 and table on the right shows correlation for Target 0.

Var1	Var2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.983089
AMT_GOODS_PRICE	AMT_ANNUITY	0.752857
AMT_ANNUITY	AMT_CREDIT	0.752195
DAYS_EMPLOYED	AMT_ANNUITY	0.082552
AMT_ANNUITY	AMT_INCOME_TOTAL	0.046421
AMT_CREDIT	AMT_INCOME_TOTAL	0.038131
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.037631
DAYS_EMPLOYED	AMT_INCOME_TOTAL	0.014979
DAYS_EMPLOYED	AMT_GOODS_PRICE	0.003626
DAYS_EMPLOYED	AMT_CREDIT	0.000968

Var1	Var2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.987239
AMT_GOODS_PRICE	AMT_ANNUITY	0.776845
AMT_ANNUITY	AMT_CREDIT	0.771315
AMT_ANNUITY	AMT_INCOME_TOTAL	0.418959
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349691
AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
DAYS_EMPLOYED	AMT_INCOME_TOTAL	0.141250
DAYS_EMPLOYED	AMT_ANNUITY	0.106425
DAYS_EMPLOYED	AMT_CREDIT	0.072515
DAYS_EMPLOYED	AMT_GOODS_PRICE	0.070881

OBSERVATIONS FROM THE CORRELATION MATRIX:

- Top 3 correlations have between the same variables in both Target1 and Target0 i.e (goods price, credit),(goods price, annuity),(Annuity, credit). This seems intuitive.
- But as move downwards on the correlation matrix, correlation between the variables decreases drastically for Target 1 while correlation almost remain stable for Target 0.
- This suggests a more comprehensive analysis conducted while assessing the credit decision in Target 0 than target 1.
- Let's take an example of 4 row in the correlation matrix. Correlation for target1 is between days employed and annuity which is around 0.08 whereas for Target 0 it is between annuity and income which is around 0.46 suggesting a clear correlation between annuity and income.
- Similar type of correlation differences are observed are observed in sub-sequent rows of correlation matrix suggesting a better assessment methods used while assessing the credit decision in target0 than in target1.
- With this, we conclude our bivariate analysis as well.