Credit EDA Case Study

Manoj Patil
Rittik Saha

Data cleaning and imputation

- Total no of columns with missing values more than 50% : 41
    - Action taken: Deleted columns from Dataset

- Total no of columns with missing values less than 15% : 16
    - Action taken: Data imputation

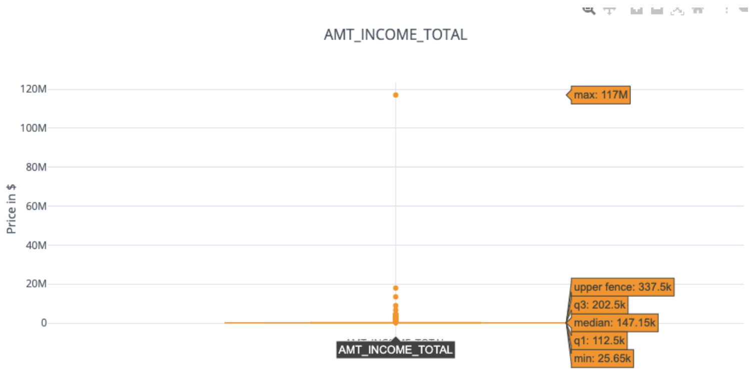
- Total number of columns we have imputed data: 16
- Data imputation methods used
    - For categorical variables: Used value with highest frequency
    - For numeric variables: Used mean/median value as per the data

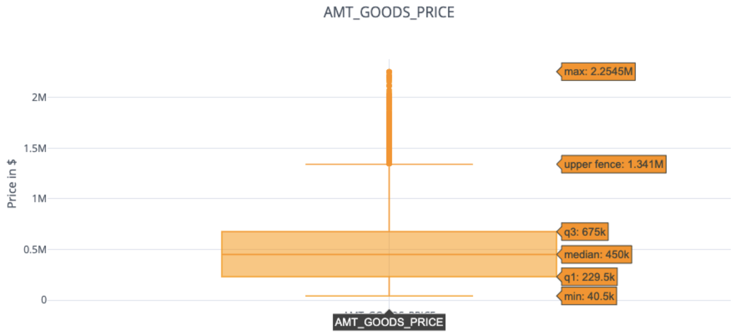
- Data cleaning and formatting
    - Changed data type from float to integer for discrete variables
    - For day fields changed negatives values to absolute values
    - For employment days data changed default values to null values

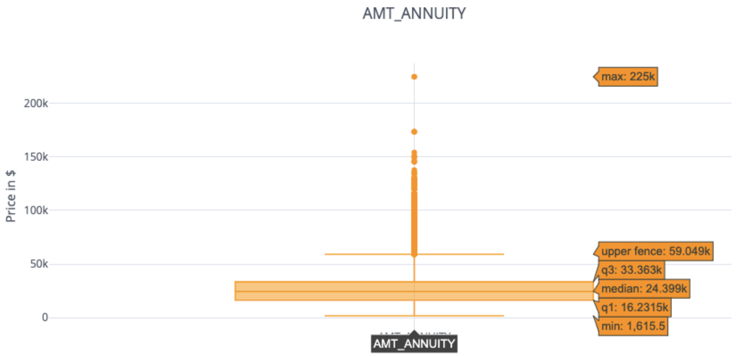
Outlier detection & analysis

- Outlier analysis: Outlier analysis done for below columns
  - ✓ AMT_INCOME_TOTAL
    - ○ Lower bound for outliers : -22500.0 & Upper bound : 337500.0
    - ○ Total no of outliers for column AMT_INCOME_TOTAL : 14035
  - ✓ AMT_ANNUITY
    - ○ Lower bound for outliers : -9465.75 & Upper bound : 59060.25
    - ○ Total no of outliers for column AMT_ANNUITY : 6250
  - ✓ AMT_GOODS_PRICE
    - ○ Lower bound for outliers : -438750.0 & Upper bound : 1343250.0
    - ○ Total no of outliers for column AMT_GOODS_PRICE : 8797
  - ✓ CNT_FAM_MEMBERS
    - ○ Lower bound for outliers : 0.0 & Upper bound : 4
    - ○ Total no of outliers for column CNT_FAM_MEMBERS : 3579

Action Taken: Deleted the outlier records from the dataset

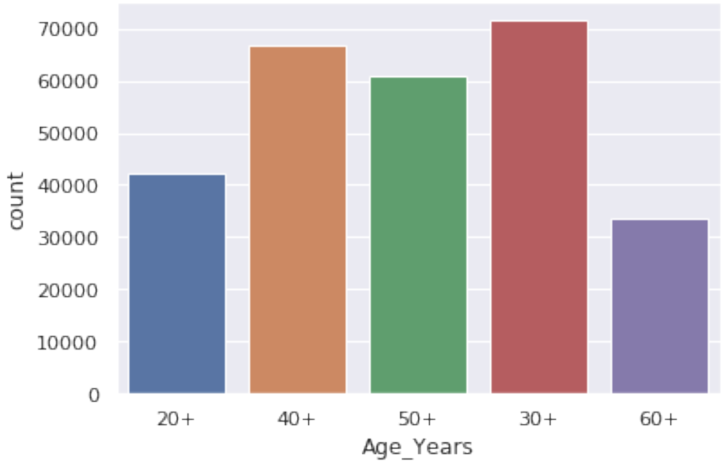

AMT_INCOME_TOTAL

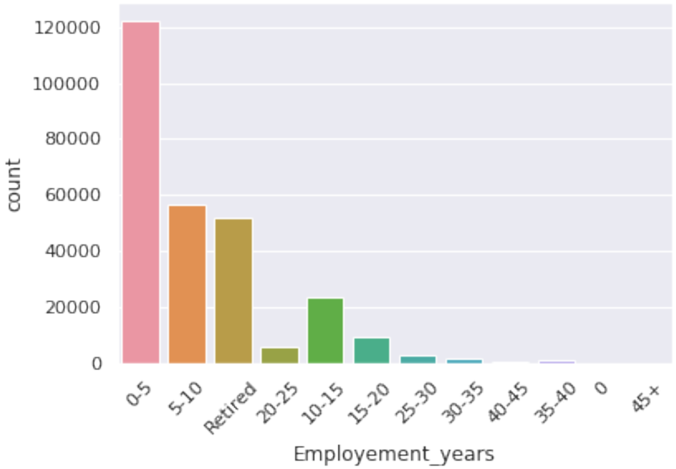

AMT_GOODS_PRICE


AMT_ANNUITY

Binning (Numeric to Categorical)

- For Age and employment days data we have done binning to convert that data into categorical data for further analysis

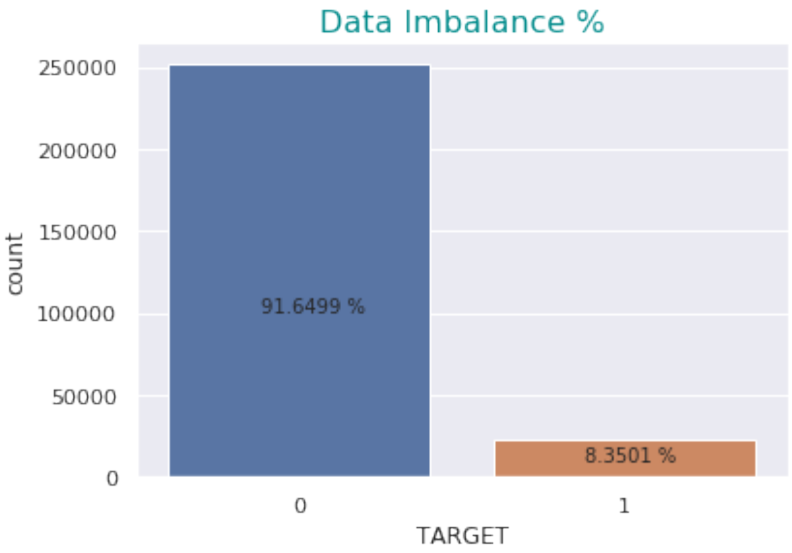




Data imbalance

The data after cleanup is highly imbalanced with around 8.35% data for loan defaulters (TARGET = 1) and remaining 91.65% data for non-defaulters.

A data-sampling technique can be used to do over sampling/under sampling which can reduce the bias introduced due to data imbalance

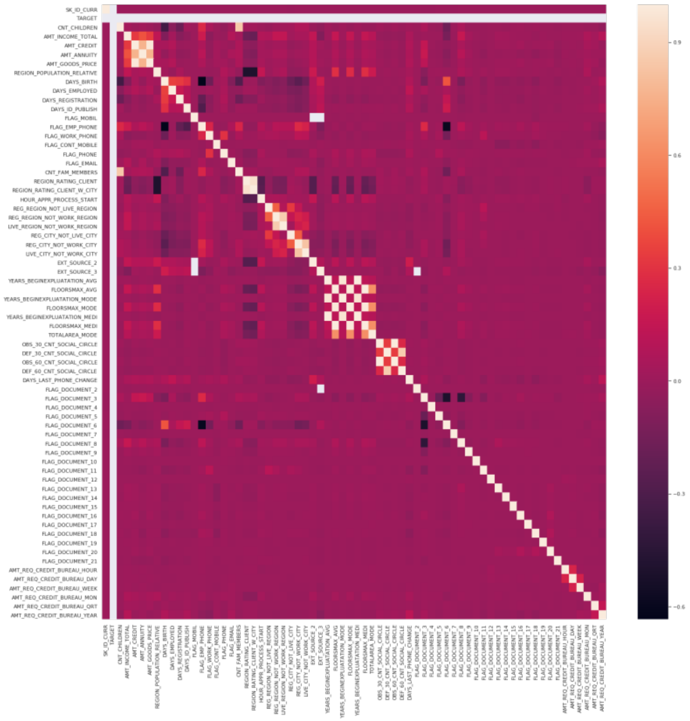


Positive correlation

Top 10 list of positive correlations between variables

Variable 1	Variable 2	Correlation for TARGET = 0	Correlation for TARGET = 1
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998518	0.998223
FLOORSMAX_AVG	FLOORSMAX_MEDI	0.997127	0.997275
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993679	0.996793
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.988479	0.989553
FLOORSMAX_AVG	FLOORSMAX_MODE	0.985978	0.986756
AMT_GOODS_PRICE	AMT_CREDIT	0.981619	0.977604
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.972117	0.981389
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.963316	0.977934
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950646	0.959696
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.861389	0.870284

Correlation diagram for Target = 0



- Top 10 correlations between variables are in the range of 0.861 - 0.998. and both datasets(defaulter and non-defaulter) have similar correlation.

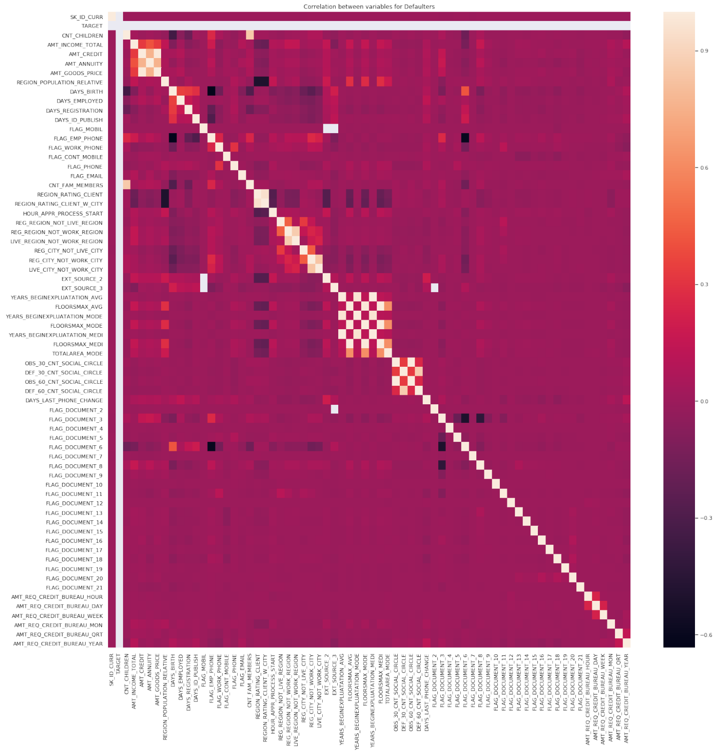
- All the variables with top 10 correlations are matching in both datasets with their correlation varying within 0.01 scale.

Negative correlation

Top 10 list of negative correlations between variables

Variable 1	Variable 2	Correlation for TARGET = 0	Correlation for TARGET = 1
FLAG_EMP_PHONE	DAYS_BIRTH	-0.633631	--0.586893
FLAG_DOCUMENT_6	FLAG_EMP_PHONE	-0.600236	-0.620202
FLAG_DOCUMENT_6I	FLAG_DOCUMENT_3	-0.506959	-0.489693
REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	-0.500267	-0.427202
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.500259	-0.422743
FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	-0.431832	-0.512800
CNT_CHILDREN	DAYS_BIRTH	-0.351201	-0.272549
FLAG_EMP_PHONE	DAYS_ID_PUBLISH	-0.290013	
CNT_FAM_MEMBERS	DAYS_BIRTH	-0.288651	
HOUR_APPR_PROCESS_START	REGION_RATING_CLIENT	-0.275983	-0.293270
REGION_RATING_CLIENT_W_CITY	HOUR_APPR_PROCESS_START		-0.275582
EXT_SOURCE_2	REGION_RATING_CLIENT		-0.240792

Correlation diagram for Target = 1

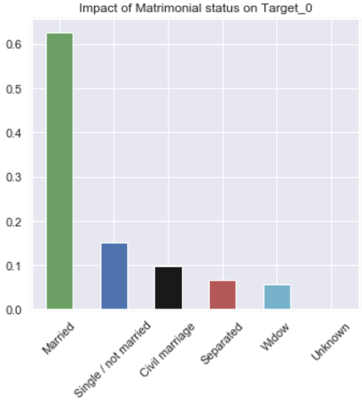


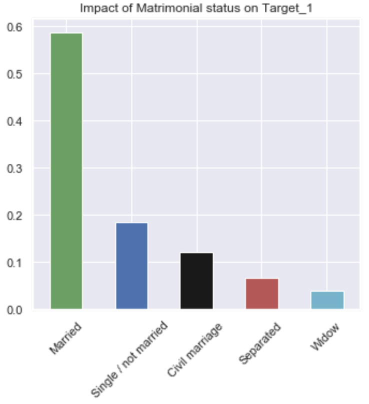
- Top 10 correlations between variables are in the range of (-0.275) to (-0.999). and both datasets(defaulter and non-defaulter) have similar correlation.
- Below their combinations of variables have higher negative correlation for defaulter dataset compared to the non-defaulter dataset
  - DAYS_EMPLOYED & FLAG_DOCUMENT_3 -0.282129
  - HOUR_APPR_PROCESS_START & REGION_RATING_CLIENT_W_CITY -0.275582

Segmented univariate analysis

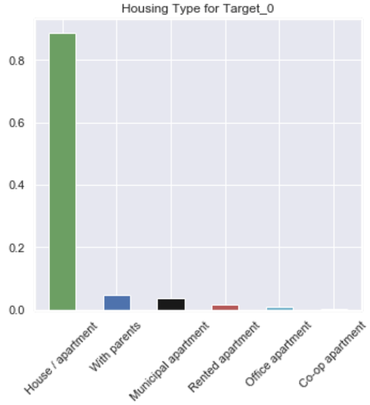
Below are few observations from segmented univariate analysis done on Family status, Housing Type, Profession,

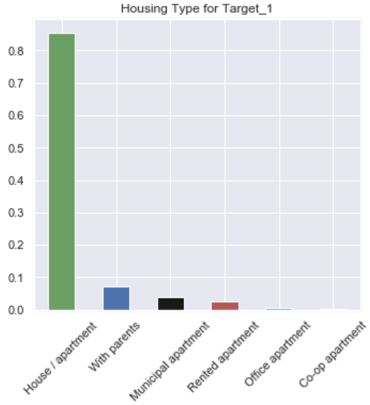
- Loan defaults are proportion is less for Married people compare to non-defaulters
- For Single/Civil Married customers, the loan defaulter proportion is little higher than the non defaulters.


Impact of Matrimonial status on Target_0

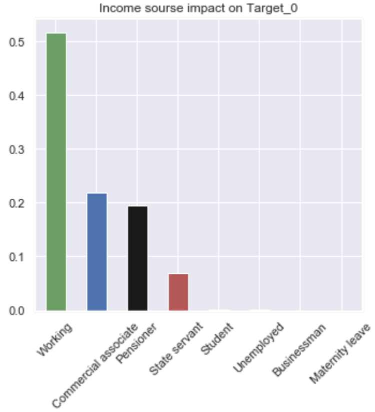

Impact of Matrimonial status on Target_1

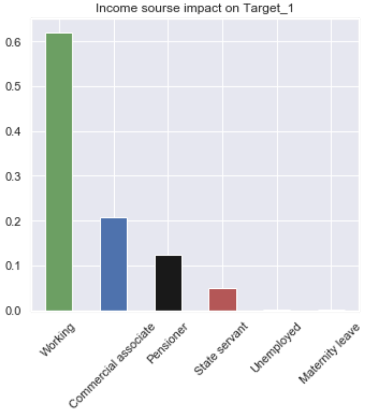
- Its seems customer living with Parents have little more proportion of defaulting compared to non-defaulters
- Similarly Municipal and Rented apartment accommodation shows slightly higher proportion towards defaulting


Housing Type for Target_0


Housing Type for Target_1

- Customers who are currently working have higher proportion of defaulters
- Pensioners seems to be pay back loan , so their proportion is less on defaulters
- State servants are comparatively show less tendency towards defaulting

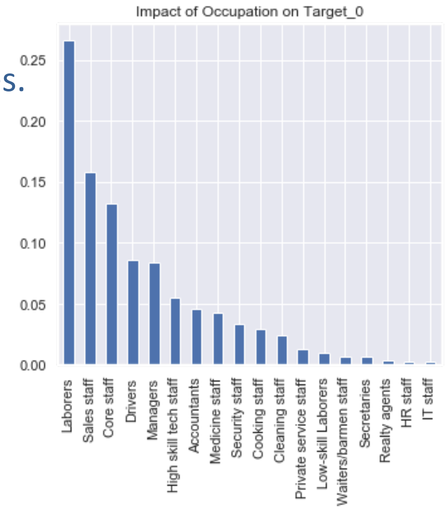

Income sourse impact on Target_0

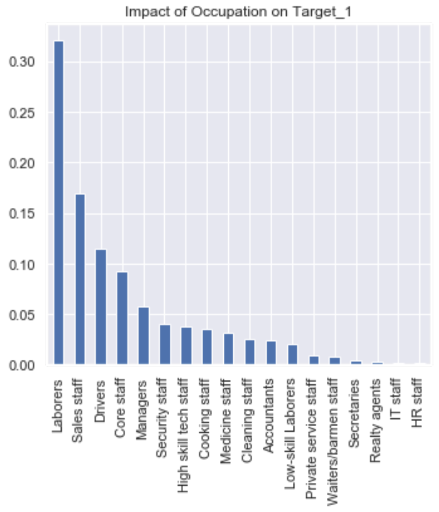

Income sourse impact on Target_1

Segmented univariate analysis

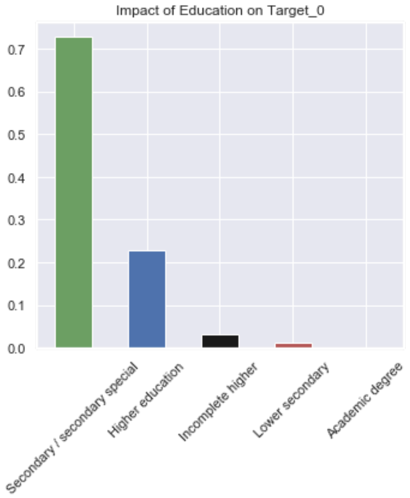
Below are few observations from segmented univariate analysis done on Employment status and education variables.

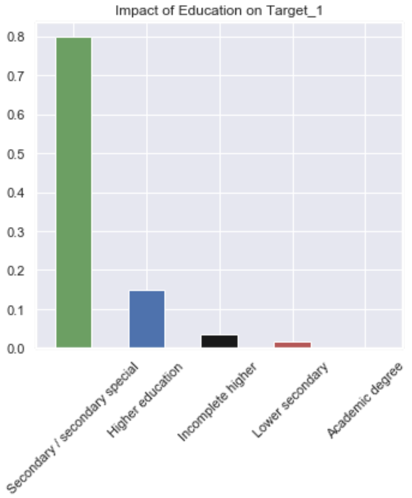
- Customers with profession as Laborer have higher proportion of defaulters
- Another observation is as IT/HR staff have lower proportion of defaulting


Impact of Occupation on Target_0


Impact of Occupation on Target_1

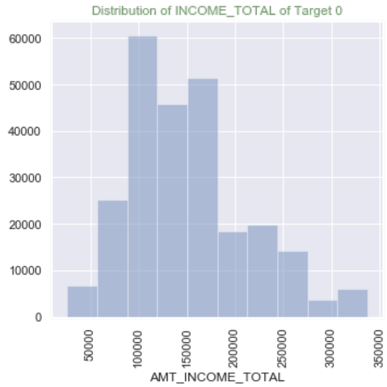
- Customers with Secondary education have high proportion of defaulting if compared to non-defaulters
- Customers with higher education tend to default less as their proportion is reduced

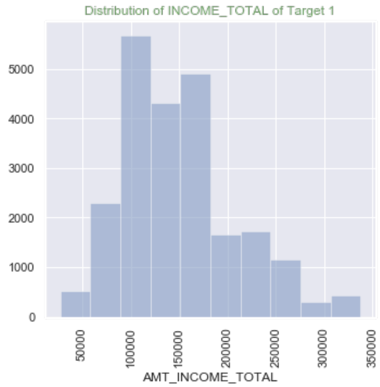

Impact of Education on Target_0


Impact of Education on Target_1

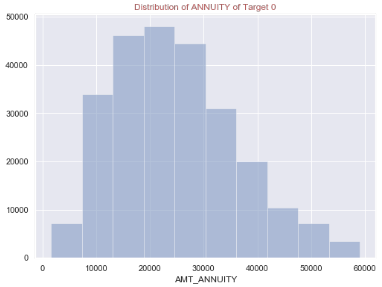
Univariate analysis

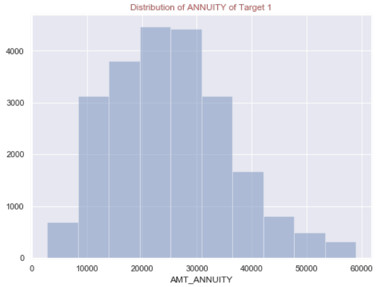
From univariate analysis on income, annuity, Loan amount, age and region population fields, below are few insights:



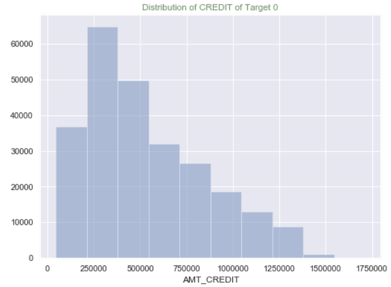


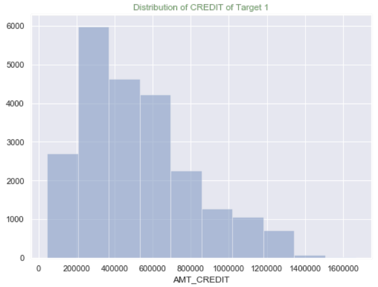
- The income of the customers seems to have similar distribution for both defaulters and non-defaulters
- The Average income seems to be around 140K for both segments





- For Annuity data The defaulters seems to have more outliers compared to non-defaulters
- The average annuity is similar for both defaulters and non defaulters around 30K

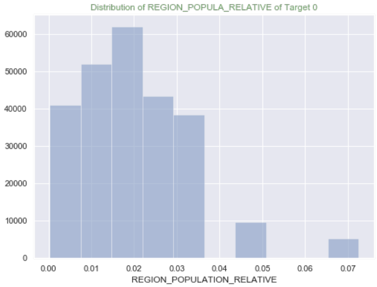


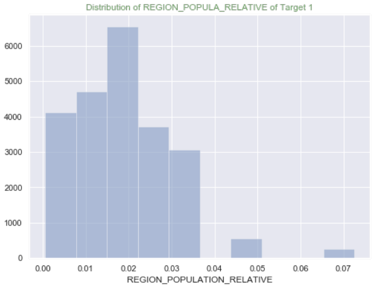


- for Loan amt. the defaulters seems to have more outliers compared to non-defaulters
- The higher fence value for defaulters is around 1.2 M compared to non-defaulters which is around 1.5M
- Large no of defaulters have credit of between 200K to 600K

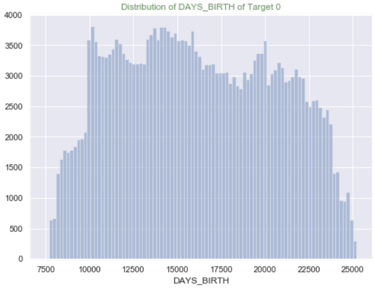
Univariate analysis

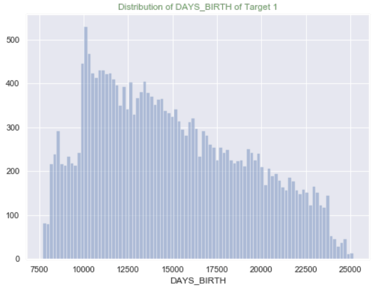
From univariate analysis on income, annuity, Loan amount, age and region population fields, below are few insights:





- It seems majority of defaulter are from low populated area, we can see that proportions for 0.05 and 0.07 are lower than that of non-defaulters



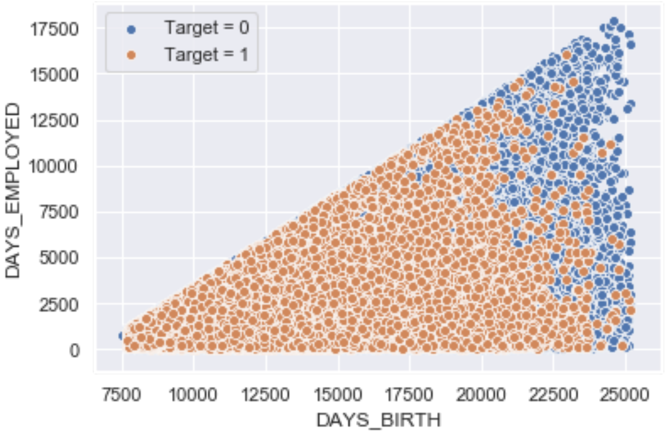


- The median age for defaulters are around 14000 days older which would be around 40 Years
- It looks like as the age increases proportion of defaulters decreases
- The younger customers seems to have higher proportion of defaulters

Bivariate analysis

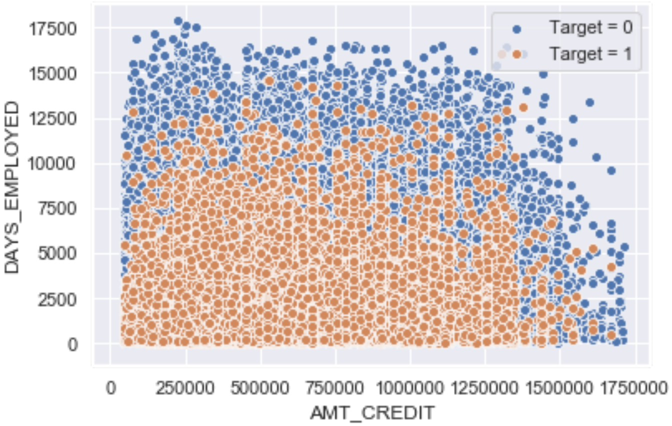
DAYS_BIRTH and DAYS_EMPLOYED

- The correlation between employment and loan amount for non defaulters is -0.0663 , but if we considered data for only non-retired applicants correlation is: 0.0818
- The correlation between employment and loan amount for loan-defaulters is 0.0018 , but if we considered data for only non-retired applicants correlation is: 0.1124
- for Payment defaulters (TARGET = 1) As age increases and days employed increases, the loan default shows reductions. So It might be case that younger people will short employment history tend to default more.
- Also for loan defaulters, their is correlation between employment period and loan amount is around 0.1124 which is significantly more than non-defaulters. (This observation is only for non-retired applicants)



AMT_CREDIT and DAYS_EMPLOYED

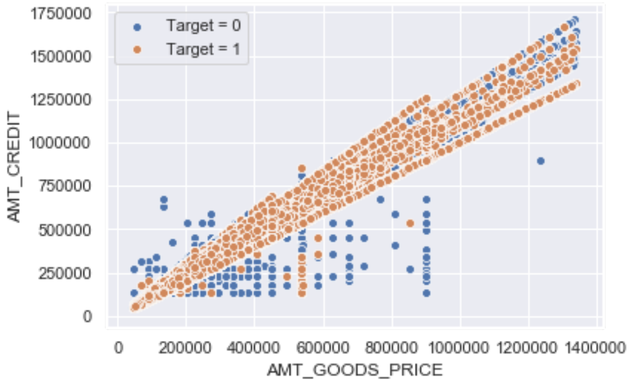
- for Payment defaulters (TARGET = 1) It seems that the credit amount of loan is low at higher experience level. Also the loan default is concentrated below 1.5M Loan amount credit and below 10000 days (around 30 Years job experience)



Bivariate analysis

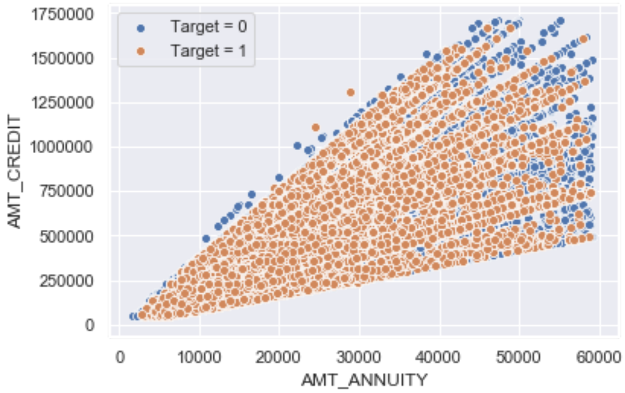
AMT_CREDIT and AMT_GOODS_PRICE

- The correlation between property price and loan amount for non defaulters is 0.9816 but for defaulters it is: 0.9776
- credit amount and goods price are highly correlated variables for both defaulters and non -defaulters. So as the home price increases the loan amount also increases



AMT_CREDIT and AMT_ANNUITY

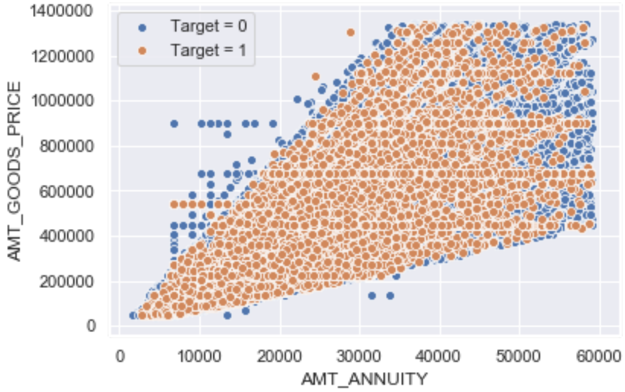
- The correlation between AMT_ANNUITY (EMI) and loan amount for non defaulters is 0.7609 but for defaulters it is: 0.7401
- credit amount and AMT_ANNUITY (EMI) are highly correlated variables for both defaulters and non -defaulters . So as the home price increases the EMI amount also increases which is logical



AMT_ANNUITY and AMT_GOODS_PRICE

The correlation between DAYS_EMPLOYED (employment) and loan amount for non defaulters is 0.7604 but for defaulters it is: 0.7374

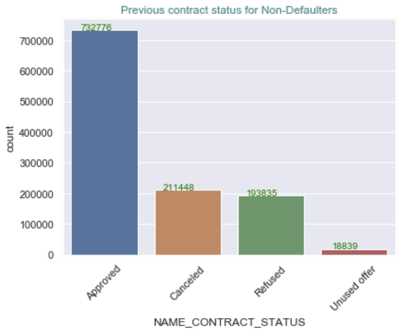
All three variables AMT_CREDIT, AMT_GOODS_PRICE and AMT_ANNUITY are highly correlated for both defaulters and non-defaulters, which might not give a good indicator for defaulter detection

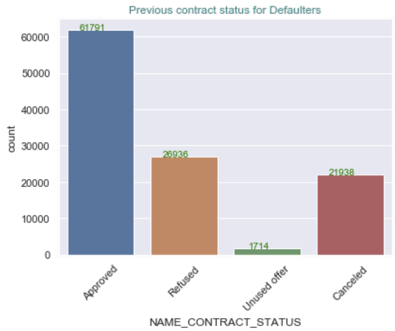


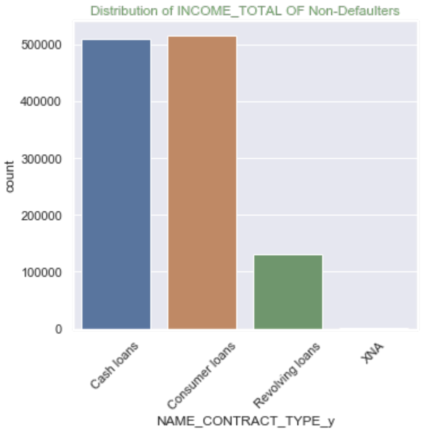
Previous application data analysis

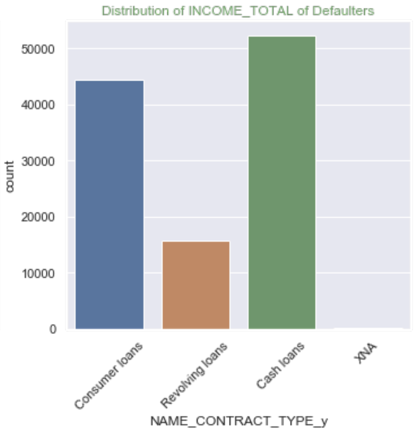
- The previously refused % of applications for non-defaulters is: 16.75
- The previously refused % of applications for defaulters is: 23.96

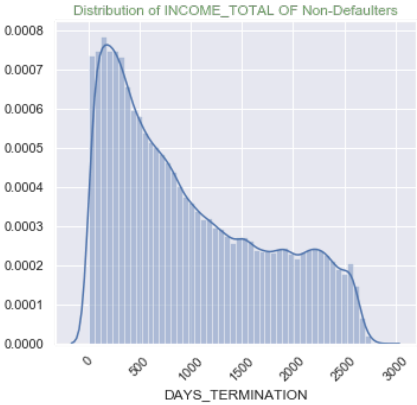
- Its seems that for TARGET = 1 clients have larger proportion of previously Refused applications
- Its seems that for defaulters more of the previous applications were for cash loans
- in case of non-defaulters the consumer loans and cash loan proportion is similar
- Its seems that for defaulters mean time of previous applications closed was smaller than that of non-defaulters
- for defaulters have previous loan duration smaller than that of non-defaulters

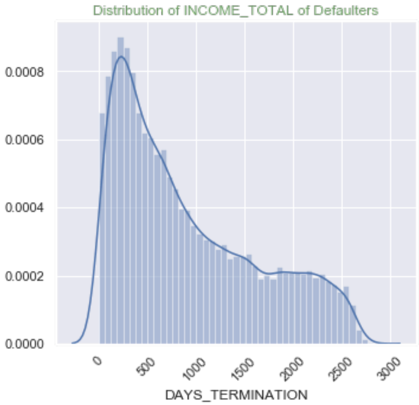

Previous contract status for Non-Defaulters

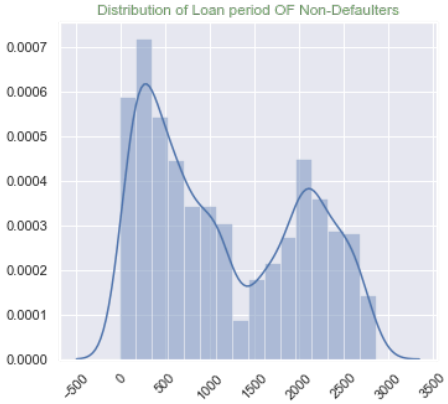

Previous contract status for Defaulters

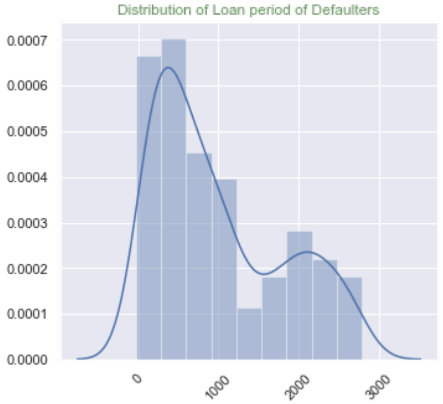

Distribution of INCOME_TOTAL OF Non-Defaulters


Distribution of INCOME_TOTAL of Defaulters


Distribution of INCOME_TOTAL OF Non-Defaulters


Distribution of INCOME_TOTAL of Defaulters

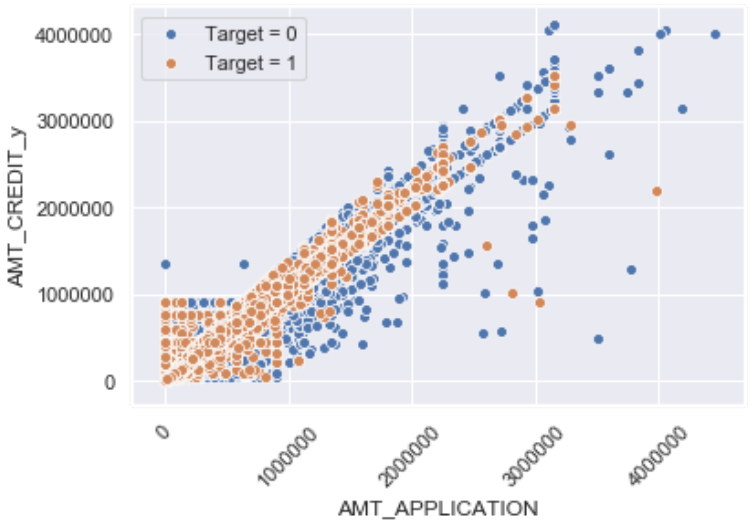

Distribution of Loan period OF Non-Defaulters


Distribution of Loan period of Defaulters

Previous application data analysis

Bivariate analysis of AMT_CREDIT_y and AMT_APPLICATION

- The previous application amount and credit amounts are show positive correlation around 0.97 Also it seems that previous application amount and current application amount have weak positive correlation of 0.092 for non-defaulters and 0.095 for defaulters



Thank You