Credit EDA Case Study

By

Saurabh Sharma

Rathish Rajendran

Business Objectives:

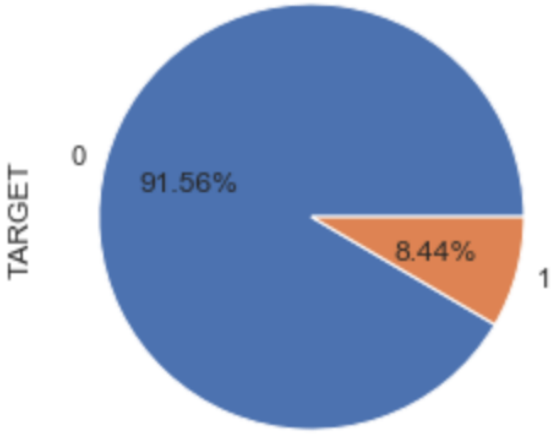
This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Datasets provided:

1. *'application_data.csv'* contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties**.
2. *'previous_application.csv'* contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer**.

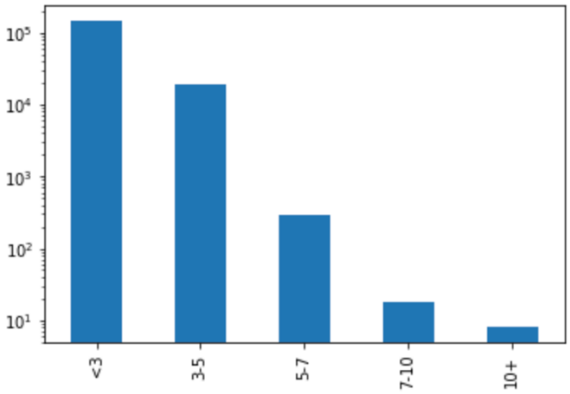
Analysis:

Plotting Imbalance between Target 0 and 1 values:



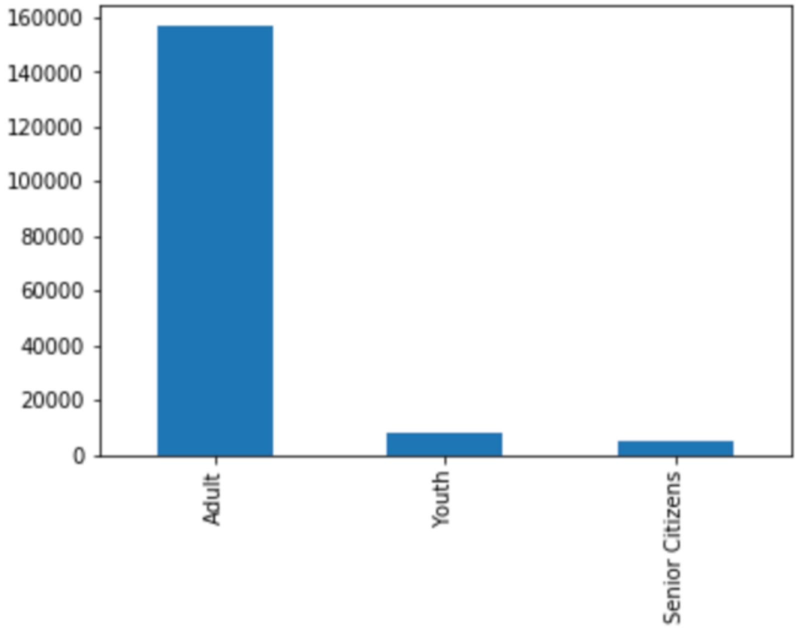
Insight: Out of total loan applications about 8.44 % defaulted.

Plotting Family members count for Loan applicants:



Insight: People with less than 3 family members are the ones with maximum number of loan applications and people with more than 10 family members are with the least number of loan applications.

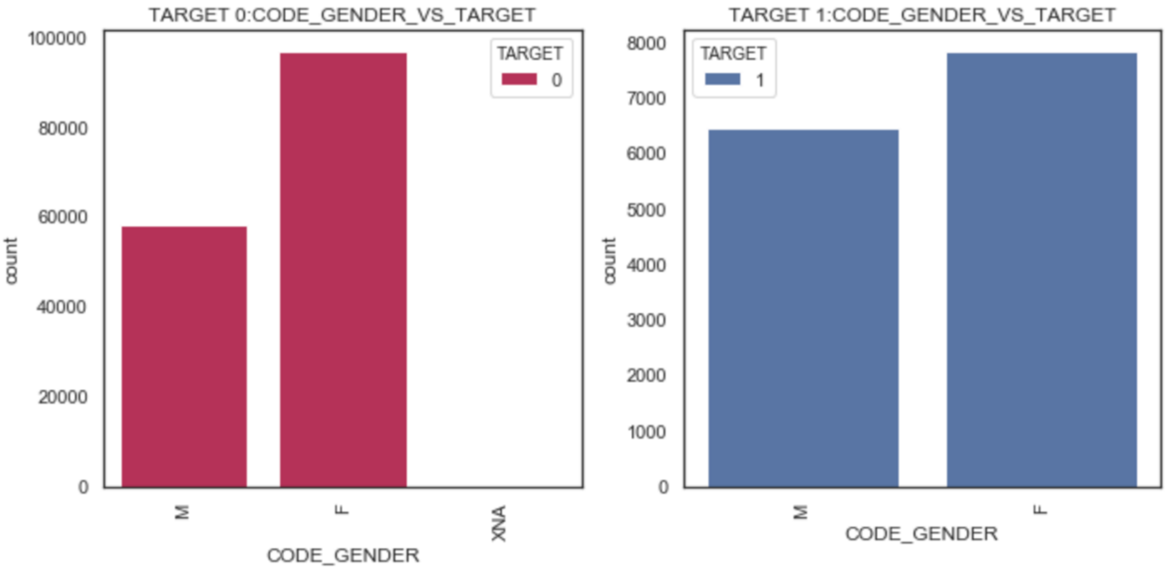
Plotting AGE:



Insight: People in Adult age group (25-59 years) are the ones with maximum number of loan applications and Senior Citizens(60+ years) are with the least number of loan applications.

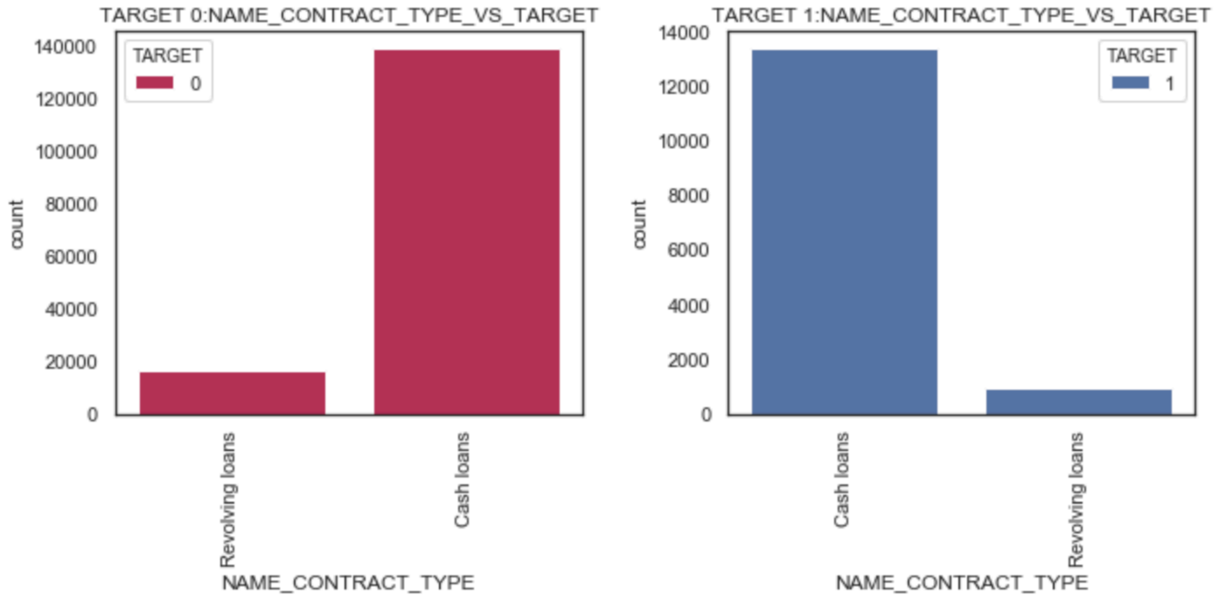
Univariate Analysis on Categorical variables for application data:

Plotting Gender Vs Target:



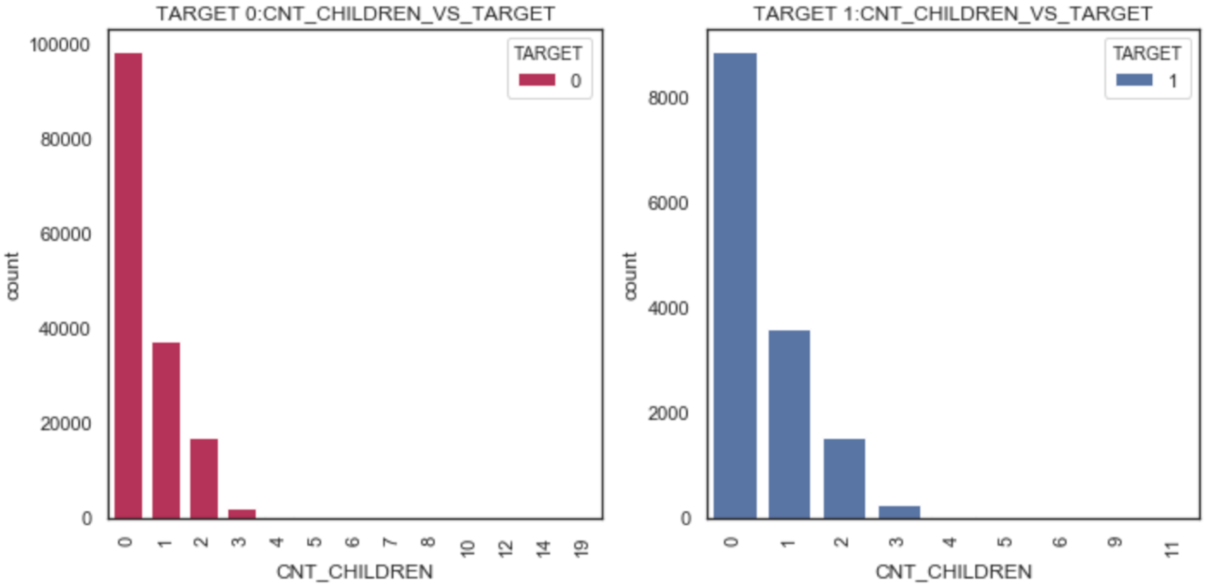
Insight: Female applicants constitute larger population than males in loan application in both the cases Target=0 as well as Target=1.

Plotting Loan Type Vs Target:



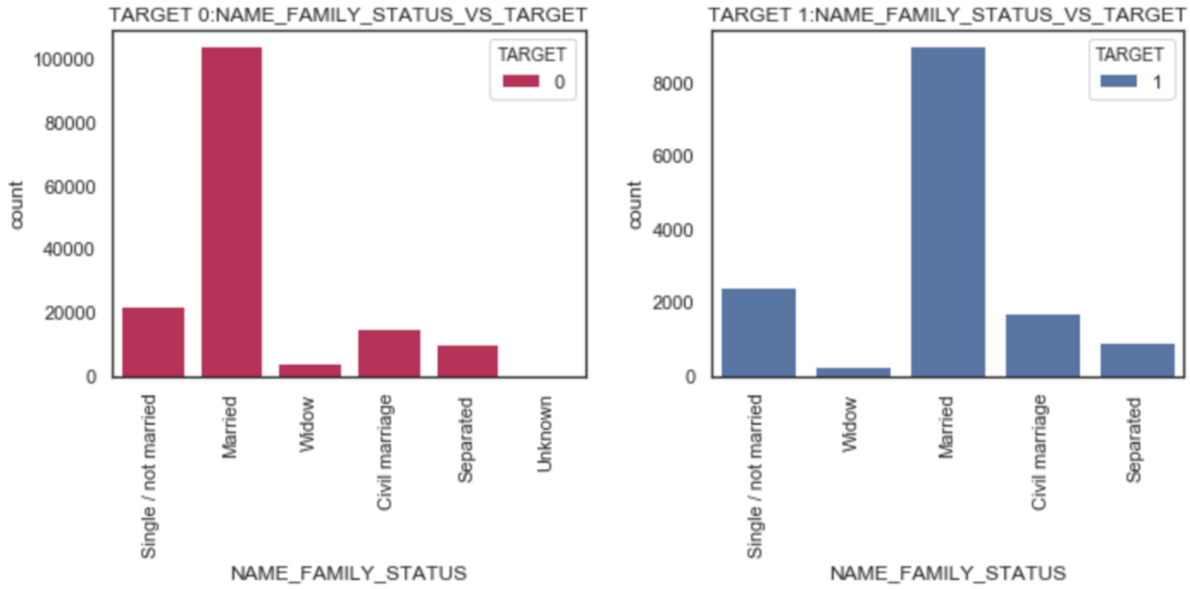
Insight: Cash loans is most preferred form of loan application in both the cases Target=0 as well as Target=1.

Plotting Count of Children Vs Target:



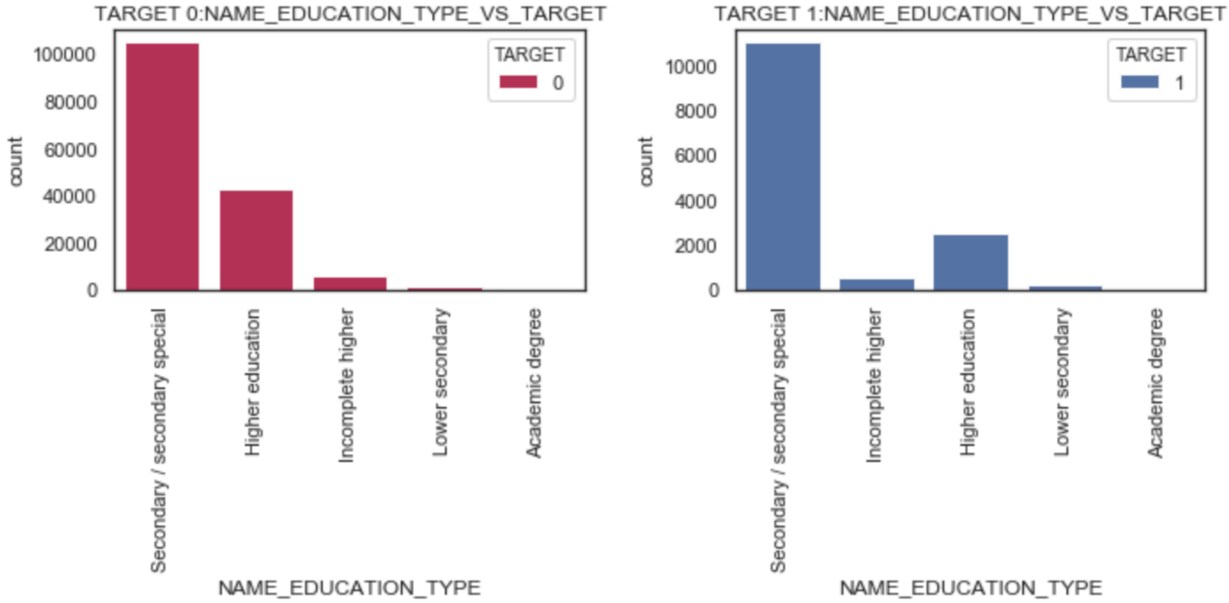
Insight: People having 0 children constitute largest population in loan applications in both the cases Target=0 as well as Target=1

Plotting Marital status Vs Target:



Insight: Married people constitute largest population in loan application in both the cases Target=0 as well as Target=1. This may arise due to additional financial overhead due to expenses and liabilities.

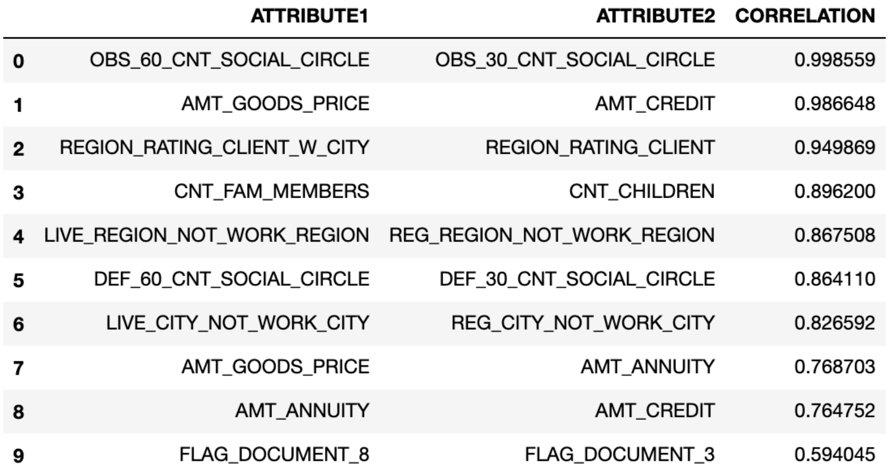
Plotting Education level Vs Target:



Insight: People with Secondary / secondary special education constitute largest population in loan application in both the cases Target=0 as well as Target=1.

Correlation:

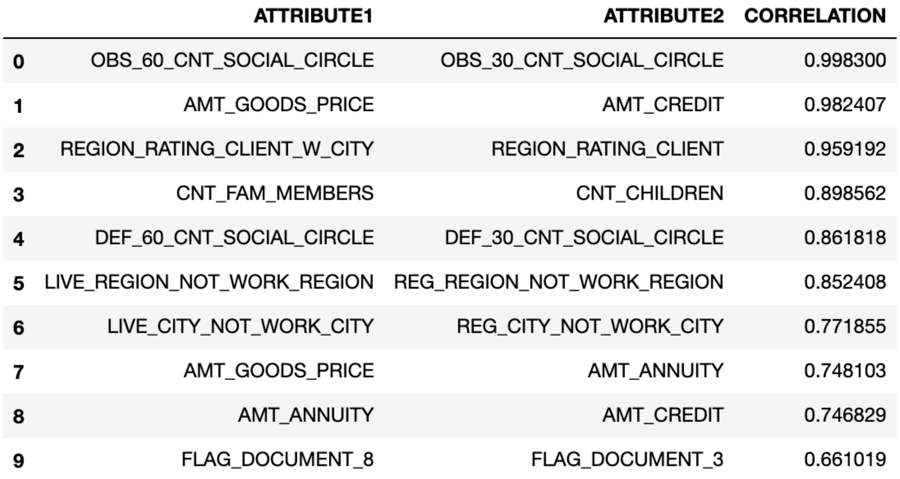
Top 10 Correlation for Target 0:



	ATTRIBUTE1	ATTRIBUTE2	CORRELATION
0	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998559
1	AMT_GOODS_PRICE	AMT_CREDIT	0.986648
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.949869
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.896200
4	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.867508
5	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.864110
6	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.826592
7	AMT_GOODS_PRICE	AMT_ANNUITY	0.768703
8	AMT_ANNUITY	AMT_CREDIT	0.764752
9	FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	0.594045

Top 10 Correlation for Target 0 (i.e. who are less likely to default the loan)

Top 10 Correlation for Target 1



	ATTRIBUTE1	ATTRIBUTE2	CORRELATION
0	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998300
1	AMT_GOODS_PRICE	AMT_CREDIT	0.982407
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.959192
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.898562
4	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.861818
5	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.852408
6	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.771855
7	AMT_GOODS_PRICE	AMT_ANNUITY	0.748103
8	AMT_ANNUITY	AMT_CREDIT	0.746829
9	FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	0.661019

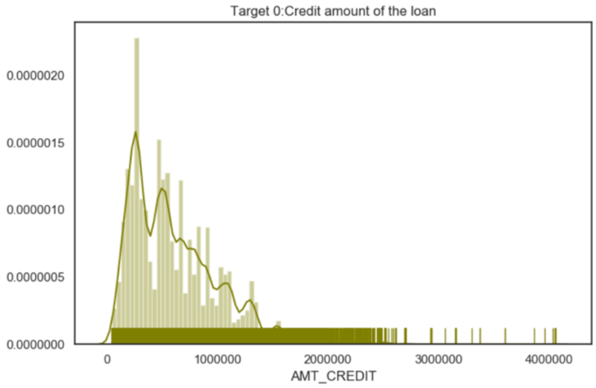
Top 10 Correlation for Target 1 (i.e. who are most likely to default the loan)

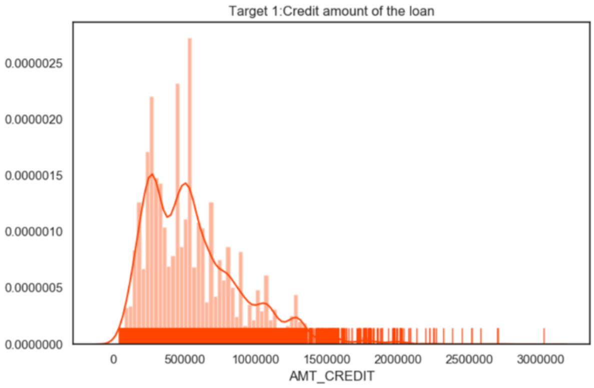
Common correlation variables between Target 0 and 1:

OBS_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE
AMT_GOODS_PRICE
AMT_CREDIT
REGION_RATING_CLIENT_W_CITY
REGION_RATING_CLIENT
CNT_FAM_MEMBERS
CNT_CHILDREN
LIVE_REGION_NOT_WORK_REGION
REG_REGION_NOT_WORK_REGION
DEF_60_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
LIVE_CITY_NOT_WORK_CITY
REG_CITY_NOT_WORK_CITY
AMT_GOODS_PRICE
AMT_ANNUITY
AMT_CREDIT
FLAG_DOCUMENT_8
FLAG_DOCUMENT_3

Univariate Analysis on Numerical variables for Target 0 and 1:

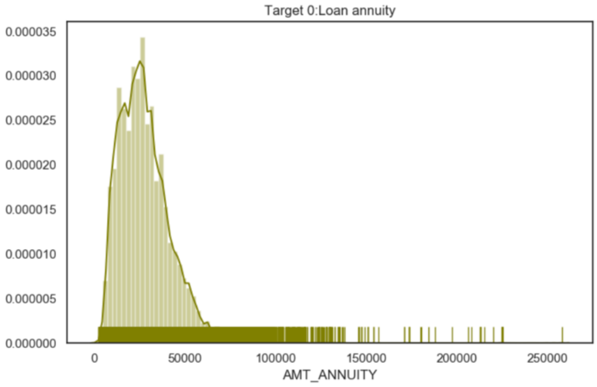
Plotting Credit Amount Vs Target:

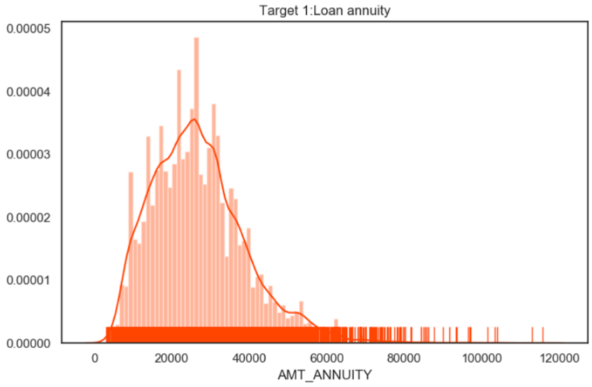




<u>**Insight:** As per above graph the amount of loan credited is more in the range of 0-1000000</u>

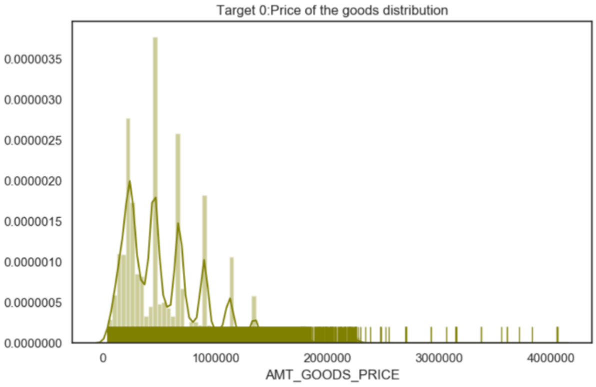
Plotting Annuity Amount Vs Target:

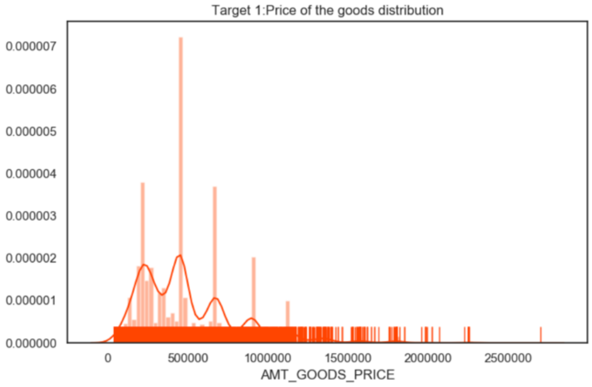




<u>**Insight:** As per above graph the Annuity amount is more in the range of 0-50000</u>

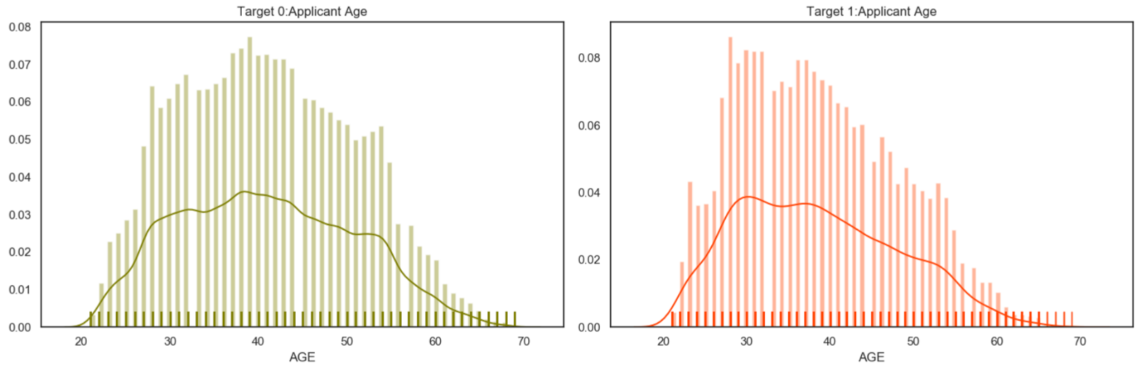
Plotting Goods Amount Vs Target:





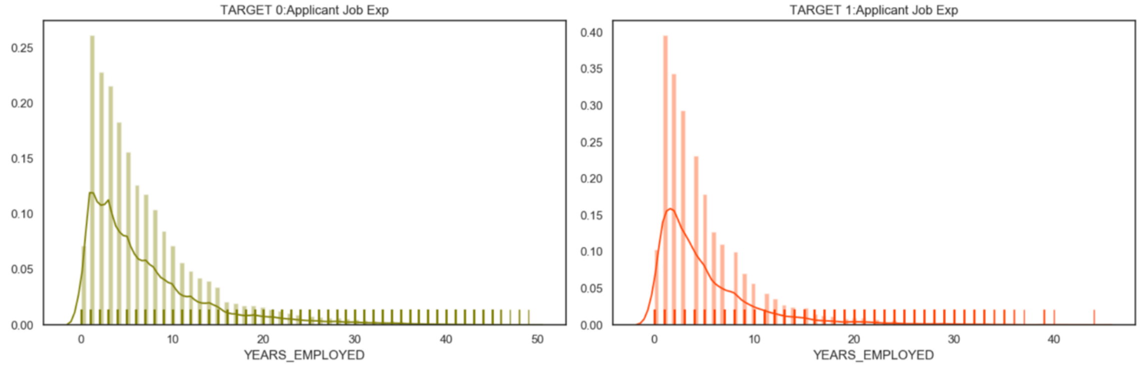
<u>**Insight:** As per above graphs the amount of Goods is more in the range of 0-1000000</u>

Plotting Age Vs Target:



Insight: As per above graphs the Age is more in the range of 25-55

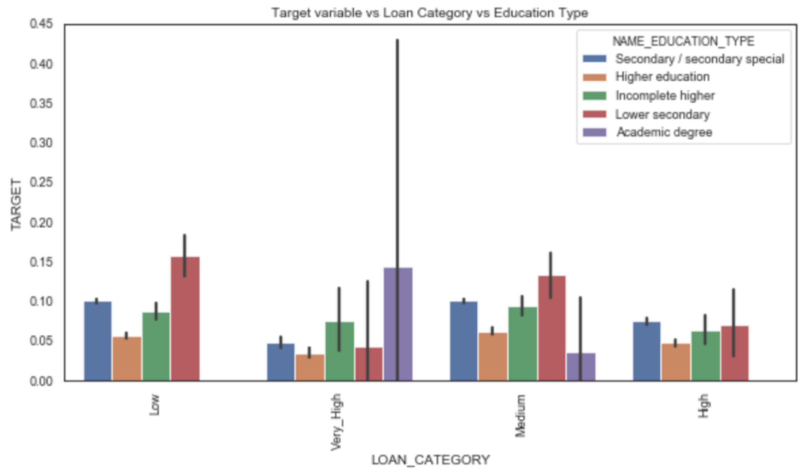
Plotting Employment Years Vs Target:



Insight: As per above graphs people having employment years ranging from 0-10 are more.

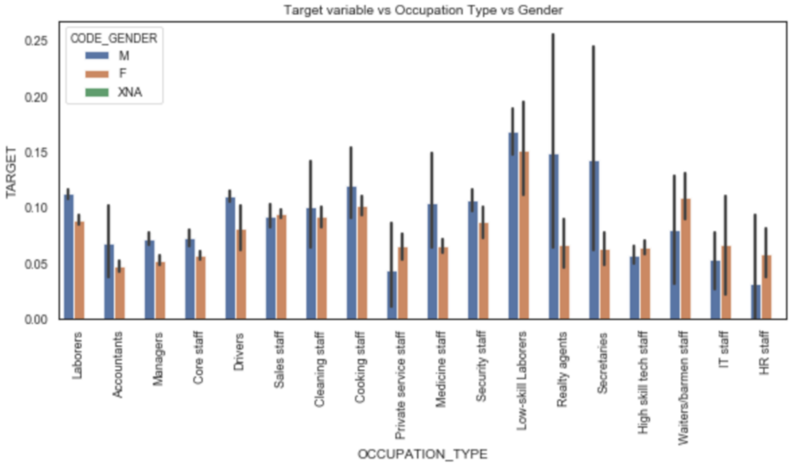
Bivariate Analysis on Categorical variables for application data:

Plotting Target Vs Loan Category Vs Education Type:



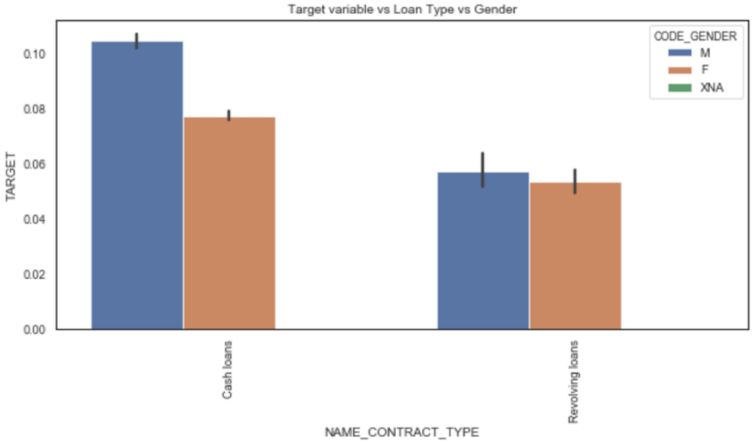
Insight: People having Lower secondary education take more loans and tends to default the loan as well.

Plotting Target Vs Occupation Type Vs Gender:



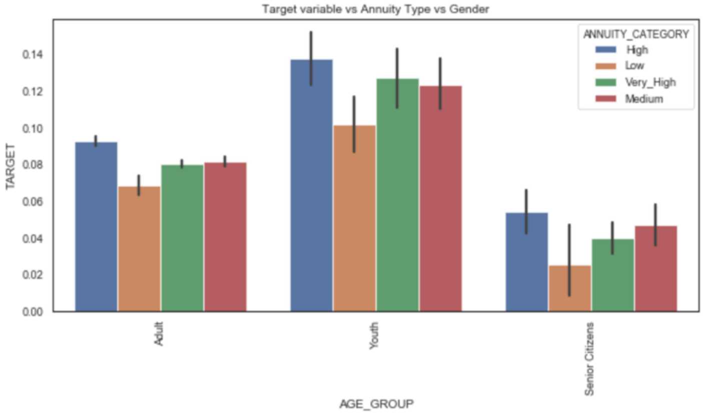
Insight: Male Low skilled Laborers are most likely to default the loan.

Plotting Target Vs Loan Type Vs Gender:



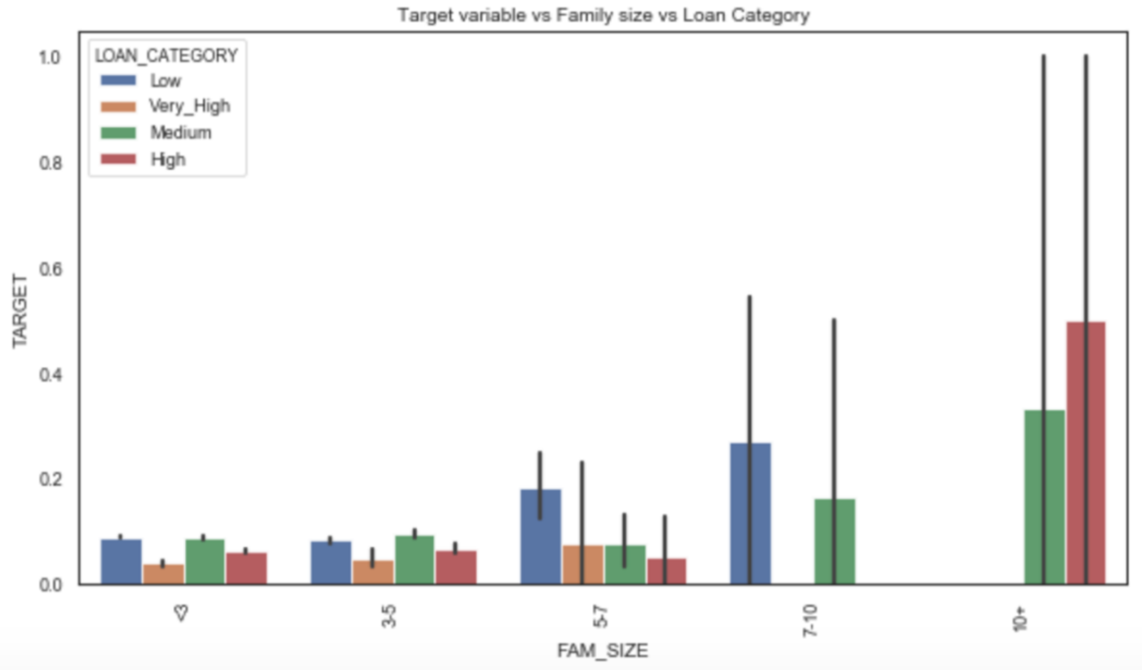
Insight: Males tend to default the loan more in case of Cash loans and Revolving loans as well

Plotting Target Vs Annuity Amount Vs Age:



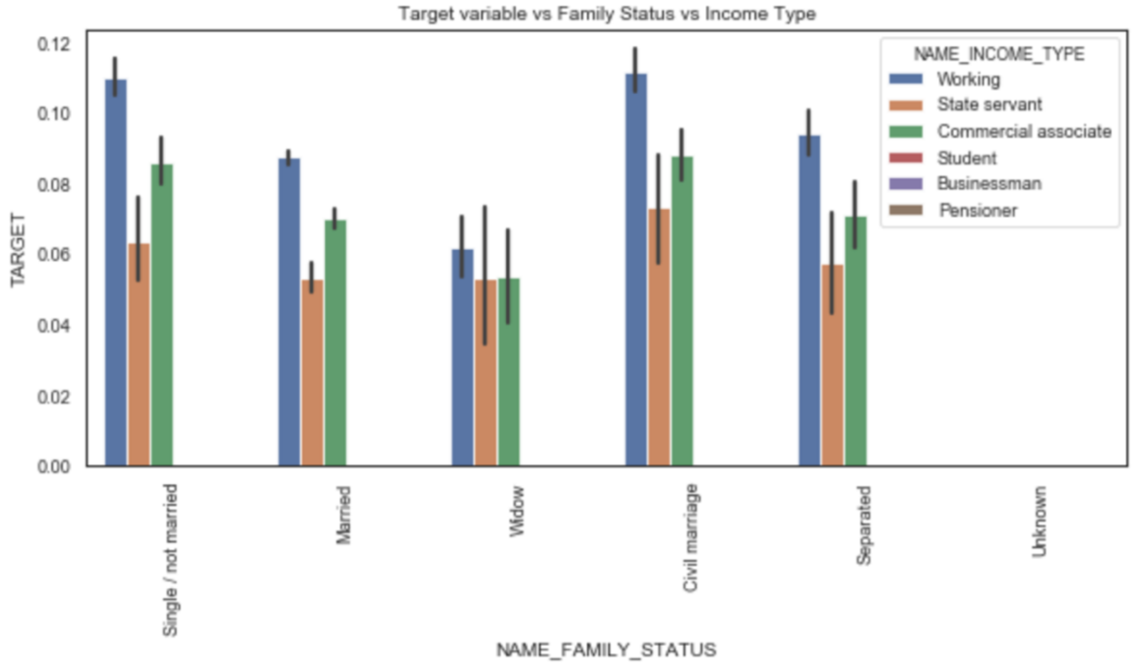
Insight: Youth population with High Annuity amount (20000-30000) tends to default the loan more

Plotting Target Vs Family Size Vs Loan Amount:



Insight: People having 10+ family members and having High and Medium Loan amount tends to default the loan more

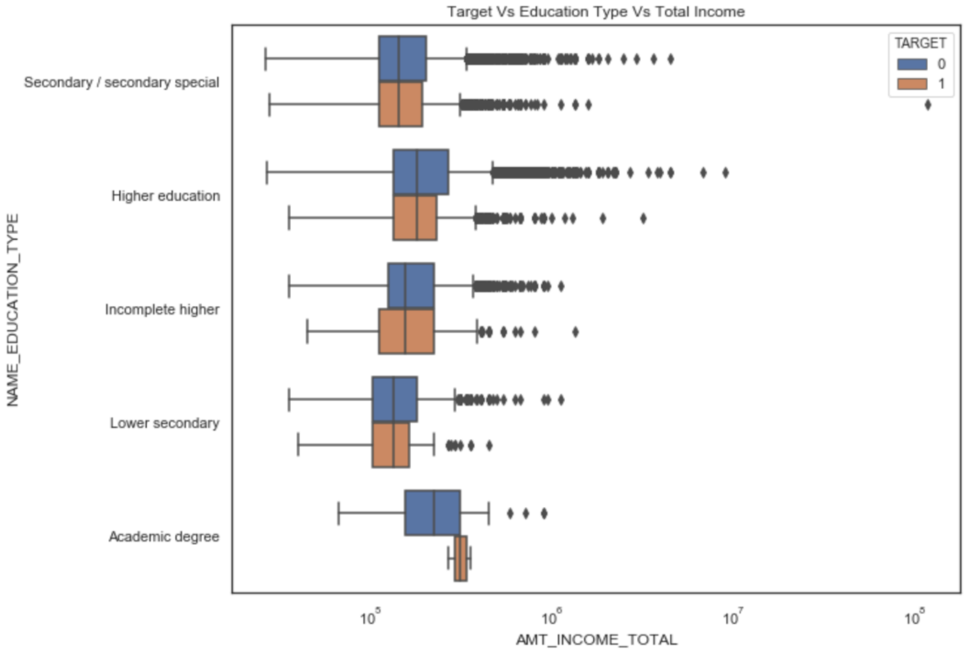
Plotting Target Vs Marital Status Vs Income Type:



Insight: Single/Civil marriage people who are Working tends to default more.

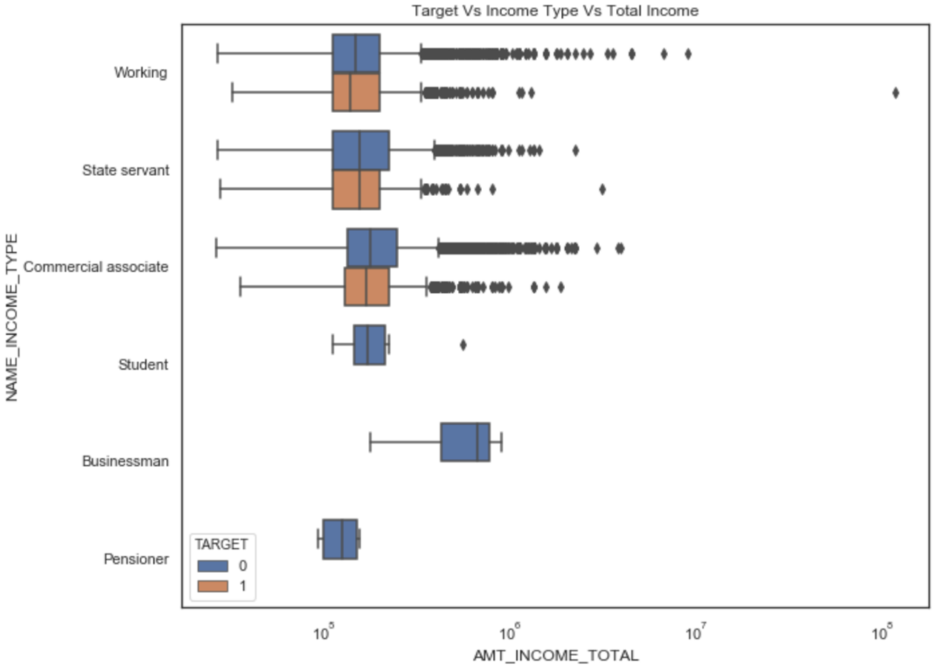
Bivariate Analysis on Numerical variables for application data:

Plotting Target Vs Education Type Vs Total Income:



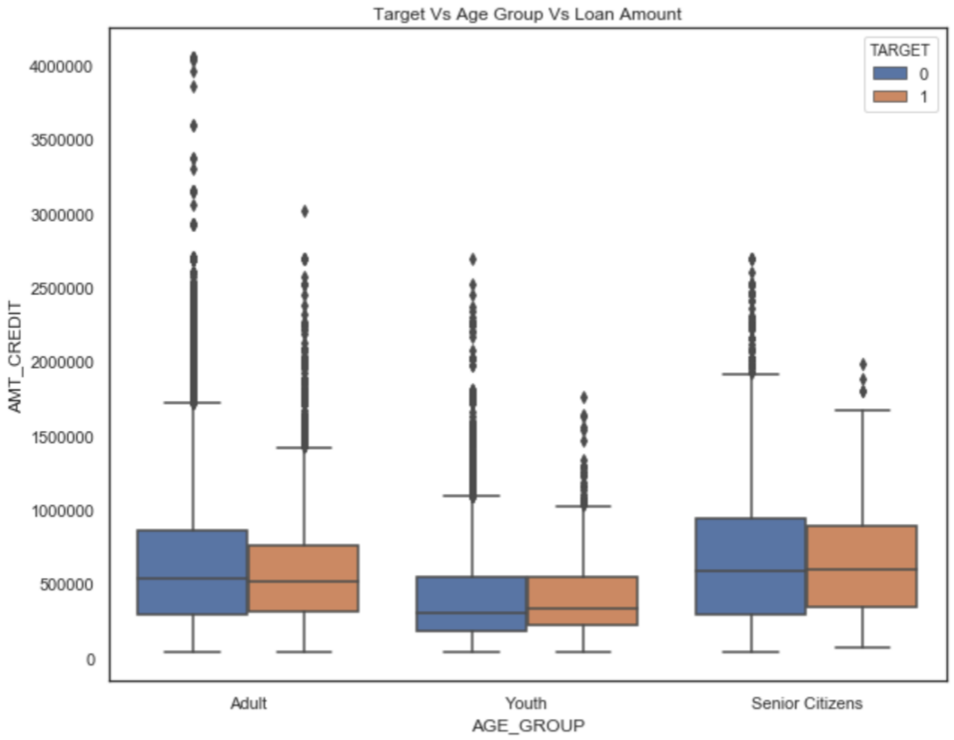
Insight: People having Academic degree are most likely to repay the loan.

Plotting Target Vs Income Type Vs Total Income:



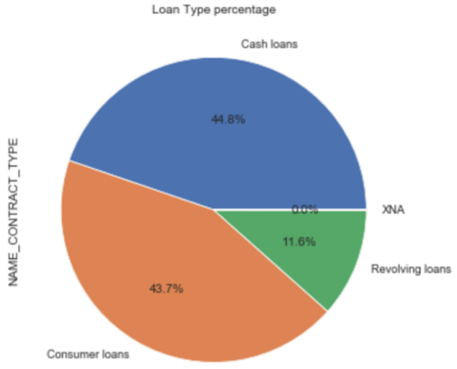
Insight: Businessman having higher income total are more likely to repay the loan

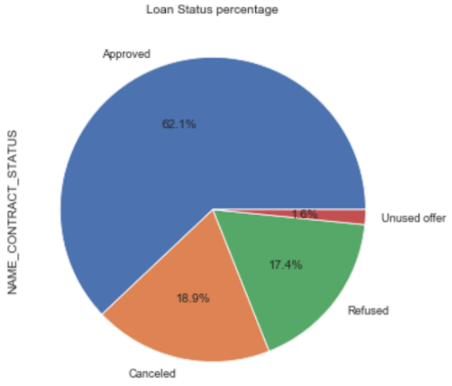
Plotting Target Vs Age Group Vs Loan Amount:

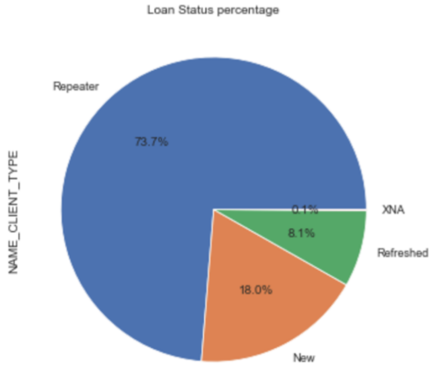


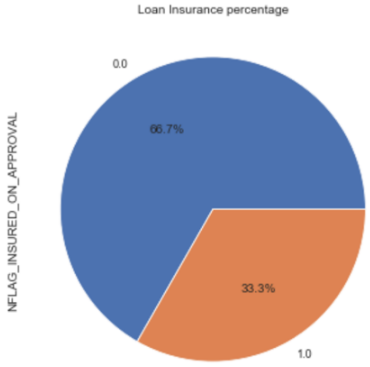
Insight: For loan amount between 2500000 and 1000000 people across all age group tends to default.

Previous Application data Analysis:







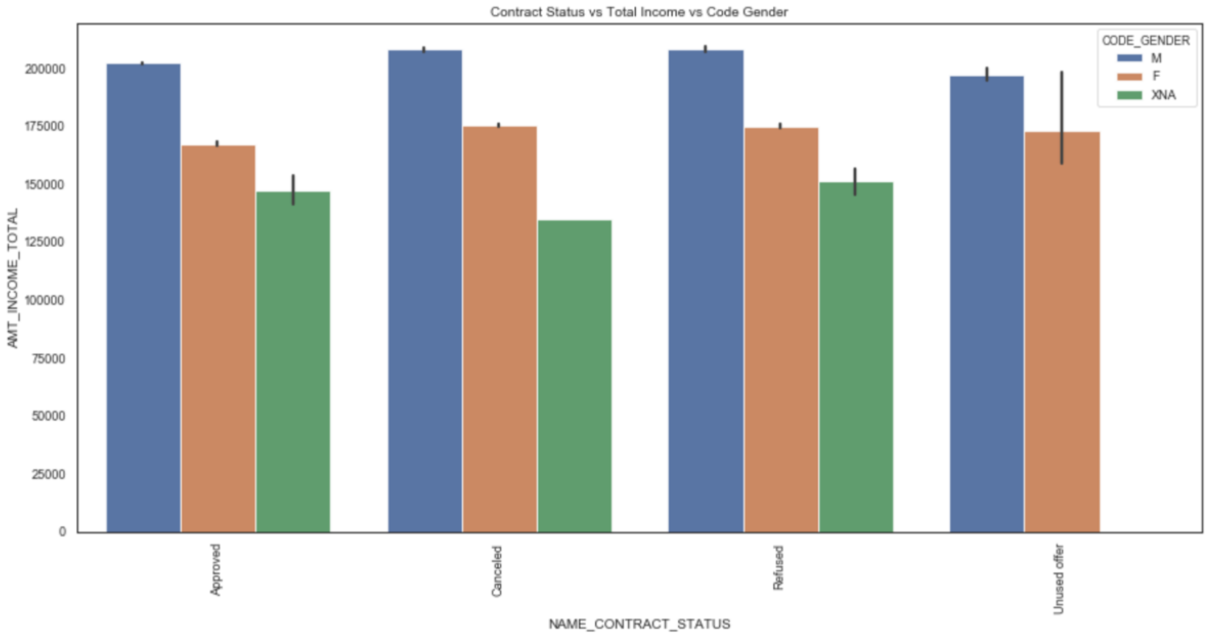


Statistics from above pie charts:

43.7 % of loans were Consumer loans
11.6 % were Revolving loans
44.8 % were Cash loans
62.1 % loans were Approved
18.9 % loans were Cancelled
17.4 % loans were Refused
1.6 % loans were Unused offer
73.7 % applicants are repeaters who have applied for loans previously.
66.7 % applicants had their loans insured

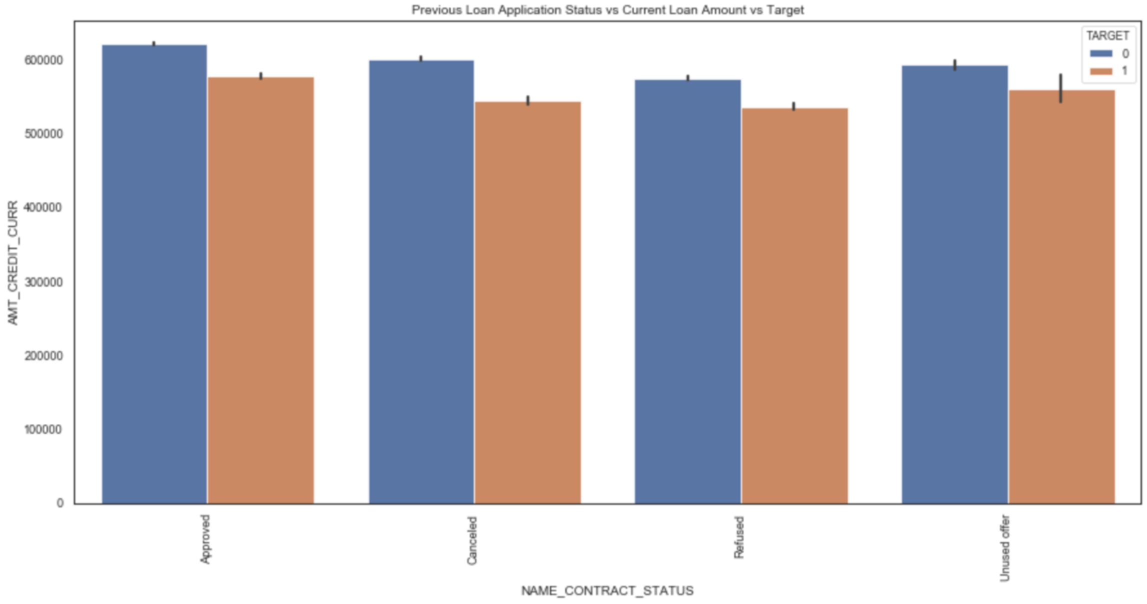
Bivariate Analysis on Merged Datasets:

Plotting Previous Loan Application Status Vs Total Income Vs Gender:



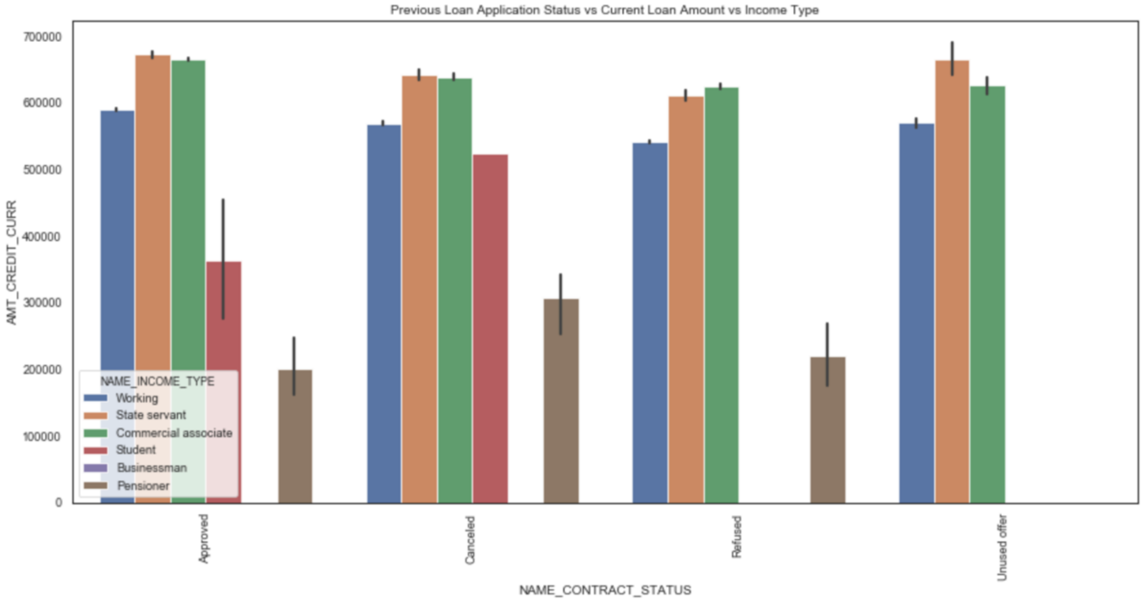
Insight: Male population across all application status values tends to have almost same Total Income i.e. 200000

Plotting Previous Loan Application Status vs Current Loan Amount vs Target:



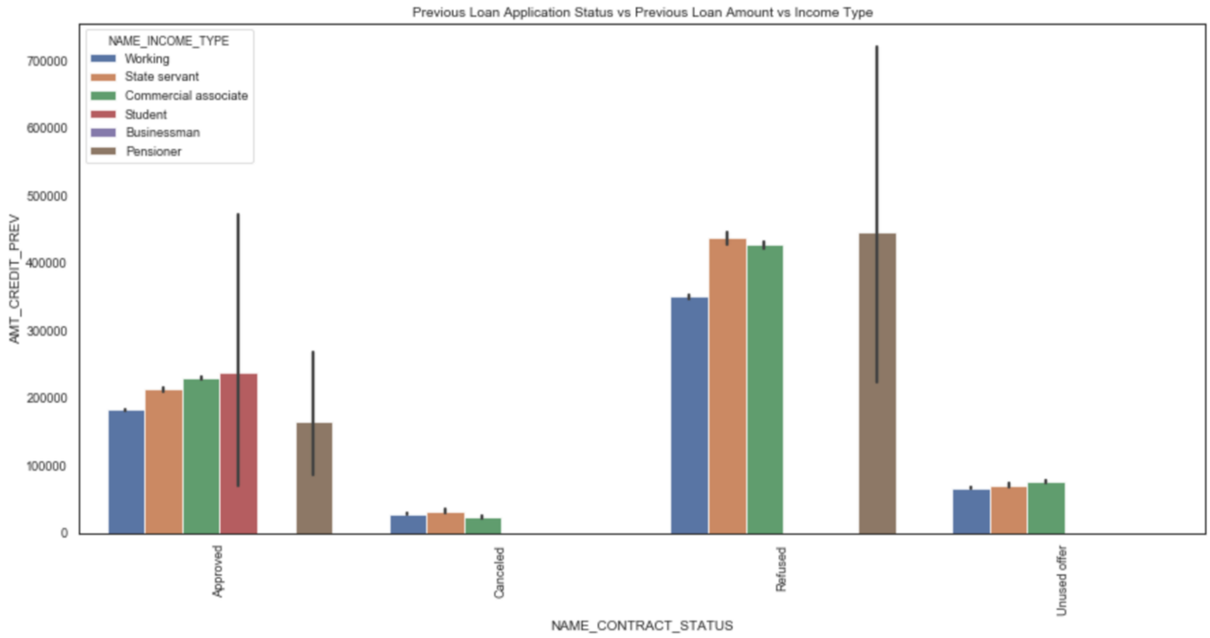
Insight: People with loan amount close to 600000 are most likely to Default the loan and repay the loan as well

Plotting Previous Loan Application Status vs Current Loan Amount vs Income Type:



Insight: State servant and Commercial associate with almost same loan application amount (i.e. 6700000) are getting their loan Approved more.

Plotting Previous Loan Application Status vs Previous Loan Amount vs Income Type:



Insight: Loan applications for amount between 0-2400000 generally gets Approved across all income types.

Conclusion:

From all our analysis, visualisation and insights provided so far, below are variables that contributes most towards loan application default:

GENDER
EDUCATION TYPE
OCCUPATION TYPE
INCOME TYPE
CONTRACT TYPE
AGE
FAMILY STATUS
FAMILY SIZE