Credit EDA CASE STUDY

Submitted by:

SaiSravya Turaga

Sajida Salam

Business Objectives

This case study aims 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.

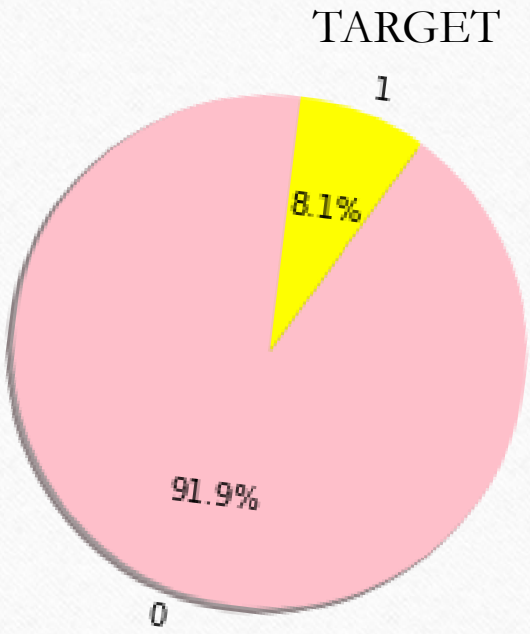
About the data set

- The data given contains the information about the loan application at the time of applying for the loan. It contains 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(TARGET=1).
- **All other cases:** All other cases when the payment is paid on time(TARGET= 0).

Imbalance Percentage

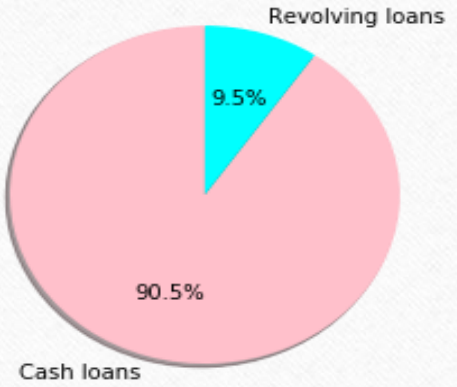
The client with payment difficulties(TARGET=1) : **8.1%**

All other cases(TARGET=0) : **91.9%**

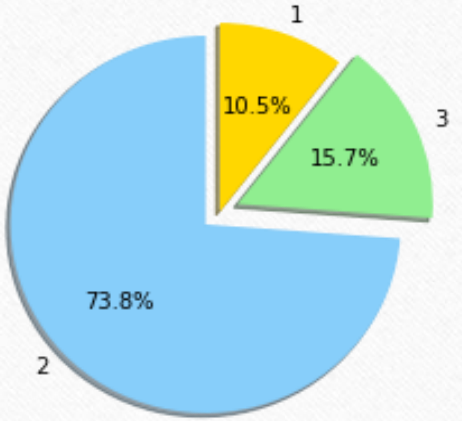


Imbalance Percentage

NAME_CONTRACT_TYPE
This column identifies if the loan is cash or revolving.
Cash loans : 90.5%
Revolving loans : 9.5%



REGION_RATING_CLIENT
The rating of the region where client lives (1,2,3)
Region 1 : 10.47%
Region 2 : 73.8%
Region 3 : 15.7%

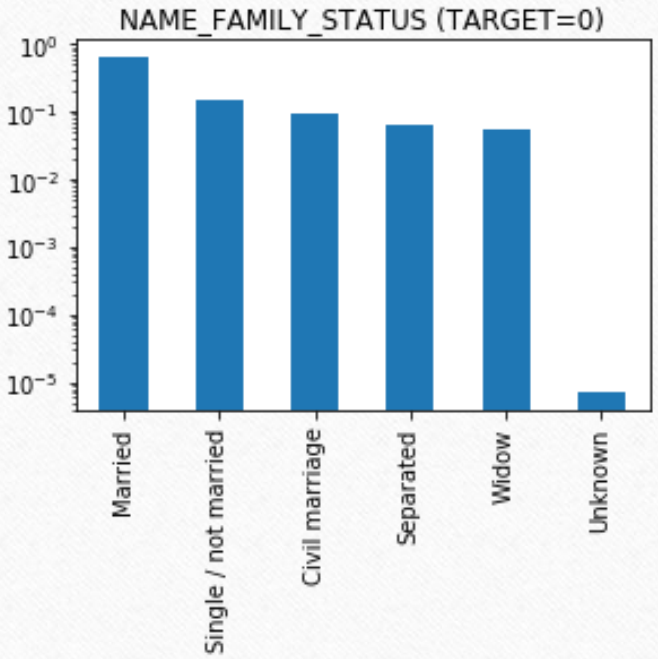


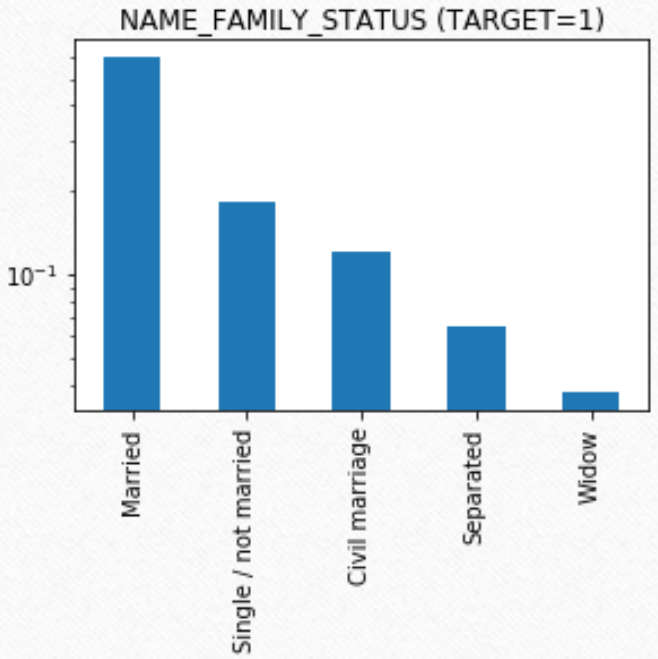
Univariate analysis for categorical variables

NAME_FAMILY_STATUS

1. Clients who are "widow", "separated", mostly tend to pay on time. So we can think in a way that the loan application can be surely approved for such category of clients.

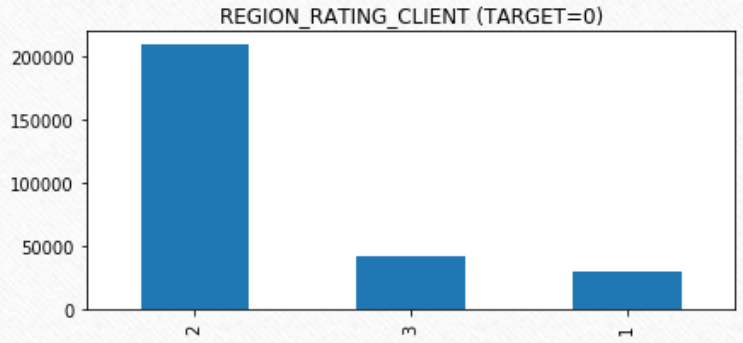
2. Clients in the "Married", "Single /not married" and "Civil marriage" status have a higher chance of defaulting.

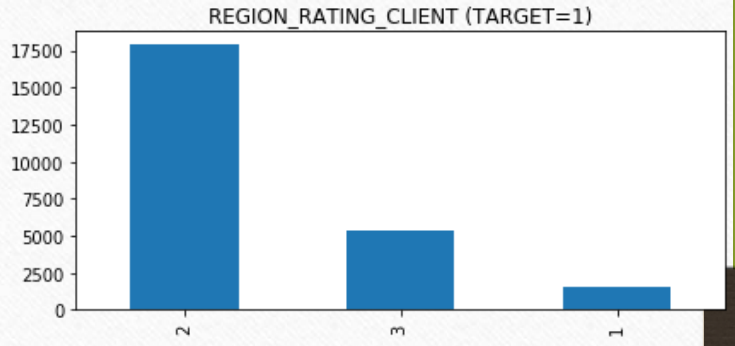




REGION_RATING_CLIENT

1. Clients with REGION_RATING_CLIENT = 2 show almost equal chances either to default or not.
2. Clients with REGION_RATING_CLIENT = 3 show a slight higher tendency to default.
3. Clients with REGION_RATING_CLIENT = 1 show a slight higher tendency to non-default.



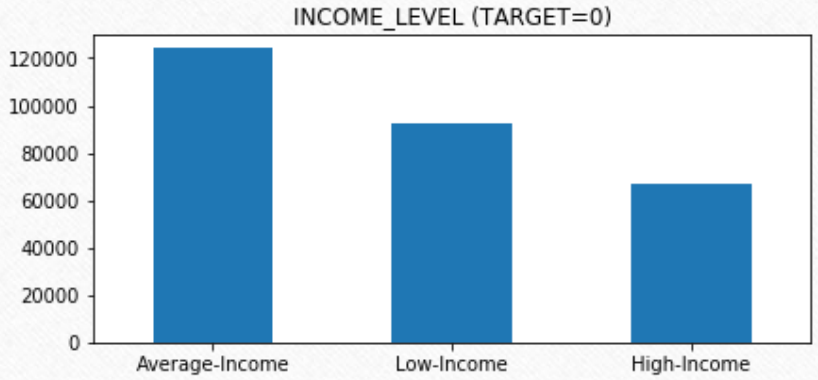


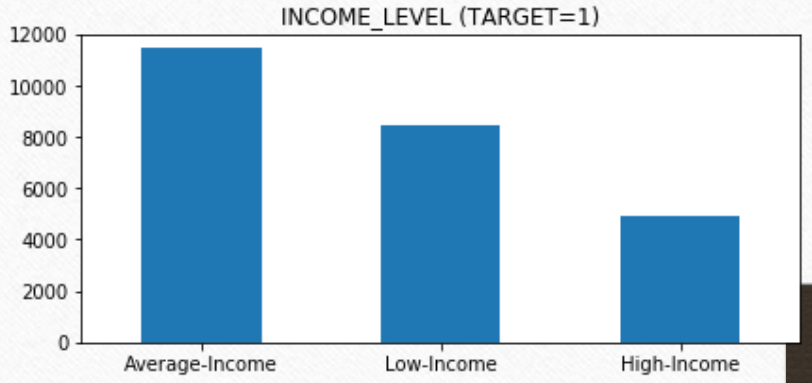
INCOME_LEVEL

1.Clients with Average-Income show almost equal chances either to default or not. Majority clients fall in this category.

2. Clients with Low-Income show almost equal chances either to default or not.

3. Clients with High-Income show a slight higher tendency to non-default.





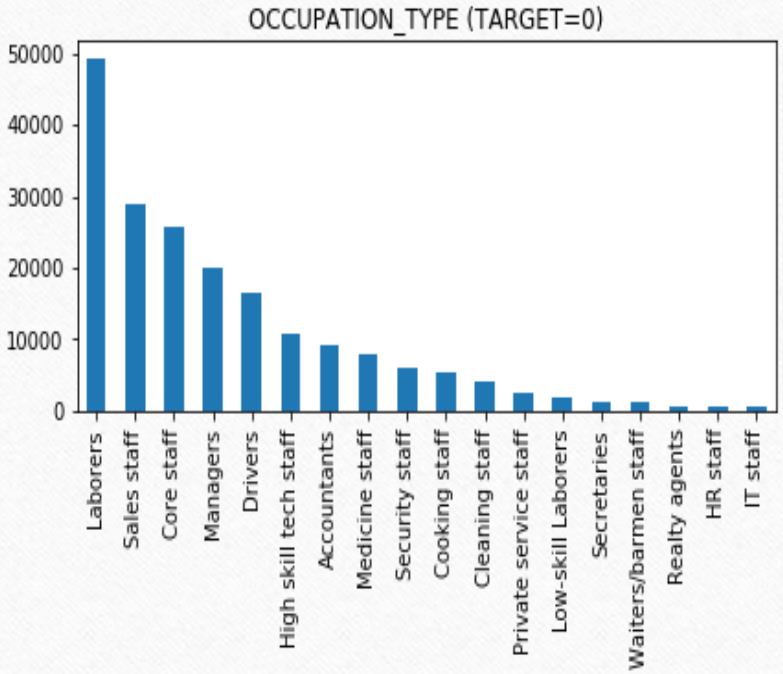
OCCUPATION_TYPE

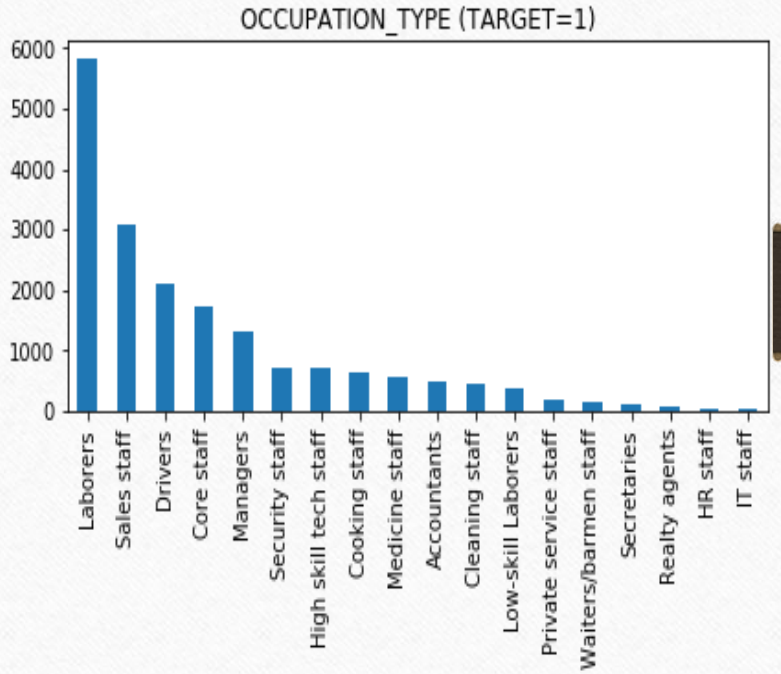
1. Laborers tend to default more. So should be more careful while approving for this category.

2. Core Staff, Managers, High Skill-tech staff tend to pay in time.

3. Accountants too tend to pay in time.

4. Secretaries tend to default slightly more.





TOP 10 CORRELATED VARIABLES FOR TARGET

The Top correlated variables for non-defaulters. (TARGET = 0)

1	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	100%
2	AMT_GOODS_PRICE	AMT_CREDIT	99%
3	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	95%
4	CNT_FAM_MEMBERS	CNT_CHILDREN	88%
6	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	86%
5	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	86%
7	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	83%
8	AMT_ANNUITY	AMT_GOODS_PRICE	78%
9	AMT_CREDIT	AMT_ANNUITY	77%
10	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	45%

The Top correlated variables for defaulters. (TARGET = 1)

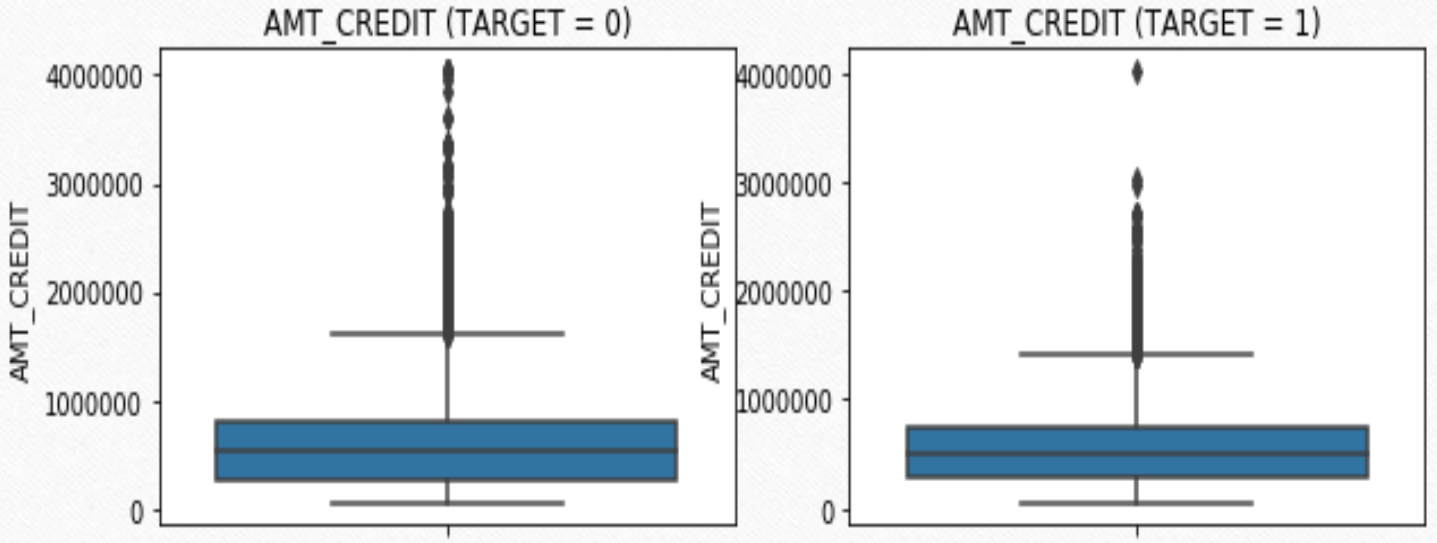
1	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	100%
2	AMT_GOODS_PRICE	AMT_CREDIT	97%
3	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	96%
4	CNT_FAM_MEMBERS	CNT_CHILDREN	89%
5	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	85%
6	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	85%
7	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	78%
8	AMT_ANNUITY	AMT_GOODS_PRICE	75%
9	AMT_CREDIT	AMT_ANNUITY	75%
10	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	50%

For both defaulters and non-defaulters, the most correlated variables are the same. Only the rate of correlations between the variables differ. For both type of clients, "the number of observation of client's social surroundings defaulted on 30 DPD" is directly related to "the number of observation of client's social surroundings with observable 60 DPD".

Univariate analysis for continuous variables

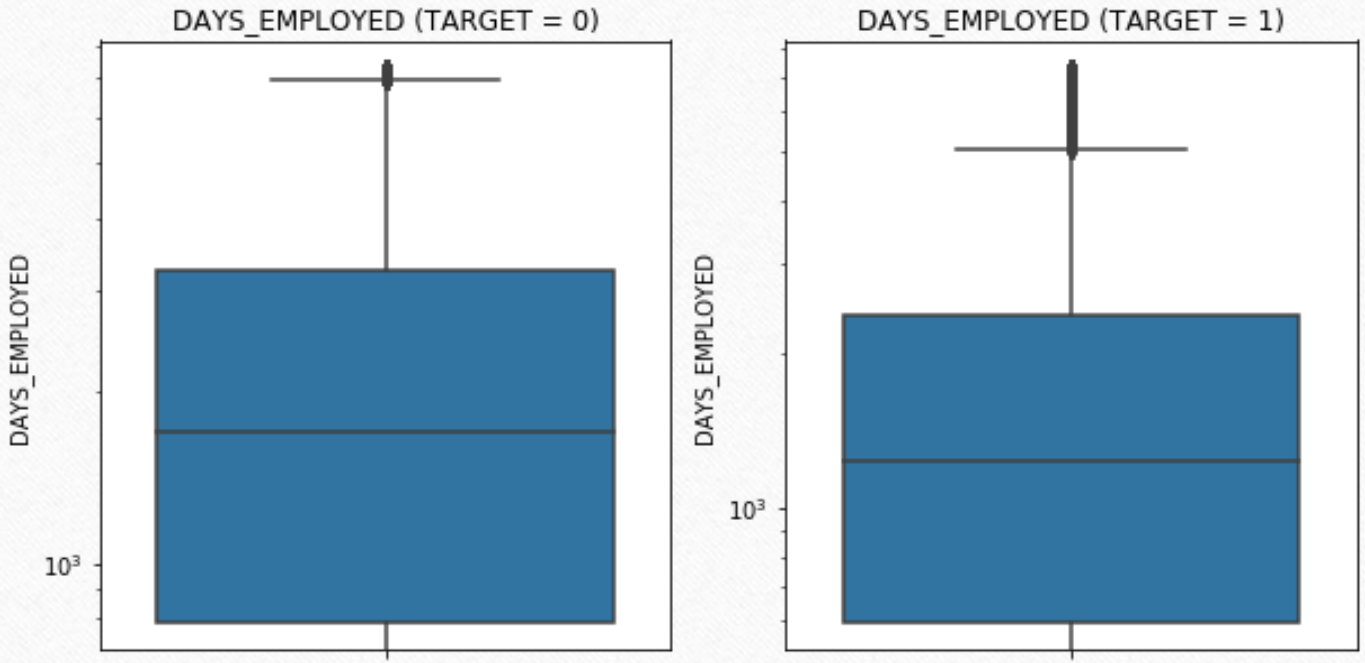
AMT_INCOME_TOTAL

It indicates income of the client. **Clients with very high income have less chances to default. Their loan application can surely be approved. The clients with lower income have higher chances to default.**



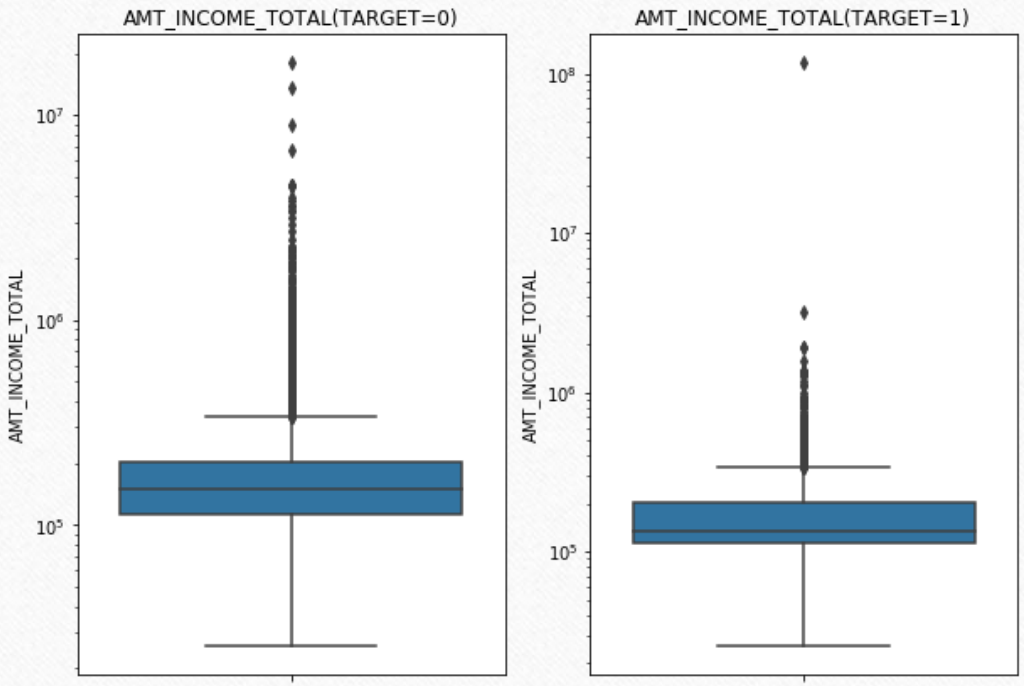
DAYS_EMPLOYED

We don't find any significant inference from the above boxplots.
The DAYS_EMPLOYED is slightly higher for non-defaulters than the defaulters.



AMT_INCOME_TOTAL

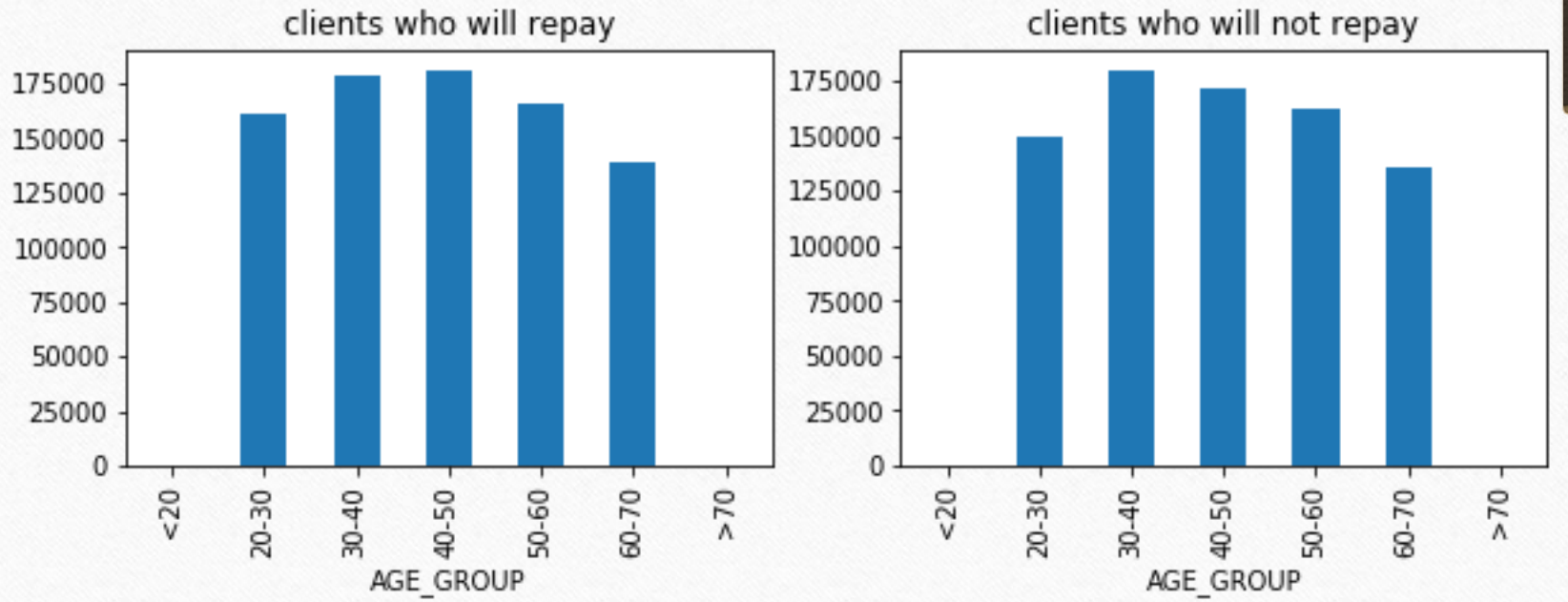
clients with very high income have less chances to default. Their loan application can surely be approved. The clients with lower income have higher chances to default.



Bivariate analysis of categorical-continuous variables

AGE_GROUP -- AMT_INCOME_TOTAL

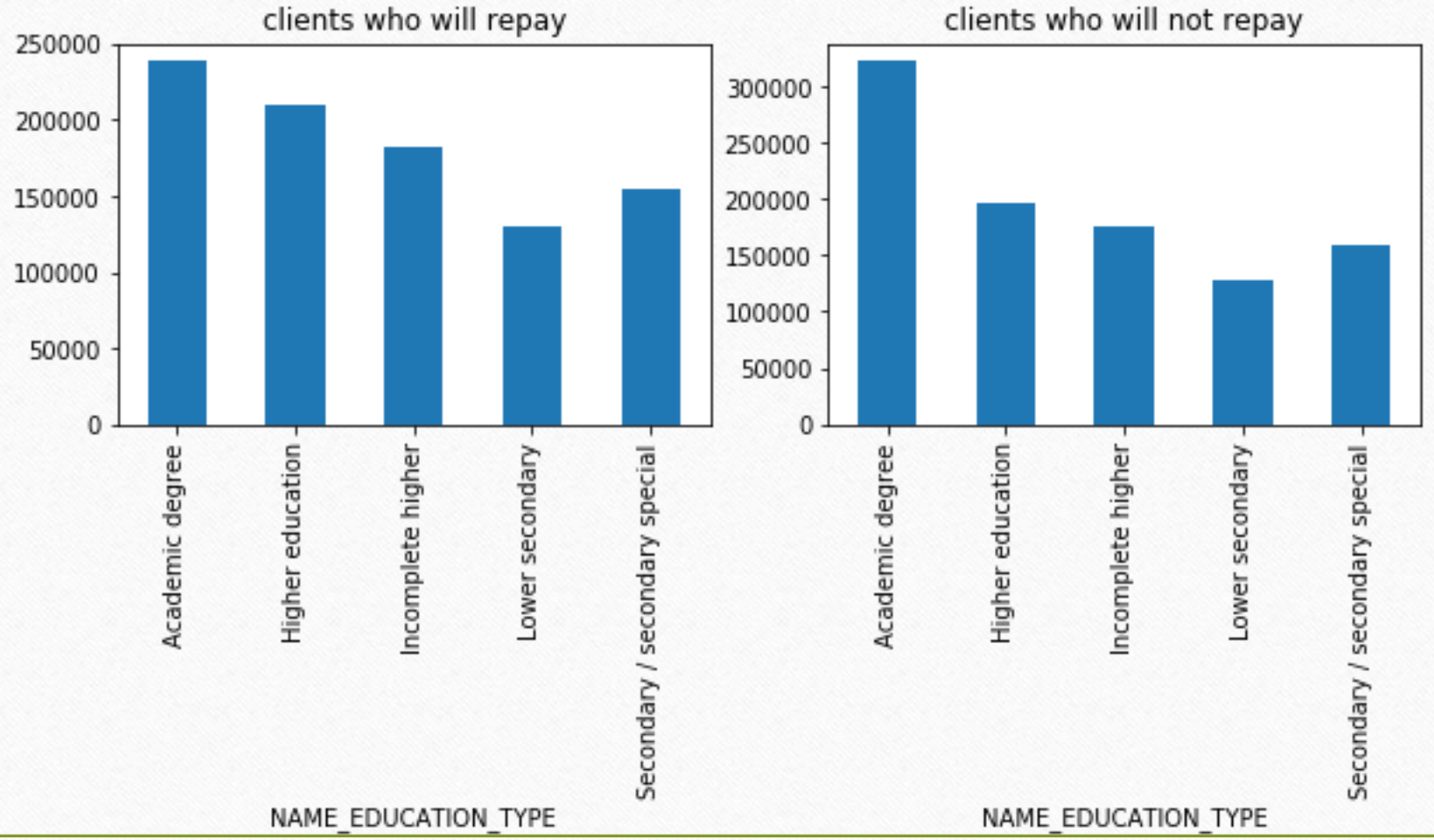
Age of 20-40 years are more likely to default



NAME_EDUCATION_TYPE -- AMT_INCOME_TOTAL

1. The clients with an academic degree and an income level above 300000 are more likely to default.

2. The clients with either higher education or incomplete higher education are more likely to repay the loan

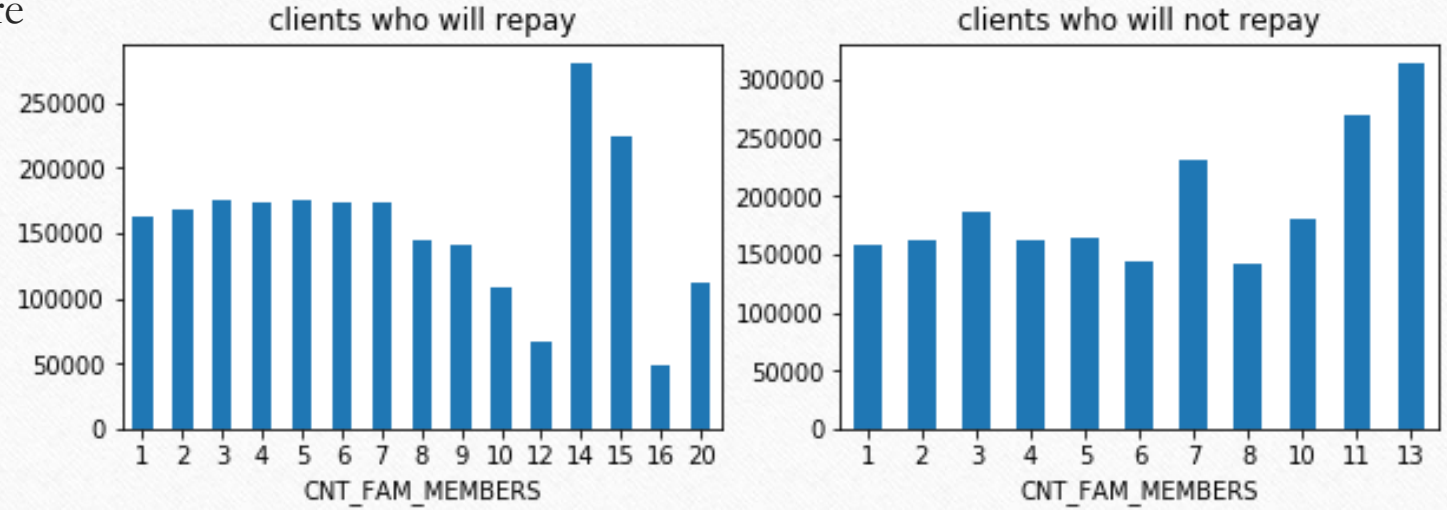


CNT_FAM_MEMBERS -- AMT_INCOME_TOTAL

1. The clients with family members 14 or more and irrespective of income are more likely to repay the loan.

2. The clients with family members between 8 and 11(inclusive) and have an income of 140000 or more tend to be defaulters.

3. The clients with family members equal to 13 are not repaying.



Bivariate analysis of categorical-categorical variables

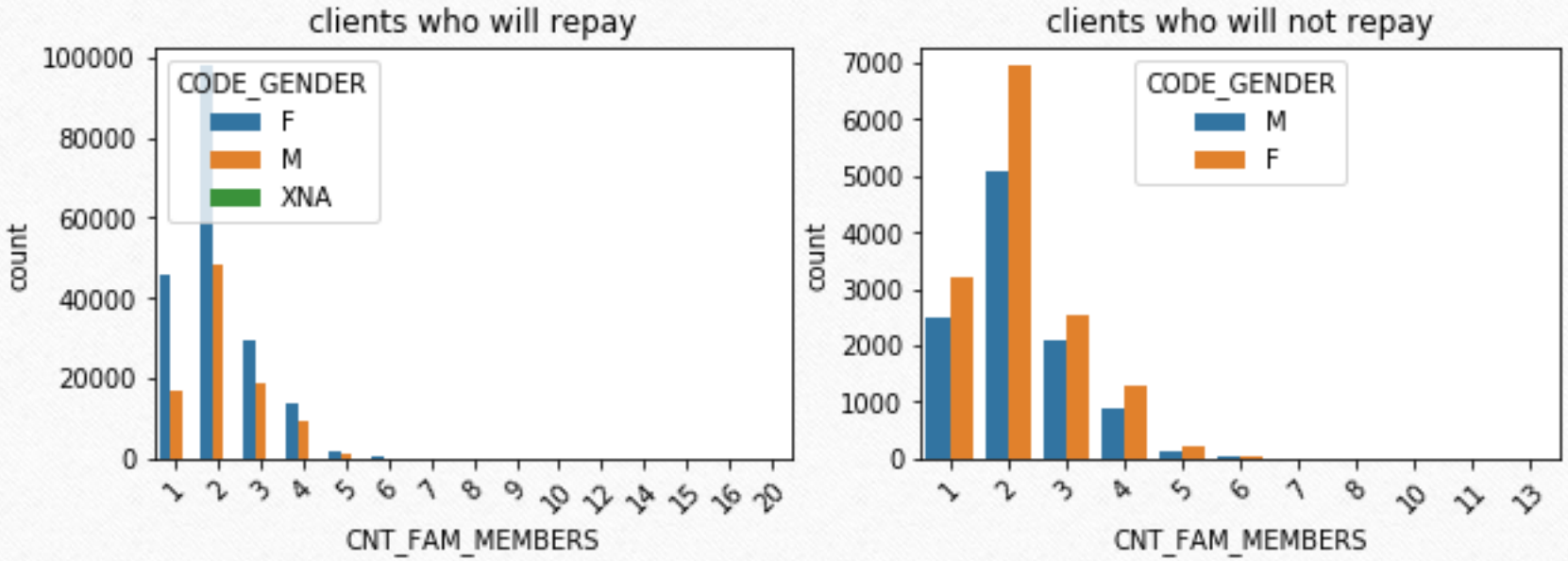
NAME_EDUCATION_TYPE -- INCOME_LEVEL

The clients with higher education have more tendency to repay.

CNT_FAM_MEMBERS -- CODE_GENDER

A female client having 2 family members are

more likely to default



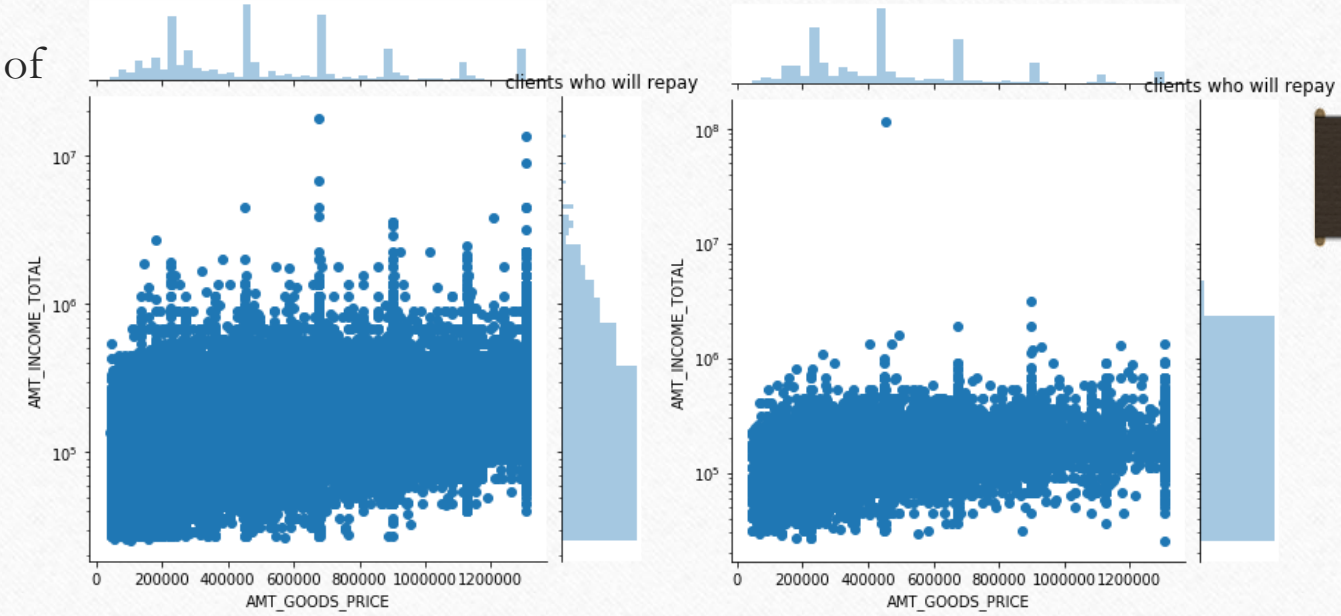
NAME_EDUCATION_TYPE -- INCOME_LEVEL

Higher education has a greater tendency to repay.

Bivariate analysis of Continuous-Continuous variables

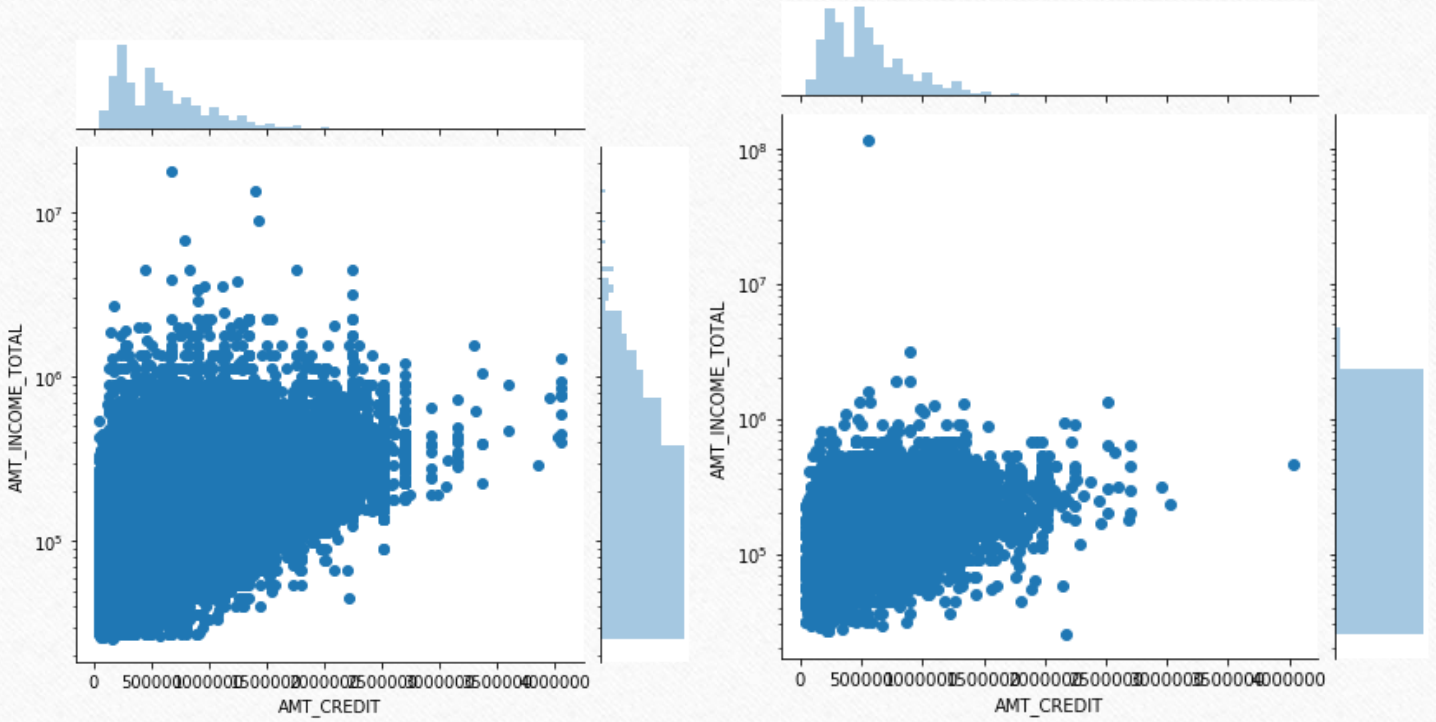
AMT_INCOME_TOTAL and AMT_GOODS_PRICE

The income level is well correlated with price of seller goods. So we can say that clients with high income are more having higher good price and vice-versa.



AMT_CREDIT and AMT_INCOME_TOTAL

the client income and the amount credit are well correlated.

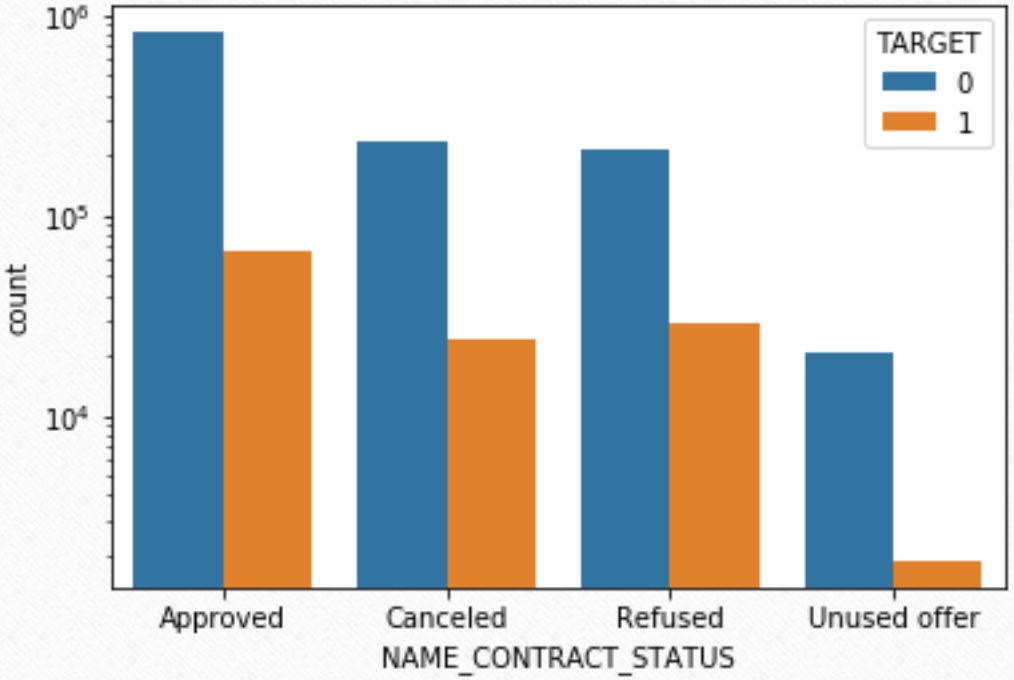


ANALYSIS OF MERGED DATA

NAME_CONTRACT_STATUS against TARGET

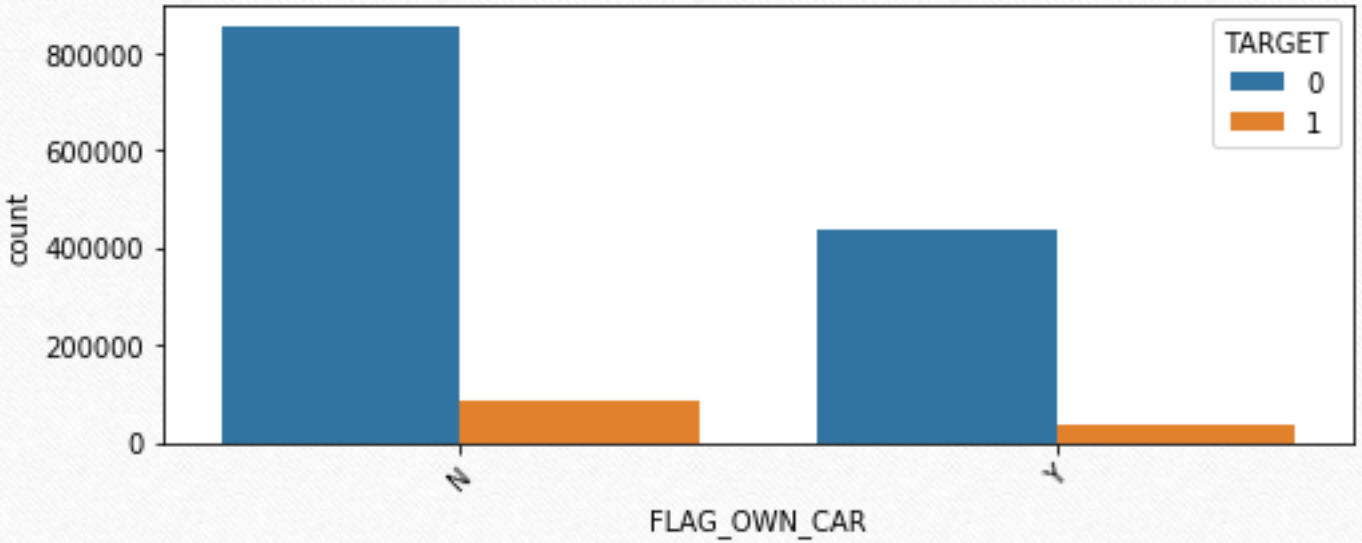
Of the total approved loans,

approximately 8% **are defaulted.**



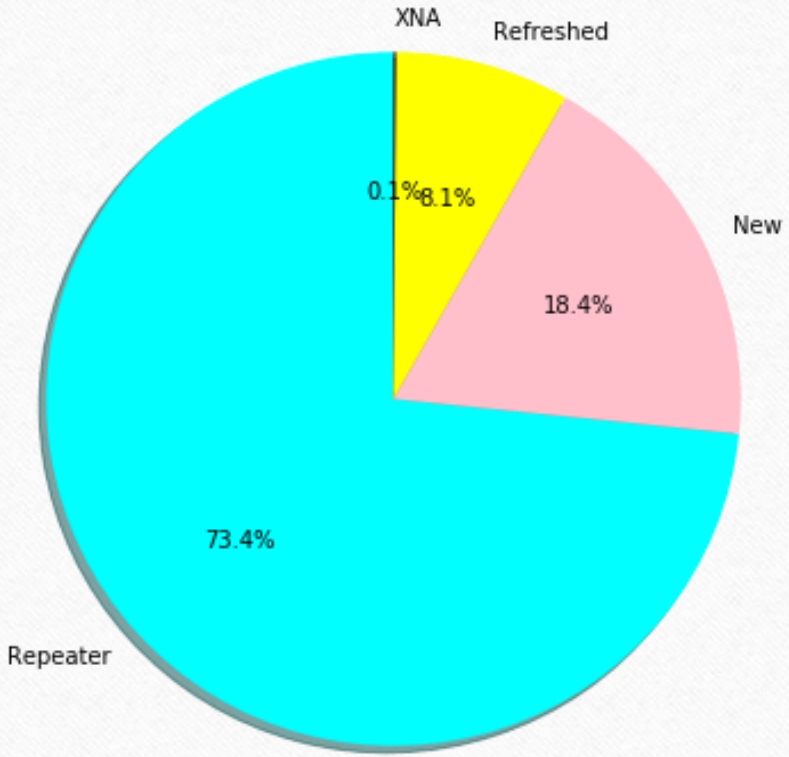
FLAG_OWN_CAR against TARGET

The clients who owns are a car are approximately half of who doesn't have a car. The one who owns a car are more likely to repay.



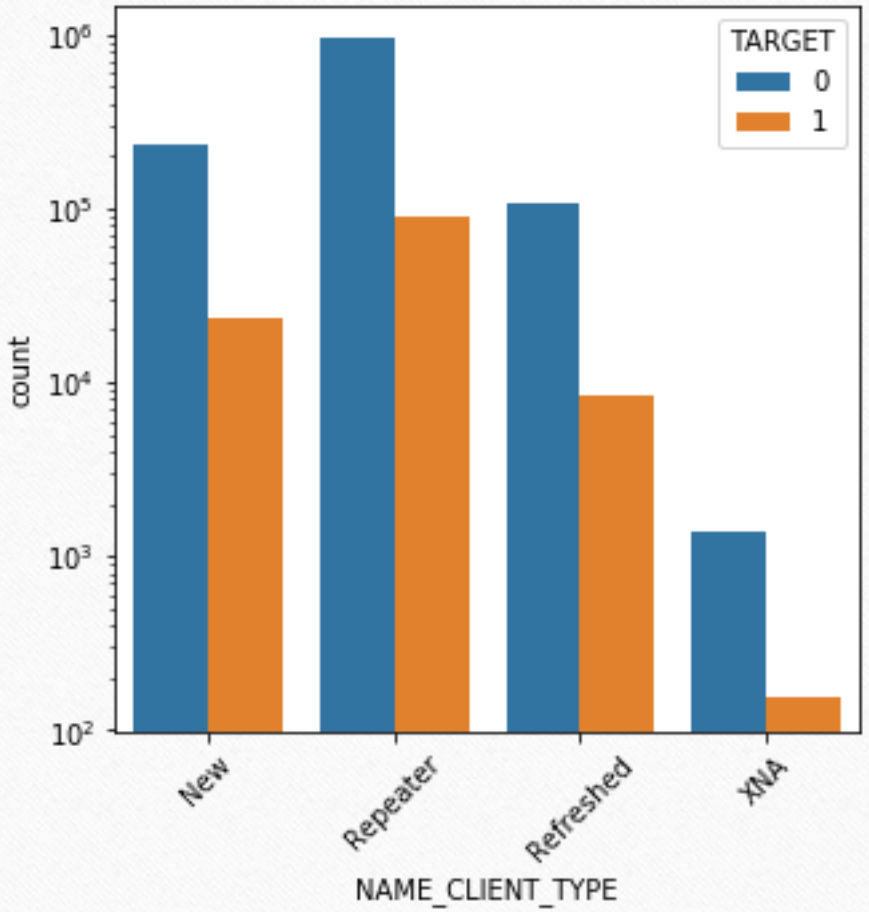
NAME_CLIENT_TYPE
It indicates whether the client old or new client when applying for the previous application

Above 70% clients are repeaters. Need to focus more on such category.



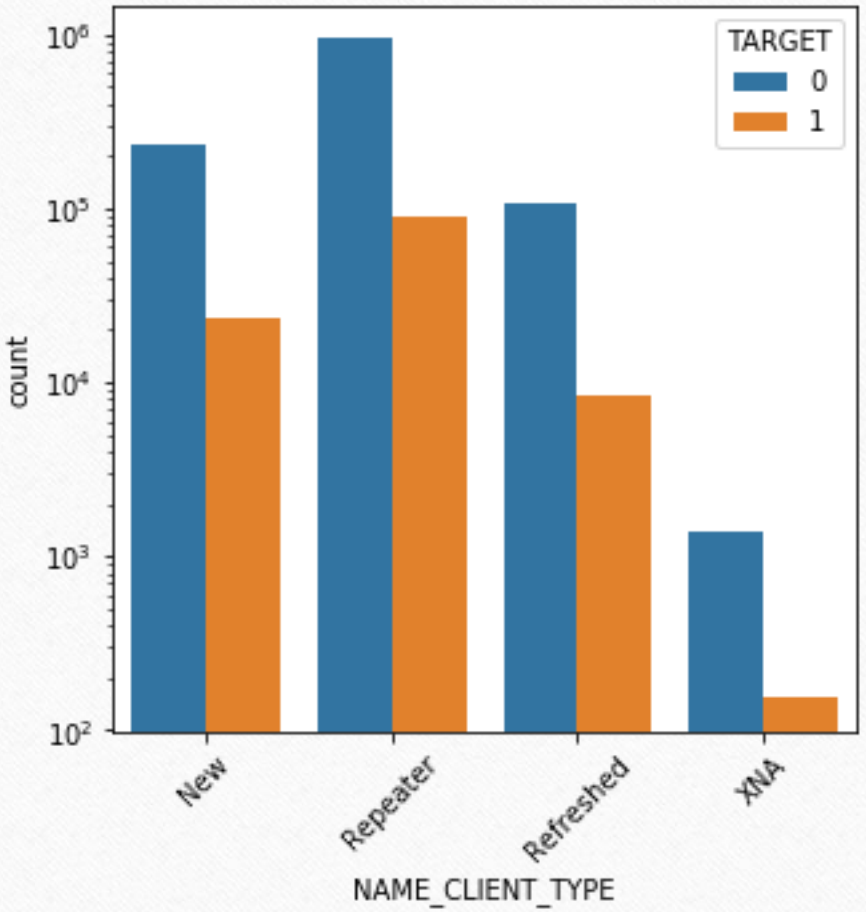
NAME_CLIENT_TYPE against TARGET

Clients the defaulters are approximately 10% less than those of non-defaulters



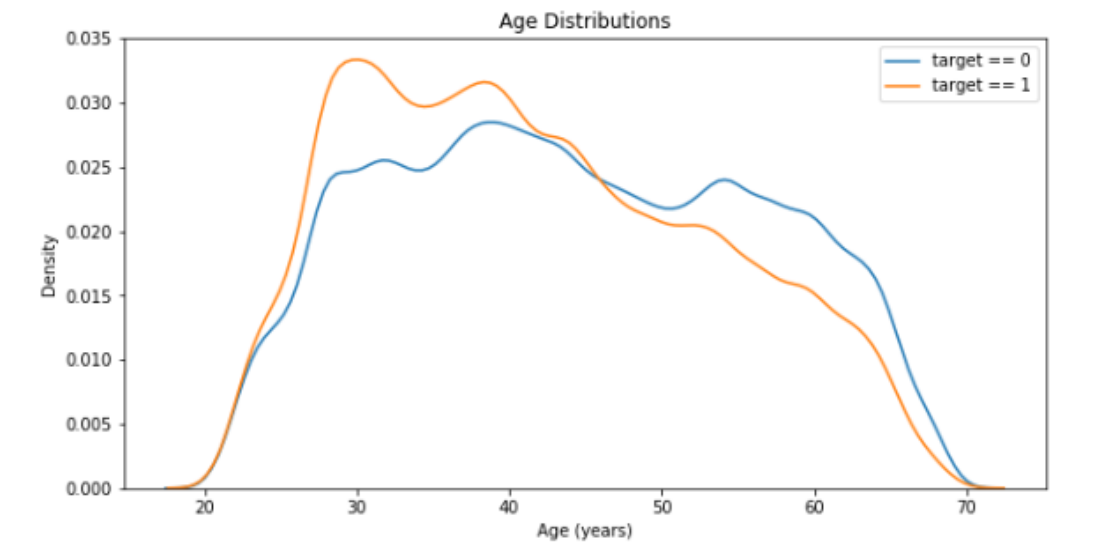
NAME_CLIENT_TYPE against TARGET

Clients the defaulters are approximately 10% less than those of non-defaulters.



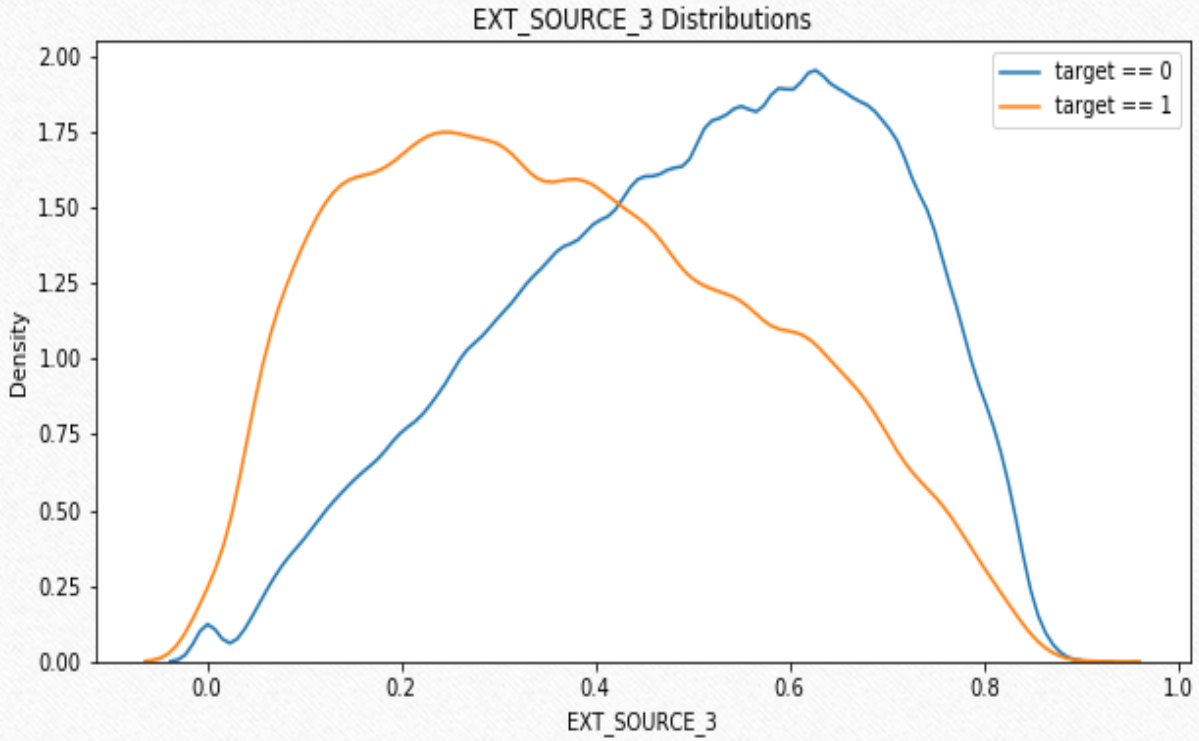
ANALYSIS OF AGE

The pattern clearly infers, "*Younger clients are more likely to default than elder once*".

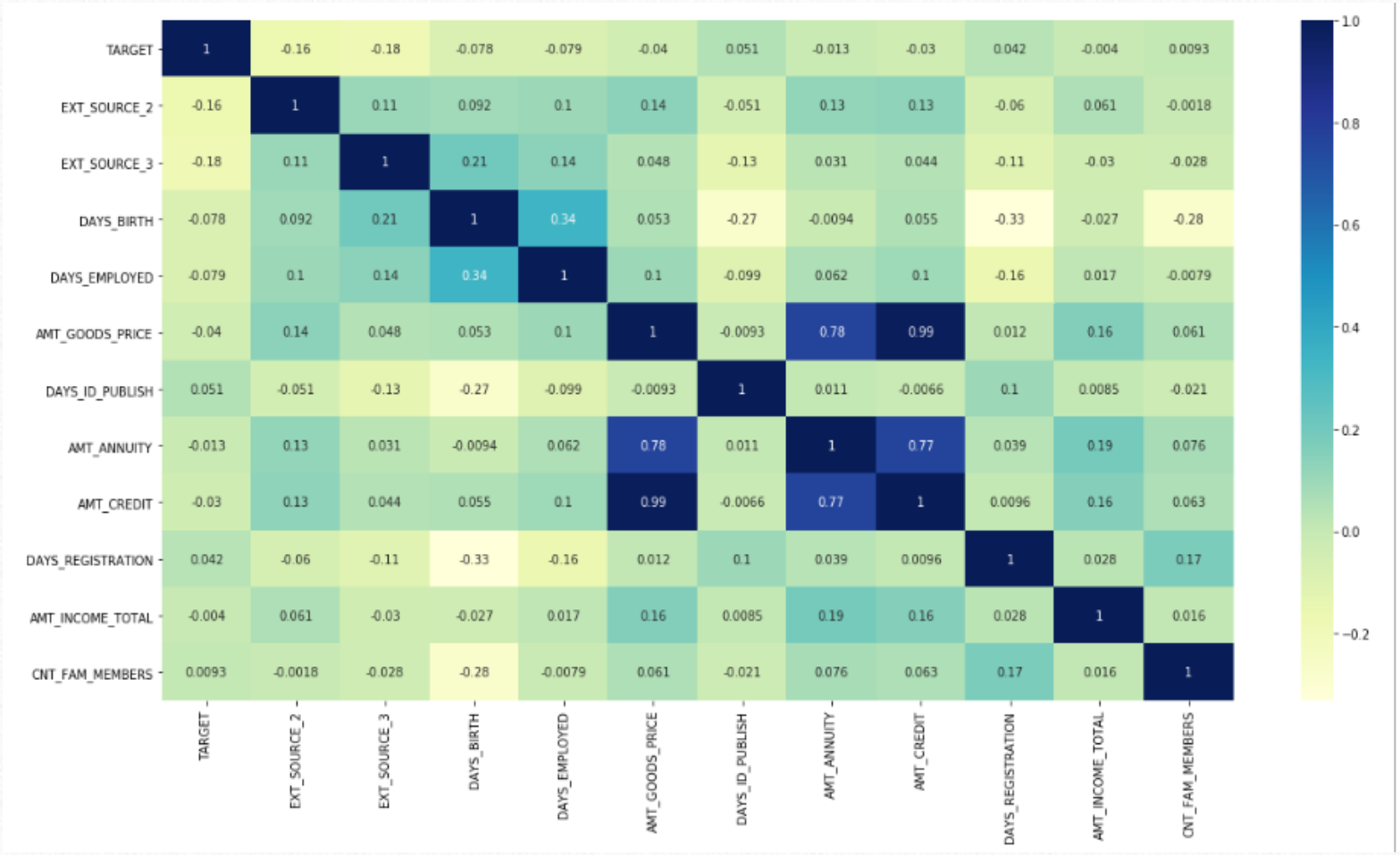


EXT_SOURCE_3 against TARGET

The clients having EXT_SOURCE_3 less than 0.4 will most likely pay and greater than 0.4 are more likely to default.

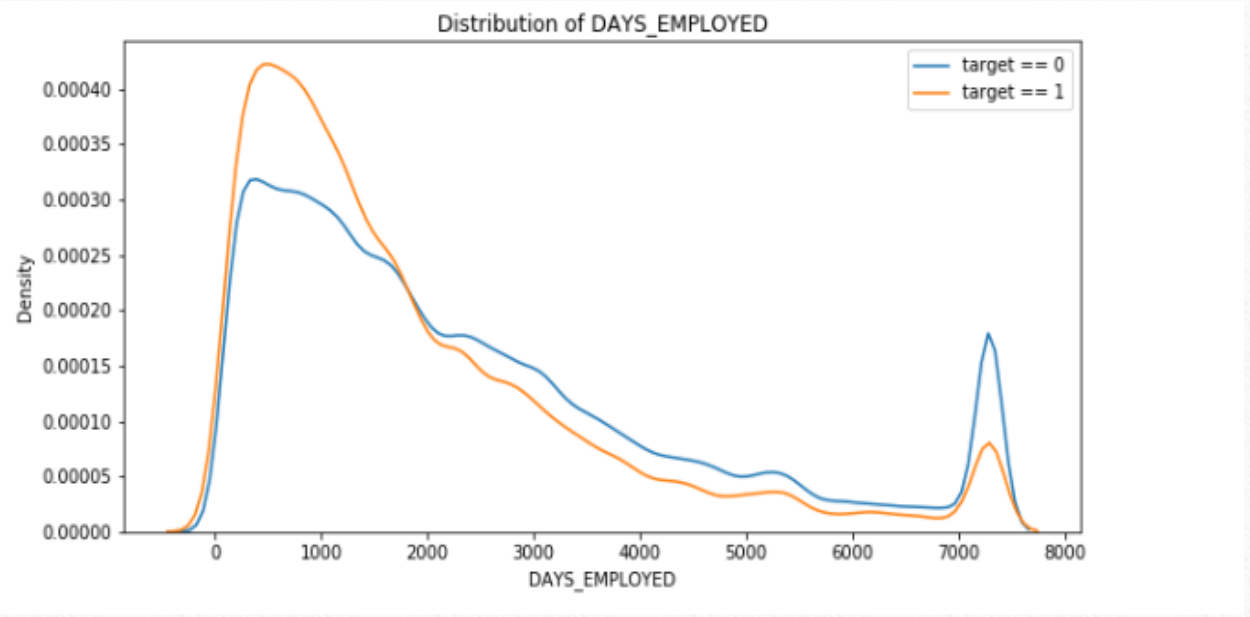


Correlation of Columns



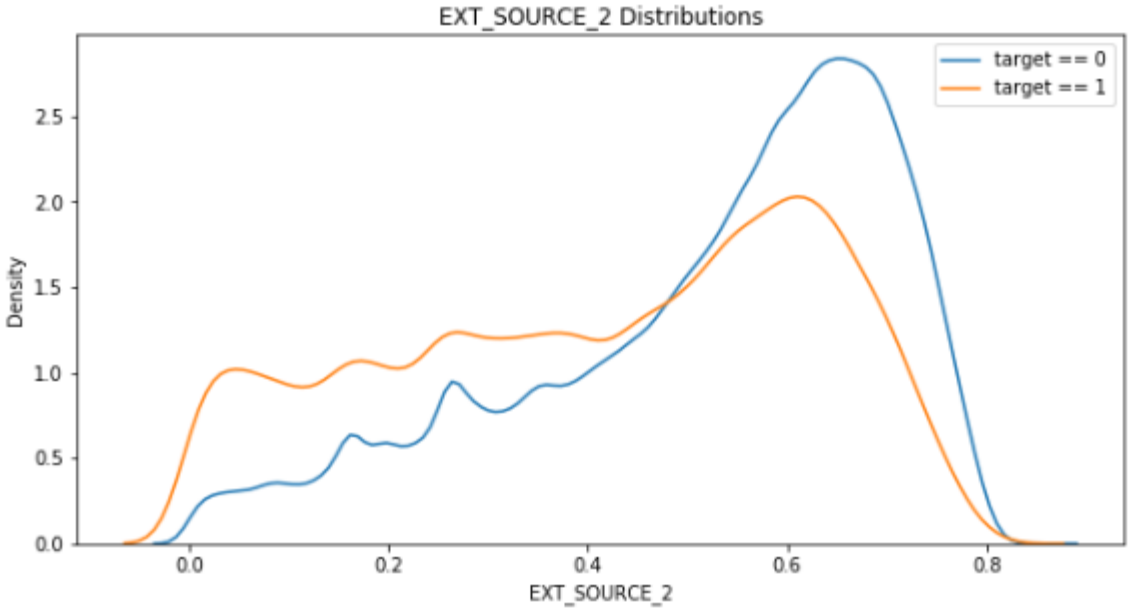
Analysis on DAYS_EMPLOYED

New joined employees are more likely of not paying the loan than the old employees of an organization.



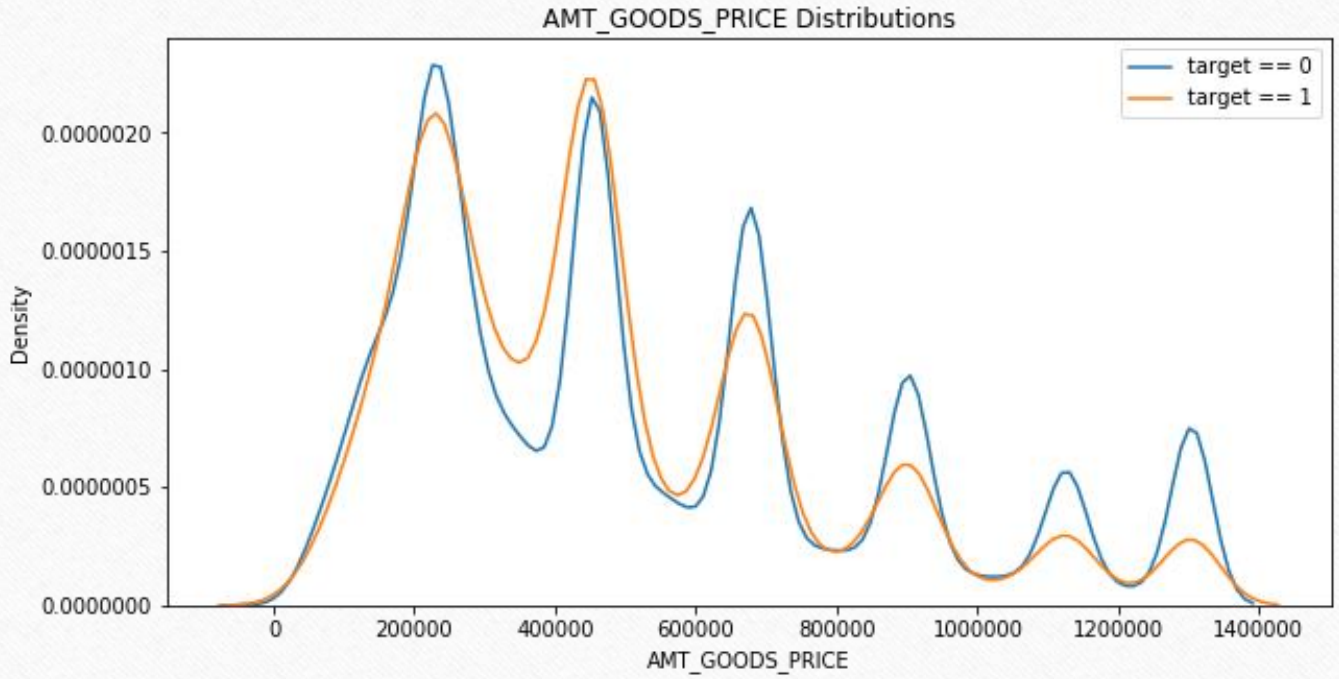
Analysis of EXT_SOURCE_2

"EXT_SOURCE_2"(Normalized score from external data source) is important feature we need to look upon. As we can notice that a client with higher values of value of EXT_SOURCE_2 is more likely not to default than once with low value.



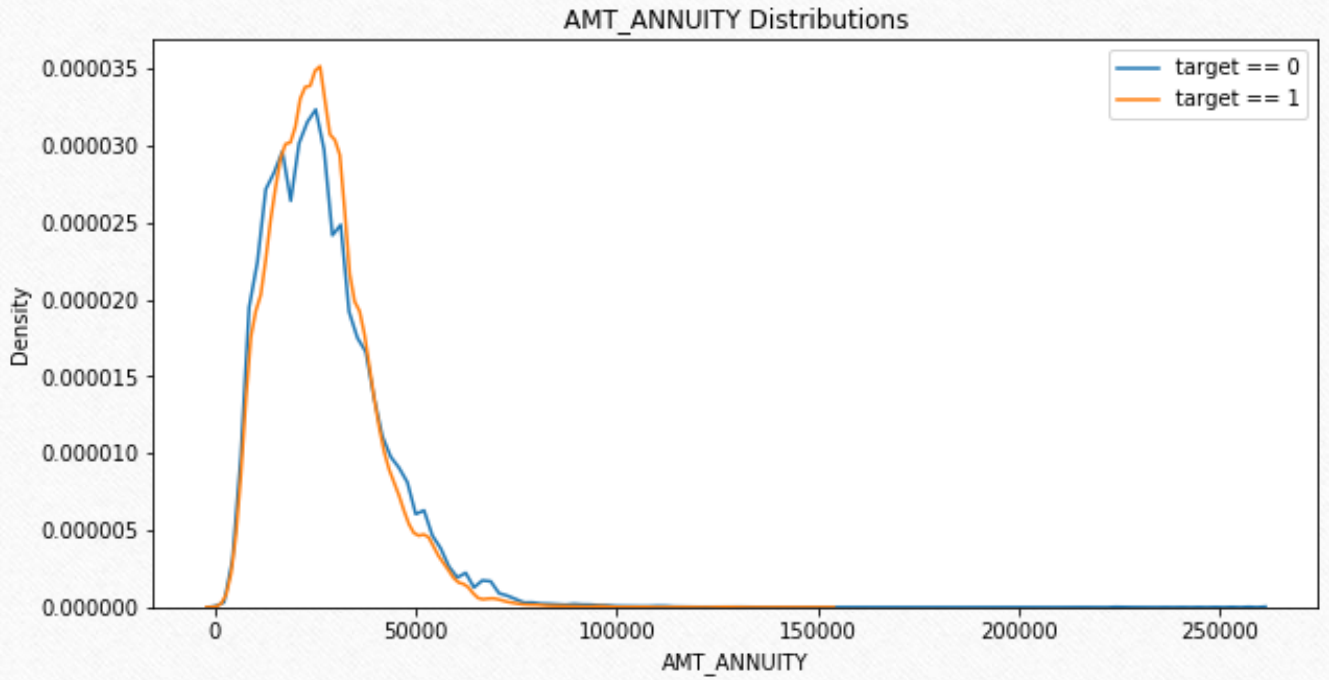
Analysis of AMT_GOODS_PRICE

The goods price increases they clients tends repay the loan.



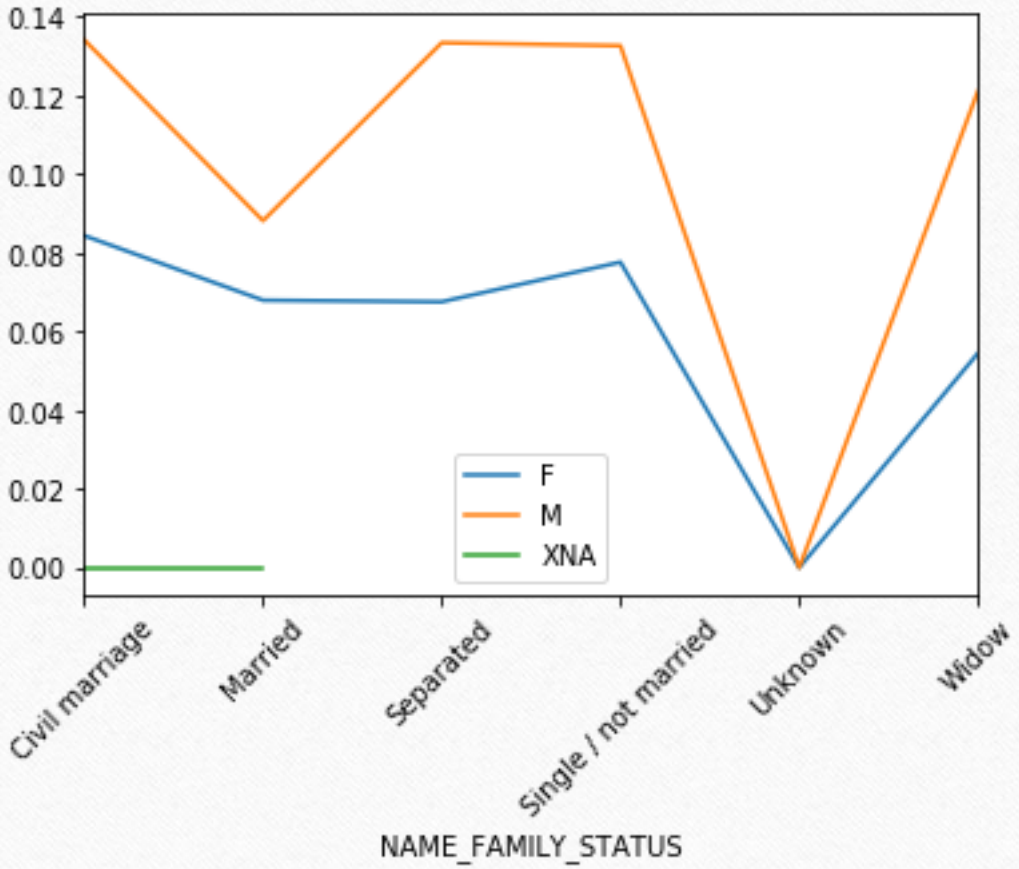
Analysis on AMT_ANNUITY

The clients with an installment of less than 50000 are more likely to default.



Analysis on NAME_FAMILY_STATUS and CODE_GENDER

Its is most likely that Males whose Family status is "Civil Marriage, Separated, Single/Not married" are more likely to default than females.



FINAL WORDS

- Variables like **EXT_SOURCE_2, EXT_SOURCE_3,DAYS_BIRTH**, **DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, AMT_GOODS_PRICE, AMT_ANNUITY** and **AMT_INCOME_TOTAL** are inversely proportional to the TARGET (1 representing a Defaulter and 0 Non-Defaulter) The above variables act as strong indicators for loan defaulting/non-defaulting.
- Besides these we also noticed another observation that higher count of family members increases probability to default