Case study for credit EDA

By Pham Minh Man

Data understanding

Basic checking

- Checking shape, info, d-type, describe, meaning of all column in the dataset
- Checking null %

Sanity

- Check data contain to identify which fields contain wrong data (“date" column contain negative value, income type have wrong value)

Data cleaning and manipulation

Data cleaning

- Checking null percent, drop all column which have null percent > 40%
- Identify fields which have many outlier (list column)
- Drop column which contain wrong data (“date")

Data manipulation

- For other fields we can replace the null value which mean/median value
- For the category fields, we fill the missing values with the common value in that column.

Data Analysis

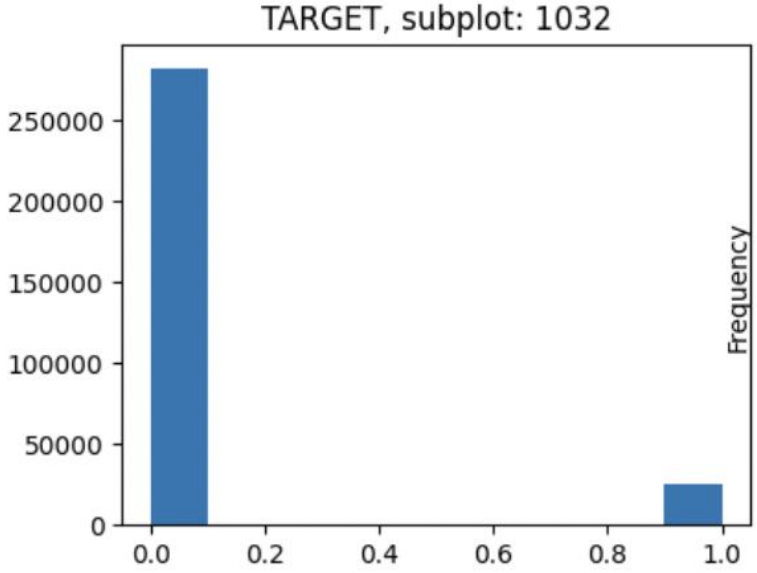
Data distribution

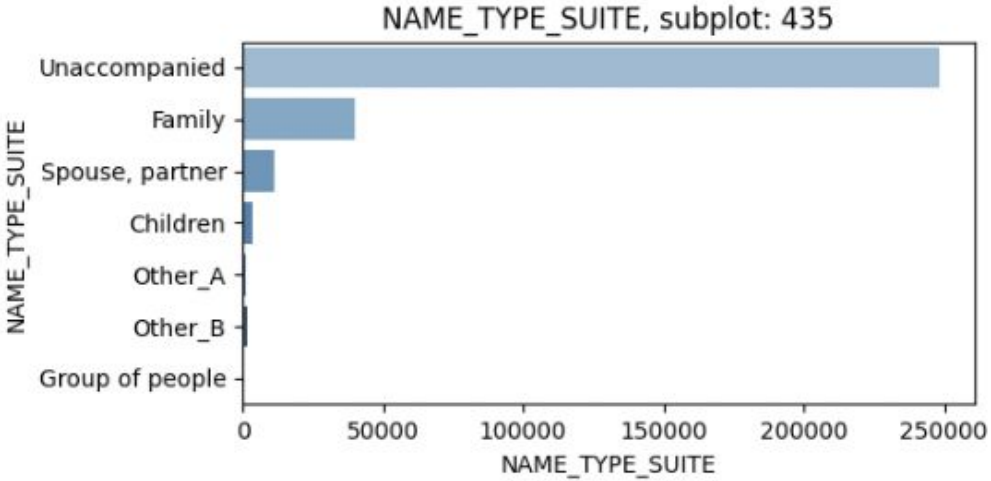
Bivariate analysis

Insight from the data

Data distribution

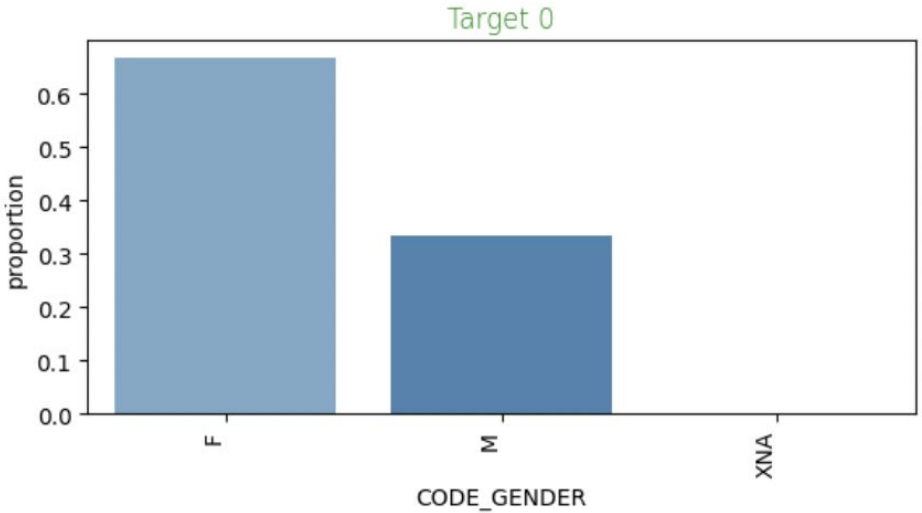
Many fields is highly imbalance like, and most importantly field “TARGET" is also imbalance, and many other fields like “NAME_CONTRACT_TYPE", “NAME_HOUSING_TYPE





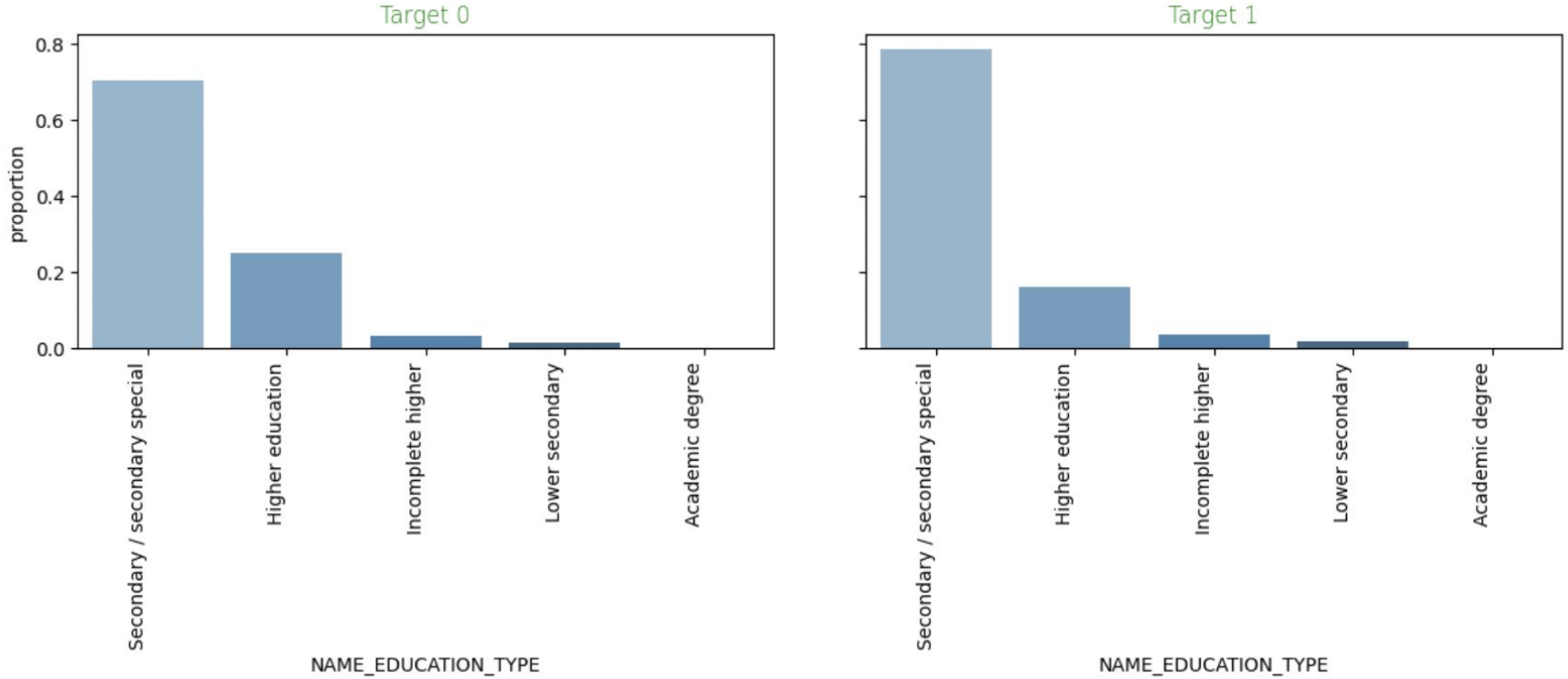
Bivariate analysis

Male have higher proportion in defauler



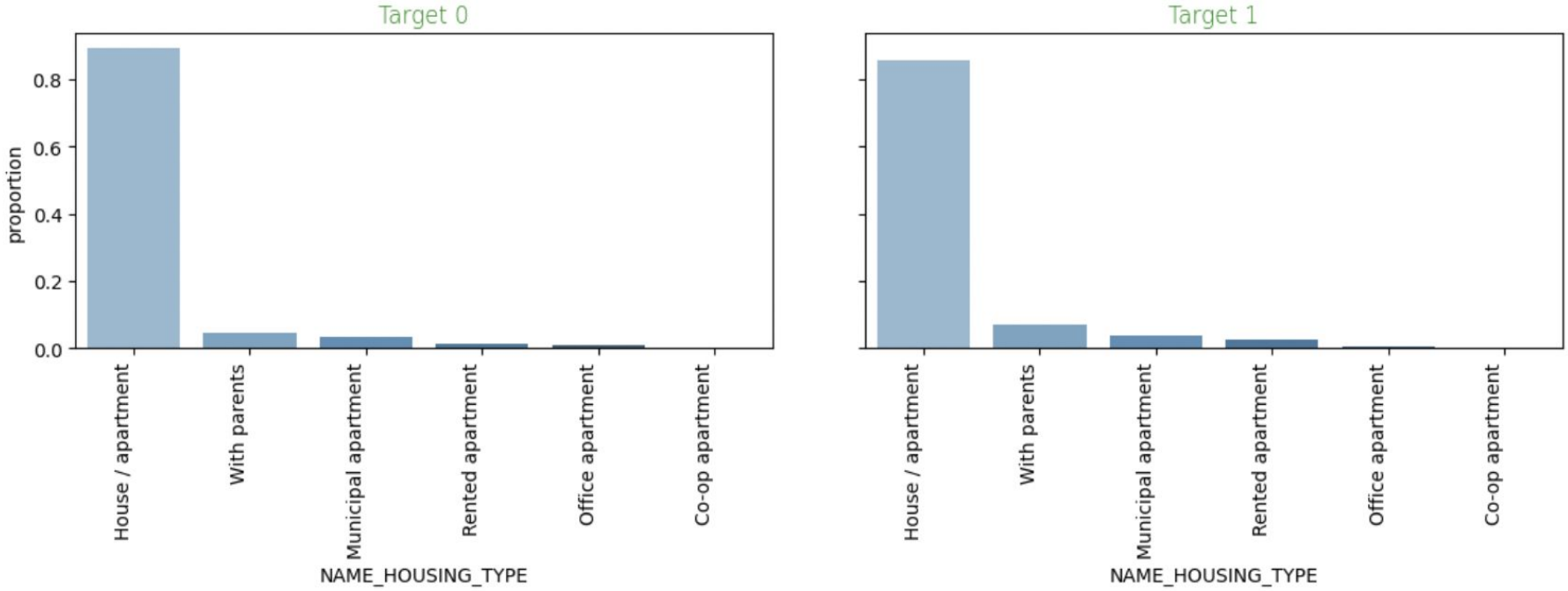
Bivariate analysis

People have higher education is less likely to default compare to secondary



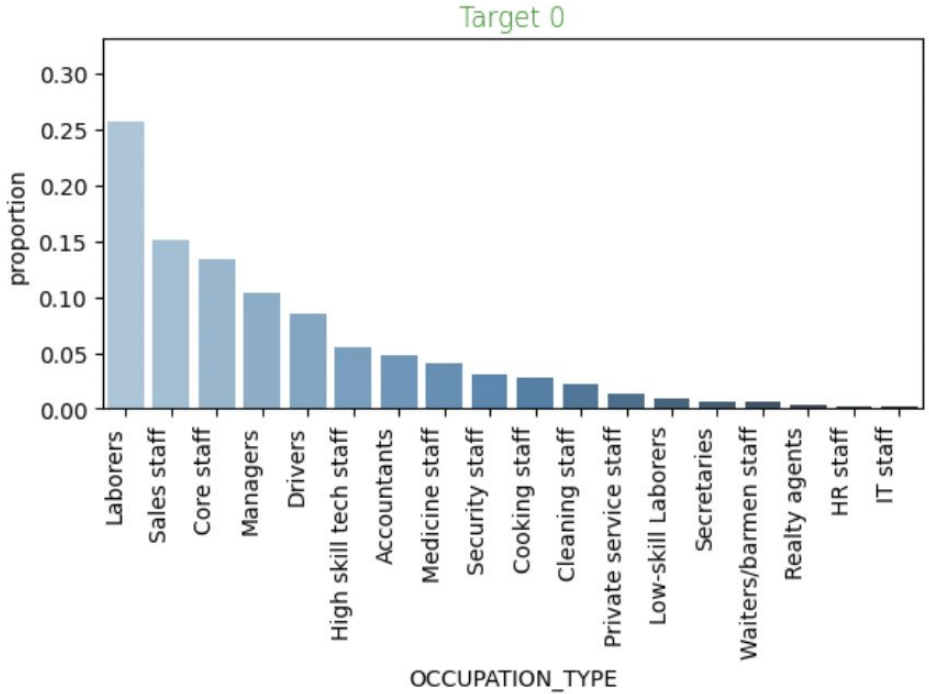
Bivariate analysis

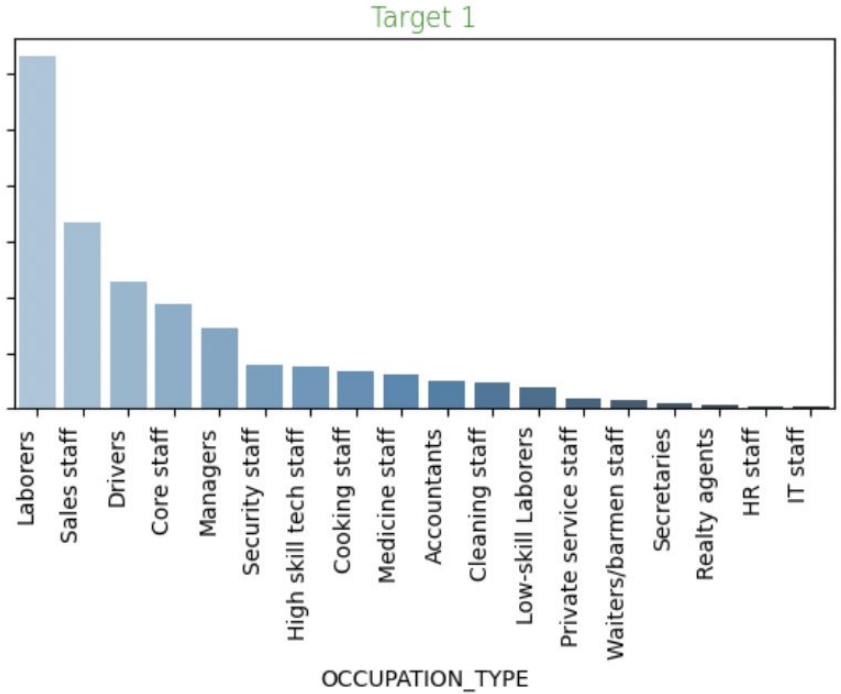
People with parent have higher chance to be default



Bivariate analysis

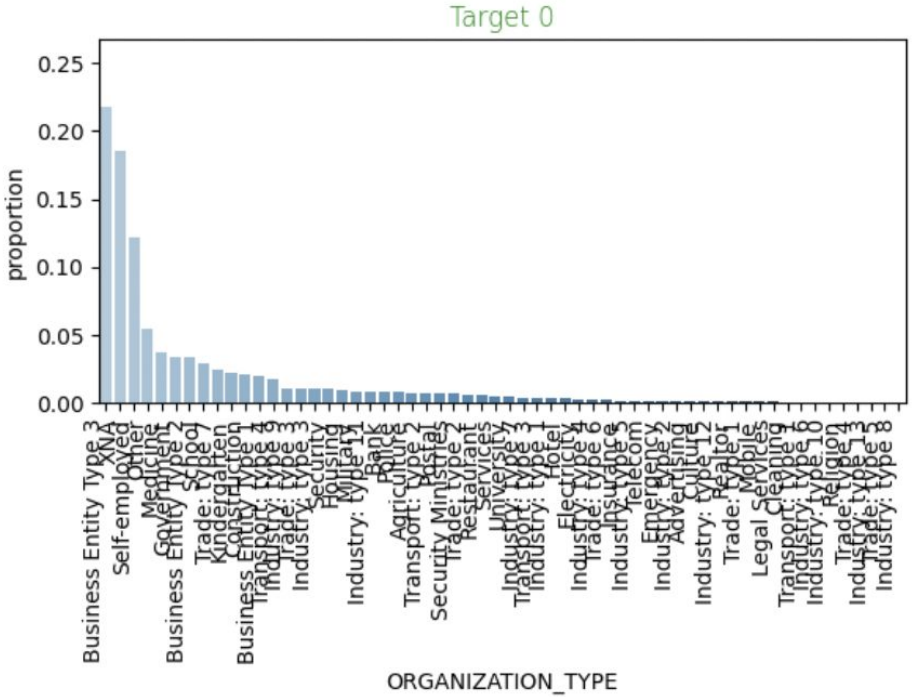
People occupation type as “Laborers" have higher chance to be defaulter compare to non-defaulter

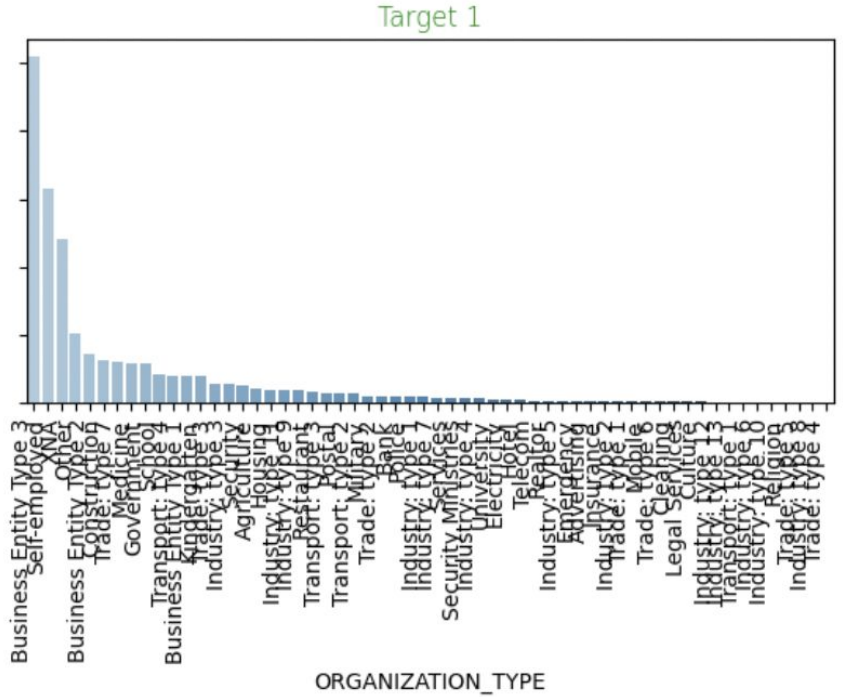




Bivariate analysis

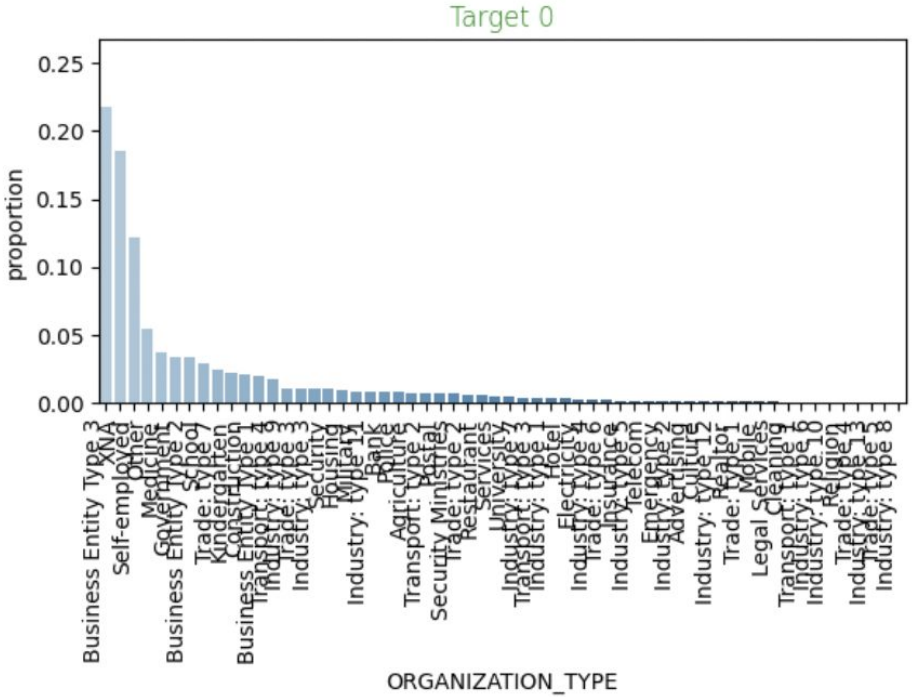
People have organization type as Business Entity Type 3 have more risk to be defauler

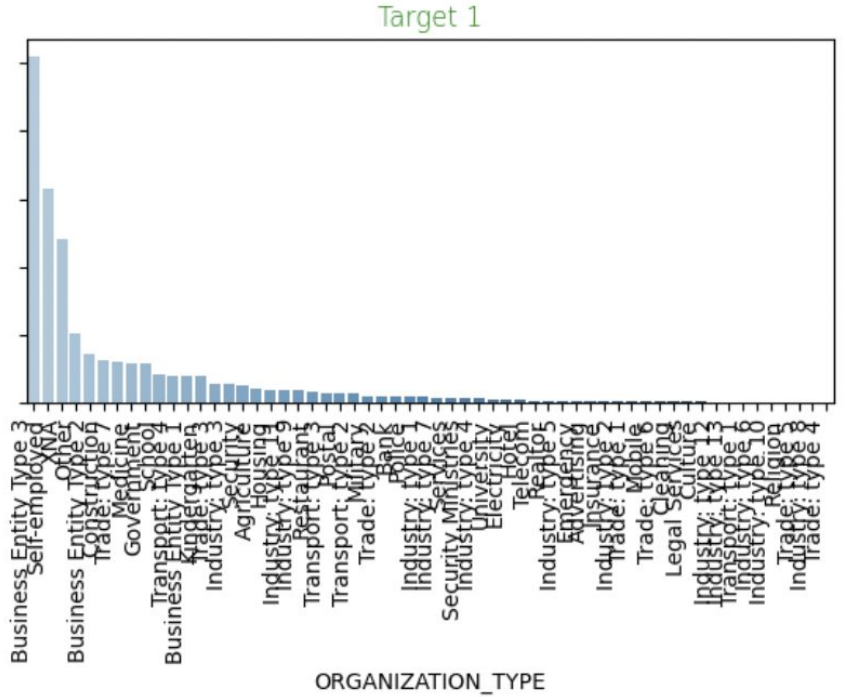




Bivariate analysis

People have organization type as Business Entity Type 3 have more risk to be defauler





Multivariates Analysis between Numerical and Categorical

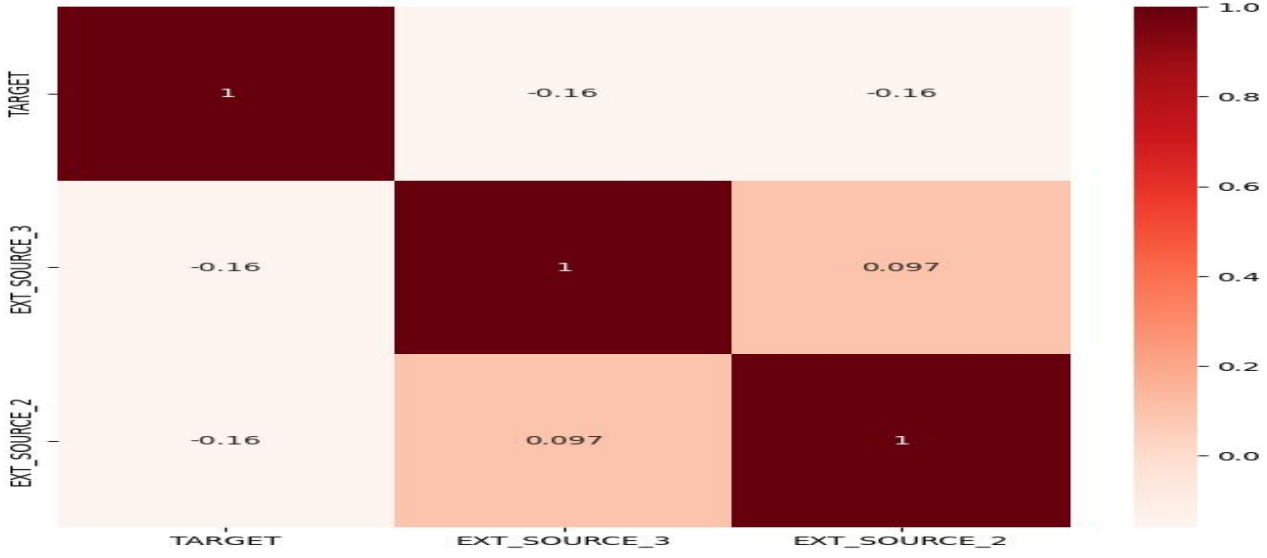
Bivariate analysis

First of all, we could using the quantify the correlations between the numerical columns and target to get the idea. the positive effect is not strong enough to be compare with the negative one. So from now we focus more on the last 2 ones.

```
[{'TARGET': 1.0},
 {'REGION_RATING_CLIENT_W_CITY': 0.060892667564823415},
 {'REGION_RATING_CLIENT': 0.05889901494571238},
 {'DAYS_LAST_PHONE_CHANGE': 0.05521847562884947},
 {'REG_CITY_NOT_WORK_CITY': 0.050994464368126 67},
 {'FLAG_EMP_PHONE': 0.04598221971659251},
 {'REG_CITY_NOT_LIVE_CITY': 0.04439537480570111},
 {'FLAG_DOCUMENT_3': 0.0443463468511448 65},
 {'LIVE_CITY_NOT_WORK_CITY': 0.0325183411014988},
 {'DEF_30_CNT_SOCIAL_CIRCLE': 0.03222153390591371},
 {'DEF_60_CNT_SOCIAL_CIRCLE': 0.03125121111173646},
 {'FLAG_WORK_PHONE': 0.0285243223632175 02},
 {'CNT_CHILDREN': 0.019187133596269994},
 {'AMT_REQ_CREDIT_BUREAU_YEAR': 0.018160138737084826},
 {'CNT_FAM_MEMBERS': 0.009307781738417764},
 {'OBS_30_CNT_SOCIAL_CIRCLE': 0.009123291153639138},
 {'OBS_60_CNT_SOCIAL_CIRCLE': 0.009014856730913 99},
 {'REG_REGION_NOT_WORK_REGION': 0.006941907545371621},
 {'REG_REGION_NOT_LIVE_REGION': 0.005575944520908466},
 {'FLAG_DOCUMENT_2': 0.005417144279619333},
 {'FLAG_DOCUMENT_21': 0.003708625029306517},
 {'LIVE_REGION_NOT_WORK_REGION': 0.002819479184158924},
 {'AMT_REQ_CREDIT_BUREAU_DAY': 0.0024642575387485936},
 {'AMT_REQ_CREDIT_BUREAU_HOUR': 0.0008478053136984312},
 {'AMT_REQ_CREDIT_BUREAU_WEEK': 0.0007177649853892554},
...
 {'AMT_CREDIT': -0.030369286461429 65},
 {'REGION_POPULATION_RELATIVE': -0.03722714854244522},
 {'AMT_GOODS_PRICE': -0.03962840801665482},
 {'EXT_SOURCE_3': -0.15739659264729502},
 {'EXT_SOURCE_2': -0.160303152496840 82}]
```

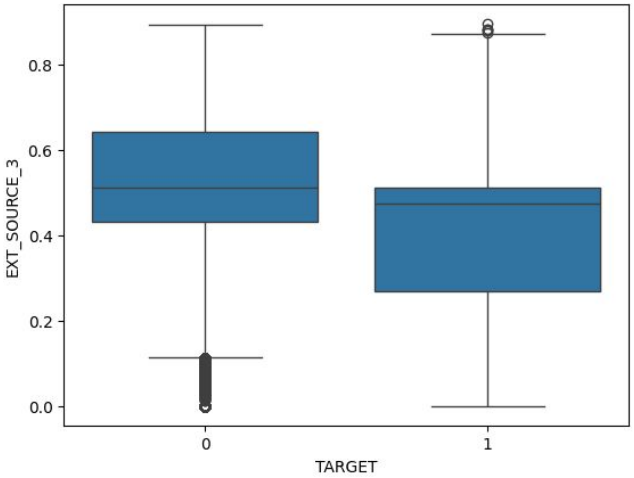
Heat map

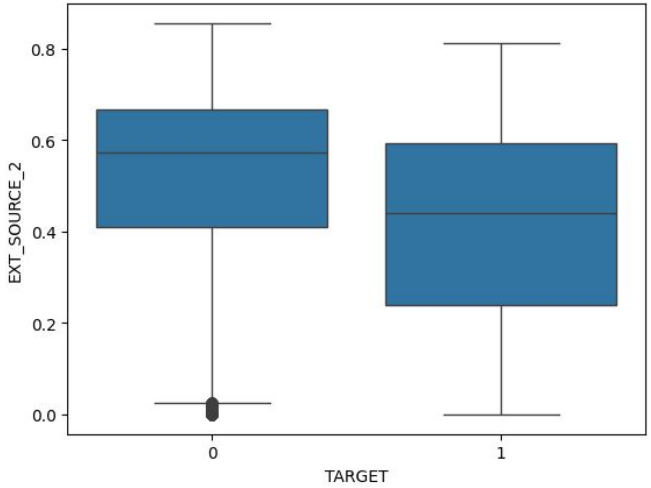
The heatmap give us a theory that the more EXT_SOURCE_3 or 2 goes down so the target values may go up -> 1.In other words, People with higher external score tends to be not a defaulter.



Box plot

The Box plot strengthen the mentioned hypothesis, the more values in TARGET = 0 has higher values than ones in TARGET =1.





Conclusion: Data is highly imbalanced on target field, this can effect on the insight. Many fields have high null value cause us cannot take advantage of all fields. Many fields contain wrong data, the reliability is not high. Many fields have outliers.
Insight: Base on the analysis if we want to reduce the defaulter we should be more caution on:

- Be more caution on male customer
- Have higher trust on people have higher education
- Be more caution on people who live with parents
- Be more caution on people have occupations as “Laborers”
- Be more caution on people have organization as Business Entity Type 3
- Have higher trust on people have high “EXT_SOURCE_3” and “EXT_SOURCE_2”