# Credit EDA Assignment

PULKIT PAREEK

# Problem Statement

To comprehend which consumer attributes and loan attributes influence the tendency of default.

To find which customer are more likely to default on their loan payment.

# Analysis Process

Data Understanding :

- Checking data rows, shape of dataset, data type of columns, number of missing values

Data Quality:

- Missing Values : We will check percentage of missing values in a columns. For all the column grater than 45% were removed and less 15% were imputed by median (as there are outliers in those columns median less affected by outliers). For all column between 15% to 45% were left as it as.

- Outlier : Outlier identification was don by boxplot and custom function. As it is not required to treat outliers for this eda. We can left them. But outliers can be treated by capping or flooring.

- Incorrect Format : All the days columns are in negative. So need to apply abs function to change them positive.

# Analysis Process : Continue

- Quasi Constant Feature : As some feature had more than 98% values as single values those feature won't be helpful in analysis. I dropped those features.

- Treating Categorical Columns : Categorical columns in dataset have many that are very less in quantity. For wherever I could do I have clubbed those categories to one type. So we have categories with enough data to analyze.

- Correlation : To remove unnecessary numerical columns, we use correlation to remove all those who correlation with target is less than 0.045. Here for correlation as our target is categorical we have to point biserial function.

Check if data is imbalance by checking TARGET variable value counts. Then divide dataset into Data with target 0 and data with target 1.

Finding Top 10 correlated features in dataset consider different values of target.

With all this our data will be more ready for Univariate, Bivariate and Multivariate Analysis

# Insights from App Data

- Dataset 122 columns initially but I remove all those columns that had more 45% of data missing, were quasi constant columns or they had low correlation with Target Variable (For this I have used point biserial function as this can be used for categorical-numerical relation ).

- With this I had finally approximately 25 columns.

- Now we can move to Univariate Analysis and gain Insights from there

# Top 10 Correlated Features (Among Themselves)

| | | |
|---|---|---|
| FLAG_EMP_PHONE | YEARS_EMPLOYED | 0.999703 |
| OBS_30_CNT_SOCIAL_CIRCLE | OBS_60_CNT_SOCIAL_CIRCLE | 0.998286 |
| AMT_CREDIT | AMT_GOODS_PRICE | 0.983065 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.956477 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.885556 |
| DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.869761 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.847260 |
| REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY | 0.778110 |
| AMT_ANNUITY | AMT_GOODS_PRICE | 0.752206 |
| AMT_CREDIT | AMT_ANNUITY | 0.751400 |

dtype: float64

- For both target We have same pair of features in list and order is almost same except these 2 pairs (LIVE_REGION_NOT_WORK_REGION ,REG_REGION_NOT_WORK_REGION), (DEF_60_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE).

- Most of correlation makes sense, as years of employed will after if we have a employee pone or not. All amount column correlating to each other. While credit and goods price has most correlation as we would take loan for goods.

- Number of children will increase number of family members

- We can take one feature out of pair we can do our analysis. As we can correlate one column from other.

# Univariate Analysis

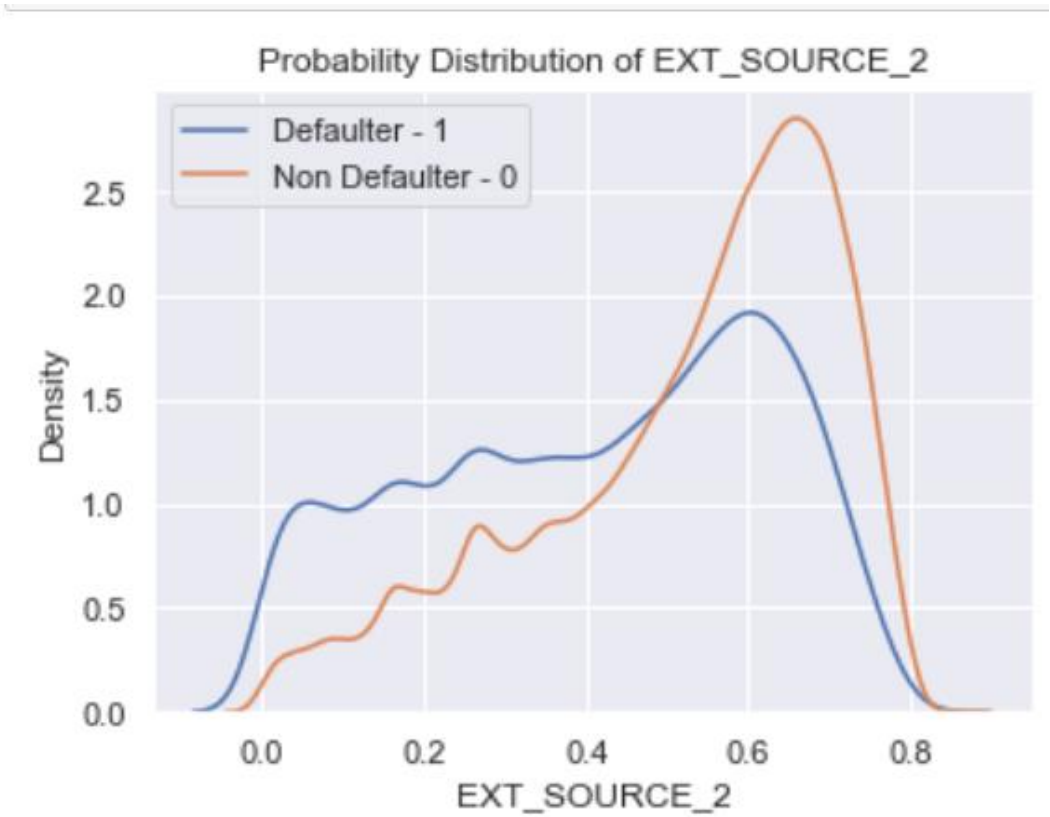
For this I have used, these

Numerical columns : [ 'EXT_SOURCE_2','YEARS_BIRTH', 'YEARS_LAST_PHONE_CHANGE', 'YEARS_ID_PUBLISH', 'YEARS_EMPLOYED', 'YEARS_REGISTRATION', 'EXT_SOURCE_3']

Categorical Columns : 'REGION_RATING_CLIENT_W_CITY', 'REGION_RATING_CLIENT', 'REG_CITY_NOT_WORK_CITY','FLAG_EMP_PHONE','REG_CITY_NOT_LIVE_CITY‘,'FLAG_DOCUMENT_3', 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE‘ ,'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE','WEEKDAY_APPR_PROCESS_START', 'ORGANIZATION_TYPE‘

In Next Few Slides we will be taking few relevant columns and their outcome

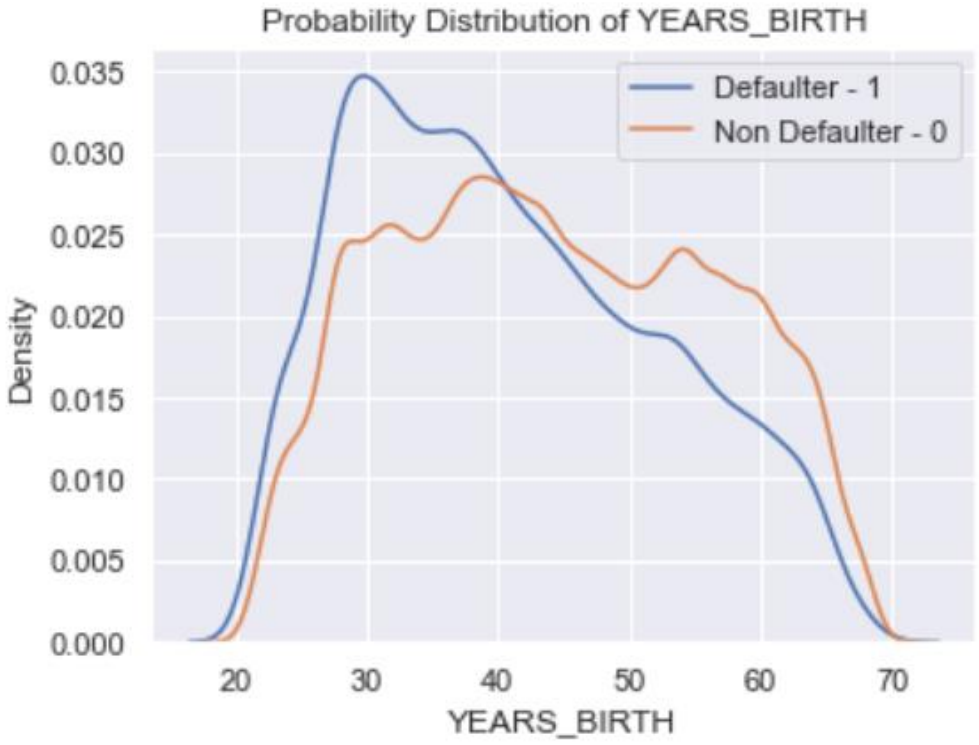
# EXT_SOURCE_2




- With EXT_SOURCE_2 we can easily distinguish Defaulter and Non Defaulter.

- Non Defaulter have higher EXT_SOURCE_2 and also spread of values are also small in comparison to Defaulter. But Non Defaulter have lower value until 0.5 (approx). We can see peak for both Targets at 0.6(approx) but Non-Defaulter has at higher side

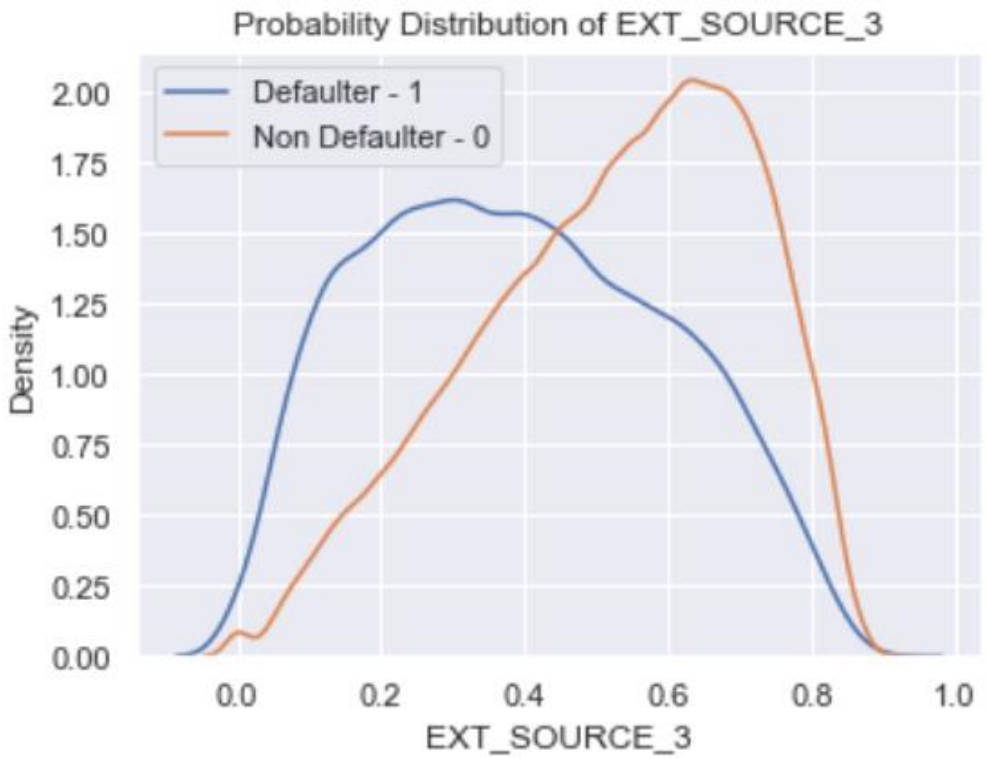
# YEARS_BIRTH




- People more than 40 years of age are less likely to default compare to people less than 40 years.

- Non Defaulter seems to higher age range also.

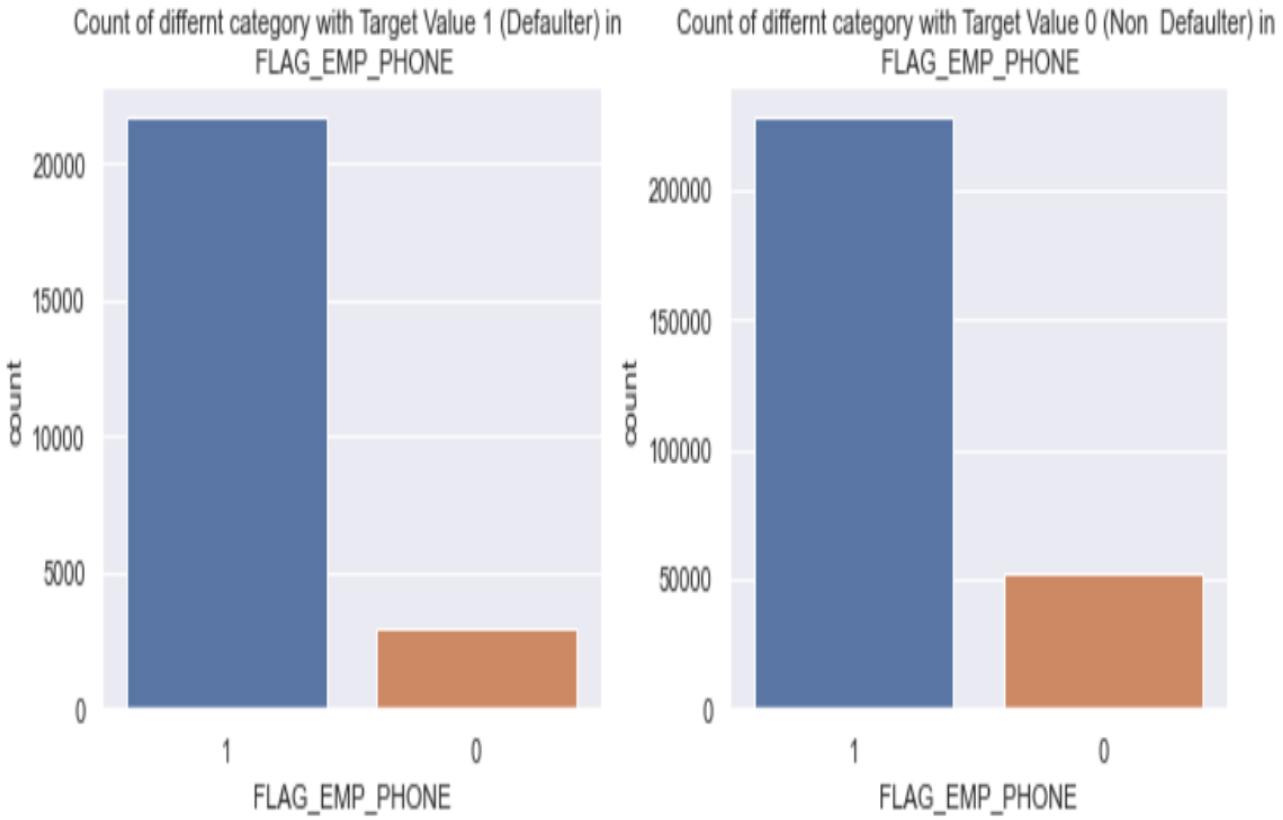
# EXT_SOURCE_3




- With this feature we can easily differentiate between Defaulter and Non-Defaulter

- Before 0.45 Deafulter have higher number Applicants, after 0.45 Non-Defaulter have more applicants

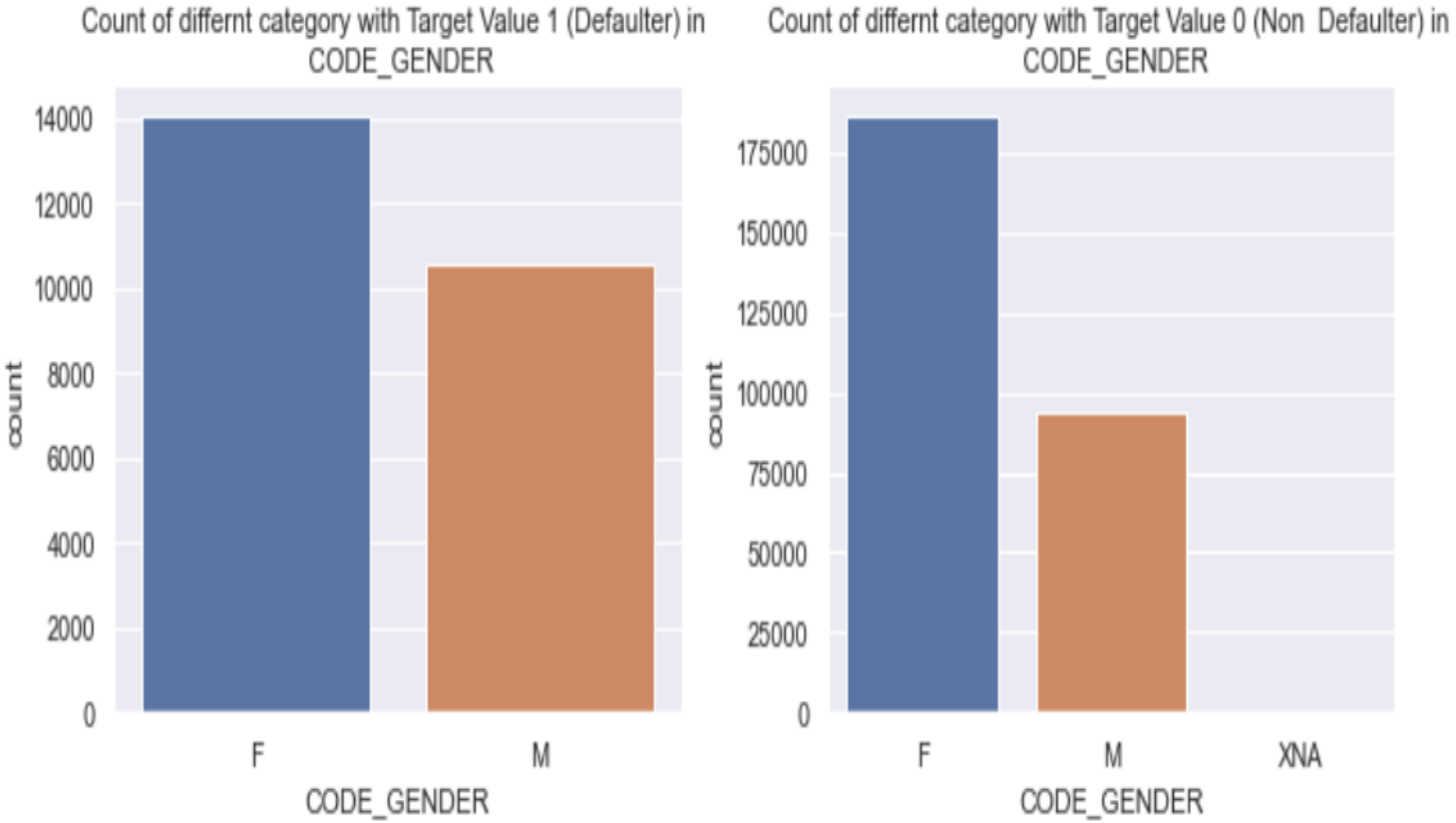
# FLAG_EMP_PHONE




- When we compare both categories in FLAG_EMP_PHONE category 0 where Applicant has employee phone, has higher number in Non Defaulter than Defaulter.

- If Applicant does not have a Employee phone there is little bit higher changes of Defaulting.

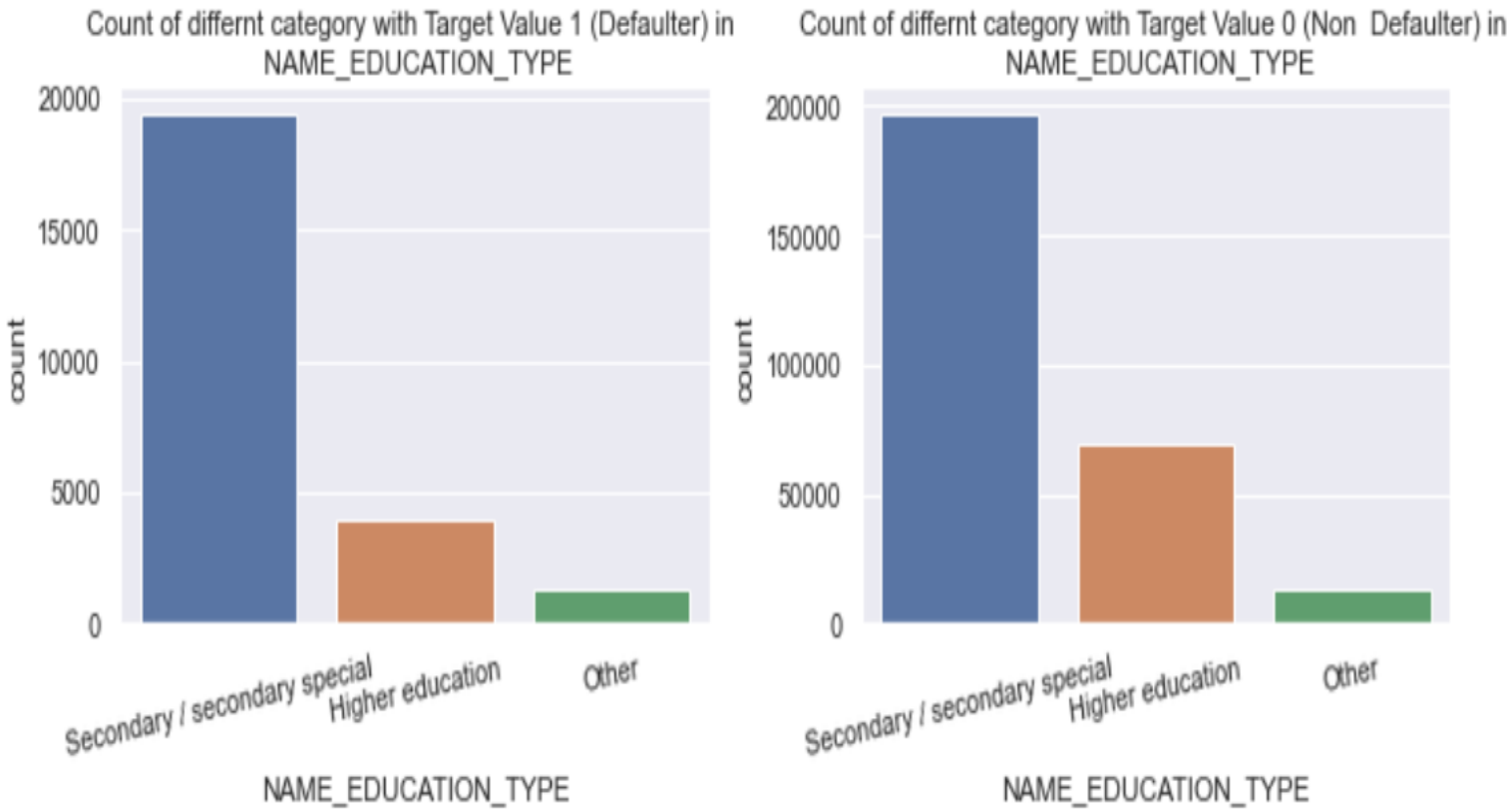
# CODE_GENDER




- Male are more likely to default on a loan.

- XNA is only chosen by Non Defaulter Applicants

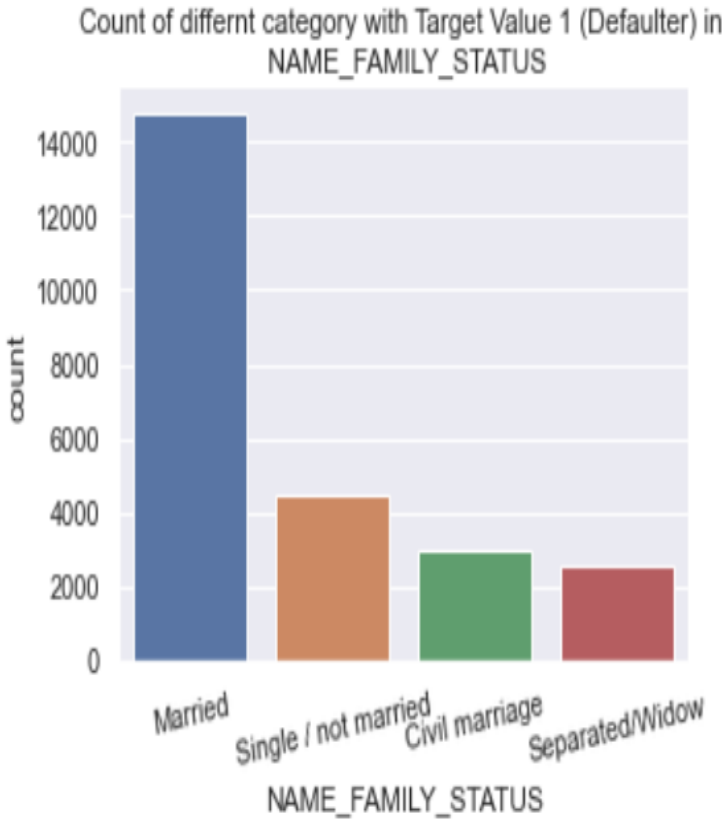
# NAME_EDUCATION_TYPE

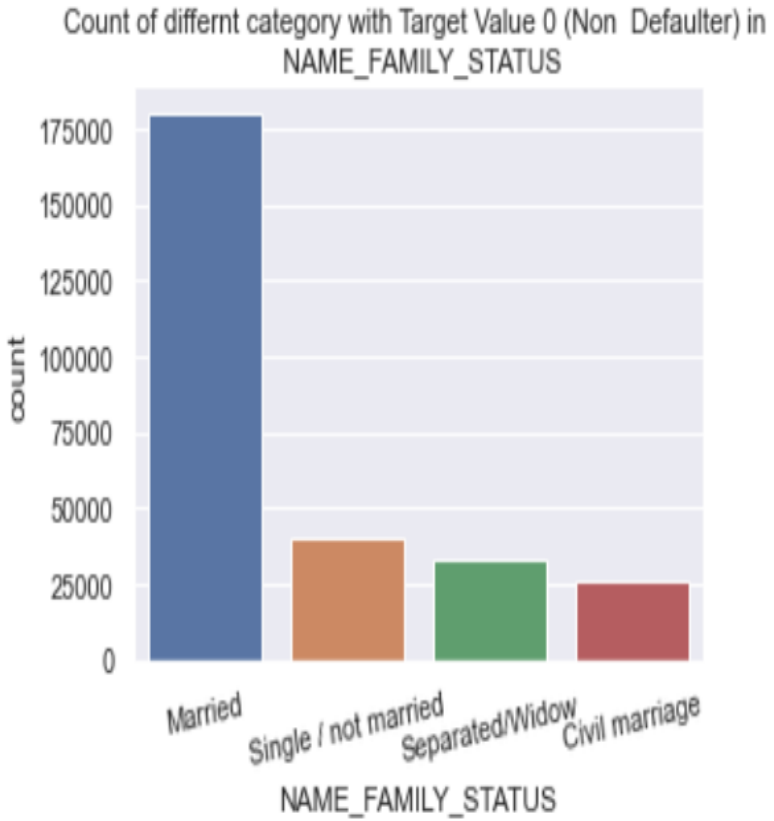



- Most applicant has only Secondary Education

- People who have higher education are less likely to default.

# NAME_FAMILY_STATUS

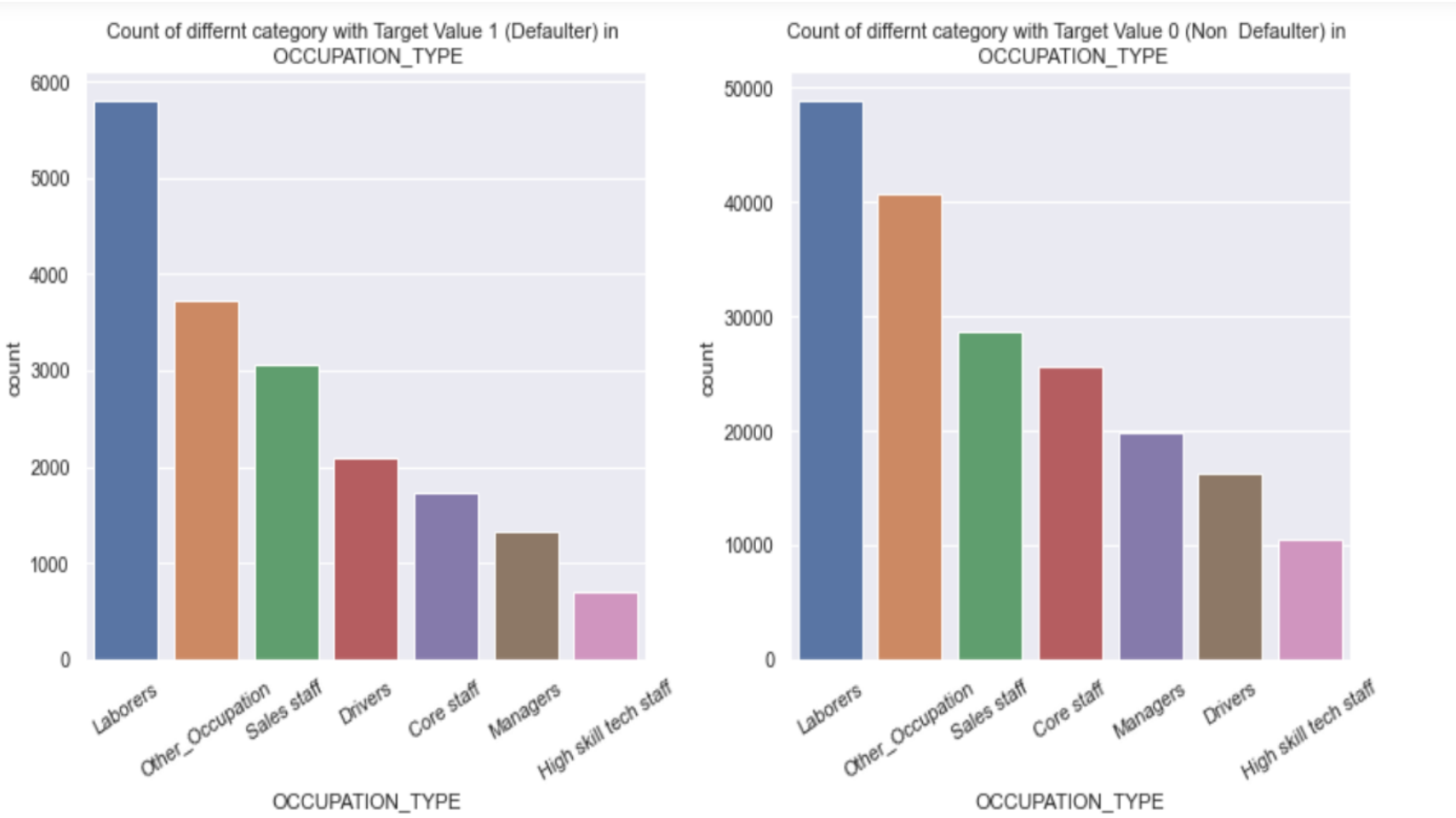



- In Non Defaulter and Defaulter There is check in order as there are more applicant in Separated /Widow section in Non Defaulter while in Defaulter in Civil Marriage.

- Most Applicant are in Married Category.

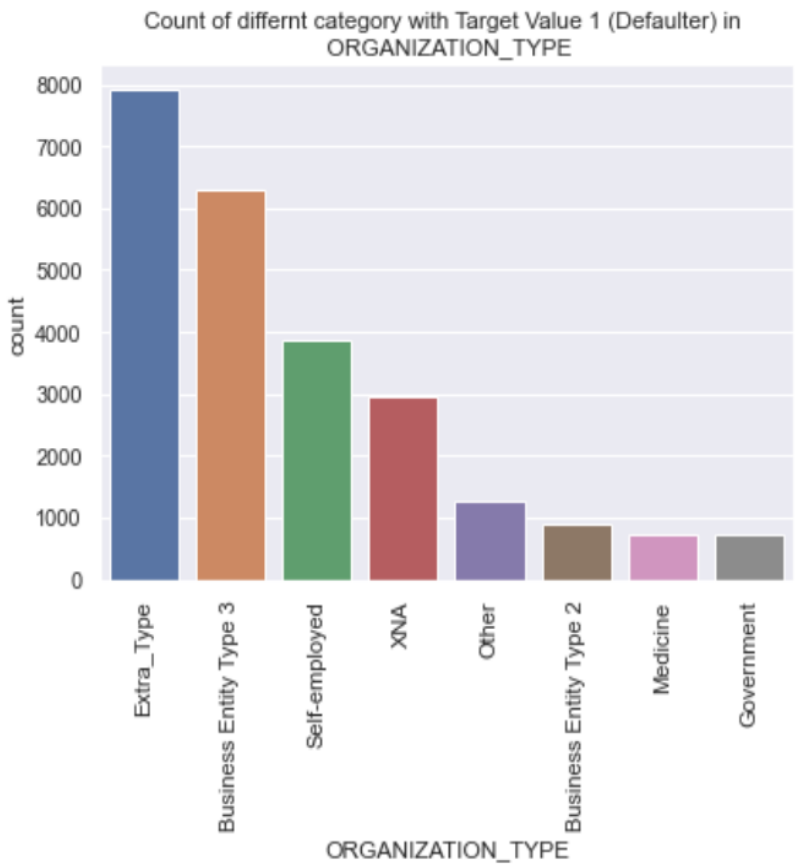
# OCCUPATION_TYPE

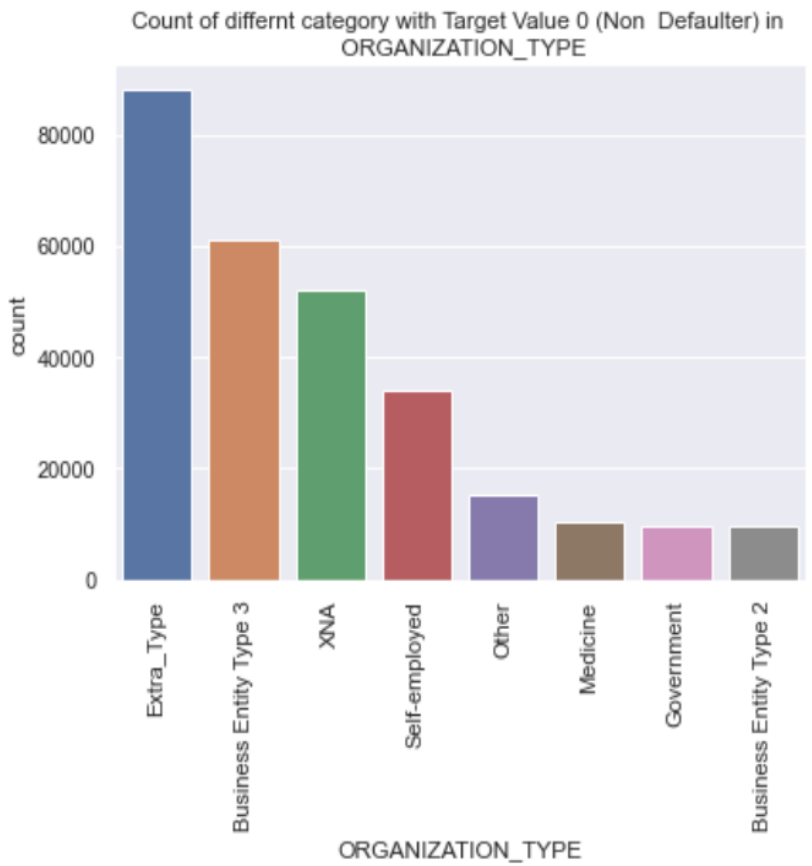



- In both Target values we have similar trend for all occupation except Driver.

- Drivers are more like to Default on a loan.

# ORGANIZATION_TYPE




- All Organization type of Application have same trend in Defaulter and Non Defaulter. Except people who have XNA are less likely to Default than who have filled Self Employed.

# Bi-Variate or Multivariate Analysis

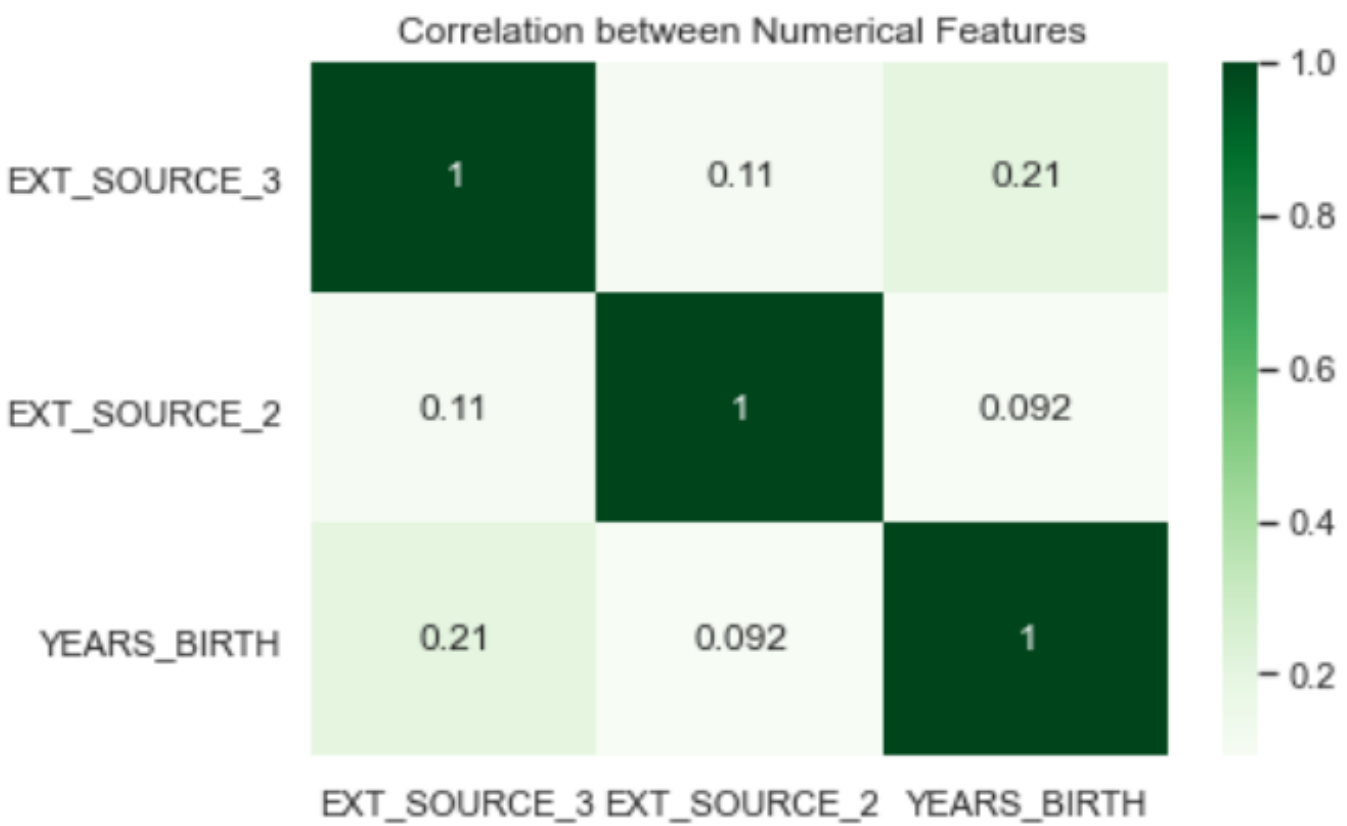
For this analysis we will using

Numerical Columns : ['EXT_SOURCE_3', 'EXT_SOURCE_2', 'YEARS_BIRTH']

Categorical Columns : ['ORGANIZATION_TYPE', 'NAME_FAMILY_STATUS','NAME_EDUCATION_TYPE','CODE_GENDER']

With these column we will using heat map, pair plot and pivot tables to gain insight

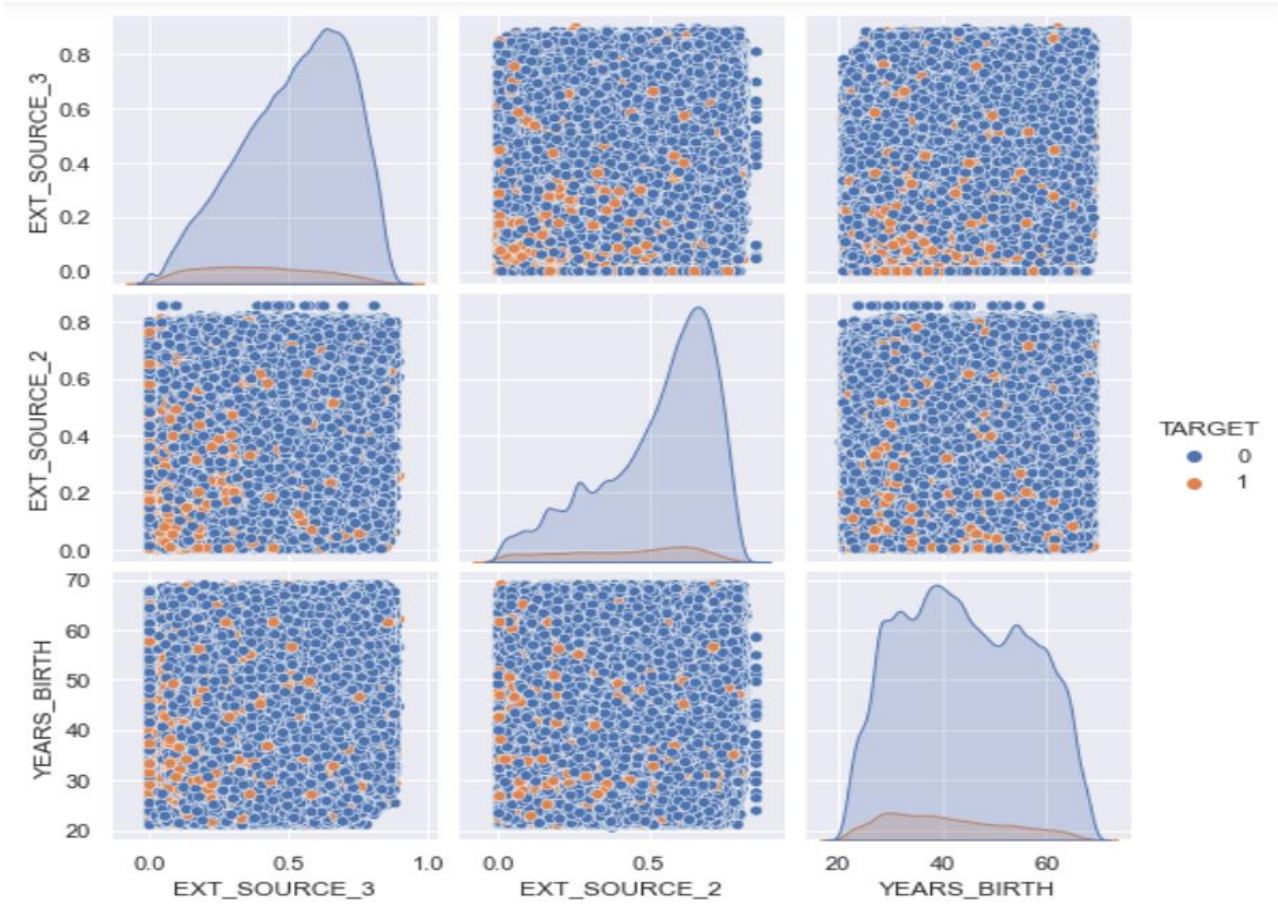
# Correlation among Numerical Features




As all three have very low correlation among themselves and also positive correlation.

For feature to be good predictor of target it should have high correlation to Target and low correlation among themselves.

# Relationship in between all numerical columns (among themselves) and target




From this graph we can see that defaulter(generally) have low values of EXT_SOURCE2 and EXT_SOURCE3 Score and have age in between 20 to 40

# ORGANIZATION_TYPE & CODE_GENDER

| ORGANIZATION_TYPE<br>CODE_GENDER | Business Entity Type 2 | Business Entity Type 3 | Extra_Type | Government | Medicine | Other | Self-employed | XNA |
|---|---|---|---|---|---|---|---|---|
| F | 0.070435 | 0.080121 | 0.072836 | 0.063098 | 0.065706 | 0.068753 | 0.089543 | 0.049706 |
| M | 0.102804 | 0.110507 | 0.097726 | 0.088205 | 0.068458 | 0.091251 | 0.127158 | 0.073789 |
| XNA | NaN | NaN | 0.000000 | NaN | 0.000000 | NaN | NaN | NaN |

With this we can see as male have higher chance of Default in that for ORGANIZATION_TYPE as self_employed it is much higher.

# NAME_FAMILY_STATUS & NAME_EDUCATION_TYPE

| NAME_EDUCATION_TYPE<br>NAME_FAMILY_STATUS | Higher education | Other | Secondary / secondary special |
|---|---|---|---|
| Civil marriage | 0.066778 | 0.104353 | 0.108373 |
| Married | 0.049875 | 0.084670 | 0.084315 |
| Separated/Widow | 0.054273 | 0.083639 | 0.075543 |
| Single / not married | 0.062611 | 0.101224 | 0.113829 |

In Education & Family status highest is for (Secondary / secondary special & Single / not married).

# Findings from applicants data

- Application External Source score 2 & 3 matters a lot in deciding Defaulter and Non-Defaulter.

They also have highest correlation (Linear relation with target variable.)

- Years Birth (Person's Age) is also can be one of decisive factor for defaulting on a loan.

- Applicant's Family status are they married or not, Education Level like Higher or secondary level, Organization type in which they work and what is their gender also matters in deciding who we can give loans.

# Using Previous application Data

- As getting a loan and If we will be able pay loan does depend on previous application, so we will be using that data with merging it to applicant data, to get more clear about which Applicants are more likely to Default on a loan.

- Similar to app data here also we did some data cleaning, after that we reached to these

Numerical columns :  ['HOUR_APPR_PROCESS_START', 'DAYS_DECISION', 'CNT_PAYMENT']

For Categorical as all are in Object for those columns, I have used count plot with to get relevance of those columns with Target

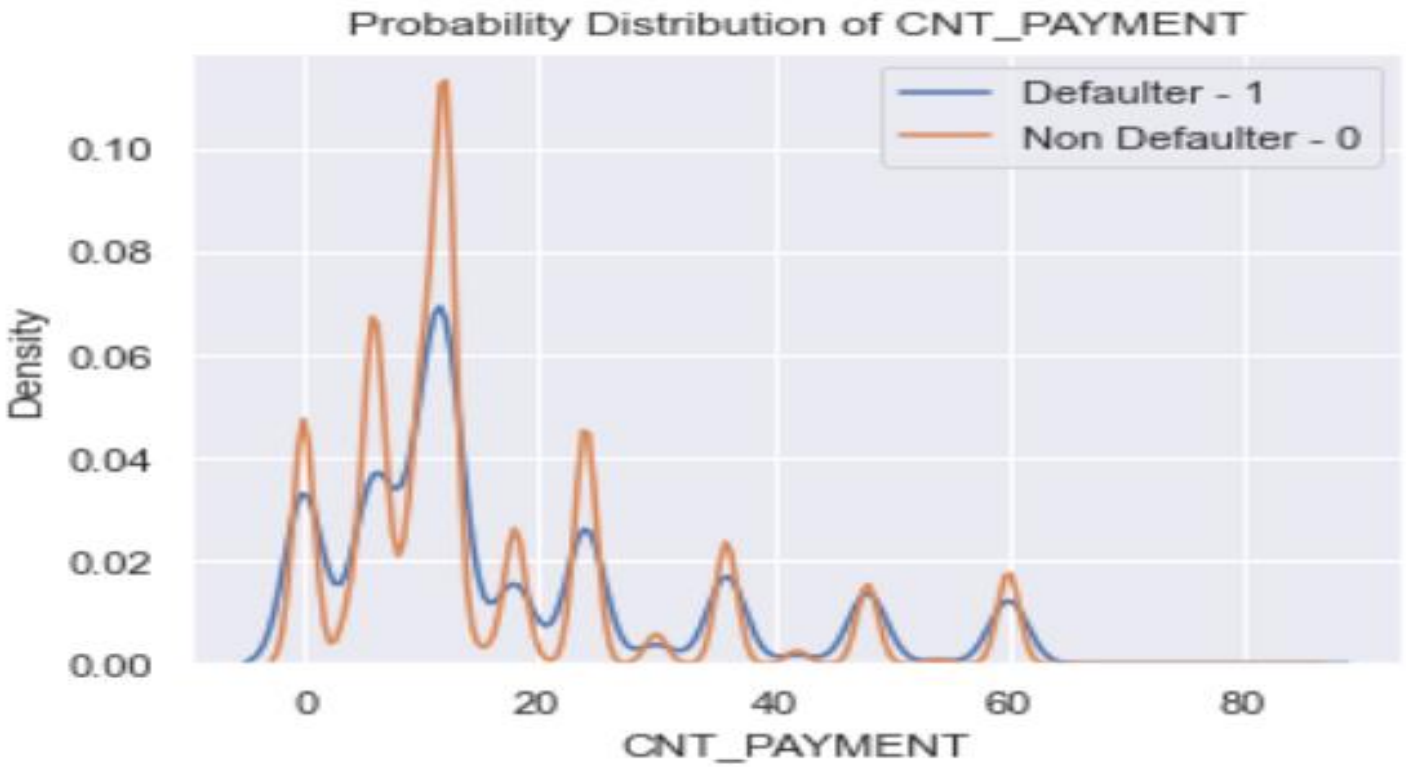
# Univariate Analysis

Numerical columns :  ['HOUR_APPR_PROCESS_START', 'DAYS_DECISION', 'CNT_PAYMENT']

I did univariate analysis all three column but only in CNT_PAYMENT there was some useful information, while two did not.

So I will leaving other two out of this presentation

# CNT_PAYMENT



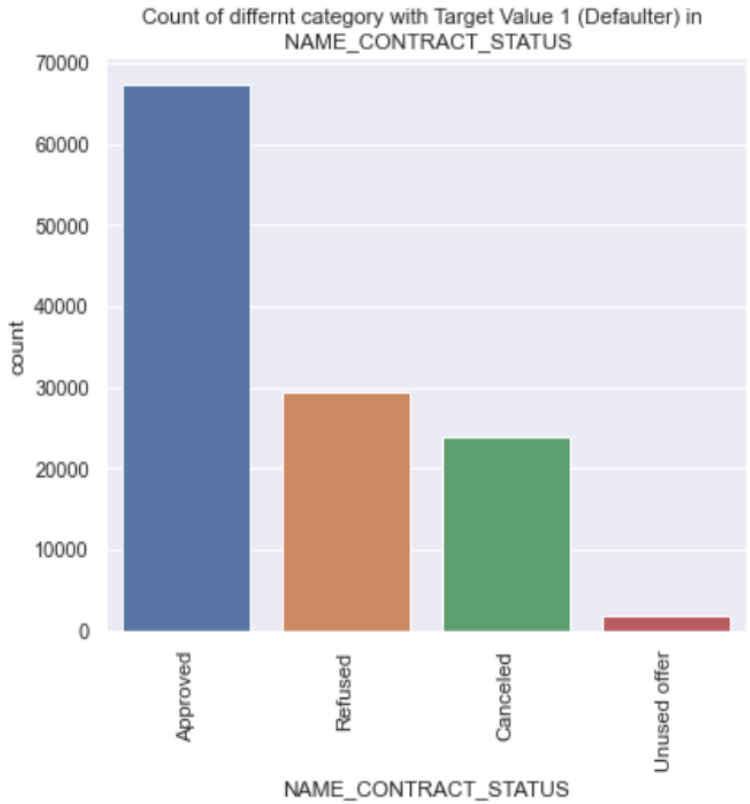

For cnt_payment we can see Defaulter and Non Defaulter seems have peak at same values but for Non-Defaulter it is high.

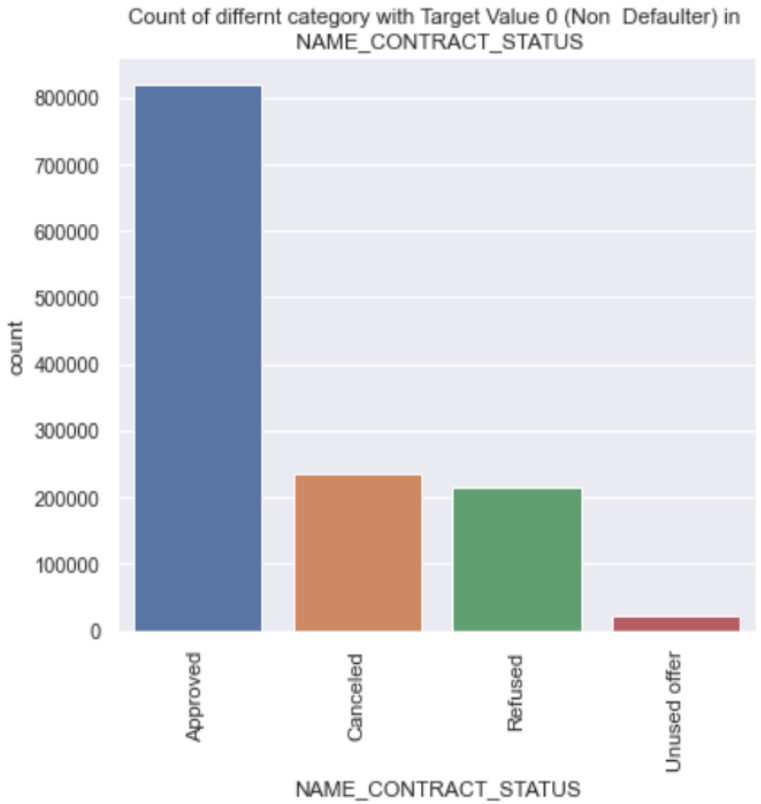
# Categorical Univariate Analysis

I did make count plot of all categorical (object data type) columns, tried to get info about of those.

Out of all columns we got some info from NAME_CONTRACT_TYPE & NAME_CONTRACT_STATUS

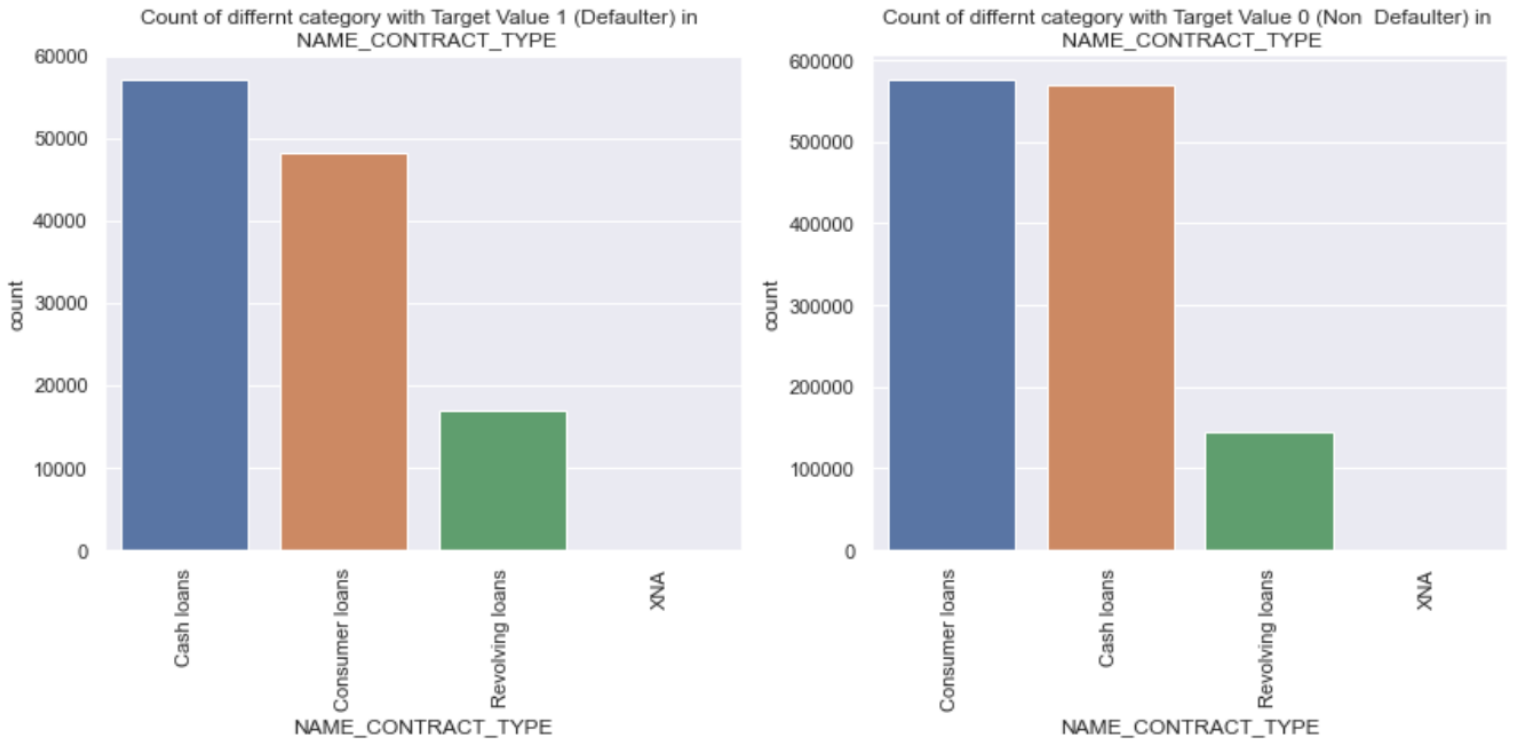
# NAME_CONTRACT_STATUS




For Contact Status, for defaulter applicant seems to have more Refused and cancelled than non-defaulter

# NAME_CONTRACT_TYPE







Defaulter seems to have more consumer loans than Non Defaulter, while for cash loans even the difference is little it is opposite of consumer loans

# NAME_CONTRACT_TYPE & NAME_CONTRACT_STATUS

| NAME_CONTRACT_STATUS<br>NAME_CONTRACT_TYPE | Approved | Canceled | Refused | Unused offer |
|---|---|---|---|---|
| Cash loans | 0.075516 | 0.088401 | 0.125810 | 0.092593 |
| Consumer loans | 0.073853 | 0.128668 | 0.101350 | 0.082337 |
| Revolving loans | 0.090343 | 0.109254 | 0.129050 | 0.000000 |
| XNA | NaN | 0.197183 | 0.241379 | NaN |

Here we can see Applicants that has refused from loan and of XNA Contract type are more likely to default on their loan.

# Findings from prev application

- From this we can find cnt payment, Type of contract and Contract’s status on their previous application as deciding factor defaulting on their current application.

# Features to consider

1. From applicants data we can use

['EXT_SOURCE_3‘: Score from external source 3, 'EXT_SOURCE_2‘: Score from external source 2, 'YEARS_BIRTH‘: Age of person(this column is converted from days_birth) ]

['ORGANIZATION_TYPE', 'NAME_FAMILY_STATUS','NAME_EDUCATION_TYPE','CODE_GENDER'] as person’s organization, their family status, education level  does gives how much they can pay back. From gender Male are more likely to Default on a loan.

2. From Previous application we can use :

Cnt_payment, as how many payments were completed, what type of contract they previously had and what was the status of that contract

Example : Applicants that has refused from loan and of XNA Contract type are more likely to default on their loan.

# Thank You

PULKIT PAREEK