CREDIT EDA CASE STUDY
Analysis

DHEERAJTHEJA ALLURU
V V SATYA SAI KIRAN KUSUMANCHI

EDA Credit Case Study

To ensure that the bad loans are reduced to greater extent by analyzing the underlying pattern of the historical and current data.

Objective1

To ensure that the intended and reliable candidates get the loan so that there is no loss to institution.

Objective 2

Data Source

1.The current applicants who have applied now
2. Historical data for the same customer with previous loan history.
Both can be found here

Implementation Plan

EDA_CaseStudy.ipynb attached in the zip file

Phase 1

Phase 2

Phase 3

Feature Engineering:

- Load of data.
- Dropping unnecessary and sparsely(<50%) populated variables.
- Data Imbalance Checks.
- Data Imputations using statistical Techniques.

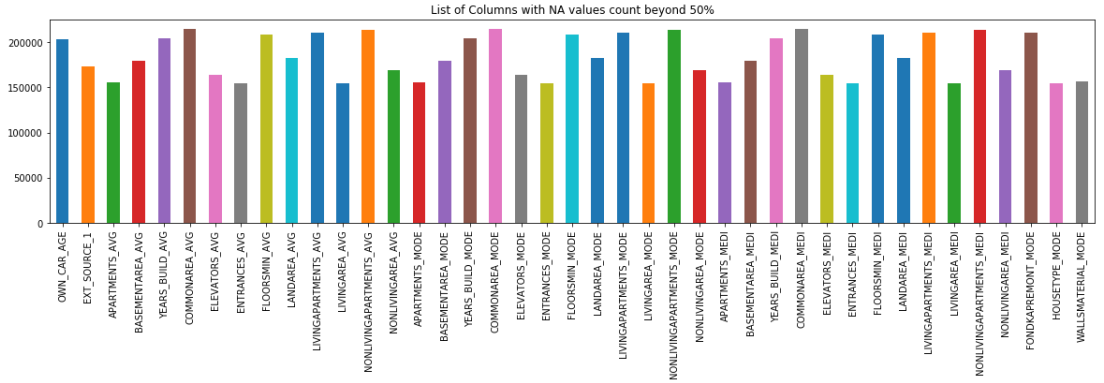
Univariate and Bivariate Analysis for the datasets in terms of TARGET variable in Application data.

Univariate and Bivariate Analysis for the datasets in terms of TARGET variable using the historical (Previous) data.

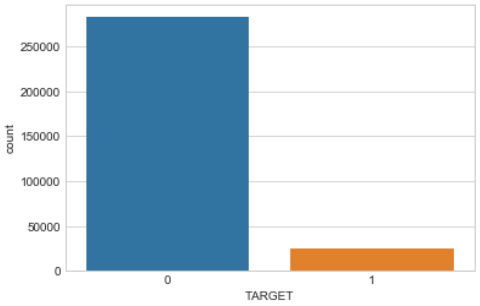
Phase 1

Feature Engineering:

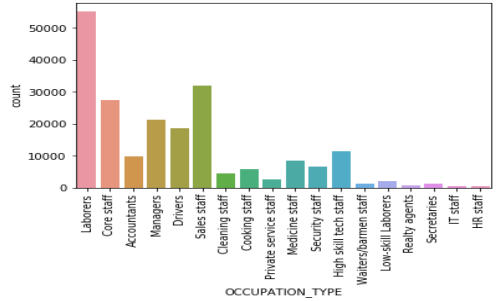
- **Dropped the Columns** which are populated less than 50 %, as these might not be helpful in giving the right understanding of the data.
- Retained only the important columns that are found to be relevant for the analysis.
- **Data Imputations:**
  - Categorical Variables: Used MODE for the imputation
    E.g.: OCCUPATION_TYPE
  - Quantitative Variables: Used Median / Mode for the imputation based on outliers.
    E.g.: AMT_ANNUITY, AMS_GOOD_PRICE and CNT_FAM_MEMBERS
- **Binned the Continuous Variables.**
- **Outliers Identification and Treatment** – Chose the columns: CNT_CHILDREN and CNT_FAM_MEMBERS.
- **Data Imbalance checks for TARGET**:
  - Percentage of clients with difficulty(%) : 8.07
  - Percentage of clients with No difficulty(%): 91.93
  - Ratio of data_imbalance: 11.387:1



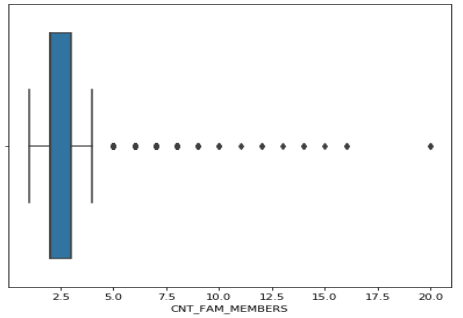
Data Imbalance



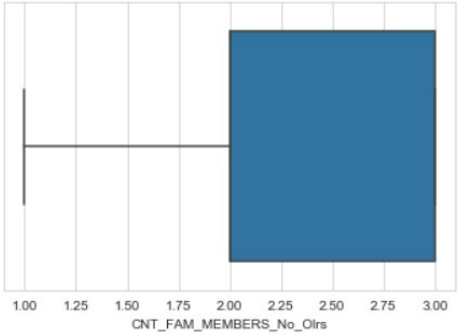
Data Imputation (using Mode)



Outliers Treatment – (Spread of Family members Count)



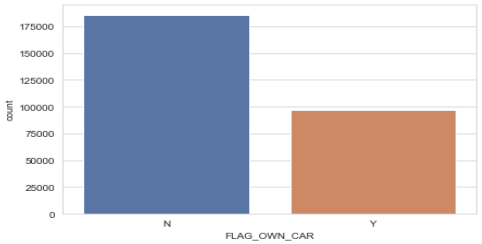
Pre

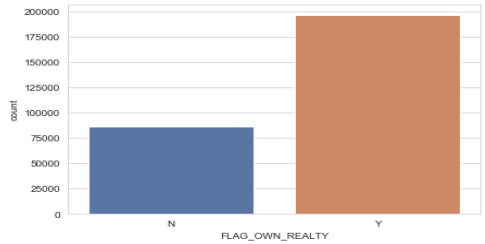


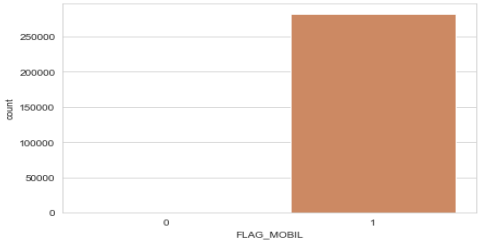
Post

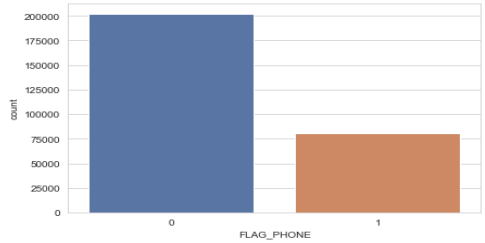
Phase 2: Univariate Analysis

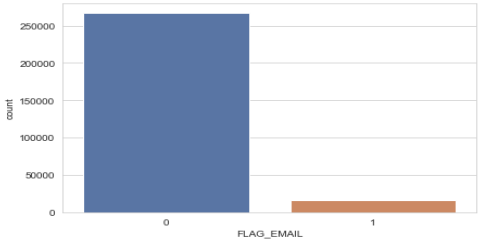
Data Set: Applicant who don't find it difficult to repay loan TARGET:0

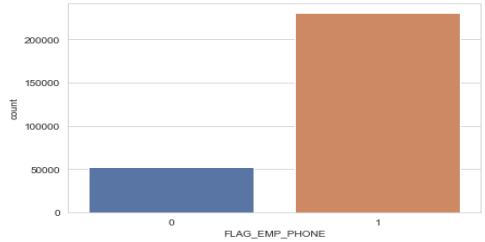


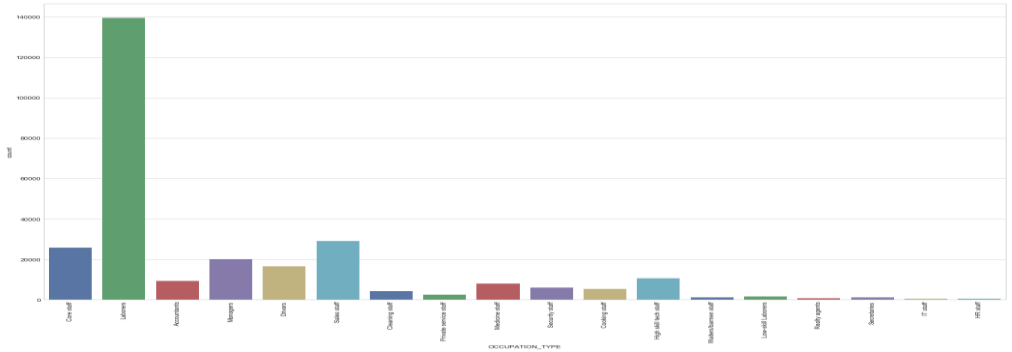












Observations:

- For the applicants who don't own a car, there is a chance that they don't find difficulty in repaying the loan. (Assumption: It could be because there wont be expense on car).
- For the applicants who own a real estate property there is a chance that they don't find difficulty in repaying the loan. (Assumption: It could be because they don't need to pay rent for living or get steady income on realty).
- For the applicants who own a mobile phone, phone, Email, work phone there is a chance that they don't find difficulty in repaying the loan. (Assumption: It could be because , it is common to own a mobile phone now a days and they get reminders about the loan payment timely).

Phase 2: Univariate Analysis

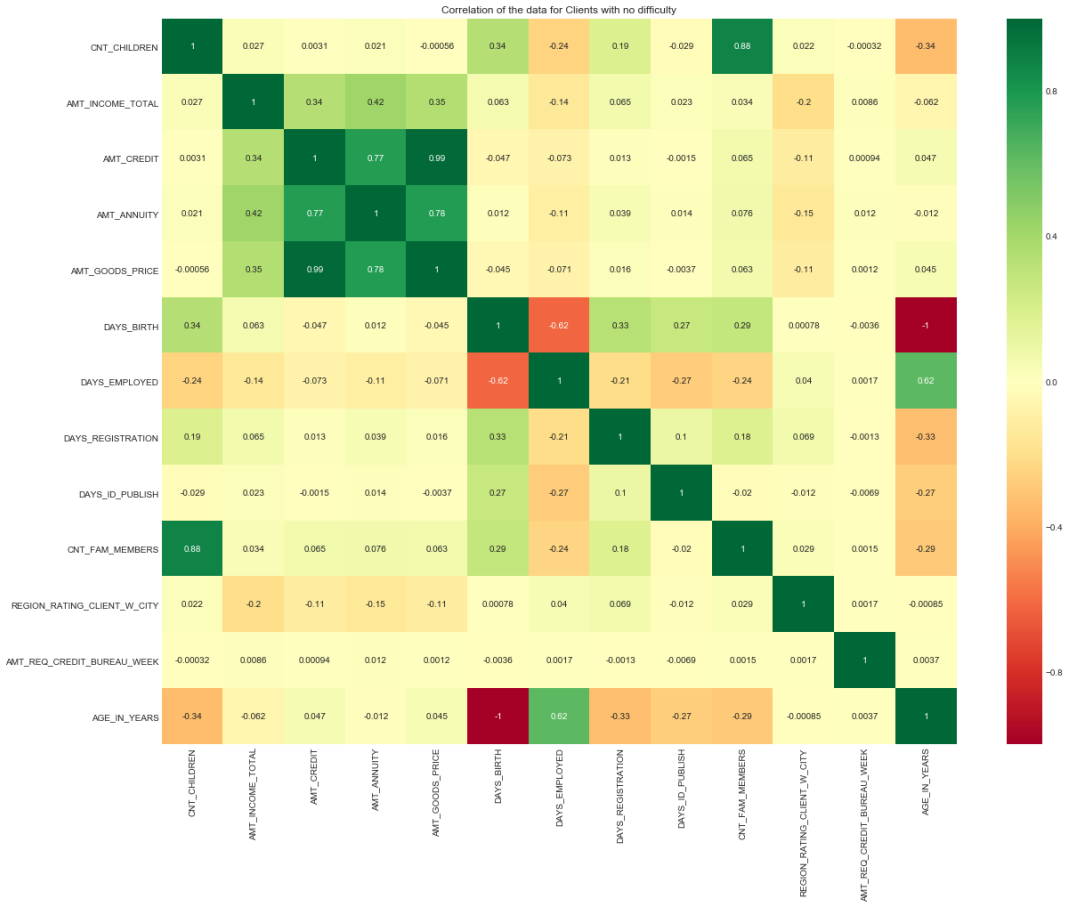
Data Set: Applicants who find it difficult to repay the loan, TARGET: 0

Observations:

- It is understood that for this category of applicants they are taking more cash loans, most of them are married and single, and they are working professionals or commercial associates.
- Most of the applicants who are in this category are lower secondary and above educated and are mostly living in a house / apartment.
- Most of the applicants who belong to this category are females and typically with a family size of two, and most of them don't have a child.
- Most of the applicants in this category reside in the type 2 and 3 regions.
- Most of the applicants in this category belong to the 30 to 60 years age group.
- Applicants in this category mostly fall under the occupation of laborers who work in the Business Entity type 3 organizations/ self employed/they didn't mention.

Phase 2: Univariate Analysis

Data Set: Applicant who find no difficult to repay loan, TARGET: 0



Observations

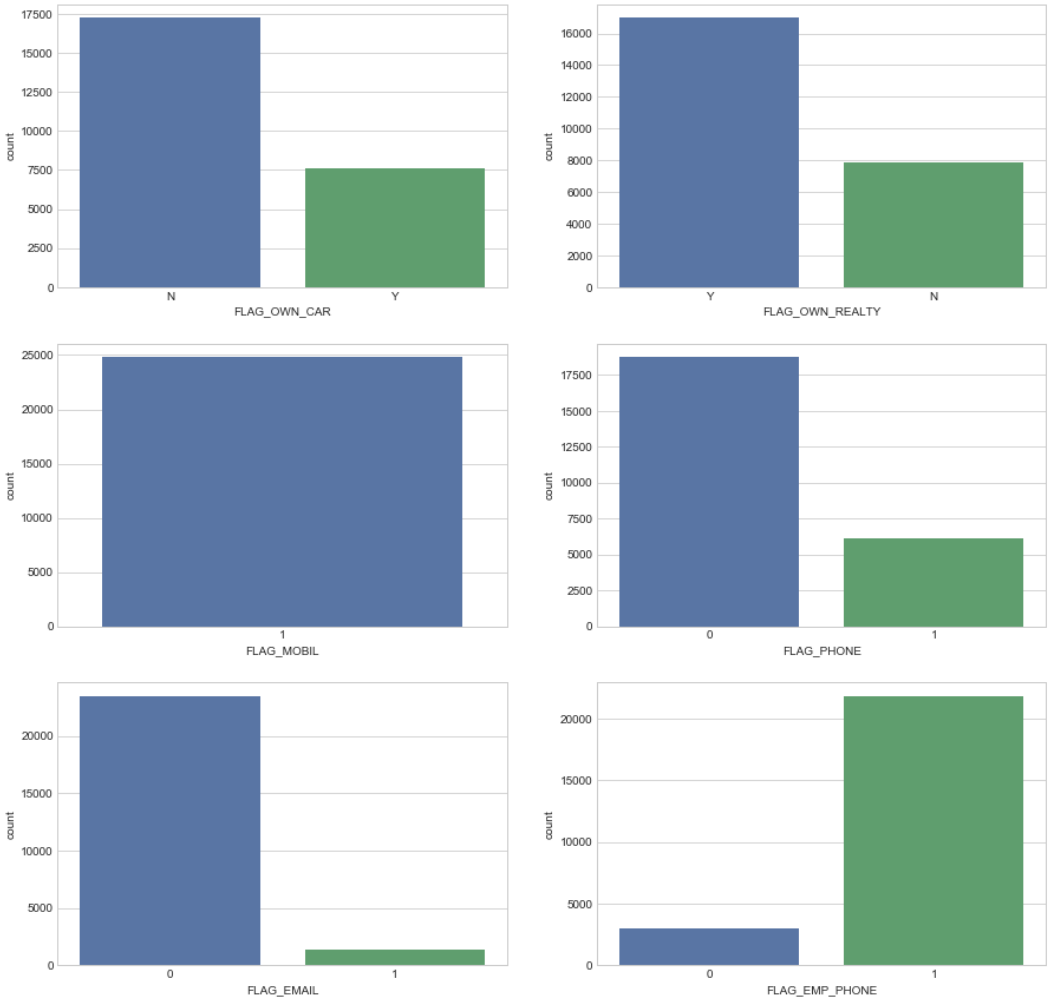
- Higher the income of Applicant, higher is the annuity amount, amount of goods' price and Credit amount.
- Higher the Good Price for the loans the applicants are applying for, higher is the credit amount they are looking for.
- In the same way, higher the price of goods for which loan is intended, higher is the annuity amount.
- Higher the amount of Credit, higher the annuity amount.

Top10 highly correlated variable in the df0 are:

1. AGE_IN_YEARS & DAYS_BIRTH
2. AMT_GOODS_PRICE & AMT_CREDIT
3. CNT_FAM_MEMBERS & CNT_CHILDREN
4. AMT_GOODS_PRICE & AMT_ANNUITY
5. AMT_ANNUITY & AMT_CREDIT
6. DAYS_EMPLOYED & DAYS_BIRTH
7. AGE_IN_YEARS & DAYS_EMPLOYED
8. AMT_ANNUITY & AMT_INCOME_TOTAL
9. AMT_GOODS_PRICE & AMT_INCOME_TOTAL
10. AMT_CREDIT & AMT_INCOME_TOTAL

Phase 2: Univariate Analysis

Data Set: Applicant who find it difficult to repay loan, TARGET: 1

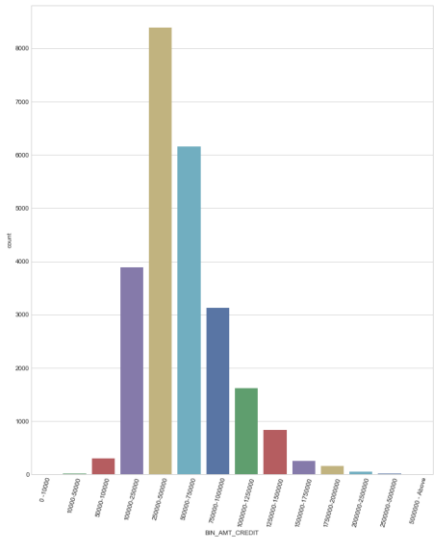


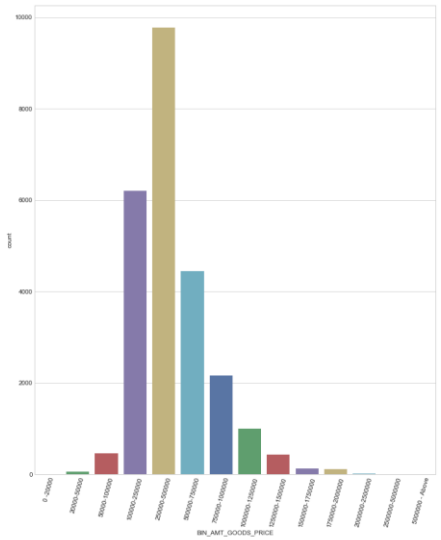
Observations:

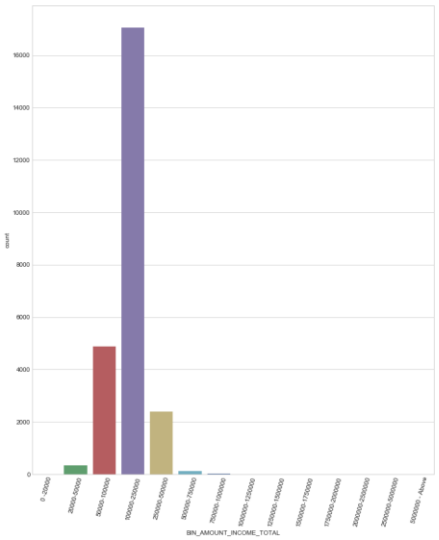
- For the applicants who doesn't own a car there is a chance that they find difficulty in repaying the loan, this might be because they are not in a condition of affording the car also
- For the applicants who own a real estate property there is a chance that they find difficulty in repaying the loan. It might be deprecating assets class which needs more investment
- we could get an inference with owning a mobile or phone or email or employee phone, which states that even they are informed about the remainder they find difficulty in repaying the loan in proper time

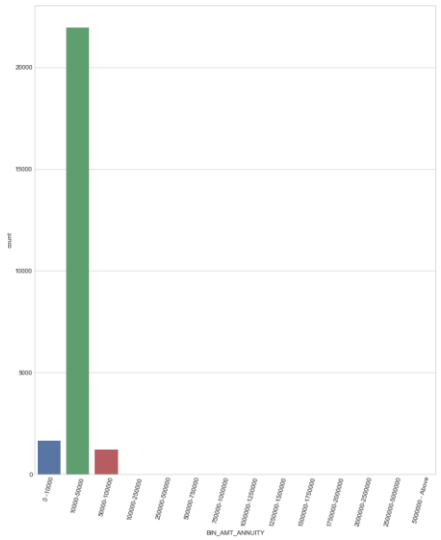
Phase 2: Univariate Analysis

Data Set: Applicant who find it difficult to repay loan, TARGET: 1







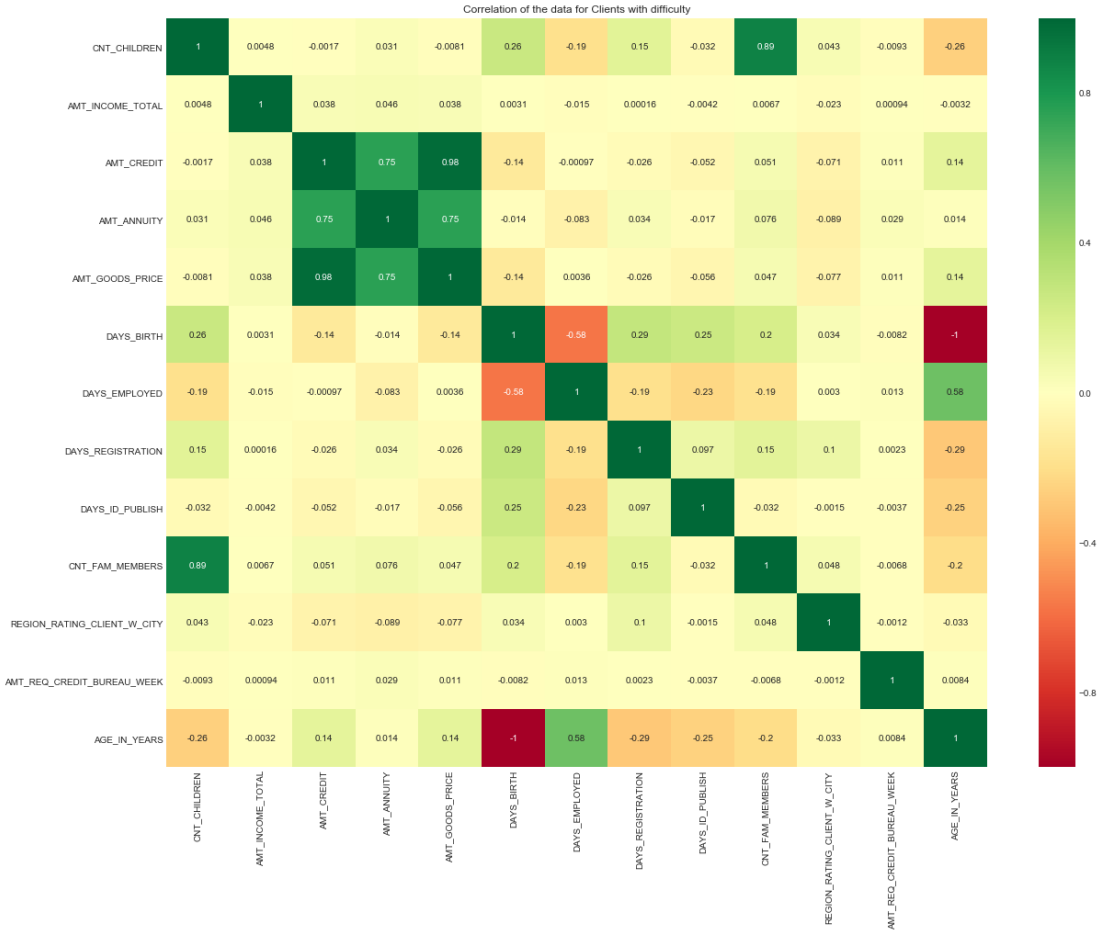


Observations

- Most of the applicants in this category reside in the type 2 and 3 regions.
- Most of the applicants in this category belong to the 20 to 50 years age group.
- Applicants in this category mostly fall under the occupation of laborer's who work in Business Entity type 3 organizations/ self employed/ they didn't mention.
- As observed, most of the applicants who are not able to repay the loans on time are having the credit somewhere between (100k to 750k) and annuity between 10k to 50k.
- Most of the applicants in the category are in the income range between 100K to 250K.
- The Goods price for which they are taking the loan are in the range of 100K, 750K.

Phase 2: Univariate Analysis

Data Set: Applicant who find it difficult to repay loan, TARGET: 1



Observations:

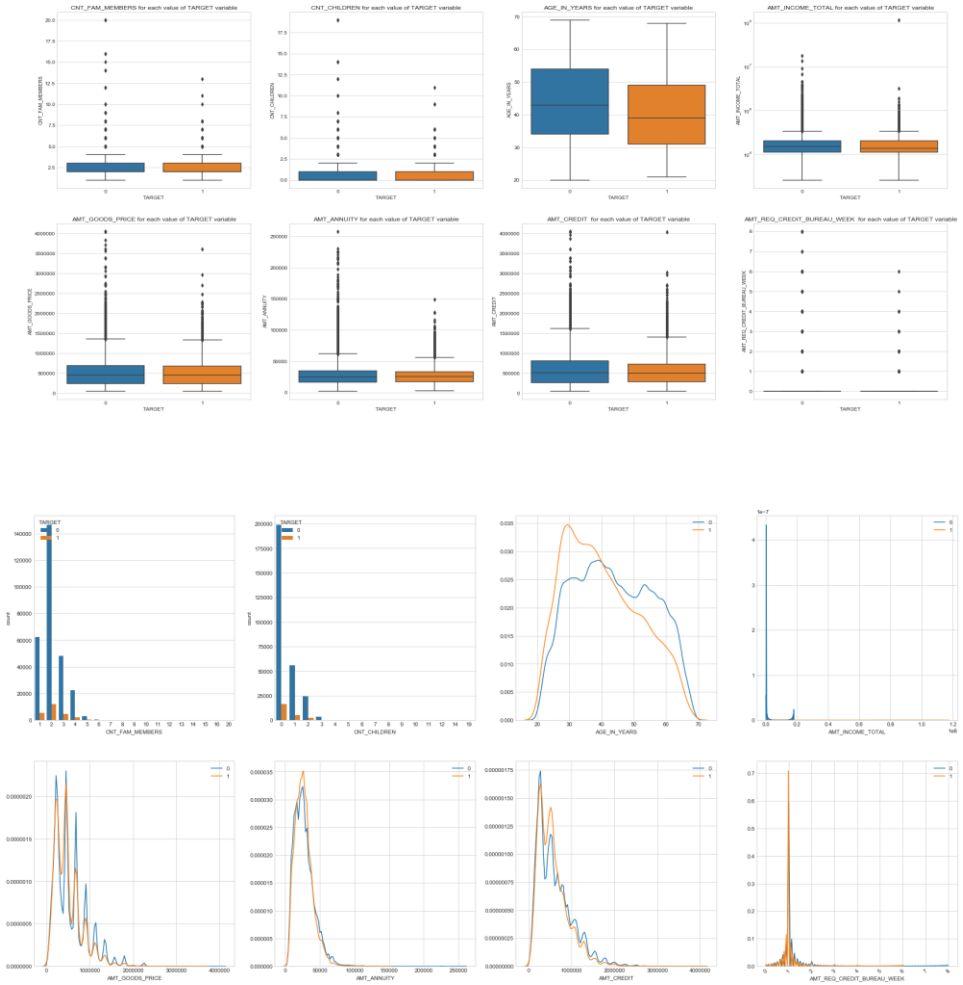
- Higher the price of Goods for which loan is taken, higher is the Credit Amount and Annuity amount.
- Higher the amount of Credit, higher is the Annuity amount.
- Observed that the correlation between the Income amount of client and Credit amount is not high.
- And there is very less correlation between Credit amount and Family size.
- And there is less significance of credit amount to number of children in a family.

Top 10 highly correlated variable in df1 are :

1. AMT_GOODS_PRICE & AMT_CREDIT
2. CNT_FAM_MEMBERS & CNT_CHILDREN
3. AMT_GOODS_PRICE & AMT_ANNUITY
4. AMT_ANNUITY & AMT_CREDIT
5. AGE_IN_YEARS & DAYS_EMPLOYED
6. DAYS_EMPLOYED & DAYS_BIRTH
7. AGE_IN_YEARS & DAYS_REGISTRATION
8. DAYS_REGISTRATION & DAYS_BIRTH
9. AGE_IN_YEARS & CNT_CHILDREN
10. DAYS_REGISTRATION & DAYS_BIRTH

Phase 2: Bivariate Analysis

Data Set: Applicant who find it difficult to repay loan, TARGET: 1,0

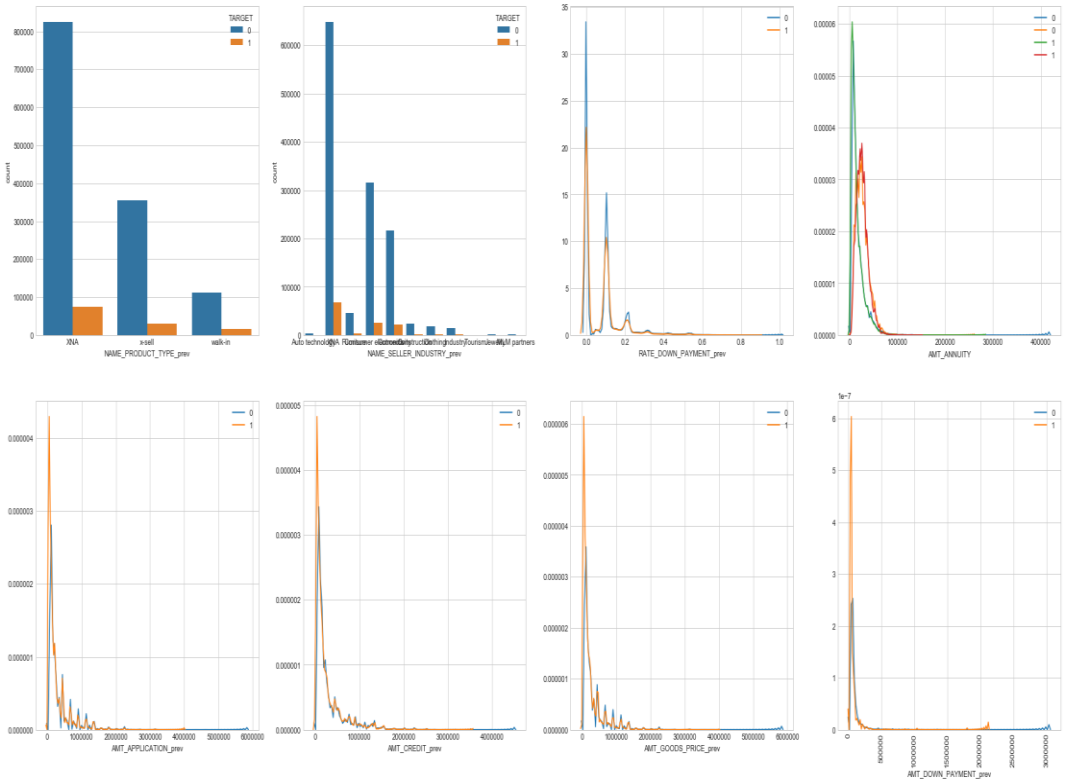


Observations:

- As observed in the **histplot** as drawn between the target and the age in years, it can be inferred that the applicants around age of 30 are more inclined to default the loan**.**
- As per the AMT_Income_Total for each value of Target variable (1st row,4th) graph it is observed that the median income of the applicant is less for the applicants who are inclined to default the loan.
- With respect to the annuity, as observed from the below distribution plot (2nd row, 2nd), most of the applicants have some annuity between 0-50000. and in that bracket, it can be inferred that there is more possibility for the people finding difficulty in repaying the loan.
- For the Amount credit, it is observed that the majority of the applicants fall under the 100000 below and can be inferred the applicants face difficulty in repaying the loans if their credit is nearer to 100000.
- As observed in the **distplot** drawn for the number of queries being sent to the credit bureau, it can be inferred that whenever there are more queries, we have seen that the applicants have more difficulty in repaying the loan.

Phase 3: Historical Analysis

Data Set: Applicant who find it difficult to repay loan, TARGET: 1



Observations:

- **The Applicants who were having trouble repaying now are having a chance that they were in X-sell for a previous loan. This might be because we have sold the credit product which they might not require.**
- Applicants who are having the Annuity below the 100k in the history are most likely to have problem in repaying the loan, It could be possibly they might be taking the new loan for the annuity itself.
- Applicants who has taken loan amount below 500K previously are having trouble paying the loan now.
- If they have a credit history less than 5 Lakhs the chances are, they might be having trouble in clearing the current loan.
- For the applicants who have taken loan less than 500K worth of goods price might face problem in repaying the loan now.
- **Applicants who has less down payment history has more chances of defaulting the loan or delaying the loan payment for current one.**

Inferences

- From prior analysis on Age_In_Years, we can infer that the applicants around age of 30 are more inclined to default. - Assumption: This could be because of the less savings or inexperienced financial planning.

- Applicants who might delay the repayment, most of the are in the income range between 100K to 250K (who fall under non-taxable bracket in India).

- Applicants with higher Income could have reactively higher credit and annuity but chances of them defaulting/ late payment is less.

- The price of Goods for which they are taking the loan is in the range of 100K, 750K (1 to 3 times of the income earned).

- From prior analysis, the correlation between the Income and Loan is not high which meant that the people with relatively low income are also taking more loans.
- As analyzed the Number of queries being sent to the credit bureau, we can infer that whenever there are more queries, the applicants have more difficulty in repaying the loan. This means when bank wanted to query more times for credit score and finally got the loan disbursed the chances are the applicant might default is high
- The Applicants who were having trouble repaying now are having a chance that they were in X-sell for a previous loan. Assumption: This might be because we have sold the financial product which they might not require.

Thank You