Credit Data EDA

Shalini Jain

Content

Problem statement

1. Aim is to identify patterns which indicate if a client had difficulty paying their installments which will help the bank in taking following actions:
   a. Denying the loan
   b. Reducing the amount of loan
   c. Lending at higher interest rate to risky applicants etc
2. Identifying the correlation between dependent variables and Target variable
3. To ensure that consumers who are capable of paying the loan are not rejected

Data loading

1. Imported all the necessary libraries for EDA like pandas, numpy, seaborn, matplotlib etc
2. We are provided with 3 Data files:
    a. Application_data.csv - it contains all the information of a client at the time of application
    b. Previous_application_data.csv - contains information about client's previous loan application
    c. Column_Description.csv - contains metadata of above two files
3. We will be working on Application_data.csv and Previous_application_data.csv for our analysis
4. There are 307511 rows and 122 columns in Application dataset
5. 1670214 rows and 37 columns are present in previous application dataset

Application Data Cleaning 1

1. Check for missing values in the dataset
   a. There are zero columns with null values in integer type columns
   b. 6 Category columns have missing values
   c. More than 50 float type columns have missing values
   d. Removed all the columns having more than 40% missing values
2. Removed All the columns with more than 40% missing values
3. Two category columns - Occupation_Type, Name_Type_Suite and 16 float type columns are left with missing values
4. Drop the columns which are insignificant for our analysis - we would drop column starting with FLAG_DOCUMENT and columns with phone related FLAGs .
5. All columns starting from days are converted to positive values
6. Binning:
   a. AMT_INCOME_TYPE and AMT_CREDIT_TYPE are bucketed in low, medium high etc to have easier mapping with Target variable
   b. Age_group column is derived from Days_Birth column and bucketed into young to senior citizens group

Application Data Cleaning 2

7. Missing data Imputation -

a. Discrete numeric missing values are imputed with mode values
b. Continuous missing values in EXT_SOURCE_3, EXT_SOURCE_2 are imputed with median values
c. Name_Type_Suite missing values and Code_Gender ‘XNA’ values are imputed with mode value
d. As ORGANIZATION_TYPE ,NAME_INCOME_TYPE and Occupation_Type columns contain work related information of client. We will impute the missing records in these columns by finding the relation between them.
e. Customer with ORGANIZATION_TYPE 'XNA' has NAME_INCOME_TYPE Pensioner, so we would replace XNA with Pensione
f. 57% values of occupation_type are missing where NAME_INCOME_TYPE is Pensioner. So we would replace these values with Pensioner
g. Most of the remaining missing values clients in occupation_type are doing job or business, so we would replace these missing values with 'Working'

Previous Application Data Cleaning

1. Removed All the columns with more than 40% missing values
2. AMT_GOODS_PRICE, AMT_ANNUITY, PRODUCT_COMBINATION,CNT_PAYMENT are left with some missing values.
3. Missing data Imputation -
    a. Imputed AMT_GOODS_PRICE, AMT_ANNUITY missing values with median values
    b. PRODUCT_COMBINATION missing values are imputed with mode
    c. CNT_PAYMENT are Term of previous credit at application of the previous application. This column contains missing values for those records where loan wasn't provided. So, we would impute these rows with 0

Outlier analysis

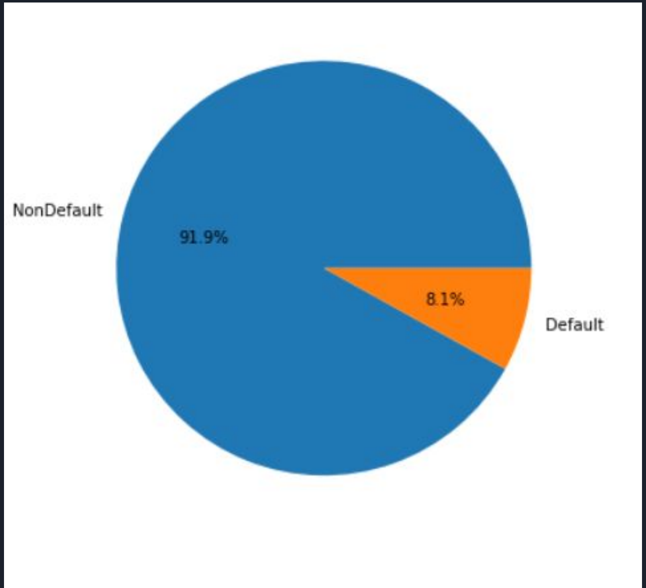
We plotted Boxplot for all the numerical columns and checked for outliers.

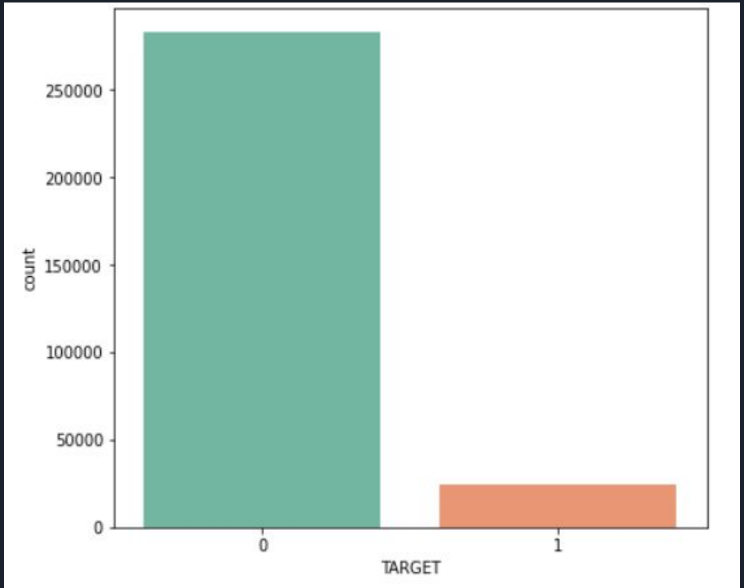
Insights:

1. AMT_INCOME_TOTAL,AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE are having much higher values than IQR, but these values shouldn't be considered as outliers as these are possible values.
2. Third quartile of DAYS_REGISTRATION AND DAYS_LAST_PHONE_CHANGE is larger as compared to the First quartile and all have a large number of outliers.
3. IQR for DAYS EMPLOYED is very slim. Most of the values are present below 25000. And an outlier is present 375000.
4. DAYS_BIRTH ,DAYS_ID_PUBLISH and EXT_SOURCE_2,EXT_SOURCE_3 don't have any outliers.
5. Boxplot for DAYS_EMPLOYED ,OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE,OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE,AMT_REQ_CREDIT_BUREAU_HOUR,AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK,AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT and AMT_REQ_CREDIT_BUREAU_YEAR are very slim and have a large number of outliers.

Data Imbalance

Application data is highly imbalanced. Default population is 8.1 % and non-default population is 91.9% and Imbalance Ratio is 11.3





Top 10 Correlation of Defaulters

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998270
AMT_CREDIT	AMT_GOODS_PRICE	0.982783
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869016
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_GOODS_PRICE	0.752295
AMT_CREDIT	AMT_ANNUITY	0.752195
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.497937

We will be using 'AMT_CREDIT','AMT_GOODS_PRICE', 'AMT_ANNUITY', 'CNT_CHILDREN' for our analysis as these are more significant for our business case

Top 10 Correlation of Non-Defaulters

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998510
AMT_CREDIT	AMT_GOODS_PRICE	0.987022
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878571
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861861
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.859371
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
AMT_ANNUITY	AMT_GOODS_PRICE	0.776400
AMT_CREDIT	AMT_ANNUITY	0.771276
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.539005

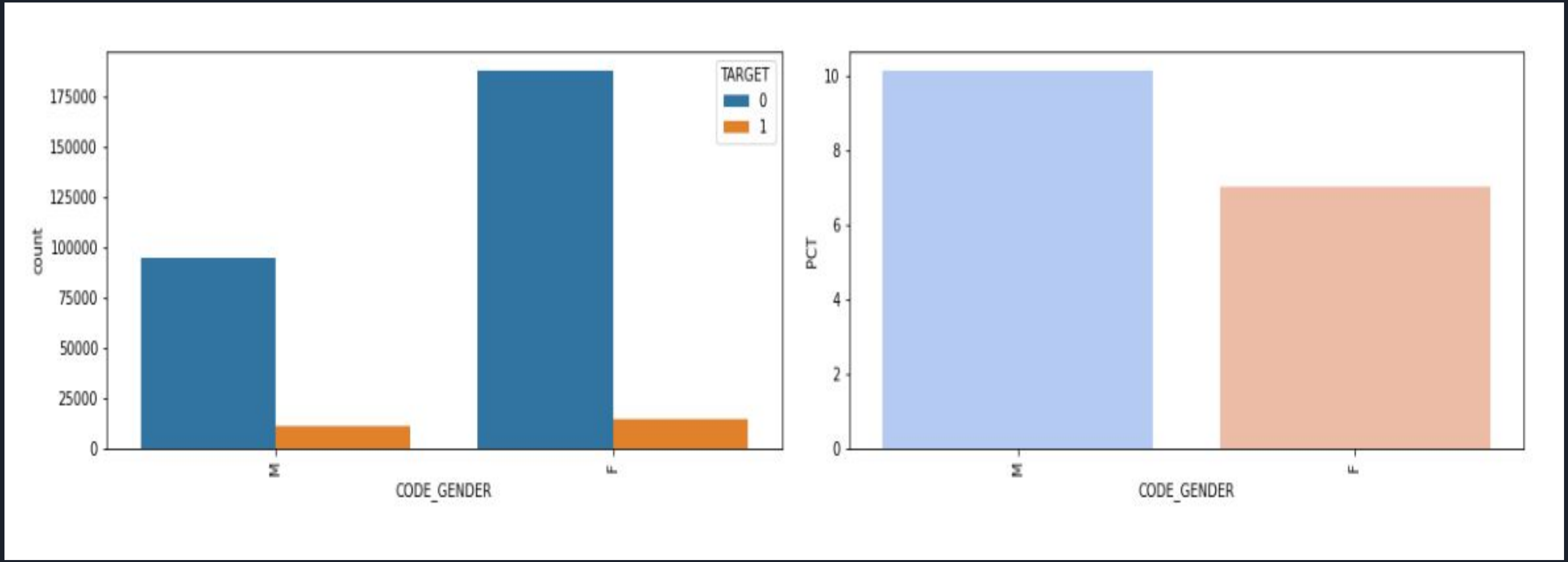
We will be using 'AMT_CREDIT','AMT_GOODS_PRICE', 'AMT_ANNUITY', 'CNT_CHILDREN' for our analysis as these are more significant for our business case

Data Visualization

1. Univariate Analysis :
    a. All the categorical columns
    b. Amount columns
    c. Top 3 correlated numerical columns
2. Bivariate Analysis
    a. Top 3 correlated numerical columns
3. Multivariate Analysis
    a. Income and education type
    b. Income and previous application status

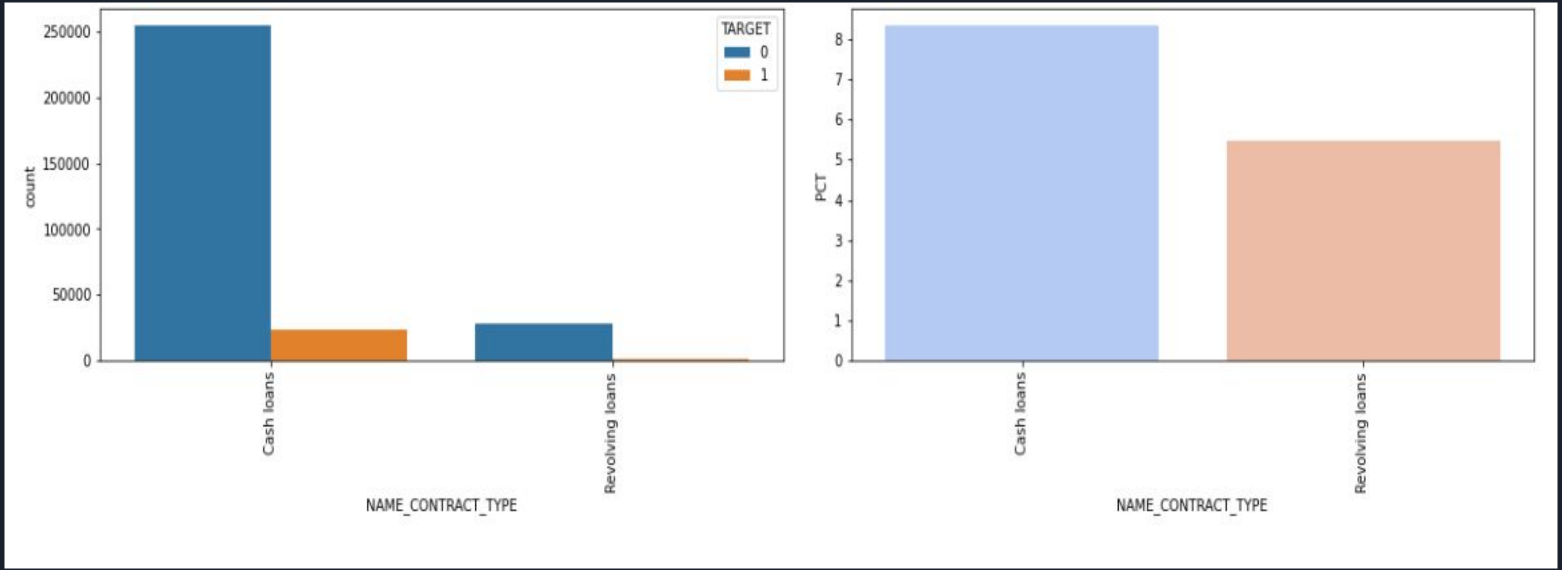
UNIVARIATE ANALYSIS

Target vs Gender



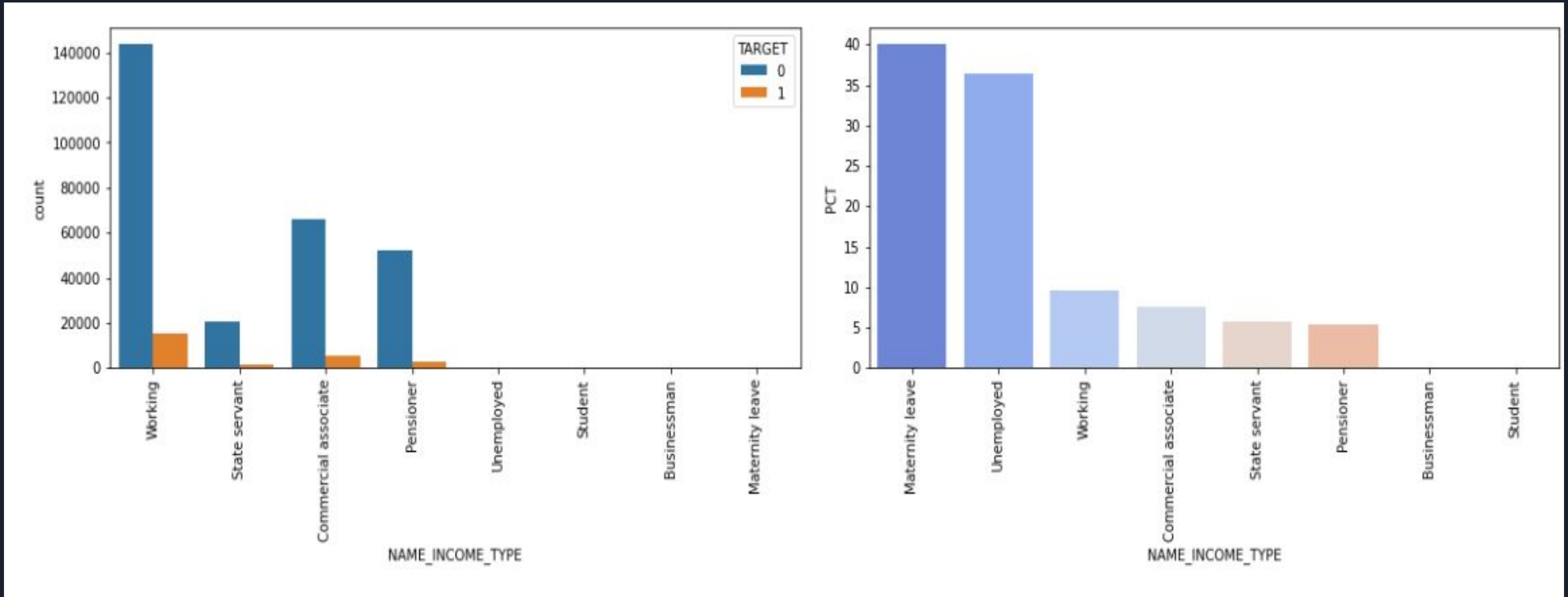
Most of the loans have been taken by female Default rate is higher in Male clients(~10%) compare to female clients(~8%)

Target vs NAME_CONTRACT_TYPE



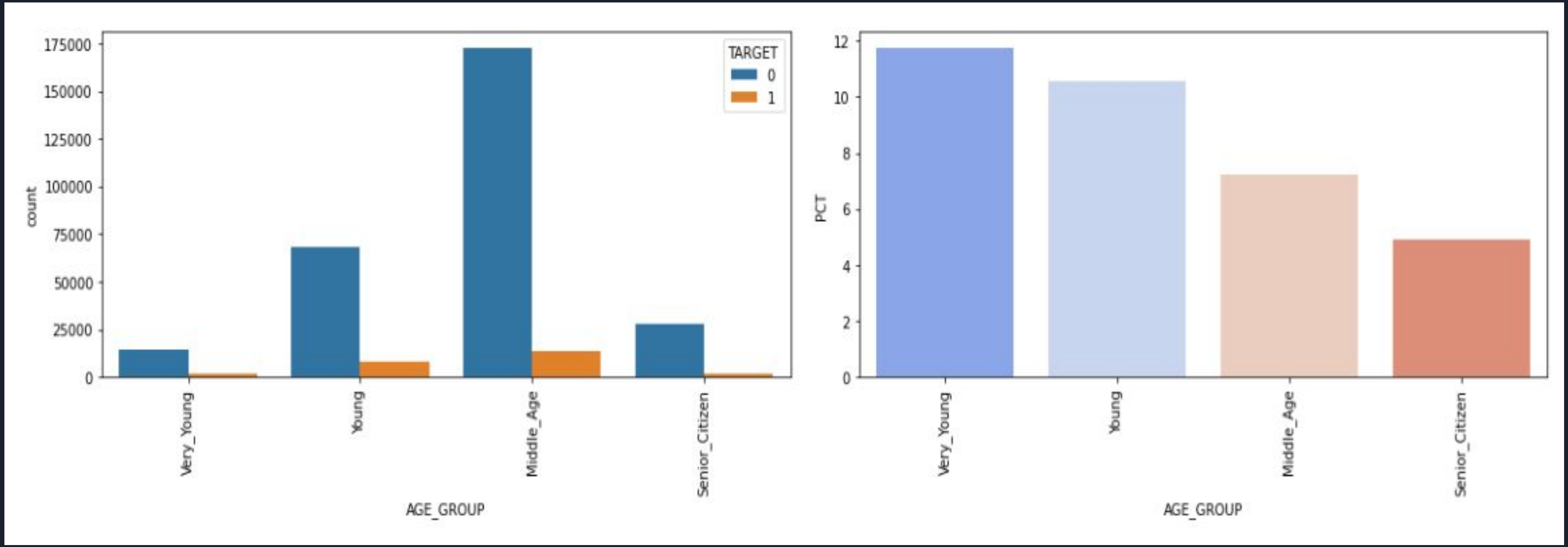
Most of the customers have taken cash loan. Customers who have taken Revolving loans are less likely to default

Target vs Income Type



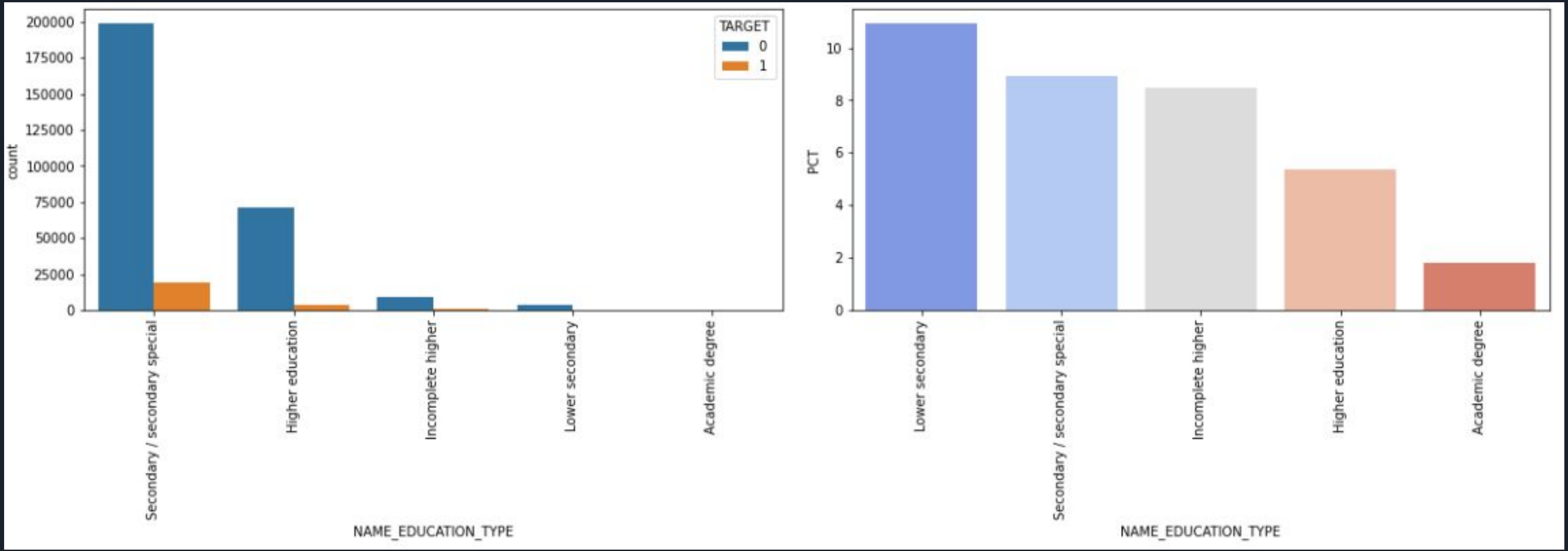
Most of the loans are taken by working, commercial associates and pensioners with very less default rates

Target vs Age Group



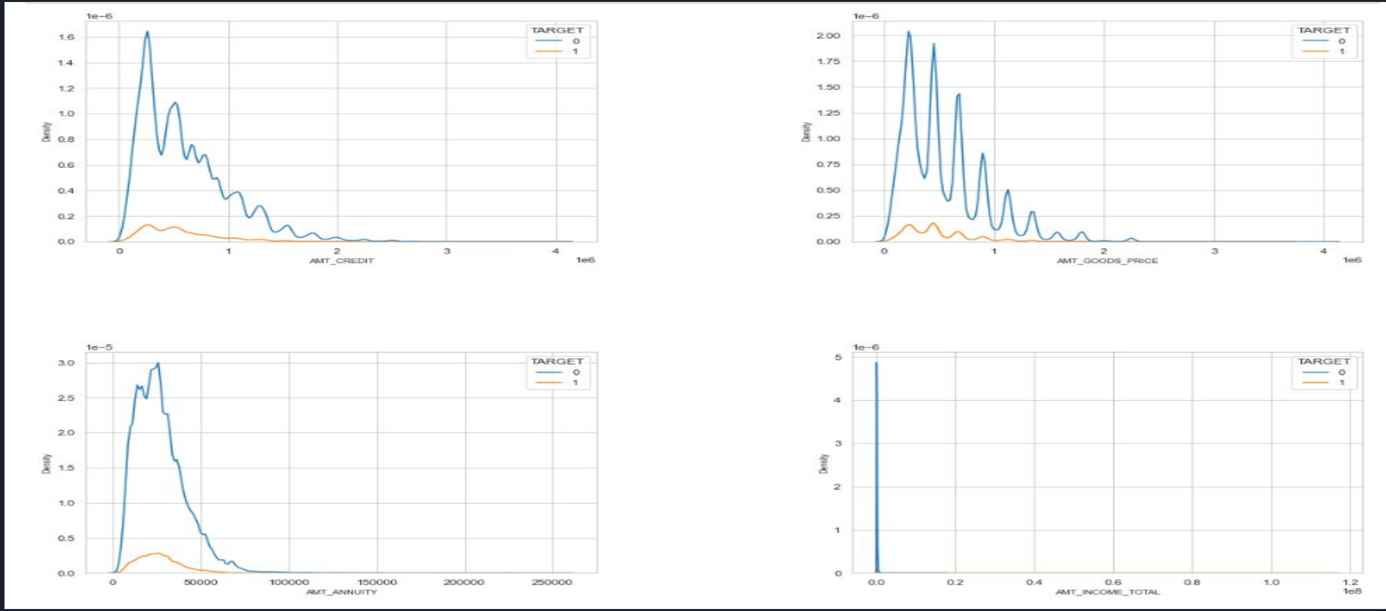
Deafult rate is decreasing with age. Middle age clients are less likely to default.

Target vs Education Type



Higher education is the safest segment to give the loan with a default rate of less than 5%

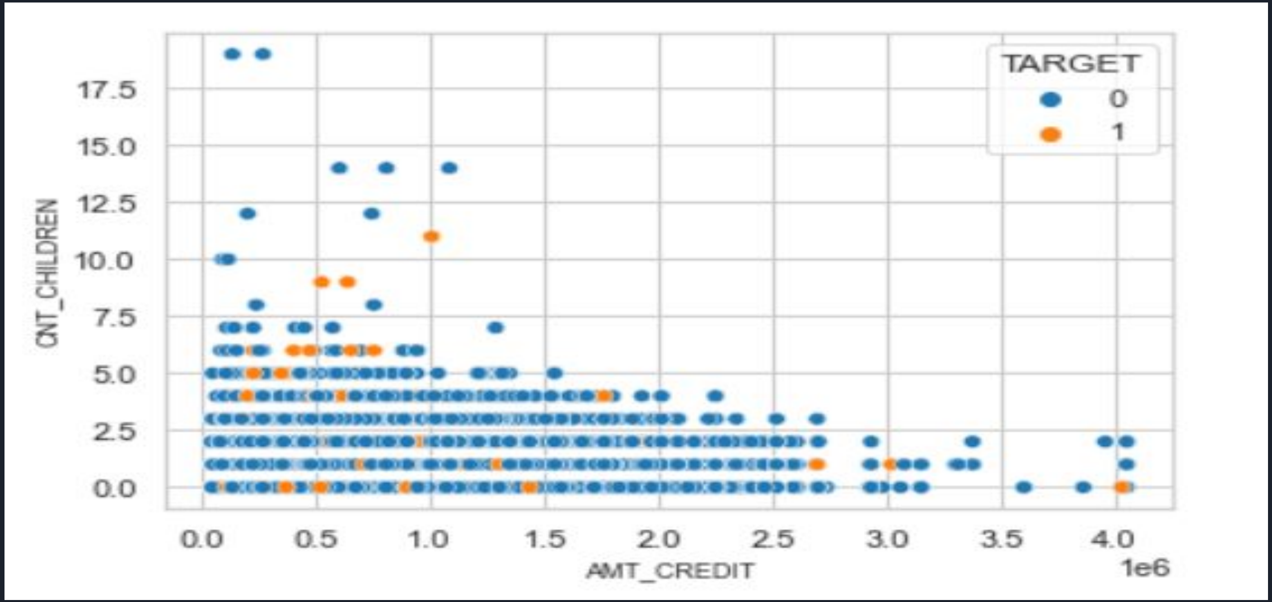
Univariate analysis on numeric variables



1. most of the loans were given for the goods price ranging between 0 to 1 mn
2. most of the loans were given for the credit amount of 0 to 1 mn
3. most of the customers are paying annuity of 0 to 50 K
4. mostly the customers have income between 0 to 1 mn

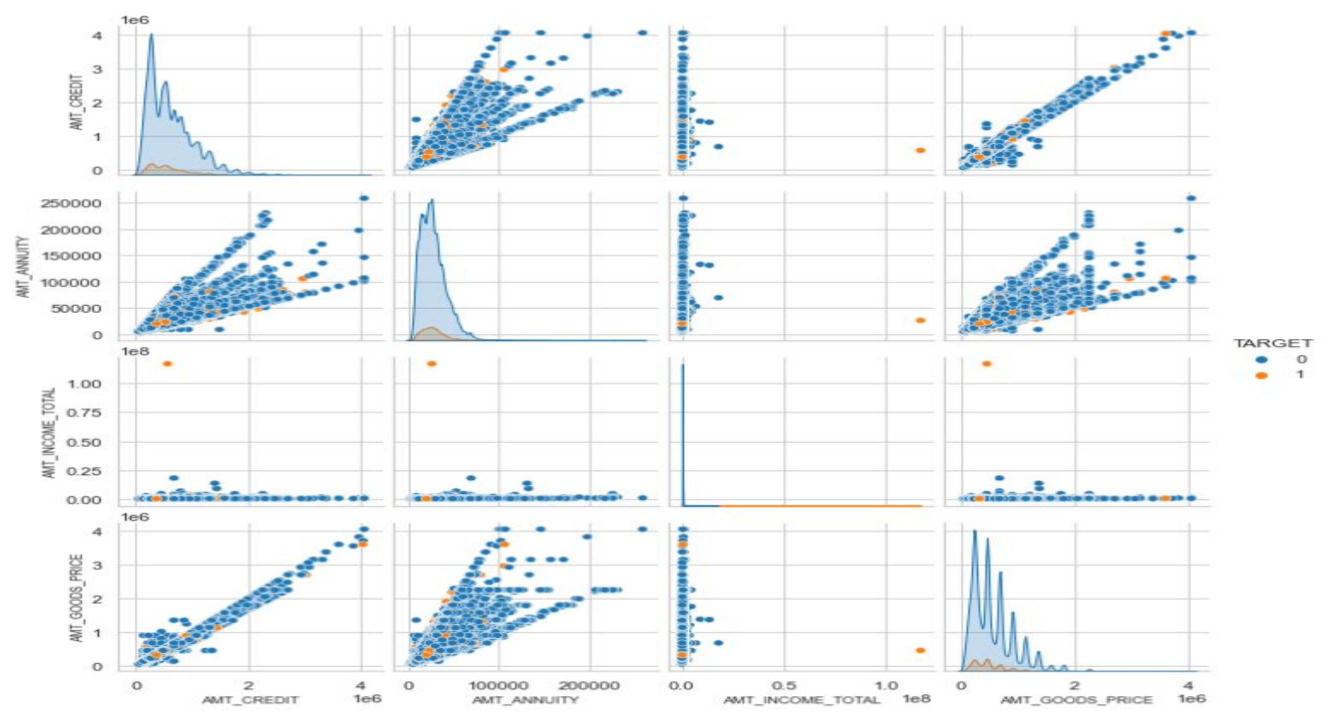
BIVARIATE ANALYSIS

AMT_CREDIT VS CNT_CHILDREN



Defaulters are high for clients having more than 4 children.

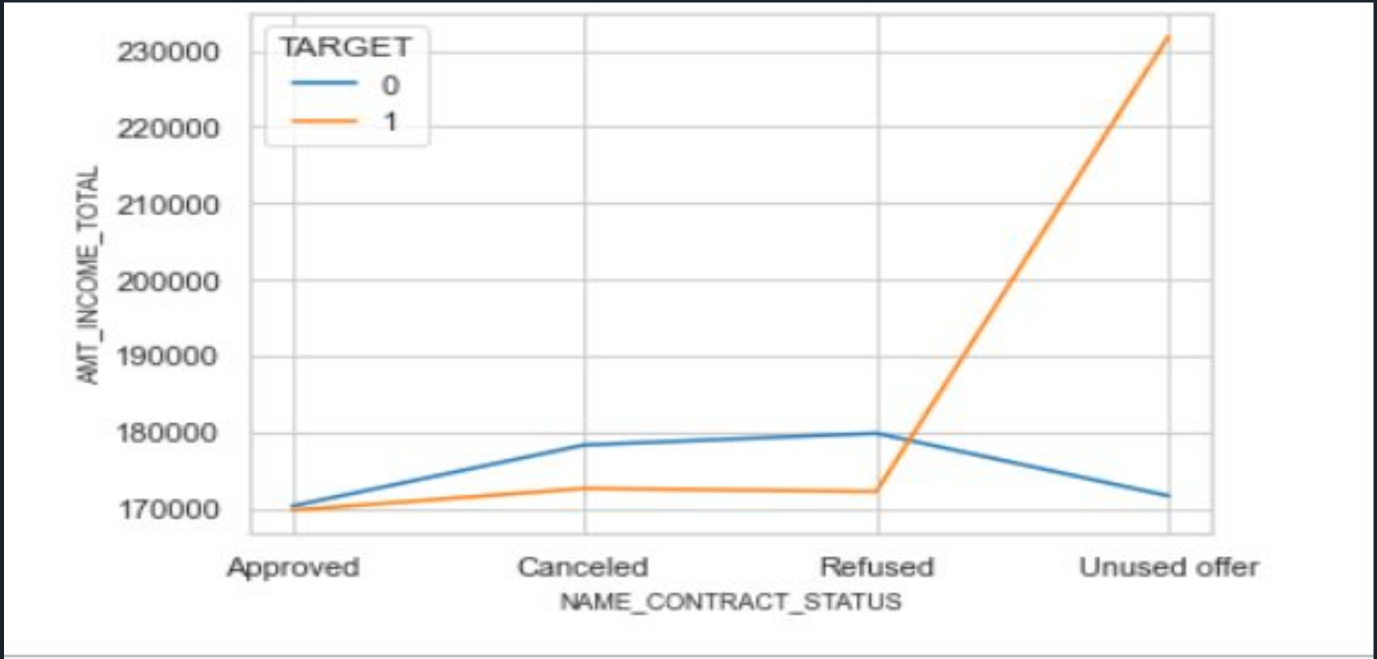
Pair plot of TARGET, AMT_CREDIT AMT_ANNUITY and AMT INCOME



1. AMT_CREDIT and AMT_GOODS_PRICE are linearly corelated, if the AMT_CREDIT increases the defaulters are decreasing
2. people having income less than or equals to 1 mn. are more like to take loans out of which who are taking loan of less than 1.5 million, coudl turn out to be defaulters. we can target income below 1 million and loan maount greater than 1.5 million
3. People who can pay the annuity of 100K are more like to get the loan and that's upto less than 2ml (safer segment)

Analysis on merged dataset

We will join these two dataset on customer ID SK_ID_CURR and check the the previous application status of the current defaulters and non-defaulters



Insights on merged dataset

1. most of the applications which were previously either canceled or refused 80-90% of them are repayer in the current data
2. offers which were unused previously, now have maximum number of defaulters despite of having high income band customers

RECOMMENDATIONS

Following are the strong indicators of default:

1. NAME_HOUSING_TYPE - People living in rented apartment
2. NAME_FAMILY_STATUS - civil marriages, single/not married
3. NAME_EDUCATION_TYPE - Lower secondary
4. OCCUPATION_TYPE - Low-Skill Laboreres and drivers
5. offers prev. unused and high income customer should be avoided
6. CNT_CHILDREN - Customers withmore than 5 children
7. NAME_INCOME_TYPE - Maternity Leave, students, unemployed

Following clients should be targeted:

1. NAME_HOUSING_TYPE - People living in their own apartment
2. NAME_FAMILY_STATUS - Married
3. NAME_EDUCATION_TYPE - Higher education
4. CODE_GENDER - FEMALE
5. OCCUPATION_TYPE - Accountants, Core staff,High skill tech staff , Managers
6. Offers prev. Cancelled or refused
7. NAME_INCOME_TYPE - Working and Pensioners

Thank You