Exploratory Data Analysis On Risk Analytics Data.

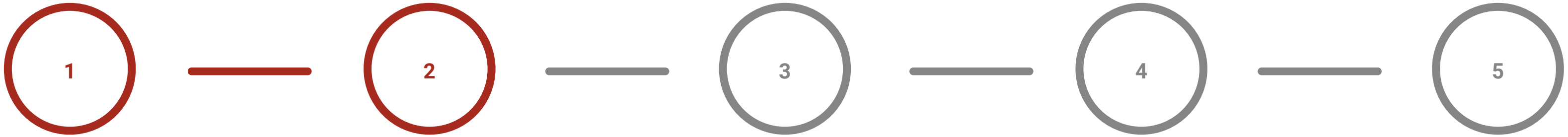
Amey Basangoudar

Problem Statement

A major problem in the fintech Industry is faced by companies that hand out loans to customers. Giving loans is a means of business for these companies. Hence, it is imperative for these companies to choose their customers smartly.

The dataset given contains details of the customer like his/her income, age, job type, etc. Based on these columns, the goal is to identify the top features that will give an estimate of which customer is most likely to default on their loans.

Methodology



1

Data Cleaning

loading the data, merging the columns, imputing null values, handling outliers and data imbalance

2

Univariate Analysis

Analysing the distribution of all the variables in the dataset

3

Bivariate Analysis

Analysing the variables with respect to the target column

4

Multivariate Analysis

Analysing multiple columns to identify best combination of features that affect the target column

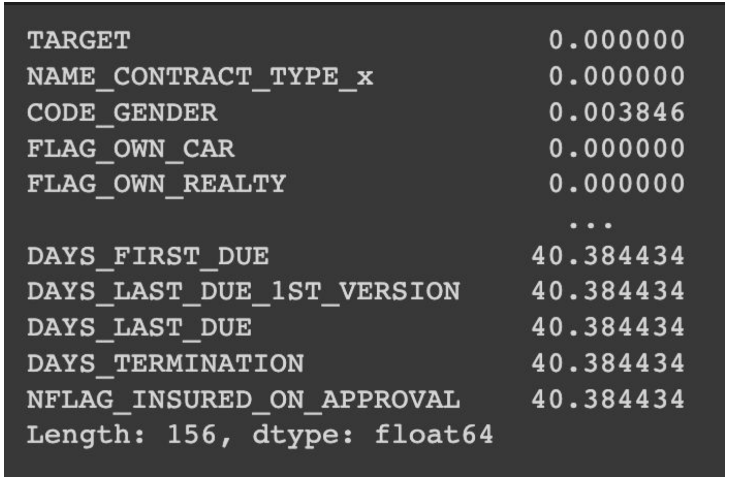
5

Conclusion

Choosing the best suited features that affect the target column

Data Cleaning - Handling Missing Values

- XNA and XAP converted to null values
- The columns with more than 20% null values are dropped
- For the columns that have less than 1% null values, the rows are dropped
- Other null values are imputed with median for numeric variables and mode for categorical variables

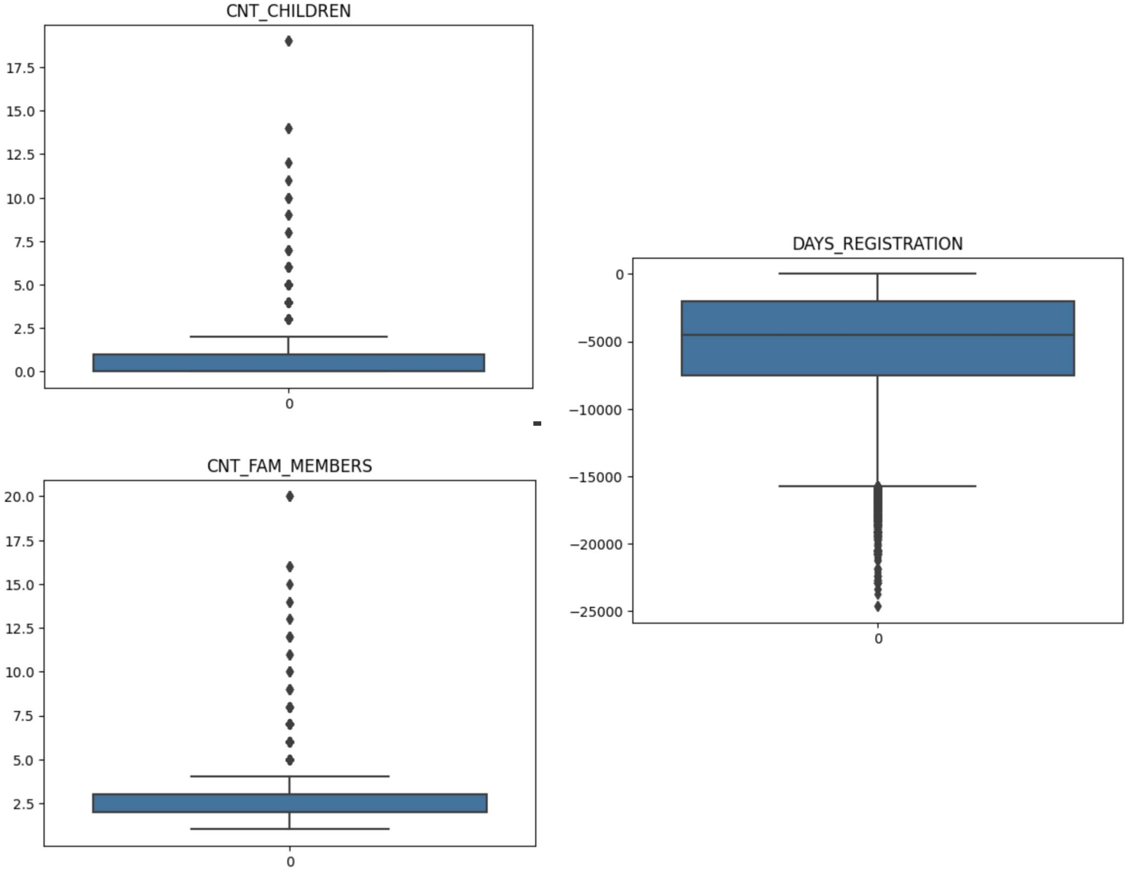


```
TARGET                          0.000000
NAME_CONTRACT_TYPE_x            0.000000
CODE_GENDER                     0.003846
FLAG_OWN_CAR                    0.000000
FLAG_OWN_REALTY                 0.000000
                                  ...
DAYS_FIRST_DUE                 40.384434
DAYS_LAST_DUE_1ST_VERSION      40.384434
DAYS_LAST_DUE                  40.384434
DAYS_TERMINATION               40.384434
NFLAG_INSURED_ON_APPROVAL      40.384434
Length: 156, dtype: float64
```

Figure 1: Subset of the columns with null values (%)

Outlier Detection

- Most of the columns have outliers
- The outliers are not imputed as there is no modelling involved
- Keeping the outliers gives a better understanding of how they affect the target variable


CNT_CHILDREN

DAYS_REGISTRATION

CNT_FAM_MEMBERS

Figure 2: Outliers of some of the columns in the dataset

Data Imbalance

- The dataset has columns that are disproportionate.
- For example the dataset has more number of females datapoints compared to males.
- Similarly more datapoints exist for people who are working and live in a house or apartment

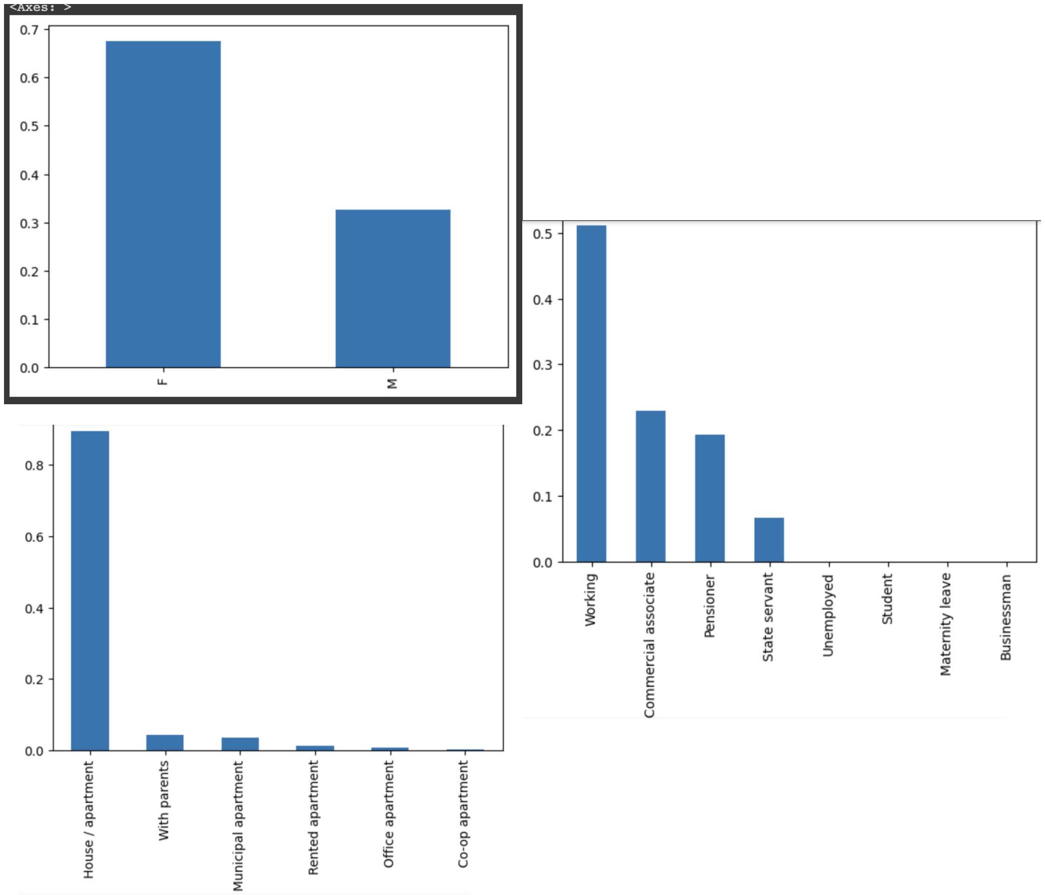


Figure 3: Data Imbalance in some columns

Univariate Analysis

- The target variable indicates if a customer is likely to default (1) or not (0).
- The dataset has more number of non-defaulters compared to defaulters.
- The age of customer is provided in days.
- The distribution of ages is gaussian where more number of customers have ages between 12500 to 17500 days i.e 35 - 50 years.

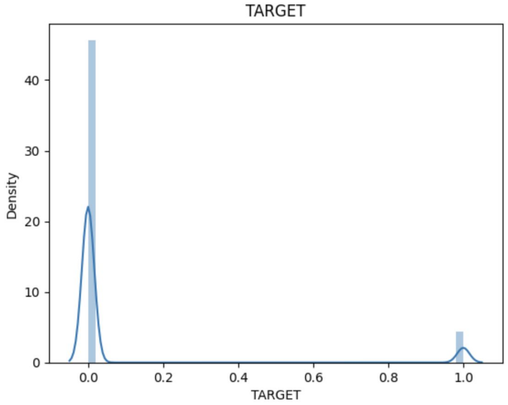


Figure 4: Distribution of Target

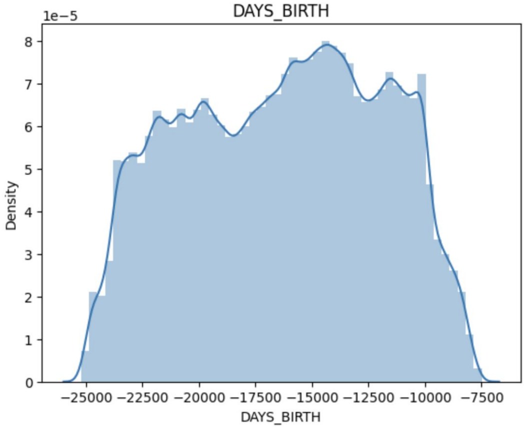
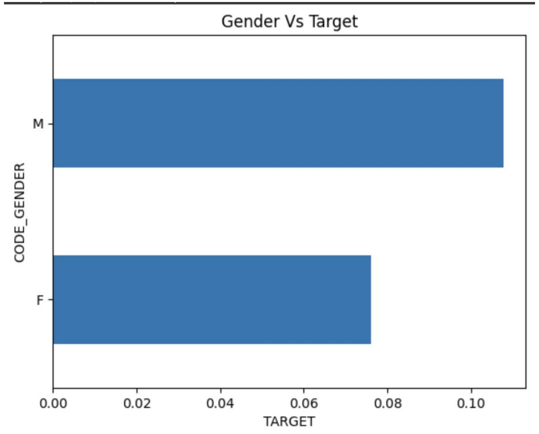
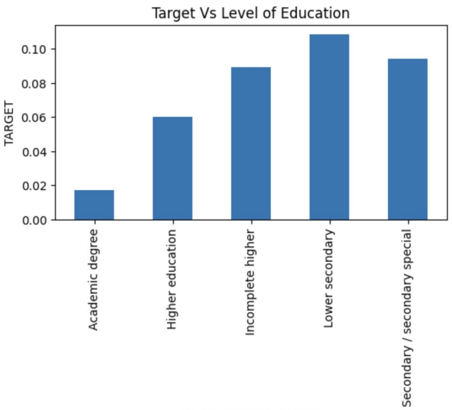


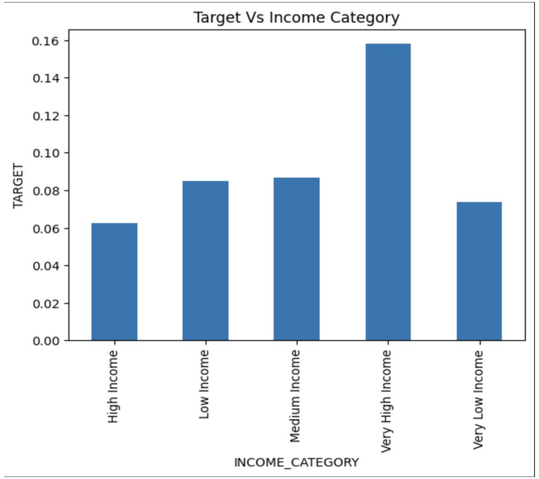
Figure 5: Distribution of Age

Bivariate Analysis

- Figure 6 shows that males have a higher chance to default
- People who have completed only lower secondary education are likely to default
- People with salaries between 10 and 50 lakhs have a high chance to default. This does not make sense as people with high incomes should be able to pay their bills on time.

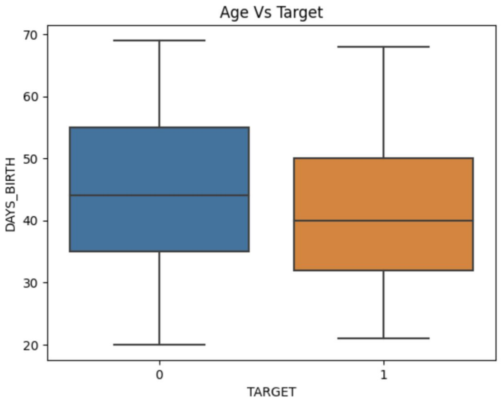

Gender Vs Target

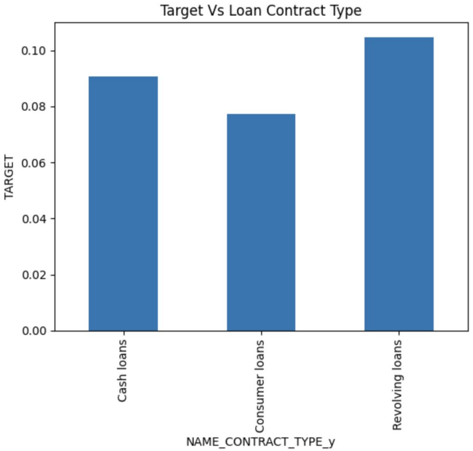

Target Vs Level of Education

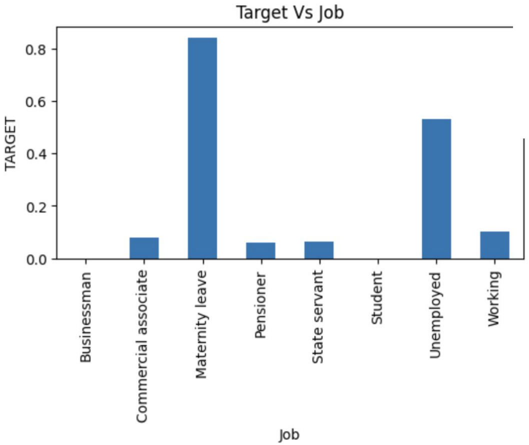

Target Vs Income Category

Bivariate Analysis

- Age does not seem to be a solid indicator of default. However people who default are slightly younger
- Revolving loans like credit cards have the highest chance of default. This makes sense as the payment is recurring and people are likely to miss their payments sometimes
- People who are unemployed or on maternity leave have the highest chance of default as they have no source of income

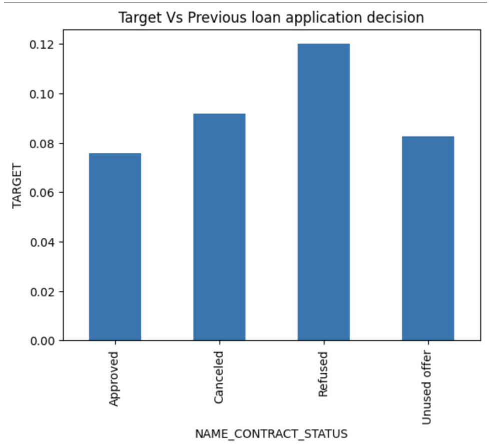

Age Vs Target

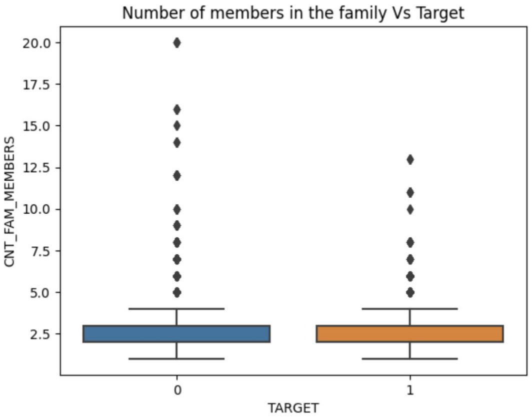

Target Vs Loan Contract Type


Target Vs Job

Bivariate Analysis

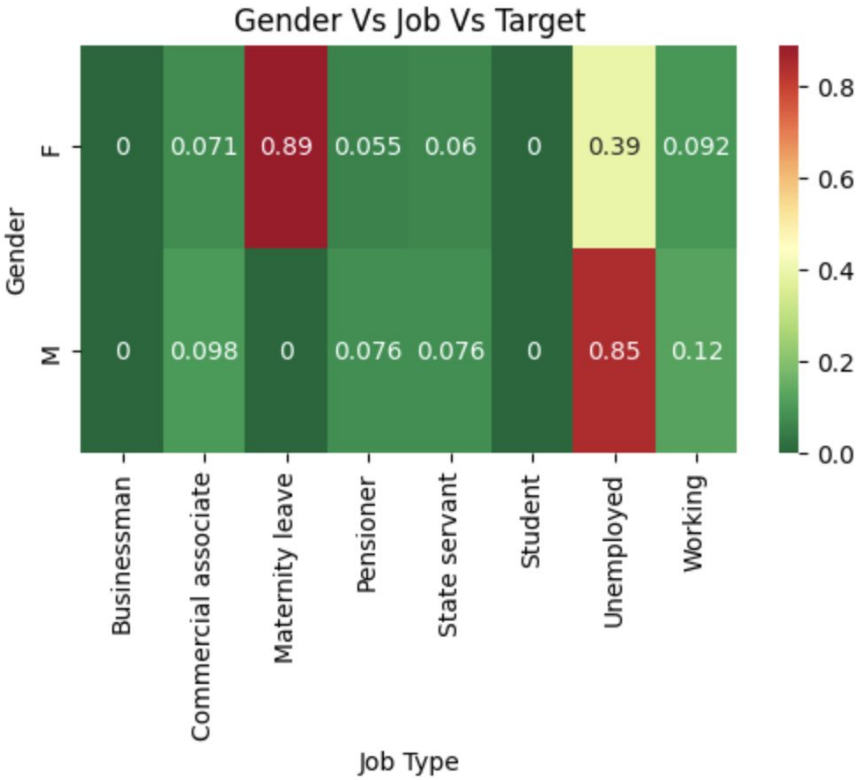
- People whose previous application was rejected have a higher chance of default
- The number of members in the family has no effect on whether a person will default or not





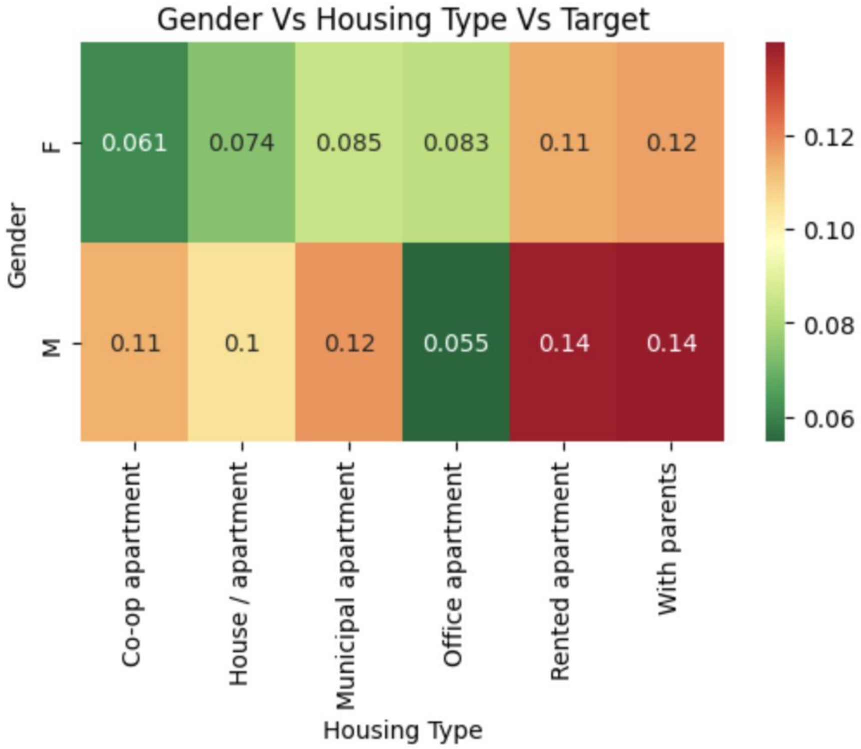
Multivariate Analysis

- Females on maternity leave have a high chance of defaulting
- Males who are unemployed also tend to default on their payments a lot
- The above two make sense as both categories have no source of income



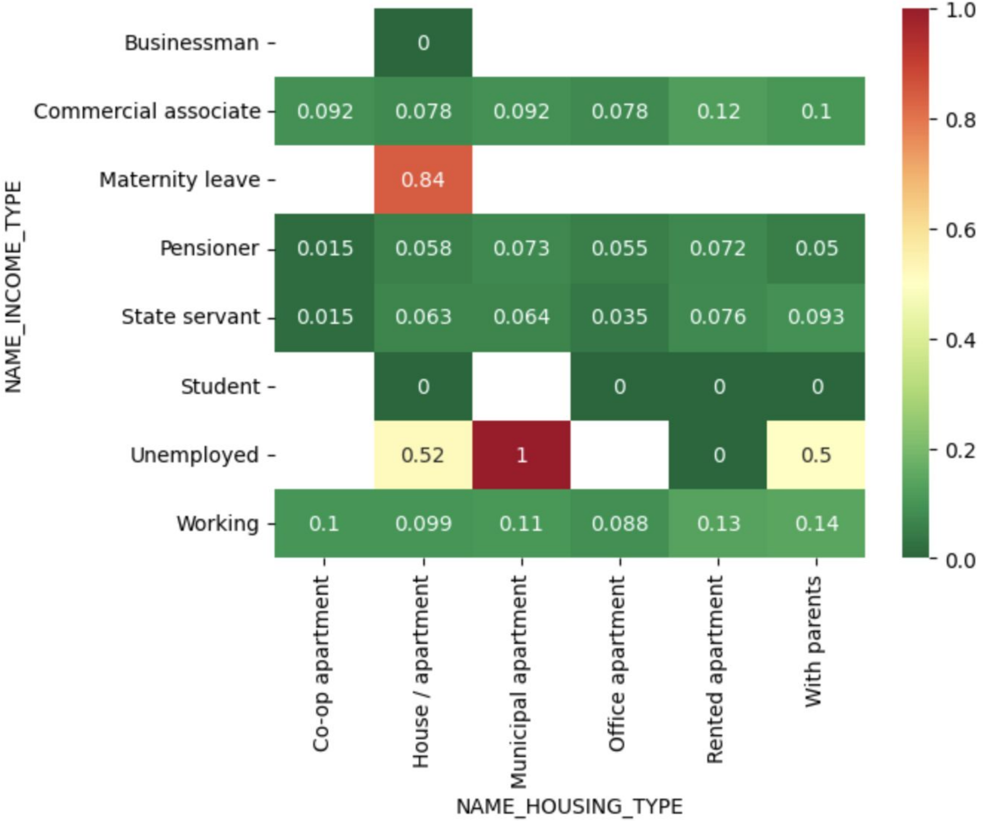
Multivariate Analysis

- From the bivariate plot we know that males have a higher chance of default
- From this plot we can see that males who either live with their parents or live in a rented apartment have a high chance of default



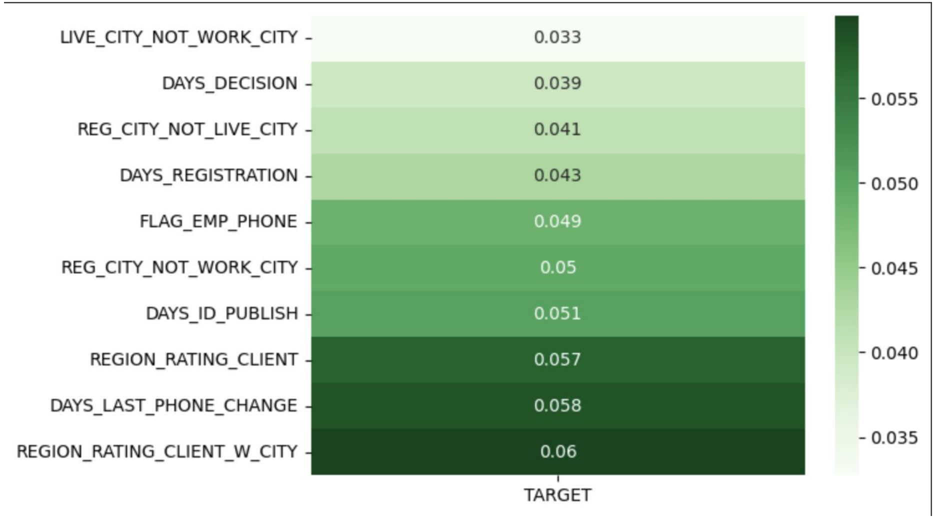
Multivariate Analysis

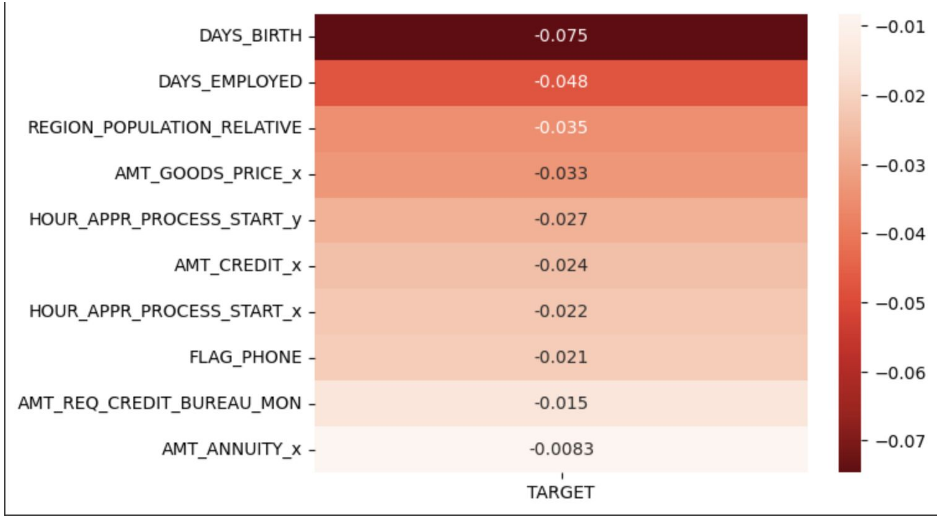
- Unemployed people who live in a Municipal apartment will definitely default on their loans
- Women on maternity leave who live in a house or apartment have a high chance to default



Correlated Variables

- The plots below show the most positively and negatively correlated features





Conclusion

- From the analysis it is evident that banks should never grant loans to unemployed people especially if they live in a municipal apartment
- The bank must also not grant loans to women on maternity leave as they have a very high chance to default
- Also male applications must be scrutinized more as they are more likely to default
- It is essential to look at the applicant’s level of education as lower the level of education higher is the chance to default