CREDIT EDA ASSIGNMENT

RAGUL

PROBLEM STATEMENT

- To understand the driving factors behind the loan default and identification of variables which are strong indicators of the default.

- Identify the patterns which indicate if a client has difficulty in paying their instalments which may be used for taking actions such as denying the loan, lending at higher interest rate, etc.

DATA SET

- application_data.csv – contains all information of the clients at the time of application.
- previous_application.csv – contains information about client's previous loan data.
- column_description.csv – describes the meaning of the variables.

IMPORTING LIBRARIES AND DATASET

Importing the libraries

```
#importing necessary libraries
import pandas as pd, numpy as np
import matplotlib.pyplot as plt, seaborn as sns
```

```
#importing warnings
import warnings
warnings.filterwarnings("ignore")
```

Reading the Data set

```
#reading the CSV file and creating a dataframe for application data
appl_data = pd.read_csv("application_data.csv")
appl_data.head()
```

Initial step involves importing the necessary libraries and warnings

> Numpy & Pandas
> Matplotlib
> Seaborn

Reading the dataset and creating the dataframes

> appl_data
> prev_appl

APPROACH AND METHODOLOGY

- Initiated the data cleaning by inspecting the dataframes. Examined the shape and information of dataframe.
- Used describe function to understand the statistical summary of the dataframe.
- Further proceeded with data cleaning by handling null values in the dataset. application_data had null values in 67 columns and previous_application has null values in 16 columns.
- Dropped the columns with null values more than 30% in both dataframes.
- Proceeded with imputation of missing values. For the categorical columns, imputed the missing values with mode of the variable. For example, in NAME_TYPE_SUITE column, Unaccompanied value has been filled in null values.

APPROACH AND METHODOLOGY

- Null values in numerical column has been handled in two ways. For the columns with outliers, null values has been imputed with median of the variable. Eg: AMT_ANNUITY column.
- For the columns without any outliers present in it, null values has been replaced with mean of the variable. Eg: EXT_SOURCE_3 column.
- Further continued the data cleaning with handling outliers in columns. Used boxplots for deduction of outliers in the columns. For the AMT_INCOME_TOTAL column, a significant difference between 98th percentile and maximum value has been witnessed. Proceeded with dropping the records beyond 98th percentile of the "AMT_INCOME_TOTAL" variable.
- Similarly, for the CNT_CHILDREN column, records beyond count of children more than 6 numbers has been dropped.

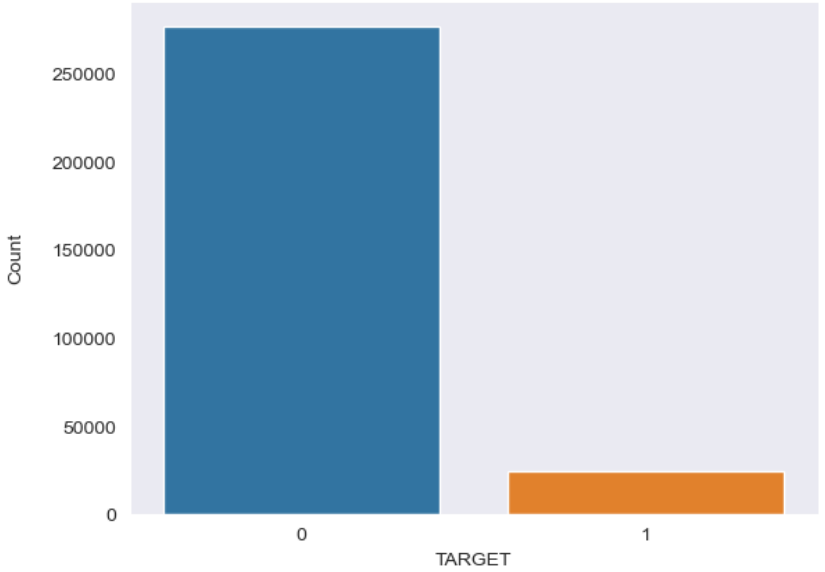
APPROACH AND METHODOLOGY

- Fixing the datatypes of the columns: some numerical columns in application dataset had to be in “int” datatype instead of float. Eg: 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE’ columns into integer.
- Proceeded with dropping the unnecessary columns in the dataset. Eg: 'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3’, etc has been dropped.
- Creation of new column: “AGE_IN_YEARS” column has been created from the column “DAYS_BIRTH” in which age of the clients has been provided in days.
- Checking the consistency of the records: Inconsistent values in the columns has been replaced with mode in categorical variable. Eg: CODE_GENDER has value XNA which is invalid. Hence, replaced as “F” which is mode of variable.

APPROACH AND METHODOLOGY

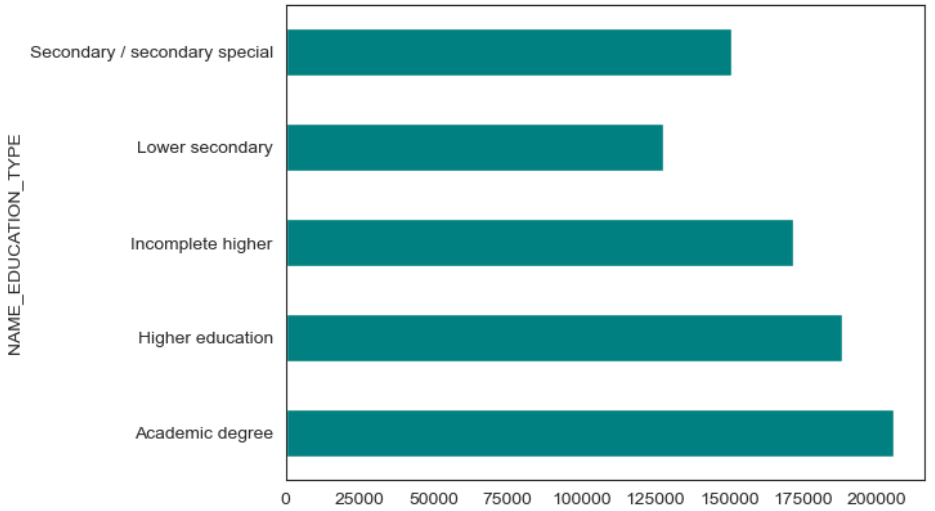
- Checking imbalance in the data: Certainly, imbalance in the data could be deducted. Imbalance Ratio (IR), defined as the ratio of the number of instances in the majority class to the number of instances in the minority class. In the given data, IR is 11.31
- Further continued with univariate, bivariate and multivariate analysis.
    - Categorial Univariate Analysis
    - Numerical Univariate Analysis
    - Numerical & Categorical Bivariate Analysis
    - Numerical & Numerical Bivariate Analysis
    - Multivariate Analysis
- Found the top 10 correlation for defaulters and non-defaulters.
- Found the insights from the graphs, plots and charts.

GRAPHS & INSIGHTS

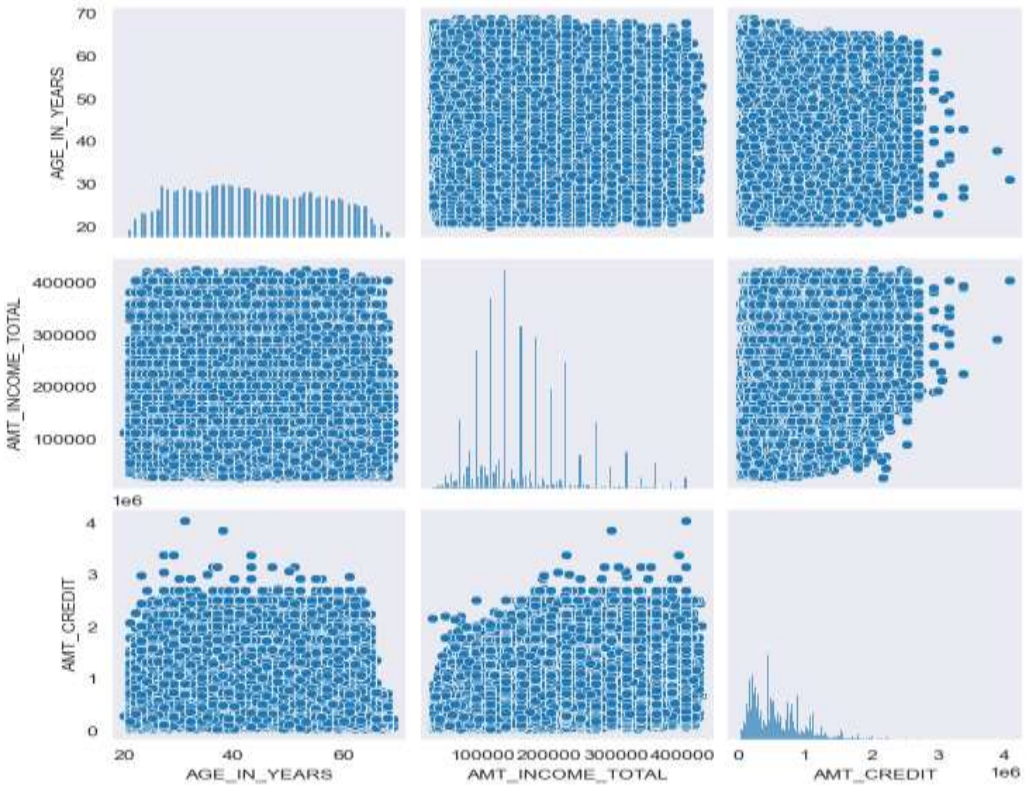


Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations. Certainly, imbalance in the data could be witnessed from the plot. 0 (Non-Defaulters) is the majority class and 1 (Defaulters) is the minority class. Imbalance Ratio for dataset is 11.31

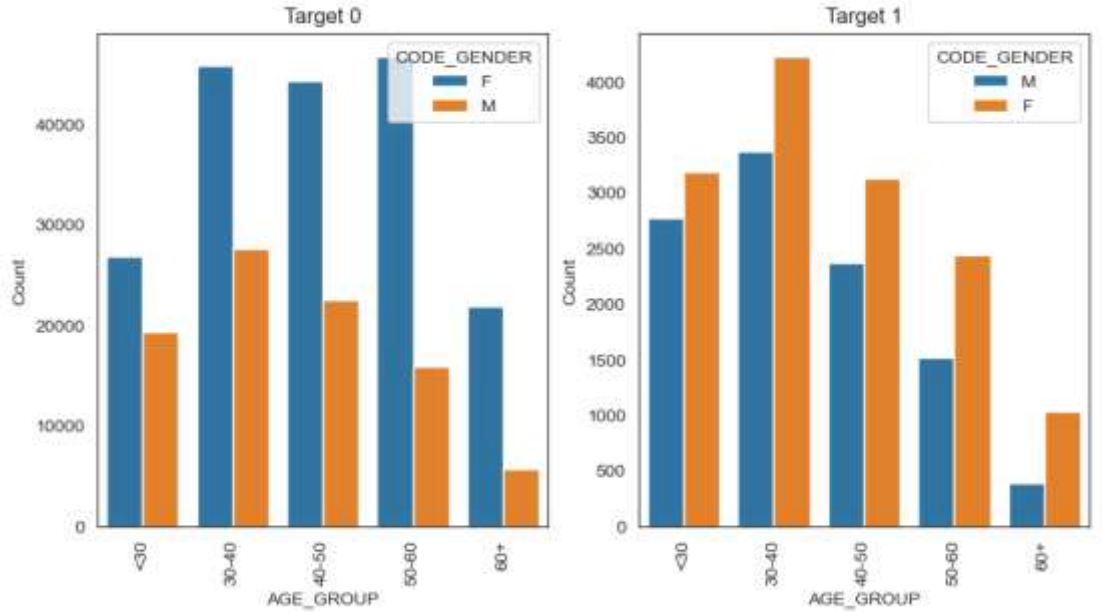
Average income of academic degree holding clients is highest among the others: 2,05,550



GRAPHS & INSIGHTS



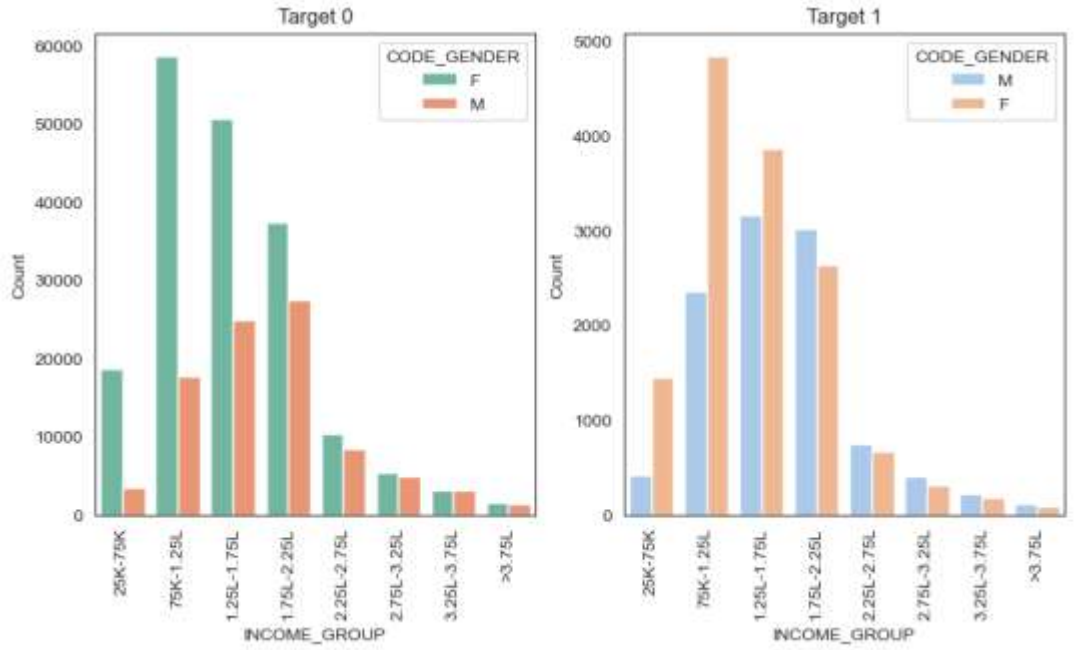
Clients of high income group (>200000) has loan credit amount greater than 27L. No strong relation or pattern could be seen between income and loan credit amount



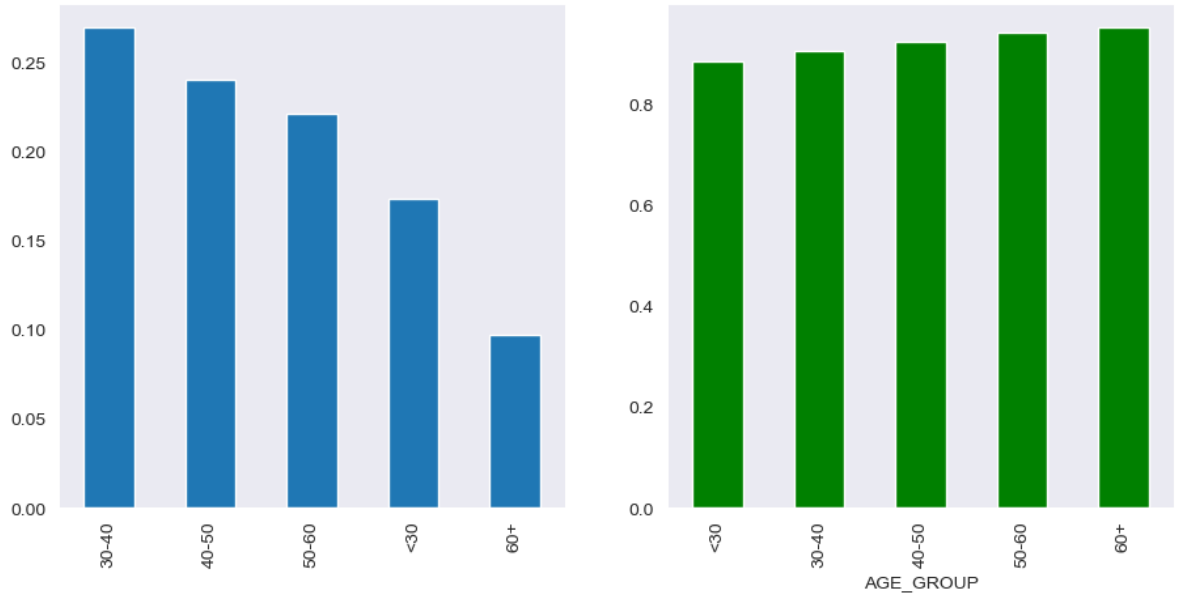
In gender based classification on age group of clients,

1. Among defaulters, 30-40 age group has high occurrence; for non-defaulters, uniform spread across various age group could be seen.
2. In defaulters, 60+ age group clients accounts a low percentage of 5.86% and 30-40 age group accounts 31.04%

GRAPHS & INSIGHTS

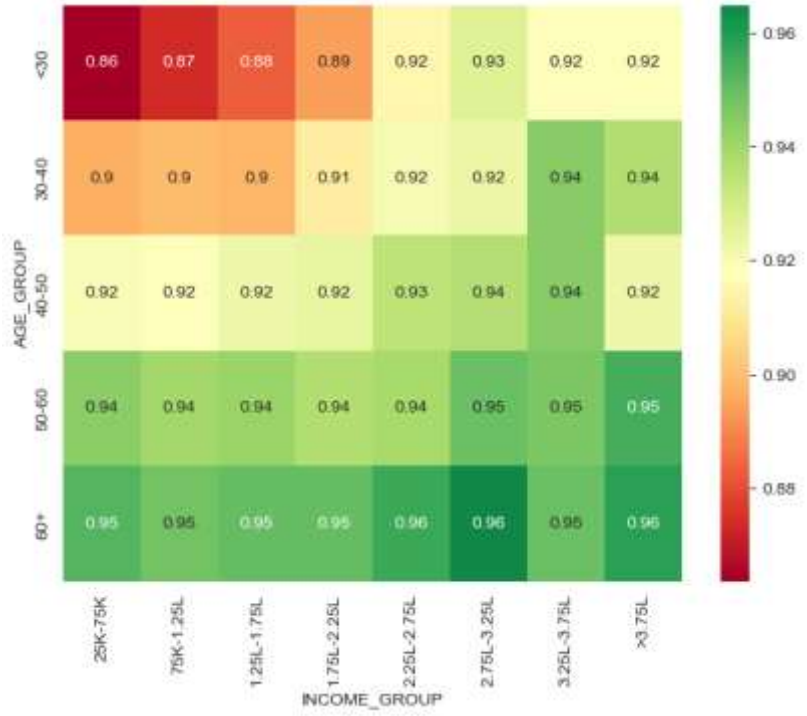


1.Both among the defaulters and non-defaulters, income range of 75,000 to 2,25,000 accounts for 80%(approx).
2.Very low occurrences could be seen beyond income range of more than 3,75,000
3.Among the income range of 1.75L to 2.25L in defaulters, male counts is higher in occurrence.

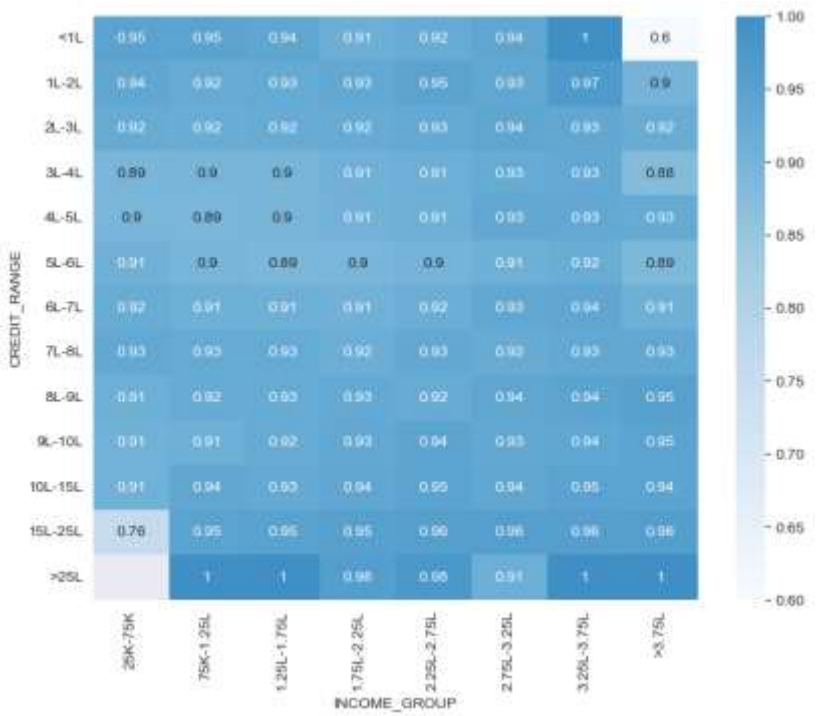


Clients among the age group 60+ age accounts as low as 9.67%, however the mean of non-default for 60+ aged population is marginally higher than others.

GRAPHS & INSIGHTS

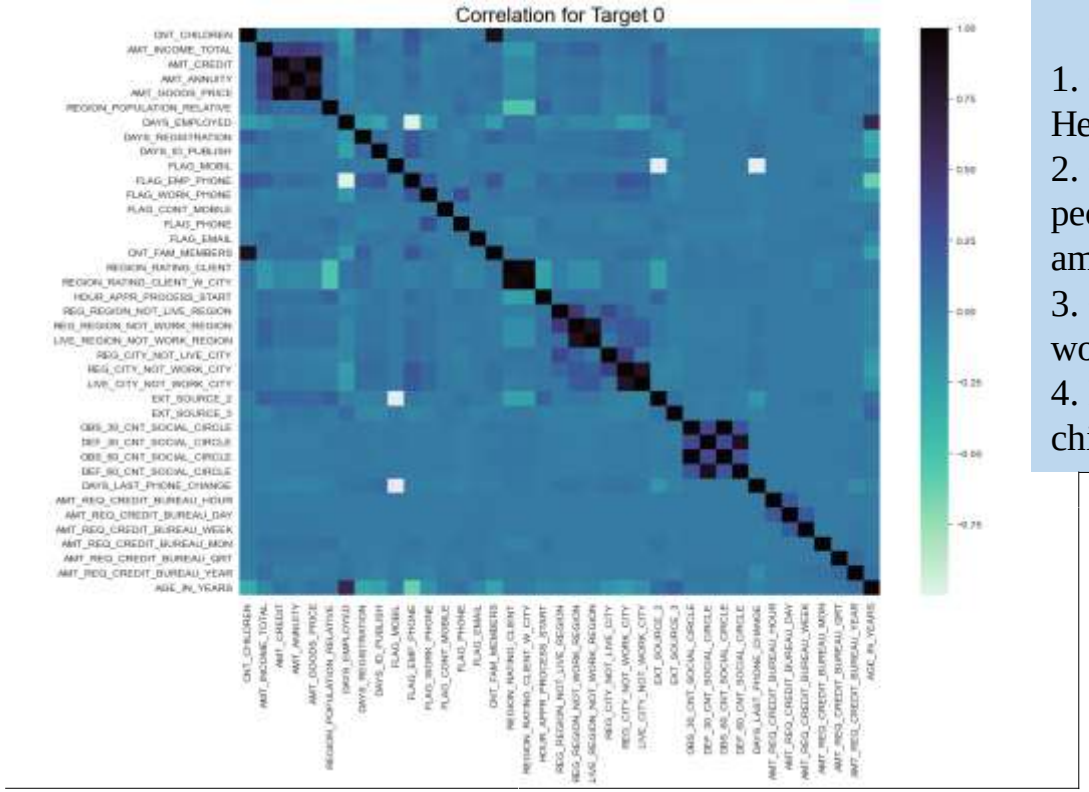


Clients of age group 60+ and belonging to income group 2.75L to 3.25L has the highest average of non-default. Similarly, clients falling under age group of <30 and belonging to income group of 25,000 to 2.25L has lowest average of non-default.



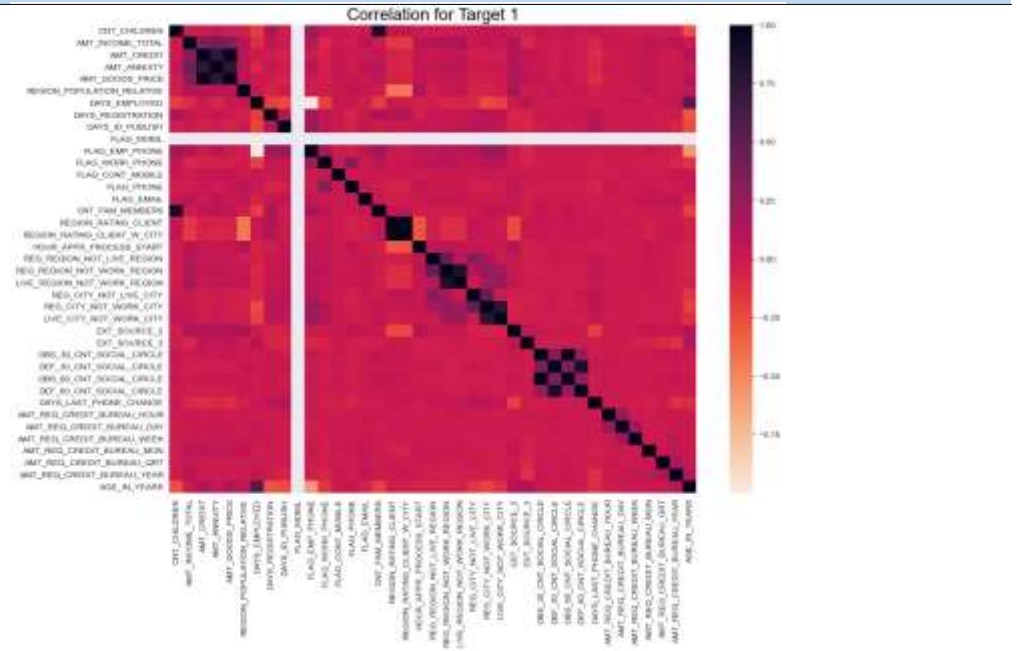
Clients belonging to income group between 25,000 to 75,000 and credit loan amount range of 15,00,000 to 25,00,000 has lowest average of non-default. Similarly, among the income group greater than 3,75,000 and credit loan amount range of less than 1,00,000 has lowest mean of non-default.

GRAPHS & INSIGHTS



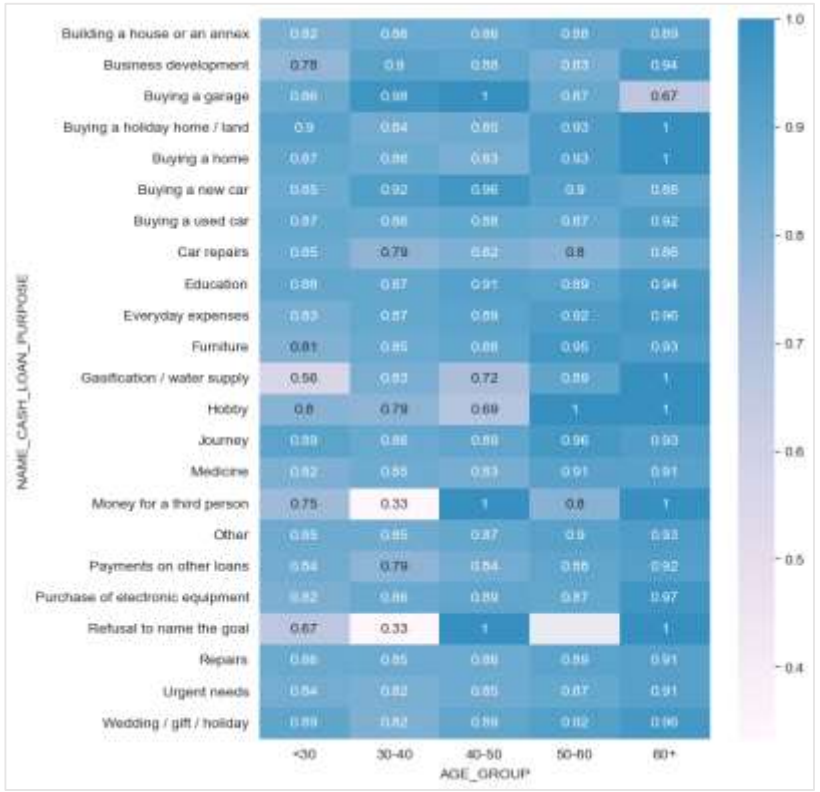
Insights from the correlation for Target 0

1. Clients belonging to age group more than 60 has negative correlation. Hence, increase in age leads to decrease in total income.
2. Credit amount is inversely proportional to the days employed, which says people who have been employed for long time has requested low credit amount
3. Clients with more number of children live in less populated region. In other words, clients with less number of children live in densely populated region.
4. Increase in the days of employment for clients leads to decrease in count of children.

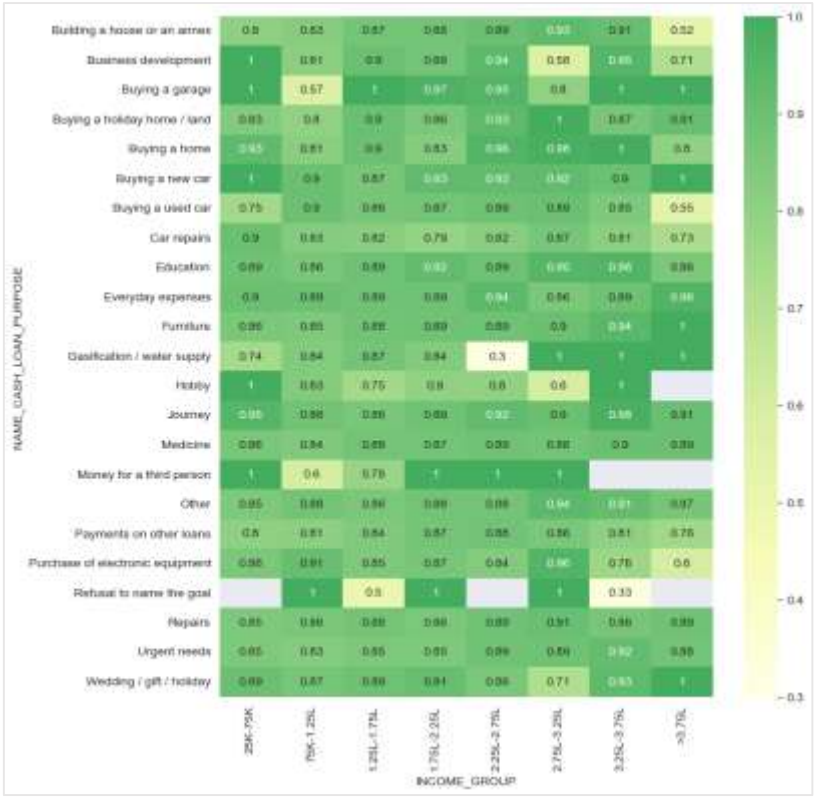


1. Increase in the number of children leads to the decrease in total income
2. Clients with less number of children have requested higher loan credit amount

GRAPHS & INSIGHTS



1. Clients belonging to age group of 40 - 50 and loan purpose as per the previous application of "Buying a garage" & "Money for third person" has highest average of non-default.
2. Similarly, clients belonging to age group of 60+ with loan purpose of "Gasification/Water Supply", "Hobby", "Buying a home", "Buying a holiday home/land" has highest mean of non-default.
3. Clients belonging to age group of 30 - 40 with loan purpose of "Money to third person" & "Refusal to name the goal" has the lowest mean of non-default.



1. It is advisable to issue loans to clients above income of 2,75,000 for the purpose of Gasification/Water Supply
2. Clients with income greater than 3,75,000 and purpose of loan for "wedding/Gift/Holiday" has higher average of non-default.
3. Similarly, it is not advisable to issue loan for purpose of "Building a house or an annex", "Buying a used car" & "Purchase of electronic equipment" for the clients belonging to income group more than 3.75L

FINDINGS & CONCLUSION

1. After cleaning the dataset, imbalance ratio of 11% could be seen in the dataset.
2. Of the total applications in the given data, 90% accounts for cash loan and remaining are revolving loans
3. 1/3rd of the clients own a car and 2/3rd does not own a car.
4. Majority of the clients are confined within the income range of 75,000 to 2,25,000.
5. Positive correlation of 0.39 could be seen between total income and credit amount
6. Most of credit loan amount range lies between 2,00,000 to 3,00,000.
7. Among defaulters, 60+ age group clients accounts a low percentage of 5.86% and 30-40 age group accounts 31.04%.
8. Generally, clients of widowed family status has highest average of non-default.
9. Similarly, it is advisable to issue loans to clients having office apartment and clients holding a academic degree of education type as they have less payment difficulties.
10. Issuance of loan to 60+ age group could be increased as it accounts for only 9.67% however clients of 60+ age group has highest average of non-default.
11. Clients who got credit in previous application and utilized it for buying a garage has the highest average income.
12. Clients with loan purpose of "Repairs" has highest count of non-default among others.

FINDINGS & CONCLUSION

Based on Education type and Family status:

> Clients with academic degree have the highest average of non-default count irrespective of different family status.

> Clients with combination of Lower secondary education and civil marriage or separated or single/unmarried has lowest mean of non-default.

Based on Family Status and Income Type:

> Clients with combination of Businessman and Married family status have good level of non-default

> Clients who are unemployed & married should not be issued loan.

Based on Income Range Group and Age Group:

>Generally, clients belonging to age group of 60+ has higher average of non-default. Specifically, clients above 60 and income group of 2,75,000 to 3,25,000 are marginally greater than others.

> Similarly, clients belonging to age group less than 30 years are relatively poor performing among than other age group. Further, age group of <30 and income of less than 2,25,000 should not be preferred.

Based on Credit Range and Income Range Group:

> Clients belonging to income group between 25,000 to 75,000 and credit loan amount range of 15,00,000 to 25,00,000 has lowest average of non-default.

FINDINGS & CONCLUSION

Based on Loan Purpose and Age Group

> Clients belonging to age group of 40 - 50 and loan purpose as per the previous application of "Buying a garage" & "Money for third person" has highest average of non-default.

> Similarly, clients belonging to age group of 60+ with loan purpose of "Gasification/Water Supply", "Hobby", "Buying a home", "Buying a holiday home/land" has highest mean of non-default.

> Clients belonging to age group of 30 - 40 with loan purpose of "Money to third person" & "Refusal to name the goal" has the lowest mean of non-default.

Based on Loan Purpose and Income Range Group

> It is advisable to issue loans to clients above income of 2,75,000 for the purpose of Gasification/Water Supply

> Clients with income greater than 3,75,000 and purpose of loan for "wedding/Gift/Holiday" has higher average of non-default.

> Similarly, it is not advisable to issue loan for purpose of "Building a house or an annex", "Buying a used car" & "Purchase of electronic equipment" for the clients belonging to income group more than 3.75L

Based on Loan Purpose and Family Status

> It is advisable to issue loan for clients of "Widow" family status for the purpose of "Hobby", "Gasification/Water Supply", "Business Development", "Buying a garage"

> Issuing loans to clients with civil marriage for purpose of "Hobby" & "Money to third person" is not advisable.