EDA Case Study

Risk Analytics in Banking & Financial services of losing money while lending to customers

Shivam Jha

31 May 2022

Introduction

The aim of this assignment is to apply EDA in a real business scenario. In this assignment, apart from applying the techniques, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. **Approved:** The Company has approved loan Application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).

4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Data Understanding

This dataset has 3 files as explained below:
https://drive.google.com/drive/folders/1Rc87Z0t4f9VTvyUn-lPatprWHQ3664_1?usp=sharing

1. '*application_data.csv*' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. '*previous_application.csv*' contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. '*columns_description.csv*' is data dictionary which describes the meaning of the variables.

Analysis on

Current DataSet - application_data.csv

Problem statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers.

Analysis Approach

To use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected and to understand how consumer attributes and loan attributes influence the tendency of default.

1. Data Collection - Process of loading data into our system

Current DataSet - application_data.csv

```
In [3]: # reading application_data dataset which is in CSV format and get the number of rows and column in dataset
        current_app_data = pd.read_csv("/Users/shivamjha/DataScience/EDA/EDACaseStudy/application_data.csv")
        current_app_data.shape

Out[3]: (307511, 122)
```

```
In [5]: # Use info() method of pandas to get rows and columns,the data type of each column, and the number of non-NaN elements
        # This will give us concise summary of the dataframe application_data
        current_app_data.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 307511 entries, 0 to 307510
        Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
        dtypes: float64(65), int64(41), object(16)
        memory usage: 286.2+ MB
```

2. Data Cleaning & Imputation of the missing Data in columns

- Removing missing values, outliers, and unnecessary rows/ columns.
- Re-indexing and reformatting our data.

3. Analyse the Data Imbalance

- Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations.

4. Univariate Analysis

- Analyse data of just one variable. A variable in dataset refers to a single feature/ column.
- We can do this either with graphical or non-graphical means by finding specific mathematical values in the data.
- Some visual methods include are Histogram and Box-plots

5. Bivariate Analysis

- Here we Analyse two variables and compare them.
- This way we can find how one feature affects the others
- It is done with scatter plots which plot individual data points or correlation matrices that plot correlation in hues.
- We can also use box plots

Cleaning & Imputation of the missing Data in columns

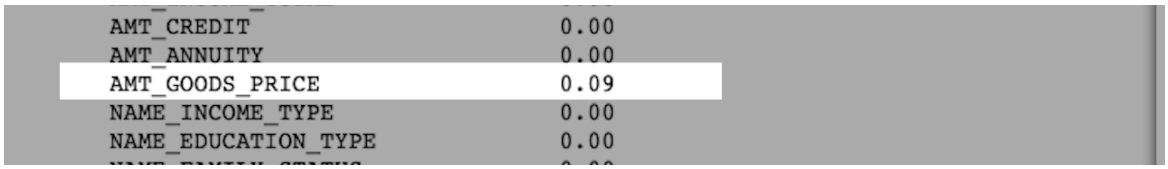
- Total number of rows and column in dataframe is (307511, 122)
- if a column contains more than 30% of its values missing, delete that column
- Original data frame had 122 columns but after deleting the columns which had more than 30% missing values the count of columns reduced to 72
- After removing the missing values now we have 307004 we lost 507 row in the process. Now dataframe with no missing values.
- At the beginning we had 307511 rows

```
In [19]: #Count non-NA cells for each column
         current_app_data.count()

Out[19]: SK_ID_CURR                  307511
         TARGET                      307511
         NAME_CONTRACT_TYPE          307511
         CODE_GENDER                 307511
         FLAG_OWN_CAR                307511
                                      ...
         AMT_REQ_CREDIT_BUREAU_DAY   265992
         AMT_REQ_CREDIT_BUREAU_WEEK  265992
         AMT_REQ_CREDIT_BUREAU_MON   265992
         AMT_REQ_CREDIT_BUREAU_QRT   265992
         AMT_REQ_CREDIT_BUREAU_YEAR  265992
         Length: 72, dtype: int64
```

- Remove the columns which are irrelevant for the analysis

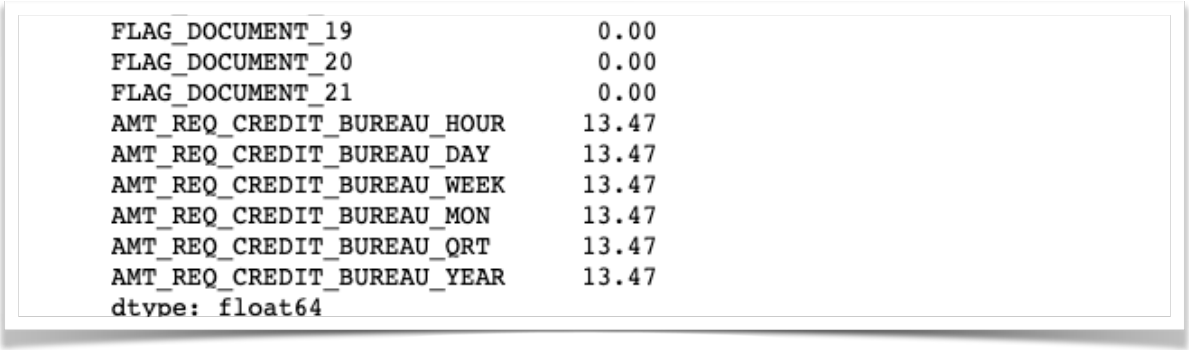
```
In [31]: # generate descriptive statistics on 'AMT_GOODS_PRICE' of dataframe
         current_app_data['AMT_GOODS_PRICE'].describe()

Out[31]: count    3.070040e+05
         mean     5.384842e+05
         std      3.694861e+05
         min      4.050000e+04
         25%      2.385000e+05
         50%      4.500000e+05
         75%      6.795000e+05
         max      4.050000e+06
         Name: AMT_GOODS_PRICE, dtype: float64

In [32]: # remove exponent notation in AMT_GOODS_PRICE
         current_app_data['AMT_GOODS_PRICE'].describe().apply(lambda x: '%.2f' % x)

Out[32]: count     307004.00
         mean      538484.20
         std       369486.08
         min        40500.00
         25%       238500.00
         50%       450000.00
         75%       679500.00
         max      4050000.00
         Name: AMT_GOODS_PRICE, dtype: object
```

- Remove exponent notation in AMT_GOODS_PRICE column
- Analyse the missing values in columns



```
AMT_CREDIT                  0.00
AMT_ANNUITY                 0.00
AMT_GOODS_PRICE             0.09
NAME_INCOME_TYPE            0.00
NAME_EDUCATION_TYPE         0.00
```

box plot from DataFrame AMT_GOODS_PRICE columns



```
FLAG_DOCUMENT_19              0.00
FLAG_DOCUMENT_20              0.00
FLAG_DOCUMENT_21              0.00
AMT_REQ_CREDIT_BUREAU_HOUR   13.47
AMT_REQ_CREDIT_BUREAU_DAY    13.47
AMT_REQ_CREDIT_BUREAU_WEEK   13.47
AMT_REQ_CREDIT_BUREAU_MON    13.47
AMT_REQ_CREDIT_BUREAU_QRT    13.47
AMT_REQ_CREDIT_BUREAU_YEAR   13.47
dtype: float64
```

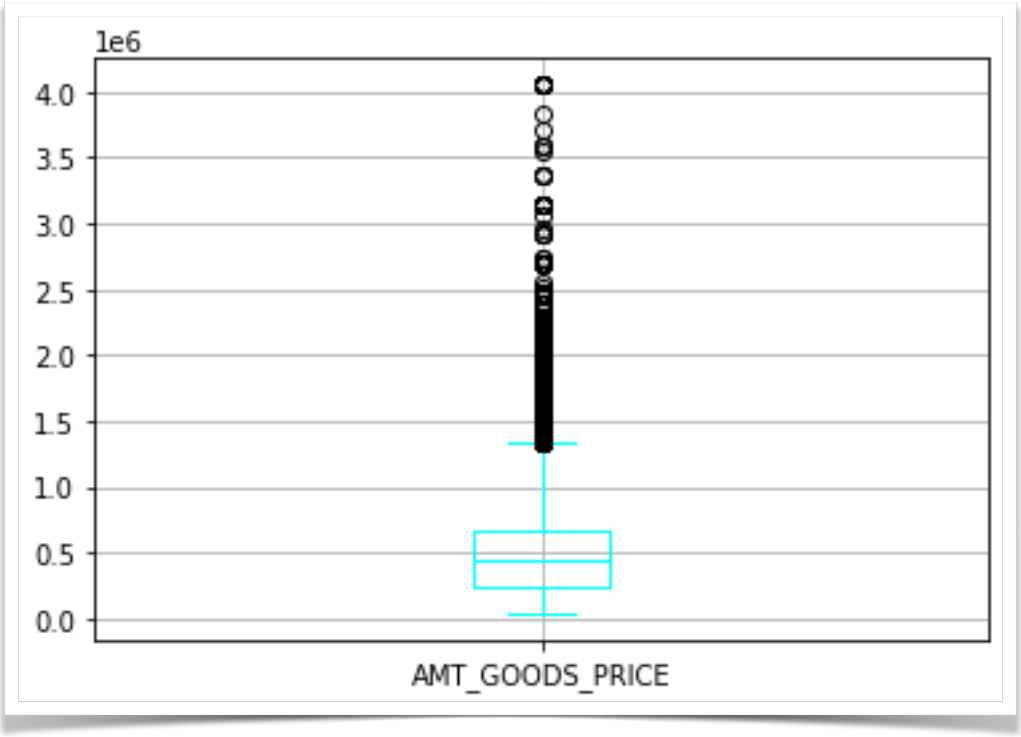
Observations and recommendations :

- There are 5 columns for the number of enquiries to Credit Bureau about the client.
- It doesn't make much sense to count the number for an hour or a day or a week.
- We will delete the column for the count of hour, day, week, month and year.

Conversion of columns for better readability and analysis

- Convert DAYS_BIRTH column into AGE column - divide DAYS_BIRTH by 365 for taking Age
- Convert DAYS_EMPLOYED column into YEARS_EMPLOYED column for readability and analysis - divide DAYS_EMPLOYED by 365 for YEAR_EMPLOYED
- Drop DAYS_BIRTH and DAYS_EMPLOYED

Identify and remove outliers



- The standard deviation of column is very high
- Dataset in columns AMT_INCOME_TOTAL,AMT_CREDIT,AMT_ANNUITY,AMT_GOODS_PRICE,AGE contain extreme values that are outside the range
- imputing missing values with mean or median will not be correct solutions
- these rows are only 0.09% of the whole record it is good to remove the rows with missing AMT_GOODS_PRICE.
- Used IQR (Interquartile Range Method) to find the outliers in all the numerical columns with 1.5 IQR rule () and removing the outlier records
- we lost around 10% rows in the outliers handling exercise which is necessary to get rid of the outliers for fair analysis of the data.

Analyse the Data Imbalance

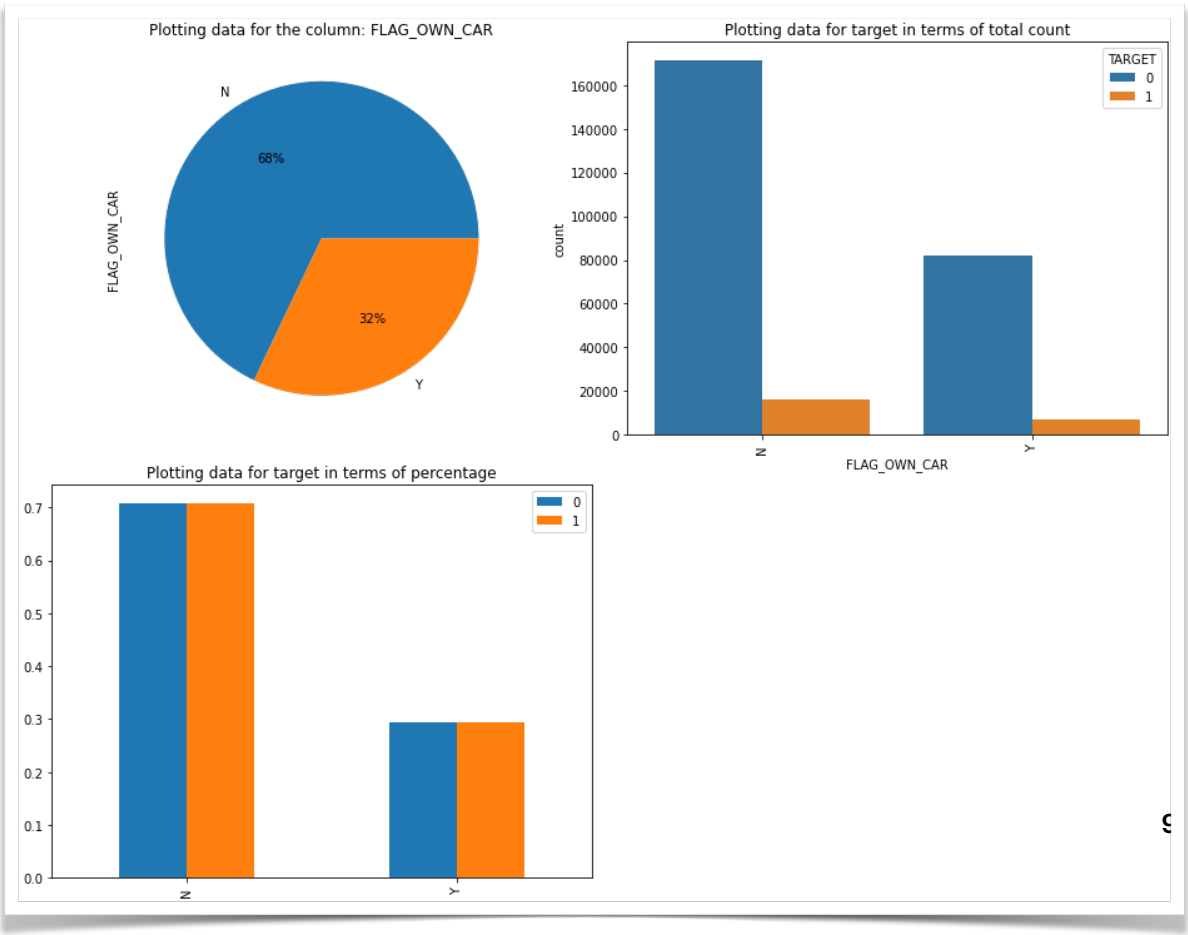
Approach to deal with the imbalanced dataset problem

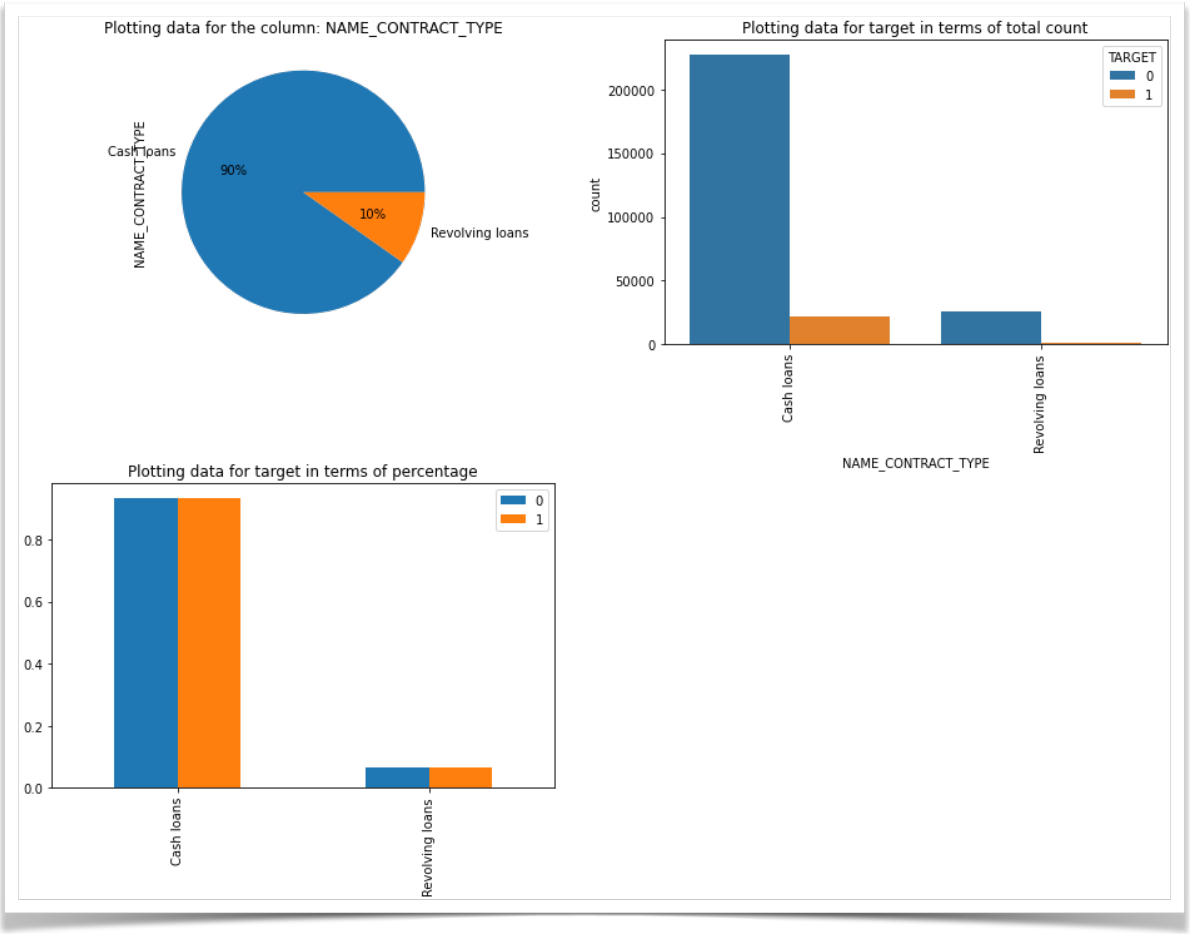
- it is good to identify the minority classes correctly.
- So model should not be biased to detect only the majority class but should give equal weight or importance towards the minority class too.
- Created two data frame Client with payment difficulties and all other cases on Target Variable
- Ratio of data imbalance is

```
0    0.916255
1    0.083745
```

```
In [78]: #relative frequencies of the unique values pandas
         current_app_data.TARGET.value_counts(normalize=True)

Out[78]: 0    0.916255
         1    0.083745
         Name: TARGET, dtype: float64
```

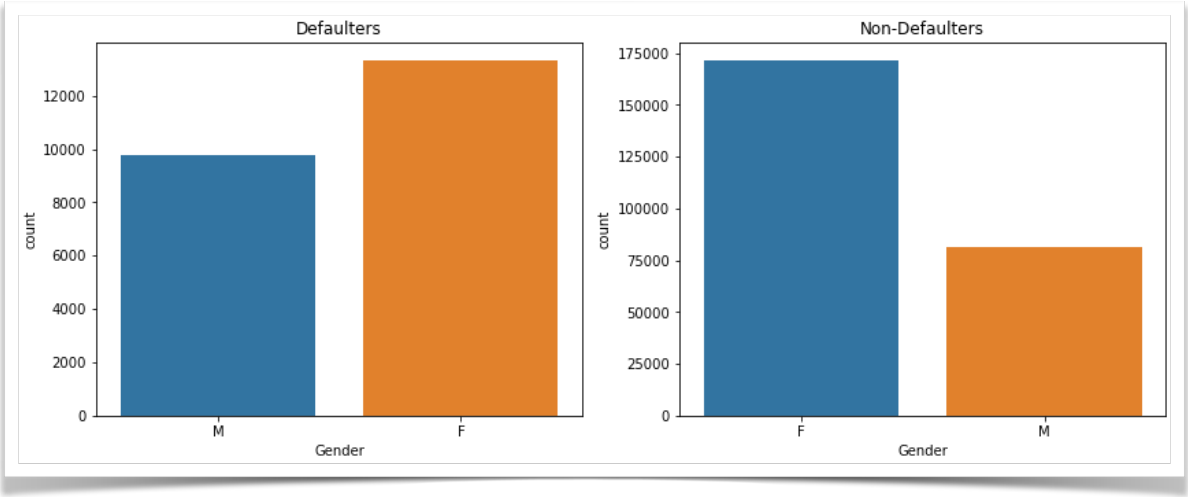
- Plots to identify the data imbalance




Plotting data for the column: NAME_CONTRACT_TYPE
NAME_CONTRACT_TYPE
Cash loans
90%
10%
Revolving loans
Plotting data for target in terms of total count
TARGET
0
1
count
200000
150000
100000
50000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
Plotting data for target in terms of percentage
0
1
0.8
0.6
0.4
0.2
0.0
Cash loans
Revolving loans

Univariate analysis

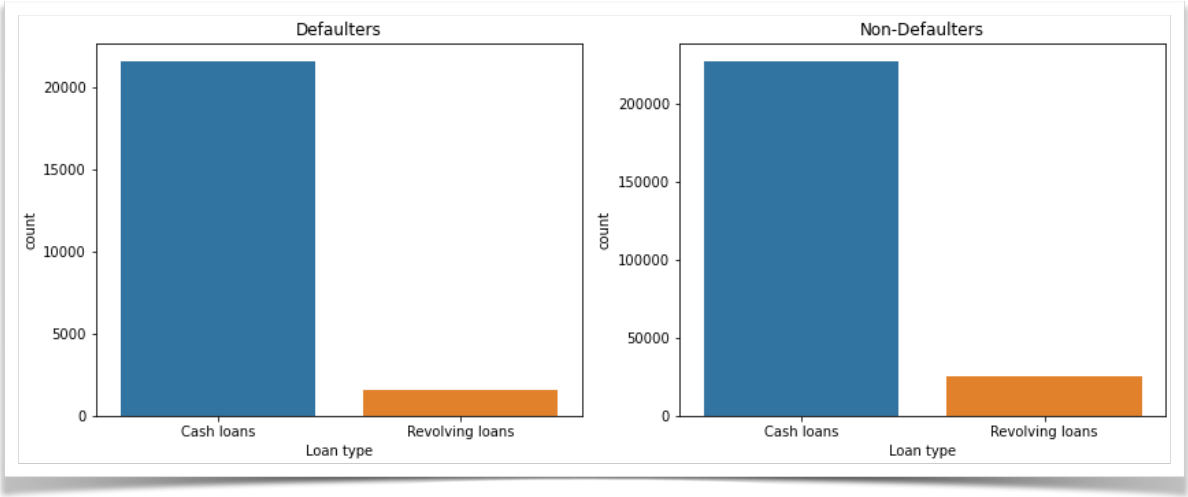
1. Count of defaulters and non-defaulters on the basis of gender



Observations :

- **Defaulters** - We can see that females are slightly more in number of defaulters than male.
- **Non-defaulters** - The same pattern continues for non-defaulters as well. The females are more in number here than male.

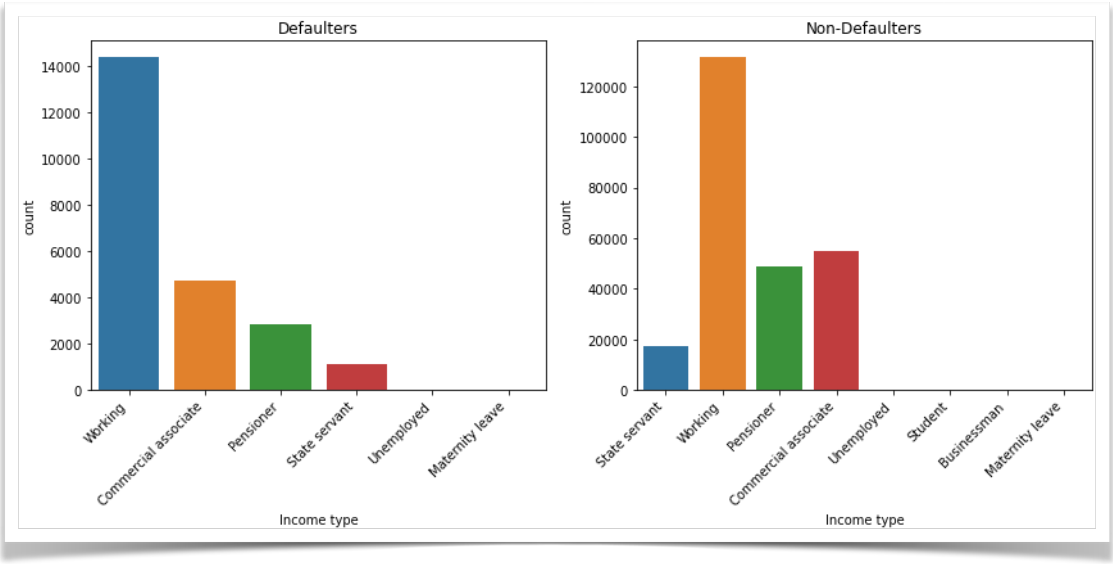
2. Defaulters and non-defaulters on the basis of Loan type



Observations :

- We see in both the cases that Revolving loans are very less in number compared to Cash loans.

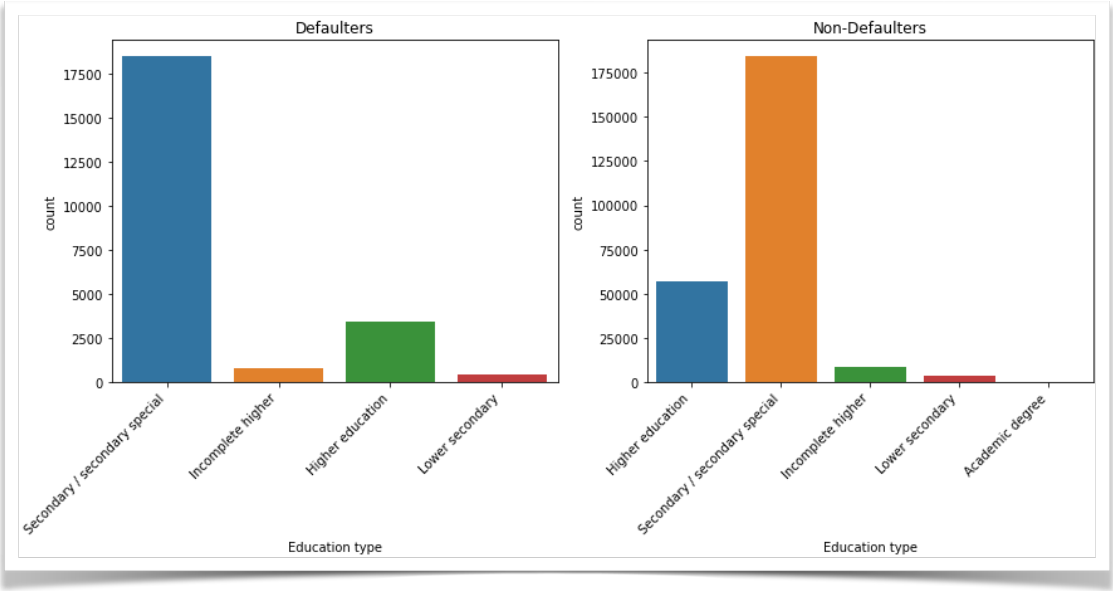
3. Defaulters and non-defaulters on the basis of Income type



Observations :

- **Defaulters** - Working people are mostly defaulted as their numbers are high with compare to other professions.
- **Non-defaulters** - Similarly here also working people are more in number who are not defaulted.

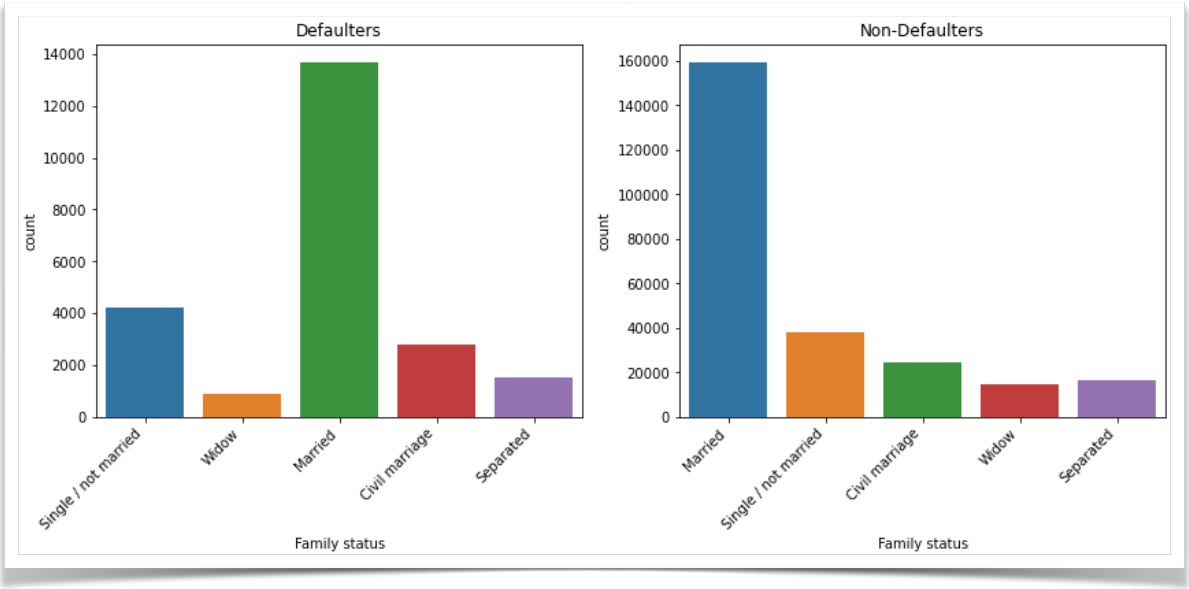
4. Defaulters and non-defaulters on the basis of Education type



Observations :

- **Defaulters** - Education with Secondary/Secondary sepcial customers are more number in defaulters comapre with other level of eduacted poeple.
- **Non defaulters** - Here also Secondary/Secondary sepcial are more in numbers.

5. Defaulters and non-defaulters on the basis of Family status

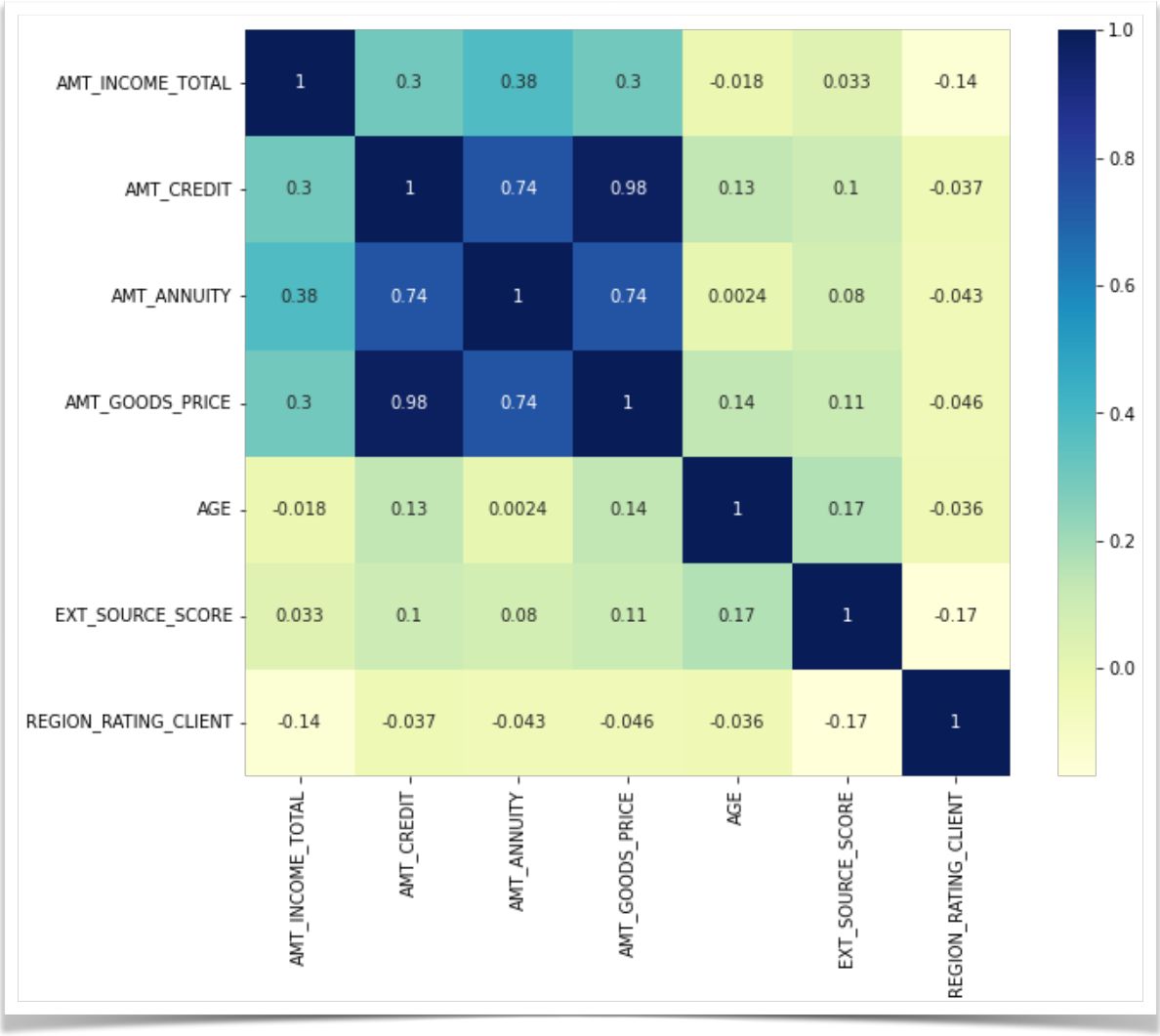


Observations :

- For both the customers (defaulters and non-defaulters) married people are more in number comapred with single, separated, widow etc.

Bivariate Analysis

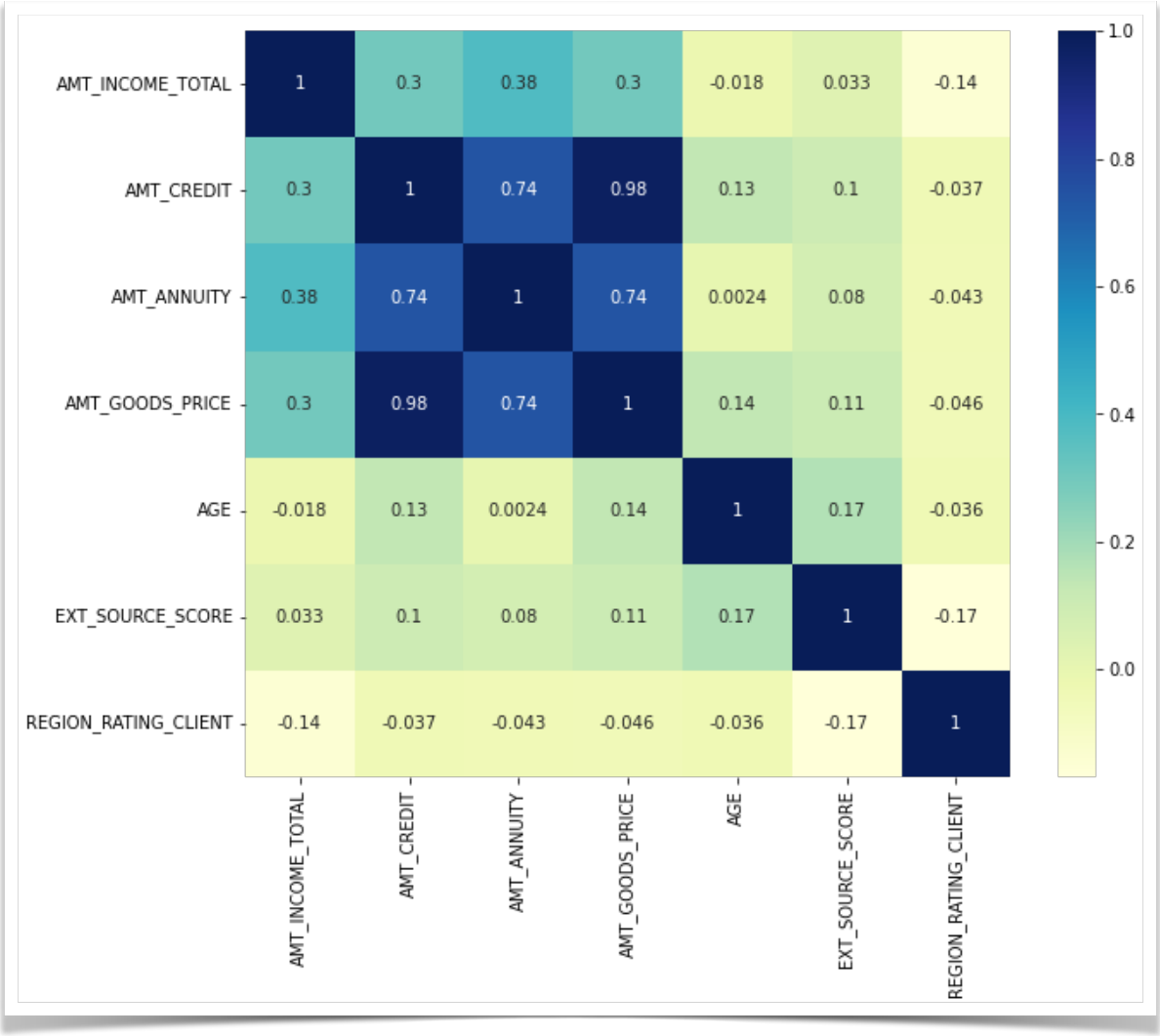
1. Correlation of relevant numerical columns for defaulters and non defaulters



Observations :

- AMT_CREDIT and AMT_ANNUITY (0.74)
- AMT_CREDIT and AMT_GOODS_PRICE (0.98)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.74)

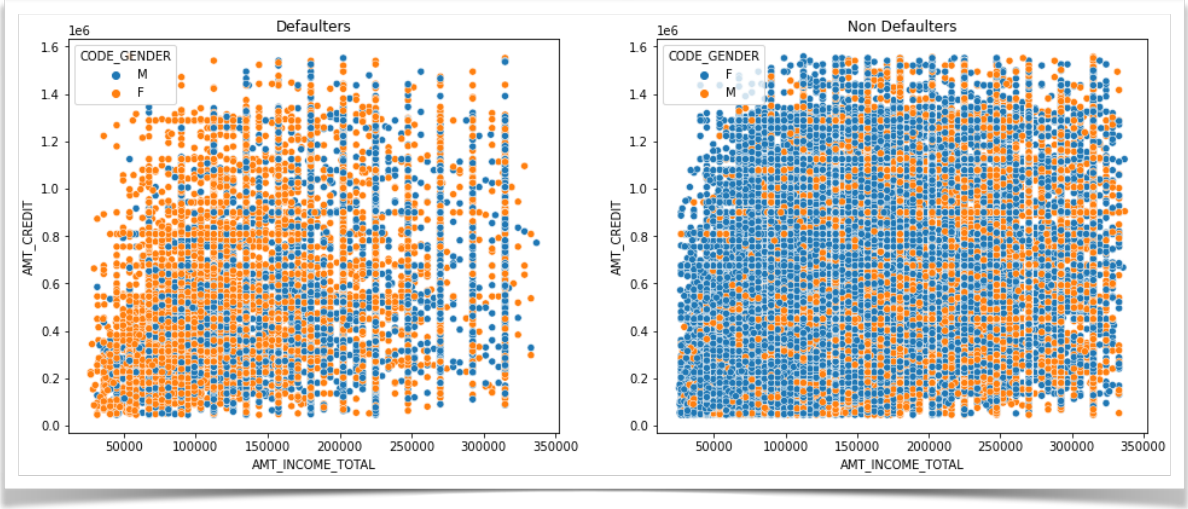
2. Correlation of relevant numerical columns for non defaulters



Observations :

- AMT_CREDIT and AMT_ANNUITY (0.76)
- AMT_CREDIT and AMT_GOODS_PRICE (0.98)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.76)

3. Credit amount of the loan on the basis of client income for both male and female¶

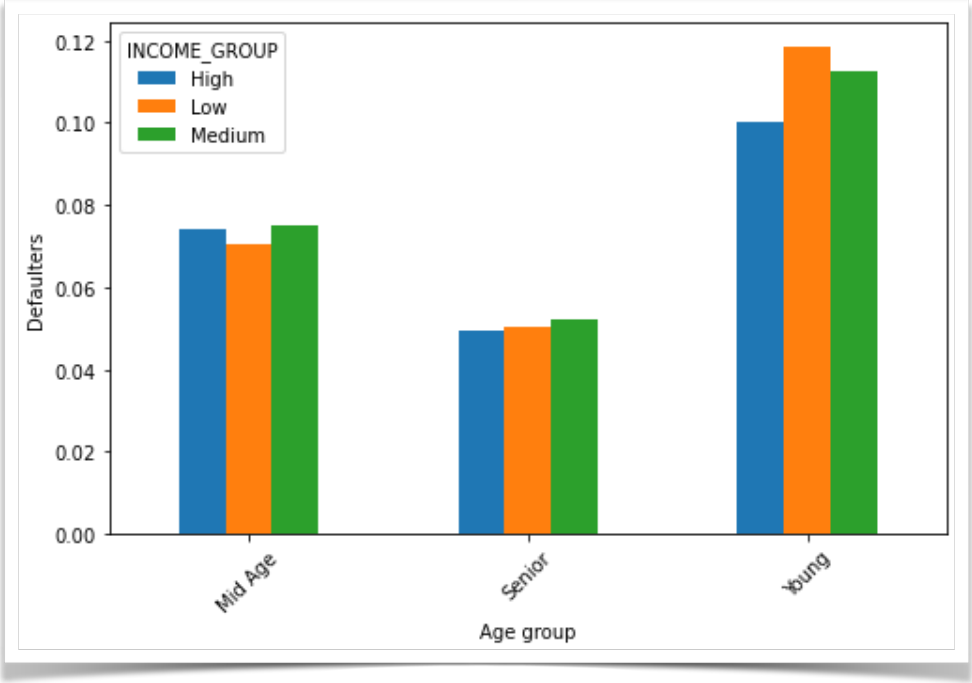


Observations :

- **Defaulters** - We can slightly figure out that the values are more concentrated on the lower income and lower credit of the loan. That means as the income is increased, the amount of loan is also increased. This is true for both the genders.
- **Non defaulters** - We can hardly figure out any pattern out of this.

Anslysis of two segmented variables

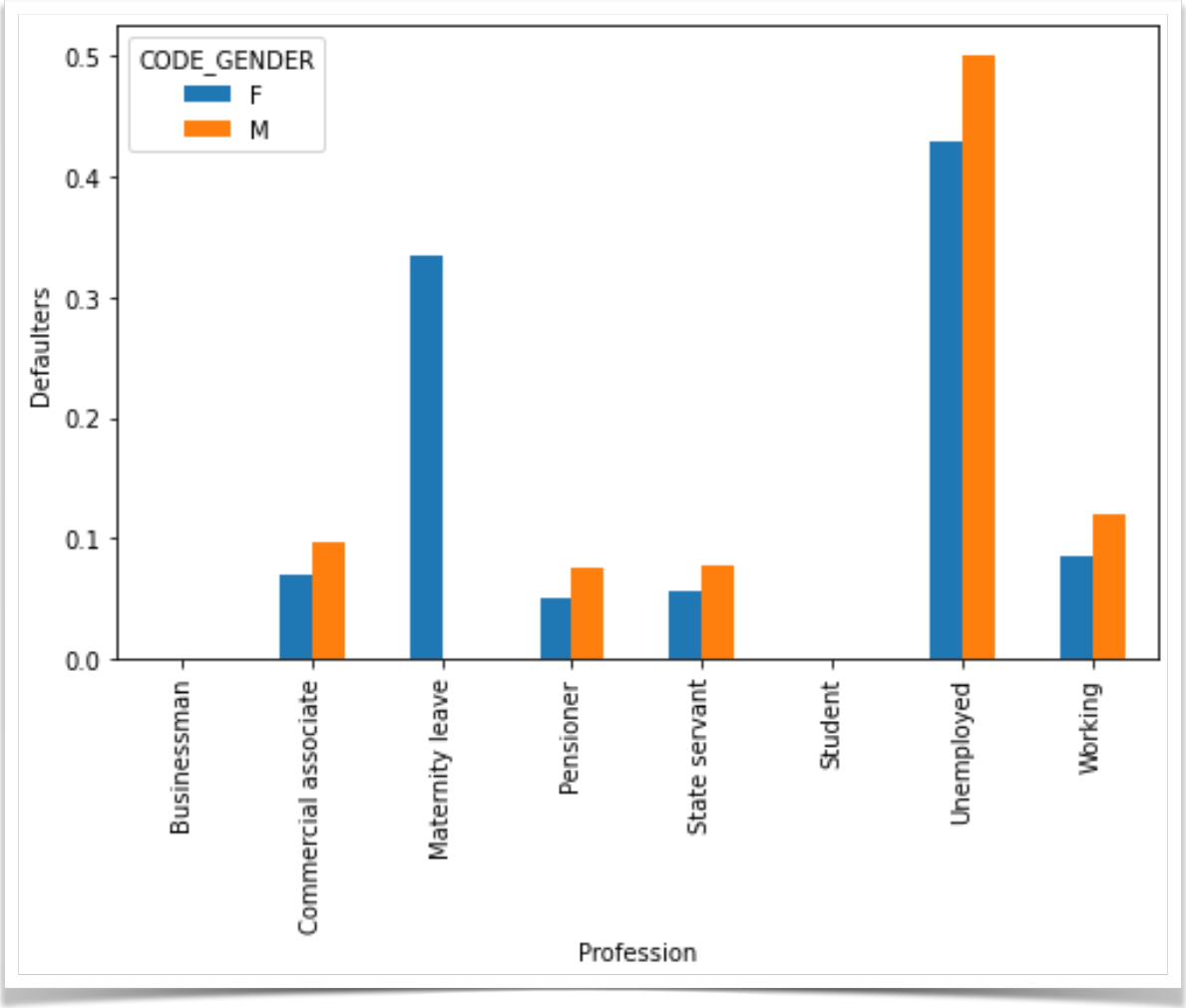
4. Age group and Income group



Observations :

- Young clients are more defaulted than Mid age and senior.
- Young low income people are more defaulted.
- For Mid age and senior people the default rate is almost same in all income group.

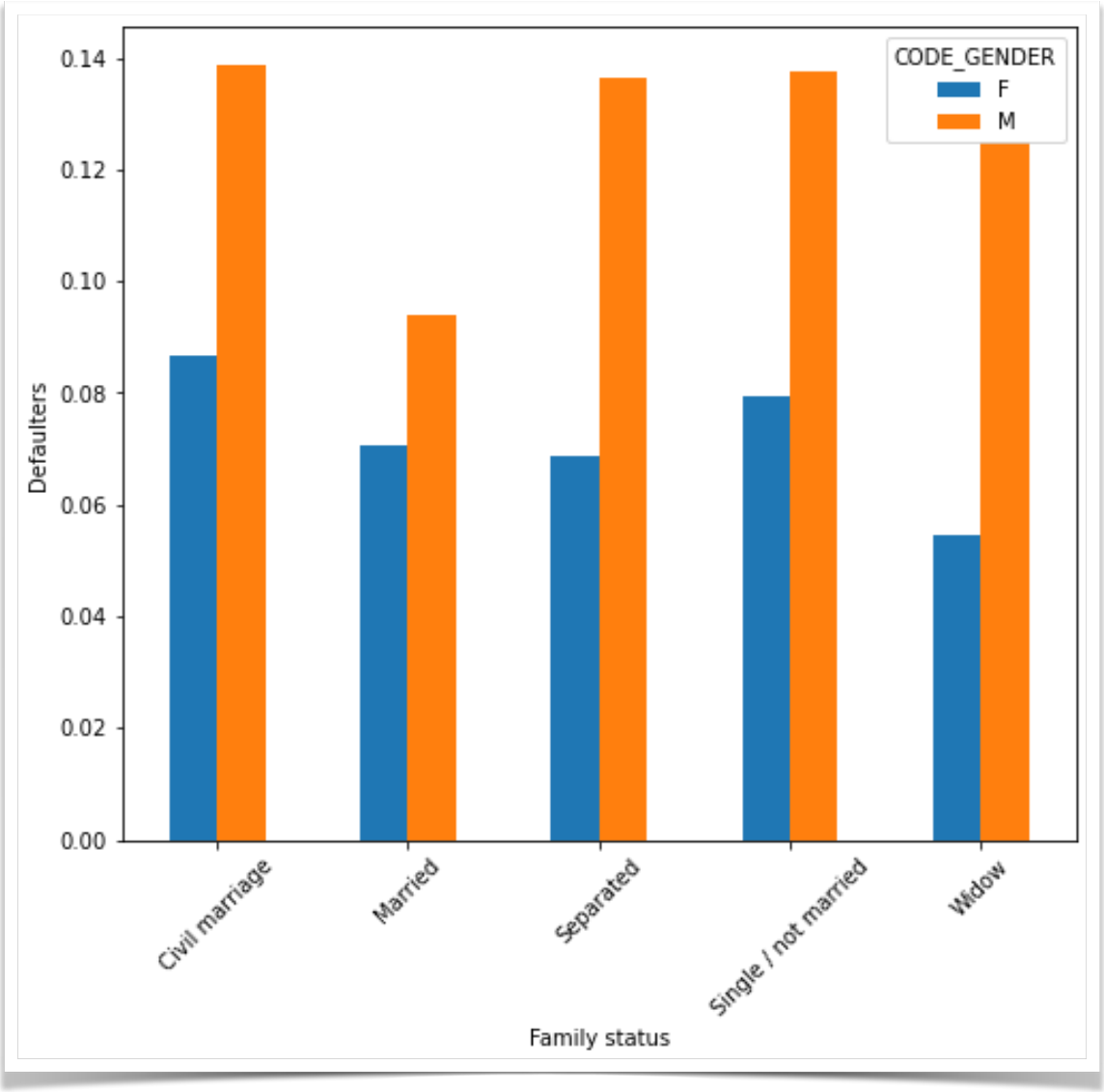
5. Profession and Gender



Observations :

- the unemployed clients are more defaulted.
- Clients with maternity leave are expected to be defaulted more.
- The default rate is lesser in all other professions.
- Males are more defaulted with their respective professions compared to females

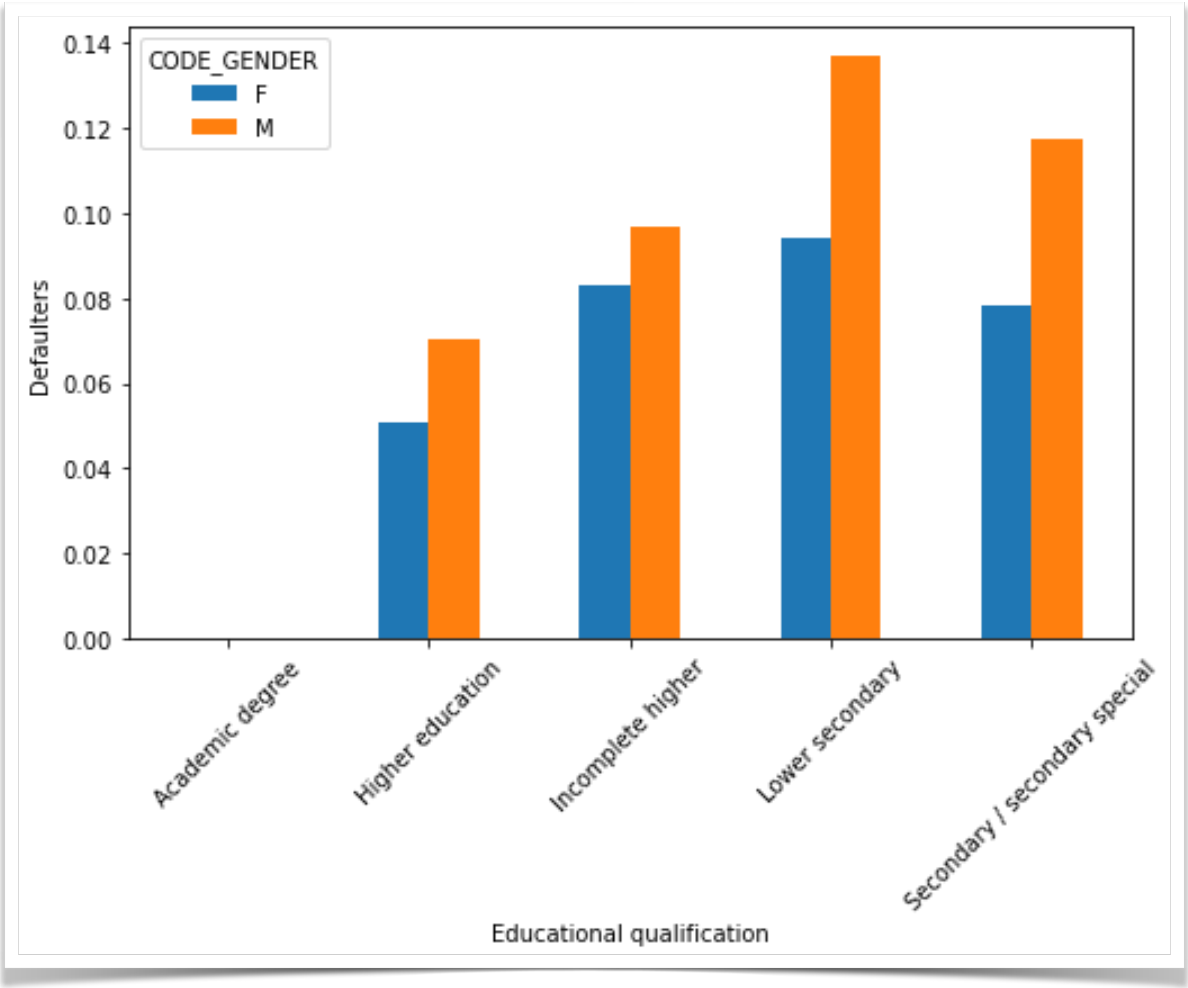
6. Family status & age group



Observations :

- Across all family status the `Male` clients are more defaulted than `Female`.

7. Education and gender

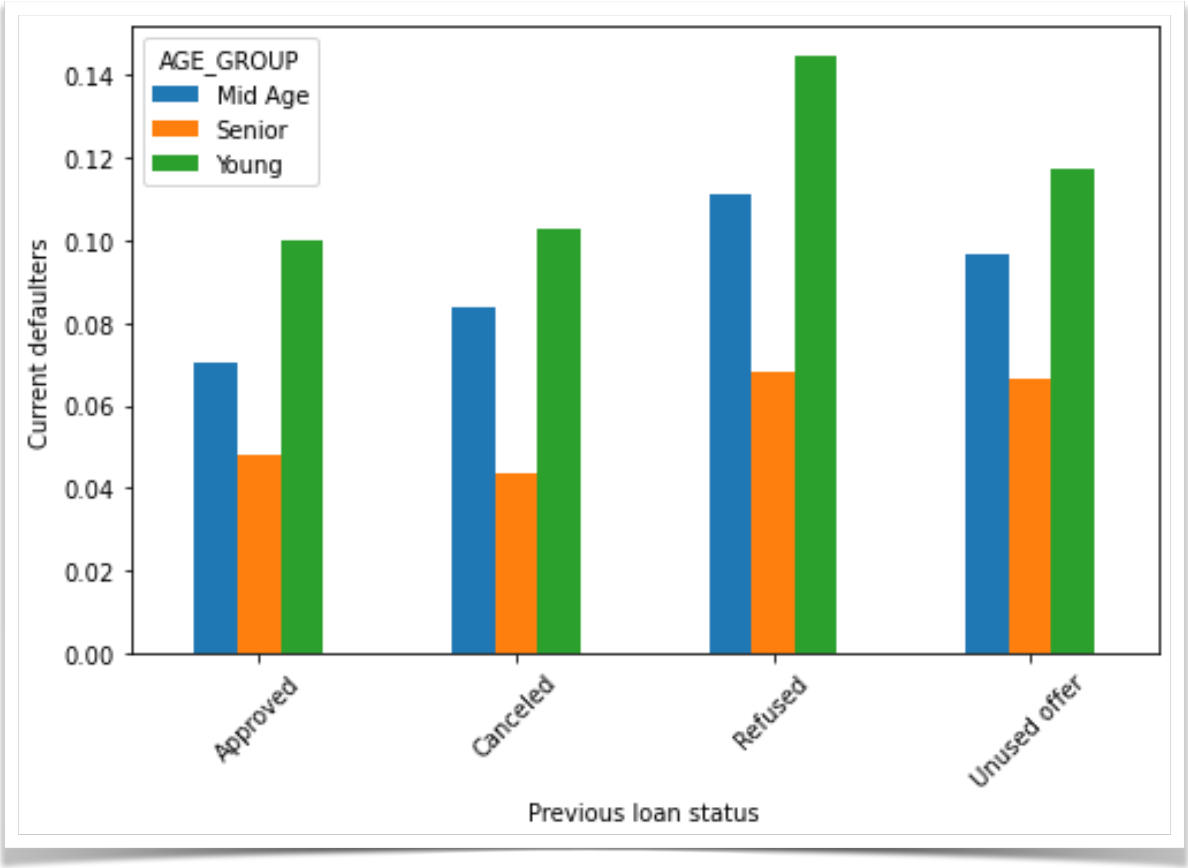


Observations :

- Lower secondary educated clients are more defaulted followed by Secondary and Incomplete higher educated clients.
- The Higher educated people are less defaulted.
- Accross all educated level Females are less defaulted than male.

Loan application status relations on Current and Previous data

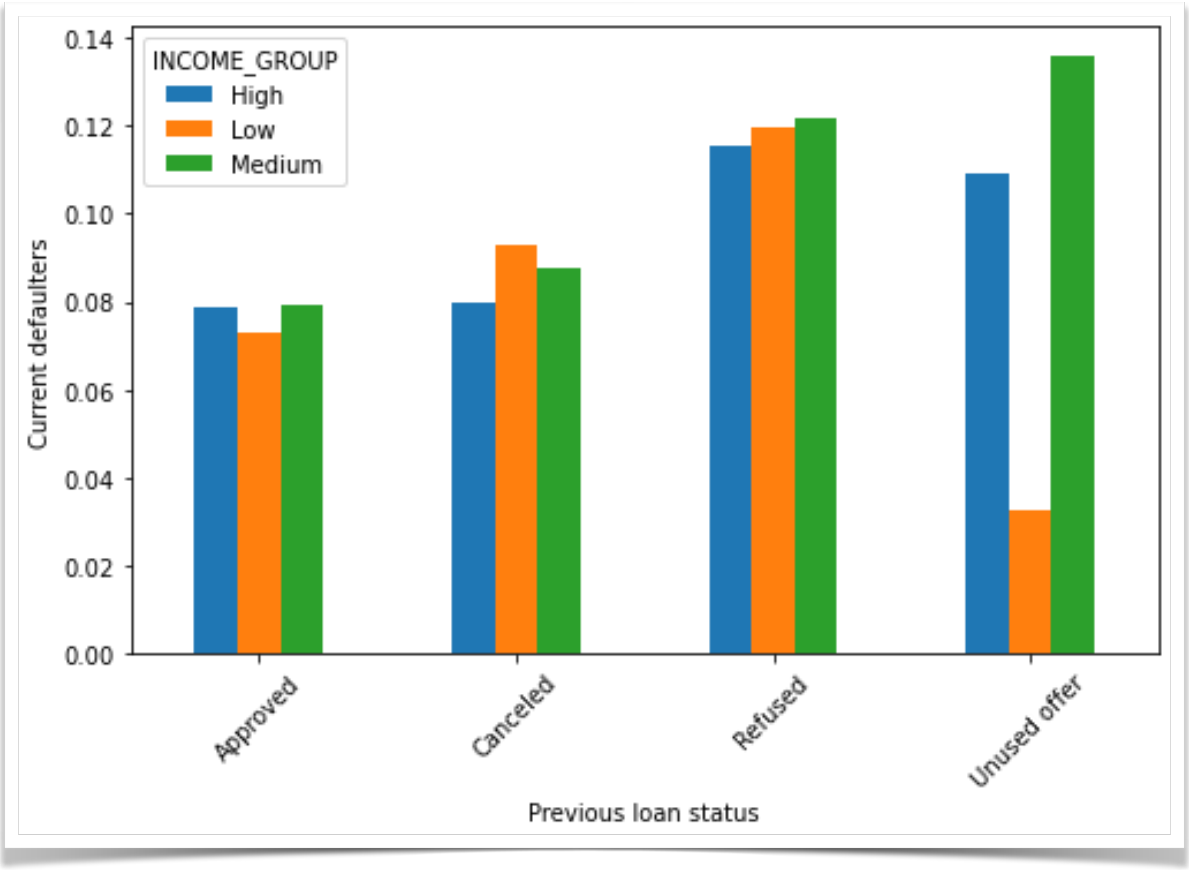
8. Current loan defaulter status with respect to previous loan application status and age group



Observations :

- For all the previous status young pplicants are more defaulted.
- For all the previous status Senior applicants are less defaulted compared to others
- In all income groups previously refused applicants are more defaulted.
- Safer to grant loans for senior citizen.

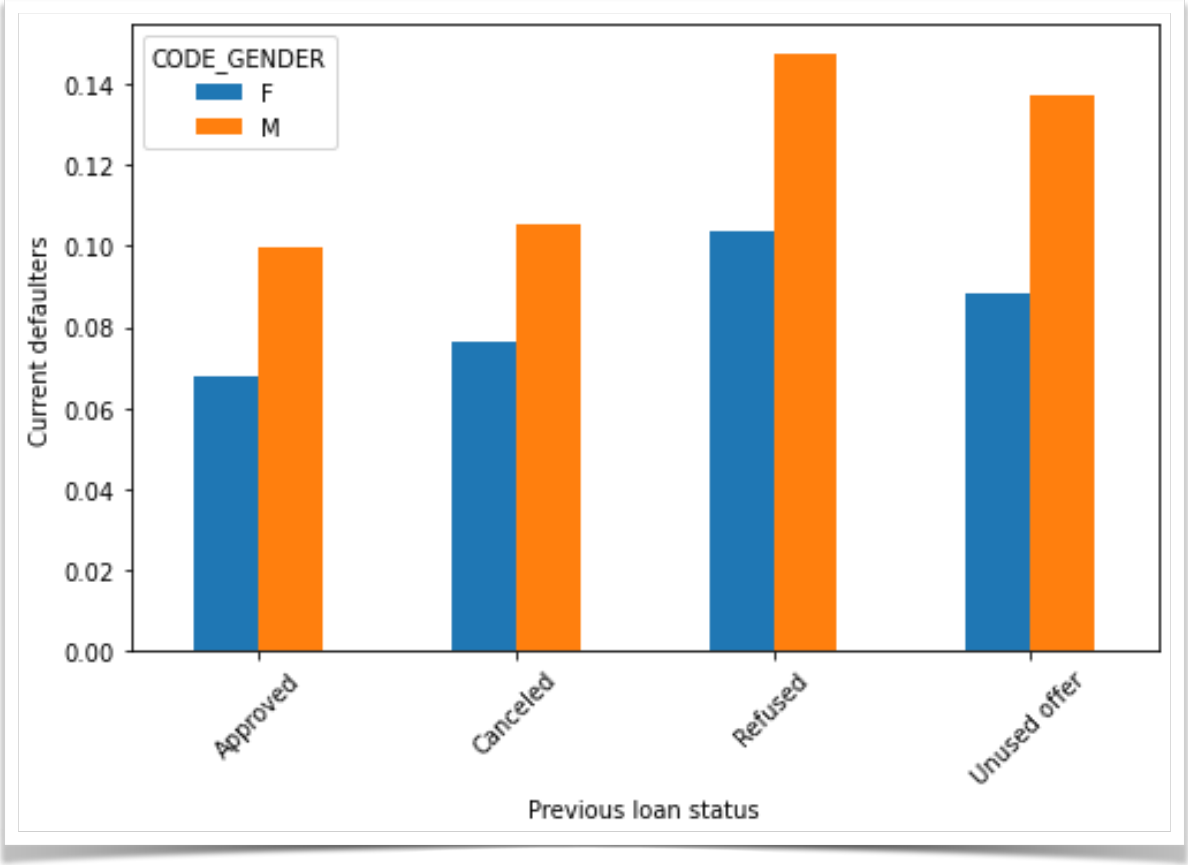
9. Current loan defaulter status with respect to previous loan application status and income group



Observations :

- For previously Unused offer the Medium income group was more defaulted and `Low` income group is the least.
- For other application status more or less all the income groups are equally defaulted
- Lesser risk to grant loans for approved applicants to all income groups.

10. Current loan defaulter status with respect to previous loan application status and income group



Observations :

- We see that previously Refused client is more defaulted than previously Approved clients. Also, in all the cases the Males are more defaulted than Females
- New clients with previously unused offer are more defaulted.
- There is a risk to grant loans for clients, whose applications were refused or unused previously.
- It is recommended to provide loans to previously approved females.

Conclusion

1. Financial institutions should provide loan more to Married customer compared to other family type.
2. Preference should be given to Senior citizen
3. Female are more reliable than Male
4. Customers with less educated should be avoided
5. Customers those are unemployed should be given loan
6. If the customer having applied the loan and is previously refused, cancelled or unused offer should be cautious
7. Loan should be provided to higher income with highly educated customer
8. Customer who are young should be given less loan if income is less because they are more riskier that other age group.