Credit EDA Case Study

Batch – CS23 Aug 2020

23/11/2020

Namitha Murugesh

and

Sunila R K

Problem Statement

- The main goal of this case study is to analyze and identify the best approach to sanction loans to clients.

- Main objective is to minimize the risks concerning
  - Not approving loan to non defaulters (Interest Loss)
  - Approving loan to defaulters (Credit Loss)

Data Available

- Data about clients with payment difficulties
- Data about previous applications

Analysis Approach

The following steps were carried out during the analysis

- Inspecting the dataset
- Understanding the data
- Inspecting the structure of data
- Quality check
  - Inspecting the percentage of missing values
  - Dropping columns with missing values more than 50%
  - Inspecting the columns with missing values less than 13%
  - Check the data types of all the columns and modify the data appropriately.

Analysis Approach - Quality Check Continued

- For numerical columns, check the outliers and suggest appropriate methods
- Binning of continuous variables
- Check the data imbalance based on the variable target

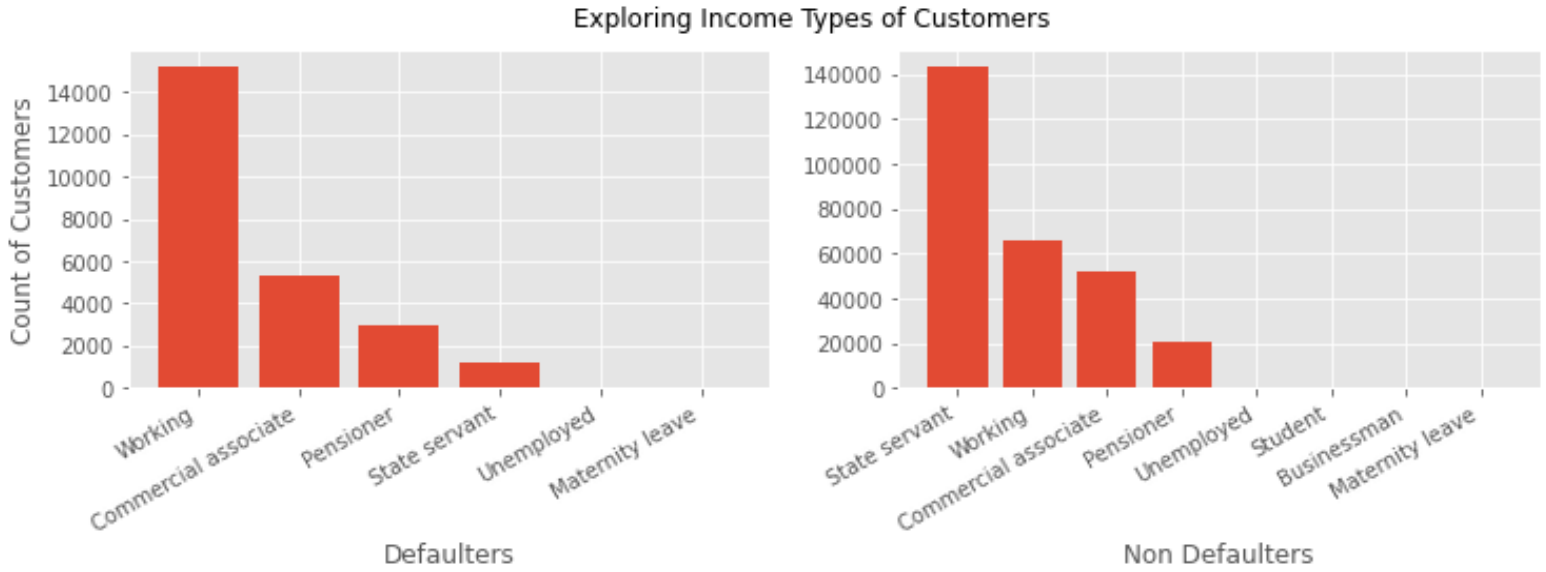
Analysis Approach

The dataset is split based on the target variable namely Target 0(Non Defaulters) and Target 1(Defaulters).

Analysis is carried out on each of these data frames by:

- Univariate
  - Categorical variables
  - Continuous variables
- Bivariate
  - Categorical Vs. Categorical
  - Categorical Vs. Continuous
  - Continuous Vs. Continuous

Univariate Categorical Analysis - Distribution of Income Type



Distribution of Income Type

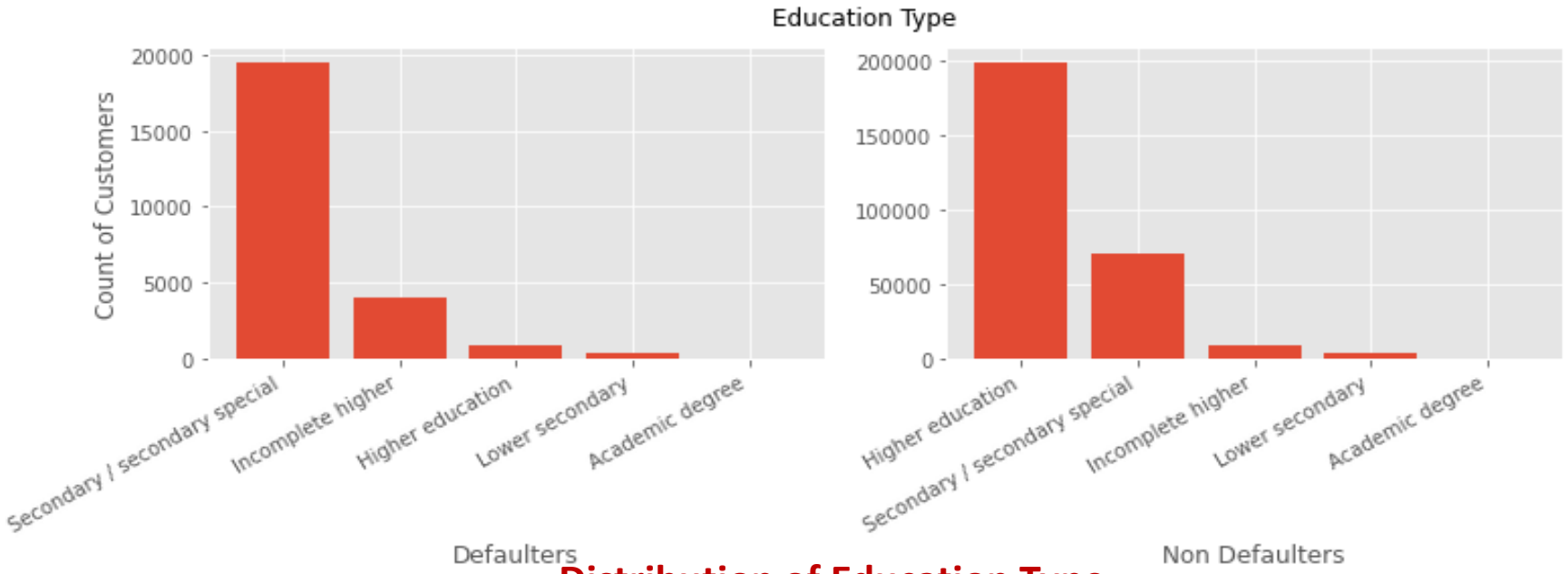
Observation

- Most of the customers who default are working class. Non defaulters belong to State servant category.

Points to conclude from the above graph:

- While sanctioning loans to the working class caution has to be exercised. Where as state servant category customers are seeming to be non defaulters. This can be associated to the job security and assured income to the state servant category.

Univariate Categorical Analysis



Distribution of Education Type

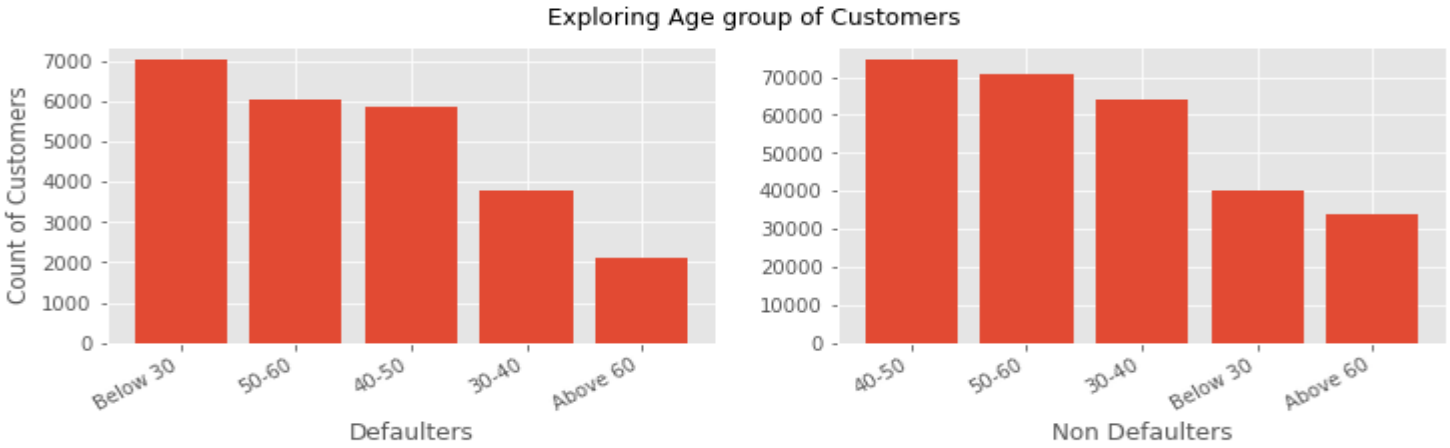
Observation:

- Maximum customers who default have education of secondary level. Non defaulters have a higher education. A data imbalance is seen in this column

Points to conclude:

- The customers with secondary education may not be placed with a very high income and secure jobs, hence explaining the defaulter behaviour.

Univariate Continuous Analysis



Exploring Age group of customers

Observation:

- Most of the customers who default are below 30 . Most of the non- defaulters are between 40 to 50 years of age.

Points to conclude from the above graph:

- Caution has to be exercised while granting loans for people below 30 years of age. The reason for defaulters having age less than 30 may be associated with an unstable job since they may be just out of college. Most of the non- defaulters are between 40 to 50 years of age relating to a well settled source of income.

Univariate Continuous Analysis

Exploring the recent change in identity document

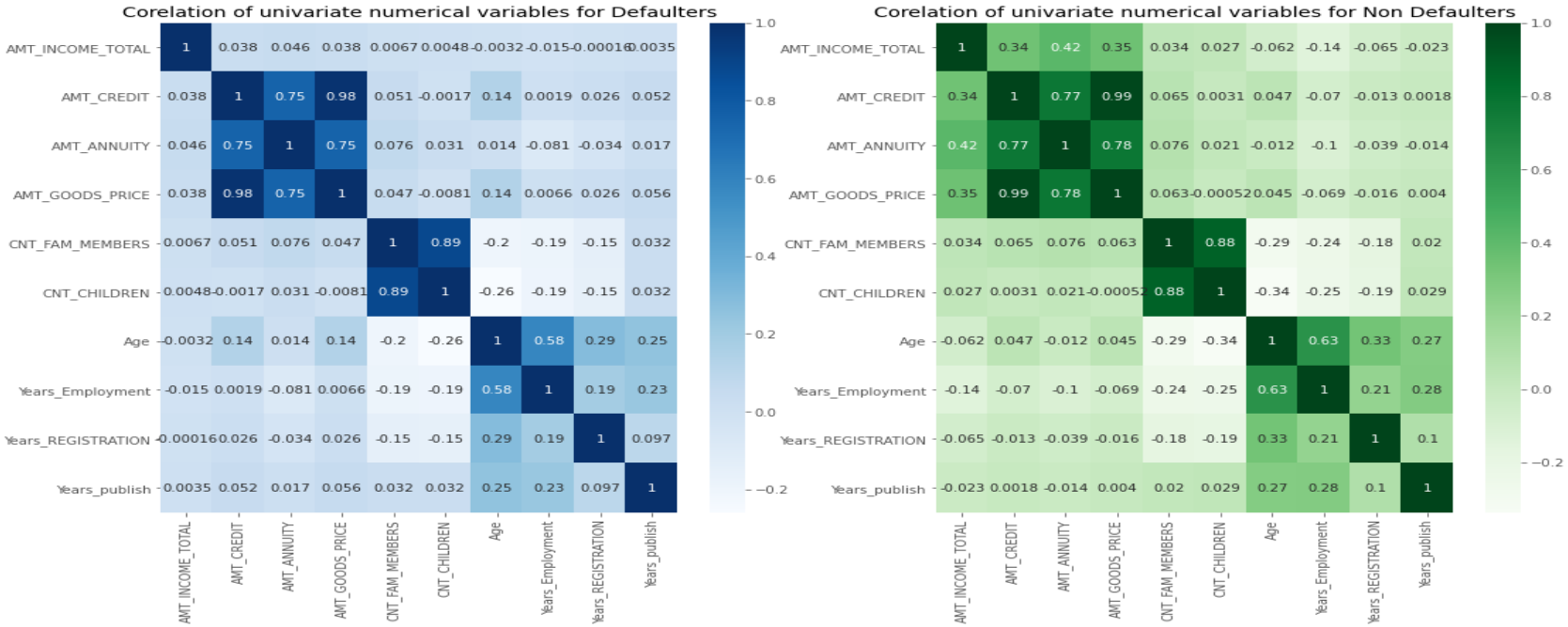
Observation:

- This grapgh shows that defaulters have changed their identity document around 7 years before the application whereas non-defaulters have changed their identity document around 9 years

Points to conclude from the above graph:

- This change of identity in recent past might raise a flag for defaulters. If there is a change in identity document of around 7 years or less then caution must be exercised before granting the loan to the individual.

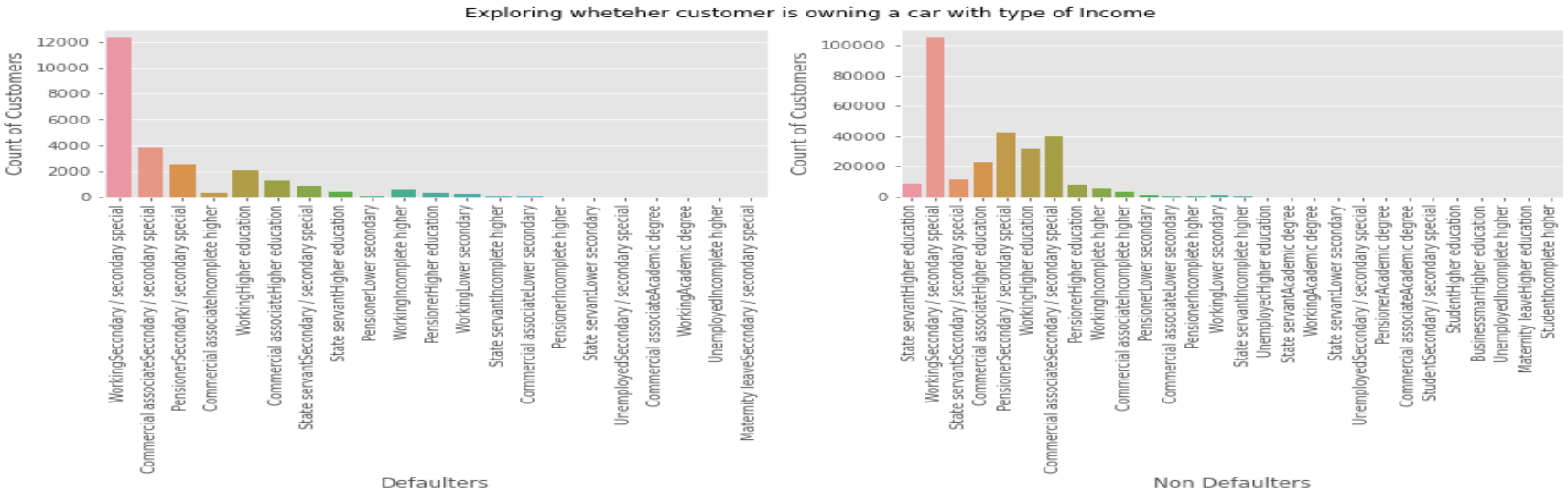
Correlation of Univariate numerical variables



Observation:

Both the data frames have similar correlation for the variables. The age has very low correlation with count of children and family members . Amount of income, credit, goods price and annuity are highly correlated hence will be looked into in detail in the following sections.

Bivariate Analysis - Categorical Vs. Categorical



Distribution of Income Type Vs. Education Type

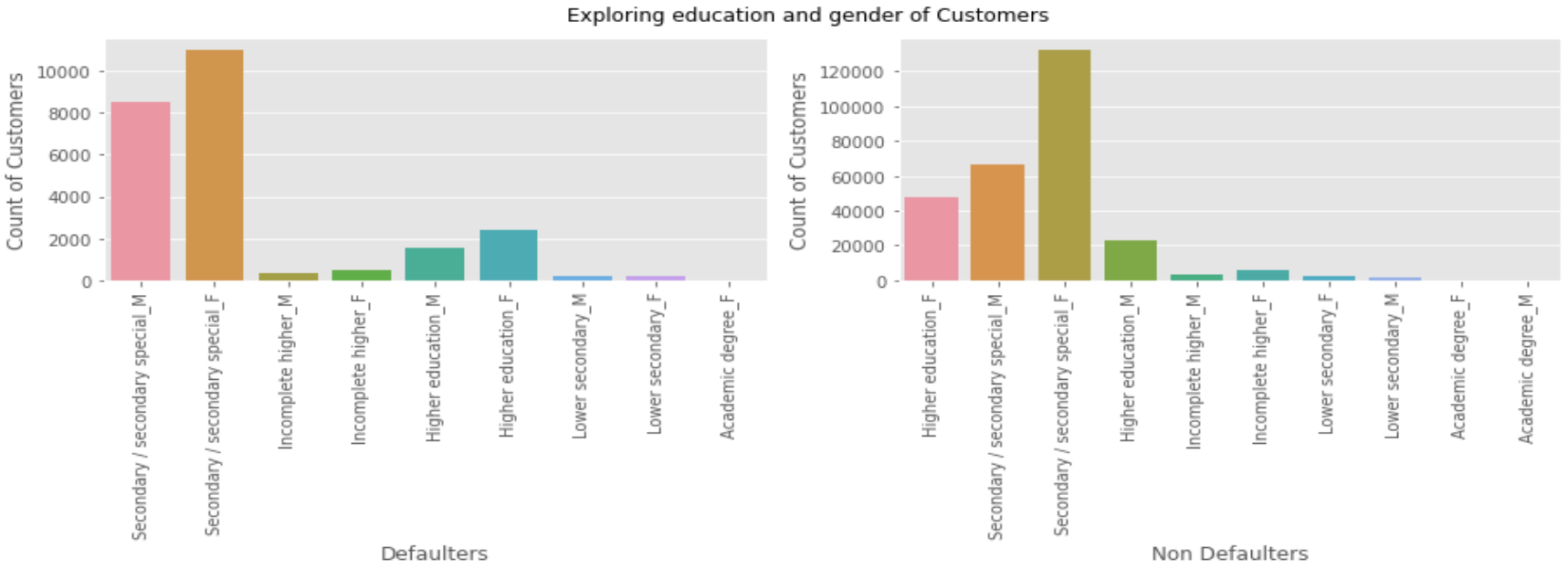
Observation:

- Working class with secondary education are the largest customer segment. They appear both in defaulters and non defaulters. Besides them Commercial associates with secondary education are among the next highest defaulters.

Points to conclude:

- If the customer is from working class or commercial associates and has secondary education, the loan has to be sanction with caution. This can be explained as people with lower education dont get very stable jobs.

Bivariate Analysis-Categorical Vs. Categorical



Distribution of Education Type vs Gender

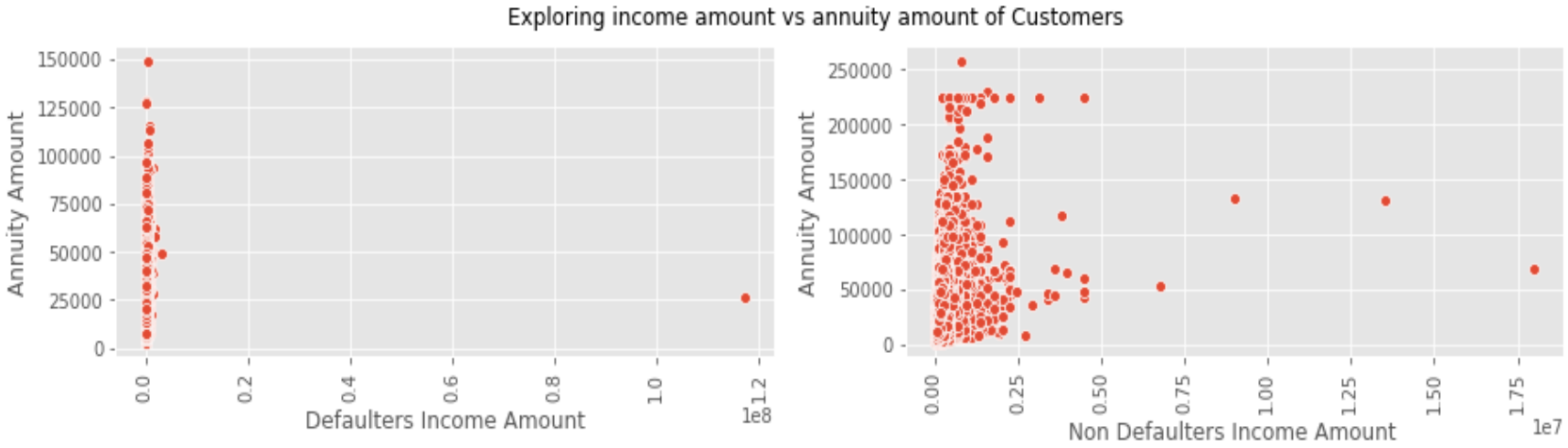
Observation:

Customers who are male and have a secondary education are high in defaulters list. Female with secondary education are present in high number in both defaulters and non defaulters.

Points to conclude:

Loans to customers with secondary education irrespective of the gender have to be sanctioned with caution.This can again be on account of unstable or low paying jobs to lower education levels.

Bivariate analysis -Continuous Vs. Continuous

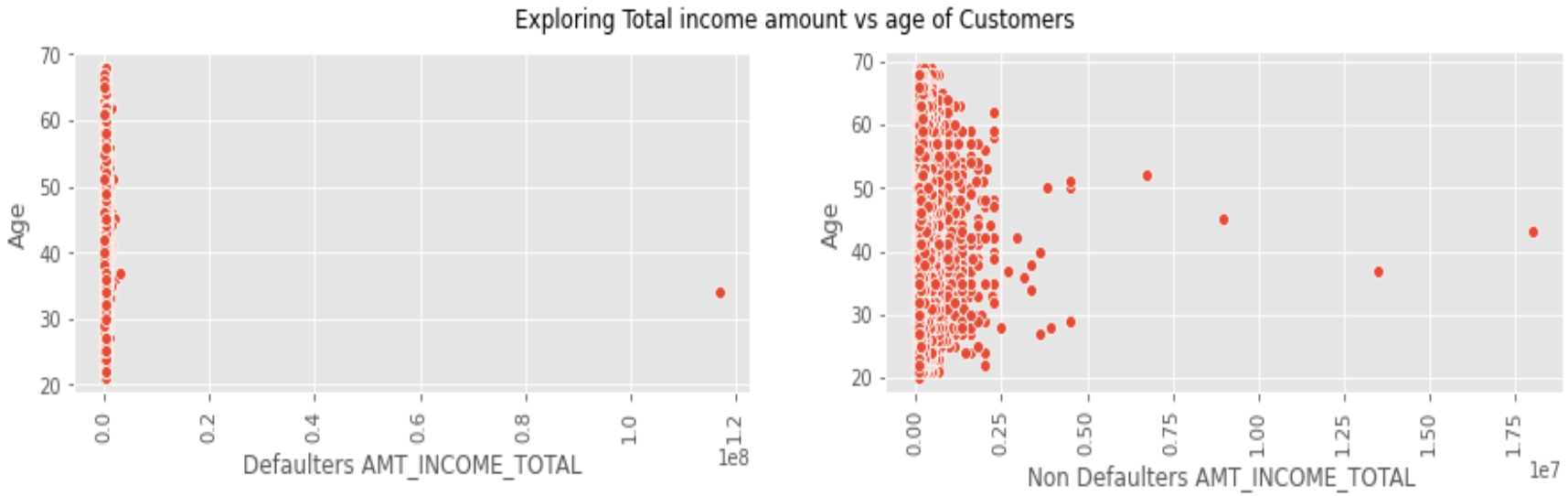


Distribution of Income Amount vs Annuity Amount

Observation:
Annuity amount is high for defaulters though the income amount is in the lower range.

Points to conclude:
Hence it may be difficult to repay with lower income and high annuity. None of the defaulters have a high income amount except for an outlier.

Bivariate analysis -Continuous Vs. Continuous



Distribution of Income Amount Vs. Age of customers

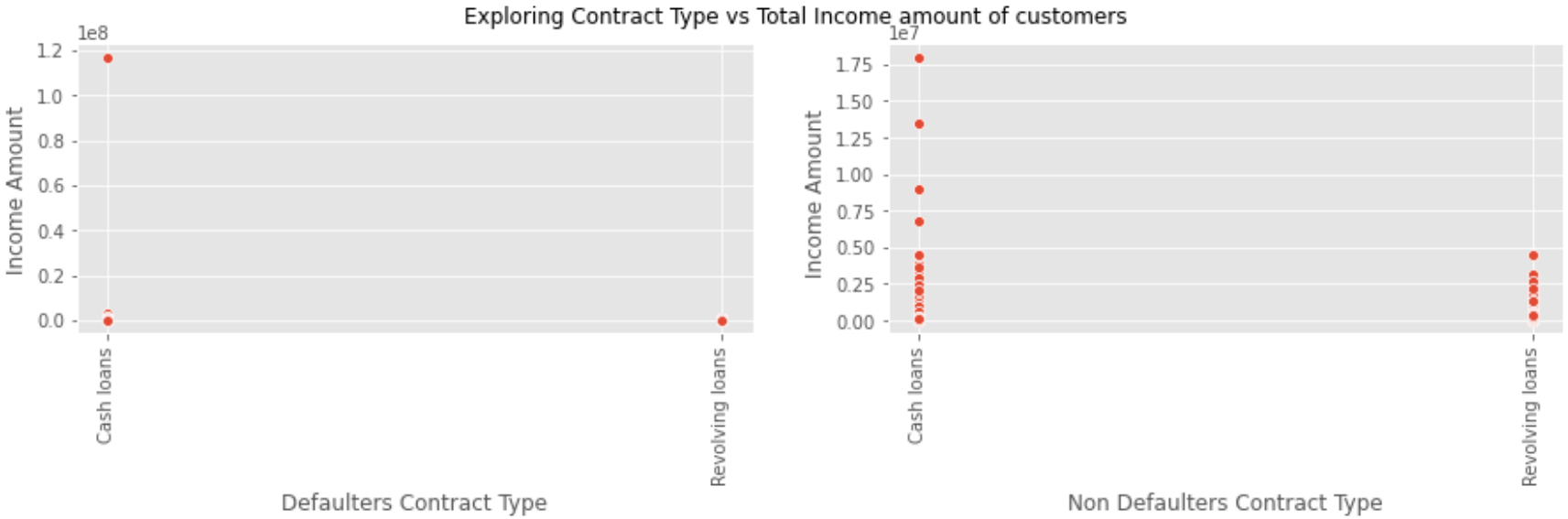
Observation:

The defaulters have lower salary compared to non defaulters except for an outlier.

Points to conclude:

Irrespective of age if the total income is low then caution has to be exercised before Sanctioning loan.

Bivariate analysis - Continuous Vs. Categorical



Distribution of Contract Type and Income Amount of customers

Observation:

Revolving loans are less in the group of defaulters compared to non defaulters.

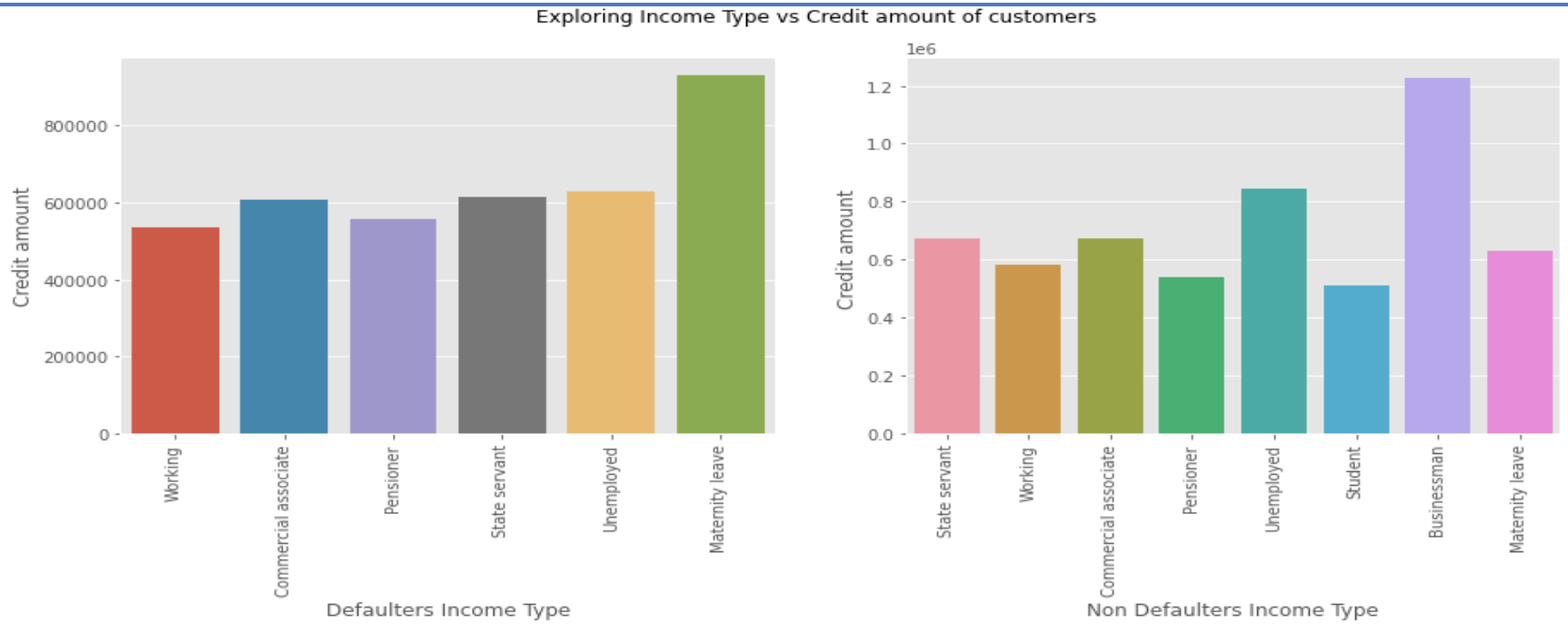
Income amount is lower for the defaulters group irrespective of the loan type. Revolving loans are less in the group of defaulters compared to non defaulters.

Points to conclude:

This graph establishes that customer with lower income irrespective of the loan type

Fall under the defaulter customers group. Hence loans should be sanctioned with caution

Bivariate analysis -Continuous Vs. Categorical



Distribution of Income Type vs Credit Amount

Observation:

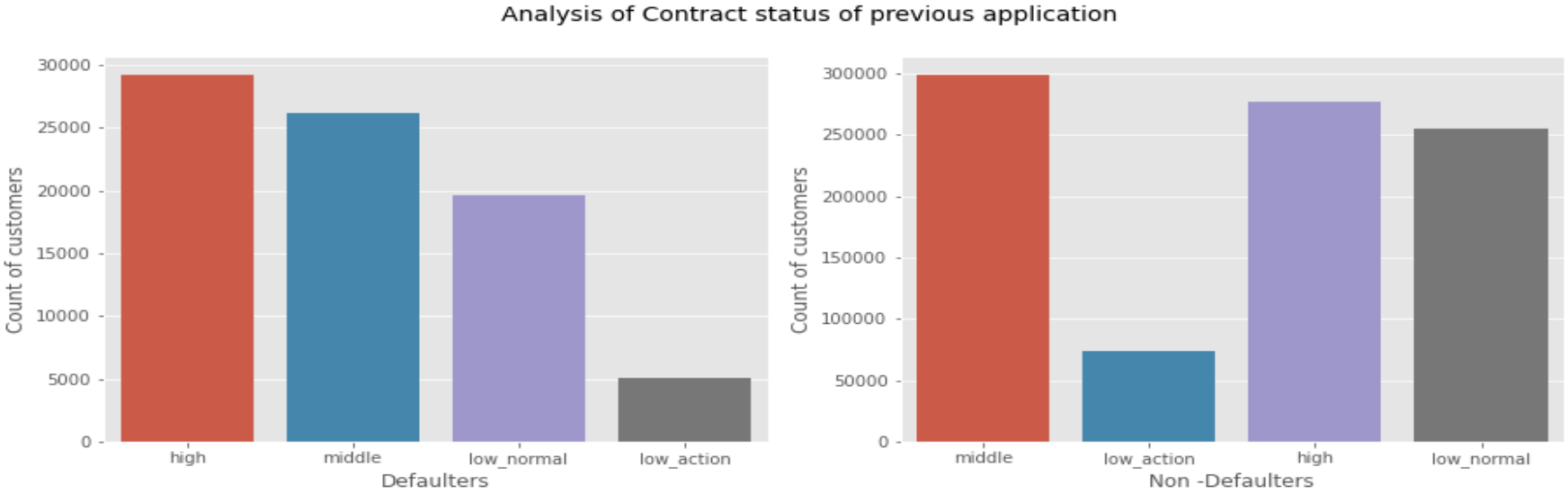
People with Maternity Leave and a credit amount of more than 8 lakhs tend to be defaulters.

Businessman and student don't feature in the defaulters list.

Points to conclude:

Since customers with matrenity leave may not continue with their jobs after the maternity leave, repaying loan maybe difficult. Sanctioning loan of amount higher than 5000000 to customers under maternity leave has to done with caution.

Analysis of Previous Application



Analysis of Contract status of previous application

Observation:

- The graph shows most of the defaulters had a interest rate of high and middle while non-defaulters had a middle interest rate.

Points to conclude:

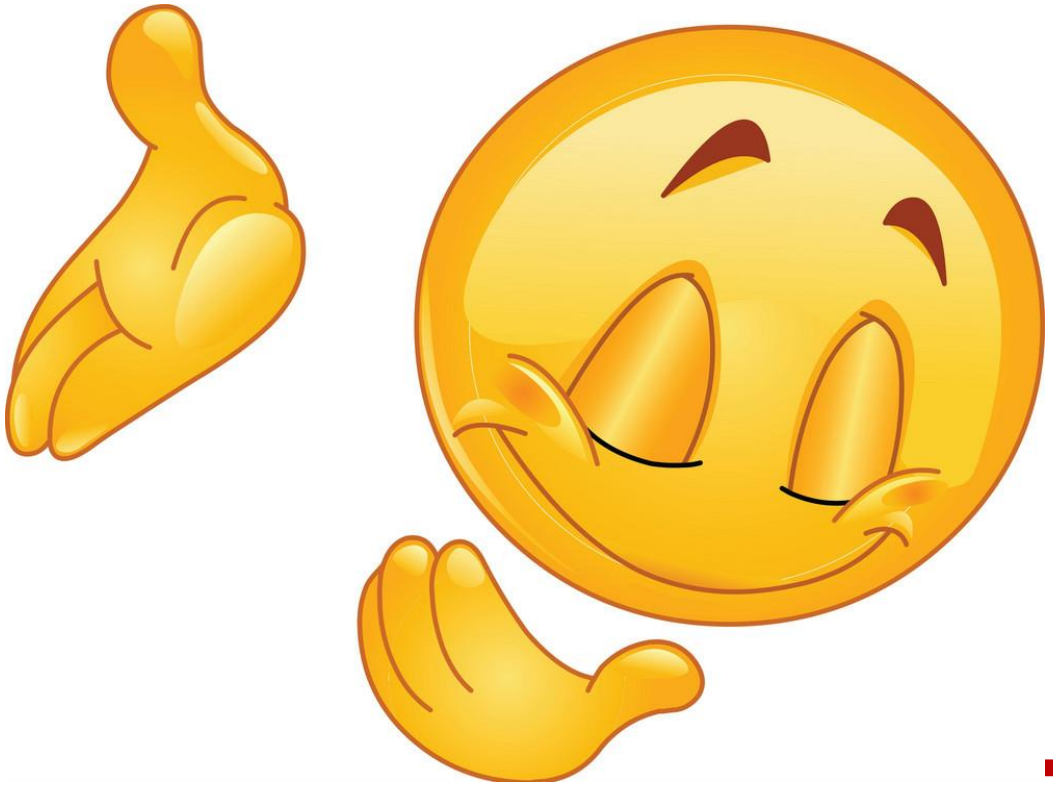
- This means with high interest rate the annuity will also increase and makes it difficult to repay the loan. For non defaulters the contract status is low action making it easy to repay the loan.

Conclusion

The following variables have a key role in deciding whether customers will default or not.

1. **Income Type**
2. **Income Amount**
3. **Annuity**
4. **Educational Type**
5. **Credit Amount**
6. **Contract Status**
7. **Contract Type**
8. **Gender**
9. **Age**
10. **Time frame in which the Identity document was changed**

These can be called as driver variables and the insight was inferred from the graphs represented.



Thank You