Credit EDA Assignment

Parijaat Sunil

January 2024

Problem Statement & Data Summary

- **Problem Statement:**
  - Identify and demonstrate factors which are strong indicators of default

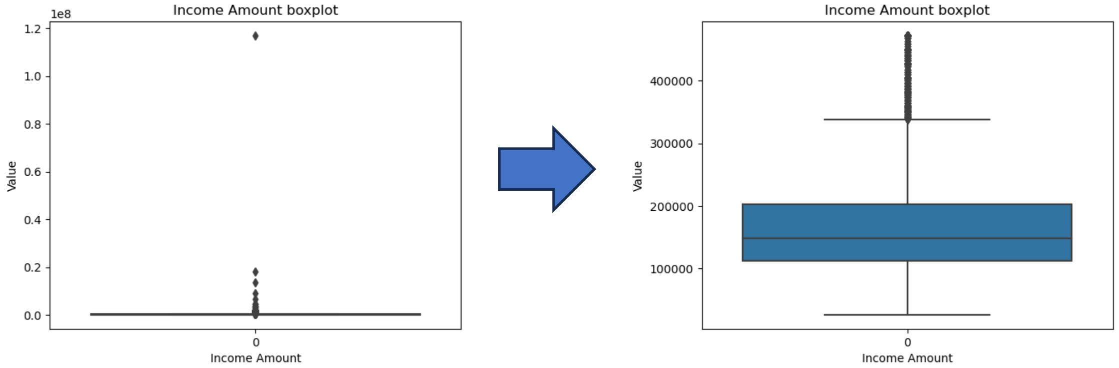
- Data Summary
  - Application Data
    - 307,511 Rows x 122 Columns
    - Column data types: Float(65), Integer(41), Object(16)
  - Previous Application
    - 1,670,214 Rows x 37 Columns
    - Column data types: Float (15), Integer(6), Object(16)
  - Column Descriptions
    - Reference data which contains the detailed column descriptions

Assumptions

- The EXT_SOURCE columns carry the normalized credit scores of the client from different credit agencies
- The REGION_RATING_CLIENT_W_CITY supersedes the REGION_RATING_CLIENT column
- Pensioner clients have the maximum work duration
- Unemployed clients have no work experience
- Civil marriage is the same as married

Missing & Outlier Values

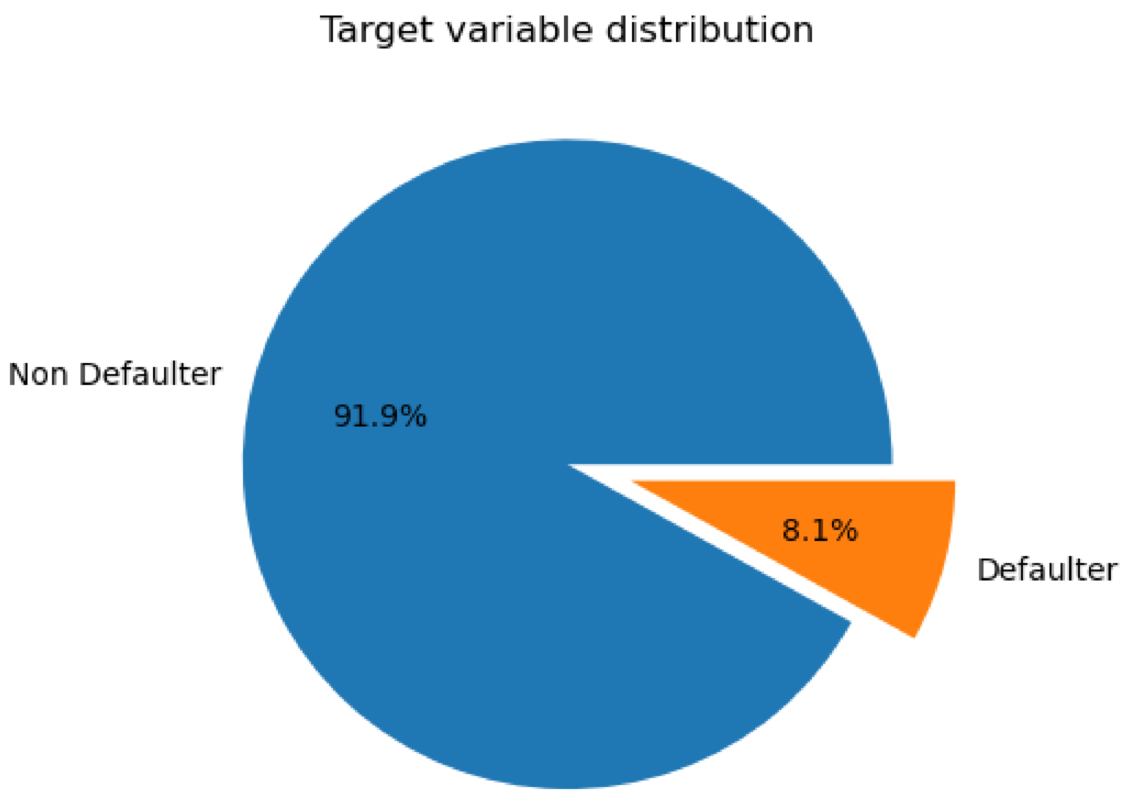
- Missing Values
  - There are 4 approaches we have used in this study to handle missing values:
    - **Column drop:** >30% missing values or weak correlation with target
      - Example: Normalized information where the client lives columns; Weekday process start column
    - **Imputing missing values:** mean (where there are not many outliers), median (where the outliers are many and far flung) or mode (where the column is categorical)
      - Example: Society circle default column; Annuity amount column; Gender column
    - **Imputing missing values:** finding relationships to other columns
      - Example: Goods value column filled in with Credit amounts
    - **Row drop:** number of missing values were <1%
      - Example: Count of family members column; Credit score column
- Outlier Values
  - The outliers have been identified by using boxplots
  - We have handled outliers by two methods:
    - Value Capping
    - Retaining Outliers



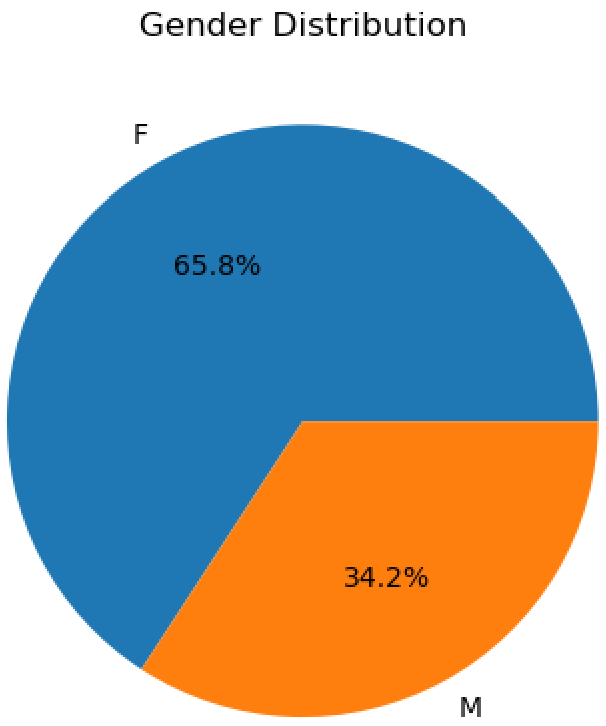
Data Imbalance – Target Variable

- There is a data imbalance with respect to the Target variable
  - The ratio of imbalance is 11.39

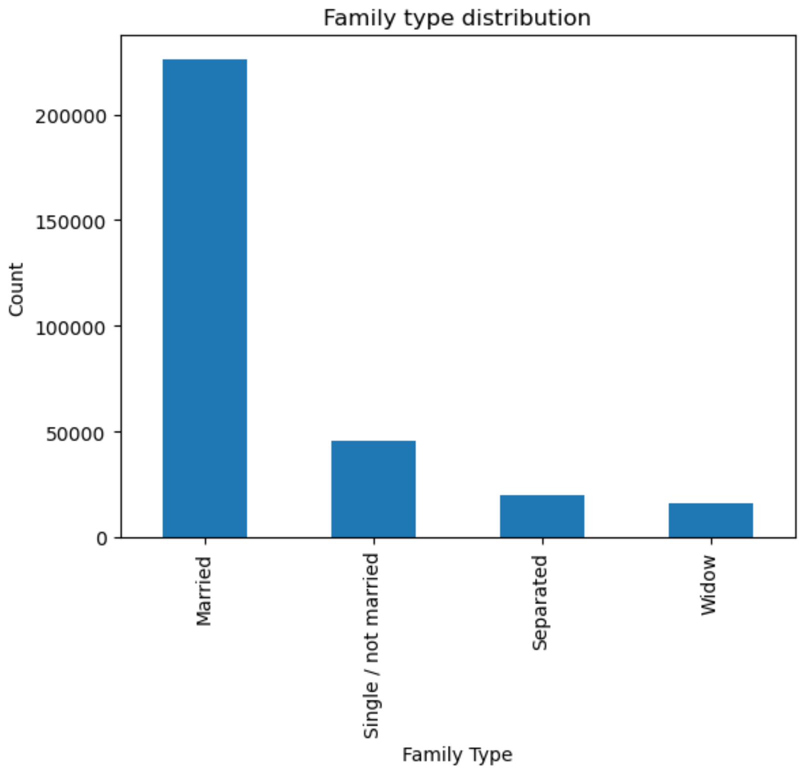
- This imbalance is expected as the number of defaulters would have to be lower than non-defaulters for the bank to fuction



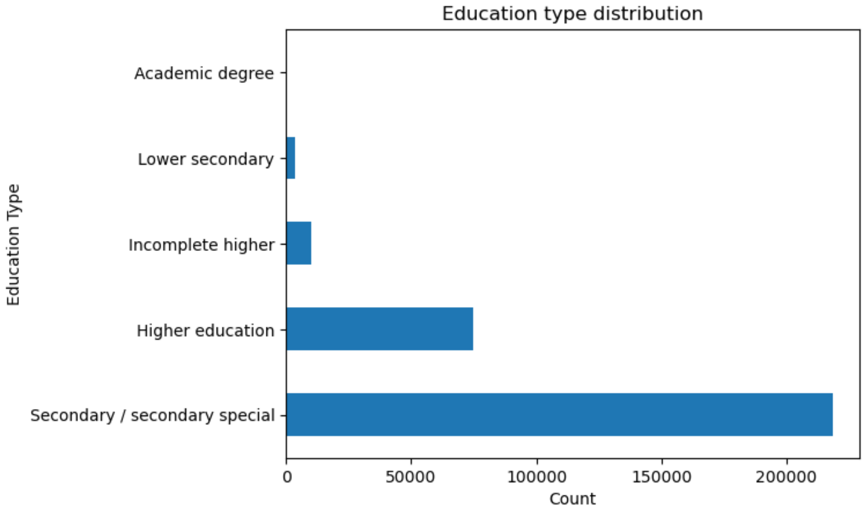
Application Data – Univariate Analysis



Inferences:
- Most applicants for loan are female

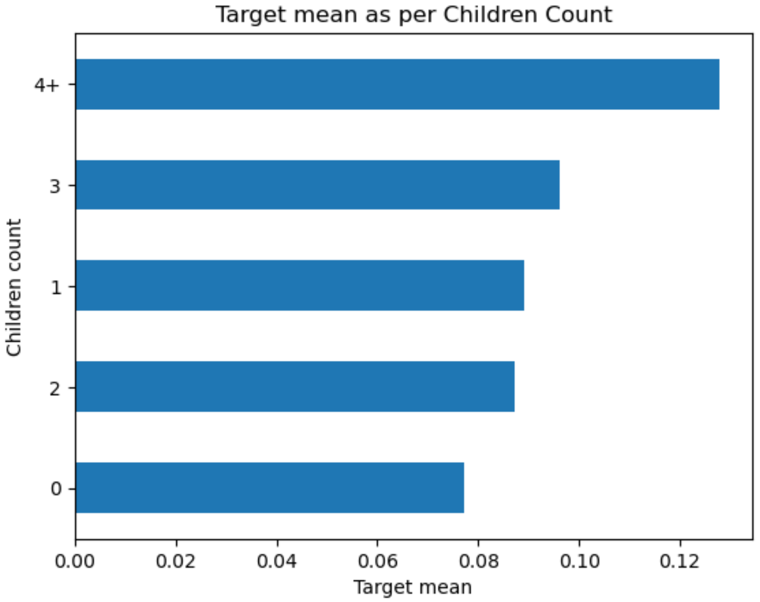


Inferences:
- Most applicants for loan are married



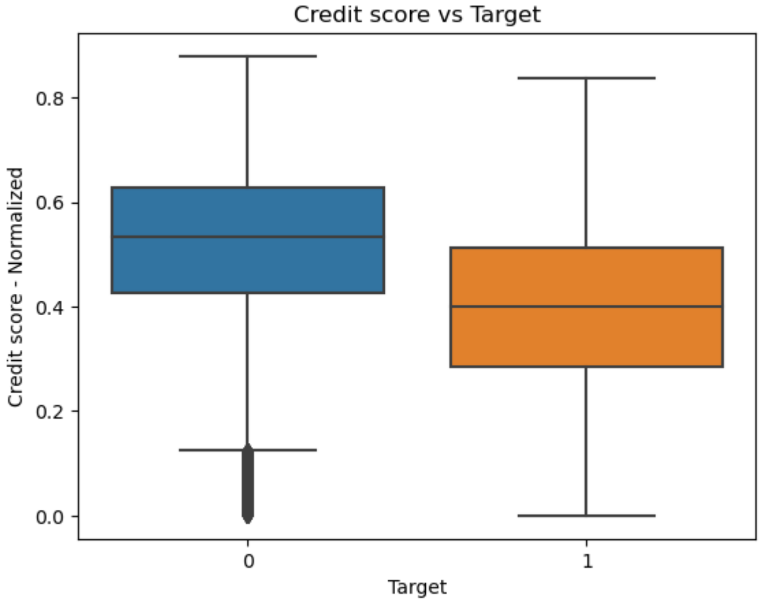
Inferences:
- Most of the clients are from the Secondary education level
- The next biggest level of education is Higher Education
- Lowest number of clients have an academic degree

Application Data – Bivariate Analysis (1/2)



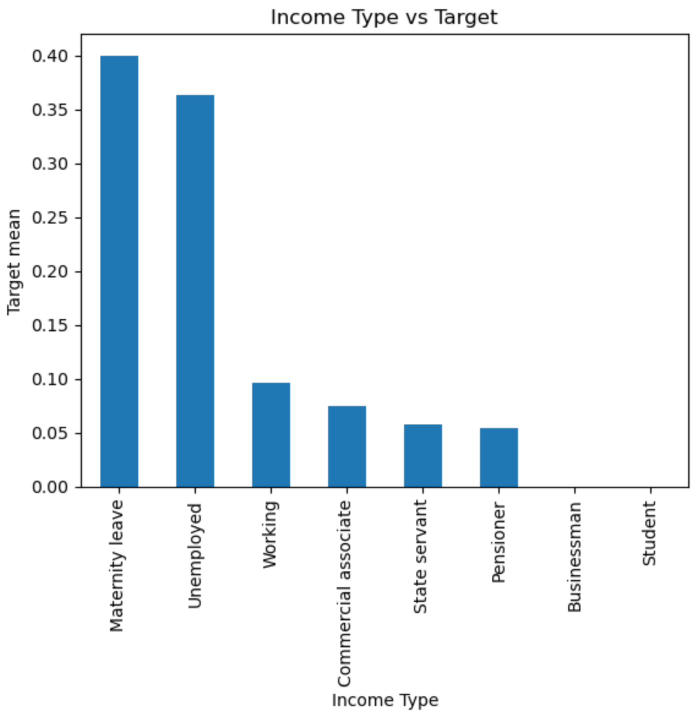
Inferences:

- The occurrence of default in case of clients with 0 to 3 children is almost similar, with clients having no children having the lowest rate of default
- Clients with 4 or more children have a much higher rate of default



Inferences:

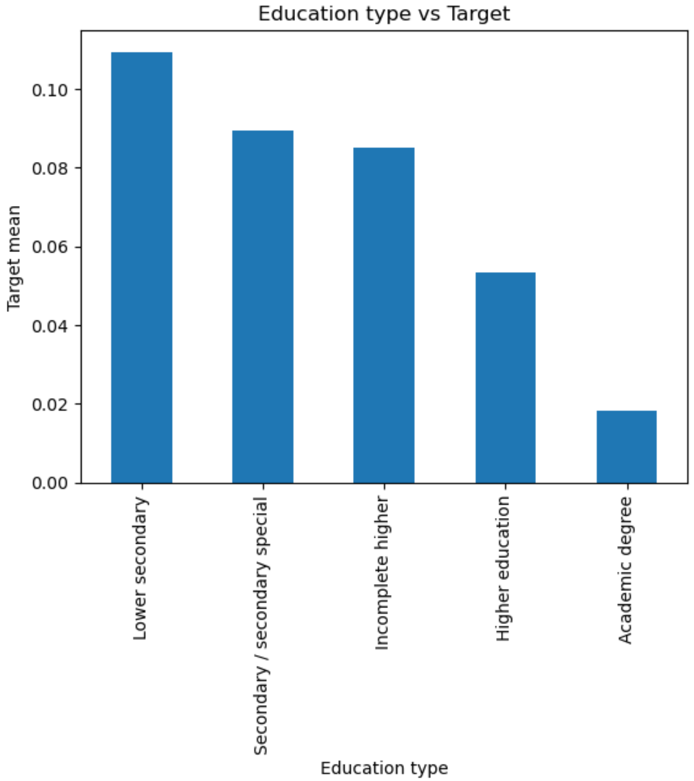
- Lower credit score is a strong indicator of loan default



Inferences:

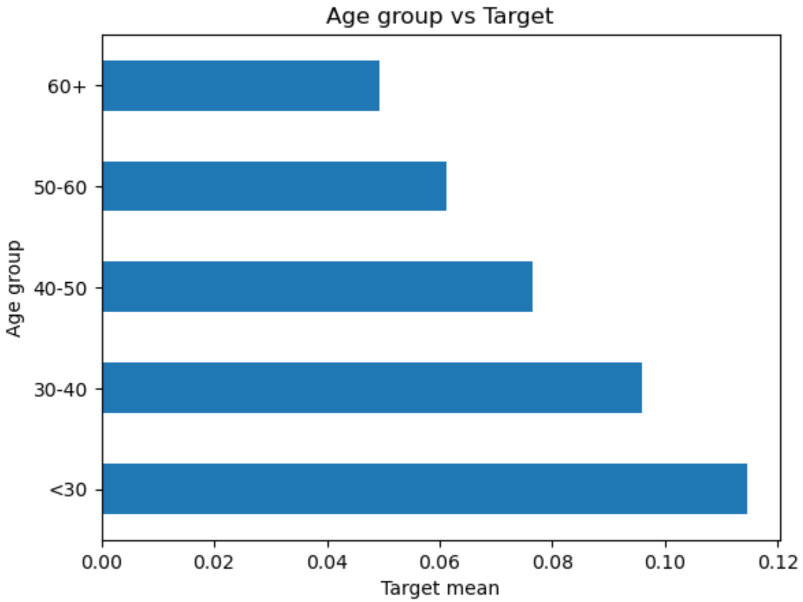
- Clients who are on maternity leave at time of application or unemployed have a higher occurence on defaulting on their loan
- Students and businessmen have the lowest occurence of loan default

Application Data – Bivariate Analysis (2/2)



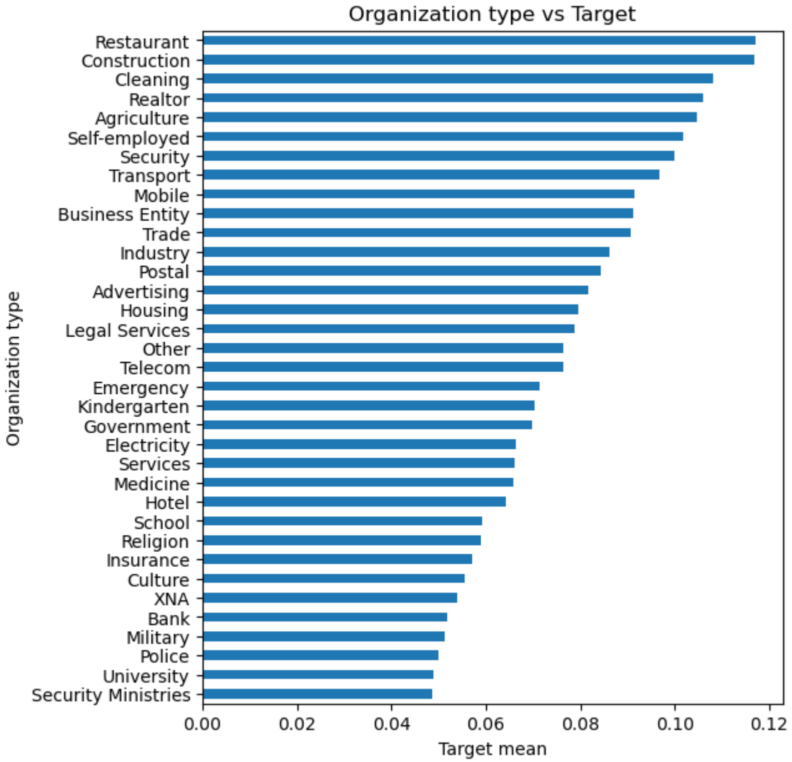
Inferences:

- Clients who have an education of lesser than higher education have higher frequency of default
- Clients with an academic degree have the lowest occurence of loan default



Inferences:

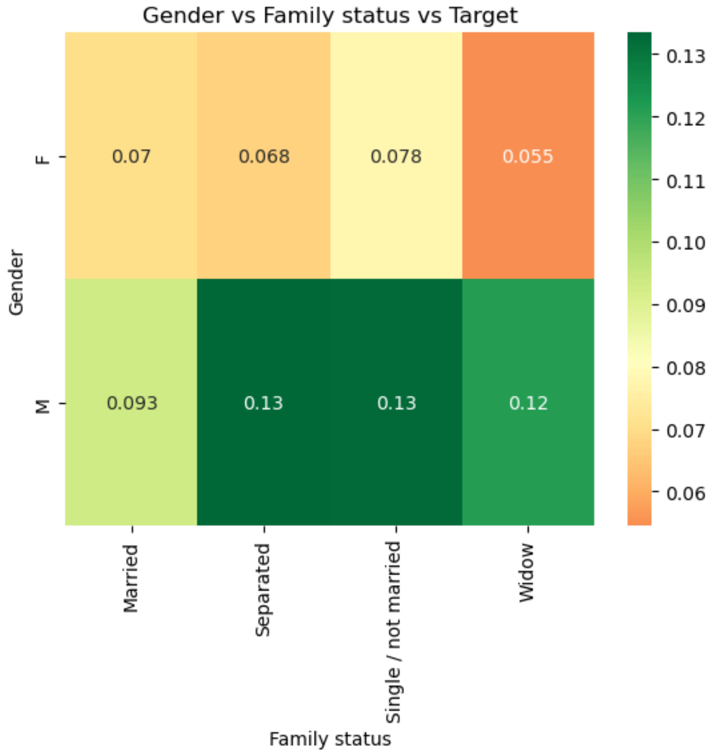
- Younger Clients have a higher occurence of defaulting on their loan
- Loan default occurence reduces with age



Inferences:

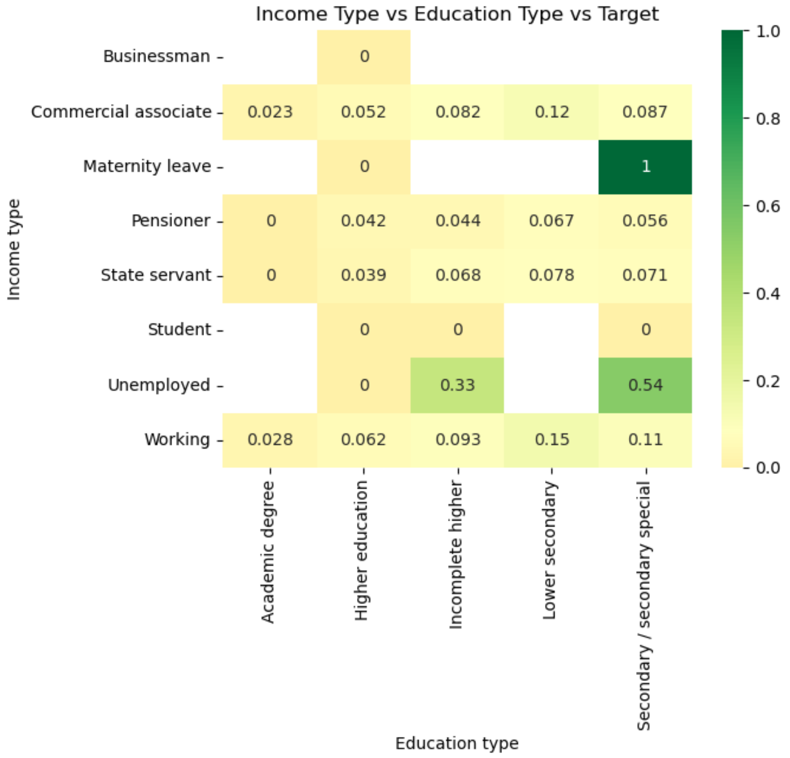
- Riskier business industries like restaurants, construction and others have the highest Occurrence of defaults
- Government and stable industries like Banks have the lowest Occurrence of defaults

Application Data – Multivariate Analysis



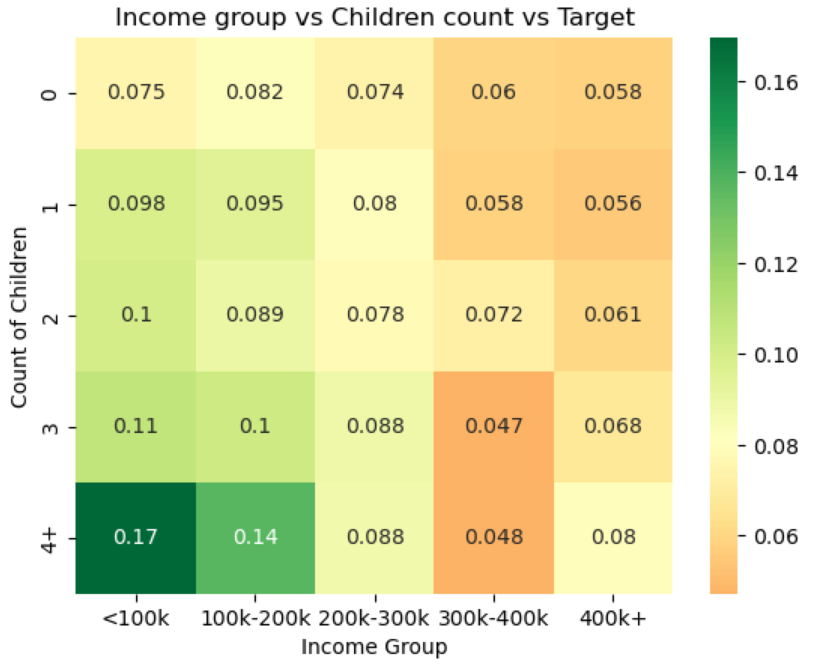
Inferences:

- Unmarried men (single, separated or widowers) have the highest occurrence of loan default
- Female widows have the lowest occurrence of loan default



Inferences:

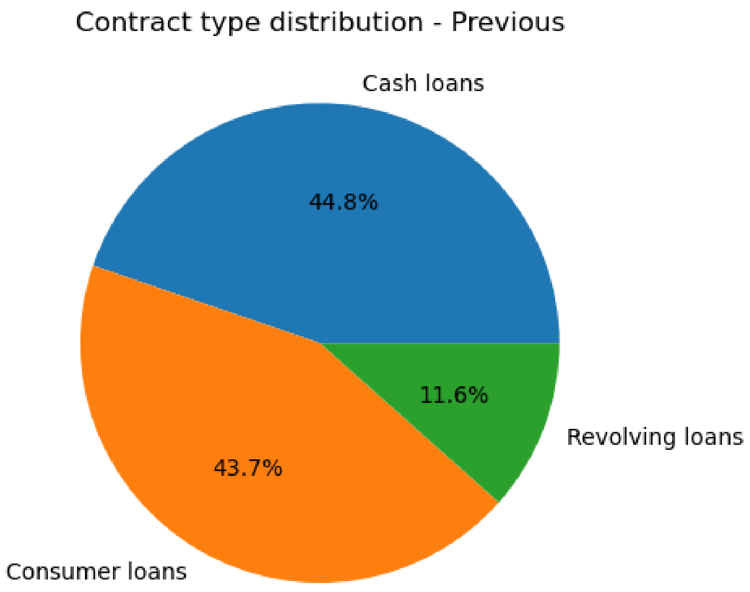
- Clients on maternity leave who have not completed higher education have the highest occurrence of loan default.
- Similarly unemployed clients who have not completed higher education have the next highest occurrence of loan default.



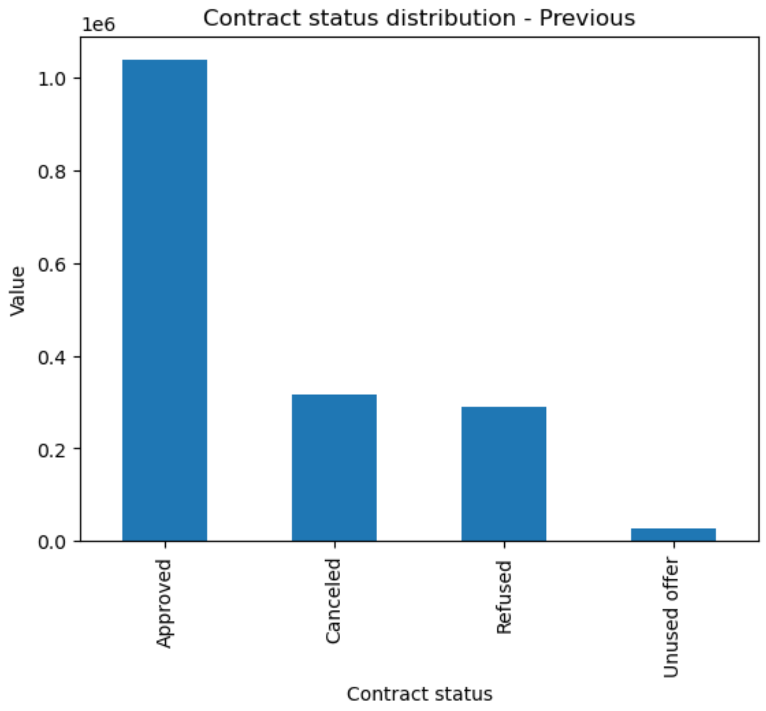
Inferences:

- There is a significant increase in the occurrence of loan default in clients having the lower income group and 4+ children
- Clients having lesser children or higher income have a lower rate of loan default

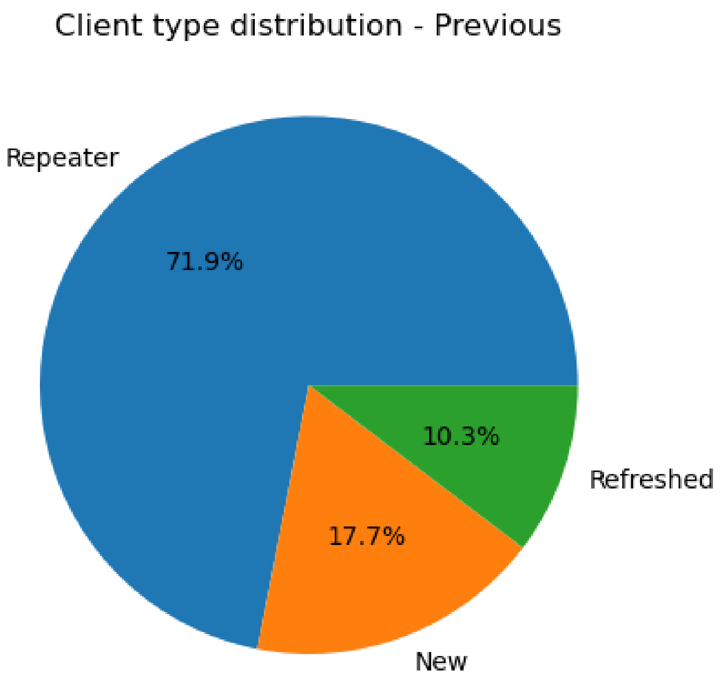
Previous Application – Univariate Analysis



Inferences:
- There is a new category of loans in the previous application dataset which accounts for approximately 44% of the data

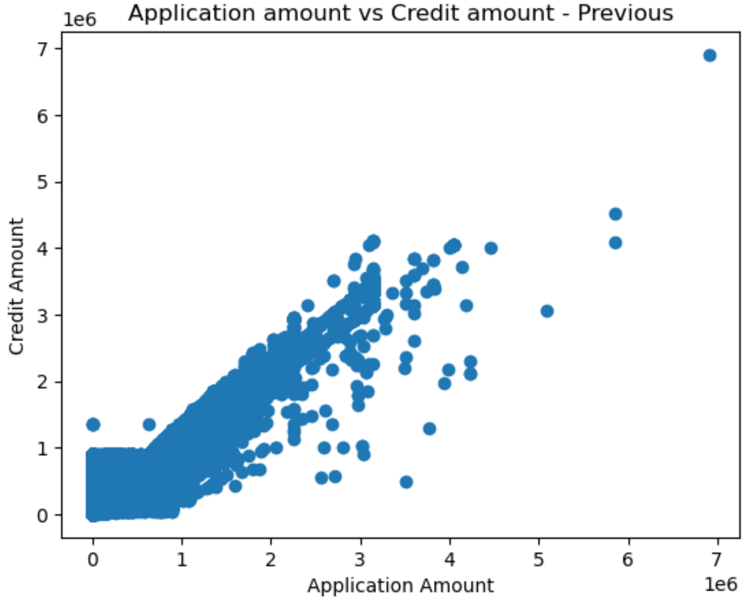


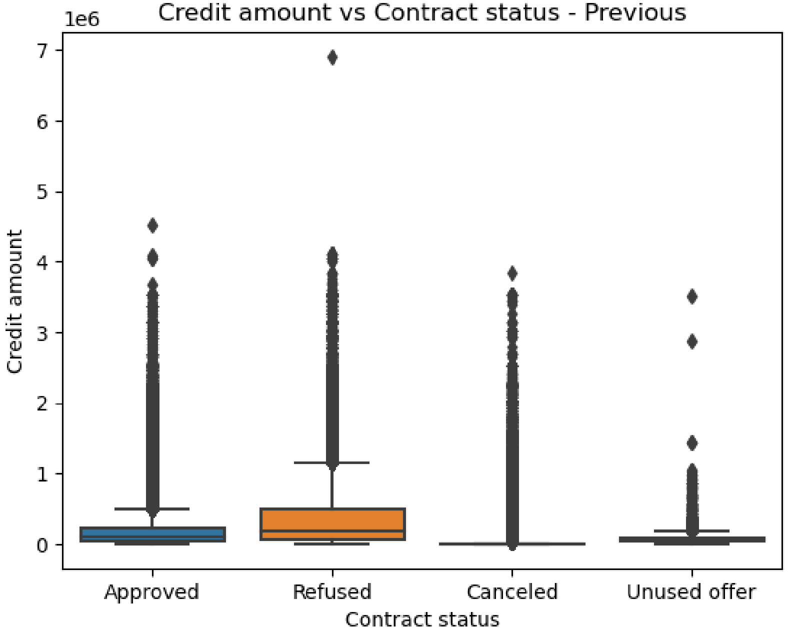
Inferences:
- Most of the previous applications were Approved and very few were unused.

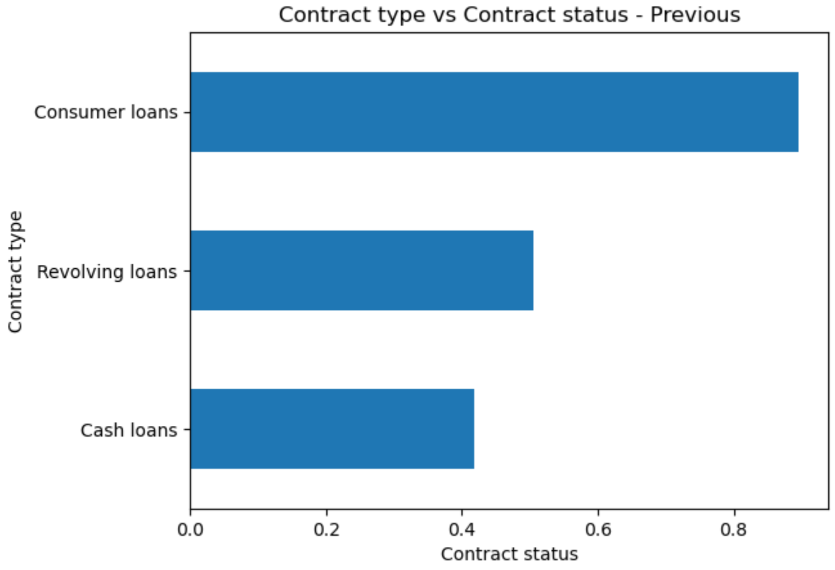


Inferences:
- Approximately 74 % of loans were from repeat clients

Previous Application – Bivariate Analysis







Inferences:

- There is a strong correlation between the application amount and credit amount, which makes sense as credit amount is dependent on application amount

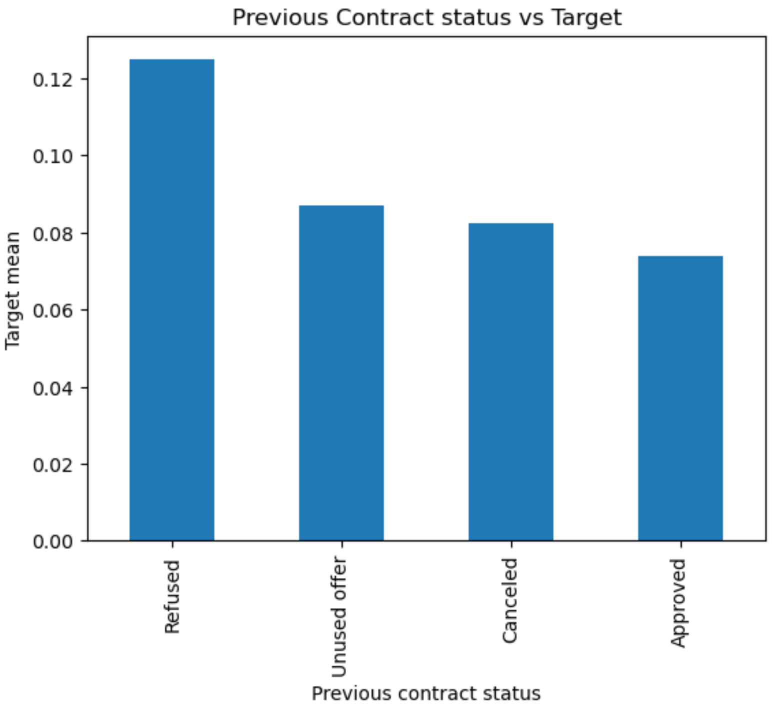
Inferences:

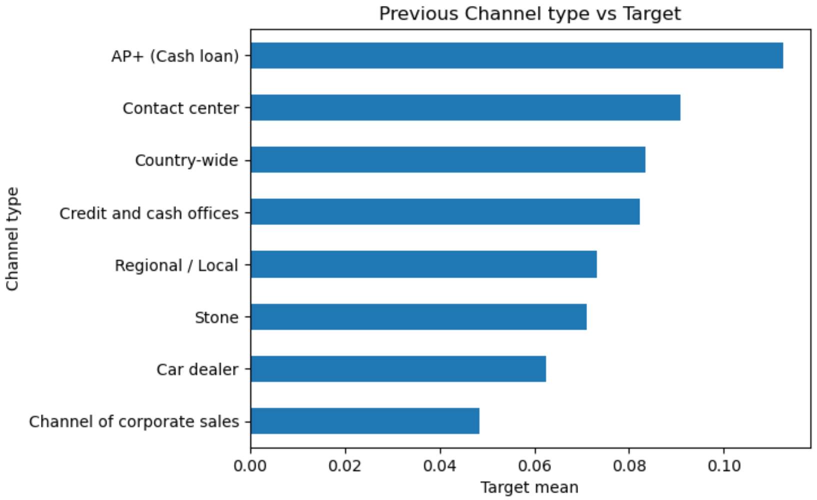
- Applications which were refused had a wider spread of the credit amount
- There are quite a few outliers in each case, with most credit amounts being lower than approximately 1,000,000

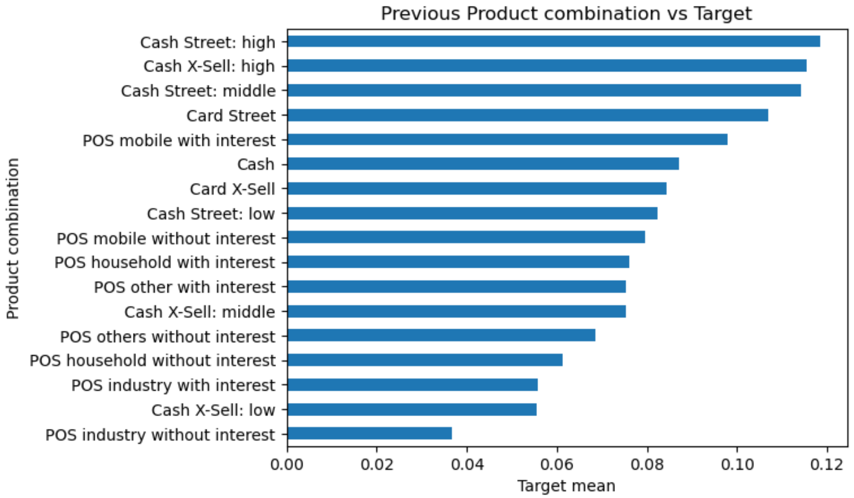
Inferences:

- Consumer loans had a higher occurrence of getting approved
- Revolving and Cash loans had lower occurrence of being approved with Cash loans being lowest

Combined Dataset – Bivariate Analysis







Inferences:

- The loan default rate is higher for cases where the previous application was refused
- For all other categories, the default rate is similar

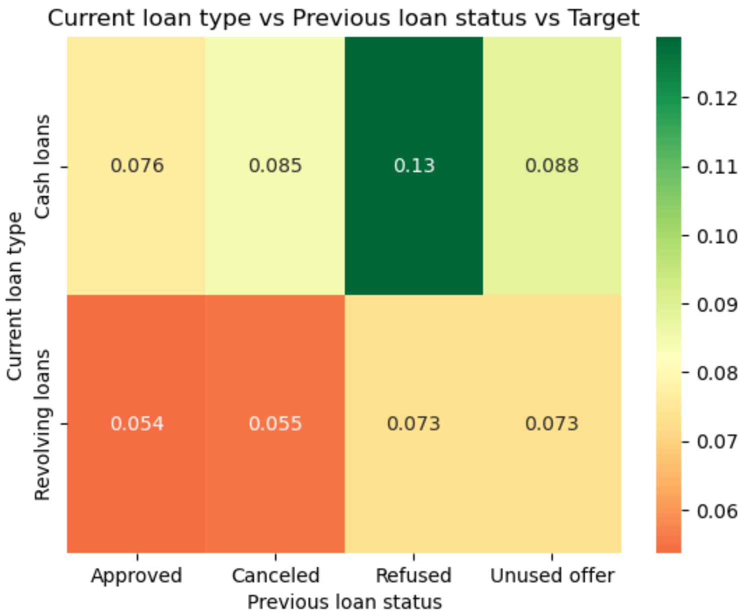
Inferences:

- Clients which have been acquired from channel AP+ (Cash loan) have the highest rate of default
- Clients which are acquired through corporate sales have lowest rate of default

Inferences:

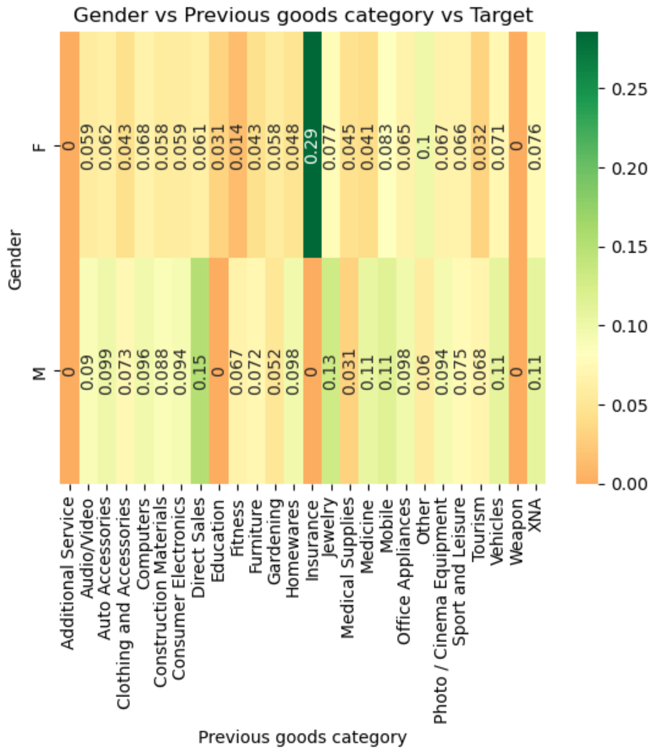
- Clients with product combination of Cash Street: high have the highest occurrences of default
- Clients with POS industry without and with industries have the lowest rates of default

Combined Dataset – Multivariate Analysis



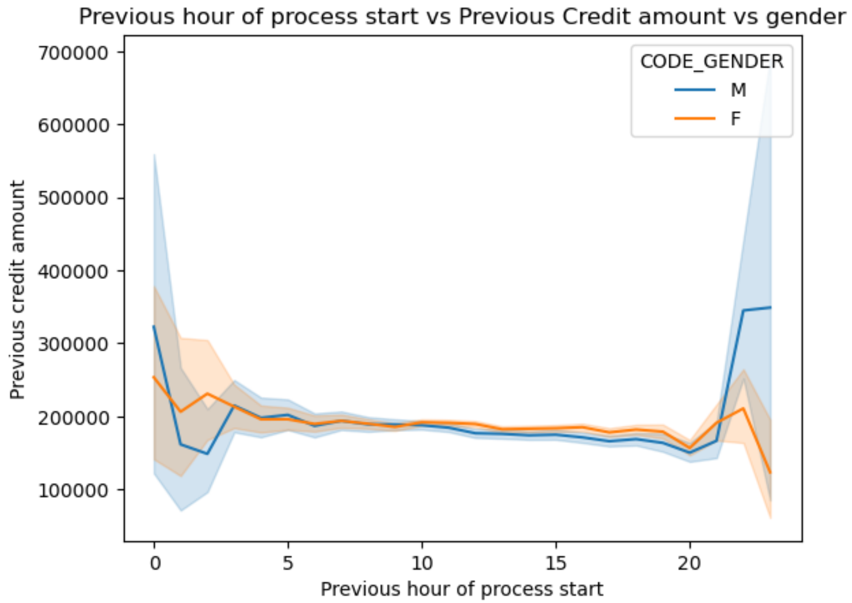
Inferences:

- Clients applying for Cash loans with a history of being refused have the highest occurrence of default



Inferences:

- Females taking loan for insurance previously have the highest rate of default
- Males have the highest occurrence of default for Direct Sales and Jewelry



Inferences:

- Male clients usually submit applications late at night
- Applications for larger credit amounts are usually placed during late nights

Top Correlations with Target

- The target variable is most impacted by the below variables:
  - **Age –** Younger clients more likely to default
  - **Income –** Lower income clients more likely to default
  - **Credit score –** Lower credit score clients more likely to default
  - **Count of Children –** Clients with 4+ children have higher occurrence of default
  - **Education –** Clients are more likely to default at lower levels of education
  - **Organization type –** Clients working in riskier industries are more likely to default
  - **Marital Status –** Unmarried clients are more likely to default
  - **Previous contract status –** Clients who have been refused loans earlier have a higher occurrence of default
  - **Channel Type –** Corporate clients and AP+ Cash clients have the lowest and highest occurrence of default respectively
  - **Goods Category –** Loans taken for jewellery and direct sales have highest rates of default

Thank You