EDA case study on Banking Dataset.

Group Partners – Vagmi Gupta & Abdul Rehman

Table of Contents

1 Project Domain and Problem Statement

2 Approach to Data cleaning

3 Approach to Data Analysis

4 Data visualisation and Insights

5 Driver variables for identifying Default

1

Project Domain and Problem Statement

Domain of the Project / Case study

The case study majorly focuses on applying the EDA techniques on real life business scenario of *Loan approvals.*

The major aim of this project is to understand the risk analytics in the Banking and Financial services and utilize the data to minimize the risk of losing money while lending to customers.

The major losses that a bank can infer is :

- ❑ Credit loss: If the bank approves the loan for the client who defaults later. This will lead to lead to a financial loss for the bank.
- ❑ Interest loss : if the bank rejects the loan for the client who can re-pay the loan. This will lead to loss of business to the bank.

Problem statement

This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.

The application data has been categorized based on the applicant's profile. There are two types 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 the client applies for the loan, there are 4 decisions that could be taken.

- ❑ Approved
- ❑ Cancelled
- ❑ Refused
- ❑ Unused Offer

“If you Torture the Data enough it will confess to anything !.”

—Ronald H. Coase

Datasets

The Datasets are provided in 3 different files

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.

Approach to Data cleaning

2

Approach to Data Cleaning

Data cleaning is one of the most crucial step in Exploratory Data Analysis. This is where the data is correctly transformed and the datasets are prepared for Data Analysis.

The major steps followed in this case study are:

- ❑ Data types : checking if all the columns have correct data types.
- ❑ Checking for missing & null values.
- ❑ Treating outliers.
- ❑ Treating missing values
    - ❑ Dropping the columns with more than **50%** missing values.
    - ❑ Imputing the rest of the column with missing values with below techniques
        - ❑ Mean for numerical columns without outliers.
        - ❑ Median for numerical columns with outliers.
        - ❑ Mode for categorical columns.

3

Approach to Data Analysis

Approach to Data Analysis

Data Analysis, this is the stage where we start understanding the message contained in Data.

The major steps followed for Data Analysis are:

- ❑ Normalizing and categorizing certain columns for aiding the Data Analysis.
- ❑ Univariate analysis to understand the composition of certain variables in the dataset
- ❑ Bi-variate & Multi-variate Analysis between certain columns to find the correlation between the columns.

Data Analysis - Normalization

Normalizing is a step used to transform certain data by standardization of units of certain variables and by bucketing for ease and better effectiveness of data analysis.

The following columns have been Normalized :

- ❑ **DAYS_BIRTH & DAYS_EMPLOYED :** Since this data had the Age & Employment experience in days and had negative values relative to the application date, the following columns were created:
    1. *AGE (IN YEARS)* & *YEARS_OF_EMPLOYMENT* columns were added converting the Days values to Years.
    2. *AGE_IN_YEARS_RANGE* & *YEARS_OF_EMPLOYMENT_RANGE* columns were added after bucketing the values in bins of 5 years & 1 years respectively.

- ❑ **AMT_INCOME_TOTAL :** *INCOME (IN LAKHS)* column was created by converting the Total income values to lakhs. '*INCOME_RANGE*' column was added by binning the values in the buckets of 5 lakhs.

- ❑ **Credit_Ratio:** '*Credit_Ratio*' column was created to describe the Credit Amount requested / Total income of the customer.

- ❑ **AVERAGE_RATING:** '*AVERAGE_RATING*' column was created to describe the Average credit scores of a customer from 2 external sources (*EXT_SOURCE_2 & EXT_SOURCE_3)*

Data visualisation & Insights

4

Approach to Data visualisation

Data visualisation is a step to identify, understand & communicate the insights from the Data analysis to the target audience.

The Graphs and insights for the datasets are in this order

- a. Application Dataset
  - i. Univariate
  - ii. Segmented Univariate
  - iii. Bivariate
- b. Previous application
- c. Merged Dataset

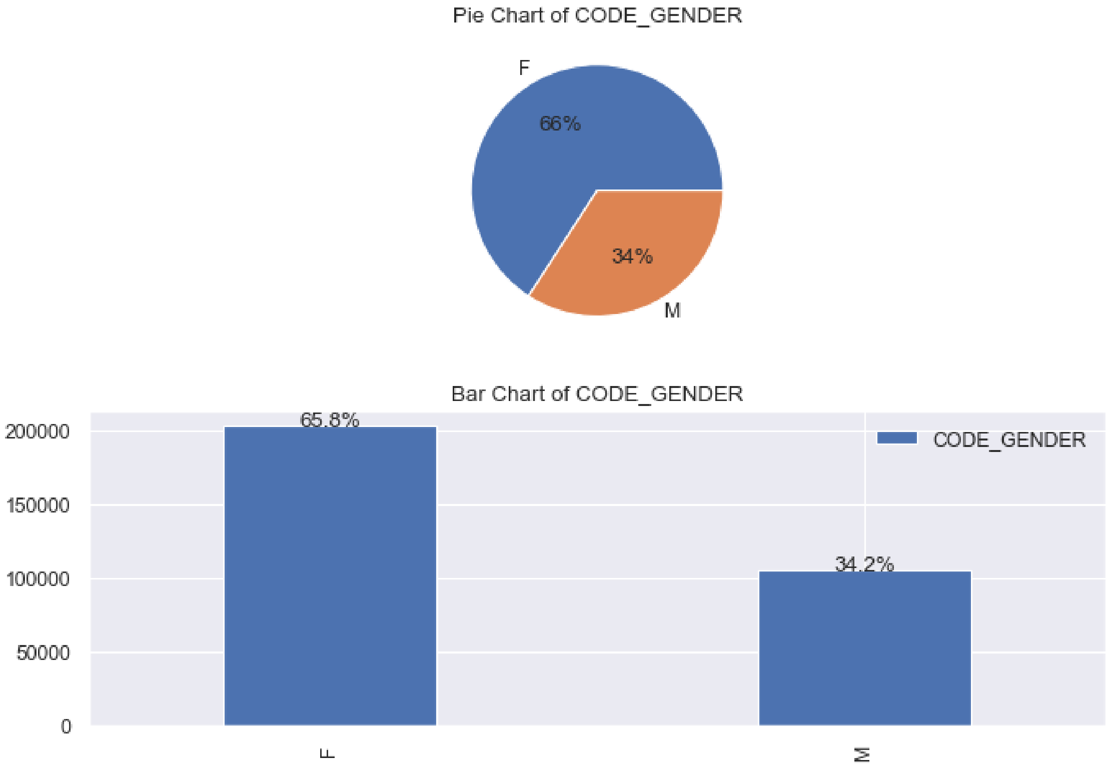
Libraries used:

Matplotlib

Seaborn

Plotly

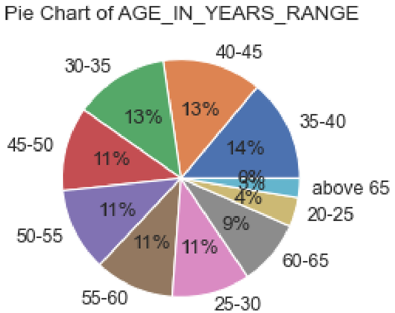
Application Dataset- Univariate Analysis

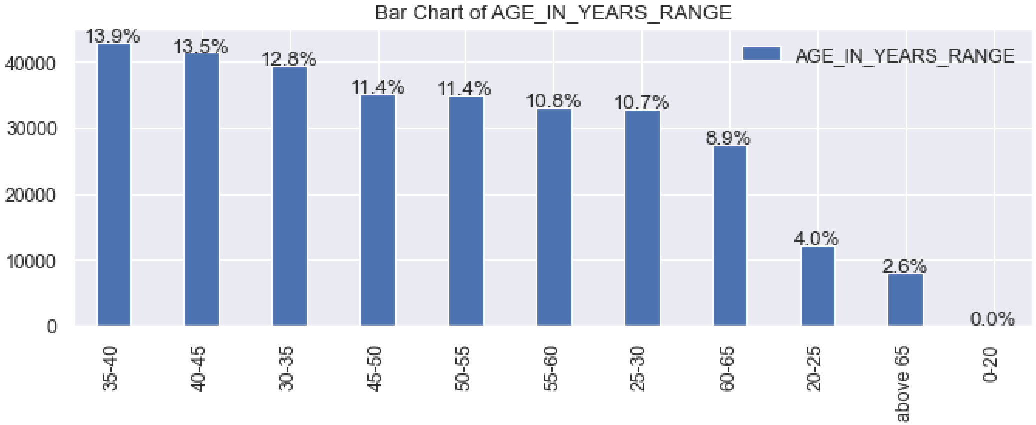


CODE_GENDER:

Female applicants are significantly (approximately double) higher than the male applicants.

Univariate Analysis (cont...)

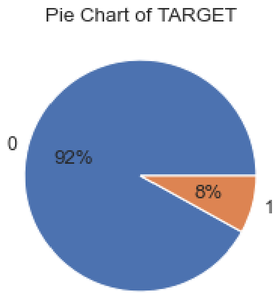




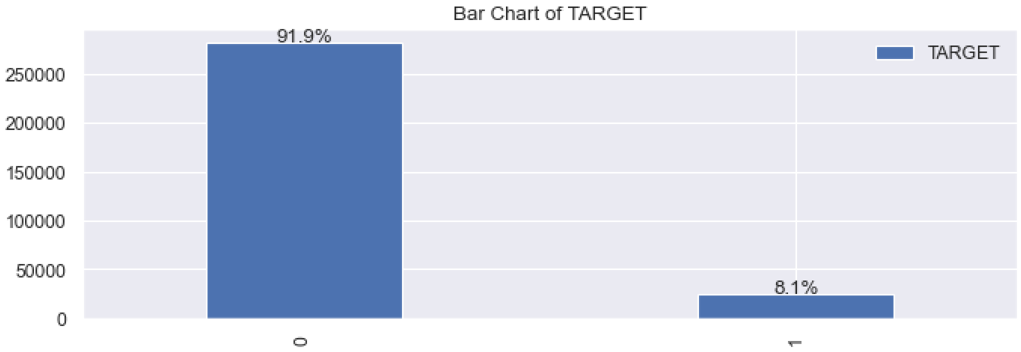
AGE_IN_YEARS_RANGE:

Majority of the applicants are in the age buckets 35 to 40 years followed by 40 to 45.

Univariate Analysis (cont...)



Pie Chart of TARGET

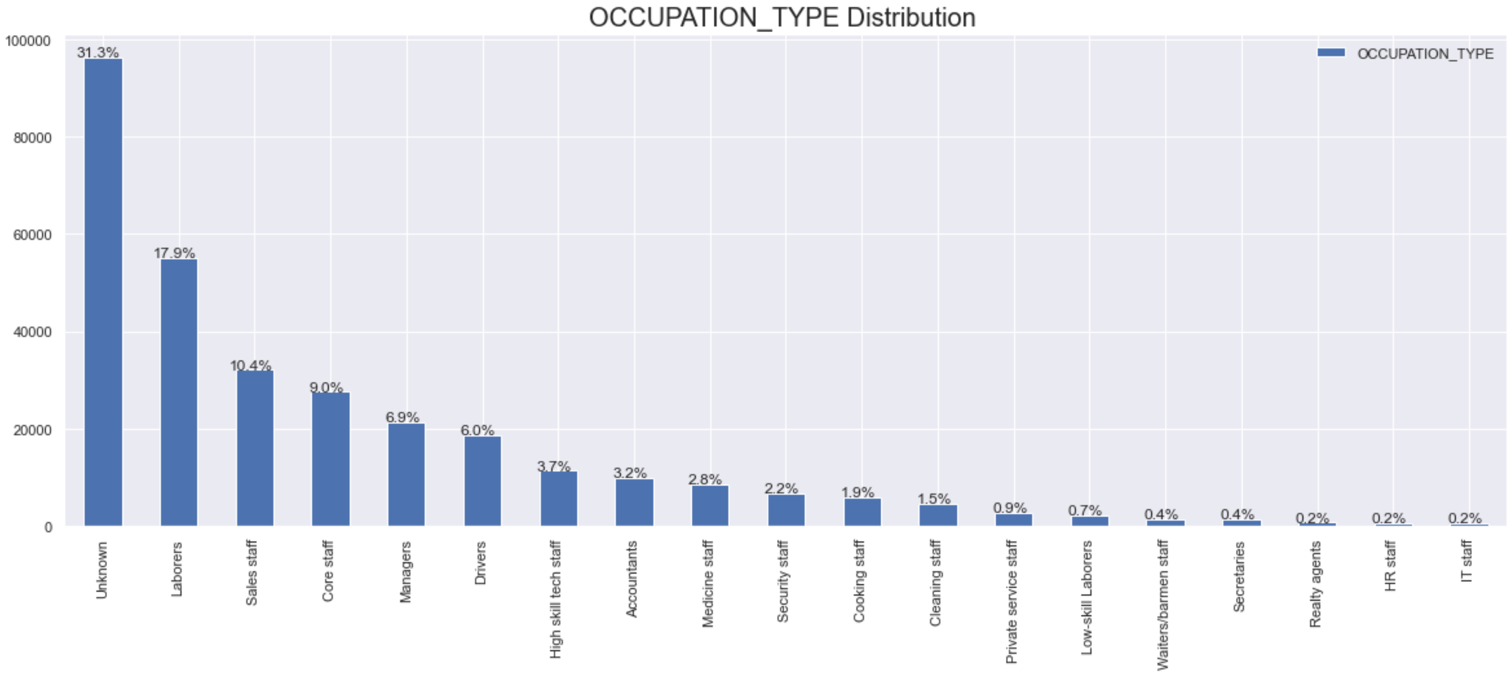


Bar Chart of TARGET

TARGET:

The proportion of applicants with payment difficulties is significantly lesser than the others.

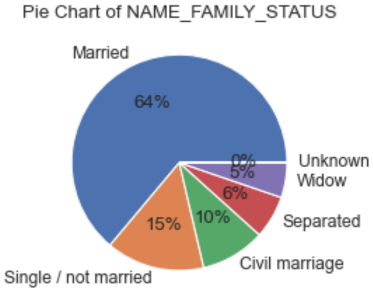
Univariate Analysis (cont...)



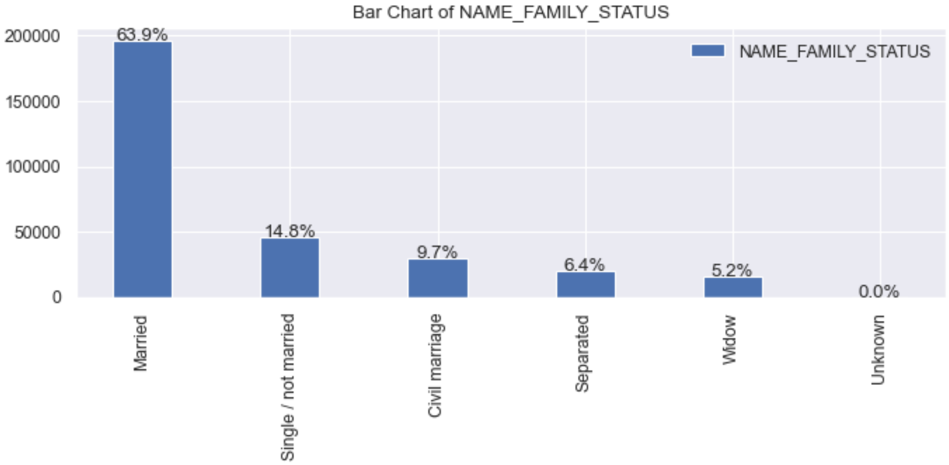
OCCUPATION_TYPE:

Laborers make up a significant proportion of the applicants followed by sales staff and core staff (Most of the data regarding in the Occupation_Type was missing and has been imputed with **Unknown**).

Univariate Analysis (cont...)



Pie Chart of NAME_FAMILY_STATUS

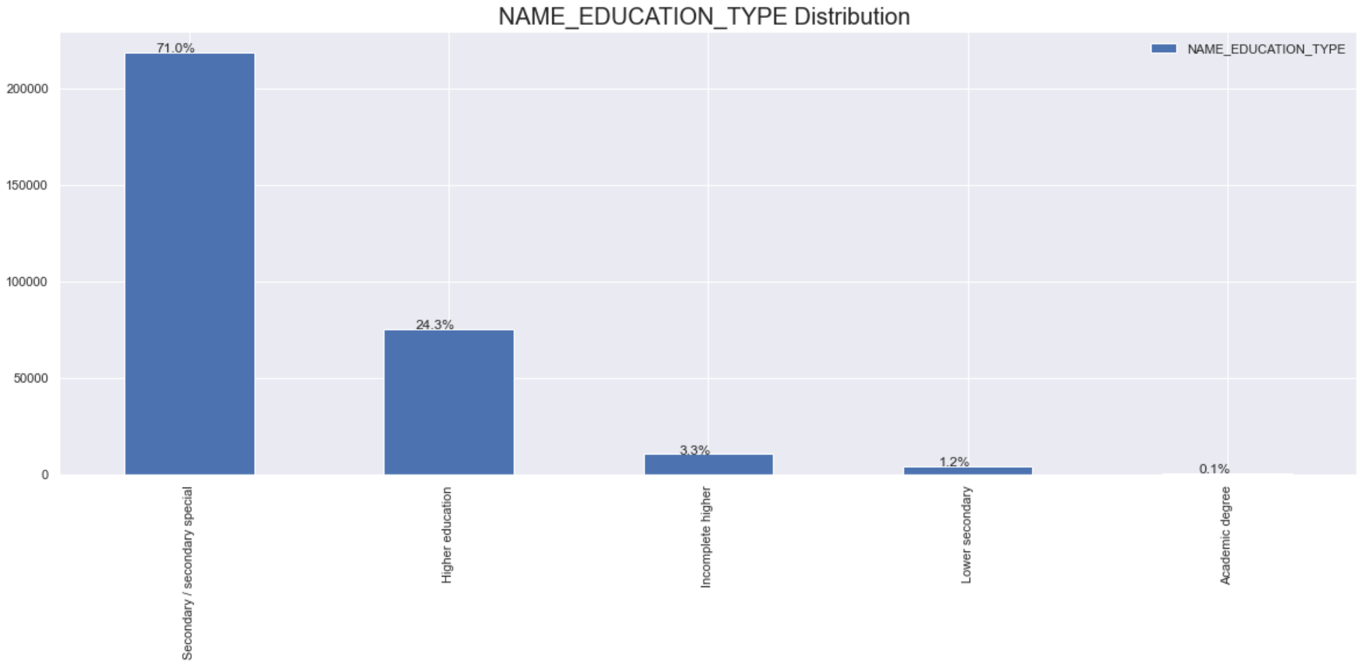


Bar Chart of NAME_FAMILY_STATUS

FAMILY_STATUS:

Major proportion of the applicants are **Married (63.9%)** followed by **Single/not married (14.8%)**.

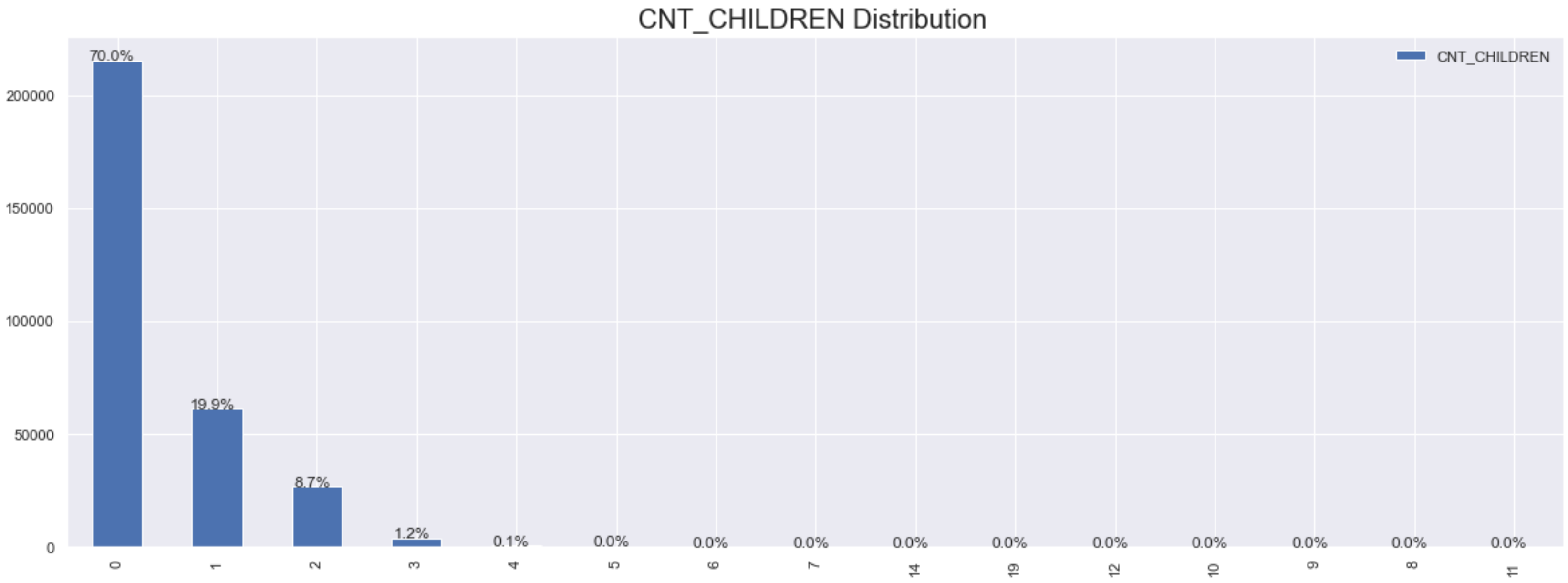
Univariate Analysis (cont...)



EDUCATION_TYPE:

Majority of the clients have secondary/secondary special education followed by higher education.

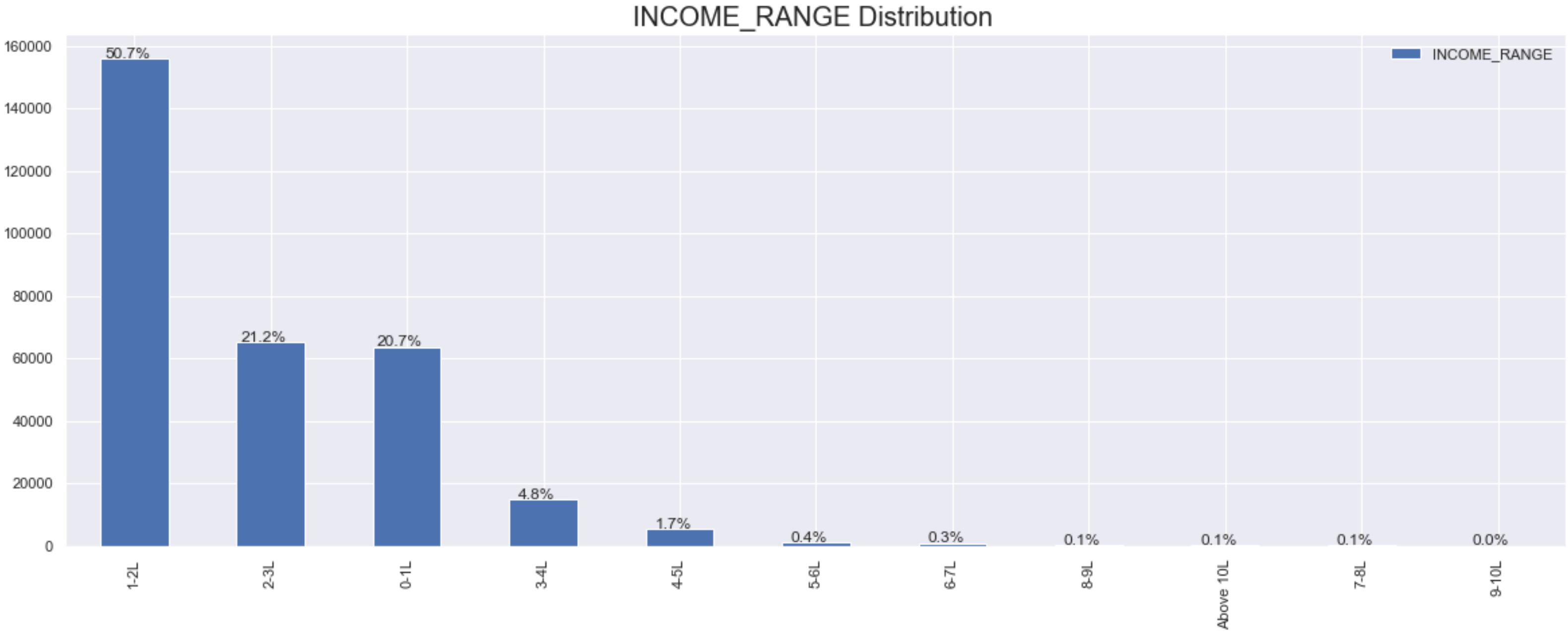
Univariate Analysis (cont...)



CNT_CHILDREN:

1. Major proportion of the applicants (70%) do not have children.
2. ~20% of applicant have single child.

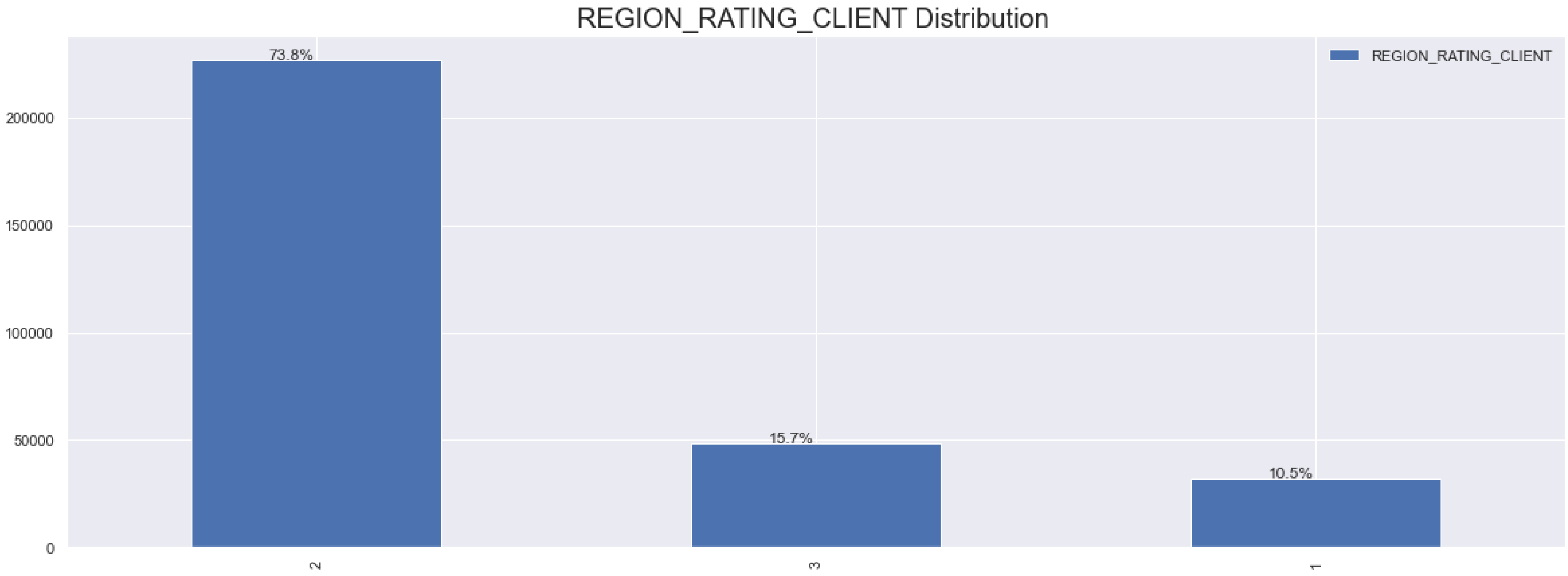
Univariate Analysis (cont...)



INCOME_RANGE:

Major proportion of applicants have income ranging from **1-2 lakhs (50.7%)**

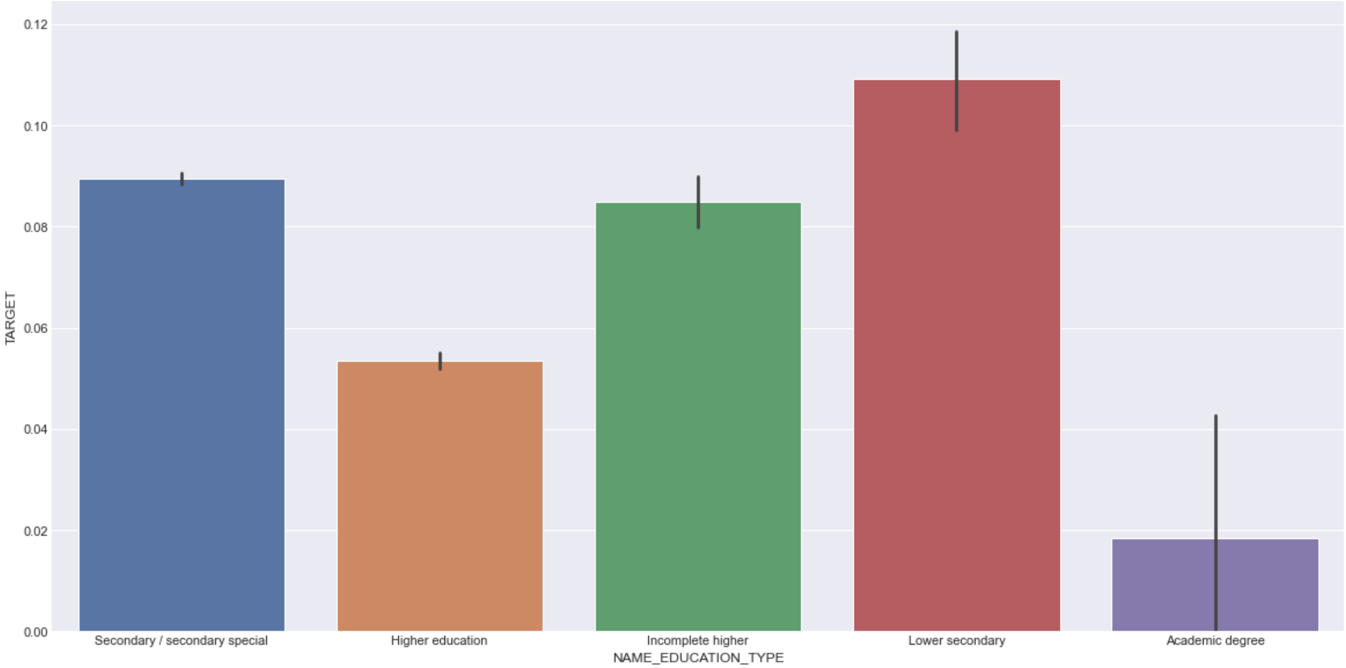
Univariate Analysis (cont...)



REGION_RATING_CLIENT:

Majority of the applicants are residing in regions with rating **2 (73.8%)**.

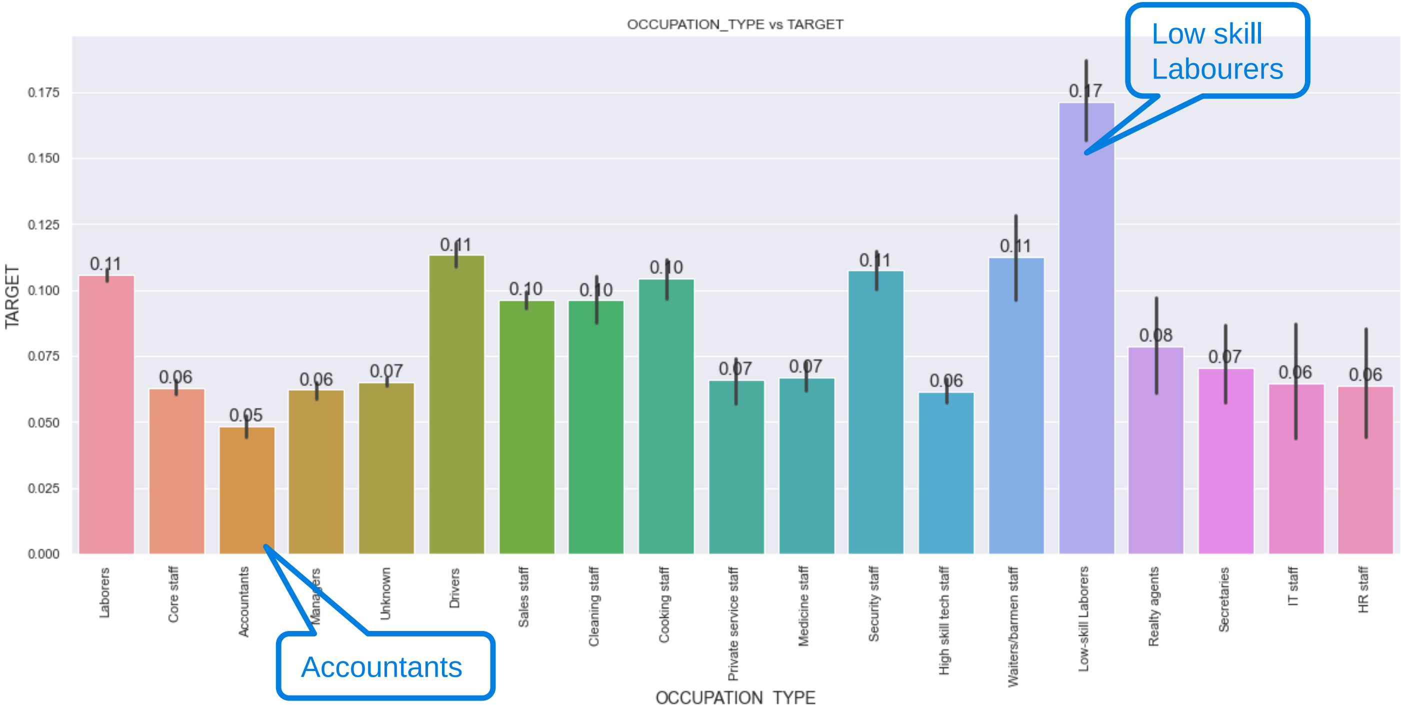
Bivariate Analysis



Target v/s Education type:

A higher percentage of applicants with an education level up to lower secondary level have payment difficulties as compared with others.

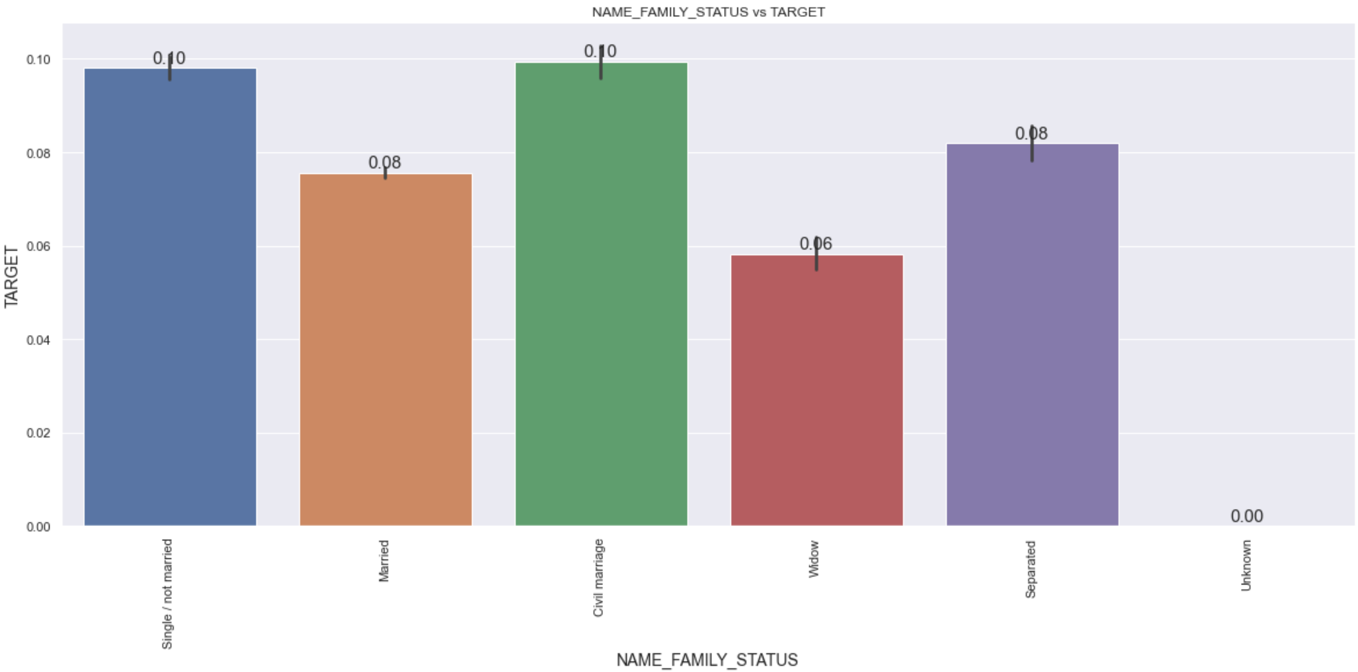
Bivariate Analysis (cont...)



Target v/s Occupation type:

- A higher percentage of applicants who are Low-skill Laborers have payment difficulties as compared with others.
- Lesser number of accountants have payment difficulties as opposed to others occupation types.

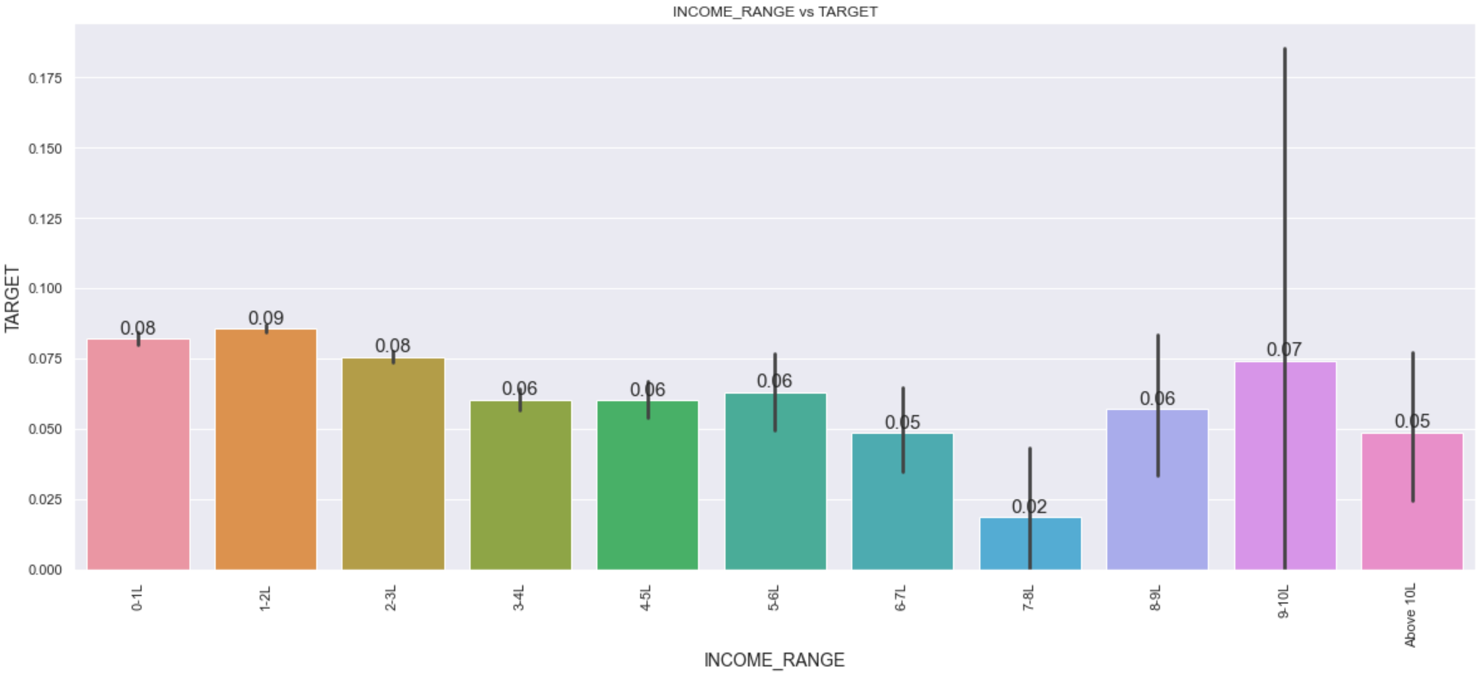
Bivariate Analysis (cont...)



Target v/s Family_Status:

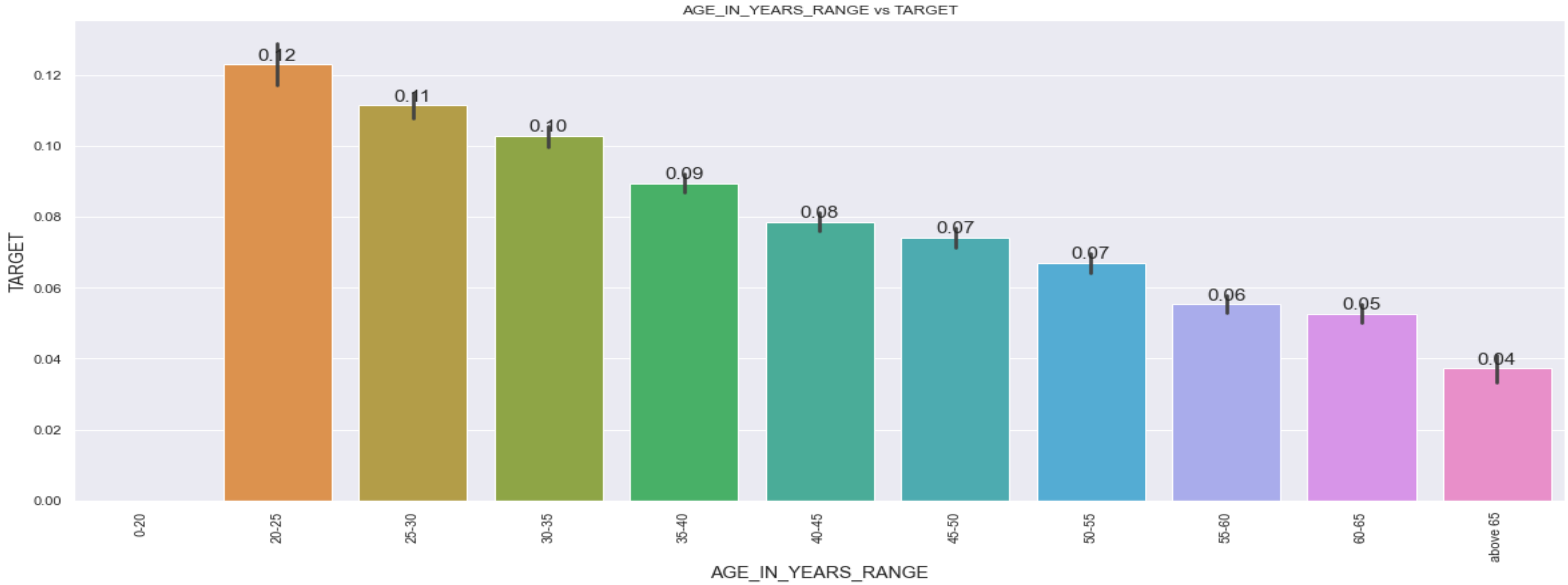
Applicants with family status **Single / unmarried** & **Civil marriage** appear to have more payment difficulties when compared to other family status

Bivariate Analysis (cont...)



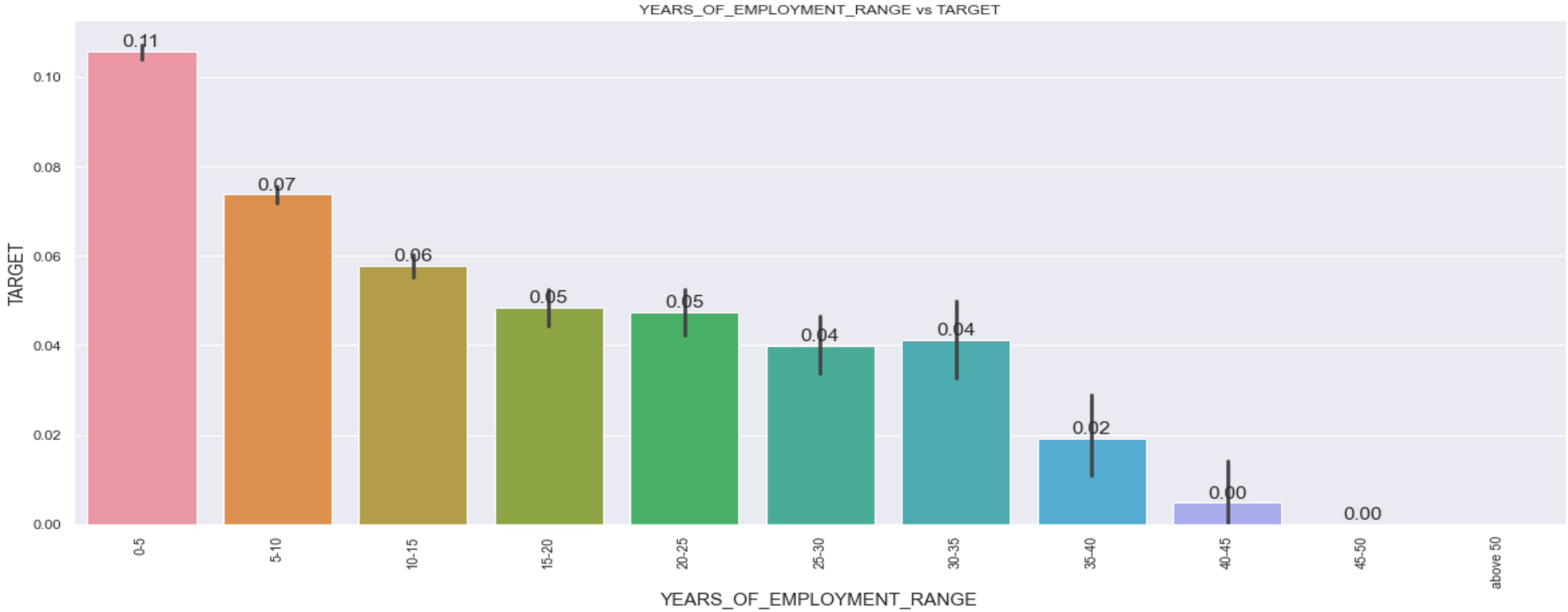
Target v/s Income_Range:
Applicants with annual income ranging from **1-2 lakhs** appear to have slightly more payment difficulties when compared to others followed by **0-1 lakh** & **2-3 lakh.** The least payment difficulties are faced by applicants in the income range **7-8L**.

Bivariate Analysis (cont...)



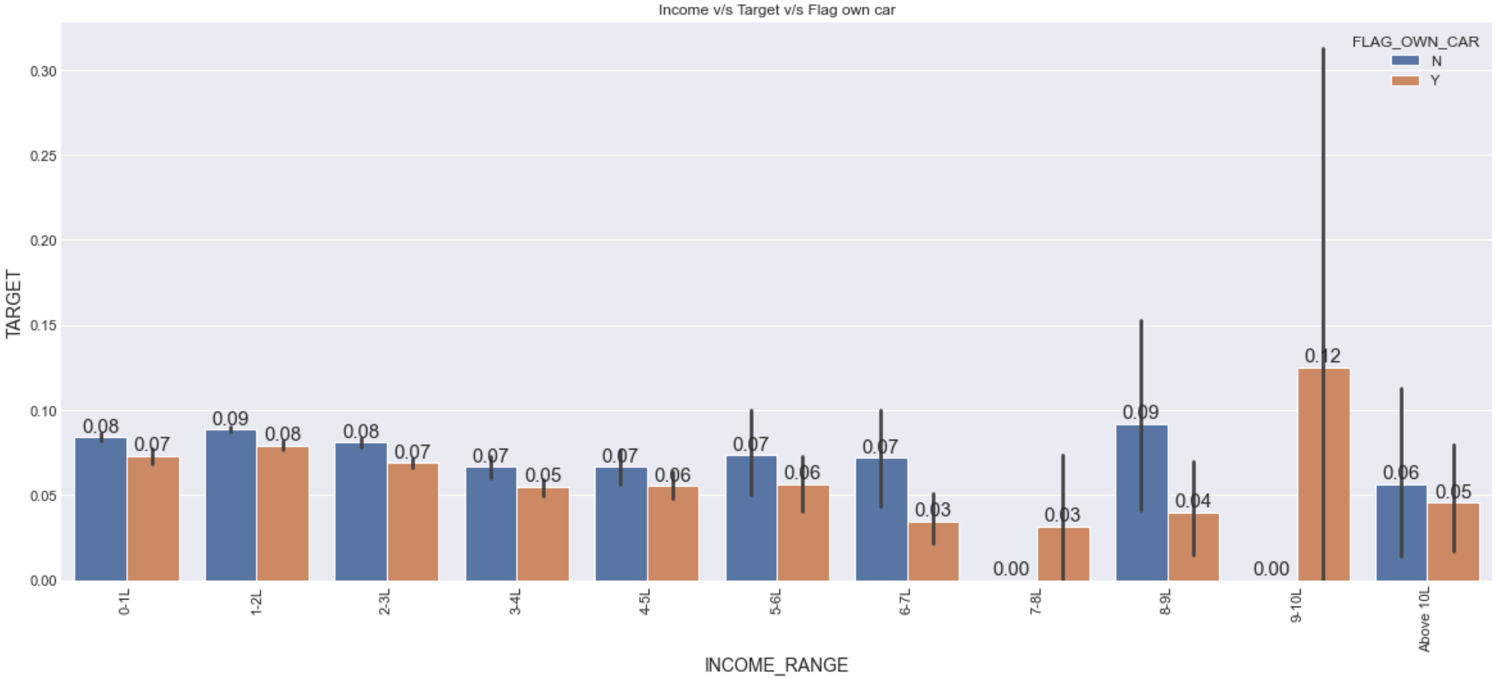
Target v/s Age_in_Years_Range:
Applicants with age ranging from **20-25** seems to have more payment difficulties when compared to others followed by **25-30.**

Bivariate Analysis (cont...)



Target v/s Years_of_Employment_Range:
Applicants in their early careers phase with employment experience ranging from **0-5** years have more payment difficulties when compared to others, followed by **5-10** years.

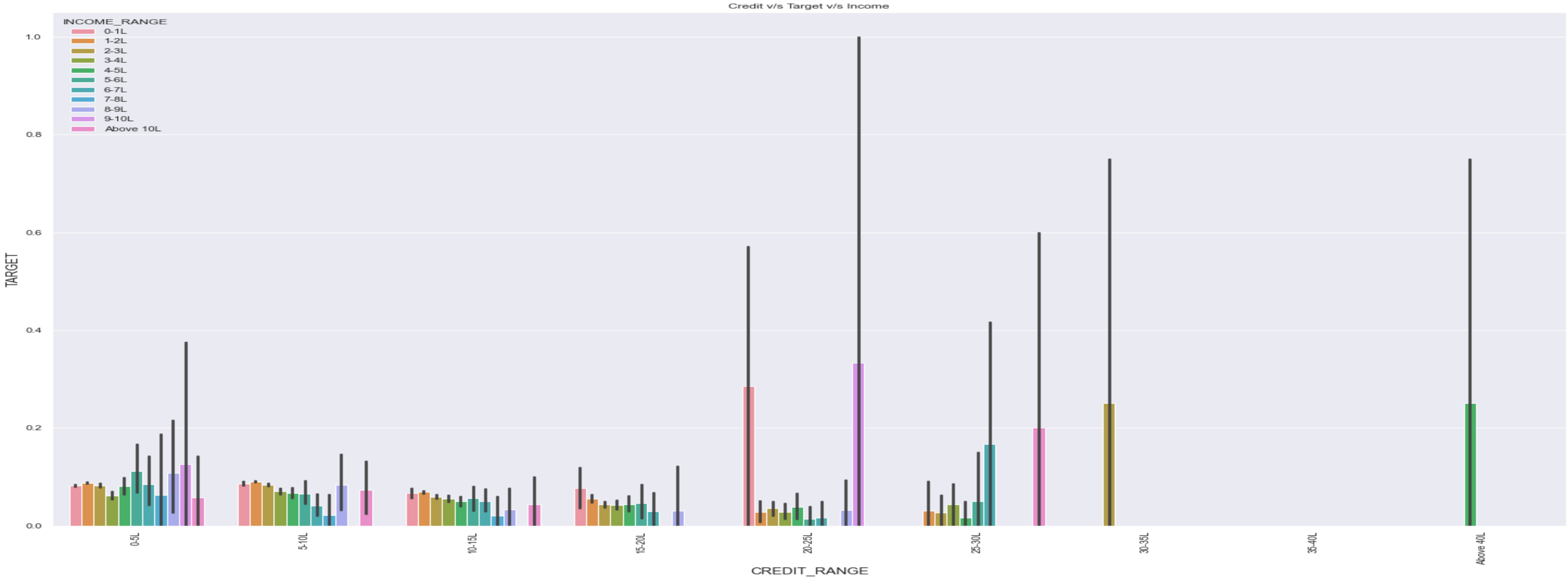
Bivariate Analysis (cont...)



Target v/s Income_Range vs Flag_Own_Car:

1. Applicantsin the income ranges between 0-3L and credit range 0-5L are less likely to have payment difficulties as compared to those in income ranges between 5-6L, 8-9L and 9-10L and in the same credit range.
2. Applicants in the income ranges 0-1L and 9-10L and credit range 20-25L are much more likely to default than those in the same credit range and different income ranges.

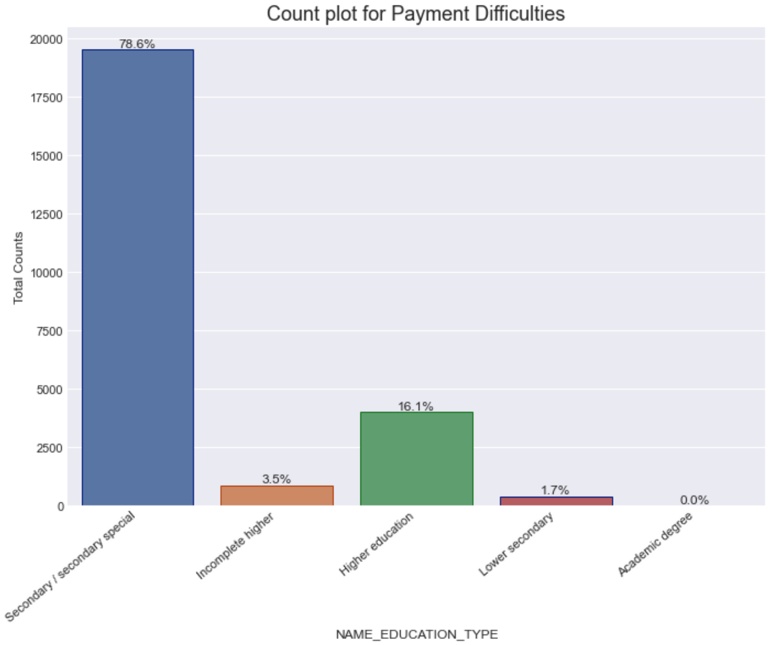
Bivariate Analysis (cont...)

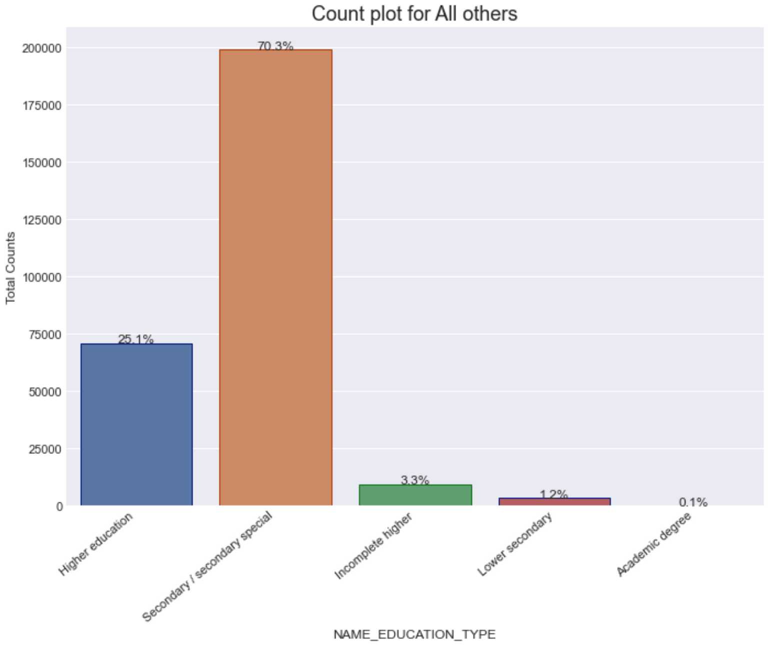


Credit v/s Target v/s Income:

1. Customers in the income ranges between 0-3L and credit range 0-5L are less likely to have payment difficulties as compared to those in income ranges between 5-6L, 8-9L and 9-10L and in the same credit range.
2. Customers in the income ranges 0-1L and 9-10L and credit range 20-25L are much more likely to default than those in the same credit range and different income ranges.

Segmented Univariate Analysis

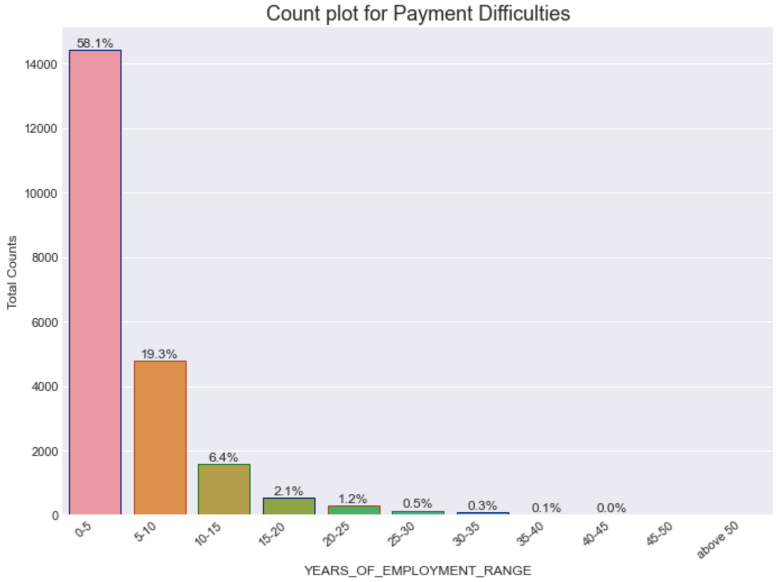


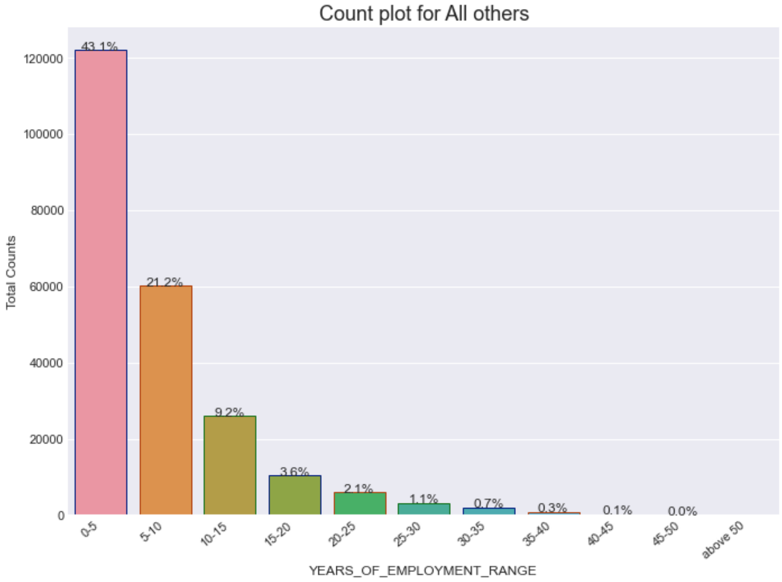


EDUCATION_TYPE:

1. For applicants in both the datasets, the most common level of education is **Secondary / Secondary Special** and the least is **Academic Degree**.
2. The proportion of applicants in the secondary education is slightly higher than in the payment difficulty segment than in all others.

Segmented Univariate Analysis (cont…)

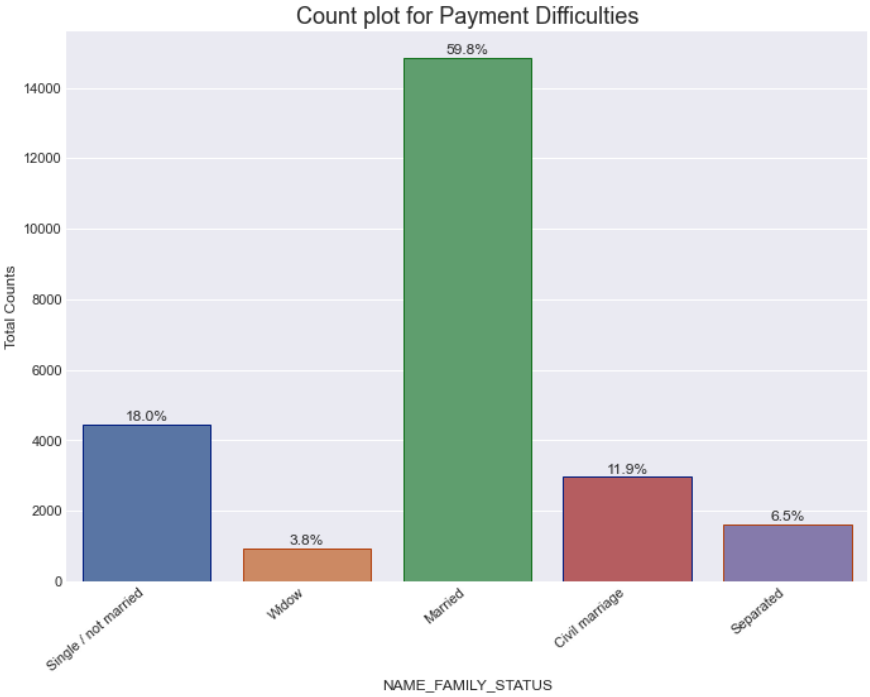


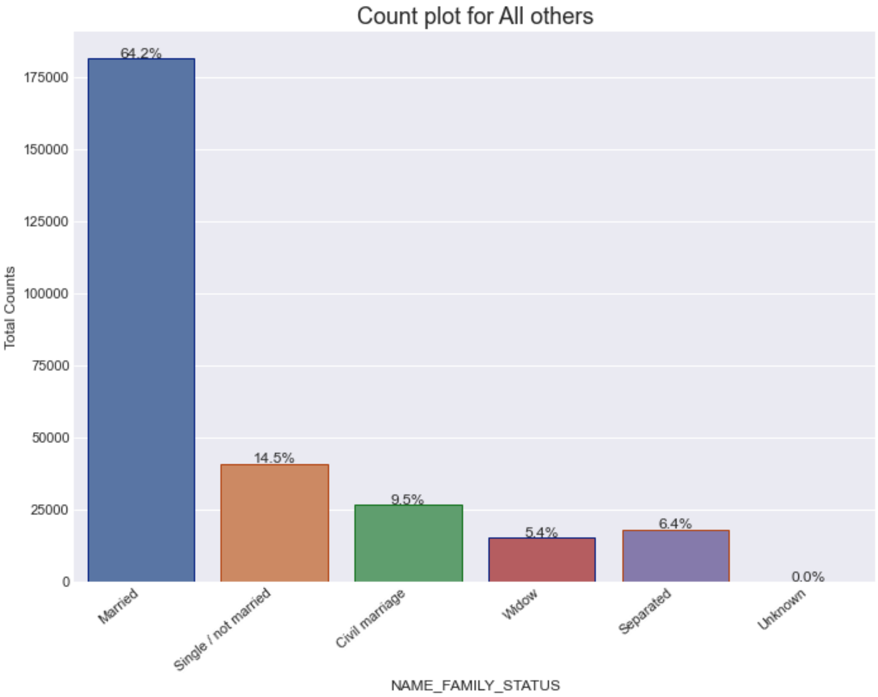


YEARS_OF_EMPLOYMENT_RANGE:

1. For both the segments, most of the applicants have employment experience between **0-5 years**.
2. The proportion of customers in the range 0-5 years is higher in the payment difficulty segment as compared to All others.

Segmented Univariate Analysis (cont...)

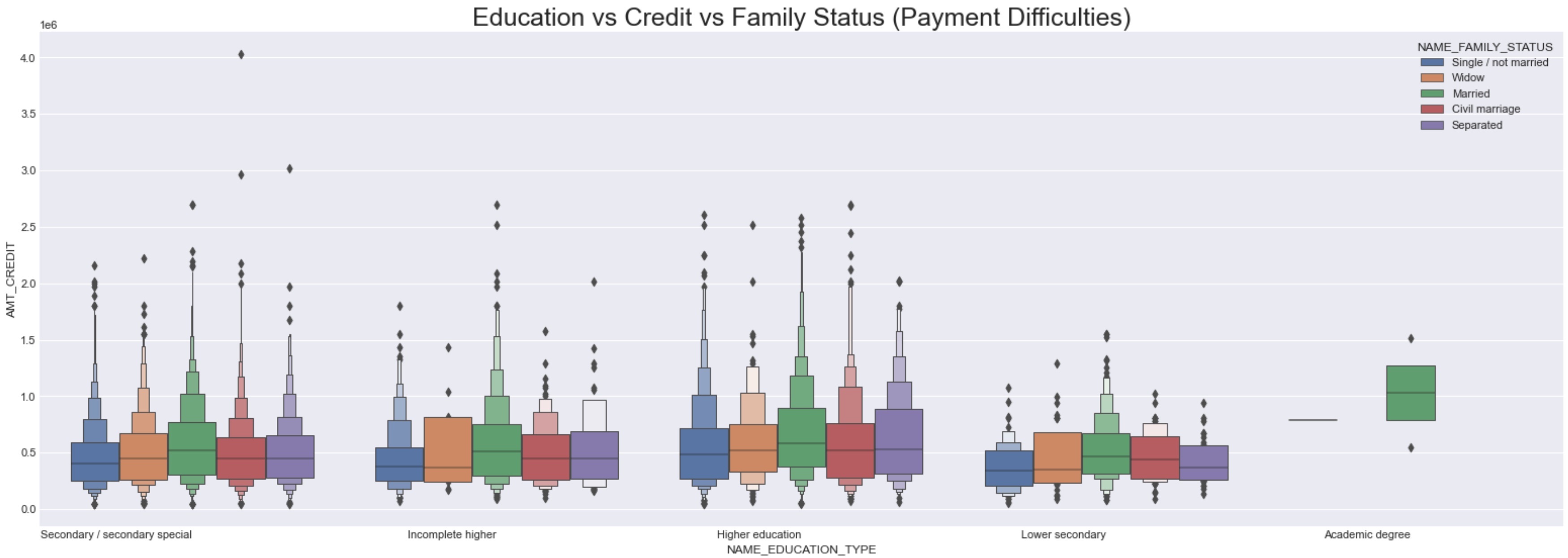




FAMILY_STATUS:

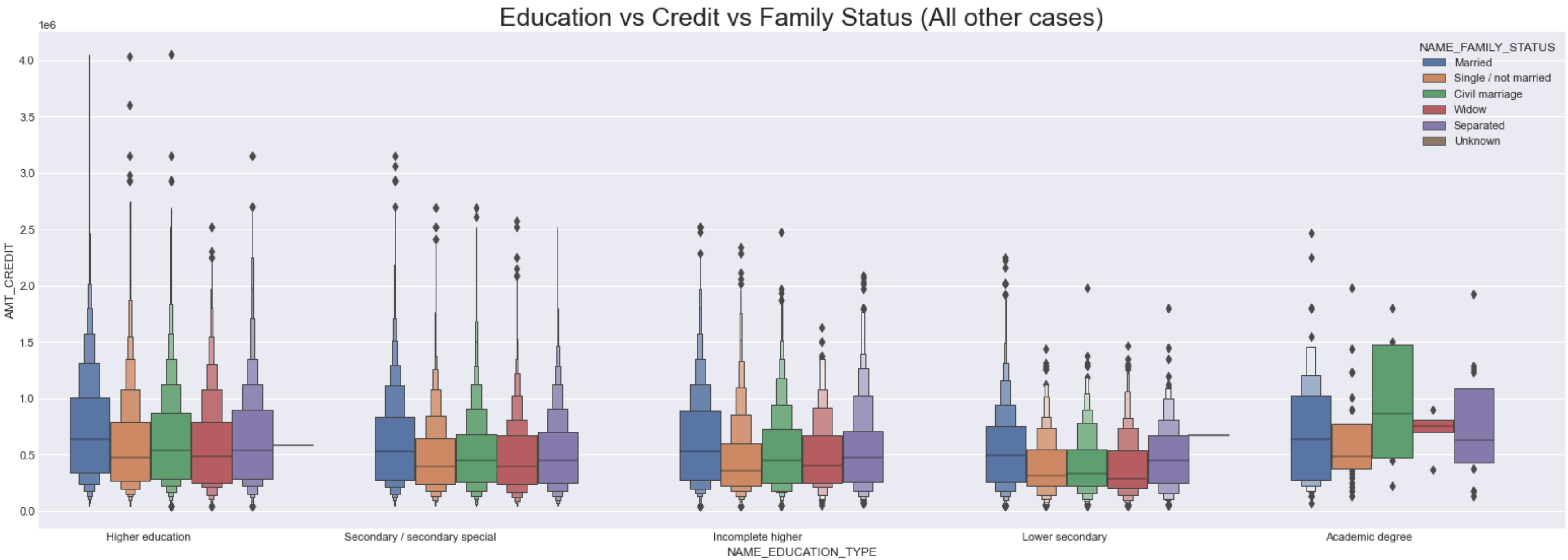
Majority of both the types of applicants are **Married**. They are followed by **Single / Not Married**.

Bivariate Analysis



CREDIT vs Education Type vs family status:
For Applicants with payment difficulties, those with family status - married, have the highest median values across all the education types. So, the median credit amount is highest as compared with others.

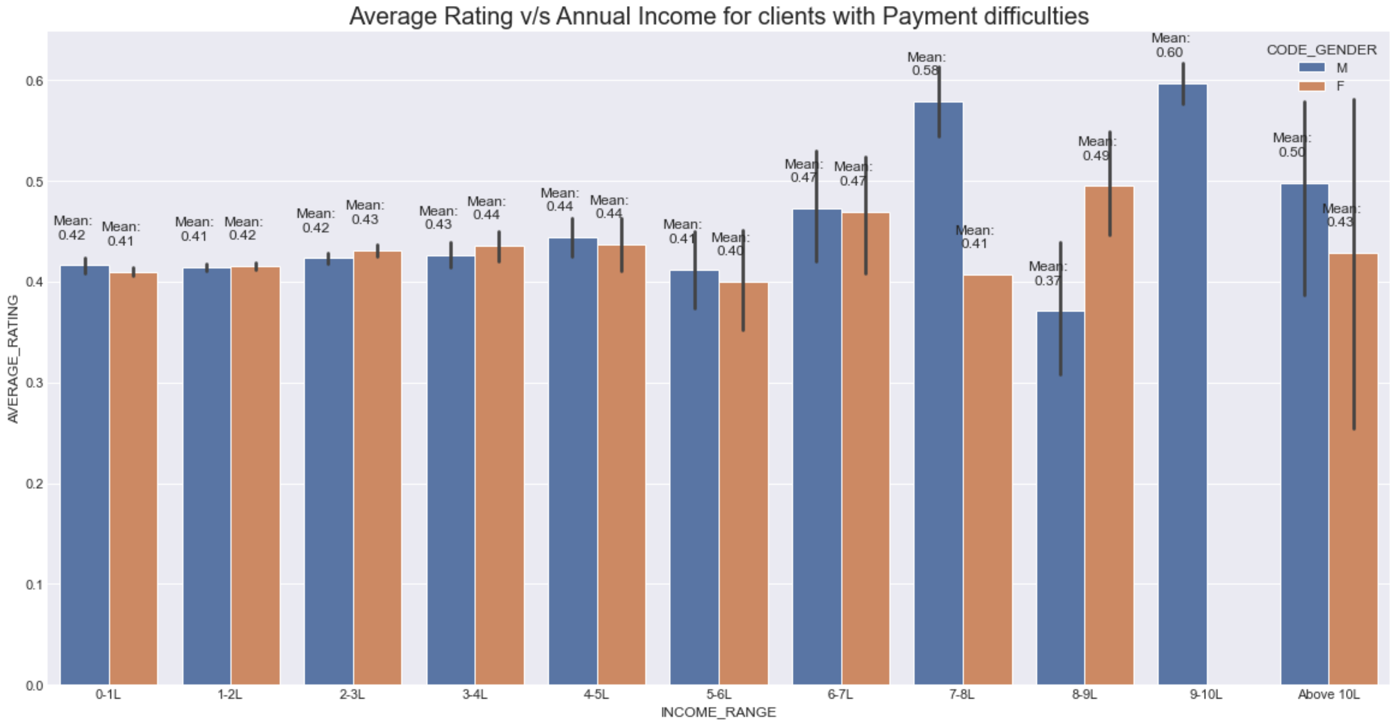
Bivariate Analysis



CREDIT vs Education Type vs family status:

1. For applicants in all the other cases, those with Family status Married have higher Median values across all education types except Academic Degree.
2. Those applicants who have an academic degree and who have had a civil marriage have the highest median value i.e. hence they've requested the highest median amount of credit.

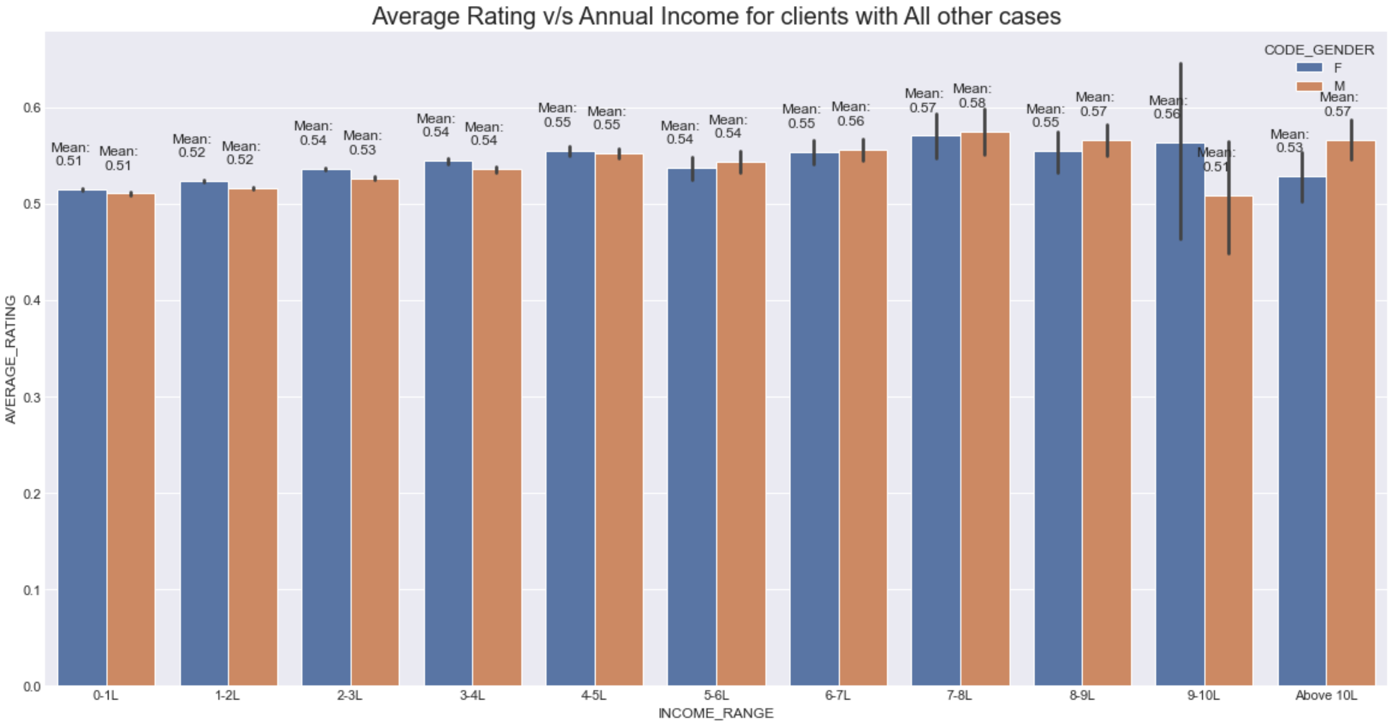
Bivariate Analysis contd..



Ext source_2 v/s Annual Income for Payment with Difficulty:

- For applicants with payment difficulties, the average rating for both the genders is very close for the Income brackets up to 7 Lakhs.
- The number of female applicants in the income bracket 9-10L is the least among all the income ranges; count = 11.

Bivariate Analysis contd..



Average Rating v/s Annual Income for clients with All other cases

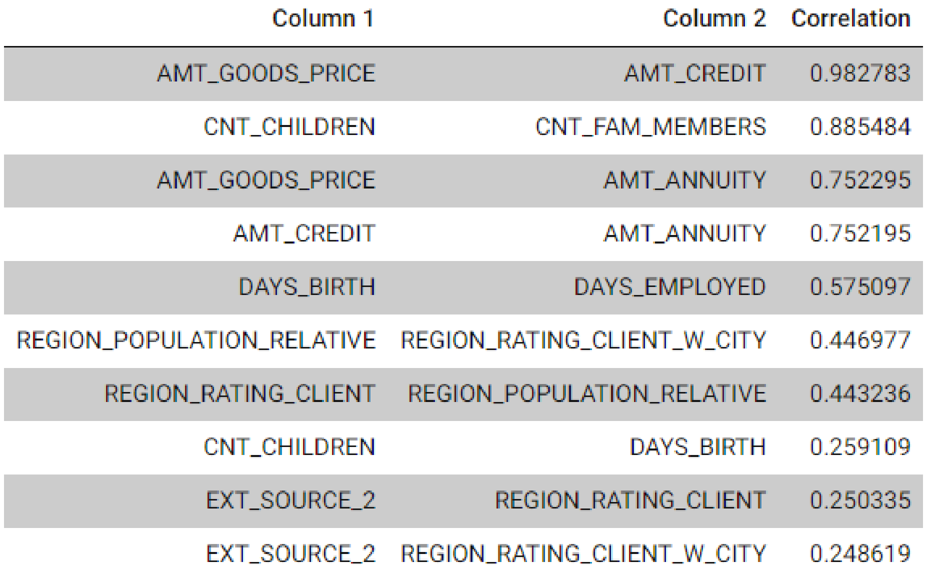
Ext source_2 v/s Annual Income for All other cases:

The average rating for applicants is higher in 'Clients with All Other cases' than the clients with Payment difficulties irrespective of gender and the income brackets.

.

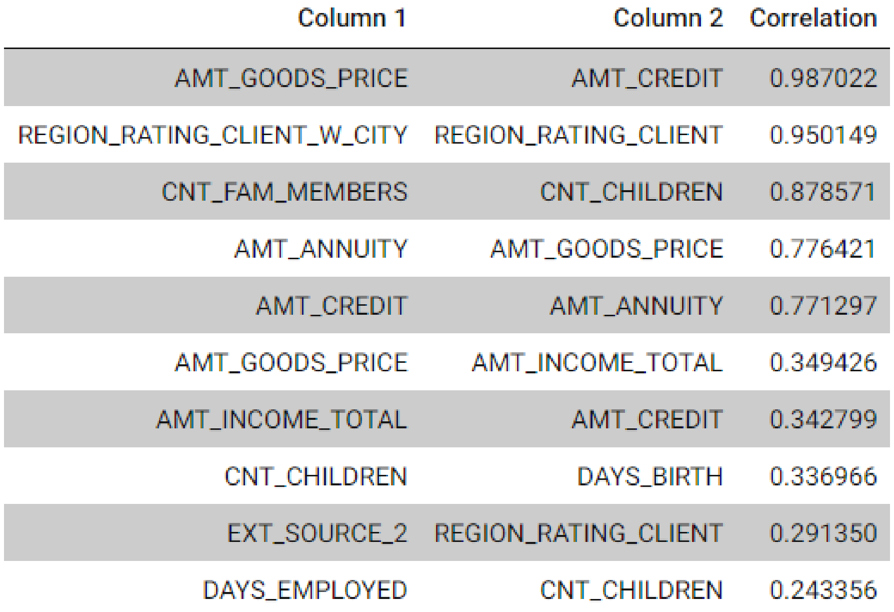
Top 10 Correlations

Payment Difficulties



Column 1	Column 2	Correlation	
AMT_GOODS_PRICE	AMT_CREDIT	0.982783	1
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484	2
AMT_GOODS_PRICE	AMT_ANNUITY	0.752295	3
AMT_CREDIT	AMT_ANNUITY	0.752195	4
DAYS_BIRTH	DAYS_EMPLOYED	0.575097	5
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT_W_CITY	0.446977	6
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.443236	7
CNT_CHILDREN	DAYS_BIRTH	0.259109	8
EXT_SOURCE_2	REGION_RATING_CLIENT	0.250335	9
EXT_SOURCE_2	REGION_RATING_CLIENT_W_CITY	0.248619	10

All Others

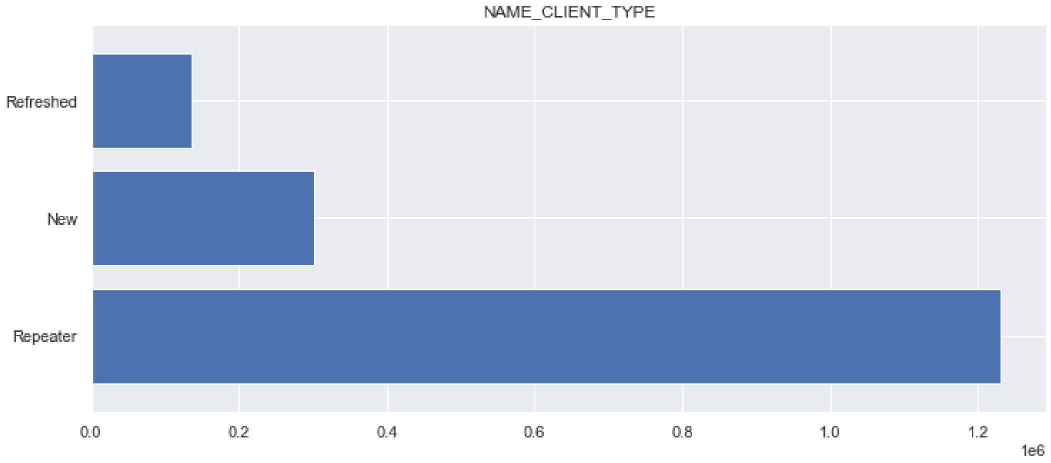


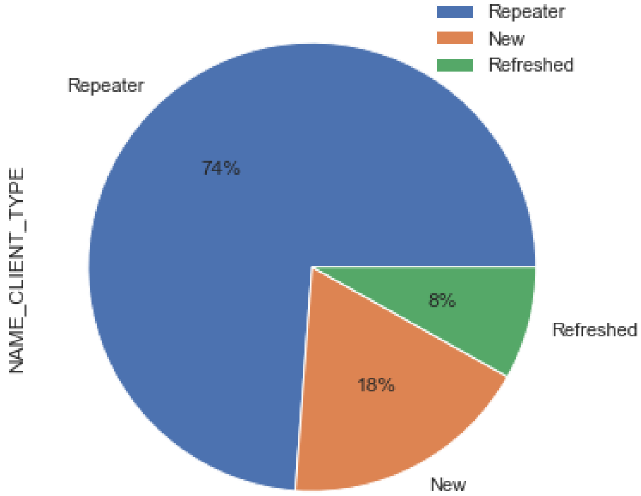
Column 1	Column 2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.987022
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
AMT_ANNUITY	AMT_GOODS_PRICE	0.776421
AMT_CREDIT	AMT_ANNUITY	0.771297
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349426
AMT_INCOME_TOTAL	AMT_CREDIT	0.342799
CNT_CHILDREN	DAYS_BIRTH	0.336966
EXT_SOURCE_2	REGION_RATING_CLIENT	0.291350
DAYS_EMPLOYED	CNT_CHILDREN	0.243356

Top 10 Correlations

- The highest correlation between 2 variables for customers with payment difficulties and the rest is goods price amount and credit amount.
- The credit score (from Ext source 2) of an applicant has a strong correlation with the rating of the region where the client lives. It has a slightly stronger correlation in those who don't have payment difficulties (0.29) as opposed to those who have payment difficulties (0.25).
- Goods price amount and total income are more strongly correlated in all the other cases as compared to payment difficulties. This could mean that those who don't have payment difficulties evaluate their income and the goods they want to buy in a better way.
- The Region_rating_client and Region_rating_client_w_city are directly proportional to each other and have higher positive correlation for customers without payment difficulties as opposed to customers with payment difficulties

Univariate – Previous Application

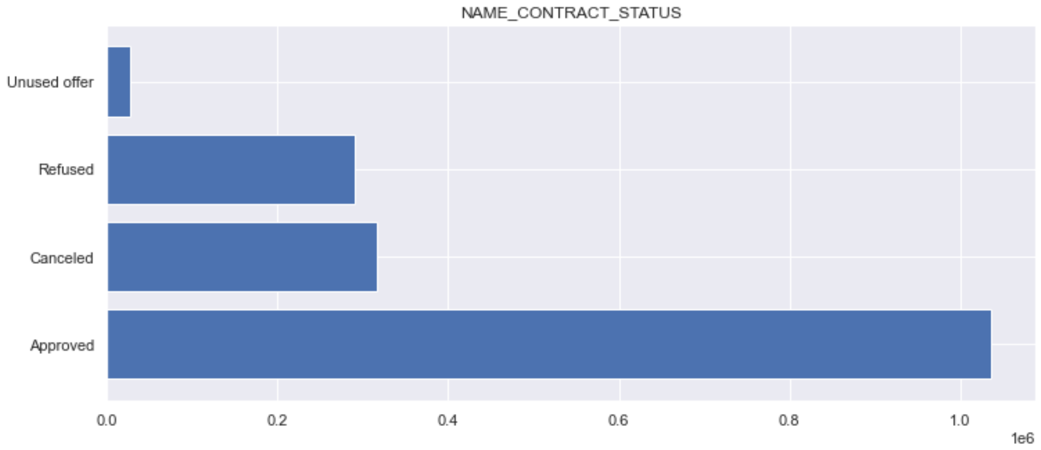


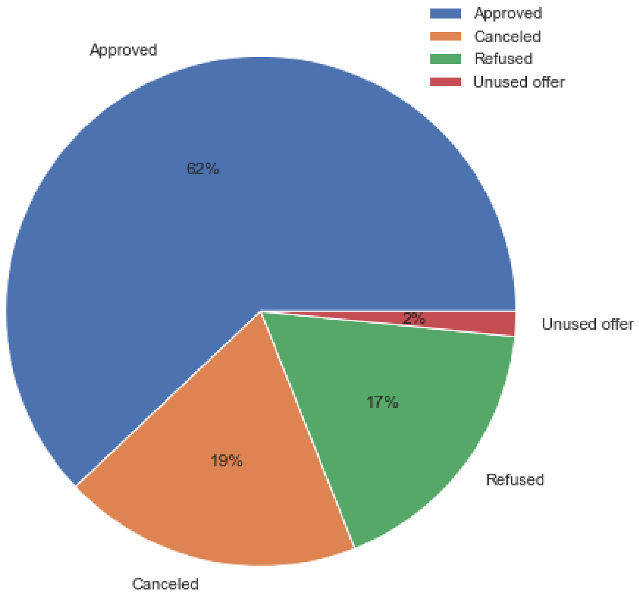


NAME_CLIENT_TYPE:

The maximum percentage of applicants are repeaters with a 74% share followed by New Applicants.

Univariate Contd..

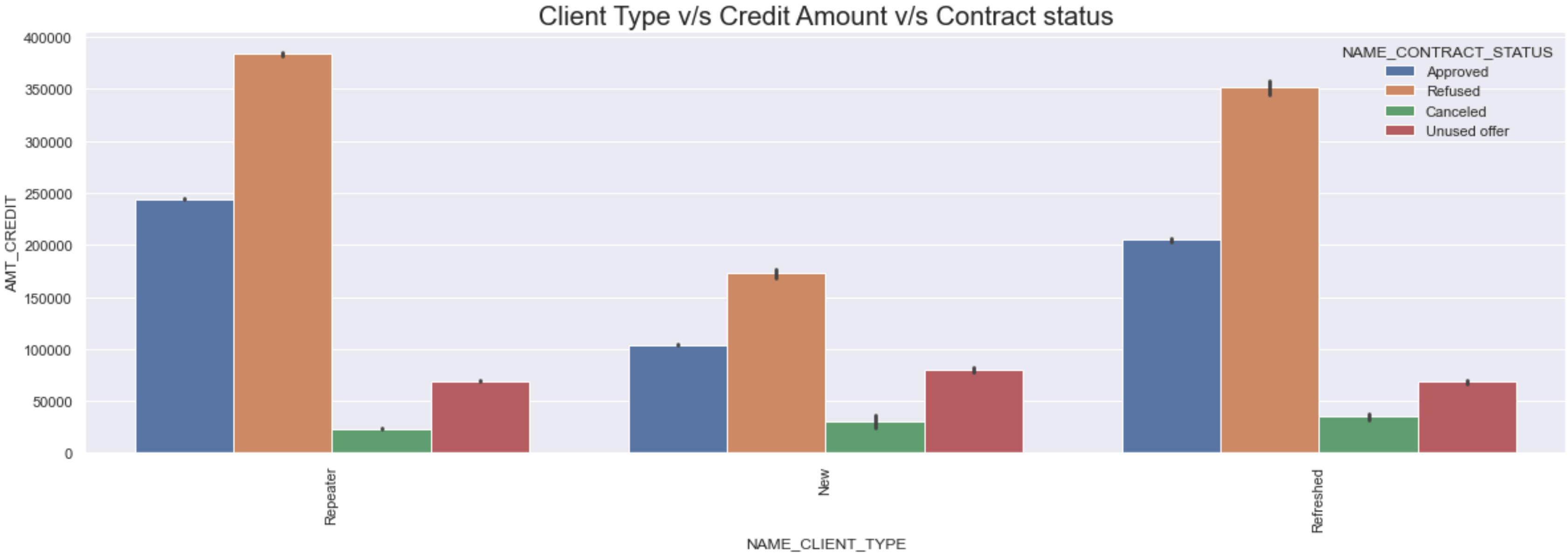




NAME_CONTRACT_STATUS:

Majority of the previous applications have been approved. This is the followed by the proportion of applicants who cancelled their applications sometime during approval.

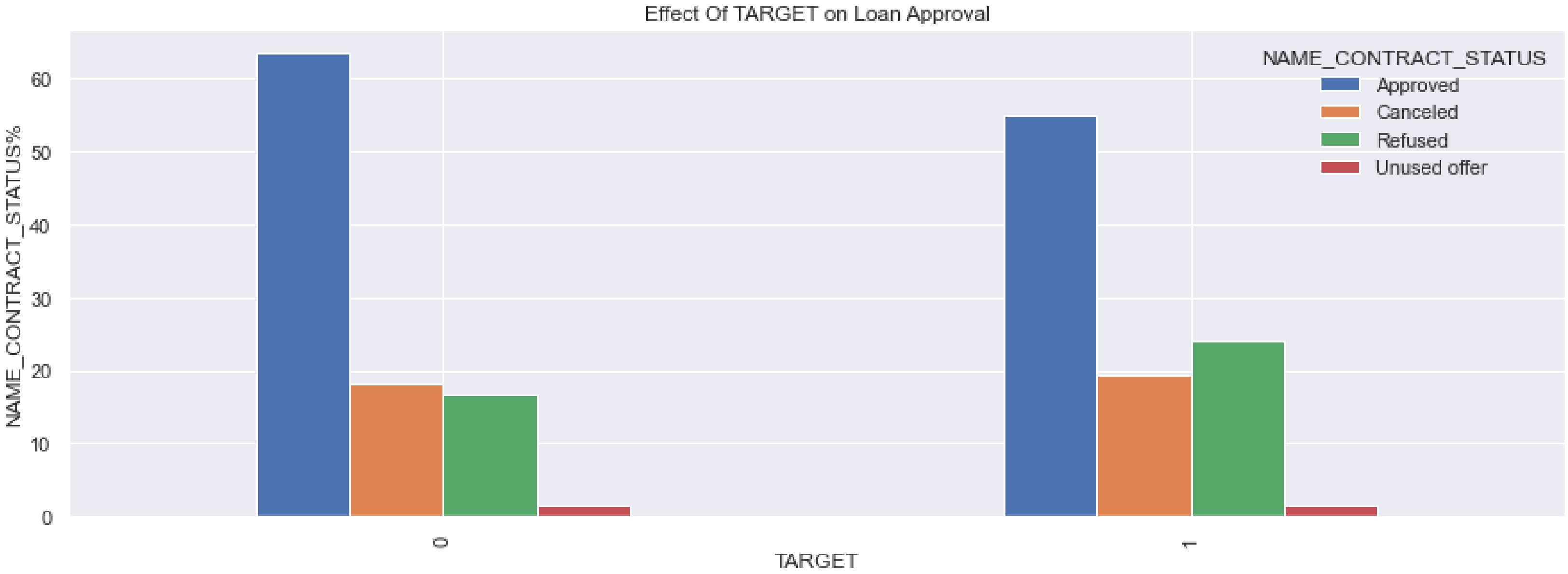
Multivariate– Previous Application



Client_Type v/s Credit Amount v/s Contract status:

- Among the approved and refused loans of all types of customers, repeat customers had the highest average amount of credit.
- We observe that the new customers had the least average amount of credit, among all the approved and refused loans of all types of customers.

Bivariate– Merged Dataset



Target v/s Name_Contract_Status:

- The percentage of customers who have payment difficulties who have had their loans approved is lower than the percentage of those who fall in all the other cases.

5

Driver variables for identifying Default

Main Driver variables

1. *Education Type* – Applicants with an academic degree are least likely to default while those with an education level up to lower secondary are most likely to default.
2. Occupation – Applicants who are low-skill labourers are most likely to default while those who are accountants are least likely to default.
3. *Age* – Applicants in the age bracket 20-25 are likely to have the most payment difficulties while those above 65 are likely to have the least.
4. *Employment experience* – Those with 0-5 years of experience are most likely to default while those with 40-45 years of experience are least likely.
5. *Type of housing* – Those with an office apartment are the least likely to default while those with a rented apartment or those living with their parents are most likely to default.
6. *Income Range* – Those with an income ranging between 1-2L are most likely to default while those in the income range 7-8L are least likely to default.

Thanks!

Email:

vagmi2009@gmail.com

rehman.me02@gmail.com