EXPLORATORY DATA ANALYSIS
HOME CREDIT DEFAULT RISK

Tran Kha Uyen   Nguyen Thi Kieu Nhung   Chu Duc Trung
Nguyen Ngoc Bang Anh   Nguyen Hoang Tu

Data Preparation and Visualization Project

Instructor : Dr. Nguyen Thi Quynh Giang

November 16, 2022

Presentation Overview

Introduction about case study

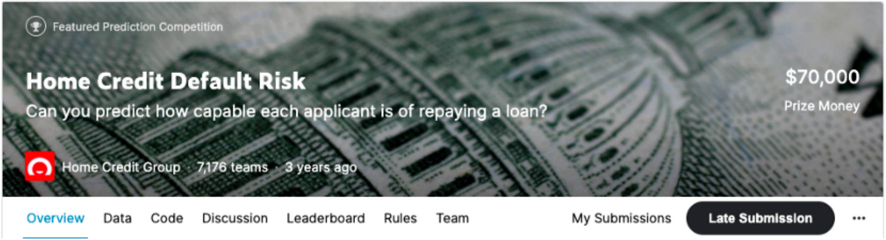
Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders for their financial needs, and are at the risk of being taken advantage of, mostly with unreasonably high rates of interest.

Introduction about case study (contd.)

In order to address this issue, 'Home Credit' uses a lot of data (including both Telco Data as well as Transactional Data) to predict the loan repayment abilities of the applicants. If an applicant is deemed fit to repay a loan, his application is accepted, and it is rejected otherwise. This will ensure that the applicants having the capability of loan repayment do not have their applications rejected.

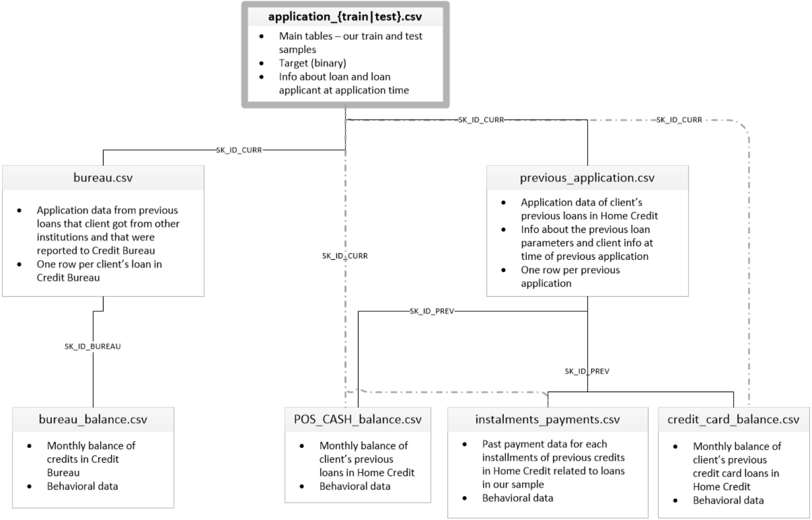
Source of Data
Home Credit Default Risk

Source:
https://www.kaggle.com/c/home-credit-default-risk



Home Credit Group provided a large dataset to motivate machine learning engineers and researchers to come up with techniques to build a predictive model for analyzing and estimating the risk associated with a given borrower through a Kaggle competition.

The Dataset Schema & Description



The Dataset Schema & Description (contd.)

application_train/test.csv

- This is the main table, broken into two files for Train (with TARGET) (ie. the prediction provided) and Test (without TARGET).
- Static data for all applications. One row represents one loan in our data sample.

The Dataset Schema & Description (contd.)

bureau.csv

- All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
- For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.

The Dataset Schema & Description (contd.)

POS_CASH_balance.csv

- Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
- This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample

The Dataset Schema & Description (contd.)

credit_card_balance.csv

- Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
- This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample

previous_application.csv

- All previous applications for Home Credit loans of clients who have loans in our sample.
- There is one row for each previous application related to loans in our data sample.

The Dataset Schema & Description (contd.)

installments_payments.csv

- Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
- There is one row for every payment that was made plus one row each for missed payment.
- One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.

Univariate Analysis : Name_Contract_Type

Most of the people are taking loans in the form of cash loans instead of revolving loans such as credit cards.

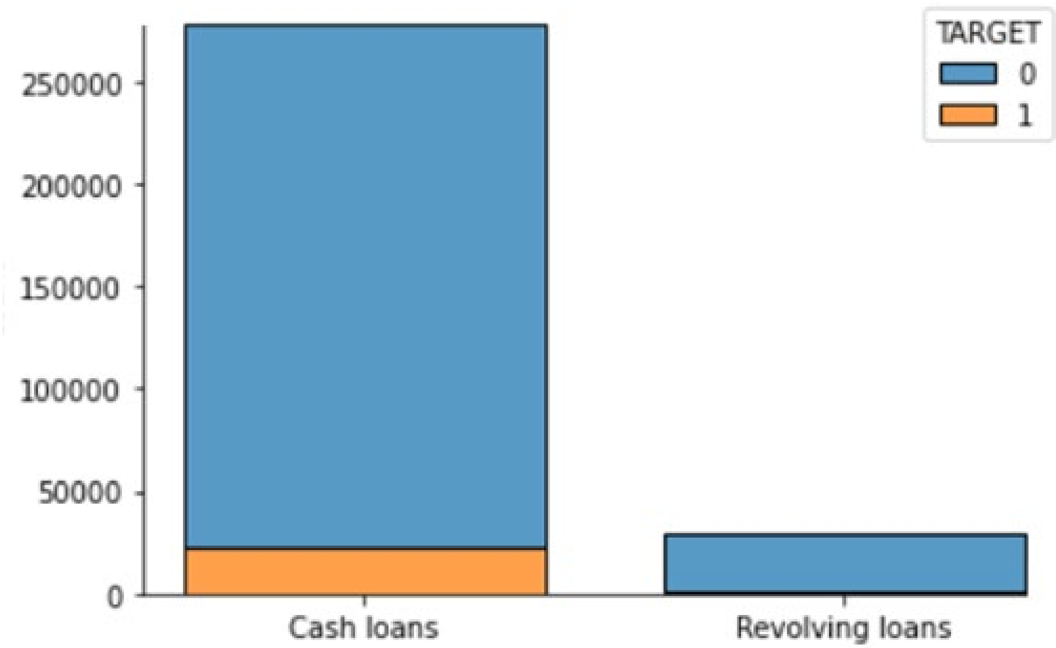
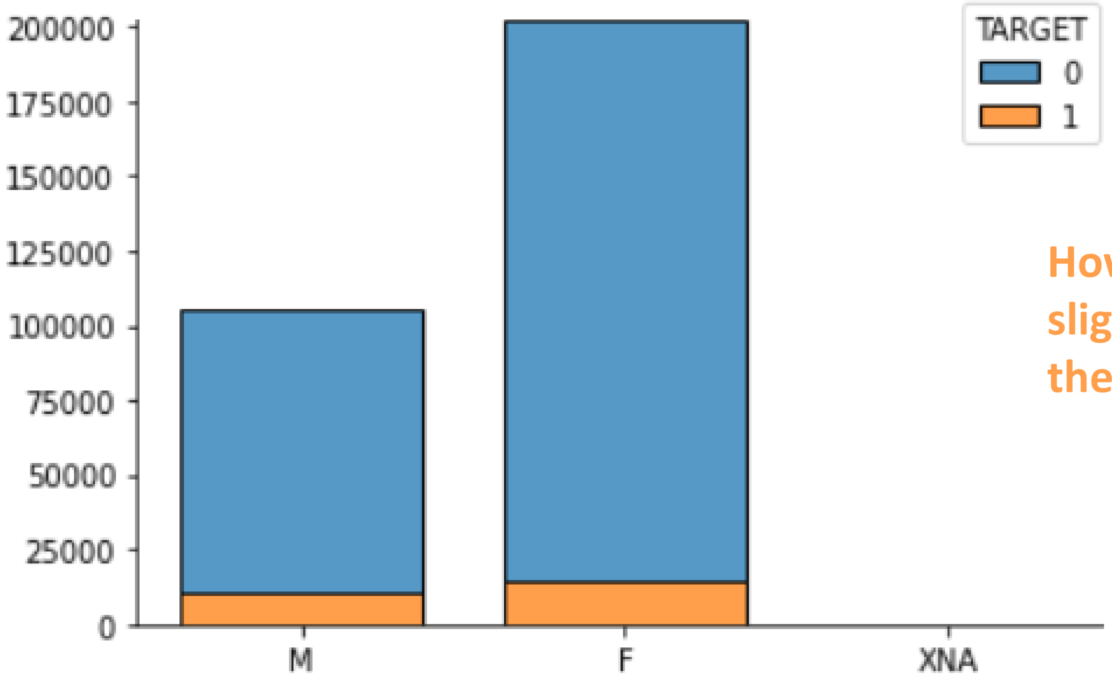


FIGURE 1 Number of loans approved vs rejected according to contract type

Univariate Analysis : CODE_GENDER

Female took much more number of loans as compared to men

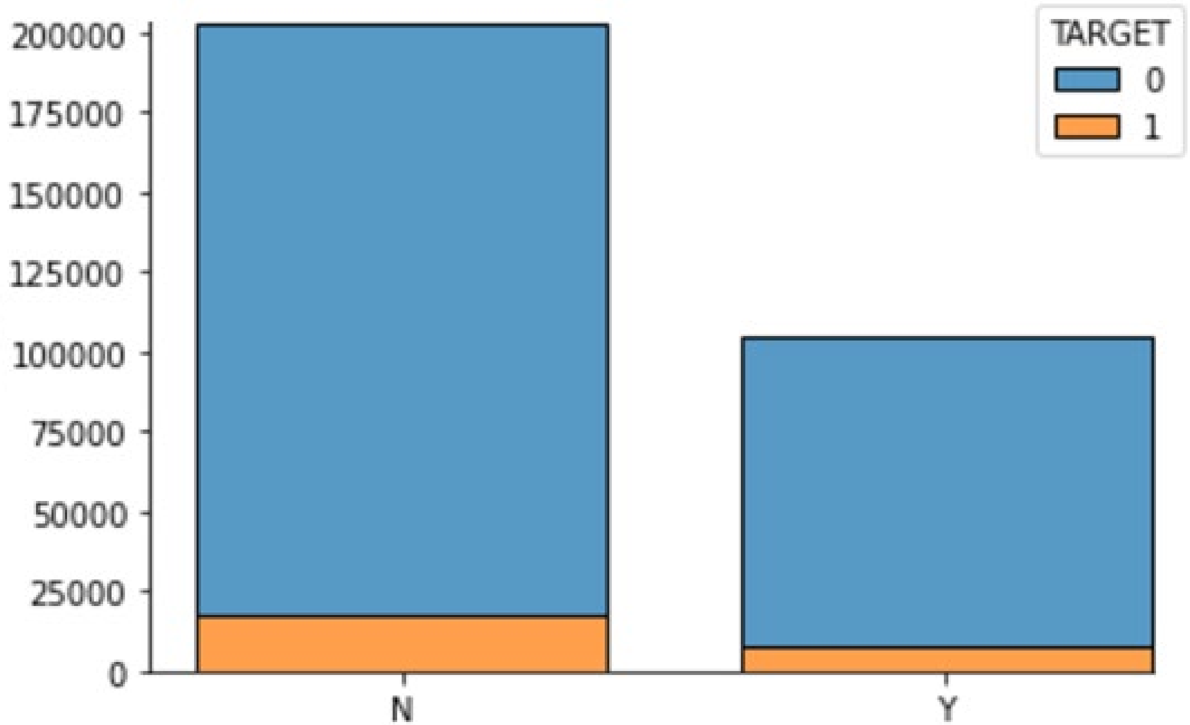


However, at the same time, Men are slightly more capable of repaying the loan as compared to female

FIGURE 2 Number of loans approved vs rejected according to gender

Univariate Analysis : FLAG_OWN_CAR

Most of the applicants for loans do not own a car.



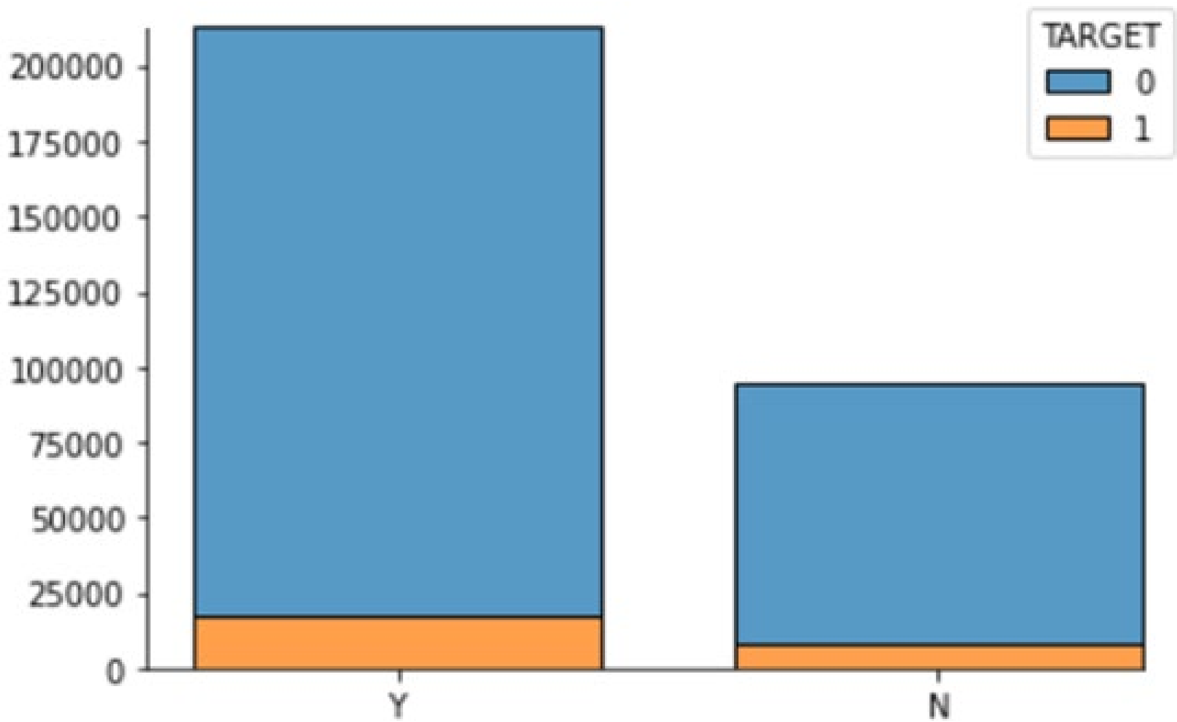
However, there is not much difference in the loan repayment status for the customer based on this information

FIGURE 3 Number of loans approved vs rejected according to own car

Exploratory Data Analysis
Application Train

Univariate Analysis : FLAG_OWN_REALTY

Most of the applicants for loans own a house, which is a little surprising.



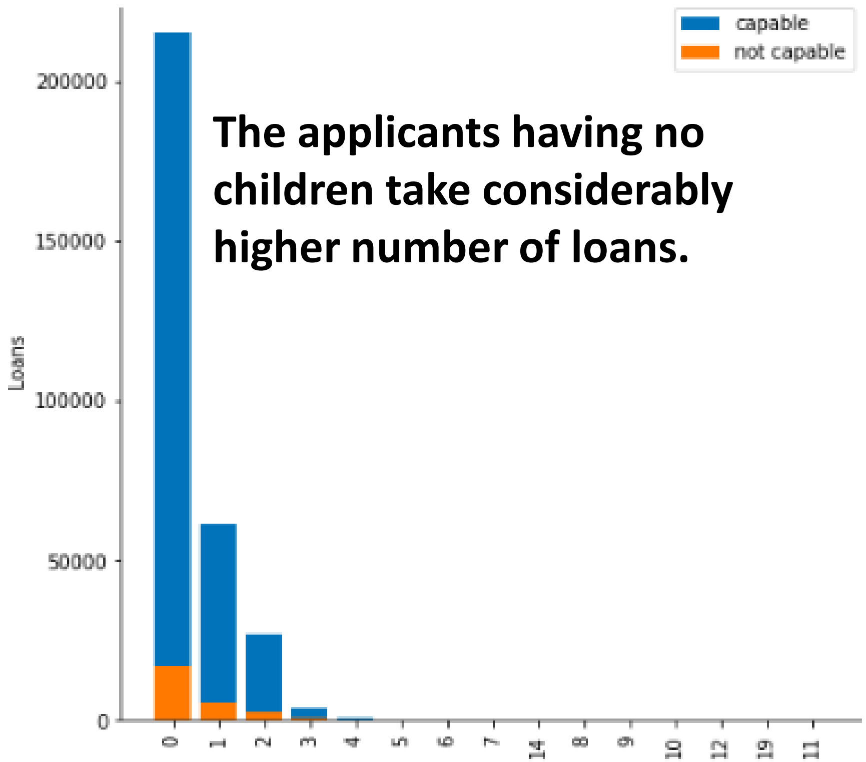
However, there is not much difference in the loan repayment status for the customer based on this information

FIGURE 4 Number of loans approved vs rejected according to own realty

Exploratory Data Analysis

Application Train

Univariate Analysis : Count of Children

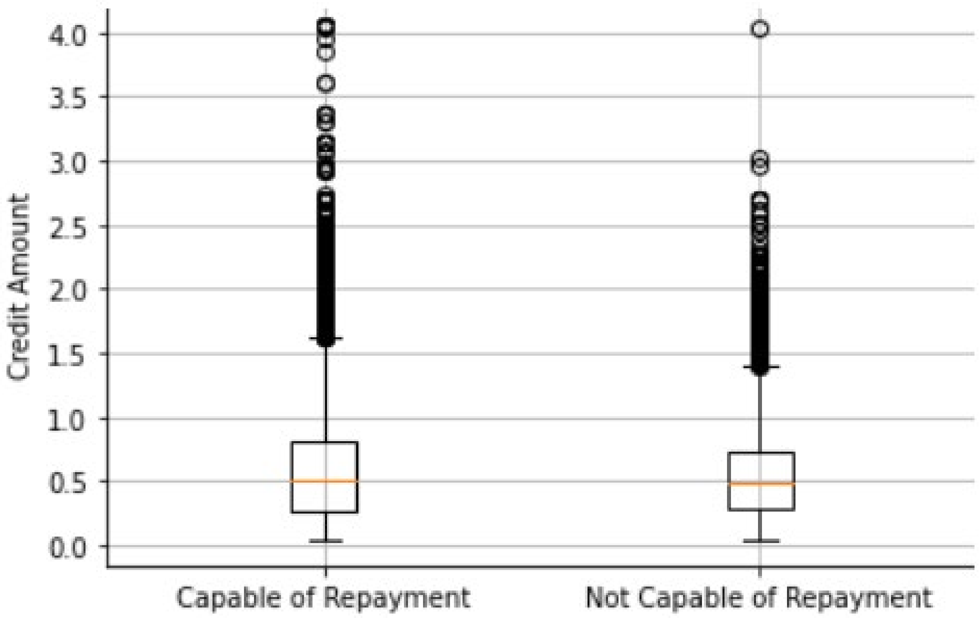


The applicants having no children take considerably higher number of loans.

However, again, there is not much difference in the loan repayment status for the customer based on this information.

FIGURE 5 Number of loans approved vs rejected according to number of children

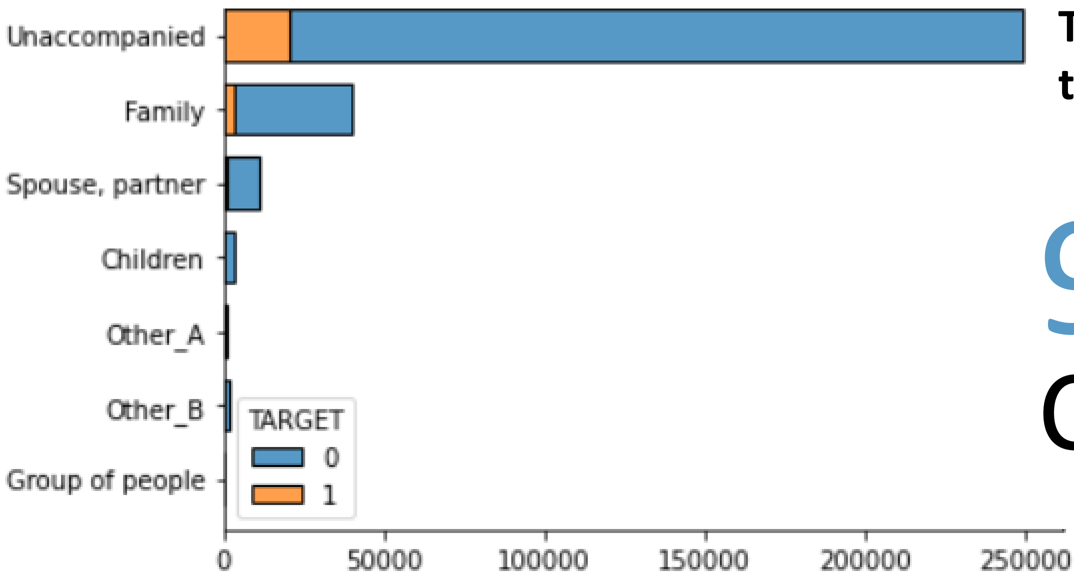
Univariate Analysis : Amt_credit



The customers with higher credit amount have a slightly higher chances of being capable of loan repayment than customers with lower credit amount

FIGURE 6 Credit amount for each loan

Univariate Analysis : NAME_TYPE_SUITE



The client comes unaccompanied to the bank in the most number of cases:

92%

CAPABLE

FIGURE 7 Number of loans approved vs rejected according to the various types of people accompanying the client for loan

Univariate Analysis : NAME_INCOME_TYPE

The people who are working take the most number of loans whereas Commercial Associates, Pensioners and State Servants take considerably lesser number of loans.

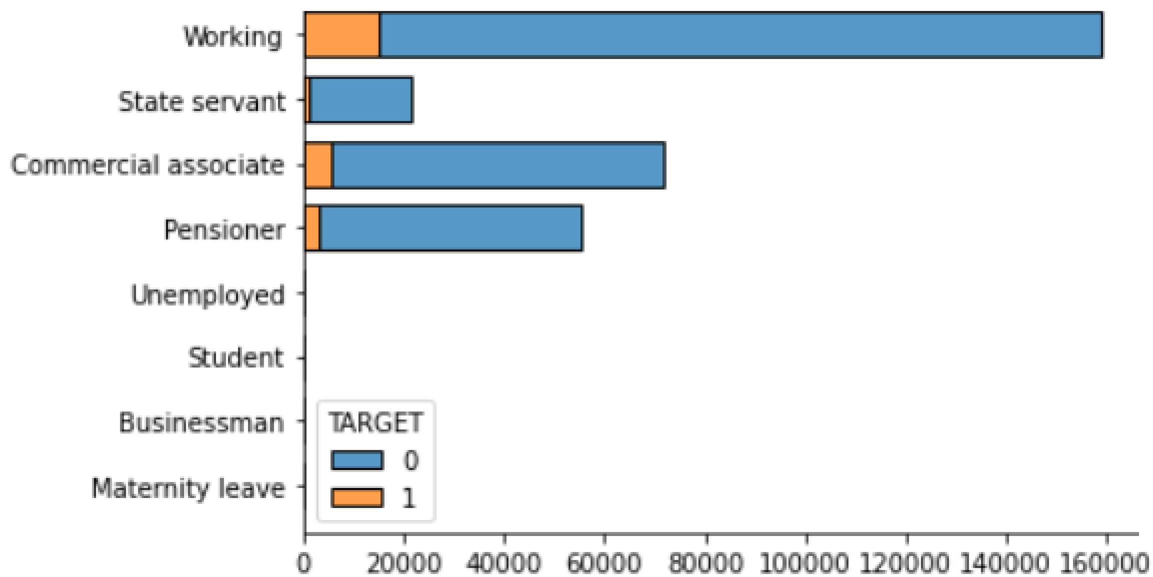


FIGURE 8 Number of loans approved vs rejected according to the various types of income

Univariate Analysis : FAMILY_STATUS

Married people apply for the most number of loans and the number of people deemed incapable of repayment is also the highest.

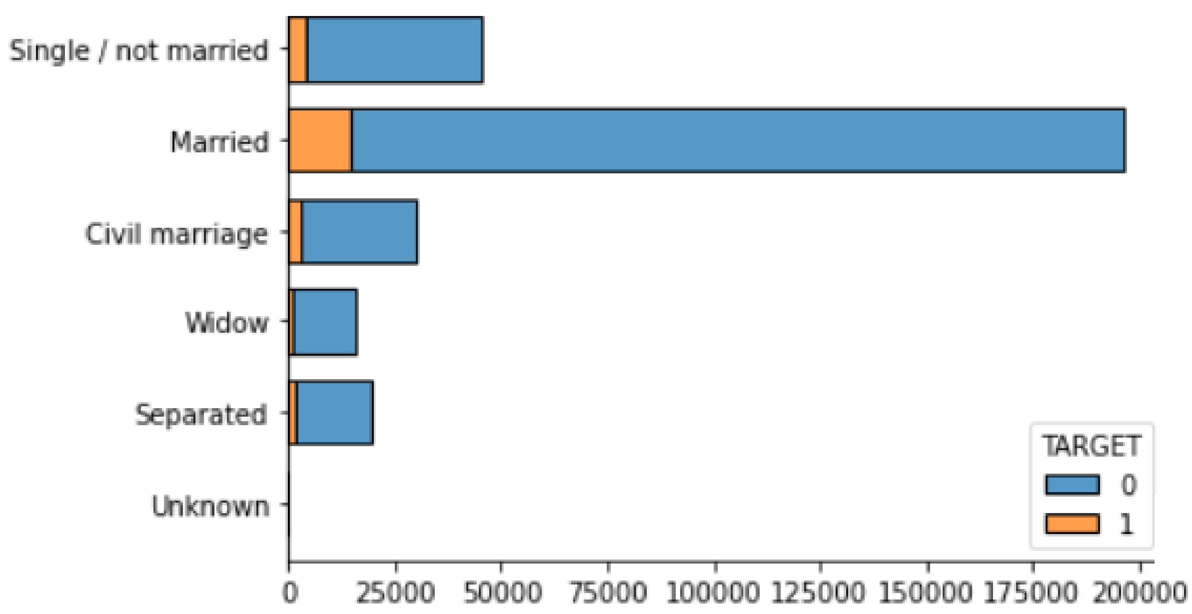


FIGURE 9 Number of loans approved vs rejected according to Family status

Univariate Analysis : NAME_HOUSING_TYPE

Most of the applications in the bureau_data is closed, following by the status of being Active

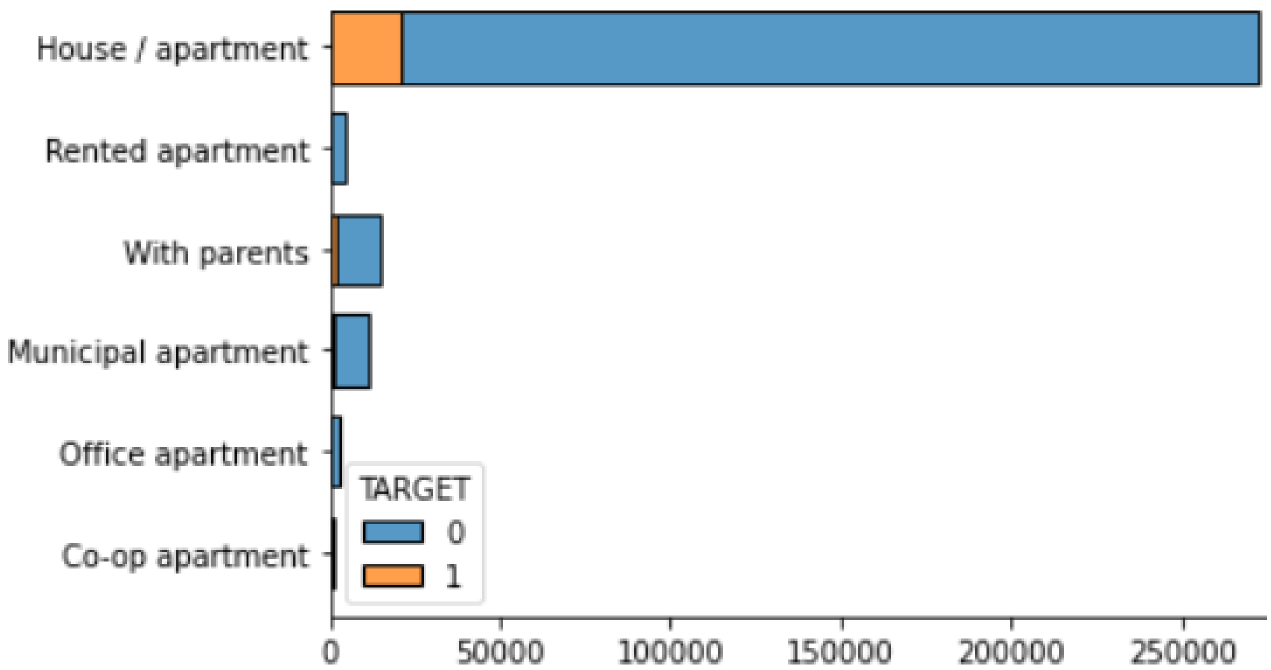
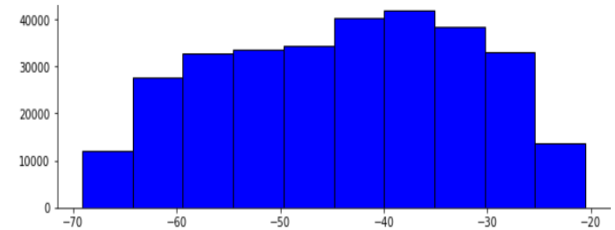


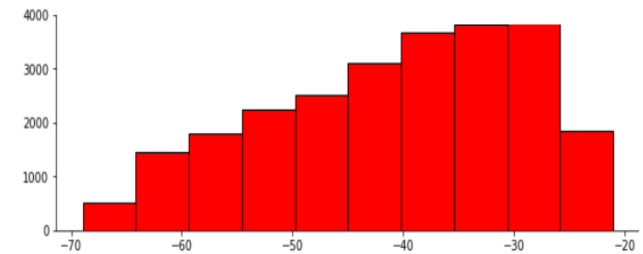
FIGURE 10 Number of loans approved vs rejected according to the type of House

Univariate Analysis : DAYS_BIRTH

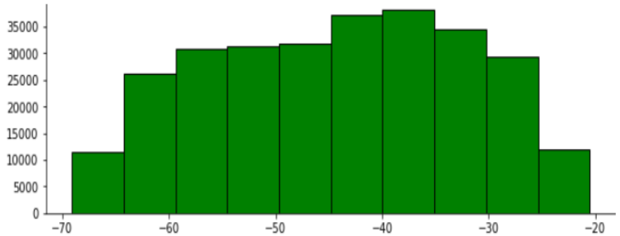


Most number of people applying for loans are in the range of (35-40) years whereas this is followed by people in the range of (40-45) years whereas the number of applicants in people aged <25 or aged>65 is very low.

Buckets of
35-45 years
are deemed to be most
capable of loan repayment



(deemed incapable)



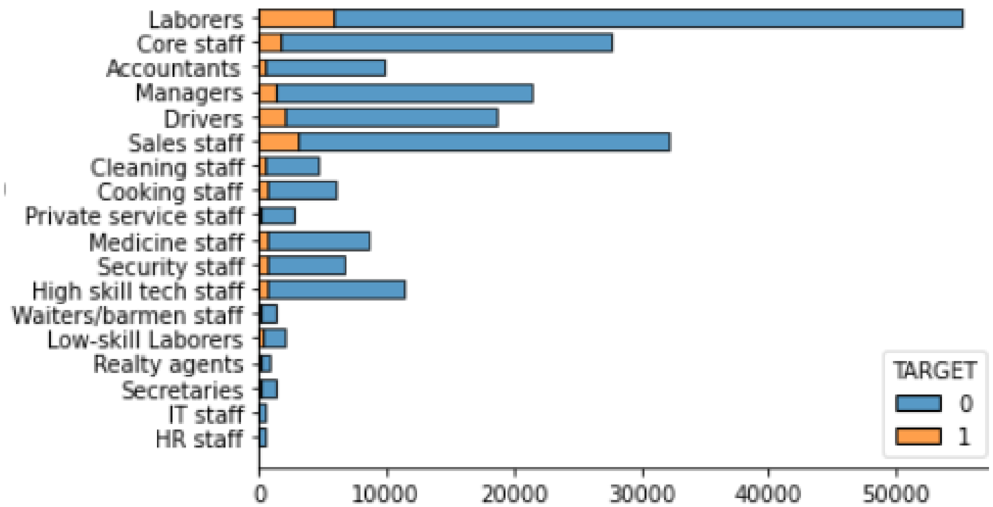
(deemed Capable)

Buckets of
25-35 years
are deemed to be most
incapable of loan repayment

FIGURE 11 Age Buckets of Client at the the time of application, deemed capable and deemed capable

Univariate Analysis : OCCUPATION_TYPE

Out of all the possible Occupation Types, the majority of applicants have not provided their Occupation Type in the application (approx. 31.39%) which is followed by Laborers (approx. 18%).

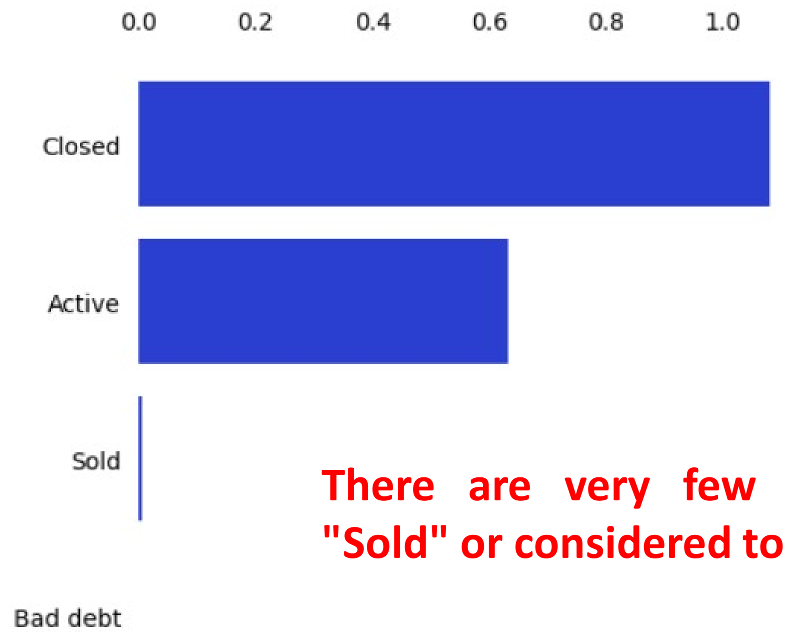


Out of all the occupations **Waiters/barmen** **least capable** of repayment followed by Laborers

FIGURE 12 Number of loans approved vs rejected according to the type of occupation

Univariate Analysis : CREDIT_ACTIVE

Most of the applications in the bureau_data is **closed**, **following by** the status of being **Active**



There are very few loans that are "Sold" or considered to be "Bad debt"

FIGURE 13 Status of the Credit Bureau

Univariate Analysis : DAYS_CREDIT

Most of clients applied for Bureau Credit is less than 500 days before the data of loan application

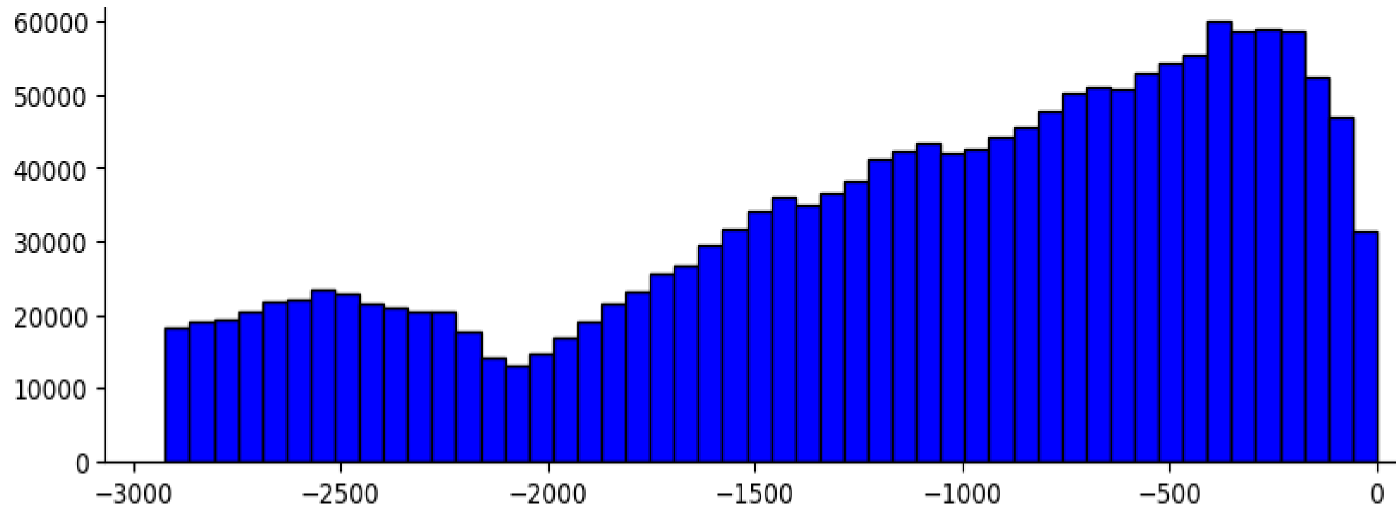


FIGURE 14 Length of days before current application that client applied for Credit Bureau Credit

Univariate Analysis : CREDIT_TYPE

Consumer Credit and **Credit Cards** are **the mostly registered** credit types in the Credit Bureau

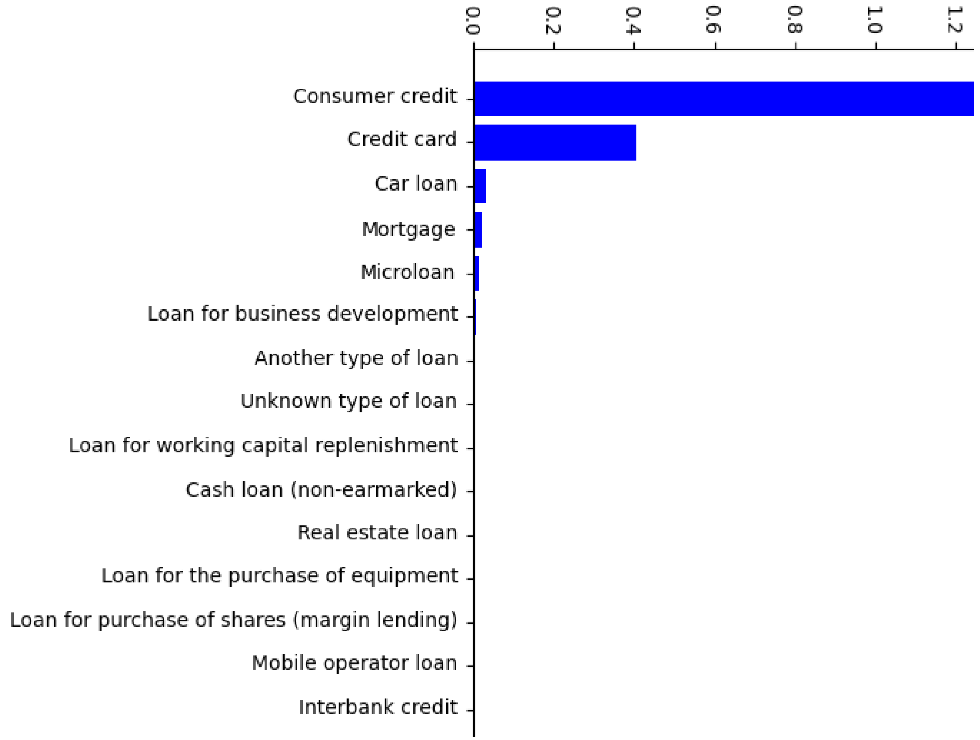
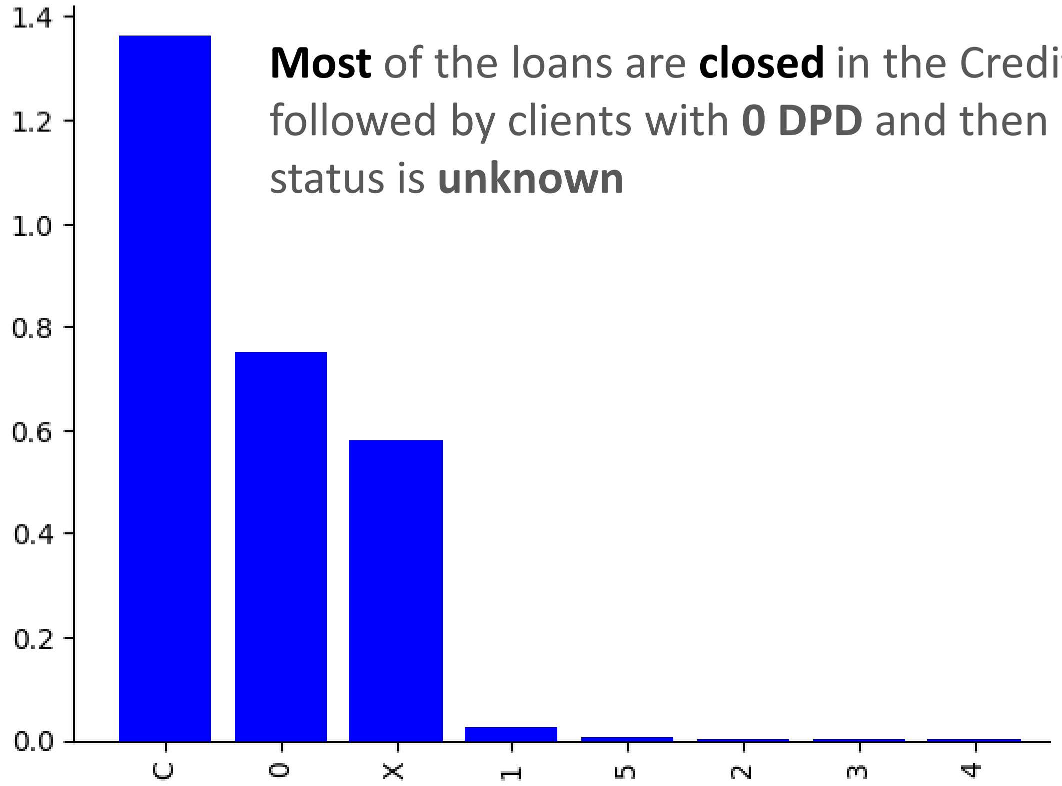


FIGURE 15 The loans according to type of credit

Univariate Analysis : STATUS



Most of the loans are **closed** in the Credit Bureau, which is followed by clients with **0 DPD** and then by applicants whose status is **unknown**

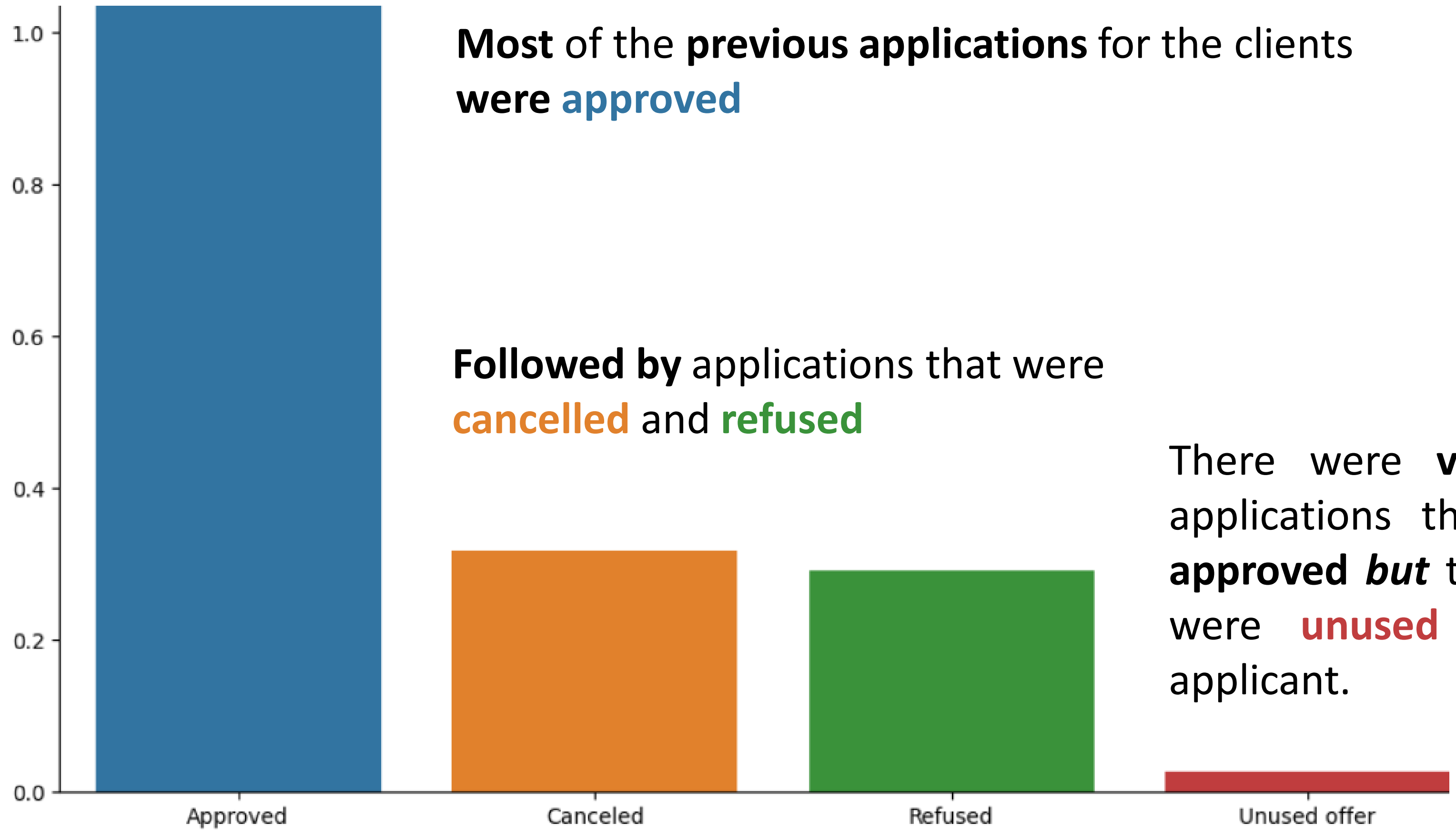
C	Closed
X	Unknown
0	No Day past Due (DPD)
1	maximal did during month between 1-30
2	DPD 31-60
3	DPD 61-90
4	DPD 91-120
5	DPD 121-150

FIGURE 16 Distribution of Status in Bureau

Exploratory Data Analysis

Previous Application

Univariate Analysis : NAME_CONTRACT_STATUS



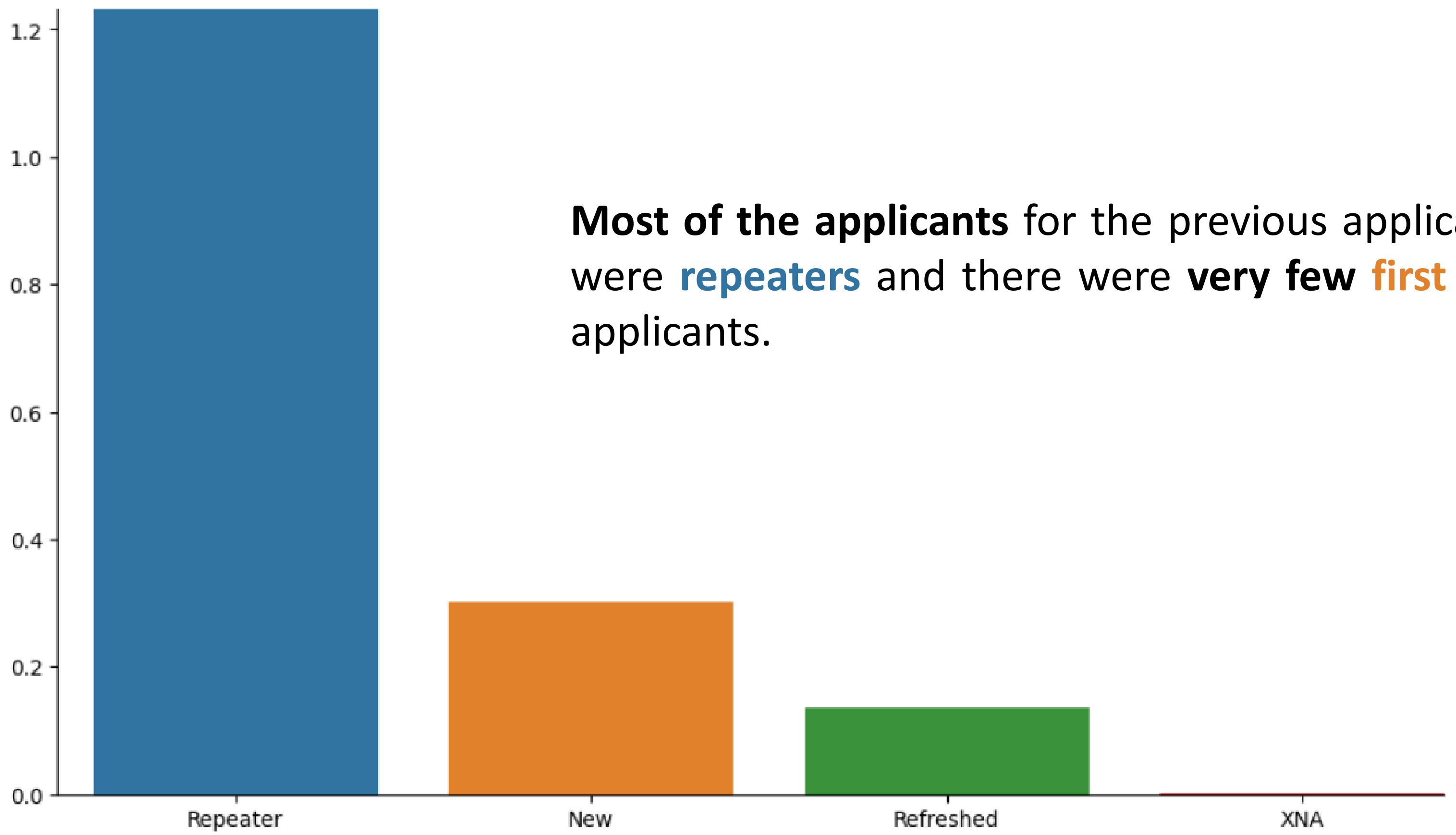
Most of the **previous applications** for the clients **were approved**

Followed by applications that were cancelled and **refused**

There were **very few** applications that **were approved** ***but*** the loans were **unused** by the applicant.

FIGURE 17 Distribution of previous application status

Univariate Analysis : NAME_CLIENT_TYPE



Most of the applicants for the previous application were **repeaters** and there were **very few** **first time** applicants.

FIGURE 18 Distribution of previous application status

Univariate Analysis : MONTH_BALANCE

Most clients has

10-20 months balance

before **the date of application.**

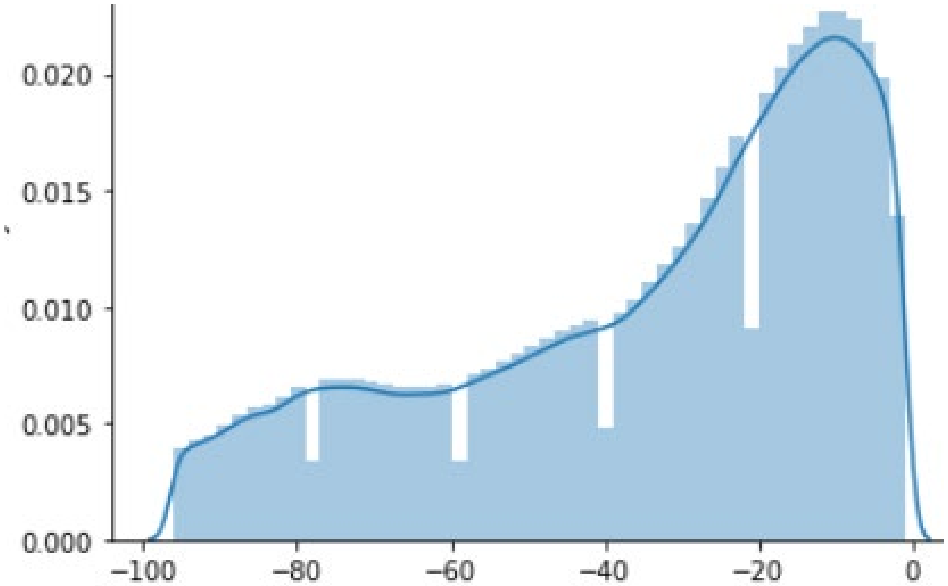
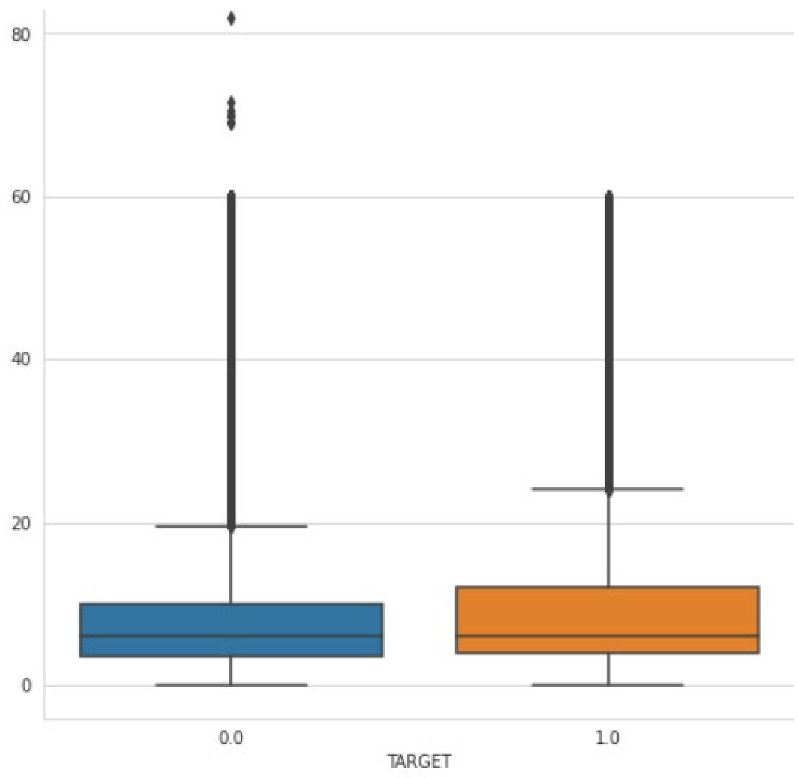


FIGURE 19 Distribution of Month balance

Univariate Analysis : CNT_INSTALMENT_FUTURE



Incapable tend to have **more number of Installments remaining** on their previous credits as compared to **Capable.**

FIGURE 20 Box-plot for cnt installment future

Univariate Analysis : DAYS_INSTALMENT

The **Incapable** tend to have lesser number of days since their last payment, while **Capable** have more number of days since their last payments.

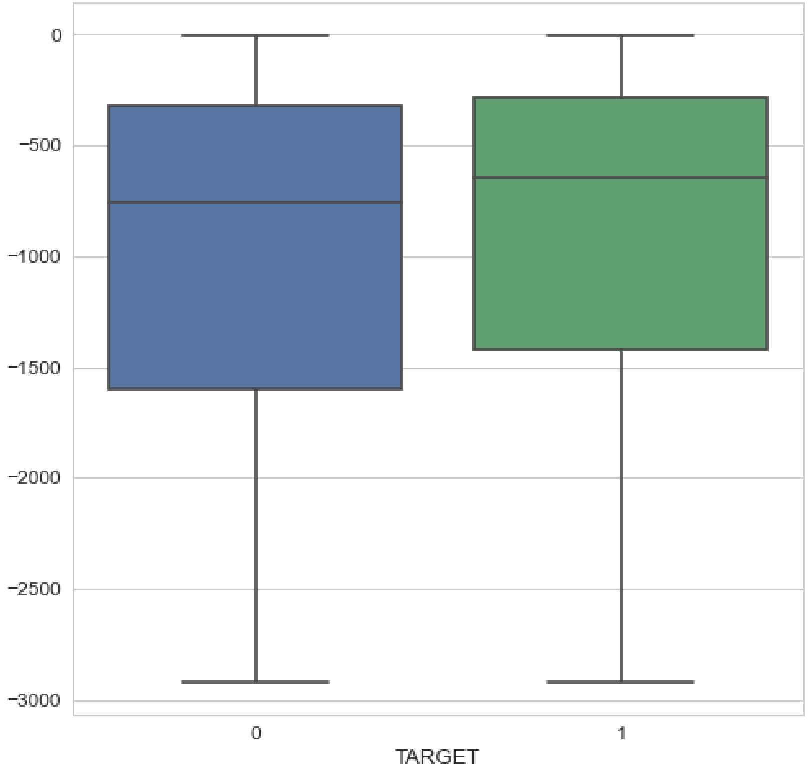


FIGURE 21 Box-plot of day instalment

Univariate Analysis : AMT_BALANCE

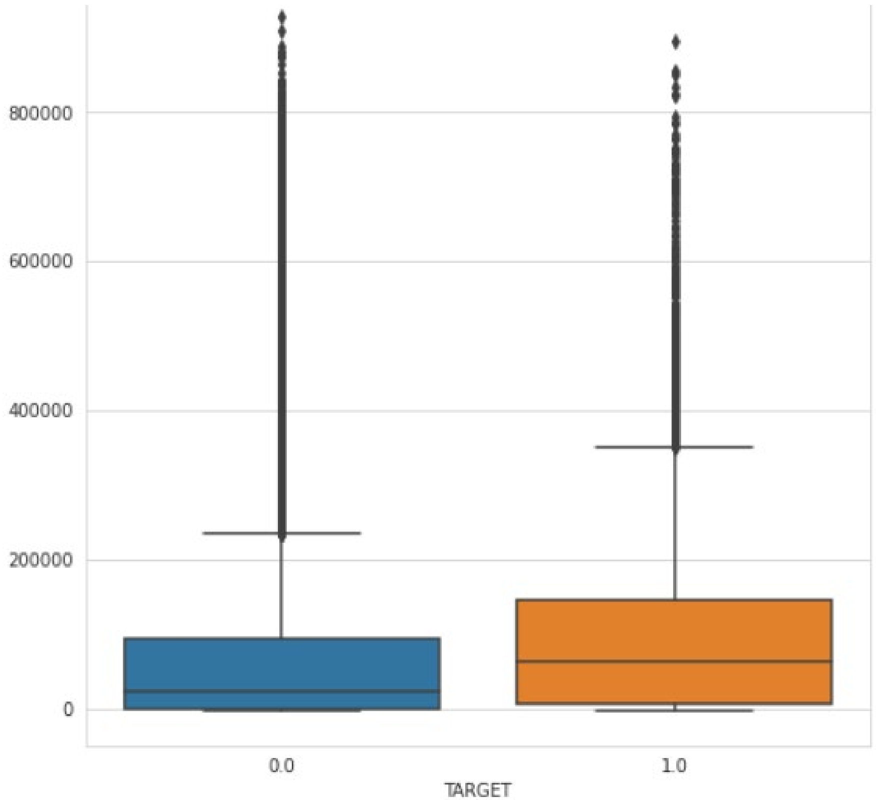


FIGURE 22 Box-plot of amount balane

Incapable here too appeared to have a **higher minimum installment** each month as compared to **Capable**

Univariate Analysis : AMT_ TOTAL_RECEIVABLE

The Incapable usually had **higher amount receivable** on their previous credit, which may imply the higher amounts of credits that they may have taken.

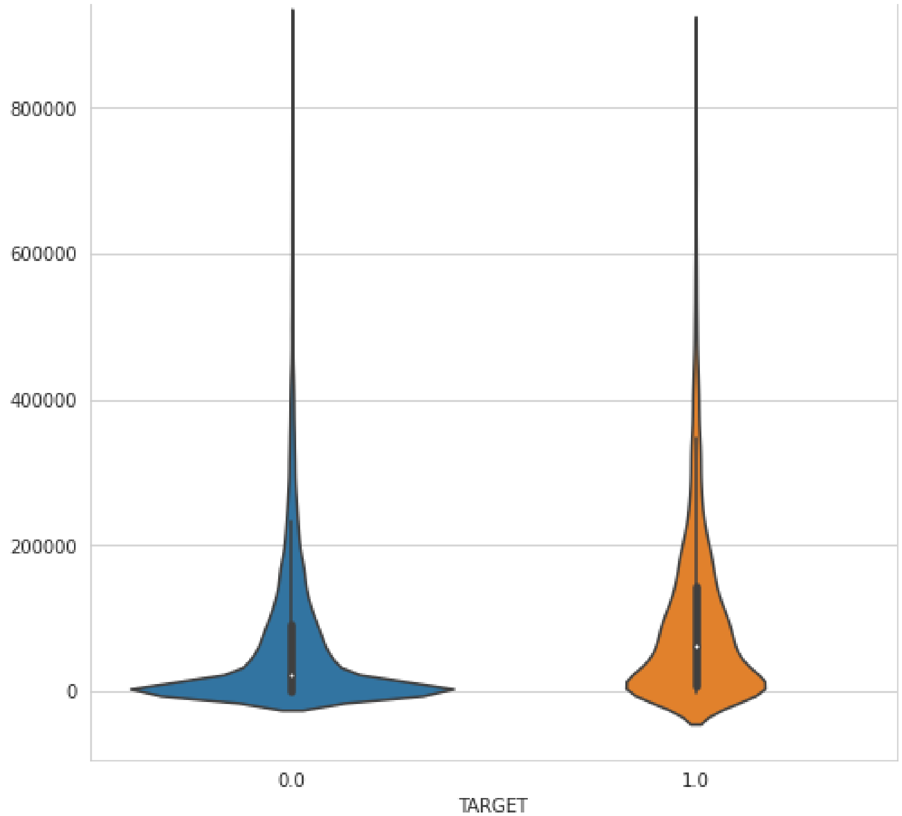


FIGURE 23 Box-plot of amount total receivable

Exploratory Data Analysis

Credit Card Balance

Univariate Analysis : CNT_INSTALMENT_MATURE_CUM

The **capable** usually had **higher range** of values for the **number of installments** paid as compared to **incapable**.

This might show the defaulting behaviour, where in the **capable** usually would **pay fewer number of installments** on their previous credit.

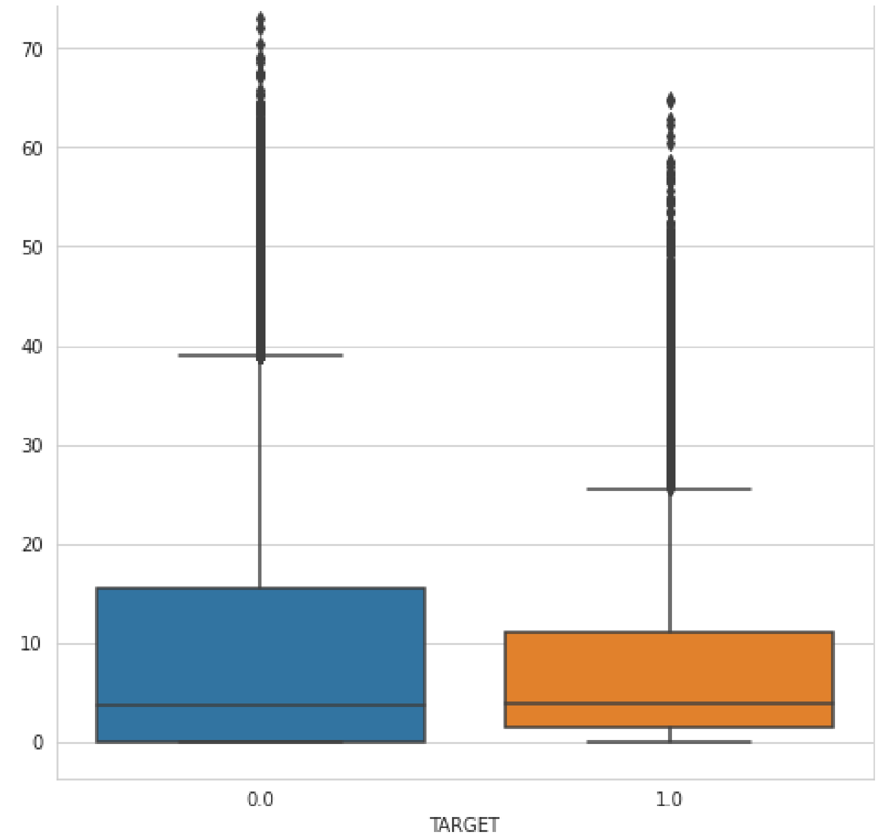


FIGURE 24 Box-plot of number of instalment previous credit