Capstone Project Milestone Report

Home Credit Default Risk: Predicting how likely each applicant is of repaying a loan?

1. Project Overview

The goal of this capstone project is to predict loan repayment ability of clients, namely the probability of default of clients. The reason behind the project is that many people struggle to get loans due to insufficient or non-existent credit histories. As a result, this population is often taken advantage by untrustworthy lenders. Hence, if lenders could predict the repayment ability better, unbanked population could have a better loan experience. The objective is to use historical data to predict whether or not an applicant will be able to repay a loan. This is a standard supervised classification problem where the label of target variable is a binary variable, 0 (will repay loan on time), 1 (will have difficulty repaying loan).

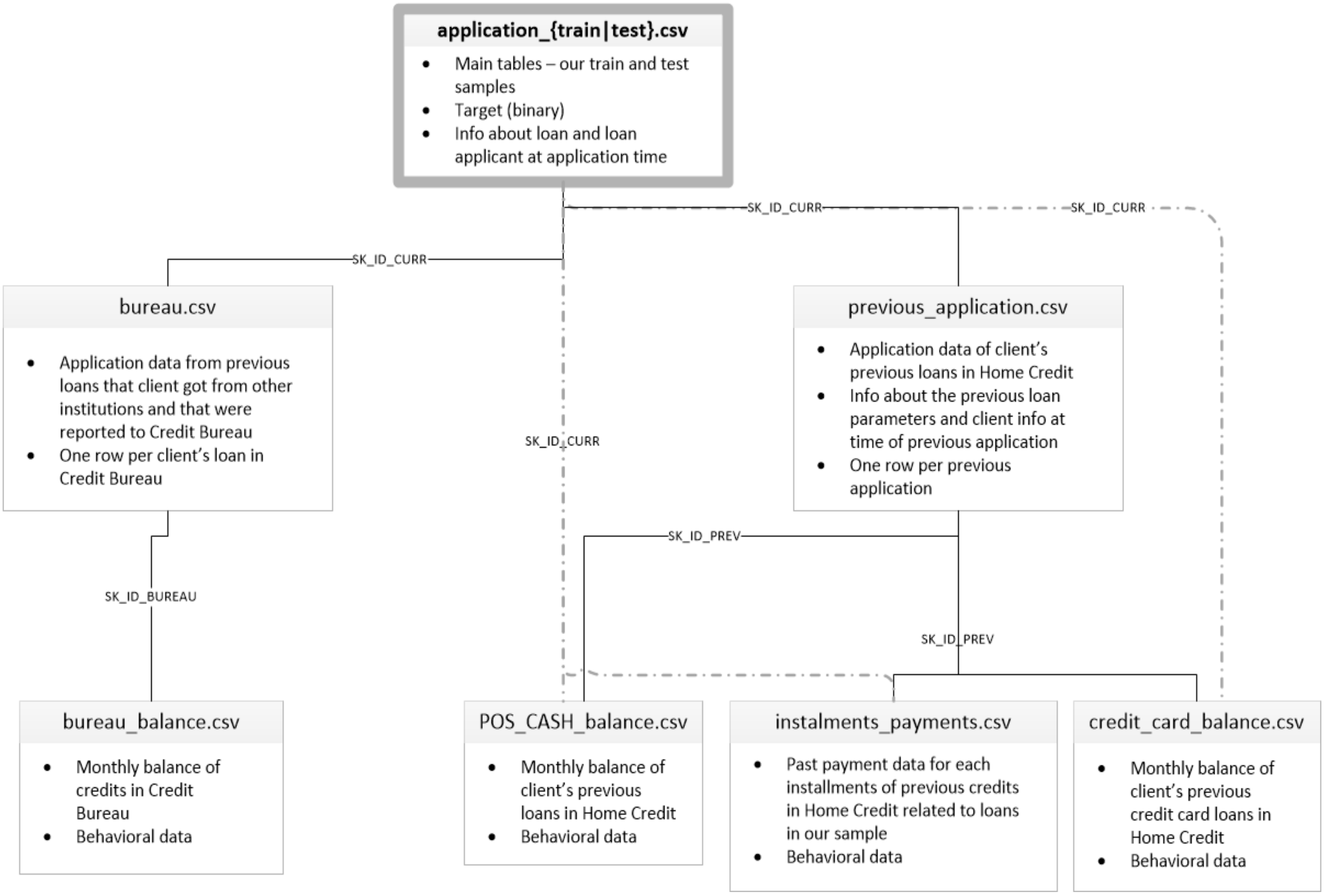
2. Dataset

The data set that we use for this project comes from publicly available Kaggle Competition dataset provided by Home Credit. Home Credit has provided a variety of datasets to predict their clients' repayment abilities.

There are five separate *.csv files were used as part of this project:

- **application_train.csv:** static data for all applications. One row represents one loan in our data sample.
- **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.
- **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 – i.e. the table has (#loans in sample # of relative previous credits # of months in which we have some history observable for the previous credits) rows.
- **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.
- **installments_payments.csv:** Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample. There is a) one row for every payment that was made plus b) 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.

The following diagram provides a brief summary:



application_{train|test}.csv
- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

bureau.csv
- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

previous_application.csv
- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

bureau_balance.csv
- Monthly balance of credits in Credit Bureau
- Behavioral data

POS_CASH_balance.csv
- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv
- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

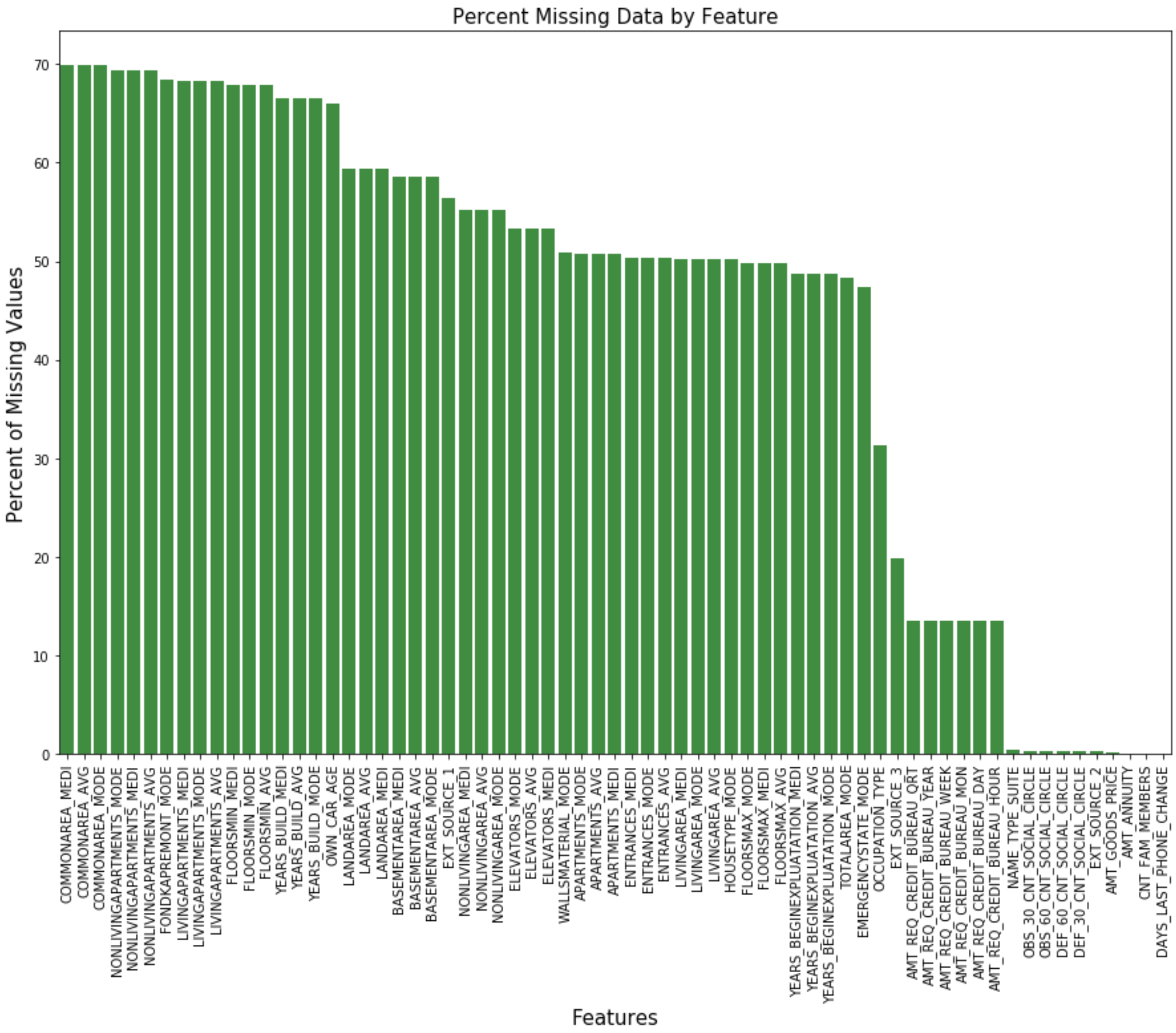
credit_card_balance.csv
- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

The dataset contains personal and financial information belonging to 356,255 individuals who had previously been recipients of loans from Home Credit. These individuals are divided into training and testing sets:

- Training dataset: 307,511 Records, 122 Columns
- Test dataset: 48,744 Records, 121 Columns.

The first two columns, SKI_ID_CURR and TARGET , represent a borrower's loan ID and target value, respectively. The loan ID is a unique identifier assigned to each borrower. A target value of 1 indicates that the borrower eventually made at least one late loan payment. A target value of 0 indicates that the borrower always paid on time. The remaining 120 columns are explanatory variables that contain information about borrowers, including history of past payment, marital status, age, type of housing, region of residence, job type, education level, amount of bill statement.

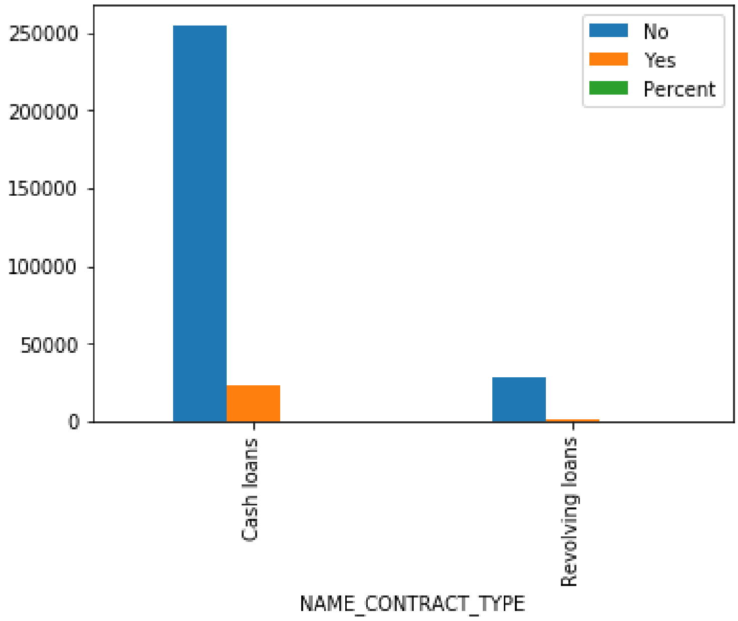
There are missing values in the datasets. Over 67 of the 120 total features have one or more entries of NaN entries. The graph below shows the features has missing value and precentage of missing value avaiable in training dataset.



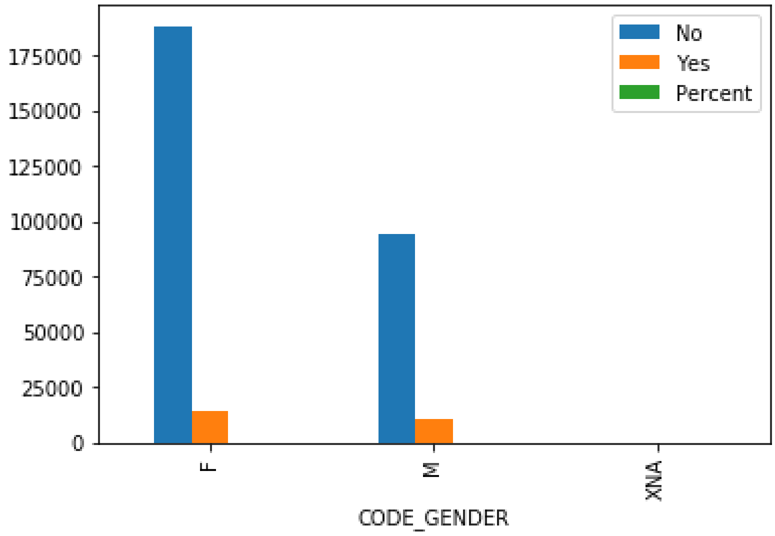
3. Exploratory Data Analysis (EDA):

We use exploratory data to summarize data sets’ main characteristics with visual methods. We answer following questions by visually inspecting data.

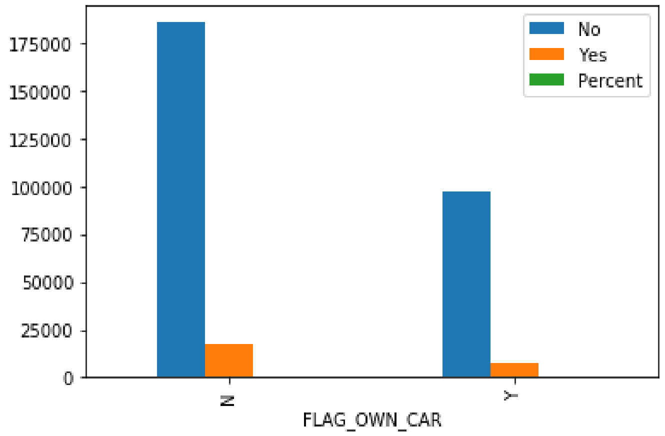
1. **Is there difference in default rate depending on loan type (cash vs revolving)?**
   Revolving loans are just a small fraction from the total number of loan. Default among revolving loans is lower, as well (8.3% vs 5.5%).



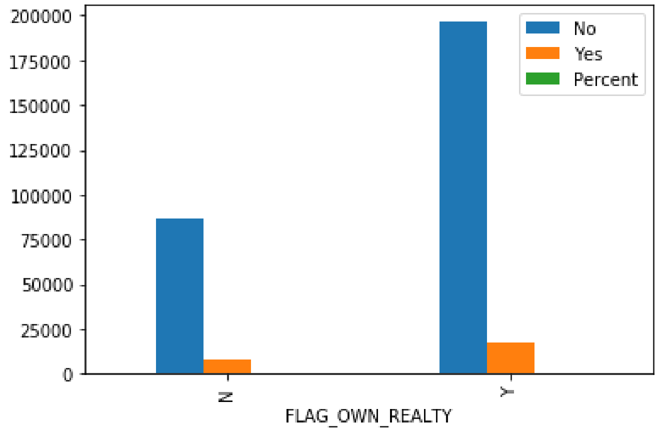
2. **Is there difference in default rate depending on gender?**
   Females have relatively more difficulties in repaying the loan back. The number of female clients is almost double the number of male clients. Looking to the percent of defaulted credits, males have a higher chance of not returning their loans (10.1%), comparing with women (7%). This could be because of the general larger population of female applicants as opposed to male applicants.



3. **If a customer owns a car, would it affect payment of loan?**
   The clients with a car are almost half of the ones that doesn't have one. The clients that owns a car are less likely to not repay loan. Both categories have close not repayment rates (8.5% vs 7.2%).

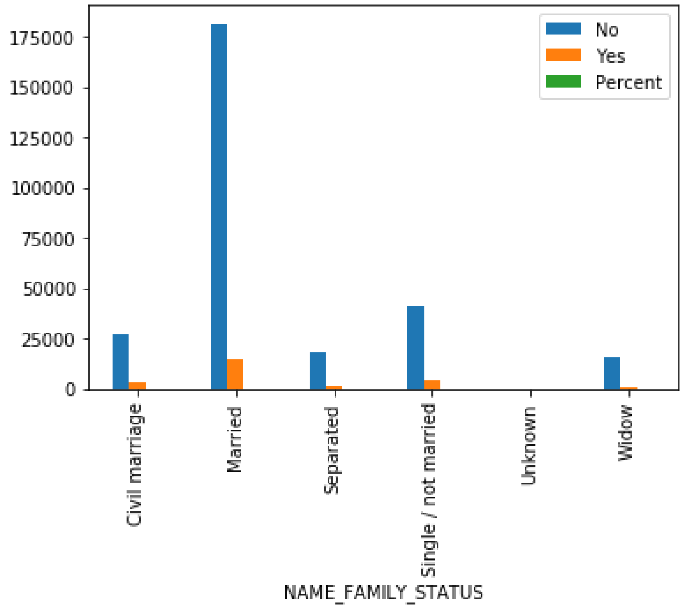


4. **If a customer owns a real estate, would it affect payment of loan?**
   The clients that own real estate are more than double of the ones that doesn't own. Both categories (owning real estate or not owning) have not-repayment rates close to 8%.

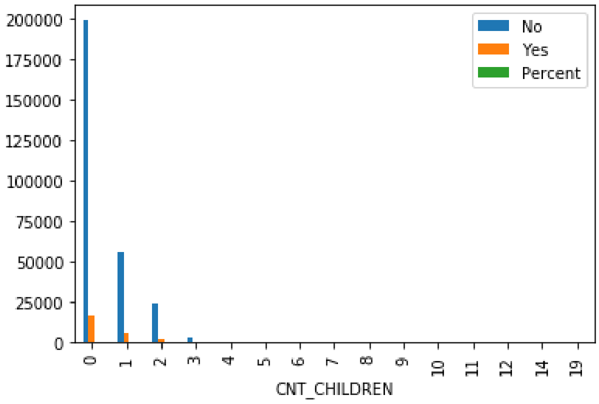


5. **Does the family status of client affects payment?**

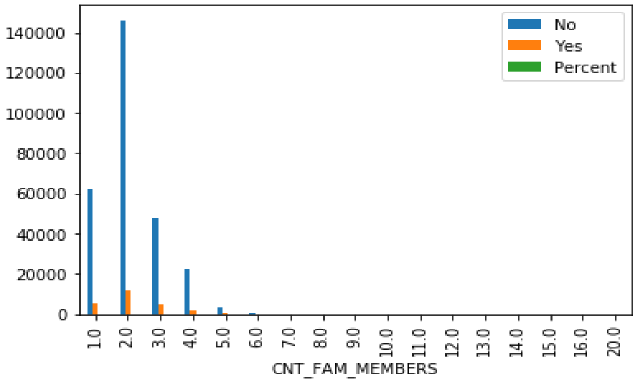
   Most of clients are married, followed by Single/not married and civil marriage. In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest



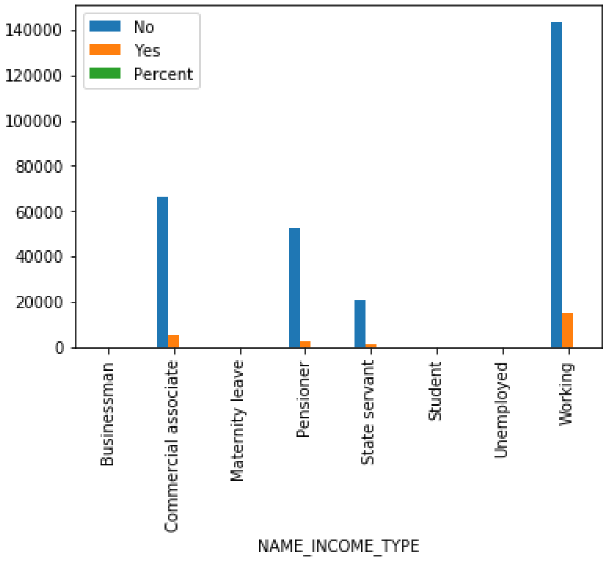
6. **Is having children affects payment?**
Most of the clients taking a loan have no children. As for repayment, clients with no children, 1, 2, 3, and 5 children have percents of no repayment around the average (10%).



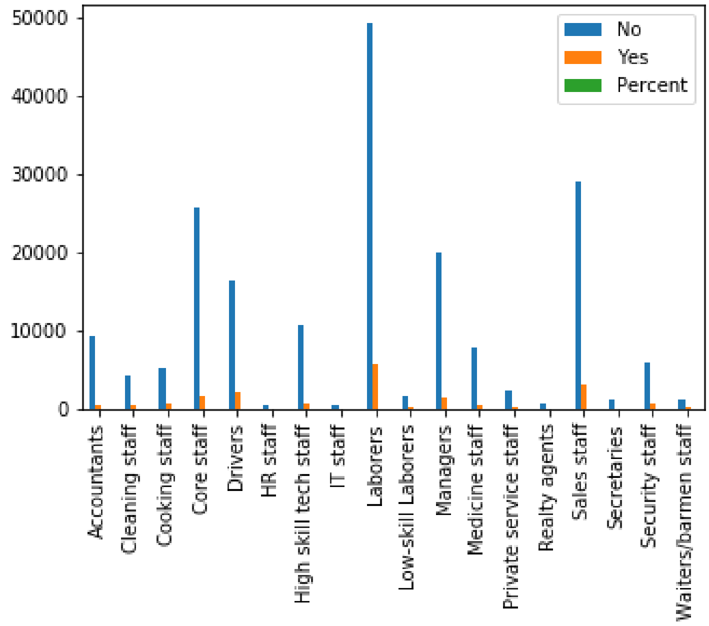
7. **Does family size have an affect on repayment?**
Clients with family members of 2 are most numerous, followed by 1 (single persons), 3 (families with one child) and 4. As for repayment, clients with up until 5 members is below 10%. Clients with family size of 11 and 13 have 100% not repayment rate. Clients with families size of 10 or 8 members have not repayment rate over 30%.



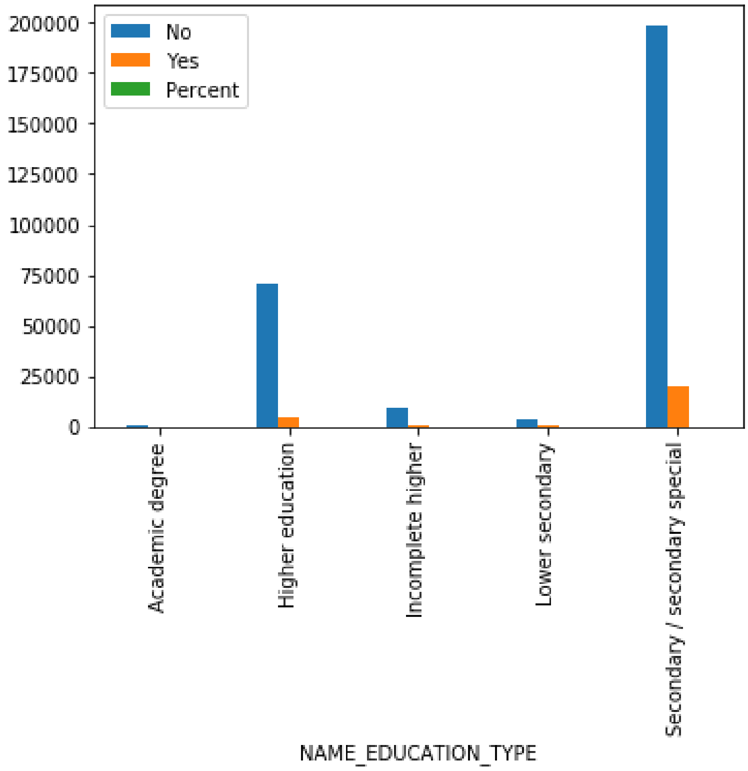
8. **Does income type have an affect on repayment?**
Most of applicants have Working job status, followed by Commercial associate, Pensioner and State servant. The applicants with the type of income Maternity and Unemployed have highest level of default. The restare under the average of 10% not paying loans.



9. **Does the occupation have an affect on repayment?**
   Most of the loans are taken by laborers, Sales staff. IT staff take the lowest amount of loans. Low-skill Laborers have the highest default rate (above 17%), followed by Drivers, Waiters/Barmen, Security staff, Laborers and Cooking staff.



10. **Does the organization type has an affect on repayment?**
    Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
11. **How the education of client affects repayment rate?**
    Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. The Lower secondary category have the highest repayment rate (11%). The people with Academic degree have less than 2% not repayment rate.


No
Yes
Percent
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE