**Capstone Project Final 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 in this project 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 intended result is to be able to predict if a consumer applying for loan will be able to repay it on time (binary variable is zero) or will delay the repayment (binary variable is 1). We will apply machine learning models to learn from the available data and predict on new unseen data of consumers. Our desired output of the model is a probability of a client belonging to one of two classes – one which is likely to repay the loan on time and the other potentially delaying payments.

**2. Dataset**

The dataset 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 seven datasets and the relationshio between the different datasets is ilustrated by the diagram on the next page.

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

* **Application\_train:** static data for all applications. One row represents one loan in our data sample. The training set consists of a label, TARGET, where 1 represents client with payment difficulties and 0 for those without. We have 120 features in this dataset, with 307,511 observations, each with a unique ID. This ID is used later to join information from the other datasets to the main dataset. The feature space covers essential key information for loan applications such as the amount of credit for the loan, annual income of the client, annuity of the loan. It also includes detailed information about the client's housing, the documents provided by the client during the application etc
* **Bureau:** 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.
* **Bureau Balance**: The bureau balance dataset gives us more details about each of the loans in the bureau dataset. This dataset gives us the monthly balance for the loans specified by the SK\_ID\_BUREAU in the bureau dataset and the status of the loan with categorical values such as Active, Closed, unknown or the binned amount of days past due.
* **POS Cash Balance:** 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.
* **Credit Card Balance:** Monthly balance of clients’ all previous credit cards in Home Credit.
* **Previous Application:** 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:** 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 of the relationship between the datasets:**



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.

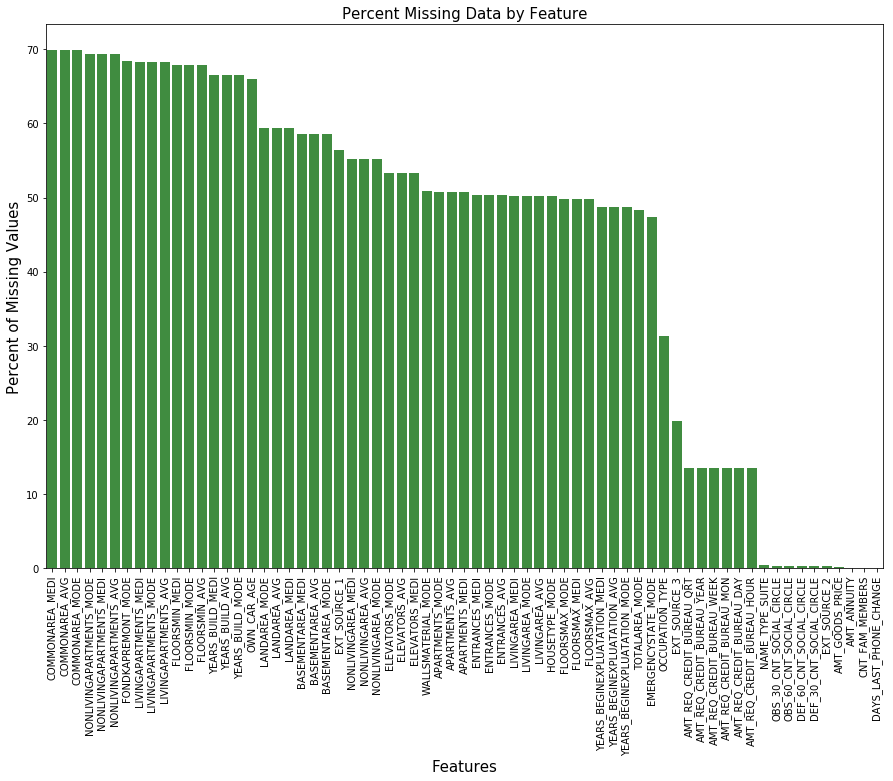
respectively. The loan ID is a unique identifier assigned to each borrower. A target value of 1 indicates that the borrower eventually made at least one late loan payment. A target value of 0 indicates that the borrower always paid on time.

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.

**Summary of the Datasets:**The following table summarizes the total number of features and their data types across all the datasets:



**Missing values in the datasets:** 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.



**Anomalies in the datasets:** We looked at the specific columns, namely days\_birth and days\_employed to detect anomalies in these columns.

* Days of birth: Ages looks reasonable, max and min ages do not show anomaly:

count 307511

mean 43.936973

std 11.956133

min 20.517808

25% 34.008219

50% 43.150685

75% 53.923288

max 69.120548

Name: DAYS\_BIRTH, dtype: float64

* Days employed: There is a problem in the maximum value, maximum value can not be 1000 years. These abnormal values were replaced with nan, to be treated as missing values.

count 307511

mean 63815.045904

std 141275.766519

min -17912.000000

25% -2760.000000

50% -1213.000000

75% -289.000000

max 365243.000000

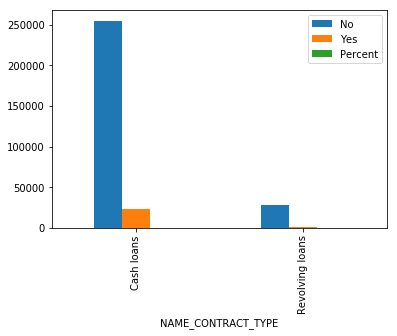
Name: DAYS\_EMPLOYED, dtype: float64

**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:

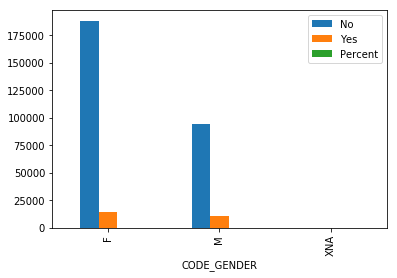
1. **Is there difference in default rate depending on loan type (cash vs revolving)?**

Revolving loans are just a small fraction from the total number of loan. Default among revolving loans is lower, as well (cash loans: 8.3% vs revolving loans: 5.5%).



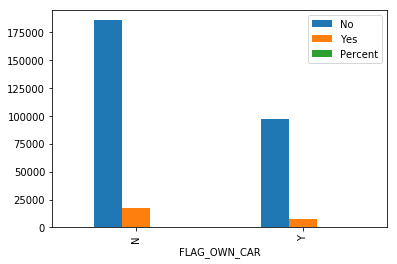
1. **Is there difference in default rate depending on gender?**

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.



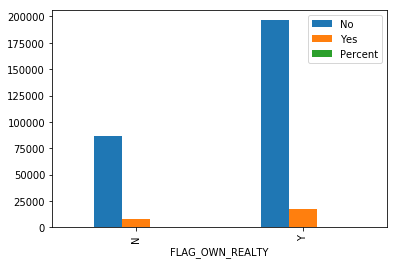
1. **If a customer owns a car, would it affect payment of loan?**

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 default rates (no car: 8.5% vs car owner: 7.2%).



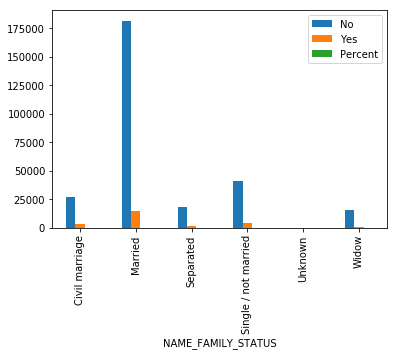
1. **If a customer owns a real estate, would it affect payment of loan?**

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 default rates close to 8%.

****

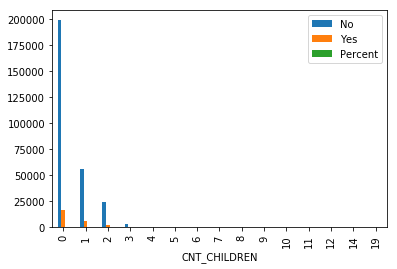
1. **Does the family status of client affects payment?**

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 default rate (10%) and Widow has the lowest.

****

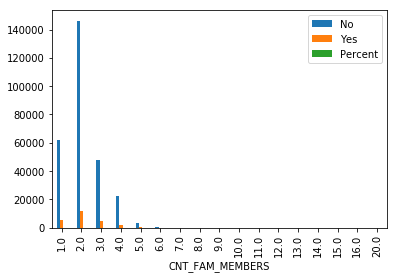
1. **Is having children affects payment?**

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 default rate around the average (10%).

****

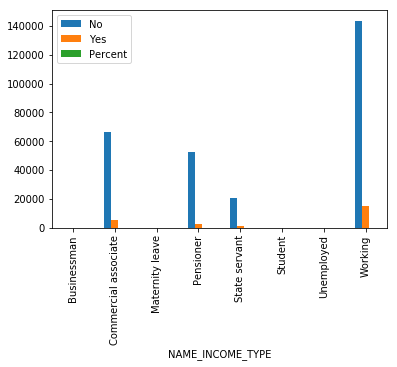
1. **Does family size have an affect on repayment?**

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 default rate over 30%.



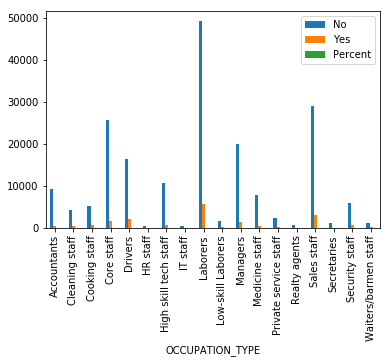
1. **Does income type have an affect on repayment?**

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.



1. **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.



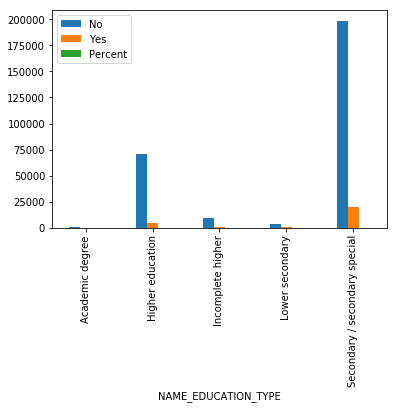
1. **Does the organization type has an affect on repayment?**

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%).

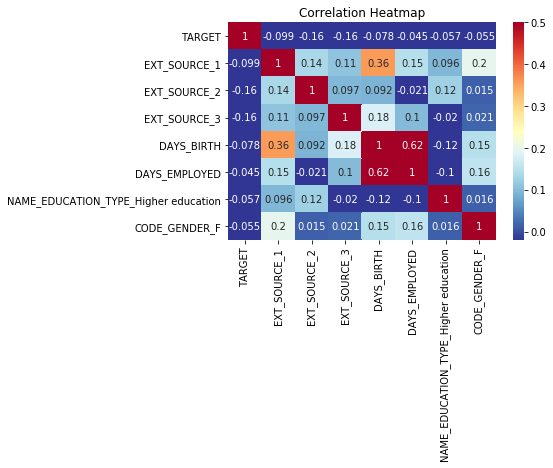


1. **How the education of client affects repayment rate?**

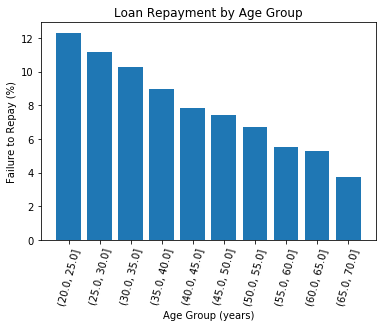
Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. The Lower secondary category have the highest repayment rate (11%). The people with Academic degree have less than 2% not repayment rate.



1. **Other Exploratory Visualization and Correlations**

****EXT\_SOURCE\_1, EXT\_SOURCE\_2 and EXT\_SOURCE\_3 show inverse influence on risk repayment ability of the applicant based on above correlation heatmap

The bar plots below show the percentage of loan repayment failure in terms of age groups and years employed groups. Younger clients, in the age group of 20 to 25, appears to default the most on their loans. The risk of defaulting decreases gradually with age.



**4. Algorithms and Techniques**

Since this is a classification problem, a naive benchmark score would be would be 0.5.

Two models were picked to be tested as our baseline: Random forest and a logistic regression classifier.

**5. Results**

xxxx

**6. Conclusion**

xxxx