## Home Credit Risk Analysis

Capstone Project Milestone Report

**Introduction**

The increasing number of loan applicants with the effects of global financial crisis lead to the increase of credit risk analysis. That aims to prevent the loss due to a borrower’s failure to make payments, to improve the overall performance and secure a loan risk. Credit risk identification require the ability to store, categorize financial data based on a variety of criteria, assess risk levels associated with borrowers’ profiles, their capability to repay the loan, and their borrowing history. Credit risk analytics is also an approach for financial firms to gather and use data in order to keep up with new demands, get more impact in the market. Building a professional model can help business teams to evaluate, quantify model risk and performance for every individual credit decision, and determine the safety constraints and key decision factors for lending risks. Credit Risk analysis is important for banks and lender to ensure that the borrower has a good credit score, the capability to repay their debt. Therefore it helps them to improve their business and serve customer better.

**Objective**

This project’s goal is to create a prediction model to help decide if the loan is risky or not, which is further help in making decision in lending loans for customers.

**Dataset**

The data is provided by Home Credit, <https://www.kaggle.com/c/home-credit-default-risk/data>

There are 7 different sources of data:

* *application\_train/application\_test*: the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK\_ID\_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid.
* *bureau*: data concerning client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.
* *bureau\_balance*: monthly data about the previous credits in bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
* *previous\_application*: previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK\_ID\_PREV.
* *POS\_CASH\_balance*: monthly data about previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows. \_ *credit\_card\_balance*: monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
* *installments\_payment*: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.

**Data Exploration**

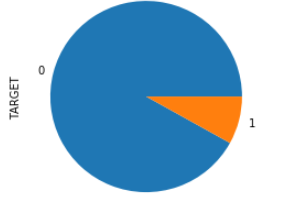
The first step of the project is to do data exploration to understand about features and correlation between features, clean missing values or outliers.

After combining files, the dataset contains 200 features, and more than 300,000 distinct records. 67 features have missing values.

1. ***Percentage of loan which is paid?***

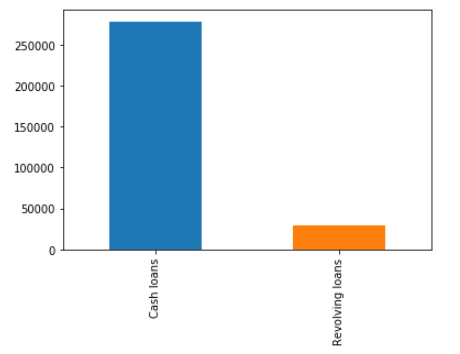
The target variable contains 2 values:

* 0 : the customer who paid the loan.
* 1 : the customer who didnot paid the loan. The target is imbalance with the majority of loans is paid (91%) and about 9% of the loans is not paid



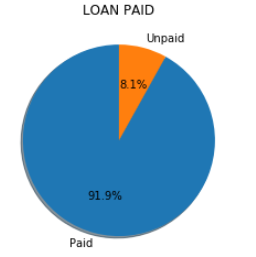
1. ***Type of loans:***

There are 2 types of loan in the data, where cash loans are more than revolving loan.



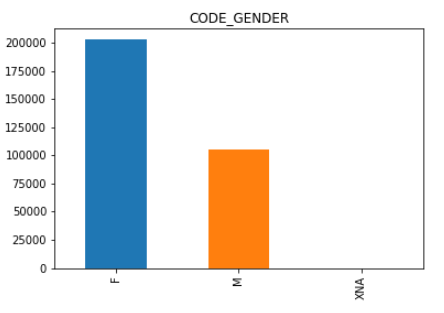
1. ***Is the loan was paid on time or not?***

91% of the loan is paid and 8% of the loan is not paid. We see this is an imbalanced problem.



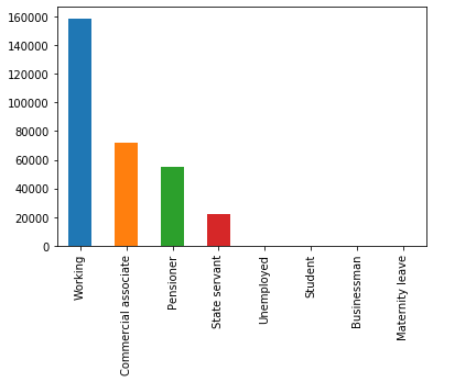
1. ***Which gender mostly to apply for the loan?***

Female tends to apply for the loan than Male.



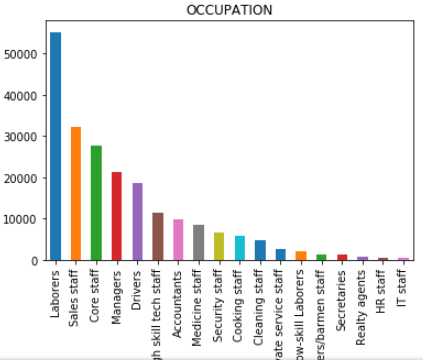
1. ***What source of income for applying a loan?***

Most of income sources are from people who working.

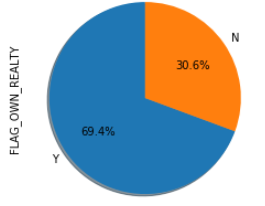
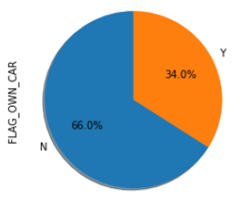
******

1. ***Applicant information analysis*** [***¶***](http://localhost:8888/notebooks/3.home_credit_EDA.ipynb#1.3:-What-source-of-income-for-applying-a-loan?)

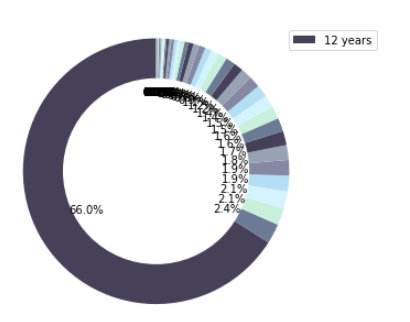
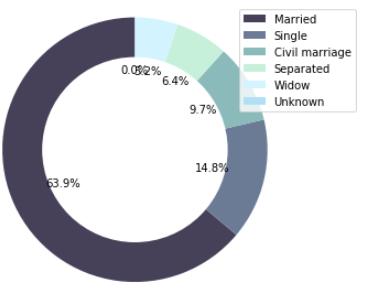
Most of applicants are laborers, sales, and staff



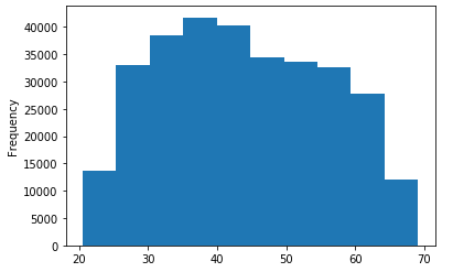
66% of people who do not have car applied for a loan. 69% people own a property tends to apply for a loan more than people who not own a property

Most of applicants own a car is 12 years. 63% of applicants are married.

Most of people in age from 30 to 45 have loan.

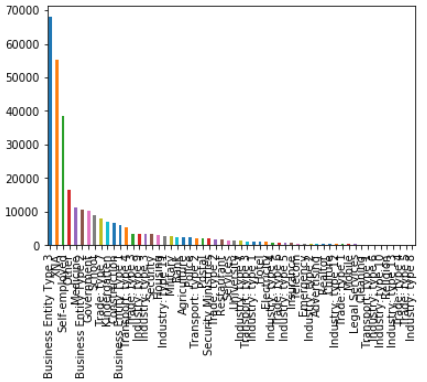


Top organization who applied for a loan:

- Business

- Self employed

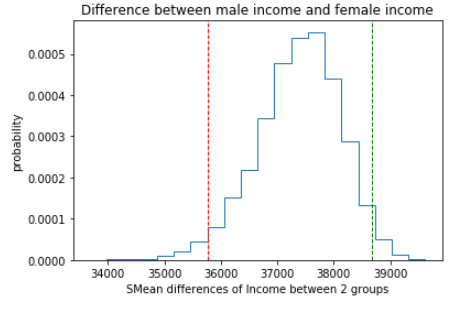
- Medicine



**Statistical Inference and Analysis**

*Question 1: Is there any difference of income between female and male?*

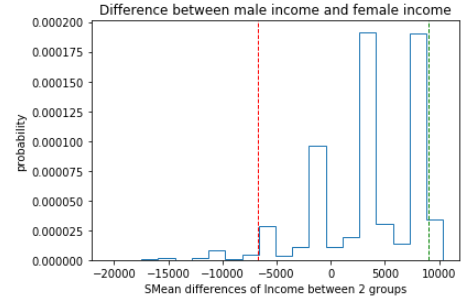
95% confidence interval for the difference between the mean of female and male income over 10,000 replicates is 35773.12 and 38670.224.



There is a statistically significant between the means of 2 groups.

*Question 2: Is there any difference between 2 means of income who paid for the loan and not paid for the loan.*

95% confidence interval for the difference between the mean of paid and non-paid income over 10000 replicates is -6747.13, 9053.15.



Q3: Perform a bootstrapped hypothesis test at the 5% significance level (α=0.05) to calculate the p-value of the observed difference between female and male income. We get the p-value = 0. Therefore, there is a statistically significance in income between female and male groups.

**Conclusion:**

Based on the exploration, there are a lot of missing data and outlier that need to be processed. Once the data is cleaned and ready for further analysis, we can applied classification model (logistic regression), tuning method to the data.