Home credit risk analysis

Many people struggle to get loans due to insufficient or non-existence credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. This phenomenon is so popular in both developed and developing countries. Even if in the USA, the financial system is very strong and well established. Loan for an individual is still hard for banks and other financial institutes. In countries with huge populations like China, small loan for many personal purposes such as start-up a business or go to schools for individuals is also difficult problem due to insufficient credit history or relevant credit in other social aspects. Thanks to the age of big data, we are now able to connect diverse types of data which is related to one person’s credit and formulate his/her personal financial profile. More and more financial firms started taking into consideration of the power of big data and method using machine learning/AI to build model for credit analysis automatically. Home Credit is one of these financial firms, founded in 1997 and headquartered in Czech Republic. As an international non-bank financial institution, Home Credit operates in 14 countries and focuses on lending primarily to people with little or no credit history. As of 2016 the company has over 15 million active customers, with two-thirds of them in Asia and 7.3 million in China. Major shareholder of company is PPF, a private held international financial and investment group, which controls an 88.62 stake.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has positive load experience, Home Credit makes use of a huge and variety of alternative data –including telco and transactional information – to predict their clients’ repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they keep seeking better models and systems that make better client experience and efficient prediction. In this report, we will present our model for calculating the probability of repaying a loan for each client based on enormous clients’ historical data from Home Credit

# Data explanation

The raw data from Home Credit is not well-cleaned and diverse. The FIG.1 shows the relationships between these primary tables and we first list the meaning of each fields of each table.

Seven tables are provided separately, they are

* Main table: the client application for credit currently, which contains current application information and TARGET (how difficult the client can repay the loan)
* Bureau table: the previous application of a client in other financial institutes that were reported to Credit Bureau (for clients who have a loan in Home Credit).For every loan in these samples, there are as many as number of credits the client had in Credit Bureau before the application date
* Bureau balance: monthly balances of previous credits in Credit Bureau. This table has one row for each month of history of every precious credit reported to Credit Bureau—i.e. the table has # of loans in sample \* # of relative previous redit \*# of months where Home Credit have some history observable for the previous credits
* POS\_CASH\_balance: monthly balance snapshots of precious POS(point of sales) and cash loans that the applicant had with Home Credit
* Credit\_card\_balance: monthly balance snapshots of previous credit cards that the applicant has with Home Credit
* Previous application: all previous applications for Home Credit loans of clients who have loans in Home Credits’ samples
* Installments\_payments: repayment history for the previously disbursed credits in Home Credit related to the loans in current samples

The overall relationships between these tables:



FIG.1

These tables can be joined into main table by different keys after preprocessing. The difficulty of these data tables is that there are many missing values in many fields and it is hard to extract the useful features based on so many interactions among fields.