Home Credit Default Risk

Chris Barcelon
August 2018

I. Definition

Project Overview

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.

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 a positive loan experience; Home Credit makes use of a 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're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

The Home Credit Default Risk competition.

Problem Statement

The problem is to build a model that will predict the probability that a loan applicant will default on their home loan.

Metrics

The model will be scored on area under the ROC curve between the predicted probability and the observed target.

II. Analysis

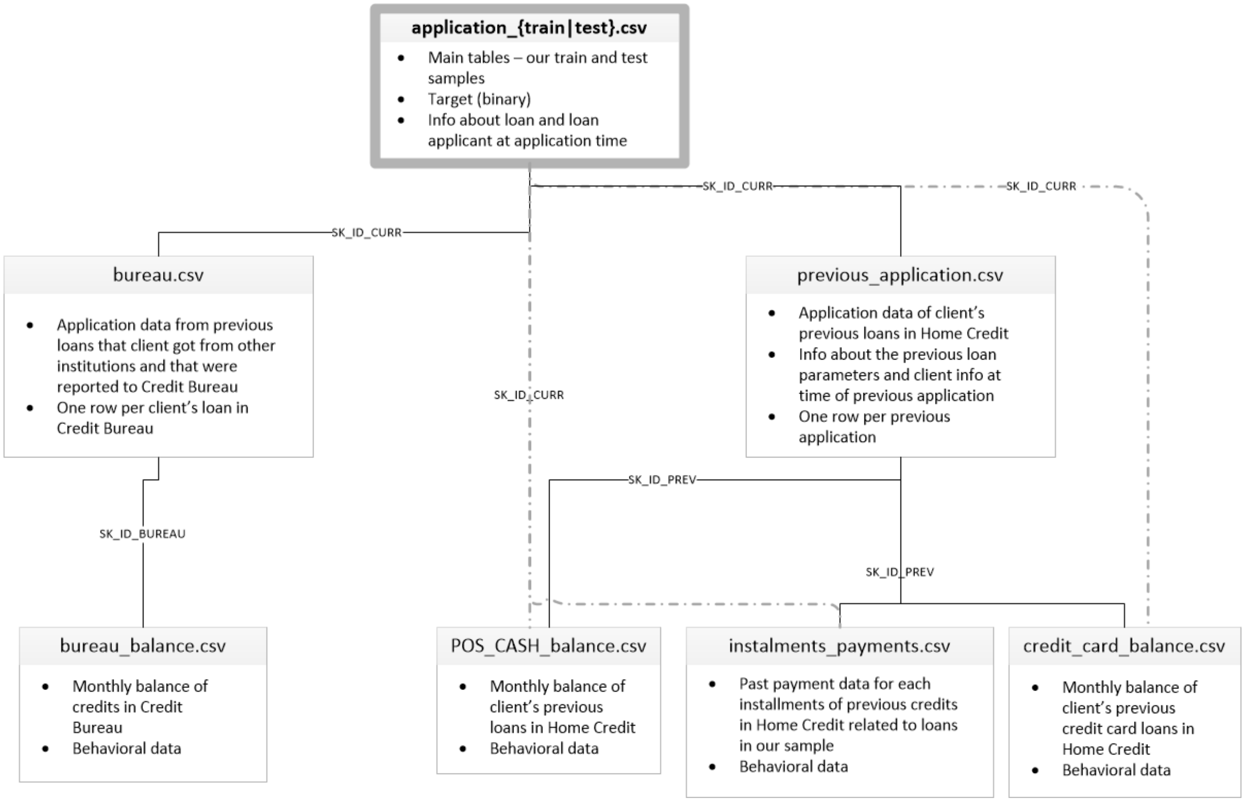
Data Exploration

The dataset is broken up into 8 csv files. There are two main files the application_train and application_test files which hold the static information for each application. The train file has a TARGET column that has 1 if the loan was repaid and 0 if the applicant had difficulty making loan payments. The 6 remaining files hold supplemental about the clients applying for the loan.

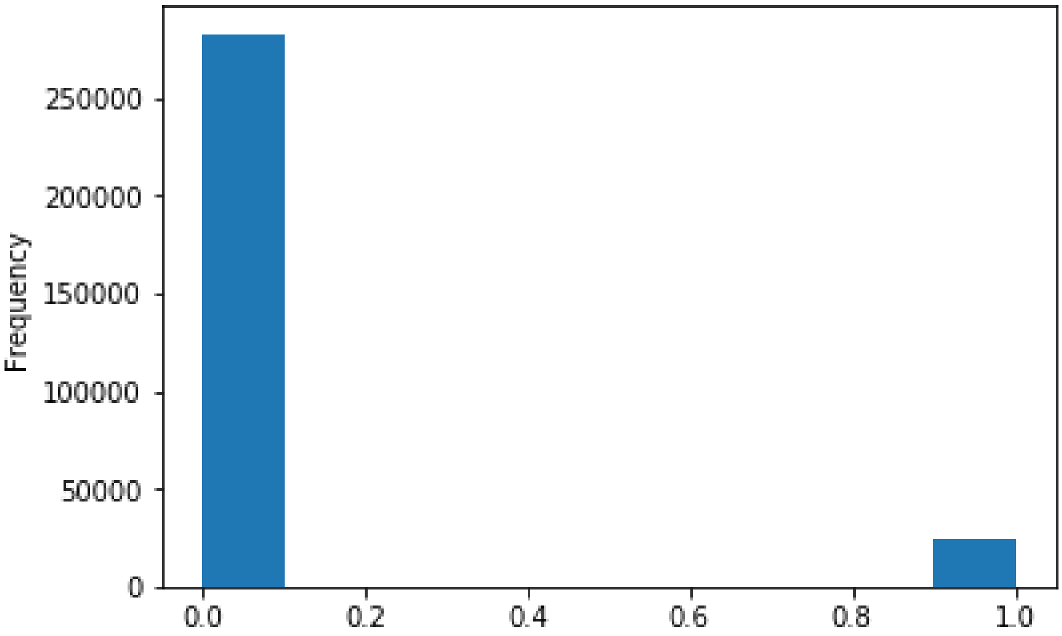
- **bureau.csv** – All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample). For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.

- **bureau_balance.csv** - Monthly balances of previous credits in Credit Bureau. This table has one row for each month of history of evey previous credit reported to Credit Bureau – i.e the table ahs (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
- **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.
- **credit_card_balances.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 – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) 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 previous 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 for each 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.

Here is a map of how the different files of the data set are related. As shown by the chart below the bureau_balance file contains informant connected to the bureau.csv which is then connected to the main application file. The previous_application file is supplemented by the POS_CASH_balance, installments_payments, and credit_card_balance files.


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

In the application_train file there are 307511 loans with 122 features including the TARGET. The data is unbalanced there are 282686 which have a TARGET value 0 indicating the loan was repaid on time, and 24825 loans that in which the client had difficulty paying back the loan.



Algorithms and Techniques

Logistic Regression Implementation