**HOME CREDIT DEFAULT RISK**

Title: Can you predict the loan repayment capability of an applicant?

Problem: 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 it would like to know whether we could predict an applicant’s loan repayment capability, in doing so it 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.

Who might care? Both involved parties- the financing instituions and customer/applicant- can benefit from such a model. Applicants who are likely to hesitate or avoid applying for loans because of misconceptions or stereotyped stigmas can be more confident knowing that such a predictive model will allow the financing instituion trust such an applicant and not pre-maturely decide an outcome. Both parties gain- the financier gains business and the customer is approved for a much needed loan.

Data: The Data has been provided by Home Credit, a financial institution who focuses on responsible lending to people with little or no credit history who may underserved by traditional banks. Seven different sources of data are provided:

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.

The following diagram demonstrates the relation between the data:



alt text

Also provided is a description file- HomeCredit\_columns\_description.csv- for the columns in the listed data files.

Modeling approach: A supervised classification algorithm will be the essential option for this predictive model to predict whether an applicant will default on a loan. Labels are included in the data thus the goal being to train the model to be able to predict the labels. This label is a binary variable with 0 referencing on time payment and 1 referencing delayed payment.

Possible limitations: Only foreseen possible limitation is incorrect or inconsistent data being provided that would not correlate between data sets to allow a proper analysis.

Deliverables:

Codes (notebooks) for: a. data acquisition

1. data cleaning
2. data exploration analysis
3. machine learning model development
4. random forest
5. grid and random search