1 Overview

The dataset comes from an international non-bank financial instituition. The task is to build a machine learning model based on this training dataset to predict whether to offer loans to clients, according to their information in the dataset. In addition to building a complete machine learning model, we also suggest you present your data mining skills (such as data pre-processing, visualisation and feature engineering). These techniques will play a major role in your PhD project.

2 Dataset description

There are two table files: `main_train.csv, additional.csv`.

1. The file `main_train.csv` contains the main table. It has static data for all clients, one row represents one client data. "SK_ID_CURR" is the id of the loan application and "TARGET" column is the prediction target (or called 'label' in supervised learning). Note that the information contained in the file are anonymised. For instance, you could find that "DAYS_BIRTH" is negative.

2. The file `additional.csv` contains credit history of the clients, provided by other financial institutions. For every loan in our sample, there are as many rows as number of credits the client had before the application date. You might use this dataset to get more features.

3 Submission requirements

1. You need to submit (1) one document and (2) all your codes. The document file describes your findings and presents related figures which you use to explain, e.g., your feature engineering process and the evaluation of your prediction on the training dataset.

2. We suggest you to use Python (with version >3.6) to perform this project. If really necessary, you can also use R.