Capstone Project

Machine Learning Engineer
Nanodegree
Meskhidze Konstantin
June 28, 2018

Definition

For the final project I've decided to join to one of competitions on Kaggle.com. It was launched by Home Credit Group, an international consumer finance provider. It focuses on responsible lending primarily to people with little or no credit history. 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.

In a nutshell I need to build a binary classifier that is used by the bank for a credit approval for people without any credit history.

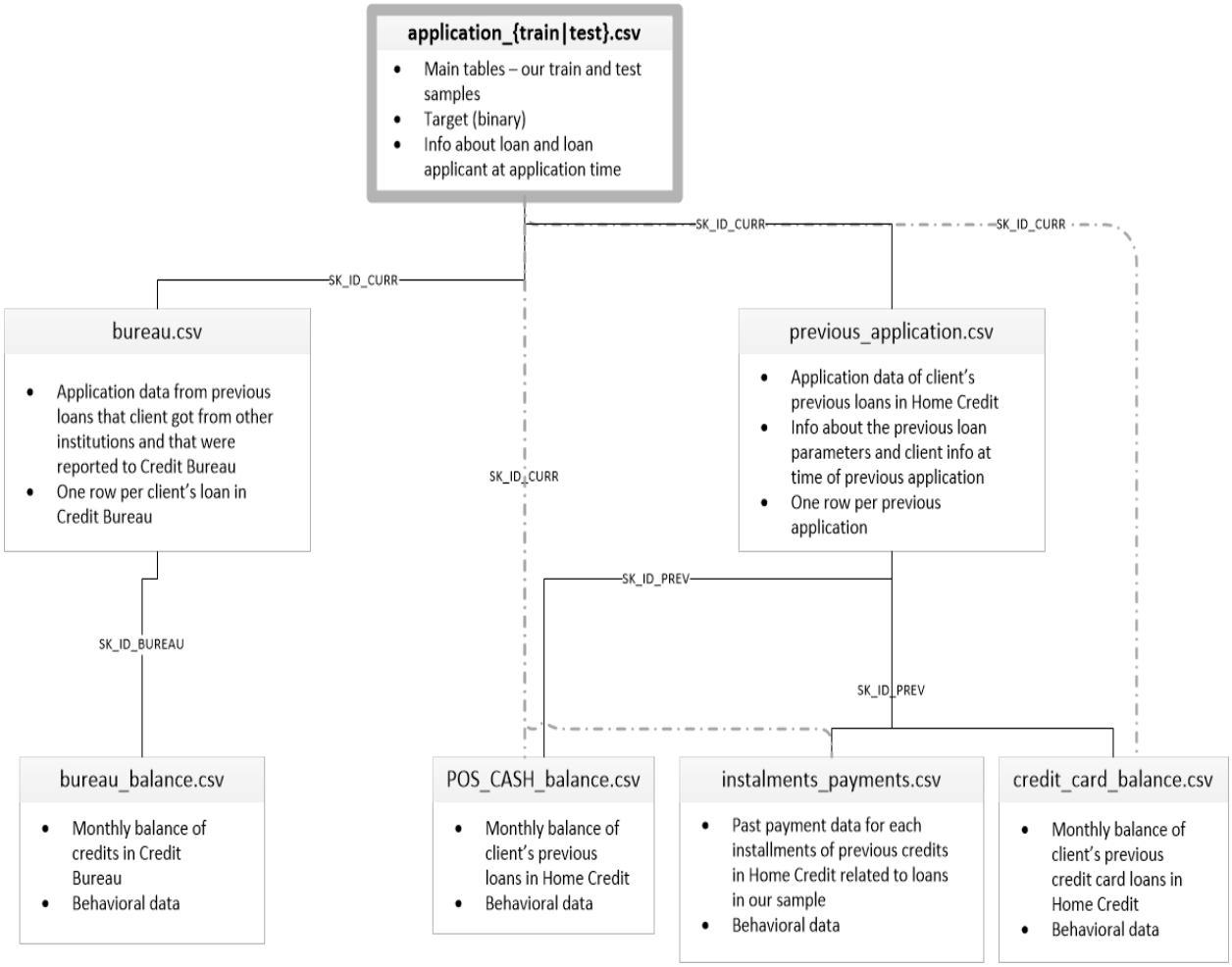
Data Description

There are the following files in a dataset:

- application_{train|test}.csv
  - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
  - Static data for all applications. One row represents one loan in our data sample.
- 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 every previous credit reported to Credit Bureau – i.e the table has (#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_balance.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 previously 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 each for 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.
- HomeCredit_columns_description.csv
  - This file contains descriptions for the columns in the various data files.



In the project I will be following the next process:

1. Data cleaning.
    - Dealing with missing data, NaN data.
    - Deleting outliers.
2. Encoding data.
    - Convert categorical data to numeric values
3. Calculate covariance matrix.
    - Variability comparison between categories of variables
    - Finding variables relationship to reduce features' dimention
4. Feature scaling and fitting.
5. Tuning functions.
    - Preparing functions for evaluation metrics for a model: confusion matrix, ROC curve.
    - Function for cross validation tuning
    - Function for GridSearch tuning
    - Function for RandomizedSearch tuning
6. Evaluating models.
    - Logistic Eegression
    - KNeighbors Classifier
    - Decision Tree Classifier
    - Random Forests
    - SVM
    - Boosting
    - Neural networks
7. Comparing models.
8. Predictions on a test set.
9. Making a conclusion.