Capstone Proposal:
Credit Risk Modelling

P. A. Ortiz Otalvaro

June 2019

Abstract

The objective of this project is to use loan data to predict whether or not an applicant will be able to repay a loan. Specifically, using historical loan applicaition data from Home Credit a set of models are built to determine credit risk for banks and supervisors to map risks correctly and consistently. Such models aim to reduce unwarranted variability in risk-weighted assets across banks

Introduction

Pillar 1 funds requirements is the minimum amount of capital a bank must hold by law. Banks can choose to use internal models to set out their own funds requirements. In 2015, the European Central Bank (ECB) decided to start a project to check Pillar 1 models at all directly supervised banks that use them. In this way the ECB aims to assess whether the internal models currently used by banks comply with regulatory requirements, and whether their results are reliable and comparable.

One of TRIM's major goals is to increase consistency and stability when banks calculate risk-weighted assets through their own internal models. This is because the evaluation of risk assets is an important metric that contributes in determining a bank's Pillar 1 funds requirements as it weighs its assets according to their riskiness.

All banks under the supervision of the ECB are therefore following a transformation in their models and have a clear need for accurate models that best match their in-house framework and the complexity of their investments. Particularly, in the Netherlands several banks are requesting consultants for internal TRIM preparation projects (for instance ING, ABN AMRO, and Rabobank). For them it is critical to get the approval of the European Central Bank on their internal models and thus the results obtained in this work would be of great use for them.

For more information see https://www.bankingsupervision.europa.eu/about/ssmexplained/html/trim.en.html

Business questions

Main questions/tasks:

1. Predict credit default: whether a loan will default. In other words: what is the credit risk (credit default) of a service given to a specific customer? (Credit risk or credit default indicates the probability of non-repayment of bank financial services that have been given to customers)
2. Predict: How much is the loss incurred when a loan defaults?

Other questions:

3. How can we measure the credit worthiness of the loan application over a time period? In other words, which are good techniques/algorithms/models to predict credit default ? (logistic regression, discriminant analysis methods, neural networks)
4. Which is the best model to predict credit default for a given institution? How can this model be selected? Which criteria can be used to select the best model for a given institution?
5. Which parameters/criteria are crucial in a model that predicts credit default? Which are the most important? Why?

Data

7 csv files with information related to Home Credit customer's past financial data:

1. *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.

2. *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.

3. *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.

4. *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.

5. *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.

6. *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.

7. *installments_payments.csv*

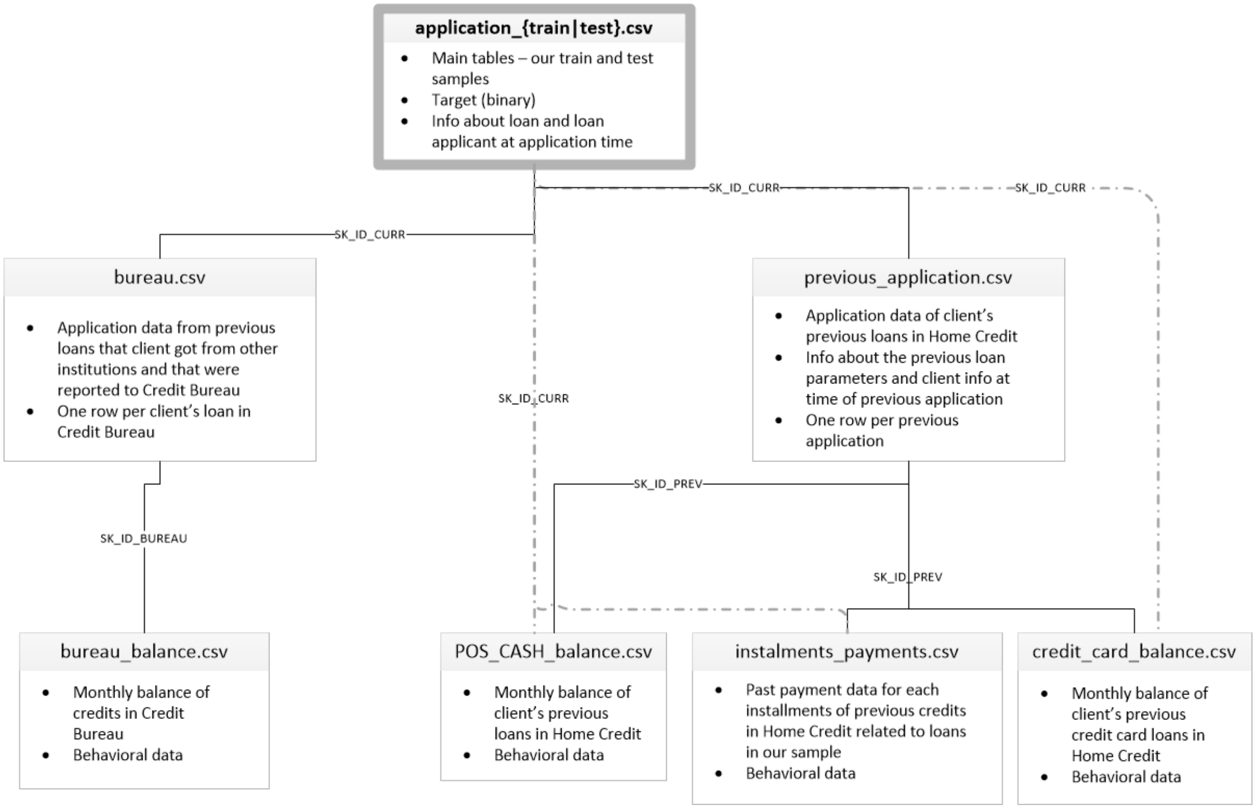
- Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
- There is
  a) one row for every payment that was made plus
  b) one row each for missed payment.
- One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.

8. *HomeCredit_columns_description.csv*

- This file contains descriptions for the columns in the various data files.

Source: https://www.kaggle.com/c/home-credit-default-risk/data

In a diagram:



application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

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

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

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

Approach

For now I present all possible approaches to the problem. Later on, after having worked with the data set and learned about risk models, the approach will be selected.

Previously used models

1. Credit risk analysis using deep learning and extreme gradient boosting This is a paper from the Bank of Greece. It doesn't give much detail on the algorithms. From this paper: "Given the extended number of employed predictors and the large scale dataset employed we resort to a methodology from the general domain of Machine Learning techniques called Extreme Gradient Boosting (henceforth XGBoost) and a Deep Learning Technique used to train, and _deploy deep neural networks (MXNET)." https://www.bis.org/ifc/publ/ifcb49_49.pdf

2. Credit risk analysis using logistic regression modelling "A loan officer at a bank wants to be able to identify characteristics that are indicative of people who are likely to default on loans, and then use those characteristics to discriminate between good and bad credit risks." https://www.smartdrill.com/pdf/Credit%20Risk%20Analysis.pdf

3. Credit risk prediction using artificial neural network https://www.datasciencecentral.com/profiles/blogs/credit-risk-prediction-using-artificial-neural-network-algorithm https://www.sciencedirect.com/science/article/abs/pii/S1062976907000762 https://www.neuraldesigner.com/learning/examples/credit-risk-management https://www.worldscientific.com/doi/abs/10.1142/S0129065709002014 https://link.springer.com/chapter/10.1007/978-3-319-72862-9_6

4. Credit risk model wiht random forest https://webofproceedings.org/proceedings_series/ECS/ICMCS%202019/icmcs02089.pdf http://article.sciencepublishinggroup.com/html/10.11648.j.ajtas.20150404.13.html

Deliverables

What comes to mind is a set of codes using different approaches to determine credit risk and a analysis over their results and evaluation of their performance giving a final advice to banks of which model is best given their specific criteria for loans.

- Script(s) used to execute the analysis/modelling
- Paper with analysis process and findings
- Slide deck: targeted to ING, ABN AMRO or Rabobank to persuade the client to implement the algorithm.
- GitHub Repository with all deliverables.