Capstone Proposal 1: Credit Risk Modelling

P. A. Ortiz Otalvaro

June 2019

## Abstract

Still written in bullet points as I don’t have a complete big picture of what the project will be.

* Build a set of models to determine credit risk for banks and supervisors to map risks correctly and consistently. Given the conditions of the bank.
* or test the models?  
  *Cross-Validation Techniques Can Lead to Substantial Improvements in Evaluating Predictive Modeling Efficacy*  
  *“I recently had the privilege of leading a team of talented people in Kaggle’s largest featured competition to-date where the objective was to predict the relative likelihood of default for a dataset of cash loans and revolving loans. In the competition, the team used a stratified k fold cross-validation (CV) approach with a constant seed. For stratified k fold validation, the training data is separated into k folds and the model is run k times. Each time, the model is trained on all the training data except one fold. The excluded fold changes each time and is treated as a validation set so that predictions are made on it. Those predictions can be stored before the next fold is run. At the end of this process, predictions are available for the whole of the training set. In the competition, the dataset was unbalanced. This means that there was not an even split between the number of people recorded as having defaulted (i.e., failed to pay on their loans), and those not in default (i.e., people still paying or who already paid off the loan). Less than 10% were in default. We decided that a stratified approach to k fold CV was prudent since it meant that the unbalanced distribution of the whole would be reflected in each fold.”* <https://www.datascience.com/blog/credit-risk-datasets-automated-testing-and-predictive-modeling>
* Such model aims to reduce unwarranted variability in risk-weighted assets across banks

## Introduction

* *“What is the problem you want to solve?”*
* *“Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?”*

PROBLEM: “*The targeted review of internal models, or TRIM, is a project to assess whether the internal models currently used by banks comply with regulatory requirements, and whether their results are reliable and comparable. One major objective of TRIM is to reduce inconsistencies and unwarranted variability when banks use internal models to calculate their risk-weighted assets (a commonly used regulatory metric that “weighs” a bank’s assets based on their riskiness and constitutes a key factor in determining the bank’s own funds requirements)."* *"Banks can use internal models to determine their Pillar 1 own funds requirements, i.e. the minimum amount of capital they must hold by law. Under the project, the ECB will check Pillar 1 approved models at all directly supervised banks that use them.*” <https://www.bankingsupervision.europa.eu/about/ssmexplained/html/trim.en.html>

CLIENT: I thought of this project because I have seen 3 banks in the Netherlands requesting consultants for internal TRIM preparation projects: ING, ABN AMRO, Rabobank. For them it is critical to get the approbal of the European Central Bank on their internal models

## Data

* *“What data are you going to use for this? How will you acquire this data?”*
* *“The project uses an existing data set. However, the data set is still raw, leaving some opportunities for wrangling, cleaning and other real-world challenges. More importantly, the student clearly states the limitations of this data set, given there’s not enough clinical context”*

I’ve found a few sources from which I could get the data:

* Open data room Banque de France: <https://www.banque-france.fr/en/statistics/access-granular-data/open-data-room>
* I like a lot the idea behind this database: <https://blog.okfn.org/2013/01/31/sovereign-credit-risk-an-open-database/>
* Data from credit risk’s book could be an option: <http://www.creditriskanalytics.net/datasets.html>
* Kagge data: <https://www.kaggle.com/c/home-credit-default-risk/data> <https://www.kaggle.com/uciml/german-credit>

## Approach

*“In brief, outline your approach to solving this problem (knowing that this might change later).”*

**Question: in my research I found different ways of approaching risk modelling. I do not know which one would be the best approach (since I do not know much of the topic or of data science algorithms). Do you perhaps have a suggestion for what would be best for me?**

1. 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>

## 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.

* code(s) used to execute the analysis/modelling
* Code to test/validate the model(s)
* 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.