Forecasting recovery rates on non-performing loans with machine learning

Anthony Bellotti[a], Damiano Brigo[a], Paolo Gambetti[b], Frédéric Vrins[b]

[a]*Department of Mathematics, Imperial College London, London SW7 2AZ, UK*
[b]*LFIN, UCLouvain, Louvain-la-Neuve, B-1348, Belgium*

Abstract

Reducing exposures to non-performing loans (NPL) is one of the major issues banks have been facing in the aftermath of the financial crisis. Although in principle this can promptly be achieved by selling defaulted exposures to specialized investors, several inefficiencies affect the NPL market. Two key factors pointed out by European regulators are the discrepancies in exposures' data availability and valuation methodologies between the bank and the investor. For example, while it is obvious that prices of NPL portfolios depend on the *recovery rate* achievable by the investor, this quantity is unknown at the time of purchase and needs to be predicted. To reduce bid-ask spreads, both parties should align on an effective modeling framework. In this paper, we investigate what are the best methods for predicting recovery rates on NPL using a private database from a debt collection agency. We benchmark 20 machine learning techniques belonging to the classes of linear, nonlinear and rule-based algorithms. Compared to previous literature, we introduce four new techniques: Gaussian processes, relevance vector machines, conditional inference trees and Cubist. Predictors include socio-demographic characteristics of defaulted clients, contract specifications and economic indicators at time of default. A special feature of our framework is that we also rely on the recovery data provided by the original loans holder such as the time-series of contacts to defaulted clients and clients' reimbursements to the bank. Using feature engineering, we seek to extract knowledge on clients' behavior that proxies for their ability or willingness to repay the debt and to characterize the pressure exercised by the bank in soliciting each client. We find these features help models to better identify debtors with different repayment ability and/or commitment, and in general with different recovery potential. Models having access to bank's recovery and contact history information exhibit better performances in predicting debt collector's recovery rates according to all performance measures. Variable importance metrics further emphasize the primary role of these predictors as well as the relevance of contract specifications. The capability of machine learning algorithms to isolate sub-groups of clients is another important quality to consider when modeling recovery rates. We find that rule-based algorithms such as Cubist, boosted trees and random forests outperform all the other methods in predicting debt collector's recovery rates. Performance improvements are registered across all model specifications and are more pronounced when bank's recovery and contact history is considered.

Abstract submitted to Credit Scoring & Credit Control XVI *March 21, 2019*