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

We compare the performances of a wide set of regression techniques and machine learning algorithms for predicting recovery rates on non-performing loans, using a private database from a European debt collection agency. We find that rule-based algorithms such as Cubist, boosted trees and random forests perform significantly better than other approaches. In addition to loan contract specificities, the predictors referring to the bank recovery process – *prior to the portfolio's sale* to the debt collector – are also proven to strongly enhance forecasting performances. These variables, derived from the time-series of contacts to defaulted clients and clients' reimbursements to the bank, help all algorithms to better identify debtors with different repayment ability and/or commitment, and in general with different recovery potential.

Keywords: Risk management, Recovery rate, Non-performing loans, Forecasting

1. Introduction

Lending is, by far, the primary business in retail banking. Whereas most loans are paid back in full and in due time, some others default, in the sense that the borrower violates the repayment schedule. The latter, commonly labeled as *non-performing loans* (NPLs), have been the focus of European regulators' attention in recent years, as many banks still face difficulties to dispose of those materialized on their balance sheets during the financial crisis. To limit impairment losses and financial stability concerns, regulators recommend banks to pool their NPLs and sell them to specialized investors, such as *debt collection* agencies. However, the prompt disposal of NPLs is hampered by the large bid-ask spreads characterizing their market, determined by discrepancies in data availability between banks and investors, and by poor valuation methodologies (ECB, 2017; ESRB, 2017, 2018).

One of the most important variables governing the price of NPLs portfolios is the *recovery rate*, the percentage of exposure that can be recovered from each borrower through

*Corresponding author: Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium. Tel.: +32 10 479 422.
E-mail addresses: a.bellotti@imperial.ac.uk (A. Bellotti), damiano.brigo@imperial.ac.uk (D. Brigo), paolo.gambetti@uclouvain.be (P. Gambetti), frederic.vrins@uclouvain.be (F. Vrins)

the debt collection process. The recovery rates achievable by the debt collector are unknown at the time of purchase of the portfolio and need to be predicted. Clearly, in order to fairly evaluate the portfolio, both parties should rely on a set of information that best allows identifying borrowers' recovery potential, and on effective forecasting methodologies. To that end, several regression methods have been proposed. More recently, machine learning techniques also started to be successfully applied to this field of research. However, most of the existing studies focus on corporate bonds or loans. Models focusing on retail credit products such as mortgages and credit cards still largely need to be investigated.

Several gaps in this specific stream of literature urge to be filled. For instance, while most references model the recovery rate obtainable by the same bank originating the loan, only the study of Ye and Bellotti (2019) focuses on the recovery rate achievable, at a second stage, by the specialized investor purchasing the NPL on secondary markets. More research is now needed in this latter direction, given the systematic character of the NPL issue. Similarly, there is no precise research quantifying the impact of asymmetries of exposures' information on recovery rates forecasting. Providing such a study would contribute to the regulator's objective of designing macro-prudential policies aimed at improving the functioning of NPL markets (ESRB, 2018).

The literature on retail loans also lacks an up-to-date benchmark study involving machine learning methods. The last research of this kind was conducted in Loterman et al. (2012), where the authors compared twenty-four different techniques and conclude that nonlinear and two-stage algorithms are associated with better predictive performances with respect to linear models used in previous analyses (Bellotti and Crook, 2007; Caselli and Querci, 2009). Nevertheless, they also document particular difficulties in forecasting recovery rates on retail loans, for which model performances are generally aligned across models and often unsatisfactory. Two-stage algorithms for consumer loans are also proposed by Bellotti and Crook (2012) and Ye and Bellotti (2019) to accommodate the multi-modal character of recovery rate distributions. The spectrum of machine learning algorithms involved in this literature is however still limited; many alternative models that have proven to be effective in other domains still have to be applied to the recovery rate problem.

Eventually, general data constraints – added to privacy protection norms and confidentiality policies promoted by banks – did not allow previous references to identify many potential predictors of recovery rates for retail loans. The literature is indeed very limited in this respect. For example, Grippa et al. (2005) and Querci (2005) understand the importance of taking regional differences into account on top of the loans characteristics or the debtor's socio-demographics which are typically used for credit scoring. Besides account's information, Bellotti and Crook (2012) further consider three macroeconomic variables measured at the time of default to model credit cards' recovery rates in the UK. They report modest improvements in forecasting performances for models including macroeconomic fac-

tors and stress that their effectiveness is strictly dependent on having enough training data to span the entire business cycle. As for the inclusion of macroeconomic factors, we observe that predictive models for retail loans' recovery rates have not reached the same level of complexity of the most recent bond-related models (Nazemi and Fabozzi, 2018; Nazemi et al., 2018).

Another critical reflection of Bellotti and Crook (2012) is the possibility of extracting information on clients' recovery potential from the account behavioral patterns observed during the debt repayment period. This is partially investigated in Ye and Bellotti (2019) thanks to a private database from a debt collection agency that purchased a large portfolio of NPLs from a European bank. The authors show that pre-purchase recovery rates and the total number of calls between the bank and defaulted clients are significant predictors of post-purchase recovery rates. Despite these novel results, however, their study does not fully exploit account behavior data in the spirit of what Bellotti and Crook (2012) suggest. Indeed, the literature still lacks a systematic study that takes into account the behavioral *dynamics* of reimbursements and/or contacts observed on defaulted accounts during the debt repayment period.

The objective of our paper is to fill all the aforementioned gaps. We structure this research as a benchmark study of machine learning methods for forecasting NPL recovery rates. This approach, together with the unique proprietary dataset in our possession, allow us to provide two key contributions to this field of research.

First, we consider four new techniques for recovery rates forecasting: Gaussian processes, relevance vector machines, conditional inference trees, and Cubist. The latter are increasingly popular in the machine learning community and have been successfully applied in various fields, such as image processing, robotics, mechanical engineering, and epidemiology. We benchmark a total of twenty different models belonging to the classes of linear, nonlinear and rule-based algorithms. In this respect, we find that the capability of rule-based algorithms to isolate sub-groups of clients is an important quality to consider when modeling retail loan recovery rates. Cubist, boosted trees and random forests outperform all the other methods in predicting debt collector's recovery rates. This evidence is preserved across all model specifications and is found to be statistically significant by the model confidence set procedure.

Second, we extend the spectrum of retail loan recovery rates predictors in two directions. On the one hand, we enlarge the number of economic indicators involved in the models. We consider more than one-hundred financial, macroeconomic, and housing market indicators measured at the time of default for each loan. We find the latter two types of indicators to be always selected as important predictors of recovery rates. On the other hand, and most importantly, we enhance the framework of Ye and Bellotti (2019) by also taking into account the dynamics of the time-series of contacts to defaulted clients and clients'

reimbursements to the bank, *prior to the portfolio's sale* to the debt collector. 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 that these features help all models to better identify debtors with different repayment ability and/or commitment, and in general with different recovery potential. Models having access to the bank recovery process information exhibit better performances in predicting debt collector's recovery rates according to all performance measures. Variable importance metrics further emphasize the role of these predictors, as well as the one of contract specifications. We eventually bring evidence that rule-based algorithms should be preferred to the other methods when recovery process data are not disclosed by the selling bank. Their forecasting ability, although is impacted, remains very performing.

The remainder of this paper is structured as follows. Section 2 includes a description of the data. Section 3 includes an overview of the different algorithms we employ, together with the relative model specifications. Section 4 discusses the results of our study. Section 5 concludes.

2. Data

The dataset involved in this study refers to a large transaction of NPLs between a European bank (on the sell-side) and a debt collection agency (on the buy-side)[1]. All exposures are entirely represented by defaulted consumer loans: mortgages and credit cards. The data we consider represent an extension of the sample used in Ye and Bellotti (2019) and derive from the connection among four different databases that we now describe.

2.1. Socio-demographic and loan file data

The first database, partly retrieved from the reference state's Bad Debt Bureau, includes socio-demographic and loan file information for the defaulted borrowers. The main socio-demographic features included in this database are the borrowers' age, gender, marital status, as well as their province of residence. All borrowers have the same nationality and are representative of all state's provinces. The database also provides further personal details that are instead not used due to excess of granularity (e.g. city of residence, postal code), class imbalance (e.g. language spoken) and redundant content (e.g. title vs gender).

Loan file information contained in this first database refers to the credit application process and the default event. As for loan applications, we recover contractual details such as the type of credit product (e.g. mortgages vs credit cards), the amount of principal and

[1]Due to a confidentiality agreement, we cannot mention the name of the two parties, nor their specific country of domicile.

interest, the over-limit and late payment fees, as well as the account open date. Similarly, we can access detailed features that express borrowers' credit risk: these include the Credit Bureau Score, credit limit and an indicator signaling whether an employer reference was provided at the time of application. As for the information related to default instead, we can retrieve the loans' default date and the outstanding exposures at the time the debt collector purchased the loans.

2.2. Bank recovery history

The second database contains information relative to the recovery process promoted by the bank that originated the loans, before the portfolio is sold to the debt collector. For each borrower, the database includes monthly summaries of net repayment amounts computed as the difference between the borrower's i repayment and the administrative fees registered at a given month t:

$$NRA_t^i = P_t^i - A_t^i \, . \tag{1}$$

Additionally, the database also includes information on the evolution of borrowers' outstanding exposure during the bank recovery period. Given this information, we hence define the recovery rate obtained by the bank on a given loan as:

$$RR_i^{Bank} = \frac{\sum_{t=t^i}^{\tau} NRA_t^i}{OB_{max}^i} \, , \tag{2}$$

where t^i denotes the start of borrower i's recovery process, τ is the selling date of the NPL, and OB_{max}^i represents the maximum outstanding balance recorded for the i-th borrower during the bank recovery process[2].

Furthermore, this second database also provides monthly information on the number of calls, visits and contacts between the bank and the defaulted borrower. As will be explained in section 3.1, many variables of interest for this study will be feature engineered from those latter.

2.3. Debt collector's recovery rates

The third database includes information on the recovery amounts obtained by the debt collector which purchased the NPLs from the originating bank: only positive recovery amounts are registered. The main quantity of interest derived from this database is the recovery rate obtained by the debt collector, after the purchase of the NPLs. We compute it as the sum of recovery payments made by a given borrower i, between the time of purchase

[2]We consider the maximum instead of the initial balance because the exposure is not monotonically decreasing for each NPL. This would be the case, for instance, if additional penalty fees are debited to the defaulted borrower during the bank recovery process.

of the NPL τ and the last repayment date T, and divided by the outstanding balance measured at the time of purchase of the NPL:

$$RR_i^{Debt\ collector} = \frac{\sum_{t=\tau}^{T} P_t^i}{OB_\tau^i} \ . \tag{3}$$

The recovery rate obtained by the debt collector is the target variable for our study.

2.4. Business cycle variables

Eventually, in order to capture the systematic component underlying recovery rate variations, we connect a fourth database of business cycle variables. Data are compiled by the reference state's National Bank and comply with the international quality standards established in the ESCB's Public Commitment on European Statistics. The original database includes more than 400 macroeconomic, financial, structural and housing market indicators for the reference state, measured at different frequencies.

Additionally, we also consider a news-based measure of economic-related uncertainty for the Euro area (Baker et al., 2016) and two uncertainty measures specific to the reference state. As it is shown by several references, in fact, economic uncertainty is a key determinant of the economic outlook (Kose and Terrones, 2012; ECB, 2016; Gieseck and Largent, 2016; Ludvigson et al., 2019). Measures of economic uncertainty have also proved to be particularly effective to model bond recovery rates (Gambetti et al., 2019). Their usefulness for retail loans is instead first investigated in this study.

2.5. Dataset definition and pre-processing

We merge the first three databases using borrowers' identifiers. 34,807 identifiers from the second database can be matched with the ones included in the first one. These are the NPLs for which some recovery attempt was registered by the bank (independently of whether the attempt was successful or not). Of these, only 10,232 can also be matched with the third database which includes debt collector recovery information.

We instead connect systematic variables to the other data using the borrowers' default dates[3]. A visual summary of the loans origination, default and recovery periods for the NPL data we consider is included in Fig. 1.

The resulting dataset is then submitted to preprocessing. Several steps are needed to correct wrongly encoded information, to remove missing values, and to discard uninformative or zero-variance predictors. We also filter the predictors' space to remove variables

[3]Given that not every measure is available for each default date, we need to subset the sample. In this respect, the knowledge provided by industry participants suggests us to prioritize the availability of housing market indicators, which are mostly accessible after March 2007. We hence retain only the loans defaulted after that date.

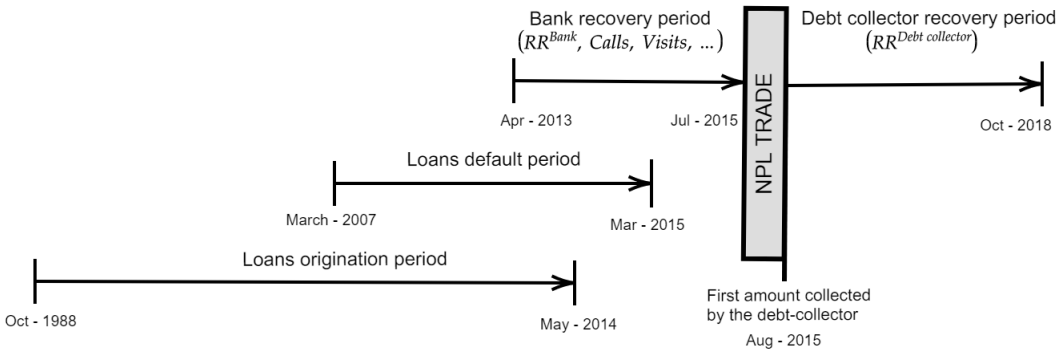

Figure 1: Time lines of loans origination, default issues and recovery periods. All borrowers opened a loan account between October 1988 and May 2014, which subsequently defaulted between August 1990 and March 2015. After default on a given loan, the originating bank started an internal recovery procedure. The bank eventually interrupted all recovery processes in August 2015 and sold all NPLs to an external investor, a specialized debt collection agency. This entity tried to recover the maximum amount out of the remaining outstanding balance for each NPL. The debt collector recovery process covers more than three years.

with perfect collinearity. Categorical predictors are eventually dummy coded. After pre-processing, the dataset features 10,152 recovery rate observations and 344 predictors. The definition of training and test samples is then undertaken using a standard 75%–25% random split. Summary statistics for both samples of recovery rates are reported in Table 1.

Table 1: Summary statistics of $RR^{Debt\ collector}$ for training and test sets.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Training set	7,614	0.395	0.298	0.0001	0.107	0.630	1.000
Test set	2,538	0.408	0.303	0.0002	0.115	0.650	1.000

We now proceed with the description of the methodology.

3. Methodology

3.1. Feature engineering on bank recovery information

We assume that much of the information about recovery potential during the debt collector recovery period can be extracted from the variables relating to the bank recovery process. By applying feature engineering, we seek to capture behavioral patterns that relate to the capacity and commitment of each borrower in reimbursing the defaulted debt. By investigating the dynamics of interactions between the bank and defaulted clients instead, we attempt to extract knowledge on the difficulty of imposing repayments on the latter. Similarly, we can gain insights about the approaches the bank used for softening potential clients' resistances, and about the relative spillover effects during the debt collector recovery period.

For each exposure, we hence undertake feature engineering to derive potentially meaningful quantities, such as the average and the variation of monthly repayments and/or calls, visits or contacts. Further examples include these variables' minimum and maximum, their first and last values, the total number of repayments, or the time, out of the total recovery period, in which the bank had the initiative of contacting the client. To the best of our knowledge, none of these features was considered in previous studies.

3.2. Model specifications and training

The identification of the best class of machine learning methods for forecasting NPL recovery rates is undertaken through a benchmark study. We compare 20 models belonging to the classes of linear, nonlinear and rule-based algorithms. In order to assess the predictive power of the bank recovery process information, we train each model using two different specifications of the predictors' space. The first specification (*specification 1*) represents the full space of predictors described above: socio-demographic variables, loan file information, bank recovery period information and systematic variables. As for the second specification (*specification 2*) instead, we define it by removing from the predictors' space all the variables related to – or derived from – the bank recovery process.

When required, model hyper-parameters are tuned by 10-fold cross-validation on the training set. In these cases, mean squared error (MSE) is used as cost function. Training routines for all models can be efficiently reproduced through the latest version of the `caret` R library (Kuhn, 2008; Kuhn and Johnson, 2013; Kuhn, 2018). We provide more details on the specific algorithms we used in Table A1 in the appendix.

3.3. Linear models

We consider seven linear models in this study: ordinary least squares regression (OLS), OLS regression with backward stepwise selection, two different versions of ridge and lasso regression, and the elastic net regression. All these linear models can be defined as particular configurations of the following minimization problem:

$$\arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right) \tag{4}$$

where $Y = (Y_1, ..., Y_N)$ denotes the vector of training set observations, X denotes the $(N \times p)$ model matrix of regressors and $\beta = (\beta_1, ..., \beta_p)$ denotes the vector of unknown regression coefficients. Different specifications of the penalty factor $\lambda \geq 0$ and the mixing factor $0 \leq \alpha \leq 1$ give rise to different models. In particular, we have that:

- $\lambda = 0$ defines the standard ordinary least squares regression (OLS) where the full model matrix X is considered. Backward stepwise selection is an iterative procedure where the predictor that has the least impact on the fit is sequentially removed from the model.;

- $\lambda > 0$ defines a penalized model. By specifying $\alpha = 0$, the model coincides with the ridge regression. By specifying $\alpha = 1$, the model generates the lasso. Elastic net regression is instead a weighted average of ridge and lasso where $0 < \alpha < 1$.

The OLS model does not feature any hyper-parameter and it is tuned on the overall training set. The maximum number of predictors in the backward stepwise selection is tuned by cross-validation. The same applies to the hyper-parameter λ in ridge and lasso regression. As for these latter, we also take a *heuristic* choice of λ based on the "one standard error rule" (Hastie et al., 2009): we choose the highest lambda (i.e. the most regularized model) for which the corresponding MSE is within one standard error of the optimum MSE. As for the elastic net, the hyper-parameters λ and α are also tuned by cross-validation.

3.4. Nonlinear models

3.4.1. Multivariate adaptive regression splines (MARS)

MARS (Friedman, 1991) is an adaptive algorithm which involves piecewise linear transformations of the original predictors. Each predictor X_j is expanded into a set of *reflected pairs* that is determined by specific cut points t according to:

$$(X_j - t)_+ = \begin{cases} X_j - t, & \text{if } X_j > t \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad (t - X_j)_+ = \begin{cases} t - X_j, & \text{if } X_j > t \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $j = (1, 2, ..., p)$ and $t \in \{x_{1j}, x_{2j}, ..., x_{Nj}\}$. A standard linear regression model is then created by applying forward-stepwise selection on the elements of the set $\mathcal{H} = \{(X_j - t)_+, (t - X_j)_+\}$, that is, the set of all predictor/cut point combinations.[4] To avoid overfitting, the MARS algorithm further applies a backward deletion procedure to the individual features that, if removed, are associated with the smallest error rate. The best model of each size is then produced. In this study, we apply a MARS algorithm of degree one and we tune the number of terms to be retained in the final model by cross-validation.

3.4.2. K-nearest neighbors

K-nearest neighbors is a non-parametric method that produces a prediction of the target variable by using the average of the K training observations that are closest in the input space. A new prediction is hence computed as:

$$\hat{f}(x) = \frac{1}{K} \sum_{x_i \in N_K(x)} y_i \ , \tag{6}$$

[4]The *degree* of the MARS algorithm represents the number of reflected pairs that are selected for each predictor during the forward-stepwise procedure. Multiple reflected pairs for the same predictor can be selected by MARS models of *degree* higher than one.

where $N_K(x)$ denotes the neighborhood of x, the set of the K closest observations in the training set. The notion of closeness between samples is generally based on the Euclidean distance. The parameter K, denoting the size of the neighborhood, can be tuned by cross-validation.

3.4.3. Model averaged neural networks

Model averaged neural networks (Ripley, 1996) is an ensemble method where the outputs of multiple neural networks are averaged to form a unique prediction. Neural networks composing the ensemble are initialized by using different starting values for the parameters to estimate. This moderates the effects of the back-propagation algorithm's convergence to local optima, and reduces the model variance with respect to that of a single neural network that may tend to over-fit the data. Individual neural networks can also be modified to include a *weight decay* λ that penalizes large coefficients. This amounts to solving the following minimization problem:

$$\underset{\beta,\,\gamma}{\arg\min} \sum_{i=1}^{N}(y_i - f_i(x))^2 + \lambda \sum_{k=1}^{h}\sum_{j=0}^{p}\beta_{jk}^2 + \lambda \sum_{k=0}^{h}\gamma_k^2 \tag{7}$$

where $f_i(x)$ is the i-th fitted value from a neural network model involving p predictors and h hidden units. In this work we apply model averaging by using 5 neural networks with weight decay. The number of hidden units and the weight decay parameter are both tuned by cross-validation.

3.4.4. Support vector machine

Support vector regression (Vapnik, 1995) is a kernel-based algorithm that has proven to be particularly effective to limit the effect of outliers on the model fit. Coefficients for support vector regression can be estimated starting from the following minimization problem:

$$\underset{\beta_0,\,\beta}{\arg\min}\; C\sum_{i=1}^{N} L_\epsilon(y_i - f(x_i)) + \sum_{j=1}^{p}\beta_j^2\;, \tag{8}$$

where L_ϵ is an ϵ-insensitive loss function and C is the penalty assigned to residuals of size larger or equal to ϵ. The solution of (8) can be written in terms of a set of unknown weights w_i, and a positive definite kernel function $K(\cdot)$ that depends on the training set data points:

$$f(x) = w_0 + \sum_{i=1}^{N} w_i K(x,\, x_i)\;. \tag{9}$$

Training samples associated to non-zero weights (i.e. the *support vectors*) determine the model fit. The most common choice for the kernel function is the radial basis kernel

$K(x,\, x') = \exp(-\sigma\|x - x'\|)$. We estimate the scaling parameter σ following the methodology of Caputo et al. (2002); we instead tune the cost parameter C via cross-validation.

3.4.5. Relevance vector machine

Relevance vector regression (Tipping, 2001) is a kernel method whose functional form is identical to (9) but where the model weights are estimated using a Bayesian learning framework. This approach has the advantage of leading to much sparser models than SVM regression because the posterior distribution of the weights is essentially zero in many cases. Additionally, it leads to a probabilistic interpretation of the model predictions. In particular, the predictive distribution of the target variable y^* for a new input vector $x*$ is given by:

$$p(y^*|x^*,\, X,\, y) \sim \mathcal{N}(\mu^T\phi(x^*),\; s^2 + \phi(x^*)^T\Sigma\phi(x^*)) \;, \tag{10}$$

where μ is the vector of posterior mean weights, $\phi(x^*) = [1, K(x^*, x_1), ..., K(x^*, x_N)]^T$, s^2 is the estimated variance of the output noise, and Σ takes account of the uncertainty around the model weights. The radial basis function is the standard choice for the kernel, but unlike SVM, the RVM kernel function is not constrained to be positive definite.

3.4.6. Gaussian processes

Gaussian processes (Williams and Rasmussen, 1996) is another Bayesian kernel method that can be considered as a non-sparse non-parametric generalization of the RVM model. In contrast to the RVM in fact, Gaussian processes impose a prior distribution directly on the function values (instead of the model weights). The joint prior distribution of the function values is assumed to be Gaussian with zero mean and covariance matrix equal to the kernel matrix $K_{ij} = K(x_i, x_j)$. Gaussian processes are entirely determined by this covariance matrix, the underlying assumption being that samples that are close in input space should also have similar values for the target variable. Given a new input vector x^*, the approach leads to the following predictive distribution of the target variable y^*:

$$p(y^*|x^*,\, X,\, y) \sim \mathcal{N}(\phi(x^*)^T K^{-1} y,\; s^2 + \phi^*(x^*) - \phi(x^*)^T K^{-1}\phi(x^*)) \;, \tag{11}$$

where $\phi(x^*)$ and s^2 have the same meaning as above and $\phi^*(x^*) = K(x^*, x^*)$. We implement Gaussian processes by using radial basis kernel in this study.

3.5. Rule-based models

3.5.1. Regression trees

By using conditional statements, regression trees (Breiman et al., 1984) partition the predictors' space into a set of non-overlapping regions and fit a simple model to each of them. In their simplest version, they fit an intercept-only model which amounts to use the average of the target variables associated with each region. To build the regions, the

algorithm employs a top-down recursive partitioning. It starts with the full dataset and divides it into two groups according to the predictor/cut point combinations that achieve the largest decrease in RSS. The process is recursively applied within the new regions until a certain stopping criterion is satisfied, such as the number of samples in the terminal nodes. To limit over-fitting, the tree is then pruned back by using *cost-complexity* pruning. The amount of regularization is determined by a complexity parameter that can be tuned via cross-validation.

3.6. Conditional inference trees

Regression trees often suffer from selection bias: predictors featuring a higher number of candidate cut points have a higher probability of being chosen during the tree growing step. Conditional inference trees (Hothorn et al., 2006a) have been conceived to overcome this limitation. For each predictor, the algorithm employs statistical hypothesis testing to assess the difference between the means of the two samples created by a candidate split. To reduce the selection bias for highly granular predictors, multiple comparison corrections are applied (Westfall and Young, 1993). The p-value threshold determining the implementation of a new split, as well as the hyper-parameter controlling the tree maximum depth, can be tuned by cross-validation.

3.6.1. Bagged trees

Bagged trees (Breiman, 1996) is an ensemble method resulting from the aggregation of the outputs of multiple regression trees that are trained on bootstrapped versions of the training set. Given a new sample of predictors, the predictions of the individual trees are then averaged to deliver a unique predicted value for the target variable. The individual trees composing the ensemble are generally not pruned, which results in low bias but high variance. The latter concern is then mitigated by the aggregation effect. The number of trees composing the ensemble, equal to the number of bootstrap samples that need to be generated, can be tuned by cross-validation. However, model training can become computationally expensive for large datasets.

3.6.2. Random forests

Random forests (Breiman, 2001) is a rule-based algorithm that was conceived to overcome the problem of high correlations among individual trees in bagged models. For bagged trees, in fact, all predictors are considered at each split during the tree-growing process, which translates in trees with very similar structures (especially in the top nodes). For random forests instead, only a subset of $m < p$ randomly selected predictors is considered at each split. This reduces tree correlation and also the variance of the ensemble prediction. The number of randomly selected predictors for this model can be tuned by cross-validation.

3.6.3. Boosted trees

Contrarily to bagging and random forests, boosted trees is an ensemble method where the base learners (i.e. regression trees) are *sequentially* fitted. The model is based on the gradient boosting machines algorithm (Friedman, 2001), a powerful procedure where the model residuals are iteratively fitted by many weak learners. To avoid overfitting, only a percentage of each fitted value (called *learning rate*) is added to the residual from the previous learner. The main hyper-parameters for this model are represented by the number of boosting iterations (equal to the number of trees), the learning rate, and the individual trees' depth. Stochastic gradient boosting is an improved version of the algorithm that also includes a random sampling scheme of the training data at each iteration step.

3.6.4. Cubist

Cubist (Quinlan, 1993) is a rule-based algorithm that extends the M5 model tree approach (Quinlan, 1992) with features borrowed from boosting and K-nearest neighbors. Similarly to M5, cubist features a tree-structure where each node contains a linear regression model. The predictors of the linear models are the same variables that satisfy the rule defining a specific node. After the tree is grown, the fits in each node are recursively *smoothed* by using the fit from the corresponding parent node.[5] Rules are then pruned and/or combined using the adjusted error rate criterion as in M5. Given a new sample, the cubist model computes a prediction as the average of the models from all the corresponding parent rules. Eventually, it can adjust the prediction using a weighted average of sample neighbors, were the weight attributed to each neighbor is proportional to the distance in input-space. Committees can be created by connecting several model trees in a particular boosting-like framework, that iteratively try to correct for positive/negative prediction errors. The number of neighbors for instance-based correction and the number of committees can be tuned via cross-validation.

3.7. Performance assessment

The best configuration of each model is used to forecast the sample of recovery rates in the test set. Model performances are assessed according to three different performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R^2, where the latter is computed as the squared correlation between the predictions $\hat{Y}$ and the corresponding test set observations Y. Additionally, we complement the comparison of performance metrics with the model confidence set (MCS) procedure proposed by Hansen et al. (2011).[6]

[5] Smoothing involves a linear combination of the two models where the one with the smallest RMSE has the largest weight. See Quinlan (1992) and Kuhn and Johnson (2013) for details.

[6] In particular, we refer to the algorithmic implementation by Bernardi and Catania (2018) included in the `MCS` R library. By default, the MSC procedure is undertaken with a value of $\alpha = 0.15$ but this quantity can be changed by the user.

Given a collection of competing models $\mathcal{M}^0$, the MCS procedure aims at identifying a *superior set of models* $\widehat{\mathcal{M}}^*_{1-\alpha} \subseteq \mathcal{M}^0$ with a confidence level $1-\alpha$. The procedure is based on sequential hypothesis testing and on the null hypothesis of equal predictive ability for the set of models under consideration. At each step, a model associated with poor forecasting performances is removed from the set of candidate best models. The test statistic and the elimination rule both depend on an arbitrary loss function $\mathcal{L}(Y, \hat{Y})$, which allows the procedure to be applied to a wide spectrum of problems. The strongest result for this procedure is represented by the case in which only one model belongs to the superior set of models.

4. Results and Discussion

4.1. Benchmarking regression algorithms

We visually document the out-of-sample performances of our algorithms in the right column of Fig. 2. White dots refer to the case in which models were allowed to access all types of variables during training, there included the ones feature engineered from the bank recovery process information (i.e. specification 1 models). Black dots instead refer to the case where models were allowed to learn from all variables except from those referring to the bank recovery process (i.e. specification 2 models). Following previous references, we mainly discuss model performances in terms of R^2 figures in what follows. In fact, while the RMSE is more expressive of forecasting performances, the R^2 remains the most intuitive measure of explanatory power. Our observations in terms of R^2 values remain valid in the case we use RMSE metrics.

By analyzing the metrics profiles for specification 1, we first notice that four models stand out for their better forecasting performances. Cubist, random forests and boosted trees (with and without random sampling) have a clear advantage on the other models in terms of lower RMSE and MAE values. They also feature a sensibly higher proportion of explained variation, with R^2 measures ranging from 18.54% to 19.79%. The gap with the remaining group of models is noticeable, with the latter barely reaching R^2 values of 15.52% and 15.70% associated with MARS and bagged trees respectively. We also find that the R^2 values of the algorithms trained on specification 1 are generally superior to those obtained in previous analyses on personal loans, and regardless of the fact the latter were based on linear regression (Bellotti and Crook, 2012; Leow et al., 2014) or more refined methods (Loterman et al., 2012).[7] This is a first indication of the importance of involving bank recovery period information in the modeling exercise by the debt collector.

[7]For example, when modeling personal loans with linear regression, Bellotti and Crook (2012) and Leow et al. (2014) obtain test set R^2 measures of 11% and 14.28%, respectively. Similarly, Loterman et al. (2012) reach 13.79% of R^2 using a combination of linear regression and support vector machines.

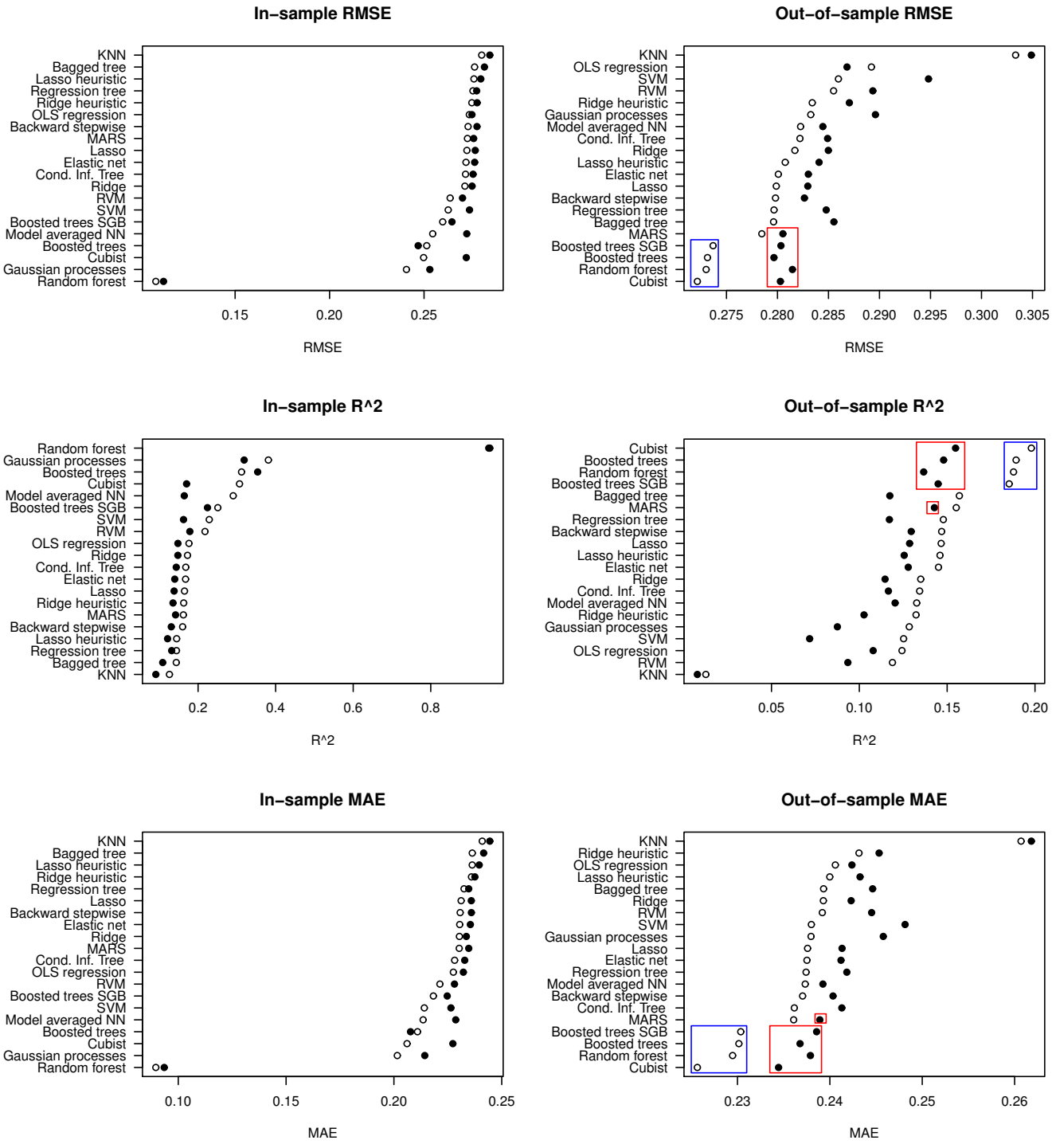

Figure 2: Comparison of model performances across models and model specifications on the test set. Specification 1 corresponds to the white dots, specification 2 to the black dots instead. Blue rectangles identify the superior set of models tested on specification 1 using the squared errors as loss function. Red rectangles instead represent the superior set of model tested on specification 2.

Moreover, when we use the squared error loss function in the model confidence set procedure, the superior set of models (depicted by blue squares) on the test set precisely coincides with Cubist, random forests and the two versions of boosted trees.[8] These four rule-based algorithms are hence significantly associated with better forecasting performances and should be preferred to other methods when the debt collector needs to forecast NPL recovery rates.

This result also essentially suggests that *groups* of borrowers with substantially different recovery potential are included in the portfolio and that once a group is determined, the recovery potential of the borrowers inside the group is relatively homogeneous. Considering the set of predictors our models are allowed to access during training, we expect this segmentation to be largely derived on the basis of behavioral features emerged during the bank recovery process. The results of specification 2 models, discussed below, seem to point in that direction.

As we expected in fact, when we prevent the algorithms to access the predictors derived and/or referable to the bank recovery process, we observe a sudden deterioration in all model performances.[9] This difference is suggestive of the high informational content all models were deprived of, prior to training them.

Nevertheless, we observe that Cubist and the two versions of boosted trees are again associated with lower RMSE and MAE values, and larger R^2 metrics with respect to those of other regression methods. We also find the MARS algorithm to be competitive in this respect, in what it even outperforms random forests.

In terms of explanatory power, the best four models trained on specification 2 feature test set R^2 ranging from 14.28% for MARS to 15.48% obtained by Cubist. We now observe that most algorithms exhibit R^2 values much more aligned with the ones reported by the aforementioned references which did not exploit our particular set of predictors.

On this specification, the model confidence set procedure with squared error loss identifies a superior set of models (represented by the red squares) still composed by Cubist, boosted trees and random forests, and augmented with MARS. However, we can also observe that the advantage of the superior set of models, with respect to the other group, is less clear cut compared to the previous case. In fact, we report a spread in the test set R^2 of 18.54% to 15.70% (between boosted trees with stochastic gradient boosting and MARS) for specification 1, and only a spread of 13.37% to 12.95% (between random forests and linear regression with backward selection) for specification 2. Similar evidence also applies in terms of forecasting performances, quantified by the other two metrics.

In particular, we observe that most of the largest losses in forecasting power are associ-

[8] In the case we use the absolute error instead, only Cubist is identified as the single best performing model.

[9] This evidence applies to all algorithms with the exception of OLS regression, whose RMSE decreases. This means that the model is now better at capturing extreme observations.

ated with the four rule-based algorithms that best performed on specification 1. It appears that these models lost most of the relevant information that was determining their advantage. We notice that declines in forecasting performances are also particularly evident for other two rule-based models: regression trees and bagged trees. Hence, while all models seem to be affected by the lack of relevant information about borrowers' repayment ability and/or commitment, the methods that *group* borrowers based on those features are the most affected ones.

Despite this, the superiority of rule-based algorithms in forecasting these data remains evident, especially in the case of models relying on ensembles and tree-like structures.[10] The practical consequence is that, when forecasting NPL recovery rates, debt collector agencies can still rely on the ability of these models to correctly identify sub-groups of clients with different recovery potential, even in the case bank recovery information is not available. While the segmentation would be performed less efficiently in this case, the specific architecture of rule-based ensembles still represents an advantage with respect to that of other approaches merely implying linear or non-linear relationships.

4.2. Variable importance rankings

The evidence reported so far suggests that groups of clients with different recovery potential can be discovered much more efficiently when the algorithms can access bank recovery information; but also that some adequate segmentation can still be made without those variables.

In this section, we answer the question on what are the specific predictors used by our algorithms and what is their relative importance inside the model. We provide this evidence in Fig. 3, which includes the top 20 variable importance rankings for the three models that best performed on the test set. Given that we use model-specific variable importance metrics, and for the sake of interpretability, all values are scaled relative to the largest inside each model.[11]

Variable importance rankings for specification 1 models are displayed in the left column of Fig. 3. In this case, we observe that the top 20 ranking is always composed by a mixture of bank recovery process information and loan file data. The recovery rate obtained by the bank, as well as the outstanding balance of the NPL when the bank started the recovery process, seem to play a particularly important role among the former type of variables. Compared to Cubist, random forest and boosted trees seem to make larger use of the information relating to the dynamics of borrowers' repayments to the bank (i.e. the

[10] We hence extend in two directions the findings of Bastos (2014) regarding the superiority of horizontal ensembles in forecasting recovery rates on defaulted corporate debt. First, by validating the aforementioned result also in the context of retail NPLs. But most importantly, by extending the analyses to non-horizontal ensembles, such as boosted trees and Cubist.

[11] See Kuhn (2018) for the details about model-specific variable importance metrics.

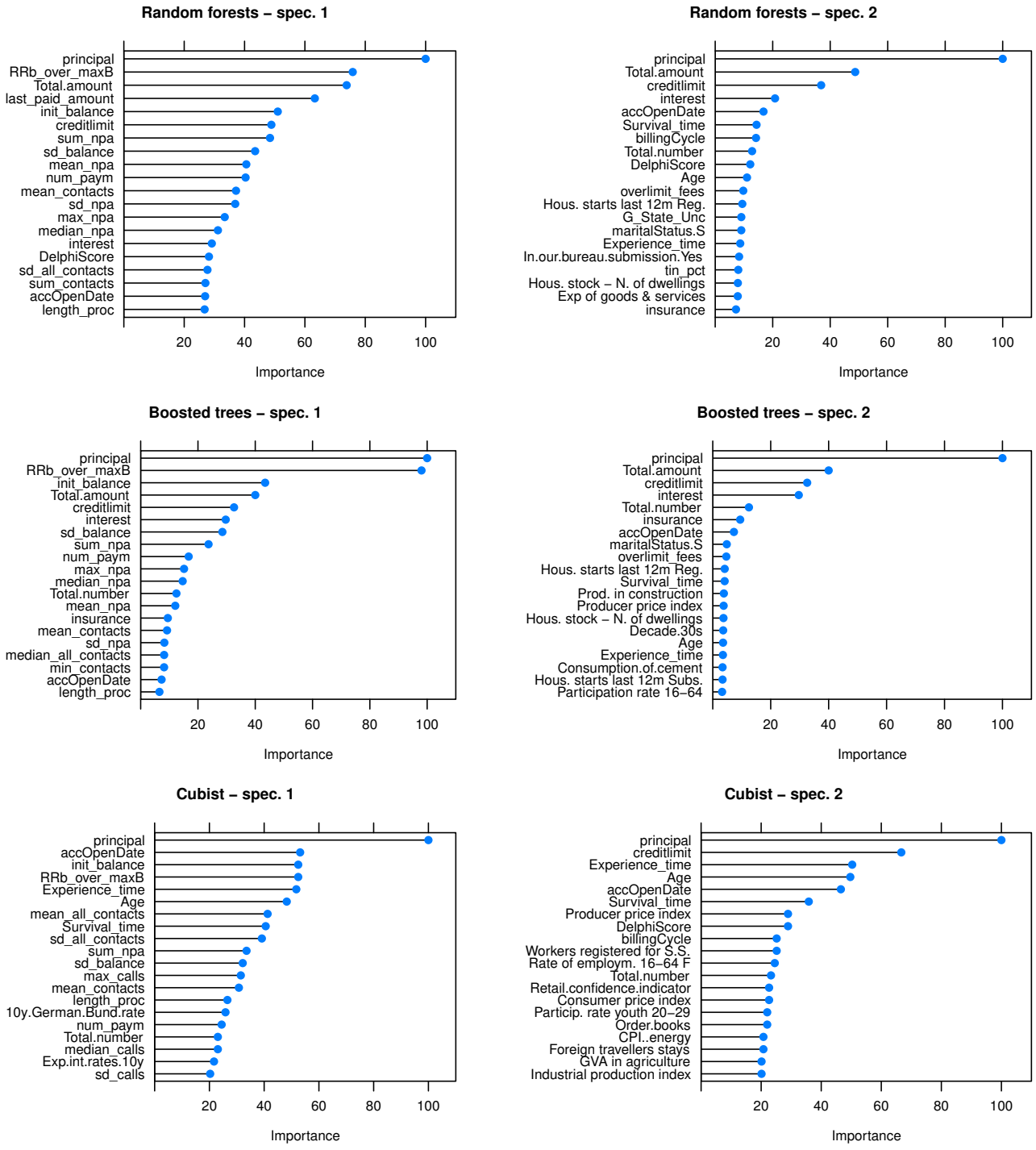

Figure 3: Top 20 variable importance rankings for Random forests, boosted trees and Cubist on the two specifications of the predictors set.

summary statistics of the net paid amounts). Variables relating to contacts with defaulted clients are present in the top 20 ranking for all models, with a particular focus on call data

in the case of Cubist. This latter is also the only model that explicitly attributes high importance to systematic variables already on specification 1, and in particular to interest rates data.

With regard to loan file information instead, the largest importance is always attributed to the loan principal. This evidence is not surprising because this variable is highly correlated with the exposure used to compute the recovery rate obtained by the debt collector. Additional loan file data figuring in the top 20 ranking for specification 1 are mostly associated with the creditworthiness of the debtor. These include the credit limit, interest, Credit Bureau score, survival time, and a proxy for clients' experience with the loan product.

The right column of Fig. 3 displays variable importance rankings for specification 2 models. We clearly recognize that bank recovery period information is substituted by socio-demographic and business cycle variables, although these latter have a significantly lower importance than the former factors. In general, loan-specific variables are the ones that play the most important role in this specification.

In terms of socio-demographic factors that gain positions in the variable importance ranking instead, we can observe that the borrowers' age is now present in the top 20 for all models, and the marital status is further displayed by random forests and boosted trees.

Particularly interesting are the business cycle variables that are displayed in the top 20 ranking on this specification. The algorithms seem in fact to understand that the NPLs for which we forecast recovery rates are referred to consumers, and also include residential mortgages. Importance rankings of boosted trees and Cubist display different measures of labor market conditions (e.g. participation rate, rate of employment and workers registered for social security) and inflationary pressure (e.g. consumer and producer price indexes). The rankings of boosted trees and random forests also feature several housing market indicators: production in construction, number of dwellings and house started, and consumption of cement.

Eventually, we observe that the random forest features less systematic variables in its top 20 ranking than the other two algorithms, but also that it displays the economic uncertainty index. We report a similar behavior also for boosted trees with stochastic gradient boosting. This latter algorithm actually includes the same economic uncertainty index also on specification 1, together with a housing market indicator. It hence appears that the findings of Gambetti et al. (2019) about the usefulness of economic uncertainty proxies for modeling recovery rates are also validated in the context of NPLs. We attribute this result to the well-established capability of uncertainty measures to anticipate economic fluctuations (see references in Section 2.4).

4.3. Effect of the variables on model predictions

In the previous sections, we have highlighted the superiority of rule-based ensembles in forecasting the recovery rates achievable by the debt collector. We have also shown

that algorithms should be trained using predictors referring to the bank recovery process, if available. The latter increase the capability of the model to understand borrowers' recovery potential, and ultimately increase its forecasting performances. When not available, the modeler should mainly refer to loan file data instead.

Especially for practical purposes, it is then appealing to investigate the relationship between model forecasts and the predictors, to understand if relationships are different across models, and also to give them an interpretation. We do this thanks to the accumulated local effects (ALE) plots displayed in Fig. 4 and 5.[12] They represent how individual inputs influence the model predictions on average: the vertical axes measure the differences with respect to the mean prediction (close to 40% for all algorithms in our case). Flat lines centered at zero indicate that the feature is not used by the algorithm. We select the plots based on their interpretability and the importance of the corresponding variables.

From the plots selected for specification 1 models (Fig. 4), we notice that the rule-based algorithms with higher forecasting power also generally agree on the shape of the approximated prediction-variable relationships. The flexibility of these models allows them to better approximate the true relationship compared to the best linear or nonlinear models (backward stepwise selection and MARS), which only capture general trends.

Some interesting intuitions that can be derived from these plots. Proceeding row-wise in the figure, we infer that the debt collector should expect higher recovery rates for those borrowers who have already shown their repayment ability and/or commitment during the bank recovery period. In particular, she should expect to collect higher than average recoveries for those borrowers with bank recovery rates higher than a threshold of roughly 30%. NPLs with outstanding balance higher than 5000 euros at the beginning of the bank recovery period also seem to promise recovery rates higher than the average for the debt collector. There are also indications of higher recovery rates for those NPLs associated with a larger number of net repayments to the bank (independently of their value). As for the interactions between the bank and defaulted borrowers, all models seem to understand that an increase in the average number of contacts (different from calls and visits) has not a positive connotation in terms of recovery potential. On the other side, borrowers that have been the object of larger efforts to enforce repayments through calls, seem to promise higher recovery rates to the debt collector. Eventually, it seems that the last registered net repayment – before the bank sold the portfolio to the debt collector – is quite representative of the recovery potential for a given client. A possible explanation is the fact that the bank has probably notified the clients before transferring their NPLs to

[12] ALE plots were developed by Apley (2016) to visualize the effects of different input variables on the model predictions also for complex algorithms. Relative to common alternatives (i.e. partial dependence and marginal plots), ALE plots do not suffer from bias in the presence of correlated predictors, do not require extrapolation and are more computationally efficient.

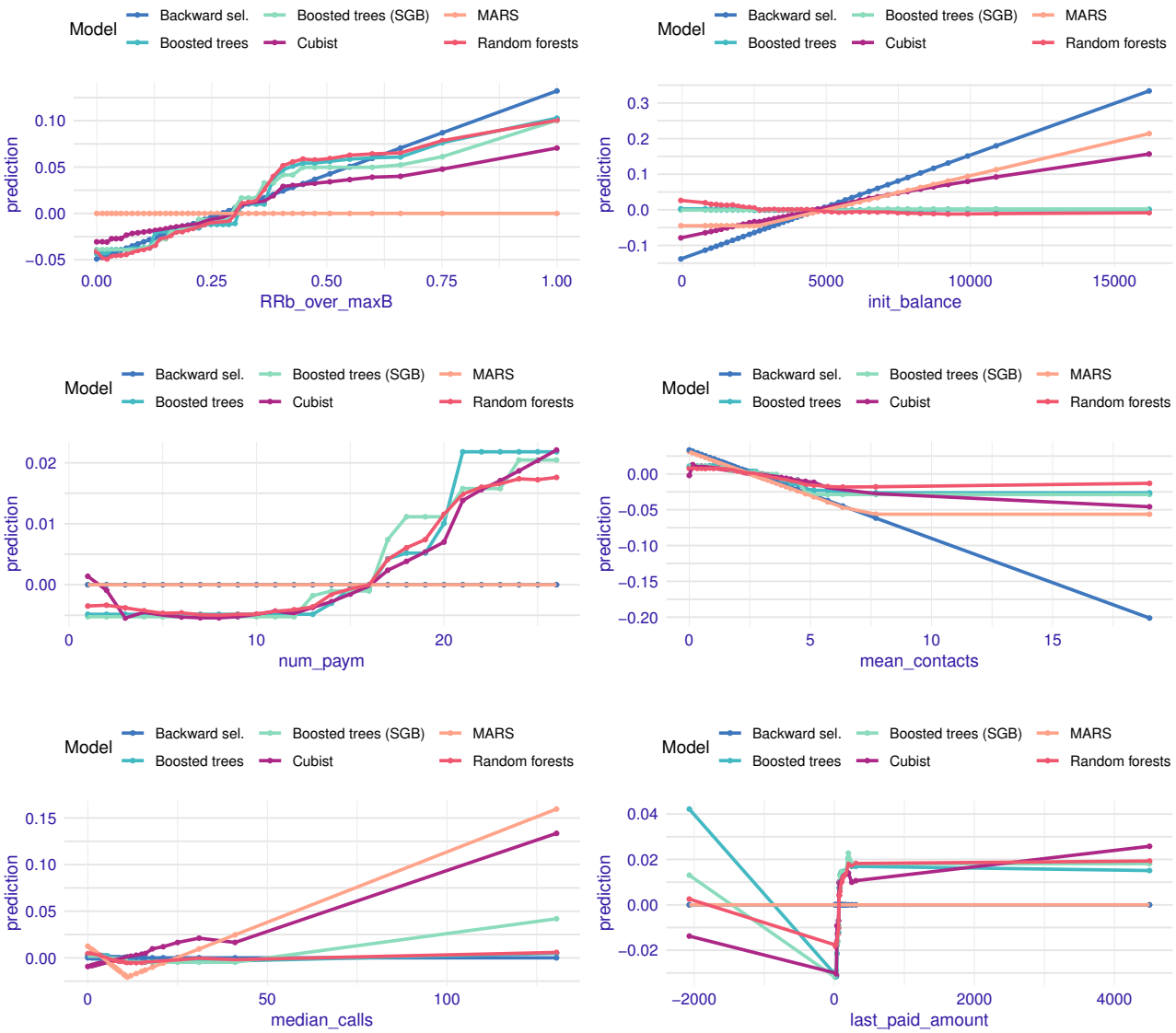

Figure 4: Accumulated local effects plots for selected predictors among bank recovery period data. Plots are derived from specification 1 models.

the specialized debt collector; borrowers with "latent" repayment ability tried to avoid this situation or to largely limit their exposure before the transfer.

By analyzing the ALE plots of Fig. 5, relative to specification 2 models, we find that the principal of the loan contributes negatively to the model predictions. We also observe one of the main shortcomings of applying algorithms that are not tailored to model recovery rates data: linear regression with backward stepwise selection and MARS can predict recovery rate values largely below the zero boundary. The credit limit clearly conveys information relative to borrowers' repayment ability: the higher the limit they were granted, the higher the recovery rate the debt collector can obtain. With respect to borrowers age, most models

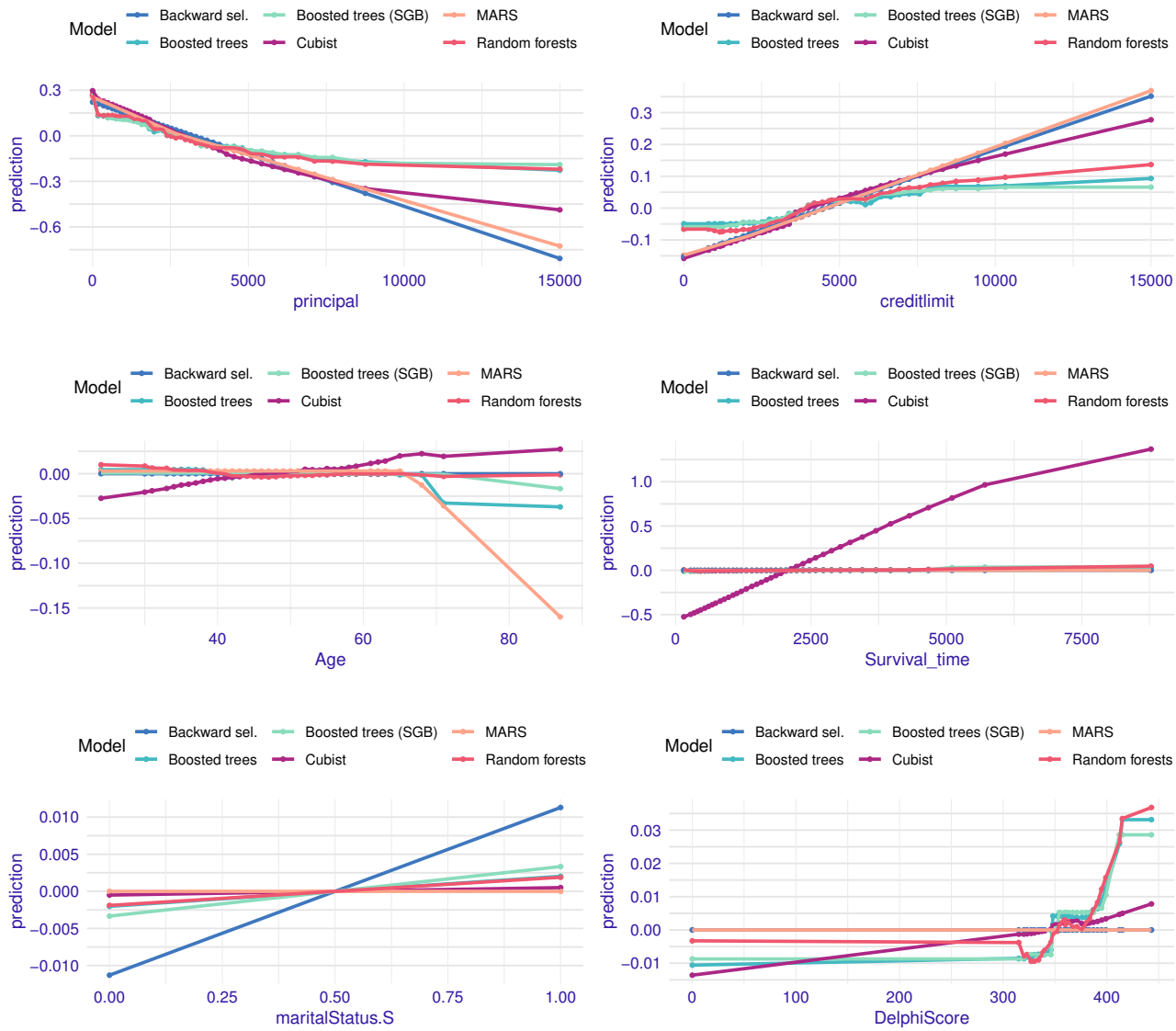

Figure 5: Accumulated local effects plots for selected predictors among socio-demographic and loan file data. Plots are derived from specification 2 models.

seem to identify a slightly negative downtrend after 65 years, which roughly corresponds to the retirement age for the reference state. These models suggest that the debt collector has to expect proportionally lower recovery rates for borrowers only perceiving a retirement income. As for the survival time, the relation displayed by all models is overshadowed by the large response of Cubist but is always positive: the more time passed before a client defaulted on his loan, the higher is the recovery rate the debt collector can expect to obtain. All algorithms also capture a positive signal, in terms of higher expectation of recovery, if the borrower is single (only the extremes of the corresponding ALE plot matter in this case). It is in fact well-known that collections are largely slowed down, in several European

legislations, when the borrower has dependent family members and especially in the case of mortgages. Furthermore, we can observe that most models predict proportionally higher recovery rates for those NPLs whose borrower was attributed a higher score by the Credit Bureau. In particular, the algorithms suggest that the debt collector should expect recovery rates higher than the average if the score was higher than 350 points.

5. Conclusion

We undertook a large scale benchmark study of machine-learning algorithms for forecasting recovery rates on non-performing loans for retail clients. We carried out model comparisons under two perspectives. First, we compared algorithms belonging to three different classes – namely linear, nonlinear and rule-based – with the objective of identifying the model structure best suited to the recovery rate problem. Second, we compared those algorithms across different specifications of the predictors set in order to shed light on the variables that should be involved in the forecasting exercise by the debt collection agency, whenever she purchases a portfolio of NPLs from a bank.

We collected a large database of predictors for this study. We involved factors that previous literature identified as important determinants of recovery rates; these latter include socio-demographic characteristics of defaulted borrowers, loan file information and business cycle variables. But most importantly, we considered a novel type of predictors feature engineered from the data relative to the recovery procedure promoted by the selling bank. We derived these predictors to characterize the *dynamics* of borrowers' repayments to the bank and those relative to the bank's contacts for enforcing repayments. We in fact assumed that the dynamics observed during the bank recovery process convey important information relative to borrowers' repayment ability and/or commitment, and are also informative of recovery potential during the recovery process promoted by the debt collector.

We found that rule-based algorithms of ensemble type, especially random forests and the newly added boosted trees and Cubist, displayed the best forecasting performances. This superiority was also found to be statistically significant by the model confidence set procedure. Regardless of the type, we found all models to perform much better when they were allowed to access the variables relative to the bank recovery process during training. Variable importance metrics highlighted the primary role of these variables in influencing the models and stressed the priority to employ loan file data when the latter factors were not available. We believe that the superiority of rule-based algorithms in forecasting recovery rates has to be explained with their natural capability to segment defaulted borrowers based on their recovery potential. This segmentation ability is much more evident when the algorithms could learn from borrowers' past repayment behavior and the bank's loan enforcement strategy.

Our research sheds light on some best practices that should be implemented to contribute to a better valuation and management of non-performing loans. It invites regulators to implement policies to limit information asymmetries on secondary markets, and in particular to encourage banks to disclose the information about the traded NPLs' recovery procedure. Furthermore, it bolsters debt collectors' interest to refine NPL assessment processes for trying to extract as much knowledge as possible about borrowers' recovery potential from bank recovery data. Eventually, it highlights the importance of involving enhanced modeling techniques in the valuation exercise.

References

Apley, W. D. (2016), "Visualizing the effects of predictor variables in black box supervised learning models," Available on arXiv at: arXiv:1612.08468.

Baker, S. R., Bloom, N., and Davis, S. J. (2016), "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 131, pp. 1593–1636.

Bastos, J. (2014), "Ensemble predictions of recovery rates," *Journal of Financial Services Research*, 46, pp. 177–193.

Bellotti, A., and Crook, J. (2007), "Modelling and predicting loss given default for credit cards," Credit Scoring and Credit Control Conference XI.

——— (2012), "Loss given default models incorporating macroeconomic variables for credit cards," *International Journal of Forecasting*, 28, pp. 171–182.

Bernardi, M., and Catania, L. (2018), "The model confidence set package for R," *International Journal of Computational Economics and Econometrics*, 8, pp. 144–158.

Breiman, L. (1996), "Bagging predictors," *Machine Learning*, 24, pp. 123–140.

——— (2001), "Random forests," *Machine Learning*, 45, pp. 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, Monterey, CA: Wadsworth and Brooks.

Caputo, B., Sim, K. L., Furesjo, F., and Smola, A. J. (2002), "Appearance-based object recognition using svms: Which kernel should I use?" in "Proc. of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision," .

Caselli, S., and Querci, F. (2009), "The sensitivity of the loss given default rate to systematic risk: new empirical evidence on bank loans," *Journal of Financial Services Research*, 34, pp. 1–34.

ECB (2016), "The impact of uncertainty on activity in the euro area," *ECB Economic Bulletin*, pp. 55–74.

——— (2017), "Guidance to banks on non-performing loans," Tech. rep., European Central Bank.

ESRB (2017), "Resolving non-performing loans in europe," Tech. rep., European Systemic Risk Board.

——— (2018), "Approaching non-performing loans from a macroprudential angle," Tech. rep., European Systemic Risk Board.

Friedman, J., Hastie, T., and Tibshirani, R. (2010a), "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software, Articles*, 33, pp. 1–22.

——— (2010b), "glmnet: Lasso and elastic-net regularized generalized linear models," *R Package Version*, 1.

Friedman, J. H. (1991), "Multivariate adaptive regression splines," *The Annals of Statistics*, 19, pp. 1–67.

——— (2001), "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, 29, pp. 1189–1232.

Gambetti, P., Gauthier, G., and Vrins, F. (2019), "Recovery rates: Uncertainty certainly matters," *Journal of Banking & Finance*, 106, pp. 371–383.

Gieseck, A., and Largent, Y. (2016), "The impact of macroeconomic uncertainty on activity in the euro area," *Review of Economics*, 67, pp. 25–52.

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G. (2018), *gbm: Generalized Boosted Regression Models*, R package version 2.1.4, URL https://CRAN.R-project.org/package=gbm.

Grippa, P., Iannotti, F., and Leandri, F. (2005), "Recovery rates in the banking industry: stylised facts emerging from the italian experience," in E. Altman, A. Resti, and A. Sironi, (Eds.) "Recovery risk," London: Risk Books.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011), "The model confidence set," *Econometrica*, 79, pp. 453–497.

Hastie, T., and Tibshirani, R. (2017), *mda: Mixture and Flexible Discriminant Analysis*, R package version 0.4-10, URL https://CRAN.R-project.org/package=mda.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning: data mining, inference and prediction*, Springer, 2nd ed.

Hothorn, T., Hornik, K., and Zeileis, A. (2006a), "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, 15, pp. 651–674.

——— (2006b), "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, 15, pp. 651–674.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004), "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, 11, pp. 1–20.

Kose, M. A., and Terrones, M. (2012), "How does uncertainty affect economic performance?" *IMF World Economic Outlook*, pp. 49–53.

Kuhn, M. (2008), "Building predictive models in R using the caret package," *Journal of Statistical Software, Articles*, 28, pp. 1–26.

——— (2018), *caret: Classification and Regression Training*, R package version 6.0-80, URL https://CRAN.R-project.org/package=caret.

Kuhn, M., and Johnson, K. (2013), *Applied Predictive Modeling*, New York, Springer.

Kuhn, M., and Quinlan, R. (2018), *Cubist: Rule- And Instance-Based Regression Modeling*, R package version 0.2.2, URL https://CRAN.R-project.org/package=Cubist.

Leow, M., Mues, C., and Thomas, L. (2014), "The economy and loss given default: evidence from two UK retail lending data sets," *Journal of the Operational Research Society*, 65, pp. 363–375.

Liaw, A., and Wiener, M. (2002), "Classification and regression by randomforest," *R News*, 2, pp. 18–22.

Loterman, G., Brown, I., Martens, D., Mues, C., and Baesens, B. (2012), "Benchmarking regression algorithms for loss given default modeling," *International Journal of Forecasting*, 28, pp. 161–170.

Ludvigson, S. C., Ma, S., and Ng, S. (2019), "Uncertainty and business cycles: exogenous impulse or endogenous response?" Working Paper.

Lumley, T. (2017), *leaps: Regression Subset Selection*, R package version 3.0, URL https://CRAN.R-project.org/package=leaps.

Milborrow, S. (2018), *earth: Multivariate Adaptive Regression Splines*, R package version 4.6.3, URL https://CRAN.R-project.org/package=earth.

Nazemi, A., and Fabozzi, F. J. (2018), "Macroeconomic variable selection for creditor recovery rates," *Journal of Banking & Finance*, 89, pp. 14–25.

Nazemi, A., Heidenreich, K., and Fabozzi, F. J. (2018), "Improving corporate bond recovery rate prediction using multi-factor support vector regressions," *European Journal of Operational Research*, 271, pp. 664–675.

Querci, F. (2005), "Loss given default on a medium-sized Italian bank's loans: an empirical exercise," *European financial management association*.

Quinlan, J. R. (1993), "Combining instance-based and model-based learning," in "Proceedings of the Tenth International Conference on International Conference on Machine Learning," San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., ICML'93, pp. 236–243.

Quinlan, R. (1992), "Learning with continuous classes," *Proceedings of the 5th Australian joint conference on artificial intelligence*.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Therneau, T., Atkinson, B., and Ripley, B. (2017), *rpart: Recursive Partitioning and Regression Trees*, R package version 4.1-11, URL https://CRAN.R-project.org/package=rpart.

Tipping, M. E. (2001), "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, 1, pp. 211–244.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag.

Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer, 4th ed., ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4.

Wang, Z. (2018), *bst: Gradient Boosting*, R package version 0.3-15, URL https://CRAN.R-project.org/package=bst.

Westfall, P., and Young, S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Wiley Series in Probability and Statistics, Wiley.

Williams, C. K. I., and Rasmussen, C. E. (1996), "Gaussian processes for regression," in "Advances in Neural Information Processing Systems 8," MIT press, pp. 514–520.

Ye, H., and Bellotti, A. (2019), "Modelling recovery rates for non-performing loans," *Risks*, 7, pp. 1–17.

Appendix A. List of algorithms

Table A1: List of prediction algorithms for linear, nonlinear and rule-based models.

Algorithm	Caret method	Description	Author
`lm`	`lm`	Linear regression	R Core Team (2017)
`leaps`	`leapsBackward`	Backward stepwise selection	Lumley (2017)
`glmnet`	`ridge`	Ridge regression	Friedman et al. (2010a) Friedman et al. (2010b)
`glmnet`	`lasso`	Lasso regression	"
`glmnet`	`glmnet`	Elastic net regression	"
`earth`	`earth`	Multivariate Adaptive Regression Splines	Milborrow (2018) Hastie and Tibshirani (2017)
`knn`	`knn`	K-Nearest neighbors	Venables and Ripley (2002)
`nnet`	`avnnet`	Model averaged neural networks	"
`ksvm`	`svmRadial`	Support vector regression	Karatzoglou et al. (2004)
`rvm`	`rvmRadial`	Relevance vector regression	"
`gausspr`	`gausssprRadial`	Gaussian processes	"
`rpart`	`rpart`	Regression trees	Therneau et al. (2017)
`ctree`	`ctree2`	Conditional inference trees	Hothorn et al. (2006b)
`bag`	`treebag`	Bagged trees	Kuhn (2018)
`bst`	`bstTree`	Boosted tree	Wang (2018)
`gbm`	`gbm`	Stochastic gradient boosting	Greenwell et al. (2018)
`randomForest`	`rf`	Random forests	Liaw and Wiener (2002)
`cubist`	`cubist`	Cubist	Kuhn and Quinlan (2018)

Appendix B. Description of variables appearing in the plots

Table B1: Explanation of variables names appearing in the plots of the main text. The second column identifies the variables that are also present in the second specification of the predictor set.

Variable name	Included in specification 2	Description
RRb_over_maxB		Recovery rate obtained during the bank recovery process
last_paid_amount		Last net paid amount registered
init_balance		Initial total balance for a loan at time of bank recovery
sum_npa		Sum of the net paid amounts
sd_balance		St. deviation of the outstanding balance during the bank recovery process
mean_npa		Borrower's average net paid amount
num_paym		Number of monthly summaries registered
mean_contacts		Average monthly contacts (different from calls and visits)
sd_npa		St. deviation of the net paid amounts
max_npa		Maximum net paid amount registered
median_npa		Median of the monthly net paid amounts
sd_all_contacts		St. deviation of all the monthly contacts
sum_contacts		Total number of contacts (different from calls and visits)
accOpenDate		Date of opening of the loan account
length_proc		Length of the bank recovery process
median_all_contacts		Median of all the monthly contacts
min_contacts		Minimum number of contacts (different from calls and visits)
sd_all_contacts		St. deviation of all monthly contacts
max_calls		Maximum number of calls
median_calls		Median number of calls
sd_calls		St. deviation of the calls
principal	✓	Original amount of the loan
Total.amount	✓	Total amount borrowed by an individual on all accounts
creditlimit	✓	Credit limit
interest	✓	Interest rate payments
DelphiScore	✓	Credit Bureau Score
Survival_time	✓	Time of survival for a given loan
billingCycle	✓	Number of months since repayments began
Total.number	✓	Number of loan accounts for a borrower
Age	✓	Age of the borrower
overlimit_fees	✓	Penalty fees for exceeding the credit limit
Hous. starts last 12m Reg.	✓	House started in the last 12 months (relative to default date) in a given region
G_State_Unc	✓	Economic policy uncertainty index for the reference state
maritalStatus.S	✓	Identifier for borrowers that are single
Experience_time	✓	Proxy for the borrower's experience with the loan product
In.our.bureau.submission.Yes	✓	Identifier for loans that are in the debt collector bureau
tin_pct	✓	Percentage interest rate
Hous. stock - N. of dwellings	✓	Housing stock - Number of dwellings
Exp of goods & services	✓	Export of goods and services
insurance	✓	Insurance fees
Prod. In construction	✓	Production in construction
Producer price Index	✓	Producer price index
Decade.30s	✓	Identifier of borrowers in their 30s
Consumption of cement	✓	Consumption of cement for the reference state
Hous. starts last 12m Subs.	✓	House started in the last 12m subsidized by the government
Participation rate 16-64	✓	Participation rate of citizens of age range 16-64
10y.German.Bund.rate	✓	Interest rate on the 10y German Bund
Exp.Int.rates.10y	✓	Expected interest rate on the 10y reference state's bond
Workers registered for S.S.	✓	Number of workers registered for social security
Rate of employm. 16-64 F	✓	Rate of employment of females in the age range 16-64
Retail.confidence.Indicator	✓	Retail confidence indicator
Consumer price Index	✓	Consumer price index
Particip. rate youth 20-29	✓	Participation rate of people in the age range 20-29
Order.books	✓	Orders of books
CPI..energy	✓	Consumer price index: energy
Foreign travellers stays	✓	Foreign travellers hotel stays
GVA in agriculture	✓	Gross value added for agriculture
Industrial production index	✓	Industrial production index