Classification of NPL with a Random Forest approach

di Massimiliano Zanoni (Iason Ltd)

Articolo sottoposto a doppio referaggio anonimo, pervenuto in data 23/12/2019 e accettato il 17/03/2020

Abstract

Artificial Intelligence has quickly entered in the financial services industry covering a wide range of applications. This work studies a structured statistical approach to classify non-performing unsecured commercial exposures according to their recovery potential, based on a Machine Learning technique known as Radom Forest.
The framework adopted is based on two different components: one identifying the cases that may be recovered and the other estimating their recovery level. In addition, the work compares the RF - introduced with a review of the underlining Decision Tree theory and its performance metrics - to the better-known Logit approach.
The framework is meant to provide an evaluation of recovery at aggregate level, for pricing and management purposes, but it is also successfully tested in comparison between two portfolios, one of which know to the analyst. Results show that the Random Forest approach is as reliable and slightly more performing than the better known Logistic approach, even with a limited set of information.

1 Introduction

Learning is a process meant to identify patterns and matching rules within available data, in order to infer the inputs-outputs mapping. Traditional modelling performs this task by presuming a functional form between the two, then estimate the key parameters through a procedure called fitting.
Statistical learning instead consists in a set of methodologies where no functional form is assumed and classification is based directly on the characteristics of events (Gareth, et al., 2013). **Machine Learning** (ML) constitutes a set of non-parametric, non-linear statistical methods to implement statistical learning.
In the case considered, the training dataset consists of short-term non-performing commercial exposures to be recovered by a servicer, which usually purchases such portfolios at a discount and then works them out, hence a correct price estimate and an efficient workout process are key to profitability.
It is worth noticing that, though forecasting occurs at single deal level, what really matters to the investor is the ability to correctly infer the amounts which can be recovered at aggregate levels.
The underlying assumption is that both the recovery event and the recovery rate can be inferred from past transactions and debtor information. In particular, the work confirms, with (Khieu, et al., 2012), that loan's characteristics are more significant determinants of the recovery rate than are borrower characteristics prior to default. On the other hand, we were not able to inquire if predictors referring to the recovery process of the processor, or the bank prior to portfolio sale, were relevant as suggested by (Bellotti, et al., 2019), since such information was not available.
The approach adopted mirrors the standard Loss Given Default (LGD) framework which splits the default event from loss estimates and was designed independently of existing literature. Focusing on the servicer, the work complements the literature which mostly estimates recovery models from the same bank originating the loan (Ciavoliello, et al., 2016). In addition, it contributes to update the benchmark study involving Machine Learning methods.
Besides its good predictive capabilities, RF can provide valuable insight into the main factors driving the recovery dynamic (Breiman, 2001). It is worth noticing that, though forecasting occurs at single deal level, what really matters to the investor is the ability to correctly infer the amounts which can be recovered at aggregate levels (e.g. at geographical area level).
After an introduction to the RF approach, a chapter is dedicated to the description of the Non-Performing Loan (NPL) dataset before analysing the modelling approach where the model for the probability of recovery is compared to a classical generalized Logistic model.
Finally, the last chapter presents the results at regional and portfolio level, showing that the RF framework is as reliable as the better-known Logistic approach.

2 Tree-based clustering

Random Forest is a machine learning approach which sets binary rules, represented as trees, to cluster the event space by recursive branching. A clustering rule is a set of splitting points in each dimension of the event space, dividing (branching) the dataset into two separate parts. A set of rules represent a tree ending in a defined boxing of the event space (terminal node or *leave*) to which a unique value is assigned (prediction). All the different trees together are called the *Forest.* The final model is obtained by averaging the value assigned to any event by over all the splitting trees composing the forest.
While the linear approach relies on a 'linearized' law, the RF does not need any a-priory knowledge of the system. For each observation, a prediction is obtained by recursively applying the following steps:

1. Randomly select *m* variables from all the ones in the dataset (*feature bagging*).
2. Randomize the selection of variables in each split[1]

[1] The continuous change of variables at each split limits the chance to include correlation in the learning process.

3. For each of the selected variables identify the split-point that minimizes the residual sum of squares of predictions in the specific tree. Formally, for a randomly selected feature j, two regions are defined by the split point *c*

$$(1) \qquad \boldsymbol{R_1(j,c) = \{X|X_j < c\}\ and\ R_2(j,c) = \{X|X_j \geq c\}}$$

such that:

$$(2) \qquad \boldsymbol{\sum_{i:x_i \in R_1(j,c)} (y_i - \hat{y}_{R1})^2 + \sum_{i:x_i \in R_2(j,c)} (y_i - \hat{y}_{R2})^2}$$

is minimized, where $\hat{y}_{R1}$ and $\hat{y}_{R2}$ are average estimated outputs for each split.

4. Take the arithmetic average each outcome over all trees in the forest.

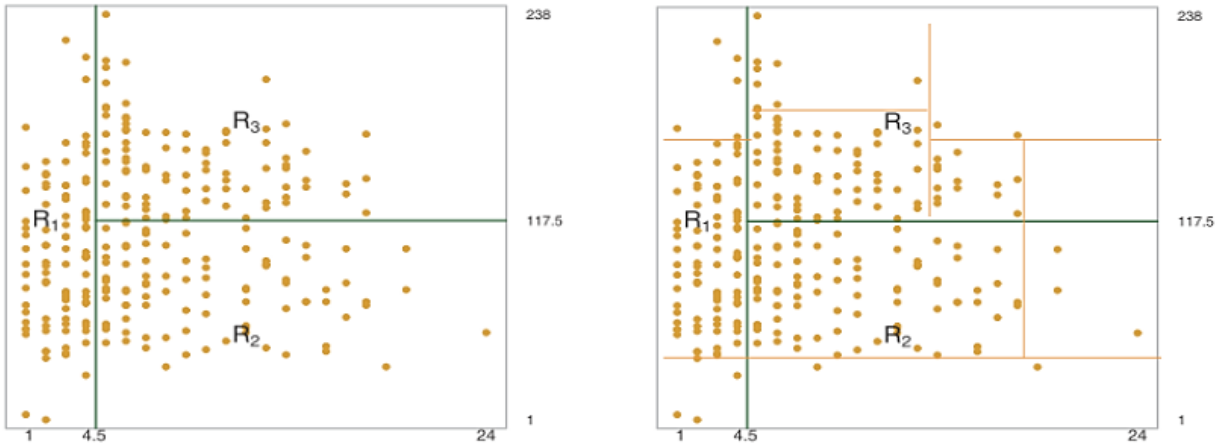

Figure 1 Splitting Process

The process is affected both by the number of variables tried at each split, and by the number of processed layers and it eventually ends up drawing a number of boxes, identified by one value of the variable, independently of the number of events included. This process differs from that of bagged trees where all predictors are considered at each split, creating highly correlated tree. RF process overcomes this problem by randomly selecting a limited number of predictors at each split, which makes the model less prone to overfitting problem (Hastie, et al., 2009).

2.1 Performance indicators

Performance measures are key to drive variable selection and model validation. Beside classical statistical measures such as

- ***Percentage Variance explained*** (PVE): it measures how close the predictions on new observations get to the real variable. Formally:

$$(3) \qquad PVE = \frac{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n} - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}}{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}$$

where $\bar{y}$ and $\hat{y}_i$ are the average of real and estimated outputs respectively[2].

- ***Receiver operating characteristic*** (ROC)[3] curve (Figure 2), which is a usual performance measure for binary models. Formally Specificity or true negative and Sensitivity or true positive are defined as:

$$(4) \qquad \text{Specificity} = \frac{\text{A}}{\text{A+B}} : \text{P}(\text{Q} = 0|\text{Y} = 0), \text{Sensitivity} = \frac{\text{D}}{\text{C+D}} : \text{P}(\text{Q} = 1|\text{Y} = 1)$$

In general, the recognition of events in model application requires the definition of a cut-off value that separates positive from negative cases. Together with their 'false' classification, they can be represented through a Confusion Table represented in Table 1.

[2] For a more rigorous disruption of variable importance in RF models refer to (Grömping, 2009) and (Breiman, 1996)
[3] For detailed analysis of the ROC method refer to (Fawcett, 2006)

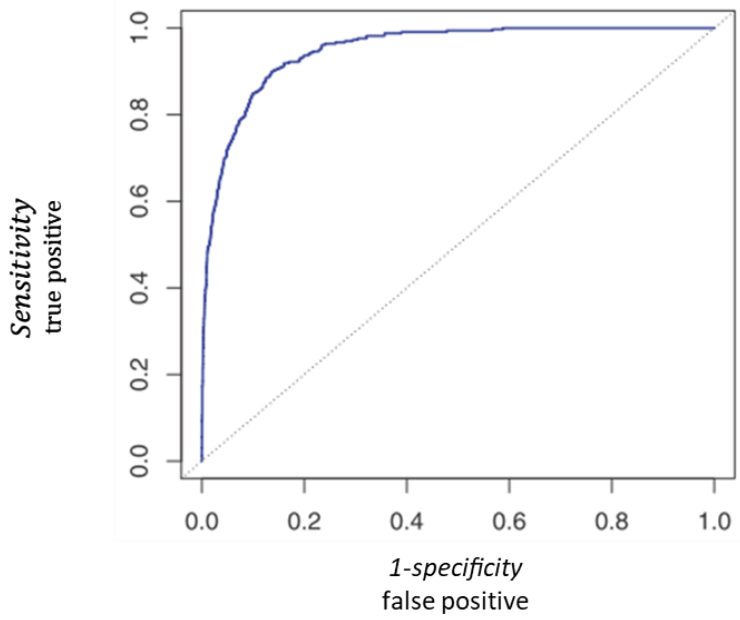

Figure 2 ROC Curve

		Predicted	
		Negative (0)	Positive (1)
Observed	Negative (0)	A	C
	Positive (1)	B	D

Table 1 Confusion Table

- ***Mean decrease in accuracy*** and ***Mean decrease in node impurity*** represent important measures of the contribution of each single variable; the former analysing the change in the prediction error due to the change in values of a given variable and the latter by comparing the "purity of the node split" in the alternative of including or excluding the given variable from the set used for splitting [4].

For completeness, the Confusion Table is extended to include the totals of each measures, the portfolio mix, real forecasted and the diagonal sum of true positive and negative.

		Fit			
		Non-Rec	Recovery	**Total**	Portfolio-mix
Real	Non-Rec	20,523	354	20,877	97.0%
	Recovery	524	123	647	3.0%
	Total	21,047	477	**21,524**	
	Portfolio-mix	97.8%	2.2%		
	OK	97.5%	25.8%		
	KO	2.5%	74.2%	**Tot. acc.**	
	OK on sample	98.3%	19.0%	**95.9%**	

Table 2 Extended Confusion Table – Example Linear Cut Off 10%

In the example (Table 2), the portfolio mix offers a rough measure of the model's ability to replicate the original portfolio, forecasting only 2.2% of recovery events versus an effective 3%. Furthermore, it underlines the fraction of correctly separately forecasted success and fail cases (OK) and overall (OK on sample)[5].

[4] For detailed definition refer to (Breiman, 2001)

[5] The overall number however does not distinguish between success and fail events, distinction that may be vital in cases where the cost of different mistakes is not homogeneous.

3 Dataset description

The NPL dataset includes 22,290 non-performing retail exposures related to unpaid utility bills, with limited information about the customer and the exposure provided both as snapshot at a given date and as payment flows. Some key information is inferred from the fiscal code (gender and age), hence exposures where this is missing are discarded. Several data quality issues imposed more exclusions, e.g. exposures related to debtors above 95.

As summarized in the following figure, recovery rates are distributed mostly towards extreme values, i.e. most dossiers present no recovery or a high recovery rate, leaving only a fraction in the middle of the recovery scale. In order to reduce statistical noise dossiers with a very small recovery rate or absolute amount, were also excluded from the analysis.

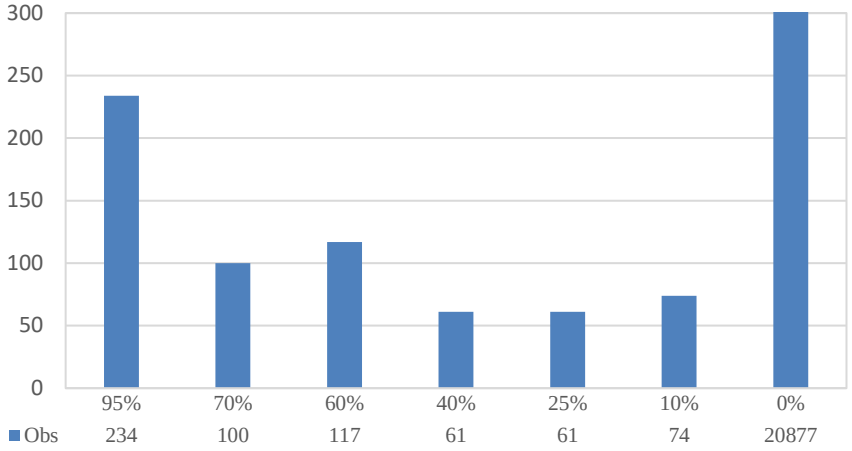

Figure 3 Number of observations by Recovery Level

The variable "*contacts"* (shown in the following histogram) represents the number of interactions with the counterparty in the recovery process, intuitively this is linked to the duration of the process and to the average recovery rate, at least up to a certain number of contacts. On average, dossiers with 140-170 contacts present a 50-55% recovery rate however, increasing the number of contacts does not improve the rate of recovery (in fact it is reduced to around 30-35%), showing a possible inefficiency when a significant effort is made on cases yielding a poor and slower recovery.

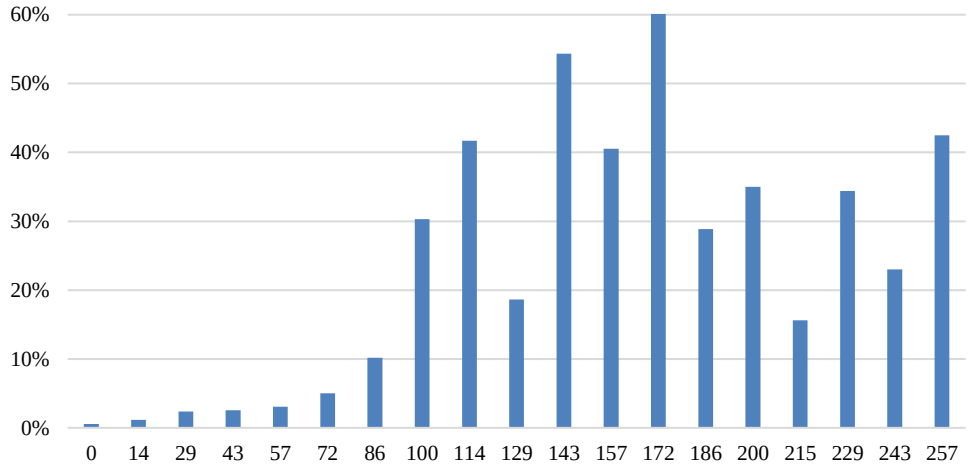

Figure 4 Average Recovery Rate by Number of Contacts

In many cases the recovery process involves a single flow but, in general, the cumulated recovery increases with time up to the final recovery amount. At any given period *T*, the *average cumulated recovery rate* is given by the sum of payments, divided by the total exposure of dossiers.

$$(5) \qquad average\ cumulated\ recovery\ rate = \frac{\sum_{t,T} \text{payment}_t}{\sum_T \text{exposure}_T}$$

Given the different recovery dynamics, the analysis of the average recovery curve at portfolio level should focus on dossiers with homogeneous durations i.e. by restricting the recovery process to dossiers with a recovery duration above D,

$$(6) \qquad \frac{\sum_{t,T,K>D} \text{payment}_{t,K}}{\sum_{T,>D} \text{exposure}_{T,K}}$$

where K selects those dossiers with a collection process exceeding duration *D*.

Clearly with D=0 the set embraces the whole portfolio, including all the quick-recovery dossiers, with each set reflecting an average of many dossiers with different recovery dynamics and durations.
The curve is steeper in the initial part, but reaches a lower total recovery rate (blue line in Figure 5) with respect to subsets that includes dossiers with a longer recovery process, e.g. over 3 months (orange line) which presents a slower rise to a higher recovery rate.

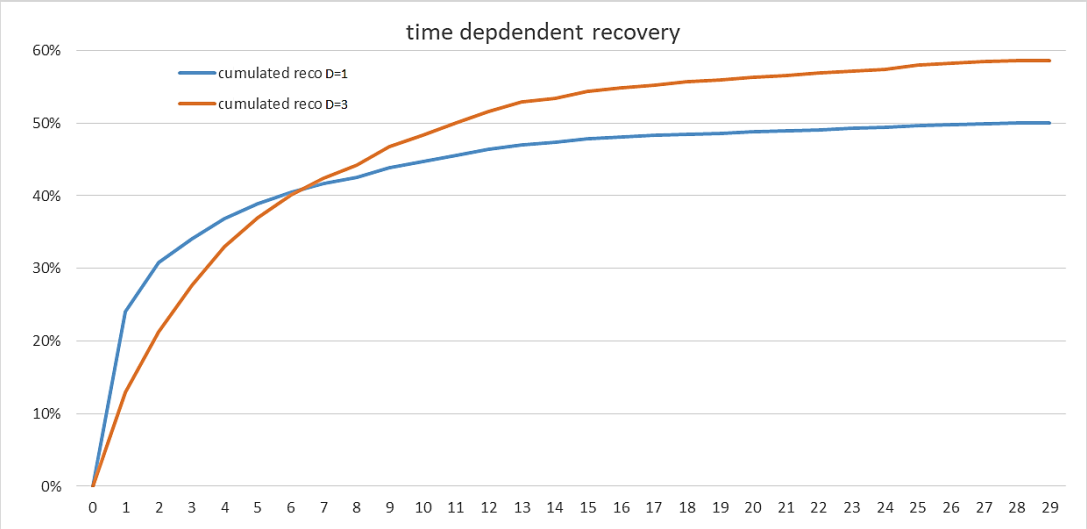

Figure 5 Time Dependent Recovery: Average Cumulated Recovery Rate Dynamic for different portions of portfolio

4 Development approach

The framework used to forecast recoveries in a portfolio of NPL exposures is based on two different models: one dedicated to forecast the probability of recovery of the single dossier and one to estimate the recovery rate in case of a recovery. Accordingly, the estimation process is divided into two parts: the first one is meant to estimate the *Probability of Recovery* (PR) of a given dossier; while the second is targeted to estimate the Recovery Rate of each recovered exposure.
The two models are estimated on the same variable set and tested separately, they will thus be statistically independent of each other, but dependent on the availability and quality of the same dataset.
In order to provide a frame for comparison, beside the RF approach, the PR model is estimated also using a generalized linear approach (Logit) while, for the recovery rate, only one multi-linear model is estimated.
Each of the two models is characterized by a specific performance level. Thus it will introduce an independent error into the final recovery estimates.
However, since the two models are statistically independent, their errors may partially cancel out when dossiers are aggregated at macro-regional level, potentially generating a more accurate total recovery with respect to what obtained at dossier level.
The performance measure used for the Recovery given Recovery (RGR) model will be a classic adjusted R-squared correlation, while to measure the quality of the PR model and compare the two approaches, a ROC curve, together with an extended Confusion table, is used, in addition to specific metrics evaluating the relevance of each variable (Percentage of Variance Explained and Increased Node Purity).

5 Probability of recovery model

The probability of recovery as a continuous variable, is estimated similarly within the Logit and the Random Forest approach, then transformed in a list of recovery events through a specific cut-off optimized for each approach. Alternatively, recovery events can be tagged directly by using the RF as a classifier. This result will be compared to the one obtained from the continuous PR.

5.1 Threshold definition

The cut-off necessary to determine which dossier should be considered a successful recovery can be defined in different ways:

A. A possible value is obtained by dividing the dataset in real recovery and non-recovery events and plotting each group along the probability of recovery assigned by the model. The value at the intersection point of these two distributions can be taken as cut-off (Figure 6).

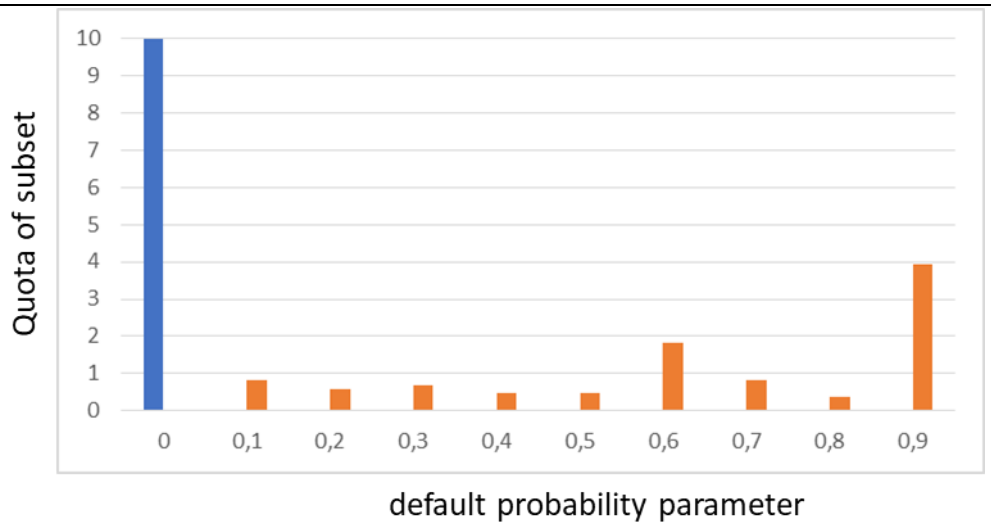

Figure 6 Quantile Distribution of Recovery (Orange) and Non-Recovery Evens (Blue)

B. An alternative and more rigorous way to define the cut-off is to set it in a way that the number of estimated recoveries is close to the real one.

C. Table below shows the number of predicted recoveries as a function of the cut-off in the portfolio analysed including 626 recoveries. This approach suggests the same cut-off selected with the previous visual approach i.e. PR = 0.1.

cut-off >=	predicted
9%	831
10%	655
11%	512

Table 3 CUT-OFFS

D. An alternative analytic approach is to define the cut-off that maximizes the overall number of success cases - both negative and positive – i.e. the value that alternatively:

- minimizes the differences (Specificity – Sensitivity)[6] considering positive values only
- maximizes the sum (Specificity – Sensitivity) commonly known as Youden's Index (Youden, 1950)

Considered that in the present work, method **A** and **B** identify the same cut-off level, only the two cut-offs will be compared.

5.2 Linear approach

The model estimated with the Logit approach is summarized in the following table:

Variable	**P-value**	**relevance**
Intercept	0.000	High
Abroad	0.000	High
Debtor age	0.000	High
Macro Region: Center-South	0.005	Medium
Macro Region: Island	0.037	Low
Macro Region: North	0.141	Low
Contact	0.000	High

Table 4 Logistic Model for the Probability of Default

[6] See equation (4) for relevant definition

The ROC curve associated to this model visually shows how much it improves the random guess represented by the straight line. Though the ROC does not depend on a specific cut-off, the performance of the model, in terms of correctly identified recovery events, is based on a cut-off chosen.
In more details, the performances of the different models and options are compared through an extended Confusion matrix which includes, besides true/false positive and negative events, the following metrics:

- **Total accuracy** (i.e. correct predictions) - (TN + TP) / ∑ all
- **Precision** = TP / (TP + FP) - (# of predictions)
- **Specificity** = TN / (TN + FP) (# num of negative occurrences)

The Confusion matrices in Table 5 and Table 6 allow to compare option **B** and **C** outlined before[7]. Adopting definition **B** for the cut-off (10%) the resulting portfolio mix is very similar to the real one (3.1% vs 2.9%) but only 21.5% of recovery signals is correct (Table 6).
Adopting definition **C** for the cut-off (3%), which grants the same performance for both positive and negative forecasts, a much larger set of recovery events is identified, at the expenses of a significant increase in wrong recovery signals (94,1%, vs 78,5% of portfolio), leading to an excessive 33,7% portfolio mix.
At the base of this approach, is not only the wrong assumption that the two event types cover, more-or-less, the same portion of the population, but also both that the same cost is associated to both types of mistakes, clearly incorrect.
In summary, approach **B**, identifying a number of recoveries close to the real one (655 vs 626), is used to compare the RF to the Logit approach, as it provides a good estimate of the portfolio mix.

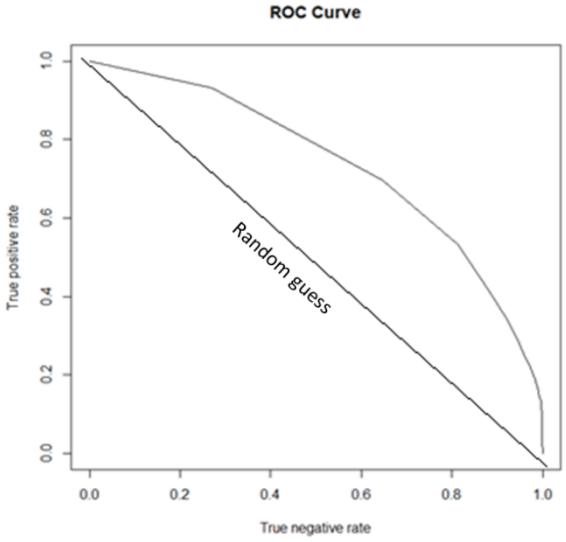

Figure 7 ROC Curve – Logistic Approach

		Fit			
		Non-Rec	Recovery	**Total**	Portfolio-mix
Real	Non-Rec	20,094	514	20,608	97.1%
	Recovery	485	141	626	2.9%
	Total	20,579	655	**21,234**	
	Portfolio-mix	96.9%	3.1%		
	Precision	97.6%	21.5%		
	KO	2.4%	78.5%	**Tot. acc.**	
	OK on sample	97.5%	22.5%	**95.3%**	

Table 5 Linear approach, cut-off method B

7 Option A is not reported as it resembles option B.

		Fit			
		Non-Rec	Recovery	**Total**	Portfolio-mix
Real	Non-Rec	13,879	6729	20,608	97.1%
	Recovery	205	421	626	2.9%
	Total	14,084	7150	**21,234**	
	Portfolio-mix	66.3%	33.7%		
	Precision	98.5%	5.9%		
	KO	1.5%	94.1%	**Tot. acc.**	
	OK on sample	67.3%	67.3%	**67.3%**	

Table 6 Linear approach, cut-off method C

6 Random Forest approach

A different model for the probability of recovery was estimated on the same dataset using a Random Forest approach in regression mode. The output is then turned into recovery events with a new cut-off, calibrated on the same principle, i.e. to obtain a number of recovery events closer to the real one.

		Fit			
		Non-Rec	Recovery	**Total**	Portfolio-mix
Real	Non-Rec	20,132	476	20,608	97.1%
	Recovery	468	158	626	2.9%
	Total	20,600	634	**21,234**	
	Portfolio-mix	97.0%	3.0%		
	Precision	97.7%	24.9%		
	KO	2.3%	75.1%	**Tot. acc.**	
	OK on sample	97.7%	25.2%	**95.6%**	

Table 7 RF approach, cut-off method B

As shown in Table 7, the RF approach identifies a higher number of true recovery cases than the Logit approach (158 vs 141) in a smaller set classified as recovered, thus hitting two scores: a portfolio mix closer to the real one, and a higher performance (24.9% vs 21,5%). It is worth noticing that, differently from the Logit approach, the RF approach includes a non-deterministic component that changes the output at each new run, generating a slightly different classification. With the RF approach, the model usually includes all variables with different relevance, as shown in the following table, where they are ranked according to their relative importance[8].

Variable	IncNodePurity	%IncMSE
Exposure	20,23	2,18E-03
Contacts	48,33	7,60E-03
Dossier age	11,87	6,84E-04
Macro region	2,45	2,86E-04
Abroad/Italy	1,00	6,60E-04
Debtor age	8,07	5,88E-04
Debtor gender	0,70	1,85E-05

Table 8 Variable performance – RF-PR model

[8] The performance and robustness depend on parameters which need to be tuned properly during the estimation process, the ones used in the present model are: the number of trees, (500) minimum size of terminal nodes in each tree, (50) the number of variables to be selected every time, (2)

As shown, the exposure is the most relevant variable in the model, fact that may create a preference for larger exposure, with the possible result of overestimating recovery exposure. The Logit and the RF models can be directly compared through their relative ROC curves, both plotted on the same graph.

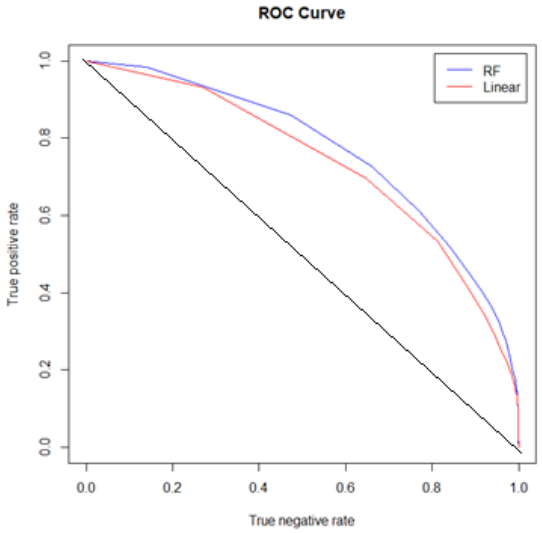

Table 9 ROC curve – comparison of RF & Logistic approach

7 Random Forest as direct classifier

Beside "regression mode" the RF approach can be used as a classifier, with no need to define a cut-off. Recovery events are directly determined by the branching algorithm. The outcome is summarized in Table 10.

		Fit			
		Non-Rec	Recovery	**Total**	Portfolio-mix
Real	Non-Rec	20,567	41	20,608	97.1%
	Recovery	541	85	626	2.9%
	Total	21,108	126	**21,234**	
	Portfolio-mix	99.4%	0.6%		
	Precision	97.4%	67.5%		
	KO	2.6%	32.5%	**Tot. acc.**	
	OK on sample	99.8%	13.6%	**97.3%**	

Table 10 RF approach as a classifier

This approach produces the best results by far in terms of true forecasting performance (67,5% vs the usual 22-24%.). However, this comes at the expenses of an extremely careful classification of positive events: only 126 are selected, generating a very different portfolio mix from the real one (0.6% vs 2.9%). Consistently, the large number of undetected recoveries reduces the performance on the recovery set 13,6% (vs 25,2% obtained in regression model).

8 Overfitting test

Overfitting happens when a model learns the noise in the data to the extent that the performance of the model on new data is poor. To test this feature, the given portfolio is divided into two sub-portfolios, referred to as train and test portfolios. The characteristics of the two portfolios are summarized in Table 11. A new model is estimated on the train portfolio with the same approach. To keep the number of recoveries close to the actual number of recovered dossiers, the 12% cut-off is chosen, yielding 415 fitted recoveries.

	Dossier	**Recovery**	**Prob. Recov**	**Exposure/000**	**Recovery/000**
Train Set	14,014	413	2.95%	14,907	284
Test Set	7,220	213	2.95%	7,993	152
Total	21,234	626	2.95%	22,900	437

Table 11 Test and Train Portfolios

		Fit			
		Non-rec	Recovery	**Tot**	Portfolio-mix
Real	Non-rec	13,285	316	13,601	97.1%
	recovery	314	99	413	2.9%
	Total	13,599	415	**14,014**	
	Portfolio-mix	97.0%	3.0%		
	Precision	97.7%	23.9%		
	KO	2.3%	76.1%	**Tot. acc.**	
	OK on sample	97.7%	24.0%	**95.5%**	

Table 12 Extended Confusion Matrix for train Portfolio

		Fit			
		Non-rec	Recovery	**Tot**	Portfolio-mix
Real	Non-rec	6,825	182	7,007	97.0%
	recovery	153	60	213	3.0%
	Total	6,978	242	**7,220**	
	Portfolio-mix	96.6%	3.4%		
	Precision	97.8%	24.8%		
	KO	2.2%	75.2%	**Tot. acc.**	
	OK on sample	97.4%	28.2%	**95.4%**	

Table 13 Extended Confusion Matrix for Test Portfolio

Table 12 and Table 13 present the extended confusion matrix for train and test portfolios. The test portfolio offers a rough measure of the model's ability to replicate the train portfolio, forecasting 3.4% of recovery events versus 3%. Furthermore, it underlines approximately similar fraction of correctly separately forecasted success and fail cases (OK) and overall (OK on sample).

9 Recovery rate model

The model estimating the recovery rate, for those dossiers for which a recovery occurs, is developed with a standard multilinear approach, on the basis of information available on relevant recovered dossiers. For this model, no RF alternative is provided. The recovery rate (RR) is defined in the same way at granular and aggregate level as:

$$(7) \qquad RR = \frac{Recovered\ Amount}{Exposure}$$

The geographic variables (macro region) are maintained in the model summarized below, even with low relevance, as they characterize the final level of aggregation. The *Adjusted R-squared* (10.4%)[9] of the selected model confirms that the overall performance is quite low but this does not prevent a good forecast of overall recoveries at aggregate level (macro-region), as shown in Table.

Variable	**P-value**	**Relevance**
Intercept	0.000	High
Exposure	0.000	High
Contacts	0.000	High
Dossier age	0.002	Middle
Center-south	0.090	Low
Island	0.104	Low
North	0.618	Low

Table 14 Recovery Rate – Linear Model

[9] The model is consistent: at a more technical level: the F-statistic yields 13.12 to be compared with the critical f-value based on 6 and 619 degrees of freedom (over 120 DF the value is 1,774) to confirm that the model is statistically significant.

Macro region	%dossier	% Exposure	Real rec. rate	Fitted rec. rate
North	31.4%	33.1%	63.77%	65.47%
Center North	10.9%	10.6%	66.91%	64.60%
Center South	42.9%	41.1%	61.11%	59.67%
Island	14.8%	15.3%	58.60%	58.16%
TOTAL	21,234	22,900,489	62.6%	62.34%

Table 15 Comparison of Recovery Rate Outcome

The overall forecast for the Recovery Rate (RGR) is quite good (62.34% vs 62.6%). The match holds also at geographic area level, with the largest gap occurring for Centre-North where the model the underestimates recovery rate by 2.3% on average.

10 Results

In order to measure the overall performance of the framework presented, recovery estimates at dossier level are summed up over the aggregating perimeter (macro-area) and then compared to the amount effectively recovered on the same perimeter,

$$\widehat{ER} = \sum_1^N E_i * I_i * \widehat{RgR_\iota} \tag{8}$$

$$I_i = \begin{cases} 1 \; if \; PR \geq \text{cut} - \text{off} \\ 0 \; if \; PR < \text{cut} - \text{off} \end{cases} \tag{9}$$

where the exposure of the given dossier selected by the recovery event (i) at granular level - a non-linear function of the estimated probability of recovery – is multiplied by the estimated recovery rate for the given dossier ($\widehat{RgR_\iota}$). This is different from the average recovered amount calculated through the average recovery rate over the given perimeter

$$\overline{\text{ER}} = \sum_{i \in \text{real rec}} E_i * \widehat{RgR_\iota} \tag{10}$$

as no exposure weighting is involved in the recovery rate estimate.
Hence the weighted average recovery rate at aggregated level, must be calculated from the expected recovered amount at granular level divided by their corresponding exposure.

$$\widehat{RgR} = \sum_{fit} E_i * I_i * \widehat{RgR_\iota} \Big/ \sum_{fit} E_i * I_i \tag{11}$$

It is worth mentioning that the average probability of recovery (RP) in a given perimeter is the ratio of dossiers recovered real or estimated

$$\overline{RP}_{real/fit} = 1/\text{N} \sum_{i=1}^N \hat{I}_\iota \tag{12}$$

The results in terms of recovery parameters are summarized in the following table, confirming a good fit at aggregated level. Also the average probability of recovery is sufficiently precise, with the RF doing better than the Logit.

	Real		RgR%			PR%		
Macro region	Dossier #	Recovery/000	Real	Fitted	Real #	Real %	Logit %	**RF optimal**
North	6,665	166,6	63.8%	65.5%	211	3.2%	3.1%	3.1%
Center North	2,308	64,9	66.9%	64.6%	87	3.8%	4.6%	3.2%
Center South	9,109	154,2	61.1%	59.7%	246	2.7%	1.7%	2.5%
Island	3.152	50,8	58.6%	58.2%	82	2.6%	1.4%	2.8%
TOTAL	**21,234**	**436,5**	**62.6%**	**62.3%**	**626**	**2.9%**	**2.4%**	**2.8%**

Table 16 Fitted vs real recovery parameters summary

When a specific cut-off is defined for each approach, the resulting PR at macro-region level (2.8% RF optimal) gets very close to the real one (2,9%) and presents a more regular behaviour compared to that of the logistic approach which is also less precise (2.4%).
In terms of amounts recovered, the results are summarized in the following table, which compares the different components. Notice that the expression '*exp*Real PR*' next to exposure of effectively recovered (*Real*), represents an average based on the real probability of recovery defined at macro-region level. The two amounts differ as the last is not weighted for the relative exposure.

	amounts /000		Exp of recovered[10] /000				Tot recovery /000			
Macro region	Exposure	%Exp	Real	exp * Real RP	Logit	**RF optimal**	Real	%Real	Logit	**RF optimal**
North	7,572	33.1%	261.2	239.7	344,1	263,4	166,6	38.2%	229,9	160,6
Ctr North	2,433	10.6%	97.0	91.7	148,3	80,1	64,9	14.9%	97,2	46,7
Ctr South	9,402	41.0%	252.4	253.9	245,5	261,1	154,2	35.3%	140,6	135,9
Iseland	3,494	15.3%	86.7	90.9	93,9	144,3	50,8	11.6%	24,7	45,9
TOTAL	22,900	100%	697,3	676,2	831.7	748.8	436.5	100%	492.4	389,1

Table 17 Fitted vs real recovery amount summary

As for the recovery rate, the exposure of recovered dossiers estimated with the RF approach, driving the total recovery, is closer to the real one with respect to the Logit (748.8 vs 831.7 €). Considering the amounts recovered instead of the dossier exposure, the RF approach undershoots the real amount, but still performs slightly better than the Logit.

Conclusions

This work analyzed the collection profiles in a portfolio of retail unsecured bills managed by a servicer, with the purpose of modelling the recovery capabilities on the basis of the limited information available on location and personal characteristics of the debtors. The exercise compares a RF approach to a classical Logistic estimate, using extended Confusion tables, besides other metrics, to compare their performances. To avoid averaging the recovery rate over a high number of non-recovery events, estimates are based on a framework that separates the recognition of recovery cases from the estimate of the recovery rate. The two components are analyzed separately. Although, in statistical terms, the performance of the model is generally low, mainly due to the very limited information available, at aggregate level the results are surprising. When the RF approach is used to identify recovery events, the results confirm the model to be robust with performance slightly better, than those of the standard Logistic, and with no need to define any specific conditions on explicative variables. Finally, the framework is suitable to estimate the portfolio value, both in absolute and relative terms, i.e. compared to a known portfolio. While there is a handful list of artificial intelligence methods which require large datasets, this work could be extended by applying some other machine learning technics, including support vector machine (SVM). As a classifier, SVM relies on the concept of distance between different points, a situation very similar to the characteristics of the data set considered in this work. However, as reported by (Wyner, et al., 2017), RF remains most likely the best classifier,.

Furthermore, the authors propose to combine Adaboost method with RF, using RF as the weak learner in the process for selecting the high weight instances during the boosting process. As reported by (Thongkam, et al., 2008), the proposed method outperforms a single classifier and other combined classifiers for the breast cancer survivability prediction.

Massimiliano Zanoni

Bibliography

Bellotti, A., Brigo, D., Gambetti, P. & Vrins, F., 2019. Forecasting recovery rates on non-performing loans with machine learning. s.l., s.n.

Breiman, L., 1996. Out-of-Bag Estimation, s.l.: Statistics Department, University of California, Berkeley.

Breiman, L., 2001. Random Forests, s.l.: Statistics Department, University of California, Berkeley.

Ciavoliello, L. G. et al., 2016. What's the value of NPLs?. s.l., Banca D'Italia, Eurosistema.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters.

Gareth, J., Witten, D., Hastie, T. & Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. s.l.:Springer.

Grömping, U., 2009. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second edition. s.l.:Springer.

Khieu, H., Mullineaux, D. J. & Yi, H.-C., 2012. The Determinants of Bank Loan Recovery Rates. Journal of Banking and Finance.

Thongkam, J., Xu, G. & Zhang, Y., 2008. AdaBoost algorithm with random forests for predicting breast cancer survivability. s.l., International Joint Conference on Neural Networks (IJCNN).

Wyner, A. J., Olson, M., Bleich, J. & Mease, D., 2017. Explaining the Success of AdaBoost and Random Forests as. Journal of Machine Learning Research, Volume 18.

Youden, W., 1950. Index for rating diagnostic tests. Cancer.

[10] The exposure of recovered dossiers in any model is the sum over dossier recovered according to the specific model and cut-off.

RISK MANAGEMENT MAGAZINE

Anno 15, numero 1 **Gennaio – Aprile 2020**

Poste Italiane - Spedizione in abbonamento postale – 70% aut. DCB / Genova nr. 569 anno 2005

TESTATA INDIPENDENTE CHE NON PERCEPISCE CONTRIBUTI PUBBLICI (legge 250/1990)

In collaborazione con REFINITIV

IN QUESTO NUMERO

Risk Management Magazine
Anno 15 n° 1 Gennaio - Aprile 2020

Direttore Responsabile:
Maurizio Vallino

Condirettore
Corrado Meglio

Proprietà, Redazione e Segreteria:
Associazione Italiana Financial Industry Risk Managers (AIFIRM), Via Sile 18, 20139 Milano

Registrazione del Tribunale di Milano n° 629 del 10/9/2004

ISSN 2612-3665

E-mail: risk.management.magazine@aifirm.it; Tel. 389 6946315

Stampa: Algraphy S.n.c. - Passo Ponte Carrega 62-62r 16141 Genova

SPEDIZIONE IN ABBONAMENTO POSTALE AI SOCI AIFIRM RESIDENTI IN ITALIA, IN REGOLA CON L'ISCRIZIONE

Rivista in stampa: 8 Aprile 2020

Rivista accreditata AIDEA