Decomposing Loss Given Default:

A Closer Look at Recovery Patterns

Aida Salko
Department of Economics and Social Sciences
Sapienza University of Rome, Italy
E-mail address: aida.salko@uniroma1.it

Rita D'Ecclesia
Department of Statistics
Sapienza University of Rome, Italy
E-mail address: rita.decclesia@uniroma1.it

Abstract

This study integrates cures, partial recoveries, and write-offs in modeling Loss Given Default (LGD) and investigates the performance of different algorithms in estimating each component of the decomposed approach. We use a unique database of defaulted real estate-backed loans in European countries. The aim of this study is to accurately estimate the ultimate recovery rate, hence the LGD, by using various machine learning methods including random forest, k-nearest neighbor, extreme gradient boosting, and multivariate adaptive regression splines. We find that the new models we used to estimate each component of the equation, outperform the traditional statistical models such as logistic regression or OLS, and in particular, random forest leads with the highest performance among all models in terms of both in-sample and out-of-sample results. The results confirm that using the random forest in this multiple-step modeling of the recovery rate could improve the whole recovery rate estimation performance.

Keywords: Loss Given Default, Global Credit Data, Probability of Cure,
Probability of Write-off, Machine Learning.

***JEL Classification**: C53, G21, G32*

1 Introduction

Finding a proper methodology that offers high predictive accuracy for Loss Given Default (LGD) has been a challenge for many academics and practitioners over the last years. The introduction of the Basel II framework has stimulated a growing literature focusing on LGD estimation, as one of the main parameters of credit risk when building the banks' internal models. Considering that the LGD appears to be usually bi-modally distributed, this makes its estimation even more challenging. Several traditional statistical models as well as recent innovative algorithms of machine learning have been used in modeling and predicting LGD. The LGD models in the existing literature are mainly divided into one-stage and two-stage modeling. The majority of the previous studies are generally focused on the one-stage modeling framework of LGD, while recently two-stage methods are becoming popular and a promising approach in the literature by offering higher accuracy and better performances in LGD estimation[1]. These methods consist of a combination of binary decision models and regression models. In other words, these techniques are composed of two models that detect full-loss or no-loss cases, considering them as classification problems, and cases in the middle ($0< \text{LGD} <1$) separately. Logistic regression is the method that is mostly used as the first model in the two-stage setting,[2] while several linear and non-linear regression methods are used in the second stage. However, one of the main drawbacks of logistic regression stands in producing biased and inconsistent results if the model is not specified correctly. In addition, the existence of non-linear relationships between explanatory variables and LGD as confirmed by some studies[3] may produce even weaker results if logistic regression is used (Tanoue et al. 2020).

One of the main weaknesses of the existing two-stage models stands in neglecting the recovery patterns between the main two modes of LGD (i.e. 0 and 1). To overcome this, Starosta (2021) proposes a new LGD decomposition by integrating cures, partial recoveries and write-offs into one equation and find that the proposed model performs better by offering higher effectivity. The author uses a traditional ordinary least squares regression

[1] Loterman et al. (2012), Tanoue et al. (2017, 2020), Starosta (2021).

[2] Bellotti & Crook (2010), Matuszyk et al. (2010), Gürtler and Hibbeln (2011), Loterman et al. (2012).

[3] Loterman et al. (2012), Yao et al. (2015).

(OLS), a logistic regression, classification and regression trees (CART), and support vector machines (SVM) to estimate each component of the proposed equation. He finds empirical evidence that the decomposed model reveals better predictions in terms of out-of-sample predictive metrics and the combination of a classification tree and the regression tree produces the best overall results.

We follow the work developed by Starosta (2021) in decomposing LGD using mixture distributions of in-default events and apply different algorithms that may lead to more precise and robust estimates for each component of the decomposed model of LGD for real estate-backed defaulted loans. We aim to improve all the components in this multiple-step LGD estimation model using several machine learning methods.

Our first contribution stands in the application of several innovative models that have proved to be successful in many fields and provide higher predicative accuracy compared to traditional statistical models. Therefore we use a random forest, extreme gradient boosting, k-nearest neighbor, and multivariate adaptive regression splines (MARS) to estimate each component of the proposed equation by Starosta (2021). Moreover, we use a new set of different accuracy measures such as the Brier score and AUC (ROC accuracy ratio) as well as the Mathews correlation coefficient to compare the predictive performance between the models for the probability of cures and write-offs. In addition, a special focus is given to the real estate serving as the collateral for the loan, by including specific information in the models including the real estate types and their location. Furthermore, to be in line with the LGD downturn estimation as required in the Basel framework, apart from several macroeconomic variables, we also include the real house price index and news-based uncertainty index which are particularly related to the real estate-backed loans. Finally, we use a unique database of defaulted bank loans provided by Global Credit Data[4] (GCD), and use the LGD data for all the European countries, instead of only one bank in one country as it is commonly found in LGD works. In this way, in terms of a regional spread, we provide new evidence of the effectiveness of this model for all the European areas.

[4] Global Credit Data provides the largest LGD data base worldwide. The association consists of 55 banks from all over the world. See http://www.globalcreditdata.org/ for further information

To the best of our knowledge, this is the first LGD study on a European level, focusing particularly on real estate-backed loans including cures, write-offs, and partial recoveries as proposed by Starosta (2021) by applying different innovative algorithms that may lead to more precise estimates.

We find that the new models we used to estimate each component of the LGD equation, outperform the traditional statistical models such as logistic regression, model trees or OLS, and in particular, random forest leads with the highest performance among all models in terms of both in-sample and out-of-sample results. The results confirm that using the random forest algorithm to model each component of the decomposition approach for the recovery rate estimation, could improve the whole LGD estimation performance.

The remainder of the paper is structured as follows. Section 2 presents a literature review of different works that have used one-stage and two-stage LGD models. Section 3 describes the decomposed model while section 4 and 5 explains the estimation methods and the measures used to assess the quality of the models. Section 6 presents a brief description of our dataset with a special focus on the recoveries on cures, write-offs, and partial recoveries as well as information related to the real estate serving as collateral. Section 7 and section 8 report the empirical results and the concluding remarks.

2 One-Step and Two-Step models of LGD

The LGD estimation models in the existing literature are mainly divided into one-step modeling approaches and two-step modeling methods. One-step methods consist of modeling LGD directly depending on different independent variables using various techniques, while two-step methods are composed of two models that separate full-loss or no-loss cases (i.e. LGD=1 or LGD=0) and cases in the middle (0< LGD <1) separately.

The main focus of one-step methods is not just offering a good predictive performance but also exploring the main potential determinants or risk drivers affecting LGD.

Dermine and De Carvalho (2006) apply a multivariate approach on defaulted bank loans to analyze the determinants of recovery rates and find several factors including the size of the loan, collateral, industry sector, and the age of the firm to be statistically significant explanatory variables.

Caselli et al. (2008) use univariate analysis and highlight the importance of LGD downturn conditions by finding evidence on the relationship between LGD and several macroeconomic conditions of bank loans in the Italian market. Grunert and Weber (2009) apply a simple linear regression analysis and find the role of quota collateral to be positively related to the recovery rates while other factors such as the risk premium of the borrower and the size of the company are negatively related. Qi and Yang (2009) investigate high loan-to-value residential mortgages and use a general regression equation to model LGD as a function of loan and property characteristics as well as housing market conditions. The authors find evidence of higher losses during distressed housing markets. Bastos (2010) use a fractional response regression and a regression tree to forecast the bank loans credit losses. In terms of the predictive performance, he finds the regression tree to produce better results in comparison to traditional parametric models which are mostly used in estimating LGD. In a later study, Bastos (2014) applies an ensemble strategy for predicting recovery rates of defaulted debt and observes better forecasting performance compare to a single model. The superiority of these innovative algorithms of machine learning in one-step models was also found by Qi and Zhao (2011) who compare six different modeling methods for LGD and find regression trees and neural networks to outperform all the other methods. Hartmann et al. (2014) investigate the LGD for leasing and show that model trees produce better results in terms of out-of-sample performance.

Many other studies have proposed two-stage models for LGD. Bellotti and Crook (2010) propose a decision tree approach where extreme cases with no-loss (LGD=0) and full loss (LGD=1) are considered as binary classification problems and are modeled through logistic regression. For the rest of the cases, when $0< \text{LGD} <1$, the authors use a simple OLS model. Also, Thomas et al. (2010) follow the same approach on splitting the LGD data considering them as classification and regression problems and applying then a logistic and linear regression.

Matuszyk et al. (2010) use a decision tree method including a two-step approach to model LGD of unsecured consumer loans by focusing particularly on the collection process. The authors highlight the importance of modeling both the decisions by the lenders and the repayment risks of the debtors when estimating LGD. To achieve this, they propose a two-step process where in the first step, logistic regression is used to estimate the class of a debtor and in a second step, the LGD for each class is estimated using a regression model.

Gürtler and Hibbeln (2011) propose a two-step LGD modeling approach based on the different influencing factors of loans that are recovered and those written-off. The authors first apply logistic regression to distinguish between two groups of loans and on a second step use regression for the LGD of each of them separately.

Loterman et al. (2012) present an LGD benchmark study by investigating 24 different regression techniques and find the non-linear techniques in combination with a linear model component in a two-stage process, to have good predictive power. They propose two approaches of the two-step setting: i) the first one by using logistic regression to model extreme cases when LGD=0 and LGD=1, and as a second stage applying different linear and non-linear techniques to model the values in between, and ii) a simple OLS regression as a first stage and estimation the residuals using a non-linear regression model as the second stage. The authors state that an advantage of the two-stage setting stands in improving the comprehensibility of the resulting models. In addition, they find a clear trend on the superiority of non-linear techniques, the support vector machines, and neural networks in particular, in terms of performance results.

Tanoue et al. (2017) apply a multi-stage model and confirm their superior predictive accuracy relative to the OLS, Tobit, and inflated beta regression models. In a later study, Tanoue et al. (2020) investigate the performance of several probability machine models as the first model in the two-step LGD estimation model. Particularly, the authors separate LGD positive values from zeroes using a random forest, k-nearest neighbors (KNN), bagged nearest neighbors (BNN), and support vector model (SVM) as the first model, and a simple linear regression estimating the positive values given LGD > 0 as the second model. They find empirical evidence that random forest results as the best model when building the first step in the two-stage setting of the LGD estimation model.

3 The decomposed model

The model proposed by Starosta (2021) integrates cures, write-offs, and partial recoveries in one equation. We present a short summary of the model and give the definition of some concepts that will be used throughout the paper.

A default is marked as a cure event when it has time to resolution less than one year, no write-off and no collateral sale or guarantee call.[5] In other words, the borrower has exit the default status and returned to performing portfolio so we consider this as Stage 1 (s_n=1, where *n* is the observation) of the client. No collateral information is required in this stage. In case the default goes further into the process, so the event is not considered as cured anymore, the collection department takes place by resulting in partial recoveries (excluding write-offs). We consider this as Stage 2 (s_n=2) and detailed collateral information is required at this point. Apart from that, information regarding the loan file (exposure at default, facility type, guarantee indicator, seniority, etc.) and borrower-related characteristics (industry, financial health information, residence, etc.) are also considered at this stage. At this stage, the default may end (if the bank demands are met) or it can proceed to Stage 3 (s_n=3) which is then considered as a write-off. Loan-related information is mainly required for stage 3.

If we denote $s_n \in \{1, 2, 3\}$ the stage for the *n*-th exposure;

Probability of cure:

$$P\ (s_n\text{=}1)\text{=}\ \psi_{CURE}(\beta_{1,n}); \qquad (1)$$

[5] This is the cure definition as per GCD methodology.

Probability of write-off:

$$P(s_n=3|s_n \neq 1)= \psi_{WRITE-OFF}(\beta_{2,n}); \tag{2}$$

Probability of partial recoveries:

$$P(s_n=2|s_n \neq 1)= 1 - P(s_n=3|s_n \neq 1); \tag{3}$$

Expected RR for cures:

$$E(RR|s_n=1) = \mathcal{E}_{CURE}(\beta_{3,n}); \tag{4}$$

Expected RR for partial recoveries:

$$E(RR|s_n=3|s_n \neq 1)=\mathcal{E}_{PARTIAL}(\beta_{4,n}); \tag{5}$$

Expected RR for write-offs:

$$E(RR|s_n=3|s_n \neq 1) = \mathcal{E}_{WRITE-OFF}(\beta_{5,n}); \tag{6}$$

where $\beta_{1,n}, \beta_{2,n}, \beta_{3,n}, \beta_{4,n}, \beta_{5,n}$ present the set of explanatory variables for *n*-th exposure for each component (Eq.1-6). Figure 1 shows a short summary of these variables that will be used for each stage of the recovery process. Macroeconomic variables are assumed to affect all the components at each stage.

Finally, in line with Loterman et al. (2012) and Starosta (2021), the expected ultimate recovery rate (EURR) is expressed as a combination of all the components mentioned above, as follows:

$$\begin{aligned} EURR = E(RR_n) = {} & \mathcal{E}_{CURE}\left(\beta_{3,n}\right) x \, \psi_{CURE}\left(\beta_{1,n}\right) + \\ & \left(1 - \psi_{CURE}\left(\beta_{1,n}\right)\right) x \, \mathcal{E}_{PARTIAL}\left(\beta_{4,n}\right) x \, \left(1 - \psi_{WRITE-OFF}\left(\beta_{2,n}\right)\right) + \\ & \mathcal{E}_{WRITE-OFF}\left(\beta_{5,n}\right) x \, \psi_{WRITE-OFF}(\beta_{2,n}) \end{aligned} \tag{7}$$

The final result of the equation gives the recovery rate, which can easily produce the LGD, as 1 – EURR. All the components of equation (7) are estimated and combined for each default event.

Figure 1. Variables used for each component of LGD decomposition.

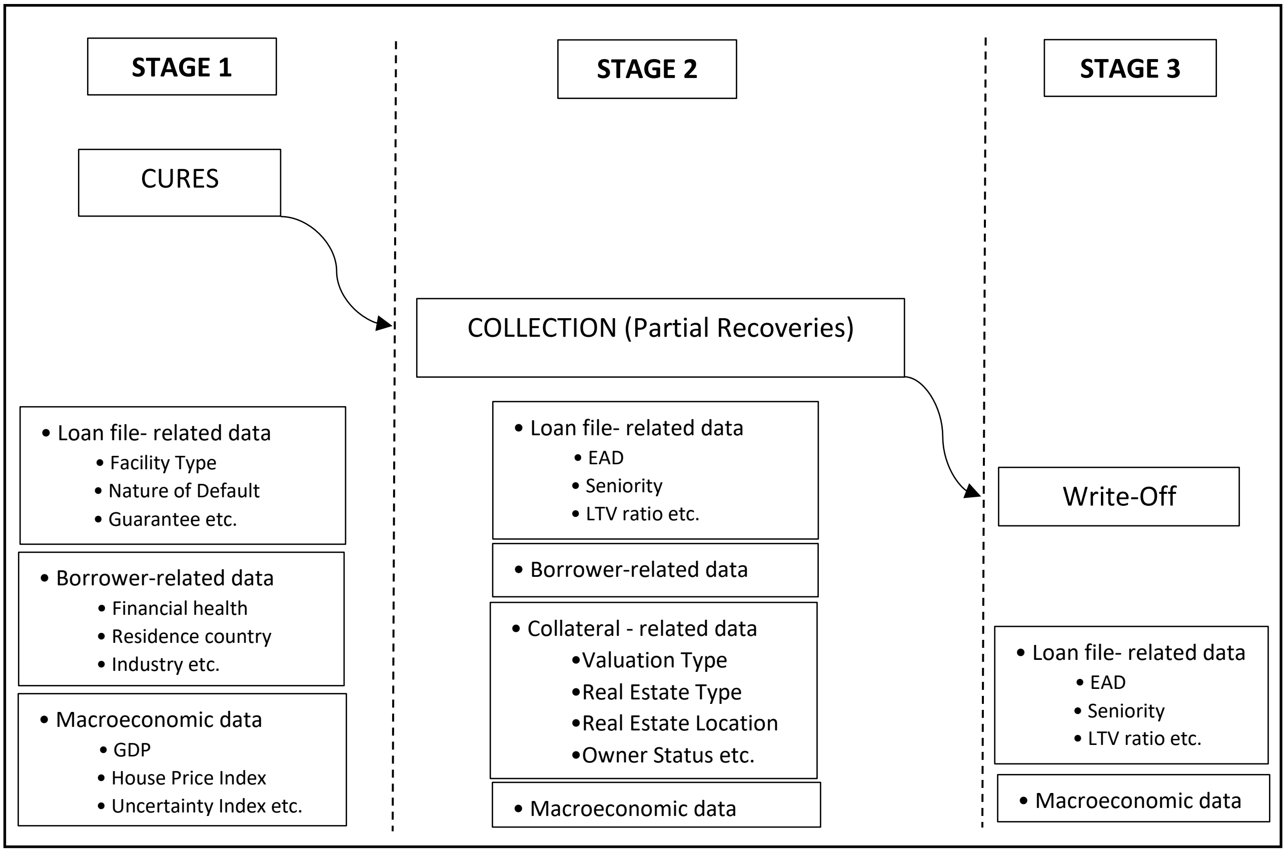


Table 1 presents a short summary of all the models that will be used for classification (ψ_{CURE} and $\psi_{WRITE-OFF}$) and regression ($\mathcal{E}_{CURE,}$ $\mathcal{E}_{WRITE-OFFS}$, and $\mathcal{E}_{PARTIAL}$) functions, as well as all the respective performance measures for each of them, which will be described in detail in the next paragraph.

Table 1. LGD decomposition, the used methods, and metrics used for assessing model quality

Objective	Components	Methods used	Performance Metrics
Expected Ultimate Recovery Rate (EURR)	1) Probability of cure $\psi_{(Cure)}$ 2) Probability of write-off $\psi_{(Write-Off)}$	• Logistic Regression • Decision Tree • K-Nearest Neighbors • Random Forest • MARS • Extreme Gradient Boosting	• ROC Curve • AUC • Brier Score • MCC
	3) Expected RR for cures $\varepsilon_{(Cures)}$ 4) Expected RR for write-offs $\varepsilon_{(Write-Offs)}$ 5) Expected RR for partial recoveries $\varepsilon_{(Partial)}$	• OLS • Regression Tree • K-Nearest Neighbors • Random Forest • MARS • Extreme Gradient Boosting	• RMSE • MAE • R^2 • ρ

4 Methodology

4.1 Logistic Regression

Logistic regression is one of the most used methods in classification problems when modeling binary responses. Apart from other models that will be explained below, for comparison purposes, we also use logistic regression to model the probability of cure and write-off, as Starosta (2021) in his study. If we denote with *p* the probability of an event, the odds of the event as *p/(1-p)*, then the logistic regression models the log odds of the event using a linear function as follows:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \qquad (8)$$

where *n* is the number of predictors. The event probability is then written as a sigmoidal function:

$$p = \frac{1}{1 + e^{[-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)]}} \qquad (9)$$

where the $\beta_0, \beta_1, \ldots \ldots \beta_n$ are estimated using the maximum likelihood method.

4.2 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is a non-parametric and non-linear regression method introduced by Friedman (1991). The main idea of this modeling technique stands in building multiple linear regression models across the range of predictor values. The MARS algorithm is considered as an extension of linear models but it makes no assumptions about the relationship between the response variable and the predictor variables. The algorithm builds the models in two steps: first, it starts by partitioning the data, and second, it runs a linear regression model on each different partition.

In the first step, the algorithms create a range of predictor values which is partitioned into several groups, and for each of these, a separate linear regression is modeled. The connections between the separate regression lines are referred to as knots. Then, the idea of the MARS algorithm is to search for the best spots to place the knots. The model can be expresses as the following problem:

If *y* is the target output and $X = (X_1, \ldots, X_N)$ is a matrix of *N* input variables, let's assume that the data are generated from an unknown "true" model which would be expresses as:

$$y = f\left(X_{1,\ldots\ldots}X_N\right) + e = f(X) + e \qquad (10)$$

where *e* is the distribution of the error. The function *f* is then approximated by applying some basis functions, which are splines (smooth polynomials), including piecewise linear and piecewise cubic functions. It can be formally written as:

$$max(0\,, x - t) = \begin{cases} x - t \;\; if\; x \geq t \\ 0, otherwise \end{cases} \qquad (11)$$

indicating that only the positive part of the equation is used otherwise it is given a zero value. Finally, the MARS model *f(X)* is expressed as a linear combination of basis functions and their interactions:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m \lambda_m(X) \tag{12}$$

where each $\lambda_m(X)$ is a basis function and the β coefficients are estimated using the least-squares method.

4.3 Classification and Regression Tree (CART)

Regression trees are used by many authors and have resulted to produce good accuracy in forecasting credit recoveries.[6] A regression tree begins with a “root” node containing all the observations and then searches all over the possible binary splits among the data to find the explanatory variable and its corresponding value for the splitting that minimizes the squared errors in case of regression or Gini index in case of classification. This approach aims to divide the data into groups in which the LGD is as homogenous as possible. Classification trees, or the so-called decision trees (DT), are used to estimate the probability events, i.e. the probability of cure and the probability of write-off, while and regression trees (RT) are used to estimate the recovery rates (4), (5) and (6).

4.4 Random Forest (RF)

Considered as an evolution of Breiman’s original bagging algorithm, the random forest is considered a popular ensemble strategy that incorporates randomized feature selection. First introduced by Breiman (2001), random forest is a powerful rule-based algorithm formed as an ensemble of decision trees where each tree is trained on a different artificially created sample. All the decision trees that form the random forest are different since each tree is built on a different random subset of data. However, random forest produces a final predictor under a different sampling mechanism. The algorithm uses only a random subset of available features is considered at each split. In other words, the random forest uses a random selection of features rather than using all features to grow the trees. This contributes to reducing the correlation and the variance of the ensemble prediction. As a

[6] Bastos (2010, 2014), Matuszyk et al.(2010), Qi & Zhao (2011), Bellotti & Crook (2012), Bellotti et al. (2019), Papoušková & Hajek, (2020)

final result, the random forest will use an average of all single predictors to make a better final prediction. That is, if we have a full set of *n* features, then only a random sample of *m* features is chosen as split candidates when building a random forest. It is important to emphasize that randomness is introduced only in the process of selecting features and not on splitting points of these features.

4.5 K-Nearest Neighbors (KNN)

The k-nearest neighbors is another popular algorithm widely used in classification and regression problems, mainly due to its simplicity. The main idea of the model stands in identifying a number of *K* points that are closest in the input space, represented as $N_k(x)$. The prediction of the target variables is computed as:

$$\hat{f}(x) = \frac{1}{K} \sum_{x_i \in N_k(x)} y_i \tag{13}$$

So the final prediction is nothing else but the average of all the *K* observations that are closest in the training set. The Euclidian distance is used for the closeness between observations.

4.6 Extreme Gradient Boosting (XGB)

Boosting is another ensemble algorithm widely used in many statistical learning methods for regression and classification. Based on the gradient boosting machines algorithms presented by Friedman (2001), this model is a powerful ensemble strategy where the residuals of the model are fitted by many weak learners iteratively. Boosted trees use the original data set to grow trees sequentially. This means that each tree is grown in sequence by using the information from the previously grown tree and therefore depending on the results of the previous trees. Stochastic gradient boosting is an improved version of this algorithm (Friedman, 2002) where at each iteration step is included a random sampling scheme. Extreme gradient boosting is another recent innovative algorithm that is built under the principle of gradient boosting framework. Popular for the speed computation that it offers, extreme gradient boosting also offers parallel tree boosting and controls overfitting through a more regularized model formalization. This is why this model can also be referred to as regularized gradient boosting (Chen et al. 2020).

5 Evaluating and comparing the performance

Evaluating the predictive accuracy of our models is an essential part of the study. In order to assess the performance of our models, we need to quantify how well the predictions actually match the observed data. In the case of classification problems, the confusion matrix, which is a two by two table, is the main tool that shows the correct and incorrect classification of cures and write-off cases. In terms of binary classification, like in our case, it shows the true negatives (TN) and true positives (TP) which are the correct predictions, versus false negatives (FN) and false positives (FP) which present the incorrect predictions. Table 2 shows an example of the confusion matrix and its elements for the classification of cured events (same for write-offs).

Table 2. Confusion Matrix

		Predicted	
		CURED	NON-CURED
Observed	CURED	TP	FN
	NON-CURED	FP	TN

There are several classification metrics that can be derived from the confusion matrix, but we use the Matthews correlation coefficient (MCC) as a statistical metric that takes into account all the four values of the confusion matrix. The MCC is calculated as:

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

The range of the MCC is between -1 and 1: closer to one indicates better classification and better model performance. The advantage of this coefficient is that it is considered as a balanced measure meaning that it can be used even if the classes are highly unbalanced, i.e. one class is over- (or under-) presented.

Another important metric based on the confusion matrix is the receiver operating characteristic (ROC) curve which is a probability curve and is one of the most important evaluation metrics that is used to check and visualize any classification model's performance. It is constructed by plotting the sensitivity (i.e. the true positive rate computed as TP/(TP + FN) against the specificity (i.e. 1-false positive rate computed as FP/(TN + FP). A standard measure to compare the ROC curves is the area under the curve (AUC). Its values are between 0 and 1. The higher the AUC, the better is the model.

Last, we use the Brier Score to evaluate the accuracy of probabilistic predictions. The Brier Score for binary classifications for a set of predictions is given as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(P_i - o_i)^2 \qquad (15)$$

where n is the number of forecasts, P_i is the predicted probability of the event i and o_i represent the occurrence of the event i.e. 1 if the event occurs or 0 if not. The results of the score take values between 0 and 1, where the lowest values of the score (close to zero) indicate better model prediction.

For the regression models ($\mathcal{E}_{CURE,}$ $\mathcal{E}_{WRITE-OFFS}$, and $\mathcal{E}_{PARTIAL}$), we use the root mean squared error (RMSE) and the mean absolute error (MAE) as the most commonly used measures of model performance. The root mean squared error (RMSE) is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (16)$$

where y_i is the actual recovery rate on loan i; the $\hat{y}$ is the predicted recovery rate on loan i and n is the number of loans in our sample. Models that have lower RMSE tend to give smaller differences between the predictor and the actual value and therefore predict recoveries more accurately. The mean absolute error (MAE) which shows on average, how far is the model prediction from the true value is given as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \, y_i - \hat{y}_i | \qquad (17)$$

Finally, we also include the statistical correlation between the actual and predicted values:

$$\rho = \frac{Cov\ (y, \hat{y})}{\sqrt{Var\ (y)\ Var\ (\hat{y})}} \tag{18}$$

However, we are interested to assess the RMSE and MAE on a sample that is independent of that used in building the models. To achieve this, we will split our sample into two sets using a standard 70% - 30% random split. The first set is used to fit the model, i.e. the training set, and the second one is used to test its accuracy, i.e. the test set. Following this, the performance measures mentioned above are assessed in both sets. In addition, each model hyper-parameters were tuned by using ten-fold cross-validation on the training set. All the models were trained using the latest version of the Caret library in R (Kuhn, 2008, 2018; Kuhn & Johnson, 2013).

6 Data

We use a unique loss database provided by GCD. The GCD association consists of 55 member banks including several global important banks, from all over the world and the data collected comes across the span of 20 years. The data that is used in our paper is of real estate-backed loan defaults and it provides us with a great variety of information on (1) the defaulted borrower, (2) the characteristics of the real estate serving as collateral, and (3) loan-related factors. In our paper, we analyze all the defaulted loans whose country of jurisdiction is located in European countries.[7]

According to Basel II definition, a default occurs if an obligor is “unlikely to pay” or “past due more than 90 days on any material credit obligation”. We refer to the default definition set by Basel II and therefore we restrict our data sample from the year 2000 in order to ensure a consistent default definition. In addition, we do not account for defaults after 2019 since the workouts of recent defaults are not necessarily completed.

[7] A total of 25 European countries are included in the study.

Including the defaults of the recent years, might lead to an unrealistically long-term average LGD since lots of cases of short workout periods are present.[8] Finally, we eliminate all loans with abnormally low and high recoveries[9] (less than −50% and higher than 150%). A total of 8,755 loans remain.

Table 3. Descriptive Statistics of Empirical Recovery Rates (%)

Statistic	Obs.	Mean	St.Dev.	Min.	Max.
Overall	8,755	80.07	32.69	-31.83	139.72
Cures	2,797	99.37	1.74	94.76	122.92
Write-offs	2,311	48.47	34.35	-31.83	130.90
Partial Recoveries	3,647	85.28	29.59	0.18	139.72

Table 3 present a short summary of recovery rates, distinguishing between cures, write-offs and partial recoveries. The mean, the standard deviation, the minimum and maximum values and the corresponding number of observations are reported for each subset.

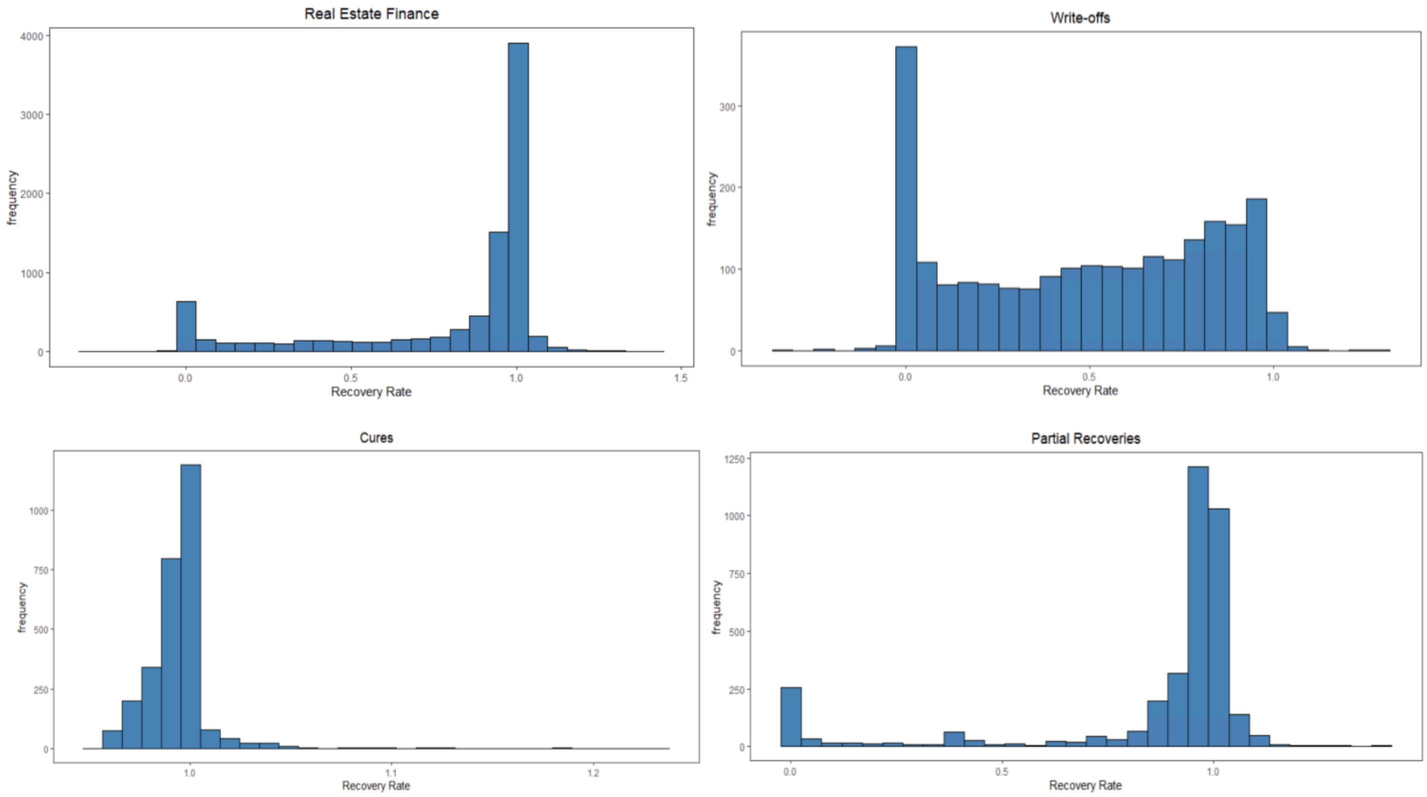
The mean of recovery rates for the overall dataset stands for 0.80 with a minimal value of -0.31 and a maximum of 1.39. We observe that these statistics change considerably for each subsample. Particularly, there is a huge difference in terms of mean recoveries between cures (0.99) and write-offs (0.48). This difference is also presented visually in Figure 3.2 which shows the distribution of the recovery rates for each category.

In addition, we observe recoveries that fall below 0 and exceed the value of 1. Negative values of LGDs as well as values greater than one may appear in some cases due to high costs such as administrative, legal, and liquidation expenses or financial penalties or high collateral recoveries. Finally, we also include a few macroeconomic factors that are also considered in the study so the models are built to be sensitive to macroeconomic characteristics as required by the European Banking Authority (2018).

[8] The resolution bias is addressed according to GCD methodology: https://www.globalcreditdata.org/system/files/documents/gcd_lgd_report_2020_appendix_01062020.pdf

[9] We use the economic recovery rate, with principal advance being part of the defaulted amount.

Figure 2. Distribution of Recovery Rates



We chose the GDP growth, the consumer price index, the unemployment rate, the news-based economic uncertainty index, and the real house price index.[10] In a recent study, Gambetti et al. (2019) find that economic uncertainty turns out to be the most important systematic determinant of recovery rate distributions.

Following Gambetti et al. (2019), we decide to use the original economic policy uncertainty index developed by Baker et al. (2015) which is based on the normalized volume of newspaper articles published in a given month containing expressions referring to economic policy uncertainty. This news-based indicator commonly referred to as the economic policy uncertainty index serves as a proxy for policy-related economic uncertainty.

[10] The macroeconomic data are retrieved from Federal Reserve Economic Data (FRED) https://fred.stlouisfed.org/
and OECD Data https://data.oecd.org/.

In addition, we decide to include the real house price index for European countries, which is the ratio of the nominal house price index (including the sales of newly-built and existing dwellings) to the consumers' expenditure deflator in each European country. [11]

Figure 3. Real Estate Types and Recovery Rates

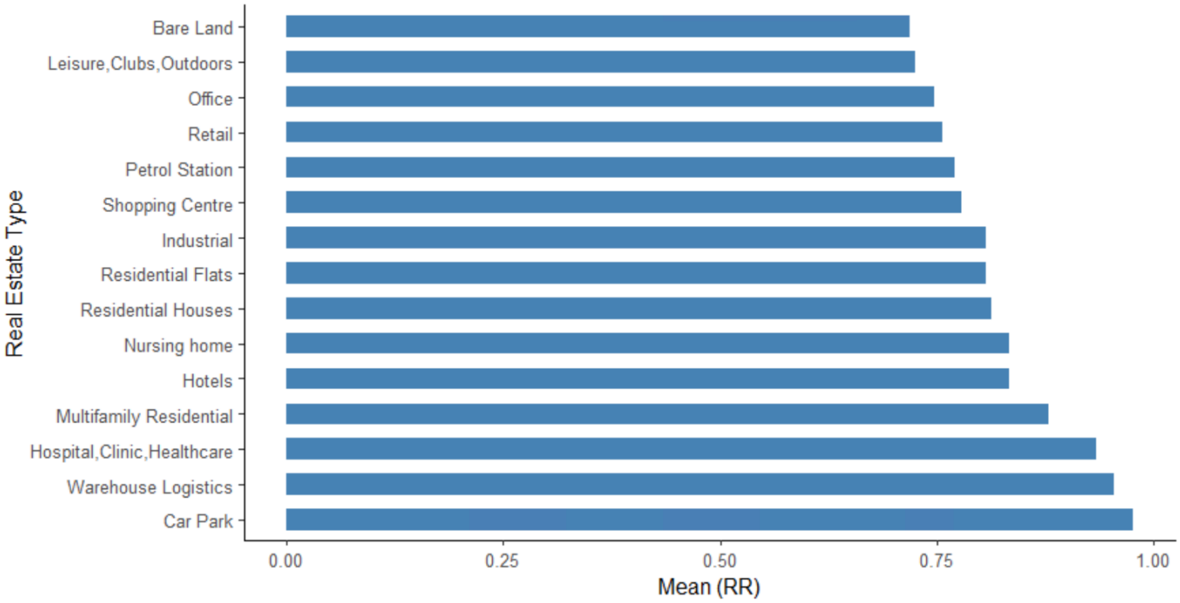


Figure 3 shows the types of the real estate serving as collateral for the defaulted loans in our database and the respective mean of recovery rates of each group. We observe that the collateral types such as the bare land that is due for development, the collateral types used for leisure such as clubs or outdoors, and the collateral used as offices present the lowest recoveries in our data with a mean of 0.71, 0.72 and 0.74 respectively. On the other side, the collaterals used for car parking, warehouse logistics and healthcare including hospitals and clinics are associated with the highest recovery rates with mean of 0.97, 0.95 and 0.93 respectively.

[11] Following the OECD Data definition of housing pricing https://data.oecd.org/price/housing-prices.htm.

7 Results

The performance of the models for the probability of cure and write-offs are presented in Table 4 and Table 5. We find that random forest reveals the best performance in terms of both in-sample and out-of-sample results for the models of the probability of cure and write-off. In addition, Figure 4 shows all the ROC curves and the respective AUC values for all the models. The ROC curves (and AUC) for the out-of-sample results are listed in the Appendix. As can be seen from Figure 4, the random forest is associated with the highest AUC (0.89 for the cure model and 0.9 for the write-off model). It also reveals the best results in terms of the lowest Brier Score (0.10 for the cure model and 0.12 for the write-off model) as well as the highest MMC (0.95 and 0.92 respectively). The other models are also well ahead of logistic regression and the decision tree in terms of all the measures. As the results show, KNN and MARS are the second and the third-best models after the random forest in terms of in-sample and out-of-sample results. We find that the predictive performance of the decision tree and the logistic regression is worse than that of the other machine learning methods.

Table 3. Model performance for the probability of cure model in decomposed approach

Model	AUC		Brier Score		MCC	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
LR	0.76	0.76	0.3465	0.3512	0.3982	0.3800
DT	0.73	0.62	0.3810	0.3839	0.3416	0.3312
KNN	0.83	0.81	0.2135	0.2619	0.6440	0.5290
RF	0.89	0.88	0.1063	0.2429	0.9533	0.6110
MARS	0.83	0.83	0.2701	0.2849	0.5522	0.5414
XGB	0.76	0.75	0.3373	0.3492	0.4032	0.3487

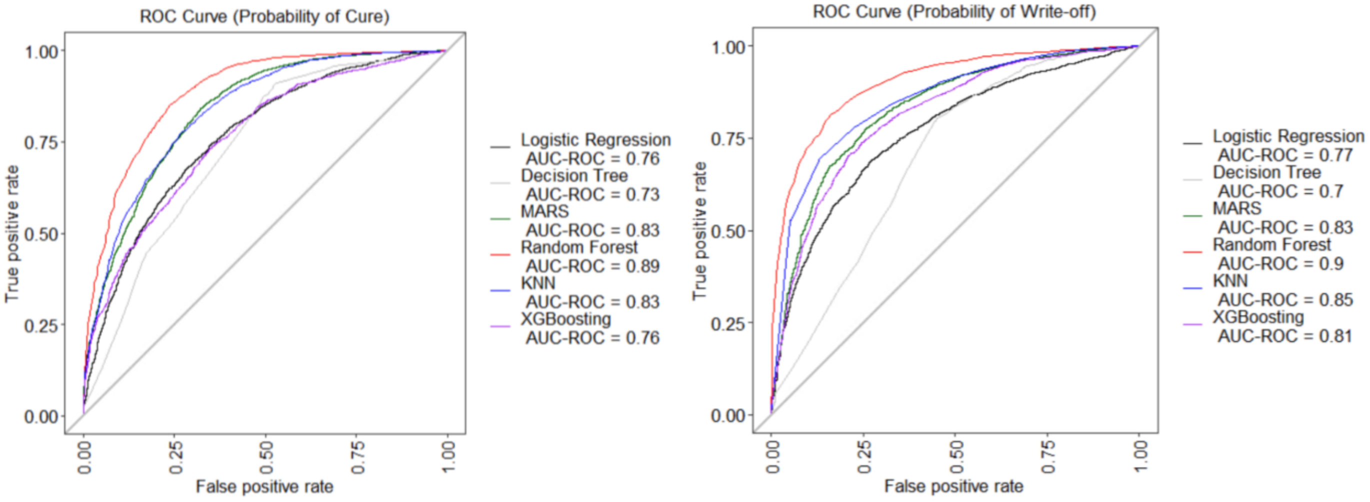
Note: The best model is underlined.

Table 4. Model performance for the probability of write-off model in decomposed approach.

Model	AUC		Brier Score		MCC	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
LR	0.77	0.76	0.3183	0.3275	0.2905	0.2822
DT	0.70	0.69	0.3218	0.3281	0.3748	0.3612
KNN	0.85	0.85	0.2034	0.2316	0.5919	0.5252
RF	0.90	0.89	0.1245	0.2315	0.9233	0.7972
MARS	0.84	0.83	0.2557	0.2609	0.4988	0.4702
XGB	0.81	0.80	0.2806	0.2859	0.4412	0.4352

Note: The best model is underlined.

Figure 4. ROC Curves for Probability of Cure and Probability of Write-off



Tables 6, 7, and 8 present the performance matrixes for estimating the recovery rates of cures, partial recoveries, and write-offs. Referring to the recovery rate of the cure model, in terms of R^2, we find that all the models are well ahead of OLS and RT, with values ranging between 0.30 to 0.89 (in-sample) and 0.29 to 0.58 (out-of-sample). Even here we find that the random forest is superior to all the models in terms of the errors (0.004 RMSE and 0.002 MAE), a correlation coefficient of 0.94 and an R^2 of 0.89. The results show that both, in terms of in-sample and out-of-sample results, the machine learning models outperform OLS and RT, and in particular, the random forest leads to remarkable predictive accuracy.

Table 5. Model performance for RR of cures model in decomposed approach.

Metric	Model	OLS	RT	KNN	RF	MARS	XGB
R^2 (%)							
	In-sample	42.69	30.18	63.99	89.40	53.31	80.86
	Out-of-sample	39.12	29.16	43.13	58.14	39.66	49.64
RMSE							
	In-sample	0.0076	0.0093	0.0067	0.0040	0.0076	0.0051
	Out-of-sample	0.0077	0.0094	0.0085	0.0073	0.0086	0.0079
MAE							
	In-sample	0.0064	0.0067	0.0046	0.0028	0.0056	0.0038
	Out-of-sample	0.0066	0.0069	0.0059	0.0050	0.0062	0.0056
ρ							
	In-sample	0.6533	0.5493	0.7999	0.9455	0.7301	0.8992
	Out-of-sample	0.6255	0.5401	0.6567	0.7625	0.6297	0.7045

Note: The best model is underlined

Considering the results for the estimation of recovery rates in the write-off model (Table 6) and partial recoveries model (Table 7), we observe that the machine learning models produce higher accuracy compared to OLS and RT, and in particular, the random forest confirms again its superiority among models.

Table 6. Model performance for RR of write-off model in decomposed approach.

Metric Model	OLS	RT	KNN	RF	MARS	XGB
R^2 (%)						
In-sample	25.98	18.30	35.97	81.16	31.34	41.54
Out-of-sample	17.70	14.80	22.27	37.21	24.34	31.87
RMSE						
In-sample	0.2017	0.3104	0.2717	0.1703	0.2806	0.2611
Out-of-sample	0.3078	0.3133	0.3040	0.2731	0.2966	0.2822
MAE						
In-sample	0.2448	0.2667	0.2219	0.1394	0.2310	0.2143
Out-of-sample	0.2563	0.2669	0.2485	0.2230	0.2433	0.2290
ρ						
In-sample	0.5097	0.4278	0.5997	0.9009	0.5598	0.6445
Out-of-sample	0.4208	0.3847	0.4720	0.6101	0.4933	0.5645

Note: The best model is underlined

Table 7. Model performance for RR of partial recoveries model in decomposed approach

Metric Model	OLS	RT	KNN	RF	MARS	XGB
R^2 (%)						
In-sample	30.34	28.42	38.73	84.97	34.26	64.41
Out-of-sample	25.26	23.35	35.41	42.08	34.01	38.09
RMSE						
In-sample	0.2451	0.2485	0.2287	0.1328	0.2304	0.1787
Out-of-sample	0.2495	0.2528	0.2325	0.2195	0.2408	0.2277
MAE						
In-sample	0.1646	0.1565	0.1377	0.0776	0.1414	0.1121
Out-of-sample	0.1662	0.1627	0.1411	0.1177	0.1478	0.1398
ρ						
In-sample	0.5508	0.5331	0.6223	0.9218	0.5853	0.8026
Out-of-sample	0.5026	0.4833	0.5950	0.6917	0.5830	0.6172

Note: The best model is underlined

The results for all the models confirmed the random forest to be the best model in estimating all the components of the recovery rate decomposed approach.

As a final result of our work, we use the random forest to assess the predictive performance of the whole decomposition approach recovery rate estimation model. Therefore, we assess the predictive performance of the decomposed approach using the random forest for each component and compare these results with a combination of logistics regression and OLS as well as the decision and regression trees used by Starosta (2021). Table 8 demonstrates the results for the different modeling methods for the final model.

Table 8. Model performance for the final recovery rate model

	Metric		
Model	RMSE	MAE	ρ
In-sample			
LR+OLS	0.3192	0.2127	0.2372
DT+RT	0.3235	0.2197	0.2096
RFD+RFR	0.2725	0.1793	0.5479
Out-of-sample			
LR+OLS	0.3738	0.2835	0.0917
DT+RT	0.3259	0.2209	0.2006
RFD+RFR	0.3053	0.2072	0.3611

We estimated three models: i) a decomposed model with logistic regression for classification and OLS for regression, ii) a decomposed model with a decision tree for classification and a regression tree for regression, and iii) a decomposed model where the random forest is used for both classification and regression.

Figure 5. Graphical Comparison

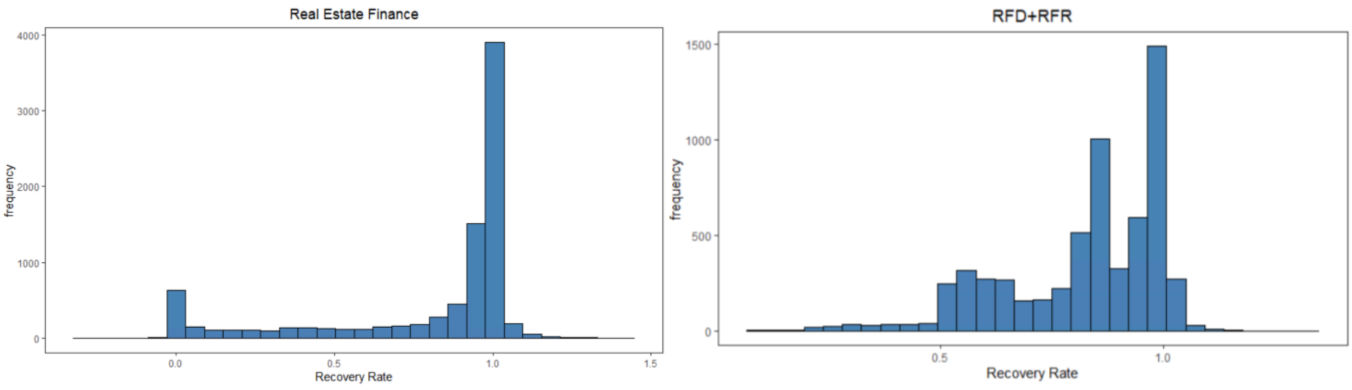


Figure 5 shows a graphical comparison between the distribution of the recovery rate in the original dataset and the one that random forest predicts. In terms of all the performance measures, we find the random forest to confirm again its superiority in modeling the final ultimate recovery rate. Therefore, this result confirms that using the random forest in the multiple-step modeling of the recovery rate proposed by Starosta (2021) could improve the whole recovery rate estimation performance.

8 Conclusion and further research

In this study, we integrated cures, partial recoveries, and write-offs in one equation to estimate the ultimate recovery rate following Starosta (2021) and applied different machine learning algorithms to predict each component of this decomposed approach. The author finds his proposed model to be effective in modeling consumer risk and moreover reveals important implications for risk management and in particular the collection department.

We used a unique database of defaulted real estate-backed loans in European countries provided by GCD. Since the investigation of alternative models is fundamental for lenders for a better prediction of losses, the aim of the study was to present improved forecasting performances of different classification algorithms in predicting the probability of cures, and probability of write-offs as well as several regression models for the recovery rate of cures, partial recoveries and write-offs. In addition to logistic regression, OLS and CART models applied by Starosta (2021), we included the KNN, the random forest, the MARS and Extreme Gradient Boosting to investigate their predictive performance in comparison to the other traditional statistical techniques.

We find that these algorithms outperform all the models used by Starosta (2021) in his study and random forest, in particular, is associated with the best performance for both classification and regression models in terms of in-sample and out-of-sample performance. To conclude with the final result of our work, which was the estimation of the recovery rate using a decomposition approach, we estimated three final models: i) a decomposed model with logistic regression for classification and OLS for regression, ii) a decomposed model with decision tree for classification and a regression tree for regression, and iii) a decomposed model where the random forest is used for both classification and regression, which revealed the best overall performance in our study. The first two combinations were chosen for comparison purposes to Starosta’s (2021) results. We find using the random forest to estimate and predict each component of the equation, leads to higher performance results compared to the other traditional statistical models for the final recovery rate, and hence the LGD.

LGD decomposition proposed by Starosta (2021) presents an innovative improvement as a multi-step modeling approach of LGD estimation and exploring other alternative models in the future that might offer higher predictive performance remain fundamental for lenders.

References

Baker, Scott R., Bloom, Nick and Davis, Stephen J., Economic Policy Uncertainty Index for Europe [EUEPUINDXM], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/EUEPUINDXM, December 1, 2021.

Basel II, C. on B.S., 2004. International convergence of capital measurement and capital standards: a revised framework. BIS report.

Basel III, 2010. Group of governors and heads of Supervision announces higher global minimum capital standards. BIS report.

Bastos, J.A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance* 34 (10): 2510–2517

Bastos, J. A. (2014). Ensemble Predictions of Recovery Rates. *Journal of Financial Services Research*, 46(2), 177–193.

Bellotti, T.,& Crook. J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society* 60 (12): 1699–1707.

Bellotti, T., & Crook, J. (2010). Loss Given Default models for UK retail credit cards, CRC Working Paper, 09/1.

Bellotti, T., & Crook J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. International Journal of Forecasting 28 (1): 171–182.

Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. D. (2019). Forecasting Recovery Rates on Non-Performing Loans with Machine Learning. SSRN Electronic Journal, August.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Caselli, S., S. Gatti, and F. Querci. (2008). The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Services*

Research 34 (1): 1–34

Chen T., He T., Benesty M., Khotilovich V. xgboost: Extreme Gradient Boosting R Package Version: 1.2.0.1 September 2, 2020. URL https:https://cran.rproject.org/web/packages/xgboost/xgboost.pdf

Dermine, J., and C.N. De Carvalho. (2006). Bank loan losses-given-default: A case study. *Journal of Banking & Finance* 30 (4): 1219–1243.

European Banking Authority (2018). Consultation paper: On guidelines for the estimation of LGD appropriate for an economic downturn ('downturn LGD estimation'). *European Banking Authority* (EBA/CP/2018/08)

Eurostat, Harmonized Index of Consumer Prices: All Items for Euro area (19 countries) [CP0000EZ19M086NEST], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CP0000EZ19M086NEST, November 29, 2021.

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

Grunert, J., and Weber M. (2009). Recovery rates of commercial lending: Empirical evidence for german companies. *Journal of Banking & Finance* 33 (3): 505–513.

Hartmann-Wendels, Thomas Miller, P., & Töws, E. (2014). Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking and Finance*, 40(1), 364–375.

Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, Articles, 28(5), 1–26.

Kuhn, M. (2018). Caret: Classification and regression training. R package version 6.0-80, URL https://CRAN.R-project.org/package=caret.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York,Springer

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of*

Forecasting, *28*(1), 161–170.

Matuszyk, A., C. Mues, and L.C. Thomas. 2010. Modelling LGD for unsecured personal loans: Decision tree approach. Journal of the Operational Research Society 61 (3): 393–398.

Papoušková, M., & Hajek, P. (2020). Modelling Loss Given Default in Peer-to-Peer Lending Using Random Forests (Czarnowski). Smart Innovation, Systems and Technologies, vol 142. Springer, Singapore.

OECD (2021), Housing prices (indicator). doi: 10.1787/63008438-en (Accessed on 28 November 2021)

OECD (2021), Quarterly GDP (indicator). doi: 10.1787/b86d1fc8-en (Accessed on 25 November 2021)

OECD, Harmonized Unemployment Rate: Total: All Persons for the European Union [LRHUTTTTEUM156S], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/LRHUTTTTEUM156S, November 29, 2021.

Qi, M., and Yang, X. (2009). Loss given default of high loan-to-value residential mortgages, *Journal of Banking & Finance* 33 (5): 788-799.

Qi, M., and X. Zhao. (2011). Comparison of modeling methods for loss given default. *Journal of Banking & Financ*e 35 (11): 2842–2855.

Starosta, W. (2021). Loss given default decomposition using mixture distributions of in-default events. *Eur. J. Oper. Res., 292*, 1187-1199.

Tanoue, Y., A. Kawada, and S. Yamashita. 2017. Forecasting loss given default of bank loans with multistage model. International Journal of Forecasting 33 (2): 513–522.

Tanoue, Y., Yamashita, S. & Nagahata, H. 2020.Comparison study of two-step LGD estimation model with probability machines. Risk Manag 22, 155–177 (2020).

Thomas, L., Mues, C., & Matuszyk, A. (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of the Operational Research Society*, 61(3). https://doi.org/10.1057/jors.2009.67.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. *BMC Bioinformatics*, **12**, p. 77.

Yao, X., J. Crook, and G. Andreeva. 2015. Support vector regression for loss given default modelling. European Journal of Operational Research 240 (2): 528–538.

Zhang, J., and L.C. Thomas. 2012. Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. International Journal of Forecasting 28 (1): 204–215.

Appendix

Figure 6. ROC Curves for the out-of-sample for the probability of cure model

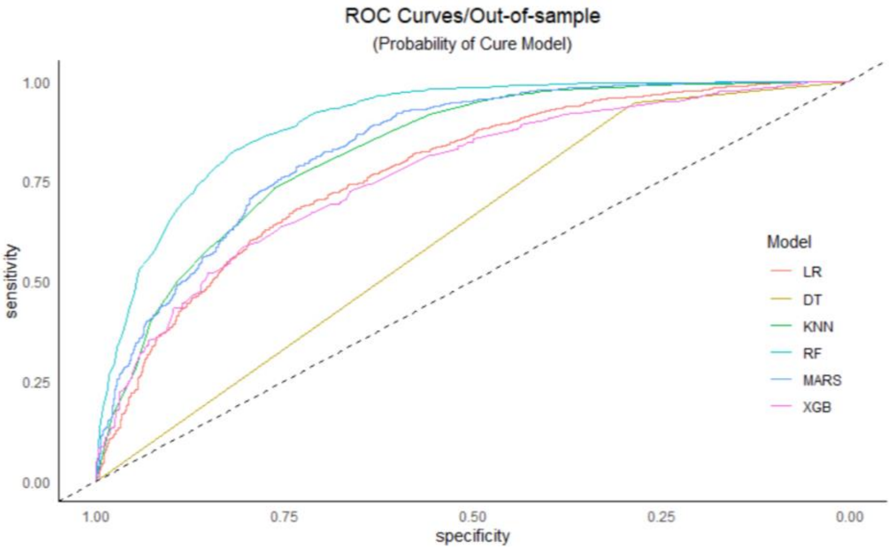
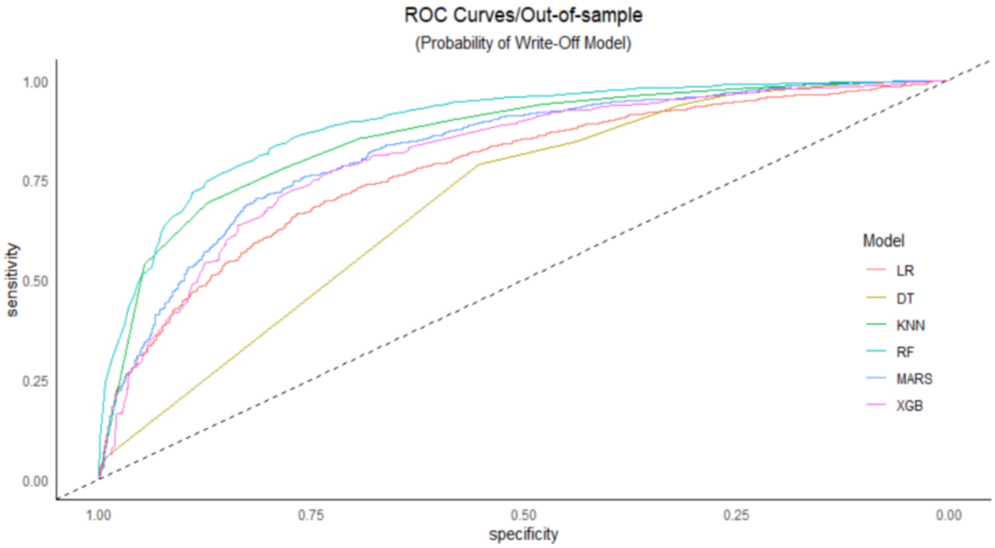


Figure 7. ROC Curves for the out-of-sample for the probability of write-off model


ROC Curves/Out-of-sample
(Probability of Write-Off Model)
sensitivity
specificity
Model
LR
DT
KNN
RF
MARS
XGB