Review of Managerial Science
https://doi.org/10.1007/s11846-023-00635-y

ORIGINAL PAPER

Machine learning due diligence evaluation to increase NPLs profitability transactions on secondary market

Maria Carannante[1] · Valeria D'Amato[1] · Paola Fersini[2] · Salvatore Forte[3] · Giuseppe Melisi[4]

Received: 25 May 2022 / Accepted: 27 February 2023

Abstract

In this paper, we contribute to the topic of the non-performing loans (NPLs) business profitability on the secondary market by developing machine learning-based due diligence. In particular, a loan became non-performing when the borrower is unlikely to pay, and we use the ability of the ML algorithms to model complex relationships between predictors and outcome variables, we set up an *ad hoc* dependent random forest regressor algorithm for projecting the recovery rate of a portfolio of the secured NPLs. Indeed the profitability of the transactions under consideration depends on forecast models of the amount of net repayments expected from receivables and related collection times. Finally, the evaluation approach we provide helps to reduce the ”lemon discount” by pricing the risky component of informational asymmetry between better-informed banks and potential investors in particular for higher quality, collateralised NPLs.

Keywords NPLs · Machine learning · Due diligence · OVERALS

Mathematics Subject Classification 91G15 · 91G70

1 Introduction

According to Banca Ifis report 2021/2022, “even though in 2020 Italy still has a Non-Performing Exposures (NPE) ratio above the EU average, we expect European NPE stock to increase by 60 billion euros in 2022–2023, worse than that estimated for the Italian financial system”. From 2017 to 2020, over 50 billion euros were

Maria Carannante, Valeria D'Amato, Paola Fersini, Salvatore Forte, and Giuseppe Melisi have contributed equally to this work.

✉ Maria Carannante
mcarannante@unisa.it

Extended author information available on the last page of the article

Published online: 15 March 2023

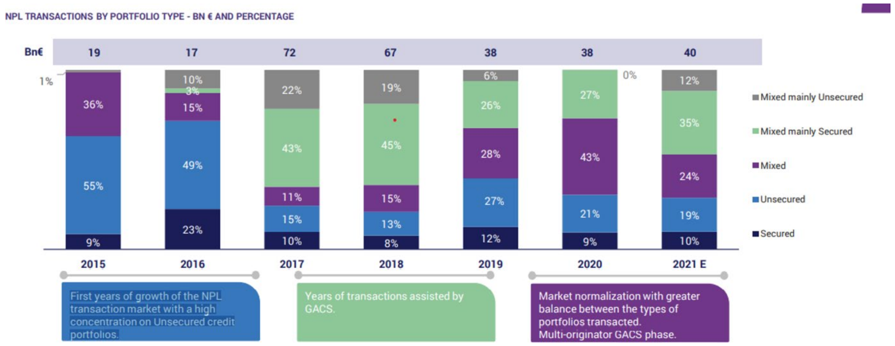

Fig. 1 Source: Ifis Bank NPLs Market Database - News and press releases - Banca Ifis internal analysis 2021

invested in the Non-Performing Loans (NPLs) market to buy approximately 214 billion euros of NPLs portfolios.

Nevertheless, the intense transactional activity on the secondary market - the incidence of 32% in 2021 (IFIS Banca (2022)) - determined the dynamism of the NPLs market. For instance, based on the annual report (Osservatorio Nazionale NPE Market (2020)) by the National Observatory NPE Market of Credit Village (Italy), the Italian secondary market recorded a significant boom as regards 2020: 284 with 12 billion in euros of Gross Book Value (GBV). The main highlight of the survey relies on advanced maturity degree, after a chaotic triennium from 2017 to 2019, which has been relevant to the contribution of foreign investors in the divestiture of whole portfolios.

The market shows deals divided between secured and unsecured loans, with a high incidence of corporate customers, with a high component of an unsecured portfolio. The market shows a sort of normalization with a greater balance between the types of portfolios transacted. On the contrary, the first years of growth of the NPLs transaction market were characterised by a high concentration on Unsecured credit portfolios (Fig. 1).

The transfer of the NPLs to the secondary market can represent a de-risking strategy for the banks operating in the primary market since it relaxes the burden on the NPLs management. Indeed it could help banks offload NPLs from their balance sheets and distribute the risk, free up bank resources, and strengthen bank stability. In other words, the transactions on the secondary market appear an attractive de-risking strategy, due to the high flexibility characterising the securitization structure. In particular, the new legal framework, the EU Directive (EU) 2021/2167 on credit servicers and credit purchasers (also known as the NPLs Directive), could promote an increase in business for specialised credit services. The new rules Directive took effect on 28 December 2021 with the deadline for implementation in all member states being 29 December 2023. They work to expedite and regulate the development of a secondary NPLs market in Europe through disclosure requirements towards the purchasers, notification requirements to the borrowers, and reporting rules to the regulators. More competition and lower costs in transactions are expected in the NPLs marketplace.

The NPLs transactions on the secondary market by portfolio type confirm the trend of divestment of unsecured shares which require specialised services. In light of these considerations, in this paper, we focus on secured NPLs transactions. The profitability of the secured segment is influenced by the liquidation value of the property of the defaulted mortgages. The idea underlying our research is to accurately estimate the profitability of transactions by fair due diligence. We propose to improve the due diligence process by developing an artificial intelligence algorithm. We set up a machine learning due diligence framework to accurately evaluate the expected recovery rates of defaulted mortgage loans.

One of the most important factors governing the price of NPLs portfolios is the recovery rate - that is, the percentage of exposure that can be recovered from each borrower through the debt collection process. To the best of our knowledge, there has been no contribution to explainable machine-learning decisions for NPLs recovery rate models. The only research on the topic focused on an empirical comparison of different classes, i.e linear, nonlinear, and rule-based algorithms for identifying the model structure best suited to the recovery rate problem (Bellotti et al. 2021). Nevertheless, most importantly the aim of the authors consists in designing “a new set of behavioural predictors based on data regarding the recovery procedure promoted by the bank selling the NPLs”. The case of the retail loan recovery rates is examined by appropriate predictive models for macroeconomic factors as in Nazemi and Fabozzi (2018); Nazemi et al. (2018).

Other interesting research investigates how to apply linear regression, beta regression, inflated beta regression, and a beta mixture model combined with logistic regression to model the recovery rate of non-performing loans (Ye and Bellotti 2019).

Unlikely the balance sheet evaluations of NPLs on the primary market, the pricing on the secondary market faces the challenges of market frictions and asymmetric information arising as buyers know less about asset quality than sellers. ”Buyers would therefore fear that assets they are bidding for are of low quality, and bid at a correspondingly low price. The sellers, being able to distinguish between low and high-quality assets, trade only in the former type - the lemons - whereas the market for the remaining assets fails. Additionally, it may be the case that sellers of NPLs may not have perfect information concerning their own assets. The resultant problems associated with informational asymmetry remain, however, as buyers cannot know whether sellers are revealing all available information.” (European Central Bank (2016)). In this context, we provide a tailor-made pricing approach on the secondary market which quotes the risky component of asymmetric information.

The layout of the paper is the following. Section 2 provides practical considerations on the new legal framework, the so-called European Union directive on non-performing loans. Section 3 proposes an *ad hoc* Dependent Random Forest Regressor algorithm i.e. dependent Forest based on Non-Linear Canonical Correlation (DF-NLCC) based on the specialized splitting rule. Section 4 proposes a pricing approach for NPLs portfolio on the secondary market. Section 5 illustrates the main outcomes of the empirical applications. Section 6 concludes.

2 New legal framework or European Union directive on non-performing loans

The NPLs Directive (EU Directive 2021/2167) regulates the activities of the key parties in the NPLs secondary market, in particular credit servicers, credit purchasers and borrowers. It introduces a new set of rules to increase the attractiveness and competitiveness of the secondary market, driving down credit servicing costs, and subsequently determining the lower cost of entry for potential credit purchasers, potentially leading to increased demand and higher NPLs sale prices for the credit purchasers.

The NPLs Directive properly acknowledges the existence of a European secondary market for loans and creates a devoted framework through the authorisation and supervision of the credit servicers.

New requirements are imposed on credit servicers in particular when interacting with borrowers.

On one hand, they have to put in place specific agreements with credit purchasers when outsourcing their credit servicing activities to the credit service providers. On the other hand, selling banks will be subject to providing all necessary information to credit purchasers to enable them to assess the value of the loan assets offered for sale and the likelihood of the recovery. In general, the main objective of the regulation is the increase of the disclosure in requirements towards the purchaser, notification requirements to the borrower and reporting to regulators.

Member states have until 29 December 2023 to implement the Directive. The implementation phase is a transition period, which leaves time for the market players to assess the implications and opportunities it offers. For instance, the different definitions in the Directive among credit purchasers, credit servicers, and credit servicing activities may seem a bit confusing at first glance, by involving some uncertainty also in the categorisation of NPLs (and how sales of combined portfolios comprising NPLs and performing loans are to be dealt with). Nevertheless, the new regulation witnesses the interest and the need to harmonise the regulation of Europe’s secondary NPL markets while protecting borrowers’ rights.

3 Dependent forest based on non-linear canonical correlation (DF-NLCC): a specialised splitting rule

In this section, we propose a Dependent Forest algorithm based on Non-Linear Canonical Correlation (DF-NLCC), defined by a specialised splitting rule for projecting the recovery rate of a portfolio of secured NPLs. We develop this technique in order to capture the variables linked to an NPLs portfolio characterised by a complex dependency structure. In particular, considering the case of a secured NPLs portfolio, the recovery rate is a time-dependent variable, since the shorter the recovery time of the credit the higher the recovery rate. In this sense, the recovery rate can be considered a random variable whose parameters are time-varying. Similarly, the covariates that determine a secured NPLs portfolio have relationships with each other and they depend on recovery time. In fact, the time required to recover credit depends on the Region, the legal entity type and the economic sector. We consider these variables as determinants of the recovery rate of an NPLs portfolio. Moreover, since we take into account secured NPLs that depend on real estate collateral, the book value of the credit is dependent both on the value of the underlying real estate and on the waiting time of the collective mandatory action.

Despite widespread use, the public availability of data on NPLs portfolios is very limited. The main reason depends on the confidentiality of the data for credit holders or for stakeholders interested in buying a portfolio. To cope with the lack of data, we propose a simulation analysis that helps to define the parameters of a series of random variables. We set up secured NPLs portfolios whose recovery is achieved through collective mandatory actions: The duration of the legal action is simulated using a two-parameters Gamma random variable, whose shape parameter k and the scale parameter θ are estimated using an iterative procedure that minimizes the difference between the ratio of the first quantile and the mean and the ratio of the third quantile and the mean of the general distribution of duration of the bankruptcy. The choice of the Gamma distribution lies in its interpretation as the waiting time until an event, i.e. the recovery of the credit through collective mandatory action. Starting from the Gamma random variable defined above we estimate a series of distributions of the waiting time for collective mandatory action, distinguishing among geographical areas, legal forms and business sectors. In particular, the scale parameter θ is assumed constant for each variable, since the differences among the expected values for categories are negligible and the parameter k is estimated using the same iterative procedure for the general distribution of the duration of the collective mandatory action. The Recovery Rate (*RR*) is estimated by taking into account the relationship between credit recovery and time so that a series of beta distributions are estimated, and parameters are determined using an iterative procedure for different classes of time. The Beta distribution is suitable for the estimation of *RR* since it is generally used for modelling a certain proportion of a given phenomenon. The Book Value (*BV*) is estimated taking into account the positive correlation with *RR*. Since the *BV* is a continuous variable, the distributions are estimated using a Gaussian distribution, with mean and variance estimated by using an iterative procedure using the data of the distribution of the *RR*. The difference between the real estate value and the *BV* is estimated taking into account the relationship between the two variables. Considering the real estate value as a percentage of the *BV*, the distributions are estimated by means of a Gaussian distribution, with mean and variance being estimated with an iterative procedure on the data of the *BV* distribution.

The random forest methodology can be described as an ensemble of decision trees. The concept underlying a random forest consists of averaging multiple decision trees affected by high variance, in order to develop a better generalization "less susceptible to overfitting" (Raschka and Mirjalili 2017). Several different approaches for creating an ensemble of classifiers combine multiple classifiers into a meta-classifier that obtains better performance than each individual.

The random forest can be defined as a general principle of classifier combination that uses L tree-structured base classifiers $(x, \Xi_k), k = 1, \dots, L$ where Ξ_k represents a family of independent identically distributed random vectors, and x is an input data. It involves that no guarantee that all those trees will cooperate effectively in the same committee. Bernard et al. (2009).

Nevertheless, the underlying data-generating process presents dependence properties, which asymptotically influence the distribution of the statistics of interest. This issue justifies the development of an algorithm framework proposal, which really mimics the dependence properties (or even the process), in order to avoid dependency risk as codified in D'Amato et al. (2013). In the random forest context, the dependency risk consists of misleading evaluations in classification or regression. In our proposal, the dependent forest we design consists of many unsupervised decision trees with a specialised splitting criterion. In particular, the specialised splitting criterion relies on nonlinear relationships of a variable with some variables.

We build the individual trees in the forest with a splitting rule specifically designed to partition the data to maximize the non-linear canonical correlation heterogeneity between nodes arranged as in OVERALS (van der Burg et al. 1994). The canonical correlations represent how much variance of the dependent variables is explained by the dimensions, where the canonical dimensions are latent variables that are analogous to factors obtained in factor analysis, except that canonical variates also maximize the correlation between the two sets of variables. In general, not all the canonical dimensions will be statistically significant. A significant dimension corresponds to a significant canonical correlation and vice versa.

We design the architecture of an unsupervised random forest based on the set of covariates Z to find subgroups of observations with non-linear canonical correlations between mean-centered multivariate datasets X and Y.

The tree-growing process is based on the Classification and Regression Tree (CART) approach (Breiman et al. 1984). The basic idea of tree growing with CART is to select the best split at each parent node among all possible splits to obtain the nodes. Inspired by Alakuş et al. (2021), where the conditional canonical correlations between two sets of variables given subject-related covariates have been estimated, we propose a splitting rule that increases the non-linear canonical correlation heterogeneity as fast as possible, being the goal to find subgroups of subjects with distinct non-linear canonical correlations. Unlike (Alakuş et al. 2021), we set up the random forest architecture in a non-linear environment, being a focus on the non-linear canonical correlation heterogeneity. The proposed splitting criterion is expressed by the following:

$$\sqrt{n_L \cdot n_R} \cdot |nr_L - nr_R| \tag{1}$$

where:

n_L is the size of the left node
n_R is the size of the right node
nr_L is the non-linear canonical correlation estimation of the left node
nr_R is the non-linear canonical correlation estimation of the right node.

To define the non-linear canonical correlation, we consider a measure of similarity with respect a undefined variable x that identifies the concept described by a dataset H. The idea is to find the weight a that makes the weighted sums of variables as similar to x as possible. In this sense, the non-linear correlation is the Sum of Squares (SSQ) of the difference between the undefined variable x and the weighted sum of the dataset variables aH:

$$nr = SSQ(x - aH) \tag{2}$$

The best split among all possible splits is the one that maximizes the formula (1).

4 In-depth pricing approach on the secondary market

In order to evaluate the NPLs portfolio price on secondary market sale, the DF-NLCC allows determining the projected RR. Nevertheless, to address our scope of computing the value of NPLs on the secondary market and the correspondent profitability for potential buyer, once estimated the RR value of a NPLs portfolio, we focus on the relevance of structural inefficiencies and information asymmetries driving a wedge between book values and market value of NLPs (European Central Bank (2016)). Indeed the bank that sells an NPLs portfolio tends to sell the worst loans it has, therefore it is reasonable to consider that the price proposed by the buyer should take this aspect into account. In light of these considerations, in this section we propose a pricing assessment approach adjustments in the secondary market, considering separately both the cost of capital of the booking in the balance sheet of the loans by the buyer and the weight in terms of information asymmetry between the bank and the potential buyer. Basically, the framework evaluation under consideration is built on the incidence of risky components of information asymmetry alike the classical balance sheet estimates.

Therefore, we distinguish between the first investor I_1 as the potential investor that ignores the information asymmetry at the time of transaction and assesses the NPLs portfolio on the basis of the cost of capital only, and the second investor I_2 as the potential investor that considers both the cost of capital and the information asymmetry.

The value of the NPLs portfolio (NPV) in the secondary market is defined as the present values of the expected cash-flows (CF) at a risk-free rate minus the Cost of Capital (CoC), defined as the by a 99.5% Value-at-Risk (VaR) of the predicted RR discounted at cost of capital.

$$NPV = \mathbb{E}[CF, i_{RF}(0,t)] - CoC \tag{3}$$

Where $\mathbb{E}[CF, i_{RF}(0,t)]$ discounted at time $t = 0$ is the expected cash-flow net of the expected recovery rate with a spot risk-free rate structure.

$$CoC = \mathbb{E}[CF, i_{RF}(0,t), VaR_{99.5\%}(RR)] - \mathbb{E}[CF, i_{RF}(0,t)]i_{CoC} \tag{4}$$

Where:

i_{CoC} is the cost of capital rate of the investor in NPLs in the secondary market;

$\mathbb{E}[CF, i_{RF}(0,t)], VaR_{99.5\%}(RR)$ discounted at time $t = 0$ is the cash flow net of recovery rate at probability level at 99.5%.

To take into account the CoC, we quantify the spot rate i_T that allows quantifying the net present value of CoC considering also the information asymmetry (IA):

$$i_{T+IA} = i_{RF+CoC+IA} \tag{5}$$

i_T is estimated by the following equation:

$$\mathbb{E}[CF, i_T(0,t)] = NPV \tag{6}$$

Then, the market value of the NPLs portfolio negotiated in the secondary market will be:

$$\mathbb{E}[CF, i_T(0,t)] - C_{IA} \tag{7}$$

where C_{IA} is the cost of information asymmetry.

So, the i_{T+IA} will be the solution of the following equation:

$$\mathbb{E}[CF, i_{T+IA}(0,t)] = \mathbb{E}[CF, i_T(0,t)] - C_{IA} \tag{8}$$

5 Numerical application

In this section, we implement the DF-NLCC algorithm. Firstly, we generate a simulated dataset of 10,000 NPLs, following the iterative procedure of random variable parameter estimation described in Sect. 4. To simulate a portfolio of NPLs coherent with the characteristics of the Italian market, we refer to a technical report released by Cerved SpA (2020), that contains data about the duration of mandatory actions, divided by Regions, legal forms and business sectors, considering both default and voluntary liquidations. Furthermore, we consider data relating to recovery rate by waiting time of recovery, estimated by Fischetto et al. (2021), based on NPLs with collaterals. In this way, it is possible to estimate not only the recovery time by selecting only the NPLs with collaterals but also the relationship between the value of collateral and recovery rate.

Tables 1, 2, 3, 4 and 5 show the estimated parameters for distributions:

As shown by Table 1 the θ parameter is assumed the same for all the Regions. For this reason, differences in durations are determined for the k parameter only, which is approximately the skewness of the distribution. In this sense, we can consider that the lower the k parameter the higher the positive skewness of the distribution and, since θ is constant, the lower the average waiting time of the mandatory actions. Furthermore, we can observe that the Northern Regions have lower k parameters

Table 1 Duration of the mandatory actions distribution parameters by region

Region	Gamma parameters	
	k	θ
Abruzzo	1.80	2.94
Basilicata	3.20	2.94
Calabria	2.99	2.94
Campania	1.97	2.94
Emilia-Romagna	1.36	2.94
Friuli V.G	1.02	2.94
Lazio	1.53	2.94
Liguria	1.19	2.94
Lombardia	1.53	2.94
Marche	1.97	2.94
Molise	2.21	2.94
Piemonte	1.36	2.94
Puglia	2.04	2.94
Sardegna	2.21	2.94
Sicilia	2.55	2.94
Toscana	1.53	2.94
Trentino A.A	1.02	2.94
Umbria	1.97	2.94
Valle D'Aosta	0.82	2.94
Veneto	1.50	2.94

Table 2 Duration of the mandatory actions distribution parameters by legal form

Legal form	Gamma parameters	
	k	θ
Sole proprietorship	2.26	2.94
Partnership company	2.30	2.94
Limited company	1.51	2.94

Table 3 Duration of the mandatory actions distribution parameters by business sector

Business sector	Gamma parameters	
	k	θ
Agriculture	1.92	2.94
Construction	1.80	2.94
Manufacturing	1.92	2.94
Services	1.56	2.94
Energy and utilities	1.90	2.94

Table 4 *RR* distribution parameters

Business sector	Beta parameters	
	α (%)	β (%)
0-2 years	55,70	44,30
3-5 years	45,08	54,92
Over 5 years	35,68	64,32

Table 5 *BV* distribution parameters

BV	Gaussian parameters	
	Mean	Variance
0-50,000	25,000	5,000
50,000-200,000	125,000	25,000
Over 200,000	350,000	70,000

Table 6 Real estate value distribution parameters

BV	Gaussian parameters	
	Mean (%)	Variance (%)
0–30%	− 30	20
30–70%	− 5	20
70–100%	20	20

than Centre and South Italy, therefore we can assume geographical differences in recovery waiting times, which is consistent with starting data.

Tables 2 and 3 follow the same logic as Table 1.

As Table 2 shows, the Limited company has a greater positive skewness and a lower average waiting time with respect to the Sole proprietorship and the Partnership company, while for the latter the distribution is very similar.

As Table 3 shows, the Services class has a greater positive skewness and a lower average waiting time with respect to the other sectors. On the contrary, Agriculture and Manufacturing show the lower skewness and the greatest average time, as well as Energy and Utilities and Construction with a lower average than the former classes (Agriculture and Manufacturing) and greater than Services.

In Table 4, the beta parameters α and β are represented in the form of the proportion of success and proportion of failure respectively. In this sense, we can observe that the greater the waiting time until the mandatory actions, the lower the proportion of credit recovery.

Table 5 shows that the greater is the value of the credit, the greater is the mean and the variance of the distribution of recovery. We can also observe that the relationship is less than proportional, so the larger the credit size the higher recovery value, but the recovery rate is not necessarily greater.

As shown in Table 6, the real estate value has a negative mean for the lower classes and a positive mean for the higher class.

Table 7 Summary statistics of *RR*

	Training set (%)	Validation set (%)
Mean	44.55	44.00
SD	35.77	35.77
Min	0	0
Max	100	100
Q1	8.35	7.90
Media	39.72	38.99
Q3	80.63	79.50

Simulated data are used to estimate a random forest to predict *RR*. To do this, we extract a training set equal to 80% of the total. Table 7 shows the summary statistics of *RR* for both training and validation sets:

As Table 7 shows, the two datasets are balanced and there are no outliers.

Table 8 describes the features used to performed tree-based algorithms

To compare the results, we perform a Decision Tree (DT), a Random Forest (RF) and an XGBoost (XGB). Table 9 shows the in-sample and out-of-sample accuracy for the three models.

As shown in Table 9, accuracy is higher for RF in in-sample forecasting with respect to the other methods, while XGB shows a greater accuracy in out-of-sample forecasting.

Table 10 compares the ranking of variable importance for DT and RF. Figure 2 shows the variable importance for RF. As criterion, we use the percentage increase in Mean Square Error (*MSE*).

As intuitively expected in secured transactions, in Table 10 and Fig. 2 the real estate value (*REV*) is the most important feature. Duration (*DUR*) and book value (*BV*) are also important variables, the former ranks first for DT and the latter for RF. The other variables have the same importance for both methods. The choice between DT or RF does not affect the variable importance.

Since XGBoost model process quantitative variables only, we create dummies for each category or class of the variables. For this reason, the importance analysis shows the relevance of each category or class. As criterion, we use the percentage increase in Mean Square Error (*MSE*).

As Table 11 and Fig. 3 show, the most important variables are the real estate value (REV), the book value (BV) and duration (DUR) with respectively an improvement of MSE of 0.563, 0.334 and 0.061. The other variables have an improvement of MSE sensitively lower.

The value of the total portfolio of NPLs is obtained by discounting the *BV* of each credit multiplied by the *RR* with a market interest rate, assumed to be 10%. The duration relative to the recovery is assumed equal to that simulated one for the construction of the Data Set, while for the *RR* is used is the present value in the data set that the estimates obtained through the three analyzed models. Results are shown in Table 12.

Table 8 Features of tree-based algorithms

ID	Name	Description	Type	Categories
REG	Region	Territorial area in which the company operates	Categorical	Abruzzo Basilicata Calabria Campania Emilia-Romagna Friuli Venezia Giulia Lazio Liguria Lombardia Marche Molise Piemonte Puglia Sardegna Sicilia Toscana Trentino Umbria Valle D’Aosta Veneto
BS	Business sector	Business sector of the company	Categorical	Agriculture Construction Manufacturing Services Energy and utilities
LF	Legal form	Legal form of the company	Categorical	Sole proprietor Partnership company Limited company
DUR	Duration	Waiting time to mandatory actions	Continuous	
BV	Book value	Credit value in the balance sheet	Continuous	
REV	Real estate value	Market value of the underlying property	Continuous	
RR	Recovery rate	Recovery rate of book value	Continuous	

Table 9 Models error estimators

	DT		RF		XGB	
	Train	Valid	Train	Valid	Train	Valid
MSE	0.062	0.059	0.060	0.033	0.057	0.041
RMSE	0.250	0.243	0.245	0.183	0.239	0.202
MAE	0.196	0.191	0.201	0.149	0.188	0.158

Table 10 Variable importance comparison

Ranking	DT	RF
1	REV	REV
2	BV	DUR
3	DUR	BV
4	REG	REG
5	LF	LF
6	BS	BS

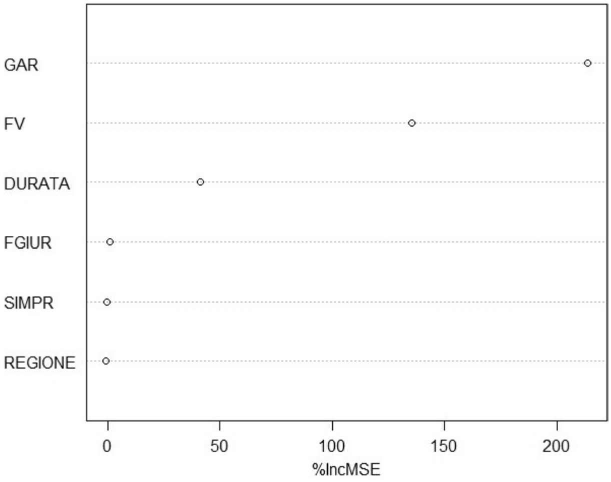

Fig. 2 Variable importance random forest

Table 11 Variable importance for XGBoost

Variable	MSE
REV	0.563
BV	0.344
DUR	0.061
Lombardia (REG)	0.003
Toscana (REG)	0.002
Partnership Company (LF)	0.002
Manufactuing (BS)	0.002
Puglia (REG)	0.002
Limited company (LF)	0.002
Campania (REG)	0.002

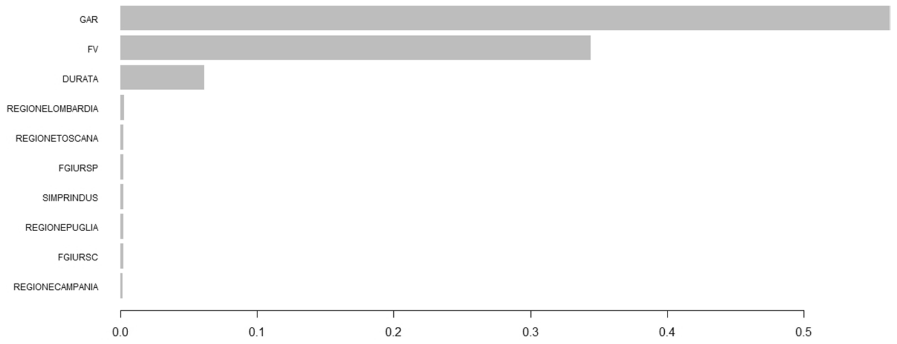

Fig. 3 Variable importance XGBoost

Table 12 Variable importance comparison

RR model	Portfolio value	Portfolio value/ Sum BV (%)
Simulated data set	581,463,952	34.55
DT	577,827,028	34.34
RF	586,799,163	34.87
XGB	581,540,805	34.56
Sum BV	1,682,798,971	

Table 13 Portfolio value according to the discount rate

Modello	4%	6%	8%	10%	12%	14%	16%
Data set	42.34%	39.39%	36.81%	34.55%	32.56%	30.78%	29.19%
DT	42.43%	39.35%	36.68%	34.34%	32.28%	30.45%	28.81%
RF	42.79%	39.78%	37.17%	34.87%	32.84%	31.04%	29.43%
XGB	42.39%	39.42%	36.83%	34.56%	32.55%	30.77%	29.17%

Table 12 shows that both the values of the portfolios and the ratios for the three models analyzed are similar and show a small deviation from the value obtained directly from the simulated data set.

Table 13 shows the sensitivity analysis of the interest rate for the portfolio evaluation:

As shown in Table 13, there are no significant deviations in the portfolio valuation depending on the model as the interest rate changes.

To define the valuation rate, we determine the capital requirement for a risk of loss due to an achieved *RR* lower than expected. In particular, the capital absorption associated with the credit recovery risk was calculated on the basis of the probability distribution of the *RR* by estimating the VaR at 99.5% of the

Table 14 Portfolio value

CoC	10.00%
Risk-free rate	0.98%
Evaluation rate	13.27 %
Portfolio value	528,444,188

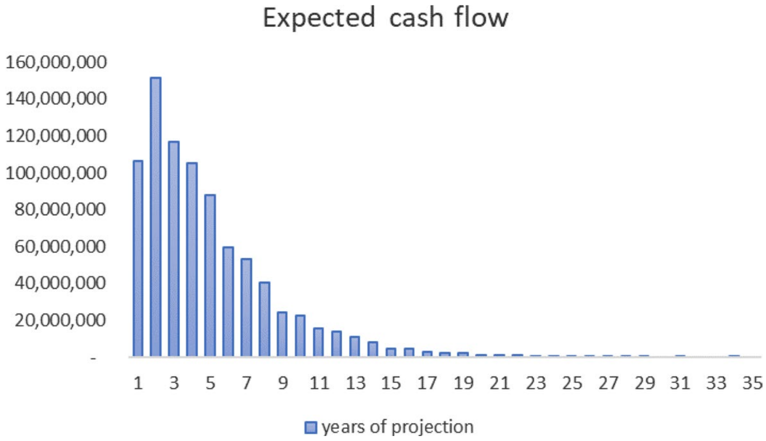

Fig. 4 Cash flow analysis by years of projection

loss. The cost of capital was therefore determined on the basis of the recovery duration assuming a percentage of the cost of capital equal to 10% per year. The determined amount defines the discount rate to be adopted for the valuation of the portfolio, to be in line with the expected remuneration required by an investor who has a risk appetite consistent with the 99.5% level adopted for the VaR.

The determined amount define the discount rate to be adopted for the valuation of the portfolio, to be in line with the expected remuneration required by an investor who has a risk appetite consistent with the 99.5% level adopted for the VaR.

The value of the total NPLs portfolio is obtained by discounting the *NBV* of each credit multiplied by the *RR* with a market cost of capital rate, assumed to be 10%. The duration of the recovery is assumed to be equal to the simulated dataset, and we compare the *RR* obtained in the dataset and in the out-of-sample forecast of the three models.

Table 14 shows the results:

In other words, by discounting cash flows at 13.27% yearly, compared to the undiscounted value, we obtain a difference equal to the total cost of capital which is 309,239,223 million of euros.

Table 15 Bid/ask spread

	Bank	First investor	Second investor
GBV	1,682,798,971	1,682,798,971	1,682,798,971
Undiscounted cash flow	837,683,411	837,683,411	837,683,411
Discount rate	4.00%	13.27%	20.00%
NBV	712,520,207		
NPV		528,444,188	445,241,388
Indirect costs (%)		5%	5%
Discounted indirect costs (Euro)		26,422,209	22,262,069
Market value		502,021,979	422,979,318

To perform the bridge analysis, we perform a projection of cash flow for 35 years. Figure 4 shows the results.

As shown in Fig. 4, the expected cash flow grows in the first years of the transaction, with a maximum in the second year and then decreases exponentially.

5.1 Profittability analysis: from the NPV to the market value

Table 15 shows the bridge analysis which highlights the differences between the Gross Book Value (*GBV*), the Net Book Value (*NBV*), the market value of the first investor and the market value that could form on the secondary market based on the price defined by the second investor:

As shown in Table 15 the difference between bank *NBV* and the market value defined for the first investor is mainly due to the discount rate used and the indirect costs charged by the investor, assuming that the cash flows estimated by the two counterparties are equal. In particular, with regard to the valuation rate, the

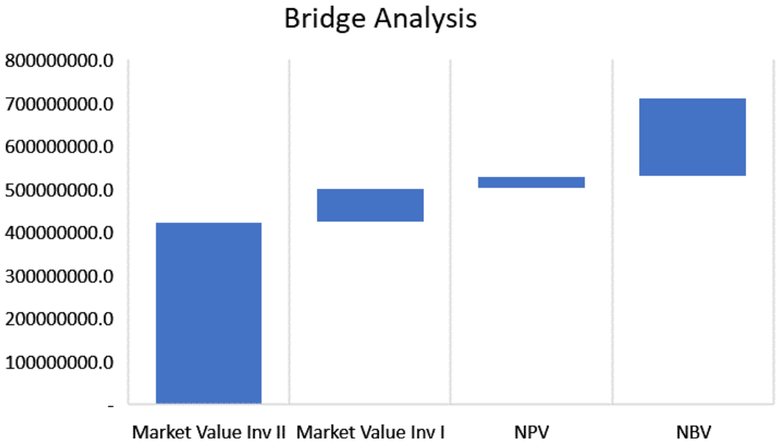

Fig. 5 Bridge analysis

Table 16 Bridge analysis sensitivity

	16%	18%	20%	22%	24%	26%
NBV	712,520,207	712,520,207	712,520,207	712,520,207	712,520,207	712,520,207
NPV	528,444,188	528,444,188	528,444,188	528,444,188	528,444,188	528,444,188
Market value Inv I	502,021,979	502,021,979	502,021,979	502,021,979	502,021,979	502,021,979
Market value Inv II	466,571,150	443,680,306	422,979,318	404,174,938	387,023,261	371,319,891

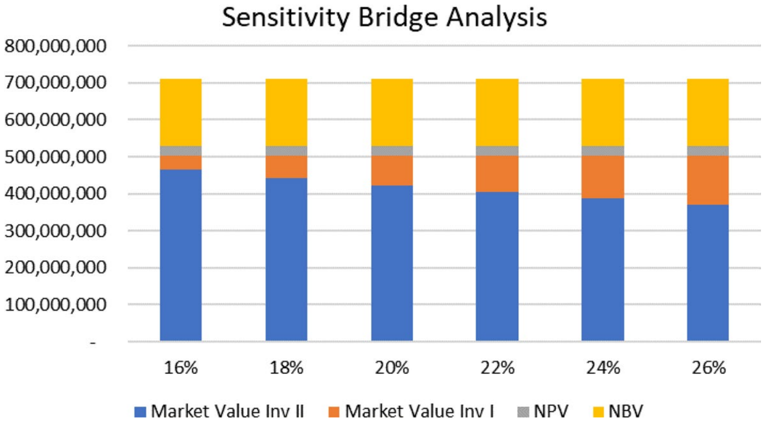

Fig. 6 Bridge analysis sensitivity

bank carries out the valuation according to the original rate of each loan, assumed to be 4%, while the investor according to a rate that must take into account the cost of capital and the specific risk of the transaction, i.e. the uncertainty inherent in the expected cash flows (also taking into account the information asymmetry). This difference is also referred to as the bid/ask spread and can be very large precisely due to the amortized cost method used by banks for the valuation of receivables in the financial statements according to the IAS/IFRS international accounting standards.

The difference between the market value of the first investor and the second investor on the secondary market mainly depends on any information asymmetry and the risk inherent in the lemon market, i.e. that the seller knowing the portfolio well could sell only the worst credits, namely lemons. This factor justifies the bid/ask spread relating to the secondary market and therefore the difference between the valuation rates adopted by the two counterparties.

Values of the bridge analysis that reconciles the value defined on the secondary market with the Net Book Value (NBV), determined by the bank are shown in Fig. 5 as follows:

- the second bar measures the difference between the market value defined by the first and second investor and it depends on the difference in the valuation rates, resulting from the information asymmetry.

- the third bar shows the difference in the valuation of the first investor with respect to the *NPV* determined on the basis of the rate of the second investor. Therefore, this component derives mainly from the indirect costs considered in the investor's assessment.
- the fourth bar reconciles the *NPV* defined by the investor to the *NBV* calculated by the bank and depends on the different valuation rates used as described above.

Figure 5 shows that with each transaction the value of the portfolio is reduced, according to the discount rate applied to the valuation in the transaction.

Finally, Table 16 and Fig. 6 show the sensitivity analysis of the bridge analysis, changing the discount rate:

As shown by Table 16 and Fig. 6, the discount rate determines the value of the portfolio, the greater is the rate, the lower is the value as the number of transactions increases.

Considering the market prices in the analysis, we assume a current discount rate between 15 and 30%. It follows that:

- the correct market price, assuming a CoC rate of 10% and a risk measure to evaluate this risk capital equal to the VaR (99.5%), is equal to the market value of the first investor's portfolio that is 502 millions of euros;
- the real profitability of the second investor is equal to the difference between the market value price of the first investor and the market value price of the second investor, which changes on the basis of the hypothesized discount rate. Considering a range from 16–26% (within 15–30% found on the market), that is always higher than 13.27%, we obtain an expected present value of extra profits cost of capital which varies from 36 million to 131 million euros or in terms of:

$$\frac{\text{current expected profit}}{NBV} = 5\% \tag{9}$$

$$\text{RORAC} = \frac{\text{current expected profit}}{\text{RC}} = \textit{from } 4.3\% \textit{ to } 15.7\% \tag{10}$$

where *RC* is the risk of capital, estimated in 834 millions of euros.

6 Conclusions

The increasing importance of NPL in the secondary market led the European Institution to regulate transactions, revealing specific features in comparison with the NPLs' primary market. The new regulatory framework could address effective due diligence to support the investment choices in NPLs portfolios. In the context of secured NPLs, due diligence can sensitively affect the business profitability from the sale on the secondary market to the dispute resolution time for recovery action.

In this paper, we propose a Dependent Forest algorithm based on Non-Linear Canonical Correlation (DF-NLCC), defined by a specialised splitting rule for projecting the recovery rate of a portfolio of secured NPLs. We develop this technique in order to capture the variables linked to an NPLs portfolio characterised by a complex dependency structure. In particular, considering the case of a secured NPLs portfolio, the recovery rate is a time-dependent variable, since the shorter the recovery time of the credit the higher the recovery rate. The technique is able to capture the variables linked to an NPLs portfolio characterised by a complex dependency structure.

Once estimated the recovery rates by artificial intelligence algorithms we provide a tailor-made approach by pricing the informational asymmetry risky component that usually drives a wedge between book values and market values of NLPs.

Indeed asymmetric information arises from banks' cherry-picking of assets for sale. Banks may be incentivised to retain the best assets, along with the best client relationships. Prices offered by investors have to account for the adverse selection of the assets up for sale. Further research will be addressed to highlight the differences between secured and unsecured NPLs transactions on the secondary market.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11846-023-00635-y.

Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

Data availibility The dataset relating to duration of mandatory actions, divided by Regions, legal forms and business sectors is available in the report ”Osservatorio su fallimenti, procedure e chiusure di impresa” provided by Cerved Spa at https://research.cerved.com/rapporti/. The dataset relating to recovery rates is available in the report ”I tassi di recupero delle sofferenze nel 2020” provided by Banca d'Italia at https://www.bancaditalia.it/pubblicazioni/note-stabilita/2021-0027/index.html.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alakuş C, Larocque D, Jacquemont S, Barlaam F, Martin CO, Agbogba K, Lippé S, Labbe A (2021) Conditional canonical correlation estimation based on covariates with random forests. Bioinformatics 37(17):2714–2721. https://doi.org/10.1093/bioinformatics/btab158

Bellotti A, Brigo D, Gambetti P, Vrins F (2021) Forecasting recovery rates on non-performing loans with machine learning. Int J Forecast 37(1):428–444. https://doi.org/10.1016/j.ijforecast.2020.06.009

Bernard S, Heutte L, Adam S (2009) On the selection of decision trees in random forests. In: 2009 International joint conference on neural networks, vol. 10802866, pp. 302–307. IEEE. https://doi.org/10.1109/IJCNN.2009.5178693

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression. Trees. https://doi.org/10.1201/9781315139470

Cerved: (2020) Osservatorio sui fallimenti, procedure e chiusure di imprese. numero 41. Technical report, Cerved SpA

D'Amato V, Haberman S, Piscopo G, Russolillo M (2013) Computational framework for longevity risk management. Comput Manag Sci 11(1–2):111–137. https://doi.org/10.1007/s10287-013-0178-2

Directive (EU) 2021/2167 of the European Parliament and of the Council of 24 November 2021 on credit servicers and credit purchasers and amending Directives 2008/48/EC and 2014/17/EU

European Central Bank (2016) Financial stability review - november 2016. Technical report, European Central Bank

Fischetto AL, Guida I, Rendina A, Santini G, Scotto di Carlo M (2021) I tassi di recupero delle sofferenze nel 2020. note di stabilità finanziaria e vigilanza. numero 27. Technical report, Banca d'Italia

IFIS Banca (2021) Npl transaction market and servicing industry. full year 2020 and forecast 2021-2022. Technical report, IFIS Banca. https://www.bancaifis.it/app/uploads/2021/01/MW_NPL_January21_ENG.pdf

IFIS Banca (2022) Mercato delle transazioni npl nell'industria del servicing. consuntivo 2021 e forecast 2022-2024. Technical report, IFIS Banca. https://www.bancaifis.it/app/uploads/2022/02/MW_NPL_Feb22_ITA.pdf

Nazemi A, Fabozzi FJ (2018) Macroeconomic variable selection for creditor recovery rates. J Bank Fin 89:14–25. https://doi.org/10.1016/j.jbankfin.2018.01.006

Nazemi A, Heidenreich K, Fabozzi FJ (2018) Improving corporate bond recovery rate prediction using multi-factor support vector regressions. Eur J Operat Res 271(2):664–675. https://doi.org/10.1016/j.ejor.2018.05.024

Osservatorio Nazionale NPE Market (2020) Credit Village. https://www.creditvillage.news/cvstudi_ricerche/

Raschka S, Mirjalili V (2017) Python machine learning - Second Edition: machine learning and deep learning with python, Scikit-learn, and TensorFlow, Second Edition. https://www.ebook.de/de/product/30113031/sebastian_raschka_vahid_mirjalili_python_machine_learning_second_edition.html

van der Burg E, de Leeuw J, Dijksterhuis G (1994) OVERALS. Comput Stat Data Anal 18(1):141–163. https://doi.org/10.1016/0167-9473(94)90136-8

Ye H, Bellotti A (2019) Modelling recovery rates for non-performing loans. Risks 7(1):19. https://doi.org/10.3390/risks7010019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Maria Carannante[1] · **Valeria D'Amato**[1] · **Paola Fersini**[2] · **Salvatore Forte**[3] · **Giuseppe Melisi**[4]

Valeria D'Amato
vdamato@unisa.it

Paola Fersini
pfersini@luiss.it

Salvatore Forte
s.forte@unifortunato.eu

Giuseppe Melisi
gimelisi@unisannio.it

[1] Department of Pharmacy, University of Salerno, Via Giovanni Paolo II 132, Fisciano 84084, Salerno, Italy

[2] Department of Business and Management, Luiss 'Guido Carli' University, Viale Romania, 32, Rome 00197, Italy

[3] Faculty of Law, Università Telematica Giustino Fortunato, Via Raffaele Delcogliano, Benevento 82100, Italy

[4] Department of Business and Management, University of Sannio, Via delle Puglie, 82, Benevento 82100, Italy