The Faculty of Economics and Sociology

University of Lodz

Modeling the Loss Given Default of Retail Contracts

Wojciech Starosta

A doctoral thesis submitted for the degree of Doctor in Economics

Supervisors:

Paweł Baranowski, Ph.D., Associate Professor

Mariusz Górajski, Ph.D.

Łódź, 2021

Contents

1 Introduction

Risk management is the fundamental concept in any modern financial institution that wants to be perceived as having a stable capital base and generating economically justified decisions. It has implications on different aspects of bank operations and organizational structure. Implementing the Advanced Internal Rating Based Approach (AIRB) for capital allocation or International Financial Reporting Standard number 9 (IFRS 9) for expected credit loss calculation regimes serves to enhance risk management practices and raises competitiveness on the market.[1] The scope of risk parameters use is not limited to assessing current risk vulnerability but also positioning the institution in the forthcoming economic environment. The importance of selection and validation of the methods to measure various types of risk seems vital not only during a downturn but in any phase of the economic cycle. This makes all units vulnerable to risk exposure, interested in having as precise parameters as possible.

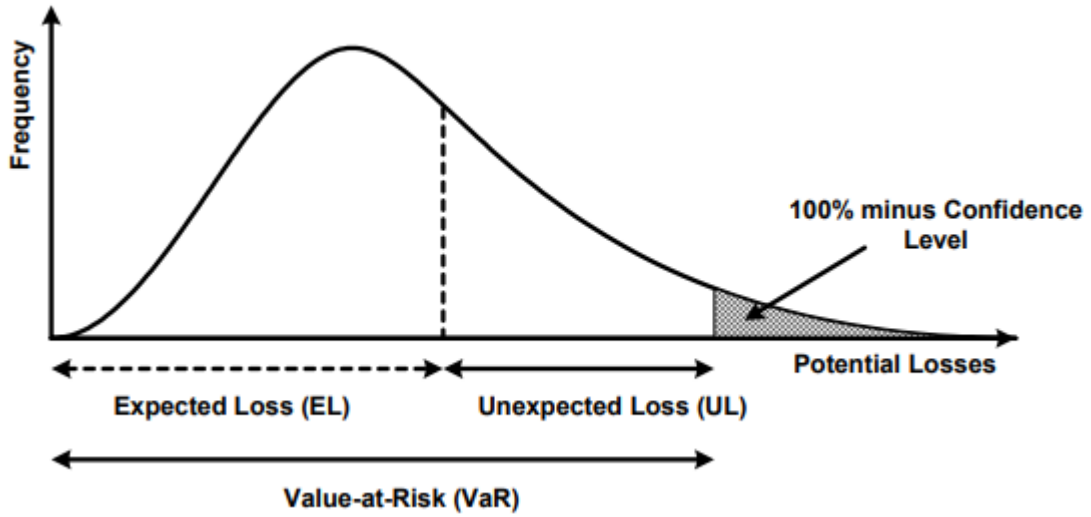
Not surprisingly, one of the major concerns of the financial institution risk framework is to assess the risk connected to credit activity correctly. The Basel II Capital Accord prescribes the minimum amount of regulatory capital an institution must hold to be resistant to unexpected losses and be in line with its risk appetite. Estimation of expected and unexpected losses associated with each exposure is possible within the Asymptotic Single Risk Factor Model (ASRF) framework (Basel Committee on Banking Supervision, 2005a). Under certain conditions Vasicek (2002) showed that, Merton's (1974) single asset model can naturally be extended to a specific ASRF credit portfolio model. The AIRB standard, which adopted these proposals, imposes estimation of the three risk parameters, which are PD probability of default (PD), loss given default (LGD) and exposure at default (EAD). The first one is defined as the likelihood that a particular client will not repay his debt and fall into default in a determined extent of time. The default event is indicated by the default indicator variable that equals one if the uncertain default occurs, and zero otherwise (Hibbeln, 2010). Loss given default stands for economic loss, expressed as a percentage of exposure, which will not be recovered if the loan goes into default. EAD is the amount expressed in a particular currency, that obligor will have to repay in case of default. It consists of the current outstanding, which was already drawn, and a part of the commitment, which can be drawn and introduces uncertainty, leading to estimate the Credit Conversion Factor (CCF) (cf. Gürtler et al., 2018, or Tong et al., 2016). Multiplication of these three elements results in *Expected Loss (EL)*, which is a part of the loan pricing and takes a substantial role in the accounting for financial instruments (specifically impairment of financial assets) as IFRS 9[2] replaced the IAS 39[3] in 2018. What is more, PD, downturn[4] LGD (dLGD) , downturn EAD (dEAD) and correlation parameter among loans are used as a part of the first pillar in *Unexpected Losses (UL)* calculation to obtain risk-adjusted capital requirements under Basel II Accord (see Figure 1).

There are many areas where competitive advantage can be gained, when underlying risk of exposure can be properly reflected by risk parameters. Firstly, pricing, which reflects true risk of a client, can be used to select an acceptance level that correctly represent institution risk appetite. It leads to flexibility in the credit policy decision-making process, as the riskiness of default event and conditional loss can be managed simultaneously. What is more, even after default, collection (debt recovery) strategies can be set according to LGD estimates, where the soft collection can be assigned to cases with a low

[1] See the gap analysis in Prorokowski (2018).
[2] International Financial Reporting Standard promulgated by International Accounting Standards Board (IASB).
[3] International Accounting Standard 39: Financial Instruments: Recognition and Measurement.
[4] Final guidelines regarding downturn LGD estimation were presented by EBA in the document "Guidelines for the estimation of LGD appropriate for an economic downturn ('Downturn LGD estimation')" (2019). Following the document we define downturn LGD as LGD estimates appropriate for an economic downturn period.

value of LGD parameter and more decisive actions otherwise. Secondly, capital requirements calculated via the advanced approach are seen as more sensitive to the underlying risk of assets, as internal models can recognize detailed risk profile absorbed by institution. Pursuing less risky assets leads to regulatory capital reduction, which can be used for other business initiatives. Last but not least, the use of in-house PD, LGD and EAD make it possible to get deep insight into the impairment process, which lead to preparing stable and forward-looking forecasts of provisions. Financial institutions that can precisely justify the value of expected losses are perceived as more valuable for potential investors, influence the market valuation and raises competitiveness on the market (Montes et al. 2018).

Figure 1 Expected and unexpected loss (Basel Committee on Banking Supervision, 2005a)



Note: the formulas for expected and unexpected losses calculated within the Vasicek one-factor model are as follows (cf. Eqs (5.37) and (5.38) in Van Gestel and Baesens, 2009):

$EL_i = EAD_i \cdot LGD_i \cdot PD_i \cdot D_i$

$UL_i = dEAD_i \cdot dLGD_i \cdot \left(\Phi \left(\frac{\Phi^{-1}(PD_i) + \sqrt{\rho(PD_i)}\Phi^{-1}(99.9\%)}{\sqrt{1-\rho(PD_i)}} \right) - PD_i \right)$,

where $\Phi(.)$ denotes the cdf of a standard normal distribution, $\rho(PD)$ is a function for the default correlation, $\gamma(M_i)$ stands for maturity adjustment, $dEAD_i$ and $dLGD_i$ are downturn EAD and downturn LGD respectively, and where the single risk factor is fixed to a confidence level of 99.9%.

Previously, researchers and practitioners mainly focused on the individual creditworthiness expressed in PD. As a result, various methods for estimating PD have been established. On the other hand, we observed a growing research on the LGD in the last few years. Despite the importance of this parameter, both in capital requirements calculation and from accounting perspective, there is still a lack of a standardized set of estimation methods or even an agreed list of potential risk drivers with rationale about directions in which LGD is pushed. The ultimate goal of this thesis is to propose an efficient methods of estimating LGD. The estimation task carries great challenge, starting from calculating the actual values, selecting sound risk drivers and functional form, ending with demonstration that an estimation method is appropriate to institutions activities and showing precise/conservative[5] calibration results at the same time. Even if the definition of LGD according to Article 4 (22) of the Capital Requirements Regulation (CRR) is straightforward and expressed as a *ratio of the loss on exposure due to the default of the counterparty to the amount outstanding at bankruptcy*, it can be measured in four different ways. These alternatives are “workout LGD”, “market LGD”,

[5] Precise in case of IFRS 9, conservative for AIRB purposes.

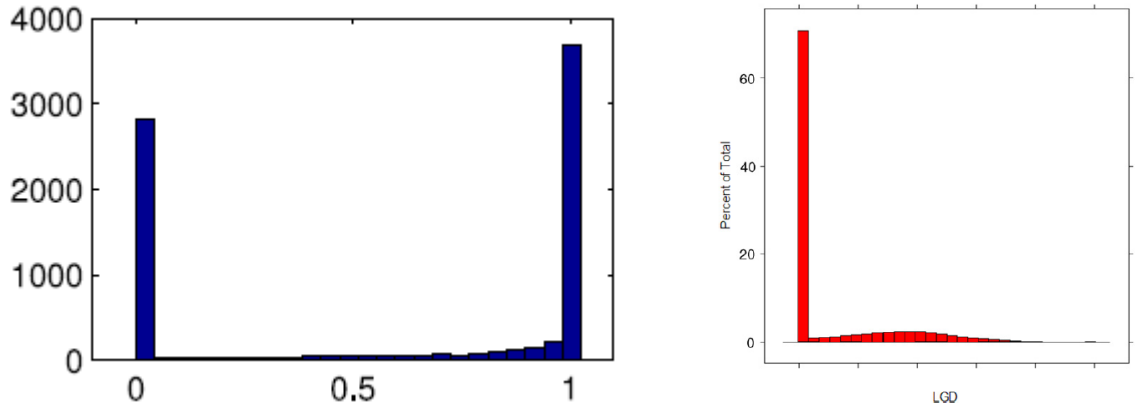
"implied market LGD" and "implied historical LGD". The latter two are considered implicit, as not established on realized LGD of defaulted facilities, but on spreads observed on non-defaulted loans, which approximate loss expectation of the market in the first case, and deriving LGD from realized losses and an estimate of default probabilities in second. Market LGD is applied by comparing market prices of bonds or commercial loans shortly after default with their par values. Finally, workout LGD is based on the institution owns loss and recovery experience. It is necessary to determine all recoveries and costs observed after default, discount them and compare with the value of defaulted exposure at the moment of default (see equation 1). Workout and market LGDs are called explicit as there is no need to extract information from selected sources, assuming that information about potential loss is accommodated inside but allows to compute it directly. It should also be noted that market and implied market LGD are measurable only on liquid markets making these methods impossible to use in specific circumstances. Given above, workout LGD should be in principle superior to other types of estimates, as it contains the most representative set of information to forecast future outcomes (European Banking Authority, 2017, p. 27). In the nominator of the workout LGD discounted cash flows are placed. Principal, interest and post-resolution payments, the book value of collateral realization, received fees, commissions, waivers and received money from selling the loan to a third party after write-off are the elements that affect its increase. Direct and indirect costs which decrease the nominator are legal expenses, administrator and receiver fees, liquidation expenses, staff salaries and other depending on institution structure. The definition of costs types that need to be included in the calculation cause serious complication, as it is not easy to assign each of it to particular loan, especially when it comes to split collection department salaries or court fees, when multiple loans are subject of the case. Although even if costs and recoveries are accessible, the workout approach suffers from the need of incorporating all defaults from the selected period into the sample. It means that also incomplete recoveries need to be taken into account, which results in having not fully observed dependent variable. Some approaches to deal with this issue can be found in Dermine and Neto (2006) or Rapisarda and Echeverry (2013). However, the common approach is based on treating incomplete defaults as completed (Baesens et al., 2016) which produces bias toward overestimation of the LGD.[6] One of the contributions of the thesis is to present a method for dealing with incomplete recovery processes, to reduce the bias coming from including/excluding them from the sample.

Contrary to PD estimation there is no consensus, what are the main risk drivers affecting LGD. Numerous studies on retail loans have been performed, but conclusions sometimes exclude each other. Most often repeated EAD (Tong and Mues and Thomas, 2013, Kruger and Rosch, 2017, Yao and Crook and Andreeva, 2017), loan amount (Thomas and Mues and Matuszyk, 2010, Bastos, 2010, Brown, 2012), LTV (Leow, 2010, Brown, 2012, Anolli and Beccalli and Giordani, 2013), or loan type (Zhang and Thomas, 2012, Tows, 2015, Hwang and Chu, 2017) still are not perceived as standard set of explanatory variables. What is more, the relationship between explanatory variables and LGD is not well established. For example, Bastos (2010) reported loan amount to have a negative effect on recovery rate (so higher loan amount indicate higher LGD), but Thorburn (2000) was not able to confirm any significant relationship. Brown (2012) found that age of exposure has a negative effect on LGD, but Bellotti and Crook (2010) found contrary results. Recent research reveals that different portfolios can be described by various set of covariates, consequently the best way to achieve decent results is to start with as a comprehensive set of variables as possible, taking into consideration the economic justification of each one. Second contribution of the thesis reveals the connection between client-behavior oriented variables and the LGD, as a new set of risk drivers.

[6] Consider the situation when current LGD is equal to 10%, but final LGD will be equal 5% as additional recoveries for open case will be obtained in future. Including 10% in the sample directly, leads to overestimation.

Finally, several functional forms are subject to consideration, as LGD suffers from strong bimodality (Schuermann, 2004) and is bounded between 0 and 1 only theoretically. Looking at the typical distribution of recoveries, two distinct humps are revealed, so recovery rate and consequently LGD is either close to zero or close to one (see Figure 2). What is more, in some cases, LGD can exceed boundaries which is a consequence of (a) including direct and indirect costs of collection (LGD higher than 1 when there is no recovery at all), (b) selling collateral at a value higher than exposure (LGD lower than 0, most common for leasing). Such unconventional distribution can be modeled with sophisticated functional forms like the fractional response regression proposed in Qi and Zhao (2011), or so-called two-stage modeling presented inter alia in Tanoue, Kawada, Yamashita (2017) with the probability of no-loss, probability of full-loss and LGD prediction when the loss occurs. Each component is estimated via a suitable method and finally, assemble predicts final loss. Among these methods, beta regression, support vector machines, regression trees, or survival analysis can be distinguished (cf. Baesens et al., 2016). Such framework can help not only to handle bimodality but also non-linearity between predictors and dependent variable, which is very common phenomena in LGD modeling. Third contribution of the thesis broaden the two-stage approach to the mixture distributions of the in-default events direction.

Figure 2 Typical distribution of the LGD (Loterman et al., 2012; Tong, Mues, Thomas, 2013)



The rest of introduction chapter is structured as follows. First, a detailed background of the regulatory environment, with emphasis on its implications on credit risk modeling, is demonstrated. Two standards play a crucial role: Basel II/III Capital Accord and IFRS 9. Then, the LGD estimation process is presented shortly. Finally, a list of contributions is given.

1.1 Regulatory environment

The financial sector is highly regulated, being under the control of external auditors as well as local authorities. This situation originates in (cf. Iwanicz-Drozdowska et al., 2017):

- being a trustee of peoples savings, as banks lend money acquired from its clients, who need certainty, that at any time there will be a possibility to withdraw it,
- being a bloodstream of the economy, realized in providing investment loans, billing or connecting business counterparts,
- fulfilling functions of the state, like confirmation of the identity or split payment implementation.

A lot of trusts must be put in place to provide these services, so both the banking sector and the state institutions make an effort to prevent a loss of confidence, as it can have severe implications to the economy as a whole. The regulators are keeping financial institutions from absorbing too much risk in their balance sheets and controlling that adopted methods are characterized by sound assumptions

and high level of understanding advantages and weaknesses. This is particularly important during assessing credit risk connected with the core activity, which is lending money.

CRR

According to Article 107(1) of the Capital Requirements Regulation, an institution shall apply either the Standardized Approach[7] or, if permitted by the competent authorities, the Internal Rating Based Approach (IRBA) to calculate their risk-weighted exposure amount for credit risk. In addition, the IRBA requires a financial institution to develop the internal models for estimating the PD, the EAD and the LGD. Chapter 3 of the CRR regulates details of each parameter, but in the thesis we focus on the LGD. These requirements are essential if institution want to meet the standard, and some of them have direct influence on the estimation process:

Article 174 allows institutions to use statistical models only if good predictive power can be proven, input variables form a reasonable and effective basis for the resulting predictors, and there are no material biases.

Article 175(4) describes that methodologies used in statistical models should provide a detailed outline of the theory, assumptions and mathematical and empirical basis of models used to estimate the exposures. Also, out-of-time and out-of-sample performance tests should be used, indicating all the circumstances under which the model does not work effectively.

Article 179(1)(f) orders to add to estimates a margin of conservatism related to the expected range of estimation errors. Consequently when methods and data are considered less satisfactory, the level of uncertainty is larger, and the margin of conservatism shall also be more extensive.

Despite the general character of the document, the specific guidelines concerning LGD estimation are addressed, which narrows the field in which the bank can operate defining its estimators. Enlargement of these ideas was prepared by European Banking Authority (EBA) in the *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures* (EBA/GL/2017/16), which was undertaken to reduce unjustified variability of risk parameters and own funds requirements. The guidelines focus on the definitions and modeling techniques used in the estimation of risk parameters. For this purpose, differentiation between model development and model calibration was proposed, so flexibility in terms of model development continues to be allowed. Still, calibration and the determination of capital requirements have to be identified objectively. Therefore, Chapter 6 of this document represents general requirements specific to LGD estimation, which are briefly introduced below.

Article 101 recognizes default observation as a separate one only if there is no return to default status in less than nine months after the first default ends. This means that two observations with time shorter than prescribed should be treated as one consequent default from the first moment when the default occurred.

Article 104 allows to use several methods to estimate LGD, especially with respect to a different type of collateral, which makes *two-stage* approach, as well as any other functional form, feasible.

Article 105 imposes an obligation to demonstrate that chosen methods are appropriate to the activities and type of exposures to which the estimates apply, and the ability to justify the theoretical assumptions underlying those methods should be proven. What is more, the methods used in LGD estimation should be consistent with the collection and recovery policies adopted by the institution

[7] Provided in Chapter 2 of the CRR.

and should take into account possible recovery scenarios. These could be particularly difficult when one wants to use a “black-box” method, like neural networks or SVM.

Article 107 concern data requirements with one crucial point about covering all the defaults from the selected observation period. This means that no removal of winsorizing is allowed (except errors in data) when it comes to dependent variable and what is made in the most of the research.

Article 108 treats incomplete recoveries as a part of the sample. Additionally, in article 153 incomplete recovery processes should be taken into account in a conservative manner.

Article 121 specify a base list of risk drivers, which need to be included in the sample. In particular, it should cover transaction-related characteristics, obligor-related characteristics, institution-related factors and external factors.

Article 151 allows weight observations (only for retail) in the estimation/calibration process to obtain more precise predictions of loss rates. But if some weights are set to zero or close to zero, there must be duly justification that it leads to more conservative estimates.

Only these few points limit modeling techniques mainly in data manipulation (like removing outliers) or estimation (the clear connection between method and recovery policies). All the above articles/requirements need to be met simultaneously if one wants to build IRB compliant model.

IFRS 9

The banks' second strategic and business challenge is adapting to the new environment under IFRS 9. Three main areas covered by this standard are:

- classification and measurement of financial instruments,
- impairment of financial assets,
- hedge accounting.

The previous standard was perceived as backwards-looking as it used incurred loss impairment model. Impairment of financial assets took place only if there was objective evidence of impairment as a result of a past event that occurred subsequent to the initial recognition of the financial asset. New standard base on forward-looking approach, which require lenders to recognize expected credit losses (ECL) over the life of financial instruments, either on 12-month or lifetime basis, depends on so-called three-bucket approach (financial instruments without significant rise in credit risk, financial instruments with significant rise in credit risk and impaired financial assets). The standard does not provide a specific method for calculating ECL, but admits that it may vary depending on the financial asset and available information. There is still no consistent approach when it comes to modeling process, but most commonly rely on using internal AIRB models and utilize them to calculate both one-year and lifetime credit losses. To adopt this solution, one should take into account a set of restrictions imposed by IFRS 9:

1. Institutions should use both internal and external information to calculate expected losses. This includes information about past events, current conditions, confirmed data about future events, and future macroeconomic situation.
2. When using historical data, an adjustment to reflect the current and future economic situation must be performed. Statistical extrapolation of historical data is not sufficient for this purpose.
3. In general, the use of the average values observed in the business cycle is not sufficient, and parameters should have more Point-in-Time philosophy (PIT) rather than Through-the-Cycle (TTC). TTC models generally leave aside the state of the overall economy by excluding

macroeconomic variables. PIT models explicitly controls for the state of the economy (Baesens et al., 2016). It is particularly important for LGD estimates as the most common approach concerns bias towards TTC for LGD and EAD also, which would make EL estimate predominantly TTC, even if PD is more cyclical.

4. The regulatory approach including a conservative buffer cannot be used. Downturn and indirect costs should be switched off in LGD calculation.
5. From the accounting point of view (unlike Basel II/III), there are no specific requirements for data. The IASB[8] expectation is based on the principle of the best available information, accessible without unnecessary costs or effort to obtain it.
6. Cash flow discounting should use effective interest rate, not interbank funding rate + add-on like in regulatory approach.

Taking above into consideration, there are strong premises to re-examine whole process of model building if financial institution want to calculate provisions in IFRS 9 manner, although logic and model structure developed under AIRB regime can be used to perform this task. However, it needs to be kept in mind that due to different samples, philosophies, discounting, etc., results will significantly differ. It is worth comparing obtained values as its compound produces Value at Risk, where accounting loss is a part of Expected Loss, and economic loss is a part of Unexpected Loss.

1.2 Loss Given Default modeling

Loss Given Default is the estimate of losses that the bank will face when customer or facility default. It is expressed as a percentage of EAD (Baesens et al., 2016):

$$LGD_i = 1 - RR_i = 1 - \frac{\sum_t CF_{it} d_t}{EAD_i} \quad (1)$$

where CF_{it} is the net cash flow at time t that comprises both positive and negative flows. Recoveries consist of principal, interest and post-resolution payments, the book value of collateral realization, received fees, commissions, waivers and received money from selling loan to third party after write-off. On the costs side, there are legal expenses, administrator and receiver fees, liquidation expenses, staff salaries and additional drawings. Second element d_t denotes discount factor, as all cash flows need to be expressed in a value appropriate at the moment of default. These could be risk-free rate plus premium in case of the AIRB approach, or effective interest rate for IFRS 9 purposes (cf. Bellini, 2019). In the "workout approach" actual LGD is calculated for each default to achieve ultimate goal, which is assigning LGD estimate to non-defaulted and currently defaulted portfolio. Thus, it is a conditional parameter that aims to approximate how significant the loss will be if a non-defaulted client goes into default or a non-conditional parameter for defaulted customers when in-default loss is estimated. The estimation process can be divided into several steps described in Figure 3.

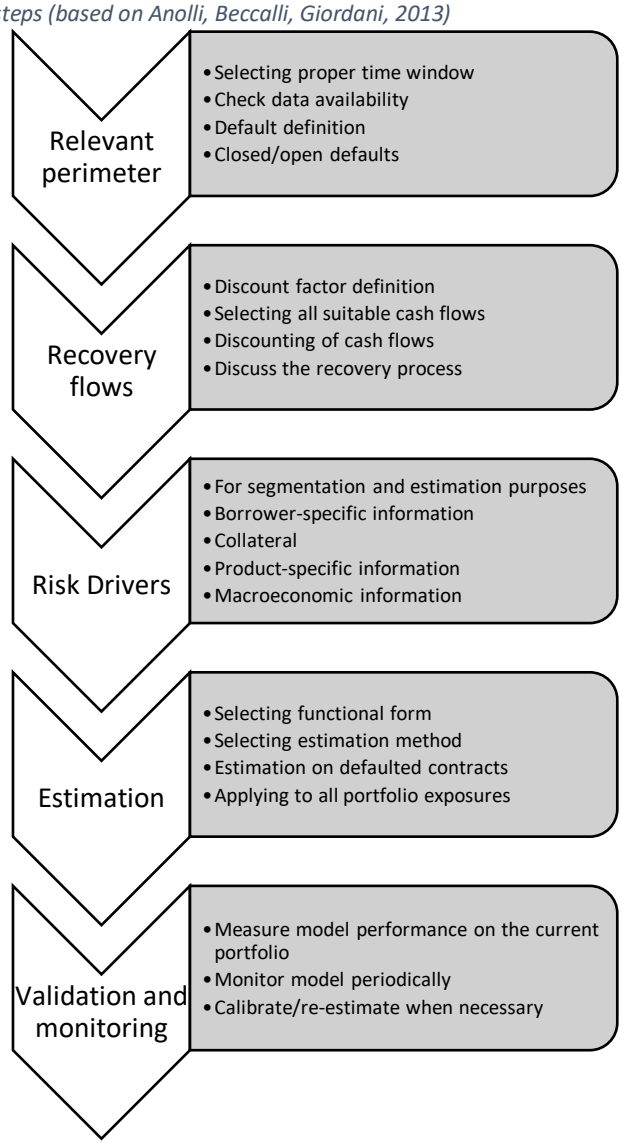
Relevant perimeter

Right at the beginning, when reference data set is created, default definition and business collection process fit should be ensured. The definition of default could be set on client level or on contract level, which implies different behavior of the contracts finally included in the sample as client level approach set all client contracts marked as default, even those without default trigger (such as days past due (DPD) or due amount). This is also connected with the collection process, as different approaches can be assigned to separate contracts or to various clients (in a client-oriented approach). Secondly, the

[8] International Accounting Standard Board.

time window should be as broad as possible (at least five years according to Basel II, but could be less in case of impairment models) to cover all recovery patterns. What is more, it should be checked how many incomplete defaults will be included in the estimation process (as a result of time window definition), how inclusion itself will look like and what is the relation of closed to open defaults in the sample (low values of this share could lead to less reliable estimates).

Figure 3 Model development steps (based on Anolli, Beccalli, Giordani, 2013)



Recovery flows

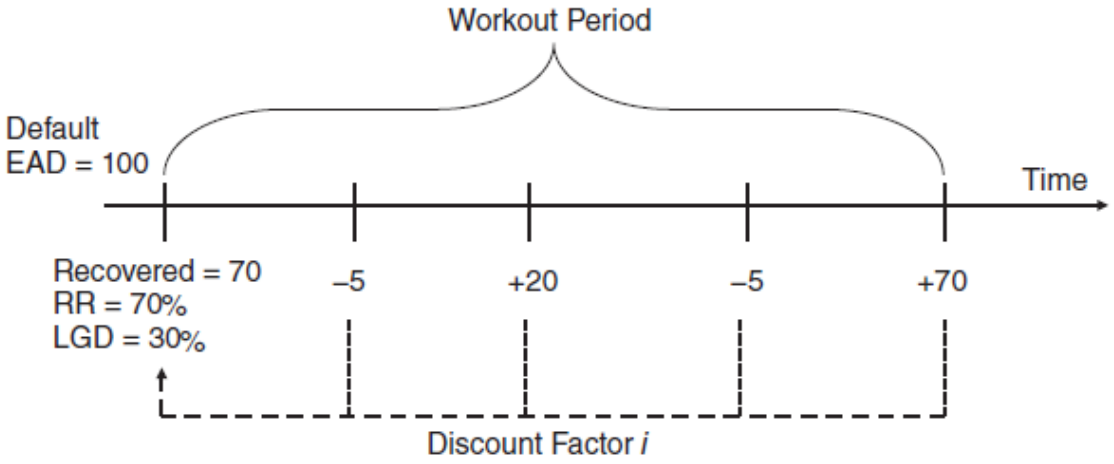
In the second stage choice of discount rate should be made. As LGD represent economic loss, it should reflect costs of holding the defaulted asset over the workout period by taking into account two aspects, (a) time value of money and (b) risk premium for undiversifiable risk (Basel Committee on Banking Supervision, 2005b). Therefore, selecting a discount rate appropriate to these points is crucial in the estimation process, as the higher the discount rate, the higher is actual LGD. Maclachlan (2004) provides a survey of different approaches, and he distinguished:

- Contractual contract rate, which is in most common usage. Interpretation is appealing as it should represent the cost of replacing the promised return on the defaulted loans. But is mixes the pre and post-default requirements for the returns and seems to have only little to do with explaining post-default LGD. For example, default can result in a change in expected cash flows

relative to promised payments (like in the case of loan term modifications) which finally leads to a higher or lesser appropriate discount rate than a contractual one.

- Lender's Cost of Equity. As Maclachlan states, *the rationale is that shareholders cover the cost of recapitalizing the bank's balance sheet. This method mistakenly replaces the systematic risk of the defaulted debt with the risk of the bank*. The risk premium mentioned in Basel II is not equal to bank's spread, so taking lender's cost of equity as a discount rate will lead to equating these two values.
- Risk-free rate (+ add-on). This approach is highly recommended by EBA (European Banking Authority, 2017, p.32) and consists of two elements. First is predefined in advance risk-free rate (interim funding rate or some equivalent), second is given add-on which target is to remove unjustified variability in LGD estimates. In this definition discount rate should be focused on the uncertainty inherent in the recovery process rather than funding costs. The main problem connected with this approach is placed in the add-on structure. Removing variability could lead to selecting one add-on for all. Still, it does not seem to be correct as each portfolio (e.g. credit cards vs mortgage loans) is characterized by different patterns when it comes to recovery processes.
- Ex-post Defaulted Bond Returns (not applicable for retail, so we will not discuss it here).

Figure 4 Workout LGD (Baesens et al., 2016, p.273)



The next step is selecting all appropriate cash flows from a given period. As stated before, cash flows consists of recoveries and costs. Recoveries can be obtained from client own payments, collaterals, guarantees, insurances and write-offs with sale. Own payments are the most typical source of recoveries in retail, as modern collection departments strategies lead to help client first, in order to continue a healthy relationship, and in case of no success, undertake legal actions. Then, when a secured loan is considered, recovery from collateral is possible, but it should be noted that such can carry more than one loan, so proper allocation need to be taken. Guarantee involves a third party willing to pay some part of the debt which is also a pattern in case of insurance, but the latter one can be initiated under some conditions (like losing a job). Finally, a write-off with sale is done mostly when there is nothing left to recover on the bank side, so an agreement with an external company is made and the package of credits is negotiated about the price. On the costs side, there is a need to include direct ones (connected strict to the analyzed contract) and indirect ones to represent the true value of economic loss. Committee of European Banking Supervisors (2005) states that *workout and collection costs should include the costs of running the institution's collection and workout department, the costs of outsourced services, and an appropriate percentage of other ongoing costs, unless an institution can demonstrate that these costs are not material*. The most common approach to deal with this issue

is about calculating the time-weighted average of the workout costs divided by the total exposure at default. Other approach replaces total exposure by the recoveries amount (Baesens et al., 2016). At last, so-called Exposure at Workout needs to be defined as a value of principal and interest with which the client comes back to non-default status in case of repaying all due amounts.

Figure 4 represents an example of the LGD calculation in the workout approach. The workout period can comprise numerously observation points at which cash flows are observed. The example given, represents calculated LGD in this approach as 1 – RR.

Risk drivers

The third stage considers risk drivers where five major quantifiable types of LGD predictors can be distinguished (Ozdemir, Miu, 2009):

- collateral,
- debt type and seniority class,
- borrower-specific information,
- industry/transaction/product-specific information,
- country and regional macroeconomic information.

In the case of retail, when collateral is considered, also guarantees and insurances are included. Besides the economic value of collateral, its character (commercial or non-commercial), market (liquid or illiquid) or type (flat, house, etc.) may serve as input. There are numerous conditions to meet if institution wants to include certain types of collaterals in their estimates described in sections 6.1.3, 6.2.2 and 6.2.3 of European Banking Authority (2017) considering adequate value of repossession, haircuts reflecting errors in valuation, legal certainty, etc. These guidelines determine the usage of collaterals in the estimation process compliant with AIRB methodology. Secondly, debt type and seniority class concern bonds and specialized lending in the highest degree. Still, then there is information about the borrower, often omitted in LGD analysis focused on contract characteristics. Typically age, net income, product structure, savings, etc., are recognized as well-performing predictors of final recovery rate and should be included in the reference data always set when possible. Regarding industry/transaction/product information, the standard group of variables should include EAD, tenor, LTV or interest rate (compare with Tong, Mues, Thomas, 2013 or Yao et al., 2017). Finally, there is macroeconomic information which is also perceived as canonical LGD predictors, but in the case of retail, it is hard to demonstrate any dependencies which hold in long-run average. Nevertheless, variables like House Price Index, Unemployment Rate, Consumer Price Index or Gross Remuneration have a clear economic relationship with the LGD, thus can be considered as LGD predictors. All risk drivers mentioned above can be a part of the estimation process, but there is another way to include them into the model, which is segmentation. It could be a significant enhancement that can easily boost its predictive ability and interpretation capabilities. For non-defaulted portfolio, the most common segmentations are by product type (secured/non-secured or with deterministic/stochastic repayment plan) or by EAD (small exposures vs big exposures). For defaulted portfolio, other dimensions are considered as time in default (0-6 months, 7-12 months, and so on) or recovery path (before/after collateral realization or in-house/agent collection).

Estimation

After observed LGDs have been computed and risk drivers were selected, the estimation process is held. Usually, bank is interested both in determining LGD forecasts before default will happen and after the event as well. First is of great importance in setting capital buffers, second in provisions, but the

whole portfolio needs to be rated in each approach. Re-introducing the connection between LGD and RR:

$$LGD = 1 - RR \tag{2}$$

it is apparent that LGD is just a part of exposure that was not recovered. In general, there is no difference if LGD or RR is modeled. The elementary way of obtaining precise estimates of LGD is to apply a conditional mean model. At the beginning, a set of segments is distinguished (based on statistical verification or as an input from a panel of experts or a combination of both), and then average LGDs are computed for each segment. Then, the portfolio is segmented in the same way as the sample, and calculated averages are assigned as the estimates of LGD. Even if simplified, this approach can give results that are good enough, when initial approach is implemented. It can also serve as the benchmark for more sophisticated methods to see if introducing more complexity is justified. The second step in LGD estimation analysis usually includes various kinds of regression models.

Below, we discuss canonical models for LGD:

Linear Regression:

$$LGD_i = X_i'\beta + \varepsilon_i \tag{3}$$

where hereinafter i indicates consecutive observation and $X_i = (X_{i1}, X_{i2} \dots X_{iK})'$ is $K \times 1$ vector of the risk drivers (explanatory variables) of observation i and β is $K \times 1$ vector of unknown regression parameters.

Even if almost none of the linear regression assumption is met in LGD estimation, the method is still widely used, mainly because of straightforward interpretability of the estimates and the results. Recently it can be found in Yao et al. (2017) or in Tanoue, Kawada, Yamashita (2017). It can be a good choice if one want to perform an initial insight into data, set the benchmark or check linear relationship between LGD and risk drivers. However due to auto-correlation, non-normal distribution and heteroscedasticity of the error term ε_i it is not recommended to use it without comparison to other approaches.

Fractional Response Regression:

$$E(LGD_i|X_i) = G(X_i'\beta) \tag{4}$$

where $E(\cdot\,|X_i)$ is the conditional expectation and G satisfies $0 < G(z) < 1$ for all $z \in \mathbb{R}$, eg., the logistic function $G(X_i'\beta) = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$ (see Papke and Wooldridge, 1996).
In this kind of model only conditional mean needs to be correctly specified in order to obtain consistent estimators, which is undoubtedly big advantage in case of LGD estimation. This method was used with success in research by Bastos (2010) and Qi and Zhao (2011). However, it should be noted that the explained variable must come from a specific range, which is not always ensured in the case of recovery rate modeling.

Beta Regression:

$$G(\mu_i) = X_i'\beta \tag{5}$$

where we assume that LGD_i admits beta distribution $\mathrm{B}(p_i, q_i)$ with parameters $p_i, q_i > 0$ such that the mean satisfies $\mu_i = E(LGD_i|X_i) = \frac{p_i}{p_i+q_i}$, and G: $(0,1) \to R$ is a link function (cf. Ferrari and Cribari-Neto, 2004). As beta distribution is supported on $(0; 1)$ there is a need to apply specific treatment to the values at the ends of the interval. Such regression is known for its ability to reflect many kinds of

probability density function just by tuning two parameters (unimodal, U-shaped, J-shaped and many more). Usage, with good performance, was shown in Chalupka and Kopecsni (2008), Bellotti and Crook (2012) or Tong, Mues, and Thomas (2013). But as Qi and Zhao (2011) argued, there is a need to investigate the model sensitivity to different parametrization resulting from data distortion or extreme values of LGD. Performance measures on the edges of the distribution can be significantly worse than in other methods. Both fractional response and beta regressions can be estimated by means of maximum likelihood methods.

In the real-life application, most of these approaches do not provide a significant upgrade compared to conditional mean and still, performance measures are relatively poor. Nowadays, a switch to more complex methods can be observed (see Thomas, Mues, Matuszyk, 2010, Loterman et al., 2012 or Nazemi et al., 2017). The most popular group consisting of Regression Trees and Support Vector Machines. We obtain the estimates of hyperparameters of these methods by minimizing a selected criterion (see Hastie, Tibshirani, Friedman, 2008). Tree-based methods recursively partition the original sample into smaller subsamples and then fit a model in each one. The algorithm evaluates every possible split (at each node, every variable and every value of this variable is checked) to find one which maximizes the decrease in impurity. Such an approach provides robust results when there are many non-linearities in data or when variables are highly correlated. On the other hand, the instability issue is well-known disadvantage, as well as the need to tune hyperparameters and lack of smoothness when it comes to the results (tree ends with a set of rules, so estimated values come from the discrete set). Models based on this approach can be found inter alia in Bastos (2010) or Brown (2012).

Support Vector Machines were proposed by Vapnik (1995), and due to their ability to solve highly non-linear problems, they have become more popular when estimating LGD (see Tobback et al., 2014 or Yao et al., 2017). Intuitively SVM maps the features to a higher dimensional space and tries to find a hyperplane that best fits the dependent variable. It produces non-linear regression by mapping the hyperplane from the transformed feature space. Mentioned kernel helps with finding non-linear dependencies, so like in tree-based methods, the structure of LGD and its predictors could be handled appropriately. SVM is perceived to be more stable than regression tree but still suffers from the need to calibrate hyperparameters, and in contrast to regression trees it is not interpretable in the desired way. Nevertheless, its usage is proven to give reliable results (see Nazemi et al., 2017 or Hurlin, Leymarie, Patin, 2018).

Literature concerning LGD estimation methods is still growing and proposals like Bagging (Nazemi, Fabozzi, 2018), Tobit Regression (Tong, Mues, Thomas, 2013), Finite Mixture Models (Tows, 2015), Regression Splines (Miller, 2017), Survival Analysis (Zhang, Thomas, 2012), Quantile Regression (Kruger and Rosch, 2017), Neural Networks (Brown, 2012) or Markov Chains (Luo, Shevchenko, 2013) were also adapted for LGD predictions.

All the approaches mentioned above form a group called one-stage LGD models, as focusing on estimating LGD as a whole in one step. However, nowadays, various kinds of decomposition of the LGD is being proposed, which is mainly about separate bi-modal distribution of the predicted phenomenon into two sub-models. There are at least two alternative approaches/two-stage LGD models:

- Cure rate modeling, which includes the probability of no-loss ($LGD = 0\%$) estimation.
- Danger rate modeling, which introduces the probability of full-loss ($LGD = 100\%$) estimation.

First, probabilities of the events and then expected conditional severity are modeled and estimated. The combination of both components leads to the final LGD estimate.

$$E(LGD_i) = Pr(cure_i) \cdot LGD_{i,LGD=0} + \big(1 - Pr(cure_i)\big) \cdot LGD_{i,LGD>0} \qquad (6)$$

In the notation above $Pr(cure_i) = Pr(LGD_i = 0\%)$ denotes the probability of being cured, so exiting default with no-loss, $LGD_{i,LGD=0}$ is estimated expected loss in case of being cured (it can be assumed to be 0, but if discounting effect is taken into account, this value can be greater than no loss at all) and $LGD_{i,LGD>0}$ is estimated loss for non-cured cases. In case of danger rate equation above is of similar structure:

$$E(LGD_i) = Pr(full\ loss_i) \cdot LGD_{i,LGD=1} + \big(1 - Pr(full\ loss_i)\big) \cdot LGD_{i,LGD<1} \qquad (7)$$

This approach is called two-staged and appear to reflect the nature of the LGD more appropriately. However, as specific status-dependent samples need to be prepared for the estimation of various stages of default, one needs to have wide observation window to monitor status changes of credit exposure at each point during default duration. When it comes to loss severity, all the methods mentioned above can be applied, and for the probability of cure or full-loss new ones were proposed. The most evident is logistic regression, as it is widely used for binary classification (cf. Gruszczyński, 2012). On the other hand, linear and quadratic discriminant analysis, decision trees (Brown, 2012) or least squares support vector classifier (Yao et al., 2017) can be distinguished. Two-stage modeling is gaining more and more popularity in academia and practice. It allows reflecting underlying recovery patterns being hidden in highlighted components, which could be the answer to reflecting bi-modal LGD distribution. Presently, developing this kind of constructions is the most promising way to obtain more precise estimates than those currently achieved. Finally, a new stream of estimation techniques arises nowadays, which is forecasts averaging. Closing contribution of the thesis utilize this concept to make a proper use of macroeconomic variables in the LGD estimation process.

Validation and monitoring

The final stage of the model development phase includes three actions that need to be undertaken right after estimation – validation, or at a specific time after model development - monitoring and re-calibration/re-estimation. Validation tests a newly developed model on out-of-sample and/or out-of-time data in order to evaluate the goodness of the estimates. In the LGD case, four areas need to be investigated (Ozdemir, Miu, 2009):

1. Discriminatory power (low vs high values);
2. The precision of the calibration (mainly realized in specified pools);
3. Realization of the LGD risk rating philosophy;
4. Testing homogeneity of the LGD risk ratings (sub-portfolio level).

Limited data and large standard deviations associated with the U-shaped distribution (see Figure 2) of the LGD may create significant noise in any test based on exposure-by-exposure analysis, so there is a need to take each test with great care and on the appropriate level of granularity. In some cases, it is worth grouping observations in some predefined ranges to reduce the noise and get a more accurate view of the model results.

The purpose of the discriminatory test is to validate the correctness of the ordinal ranking of the exposures by the model. If the analysis is done on the exposure level, we expect observations with high LGD to have, on average higher model value than otherwise. To assess the discriminatory power of a LGD model, the following tests are used:

- Gini coefficient. It proves its usage in PD models, so adapting it in the LGD framework should be considered.

- The cumulative LGD accuracy ratio (CLAR) curve (see Ozdemir, Miu, 2009) can be treated as the equivalent of the Cumulative Accuracy Profile (CAP)[9] curve in PD models performance analysis. In short, this construction is about the comparison of the cumulative percentage of correctly assigned realized LGD (in defined bands) on the vertical axis and the cumulative rate of observations in the predicted LGD bands on the horizontal axis. Given perfect discrimination drawn line will is located on a 45-degree line.

In the case of testing calibration of the LGD model, many different statistical tests were developed. Still, neither of them should be viewed in isolation but instead combined with other results to draw any definitive conclusion. Again, exposure-based or pool-based methodology can be assigned to each test but scarcity of data should be always taken into consideration:

- MSE (Mean Squared Error) between realized and predicted LGD values, one of the most often used error-based metrics (Loterman et al., 2014), suffers from sensitiveness to extreme values and lack of reference level.
- SSE (Sum of Squared Errors), which is valid only when developing several LGD models on the same sample. No reference level is provided.
- Correlation coefficients between realized and predicted RRs, mainly Pearson's, Spearman's and Kendall's.
- TIC (Theil inequality coefficient) sets the mean squared error relative to the sum of the average quadratic realized and estimated LGD. It has two useful features. First, it accounts for the goodness of fit and robustness at the same time. Second, Theil found out that useful forecast can be made up to $TIC \approx 0.15$, which allows constructing a benchmark.
- Regression Error Characteristic Curve, an equivalent for the receiver operating characteristic curve (ROC curve), provides a powerful tool for visualizing the distribution of regression errors. It plots the error tolerance on the horizontal axis and the percentage of points predicted within the tolerance on the vertical axis. The resulting curve estimates the cumulative distribution function of the error.
- Central tendency error tests help assess whether the model tends to under- or over-estimate the true LGD. Mean of the test-set error is tested under a hypothesis set as $H_0: \mu_E = 0, H_\alpha: \mu_E > 0$, which can be done via t-test or Wilcoxon signed rank test.
- Error dispersion tests are meant to detect whether test-set error distribution is getting wider. The F test and the Ansari-Bradley test allow to evaluate if this dispersion is wider on the new collected set than on the previous one and the test hypothesis is formulated like this: $H_0: \sigma_{in}^2 = \sigma_{out}^2, H_\alpha: \sigma_{in}^2 > \sigma_{out}^2$, where σ_{in}^2 is a variance of test-set error and σ_{out}^2 is a variance of training-set error.

Testing the realization of risk rating philosophy concerns checking consistency with an adopted cyclical or acyclical approach over an economic cycle. Under a cyclical approach, predicted LGD should be synchronized with the cycle (PIT), whereas under acyclical predicted LGD should remain constant over the cycle (TTC). To assess the degree of cyclicality, mobility metric, like evaluating the absolute deviation and Euclidean distance between composed matrix and identity matrix, could be calculated (see Jafry and Schuermann, 2004). Finally, homogeneity testing provides further insight into the performance of the LGD model on the sub-portfolio level. When feasible this analysis can show if the homogeneity assumption of various LGD risk drivers is valid.

After the validation, the developed solution needs to be subject to the monitoring process. The primary purpose of such is to verify the assumptions of the model (as a part of the qualitative analysis) and the

[9] See Basel Committee on Banking Supervision (2005c) for details.

correctness and stability of the parameters over time (as a part of the quantitative analysis). Regular monitoring should be handled at least annually, preferably quarterly. A qualitative analysis should include verification whether since last monitoring/calibration there have been significant changes in credit standards, credit policy or in legal environment regarding debt collection, bankruptcy law or collateral repossession that may affect the level of risk distribution or the scope of risk drivers in the portfolio covered by the model. Quantitative analysis should include backtesting broken down into all data gathered and only cases opened or resolved after model building. The same set of measures as in the validation process can be used to perform this task. Additionally, it is necessary to complete (Basel Committee on Banking Supervision, 2005c):

- stability analysis of the estimates after changing timeframe,
- comparison between the LGD estimates and relevant external data sources, taking into consideration different default definition, biases in external data samples and different measures of losses,
- comparison between realized LGD of new defaulted exposures and their LGD estimates with consideration of model philosophy as realized losses are point in time and LGD model usually generates through the cycle estimates.

All these examinations should lead to a decision if the model assumptions still hold or there is an area to implement modifications. The remedial actions set should consist of model calibration on new/changed timeframe, re-estimation or changing the model structure.

The procedure of LGD model building, validation and monitoring imposes a complete understanding of each stage as the model developer should know where and how any change could be implemented. This is another argument against “black-box” approaches which could always lead to starting from the scratch, as there may not be an easy way to tune just a part of the model. Comprehensibility is often the key requirement, as all the users should be able to understand the logic behind the prediction of the model. Nevertheless, it should be noted that interpretable machine learning is still growing and can lead to the revision of the current perspective (cf. Chlebus, Kłosok, Biecek, 2020).

1.3 Contributions

The presented thesis consists of four essays dealing with the modeling and estimation of the LGD for retail contracts. In brief, four concepts are presented. First, the recommendation for unfinished defaults inclusion in the modeling sample is determined, as an inevitable part of the process not well developed in the literature so far. Second, the inclusion of new risk drivers connected to client behavior after granting credit is analyzed. Third, a new form of LGD decomposition is proposed, based not directly on the LGD distribution but rather on events that leads to the bi-modal shape. At last, forecast averaging way of macroeconomic variables inclusion in the LGD model is presented as a possible technique to combine idiosyncratic bank data with systematic factors related to macroeconomics.

1.3.1 Modeling Recovery Rate for Incomplete Defaults using Time Varying Predictors

The Internal Rating Based (IRB) approach requires that financial institutions estimate the Loss Given Default (LGD) parameter not only based on closed defaults but also considering partial recoveries from incomplete workouts. This is one of the key issues in preparing bias-free samples, as there is a need to estimate the remaining part of the recovery for incomplete defaults before including them in the modeling process. In this paper, a new approach is proposed, where parametric and non-parametric methods are presented to estimate the remaining part of the recovery for incomplete defaults, in pre-defined intervals concerning sample selection bias. Additionally it is shown that recoveries are driven by different set of characteristics when default is aging. As an example, a study of major Polish bank is

presented, where regression tree outperforms other methods in the secured products segment, and fractional regression provides the best results for non-secured ones.

Hypothesis 1: Remaining part of the incomplete default is driven by different characteristics, depending on the current status of credit exposure.

Hypothesis 2: Secured and non-secured loans include different patterns, which can be reflected by non-parametric and parametric method consecutively.

1.3.2 Beyond the contract. Client behavior from origination to default as the new set of the Loss Given Default risk drivers.

Studies on modelling Loss Given Default (LGD) are becoming increasingly popular as it becomes the crucial parameter for setting the capital buffers under Basel II/Basel III and for calculating impairment of the financial assets under IFRS 9. The most recent literature on this topic, mainly focuses on the estimation methods and less on variables used in explaining LGD variability. The following study attempts to expand the part of modelling process by constructing a set of client behavior based predictors, which can be used to construct more precise models. The paper investigates economic justifications by means of empirical studies to examine the potential usage. The main novelty introduced in the paper is establishing connection between LGD and behavior of contract owner, not just the contract itself. Such approach results in the reduction of the values of selected error measures and consecutively improves forecasting ability. The effect is more visible in a parametric method (Ordinary Least Squares) than in a non-parametric (Regression Tree). The research suggests incorporating client-oriented information into LGD models.

Hypothesis 3: Client behavior after loan granting becomes important part of loss given default risk drivers set.

1.3.3 LGD decomposition using mixture distributions of in-default events

Modeling loss in the case of default is a crucial task for financial institutions to support the decision making process in the risk management framework. It has become an inevitable part of modern debt collection strategies to keep promising loans on the banking book and to write off those that are not expected to be recovered at a satisfactory level. Research tends to model Loss Given Default directly or to decompose it based on the dependent variable distribution. Such an approach neglects the patterns which exist beneath the recovery process and are mainly driven by the activities made by collectors in the event of default. To overcome this problem, we propose a decomposition of the LGD model that integrates cures, partial recoveries, and write-offs into one equation, defined based on common collection strategies. Furthermore, various levels of data aggregation are applied to each component to reflect the domain that influences each stage of the default process. To assess the robustness of our approach, we propose a comparison with two benchmark models on two different datasets. We assess the goodness of fit on out-of-sample data and show that the proposed decomposition is more effective than state-of-the-art methods, maintaining a strong level of interpretability.

Hypothesis 4: LGD decomposition based on in-default event leads to precision uplift.

1.3.4 Forecast combination approach in the Loss Given Default estimation

This paper examines a novel method of including macroeconomic variables into Loss Given Default models. The approach is transparent, and it easily translates changes in the overall credit environment into Expected Loss estimates, which is one of the crucial points that was recently introduced in the IFRS 9. We propose a forecast combination procedure that, separates the contract-based variables from the macroeconomic indicators. Two models are prepared and benchmarked to a single ordinary

least-squares (OLS) model. To combine the forecasts we use three approaches: the equal weighting scheme, the Granger-Ramanathan Method, and Mallows Model Averaging. We tested our predictions on out-of-time data and found that the forecast combination outperforms the single OLS model in terms of the selected forecast quality metrics.

Hypothesis 5: Forecast averaging from models based on idiosyncratic and systematic variables separately, leads to precision improvement of long-term forecasts for LGD parameter.

Rererences

Anolli, M., Beccalli, E., Giordani, T., *Retail Credit Risk Management*, Palgrave MacMillan, New York 2013.

Baesens, B., Roesch, D., Scheule, H., *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*, John Wiley & Sons, 2016.

Basel Committee on Banking Supervision, *An Explanatory Note on the Basel II IRB Risk Weight Functions*, 2005a.

Basel Committee on Banking Supervision, *Guidance on Paragraph 468 of the Framework Document*, 2005b.

Basel Committee on Banking Supervision, *Studies on the Validation of Internal Rating Systems*, 2005c.

Bastos, J., *Forecasting bank loans loss-given-default*, Technical University of Lisbon, 2010.

Bellini, T., *IFRS 9 and CECL Credit Risk Modelling and Validation*, Academic Press, 2019.

Brown, I., *Basel II Compliant Credit Risk Modelling,* University of Southampton, Southampton 2012.

Chalupka, R., Kopecsni, J., *Modelling Bank Loan LGD of Corporate and SME Segments*, IES Working Paper, 2008.

Chlebus, M., Kłosok, M., Biecek, P., *Model interpretability and explainability. Credit Scoring in Context of Interpretable Machine Learning* edited by Kaszyński, D., Kamiński, B., Szapiro, T, Oficyna Wydawnicza SGH, 2020.

Committee of European Banking Supervisors (CEBS), *Guidelines on the Implementation, Validation and Assessment of Advanced Measurement (AMA) and Internal Ratings Based (IRB) Approaches*, Technical report, CP10 consultation paper, 2005.

Dermine, J., Neto de Carvahlo, C., *Bank Loan Losses-Given-Default: a Case Study*, Journal of Banking and Finance, Vol. 30/4, 2006.

European Banking Authority, *Guidelines for the estimation of LGD appropriate for an economic downturn ("Downturn LGD estimation")*, 2019.

European Banking Authority, *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures*, 2017.

Ferrari, S., Cribari-Neto, F., *Beta Regression for Modelling Rates and Proportions*, Journal of Applied Statistics, 31:7, 2004.

Gruszczyński, M., *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Wolters Kluwer Polska, 2012.

Gürtler, M., Hibbeln, M., Usselmann, P., *Exposure at default modeling – A theoretical and empirical assessment of estimation approaches and parameter choice*, Journal of Banking and Finance 91, 2018.

Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning*, Springer, 2008.

Hibbeln, M., *Risk Management in Credit Portfolios: Concentration Risk and Basel II*, Springer Science & Business Media, 2010.

Hurlin, C., Leymarie, J., Patin, A., *Loss functions for Loss Given Default model comparison*, European Journal of Operational Research, 2018.

Hwang, R., Chu, C., *A logistic regression point of view toward loss given default distribution estimation*, Quantitative Finance 18 (3), 2017.

Iwanicz-Drozdowska, M., Jaworski, W., Szelągowska, A., Zawadzka, Z., *Bankowość*, Poltex, Warszawa, 2017.

Jafry, Y., Schuermann, T., *Measurment, estimation and comparison of credit migration matrices*, Journal of Banking and Finance 28, 2603-2639, 2004.

Kruger, S,. Rosch, D., *Downturn LGD modeling using quantile regression*, Journal of Banking and Finance 79, 2017.

Leow, M., *Credit Risk Models for Mortgage Loan Loss Given Default*, University of Southampton, Southampton 2010.

Luo, X., Shevchenko, P., *Markov chain Monte Carlo estimation of default and recovery: dependent via the latent systematic factor*, Journal of Credit Risk 9 (3), 41-76, 2013.

Loterman, G., Brown, I., Martens, D., Mues, C. and Baesens, B., *Benchmarking Regression Algorithms for Loss Given Default Modelling*, International Journal of Forecasting, Vol. 28, No. 2012, 161–170, 2012.

Loterman, G., Debruyne, M., Vanden Branden, K., Van Gestel, T., Mues, C., *A Proposed Framework for Backtesting Loss Given Default Models*, Journal of Risk Model Validation, Vol. 8, No. 1, 69-90, 2014.

Maclachlan, I., *Choosing the Discount Factor for Estimating Economic LGD*, In: Altman E, Resti A, Sironi A (Eds.): Recovery Risk, The Next Challenge in Credit Risk Management. Risk Books, London, 285–305. 2004.

Merton, R. C., *On the pricing of corporate debt: The risk structure of interest rates*, Journal of Finance 29, 1974.

Miller, P., *Modeling and Estimating the Loss Given Default for Leasing Contracts*, Universität zu Köln, 2017.

Montes, C., Artigas, C., Cristófoli, M., Segundo, N., *The impact of the IRB approach on the risk weights of European banks*, Journal of Financial Stability 39, 2018.

Nazemi, A., Fabozzi, F., *Macroeconomic variable selection for creditor recovery rates*, Journal of Banking and Finance, 2018.

Nazemi, A., Pour, F., Heidenreich, K., Fabozz,i F., *Fuzzy decision fusion approach for loss-given-default modeling*, European Journal of Operation Research 262, 2017.

Ozdemir, B., Miu, P., *Basel II Implementation. A Guide to Developing and Validating a Compliant, Internal Risk Rating System*, McGraw-Hill, 2009.

Papke, L., Woolridge, J., *Econometric method for fractional response variable with an application to 401(K) plan participation rates*, Journal of Applied Econometrics, 1996.

Prorokowski, L., *IFRS 9 in credit risk modelling*, Bank i Kredyt 49(6), 2018.

Rapisarda, G., Echeverry, D., *A Non-parametric Approach to Incorporating Incomplete Workouts Into Loss Given Default Estimates*, Journal of Credit Risk, 2013.

Qi, M., Zhao, X., *Comparison of modeling methods for Loss Given Default*, Journal of Banking & Finance, 2011.

Schuermann, T., *What Do We Know About Loss Given Default?*, Wharton Financial Institutions Center Working Paper No. 04-01, 2004.

Starosta, W., *Beyond the Contract: Client Behavior from Origination to Default as the New Set of the Loss Given Default Risk Drivers*, Journal of Risk Model Validation, Vol. 15, No. 1, 2021.

Starosta, W., *Forecast combination approach in the loss given default estimation*, Applied Economics Letters, 2020.

Starosta, W., *Loss given default decomposition using mixture distributions of in-default events*, European Journal of Operational Research 292, 2021.

Starosta, W., *Modelling Recovery Rate for Incomplete Defaults Using Time Varying Predictors*, Central European Journal of Economic Modelling and Econometrics 12, 2020.

Tanoue, Y., Kawada, A., Yamashita, S., *Forecasting Loss Given Default of bank loans with multi-stage models*, International Journal of Forecasting 33, 2017.

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

Thorburn, K.S., *Bankruptcy auctions: costs, debt recovery, and firm survival*, Journal of Financial Economics 58, 337 – 368, 2000.

Tong, E., Mues, C., Brown, I., Thomas, L., *Exposure at default models with and without the credit conversion factor*, European Journal of Operational Research 252, 2016.

Tong, E., Mues, C., Thomas, L., *A zero-adjusted gamma model for mortgage loss given default*, International Journal of Forecasting 29, 2013.

Tows, E., *Advanced Methods for Loss Given Default Estimation*, Universitat zu Koln, 2015.

Van Gestel, T., Baesens, B., *Credit Risk Management. Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*, New York, NY: Oxford University Press, 2009.

Vapnik, V., *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.

Vasicek, O., *Loan Portfolio Value*, Risk December, 160-162, 2002.

Yao X., Crook J., Andreeva G., *Enhancing two-stage modelling methodology for loss given default with support vector machines*, European Journal of Operational Research 263, 2017.

Zhang, J., Thomas, L., *Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD*, International Journal of Forecasting, Volume 28/1, 2012.

Central European Journal of Economic Modelling and Econometrics

Modelling Recovery Rate for Incomplete Defaults Using Time Varying Predictors

Wojciech Starosta*

Submitted: 16.01.2020, Accepted: 10.04.2020

Abstract

The Internal Rating Based (IRB) approach requires that financial institutions estimate the Loss Given Default (LGD) parameter not only based on closed defaults but also considering partial recoveries from incomplete workouts. This is one of the key issues in preparing bias-free samples, as there is a need to estimate the remaining part of the recovery for incomplete defaults before including them in the modeling process. In this paper, a new approach is proposed, where parametric and non-parametric methods are presented to estimate the remaining part of the recovery for incomplete defaults, in predefined intervals concerning sample selection bias. Additionally it is shown that recoveries are driven by different set of characteristics when default is aging. As an example, a study of major Polish bank is presented, where regression tree outperforms other methods in the secured products segment, and fractional regression provides the best results for non-secured ones.

Keywords: LGD, workout approach, incomplete defaults, partial recovery rate

JEL Classification: C51, G32

*University of Lodz, Poland; e-mail: w.starosta@wp.pl; ORCID: 0000-0002-2306-0263

1 Introduction

Basel II regulations on the Advanced Internal Rating Based approach permit financial institutions calculate three risk parameters (Probability of Default - PD, Exposure at Default - EAD, and Loss Given Default - LGD) in-house. Simultaneously with this option, minimal technical standards and guidelines concerning estimation have been described (Basel Committee on Banking Supervision 2017). Among them, four methods of LGD calculation can be found. The first and, at the same time, the most popular practice is "workout approach" (Basel Committee on Banking Supervision 2005, p. 4), which is based on discounting cash flows up to the moment of default in reference to the amount of exposure from the same date. The second technique is the implied historical LGD, based on the experience of total losses and PD estimates. The third and fourth methods are market LGD, based on the prices of traded defaulted loans, and implied market LGD, which is derived from non-defaulted bond prices by means of an asset pricing model (Basel Committee on Banking Supervision 2005, p. 12).

Due to the quality of estimates, the workout approach is preferred both by supervisors and in the literature (Basel Committee on Banking Supervision 2017, p. 114 and Anolli, Becalli, Giordani 2013, p. 92). However, for the sake of a complicated way of defining and calculating the mentioned recovery amounts, as well as determining the exposure at the moment of default, the workout is governed by a non-standard number of guidelines. One of them states that it is essential to take into account all observed defaults from the selected period (Basel Committee on Banking Supervision 2017, p. 34). Such a period should cover as broad information as possible so that the financial institution can reflect the current debt collection process and policies in the LGD model. Taking into consideration that debt recovery can last for several years, in the selected sample there are cases where the process has started but not yet finished at the moment of model preparation (so called open or incomplete default). It leads to the state in which the value of a dependent variable is not known for part of the observations, which is a consequence of its definition, usually referred to as the recovery rate (RR):

$$RR = \frac{\sum_{t=1}^{n} CF_t/(1+d)^t}{EAD}, \tag{1}$$

wherein the nominator sum of discounted cash flows is located and the denominator contains Exposure at Default (Anolli, Becalli, Giordani 2013, p. 92). The need of taking all defaults from selected period is problematic in cases where the final value of the nominator is not known due to the open debt collection process. Even if regulatory issues (Article 181(1)(a) of the Capital Requirements Regulation (CRR)) were not present, including only completed workouts would be not representative for the modeled parameter, and also unjustified bias would be introduced connected to the omitted cases. As stated by Rapisarda and Echeverry (2013), profiles of closed and open defaults can result in different LGD, so properly reflecting such situation lead

to more reliable estimates. In particular using only resolved cases in building LGD model introduce downward bias as more short-lasted high-RR cases would be taken into sample. On the other hand using unresolved cases as-is, end with upward bias, as unresolved cases will on average have higher final RR as observed at the moment of model preparation. In this paper we present a method of inclusion the unresolved cases, using the estimate of the remaining part of RR, which will be realized in future, to the resolved part of the sample. This leads to non-biased sample, which produces non-biased LGD estimates. The more reliable are the results of partial RR estimation, the more precise the final LGD output is, as then it possess all the patterns observed during the historical period used in the model preparation.

Our first contribution is a time and collateral dependent sample preparation, which aims to reduce bias connected primarily with the occurrence of different recovery patterns in closed and open defaults samples. Direct estimation from closed cases may lead to downward estimation bias. This is due to the fact that, among completed cases, there are usually more relatively short ones which ended with full recovery. On the other hand, among open defaults, reverse dependence is possible, so cases that are in default status for a long time with a low recovery rate may prevail. To solve this issue, we separately estimate partial recovery rate models in pre-defined sub-samples to reflect inherent features of each. We split the sample by the time in default such that different variables drives the recovery of 3-month default opposite to 30-month default. What is more, we differentiate the state before and after collateral realization for secured credits to include the change in client recovery pattern, when tangible asset is lost. The second contribution is related to the potential superiority of non-parametric methods in estimating the partial recovery rate over parametric ones in terms of the precision of the estimates given. In Dermine and Neto (2006) or Bastos (2010) additive or multiplicative version of the Kaplan-Meier estimator was used to incorporate unresolved cases, as time in default and marginal recoveries were used to estimation. Our idea is to build a different parametric and non-parametric models, potentially using various predictors which should address the problem of non-linearity between dependent and independent variables. Such methods are widely used in the LGD modeling, but according to the author's knowledge will be used for the first time in the partial recoveries estimation. The conservativeness of the estimates could be easily obtained in each solution (which is one of the major assumptions concerning Basel II/III regulations), so the presented approach can be treated as part of the discussion about the upcoming adjustments in Basel IV.

The process of estimating the recovery rate is held via fractional regression, beta regression, regression trees and support vector machines, which are gaining more and more popularity in both academic and business applications, as an alternative to the canonical regression methods. All four approaches were previously used in the LGD estimation, so we adopt them to predict the partial recovery rates not coincidentally. The ultimate goal is to prepare a sample containing all defaults together with an estimation of the remaining part of the recovery rate for open cases. This leads to

more precise LGD estimates than those resulting from the estimation only based on the sample with closed cases. In addition, performance of the methods is checked on out-of-time data, which refers to defaults that are not closed at the moment of estimation, but their realized value is already known in the validation set.
The structure of the paper is as follows. First, a review of the existing literature both on the subject of overall LGD estimations, as well as those studies where the problem of open cases is raised is presented. The second section discusses the sample preparation method to take into account the problem of the different statuses of open cases. The third section contains a brief description of the methods used in the estimation. The fourth section demonstrates the conducted study on the training sample carried out in 2015 and check the effectiveness of the methods on the out-of-sample data from 2017. Finally, a summary of the results is presented along with a suggested direction for further research.

2 Literature review

With the appearance of the settlements enclosed in the New Basel Capital Accord (Basel II), the interest in modeling credit risk parameters both among practitioners and in the academic environment increased dramatically. Although the approach to each of them has been standardized since 2004, methods that are often a combination of techniques previously described, or the application in a particular area of solutions known from other fields, are still being developed. In the LGD parameter modeling as classical methods averaging in pools (Izzi, Oricchio, Vitale 2012), linear regression (Anolli, Becalli, Giordani 2013) and beta regression (Huang and Oosterlee 2011) can be considered. These are also the methods most preferred by supervisors as well recognized both in theoretical and interpretative terms. However, it is necessary to notice the shift to more complex or even non-parametric methods, often inadequately referred to as "black boxes". Some of the most interesting proposals have been included in works of Belotti and Crook (2007, decision trees), Luo and Shevchenko (2013, Markov Chains), Brown (2012, neural networks and two-staged models) and Siddiqi and Van Berkel (2012, scoring based methods usage).
However, the literature mentioned above in most cases does not discuss the subject of the inadequacy of the sample; the modeling process begins when the dependent variable is already completely prepared. The subject of open cases was initially discussed in the paper by Dermine and Neto (2006), where the actuarial-based mortality approach with the Kaplan-Meier estimator was used to determine the recovery rate. Initially, Marginal Recovery Rates (MRR) in period t were determined as cash flow paid at the end of period t divided by loan outstanding at time t. Secondly, PULB (Percentage Unpaid Loan Balance at the end of period t) was calculated as $1 - MRR$, and finally, Cumulative Recovery Rate T periods after the default was recognized as $1 - \prod_{t=1}^{T} PULB_t$. By using both completed and open cases, recovery rate curves and exposure-weighted recovery rate curves for each period t

were determined. A similar approach was used in Bastos (2010) with its explication in Rapisarda and Echeverry (2013), where a reformulation from the exposure-weighted Kaplan-Meier estimator to a default-weighted one was shown. This is viewed as more appropriate to ensure compliance with supervisor guidelines. A second change was the transition from the aggregation of recovery rates over time and then across exposure, to aggregation recovery rates across exposure and then over time. The difference is situated in the statement that *in the first case, ultimate recovery rates must be realizations of the same random variable whereas in the second recovery, profiles need to be realizations of the same stochastic process* (Rapisarda and Echeverry 2013, p.1). Finally, the authors show distributions of recovery rates over time, which leads to more precise LGD estimators than those based only on completed cases. An overview of methods like the use of external databases, time criteria or the extrapolation of future recoveries was described in Zięba (2017), where it was stated that extrapolation gives the best results, both in terms of increasing the sample size and the impact of the final LGD estimators. The most conservative approach has been presented in Baesens, Roesch and Scheule (2016), where one of the proposals is to take account of incomplete cases as if they were completed; however, it may lead to a revaluation of the final LGD values. On the regulatory side, precise assumptions regarding the treatment of open cases should appear together with the records of Basel IV (Nielsen and Roth 2017, p. 72).
One of the aims of this study is to extend the existing literature with further methods of estimating recovery rates for open cases, which is also in line with upcoming regulations. Additionally, an attempt is made to reduce bias coming from the possibility of differences in populations of open and closed defaults and the potential revaluation of recovery rates based only on closed cases. Finally, described approach is validated on out-of-time data.

3 A bias free sample design

This section provides an overview of the sample preparation process. The nature of the recovery rate imposes at least three states in which an exposure with the premise of default can be found.

Figure 1: Closed default. All recoveries were obtained before the reference date

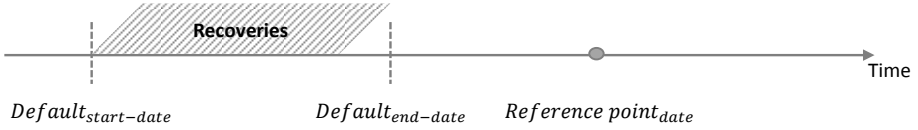


Figure 1 illustrates a standard example in which the final recovery rate is known, regardless of where the recoveries come from. Figures 2 and 3 demonstrates the

Figure 2: Incomplete default with collateral realization before the reference date. It is still possible to obtain recoveries from the client's own payments

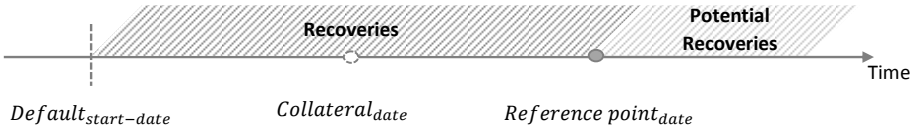
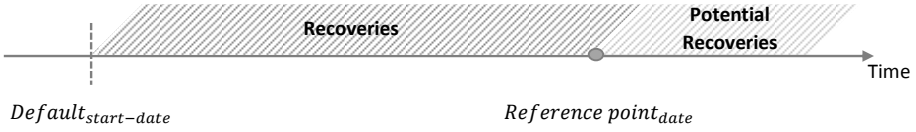


Figure 3: Incomplete default without collateral realization before the reference date. It is still possible to obtain recoveries both from the client's own payments and collateral realization in the case of a secured product



situations where the recovery process has not been finalized and it is necessary to estimate the remaining part of the recovery rate. At this stage, it seems essential to distinguish secured exposures (e.g. by mortgage or vehicle), for which the process differs radically before and after collateral realization. Before realization, the recovery rate consists of the consumer's own payments and a theoretically possible repayment from the collateral; after realization, only the consumer's own payments are possible, but their motivation is significantly different from before. The recovery rate, taking into account the above distinction, is calculated as follows:

$$RR = \begin{cases} RR_{pay} + RR_{coll}, & \text{for closed default (Fig. 1),} \\ RR_{pay} + RR_{coll} + \widehat{RR}_{pay}, & \text{for open default with collateral realization (Fig. 2),} \\ RR_{pay} + \widehat{RR}_{pay} + \widehat{RR}_{coll}, & \text{for open default without collateral realization (Fig. 3),} \end{cases} \tag{2}$$

where:
RR – recovery rate, as the dependent variable in the LGD model,
RR_{pay} – actual value of the recovery rate from the client's own payments,
RR_{coll} – actual value of the recovery rate from collateral realization,
$\widehat{RR}_{pay}$ – predicted value of the partial recovery rate from the client's own payments for the period from the reference point to the end of recovery process,

$\widehat{RR}_{coll}$ – predicted value of partial recovery rate from collateral realization for the period from the reference point to the end of recovery process.
The reference point is understood as the date from which the data originate. The actual values come from recoveries obtained before this date. The predicted values are values estimated for the period from the reference point till the end of the recovery process (it is not defined as a time period, rather any point in the future when the process will finish). And although the reference date is the same for all cases in the sample, for incomplete ones, the period from the moment of default till the reference date is different. This is key information in the recovery process, because the estimated recovery rate will be different for cases in which the default occurred a month before the reference date to the cases where the default occurred five years before the reference date. Therefore, for the needs of estimation, both closed and open cases should be divided into sub-periods in which the estimation of parameters will take place. The more granular the period selected, the more accurate the possible results will be; however, excessive fragmentation may lead to instability of estimates, as fewer and fewer observations will be involved in subsequent intervals.
Taking into consideration the remarks above, the recovery rate formula for open cases can be transformed in a manner that depends on the time in default and the collateral realization:

$$RR = \frac{\sum_{t=1}^{l} CFpay_t/(1+d)^t}{EAD} + \frac{\sum_{t=1}^{l} CFcoll_t/(1+d)^t}{EAD} + \widehat{RR}_{pay}^{l+1} + \widehat{RR}_{coll}^{l+1}, \quad (3)$$

where:
$CFpay_t$ – cash flows from own payments up to the reference date carried out in period t,
EAD – exposure at default,
l – the number of periods from the date of the default to the reference date,
$CFcoll_t$ – cash flows from the collateral realization up to the reference date carried out in period t,
$\widehat{RR}_{pay}^{l+1}$ – estimated value of the partial recovery rate from own payments from the moment $l+1$ until the end of recovery process,
$\widehat{RR}_{coll}^{l+1}$ – estimated value of the partial recovery rate from the collateral realization from the moment $l+1$ until the end of recovery process in cases where the collateral realization has not yet taken place.

This method of recovery rate construction is free from the bias caused by the selection of the sample, as it contains appropriate patterns both for complete and open cases.

At this point, it is possible to determine a way to estimate $\widehat{RR}_{pay}^{l+1}$ and $\widehat{RR}_{coll}^{l+1}$. At each time interval, the actual recovery rates are calculated from the start time of the interval (m) to the end of recovery process window (n) on the basis of complete cases.

$$RR_{pay}^{m} = \sum_{t=m}^{n} \frac{CFpay_t}{EAD(1+d)^t}. \quad (4)$$

The result of this equation is population divided into sub-samples consists of cases which lived long enough to be a part of each. Taking 6-months intervals as an example, we can see that all cases are used to determine $\widehat{RR}_{pay}^{l+1}$ for defaults being in interval from 0 till 6 month, but only defaults which lived at least till month 60 are used to estimate $\widehat{RR}_{pay}^{l+1}$ for open cases being in interval from 60 till 66 month. The recovery rate for $\widehat{RR}_{coll}^{m}$ is calculated analogously, where m is time interval for which the variable value is calculated. For example for 6-months periods, sum of recoveries from the beginning of default to the end of recovery process is determined first. The second period runs from the sixth month of recovery until the end of recovery process, and so on. Such a construction allows us to create a set, on the basis of which it is possible to estimate the partial recovery rate for each open case depending on: (i) time in default, (ii) hitherto recovery from own payments, and (iii) recovery from collateral realization.

4 Recovery rate estimation methods for open cases

The following section presents a brief summary of the methods used in recovery rates modeling, which in our study are used in the process of partial recovery rate estimation and is divided into two sub-sections corresponding to the groups convergent in terms of theoretical assumptions. The first category consists of parametric methods in which fractional regression and beta regression are presented. The second one contains regression trees and support vector machines.

4.1 Parametric methods

The first method discussed in this subsection is fractional regression (FR). Its use for LGD modeling was proven to give reasonable results, inter alia, in Belotti and Crook (2009) or Bastos (2010). Detailed assumptions about this type of regression can be found in Papke and Woolridge (1996). For the problem of estimating recovery rates, a lack of assumptions about the distribution is crucial; only the conditional mean must be correctly specified in order to obtain consistent estimators. Assuming that

$$E(y_i \mid \mathbf{x_i}) = G(\mathbf{x_i}\boldsymbol{\beta}) = 1/[1 + \exp(-\mathbf{x_i}\boldsymbol{\beta})] \tag{5}$$

the fractional logit model parameters $\hat{\boldsymbol{\beta}}$ can be estimated by maximizing the Bernoulli log-likelihood function (as in binary logistic regression) (Papke and Woolridge 1996, p. 621):

$$L(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} y_i \log[G(\mathbf{x_i}\hat{\boldsymbol{\beta}})] + (1 - y_i) \log[1 - G(\mathbf{x_i}\hat{\boldsymbol{\beta}})], \tag{6}$$

where $i = 1, ..., n$, n is a sample size and $\mathbf{x_i}$ is a vector of explanatory variables for case i. However, it should be noted that the explained variable must come from a

specific range ($0 \leq y_i \leq 1$), which is not always ensured in the case of recovery rate modeling (like where direct and indirect costs were added or collateral was sold at price higher than EAD). The solution is to apply a linear transformation in the form of classical unitarization:

$$\widetilde{RR_i} = \frac{RR_i - \min\limits_i\{RR_i\}}{\max\limits_i\{RR_i\} - \min\limits_i\{RR_i\}}. \tag{7}$$

As a result of the above-mentioned normalization formula, the obtained transformed recovery rates $\widetilde{RR_i}$ belong to the interval $[0; 1]$. Backward transformation is done during out-of-sample verification.

The second method, which is gaining more and more popularity in LGD estimation, is Beta Regression (BR). Besides the publications mentioned in Section 1, it can be found in Chalupka and Kopecsni (2008), Stoyanov (2009) or Tong, Mues, and Thomas (2013). What makes Beta Regression so popular is its flexibility in the case of modeling quantities constrained in the interval $(0; 1)$. Depending on the choice of parameters, the probability density function can be unimodal, U-shaped, J-shaped or uniform:

$$f(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \tag{8}$$

where $\Gamma(\cdot)$ denotes the gamma function. It is assumed that $\alpha > 0$ and $\beta > 0$. In such a formulation, α pushes the density toward 0 and β toward 1. Without loss of generality, these two parameters can be reformulated in terms of mean (μ) and dispersion (assuming $\varphi = \alpha + \beta$) in the following way (Huang and Oosterlee 2011):

$$\alpha = \mu\varphi, \qquad \beta = (1-\mu)\varphi. \tag{9}$$

Within the framework of Generalized Linear Models (GLM), both μ and φ can be modeled separately, with a location model for μ and a dispersion model for φ, using two different or identical sets of covariates (Liu and Xin 2014). The mean model can be expressed as:

$$g(\mu) = \gamma_0 + \sum_i \gamma_i a_i, \tag{10}$$

where a_i denotes explanatory variables, γ_i coefficients and g is the monotonic, differentiable link function. Since the expected mean μ is bounded by 0 and 1, logit can be used as the link function:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right). \tag{11}$$

Dispersion parameter φ can be treated as fixed or it can be modeled by another GLM (Huang, Oosterlee 2011):

$$h(\varphi) = \zeta_0 + \sum_i \zeta_i a_i, \tag{12}$$

where h is a link function and ζ_i are coefficients. The simplest way to achieve it is to use:

$$\varphi = e^{\zeta_0 + \sum \zeta_i a_i}. \tag{13}$$

4.2 Non-parametric methods

Tree-based methods (RT) recursively partition the original sample into smaller subsamples and then fit a model in each one. The concept is clear and easy to implement, yet the method is powerful and was adopted for LGD purposes inter alia in Qi and Zhao (2011) or Van Berkel and Siddiqi (2012). To build a tree, an algorithm is needed which, at each node t, evaluates the set of variable splits to find the best one, i.e., the split s that maximizes the decrease in impurity (im) (Brown 2012, p.51):

$$\Delta im(s,t) = im(t) - p_L im(t_L) - p_R im(t_R), \tag{14}$$

where p_L and p_R denote the proportion of observations associated with node t that are sent to the left child node t_L or to the right child node t_R. In the case of a continuous variable, like a recovery rate, regression trees are used and a standard criterion for this type of model is minimizing the sum of squares $\sum (y_i - \hat{y}_l)^2$, which leads to averaging recovery rate in region R_m as the value of each leaf:

$$\hat{c}_m = avg(y_i \mid x_i \in R_m). \tag{15}$$

Finding the best partition is quite straightforward. First, splitting variable j and split point s are selected, so a pair of half-planes can be defined:

$$R_1(j,s) = \{X \mid X_j \leq s\} \text{ and } R_2(j,s) = \{X \mid X_j > s\}. \tag{16}$$

The second splitting variable j and split point s are searched to solve:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]. \tag{17}$$

For the choice of j and s, the inner minimization is solved by:

$$\hat{c}_1 = avg(y_i \mid x_i \in R_1(j,s)) \text{ and } \hat{c}_2 = avg(y_i \mid x_i \in R_2(j,s)). \tag{18}$$

After the first split is determined, the procedure is repeated on all regions (Hastie, Tibshirani, and Friedman 2008, p. 307). The question arises when one should stop growing each tree. This is another advantage of the described approach, as there are many elegant methods to achieve this:

1. establishing a minimal impurity decrease,
2. fixing the maximal depth,

3. selecting the minimal number of observation in a leaf.

These are also the most common methods of solving the instability issue, which is often raised when tree-based models are used. The lack of estimates smoothness can be considered as another drawback, as it can deteriorate performance in the regression setting, where underlying function is expected to be smooth (Hastie, Tibshirani and Friedman 2008). However in the case of partial recovery rate estimation, it is not an issue, because we can define each region as different recovery pattern (specific scenario which leads to particular value of partial RR).
The Support Vector Machine (SVM) is another non-parametric technique for classification and regression problems used in LGD modeling more and more frequently (see Loterman et al., 2012 or Yao, Crook, and Andreeva 2017). It produces nonlinear boundaries by constructing a linear boundary in the transformed version of the feature space. Formally, an SVM constructs a hyperplane or set of hyperplanes in a potentially infinite dimensional space. The SVM finds this hyperplane using support vectors and margins (defined by support vectors). In a regression model (Hastie, Tibshirani, and Friedman 2008, p. 434):

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x), \tag{19}$$

where $h_m(x)$ is a set of basis functions (by which we denote a function that augments vector of $\mathbf{X}$ by additional variables via selected transformation, like $h_m(x) = x_j x_k$ or $h_m(x) = \log(x_j)$) and $m = 1, 2, ..., M$, the goal is to minimize:

$$L(\beta) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \|\beta\|^2 \tag{20}$$

for some general error measure $V(r)$. Regardless of $V(r)$ the solution of has the form:

$$\hat{f}(x) = \sum_{i}^{n} \hat{\alpha}_i K(x, x_i), \tag{21}$$

where $K(x, y) = \sum_{m=1}^{M} h_m(x) h_m(y)$ and it denotes specific kernel. This allows SVM to easily capture non-linear dependencies by using different kernel function. There are many possible kernels, but in this study, the radial one is used with squared Euclidean distance:

$$K(x, x') = e^{-\|x - x'\|^2 / 2\sigma^2}. \tag{22}$$

5 Empirical analysis of partial Recovery Rates

In this section, an attempt to estimate the partial recovery rate is made using data from one of the largest Polish banks applying the AIRB regime. A sample of completed defaults from 2003 to 2015 is used for models preparation. These models then predicts the recovery rate for open cases from the same time period, and finally, goodness of fit is checked on a part of the sample where the recovery process finished during the 2015 – 2017 period. The process of parameter estimation is conducted in 6-months intervals, so the first interval predicts the final recovery rate for cases whose default lasted from 0 to 5 months, the second from 6 to 11 months, etc. We assume that such a split is granular enough and allows to prepare stable models in each interval. In presented models part of the recovery connected with the collateral ($\widehat{RR}_{coll}$) is included via Loan To Value (LTV) variable calculated at each point after default. To reflect both possibilities drawn on Figure 2 and Figure 3, its construction is as follows:

$$LTV_l = \begin{cases} \dfrac{\text{Loan value}_l}{\text{Collateral value}_l}, & \text{if collateral was not sold in selected interval,} \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

Additionally, we benchmarked our models to simple Naïve Markov chain in the form of transition matrix (cf. Jarrow et al., 1997). We divided partial recoveries into classes, taking into account only months since default, and estimate the final class according to the initial class for each case. It is an equivalent of "mean prediction", frequently treated as a benchmark to more sophisticated methods or recovery rates estimation.

5.1 Sample description

As mentioned above, the sample is made up of default events which occurred between 2003 and 2015 and contains both secured (ML) and non-secured loans (NML). The predictors were obtained at the moment of default and then at each point respectively. This allows us to show the dynamic nature of characteristics during default window and access variables specific to this stage of the process (like DPD or due amounts). The proportion of completed and open cases is presented in Table 1. It clearly demonstrates that this specific portfolio suffers from a huge share of open defaults. Taking into account only the completed ones would lead to the removal of 46.53% of secured and 62.15% of unsecured contracts, so there is no doubt that data selection bias would be introduced. What is more, there is a significant difference in the distribution of explanatory variables, as shown in Tables 2 and 3. This may cause another problem with data representativeness.
The variables RR^m_{pay} and RR^m_{coll} are prepared according to formulas from Section 2 in 6-month time intervals, and consist of both principal and interest recoveries. So, each point in Figure 4 is a mean recovery rate from the beginning of the interval till the end of the recovery process.

Table 1: Proportion of closed cases in the sample by type of credit

	Closed	Open	All
Secured	53.47%	46.53%	6 953
Non-secured	37.85%	62.15%	122 353

Table 2: Descriptive statistics for secured credits by label

Variable	Label	Mean	5^{th} Pctl	25^{th} Pctl	50^{th} Pctl	75^{th} Pctl	95^{th} Pctl	Max
EAD	Closed	318k	38k	105k	207k	394k	942k	6.598k
	Open	421k	60k	161k	297k	509k	1.167k	9.505k
Interest rate	Closed	0.042	0.009	0.025	0.040	0.054	0.090	0.128
	Open	0.037	0.009	0.013	0.034	0.050	0.089	0.160
Days past due (DPD)	Closed	46.00	0	0	33	91	92	443
	Open	49.20	0	13	46	91	91	1681
Tenor	Closed	294	120	239	336	359	360	360
	Open	306	155	240	358	359	360	360
Requested amount	Closed	359k	54k	123k	236k	438k	1015k	7617k
	Open	465k	78k	184k	330k	561k	1281k	9977k
Months on book (MOB)	Closed	42.99	8	22	39	60	93	144
	Open	50.19	12	31	47	68	97	143
Due principal	Closed	3.4k	0	0	648	1843	7755	893k
	Open	4.5k	0	322	1254	3006	11k	722k
Due interest	Closed	1.6k	0	0	584	1592	6217	109k
	Open	2.8k	0	210	891	2192	8k	135k
Principal	Closed	321k	40k	106k	208k	340k	945k	7209k
	Open	427k	60k	162k	302k	514k	1189k	12354k
Interest	Closed	2.1k	23	347	868	2056	7482	130k
	Open	2.8k	70	490	1214	2806	9662	172k
Due amount	Closed	5.1k	0	0	1442	3709	14k	910k
	Open	6.7k	0	782	2476	5531	19k	722k
LTV	Closed	0.95	0.03	0.27	0.69	1.00	1.06	1.63
	Open	1.06	0.05	0.30	0.76	1.00	1.32	1.78
Foreign currency	Closed	0.77	0	1	1	1	1	1
	Open	0.77	0	1	1	1	1	1

It can be seen that recoveries from the consumer's own payments decrease over time, which seems reasonable, as client motivation to repay diminishes with duration of default and longer defaults are seen as more problematic (poor financial situation, difficulties with reaching the customer, client goes into litigation, etc.). Also, recoveries from secured loans are greater than non-secured ones, which indeed is consistent with the findings from the previous studies (see e.g., Gurtler and Hibbeln 2013), as clients care more about losing their home or car as a consequence of a default. Finally, the shape of the collateral RR curve results from the more discrete

Table 3: Descriptive statistics for non-secured credits by label

Variable	Label	Mean	5^{th} Pctl	25^{th} Pctl	50^{th} Pctl	75^{th} Pctl	95^{th} Pctl	Max
EAD	Closed	7.1k	85	2.4k	4k	5.8k	23k	808k
	Open	12k	910	3.2k	5.5k	11k	47k	243k
Interest rate	Closed	0.184	0.110	0.160	0.200	0.210	0.227	0.662
	Open	0.168	0.098	0.144	0.169	0.200	0.230	0.590
Days past due (DPD)	Closed	126.4	25	91	91	92	474	3132
	Open	75.68	0	43	88	91	123	2201
Tenor	Closed	16.06	12	12	12	12	60	120
	Open	24.6	12	12	12	36	60	156
Requested amount	Closed	8.1k	1.2k	3k	5k	6.9k	27k	1000k
	Open	13.7k	1.3k	3.5k	5.6k	13k	50k	2400k
Months on book (MOB)	Closed	22.36	5	9	16	30	58	138
	Open	29.18	5	12	23	40	74	154
Due principal	Closed	1.9k	1.92	431	623	1.1k	5.9k	800k
	Open	1.6k	0	138	523	961	4.2k	425k
Due interest	Closed	196	0.01	26	100	204	650	17k
	Open	297	0	39	128	304	1.2k	43k
Principal	Closed	6.9k	73	2.3k	3.8k	5.7k	23k	800k
	Open	12k	889	3.1k	5.2k	11k	46k	2399k
Interest	Closed	267	2.66	69	164	301	792	19k
	Open	394	19	90	204	414	1.4k	43k
Due amount	Closed	2.1k	26	526	746	1.4k	6.3k	808k
	Open	1.9k	1.33	327	682	1.3k	5.2k	452k

construction of the process where collateral realization can happen (in general) once in the default window, in contrast to RR from the client's own payments, where the client can repay the due amount using more than one transaction. The structure of the sample in division by payers, non-payers and partial payers is shown in Figure 5. Diminishing number of payers along with relatively stable share of non-payers, in great extend supports the conclusions drawn from analyzing the mean recovery rate curves.

5.2 Models

We estimate RR till month in default equal to 60, as a result of sharp observation number decrease after this interval. This effects with assigning values from the 60m interval to every observation with months in default greater than 60, but no greater than 96, when the values of average recoveries are set to zero. This is in line with the general assumption that after a certain time, financial institutions no longer expect any repayments (Basel Committee on Banking Supervision 2017, p. 34). Following Section 2, twelve models for each method are prepared based on static and dynamic variables presented in Tables 2 and 3. Point estimates are provided in the Appendix.

Figure 4: Mean of the recovery rate at each point till the end of recovery process for cases which lived in particular interval

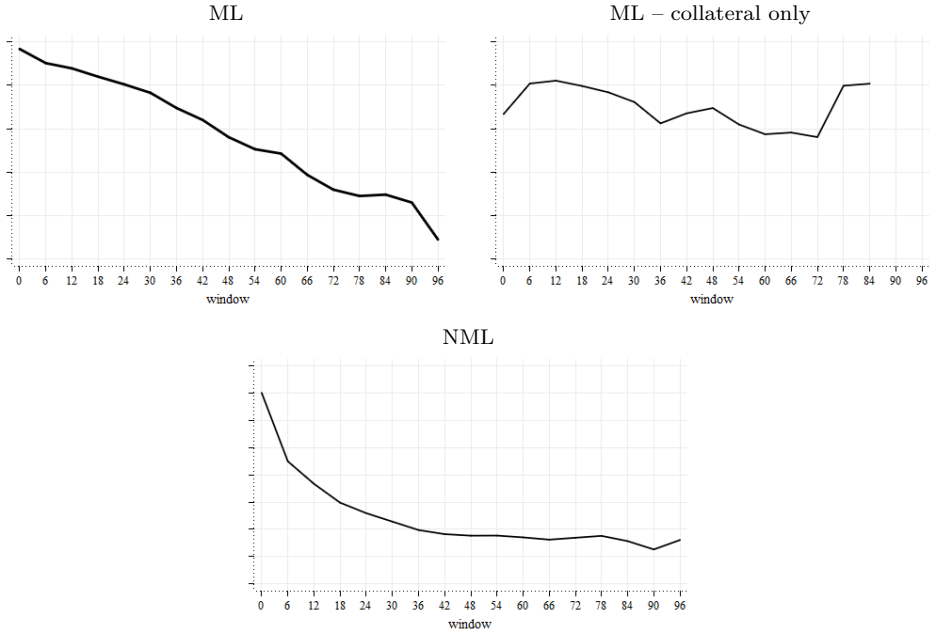
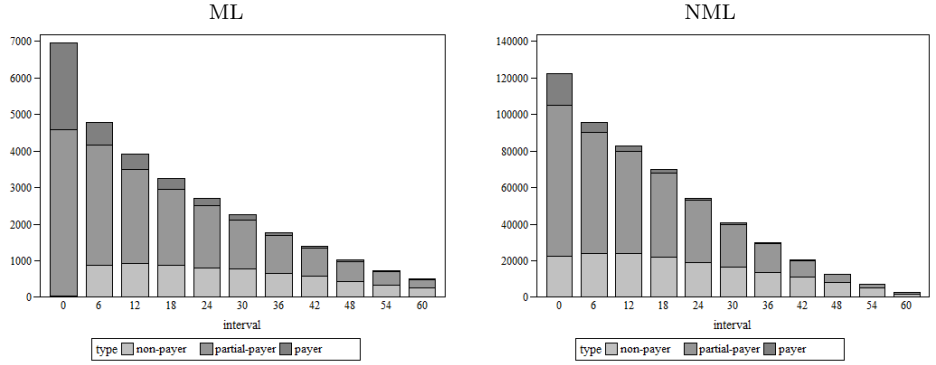


Figure 5: Number of observations in consecutive intervals in division by payers, non-payers and partial payers



Each column shows the information from the beginning of the interval till the end of recovery process

Fractional regression

Due to highly correlated variables in our data set, we use L1 criterion for regularization scheme to select the best set of predictors. Tables 4 and 5 summarizes the results in the case of variables used, RMSE and the selected correlation measures (Pearson and Spearman coefficients).

Table 4: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Fractional Regression – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√										
DPD	√	√	√	√	√	√	√	√	√	√	√
TENOR											
REQ. AMOUNT											
MOB	√	√	√	√				√			
DUE PRINCIPAL											
DUE INTEREST											
PRINCIPAL											
INTEREST											
DUE AMOUNT											
LTV								√	√	√	√
FOREIGN CURRENCY											
RMSE	.1329	.1795	.1879	.1962	.2265	.2419	.2612	.2592	.2998	.2649	.2673
PEARSON	.2388	.3207	.3988	.4692	.4418	.4761	.5418	.6097	.5281	.5740	.6532
SPEARMAN	.1956	.1084	.1547	.2157	.1758	.2052	.3490	.4653	.4411	.5023	.5680

First conclusion, that we can draw from Table 4 and Table 5, consist in recovery pattern changes observable across time in default. The only variable significantly important in all regressions is DPD, which means that along with the increase of days-past-due partial recovery rate is decreasing over time. But as time in default rise, we can see switch from contract based variables (interest rate, months on book) to LTV. This is definitely something worth to examine. When client goes into default, at the beginning his recoveries consist mainly of own payments (RR_{pay}) and majority of them deal with debt with their own strengths. But if default last for more than 42

Table 5: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Fractional Regression – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√		√	√	√	√	√	√	√		
DPD				√	√					√	√
TENOR	√	√									
REQ. AMOUNT											
MOB	√	√	√	√	√	√	√	√	√	√	
DUE PRINCIPAL											
DUE INTEREST											
PRINCIPAL							√	√	√	√	
INTEREST											
DUE AMOUNT											
RMSE	.2995	.2883	.2682	.1750	.2345	.2215	.2139	.2120	.2266	.2422	.2385
PEARSON	.3054	.3098	.2664	.2563	.2341	.2089	.2026	.2157	.2092	.1981	.2432
SPEARMAN	.2303	.2824	.2321	.1521	.1154	.1046	.1079	.1613	.2076	.3064	.3783

months, then capability to repay worsens and the collateral is used more frequently to compensate the remaining part of the debt. Collateral realization can introduce non-linearity into the model, which could not be easily captured by fractional regression and can be the reason for RMSE rising in latter intervals. It is also an unique characteristic of secured credits, as being in default for more and more months usually leads to involving court, bailiff or consumer bankruptcy. This events are not efficiently modeled by contract characteristics only or even if so, then non-linear approach to each case should be handled by different method, which is able to produce more robust estimates for this intervals.

Non-secured credits behave similarly when it comes to recovery pattern change. At the beginning we can see that static variables, like interest rate or tenor, are used more frequently to estimate partial recovery rate. But in closing intervals DPD, months on book and principal are of greater importance. We can conclude that at the beginning it is difficult to predict partial recovery rate as similar contracts are driven by the same characteristics to different RR levels, which is confirmed by higher RMSE in initial intervals. After that stronger patterns appear derived by DPD and MOB mainly and RMSE decreases (opposite to secured loans). The reason for this could also be situated in specific collection department policy, which could be the part of the process from some point (after DPD threshold exceeded for example).

Beta regression

Due to highly correlated variables in our data set, we use L1 criterion for regularization scheme to select the best set of predictors. Tables 6 and 7 summarizes the results in the case of variables used, RMSE and the correlation parameters.

Table 6: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Beta Regression – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√									√	
DPD	√	√	√	√	√	√	√	√	√	√	√
TENOR										√	
REQ. AMOUNT								√	√		
MOB	√	√	√	√	√	√	√				
DUE PRINCIPAL											
DUE INTEREST											
PRINCIPAL											
INTEREST											√
DUE AMOUNT											
LTV			√					√	√		
FOREIGN CURRENCY								√	√		
RMSE	.1415	.1797	.1875	.1983	.2268	.2442	.2683	.2733	.3003	.2817	.2793
PEARSON	.2150	.3231	.4082	.4680	.4485	.4862	.5247	.5779	.5591	.5089	.6331
SPEARMAN	.2139	.1062	.1685	.2083	.2238	.2840	.4086	.5998	.5241	.4856	.5383

Beta regression is able to find more relationships with predictors than fractional regression but it did not translate into better results on average. What is interesting is a fact, that in BR collateral is important in one of the first stages, then this importance is lost for a while, but finally like in FR it is main driver along with DPD when it comes to secured loans. RMSE rises with time, which can be the result of high complexity of defaults lasting years in default (like in FR).
An opposite arises with NML loans, where the biggest errors are observed again at the beginning of the default. Here, motivation to repay is significantly different from secured loans, so finding proper patterns in the data seems to be harder for the first

Table 7: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Beta Regression – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√	√	√	√	√	√	√	√	√	√	√
DPD	√				√					√	
TENOR	√	√	√							√	√
REQ. AMOUNT									√	√	√
MOB	√	√	√	√	√	√	√	√	√		
DUE PRINCIPAL											
DUE INTEREST		√	√								
PRINCIPAL						√	√	√	√	√	√
INTEREST	√										
DUE AMOUNT											
RMSE	.3048	.3048	.2865	.1842	.2540	.2411	.2361	.2356	.2507	.2634	.2592
PEARSON	.3118	.3171	.2796	.2401	.2186	.2051	.1986	.2123	.2287	.2381	.2262
SPEARMAN	.2736	.2955	.2405	.1635	.1396	.1094	.0990	.1599	.2996	.3582	.4856

year in default. Then there is a meaningful decrease in error for month 18, which may suggest that the debt collection policy might result in write-offs or termination at that time, and the model captures it. For month 24 and later, the RMSE is quite stable, mainly due to the fact that there is no factor of collateral, so only customers' own payments are modeled.

Regression Trees

As stated in Section 4.2, parametrization needs to be made to build a tree. For the sake of the results comparison, we decide to select the same parameters for every tree, which are ANOVA as the splitting selection method (applicable for continuous variables), complexity parameter selected based on a 10-fold cross-validation, 10 as the maximum depth (selected arbitrarily, but this constraint is not binding as no tree grew so deep) and 30 as the minimum observations in a leaf (to get the statistical significance of the mean). Such parametrization allows to avoid overfitting with building precise tree at the same time.

On the sample selected for model building, it can be seen that regression trees give a lower RMSE for latter intervals (compared to fractional and beta regression), which suggests strong non-linearity between the recovery rate and the explanatory variables. Initial intervals are comparable when it comes to error measure, although regression tree is not limited by not correlated variable selection, which can be easily seen when

Table 8: Variables used in particular tree for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Regression Tree – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD	√	√	√	√	√	√			√		
INTEREST RATE	√		√	√	√	√	√	√	√		
DPD	√	√	√	√					√		
TENOR	√	√						√		√	
REQ. AMOUNT	√	√	√	√	√			√			
MOB	√	√	√	√	√	√	√				
DUE PRINCIPAL	√	√			√			√	√		
DUE INTEREST	√	√			√		√	√	√	√	
PRINCIPAL	√		√	√	√			√	√		
INTEREST		√	√	√		√					
DUE AMOUNT	√					√					√
LTV	√	√	√	√				√	√		
FOREIGN CURRENCY											
RMSE	.1268	.1726	.1790	.1870	.2078	.2168	.2373	.1923	.2103	.2202	.2023
PEARSON	.3811	.4846	.5296	.5755	.5676	.6157	.6492	.8087	.8005	.7310	.8021
SPEARMAN	.3199	.2409	.3670	.4567	.4599	.4643	.5586	.7402	.7652	.7114	.7706

one collate Table 8 with Table 6 or Table 9 with Table 7. Regression tree like BR find collateral, expressed in terms of LTV, significant at the beginning and at the end of recovery process.
For non-secured products EAD, interest rate, DPD and MOB are the main drivers, significant in almost all intervals. However there are also variables which differentiate RR at the initial stages of default, like tenor, requested amount or interest. This is also coherent with statement, that each interval's inherent features should be taken into account, when partial RR are estimated.

Support Vector Machines

The final method also requires parametrization; however, in the plain version, only variable classification (continuous) and the kernel (radial) need to be specified. Tables 10 and 11 summarize the results based on these assumptions.
It is particularly clear, that according to SVM, for secured products partial RR is mainly driven by delinquencies (due principal, due interest, due amount). At the

Table 9: Variables used in particular tree for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Regression Tree – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD	√	√	√	√	√	√	√	√	√	√	
INTEREST RATE	√	√	√	√	√	√	√	√	√		
DPD	√	√	√	√	√	√	√	√	√	√	√
TENOR	√	√	√	√	√						
REQ. AMOUNT	√	√									
MOB	√	√	√	√	√	√	√	√	√	√	
DUE PRINCIPAL	√	√	√	√	√	√	√	√			
DUE INTEREST	√	√	√								
PRINCIPAL	√	√	√	√	√	√					√
INTEREST	√	√	√								
DUE AMOUNT							√	√			
RMSE	.3245	.3003	.2783	.2548	.2407	.2304	.2171	.2039	.2116	.2123	.1885
PEARSON	.5004	.5563	.5417	.4953	.4696	.4347	.4281	.5203	.4781	.5113	.6412
SPEARMAN	.3815	.5114	.4845	.4090	.4152	.3981	.4235	.5429	.5763	.6511	.6721

beginning more importance is found in EAD and principal, but finally DPD and requested amount took its place. It seems that when client goes into default the main drivers consist in how much he owe at this point and how much of this exposure is past due. But after some time not going back to performing portfolio, his repaying pattern is more dependent on number of days past due and initial amount as higher amounts are generally harder to repay. RMSE for SVM shows lower values than other methods on average, especially for latter intervals, which may be a good reason to consider ensemble of models (but it is beyond this paper).
The most stable results, when it comes to variable selection, are made by SVM for non-secured products. Due principal find its place in 11 out of 12 intervals, EAD in 10/12, interest and due amount in 9/12. Months on book and principal seems to be more important after some time in default, but there is no clear evidence that some variable is particularly meaningful only at the beginning stages. This can support the fact that finding different patterns for the group of close intervals is crucial, as RMSE for SVM achieve higher levels than for the other methods.

5.3 Out-of-sample verification

Using the data from 2017, compiled from cases marked as open in 2015 and closed in 2017, Table 12 show the results of four considered estimation methods. For each case, the time in default is calculated so that assignment to the proper interval could

Table 10: Variables used in particular model for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (SVM – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD	√	√	√								
INTEREST RATE											
DPD					√		√	√		√	√
TENOR	√										
REQ. AMOUNT						√	√	√	√		
MOB		√	√	√					√	√	
DUE PRINCIPAL	√	√		√	√	√	√	√	√	√	√
DUE INTEREST			√	√	√	√	√		√	√	√
PRINCIPAL	√	√	√								
INTEREST			√	√	√	√			√		√
DUE AMOUNT	√	√		√	√	√	√	√		√	√
LTV								√			
FOREIGN CURRENCY											
RMSE	.1370	.1886	.1869	.1720	.1885	.1920	.2010	.1859	.2059	.1780	.2059
PEARSON	.2541	.4700	.5873	.6761	.6732	.7183	.7662	.8232	.8102	.8399	.8000
SPEARMAN	.3727	.6027	.6493	.6208	.6520	.6861	.7093	.7720	.7552	.8054	.7532

As SVM uses combination of all variables in each interval, five strongest are selected to show meaningful results

be made. Then, the final estimated recovery rate is computed as:

$$\widehat{RR} = \min(RR^l + \widehat{RR}_{pay}^{l+1} + \widehat{RR}_{coll}^{l+1}, 1), \tag{24}$$

where RR^l denotes the recovery rate obtained till the moment of the reference point, which is fixed as 02.2015. We limit the estimated value to 1, to avoid the recovery being higher than the value of the clients' obligations. Table 12 shows the values of RMSE for each method with confidence intervals computed on 100 bootstrapped samples.

The out-of-time predictions shows that for secured credits regression trees seems to capture non-linearity in partial RR modeling with the highest accuracy, but SVM also performs well. Regression trees are supported by its stability, when it comes to comparing RMSE on whole sample and bootstrapped. Fractional regression and beta regression performs significantly worse, as even confidence interval are not overlapping

Table 11: Variables used in particular model for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (SVM – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD		√	√	√	√		√	√	√	√	√
INTEREST RATE	√										
DPD	√						√				
TENOR											
REQ. AMOUNT	√										
MOB						√	√	√	√		√
DUE PRINCIPAL	√	√	√	√	√	√	√	√	√	√	
DUE INTEREST		√	√	√	√	√					√
PRINCIPAL							√	√	√	√	√
INTEREST	√	√	√	√	√	√				√	√
DUE AMOUNT		√	√	√	√	√		√	√	√	
RMSE	.3574	.3180	.2961	.2713	.2574	.2449	.2296	.2258	.2330	.2346	.2334
PEARSON	.4156	.5270	.5129	.4667	.4377	.4220	.4236	.4537	.4460	.4835	.5106
SPEARMAN	.3588	.4891	.4688	.4445	.4973	.5560	.5817	.6158	.6285	.6760	.6993

As SVM uses combination of all variables in each interval, five strongest are selected to show meaningful results

Table 12: RMSE level for consecutive methods on the out-of-sample set. ML denotes secured loans, and NML non-secured. Best measure is underlined

	Method	RMSE	LCLM	RMSE Bootstrap	UCLM
ML	Fractional Regression	0.2361	0.2336	0.2355	0.2375
	Beta Regression	0.2413	0.2389	0.2410	0.2430
	Regression Trees	<u>0.2168</u>	<u>0.2149</u>	<u>0.2168</u>	<u>0.2187</u>
	Support Vector Machines	0.2197	0.2162	0.2179	0.2197
	Naïve Markov Chain	0.2499	0.2477	0.2492	0.2507
NML	Fractional Regression	<u>0.2871</u>	<u>0.2871</u>	<u>0.2875</u>	<u>0.2879</u>
	Beta Regression	0.3068	0.3064	0.3068	0.3071
	Regression Trees	0.2891	0.2890	0.2895	0.2900
	Support Vector Machines	0.3001	0.3000	0.3006	0.3012
	Naïve Markov Chain	0.4336	0.4331	0.3236	0.4341

non-parametric methods values. The crucial thing here is the presence of collateral, which can be realized in almost any point in time and paths leading to this scenario are not captured well by parametric methods. On particular it can be a result of collection strategy performed by the financial institution or the real estate market

liquidity (or combination of both).
For non-secured loans fractional regression outperforms all other methods both in case of whole sample RMSE, like in non-overlapping confidence intervals. As an alternative regression trees can be viewed. Beta regression and SVMs give significantly worse results, so here the choice is straightforward. The recovery process for credits without collateral seems to be more linear, as it consists only of own payments made by the client during default window. Such patterns can be described mainly by due amounts and DPD, which in fact is done by every method used. And because relationship between RR and these variables can be well described by distribution underlying parametric method, these advantage is moved on out-of-sample data, where non-parametric methods are slightly worse (regression tree) or significantly worse (SVM). Comparing the results to the Naïve Markov Chain, it can be clearly seen, that the rise in quality is relevant. Even the worse method in each segment is not comparable to the selected benchmark (in terms of not overlapping confidence levels), which shows material upgrade of presented approach.
The tasks for future research on partial RR estimation are as follows. Another parametrization of Regression Trees and SVM, like choosing a different splitting method or kernel, should be studied. The trees built in this paper are relatively small, to prevent overfitting, but it looks like there is room to make it more complex to obtain better estimates. Techniques like Random Forest or Gradient Boosting, used, inter alia, in Papouskova and Hajek (2019), can lead to an improvement in performance at the expense of interpretability. SVMs can also be reparametrized with another kernel, like Sigmoid or Hyperbolic, which may reflect the pattern more accurately. Partial Dependency Plots along with Individual Conditional Expectation plots could be added to compare the results with our study. Secondly, interval range is selected arbitrarily, so what is good for one institution, will not always work well for another. Next studies can be broadened by interval selection basing on the recovery patterns specific to the collection process. Thirdly, a reference data set containing information about other risk drivers (such as credit bureau data or detailed collateral characteristics) should be studied to find additional relevant dependencies for consecutive intervals in partial recovery rates estimation.

6 Conclusions

This paper consider a method of estimating partial recovery rates for open cases, basing on modeling recoveries in intervals, where explained variable consist of all cash flows observed from the beginning of the interval till the end of recovery process window. Two parametric and two non-parametric methods are applied on a sample from a Polish commercial bank using AIRB regime to calculate LGD. The selection of the methods was dictated by their robustness confirmed in previous studies (compare with Bastos, 2010 and Yao, Crook, and Andreeva, 2017). Models are built on data from 2003-2015 and validated on defaults closed during the 2015-2017 period. This

study shows that different features drives recoveries when time in default progresses. Recovery patterns are changing and reflect them properly can lead to producing more precise estimates, which finally leads to bias reduction in LGD model, which is in line with Rapisarda and Echeverry (2013) findings. In addition, it is confirmed which method is more suitable to model partial recovery rate. We find that when secured loans are considered, non-parametric methods are able to capture non-linearity, mostly coming from collateral inclusion. Superiority of non-parametric methods was also confirmed in other studies regarding LGD estimation, mention the Loterman et al. (2012) or Tobback et al. (2014). Opposite is true for non-secured loans, where fractional regression gave the best result, but regression trees are only slightly worse. This finding is in opposition to some of the newest studies, but has support in Belotti and Crook (2009), who shows superiority of OLS over selected non-parametric methods. Our solution can be adopted as part of the planned Basel IV framework. The Basel IV will require extended treatment of incomplete defaults compared to previous regulations and consequently leads to more appropriate risk quantification both for setting capital buffers and provision level in the current regulatory and economic environment.

Acknowledgments

The author reports no conflicts of interest. The author alone is responsible for the content and the writing of the paper. We gratefully acknowledge the insightful comments provided by Paweł Baranowski.

References

[1] Anolli M., Beccalli E., Giordani T., (2013), *Retail Credit Risk Management*, Palgrave MacMillan, New York, DOI: 10.1057/9781137006769.

[2] Baesens B., Roesch D., Scheule H., (2016), *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*, John Wiley & Sons.

[3] Basel Committee on Banking Supervision (2005), Studies on the validation of Internal Rating System, available at: `https://www.bis.org/publ/bcbs_wp14.htm`.

[4] Basel Committee on Banking Supervision (2017), Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures (EBA/GL/2017/16), available at: `https://eba.europa.eu/documents/10180/2033363/Guidelines+on+PD+and+LGD+estimation+%28EBA-GL-2017-16%29.pdf`.

[5] Bastos J., (2010), Forecasting bank loans loss-given-default, *Journal of Banking and Finance* 34(10), 2510-2517, DOI: 10.1016/j.jbankfin.2010.04.011.

[6] Belotti T., Crook J., (2007), Modelling and predicting loss given default for credit cards, *Quantitative Financial Risk Management Centre* 28(1), 171–182.

[7] Belotti T., Crook J., (2009), Loss Given Default models for UK retail credit cards, *CRC Working Paper* 09/1.

[8] Brown I., (2012), *Basel II Compliant Credit Risk Modelling*, University of Southampton, Southampton.

[9] Chalupka R., Kopecsni J., (2008), Modelling Bank Loan LGD of Corporate and SME Segments, *IES Working Paper*.

[10] Dermine J., Neto de Carvahlo C., (2006), Bank Loan Losses-Given-Default: a Case Study, *Journal of Banking and Finance* 30(4), 1219–1243.

[11] Gurtler M., Hibbeln M., (2013), Improvements in loss given default forecasts for bank loans, *Journal of Banking and Finance* 37, 2354-2366, DOI: 10.2139/ssrn.1757714.

[12] Hastie T., Tibshirani R., Friedman J., (2008), *The Elements of Statistical Learning*, Springer, DOI: 10.1007/978-0-387-84858-7.

[13] Huang X., Oosterlee C., (2011), Generalized beta regression models for random loss given default, *The Journal of Credit Risk* 7(4), DOI: 10.21314/JCR.2011.150.

[14] Izzi L., Oricchio G., Vitale L., (2012), *Basel III Credit Rating Systems, Palgrave MacMillan*, New York, DOI: 10.1057/9780230361188.

[15] Jarrow R., Lando D., Turnbull S., (1997), Markov model for the term structure of credit risk spreads, *Review of Financial Studies* 10, 481–523.

[16] Liu W., Xin J., (2014), Modeling Fractional Outcomes with SAS, *SAS Paper* 1304–2014.

[17] Loterman G., Brown I., Martens D., Mues C. and Baesens B., (2012), Benchmarking Regression Algorithms for Loss Given Default Modelling, *International Journal of Forecasting* 28(1), 161–170.

[18] Luo X., Shevchenko P., (2013), Markov chain Monte Carlo estimation of default and recovery: dependent via the latent systematic factor, *Journal of Credit Risk* 9(3), 41–76.

[19] Nielsen M., Roth S., (2017), *Basel IV: The Next Generation of Risk Weighted Assets*, John Wiley & Sons.

[20] Papke L., Woolridge J., (1996), Econometric method for fractional response variable with an application to 401(K) plan participation rates, *Journal of Applied Econometrics*, DOI: 10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1.

[21] Papouskova M., Hajek P., (2019), Two-stage consumer credit risk modelling using heterogeneous ensemble learning, *Decision Support Systems* 118, 33–45, DOI: 10.1016/j.dss.2019.01.002.

[22] Qi M., Zhao X., (2011), Comparison of modeling methods for Loss Given Default, *Journal of Banking and Finance* 35(11), 2842–2855, DOI: 10.1016/j.jbankfin.2011.03.011.

[23] Rapisarda G., Echeverry D., (2013), A Non-parametric Approach to Incorporating Incomplete Workouts Into Loss Given Default Estimates, *Journal of Credit Risk* 9(2), DOI: 10.21314/JCR.2013.159

[24] Regulation (EU) No 575/2013 of the European Parliament and of the council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012.

[25] Stoyanov S., (2009), Application LGD Model Development, *Credit Scoring and Credit Control XI Conference*, available at: `https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2017/03/Application-LGD-Model-Development-Nistico-and-Stoyanov.pdf.`

[26] Tobback E., Martens D., Van Gestel T., Baesens B., (2014), Forecasting Loss Given Default models: impact of account characteristics and the macroeconomic state, *Journal of the Operational Research Society* 65(3), DOI: 10.1057/jors.2013.158.

[27] Tong E., Mues C., Thomas L., (2013), A zero-adjusted gamma model for mortgage loss given default, *International Journal of Forecasting* 29(4), 548–562, DOI: 10.1016/j.ijforecast.2013.03.003.

[28] Van Berkel A., Siddiqi N., (2012), Building Loss Given Default Scorecard Using Weight of Evidence Bins, *SAS Global Forum*, available at: `https://support.sas.com/resources/papers/proceedings12/141-2012.pdf.`

[29] Yao X., Crook J., Andreeva G., (2017), Enhancing two-stage modelling methodology for loss given default with support vector machines, *European Journal of Operational Research* 263(2), 679-689, DOI: 10.1016/j.ejor.2017.05.017.

[30] Zięba P., (2017), Methods of Extension of Databases Used to Estimate LGD Parameter, *Studia i Prace Kolegium Zarzadzania i Finansów* 150, 31–55.

Table 13: Point estimates for ML products (fractional regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	LTV	Foreign currency	AIC
0		−1.2397	−5.3147			4.7684								856.2
6			−2.7503			2.2711								922.8
12			−1.8843			1.9214								684.6
18			−1.9907			1.0843								562.2
24			−1.9232											470.0
30			−1.9091											368.6
36			−2.1006											281.2
42			−1.4796			1.4982						2.3506		214.9
48			−1.5652									2.7410		172.6
54			−1.5384									2.4625		111.5
60			−2.3784									0.9706		83.6

Table 14: Point estimates for NML products (fractional regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	AIC
0		−6.0128		2.4939		3.5752						49264
6				1.4825		3.1668						42487
12		−1.1924				2.4163						37379
18		−1.0850	−0.6197			1.1666						31277
24		−1.1985	−0.4834			1.3169						21616
30		−0.7032				1.4386						14170
36		−0.3831				1.4426			1.8663			8549.0
42		−0.4507				1.3873			1.6683			4500.7
48		−0.8006				0.9894			2.0679			1905.6
54			1.0519			1.1111			1.4902			897.8
60			1.7044									535.2

Table 15: Point estimates for ML products (beta regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	LTV	Foreign currency	AIC
0		−1.4276	−8.9717			4.5474								−36986
6			−2.5036			1.9032								−8814
12			−1.8090			1.8197						−2.1738		−5333
18			−1.7366			0.6983								−2925
24			−1.4090			0.8710								−1707
30			−1.3399			1.2221								−964.3
36			−1.0940			1.4032								−481.5
42			−1.8745		−1.5825							1.7143	−0.9733	−288.3
48			−1.8366		−1.3086							1.4967	−0.8371	−161.0
54		2.7190	−2.0339	1.5944										−65.4
60			−2.5486							0.9476				−70.0

Table 16: Point estimates for NML products (beta regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	AIC
0		−3.5808	−3.7917	1.1895		2.8365				−4.3850		−196000
6		−0.7727		0.5931		2.1413		−2.0596				−27684
12		−1.1711		0.2217		1.8590		−1.2612				−9219
18		−1.3018				1.3188						−7715
24		−1.6910	0.2743			1.4502						−3537
30		−1.1769				1.4664			3.2828			−2736
36		−0.9329				1.7543			3.7141			−2917
42		−0.8839				1.4680			2.3343			−1644
48		−0.8798			−8.7441	0.3529			10.2564			−1438
54		−0.9045	0.4408	1.2967	−7.6474				5.6963			−2668
60		−0.8611		1.8029	−11.1726				8.0010			−1772

Journal of Risk Model Validation **15**(1), 69–91
DOI: 10.21314/JRMV.2020.234

Research Paper

Beyond the contract: client behavior from origination to default as the new set of the loss given default risk drivers

Wojciech Starosta

Department of Economics and Sociology, Institute of Econometrics, University of Łódź, ul. POW 3/5, 90-255 Łódź, Poland; email: w.starosta@wp.pl

(Received October 7, 2019; revised and accepted November 16, 2020)

ABSTRACT

Modeling loss given default has increased in popularity as it has become a crucial parameter for establishing capital buffers under Basel II and III and for calculating the impairment of financial assets under the International Financial Reporting Standard 9. The most recent literature on this topic focuses mainly on estimation methods and less on the variables used to explain the variability in loss given default. In this paper, we expand this part of the modeling process by constructing a set of client-behavior-based predictors that can be used to construct more precise models, and we investigate the economic justifications empirically to examine their potential usage. The main novelty introduced in this paper is the connection between loss given default and the behavior of the contract owner, not just the contract itself. This approach results in the reduction of the values of selected error measures and progressively improves the forecasting ability. The effect is more visible in a parametric method (fractional regression) than in a nonparametric method (regression

Print ISSN 1753-9579 | Online ISSN 1753-9587

tree). Our findings support incorporating client-oriented information into loss given default models.

Keywords: credit risk; retail; variable selection; recovery rate; loss given default.

1 INTRODUCTION AND LITERATURE REVIEW

Loss given default (LGD) estimation causes many methodological and calculation problems, including the bimodal distribution (see Loterman *et al* 2012; Yao *et al* 2017), the need for a wide observation window,[1] the inclusion of partial recoveries, and difficulties in identifying proper predictors. Current studies mainly focus on finding new estimation methods, which can be more precise due to nonstandard statistical procedures, including the recently exploited two-stage modeling approach. In terms of applying new techniques, studies by Qi and Zhao (2011) and Brown (2012) may serve as an example: Brown (2012) compared methods such as ordinary least squares (OLS), beta regression, regression trees, least squares support vector machines (SVMs) and neural networks; Qi and Zhao (2011) focused on fractional response regression, inverse Gaussian regression, regression trees and neural networks. In addition, interesting approaches were adopted by Luo and Shevchenko (2013), who used the Markov chain Monte Carlo method, and Witzany *et al* (2012), who attempted to use survival time analysis techniques, in particular the proportional Cox model and its modifications, for LGD estimation. Alternatively, Loterman *et al* (2012) used two-stage modeling, which is becoming increasingly popular, by combining logistic regression with neural networks, OLS with regression trees and OLS with least squares SVMs, as well as many other methods.

The main issue in these types of models is the need to distinguish extreme LGD values from the middle range of the distribution. Yao *et al* (2017) investigated the problem in terms of classification (high and low values of LGD) and regression (values between high and low). They used the least squares support vector classifier and a set of regression methods (OLS, fractional response regression, etc). The same framework was also suggested by Gürtler and Hibbeln (2013), who tested a hypothesis about the superiority of a two-step model over direct LGD regression in terms of the coefficient of the determination measure. Based on these studies, as well as the work of Huang and Oosterlee (2011), Liu and Xin (2014) and Nazemi and Fabozzi (2018), we see a switch from canonical methods, such as OLS or historical averaging, to regression trees, SVM or beta regression. These new LGD practices can help with the bimodal LGD distribution either by being more flexible (eg, using beta

[1] Some recoveries can last eight years or more.

regression or regression trees) or by dividing it into regions in which the estimation will give better results (two-stage models).

Less importance is attached to the issue of finding appropriate predictors that can better explain the LGD variance and improve the forecasting ability.[2] According to the Basel Committee on Banking Supervision (2017, p. 30), institutions can consider the factors related to transactions, obligors and institutions (in terms of an organization's recovery processes or legal frameworks). Significant attention has been paid to choosing only the meaningful differentiating risk factors of transactions. Most of the latest studies focusing on retail banking consider contract-driven information as the predictors.[3]

There are three main sources of information when the set of predictors is to be established in LGD modeling (Ozdemir and Miu 2009, p. 17).

(1) Contract: loan-to-value (LTV), exposure at default (EAD), loan term, etc.

(2) Client: previous default indicator, employment status, monthly income, etc.

(3) Macroeconomics: house price index (HPI), consumer price index (CPI), etc.

The main contribution of our study is that it investigates the potential sources of risk driven by consumer behavior after credit is granted. Behavioral data is widely used in other areas of risk management, such as probability of default (PD) estimation (see Izzi *et al* 2012, p. 63; West 2000) or fraud detection (see Kovach and Ruggiero 2011; Fiore *et al* 2019). However, our study is, to the best of our knowledge, the first to add such wide-ranging data to the LGD model. We used transactional data, application data (about credit but also the full spectrum of banking services) and bank–client relation data to create new risk drivers, all with economic justification explicitly given and checked using out-of-sample data to confirm robustness. Our data set originated from a major Polish bank, where online communication (via a personal computer or a cell phone) between the bank and its customers is the primary contact channel. This allowed us to state that the fulfillment of the variables based on internet activity is sufficient[4] and covers the complete business cycle.[5]

We checked whether the inclusion of the predictors describing the contract owner's behavior leads to an increase in the precision and discrimination of LGD estimates. The research on estimation methods employs increasingly complicated structures,

[2] Corporate bonds were studied in Schuermann (2004), and small and medium-sized enterprise segments were studied in Chalupka and Kopecsni (2008).

[3] See Table 2 for a comparison.

[4] This is not always ensured in the case of LGD models, where at least a five-year time series is needed to perform an estimation.

[5] In line with the Basel II regulations (Basel Committee on Banking Supervision 2005, p. 65).

in which the appropriateness of, and consistency with, collection and recovery policies can be difficult to demonstrate (as requested in Basel Committee on Banking Supervision (2017, p. 74) and discussed in academia (see, for example, Martens *et al* 2011)). Thus, expanding the scope of information seems to be a reasonable choice to provide the best estimates possible. While a comprehensive approach to clients in LGD estimation is not widely recognized, it can be especially important in the nonmortgage loan (NML) segment, because recovery rate (RR) variability cannot be described by security possession. Assuming the new risk drivers have a significant impact on the LGD estimates, we evaluated whether parametric or nonparametric methods gave a larger increase in precision and discrimination.

The remainder of this paper is structured as follows. In Section 2, we propose the set of predictors with a calculation method and a link to LGD. In Section 3, we discuss champion and challenger approaches and goodness-of-fit measures. In Section 4, we estimate a model with standard explanatory variables as a champion approach. In Section 5, we discuss the challenger models with one or more new variables and goodness-of-fit measures to compare all variants. In Section 6, we present our conclusions.

2 DATA AND A DESCRIPTION OF THE VARIABLES

Considering various ways of determining a dependent variable (Basel Committee on Banking Supervision 2005, p. 4), let LGD be defined as

$$\mathrm{LGD} = 1 - \mathrm{RR} = 1 - \sum_{t=1}^{n} \frac{\mathrm{CF}_t}{(1+d)^t} \left(\frac{1}{\mathrm{EAD}} \right).$$

This is a so-called workout approach, where the LGD is determined as 1 minus the sum of the discounted cashflows divided by the exposure at default (EAD) (Anolli *et al* 2013, p. 92).

A data set of NMLs used in the analysis (with the LGD meeting the definition above) was provided by a major Polish bank and consists of around 135 000 observations from May 2011 to December 2017. All the defaults come from nonsecured portfolios containing credit cards, cash loans and revolving loans given to private individuals or small and medium-sized enterprises (SMEs). Complete and incomplete cases were included in the reference data set, as the mean workout period was relatively long (30 months for complete defaults) (see Tanoue *et al* 2017). In such circumstances, the removal of incomplete defaults could lead to sample selection bias.

TABLE 1 Descriptive statistics of recovery rate for closed cases only and full sample.

Variable	Minimum	Q1	Mean	Q3	Maximum
RR for closed cases	0.0000	0.2372	0.6764	0.9980	1.0000
RR for full sample	0.0000	0.2471	0.5527	0.9528	1.0000

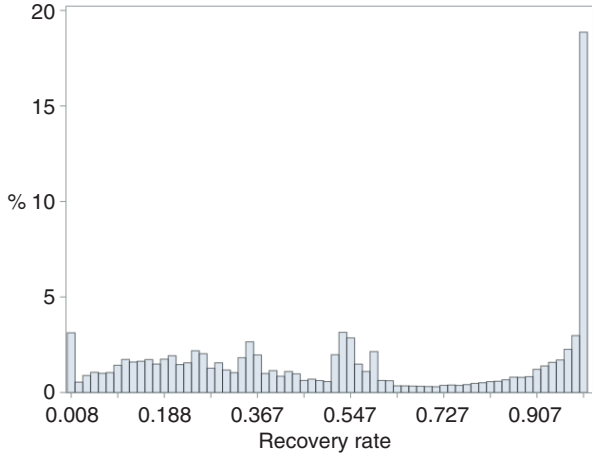
To estimate the partial RR for incomplete defaults from the moment of data origination until the end of the default, we used an approach presented in Starosta (2020). From the given methods, we chose the fractional regression model, which was the best approach reported for NMLs. In a second step, we divided the sample into 10 subsamples (based on the time-in-default characteristic) and calculated the partial RR from the start of the interval until the end of the default. Finally, we estimated the models and assigned predicted values to the actual values.

The structure of the sample is different after including the incomplete cases, as presented in Table 1. This is mainly the result of the long-lasting defaults characterized by low recoveries, where the litigation process is still in progress. Removal of these cases could cause a serious underestimation of the LGD (Gürtler and Hibbeln 2013). The final distribution of the RR is presented in Figure 1.

As a benchmark, we estimated the model using the following standard set of predictors:

- type of product (credit card, cash loan, revolving loan);
- interest rate;
- requested amount;
- time on the books;
- principal amount;
- interest amount;
- EAD;
- tenor;
- decreasing installment indicator;
- second applicant indicator;
- length of relationship;
- client age.

FIGURE 1 Distribution of the recovery rate for full sample.



The predictors in the above list generally coincide with the features from studies on retail banking in which LTV, EAD, time on the books and loan size are the most common risk drivers. Table 2 summarizes these studies. We mainly considered contract-driven indicators in our study, but we also included some client-based indicators. At this stage, macroeconomic variables were not included in any version of the model. This decision was driven mainly by the fact that there was no serious downturn period in the sample window for the Polish market.

As shown in the next section, we extended the basic set of predictors by including new factors that are not widely used in RR estimation but still have an economic link with a modeled phenomenon. We mainly focus on the client-based indicators, as there is almost nothing to add to the contract-driven indicators. We distinguished three main data sources, also mentioned in the credit scoring literature (see Anderson 2007, p. 275): transactional data, behavior on credit accounts and applications for other bank products.

2.1 Transactional data

The first new risk driver (cash_dep_acc) is the amount that the owner of the defaulted contract has in any other deposit accounts. Here, the link to the RR is straightforward. Having a "financial cushion" in the form of liquid assets such as cash makes it easier to return to a nondefault portfolio in a time of financial distress. In this scenario, the

TABLE 2 The main risk drivers and goodness-of-fit measures from selected studies.

Study	Main risk drivers	Goodness-of-fit measures		
		R^2	MAE	RMSE
Leow (2010)	LTV; Previous default indicator; Time on books; Security indicator; Age of property; Region of property	0.233–0.268	0.101–0.121	0.158–0.161
Brown (2012)	Age of exposure; Months in arrears; Loan amount; Application score; Joint application indicator; Time at bank; Residual status; Employment status; LTV; Time on books; Loan term; EAD; Utilization	−0.695–0.497	0.034–0.431	0.122–0.693
Anolli *et al* (2013)	Geographic area; Income; Age; Marital status; Vintage of contract; Duration; Installments; Delinquency; LTV; Gross Domestic Product; Unemployment rate; Saving ratio; HPI	N/A	N/A	N/A
Tong *et al* (2013)	EAD; Loan amount; Property valuation; HPI; Time on books; Debt to value; Previous default indicator; Second applicant indicator; Loan term; Property age; Security type; Geographical region	0.298–0.338	N/A	N/A
Zhang and Thomas (2012)	Employment status;Mortgage indicator; Visa card indicator; Insurance indicator; Number of dependents; Personal loan account indicator; Residential status; Loan term; Loan purpose; Time at address; Time with bank; Time in occupation; Monthly expenditure; Monthly income	0.029–0.107	0.352–0.408	0.406–0.493
Thomas *et al* (2010)	Number of months in arrears; Application score; Loan amount; Time until default of the loan	0.083–0.227	N/A	N/A
Yao *et al* (2017)	Loan term; Time on books; Sum of transactions across all current accounts; Month in arrears; EAD; Delinquency status; Number of payments made; Most recent payment received; HPI; CPI	0.069–0.647	0.154–0.318	0.211–0.356

client's motivation to use these funds plays a significant role; the client's decision can be driven by the term of the deposit and the penalty for breaking it. Potentially, the expected benefits from longer deposits, which are more difficult to withdraw in the short term, could be greater than the benefits from repaying a loan. Although the size of the impact must be measured, the general rule is that the final RR increases with the amount of cash available on the deposit accounts. Variables can be determined for different time horizons. In the basic version, this is the sum of the balances of the deposit accounts at the time of the default. In a more complicated alternative, this could be the mean balance from the last N months before the default.

The number of log-ins (n_login) is the second variable analyzed. If a financial institution has developed an effective internet banking system, at some point this becomes the easiest channel of communication for its clients. Nowadays, functionalities such as fully automatic credit analysis, chats/video chats or brokerage account management can be accessed via the internet without leaving home. Many clients use the internet as their main form of communication; they only go to a bank branch as a last resort. It is also useful for banks to know the best way to reach their clients. The hypothesis concerning this point is about higher RRs for clients who use internet banking more often, as they can be reached more easily by debt collection departments in the case of any financial difficulties. Similarly to cash_dep_acc, n_login can be checked at the time of default or during the last N months.

2.2 Behavior on credit products

There is less information on NMLs than secured loans; thus, every piece of data should be considered as a potential factor. We therefore examined the connection between the contract owners and the financial institution. It should be a positive signal for the RR that an NML owner also has a mortgage loan (ML). The motivation related to repaying secured credit, where a house or a car is set as collateral, is significantly different from the case where a client has nothing tangible to lose. Empirical studies have confirmed that RR values are higher for secured credit than for nonsecured credit (Basel Committee on Banking Supervision 2005, p. 74), which leads to the hypothesis that possession of an ML contract by at least one owner leads to an increased RR when the NML contract is analyzed:

$$\text{ML\_indic} = \begin{cases} 1 & \text{if at least one contract owner has ML,} \\ 0 & \text{otherwise.} \end{cases}$$

In a more conservative version, in order to assign an ML_indic $= 1$ every owner has to have an ML contract.

Tenor and time on the books are widely known LGD predictors, and they are often used in studies. However, when a financial institution has many different tenors on offer, this could lead to a discrepancy in estimating the LGD for contracts with low tenors when most of a sample consists of long tenors, and vice versa. One way to deal with this situation is to use the credit life cycle phase instead of nominal tenor and time on the books:

$$\text{credit\_life\_cycle} = \frac{\text{time on the books}}{\text{tenor}}.$$

Time on the books should always lie in the interval [0; 1], and it should have a simple interpretation in almost all cases. If the value of a variable is larger, is should be possible to obtain a greater recovery. It is assumed that the clients' motivation to repay credit is higher at the end of a schedule than at the beginning, when the perspective of getting rid of a financial burden is farther away. Thus, instead of taking the contract perspective expressed by the length of credit, the variable is presented from a client's perspective and is expressed by how many months are left on the loan or how close the client is to full repayment.

Finally, we checked the information about delinquencies on any client contract in the selected historical period. Stating the past due amount on the analyzed contract as well as the other products owned can lead to lower RRs in comparison with clients with a clean delinquency history. However, if another contract is past due, then the collection department can take care of the rest of the client contract earlier. Thus, depending on the collection policy, a positive impact can also be valid. We define material delinquency as 30 days past due on an amount higher than 1% of EAD:

$$\text{n\_delinq} = \frac{\text{number of days with delinquency}}{\text{number of days possessing credit product within last year}}.$$

This variable can also be examined in the selected time period.

2.3 Requesting other bank products

During the life of a credit, the client's cooperation with the bank can develop in different directions. Other credit products can be granted (or requested but not granted), an application for deposit products can be filled out or insurance products can be requested.[6] All these events can affect the LGD in various ways, so we examined

[6] Cooperation between banks and insurance companies is getting closer, and insurance connected to repayment in the case of an unexpected event is now a standard add-on to any credit product.

each of them with the use of binary variables, defined as follows:

$$\text{insurance} = \begin{cases} 1 & \text{if client bought insurance after initial credit was granted,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{next\_credit\_granted} = \begin{cases} 1 & \text{if client got another loan after initial credit was granted,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{next\_credit\_app} = \begin{cases} 1 & \text{if client applied for another loan after initial credit was granted,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{deposit} = \begin{cases} 1 & \text{if client put a deposit down after initial credit was granted,} \\ 0 & \text{otherwise.} \end{cases}$$

The above features were selected to focus on expanding the potential determinants of LGD in the following directions:

- client behavior on a deposit account (managing inflows and outflows, saving propensity);
- degree of relationship with the bank (log-ons, but also channels of communication, possession of mobile applications, etc);
- product structure (requests for other products, both credit and insurance);
- seeking new relationships in core variables (inverting the perspective toward the client).

In our study, at least one new proposition from the list presented above is formulated to check the potential of each area. The results could suggest the most promising area of investigation for future research.

3 METHODOLOGY

The champion and challenger approaches are used to evaluate performance or to determine a set of benchmarking values (Apeh *et al* 2014). In our study, the champion method corresponds to the most effective model built with a standard set of explanatory variables. Estimation occurs via two methods.

The first method is fractional regression (see Belotti and Crook 2009; Bastos 2010), which is viewed as being a good benchmark for more complicated methods, but also as having the desired properties and giving interpretable results that are

preferred from a regulatory point of view:

$$E(\mathrm{RR}_i \mid \boldsymbol{x}_i) = \frac{1}{[1 + \exp(-\boldsymbol{x}_i \boldsymbol{\beta})]},$$

where $i = 1, \ldots, n$, n is the sample size, $\boldsymbol{\beta}$ are model parameters and $\boldsymbol{x}_i$ is a vector of the explanatory variables for case i. In the estimation process, forward, backward and stepwise methods are used with a p-value of less than 0.05 as the criterion for adding a variable. The maximum value of the correlation between the final set of predictors is fixed at 80%. The variable with the higher loglikelihood is ultimately used, and the second one is removed.

A second method is the regression tree method (see Bastos 2010; Qi and Zhao 2011), a nonparametric approach that, in theory, better reflects nonlinear dependencies, such as in the LGD case. The interpretation is also straightforward, as this model results in a set of rules (binary splits) that can be shown as a combination of "if–else" statements. However, the method has some drawbacks, such as the potential instability connected to changes in the population or overfitting (Hastie *et al* 2008). It is necessary to take great care when tuning hyperparameters of a tree to limit these two issues. In this study, we implemented a standard set of tuning rules (see Qi and Zhao 2011; Nazemi and Fabozzi 2018).

(1) Select analysis of variance as the splitting selection method (applicable for continuous variables).

(2) Determine the complexity parameter based on tenfold cross-validation.

(3) Calculate the value of the minimum observations in a leaf in the manner proposed by Israel (1992):

$$n = \frac{z^2 \sigma^2}{\varepsilon^2},$$

where z is the abscissa of the normal curve that cuts an area, ε is the desired level of precision at the tails and σ^2 is the variance of the LGD. The value obtained is rounded up to the nearest integer.

(4) Assign a value of 5 to the hyperparameter to achieve stable and intuitive results. This parametrization should lead to the tree, where the maximum depth value should not be binding.

Each model is checked for precision and discrimination using the methods described in Table 3 on a holdout sample consisting of 30% of the total. The remaining 70% is used to train the model. We chose two methods – root mean square error (RMSE) and mean absolute error (MAE) – to evaluate precision, and two methods – the Gini index and the cumulative LGD accuracy ratio (CLAR) – to assess discrimination. Each of these proved useful in previous studies.

TABLE 3 Precision and discrimination measures used in LGD model evaluation.

Measure	Calculation	Usage
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{LGD}_i - \widehat{\mathrm{LGD}}_i)^2}$	Bastos (2010)
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert\mathrm{LGD}_i - \widehat{\mathrm{LGD}}_i\rvert$	Tanoue *et al* (2017)
Gini index*	2AUROC − 1	Zhang and Thomas (2012)
CLAR	Bucketing predicted and realized LGD	Ozdemir and Miu (2009)

*AUROC denotes area under receiver operating characteristic.

The precision values should be as low as possible, and the discrimination values should be as high as possible. The calculation of the RMSE, MAE and CLAR does not cause a major problem. However, in the case of the Gini coefficient we need to make a clarification: because a variable is expected to be binary, we divided each observation into two occurrences, and we assigned 1 to the first occurrence and 0 to the second. We then calculated the weighted Gini index, where the weight for the first case is the value of the RR and the weight for the second case is $1 - \mathrm{RR}$. As there are no official benchmarks for any of the selected measures, we compared each model with all the others for all four measures. Additionally, we checked the increase in performance compared with the naive classifier expressed as the average value from the sample.

4 CHAMPION MODELS: EMPIRICAL RESULTS

4.1 Fractional regression

In each approach, we treated the RR as a dependent variable for the estimations. The final form of the fractional regression model is presented in Table 4.

From the table, we can draw some straightforward conclusions related to the estimates of the parameters. Both the length of relationship and the requested amount positively influenced RR. The longer a customer stays with the bank, the higher the proportion of debt that will be recovered after default, which is in line with the findings reported in Tong *et al* (2013) and Yao *et al* (2017). The same can be stated for the requested amount, where a positive sign was reported in Brown (2012) (in the same study, the sign was negative for another data set). This may be connected with the EAD estimate, which is negative, as reported by Tong *et al* (2013) and Tanoue *et al* (2017): high EAD values result in fewer recoveries. The model aims to differentiate cases with a high requested amount and a high EAD (there is still a high due amount,

TABLE 4 Parameter estimates of regression model in the champion approach.

Parameter	Estimate	Standard error	$p\ (>\chi^2)$
Intercept	0.6254	0.0269	<0.0001
EAD (in thousands)	−0.0038	0.0012	0.0051
Length of relationship	0.00296	0.0002	<0.0001
Client age	−0.0081	0.0006	<0.0001
Requested amount (in thousands)	0.0099	0.0009	<0.0001
Tenor	−0.0077	0.0003	<0.0001
Interest amount (in thousands)	−0.0007	0.0001	<0.0001

so there are lower recoveries) from a high requested amount and a low EAD (where the greater part of the outstanding credit has already been repaid or limit usage was low, so there is a chance to recover more, as the client could be more willing to complete the repayment). At this point, a new variable, such as the EAD divided by the requested amount, which can connect the two values mentioned above, may be useful. However, to compare the full specification, we did not consider this. Then, there is client age, which decreases the RR by 0.81% every year. To some extent, the same conclusion can be found in Belotti and Crook (2009), where the impact was estimated at the 0.346% level. The interest amount is the last variable included in the model; it has a negative influence on the RR.

4.2 Regression trees

The second model, based on regression trees, was tuned as follows:

- the minimum number of observations in a leaf was set equal to 181;
- the number of cross-validations was set equal to 10;
- the complexity parameter was set equal to 0.00063483691;
- the maximum depth was set equal to 5.

In comparison with the fractional regression model, four additional variables are used to construct a tree; this is mainly due to the inclusion of nonlinear dependencies and no further assumptions being made about the correlation between the predictors.

TABLE 5 The importance of the variables used in tree construction.

Variable	Importance
Interest amount	659.3
Principal amount	612.9
EAD	584.7
Requested amount	310.3
Interest rate	271.2
Tenor	242.1
Length of relationship	134.8
Months on book	49.4
Client age	5.9
Decreasing installment indicator	2.5

Here, importance is the sum of the goodness of split measures for each split for which it was the primary variable, plus the goodness for all splits in which it was a surrogate.

TABLE 6 Performance measures for the champion approaches.

Measure	Fractional regression	Regression tree
RMSE	0.33242	0.32002
MAE	0.29378	0.27700
Gini (%)	20.08	26.01
CLAR	0.7338	0.7233

Table 5 shows the importance of each predictor; similar to the fractional regression method, the interest amount is one of the strongest predictors. Unsurprisingly, the variables connected to principal and interest, and those derived from these two variables, have the greatest impact on the RR prediction.

4.3 Comparison of the champion approaches

At this point, we analyzed the predictive accuracy to establish benchmarks that the challenger approaches need to beat.

The results suggest that both methods perform well on this particular data set (Table 6). The selected metrics are slightly in favor of the regression tree, but we are not neglecting any method at this point. Globally, the performance indicators are not very different from those reported in similar studies on nonmortgage products (see, for example, Yao *et al* 2017). As previously mentioned, there is a body of literature

about other sophisticated methods, but we did not focus on finding the best statistical model. Thus, further research may be needed to obtain more insight about LGD modeling for unsecured loans, both here and when using client behavior variables.

5 CHALLENGER APPROACH: EMPIRICAL RESULTS

In this section, we assess the performance of the newly created variables, first individually, then in terms of their interactions. In the first path, the data set is reestimated, and extended one variable at a time. This allows us to determine whether a specific risk driver is relevant in the RR modeling process. Second, the entire set of variables, both contract-level and client-level, is included in the estimation to verify the hypothesis of a potential relationship between the RR and the contract owner behavior.

The strongest influence, confirmed by six of the eight measures in both the parametric and nonparametric models, is exerted by credit_life_cycle. The precision of the model was increased by 0.00473 in the case of MAE, and by 3.65 percentage points in the case of Gini. This result suggests that inverting the perspective to a client view can elicit new insight from the raw contract data. The sign is strongly positive, which is compliant with the general assumption that being closer to full repayment has a positive effect on the RR. Another "changing perspective" variable is the share of EAD in a requested amount (or limit usage in the case of credit lines/credit cards), which determines the profile of the repayment pattern, such as decreasing installments, prepayments or high-volume usage. Each new variable achieves statistical significance at a 0.05 p-value level when analyzed separately. In the regression tree model, some of the variables (eg, the number of log-ins or the ML indicator) were not used in any division.

In the next step of the challenger approach, we examine the estimation of the entire training set using both base and new variables as predictors. In the fractional regression model, the base specification was used and all the new variables were added. Because the correlation coefficient between the indicators next_credit_granted and next_credit_approved was found to be 92%, the variable with the higher loglikelihood was used. The assumptions discussed in Section 3 hold for a regression tree.

Seven of the eight new variables were found to be statistically significant, which supports the hypothesis that the contract owner behavior is connected to the RR level (see Table 8). The same can be stated for the regression tree model (Table 9).

Considering the signs of the parameters, three of the four assumed directions hold; for the next three parameters, this direction is arguable. Being close to the end of the credit term, with all other factors unchanged, has a positive effect on the RR, which is a desirable property, as we expected it to work in the assumed direction. Taking into

TABLE 7 Performance measures for challenger approach: one variable at a time.

(a) Fractional regression

Variable	RMSE	MAE	Gini (%)	CLAR
cash_dep_acc	0.33142	0.29273	20.79	0.7354
n_login	0.33074	0.29270	21.94	0.7372
ML_indic	0.33194	0.29292	20.34	0.7315
credit_life_cycle	0.32807	0.28905	23.73	0.7417
n_delinq	0.33177	0.29295	20.61	0.7354
next_credit_granted	0.33232	0.29353	20.42	0.7344
next_credit_app	0.33219	0.29329	20.60	0.7349
deposit	0.33234	0.29355	19.81	0.7319
insurance	0.33171	0.29212	20.92	0.7361

(b) Regression tree

Variable	RMSE	MAE	Gini (%)	CLAR
cash_dep_acc	0.31982	0.27667	25.89	0.7189
n_login	0.32002	0.27700	26.01	0.7233
ML_indic	0.32002	0.27700	26.01	0.7233
credit_life_cycle	0.31796	0.27499	27.87	0.7275
n_delinq	0.32035	0.27758	25.69	0.7311
next_credit_granted	0.32022	0.27512	26.12	0.7579
next_credit_app	0.32033	0.27487	26.06	0.7416
deposit	0.31978	0.27659	26.08	0.7241
insurance	0.31978	0.27659	26.08	0.7241

account the sum of cash on deposit accounts, we also found a positive coefficient sign, which confirms the hypothesis presented in Section 2. We used the average amount from the last three months before default in our study, but the analysis could definitely be widened to a six- or even a twelve-month window to select the best predictor. We should also consider that, in the selected portfolio, some clients may only have a credit product at a financial institution (with no debit account). It appears that the average RR for such clients can be found among clients with a positive amount of cash in their deposit accounts.

The third and fourth variables discussed are the number of log-ins and the ML indicator. Regarding the ML indicator, undoubtedly having an ML has a positive effect on the RR. However, the number of log-ins reveals a different behavior than expected. At the beginning, there is indeed a positive relationship, but a negative

TABLE 8 Parameter estimates of regression model in the challenger approach (full set of variables).

Parameter	Estimate	Standard error	$p\ (>\chi^2)$
Intercept	−0.2587	0.0437	<0.0001
EAD (in thousands)	−0.0036	0.0004	<0.0001
Length of relationship	0.0030	0.0002	<0.0001
Client age	−0.0094	0.0007	<0.0001
Tenor	0.0033	0.0004	<0.0001
Months on book	−0.0045	0.0003	<0.0001
Sum of cash on deposit accounts (in thousands)	0.0041	0.0003	<0.0001
ML indicator	0.4546	0.0346	<0.0001
Credit life cycle	1.5062	0.0426	<0.0001
Number of log-ins	−0.0206	0.0012	<0.0001
Delinquencies	0.2042	0.0273	<0.0001
Insurance	−0.1617	0.0144	<0.0001
Deposit	−0.1588	0.0237	<0.0001

relationship develops as the number of log-ins increases. This suggests that, at some point, the client may try to stabilize their situation by logging in frequently, but the RR is not pushed in the desired direction. In this case, a change in the nonlinear specification can be considered, such as adding the squared version of a particular variable. Nevertheless, this behavior can be treated as unexpected, and it should be more fully investigated in further studies. Delinquencies on the other contracts possessed by the client have a positive effect on the RR, which could be due to the collection policy, as mentioned in Section 2.

The parameter sign for the last two variables is negative, but we did not make any initial assumptions about their influence, which implies the need to confirm the direction in further research.

In the regression tree model, the new variables have less influence, as the principal amount and EAD still play a major role. However, the credit life cycle is presented as one of the strongest risk drivers, and the indicator of the next credit application is in the top half of all the variables in Table 9. Finally, the information presented in Table 10 demonstrates the performance measures using behavioral characteristics.

The selected measures indicate that the new specification performs well. However, this could still be the result of incorporating more independent variables, so more research is required. To check the robustness of these metrics, a three-step validation was performed.

TABLE 9 The importance of the variables used in tree construction for the challenger approach (full set of variables).

Variable	Importance
Principal amount	616.3
EAD	614.5
Credit life cycle	555.2
Interest amount	526.9
Tenor	282.2
Requested amount	190.3
Indicator of next credit application	170.7
Indicator of next credit granted	164.9
Sum of cash on deposit accounts	150.7
Months on book	148.9
Insurance	94.7
Number of log-ins	37.3
Length of relationship	19.7
Delinquencies	18.6
Decreasing installment indicator	1.1
Client age	0.1

(1) Calculate the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which take the number of parameters into account and penalize it. As the regression tree is a nonparametric estimation method, we used the following equations to calculate the selected measures (Kuhn and Johnson 2013):

$$\text{AIC} = n \log(\text{RSS}) + \alpha \, \text{number\_of\_leaves},$$
$$\text{BIC} = n \log(\text{RSS}) + \log(n) \, \text{number\_of\_leaves},$$

where RSS denotes the residual sum square.

(2) Estimate a naive classifier expressed as a mean and calculate the relative change from this classifier to the champion model and to the challenger model

TABLE 10 Performance measures for the challenger approach.

Measure	Fractional regression	Regression tree
RMSE	0.32790	0.31746
MAE	0.28845	0.27332
Gini (%)	23.89	28.52
CLAR	0.7426	0.7931

TABLE 11 The AIC and BIC for the champion and challenger approaches.

Measure	Champion	Challenger
AIC fractional regression	128 203.74	126 198.68
BIC fractional regression	128 274.18	126 329.42
AIC regression tree	−92 938.17	−93 585.30
BIC regression tree	−95 852.01	−93 456.06

TABLE 12 Relative change (in percent) from naive classifier to the champion approach, the challenger approach and the approach based only on new variables.

Measure	Champion FR	Champion RT	Challenger FR	Challenger RT	Only new FR	Only new RT
RMSE	−3.10	−6.71	−4.41	−7.46	−3.80	−4.38
MAE	−3.39	−8.91	−5.14	−10.12	−4.20	−5.09
CLAR	22.14	20.39	23.60	32.01	22.79	23.44

The first two rows are expected to be negative, and the last one positive. The Gini index cannot be computed for the naive classifier as there is only one level of estimated value. FR, fractional regression. RT, regression tree.

for error measures, where

$$\text{relative change} = \frac{\text{measure} - \text{measure}_{\text{reference}}}{\text{measure}_{\text{reference}}}.$$

(3) Perform the estimation only on new variables to check whether the relative change between this model and the naive classifier is at least as good as the change between the champion model and the naive classifier.

The results in Tables 11 and 12 suggest that each approach performs better than a sample mean, but the degree of goodness-of-fit is wide. Taking fractional regression into consideration, the challenger model can be characterized by a material upgrade, and adding new variables significantly boosts the precision and discrimination. The

robustness of this approach is confirmed by the AIC, the BIC and the out-of-sample precision. Moreover, only selecting the newly created variables seems to be better than using the champion approach (for example, a −3.80 versus a −3.10 gain on RMSE). This could stem from a greater linear dependency between the client-related variables than between the contract-related variables for changes in the RR, which is one of the assumptions of the fractional regression model. In addition, the AIC and the BIC confirm a better fit for the model with the new predictors, which allows us to say that the second model outperforms the base model, given the different specifications. These findings are slightly less obvious for the regression tree model. Even if our interpretation of the AIC and the BIC leads to the same conclusion as for the fractional regression, the relative change between the champion and challenger approaches is a little smaller, as is the influence of the new variables. In this situation, the gain in performance measures is clearly smaller for the "only new" approach than for the champion approach. We can argue that, taking into consideration this particular data set, the regression tree benefited from nonlinearity in the predictors, which reduced information gain from the client-oriented predictors. The same holds in terms of the discrimination measure, where "only new" is less effective than the champion approach but is still substantially better than a sample mean.

We believe that much more information, from the client, not just the contract, could be used for the LGD/RR estimation. Different recovery patterns can be seen for clients that are self-employed versus clients that are employed full time, or those that have already repaid some of their other obligations in their credit history versus new borrowers. More transactional data (such as the amount of inflows or payment patterns) or geolocation data (indicating changing jobs) can be adopted. Even information that is already available can be useful when used in a new manner, such as a share of EAD in the requested amount or the dynamics of log-ins, not just the mean. However, when creating new predictors, we should always consider their usefulness for the PD models, so the correlation between PD and LGD can be properly reflected in the capital requirements calculation.

6 CONCLUSION

The main aim of this study was to establish a connection between LGD and a new set of contract owner-oriented variables. We based the study on a large sample of NMLs from a Polish bank with an online communication system as a primary client contact channel. Evaluation of their performance shows that these new predictors are an effective supplement to the standard predictors, improving the LGD precision.

First, two techniques were applied to build champion models based on a standard set of predictors (fractional regression and regression tree), and then the performance of each new variable was checked and the estimation was performed on the entire set

of variables. The regression tree method performed better than the fractional regression method in terms of the selected measures in both the champion and challenger approaches.

Second, adding client-based variables significantly reduced the error measures and increased discrimination in comparison with the champion approach; this upgrade is greater when using the fractional regression method than the regression tree method.

Third, in comparison with a naive classifier, such as the mean, client-based variables can have a significant influence on LGD precision. Moreover, in the fractional regression method, the impact of the client-based variables is the same as that of the contract variables. For the regression tree method, the impact of the client-based variables is half that of the contract variables.

We conclude that incorporating information about the contract owner's behaviors plays a crucial role in the predictive accuracy of LGD modeling, and great care should be taken when choosing an appropriate estimation method.

DECLARATION OF INTEREST

The author reports no conflicts of interest. The author alone is responsible for the content and writing of the paper.

REFERENCES

Anderson, R. (2007). *The Credit Scoring Toolkit*. Oxford University Press.

Anolli, M., Beccalli, E., and Giordani, T. (2013). *Retail Credit Risk Management*. Palgrave MacMillan, New York (https://doi.org/10.1057/9781137006769).

Apeh, E., Gabrys, B., and Schierz, A. (2014). Customer profile classification: to adapt classifiers or to relabel customer profiles? *Neurocomputing* **132**, 3–13 (https://doi.org/10.1016/j.neucom.2013.07.048).

Basel Committee on Banking Supervision (2005). Studies on the validation of internal rating system (revised). Working Paper, Bank for International Settlements, Basel. URL: https://www.bis.org/publ/bcbs_wp14.htm.

Basel Committee on Banking Supervision (2017). Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. Report EBA/GL/2017/16, European Banking Authority. URL: https://bit.ly/2Lv24Il.

Bastos, J. (2010). Forecasting bank loans loss-given-default. *Journal of Banking and Finance* **34**(10), 2510–2517 (https://doi.org/10.1016/j.jbankfin.2010.04.011).

Belotti, T., and Crook, J. (2009). Loss given default models for UK retail credit cards. Working Paper 09/1, Credit Research Centre, University of Edinburgh Business School.

Brown, I. L. J. (2012). Basel II compliant credit risk modelling: model development for imbalanced credit scoring data sets, loss given default (LGD) and exposure at default (EAD). PhD Thesis, University of Southampton. URL: https://bit.ly/2OA3jkv.

Chalupka, R., and Kopecsni, J. (2008). Modelling bank loan LGD of corporate and SME segments: a case study. Working Paper 27/2008, Institute of Economic Studies, Charles University, Prague.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., and Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences* **479**, 448–455 (https://doi.org/10.1016/j.ins.2017.12.030).

Gürtler, M., and Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking and Finance* **37**(7), 2354–2366 (https://doi.org/10.1016/j.jbankfin.2013.01.031).

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer.

Huang, X., and Oosterlee, C. (2011). Generalized beta regression models for random loss given default. *The Journal of Credit Risk* **7**(4), 45–70 (https://doi.org/10.21314/JCR.2011.150).

Israel, G. (1992). Determining sample size. Working Paper PEOD6, Institute of Food and Agriculture Sciences, University of Florida, Gainesville. URL: https://bit.ly/3tFBmrC.

Izzi, L., Oricchio, G., and Vitale, L. (2012). *Basel III Credit Rating Systems*. Palgrave Macmillan, London (https://doi.org/10.1057/9780230361188).

Kovach, S., and Ruggiero, W. V. (2011). Online banking fraud detection based on local and behavior data. In *Proceedings of the Fifth International Conference on Digital Society*, Berntzen, L. (ed), pp. 166–171. IARIA/Curran Associates, Red Hook, NY.

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*. Springer (https://doi.org/10.1007/978-1-4614-6849-3).

Leow, M. (2010). Credit risk models for mortgage loan loss given default. PhD Thesis, University of Southampton.

Liu, W., and Xin, J. (2014). Modeling fractional outcomes with SAS. In *Proceedings of the SAS Global Forum 2014 Conference*, pp. 1304–2014. SAS Institute Inc, Cary, NC.

Loterman, G., Brown, I., Martens, D., Mues, C., and Baesens, B. (2012). Benchmarking regression algorithms for loss given default modelling. *International Journal of Forecasting* **28**(1), 161–170 (https://doi.org/10.1016/j.ijforecast.2011.01.006).

Luo, X., and Shevchenko, P. (2013). Markov chain Monte Carlo estimation of default and recovery: dependent via the latent systematic factor. *The Journal of Credit Risk* **9**(3), 41–76 (https://doi.org/10.21314/JCR.2013.166).

Martens, D., Vanthienen, J., Verbeke, W., and Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems* **51**, 782–793 (https://doi.org/10.1016/j.dss.2011.01.013).

Nazemi, A., and Fabozzi, F. (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking and Finance* **89**, 14–25 (https://doi.org/10.1016/j.jbankfin.2018.01.006).

Ozdemir, B., and Miu, P. (2009). *Basel II Implementation: A Guide to Developing and Validating a Compliant, Internal Risk Rating System*. McGraw Hill, New York.

Qi, M., and Zhao, X. (2011). Comparison of modeling methods for loss given default. *Journal of Banking and Finance* **35**(11), 2842–2855 (https://doi.org/10.1016/j.jbankfin.2011.03.011).

Schuermann, T. (2004). What do we know about loss given default? Working Paper 04-01, Wharton Financial Institutions Center (https://doi.org/10.2139/ssrn.525702).

Starosta, W. (2020). Modelling recovery rate for incomplete defaults using time varying predictors. *Central European Journal of Economic Modelling and Econometrics* **12**(2), 195–225 (https://doi.org/10.24425/cejeme.2020.133721).

Tanoue, Y., Kawada, A., and Yamashita, S. (2017). Forecasting loss given default of bank loans with multi-stage models. *International Journal of Forecasting* **33**(2), 513–522 (https://doi.org/10.1016/j.ijforecast.2016.11.005).

Thomas, L., Mues, C., and Matuszyk, A. (2010). Modelling LGD for unsecured personal loans: decision tree approach. *Journal of the Operational Research Society* **61**(3), 393–398 (https://doi.org/10.1057/jors.2009.67).

Tong, E., Mues, C., and Thomas, L. (2013). A zero-adjusted gamma model for mortgage loss given default. *International Journal of Forecasting* **29**(4), 548–562 (https://doi.org/10.1016/j.ijforecast.2013.03.003).

West, D. (2000). Neural networks in credit scoring models. *Computers and Operations Research* **27**(11), 1131–1152 (https://doi.org/10.1016/S0305-0548(99)00149-5).

Witzany, J., Rychnovsky, M., and Charamza, P. (2012). Survival analysis in LGD modeling. *European Financial and Accounting Journal* **7**(1), 6–27 (https://doi.org/10.18267/j.efaj.12).

Yao, X., Crook, J., and Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research* **263**(2), 679–689 (https://doi.org/10.1016/j.ejor.2017.05.017).

Zhang, J., and Thomas, L. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting* **28**(1), 2094–215 (https://doi.org/10.1016/j.ijforecast.2010.06.002).

European Journal of Operational Research 292 (2021) 1187–1199

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Loss given default decomposition using mixture distributions of in-default events

Wojciech Starosta

Chair of Econometrics, Institute of Econometrics, Department of Economics and Sociology, University of Lodz, POW 3/5 Street, 90-255 Lodz, Poland

ARTICLE INFO

Article history:
Received 18 May 2020
Accepted 20 November 2020
Available online 28 November 2020

Keywords:
Risk management
Loss given default
Probability of cure
Probability of write-off
Recovery rate

ABSTRACT

Modeling loss in the case of default is a crucial task for financial institutions to support the decision making process in the risk management framework. It has become an inevitable part of modern debt collection strategies to keep promising loans on the banking book and to write off those that are not expected to be recovered at a satisfactory level. Research tends to model Loss Given Default directly or to decompose it based on the dependent variable distribution. Such an approach neglects the patterns which exist beneath the recovery process and are mainly driven by the activities made by collectors in the event of default. To overcome this problem, we propose a decomposition of the LGD model that integrates cures, partial recoveries, and write-offs into one equation, defined based on common collection strategies. Furthermore, various levels of data aggregation are applied to each component to reflect the domain that influences each stage of the default process. To assess the robustness of our approach, we propose a comparison with two benchmark models on two different datasets. We assess the goodness of fit on out-of-sample data and show that the proposed decomposition is more effective than state-of-the-art methods, maintaining a strong level of interpretability.

1. Introduction

Loss Given Default (LGD), determined in a workout approach, is a high uni-, bi-, or even multi-modally distributed credit risk parameter that is calculated in the advanced internal rating-based (AIRB) approach. It makes estimation with well-known parametric and non-parametric methods difficult, as even highly flexible approaches like beta regression, which are capable of reproducing the density of LGDs, are not able to reduce errors or explain variability to a sufficient degree. Nowadays, more and more two-stage modeling frameworks are presented in the literature as a remedy for discriminating no-loss cases and full-loss cases, which translates to two modes of the LGD distribution. However, such an approach has significant weaknesses. Dividing the LGD into a mixture distribution of $LGD = 0$, $LGD = 1$, and a distribution center does not take into account the recovery pattern which exists beneath the modes mentioned above. A no-loss case can be the result of (a) all due amounts being repaid by the client with little or no support from the collection department, (b) a new repayment schedule agreement between the client and the bank, or (c) collateral seizure which covers all remaining debt. These three paths are mainly driven by different patterns, which are put in the same basket by simply placing $LGD = 0$ cases in it. Secondly, a typical LGD distribution is concentrated in zero and one, but there are many values close to 0 (but not exactly 0, due to discounting) for clients who repaid the full amount some time after default, or close to 1 for clients who paid just once or who did not pay at all, which implies LGD over 1 when the direct and indirect costs of the workout process are added. Truncating observations in 1 and 0 leads to the removal of potentially useful data that are close to the peaks of the histogram and have the same recovery pattern. Thirdly, full-loss does not translate into a write-off, as write-offs can be done at any time during a default when the bank does not expect any further recoveries (however, previous repayments are possible and even frequent) or potential recoveries are threatened by a lack of possibility to sell the collateral. The collateral impact on the recovery rate (RR) is different when we expect the client to capitalize it from when the obligation is put on sale to a third-party.

Our first contribution to the field of LGD modeling is a new LGD decomposition using a mixture distribution of events that are connected to recovery patterns. We expect that this will lead to more precise and robust estimates, mainly because better predictors fit the modeled phenomenon. Additionally, the proposed model can be useful in setting collection strategies and performing responsive risk management actions, as selected events come from the collection policy itself. We consider three patterns client could follow: cures, write-offs and partial recoveries (not cures, nor write-offs, but cases connected with collateral seizure, partial write-off,

E-mail address: w.starosta@wp.pl

https://doi.org/10.1016/j.ejor.2020.11.034
0377-2217/

restructuring, etc.). First, we estimated the probability of cure, defined based on activities widely used in the banking sector to manage relationships with clients in default.[2] Secondly, for non-cured cases, we defined and estimated the probability of write-off. Finally, we prepared models conditional on the recovery pattern for cures, write-offs, and partial recoveries. The timeline connected with these episodes is essential, which translates to the assumption that the events do not overlap. It means that cure is possible only for cases that meet the definition of returning to the performing portfolio without major actions by the collection department, including write-off. Then, defaults that are not cured and not written-off are treated as partial recoveries. Finally, a write-off is possible only for the not cured part of the portfolio. The severities of the losses are modeled based on a combination of probabilities of the events and conditional recovery rate estimated on cases belonging to a specific path. Properly defining cure and write-off events allows us to prepare each part of the model based on an inherent set of information, which is our second contribution. To be specific, we define cure as a component that is connected directly to the client, write-off as mainly outside the bank part of the process and the rest as contract and contract owner composition, as in standard LGD models presented in other studies. This allows us to link each component with different levels of data aggregation and origination, and it reflects future collection department actions, which will finally lead to more precise LGD estimates.

The rest of the paper is structured as follows. Section 2 presents a literature review of both the factors that influence LGD as well as the two-stage models previously used in modeling the aforementioned risk parameter. Section 3 defines each event used in LGD decomposition, the level of the information used and the combining mechanism of the predictions. Section 4 presents a brief description of the estimation methods and the process of assessing the quality of the model. Section 5 describes the dataset of the bank loans used in this study. In Section 6, empirical evidence is demonstrated, including interpretation of the parameters and comparison of the model performances. Section 7 concludes.

2. Literature review

2.1. LGD predictors

This paper relates to the literature of retail LGD estimation in the case of explanatory variables in several ways. First, we analyze contract-based information. Collateral influence, which is supposed to be the main risk driver for secured loans, has been confirmed in Dermine and de Carvalho (2006) and Krüger and Rösch (2017). Loan to Value (LTV), which is inextricably linked to collateral, has also been the subject of many studies, and its effect was checked both for the haircut model (Leow, 2010) and LGD (Brown, 2012). Another widely used variables are time on books (Tong, Mues, and Thomas, 2013, or Yao, Crook, and Andreeva, 2017), loan size (Brown, 2012 or Do, Rösch, and Scheule, 2018) and interest rates (Leow, 2010). Additionally, in the set of LGD predictors we can also distinguish arrears (Brown, 2012), loan tenor (Zhang, Thomas, 2012), or debt to value (Tong et al., 2013). The results of all analyses varied, as different samples and different methods were used to build the models.

Moving on to information about the contract owner, first, the study of Belotti and Crook (2010) needs to be mentioned, as they used various types of client data to build their LGD model, like time with bank, income, or credit bureau score. Each proved its significance and intuitive influence on the dependent variable. The application score was one of the main risk drivers for Thomas, Mues, and Matuszyk (2010), where it was used to predict LGD value for those with $LGD > 0$, and in Tanoue, Kawada, and Yamashita (2017) where higher scores decreased the probability of full recovery. Brown (2012) was able to find a relationship with the time at the bank and employment status. Yao et al. (2017) also used client characteristics, such as binary variables that indicate a customer returning to order, or another indicating a customer being on a repayment plan. Client specific information is also recommended in other studies, such as Ozdemir and Miu (2009, p. 17) or Anolli, Beccalli, and Giordani (2013, p. 102). It might include the age of the customer, years in their current job, marital status, and the borrower's rating.

The final set of predictors consists of macroeconomic indicators. Their influence is ambiguous, as most LGD variability, especially in the case of retail, is explained by banks' idiosyncratic characteristics (mainly changes in collection policy), which makes a connection with the economic cycle limited. One of the first approaches was made by Qi and Yang (2009), who studied the recovery rate for residential mortgages. In addition to contract variables, they tried to find a relationship between House Price Index (HPI), House Price Ratio (HPR), and the economic downturn indicator. At the same time, Belotti and Crook (2010) indicate the significance of bank interest rates and unemployment level. Leow (2010) checked a wide range of macroeconomic variables influence on different components of the model. For example, the unemployment rate and HPI decrease the probability of repossession, and GDP growth, with purchasing power growth affecting the haircut model. The direct impact of account characteristics vs. macroeconomic state was studied in Tobback, Martens, Van Gestel, and Baesens (2014), where a set of 11 macro indicators was used to predict recovery rates. According to Yao et al. (2017) including the monthly unemployment rate, monthly Consumer Price Index (CPI), and monthly HPI increases the R^2 measure only slightly. However, the direction of some covariates was discordant with the previous findings. Finally, in the cross-regional work by Betz, Kellner, and Rösch (2018), systematic effects were found to be region- and macroeconomic cycle specific.

2.2. LGD models

LGD distribution typically has two modes at zero and one (or just one in the unimodal case). Previous research studied various kinds of one-stage models (see Tong et al., 2013; Van Berkel, Siddiqi, 2012), but in comparison to two-stage model, in many cases, their ability to explain LGD variance was limited. Here, we focus on the literature concerning the two-stage approach as the basis for further considerations. To the best of our knowledge, the first paper to introduce such an idea was Belotti and Crook (2010). They proposed a decision tree model that uses logistic regression to model the special cases for full-loss and no-loss ($LGD = 1$ and $LGD = 0$) as binary classification problems. Then, linear regression model with OLS estimator was used to model cases in the middle ($0 < LGD < 1$). They suggest that there are special conditions that would make a customer repay either the full amount or nothing, rather than just a portion. The forecast from this approach was set as the expected value from the three sub-models: $(1 - p_0)(p_1 + (1 - p_1)LGD_i)$ where p_0 was the probability of no-loss, p_1 was the probability of full-loss, and LGD_i was a loss estimated by a regression model. The accuracy of the model was checked on a hold-out sample, and it was found that simple linear regression performed better than the two-stage model in the case of the R^2 measure. A similar approach was suggested in Thomas et al. (2010). The data were split into cases where $LGD \leq 0$ and $LGD > 0$. A logistic model was built to separate these two groups, and a linear regression was selected to estimate the values of LGD

[2] Debt collection activities and strategies can be found, inter alia, in Cornejo (2007, p. 191) or in Finlay (2009, p. 207).

belonging to $0 < LGD < 1$ interval. The next study to consider was the large research conducted by Loterman, Brown, Martens, Mues, and Baesens (2012). They proposed two alternatives to two-stage modeling. The first was the use of logistic regression to estimate the probability of LGD ending up at one of the peaks, 0 or 1. Then, a linear or non-linear technique was built using only the observations beyond the peaks. An LGD estimate was then estimated as the weighted average of the LGD at the peak and the estimate produced by the second-stage model, where probabilities were the weights. The second proposition involved building an OLS model and then, in the second stage, estimating the residuals from a linear regression using a non-linear regression model. The estimate of the residuals was added to the OLS estimate to obtain a more accurate LGD prediction. The least squares support vector classifier (LS-SVC) was incorporated into the two-stage structure by Yao et al. (2017). The framework was similar to that by Belotti and Crook (2010), with the following notation: $RR = P_a \cdot (P_c + P_b \cdot RR_{reg})$, where RR is an expected recovery rate, and $P_a = P(RR > 0)$, $P_b = P(0 < RR < 1 | RR > 0)$, $P_c = P(RR = 1 | RR > 0)$ and RR_{reg} denotes the predicted value from the regression model. LS-SVC gave better results than the logistic regression for the classification part, but there were no significant differences between models predicting RR_{reg}. The conclusion was that LS-SVC combined with OLS regression gave the best results compared to both the one-stage method and different combinations of two-stage methods. The indirect approach for in-default exposures was also proposed by Joubert et al. (2018). They use survival analysis to produce estimates for cure and write-off probabilities, and the haircut regression model for the severity part. Finally, a three-step approach was proposed by Do et al. (2018) where a joint probability framework for a default event, cure event ($LGD = 0$), and non-zero LGD event was set up. In the dependent model (assuming dependence between stochastic processes), all the components were derived in one regression. In the independent structure (assuming independence between stochastic processes), three separate regressions were used (probit for both default and cure, and OLS for non-zero LGD). Furthermore, the OLS model for LGD was prepared. Verification was conducted on time-variant samples, and based on RMSE, they found that the dependent structure outperforms the independent structure and OLS.

Additionally, more and more attention is put on the interpretability of the LGD models. Modeling non-linearities by SVM can lead to high accuracy, but the comprehensibility of such an approach is limited (Martens, Baesens, Van Gestel, and Vanthienen, 2007). It is particularly important when modeling risk parameters, as both the users and the regulators should be able to understand the logic behind the predictions. Also, consistency with existing domain knowledge should be maintained (see Martens, Vanthienen, Verbeke, and Baesens, 2011 or Maldonado, Bravo, Lopez, and Perez, 2017). In a financial context, all of these are inevitable from the perspective not only of model developers but also internal validators, the management responsible for the application, model users on daily basis (credit analyst, debt collection analysts, etc.), conduct regulators, and prudential regulators (see Bracke, Datta, Jung, and Sen, 2019 for further discussion). According to the BCBS (Basel Committee on Banking Supervision, 2017) (a) the estimation methods should be appropriate to institutions activities and type of exposures, (b) institutions should be able to justify the theoretical assumptions underlying those methods, (c) the methods should be consistent with collection and recovery policies, and (d) should take into account possible recovery scenarios along with legal environment.

In this paper, we expand the two-stage approach by modeling cure and write-off probabilities, but also severities connected with each default ending status. We shed light on cure and partial recovery severities for the first time in this type of study. The presented methodology can be used for any type of exposure and banking activity, but only if there is a possibility to properly connect the collection strategy to the cure and write-off events. Our approach is consistent with postulates concerning comprehensibility, understandable logic behind predictors, and existing domain knowledge. Additionally, we add to the above (a) an extended definition of a cure event connected with widely used actions taken by the collection department, (b) a combining mechanism that results from LGD decomposition based on the aforementioned cure and write-off events, (c) a justification for using an inherent set of variables to estimate different components based on defined events, and (d) connecting each component with a particular moment in the recovery process. Such a framework not only leads to the efficient management of in-default exposures, but it also helps with facilitating capital allocation on performing assets.

3. LGD decomposition

Taking into consideration both interpretability and consistency with recovery policies, we propose a decomposition of the LGD parameter, derived from widely known collection department activities driving the level of the recovery rate and the timeline of the defaulted asset in the state of default (cf. Finlay, 2009). The standard path can be described as in Fig. 1.

We define a cure event (denoted as $s_i = 1$, where i indicates consecutive observation) as a default (a) which returned to the performing portfolio, (b) with no major action taken by the collection department, and (c) where there was no write-off. This construction allows the default to be marked as a cure only when the client exited the default status with little or no help from the bank. So, the first path is based on the client and his behavior, which prevents the default from going further into the process. Collateral information is not compulsory at this point (except for the information that credit itself is secured), as using collateral automatically moves the client into stage 2 ($s_i = 2$).

Partial recovery results from numerous actions that can take place in the collection department, excluding write-off (like collateral seizure, restructuring, partial write-off, termination of the agreement, litigation, etc.). They mostly depend on the contract information (Days Past Due (DPD), exposure amount, due principal, etc.), collateral (type, value, construction year, etc.), and the client (relation time with the bank, reachability, age, etc.). The important thing to note is that the LGD for stage 2 can still be 0 in the case of collateral seizure, which covers all debts or restructuring with a new repayment schedule when the client and the bank change the agreement, but future inflows are secured. Still, partial write-offs may occur, where a fraction of the exposure is treated as lost, although the rest is paid in full. This path aggregates the contract, collateral, and contract owner information to estimate the recovery, bounded between cure and write-off. If the bank's demands are met, the default ends. If not, the contract is moved to stage 3 ($s_i = 3$).

A write-off, treated as an event that formally recognizes that an asset no longer has value, does not imply $LGD = 1$ in our notation. This action can be performed at any moment of the process, excluding cured cases. Except for when there was no repay at all, a write-off can be executed when (a) there is still a significant unrepayable amount of the exposure after collateral realization, or (b) the client repaid only a fraction of the exposure and then stopped. What is more, the bank can complete a write-off by selling the debt to a third-party. These amounts received from a specialized company should also be treated as positive cash flow in LGD determination. In stage 3 ($s_i = 3$), we state that it is mainly contract information that is required. Collateral, if applicable and the option of voluntary repayment had been exhausted, was used in stage 2. The price that can be obtained from the third-party also depends

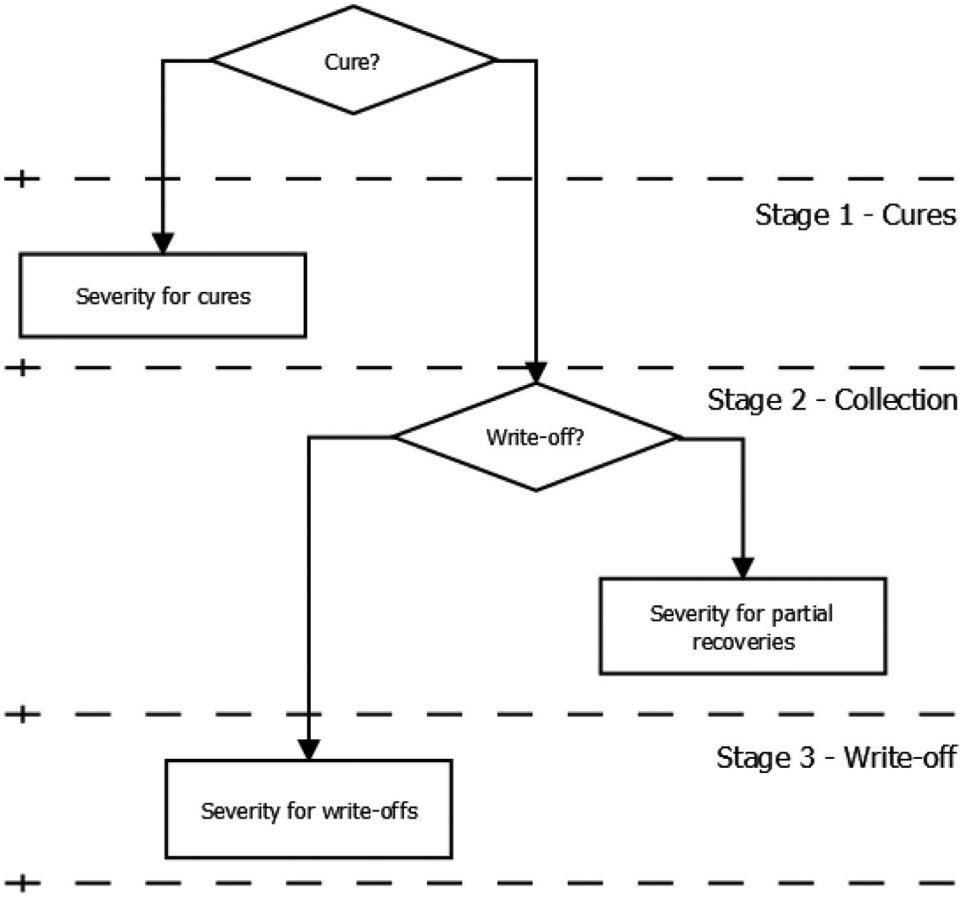


Fig. 1. Stages (s) of the asset in a state of default.

Table 1
Connection between the origination of the explanatory variables and each component of the LGD decomposition.

Stage 1: Cures	Stage 2: Collection department	Stage 3: Write-offs
Aggregation level: Client - whole product structure, - behavioral data, - socio-demographic data.	Aggregation level: Contract - contract-level data like tenor, Exposure at Default (EAD), etc., - collateral information, - contract owner information connected to the specific contract.	Aggregation level: Contract - contract-level data like tenor, EAD, etc.
Macroeconomic indicators		

primarily on the characteristic of the contract (DPD, exposure, previous payments, etc.).

The stages of the recovery process are summarized in Table 1. We assume that the macroeconomic environment can influence all components but that disparate variables can affect each stage.

Such distinctions leads to the framework in which five models are required: (a) the probability of the cure estimation operating on cures and non-cures, (b) the probability of the write-off estimation operating on non-cures, (c) the recovery rate estimation for cured cases operating on cures, (d) the recovery rate estimation for write-offs operating on write-offs, (e) and the recovery rate estimation for partial recoveries operating on not cures nor write-offs. In this framework, the cures (stage 1), write-offs (stage 3), and partial recoveries (stage 2) are modeled separately and sequentially with the use of the methods presented in Section 4. Here, we introduce the notation for the probability of cure:

$$Pr(s_i = 1) = \varphi_{p(cure)}\left(\boldsymbol{x}_{1,\boldsymbol{i}}\right) \tag{1}$$

probability of partial recoveries:

$$Pr(s_i = 2|s_i \neq 1) = 1 - P(s_i = 3|s_i \neq 1) \tag{2}$$

probability of write-off:

$$Pr(s_i = 3|s_i \neq 1) = \varphi_{p(write-off)}\left(\boldsymbol{x}_{2,\boldsymbol{i}}\right) \tag{3}$$

expected RR for cures:

$$E(RR_i|(s_i = 1)) = \omega_{cure}\left(\boldsymbol{x}_{3,\boldsymbol{i}},\right) \tag{4}$$

expected RR for partial recoveries:

$$E(RR_i|(s_i = 2, s_i \neq 1)) = \omega_{partial}\left(\boldsymbol{x}_{4,\boldsymbol{i}}\right) \tag{5}$$

expected RR for write-offs:

$$E(RR_i|(s_i = 3, s_i \neq 1)) = \omega_{write-off}\left(\boldsymbol{x}_{5,\boldsymbol{i}}\right) \tag{6}$$

where $\varphi_{p(cure)}$, $\varphi_{p(write-off)}$, ω_{cure}, $\omega_{partial}$, $\omega_{write-off}$ are regression or classification functions chosen to minimize the model cost functions or classification performance metrics (eg., likelihood func-

tion or residual sum of squares); $\boldsymbol{x}_{1,i}, \boldsymbol{x}_{2,i}, \boldsymbol{x}_{3,i}, \boldsymbol{x}_{4,i}, \boldsymbol{x}_{5,i}$ denotes explanatory variables[3] for the i th exposure in Eqs. (1), (3), (4), (5) and (6), $s_i \in \{1, 2, 3\}$ denotes the stage for the i th exposure; i indicates the consecutive observation. In order to complete the expected ultimate recovery rate (EURR), a parallel combination is used in a manner consistent with Loterman et al. (2012):

$$\begin{aligned} EURR &= E(RR_i) = E(RR_i|s_i = 1)Pr(s_i = 1) + E(RR_i|s_i \neq 1)Pr(s_i \neq 1) \\ &= E(RR_i|s_i = 1)Pr(s_i = 1) + Pr(s_i \neq 1)(E(RR_i|s_i = 2, s_i \neq 1)Pr(s_i = 2|s_i \neq 1) \\ &\quad + E(RR_i|s = 3, s \neq 1)Pr(s_i = 3|s_i \neq 1)) = \omega_{cure}\left(\boldsymbol{x}_{3,i}\right)\varphi_{p(cure)}\left(\boldsymbol{x}_{1,i}\right) \\ &\quad + \left(1 - \varphi_{p(cure)}\left(\boldsymbol{x}_{1,i}\right)\right)\left(\omega_{partial}\left(\boldsymbol{x}_{4,i}\right)\left(1 - \varphi_{p(write-off)}\left(\boldsymbol{x}_{2,i}\right)\right)\right. \\ &\quad \left. + \omega_{write-off}\left(\boldsymbol{x}_{5,i}\right)\varphi_{p(write-off)}\left(\boldsymbol{x}_{2,i}\right)\right) \end{aligned} \tag{7}$$

By adopting such a mechanism we can assess the influence of each component directly and in an intuitive way that reflects the activities performed by the collection departments. What is more, the collection strategy can be based directly on the results. Cases with a high probability of cure can be treated with less attention, which can release the resources for more complicated defaults. Recovery from a write-off, on the other hand, can indicate which contracts should be removed from the banking book at the first possible term and which ones it is worth working with a little more. The first part of the equation is responsible for quantifying the fraction of the recovery in the case of a cure event. Secondly, for the non-cured part, two more paths are possible. First, assess the fraction of the recovery in the case of a write-off; second, assess another fraction when there is neither a cure nor a write-off event. For each default event, all components need to be estimated and combined with the rest. The result is that the whole equation gives the recovery rate, but it can be easily transformed into LGD as $1 - EURR$.

4. Methods

Within this section, we describe the set of methods used to estimate the functions $\varphi_{p(cure)}$, $\varphi_{p(write-off)}$, ω_{cure}, $\omega_{partial}$, $\omega_{write-off}$ in (1), (3), (4), (5), (6), respectively and assess their quality. We use one parametric and one non-parametric technique to model each component. When it comes to performance metrics, the final LGD predictive power evaluation is the ultimate goal, but each component is also examined with a suitable metric. Discrimination indicates how well the model ranks the observations. Calibration refers to the ability to provide precise estimates (as close to the observed values as possible). A well-calibrated model should always discriminate, but good discrimination ability does not imply good calibration (Loterman et al., 2012). To observe if the models are both discriminative and precise, each component is checked in both dimensions.

4.1. Ordinary least squares (OLS) regression

Ordinary least squares regression is the most common technique used when estimating LGD; initially, it was the leading method, but nowadays, it is used mostly as a benchmark. OLS model as a primary approach can be found in Covitz and Han (2004) or in Acharya, Bharath, and Srinivasan (2007), among others, while for its use as a supportive approach, the studies of Qi and Zhao (2011) or Belotti and Crook (2010) may serve as an example. However, it is worth pointing out that in the study by Belotti and Crook, OLS regression outperforms more sophisticated methods like the Tobit regression model or regression tree. In our framework we use the simplest version of the OLS regression in the form of (considering RR for cures as an example):

$$\omega_{cure}(\boldsymbol{x}) = \hat{\beta}_0 + \sum_i \hat{\beta}_i x_{3,i} \tag{8}$$

where $\boldsymbol{x}_3 = [1,\ x_{3,1}, \ldots, x_{3,k}]'$ represents the vector of covariates, and $\hat{\beta}_i$ represents the OLS estimates of parameter at $x_{3,i}$.

4.2. Logistic regression (LR)

Logistic regression is the first choice when it comes to modeling binary variables. It is widely used in estimating Probability of default (PD) (Anderson, 2007, p. 42), and it is becoming increasingly popular in the case of LGD when binary events are being modeled. In our approach, logistic regression is used to model cure and write-off events in the standard form (for cure probability as an example):

$$\varphi_{p(cure)}(\boldsymbol{x}) = 1/\left(1 + exp\left(-\left(\hat{\beta}_0 + \sum_i \hat{\beta}_i x_{1,i}\right)\right)\right). \tag{9}$$

It is assumed that each event takes only two values (0 or 1), which translates into cure/non-cure and write-off/not written-off cases.

4.3. Classification and regression trees (CART)

Tree-based methods recursively partition the original sample into smaller subsamples and then fit a model in each one, and they are one of the most commonly used methods in LGD estimation nowadays (see Loterman et al., 2012 or Nazemi, Fatemi Pour, Heidenreich, and Fabozzi, 2017). CART can be used both for classification and regression problems. The initial idea is the same and involves reducing impurity by finding the best split which minimizes squared-errors in case of regression or Gini index in case of classification. After determining the first split, the procedure is repeated in all regions (Hastie, Tibshirani, and Friedman, 2008, p. 307). We use CART to estimate events probability (1) and (3), and recovery rates (4), (5) and (6) using the standard set of tuning rules (see Qi and Zhao, 2011 or Nazemi et al., 2017) described in Section 6.1 in detail.

4.4. Support vector machines (SVM)

SVMs were proposed by Vapnik (1995) and due to their ability to solve highly non-linear problems, they have become more popular when estimating LGD (see Tobback et al., 2014 or Yao et al., 2017). With this technique the input vectors ($\boldsymbol{x_i}$) are mapped into a high dimensional space using one of the selected kernels. By means of this kernel mapping the problem is also transformed from non-linear into linear settings to provide more accurate predictions. As in CART, this method can be used both for classification (in the form of least squares support vector classifier (LS-SVC)) and regression (the least squares support vector regression (LS-SVR)), but in this research only LS-SVC will be adopted. In our study, we uses a similar approach to the one presented in Yao et al. (2017), where the LS-SVM classification function (LS-SVC) is given as:

$$f(x) = sgn\left(\sum_{i=1}^{N} \hat{\alpha}_i y_i \boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x}) + \hat{\beta}_0\right) \tag{10}$$

where $\boldsymbol{x}$ is the current vector of inputs, $\boldsymbol{x}_i$ are ith covariates and $y_i \in \{-1, 1\}$ determines the classes for cure events i.e. $y_i = 1$ if $s_i = 1$ and $y_i = -1$ if $s_i \neq 1$ in the training sample, $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x})$ is the kernel, $\hat{\alpha}_i$ is the estimated value of the Lagrange multiplier associated with LS-SVC and $\hat{\beta}_0$ denotes the intercept term. Among possible kernels we used the Gaussian radial basis function which is

[3] A full list of explanatory variables for each component is a part of Section 6.2.

Table 2
Performance metrics used in the study.

Metric	Description	Type
RMSE	The Root Mean Squared Error measures the distance between the actual and predicted values.	Calibration
MAE	The Mean Absolute Error takes a mean absolute difference instead of squared error like in RMSE.	Calibration
DM	The Diebold-Mariano test determines whether forecasts are significantly different.	Calibration
GINI	Calculates the area between the curve and the diagonal in the Lorenz curve. Used to assess both LGD and binary events.	Discrimination
ρ	Spearman's correlation coefficient measures the degree of the relationship between the actual and predicted values using a monotonic function.	Discrimination

of the form:

$$\boldsymbol{K} = exp\left(-\sigma \|\boldsymbol{x_i} - \boldsymbol{x}\|^2\right) \tag{11}$$

where σ is a scaling parameter. Following Platt (1999) approach, we map SVM outputs into probabilities by applying sigmoid map:

$$\varphi_{p(cure)}(\boldsymbol{x}) = \frac{1}{1 + \exp\left(\hat{a} f(x) + \hat{b}\right)} \tag{12}$$

where $\hat{a}$, $\hat{b}$ are maximum likelihood estimate of the sigmoid function parameters obtained on the training data set.

4.5. Performance metrics

In this research, we use a mix of metrics to assess both discriminatory power and the calibration of each model. Following the studies performed by Loterman et al. (2012), Zhang and Thomas (2012) and Yao et al. (2017), we list the selected set in Table 2. Additionally, we add the Diebold-Mariano test to check if the difference in error measures is statistically significant. The calibration measures are expected to be as low as possible. On the other hand, the GINI Index and Spearman Correlation coefficient should be maximized to get the best model.

5. Data

Our study utilizes two consumer credit datasets, one with mortgage loans and, another with cash loans, provided by a Middle-European AIRB bank. The first set consists of 6798 defaults, second comprise 68 129 defaults. The observations come from January 2010 to December 2017. All direct and indirect costs are added to the dependent variable, so the observed recovery rate can be less than zero for some observations, but we still include them in the final data set. Figs. 2 and 3 present a histogram of the recovery rates for the both samples. The two segments have modes near 1. The cash loan segment has a more U-shaped distribution in contrast to the J-shaped mortgage loan distribution.

We divided each segment into five sub-samples equivalent to five functions $\varphi_{p(cure)}$, $\varphi_{p(write-off)}$, ω_{cure}, $\omega_{partial}$, $\omega_{write-off}$, according to the following cure definition: if there was no termination of the agreement, no collateral realization, no write-off, and the contract returned to the performing portfolio, then the case is treated as a cure. Otherwise, it is a non-cure. The write-off event is a cancelation from an account of a bad debt. Descriptive statistics for each event are presented in Table 3.

Table 3
Descriptive statistics of the recovery rate for each event (cures, partial, write-off). P1, P5, P50, P95, P99 denotes consecutive percentiles.

Segment	Statistic	Cures	Partial recoveries	Write-offs
Mortgage Loan	Mean	95.55%	63.70%	39.16%
	P1	75.00%	0.00%	−2.00%
	P5	84.00%	18.00%	2.00%
	P50	98.00%	60.29%	35.00%
	P95	100.00%	100.00%	92.00%
	P99	100.00%	100.00%	100.00%
Cash Loan	Mean	94.39%	50.61%	13.84%
	P1	52.32%	−2.17%	−2.33%
	P5	78.73%	8.73%	0.00%
	P50	98.12%	38.14%	7.75%
	P95	100.00%	99.97%	49.94%
	P99	100.00%	100.00%	83.74%

For cured cases, the mean of RR is relatively high; however, there are exceptions, with RR as low as 75% for mortgage loans and 52% for cash loans. This is connected mainly with a discounting issue for longstanding defaults. For the second event (partial recoveries), recoveries are significantly smaller, about 30 percentage points for mortgages and 45 percentage points for cash loans. Finally, write-offs are characterized by the lowest RR, but still, there are some cases in which selling to third-party complements the recovery obtained from the client and collateral to 100%.

As stated before, a different set of variables is assigned to each event (except for macroeconomic indicators for which the set is constant). However, as the LGD is estimated at the contract level, we propose the following aggregation levels. For the probability of write-off and recovery rate for contracts that are written-off, we are considering only contract information, and we do not need aggregation schema. The first operation comes with the recovery rate for cured contracts and the recovery rate for partial recoveries. We use a mix of contract and client information, so the contract is treated as before, but for instances where there is more than one contract owner, we need to prepare aggregation schema. In the simplest version, we are considering three aggregation functions, which are minimum, maximum, and mean. Finally, for the probability of cure, the data composition is run through the analyzed contract, through the owners of this contract, and eventually, all contracts that are owned by these clients who are connected to the original contract. The aggregation schema takes place via minimum, maximum, mean, and sum (depending on the analyzed variable). Fig. 4 summarizes all possibilities.

The samples include 24 candidate predictors for the probability of cure, as well as 14 for the probability of write-off and RR for partial recoveries, and 12 for RR for write-offs and RR for cures. Additionally, the missing values are imputed with the mean.[4] Finally, our set of variables consists of account-level indicators (such as the principal amount, tenor, time on book, etc.), client-oriented indicators (client age, relationship time, number of credit cards, etc.), collateral description (value and type), and macroeconomic indicators (Gross Domestic Product, Consumer Price Index, average wages, etc.). Most have been demonstrated to be important LGD predictors in previous studies. Tong et al. (2013) studied balance at default, time on books, valuation of collateral, and tenor. Belotti and Crook (2012) showed the influence of relationship time, the number of credit cards, time on books, balance at default, and bank interest rates on LGD. Similar variables were used by Yao et al. (2017), and statistical significance was reported for, inter alia, time on books, the Consumer Price Index, and the num-

[4] Most missing information (8%) concerned client age.

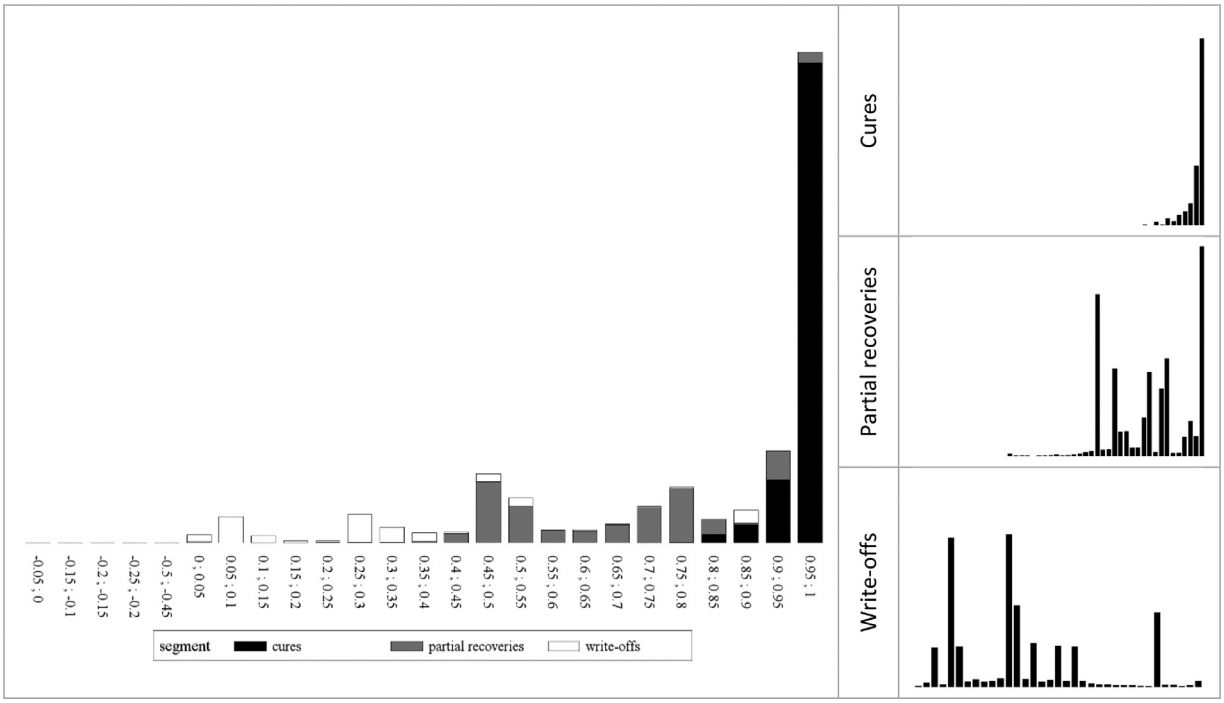


Fig. 2. Distribution of recovery rates for mortgage loans broken down into stages. For the right panel, the x-axis values were removed for clarity.

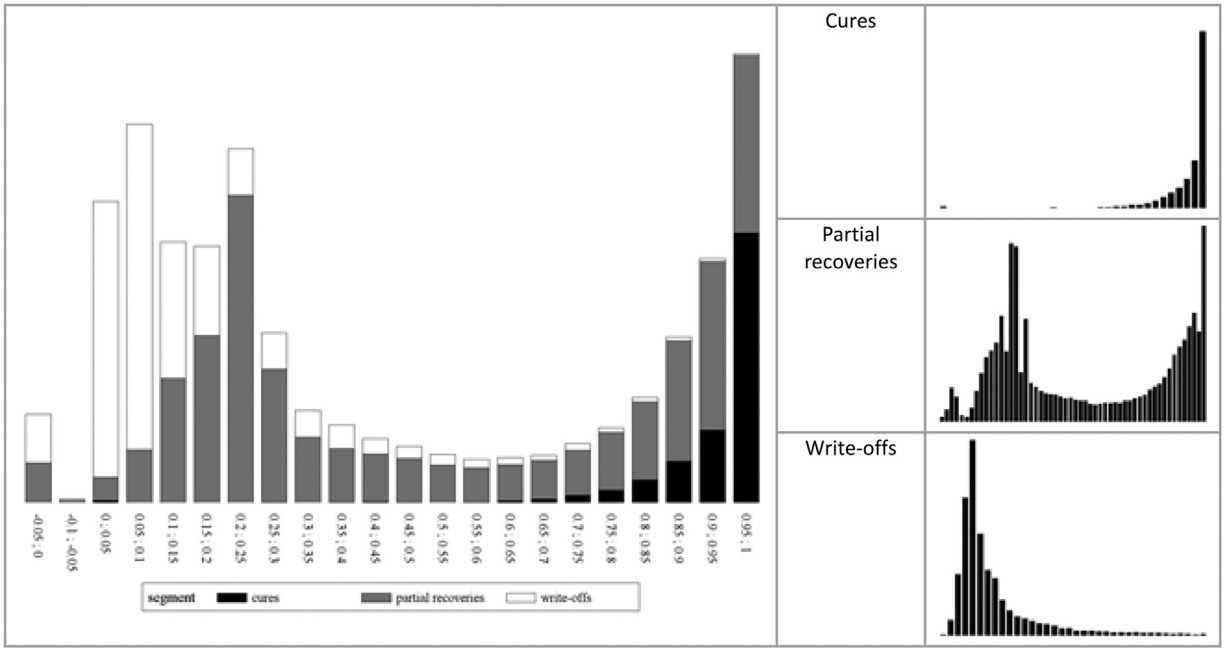


Fig. 3. Distribution of recovery rates for cash loans broken down into stages. For the right panel, the x-axis values were removed for clarity.

ber of months in arrears. Finally, in Tanoue et al. (2017), collateral quota and EAD were the main risk drivers. In this context, the list of variables used in the study is consistent with the previous research.

There are also some new variables never investigated before, like the flag of different installment plans (equal or decreasing) or the flag of EAD higher than the requested amount (valid in the case of revolving loans or credits denominated in foreign currencies). Additionally, to the best of the author's knowledge, the influence of the yield of bonds and stock exchange indices is checked for the first time in retail LGD. The full list of variables used in the study is a part of Section 6.

6. Experiment

In this section, we first describe setting up the experiment, then present the decomposed models and asses their quality. We detail the dataset preprocessing, set a benchmark, describe the process of tuning hyperparameters (where applicable), and evaluate each composition to rank them in terms of the final LGD prediction.

6.1. Experimental set-up

When analyzing LGD, the problem of collecting data while the workout process is in progress arises. Using only completed work-

Table 4
Probability of cure model in decomposed approach.

	Logistic Regression		Classification Tree	
	Mortgage Loans	Cash Loans	Mortgage Loans	Cash Loans
(Intercept)	3.2442***	3.6819***		
Principal	−4.41E-7***	−2.52E-6***		
Tenor	0.0013***	0.0120***		9.34
Flag of possessing contract in currency other than domestic	0.0151***	−0.6953***		
Flag of possessing contract with decreasing installment plan	−0.2250*			
Number of co-applicants	−1.2580***	−0.5006***	8.25	
Flag of EAD higher than the requested amount on any client contract	0.0334***	0.3966***		3.54
Due principal	4.06E-6***		5.36	3.72
Due interest	1.40E-5***	7.90E-5***	14.54	4.01
Due amount			10.41	3.37
Age of the client	0.0300***	0.0135***		6.17
Relationship time	−0.0075***	−0.0063***		3.50
DPD	0.0047***	0.0108***	23.48	17.09
Number of mortgage loans			13.51	
Number of credit cards		0.0031*	2.84	
Number of revolving loans		0.0148***		1.18
Number of cash loans				4.69
Sum of owned contracts				3.55
Contractual IR#		1.9612***	13.76	5.88
Requested amount			5.03	1.68
Months on book		−0.0165***	2.83	10.59
EAD				4.49
Export				2.44
CPI	0.2713***	0.3578***		2.05
Yield on 10-year bonds		−1.8487***		4.91
Yield on 5-year bonds		1.7451***		
WIBOR 1Y	−0.6423***	−0.6494***		4.33
LIBOR 3M		−0.4882***		
Average wages		−0.0997***		1.96
Warsaw Stock Exchange Index	−0.0001*			
GINI	.4362	.4633	.5293	.5089

#Interest Rate.
*** indicates 1% level of significance, ** 2.5% and * 5%. The importance of the classification tree variables are rescaled to total 100.

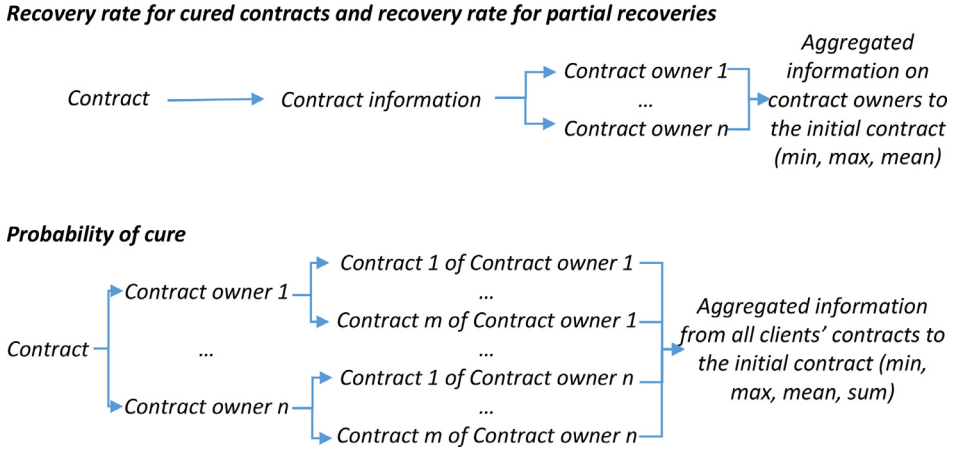


Fig. 4. Aggregation schemas for the components.

out leads to sample selection bias, as does including it as-is. To avoid this bias, we include all defaults in the sample using scenario analysis to estimate the partial recovery rate that should be added to the actual value as a prediction of the final RR.[5] To address the issue of potentially inappropriate estimates of the partial recovery rate, we estimated each model on a shorter window including default from the 2010–2014 period only. We reported the results in Section 6.3. Additionally, we do not exclude any case from the sample if the LGD is outside the interval [0, 1]. The model should also be able to find those patterns that lead to extreme values. On the other hand, the predictors were winsorized at the 1st and 99th percentile, which is crucial in the case of regression analysis sensitiveness.

[5] We use the bank's internal procedure, which is part of the institution's know-how and cannot be revealed here due to confidentiality issues.

Table 5
Probability of write-off model in decomposed approach.

	Logistic Regression		Classification Tree	
	Mortgage Loans	Cash Loans	Mortgage Loans	Cash Loans
(Intercept)	4.2459***	6.8520***		
Flag of SME contract	−0.4303***			
Contractual interest rate	8.2102***	2.8744***	7.97	6.95
DPD	−0.0014*	−0.0042***	1.54	2.54
Tenor	−0.0012***	−0.0110***	9.15	5.10
Time on books	0.0031***	0.0237***	9.97	4.25
Age of the client		−0.0151***	5.41	5.25
EAD		6.38E-6***	3.04	9.61
Due principal		2.50E-5***		
Due interest	3.50E-5***			1.37
Principal			6.86	0.49
Interest	−6.00E-5***	−0.0004***		3.77
Requested amount	7.75E-8***	−1.00E-5***		1.98
Collateral quota			7.40	
Type of collateral				
Export		−0.0806***		1.67
Import	−0.0147***	0.1328***		3.39
GDP		−0.0456***	7.01	1.13
Domestic Demand	0.0882***	−0.1058***		0.77
CPI	0.2439***	0.2727***		9.69
Yield on 10-year bonds		−0.0585***		
Yield on 5-year bonds	−0.3143***			1.27
Yield on 2-year bonds			37.84	1.83
WIBOR 1Y	−0.7712***	−2.0308***	3.82	29.02
LIBOR 3M	−1.6640***	−1.4283***		2.32
Average wages		0.0244***		
Warsaw Stock Exchange Index		0.0001***		7.58
GINI	.5471	.6155	.5230	.6391

*** indicates 1% level of significance, ** 2.5% and * 5%. The importance of classification tree variables are rescaled to total 100.

For each method, we divided the dataset randomly[6] into a training set and a test set (70% and 30% respectively). In the case of regressions, the forward selection method was used to remove irrelevant independent variables with the p-value cut-off equal to 0.05. SVM and CART manage selection internally. The hyperparameters for CART and SVM[7] are chosen based on 10-fold cross-validation performed on the training set. The mean squared error is minimized to retrieve final parametrization. ANOVA/Gini was selected as the splitting criterion for CART, and the radial basis function as the SVM kernel was used. When estimating parameters of logistic classification, we transformed the explanatory variables (except for the macroeconomic indicators) according to the widely used "weight of evidence" approach (see Anderson, 2007, p. 192). Such an approach converts the risk associated with a particular variable onto a linear scale. Next, for those explanatory variables that correlated the most[8] with each other, we chose the one with the highest information value (see Anderson, 2007, p. 193). Such a set was suffice as the final input for the logistic regression.

When it comes to the estimation part, first, we performed an OLS regression, to set an opening benchmark for our further considerations. In the OLS model (and any other method), the recovery rate is treated as the dependent variable, and the full set of characteristics take part in the estimation in the first place, and then is reduced by selection method adequate to estimation technique. Recoveries are estimated directly (in one step) in this approach without any transformation applied to dependent variable.[9] In the second step, we estimate a two-stage model, using a modeling framework introduced in Yao et al. (2017). As a consequence, we divide the dataset into (a) a full sample on which we estimate the probability of $RR = 0$ vs. $RR > 0$, (b) a sub-sample with $RR > 0$ cases, where we estimate the probability of $RR = 1$ vs. $0 < RR < 1$, and (c) a sub-sample with $0 < RR < 1$ cases to estimate the recovery rate. We chose the best-reported specification based on an out-of-sample RMSE value, which was LS-SVC for classification problems and OLS for regression problems.[10] Finally, we built a decomposed model based on the methodology demonstrated in Section 3. For classification problems, we use logistic regression and classification trees. For regression problems, we use OLS and regression trees. Then, we compare the resulting models' performance on the test set according to the selected measures. We also investigate the robustness of each algorithm with a bootstrap approach, drawing 5% of the test sample for ML and 1% of the test sample for NML 500 times and then calculating the upper and lower confidence limits for RMSE, MAE, and GINI.

6.2. Decomposed model

First, we estimate the probability of cure (Table 4) according to the definition presented in Section 5. Client-level variables and macroeconomic variables with statistical significance are included in the final form. We notice many similarities between mortgages and cash loan portfolios. The conjunction of three variables comes to the fore: the principal amount (negative influence), tenor (positive influence), and decreasing installment plan (negative influence). It can be seen that when a client defaults, installment (which is an offshoot of principal and tenor) is important in describing willingness to repay. If the initial value is relatively

[6] Additionally, we performed stratified random sampling to check if time systematic patterns affect LGD. The findings remain the same, and the results are available upon request.

[7] Hyperparameters for CART included a cost-complexity parameter. Hyperparameters for SVM included a gamma parameter from the range of (2^{-32}, 2^{-16}, 2^{-8}, 2^{-4}, 2^{-2}, 2^{0}) and a cost parameter = (2^{0}, 2^{9}, 2^{10}).

[8] Pearson correlation higher than 80%.

[9] Detailed results of the OLS regression model are available upon request.

[10] Detailed results of the LS-SVC with the OLS benchmark model are available upon request.

Table 6
The severity for consecutive stages in the decomposed approach for the regression models.

	Severity for cures		Severity for write-offs		Severity for partial RR	
	Mortgage loans	Cash loans	Mortgage loans	Cash loans	Mortgage loans	Cash loans
(Intercept)	1.0832***	0.8979***	0.2120***	0.2227***	0.8330***	−0.1133***
Flag of SME contract			0.0366***		−0.0478***	
Contractual interest rate	−1.3758***	−0.2566***		−0.4243***	0.7381***	−0.2091***
DPD	6.92E-5***	−0.0007***		−4.32E-5***	−0.0004***	0.0004***
Tenor	−3.87E-5***	0.0003***		−0.0004***	−0.0007***	−0.0022***
Time on books	−7.93E-5***	−0.0003***	0.0022***	0.0014***	0.0013***	0.0027***
Flag of possessing contract in currency other than domestic	−0.0078***					
Flag of possessing contract with decreasing installment plan		−0.0061***				
Number of co-applicants	−0.0030***					
Flag of EAD higher than the requested amount	−0.0246***	0.0055***				
Age of the client	0.0002***	−0.0005***		−0.0002***	0.0007**	−0.0019***
EAD		5.40E-7***	0.0015***		3.76E-6***	
Due principal			−4.81E-8***		2.88E-6***	9.38E-6***
Due interest		6.23E-5***	−8.49E-7***	1.12E-6***	−6.58E-6***	−2.24E-5***
Principal						
Interest	−1.44E-6***	1.84E-5***			3.49E-6***	1.43E-6***
Requested amount		3.04E-7***		−9.97E-8***		
Collateral quota					1.50E-7***	
Type of collateral					−0.1712***	
Number of cash loans		0.1776***				
Number of credit cards		0.1417***				
Number of revolving loans		0.1089***				
Export			−0.0019***		0.0011*	−0.0013*
Import		0.0003***		0.0020***		−0.0019***
GDP	0.0026***	−0.0039***		−0.0049***		0.0274***
Domestic Demand	−0.0011***			−0.0024***		−0.0190***
CPI	0.0014**	0.0074***		0.0109***	0.0212***	−0.1121***
Yield on 10-year bonds		−0.0028***				0.0092***
Yield on 2-year bonds	−0.0218***		0.0130***	0.0356***	−0.0272***	
WIBOR 1Y	0.0047	−0.0270***		−0.0105***		0.2256***
LIBOR 3M	0.0164***	0.0086***		−0.0103***	−0.1011***	0.2256***
Average wages						0.0017**
Warsaw Stock Exchange Index		4.69E-7***		−2.00E-6***		7.95E-6***
RMSE	0.0589	.0838	.2461	.1593	0.2165	.2775
MAE	0.0332	.0523	.1917	.1076	0.1676	.2225

*** indicates 1% level of significance, ** 2.5% and * 5%.

high (a high principal with a short tenor), then the cure path becomes hard to follow, and a decreasing installment plan (indicating higher installments at the beginning of the crediting period) is another disadvantage. On the other hand, a positive coefficient sign next to due principal and due interest can be counterintuitive, although taking the collection department's point of view into consideration, we can state that higher due amounts are assigned to well-performing employees who are better prepared to direct the client back to the performing portfolio. These are also defaults, that can recover and pay higher returns on average, which is indicated by the interest amount directly. What is also becoming more popular is relegating low due amount to external companies, which usually performs worse than in-house collection. Finally, it needs to be taken into account that due amounts are also part of the probability of write-off and partial RR models, and they affect the aforementioned components as well.

In the second step, we prepare the write-off probability model (Table 5). Although most variables coincide between portfolios, some discrepancies arise. The requested amount can serve as an example: as in the case of mortgage loans, bigger credit implies a greater chance of a write-off. However, in the case of cash loans, the smaller the loan the greater the chance of a write-off. It reveals that the collection policy is constructed to dispose of small exposures, and it works with only the most profitable cases. On the other hand, the macroeconomy seems to have a similar influence both on mortgages and cash loans. Yield on bonds, Interbank Offered Rates, and CPI drive the probability of write-offs, which demonstrates that financial institutions connect their sells with the overall state of the economy.

The severity of cured cases (Tables 6 and 7) is estimated for the first time in this kind of research.[11] Covariate signs are not well established, which makes interpretation difficult. The final RR value should oscillate around one with regard to discounting, which makes it possible to treat this part of recovery rate estimation as an equivalent of time in default estimate to some degree.[12] It is evident that many factors determine the RR for cured cases, and neglecting this part of the distribution can lead to a decrease in the precision of the whole model. Taking the results from the regression tree into consideration, we can conclude that levels of RR are connected to the economic cycle. The yield on bonds or the Interbank Offered Rates are among the most important risk drivers, which suggests that the severity of the losses for cures is an offshoot of the state of the economy.

The severity of write-offs (Tables 6 and 7) is also determined by a series of factors. Additionally, for a large number, this influence does not overlap with the severity of cures, like in the case of the interest amount or due principal. It can also be inferred that cash loans are affected by the state of the economy in various ways, but only the yield on 2-year bonds and export affect mortgages. The severity for write-offs is also the only RR model which gives bet-

[11] To the best of the author's knowledge.

[12] Lower values of the recovery rate in this equation indicate a longer default time and smaller discount factors applied to the coming cash flows.

Table 7
Severity for consecutive stages in the decomposed approach for regression tree models.

	Severity for cures		Severity for write-offs		Severity for partial RR	
	Mortgage loans	Cash loans	Mortgage loans	Cash loans	Mortgage loans	Cash loans
(Intercept)						
Flag of SME contract		1.03				
Contractual interest rate	26.26	7.20	10.84	10.07	5.50	3.63
DPD	2.42	33.91		1.78		4.44
Tenor	6.61	11.21		10.50	10.45	10.89
Time on books	3.72	4.90	25.43	7.41	7.88	6.42
Flag of possessing contract with decreasing installment plan	1.68					
Number of co-applicants	1.48					
Age of the client	6.08	5.68	16.03	9.22	4.28	5.28
EAD	1.31	1.90		3.67	7.78	4.33
Due interest	1.87	0.76		1.56		0.95
Principal	1.21	1.08	19.16	3.31	20.43	6.22
Interest	3.80	1.89	5.16	0.56		0.55
Requested amount	2.91	2.61	4.90	6.88	4.59	2.10
Collateral quota					19.68	
Relationship time	1.90	3.49				
Export	0.14			0.82		1.21
Import		0.50		5.73		3.55
GDP	0.42	1.75		1.02	1.16	2.08
Domestic Demand	5.51		11.32	3.30		0.93
CPI	0.92	0.87		3.16	2.76	7.28
Yield on 10-year bonds	0.04	1.51		1.48		0.65
Yield on 5-year bonds	13.46	2.10		5.78		0.99
Yield on 2-year bonds	3.92	0.78				0.37
WIBOR 1Y	5.90	14.51	7.17	20.09	10.74	28.45
LIBOR 3M	7.19			1.19	1.97	1.27
Average wages	0.86	0.93		0.28		1.30
Warsaw Stock Exchange Index	0.38	1.38		2.21	2.78	7.11
RMSE	0.0597	.0732	.2448	.1602	.2108	.2750
MAE	0.0336	.0457	.1924	.1072	.1578	.2157

*** indicates 1% level of significance, ** 2.5% and * 5%. The importance of the variables are rescaled to total 100.

Table 8
Model performance results. LS-SVC denotes Least-Squares Support Vector Classifier, LR denotes Logistic Regression, CT denotes Classification Trees, RT denotes Regression Trees. The best model in each metric is underlined. The Diebold-Mariano (DM) test verifies the null hypothesis that RMSE from the two models is equal. All comparisons are made to the model with the lowest RMSE. *** indicates p-value lower than 1%, ** 2.5% and * 5%.

		RMSE	MAE	GINI	ρ	DM
Mortgage loans	OLS	.2549	.2048	.3179	.3357	24.81***
	LS-SVC + OLS	.2661	.2155	.2999	.3069	28.62***
	LR + OLS decomposition	.2468	.1917	.3816	.4117	14.56***
	CT + RT decomposition	.2357	.1778	.4385	.4450	–
Cash loans	OLS	.3571	.3175	.2768	.2621	54.72***
	LS-SVC + OLS	.3582	.3174	.2731	.2544	59.32***
	LR + OLS decomposition	.3516	.3113	.3096	.2948	31.16***
	CT + RT decomposition	.3458	.2996	.3356	.3283	–

ter results for cash loans than for mortgages when RMSE/MAE is concerned, which might be connected with the more liquid market for non-mortgage loans, which are put on sale more frequently than mortgage loans.

For partial RR (Tables 6 and 7), the vast majority of contract-based predictors work in the opposite direction, except for time on books, due principal amount and principal amount. Another differences can be found in the macro economy section, where different sets of variables are used both for mortgages and cash loans, and again, cash loan portfolio is more influenced. This can be a result of collateral presence, which highly differentiate the importance of other predictors.

6.3. Comparative performance

Tables 8 and 9 demonstrate the forecast results for the different modeling methods broken down into mortgage loans and cash loans. We estimated four models in each portfolio: (a) a one-step OLS regression model, (b) a two-step model as a combination of LS-SVC and OLS, (c) a decomposed model with logistic regression for classification and OLS for severity, and (d) a decomposed model with a classification tree for classification and a regression tree for severity. We evaluated model prediction effectiveness considering five commonly used measures described in Section 4.5, based a hold-out sample of 30% of the total. The best performing model according to each metric is underlined. Focusing on the first table, it is clear that the decomposed model performs best for each measure of forecast effectiveness. We see an increase in quality, mainly in well-established connections between LGD and the independent variables. Even if this connection is more non-linear (which can be inferred from better results for the CT + RT alternative than for LR + OLS), the intrinsic features of these portfolios are better reflected by associating the recovery paths to events that drive the final LGD value rather than to the LGD distribution. This is most apparent for the write-off severity; assuming that these are just full-loss cases leads to a great part of the LGD variability be-

Table 9
Model performance results for the period 2010–2014. LS-SVC denotes Least-Squares Support Vector Classifier, LR denotes Logistic Regression, CT denotes Classification Trees, RT denotes Regression Trees. The best model in each metric is underlined. The Diebold-Mariano (DM) test verifies the null hypothesis that RMSE from the two models is equal. All comparisons are made to the model with the lowest RMSE. *** indicates p-value lower than 1%, ** 2.5% and * 5%.

		RMSE	MAE	GINI	ρ	DM
Mortgage loans	OLS	.2635	.2158	.3388	.3438	18.67***
	LS-SVC + OLS	.2726	.2252	.3238	.3171	21.19***
	LR + OLS decomposition	.2565	.2021	.3864	.4054	12.70***
	CT + RT decomposition	.2506	.1941	.4165	.4312	–
Cash loans	OLS	.3800	.3497	.2663	.2743	48.80***
	LS-SVC + OLS	.3806	.3489	.2638	.2678	49.34***
	LR + OLS decomposition	.3690	.3383	.3397	.3496	28.30***
	CT + RT decomposition	.3619	.3255	.3691	.3800	–

Table 10
500-fold Bootstrap for RMSE, MAE, and GINI values. LCL denotes the lower confidence limit for the mean, and UCL denotes the upper confidence limit for the mean.

		RMSE		MAE		GINI	
		LCL	UCL	LCL	UCL	LCL	UCL
Mortgage loans	OLS	.2545	.2555	.2046	.2053	.3161	.3202
	LS-SVC + OLS	.2659	.2669	.2154	.2161	.2972	.3011
	LR + OLS decomposition	.2462	.2472	.1913	.1920	.3787	.3824
	CT + RT decomposition	.2351	.2361	.1773	.1780	.4357	.4393
Cash loans	OLS	.3567	.3573	.3172	.3179	.2758	.2787
	LS-SVC + OLS	.3578	.3584	.3170	.3176	.2718	.2747
	LR + OLS decomposition	.3513	.3519	.3109	.3115	.3084	.3114
	CT + RT decomposition	.3455	.3461	.2993	.2999	.3338	.3366

ing left unexplained.[13] Considering the mortgage portfolio, CT + RT produces estimates which reduce the prediction errors (the lowest RMSE value) to the greatest extent. Discrimination measures also indicate with no doubt that decomposition based on trees works best with assigning high predictions to high realizations (and vice-versa). For cash loans portfolio, the tree-based version outperforms others, regardless of the measure selected. It should also be noted that the most sophisticated approach using SVM does not seem to show any benefits in performance compared to simple OLS. In the case of mortgage loans, it may stem from the more J-shaped distribution, which makes it difficult to properly classify cases with full-loss, as their number is limited. For cash loans, we also notice many cases around the peaks, so patterns blur in the immediate surroundings of zero and one, where the two-stage benchmark presents its power.[14] The Diebold-Mariano test also supports the results, as it indicates rejecting the null hypothesis regarding mean equality.

To check the impact of open defaults for which the partial recovery rate was added to handle resolution bias, we re-estimated all models for the 2010–2014 period. Generally, the out-of-sample verification confirms the results obtained for the whole sample. The decomposed model still outperforms selected benchmarks, indicating that including short-lasting defaults from the latest period does not influence the stability of the proposed approach. Additionally, we use Bootstrap to determine the confidence intervals for the selected measures. For almost every criterion, these intervals do not overlap between the benchmark models and the decomposed models, which allows us to state that the values reported in Table 10 truly represent the difference in quality. However, we need to note that each component of this approach needs to perform reasonably well to achieve satisfying results. Following Yao et al. (2017) findings, the power of the classification parts (cure probability and write-off probability) plays a crucial role, and even excellent results obtained by the severity models will not correct the weaknesses of the binary-event models. However, if the institution can clearly distinguish the recovery paths, then preparing models simply on collection events can improve the quality significantly. What is more, such results are achievable with the usage of a highly interpretable set of methods.

[13] See descriptive statistics and distributions in Section 5 to compare.
[14] See the discussion in Section 4 of Yao, Crook, Andreeva (2017).

7. Conclusion

This paper developed an event-based decomposition of the LGD model with a framework for the probability of cure, probability of write-off, severity of cures, severity of write-off, and severity of partial recoveries. The results support the idea that LGD decomposition using a mixture distribution of the events is effective in modeling consumer risk. In addition, this study shows that using a specific set of variables and their transformations can lead to more precise and interpretable results than using the same set regardless of the nature of the event. Starting with the two-stage framework presented in Belotti and Crook (2010) or Yao et al. (2017), among others, we developed models for consecutive stages of the workout process. First the probabilities of cure and write-off were modeled with a logistic regression and classification trees. Next, the severities of the cures, partial recoveries, and write-offs were estimated with the use of ordinary least squares and regression trees. Each model was prepared based on a specific set of variables that reflects the intrinsic features connected to the stage. Finally, we proposed a combination mechanism to calculate the ultimate recovery, taking into account all five elements of the equation. To check the effectiveness, two benchmarks were prepared: a direct LGD estimation with an OLS model and a two-stage model based on the assumptions defined in Yao et al. (2017). The decomposed model demonstrates better predictions than any benchmark in terms of out-of-sample predictive metrics, although the improvements in cash loan portfolio are not as remarkable as for mortgage portfolio. Of the two propositions, combining the classification tree and the regression tree outperforms the combination of logistic regression and ordinary least squares. We were not able to confirm the superiority of LS-SVC + OLS combination over simple OLS, as presented in Yao et al. (2017). However, our finding about a two-stage event-based approach advantage over one-stage is in line with Tanoue et al. (2017), as an example. Additionally, we examined the im-

pact of risk factors on both portfolios broken down into stages. The results show that the probability of cure is driven mostly by DPD; the probability of write-off is driven by Interbank Offered Rates; cure severity is driven by contractual interest rate and DPD; write-off severity is driven by time on books and Interbank Offered Rates; and partial RR severity is driven by principal and Interbank Offered Rates.

Our model has important implications for bank risk management policy and collection department policy. Specifically, it provides a way to optimize capital allocation by means of the risk-weighted assets value. This can be achieved by directing the credit policy to accept low-risk clients with a high probability of cure value. Secondly, the loan loss provisioning process can be managed more efficiently by the faster selling of defaults with a high probability of write-off values or by putting less effort into defaults likely to be cured. Additionally, linking events to the actual actions taken by the collection departments makes it possible to differentiate cures, as low-cost cases need only minor help from the institution to get back on track, and write-offs are seen as events that take place even after some paybacks from the client side, which is also consistent with economic intuition. We suggest that the choice of event that leads to a specific recovery path provides a better fit and better understanding of the LGD parameter than a one-equation estimation or defining cures and write-offs based only on the dependent variable distribution.

Acknowledgments

I am grateful to the editor and two anonymous referees for their valuable comments. I am also grateful to Paweł Baranowski and Mariusz Górajski, for their valuable comments on the draft of the paper and help when preparing revision.

References

Acharya, V., Bharath, S., & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries. *Journal of Financial Economics, 85*(3) doi: 10.1016/j.jfineco.2006.05.011.

Anderson, R. (2007). *The credit scoring toolkit*. Oxford University Press.

Anolli, M., Beccalli, E., & Giordani, T. (2013). *Retail credit risk management*. New York: Palgrave MacMillan.

Basel Committee on Banking Supervision (2017). Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures (EBA/GL/2017/16). Available at https://eba.europa.eu/documents/10180/2033363/Guidelines+on+PD+and+LGD+estimation+%28EBA-GL-2017-16%29.pdf, accessed at 2020-05-16.

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

Belotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting, 28*(1). https://doi.org/10.1016/j.ijforecast.2010.08.005.

Betz, J., Kellner, R., & Rösch, D. (2018). Systematic effects among loss given defaults and their implications on downturn estimation. *European Journal of Operational Research, 271*. https://doi.org/10.1016/j.ejor.2018.05.059.

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis, *Bank of England Staff Working Paper*, No. 816.

Brown, I. (2012). *Basel II compliant credit risk modelling*. Southampton: University of Southampton.

Cornejo, J. (2007). *Risk management in consumer financial institutions*. BookSurge Publishing.

Covitz, D., & Han, S. (2004). An empirical analysis of bond recovery rates: Exploring a structural view of default, Working Paper, The Federal Reserve Board, doi: 10.17016/FEDS.2005.10

Dermine, J., & Neto de Carvahlo, C. (2006). Bank loan losses-given-default: A case study. *Journal of Banking and Finance, 30*(4). https://doi.org/10.1016/j.jbankfin.2005.05.005.

Do, H., Rösch, D., & Scheule, H. (2018). Predicting loss severities for residential mortgage loans: A three-step selection approach. *European Journal of Operation Research, 270*. https://doi.org/10.1016/j.ejor.2018.02.057.

Finlay, S. (2009). *Consumer credit fundamentals*. New York: Palgrave and MacMillan.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Joubert, M., Verster, T., & Raubenheimer, H. (2018). Making use of survival analysis to indirectly model loss given default. *ORiON, 34*(2). https://doi.org/10.5784/34-2-588.

Krüger, S., & Rösch, D. (2017). Downturn LGD modelling using quantile regression. *Journal of Banking and Finance, 79*. https://doi.org/10.1016/j.jbankfin.2017.03.001.

Leow, M. (2010). *Credit risk models for mortgage loan loss given default*. Southampton: University of Southampton.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modelling. *International Journal of Forecasting, 28*(1). https://doi.org/10.1016/j.ijforecast.2011.01.006.

Maldonado, S., Bravo, C., Lopez, J., & Perez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. *Decision Support Systems, 104*. https://doi.org/10.1016/j.dss.2017.10.007.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operations Research, 183*(3). https://doi.org/10.1016/j.ejor.2006.04.051.

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems, 51*(4). https://doi.org/10.1016/j.dss.2011.01.013.

Nazemi, A., Fatemi Pour, F., Heidenreich, K., & Fabozzi, F. J. (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research, 262*(2). https://doi.org/10.1016/j.ejor.2017.04.008.

Ozdemir, B., & Miu, P. (2009). Basel II implementation: A guide to developing and validating a compliant. *Internal risk rating system*. McGraw-Hill. https://doi.org/10.1036/0071591303.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola, P. Bartlett, B. Sholkopf, & D. Schuurmans (Eds.), *Advanced in large margin classifiers* (pp. 61–74). MIT Press.

Qi, M., & Yang, X. (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking and Finance, 33*(5). https://doi.org/10.1016/j.jbankfin.2008.09.010.

Qi, M., & Zhao, X. (2011). Comparison of modeling methods for Loss Given Default. *Journal of Banking and Finance, 35*(11). https://doi.org/10.1016/j.jbankfin.2011.03.011.

Tanoue, Y., Kawada, A., & Yamashita, S. (2017). Forecasting Loss Given Default of bank loans with multi-stage models. *International Journal of Forecasting, 33*(2). https://doi.org/10.1016/j.ijforecast.2016.11.005.

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.

Tobback, E., Martens, D., Van Gestel, T., & Baesens, B. (2014). Forecasting loss given default models: Impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society, 65*(3). https://doi.org/10.1057/jors.2013.158.

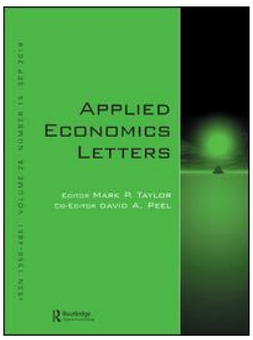
Tong, E., Mues, C., & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loss given default. *International Journal of Forecasting, 29*(4). https://doi.org/10.1016/j.ijforecast.2013.03.003.

Van Berkel, A., & Siddiqi, N. (2012). building loss given default scorecard using weight of evidence bins in SAS enterprise miner. *SAS Global Forum* 2012.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Yao, X., Crook, J., & Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research, 263*(2). https://doi.org/10.1016/j.ejor.2017.05.017.

Zhang, J., & Thomas, L. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting, 28*(1). https://doi.org/10.1016/j.ijforecast.2010.06.002.



Routledge
Taylor & Francis Group

Applied Economics Letters

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rael20

Forecast combination approach in the loss given default estimation

Wojciech Starosta

To cite this article: Wojciech Starosta (2020): Forecast combination approach in the loss given default estimation, Applied Economics Letters, DOI: 10.1080/13504851.2020.1854438

To link to this article: https://doi.org/10.1080/13504851.2020.1854438

Published online: 30 Nov 2020.

APPLIED ECONOMICS LETTERS
https://doi.org/10.1080/13504851.2020.1854438

Routledge
Taylor & Francis Group

ARTICLE

Forecast combination approach in the loss given default estimation

Wojciech Starosta

Institute of Econometrics, Department of Economics and Sociology, University of Lodz, Lodz, Poland

ABSTRACT

This paper examines a novel method of including macroeconomic variables into Loss Given Default models. The approach is transparent, and it easily translates changes in the overall credit environment into Expected Loss estimates, which is one of the crucial points that was recently introduced in the International Financial Reporting Standard 9. We propose a forecast combination procedure that separates the contract-based variables from the macroeconomic indicators. Two models are prepared and benchmarked to a single ordinary least-squares (OLS) model. To combine the forecasts we use three approaches: simple average, the Granger–Ramanathan Method, and Mallows Model Averaging. We tested our predictions on out-of-time data and found that the forecast combination outperforms the single OLS model in terms of the selected forecast quality metrics.

KEYWORDS

Loss given default; forecast combination; IFRS 9; model averaging; loan loss provisioning

JEL CLASSIFICATION

C51; C53; G32

I. Introduction

The recent global financial crisis revealed many weaknesses in the incurred loss models, which were the main part of the existing International Accounting Standard 39 (IAS 39). The backward-looking approach resulted in delaying the recognition of credit losses and balance sheet financial asset overstatement. A review of accounting standards produced International Financial Reporting Standard 9 (IFRS 9), which went live in 2018. The key innovation incorporates a shift from the backward-incurred-loss perspective into a forward-looking Expected Credit Loss (ECL) calculation (Bellini 2019, 2). The parameters (Probability of Default, Exposure at Default, Loss Given Default) are expected to be simultaneously unbiased, point-in-time (PIT), and forward-looking. Because the unbiasedness and PIT character have already been deliberated in the literature (see Ozdemir and Miu 2009; Anolli, Beccalli, and Giordani 2013) as a part of Basel models, we focused on the forward-looking concept in the context of the Loss Given Default (LGD) models (probability of default was studied inter alia in Durovic 2019). The forward-looking perspective relies on the proper inclusion of credit and the macroeconomic environment into the risk models. It should reflect the potential downturn or upturn in the quality of the acquired credits. Miu and Ozdemir (2017) argued that 'a replicable, transparent and defendable mechanism to translate the change in the credit environment to the change in the portfolio's Expected Loss estimation is needed'.

In this study, we proposed a forecast combination approach, widely used in other fields of economic modelling (e.g. Jumah and Kunst 2016), and meeting the postulates presented above. We used a sample of defaulted assets to estimate the LGD parameter in two ways. First, we used contract-based and macroeconomic variables and we prepared a full specification model. In the second step, we divided the set of variables into two categories, contract information and macroeconomic indicators, and we estimated the two models separately. Then, we combined both forecasts using three methods: (a) simple average, (b) the Granger–Ramanathan Method, and (c) Mallows Model Averaging. Finally, we checked the performance of four forecasts on an out-of-time sample.

II. Estimation methodology

Our estimation methodology consisted of four steps. First, having a dataset with dependent

CONTACT Wojciech Starosta w.starosta@wp.pl Institute of Econometrics, Department of Economics and Sociology, University of Lodz, Lodz 90-255, Poland

variable LGD (Y) and the set of explanatory variables of contract characteristics (X) and macroeconomic indicators (Z), we estimated a linear regression model:

$$y_i = \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{l=1}^{L} \gamma_l z_{il} + e_i$$

where β_k represents a regression parameter for the contract variables, γ_l represents a regression parameter for the macroeconomic indicators, and e_i are the error terms. This approach is popular in the LGD literature (see Qi and Yang 2009, or Belotti and Crook 2010), and it gives reasonably stable and precise results. This model serves as a benchmark for our further considerations.

In the second step, we divided our dataset into two sub-sets. Dataset A includes contract information with n observations and K explanatory variables. Dataset B contains macroeconomic indicators with n observations and L explanatory variables. We followed the Granger finding that 'it is more usual for combining to produce a better forecast when the individual forecasts are based on different information sets, and each may be optimal for their particular set' (Granger 1989, 168). A similar approach, based on random variable attribution to the datasets, is a part of the Random Forest algorithm (Bellini 2019, 59). Next, we fit a linear regression model on each dataset:

$$y_i^{ci} = \sum_{k=1}^{K} \delta_k x_{ik} + e_i^{ci}$$

$$y_i^{mi} = \sum_{l=1}^{L} \theta_l z_{il} + e_i^{mi}$$

where y_i^{ci} denotes LGD for the regression with contract information only, y_i^{mi} denotes LGD for regression with macroeconomic indicators only, δ_k represents the regression coefficients in the contract approach, and θ_l represents the regression coefficients in the macroeconomic approach. The crucial point is to combine these two forecasts into one forecast to prepare the best model on in-sample data, which is our third step. Let $w(m)$ be a vector of weight assigned to the mth forecast and $\hat{f}_t(m)$ be the vector of values of the forecast (in our framework these are $\hat{y}_i^{ci}$ and $\hat{y}_i^{mi}$). The combination of forecast of y_{n+1} is given by (Hansen 2008):

$$\widehat{w'f}_{n+1} = \sum_{m=1}^{M} w(m)\hat{f}_{n+1}(m)$$

We consider three weighting schemes, each satisfying the convexity constraints (weights to be in interval [0,1] and to sum up to 1). The selected schemes go from (a) choosing one simple method to check if estimation error connected with subsequent two do not cause forecast deterioration, (b) using approach often preferred in similar studies, (c) check one novel method never investigated in credit risk context before.

Approach A: Simple average

An equally weighted combination of forecasts is our first choice, as, in practice, it often performs better than more sophisticated approaches (Genre et al. 2013). The vector of weights is determined as:

$$\widehat{w} = \frac{1}{M}.$$

Approach B: The Granger–Ramanathan Method

Granger and Ramanathan (1984) introduced another way to combine forecasts, which is selecting weights by minimizing the sum of the squared forecast errors:

$$Q(w) = \sum_{i=1}^{n} \left(y_i - \widehat{f}_i' w\right)^2$$

Additionally, we impose the convexity constraints:

$$\widehat{w} = \min_{0 \le w(m) \le 1, \sum_{m=1}^{M} w(m) = 1} Q(w)$$

Approach C: Mallows Model Averaging

Finally, we used a scheme presented in Hansen (2007, 2008), which is the Mallows criterion for model selection and its extension. Assuming the full-sample averaging estimator of the conditional mean μ_i to be $\widehat{\mu}_i' w = \widehat{a}(w)' x_i$, where $\widehat{\mu}_i = \left(\hat{\mu}_i(1), \hat{\mu}_i(2), \ldots, \hat{\mu}_i(M)\right)'$ and $\hat{\mu}_i(m) = x_i(m)' \widehat{a}(m)$, the MMA criterion is as follows:

$$C_n(w) = \sum_{i=1}^{n} \left(y_i - \widehat{\mu}_i' w\right)^2 + 2w(m)r(m)s^2$$

where $R = (r(1), r(2), \ldots, r(M))'$ is a vector with a consecutive number of regressors and $s^2 = \frac{1}{n-r(M)}\widehat{e}(M)'\widehat{e}(M)$. In this notation, the weight vector is the value of w that minimizes $C_n(w)$:

$$\widehat{w} = \min_{w\in[0,1]^M:\sum_{m-1}^{M} w^m=1} C_n(w)$$

which translates to forecasts defined as $\widehat{\mu}_i'\widehat{w} = \widehat{a}(\widehat{w})'x_i$. As in our exercise $M = 2$; and this equation can be solved analytically.

In the fourth step, we assessed the quality of each forecast using the set of methods containing Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Spearman's correlation coefficient (ρ). The forecasting error ε_t is defined as the difference between the measured value at time t (y_t) and the predicted value at time t obtained from forecast g $\left(f_t^{(g)}\right)$. Metrics based on errors should be as low as possible, in contrast to metrics based on correlation. This set allows us to gain insight into how the forecasts perform in terms of overall effectiveness, detecting outliers, comparability, and the ability to rank observations. To properly assess the quality of the forecasts, we divided our dataset into a training sample, on which we estimated our models, and the hold out sample, on which we calculated the selected metrics. The sample is divided at the year of default. The period ranging from 2010 to 2015 was used for estimation purposes; the period ranging from 2016 to 2018 was used to assess forecast quality.[1]

Table 1. Regression results (point estimate and standard error in parenthesis).

	Full specification	Forecast combination: contract	Forecast combination: macro
(Intercept)	0.7301***	0.9045***	0.4864***
	(0.0294)	(0.0097)	(0.0299)
SME flag	−0.0802***	−0.063***	
	(0.0055)	(0.0054)	
Interest rate	−0.2917***	−0.3965***	
	(0.056)	(0.0558)	
DPD	−0.0014***	−0.0014***	
	(0.0001)	(0.0001)	
Tenor	−0.0004***	−0.0003***	
	(0.0001)	(0.0001)	
Time on books	0.0015***	0.0019***	
	(0.0001)	(0.0001)	
Age of the contract owner	−0.0019***	−0.0017***	
	(0.0002)	(0.0002)	
EAD	−1.11E-7***	−1.34E-7***	
	(3.37E-9)	(1.1E-8)	
Due principal	−4.06E-7***	−4.02E-7***	
	(7.90E-8)	(7.93E-8)	
Interest	−7.72E-6***	−7.27E-6***	
	(4.89E-7)	(4.92E-8)	
Requested amount		2.18E-8**	
		(9.54E-9)	
Collateral quota	5.35E-8***	5.28E-8***	
	(1.58E-9)	(1.62E-9)	
Type of collateral	−0.0171**	−0.0166**	
	(0.0075)	(0.0075)	
Export			0.003***
			(0.0008)
Import	−0.0024***		−0.0068***
	(0.0002)		(0.0008)
GDP	0.0033***		
	(0.0009)		
Domestic Demand			0.0081***
			(0.0012)
CPI	−0.0418***		−0.0608***
	(0.0038)		(0.0042)
Yield on 10-year bonds	−0.0086***		−0.0128***
	(0.0025)		(0.0028)
WIBOR 3 M	0.0704***		0.0955***
	(0.0052)		(0.0056)
LIBOR 3 M	−0.0252***		−0.0120**
	(0.007)		(0.0079)
Nominal wages			−0.0044***
			(0.0012)
Warsaw Stock Exchange Index	2.23E-6***		3.35E-6***
	(3.11E-7)		(3.32E-7)

Independent variable is set as Recover Rate (1 – LGD). *** indicates 1% level of significance, ** 2.5% and * 5%.

III. Data set and empirical results

Our data set includes 7244 observations of the mortgage loans defaulted during the 2010–2018 period. The division into contract information and macroeconomic indicators is driven by the IFRS 9 requirements, but it also isolates the intrinsic portfolio characteristics from the changes in the outer environment. In our considerations, we used a mix of static variables (Small and Medium-sized Enterprises (SME) flag or requested amount), the dynamic variables (months on book or age of the contract owner), and a macroeconomic variable (export, import, etc.). Table 1 provides the standard OLS estimates for each approach. The stepwise selection method was used to remove irrelevant independent variables, with a p-value cut-off equal to.05.

Because we only considered the forecast ability of each approach, we used general statements when interpreting our results. The signs of the coefficients have the same direction between the approaches and agree with the previous research (see Belotti and Crook 2010; Tong, Mues, and Thomas 2013),

[1]To assess the robustness, we also compared the following periods: 2010–2014 vs. 2015–2018 and 2010–2016 vs. 2017–2018. The conclusions remain the same.

Table 2. Model performance results.

Method	RMSE	MAE	ρ
Full specification OLS	.19006	.14896	.29499
Forecast combination (Simple average)	.18153	.15431	.25748
Forecast combination (Granger–Ramanathan)	.18109	.13974	.31616
Forecast combination (MMA)	.18706	.14380	.29313

Best model in each metric is underlined.

which is an expected property indicating robustness. In the forecast-combination approach, three new variables were included in the regression models: the requested amount in the contract section and the domestic demand, with the nominal wages in the macro section.[2] Without further considerations, the rest of this section addresses the forecast quality part of our findings.

Table 2 presents the performance measures for each approach. The combination with weights calibrated gave better out-of-sample predictions than the full specification OLS model, which supports our statement about the superiority of a combined approach. Additionally, the simple average performed worse, in terms of MAE and correlation, which indicates that tuning the weights could lead to better forecasts. Within the three combined models, the MMA was slightly worse than the Granger–Ramanathan, but still outperformed simple average. We find the reason of such behaviour in better fit of short-term dependencies, reflected by the contract-based variables, in contrast to long-term dependencies expressed by the macro-economic indicators. The results of forecast combination suggest that there are some limitations of macroeconomic variables, which can explain only limited part of the LGD variability, especially in the case of retail exposures. This leads to unequal weight attribution, favouring one of the components.

IV. Conclusion

Inclusion of forward-looking information is not an easy task, and it should be treated with great care to obtain robust and predictive models. Our results confirm that dividing the variables into contract information and macroeconomic indicators can lead to a higher forecast quality than estimating the full specification model. Additionally, the following advantages could be appealing for risk managers:

- The influence of the macroeconomic environment can be easily measured and validated, especially if assessing the quality of the macroeconomic indicators separately from the whole LGD model.
- The complete economic cycle does not have to meet the current collection policy practices. The presented approach allows for estimating each part of the model separately based on different time windows, so a longer observation period can be assigned to the macro part, and a shorter observation period can be assigned to the contract part to reflect the ongoing organizational principles.
- The stress tests exercise can be effectively facilitated.

We believe that our proposition not only meets this expectation, it is intuitive and transparent and easily transforms situations in the credit environment into expected loss values, which is desirable from the accounting view imposed by IFRS 9.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Wojciech Starosta http://orcid.org/0000-0002-2306-0263

References

Anolli, M., E. Beccalli, and T. Giordani. 2013. *Retail Credit Risk Management*. New York: Palgrave MacMillan.

Bellini, T. 2019. *IFRS 9 and CECL Credit Risk Modelling and Validation*. San Diego: Academic Press.

Belotti, T., and J. Crook. 2010. "Loss Given Default Models for UK Retail Credit Cards." *CRC Working Paper* 09/1.

Durovic, A. 2019. "Macroeconomic Approach to Point in Time Probability of Default Modeling – IFRS 9 Challenges." *Journal of Central Banking Theory and Practice* 1: 209–223. doi:10.2478/jcbtp-2019-0010.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann. 2013. "Combining Expert Forecasts: Can Anything Beat the

[2]To be fully compliant, we also estimated the forecast combination approach only using variables included in the one-stage OLS. The conclusions are the same; the results are available upon request.

Simple Average?" *International Journal of Forecasting* 29: 108–121. doi:10.1016/j.ijforecast.2012.06.004.

Granger, C. W. J. 1989. "Combining Forecasts – Twenty Years Later." *Journal of Forecasting* 8: 167–173. doi:10.1002/for.3980080303.

Granger, C. W. J., and R. Ramanathan. 1984. "Improved Methods of Combining Forecast Accuracy." *Journal of Forecasting* 19: 197–204. doi:10.1002/for.3980030207.

Hansen, B. 2007. "Least-squares Model Averaging." *Econometrica* 75: 1175–1189. doi:10.1111/j.1468-0262.2007.00785.x.

Hansen, B. 2008. "Least-squares Forecast Averaging." *Journal of Econometrics* 146: 342–350. doi:10.1016/j.jeconom.2008.08.022.

Jumah, A., and R. Kunst. 2016. "Optimizing Time-series Forecasts for Inflation and Interest Rates Using Simulation and Model Averaging." *Applied Economics* 48: 4366–4378. doi:10.1080/00036846.2016.1158915.

Miu, P., and B. Ozdemir. 2017. "Adapting the Basel II Advanced Internal-ratings-based Models for International Financial Reporting Standard 9." *Journal of Credit Risk* 13 (2): 53–83. doi:10.21314/JCR.2017.224.

Ozdemir, B., and P. Miu. 2009. *Basel II Implementation. A Guide to Developing and Validating A Compliant, Internal Risk Rating System*. New York: McGraw-Hill.

Qi, M., and X. Yang. 2009. "Loss Given Default of High Loan-to-value Residential Mortgages." *Journal of Banking and Finance* 33 (5): 788–799. doi:10.1016/j.jbankfin.2008.09.010.

Tong, E., C. Mues, and L. Thomas. 2013. "A Zero-adjusted Gamma Model for Mortgage Loss Given Default." *International Journal of Forecasting* 29: 548–562. doi:10.1016/j.ijforecast.2013.03.003.