Politecnico
di Torino

Politecnico di Torino

Laurea Magistrale in Ingegneria Matematica

Anno Accademico 2024/25

Machine learning approaches to LGD estimation: a methodological evolution and applications to the *Intesa Sanpaolo* case

Candidato:	**Relatore:**	**Correlatore:**
Giulio Ruffinello	Patrizia Semeraro	Francesco Grande

Contents

Introduction

Credit risk constitutes the primary financial risk for intermediaries and is defined as the risk of loss arising from a borrower's failure to meet contractual obligations. Developing models to estimate and predict this risk is therefore a fundamental task of a bank's risk management function. This involves creating statistical models based on historical defaults to assess the risk of current borrowers, which in turn informs risk management strategies and lending policies. Key parameters include the Probability of Default (PD), Exposure at Default (EAD) and Loss Given Default (LGD). While credit risk research has focused extensively on the estimation of PD, LGD modelling - the focus of this thesis - has received comparatively less attention.

Loss Given Default (LGD) measures the proportion of an exposure that is lost following a default. Its estimation in industry practice is characterized by a classic trade-off between model complexity and interpretability. Standard practice favours parsimonious, grid-based models that are easily understood but impose severe constraints on predictive power, often resulting in high bias. Conversely, flexible machine learning models, widely explored in academic literature, can achieve lower estimation error but suffer from a lack of interpretability and a higher risk of overfitting.

The primary objective of this work is to develop and validate a new methodological paradigm that resolves this conflict, yielding estimates that are both more accurate and operationally manageable. The proposed approach first employs unconstrained, high-capacity estimation models to accurately capture the complex relationships within the data. Subsequently, a post-processing aggregation step, based on hierarchical clustering, is applied to the continuous predictions. This second phase reduces the granularity of the output to a manageable number of discrete grades, thereby restoring parsimony while preserving the predictive accuracy of the initial, more complex model.

The thesis is structured to logically develop and test this methodology. Chapter 1 provides the general context, introducing the regulatory framework and the core parameters of the AIRB approach. The analysis then proceeds in Chapter 2 to establish the theoretical foundation, framing LGD estimation as a supervised learning problem. In particular, two model designs are examined: the single-stage approach and the two-stage framework, with the latter specifically designed to address the latent heterogeneity often found in borrower populations. Chapter 3 focuses on the validation methodology, presenting a critical analysis of the Somers' d metric. We analyze how this metric can exhibit potential biases in certain scenarios, suggesting that both of its asymmetric formulations should be considered, contrary to common practice. This groundwork supports the core empirical validation in Chapter 4, where the proposed models are implemented and the effectiveness of the post-aggregation procedure is illustrated.

Chapter 1

General context and research objective

This chapter establishes the context and research objective of the thesis. It begins by outlining the fundamentals of credit risk modelling, its key parameters, and the regulatory framework (Section 1.1). It then details the standard methodology for LGD model development, from historical data processing to the phases of risk differentiation and quantification (Section 1.2). The chapter concludes by focusing on the Intesa Sanpaolo case, defining the core objective of this work: to explore a methodological evolution toward a more granular LGD estimation for the Other Retail segment (Section 1.3).

1.1 Credit risk modelling: purposes, parameters and regulation

The activity of financial intermediation lies at the core of the business model of commercial banks. It consists in collecting funds from different sources - such as retail deposits, wholesale funding and shareholder equity - and then, on the asset side, allocating these resources across various investments - most notably through the extension of credit to households and firms.

Credit risk arises directly from the activity of financial intermediation and is generally defined as the **risk of financial loss resulting from a borrower's inability to meet contractual obligations**. Since banks inevitably assume this risk through their lending activities, regulation requires them to hold a minimum amount of capital as a safeguard known as capital requirement - also referred to as regulatory capital, capital adequacy or capital base. This amount is designed to absorb *unexpected losses* and to ensure that the institution does not become excessively leveraged or insolvent.

In practice, a distinction is made between expected and unexpected losses. Expected losses are already recognised in accounting terms and are covered through provisions, which appear as adjustments on the asset side of the balance sheet. Unexpected losses, on the other hand, are those that exceed expectations and represent the main focus of credit risk management and regulatory requirements, that aim to absorb these shocks, since provisions alone may be insufficient. Unlike provisions, the regulatory capital is not part of the balance sheet itself, but consists of specific balance sheet elements: CET1, AT1 and Tier 2 capital.

The required amount of regulatory capital is usually expressed as a Capital Adequacy Ratio (CAR), which measures the bank's regulatory capital divided by its *risk-weighted*

assets (RWA):

$$\text{CAR} = \frac{\text{Regulatory Capital}}{\text{RWA}} \times 100\% \tag{1.1}$$

According to Basel regulations, banks are required to maintain a minimum CAR of 8% - meaning that **the regulatory capital must be at least 8% of the total RWA**. RWA represent the bank's total assets, weighted by the risk associated to each exposure - so that riskier assets require more capital and safer assets require less, ensuring that the capital held by the bank is proportional to the actual risk it carries on its balance sheet.

In the Italian banking system, as in other European jurisdictions, there are rules governing credit risk management and capital adequacy that originate from international, European and national sources. At the international level, the most influential body is the Basel Committee on Banking Supervision (BCBS) - operating under the auspices of the Bank for International Settlements (BIS). Within the European Union, the Basel framework is transposed into European law through the Capital Requirements Regulation (CRR) and the Capital Requirements Directive (CRD). Building on these provisions, the European Banking Authority (EBA) issues Guidelines and Regulatory Technical Standards - which provide further specifications and ensure consistent implementation across Member States. Finally, at the national level, the Bank of Italy issues supervisory provisions and ensures the concrete application of the European regulatory framework within the whole Italian banking system, except for the *significant institutions* which are under the direct control of the ECB.

The regulatory framework has evolved significantly over time. The first step towards international cooperation in banking supervision was the establishment of the BCBS in 1974. Since then, the Basel Accords have been revised multiple times: Basel I (1988) introduced the first global capital standards, Basel II (2004) refined risk sensitivity and incorporated internal model approaches, Basel III (2010) strengthened capital and liquidity requirements in response to the global financial crisis and Basel IV (2017) further addressed model risk and standardisation in capital calculation.

Given the central role of RWA in determining capital requirements, to prevent inconsistencies across institutions and jurisdictions, regulators have established detailed rules and methodologies governing their calculation, including how risk weights should be assigned to different exposures. Let us now focus on assets exposed to credit risk, and therefore on the computation of credit risk weights. Under Basel I, capital requirements were based on simple, standardized risk-weight coefficients applied to broad asset classes, without differentiating between borrowers' individual creditworthiness. With Basel II, the determination of credit risk weights has undergone a significant evolution, with the introduction of both the Standardised Approach (SA) and the Internal Ratings-Based (IRB) approach.

The SA approach substitutes the standardized risk-weight coefficients of Basel I with externally determined risk weights and credit ratings provided by agencies such as Standard & Poor's, Moody's or Fitch. The IRB framework - which is at the core of this work - goes further, allowing banks to **assign a risk weight to each exposure based on their own internal assessments of borrowers' creditworthiness**, providing a more granular measure of capital requirements. Within this framework, a distinction is made between the Foundation IRB (FIRB) and the Advanced IRB (AIRB) approaches - depending on the degree of autonomy granted to banks in estimating the parameters used for risk-weighting. Under FIRB, certain parameters are prescribed by regulation (LGD, CCF) whereas under AIRB banks must develop internal models for their estimation.

Formally, for each credit exposure i, both in FIRB and in AIRB it is not computed the

associated risk weight, but directly its contribution to regulatory capital - also referred to as Risk Contribution (RC) - by applying the so-called IRB formula:

$$RC_i = f_{\text{IRB}}(PD_i, EAD_i, LGD_i, M_i) \tag{1.2}$$

where the function f_{IRB} is specified by the regulatory framework while the inputs - the risk parameters - must be estimated internally - at least in part, depending on FIRB or AIRB.

Before defining these parameters, it is necessary to clarify the notion of *default*. According to the BCBS, a default is deemed to have occurred when a counterparty has credit obligations past due for more than 90 consecutive days and/or when strong signals of financial difficulties are detected. On this basis, the IRB framework requires banks to estimate the following parameters:

- **Probability of Default (PD)**: the estimated probability that a counterparty will default within one year. Both FIRB and AIRB banks must rely on historical data to calibrate PDs for each internal rating grade.
- **Exposure At Default (EAD)**: it represents the amount of exposure that remains outstanding at the moment of default. Its estimation largely depends on the contractual features of the credit facility and can be straightforward or complex depending on the case. For a simple amortizing loan, the EAD corresponds to the outstanding principal that has not yet been repaid through scheduled instalments. For revolving facilities - such as credit lines - the calculation is more intricate: part of the exposure may already be drawn by the borrower - and is therefore certainly at risk - while another portion remains undrawn but can still be drawn down before default - thus becoming at risk as well. In general, EAD is computed as the sum of the drawn portion and the undrawn portion multiplied by a Credit Conversion Factor (CCF) - which represents the fraction of the unused credit line that is expected to be utilized at the time of default. Under FIRB, CCF values are fixed by regulation whereas AIRB banks develop their own methodologies.
- **Loss Given Default (LGD)**: it corresponds to the proportion of the exposure that is lost in the event of default. As with CCF, LGD is prescribed by regulation under FIRB whereas AIRB banks must estimate it internally.
- **Maturity (M)**: the effective remaining life of the exposure, reflecting both the timing and size of future cash flows. Under FIRB, M is set to fixed regulatory values whereas under AIRB it is calculated using the Macaulay duration formula - subject to regulatory bounds ($1 \leq M \leq 5$ years).

The institution - after having estimated the previous parameters for a credit exposure i - obtains the related risk contribution using Equation 1.2, which in its explicit version according to BCBS is:

$$\begin{aligned} RC_i = EAD_i \cdot \left[LGD_i \cdot \Phi\left(\frac{\Phi^{-1}(PD_i) + \sqrt{\rho(PD_i)}\,\Phi^{-1}(0.999)}{\sqrt{1-\rho(PD_i)}} \right) - LGD_i \cdot PD_i \right] \\ \cdot \frac{1 + (M_i - 2.5)\, b(PD_i)}{1 - 1.5\, b(PD_i)} \end{aligned} \tag{1.3}$$

The derivation of the previous formula and the definitions of its individual components are discussed in detail in Appendix A. Because RC_i is directly the capital requirement of the exposure i, and $RC_i = 8\% \cdot \text{RWA}_i$, we deduce that the risk-weighted asset amount for

the credit exposure i according to the IRB approach is equal to:

$$\begin{aligned} \text{RWA}_i = \frac{1}{8\%} \cdot EAD_i \cdot \Bigg[LGD_i \cdot \Phi\left(\frac{\Phi^{-1}(PD_i) + \sqrt{\rho(PD_i)}\,\Phi^{-1}(0.999)}{\sqrt{1-\rho(PD_i)}} \right) \\ - LGD_i \cdot PD_i \Bigg] \cdot \frac{1 + (M_i - 2.5)\, b(PD_i)}{1 - 1.5\, b(PD_i)} \end{aligned} \tag{1.4}$$

Therefore, as outlined in this section, **the primary purpose of credit risk modelling - with a specific focus on the AIRB approach at the core of this work - is the quantification of the fundamental credit risk parameters** (PD_i, LGD_i, EAD_i and M_i). The accurate estimation of these components for each exposure i is the crucial step that enables the institution to calculate its risk-weighted assets (RWA_i) according to the formulas imposed by the regulatory framework, thereby fulfilling capital adequacy requirements in a more precise manner that is aligned with the risk actually assumed.

1.2 Main steps in LGD models development

Having in mind the general purposes of credit risk modelling, from now on this work will focus on one of the risk parameters: the Loss Given Default (LGD). As previously mentioned, LGD represents the proportion of the exposure that is lost in the event of default. Specifically, **LGD refers to the economic loss rather than the accounting loss. Hence, all costs - and potentially also benefits - must be properly taken into account when defining the LGD.** Examples of costs include the expenses for realising collateral value, administrative costs related to collection activities - such as sending letters or making telephone calls to defaulted obligors - and legal costs. At the same time, benefits such as late-payment interest, penalty fees or other commissions may also be considered.

The development of an LGD estimation and prediction model is preceded by the phase in which observed LGDs - related to the institution's portfolio of credit exposures - are calculated. These observed LGDs serve as the ground truth for the development of statistical models. There are two main methodologies for computing LGD: the *workout* method and the *market* approach.

The workout method is based on discounting, at the time of default, the net cash flows actually realised during the recovery process. Therefore, in order to estimate LGD through this methodology, each defaulted exposure must be monitored over time, with all relevant information explicitly recorded, such as balances, cash flows, recovery costs and any other useful details. It is important to note that the workout LGD is not known either at the time the facility is granted or when the borrower defaults, but only once the recovery process has been completed. Under this approach, the LGD is given by:

$$LGD = \frac{EAD - \sum_{t=1}^{T} CF_t \cdot (1 + r_t)^{-t}}{EAD} \tag{1.5}$$

where CF_t and r_t denote, respectively, the cash flow and the discount rate at time t.

Considering Equation 1.5, the value of LGD typically ranges between 0 and 1: it equals 0 in cases where the recovery process leads to the full repayment of the defaulted exposure, and 1 when the entire exposure is lost. However, in some situations LGD may fall outside this range. For instance, it can take negative values if the borrower returns to performing status and late-payment interests collected exceed the recovery costs. Conversely, LGD

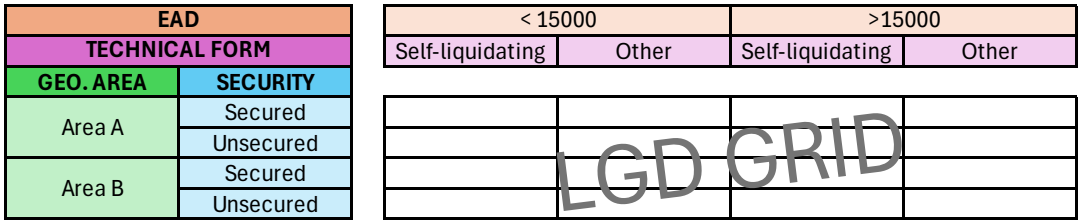


Figure 1.1: Example of an LGD grid where the risk drivers considered are the EAD, the credit facility type, the presence of collateral and the geographical area. As shown in the table, all risk drivers are treated as discrete variables with two distinct values each, including EAD, which could in principle be a continuous variable. This discretisation serves to reduce the granularity of values, mitigate noise and ensure a tractable grid representation, a topic that will be discussed in more detail later.

may exceed 1 when no recovery is achieved and, in addition to the total loss of the exposure, further indirect costs are incurred.

On the other hand, the market approach, relies on observing the prices of debt instruments issued by firms that have defaulted. Once bankruptcy occurs, outstanding bonds or loans become distressed instruments and investors trade them at prices that reflect their expectations about recoveries. These market prices are then used as a proxy for LGD estimation.

According to paragraph 6.1.1 of the EBA Guidelines [1], institutions applying the AIRB approach are not allowed to rely exclusively on the market-based methodology. Instead, the workout method is considered the reference approach, although external data may be integrated with internal observations.

Once the historical sample of observed LGDs has been collected, the next step under the AIRB approach is to develop an internal model capable of estimating LGD for each facility on the basis of a set of observed risk factors. According to the EBA Guidelines [1], model development generally follows two main steps:

1. **Risk differentiation**: The purpose of risk differentiation is to **identify the best set of risk drivers that make it possible to discriminat LGD values**. The outcome consists of groups of exposures that exhibit homogeneous risk drivers within each class, heterogeneous risk drivers across classes and different riskiness in terms of LGD. Typically, the output of this phase is referred to as the *LGD grid*, where the significant risk drivers are crossed to generate all possible combinations of classes, each of which is assigned its final LGD value. An example of a possible LGD grid is shown in Figure 1.1. In this phase, only exposures with completed recovery processes are used, as they are the only ones providing LGD values that are representative of the entire recovery process.

2. **Risk quantification**: This phase of the estimation process consists in applying calibration factors to the LGD level of each class obtained in the previous step, in order to reflect either the *long-run average LGD* (LRALGD) or, where more conservative, to the *downturn LGD* appropriate for a recessionary period. At this stage, all defaults observed in the available historical period (falling within the scope of the model) must be considered, including both closed and still active recovery processes - referred to as open positions. The long-run average LGD is then calculated as the arithmetic mean of realized LGDs, weighted by the number of defaults. This value is subsequently calibrated to incorporate the potential impact of a recessionary phase,

obtaining downturn LGD.

In this thesis, we will focus specifically on the risk differentiation phase, developing statistical and machine learning models capable of identifying not only the most relevant risk drivers for discriminating observed LGD values, but also the underlying relationships between such drivers and LGD itself. In the following section, we will analyse the case of Intesa Sanpaolo, examining how its internal practices fit into the general framework just outlined and discussing the challenges that the models proposed in the following may help to address.

1.3 A focus on the *Intesa Sanpaolo* case

Intesa Sanpaolo Group is one of the leading banking groups in Italy and within the euro area and it is classified among the *significant institutions* under the European Union's prudential supervisory framework. According to the public disclosure as of 31 March 2025, the Group reported a total capital ratio (CAR defined in Equation 1.1) of 18.50%. Specifically, this reflects total own funds of € 56,370 million against RWA amounting to €304,636 million, of which € 226,899 million stem from credit risk.

With regard to credit risk, Intesa Sanpaolo applies the AIRB approach to almost its entire portfolio, with the exception of certain segments - such as Non-Banking Financial Institutions and Sovereigns - where the very limited number of defaults makes the use of internal models less meaningful. Concerning the LGD parameter, according to [2], the Group has been authorised to use the AIRB method for the following portfolio segments:

- **Mortgage**: including residential mortgage loans to private individuals.
- **Corporate**: including banking products, leasing and factoring to companies with exposures exceeding €1 million or consolidated turnover above €2.5 million.
- **SME Retail**: including banking products, leasing and factoring to small and medium-sized enterprises not included in the corporate segment.
- **Other Retail**: including loans to private individuals other than residential mortgages.
- **Banks and Public Sector Entities**.

In this work, in collaboration with the *Credit Risk Accelerator* of Intesa Sanpaolo, a potential methodological evolution for LGD estimation will be explored. Specifically, focusing on the risk differentiation of the *Other Retail* segment, the aim is to move away from the traditional LGD grid approach (see Section 1.2) in favour of a less parsimonious method.

On the one hand, the grid-based approach offers several advantages. It allows for an intuitive interpretation of LGD estimates and an easy assessment of their economic consistency. For instance, one can readily verify that LGD tends to be higher in cases where the risk drivers indicate economically adverse conditions, and lower in more favourable scenarios (e.g., higher LGD is expected for unsecured exposures or loans issued in geographical areas associated with weaker economic conditions). These features are appreciated by regulators and are fundamental for pricing credit instruments coherently, thus maintaining a competitive advantage in the credit market.

On the other hand, the current implementation of the grid-based approach imposes stringent parsimony constraints in the model development pipeline, limiting the potential of the underlying estimation models. As anticipated in Figure 1.1, it also requires discretization of the input risk drivers, which restricts their expressivity. **This study therefore**

investigates whether removing these parsimony constraints and abandoning the requirement for a predefined LGD grid can yield models with substantially improved predictive performance, potentially justifying the resulting reduction in interpretability.

To this end, the proposed methodologies are compared with a baseline model consisting of a highly parsimonious Decision Tree. By controlling the number of terminal leaves and the tree depth, this baseline model naturally produces a grid-like output, with each split of the tree assigning facilities to the corresponding LGD cell. In contrast, we relax the parsimony constraints on the number of leaves and maximum depth, aiming to achieve more deep and accurate models while sacrificing the strict grid representation. This allows us to further experiment with more complex ensemble methods, such as Random Forest and Gradient Boosting, and to introduce a novel *two-stage model* design, in which estimation occurs in two sequential phases - hence the name two-stage models.

It is important to note that, when parsimony constraints are relaxed, the estimation may become excessively granular, leading to a very high number of distinct LGD values. Since such an output is difficult to manage - and precisely for this reason a parsimonious grid is usually employed - the approach explored in this study consists of two phases. The first is an unconstrained estimation, based on the models discussed above, where no restrictions on granularity are imposed; by construction, this step is expected to deliver more accurate results than the baseline. The second is an aggregation phase, designed to reduce the number of distinct LGD values produced by the models. **The main finding will be that, even after this aggregation - which brings the level of granularity down to one comparable with the baseline - the higher performance is preserved, thus achieving a balance between accuracy and tractability**.

Chapter 2

Theoretical framework for LGD modelling

In this chapter, we provide the theoretical framework of LGD estimation by framing it as a supervised learning problem. We begin by defining the mathematical notation and stating the core underlying assumptions (see Section 2.1). Secondly, we formalize the estimation problem using standard statistical learning theory (see Section 2.2). Then, we present an overview of the main modelling approaches found in the literature, distinguishing between single-stage and two-stage frameworks (see Section 2.3). Finally, we compare these methodologies through the lens of statistical learning theory, with a focus on their learnability and generalization properties (see Section 2.4).

2.1 Assumptions and learning settings

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{X} \subseteq \mathbb{R}^d$ denote the d-dimensional input space, which includes the relevant explanatory features - referred to as *risk drivers* - such as borrower characteristics, loan details and macroeconomic indicators, and $\mathcal{Y} \subseteq \mathbb{R}$ denote the output space, representing continuous LGD target values, usually between 0 and 1 (see Section 1.2).

We denote by:

$$\mathbf{X} : (\Omega, \mathcal{F}) \to \mathcal{X}$$

a d-dimensional random vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in the risk drivers space $\mathcal{X}$. We write $\mathbf{x} \in \mathcal{X}$ to denote a realization of $\mathbf{X}$ - i.e., a realization of the random risk drivers - , and $\mathcal{D}_{\mathcal{X}}$ to denote its probability distribution:

$$\mathcal{D}_{\mathcal{X}}(A) = \mathbb{P}(\mathbf{X} \in A), \quad A \subseteq \mathcal{X}.$$

Finally, we denote by:

$$Y : (\Omega, \mathcal{F}) \to \mathcal{Y}$$

the one-dimensional random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in the output space $\mathcal{Y}$. We write $y \in \mathcal{Y}$ to denote a realization of Y - i.e., a realization of LGD target value.

The learning framework considered in this work relies on three standard assumptions, widely adopted in the statistical learning literature: the existence of a deterministic target function, a non-realizable hypothesis class and an i.i.d. sample forming the training set. We detail each below.

Assumption 1 (Deterministic target function)**.** *We assume the existence of a fixed but unknown target function $f : \mathcal{X} \to \mathcal{Y}$ such that:*

$$Y = f(\mathbf{X}) \quad \mathbb{P}\text{-a.s.}$$

Equivalently, any realization $\mathbf{x} \in \mathcal{X}$ is mapped deterministically to a unique $y = f(\mathbf{x})$.

This implies that two credit exposures with identical features in $\mathcal{X}$ will yield the same LGD outcome. While this promotes consistency and interpretability - both highly desirable in credit risk modelling - it is also a strong assumption. In practice, latent variables and inherent randomness may introduce noise that $\mathcal{X}$ alone cannot fully capture. To address this modelling simplification, one often abandons the deterministic setting in favour of a stochastic one, where the conditional distribution of the target variable Y given $\mathbf{X} = \mathbf{x}$ is not degenerate, i.e., Y is not uniquely determined by $\mathbf{x}$ via a deterministic function. However, in this work, we adopt the deterministic assumption as a simplifying foundation for theoretical clarity.

The aim of the estimation process is to approximate the target function f using a suitable function h from a fixed set of functions $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$, called *hypothesis class*.

Assumption 2 (Non-realizable setting)**.** *We assume that the true target function f may not belong to the hypothesis class $\mathcal{H}$, i.e., $f \notin \mathcal{H}$. Consequently, even with access to infinite data, the algorithm may not be able to exactly recover f.*

This reflects a key limitation in practical learning: hypothesis classes are often constrained by statistical, computational or domain-specific considerations and may lack the expressiveness to fully capture the true data-generating process.

Assumption 3 (i.i.d. credit portfolio)**.** *We assume access to a credit portfolio $\mathcal{P}$ consisting of n realizations of the pair $(\mathbf{X}, Y)$, drawn iid from $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$, with $Y = f(\mathbf{X})$ $\mathbb{P}$-a.s.. Denoting by $\mathbf{x}_i$ and y_i the observed realizations of $\mathbf{X}$ and Y, we have $y_i = f(\mathbf{x}_i)$.*

The credit portfolio $\mathcal{P}$ is partitioned into three disjoint subsets: the training or in-sample set $\mathcal{P}_{\text{IS}}$, the test or out-of-sample set $\mathcal{P}_{\text{OOS}}$ and the out-of-time set $\mathcal{P}_{\text{OOT}}$, with $\mathcal{P}_{\text{IS}}$ used for model development [1]. Without loss of generality, we assume that the observations in the training set correspond to the first m entries of the credit portfolio, i.e., $\mathcal{P}_{\text{IS}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$.

On the one hand, the i.i.d. assumption facilitates the application of classical learning theory results, such as generalization bounds. On the other hand, in real-world credit portfolios, mild violations of this assumption may arise. A common example is when the overall population comprises distinct homogeneous subgroups of borrowers. In such cases, the data are not identically distributed, as the population exhibits latent heterogeneity. In this work, we adopt the i.i.d. assumption as the default setting to ensure analytical tractability, but we will explicitly relax it in those modelling frameworks where heterogeneity is accounted for.

2.2 Formalization of the LGD estimation problem

Given the assumptions and using the notation introduced in the previous section, the LGD estimation task can be formalized as follows. A learning algorithm $\mathcal{A}$ receives as

[1] While the in-sample and out-of-sample sets are always disjoint, the out-of-time set may overlap with either the in-sample or the out-of-sample set, or with both.

input a finite training set $\mathcal{P}_{\text{IS}}$ and returns as output the hypothesis $h_{\mathcal{P}_{\text{IS}}} = \mathcal{A}(\mathcal{P}_{\text{IS}})$, which corresponds to the *best* function within a predefined hypothesis class $\mathcal{H}$. This function $h_{\mathcal{P}_{\text{IS}}}$ represents the trained regression model used to estimate LGD values across the entire dataset, including in-sample, out-of-sample and out-of-time observations[2]:

$$\hat{y}_i = h_{\mathcal{P}_{\text{IS}}}(\mathbf{x}_i), \quad i \in [n].$$

To assess the quality of a given hypothesis $h \in \mathcal{H}$ and to select the *best* one, we ideally rely on its generalization error $R(h)$, which measures the expected loss over the data-generating distribution. Given a suitable loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, it is defined as:

$$R(h) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}} \left[\ell(h(\mathbf{x}), f(\mathbf{x})) \right].$$

where $h(\mathbf{x})$ is the predicted value by the hypothesis h and $f(\mathbf{x})$ is the observed value. The optimal hypothesis $h^\star$ within the hypothesis set $\mathcal{H}$ is the one that minimizes this quantity:

$$h^* := \arg\min_{h \in \mathcal{H}} R(h).$$

Since both the distribution $\mathcal{D}_{\mathcal{X}}$ and the true function f are unknown, the generalization error cannot be computed directly. In practice, for a given hypothesis $h \in \mathcal{H}$ the algorithm evaluates the empirical error $\hat{R}_{\mathcal{P}_{\text{IS}}}(h)$ over the training sample $\mathcal{P}_{\text{IS}}$:

$$\hat{R}_{\mathcal{P}_{\text{IS}}}(h) := \frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)^{3}, \quad (\mathbf{x}_i, y_i) \in \mathcal{P}_{\text{IS}}, \, \forall \, i \in [m]$$

and, according to the Empirical Risk Minimization (ERM) principle, it returns the empirical risk minimizer:

$$h_{\mathcal{P}_{\text{IS}}} = \mathcal{A}(\mathcal{P}_{\text{IS}}) := \arg\min_{h \in \mathcal{H}} \hat{R}_{\mathcal{P}_{\text{IS}}}(h).$$

Under ERM, the selected hypothesis $h_{\mathcal{P}_{\text{IS}}}$ is not necessarily the one with lowest generalization error h^*, but the one that best fits the training data. As a result, the model may also capture noise or spurious patterns, leading to overfitting and poor out-of-sample performance. A common approach to mitigate this risk is to constrain the complexity of the hypothesis class $\mathcal{H}$ so as to balance model flexibility and robustness.

This consideration highlights the importance of studying the so-called learnability of the chosen hypothesis class $\mathcal{H}$, that is, its capacity to yield hypotheses that generalize well when trained on sufficiently large samples. In this context, **learnability ensures that minimizing the empirical error translates into strong generalization performance, namely a low generalization error**. We now review the models commonly used in the LGD literature and examine their learnability properties.

2.3 Academic state of art in LGD estimation

The introduction of the Basel II framework has stimulated a growing body of literature focused on LGD estimation. The proposed models can be broadly categorized into two families: *single-stage* and *two-stage* modelling frameworks.

[2]This highlights a key aspect of statistical and machine learning model development: only the training data is used to fit the model, but predictions are made on the full dataset, including in-sample, out-of-sample (OOS) and out-of-time (OOT) examples.

[3]Specifying the loss function $\ell(\cdot, \cdot)$ leads to common evaluation metrics. In this work, we adopt the Mean Squared Error (MSE), defined as $\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} \left(h(\mathbf{x}_i) - y_i \right)^2$, which corresponds to the ℓ_2 loss, and the Mean Absolute Error (MAE), defined as $\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |h(\mathbf{x}_i) - y_i|$, corresponding to the ℓ_1 loss. These metrics allow us to assess LGD estimation accuracy under different sensitivity assumptions.

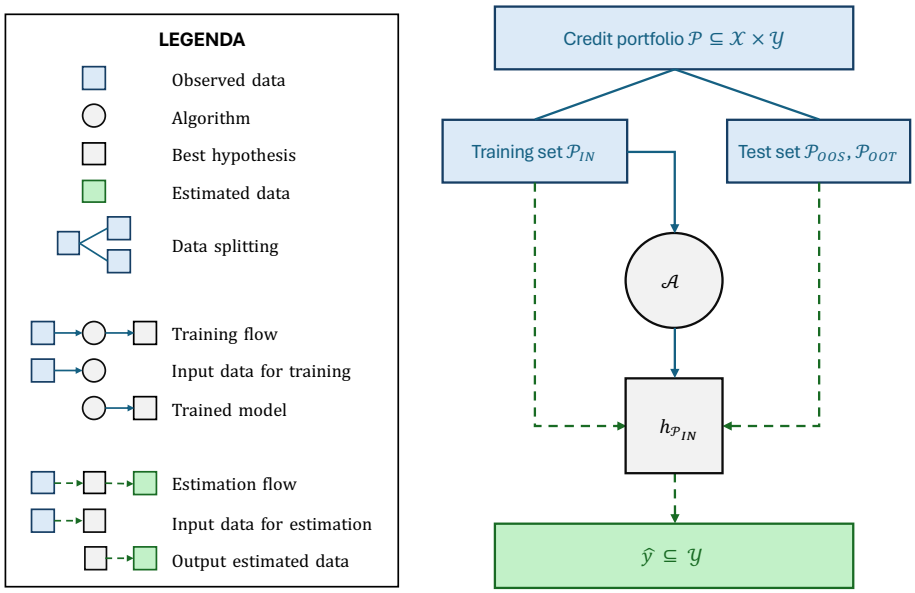


Figure 2.1: The training and estimation processes of single-stage models are illustrated with a diagram consisting of arrows and blocks, whose meaning is explained in the legend.

In single-stage frameworks, a unique algorithm $\mathcal{A}$ receives as input the training credit portfolio $\mathcal{P}_{\text{IS}} = \{(\mathbf{x}_i, y_i)\}_{i\in[m]}$ and outputs a unique regression model $h_{\mathcal{P}_{\text{IS}}}$. Then, LGD is directly estimated by applying the regression model to a set of realizations of the explanatory variables:

$$\hat{y}_i = h_{\mathcal{P}_{\text{IS}}}(\mathbf{x}_i), \quad i \in [n]. \tag{2.1}$$

Both the training and estimation workflows are illustrated in Figure 2.1. Specifically, the credit portfolio is split into a training set and a test set (blue line). The training set serves as input to the algorithm $\mathcal{A}$ to produce and tune the model $h_{\mathcal{P}_{\text{IS}}}$, thus concluding the training phase (blue arrow). During the estimation phase (green dashed arrow), both the training and test datasets are fed into the model $h_{\mathcal{P}_{\text{IS}}}$, which returns the estimated values $\hat{y}_i$, for $i \in [n]$. The test estimates are then used to evaluate the robustness of the model, that is, to assess whether the model chosen using the training set is sufficiently stable and able to generalize well to unseen data.

The approaches in this category include linear regression, beta regression, robust regression, regularized regression, regression trees, support vector regression and artificial neural networks, among others. Several studies have highlighted the superior performance of non-parametric models - particularly non-linear ones - such as those reported by Qi and Zhao (2011) [3] and Loterman et al. (2012) [4].

More recently, two-stage frameworks have gained increasing popularity. This trend is largely driven by the complex distributional characteristics of LGD, which typically deviates from normality and often exhibits bi-modal or multi-modal patterns. **Unlike single-stage models, they do not rely on the i.i.d. assumption for the population of defaulted borrowers** (see Assumption 3), **attributing the observed multi-modality to latent heterogeneity in the data**. As a result, these approaches introduce a preliminary segmentation step before estimation. **In the first stage, the population is divided into multiple homogeneous subgroups and, in the second stage, a separate LGD estimation model is trained for each subgroup**. Figure 2.2 illustrates the training and estimation workflows of the two-stage models.

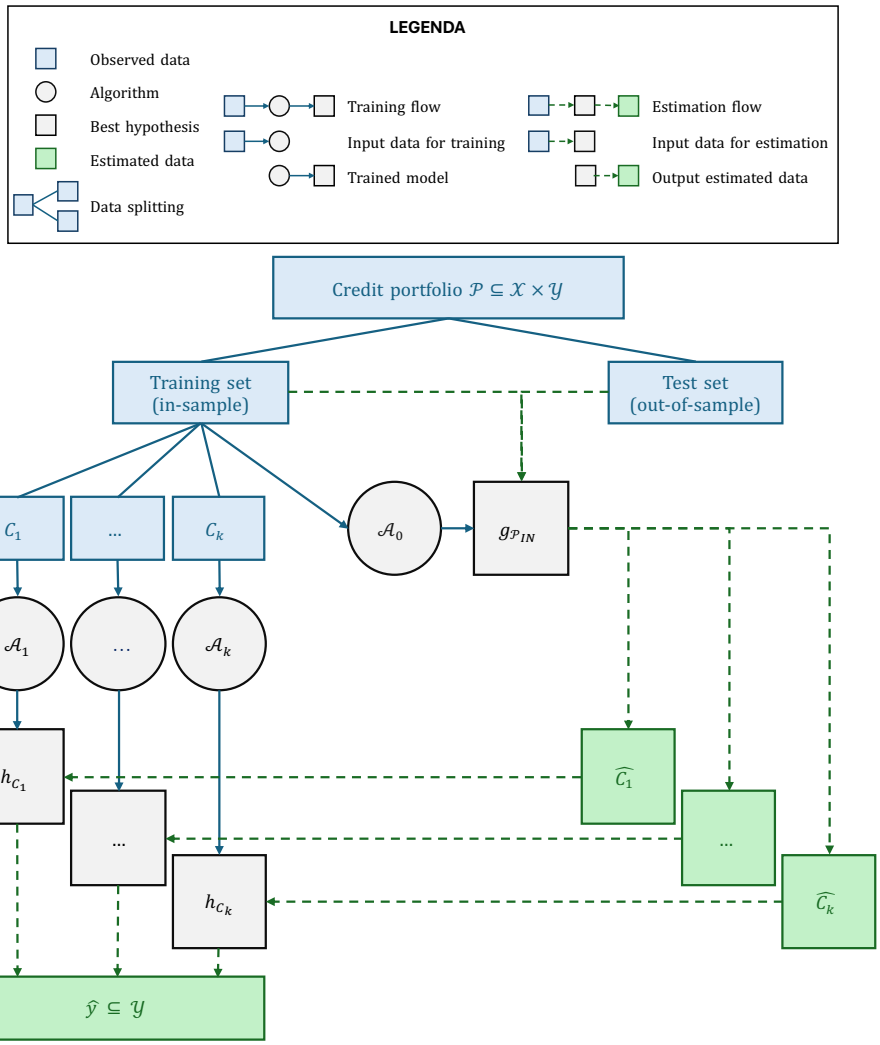


Figure 2.2: The training and estimation processes of two-stage models are illustrated with a diagram consisting of arrows and blocks, whose meaning is explained in the legend.

More formally, assuming the presence of k distinct homogeneous classes within the population of defaulted borrowers, the training set can be partitioned into k disjoint subsets based on the *observed* class membership: $\mathcal{P}_{\text{IS}} = C_1 \cup \cdots \cup C_k = \bigcup_{j=1}^{k} C_j$. A separate learning algorithm $\mathcal{A}_j$ is then applied to each subset C_j, resulting in a class-specific regression model h_{C_j} for $j = 1, \ldots, k$.

To assign observations to the appropriate class, class membership must be estimated rather than observed, as would be required when processing new data. For this purpose, a classification model is trained. Specifically, an algorithm $\mathcal{A}_0$ receives the entire training set $\mathcal{P}_{\text{IS}}$ and outputs a classifier $g_{\mathcal{P}_{\text{IS}}}$, which predicts the class membership for each observation in the form of a label in $\{1, \ldots, k\}$.

The final LGD estimate is obtained through a two-step prediction process: first, the classifier $g_{\mathcal{P}_{\text{IS}}}$ assigns each observation to a class constructing an estimated partition of the data, i.e. $\mathcal{P}_{\text{IS}} = \bigcup_{j=1}^{k} \hat{C}_j$, and then the corresponding regression model h_{C_j} is applied considering the *estimated* class membership. Formally:

$$\hat{y}_i = \sum_{j=1}^{k} \mathbb{1}\left(g_{\mathcal{P}_{\text{IS}}}(\mathbf{x}_i) = j\right) h_{C_j}(\mathbf{x}_i). \tag{2.2}$$

The main distinction among two-stage models lies in the methodology used for the initial segmentation in classes. One of the earliest examples is presented by Matuszyk et al. (2010) [5], who used logistic regression in the first stage to distinguish defaults with strictly positive LGD from those with zero or negative LGD. A separate regression model is then applied in the second stage to estimate LGD for the positive-LGD group.

Recognizing and demonstrating that a simple logistic regression may not adequately capture complex non-linear relationships between explanatory variables and LGD, Tanoue et al. (2020) [6] proposed using more advanced models for the segmentation step, but preserving the partitioning into non-positive and positive LGD observations. They experimented with random forests, k-nearest neighbours and support vector machines and found that random forests yielded the best overall performance.

Salko and D'Ecclesia (2021) [7] proposed a segmentation strategy based on the *work-out process* [4], classifying borrowers into three distinct categories: cures, partial recoveries and write-offs . Each class is handled with a dedicated regression model in the second stage, following multi-class classification using machine learning algorithms in the first stage.

Bosker et al. (2024) [8] used unsupervised cluster analysis to group borrowers according to their intrinsic similarities. This is a fully data-driven segmentation, without assuming *a priori* a class-wise partition. This procedure yielded economically meaningful and interpretable clusters, for each of which a separate predictive model was subsequently developed.

It is evident that two-stage frameworks offer numerous opportunities for extension, both thanks to the wide variety of possible combinations between parametric, non-parametric, linear and non-linear models - as explored by Loterman et al. (2012) [4] - and the availability of many segmentation criteria based on different underlying principles.

2.4 A theoretical comparison of modelling architectures

Given the different methodological approaches discussed earlier, we now present a theoretical comparison of the learnability of the corresponding hypothesis classes, using the formal framework introduced in Section 2.2. The goal is to assess the theoretical applicability and limitations of each modelling architecture and to provide a foundation that supports the empirical evaluation in the following sections.

In the context of binary classification, the conditions for learnability of a hypothesis class $\mathcal{H}$ are established by the **fundamental theorem of statistical learning**. This theorem allows us to reason, at least intuitively, about the relative merits of different modelling strategies, even beyond strictly binary classification tasks.

The theorem relies on two central notions. The first is the **VC-dimension** $VC(\mathcal{H})$, which measures the capacity or expressiveness of the hypothesis class $\mathcal{H}$. Intuitively, the more complex the hypothesis class, the higher its VC-dimension. Although originally defined for binary classifiers, the concept can be extended to real-valued functions through Pollard's pseudo-dimension. For simplicity, we refer only to the VC-dimension in this discussion.

The second notion is that of **sample complexity**, which quantifies the minimum number of training examples required to ensure that a model generalizes well. More precisely,

[4]The *work-out process* refers to the post-default recovery procedures through which the lender attempts to recover the outstanding exposure. This process typically ended in borrowers being classified into three categories: *cures*, when the borrower resumes regular payments; *partial recoveries*, when only a portion of the exposure is recovered; and *write-offs*, when the loss is considered irrecoverable.

given a hypothesis class $\mathcal{H}$, its sample complexity $n_{\mathcal{H}}(\epsilon, \delta)$ denotes the number of samples needed to guarantee, with probability at least $1 - \delta$, that the absolute difference between the true risk $R(h_{\mathcal{P}_{\text{IS}}})$ and the empirical risk $\hat{R}_{\mathcal{P}_{\text{IS}}}(h_{\mathcal{P}_{\text{IS}}})$ is at most ϵ.

Theorem 1 (Fundamental Theorem of Statistical Learning)**.** *In a binary classification setting, if a hypothesis class $\mathcal{H}$ is such that $VC(\mathcal{H}) < \infty$, then $\mathcal{H}$ is (Probably Approximately Correct) learnable with sample complexity $n_{\mathcal{H}}(\epsilon, \delta)$ satisfying:*

$$O\left(\frac{VC(\mathcal{H}) + \log(1/\delta)}{\epsilon}\right) \leq n_{\mathcal{H}}(\epsilon, \delta) \leq O\left(\frac{VC(\mathcal{H})\log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right) \tag{2.3}$$

This result has three main implications. First, a given hypothesis class $\mathcal{H}$ is learnable whenever its VC-dimension is finite; in other words, $\mathcal{H}$ is not too complex. Second, assuming a finite $VC(\mathcal{H})$, both the lower and upper bounds on the sample complexity scale linearly with it. Intuitively, this means that reducing the complexity of the hypothesis class - i.e., lowering $VC(\mathcal{H})$ - also reduces the number of samples required for successful learning roughly by the same factor.[5] Third, assuming that the sample size m is at least as large as the required sample complexity, the theorem gives a bound on the generalization error:

$$R(h_{\mathcal{P}_{\text{IS}}}) \leq \hat{R}_{\mathcal{P}_{\text{IS}}}(h_{\mathcal{P}_{\text{IS}}}) + O\left(\frac{\log\left(\frac{m}{VC(\mathcal{H})}\right)}{m/VC(\mathcal{H})}\right). \tag{2.4}$$

These theoretical relationships allow us to compare modelling architectures based on their expected generalization performance and data requirements. In a one-stage approach, a single global model is trained to approximate a function over the entire dataset $\mathcal{P}_{\text{IS}}$, using a hypothesis class $\mathcal{H}_0$. If we follow empirical evidences and we do not assume the validity of the Assumption 3, but we instead acknowledge the heterogeneity of the defaulted portfolio, the hypothesis class $\mathcal{H}_0$ may need to be highly expressive to achieve low approximation error $\hat{R}_{\mathcal{P}_{\text{IS}}}(h_{\mathcal{P}_{\text{IS}}})$. This typically results in a large VC-dimension $VC(\mathcal{H}_0)$, which in turn increases both the required sample complexity (see Equation 2.3) and the generalization gap (see Equation 2.4). In scenarios with limited training data, this can lead to overfitting.

In contrast, a two-stage approach first partitions the data into k smaller, potentially more homogeneous subsets $\{C_j\}_{j=1}^{k}$. A simpler model with hypothesis class $\mathcal{H}_j$ is then trained independently on each subset. Regardless of how the partitioning is performed (see Section 2.3), the underlying theoretical rationale remains the same: each sub-problem is expected to be simpler, allowing for a less complex hypothesis class such that $VC(\mathcal{H}_j) \ll VC(\mathcal{H}_0)$. This leads to a reduction in sample complexity and a tighter generalization bound for each sub-model.

However, this benefit comes at the cost of having fewer training examples per sub-model. Since each model is trained on a smaller subset of data, the effective sample size m_j satisfies $m_j \ll m$. The learnability is ensured only if the sample size is reduced by the same factor of the model complexity, in such a way not to violate Equation 2.3.

Against this background, **we can conclude that two-stage models have the potential to outperform single-model approaches in heterogeneous data settings, provided that the partitioning leads to sufficiently simpler and learnable sub-problems**. However, this

[5] This fact aligns with common intuition, where simpler hypothesis classes require less data to be learned. The interesting point is that both bounds scale at the same rate.

theoretical advantage does not account for the complexity of the classification task in the first stage. In fact, while the second-stage problems may indeed be simpler and more learnable, assigning each observation to the correct subgroup C_j is itself a non-trivial challenge, one that remains underexplored in the literature.

For instance, considering the segmentation proposed by Salko and D'Ecclesia (2021) [7], from a credit risk perspective, predicting the likely work-out outcome for a loan before it even defaults - as required by a two-stage model - is inherently complex. First, if it were already known that a borrower would be written off in case of default, the loan might not be granted in the first place. Second, the work-out process evolves over time with the evolution of borrowers' internal policies, implying that models must be continuously updated to duly reflect current practices and institutional policies.

Building on these considerations, we will employ the reviewed approaches as a foundation for the development of the novel methodologies proposed in this thesis. Before doing so, however, we introduce empirical validation metrics that will complement the theoretical analysis, thereby completing the framework for a practical comparison and evaluation of the different modelling strategies.

Chapter 3

Model validation under the ECB regulatory framework

In this chapter, we focus on the practical validation of LGD models in accordance with regulatory requirements. We begin by outlining the regulatory framework and the standardized validation tools established by the European Central Bank (see Section 3.1). Special emphasis is placed on tools aimed at assessing discriminatory power, including the generalized Area Under the Curve (gAUC) and Somers' d metrics, with a discussion of how these measures apply to LGD estimation (see Section 3.2 and Section 3.3). Finally, we examine potential pitfalls in the use of Somers' d and their implications for model evaluation and hyperparameter tuning (see Section 3.4).

3.1 LGD model validation tools

In accordance with the Capital Requirements Regulation (CRR), established by Regulation (EU) No 575/2013 of the European Parliament, credit institutions are required to subject their internal credit risk models to a formal validation process (see Article 185 of the CRR). This process, firstly conducted by the model developers and then by the internal validation function, aims to assess the overall adequacy, robustness and reliability of the internal estimates used to calculate own funds requirements. Furthermore, the validation review serves as the primary reference for supervisory authorities when carrying out tasks related to internal models, such as model approvals and ongoing performance monitoring.

Although all institutions must follow the same regulatory framework when validating credit risk models, comparing models across different banks remains challenging. To address this, the European Central Bank (ECB) requires institutions to submit additional validation reports using a standardised set of statistical tests and accuracy measures, known as *validation tools*. These tools provide a common basis for evaluating and comparing the predictive performance and robustness of internal models across institutions.

In February 2019, the ECB published the *Instructions for reporting the validation results of internal models* [9], which describes the scope of these validation tools, their methodological application and the associated reporting requirements. Regarding the LGD parameter, the validation tools focus on monitoring model performance in two key areas: predictive ability and discriminatory power. The analysis of predictive ability aims to ensure that LGD estimates provide reliable forecasts of realized loss rates. **The analysis of discriminatory power assesses the model's ability to distinguish between facilities** - i.e., credit exposures in the dataset - **with high and low LGD values and is particularly crucial for**

the validation of the risk differentiation phase. Since risk differentiation is the specific model development phase on which this work focuses (see Section 1.2), we will therefore place greater emphasis on the measurement of discriminatory power.

The measure used for this purpose is the generalized Area Under the Curve (gAUC), which extends the classical AUC metric to multi-class classification problems. Let us now further explore the discriminatory power related validation tools in depth.

3.2 Analysis of discriminatory power

As mentioned earlier, the analysis of discriminatory power is carried out using gAUC, a metric typically employed in multi-class classification problems. At first, referring to multi-class *classification* may seem inconsistent with the learning framework introduced in Chapter 2, where LGD is modelled as a continuous target variable and estimated through *regression* techniques. This apparent mismatch raises the question of how a classification-based metric like gAUC can be meaningfully applied in our regression context.

In practice, however, the ECB validation tool instructions [9] prescribe a post-processing discretisation procedure to be applied to the LGD estimates produced by the chosen regression model, making the regression task resemble an ordinal classification one, where gAUC becomes applicable.

To clarify this point, let us consider as example the simplified LGD grid in Figure 3.1, derived from the example in Figure 1.1 by excluding the geographical area driver. In this grid, each cell is associated with an LGD estimate, as if produced by a given regression model. The resulting structure of the example yields eight distinct estimated LGD values, or *grades*:

$$0.10,\ 0.15,\ 0.20,\ 0.25,\ 0.50,\ 0.60,\ 0.70,\ 0.80.$$

According to the post-processing discretisation procedure prescribed by [9], when the model produces 20 or fewer grades as in this example, the estimated LGDs are first sorted in ascending order and then mapped into *LGD segments* defined as follows:

- **Segment 0**: $LGD \leq 0.10$
- **Segment 1**: $0.10 < LGD \leq 0.15$
- **Segment 2**: $0.15 < LGD \leq 0.20$
- **Segment 3**: $0.20 < LGD \leq 0.25$
- **Segment 4**: $0.25 < LGD \leq 0.50$
- **Segment 5**: $0.50 < LGD \leq 0.60$
- **Segment 6**: $0.60 < LGD \leq 0.70$
- **Segment 7**: $0.70 < LGD \leq 0.80$
- **Segment 8**: $LGD > 0.80$

Observed LGDs - typically more granular and quasi-continuous - are then discretised consistently, by mapping them into the same segments defined from the estimated LGDs. At this stage, each credit exposure can be associated not only with its observed and estimated continuous LGDs, but also with the corresponding observed and estimated discrete segments, as shown in Figure 3.2.

This discretised representation makes the regression problem effectively resemble an ordinal classification task, where the target is the LGD segment. In such a setting,

EAD	< 15000		>15000	
TECHNICAL FORM	Self-liquidating	Other	Self-liquidating	Other
SECURITY				
Secured	20%	25%	10%	15%
Unsecured	70%	80%	50%	60%

Figure 3.1: Example of an LGD grid where the risk drivers considered are EAD, credit facility type and presence of collateral. The percentage values in the grey grid represent the LGD estimates.

SECURITY	EAD	TEC. FORM	EST. LGD	OBS. LGD	EST. SEGMENT	OBS. SEGMENT
Secured	< 15000	Self-liquidating	20%	22%	2	3
Unecured	> 15000	Other	60%	30%	5	4
Secured	> 15000	Other	15%	12%	1	1
Secured	< 15000	Other	25%	25%	3	3
Unecured	> 15000	Self-liquidating	50%	90%	4	8

Figure 3.2: Example of a credit portfolio containing three risk drivers (EAD, credit facility type and presence of collateral), along with observed LGDs, estimated LGDs according to the grid in Figure 3.1, and the corresponding observed and estimated LGD segments. The colours used in this table are designed to facilitate the cross-reference with the table in Figure 3.1.

classification-based metrics such as gAUC become both applicable and meaningful, as they measure the model's ability to correctly rank exposures by expected loss severity segment.

When the model produces more than 20 grades - as in the case of a more complex LGD grid or in the case of fully continuous LGD estimates not expressed as grids - the ECB instead prescribes a different discretization of the estimated LGD into 12 predefined segments, followed by the application of the gAUC metric. These segments are independent from the estimated values and are always defined as follows: $[0, 0.05)$, $[0.05, 0.10)$, $[0.10, 0.20)$, and then successive 10% steps up to $[0.90, 1.00)$, with a final segment for $[1.00, +\infty)$. This discretization makes it possible to apply gAUC even to continuous models by mapping their outputs into ordered bins, thus enabling a consistent assessment of discriminatory power using classification-based metrics like gAUC also in this case.

The models that will be proposed in this work do not take into account constraints of parsimony and they generate LGD estimates on a continuous scale, instead of on a LGD grid. However, as anticipated in Section 1.3, for tractability purposes these outputs are subsequently grouped into a *discrete LGD scale* using a hierarchical clustering–inspired technique, resulting in a discretised output. As will be shown in the empirical analysis presented in Chapter 4, this hierarchical clustering step will yield a relatively small number of grades - though always higher than 20. As a result, the validation procedure prescribed for models with more than 20 grades is consistently applied throughout this work.

In conclusion, in our regression context the use of the classification-related metric gAUC is justified by the discretisation step discussed above. Now, to better understand what the gAUC actually measures in practice, it is useful to explore its relationship with Somers' d statistic since it offers a more interpretable and intuitive perspective.

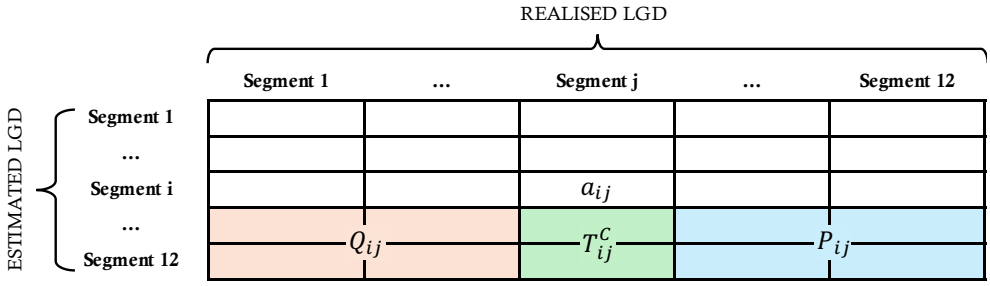


Figure 3.3: Two-way contingency table constructed according to regulatory guidelines for models producing more than 20 grades. Rows represent the 12 segments of estimated LGD, while columns represent the 12 segments of realised LGD. For each observation in a_{ij}, the corresponding concordant pairs are highlighted in blue, the discordant pairs are highlighted in orange and the columns tied pairs tied in green; in accordance with the definition of concordance, discordance and tie reported in the following.

3.3 gAUC and Somers' d in discriminatory power measurement

Somers introduced in 1962 a pair of *asymmetric* association coefficients [10], d_{YX} and d_{XY}, which are closely related to both Kendall's τ and Goodman and Kruskal's γ. These coefficients are specifically designed to measure the degree of monotonic association between two ordinal variables in a contingency table [1]. In our context, **Somers' d is used to quantify the ordinal relationship, or monotonic association, between predicted and observed LGD segments**.

The asymmetry of the metric lies in the fact that d_{YX} is used when X is the independent variable and Y is the dependent one, and vice versa for d_{XY}. Unlike Kendall's τ and Goodman-Kruskal's γ, which treat both variables symmetrically, Somers' D explicitly reflects the direction of prediction, from the independent to the dependent. This characteristic makes it particularly suitable for evaluating categorical prediction, as **it captures not only the ordinal relationship but also the extent to which knowledge of the independent variable improves the prediction of the dependent one**.

Let us now illustrate how this metric is computed, focusing on the case of validating a model that produces more than 20 grades, as in the models examined in this work. As discussed in Section 3.2, when dealing with models with more than 20 grades, a discretisation step into 12 predefined LGD segments must be performed. In this cases, according to the regulatory guidelines, the computation of Somers' d is based on the construction of a two-way contingency table that cross-classifies all possible combinations of discretised estimated LGD segments (LGD_i^E) as rows and realised LGD segments (LGD_i^R) as columns. As shown in Figure 3.3, this results in a 12×12 table, since the predefined LGD segments are 12 both for estimated and realised LGDs.

Let a_{ij} denote the observed frequency in cell (i, j), i.e., the number of exposures for which the estimated LGD belongs to segment i while the realised LGD belongs to segment j.

[1] Perfect monotonic correlation refers to a situation in which, for two variables X and Y, the value of X increases whenever Y increases, and conversely, regardless of the (possibly varying) rate of increase. Linear correlation is a special case of monotonic correlation. In general, the concept allows for curvilinear relationships, provided that the function does not "double back" on itself. More formally, for each X there is a unique corresponding value of Y, and vice versa.

On this basis, we define the following quantities:

$$P_{ij} := \sum_{k>i} \sum_{l>j} a_{kl}, \quad Q_{ij} := \sum_{k>i} \sum_{l<j} a_{kl}, \quad T^C_{ij} := \sum_{k>i} a_{kj}$$

$$P := \sum_i \sum_j a_{ij} P_{ij}, \quad Q := \sum_i \sum_j a_{ij} Q_{ij}, \quad T^C := \sum_i \sum_j a_{ij} T^C_{ij}$$

Specifically, P represents the total number of *concordant pairs*. For a given observation in cell (i, j), a concordant pair is any other observation with both a higher estimated segment ($k > i$) and a higher realised segment ($l > j$), corresponding to the elements in the cell highlighted in blue in Figure 3.3. Thus, P_{ij} counts the number of concordant pairs associated with a single observation in (i, j). By multiplying this quantity by a_{ij}, we obtain the total number of concordant pairs with respect to all observations contained in that cell. Summing over all cells finally yields P, the overall number of concordant pairs.

Analogously, Q denotes the total number of *discordant pairs*. For an observation in cell (i, j), a discordant pair is any other observation with a higher estimated index ($k > i$) but a lower realised index ($l < j$), corresponding to the elements in the cell highlighted in orange in Figure 3.3. Here Q_{ij} gives the number of discordant pairs for a single observation in (i, j), and multiplying by a_{ij} extends this to all observations in the cell. Summing across all cells provides Q, the total number of discordant pairs.

Finally, T^C corresponds to the total number of *ties on the column variable*. For an observation in cell (i, j), the ties on the column variable are the cases where exposures share the same realised segment j (column) but differ in their estimated segment i (row), corresponding to the elements in the cell highlighted in green in Figure 3.3. As before, T^C_{ij} measures such ties for a single observation in cell (i, j), and multiplying by a_{ij} gives the contribution of the whole cell.

In accordance with the validation tool instructions provided by the ECB [9], we treat the row variable LGD^E as the independent variable and the column variable LGD^R as the dependent variable, and we compute the metrics gAUC and Somers' $d(C|R)$ - that is, Somers' d of the column *given* the row, which reflects the assumed direction of dependency - as follows:

$$d(C|R) := \frac{P - Q}{P + Q + T^C} \in [-1, 1] \tag{3.1}$$

$$gAUC := \frac{d(C|R) + 1}{2} \in [0, 1] \tag{3.2}$$

Equation 3.2 highlights the direct connection between the two metrics, justifying the use of Somers' d as an alternative to gAUC. In fact, since one can be derived from the other, all insights obtained from the gAUC can be equivalently expressed through Somers' d, and vice versa. Somers' d, which is often directly referred to as the *Accuracy Ratio* (AR), is formally defined in Equation 3.1 as the **difference between the number of concordant and discordant pairs, normalized by the total number of *comparable pairs* - i.e., all pairs not tied on the independent variable. Importantly, it is precisely this definition of comparable pairs that introduces the asymmetry in the metric: when one of the two variables is designated as the independent variable, all pairs tied on that variable are excluded from the count of comparable pairs.**

The (empirical) model validation using Somers' $d(C|R)$ will complement the (theoretical) assessment in terms of learnability and sample complexity introduced in Chapter 2, and will also serve as the main reference metric for hyperparameter tuning of the models

under analysis. Before proceeding, however, it is important to highlight some potential pitfalls associated with the use of Somers' $d(C|R)$, which are discussed in the following section.

3.4 Pitfalls in using Somers' d

In his 1962 paper [10], Somers investigates these two asymmetric association measures in a sociological context, where the distinction between independent and dependent variables is well defined. Specifically, he refers to the independent variables as *stimuli* or *indicators*, and the dependent ones as *responses*. This framing clearly establishes the direction in which the asymmetry of Somers' metric should be interpreted: in d_{YX}, the variable X represents the independent stimulus, while Y represents the dependent response. A classical example provided by Somers illustrates this distinction: the variable *years of school completed* can be considered a stimulus or indicator X, while a corresponding response variable Y might be the *amount of knowledge acquired*. The assumption is that the former contributes to or influences the latter, justifying its role as the independent variable in the analysis.

As previously mentioned, within the context of LGD model validation, the regulatory guidelines [9] specify that the estimated LGD should be treated as the independent variable and the realised LGD as the dependent one, leading to Somers' $d(C|R)$. At first, this might seem counterintuitive, since it reverses the usual modelling perspective in which the observed value is considered the input and the predicted value the output. Despite the direction of the asymmetry is less clear than in the sociological context, this choice has a clear operational justification [2].

Indeed, since by construction Somers' $d(C|R)$ captures both ordinal association and the ability to predict categorical outcomes (see Section 3.3), the underlying question being asked becomes: how much more likely is it that a facility with a given higher estimated LGD (R, rows) will result in a higher realised LGD (C, columns)? If the roles were inverted - and Somers' $d(R|C)$ is considered treating the observed LGD as the independent variable - the question would become: how well does a higher given realised LGD (C) help in predicting a higher estimated one (R)? This would defeat the purpose of model validation, as it no longer tests the model's ability to predict loss severity.

Despite the operational justification discussed above, **treating the estimated LGD as the independent variable may introduce a potential bias**. As shown in Equation 3.1, the effect of dependency direction mainly lies in the computation of Somers' d denominator: **in addition to concordant and discordant pairs, only pairs tied on the dependent variable (i.e., realised LGD in the columns) are included, while ties on the independent variable (i.e., estimated LGD in the rows) are excluded**. This choice reflects the idea that pairs tied on the independent variable do not provide information about concordance or discordance.

As a result, if a model produces "lazy" estimates that fall into only a small number of estimated LGD segments - leading to many ties on the independent variable (i.e., along the rows) - the total number of comparable pairs in the denominator decreases significantly. In fact, most pairs end up tied on the rows and therefore do not fall into ties on the columns (T^C), nor into the sets of concordant (C) or discordant (D) pairs. This reduces the number of comparable pairs and, consequently, the denominator. This can artificially inflate the

[2] It is important to note that no direction of dependence is justified by causal relationships; this is precisely what makes the choice of dependence direction less straightforward, and allows, as we will see later, the use of both directions to fully assess the model under validation.

ESTIMATED LGD \ REALISED LGD	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6	Segment 7	Segment 8	Segment 9	Segment 10	Segment 11	Segment 12
Segment 1	-	-	-	-	-	-	-	-	-	-	-	-
Segment 2	-	-	-	-	-	-	-	-	-	-	-	-
Segment 3	1	1	2	2	2	2	2	2	2	2	2	-
Segment 4	-	-	-	-	-	-	-	-	-	-	-	1
Segment 5	-	-	-	-	-	-	-	-	-	-	-	-
Segment 6	-	-	-	-	-	-	-	-	-	-	-	-
Segment 7	-	-	-	-	-	-	-	-	-	-	-	-
Segment 8	-	-	-	-	-	-	-	-	-	-	-	-
Segment 9	-	-	-	-	-	-	-	-	-	-	-	-
Segment 10	-	-	-	-	-	-	-	-	-	-	-	-
Segment 11	-	-	-	-	-	-	-	-	-	-	-	-
Segment 12	-	-	-	-	-	-	-	-	-	-	-	-

Figure 3.4: A two-way contingency table constructed according to regulatory guidelines, with realized LGD segments as columns and estimated LGD segments as rows. Here, we illustrate an example of a lazy estimate, i.e., an estimate where most of the estimated LGD values fall into a few segments - Segment 3 in this example - with only a single observation in Segment 4, despite a range of realized LGD values across different segments.

value of Somers' D, even when the model lacks real discriminative and predictive power. In such cases, a lazy model producing predictions concentrated in only a few segments, may appear to perform well under Somers' $d(C|R)$ despite limited predictive usefulness.

To better illustrate this effect, consider Figure 3.4, in which estimated LGD values fall exclusively into segments 3 and 4, producing what we have called a "lazy" estimate. In terms of monotonic association, the relationship appears perfect, resulting in a Somers' $d(C|R) = 1$:

$$P = 20, \quad Q = 0, \quad T^C = 0$$
$$d(C|R) = \frac{20 - 0}{20 + 0 + 0} = 1$$

Does perfect accuracy imply a perfect model? On the one hand, the monotonic correlation is indeed perfectly satisfied. On the other hand, because this metric also aims to capture how well the independent variable can predict the dependent one, it becomes clear that the model's predictive power is limited. For example, knowing that an LGD estimate falls in segment 4 allows for a precise prediction, but knowing that it falls in segment 3 offers no such certainty.

Exploring the issue of considering both predictability power and monotonic correlation when the prediction is made from a smaller number of segments (Segment 3 and Segment 4 in the example) to a wider range of outcomes (from Segment 1 to Segment 12 in the example), Somers [10] suggests - perhaps counterintuitively - to reverse the roles of the independent and dependent variables.

Returning to the example shown in Figure 3.4, according with this proposed solution, we now consider the realised LGD (columns) as the independent variable. In this way, P and Q remain unchanged, while the tied pairs that should be now taken into account are the ties on estimated LGD (rows), which is equal to $T^R = 181$. Therefore, the reversed Somers' d is:

$$d(R|C) = \frac{20 - 0}{20 + 0 + 181} = 0.099$$

As seen in the example, this proposed solution would include in the denominator all pairs tied on the estimated LGD (ties on the rows), substantially increasing the number of comparable pairs and, consequently, decreasing the value of Somers' d. In this way,

models that produce lazy estimates - i.e., concentrated in a small number of segments - are no longer rewarded, and the bias is mitigated.

This result supports the conclusion that **while Somers' $d(C|R)$ - as prescribed by regulatory guidelines - is certainly a valid performance metric, in specific model designs where lazy estimates are possible to be produced it should be complemented by the analysis of Somers' $d(R|C)$ to detect and avoid the bias described above.**

One could argue that, even when reversing the direction of dependence, the same issue might arise: to increase the AR, one could create more tied pairs on the variable considered as independent. The crucial point is that if the estimate is treated as independent, the model could exploit this bias by producing "lazy" and tied estimates. Conversely, if the realised observations are treated as independent, it is not possible to artificially create ties, since the observations are not under the model's control.

One notable example of model design that could exploit this bias is Gradient Boosting. This method builds a sequence of decision trees that iteratively refine predictions, starting from an initial constant estimate and gradually introducing complexity through a learning process. If Somers' $d(C|R)$ is used to determine when to stop this learning process - e.g., for hyperparameter tuning - it may favour early stopping, when the estimates are still overly coarse and concentrated in a few segments, resulting in inflated scores.

A second example, referring back to the model classes introduced in Section 2.3, is a two-stage model where the first stage assigns nearly all observations to the same class C_i, and the associated regressor returns LGD estimates within a very narrow range. This design would again lead to a large number of tied estimates and an artificially high value of Somers' $d(C|R)$, despite limited practical usefulness.

Chapter 4

Empirical validation of the proposed methodologies

This chapter represents the empirical core of the thesis, where the theoretical concepts, modelling architectures, and validation frameworks discussed previously are applied and tested. The primary objective is to empirically validate the proposed methodological evolution, demonstrating its ability to produce LGD estimates that are both more accurate and more parsimonious than those derived from traditional approaches.

The analysis is structured in three main parts. We begin with an exploratory data analysis, outlining the key statistical features of the illustrative dataset (see Section 4.1). Subsequently, we present the implementation and the comparative performances of the proposed single-stage and two-stage models against the parsimonious baseline (see Section 4.2). Finally, we demonstrate how the proposed post-aggregation technique effectively resolves the trade-off between accuracy and granularity, reducing the number of distinct LGD estimates to a manageable level while preserving the significant performance improvements achieved in the unconstrained estimation phase (see Section 4.3).

The models developed in this work are based on Intesa Sanpaolo's *Other retail* dataset. During the collaboration with the Group, the original data were analysed and used to derive insights that directly informed the model development. For confidentiality reasons, however, the dataset underlying those analyses cannot be disclosed in this thesis.

Instead, the empirical evidence presented here is based on a distorted sample, constructed by applying subsampling procedures and distortion factors to the original portfolio. As such, results, numerical values and figures reported in this chapter should not be interpreted as a faithful representation of Intesa Sanpaolo's data; rather, they are provided for illustrative purposes only, to reproduce in a consistent way the main conclusions of analysis on the true dataset.

4.1 Empirical data analysis

The here considered sample consists of 5957 facilities, for each of which the following information is available: the observed LGD (calculated using the workout method), a flag indicating whether the recovery process ended in cure or non-cure, and a set of 24 risk drivers.

In Figure 4.1, a histogram showing the empirical distribution of the observed LGD is presented together with the corresponding boxplot. The distribution clearly exhibits a

bimodal shape, with one mode at 0 and another around 1, which is a typical feature of LGD data. The predominance of the mode near zero, further confirmed by a median LGD of approximately 0.225, suggests that the majority of defaults in this portfolio ultimately end either in a cure or in relatively low loss severity, rather than resulting in a total loss.

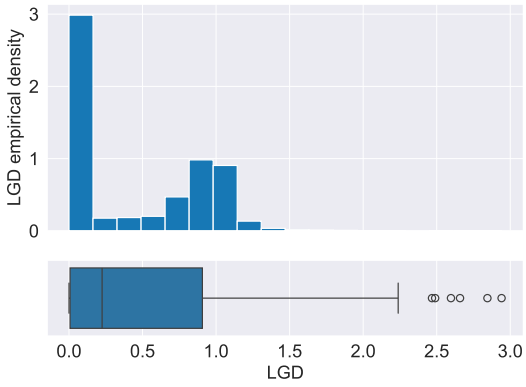


Figure 4.1: Empirical distribution of observed LGDs in the illustrative sample, together with the corresponding boxplot.

Statistic	Value
Mean	0.453
Std. Dev.	0.466
Min	0.000
25th Pctl	0.007
Median	0.225
75th Pctl	0.907
Max	2.941

Table 4.1: Descriptive statistics of observed LGDs in the illustrative sample.

In addition to the median, further descriptive statistics are reported in Table 4.1. They show that LGD values range from 0 to 2.941, although 50% of the distribution is concentrated between 0.007 (the 25th percentile) and 0.907 (the 75th percentile), while the 99th percentile reaches 1.40. This indicates the presence of outliers above one, also confirmed by the boxplot in Figure 4.1. By contrast, no negative LGD values are observed, since the results were floored at zero. This flooring procedure also accounts for the high concentration of mass around zero, which would otherwise have extended into the negative domain.

The 24 risk drivers available in the dataset can be classified into four broad categories:

1. **Customer information**: variables that describe the characteristics of the counterparty, which in the case of the *Other Retail* segment is always an individual. Typical examples include geographical area, age, years of banking relationship (banking tenure) and occupation, among others.

2. **Facility information**: variables that describe the contractual features of the specific credit facility under consideration. These usually include the technical form of the loan, the exposure amount and the type of product.

3. **Collateral-related variables**: indicators that capture the presence or absence of collateral, its type and its nominal value.

4. **Financial indicators of the counterparty**: measures of the economic and financial standing of the borrower. In the case of retail customers, these typically refer to statistics related to account balances, while for corporates they may include financial statement ratios.

In the original development of the internal LGD models, all 24 drivers were jointly considered in order to build the estimation framework. However, for the purposes of this empirical data analysis we restrict attention to only four illustrative features: the boolean indicating whether the recovery process ended in cure or non-cure, hereafter referred to as the *flag cured*; *banking tenure*, representing customer information; *EAD*, representing facility information; and a statistic derived from current account balances, representing

financial indicators of the counterparty, here in after referred to as the *liquidity indicator*. Collateral-related variables are not further considered in this section.

In Figure 4.2, we report the empirical distribution of the four selected risk drivers along with their relationship to the observed LGD[1]. For each driver, the left-hand subfigure reports the empirical distribution of its discrete values in the form of a bar plot, where categories are ordered in ascending order of the driver value. On top of each bar, the median LGD and the 25th and 75th percentiles are superimposed. This graphical representation allows us to assess whether LGD exhibits a systematic trend with respect to the driver. The right-hand subfigure shows the empirical distribution of observed LGD conditional on the extreme values of the driver - the minimum and the maximum values - to highlight differences in behaviour between the two ends of the scale.

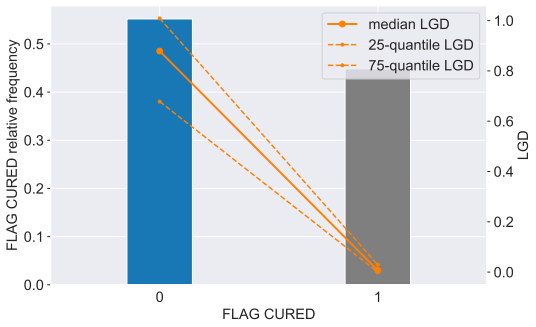
Starting from the *flag cured* (Figures 4.2a and 4.2b), this binary variable indicates whether the recovery process ended in a cure (value = 1) or a non-cure (value = 0). In the dataset, about 55% of facilities are classified as non-cured and 45% as cured. The bar plot in Figure 4.2a shows that the median LGD drops dramatically from approximately 0.5 in the case of non-cured exposures to nearly 0.005 for cured ones. Moreover, for cured exposures the interquartile range is extremely narrow, indicating that LGD values are consistently very low, as expected by definition. Conversely, non-cured exposures show a much wider interquartile range, reflecting the fact that this group includes both partial recoveries and complete losses. This interpretation is confirmed by the LGD histograms in Figure 4.2b: cured exposures concentrate almost entirely near zero, while non-cured exposures are spread across the full LGD range, with peaks near both 0 and 1; the peak near 0 is very small, corresponding to non-cured defaults that nevertheless resulted in very limited or almost negligible losses.

Turning to the *EAD* (Figures 4.2c and 4.2d), the discretisation process produced 11 distinct values ranging from 150 to 12.811. Approximately 40% of observations have the minimum exposure ($EAD = 150$), while the remainder are distributed across the other classes, with slight over-representation at 1.792 and 12.811. Figure 4.2c shows a generally monotonic increasing relationship between EAD and the median LGD, with higher exposures tending to be associated with larger relative losses. Two intermediate categories (2.544 and 2.959) deviate slightly from this trend, displaying unexpectedly high median LGD values, yet the overall increasing pattern remains clear. This result is economically intuitive: lower exposures are easier to recover, whereas very high exposures are more likely to reflect financial conditions that hinder recovery.

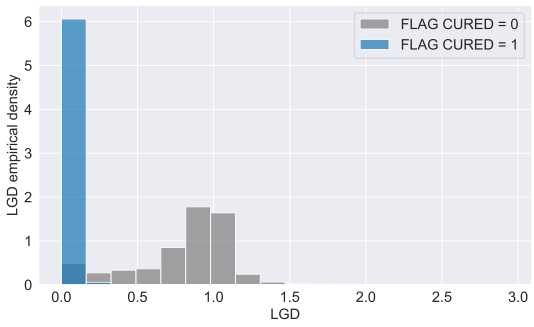
While the median shifts across classes, the 25th and 75th percentiles remain largely unchanged, resulting in a stable interquartile range. This indicates that the bimodality of the LGD distribution persists: **the relative weight of the two modes changes, causing the median to move, but no mode disappears and the interquartile range remains essentially unaffected**. Figure 4.2d further illustrates this pattern: for $EAD = 150$, the empirical LGD distribution is heavily skewed towards zero (with the mode at zero nearly three times higher than that at one), whereas for $EAD = 12.811$ the two modes at 0 and 1 are almost equally frequent, and partial recovery outcomes become more common.

Finally, the same type of considerations can also be drawn for the other two drivers,

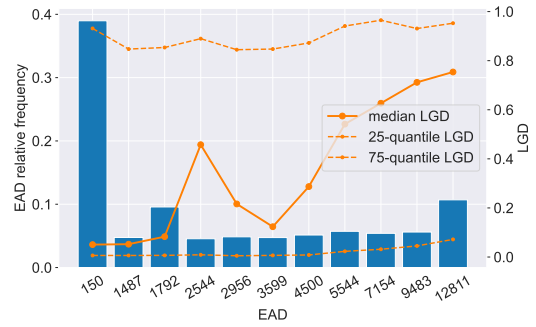
[1]It is important to note that all drivers in the dataset were already discretized prior to this analysis. This discretization reflects the format in which the data were originally provided, rather than a preprocessing step performed here. As mentioned in Section 1.3, discretization of all risk drivers is required to obtain a grid-based output. This aspect is relevant because, even though the proposed methodologies go beyond a grid-based output and this discretization requirement, it is not possible to evaluate the potential benefit of maintaining sufficient input granularity, since the raw data are already aggregated.



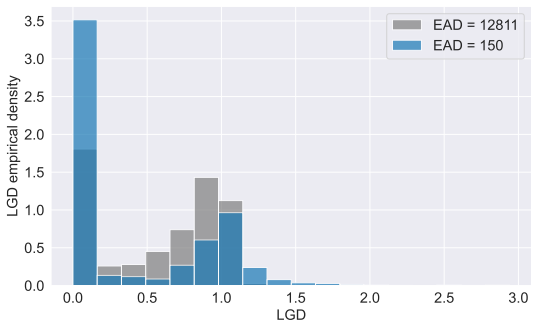
(a) Distribution of the *cured flag* with median and interquartile LGD.



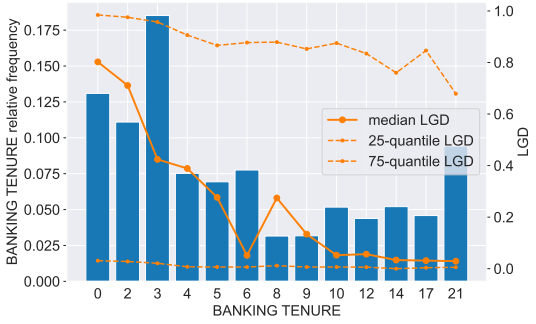
(b) Empirical LGD distribution for cured and non-cured exposures.



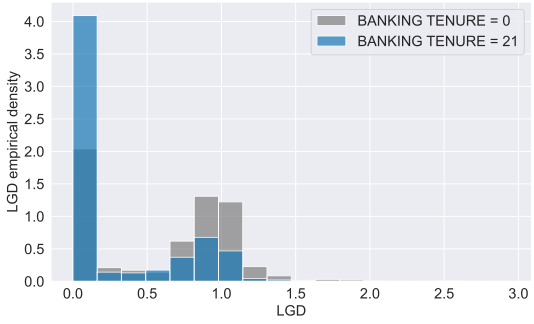
(c) Distribution of EAD categories with median and interquartile LGD.



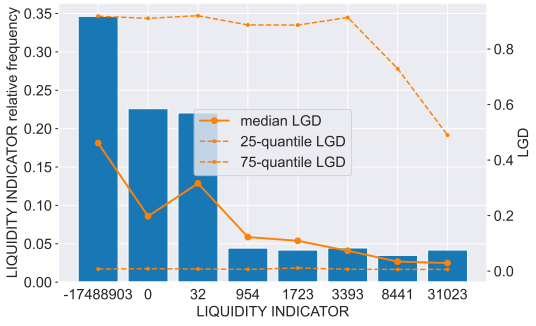
(d) Empirical LGD distribution for minimum and maximum EAD.



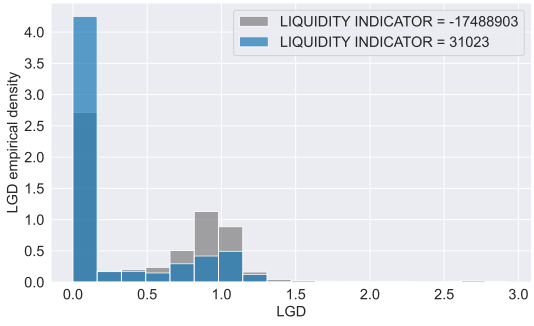
(e) Distribution of banking tenure with median and interquartile LGD.



(f) Empirical LGD distribution for minimum and maximum banking tenure.



(g) Distribution of the liquidity indicator with median and interquartile LGD.



(h) Empirical LGD distribution for minimum and maximum liquidity indicator.

Figure 4.2: Exploratory analysis of the relationship between observed LGD and selected risk drivers: *flag cured* (a–b), *EAD* (c–d), *banking tenure* (e–f) and *liquidity indicator* (g–h). Each row corresponds to one driver. The left panel shows its empirical distribution together with median and interquartile LGD, while the right panel shows the empirical LGD distribution conditional on its extreme values.

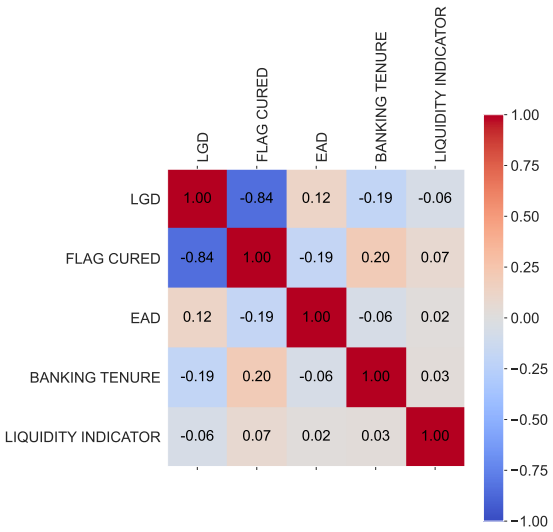


Figure 4.3: Correlation matrix between the selected features and LGD.

namely *banking tenure* (Figures 4.2e and 4.2f) and the *liquidity indicator* (Figures 4.2g and 4.2h). In both cases, observed median LGD tends to increase as the underlying economic condition captured by the driver deteriorates: LGD rises as the banking tenure decreases, and likewise LGD increases when the liquidity indicator decreases. These patterns are consistent with the economic intuition that shorter relationships with the bank and lower levels of liquidity are associated with weaker creditworthiness and more severe losses in the event of default.

In conclusion, in Figure 4.3 we analyse the correlation between these four features and the observed LGD. The observations made previously are confirmed: the sign of the correlation coefficient is consistent with the patterns shown in Figure 4.2. However, apart from a strong negative correlation between LGD and the *flag cured* equal to -0.84, the other correlations in absolute value remain low: 0.12 with *EAD*, -0.19 with *banking tenure* and -0.06 with the *liquidity indicator*. Finally, we observe generally weak correlations among the features themselves, except for a moderate correlation of 0.20 between *banking tenure* and the *flag cured*, suggesting a linear relationship between cured exposures and those with longer banking tenure.

4.2 Proposed unconstrained methodological evolution

As anticipated in Section 1.3, the proposed methodologies are compared with a baseline model consisting of a highly parsimonious Decision Tree that, controlling the number of terminal leaves and the tree depth, naturally produces a grid-like output. In this part, we relax the parsimony constraints, with the goal of improving predictive accuracy at the expense of the grid representation. Two model designs are considered: single-stage and two-stage.

The single-stage models, presented in Subsection 4.2.1, include the natural non-parsimonious extensions of the baseline: an *unconstrained* Decision Tree, a Random Forest and an Extreme Gradient Boosting (XGBoost) model.

The two-stage models, described in Subsection 4.2.2, consist of two models that differ in the segmentation performed in the first stage (see Section 2.3). The first two-stage model divides the sample according to whether the observed LGD is above or below a

given threshold; we refer to this as the *two-stage threshold model*. The second instead splits observations based on whether the recovery process ends in cure or non-cure, which we denote as the *two-stage flag cured model*.

4.2.1 Single stage methods

From an implementation perspective, the single-stage models were developed using python standard machine learning routines:

1. The `DecisionTreeRegressor` from `scikit-learn` was employed. The hyperparameters `max_depth` and `max_leaf_nodes` were tuned, while `min_samples_leaf` was fixed to a constant value determined according to the size of the in-sample dataset.

2. The `RandomForestRegressor` from `scikit-learn` was used, tuning the hyperparameters `n_estimators`, `max_depth` and `max_leaf_nodes`. Both `min_samples_leaf` and `max_features` were fixed: the former following the same rationale as in the Decision Tree case (i.e., based on the size of the in-sample dataset), and the latter set to `'sqrt'`, so that the number of features considered at each split equals the square root of the total number of features.

3. The `XGBRegressor` from `xgboost` was utilised. The tuned hyperparameters included `max_depth`, `max_leaf_nodes`, `n_estimators`, `colsample_bytree`, `subsample` and `learning_rate`. In addition, the `base_score` parameter was set equal to the in-sample average LGD.

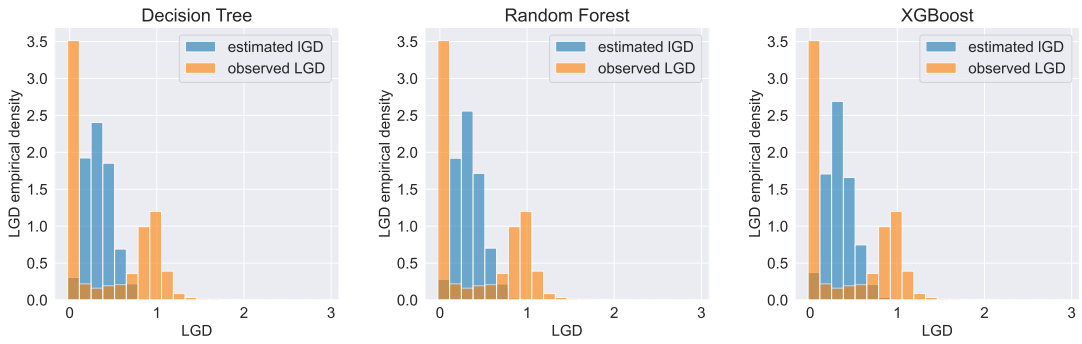
Hyperparameter tuning was performed through a *sequential* random search strategy, implemented with `RandomizedSearchCV` combined with a `ShuffleSplit` cross-validation scheme, both in `scikit-learn`. The procedure began with broad ranges of admissible values for each hyperparameter and an initial random exploration within these ranges. Based on the results, the search region was progressively narrowed to the most promising intervals, where additional random searches were carried out. This sequential strategy improves computational efficiency by allocating resources to regions of the parameter space where high-performing combinations are more likely to be found.

For each candidate tuple of hyperparameter values, the average Somers' d across the validation folds was computed. The optimal hyperparameter configuration was identified as the one maximising this metric, subject to the condition that the difference between training and validation performance did not exceed a predefined overfitting threshold.

The `ShuffleSplit` scheme was preferred to the standard k-fold approach because it reassigns observations to training and validation sets both across candidate tuples and at each CV iteration. This repeated randomisation reduces variance and provides more robust parameter selection, especially in the context of iterative search refinements.

After completing the tuning phase, feature importance was computed in order to remove irrelevant risk drivers. The model was then retrained using the optimal hyperparameters and only the most relevant features. With this refined specification, we obtained final LGD predictions on the in-sample, out-of-sample, and out-of-time datasets, and evaluated predictive performance in terms of Somers' d.

In Figure 4.4, the histograms of predicted LGDs (in blue) are compared with the observed distribution (in orange) for each of the three proposed models. For confidentiality reasons, the corresponding histogram for the baseline model cannot be reported. In any case, **both the baseline and the proposed models fail to capture the bimodality of the LGD**



(a) Decision Tree (b) Random Forest (c) XGBoost

Figure 4.4: Histograms of predicted LGDs (blue) versus observed LGDs (orange) for the three single-stage models.

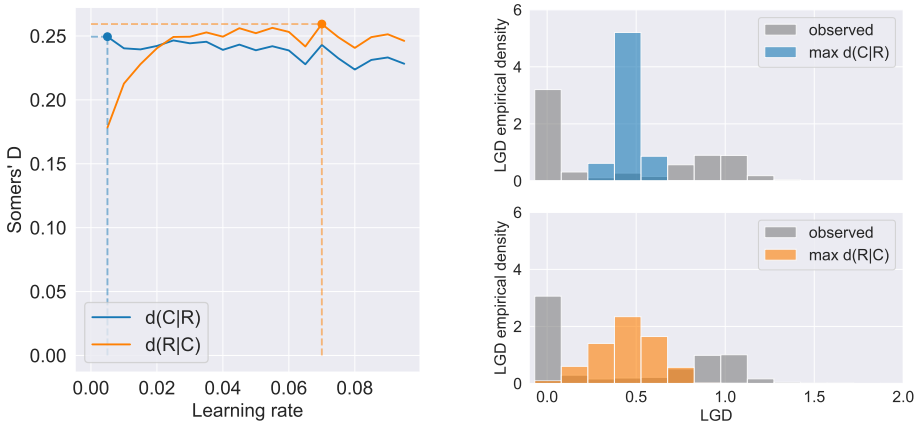
distribution. In particular, the resulting predicted distributions provided by all the models are unimodal with the mode concentrated around the average LGD value.

This behaviour can be explained, for example in the case of the Decision Tree, by the presence of terminal leaves that group together exposures with highly heterogeneous LGD values. The prediction for each leaf, being the average of these diverse observations, therefore tends to approximate the overall sample mean, as also visible from the histogram. As a result, most predicted values concentrate around this mean, revealing the model's limited ability - within the depth and split constraints imposed to avoid overfitting - to capture non-linearities and relationships that would enable a finer discrimination of LGD levels.

Despite this limitation, **we still observe an improvement in terms of predictive power compared with the baseline**. Specifically, the accuracy ratio (AR) increases from 22.3% for the baseline model to 24.84% for the Decision Tree, 25.50% for the Random Forest, and 25.99% for XGBoost (see Table 4.2).

It is worth stressing that, as prescribed by the regulatory guidelines [9] and recalled in Chapter 3, **the performance metrics reported in Table 4.2 and used to compare models refer to Somers' $d(C|R)$. However, during the hyperparameter tuning phase both versions of Somers' d - namely $d(C|R)$ and $d(R|C)$ - were considered.** Specifically, for the Decision Tree and the Random Forest models, only $d(C|R)$ was used both to evaluate their performances and to select the optimal hyperparameters. In contrast, for XGBoost only the use of $d(R|C)$ led to meaningful hyperparameter configurations. As discussed in Section 3.4, relying exclusively on $d(C|R)$ to determine parameters, such as the learning rate, may favour very slow learning, where the estimates remain overly coarse and concentrated around the base score.

To illustrate this issue empirically, Figure 4.5 presents an example where only the learning rate of XGBoost is tuned while monitoring both metrics. In Figure 4.5a, the orange curve shows that $d(C|R)$ decreases as the learning rate increases within the displayed range, reaching its maximum at the smallest value of the learning rate. Conversely, $d(R|C)$ (blue curve) reaches its minimum at this point and attains the maximum around 0.07. As a consequence, relying on $d(C|R)$ would suggest selecting an extremely small learning rate, resulting in overly "lazy" predictions that remain close to the base score, as shown in the top histogram of Figure 4.5b. In contrast, using $d(R|C)$ points to a more appropriate learning rate, producing less concentrated predictions, as illustrated in the bottom histogram of Figure 4.5b.



(a) Accuracy ratio measured with $d(C|R)$ (orange) and $d(R|C)$ (blue) as a function of the learning rate.

(b) Histograms of predicted LGDs under the two tuning criteria: using $d(C|R)$ (top) vs. using $d(R|C)$ (bottom).

Figure 4.5: Illustration of the pitfalls arising when XGBoost hyperparameters are tuned using only Somers' $d(C|R)$. Panel (a) shows the divergent behaviour of the two accuracy ratios as the learning rate varies. Panel (b) highlights the impact on the distribution of predicted LGDs, where $d(C|R)$ produces overly concentrated estimates close to the base score, while $d(R|C)$ yields more informative predictions.

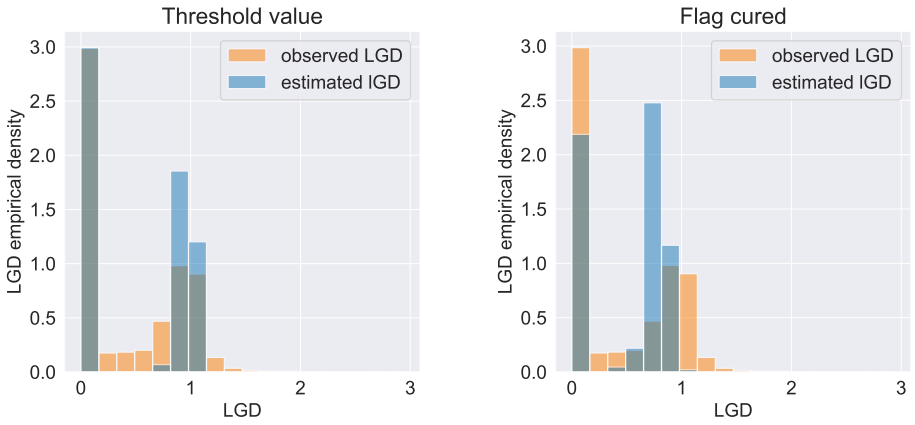
4.2.2 Two-stage methods

Two-stage models are inherently more complex than single-stage models, as they consist of multiple components: a first-stage classifier g and one or more second-stage regressors h, one for each considered class (see Figure 2.2). In this work, we employed tree-based models for both stages: `DecisionTreeClassifier` at the first stage and `DecisionTreeRegressor` instances at the second stage. Specifically, we consider for both models only two classes: for the two-stage threshold model, the two classes correspond to above-threshold and below-threshold observations; for the two-stage flag-cured model, the classes represent cured and non-cured exposures.

It is crucial to note that neither the threshold indicator nor the cured flag is directly observable; they must be predicted and this is exactly how two-stage models work. The classifier assigns a predicted class to each observation and the second-stage regressors are applied conditionally on this predicted class. Each regressor is trained exclusively on the observations belonging to its corresponding class (*observed* class) and, at prediction time, the appropriate regressor is chosen based on the first-stage classification output (*predicted* class).

Hyperparameter tuning follows the same methodology as described in Subsection 4.2.1, including sequential random search with shuffle split cross-validation and subsequent feature importance analysis to remove irrelevant risk drivers.

As shown in Figure 4.6, **the predicted distributions produced by the two-stage models successfully reflect the bimodality of the observed LGD**. However, several limitations remain. First, we observe an excessive polarization of predictions, which prevents the model from accurately capturing observations near the threshold or partially cured cases. Second, classification errors in the first stage can propagate, assigning an observation to the wrong class and producing an LGD estimate that may deviate substantially from the true value. This limitation is particularly relevant for risk management, which is not



(a) Two-stage threshold model.

(b) Two-stage flag-cured model.

Figure 4.6: Histograms of predicted LGDs (blue) versus observed LGDs (orange) for two-stage models.

Model	**AR (%)**
Baseline	22.30
Decision Tree	24.84
Random Forest	25.50
XGBoost	25.99
Two-stage Threshold	22.90
Two-stage Flag Cured	24.28

Table 4.2: Accuracy Ratios (AR), i.e., Somers' d, for baseline, single-stage, and two-stage models.

present in single-stage models, but at the cost of ignoring bimodality.

These issues could be partially addressed by increasing the number of classes, which would reduce polarization and improve granularity. However, doing so may result in insufficient data for some classes, raising learnability and overfitting concerns as discussed in Chapter 2. Another possible improvement consists of considering more powerful classifier as first stage model such as Random Forest classifier and XGBoost which could improve class prediction reducing misclassification.

In Table 4.2, we report the performance of the single-stage models, two-stage models and the baseline. Specifically, we observe that **two-stage models outperform the baseline but are less performant than the single-stage models, with XGBoost achieving the highest performance overall**.

In summary, **two-stage models provide a better representation of LGD bimodality compared to single-stage models, but they introduce challenges related to first-stage classification errors, prediction polarization and potential data scarcity when further increasing class granularity**. These factors negatively affect their overall performance, making them less effective than single-stage models; however, they still offer substantial space for improvement through more accurate classifier or a different and more suitable class design.

Also in this case, **during model development it was necessary to use both forms of Somers' d to avoid producing "lazy" estimates while defining an important parameter:**

the cut-off threshold used by the classifier. Specifically, a `DecisionTreeClassifier` assigns the predicted class to each leaf based on the composition of the observations contained in that leaf. In a standard binary classification setting, the default cut-off is 50%, meaning that if more than 50% of the observations in a leaf belong to a particular class, that leaf is assigned this class.

In the case of the two-stage threshold model, the cut-off refers to the minimum proportion of observations "above threshold" required for the leaf to be assigned the "above threshold" class; we denote this quantity as $P(\text{above})$. For the flag cured model, the cut-off represents the minimum proportion of cured observations necessary for the leaf to be assigned the cured class, denoted as $P(\text{cured})$.

In Figure 4.7 the effect of the cut-off threshold on accuracy ratio is illustrated for both two-stage models: the threshold model on the top-left Figure 4.7a and the flag cured model on the bottom-left Figure 4.7c. In neither case does the default 50% cut-off perform adequately. For the threshold model, a $P(\text{above})$ of 50% would yield a very high $d(C|R)$, almost 30%, but a much lower $d(R|C)$, below 15%. This occurs because such a threshold produces "lazy" predictions forcing the model to assign most observations to a single class, inflating $d(C|R)$ as already discussed.

In fact, in the top-right panel Figure 4.7b, we show the predicted LGD distributions obtained both when using a cut-off that maximizes $d(C|R)$ (between 50% and 60%) and when using a cut-off that maximizes $d(R|C)$ (between 30% and 40%). As anticipated, in the first case the predictions are "lazy" and extremely concentrated near zero. In the second case, predictions covers the entire range of values and are less concentrated.

Specifically this happens because, requiring a stronger majority for the above-threshold class, already under-represented in the dataset, effectively combines two effects: it makes it even less likely for an observation to be classified as above-threshold, and it inflates $d(C|R)$ by concentrating predictions in the below-threshold class.

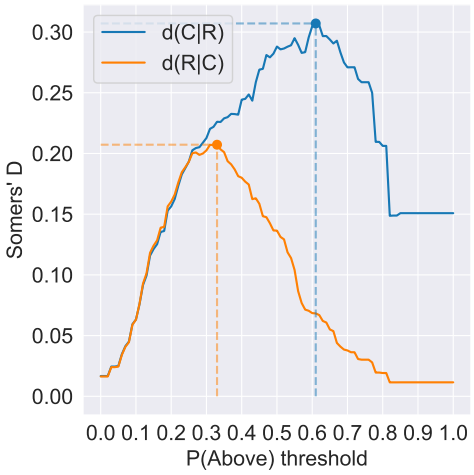
A similar mechanism occurs for the $P(\text{cured})$ in the flag cured model. Maximizing $d(C|R)$ would lead to an artificially high accuracy ratio (up to 35%) simply because the threshold is set lower than 50% and cured observations are more numerous. As a result, the majority of observations are assigned to the cured class, producing "lazy" predictions that overstate the model's apparent discriminative power.

In order to avoid this issue, literature suggests adjusting the cut-off to reflect the class proportions in the in-sample dataset, balancing the evidence required within each leaf. In both cases, the cut-off that maximizes $d(R|C)$ is close to this theoretical value, which is why we selected it for the final models.

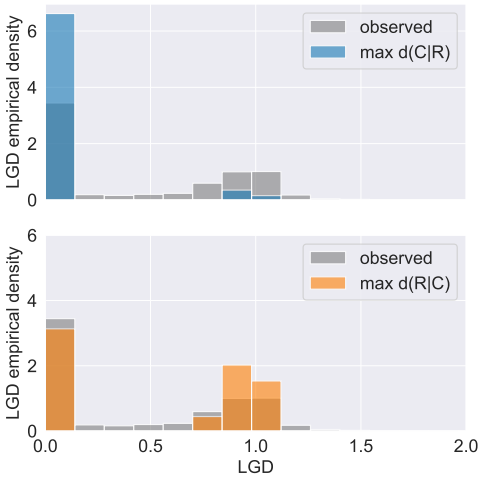
4.3 Post processing aggregation

The higher performance of the proposed models compared to the baseline is not surprising. The stringent parsimony constraints imposed on the baseline, by construction, limit its predictive power. This is evident by simply comparing the number of distinct estimates produced by the models. For instance, the best-performing model, XGBoost, raises the accuracy from 22.3% (baseline) to 25.99%, but at the cost of producing 4261 distinct predicted values instead of only 22. The substantial increase in prediction granularity explains the accuracy gain of 3.69%, yet it naturally raises the question of whether such a trade-off between accuracy and parsimony should be considered acceptable.

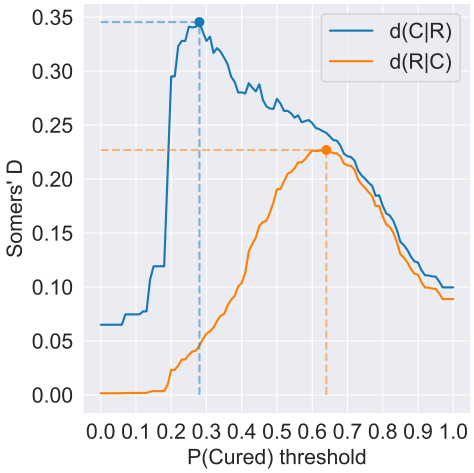
As a final methodological contribution, this work explores the possibility of reducing



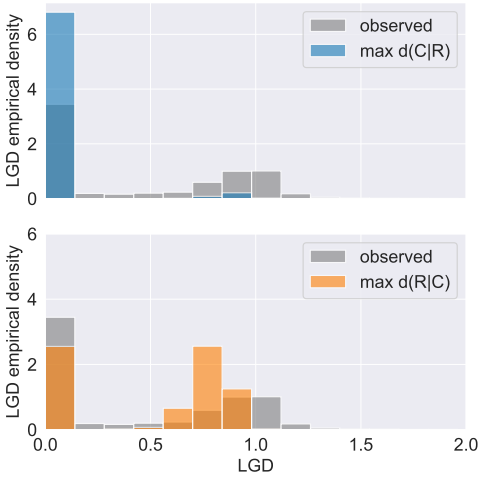
(a) $d(C|R)$ and $d(R|C)$ as a function of P(above) for the two-stage threshold model.



(b) Histograms of predicted LGDs under the two cut-off in two stage threshold: using $d(C|R)$ maximizer (top) vs. using $d(R|C)$ maximizer(bottom).



(c) $d(C|R)$ and $d(R|C)$ as a function of P(cured) for the flag cured model.



(d) Histograms of predicted LGDs under the two cut-off in two stage flag cured: using $d(C|R)$ maximizer (top) vs. using $d(R|C)$ maximizer(bottom)

Figure 4.7: Effect of the classifier cut-off threshold on two-stage models. In each row, respectively for two-stage threshold and two-stage flag cured, left panels show the impact of varying the threshold on the accuracy ratios $d(C|R)$ (orange) and $d(R|C)$ (blue). Right panels display the resulting predicted LGD distributions for cut-offs that maximize $d(C|R)$ (top) and $d(R|C)$ (bottom).

the granularity of the predictions *ex post*, i.e., after the model has been trained, rather than *a priori* by imposing parsimony constraints during training. **The key idea is to let the model grow freely and capture complex patterns without restrictions, and only afterwards simplify the predictions by aggregating them into a smaller number of distinct values.** The aim of this approach is to retain as much as possible of the accuracy gains while restoring a more parsimonious structure, although the actual impact on performance must be empirically assessed.

In order to do so, we apply techniques *inspired* by hierarchical clustering. In general, hierarchical clustering is a stepwise process that aggregates data points by progressively merging them, first combining individual points into small clusters and then gradually merging these smaller clusters into larger ones, until eventually obtaining a single cluster containing all observations. In this context, the procedure is applied to the large number of distinct LGD predictions produced by the models, with the idea of grouping similar values into a single representative one. In practice, predicted LGDs that are close to each other are merged into clusters, and each cluster is then associated with a representative value, which becomes the new estimated LGD for all the observations belonging to that cluster.

The hierarchical approach has the advantage of not requiring the final number of clusters to be fixed a priori (as in the case of k-means), but rather allows this number to be chosen ex post by monitoring how specific metrics evolve throughout the successive aggregation steps. In this way, the final number of clusters - and therefore the degree of parsimony - can be determined by balancing the trade-off between reducing the number of distinct predictions and preserving prediction quality.

In order to perform the aggregation, we adopt Ward's linkage as the merging criterion. Ward's method minimizes the increase in within-cluster variance at each step, which is particularly suitable in our context because it ensures that LGD values merged together are as homogeneous as possible. To evaluate the quality of the aggregation at different stages, we monitor both the pseudo-t^2 statistic and the cluster inertia. The pseudo-t^2 statistic measures the relative increase in total within-cluster variance that would result from a potential merge, while the inertia quantifies the total within-cluster variance if the merge were performed. These metrics allow us to assess how compact the clusters remain as the hierarchy progresses and to determine the optimal stopping point, i.e., when further merging would substantially degrade cluster quality by either significantly increasing the within-cluster variance producing an excessive relative jump in the pseudo-t^2 statistic or by resulting in an excessively high inertia.

In Figure 4.8, we provide an example of how the pseudo-t^2 and inertia metrics can be used to select the optimal number of clusters for LGD aggregation. In the top panel, the pseudo-t^2 statistic is plotted for a range of cluster numbers from 15 to 50. Peaks in the pseudo-t^2, such as the maximum at 27 (highlighted with an orange arrow), indicate merges that would produce a relatively large increase in within-cluster variance. This signals that such merges should be avoided, suggesting an good stopping point at 28 clusters, as marked by the orange dot.

The bottom panel shows the cluster inertia across the same range. Using the elbow method, a clear "elbow" appears at 28 clusters, confirming the pseudo-t^2 suggestion. Before this point - clusters from 50 to 29 - merging clusters increases the within-cluster variance gradually, but beyond the elbow - clusters from 27 to 1 - the variance rises sharply, indicating a substantial loss of cluster homogeneity. Together, these two metrics provide a robust way to determine the number of clusters that balances reducing the number of distinct LGD predictions while maintaining high prediction quality.

Once clusters are defined, each cluster is represented by its centroid, i.e., the mean of the LGD values within the cluster, reducing granularity and providing a natural and interpretable summary for all LGDs contained in the cluster.

In Table 4.3, we report the number of distinct grades and the Accuracy Ratio (AR) before and after post-aggregation. We observe that this aggregation process allows us to reduce the number of grades dramatically - for instance, XGBoost decreases from 4,261

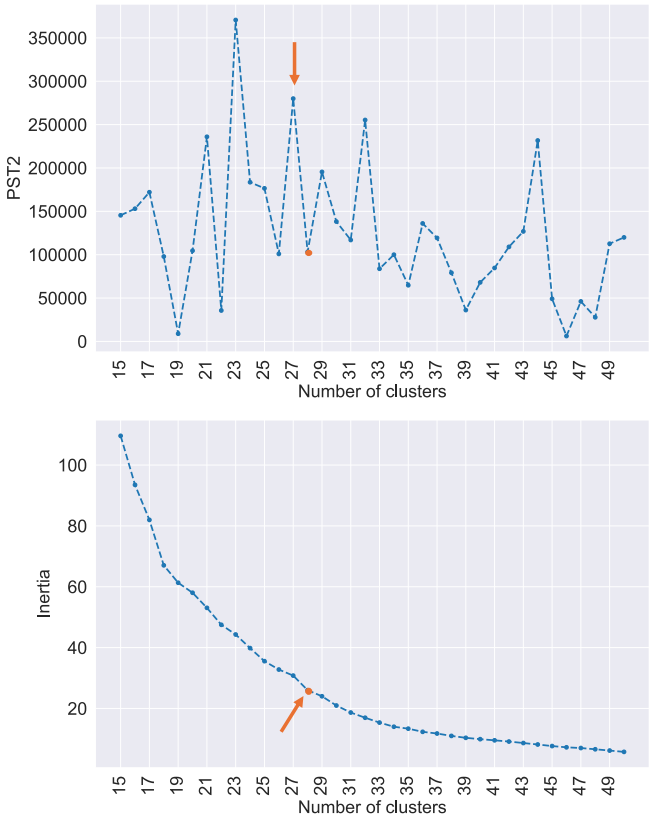
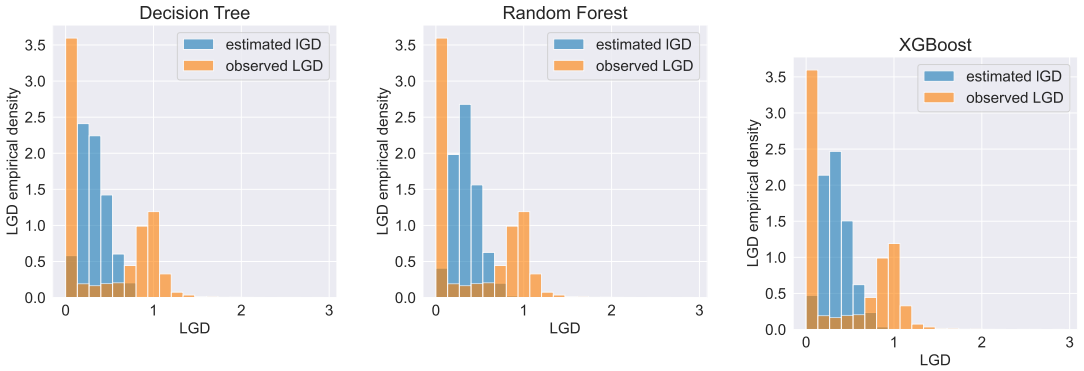


Figure 4.8: Example of using pseudo-t^2 (top) and inertia (bottom) metrics to guide the selection of the final number of clusters.

Model	**Pre-aggregation**		**Post-aggregation**	
	grades	**AR (%)**	**grades**	**AR (%)**
Baseline	22	22.30	-	-
Decision Tree	4455	24.82	24	25.03
Random Forest	4349	25.50	22	25.84
XGBoost	4261	25.99	27	25.97
Two-stage Threshold	331	22.90	22	22.46
Two-stage Flag Cured	354	24.28	26	24.32

Table 4.3: Comparison of model performance before and after post-aggregation. "Grades" indicates the number of distinct LGD predictions and AR refers to the Accuracy Ratio (Somers' $d(C|R)$).

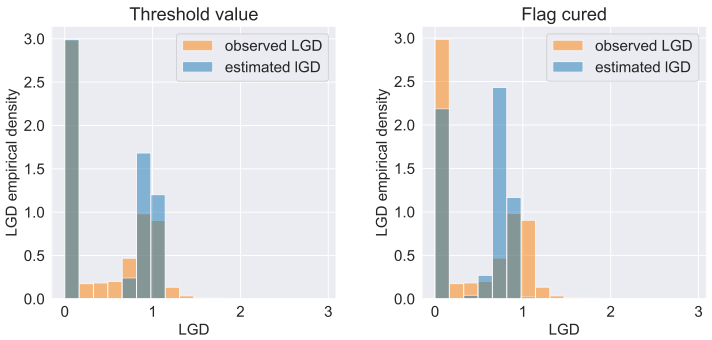
pre-aggregation to just 27 post-aggregation - while preserving nearly the same level of predictive accuracy. In addition, Figure 4.9, we show the histograms of the LGD predictions after the aggregation process compared to the observed LGDs. The key distributional features identified in the pre-aggregation estimates are maintained, indicating that the aggregation preserves the overall structure and variability of the predictions. Overall, these results demonstrate that our proposed methods enable the generation of parsimonious LGD estimates that improve upon the baseline accuracy.



(a) Decision Tree post-aggregation predictions vs. observed LGDs

(b) Random Forest post-aggregation predictions vs. observed LGDs

(c) XGBoost post-aggregation predictions vs. observed LGDs



(d) Two-stage Threshold post-aggregation predictions vs. observed LGDs

(e) Two-stage Flag Cured post-aggregation predictions vs. observed LGDs

Figure 4.9: Histograms of post-aggregation LGD predictions compared to observed LGDs for all models.

Conclusions

The central finding of this thesis is that the accuracy-parsimony trade-off can be overcome through a new paradigm consisting of an initial unconstrained estimation followed by ex-post aggregation. This was empirically illustrated by an XGBoost model, which, after being trained without constraints and subsequently simplified, achieved an Accuracy Ratio (AR) of 25.97% while being reduced to just 27 final grades. This result significantly outperforms the parsimonious baseline, which recorded an AR of 22.30% with 22 grades.

A second critical contribution concerns model validation. This work has empirically illustrated that the exclusive use of the standard Somers' $d(C|R)$ metric is inappropriate for hyperparameter tuning, as it rewards overly simplistic or "lazy" estimates, leading to the selection of suboptimal models. The analysis found that a dual-metric approach, which complements $d(C|R)$ with its reversed form $d(R|C)$, is essential for robust model selection, representing a significant methodological refinement for industry practice.

The analysis also highlighted the unique strength of two-stage models in capturing the bimodal nature of the LGD distribution, an achievement that single-stage models did not reach. Although their final AR was lower, these models offer a conceptually sound framework with significant potential for future improvement, particularly by enhancing the classifiers used in the first stage.

A key limitation of the empirical study was the use of pre-discretized risk drivers. This prevented the proposed models from leveraging their full potential in handling continuous variables, suggesting that the reported performance gains are a conservative estimate of what could be achieved with raw data.

Future research should therefore focus on three main areas: validating the methodology on continuous data to quantify its full potential; enhancing two-stage models with more advanced classifiers to combine distributional accuracy with predictive power; and promoting the adoption of the dual-metric approach to refine industry validation standards.

Acknowledgements

Desidero esprimere la mia più sincera gratitudine alla mia relatrice, la Prof.ssa Patrizia Semeraro, per la sua disponibilità e i preziosi consigli che hanno indirizzato questo lavoro di tesi.

Un ringraziamento speciale va anche al mio correlatore, il Dott. Francesco Grande, per il supporto fondamentale nella parte applicativa e per aver condiviso la sua esperienza nel settore, permettendomi di contestualizzare la ricerca in un ambito aziendale concreto. La sua visione e il suo rigore metodologico sono stati per me una fonte di grande ispirazione.

Ringrazio inoltre il Gruppo Intesa Sanpaolo per la preziosa opportunità offerta e, in particolare, la struttura Credit Risk Accelerator per avermi guidato e supportato in questo percorso, insegnandomi tanto con metodo e cura.

Un grazie sentito a Paolo, per il suo interesse sincero e il continuo confronto a tutto tondo che mi ha stimolato nel produrre un lavoro completo e accurato. Grazie anche a Riccardo, per la sua guida nel collegare l'analisi teorica al contesto applicativo, offrendo una prospettiva complementare che ha arricchito significativamente la tesi.

Un pensiero speciale va a chi è con me da sempre, i miei genitori, che con il loro incrollabile sostegno e la loro fiducia hanno permesso che tutto questo potesse realizzarsi. Grazie a Federica, la mia fidanzata, per avermi supportato in ogni momento e per essere al mio fianco; il confronto con lei è una costante fonte di ispirazione e di nuove idee.

Infine, un ringraziamento ai miei amici, Alberto, Matteo e Francesco, compagni di un percorso ormai decennale fatto di momenti leggeri e non, e con cui spero di condividere molti altri traguardi in futuro.

Appendix A

The IRB formula

In Section 1.1, the concept of capital requirement was introduced. Under the IRB approach, the risk contribution RC_i of each credit exposure i to the overall capital requirement is calculated using the IRB formula shown in Equation 1.3. The purpose of this appendix is to show how the IRB formula is derived, to interpret it and to state the underlying assumptions.

Let us start by considering a portfolio with n credit exposures, where the i-th exposure is characterised by a residual maturity T_i and an exposure at default EAD_i. Let $\tilde{\tau}_i$ and $L\tilde{G}D_i$ denote the default time and the loss given default of exposure i, respectively. We make the following assumptions:

1. The residual maturity T_i and the exposure at default EAD_i are treated as deterministic.

2. The default time $\tilde{\tau}_i$ and the loss given default $L\tilde{G}D_i$ are random variables (hence the tilde $\sim$).

3. The default time $\tilde{\tau}_i$ depends on a single systematic risk factor $\tilde{X}$, a random variable with distribution function H.

4. The random variable $\tilde{D}_i = \mathbb{1}\{\tilde{\tau}_i \leq T_i\}$ is the default indicator. Conditionally on $\tilde{X}$, $\tilde{D}_i$ is Bernoulli with conditional default probability $p_i(\tilde{X})$.

5. The loss given default is independent of the default time and of systematic risk factors.

6. The portfolio is infinitely granular, i.e. no single exposure is large enough to generate concentration risk:
$$\lim_{n\to\infty} \max_{i\in[n]} \frac{EAD_i}{\sum_{j=1}^{n} EAD_j} = 0.$$

Given these assumptions, the IRB formula can be derived in several steps: first, by obtaining a closed-form expression for the value-at-risk of the portfolio loss; second, by introducing additional assumptions inspired by Merton's model; third, by incorporating annual default probabilities PD_i; finally, by applying adjustments prescribed by the BCBS. In the following we study these steps one by one and make the related assumptions explicit.

Step 1: Closed-form expression of portfolio loss VaR The total portfolio loss is represented by the random variable $\tilde{L}$, defined as the sum of the n individual losses:

$$\tilde{L} = \sum_{i=1}^{n} EAD_i \cdot L\tilde{G}D_i \cdot \tilde{D}_i$$

Therefore, the expected loss conditioned on the systematic risk factor $\tilde{X}$ is equal to:

$$\mathbb{E}[\tilde{L} \mid \tilde{X}] = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot \mathbb{E}[\tilde{D}_i \mid \tilde{X}] = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot p_i(\tilde{X})$$

Under the assumption of infinitely granular portfolio, it can be shown that the conditional distribution of the loss $\tilde{L} \mid \tilde{X}$ collapses to its conditional expectation $\mathbb{E}[\tilde{L} \mid \tilde{X}]$ [11]. Intuitively, the absence of concentration risk and the infinitely fine-grained structure imply that idiosyncratic uncertainty associated with individual debtors vanishes, leaving only the systematic risk factor $\tilde{X}$ as relevant. Then, the cumulative distribution function of the loss $\tilde{L}$ is:

$$\mathbb{P}\left(\tilde{L} \leq l\right) = \mathbb{P}\left(\mathbb{E}[\tilde{L} \mid \tilde{X}] \leq l\right) = \mathbb{P}\left(\sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot p_i(\tilde{X}) \leq l\right) = \mathbb{P}\left(g(\tilde{X}) \leq l\right)$$

where, for simplicity, $g(\cdot)$ is defined as:

$$g(\tilde{X}) := \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot p_i(\tilde{X}),$$

Let us now assume $g(\cdot)$ to be an increasing transformation of the random variable $\tilde{X}$. Then, let us consider the value-at-risk $VaR_{1-\alpha}$ of the loss $\tilde{L}$ defined as:

$$\mathbb{P}\left(\tilde{L} > VaR_{1-\alpha}(\tilde{L})\right) = \alpha$$

hence:

$$\begin{aligned}\mathbb{P}\left(\tilde{L} \leq VaR_{1-\alpha}(\tilde{L})\right) &= \mathbb{P}\left(g(\tilde{X}) \leq VaR_{1-\alpha}(\tilde{L})\right) = \\ &= \mathbb{P}\left(\tilde{X} \leq g^{-1}(VaR_{1-\alpha}(\tilde{L}))\right) = H\left(g^{-1}(VaR_{1-\alpha}(\tilde{L})\right) = 1 - \alpha\end{aligned}$$

which leads to $VaR_{1-\alpha}(\tilde{L})$ expressed in a closed form as:

$$VaR_{1-\alpha}(\tilde{L}) = g\left(H^{-1}(1-\alpha)\right) = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot p_i(H^{-1}(1-\alpha)) \tag{A.1}$$

For completeness, in the case $g(\cdot)$ is a decreasing transformation, $VaR_{1-\alpha}(\tilde{L})$ is given by:

$$VaR_{1-\alpha}(\tilde{L}) = g\left(H^{-1}(\alpha)\right) = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot p_i(H^{-1}(\alpha)) \tag{A.2}$$

Step 2: Introduction of Merton-based assumptions In Merton's model, the default occurs when the asset value $\tilde{Z}_i$ is below a given barrier B_i: $\tilde{D}_i = 1 \iff \tilde{Z}_i < B_i$. Let us now further extend the set of assumptions as follows:

7. The asset value $\tilde{Z}_i$ is a Gaussian random variable that depends on the common risk factor $\tilde{X}$ and on an idiosyncratic risk factor $\tilde{\epsilon}_i$ as follows:

$$\tilde{Z}_i = \sqrt{\rho}\,\tilde{X} + \sqrt{1-\rho}\,\tilde{\epsilon}_i$$

where ρ is a non negative constant.

8. $\tilde{X}$ and $\tilde{\epsilon}_i$ are independent standard normal random variables

Specifically, it is easy to show that ρ is the constant asset value correlation:

$$\begin{aligned}\mathbb{E}[\tilde{Z}_i\tilde{Z}_j] &= \mathbb{E}\left[\left(\sqrt{\rho}\,\tilde{X} + \sqrt{1-\rho}\,\tilde{\epsilon}_i\right)\left(\sqrt{\rho}\,\tilde{X} + \sqrt{1-\rho}\,\tilde{\epsilon}_j\right)\right] = \\ &= \mathbb{E}\left[\rho\tilde{X}^2 + \sqrt{\rho}\sqrt{1-\rho}\,\tilde{X}\,\tilde{\epsilon}_i + \sqrt{\rho}\sqrt{1-\rho}\,\tilde{X}\,\tilde{\epsilon}_j + (1-\rho)\tilde{\epsilon}_i\tilde{\epsilon}_j\right] = \\ &= \mathbb{E}\left[\rho\tilde{X}^2\right] = \rho\left(Var(X) - \mathbb{E}[X]^2\right) = \rho\end{aligned}$$

Considering these new assumptions, the unconditional default probability p_i becomes:

$$p_i := \mathbb{P}(\tilde{D}_i = 1) = \mathbb{P}(\tilde{Z}_i < B_i) = \Phi(B_i)$$

and the conditional default probability $p_i(x)$ is equal to:

$$\begin{aligned}p_i(x) &:= \mathbb{P}(\tilde{D}_i = 1 \mid \tilde{X} = x) = \mathbb{P}(\tilde{Z}_i < B_i \mid \tilde{X} = x) = \\ &= \mathbb{P}(\sqrt{\rho}\,\tilde{X} + \sqrt{1-\rho}\,\tilde{\epsilon}_i < B_i \mid \tilde{X} = x) = \\ &= \mathbb{P}\left(\tilde{\epsilon}_i < \frac{B_i - \sqrt{\rho}\,x}{\sqrt{1-\rho}}\right) = \Phi\left(\frac{B_i - \sqrt{\rho}\,x}{\sqrt{1-\rho}}\right)\end{aligned}$$

With this new expression of conditional default probability, the function $g(\cdot)$ could be expressed as:

$$g(x) = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot p_i(x) = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot \Phi\left(\frac{\Phi^{-1}(p_i) - \sqrt{\rho}\,x}{\sqrt{1-\rho}}\right)$$

Since the obtained function $g(\cdot)$ is decreasing if $EAD_i \geq 0 \,\forall\, i$, then the value-at-risk expression to be considered is the Equation A.2, that becomes:

$$\begin{aligned}VaR_{1-\alpha}(\tilde{L}) &= g\left(H^{-1}(\alpha)\right) = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot \Phi\left(\frac{\Phi^{-1}(p_i) - \sqrt{\rho}\,H^{-1}(\alpha)}{\sqrt{1-\rho}}\right) = \\ &= \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot \Phi\left(\frac{\Phi^{-1}(p_i) + \sqrt{\rho}\,\Phi^{-1}(1-\alpha)}{\sqrt{1-\rho}}\right)\end{aligned} \tag{A.3}$$

Step 3: Incorporation of the annual default probability (PD) At this point, a new assumption is necessary:

9. The default time is assumed to be Markovian, hence we have the following relationship:

$$p_i = 1 - \mathbb{P}(\tilde{\tau}_i > T_i) = 1 - (1 - PD_i)^{T_i}$$

Finally, Equation A.3 is equal to:

$$VaR_{1-\alpha}(\tilde{L}) = \sum_{i=1}^{n} EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot \Phi\left(\frac{\Phi^{-1}\left(1 - (1 - PD_i)^{T_i}\right) + \sqrt{\rho}\,\Phi^{-1}(1-\alpha)}{\sqrt{1-\rho}}\right) \tag{A.4}$$

where the risk contribution of the exposure i is given by:

$$RC_i = EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot \Phi\left(\frac{\Phi^{-1}\left(1-(1-PD_i)^{T_i}\right) + \sqrt{\rho}\,\Phi^{-1}(1-\alpha)}{\sqrt{1-\rho}}\right) \tag{A.5}$$

Step 4: Regulatory adjustments (BCBS) In order to obtain the finalized IRB formula, the BCBS introduced the following modifications:

- The formula was simplified as:

$$RC_i \approx EAD_i \cdot \mathbb{E}[L\tilde{G}D_i] \cdot \Phi\left(\frac{\Phi^{-1}(PD_i) + \sqrt{\rho}\,\Phi^{-1}(1-\alpha)}{\sqrt{1-\rho}}\right) \cdot \varphi(M), \tag{A.6}$$

  where the conditional default probability p_i is direclty replaced by the unconditional annual probability of default PD_i, while the contribution of maturity T_i is isolated in the function $\varphi(M)$, with M denoting the effective maturity - following BCBS notation. The maturity adjustment is defined as:

$$\varphi(M) = \frac{1 + (M-2.5)\cdot b(PD_i)}{1 - 1.5 \cdot b(PD_i)},$$

  with

$$b(PD_i) = \left(0.11852 - 0.05478 \cdot \ln(PD_i)\right)^2.$$

- The confidence level α used in the value-at-risk is fixed at 0.1%.
- The asset value correlation - also referred to as default correlation - is modelled as a parametric function $\rho(PD)$ defined as:

$$\rho(PD) = 12\% \cdot \frac{1-e^{-50\cdot PD}}{1-e^{-50}} + 24\% \cdot \left(1 - \frac{1-e^{-50\cdot PD}}{1-e^{-50}}\right)$$

- The measure of credit risk is the unexpected loss, defined as:

$$UL_\alpha = VaR_\alpha(\tilde{L}) - \mathbb{E}[\tilde{L}].$$

Finally, the IRB formula is given by:

$$\begin{aligned}
&f_{\text{IRB}} = EAD_i \quad \longrightarrow \quad \text{Exposure at default} \\
&\cdot \Bigg[\mathbb{E}[L\tilde{G}D_i] \cdot \Phi\left(\frac{\Phi^{-1}(PD_i) + \sqrt{\rho(PD)}\,\Phi^{-1}(0.999)}{\sqrt{1-\rho(PD)}}\right) \quad \longrightarrow \quad \text{Expected percentage loss over one year under } \textit{extreme} \text{ conditions.} \\
&- \mathbb{E}[L\tilde{G}D_i] \cdot PD_i \Bigg] \quad \longrightarrow \quad \text{Expected percentage loss over one year under } \textit{normal} \text{ conditions.} \\
&\cdot \varphi(M) \quad \longrightarrow \quad \text{Maturity adjustment factor that extends the horizon beyond one year.}
\end{aligned} \tag{A.7}$$

Bibliography

[1] "Guidelines on pd estimation, lgd estimation and the treatment of defaulted exposures," Tech. Rep. EBA/GL/2017/16, European Banking Authority, November 2017.

[2] J. Brigando, "Loss given default: Framework normativo ed impatto sulle metodologie di intesa sanpaolo," master's thesis, Università degli Studi di Torino, 2023.

[3] M. Qi and X. Zhao, "Comparison of modeling methods for loss given default," Journal of Banking & Finance, vol. 35, no. 11, pp. 2842–2855, 2011.

[4] G. Loterman, I. Brown, D. Martens, C. Mues, and B. Baesens, "Benchmarking regression algorithms for loss given default modeling," International Journal of Forecasting, vol. 28, no. 1, pp. 161–170, 2012.

[5] A. Matuszyk, C. Mues, and L. C. Thomas, "Modelling lgd for unsecured personal loans: Decision tree approach," Journal of the Operational Research Society, vol. 61, no. 3, pp. 393–398, 2010.

[6] Y. Tanoue, S. Yamashita, and H. Nagahata, "Comparison study of two-step lgd estimation model with probability machines," Risk Management, vol. 22, no. 3, pp. 155–177, 2020.

[7] A. Salko and R. D'Ecclesia, "Decomposing loss given default: A closer look at recovery patterns," tech. rep., Sapienza University of Rome, Rome, Italy. Working paper, year unknown.

[8] J. Bosker, M. Gürtler, and M. Zöllner, "Machine learning-based variable selection for clustered credit risk modeling," Journal of Business Economics, pp. 1–36, 2024.

[9] "Instructions for institutions on the reporting for the validation of internal models for credit risk," tech. rep., European Central Bank, February 2019.

[10] R. H. Somers, "A new asymmetric measure of association for ordinal variables," American sociological review, pp. 799–811, 1962.

[11] T. Roncalli, Handbook of financial risk management. Chapman and Hall/CRC, 2020.